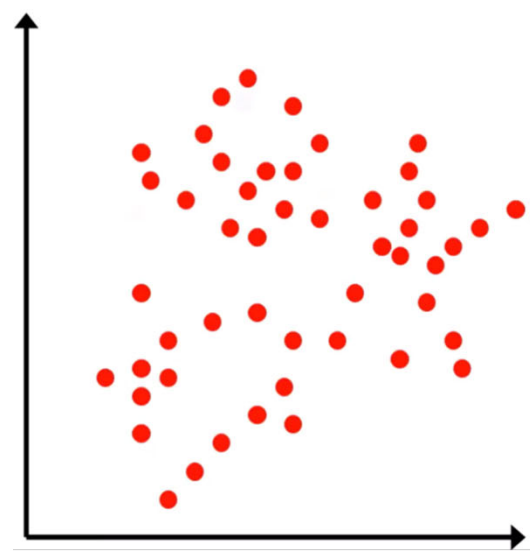
Introduction to Machine Learning

Clustering Revisited

Dimensionality Reduction

Algorithm 4 K-MEANS($\mathbf{D}$, K)

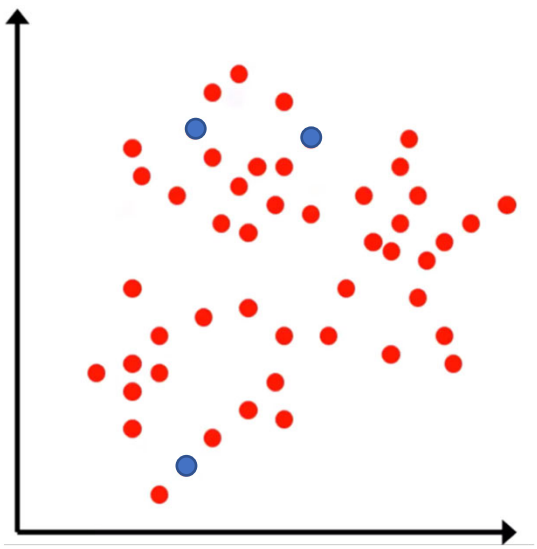
```
for k = 1 to K do
    μ_k ← some random location          // randomly initialize center for kth cluster
end for
repeat
    for n = 1 to N do
        z_n ← argmin_k ||μ_k − x_n||    // assign example n to closest center
    end for
    for k = 1 to K do
        X_k ← { x_n : z_n = k }         // points assigned to cluster k
        μ_k ← MEAN(X_k)                 // re-estimate center of cluster k
    end for
until μs stop changing
return z                                // return cluster assignments
```



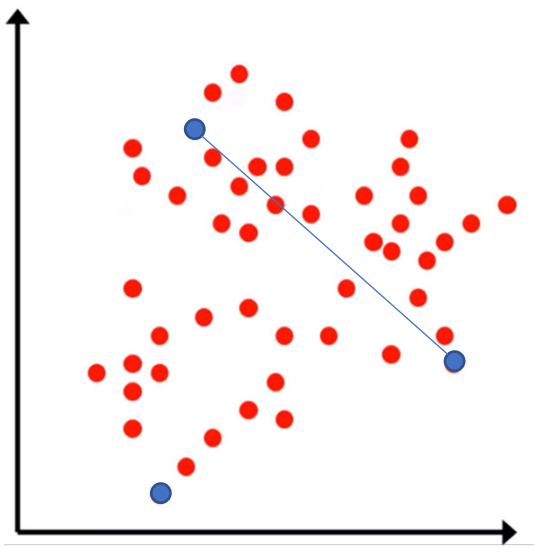
Most important K-Means choice

- Furthest-first heuristic

Random



Furthest-first



How to pick the initial centroids?

I'll Choose

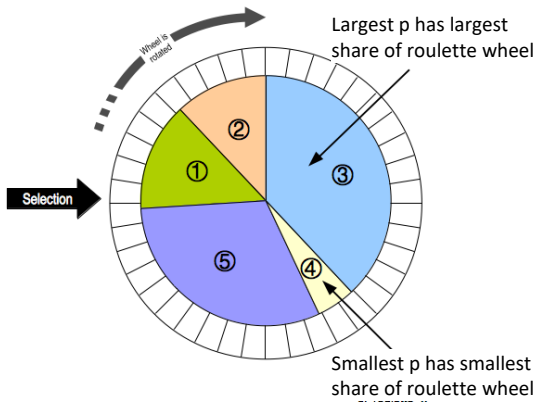
Randomly

Farthest Point

Restart

Algorithm 36 K-MEANS++(D, K)

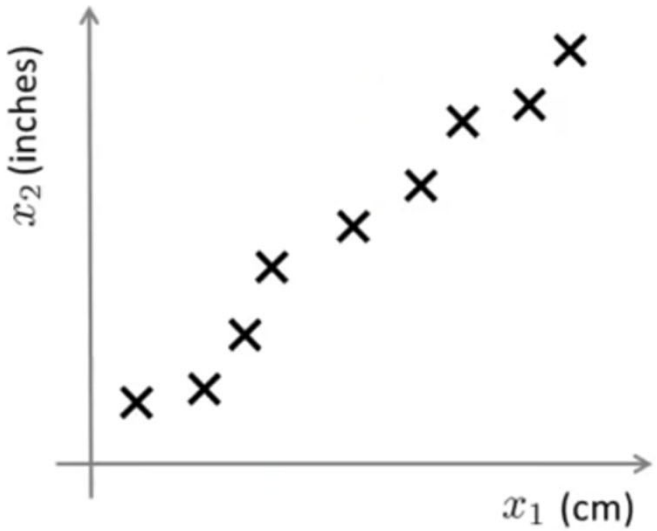
$\boldsymbol{\mu}_1 \leftarrow x_m$ for m chosen uniformly at random // randomly initialize first point
for $k = 2$ **to** K **do**
  $d_n \leftarrow \min_{k' < k} ||x_n - \boldsymbol{\mu}_{k'}||^2, \forall n$ // compute distances
  $p \leftarrow \frac{1}{\sum_n n_d} d$ // normalize to probability distribution
  $m \leftarrow$ random sample from p // pick an example at random
  $\boldsymbol{\mu}_k \leftarrow x_m$
end for
run K-MEANS using $\boldsymbol{\mu}$ as initial centers



Dimensionality reduction

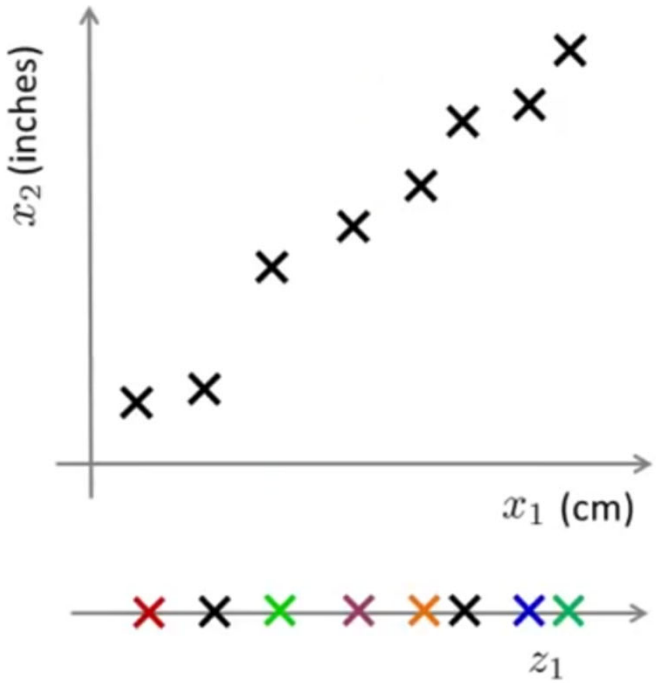
Dimensionality reduction via Principal Component Analysis

Motivation: Data compression



Reduce data from
2D to 1D

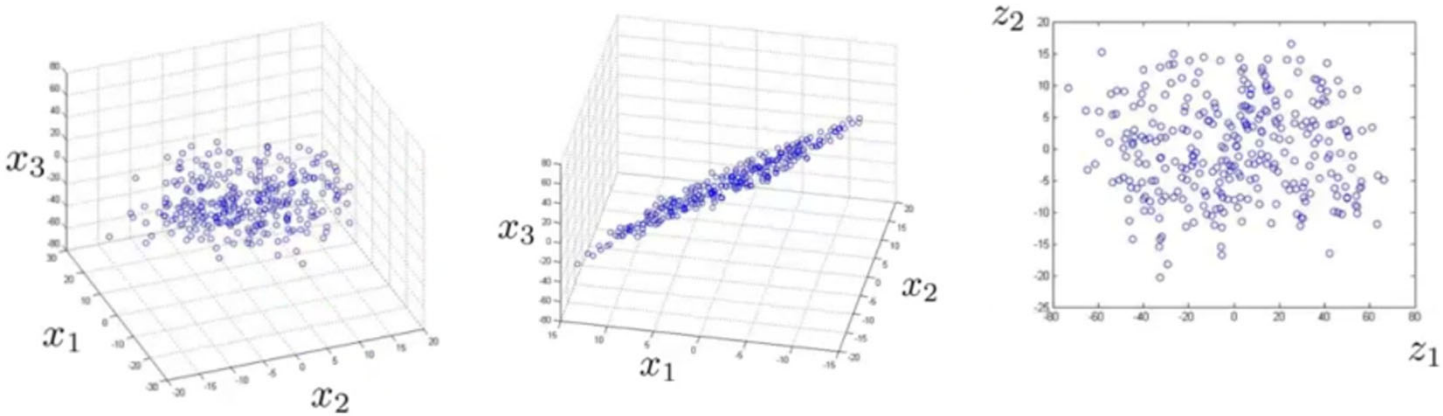
Motivation: Data compression



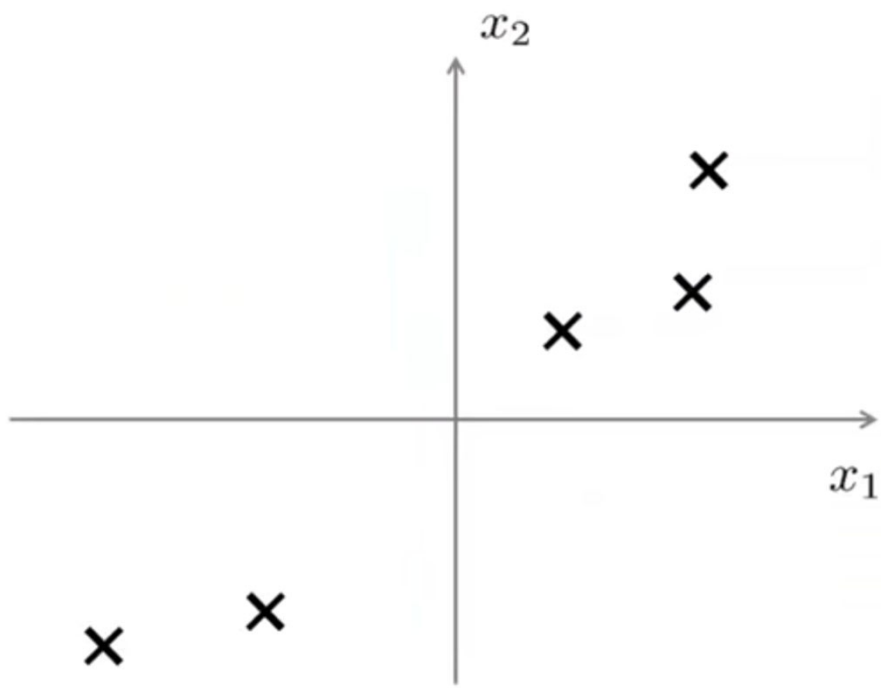
Reduce data from
2D to 1D

Motivation: Data compression

Reduce data from 3D to 2D

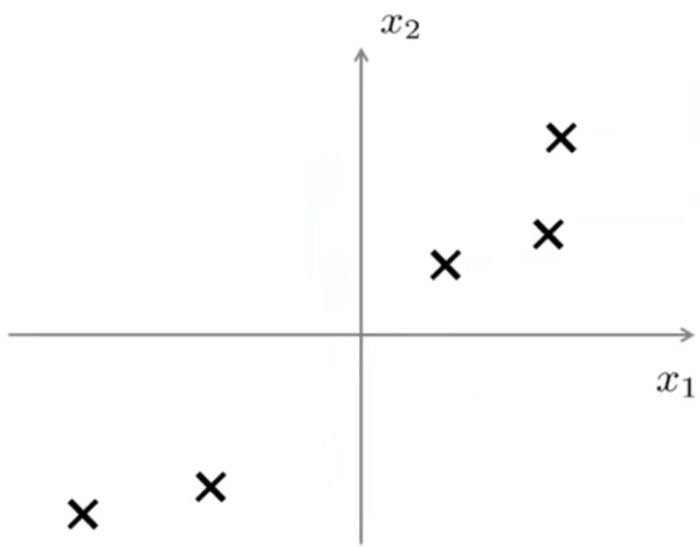


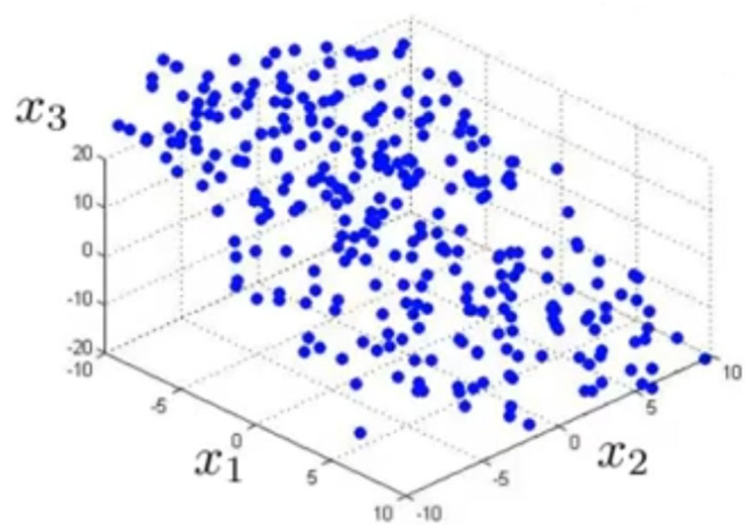
Principal Component Analysis problem



Principal Component Analysis problem

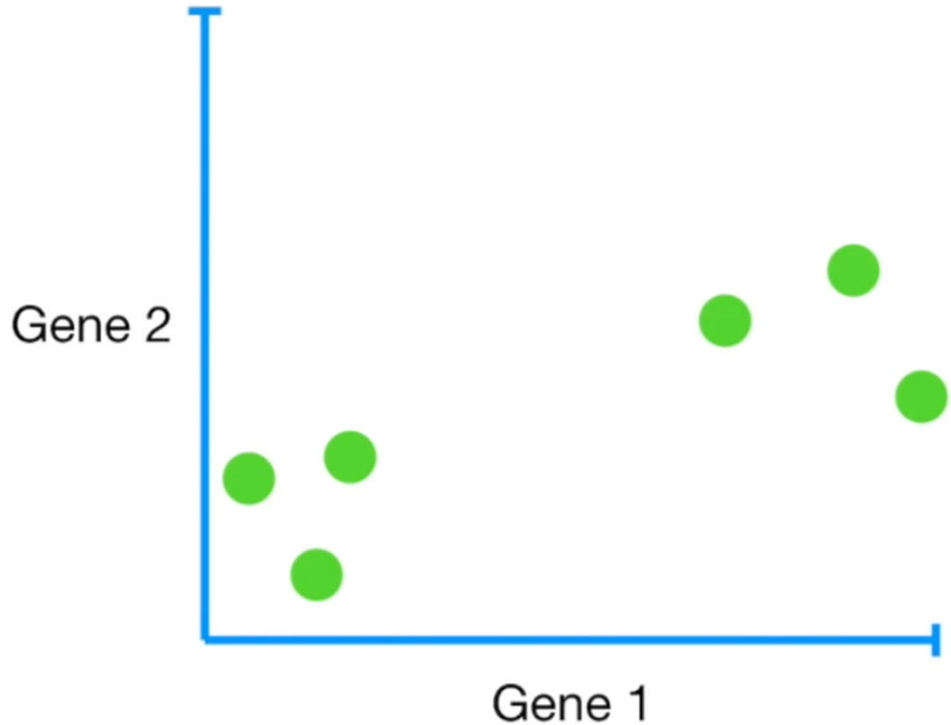
- Find a direction to project data that minimizes projection error



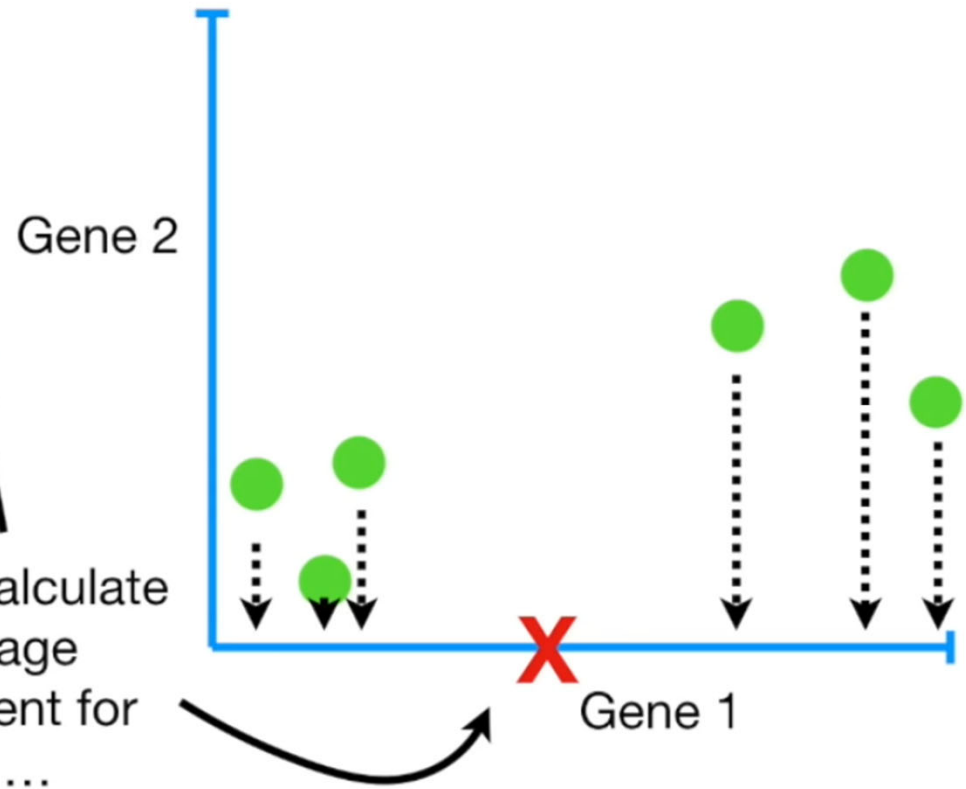


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

We'll start by plotting the data…

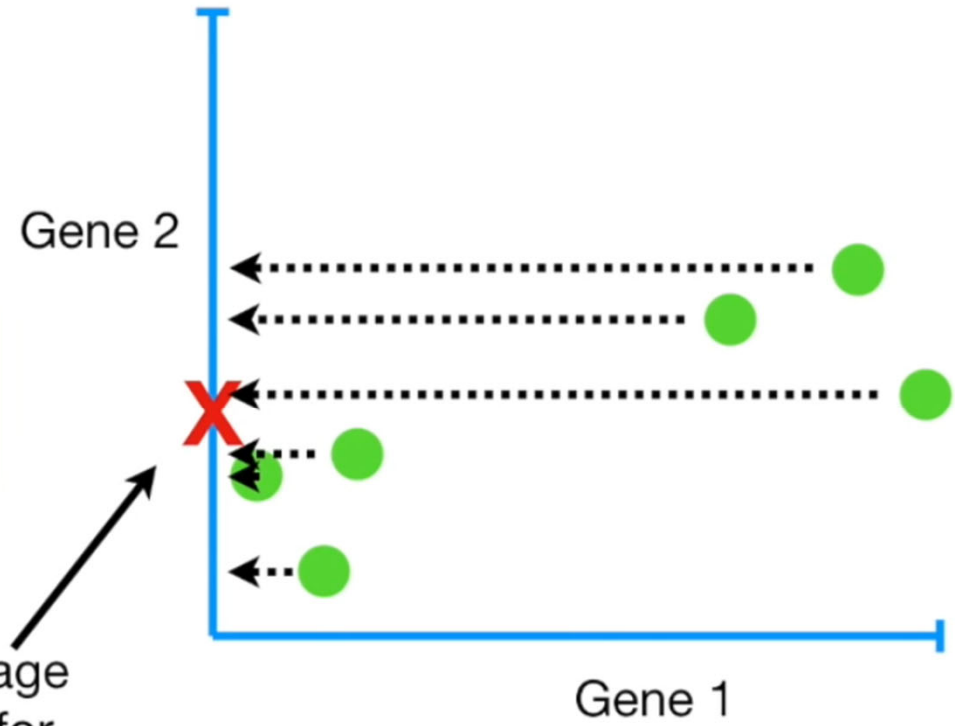


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



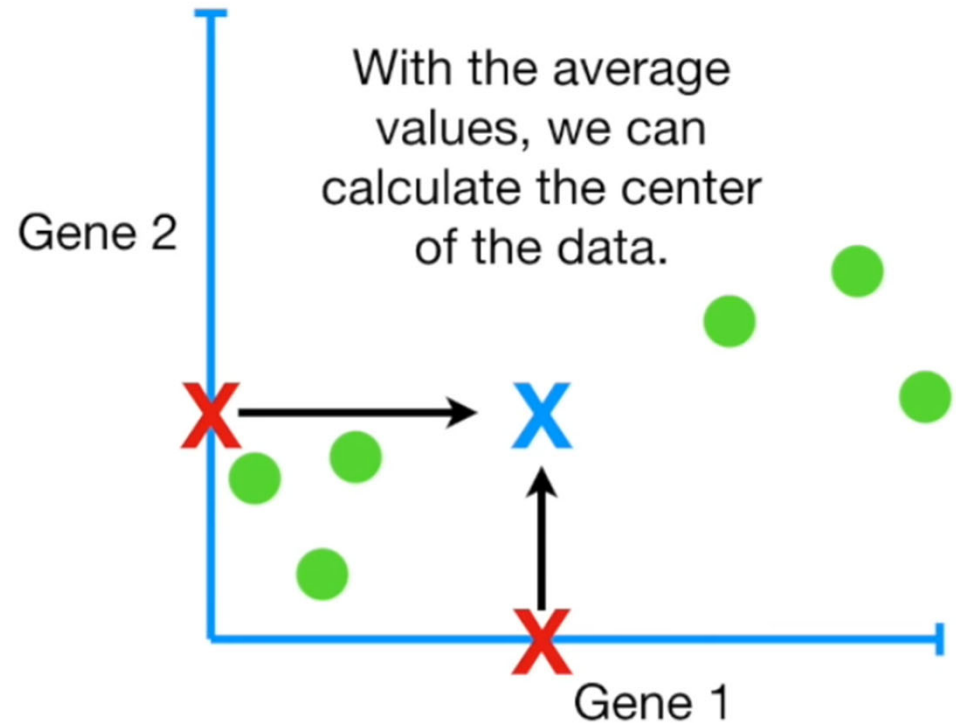
Then we'll calculate the average measurement for Gene 1…

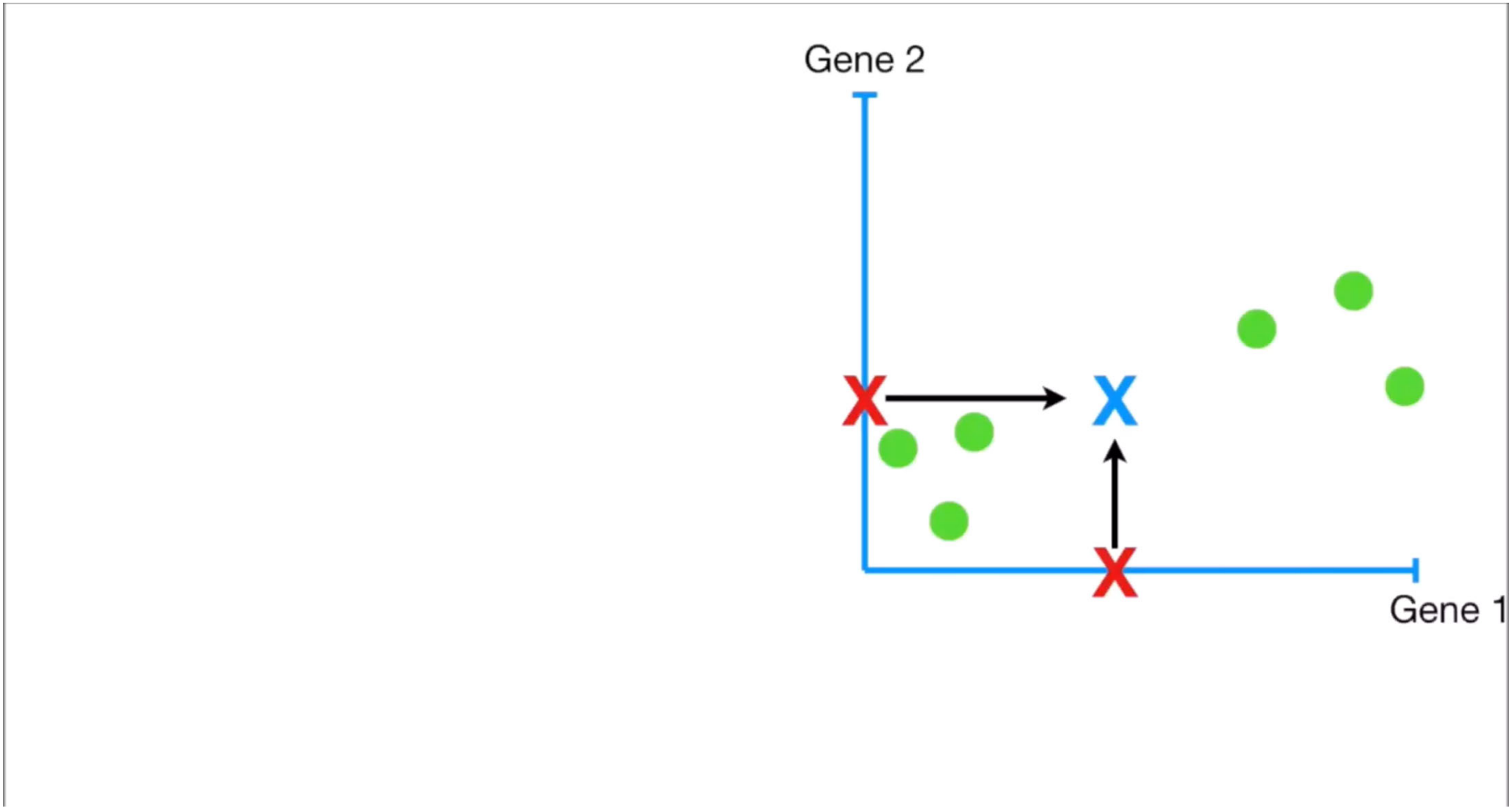
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

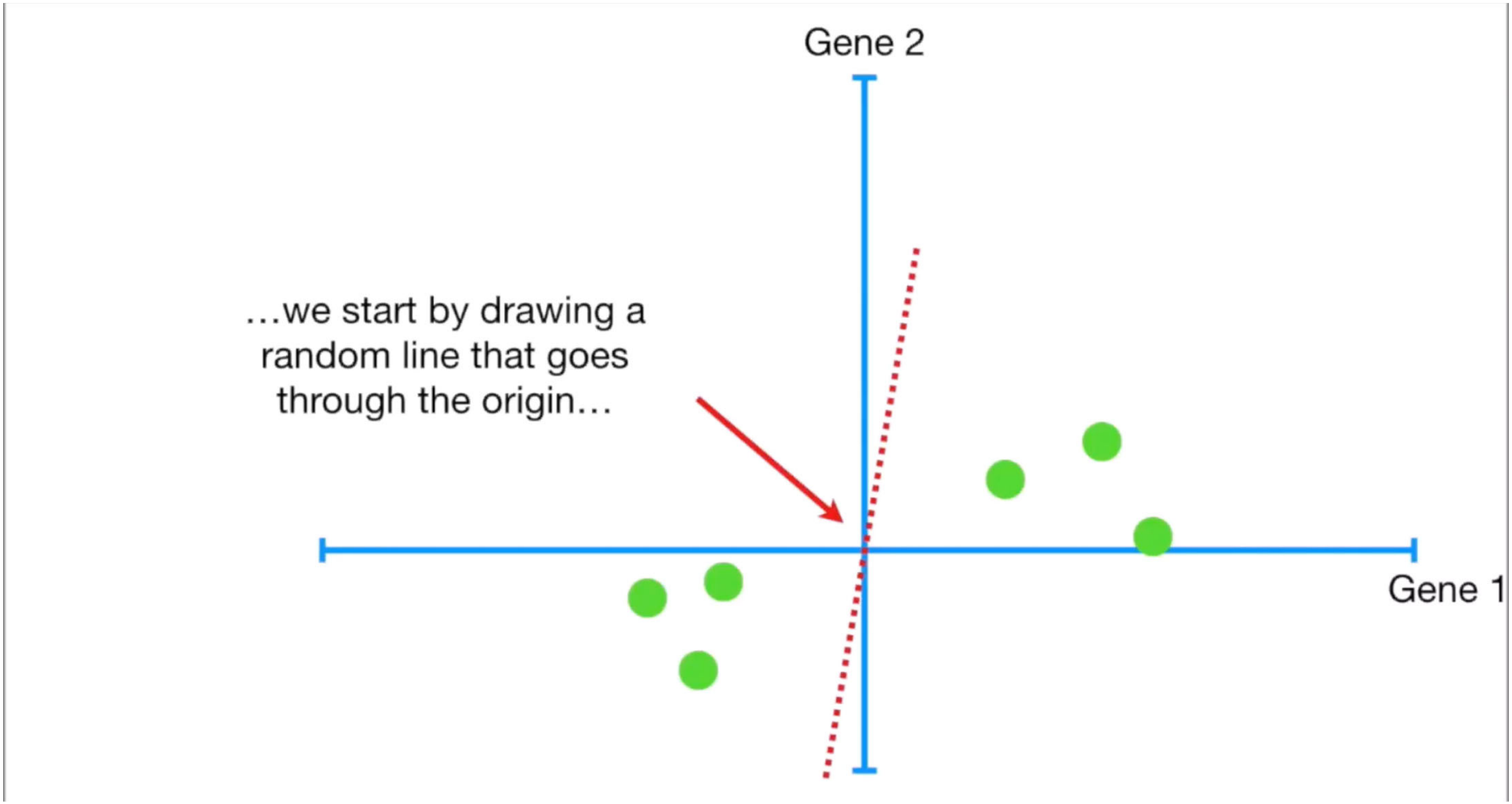


…and the average measurement for Gene 2.

	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

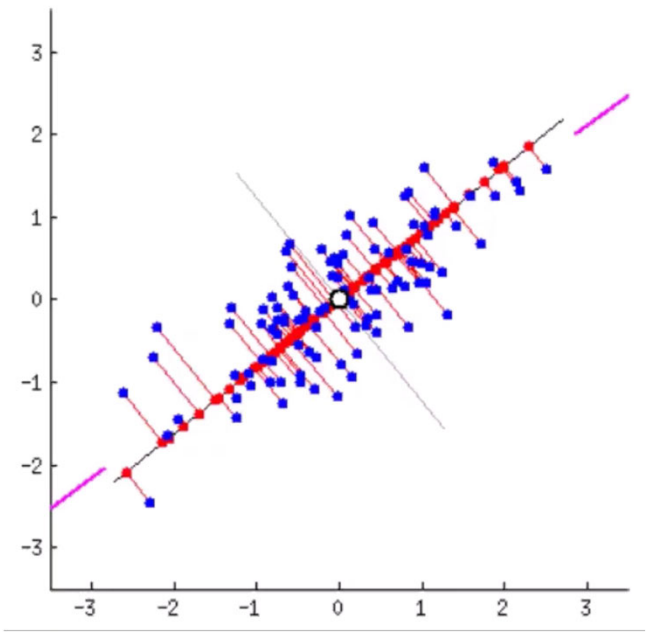



Gene 2
Gene 1

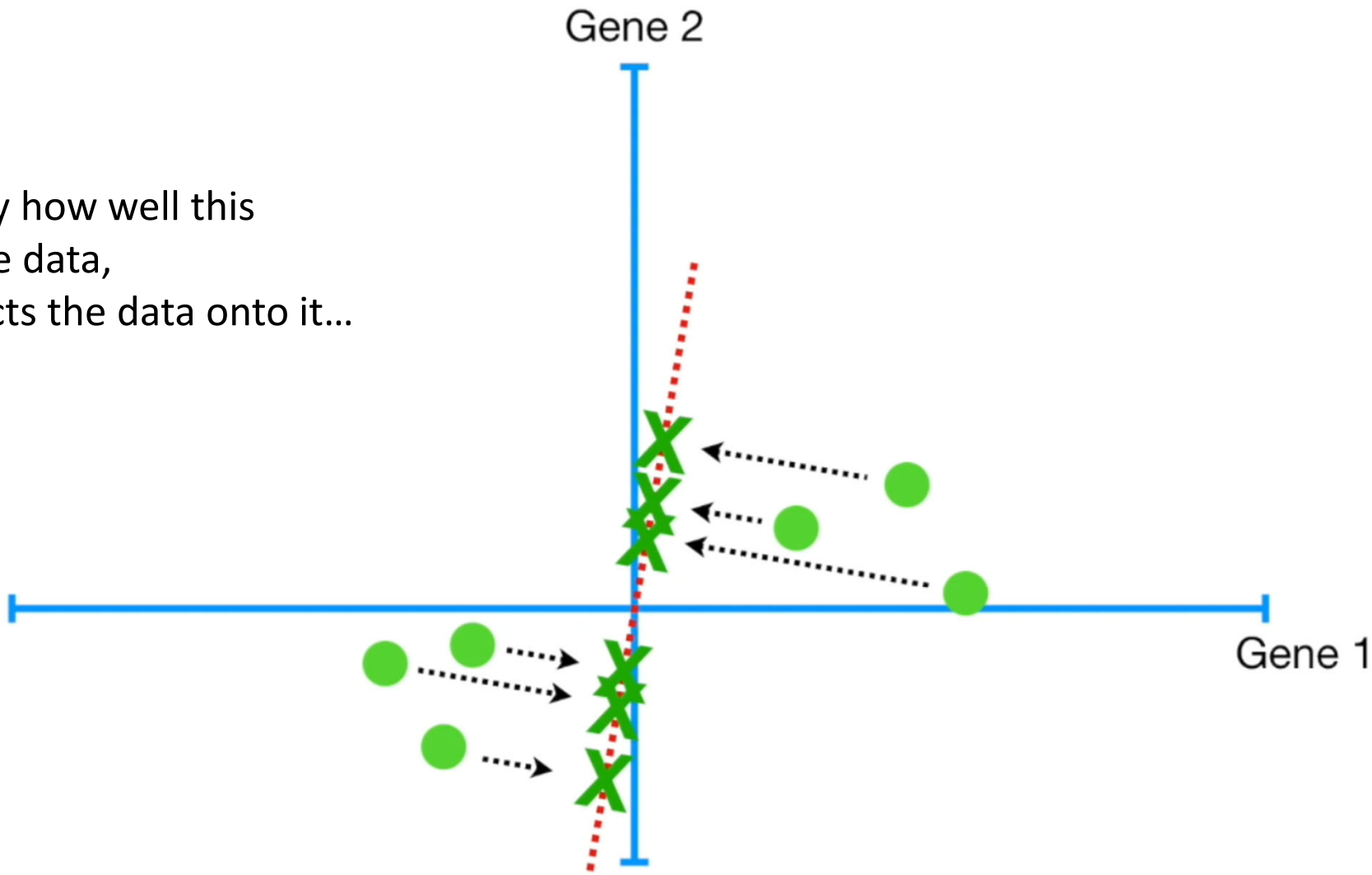

Gene 2
…we start by drawing a random line that goes through the origin…
Gene 1

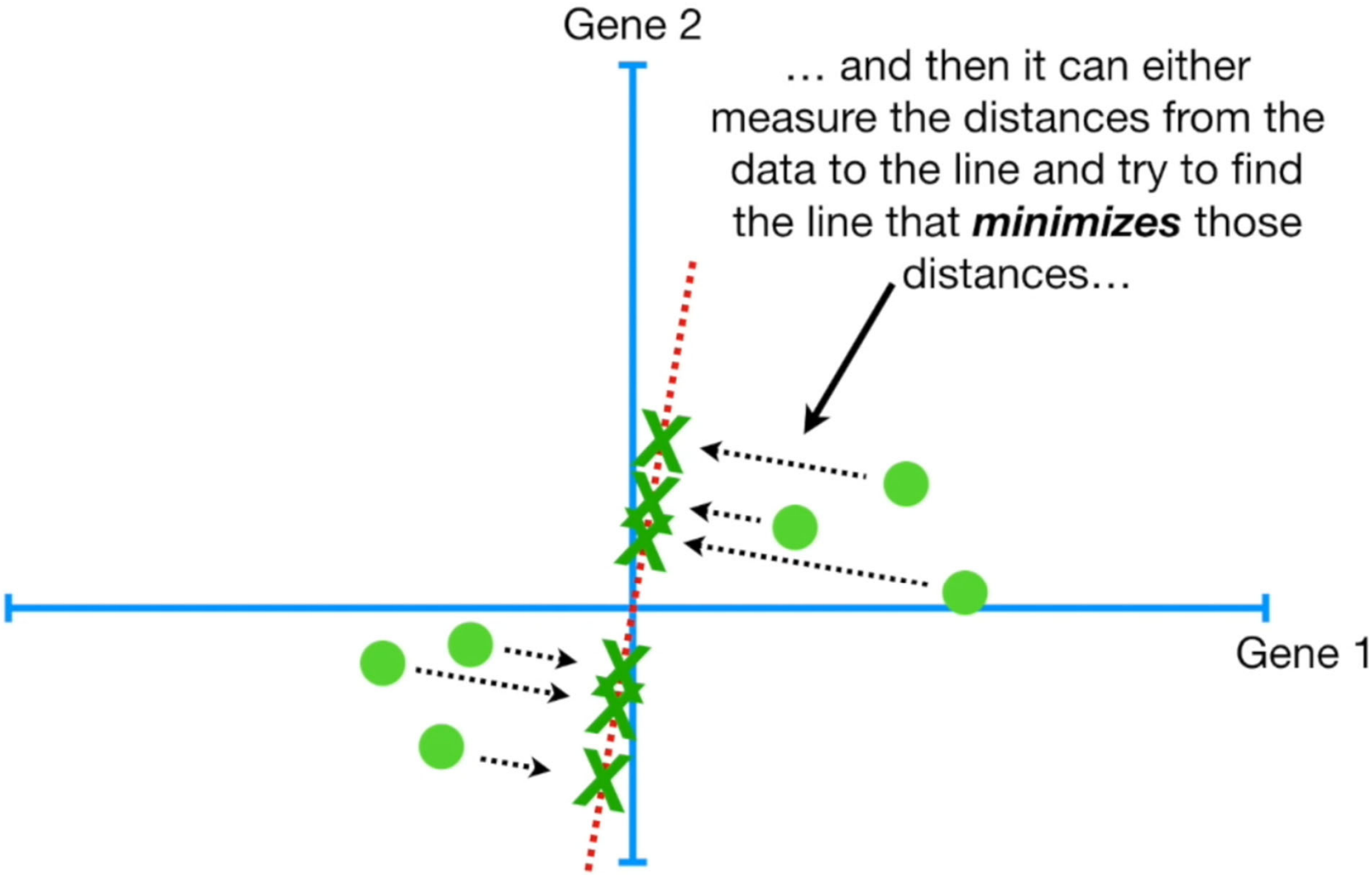
Step 2: Compute data covariance

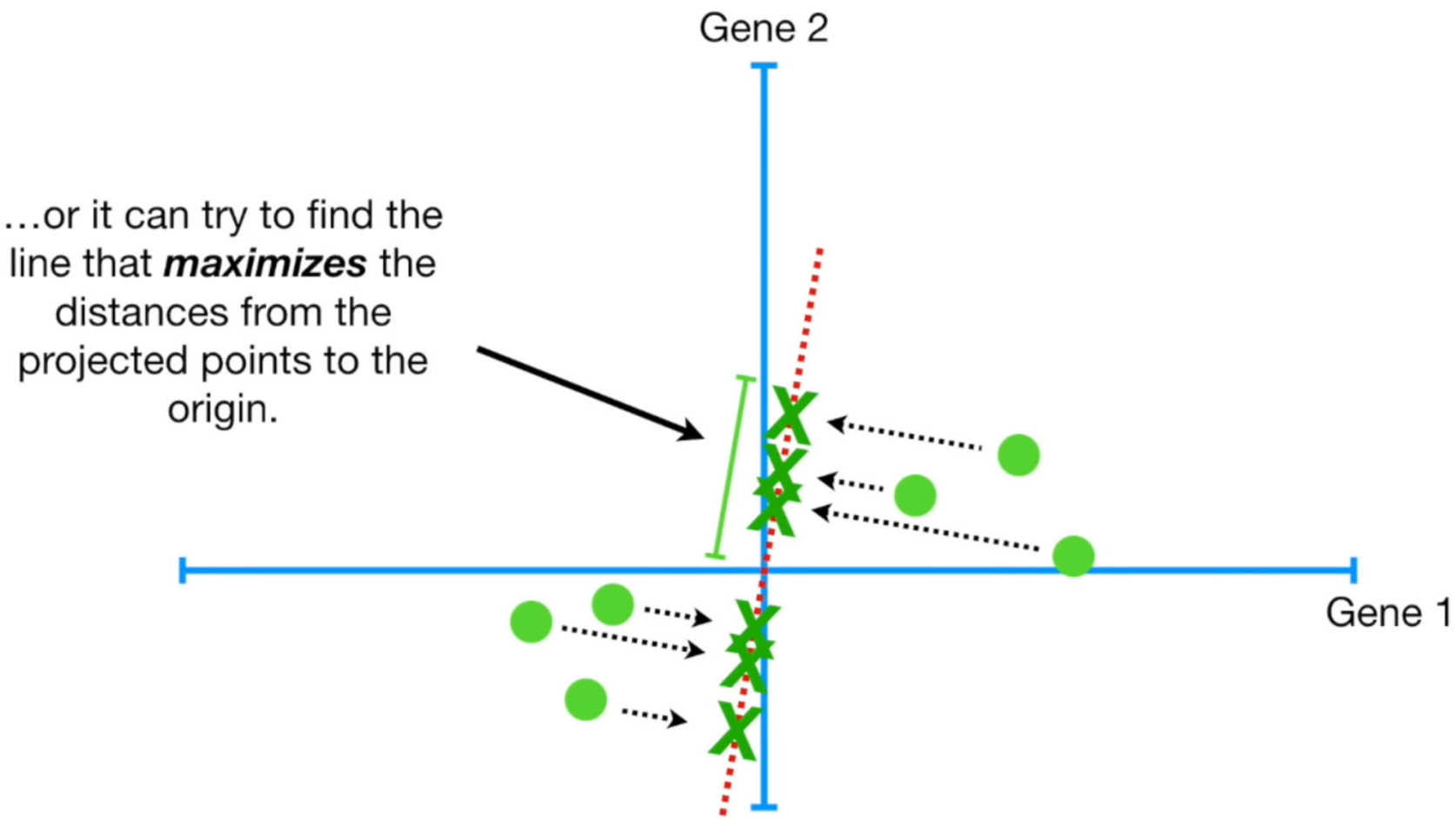
- Intuition: find vector that points in direction of maximal variance

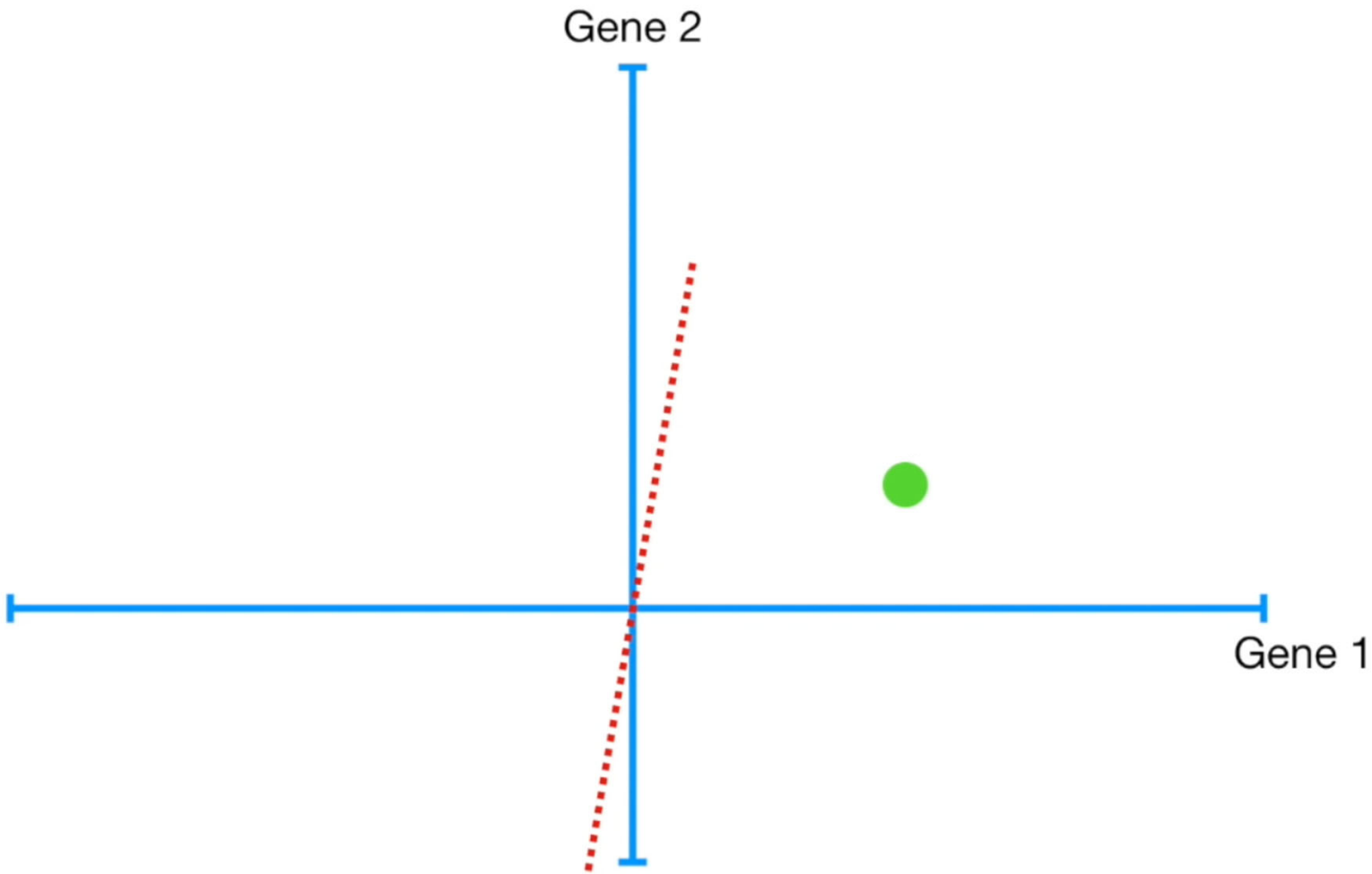


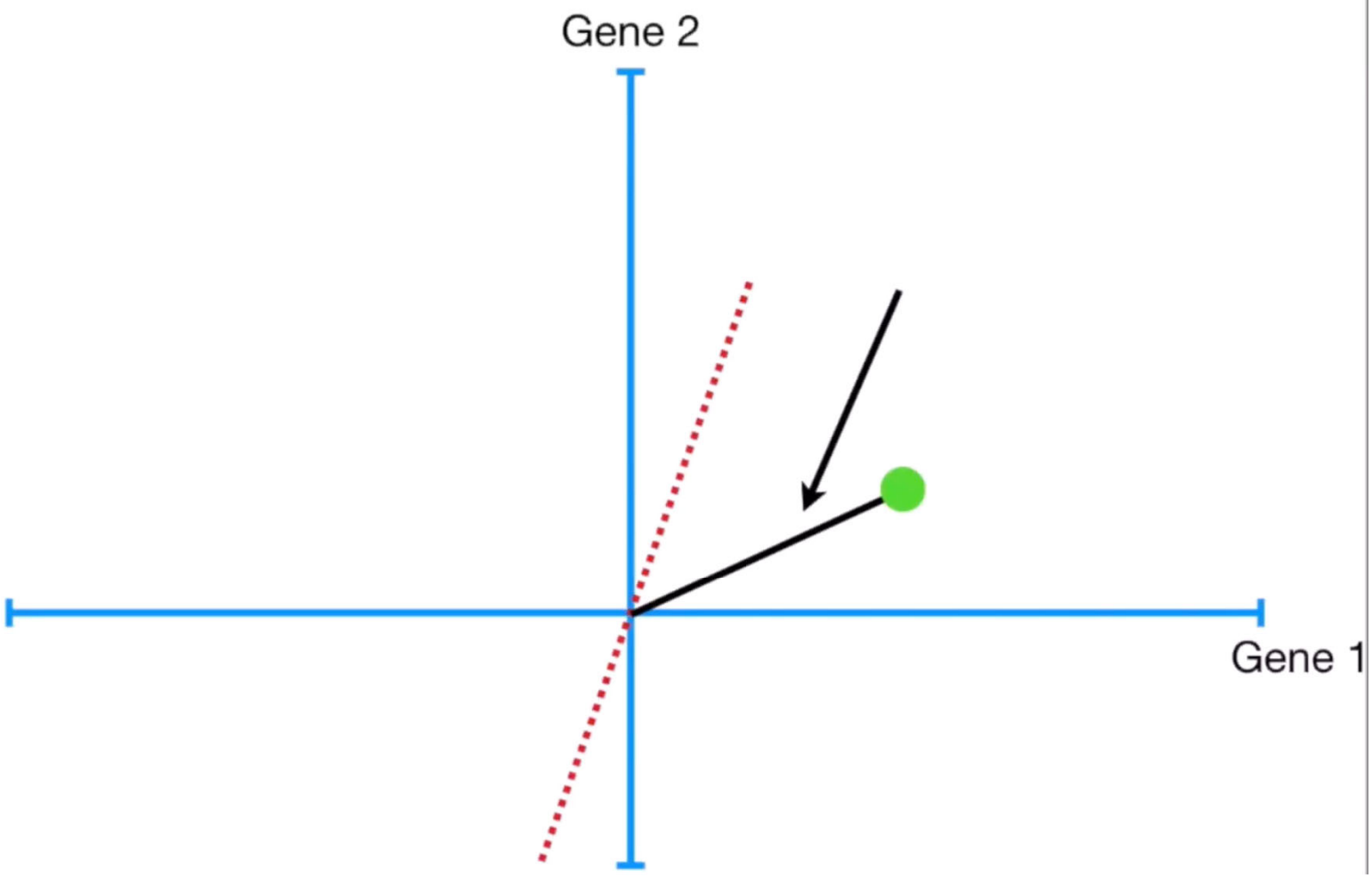
To quantify how well this
line fits the data,
PCA projects the data onto it…

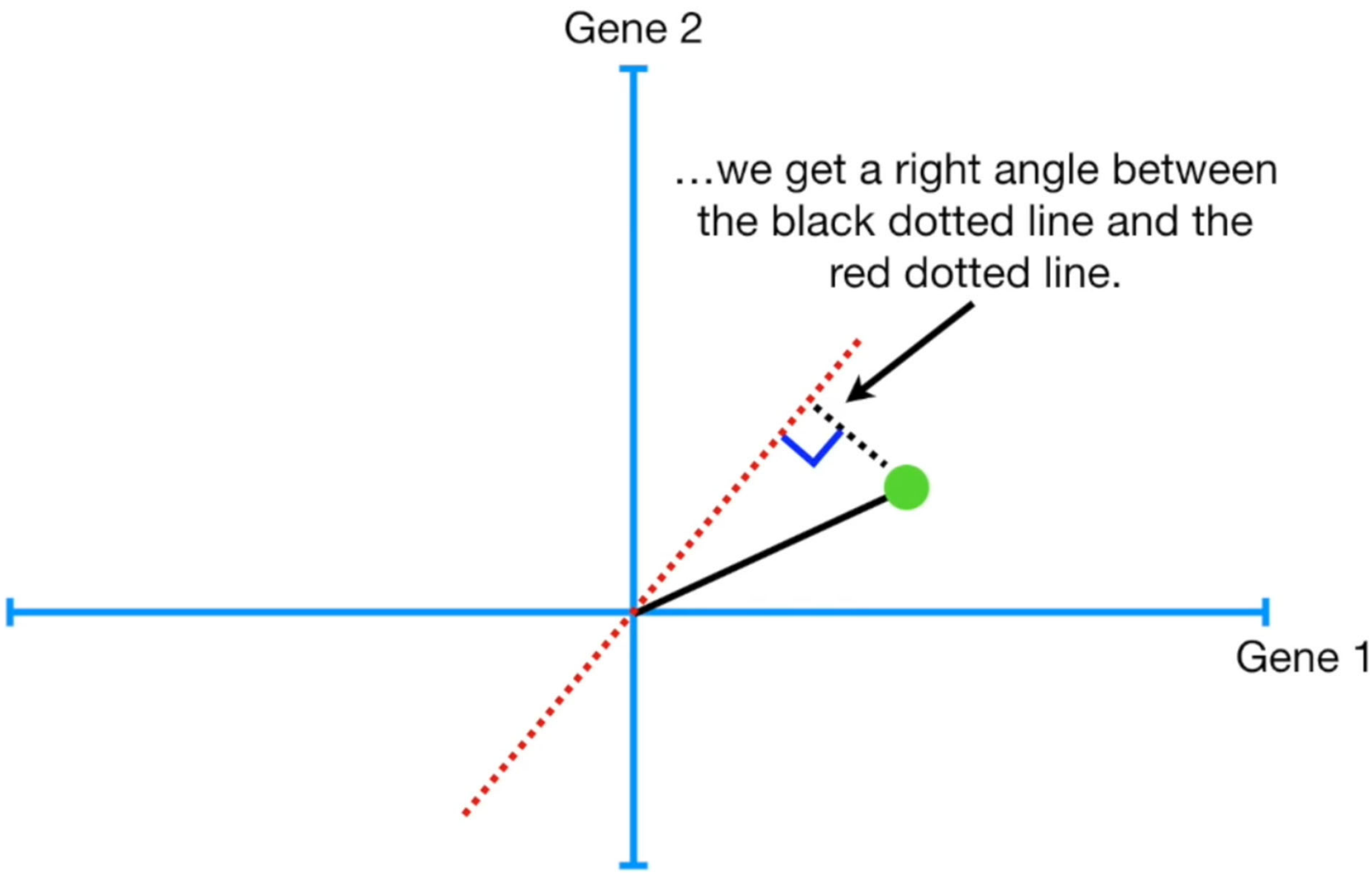



Gene 2
… and then it can either measure the distances from the data to the line and try to find the line that ***minimizes*** those distances…
Gene 1


Gene 2
…or it can try to find the line that ***maximizes*** the distances from the projected points to the origin.
Gene 1

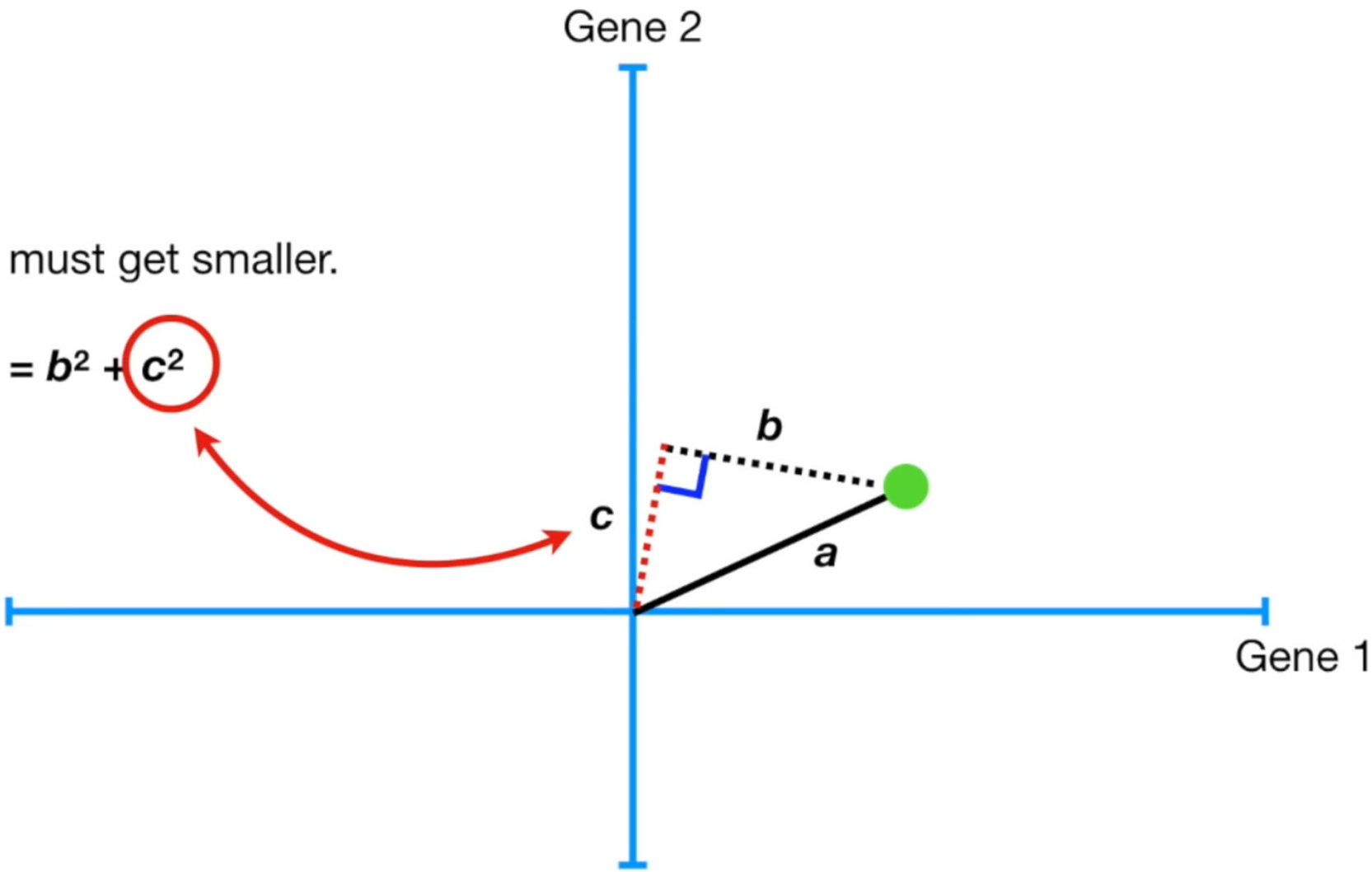

Gene 2
Gene 1


Gene 2
Gene 1


Gene 2
…we get a right angle between the black dotted line and the red dotted line.
Gene 1

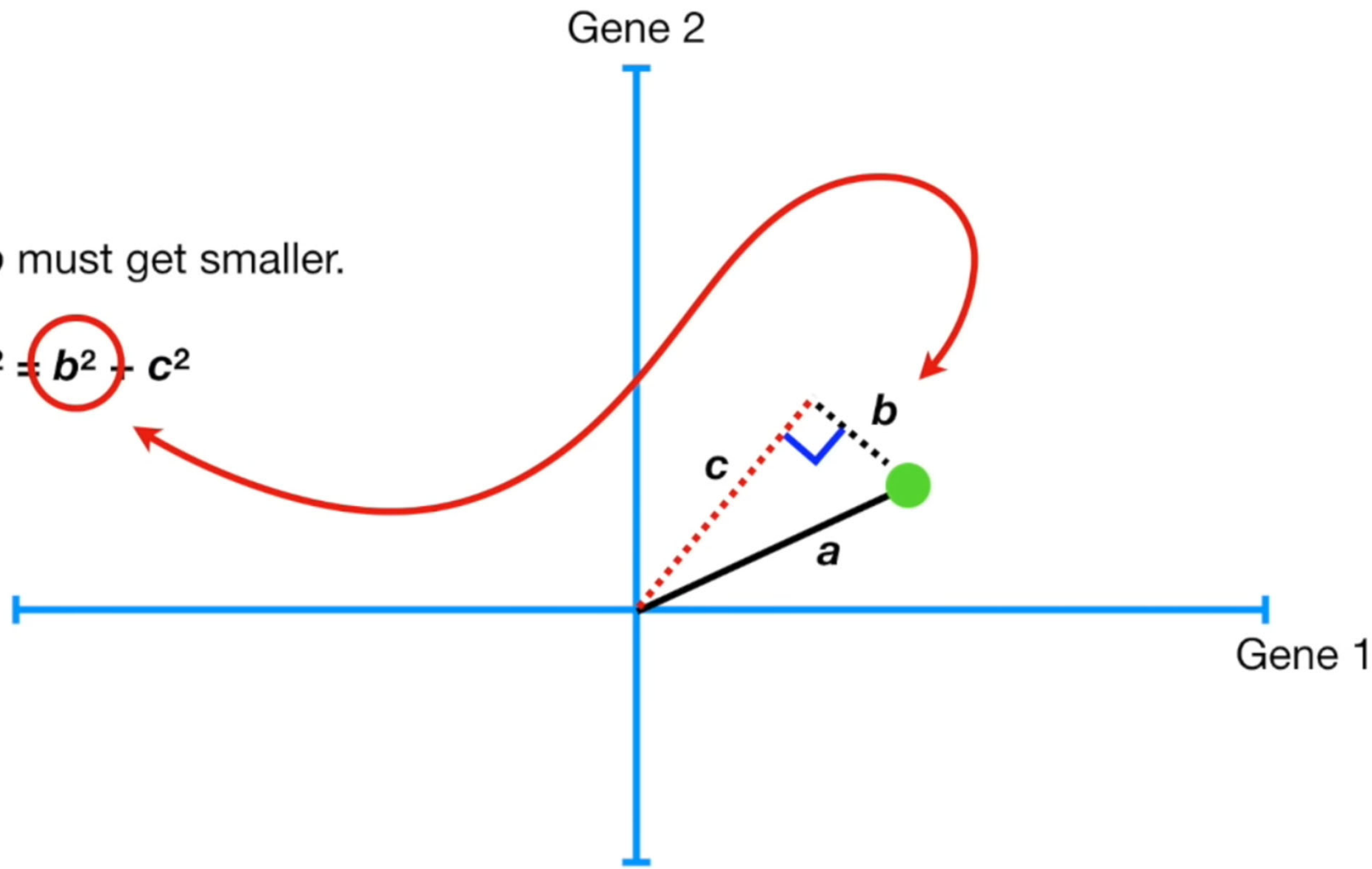
…then ***c*** must get smaller.

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



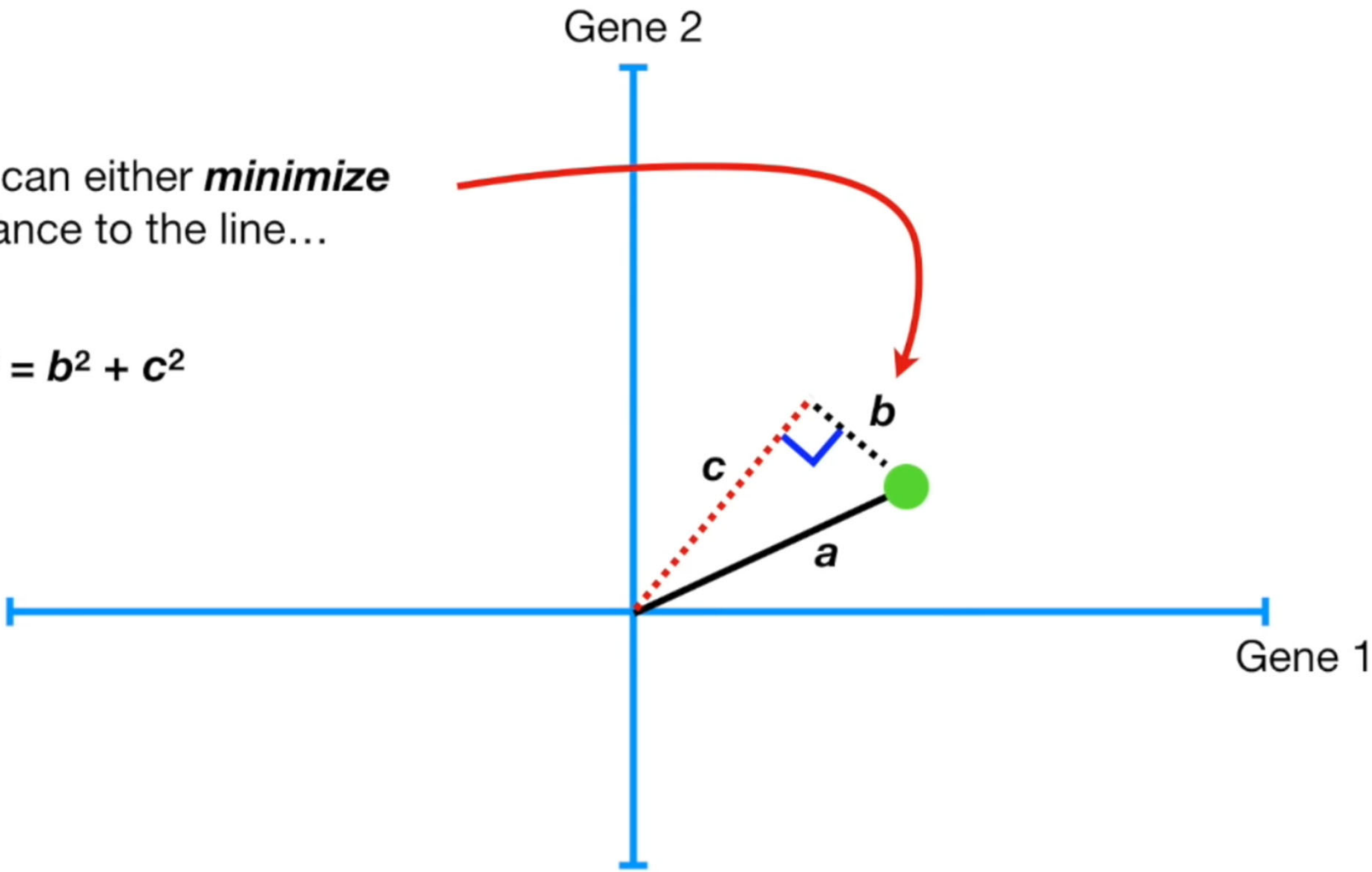
…then ***b*** must get smaller.

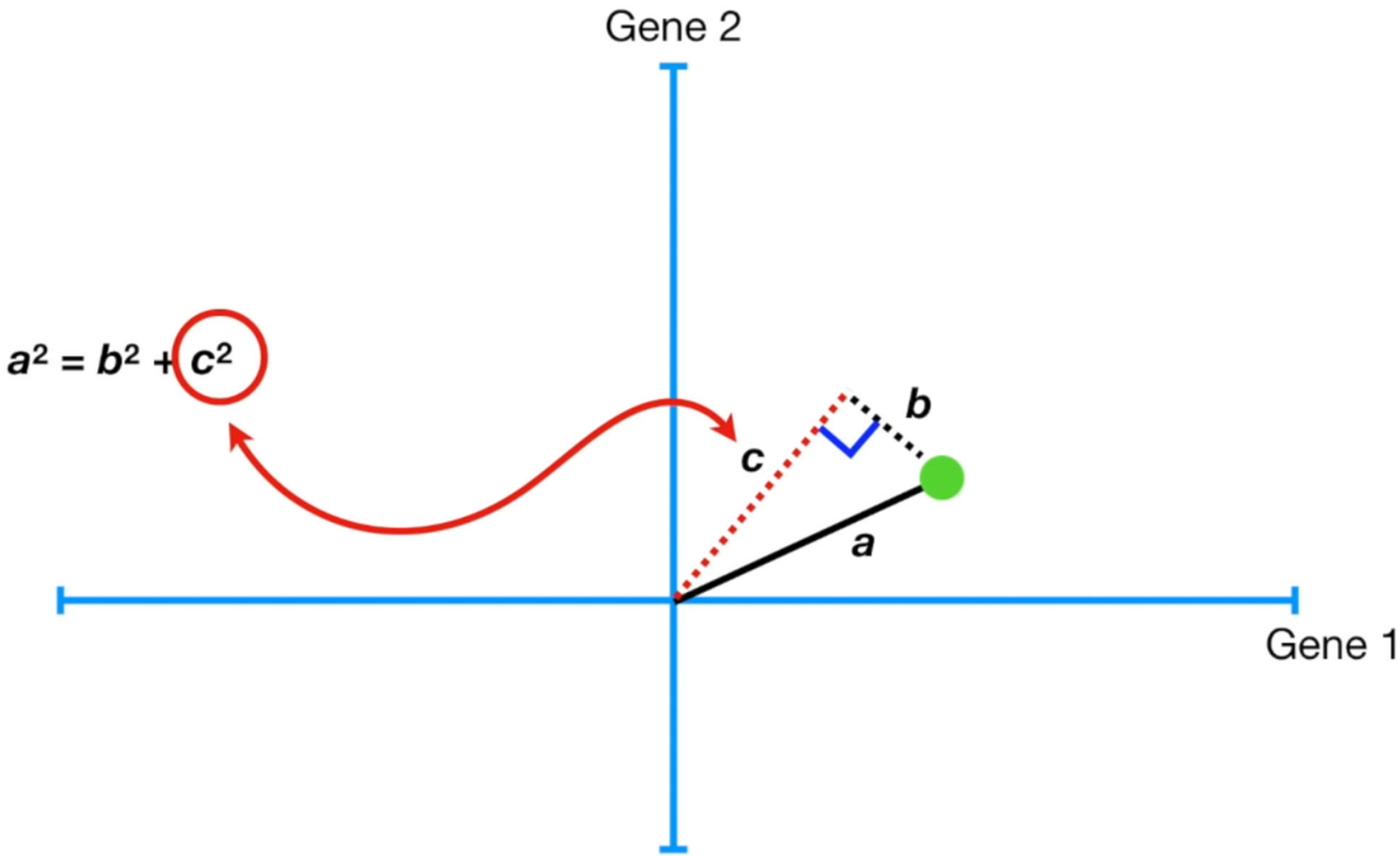
$$a^2 = b^2 + c^2$$

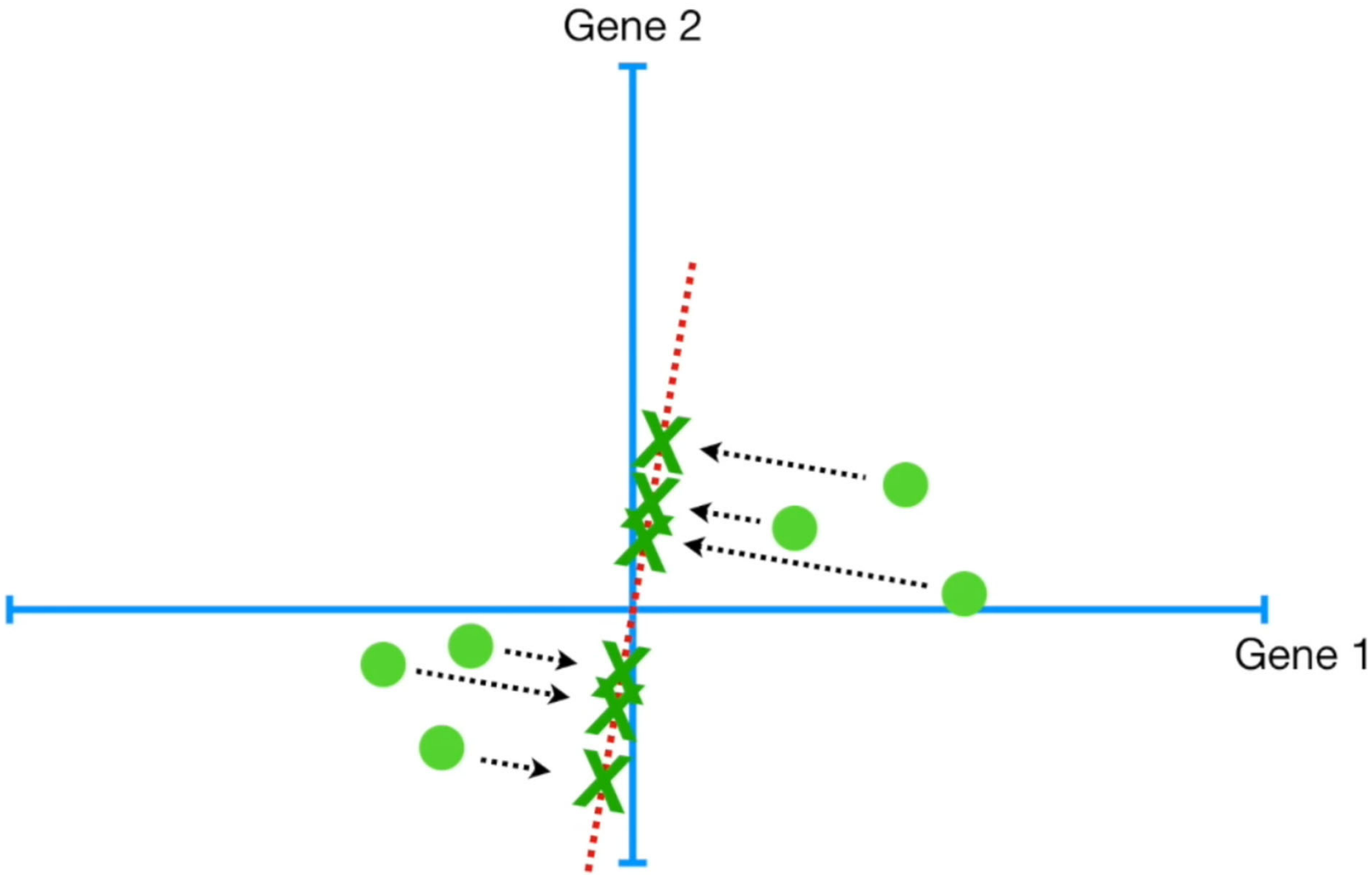


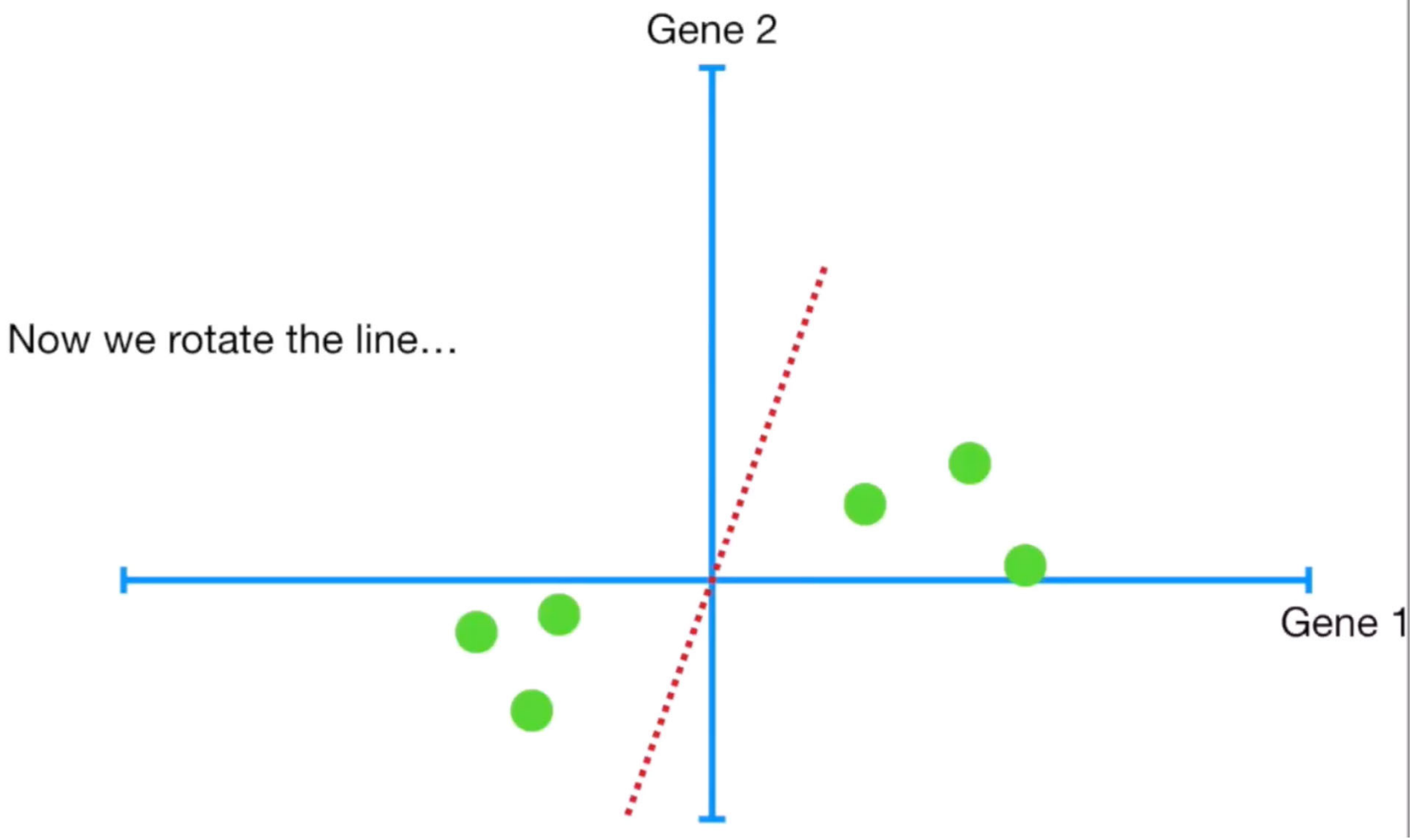
Thus, PCA can either ***minimize*** the distance to the line…

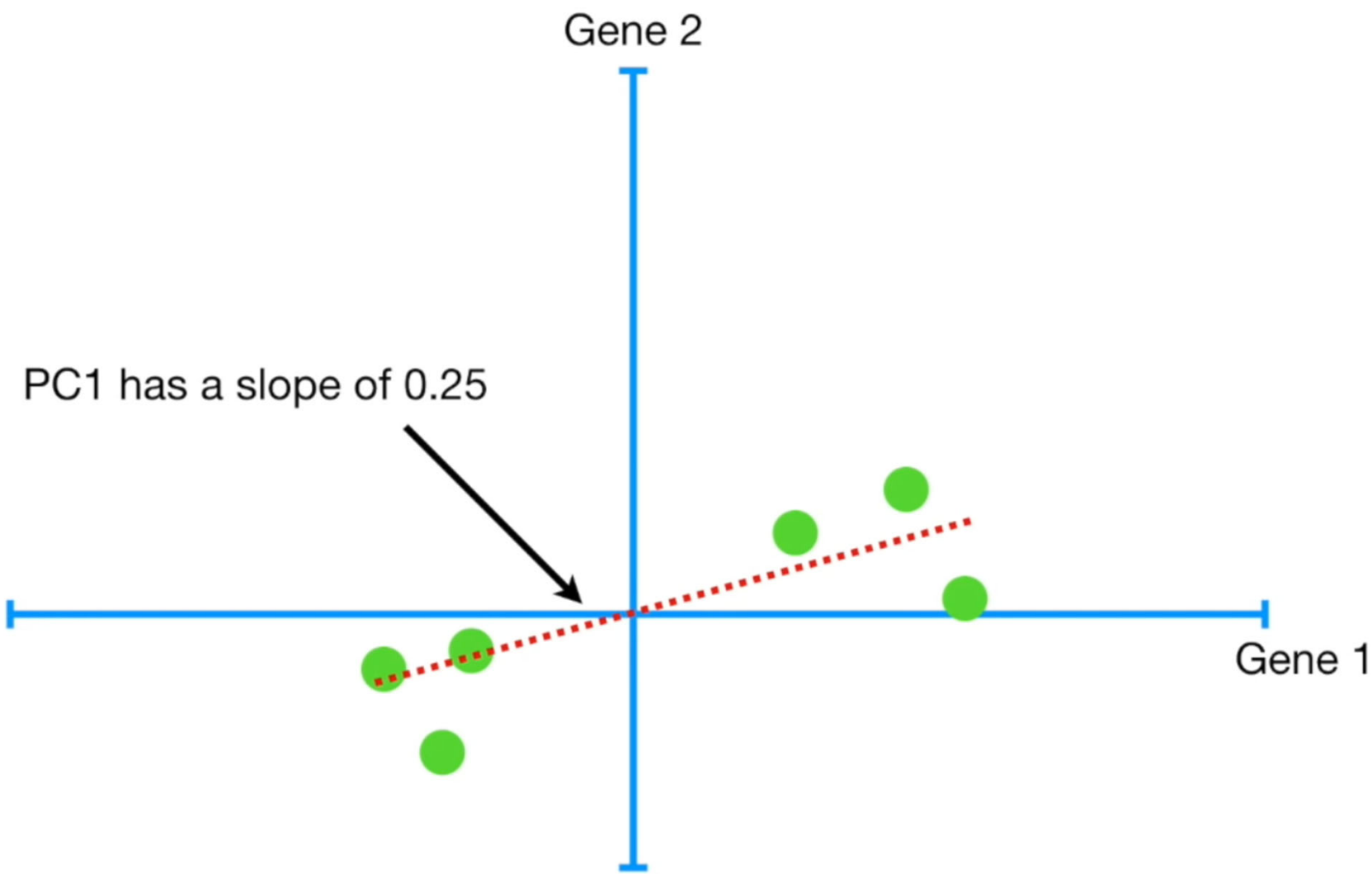
$$a^2 = b^2 + c^2$$

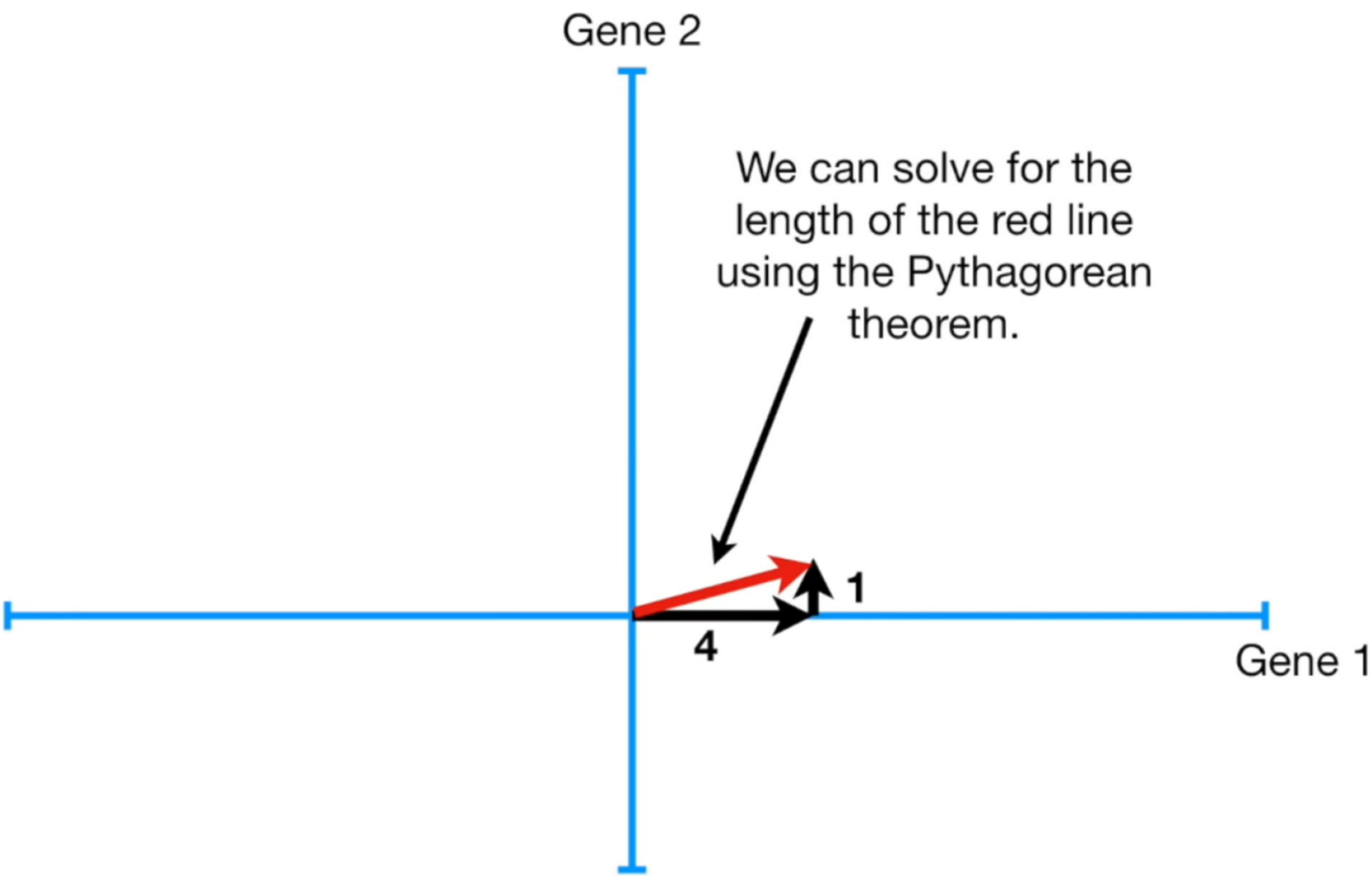


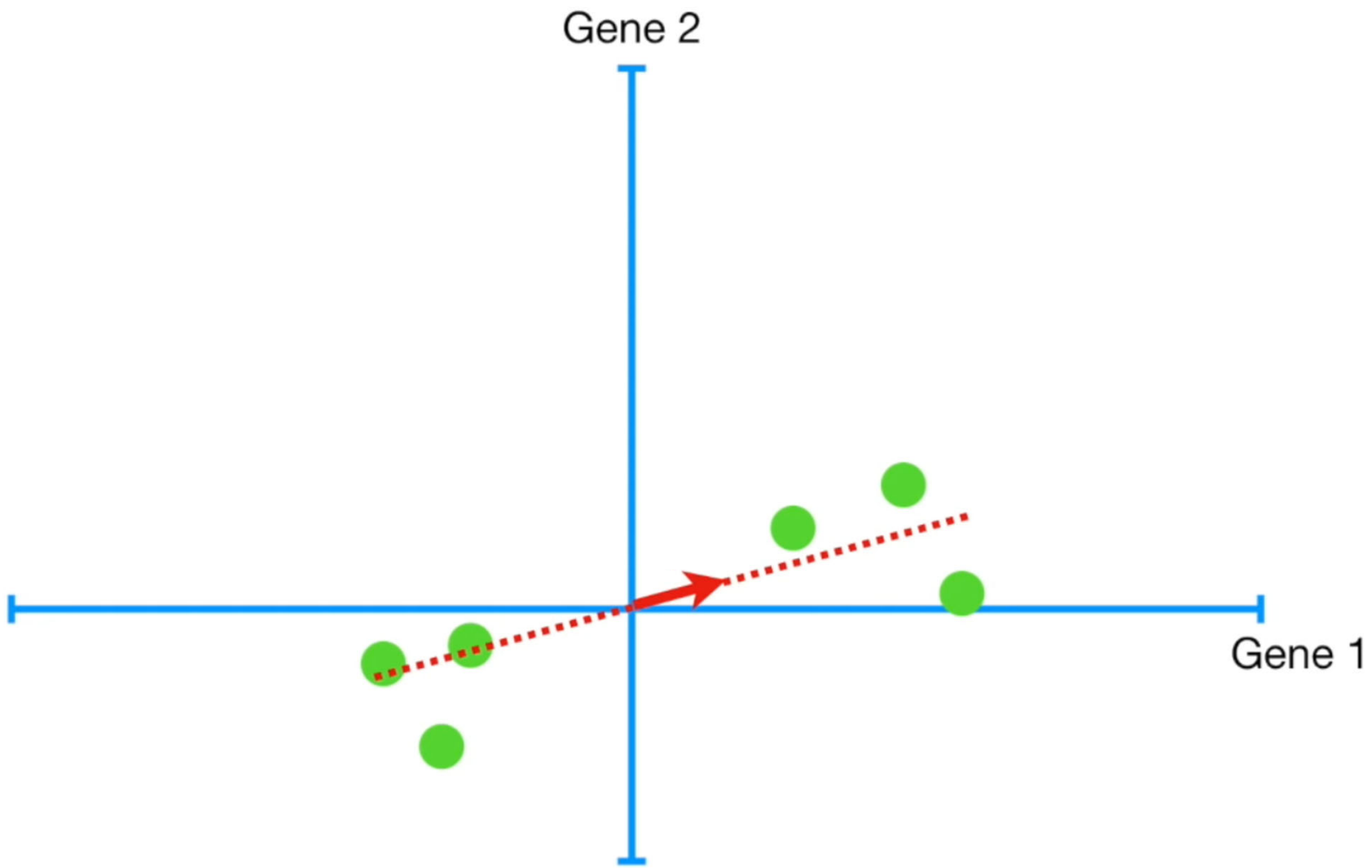

Gene 2
$a^2 = b^2 + c^2$
b
c
a
Gene 1


Gene 2
Gene 1

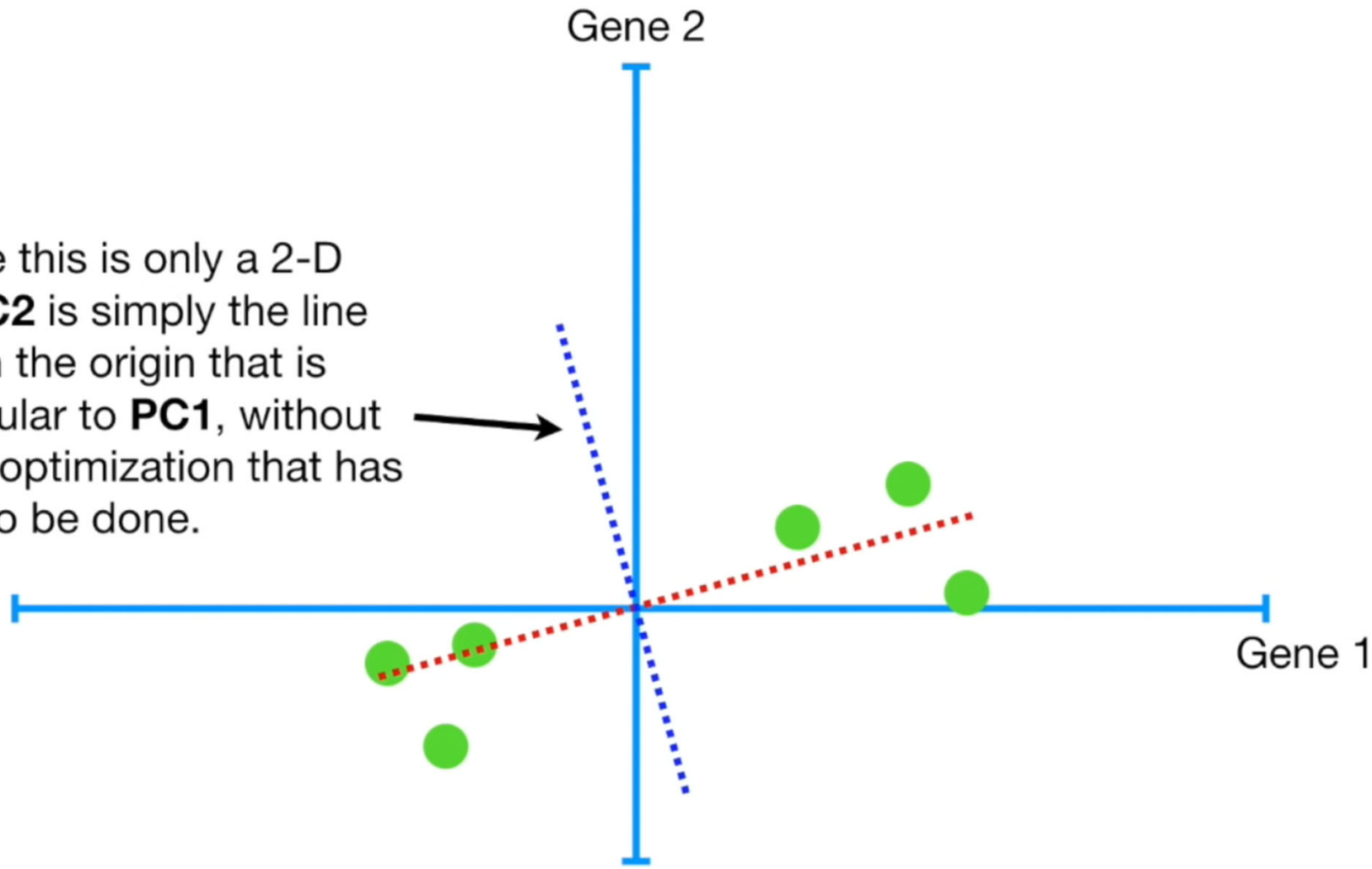

Gene 2
Now we rotate the line…
Gene 1

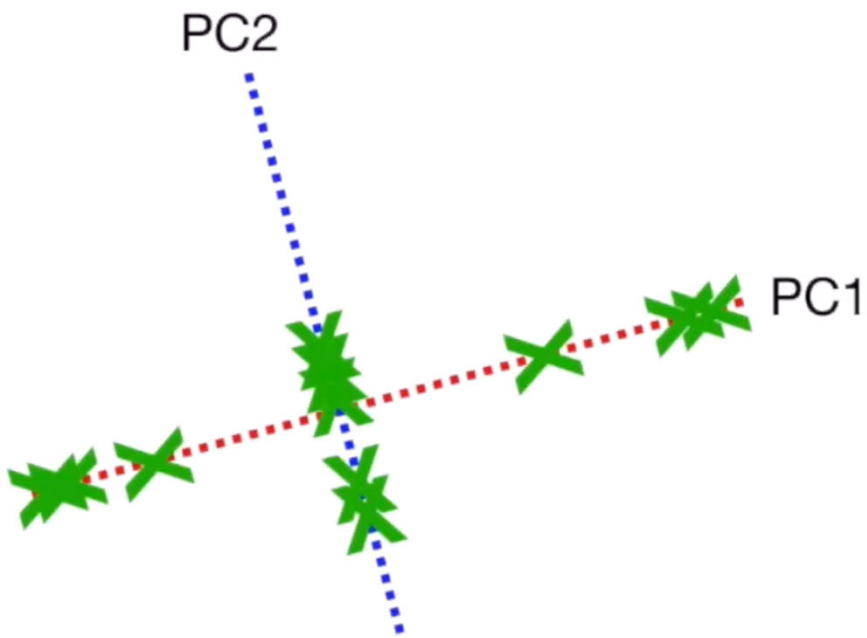

Gene 2
PC1 has a slope of 0.25
Gene 1

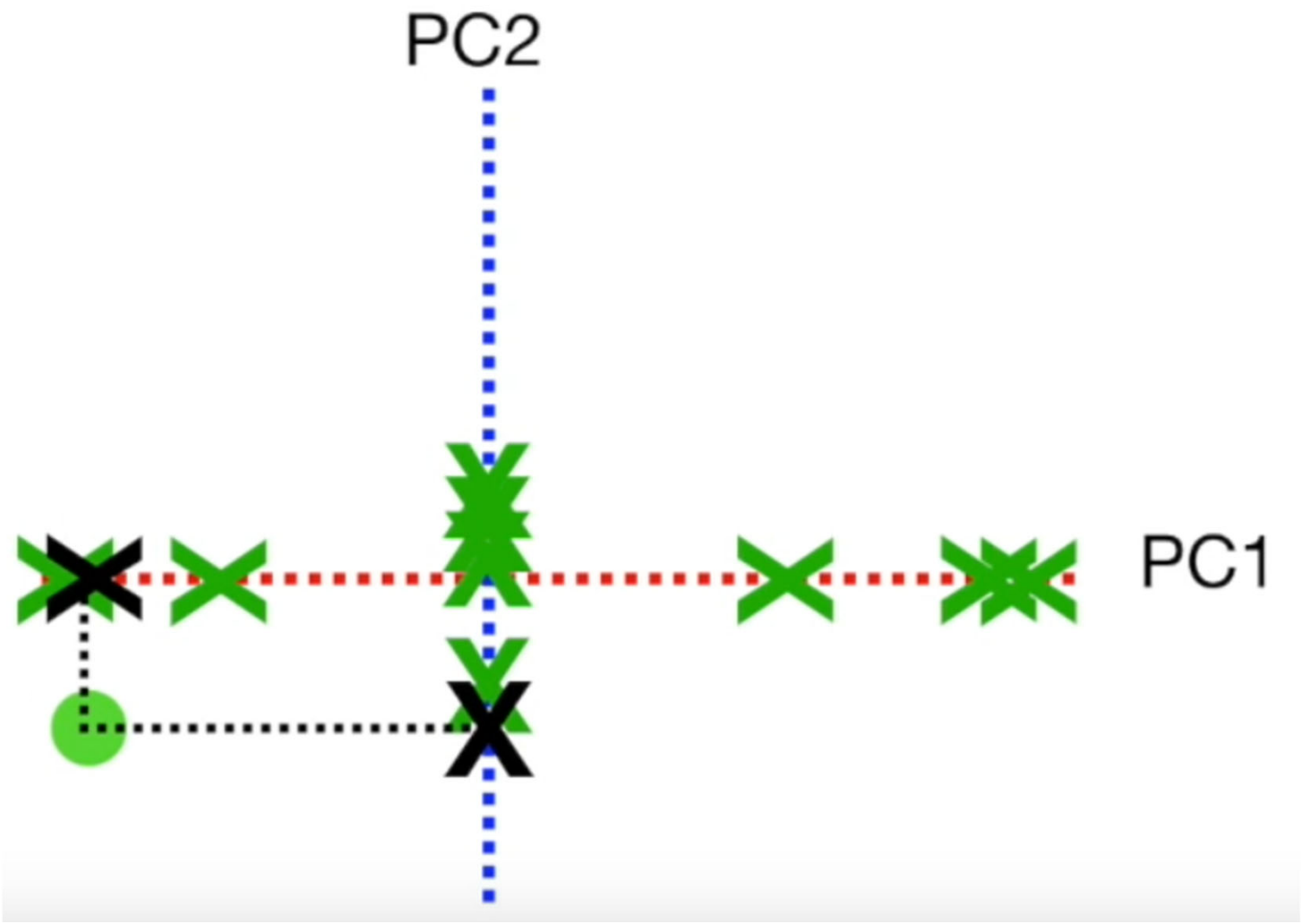

Gene 2
We can solve for the length of the red line using the Pythagorean theorem.
1
4
Gene 1


Gene 2
Gene 1

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.



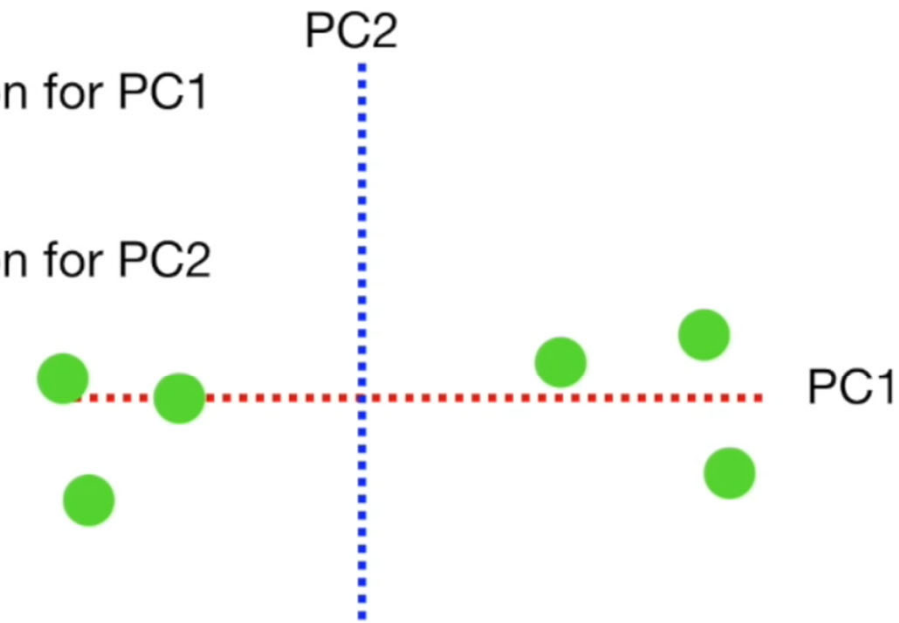

PC2
PC1


PC2
PC1

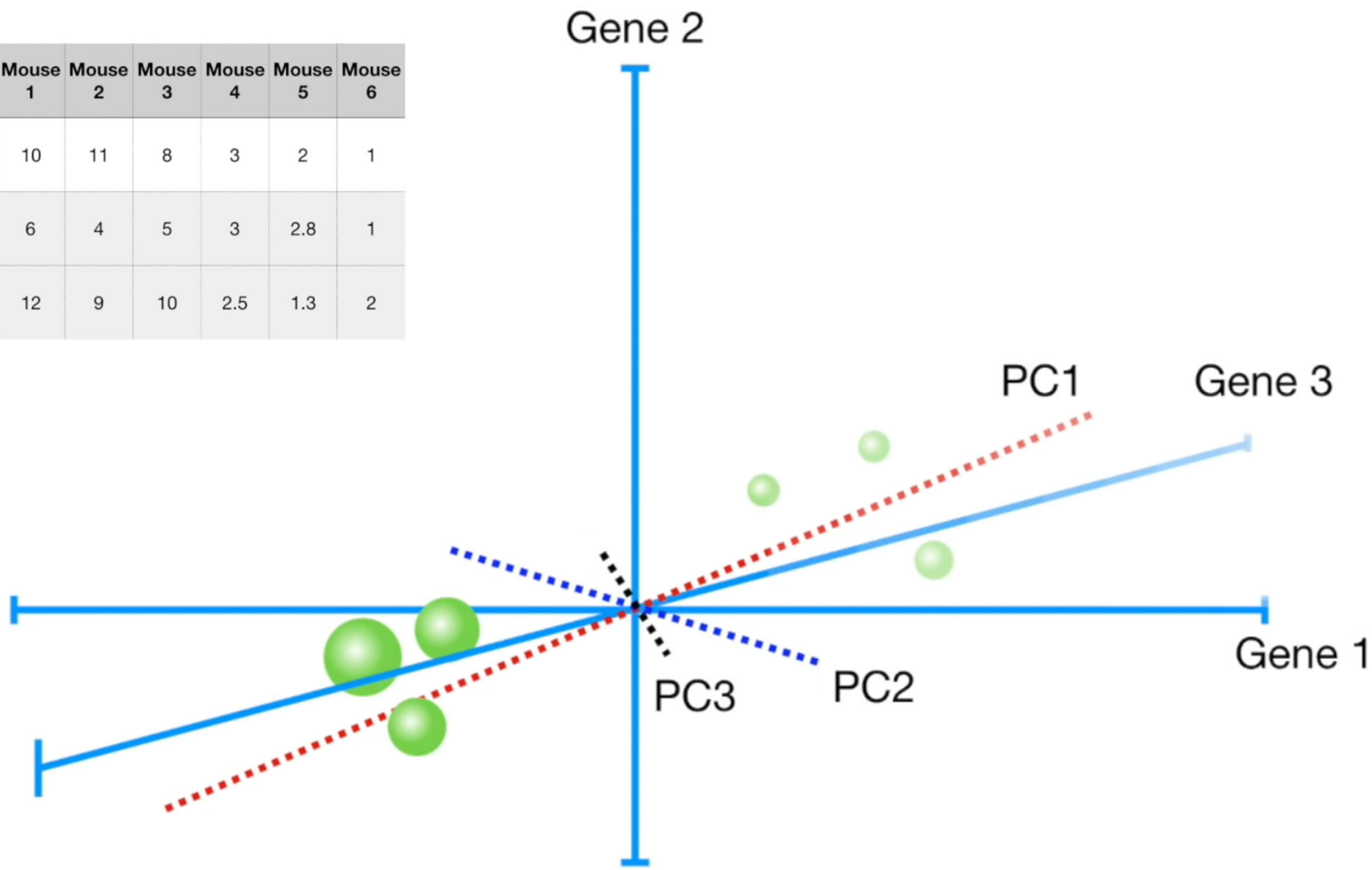
We can convert them into variation around the origin (0, 0) by dividing by the sample size minus 1 (i.e. n - 1).

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

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

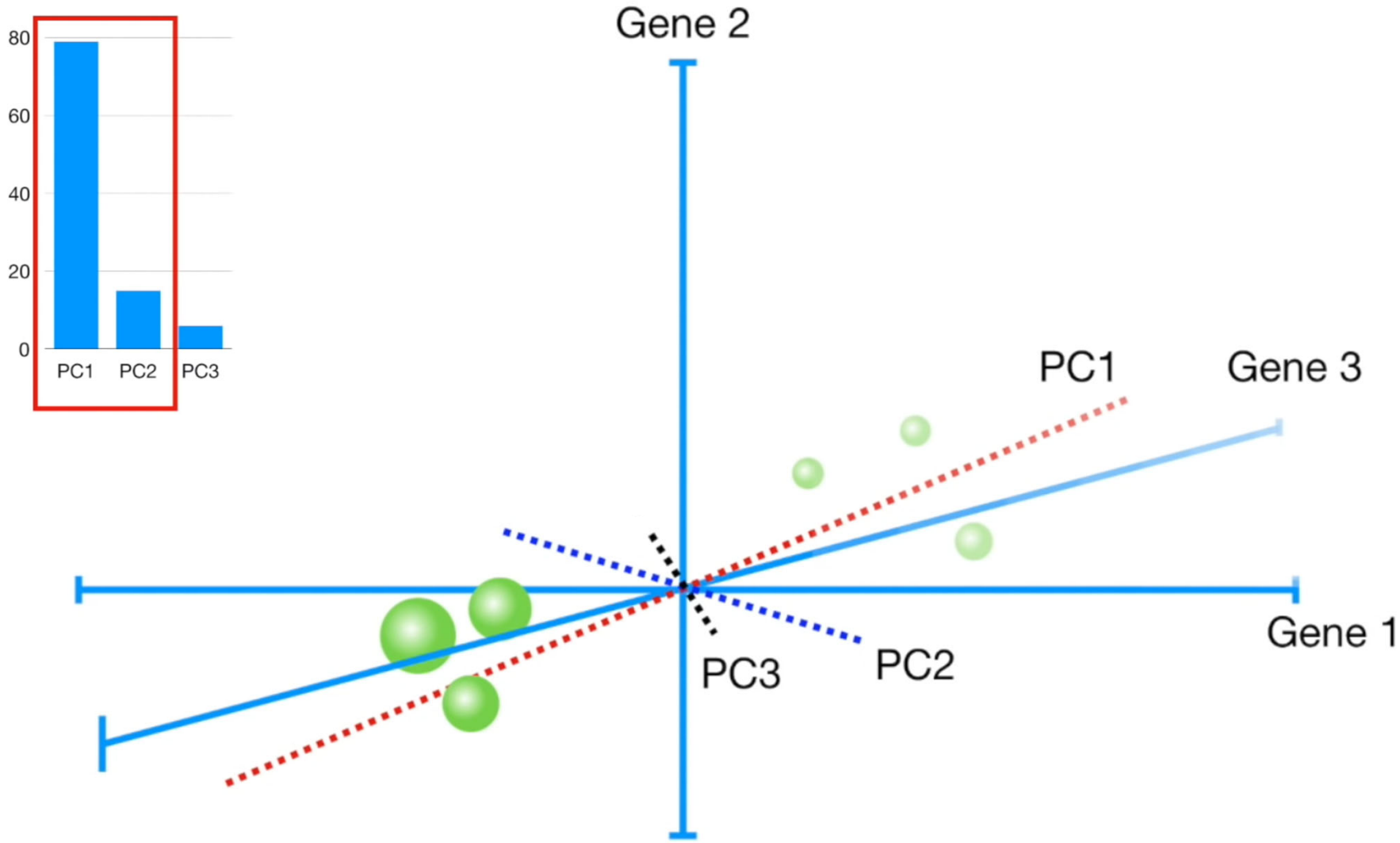


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1
Gene 3	12	9	10	2.5	1.3	2



Selecting components

- Use first components
- How many?

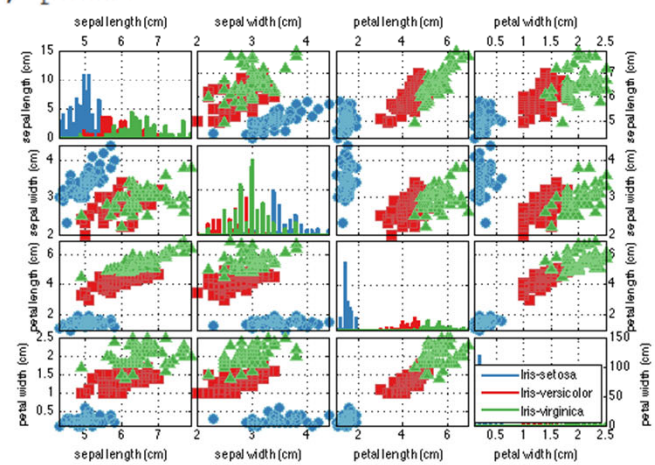

80
60
40
20
0
PC1
PC2
PC3
Gene 2
PC1
Gene 3
Gene 1
PC3
PC2

Let's try it out

- Iris dataset comes from UCI machine learning repository

```
import pandas as pd

url = "https://archive.ics.uci.edu/ml/machine-learning-databases
/iris/iris.data"

# load dataset into Pandas DataFrame
df = pd.read_csv(url, names=['sepal length','sepal width','petal
length','petal width','target'])
```

	sepal length	sepal width	petal length	petal width	target
0	5.1	3.5	1.4	0.2	Iris-setosa
1	4.9	3.0	1.4	0.2	Iris-setosa
2	4.7	3.2	1.3	0.2	Iris-setosa
3	4.6	3.1	1.5	0.2	Iris-setosa
4	5.0	3.6	1.4	0.2	Iris-setosa