DATA ANALYSIS AND VISUALIZATION

INSTRUCTOR: UMME AMMARAH

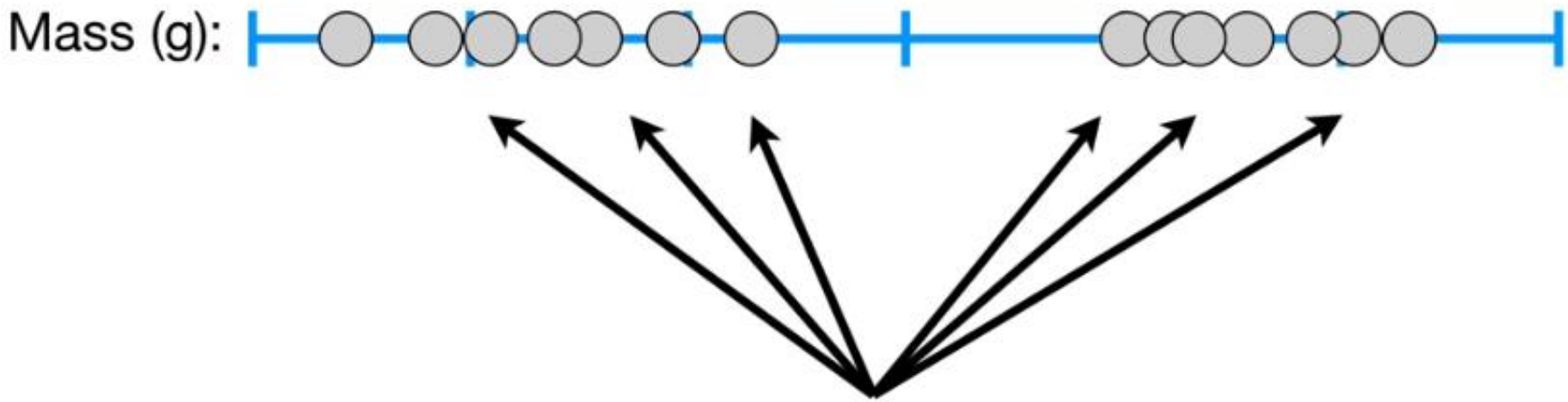
SUPPORT VECTOR MACHINE (SVM)

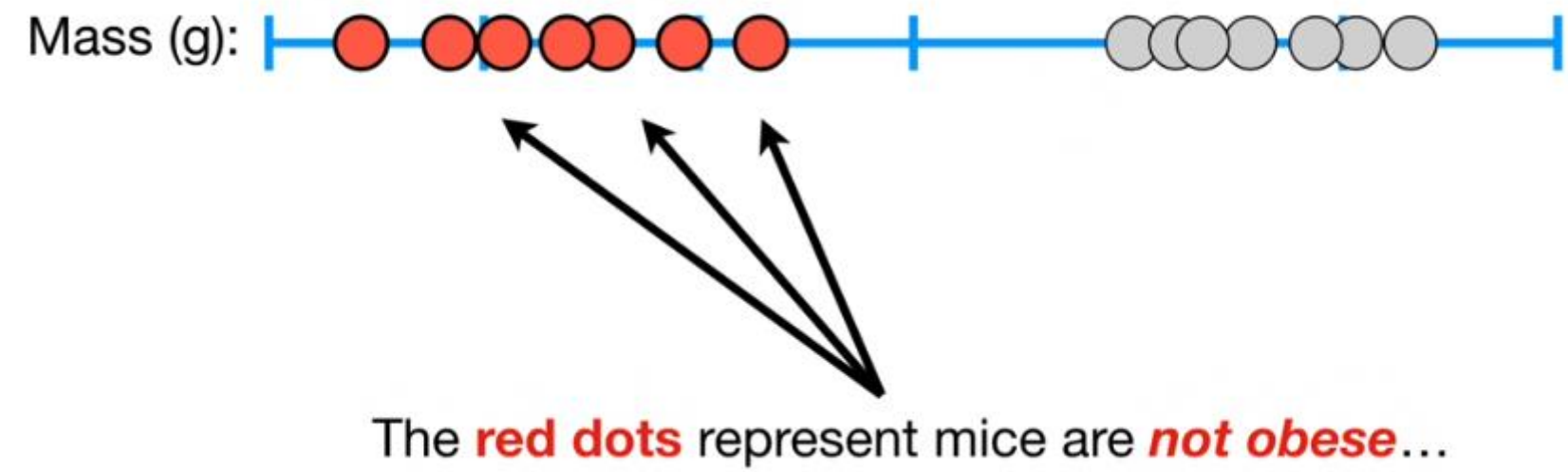
SVM

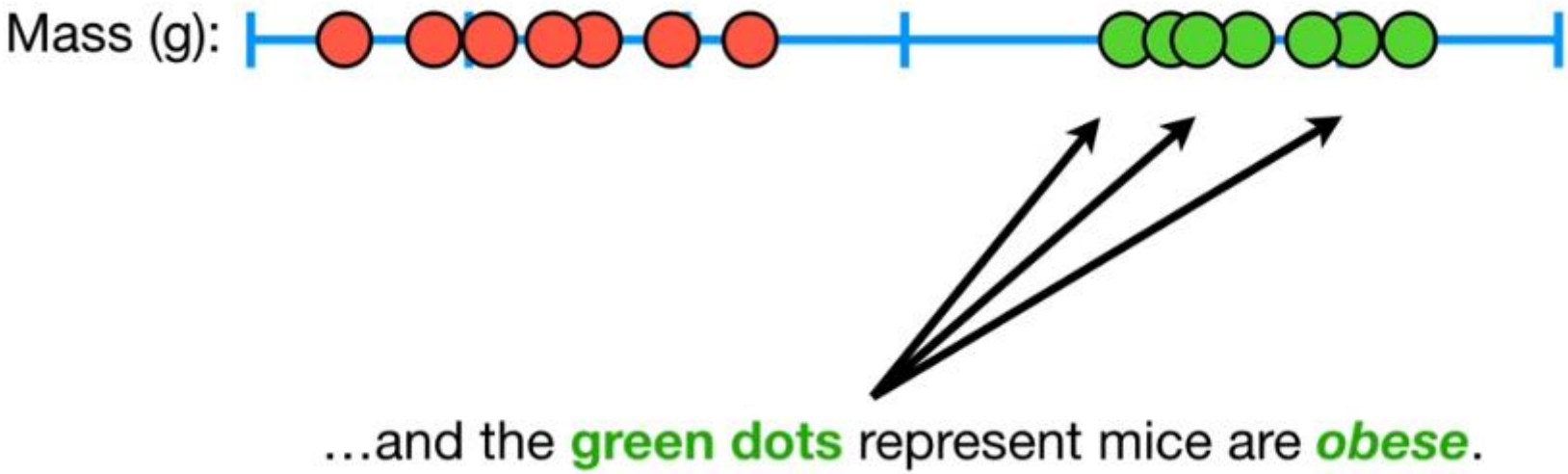
- Supervised machine learning algorithm.
- Can be used for classification and regression both, but works better for classification.
- Can solve binary class problem only.
- It works by finding the best hyperplane between the two classes.

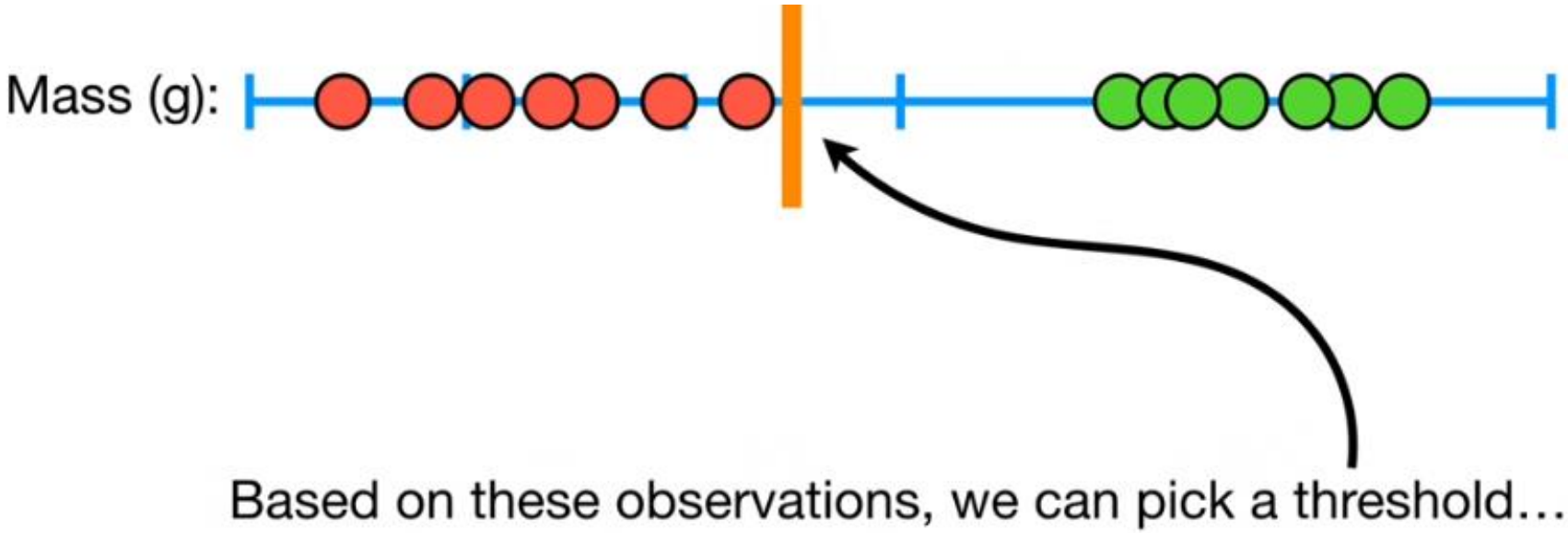
TYPES

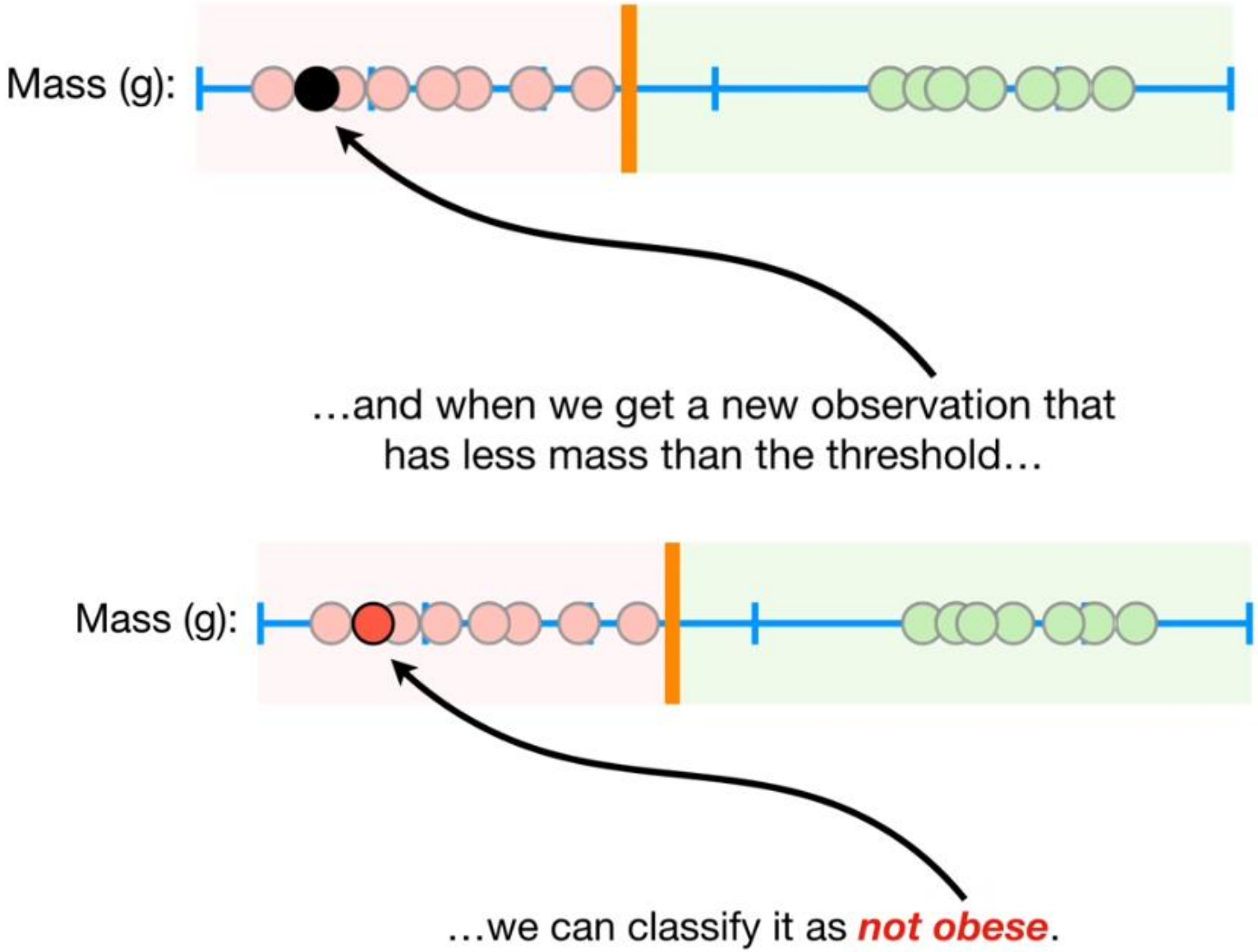
- Linear SVM
  - When the data is perfectly linearly separable only then we can use Linear SVM.
- Non-Linear SVM
  - When the data is not linearly separable then we can use Non-Linear SVM.

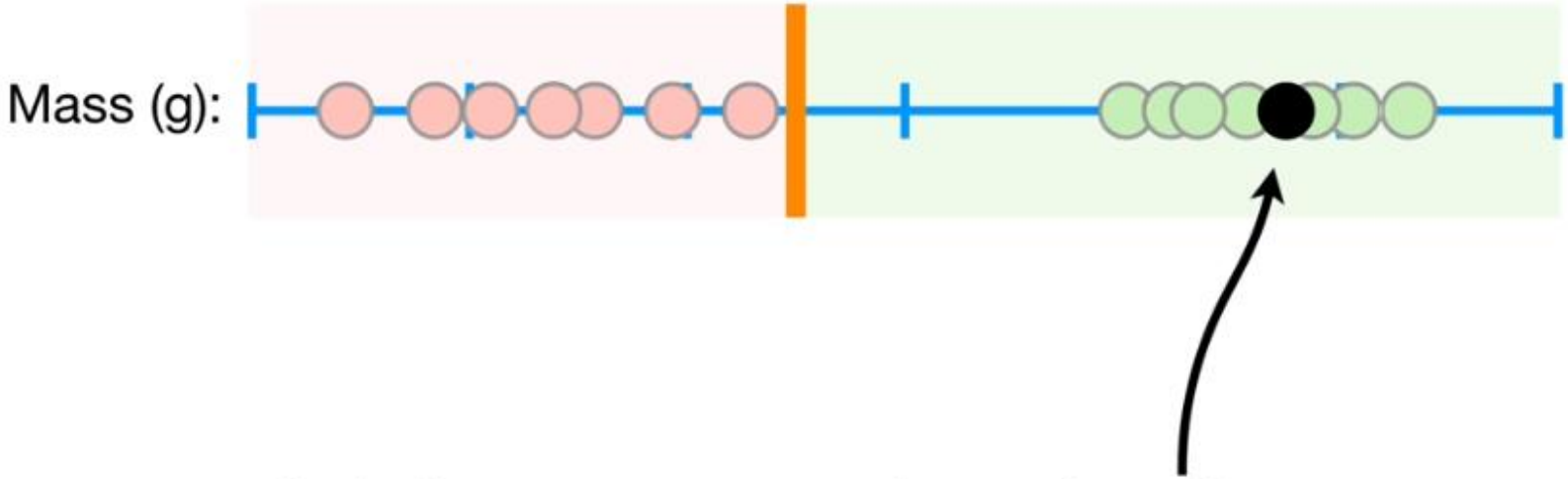

Mass (g):
Let's start by imagining we measured
the mass of a bunch of mice…

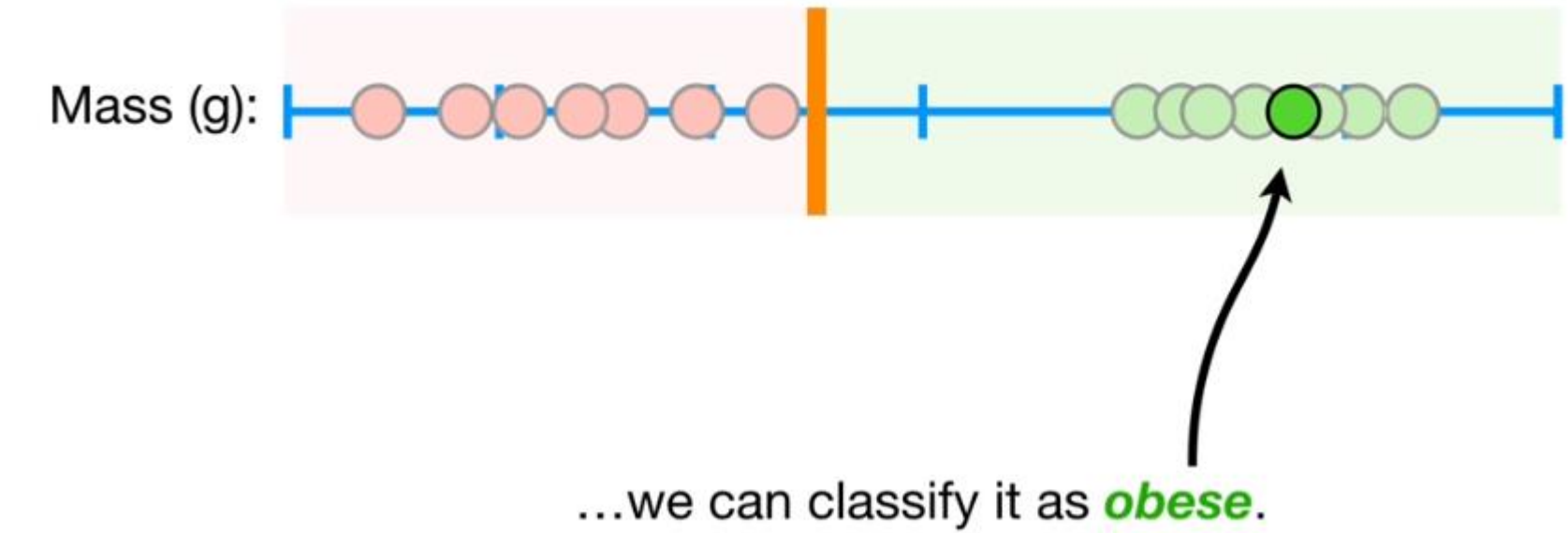

Mass (g):
The red dots represent mice are not obese…

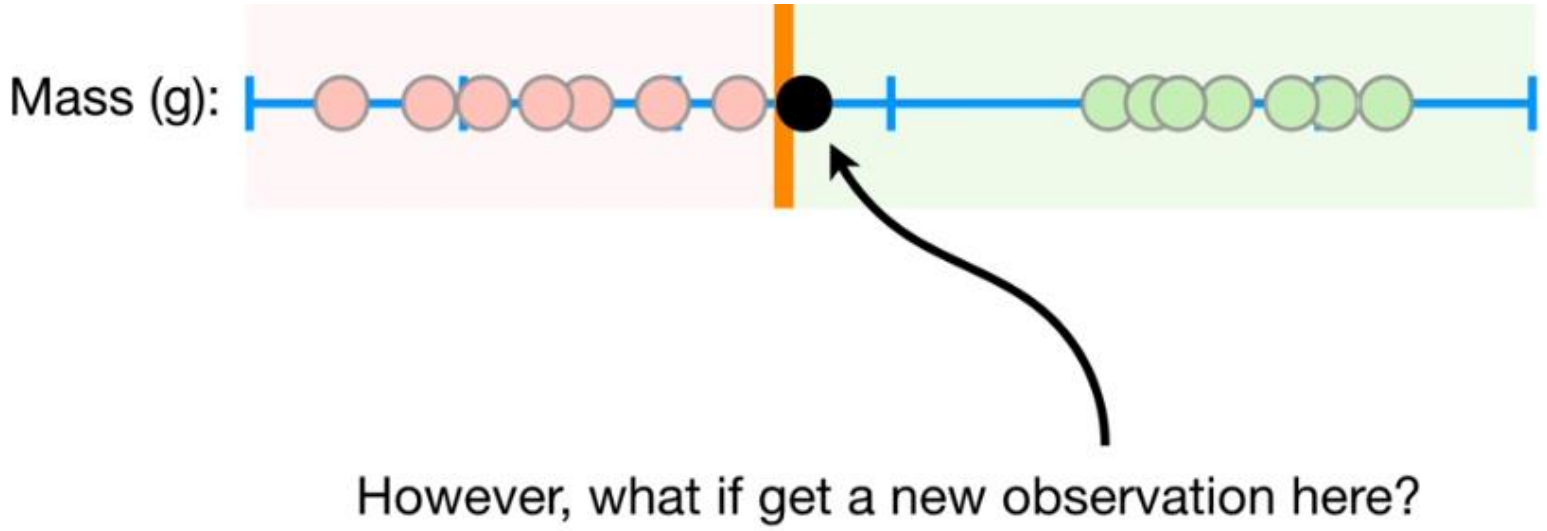

Mass (g):
…and the green dots represent mice are obese.


Mass (g):
Based on these observations, we can pick a threshold…

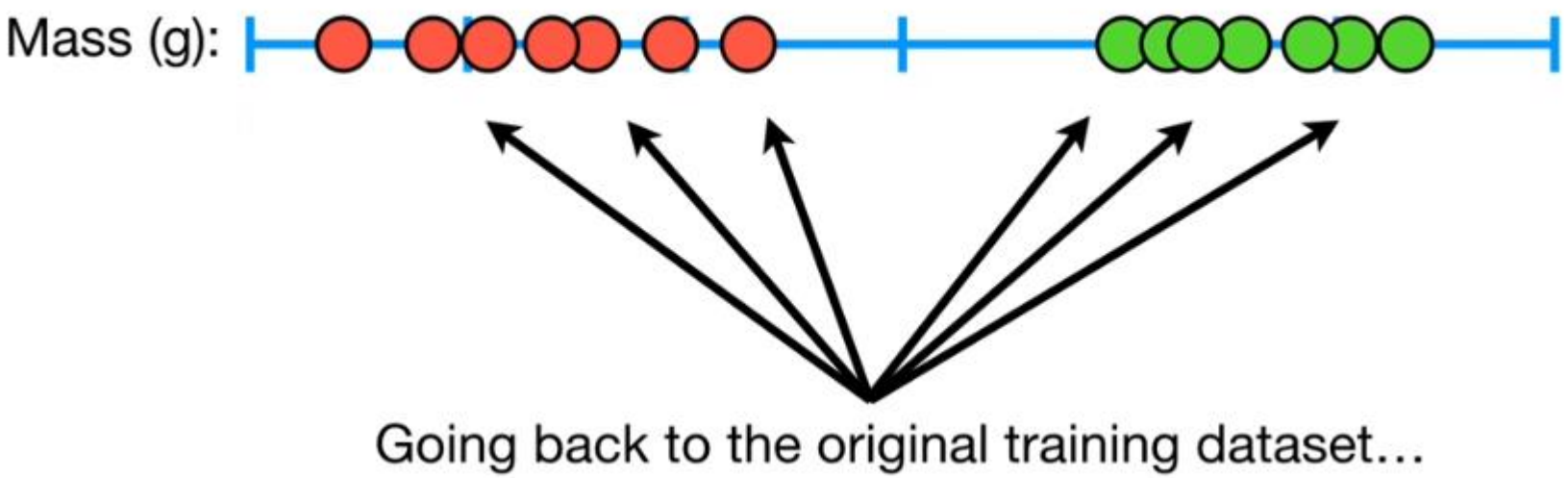

Mass (g):
…and when we get a new observation that has less mass than the threshold…
Mass (g):
…we can classify it as not obese.

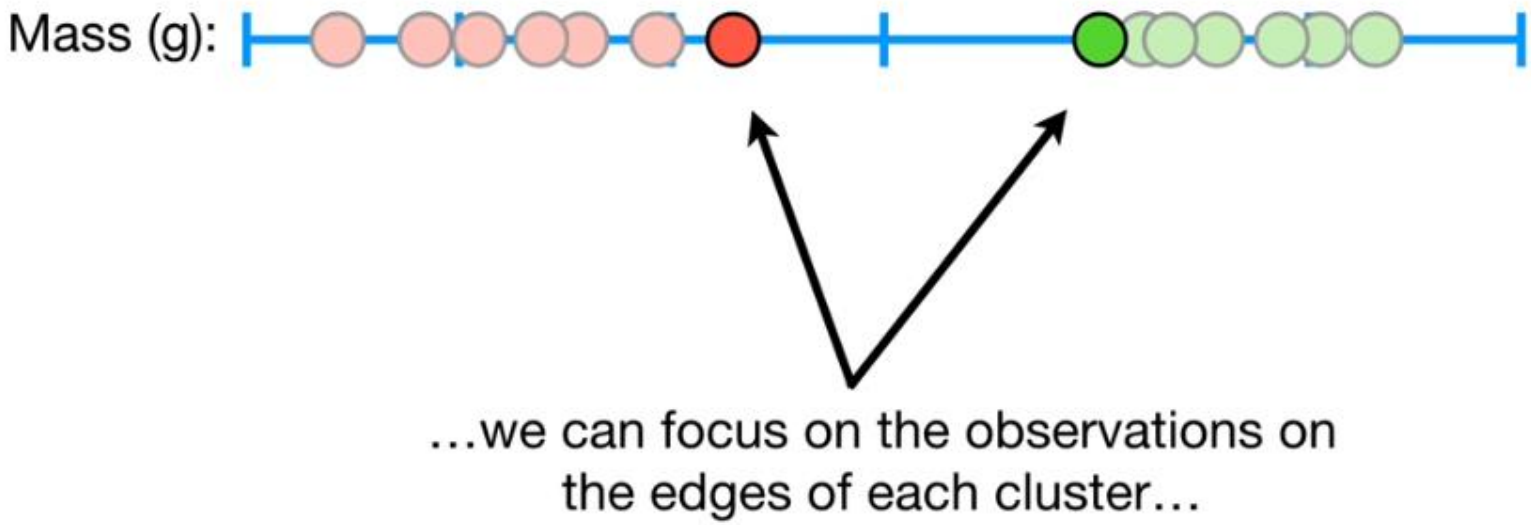

Mass (g):
And when we get a new observation with more mass than the threshold…

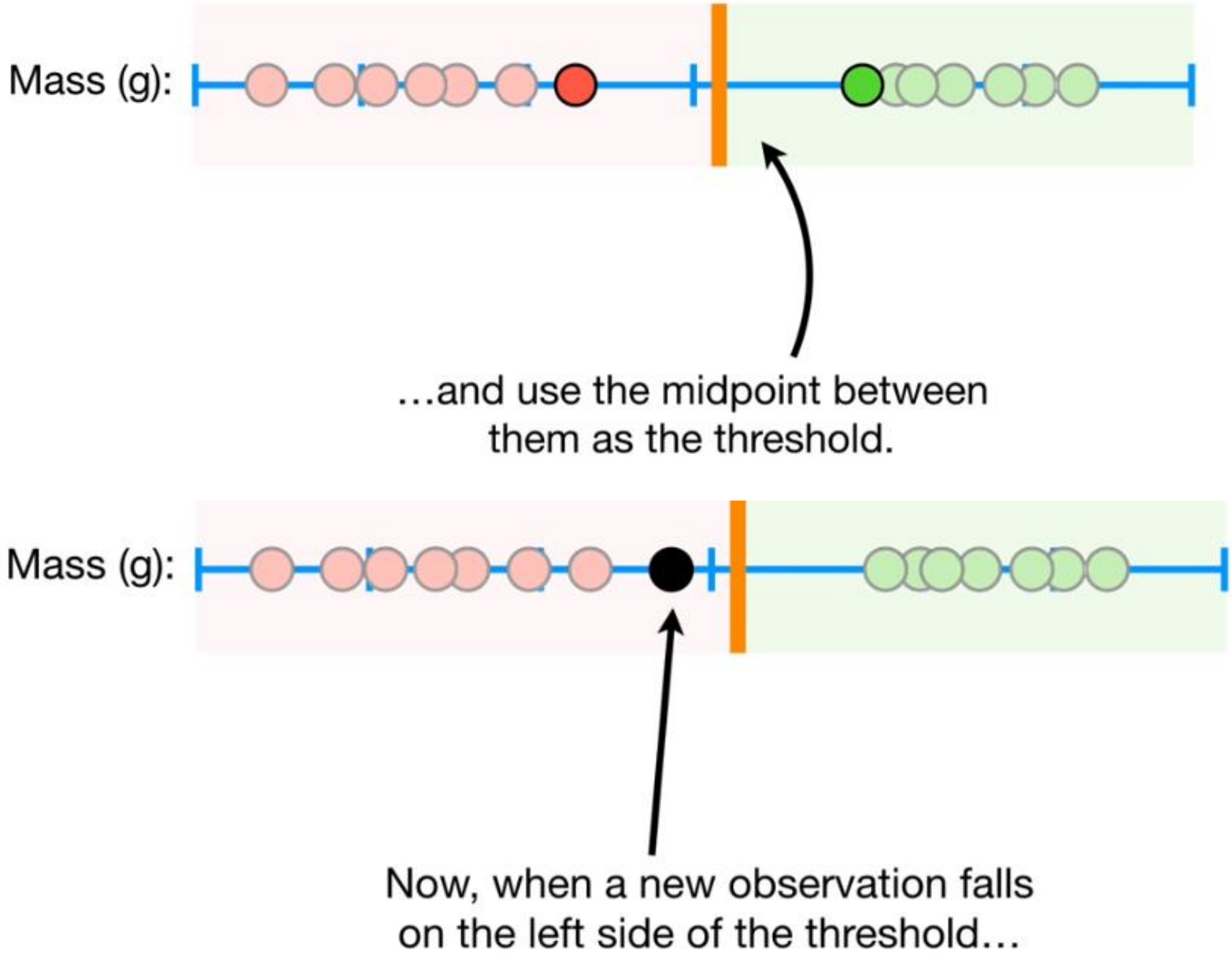

Mass (g):
…we can classify it as obese.

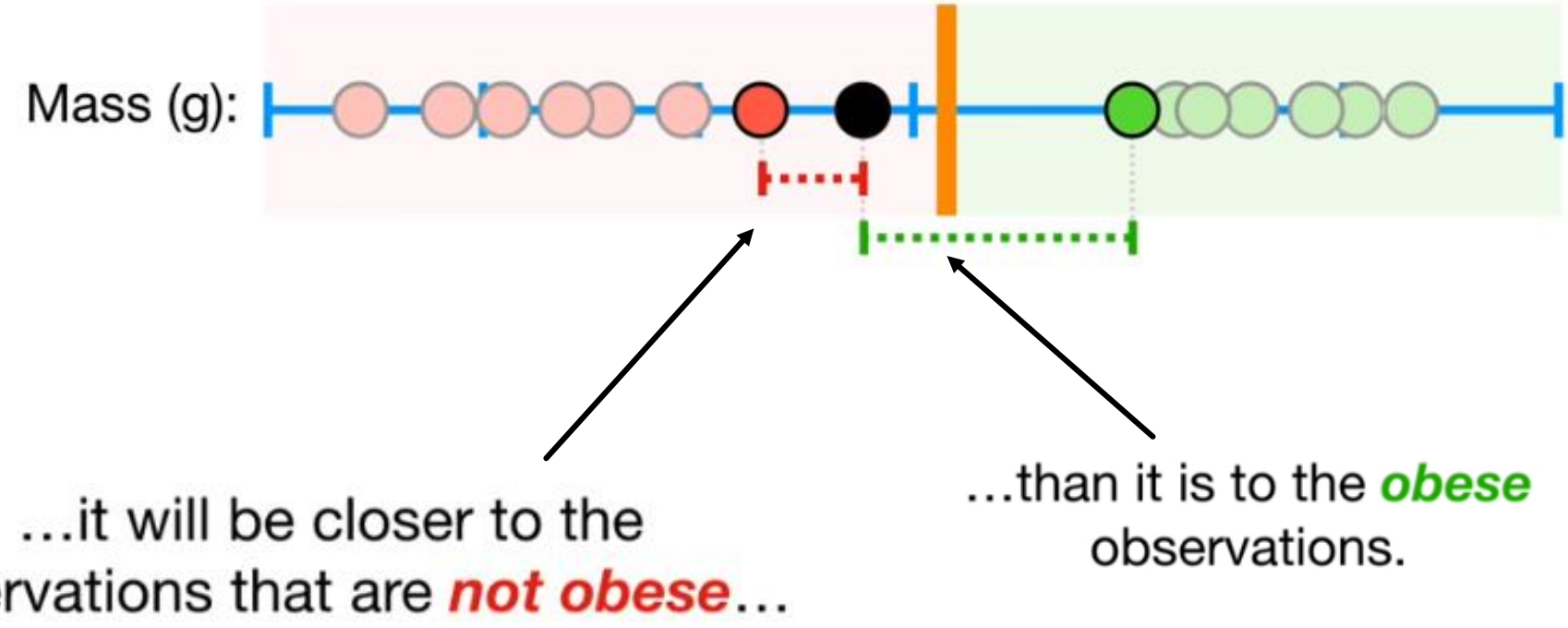


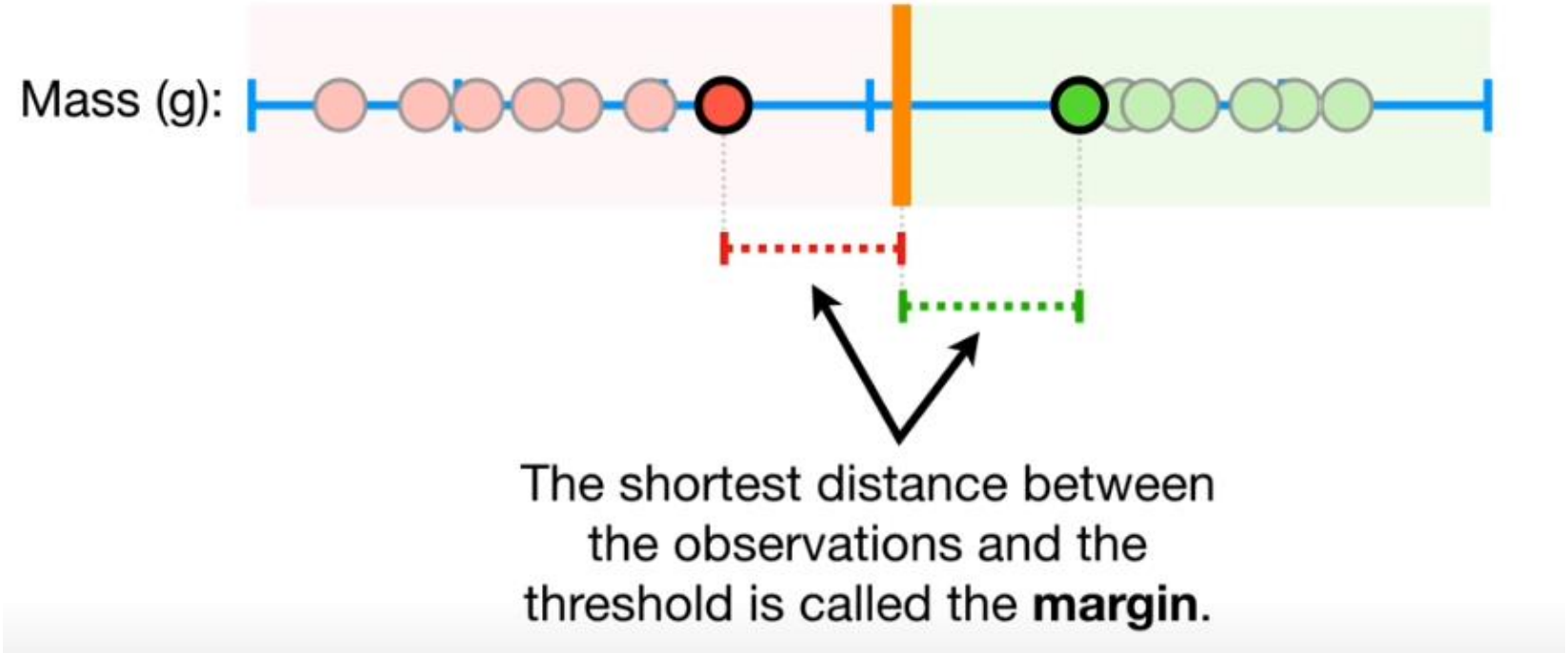
Because this observation has more mass than the threshold we classify it as obese. But that Does not make sense because this point is much closer to the observations that are not obese.


Mass (g):
Going back to the original training dataset…


Mass (g):
…we can focus on the observations on
the edges of each cluster…


Mass (g):
…and use the midpoint between them as the threshold.
Mass (g):
Now, when a new observation falls on the left side of the threshold…

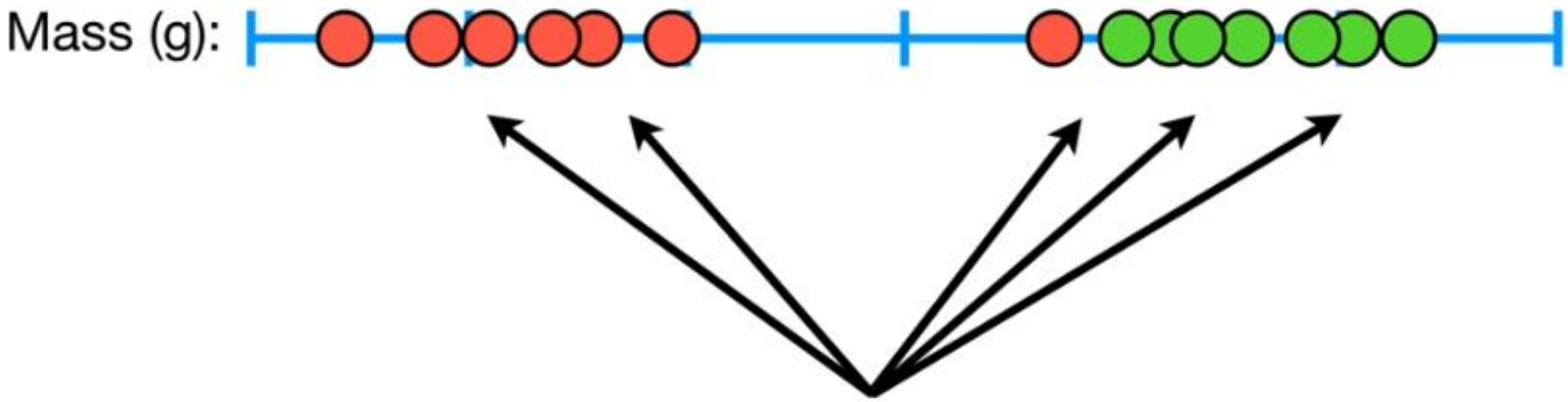

Mass (g):
…it will be closer to the observations that are not obese…
…than it is to the obese observations.



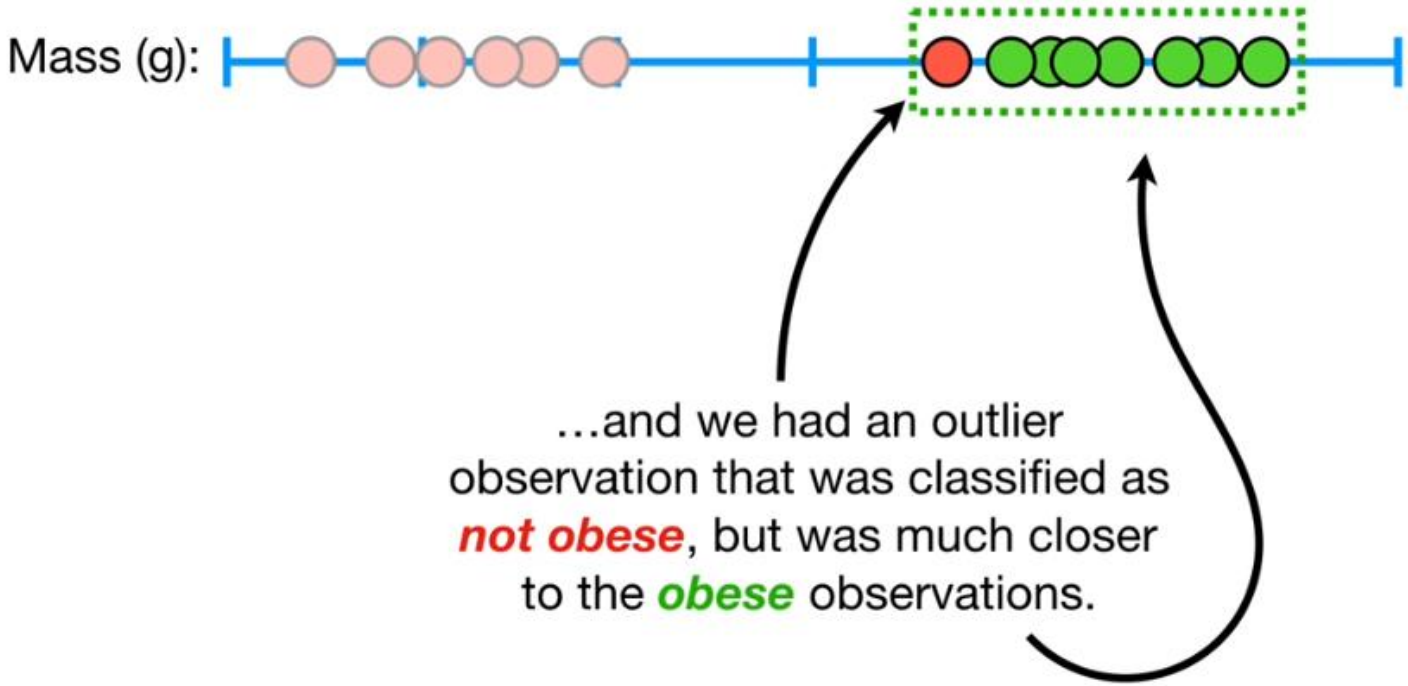
When we use the threshold that gives us the largest margin to make classifications we are using Maximal Margin Classifier.

MARGIN

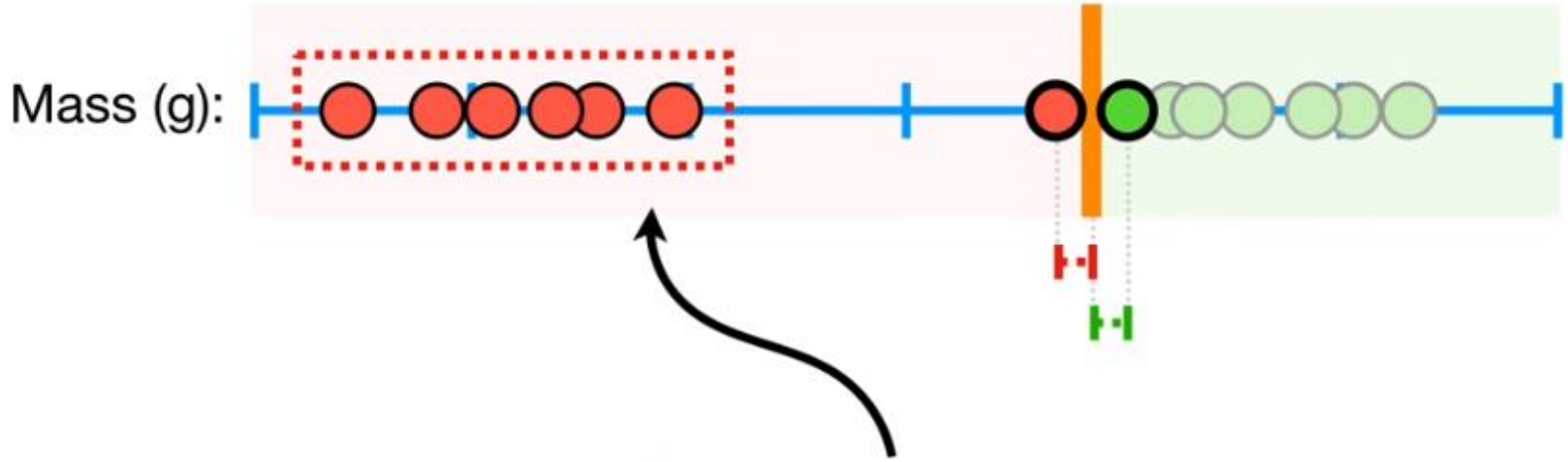
- It is the distance between the hyperplane and the observations closest to the hyperplane (support vectors).
- In SVM large margin is considered a good margin.
- A larger margin indicates a greater degree of confidence in the classification, as it means that there is a larger gap between the decision boundary and the closest data points from each class.
- The margin is a measure of how well-separated the classes are in feature space.



…but what if our training data looked like this….

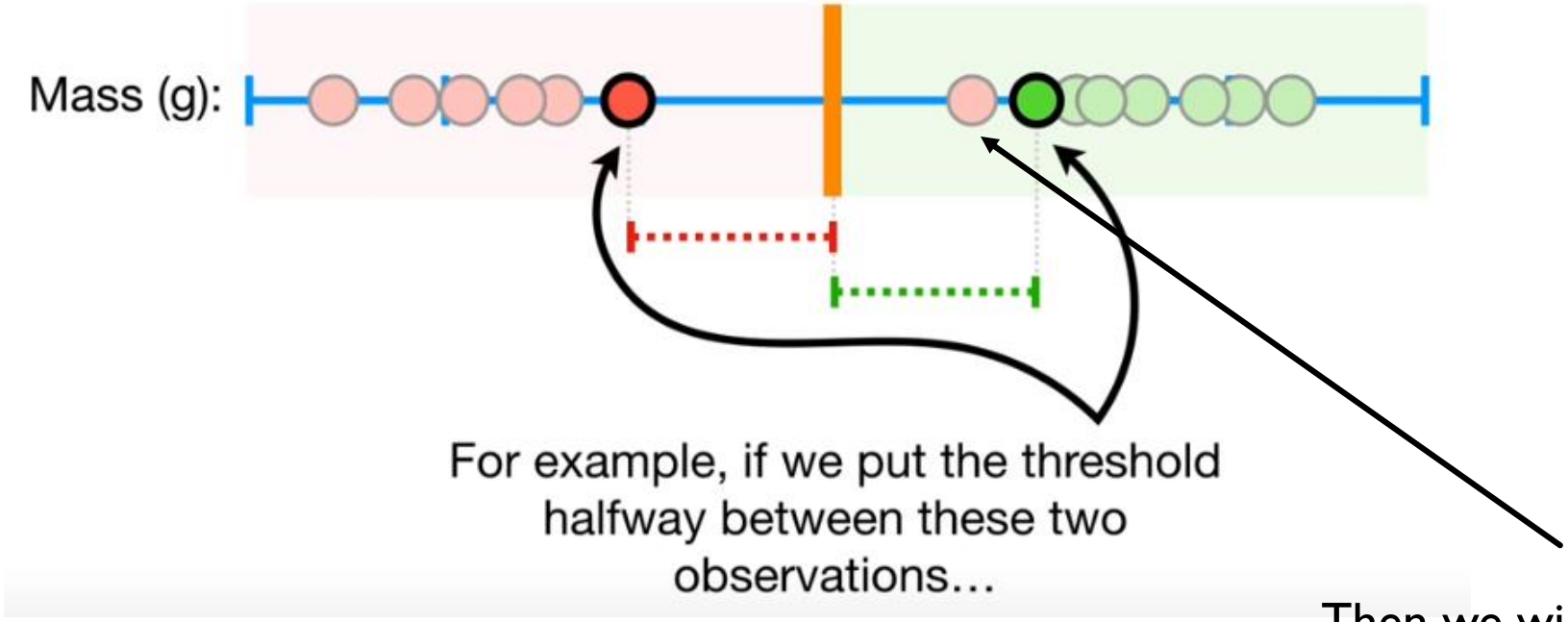


…and we had an outlier observation that was classified as ***not obese***, but was much closer to the ***obese*** observations.



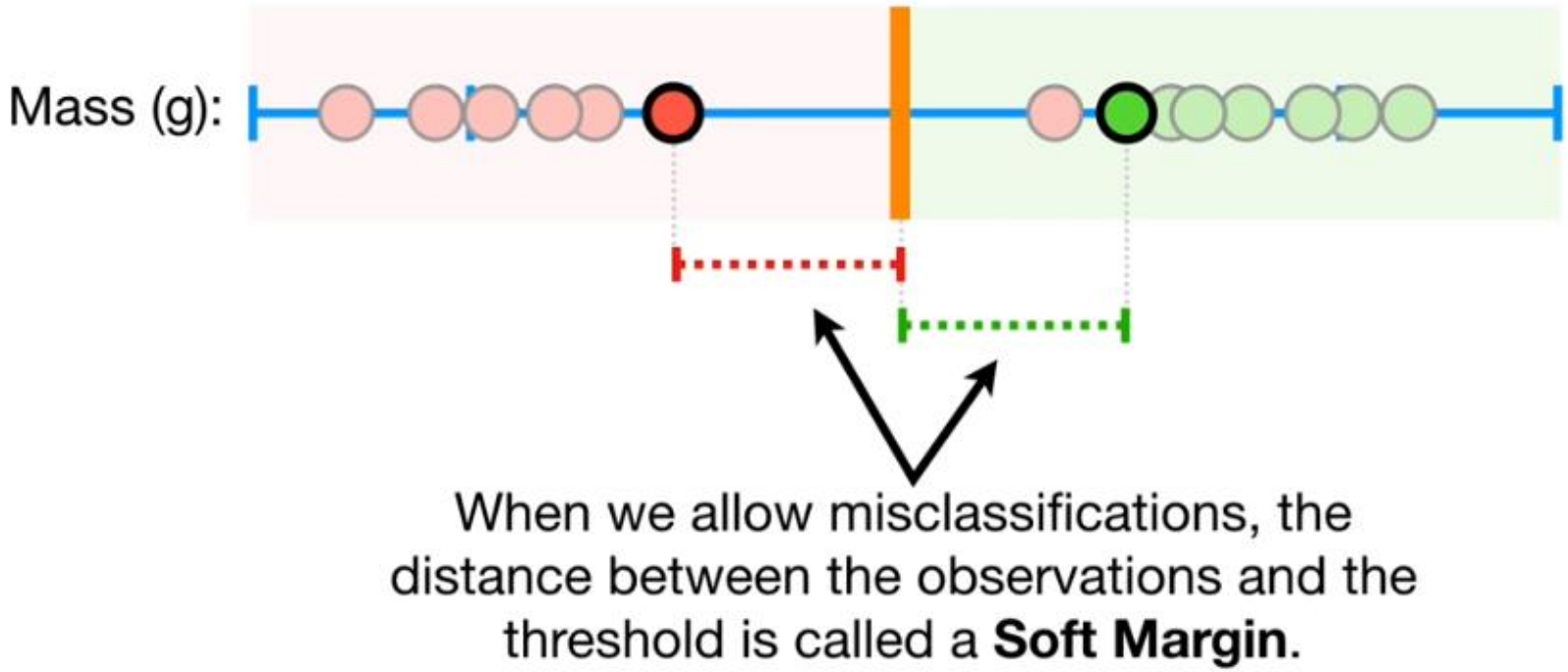
In this case maximal margin classifier would be super close to the obese observations and really far from the majority of not obese observations.

So maximal margin classifiers are very sensitive to outliers. To make a threshold that is not so sensitive to outliers we must allow misclassifications.



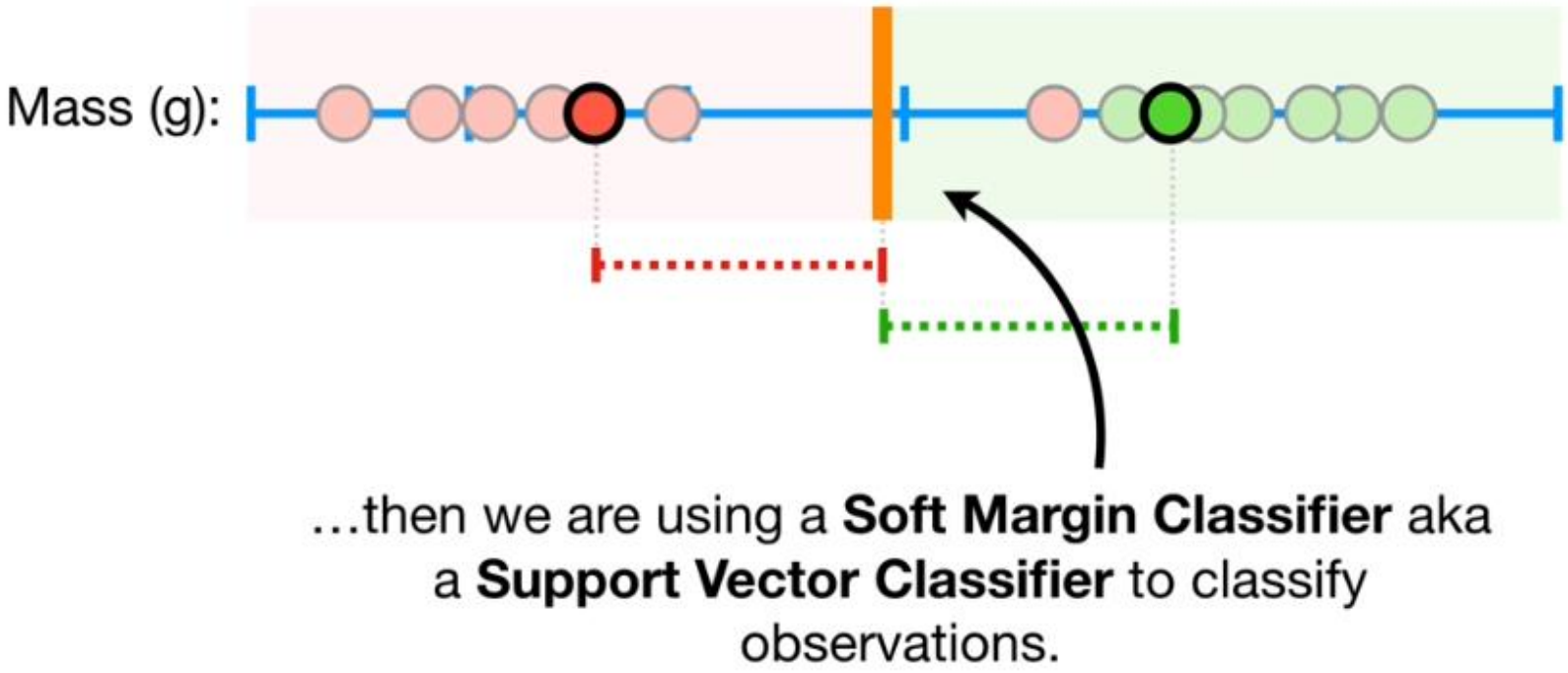
Then we will misclassify this point.
But we have generalized threshold now,

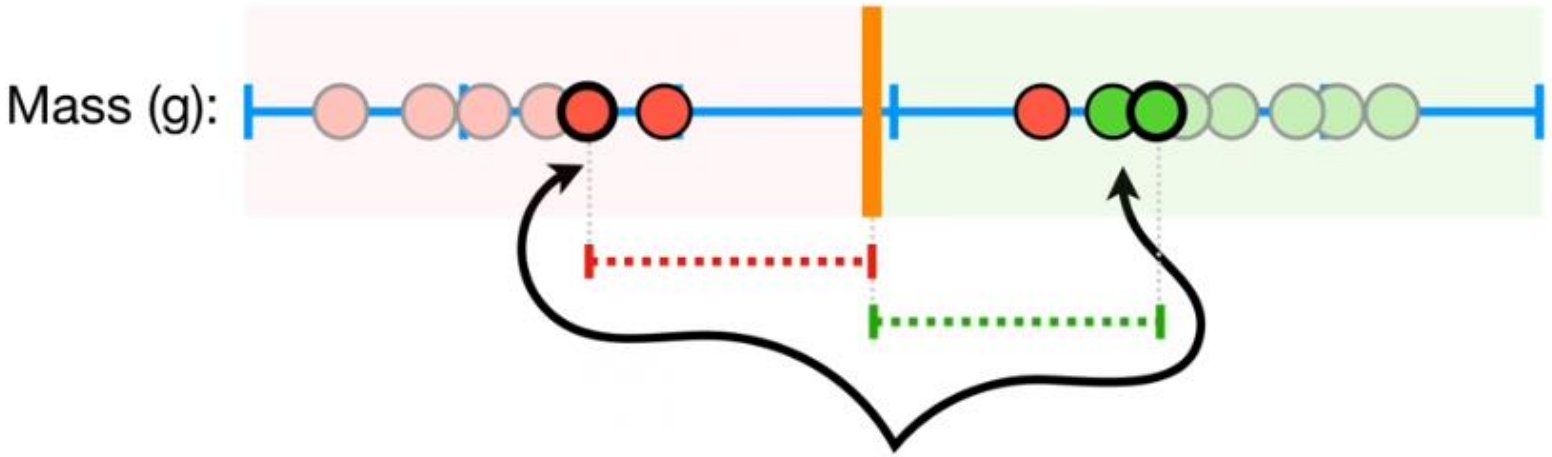
Choosing a threshold that allows error is an example of bias/variance tradeoff.



We use cross validation to determine how many misclassifications and observations to allow inside a soft margin to get the best classification.

When we use soft margin to determine the location of threshold





The name **Support Vector Classifier** comes from the fact that the observations on the edge *and within* the **Soft Margin** are called **Support Vectors**.

SUPPORT VECTORS

- They are the data points that lie closest to the decision boundary (hyperplane).

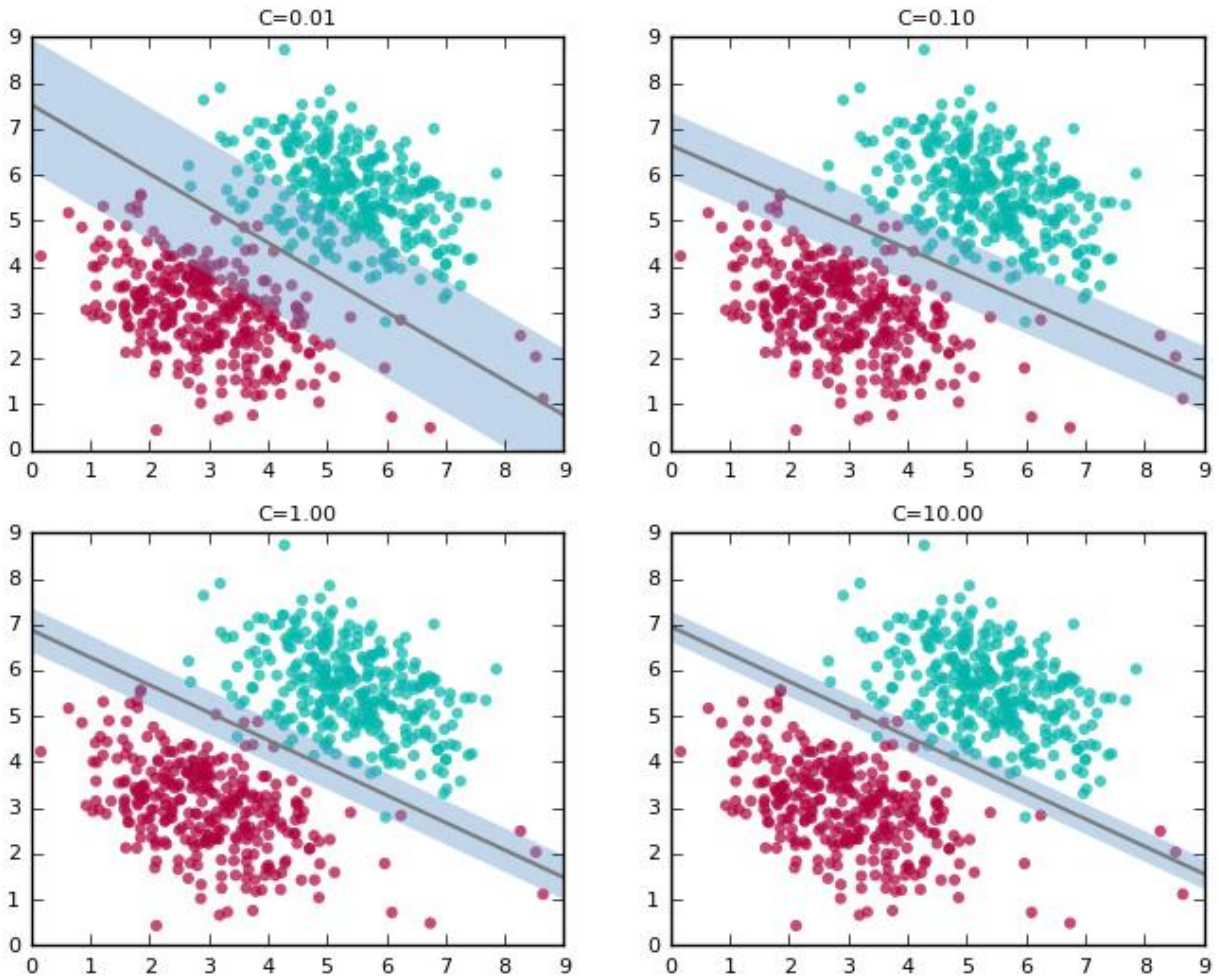
- These data points are important because they determine the position and orientation of the hyperplane, and thus have a significant impact on the classification accuracy of the SVM.

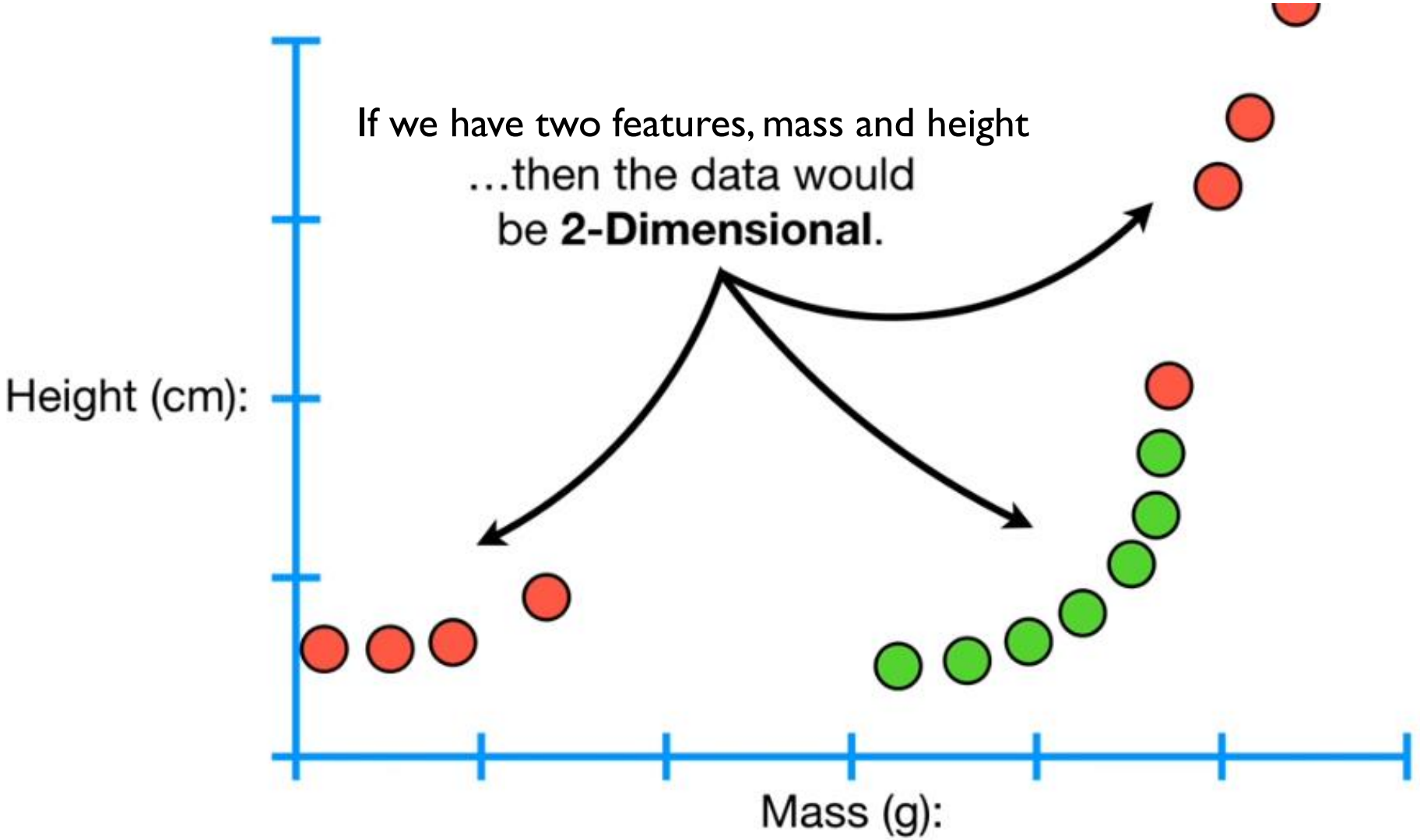
ALLOW ERRORS

A higher value of C implies you want lesser errors on the training data.

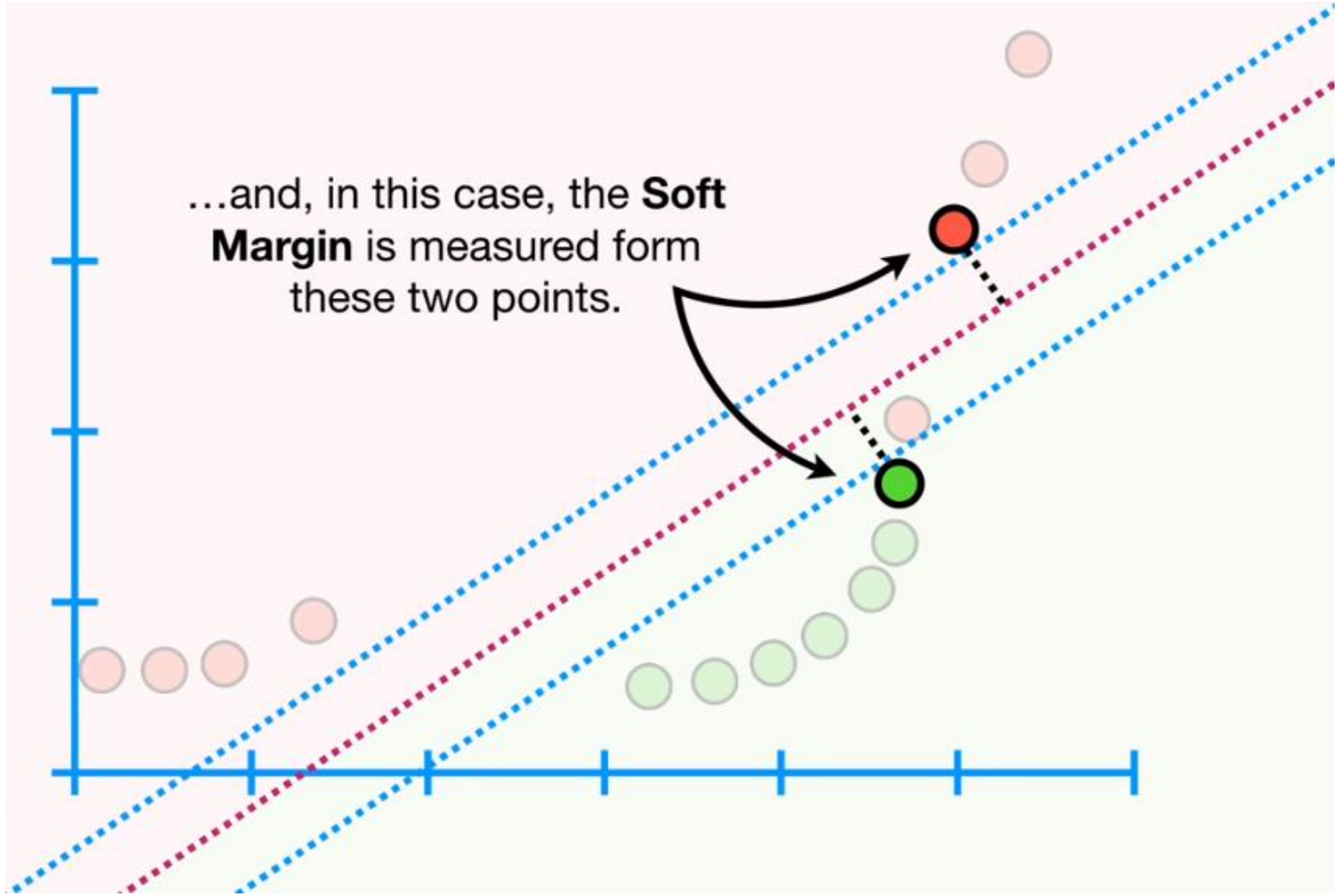
It allows you to dictate the tradeoff between:

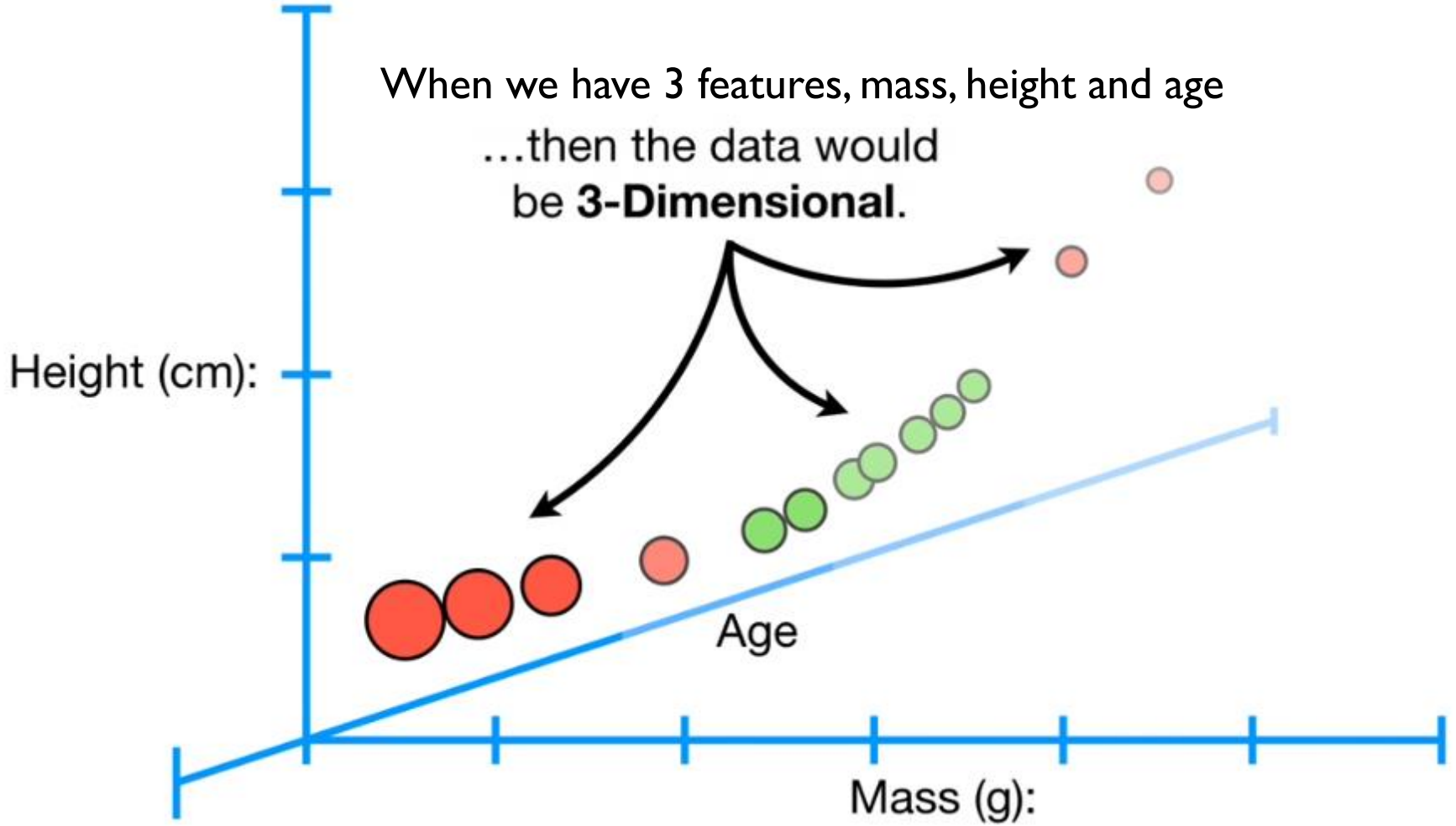
Having a wide margin and
Correctly classifying **training** data.

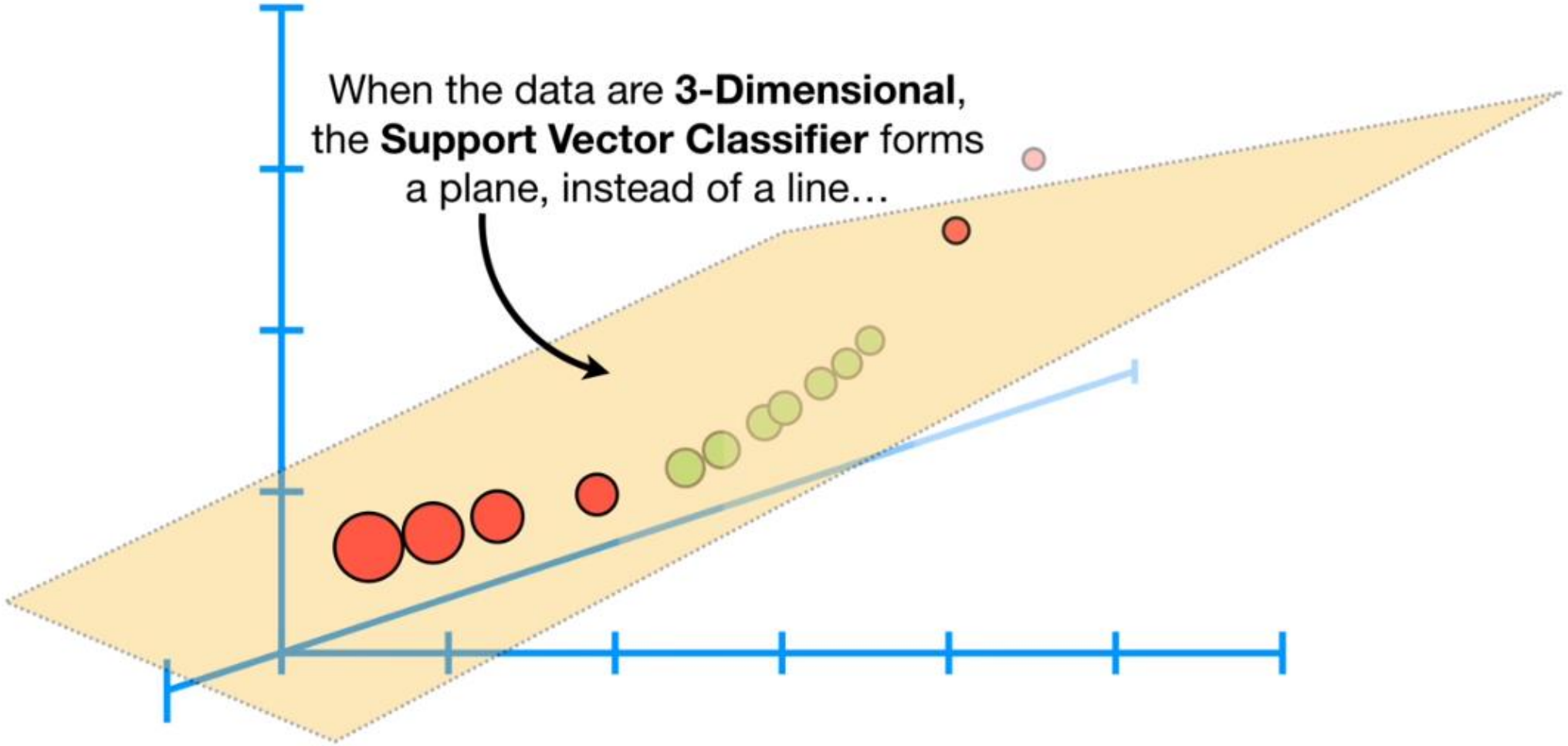



If we have two features, mass and height
…then the data would
be **2-Dimensional**.
Height (cm):
Mass (g):

When the data is 2 dimensional, Support vector classifier is a line



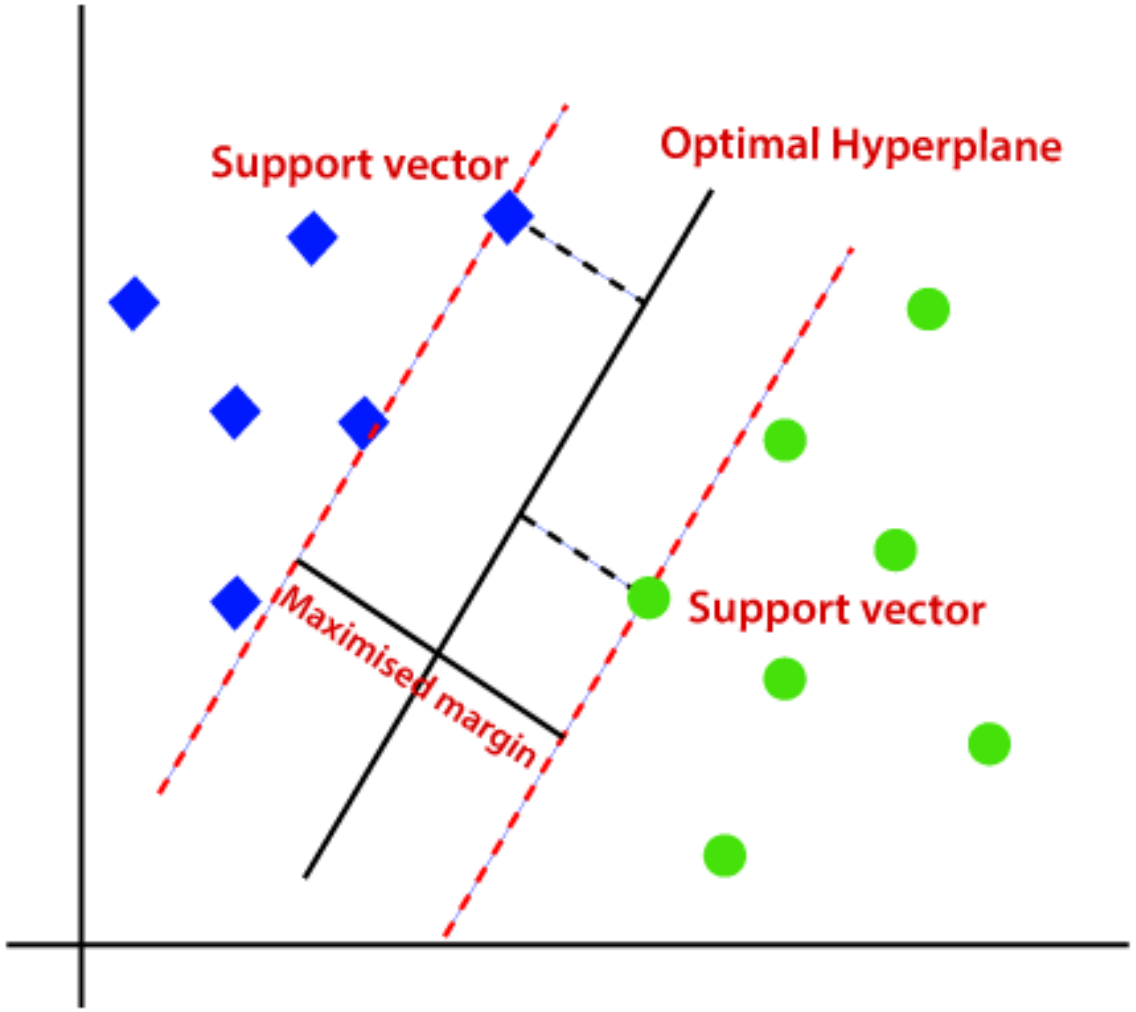

When we have 3 features, mass, height and age
…then the data would
be 3-Dimensional.
Height (cm):
Age
Mass (g):


When the data are **3-Dimensional**,
the **Support Vector Classifier** forms
a plane, instead of a line…

HYPERPLANE

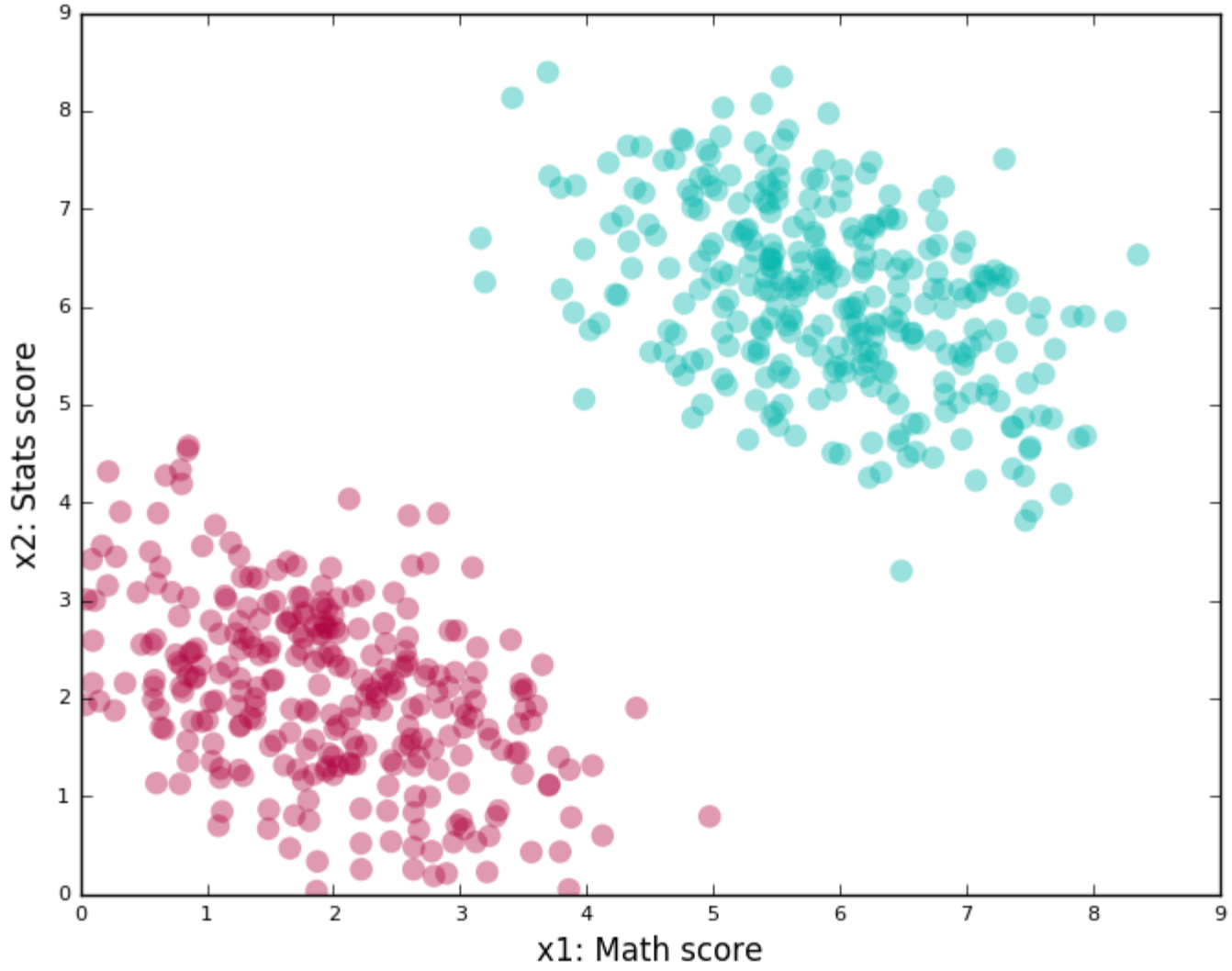
- A hyperplane is a decision boundary that separates data points into different classes in a high-dimensional space.
- In two-dimensional space, a hyperplane is simply a line that separates the data points into two classes.
- In three-dimensional space, a hyperplane is a plane that separates the data points into two classes.
- Similarly, in N-dimensional space, a hyperplane has (N-1)-dimensions.

SVM

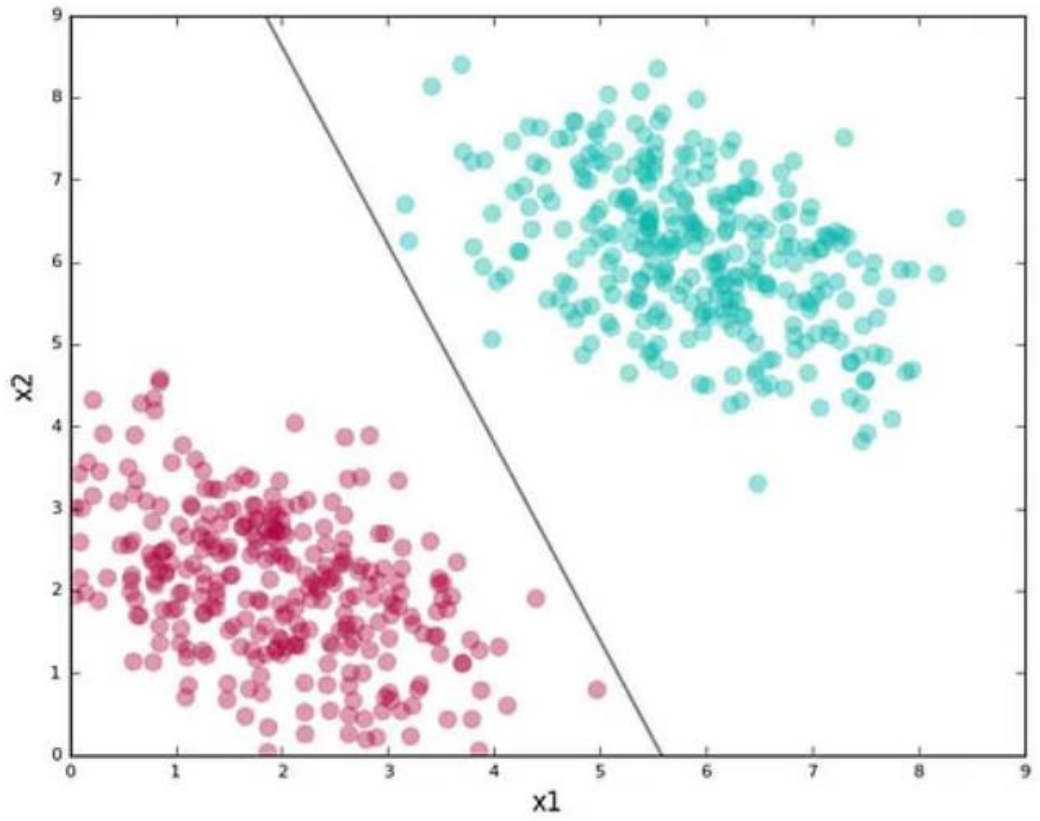

Support vector
Optimal Hyperplane
Maximised margin
Support vector

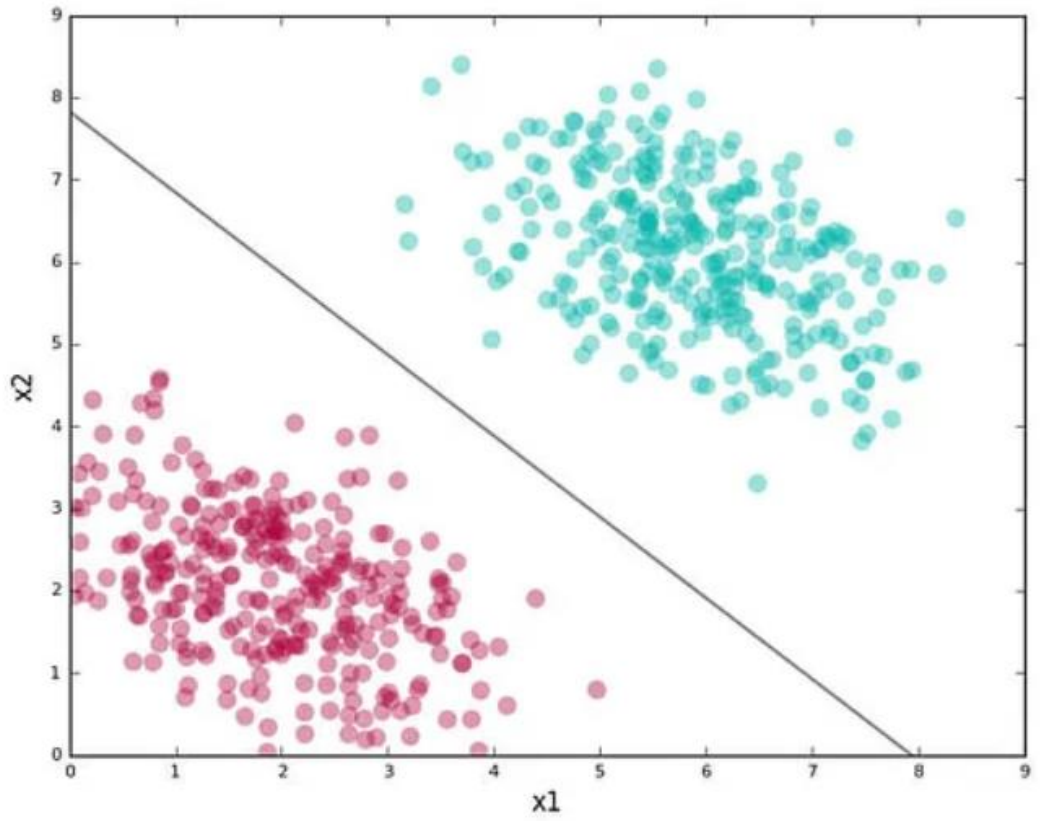
EXAMPLE

- The color of the point green or red represents how students did on the ML course: "Good" or "Bad" respectively.

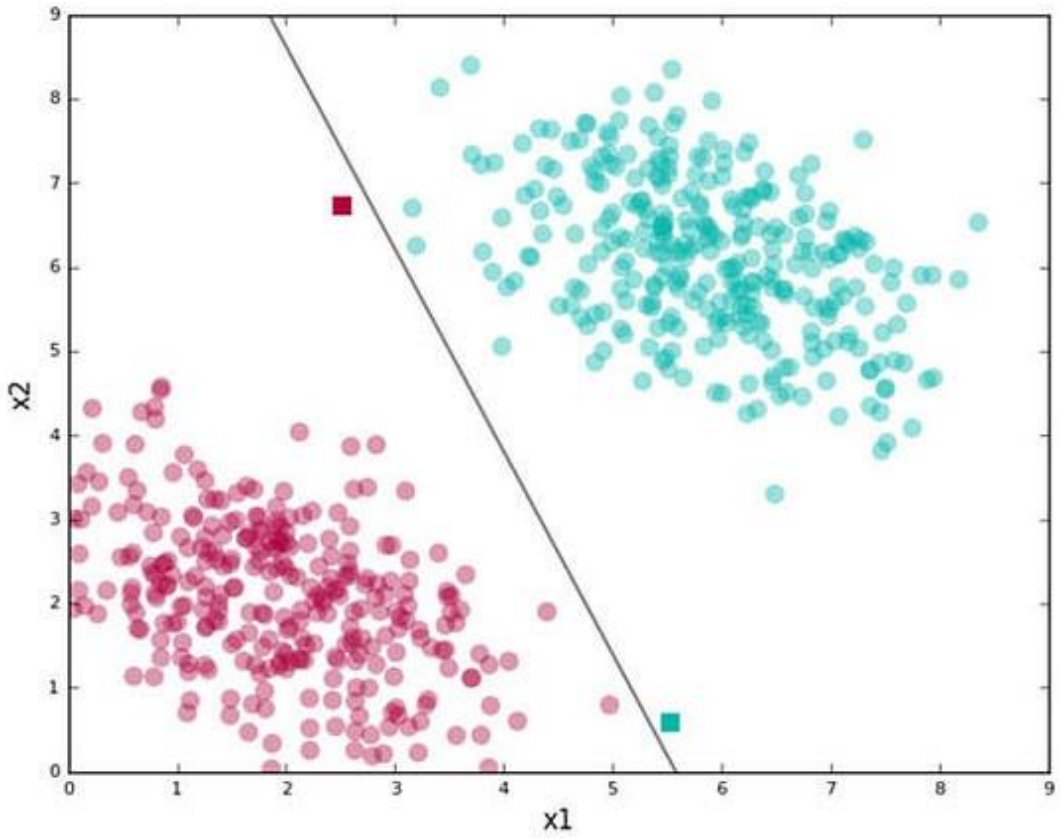


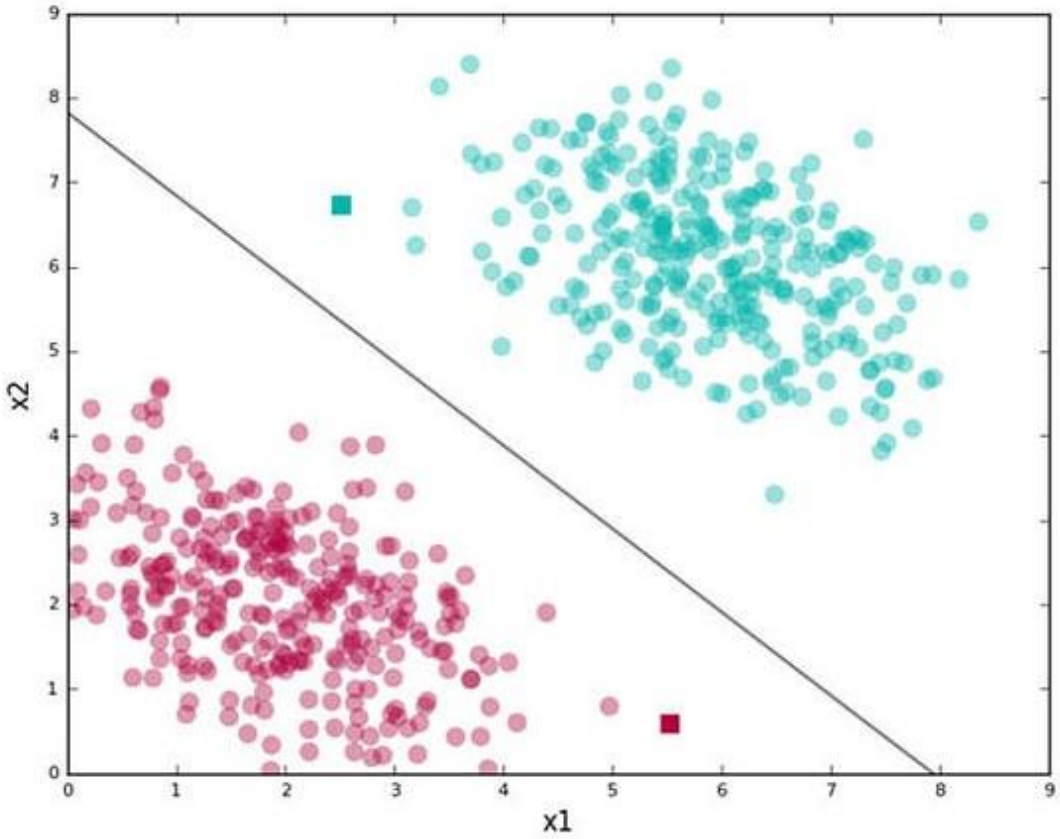
WHICH IS THE BEST LINE?





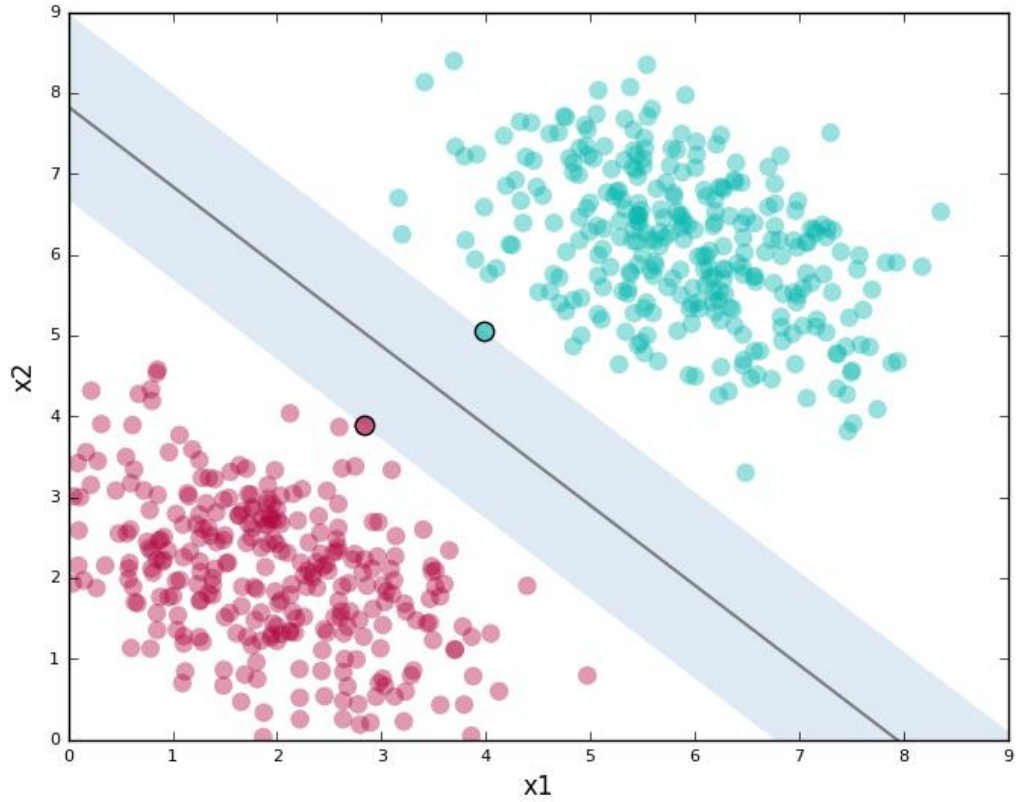
TEST POINTS

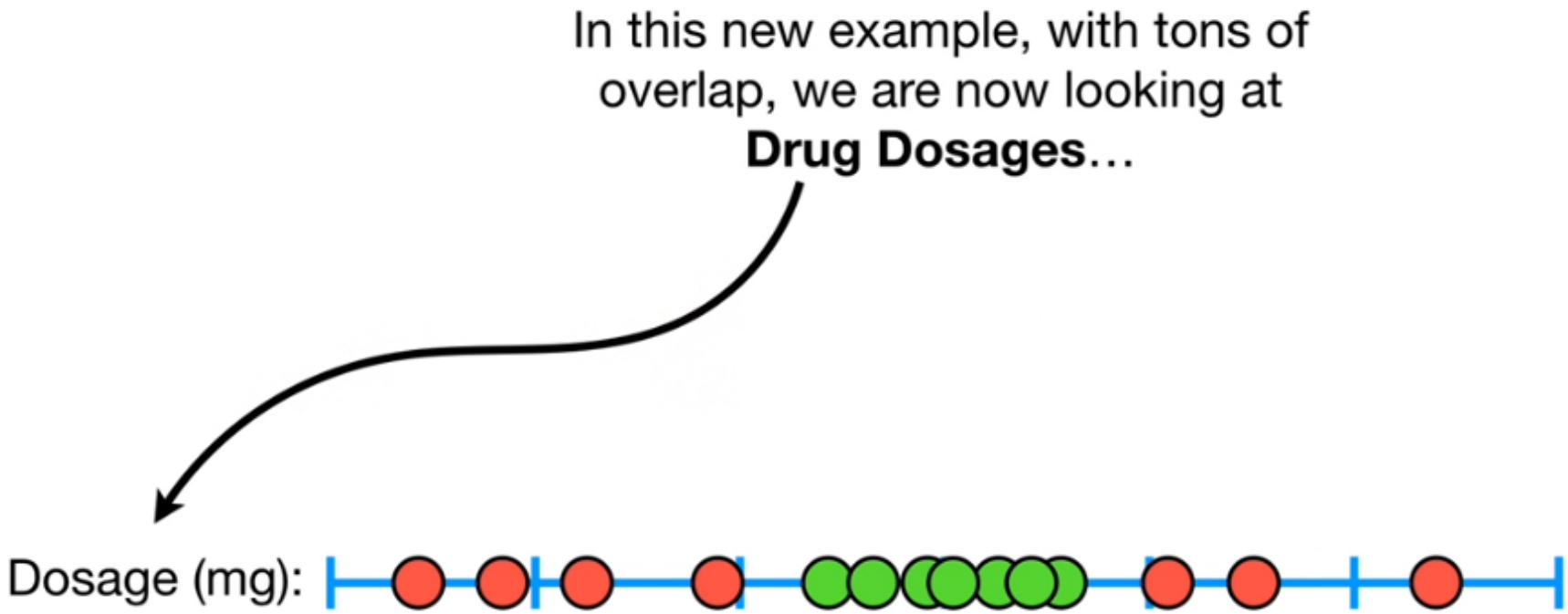




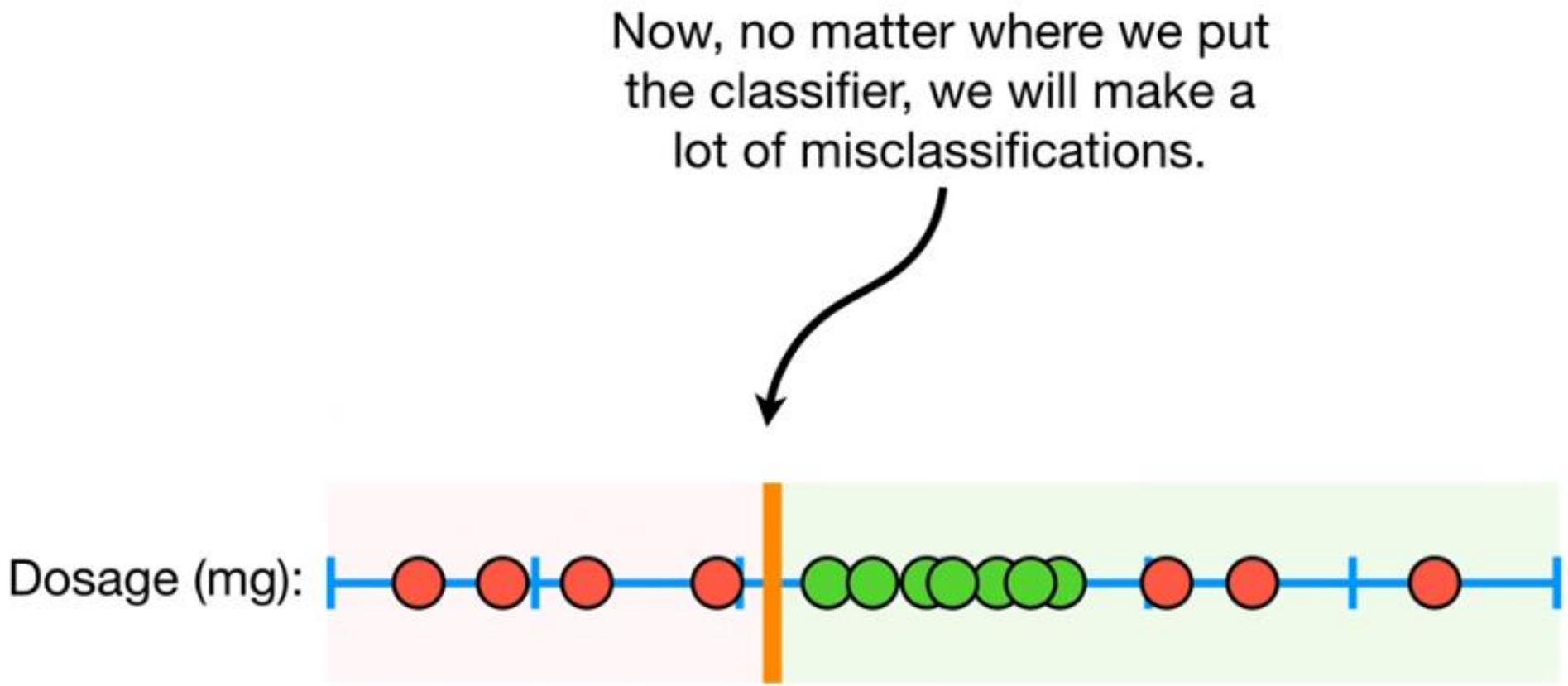
WHAT SVM DO?

- Find lines that correctly classify the training data
- Among all such lines, pick the one that has the greatest distance to the points closest to it.

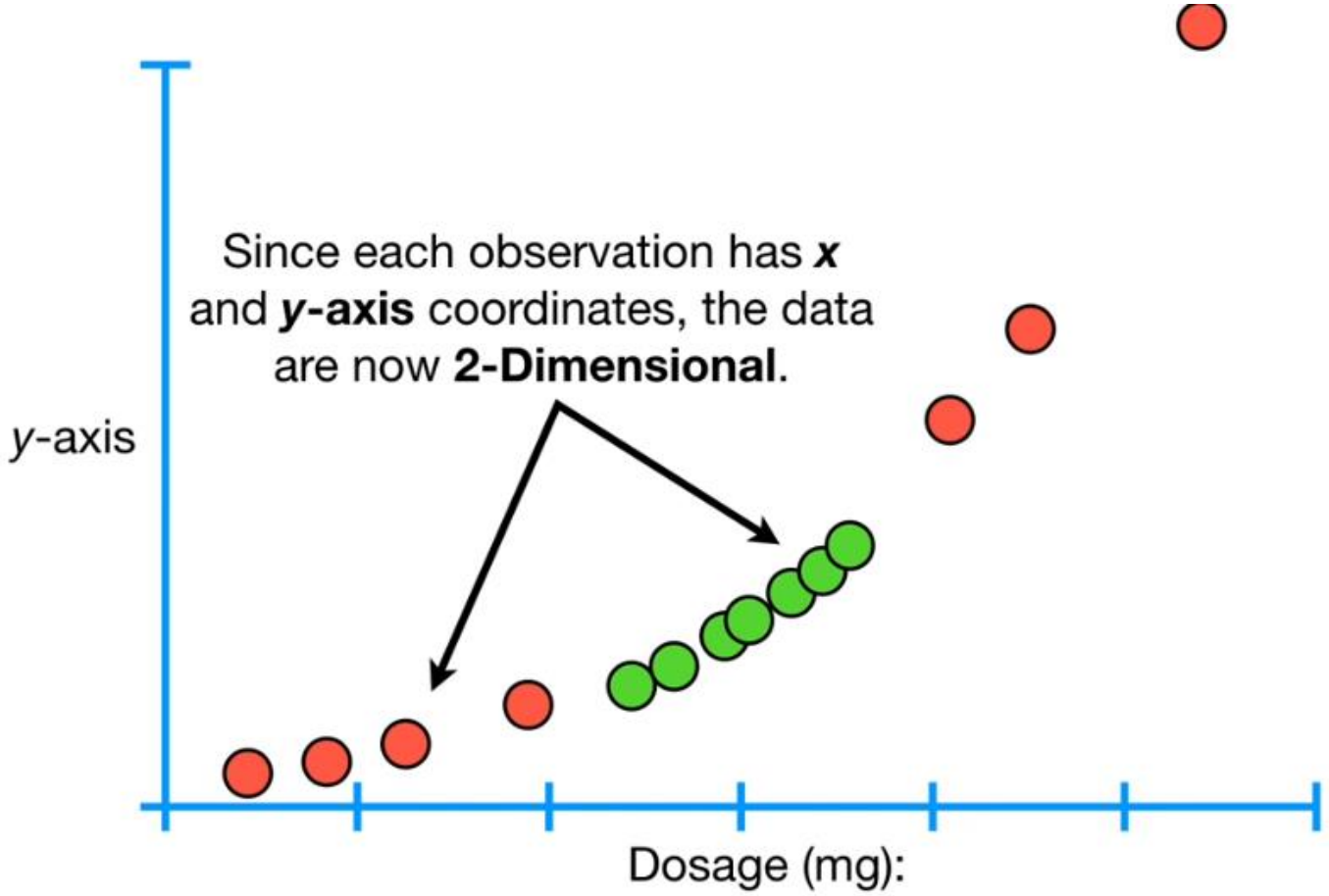


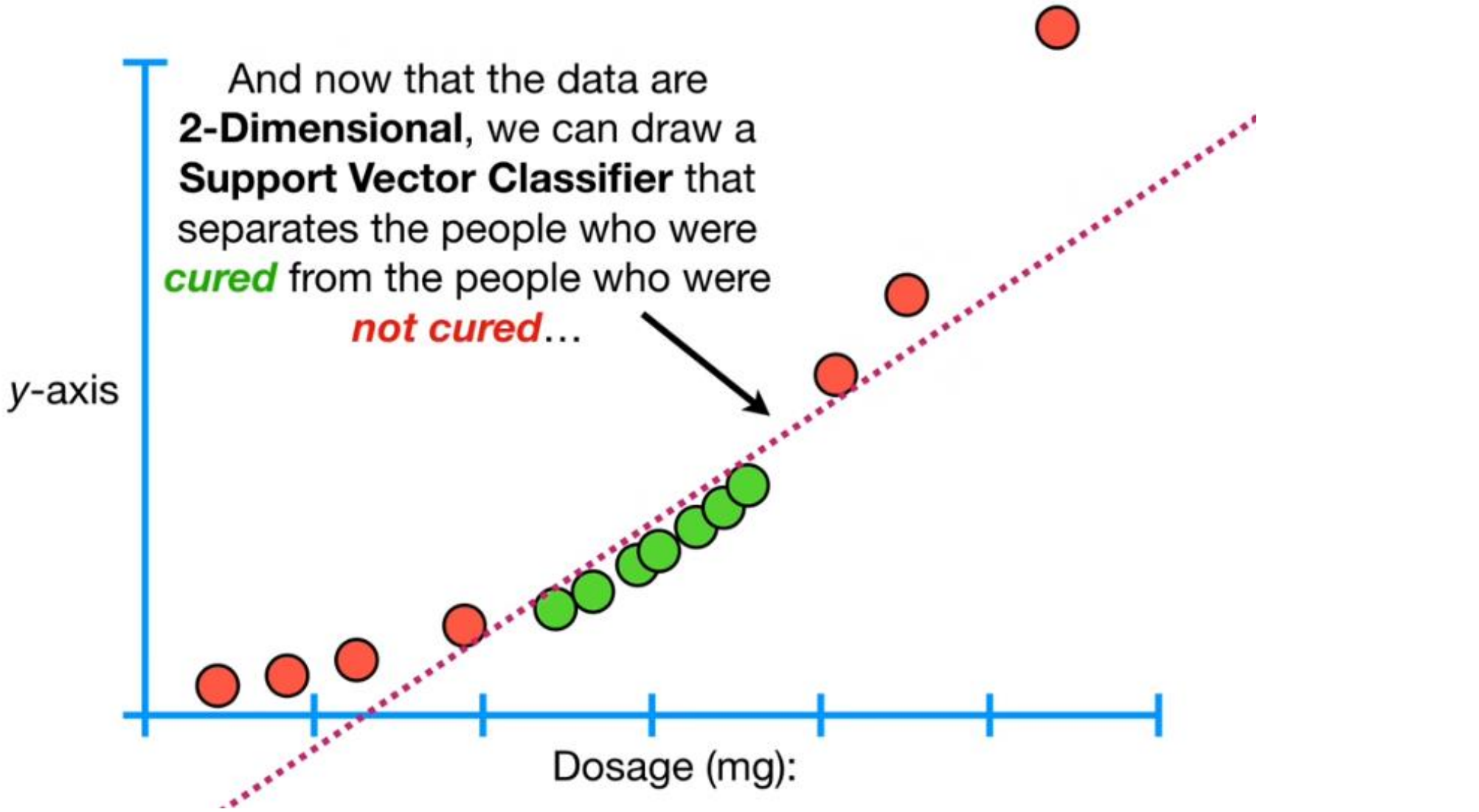


The red dots represents patients that were not cured, and green dots represents patients that are cured.

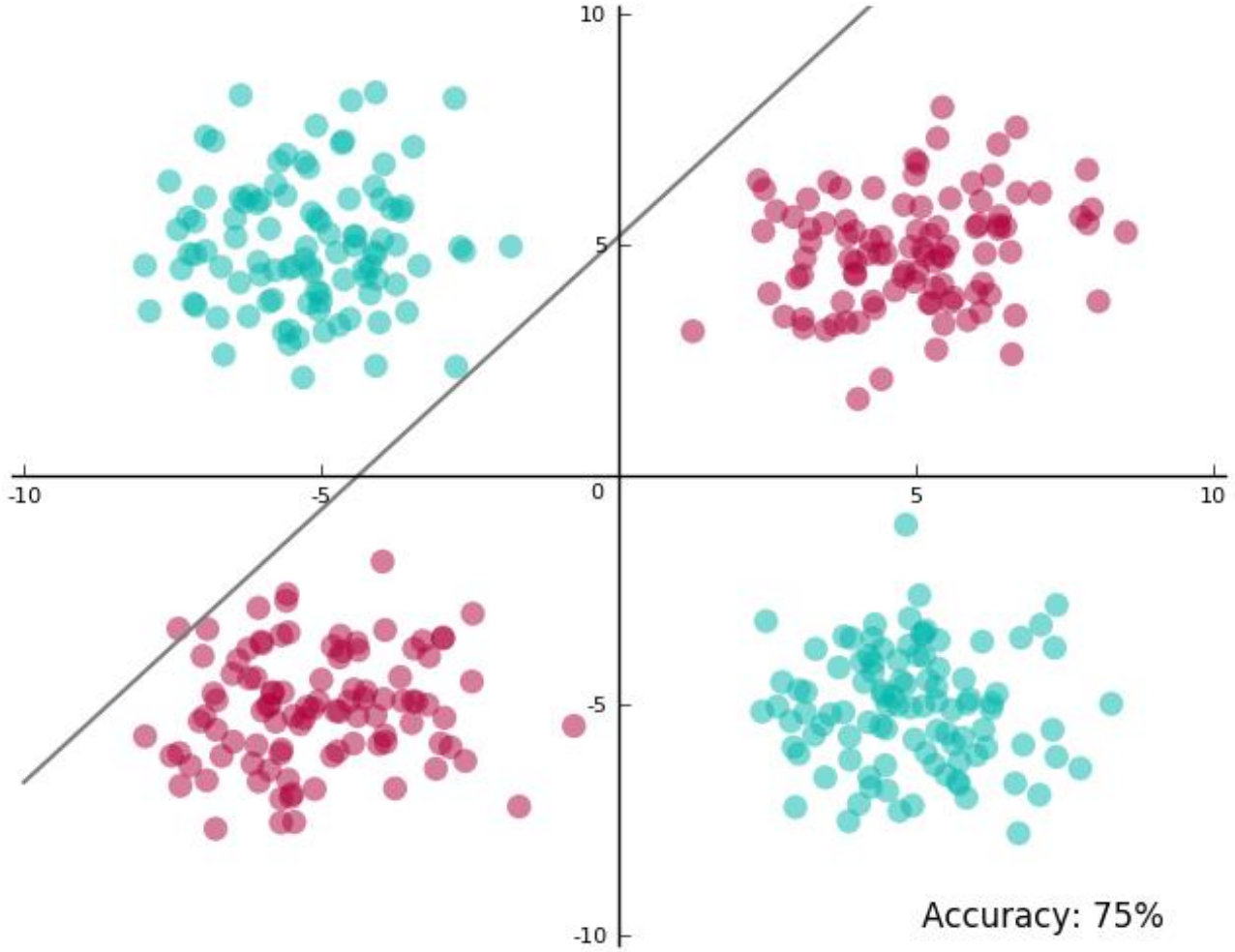

Now, no matter where we put the classifier, we will make a lot of misclassifications.
Dosage (mg):

We have transformed our data from 1d to 2D, by adding one feature, Y-axis here represents the 2nd feature, that is $Dosage^2$.



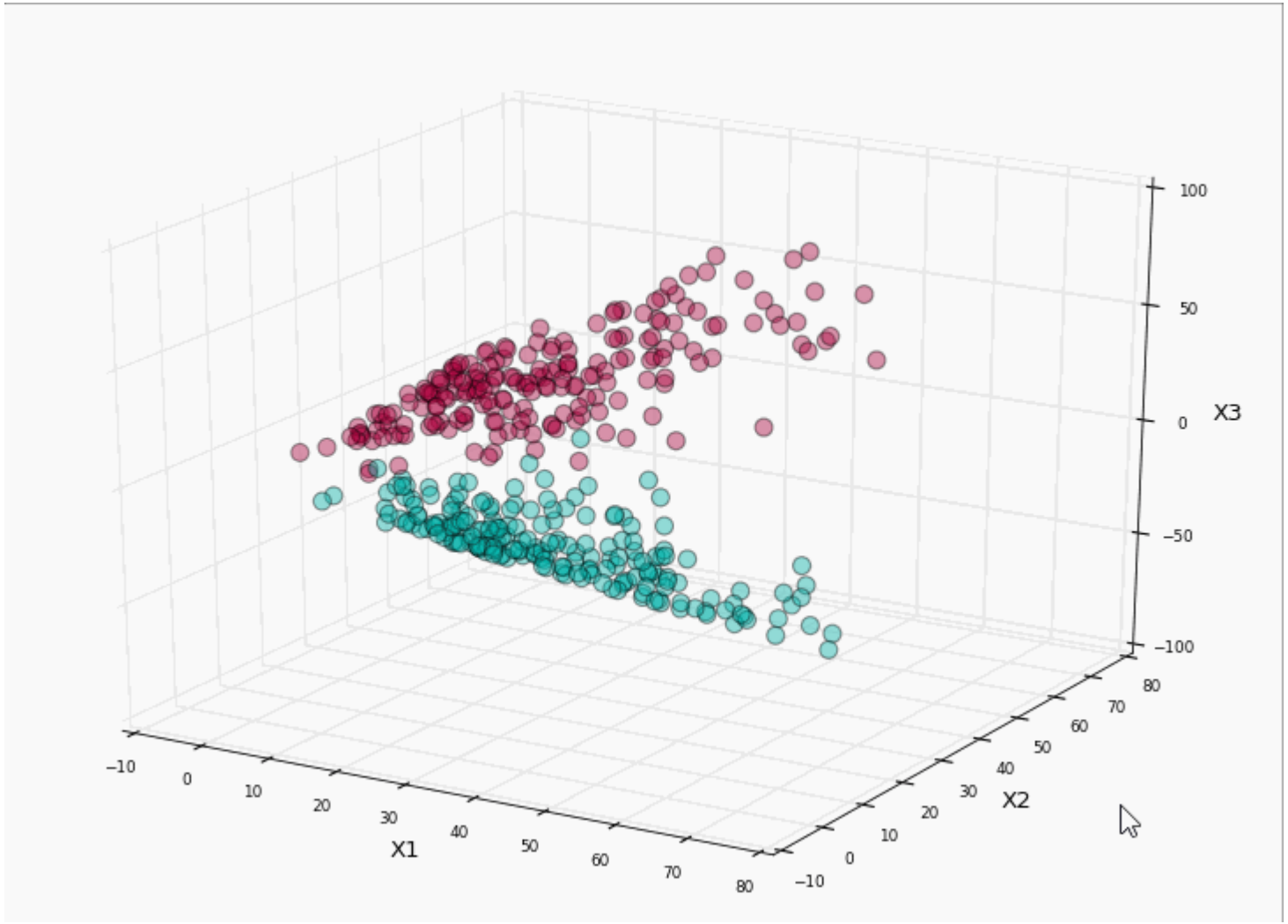

And now that the data are
2-Dimensional, we can draw a
Support Vector Classifier that
separates the people who were
cured from the people who were
not cured…
y-axis
Dosage (mg):

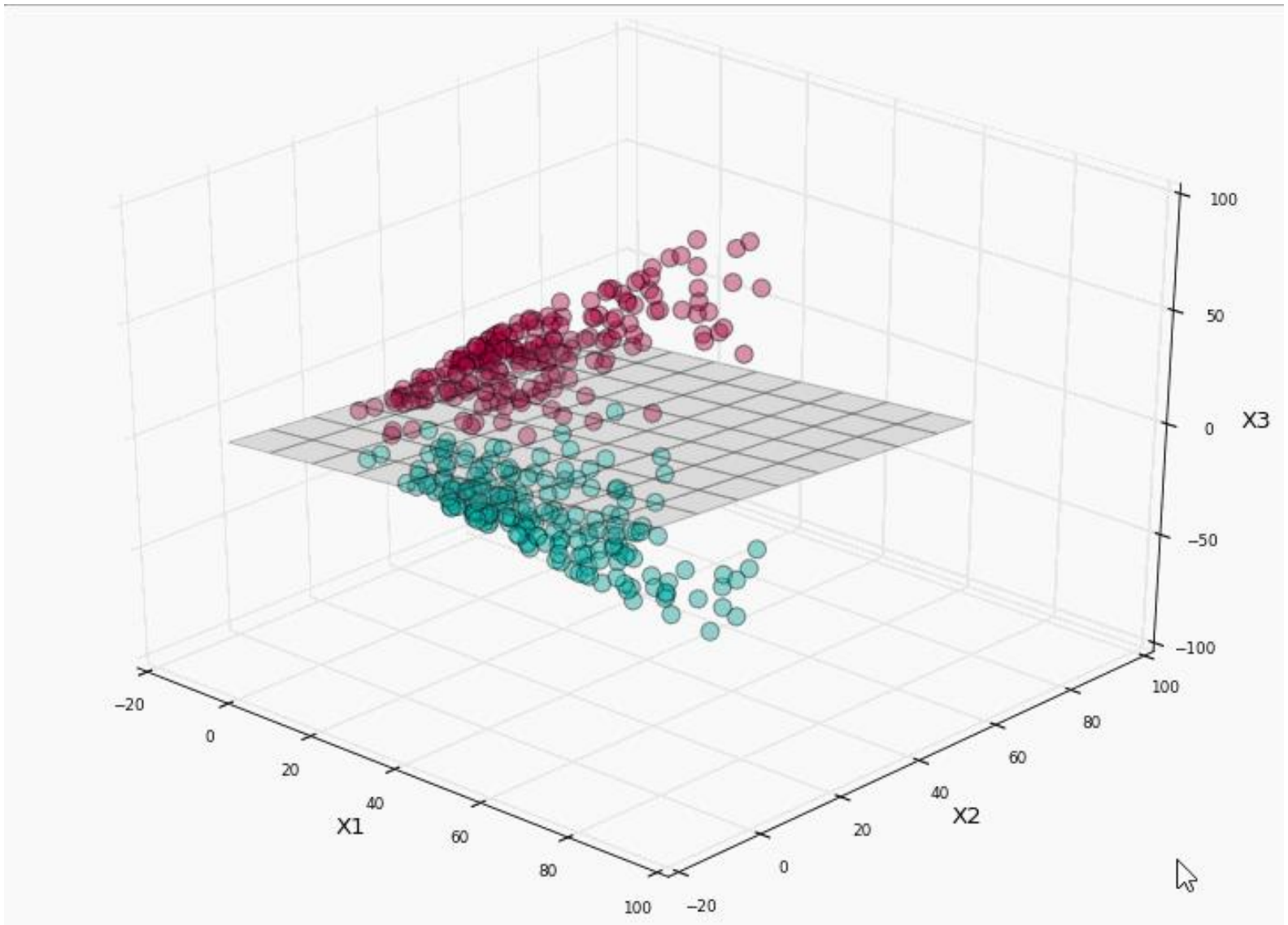
EXAMPLE



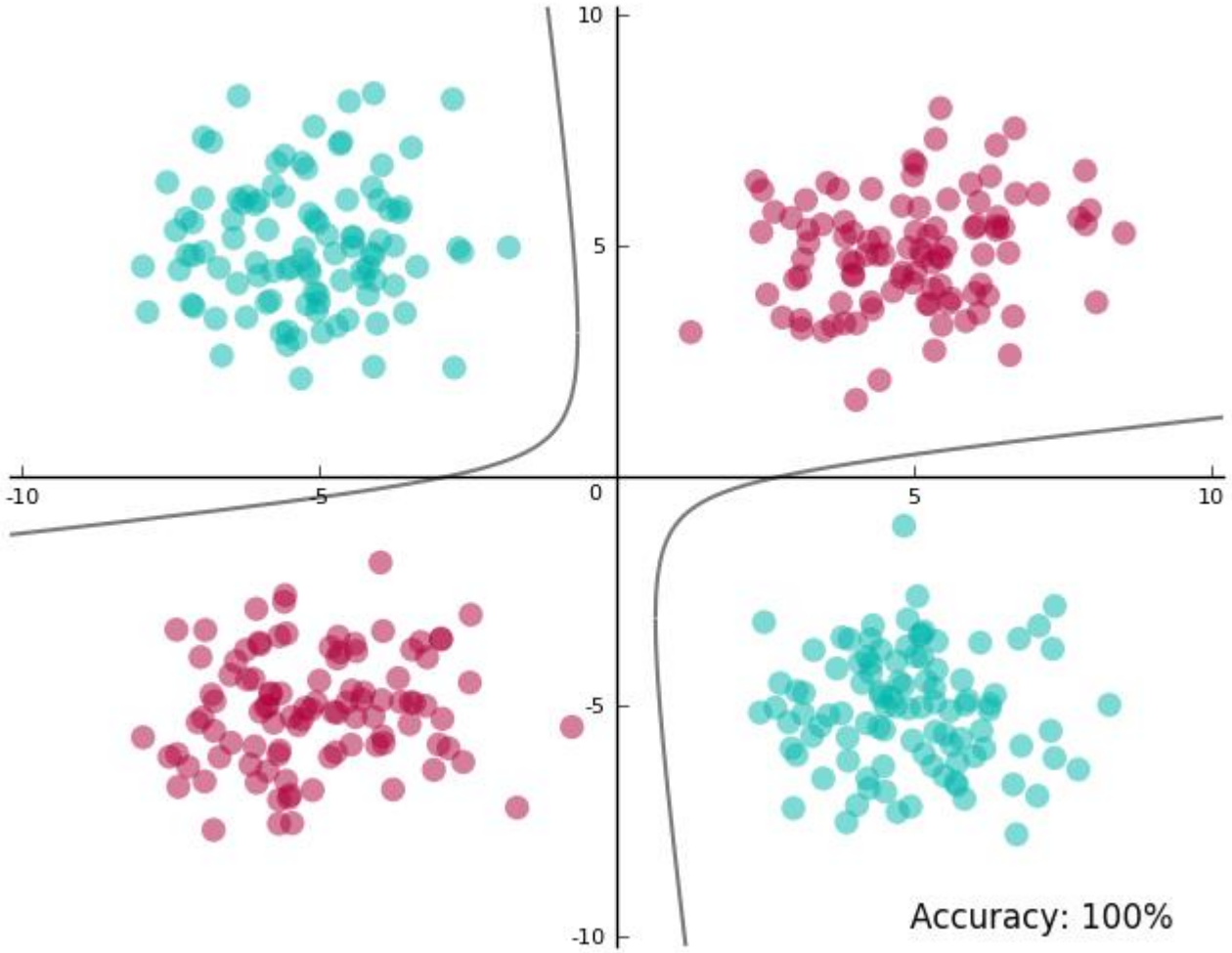
PROJECT TO HIGH DIMENSION

$$X_1 = x_1^2$$
$$X_2 = x_2^2$$
$$X_3 = \sqrt{2}x_1x_2$$




100
50
0
−50
−100
X3
−20
0
20
40
60
80
100
X1
−20
0
20
40
60
80
100
X2

ORIGINAL 2D SPACE



KERNEL

- In order to make mathematics possible, SVM use something called kernel functions to systematically find the support vector classifier (decision boundary) in higher dimensions.
  - Polynomial Kernel
  - Radial Basis Function Kernel(rbf)

THE KERNEL TRICK

- Kernel function only calculate relationships between every pair of points as if they are in the higher dimensions; they don't actually do the transformation.

- This trick, calculating higher dimensional relationships without transforming the data to the higher dimension is called the kernel trick.