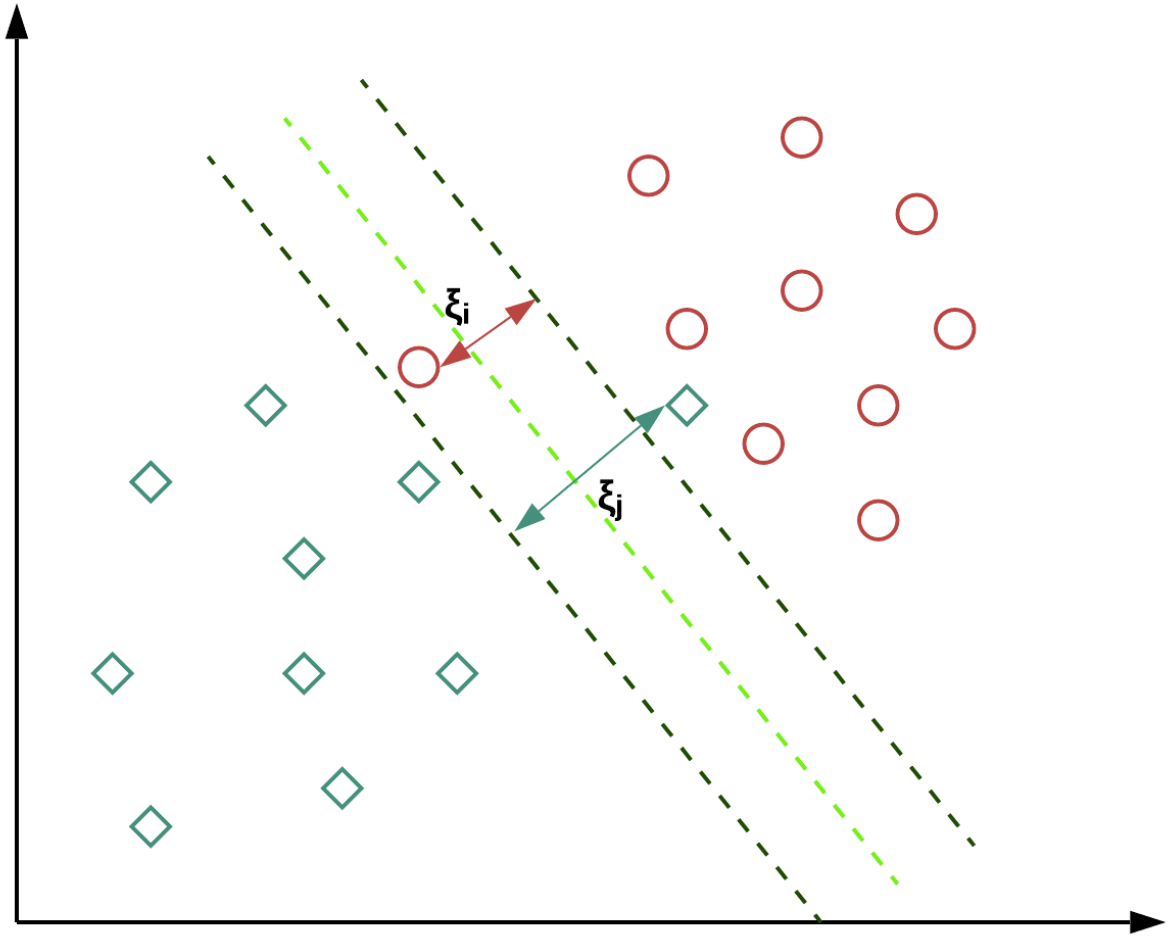
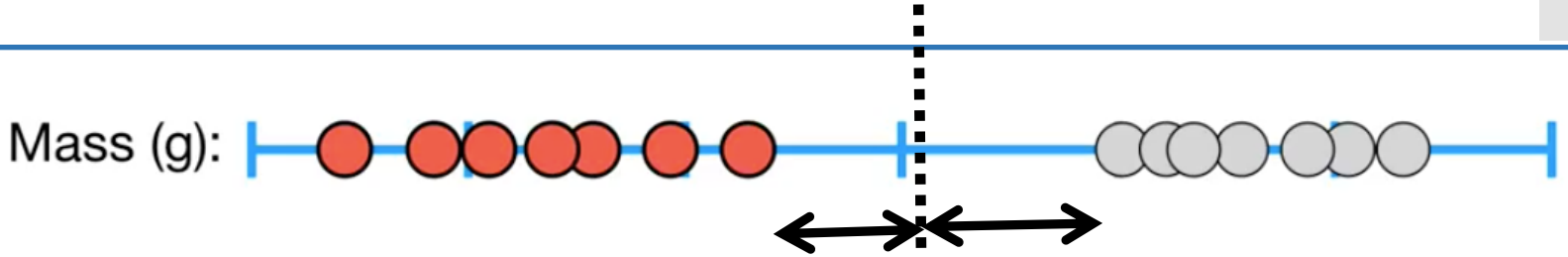
Support Vector Machines (SVM)



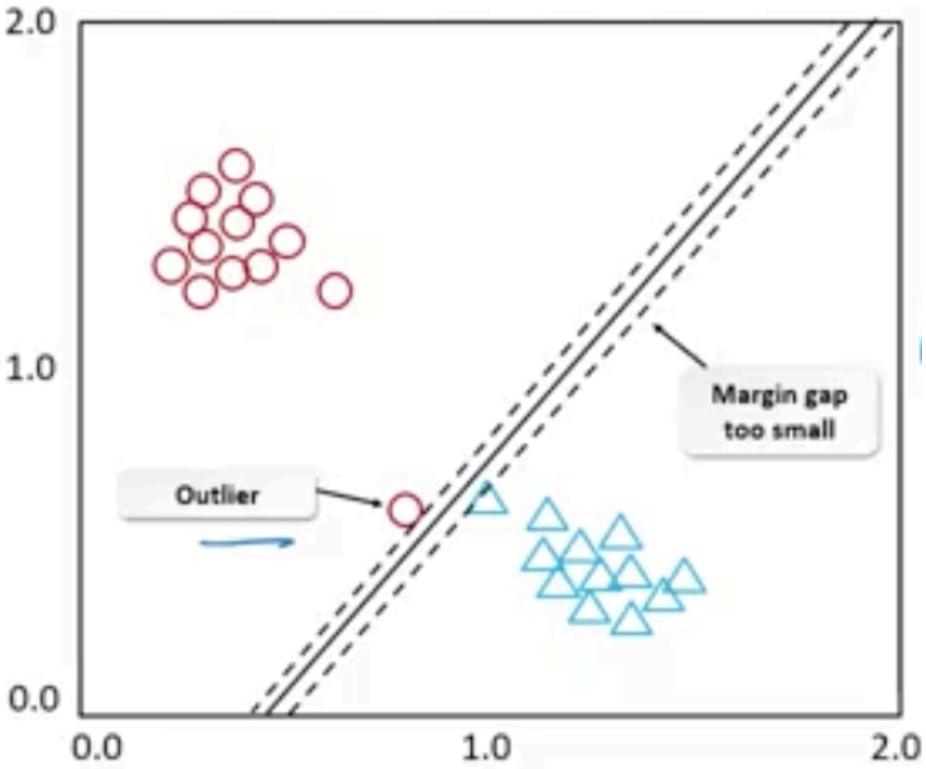


- The red dots represent mice that were not obese and the green dots represent mice that were obese
- Based on these observation, we can pick a threshold and when we get a new observation that has less mass than the threshold, we can classify it as not obese. For new observation with more mass than the threshold we can classify it as obese

- We can focus on the observation on the edges of each cluster and use the midpoint between them as the threshold. The shortest distance between the observation and the threshold is called the margin. When we use the threshold that gives us the largest margin to make classifications: Maximum Margin Classifier

Maximal Margin Classifiers

- The Maximal Margin Classifiers are super sensitive to outliers in the training data and that makes them defective

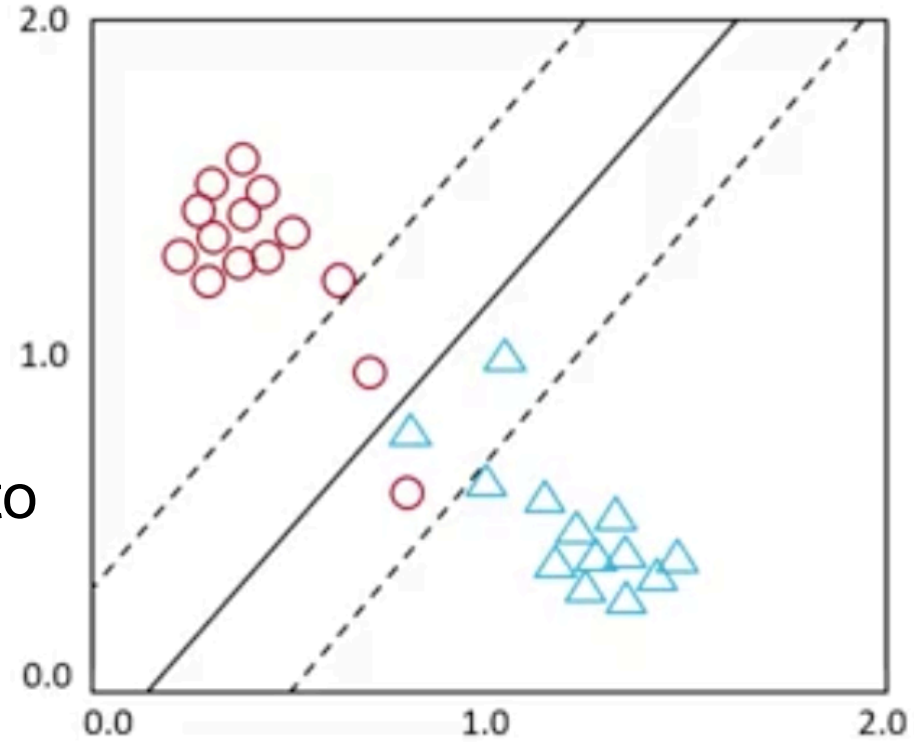


Soft Margin Classifiers or Support Vector Classifier

- To make a threshold that is not so sensitive to outliers we must allow misclassification
- Choosing a threshold that allows misclassification is an example of the Bias/Variance Tradeoff that plagues all of machine learning

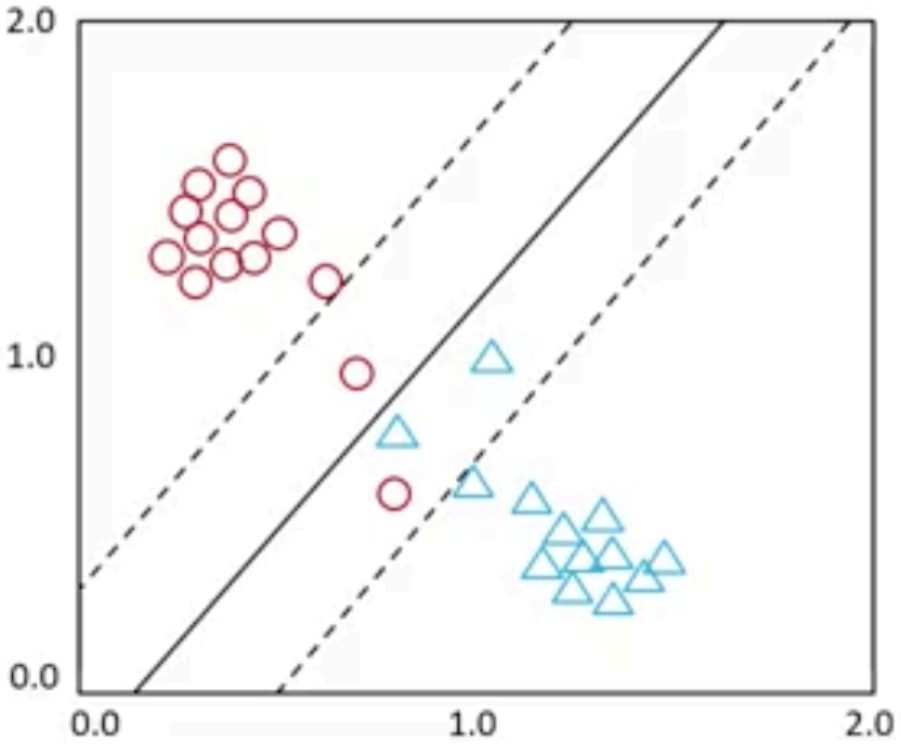
When we allow misclassifications, the distance between the observations and the threshold is called a Soft Margin

We use Cross Validation to determine how many misclassification and observation to allow inside of the Soft Margin to get the best classification



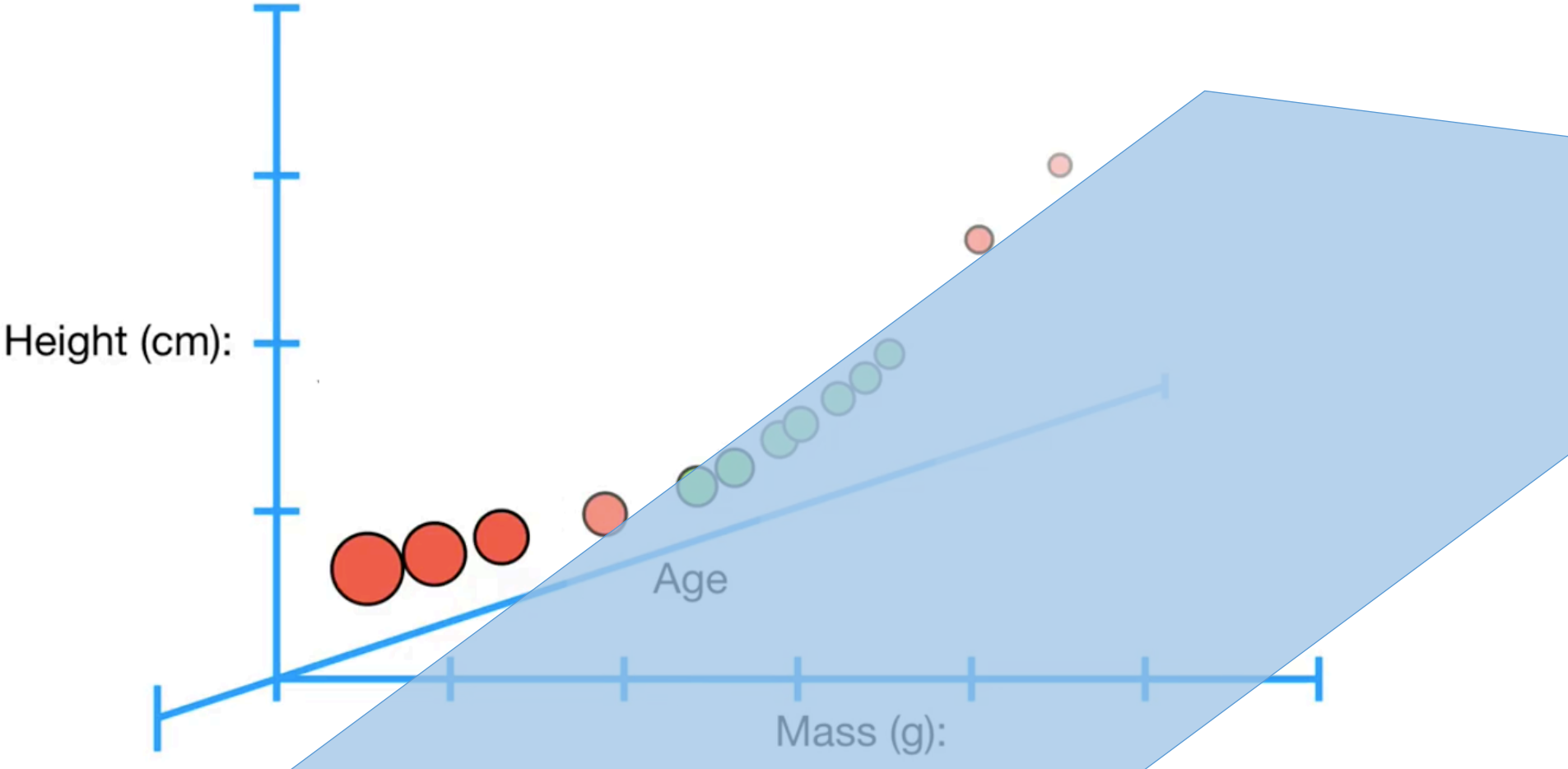
Soft Margin Classifiers or Support Vector Classifier

- Support Vector Classifier comes from the fact that the observations on the edge and within the Soft Margin are called Support Vectors
- When we use a Soft Margin to determine the location of threshold then we are using a Support vector Classifier to classify observation
- We used Cross Validation to determine that allowing this misclassification results in better classification in the long run



Support Vector Classifier

- When the data are 3-Dimensional, the Support Vector Classifier forms a plane, instead of line and we classify new observations by determining which side of the plane they are on



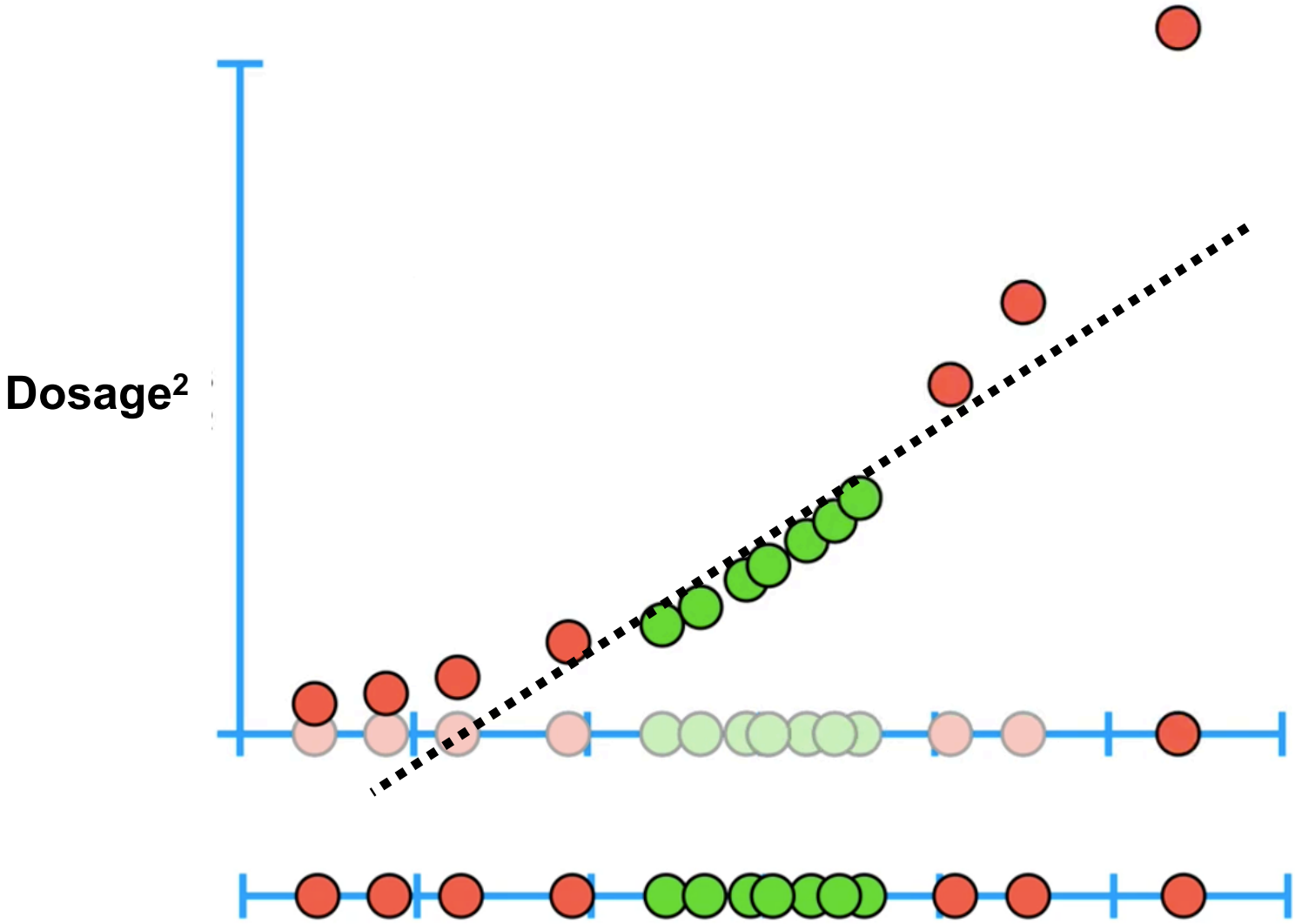
- When the data are in 2-Dimensions, the support Vector Classifier is a 1-Dimensional line in a 2-Dimensional space
- And when the data are 3-Dimensional, the Support Vector Classifier is a 2-Dimensional plane in a 3-Dimensional space.
- And when the data are in 4 or more Dimension, the Support Vector Classifier is a hyperplane

Support Vector Machines

- The red dots represented patients that were not cured and the green dots represented patients that were cured
- The drug doesn't work if the dosage is too small or too large. It only works when the dosage is just right
- The training dataset had so much overlap, we are unable to find a satisfying Support Vector Classifier to separate the patients that were cured from the patients that were not cured

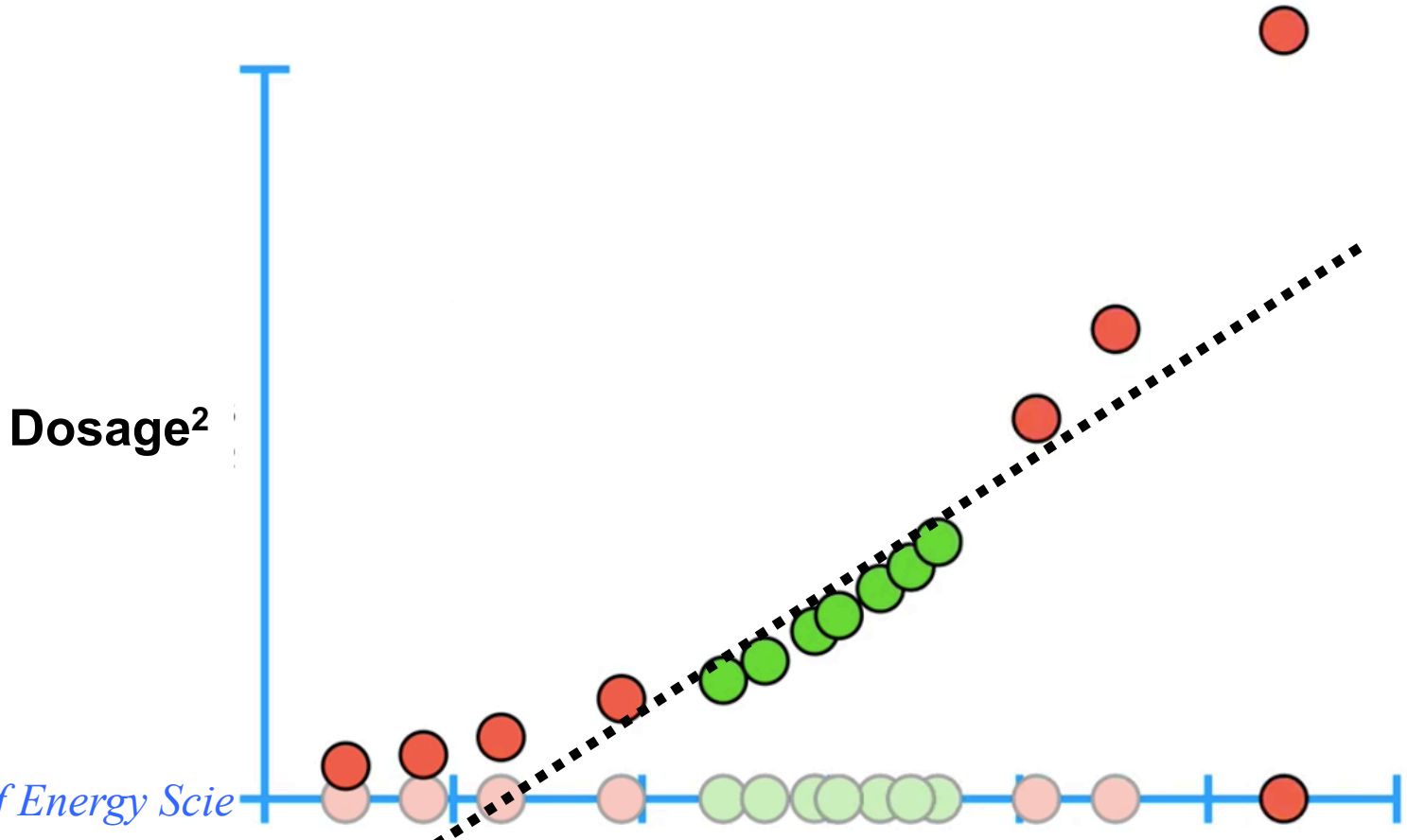
Dosage (mg)

When we gave each point a y-axis coordinate by squaring the original Dosage measurements

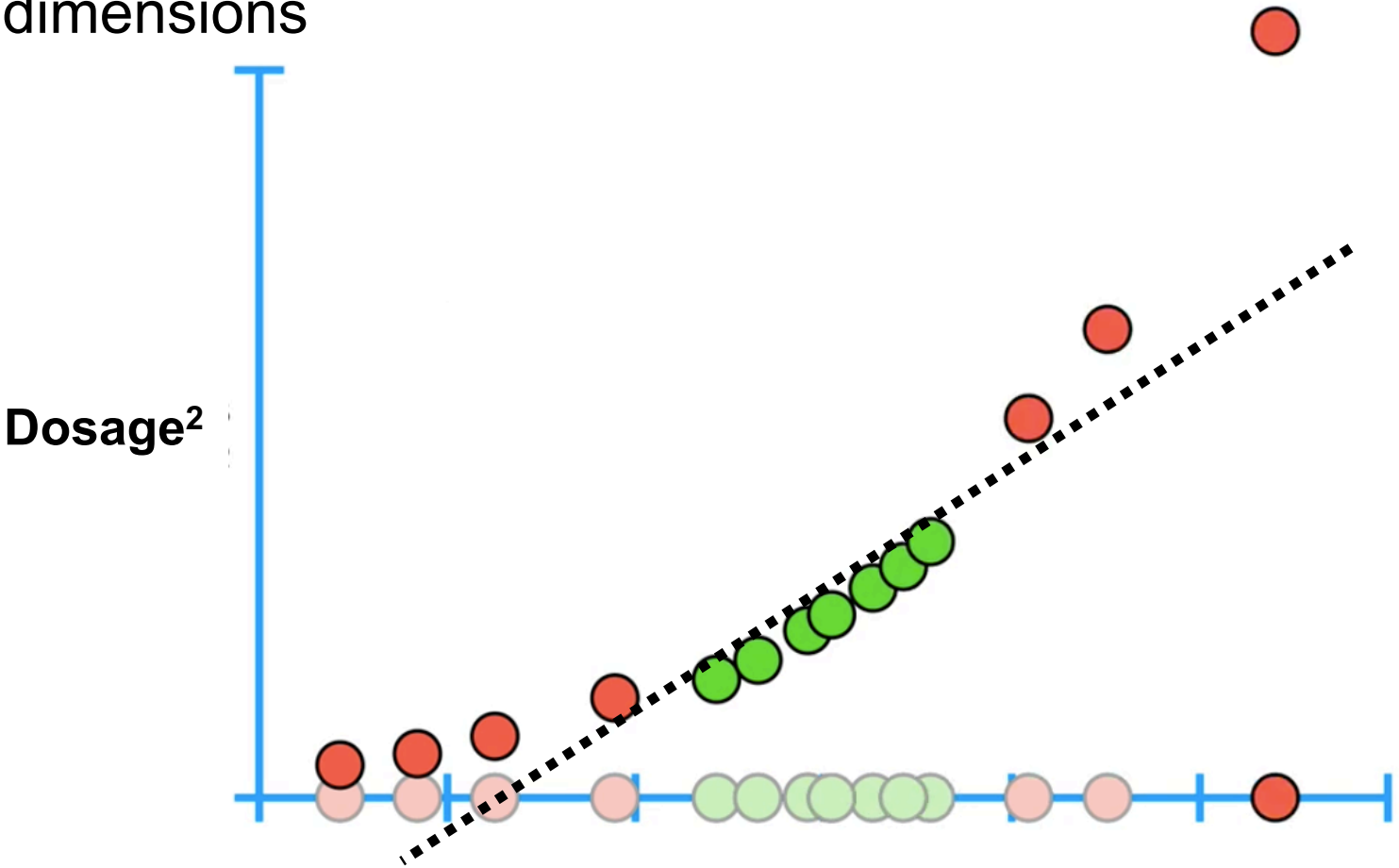


Support Vector Machines

- Start with data in a relatively low dimension
- Move the data into higher dimension
- Find a Support Vector Classifier that separates the higher dimension data into two groups

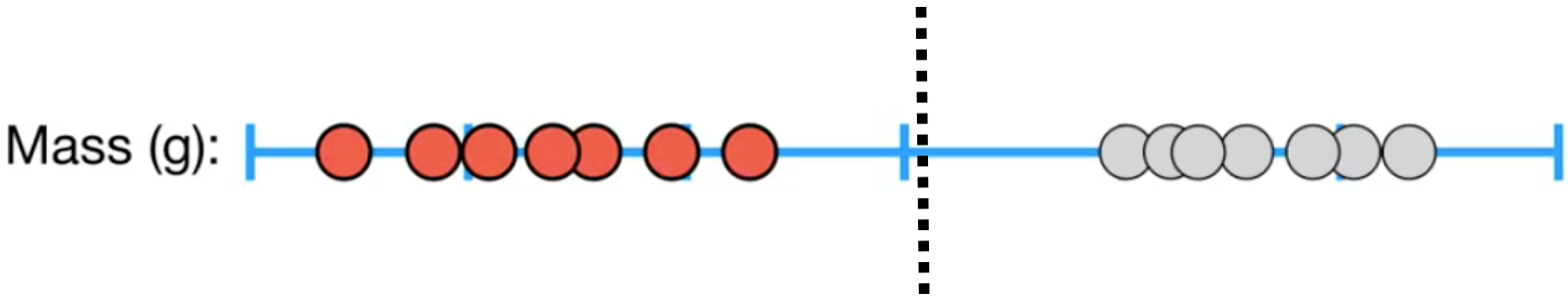


- In order to make the mathematics possible, Support Vector Machines use something called Kernel Functions to systematically find Support Vector Classifiers in higher dimensions



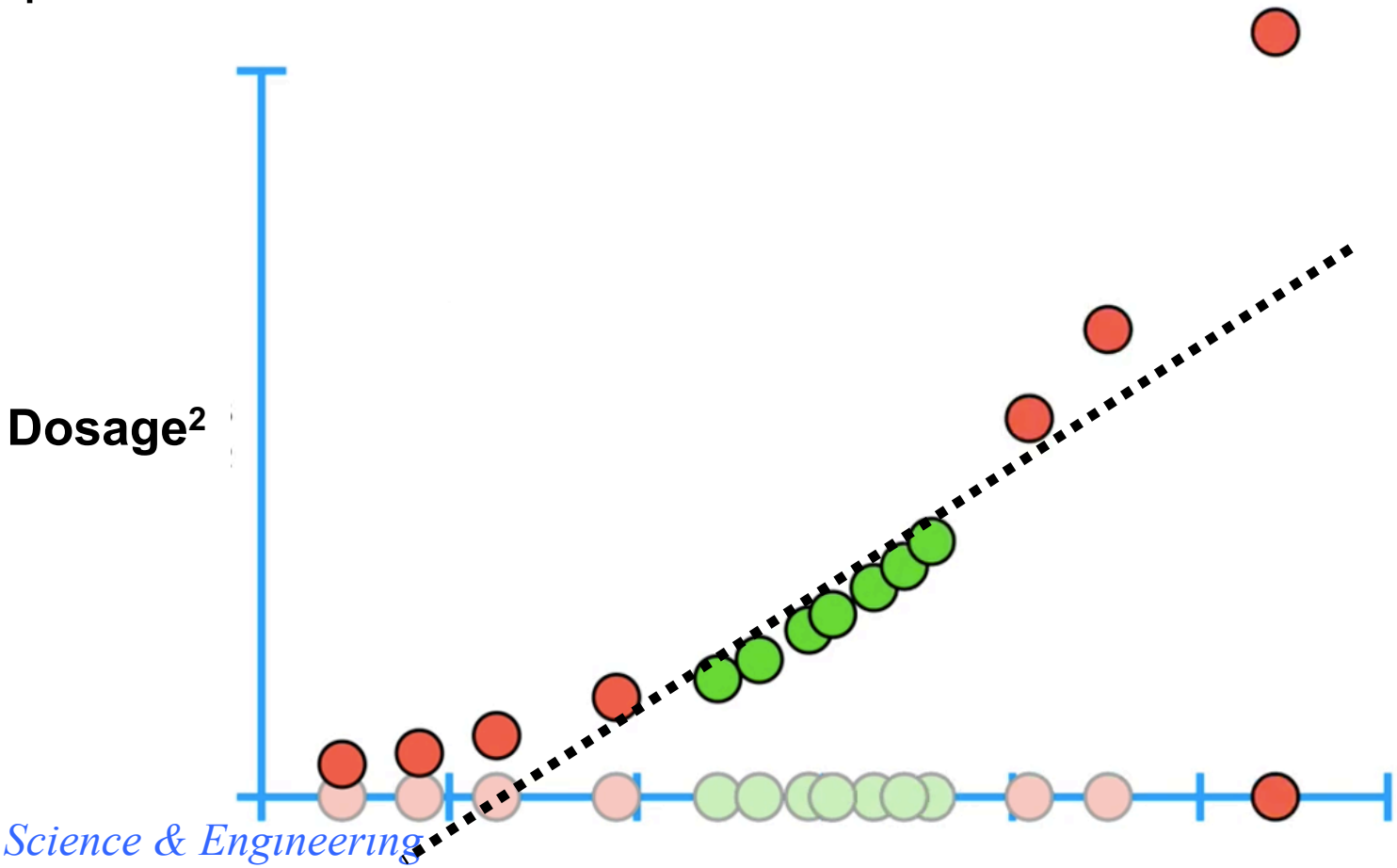
Kernel Functions to systematically finds Support Vector Classifiers in higher dimensions

- For this example, We use the Polynomial Kernel, which has a parameter, d, which stands for the degree of the polynomial
- When d = 1, the Polynomial Kernel computes the relationships between each pair of observations in 1-Dimension



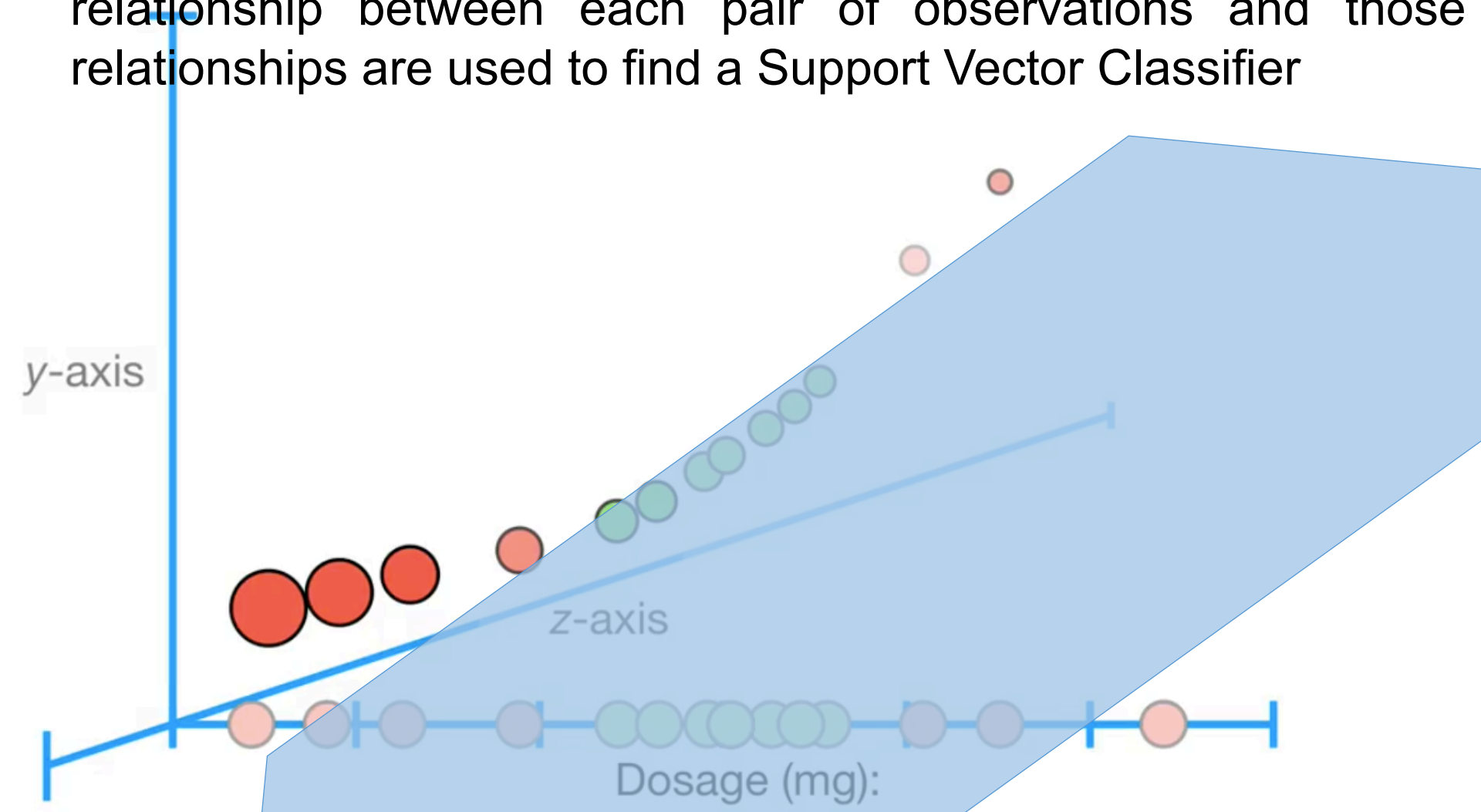
Support Vector Machines

- When d = 2, we get a 2nd dimension based on Dosage^2
- The Polynomial Kernel computes the 2-Dimensional relationship between each pair of observations and those relationships are used to find a Support Vector Classifier



Support Vector Machines

- When d = 3, we get a 3rd dimension based on Dosage^3
- The Polynomial Kernel computes the 3-Dimensional relationship between each pair of observations and those relationships are used to find a Support Vector Classifier



Polynomial Kernel

- Polynomial kernel calculates high-dimensional relationship

$$\left(a \times b + r\right)^d$$

$$\left(a \times b + \frac{1}{2}\right)^2 = \left(a \times b + \frac{1}{2}\right)\left(a \times b + \frac{1}{2}\right)$$

$$= ab + a^2 b^2 + \frac{1}{4}$$

$$= \left(a, a^2, \frac{1}{2}\right) \bullet \left(b, b^2, \frac{1}{2}\right)$$

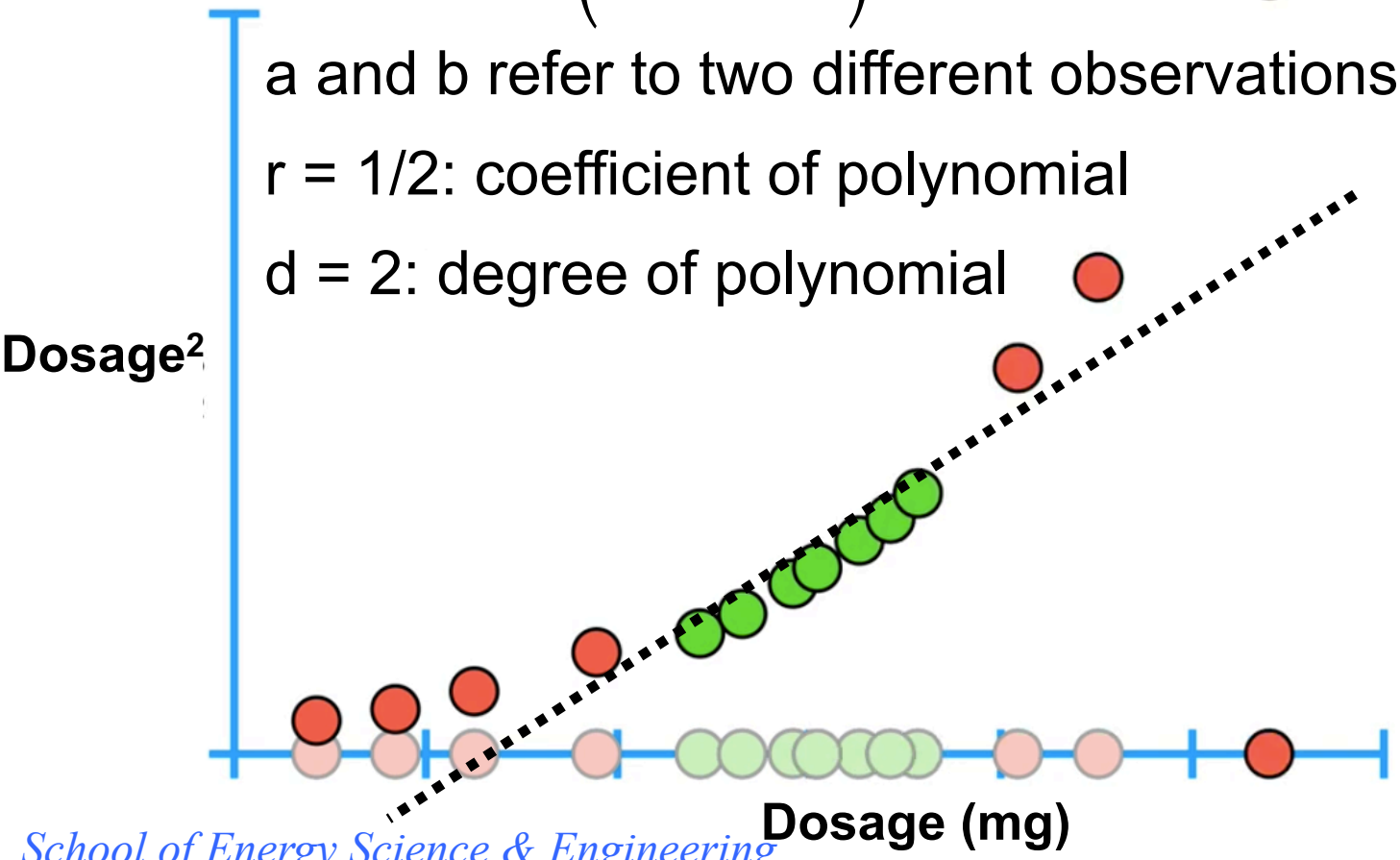
Support Vector Machine with a Polynomial Kernel to compute the observations in higher dimension and then found a good Support Vector Classifier based on the high dimension relationships

$$\left(a \times b + r\right)^d$$

a and b refer to two different observations in the dataset

r = 1/2: coefficient of polynomial

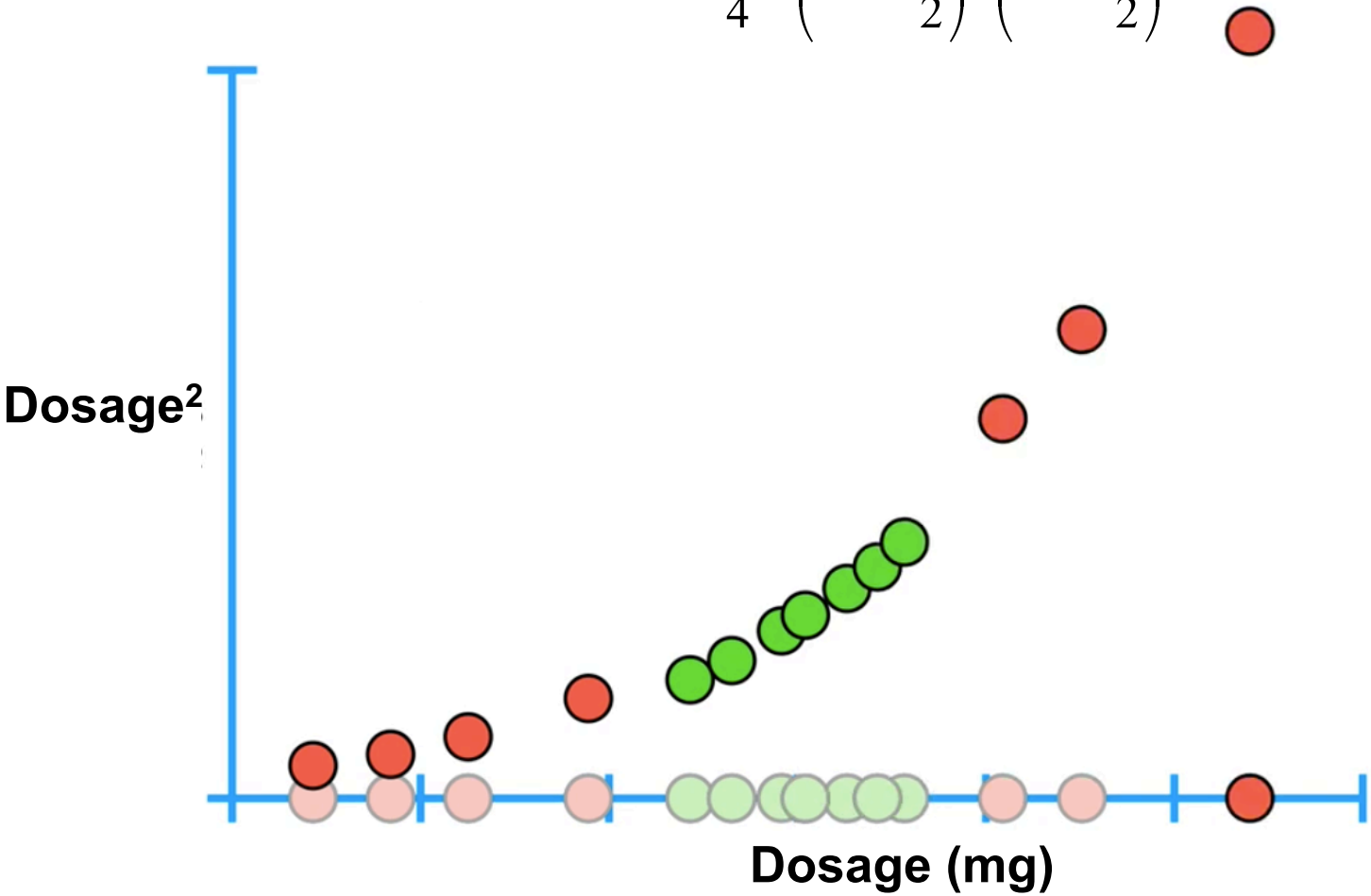
d = 2: degree of polynomial



The Dot Product gives the high dimension coordinates for the data

$$\left(a \times b + \frac{1}{2}\right)^2 = \left(a \times b + \frac{1}{2}\right)\left(a \times b + \frac{1}{2}\right)$$

$$= ab + a^2 b^2 + \frac{1}{4} = \left(a, a^2, \frac{1}{2}\right) \bullet \left(b, b^2, \frac{1}{2}\right)$$



The high dimensional relationships between these two observations: a = 9 and b =14

$$\left(9\times 14+\frac{1}{2}\right)^2=\left(126+\frac{1}{2}\right)^2=126.5^2=16002.25$$

Support Vector Machines with a Radial Kernel

- Radial Kernel or the Radial Basis Function (RBF)

$$e^{-\gamma(a-b)^2}$$

The Radial Kernel finds Support Vector Classifiers in infinite dimensions, its not possible to visualize what it does

The Radial Kernel behaves like a Weighted Nearest Neighbor model

The closest observations (the nearest neighbors) have a lot of influence on how we classify the new observation and observations that are further away have relatively little influence on the classification

Radial Kernel determines the influence of each observation in the Training Dataset has on classifying new observations

$$e^{-\gamma(a-b)^2}$$

- Just like Polynomial Kernel, a and b refer to two different Dosage measurements
- The difference between the measurements is then squared, giving us the squared distance between the two observations
- Thus the amount of influence one observation has on another is a function of the squared distance
- Gamma, which is determined by Cross Validation, scales the squared distance and thus scales the influence

Dosage (mg)

Support Vector Machines with a Radial Kernel

$$e^{-\gamma(a-b)^2}$$

- For example, if we set gamma (γ) = 1 and plug in the Dosages from two observations that are relatively close to each other

gamma (γ) = 1 $$e^{-(2.5-4)^2} = e^{-(-1.5)^2} = e^{-2.25} = 0.11$$

gamma (γ) = 2 $$e^{-2(2.5-4)^2} = e^{-2(-1.5)^2} = e^{-2\times 2.25} = 0.01$$

- So we see that by scaling the distance, gamma scales the amount of influence two points have on each other

Support Vector Machines with a Radial Kernel

- For example, if we set gamma (γ) = 1 and plug in the Dosages from two observations (2.5 and 16) when they are relatively far from each other

gamma (γ) = 1

$$e^{-(2.5-16)^2} = e^{-(-13.5)^2} = e^{-182.25} =$$ A number very close to zero

- Thus, the further two observations are far from each other, the less influence they have on each other
- When we plug values into the Radial Kernel, we get the high-dimensional relationship
- Thus, 0.11 is the high-dimensional relationship between these two observations that are relatively close ti each other

Radial Kernel works in Infinite-Dimensions

$$\left(a \times b + r\right)^d$$

- We will use the Polynomial Kernel to give us intuition into how the Radial Kernel works in Infinite-Dimensions

When r = 0, the Polynomial Kernel simplifies to a single term and that gives us a Dot Product with a single coordinate

$$\left(a \times b + r\right)^d = \left(a \times b\right)^d = a^d b^d = \left(a^d\right) \bullet \left(b^d\right)$$

When d = 2, we get

$$a^2 b^2 = \left(a^2\right) \bullet \left(b^2\right)$$

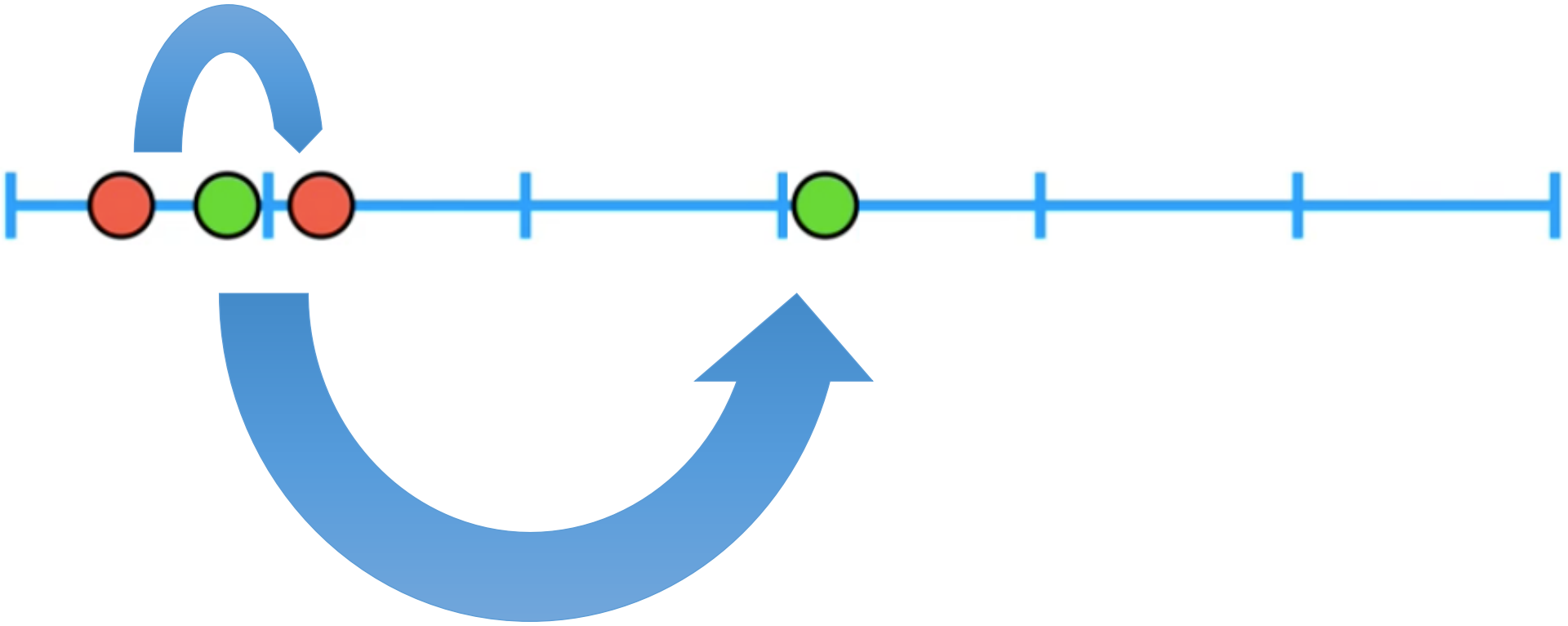
The new coordinate is just the square of the original measurement on the original axis

Polynomial Kernel

$$a^2b^2 = \left(a^2\right) \bullet \left(b^2\right)$$

The new coordinate is just the square of the original measurement on the original axis

The polynomial Kernel shift the data down the original axis



Polynomial Kernel

When r = 0, $\left(a \times b + r\right)^d = \left(a \times b\right)^d = a^d b^d = \left(a^d\right) \bullet \left(b^d\right)$

When r = 0, d = 1 $a^1 b^1 = \left(a\right) \bullet \left(b\right)$

When r = 0, d = 2 $a^2 b^2 = \left(a^2\right) \bullet \left(b^2\right)$

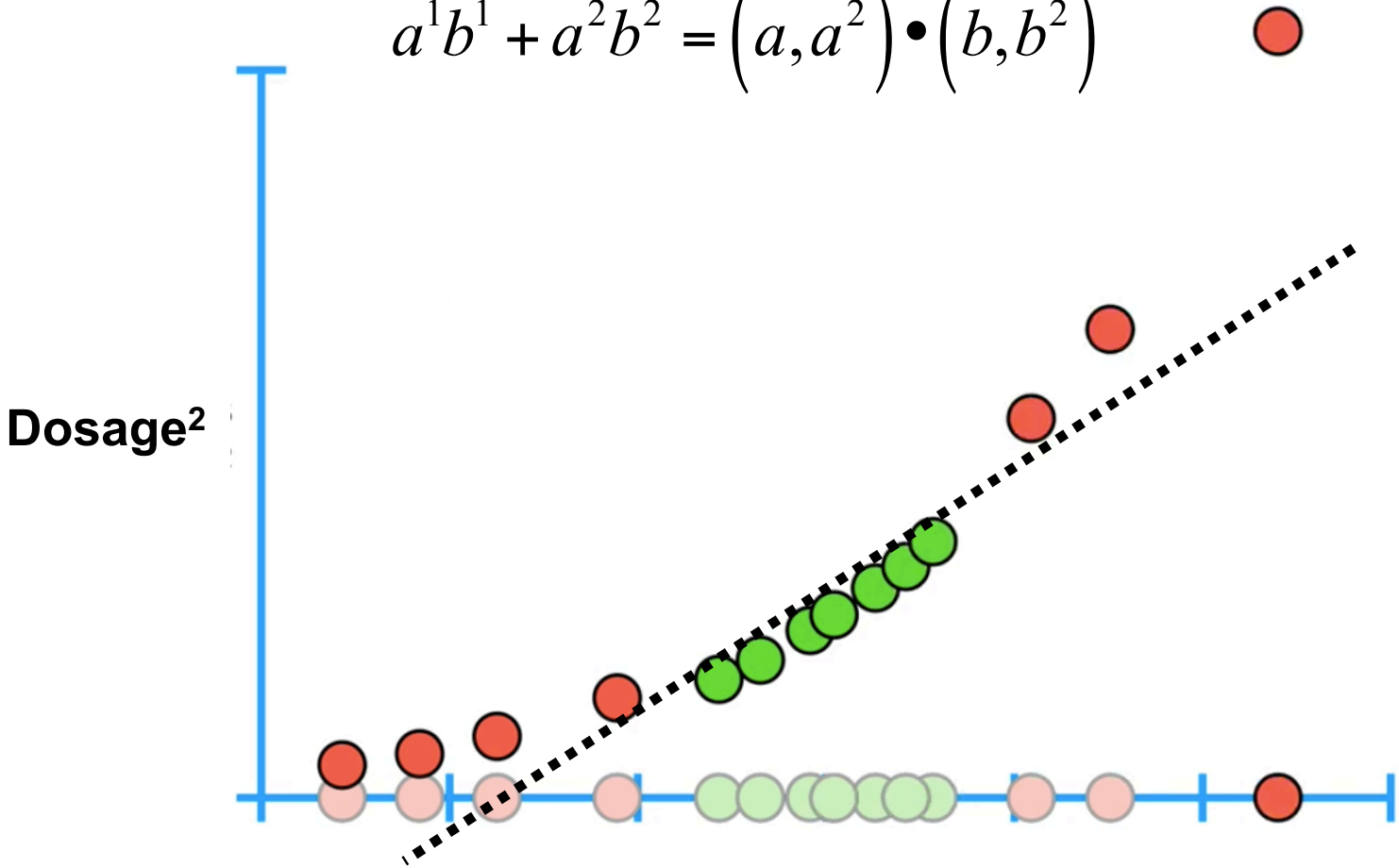
When r = 0, d = 3 $a^3 b^3 = \left(a^3\right) \bullet \left(b^3\right)$

In this example, the data stays on the same 1-Dimensional line

Polynomial Kernel

If we take a Polynomial Kernel with r = 0, d = 1 and add another Polynomial Kernel with with r = 0, d = 2

$$a^1b^1 + a^2b^2 = \left(a, a^2\right) \bullet \left(b, b^2\right)$$

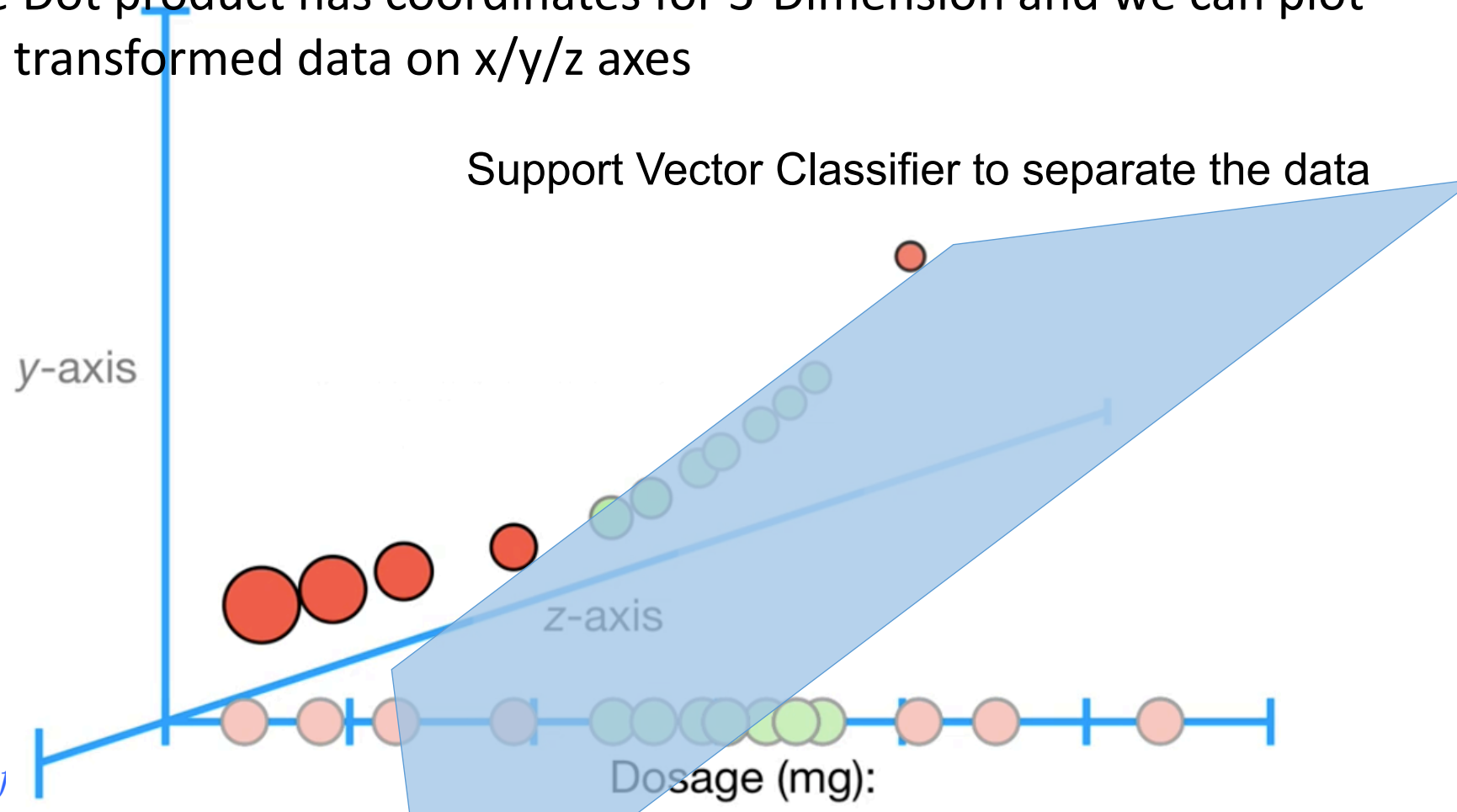


Polynomial Kernel

If we add three Polynomial Kernel with r = 0, d = 1; r = 0, d = 2; r = 0, d = 3

$$a^1b^1 + a^2b^2 + a^3b^3 = \left(a, a^2, a^3\right) \bullet \left(b, b^2, b^3\right)$$

The Dot product has coordinates for 3-Dimension and we can plot the transformed data on x/y/z axes



Polynomial Kernels with r =0 and increasing d until d = infinity

$$a^1b^1 + a^2b^2 + a^3b^3 + \cdots + a^\infty b^\infty = \left(a + a^2 + a^3 + \cdots + a^\infty\right) \bullet \left(b + b^2 + b^3 + \cdots + b^\infty\right)$$

That would give us a Dot product with coordinates for an infinite number of dimensions: that's exactly what the Radial Kernel does

$$e^{-\gamma(a-b)^2}$$

$$e^{-\frac{1}{2}(a-b)^2} = e^{-\frac{1}{2}(a^2+b^2-2ab)} = e^{-\frac{1}{2}(a^2+b^2)} \bullet e^{ab}$$

Taylor Series

$$f(x) = f(a) + \frac{f'(x)}{1!}(x-a) + \frac{f''(x)}{2!}(x-a)^2 + \frac{f'''(x)}{3!}(x-a)^3 + \cdots + \frac{f^{\infty}(x)}{\infty!}(x-a)^{\infty}$$

$$e^x = e^a + \frac{e^a}{1!}(x-a) + \frac{e^a}{2!}(x-a)^2 + \frac{e^a}{3!}(x-a)^3 + \cdots + \frac{e^a}{\infty!}(x-a)^{\infty}$$

$$e^x = 1 + \frac{1}{1!}x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots + \frac{1}{\infty!}x^{\infty}$$

$$e^{ab} = 1 + \frac{1}{1!}ab + \frac{1}{2!}(ab)^2 + \frac{1}{3!}(ab)^3 + \cdots + \frac{1}{\infty!}(ab)^{\infty}$$

$$e^{ab} = \left(1, \sqrt{\frac{1}{1!}}a, \sqrt{\frac{1}{2!}}a^2, \sqrt{\frac{1}{3!}}a^3, \cdots, \sqrt{\frac{1}{\infty!}}a^{\infty}\right) \bullet \left(1, \sqrt{\frac{1}{1!}}b, \sqrt{\frac{1}{2!}}b^2, \sqrt{\frac{1}{3!}}b^3, \cdots, \sqrt{\frac{1}{\infty!}}b^{\infty}\right)$$

Support Vector Machines with a Radial Kernel

$$e^{-\frac{1}{2}(a-b)^2} = e^{-\frac{1}{2}(a^2+b^2-2ab)} = e^{-\frac{1}{2}(a^2+b^2)} \bullet e^{ab}$$

$$e^{-\frac{1}{2}(a-b)^2} = e^{-\frac{1}{2}(a^2+b^2)}\left[\left(1,\sqrt{\frac{1}{1!}}a,\sqrt{\frac{1}{2!}}a^2,\sqrt{\frac{1}{3!}}a^3,\cdots,\sqrt{\frac{1}{\infty!}}a^\infty\right)\bullet\left(1,\sqrt{\frac{1}{1!}}b,\sqrt{\frac{1}{2!}}b^2,\sqrt{\frac{1}{3!}}b^3,\cdots,\sqrt{\frac{1}{\infty!}}b^\infty\right)\right]$$

$$s = \sqrt{e^{-\frac{1}{2}\left(a^2+b^2\right)}}$$

$$e^{-\frac{1}{2}(a-b)^2} = \left(s,s\sqrt{\frac{1}{1!}}a,s\sqrt{\frac{1}{2!}}a^2,s\sqrt{\frac{1}{3!}}a^3,\cdots,s\sqrt{\frac{1}{\infty!}}a^\infty\right)\bullet\left(s,s\sqrt{\frac{1}{1!}}b,s\sqrt{\frac{1}{2!}}b^2,s\sqrt{\frac{1}{3!}}b^3,\cdots,s\sqrt{\frac{1}{\infty!}}b^\infty\right)$$

The Radial Kernel is equal to a Dot Product that has coordinates for an infinite number of dimensions

The value we get at the end is the relationship between the two points in infinite-dimensions

$$e^{-(2.5-4)^2} = e^{-(-1.5)^2} = e^{-2.25} = 0.11$$

Numerical (SVM)

Suppose we have two different dosage levels of a medication for treating diabetes, i.e., x=3 and y=8. Using the **Radial Basis Function (RBF) kernel**, calculate the influence of x on y with γ=1.5, which scales the influence of the distance between the two dosage levels.