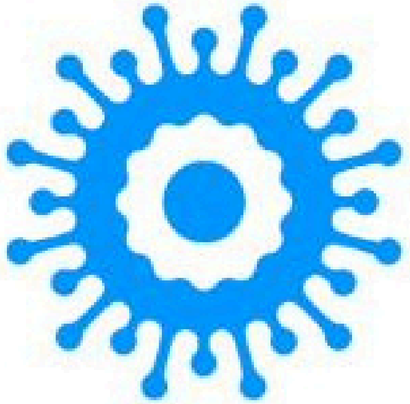
Tree Based Regressions

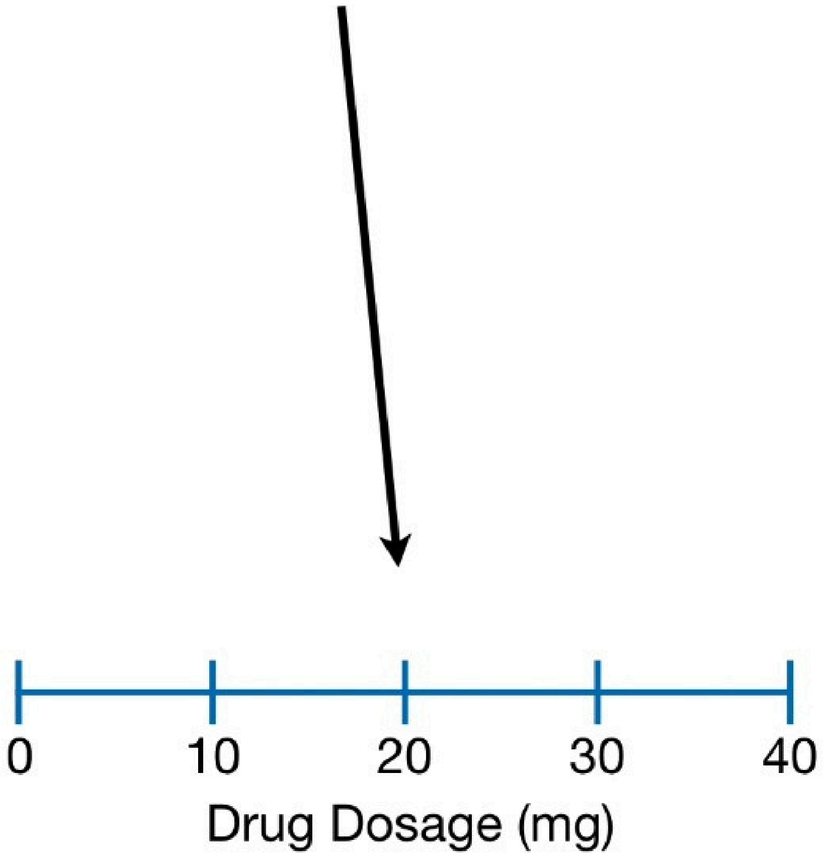
Decision Trees & Random Forest Regression

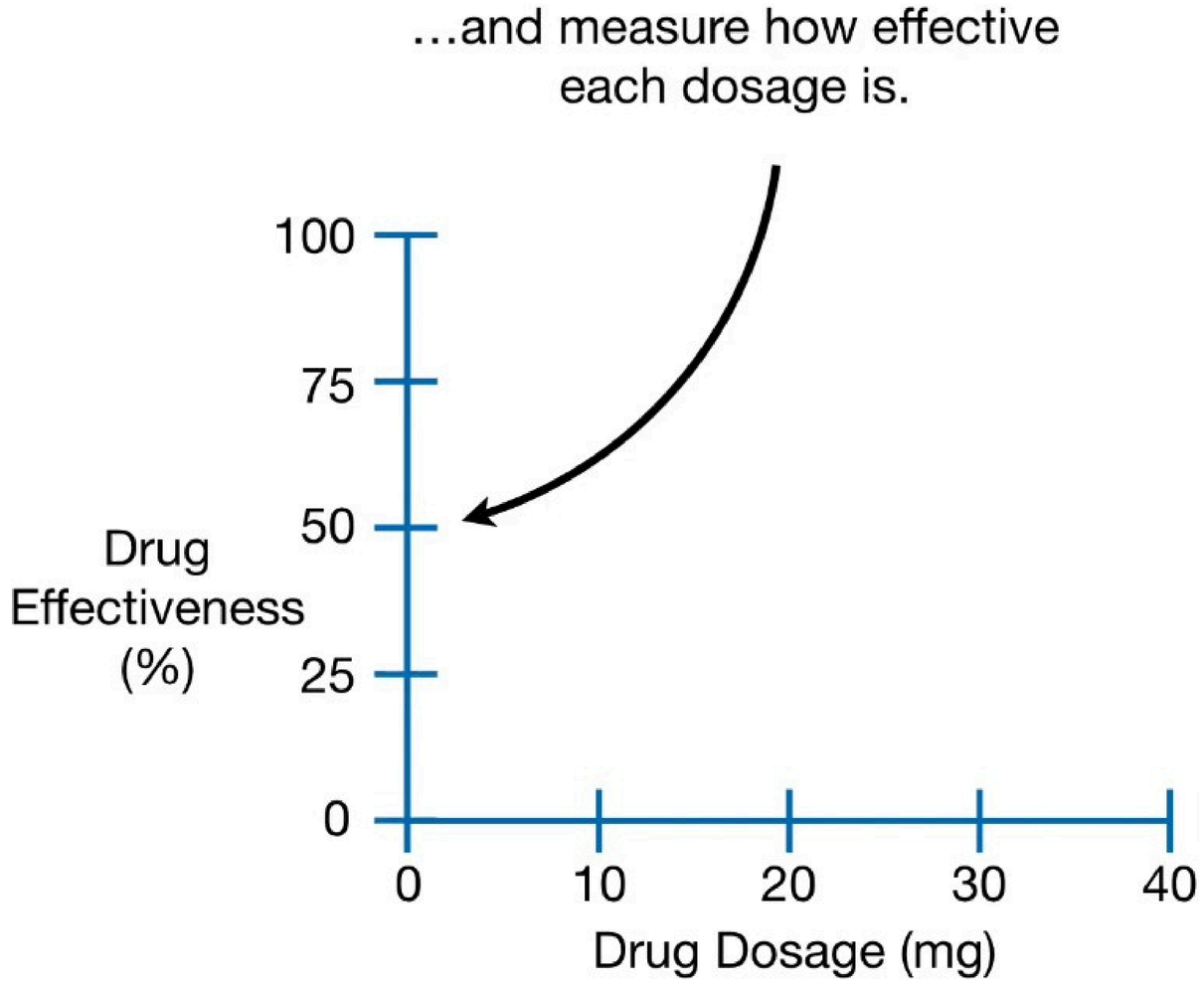
However, we don’t know the optimal dosage to give to patients.

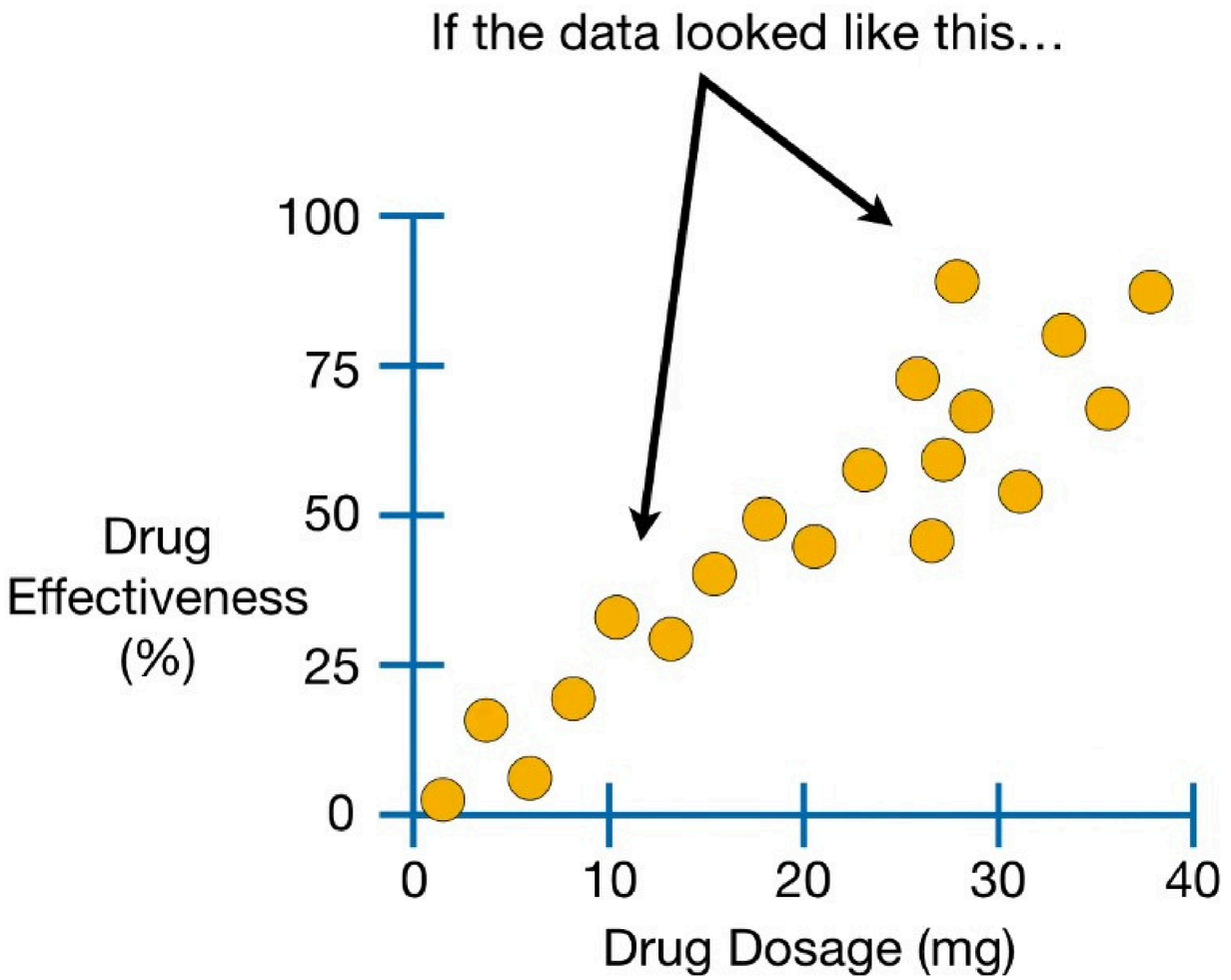
vs.

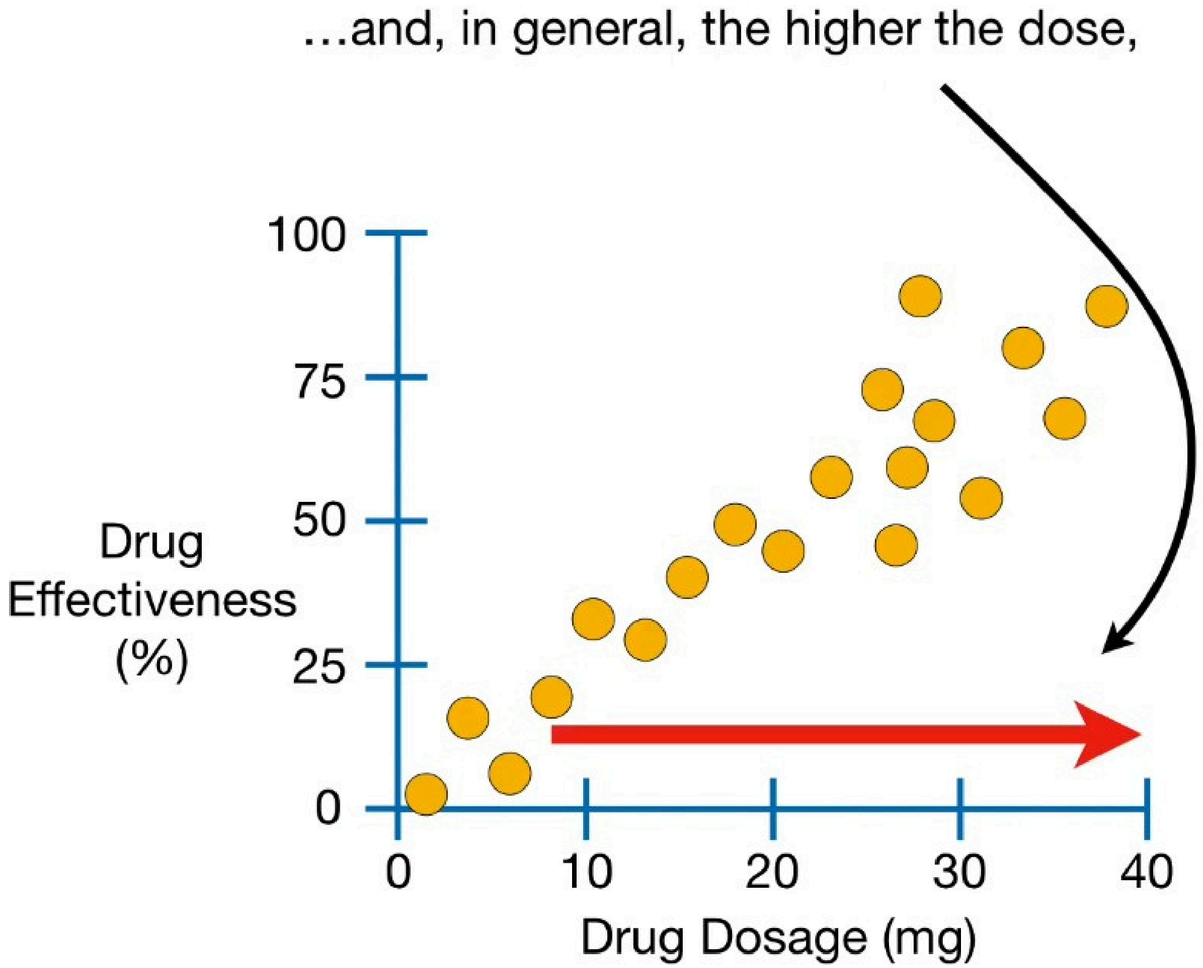


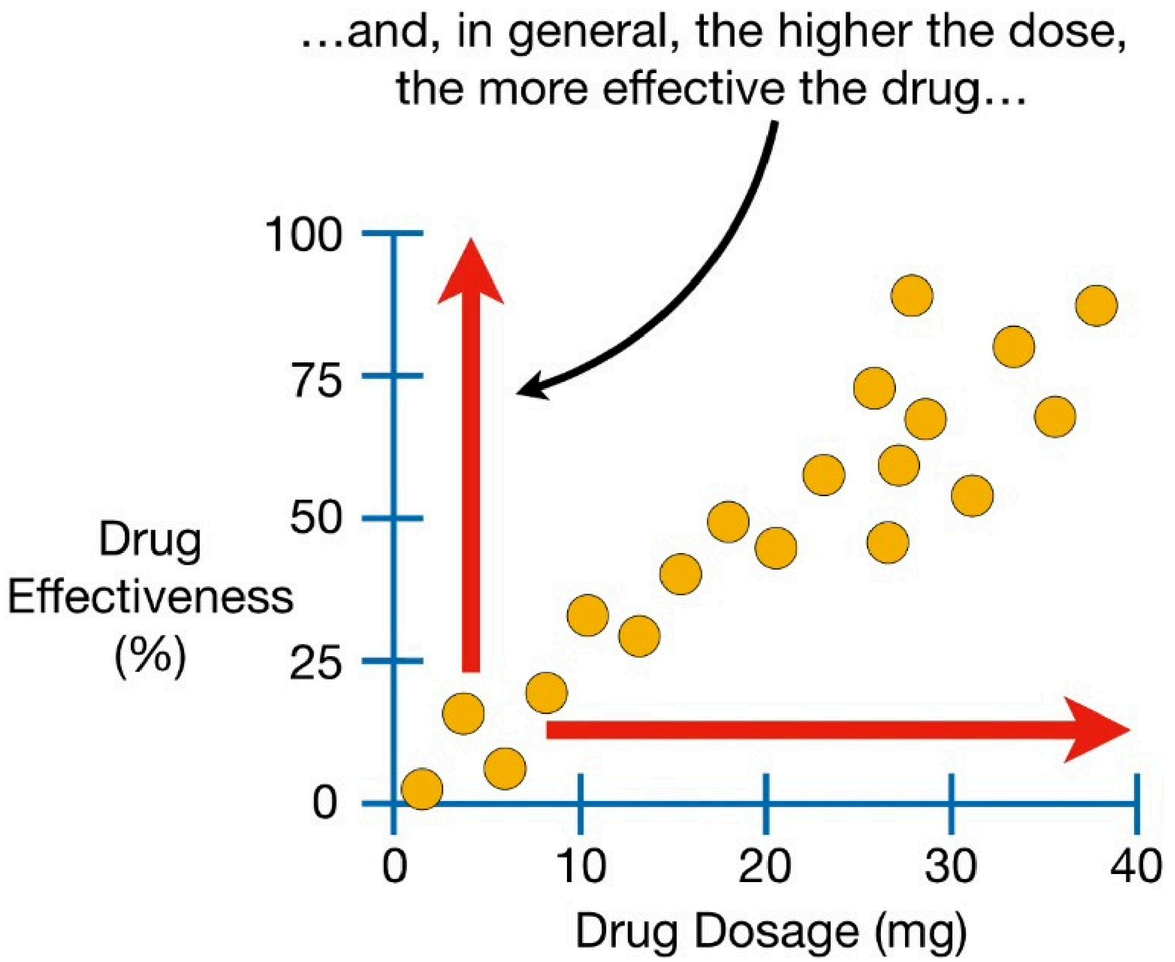
So we do a clinical trial with different dosages…



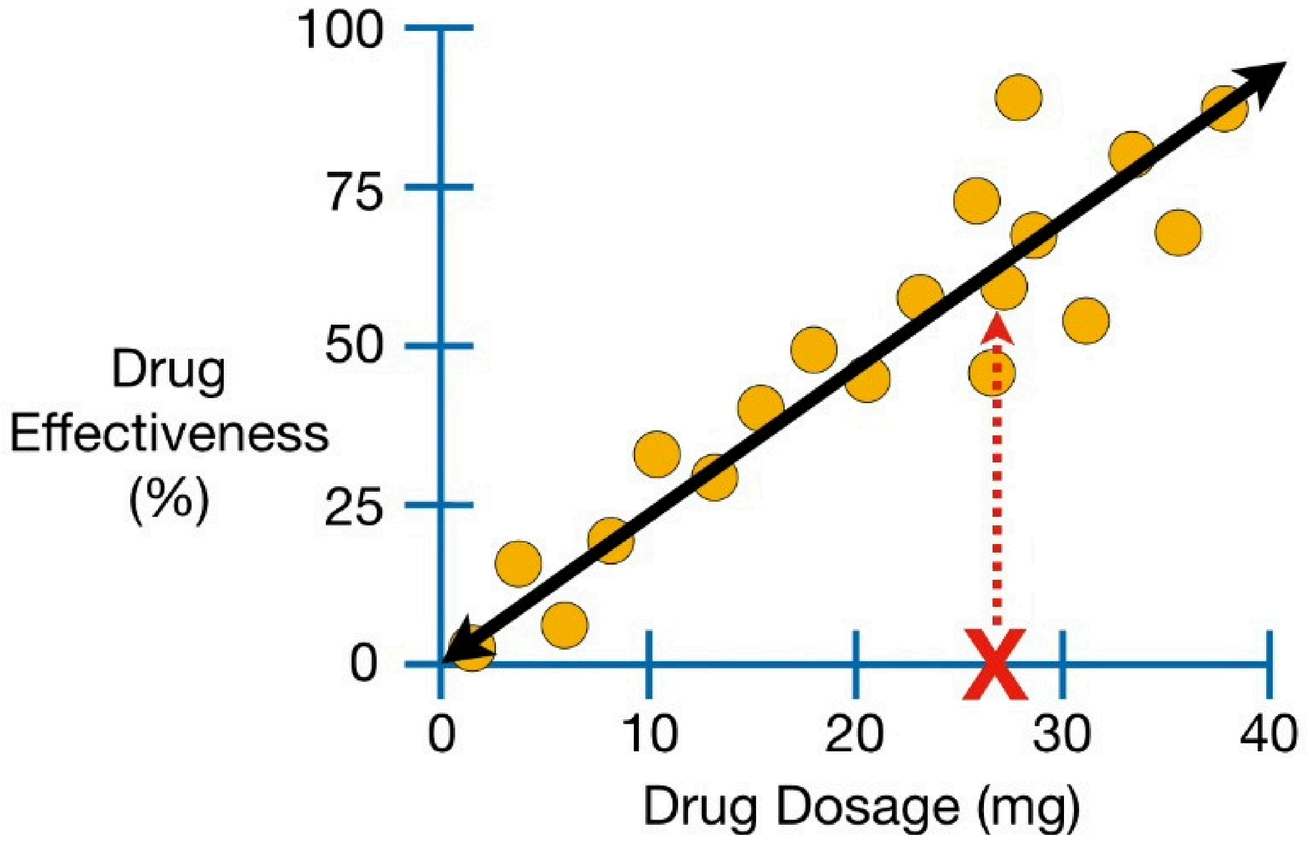

…and measure how effective each dosage is.
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)


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

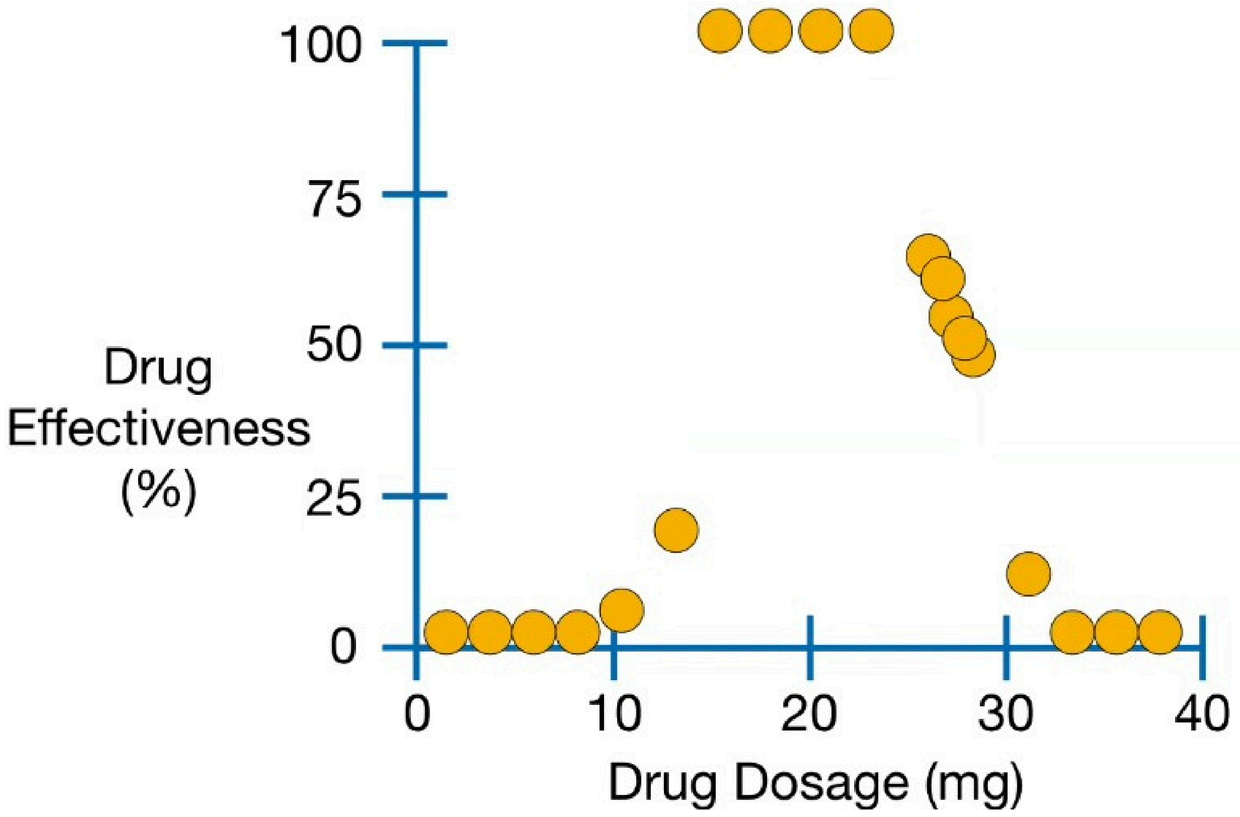

…and, in general, the higher the dose,
100
75
50
25
0
Drug
Effectiveness
(%)
0
10
20
30
40
Drug Dosage (mg)

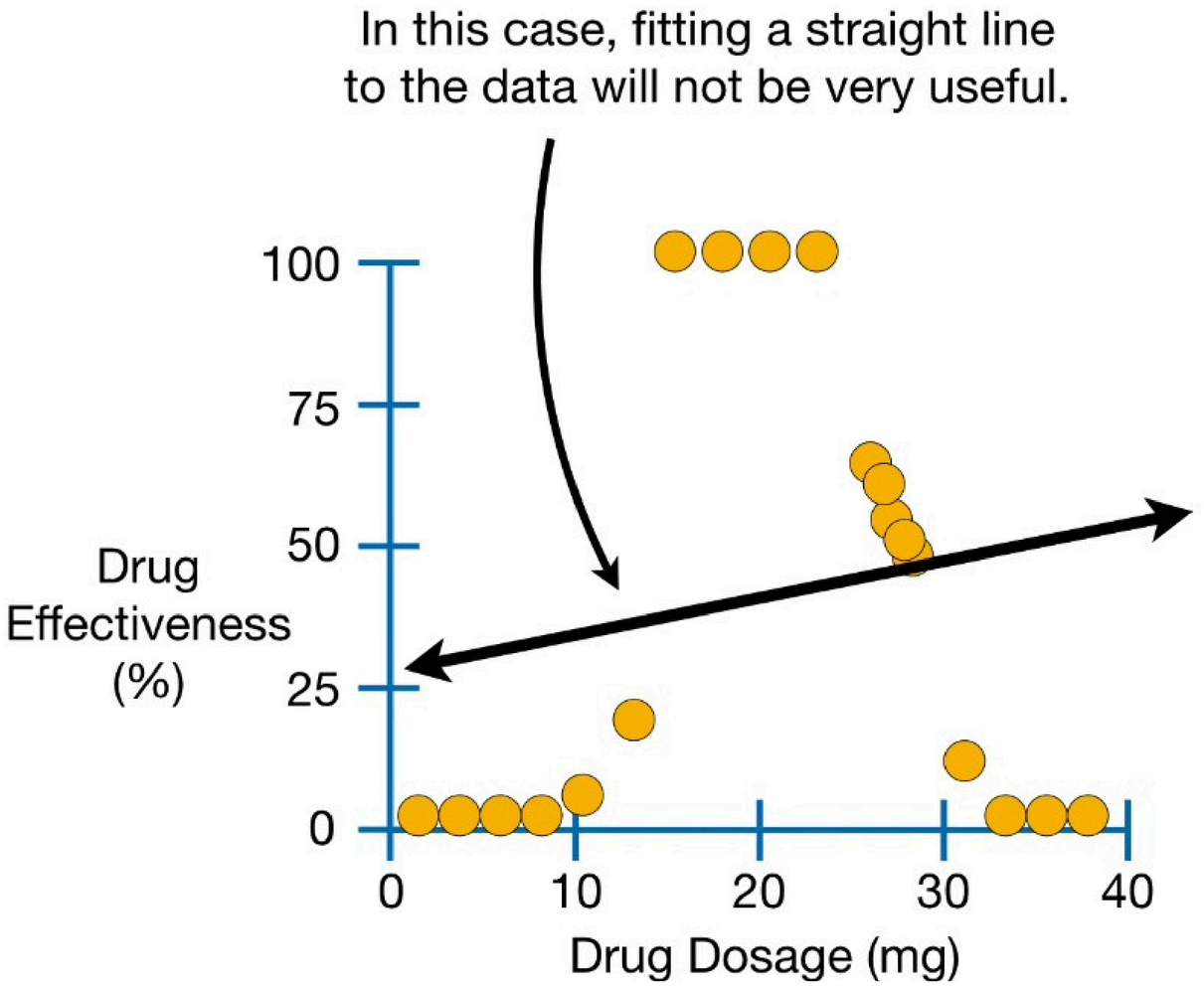

…and, in general, the higher the dose,
the more effective the drug…
100
75
50
25
0
Drug
Effectiveness
(%)
0
10
20
30
40
Drug Dosage (mg)

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

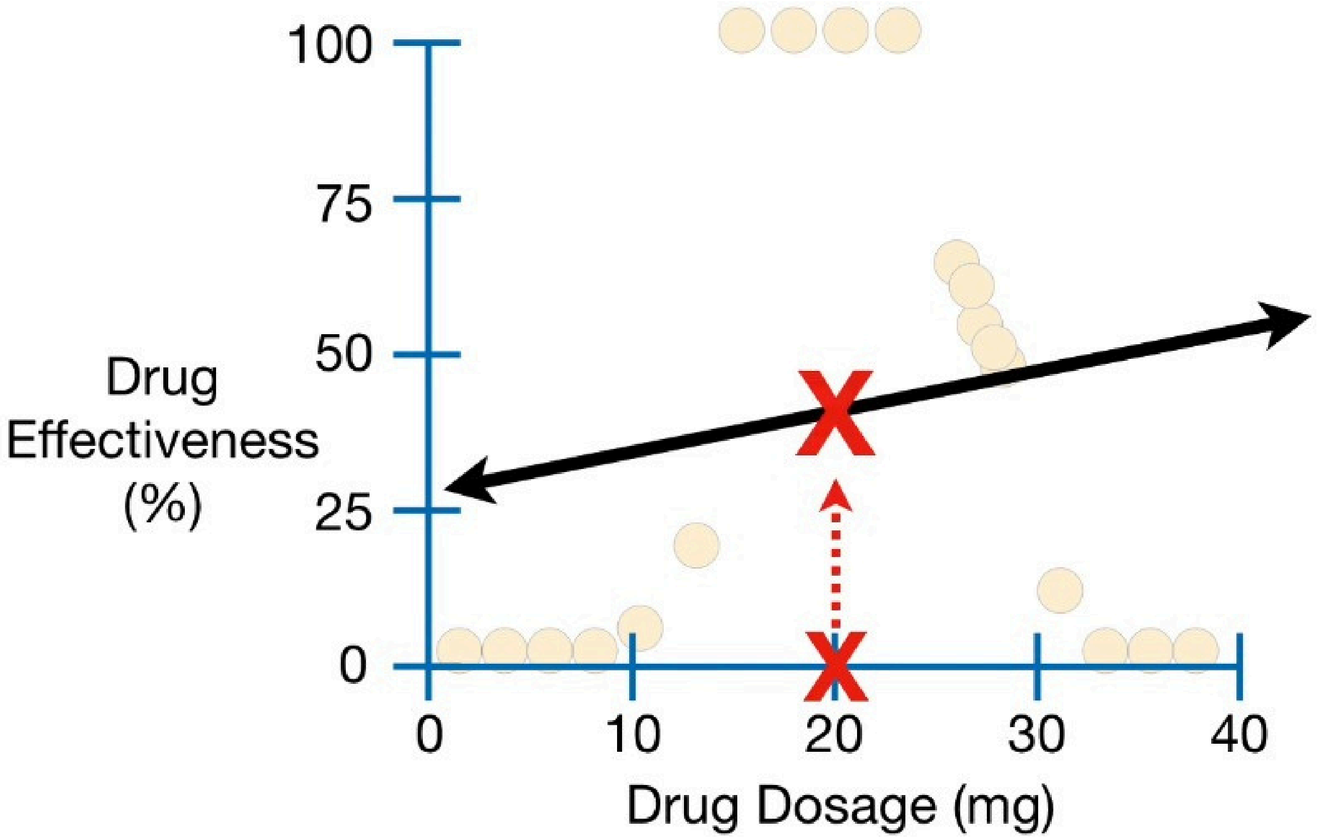


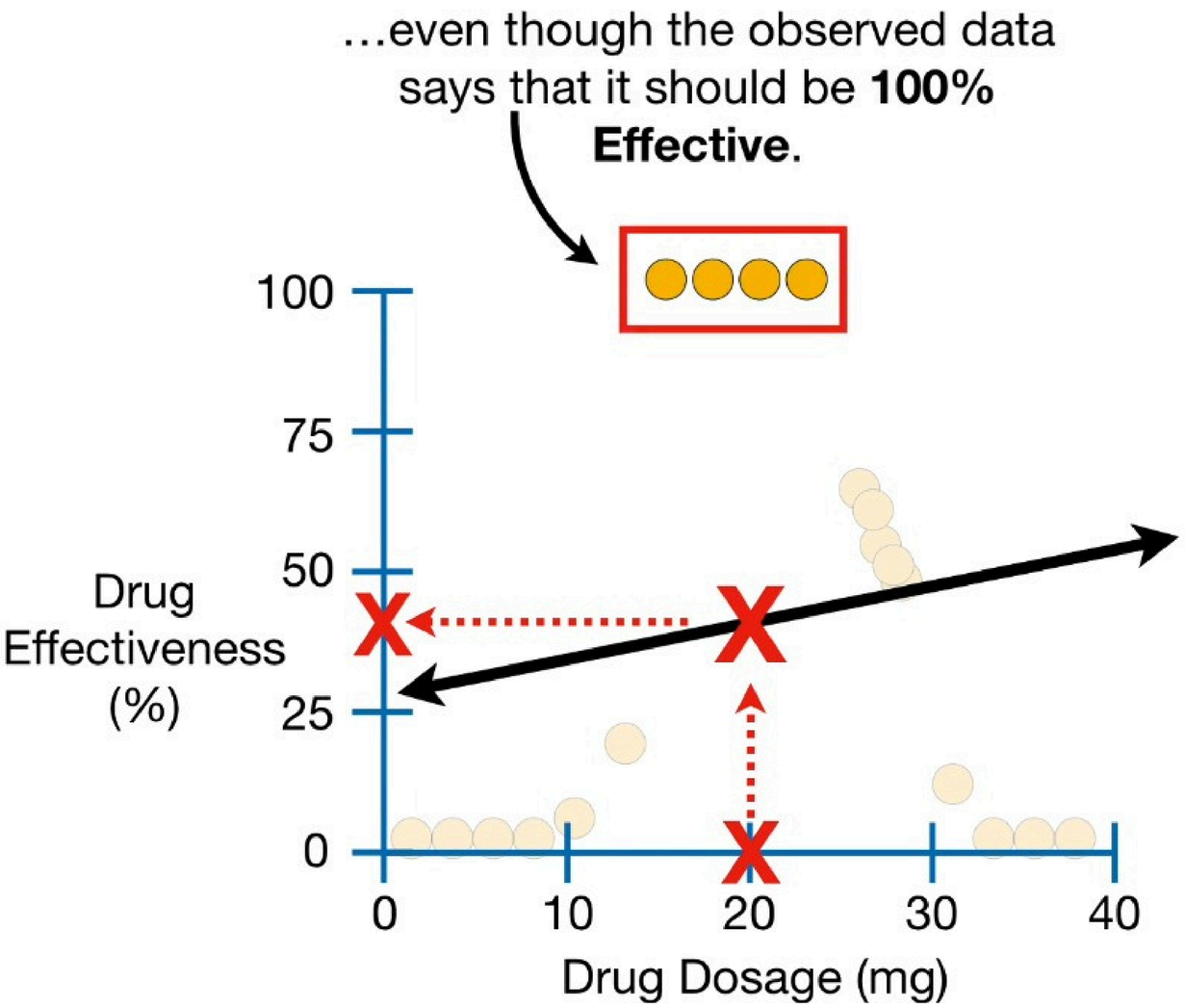
However, what if the data looked like this?



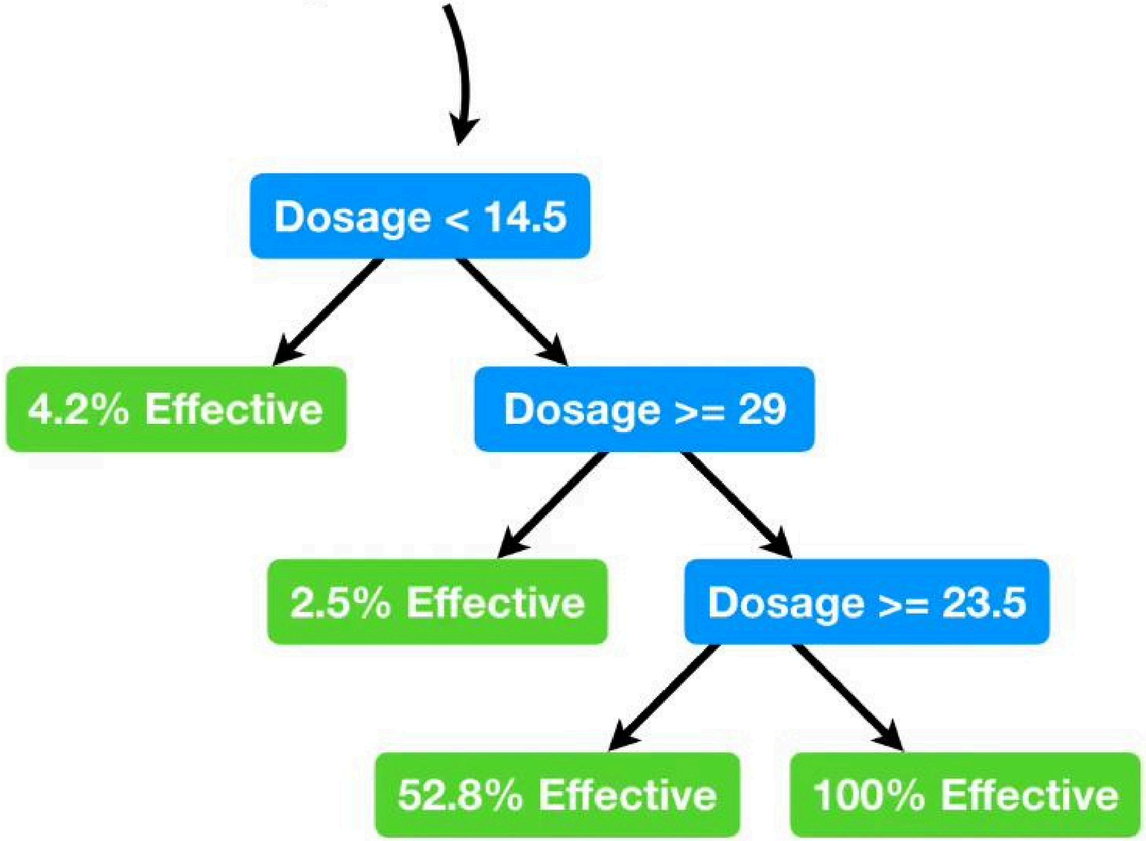

In this case, fitting a straight line
to the data will not be very useful.
100
75
50
25
0
Drug
Effectiveness
(%)
0
10
20
30
40
Drug Dosage (mg)

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

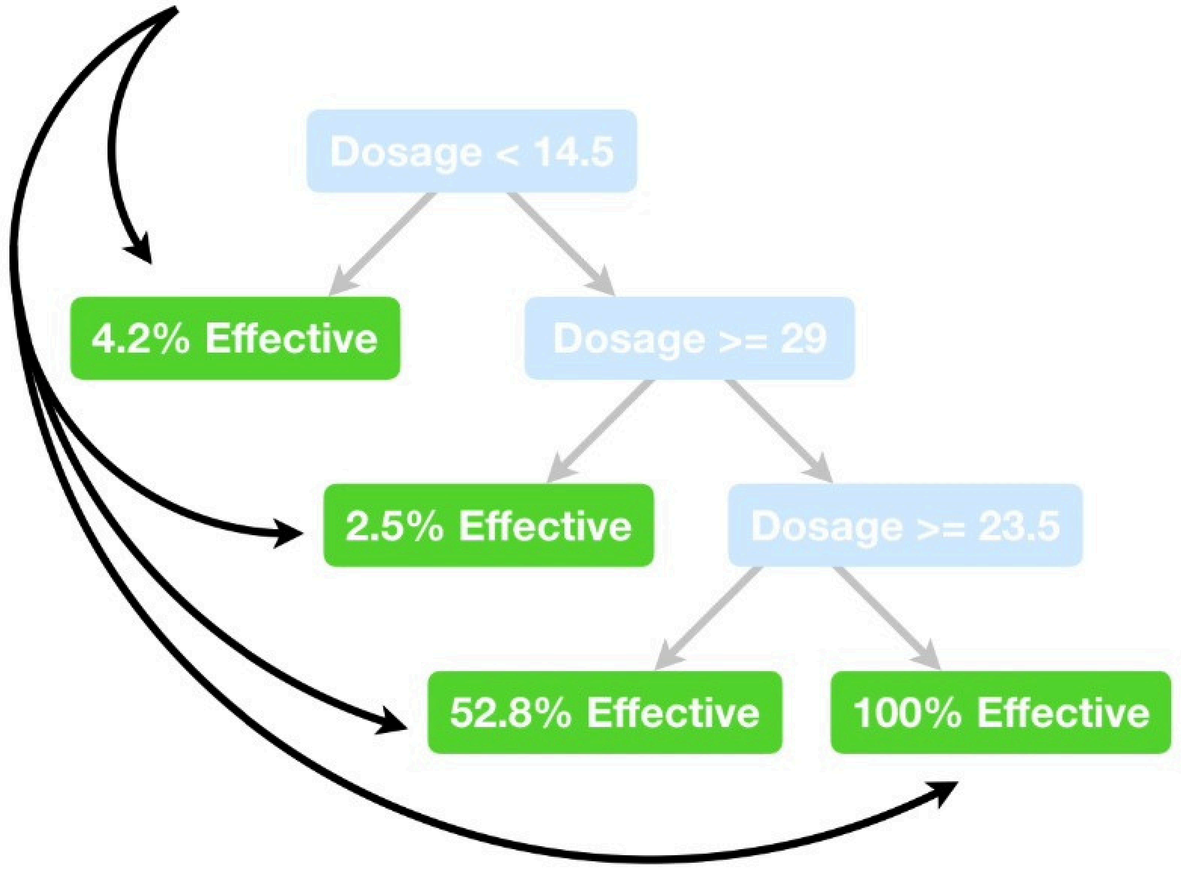



…even though the observed data says that it should be **100% Effective**.
100
75
50
25
0
Drug Effectiveness (%)
0
10
20
30
40
Drug Dosage (mg)

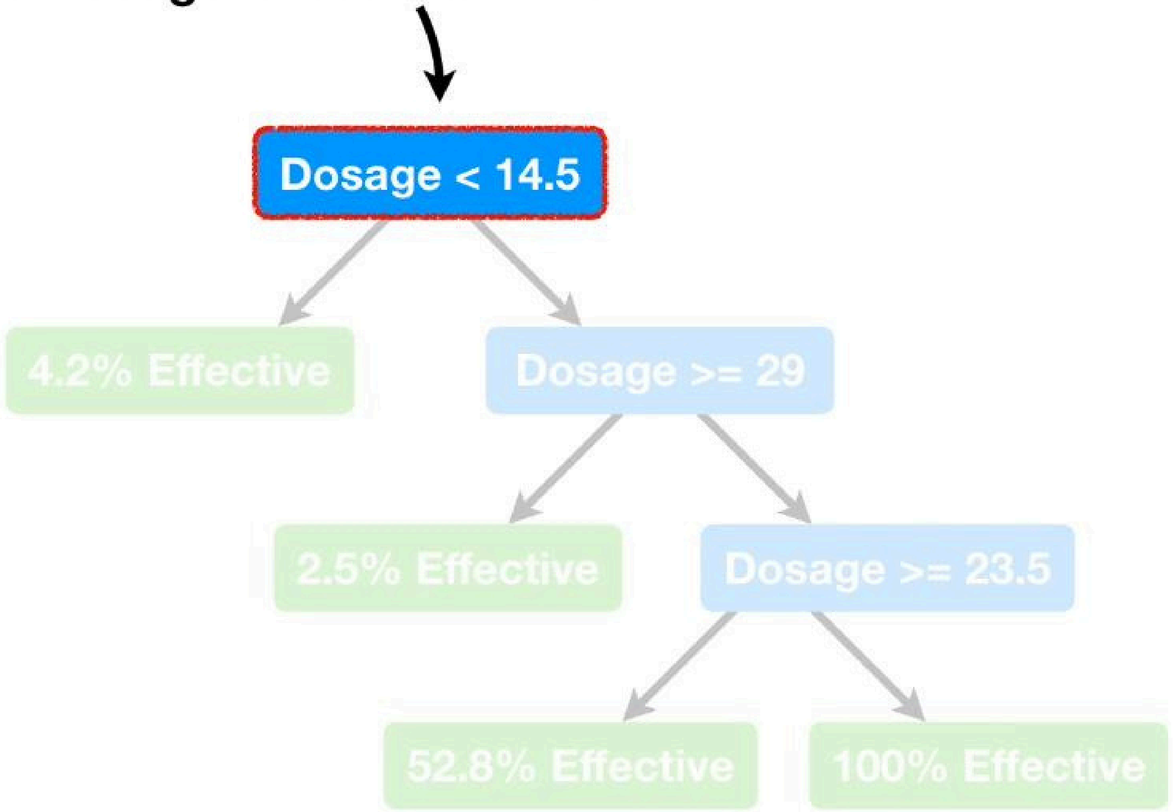
One option is to use a **Regression Tree**.



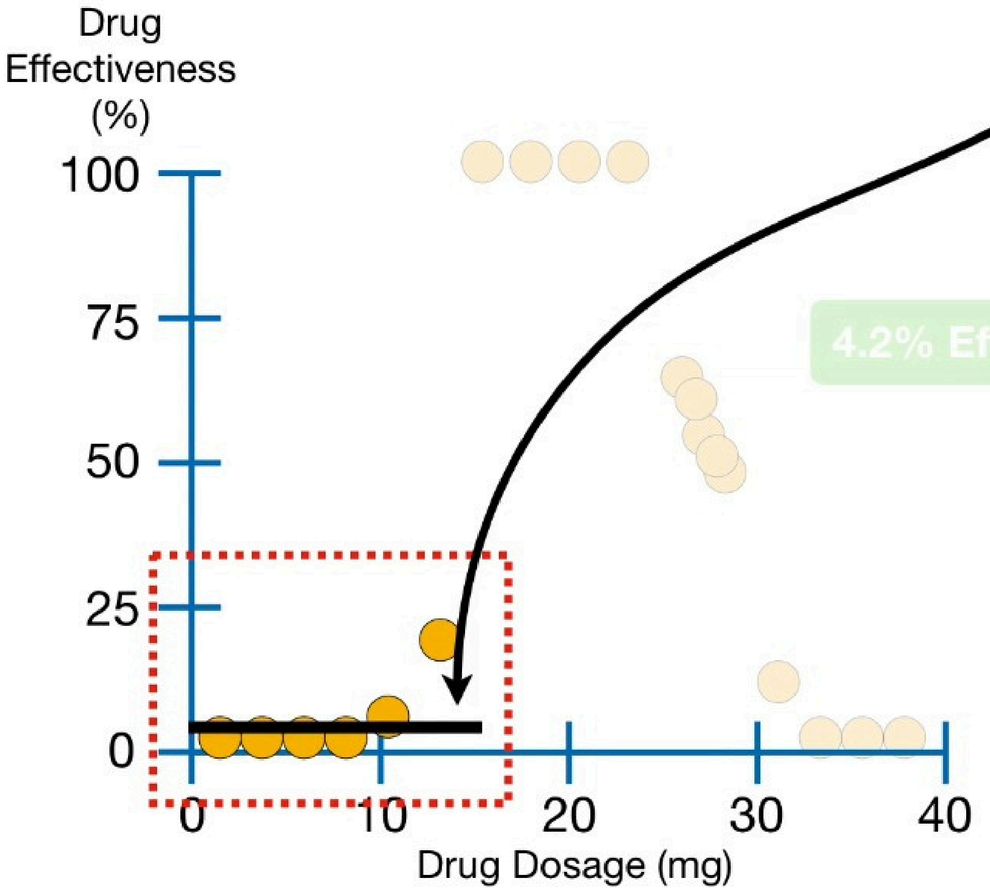
In a **Regression Tree**, each leaf represents a numeric value.

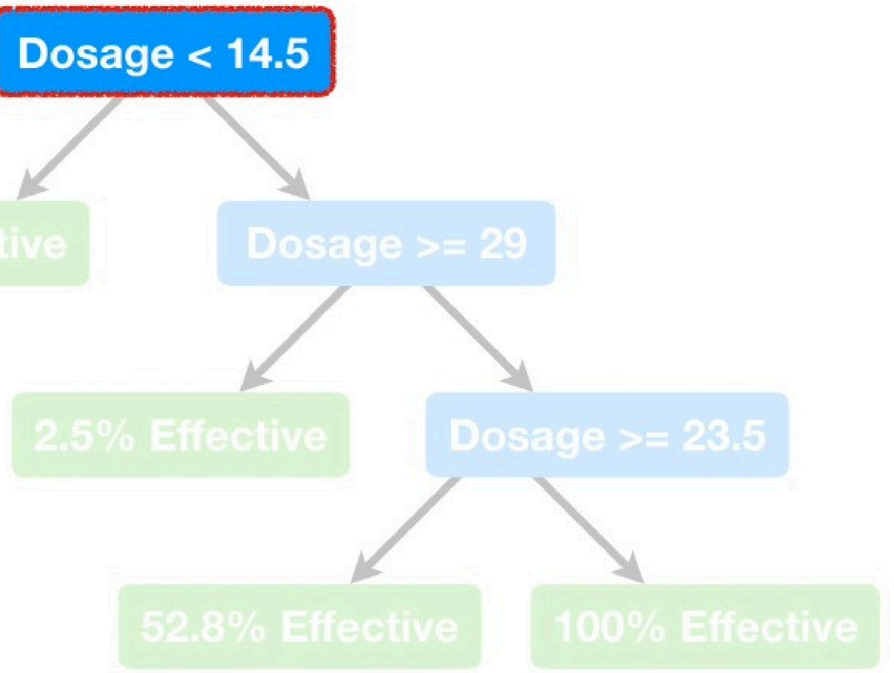


With *this* **Regression Tree**, we start by asking if the **Dosage** is less than **14.5**.

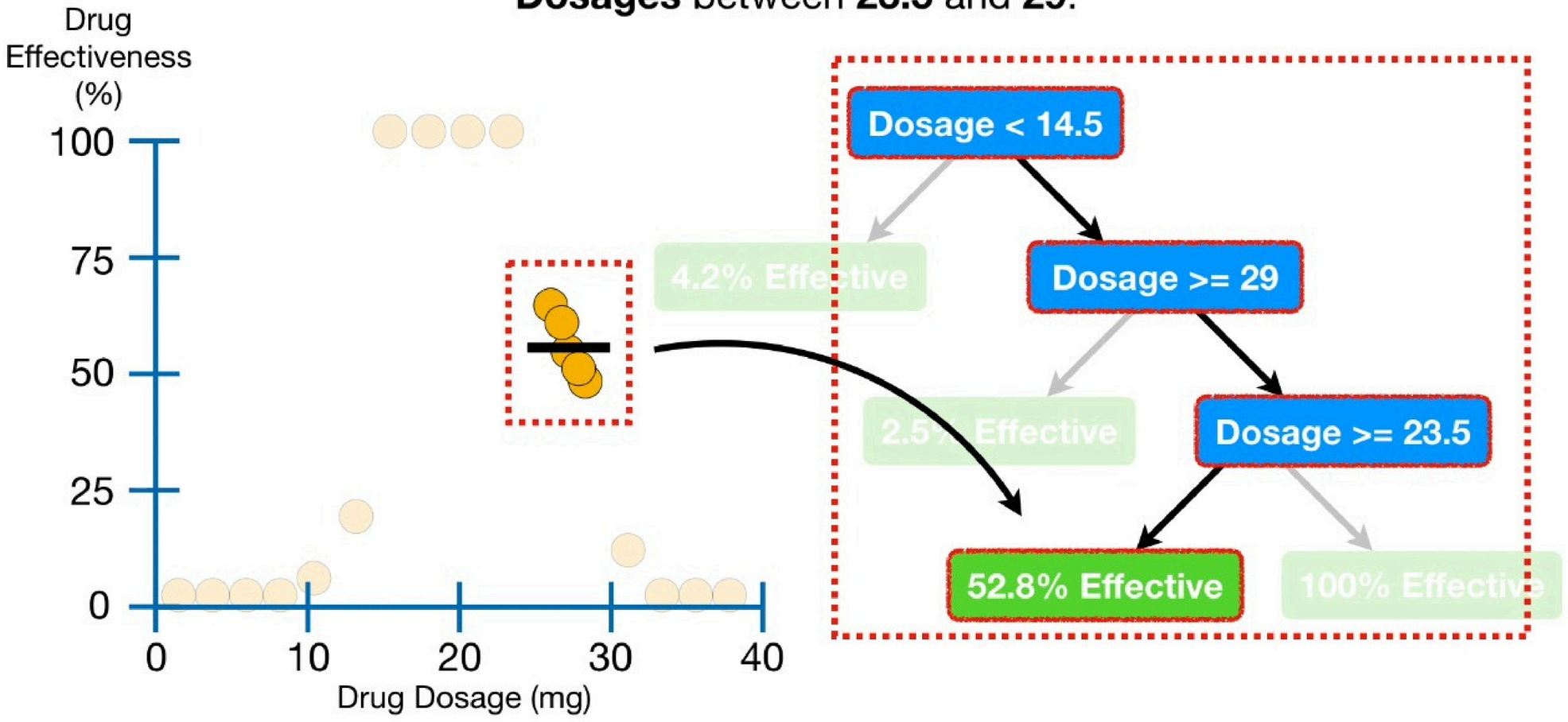


…and the average **Drug Effectiveness** for these **6** observations is **4.2%**…

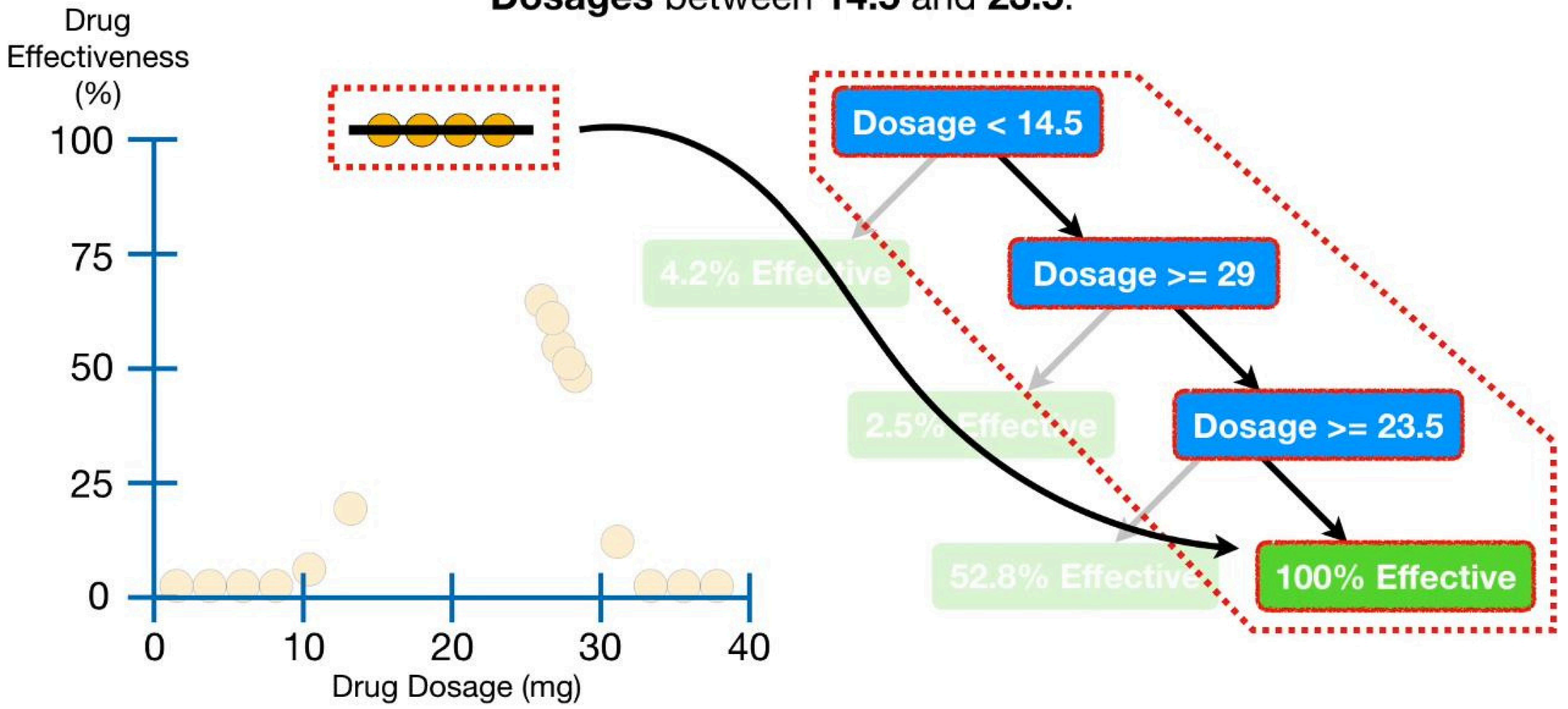




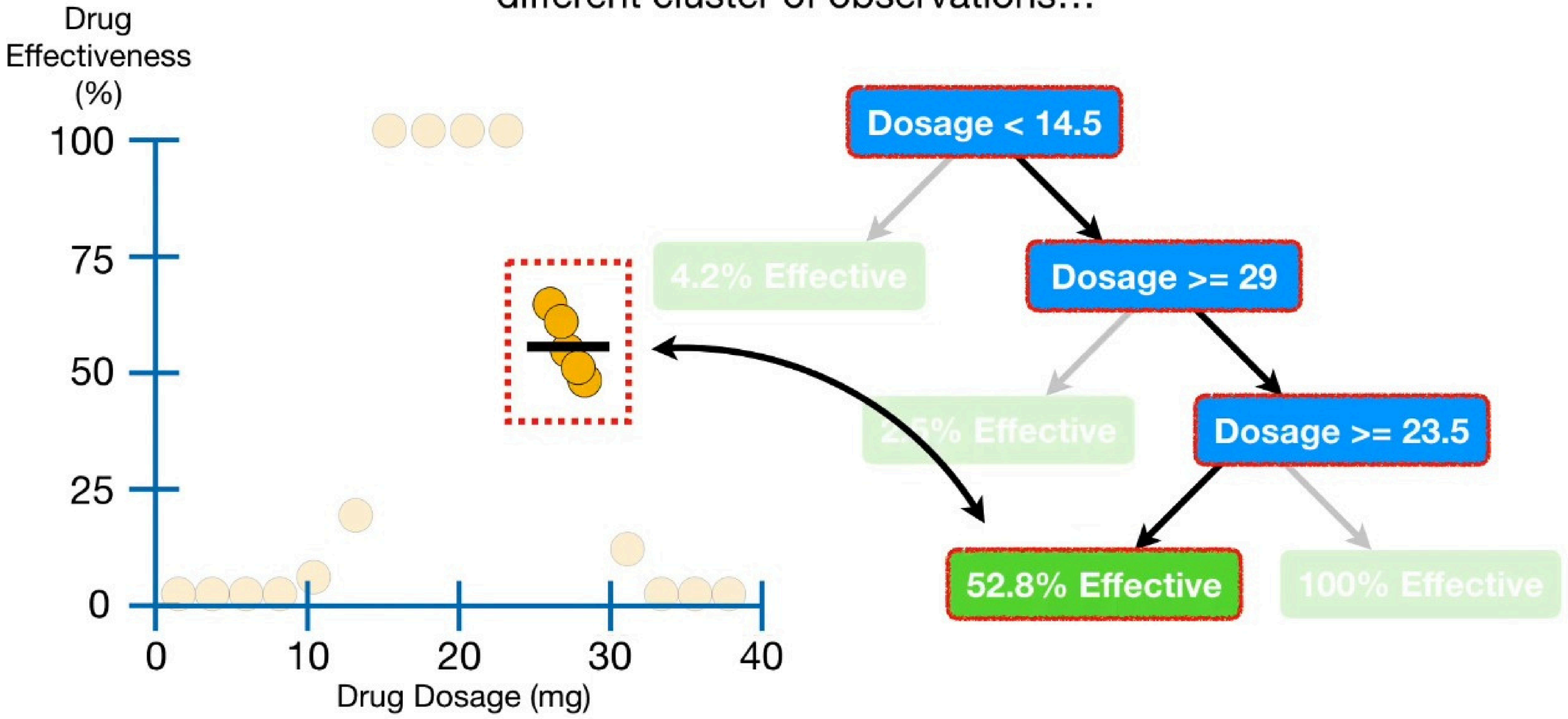
…so the tree uses the average value, **52.8%**, as its prediction for people with **Dosages** between **23.5** and **29**.



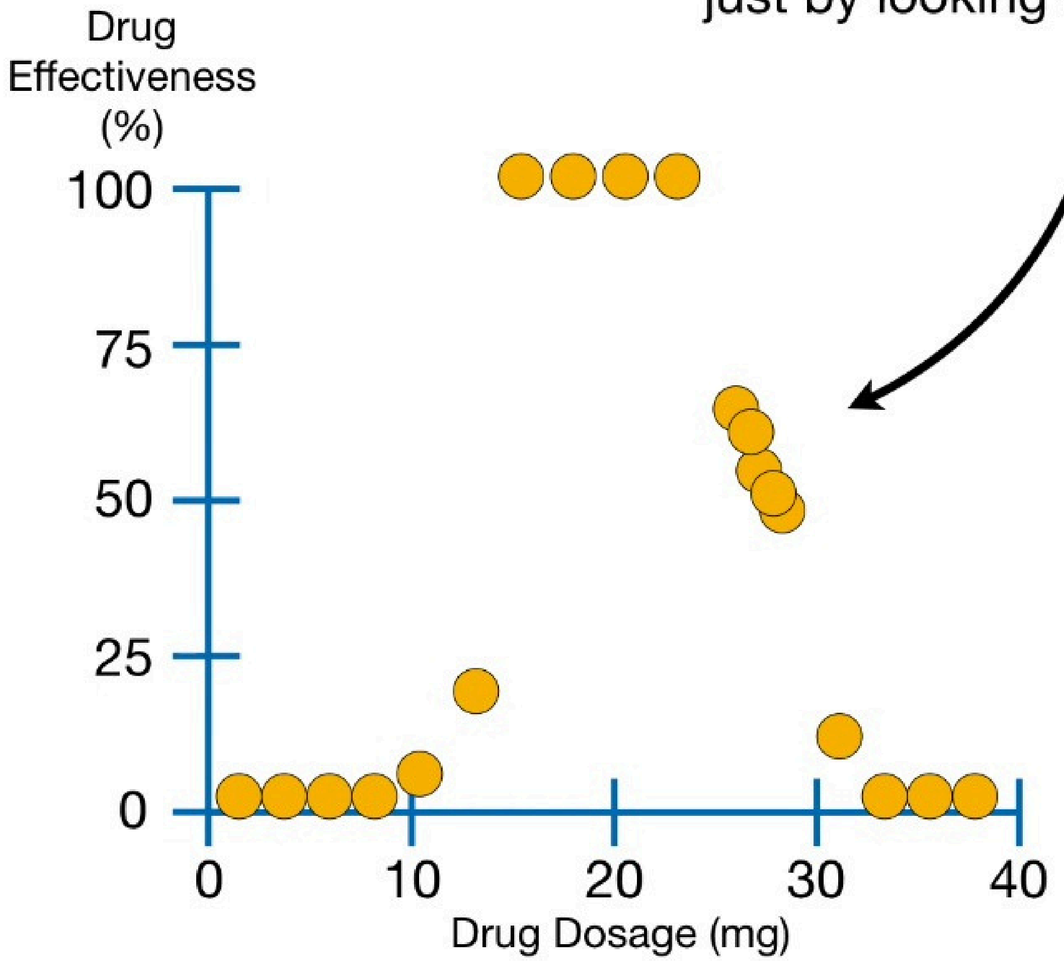
…so the tree uses the average value, **100%**, as its prediction for people with **Dosages** between **14.5** and **23.5**.



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



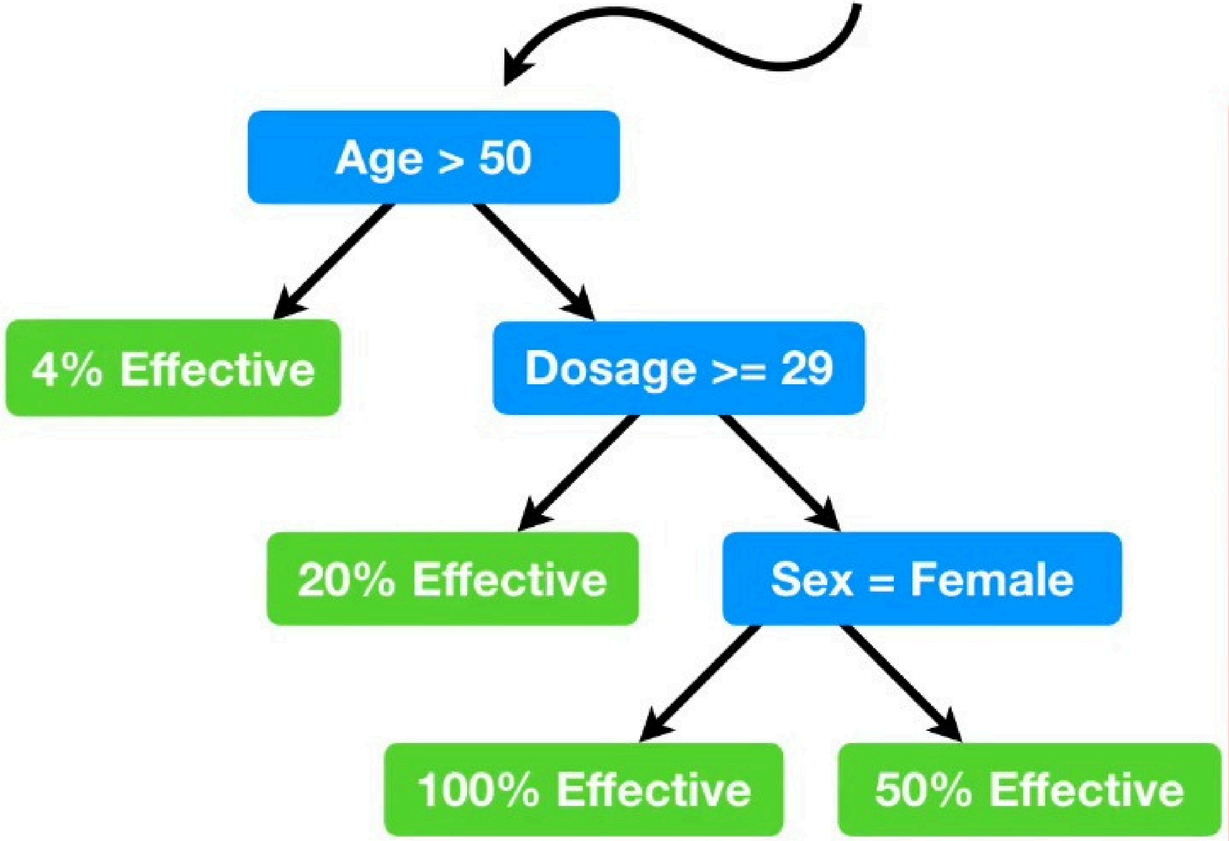
At this point you might be thinking, “The **Regression Tree** is cool, but I can also predict **Drug Effectiveness** just by looking at the graph…”



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…

Now you fully understand the concept

Decision Trees

Random Forest