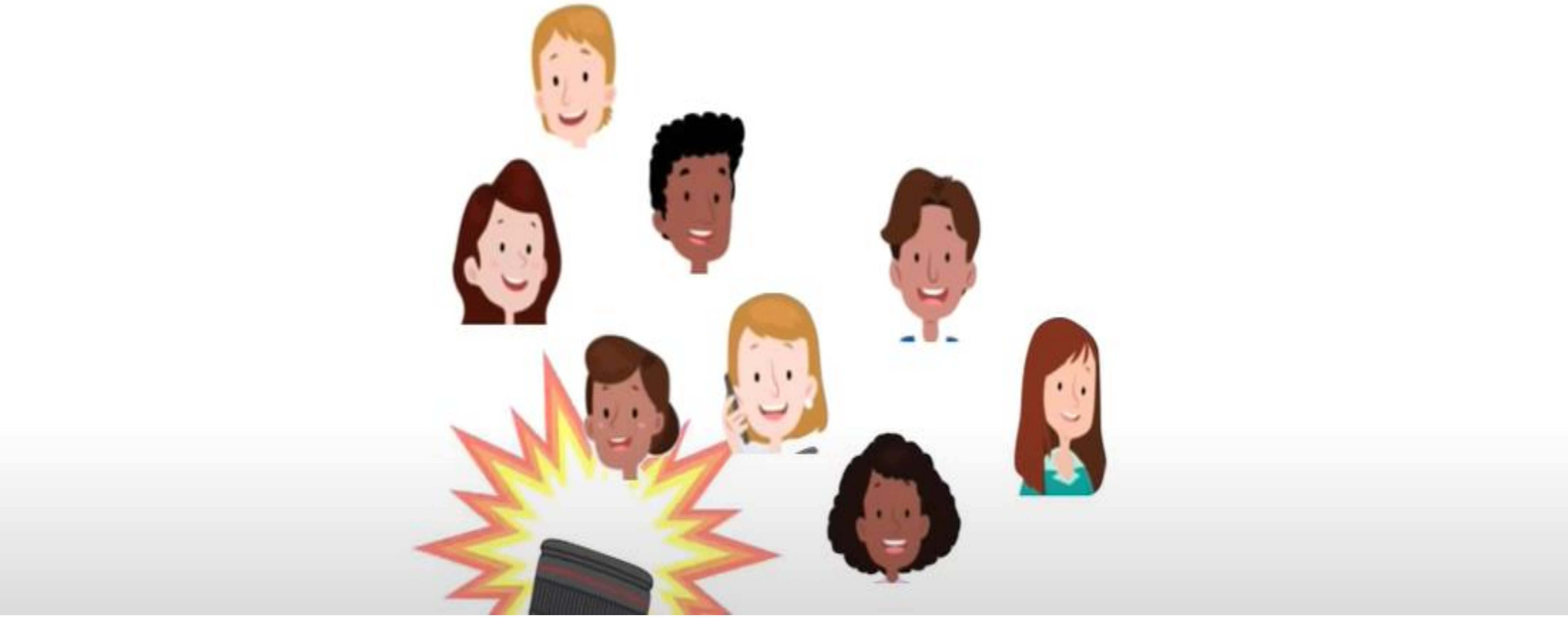
PCA

PRINCIPAL COMPONENT ANALYSIS

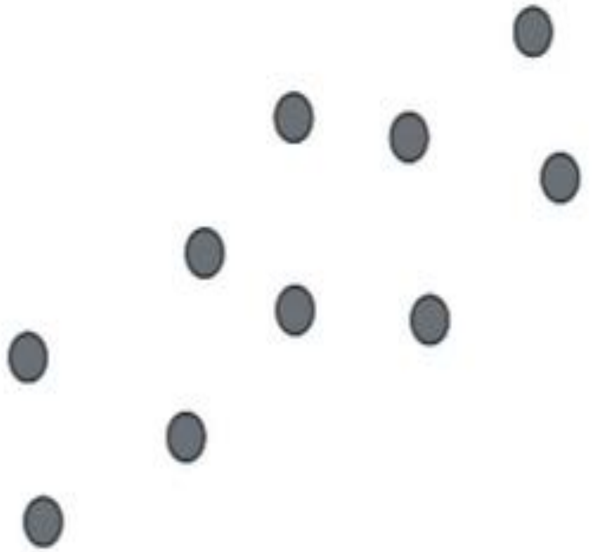
Lets say you want to take a picture of them, then you have to look for better angle….isn’t it??

Lets put the camera down to capture best possible picture to cover all of them...is it fine angle??

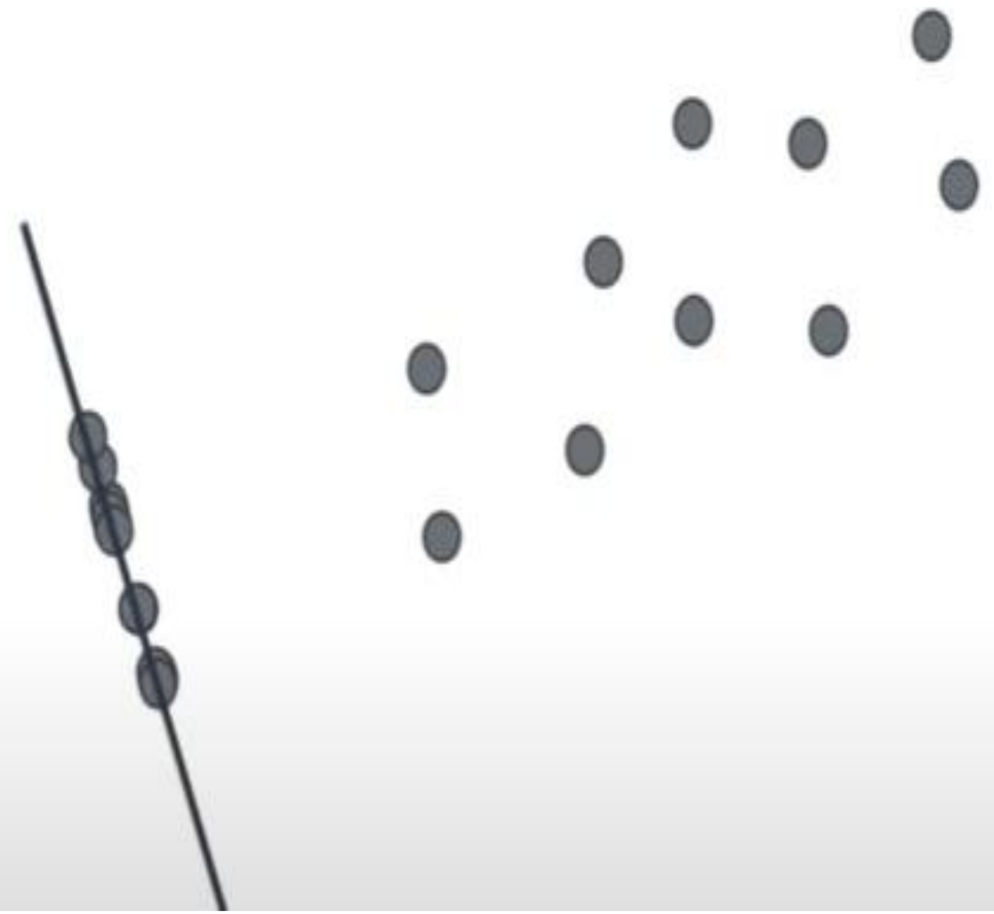


Hmmm…seems to be perfect place to shot all of them in a single frame.

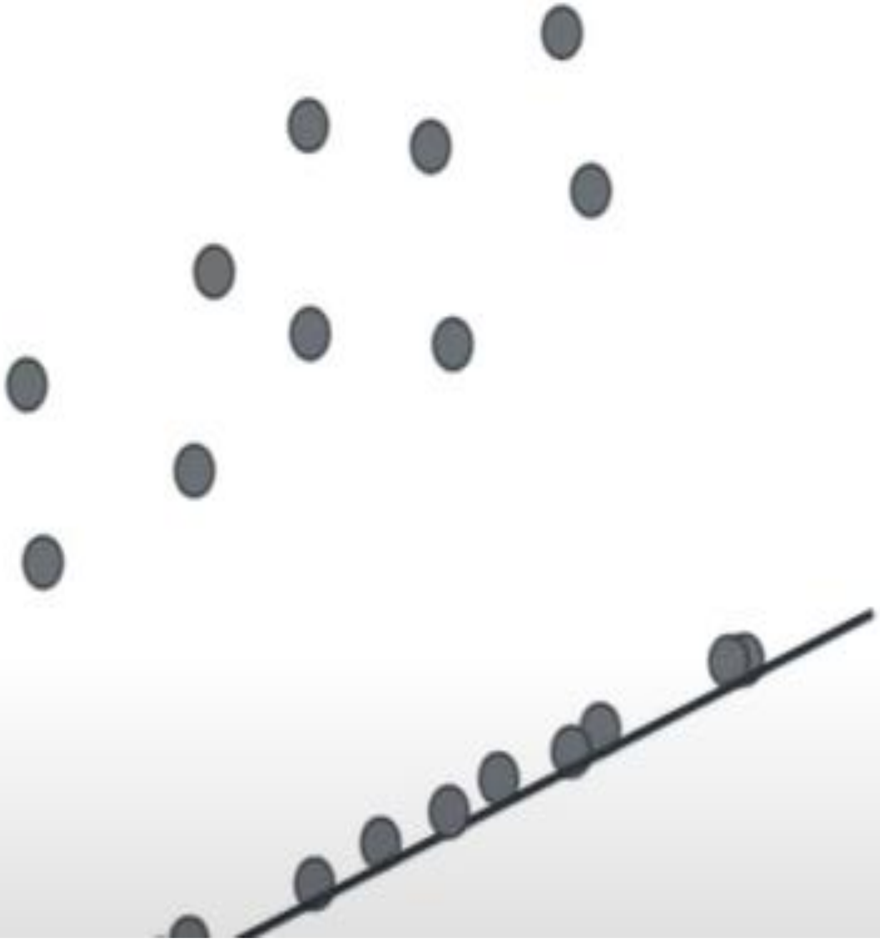
“This is nothing but PCA…capturing data images with best possible way”



Suppose this is your data spread and you want to capture the best possible way to it. What are the options?

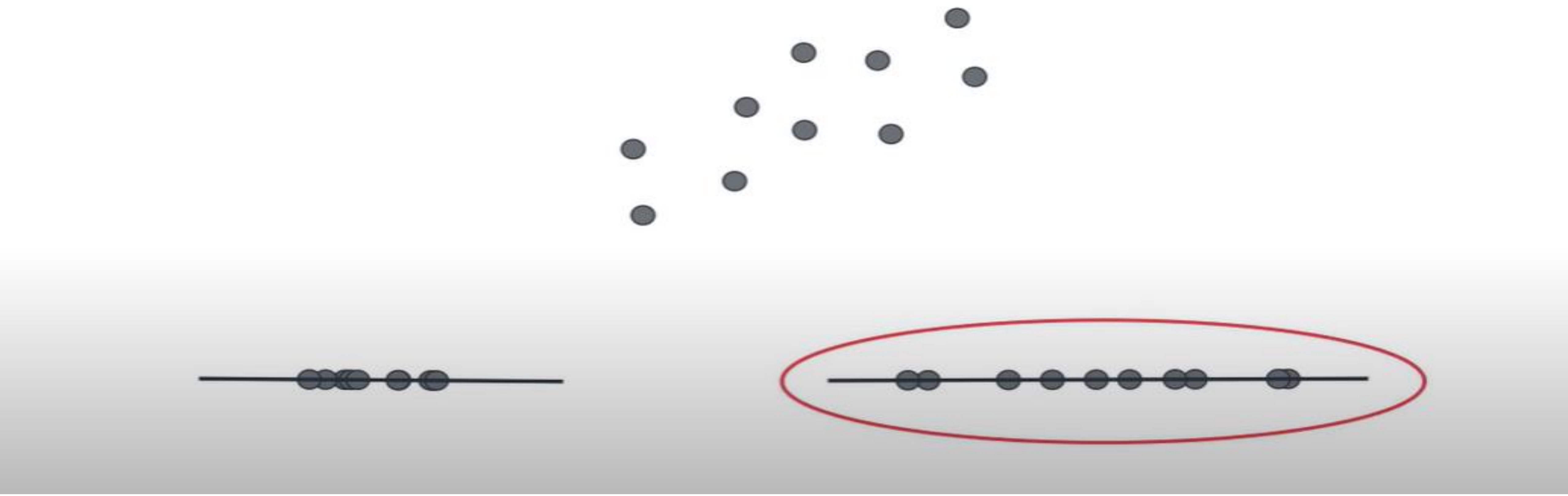


This is one way to capture the data on a single line but is it the best way??

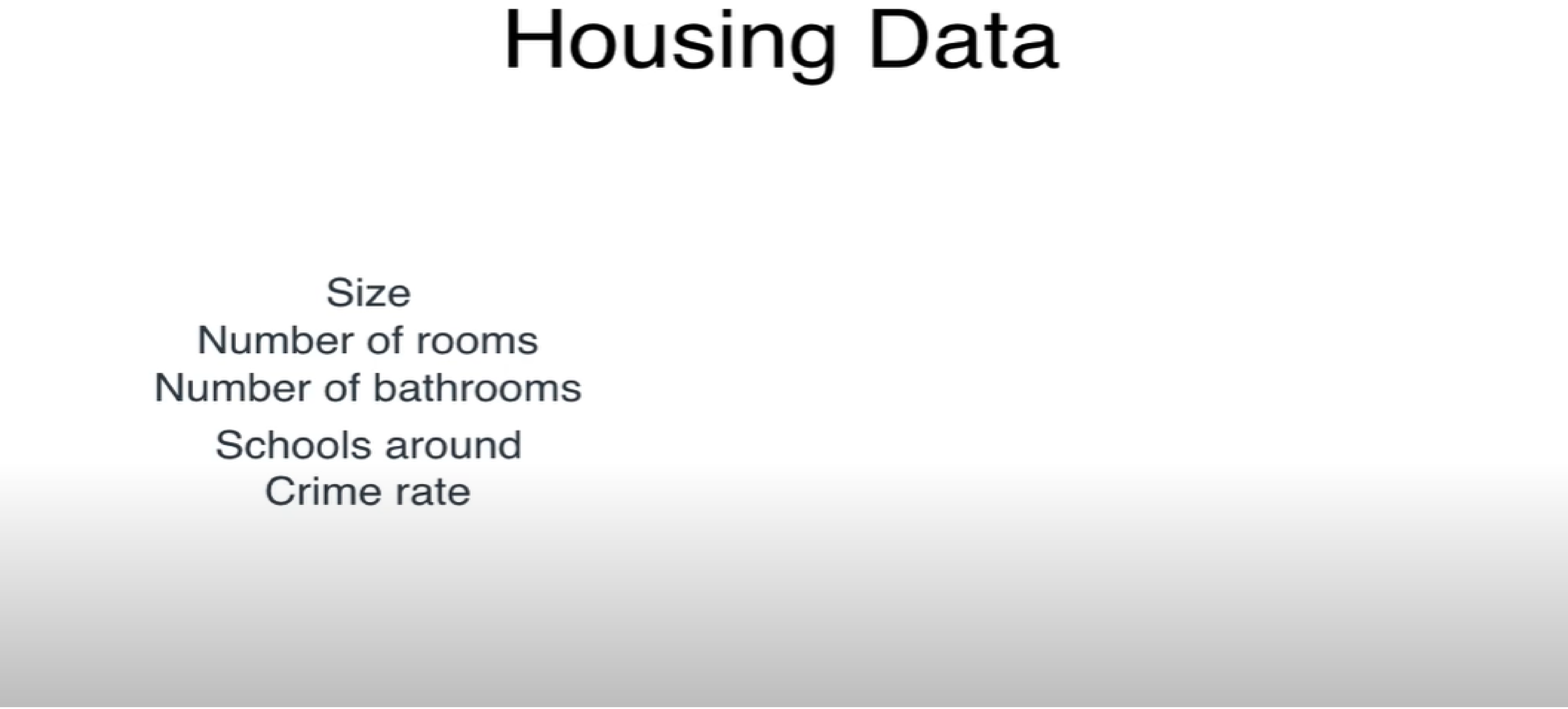


How about this line for capturing data effectively?

Dimensionality Reduction



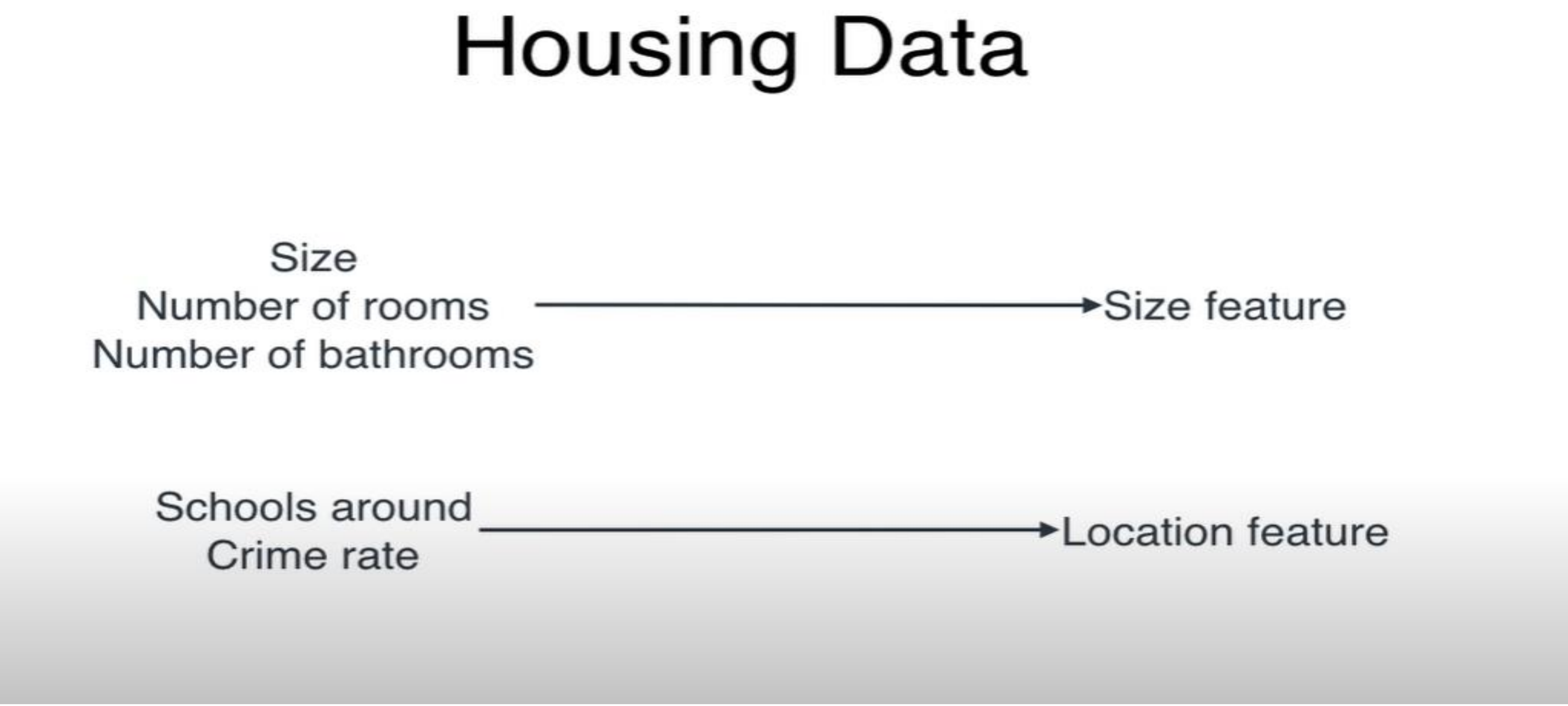
Here it’s clear that the right side plotting of capturing data covers all the points with spaces. So, that’s our best choice.



Housing Data

Size
Number of rooms
Number of bathrooms
Schools around
Crime rate

Lets say, we have a dataset with 5 attributes and we want to reduce it to 2 as part of PCA (dimensionality reduction)

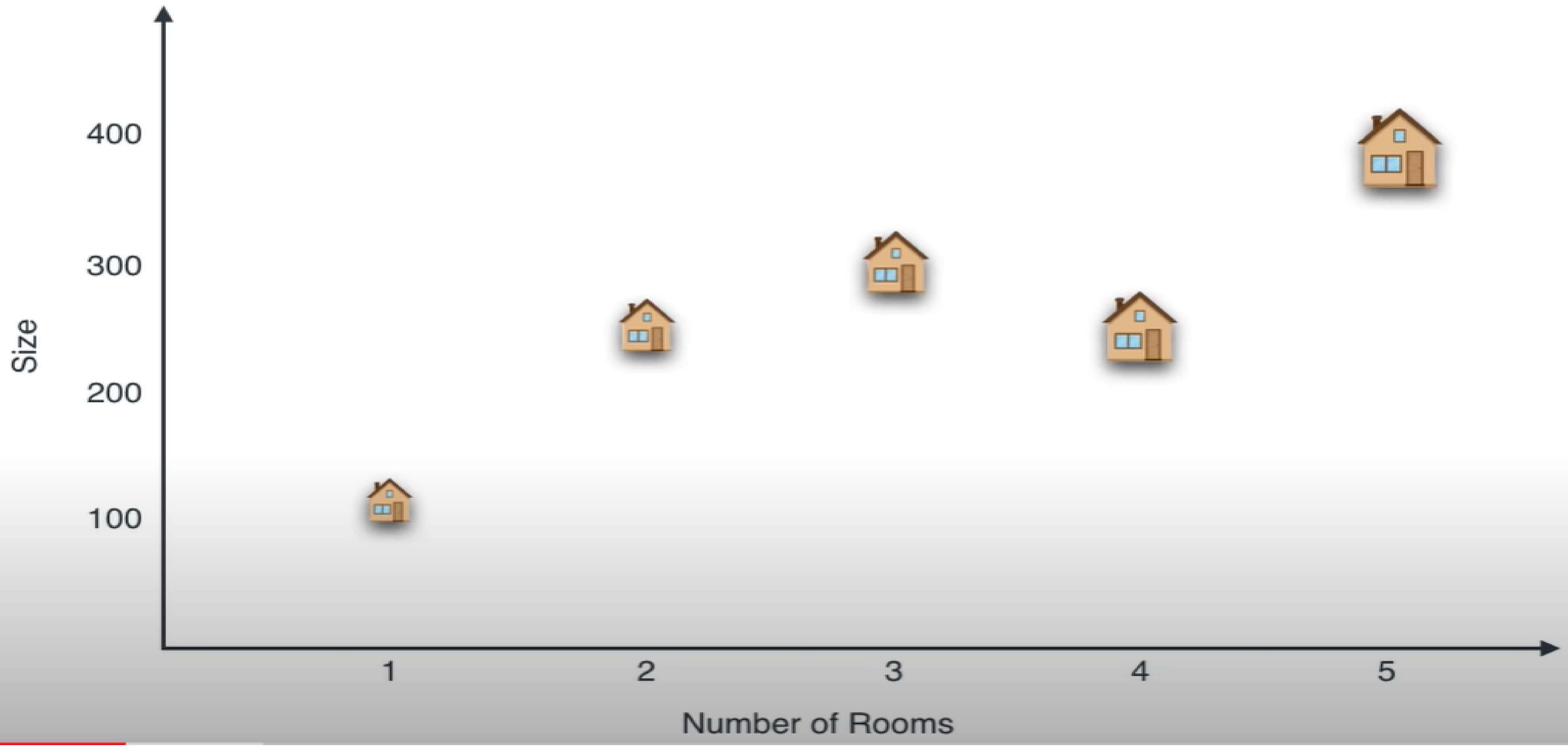


Housing Data

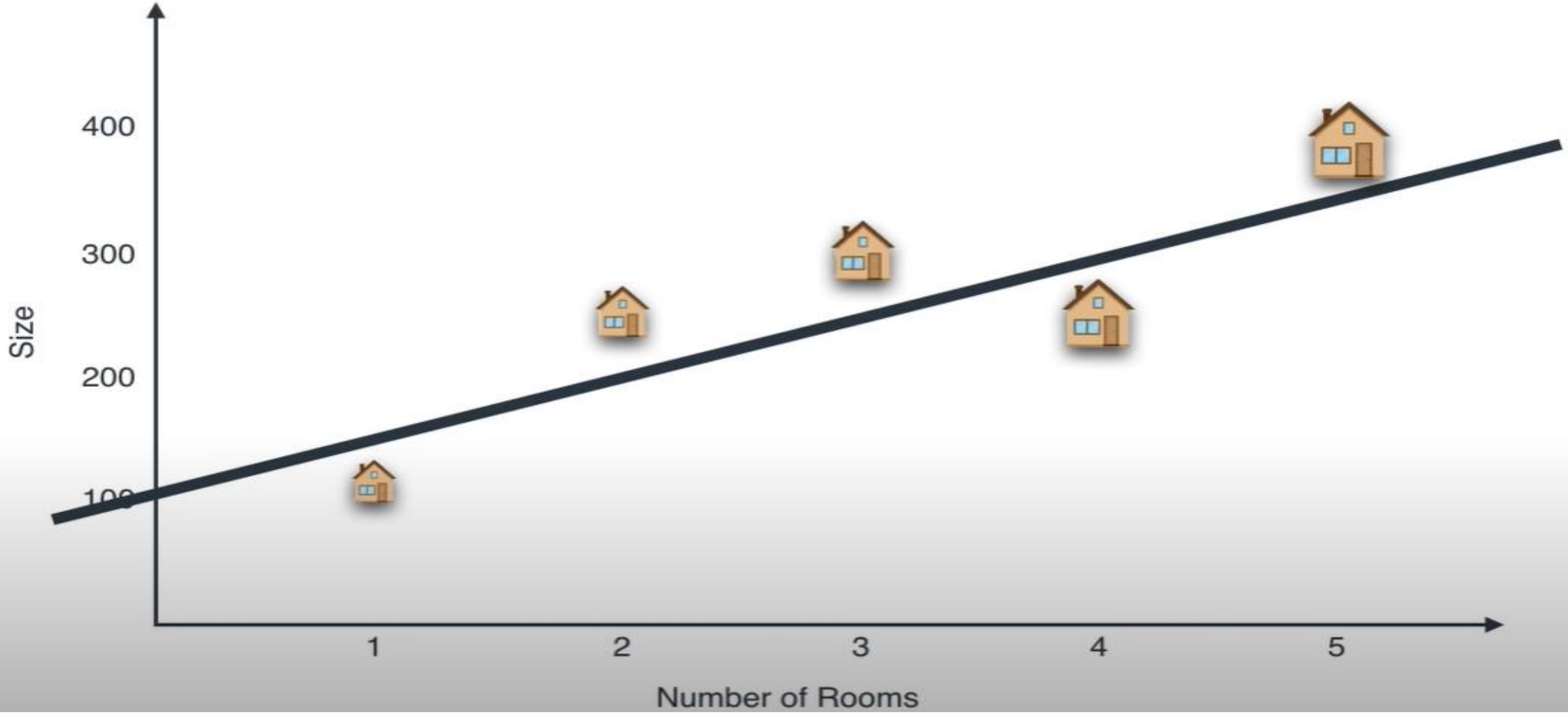
Size
Number of rooms
Number of bathrooms → Size feature

Schools around
Crime rate → Location feature

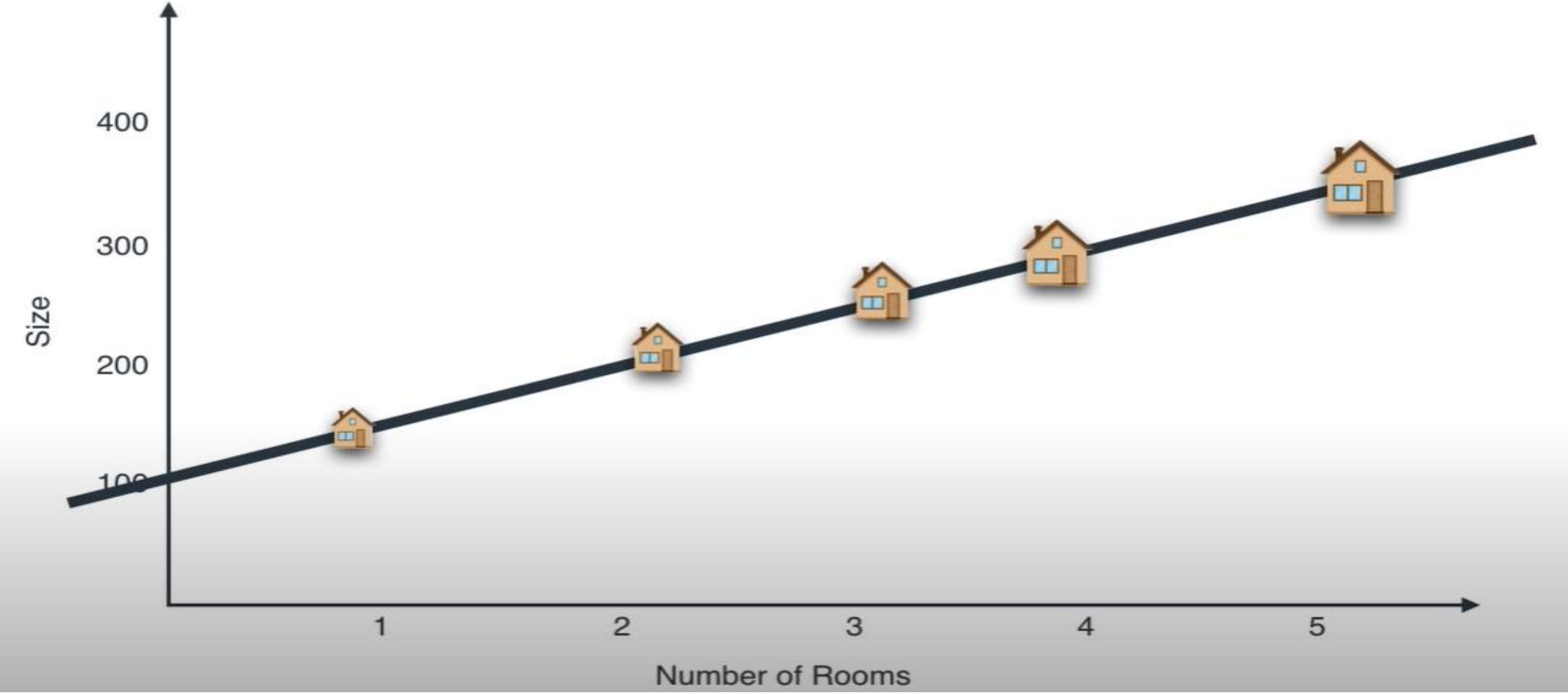
This is how we can achieve that from 5d to 2d



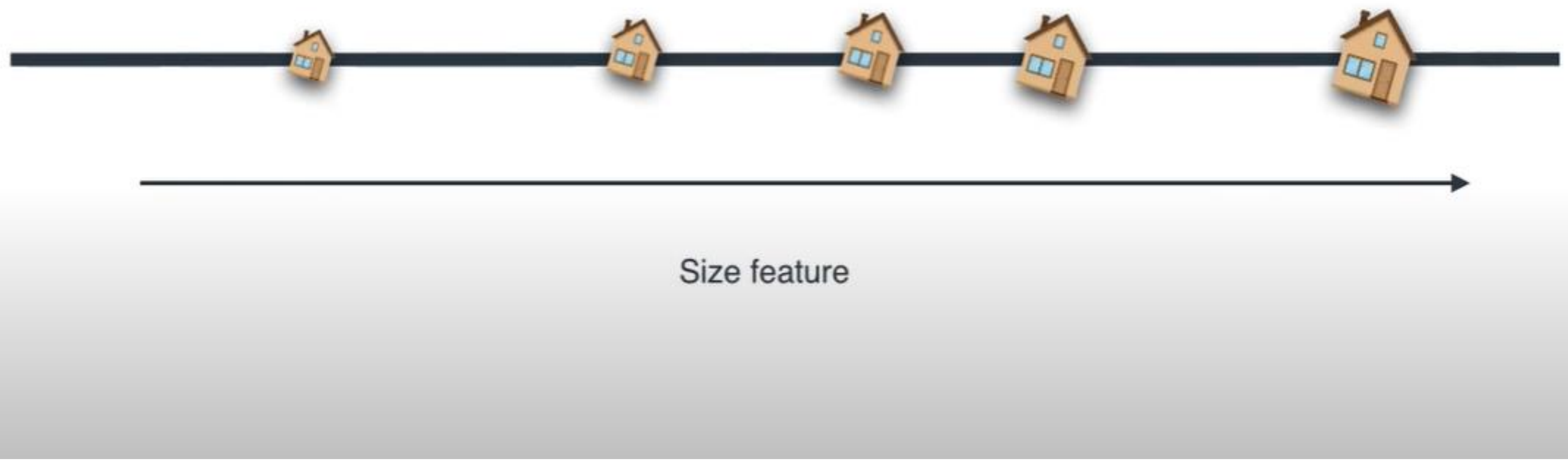
Lets plot graph with 2 attributes that we derived before.



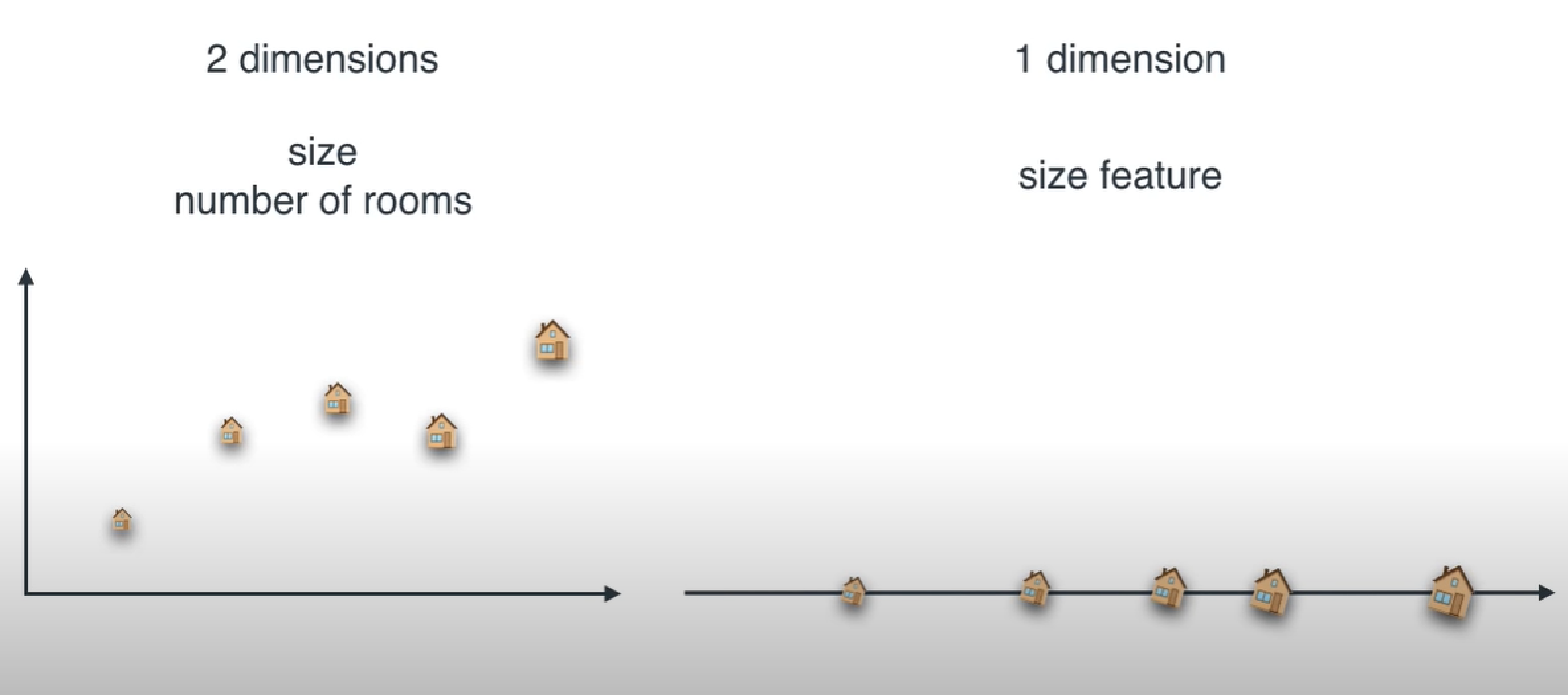
That is the best line to capture all the data points. NOTE: this is not linear regression line



Now draw perpendicular lines from data points to line and project the data points on that line.



Now it is simply one dimension.



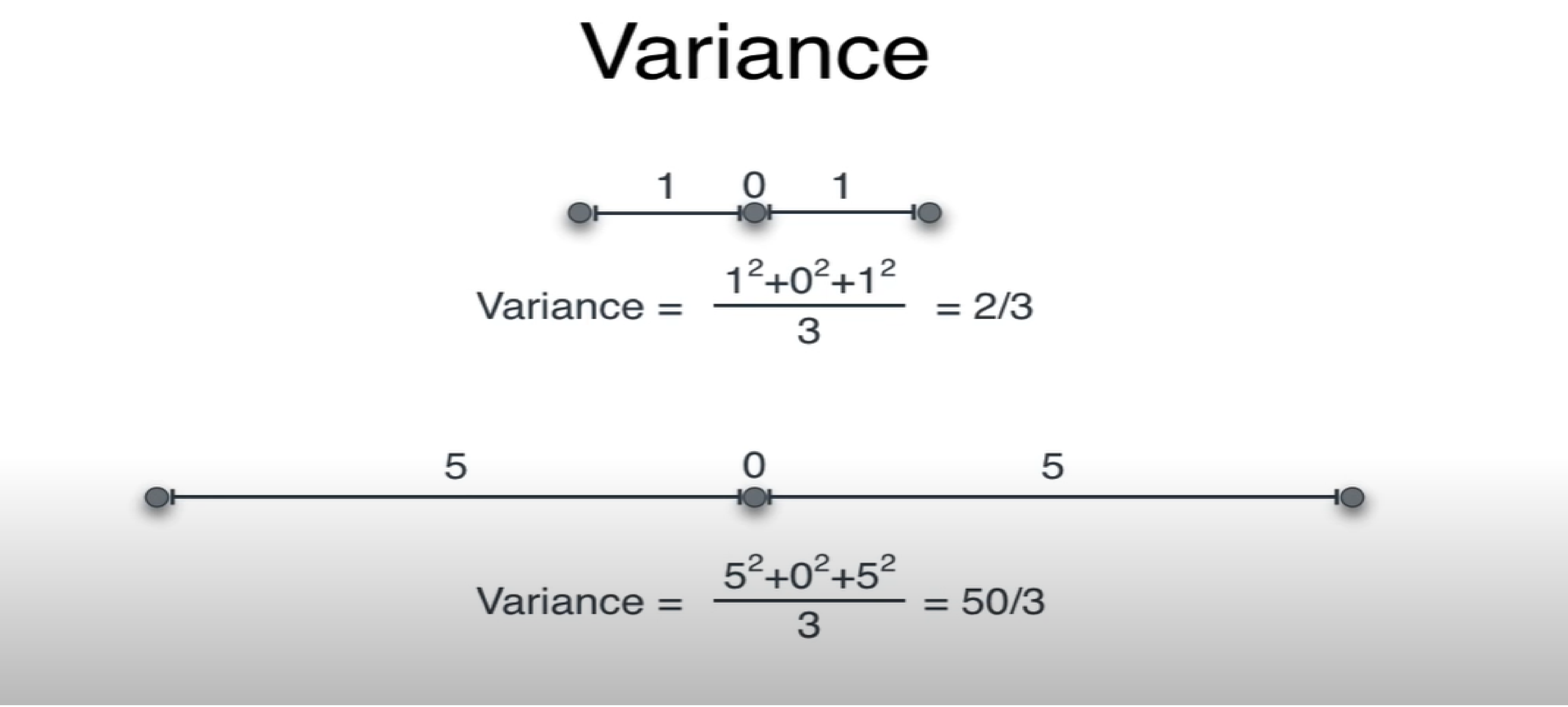
2 dimensions

size
number of rooms

1 dimension

size feature

This is how it will changes from 2d to 1d



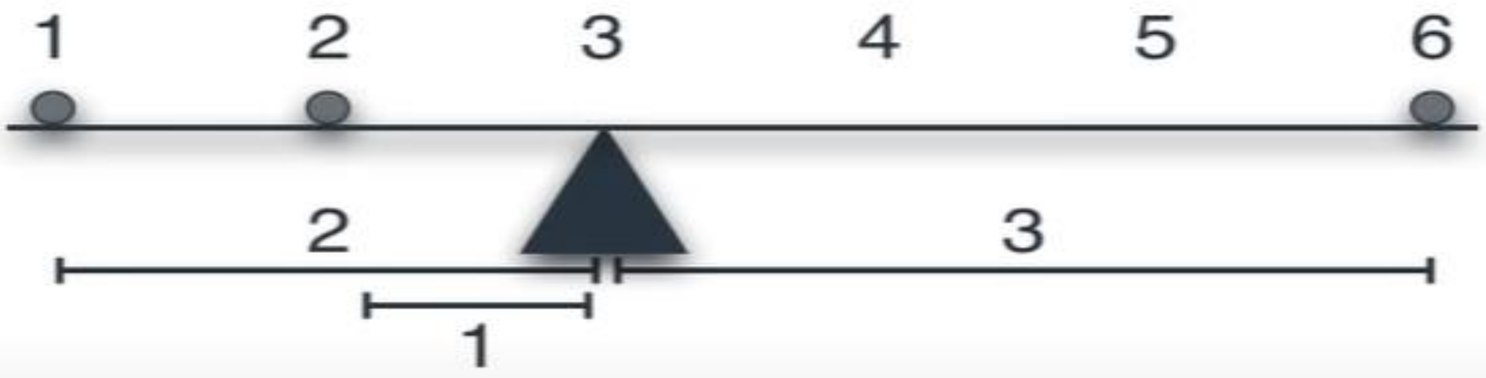
Variance

$$\text{Variance} = \frac{1^2+0^2+1^2}{3} = 2/3$$

$$\text{Variance} = \frac{5^2+0^2+5^2}{3} = 50/3$$

If you want to perform a variance between the points spread on space.

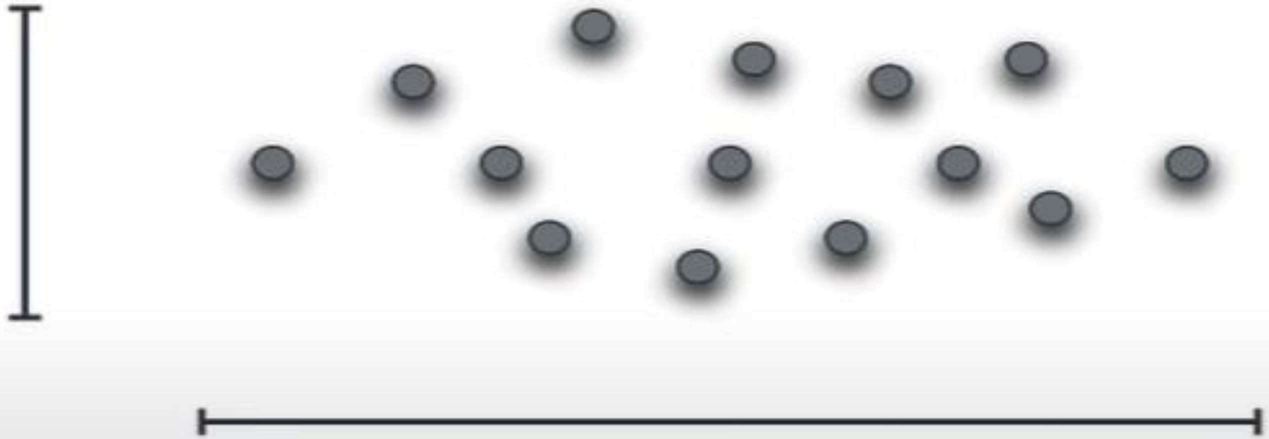
Mean



$$\text{Variance} = \frac{2^2+1^2+3^2}{3}$$

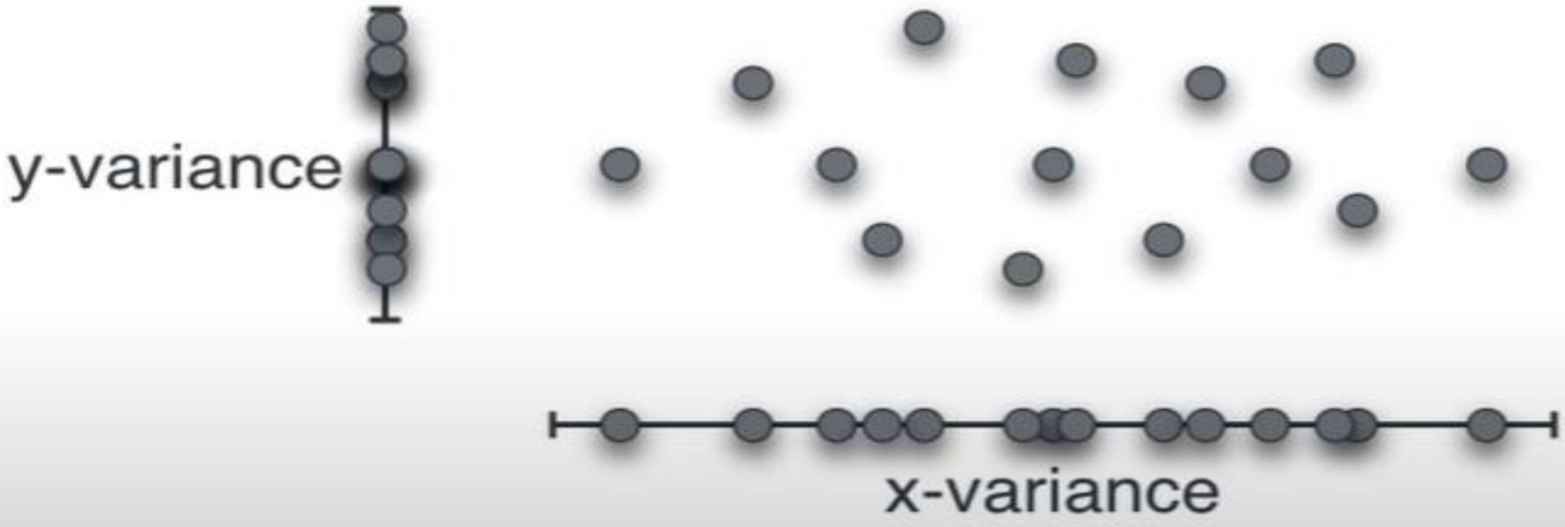
If you want to perform Mean and Variance of data points in a space.

Variance?

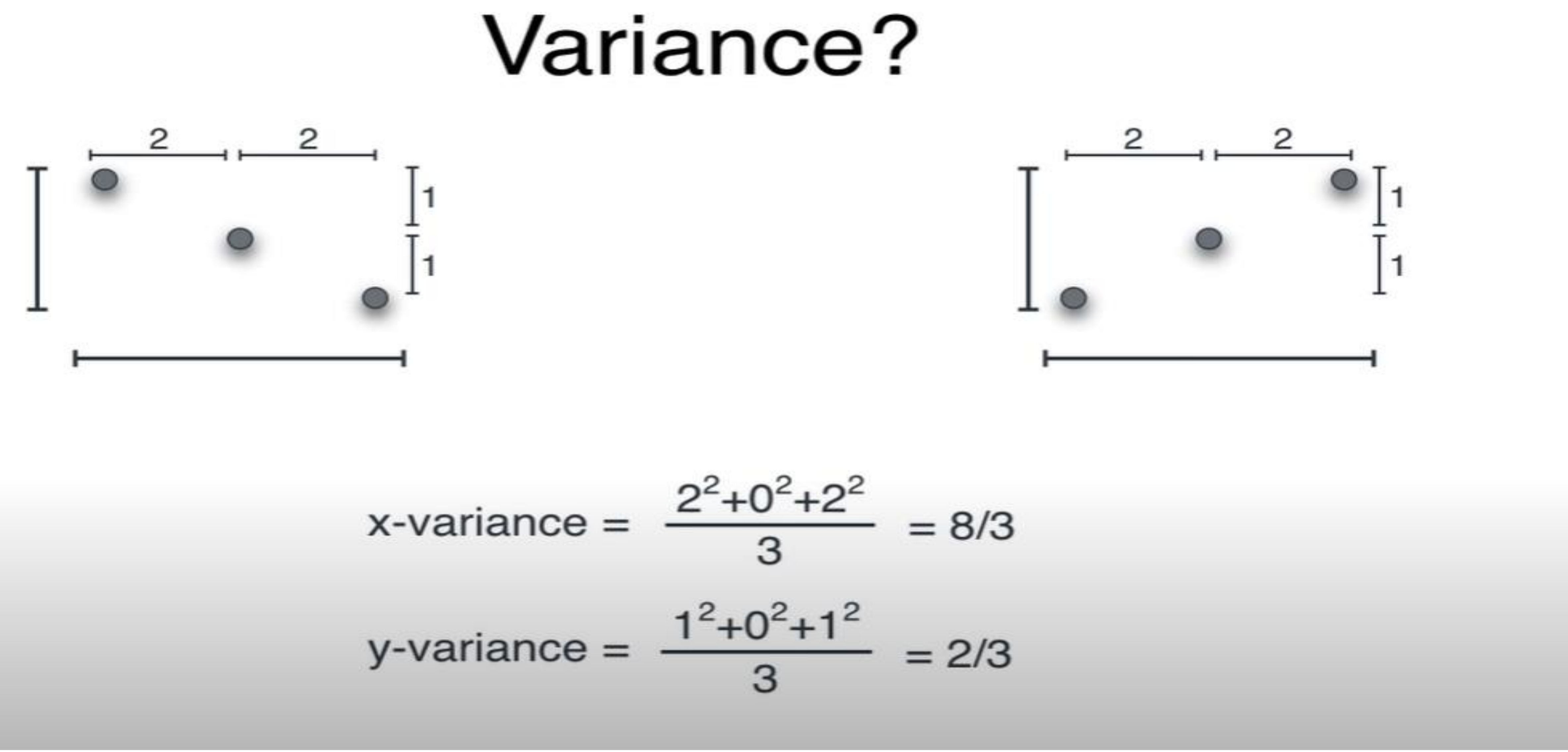


If you want to perform Variance or Spread of these kind of data points in a space then how can we do that?

Variance?



We have to calculate X-variance and Y-variance by projecting data points on respective axis.

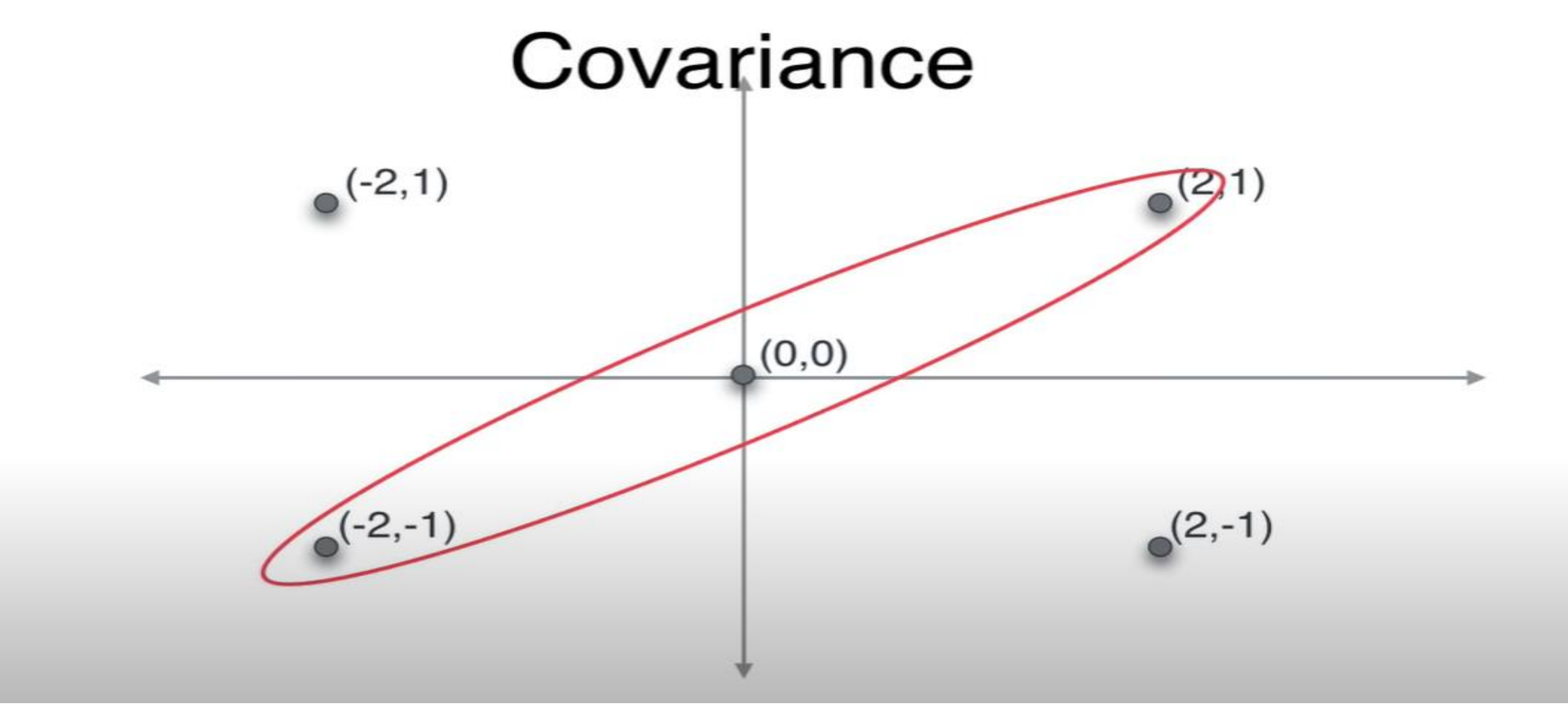


Variance?

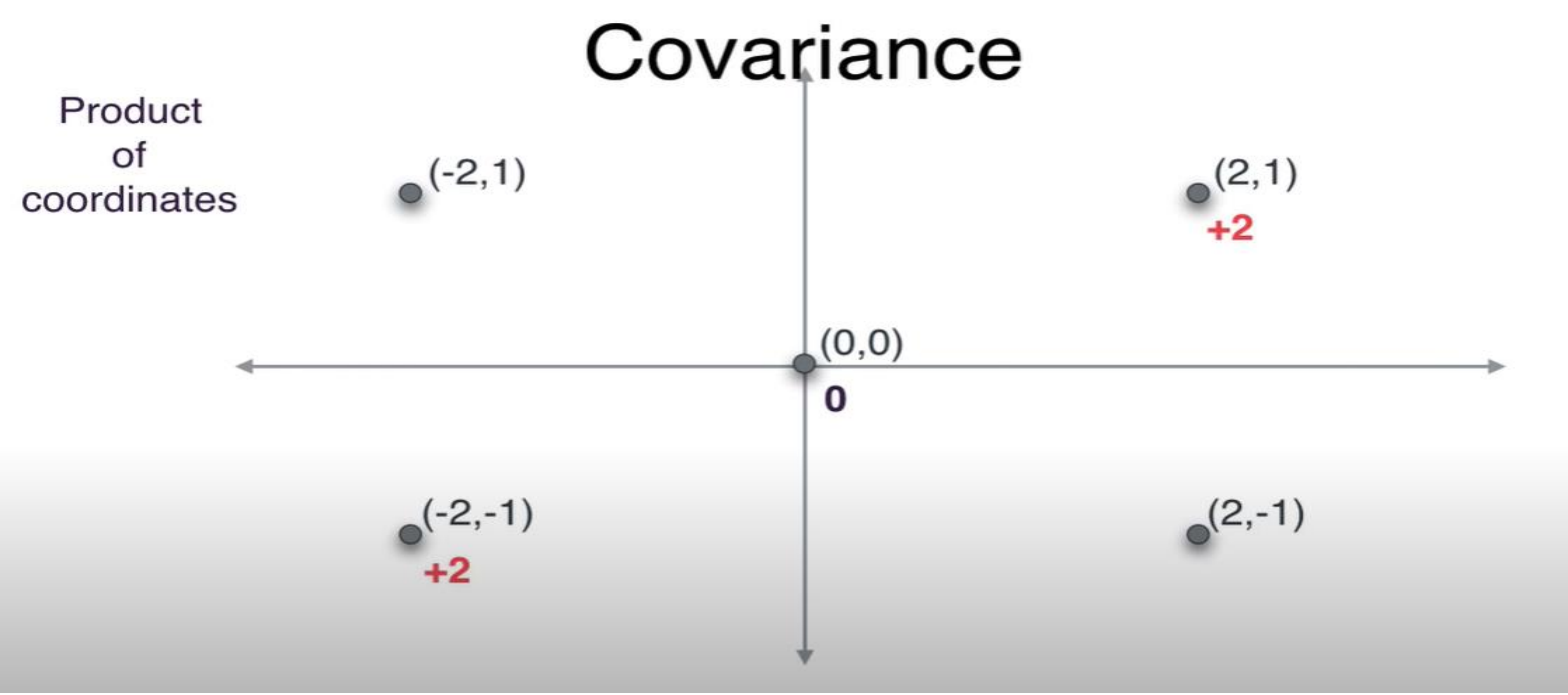
$$\text{x-variance} = \frac{2^2+0^2+2^2}{3} = 8/3$$

$$\text{y-variance} = \frac{1^2+0^2+1^2}{3} = 2/3$$

What if we have to calculate x-variance and y-variance of data points like this in free space. Even though they both have same variance but they are actually different. We need third quantity to tell that difference.

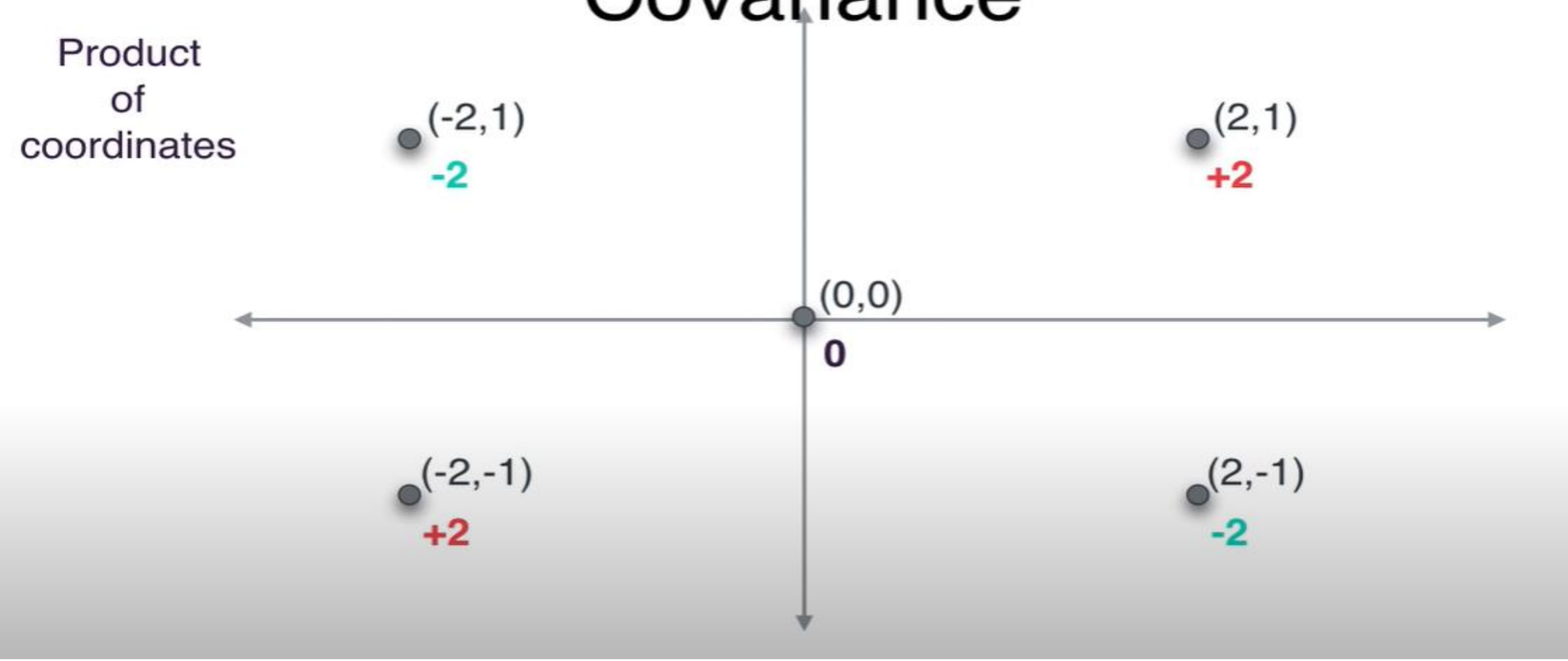


It is covariance which will tell you the difference of those points in free space.



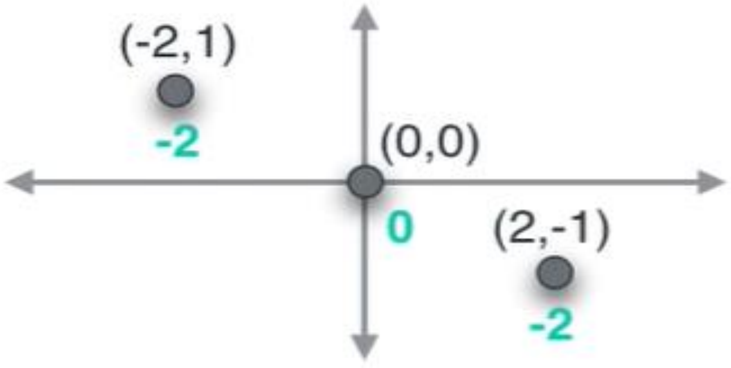
This is how we start the things to show the difference…Product of Coordinates.

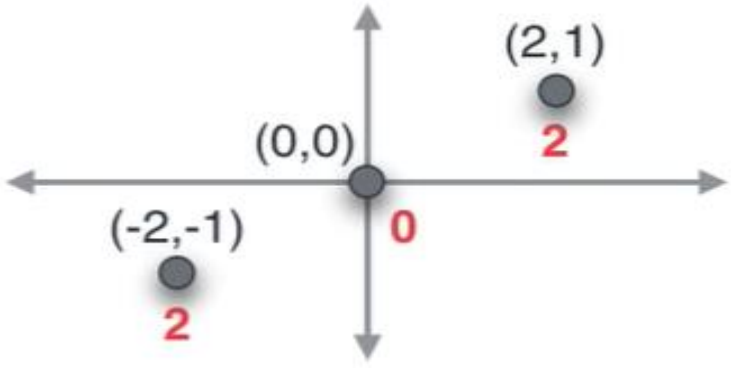
Covariance



Do the same thing for other points in space coordinate.

Covariance

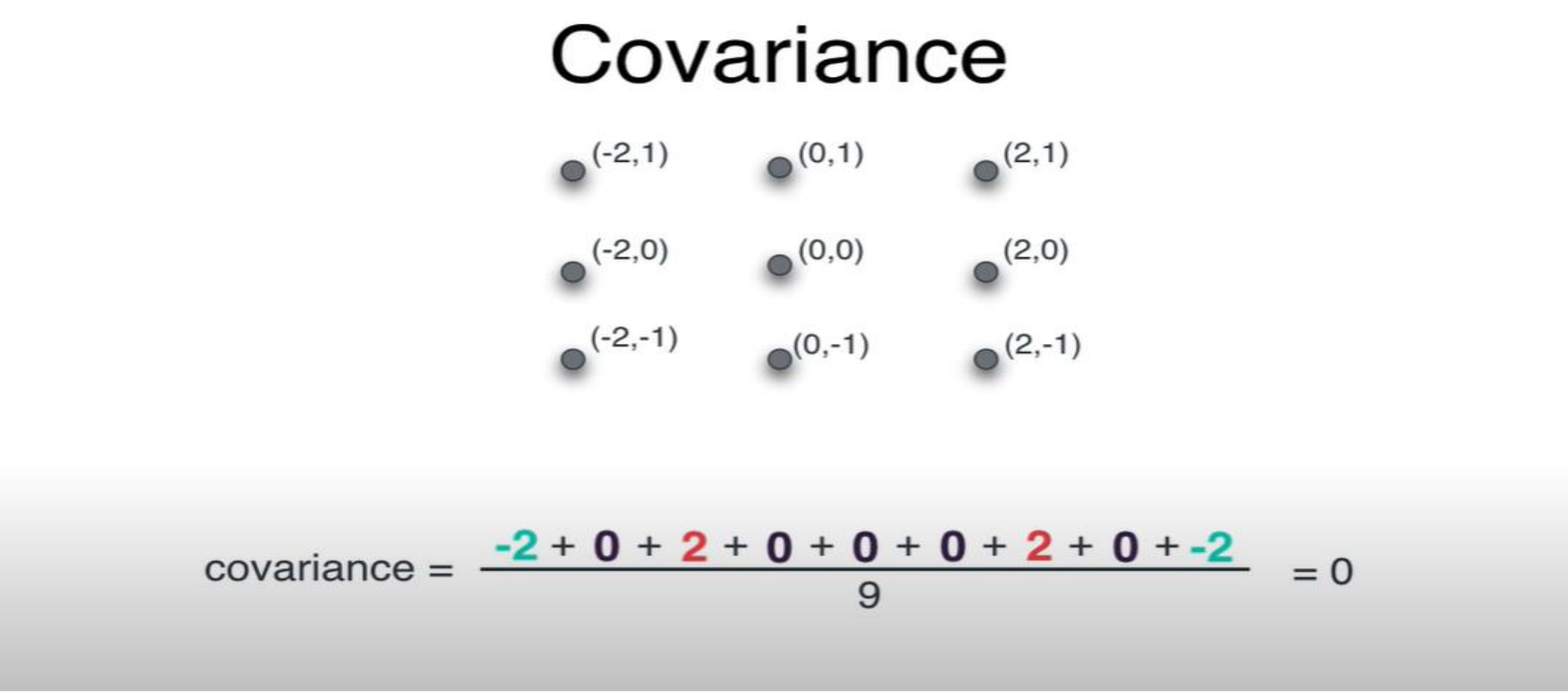




$$\text{covariance} = \frac{(-2) + 0 + (-2)}{3} = -4/3$$

$$\text{covariance} = \frac{2 + 0 + 2}{3} = 4/3$$

Covariance : sum of the products of coordinates. Now you can see the difference between positive and negative covariance.

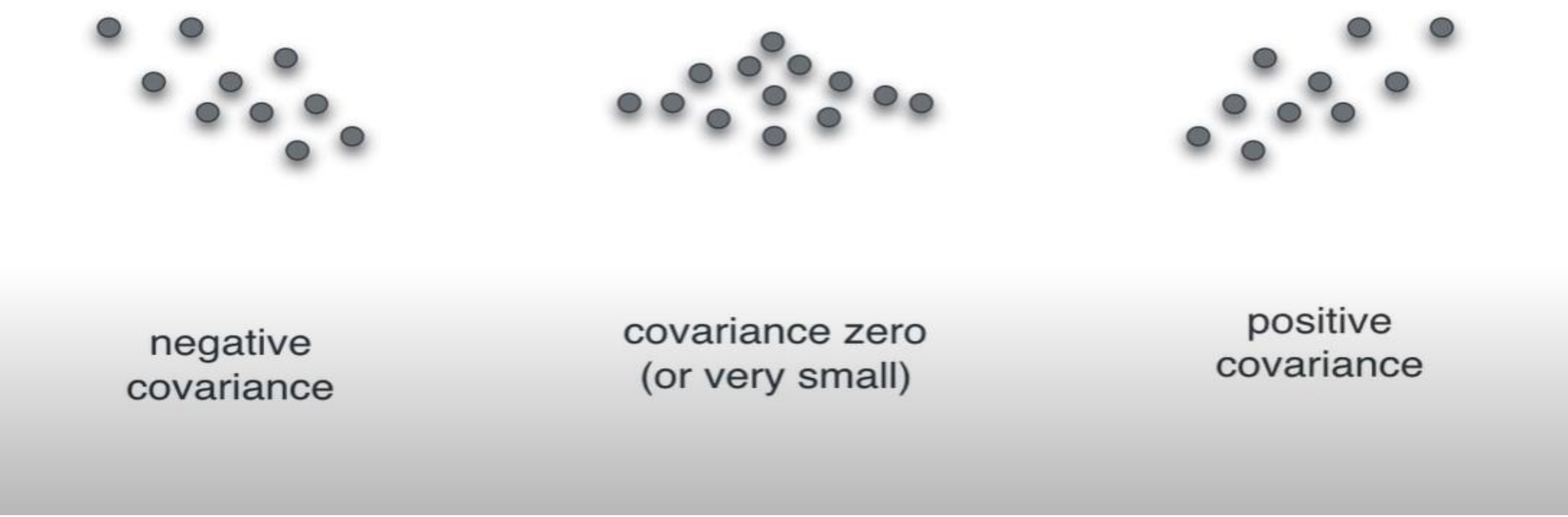


Covariance

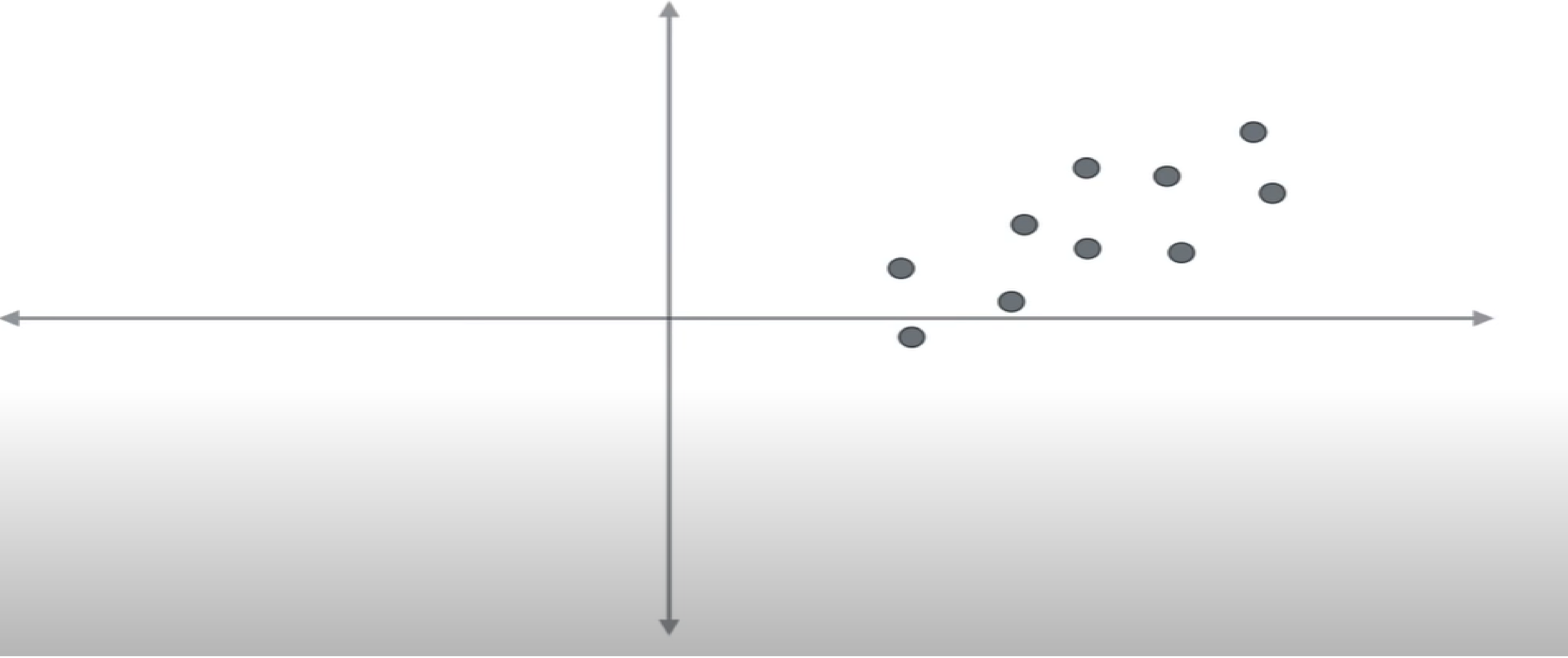
$$\text{covariance} = \frac{-2 + 0 + 2 + 0 + 0 + 0 + 2 + 0 + -2}{9} = 0$$

What is the covariance of equally spaced data points?

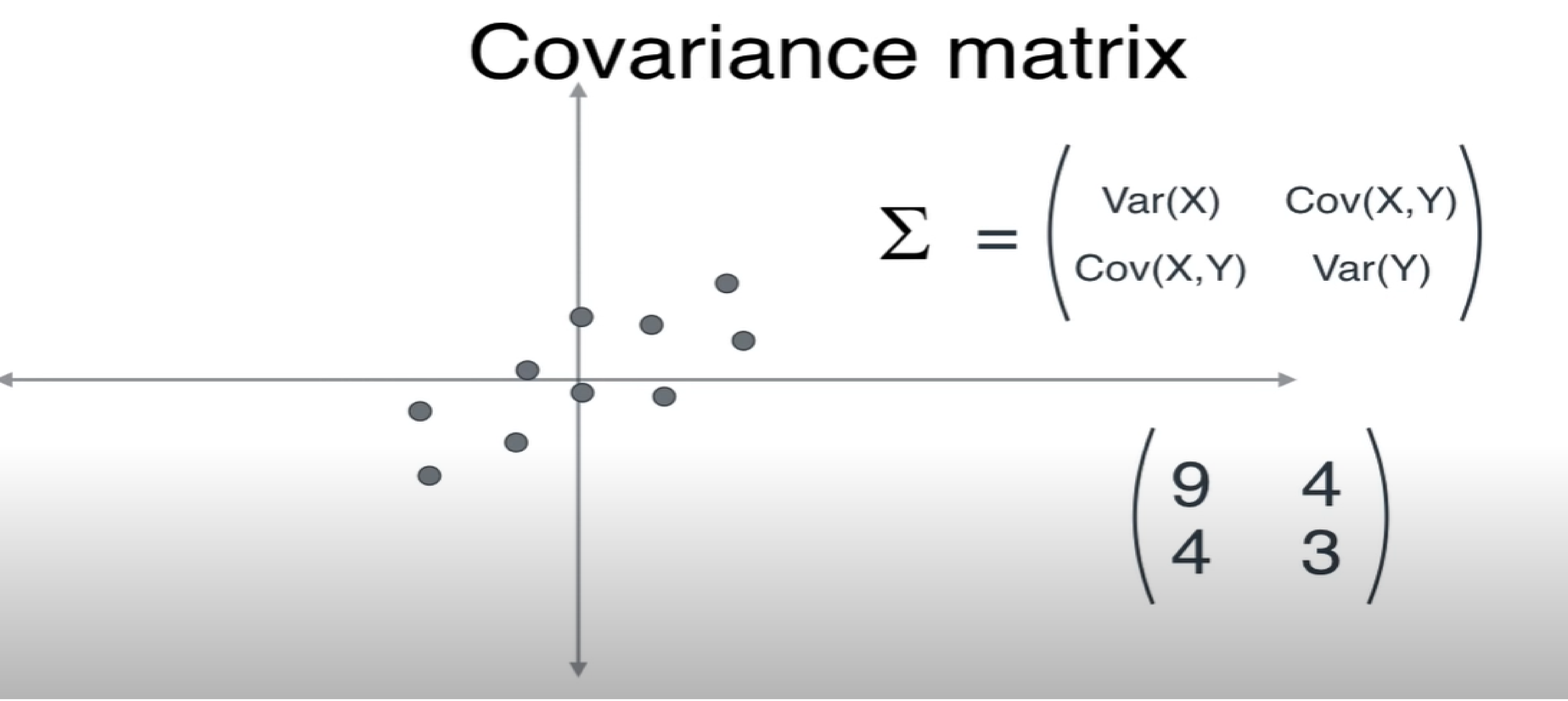
Covariance



Guess what kind of covariance will exhibit by above points in space?



Lets do PCA analysis on these data points. First randomly assign the position of data points on coordinate system.

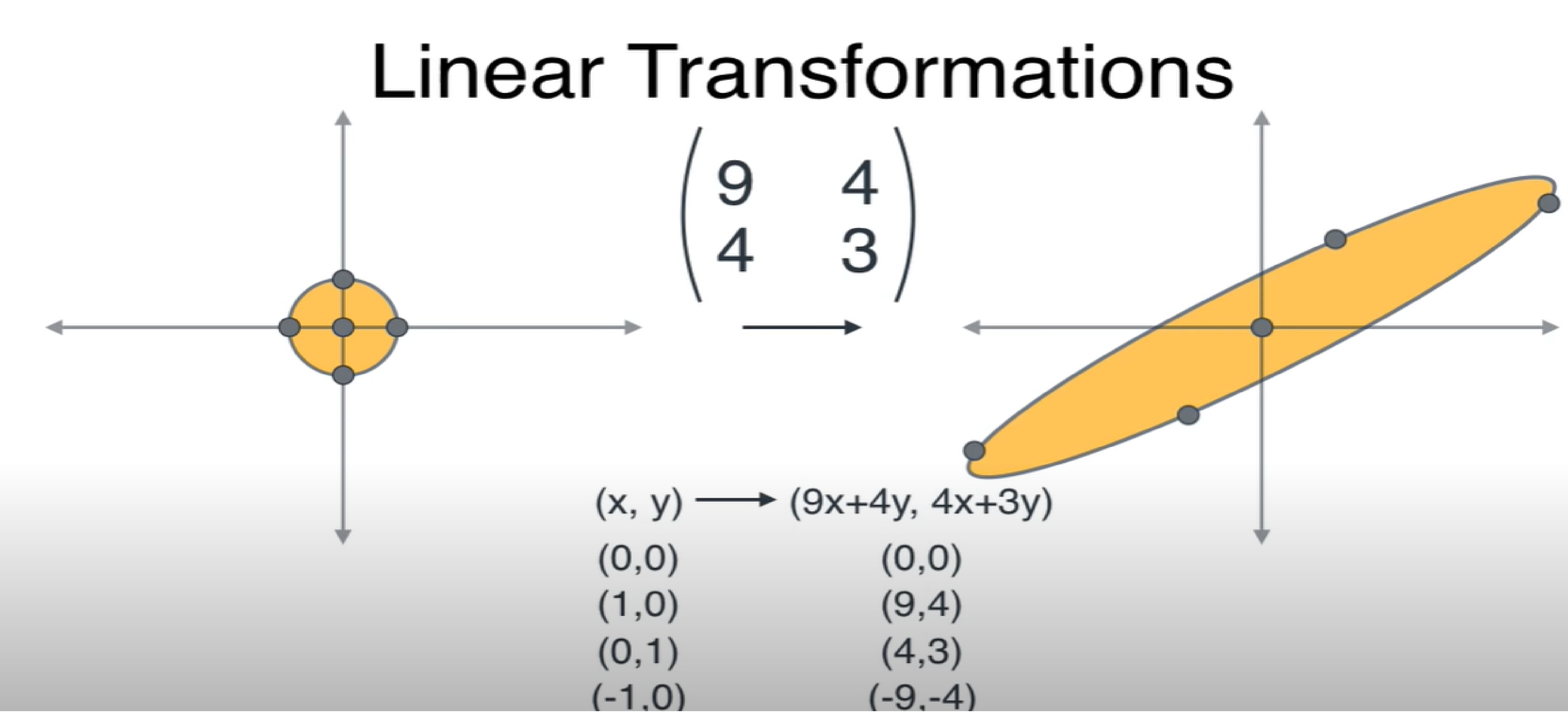


Covariance matrix

$$\Sigma = \begin{pmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(X,Y) & \mathrm{Var}(Y) \end{pmatrix}$$

$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

Now we have to create a covariance matrix and lets assume (9,4,4,3) as that matrix. See it’s a positive covariance of 4.



Linear Transformations

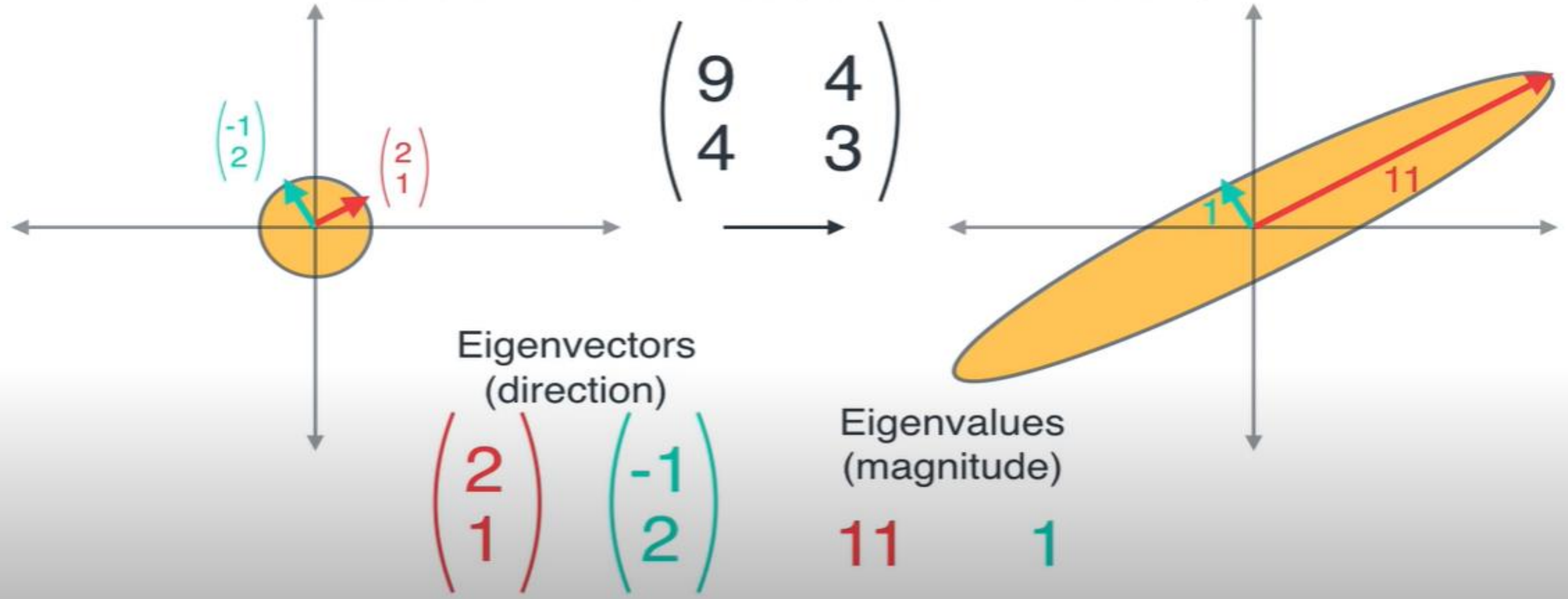
$$\begin{pmatrix}9 & 4\\4 & 3\end{pmatrix} \longrightarrow$$

$$(x, y) \longrightarrow (9x+4y, 4x+3y)$$

(0,0)	(0,0)
(1,0)	(9,4)
(0,1)	(4,3)
(-1,0)	(-9,-4)

Now do the linear transformation from one space to coordinate system of line equations.

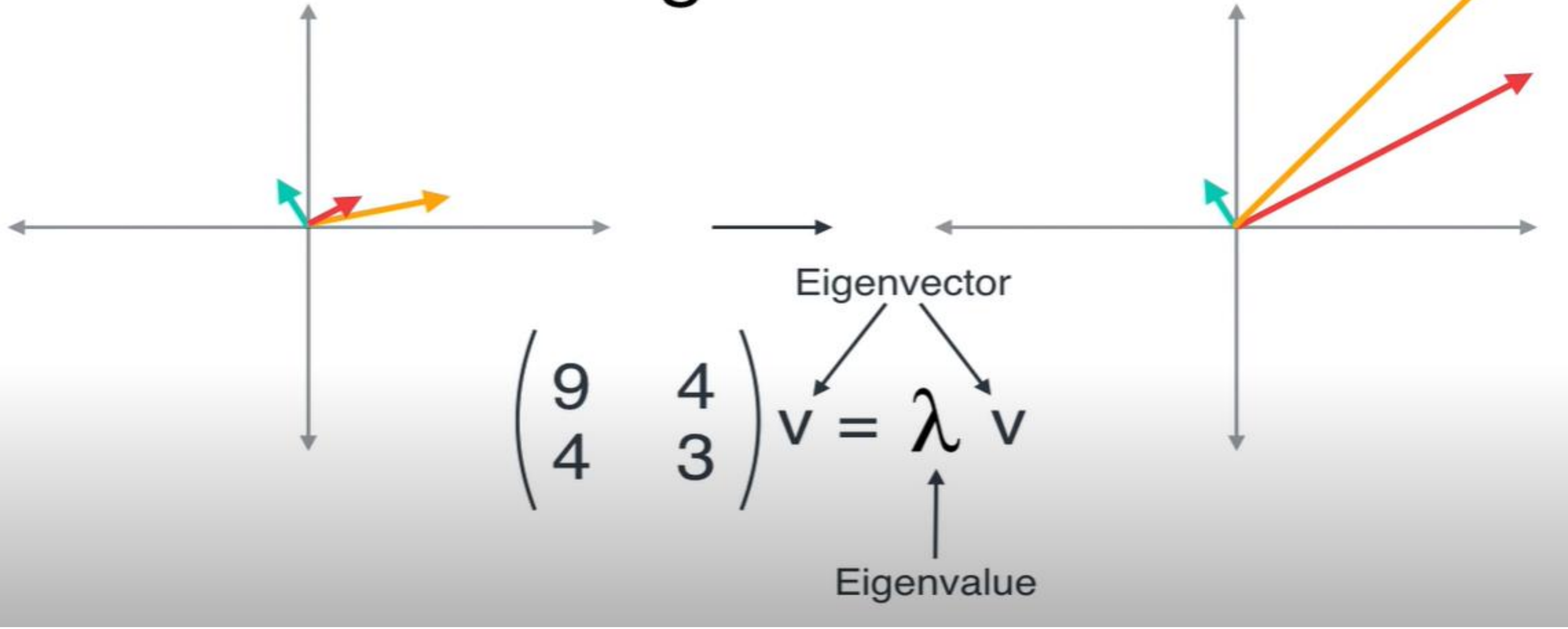
Linear Transformations



$\begin{pmatrix}-1\\2\end{pmatrix}$
$\begin{pmatrix}2\\1\end{pmatrix}$
$\begin{pmatrix}9 & 4\\4 & 3\end{pmatrix}$
11
1
Eigenvectors
(direction)
$\begin{pmatrix}2\\1\end{pmatrix}$
$\begin{pmatrix}-1\\2\end{pmatrix}$
Eigenvalues
(magnitude)
11
1

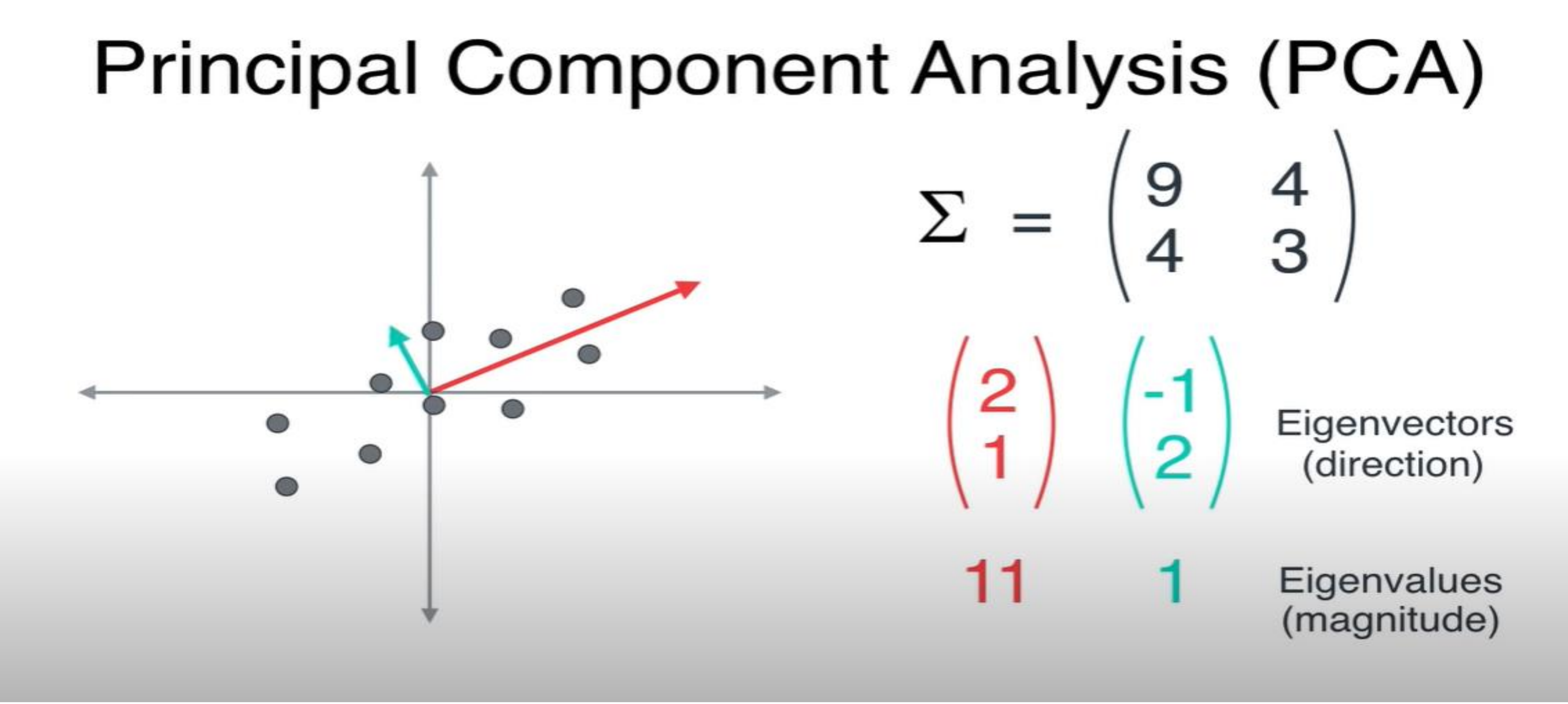
For PCA, we have to calculate covariance matrix , eigen vectors and eigen values. The eigen vectors in a free space will convert into eigen values as linear transformation. Remember those are orthogonal in nature.

Eigenstuff



$$\begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix} v = \lambda \; v$$

Generally if you perform linear transformation for normal vectors it will behave different as shown in orange line but in order to achieve same direction and spread we have to multiply with eigen vector both sides will do the needful.



Principal Component Analysis (PCA)

$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

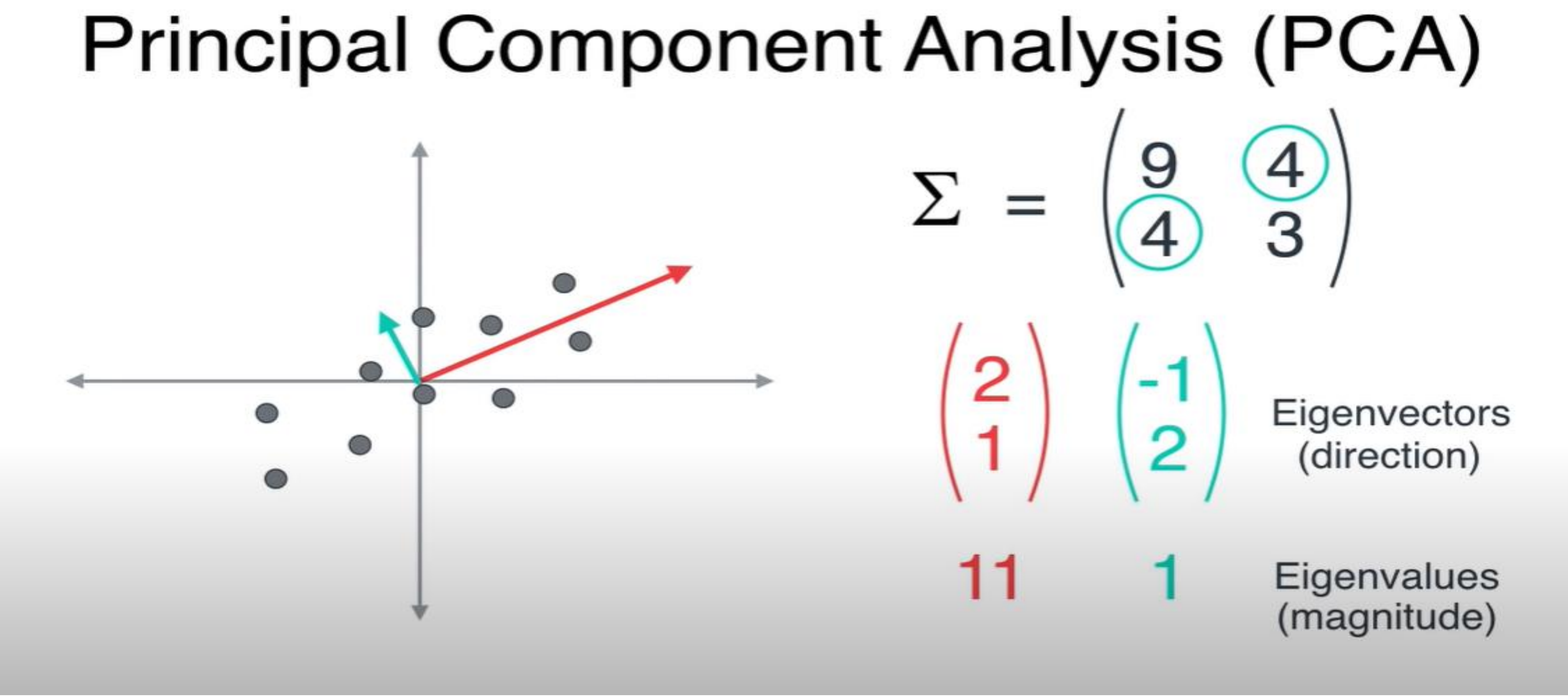
$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$

Eigenvectors (direction)

$$11 \quad 1$$

Eigenvalues (magnitude)

From this it is so clear that eigenvectors will give direction and eigenvalues will give magnitude.



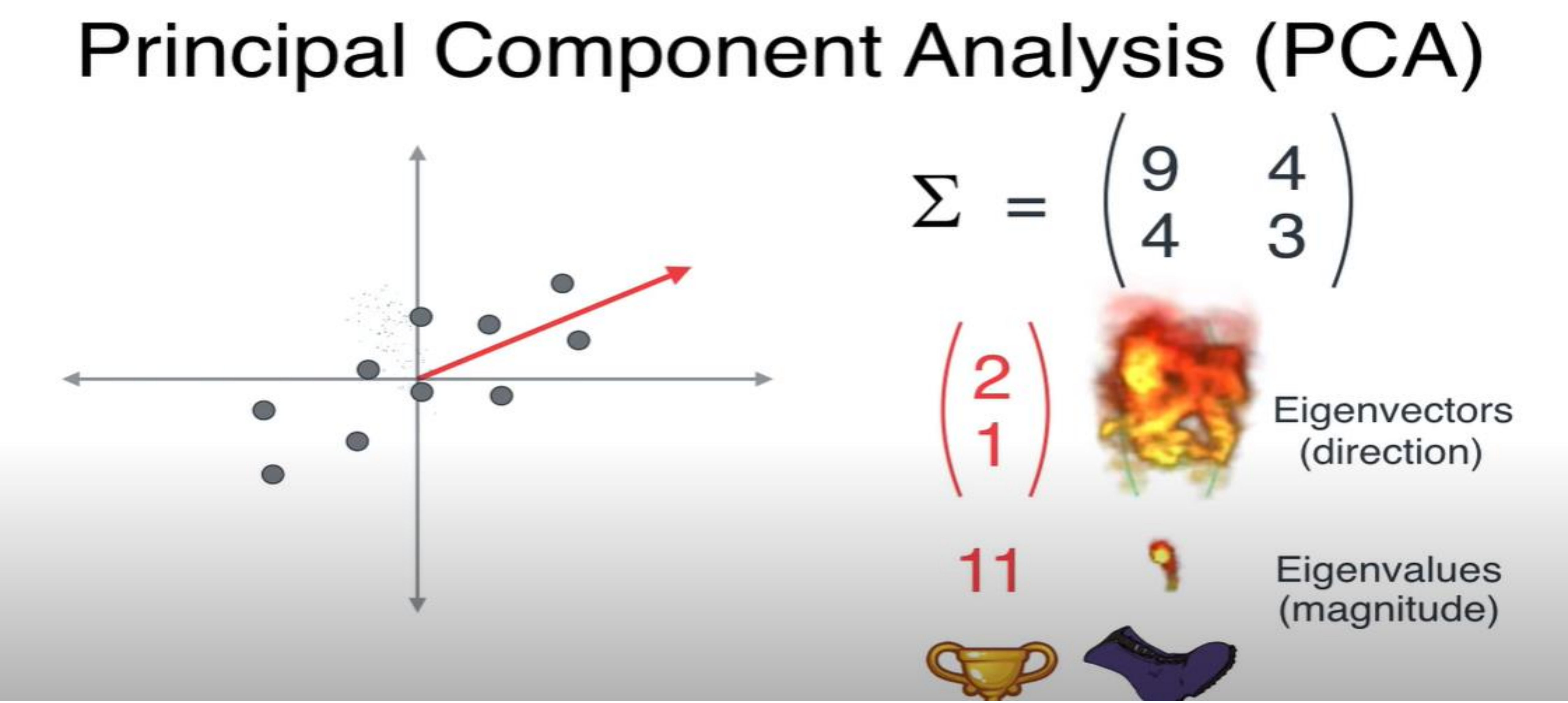
Principal Component Analysis (PCA)

$$\Sigma = \begin{pmatrix} 9 & \boxed{4} \\ \boxed{4} & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix} \quad \begin{pmatrix} -1 \\ 2 \end{pmatrix}$$ Eigenvectors (direction)

$11 \quad 1$ Eigenvalues (magnitude)

Its not luck that two vectors are orthogonal but because of symmetry matrix. ..i.e $A^T = A$ | $A^T – A^{-1} = I$ (identity matrix)



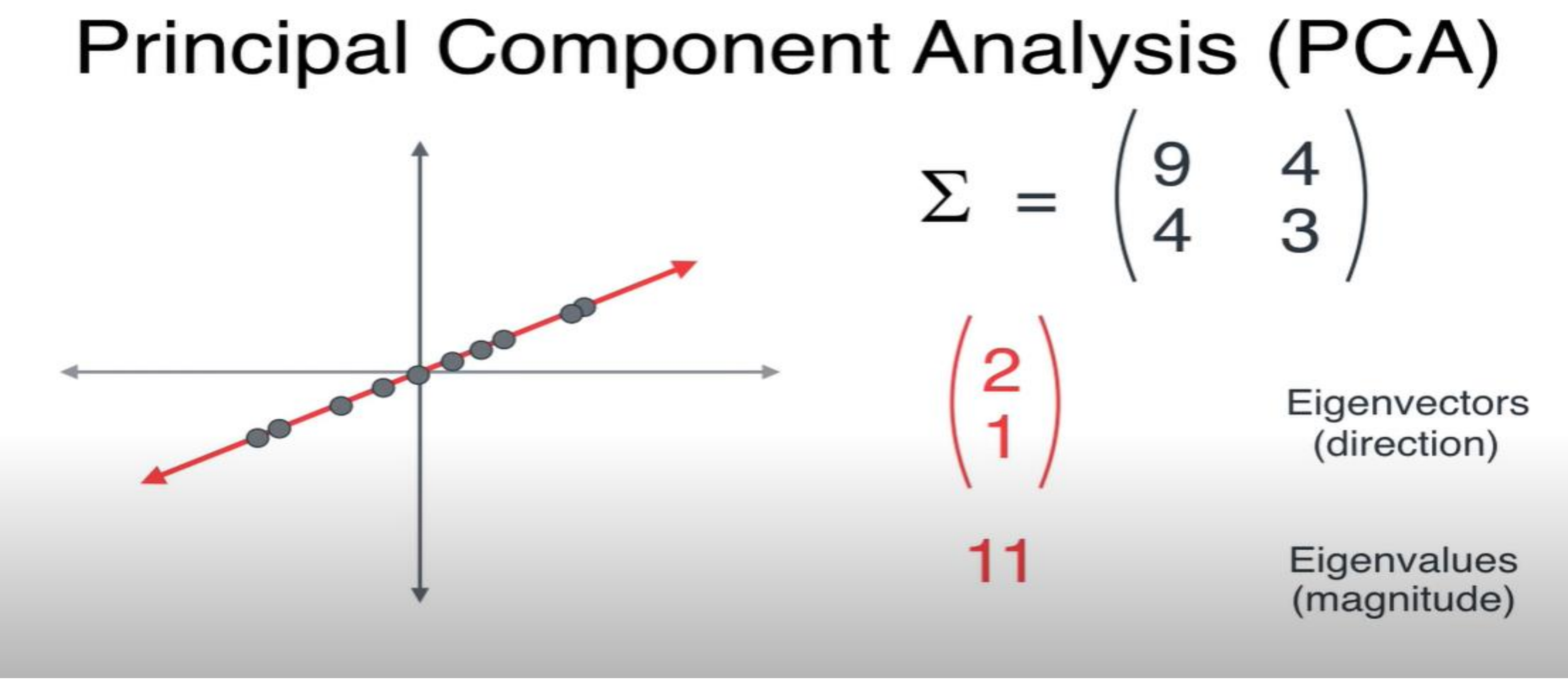
Principal Component Analysis (PCA)

$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$$ Eigenvectors (direction)

11 Eigenvalues (magnitude)

Here is the catch, which eigen value(principal component) is important?. With large value of course wins because that will explain the maximum input variance.



Principal Component Analysis (PCA)

$$\Sigma = \begin{pmatrix} 9 & 4 \\ 4 & 3 \end{pmatrix}$$

$$\begin{pmatrix} 2 \\ 1 \end{pmatrix}$$

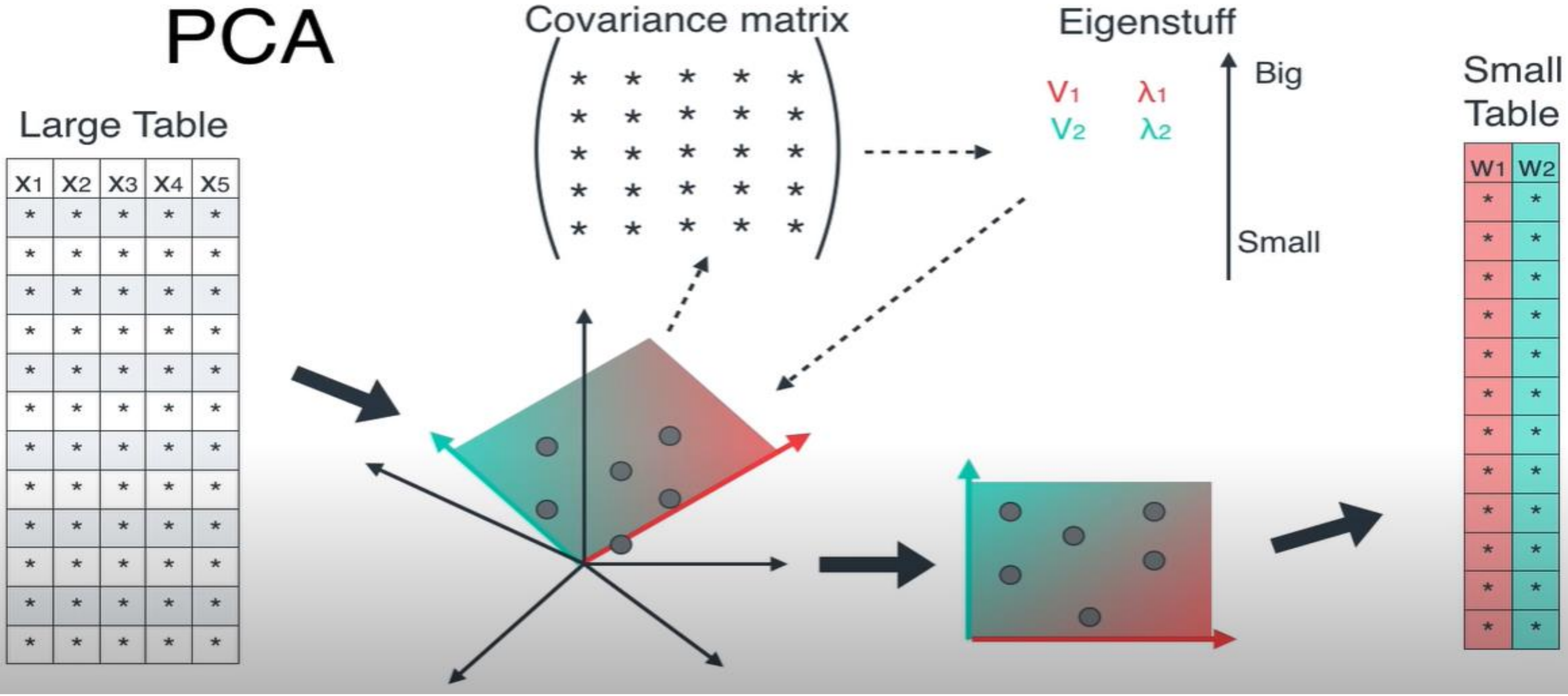
Eigenvectors (direction)

11

Eigenvalues (magnitude)

We found which component was important. So, lets project all points onto that line and that’s our principal component.

PCA



Large Table

X1	X2	X3	X4	X5
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*
*	*	*	*	*

Covariance matrix

$$\begin{pmatrix} * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \\ * & * & * & * & * \end{pmatrix}$$

Eigenstuff

Small Table

W1	W2
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*
*	*

This is the whole drama of PCA (dimensionality reduction technique) from 5d to 2d. Its clear that we reduced the data size from 5 attributes to 2 attributes which will explain the maximum variance of the original data.