Regression Tress: Clearly Explained!!!

Imagine we developed
a new drug…

Imagine we developed a new drug…

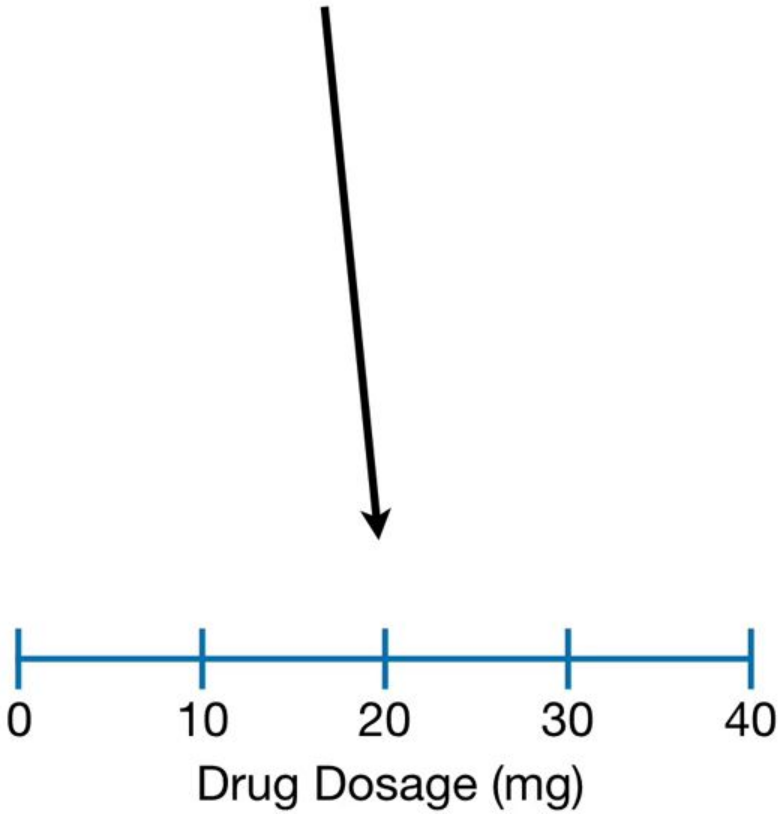
VS.

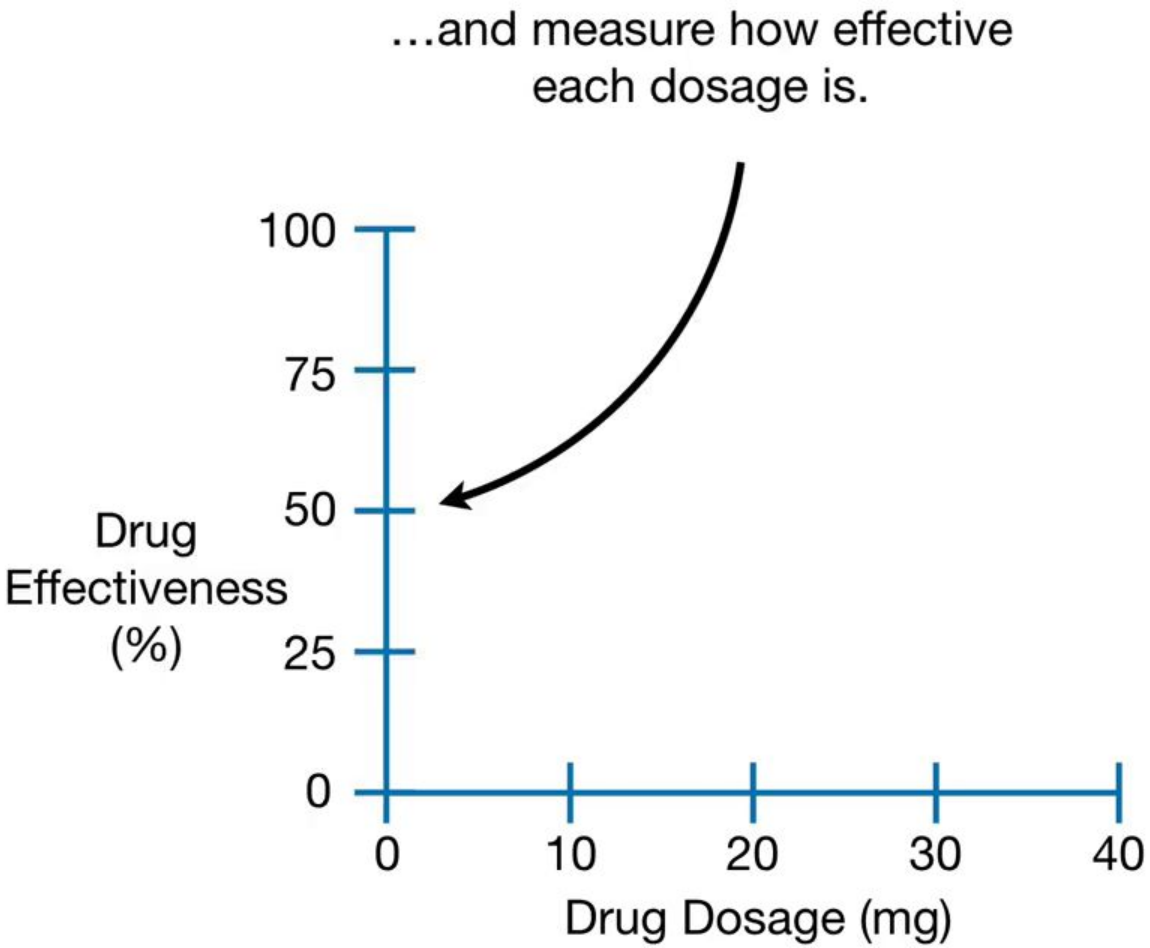
…to cure the common cold.

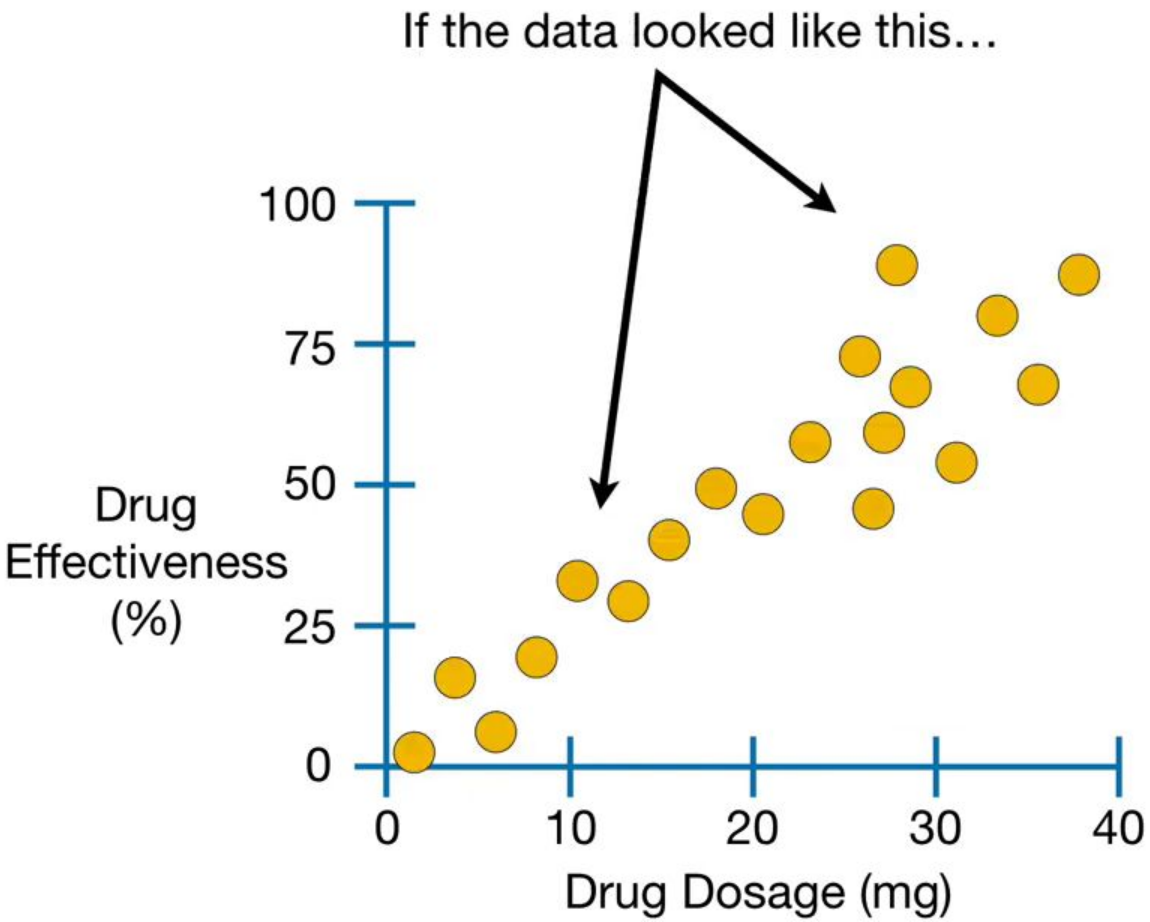
However, we don’t know the optimal dosage to give to patients.

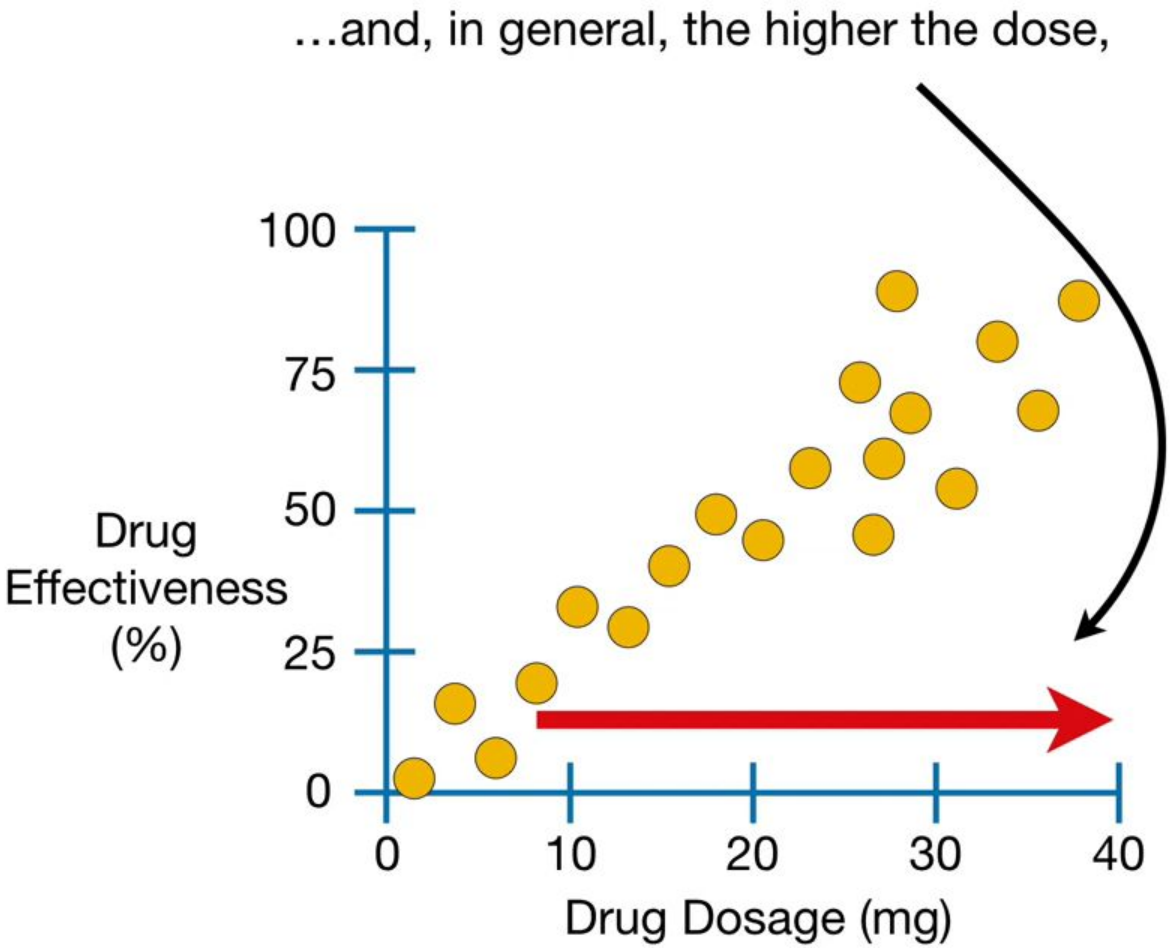
vs.

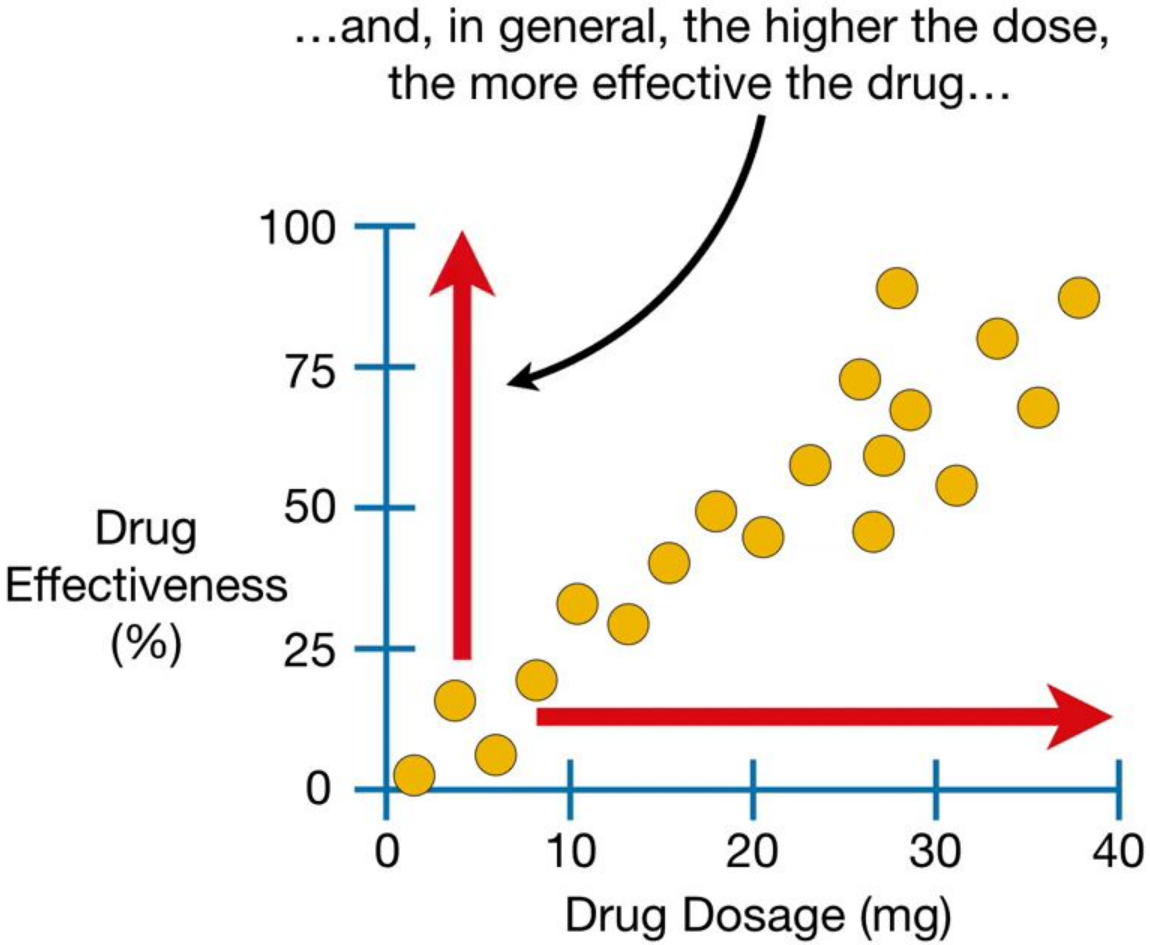
So we do a clinical trial with different dosages…

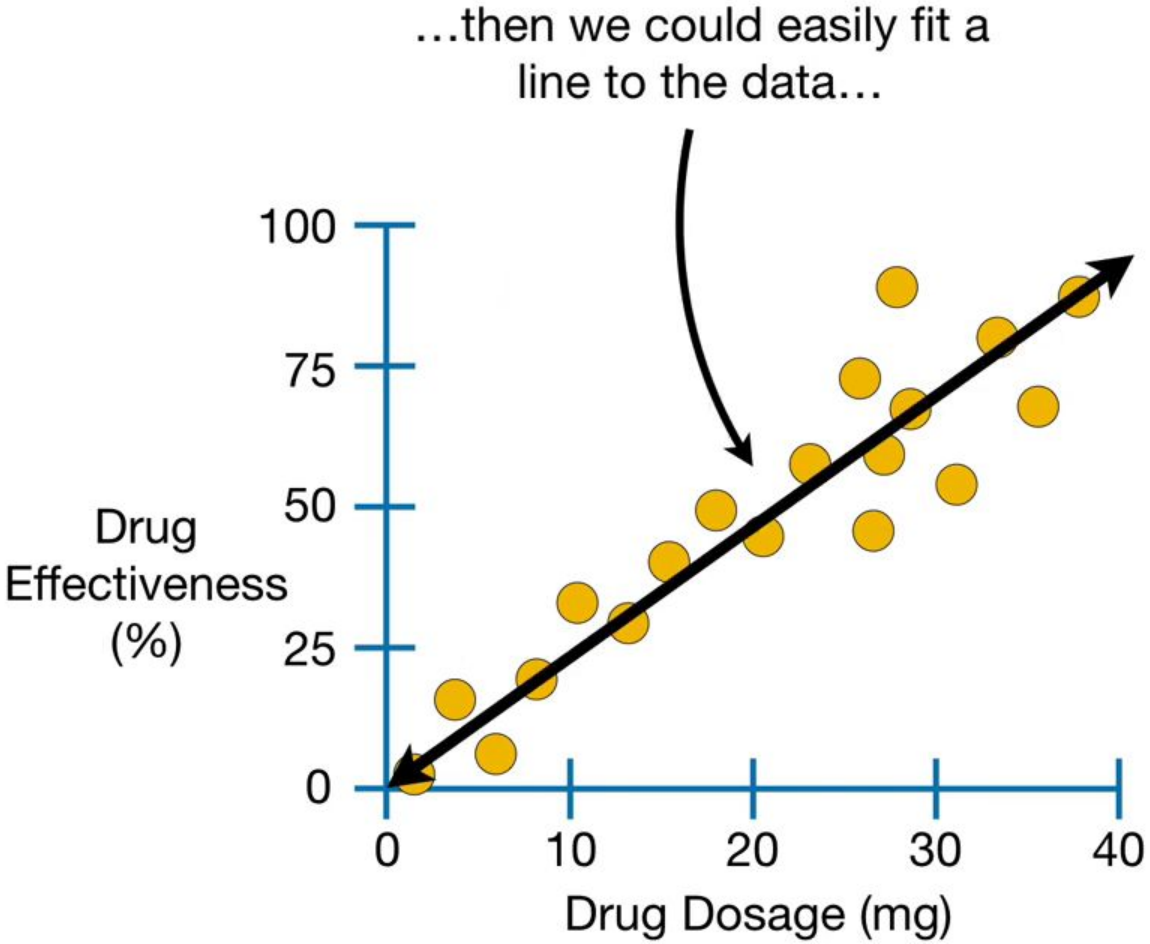


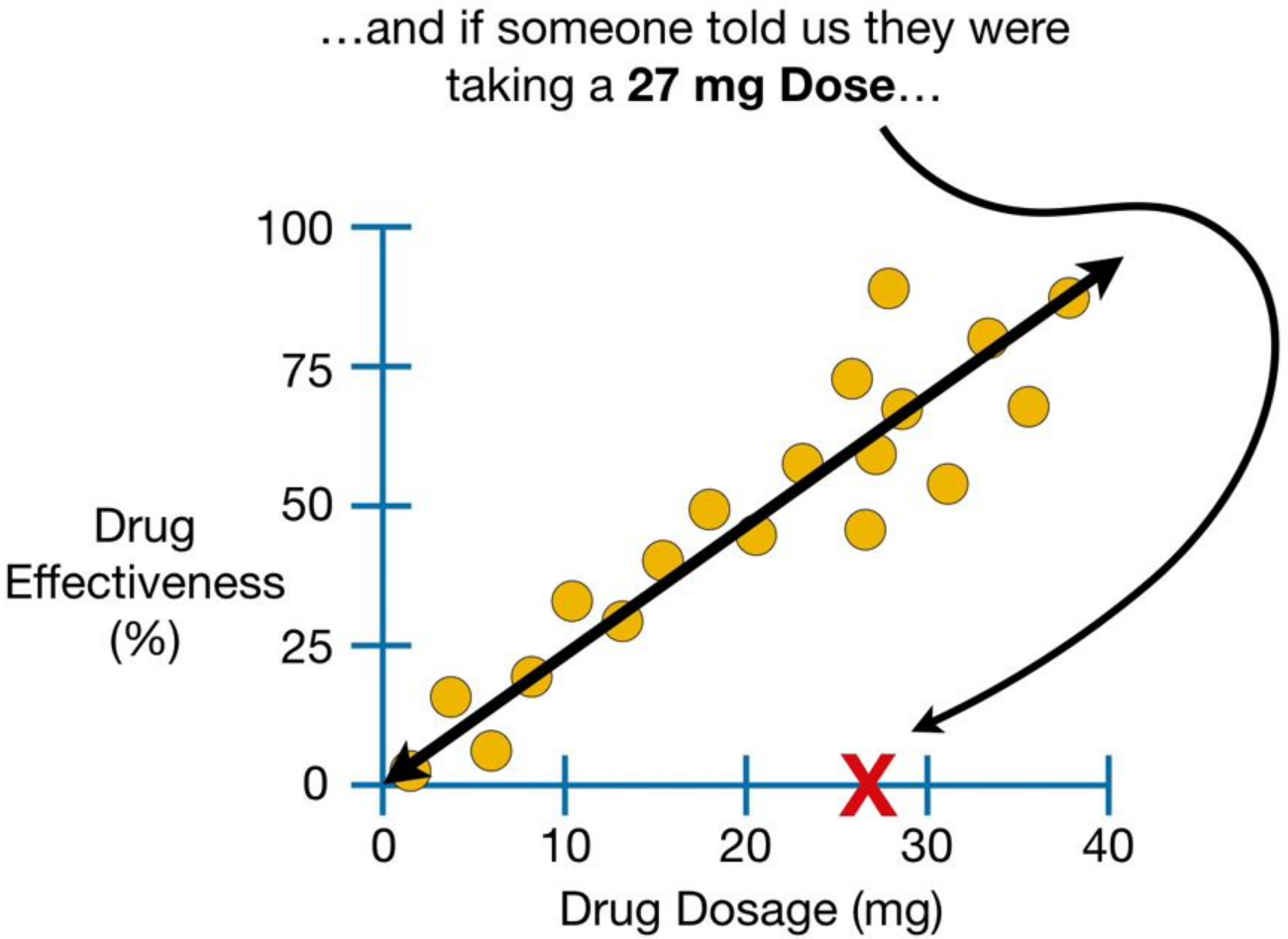

…and measure how effective each dosage is.
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)


If the data looked like this…
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)

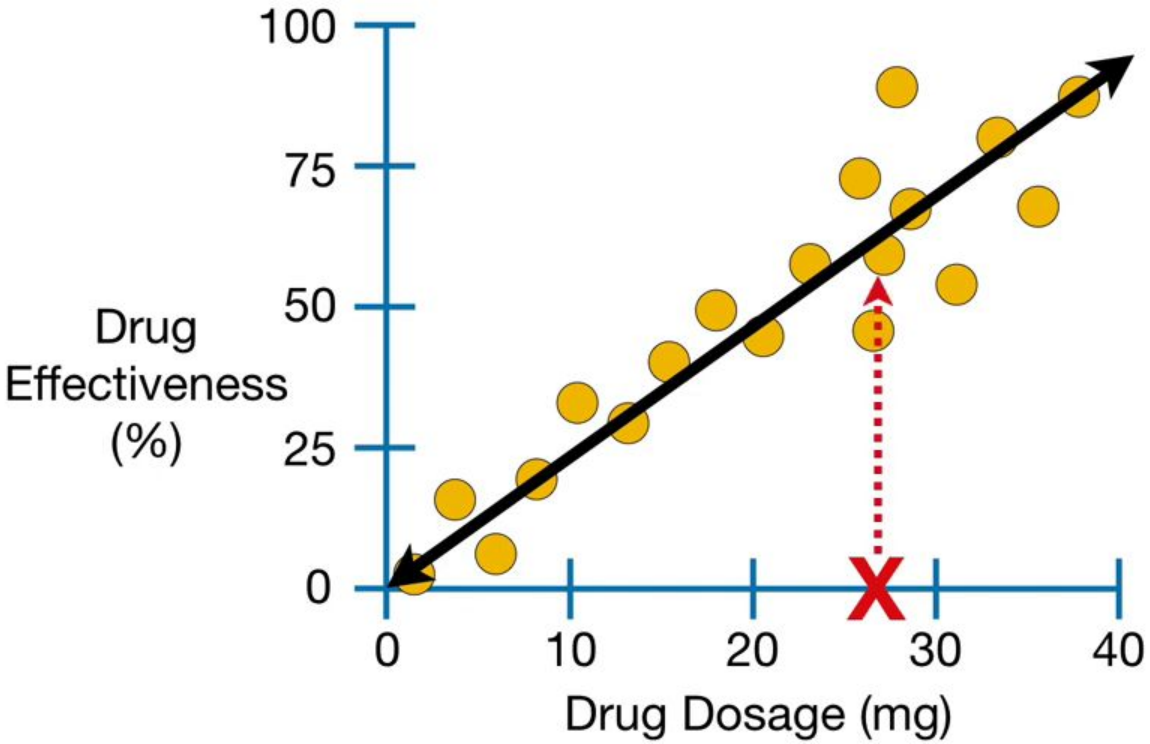

…and, in general, the higher the dose,
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)


…and, in general, the higher the dose,
the more effective the drug…
100
75
50
25
0
Drug
Effectiveness
(%)
0
10
20
30
40
Drug Dosage (mg)

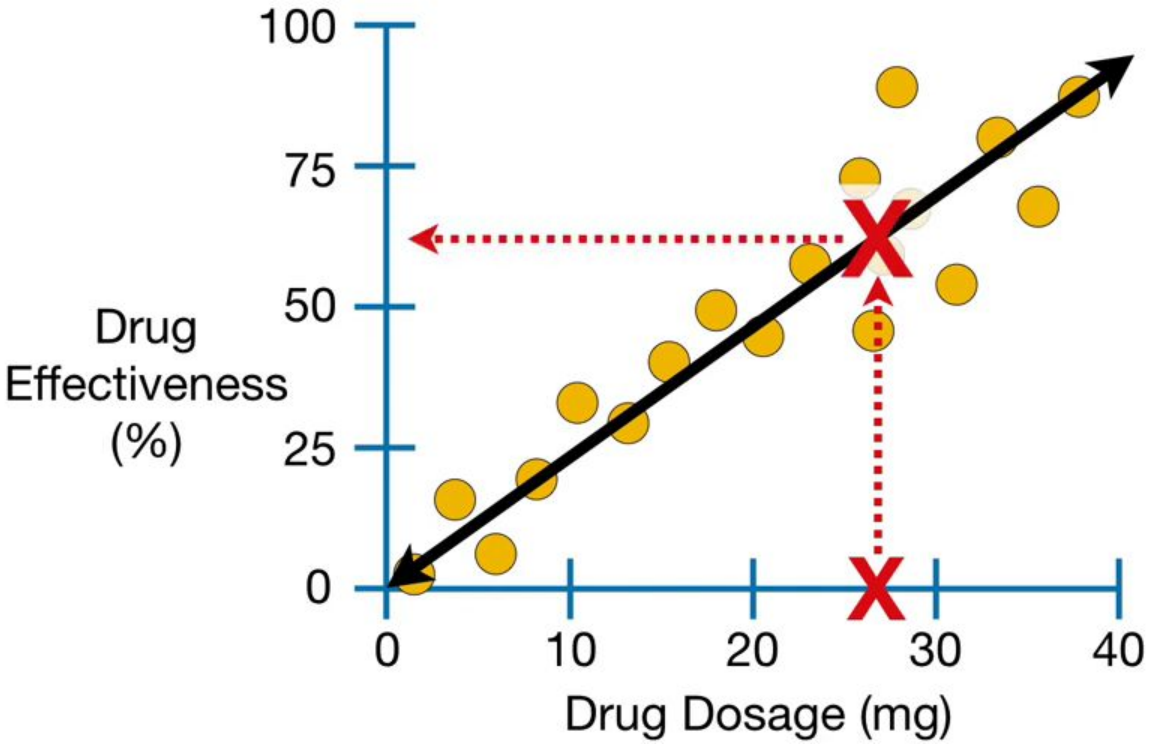

…then we could easily fit a line to the data…
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)

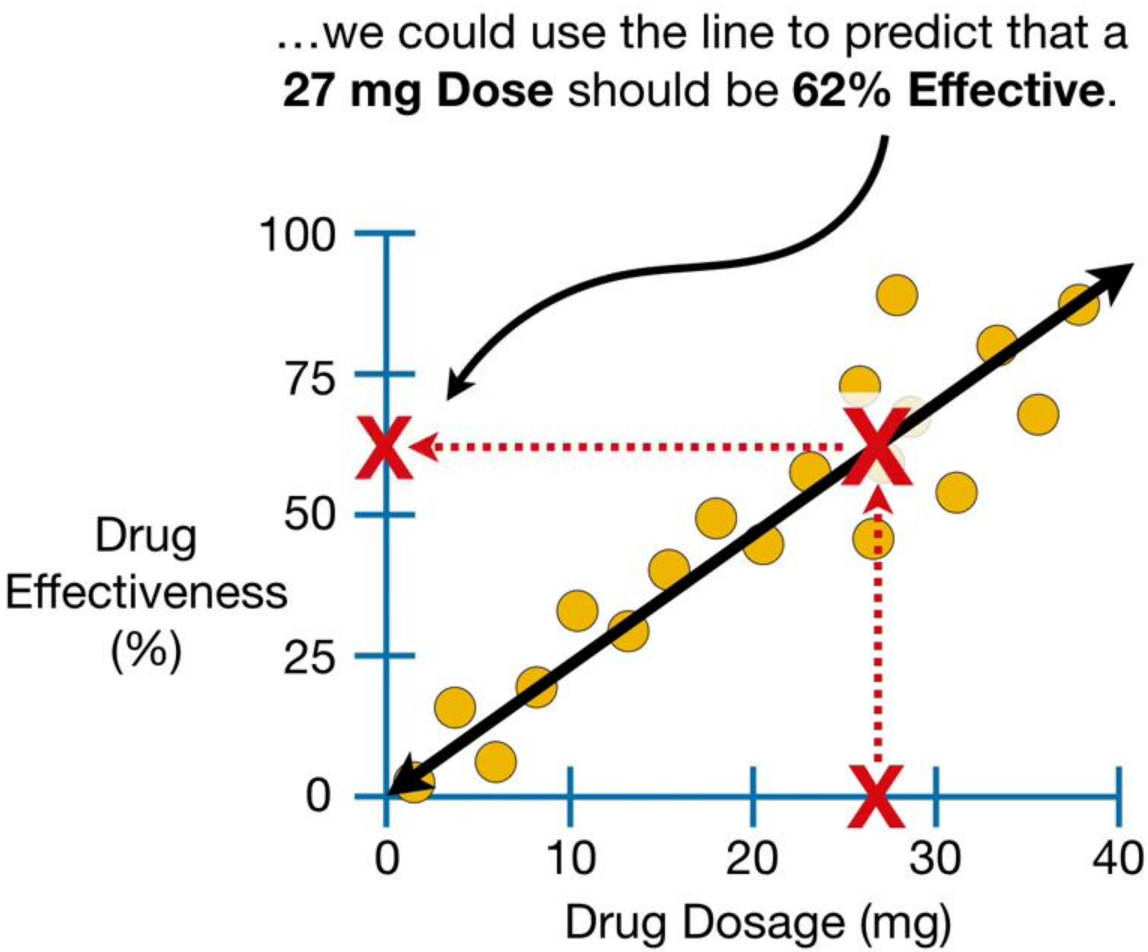

…and if someone told us they were taking a **27 mg Dose**…
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)

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

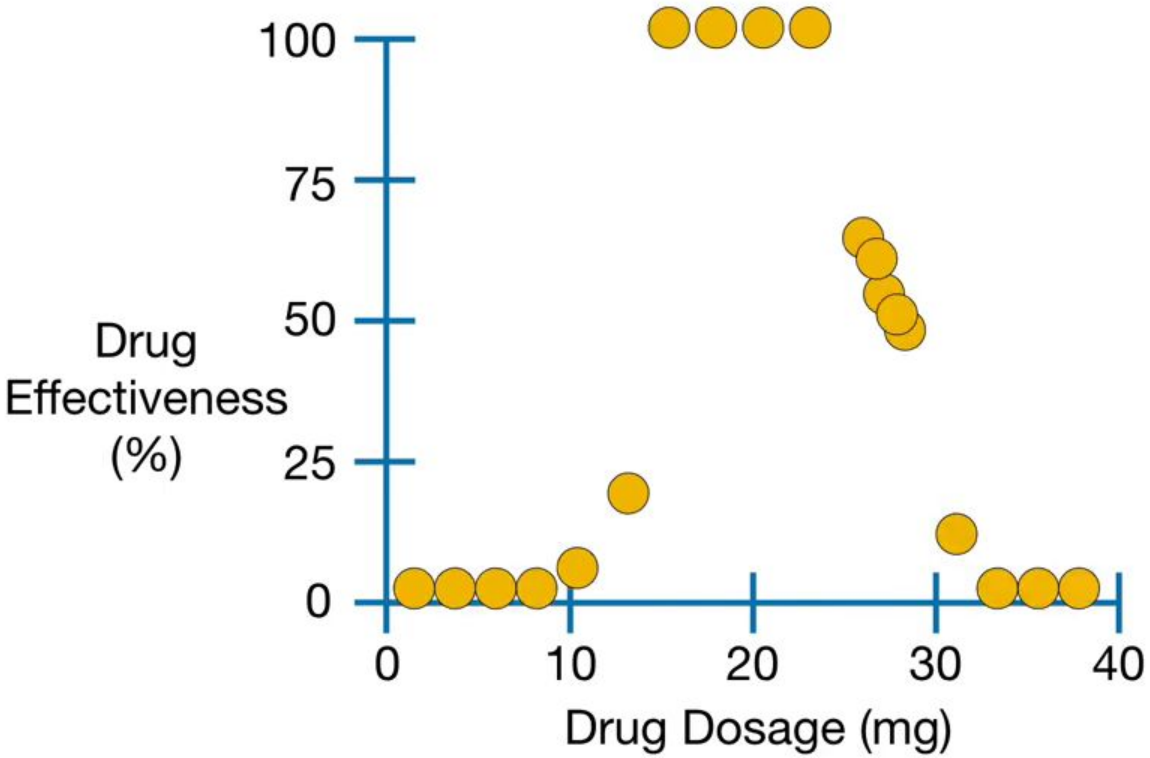


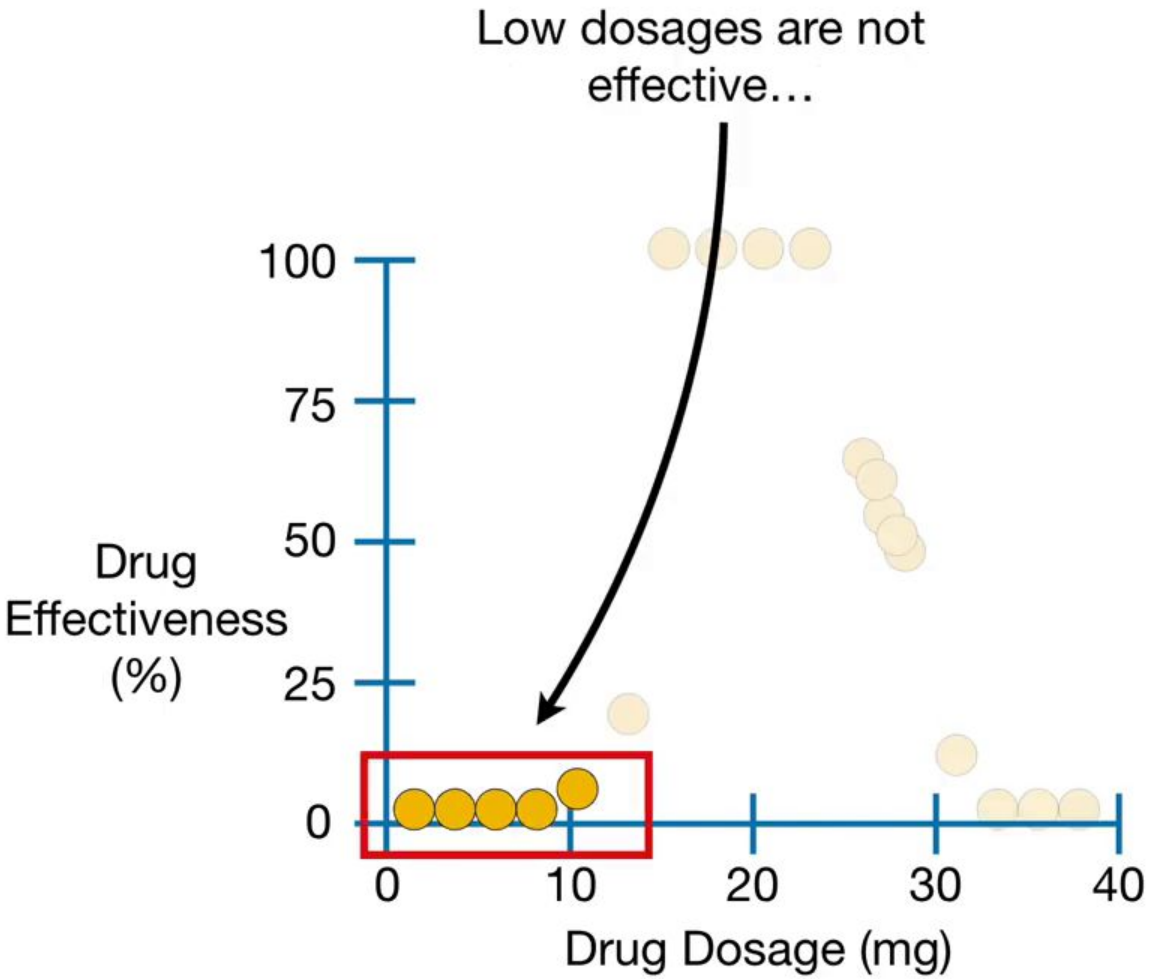
…we could use the line to predict that a **27 mg Dose** should be **62% Effective**.

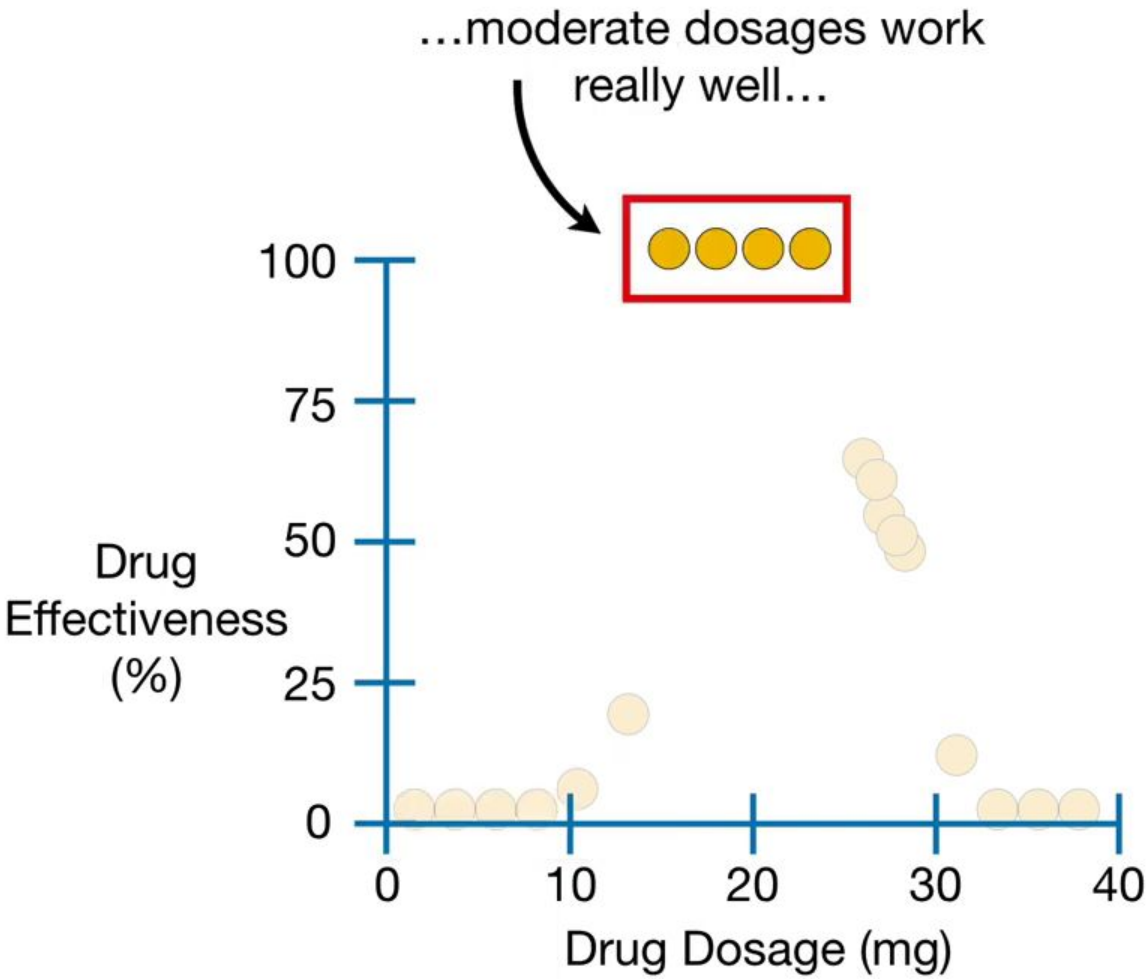


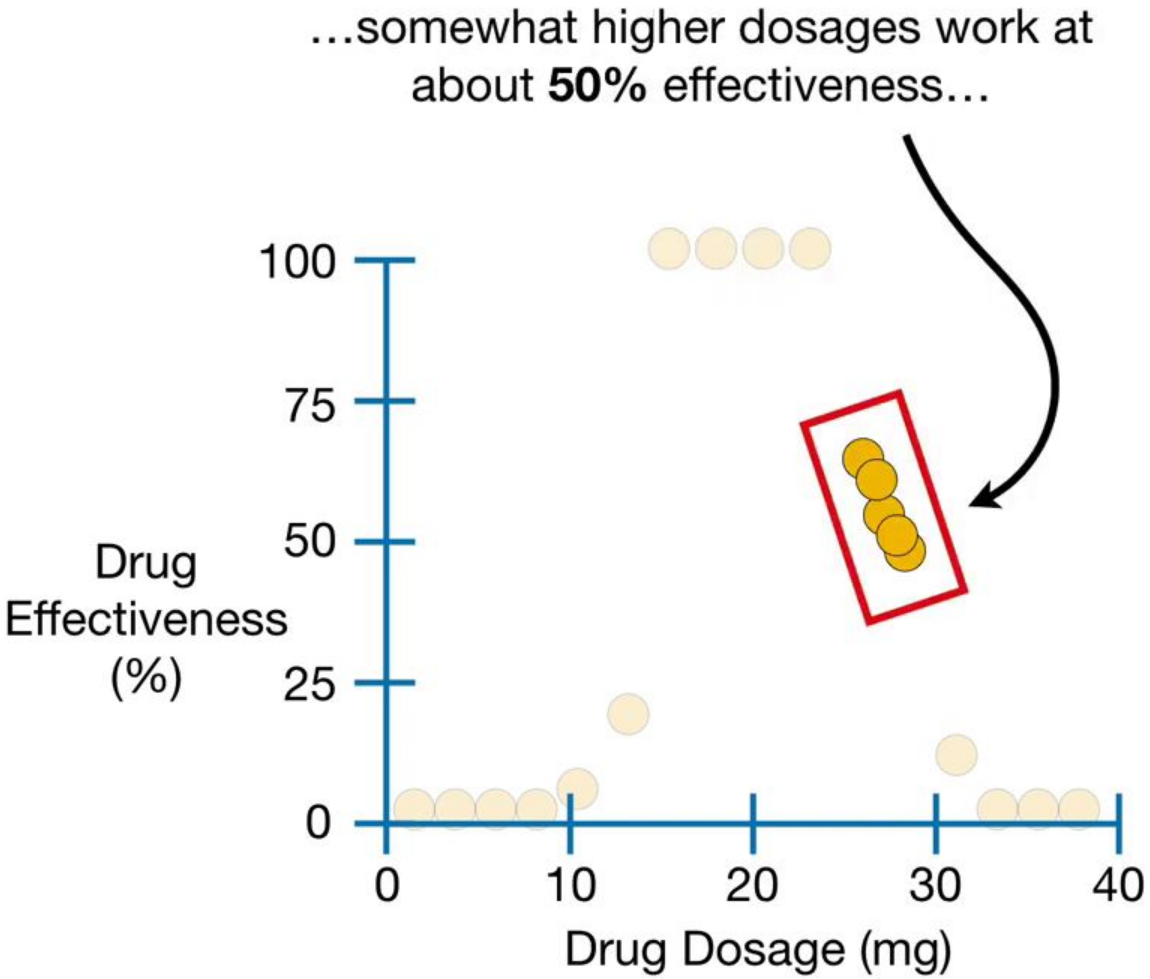

…we could use the line to predict that a
27 mg Dose should be 62% Effective.
100
75
50
25
0
Drug
Effectiveness
(%)
0
10
20
30
40
Drug Dosage (mg)

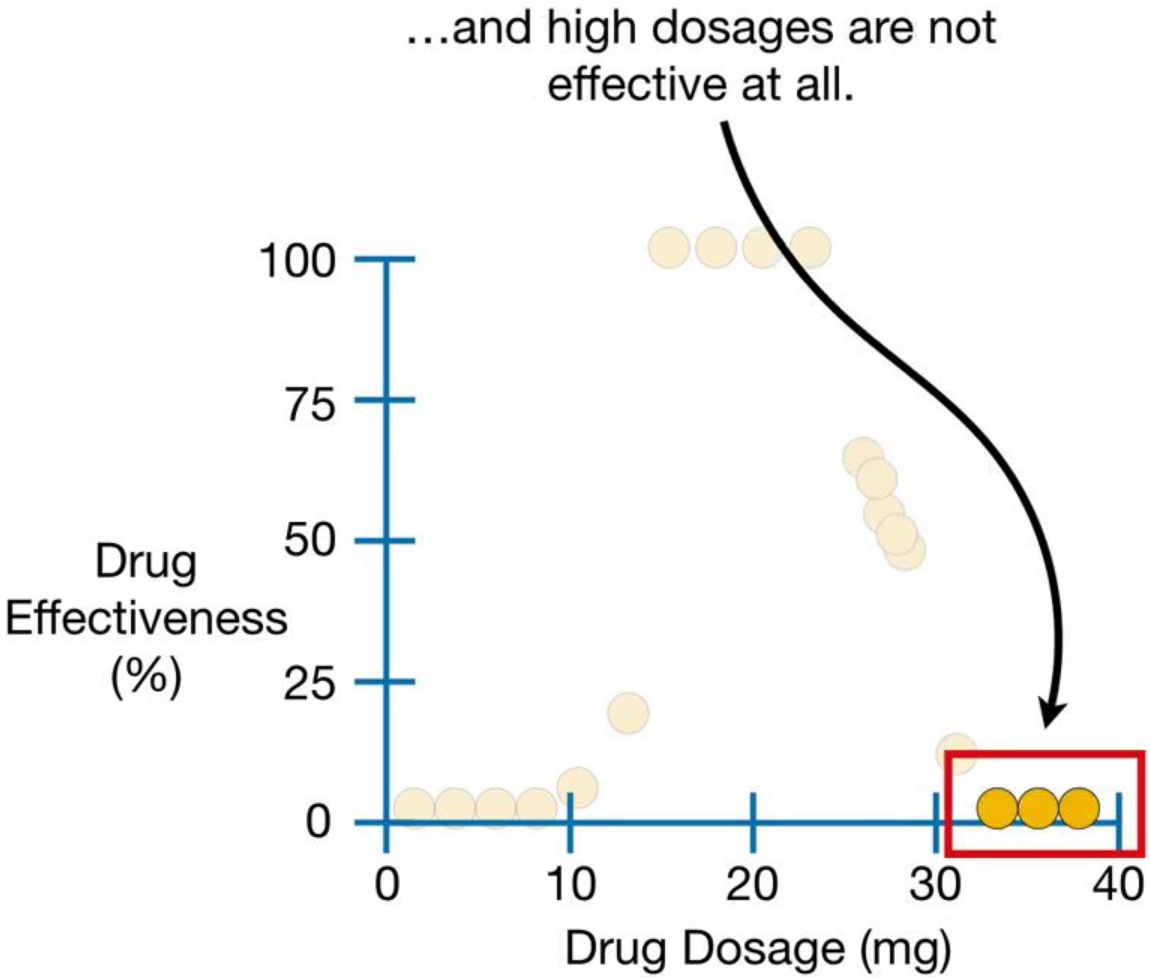
However, what if the data looked like this?

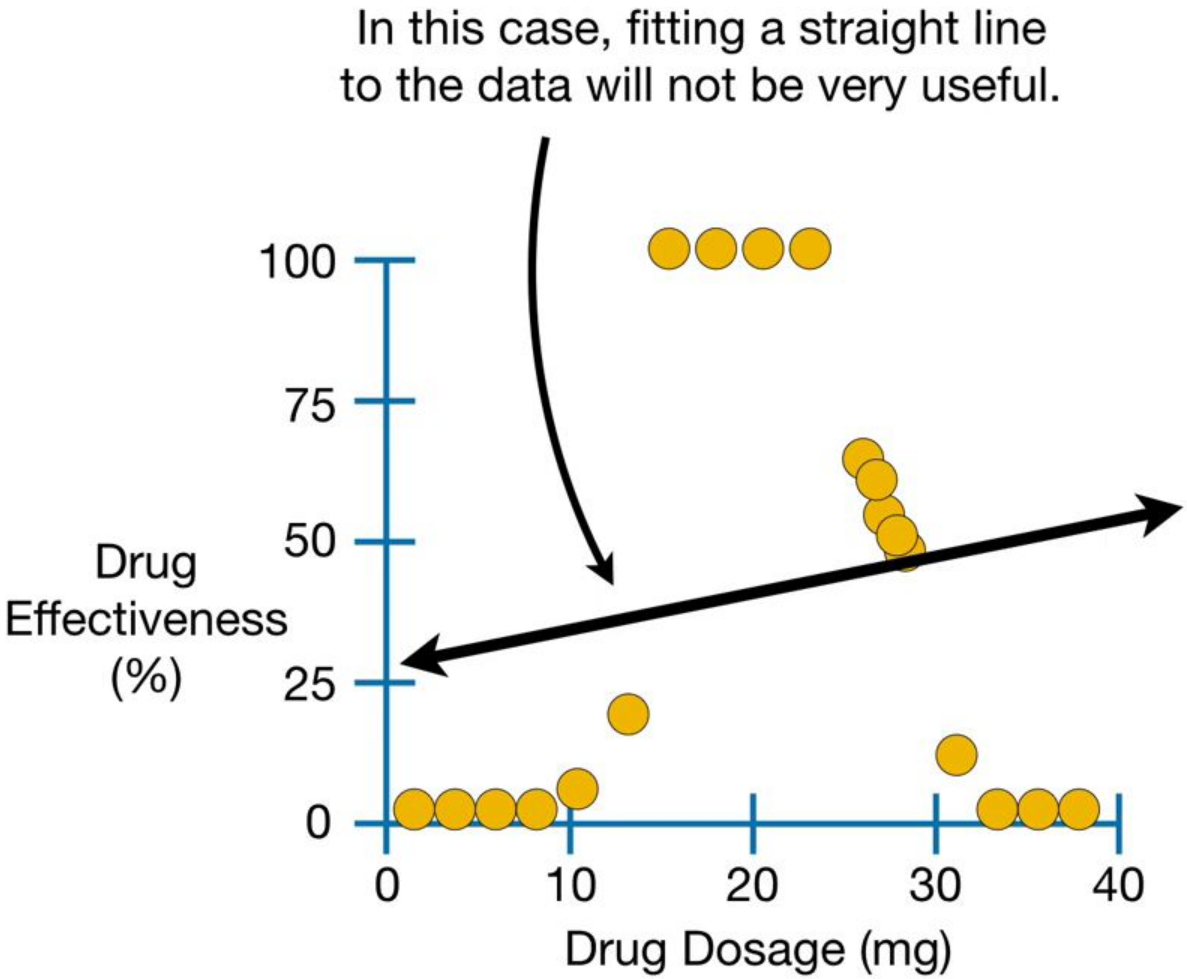


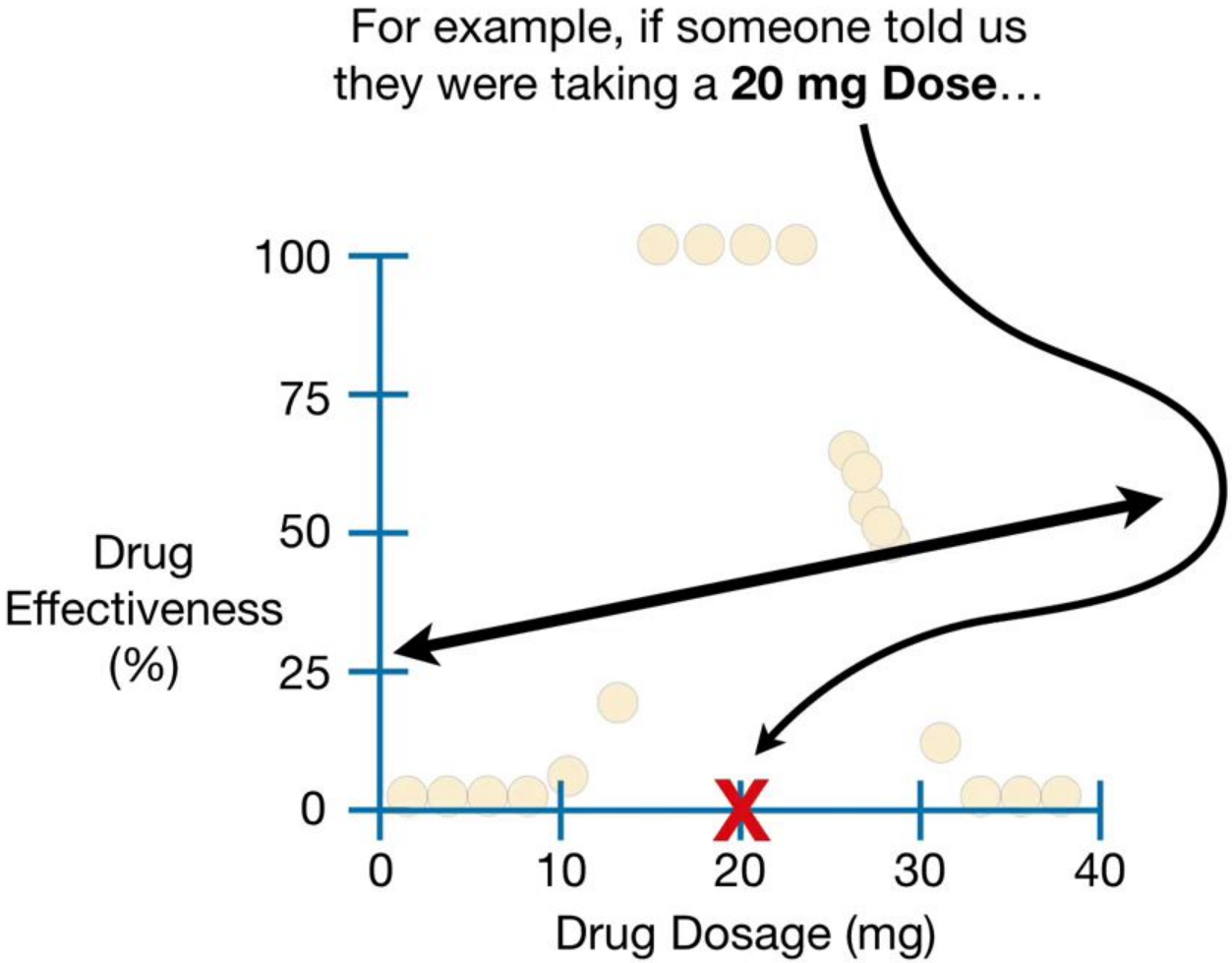

Low dosages are not effective…
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)


…moderate dosages work really well…
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)

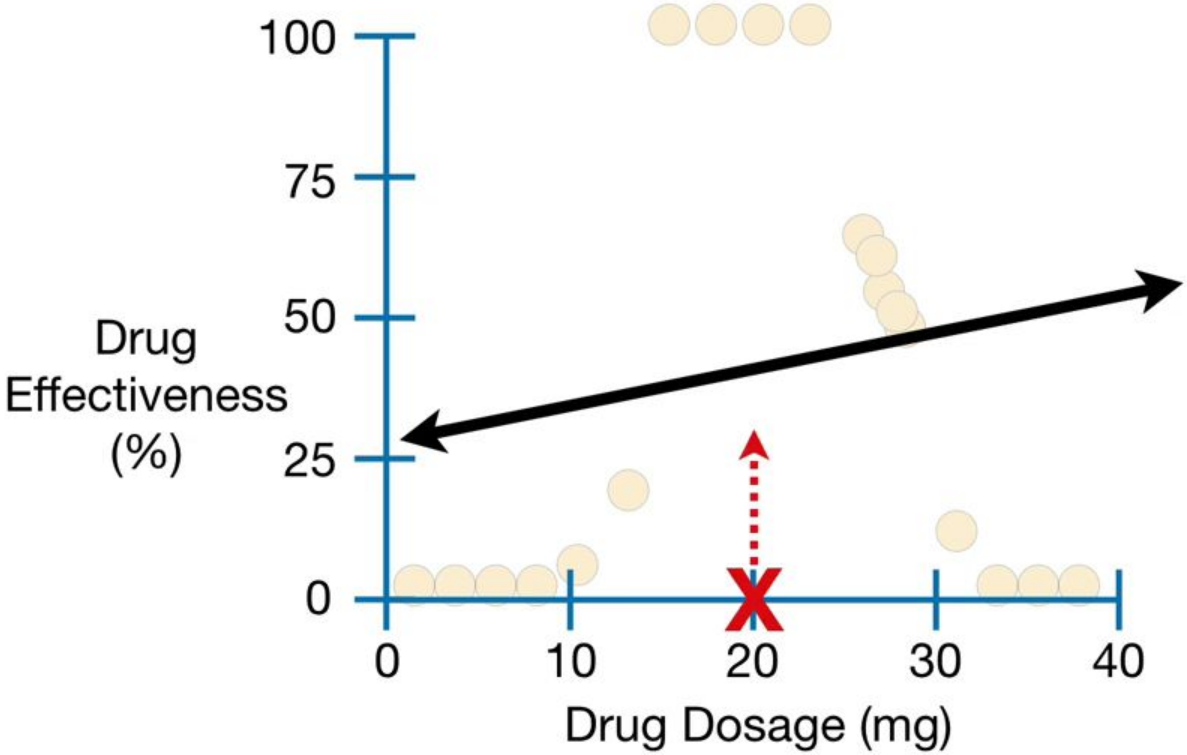

…somewhat higher dosages work at about **50%** effectiveness…
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)


…and high dosages are not effective at all.
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)

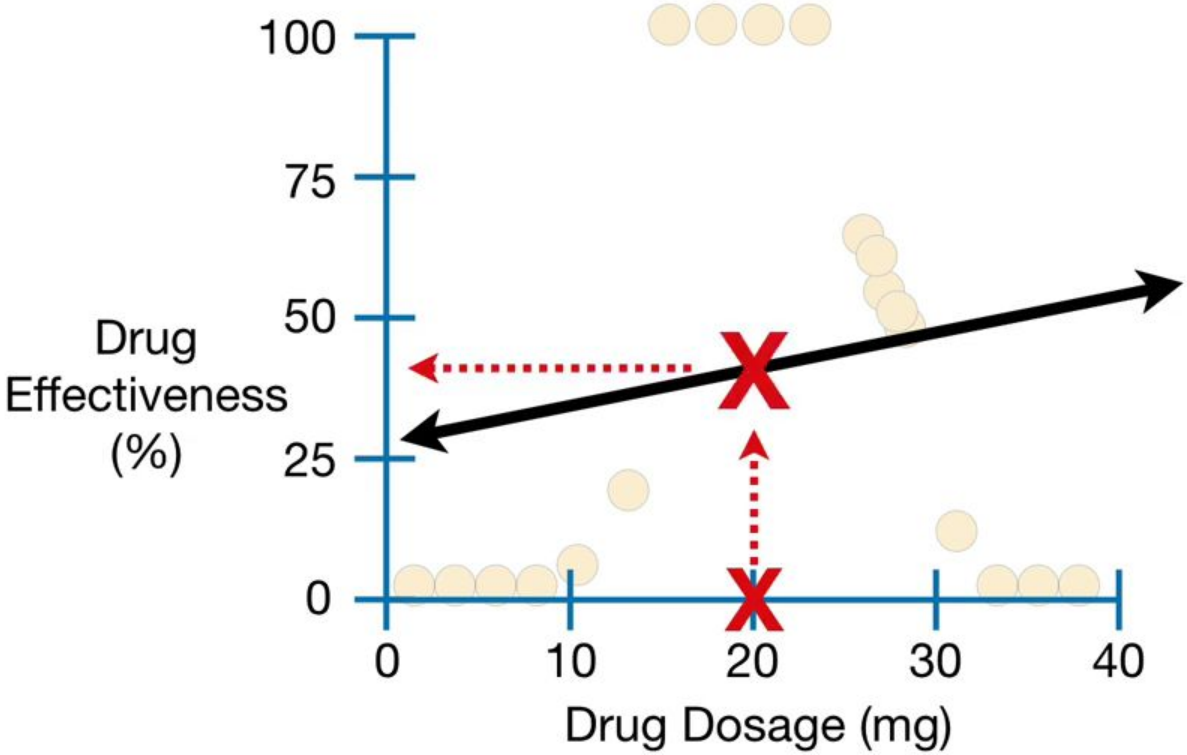

In this case, fitting a straight line
to the data will not be very useful.
100
75
50
25
0
Drug
Effectiveness
(%)
0
10
20
30
40
Drug Dosage (mg)

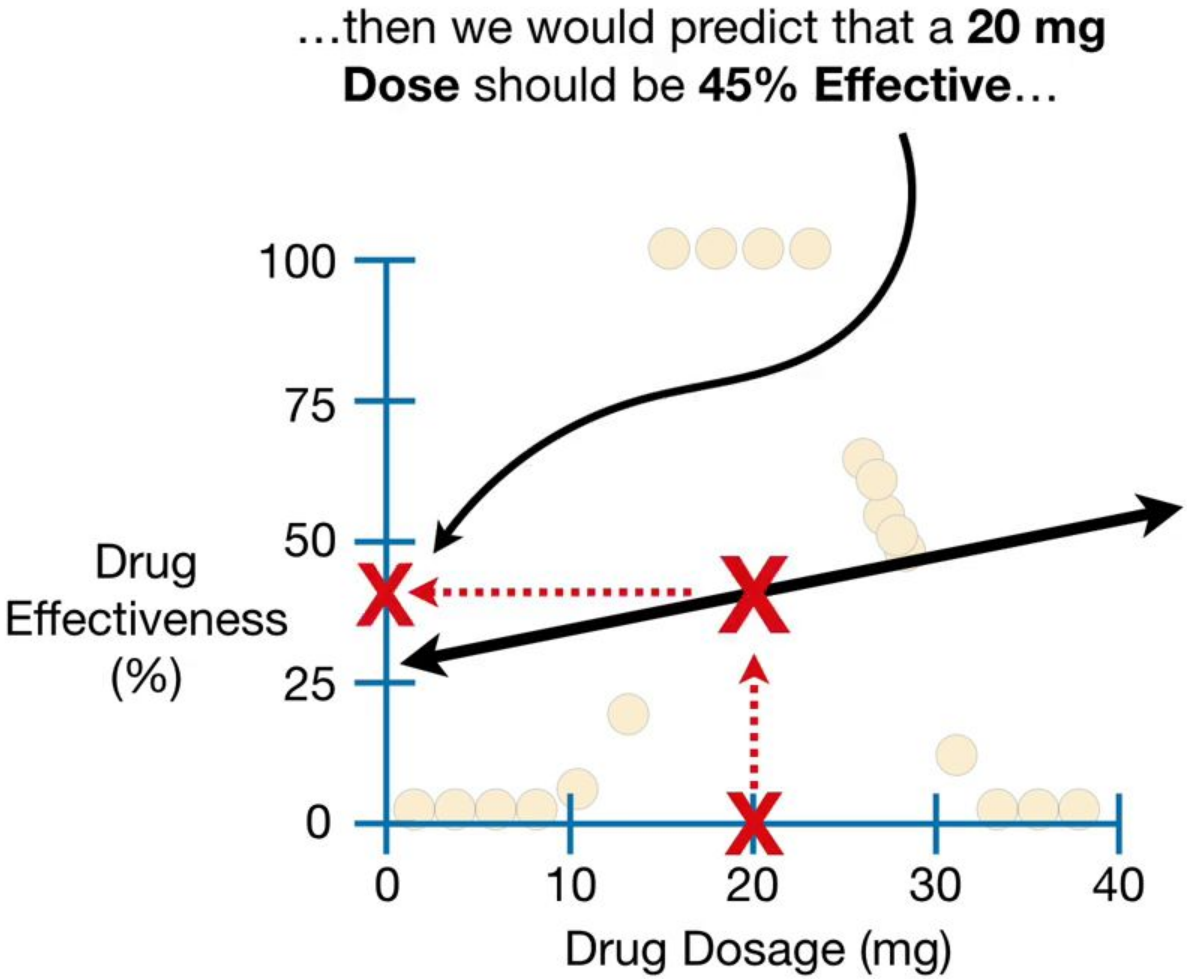

For example, if someone told us they were taking a **20 mg Dose**…
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)

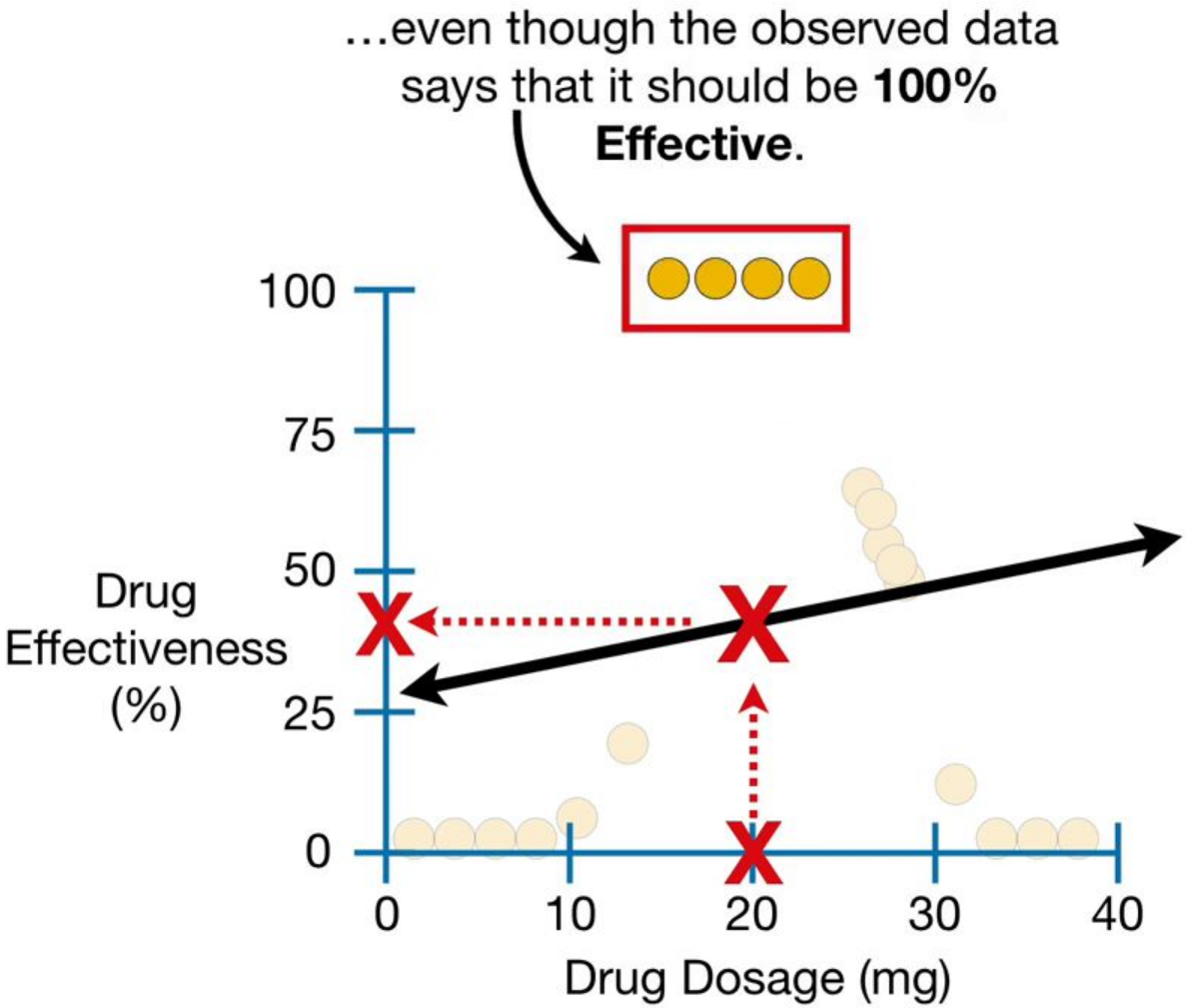
…then we would predict that a **20 mg Dose** should be **45% Effective**…



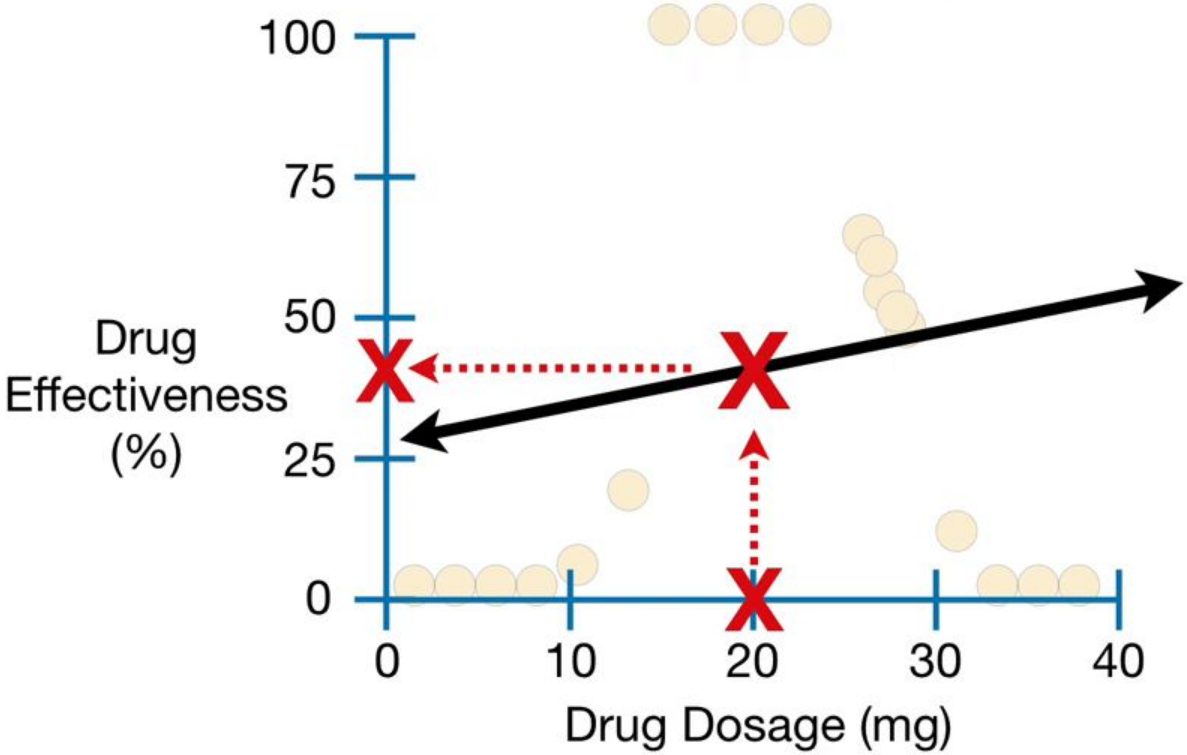
…then we would predict that a **20 mg Dose** should be **45% Effective**…



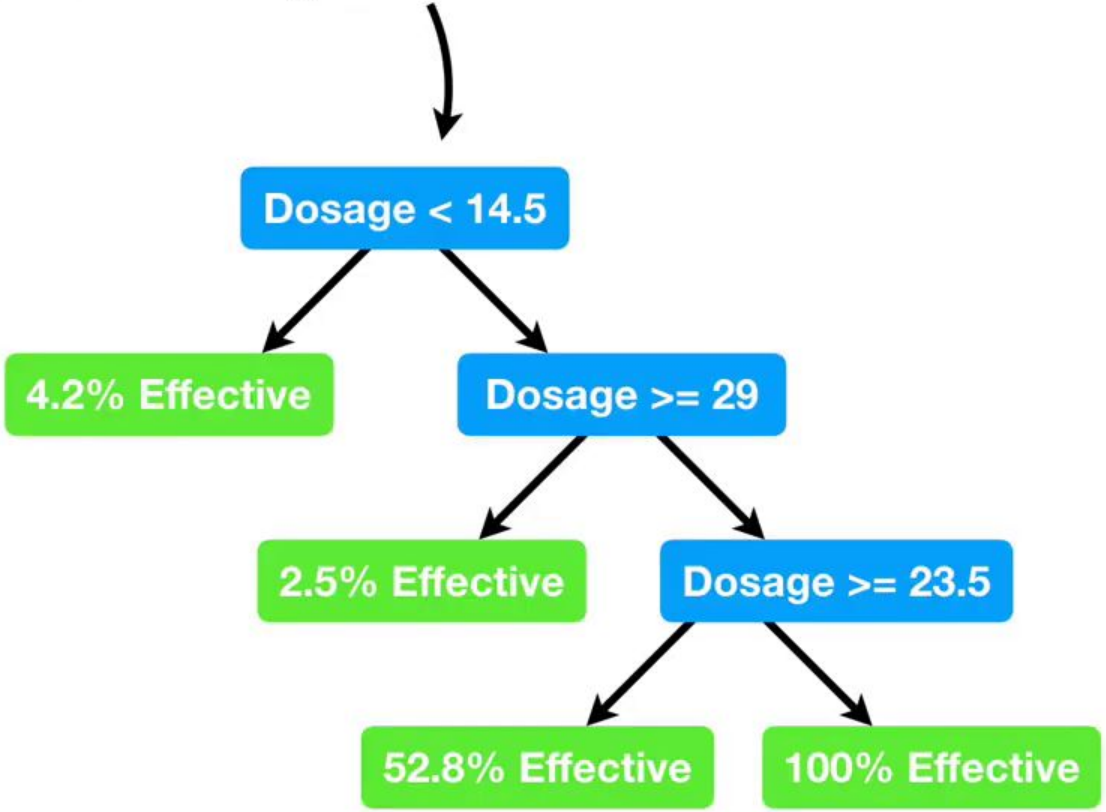

…then we would predict that a **20 mg Dose** should be **45% Effective**…
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)


…even though the observed data says that it should be **100% Effective**.
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)

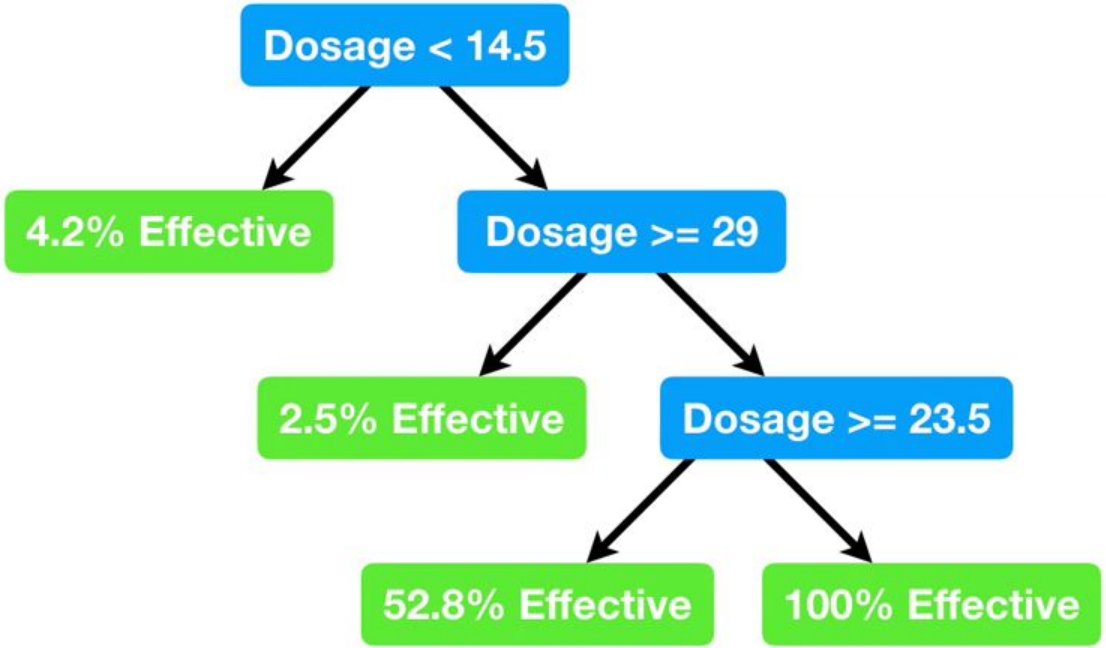
So we need to use something other than a straight line to make predictions.



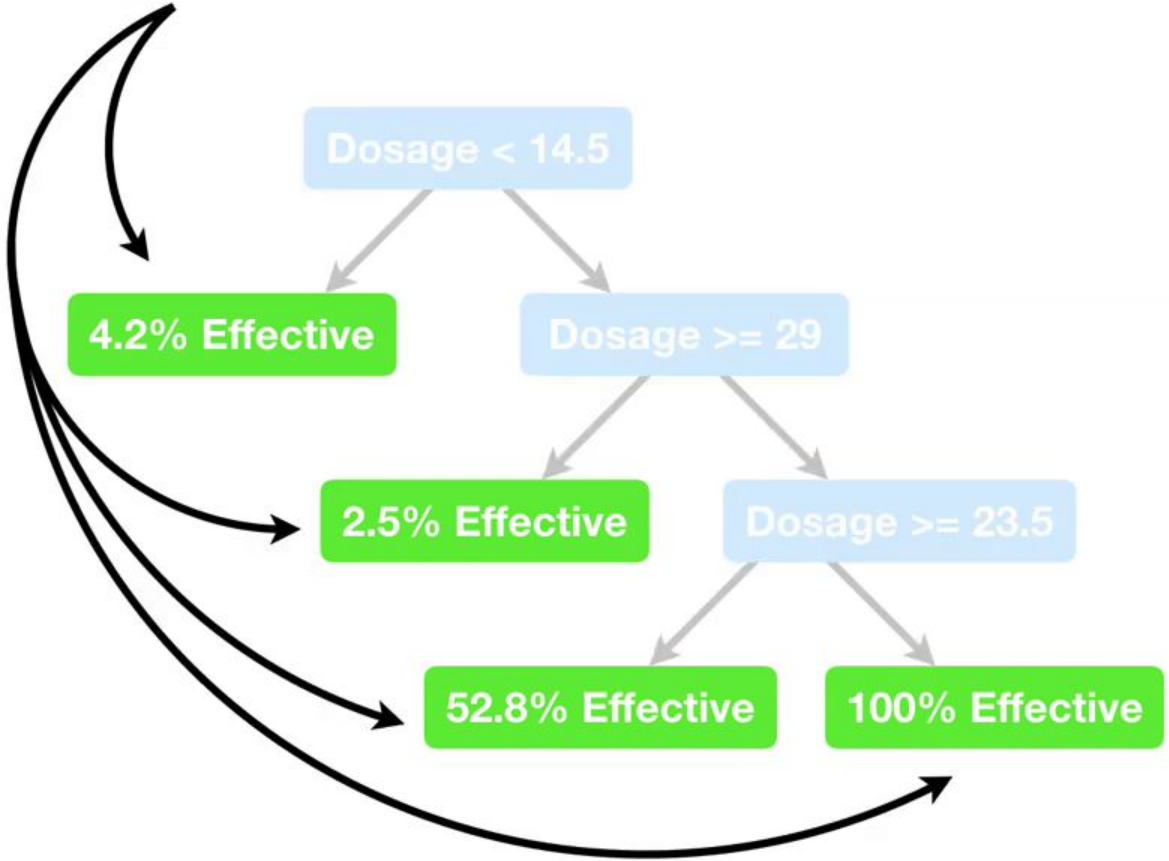
One option is to use a **Regression Tree**.

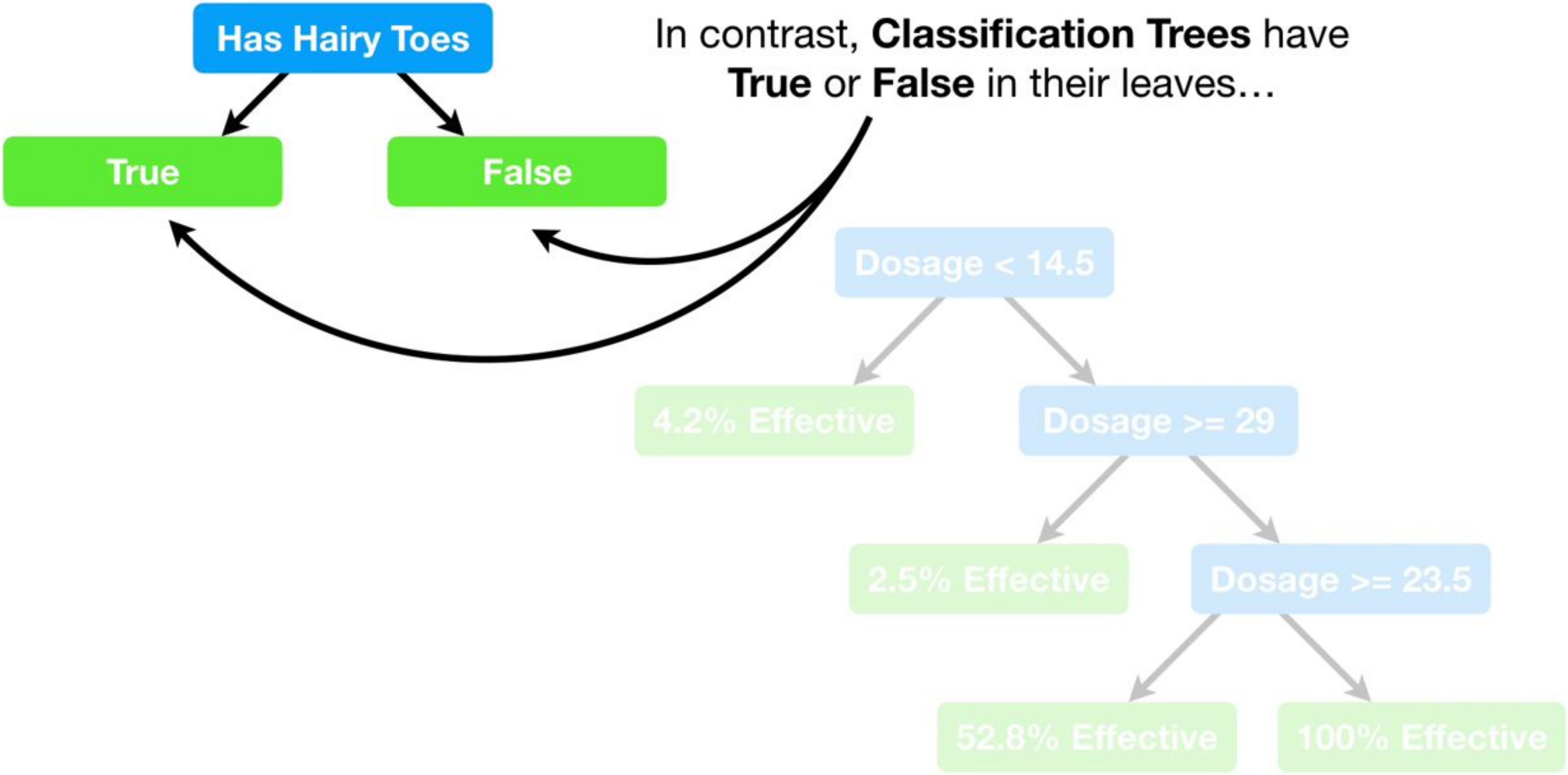


Regression Trees are a type of **Decision Tree.**

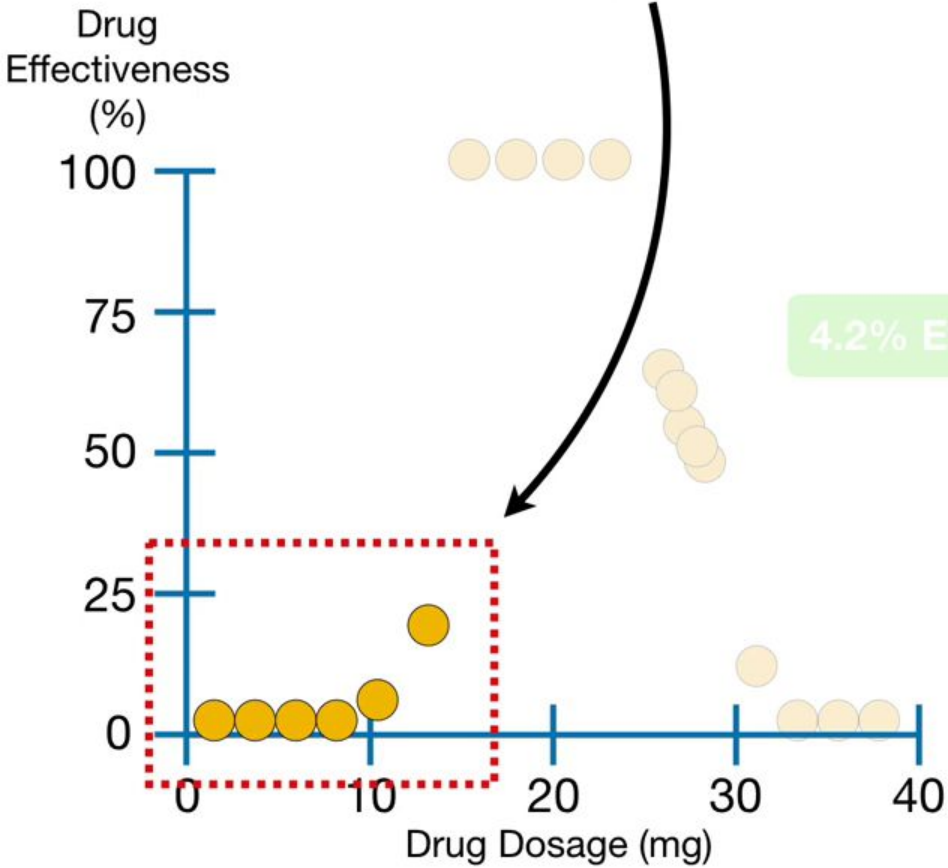


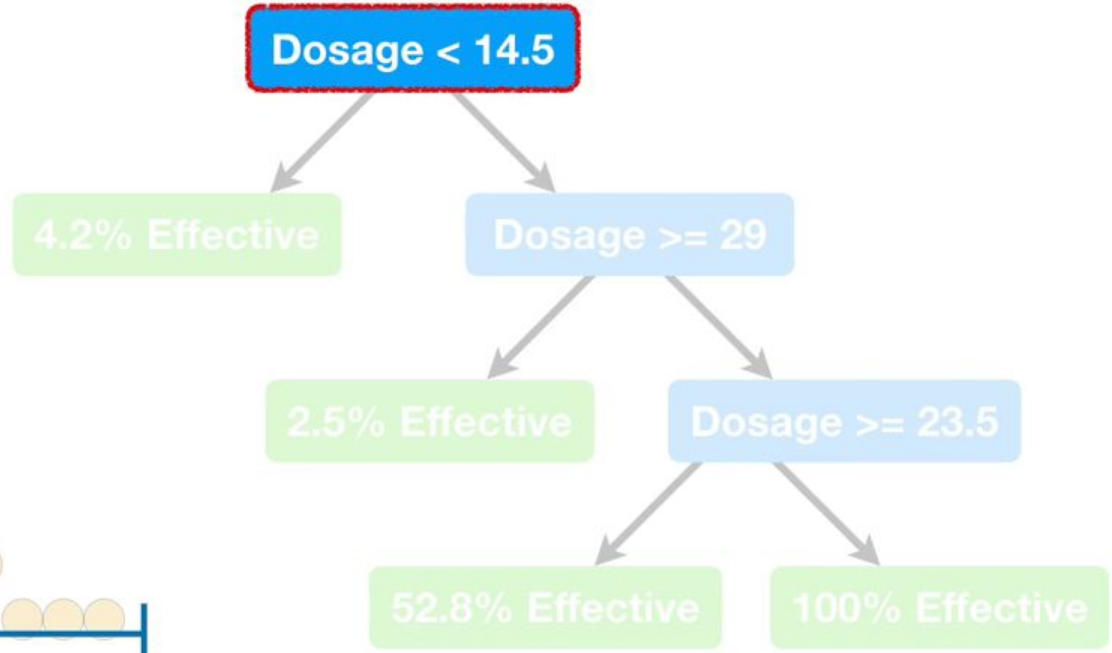
In a **Regression Tree**, each leaf represents a numeric value.



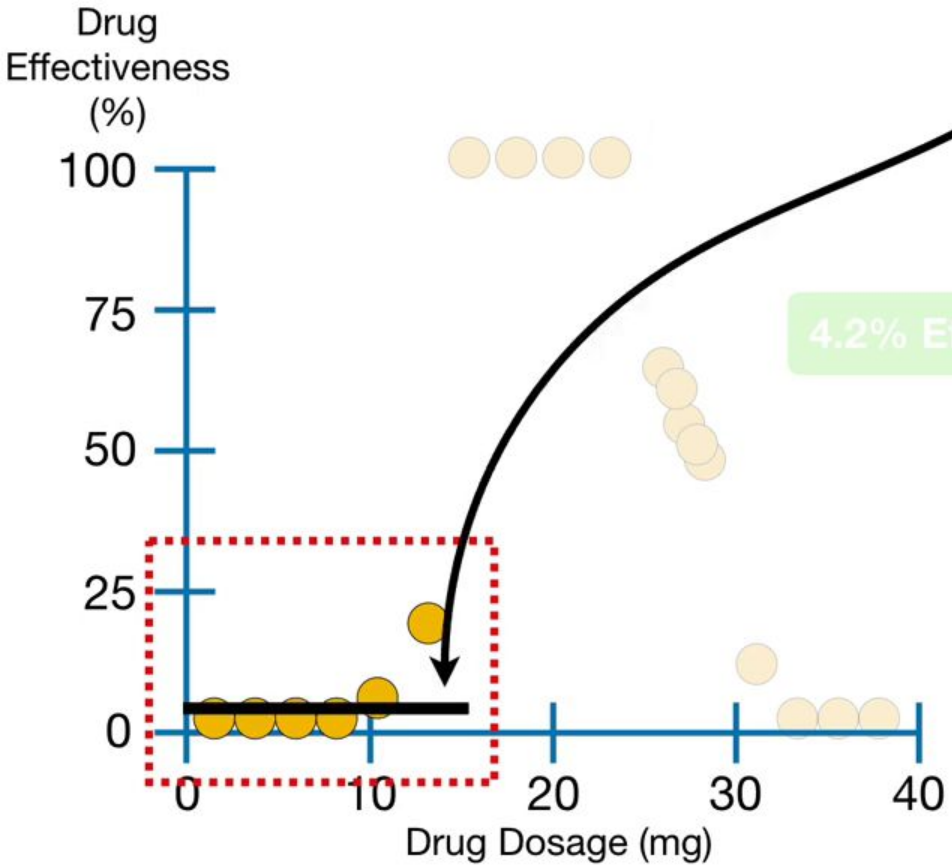

Has Hairy Toes
True
False
In contrast, **Classification Trees** have **True** or **False** in their leaves…
Dosage < 14.5
4.2% Effective
Dosage >= 29
2.5% Effective
Dosage >= 23.5
52.8% Effective
100% Effective

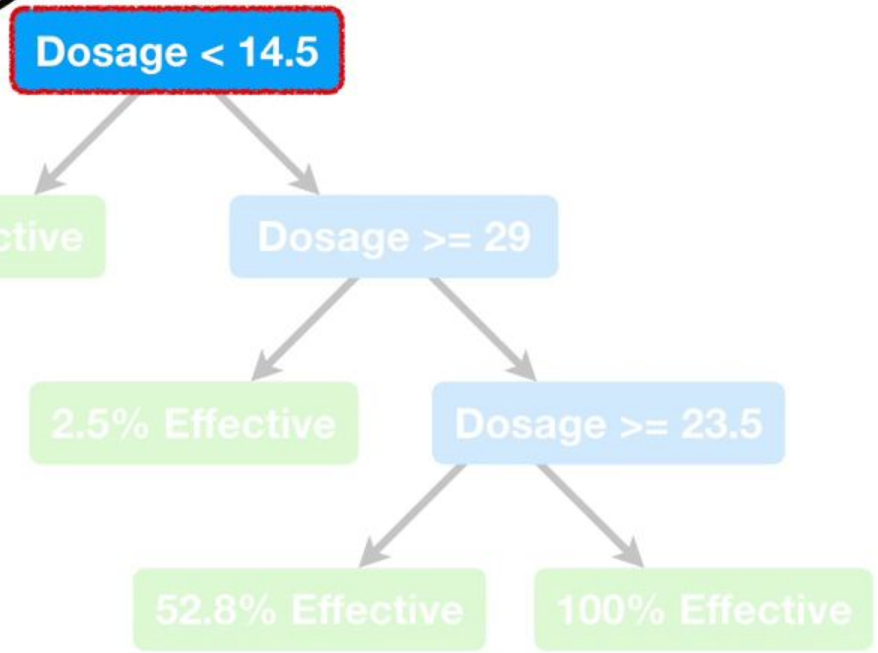
…if so, then we are talking about these
6 observations in the training data…



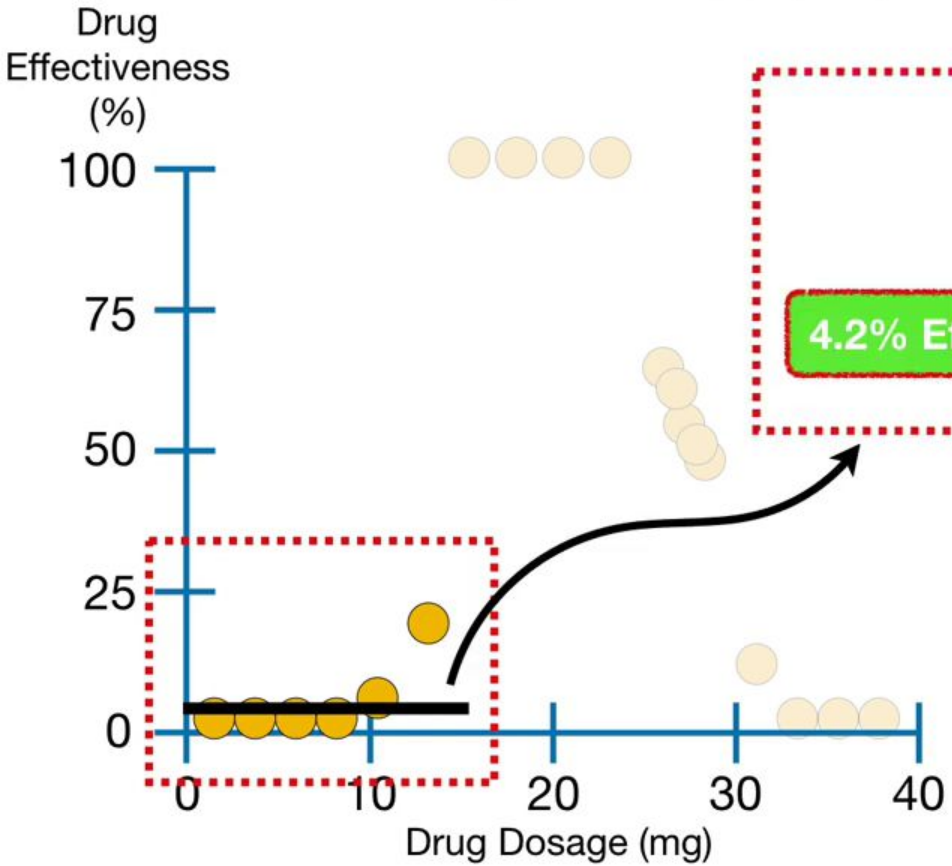


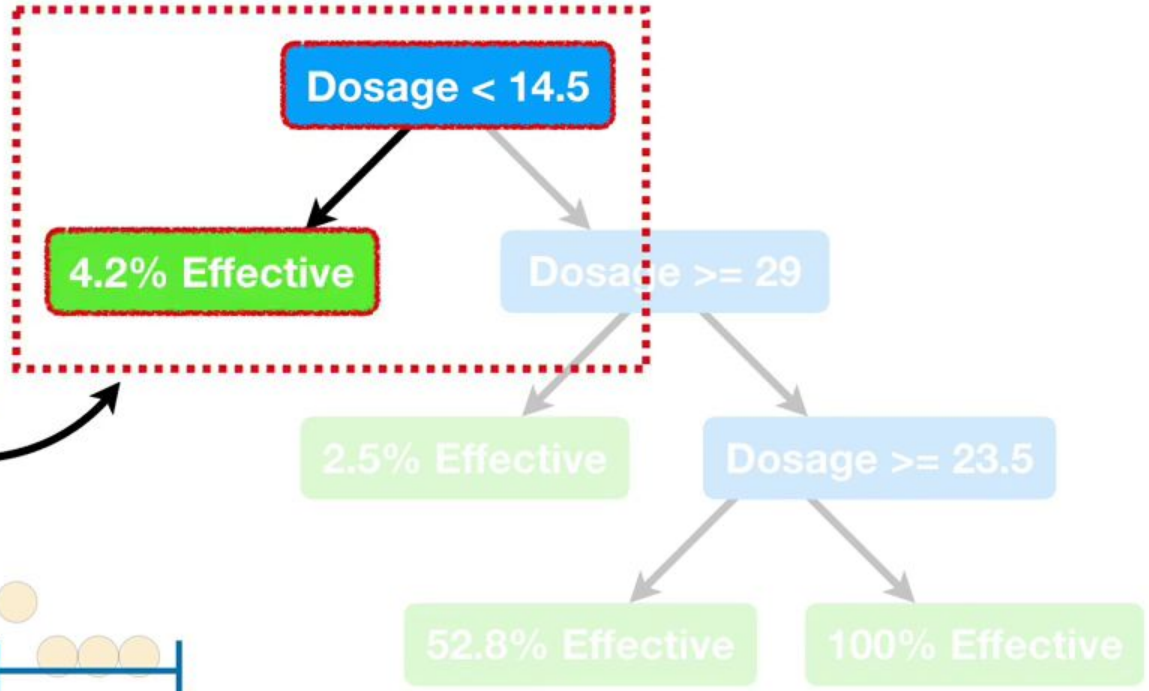
…and the average **Drug Effectiveness** for these **6** observations is **4.2%**…



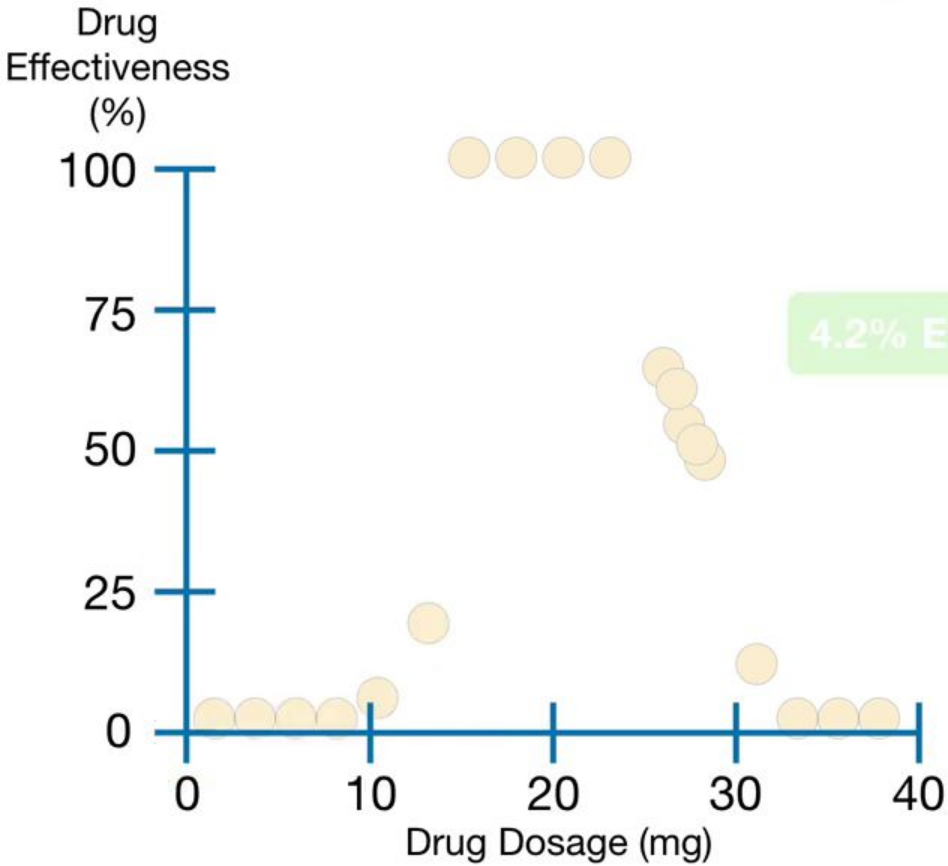


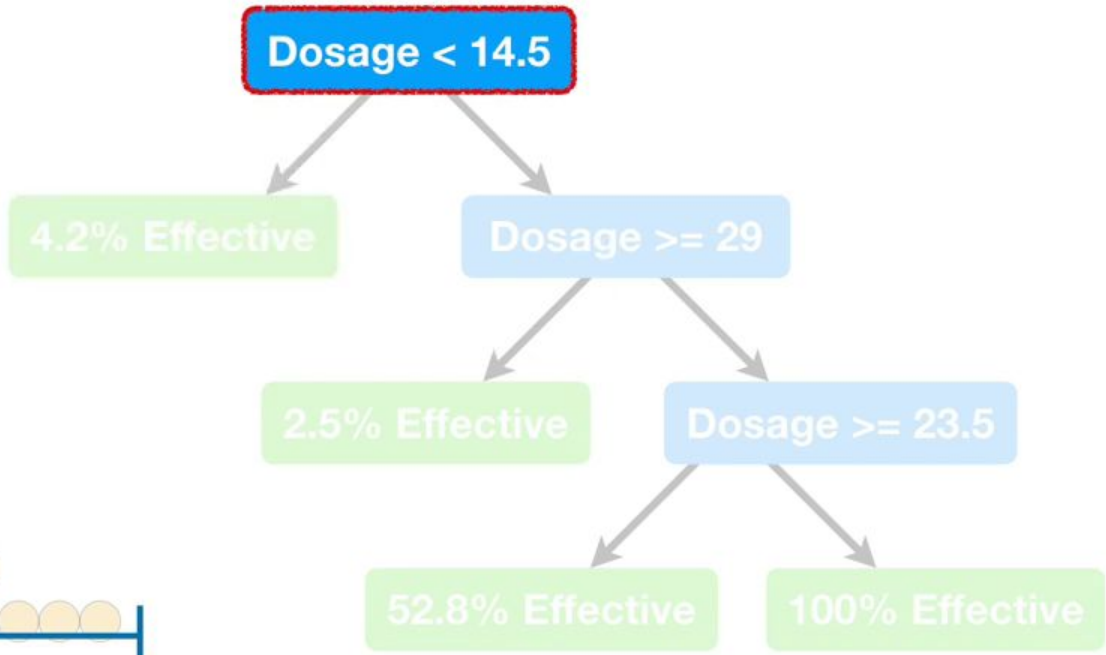
…so the tree uses the average value, **4.2%**, as its prediction for people with **Dosages < 14.5**.

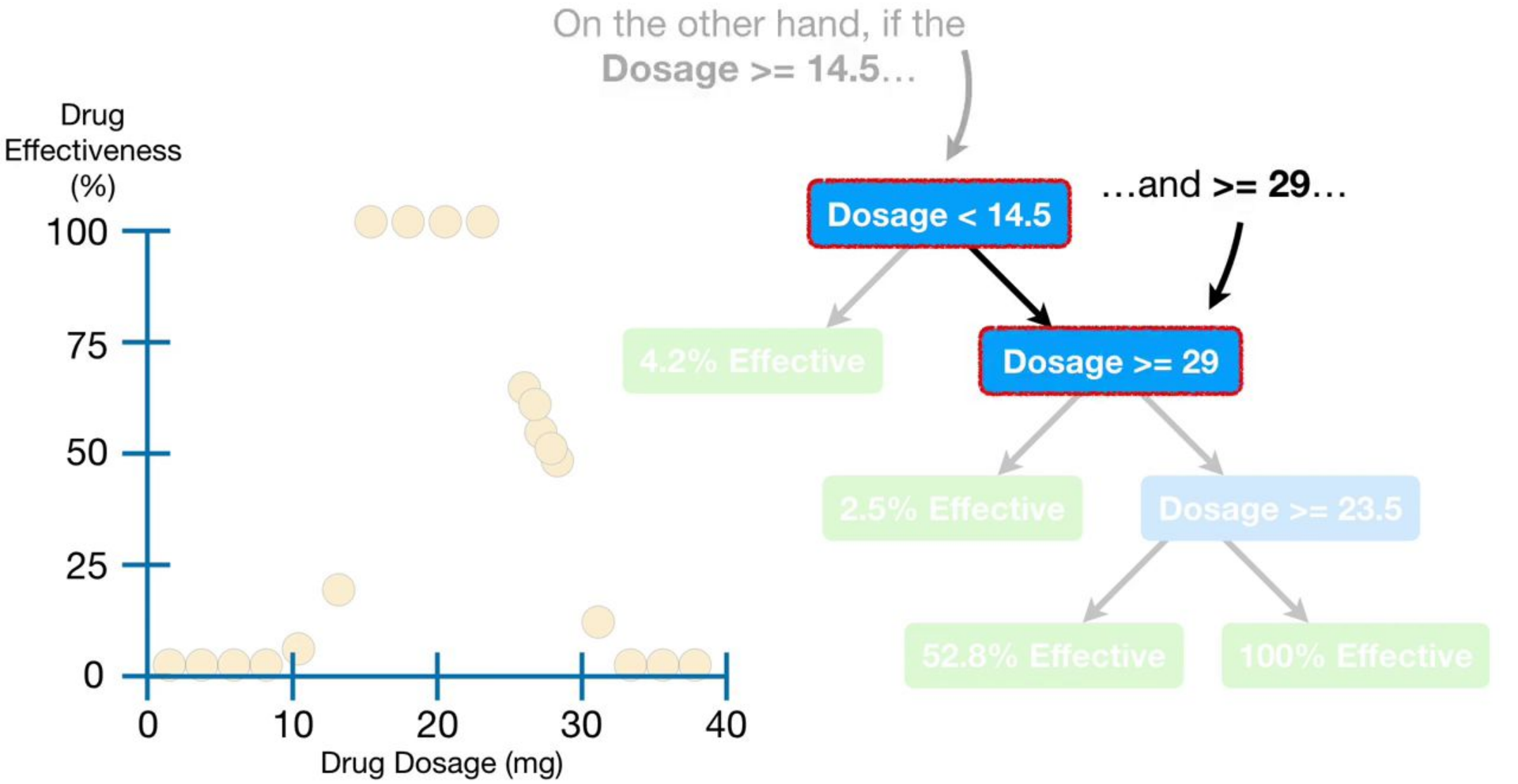




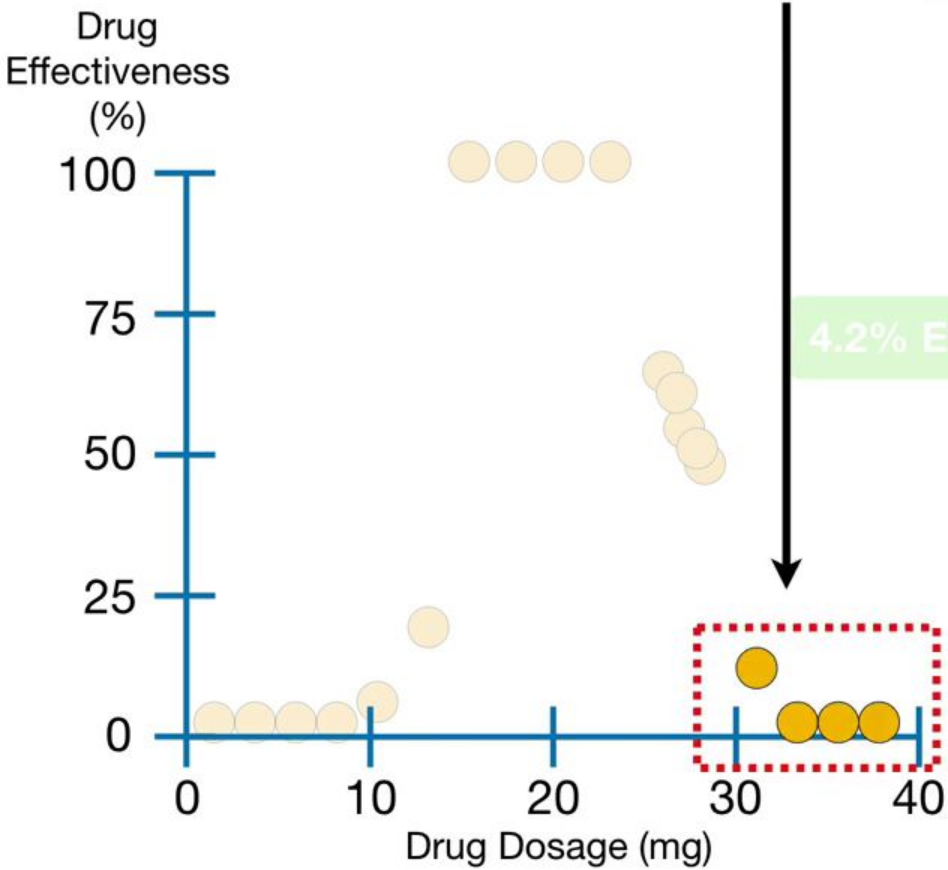
On the other hand, if the
Dosage >= 14.5…

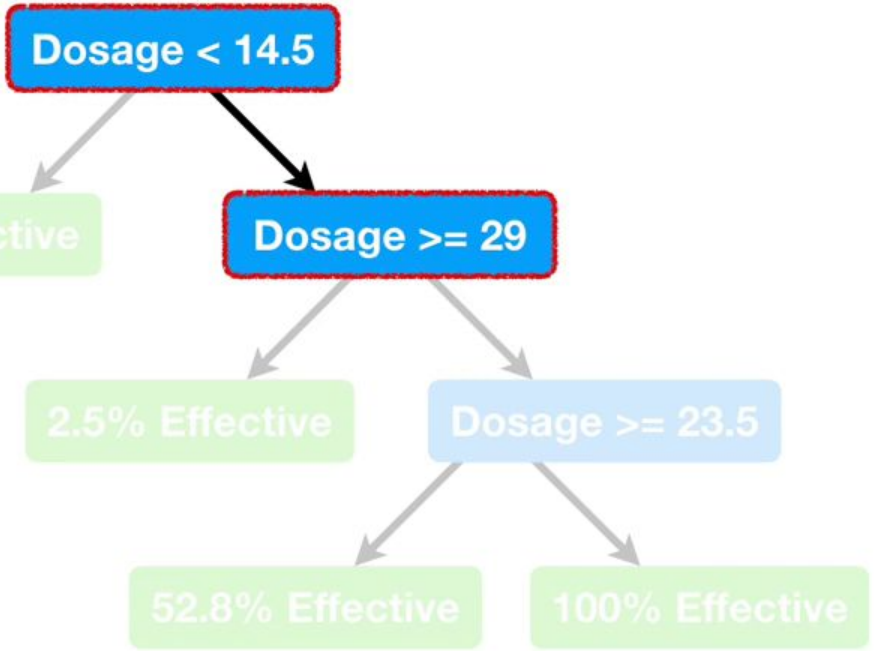




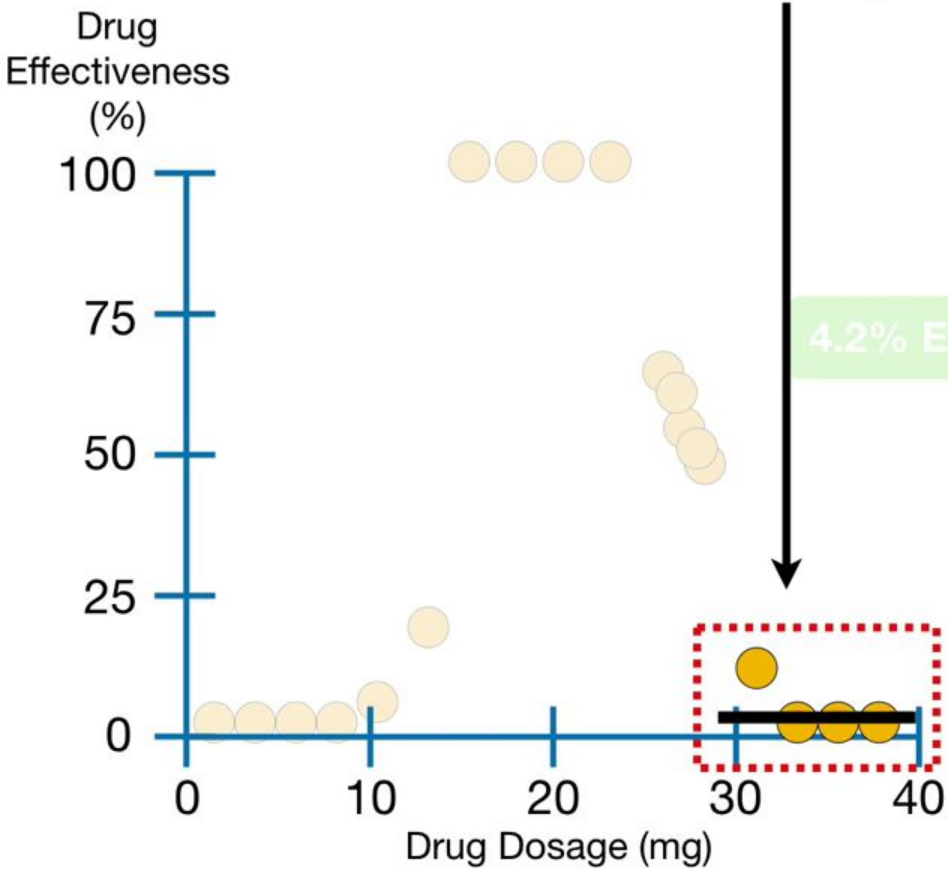

On the other hand, if the
Dosage >= 14.5…
Drug
Effectiveness
(%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
…and >= 29…
4.2% Effective
Dosage >= 29
2.5% Effective
Dosage >= 23.5
52.8% Effective
100% Effective

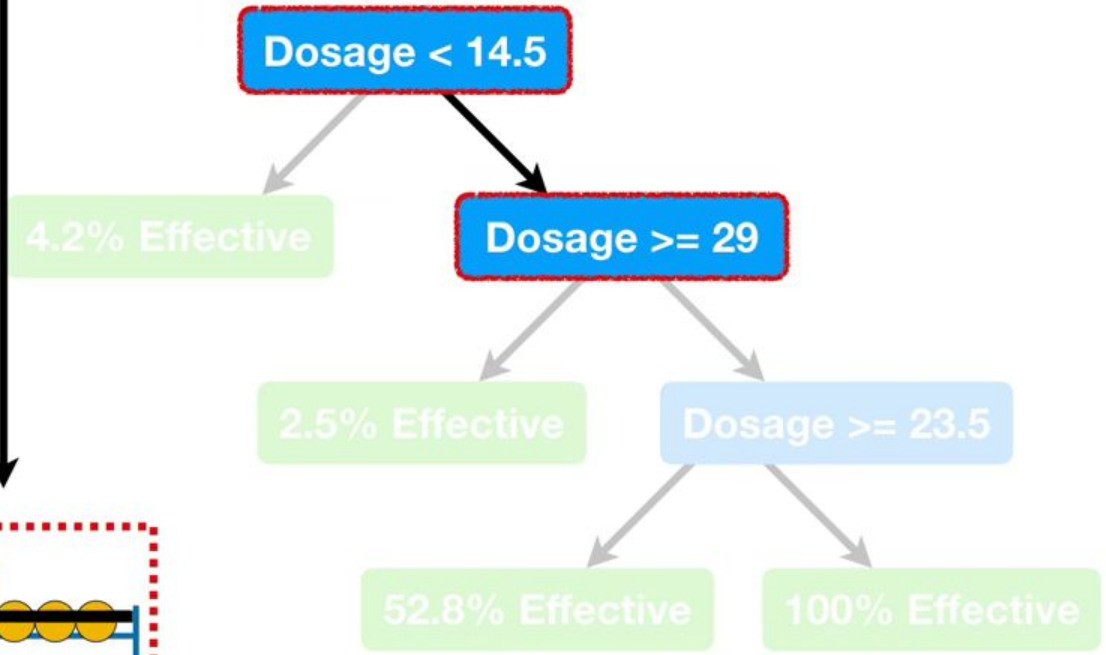
…then we are talking about these **4** observations in the training dataset…



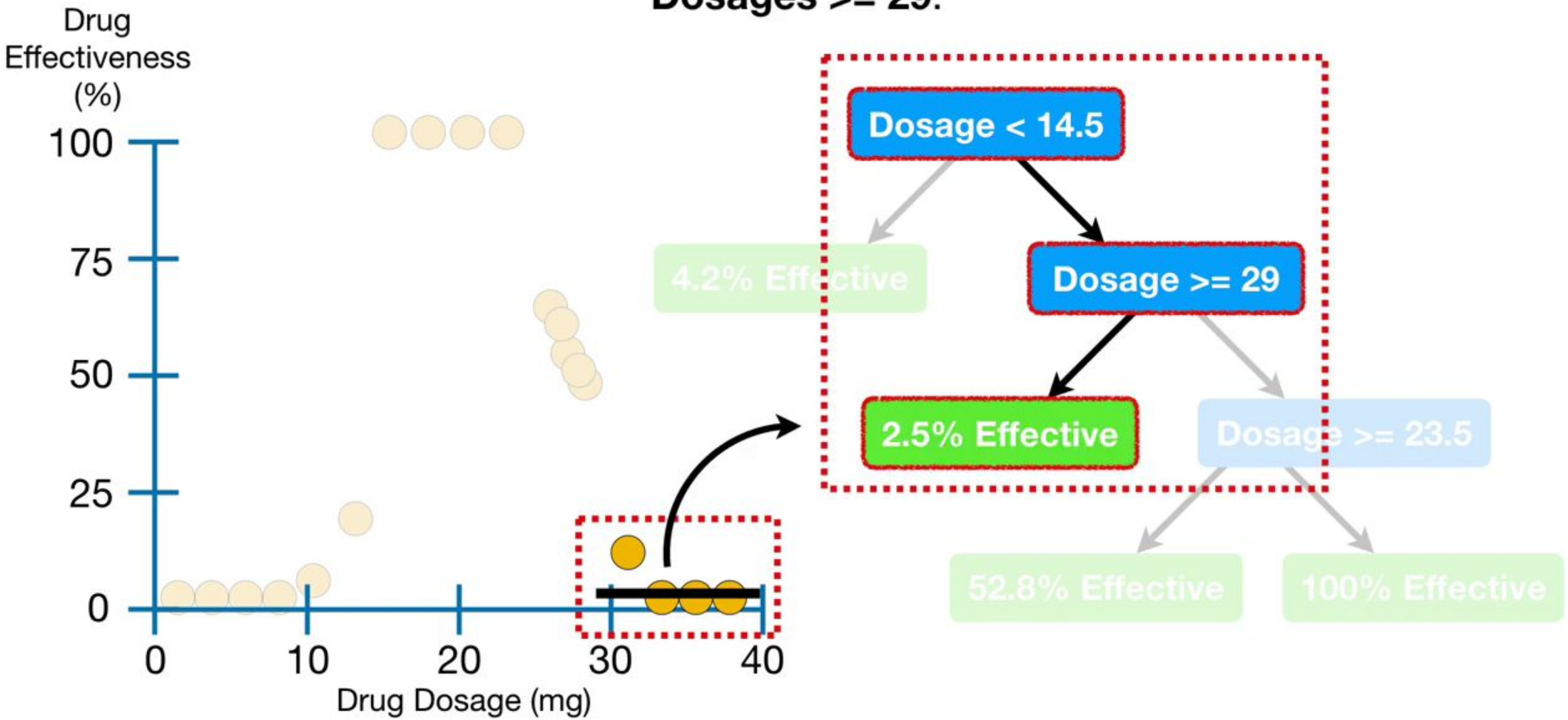


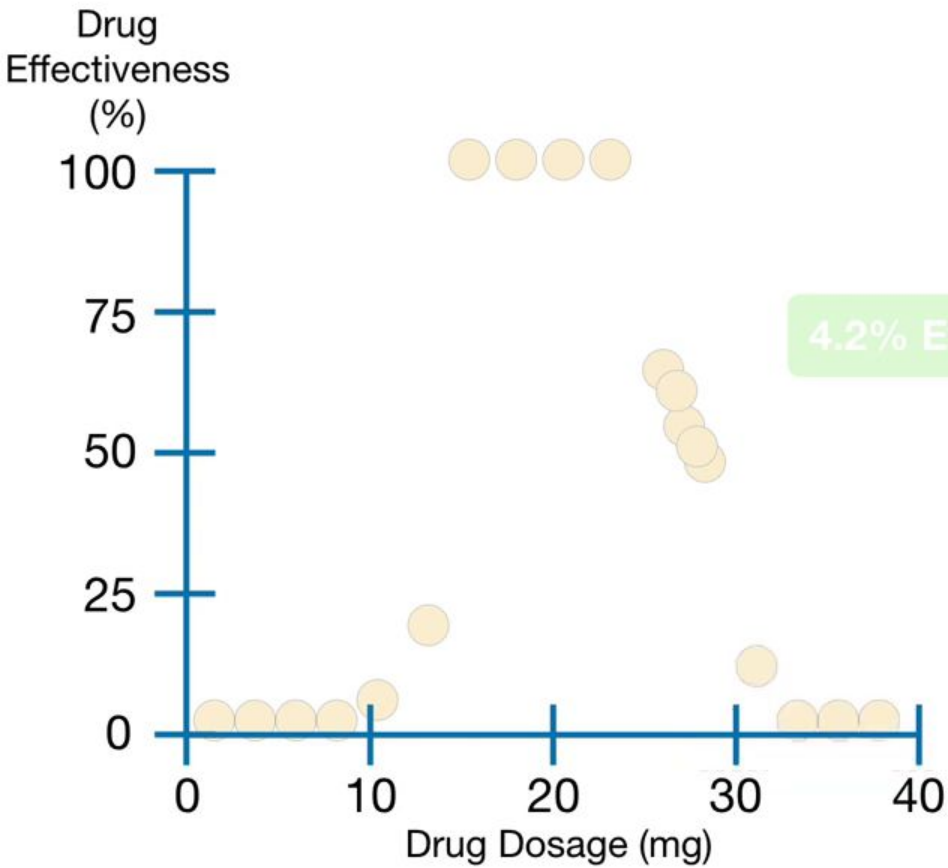
…and the average **Drug Effectiveness** for these **4** observations is **2.5%**…



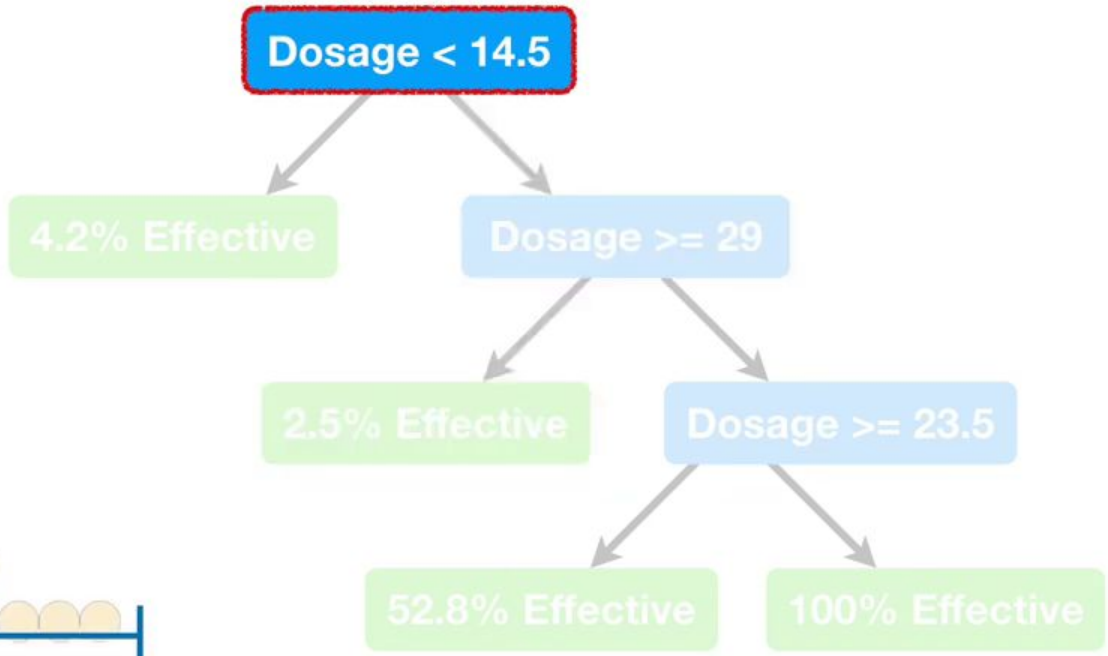


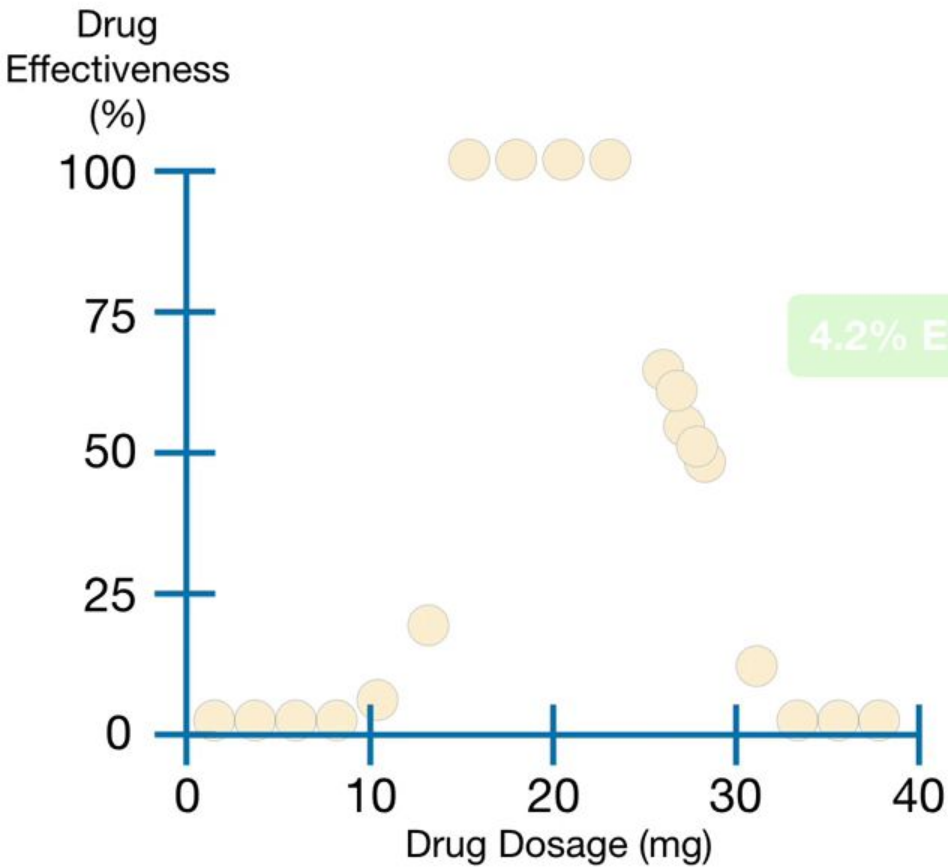
…so the tree uses the average value, **2.5%**, as its prediction for people with **Dosages >= 29**.





Now, if the **Dosage >= 14.5**…





Now, if the **Dosage >= 14.5**…

Dosage < 14.5

4.2% Effective

…and **< 29**…

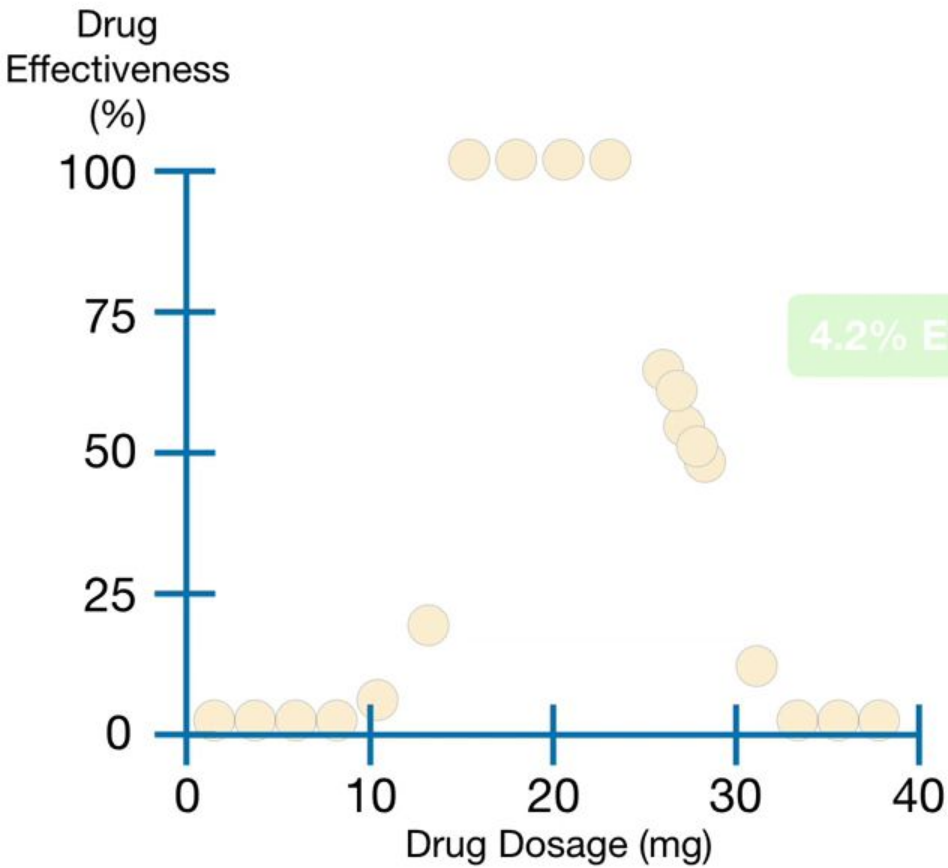
Dosage >= 29

2.5% Effective

Dosage >= 23.5

52.8% Effective

100% Effective



Now, if the **Dosage >= 14.5**…

Dosage < 14.5

4.2% Effective

…and **< 29**…

Dosage >= 29

2.5% Effective

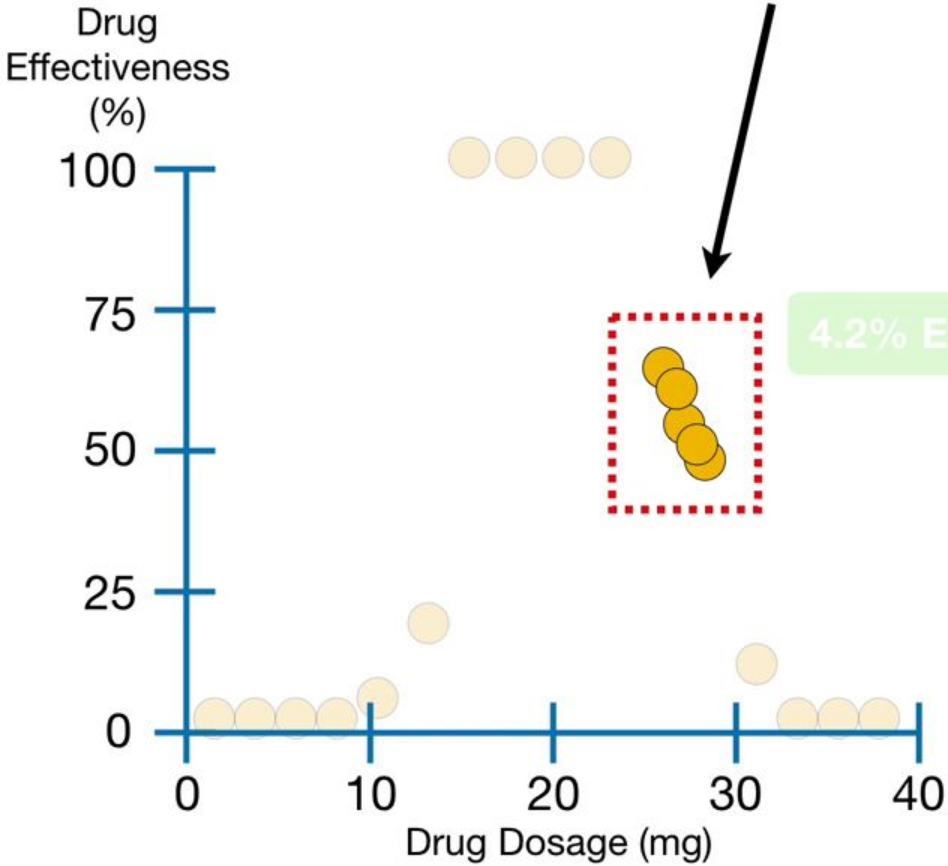
…and **>= 23.5**…

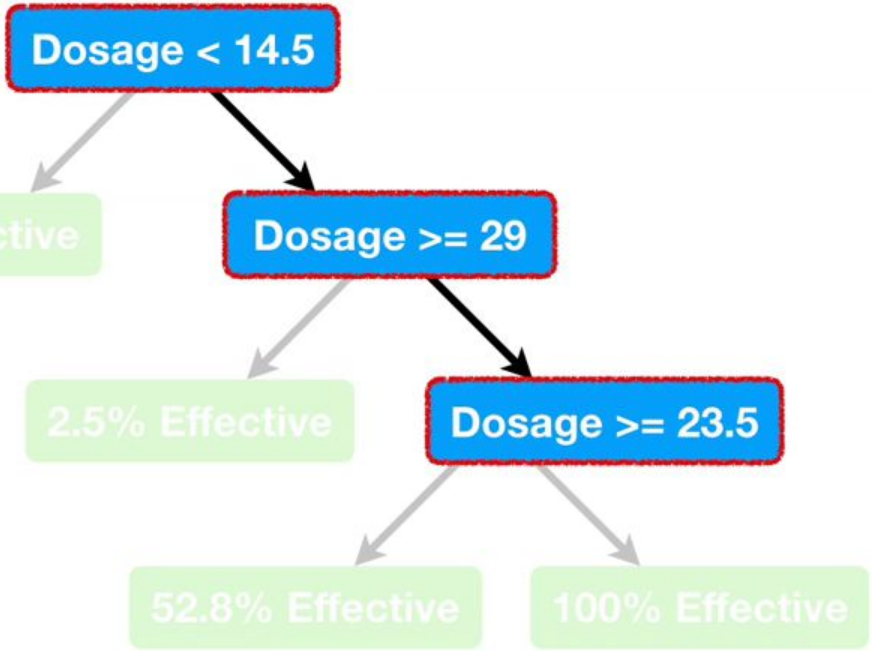
Dosage >= 23.5

52.8% Effective

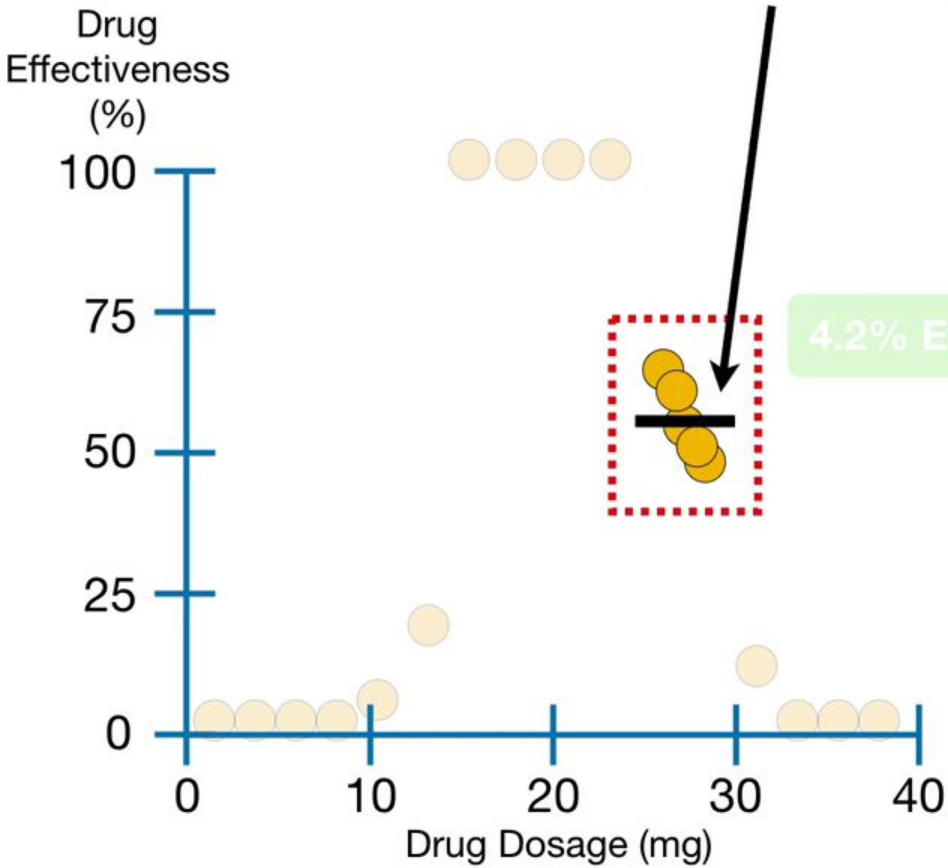
100% Effective

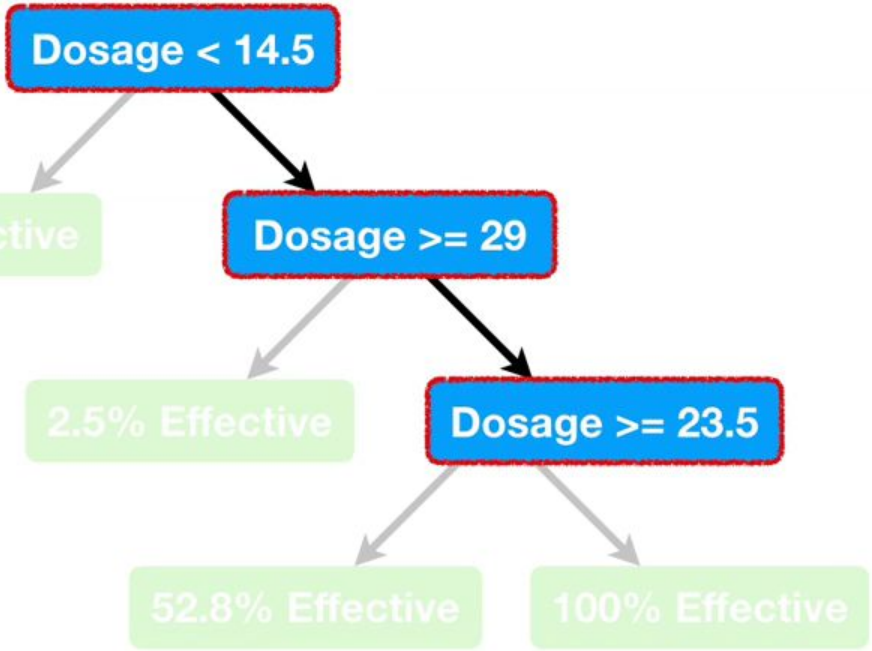
…then we are talking about these **5** observations in the training dataset…



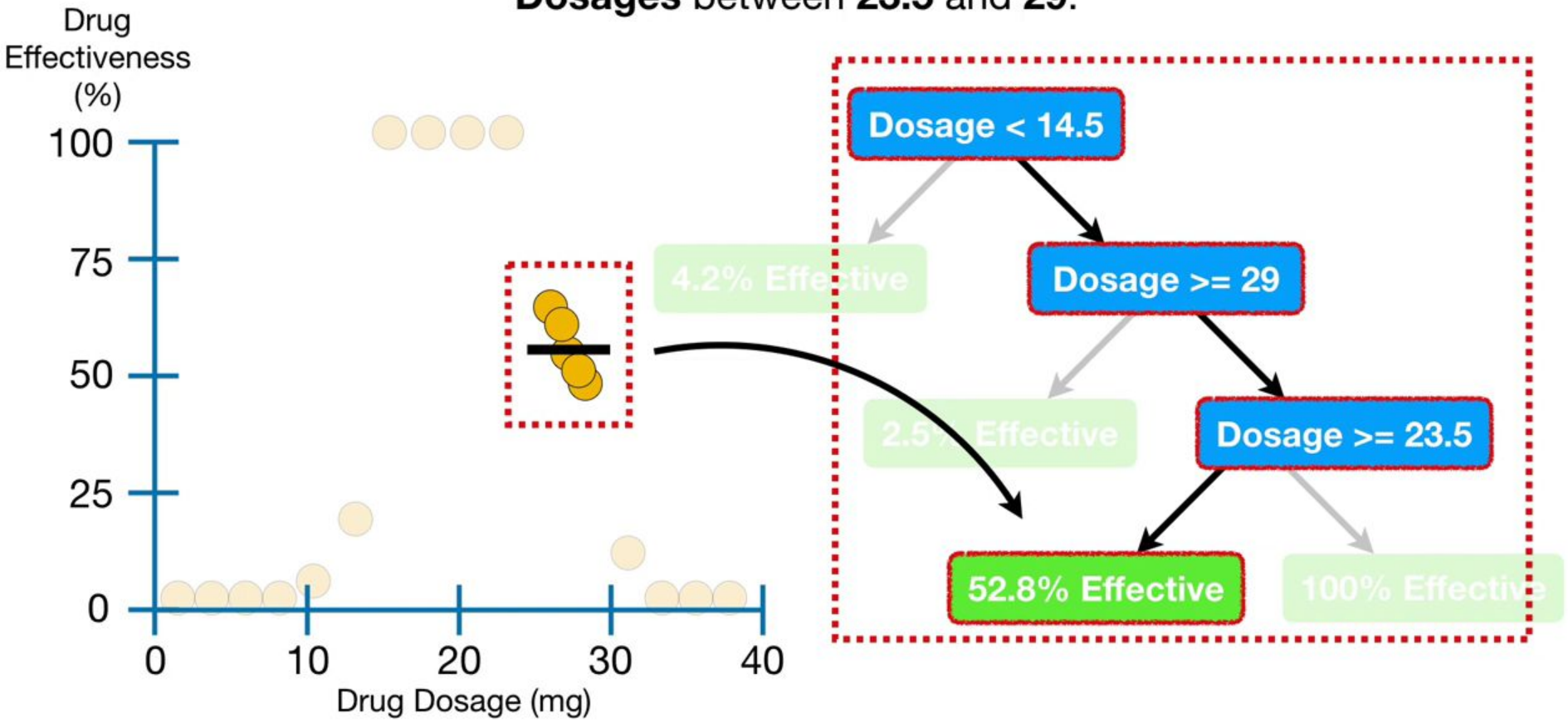


…and the average **Drug Effectiveness** for these **5** observations is **52.8%**…

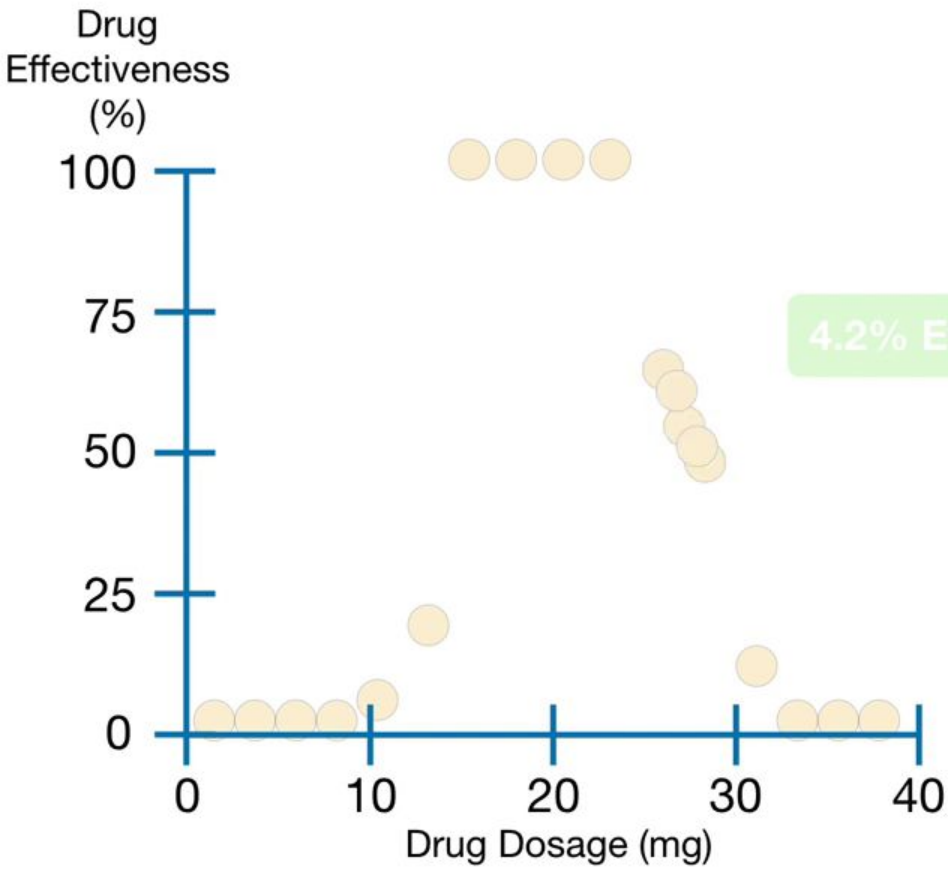


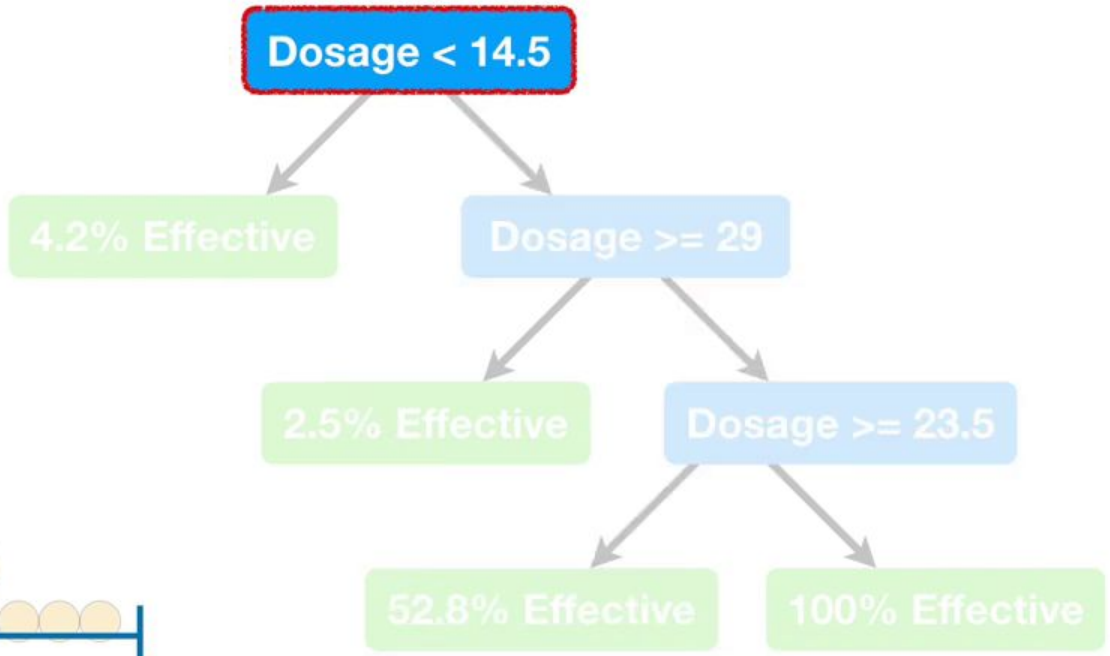


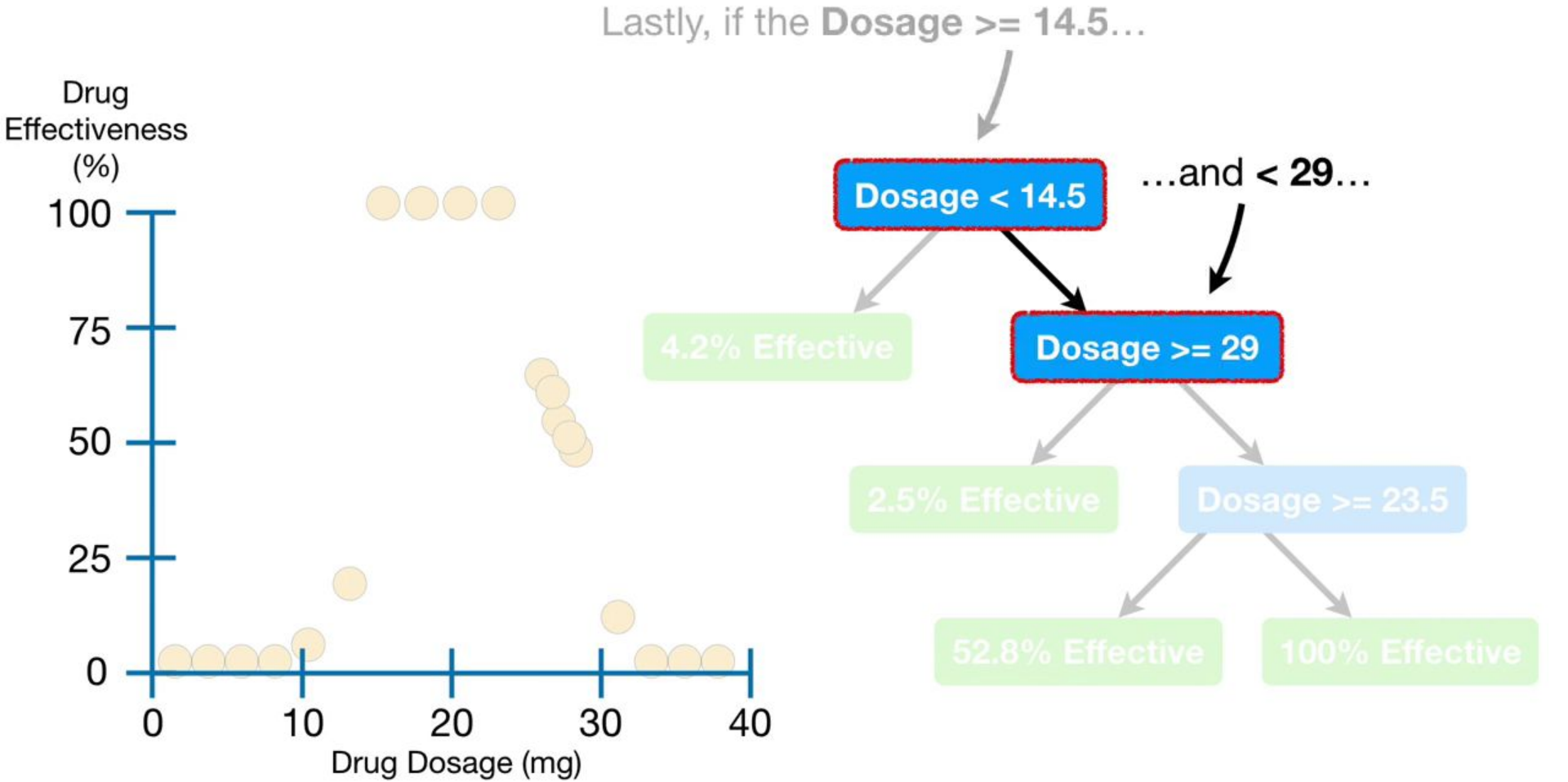
…so the tree uses the average value, **52.8%**, as its prediction for people with **Dosages** between **23.5** and **29**.

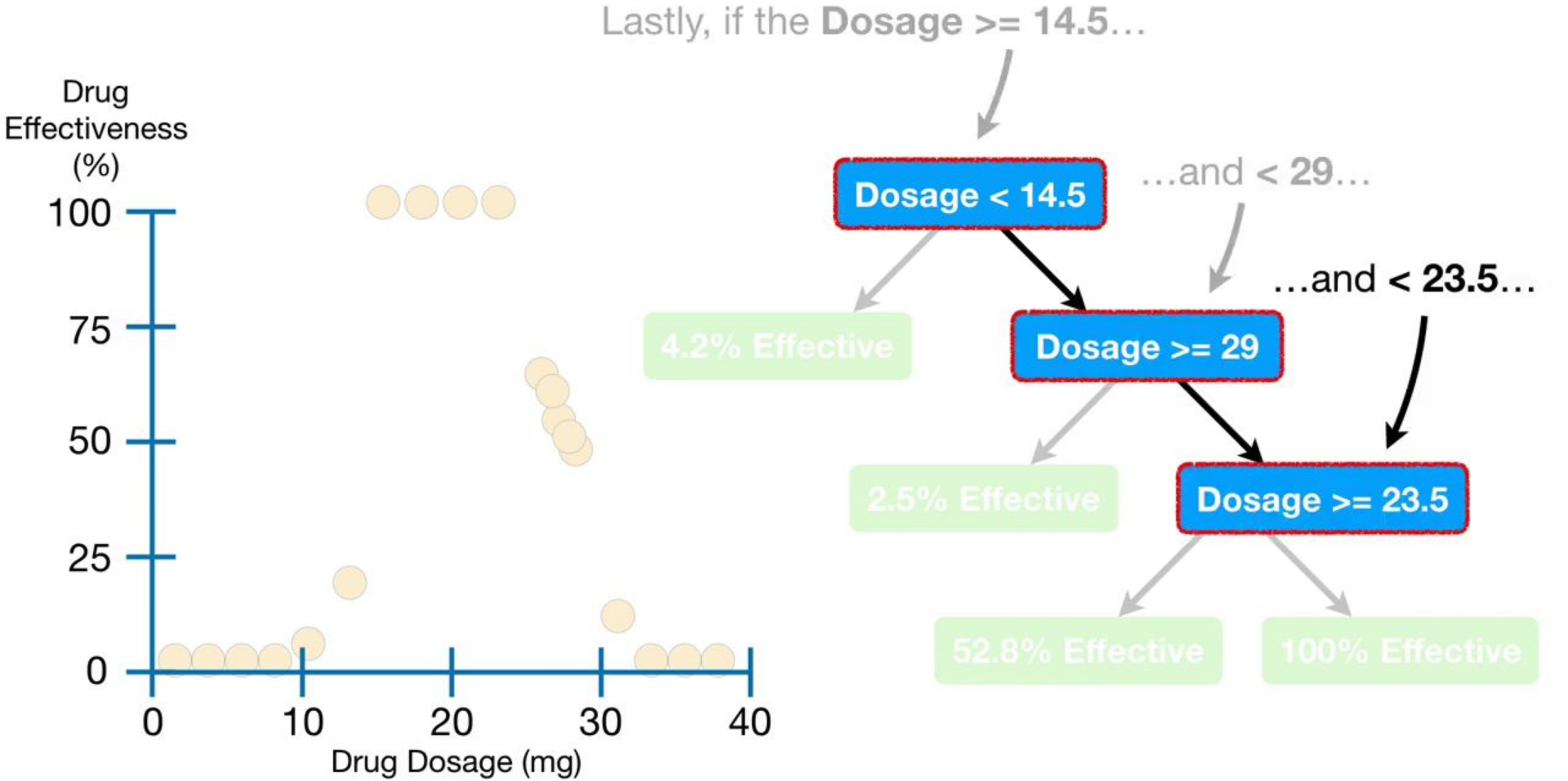


Lastly, if the **Dosage >= 14.5**…

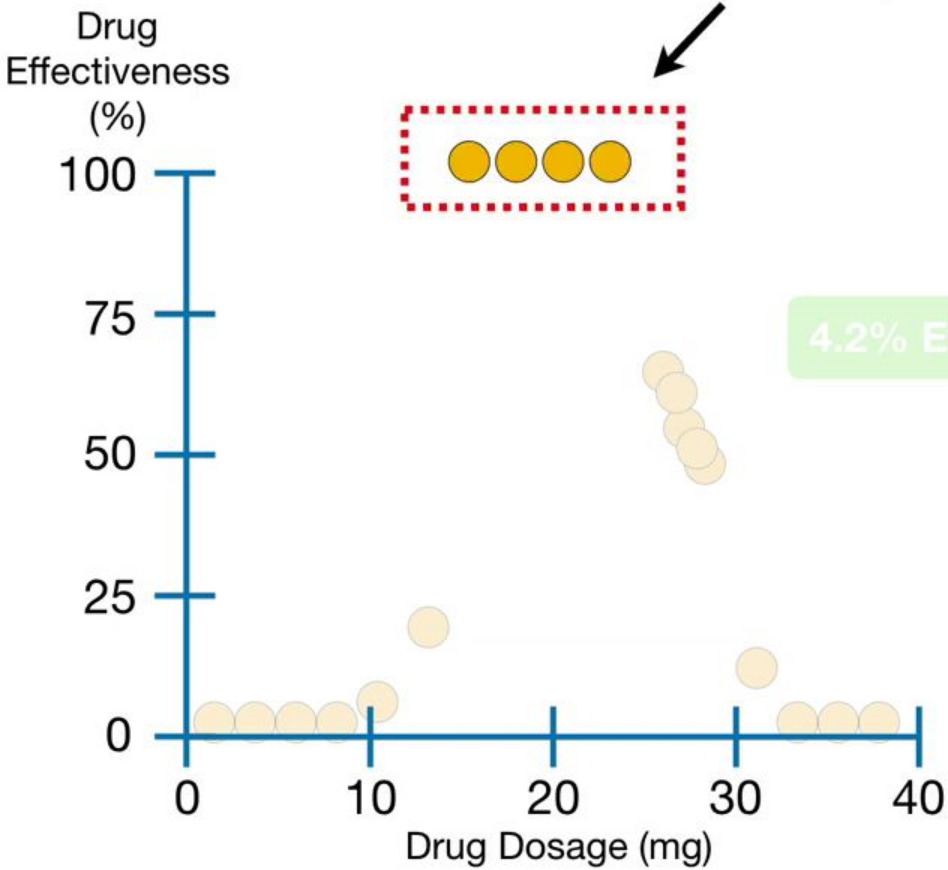


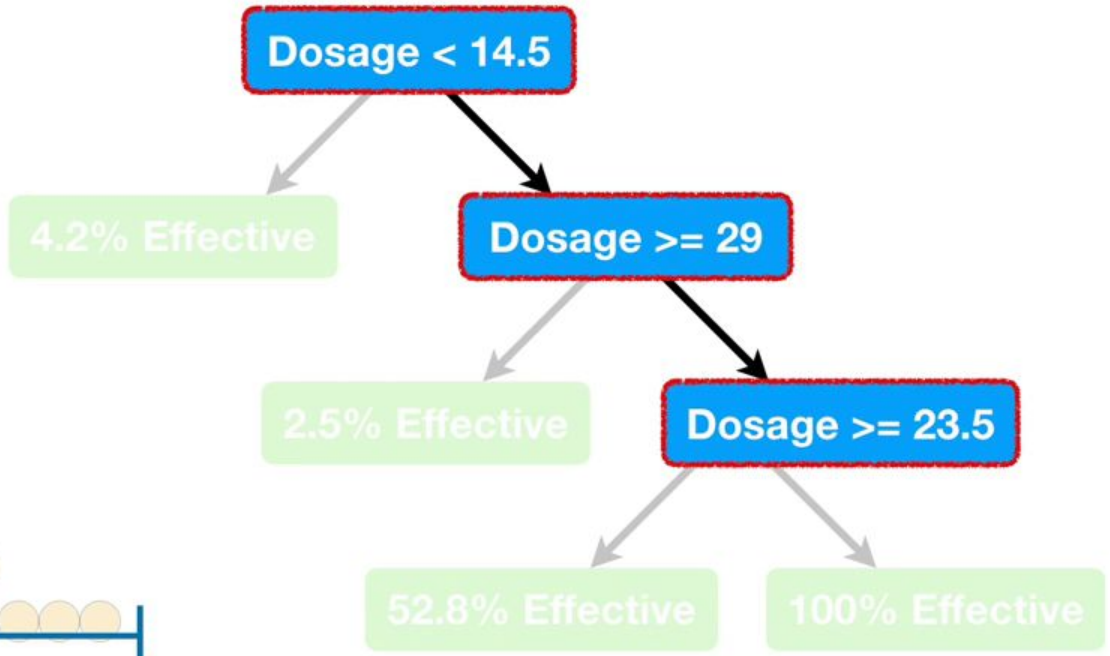



Lastly, if the Dosage >= 14.5…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
…and < 29…
4.2% Effective
Dosage >= 29
2.5% Effective
Dosage >= 23.5
52.8% Effective
100% Effective

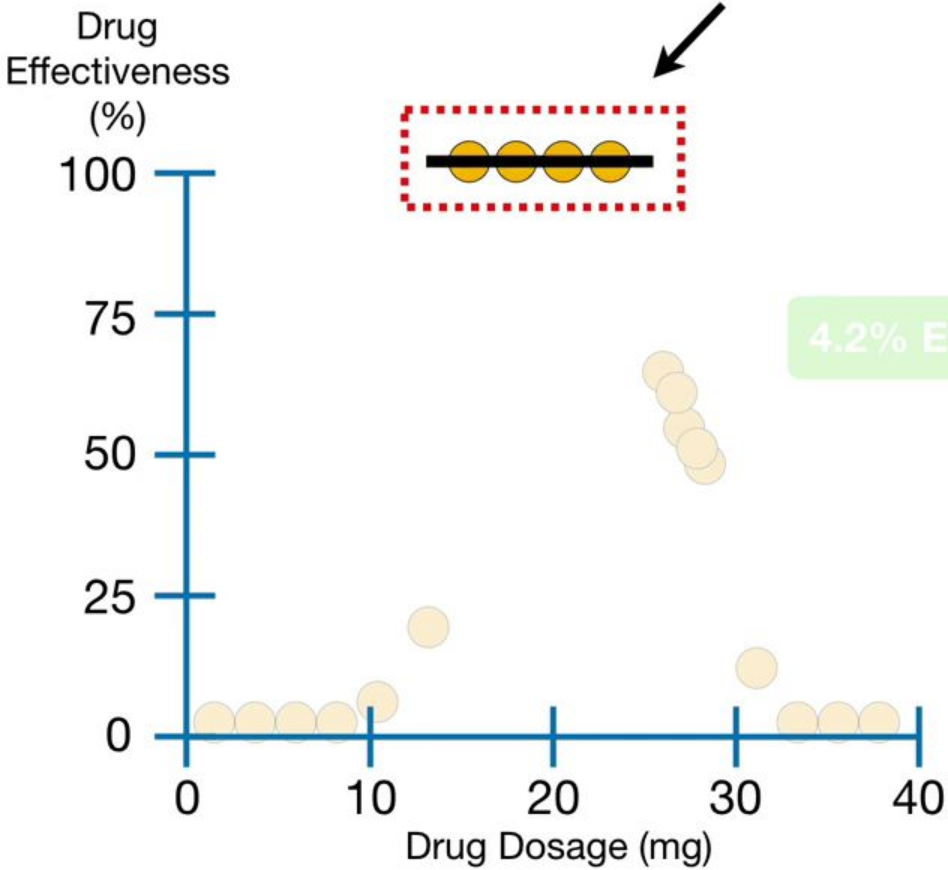

Lastly, if the Dosage >= 14.5…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
…and < 29…
…and < 23.5…
4.2% Effective
Dosage >= 29
2.5% Effective
Dosage >= 23.5
52.8% Effective
100% Effective

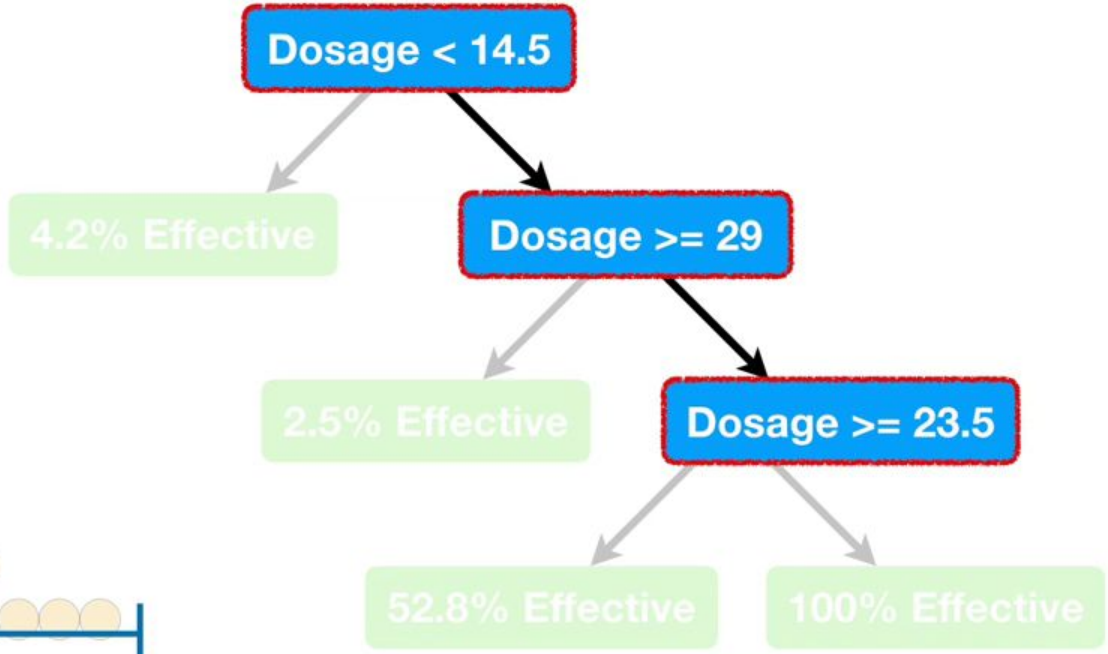
…then we are talking about these **4** observations in the training dataset…



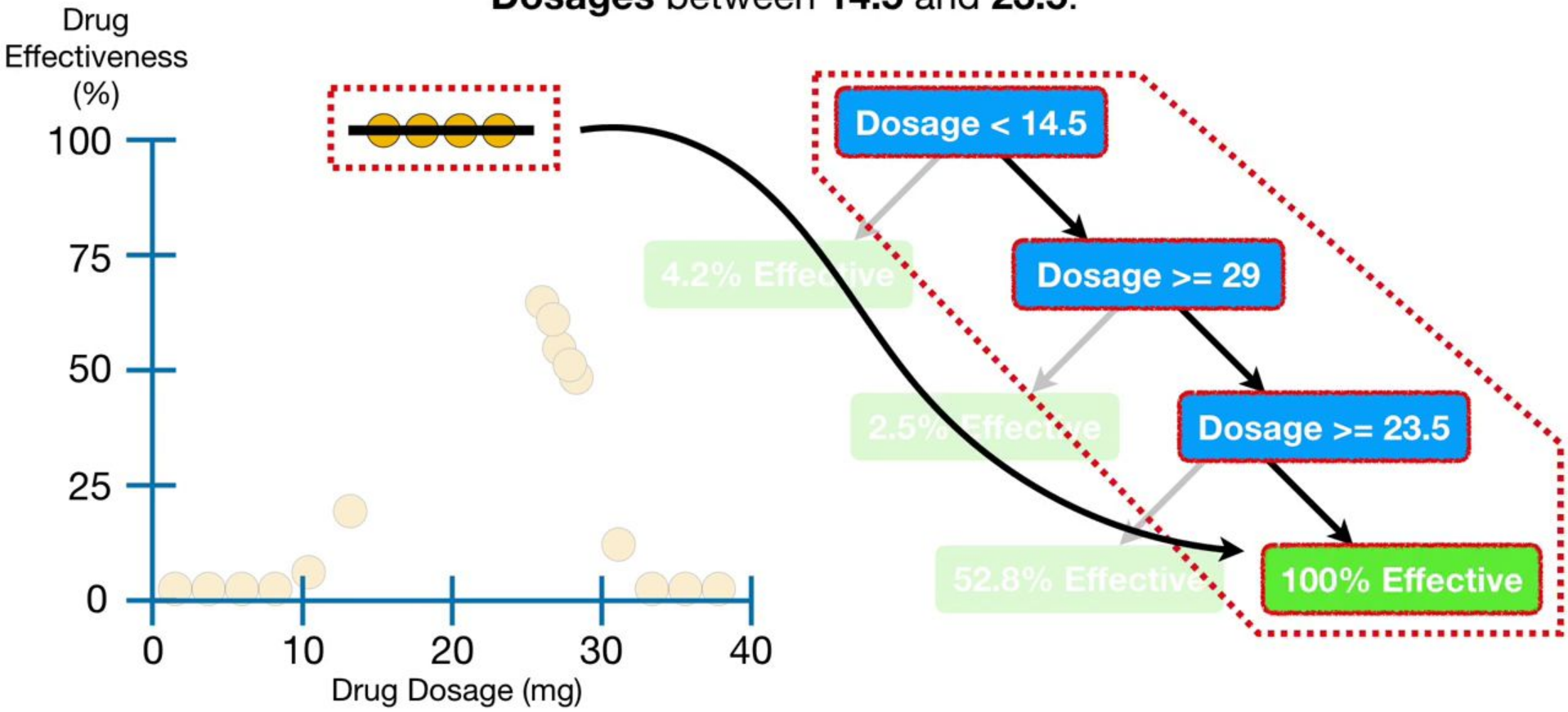


…and the average **Drug Effectiveness** for these **4** observations is **100%**…

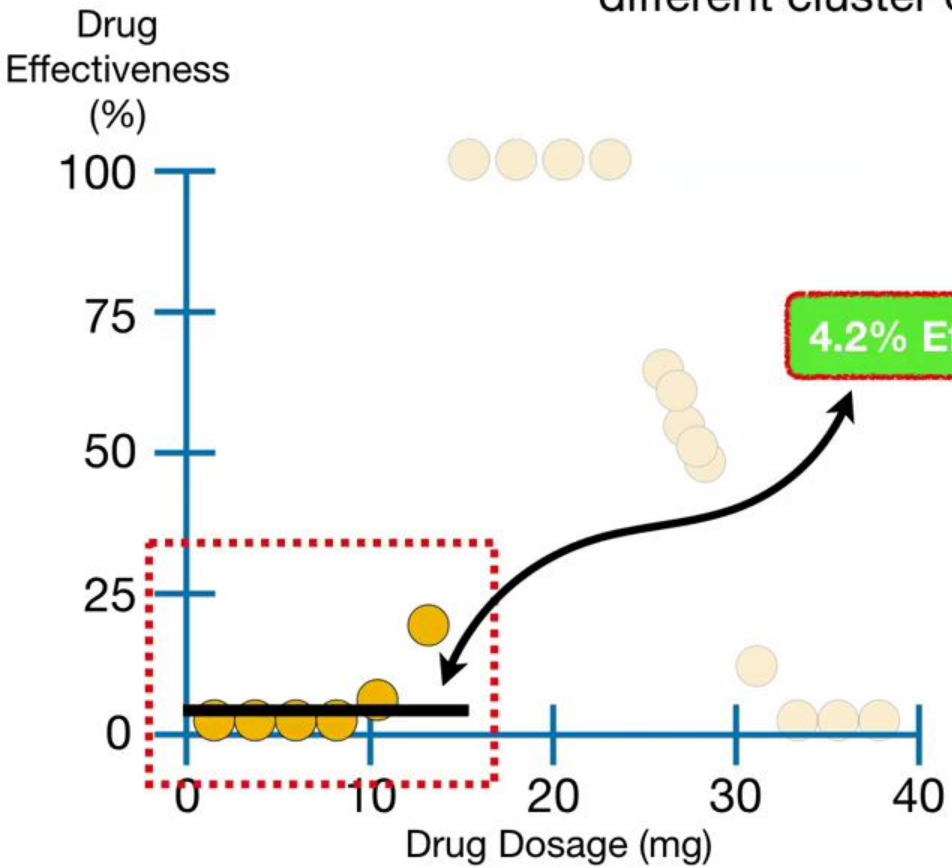


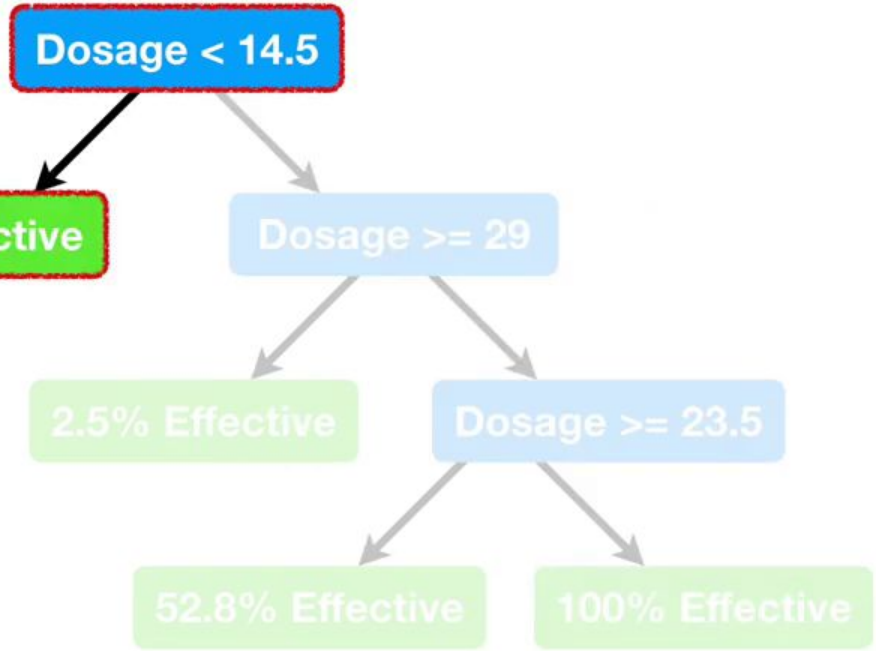


…so the tree uses the average value, **100%**, as its prediction for people with **Dosages** between **14.5** and **23.5**.

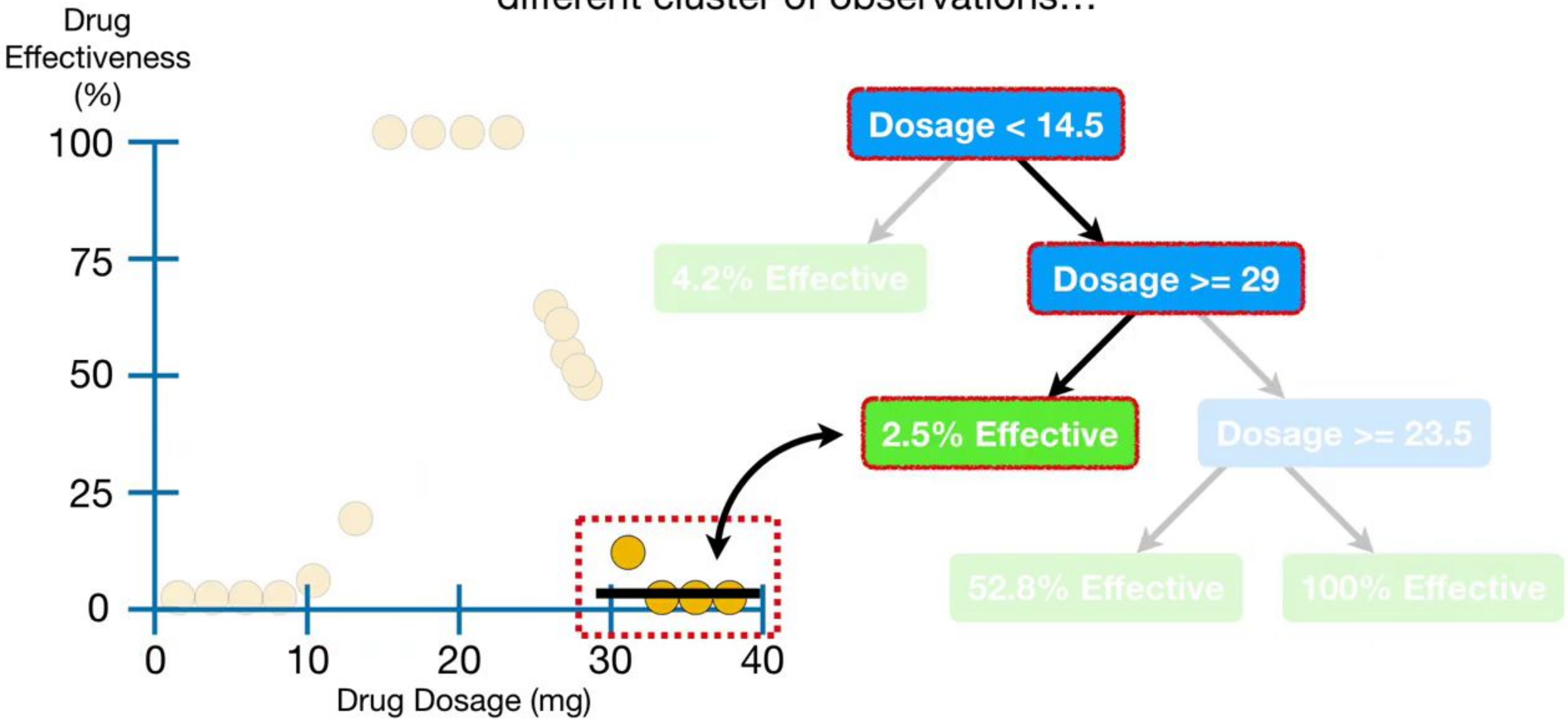


Since each leaf corresponds to the average **Drug Effectiveness** in a different cluster of observations…

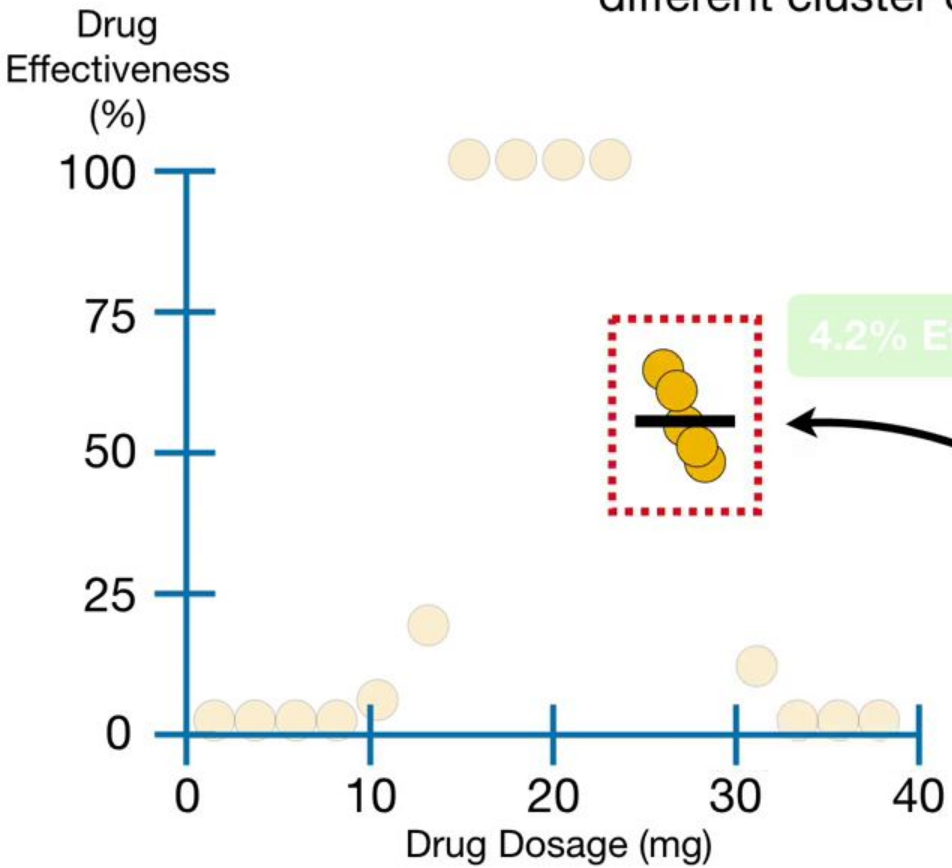


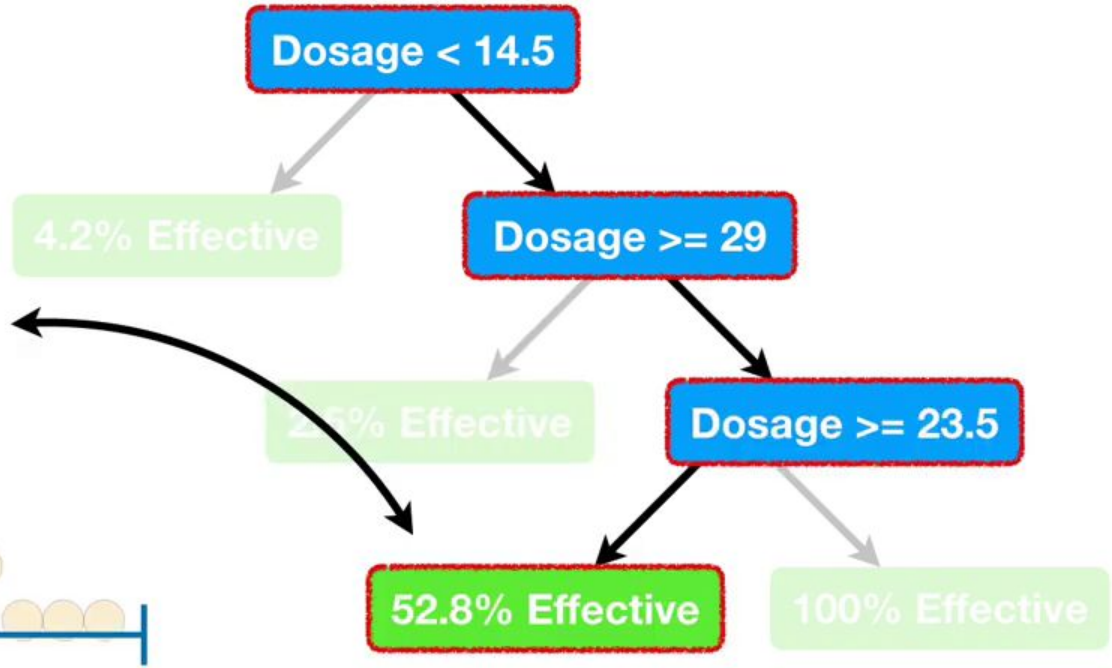


Since each leaf corresponds to the average **Drug Effectiveness** in a different cluster of observations…

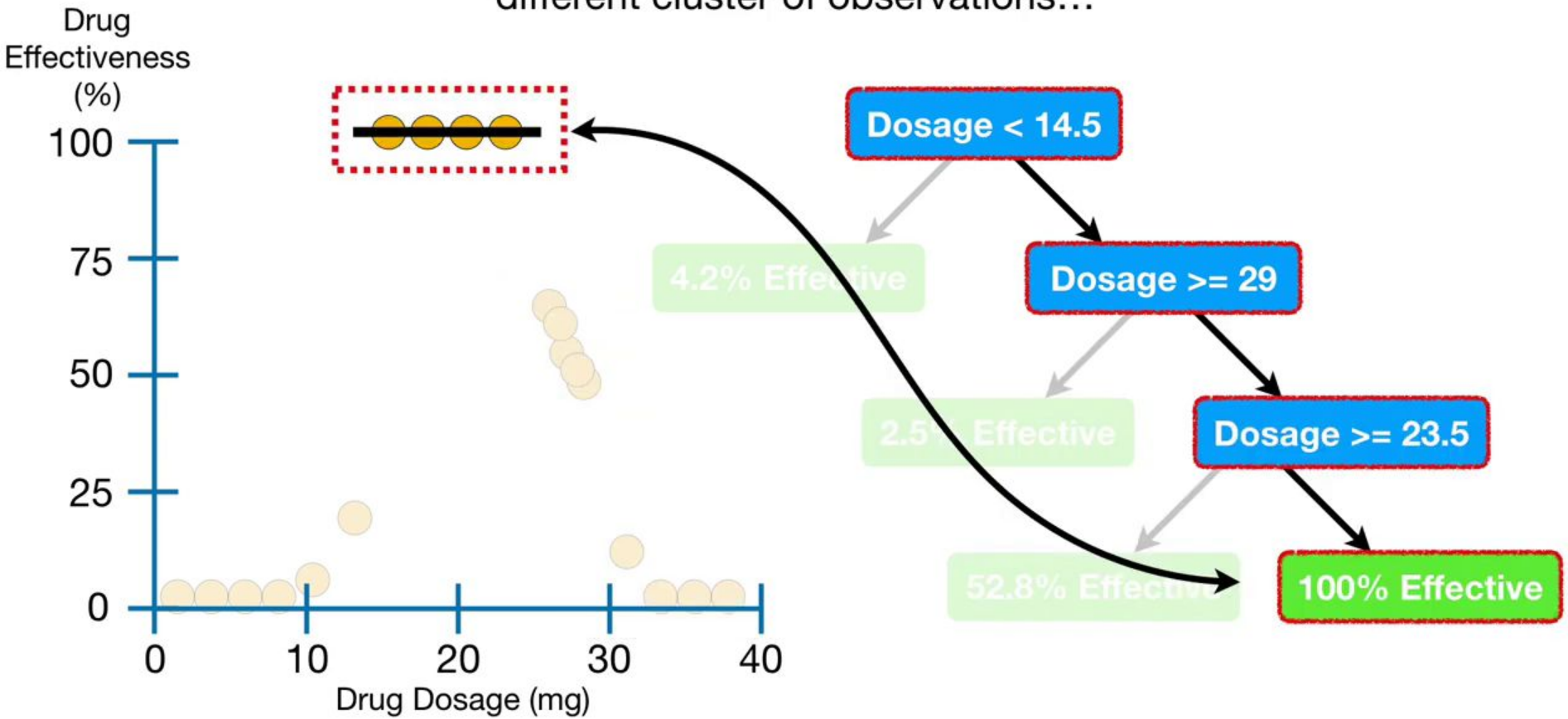


Since each leaf corresponds to the average **Drug Effectiveness** in a different cluster of observations…

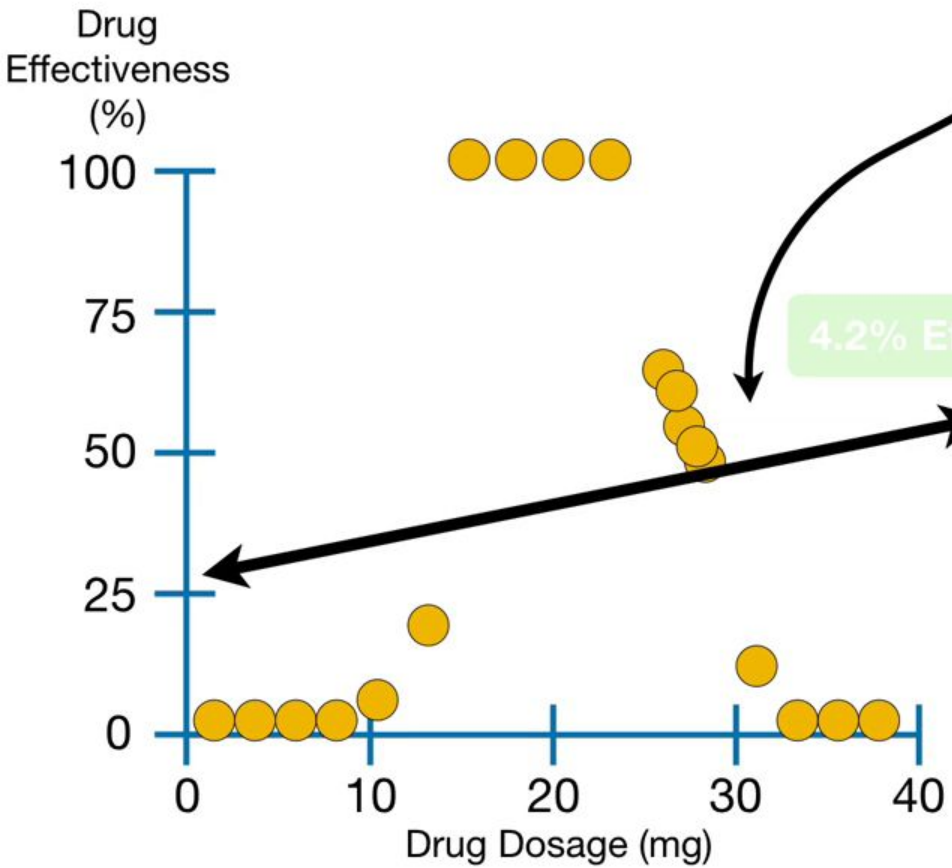


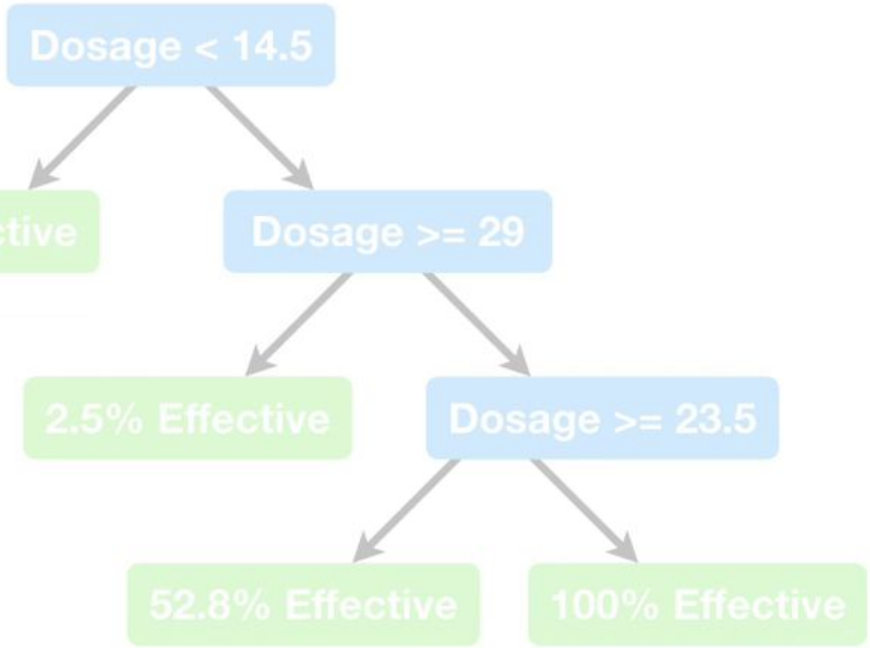


Since each leaf corresponds to the average **Drug Effectiveness** in a different cluster of observations…

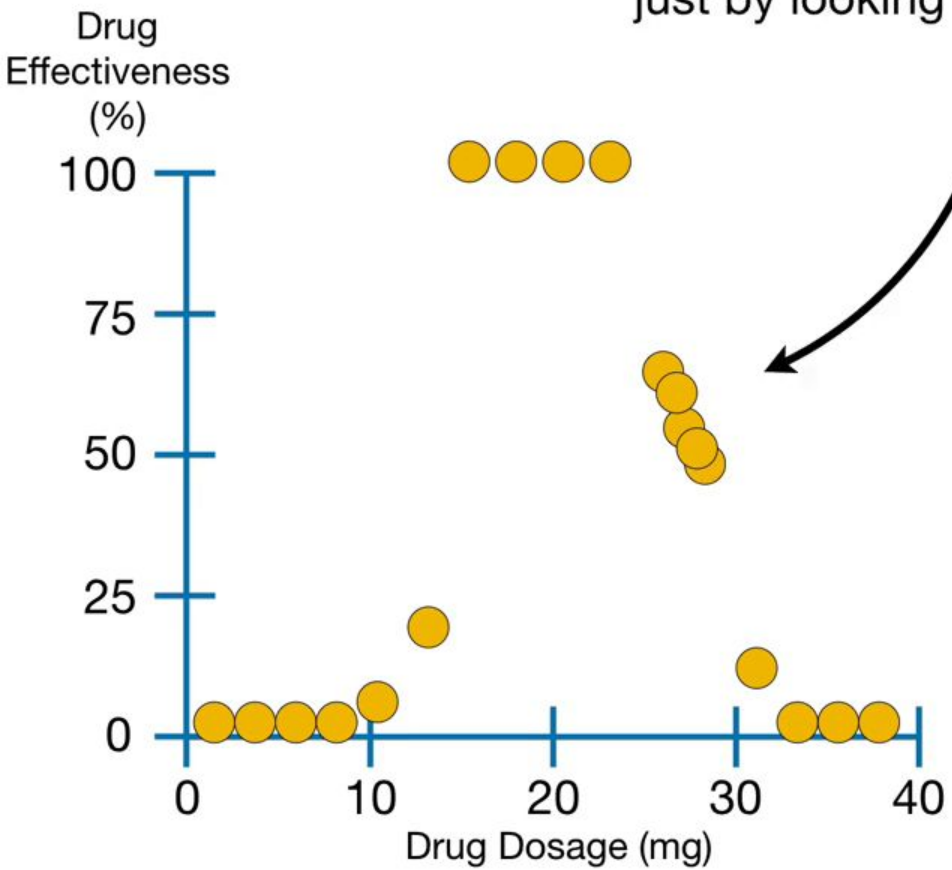


…the tree does a better job reflecting the data than the straight line.

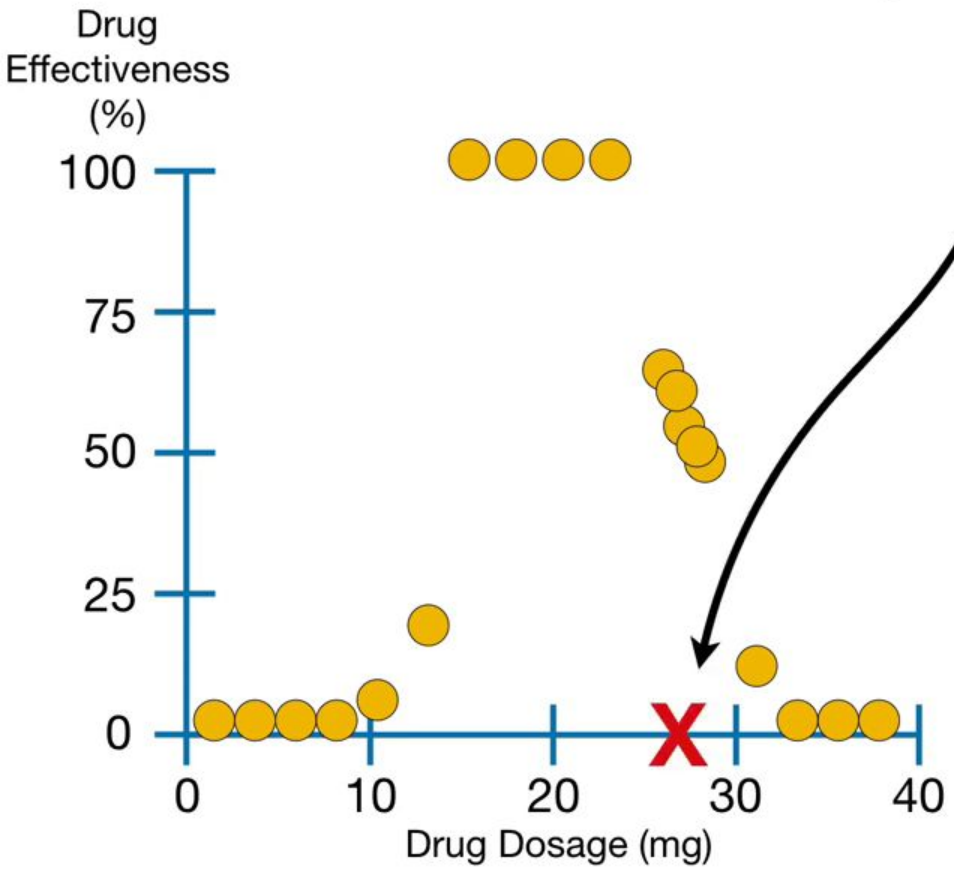




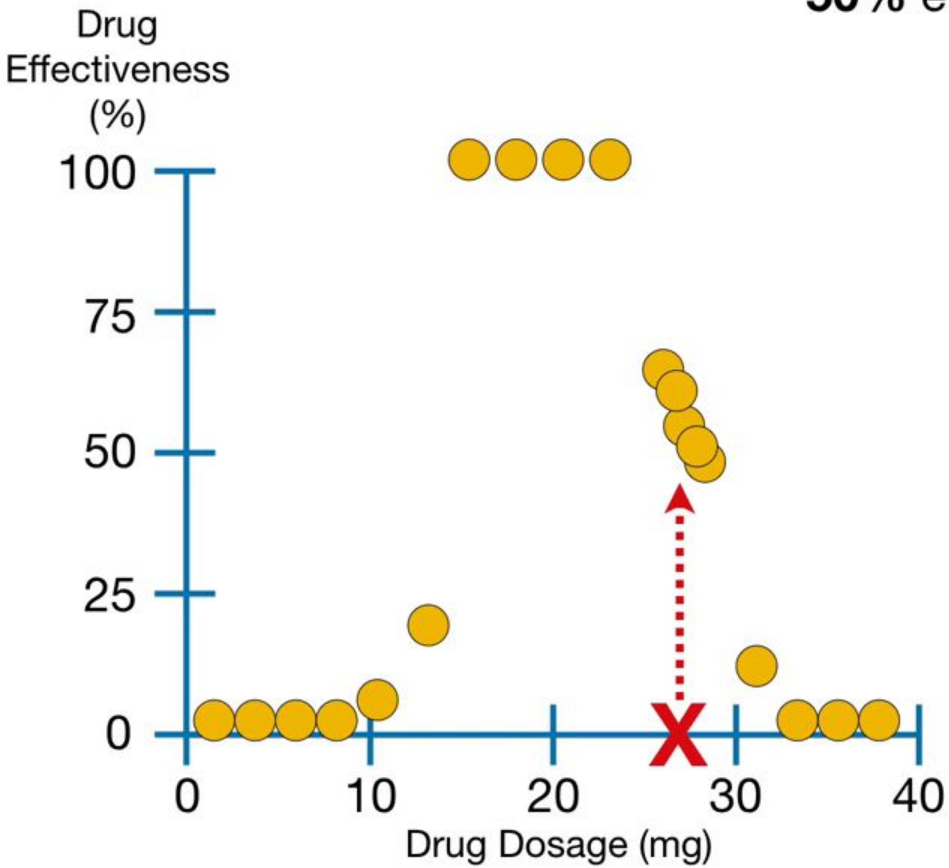
At this point you might be thinking, “The **Regression Tree** is cool, but I can also predict **Drug Effectiveness** just by looking at the graph…”



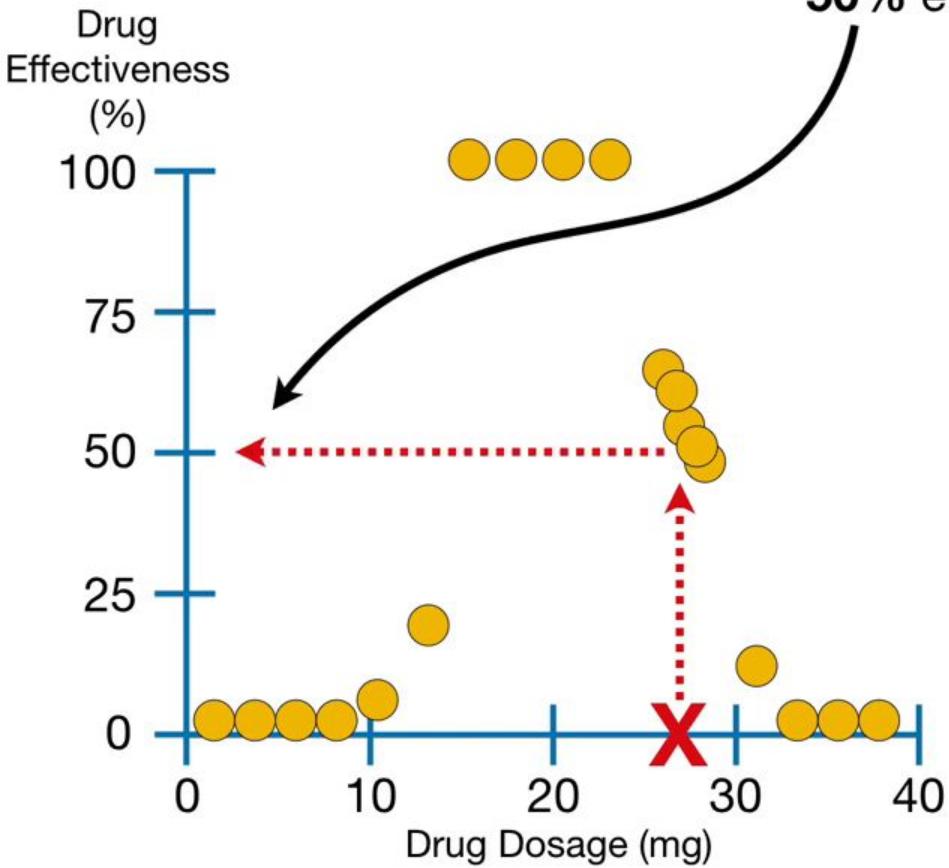
For example, if someone said they were taking a **27 mg** dose…



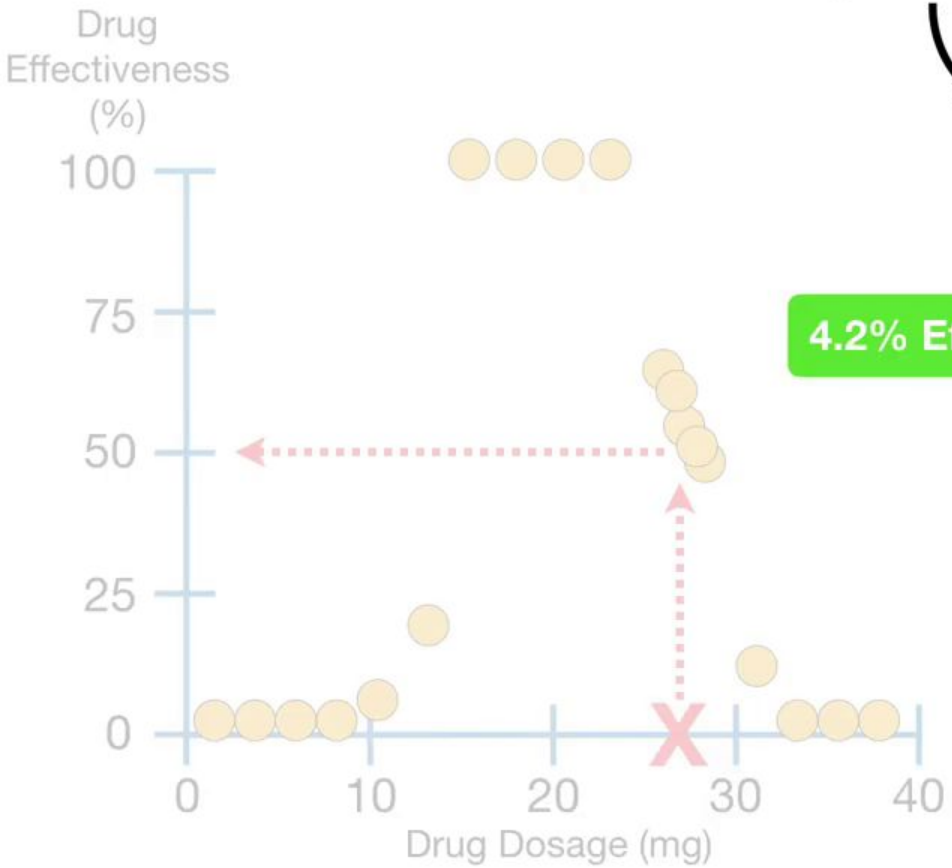
…then, just by looking at the graph,
I can tell that the drug will be about
50% effective.

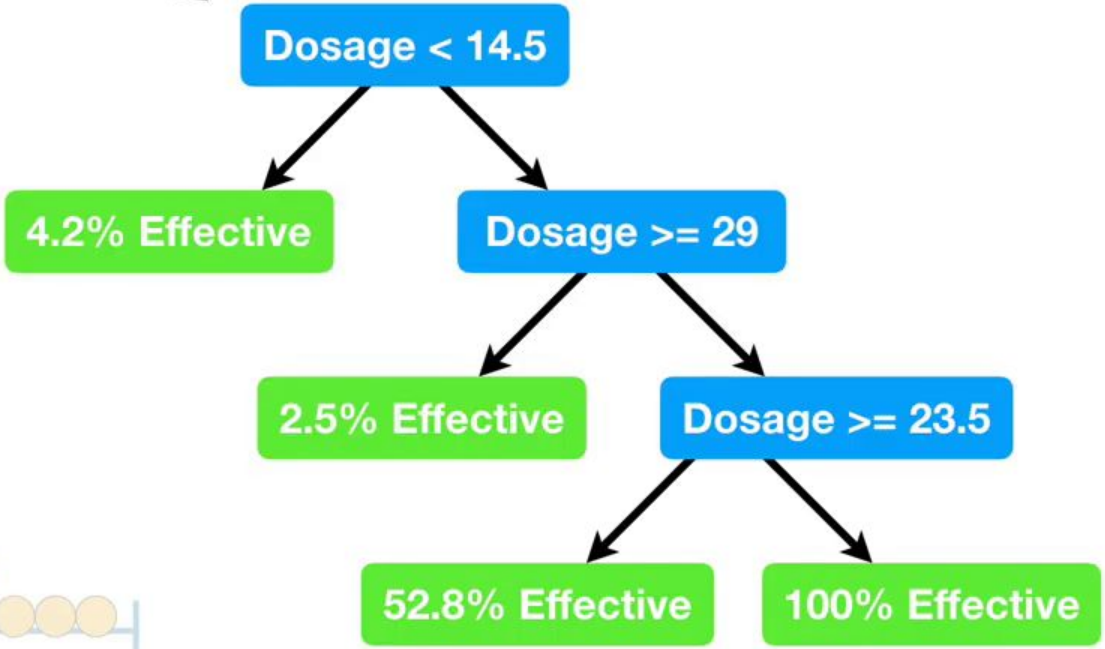


…then, just by looking at the graph,
I can tell that the drug will be about
50% effective.

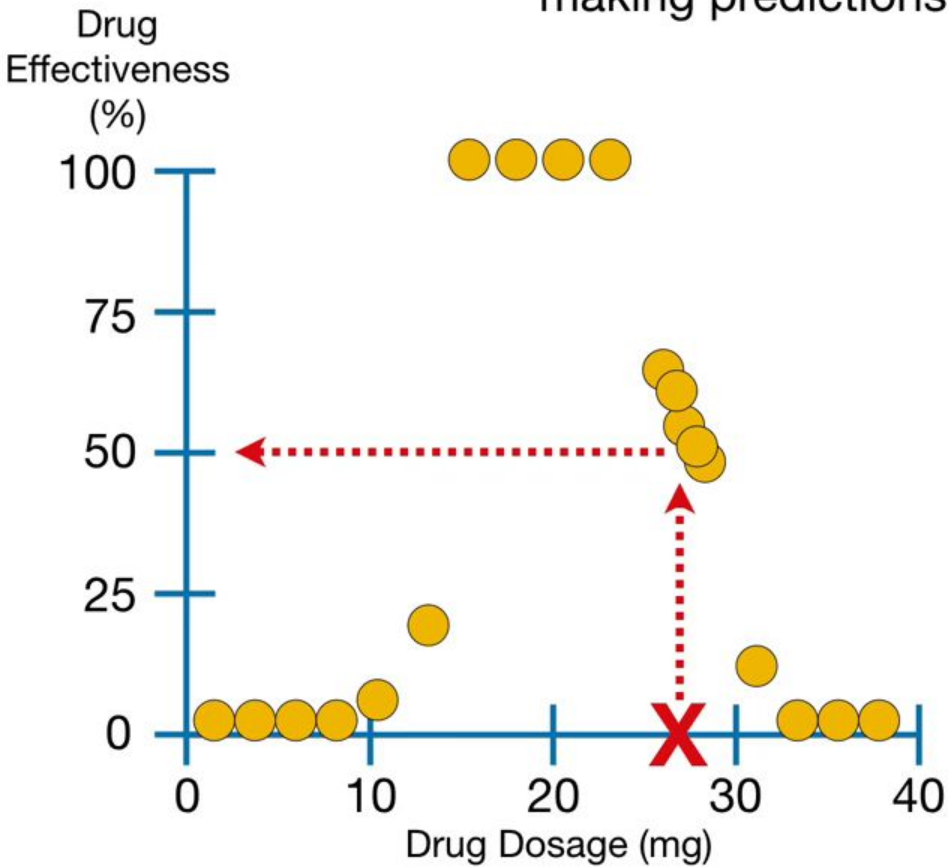


So why make a big deal about the **Regression Tree**?





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.

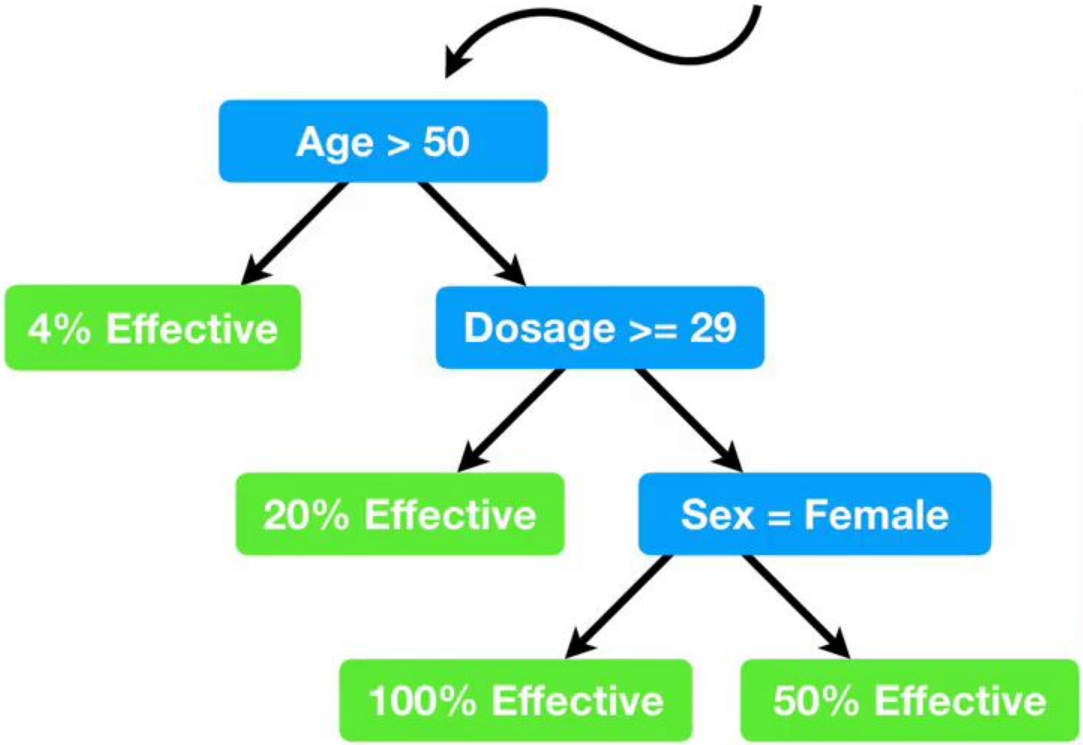


Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

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.

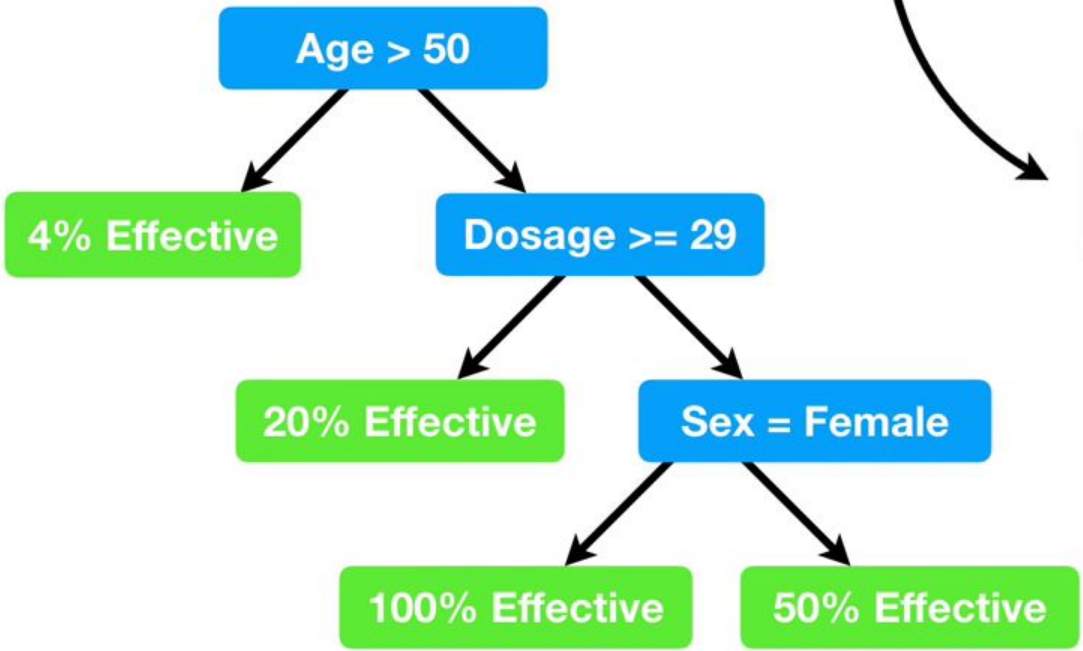
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

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



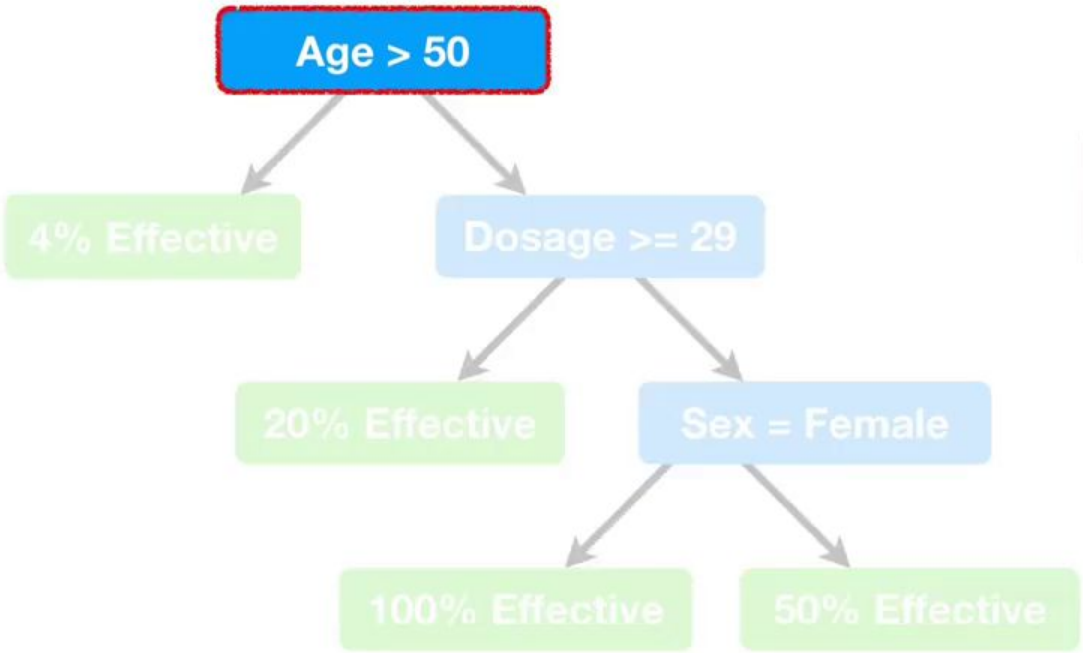
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

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



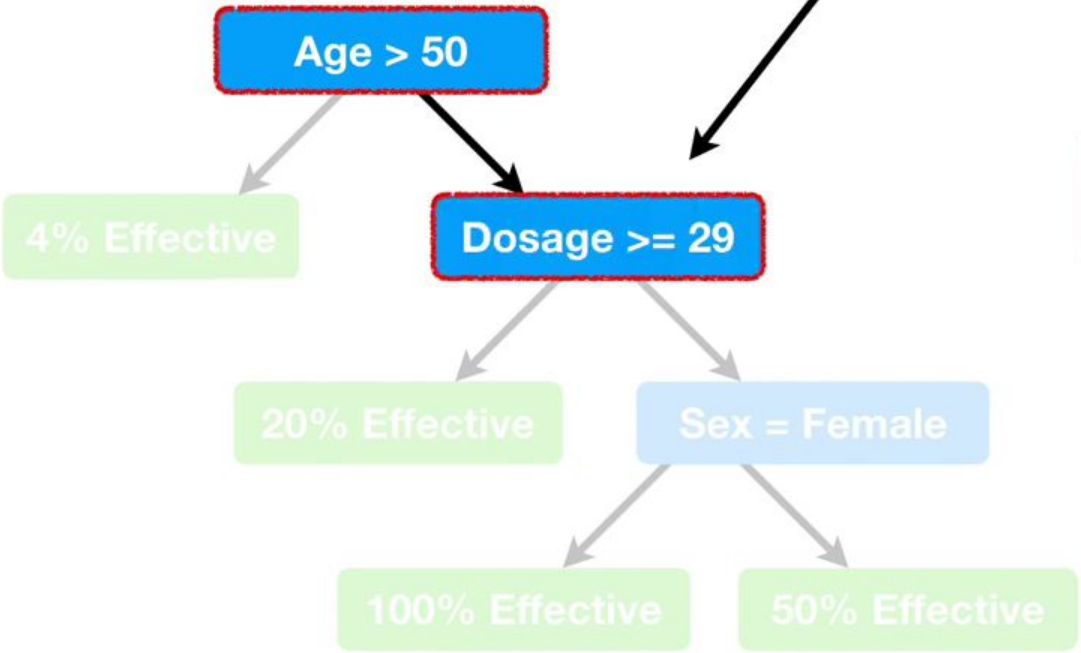
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

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



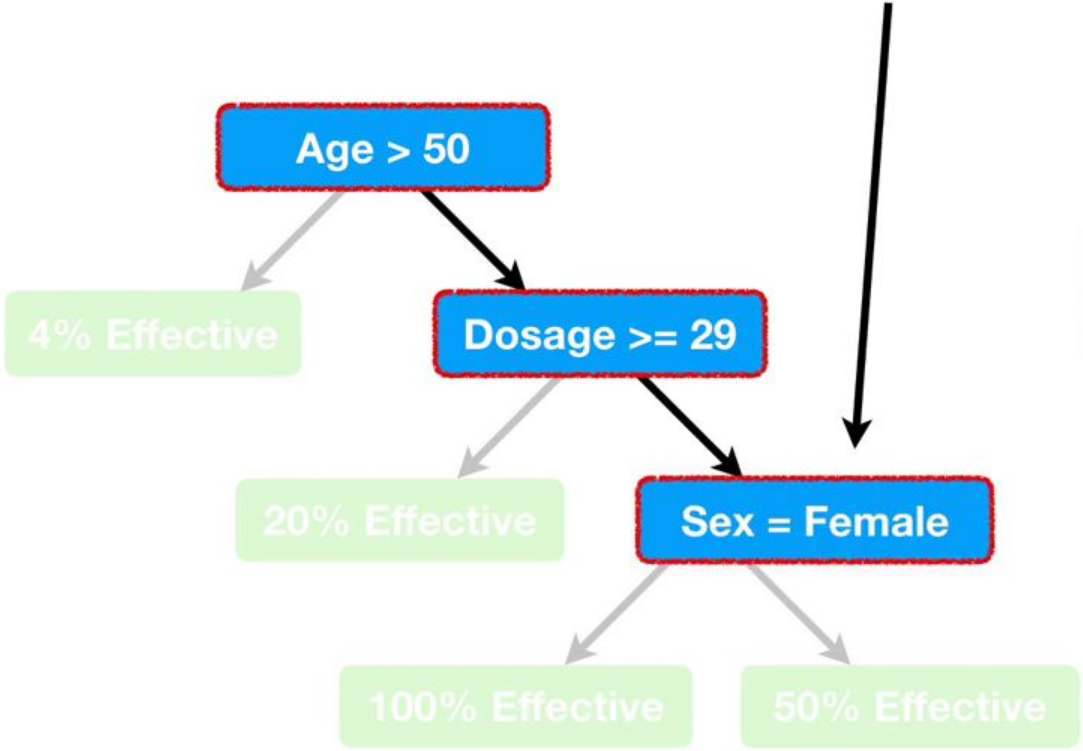
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

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



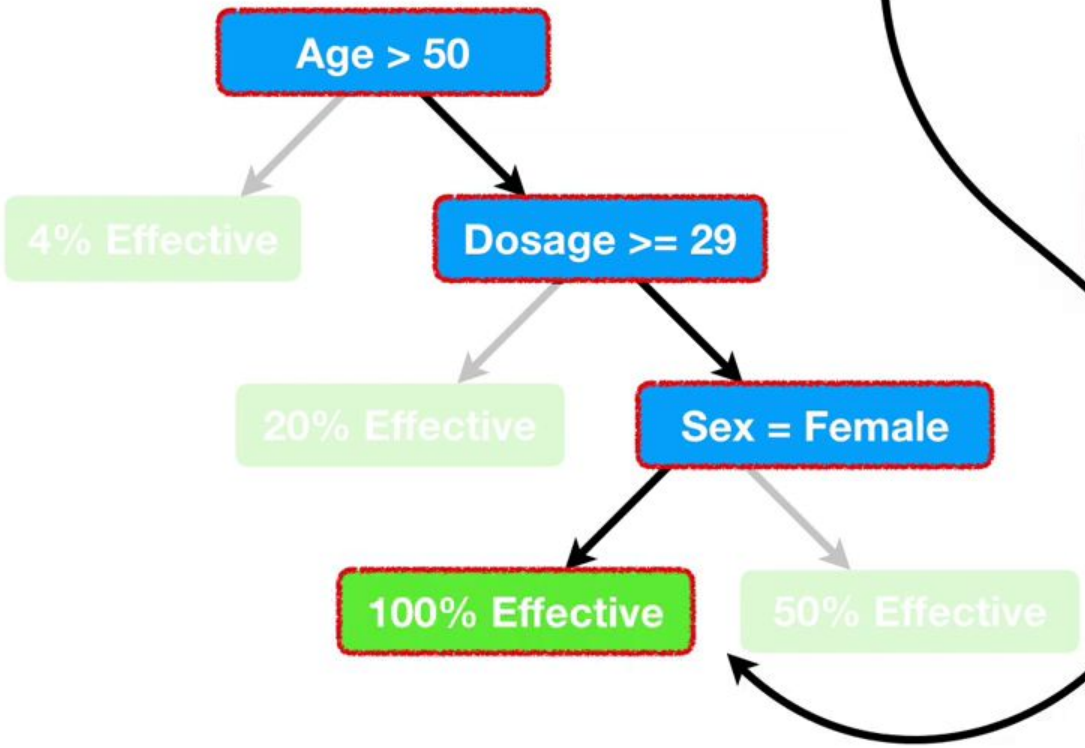
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

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



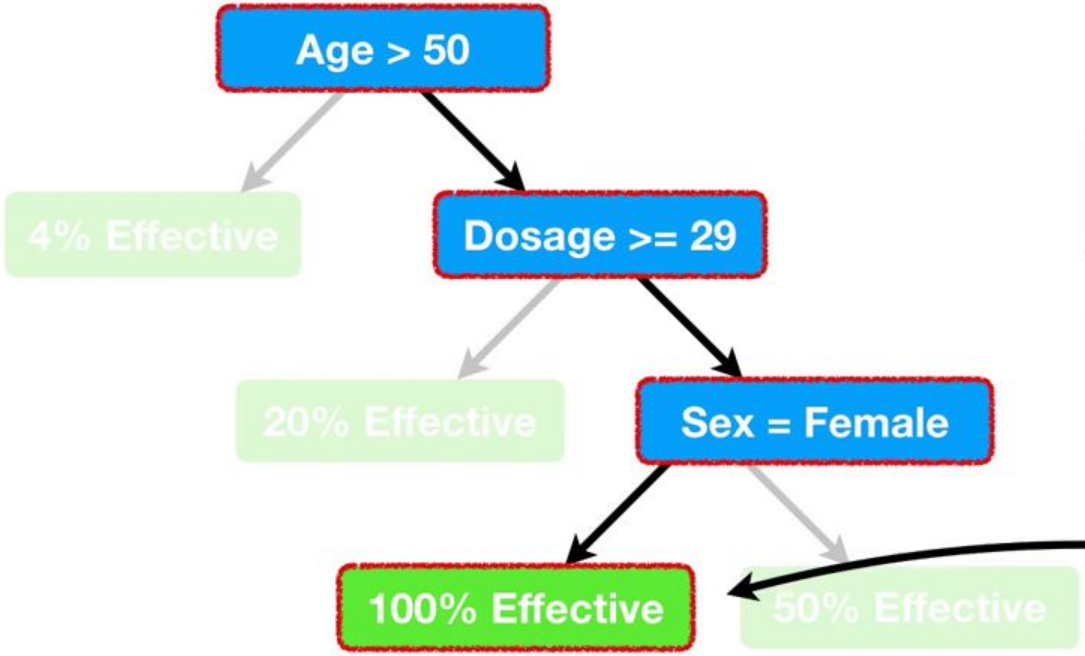
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

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



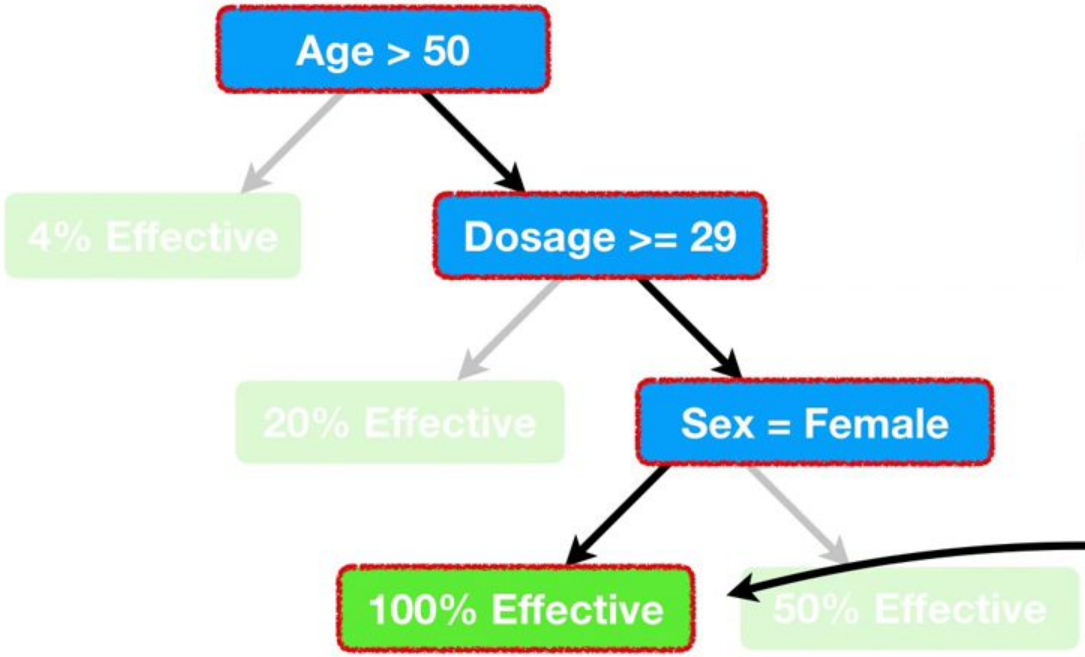
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

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



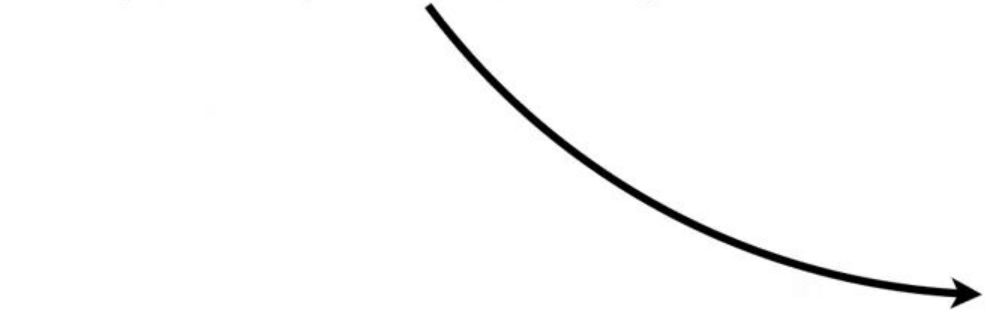
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

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



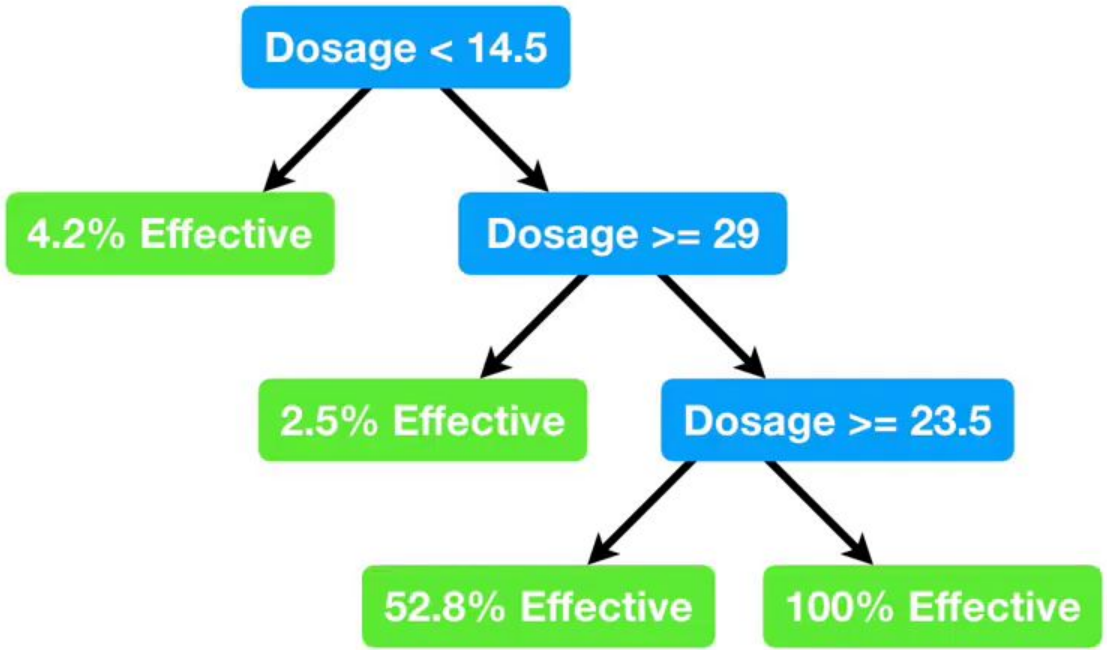
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

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



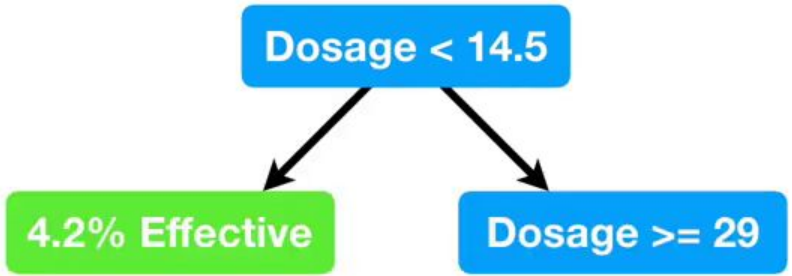
Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

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



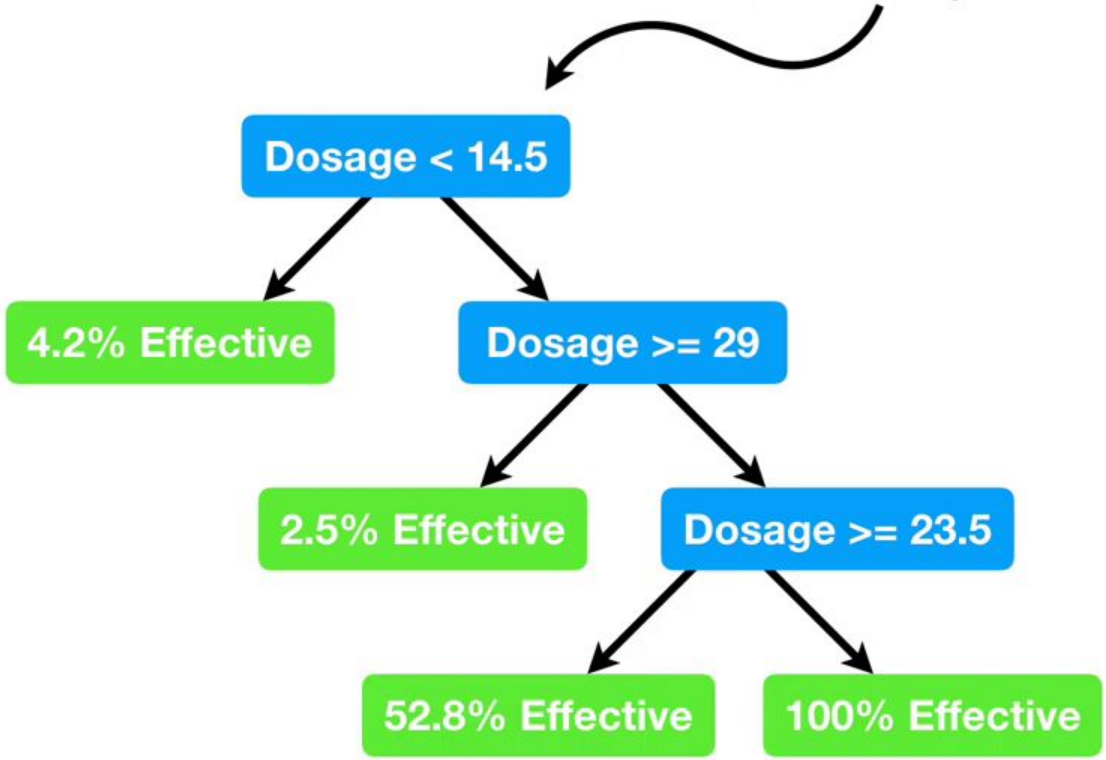
Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

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



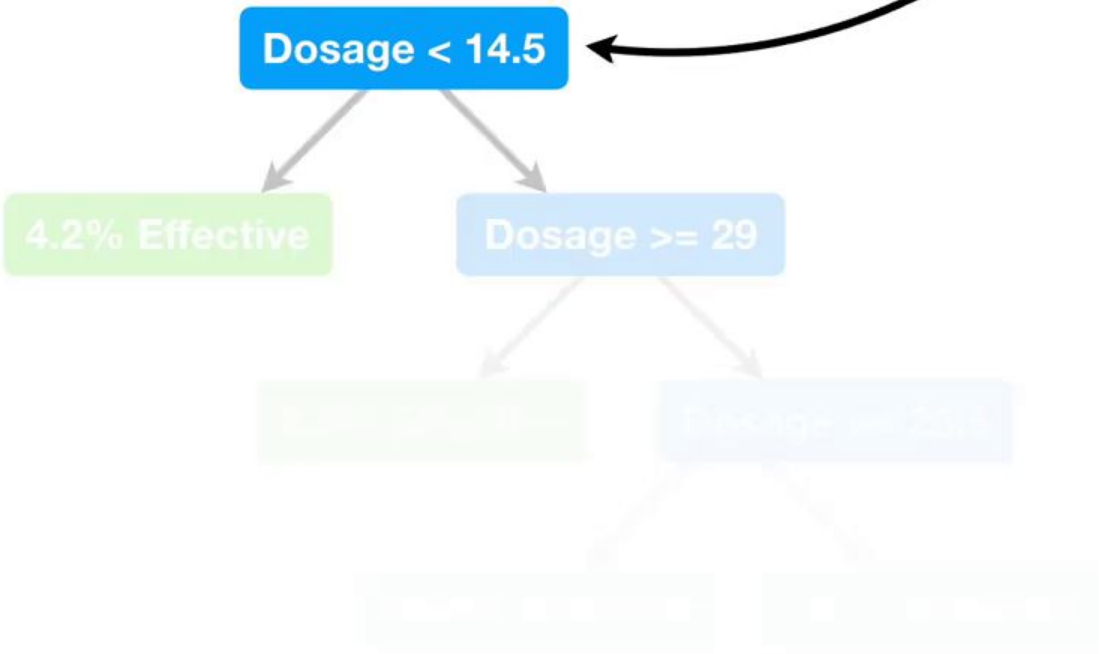
Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

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

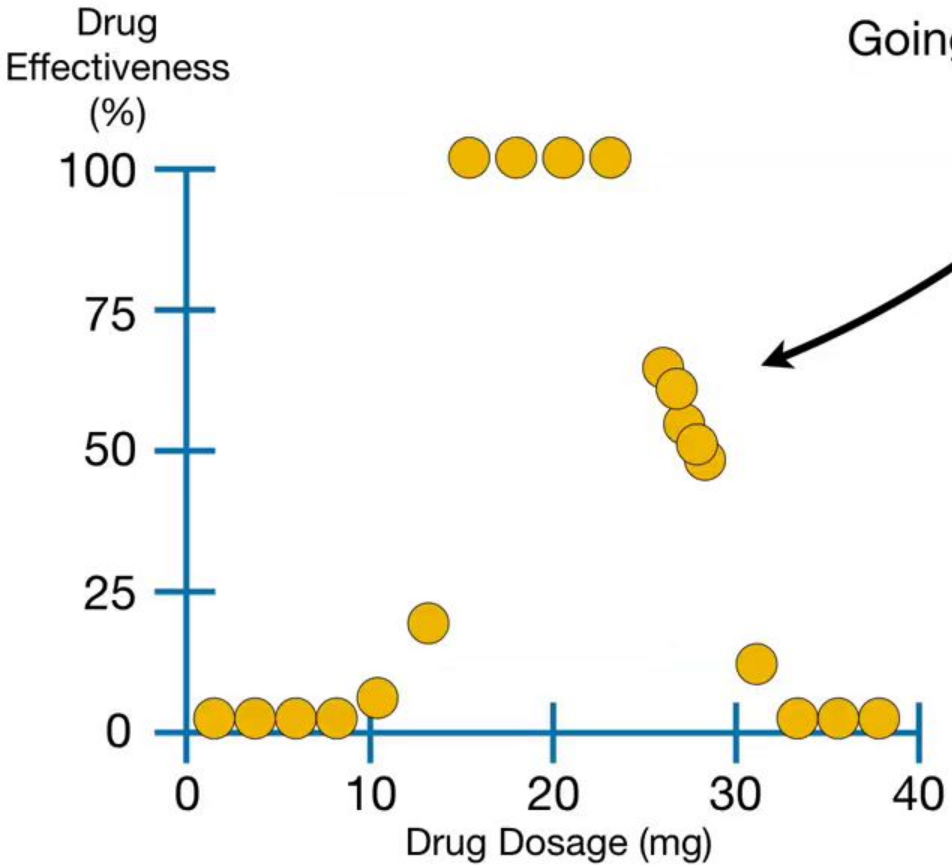


Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

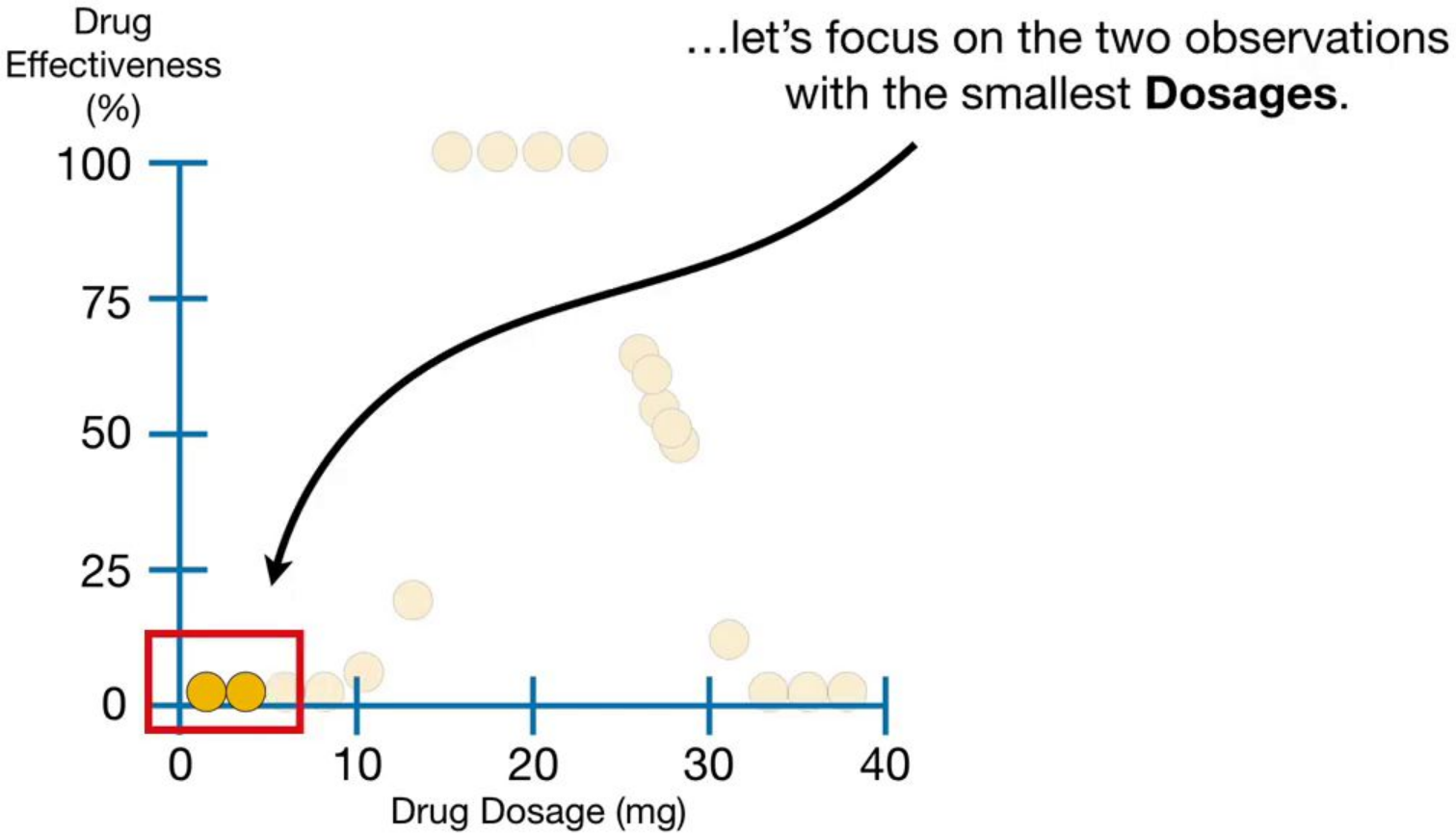
…the first thing we do is figure out why we start by asking if **Dosage < 14.5**.

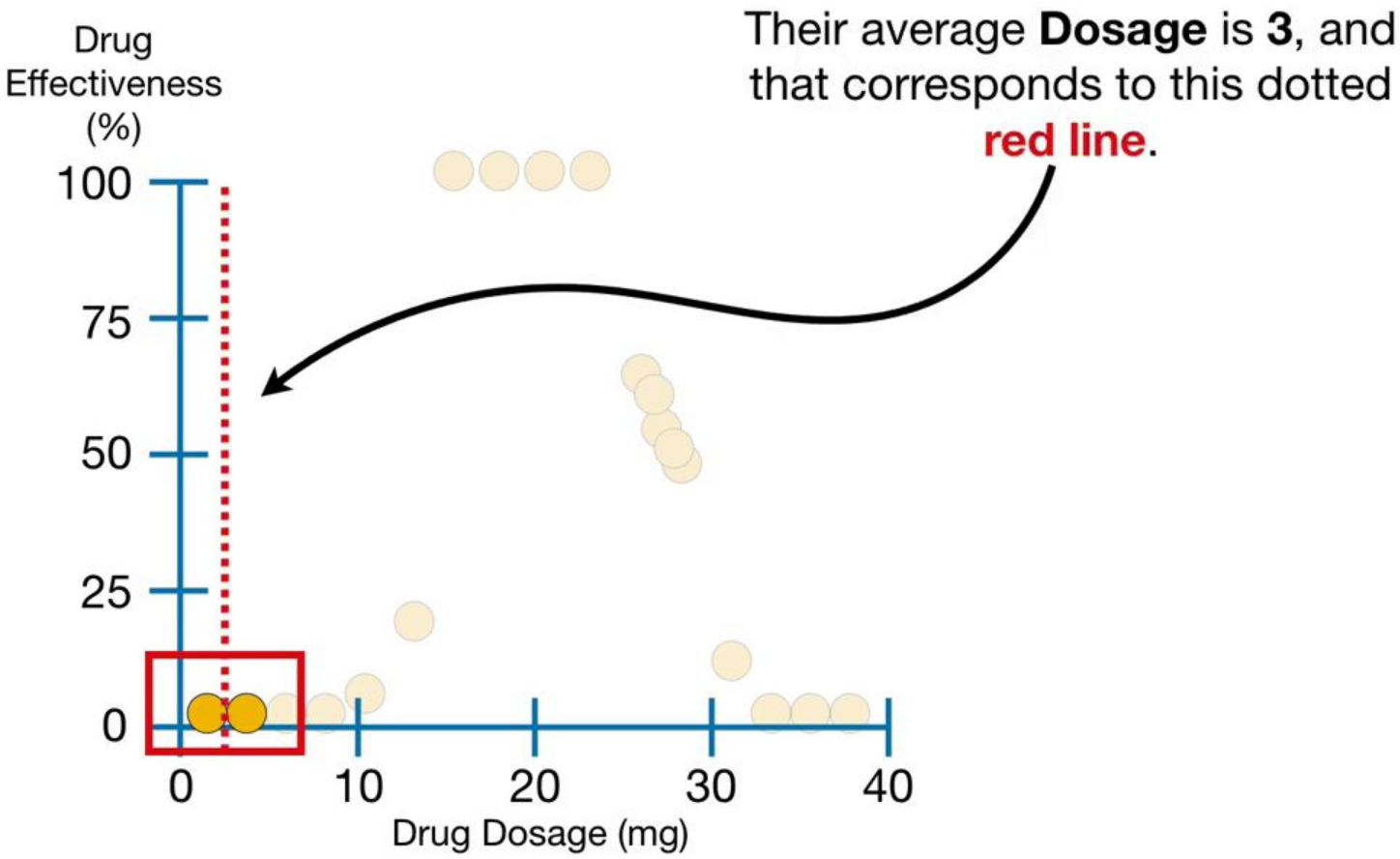


Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

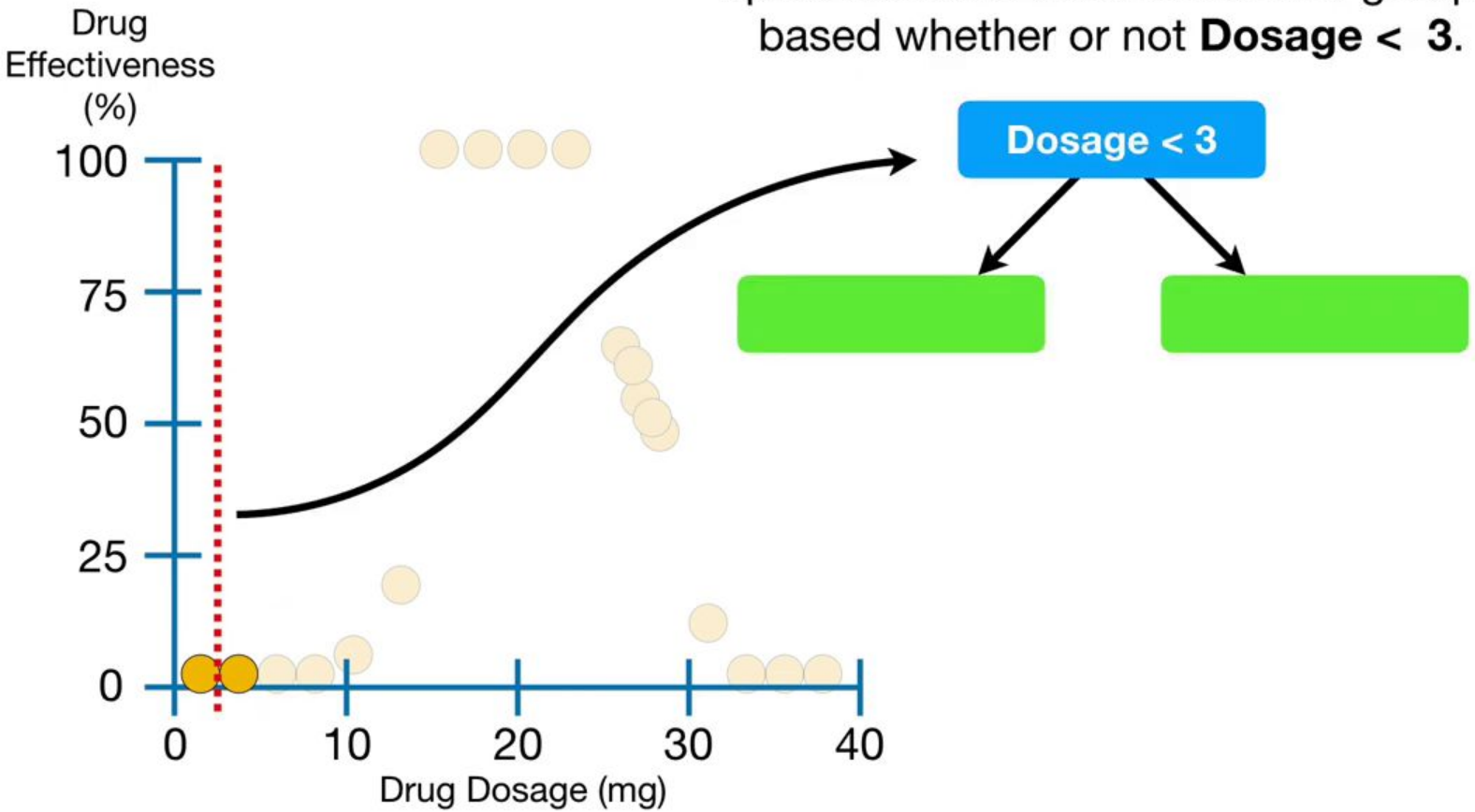


Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

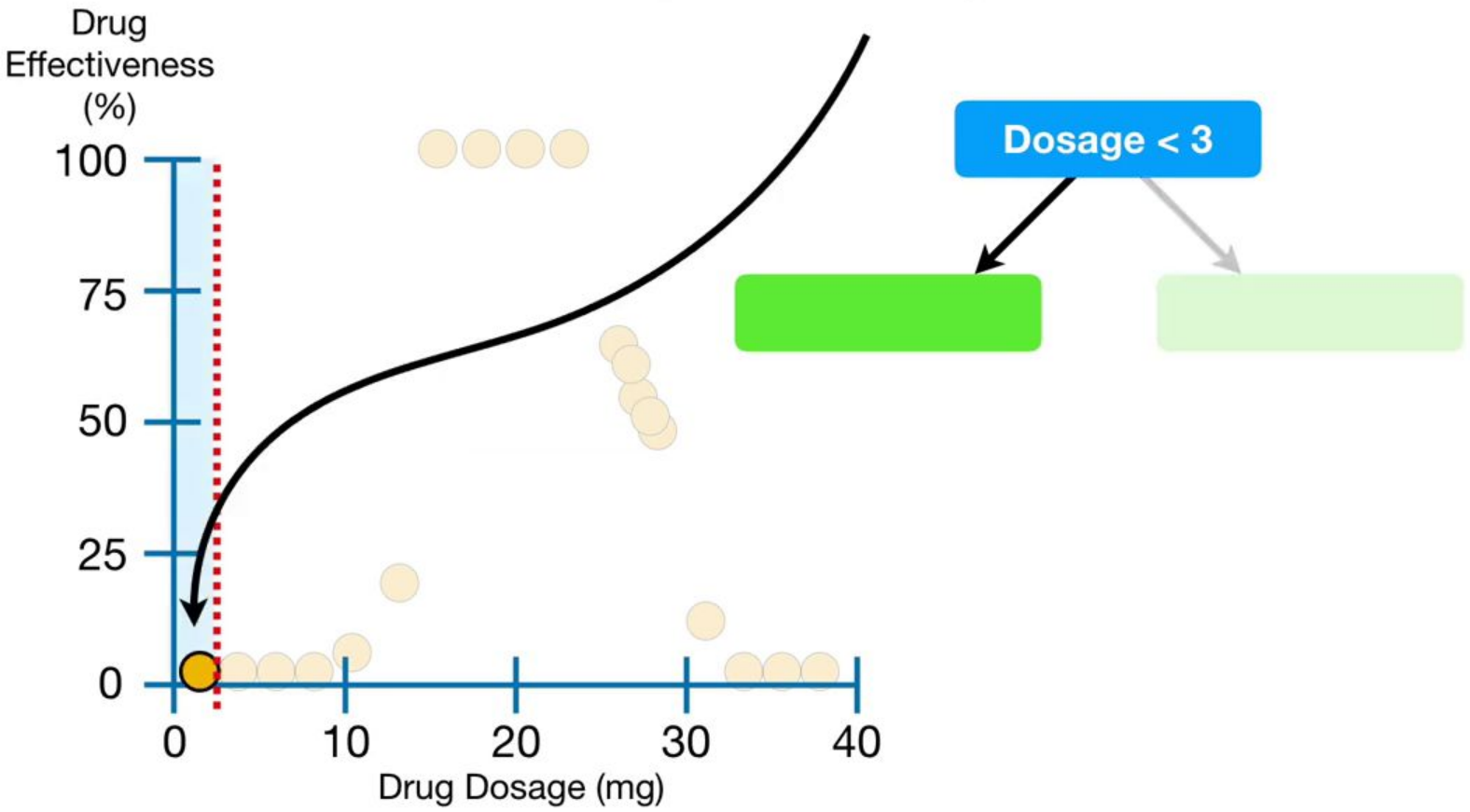

Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
…let's focus on the two observations with the smallest **Dosages**.


Drug
Effectiveness
(%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Their average **Dosage** is **3**, and
that corresponds to this dotted
red line.

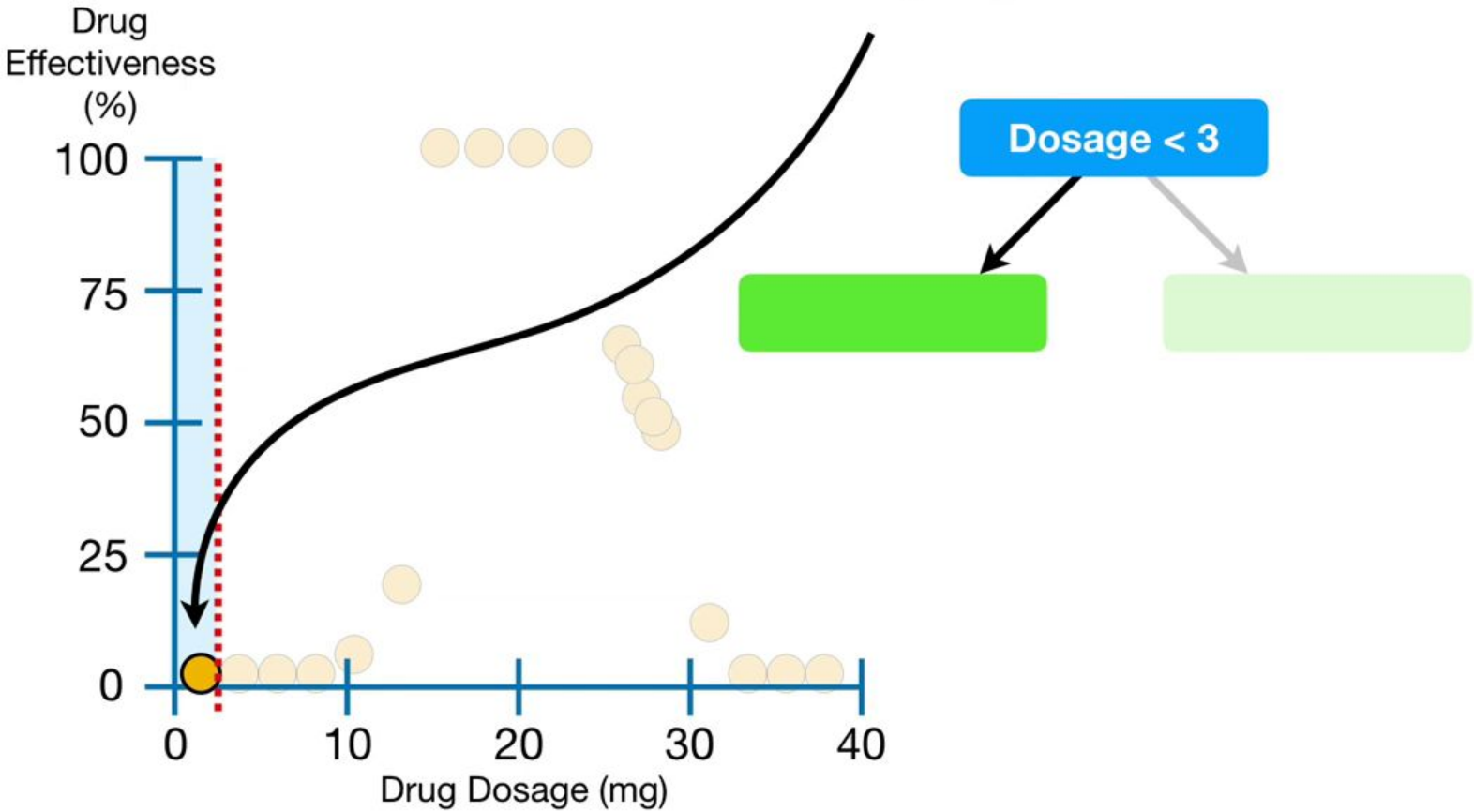
Now we can build a very simple tree that splits the observations into two groups based whether or not **Dosage < 3**.



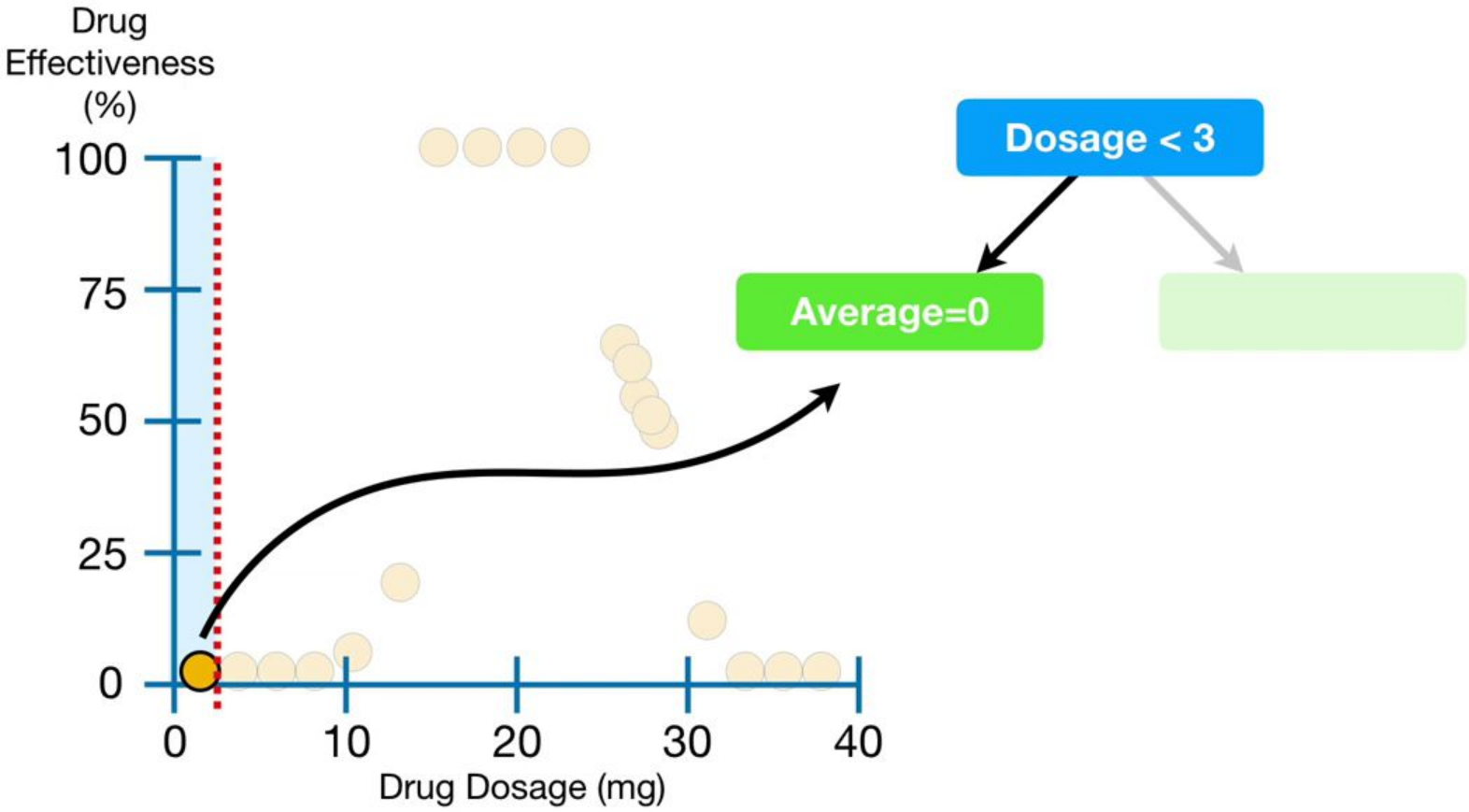
The point on the far left is the only one with **Dosage < 3**…



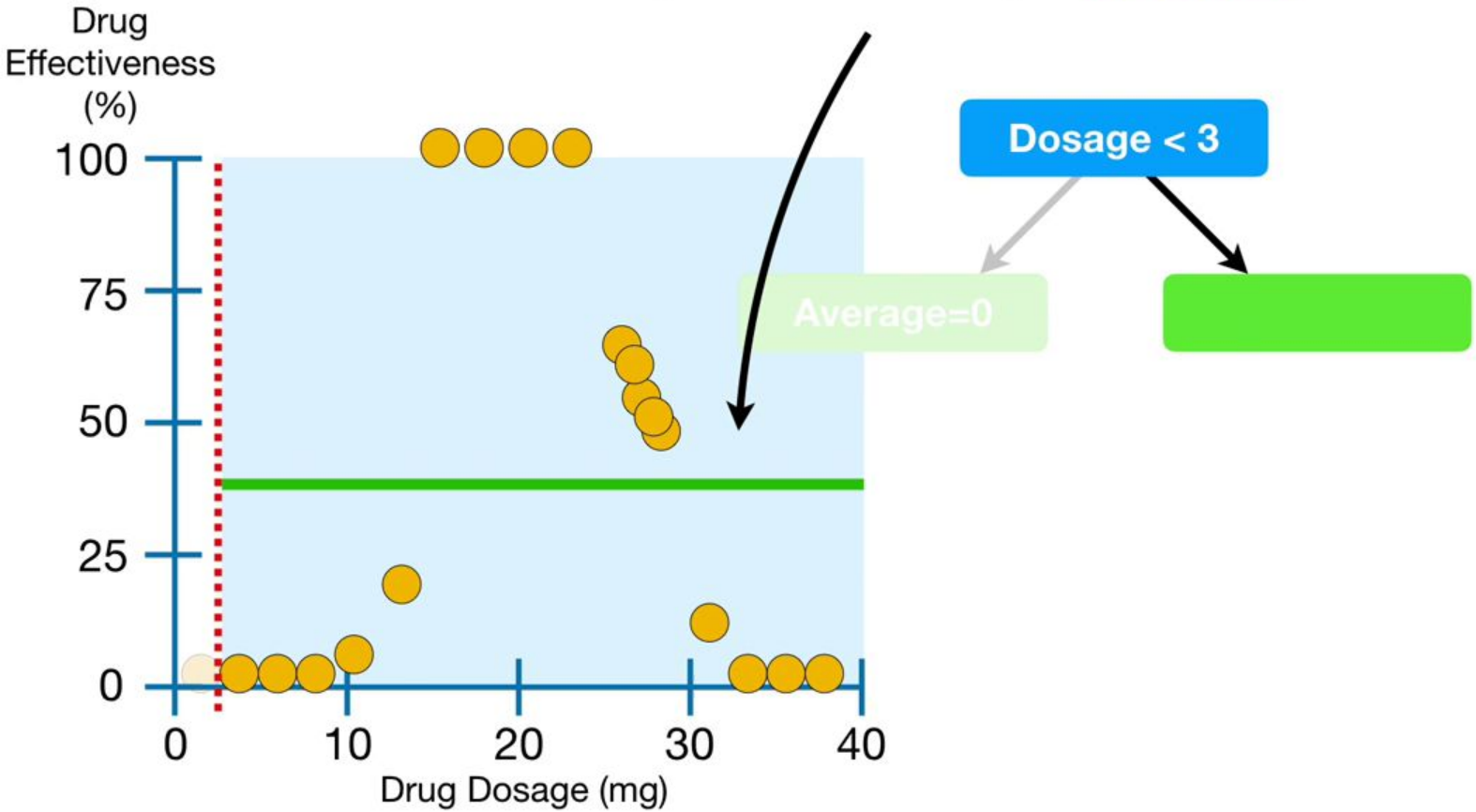
…and the average **Drug Effectiveness** for that one point is **0**…



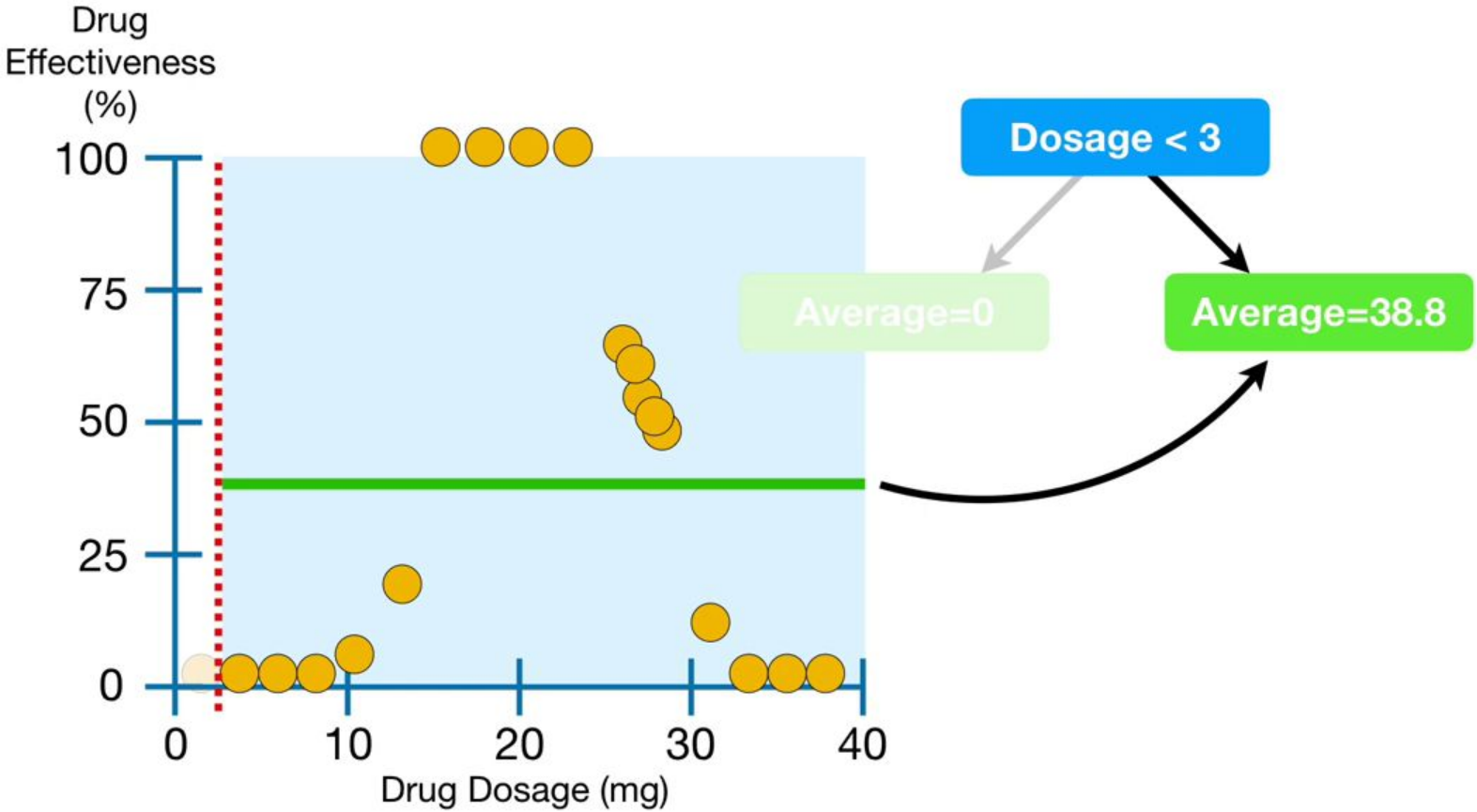
…so we put **0** in the leaf on the left side, for when **Dosage < 3**.



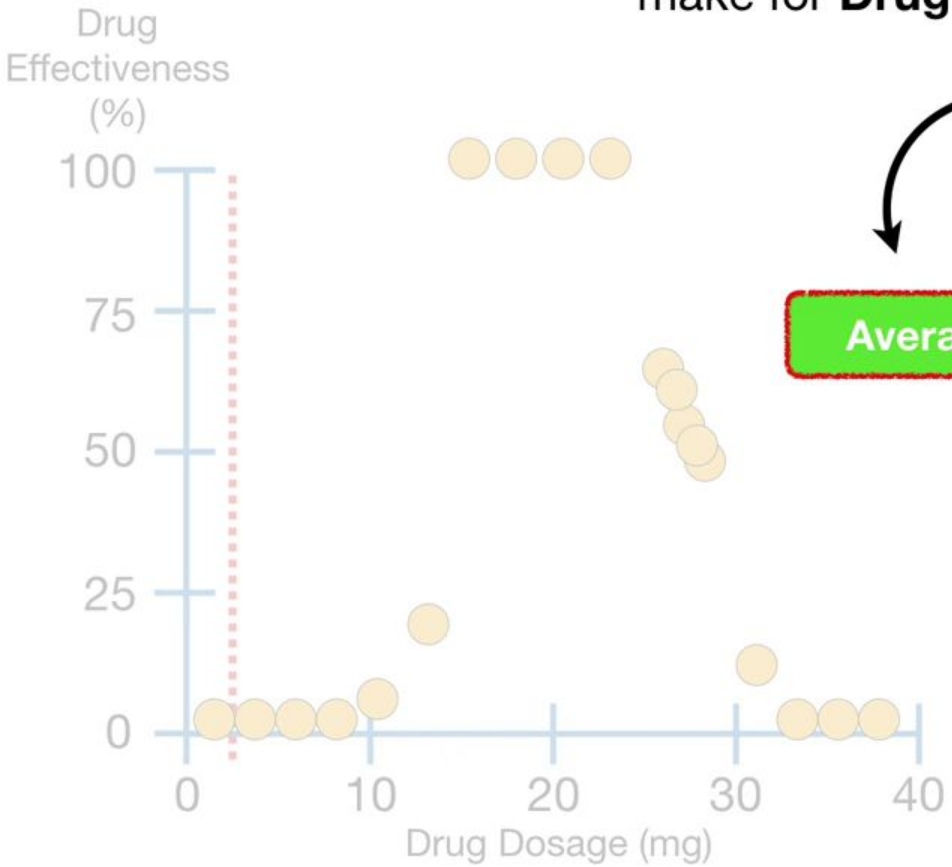
…and the average **Drug Effectiveness** for all of the points with **Dosages >= 3** is **38.8**, (the **green line**)…

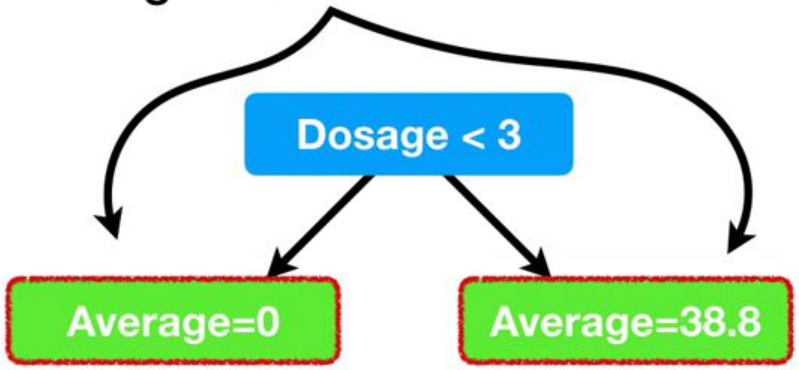


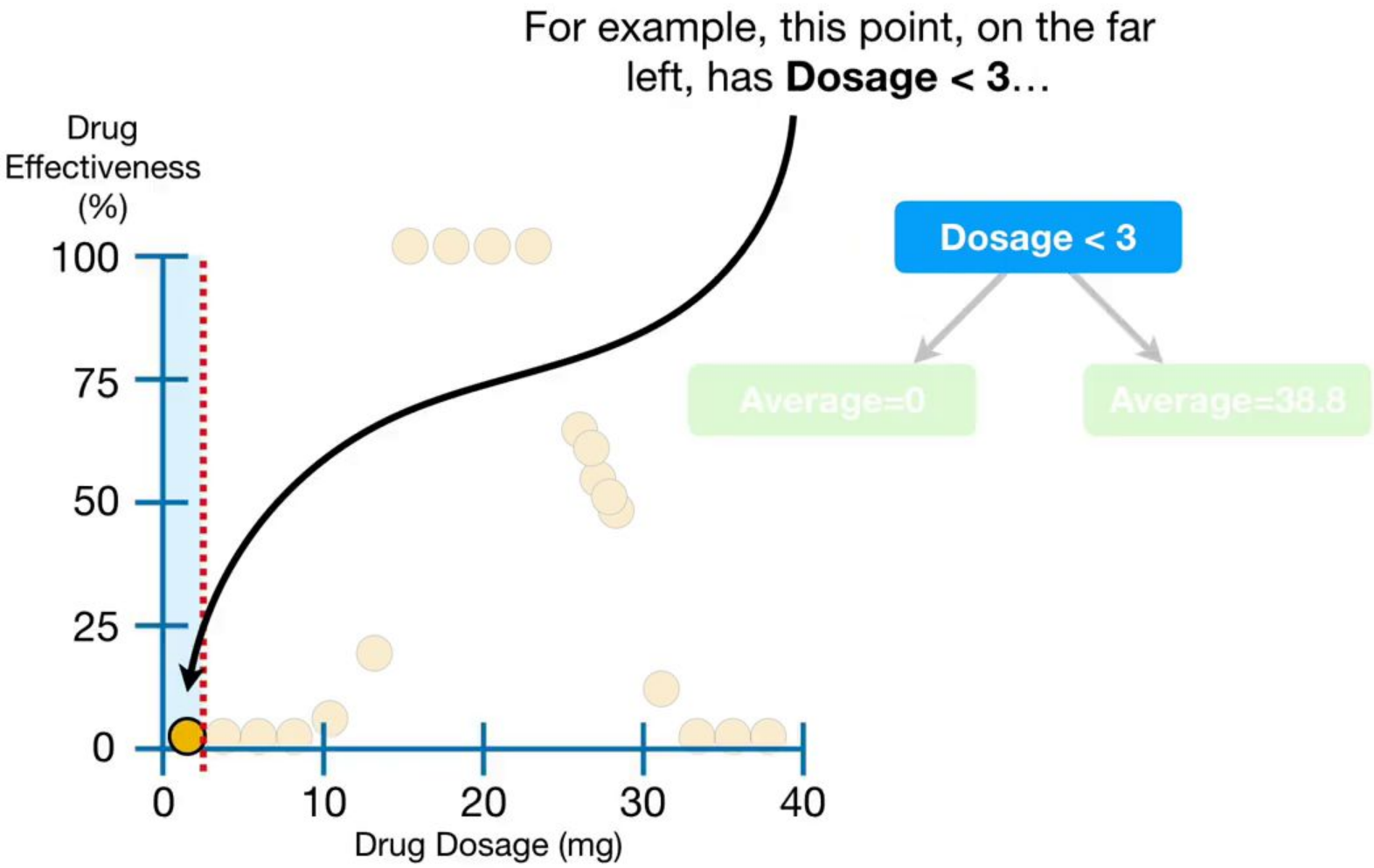
…so we put **38.8** in the leaf on the right side, for when the **Dosage >= 3**.



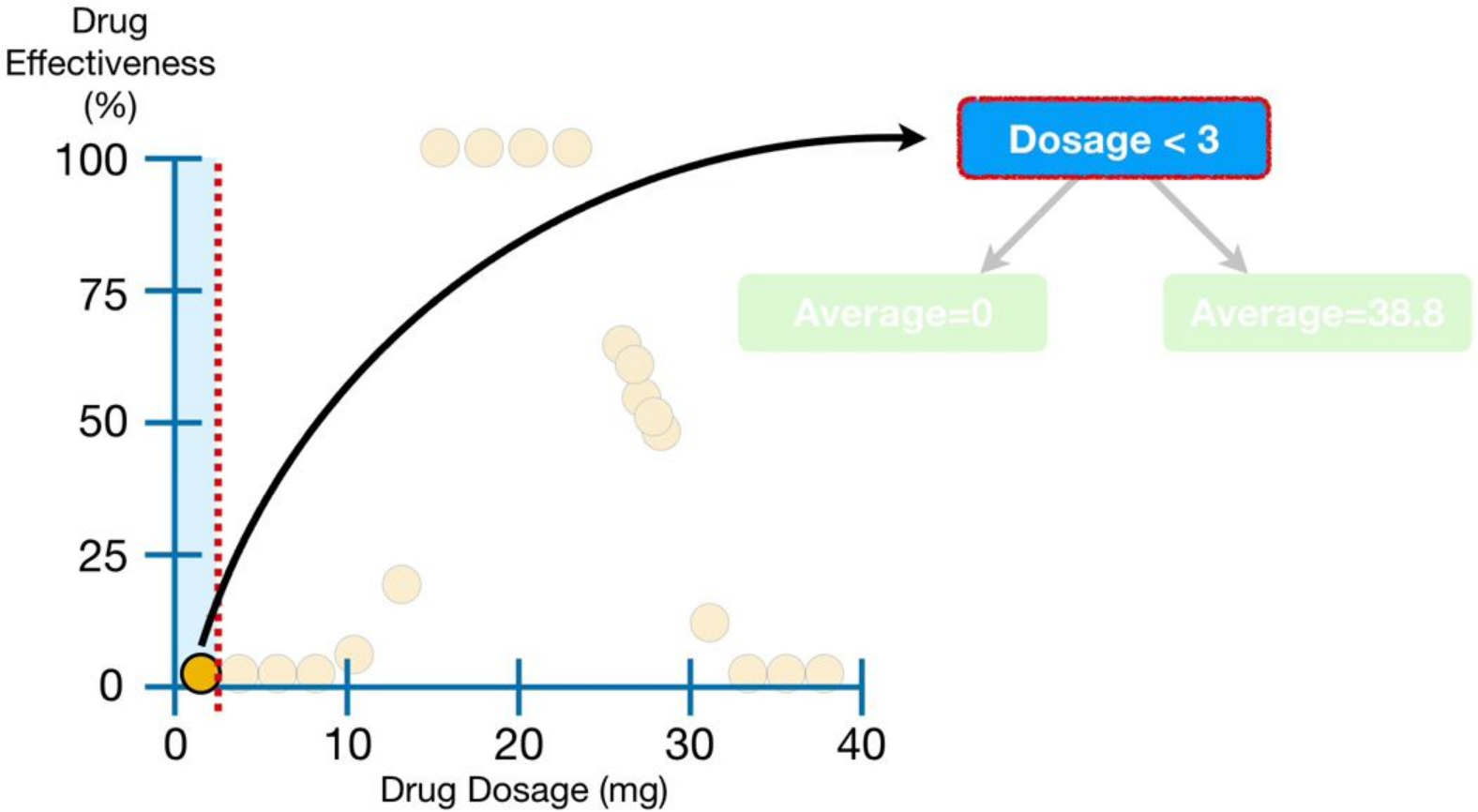
The values in each leaf are the predictions that this simple tree will make for **Drug Effectiveness**.



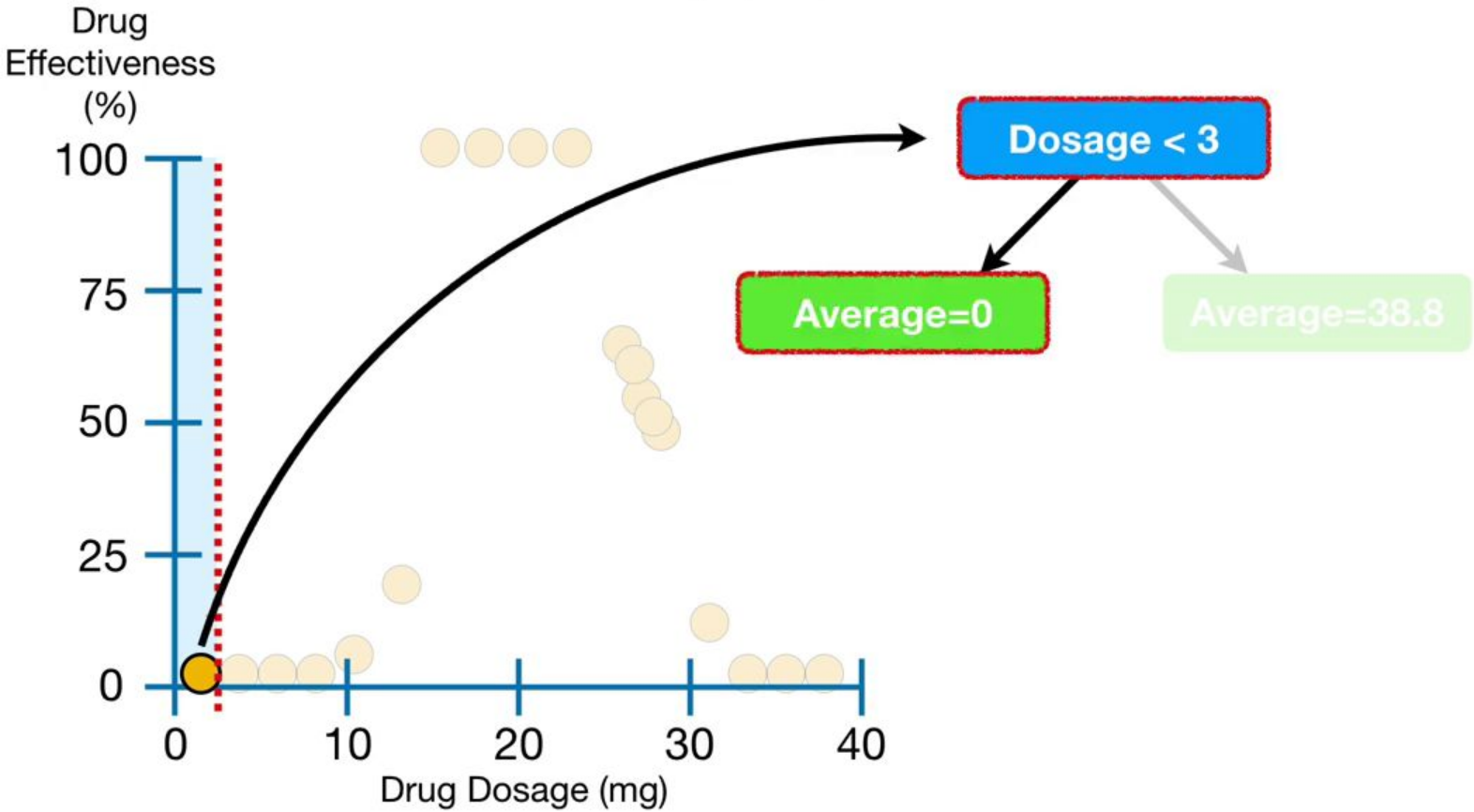



For example, this point, on the far left, has **Dosage < 3**…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 3
Average=0
Average=38.8

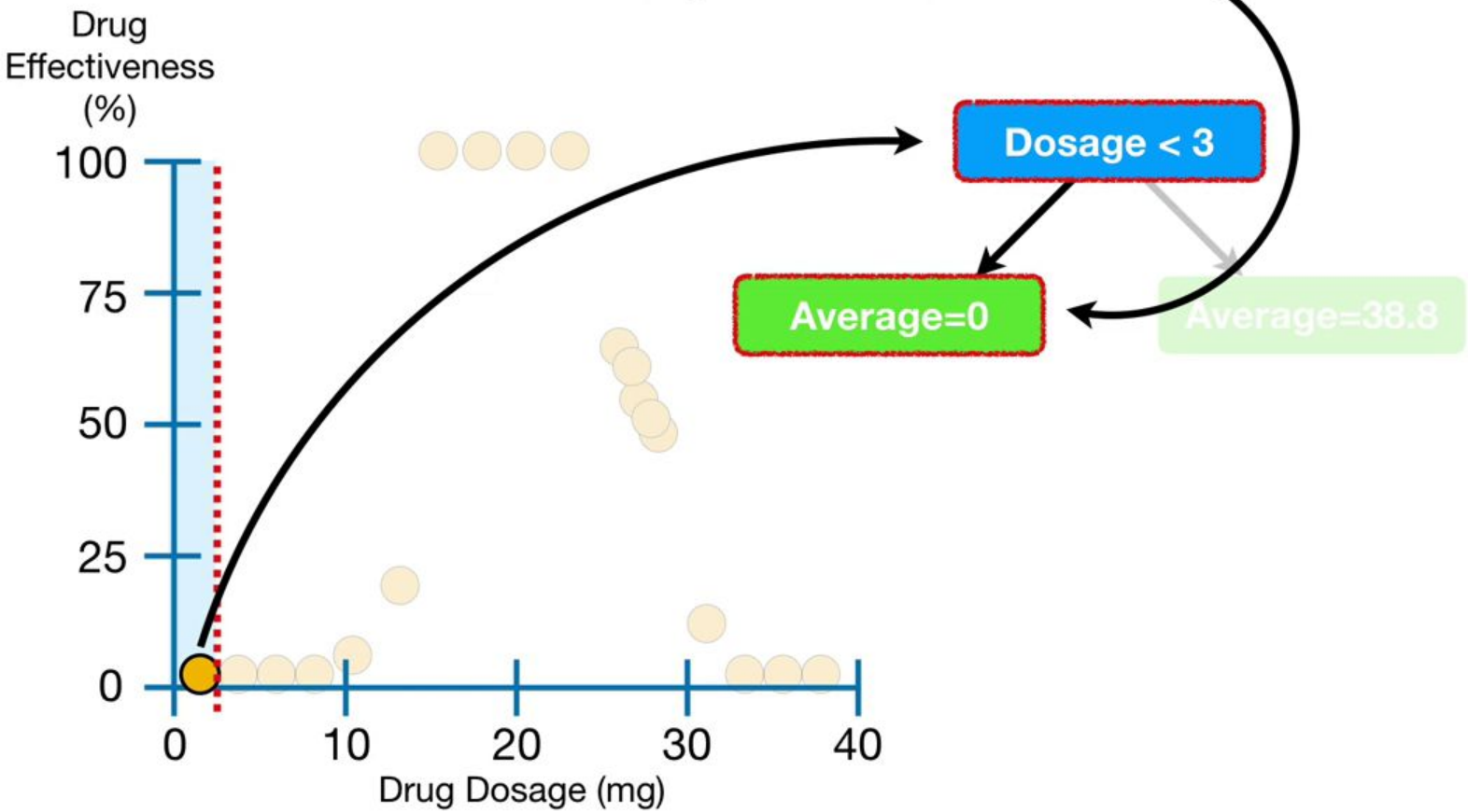
…and the tree predicts that the
Drug Effectiveness will be **0**.



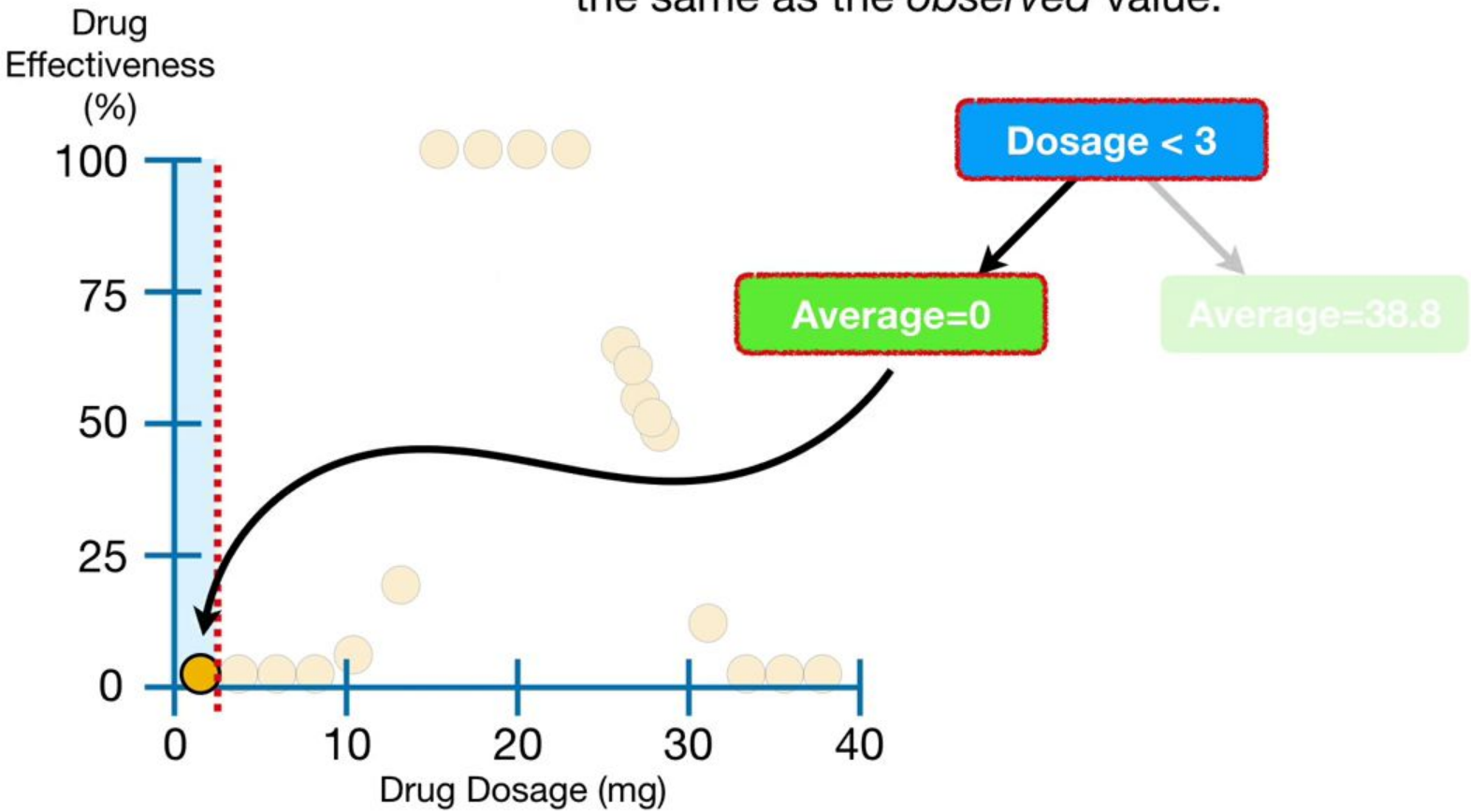
…and the tree predicts that the
Drug Effectiveness will be **0**.



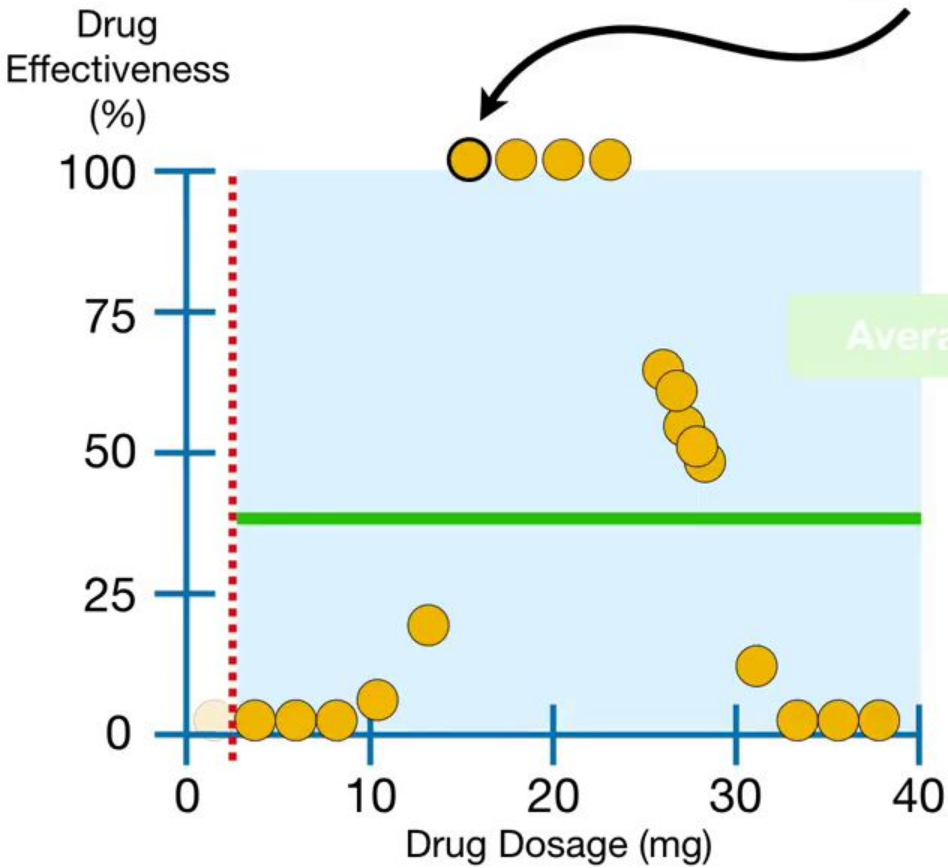
…and the tree predicts that the **Drug Effectiveness** will be **0**.

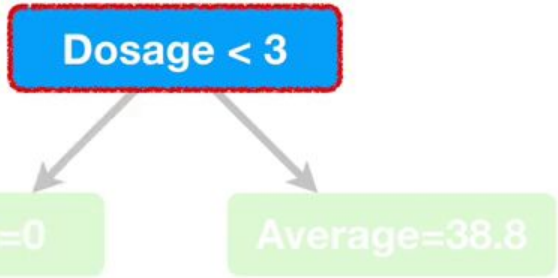


The *prediction* for this point, **Drug Effectiveness = 0**, is pretty good since it is the same as the *observed* value.

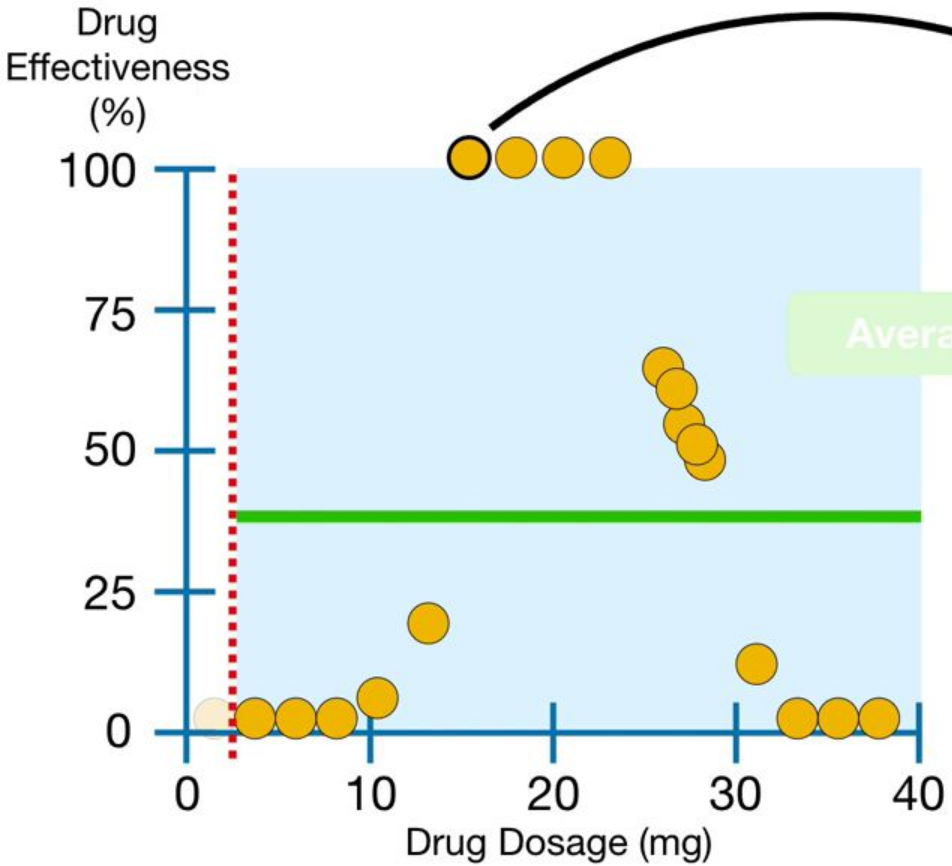


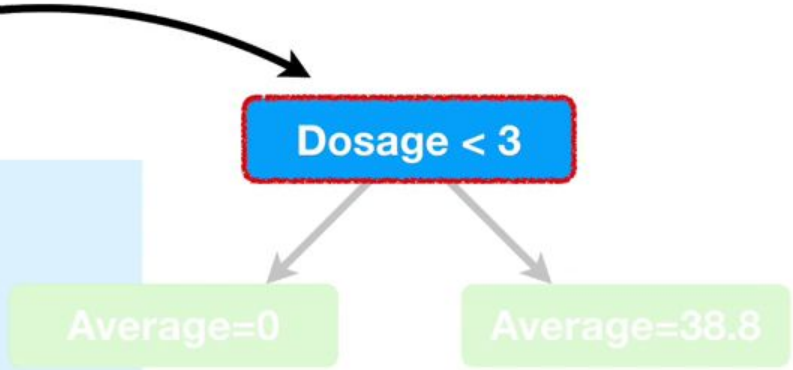
In contrast, for this point, which has **Dosage > 3**…

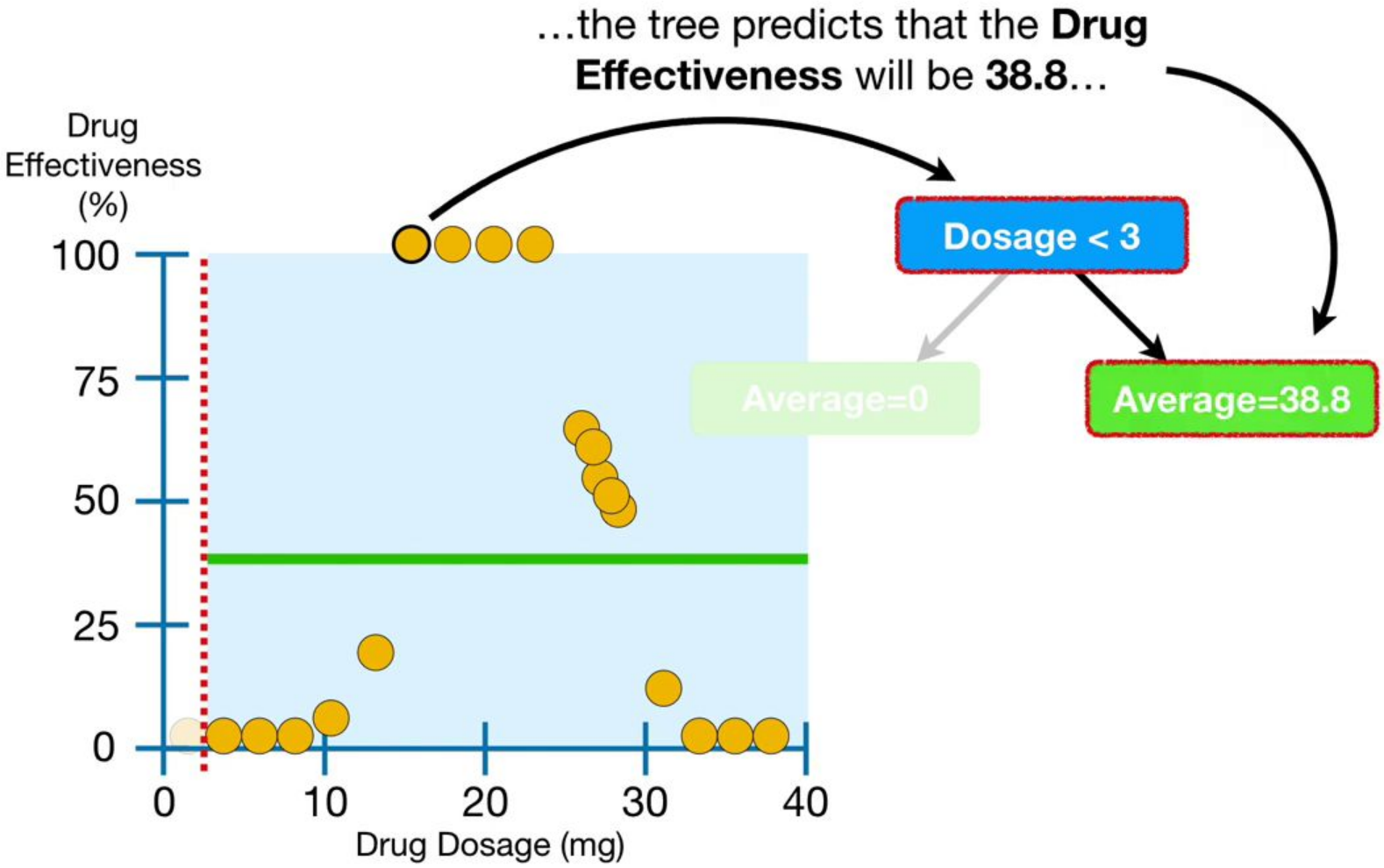




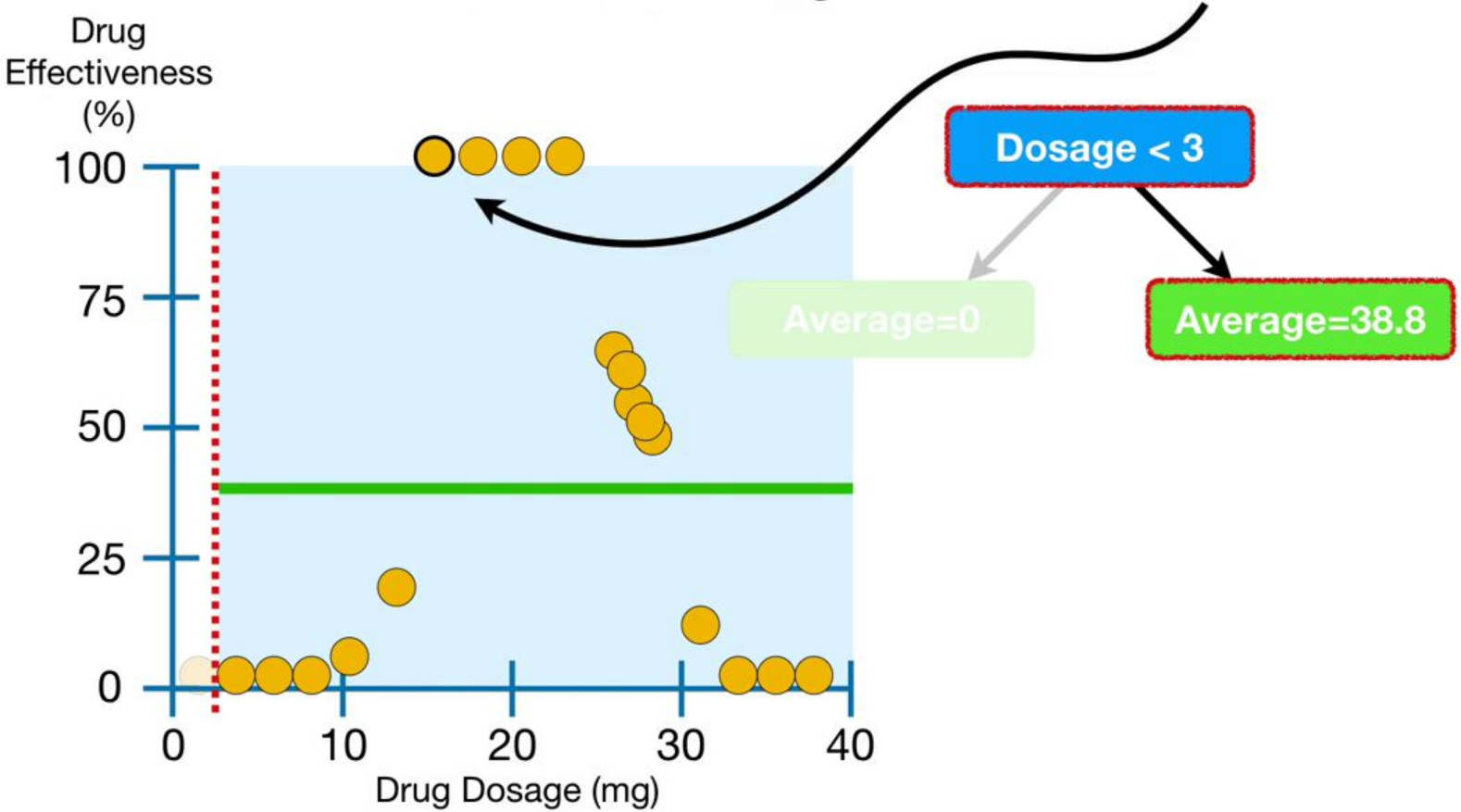
…the tree predicts that the **Drug Effectiveness** will be **38.8**…



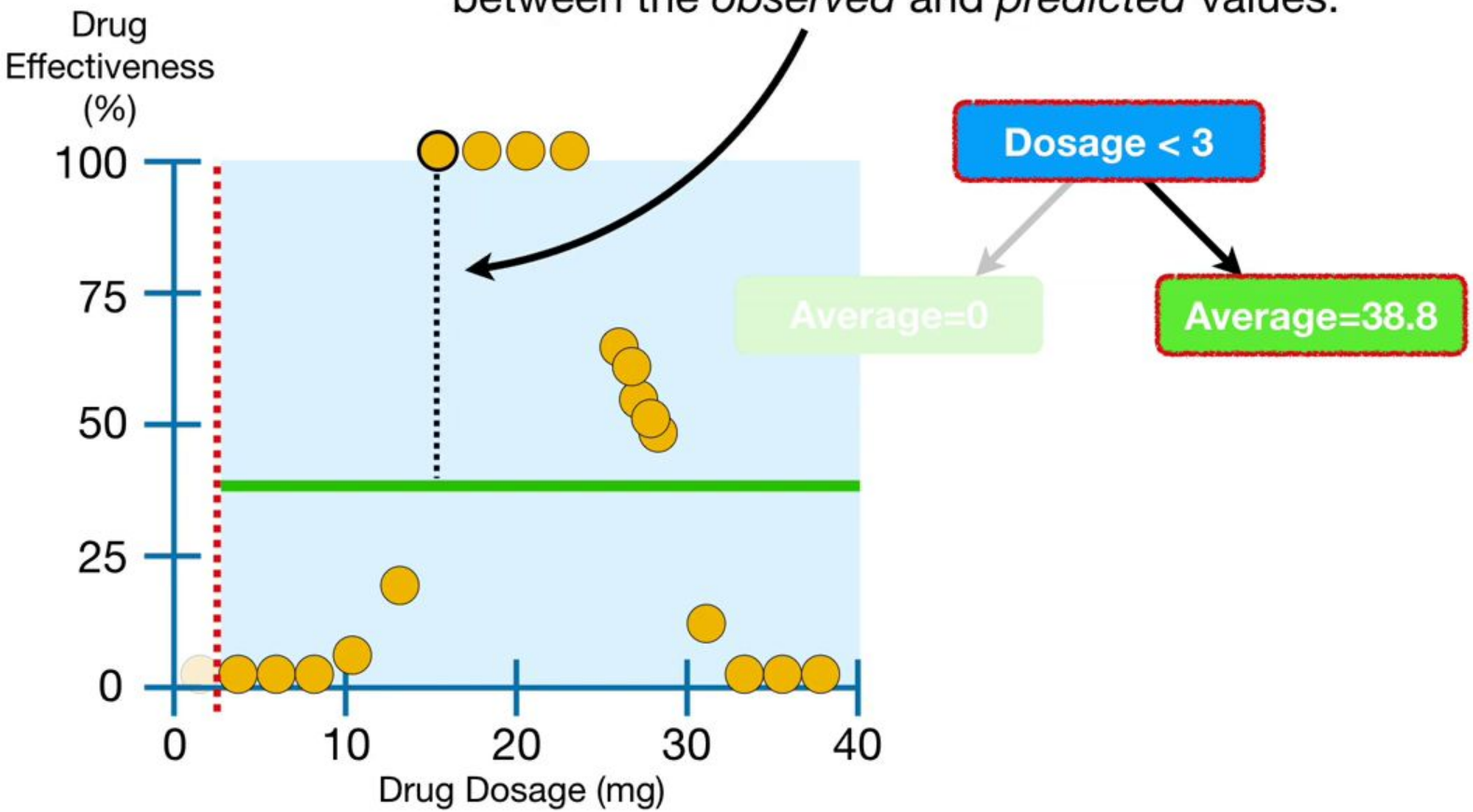



…the tree predicts that the **Drug Effectiveness** will be **38.8**…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 3
Average=0
Average=38.8

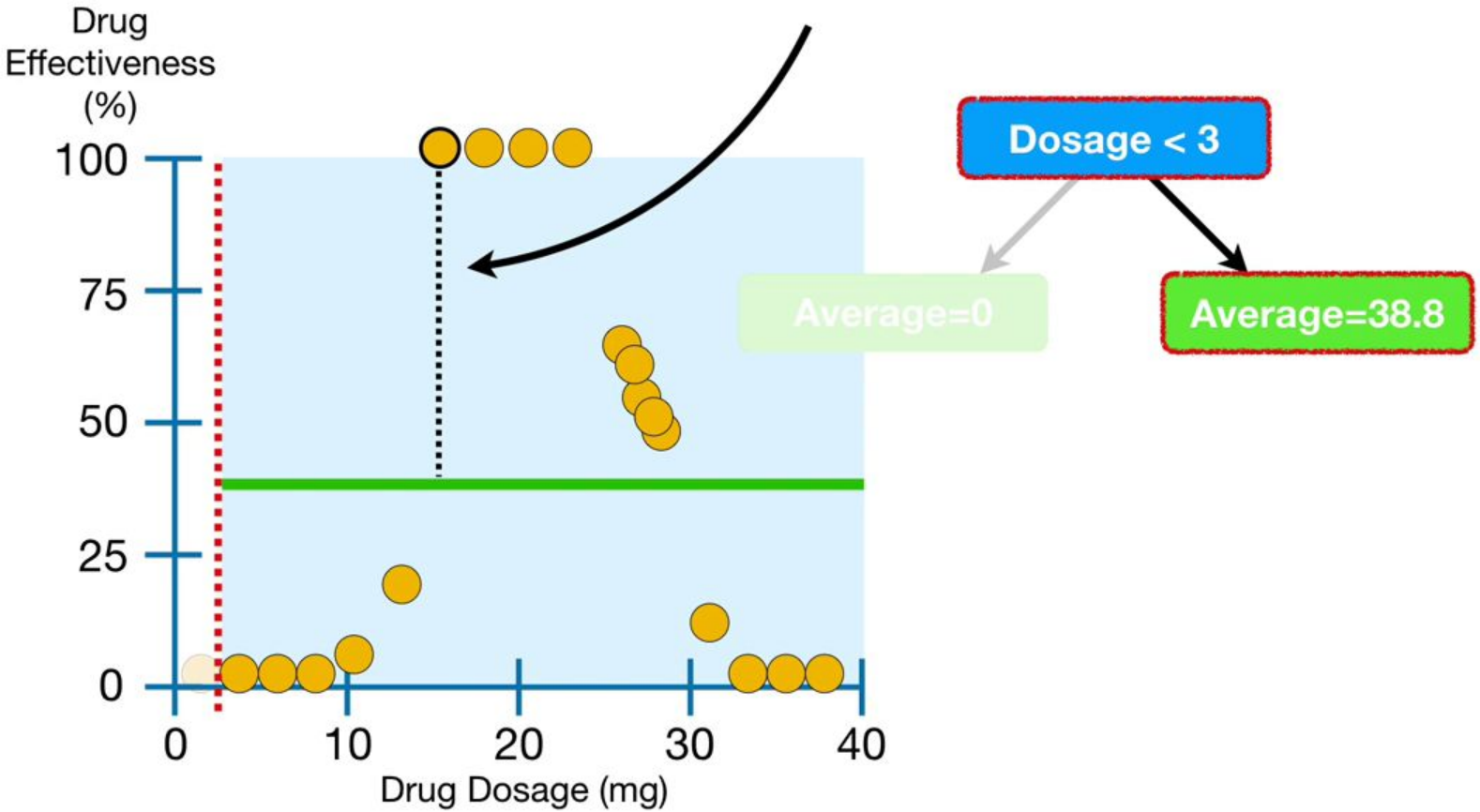
…and that *prediction* is not very good, since the *observed* **Drug Effectiveness** is **100%**.



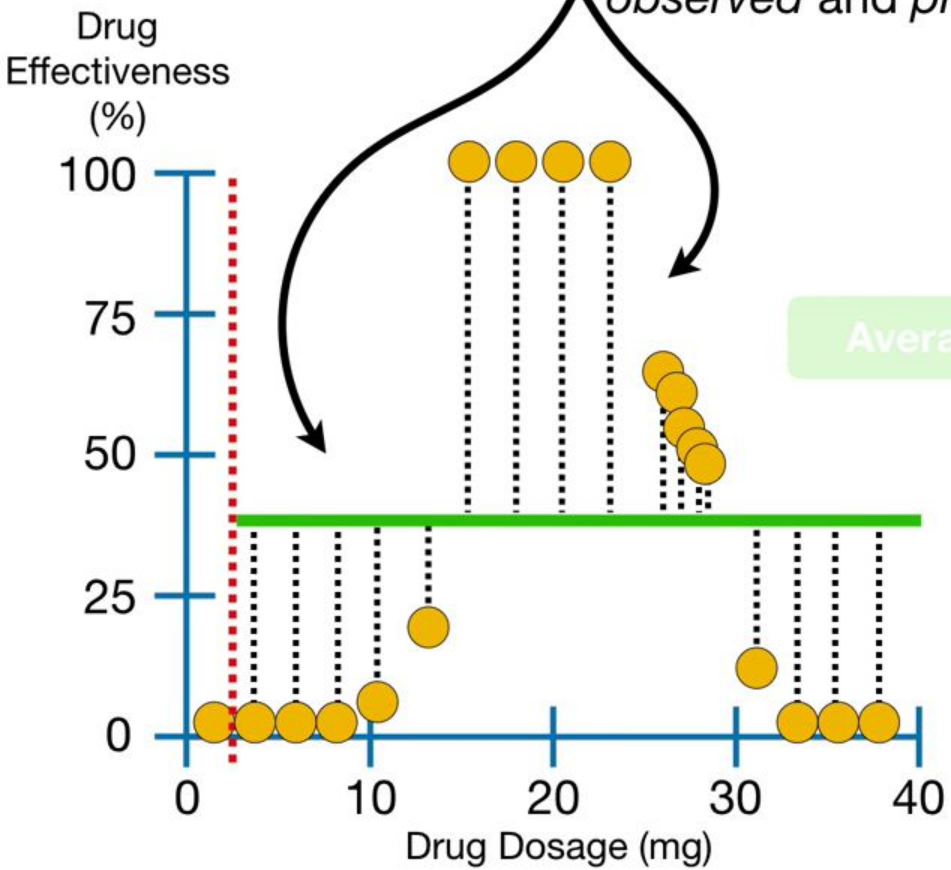
NOTE: We can visualize how bad the prediction is by drawing a dotted line between the *observed* and *predicted* values.

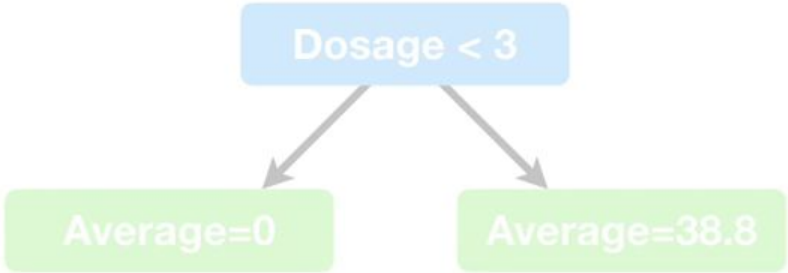


In other words, this dotted line is a **residual**.

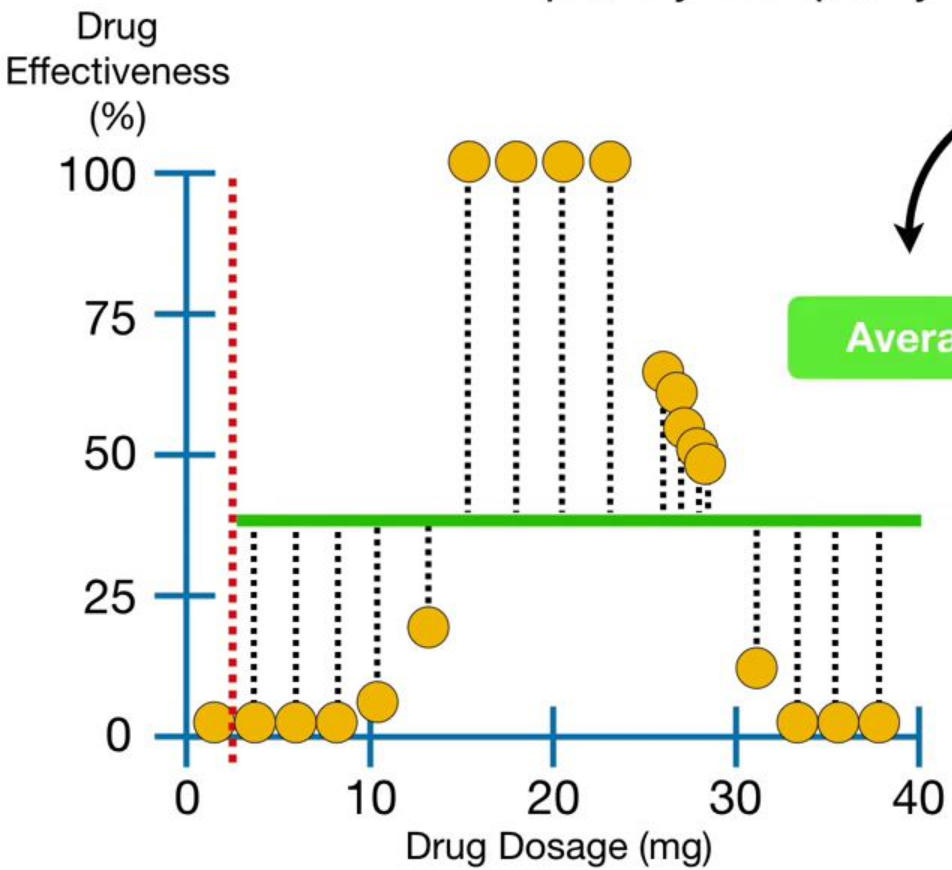


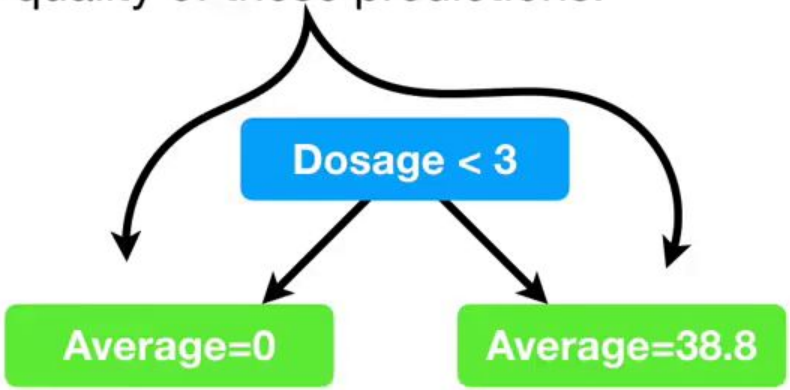
For each point in the data, we can draw its **residual**, the difference between the *observed* and *predicted* values…



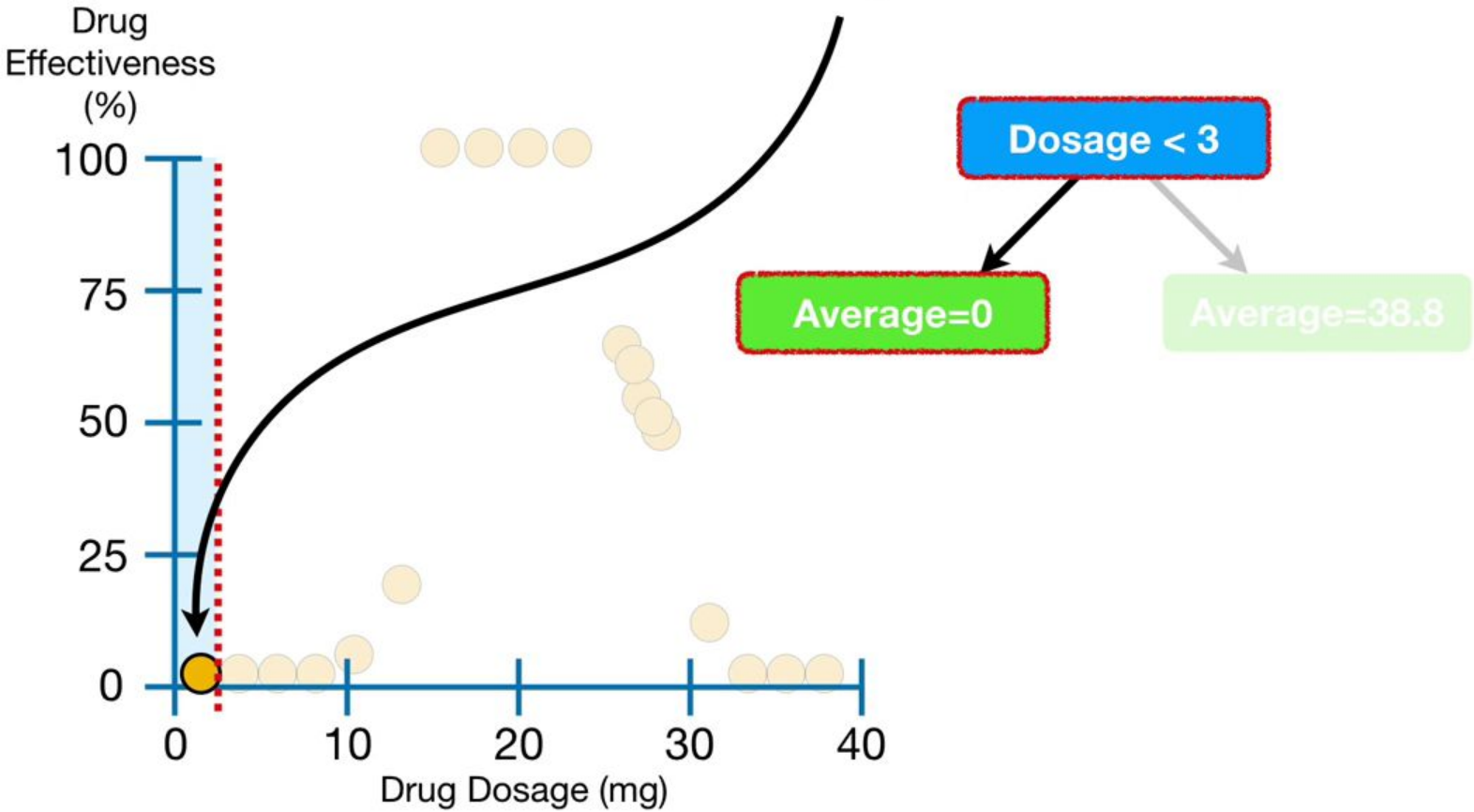


…and we can use the **residuals** to *quantify* the quality of these predictions.

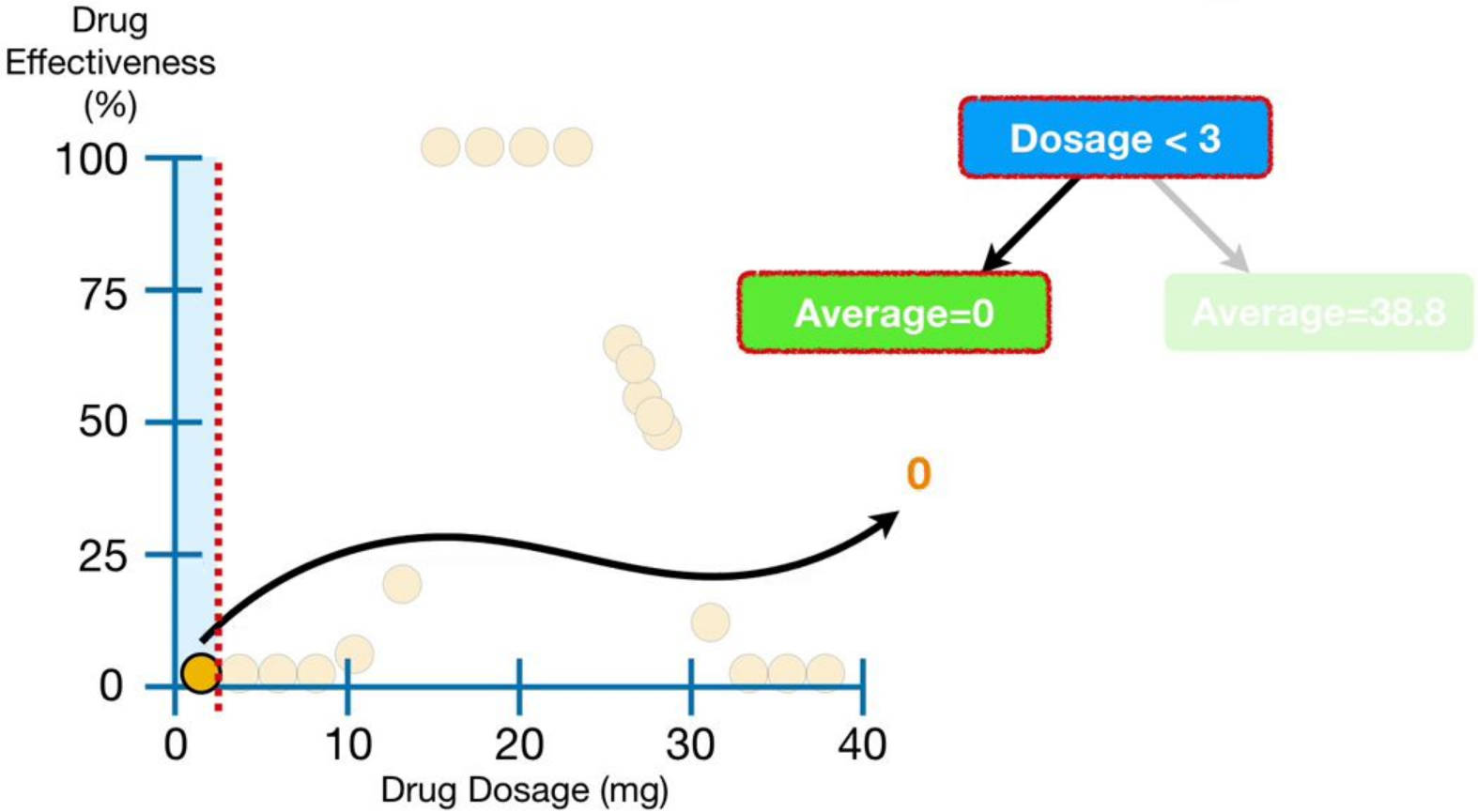




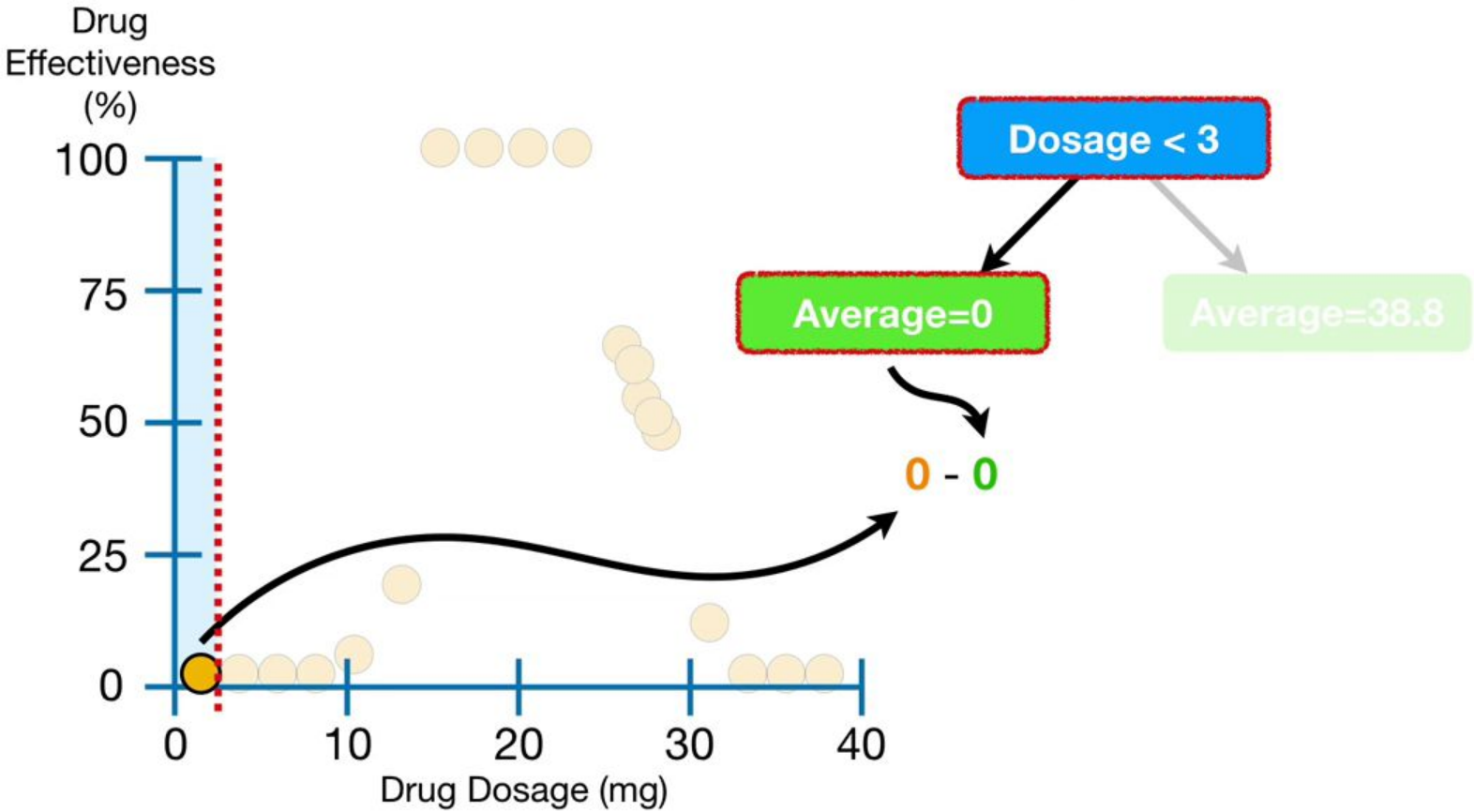
Starting with the only point with
Dosage < 3…



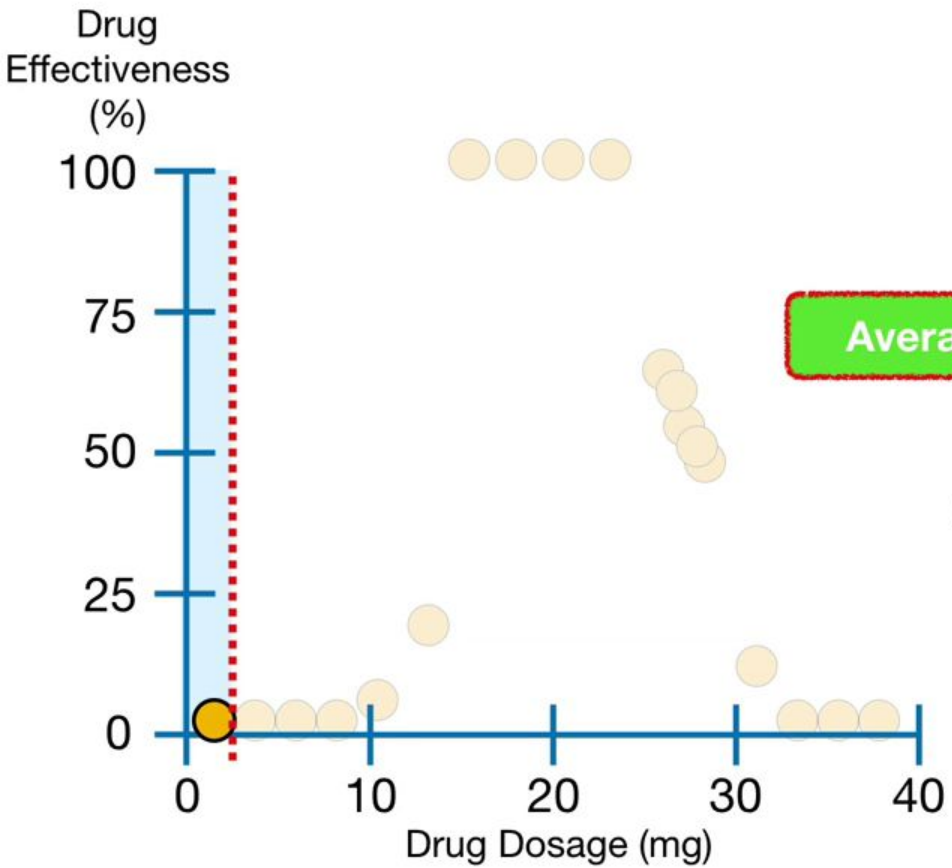
…we calculate the difference between its *observed* **Drug Effectiveness**, **0**,…



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



…and then square the difference.



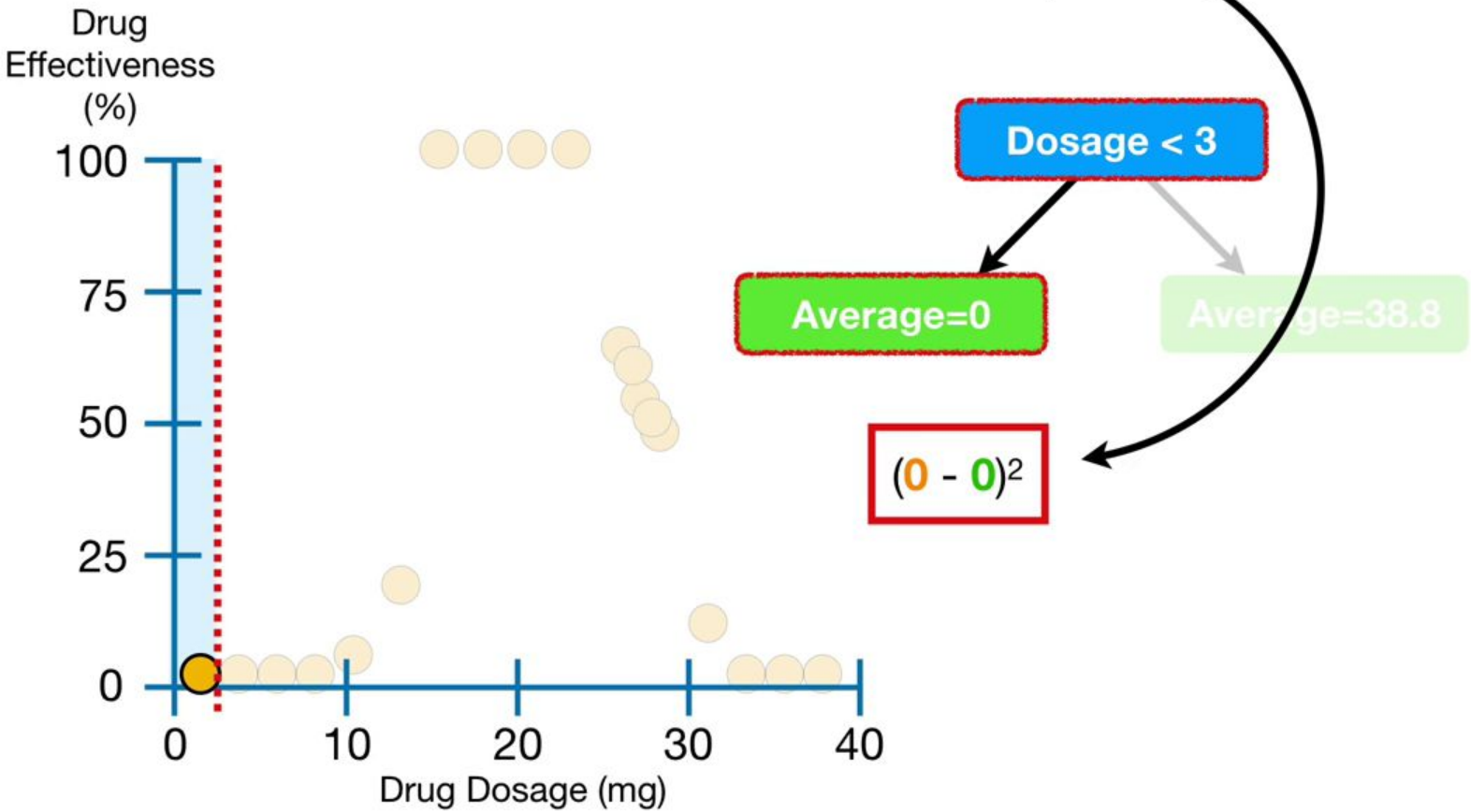
Dosage < 3

Average=0

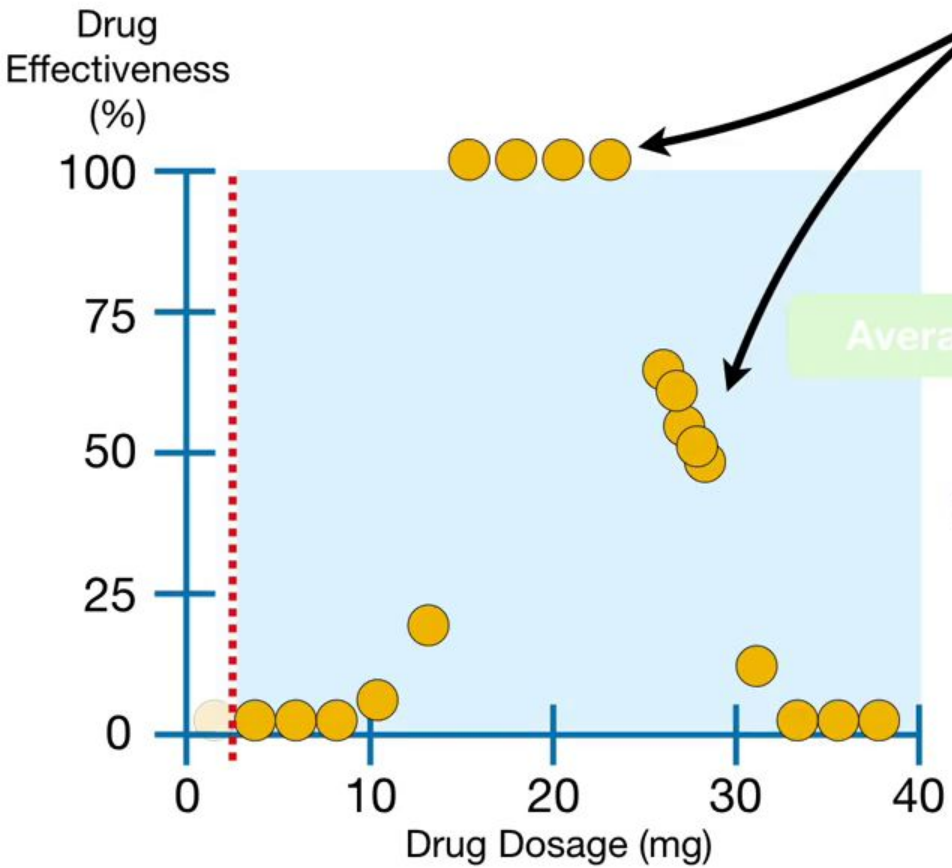
Average=38.8

$(0 - 0)^2$

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**.

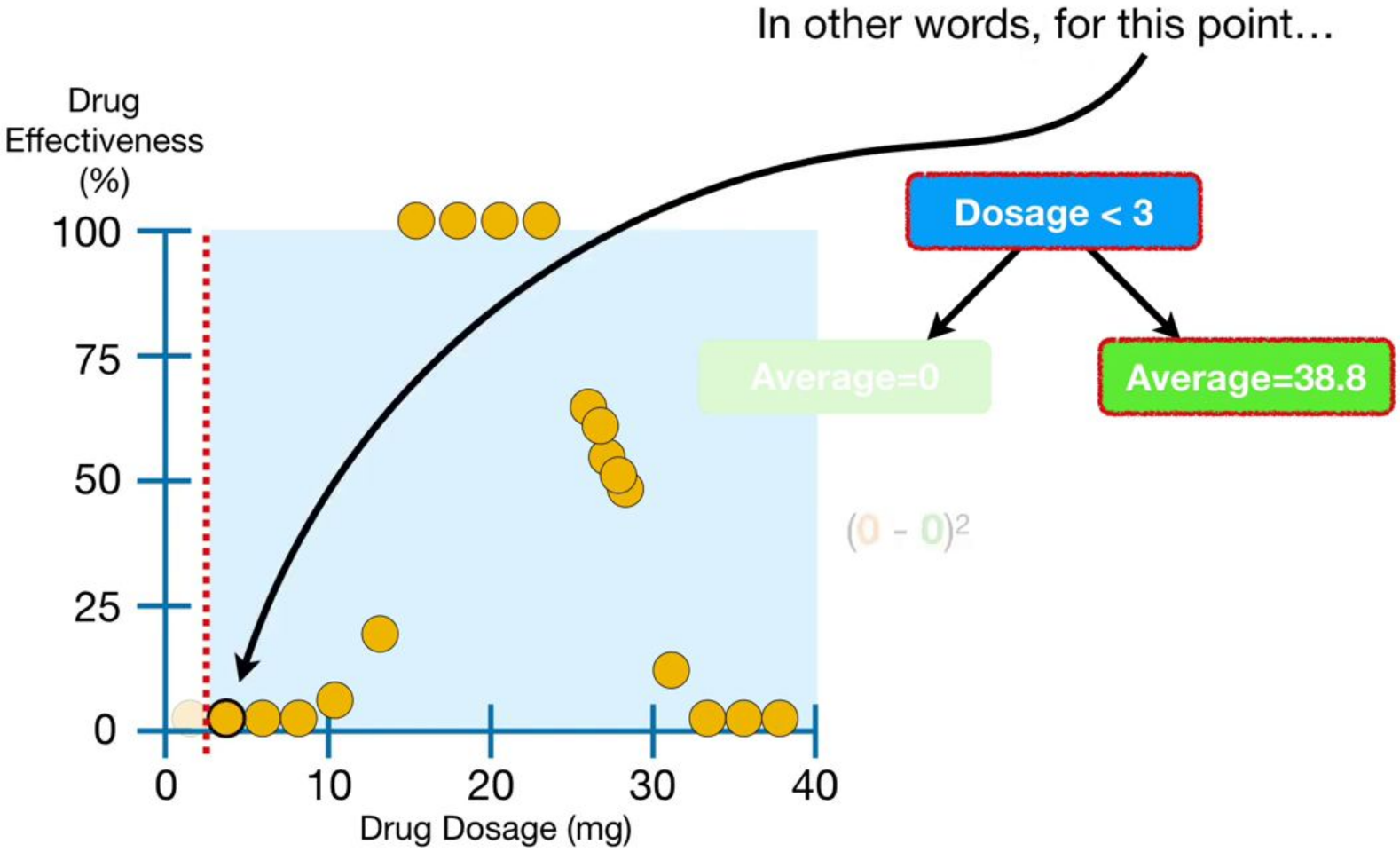


Dosage < 3

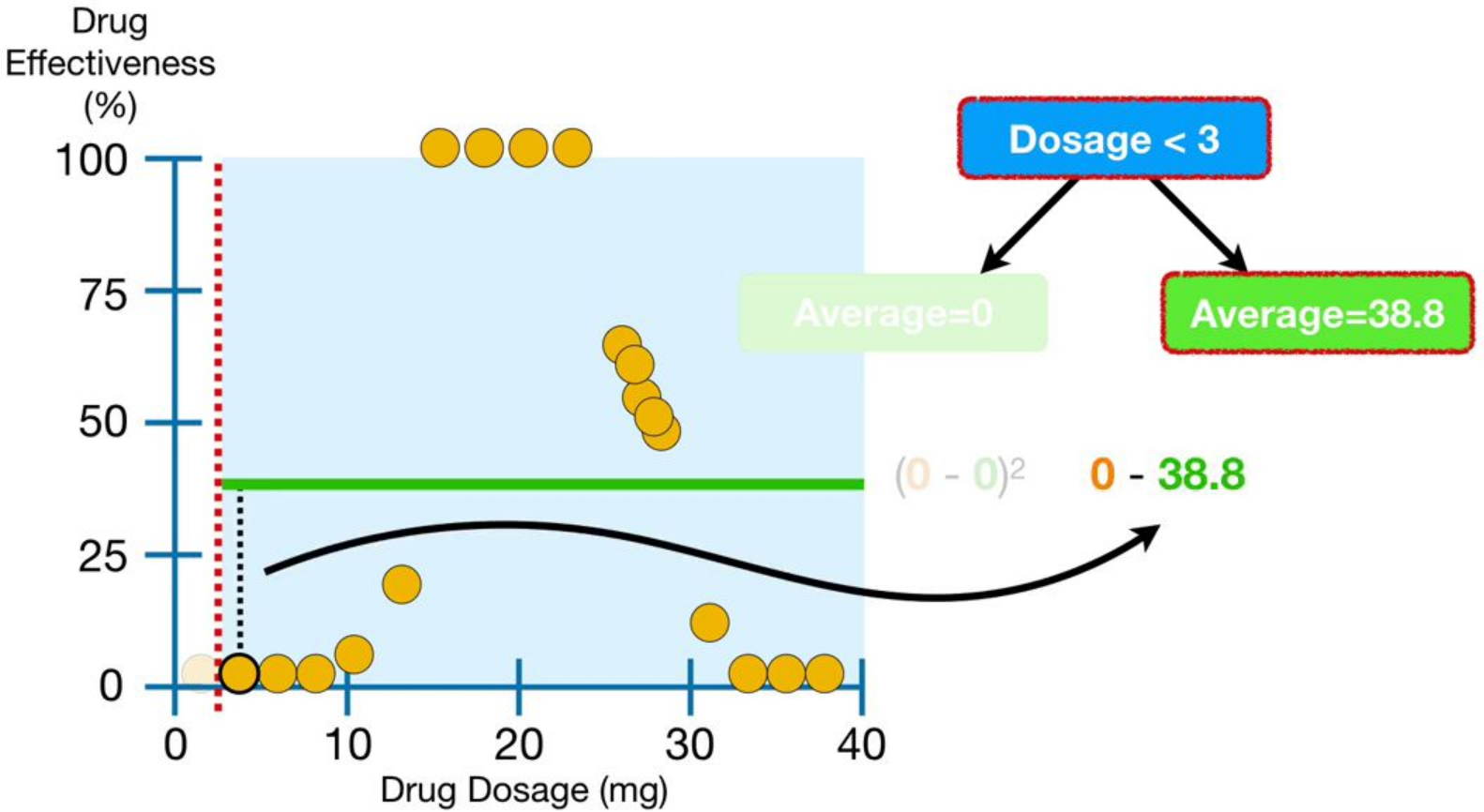
Average=0

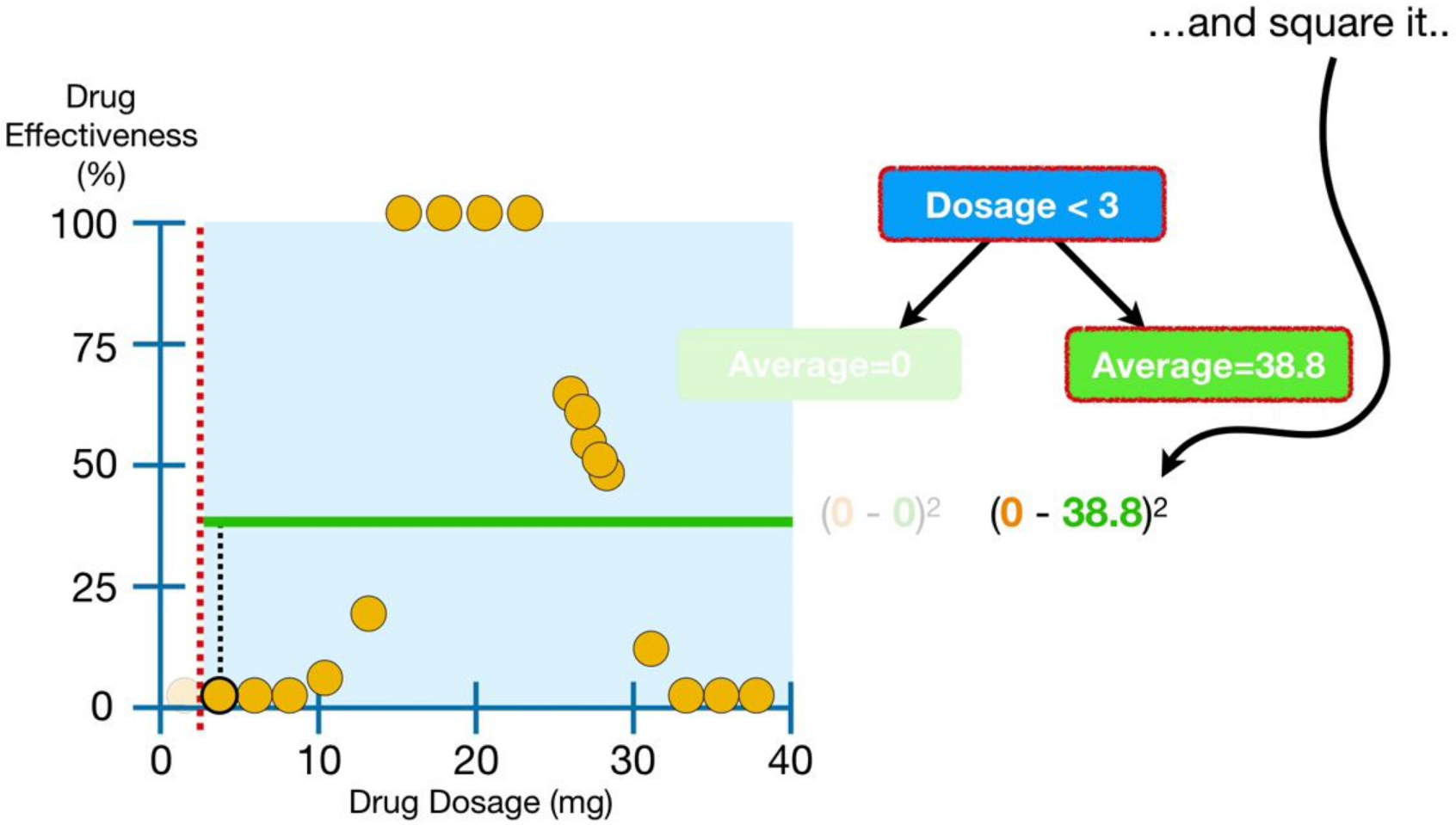
Average=38.8

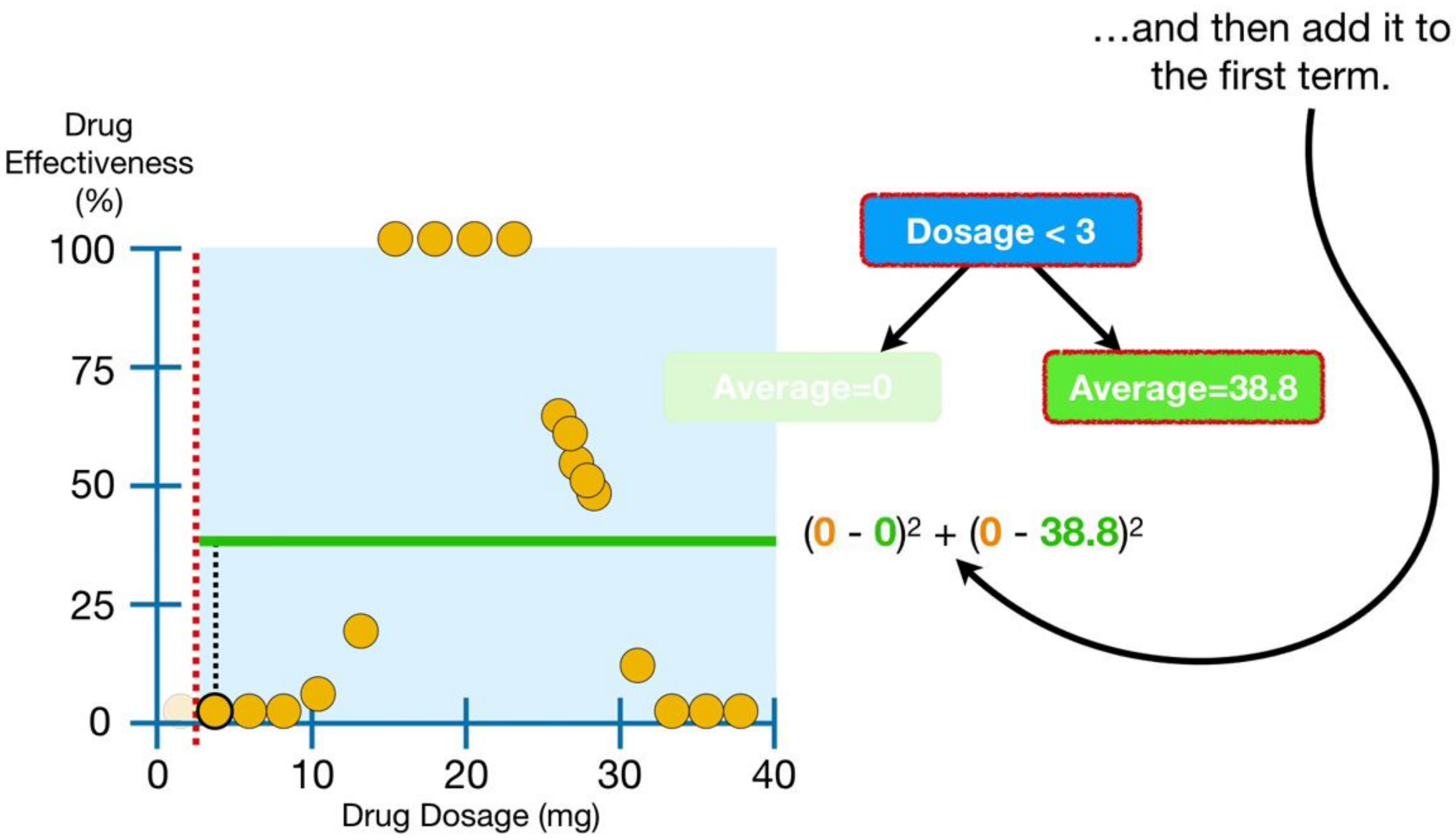
$(0 - 0)^2$


In other words, for this point…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 3
Average=0
Average=38.8
(0 - 0)²

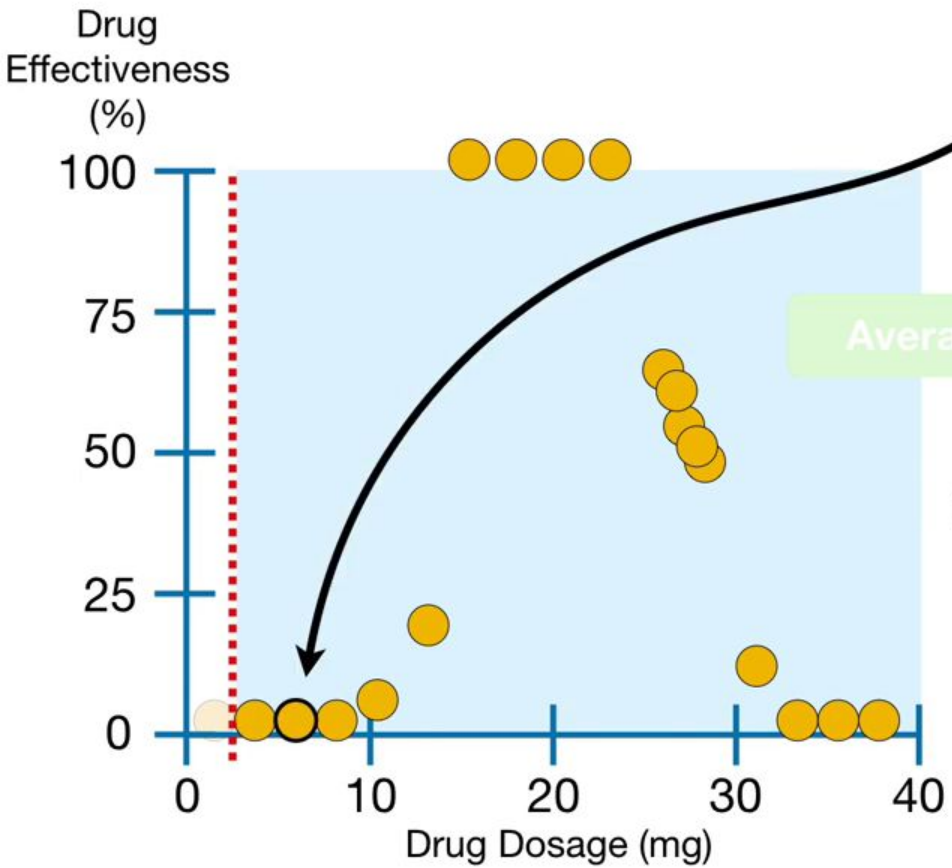
…we calculate the the difference between the *observed* and *predicted* values…




…and square it..
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 3
Average=0
Average=38.8
(0 - 0)²
(0 - 38.8)²


…and then add it to the first term.
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 3
Average=0
Average=38.8
(0 - 0)² + (0 - 38.8)²

Then we do the same thing for the next point…



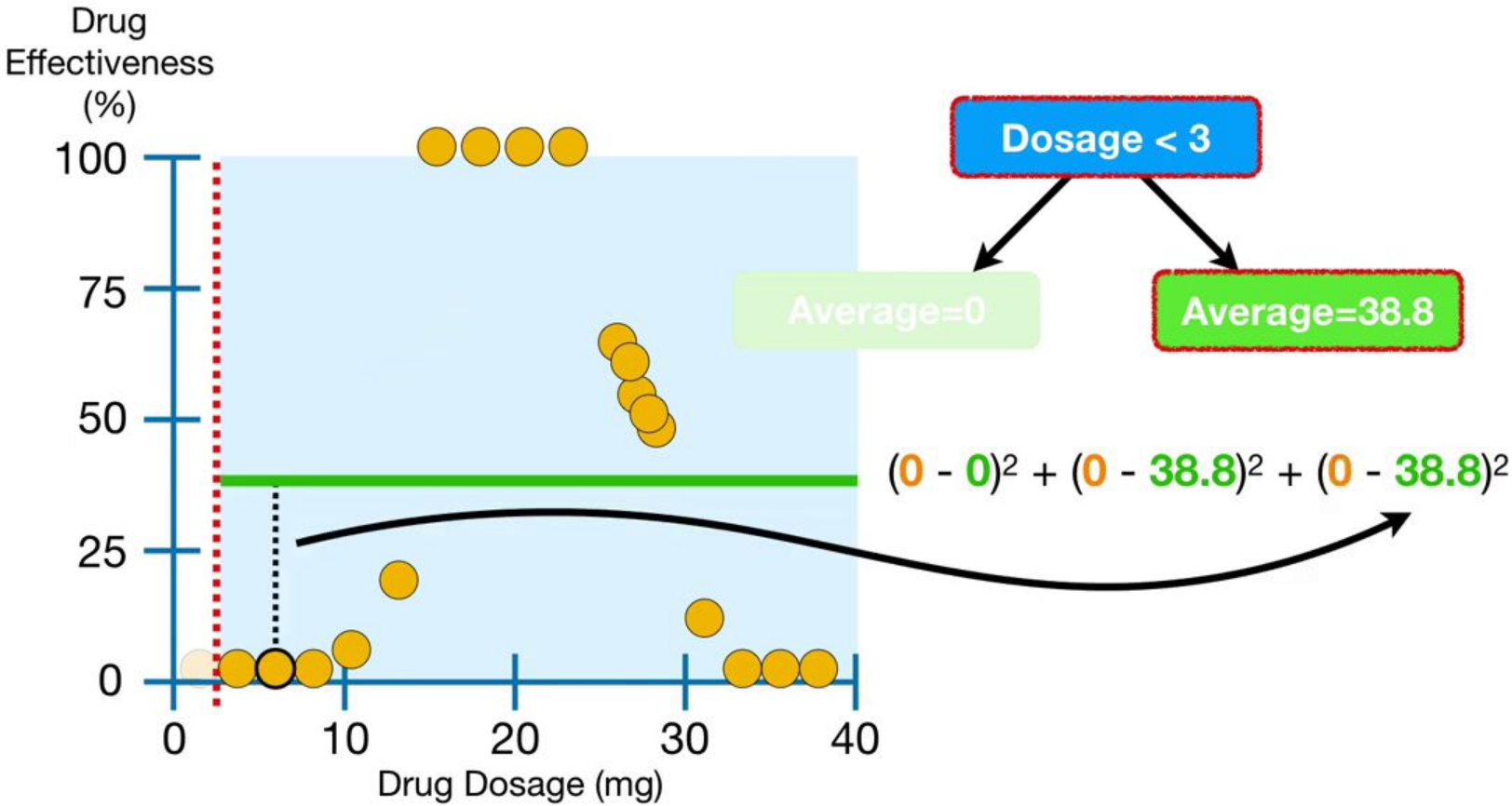
Dosage < 3

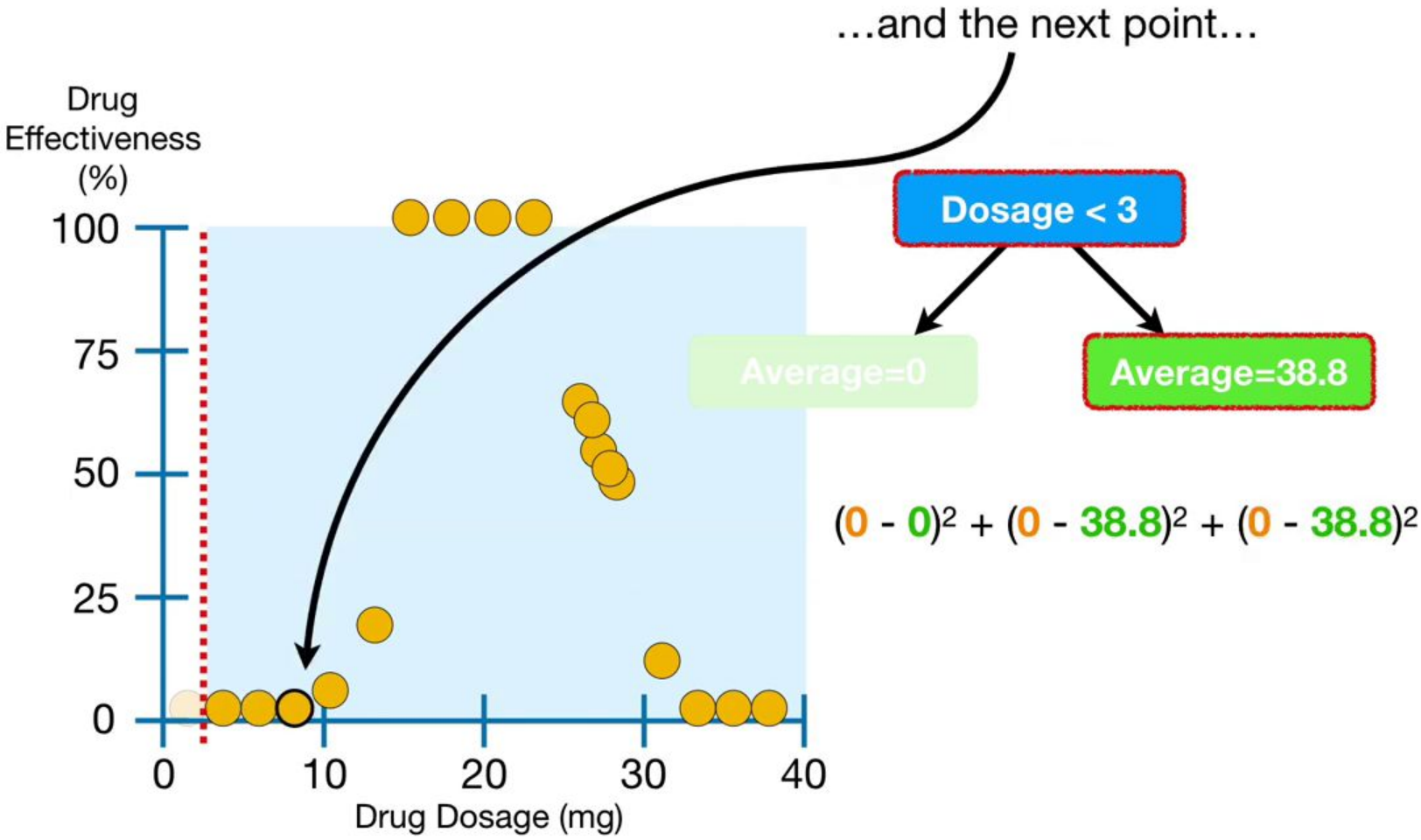
Average=0

Average=38.8

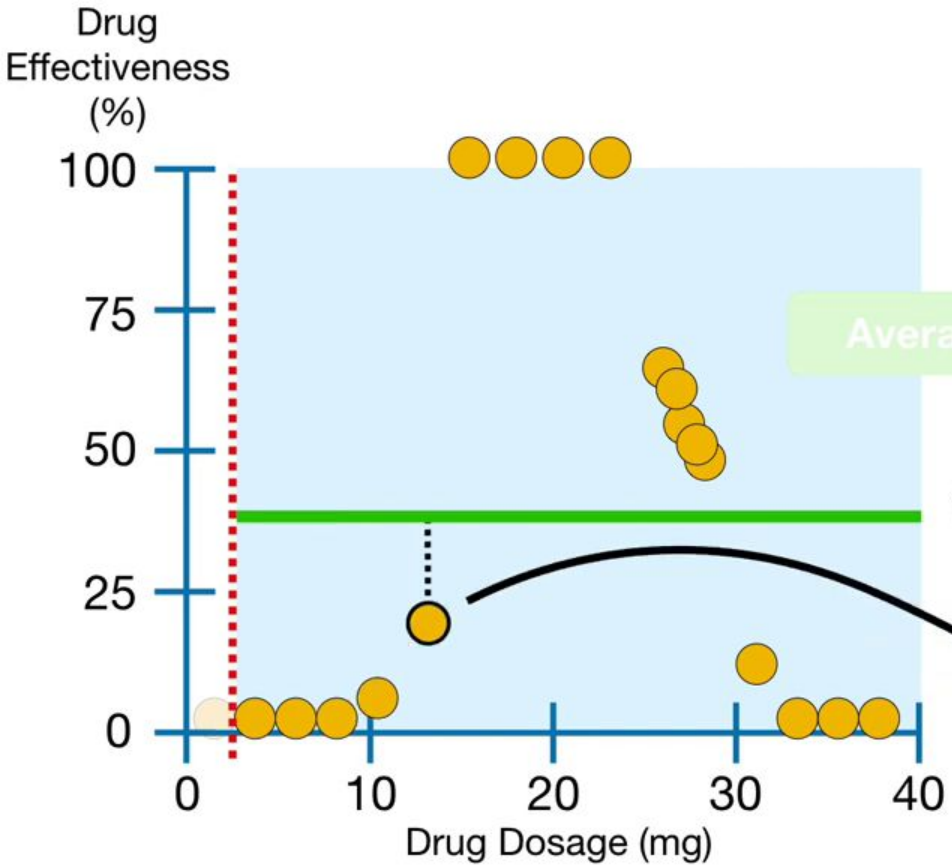
$(0 - 0)^2 + (0 - 38.8)^2$

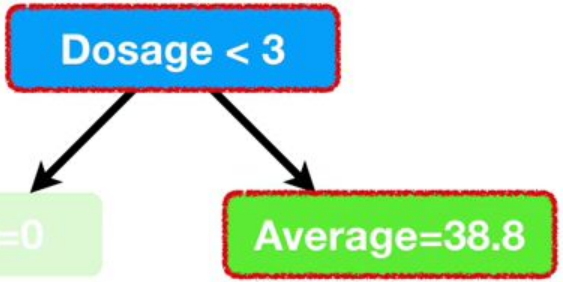
Then we do the same thing for the next point…




…and the next point…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 3
Average=0
Average=38.8
(0 - 0)² + (0 - 38.8)² + (0 - 38.8)²

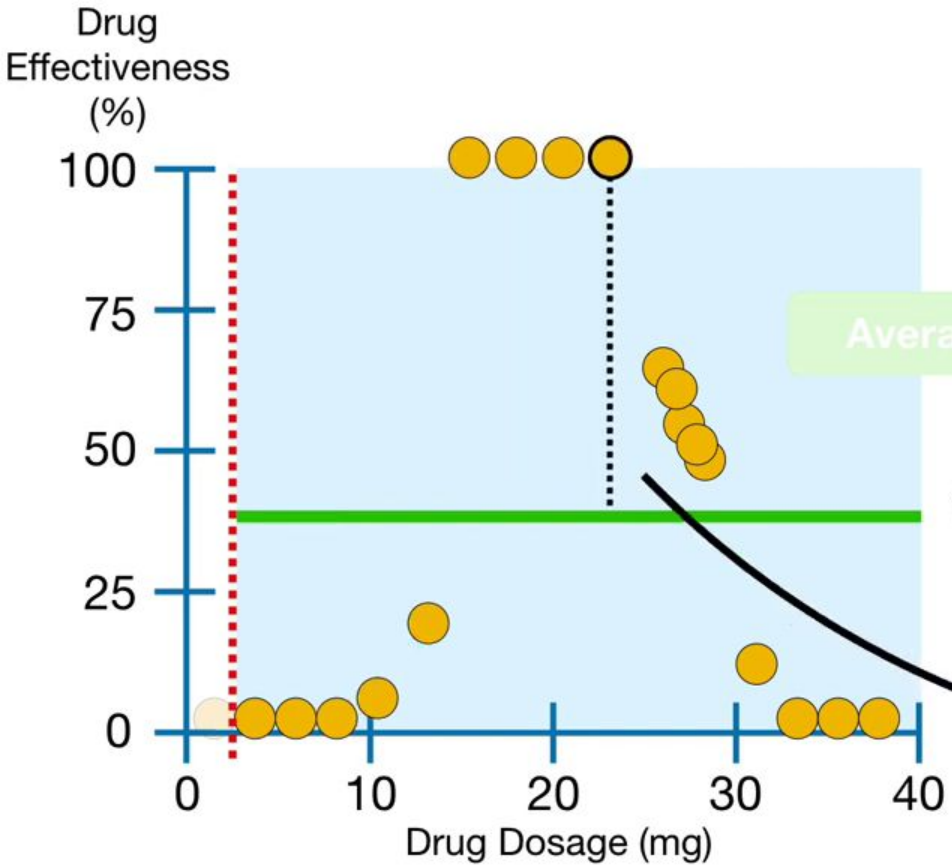
…and the rest of the points…

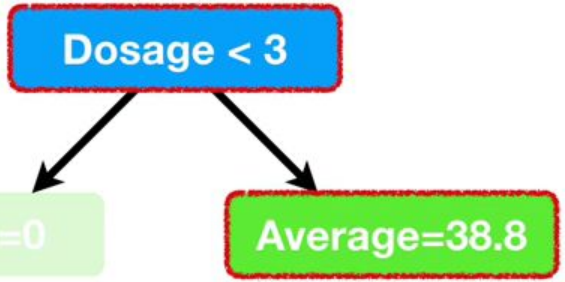




$$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (5 - 38.8)^2 + (20 - 38.8)^2$$

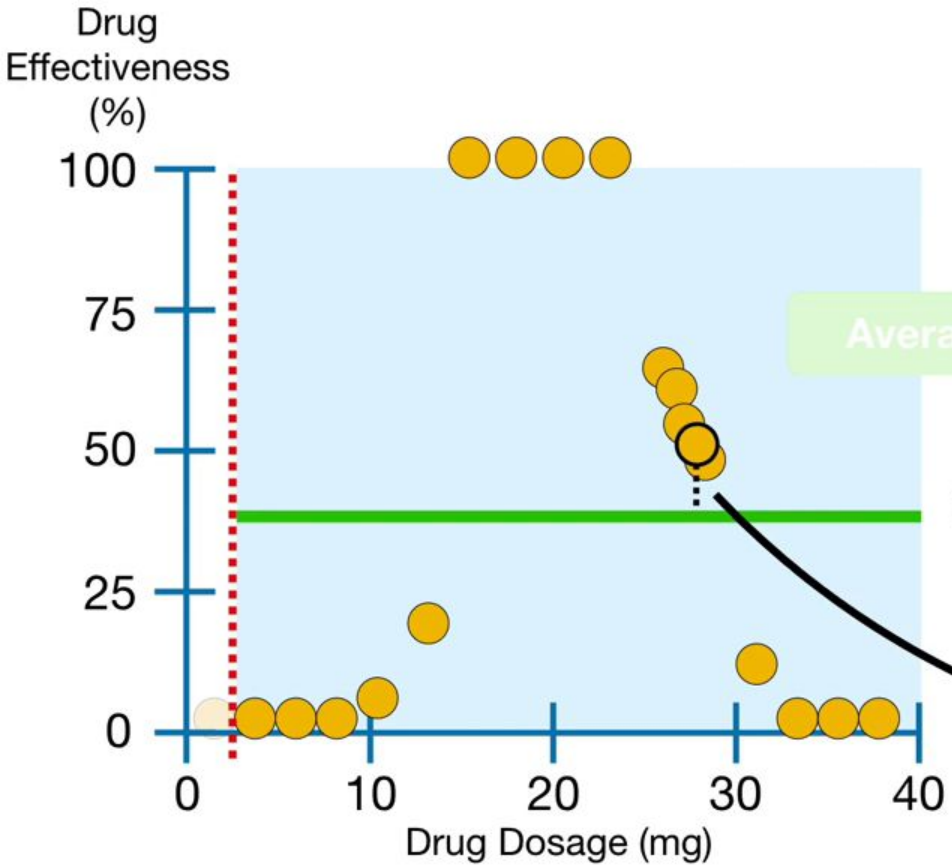
…and the rest of the points…

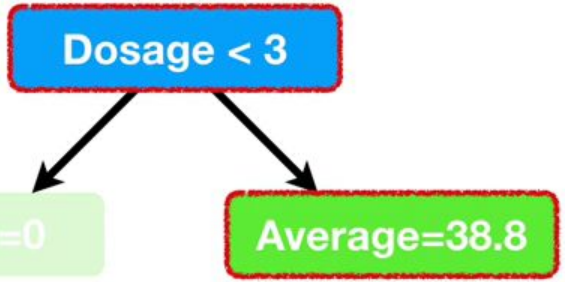




$$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2$$
$$+ (5 - 38.8)^2 + (20 - 38.8)^2 + (100 - 38.8)^2$$
$$+ (100 - 38.8)^2 + \ldots$$

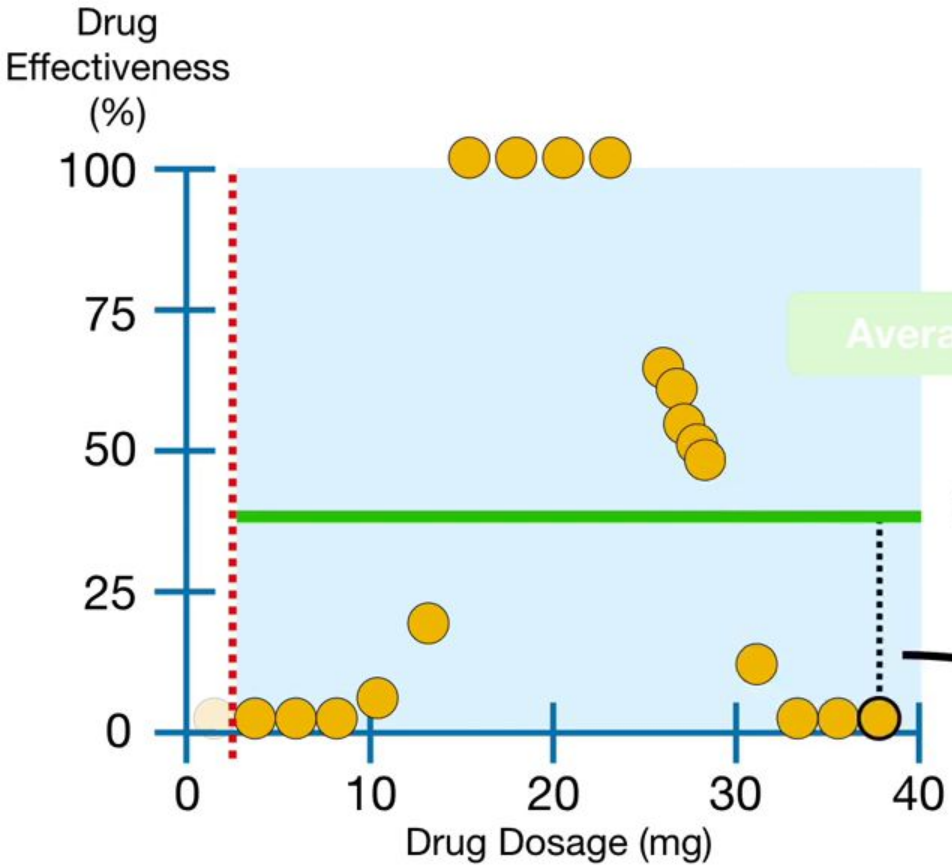
…and the rest of the points…

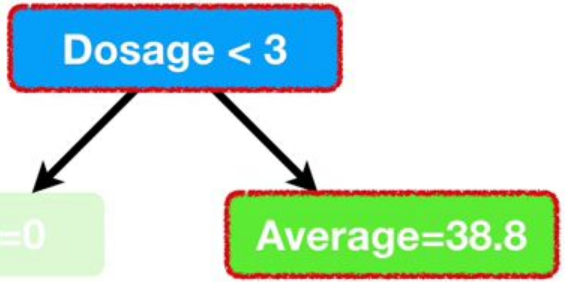




$$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2$$
$$+ (5 - 38.8)^2 + (20 - 38.8)^2 + (100 - 38.8)^2$$
$$+ (100 - 38.8)^2 + \ldots$$

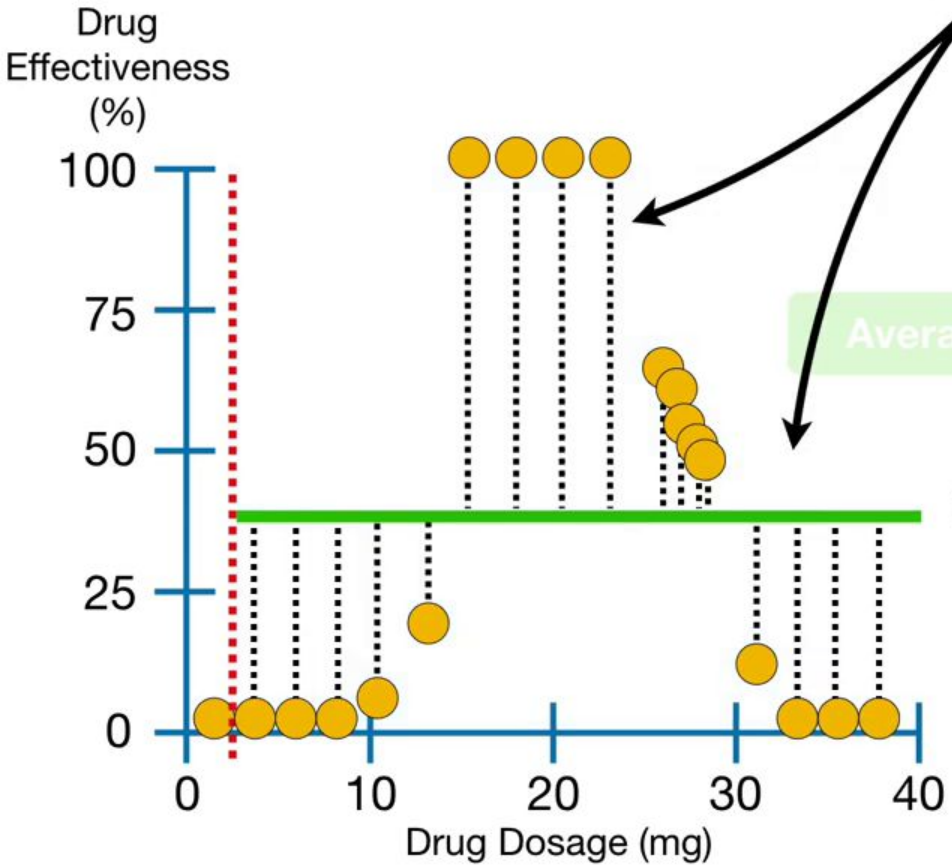
…and the rest of the points…





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

…until we have added squared residuals for every point.



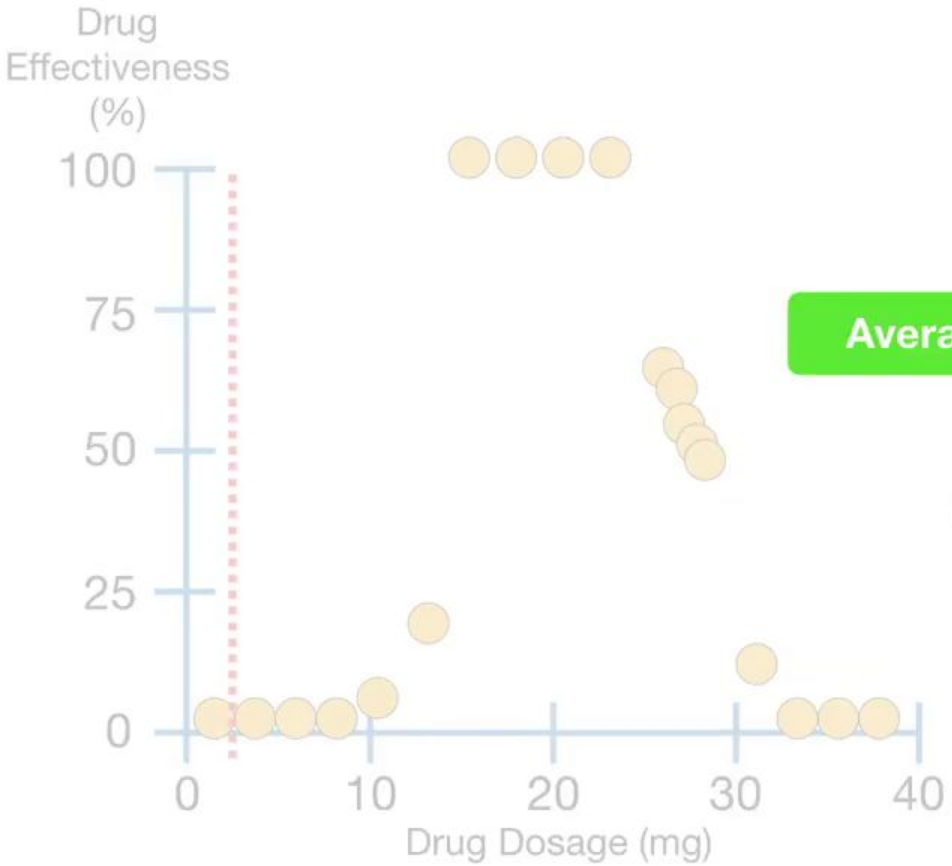
Dosage < 3

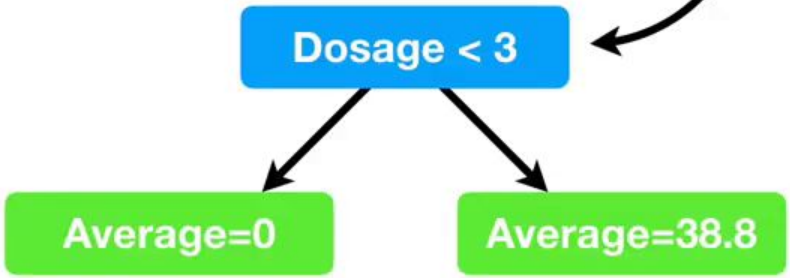
Average=0

Average=38.8

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

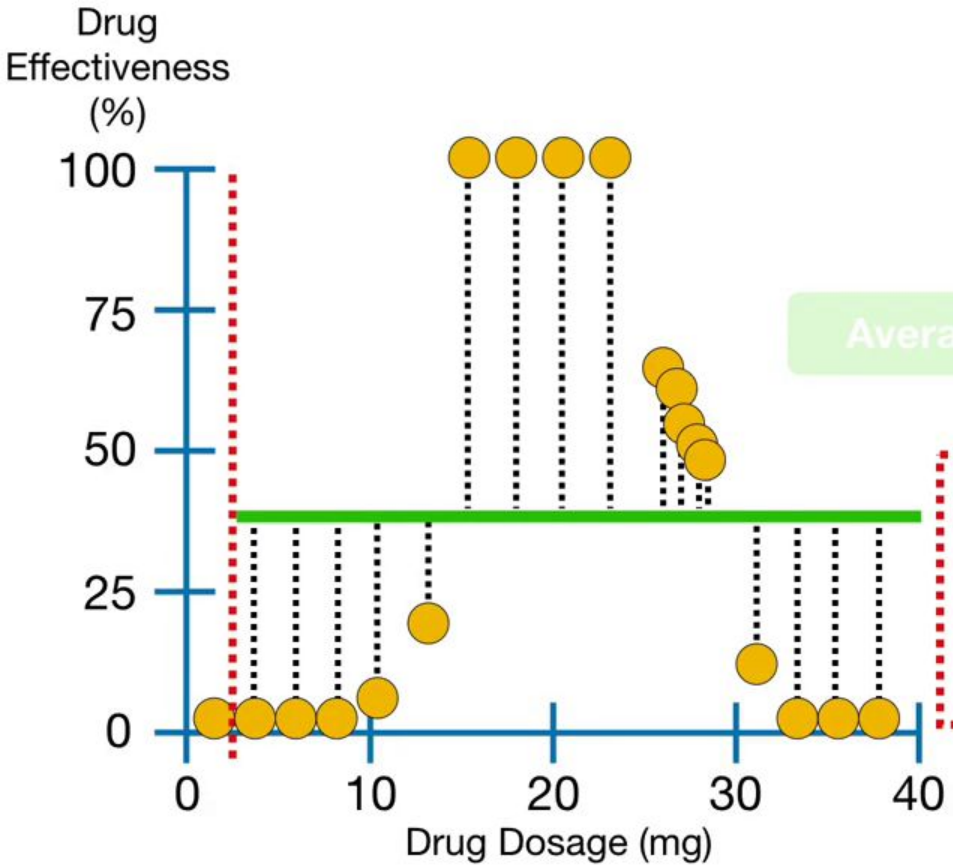
Thus, to evaluate the predictions made when the threshold is **Dosage < 3**…

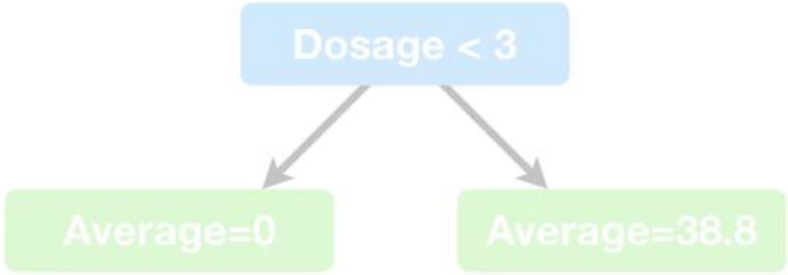




$$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2$$
$$+ (5 - 38.8)^2 + (20 - 38.8)^2 + (100 - 38.8)^2$$
$$+ (100 - 38.8)^2 + \ldots + (0 - 38.8)^2$$

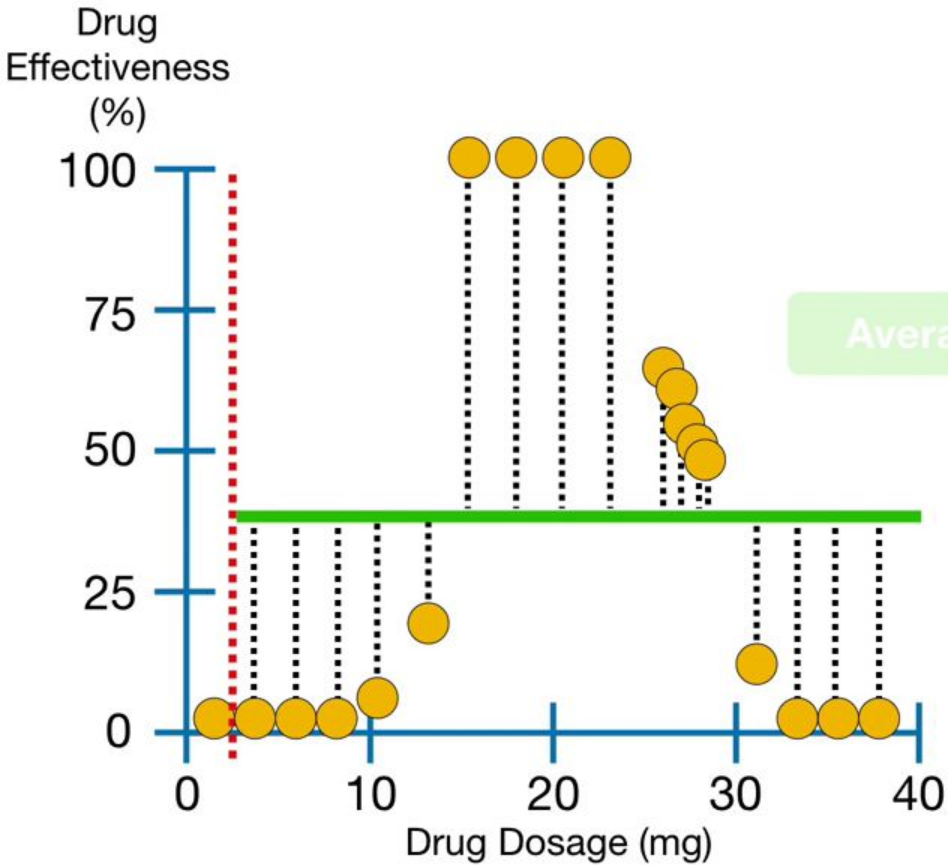
…we add up the squared residuals for every point…

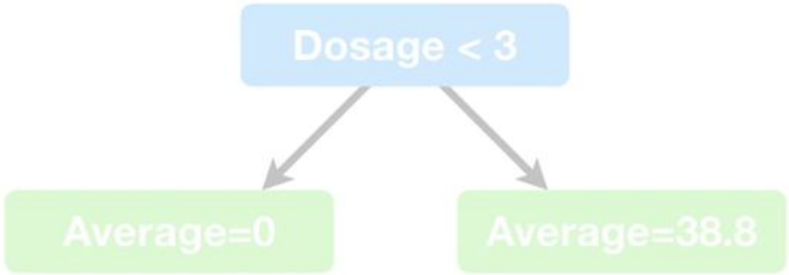




$$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2$$
$$+ (5 - 38.8)^2 + (20 - 38.8)^2 + (100 - 38.8)^2$$
$$+ (100 - 38.8)^2 + \ldots + (0 - 38.8)^2$$

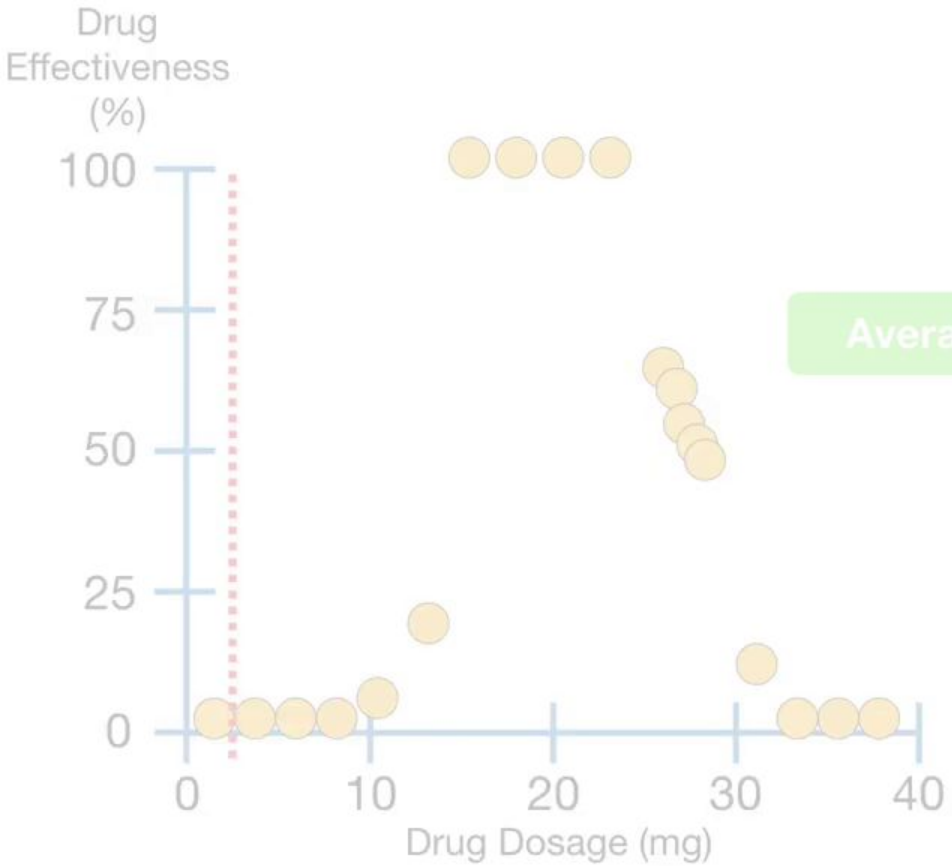
…and get **27,468.5**.

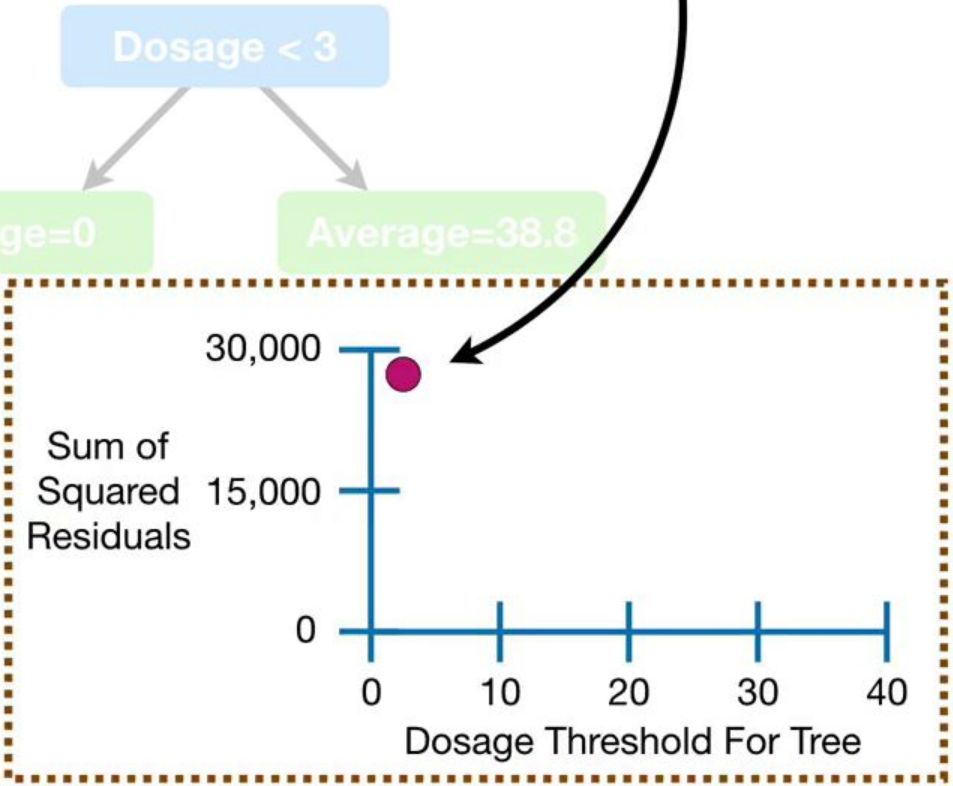


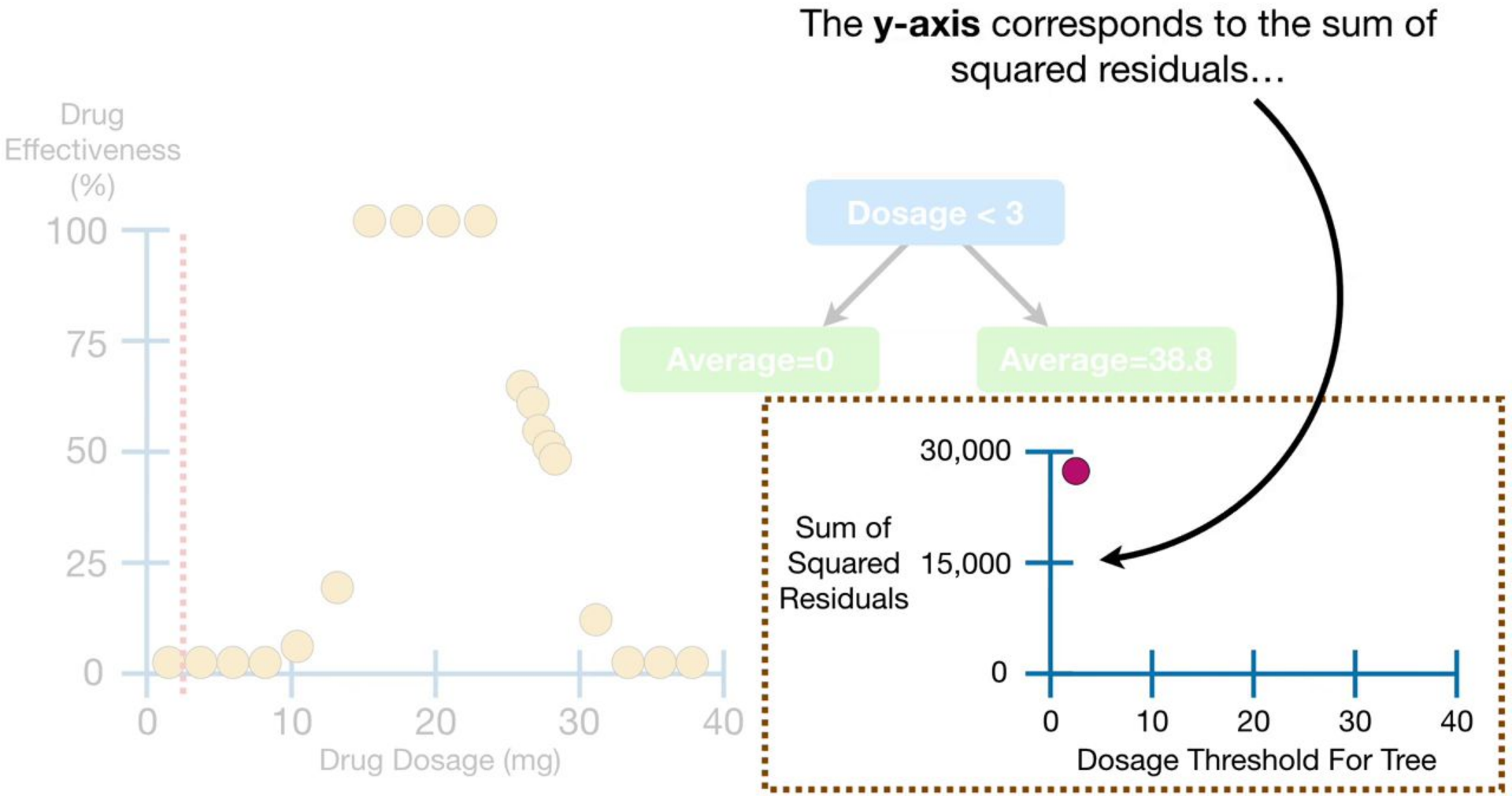


$$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2$$
$$+ (5 - 38.8)^2 + (20 - 38.8)^2 + (\ 00 - 38.8)^2$$
$$+ (100 - 38.8)^2 + \ldots + (0 - 38.8)^2$$
$$= 27{,}468.5$$

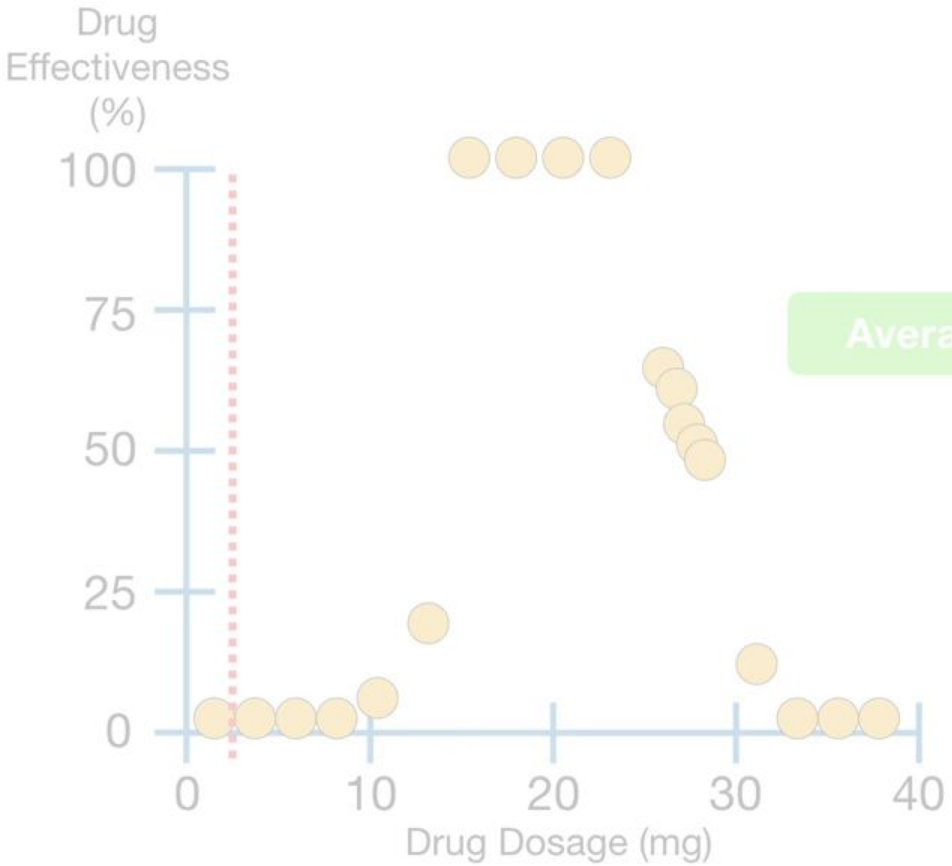
NOTE: We can plot the sum of squared residuals on this graph.

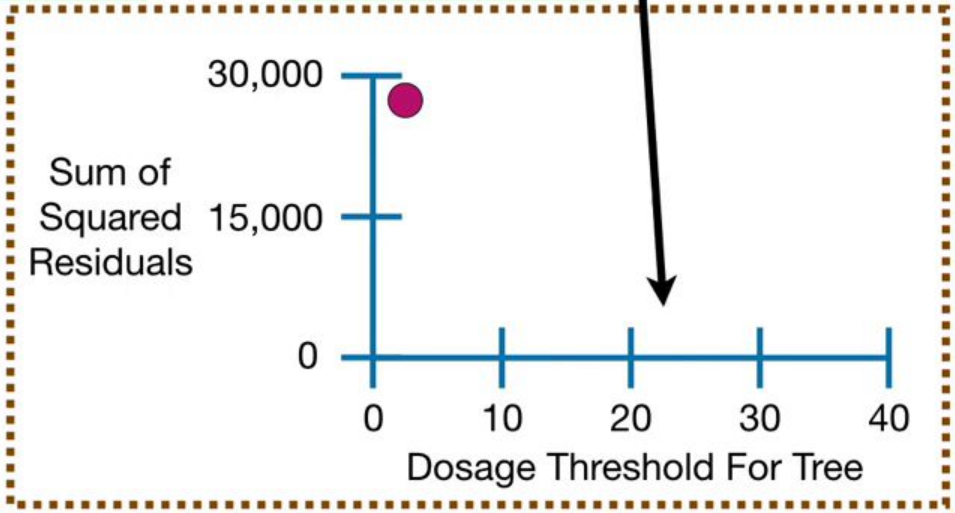


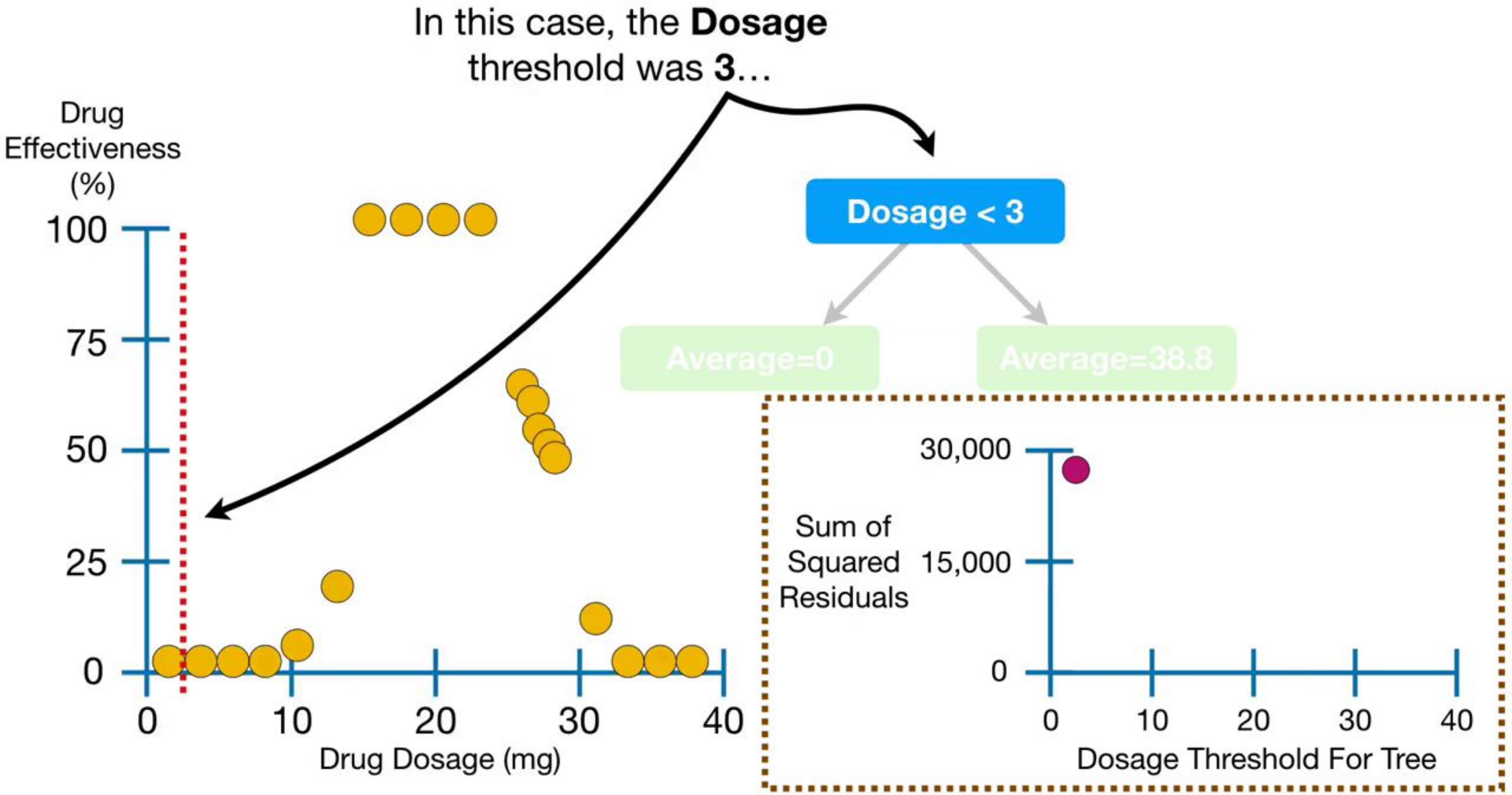



The **y-axis** corresponds to the sum of squared residuals…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 3
Average=0
Average=38.8
30,000
Sum of Squared Residuals
15,000
0
0
10
20
30
40
Dosage Threshold For Tree

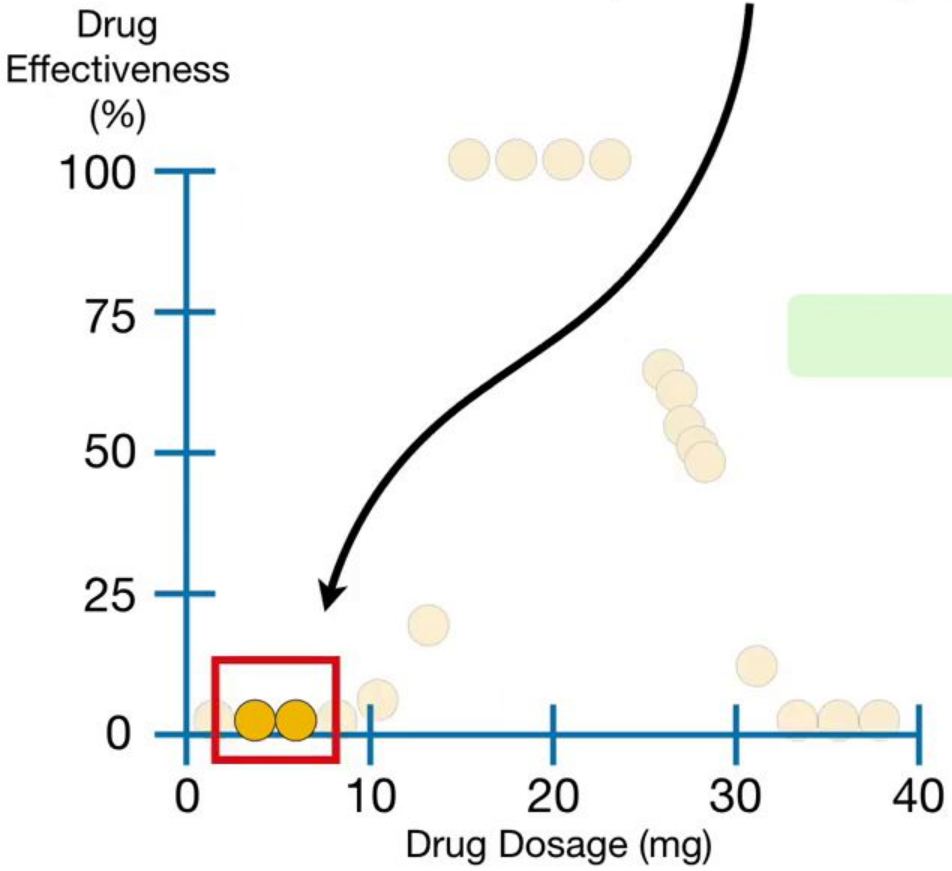
…and the **x-axis** corresponds to **Dosage** thresholds.

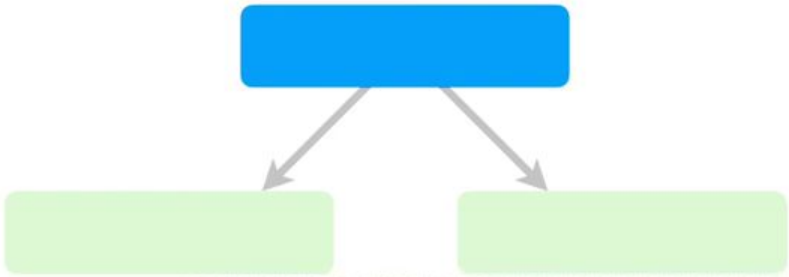


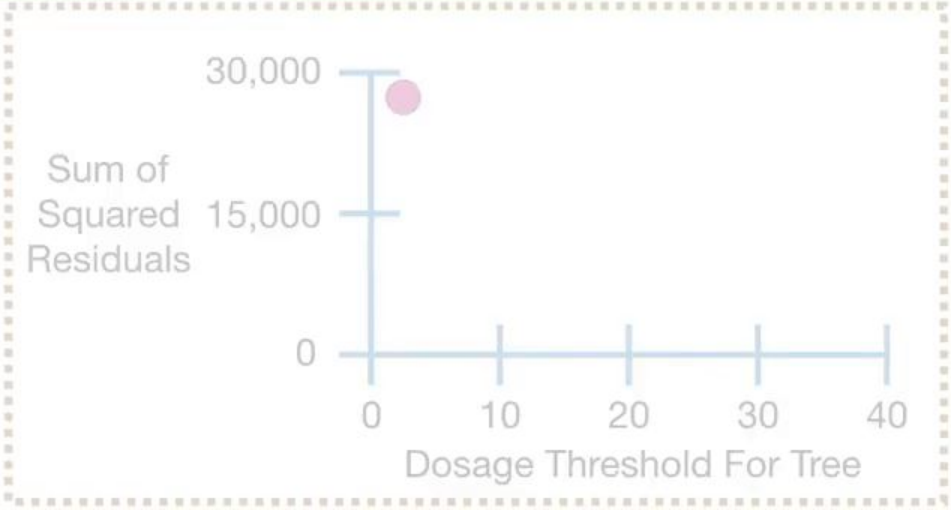



In this case, the **Dosage** threshold was **3**…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 3
Average=0
Average=38.8
30,000
15,000
0
Sum of Squared Residuals
0
10
20
30
40
Dosage Threshold For Tree

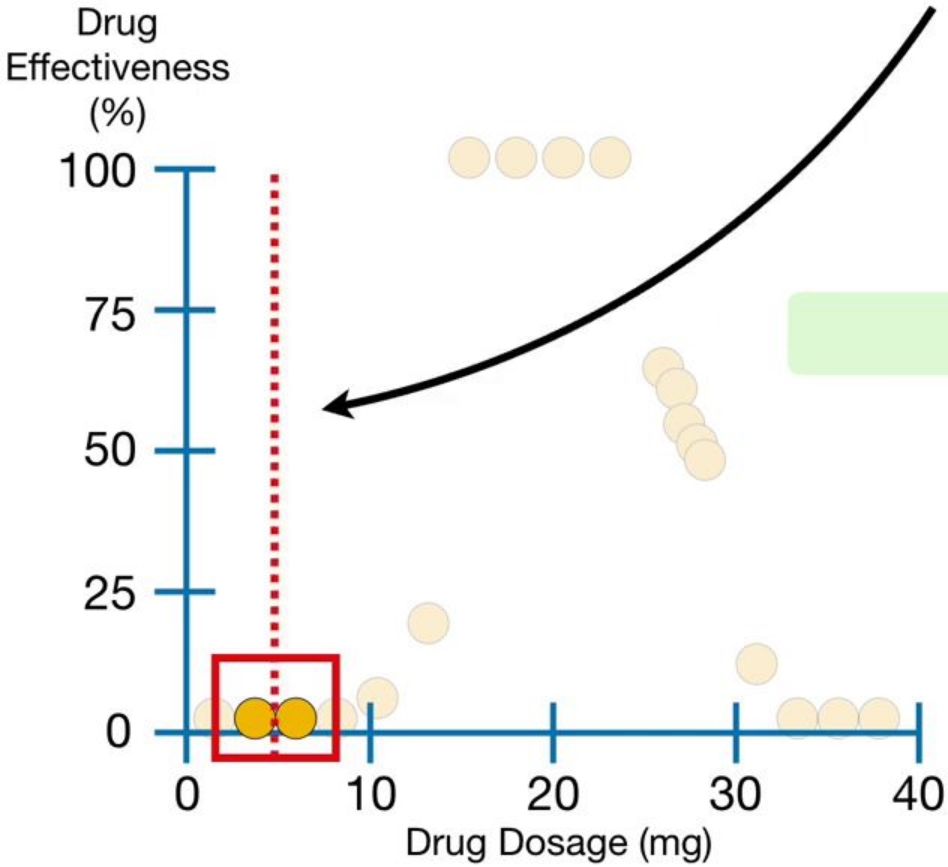
…but if we focus on the next two points in the graph…

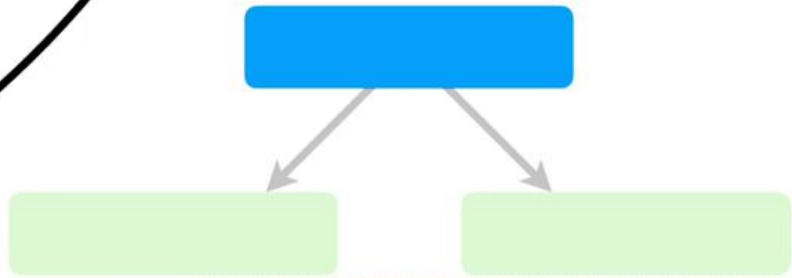


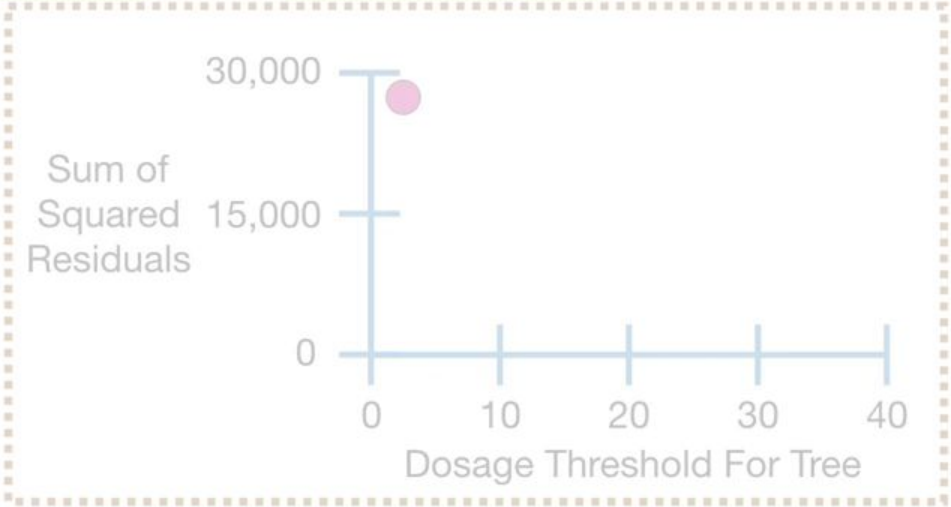




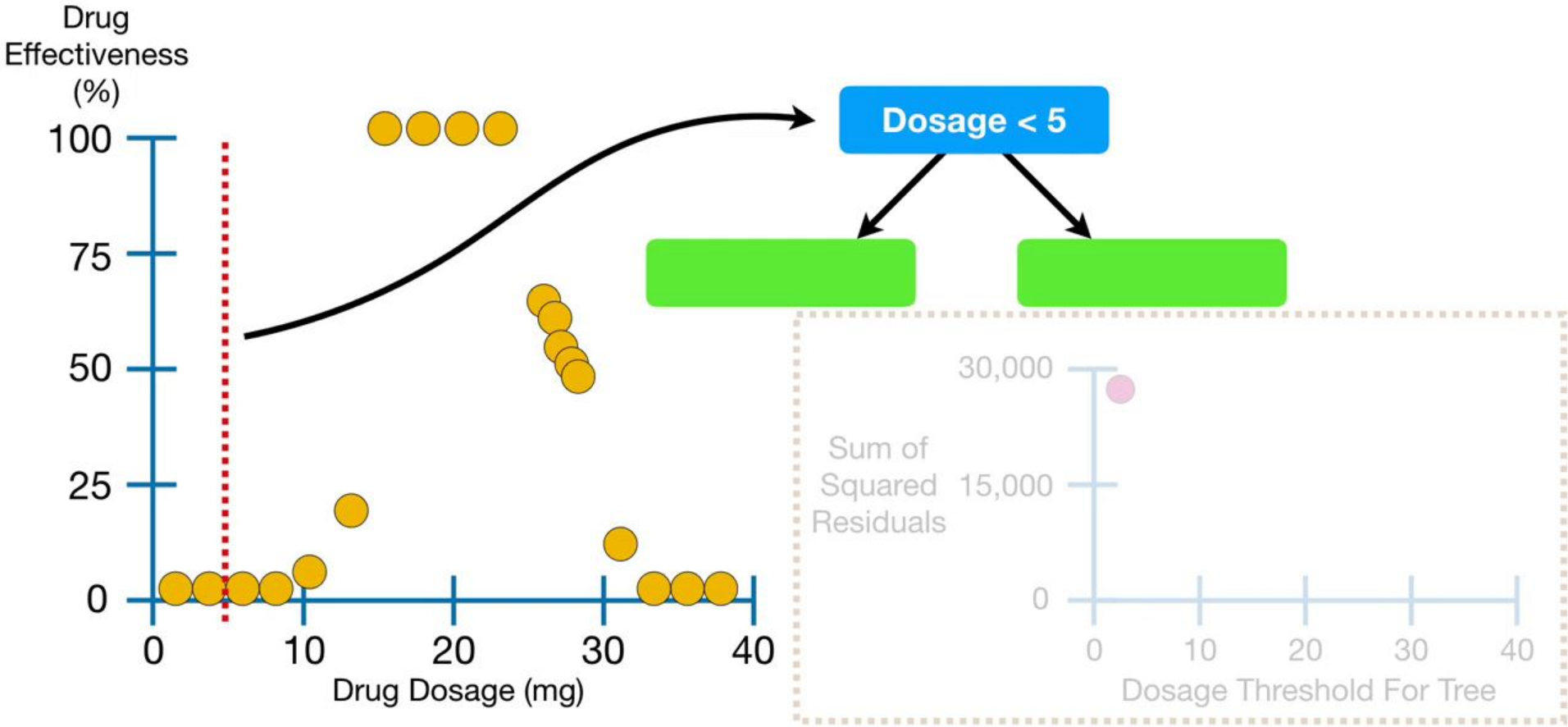
…and calculate their average **Dosage**, which is **5**…



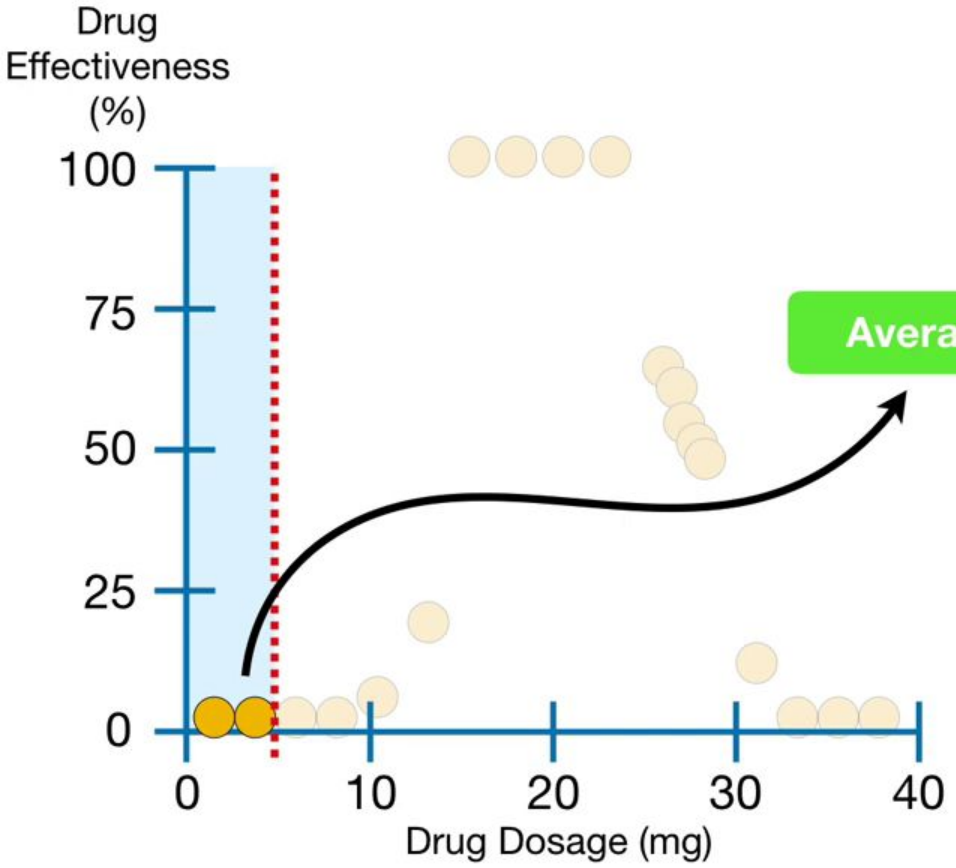


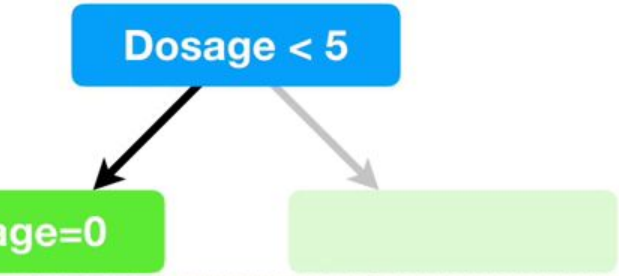


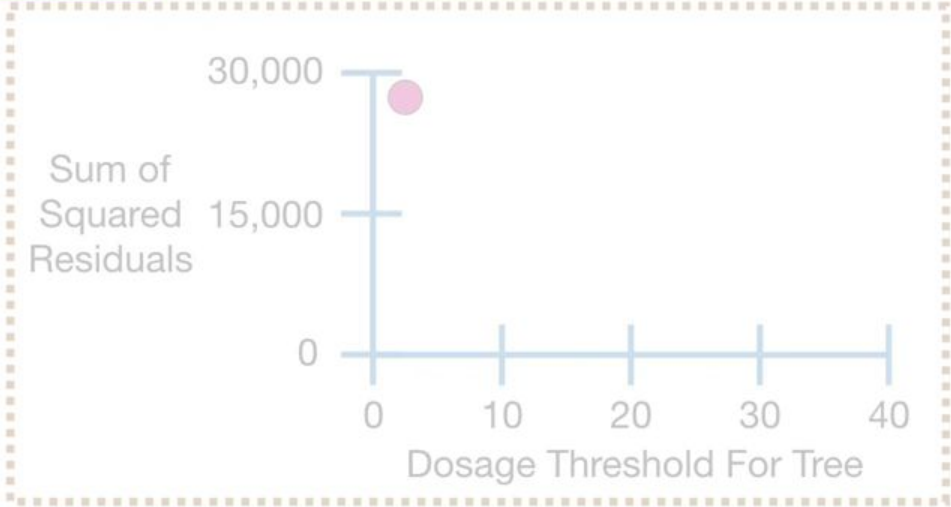
…then we can use **Dosage < 5**
as a new threshold.



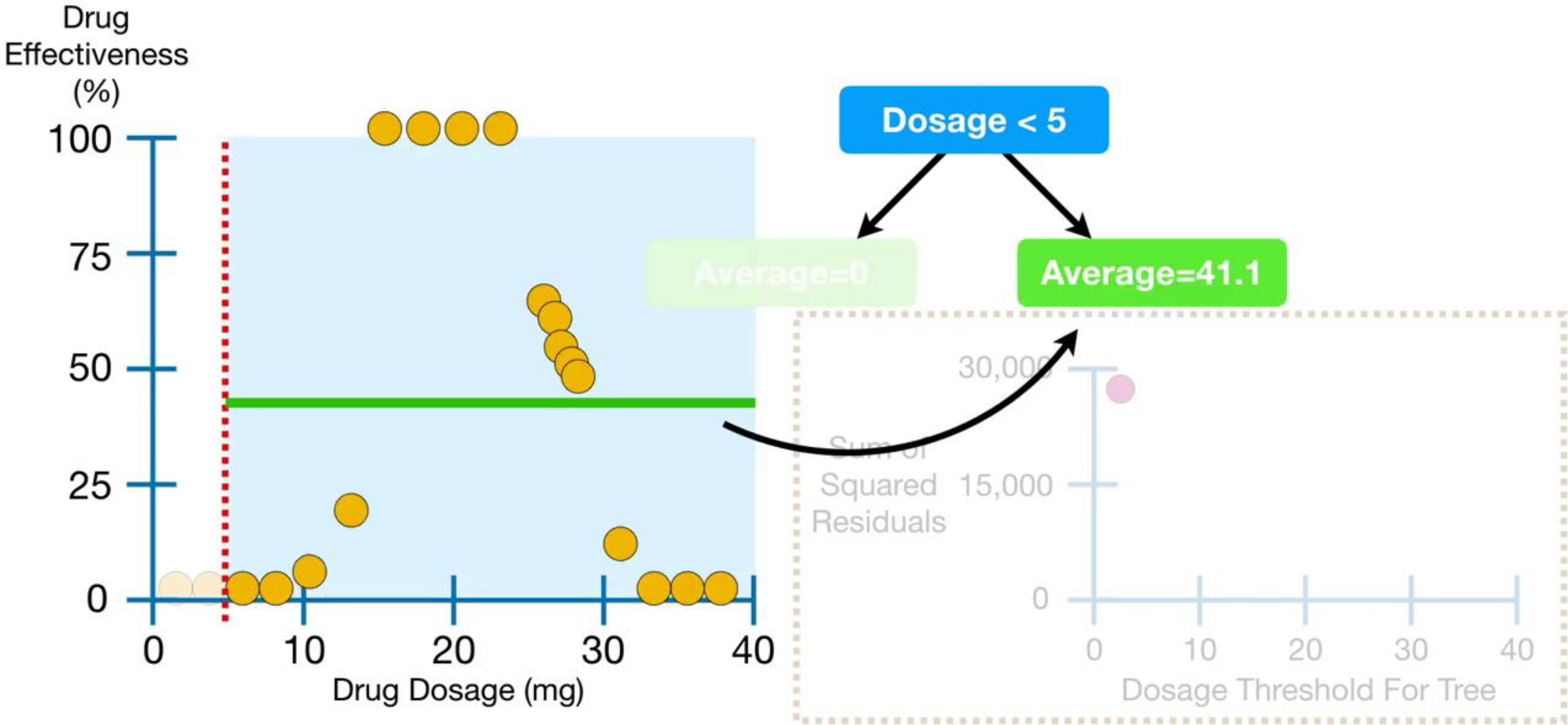
Using **Dosage < 5** gives us new predictions…



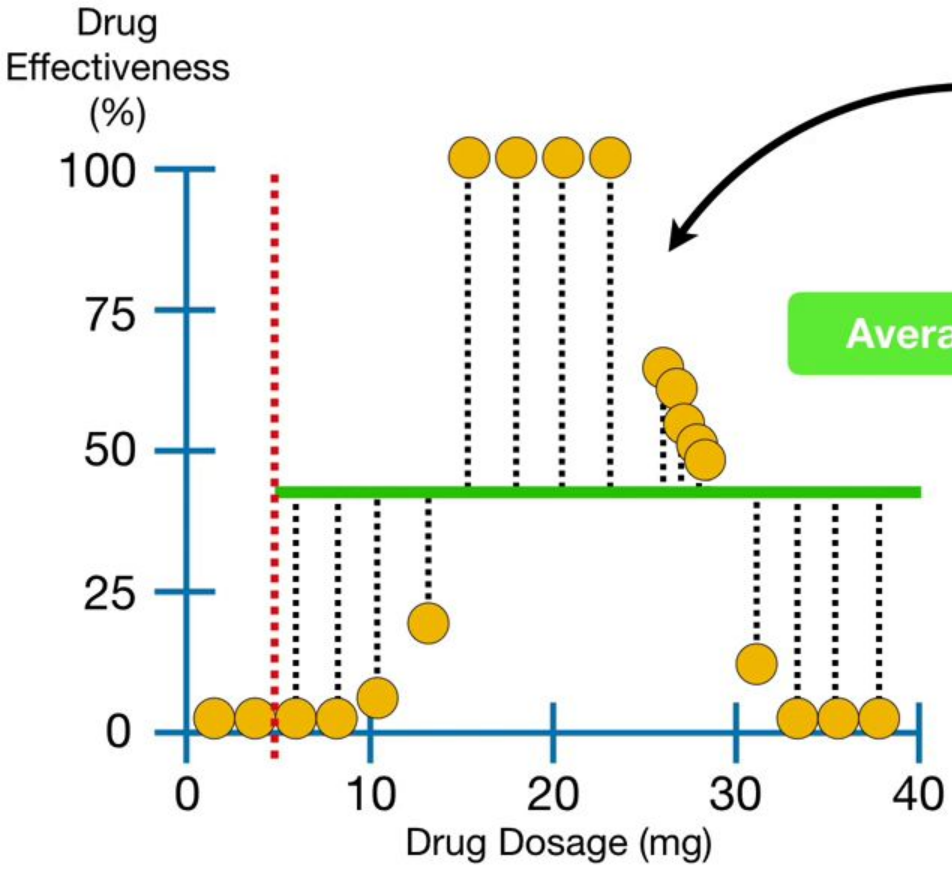


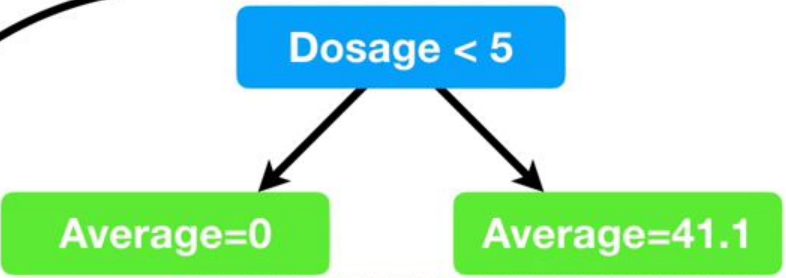


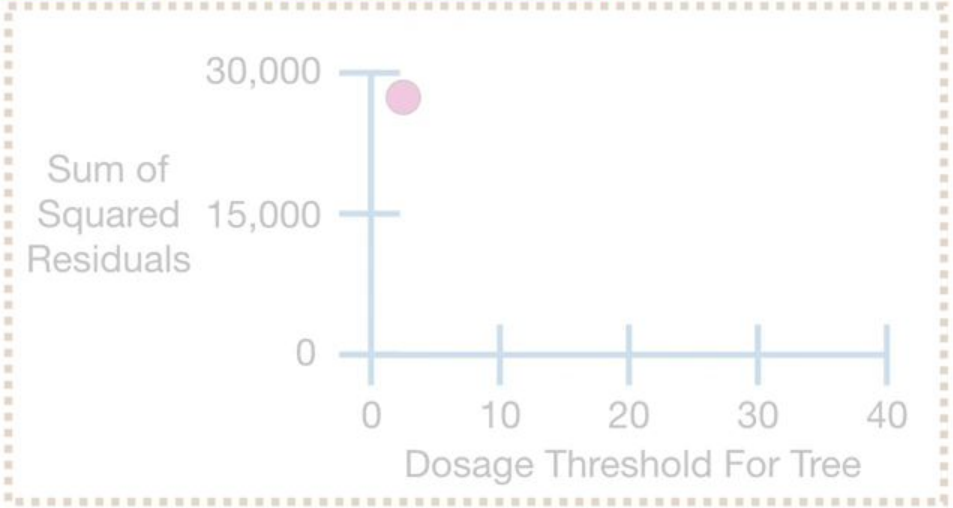
Using **Dosage < 5** gives us new predictions…



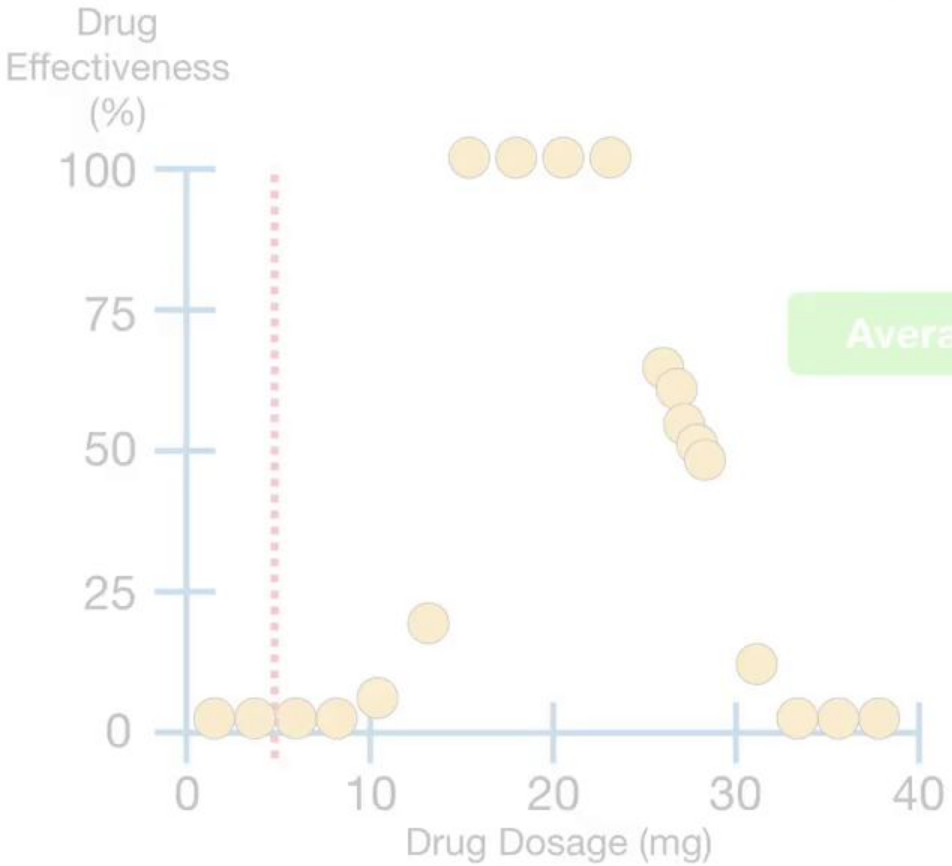
…and new residuals…

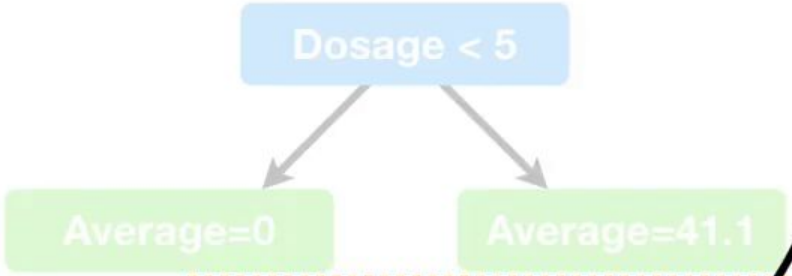


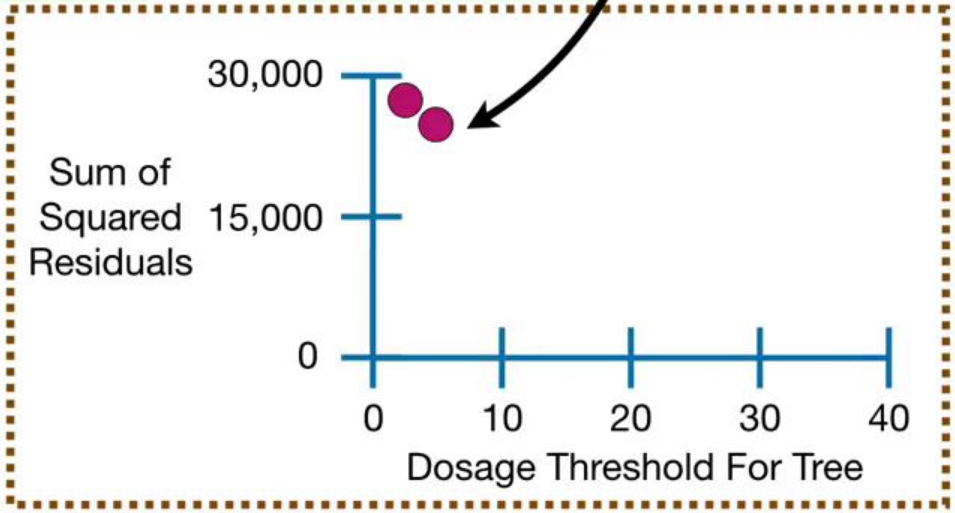




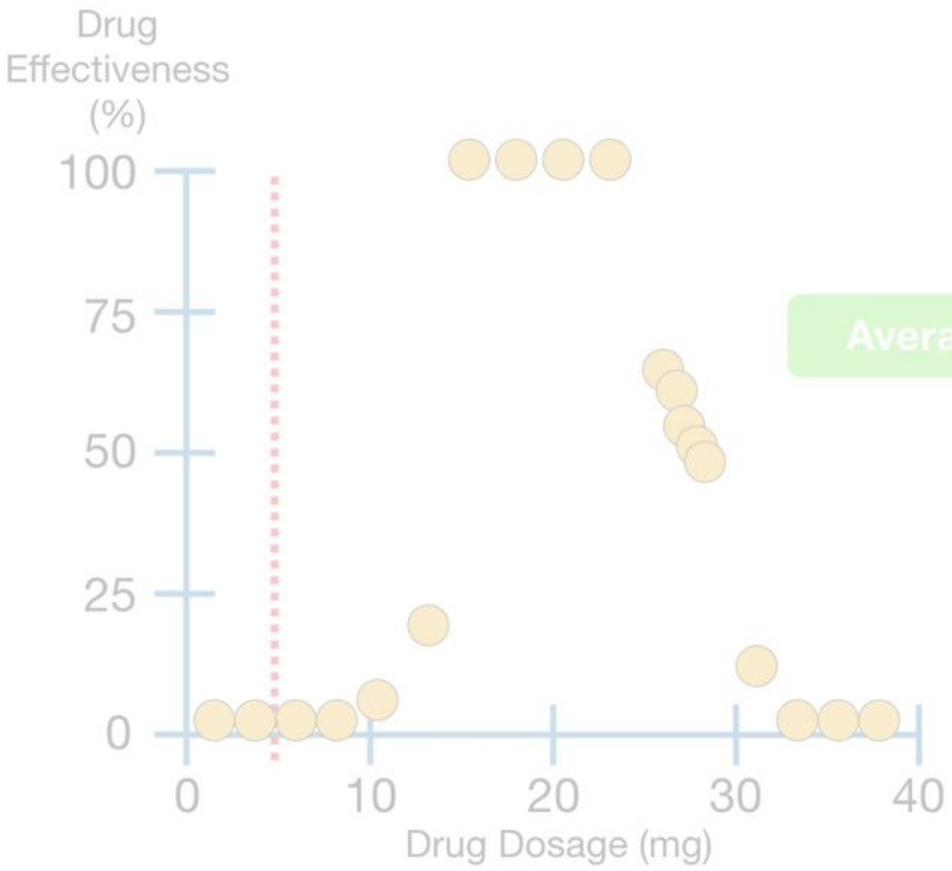
…and that means we can add a new sum of squared residuals to our graph.

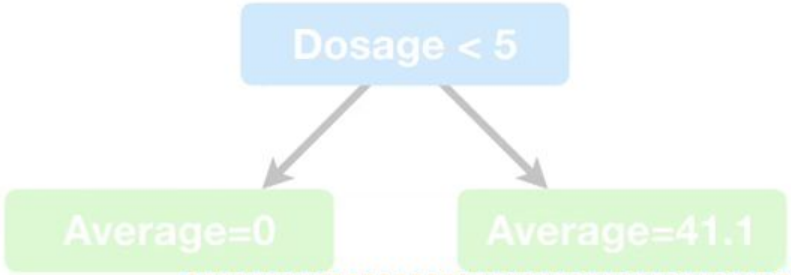


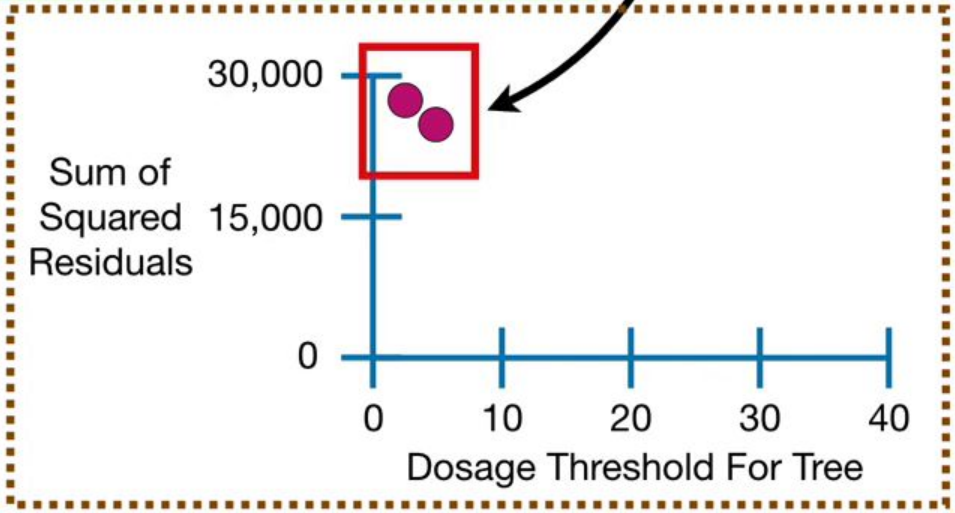




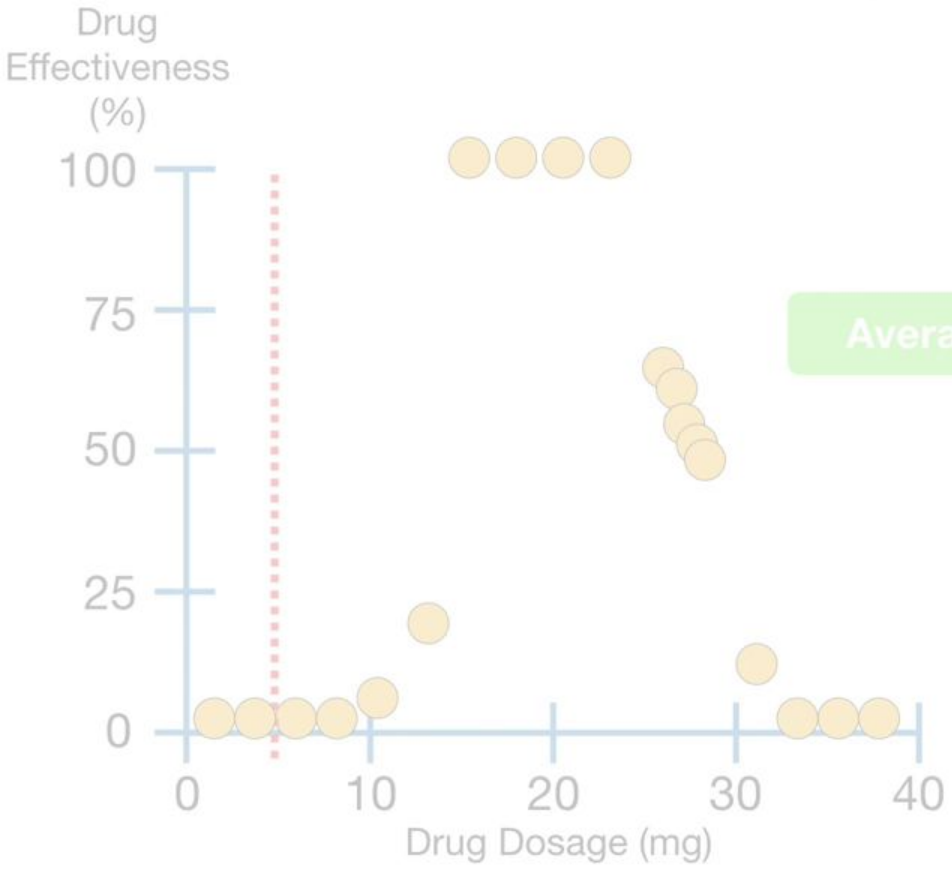
In this case, the new threshold, **Dosage < 5**, results in a smaller sum of squared residuals…

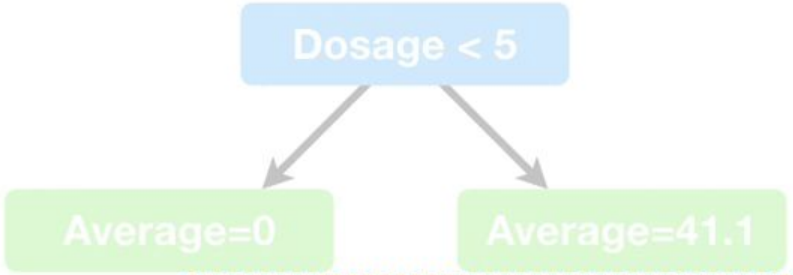


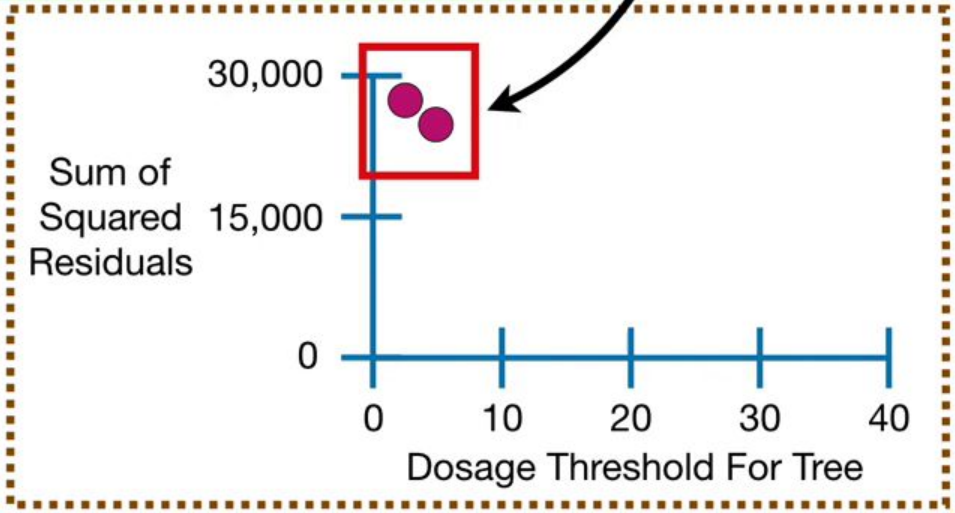


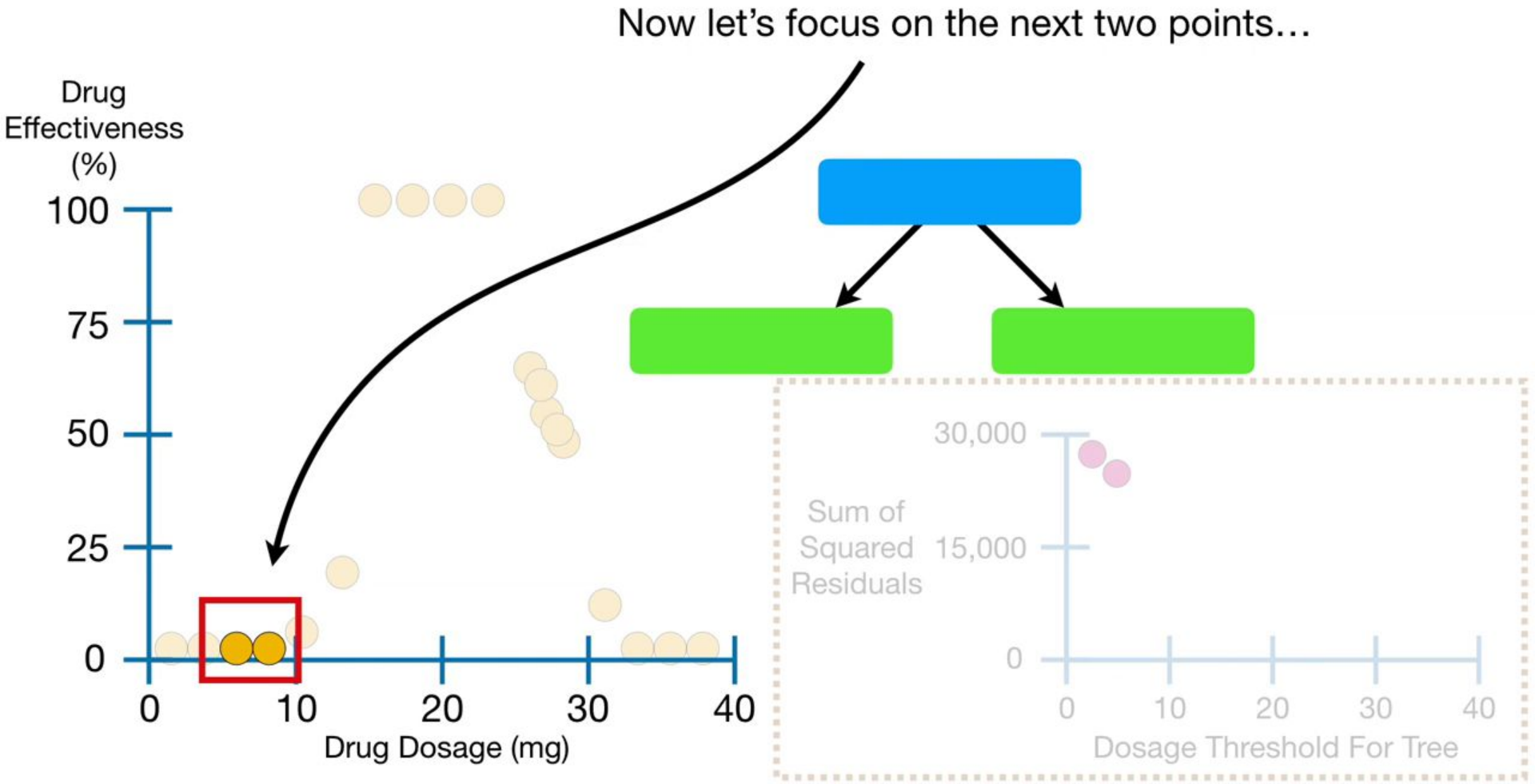


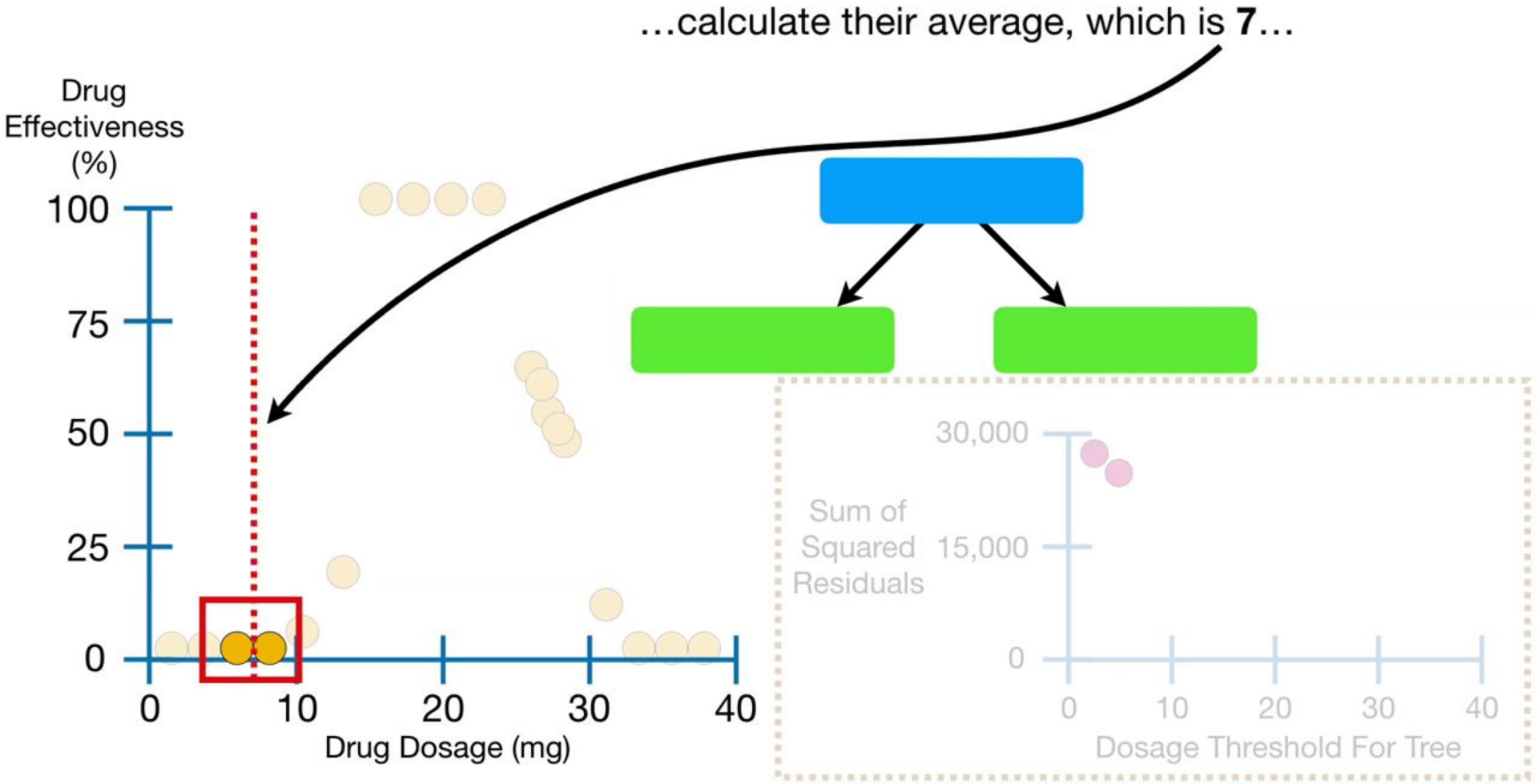
...and that means using **Dosage < 5** as the threshold resulted in better predictions over all.



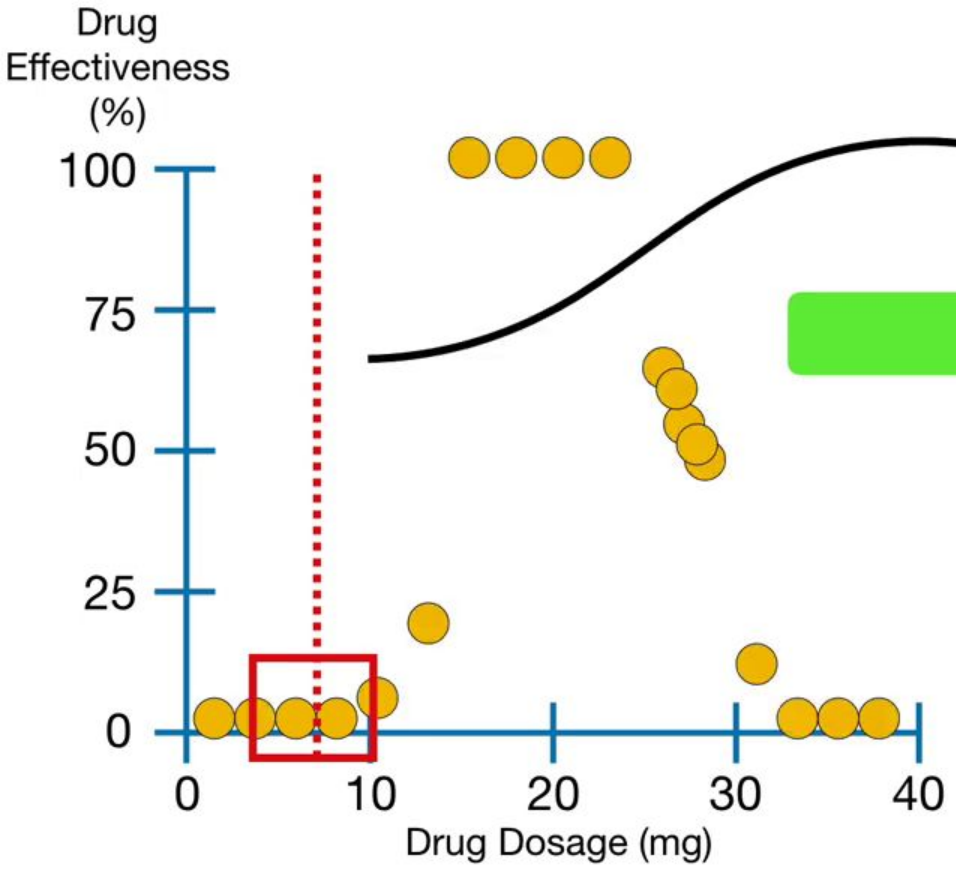


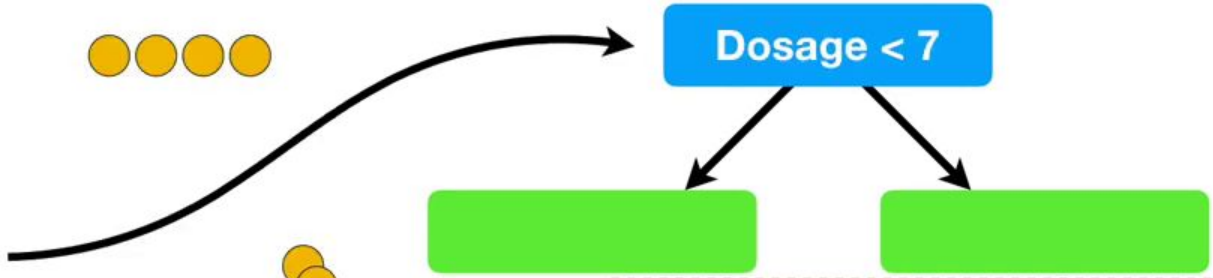


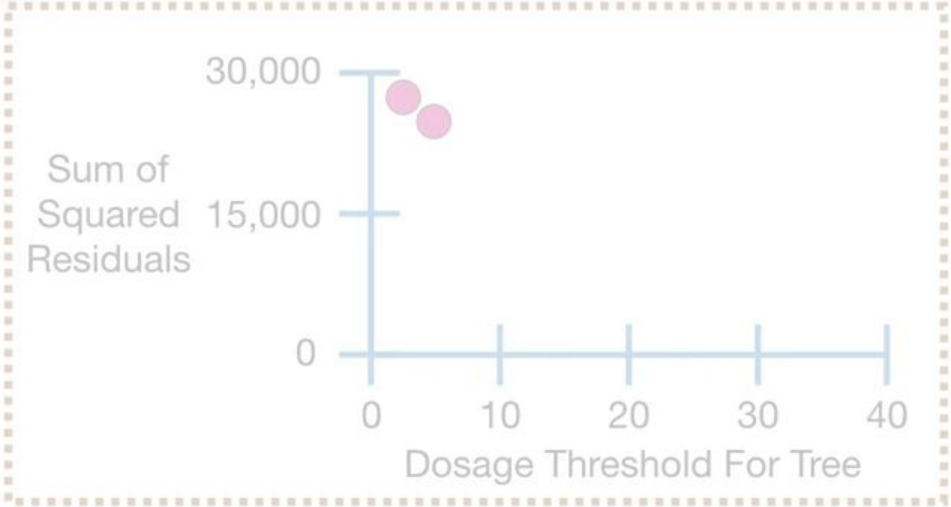

Now let's focus on the next two points…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
30,000
Sum of Squared Residuals
15,000
0
0
10
20
30
40
Dosage Threshold For Tree


…calculate their average, which is 7…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
30,000
Sum of Squared Residuals
15,000
0
0
10
20
30
40
Dosage Threshold For Tree

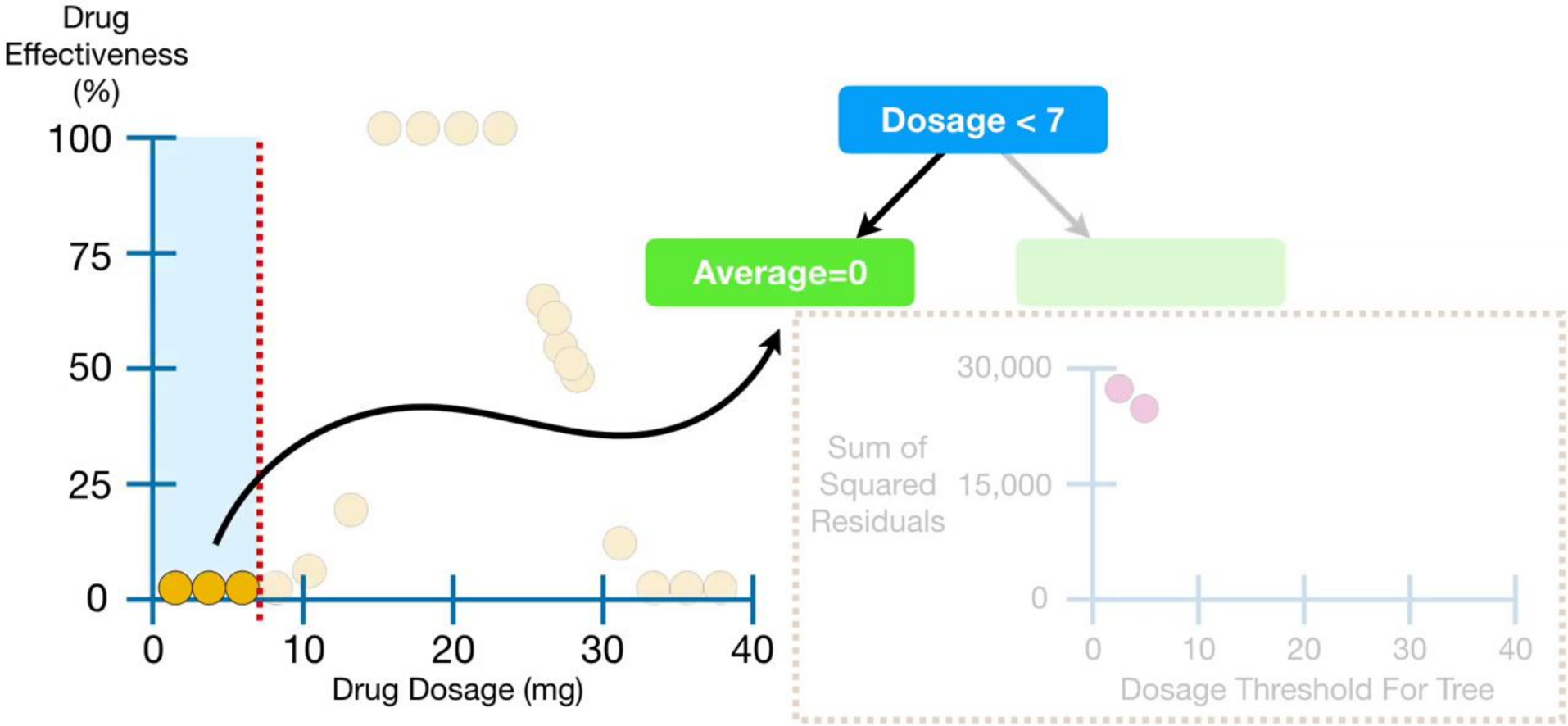
…and use **Dosage < 7** as a new threshold.



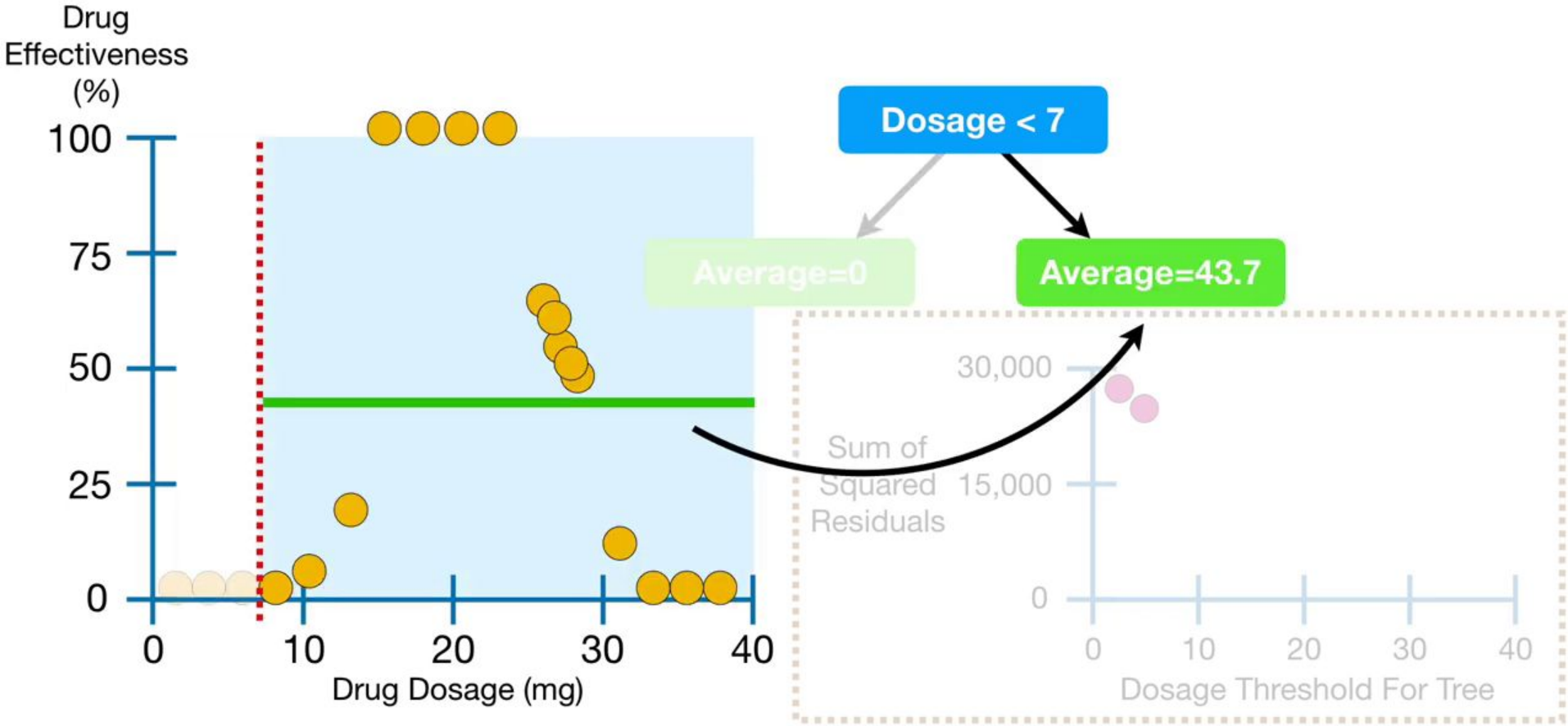


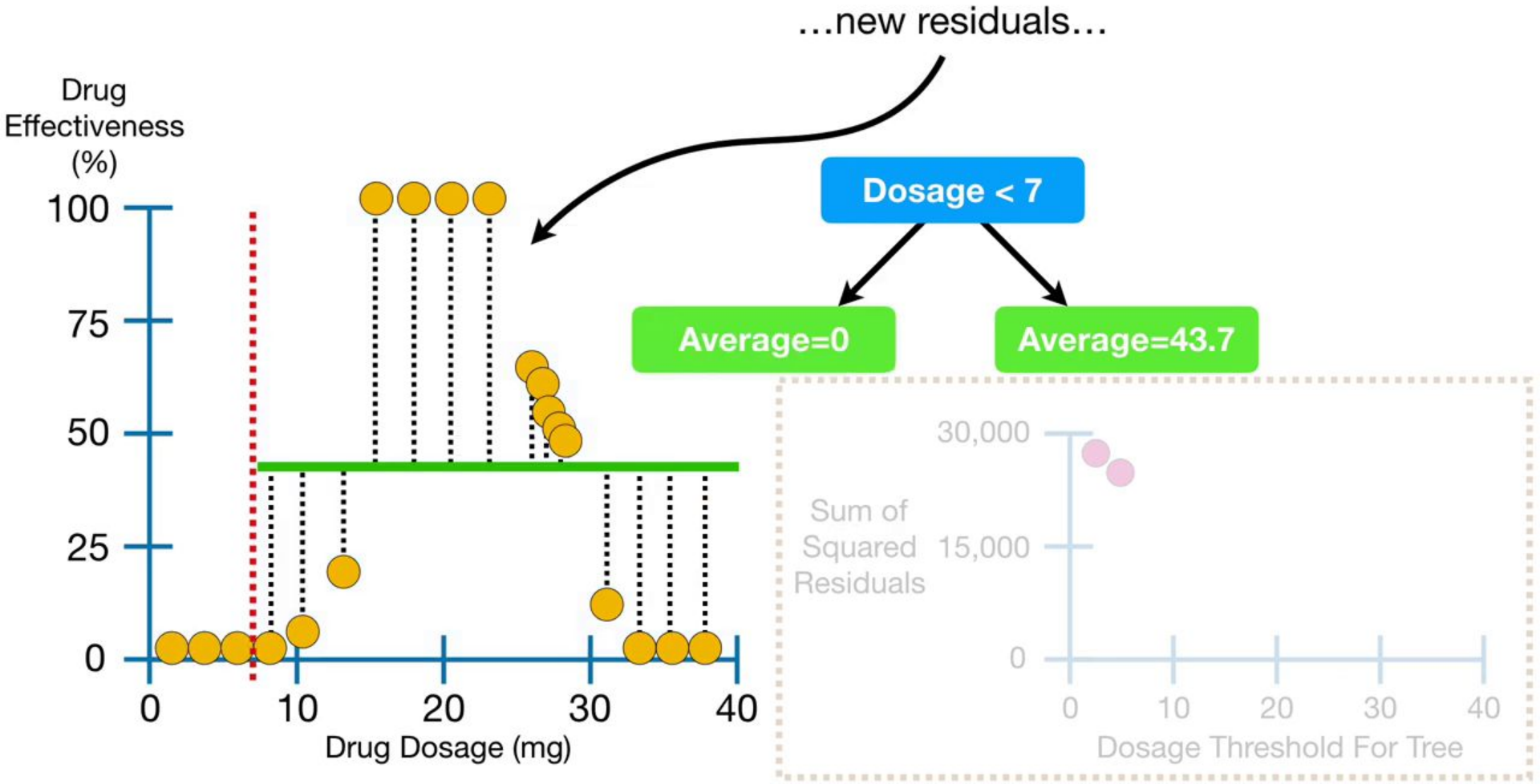


Again, the new threshold gives us new predictions…

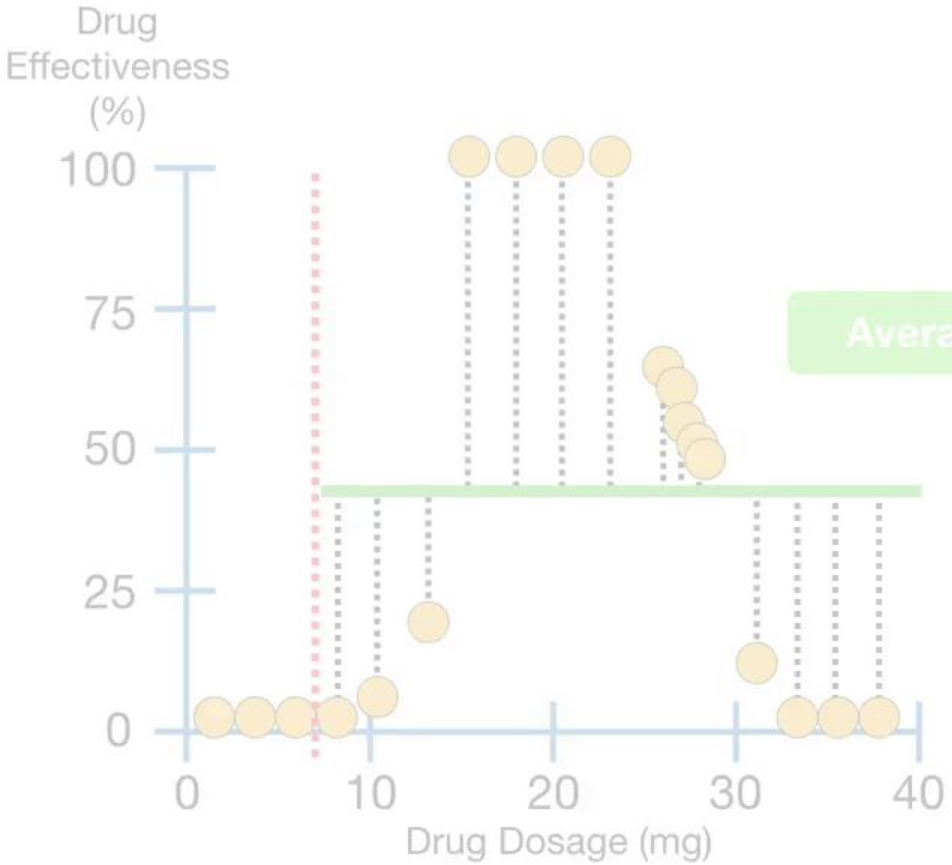


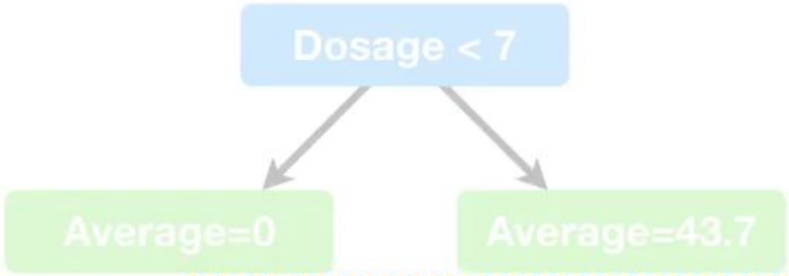
Again, the new threshold gives us new predictions…

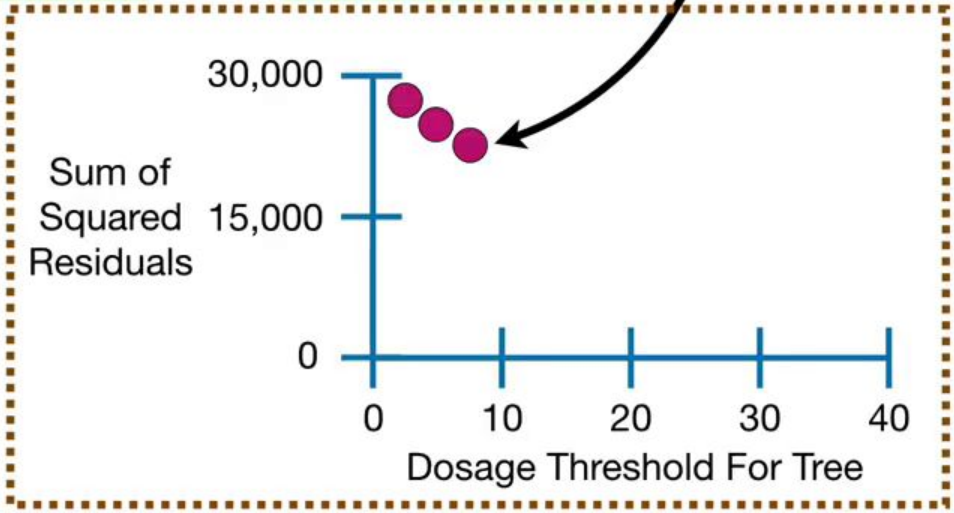


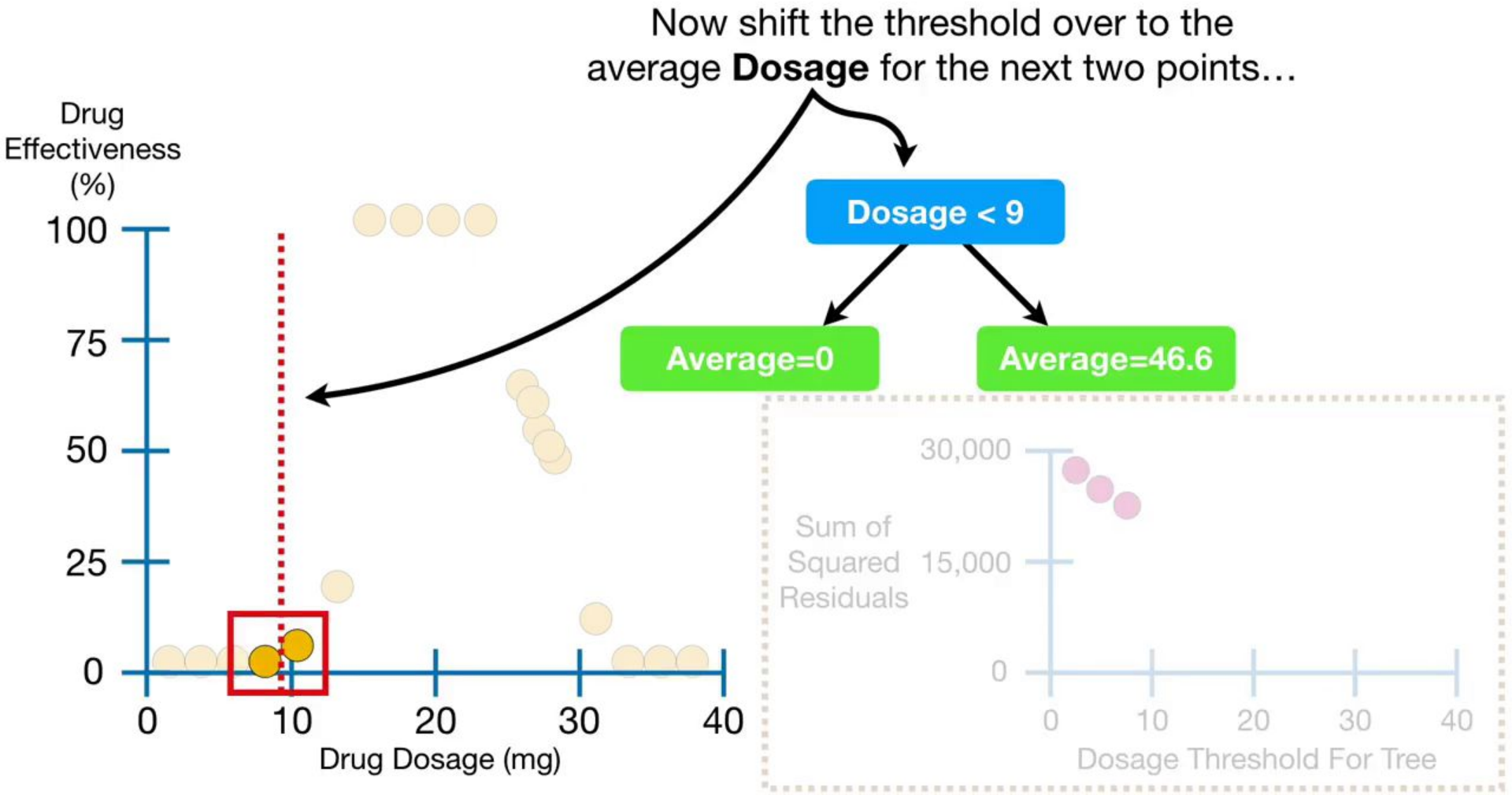

…new residuals…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 7
Average=0
Average=43.7
30,000
Sum of Squared Residuals
15,000
0
0
10
20
30
40
Dosage Threshold For Tree

…and a new sum of squared residuals.

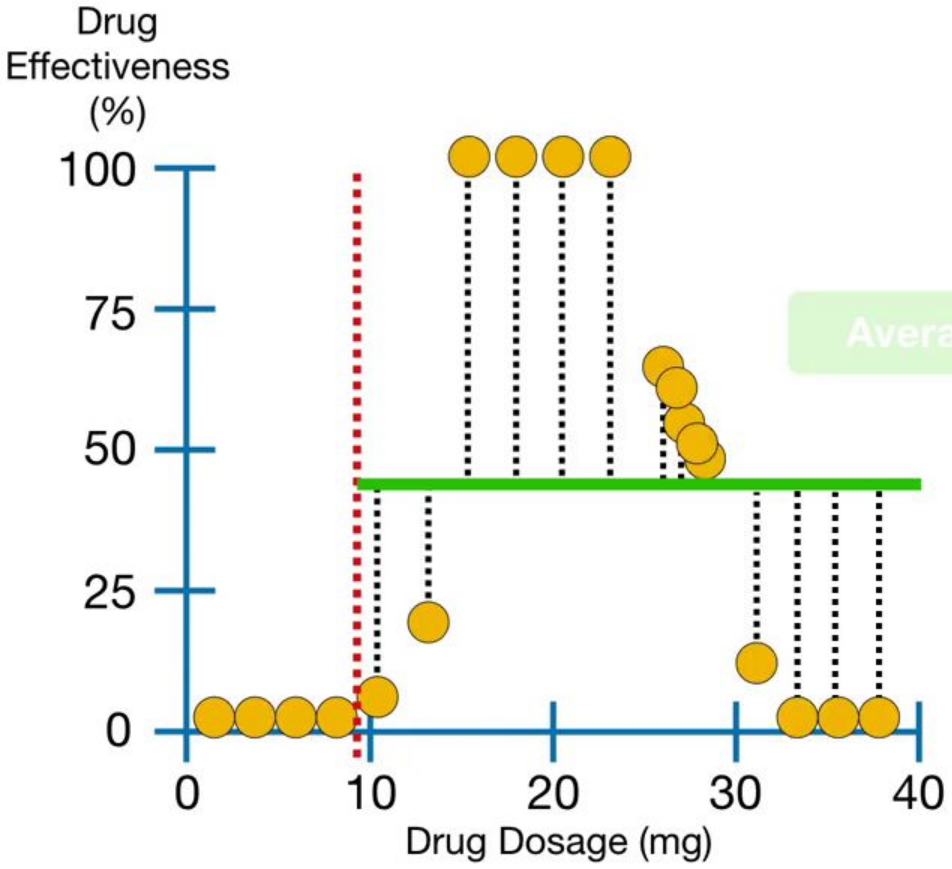


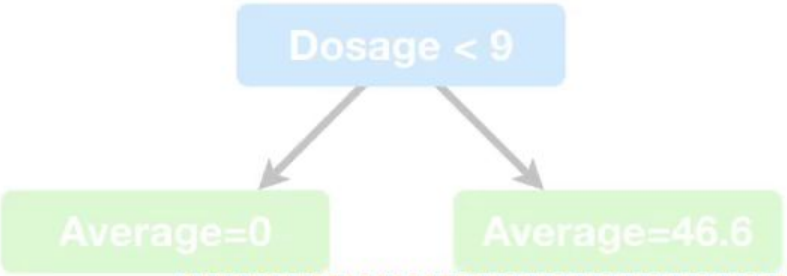


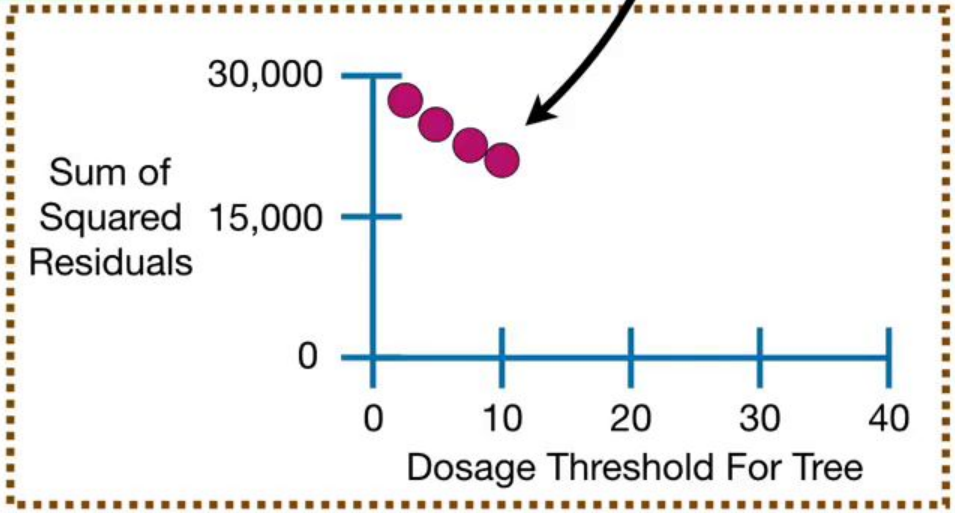



Now shift the threshold over to the average **Dosage** for the next two points…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 9
Average=0
Average=46.6
30,000
Sum of Squared Residuals
15,000
0
0
10
20
30
40
Dosage Threshold For Tree

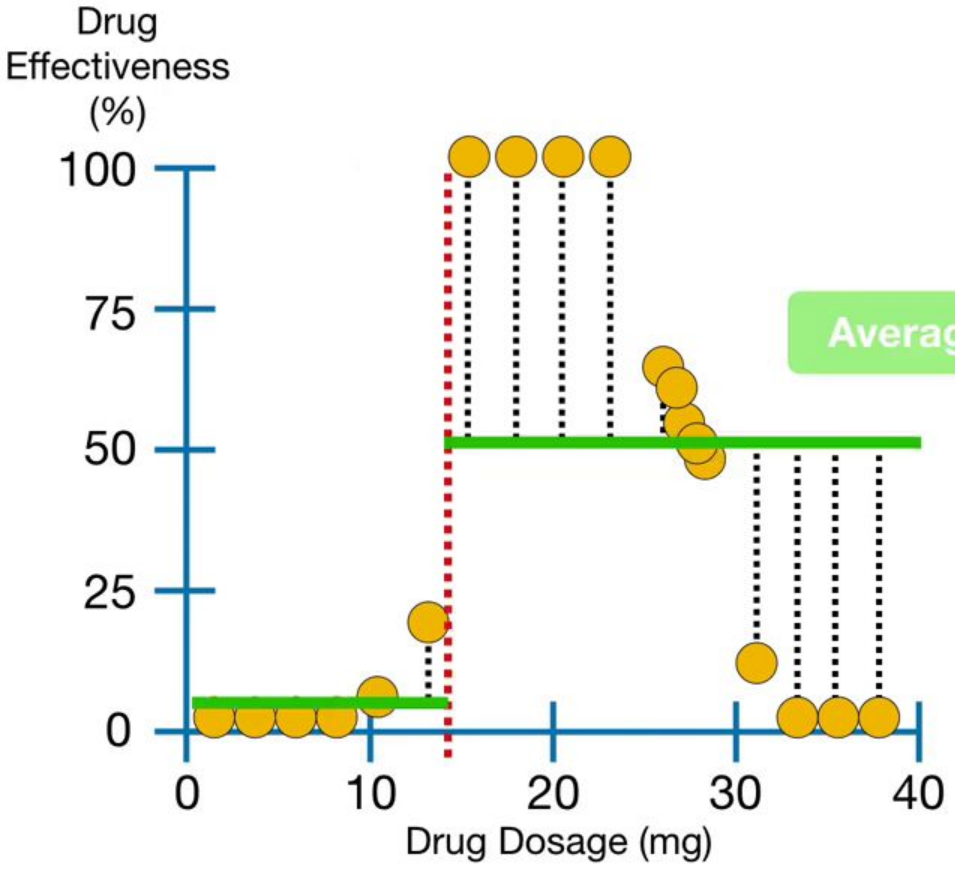
…and add the new sum of squared residuals to the graph.

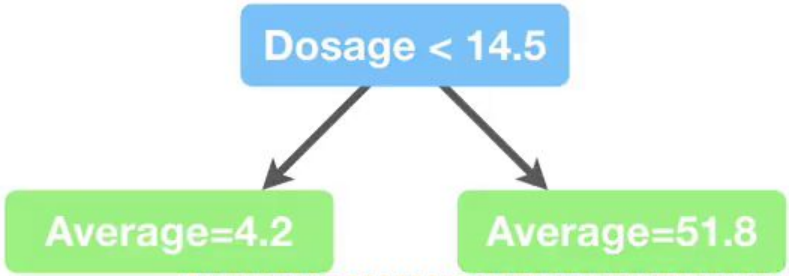


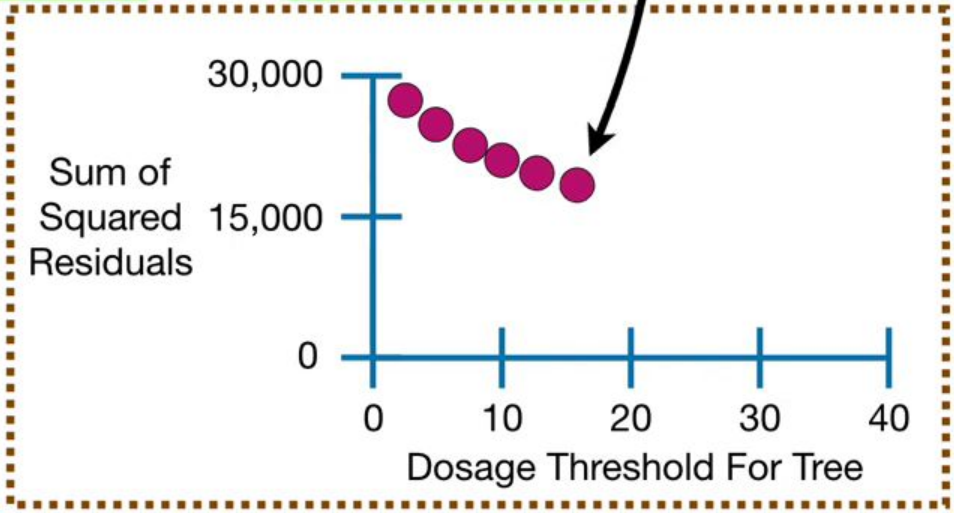




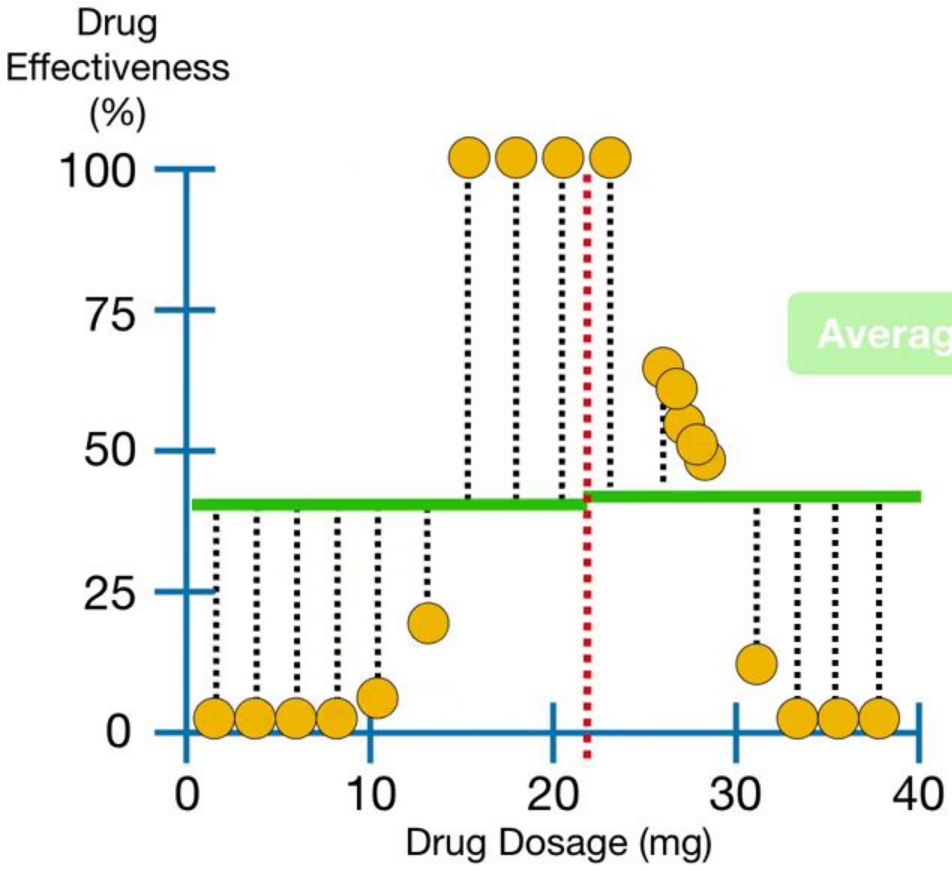
And we repeat until we have calculated the sum of squared residuals for all of the remaining thresholds.

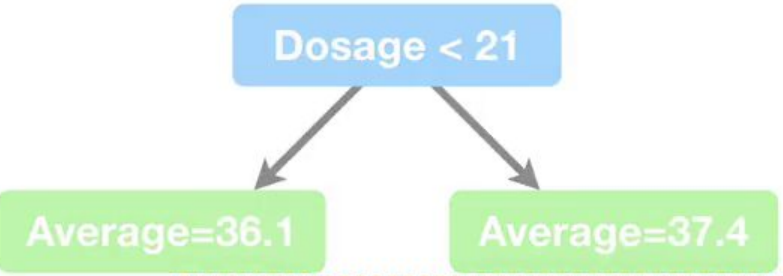


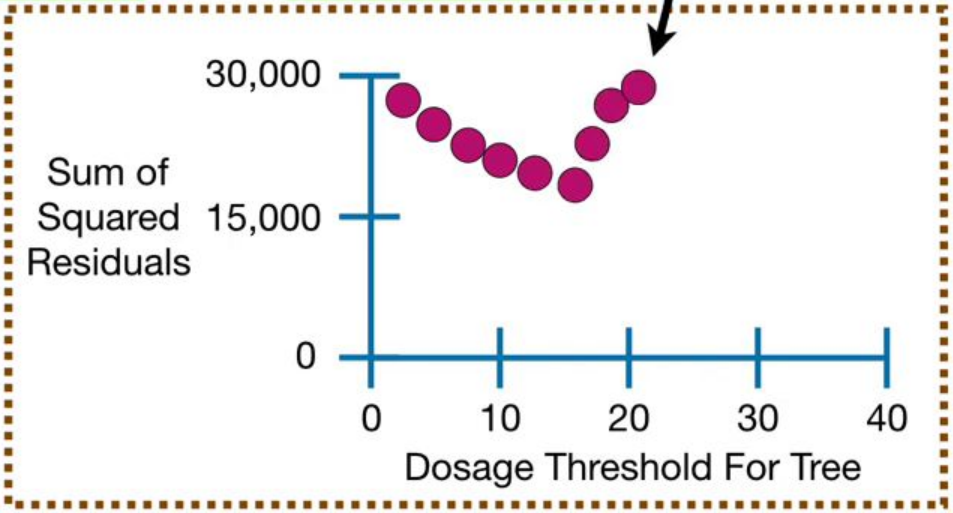




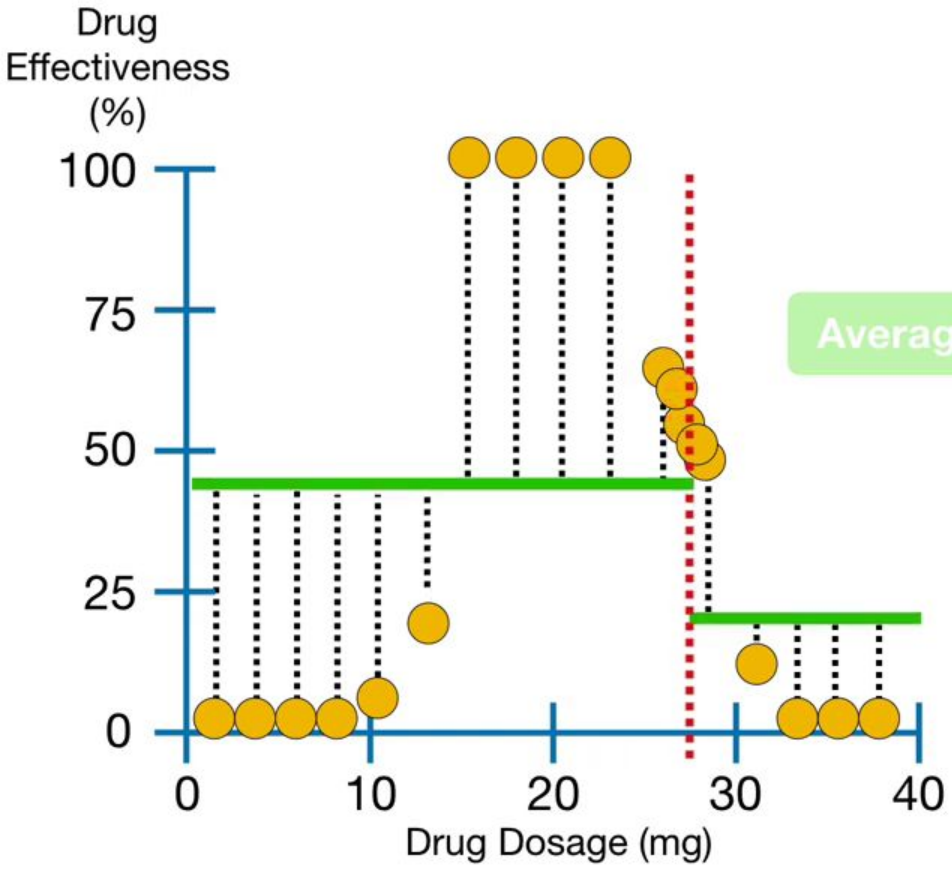
And we repeat until we have calculated the sum of squared residuals for all of the remaining thresholds.

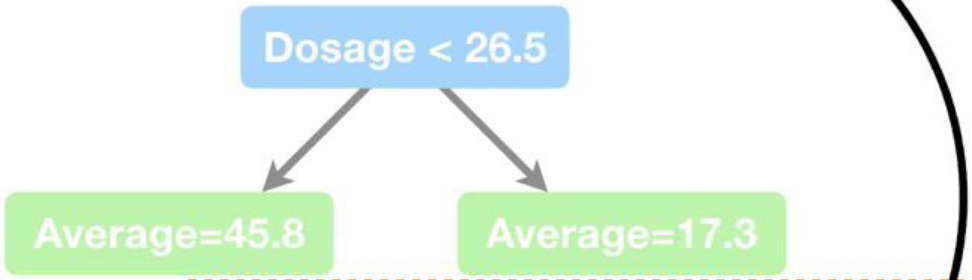


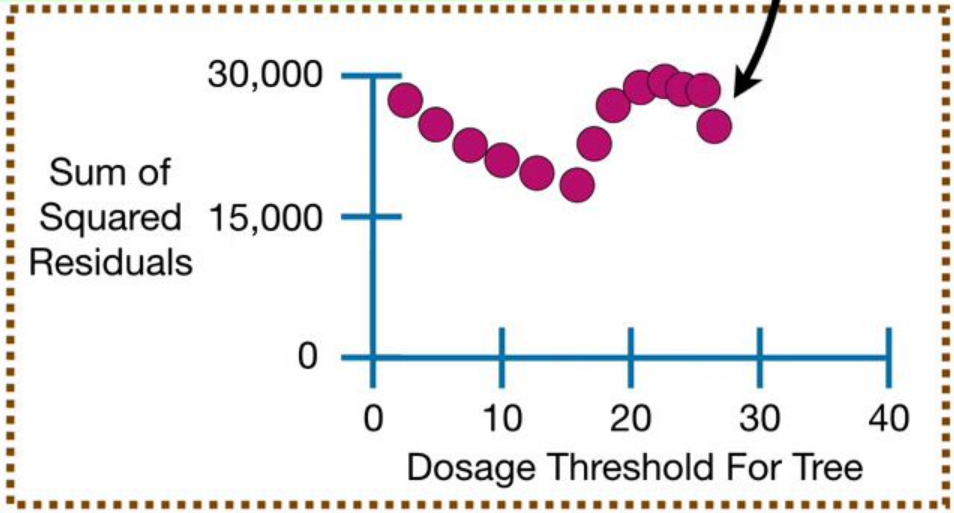




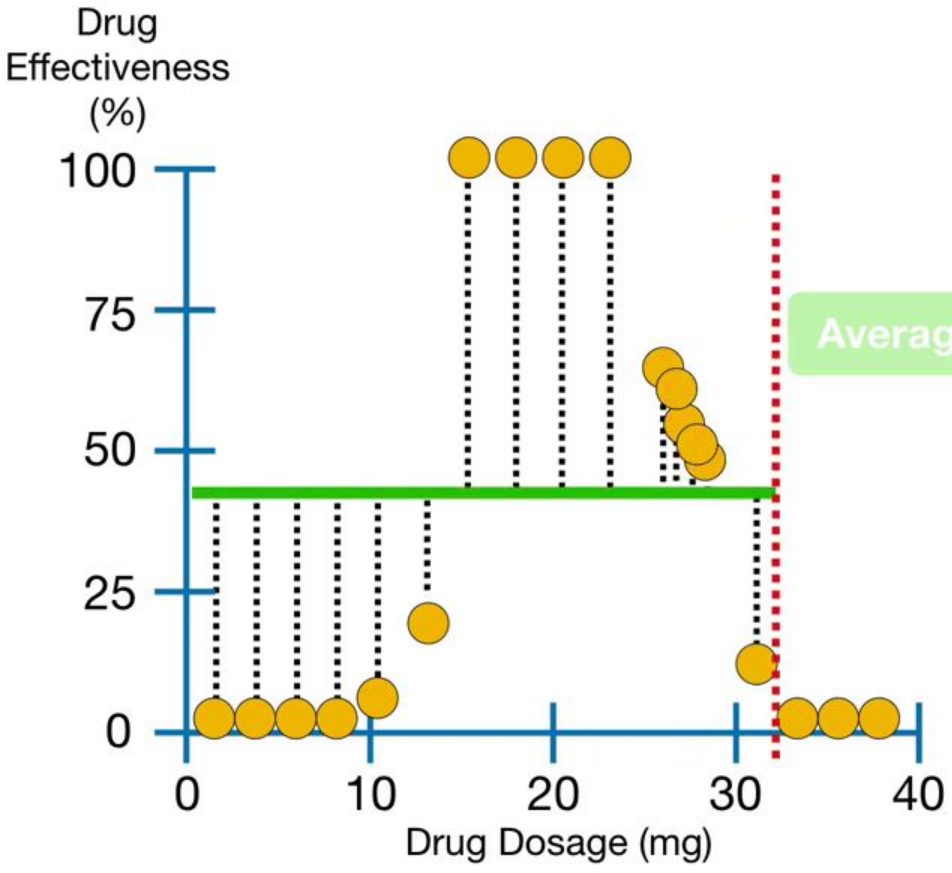
And we repeat until we have calculated the sum of squared residuals for all of the remaining thresholds.

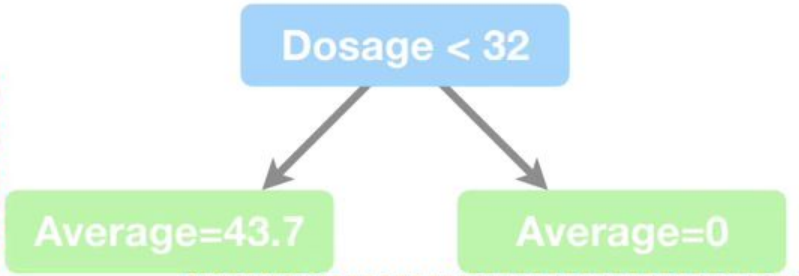


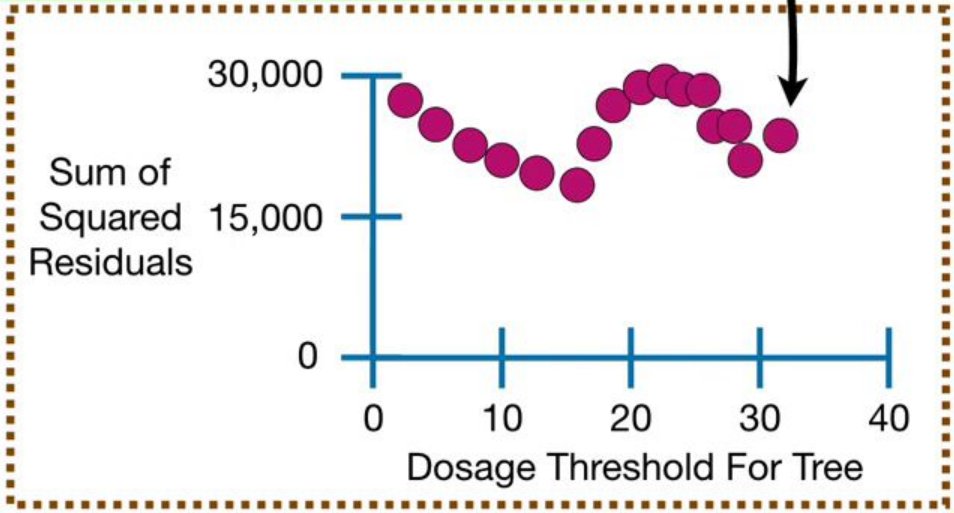




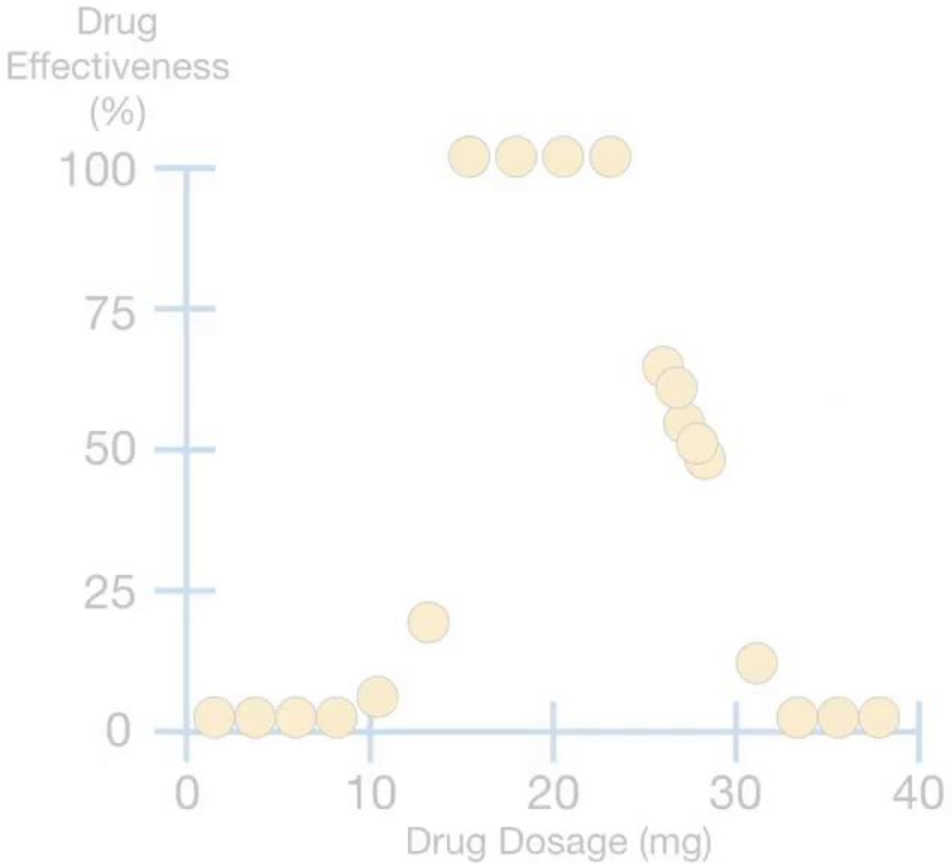
And we repeat until we have calculated the sum of squared residuals for all of the remaining thresholds.

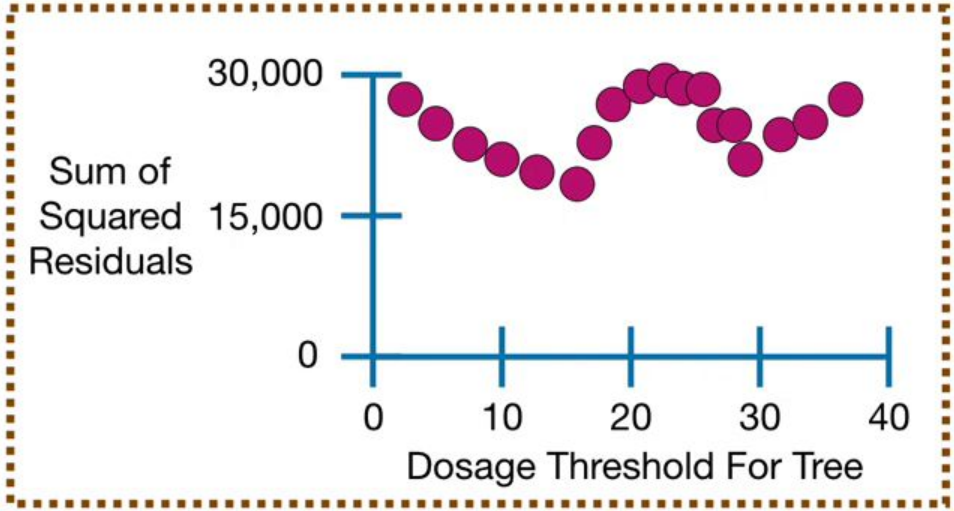


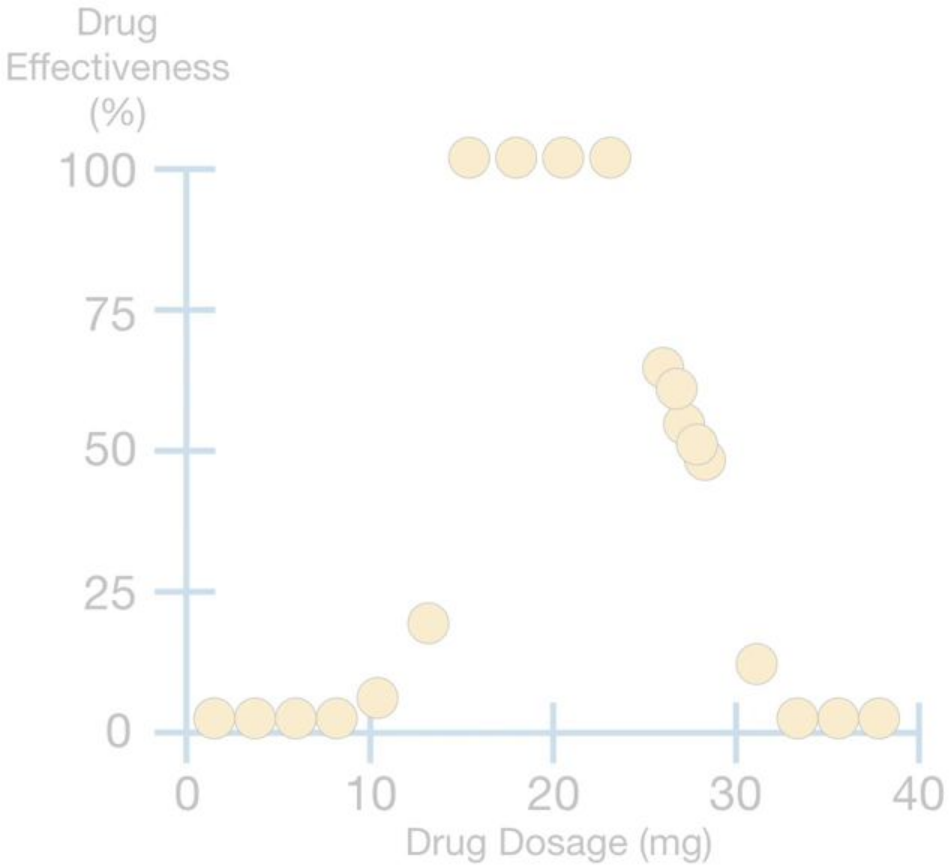




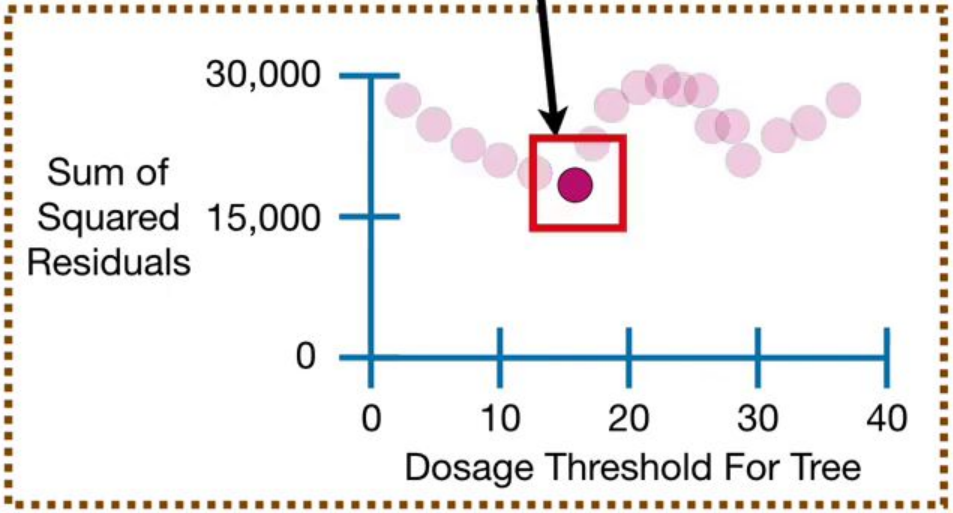
Now we can see the sum of squared residuals for all of the thresholds…



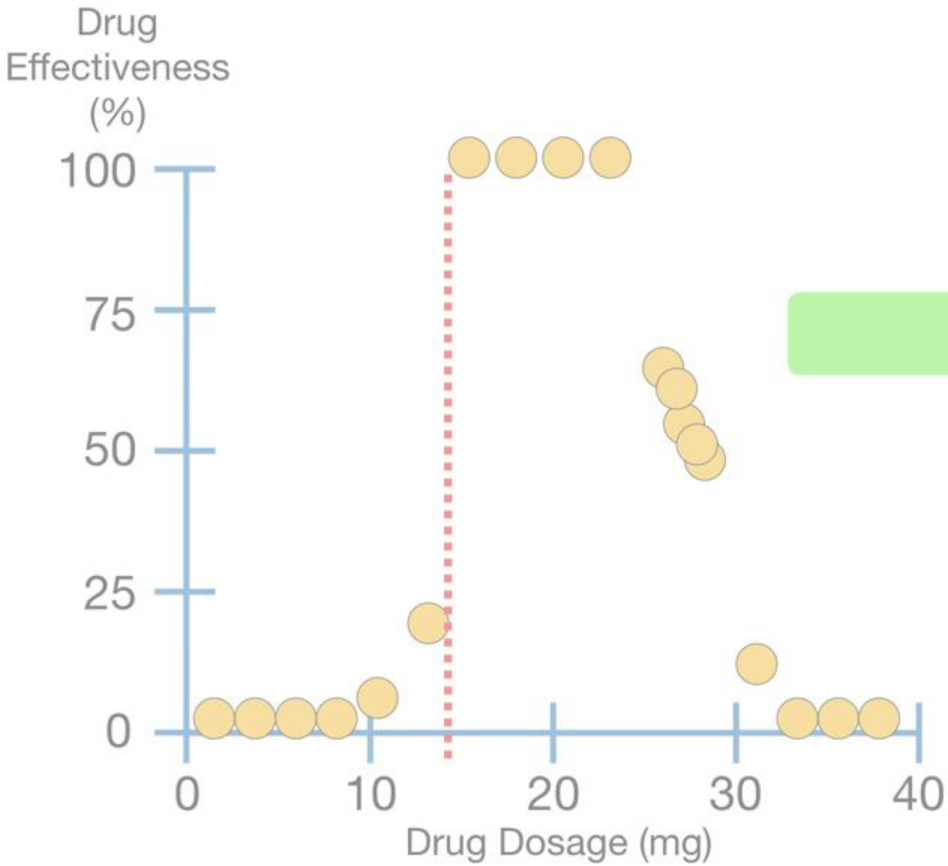


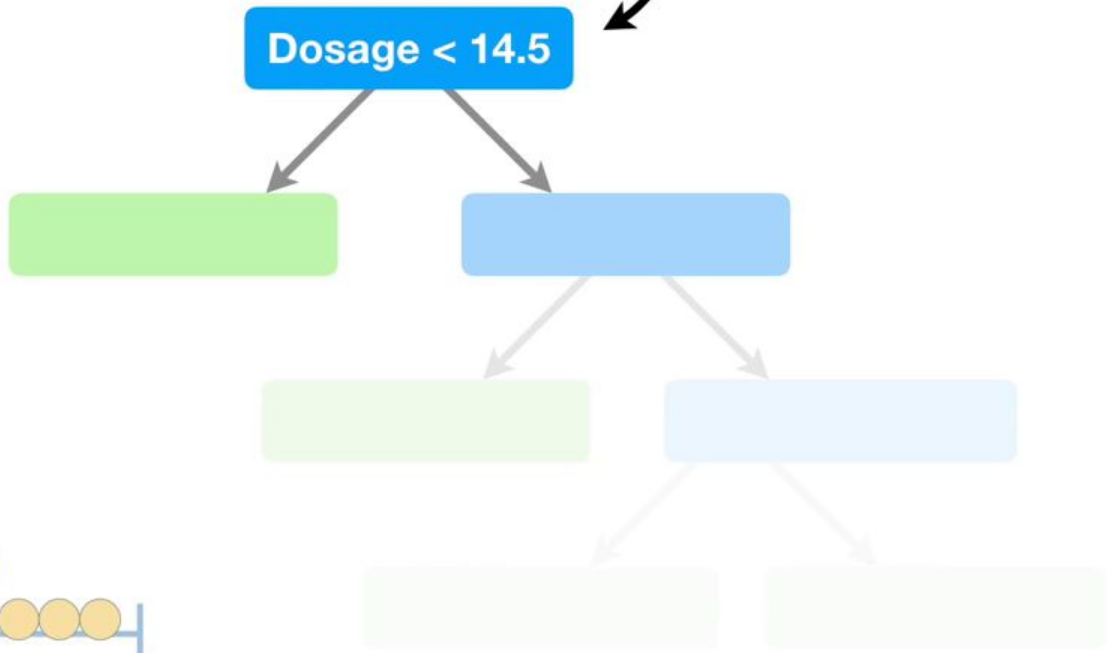


…and **Dosage < 14.5** had the smallest sum of squared residuals…

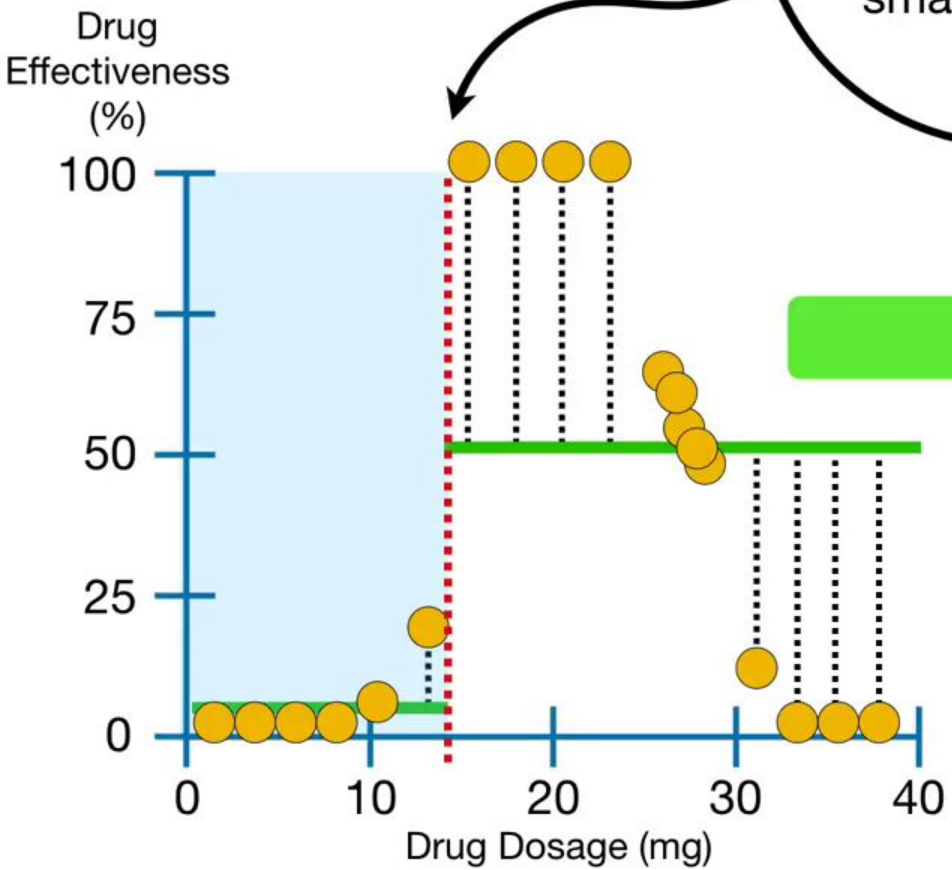


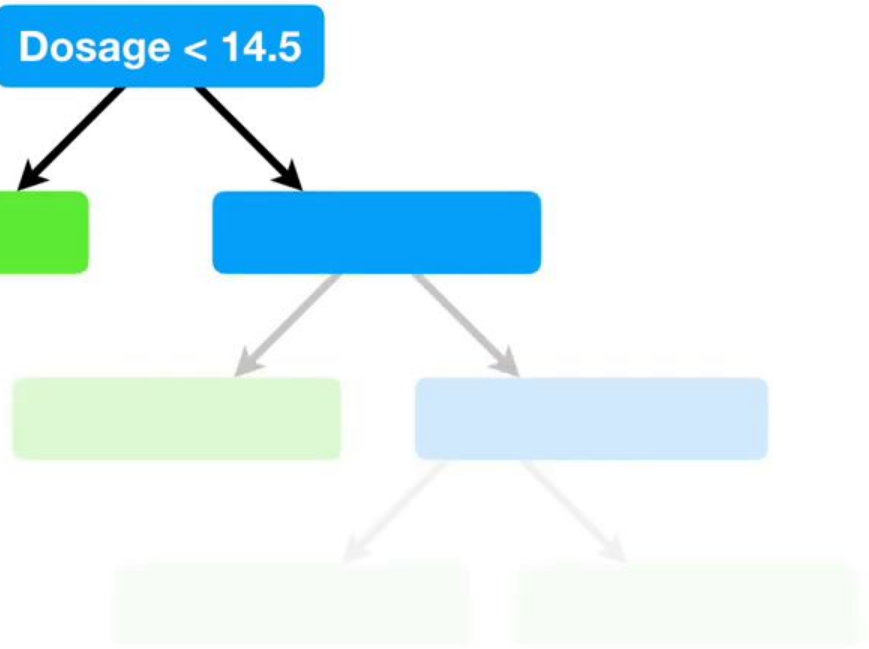
…so **Dosage < 14.5** will be root of the tree.



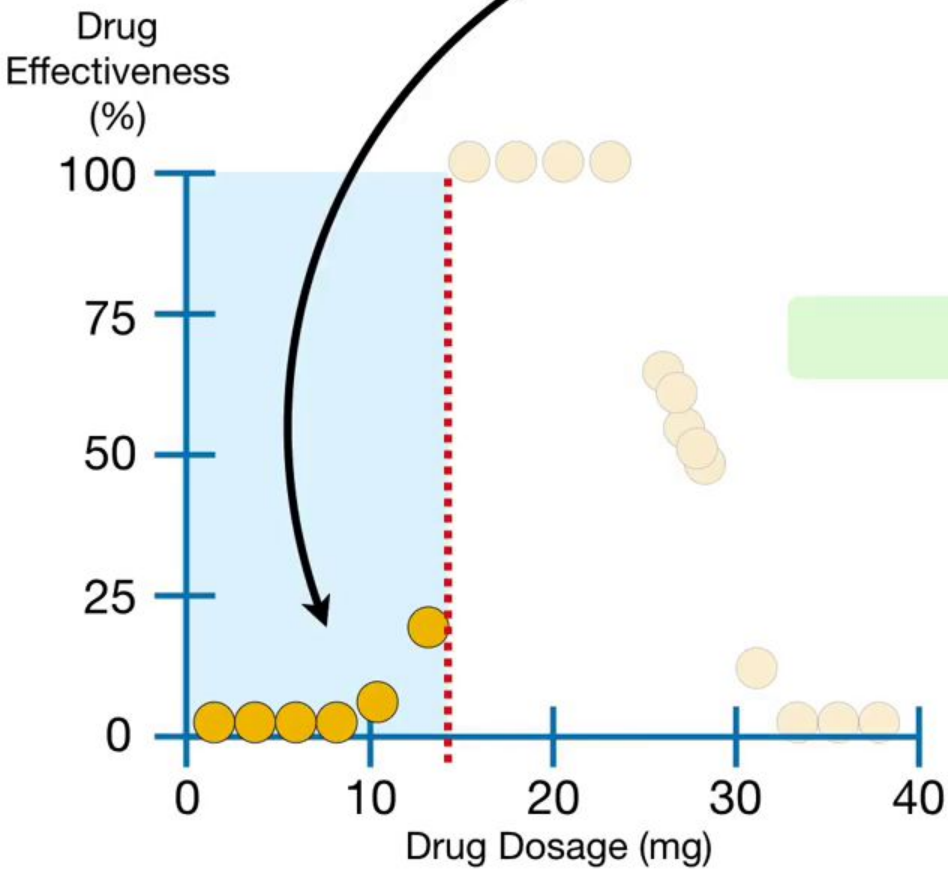


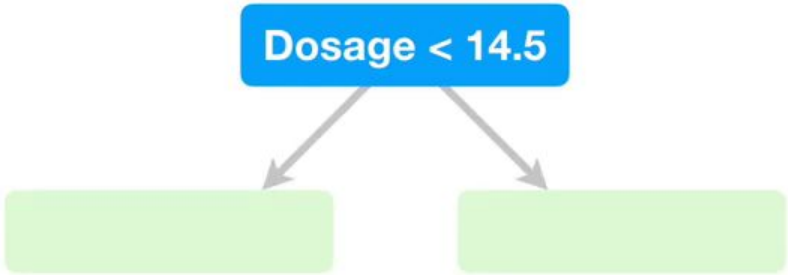
In summary, we split the data into two groups by finding the threshold that gave us the smallest sum of squared residuals.

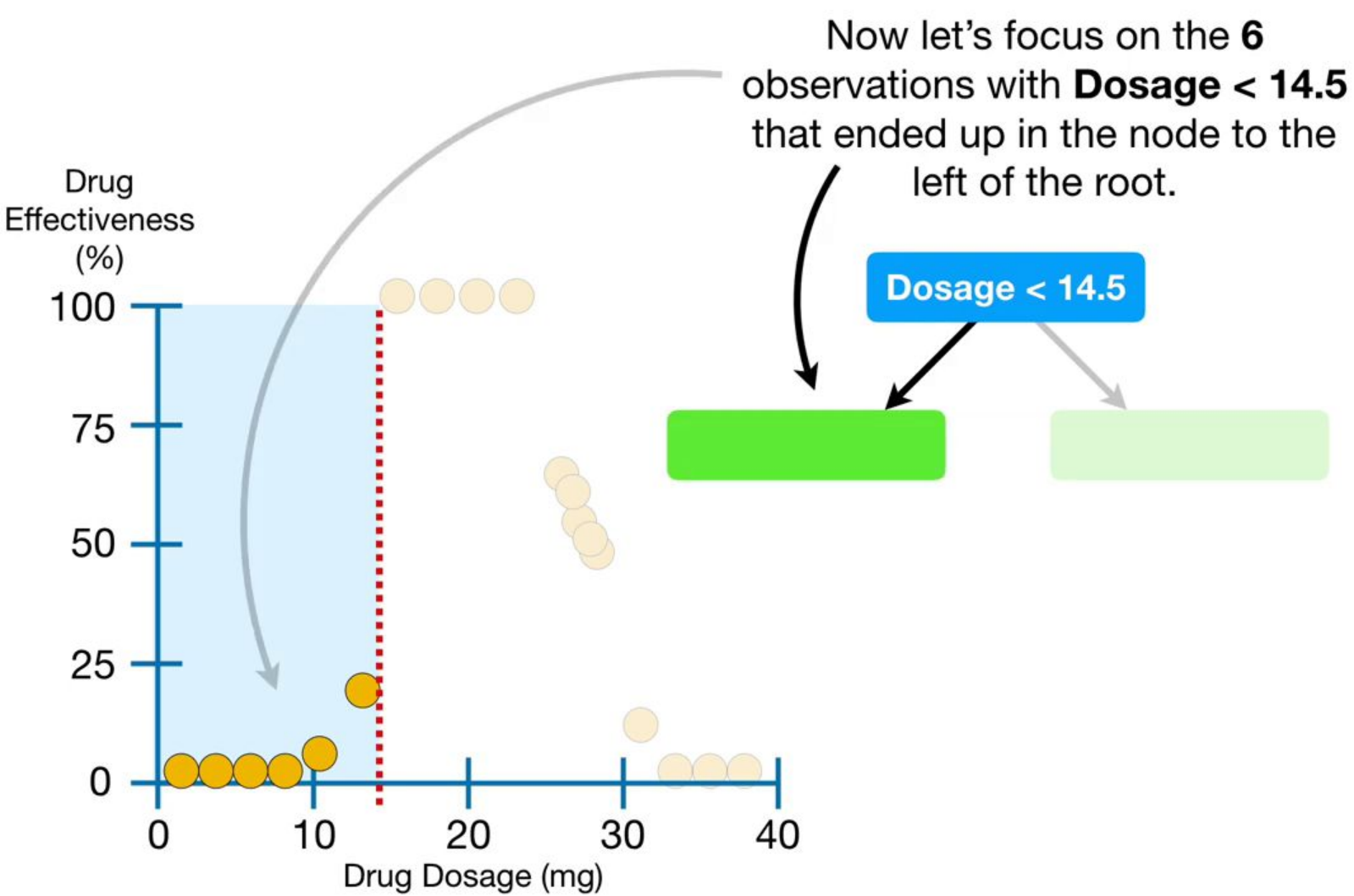




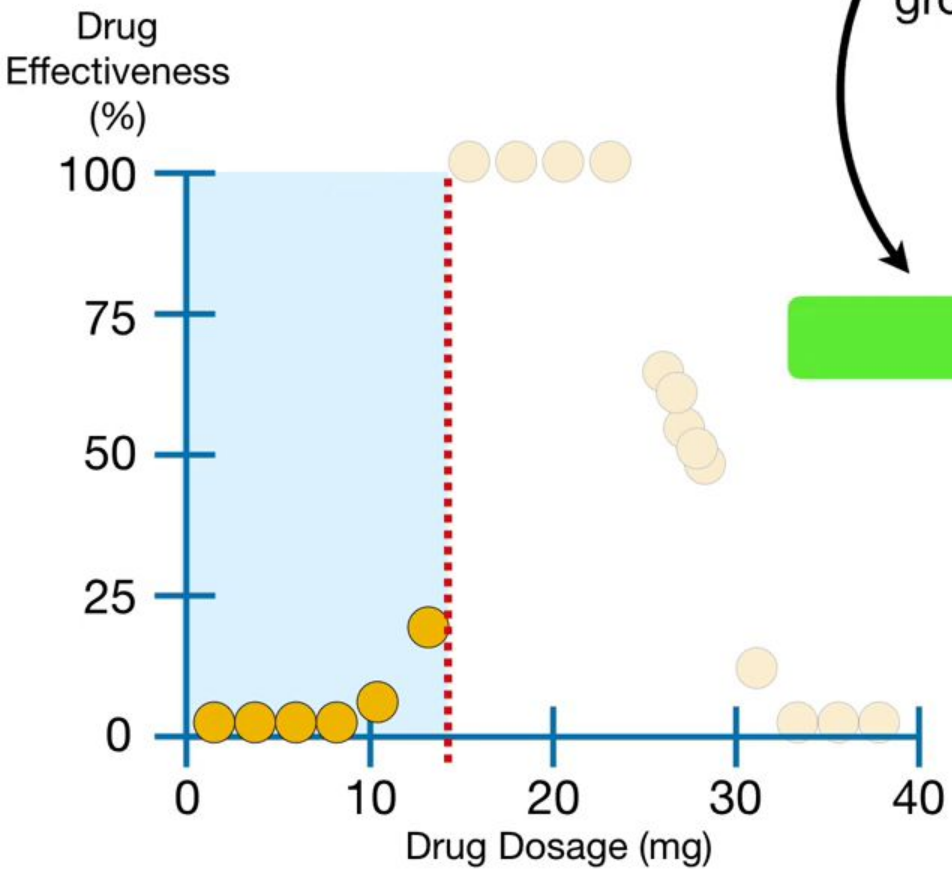
Now let's focus on the **6** observations with **Dosage < 14.5** that ended up in the node to the left of the root.

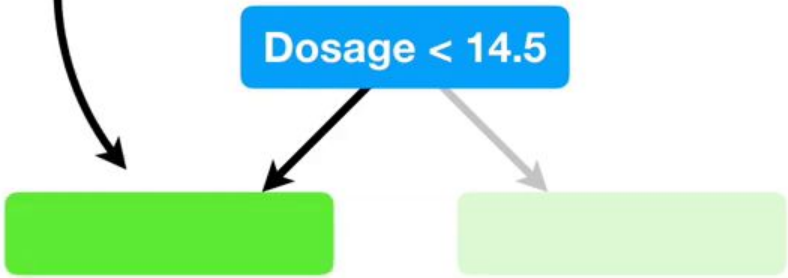


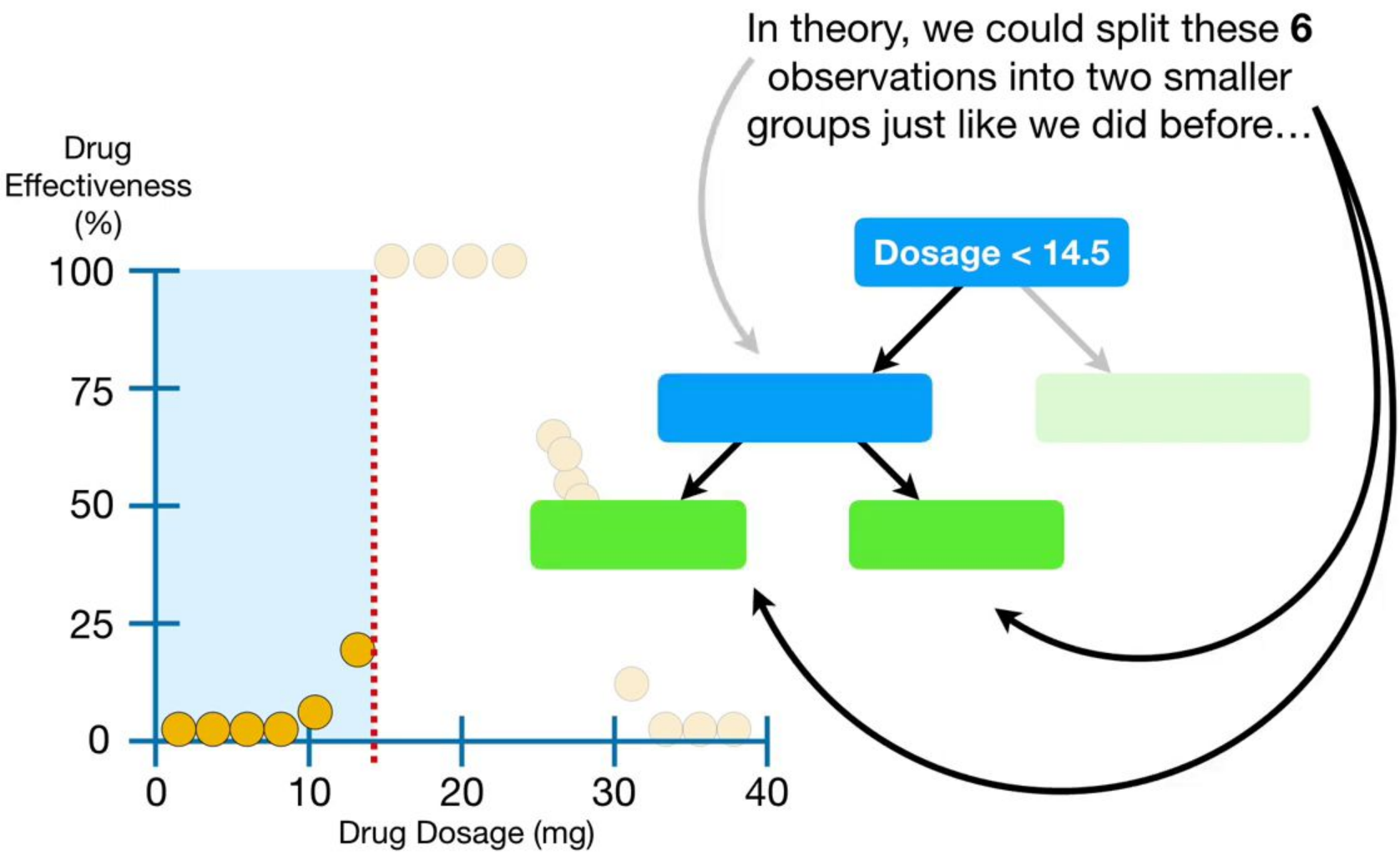


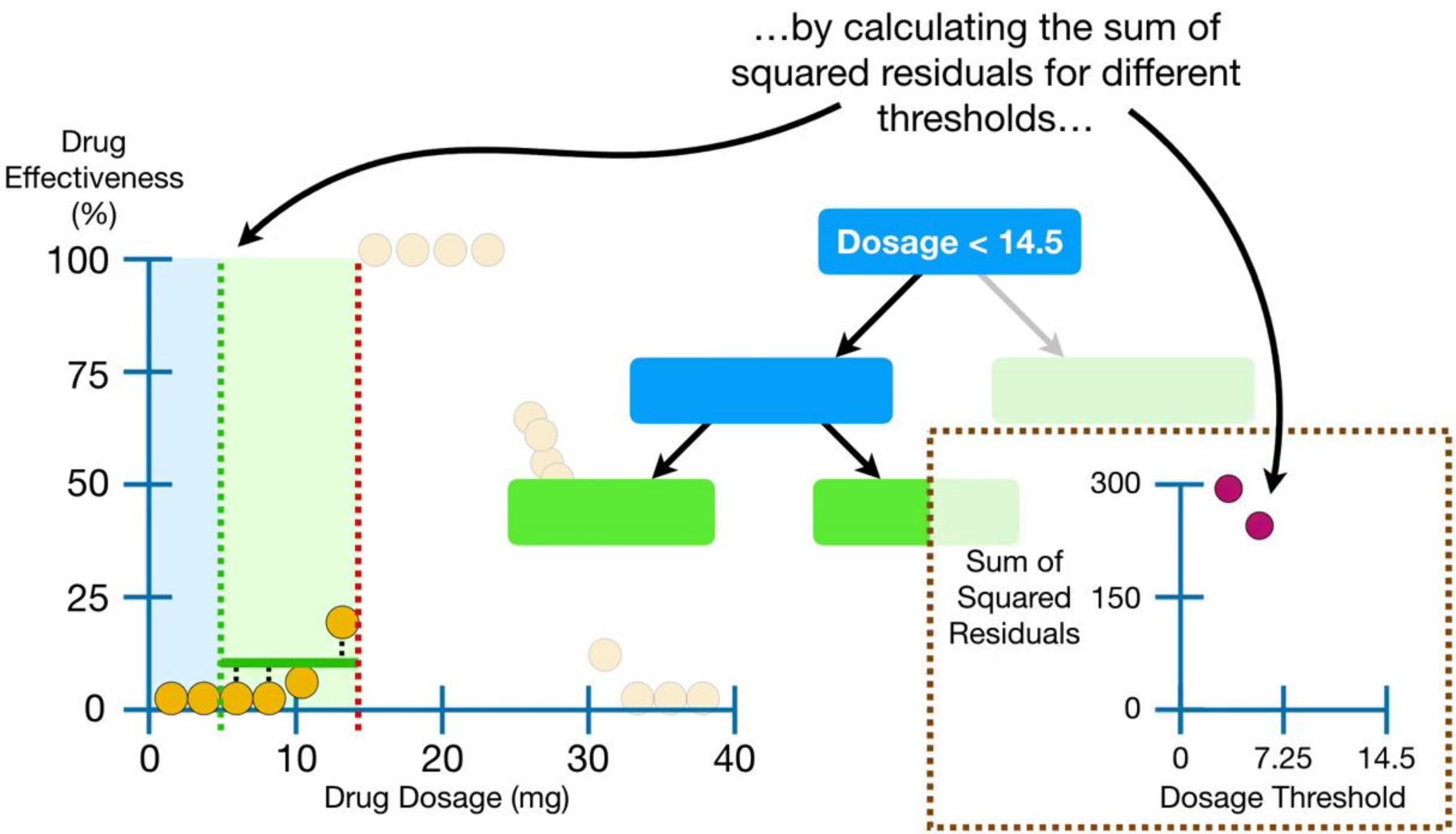

Now let's focus on the **6** observations with **Dosage < 14.5** that ended up in the node to the left of the root.
Dosage < 14.5
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)

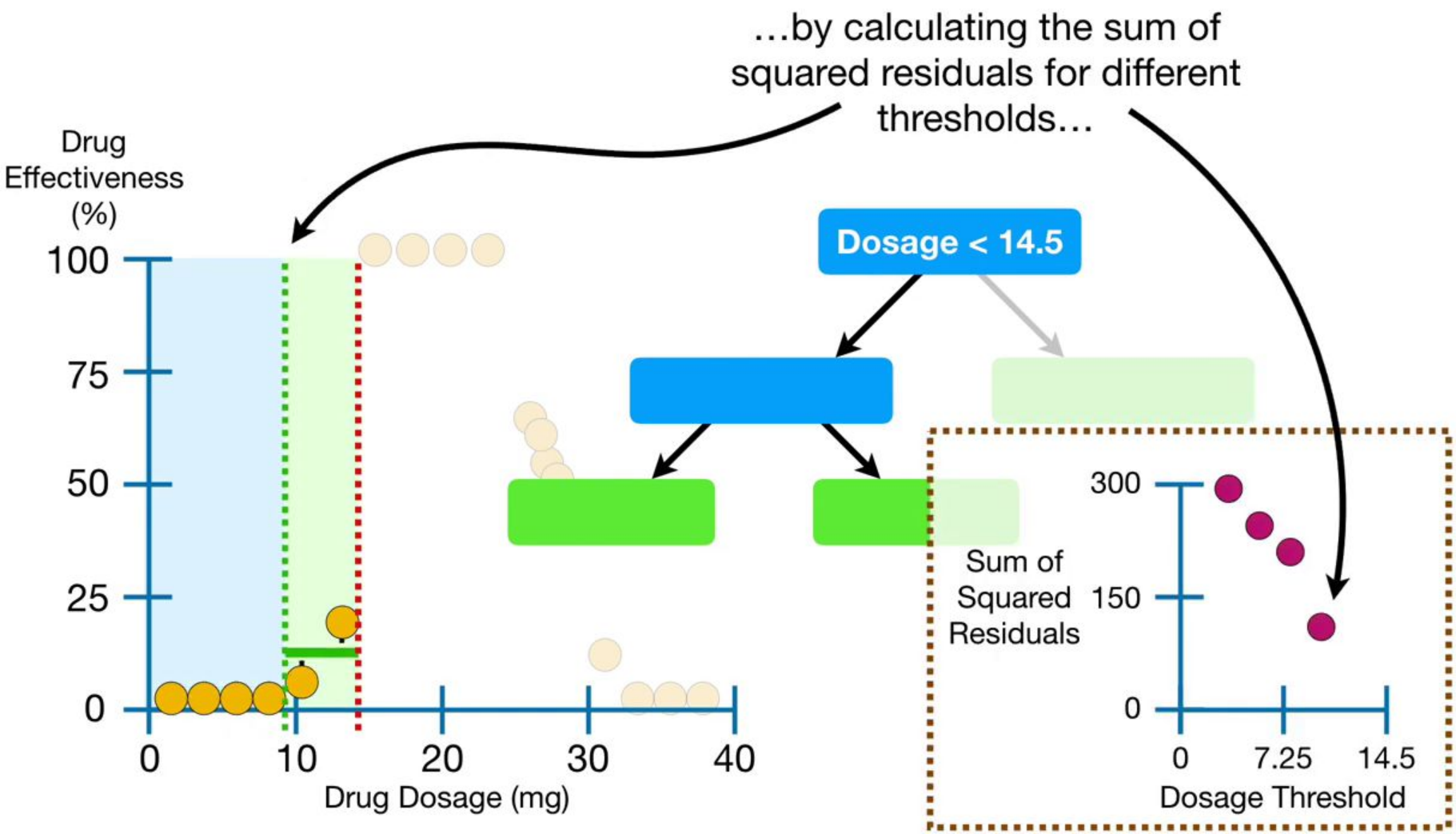
In theory, we could split these **6** observations into two smaller groups just like we did before…

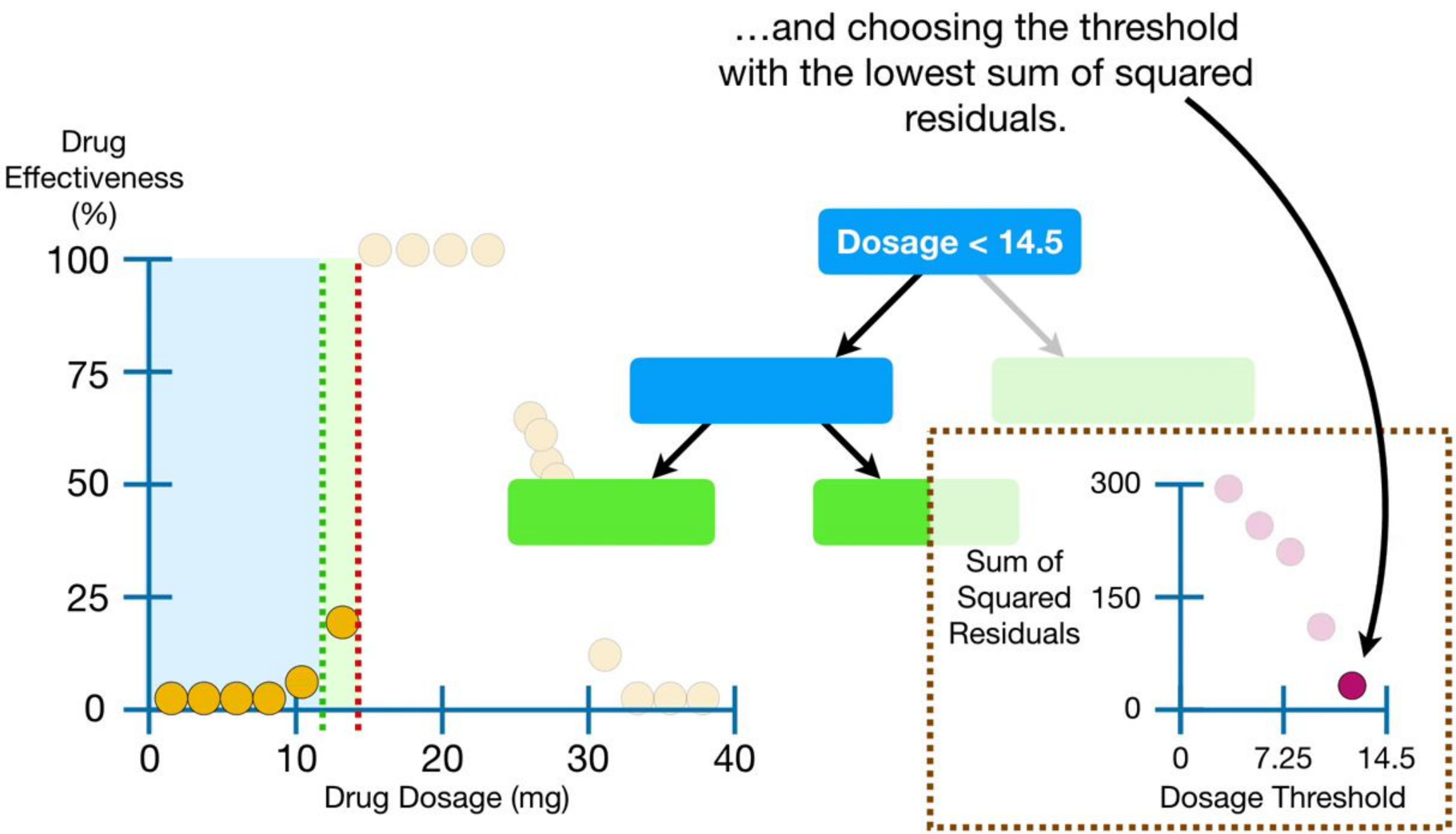


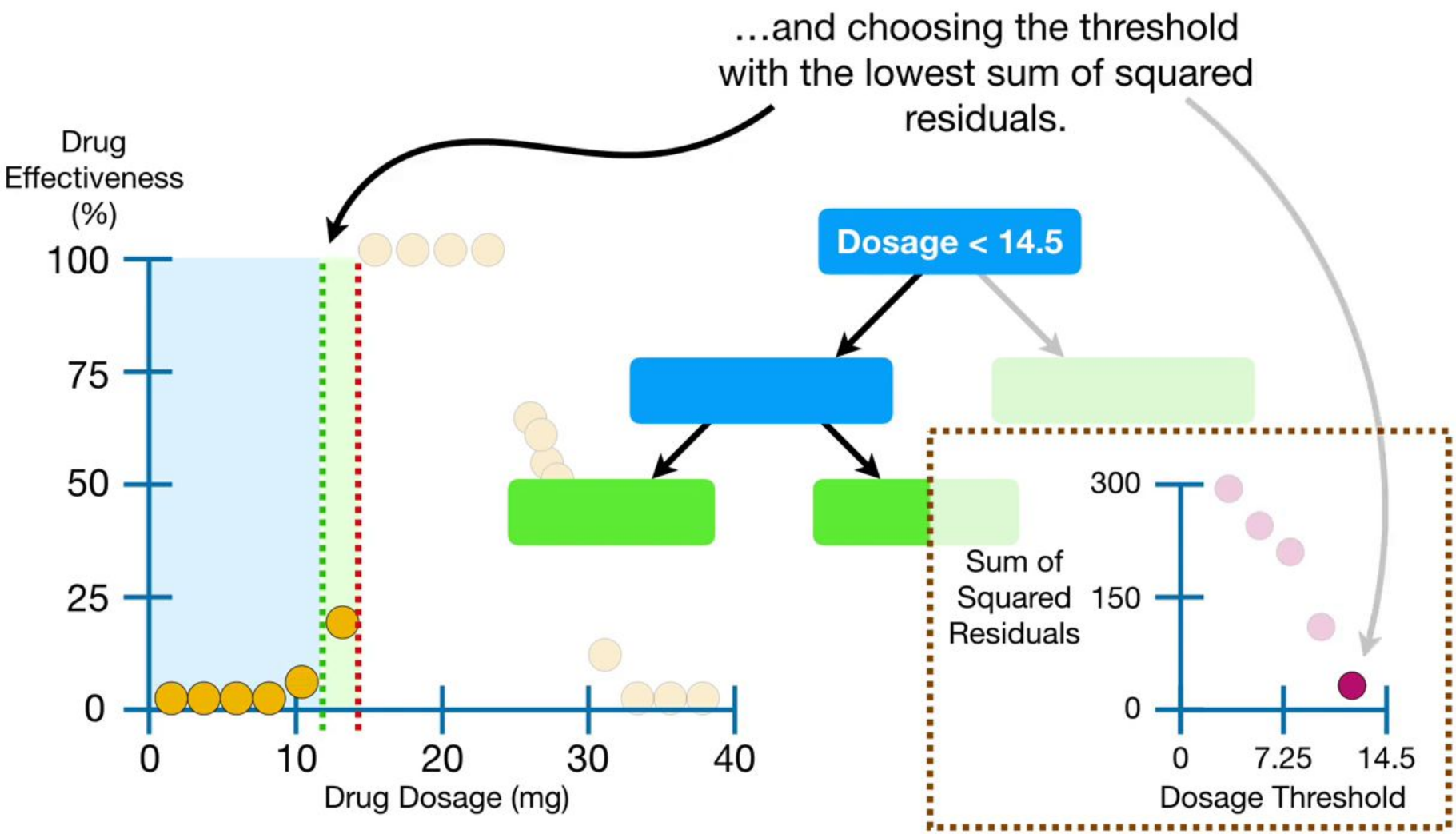



In theory, we could split these **6** observations into two smaller groups just like we did before…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5

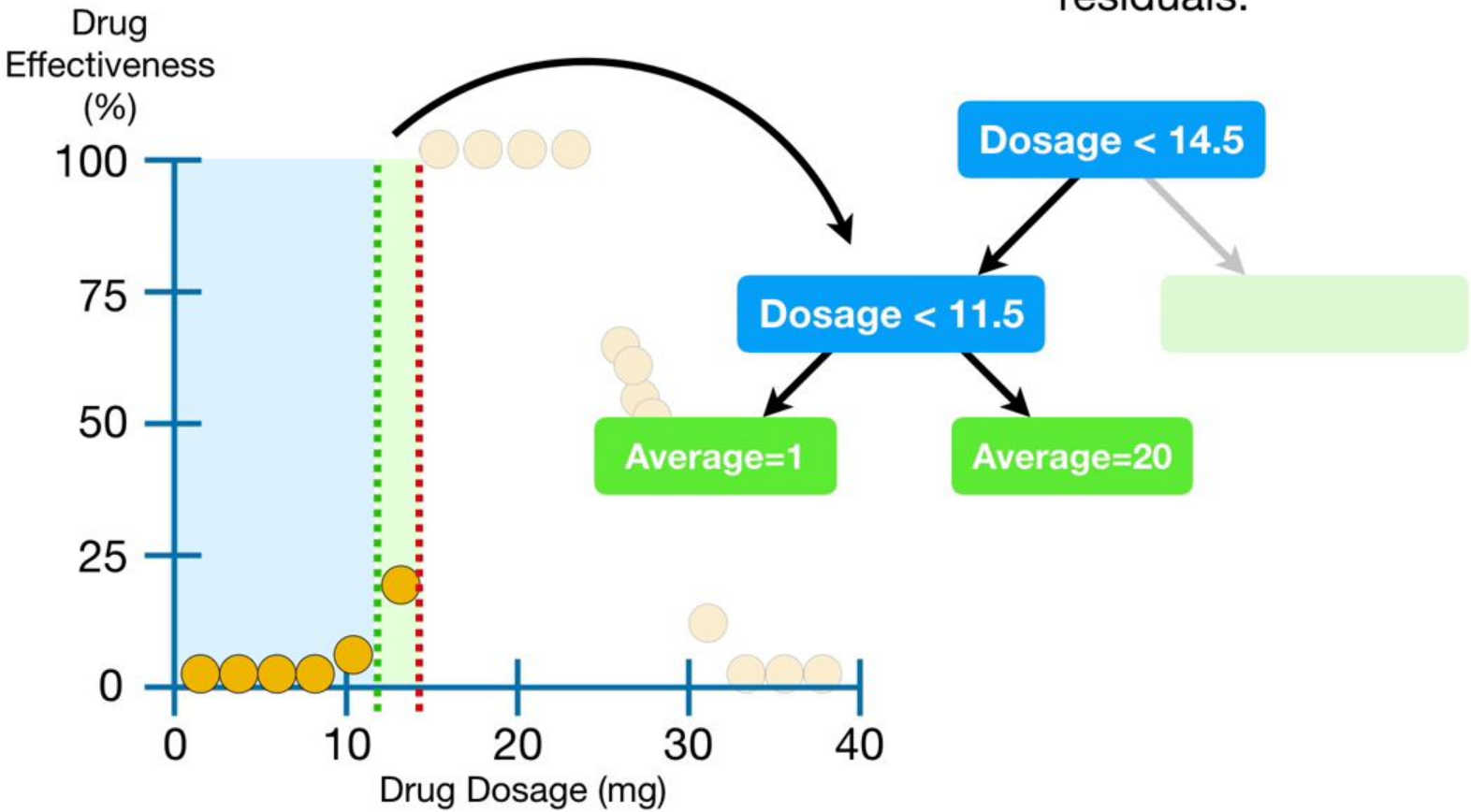

…by calculating the sum of squared residuals for different thresholds…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Sum of Squared Residuals
300
150
0
0
7.25
14.5
Dosage Threshold

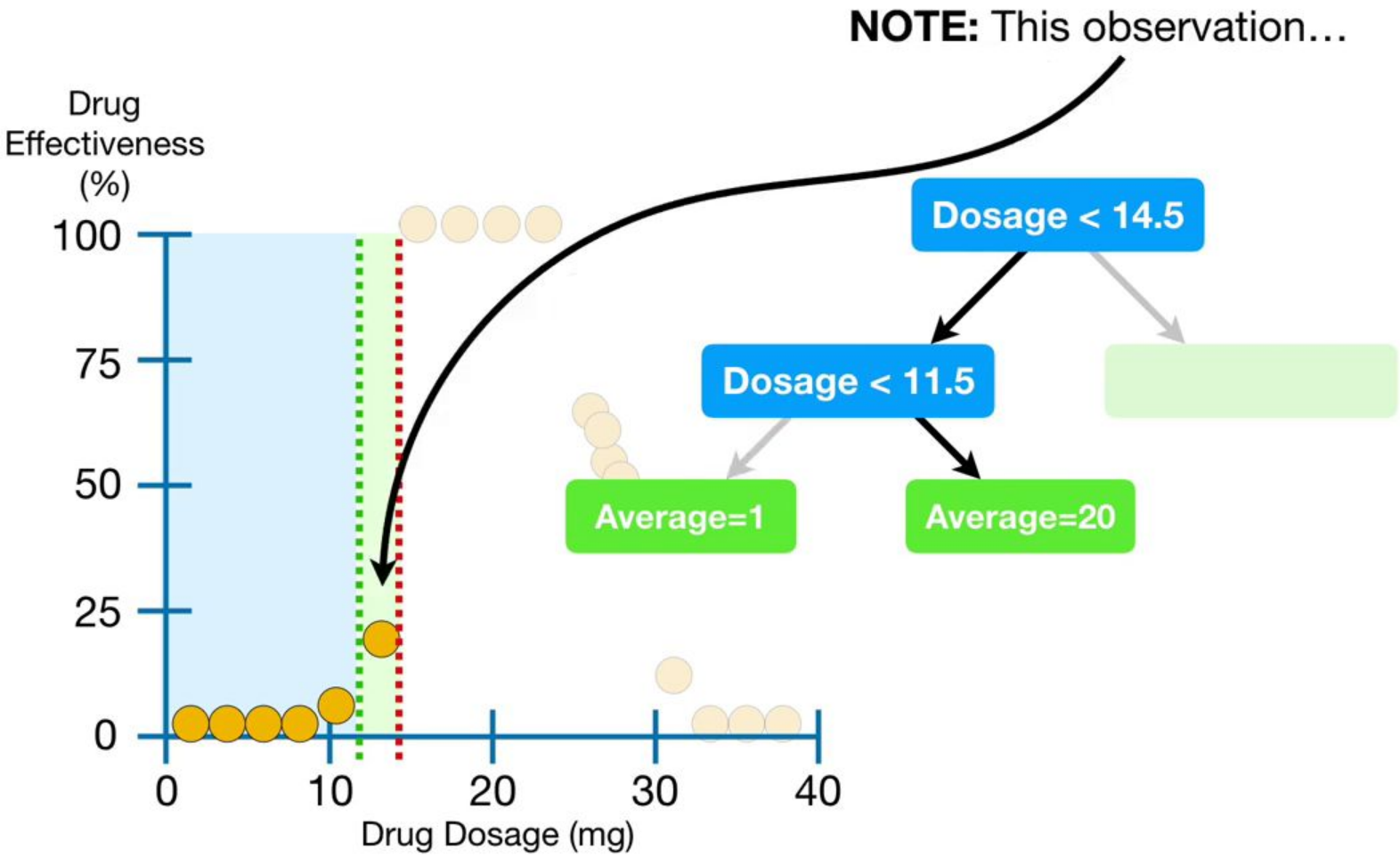

…by calculating the sum of squared residuals for different thresholds…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Sum of Squared Residuals
300
150
0
0
7.25
14.5
Dosage Threshold


…and choosing the threshold with the lowest sum of squared residuals.
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Sum of Squared Residuals
300
150
0
0
7.25
14.5
Dosage Threshold

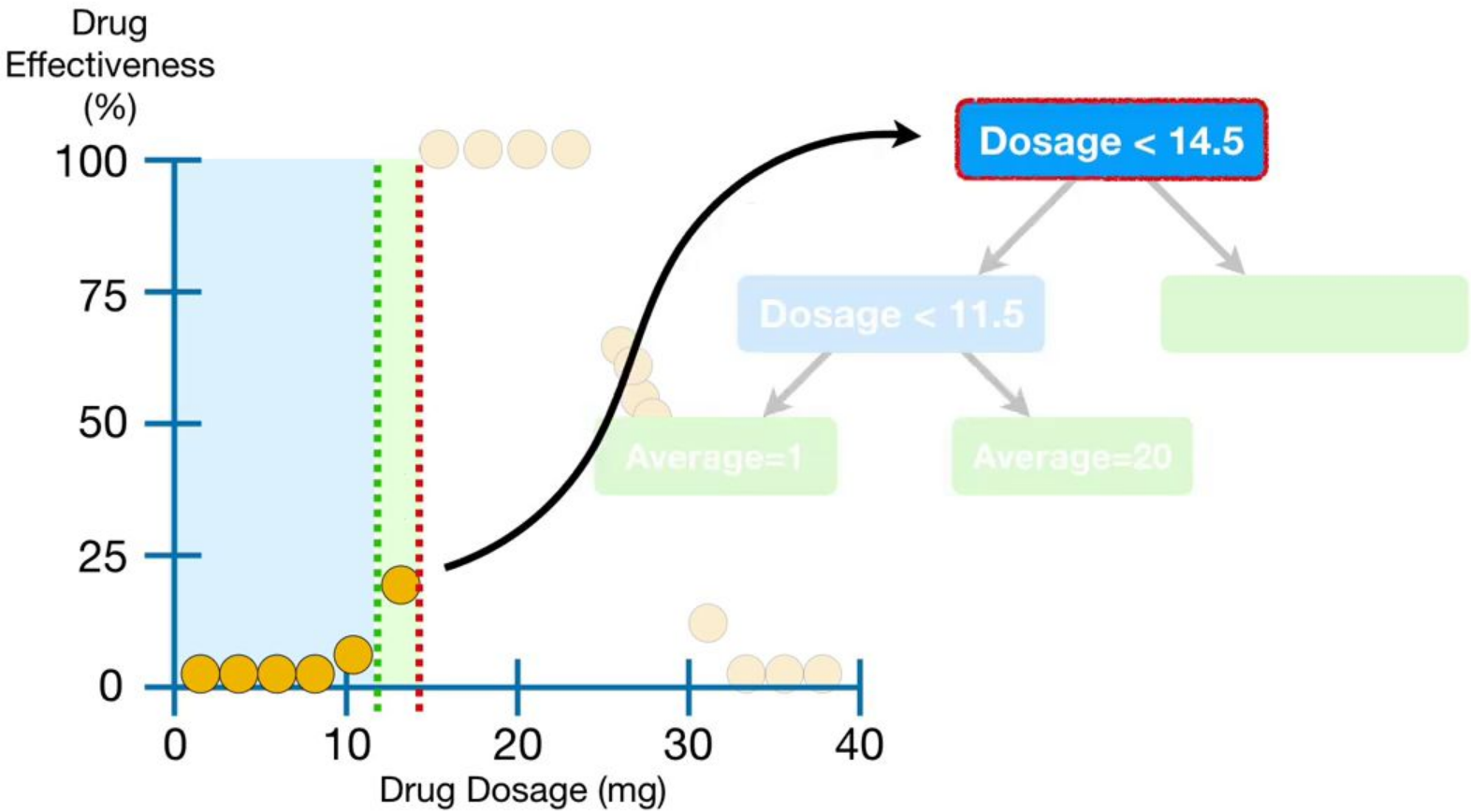

…and choosing the threshold with the lowest sum of squared residuals.
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Sum of Squared Residuals
300
150
0
0
7.25
14.5
Dosage Threshold

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

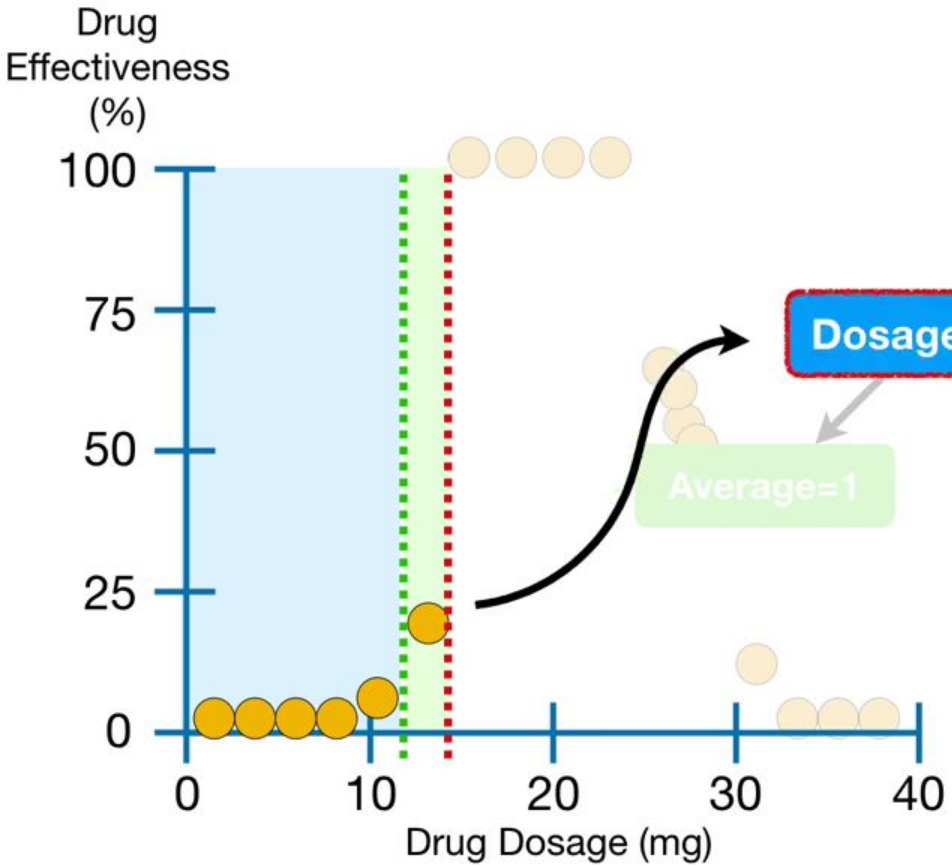


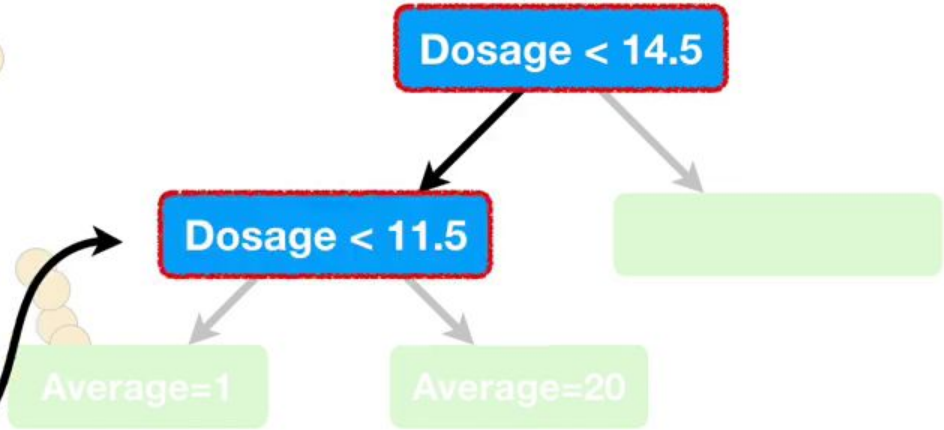

NOTE: This observation…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Dosage < 11.5
Average=1
Average=20

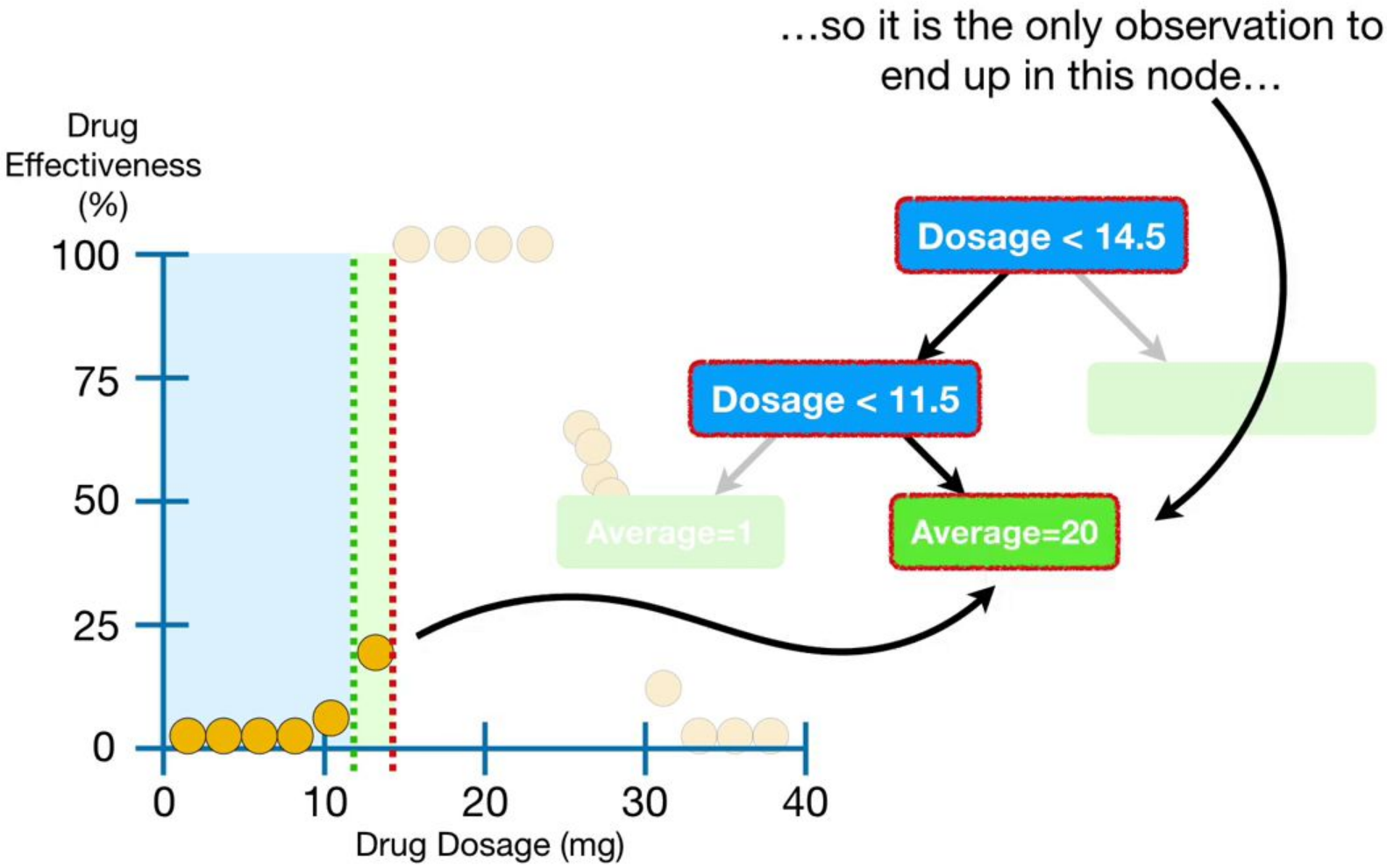
…has **Dosage < 14.5**…



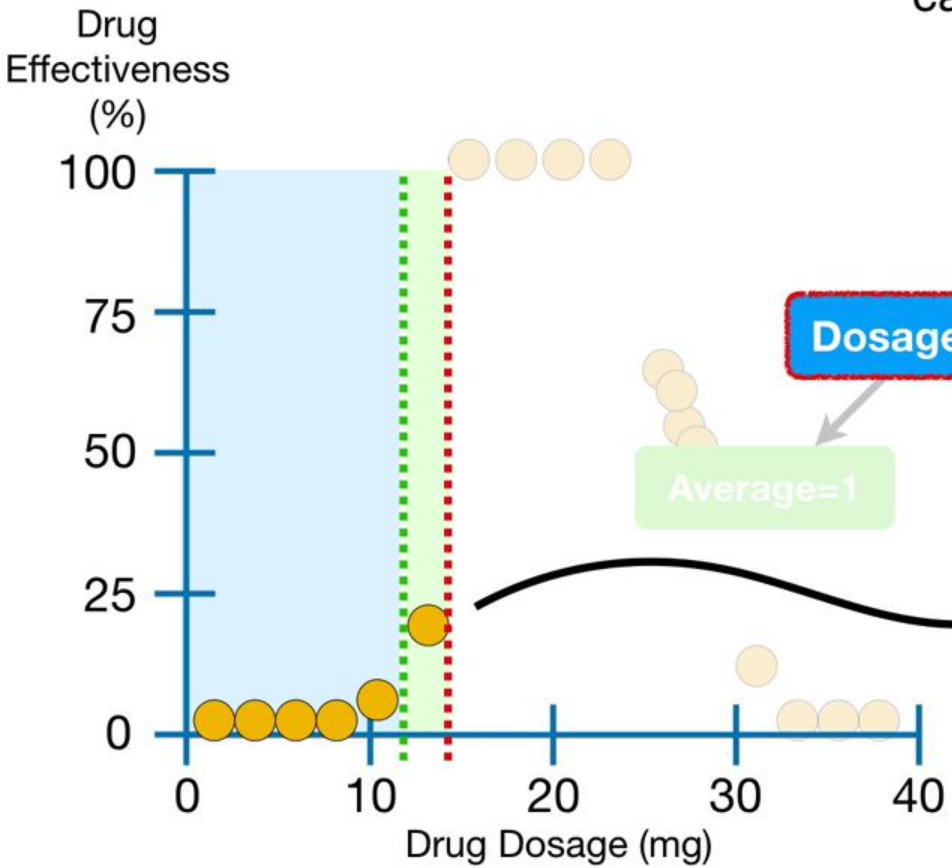
…and does *not* have
Dosage < 11.5…

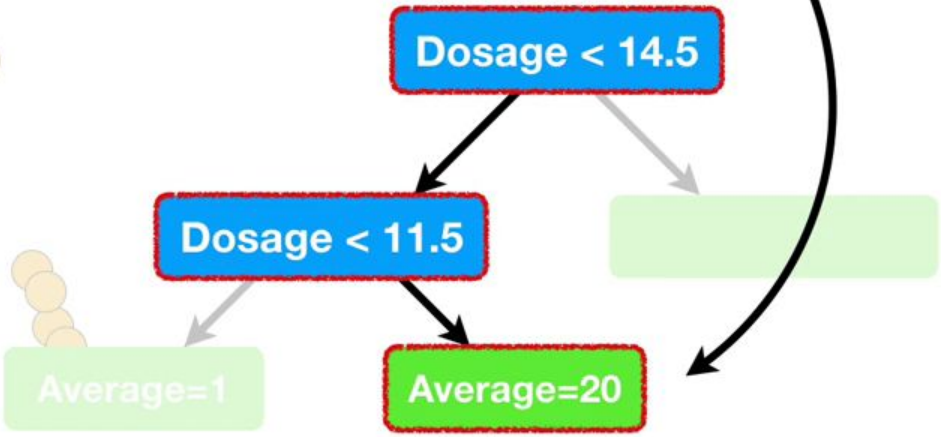




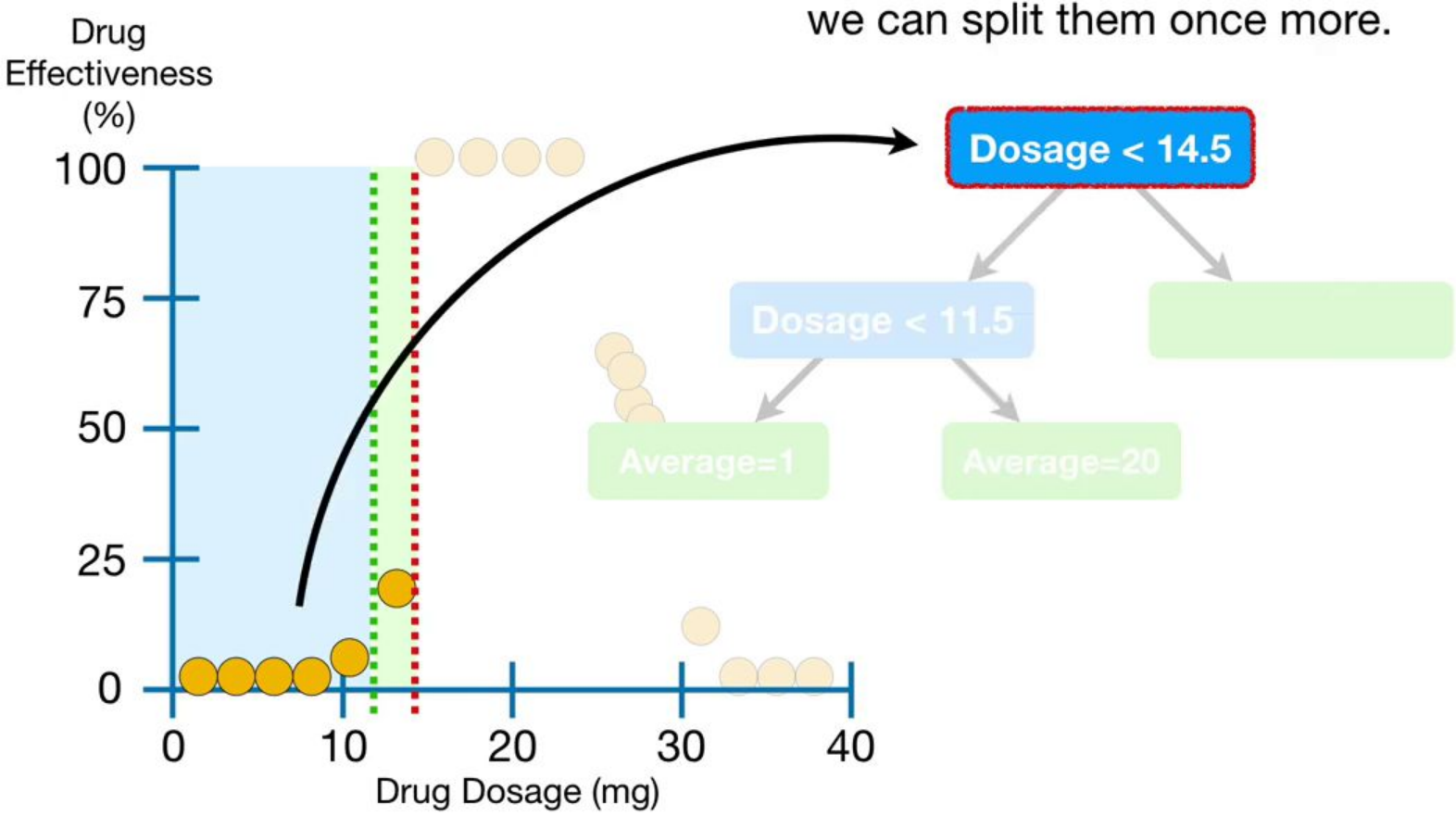

…so it is the only observation to end up in this node…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Dosage < 11.5
Average=1
Average=20

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

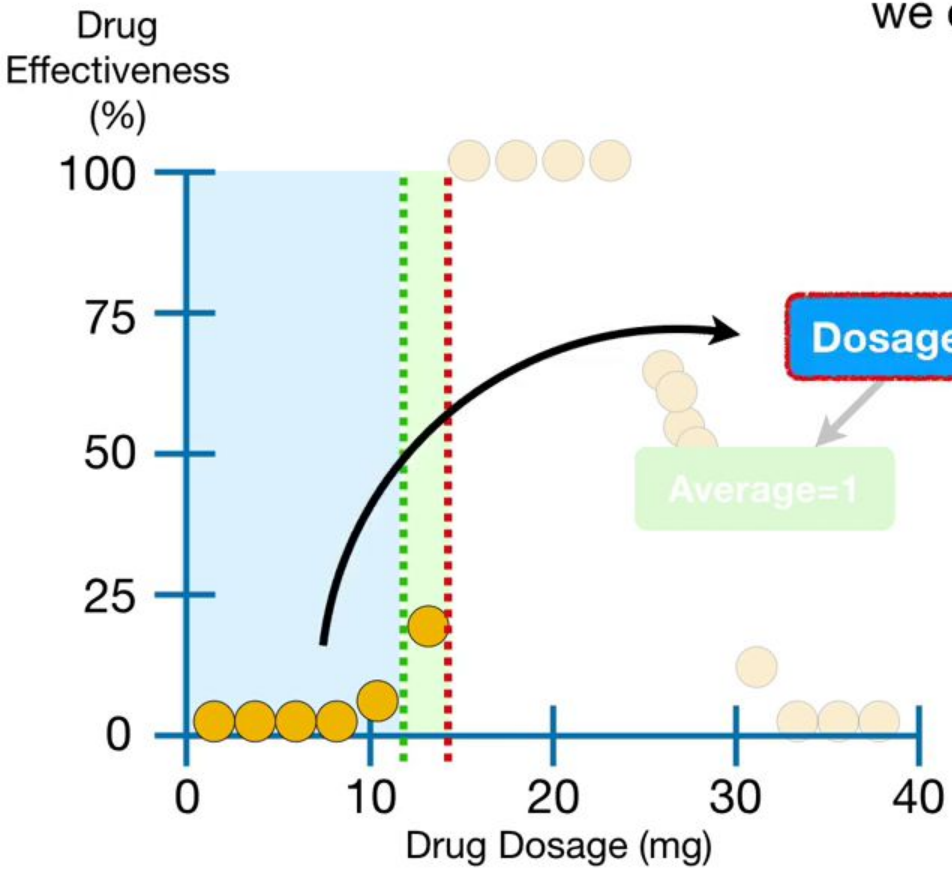


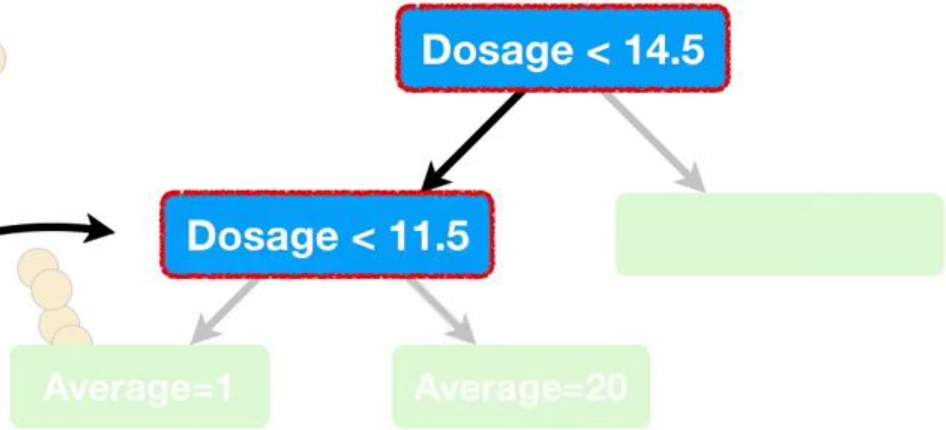


However, since the remaining **5** observations go to the other node, we can split them once more.

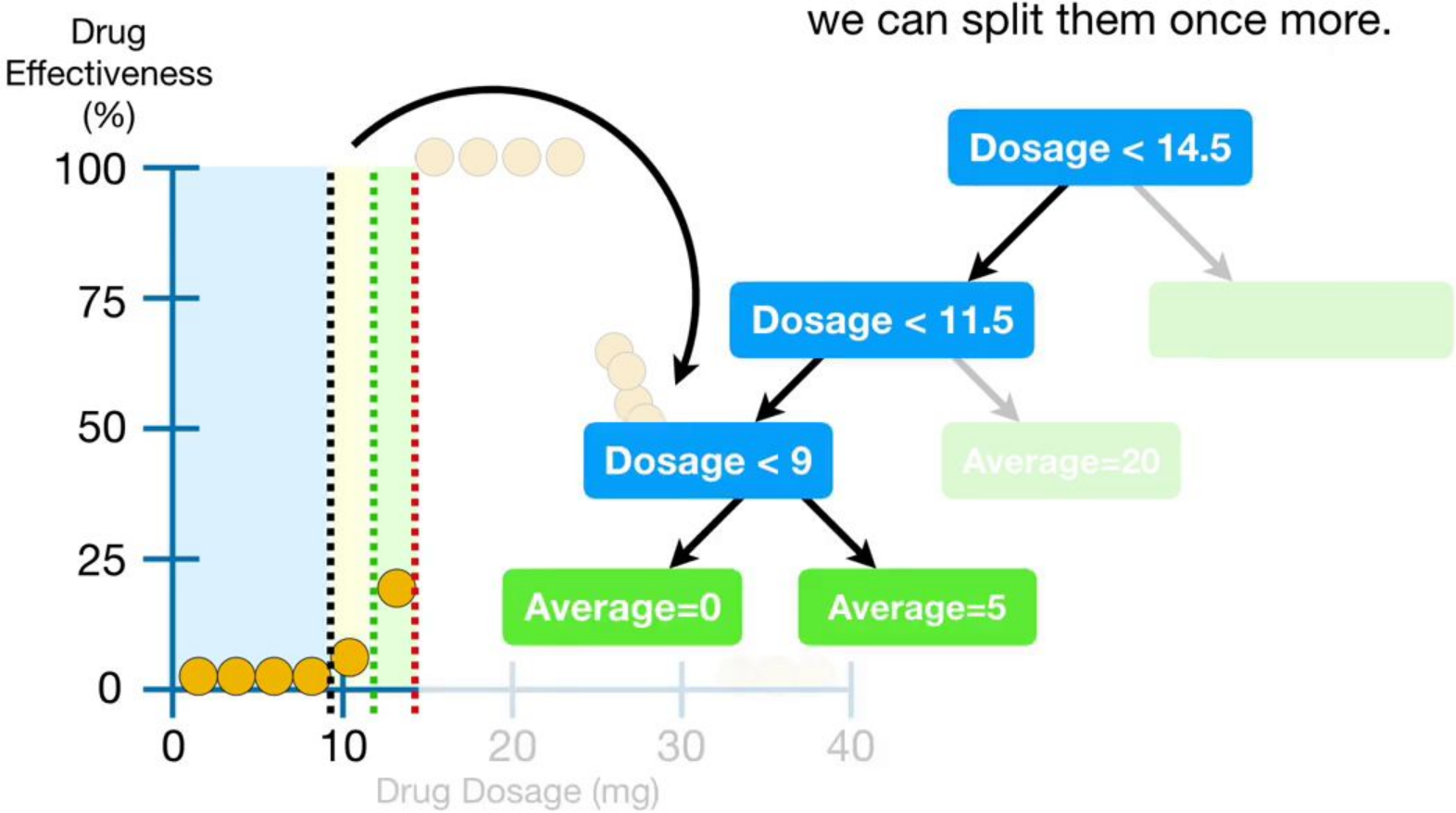


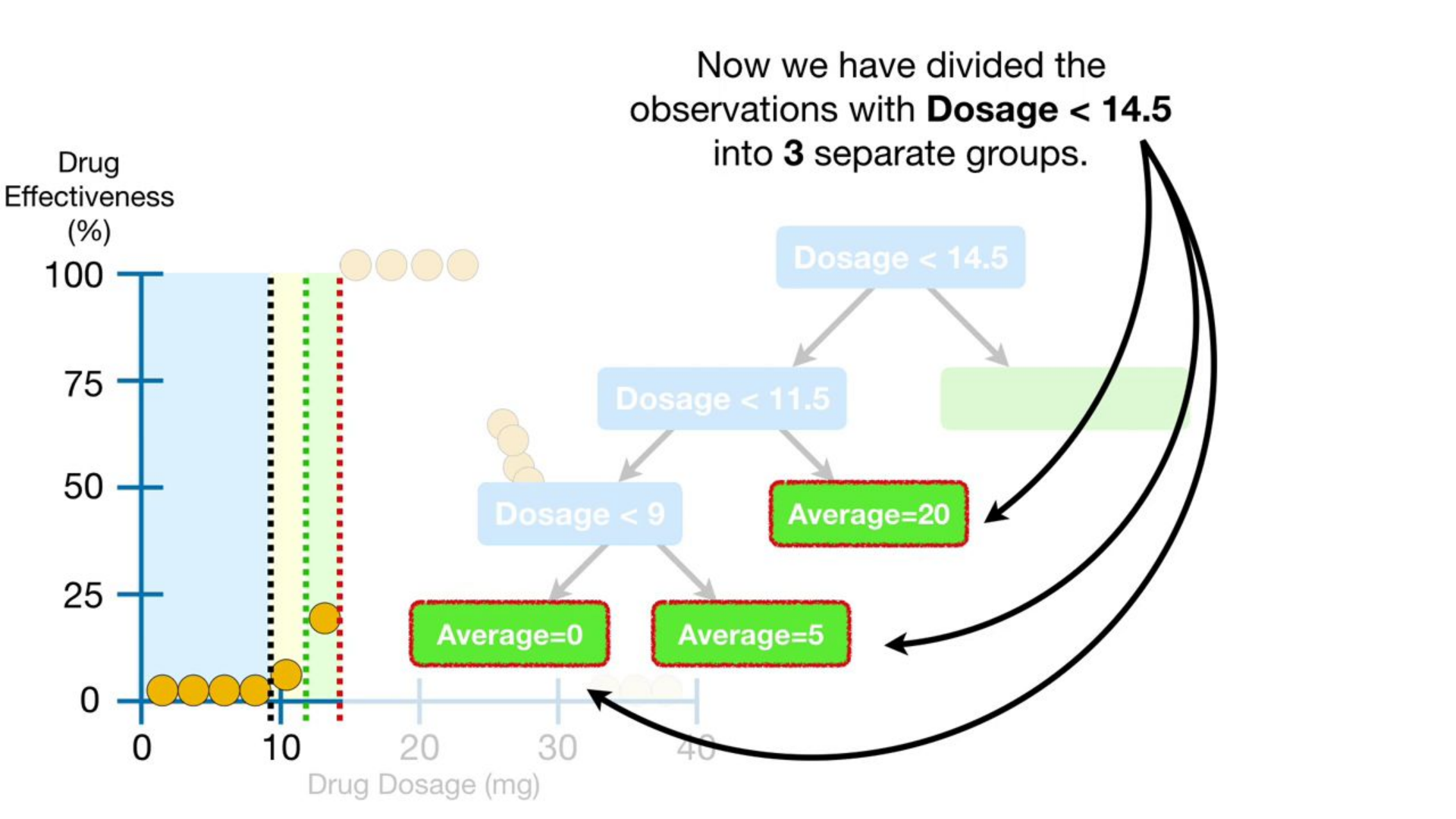
However, since the remaining **5** observations go to the other node, we can split them once more.

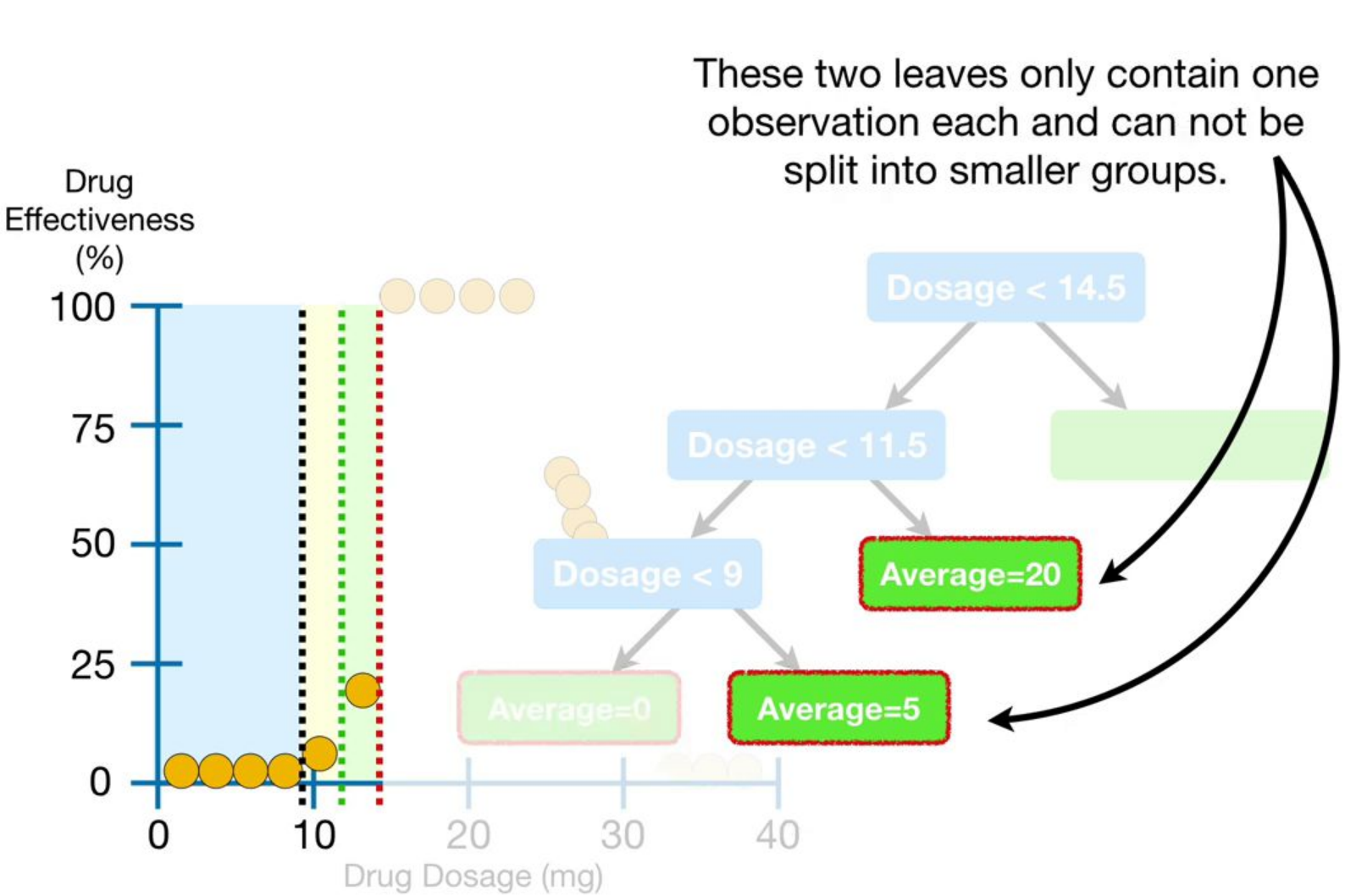


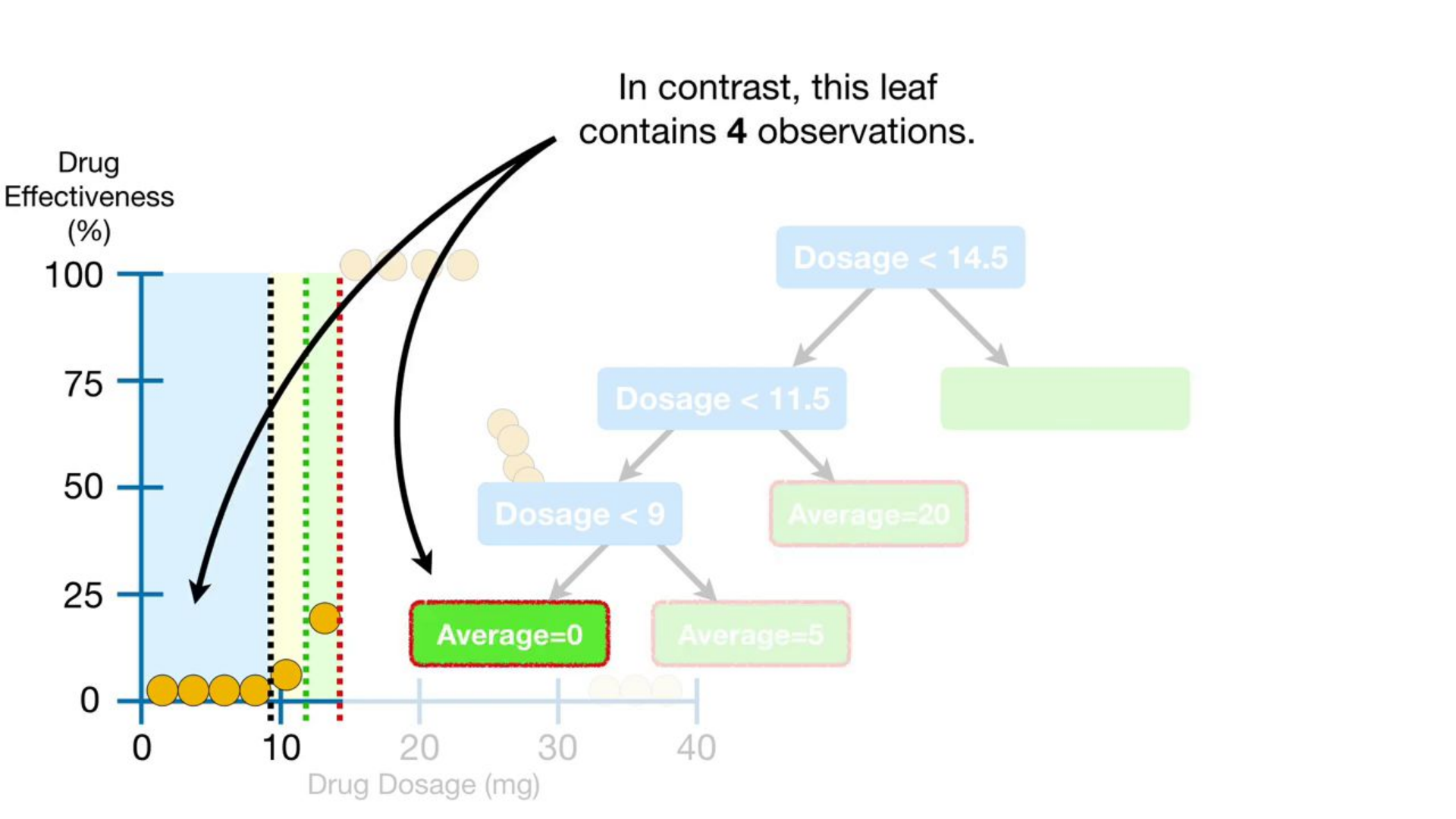


However, since the remaining **5** observations go to the other node, we can split them once more.

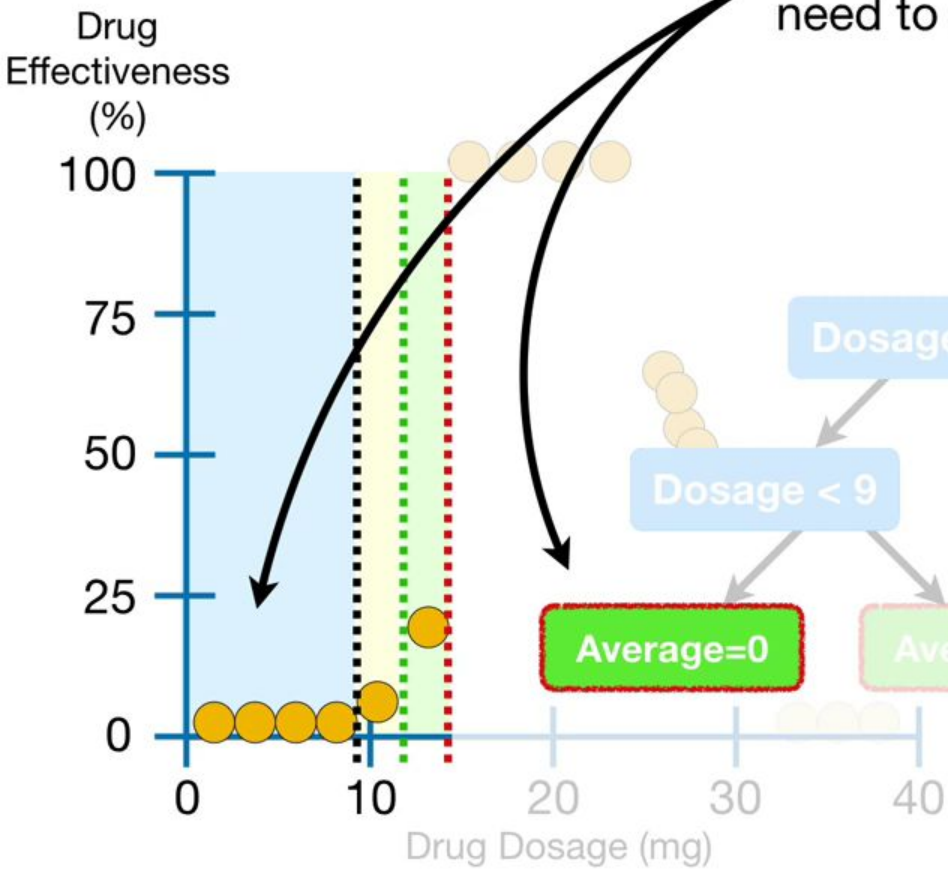


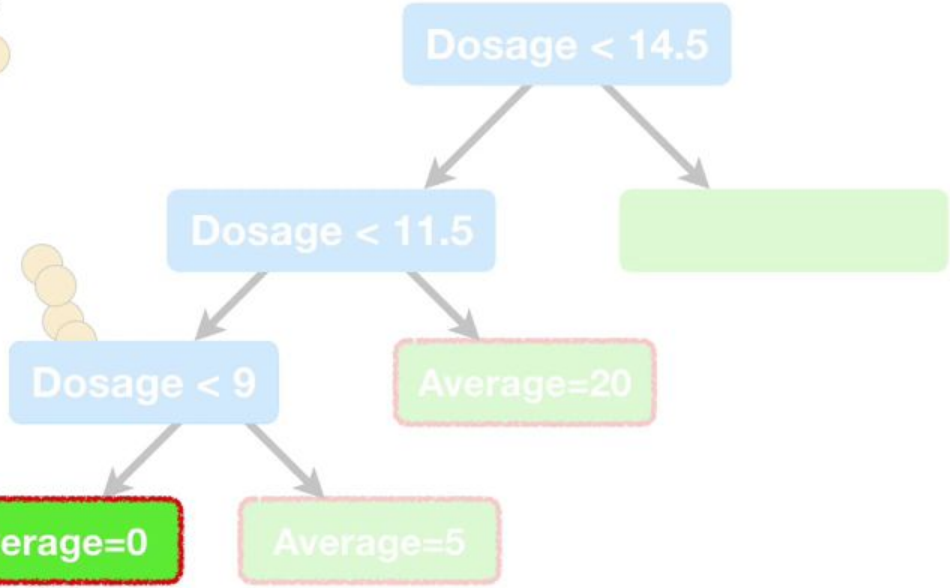

Now we have divided the observations with **Dosage < 14.5** into **3** separate groups.
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Dosage < 11.5
Dosage < 9
Average=20
Average=0
Average=5


These two leaves only contain one observation each and can not be split into smaller groups.
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Dosage < 11.5
Dosage < 9
Average=20
Average=0
Average=5

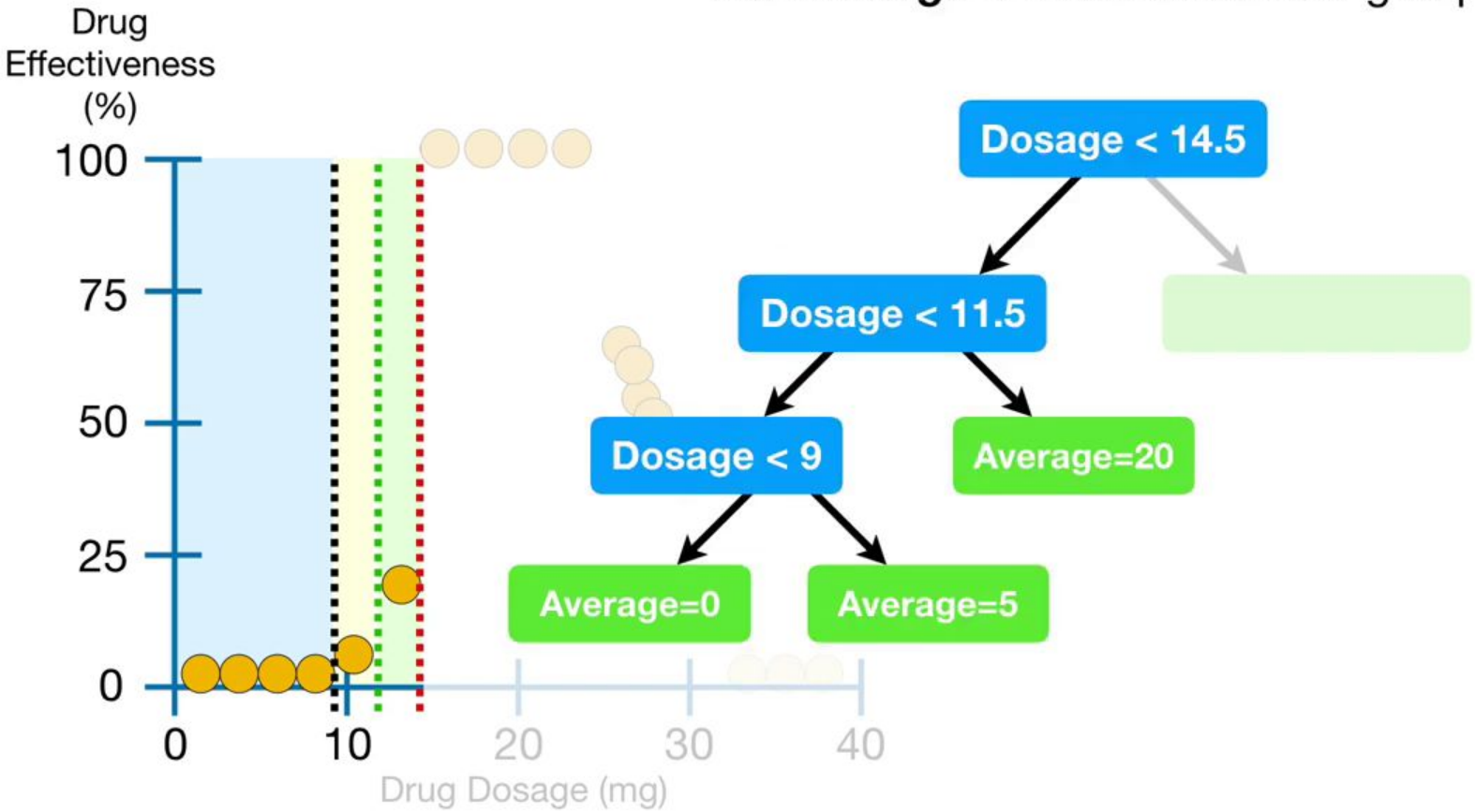

In contrast, this leaf contains 4 observations.
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Dosage < 11.5
Dosage < 9
Average=20
Average=0
Average=5

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

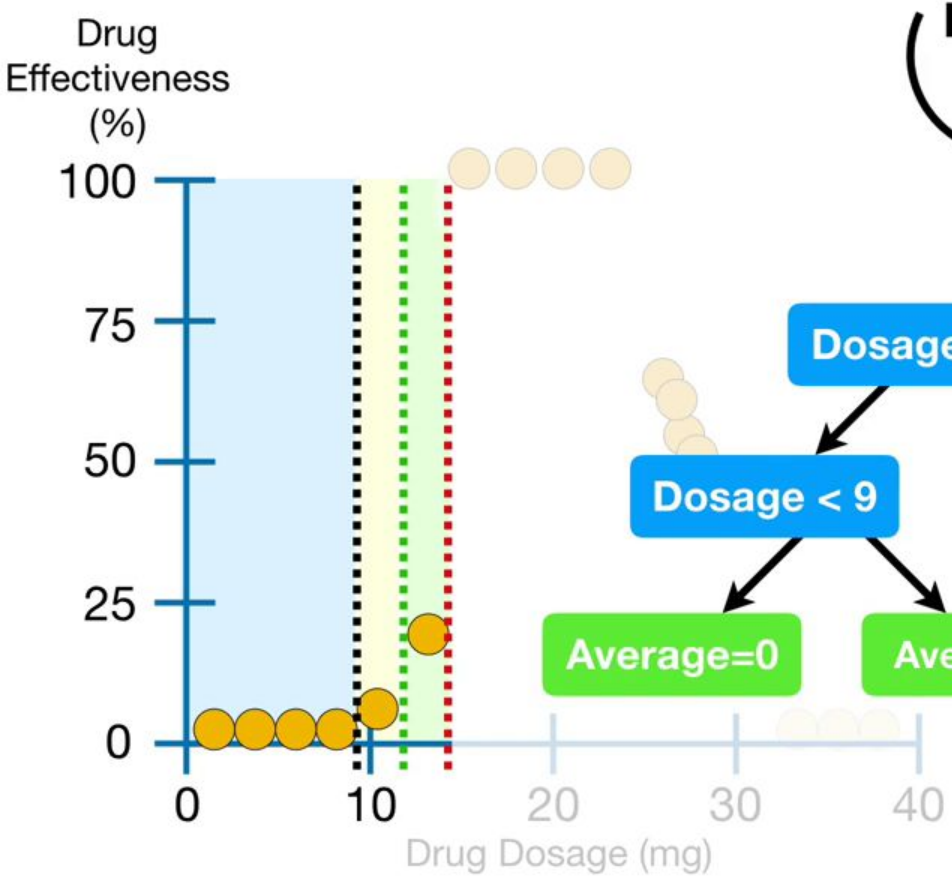


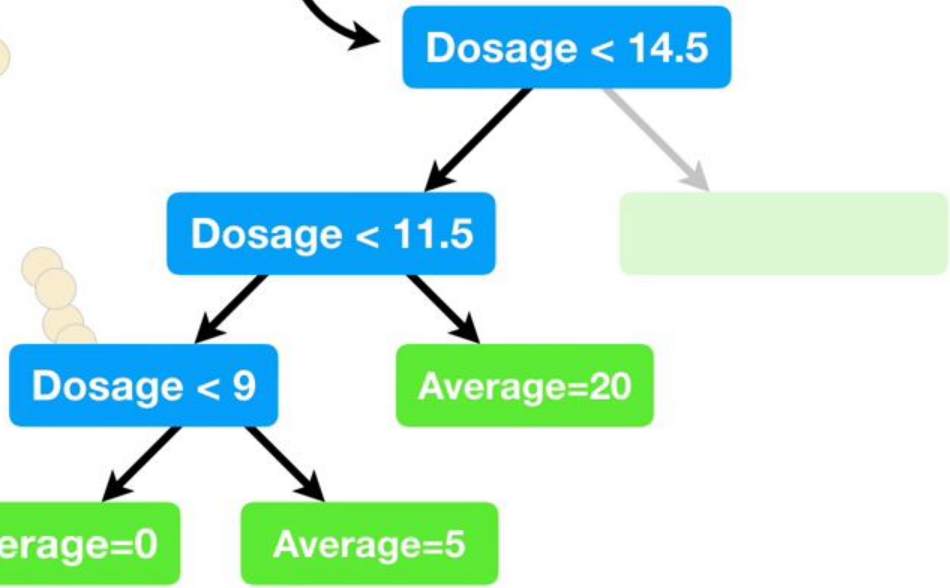


So we are done splitting the observations with **Dosage < 14.5** into smaller groups.

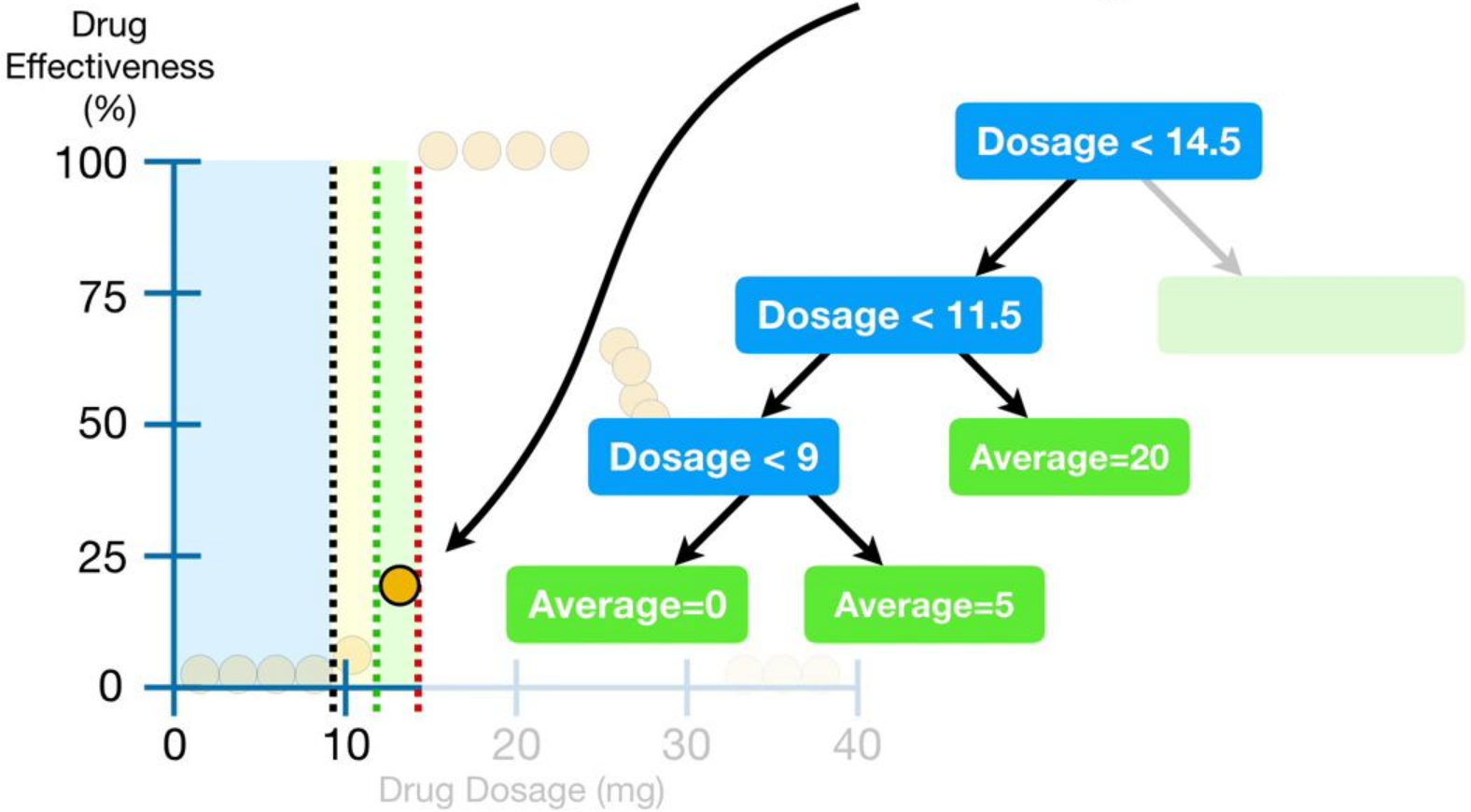


NOTE: The *predictions* that this tree makes for all observations with **Dosage < 14.5** are perfect.

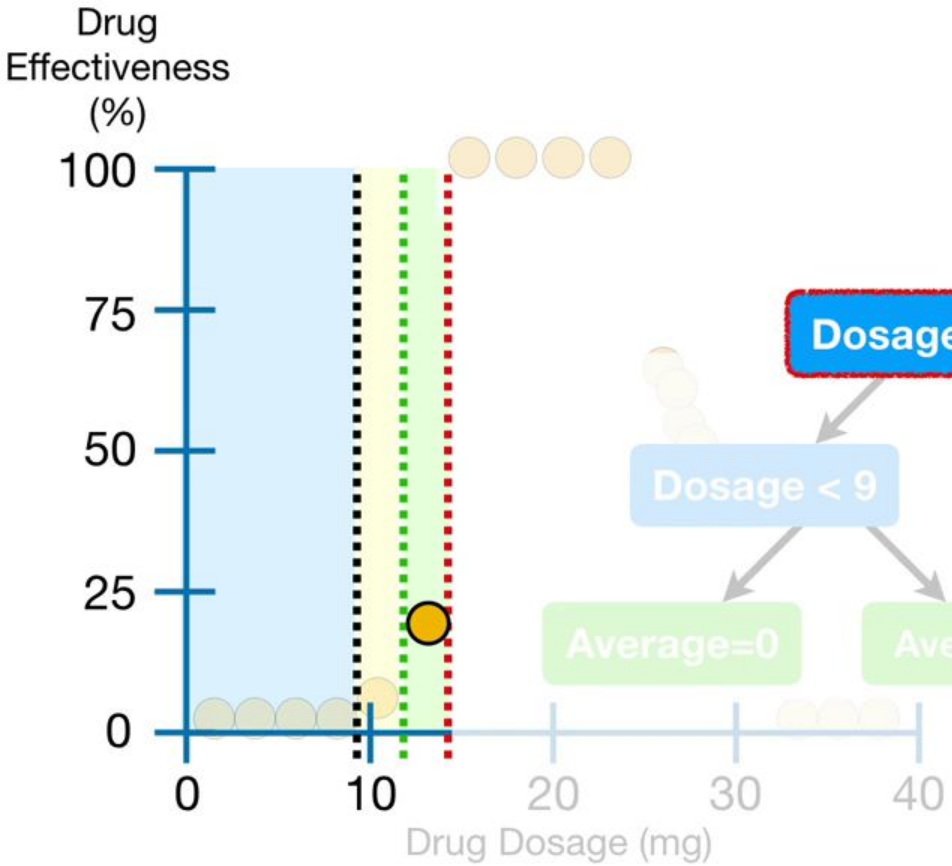


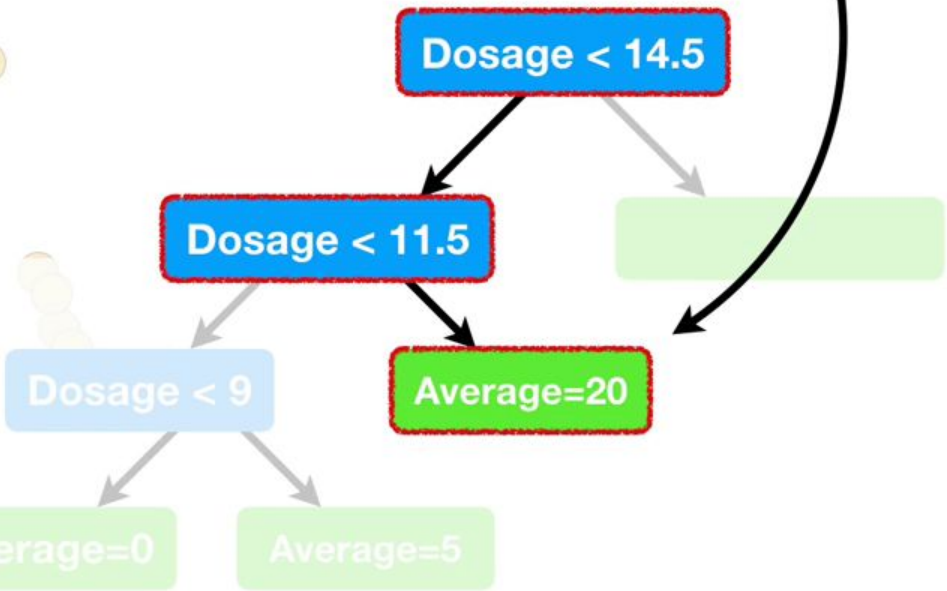


In other words, this observation has **20% Drug Effectiveness**…

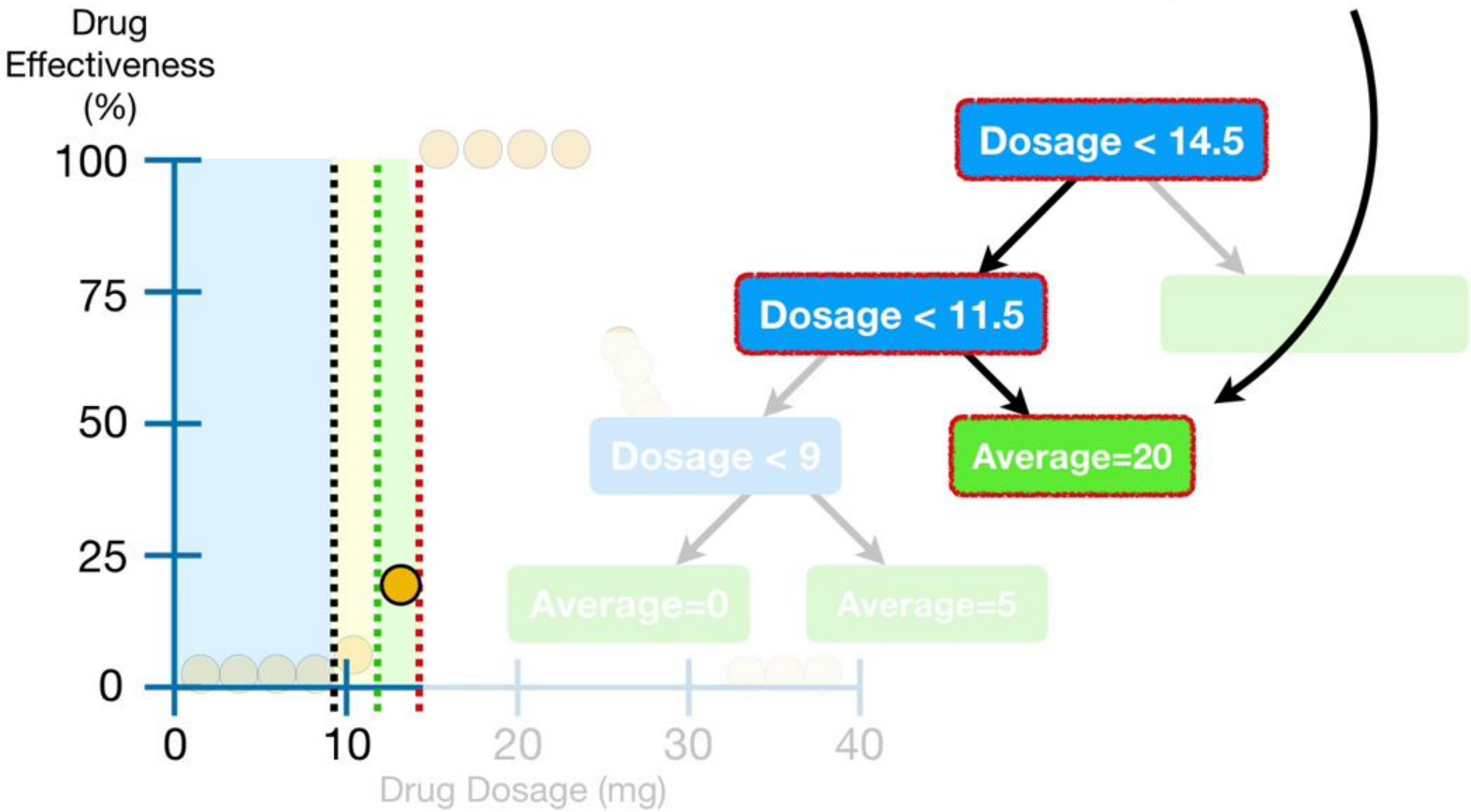


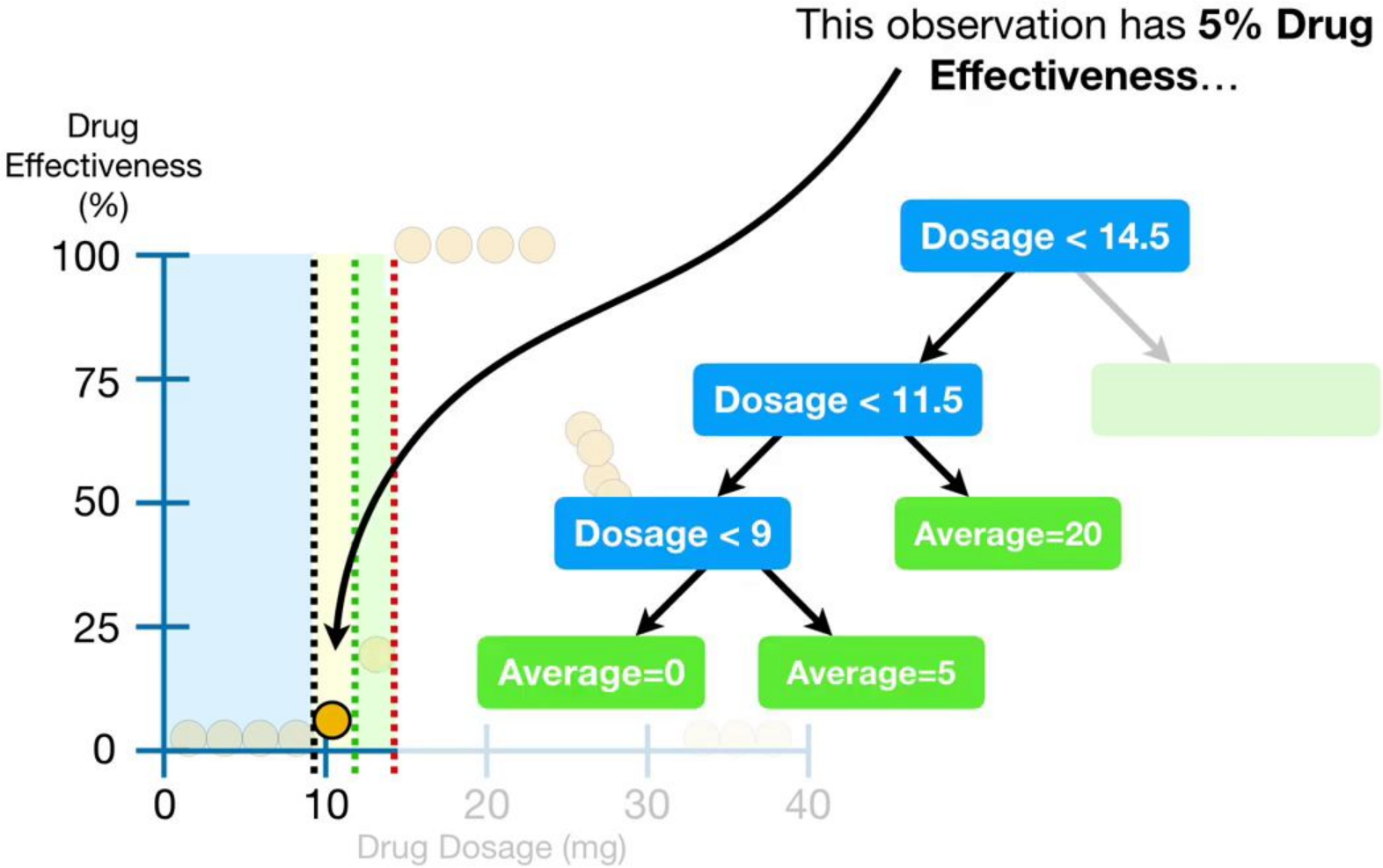
…and the tree predicts **20% Drug Effectiveness**…



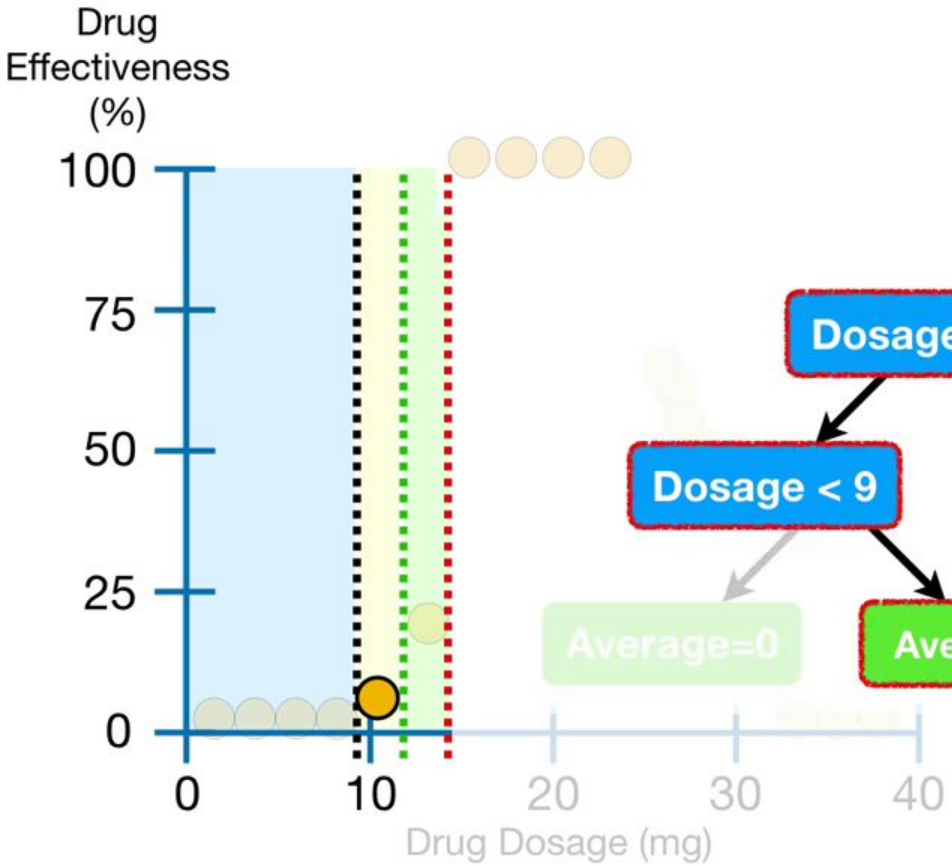


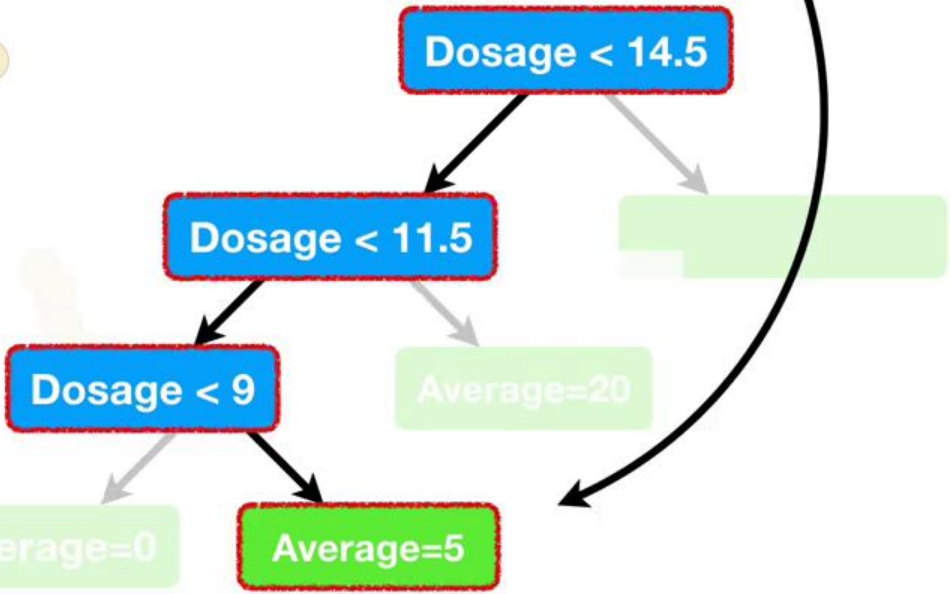
…so the *observed* and *predicted* values are the same.

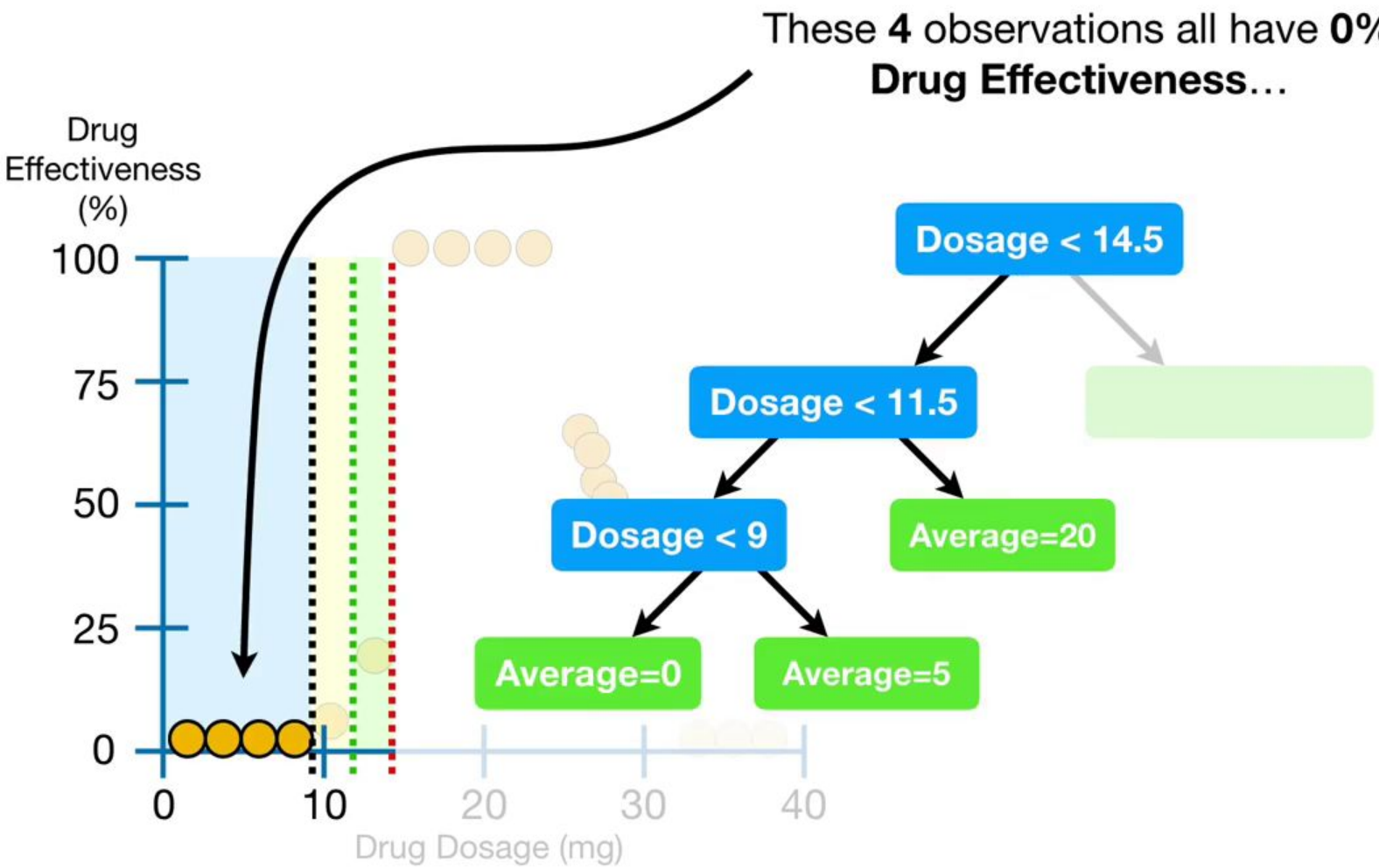



This observation has **5% Drug Effectiveness**…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Dosage < 11.5
Dosage < 9
Average=20
Average=0
Average=5

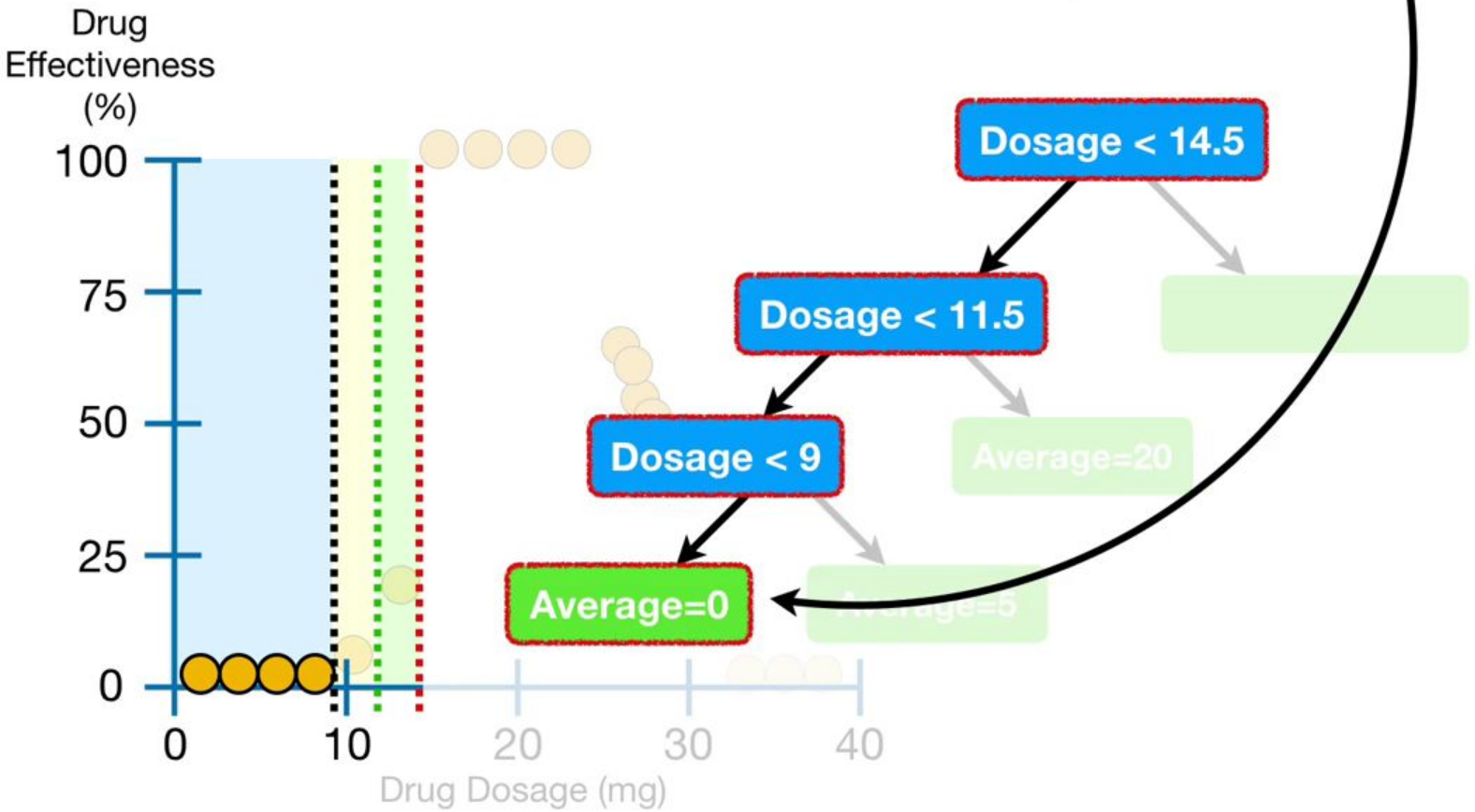
…and that's exactly what the tree predicts.



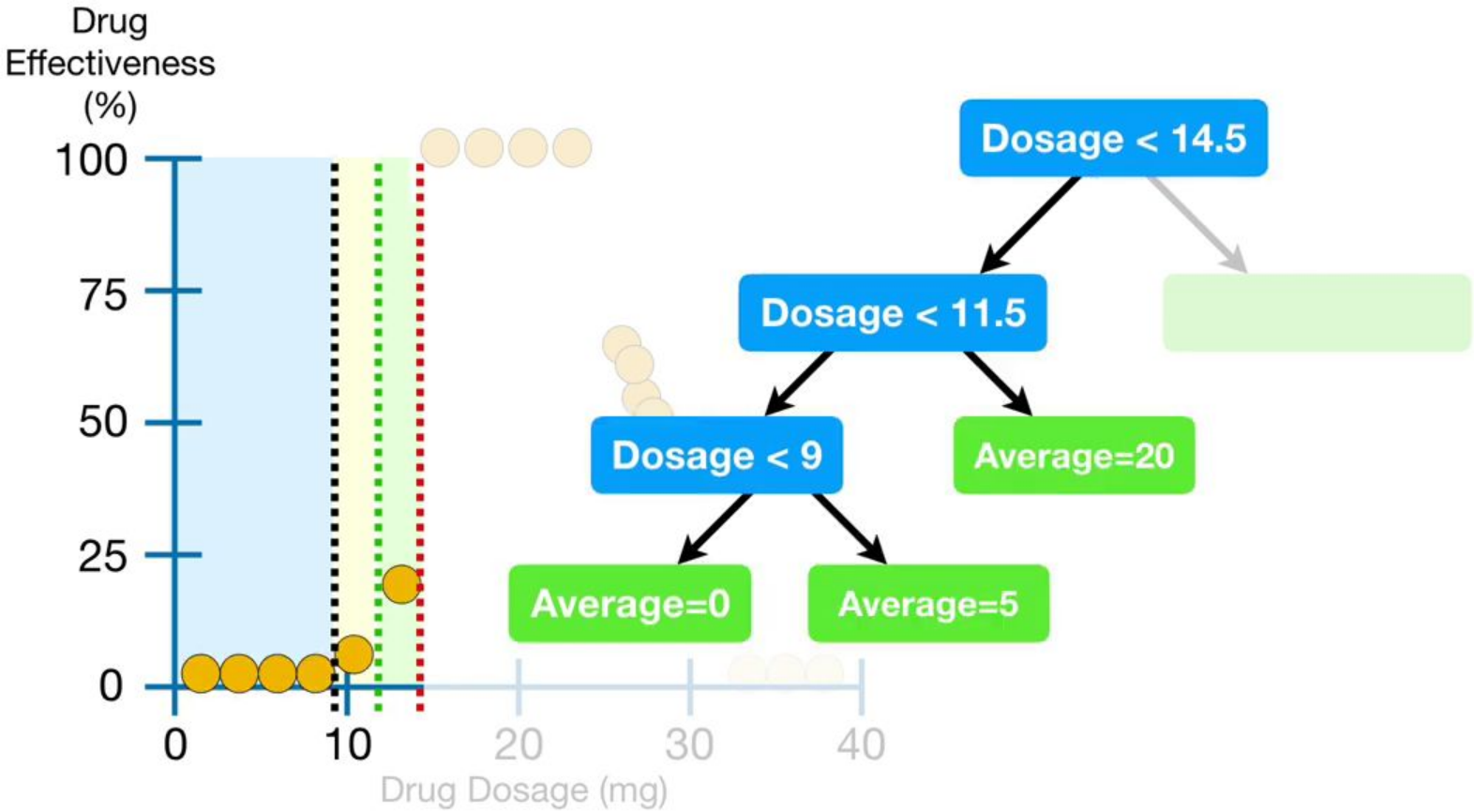



These 4 observations all have 0% Drug Effectiveness…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Dosage < 11.5
Dosage < 9
Average=20
Average=0
Average=5

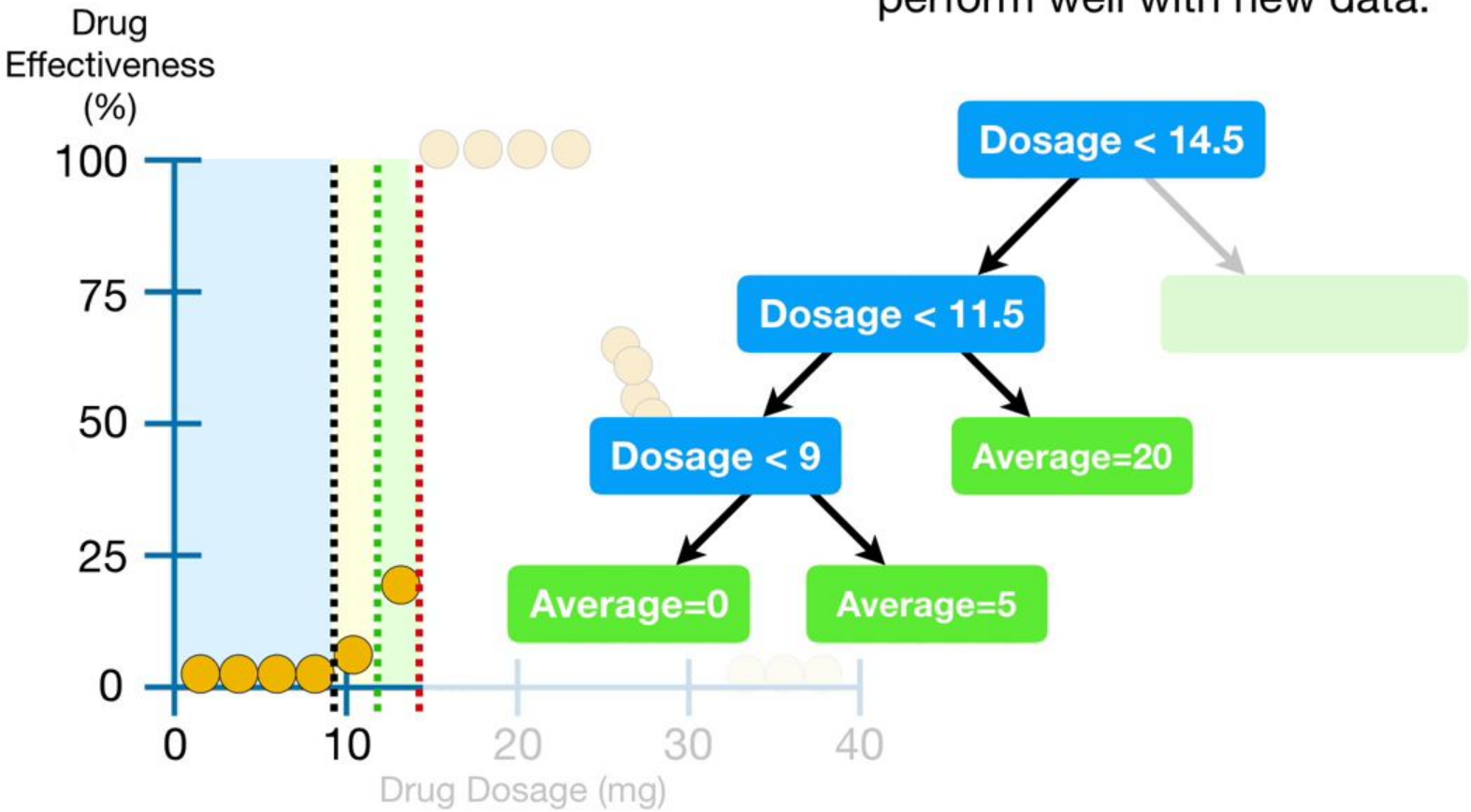
…and that's exactly what the tree predicts.



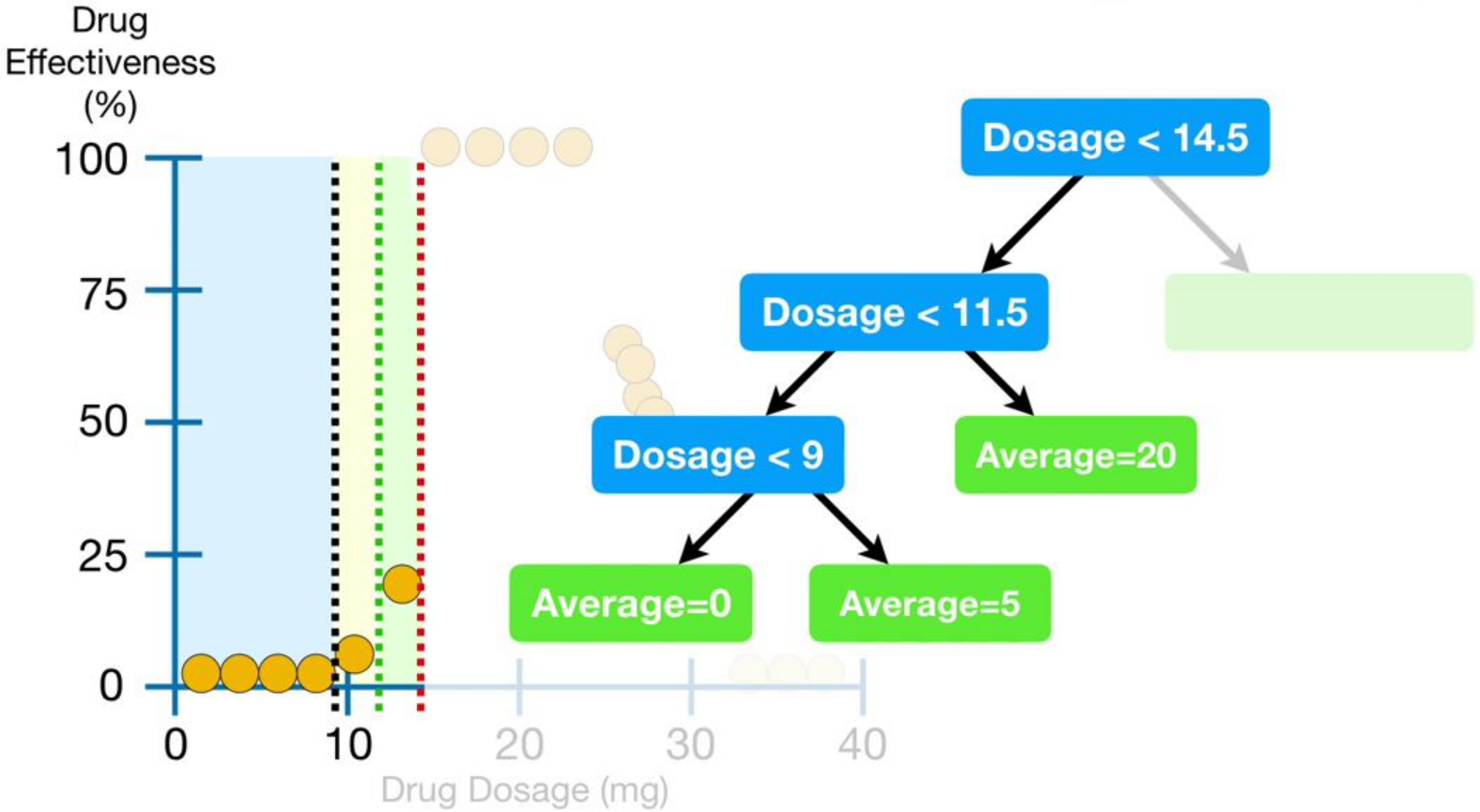
Is that awesome?



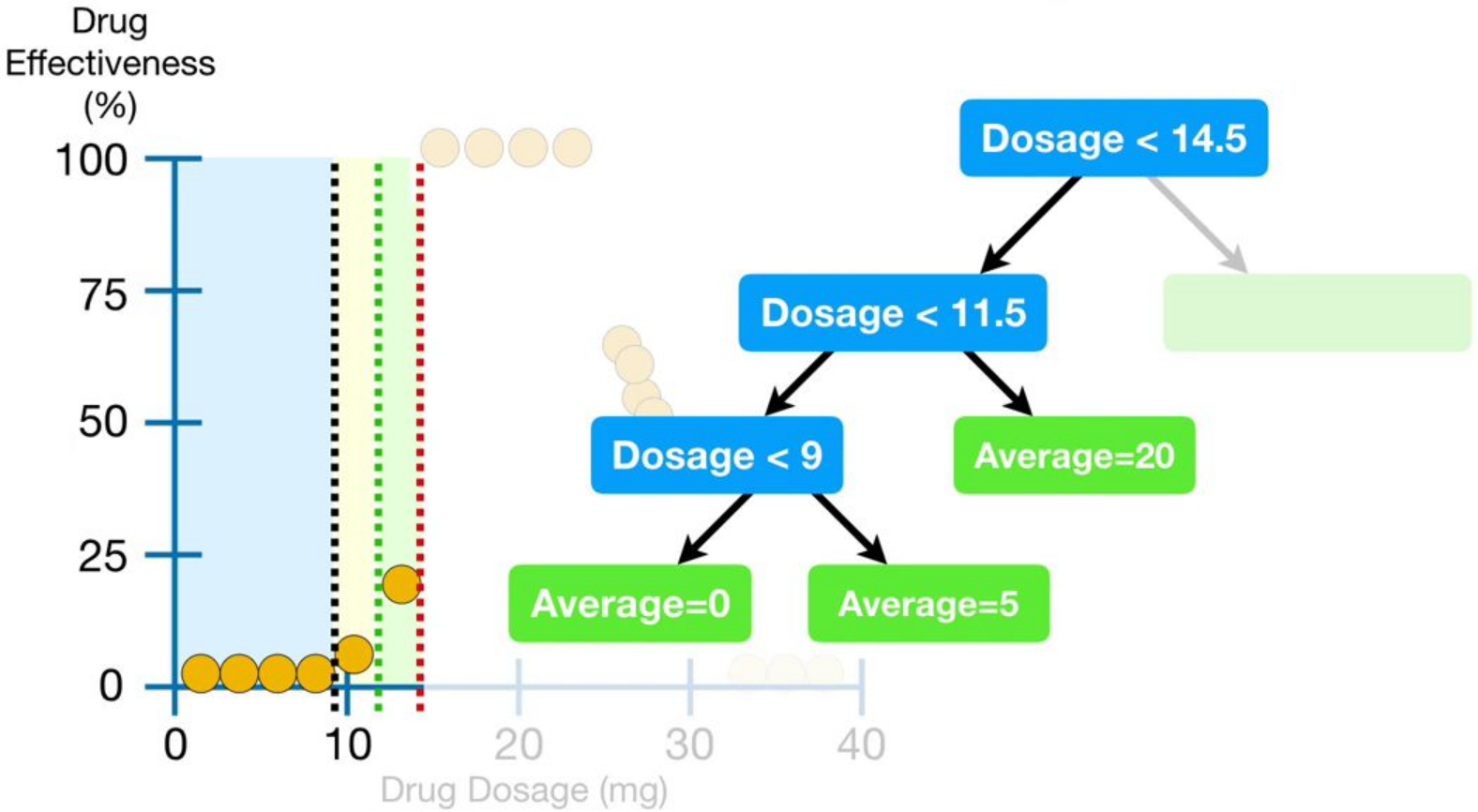
When a model fits the training data perfectly, it probably means it is overfit and will not perform well with new data.



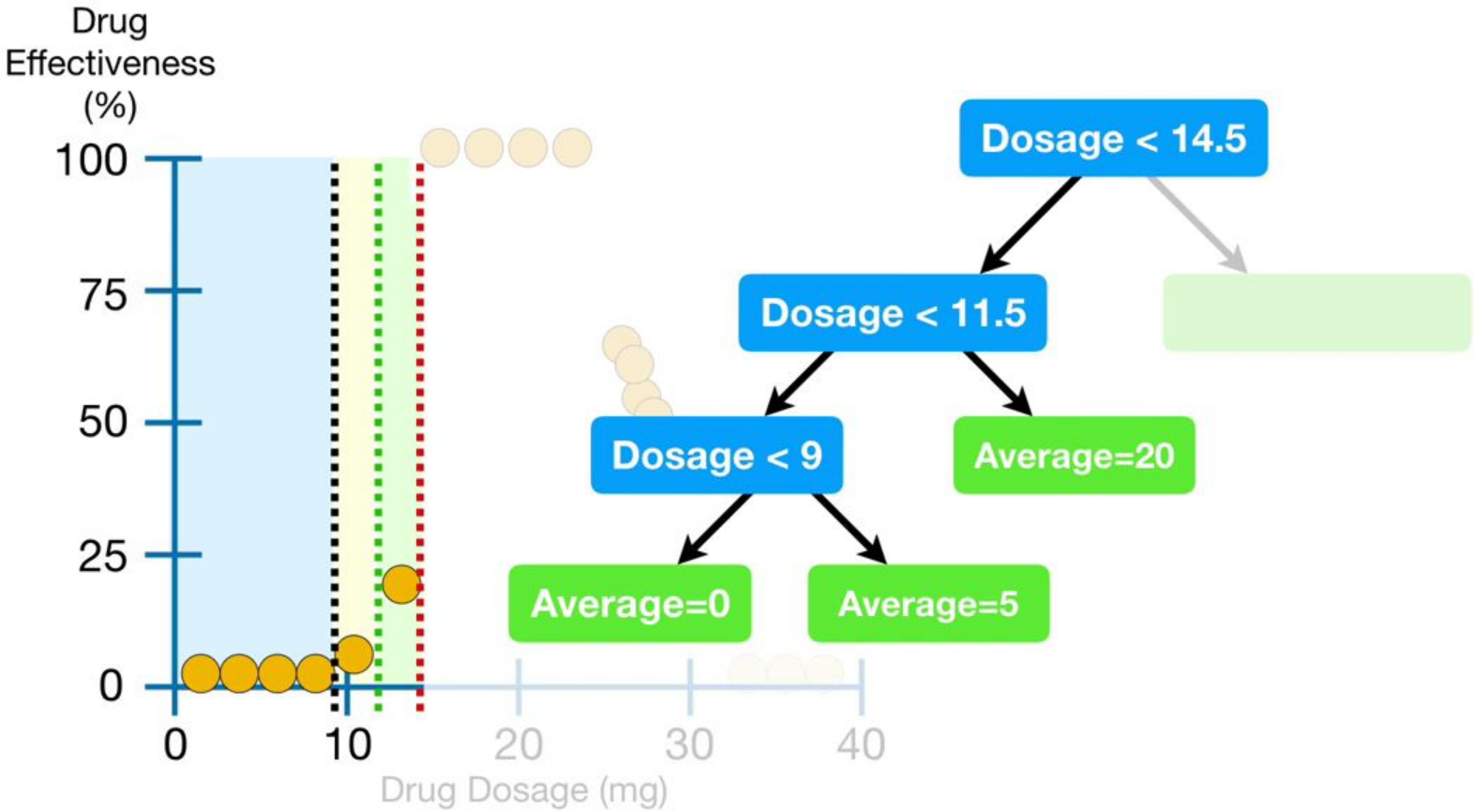
In **Machine Learning Lingo**, the model has no ***Bias***, but potentially large ***Variance***.



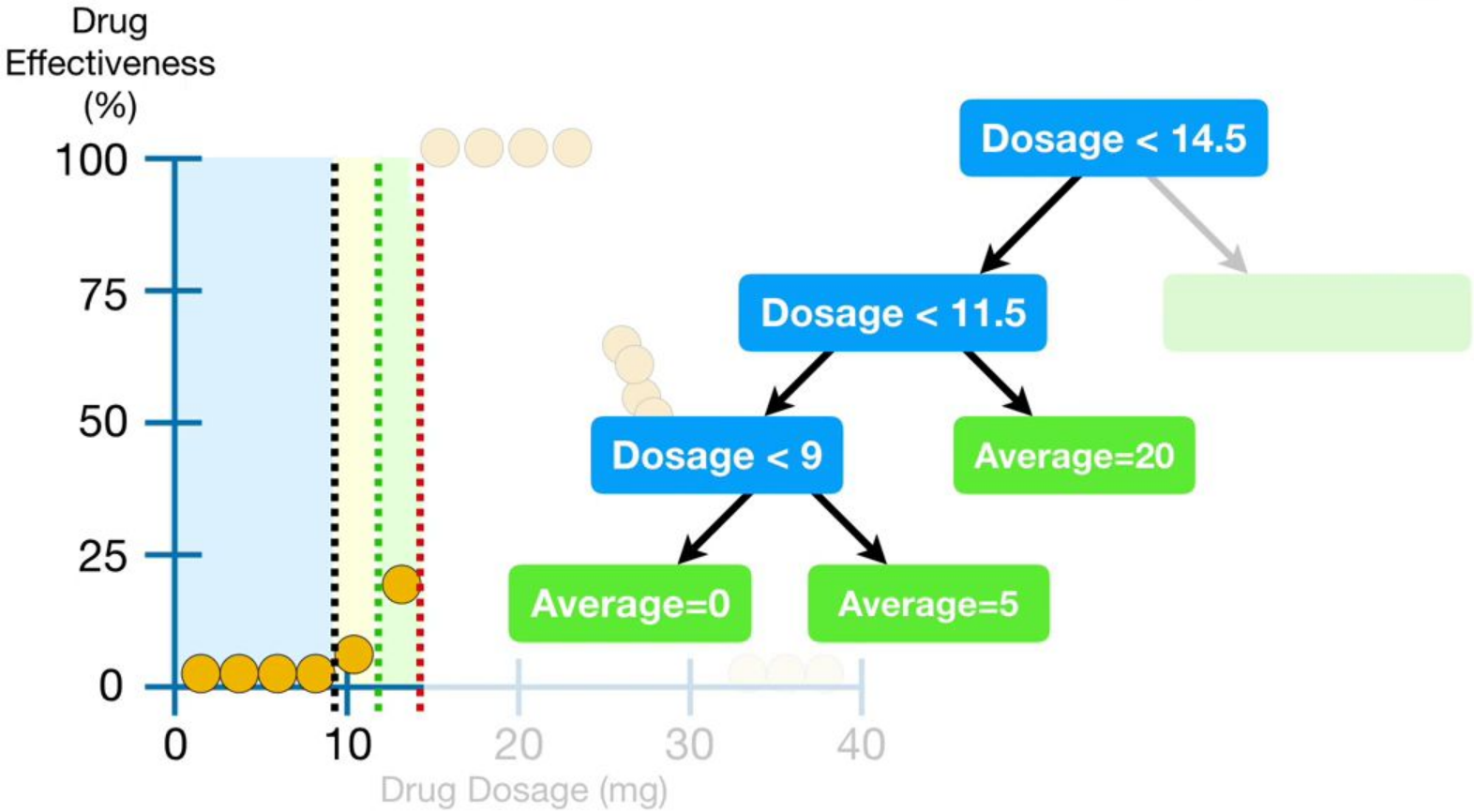
Is there a way to prevent our tree from overfitting the training data?



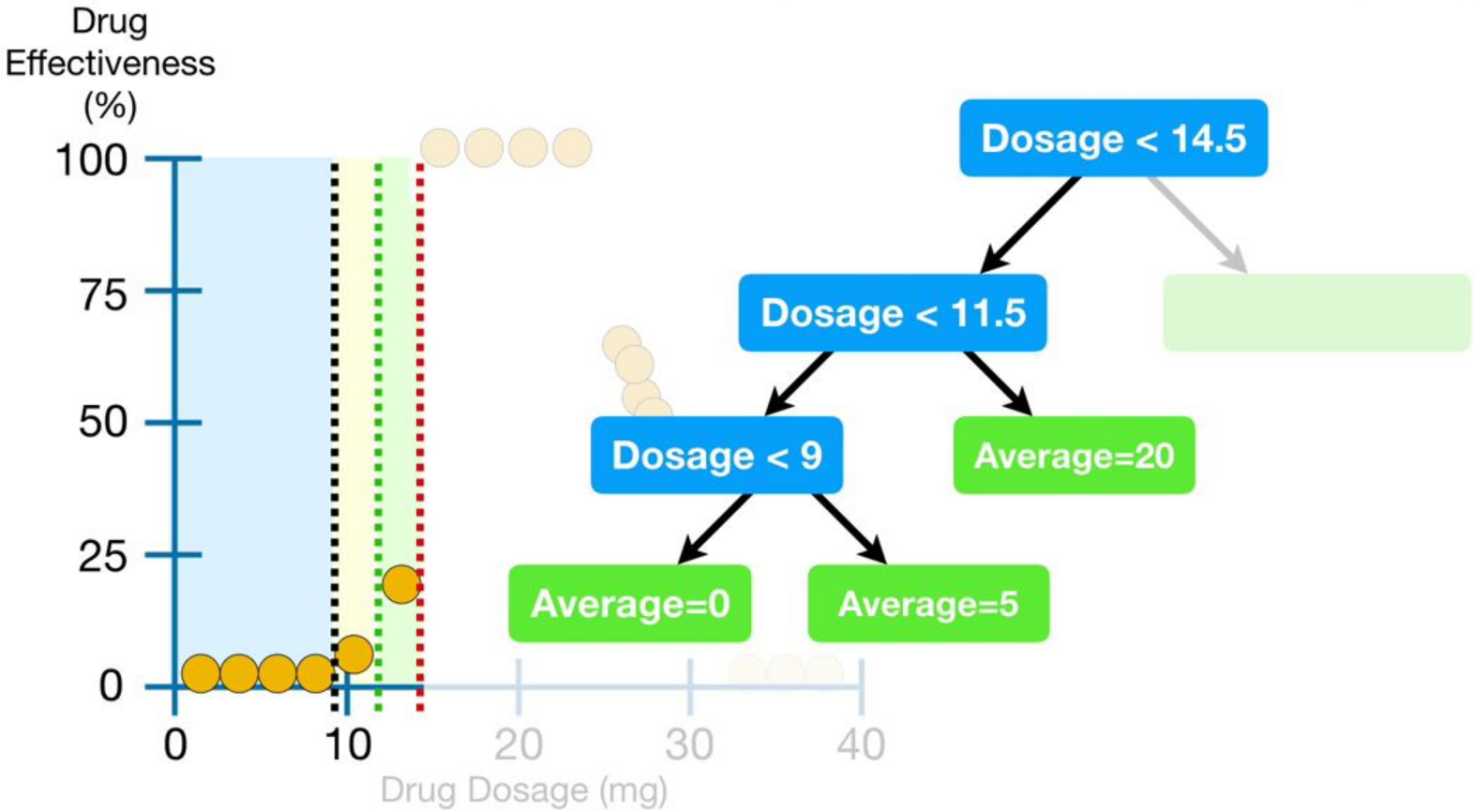
Yes, there are a bunch of techniques.



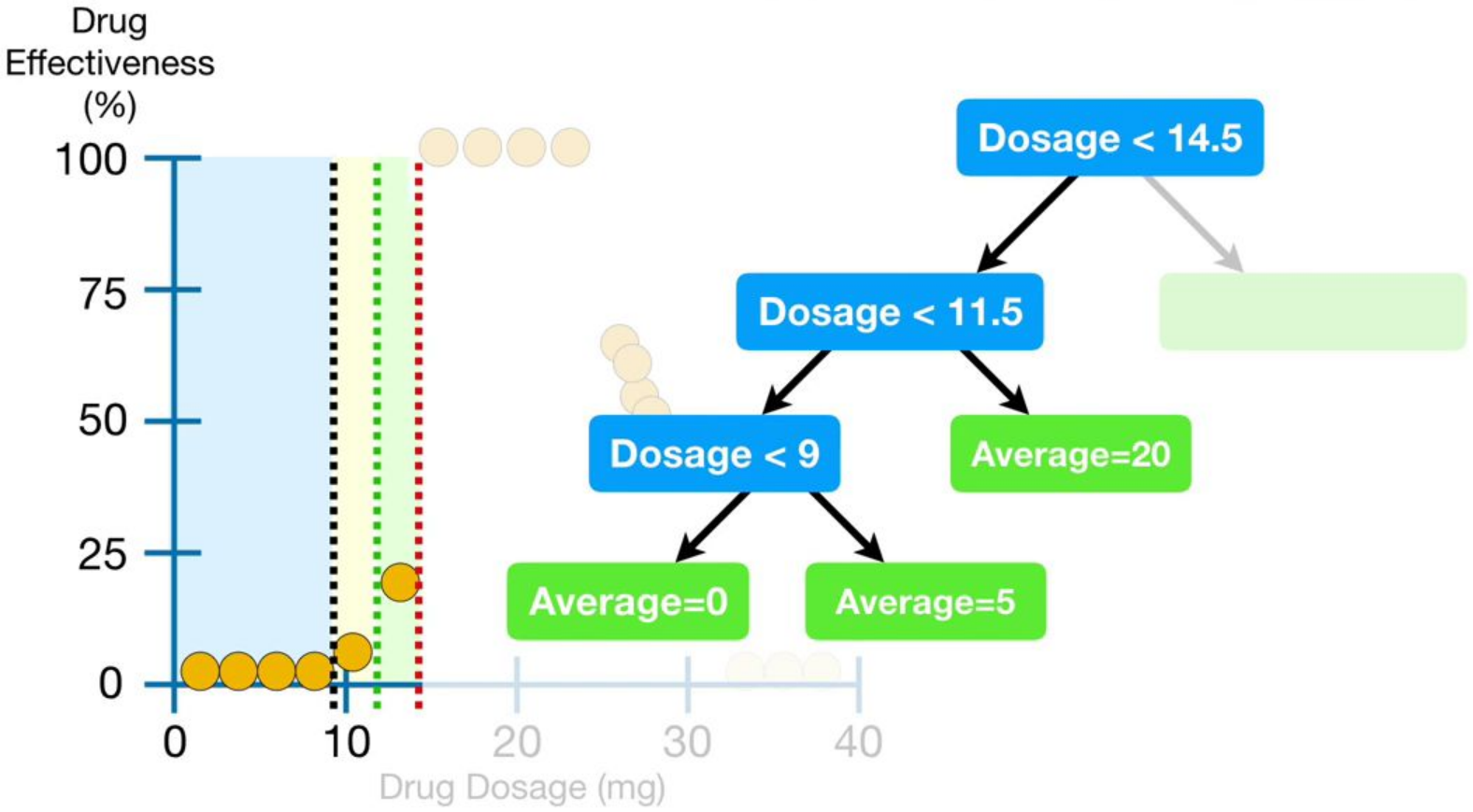
The simplest is to only split observations when there are more than some minimum number.

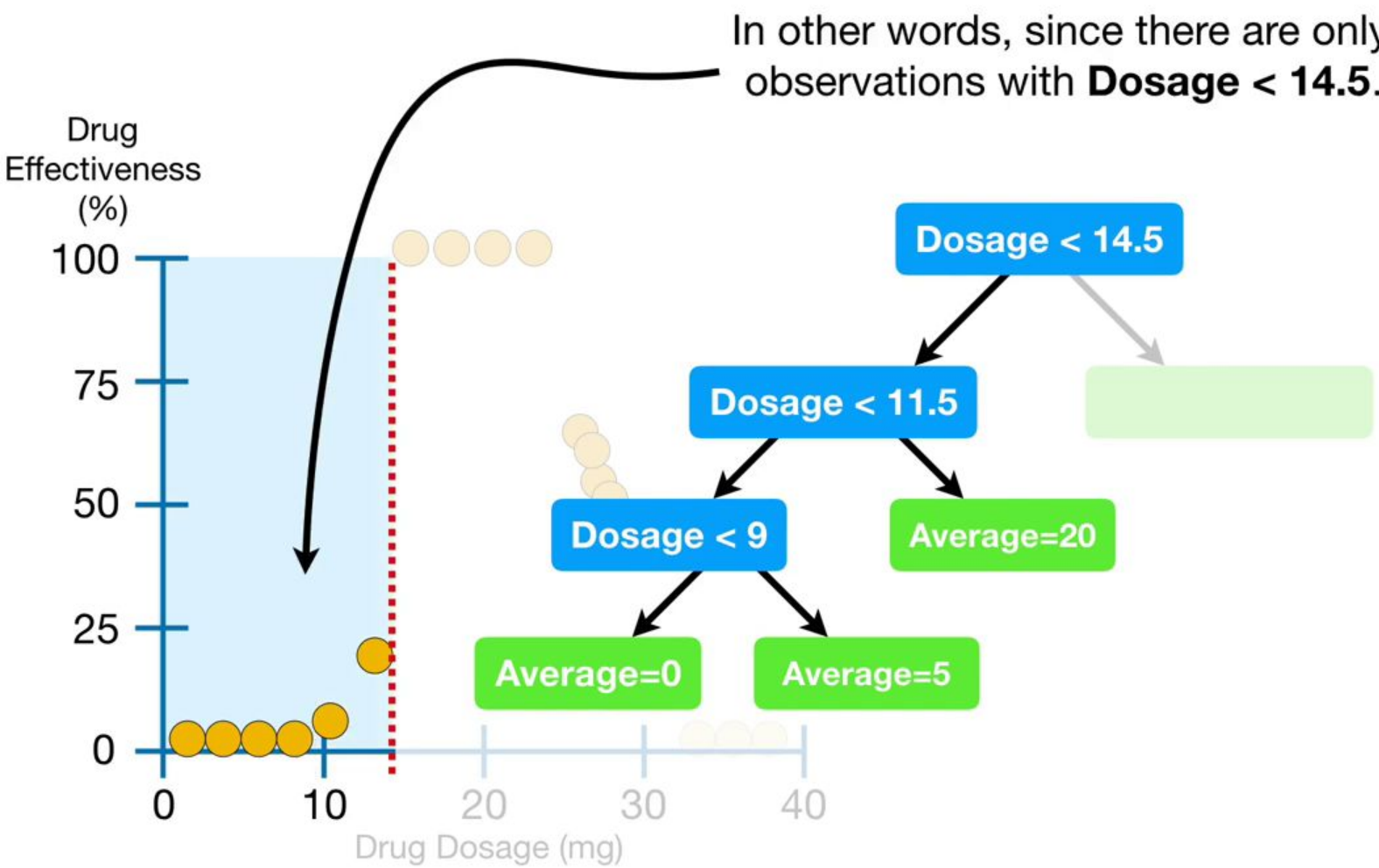


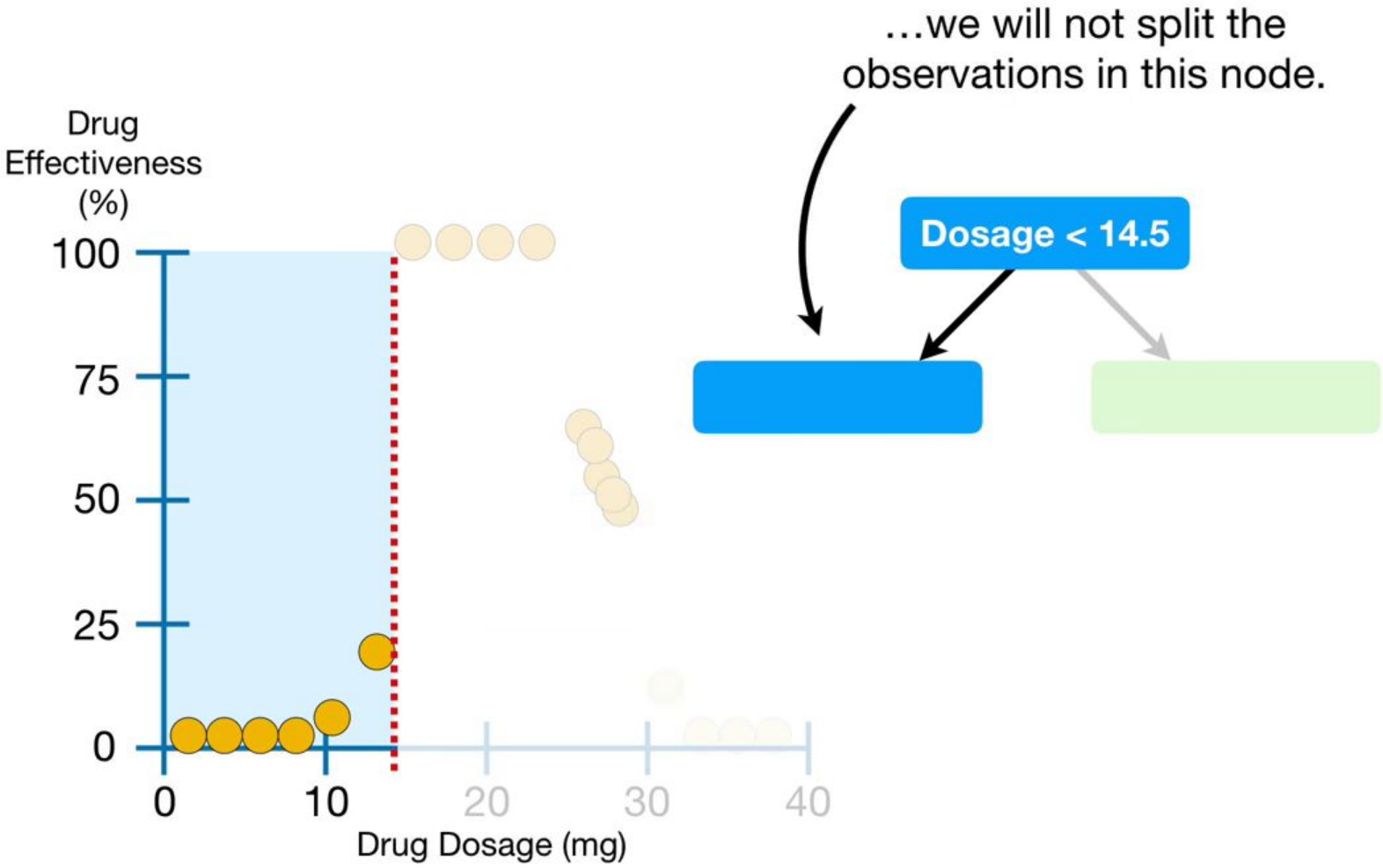
Typically, the minimum number of observations to allow for a split is **20**.

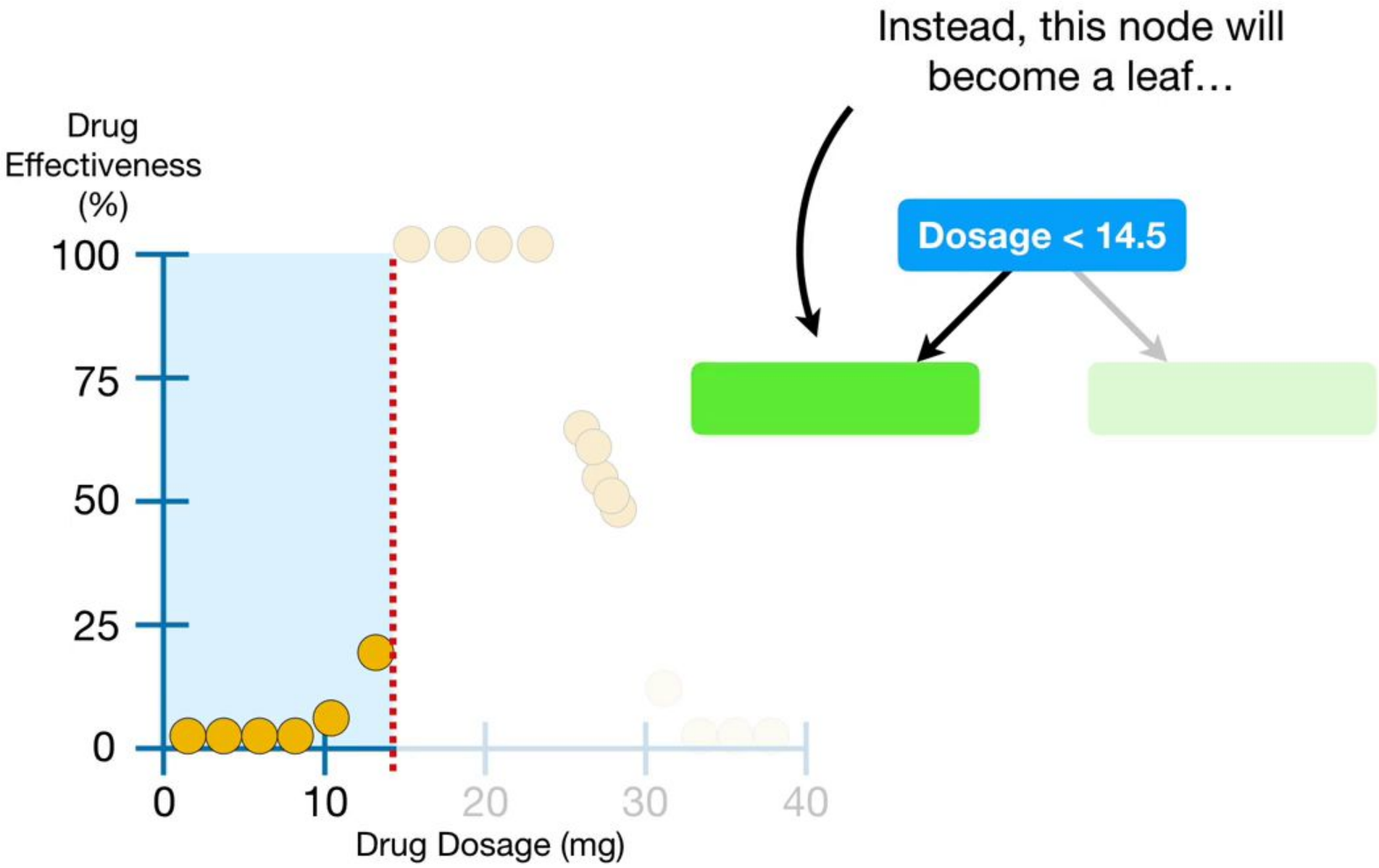


However, since this example doesn't have many observations, I set the minimum to **7**.

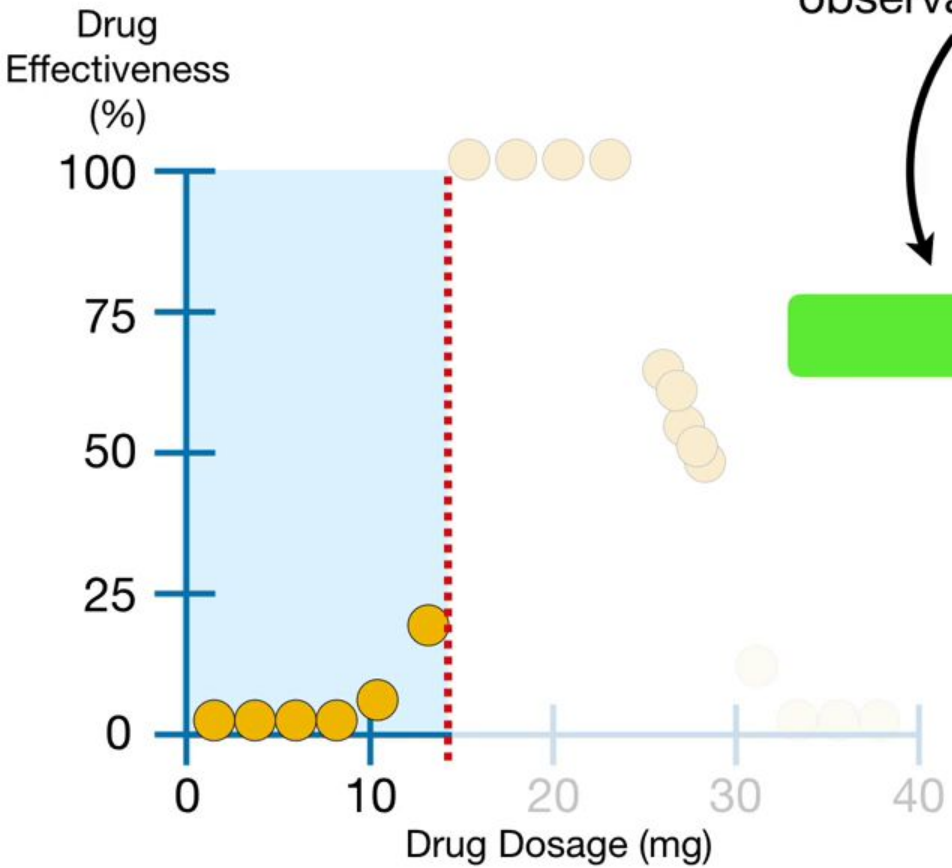


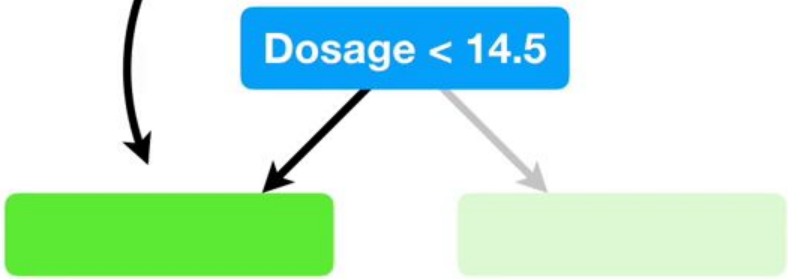

In other words, since there are only 6 observations with Dosage < 14.5…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Dosage < 11.5
Dosage < 9
Average=20
Average=0
Average=5


…we will not split the observations in this node.
Dosage < 14.5
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)

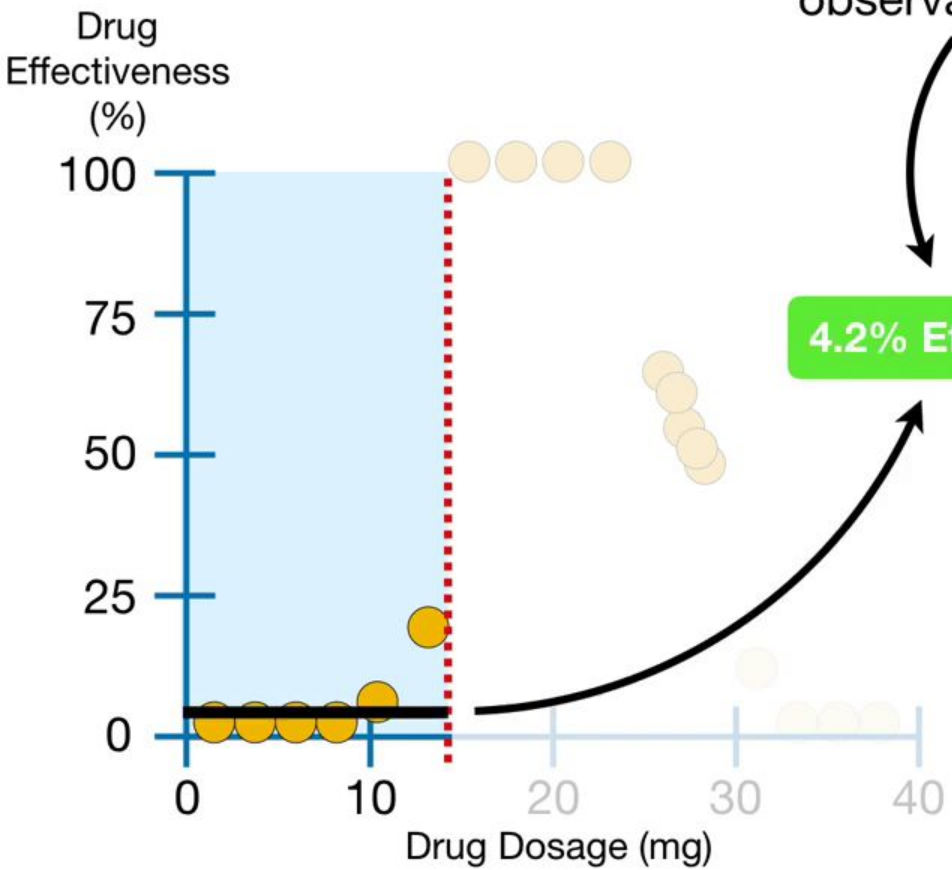

Instead, this node will become a leaf…
Dosage < 14.5
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)

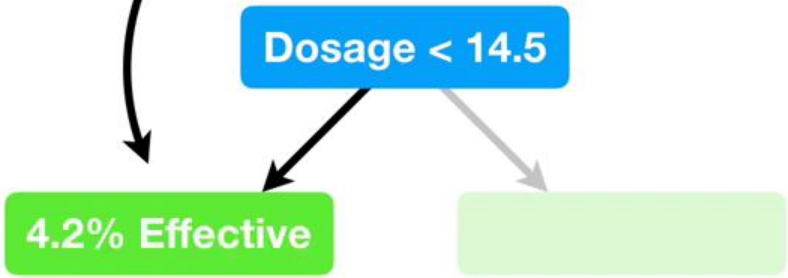
…and the output will be the average **Drug Effectiveness** for the **6** observations with **Dosage < 14.5**, **4.2%**.



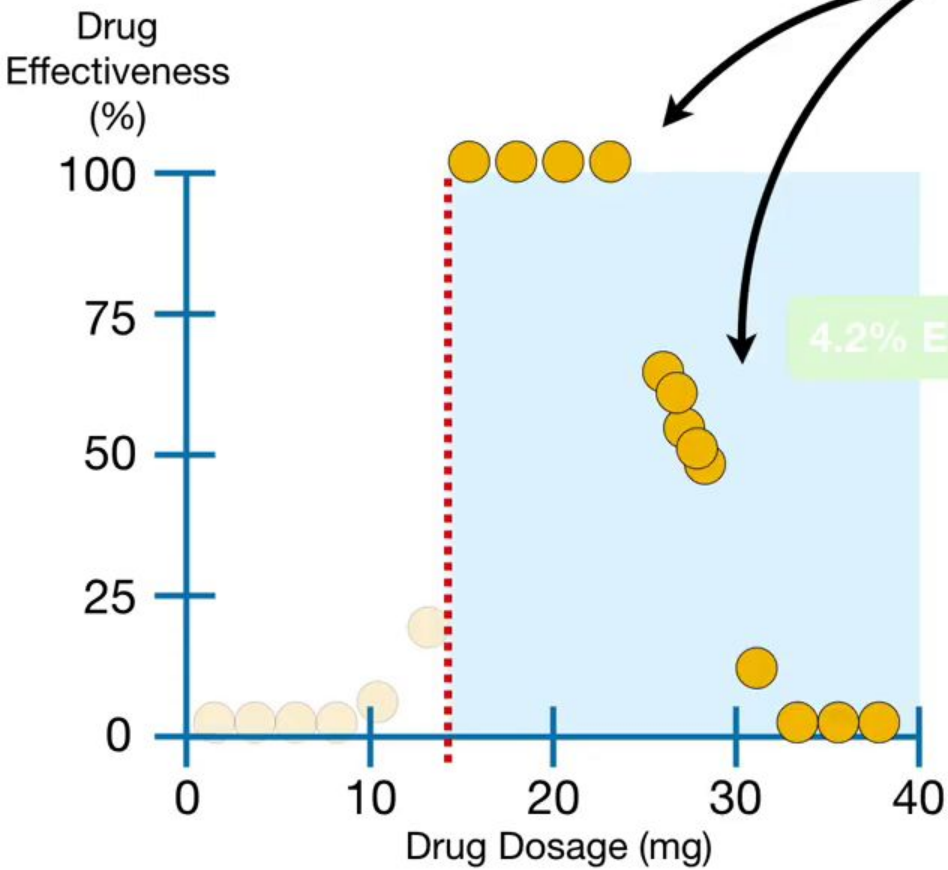


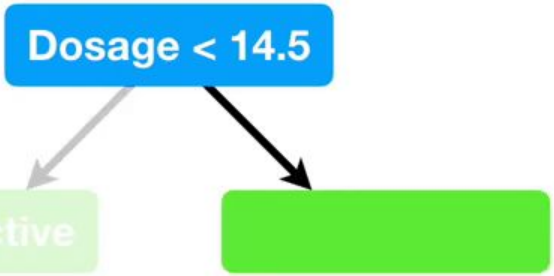
…and the output will be the average **Drug Effectiveness** for the **6** observations with **Dosage < 14.5**, **4.2%**.



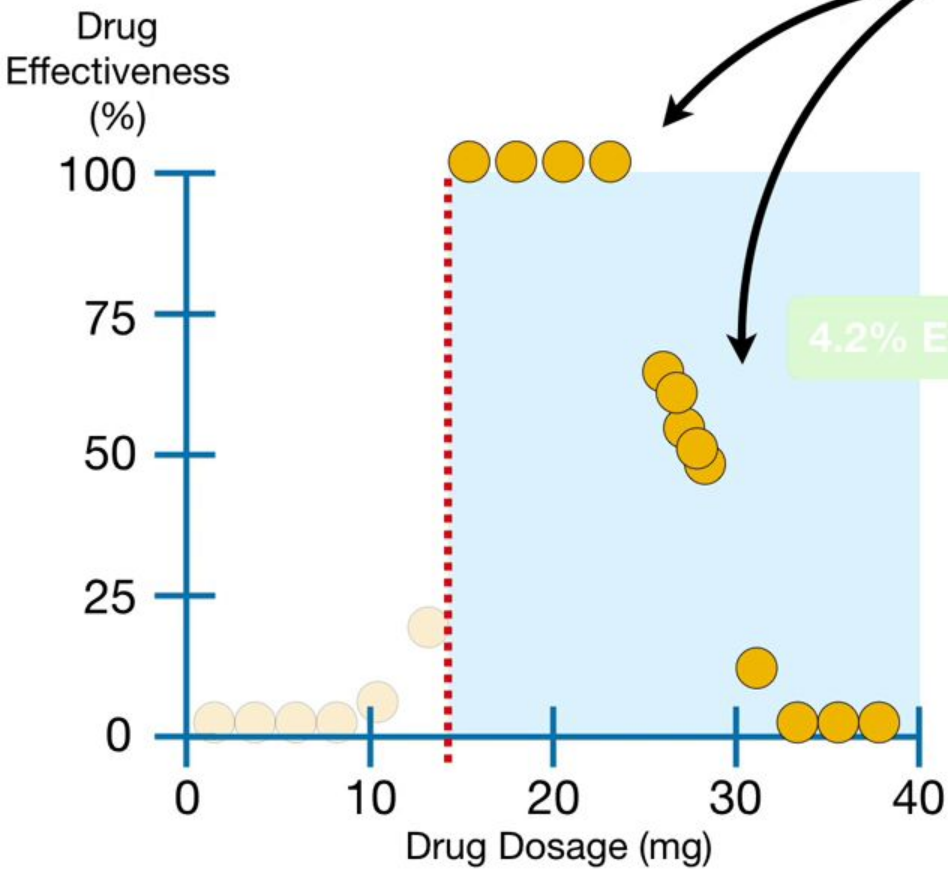


Now we need to figure out what to do with the remaining **13** observations with **Dosages >= 14.5**.





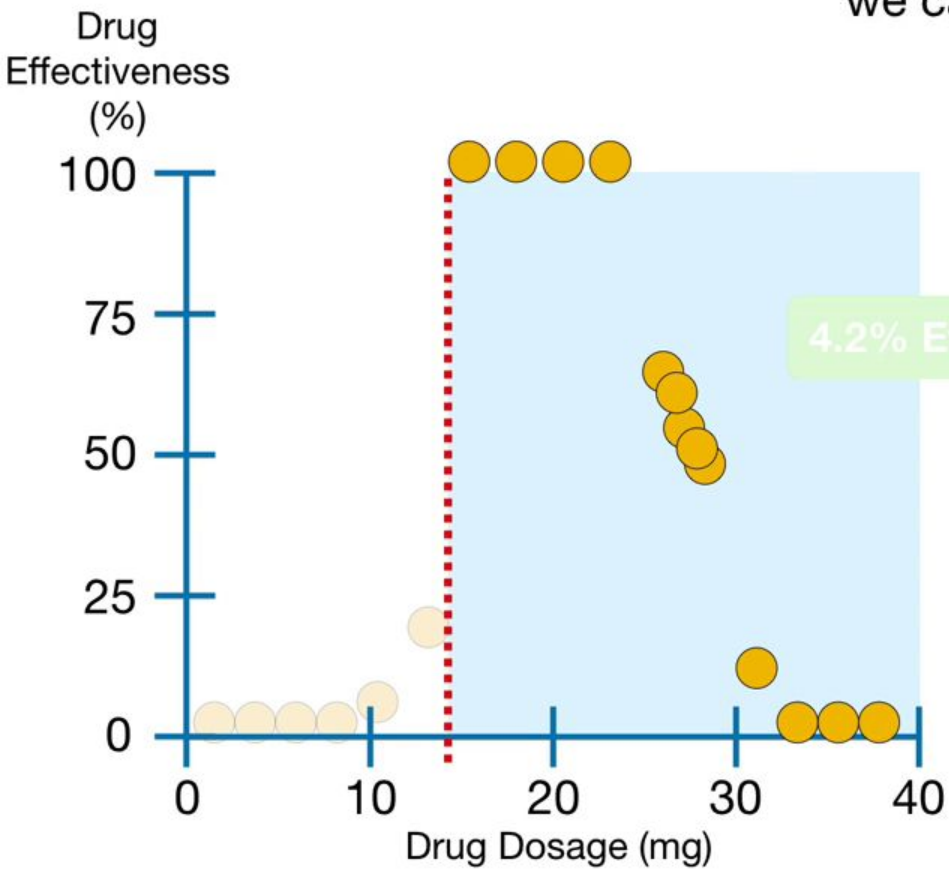
Now we need to figure out what to do with the remaining **13** observations with **Dosages >= 14.5**.

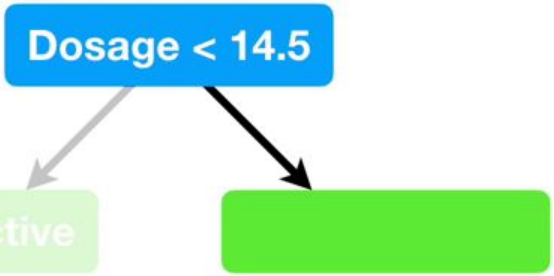


Dosage < 14.5

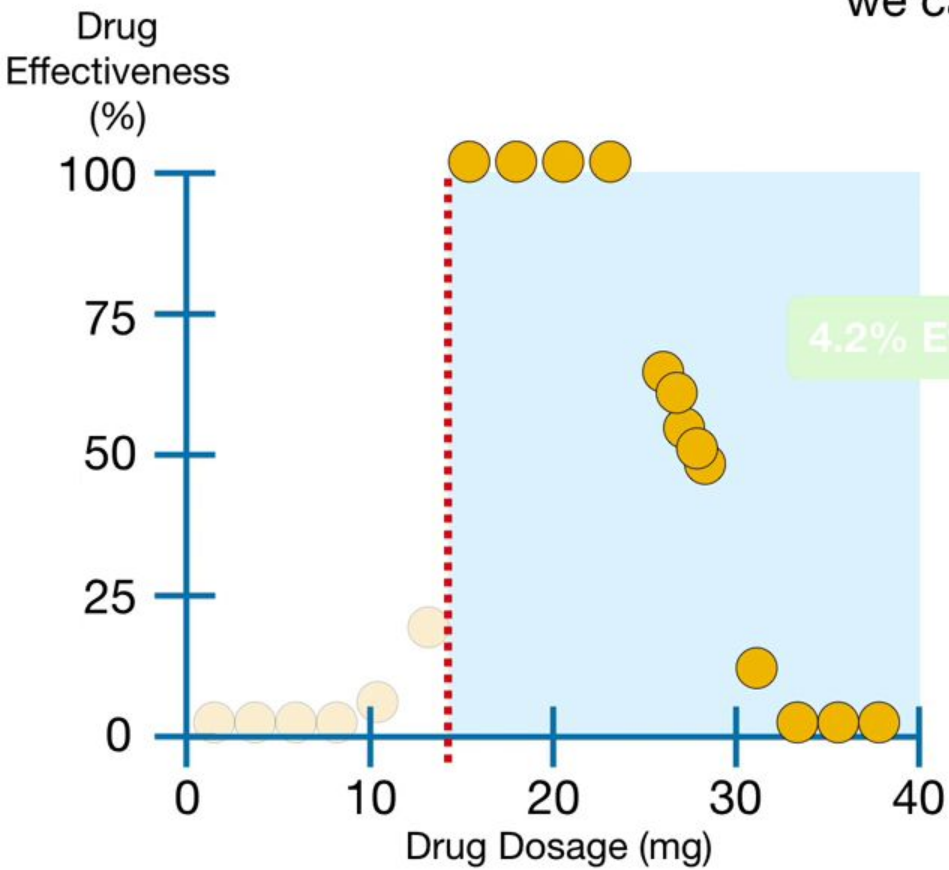
4.2% Effective

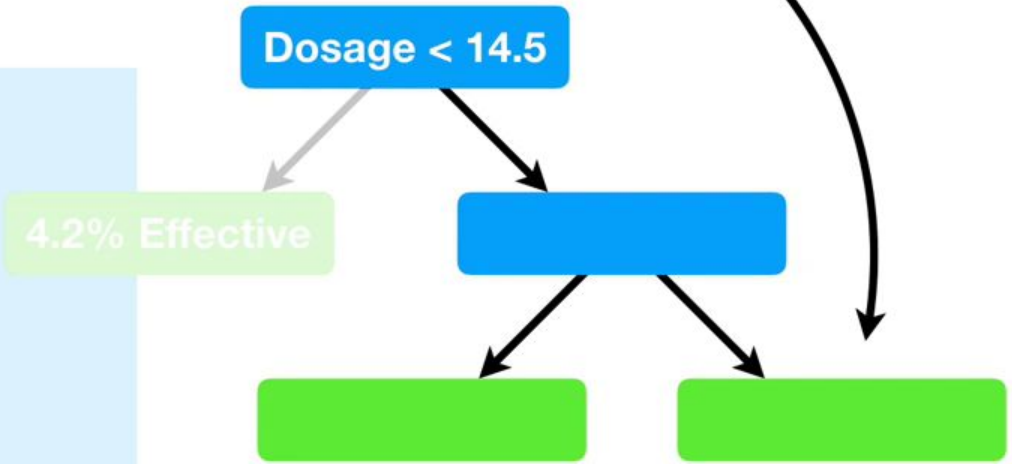
Since we have more than **7** observations on the right side (with **Dosage >= 14.5**), we can split them into two groups…



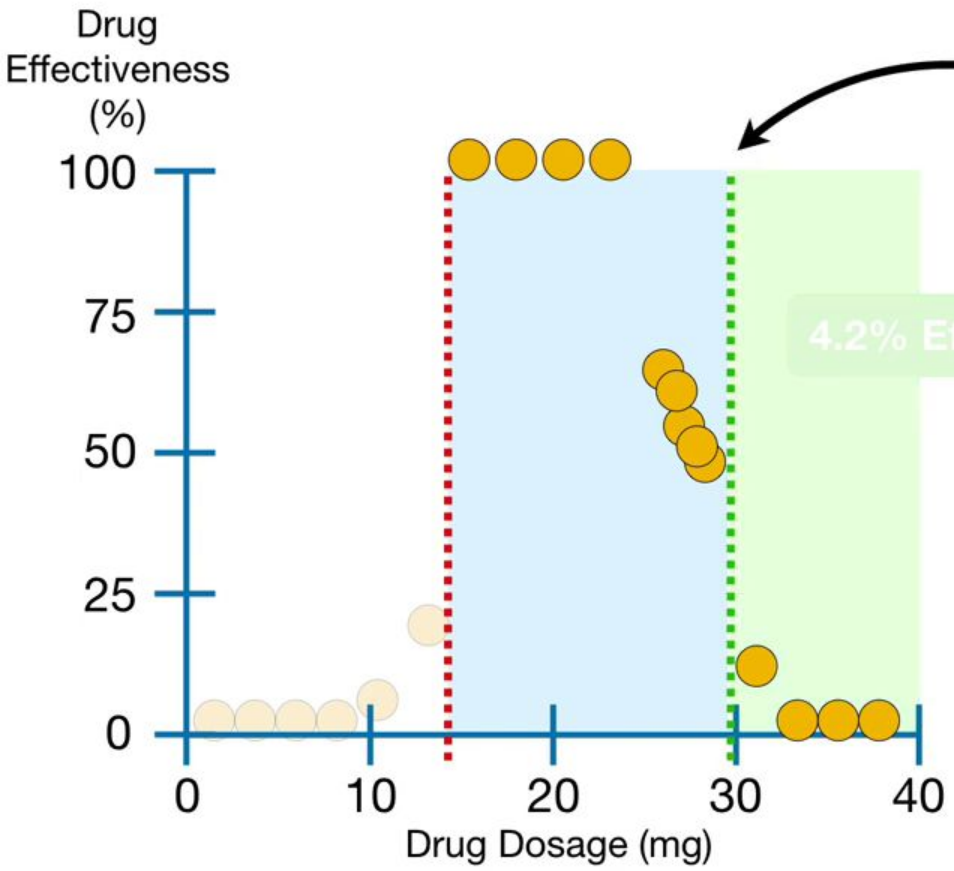


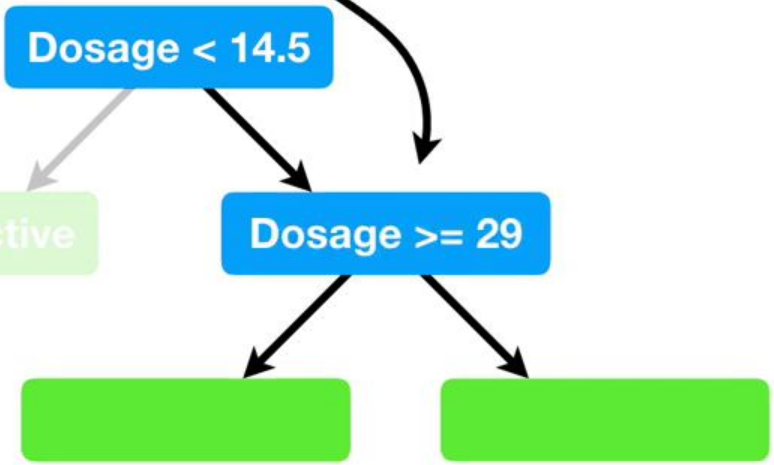
Since we have more than **7** observations on the right side (with **Dosage >= 14.5**), we can split them into two groups…



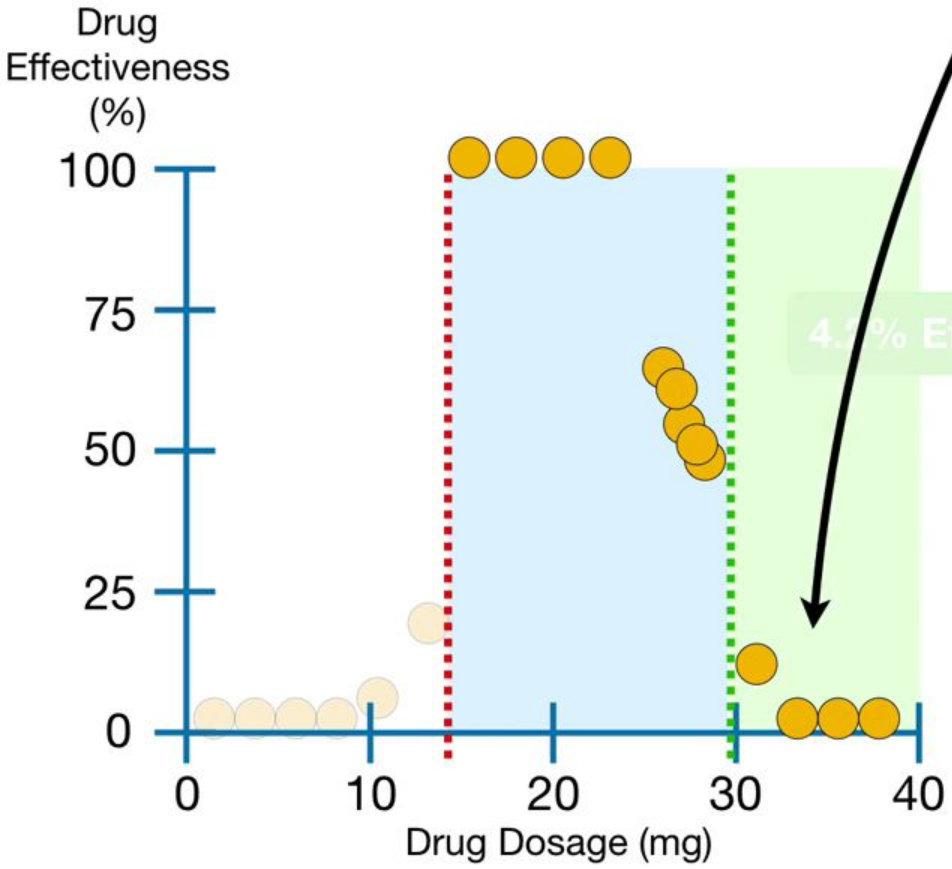


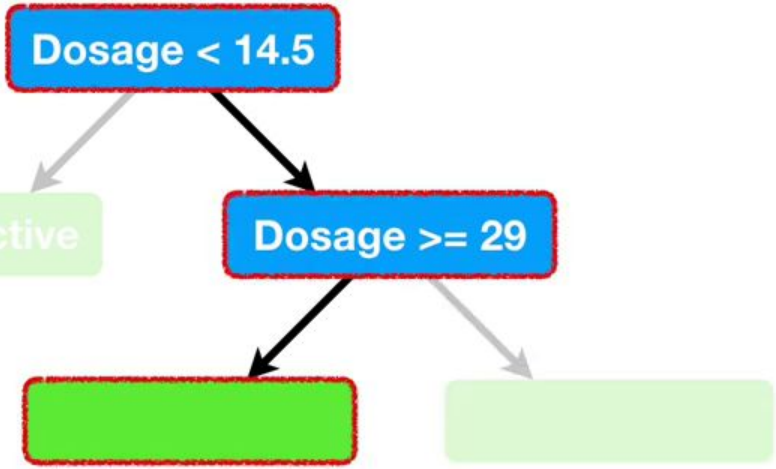
…and we do that by finding the threshold that gives us the smallest sum of squared residuals.



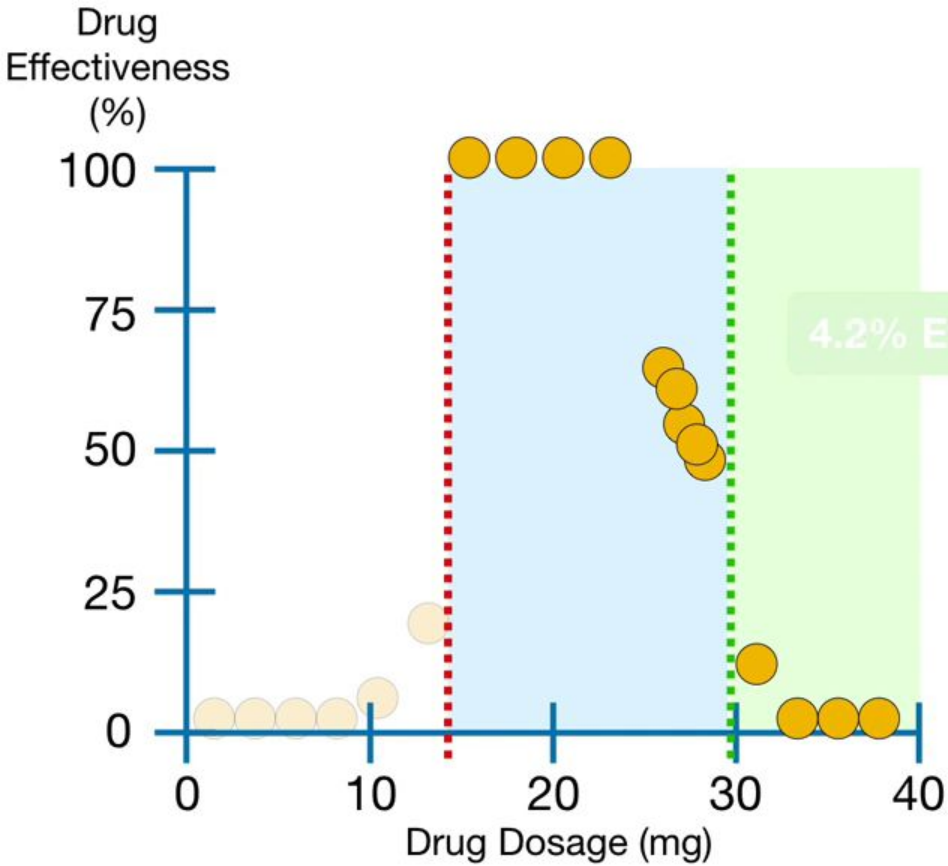


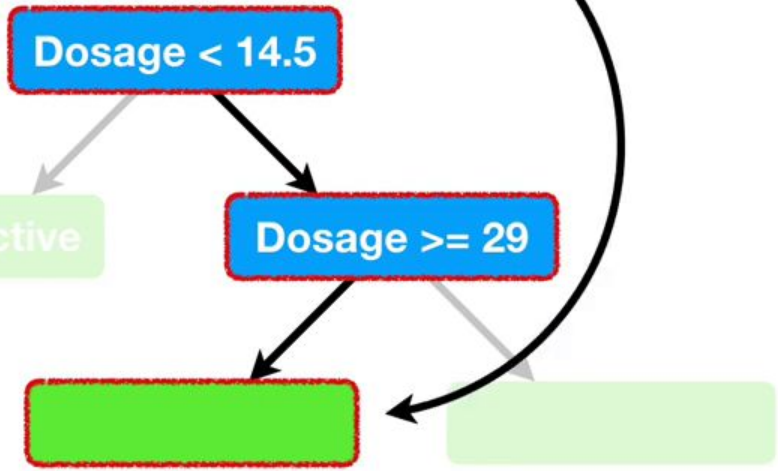
NOTE: There are only **4** observations with **Dosage >= 29**…



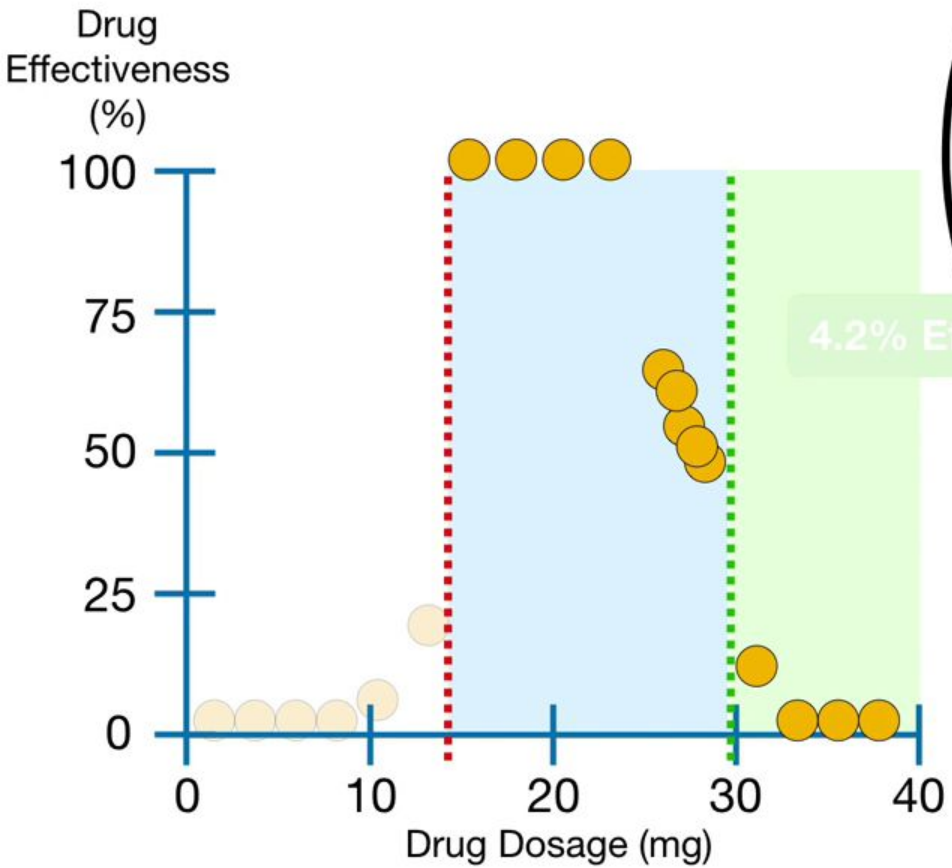


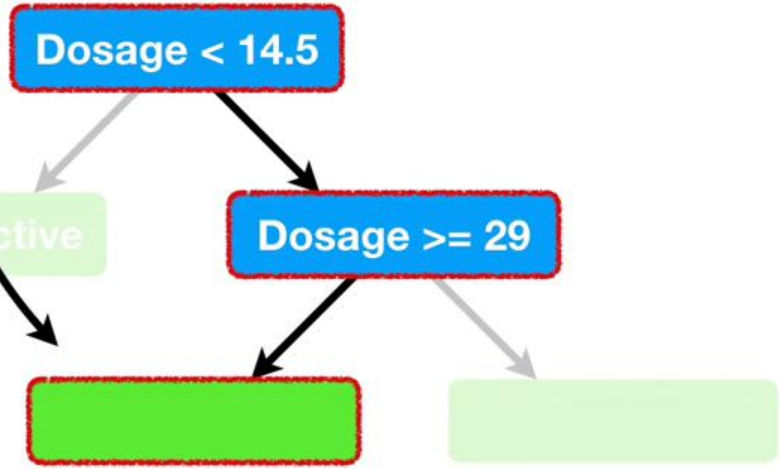
…thus, there are only **4** observations in this node…



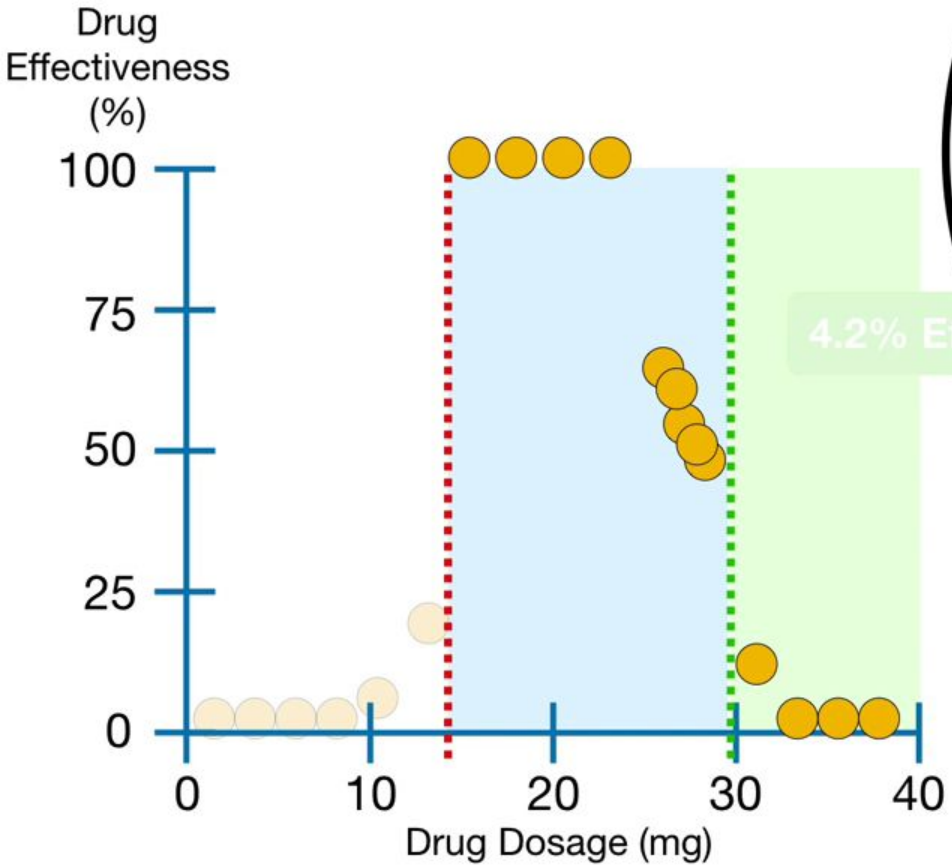


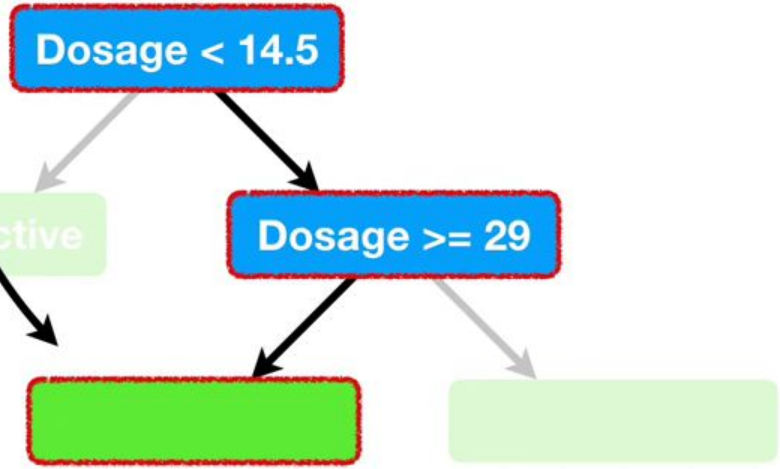
…thus, we will make this a leaf because it contains fewer than **7** observations…

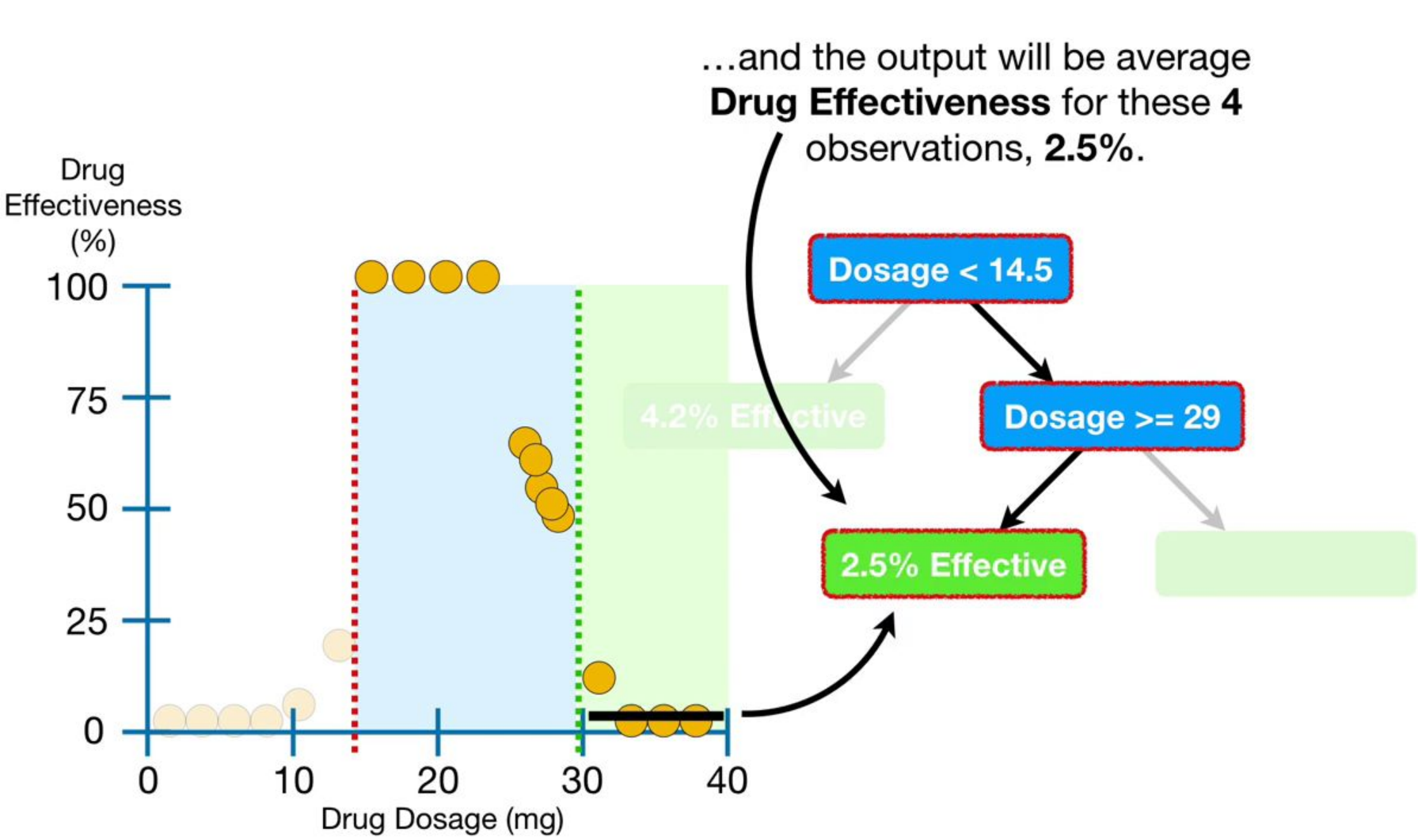




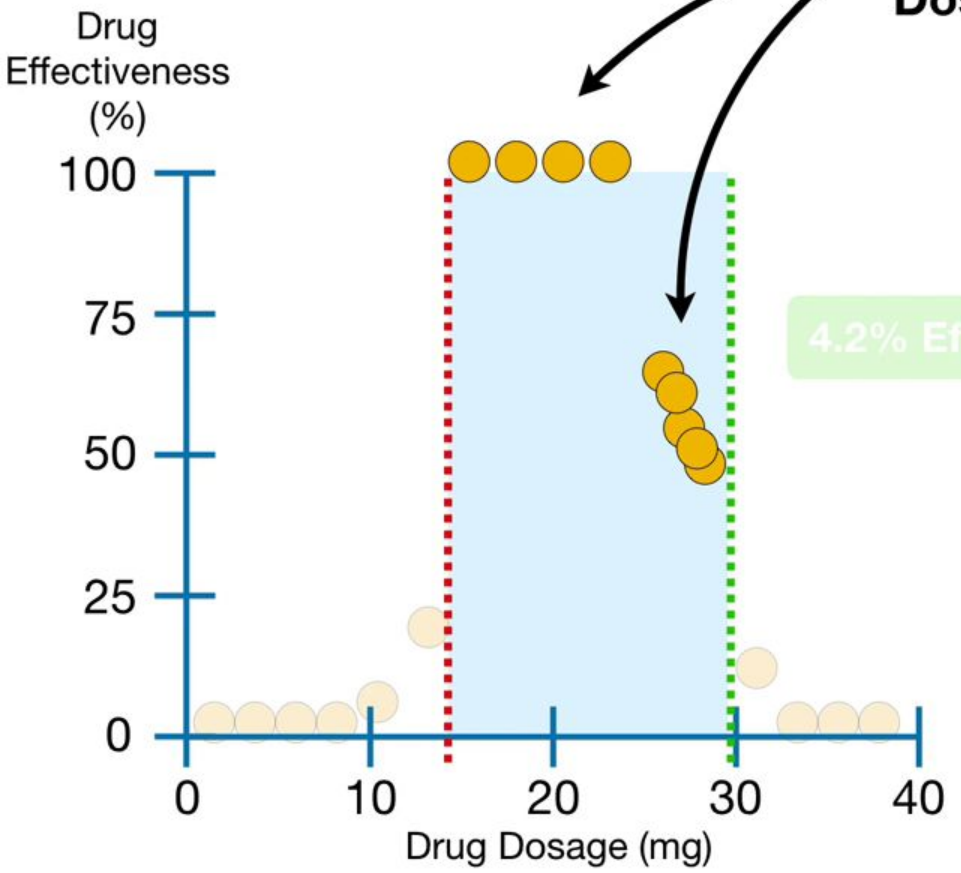
…and the output will be average **Drug Effectiveness** for these **4** observations, **2.5%**.

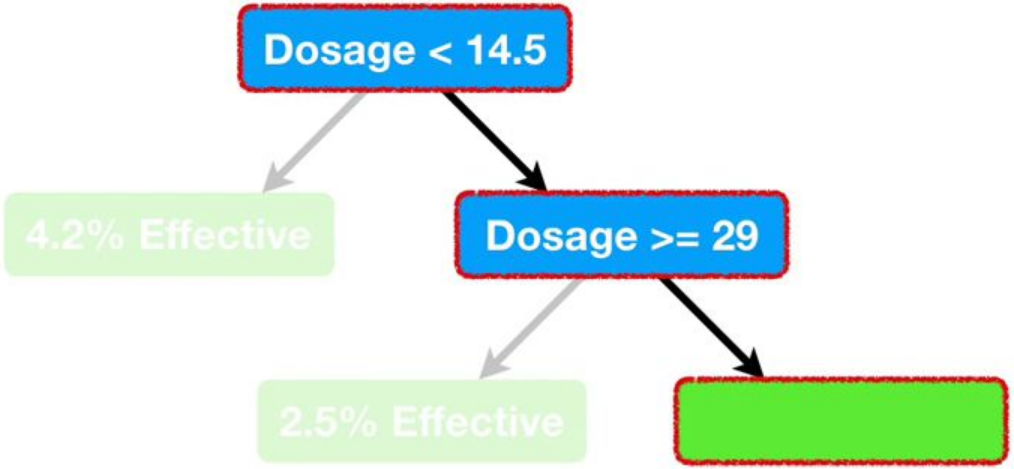


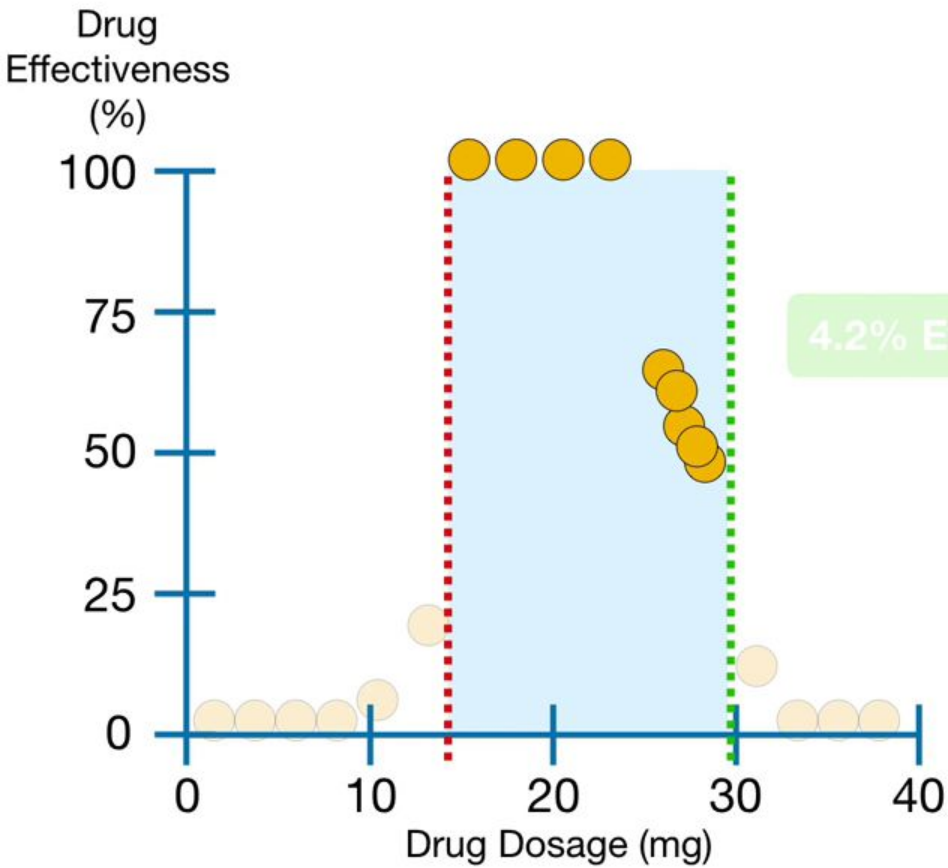



…and the output will be average
Drug Effectiveness for these 4
observations, 2.5%.
Drug
Effectiveness
(%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
4.2% Effective
Dosage >= 29
2.5% Effective

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







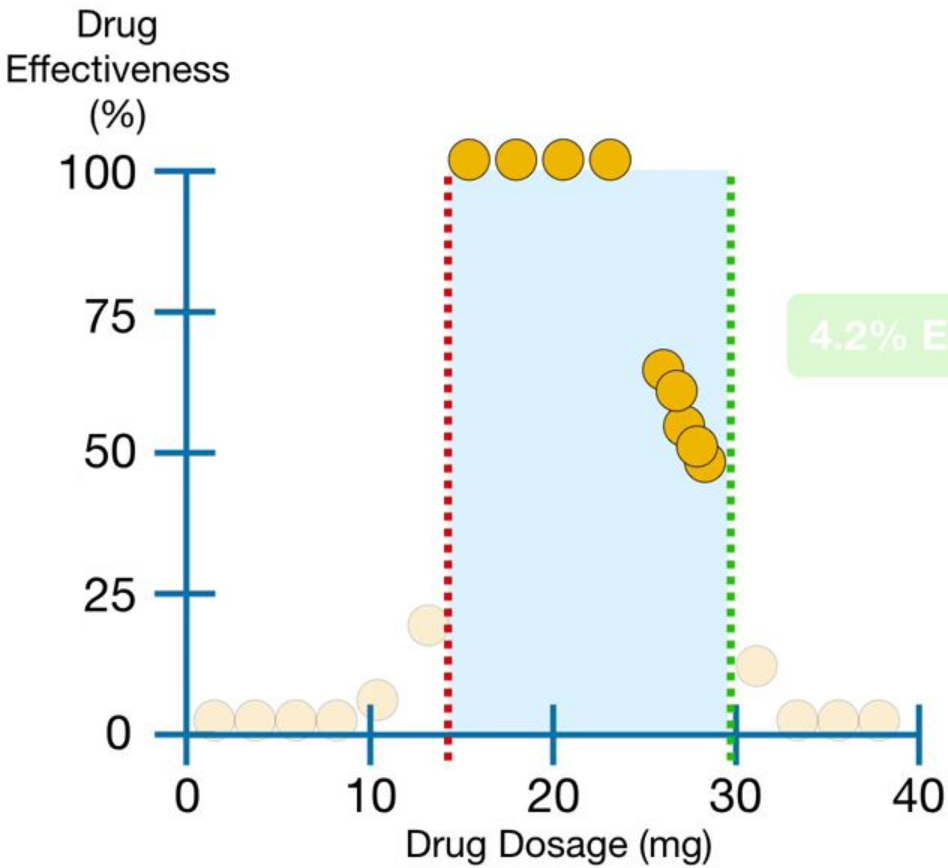
Since we have more than **7** observations, we can split them into two groups…

Dosage < 14.5

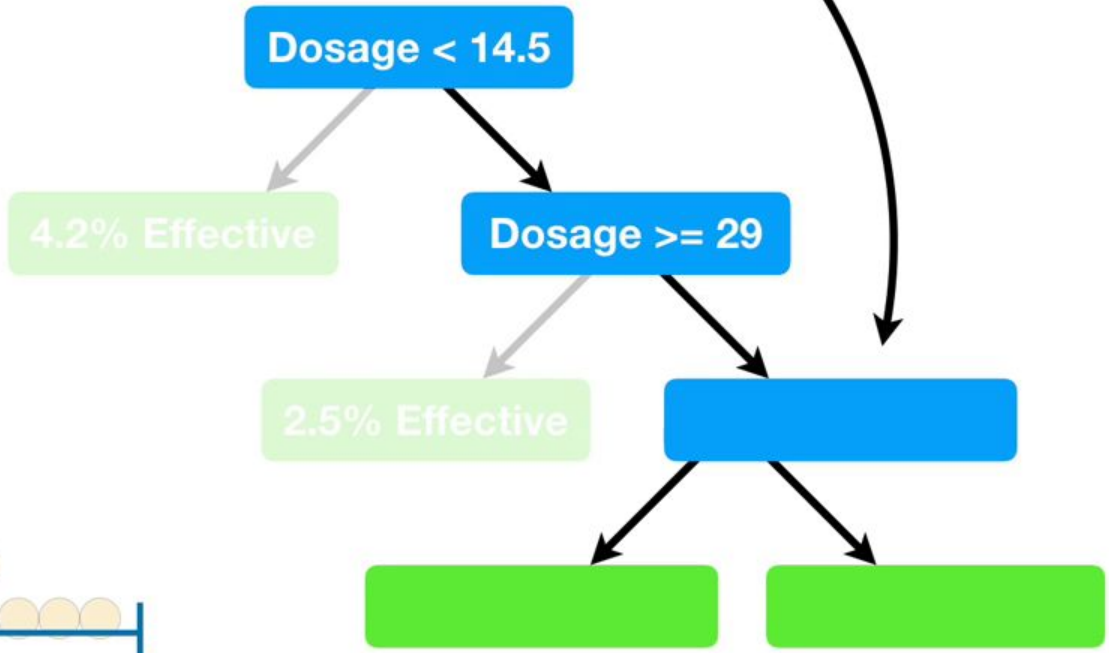
4.2% Effective

Dosage >= 29

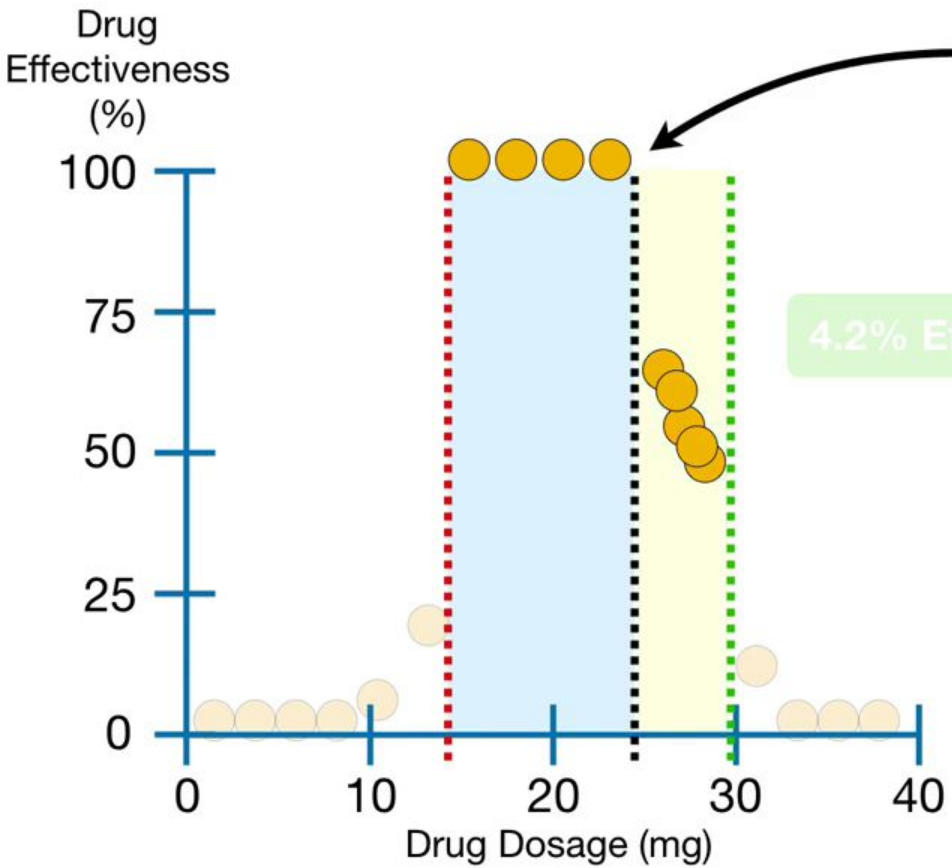
2.5% Effective

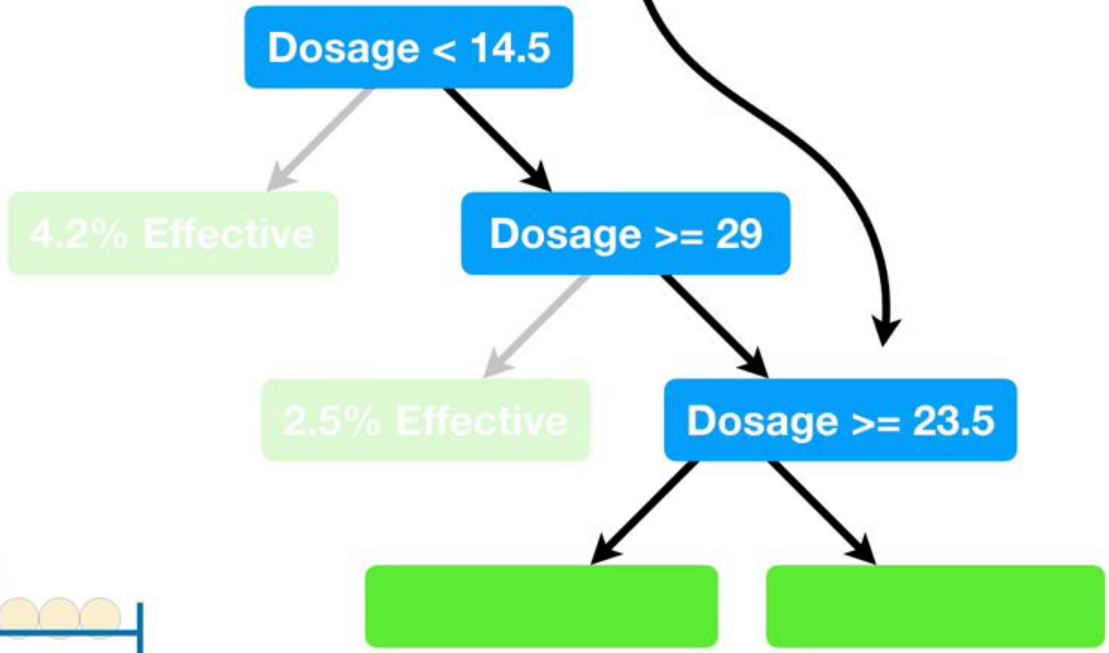


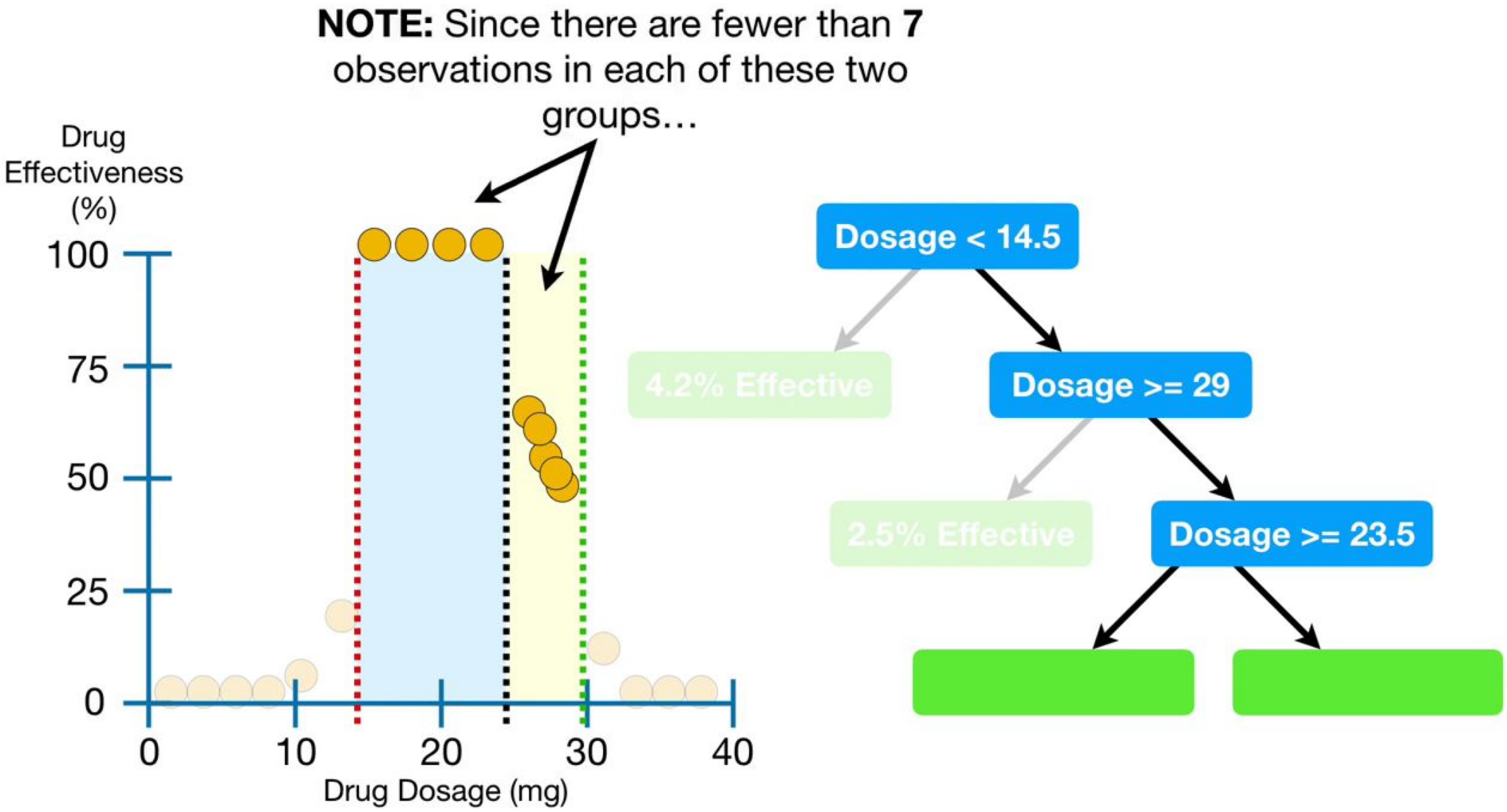
Since we have more than **7** observations, we can split them into two groups…

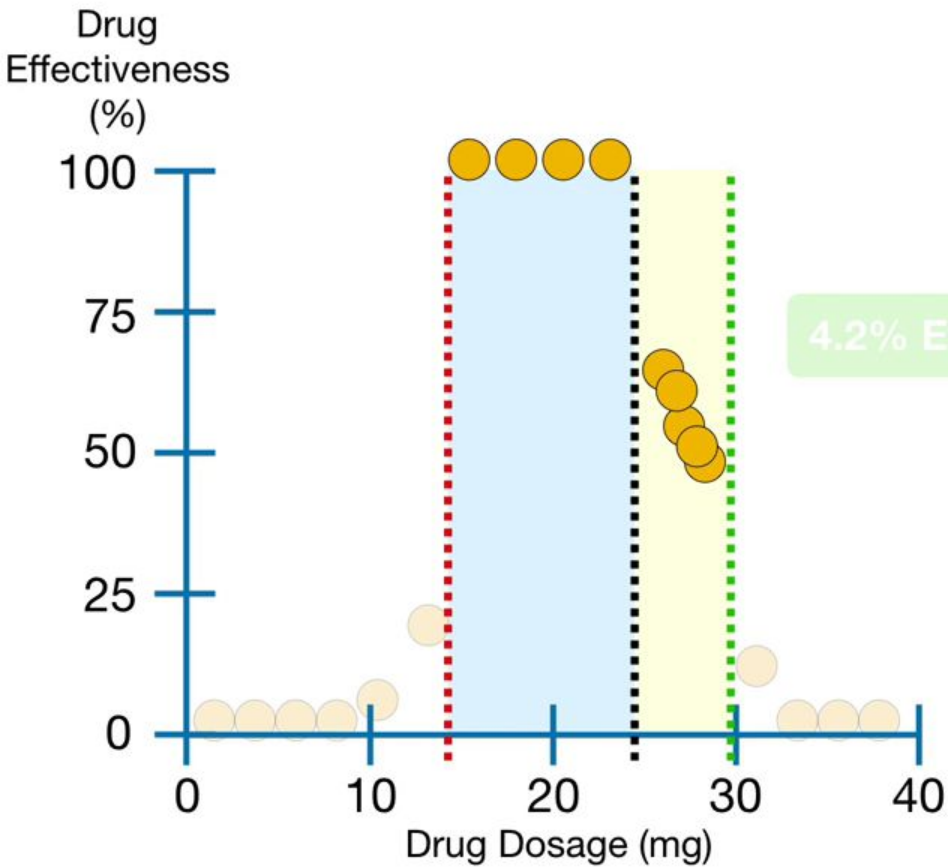


…by finding the threshold that gives us the minimum sum of squared residuals.

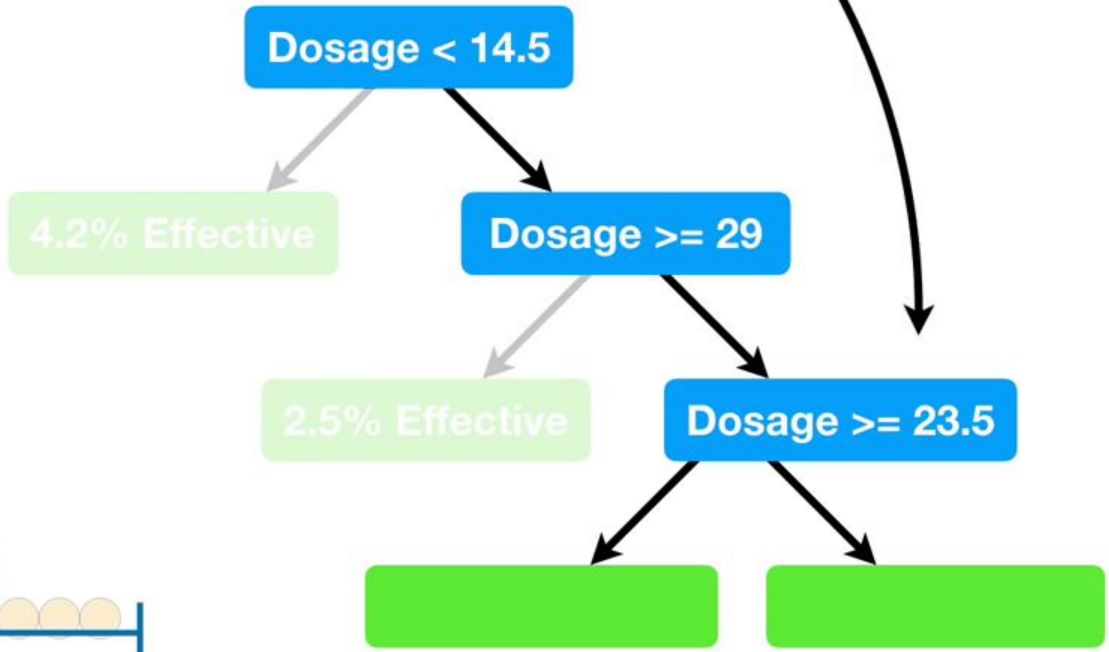




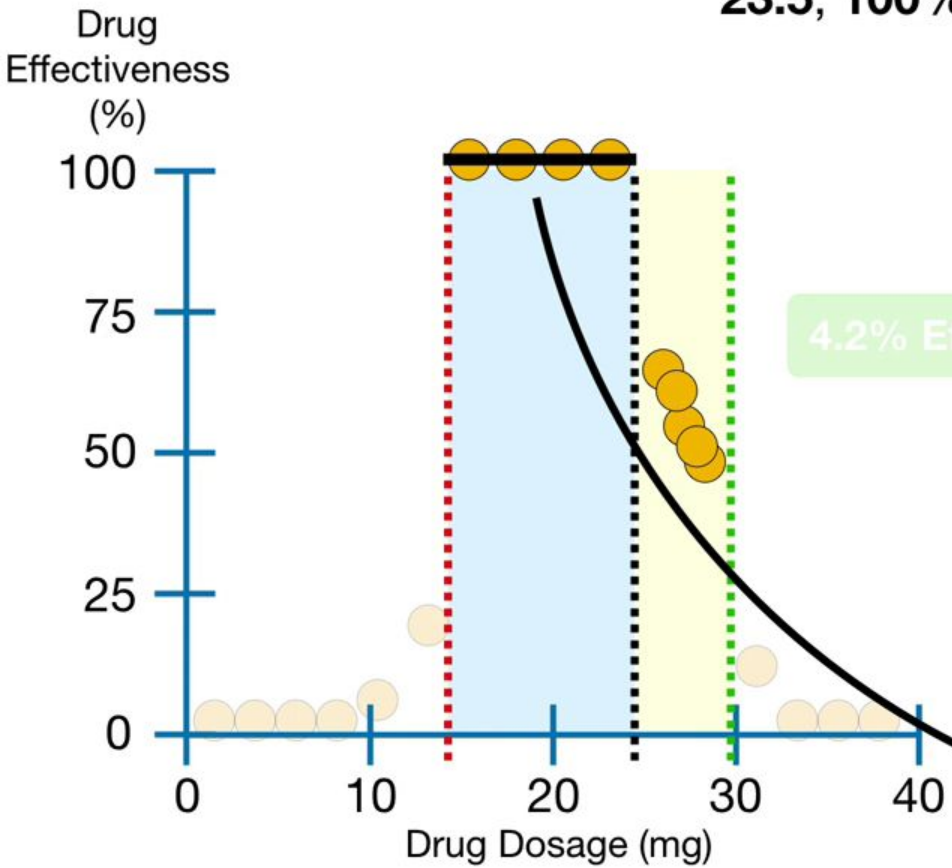

NOTE: Since there are fewer than 7 observations in each of these two groups…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
4.2% Effective
Dosage >= 29
2.5% Effective
Dosage >= 23.5

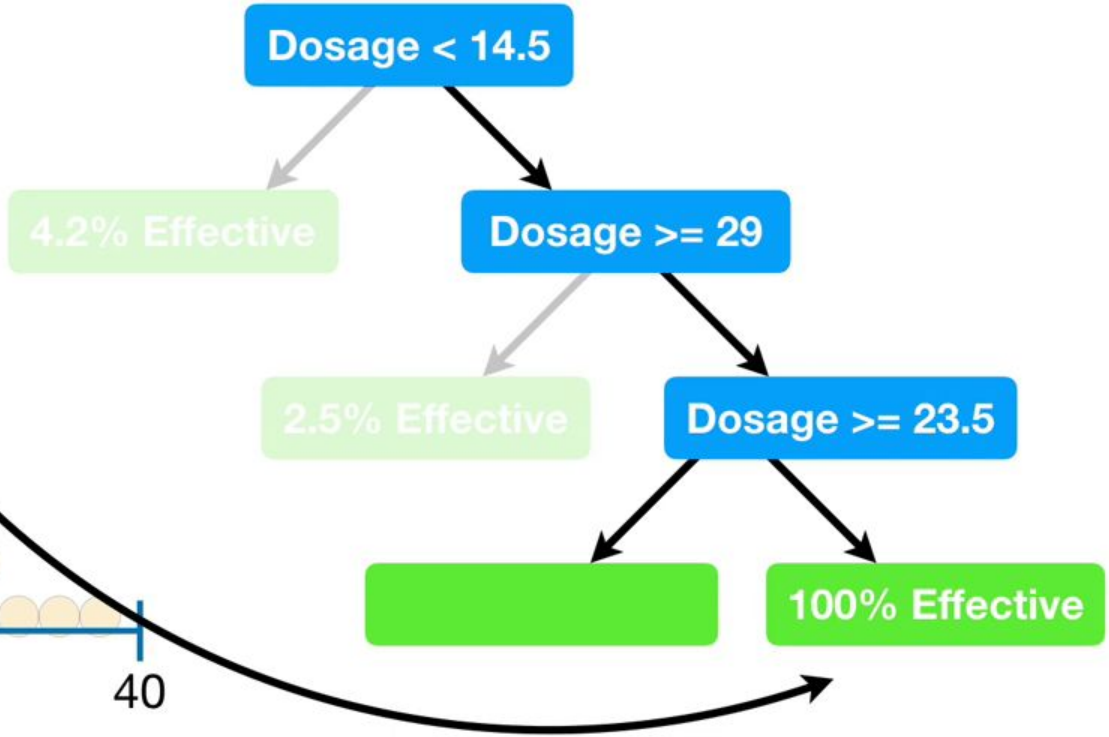


….this is the last split, because none of the leaves have more than **7** observations in them.

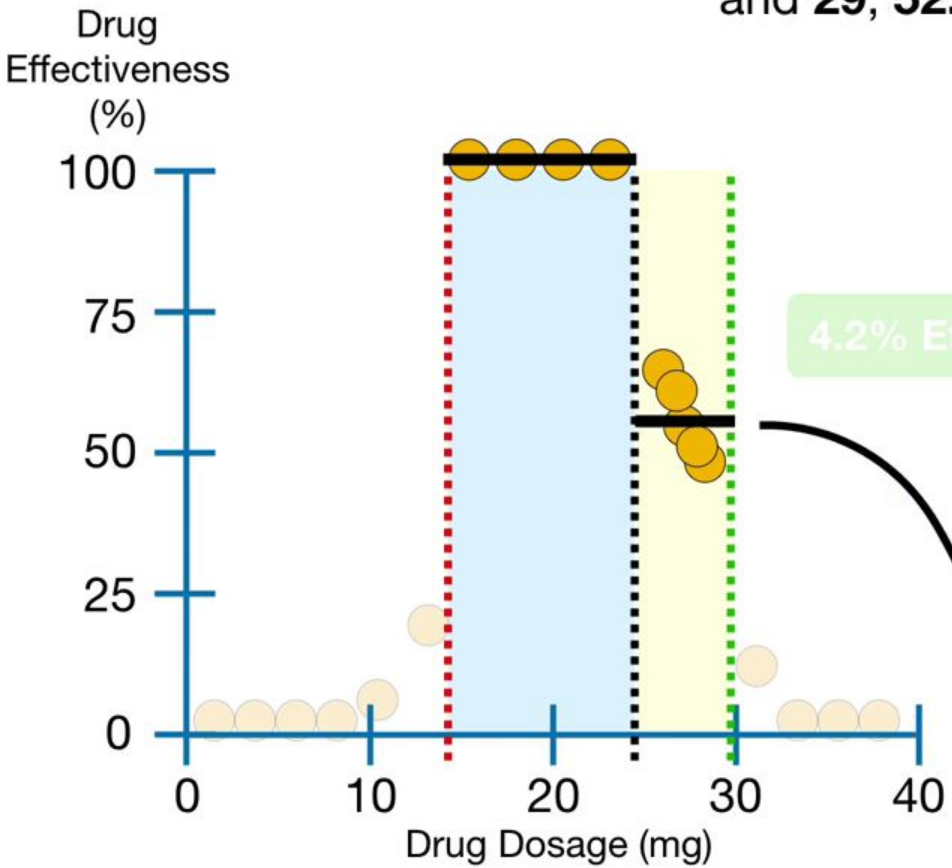


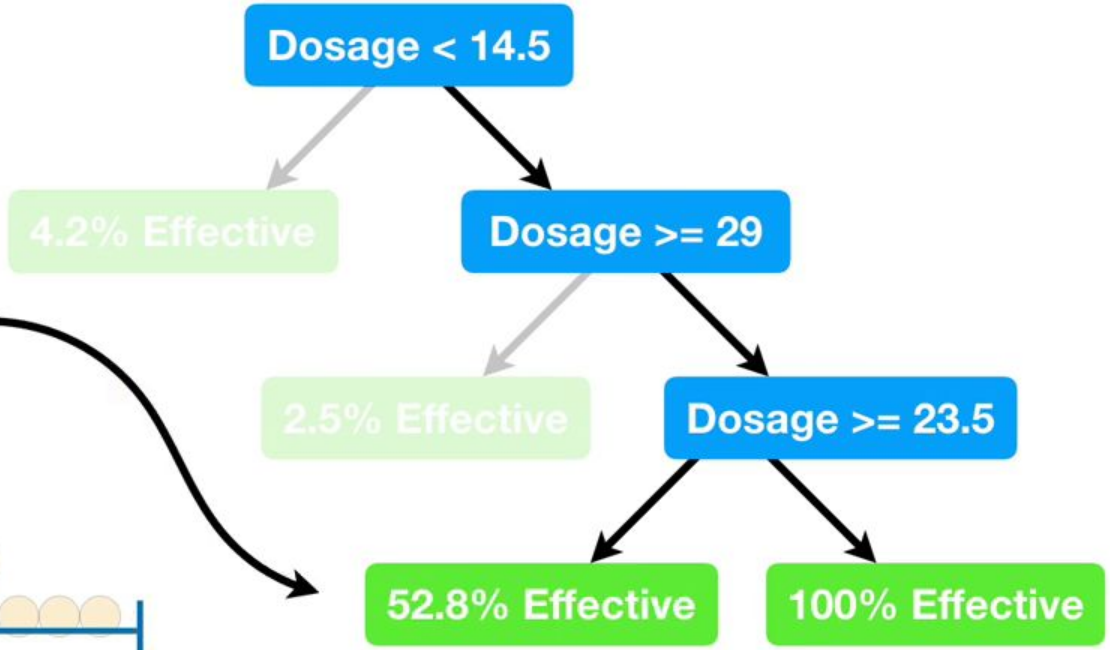
So we use the average **Drug Effectiveness** for observations with **Dosages** between **14.5** and **23.5**, **100%**, as the output for leaf on the right…



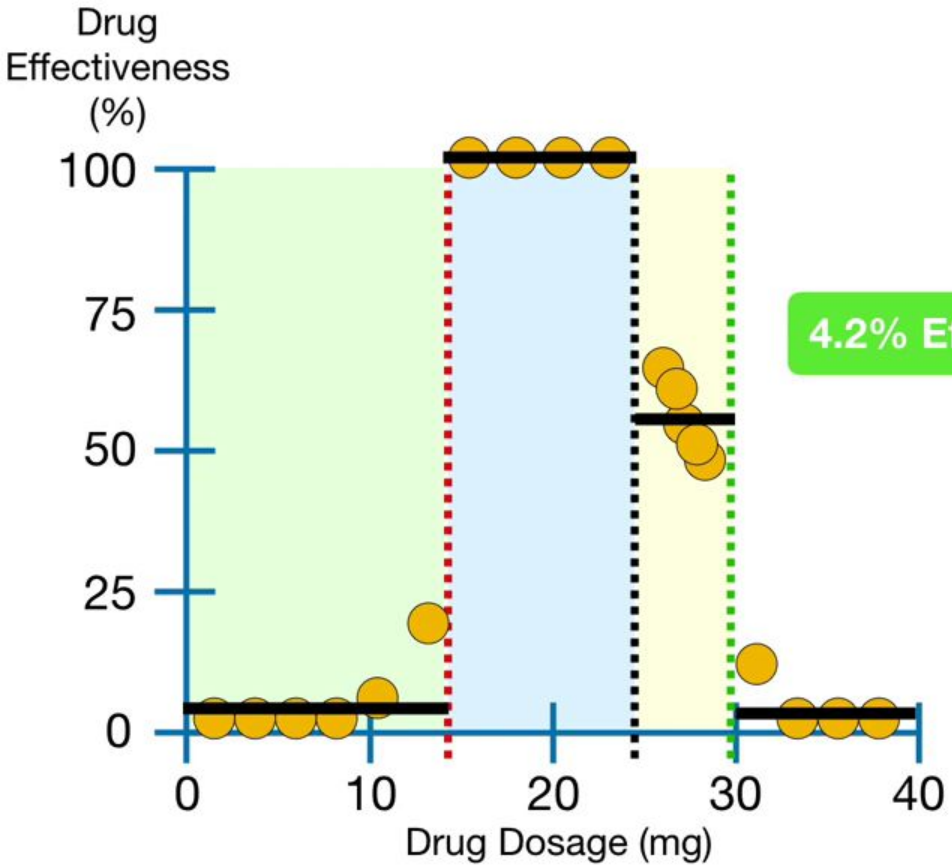


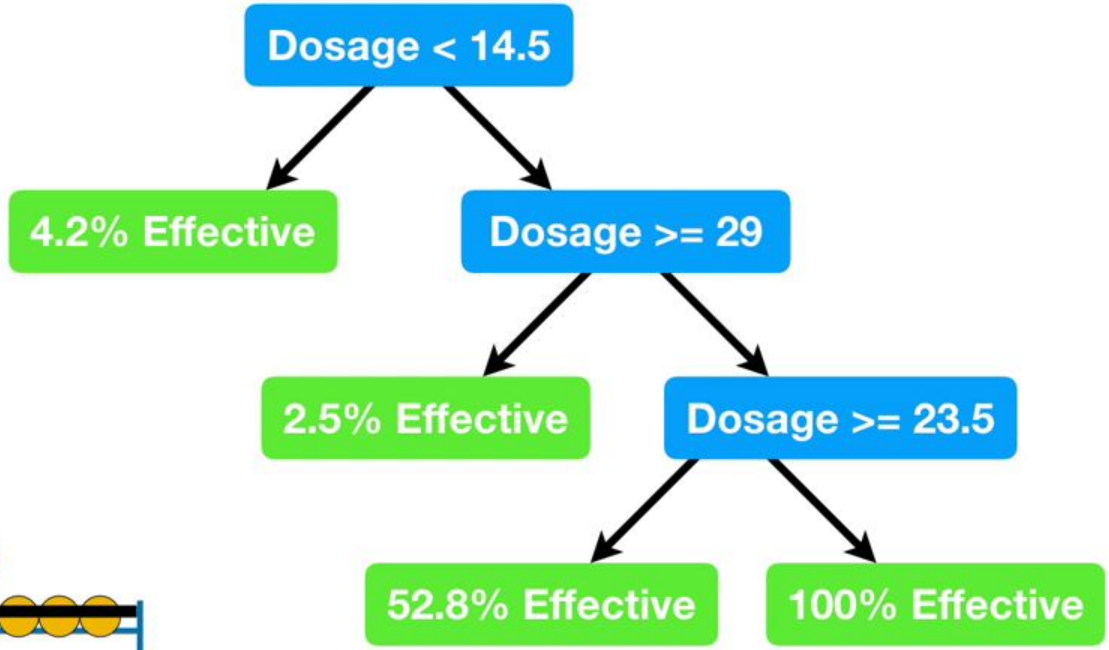
…and we use the average **Drug Effectiveness** for observations with **Dosages** between **23.5** and **29**, **52.8%**, as the output for leaf on the left.



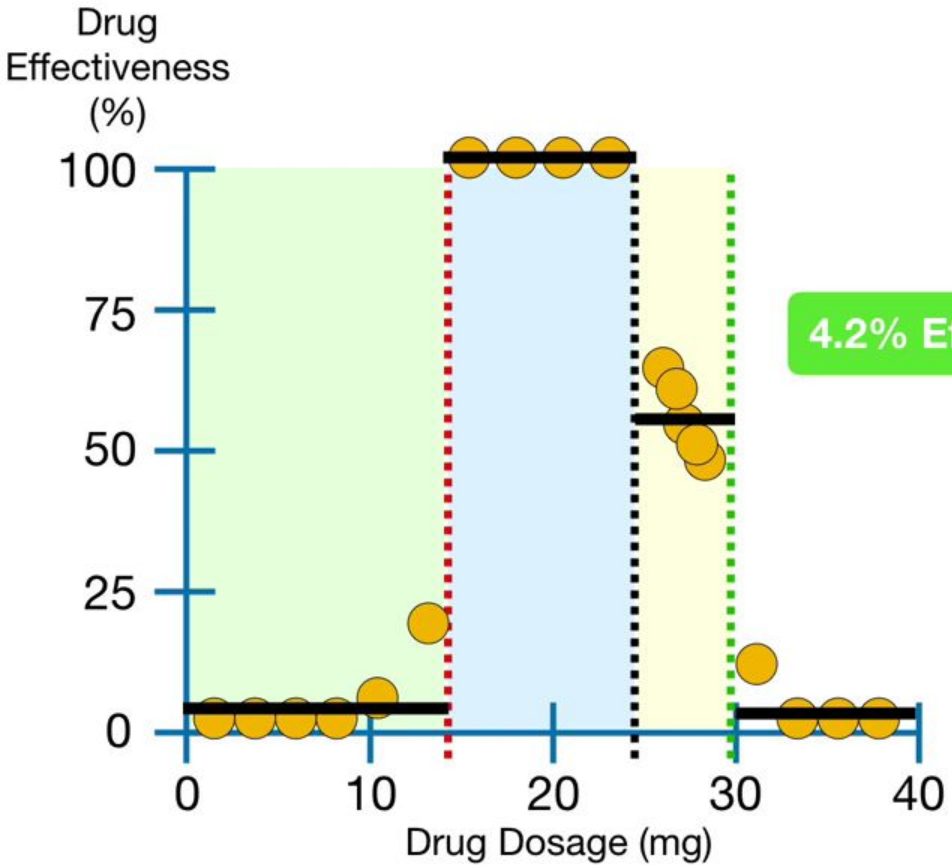


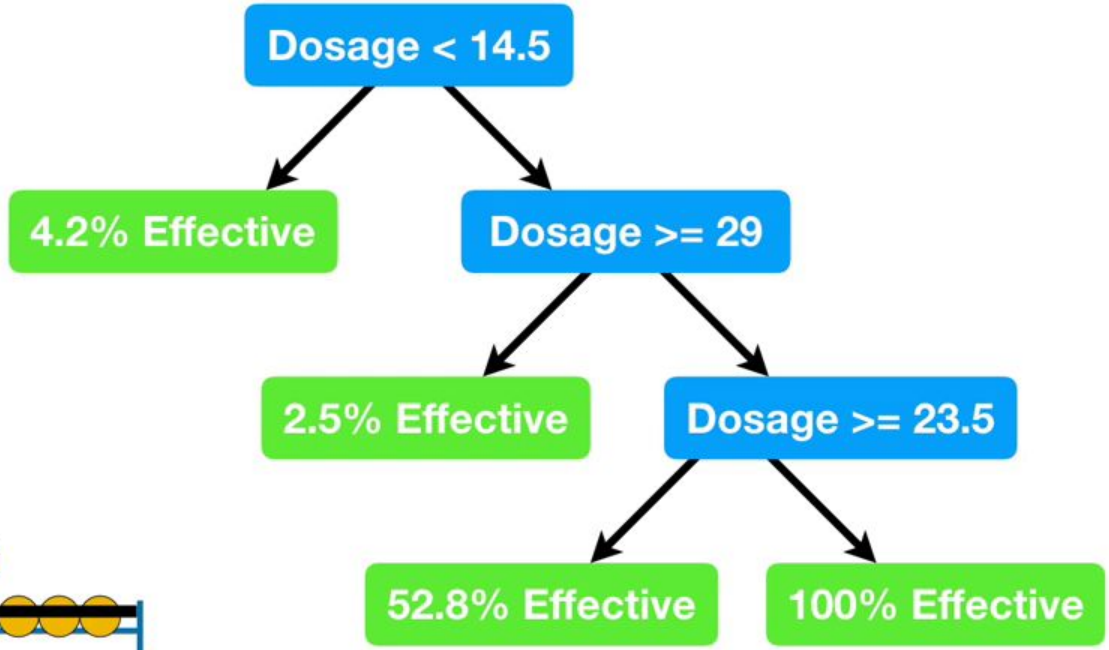
Since no leaf has more than **7** observations in it,…



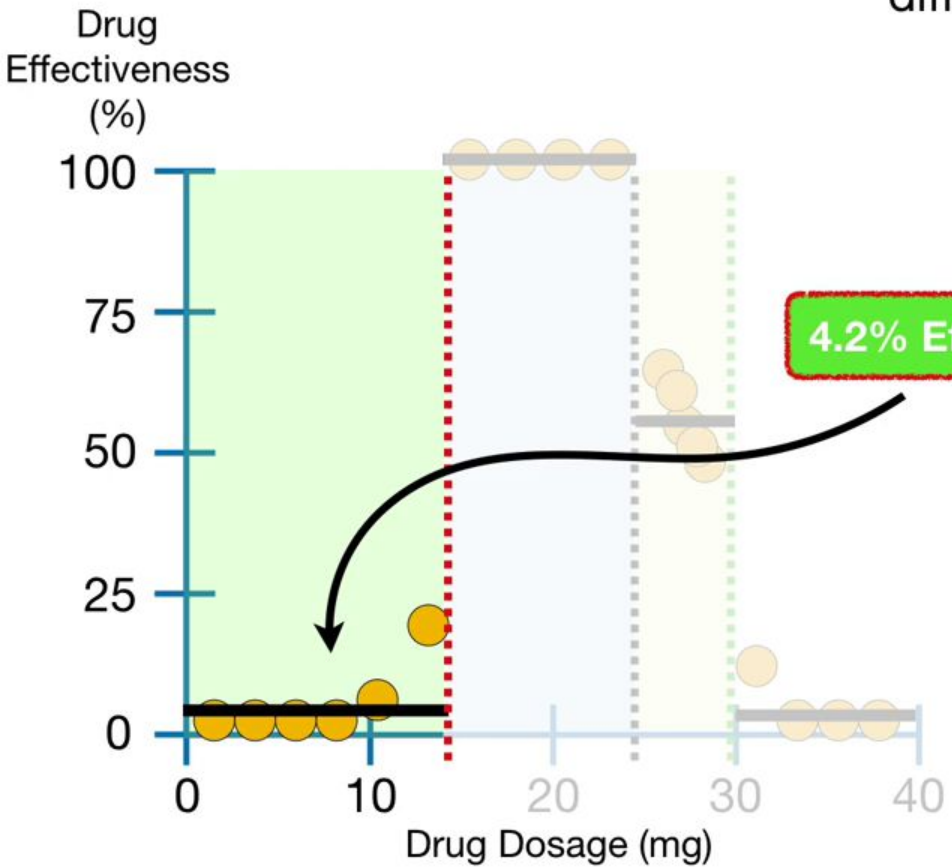


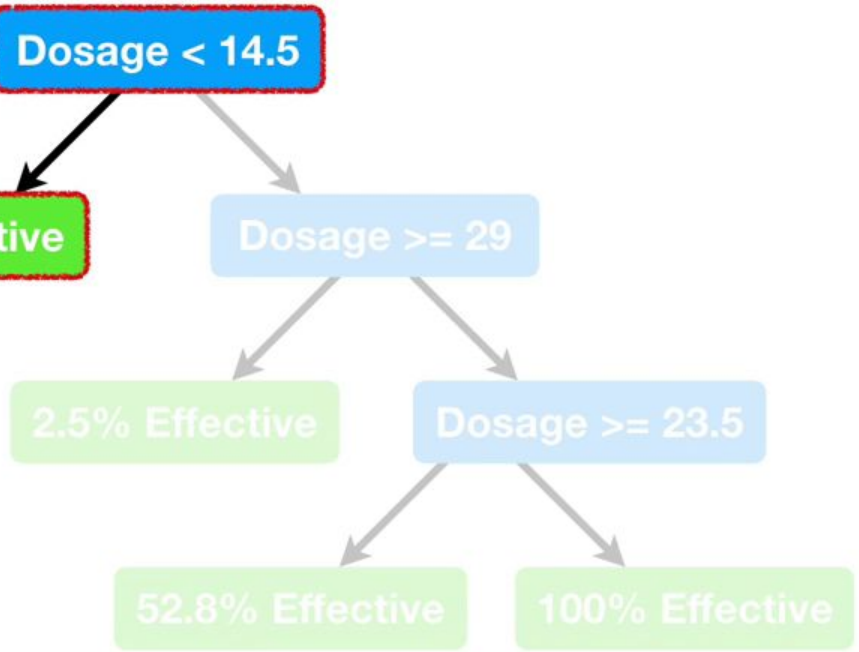
…we're done building the tree…



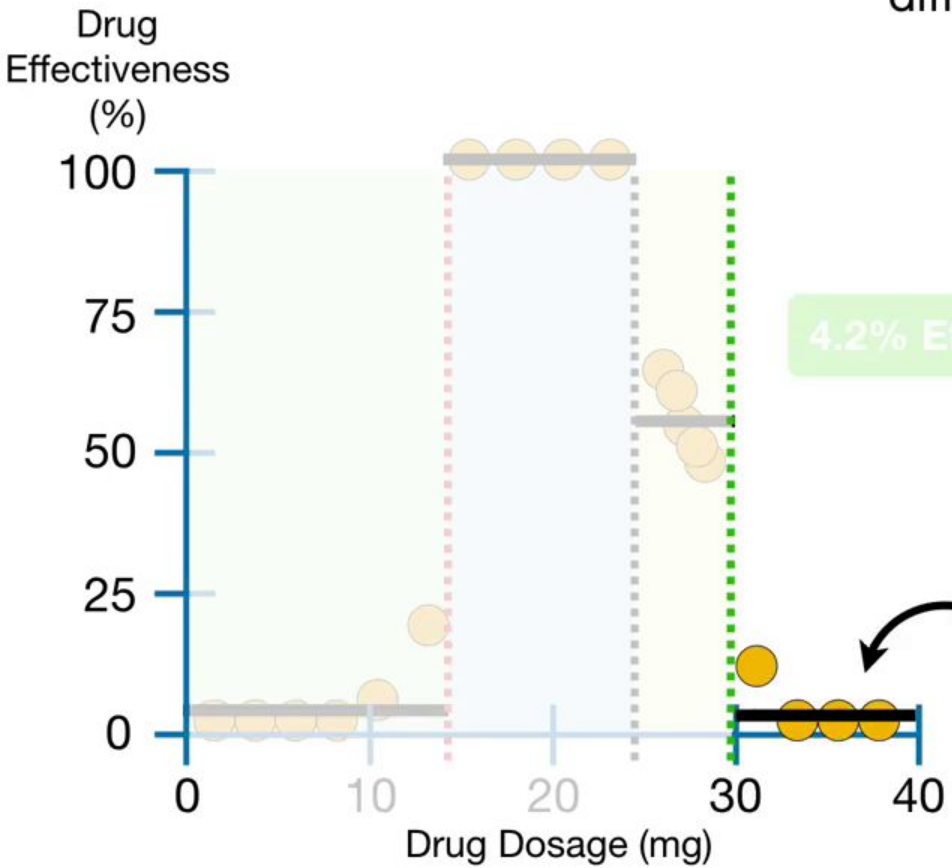


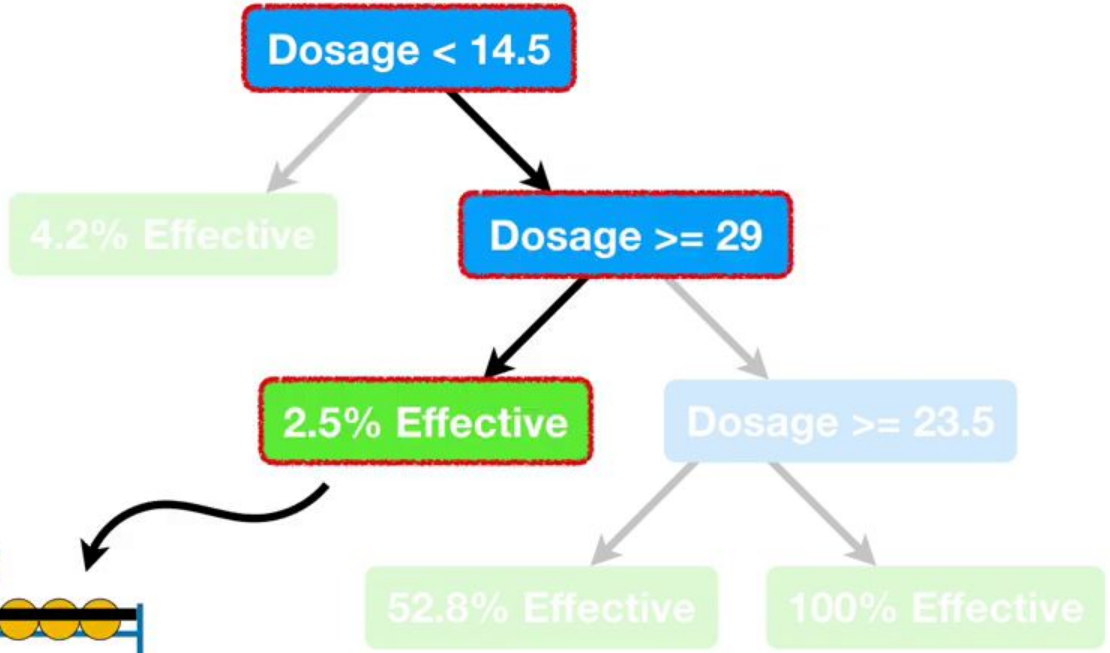
…and each leaf corresponds to the average **Drug Effectiveness** from a different cluster of observations.



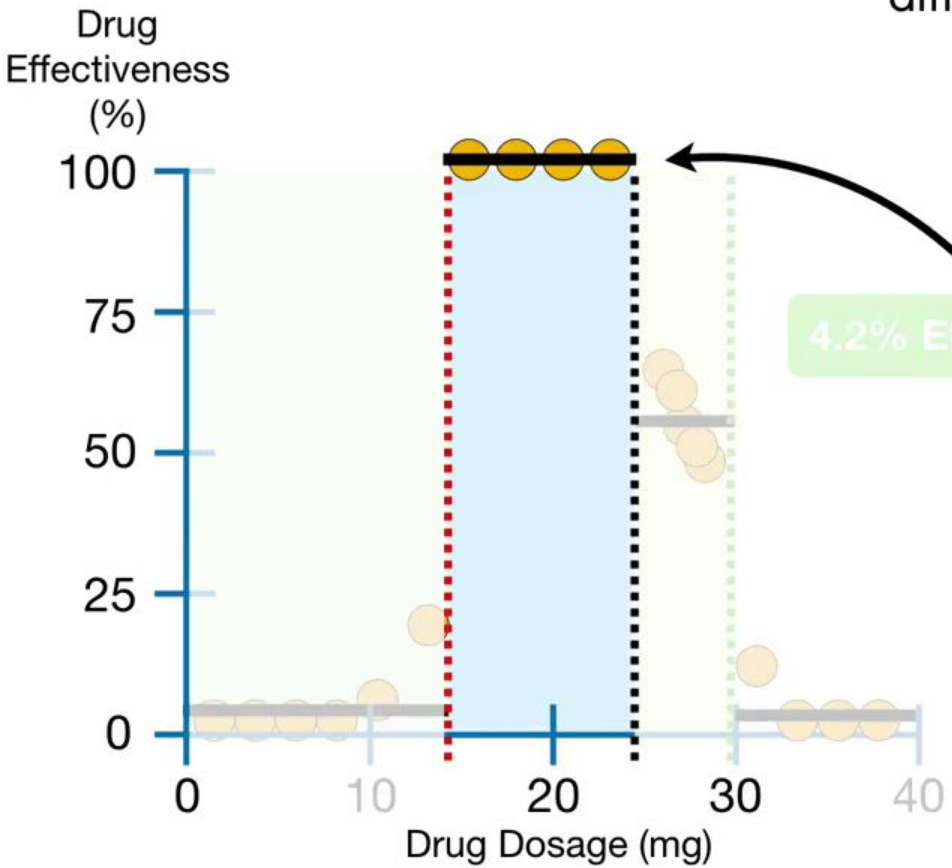


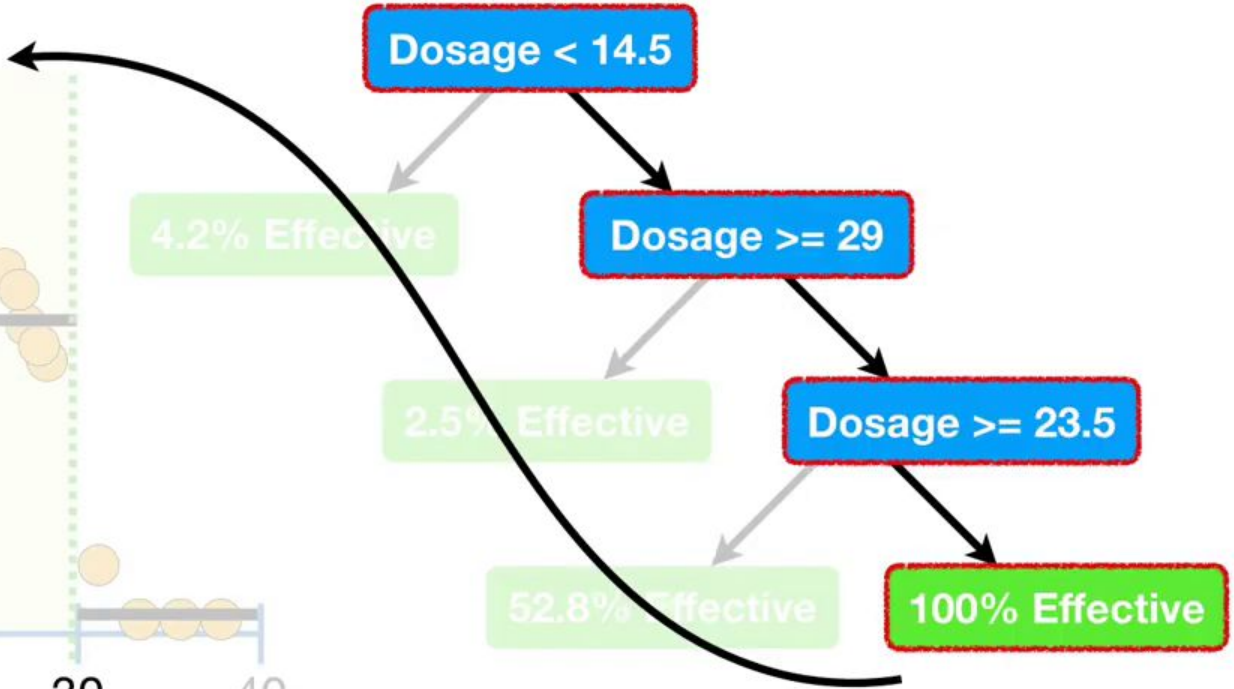
…and each leaf corresponds to the average **Drug Effectiveness** from a different cluster of observations.



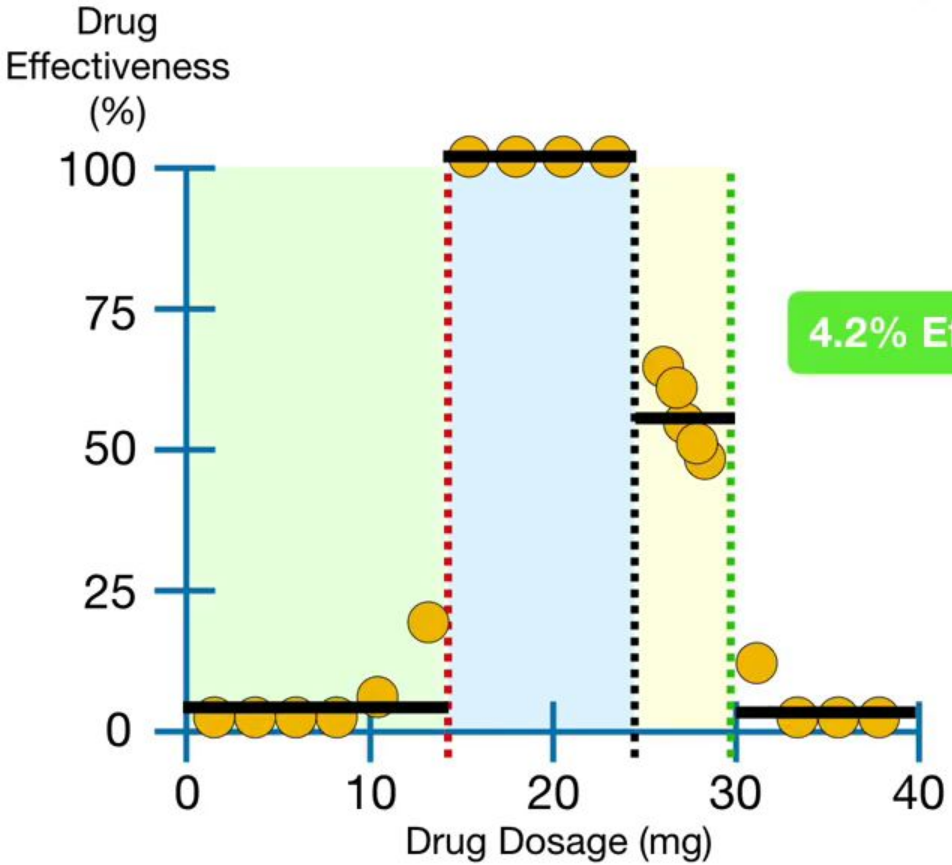


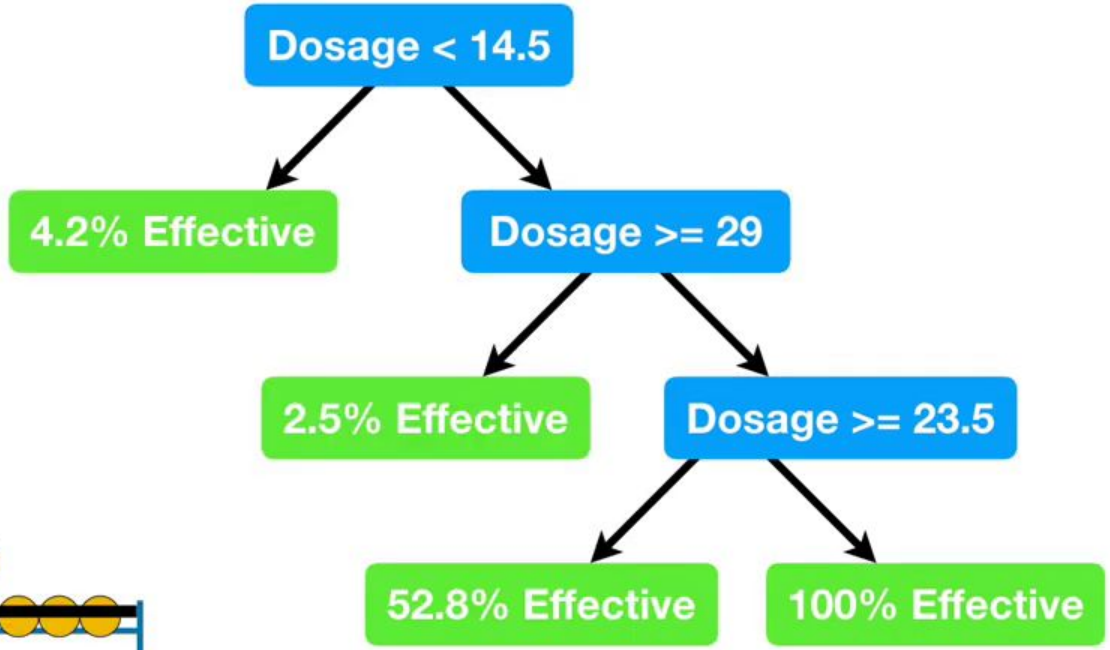
…and each leaf corresponds to the average **Drug Effectiveness** from a different cluster of observations.



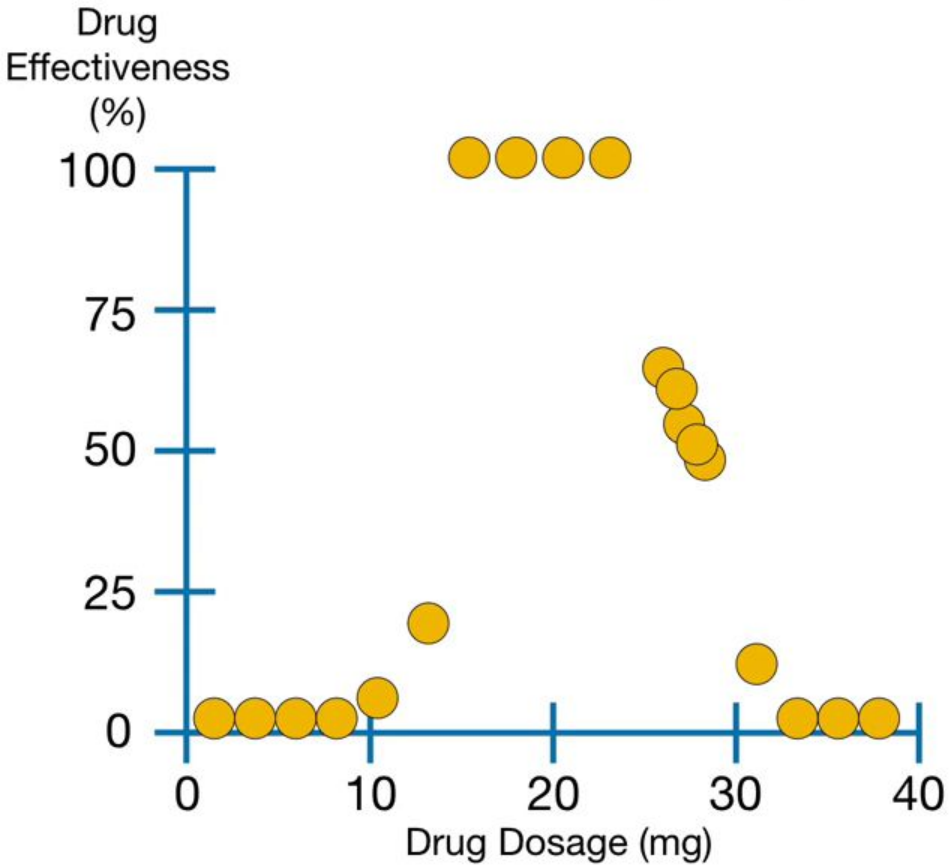


DOUBLE BAM!!!


Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)


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**.

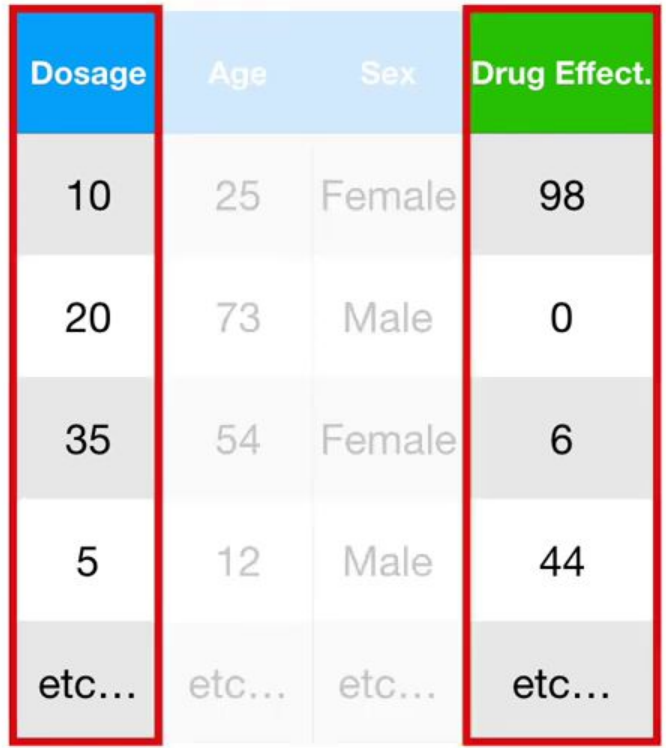


Dosage	Drug Effect.
10	58
20	60
35	57
5	44
etc…	etc…

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

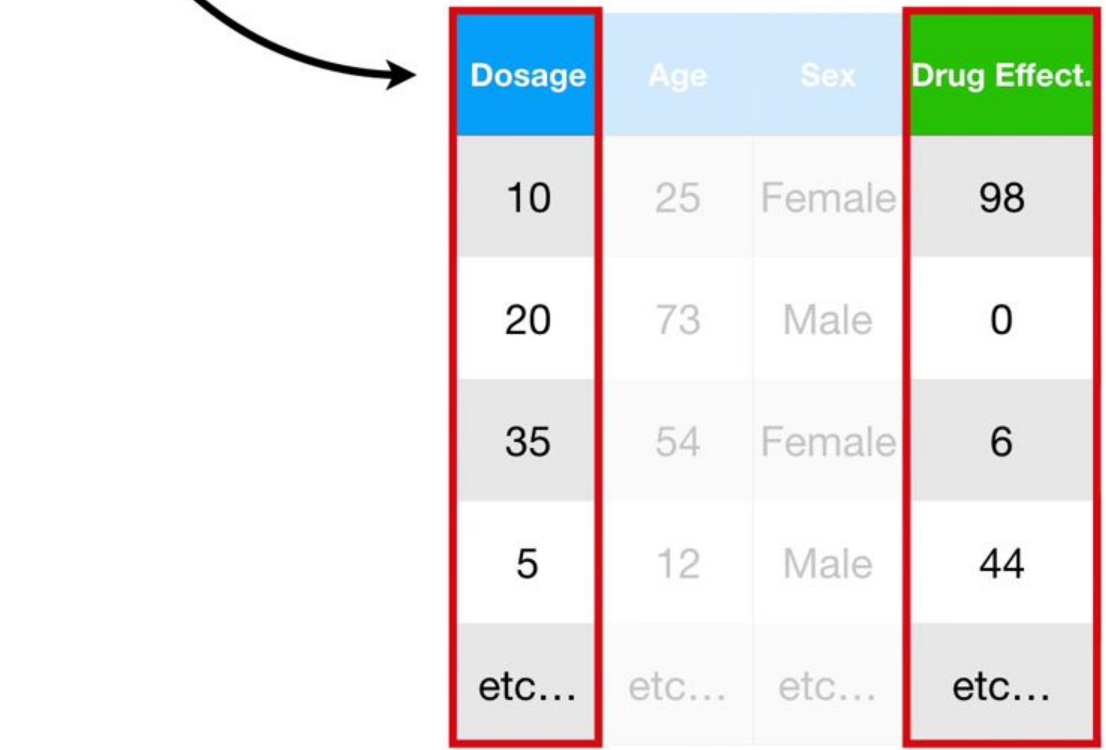
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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



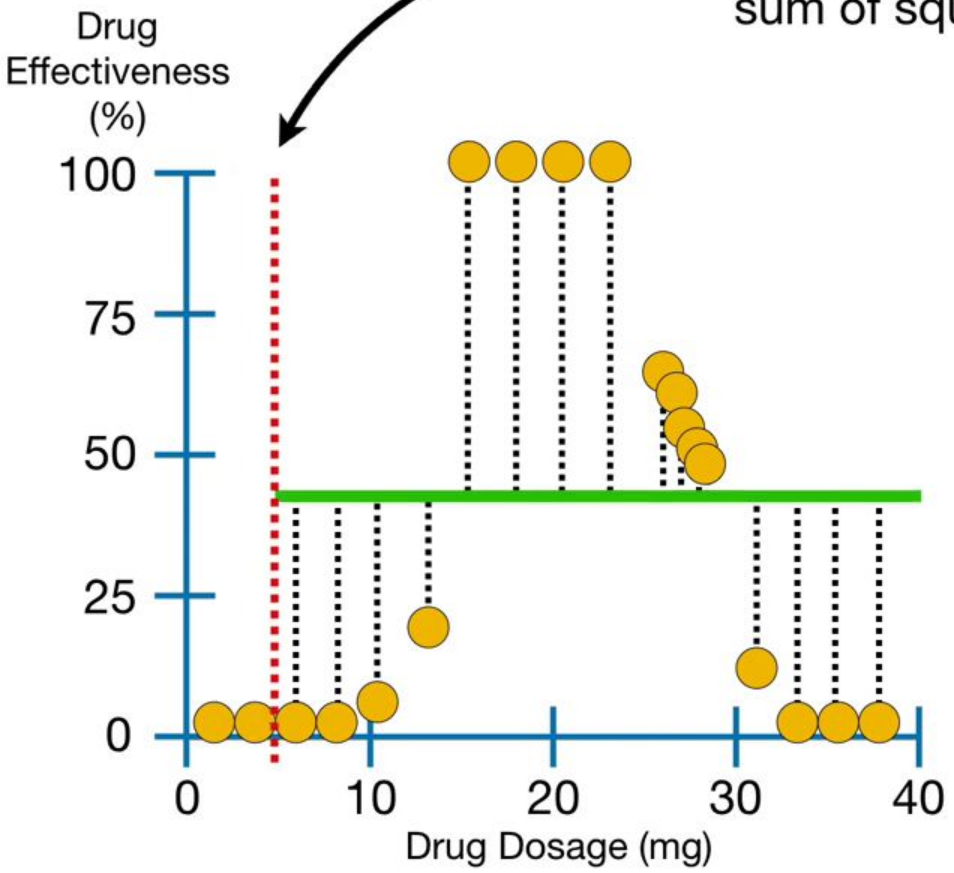
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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



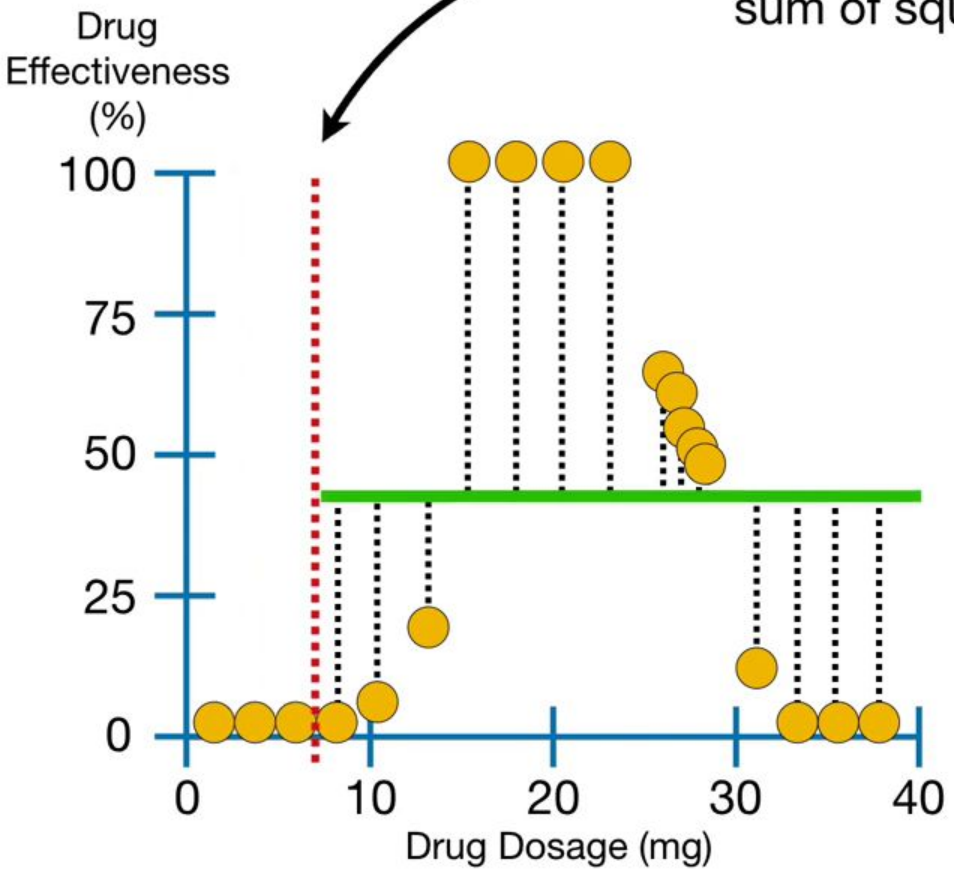
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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



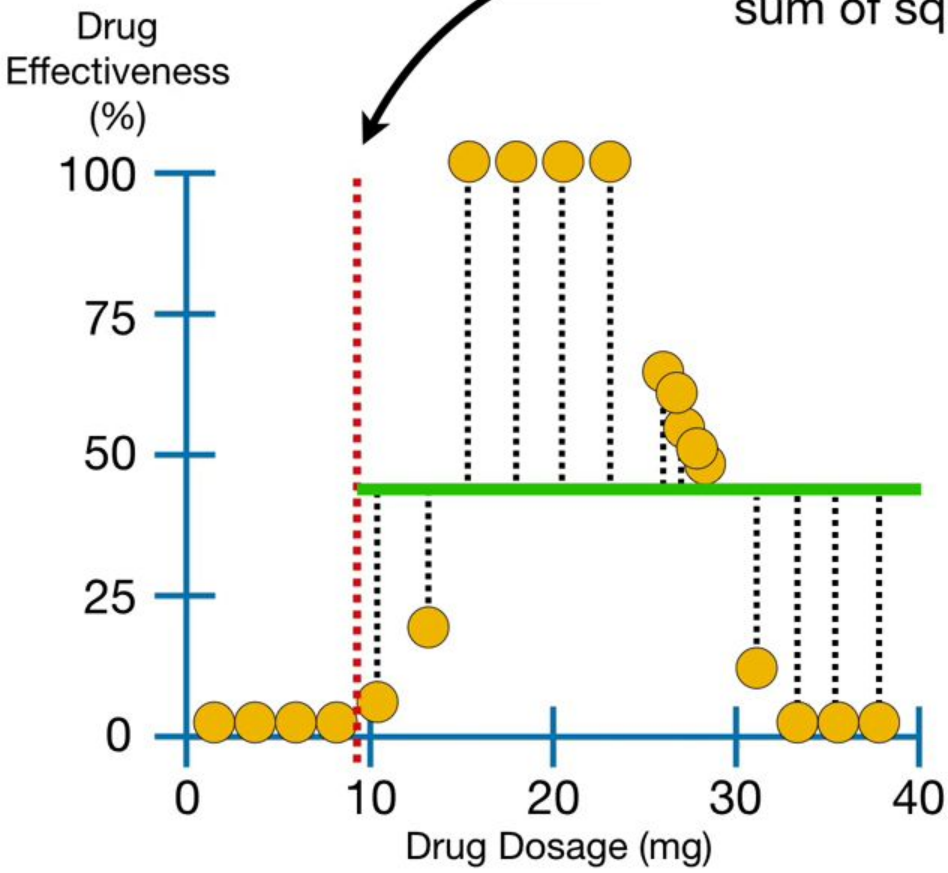
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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



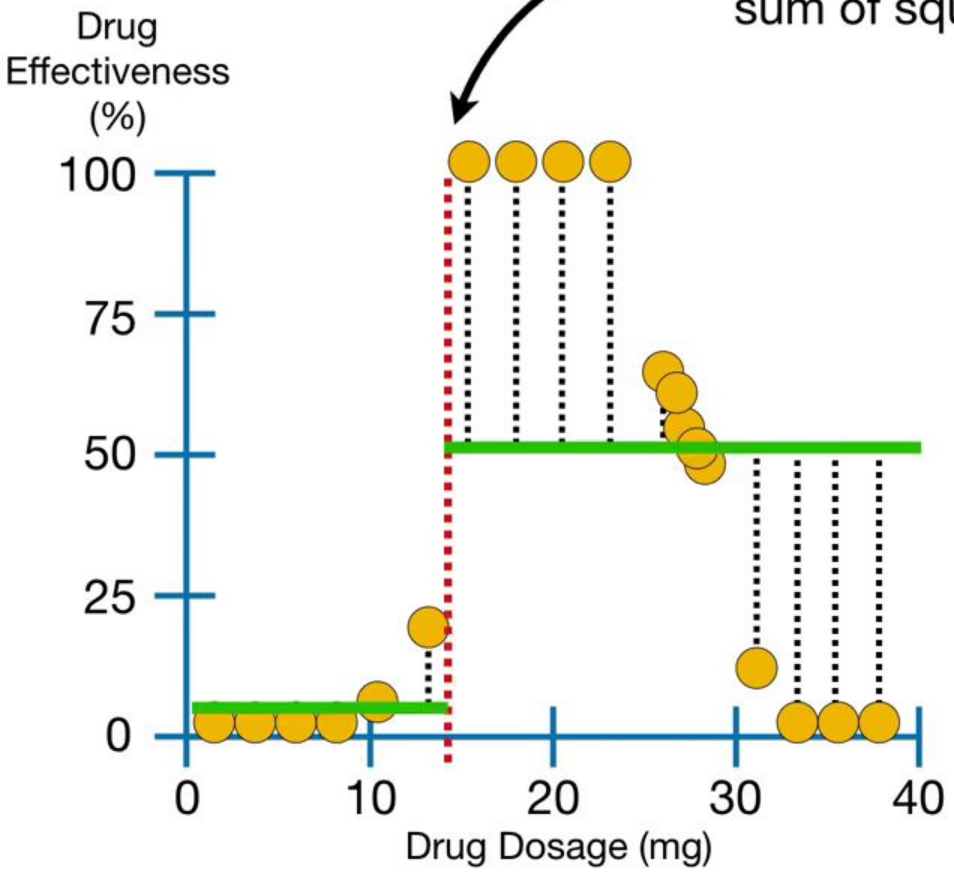
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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



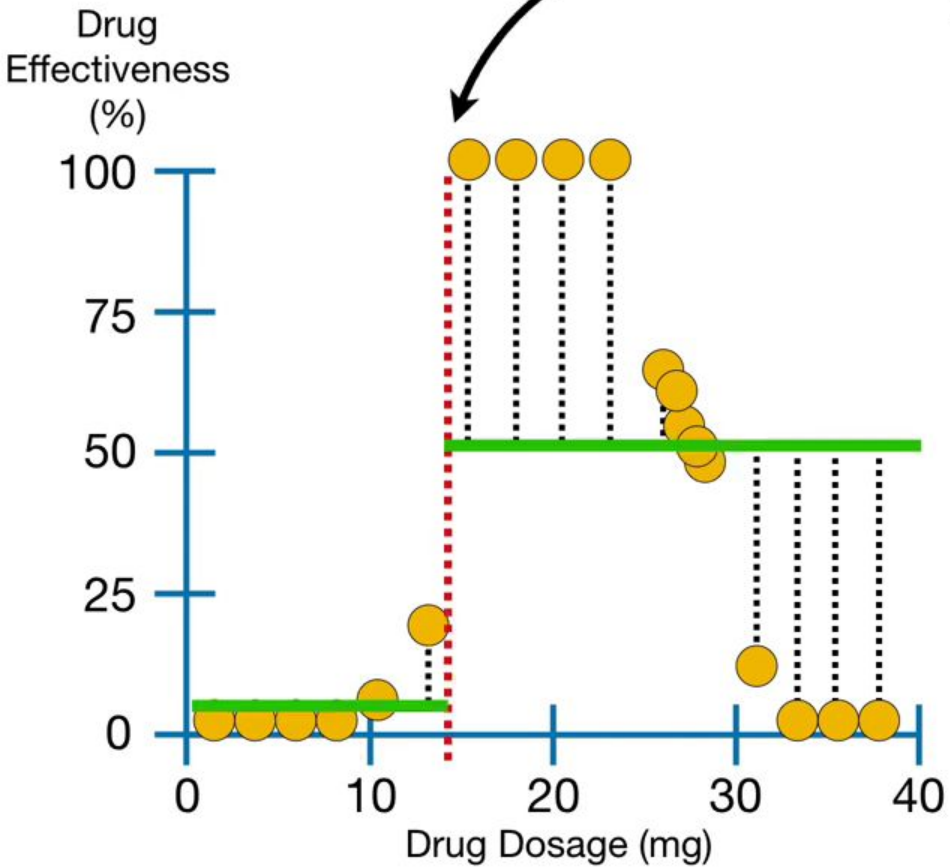
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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

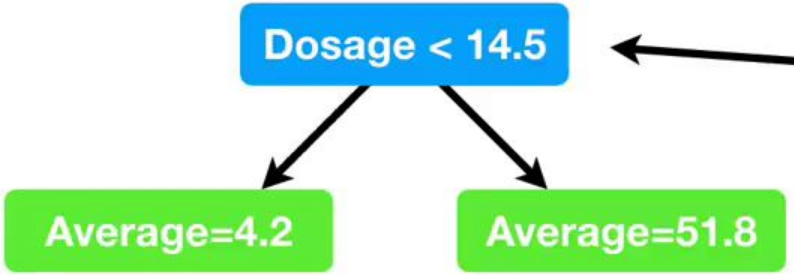


Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

…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…



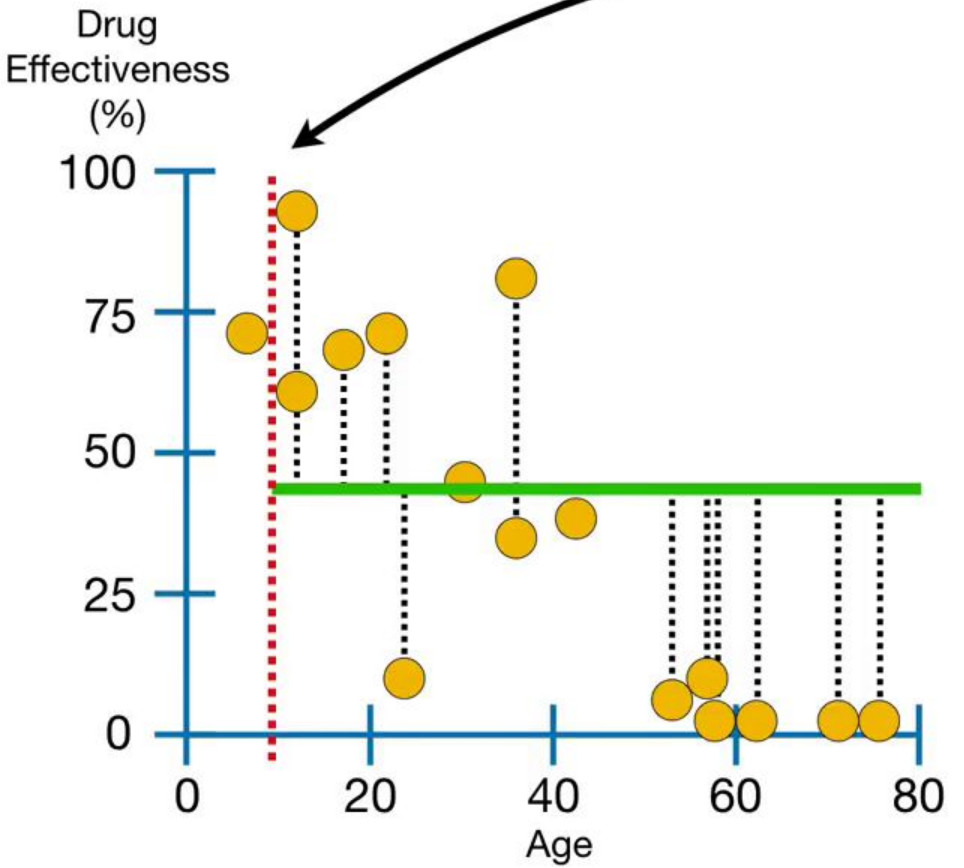
The best threshold becomes a *candidate* for the root.

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**.

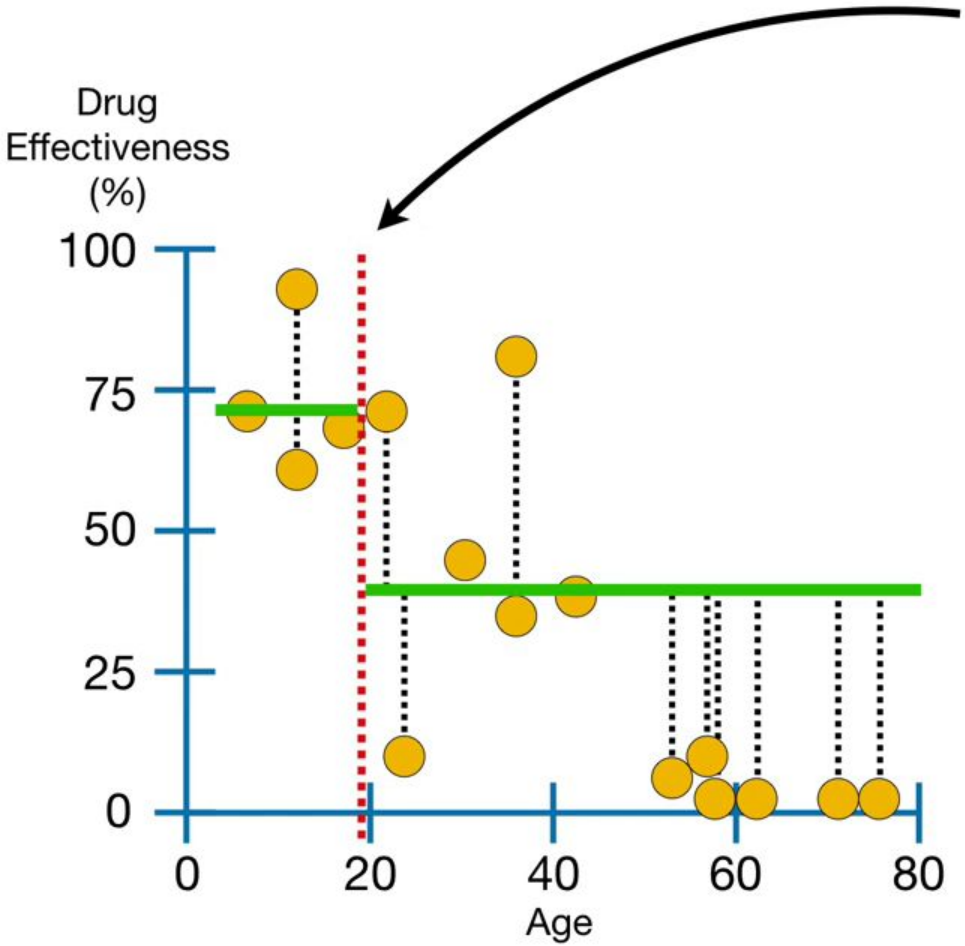
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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



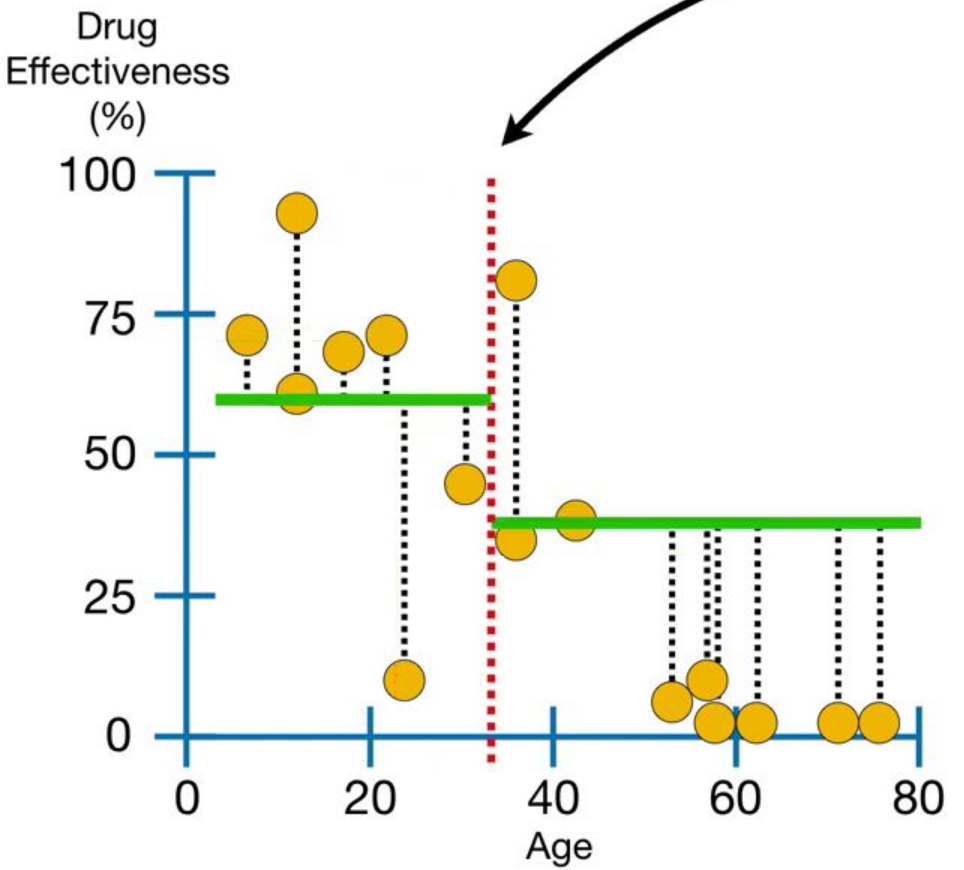
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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



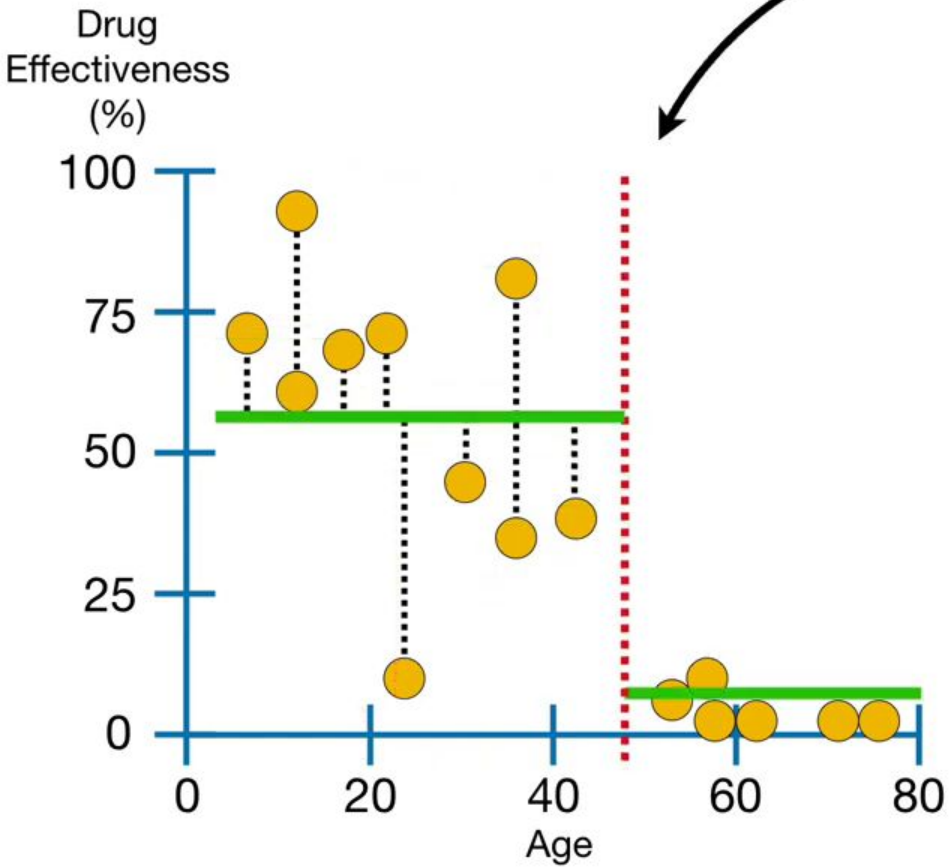
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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



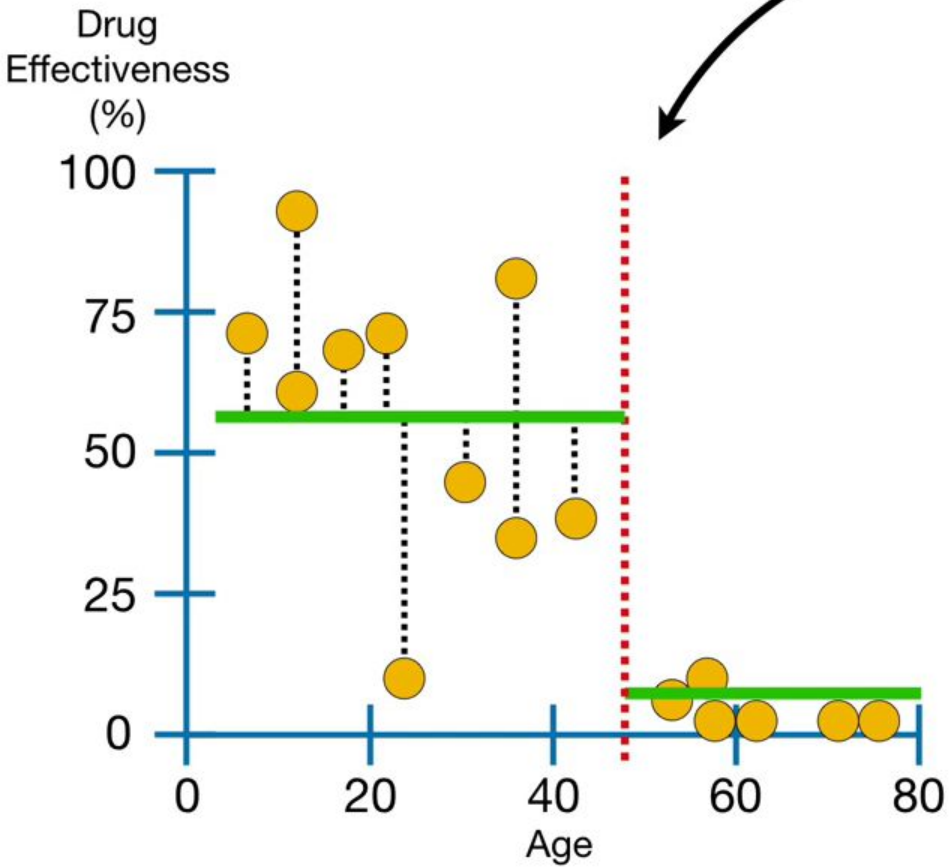
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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

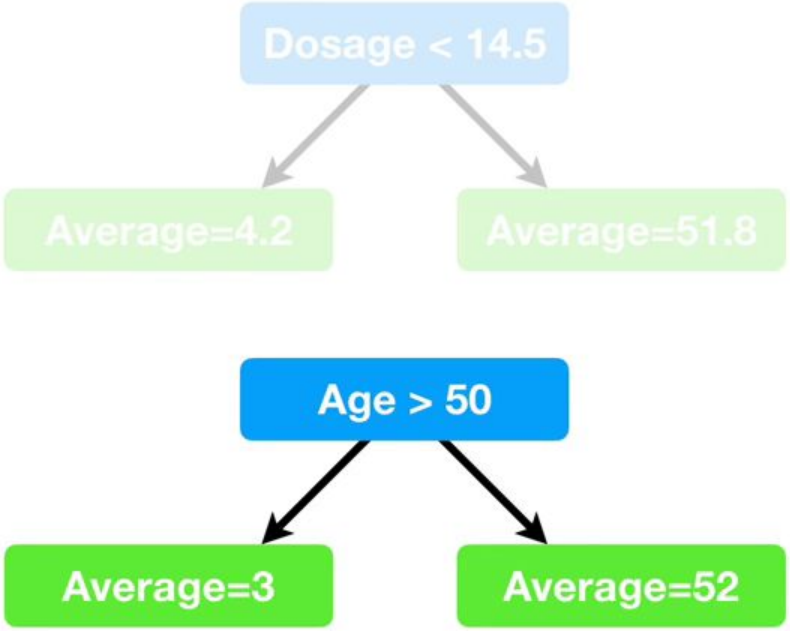


Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

…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…



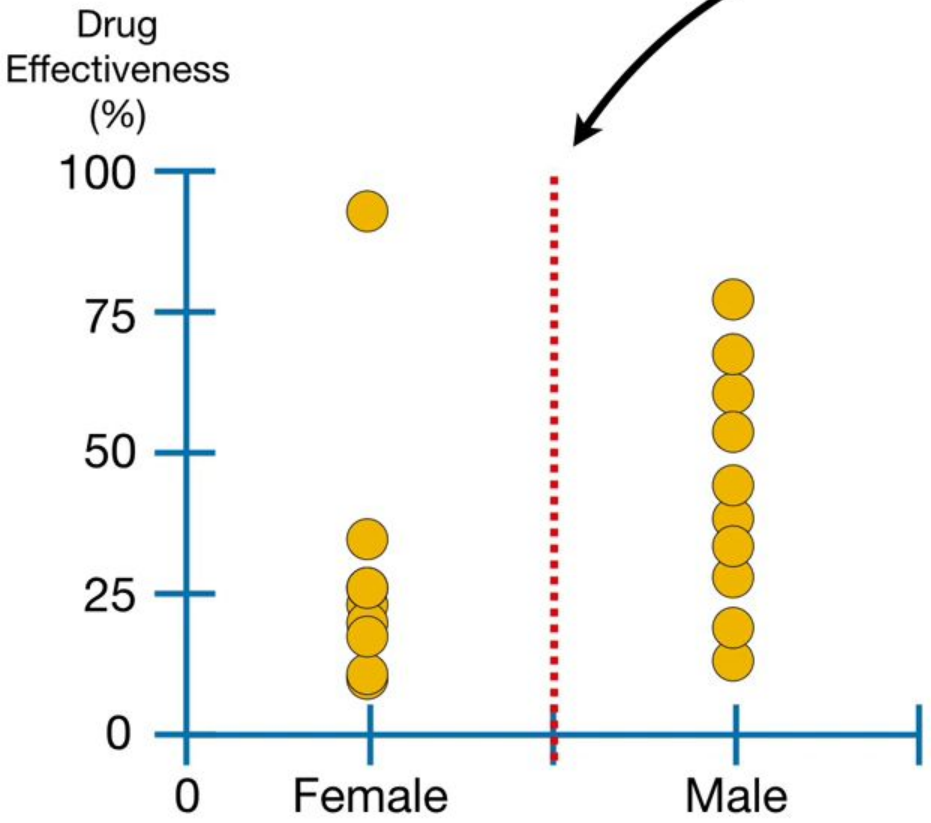
The best threshold becomes another *candidate* for the root.

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**.

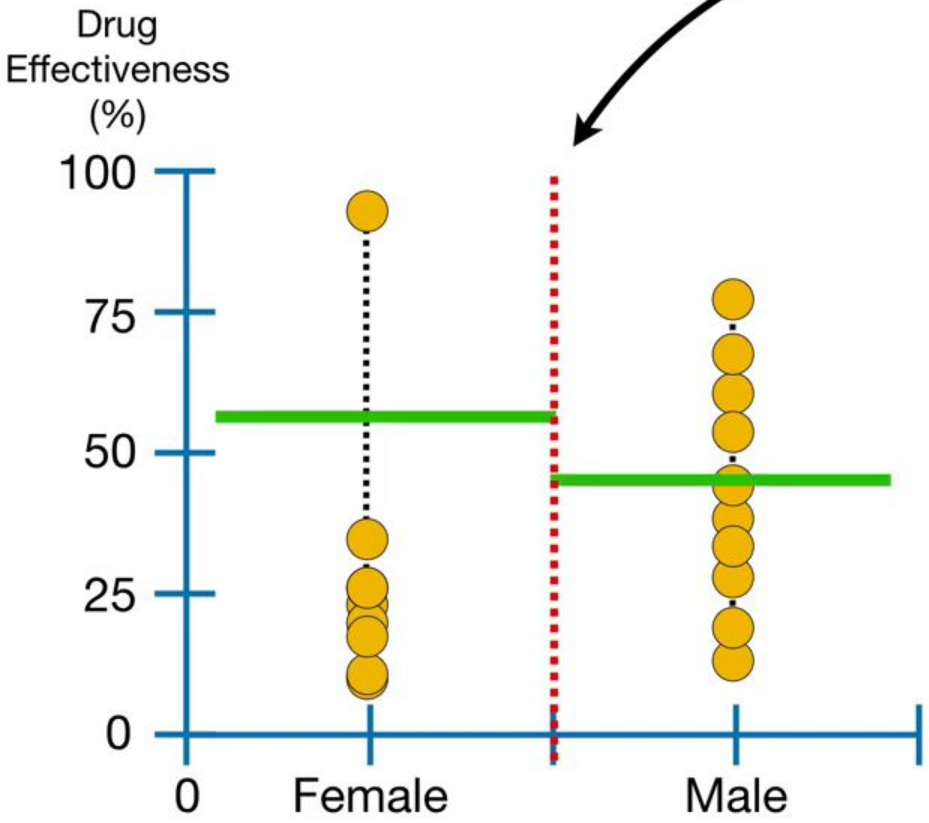
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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

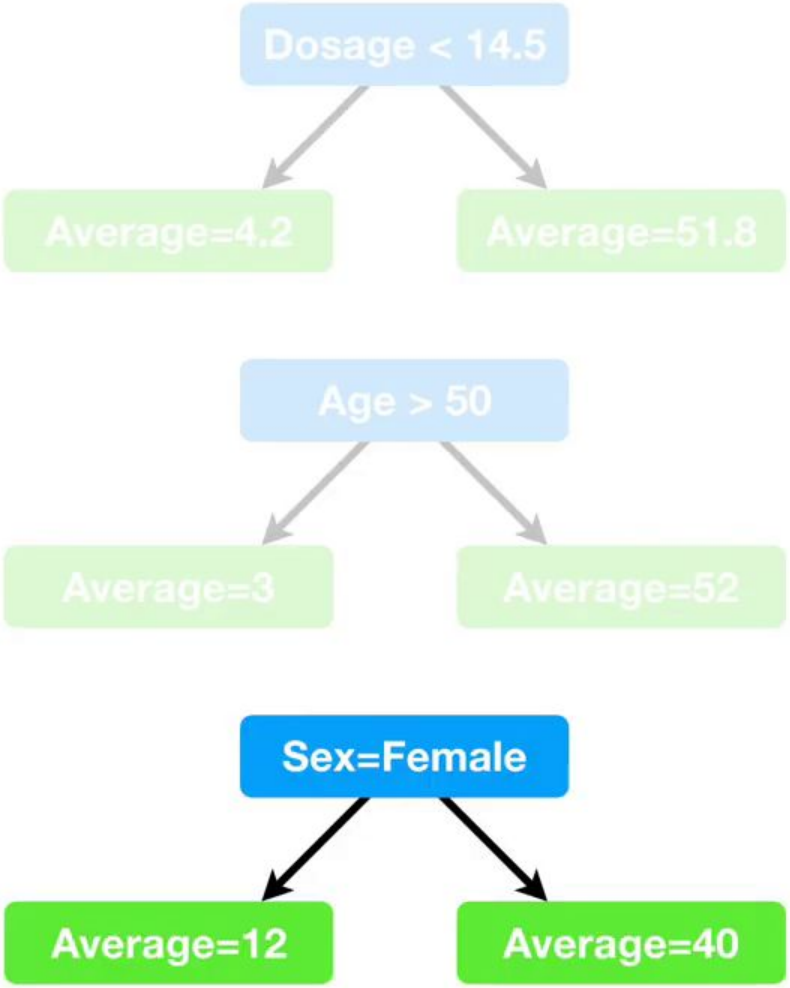


Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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

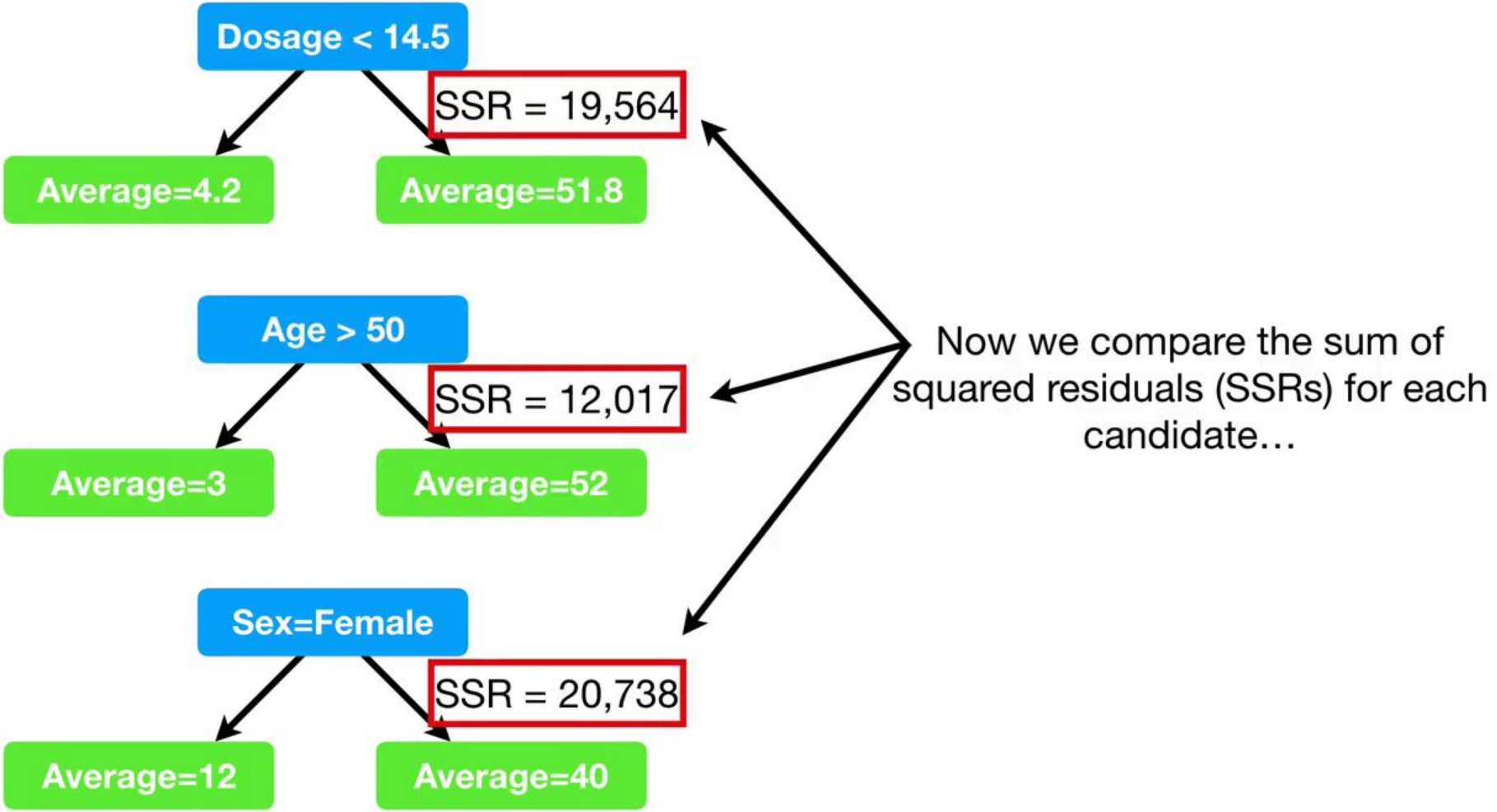


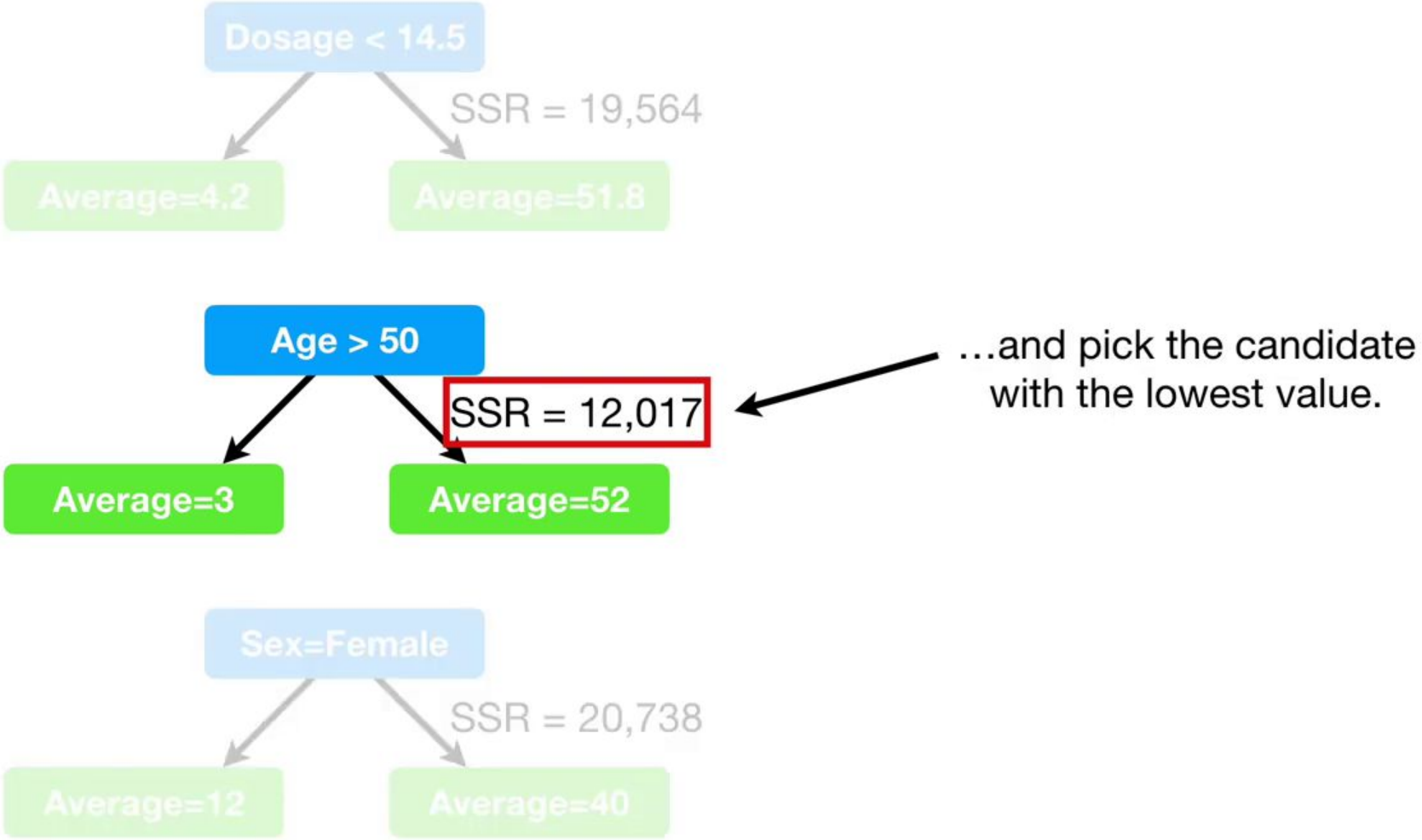
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…



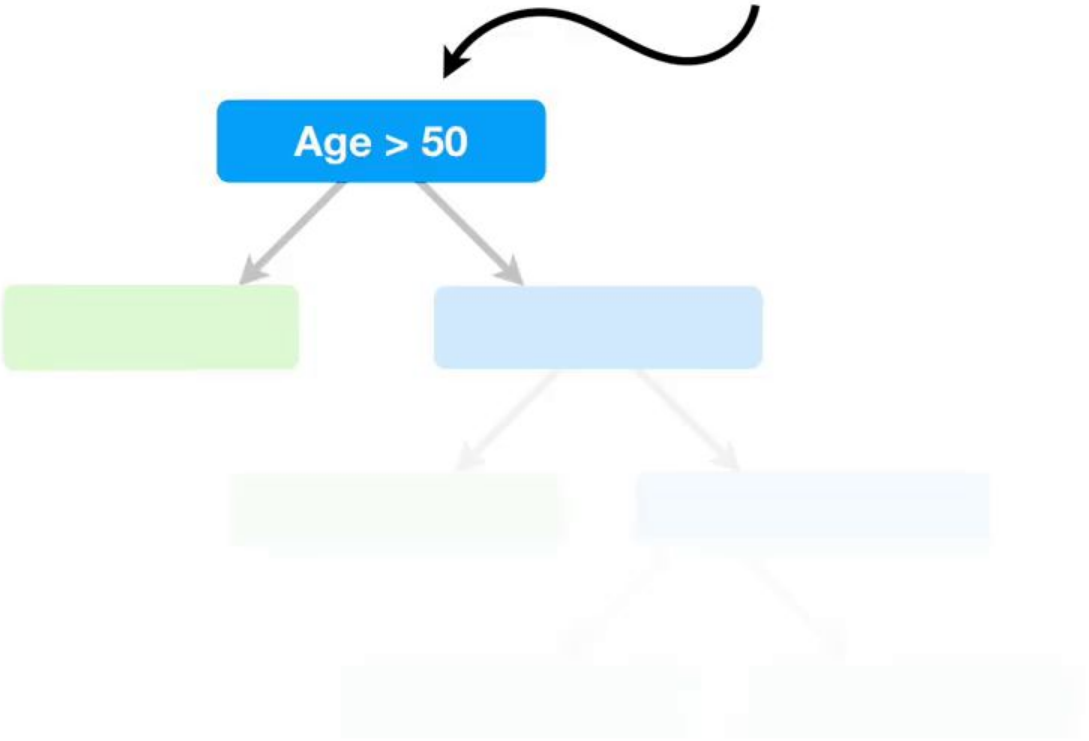
…and that becomes another *candidate* for the root.

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

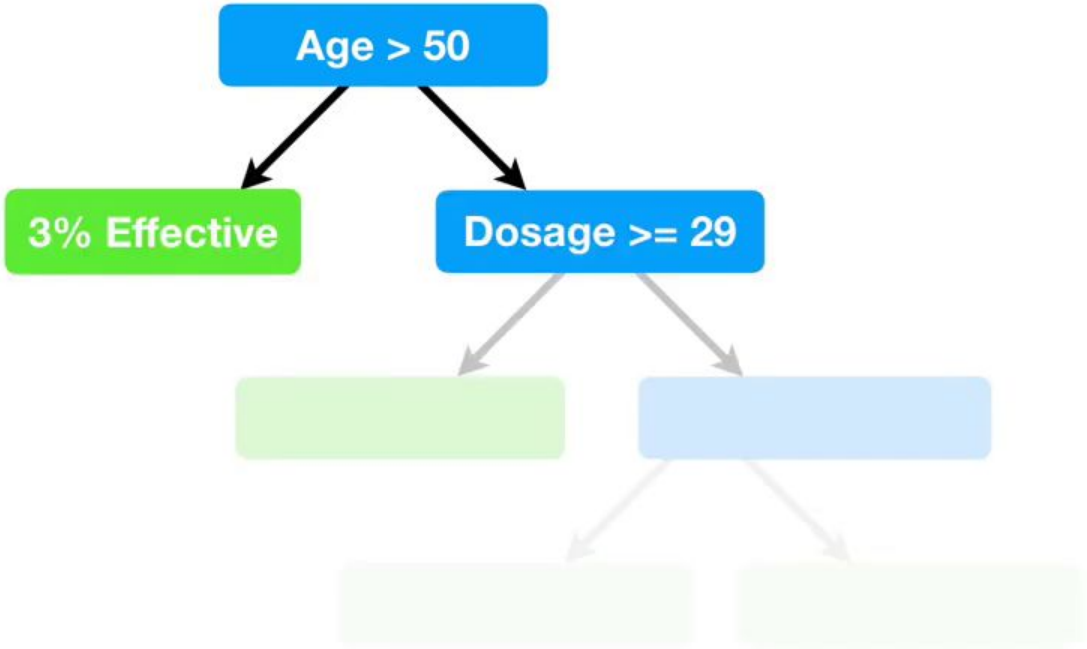

Dosage < 14.5
SSR = 19,564
Average=4.2
Average=51.8
Age > 50
SSR = 12,017
Average=3
Average=52
…and pick the candidate with the lowest value.
Sex=Female
SSR = 20,738
Average=12
Average=40

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



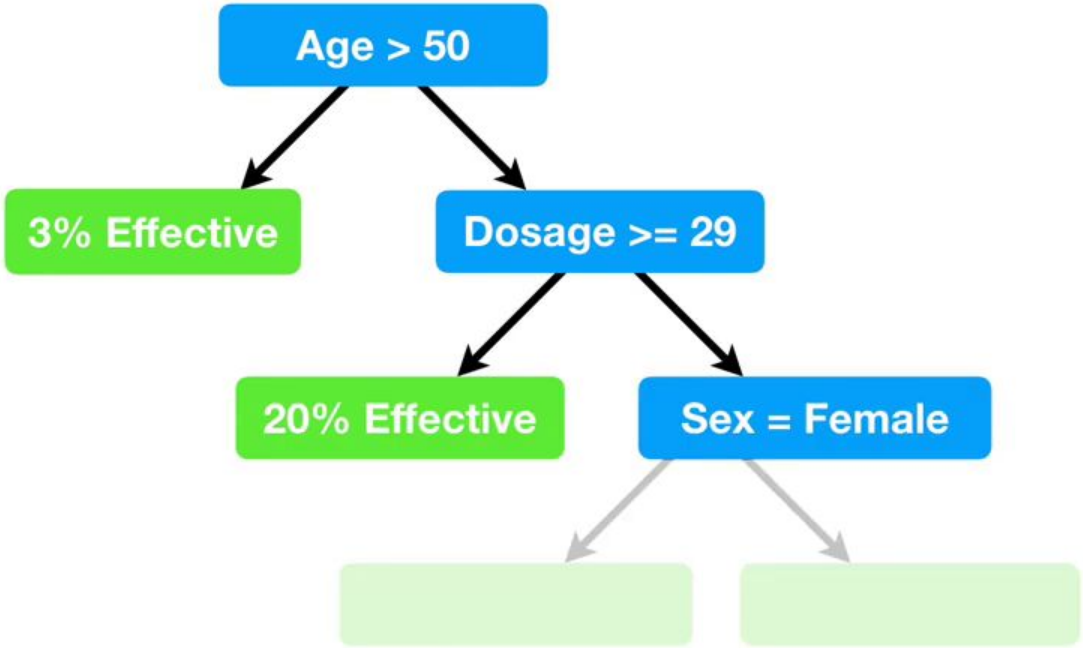
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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



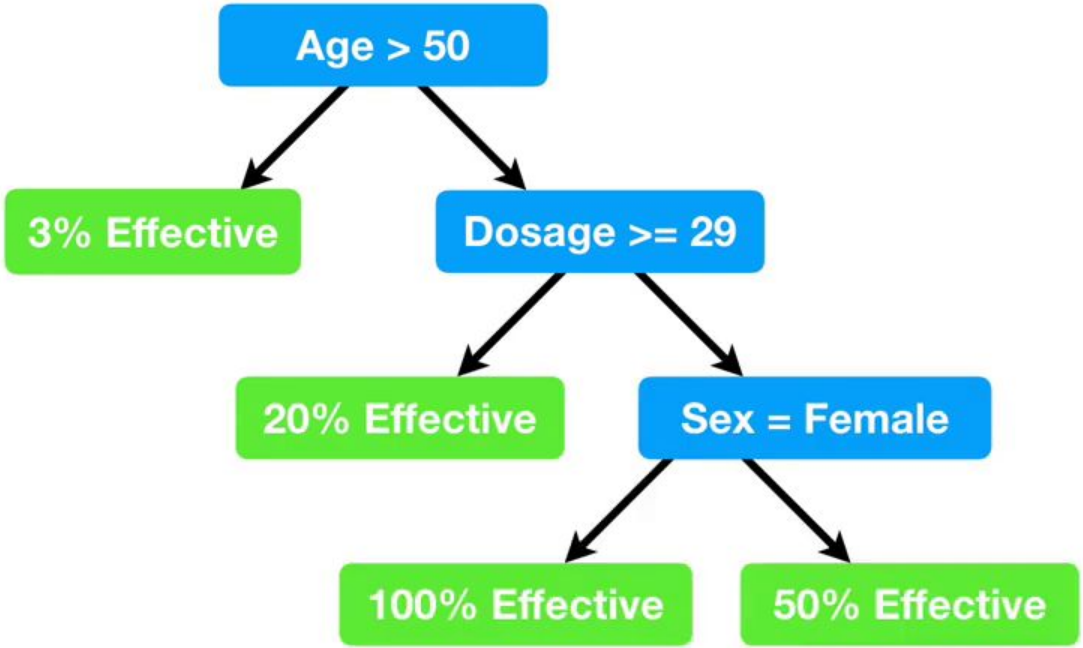
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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



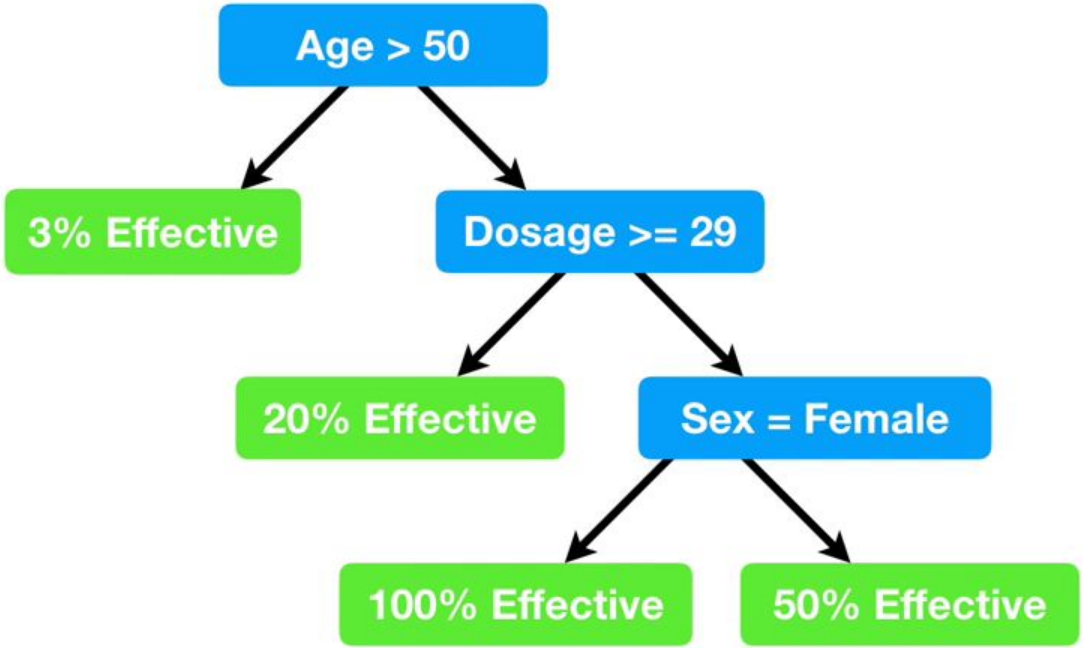
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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



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…

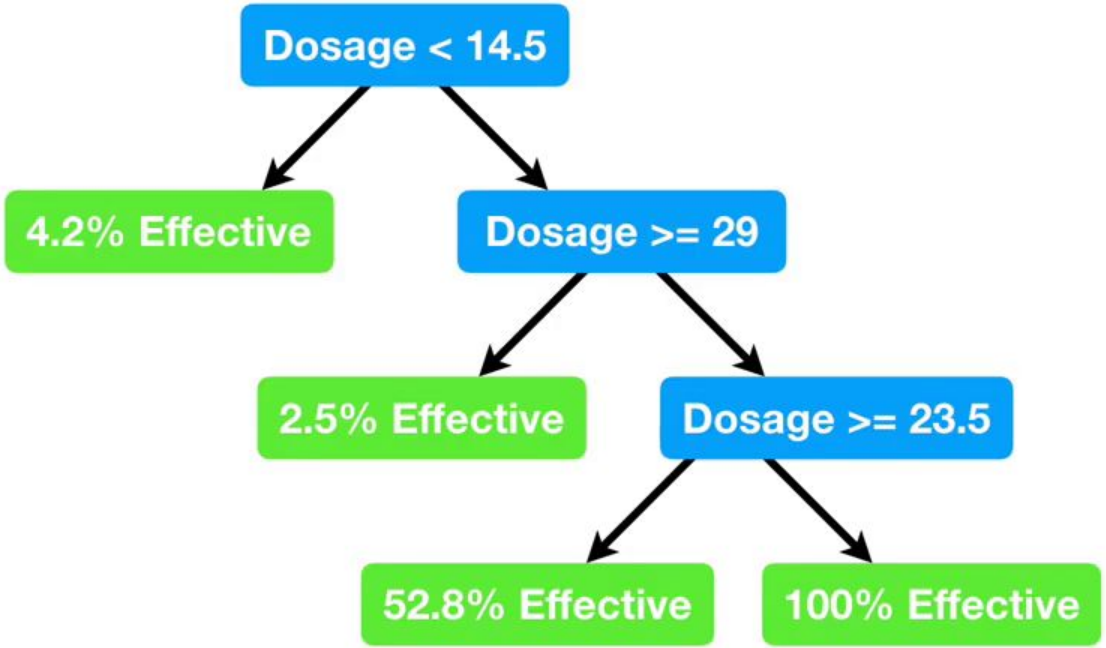
And just like before, when a leaf has less than a minimum number of observations, which is usually **20**, but we are using **7**, we stop trying to divide them.



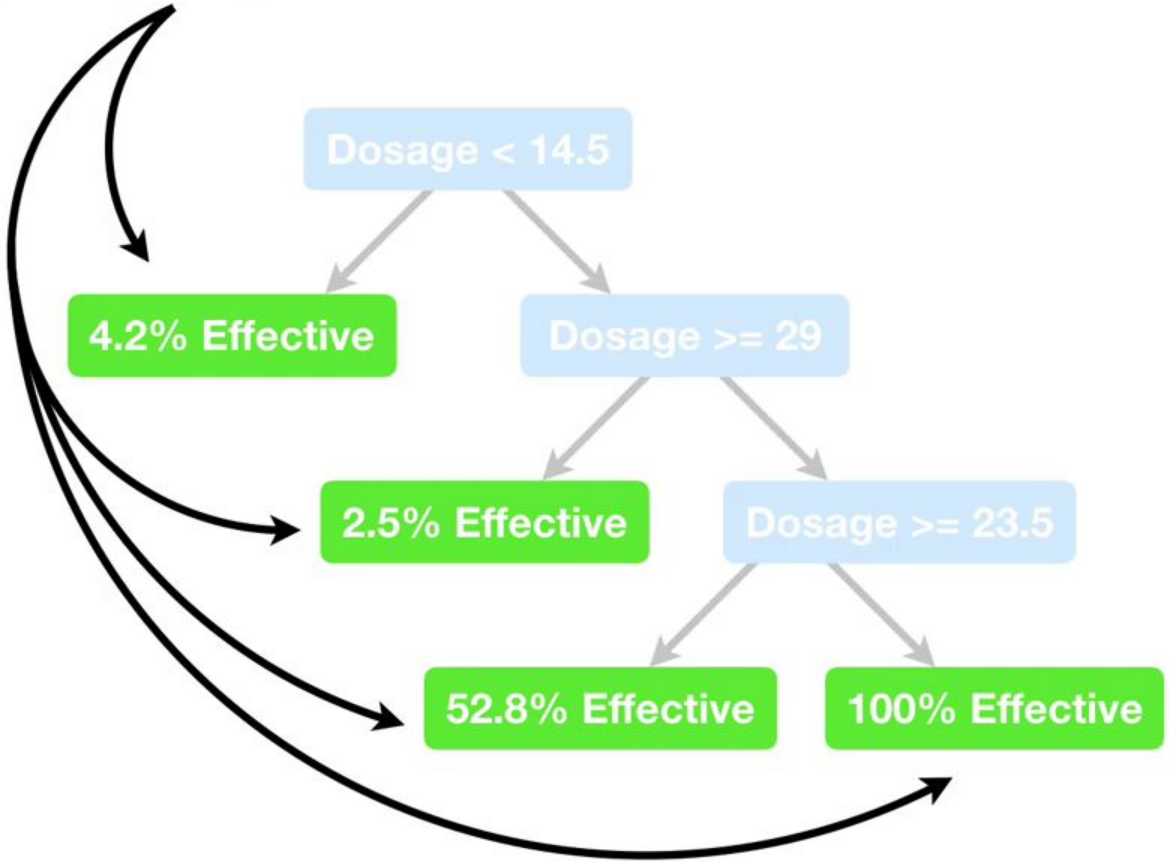
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

In summary…

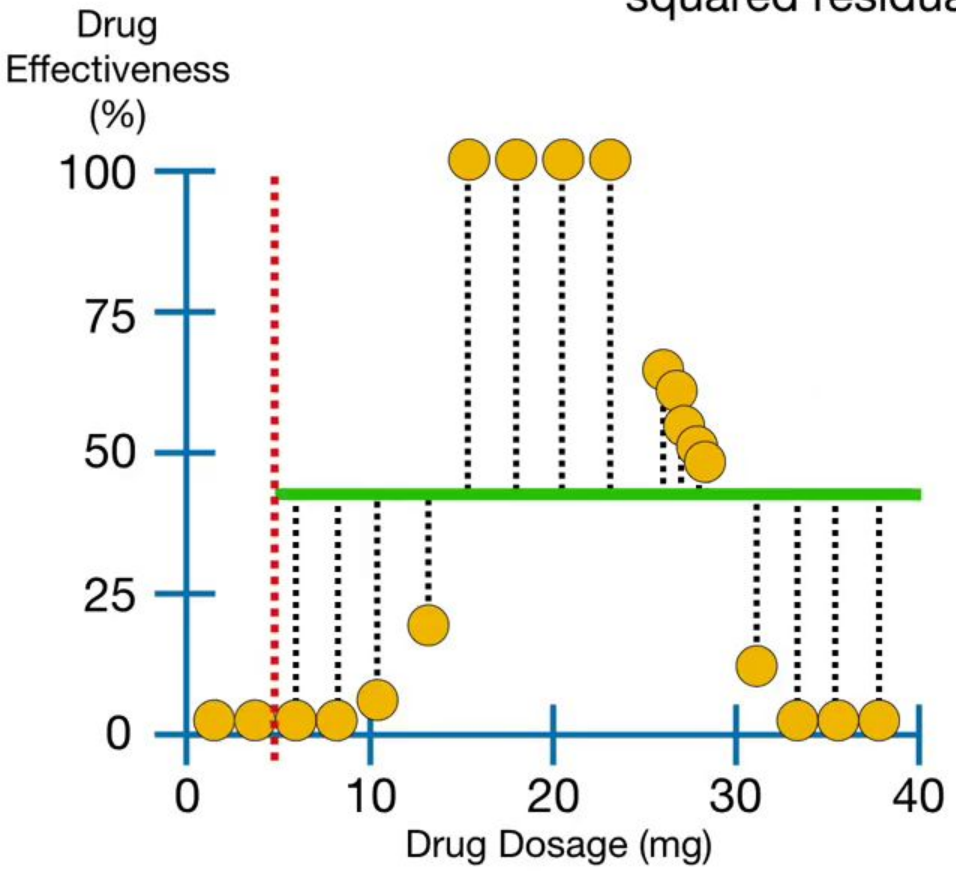
Regression Trees are a type of
Decision Tree.

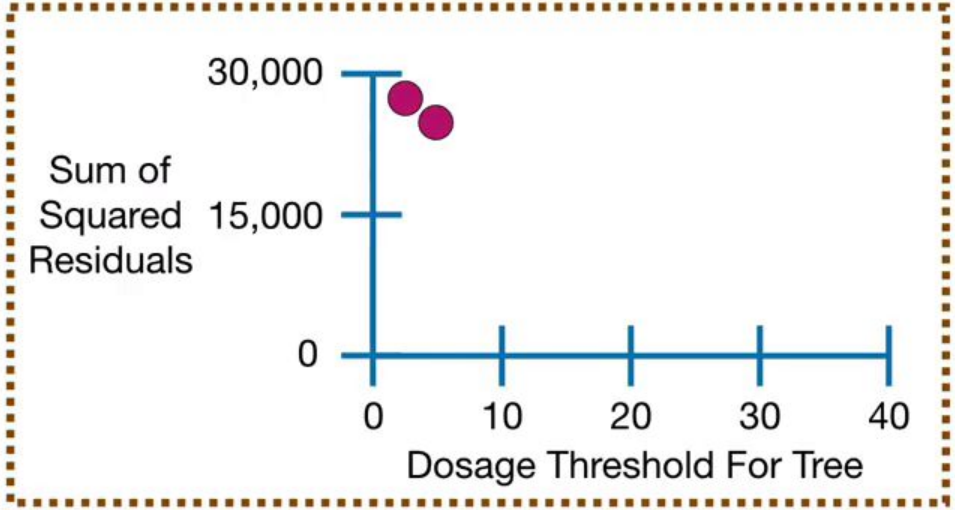


In a **Regression Tree**, each leaf represents a numeric value.

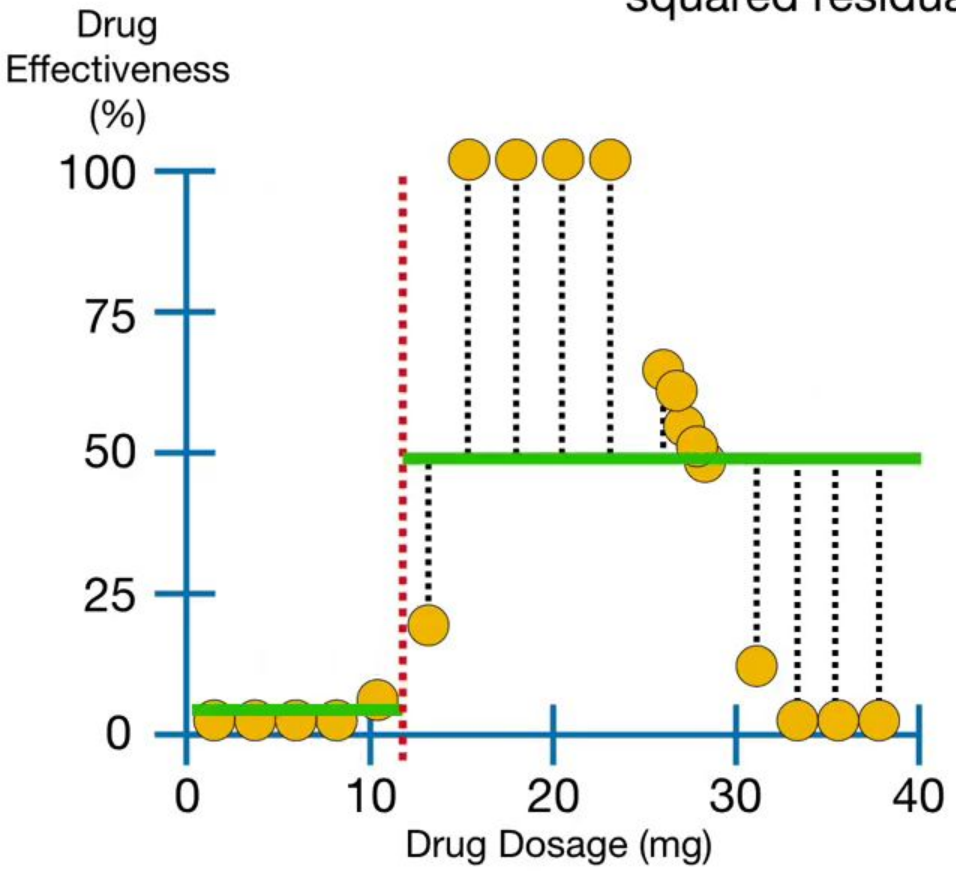


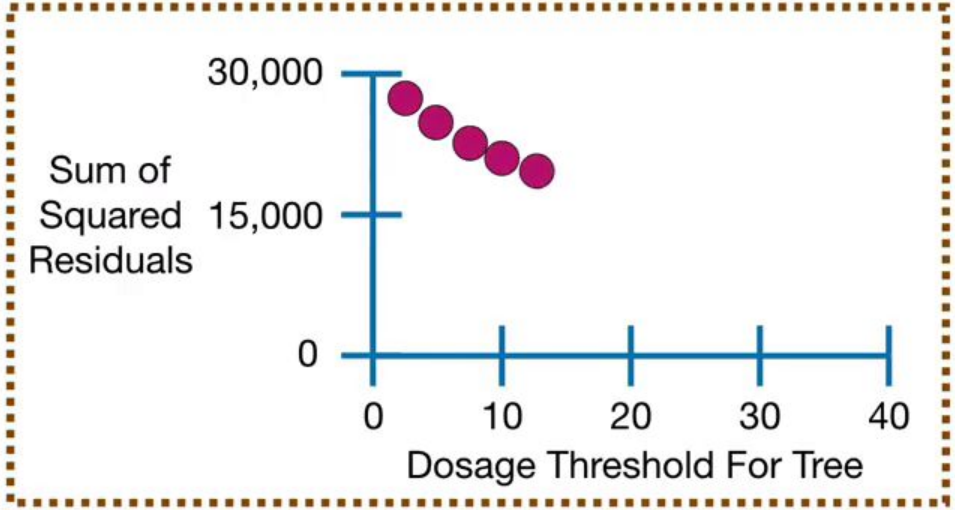
We determine how to divide the observations by trying different thresholds and calculating the sum of squared residuals at each step.



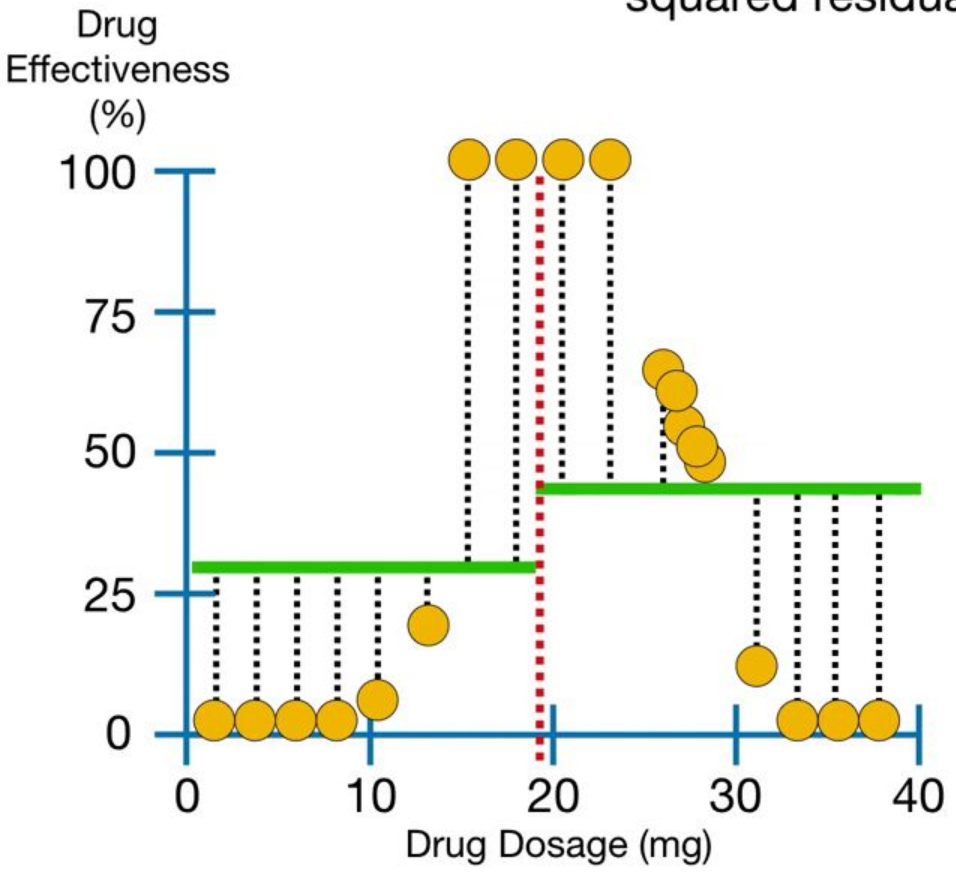


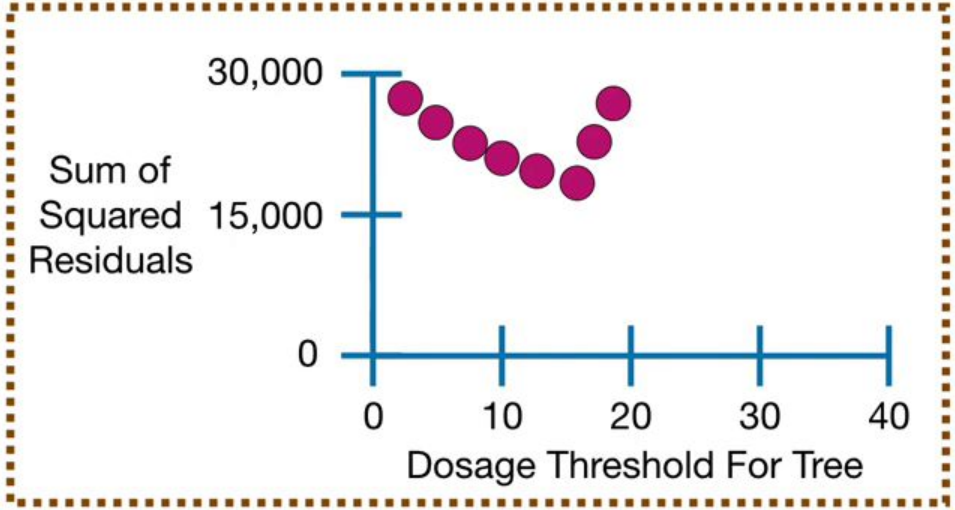
We determine how to divide the observations by trying different thresholds and calculating the sum of squared residuals at each step.



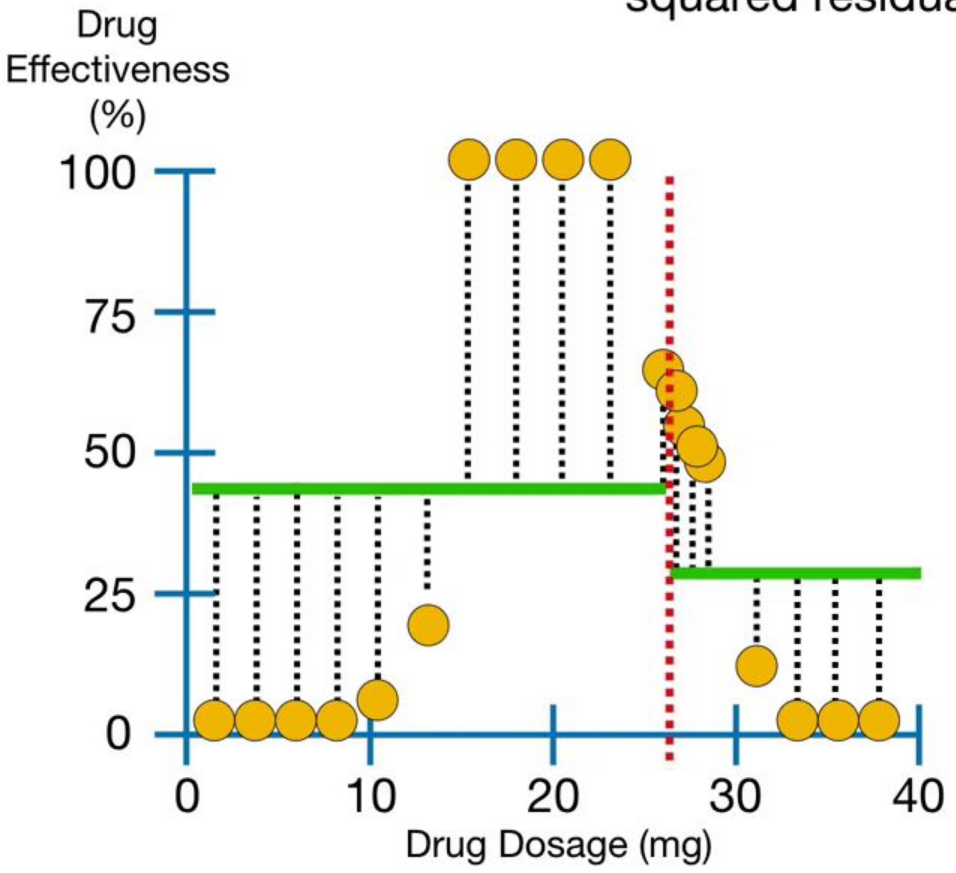


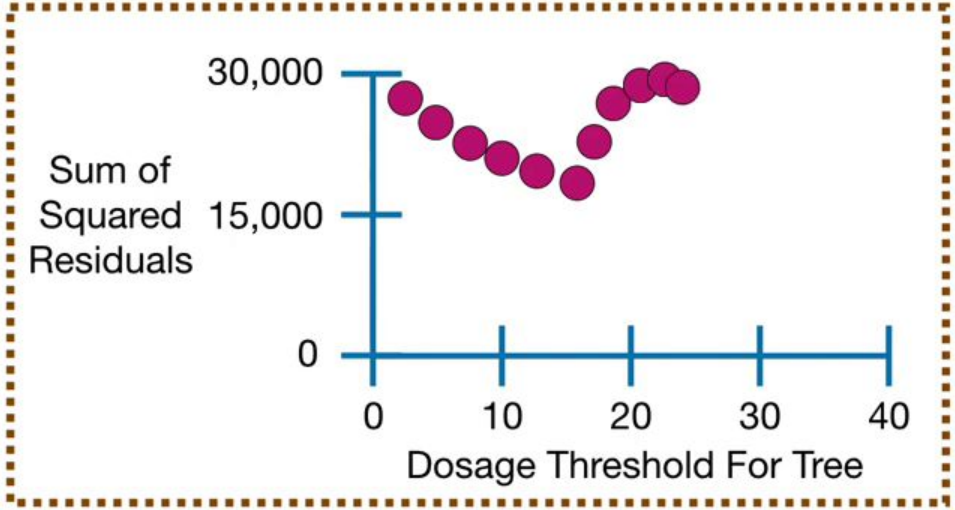
We determine how to divide the observations by trying different thresholds and calculating the sum of squared residuals at each step.



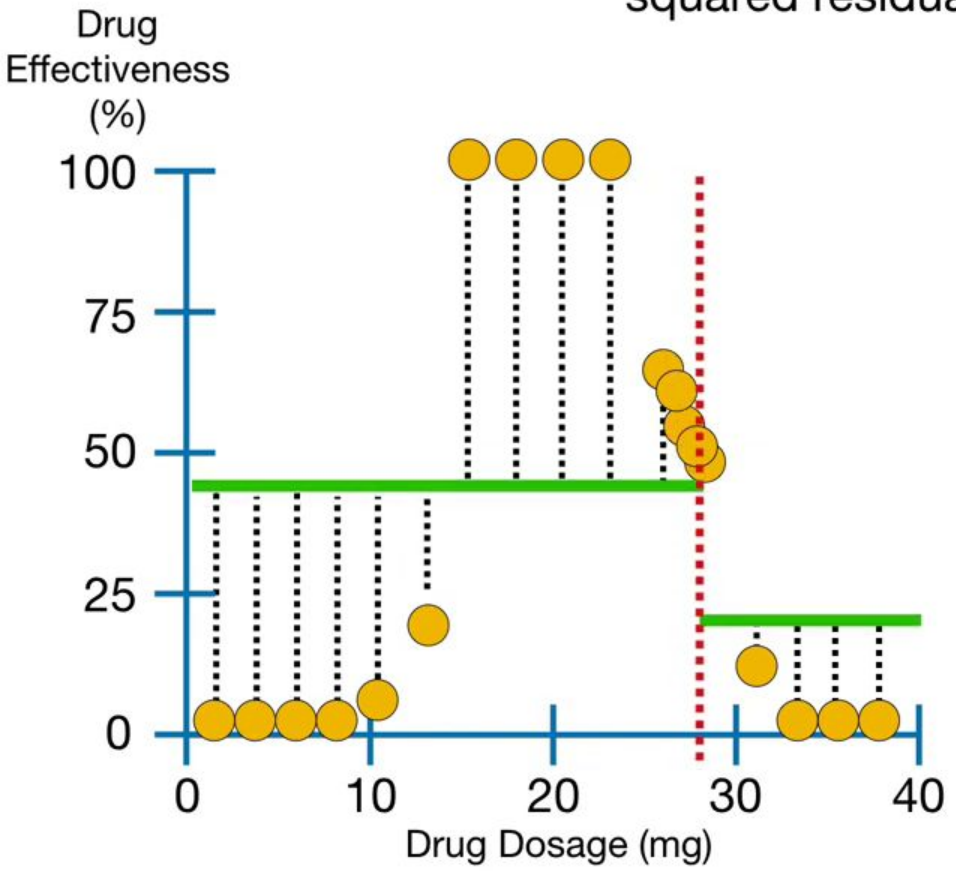


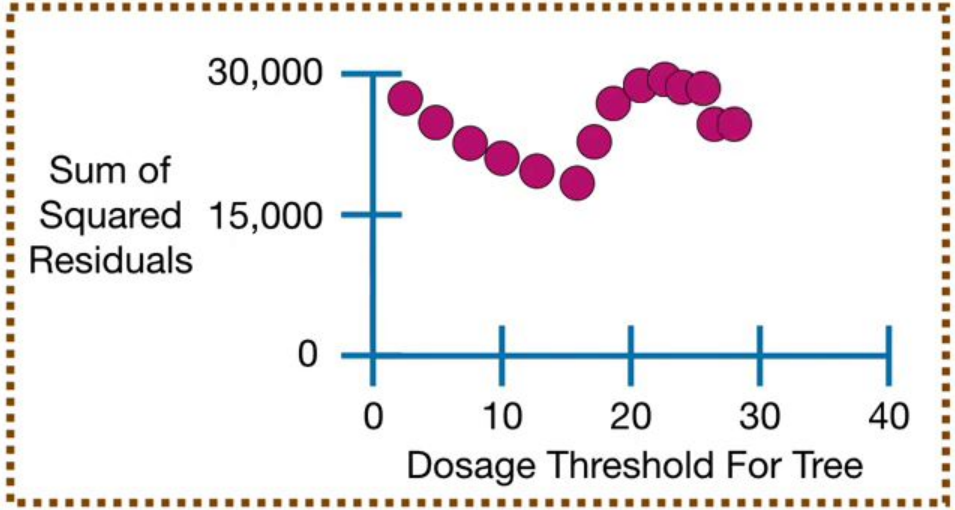
We determine how to divide the observations by trying different thresholds and calculating the sum of squared residuals at each step.



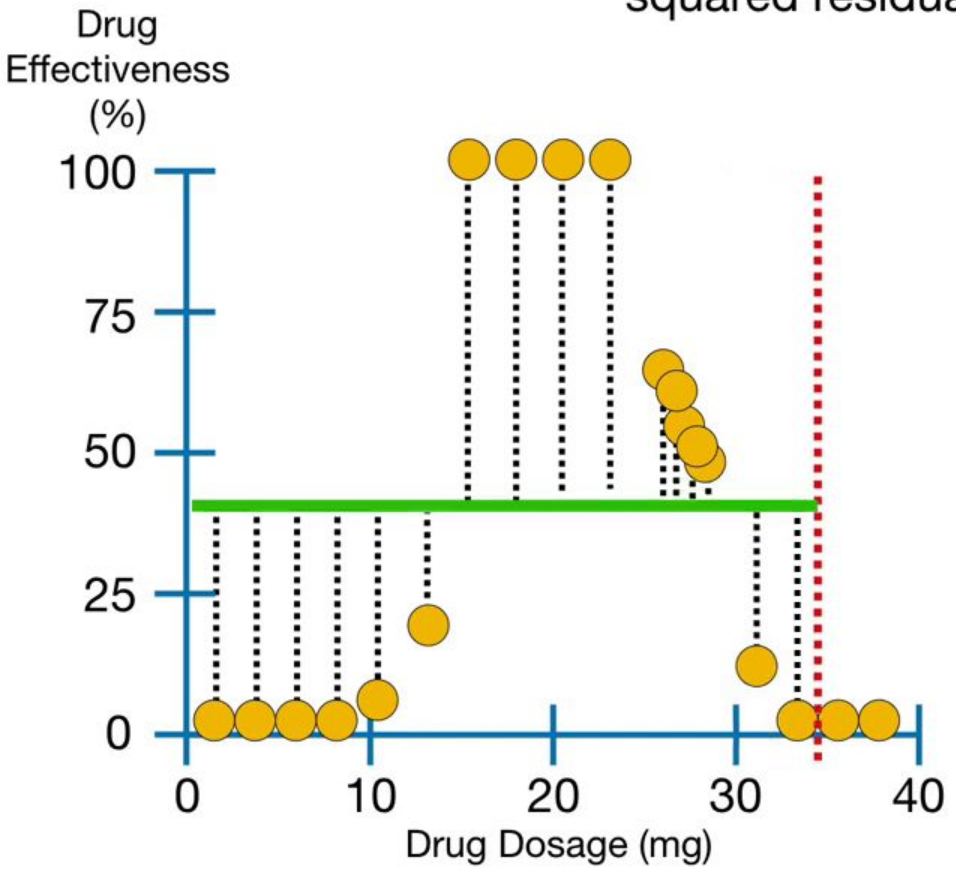


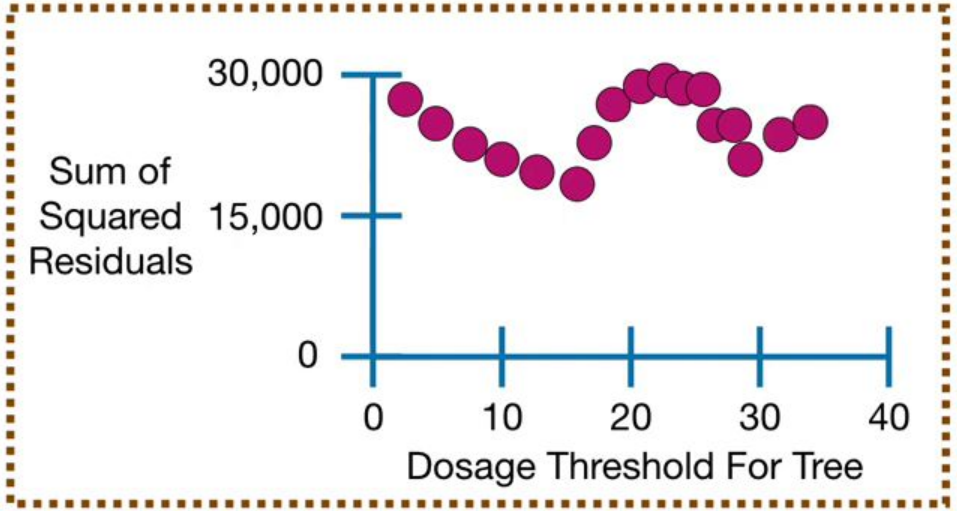
We determine how to divide the observations by trying different thresholds and calculating the sum of squared residuals at each step.

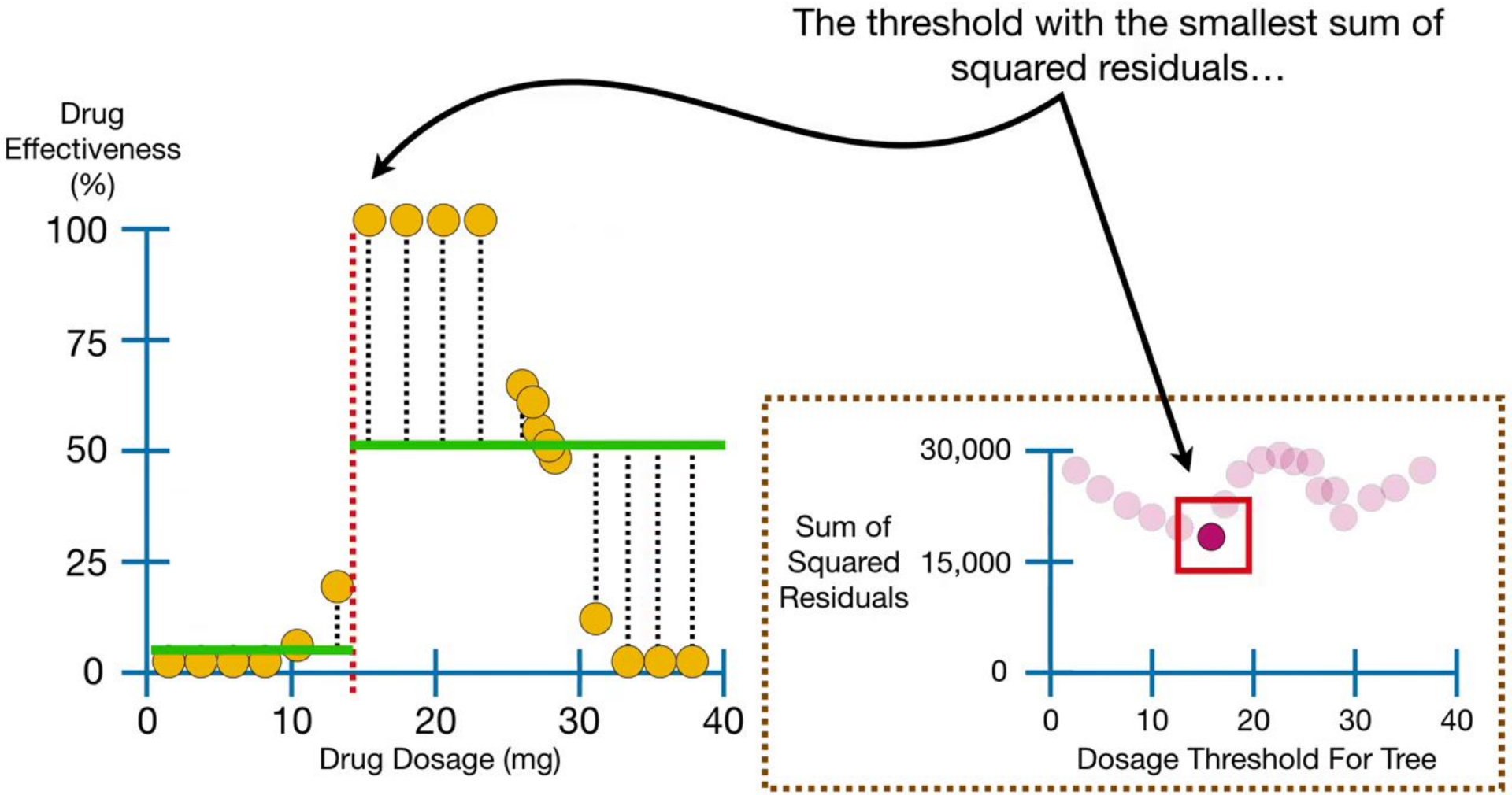


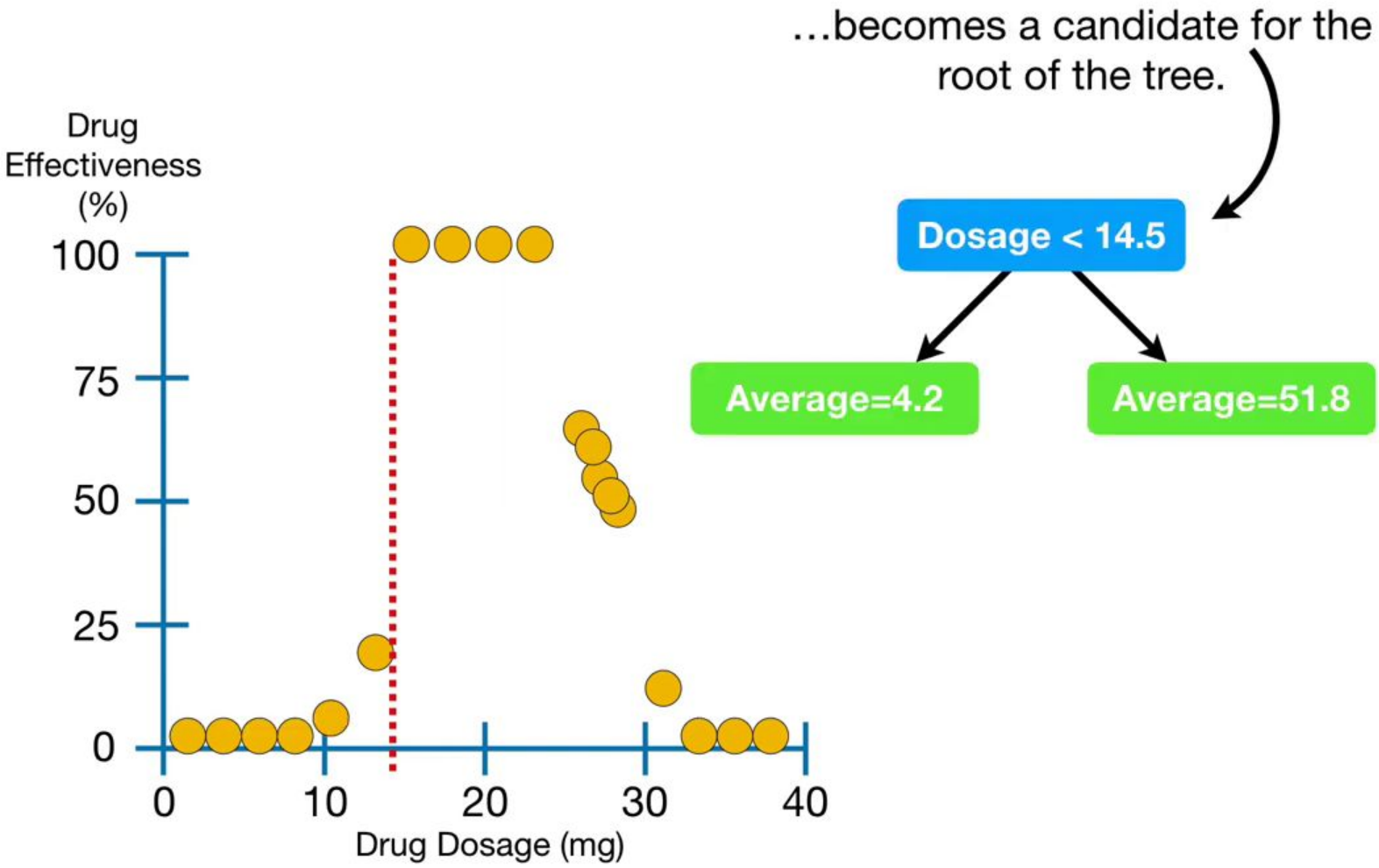


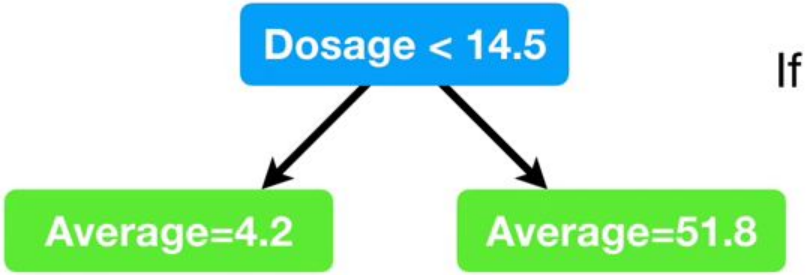
We determine how to divide the observations by trying different thresholds and calculating the sum of squared residuals at each step.





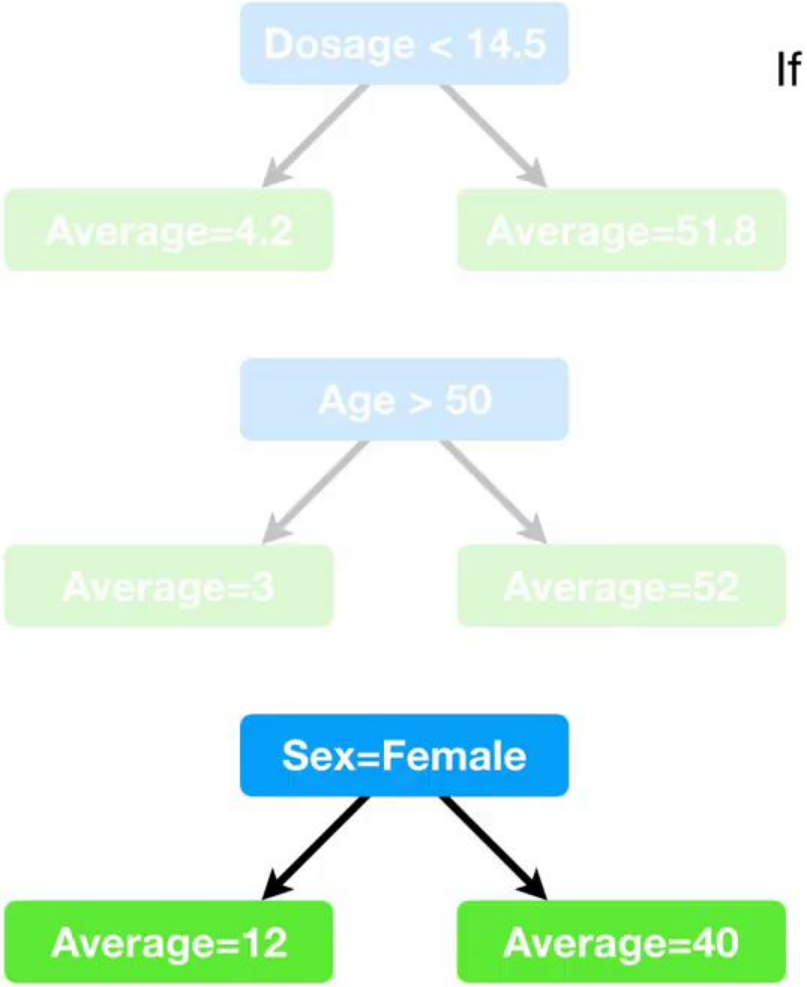

The threshold with the smallest sum of squared residuals…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Sum of Squared Residuals
30,000
15,000
0
0
10
20
30
40
Dosage Threshold For Tree


…becomes a candidate for the root of the tree.
Dosage < 14.5
Average=4.2
Average=51.8
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)



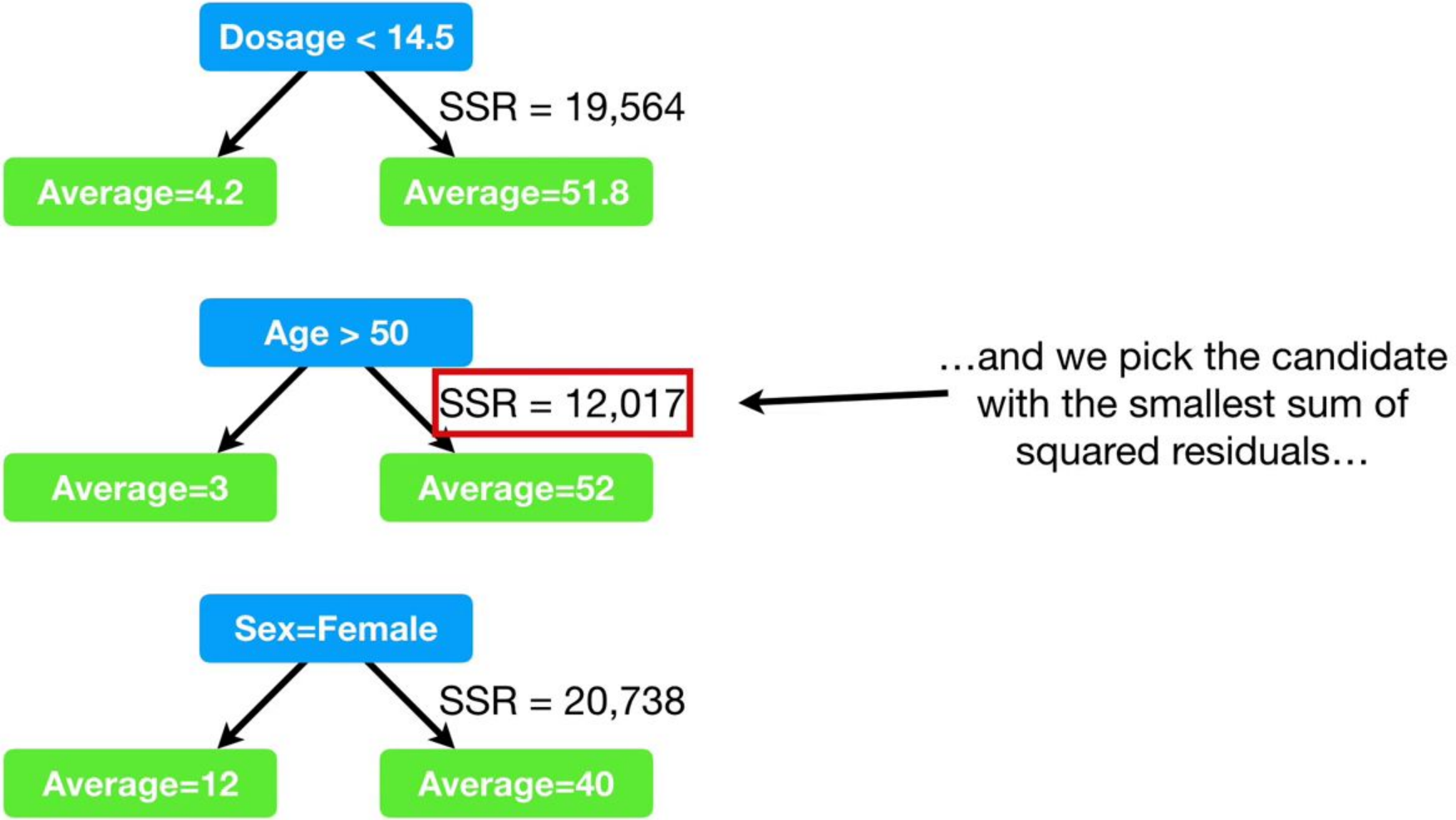
If we have more than one predictor, we find the optimal threshold for each one…

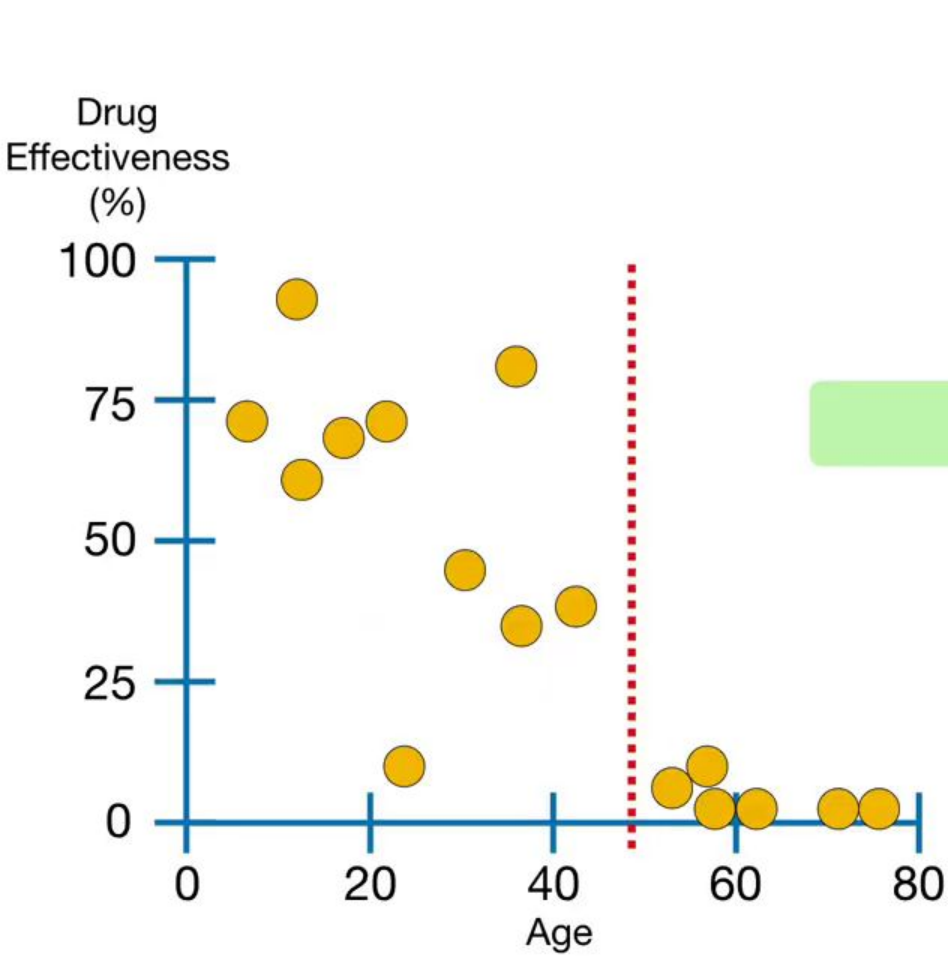
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…



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…

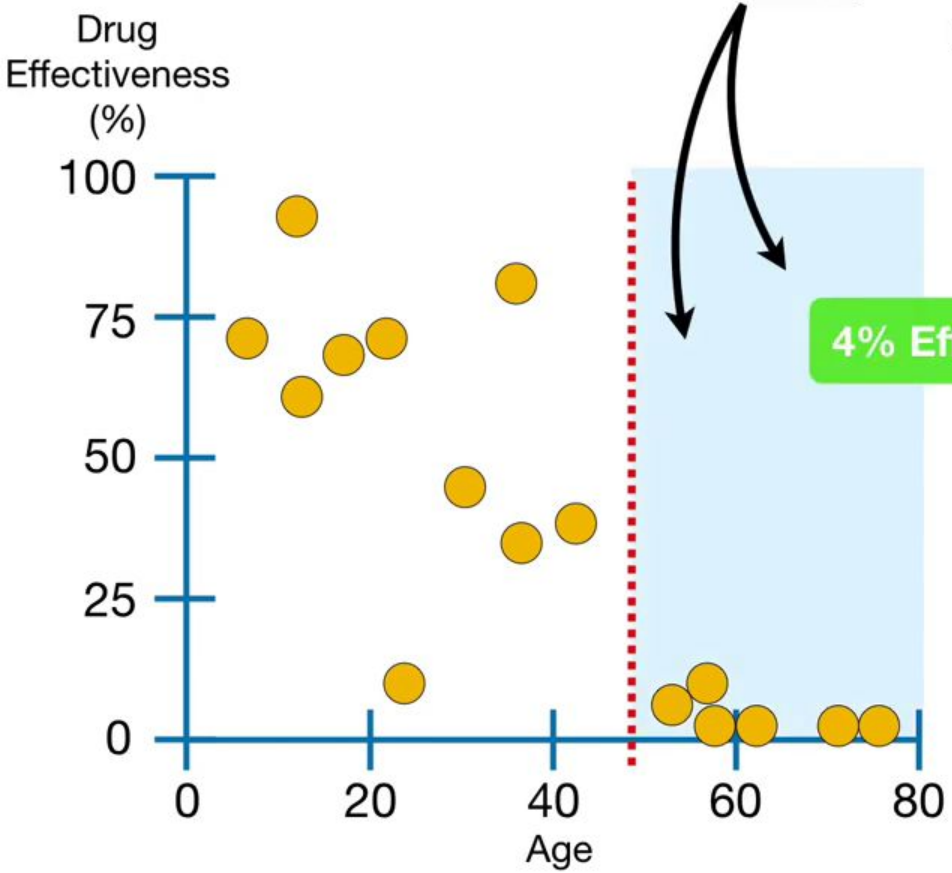

Dosage < 14.5
SSR = 19,564
Average=4.2
Average=51.8
Age > 50
SSR = 12,017
Average=3
Average=52
…and we pick the candidate with the smallest sum of squared residuals…
Sex=Female
SSR = 20,738
Average=12
Average=40

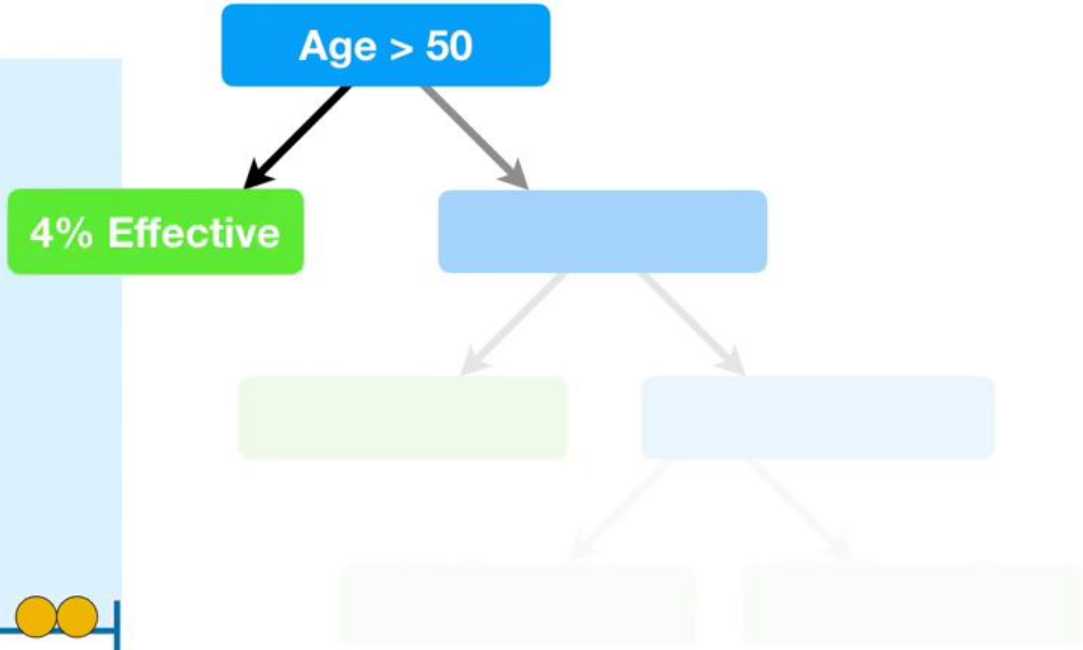


…to be the root.

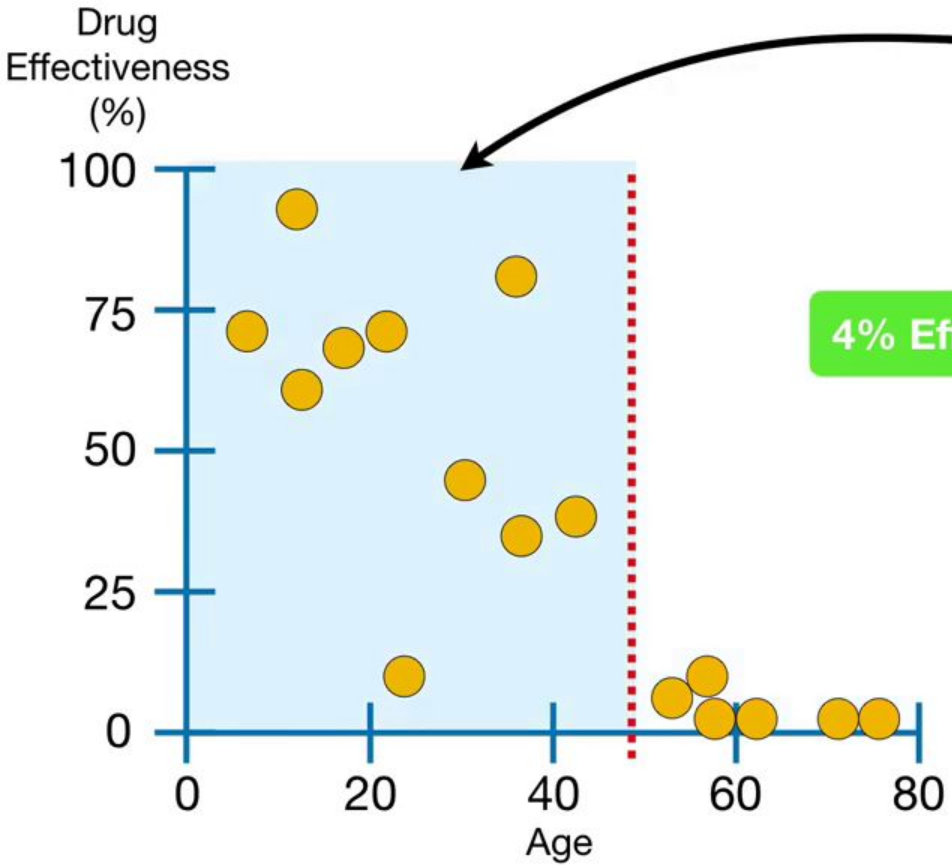
Age > 50

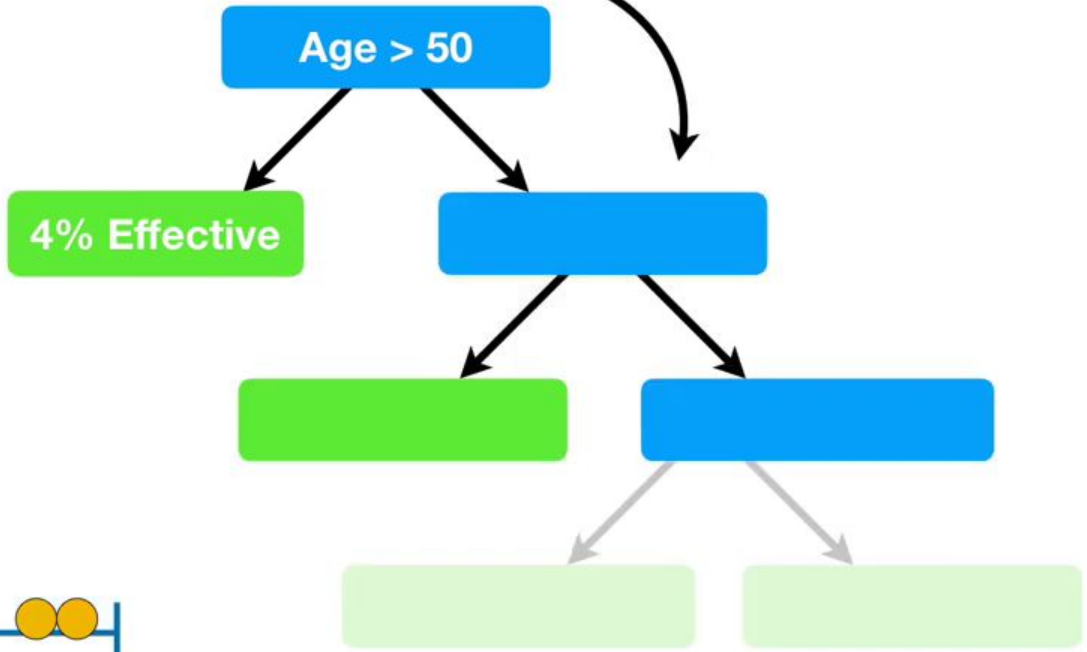
When we have fewer than some minimum number of observations in a node (**7** in this example, but more commonly **20**), then that node becomes a leaf….



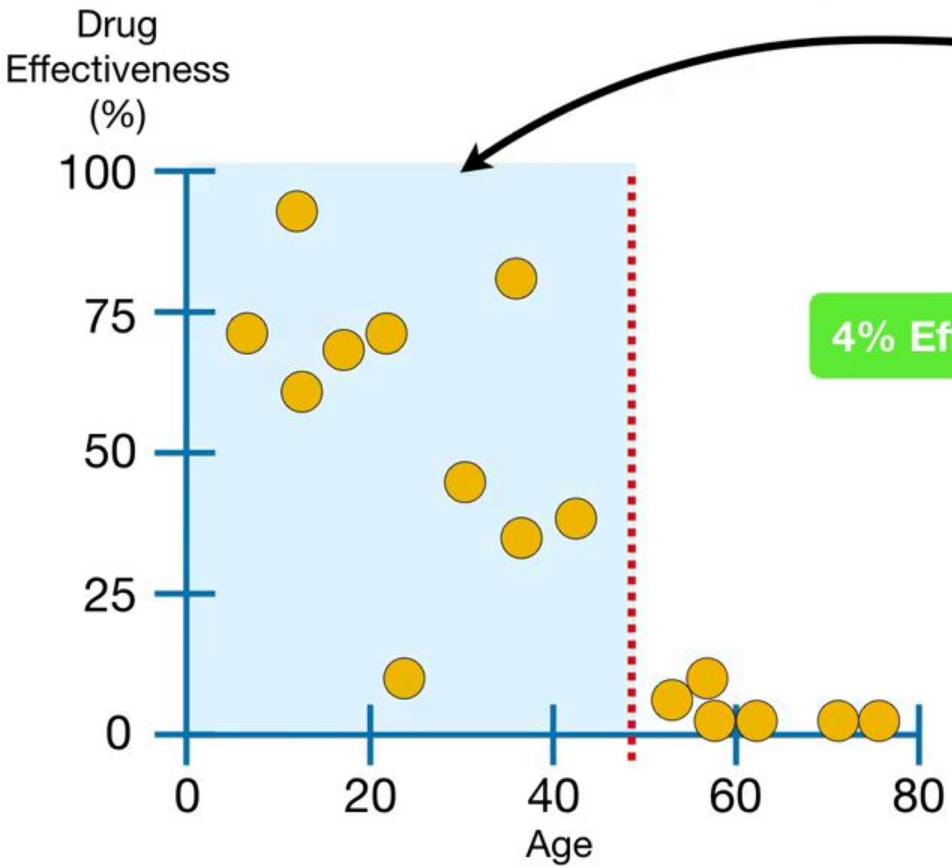


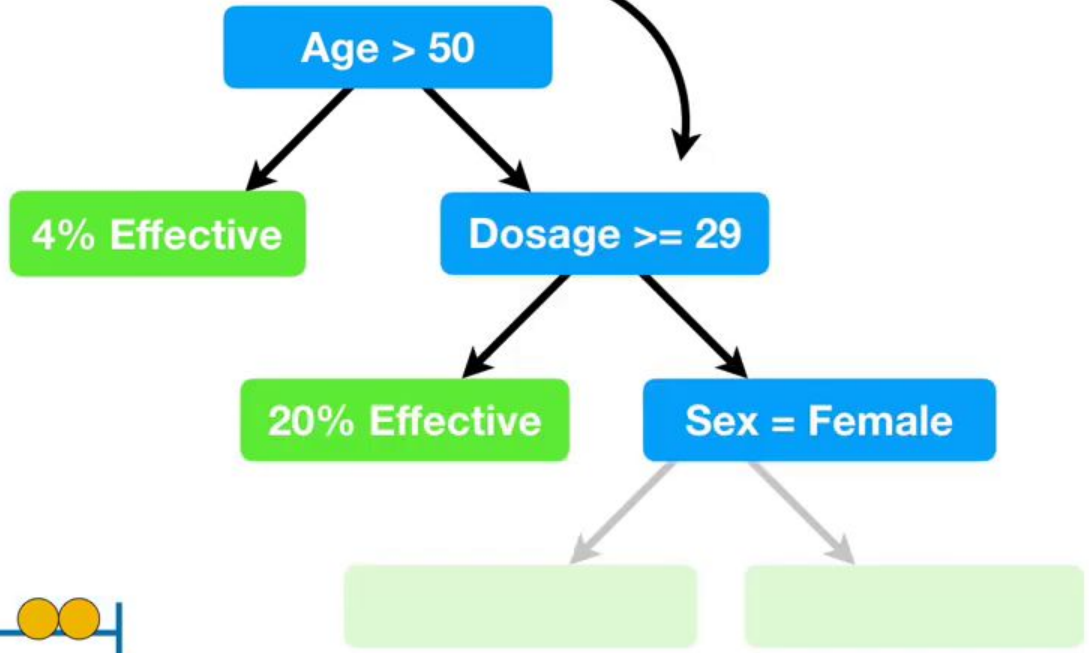
…otherwise we repeat the process to split the remaining observations…



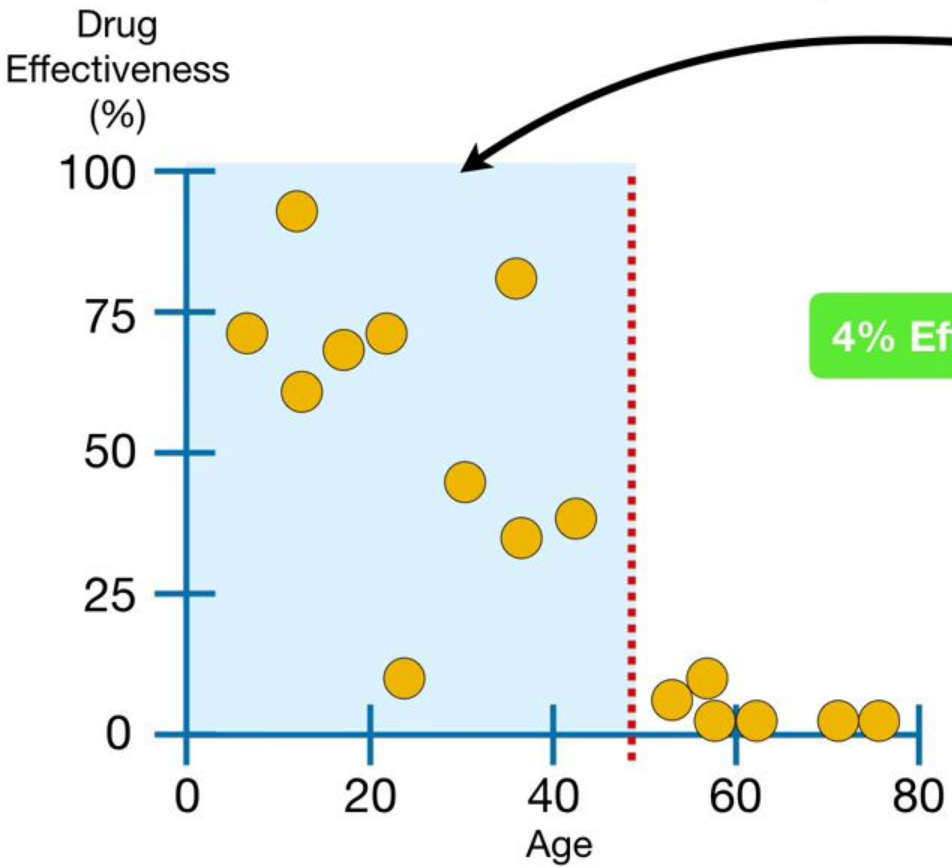


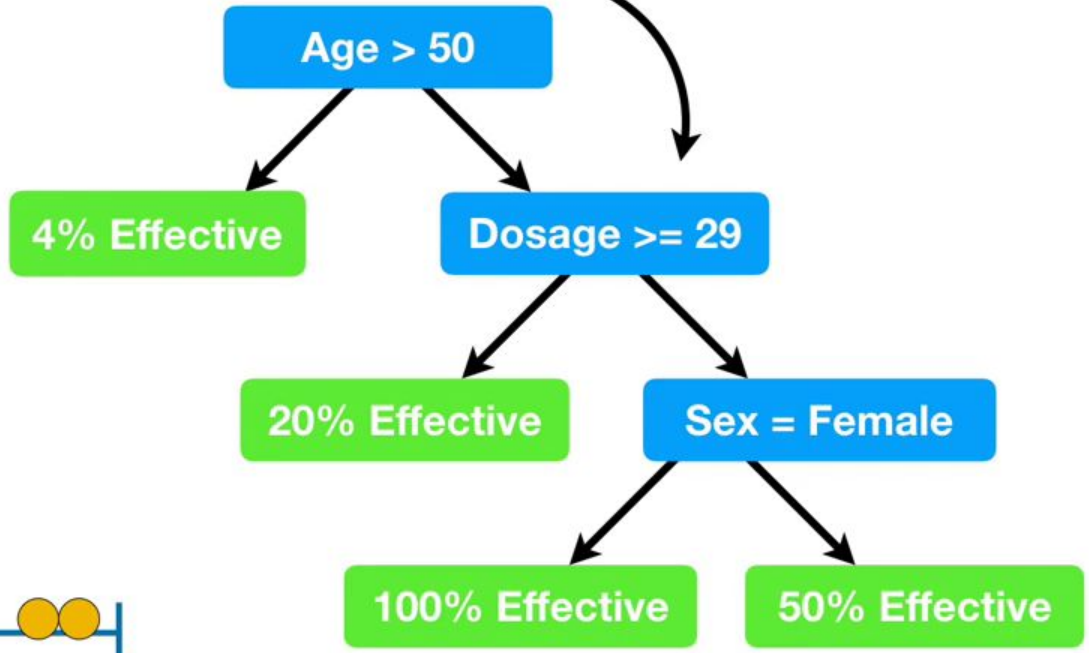
…otherwise we repeat the process to split the remaining observations…



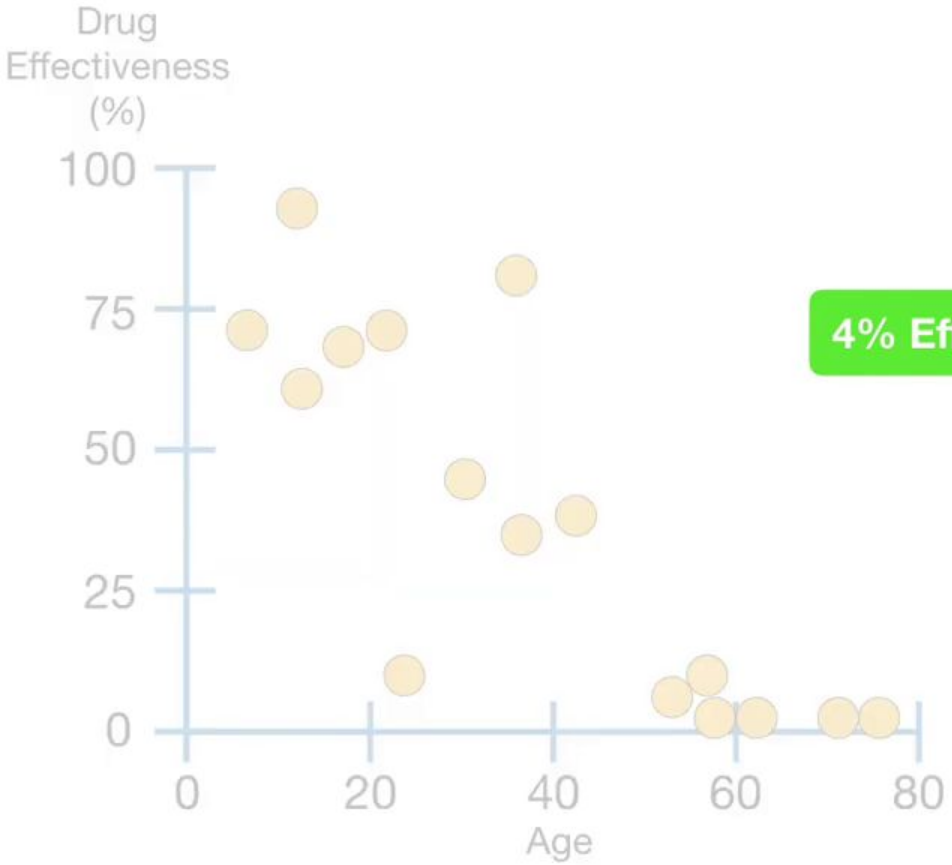


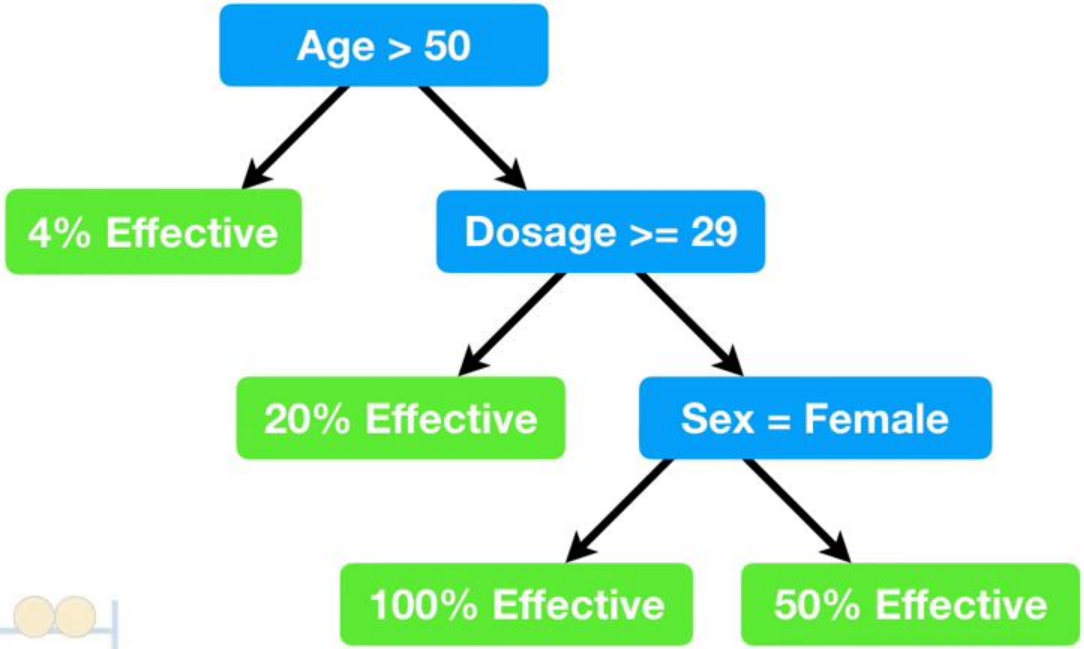
…otherwise we repeat the process to split the remaining observations…



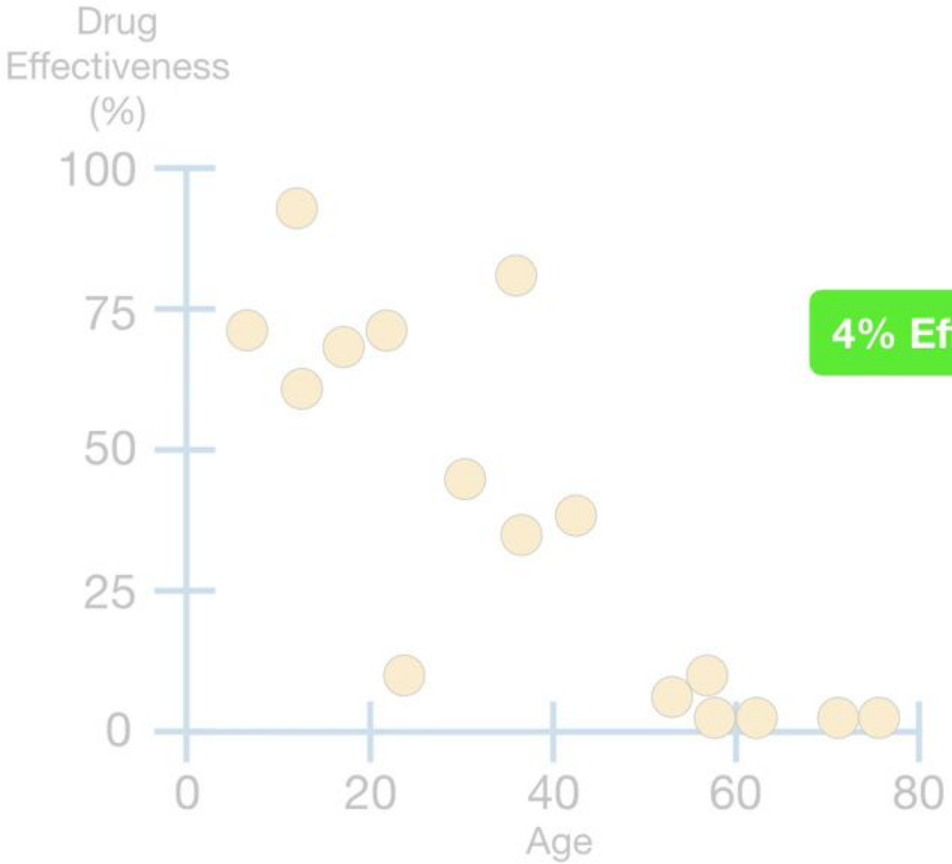


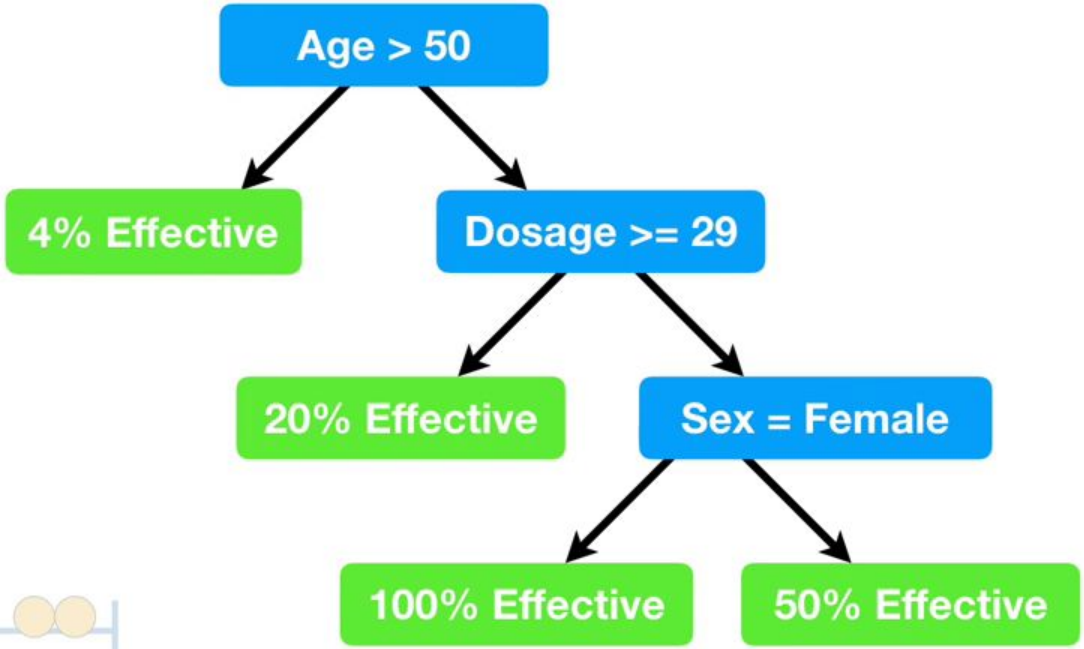
…until we can no longer split the observations into smaller groups…





…and then we are done.





The End!!!