Classification and Regression Trees… a bit more

Categorical and Numerical values in Classification trees

Categorical and Numerical values in Classification trees (1/4)

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

Sort rows

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

Categorical and Numerical values in Classification trees (2/4)

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

Compute average for all adjacent rows

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5 15 26.5 36.5 44 66.5

Categorical and Numerical values in Classification trees (3/4)

Age	Loves Cool As Ice
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

9.5
15
26.5
36.5
44
66.5

Compute the Information Gain for all possible binary options.

Consider the value with the highest gain as representative of the feature

Age < 9.5

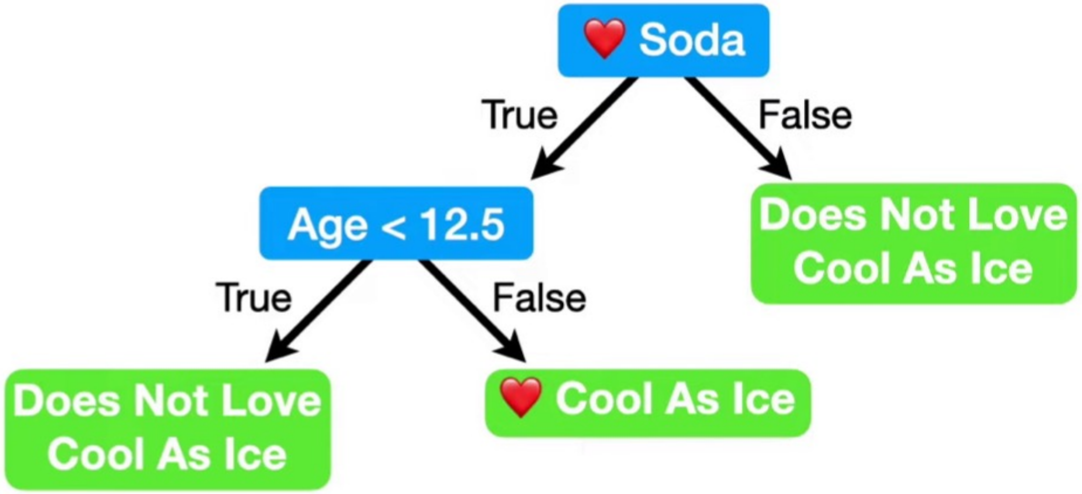
Age < 15

Categorical and Numerical values in Classification trees (4/4)

Loves Popcorn	Loves Soda	Age	Loves Cool As Ice
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

Compare the Information Gain for all the features and select the one with the highest value.

Continue until you build the whole tree.



Missing values

Missing categorical values

Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...

Missing categorical values

Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
Yes	Yes	No	No
Yes	No	**YES**	Yes
etc...	etc...	etc...	etc...

Add the most frequent value

Missing categorical values

Find a correlated feature and use it as guideline

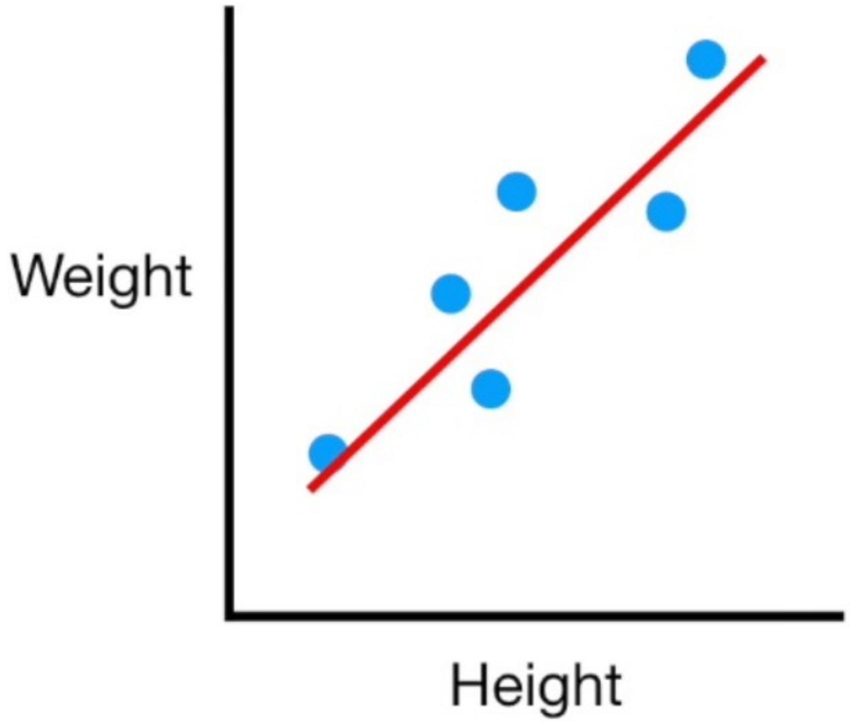
Chest Pain	Good Blood Circulation	Blocked Arteries	Heart Disease
No	No	No	No
Yes	Yes	Yes	Yes
No	Yes	No	No
Yes	No	???	Yes
etc...	etc...	etc...	etc...

Missing continuous values

Height	Good Blood Circulation	Weight	Heart Disease
5'7"	No	155	No
6'	Yes	180	Yes
5'4"	Yes	120	No
5'8"	No	???	Yes
etc...	etc...	etc...	etc...

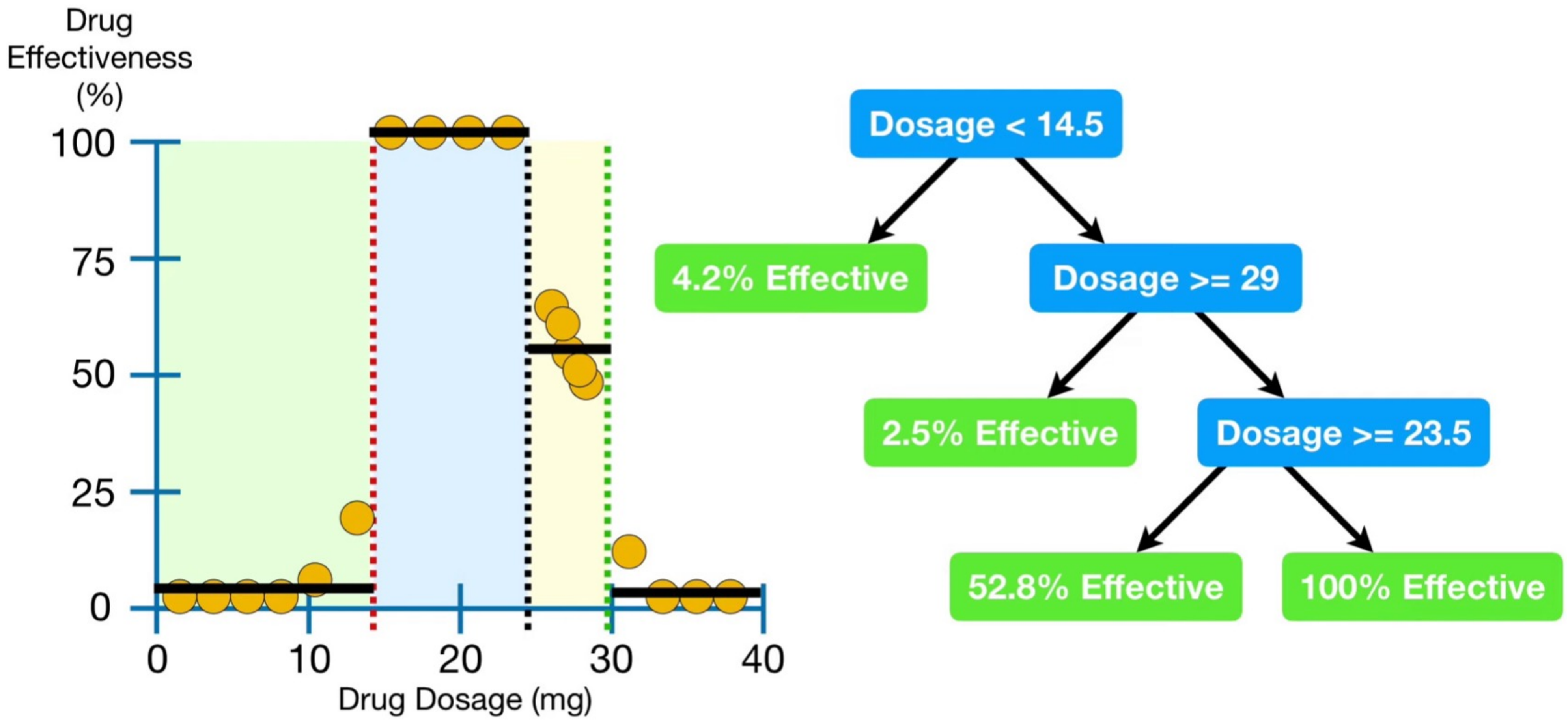
Missing continuos values

Height	Good Blood Circulation	Weight	Heart Disease
5’7”	No	155	No
6’	Yes	180	Yes
5’4”	Yes	120	No
5’8”	No	???	Yes
etc…	etc…	etc…	etc…

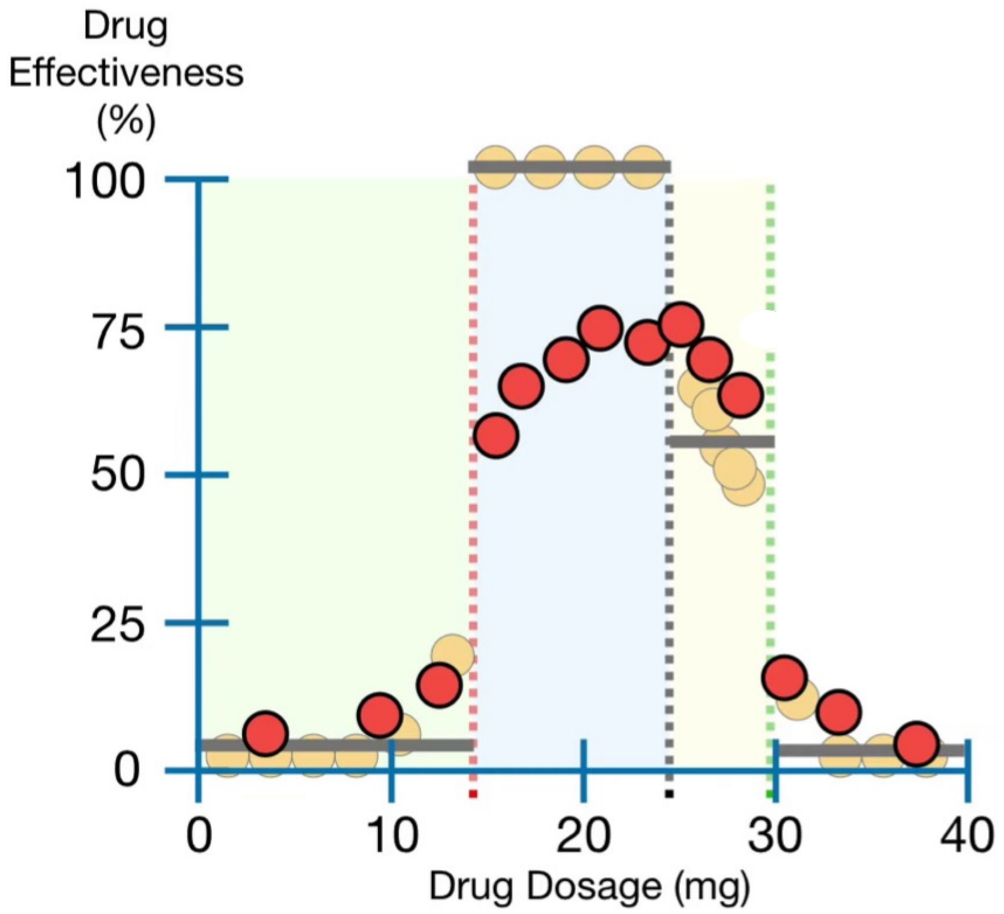


Pruning Regression Trees

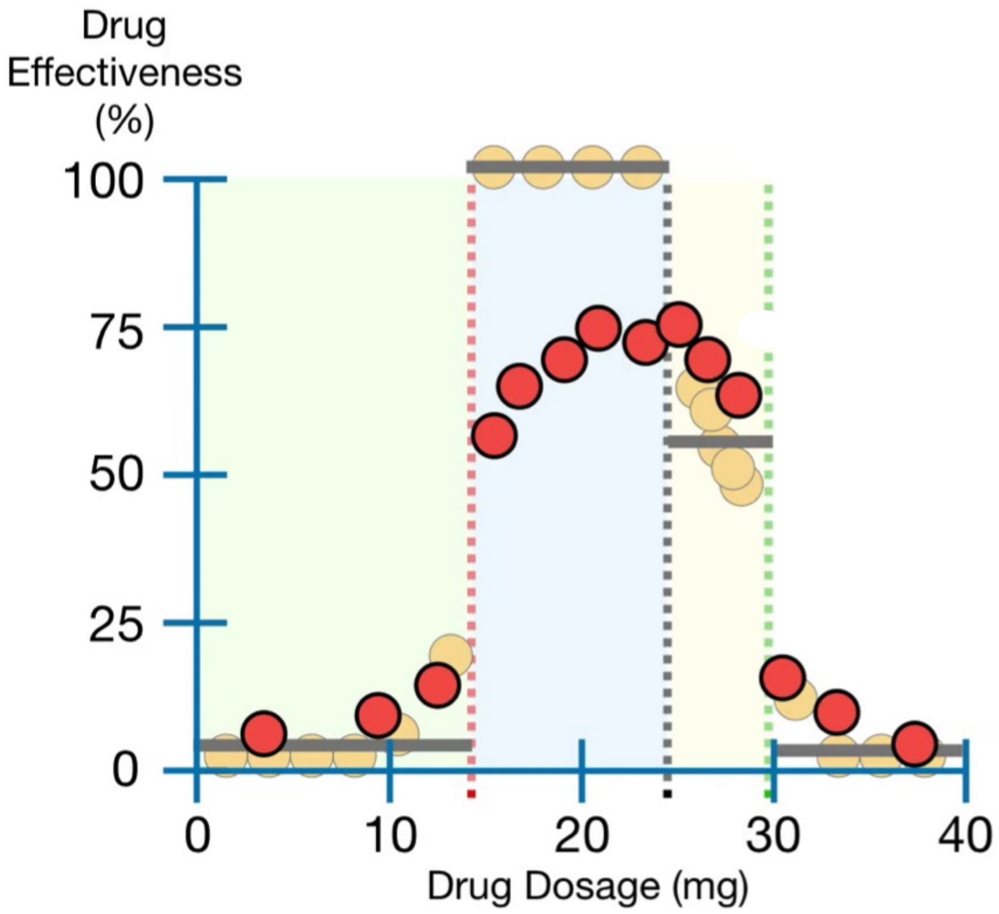
Regression Tree

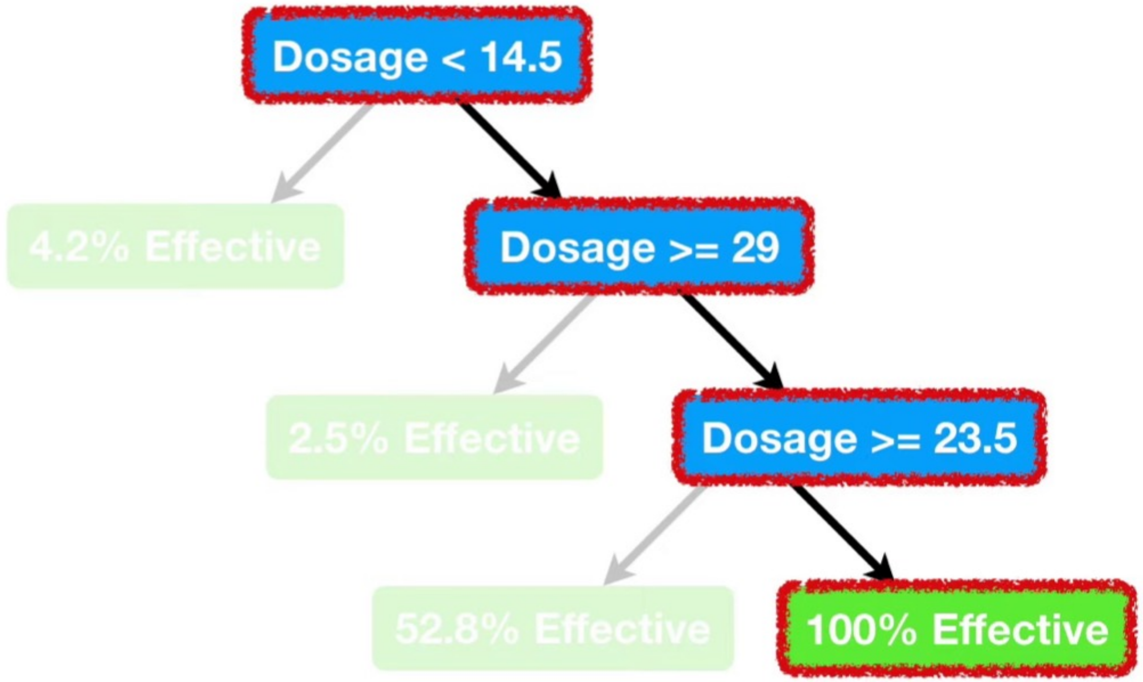


Residual Sum of Squares on the Test Set

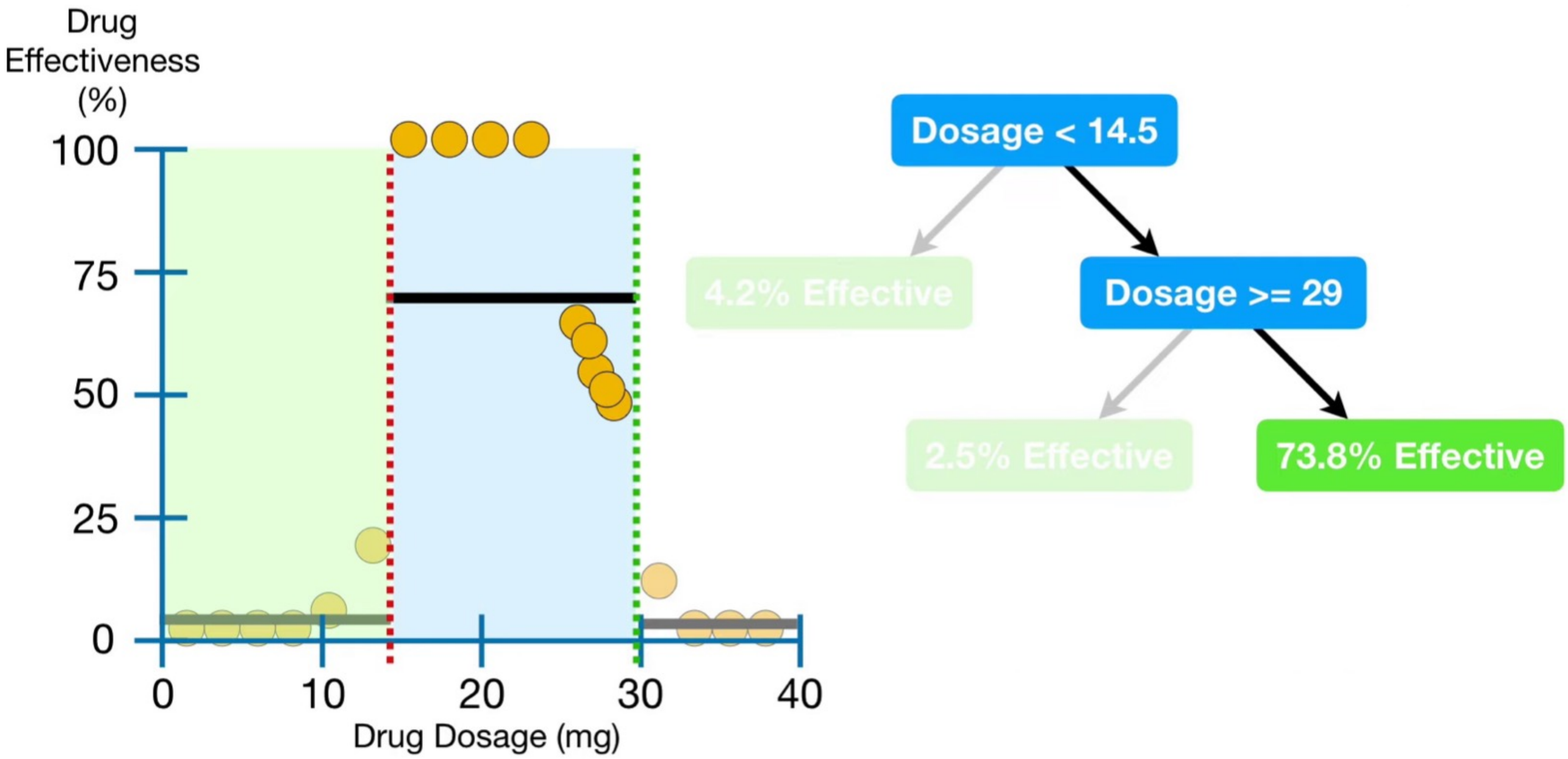


Sum of Squared Residuals on the Test Set

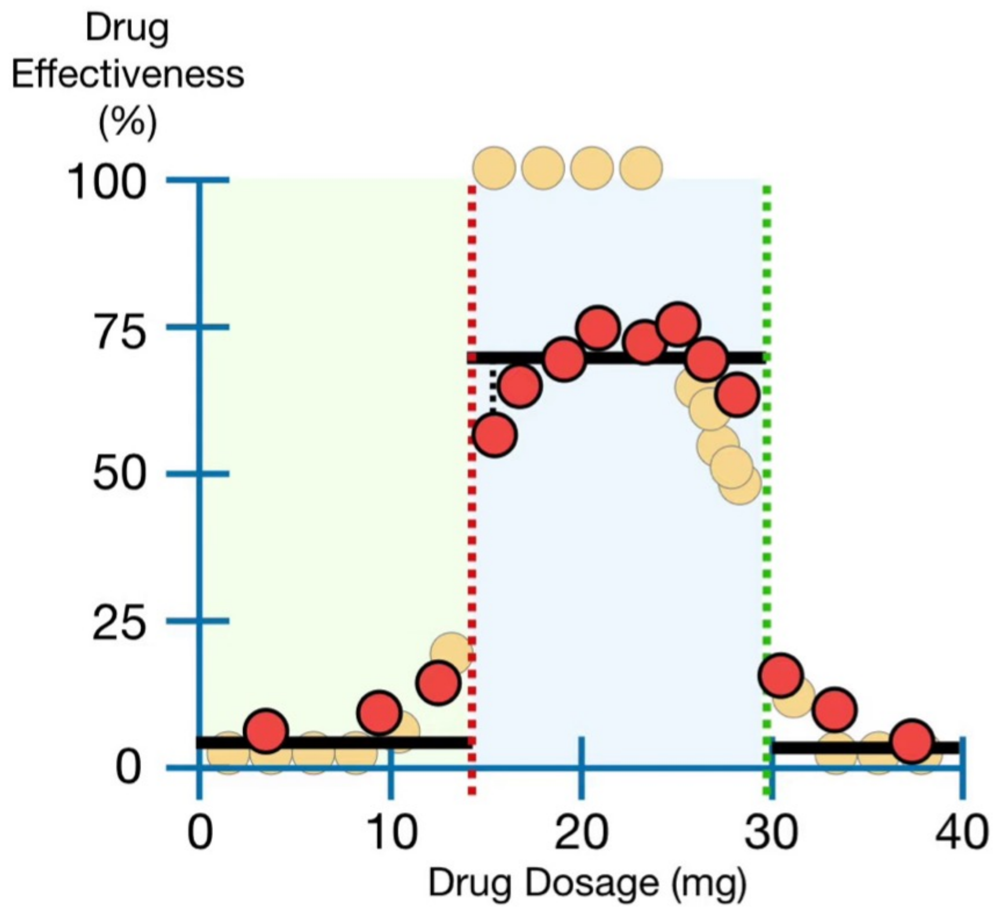




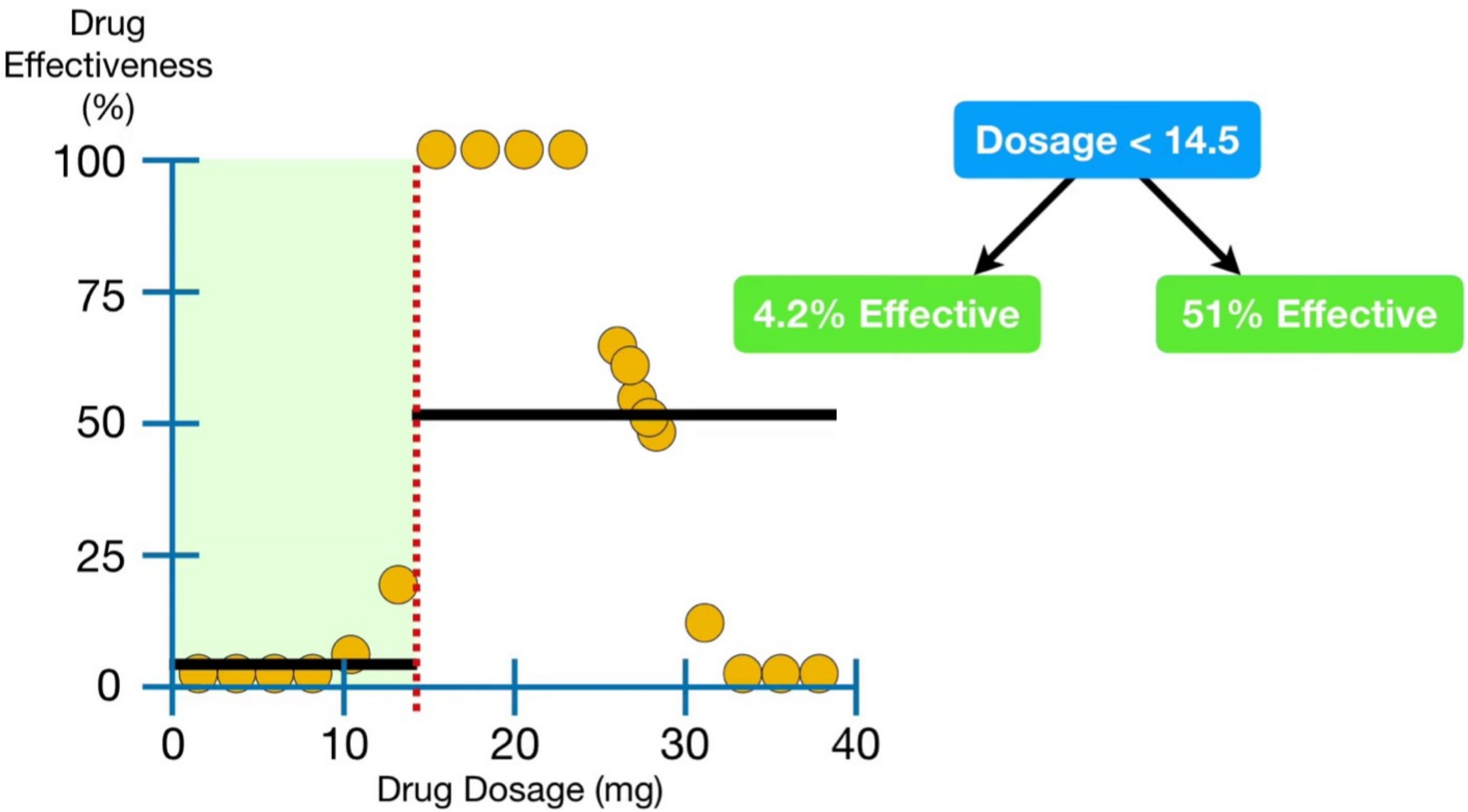
Reduce the precision of the tree



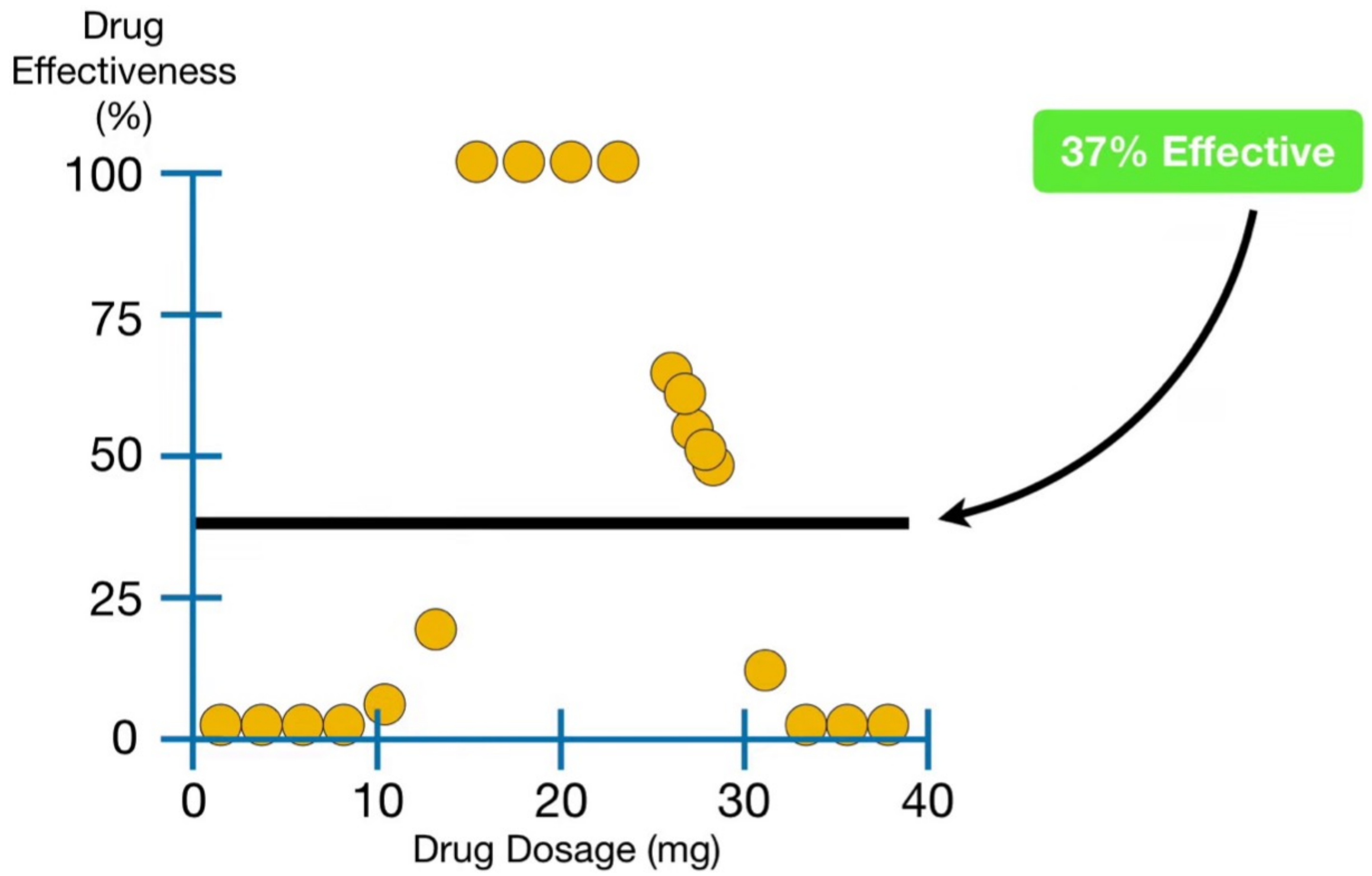
Reduce overfitting



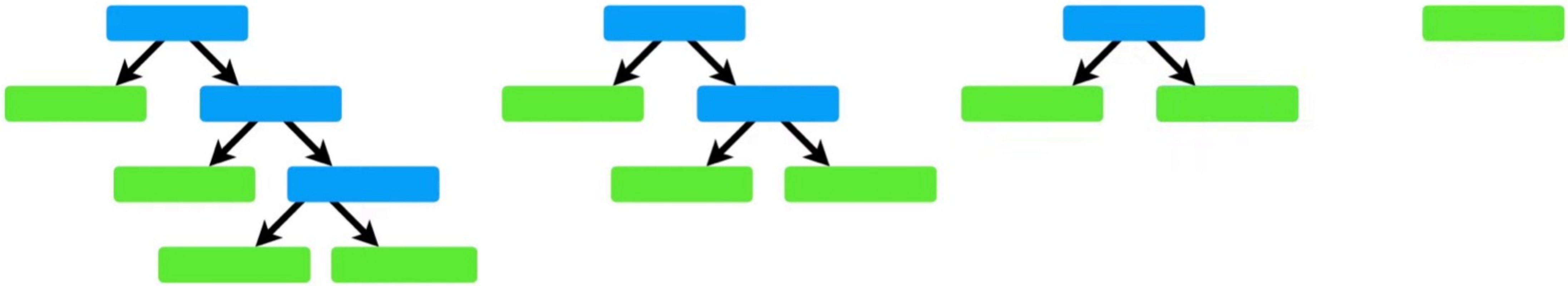
Further pruning



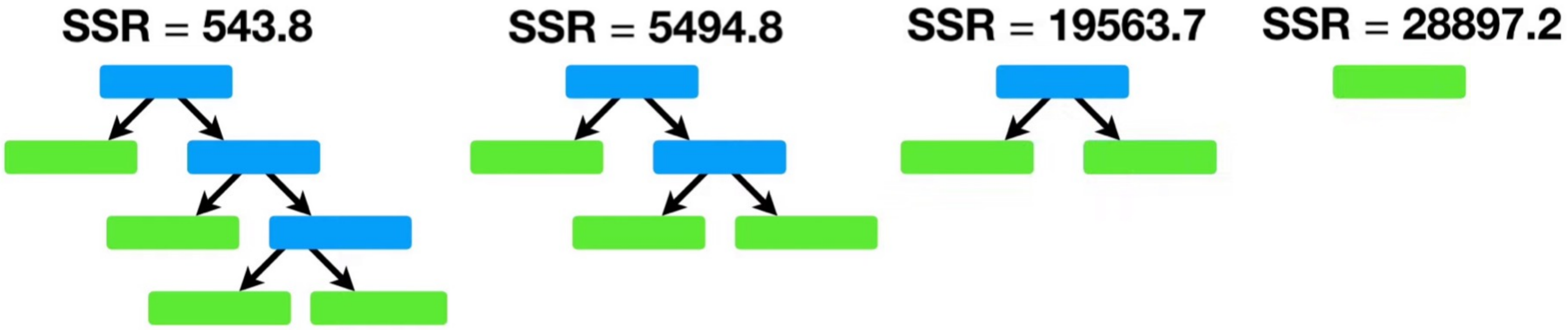
Further pruning



What is the best tree?



Compute the SSR for each tree



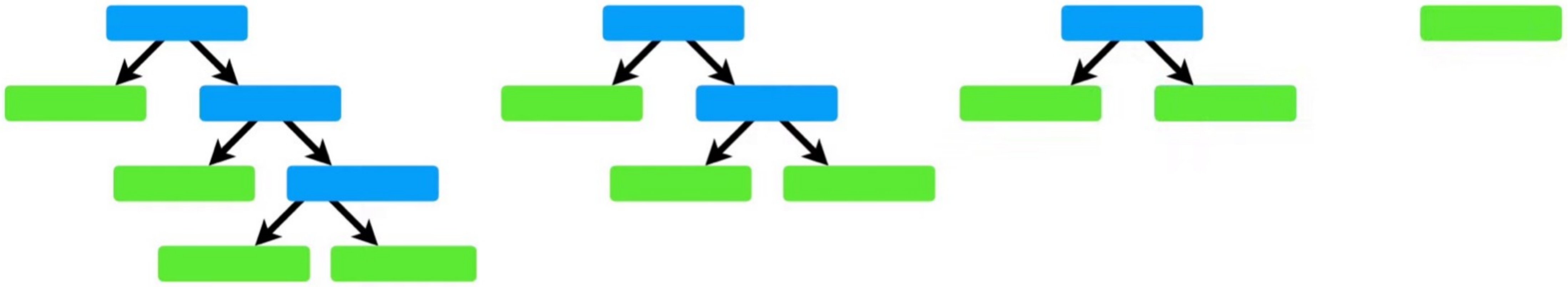
Weakest Link Pruning

introducing a penalty proportionally of the depth of the tree

Tree Score = SSR + Tree Complexity Penalty

$$= \text{SSR} + \alpha \cdot T$$

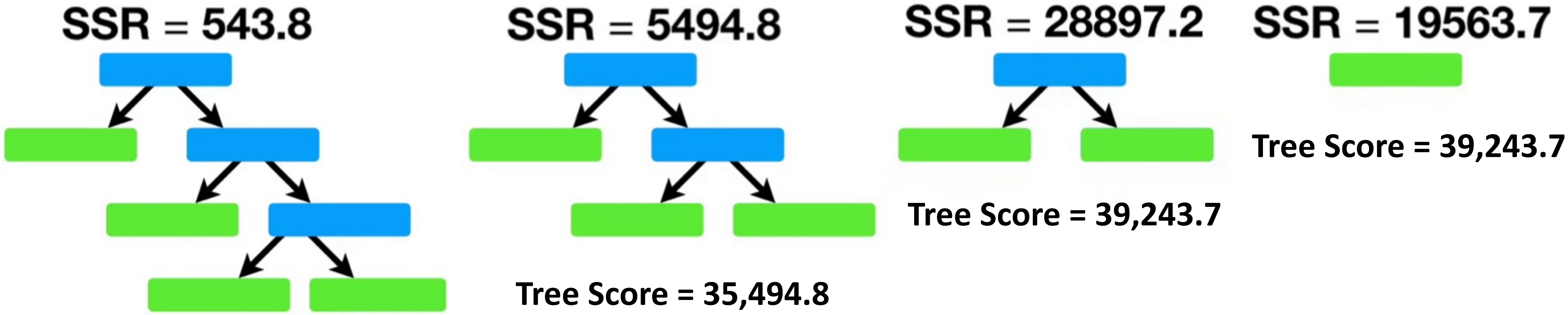
alpha is a higher parameter that needs to be tuned



Weakest Link Pruning

Tree Score = SSR + $\alpha \cdot T$
Example with $\alpha = 10.000$

the second seems to be the best. because it balance the SSR and penalty



Tree Score = 39,243.7

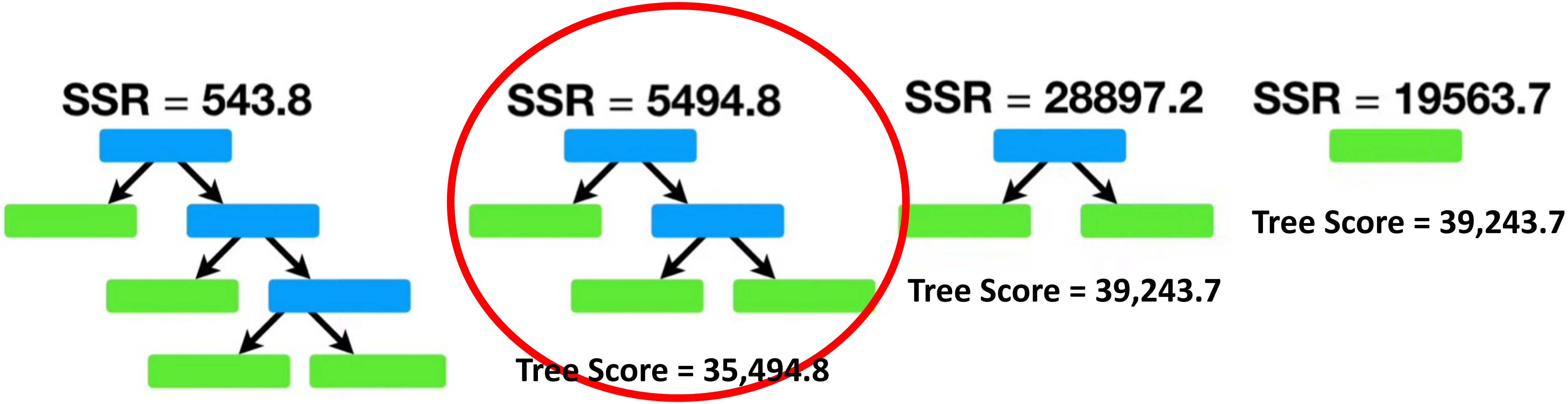
Tree Score = 39,243.7

Tree Score = 35,494.8

Tree Score = 40,543.8

Weakest Link Pruning

Tree Score = SSR + $\alpha \cdot T$
Example with $\alpha = 10.000$



Tree Score = 40,543.8

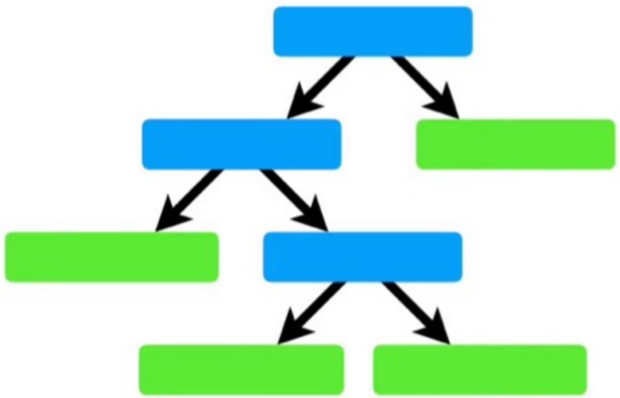
Tree Score = 35,494.8

Tree Score = 39,243.7

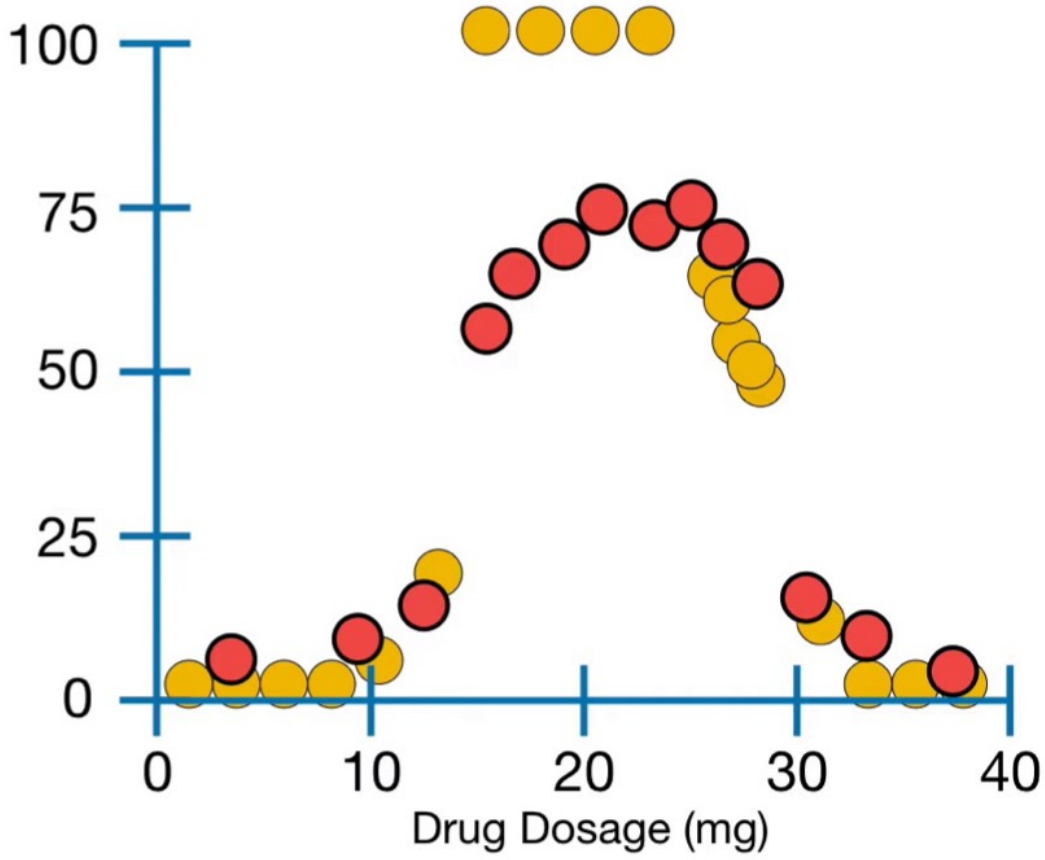
Tree Score = 39,243.7

How to evaluate the best α

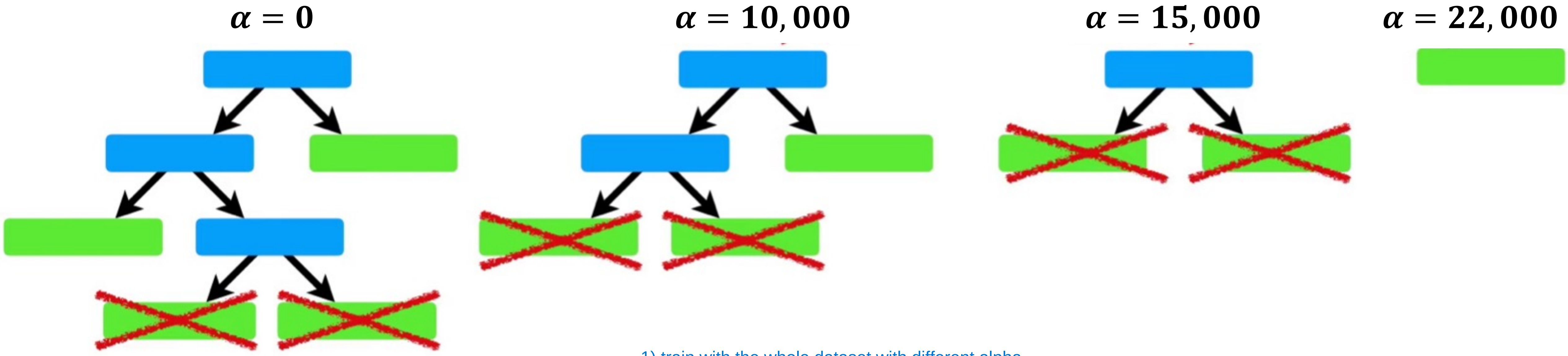
Build a tree considering all the data



we use all the data()



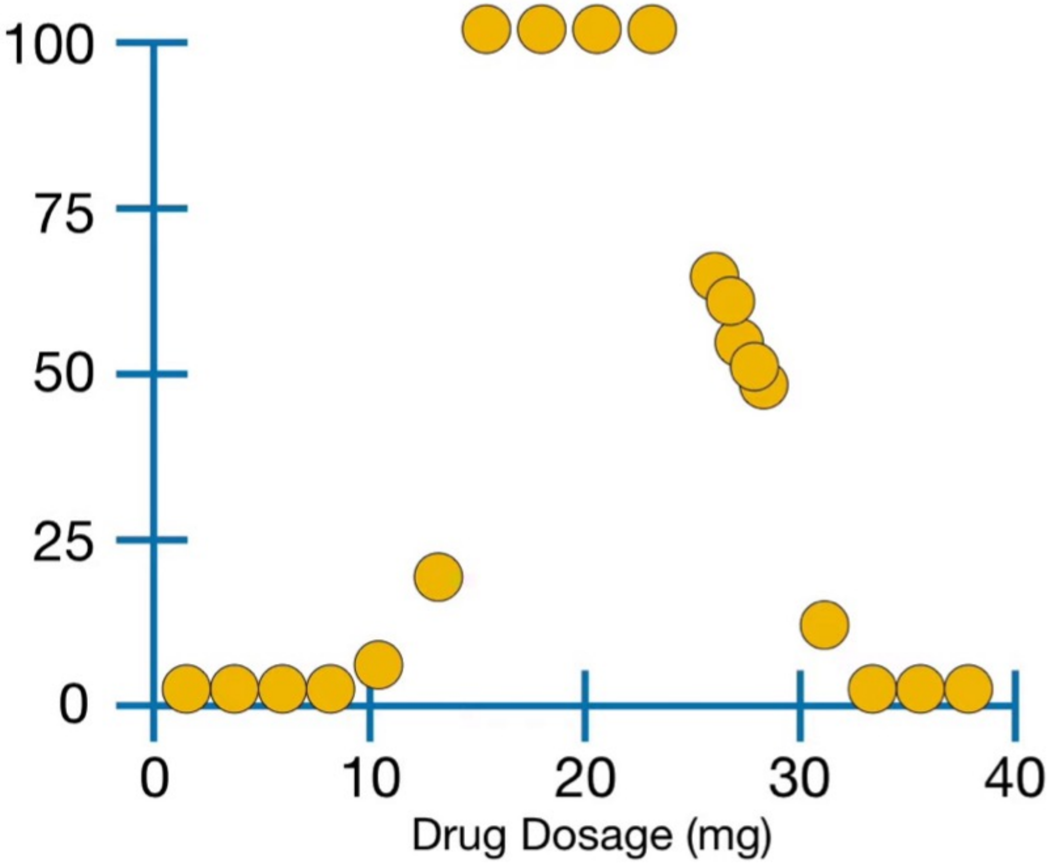
How to evaluate the best α

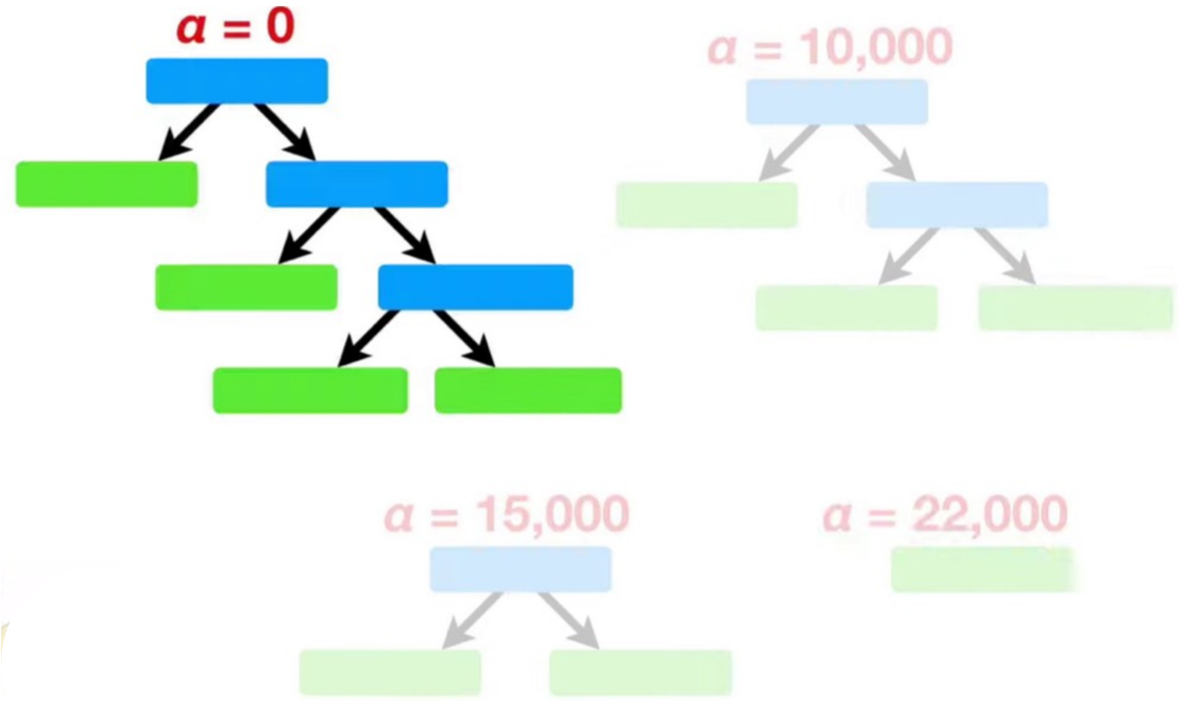


1) train with the whole dataset with different alpha
2)then i considered only the training set
3) i calculate the SSR with the test set
4)choose the tree with the lowest SSR in the TEST set

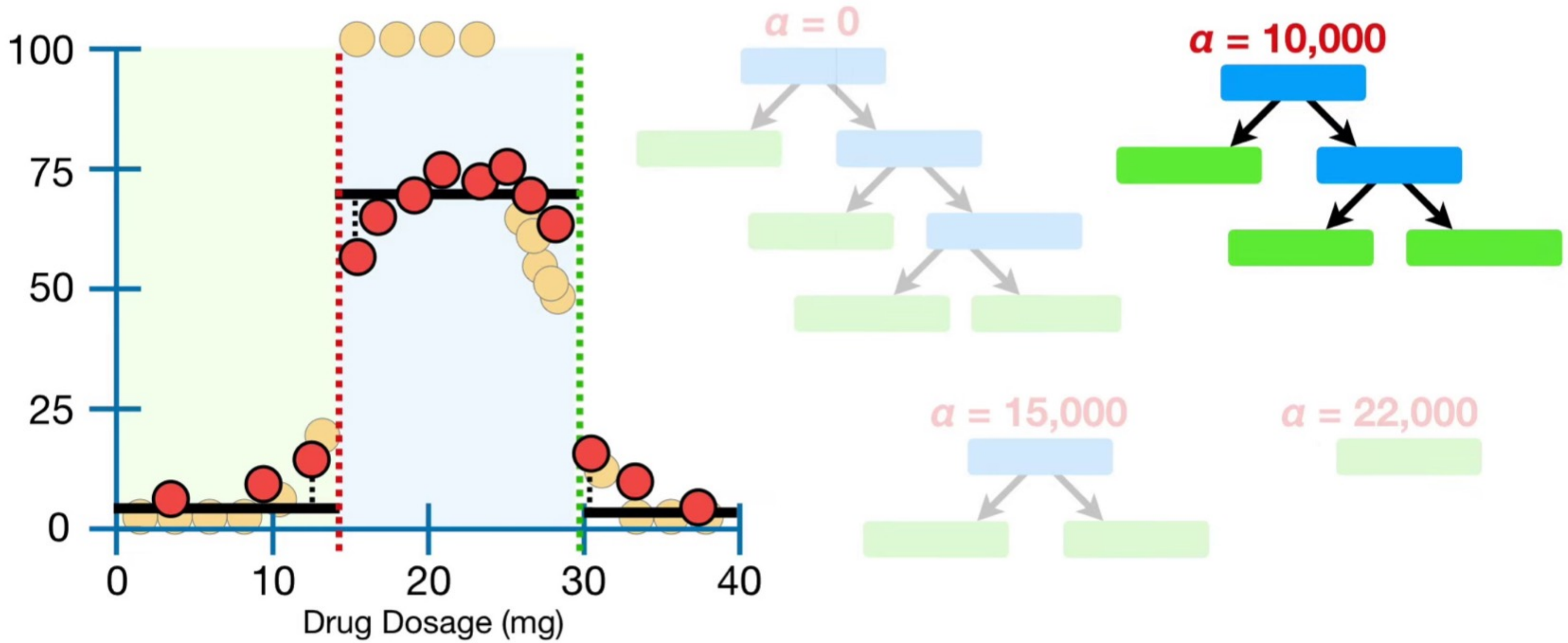
Train the tree again using the Training set only

i train the different tree, and when we train this new tree we compute the error in the





Compute SSR on the Test set for all the trees



<u>Vote for the one with the lowest SSR</u>

Repeat the process with new Training and Test sets --> K-fld cross-validation