САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ

Метод главных компонент (PCA)

Занятие 4

Глазунова Е.В.

PCA

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	1	2
Gene 2	6	4	5	3	2.8	1

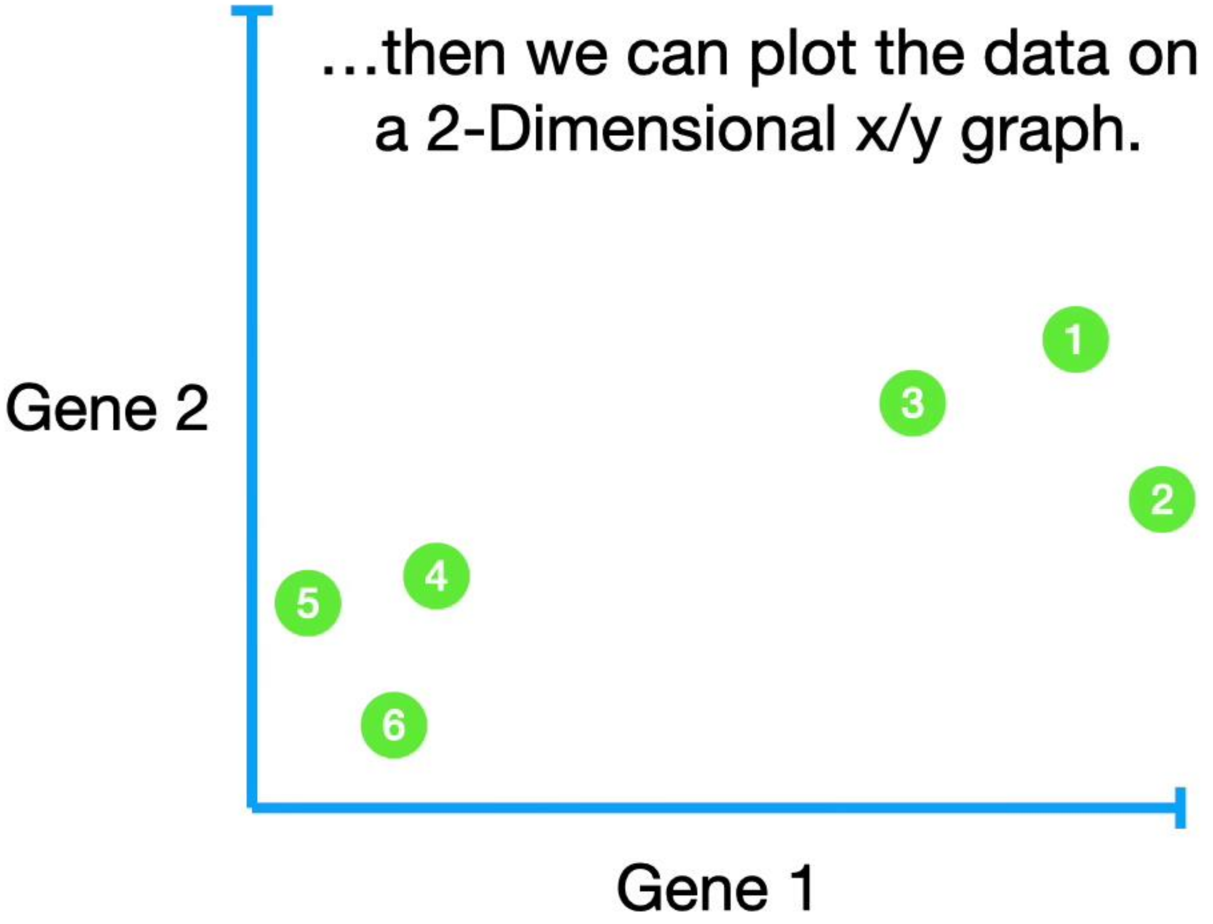
PCA

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	1	2

If we only measure 1 gene, we can plot the data on a number line…

PCA

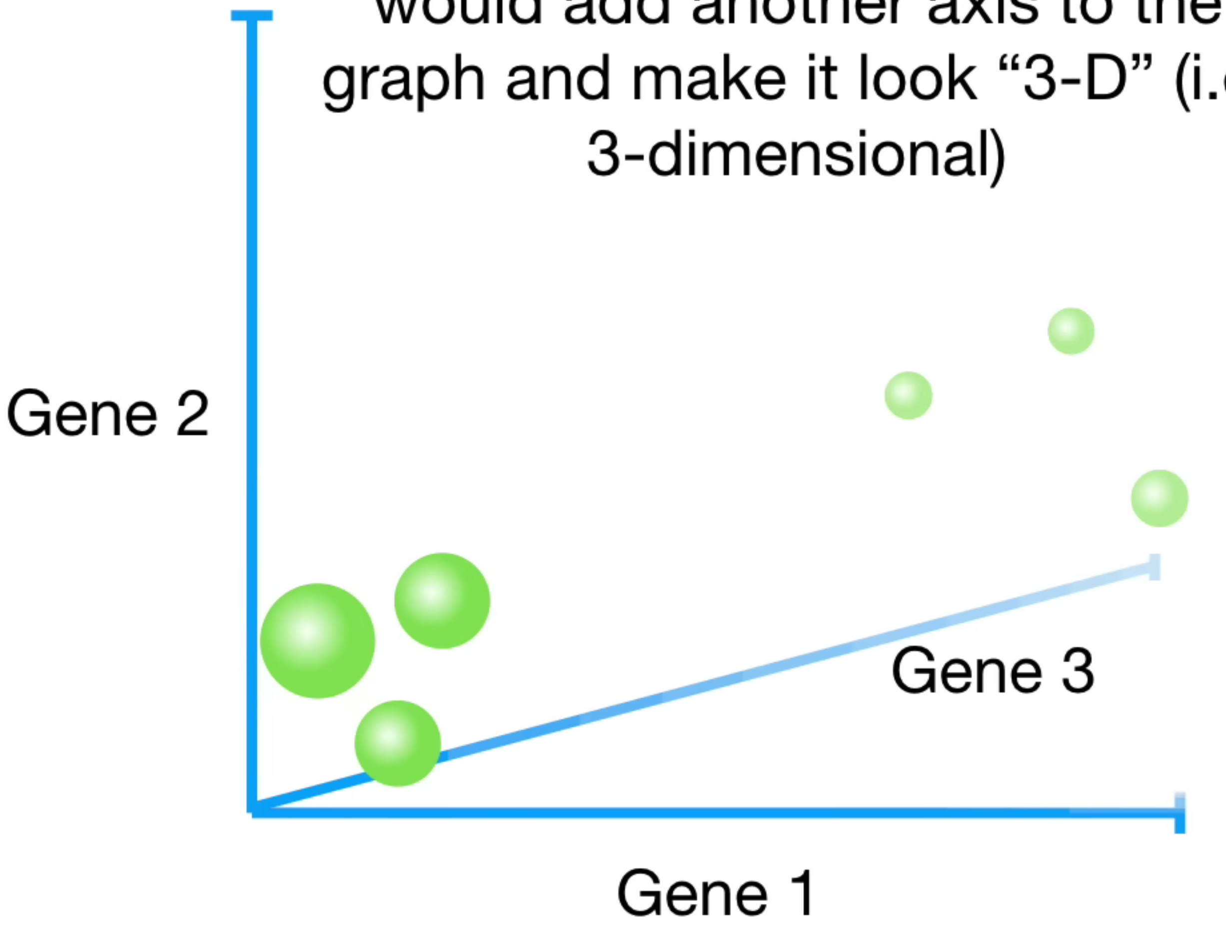
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	1	2
Gene 2	6	4	5	3	2.8	1



PCA

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2

If we measured 3 genes, we would add another axis to the graph and make it look "3-D" (i.e. 3-dimensional)



PCA

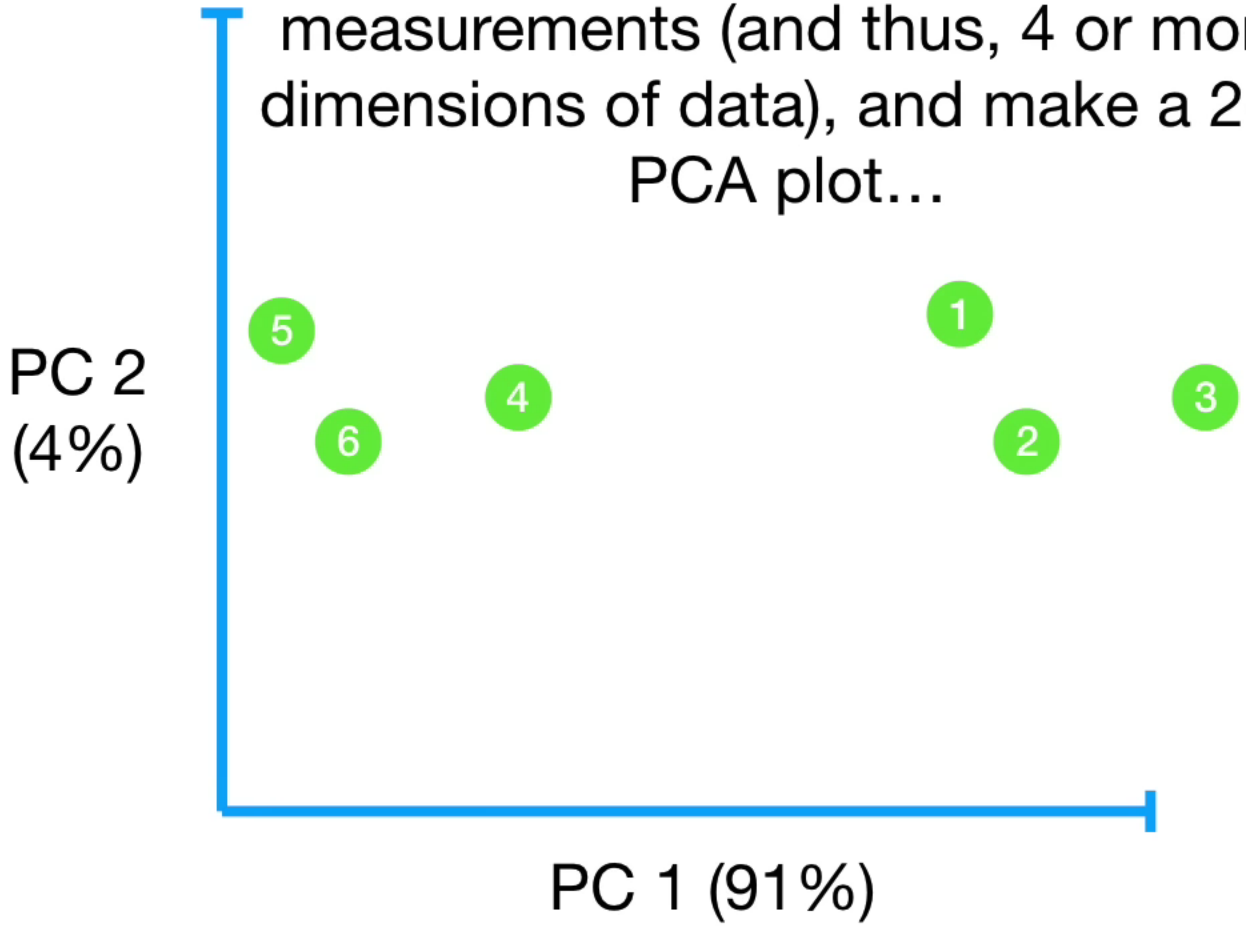
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

If we measured 4 genes, however, we can no longer plot the data - 4 genes require 4 dimensions.

PCA

	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2
Gene 4	5	7	6	2	4	7

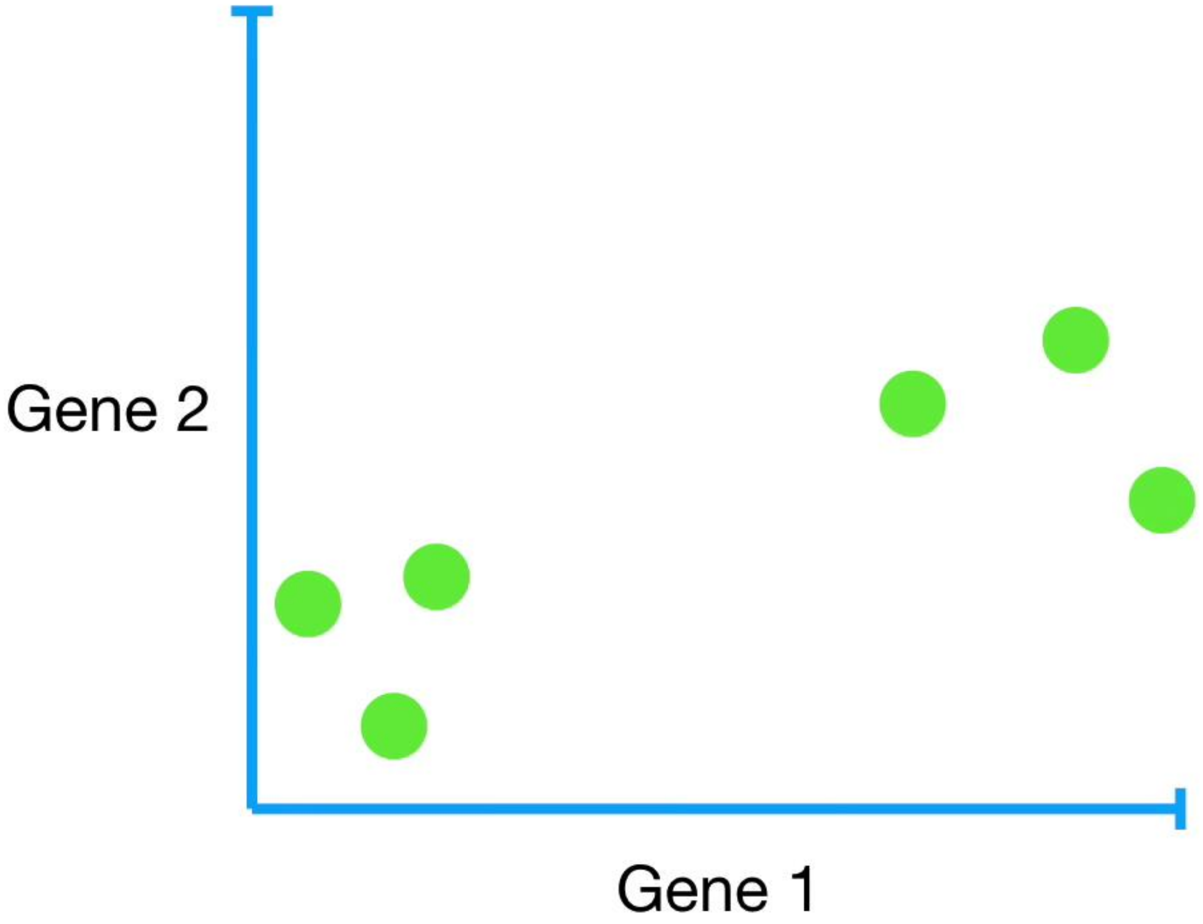
So we're going to talk about how PCA can take 4 or more gene measurements (and thus, 4 or more dimensions of data), and make a 2-D PCA plot…



PCA

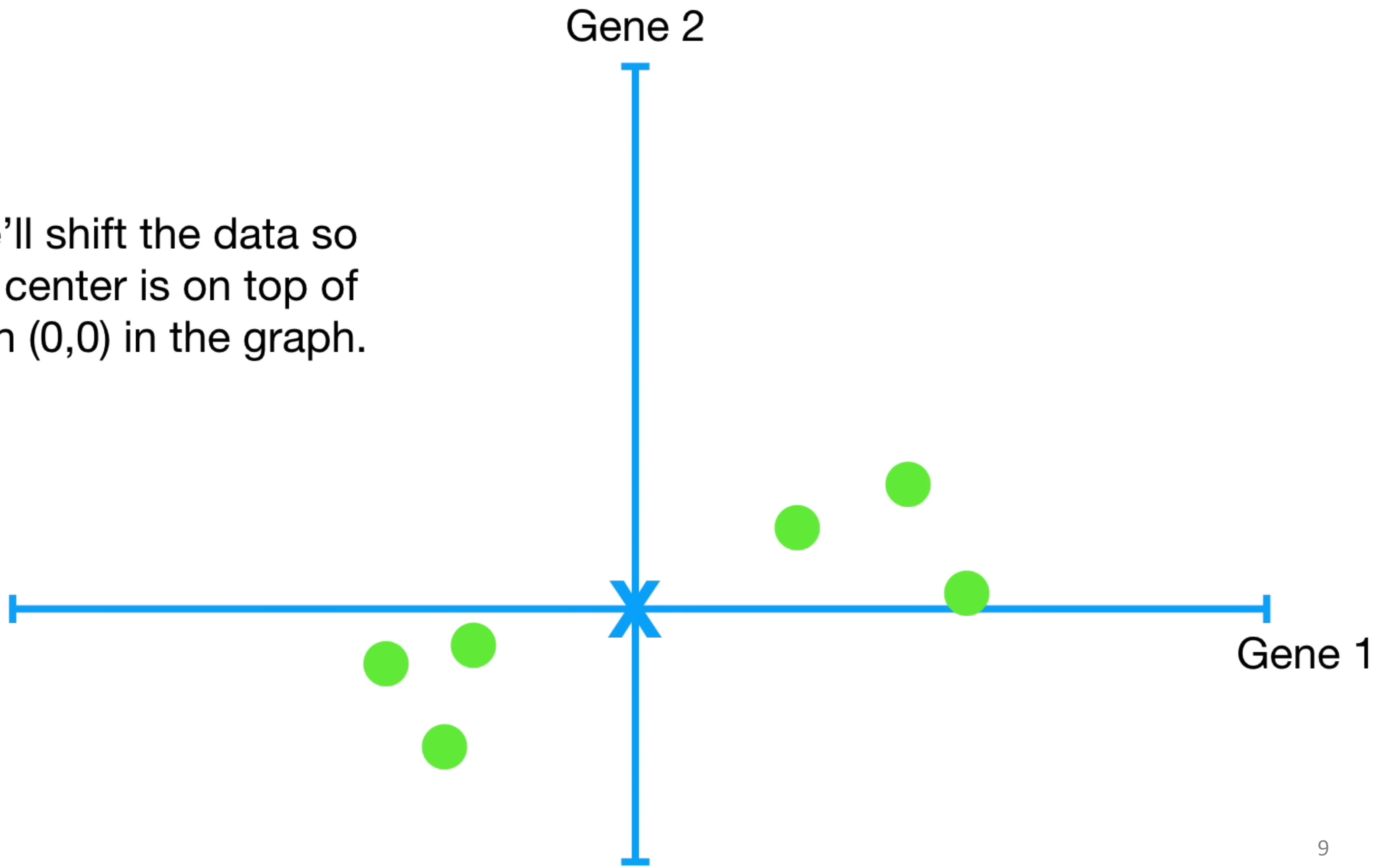
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

We'll start by plotting the data…



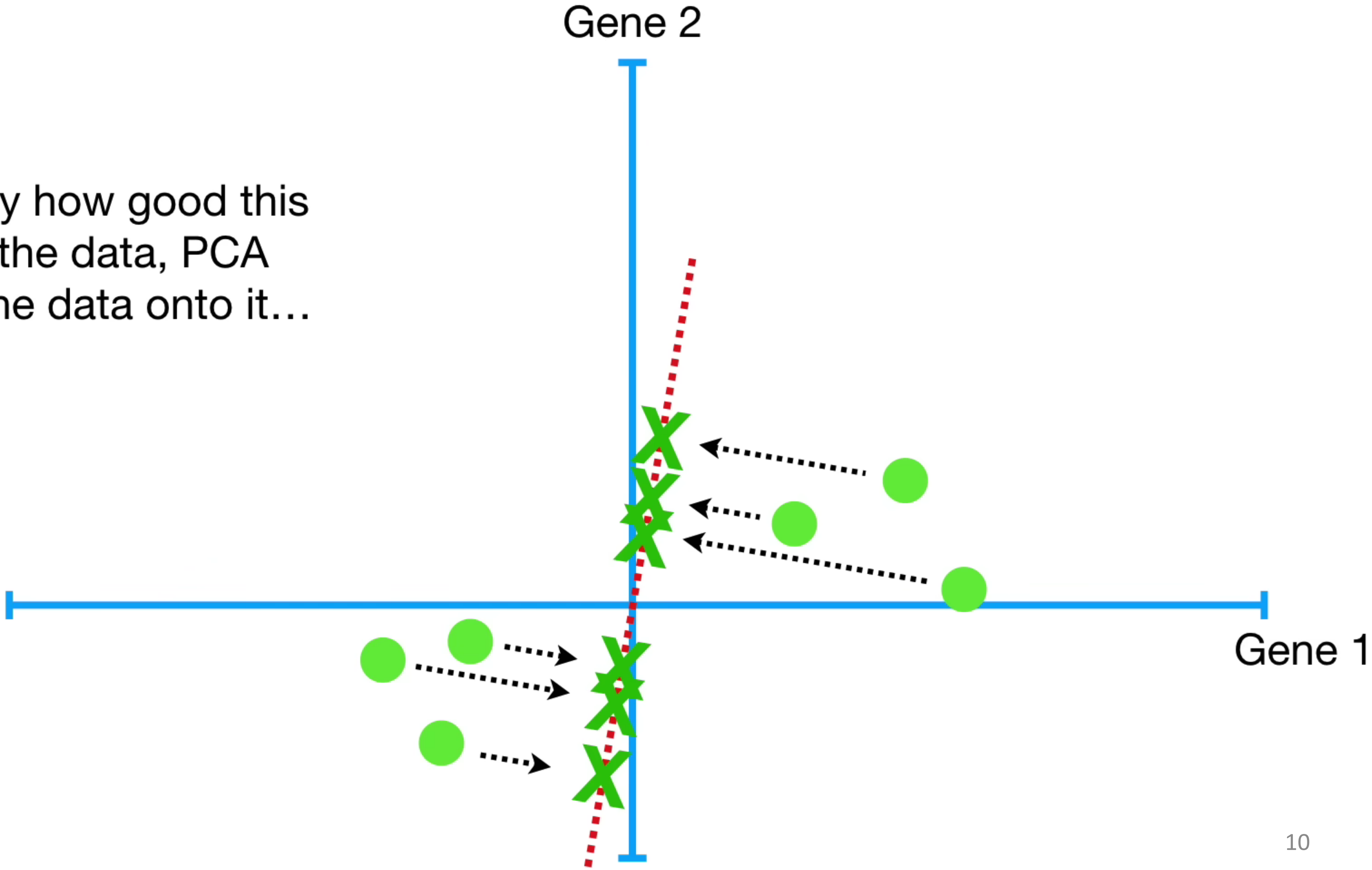
PCA

Now we'll shift the data so that the center is on top of the origin (0,0) in the graph.

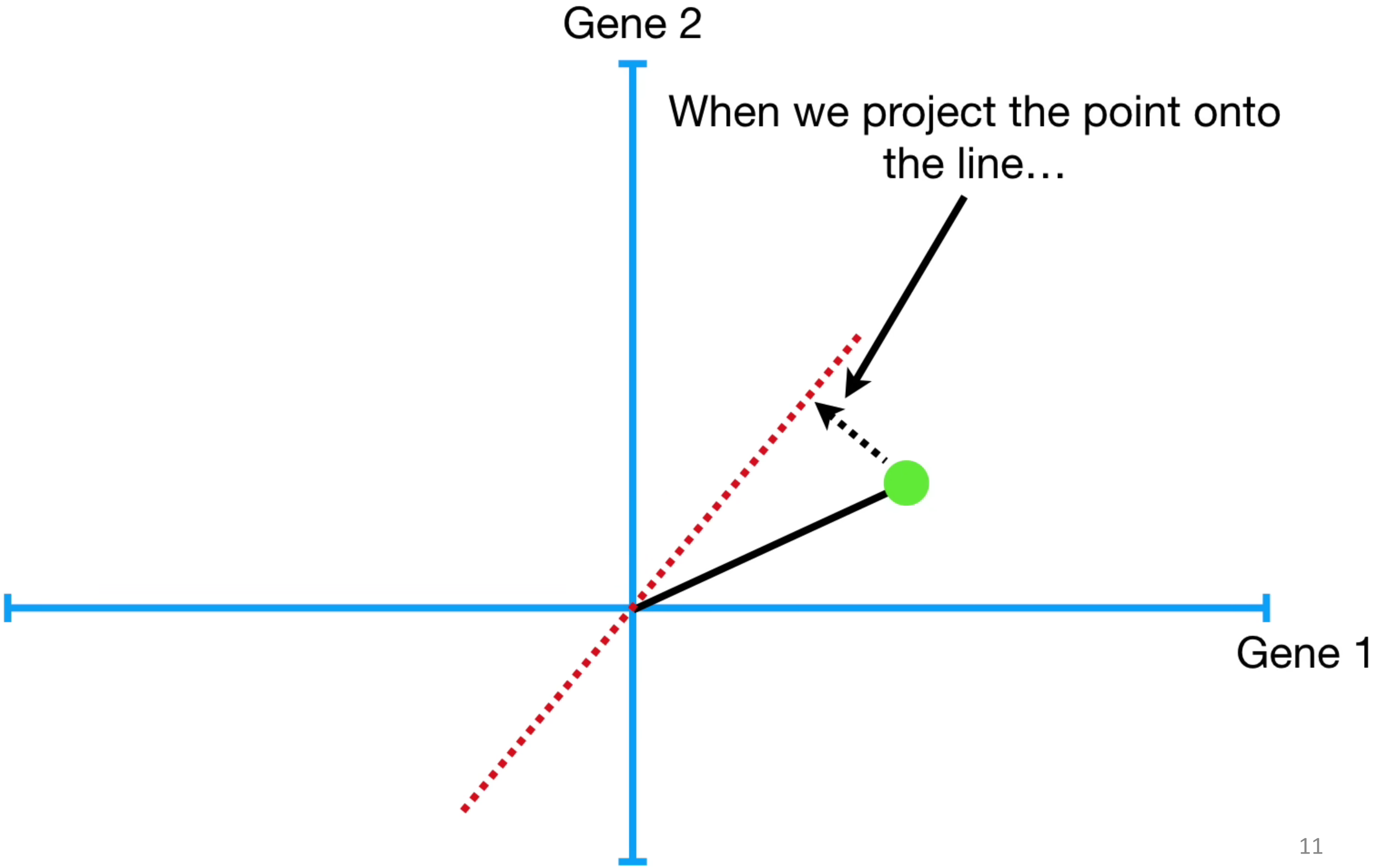


PCA

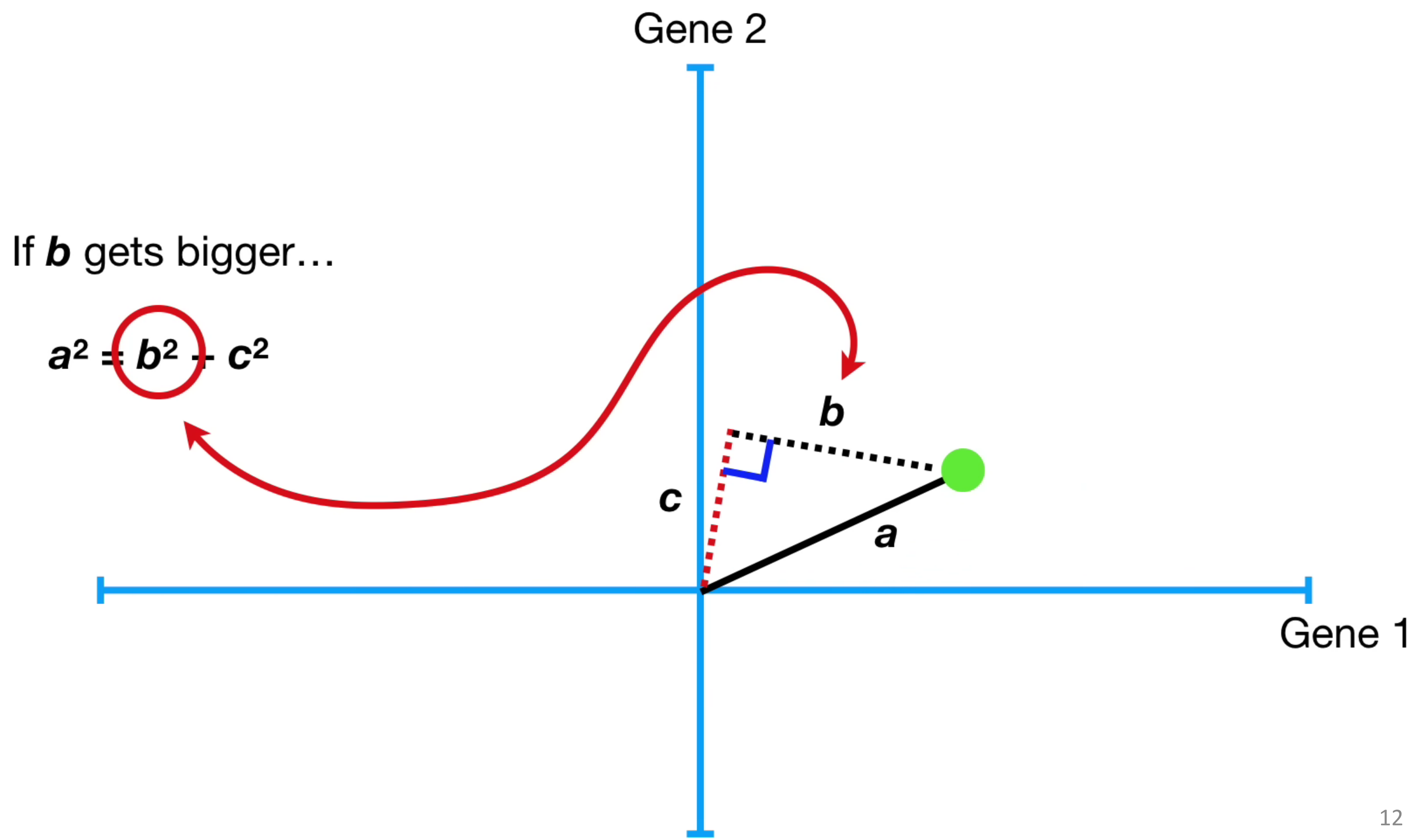
To quantify how good this line fits the data, PCA projects the data onto it…



PCA



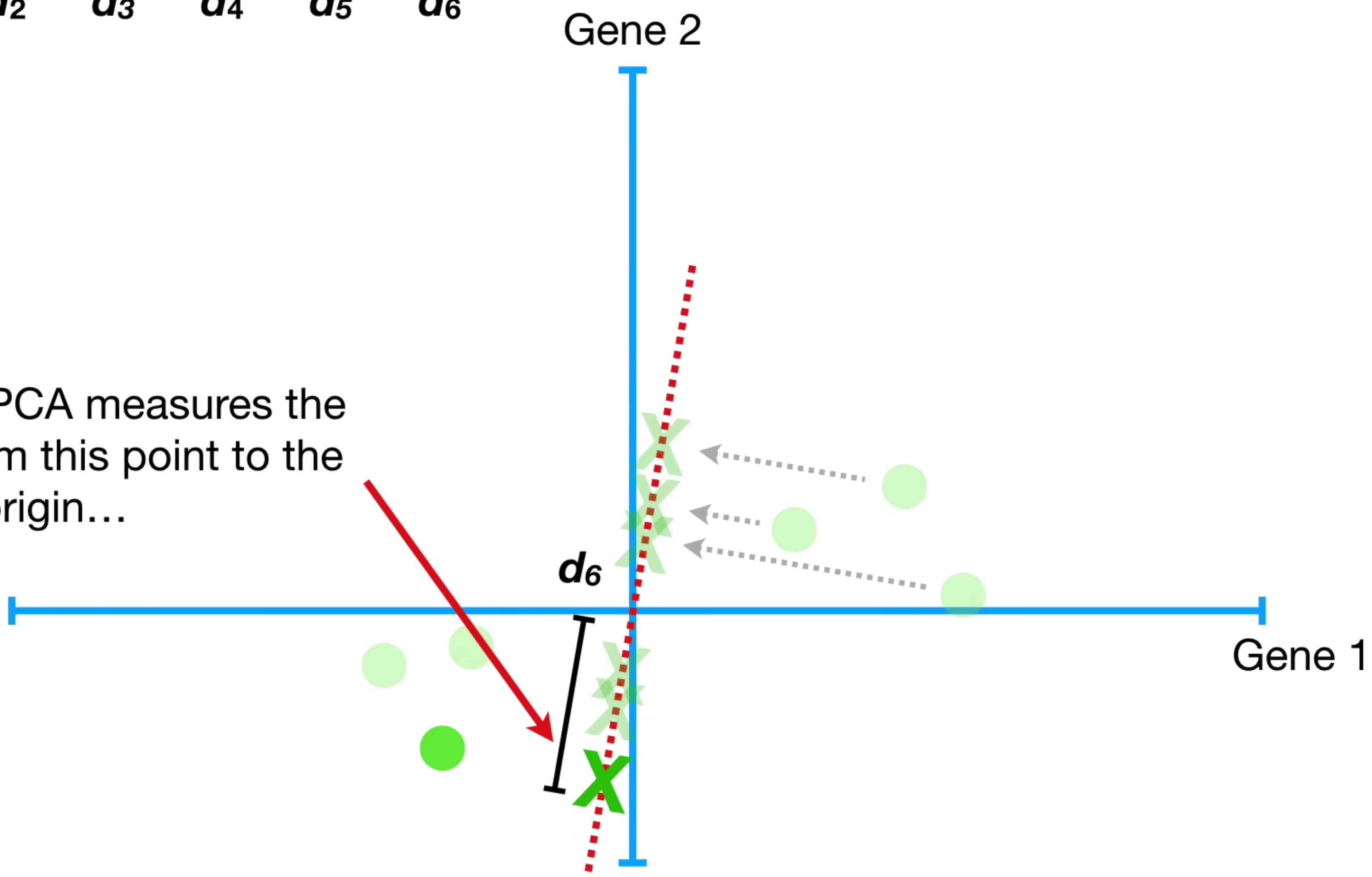
PCA



PCA

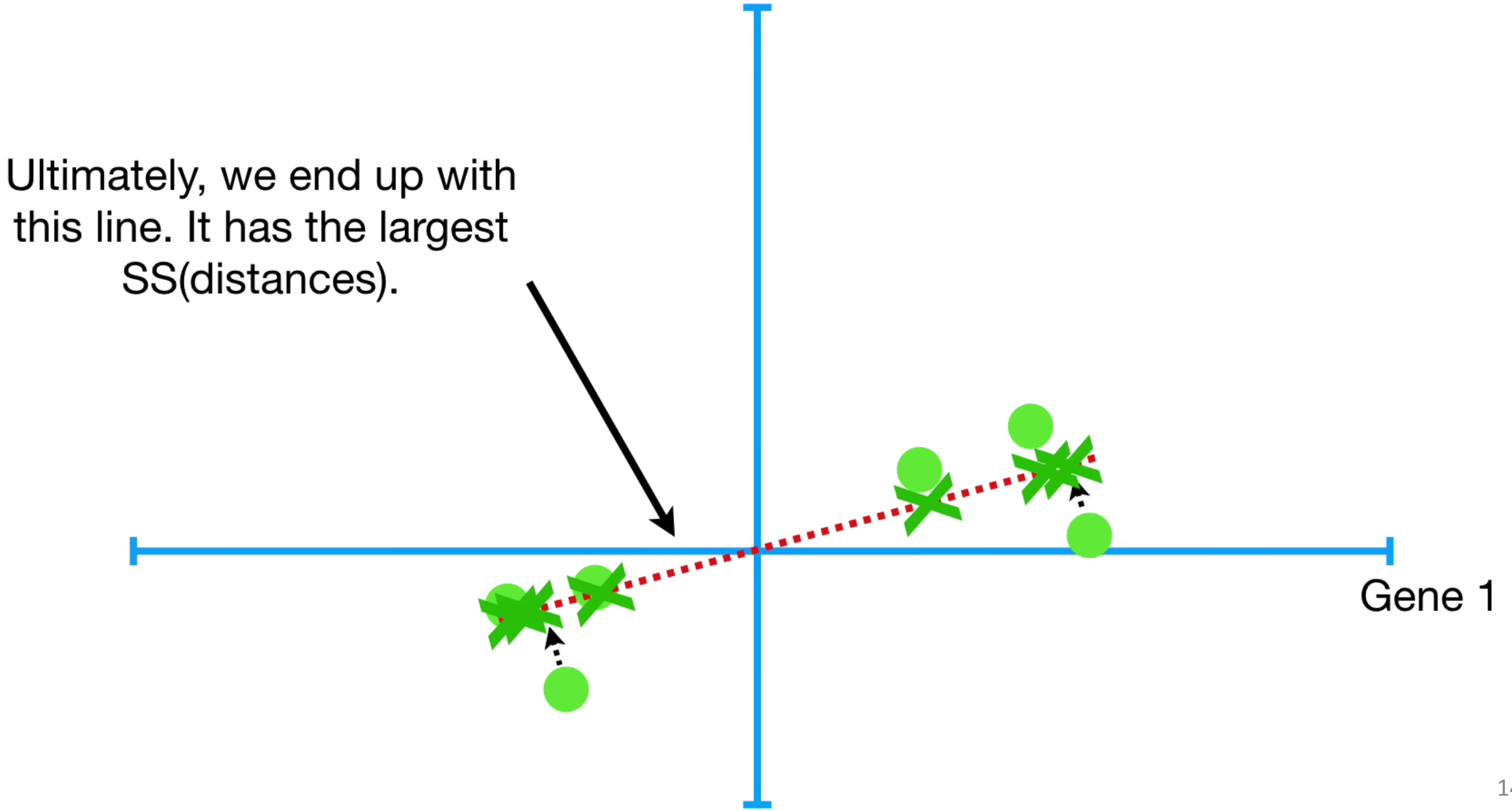
d_1 d_2 d_3 d_4 d_5 d_6

…and then PCA measures the distance from this point to the origin…

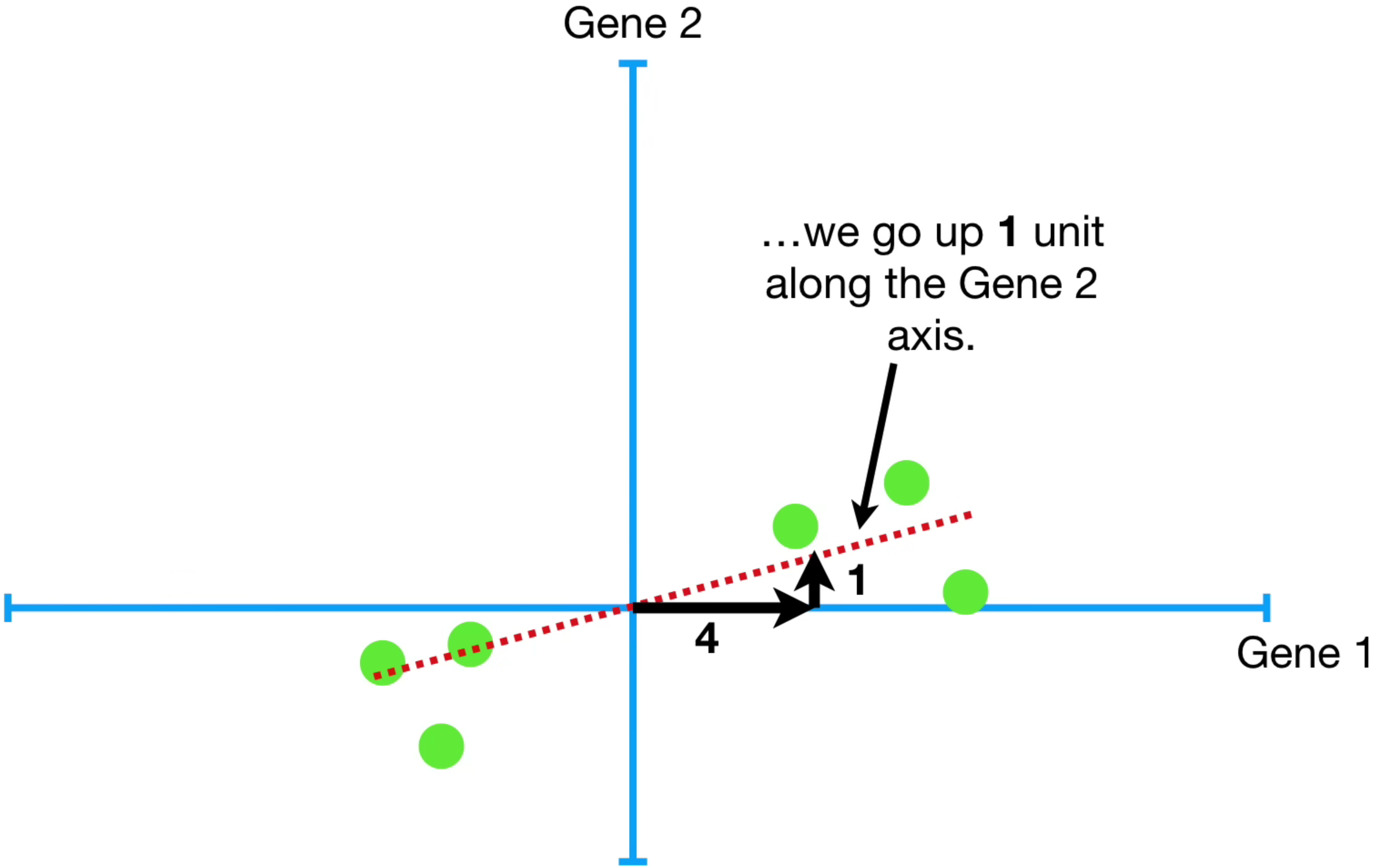


PCA

$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = \text{SS(distances)}$$



PCA

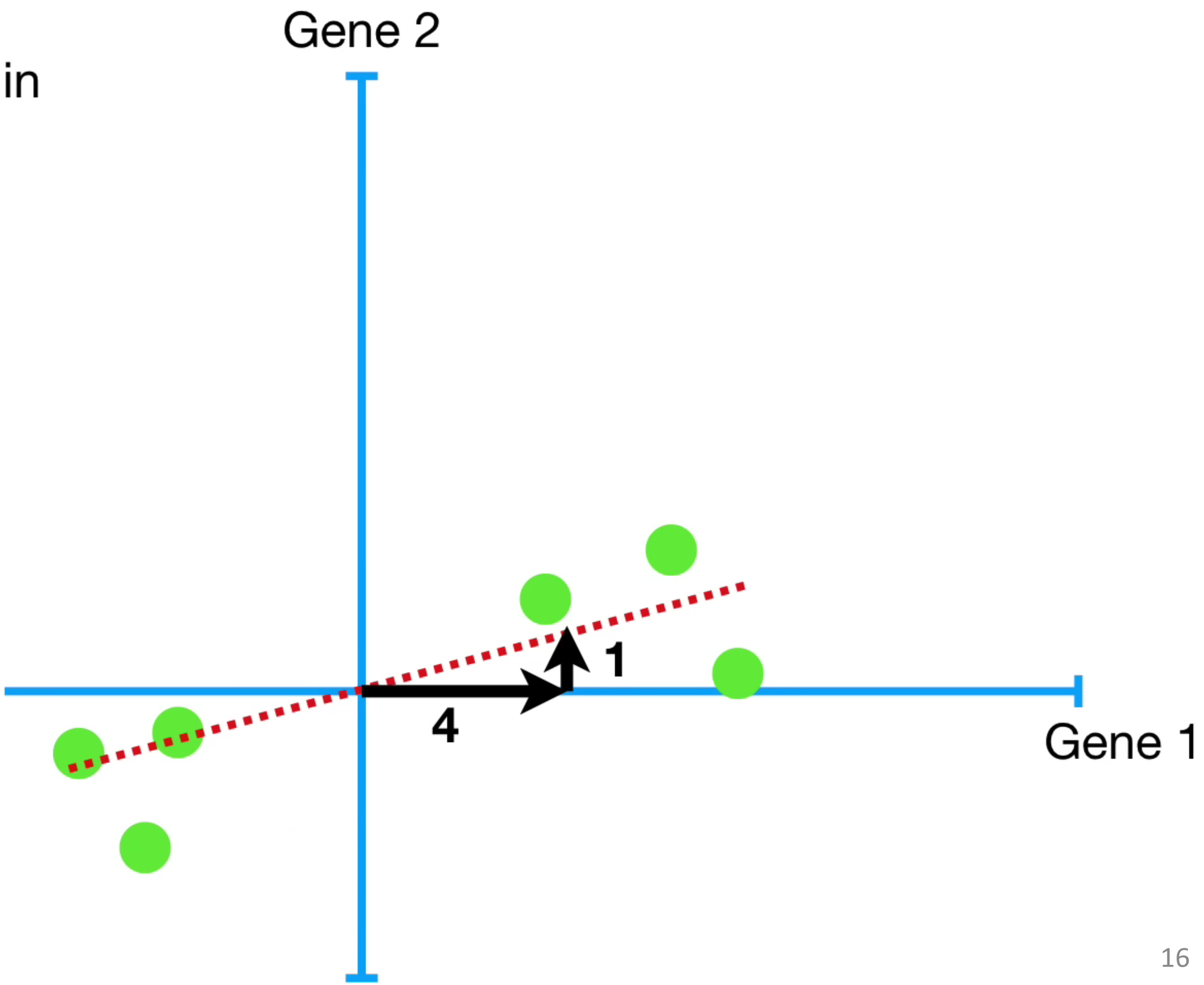


PCA

One way to think about PC1 is in terms of a cocktail recipe…

To make PC1
Mix **4** parts Gene 1
with **1** part Gene 2

The ratio of Gene 1 to Gene 2 tells you that Gene 1 is more important when it comes to describing how the data are spread out..

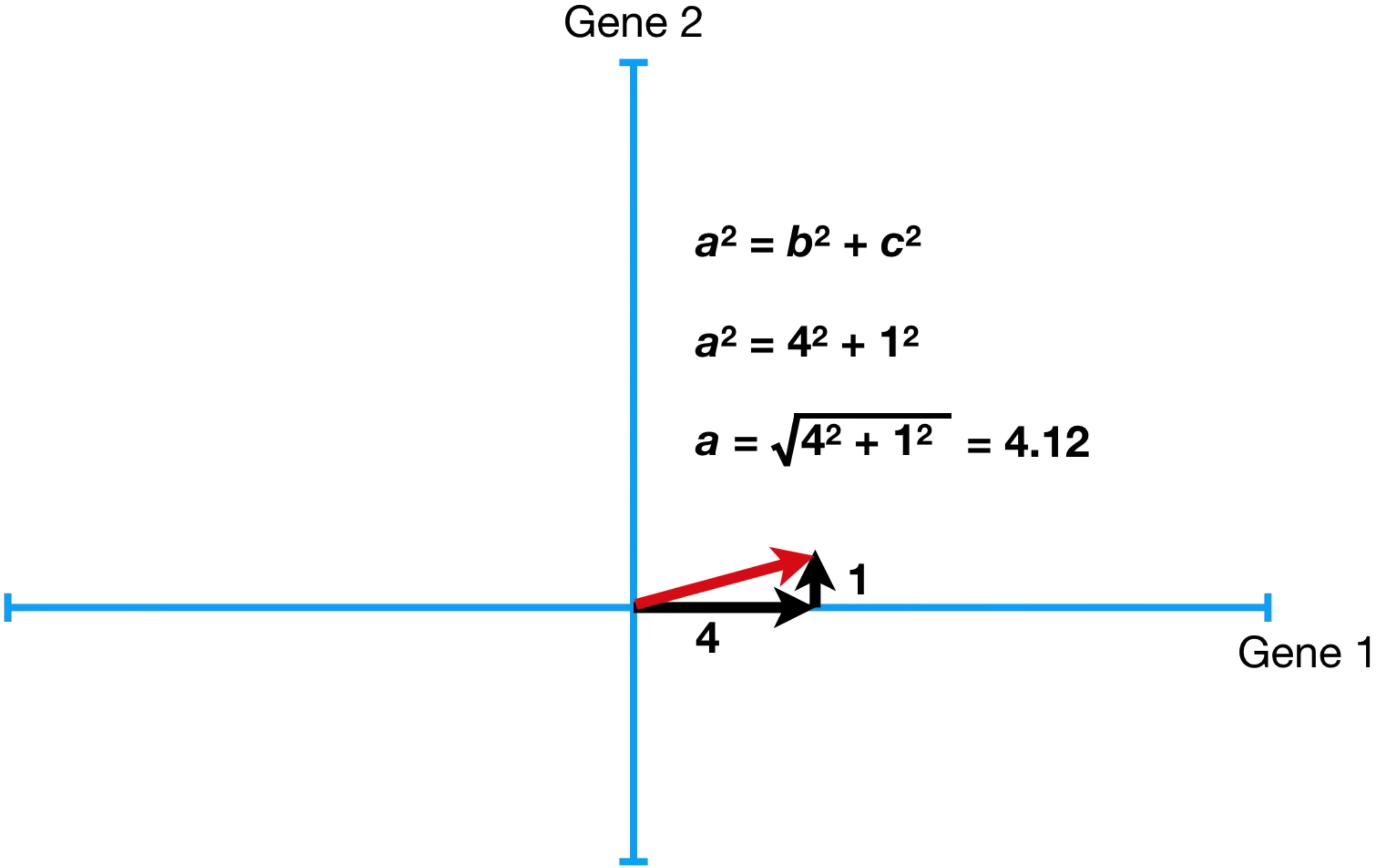


PCA

$$a^2 = b^2 + c^2$$

$$a^2 = 4^2 + 1^2$$

$$a = \sqrt{4^2 + 1^2} = 4.12$$

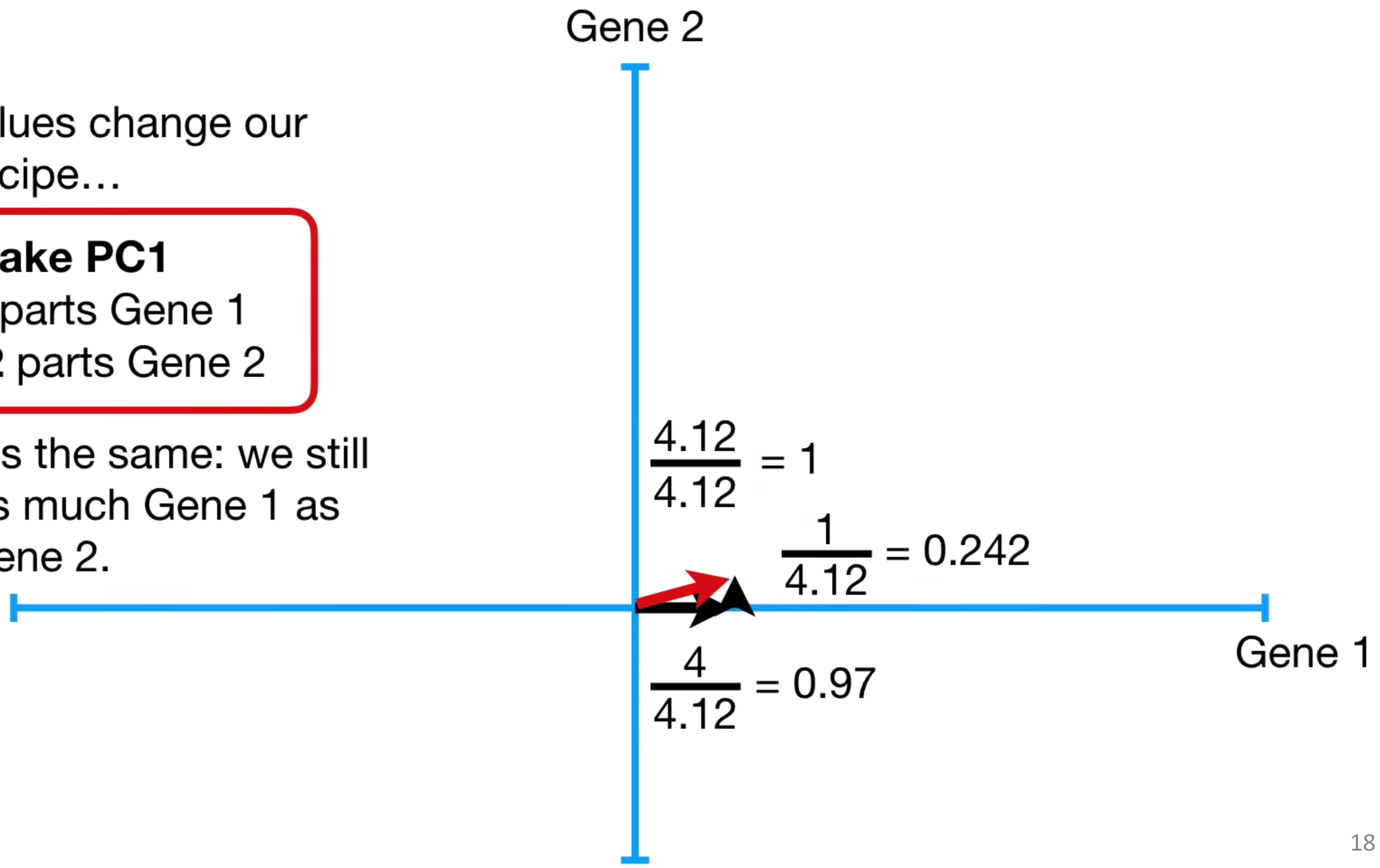


PCA

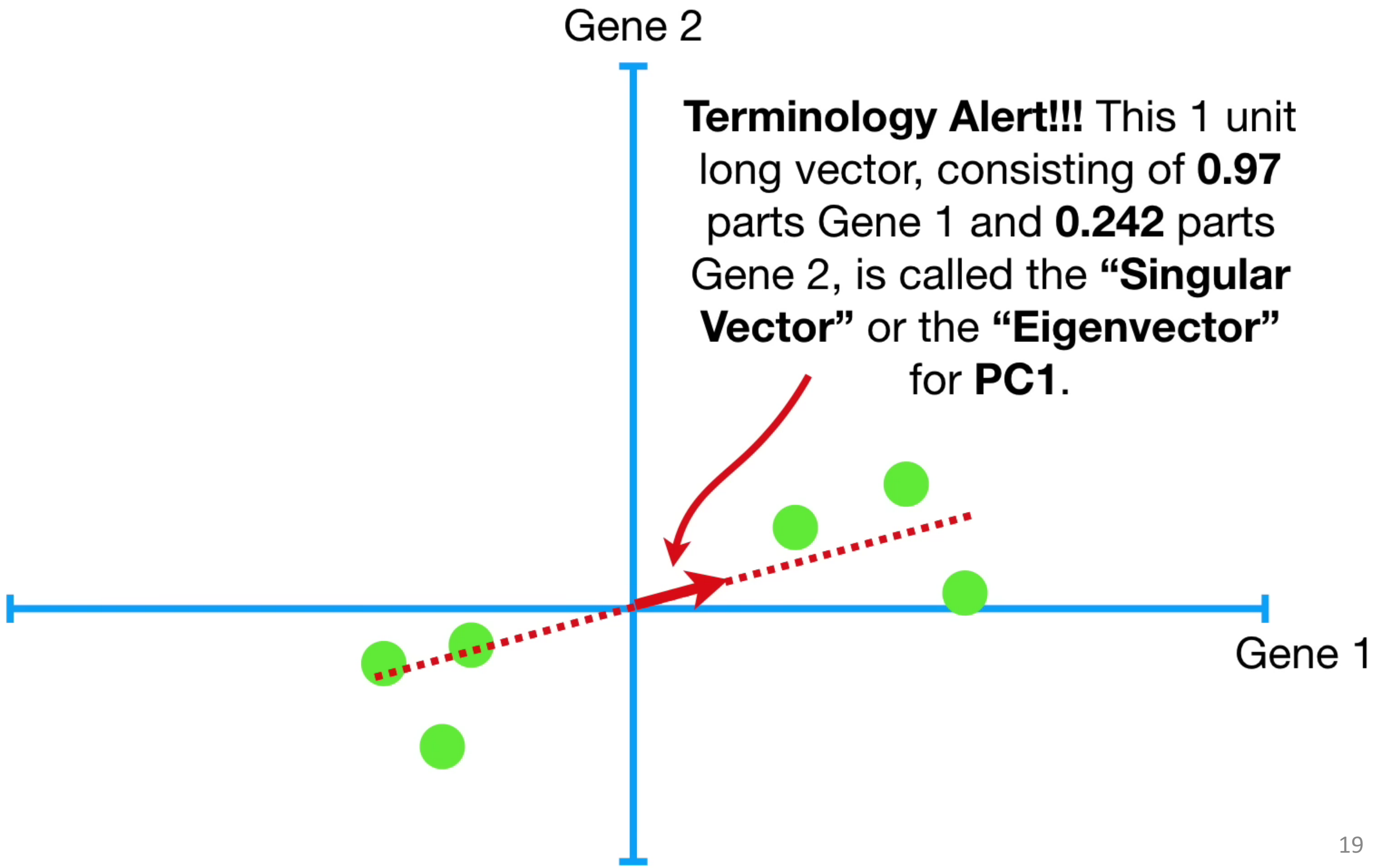
The new values change our recipe…

To make PC1
Mix **0.97** parts Gene 1
with **0.242** parts Gene 2

…but the ratio is the same: we still use 4 times as much Gene 1 as Gene 2.



PCA

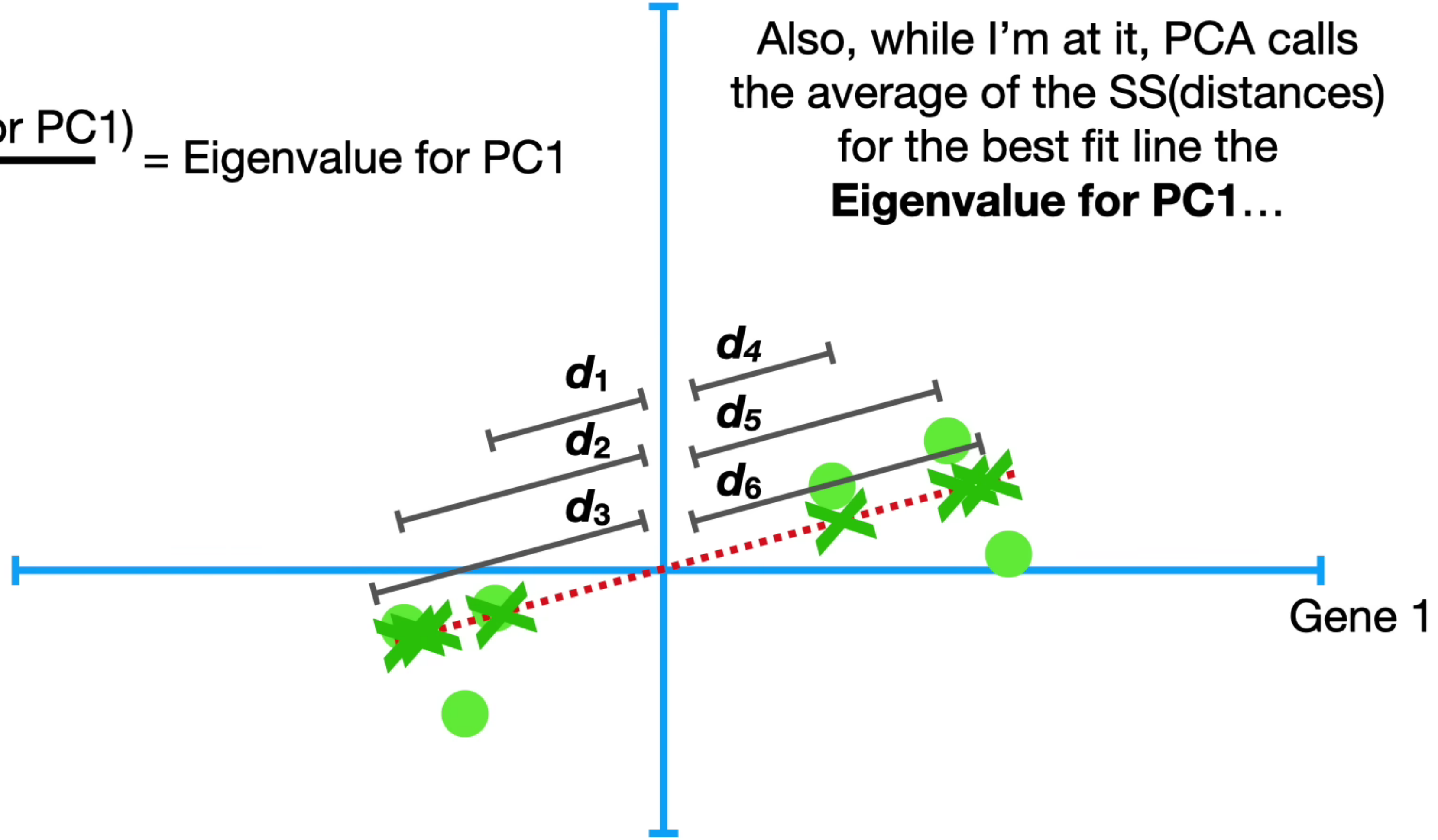

Gene 2
Terminology Alert!!! This 1 unit long vector, consisting of 0.97 parts Gene 1 and 0.242 parts Gene 2, is called the "Singular Vector" or the "Eigenvector" for PC1.
Gene 1

PCA

$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2$ = sum of squared distances = SS(distances)

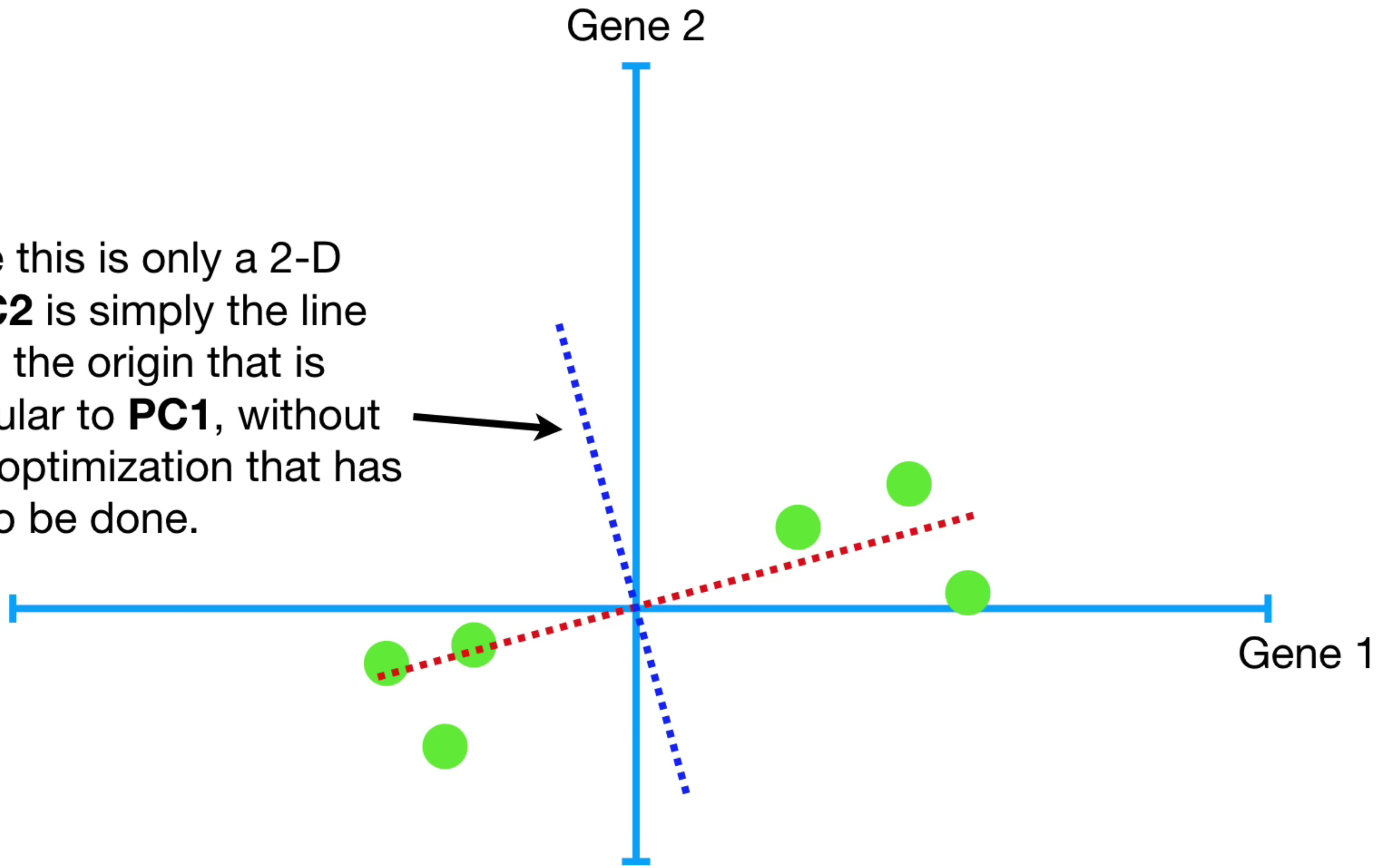
$$\frac{\text{SS(distances for PC1)}}{n - 1} = \text{Eigenvalue for PC1}$$

Also, while I'm at it, PCA calls the average of the SS(distances) for the best fit line the **Eigenvalue for PC1**…

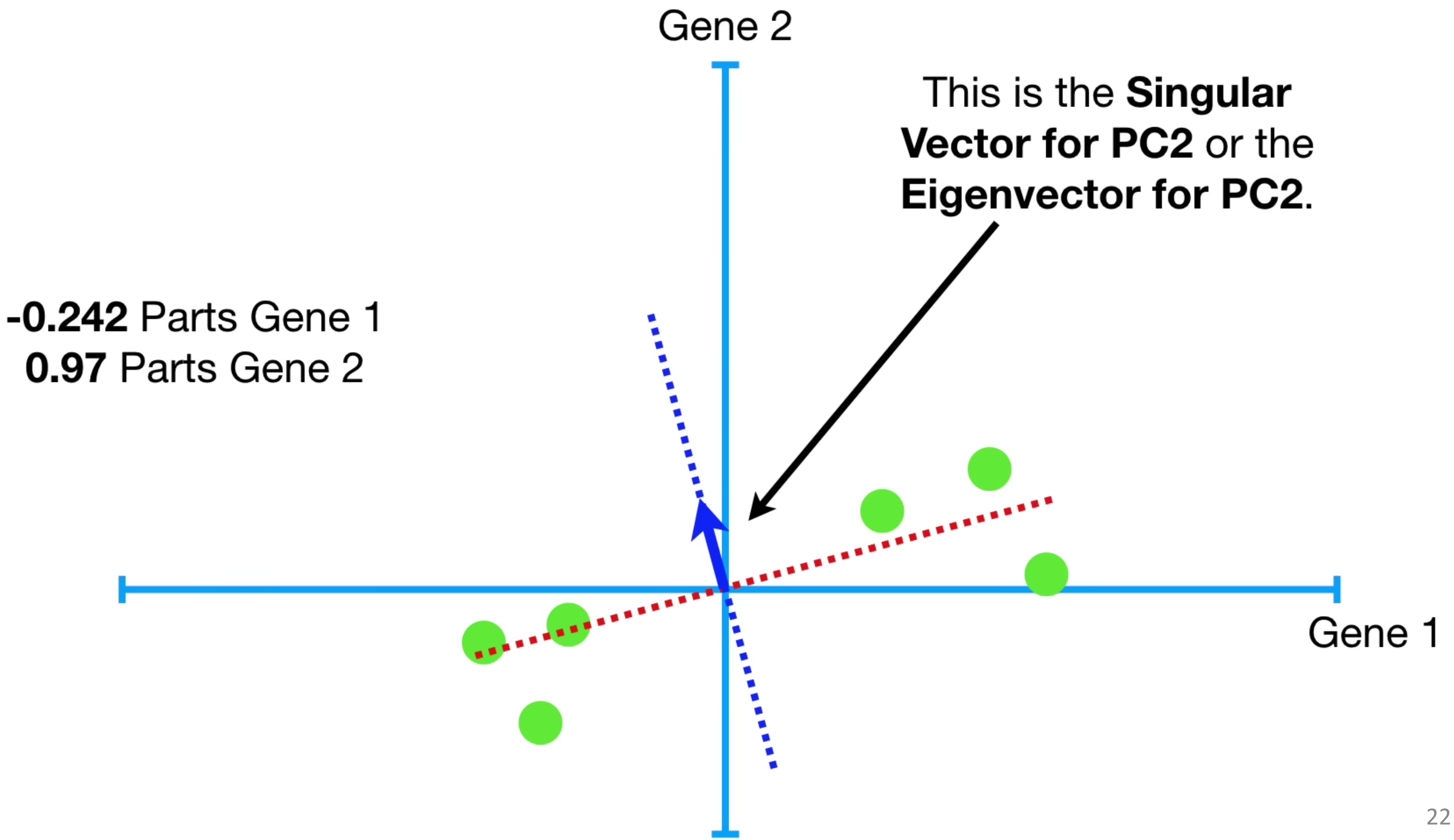


PCA

Because this is only a 2-D graph, **PC2** is simply the line through the origin that is perpendicular to **PC1**, without any further optimization that has to be done.



PCA



Gene 2
This is the **Singular Vector for PC2** or the **Eigenvector for PC2**.
-0.242 Parts Gene 1
0.97 Parts Gene 2
Gene 1

PCA

For the sake of the example, imagine that the Variation for **PC1 = 15**, and the variation for **PC2 = 3**.

That means that the total variation around both PCs is **15 + 3 = 18**…

$$\frac{\text{SS(distances for PC1)}}{n - 1} = \text{Variation for PC1}$$

$$\frac{\text{SS(distances for PC2)}}{n - 1} = \text{Variation for PC2}$$

…and that means PC1 accounts for **15 / 18 = 0.83 = 83%** of the total variation around the PCs.