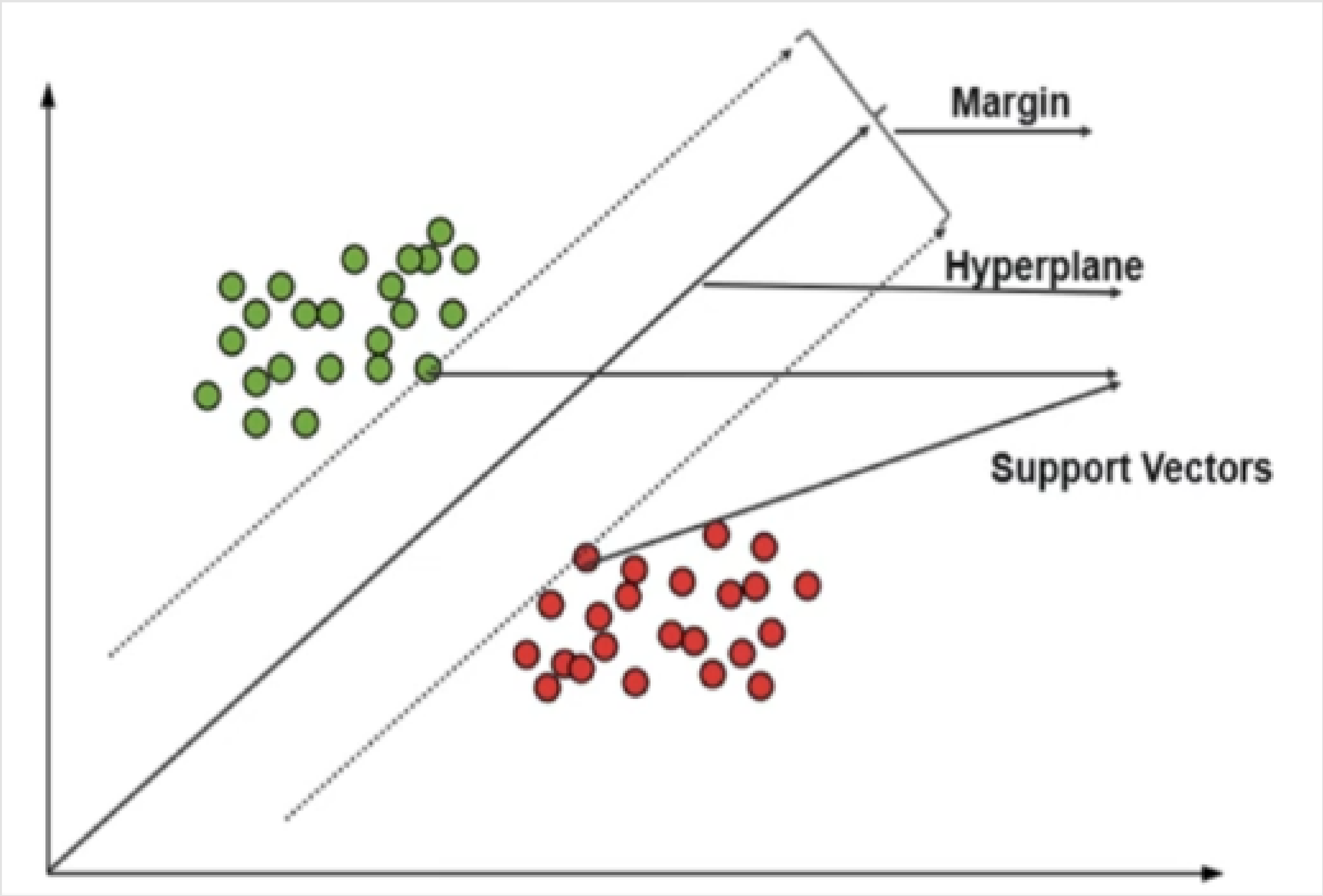
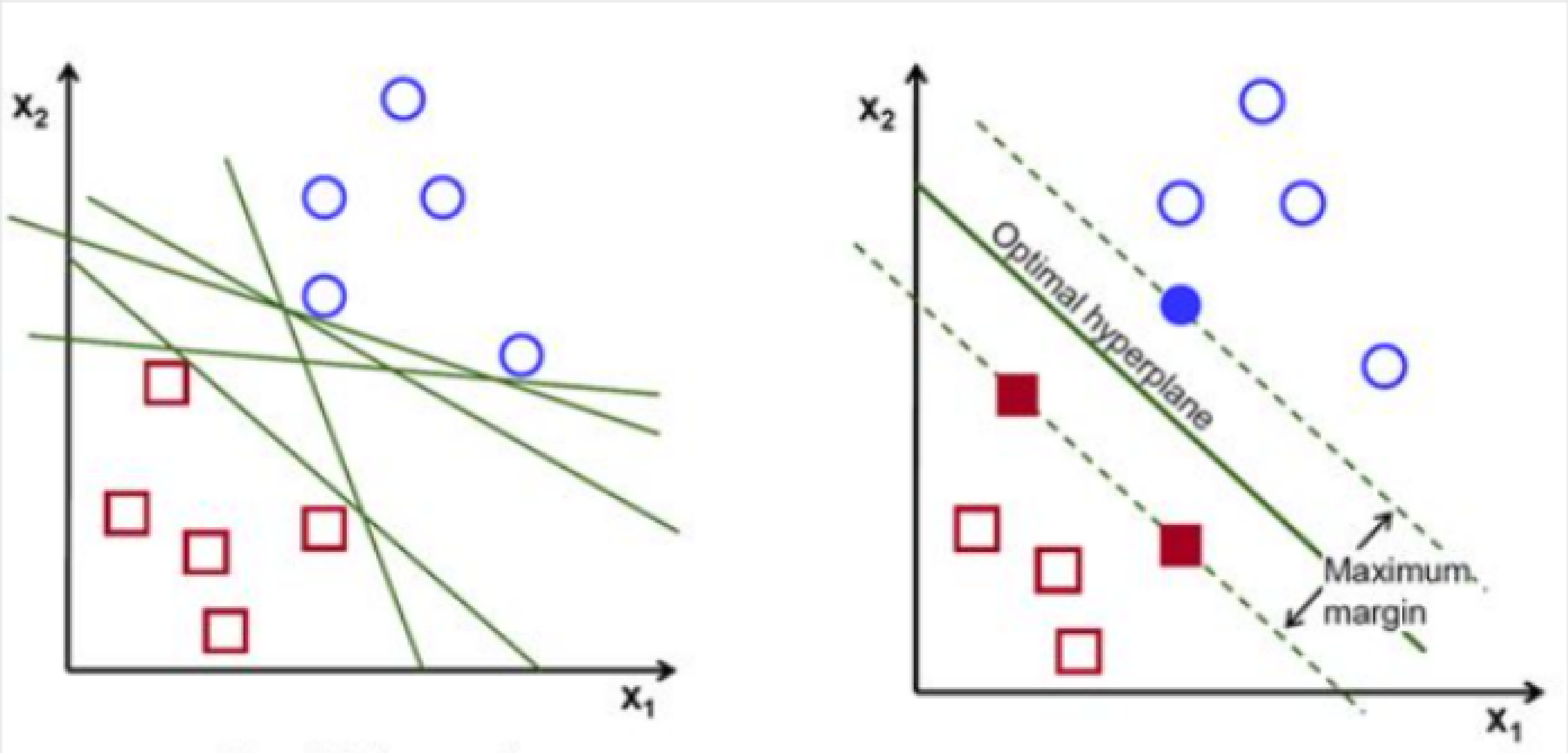
Support Vector Machine

The Support Vector Machines(SVM) aim to find the best hyperplane (also called decision boundary) that best separates (splits) a dataset into two classes/groups (binary classification problem) with the intent to maximize the margin (Maximum Margin Classifier)

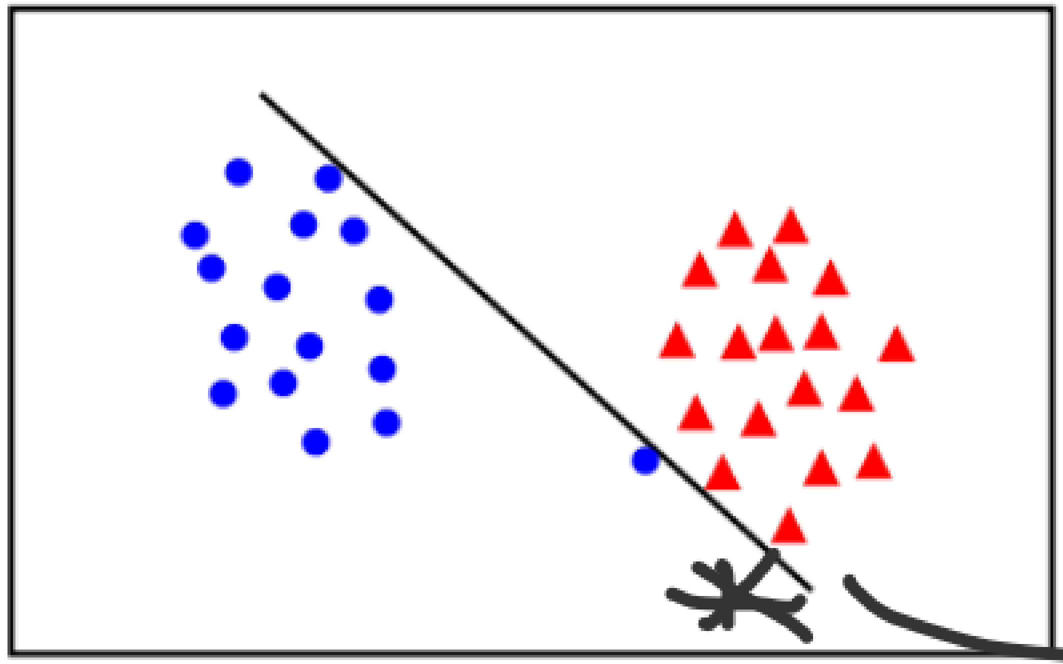
Support vectors are just the samples (data-points) that are located nearest to the separating hyperplane. These samples would alter the position of the separating hyperplane, in the event of their removal. Thus, these are the most important samples that define the location and orientation of best decision boundary.



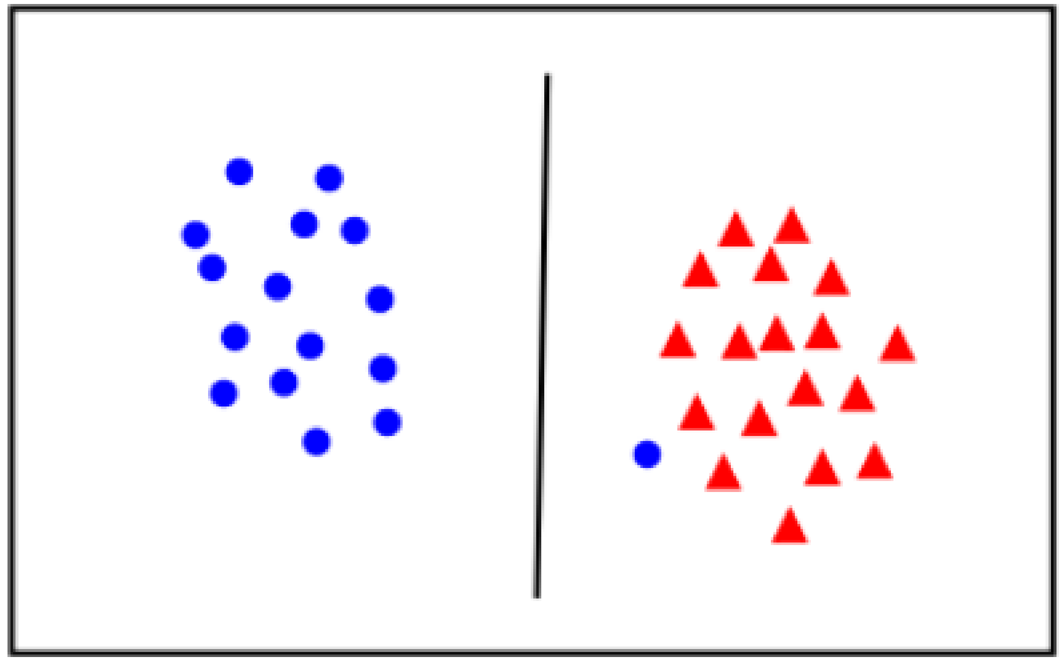


Possible hyperplanes

Linear separability again: What is the best w?



• the points can be linearly separated but there is a very narrow margin



• but possibly the large margin solution is better, even though one constraint is violated

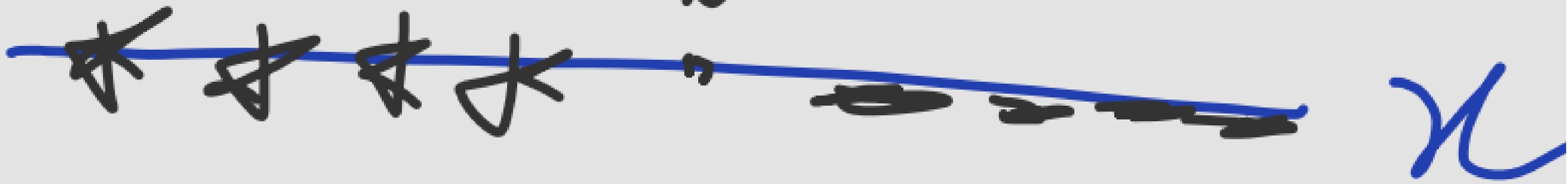
?

1. Aims to maximize the margin w.r.t. Hyperplane. Can be many hyperplanes but choose one that maximizes this margin.
2. Support vectors are orthogonal vectors from the points close to decision boundary as they adversely affect /impact the decision.

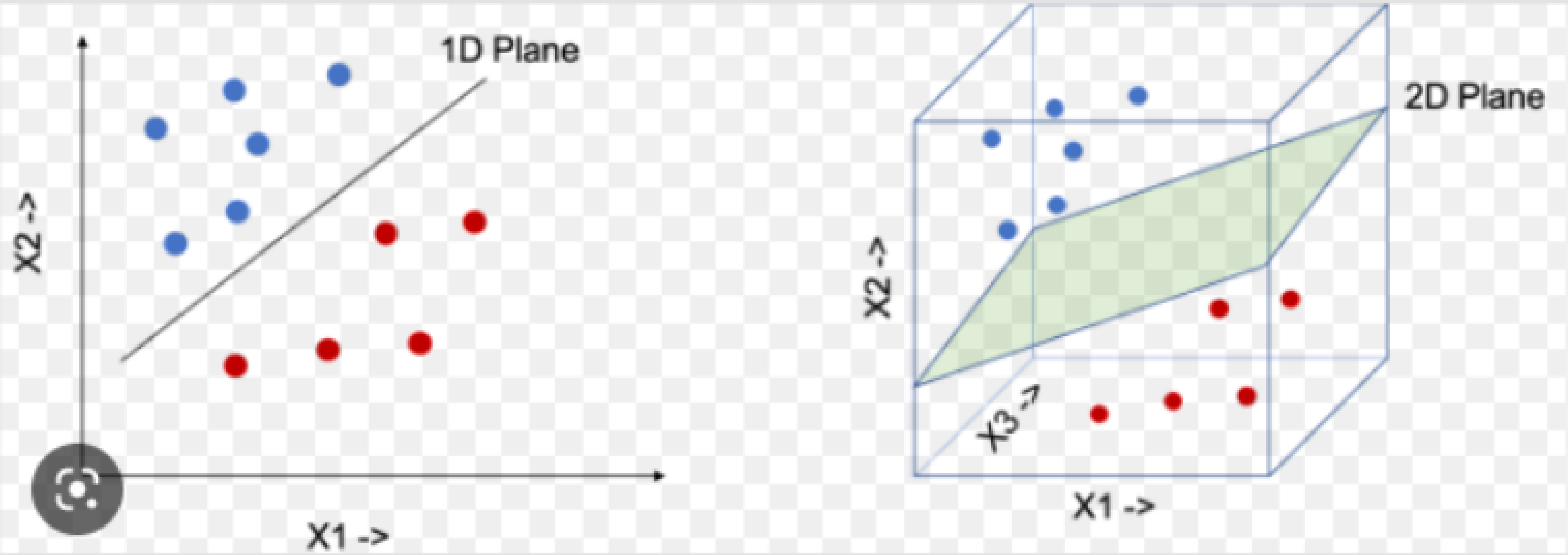
Margin: Perpendicular distance of closest points (support vectors) from the decision boundary.

Ques What will be the hyperplane for Print

① 1D data



② 2D data

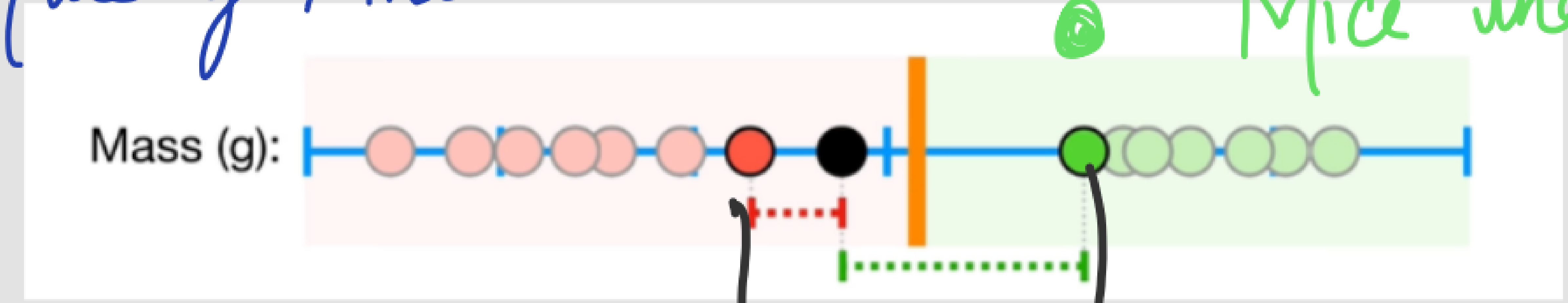


③ 3D data

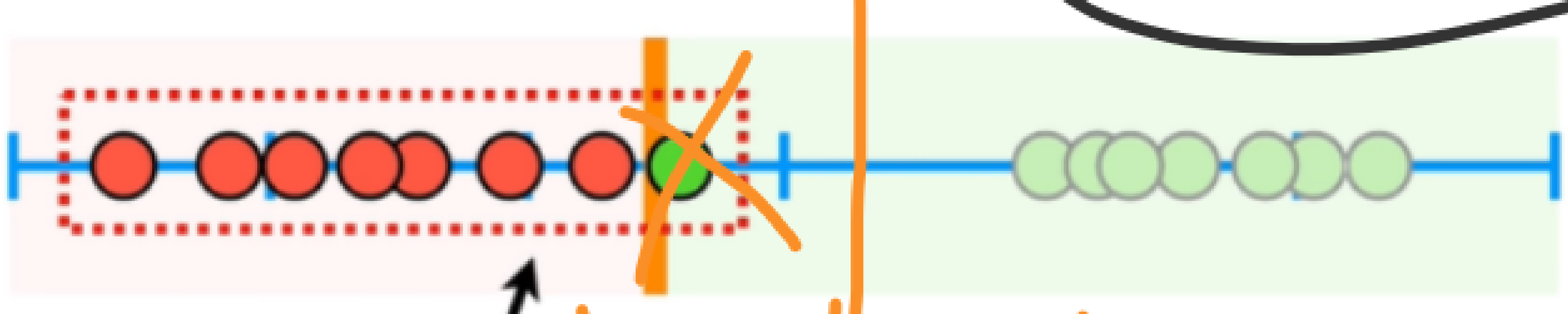
Mass of Mice

Mice that are not obese

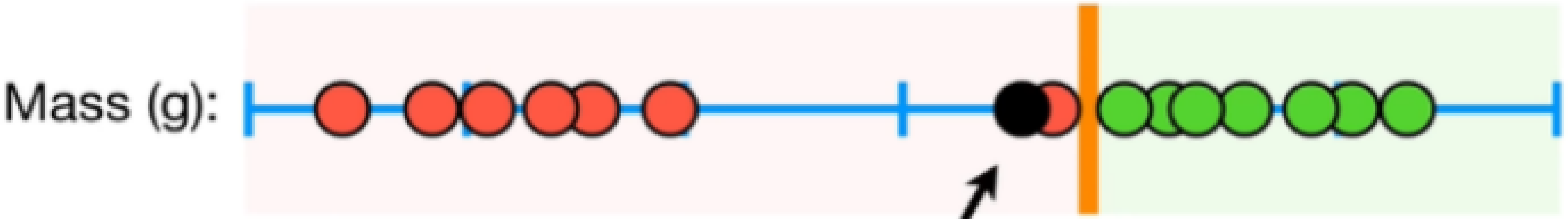
Mice that are obese



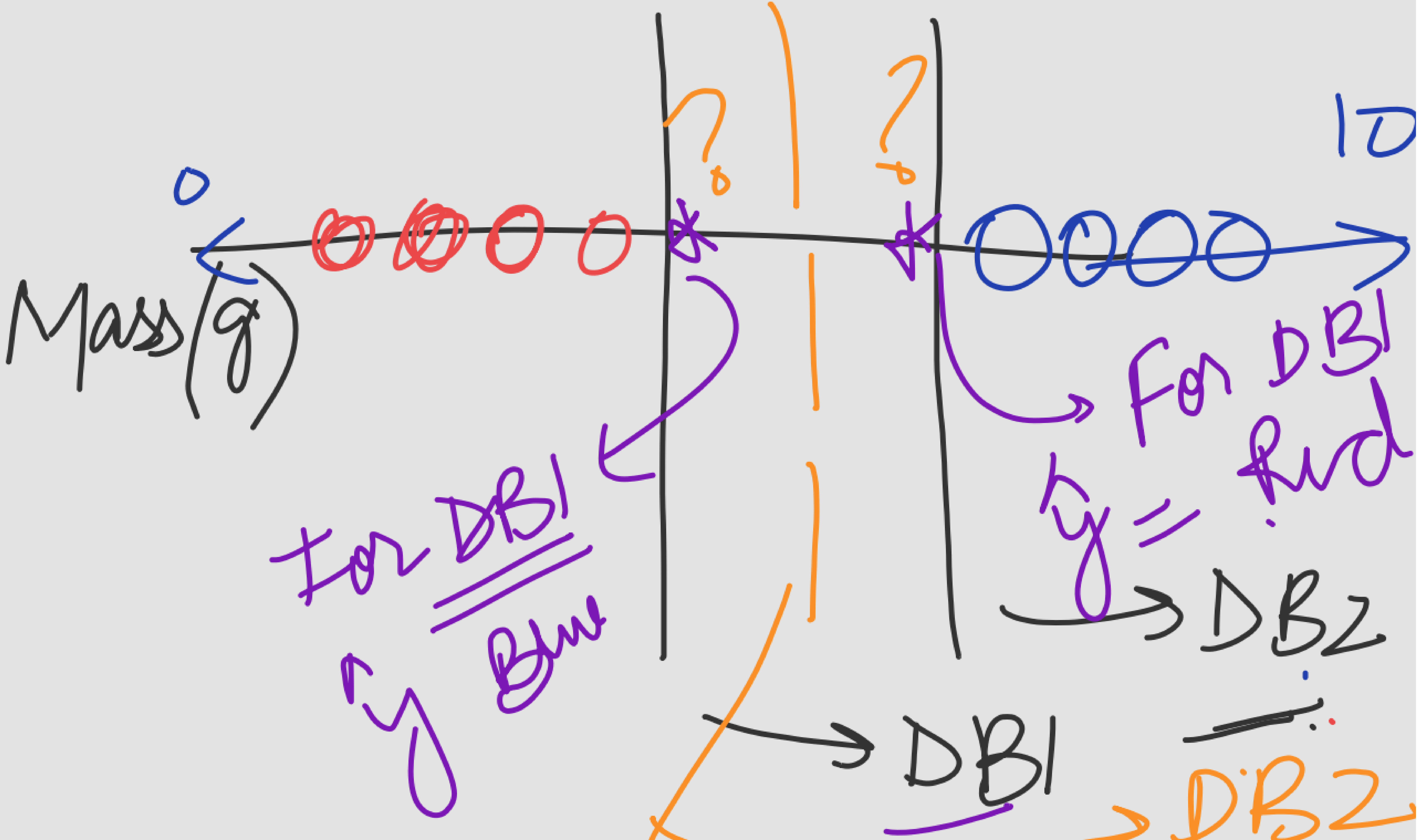
4 6 → 4.7/5.1



In this case threshold is not a good estimator.
The observation is above threshold, but closer to
the non obese mice

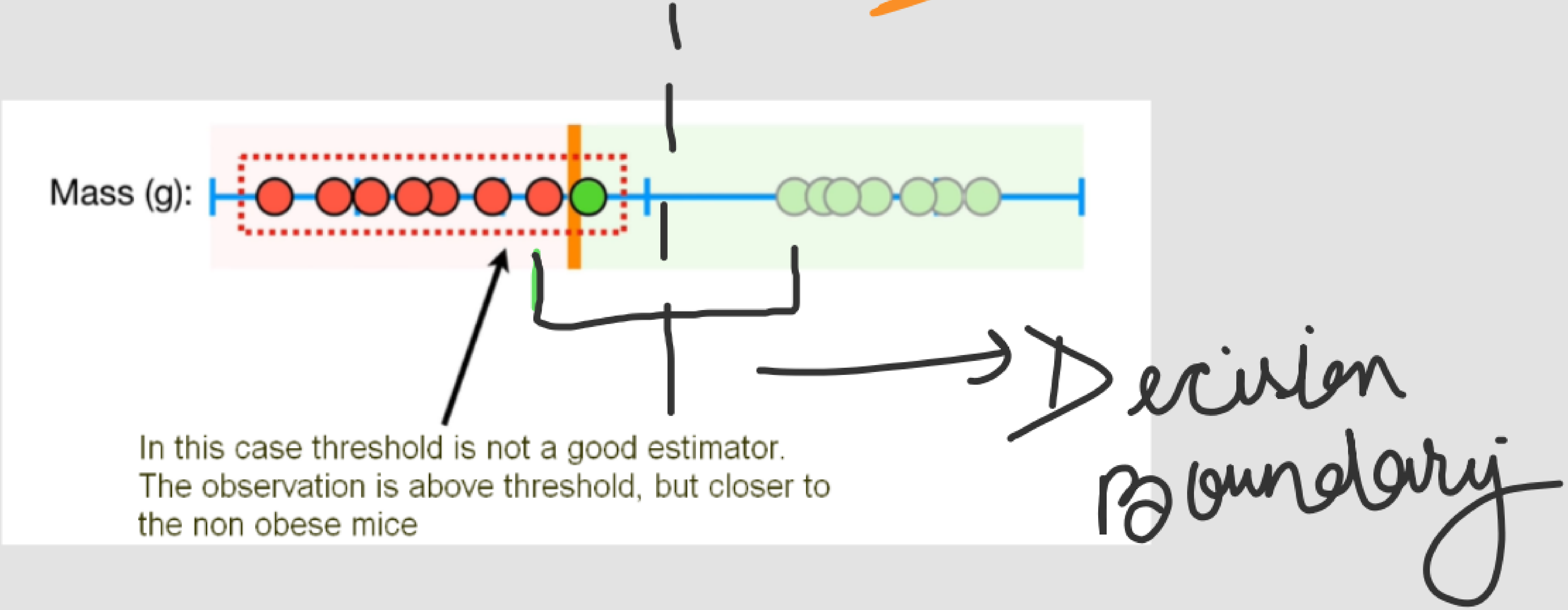


The new observation is wrongly classified
as not obese because the presence of an
outlier

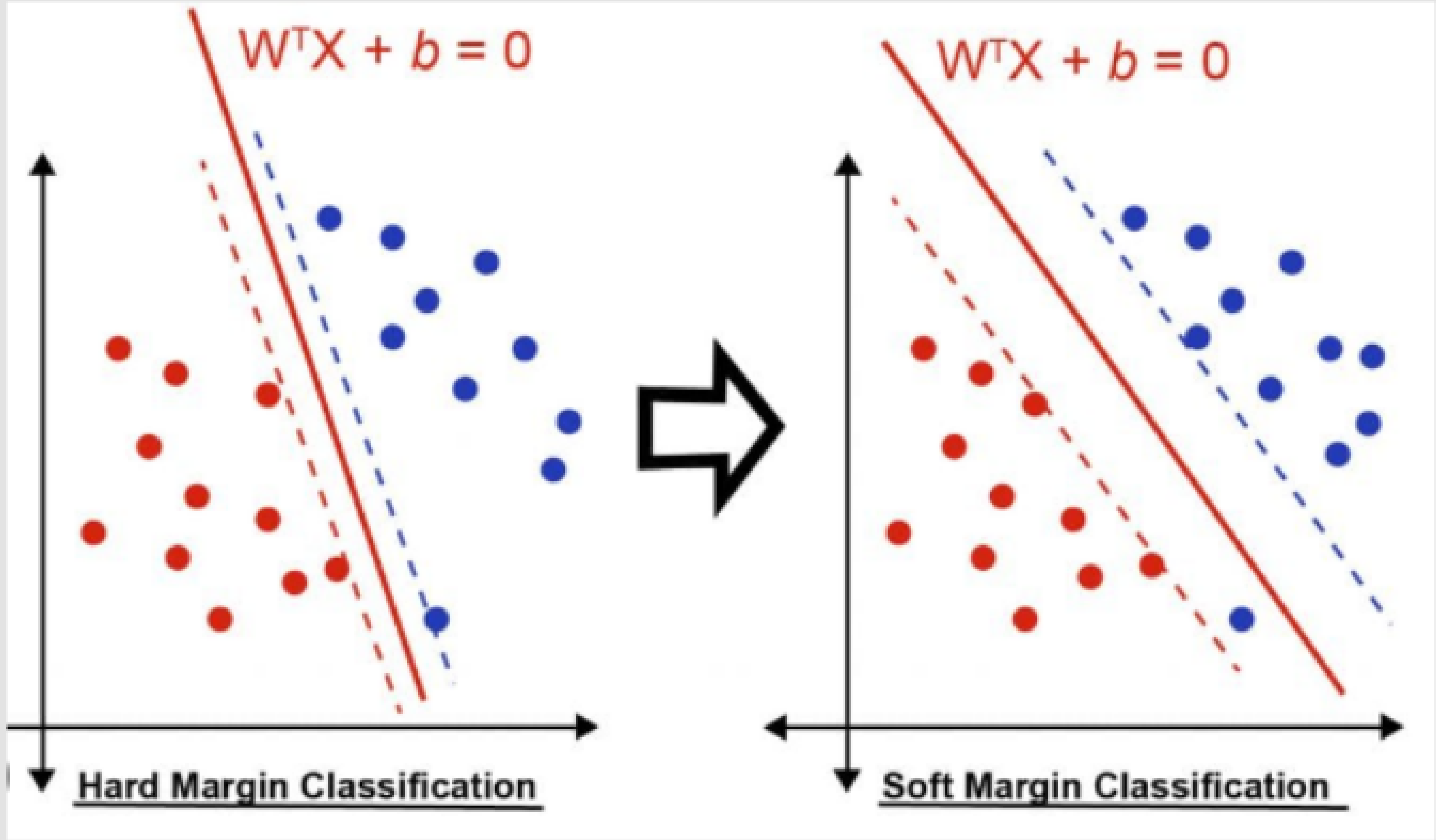

0
10
Mass (g)
?
?
For DB1
y = Blue
For DB1
y = Red
DB2
DB1
DB2

To solve this problem, we can focus on the observations on the edges of each cluster and use the midpoint between them as the threshold called Maximal Margin Classifier.

When we allow misclassification, the distance between the observations and the threshold/decision boundary is called a Soft Margin.



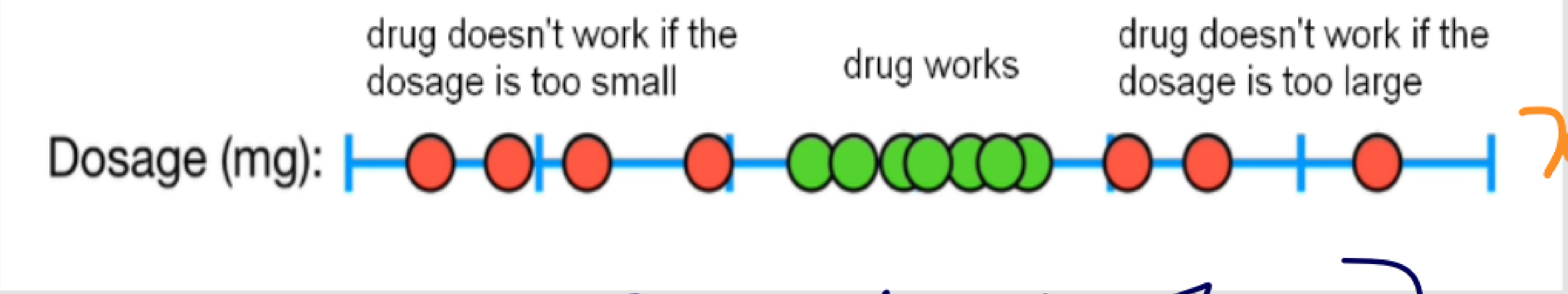
Hard Margin Vs Soft Margin



Soft Margin: try to find a line to separate, but tolerate one or few misclassified dots (e.g. the dots circled in red)

What happens when there is no clear separating hyperplane (kernel SVM) ?

Data not linearly separable !



x

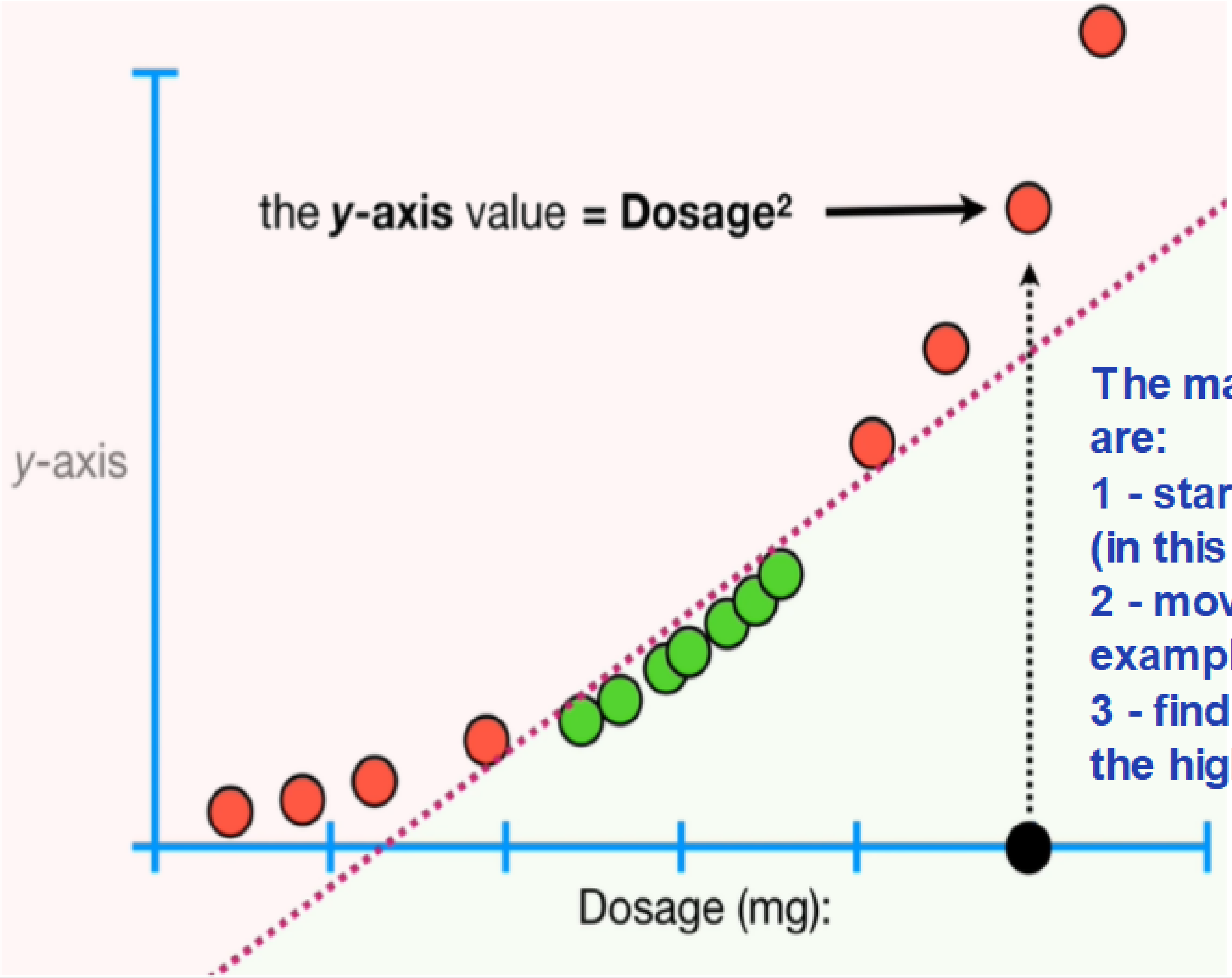
Red Dots : People who are not cured
Green Dots : People who are cured.

x < Drug will work
> E
Not

Support Vector Classifiers don't perform well with this type of data. The solution is to use the Support Vector Machines!!

If the hyperplane classifies the dataset linearly then the algorithm we call it as SVC and the algorithm that separates the dataset by non-linear approach then we call it as SVM

We use the x-axis which represent the dosages we observed, but we also add an y-axis that will be the **square of the dosages**.

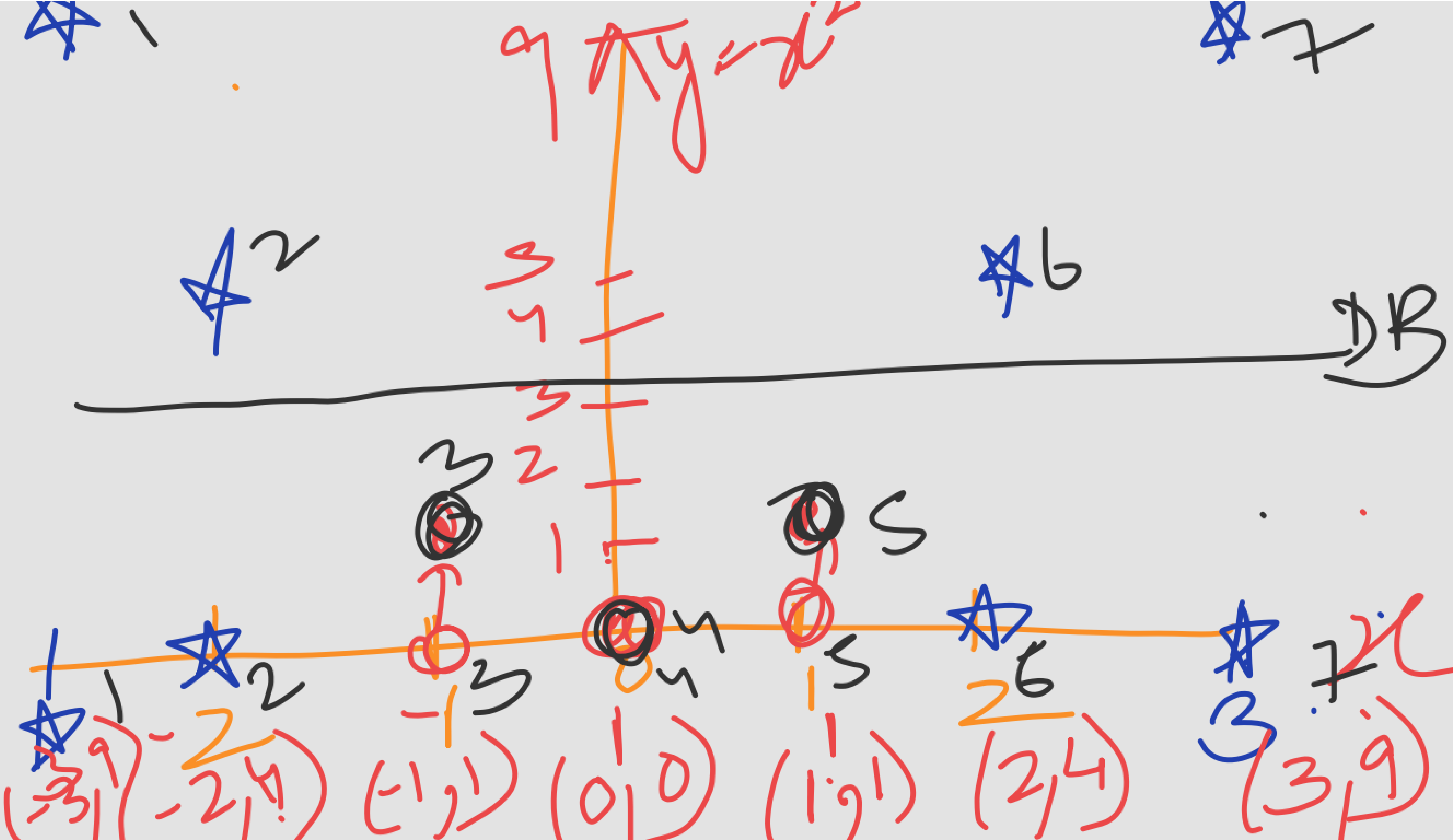


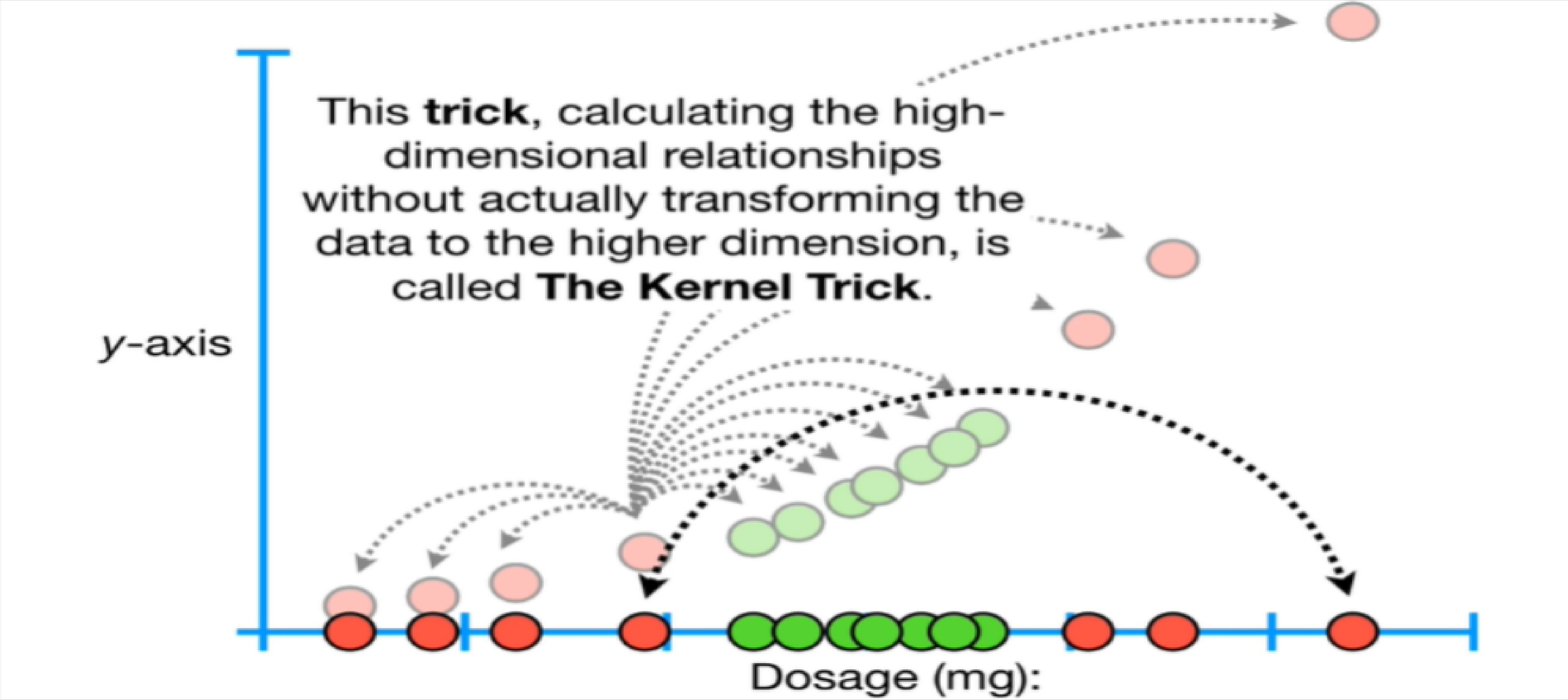
The main idea behind Support Vector Machines are:

1 - start with data in a relatively low dimension (in this example one dimension dosage in mg)

2 - move the data into a higher dimension (in this example from one to two dimensions)

3 - find a Support Vector Classifier that separates the higher dimensional data into two groups

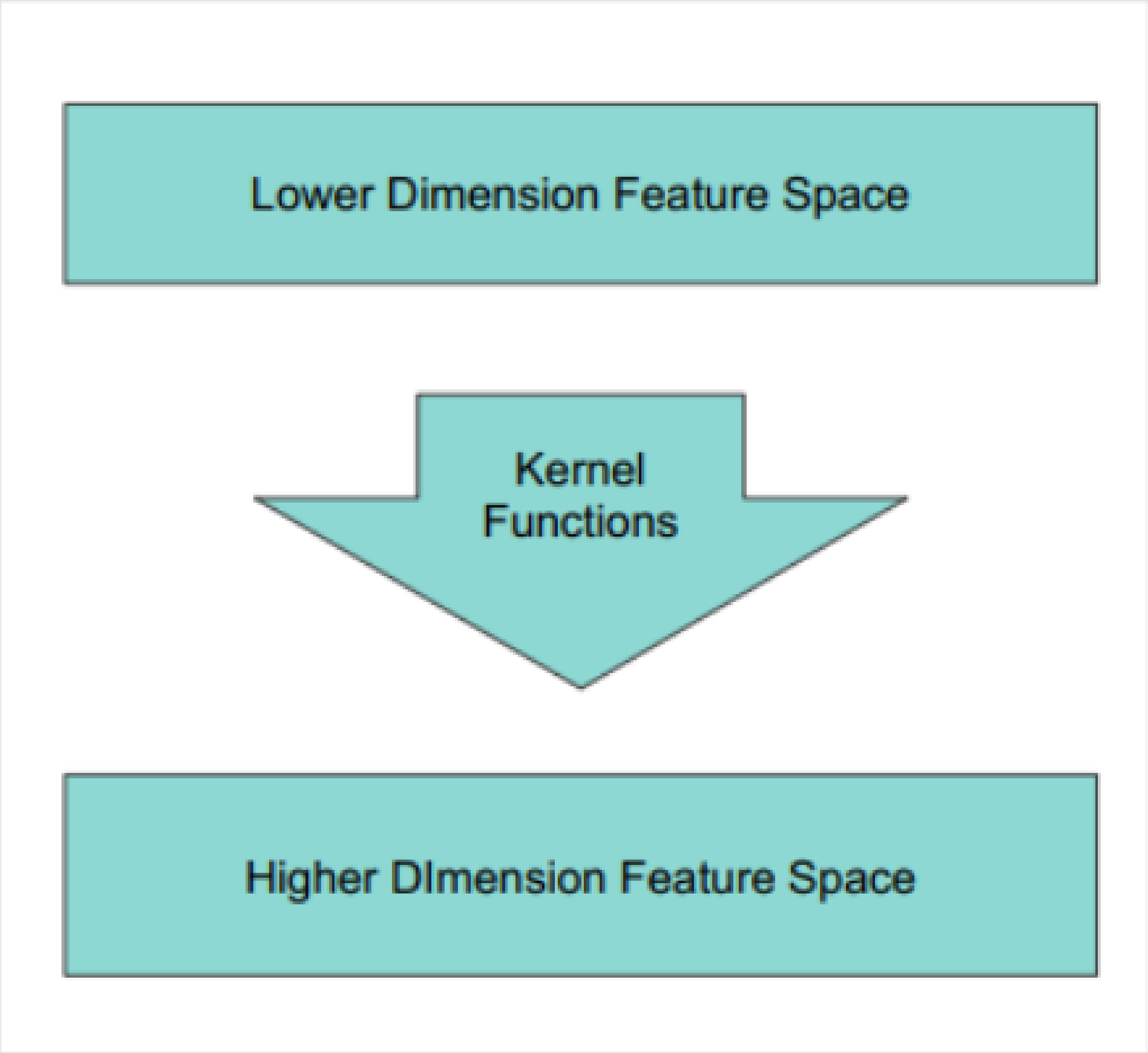

y = x²
DB
(-3, 9)
(-2, 4)
(-1, 1)
(0, 0)
(1, 1)
(2, 4)
(3, 9)

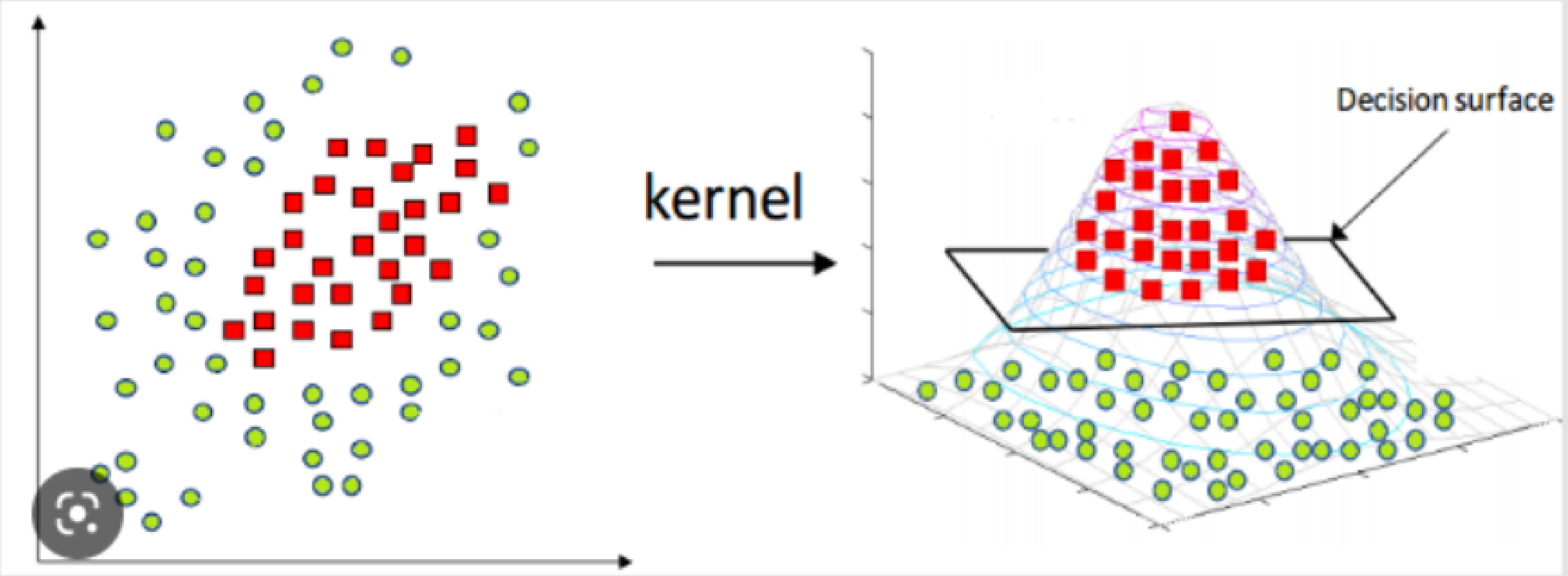


Kernel Trick: try to find a non-linear decision boundary

The **kernel trick** projects the original data points in a higher dimensional space in order to make them linearly separable (in that higher dimensional space).

> *Thus, by using the kernel trick we can make our non linearly-separable data, linearly separable in a higher dimensional space.*


Lower Dimension Feature Space
Kernel
Functions
Higher DImension Feature Space


kernel
Decision surface

* The kernel trick is based on some Kernel functions that measure similarity of the samples.
* The trick does not actually transform the data points to a new, high dimensional feature space, explicitly
* The kernel-SVM computes the decision boundary in terms of similarity measures in a high-dimensional feature space without actually doing the projection.
* Some famous kernel functions include linear, polynomial, radial basis function (RBF), and sigmoid kernels.

Also known as radial basis function

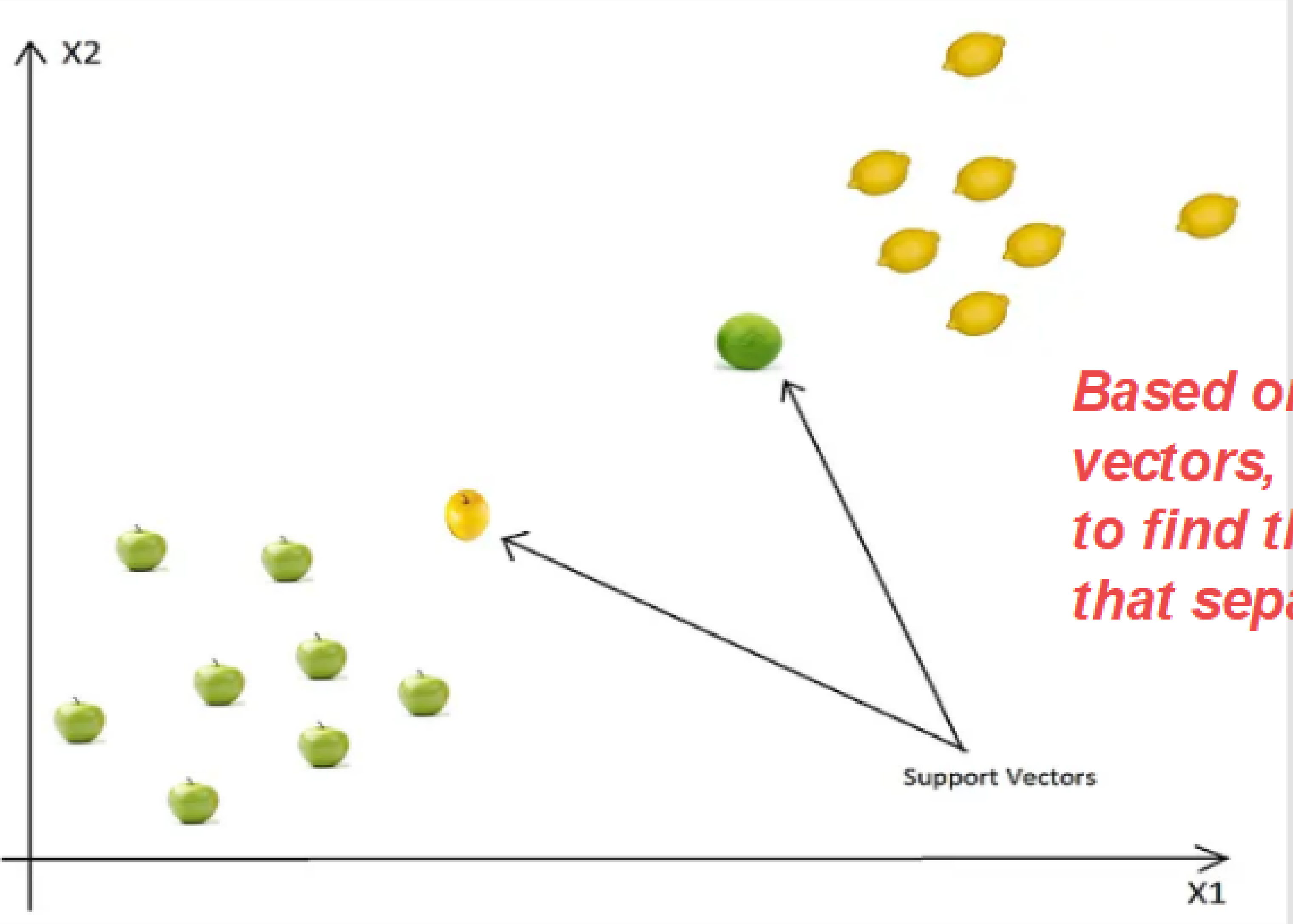
Type of Kernel	Inner product kernel $K(\vec{x}, \vec{x}_i),\ i = 1, 2, \dots, N$	Comments
Polynomial Kernel	$K(\vec{x}, \vec{x}_i) = (\vec{x}^T \vec{x}_i + \theta)^d$	Power p and threshold θ is specified a priori by the user
Gaussian Kernel	$K(\vec{x}, \vec{x}_i) = e^{-\frac{1}{2\sigma^2}\lVert\vec{x}-\vec{x}_i\rVert^2}$	Width σ^2 is specified a priori by the user
Sigmoid Kernel	$K(\vec{x}, \vec{x}_i) = tanh(\eta\, \vec{x}\, \vec{x}_i + \theta)$	Mercer's Theorem is satisfied only for some values of η and θ
Kernels for Sets	$K(\chi, \chi') = \sum_{i=1}^{N_\chi} \sum_{j=1}^{N_{\chi'}} k(x_i, x'_j)$	Where $k(x_i, x'_j)$ is a kernel on elements in the sets χ, χ'
Spectrum Kernel for strings	count number of substrings in common	It is a kernel, since it is a dot product between vectors of indicators of all the substrings.

Mathematical Formulation of SVM

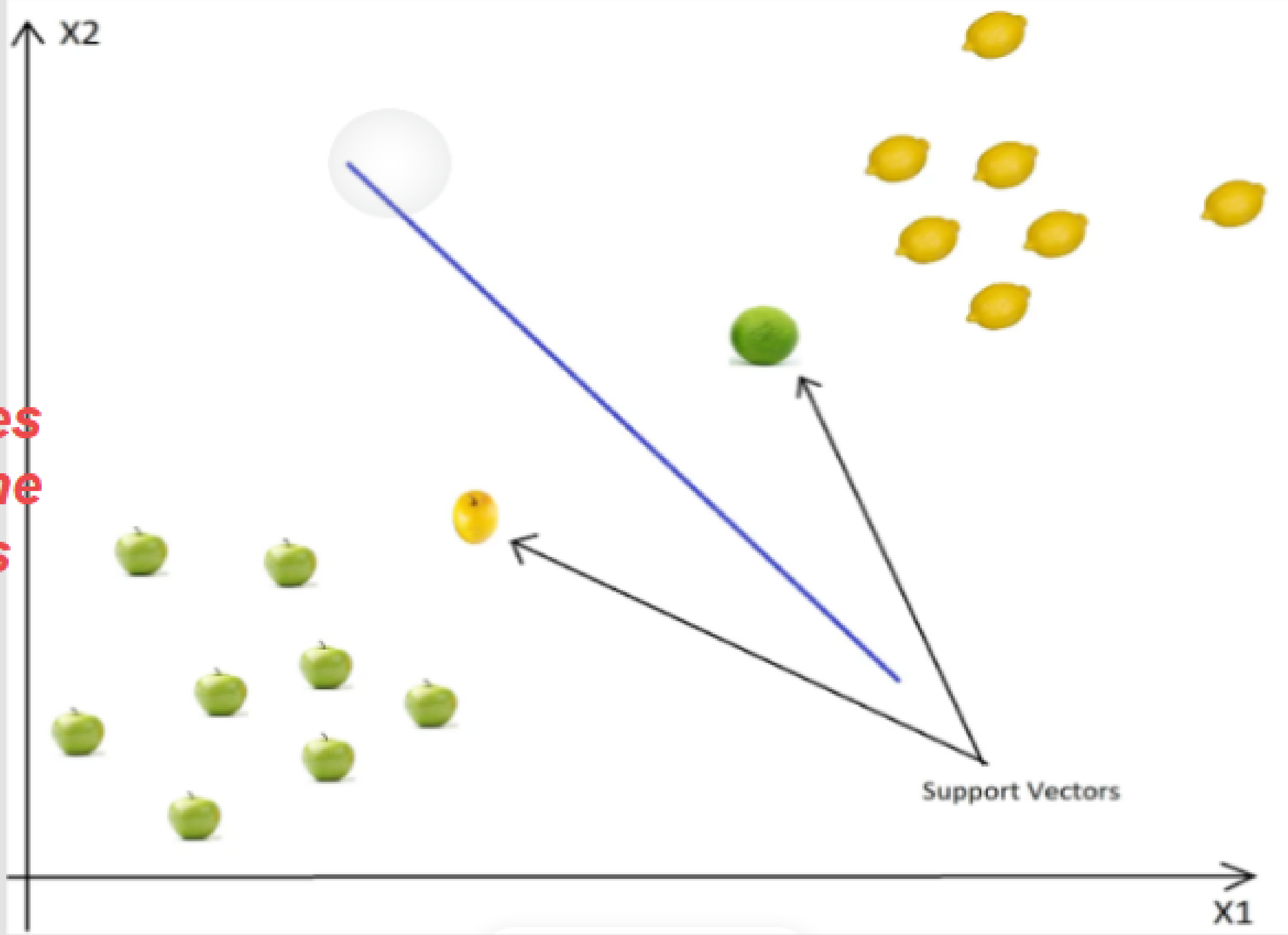
Classification : Apple vs Lemon

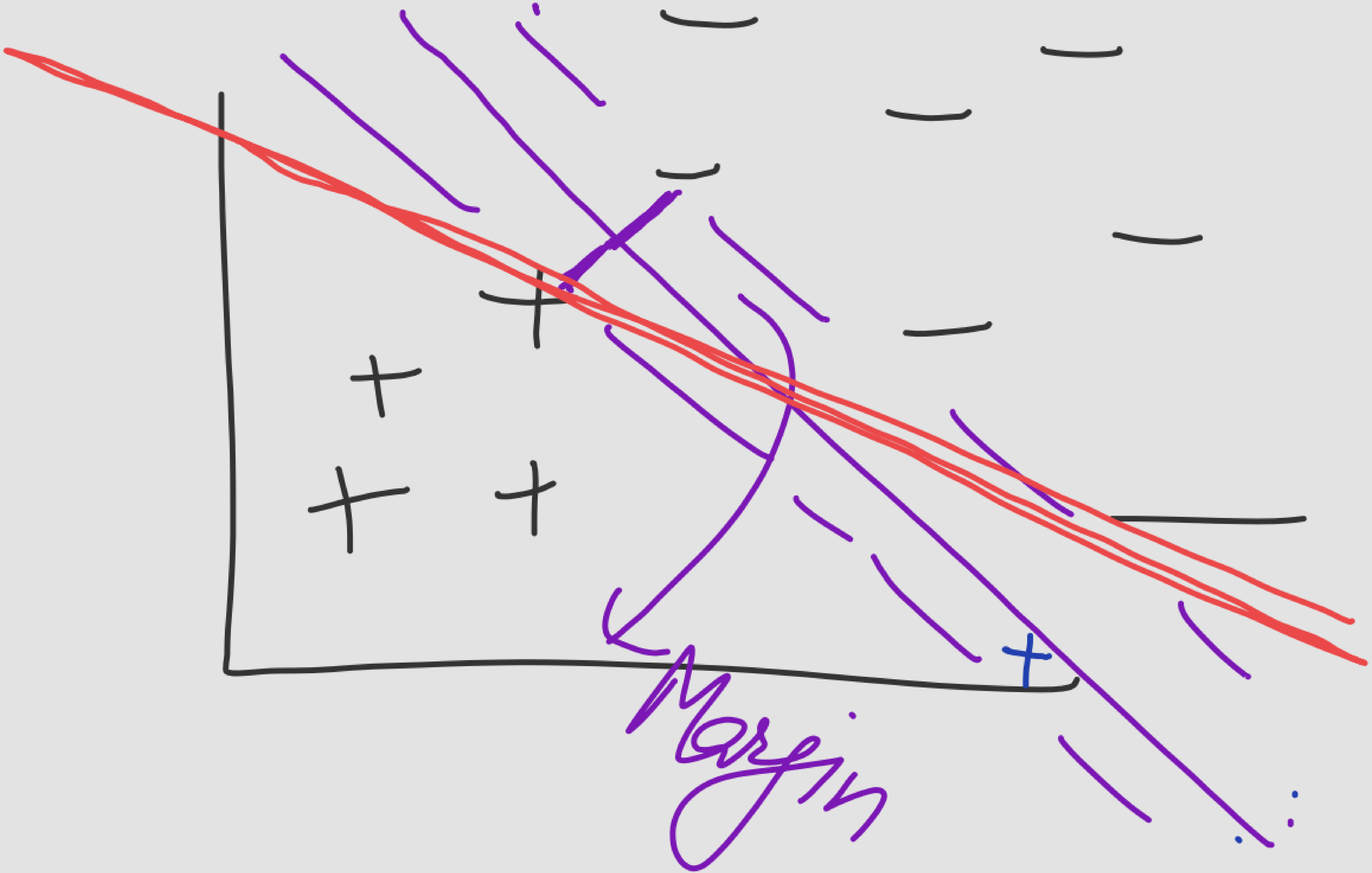
→ Search for apples which are similar to lemons

* SVM learns similarities.



Based on these support vectors, the algorithm tries to find the best hyperplane that separates the classes




Margin