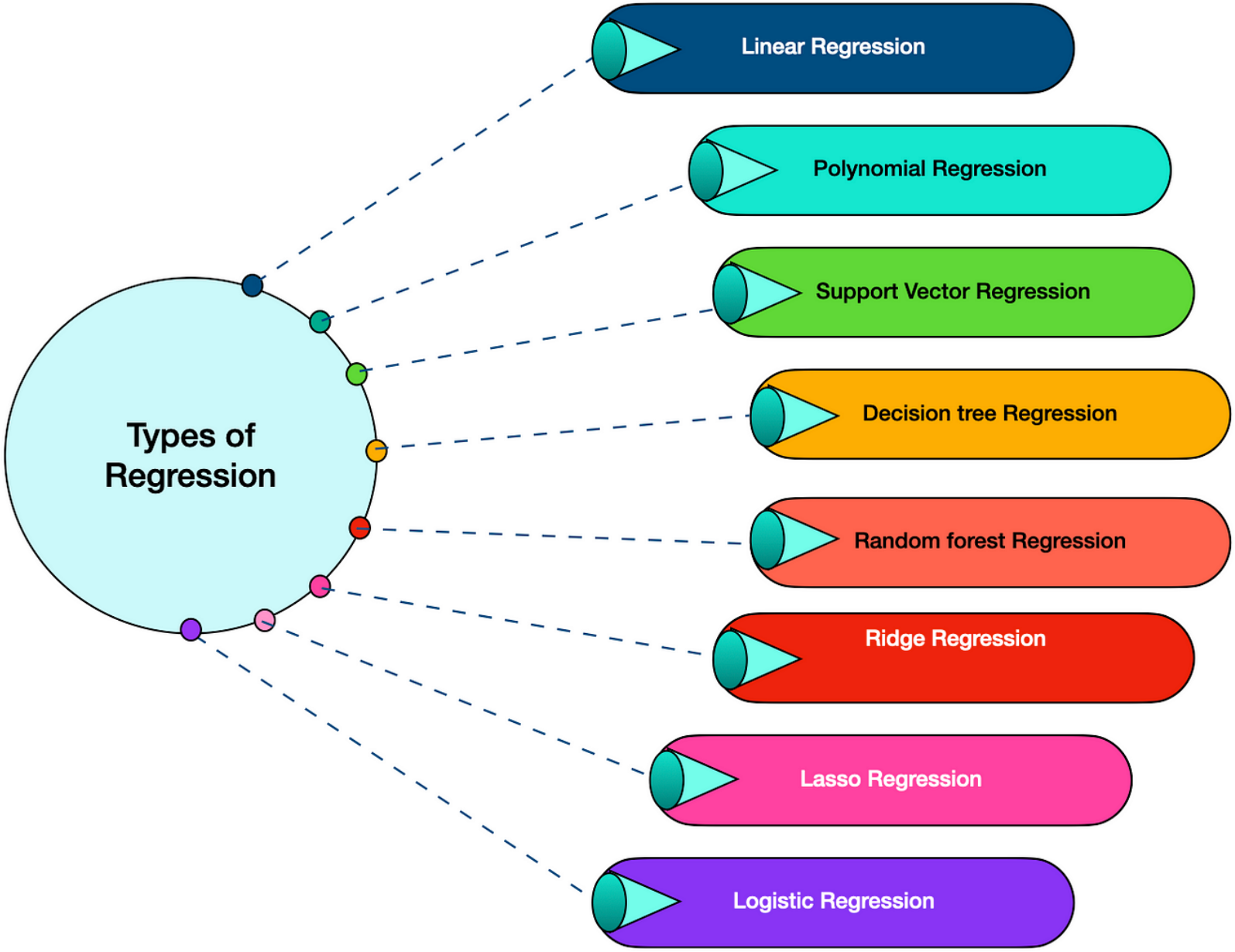
მონაცემთა ანალიტიკა Python

ლექცია 13:რეგრესია. გადაწყვეტილების ხე.

ლიკა სვანაძე
lika.svanadze@btu.edu.ge


Types of Regression
Linear Regression
Polynomial Regression
Support Vector Regression
Decision tree Regression
Random forest Regression
Ridge Regression
Lasso Regression
Logistic Regression

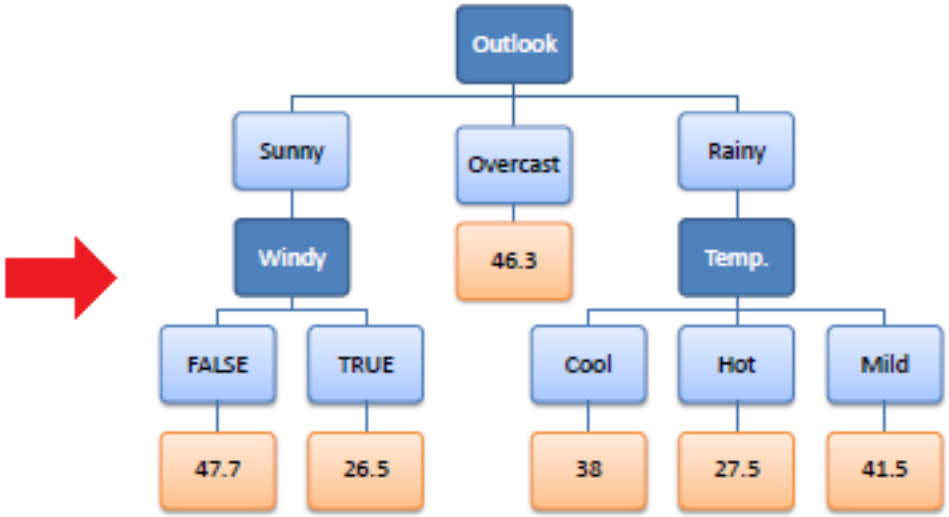
@arunp77

Decision Trees

Decision Tree - Regression

Decision tree builds regression or classification models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with **decision nodes** and **leaf nodes**. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called **root node**. Decision trees can handle both categorical and numerical data.

Predictors				Target
Outlook	**Temp.**	**Humidity**	**Windy**	**Hours Played**
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
Overcast	Hot	High	False	48
Sunny	Mild	High	False	46
Sunny	Cool	Normal	False	62
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	36
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	48
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	62
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30



Imagine we developed a new drug…

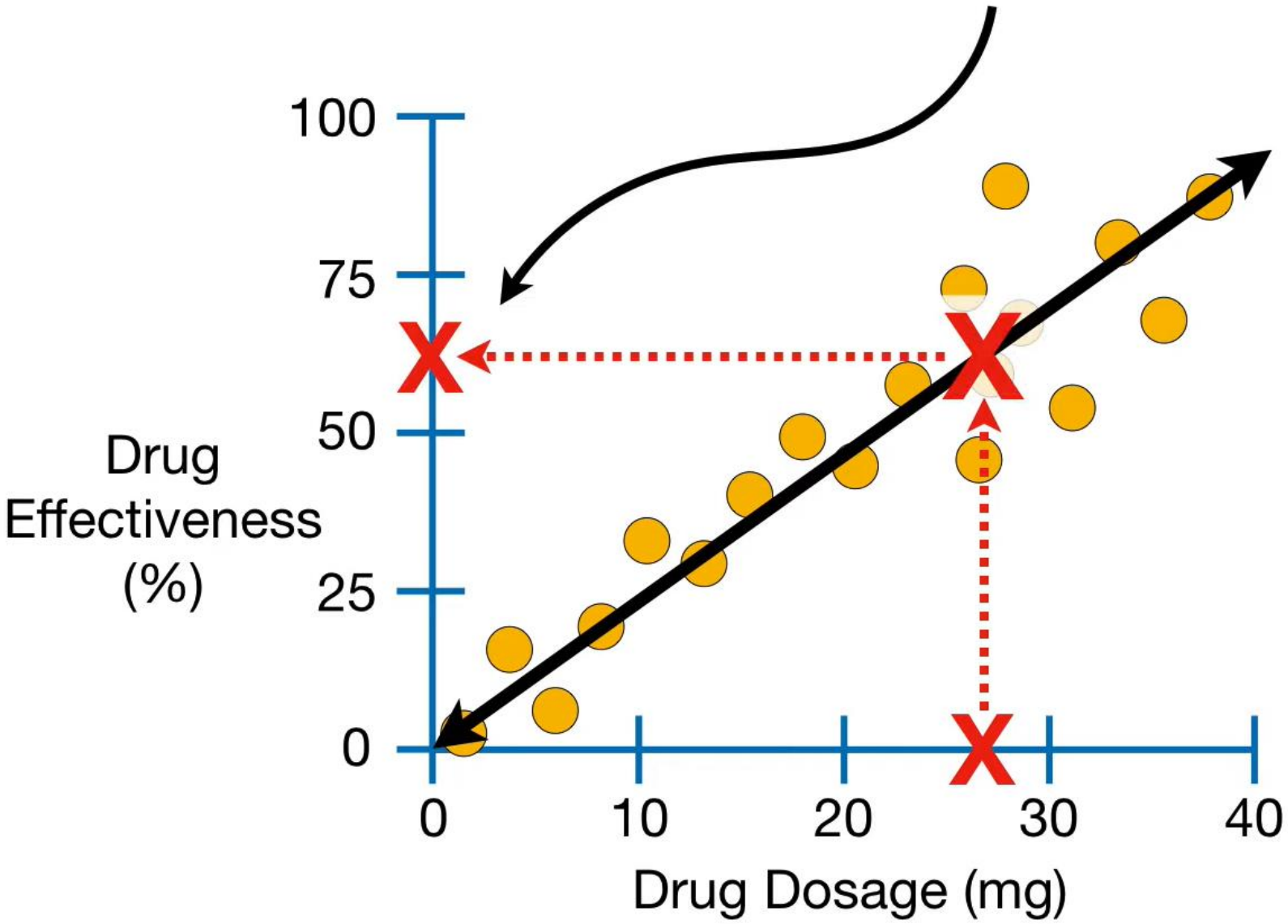
…to cure the common cold.

VS.

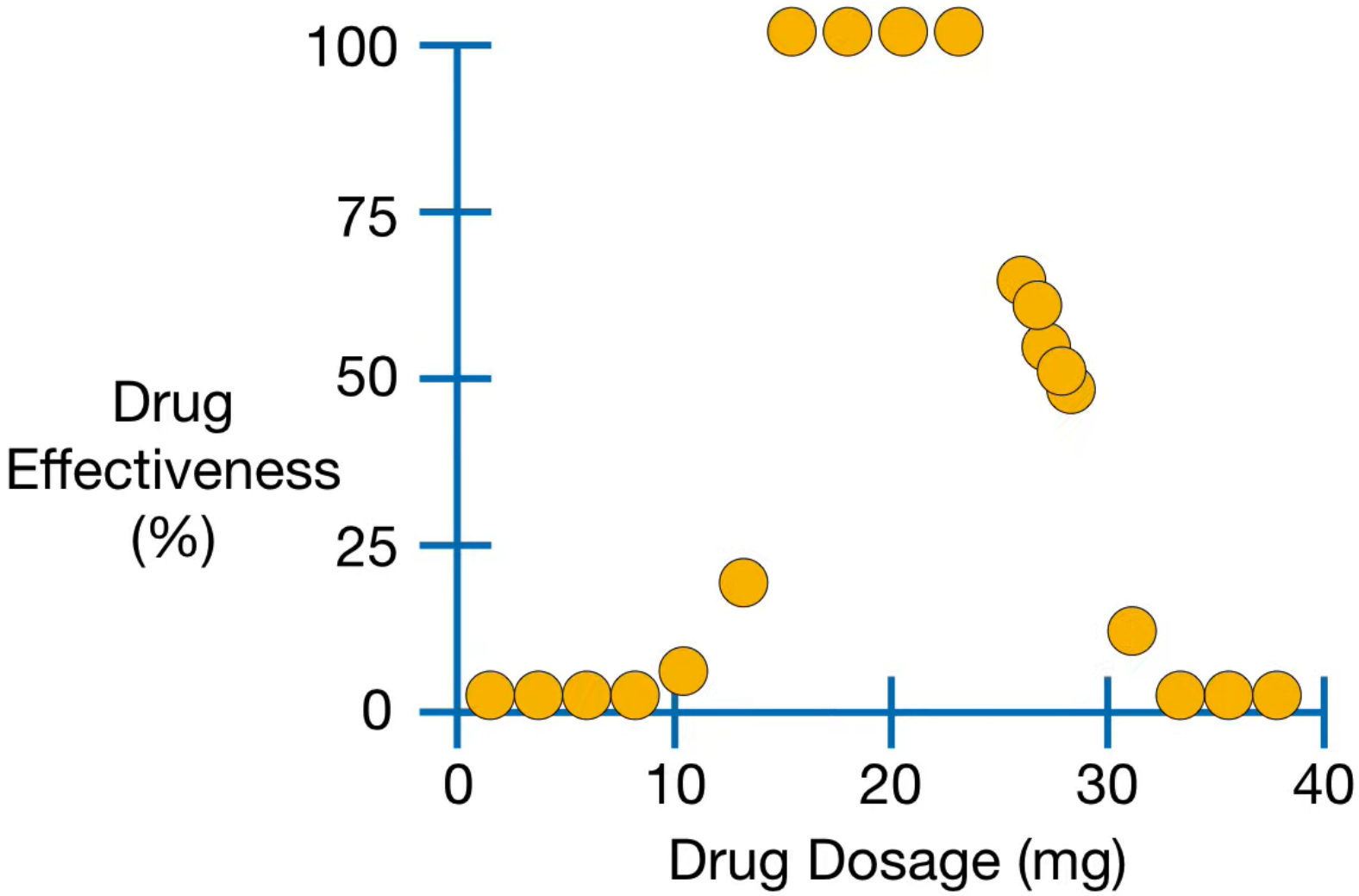
However, we don't know the optimal dosage to give to patients.

VS.

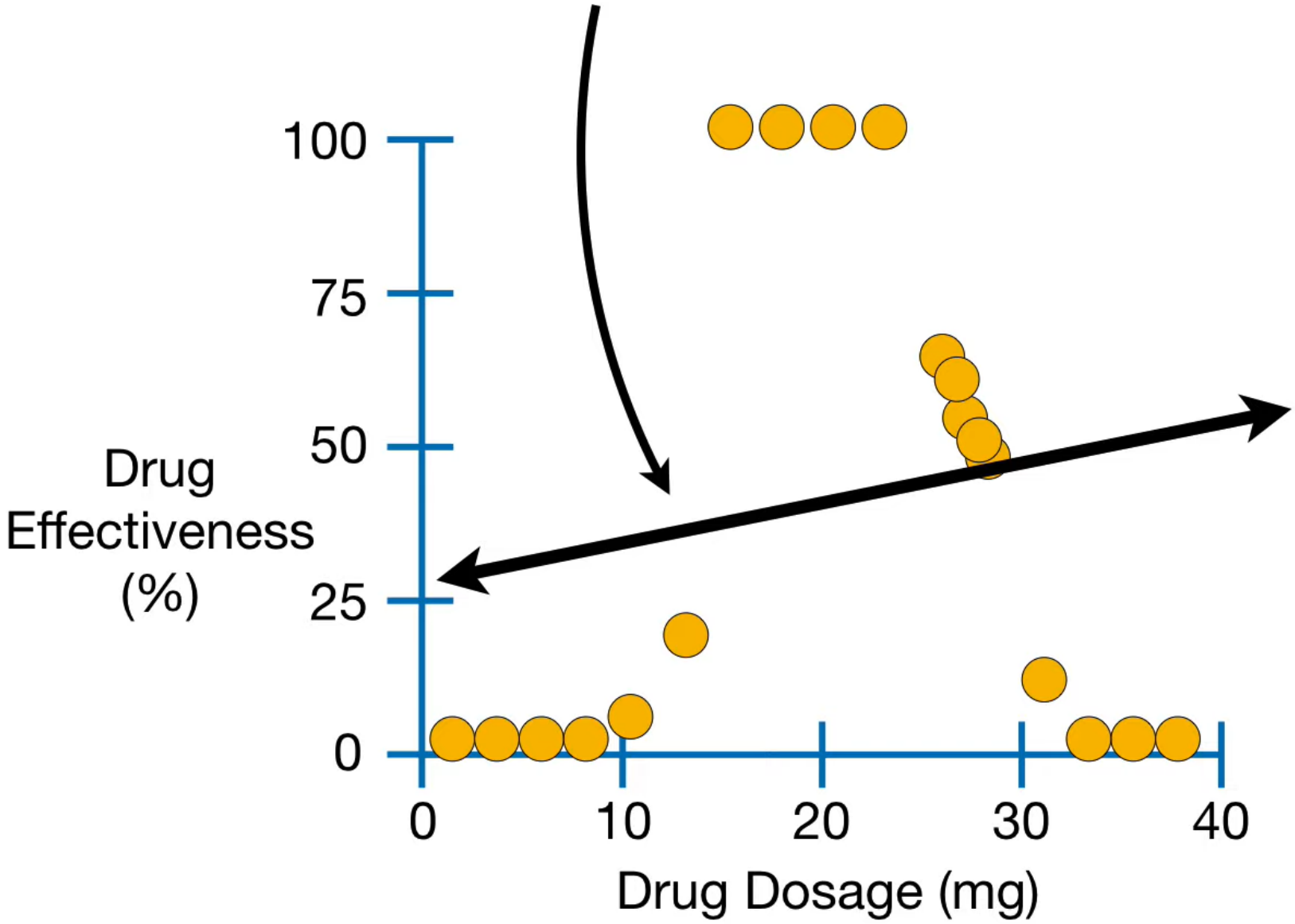
…we could use the line to predict that a **27 mg Dose** should be **62% Effective**.

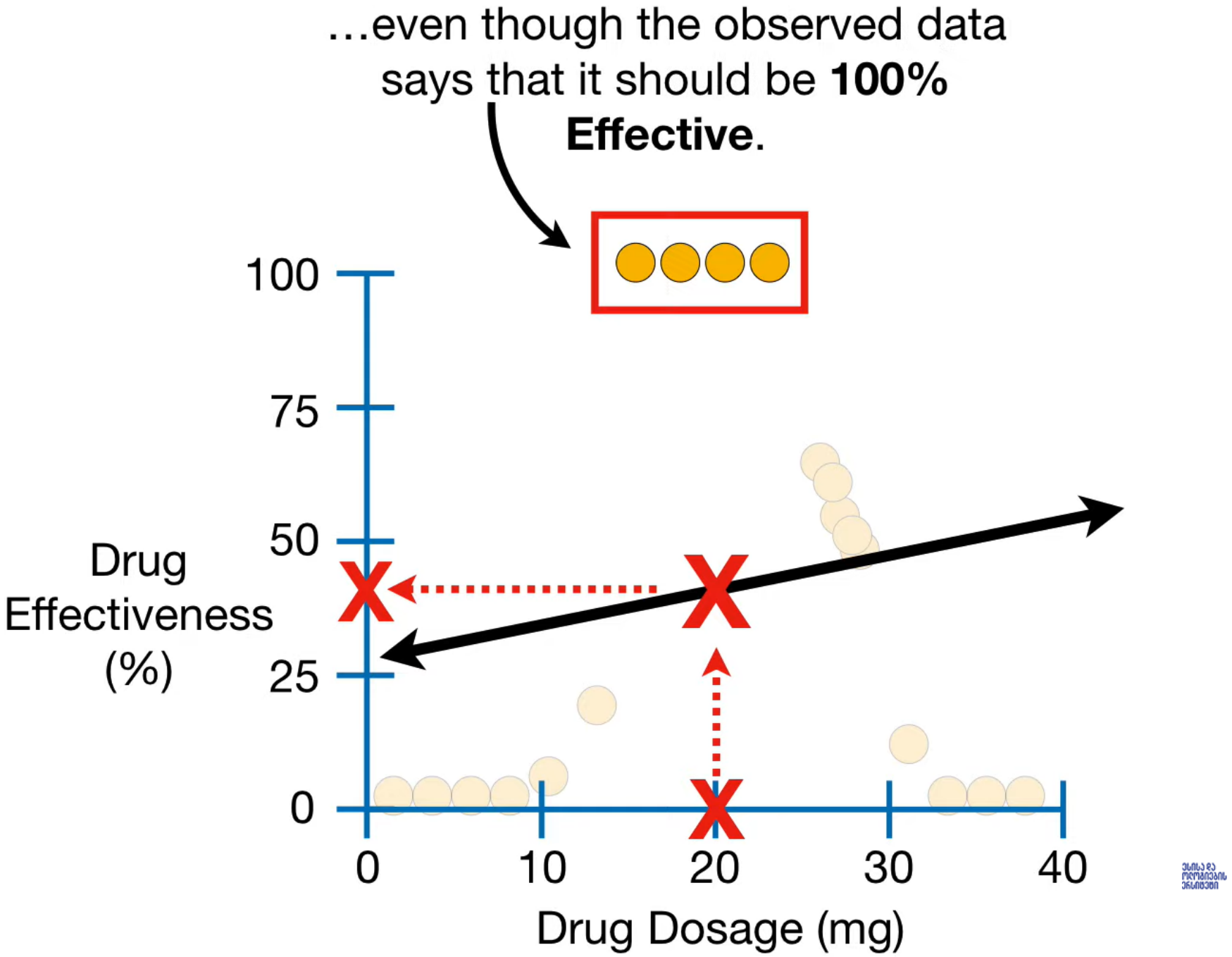


However, what if the data looked like this?

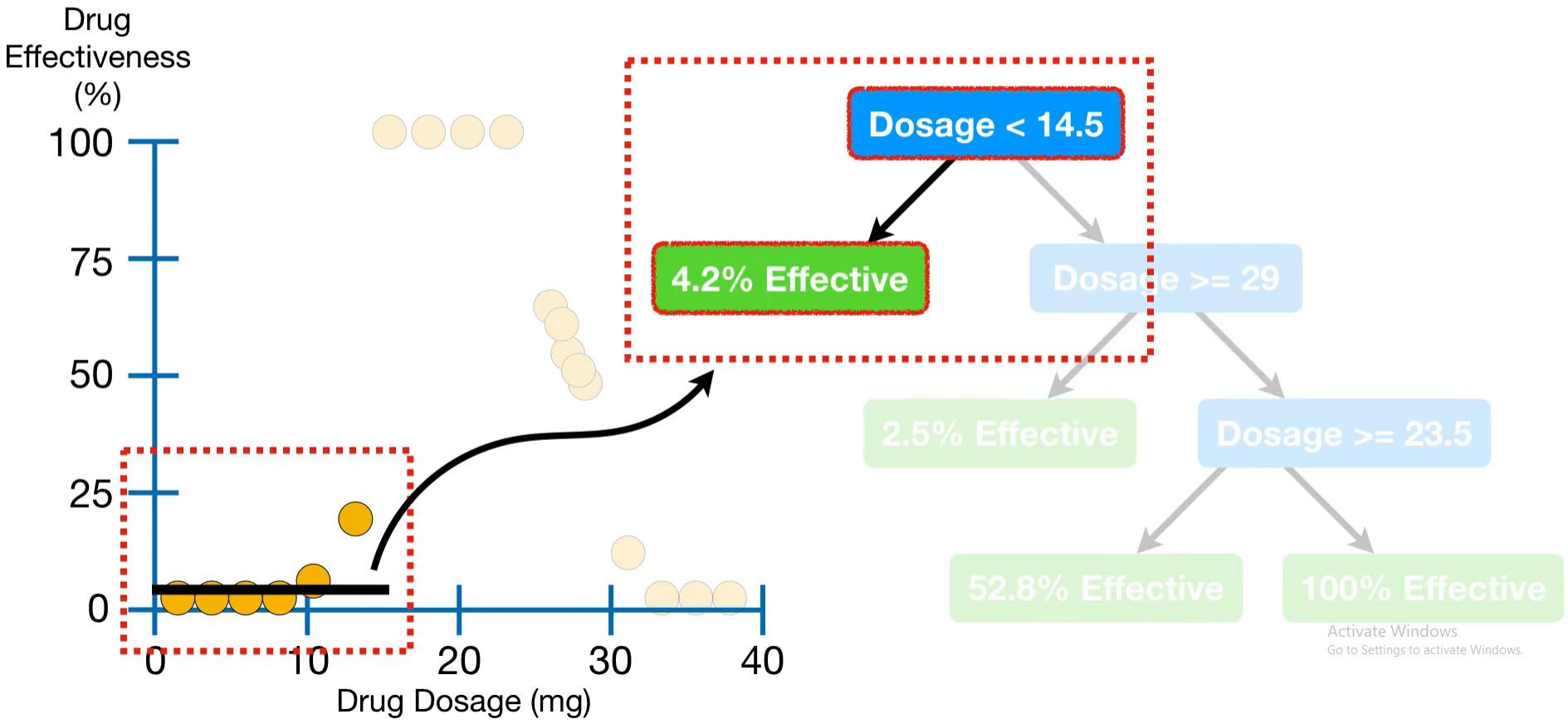


In this case, fitting a straight line to the data will not be very useful.

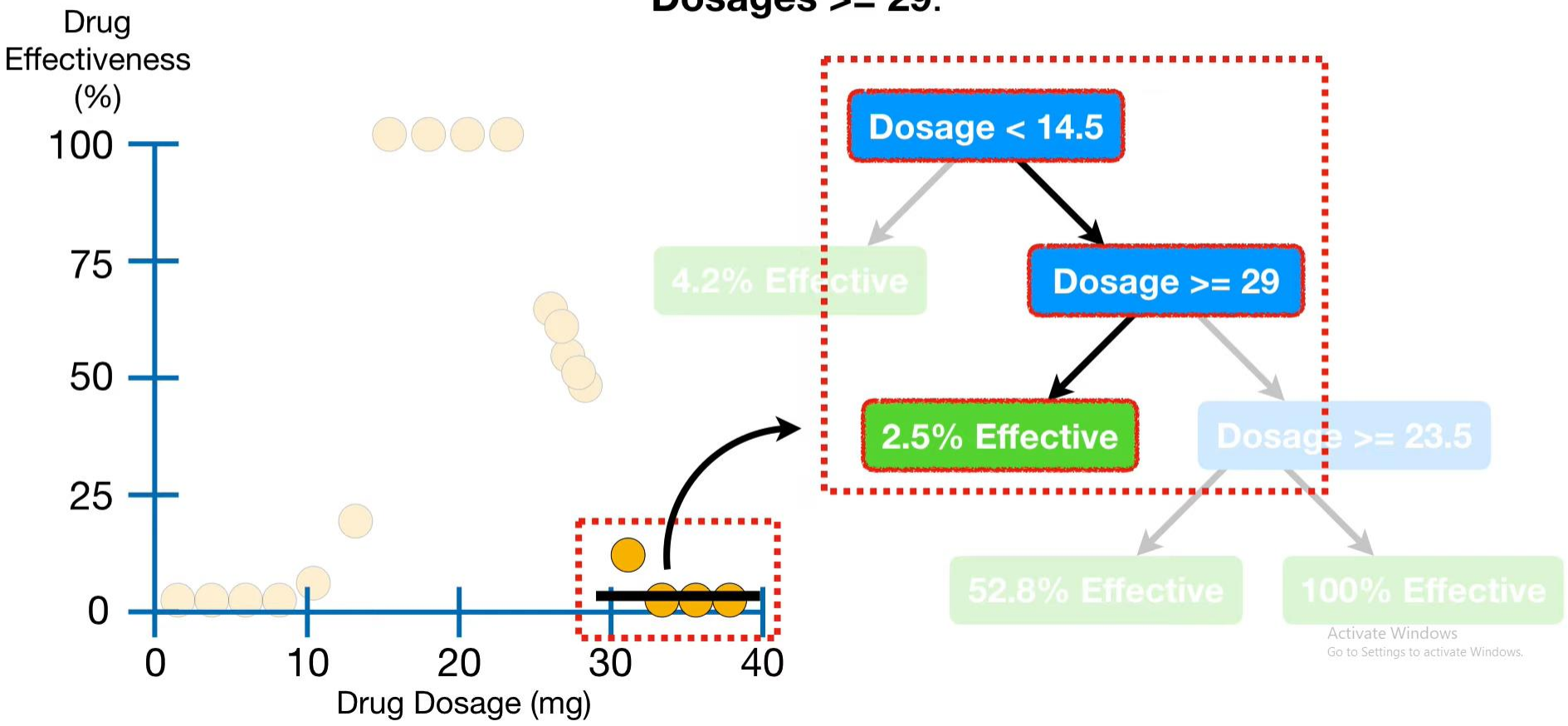



…even though the observed data says that it should be **100% Effective**.
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)

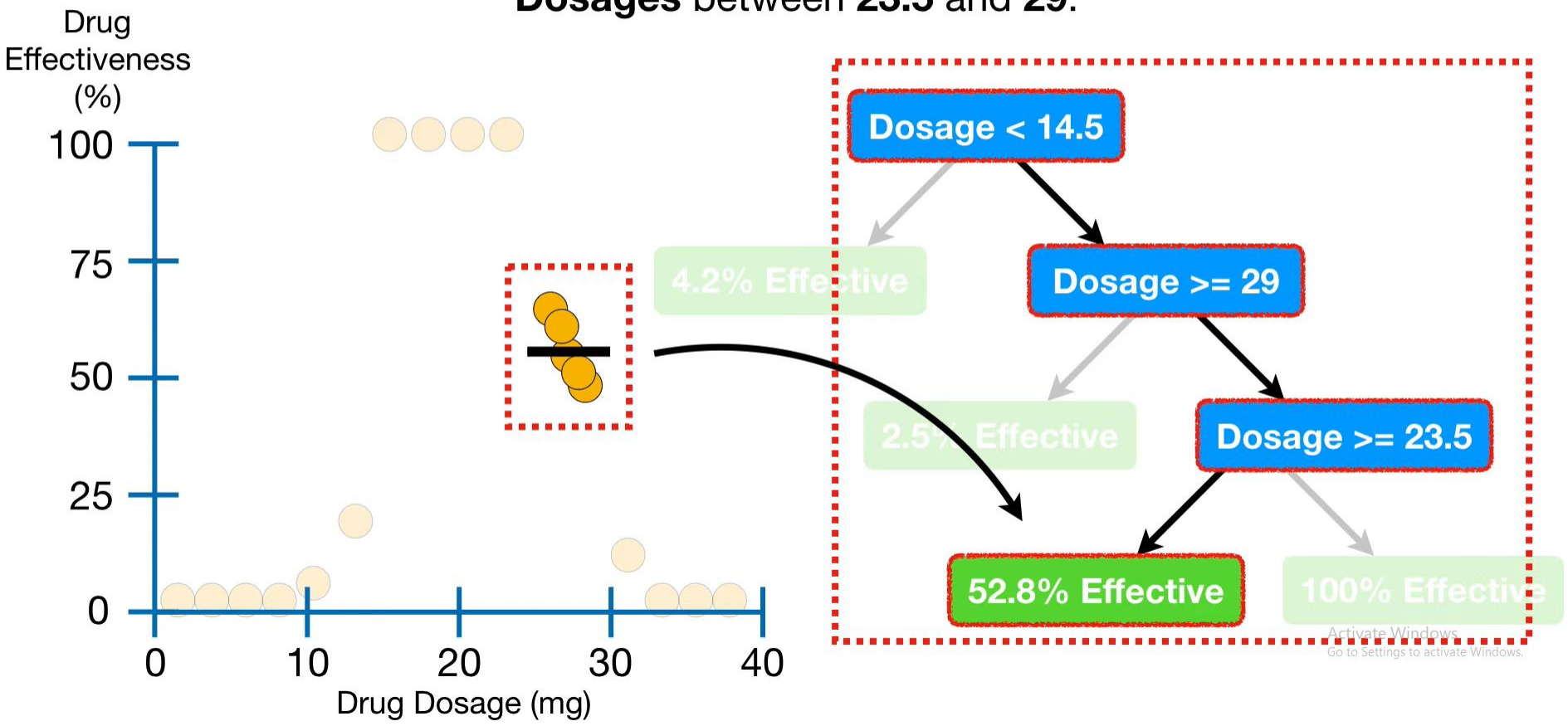
…so the tree uses the average value, **4.2%**, as its prediction for people with **Dosages < 14.5**.



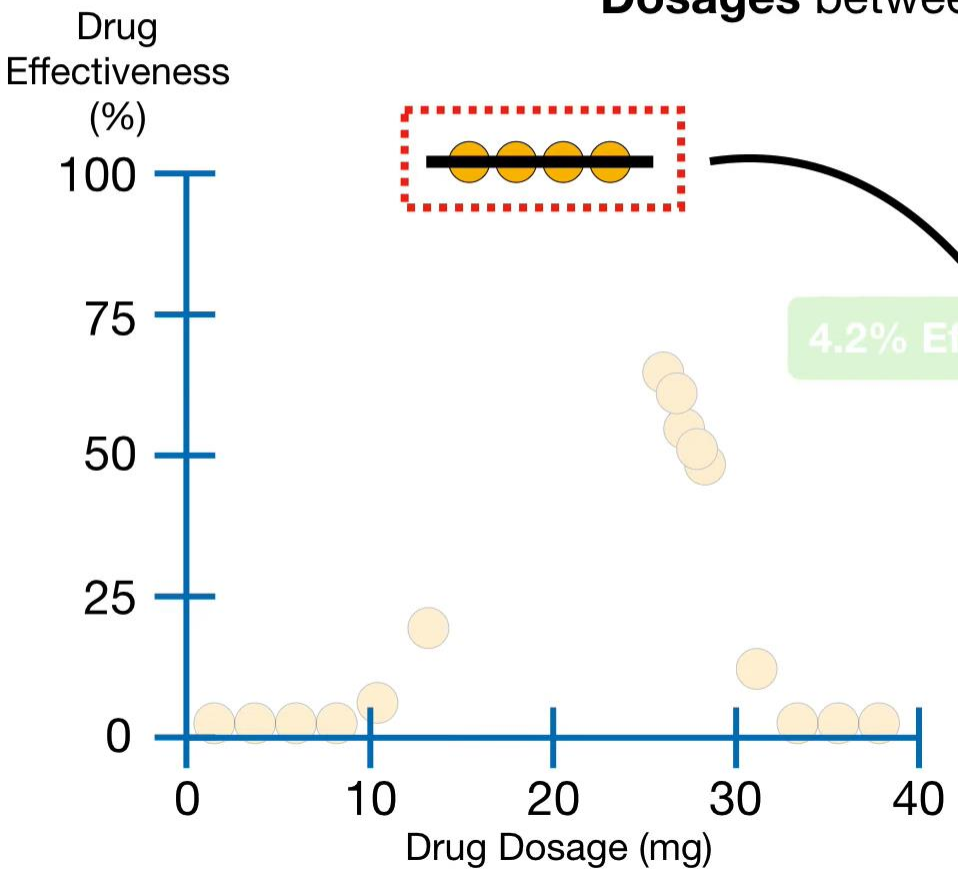
…so the tree uses the average value, **2.5%**, as its prediction for people with **Dosages >= 29**.

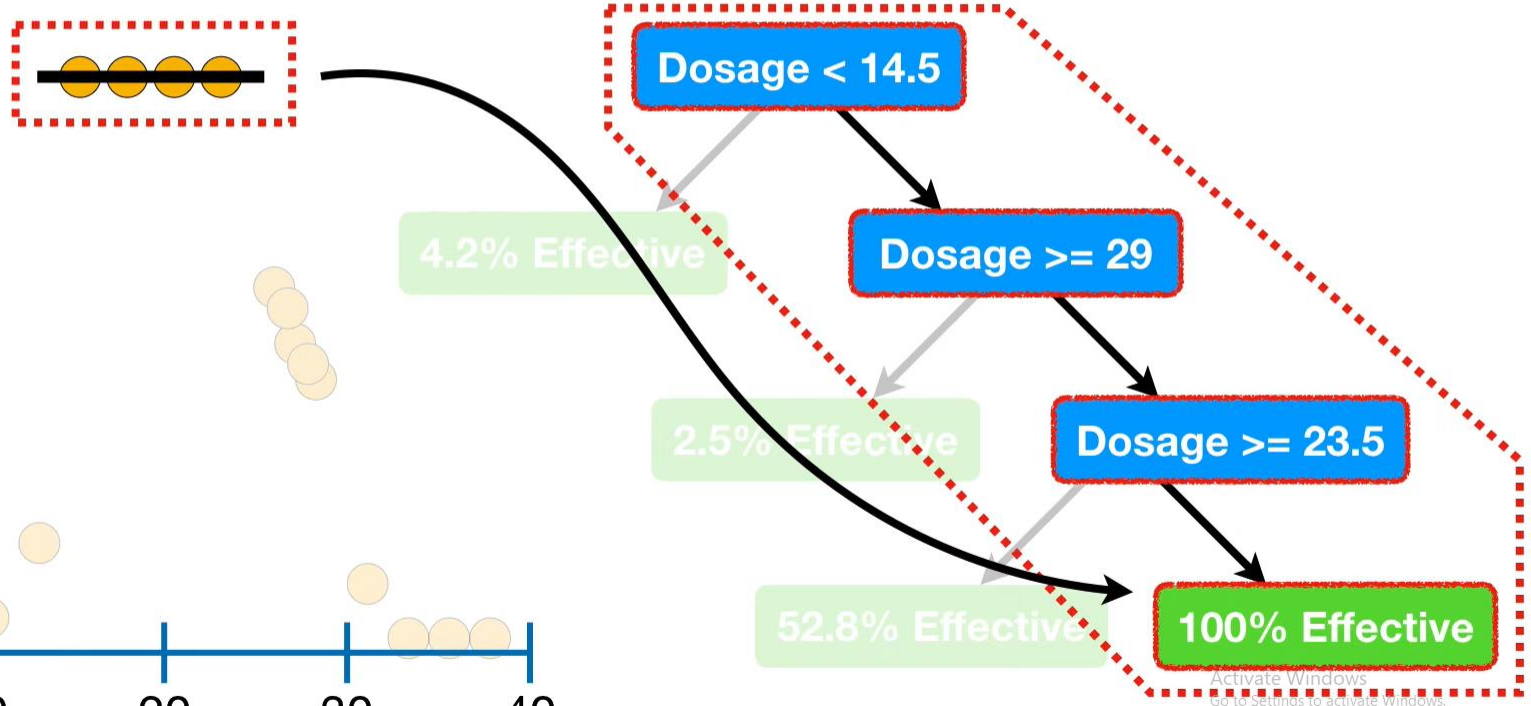


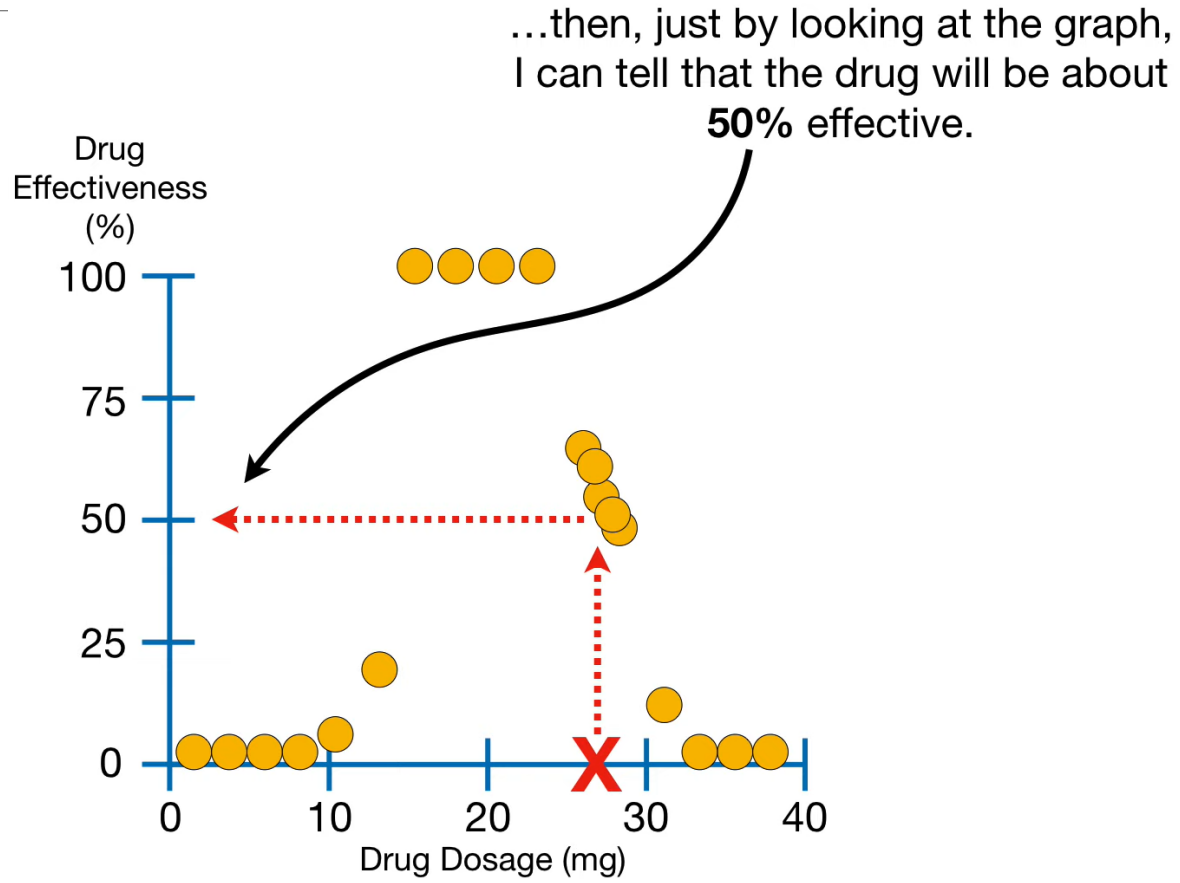
…so the tree uses the average value, **52.8%**, as its prediction for people with **Dosages** between **23.5** and **29**.



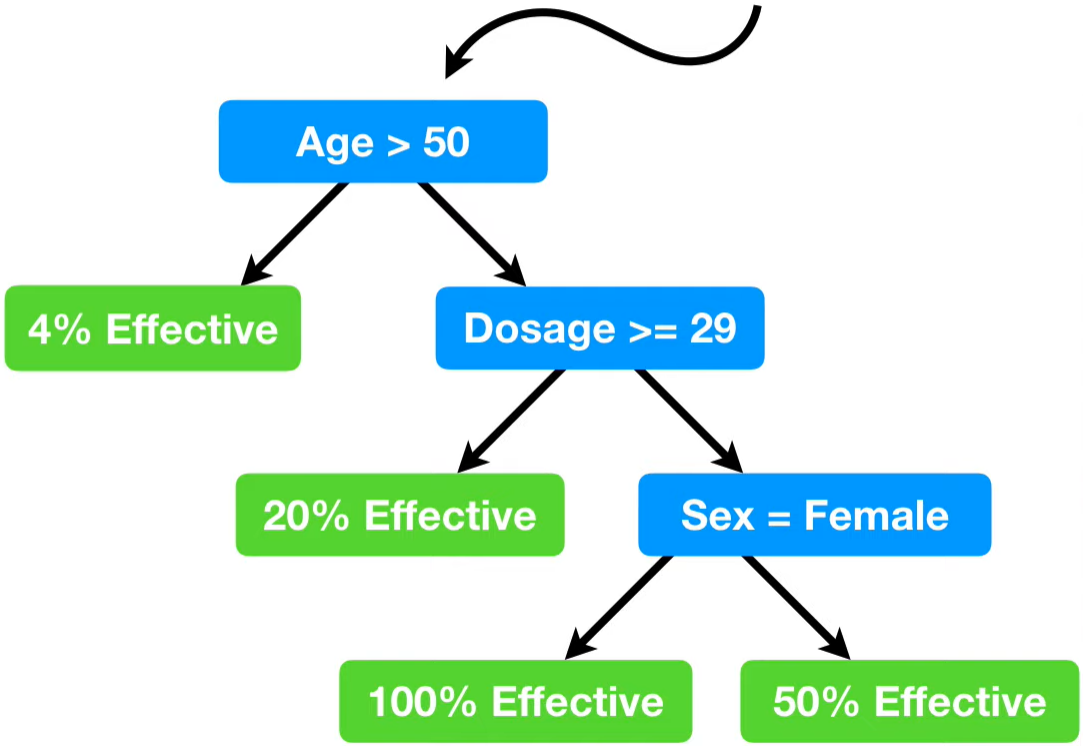
…so the tree uses the average value, **100%**, as its prediction for people with **Dosages** between **14.5** and **23.5**.





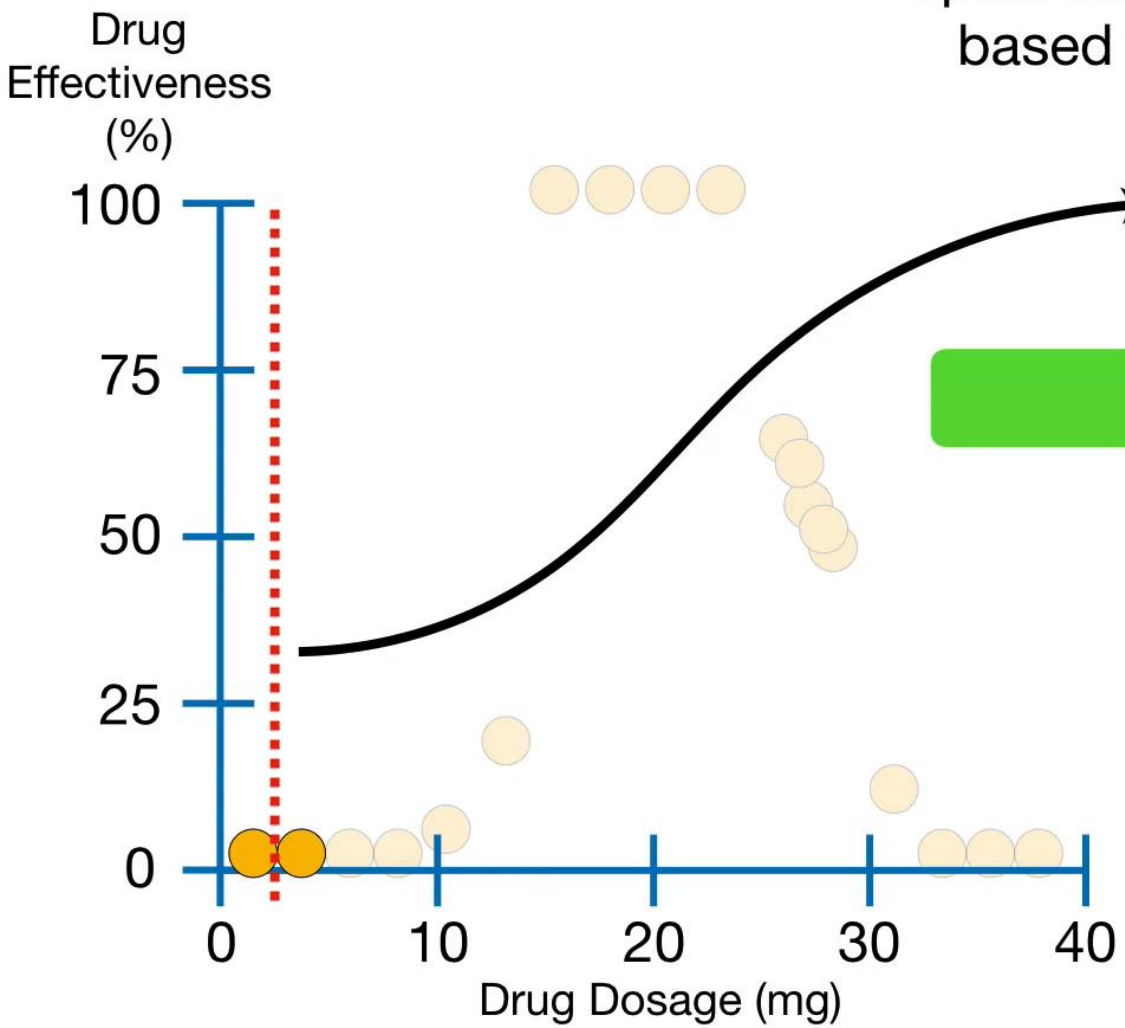

…then, just by looking at the graph,
I can tell that the drug will be about
50% effective.
Drug
Effectiveness
(%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)

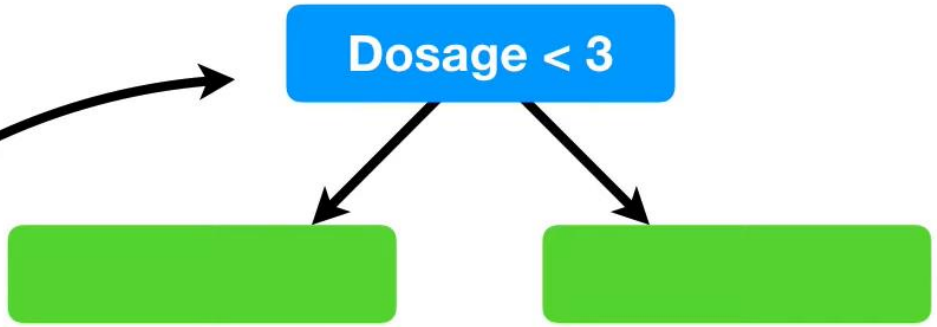
In contrast, a **Regression Tree** easily accommodates the additional predictors.



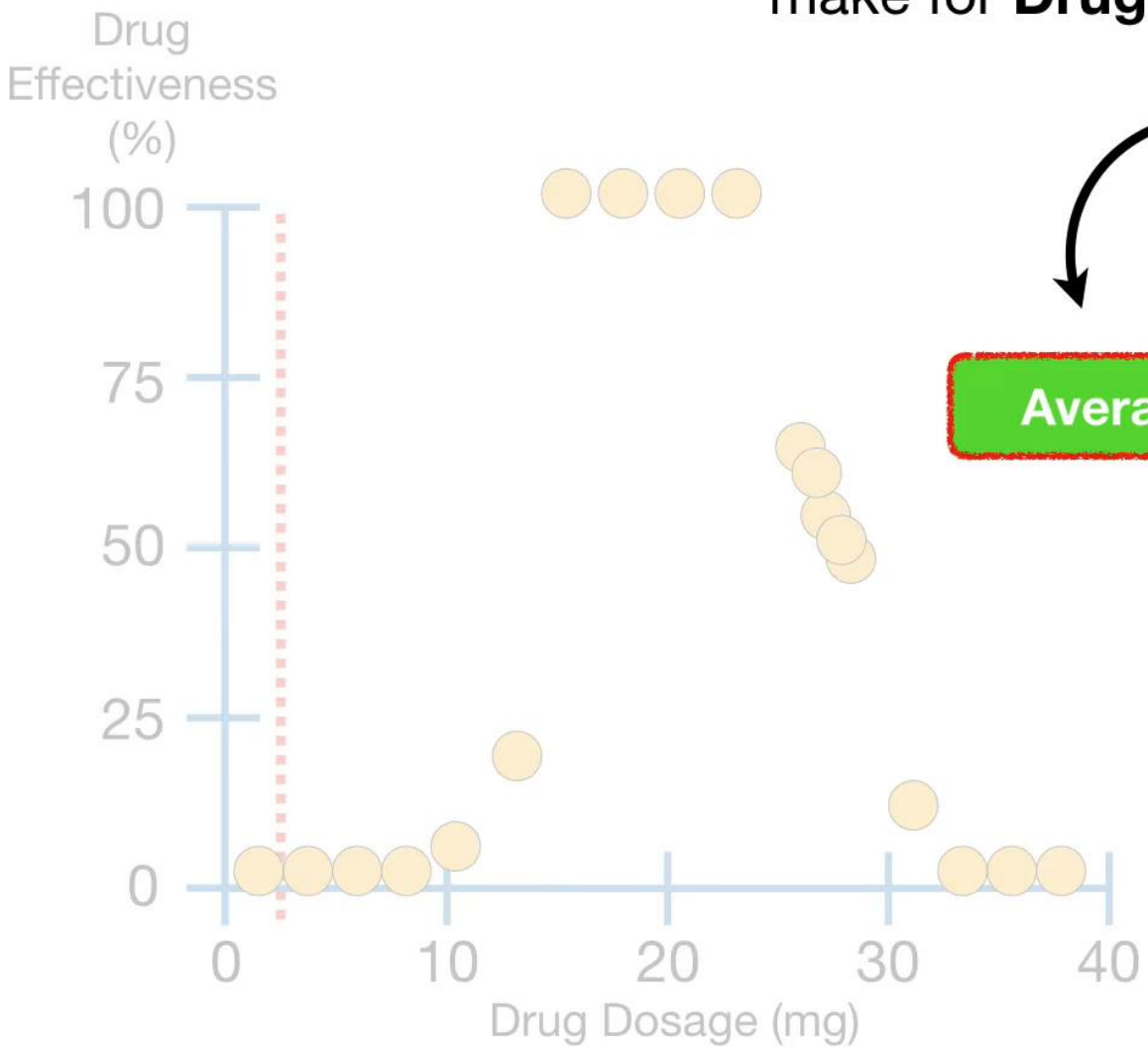
Dosage	Age	Sex	Etc.	Drug Effect.
10	25	Female	…	98
20	73	Male	…	0
35	54	Female	…	100
5	12	Male	…	44
etc…	etc…	etc…	etc…	etc…

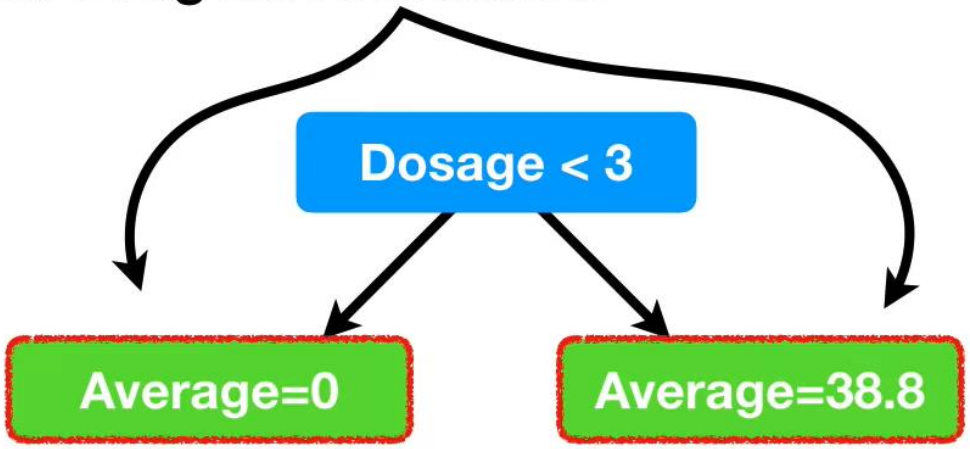
Now we can build a very simple tree that splits the observations into two groups based whether or not **Dosage < 3**.

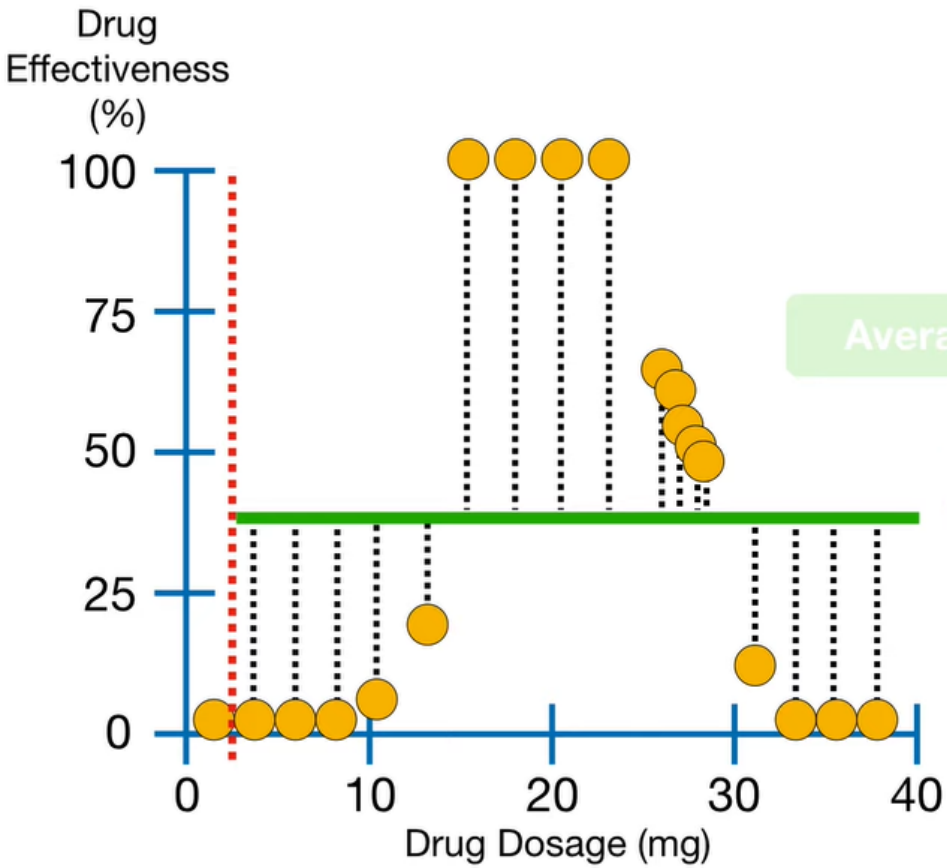




The values in each leaf are the predictions that this simple tree will make for **Drug Effectiveness**.







…and get **27,468.5**.

Dosage < 3

Average=0

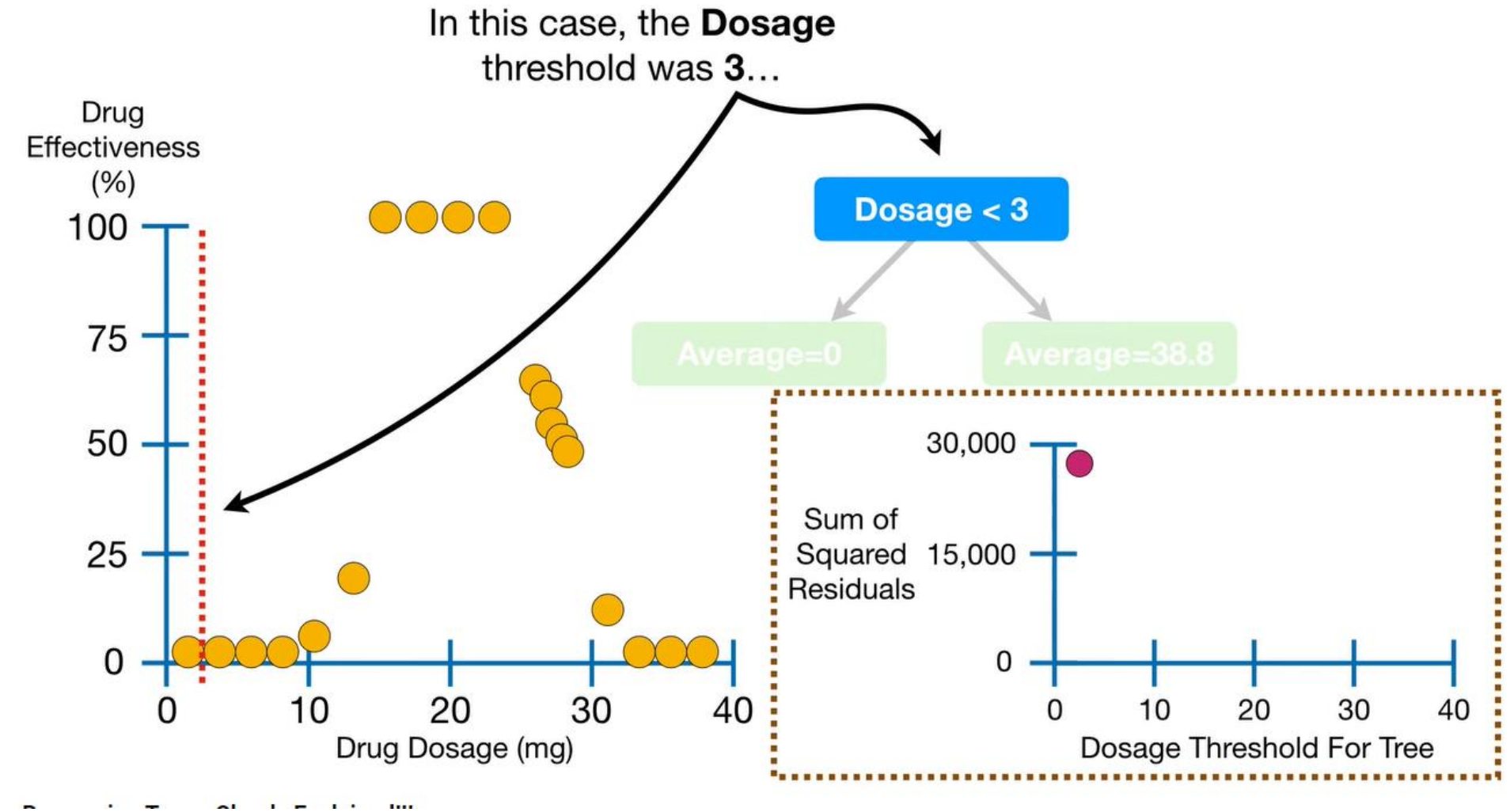
Average=38.8

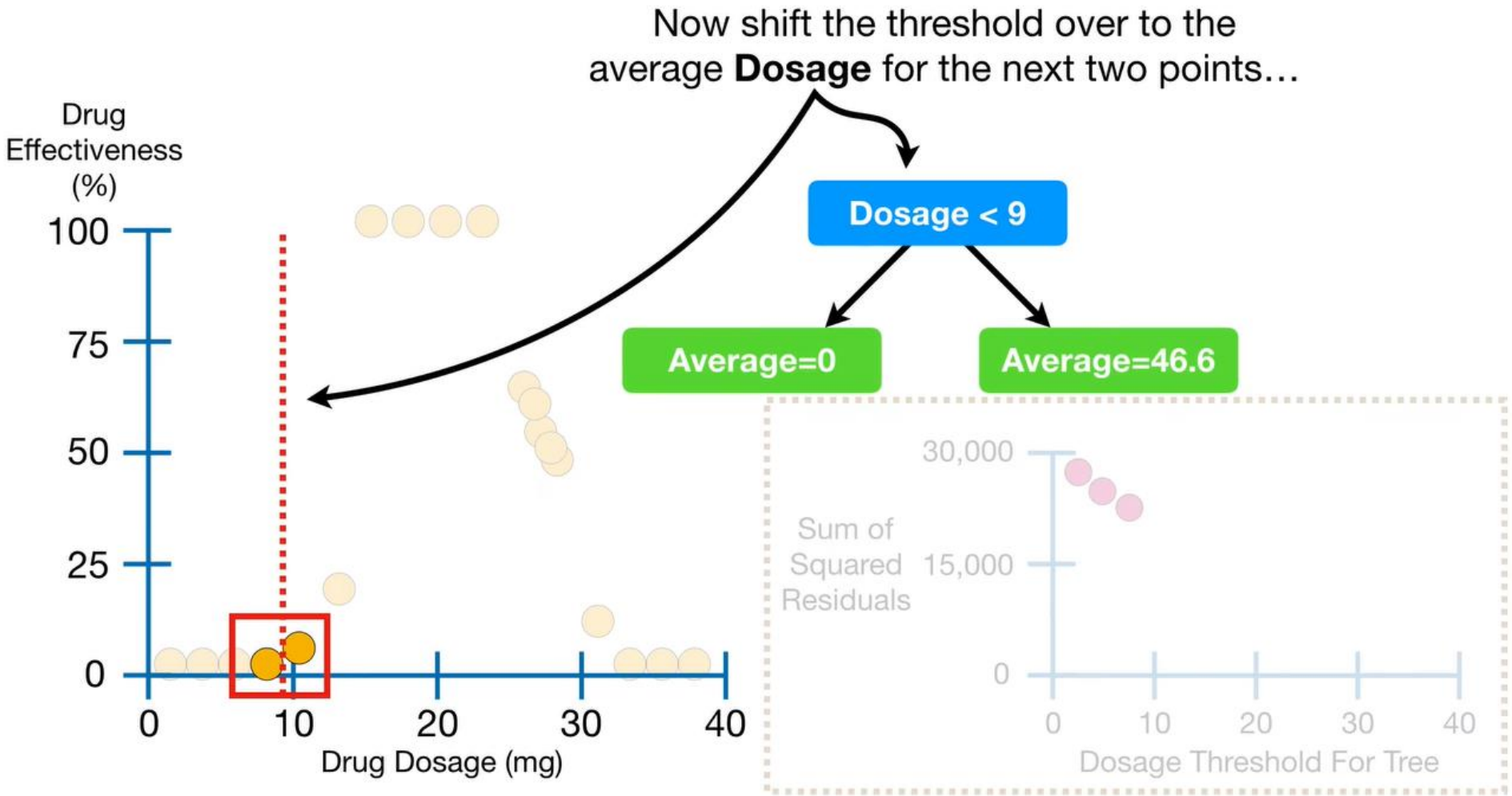
$(0 - 0)^2 + (0 - 38.8)^2 + (0 - 38.8)^2 + (0 - 38.8)^2$

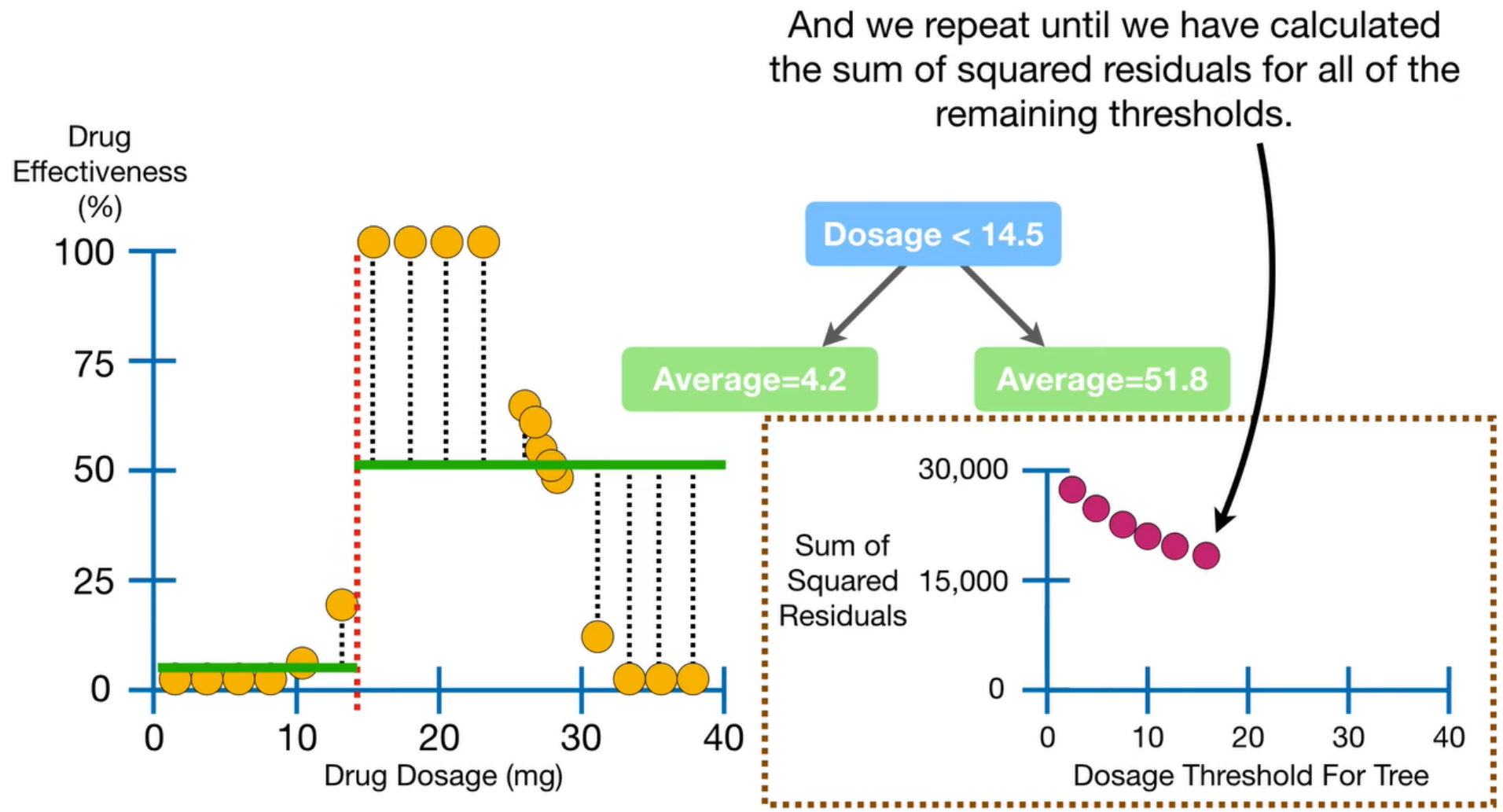
$+ (5 - 38.8)^2 + (20 - 38.8)^2 + (00 - 38.8)^2$

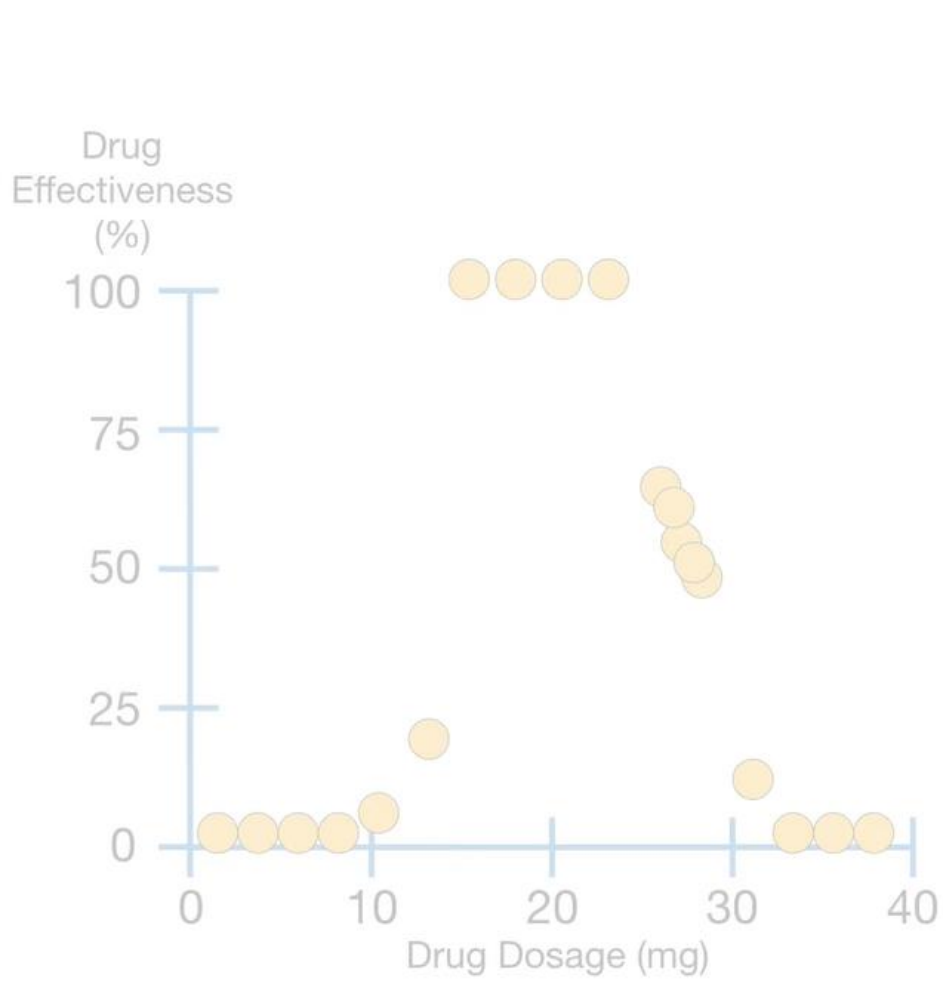
$+ (100 - 38.8)^2 + \ldots + (0 - 38.8)^2$

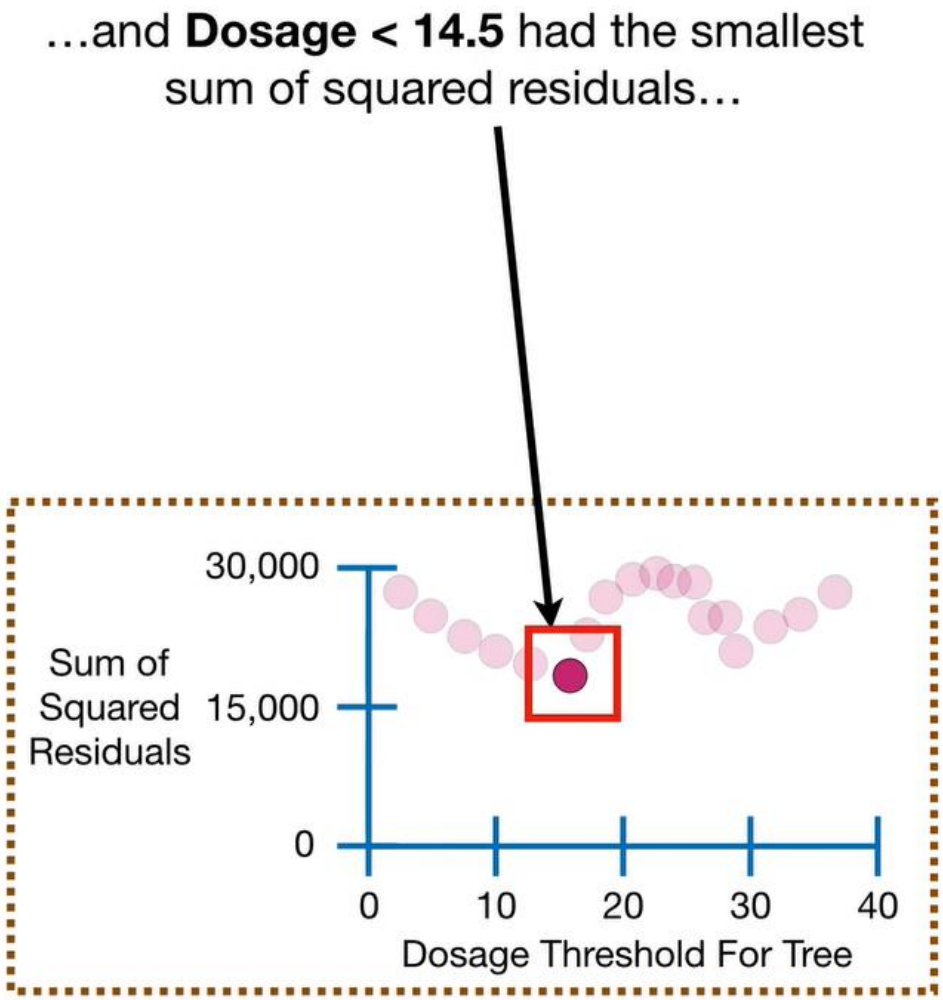
$= 27,468.5$


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

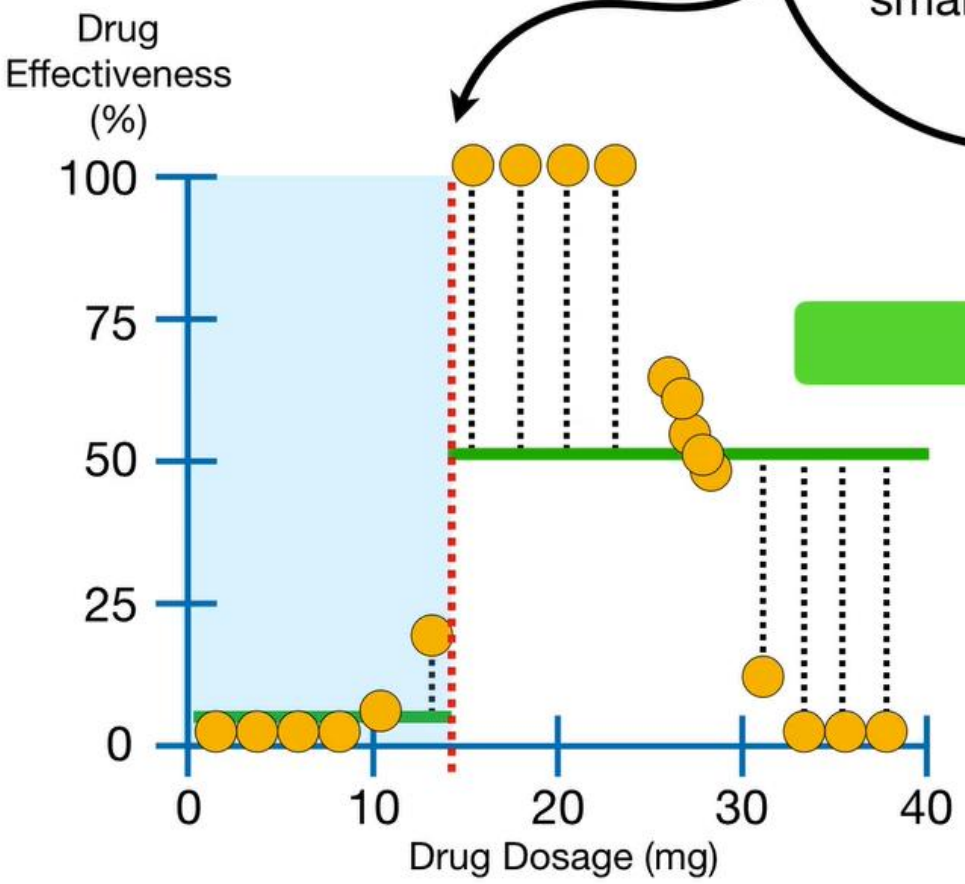

Now shift the threshold over to the average **Dosage** for the next two points…
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 9
Average=0
Average=46.6
30,000
15,000
0
Sum of Squared Residuals
0
10
20
30
40
Dosage Threshold For Tree

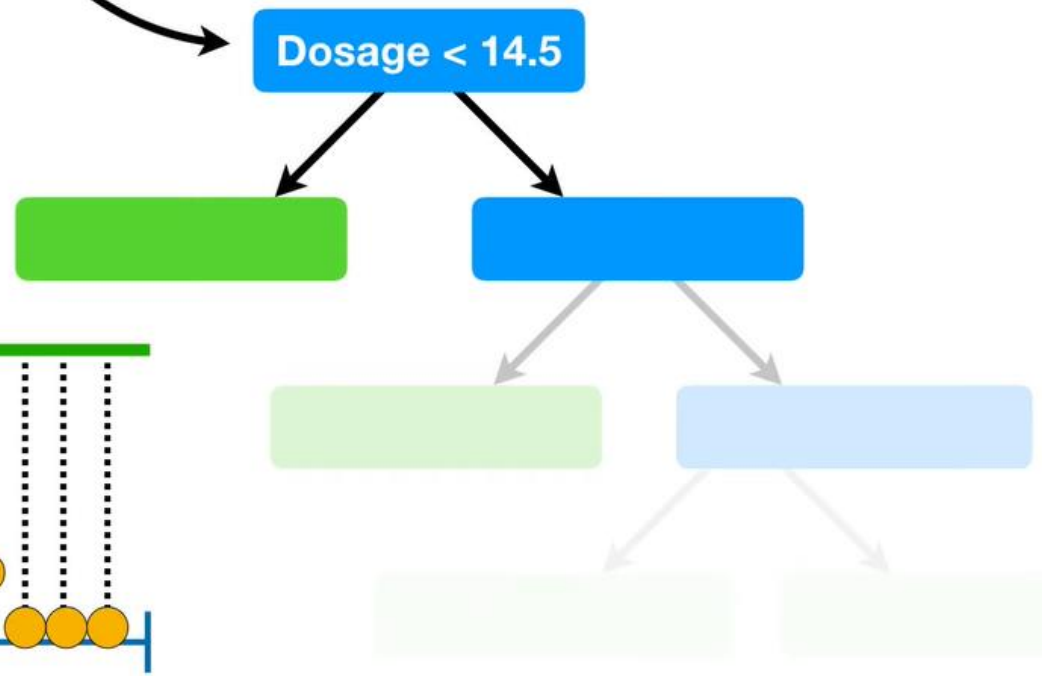

And we repeat until we have calculated the sum of squared residuals for all of the remaining thresholds.
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
Average=4.2
Average=51.8
Sum of Squared Residuals
30,000
15,000
0
0
10
20
30
40
Dosage Threshold For Tree


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

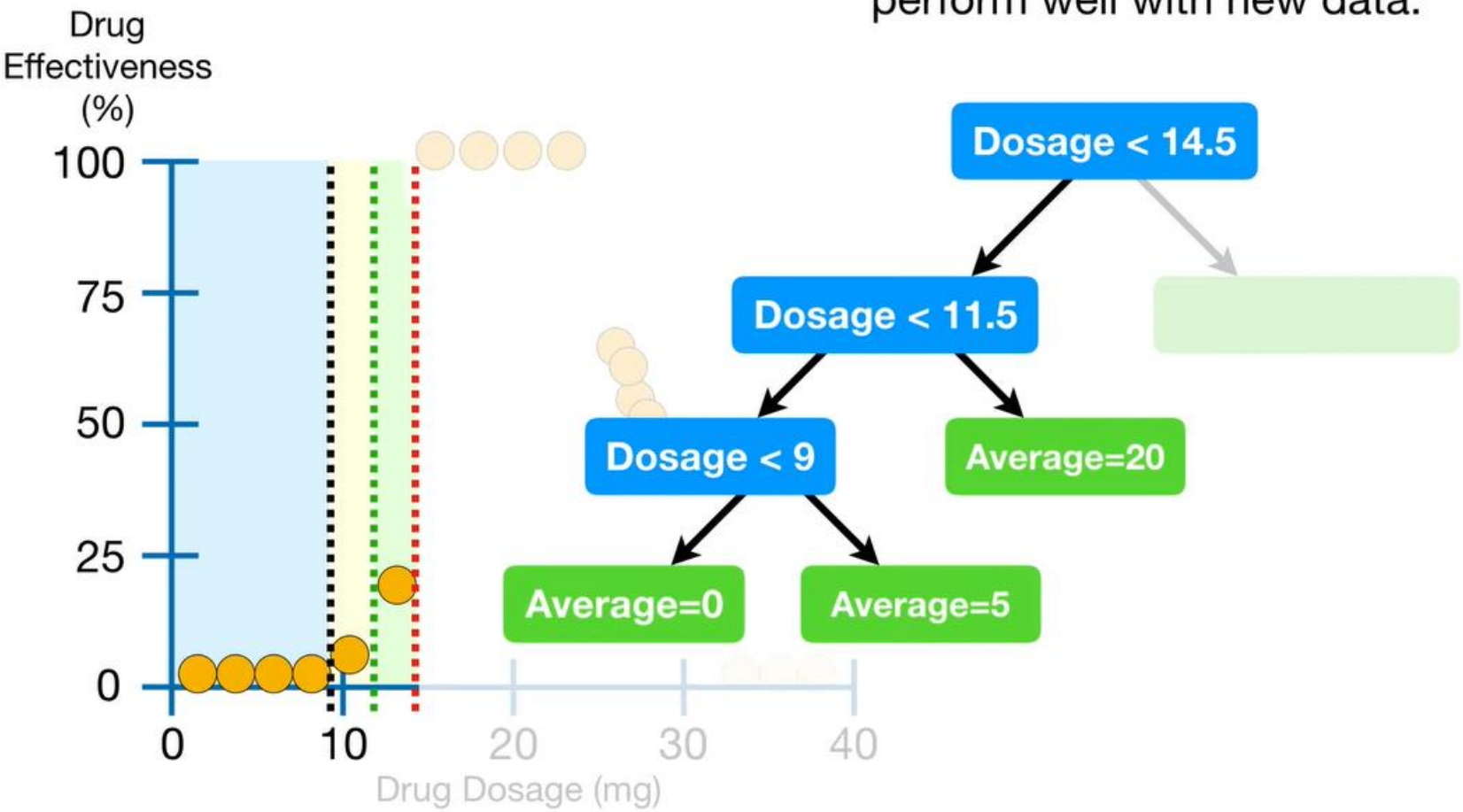

…and Dosage < 14.5 had the smallest sum of squared residuals…
30,000
15,000
0
Sum of Squared Residuals
0
10
20
30
40
Dosage Threshold For Tree

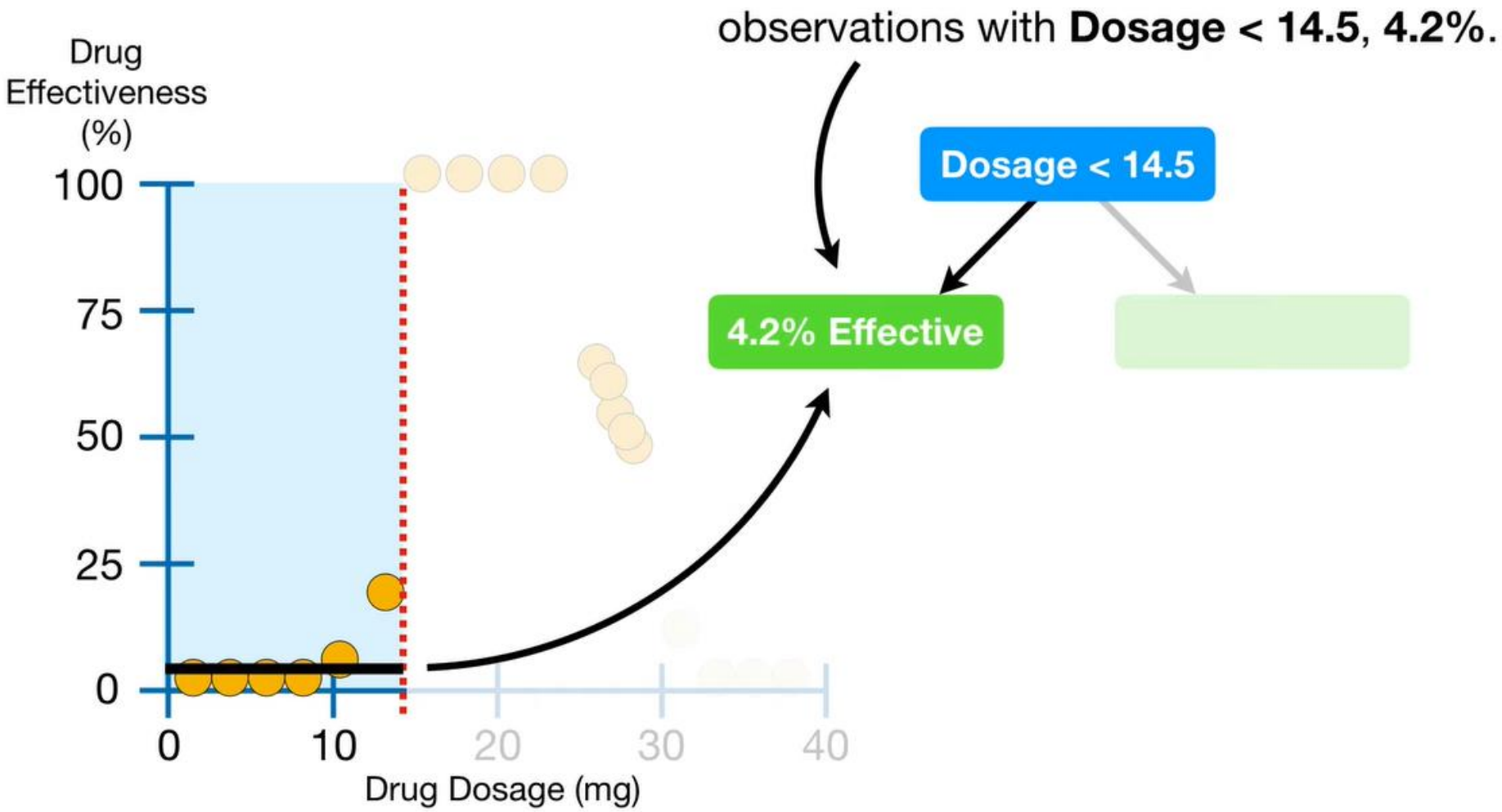
In summary, we split the data into two groups by finding the threshold that gave us the smallest sum of squared residuals.

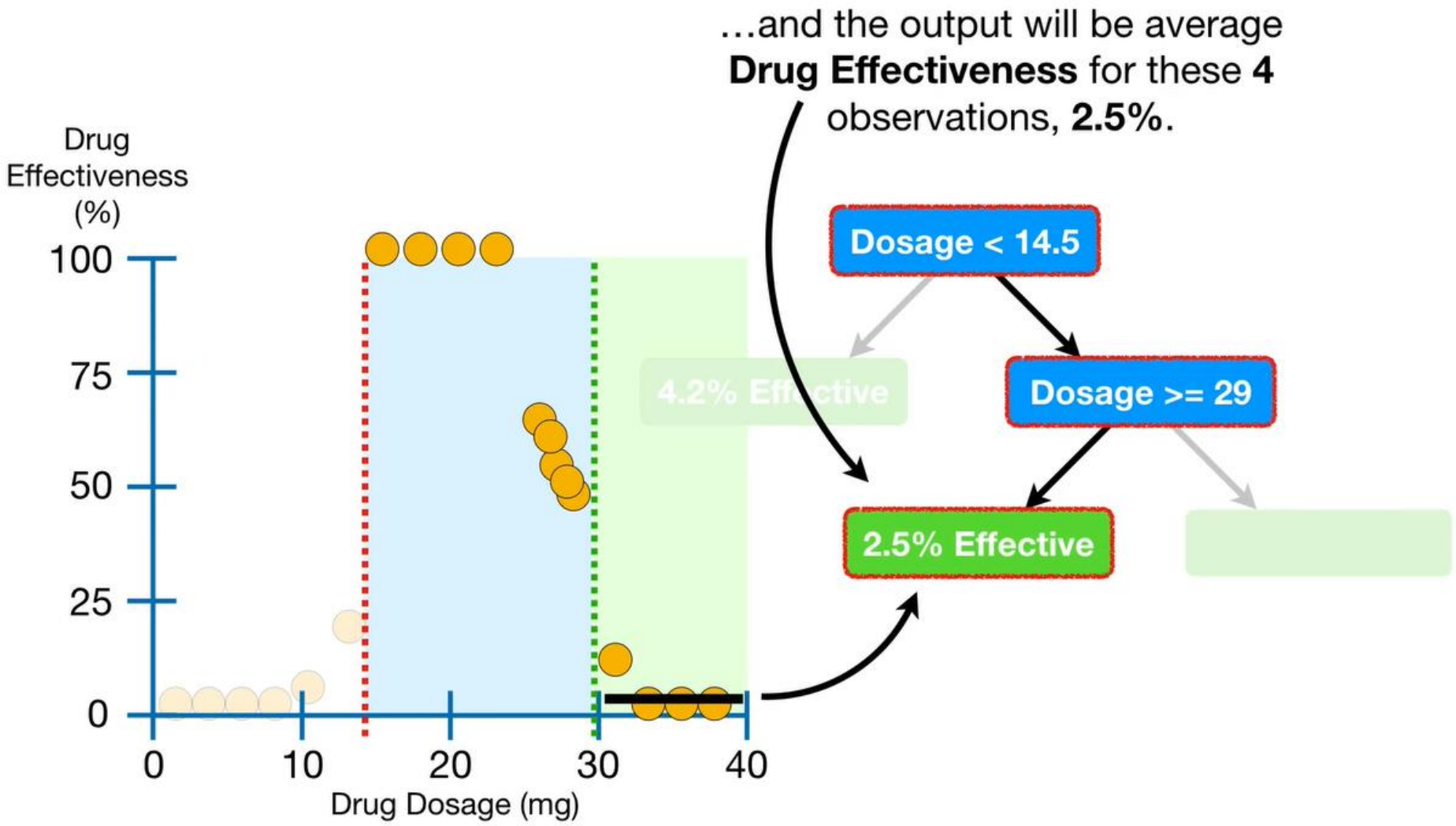


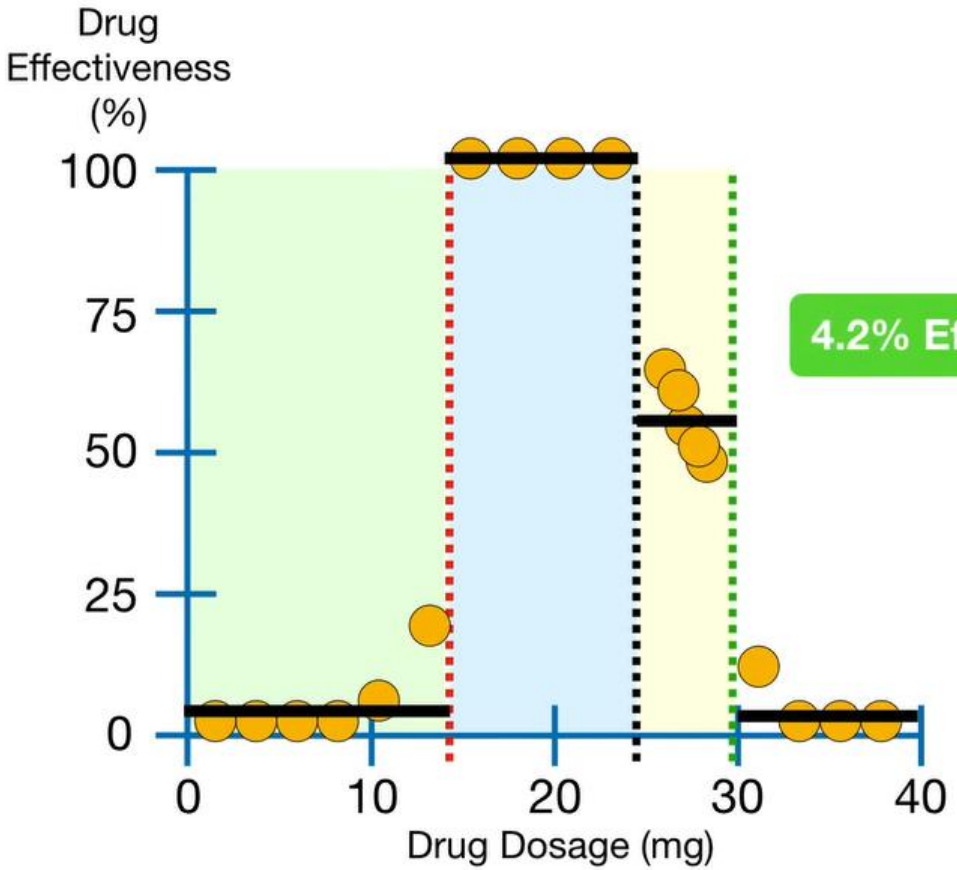


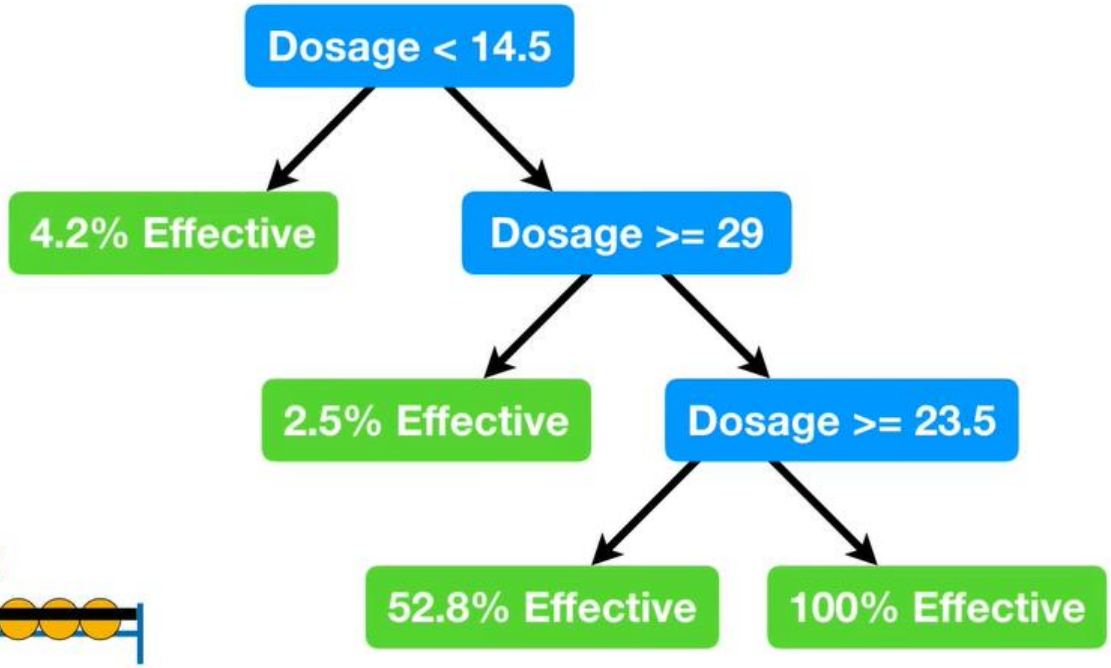
When a model fits the training data perfectly, it probably means it is overfit and will not perform well with new data.



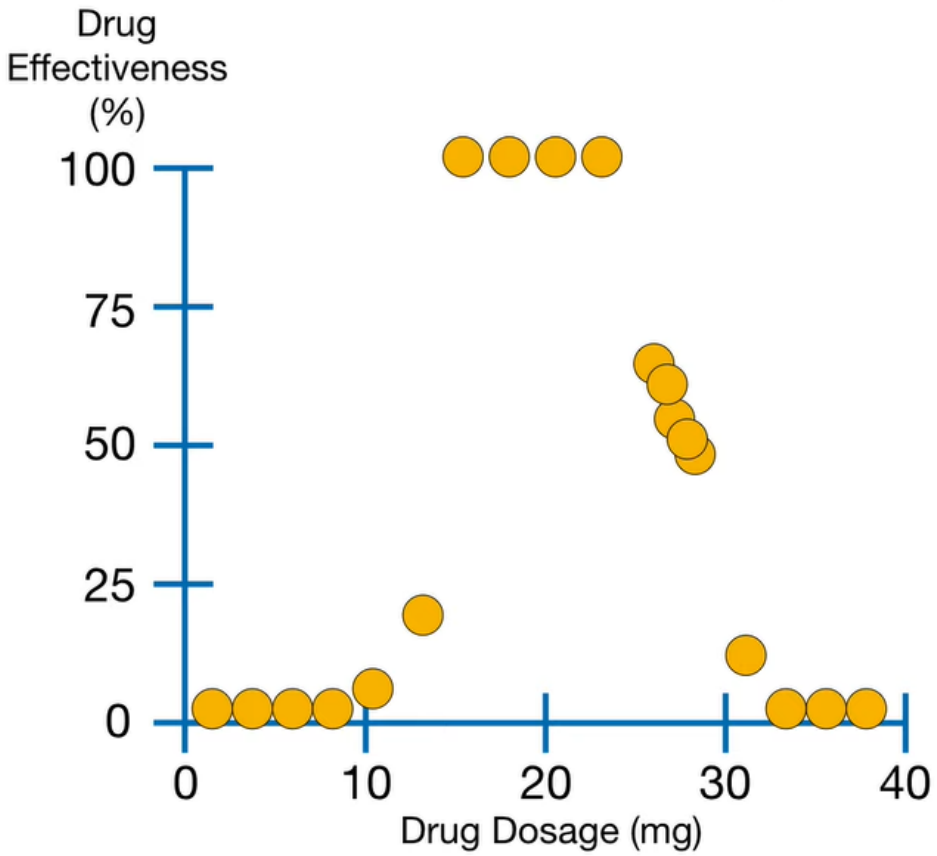

observations with Dosage < 14.5, 4.2%.
Drug Effectiveness (%)
100
75
50
25
0
0
10
20
30
40
Drug Dosage (mg)
Dosage < 14.5
4.2% Effective


…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


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

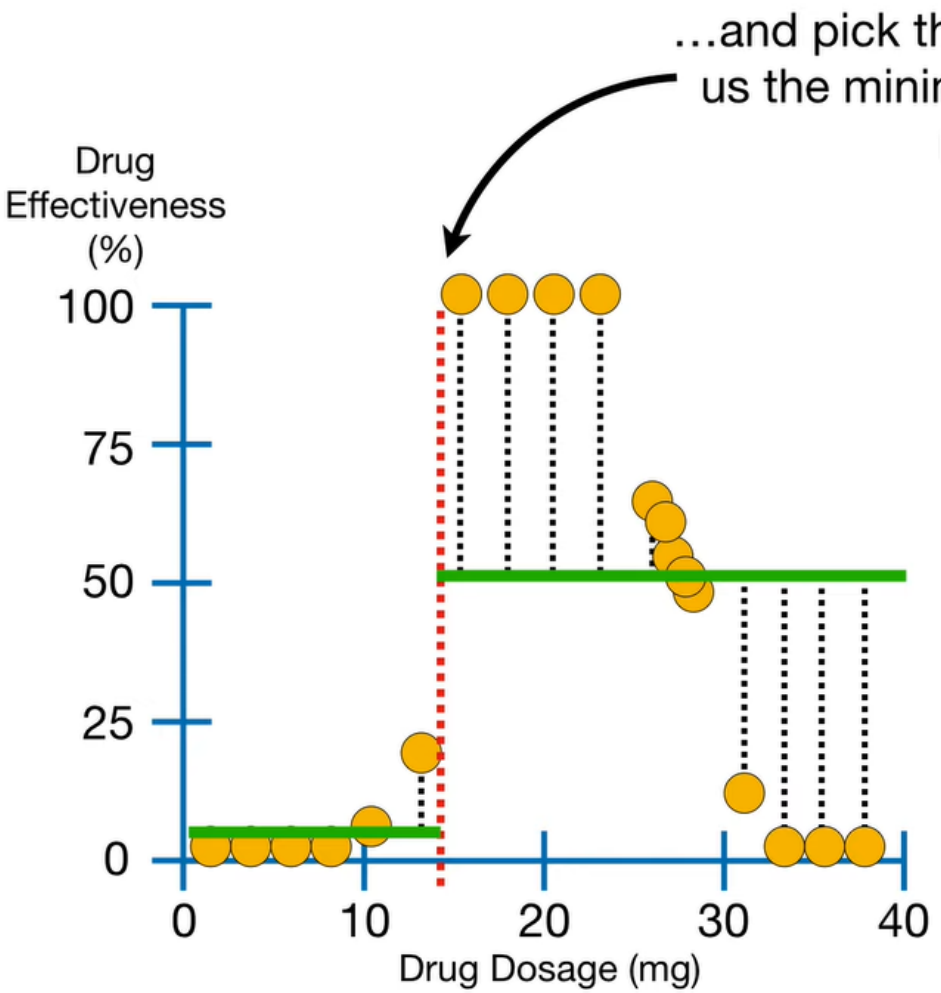
So far we have built a tree using a single predictor, **Dosage**, to predict **Drug Effectiveness**.



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

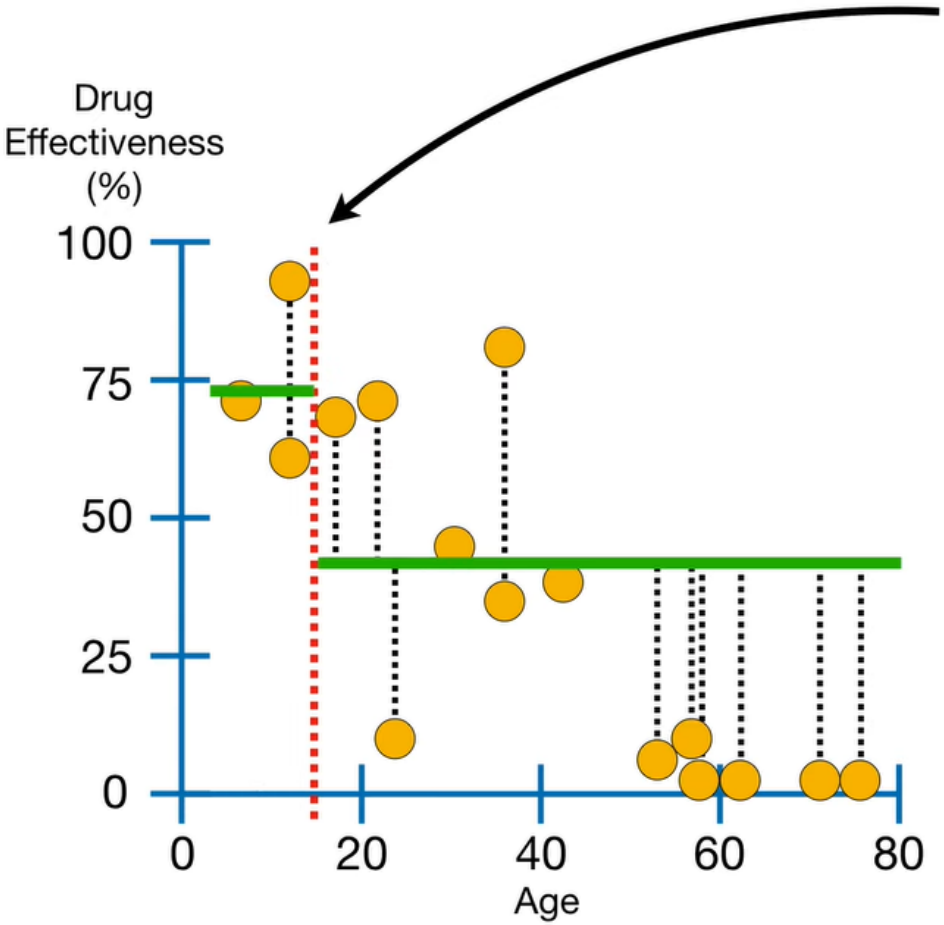
Now let's talk about how to build a tree to predict **Drug Effectiveness** using a bunch of predictors.

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…

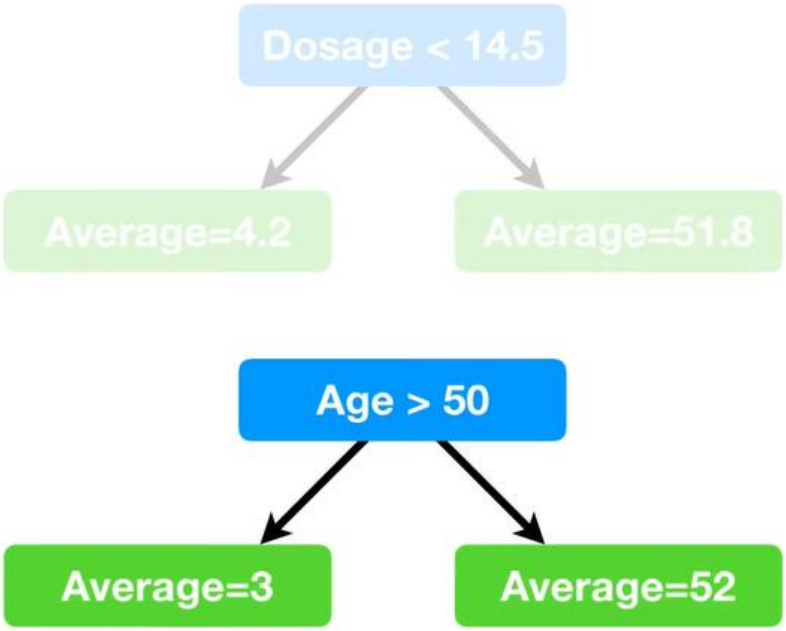


Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…

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



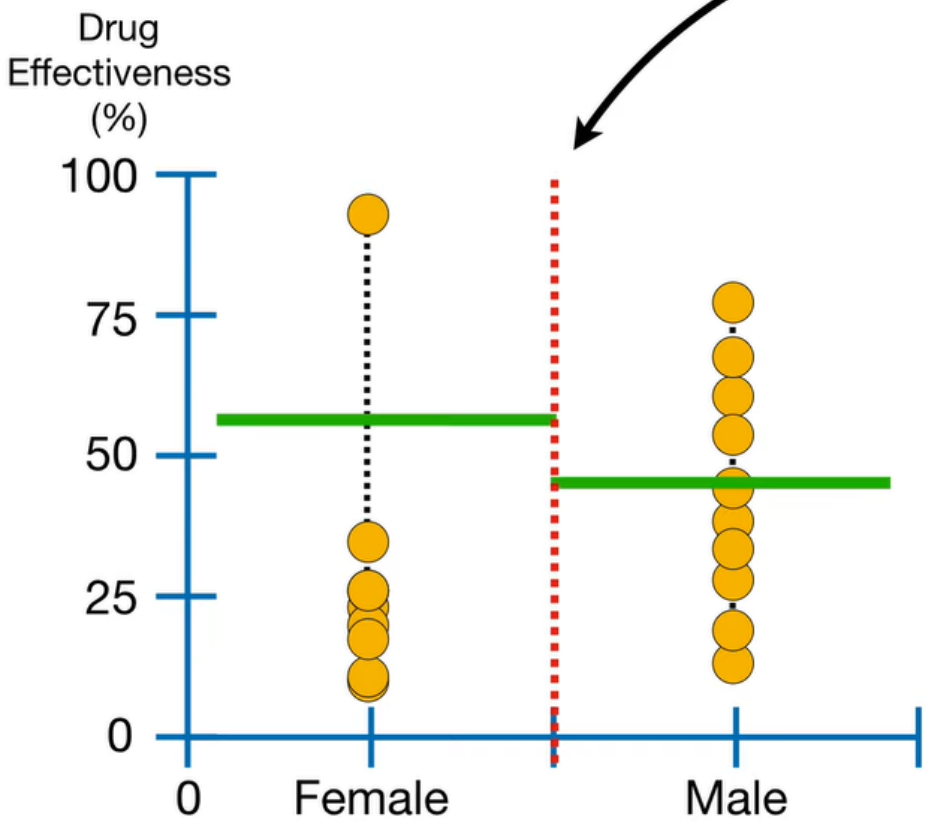
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…



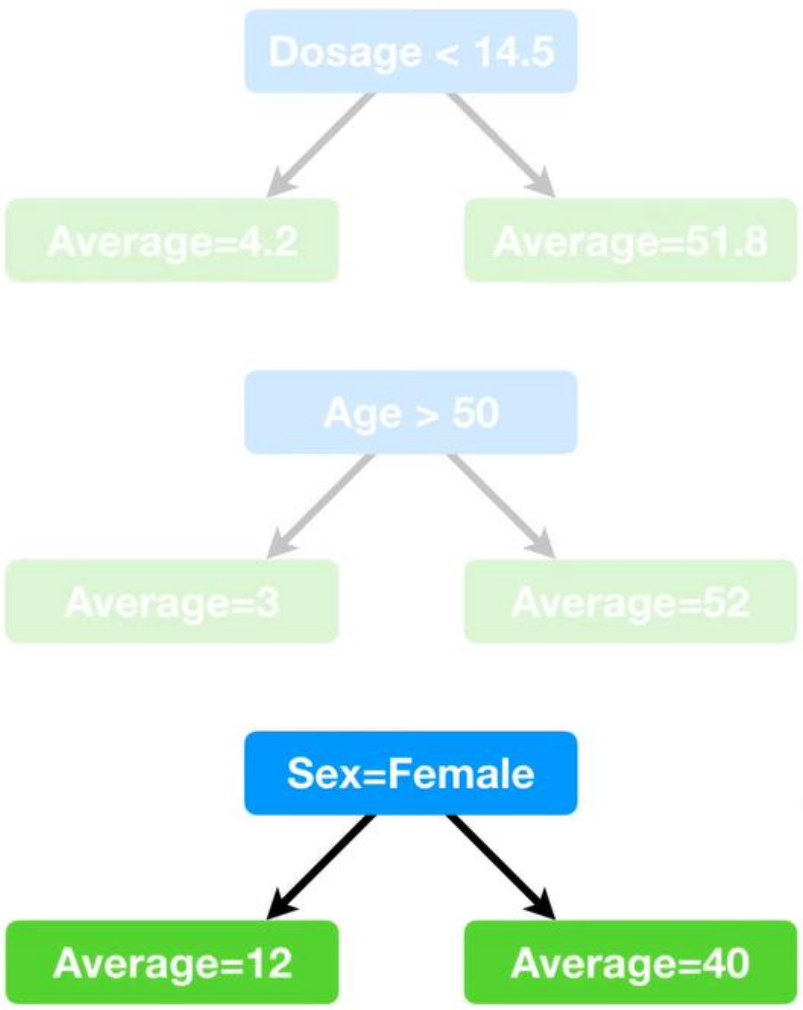
The best threshold becomes another *candidate* for the root.

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…

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

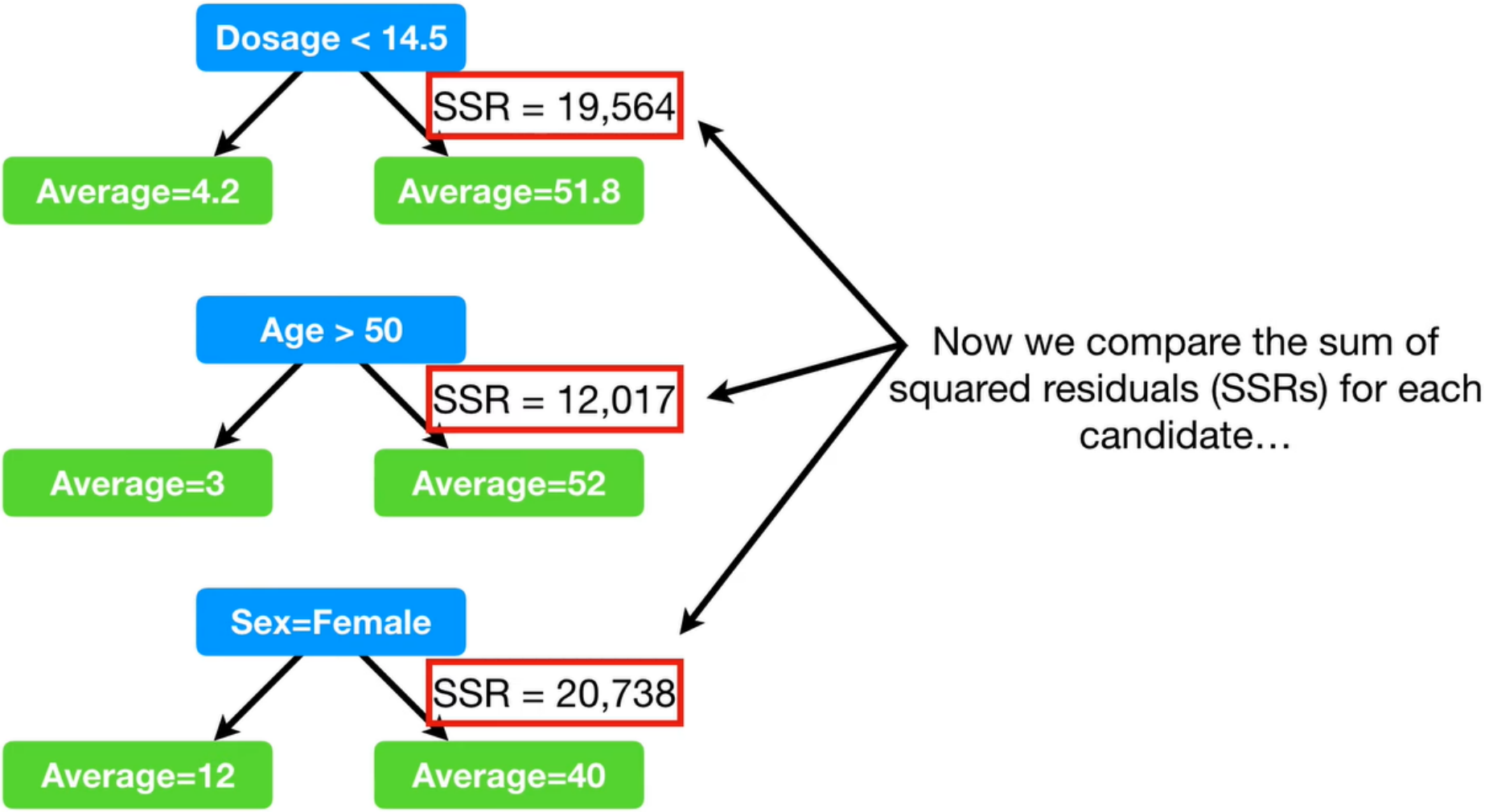


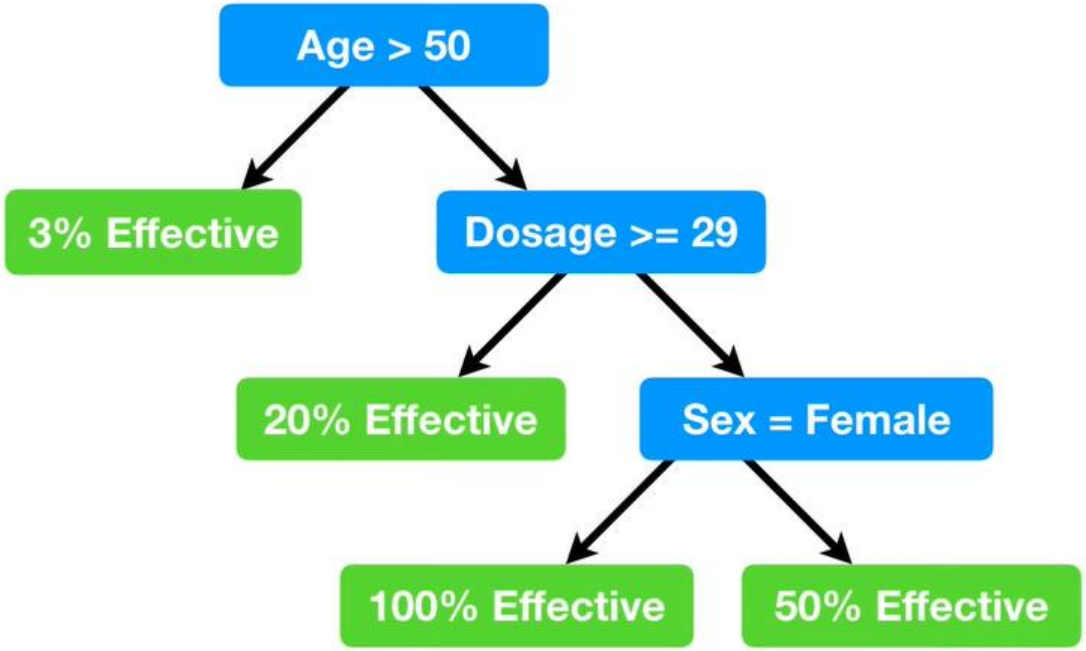
Dosage	Age	Sex	Drug Effect.
10	25	Female	98
20	73	Male	0
35	54	Female	6
5	12	Male	44
etc…	etc…	etc…	etc…



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

Python implementation (See python files)