

PCA Geometry

1D – Plot: Clear separation and depiction of data

One sample of data -

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

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



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

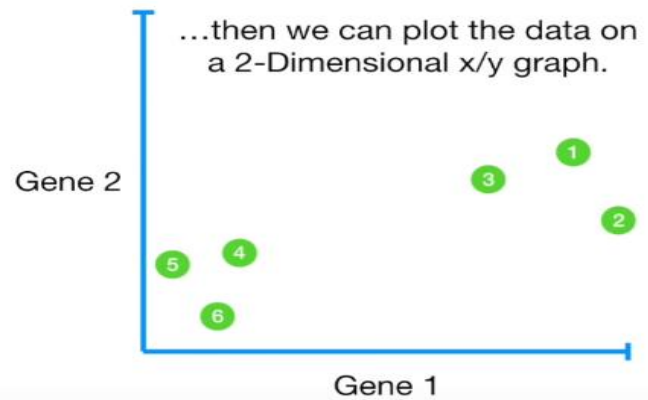
Even though it's a simple
graph, it shows us that mice 1,
2 and 3 are more similar to
each other than they are to
mice 4, 5 6.



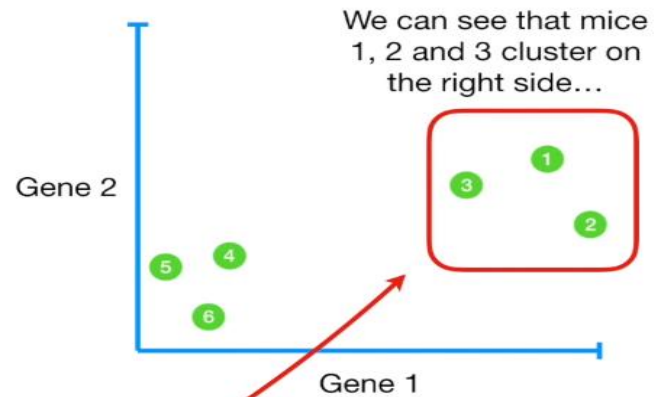
2D – Plot: Clear separation and depiction of data

Two sample of data -

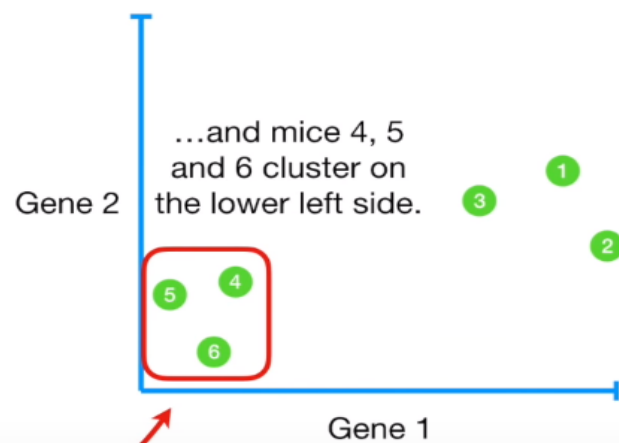
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



	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

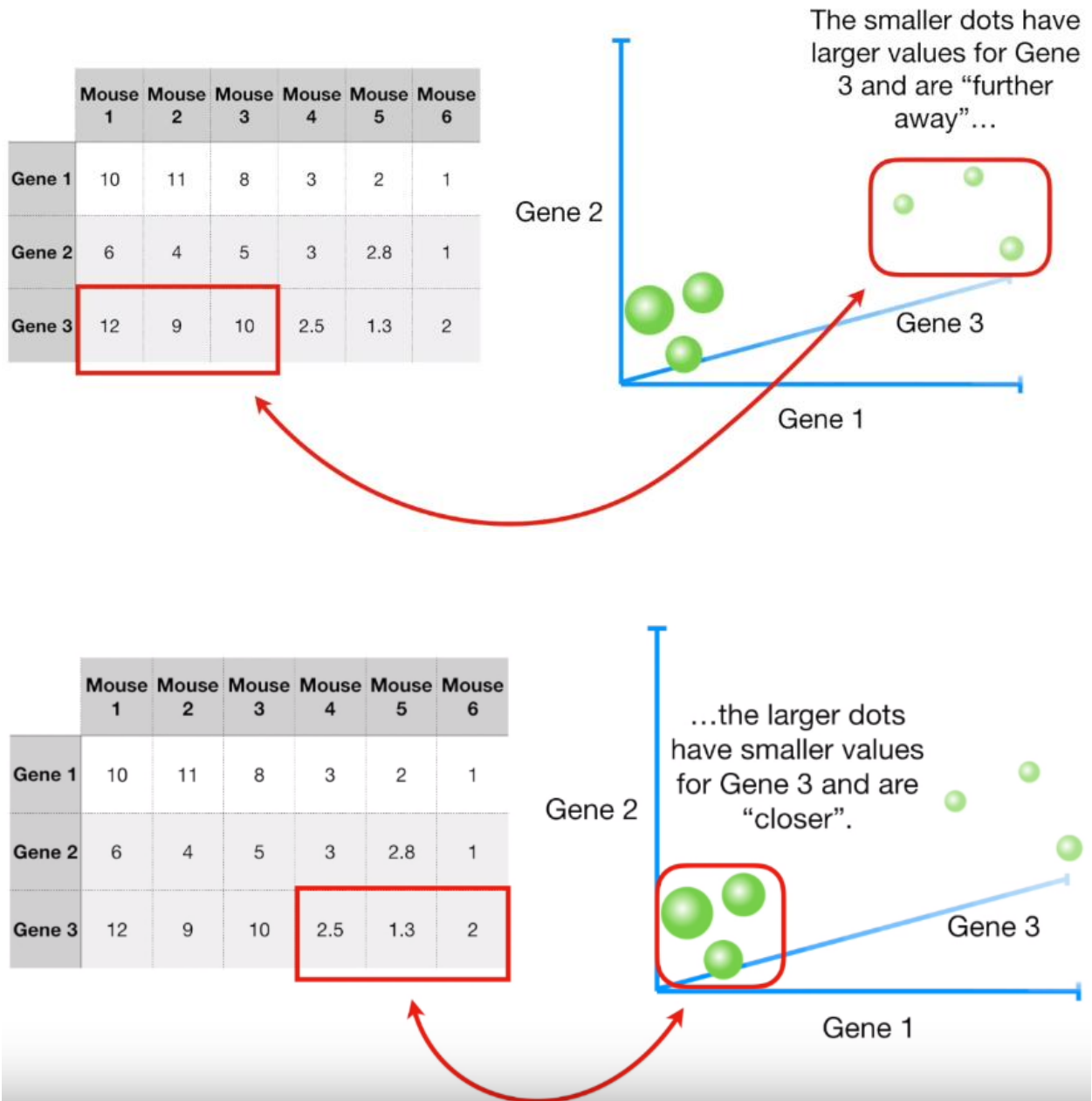


	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1



3D – Plot: Clear separation and depiction of data

Three sample of data -



4D – Plot: Not Possible

Introducing four sample of 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
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.

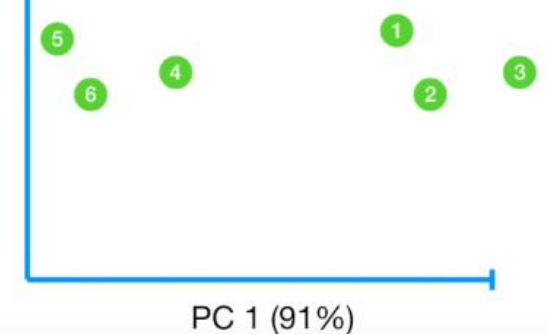
:(

Need for PCA for representing high dimensional data in lower space (dimension)

	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

PC 2
(4%)

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...

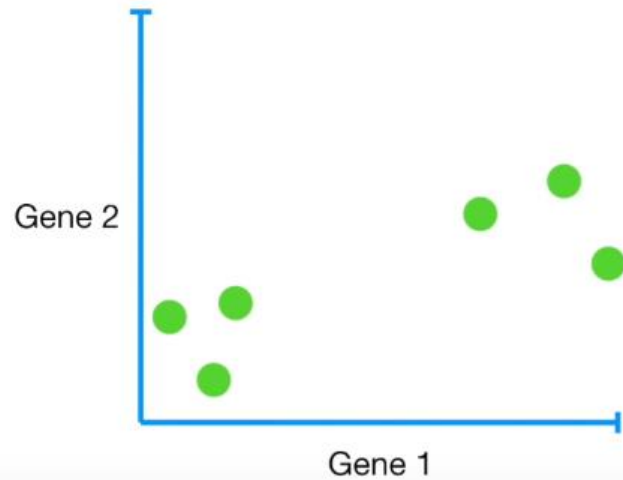


Example with two-dimensional data for illustrating PCA concept!!

PCA Functioning with 2D

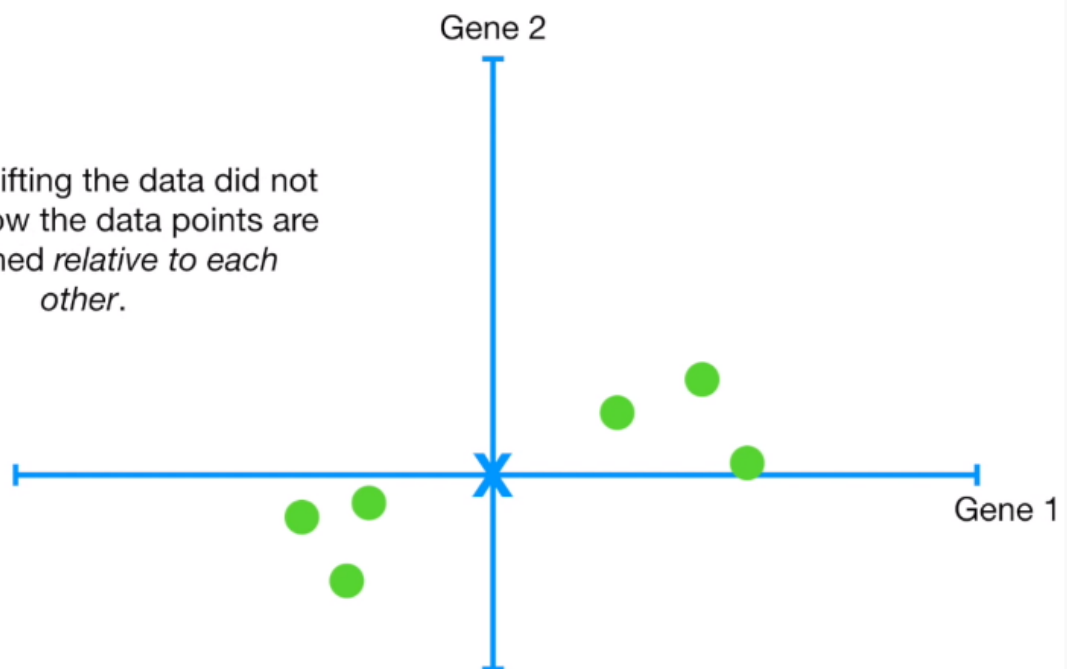
	Mouse 1	Mouse 2	Mouse 3	Mouse 4	Mouse 5	Mouse 6
Gene 1	10	11	8	3	2	1
Gene 2	6	4	5	3	2.8	1

We'll start by plotting the data...

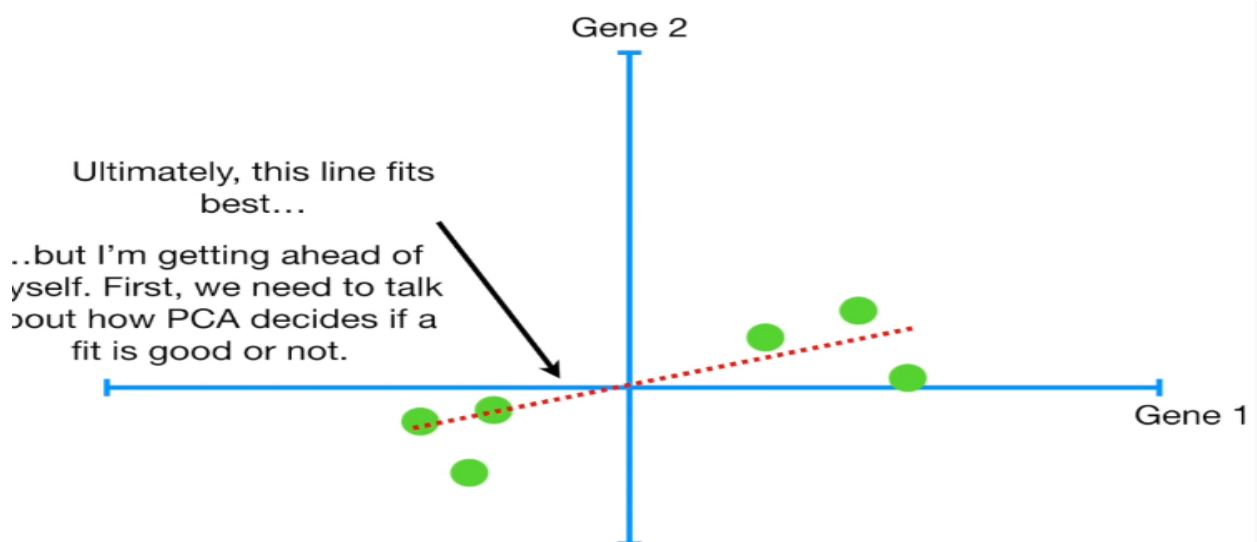
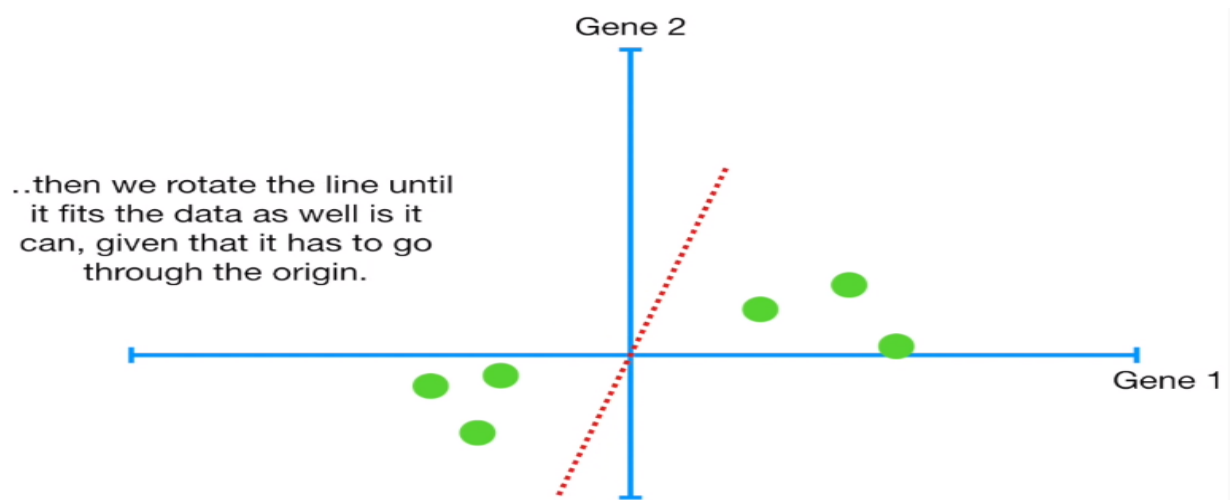
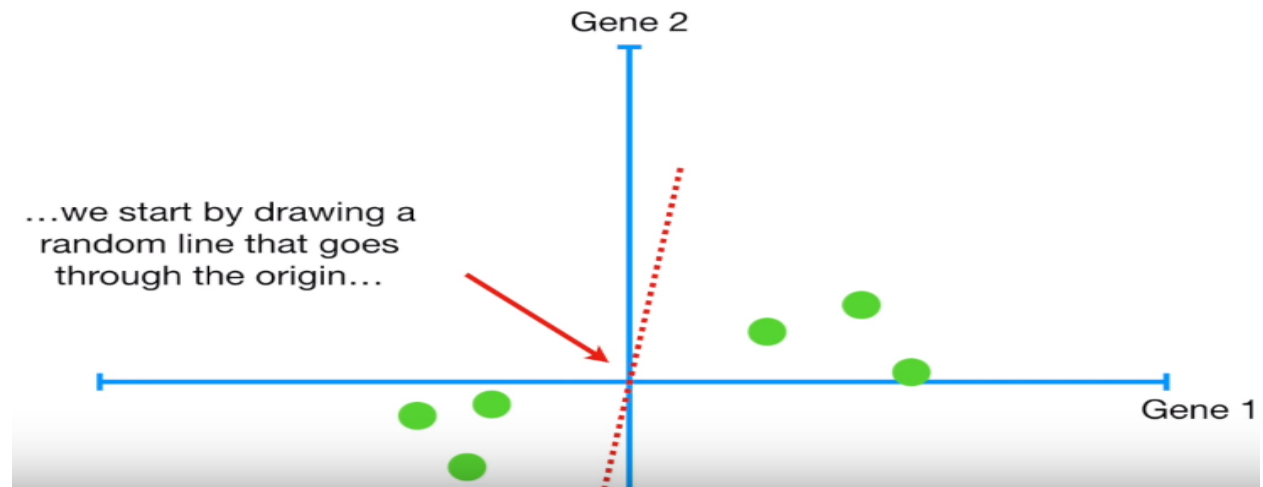


Column standardization: - Shifting to the origin

NOTE: Shifting the data did not change how the data points are positioned *relative to each other*.

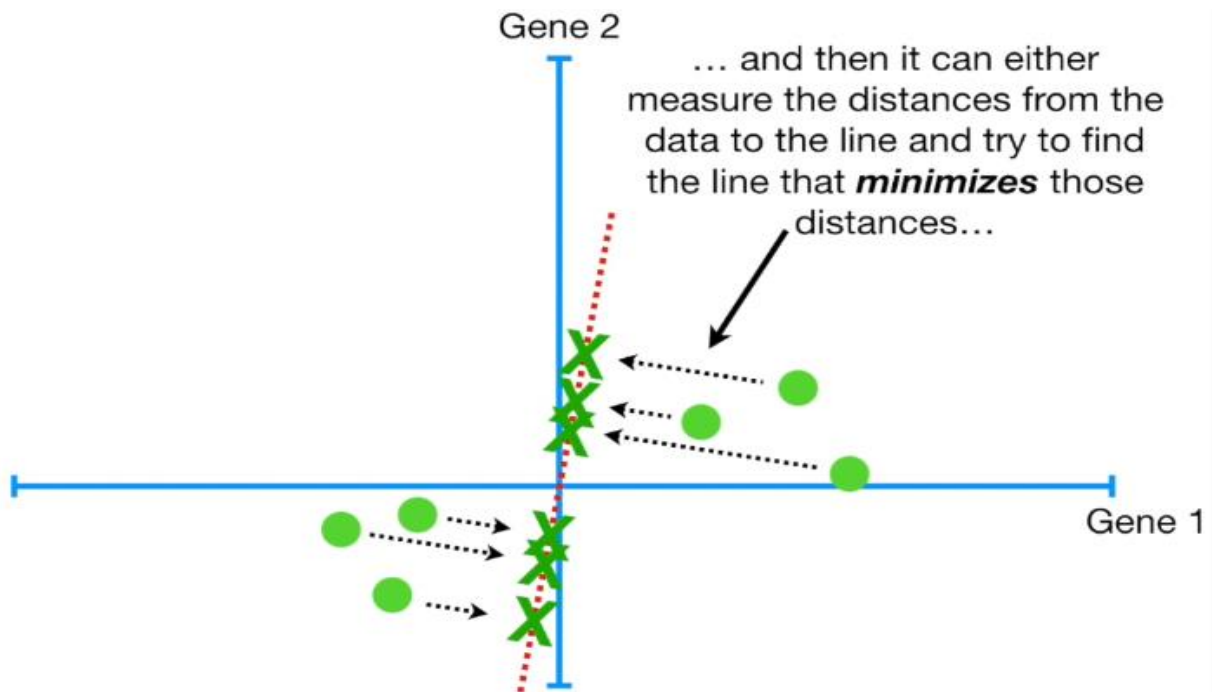


Estimating a best fit line: -

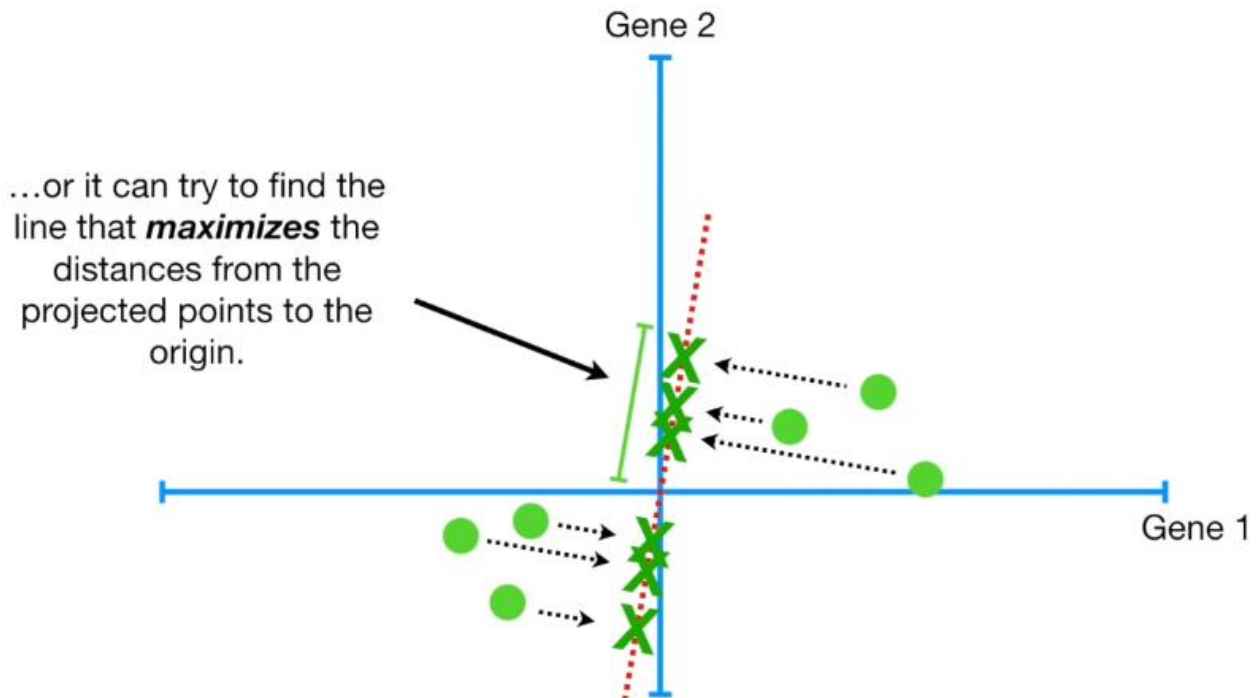


Method for finding the best fit line: - Projection of data onto the line

Optimization 1: Minimize sum of distances of the projected data from the line



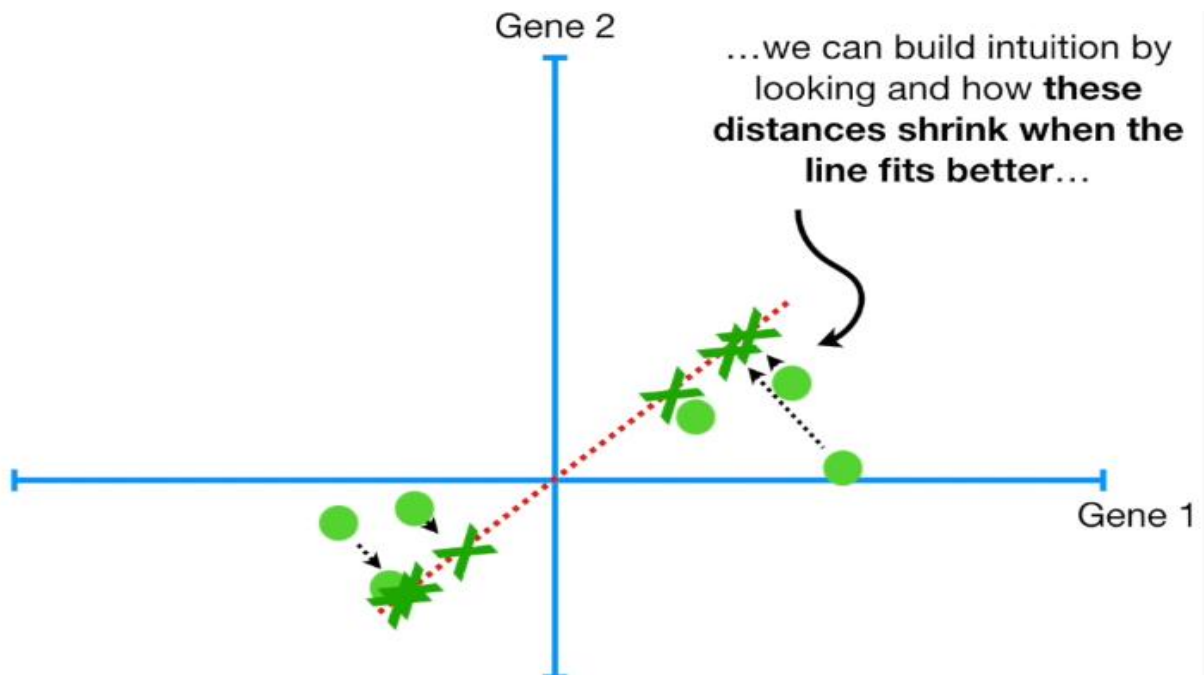
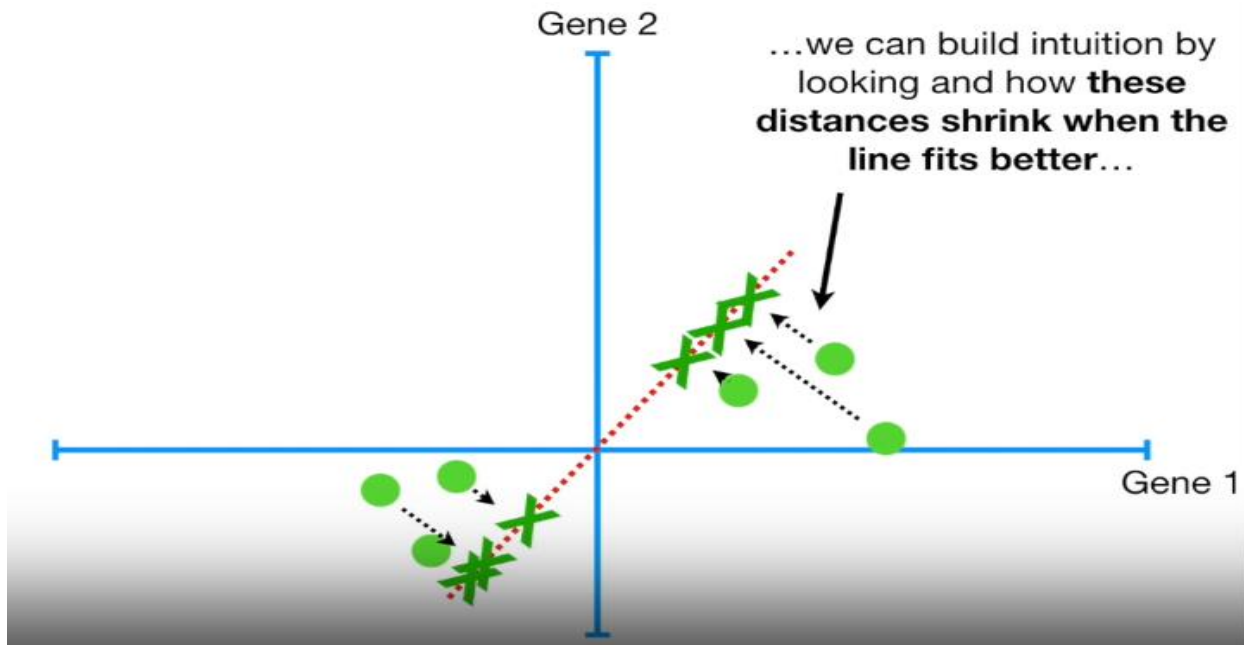
Optimization 2: Maximize the distance of projected data from the origin



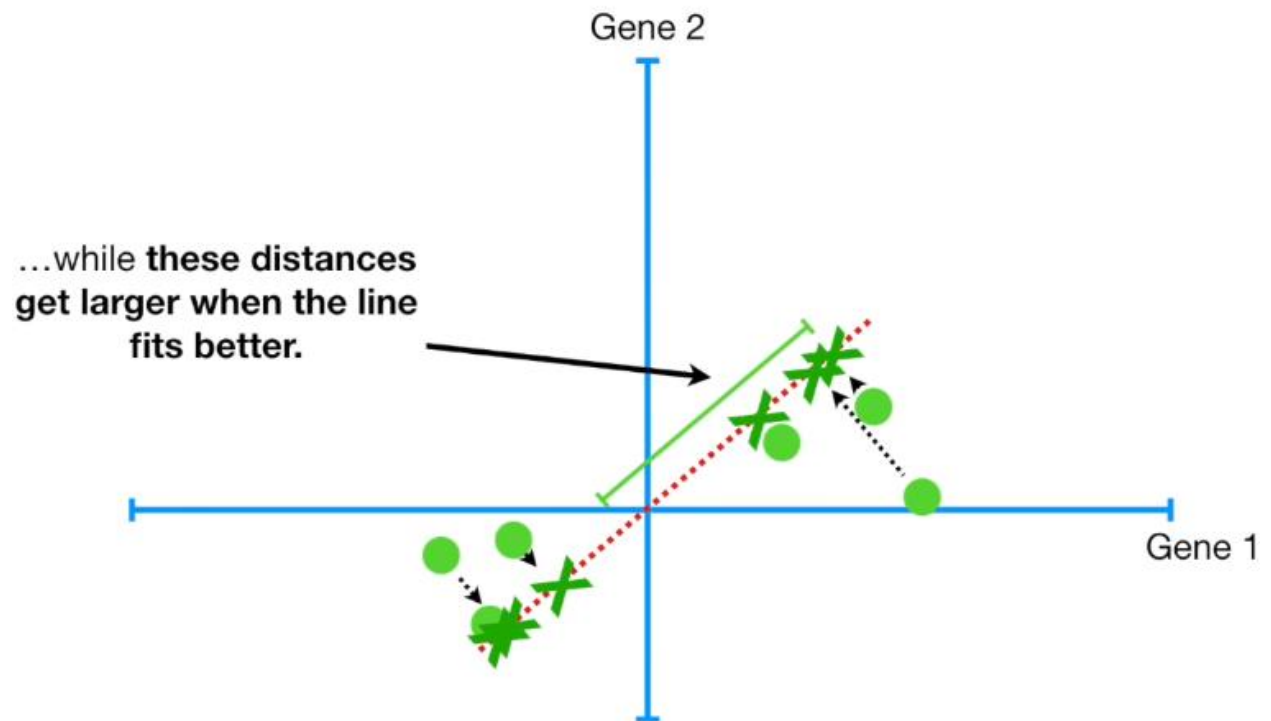
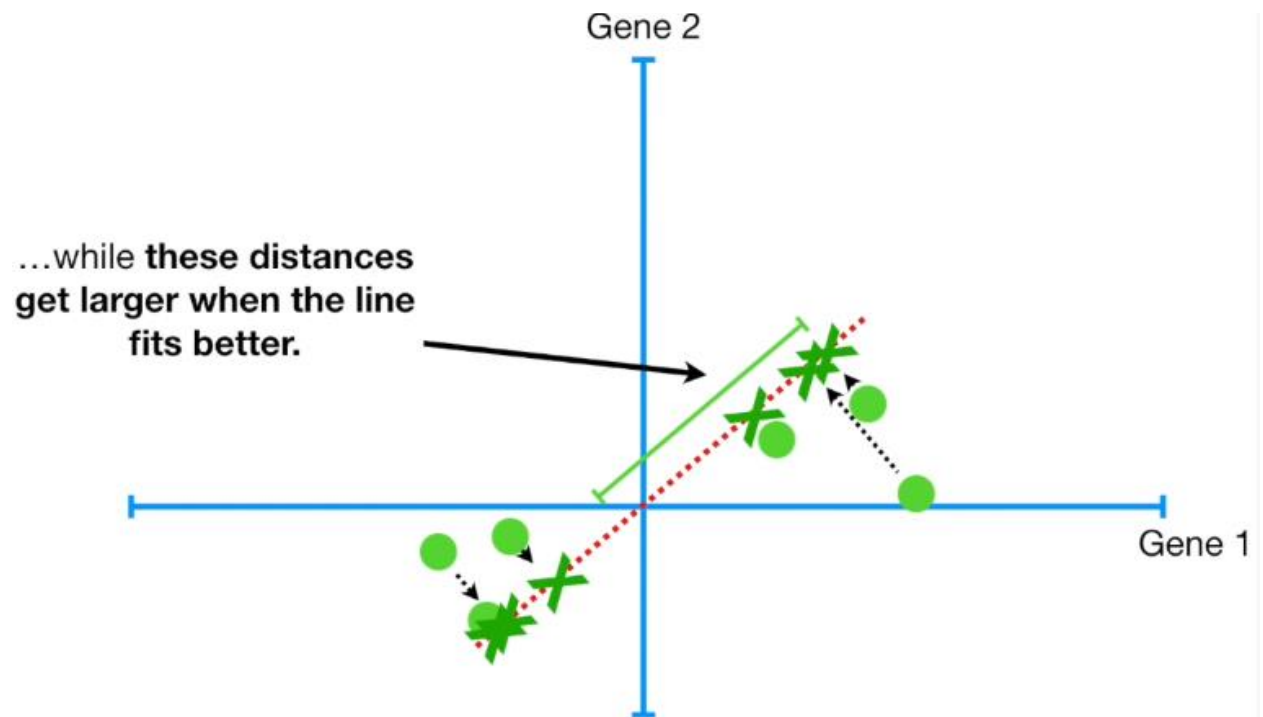
Both optimization 1 and 2 are equivalent, can be proved by below intuition

Visualize by rotation – observe the shrinking of the distances from the line

- Optimization 1



- Optimization 2

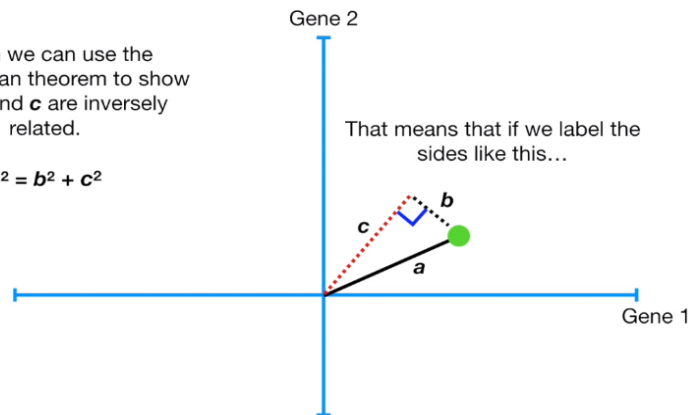


Geometric proof - optimization 1 and 2 are equivalent

...then we can use the Pythagorean theorem to show how **b** and **c** are inversely related.

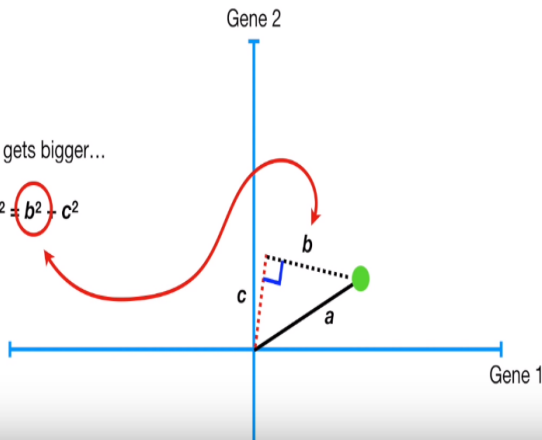
$$a^2 = b^2 + c^2$$

That means that if we label the sides like this...



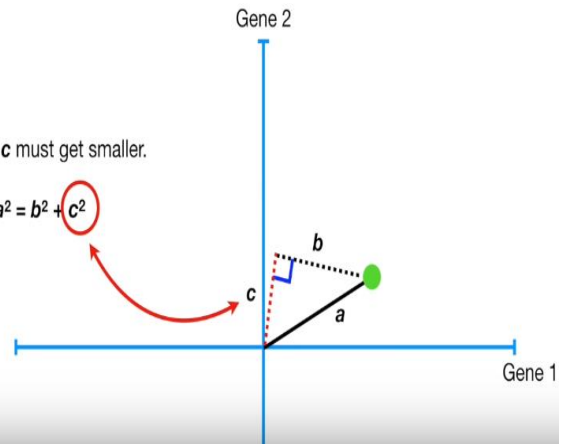
If **b** gets bigger...

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



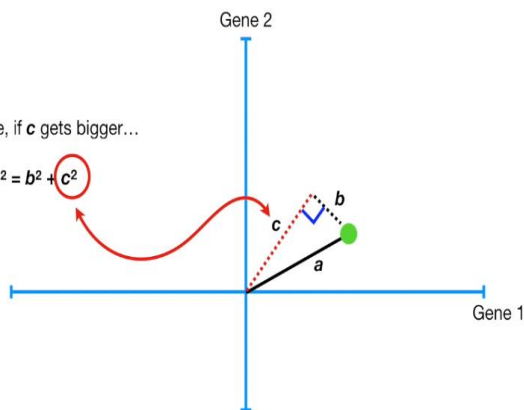
...then **c** must get smaller.

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



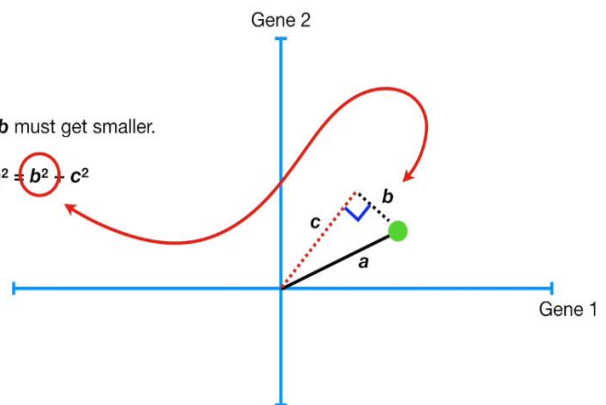
Likewise, if **c** gets bigger...

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



...then **b** must get smaller.

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

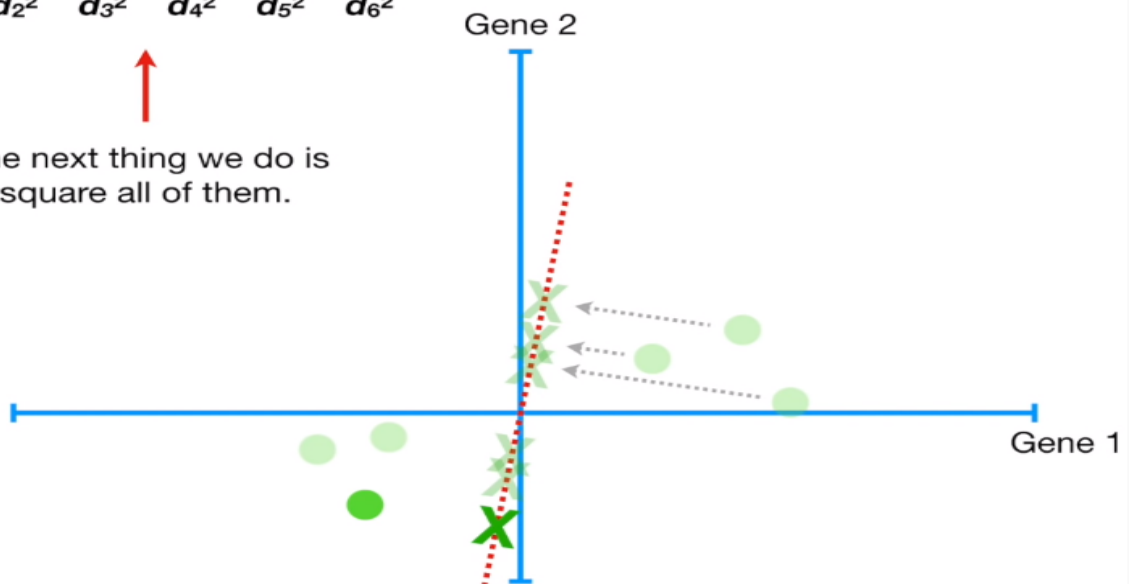


Optimization 1: Minimize sum of distances of the projected data from the line

Find: first principle component

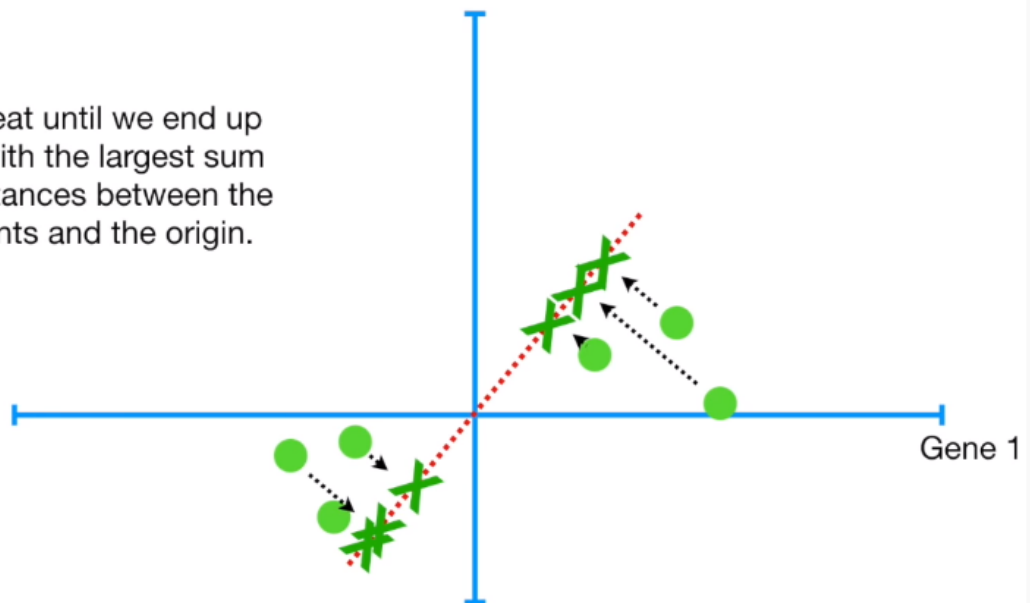
d_1^2 d_2^2 d_3^2 d_4^2 d_5^2 d_6^2

The next thing we do is
square all of them.



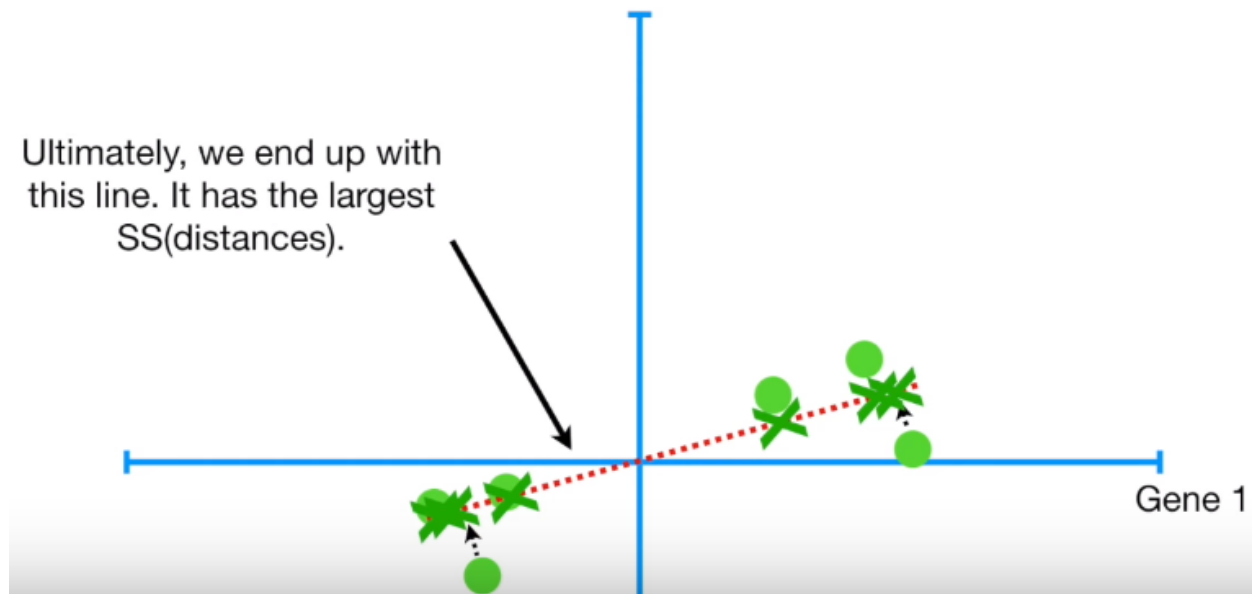
$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = \text{SS}(\text{distances})$$

...and we repeat until we end up
with the line with the largest sum
of squared distances between the
projected points and the origin.

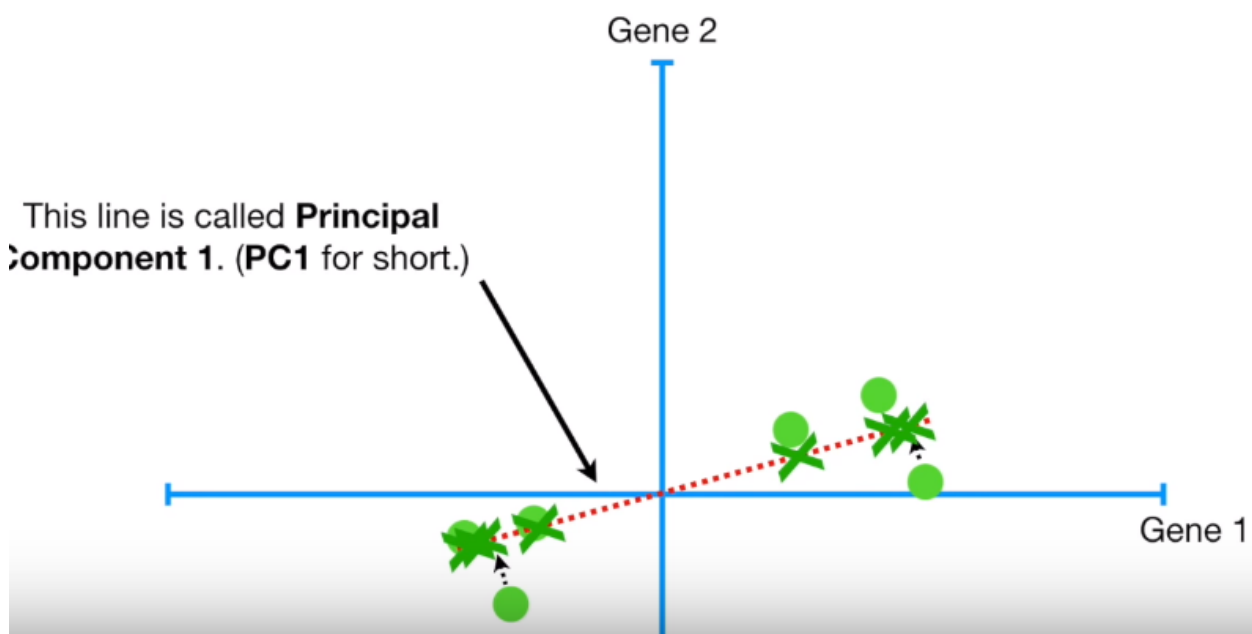


$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = \text{SS}(\text{distances})$$

Ultimately, we end up with this line. It has the largest SS(distances).



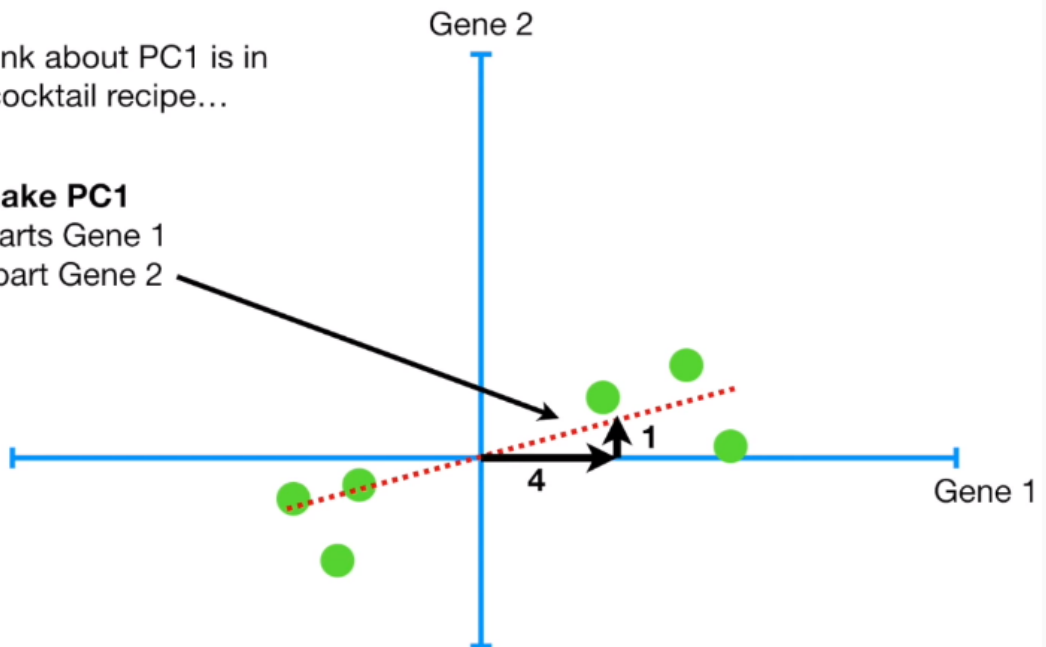
This line is called **Principal Component 1. (PC1 for short.)**



Geometric interpretation of projected data on principle component

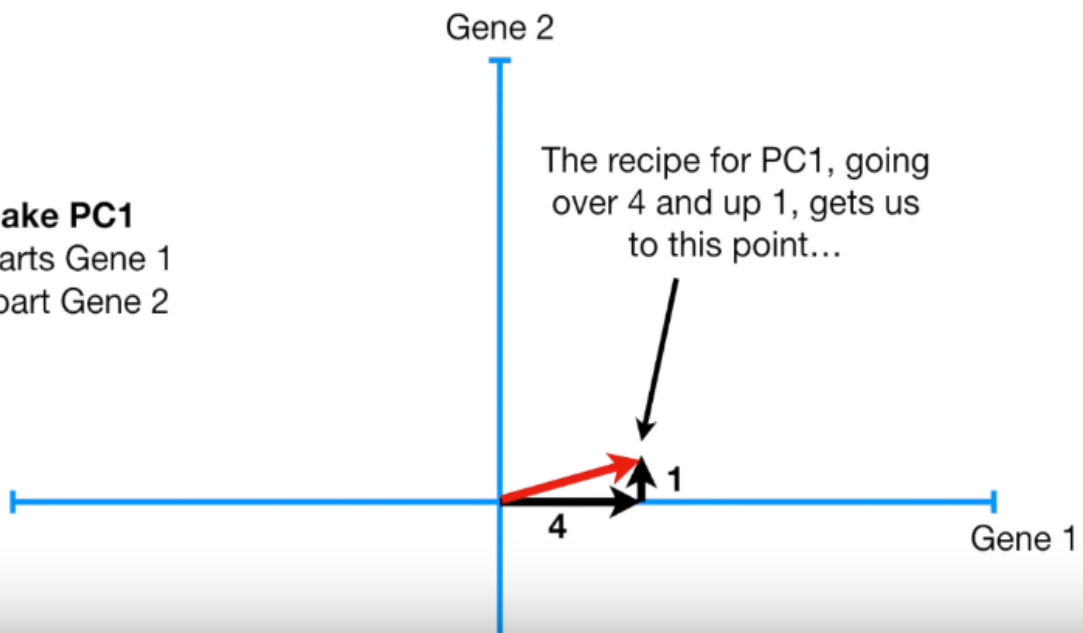
One way to think about PC1 is in terms of a cocktail recipe...

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

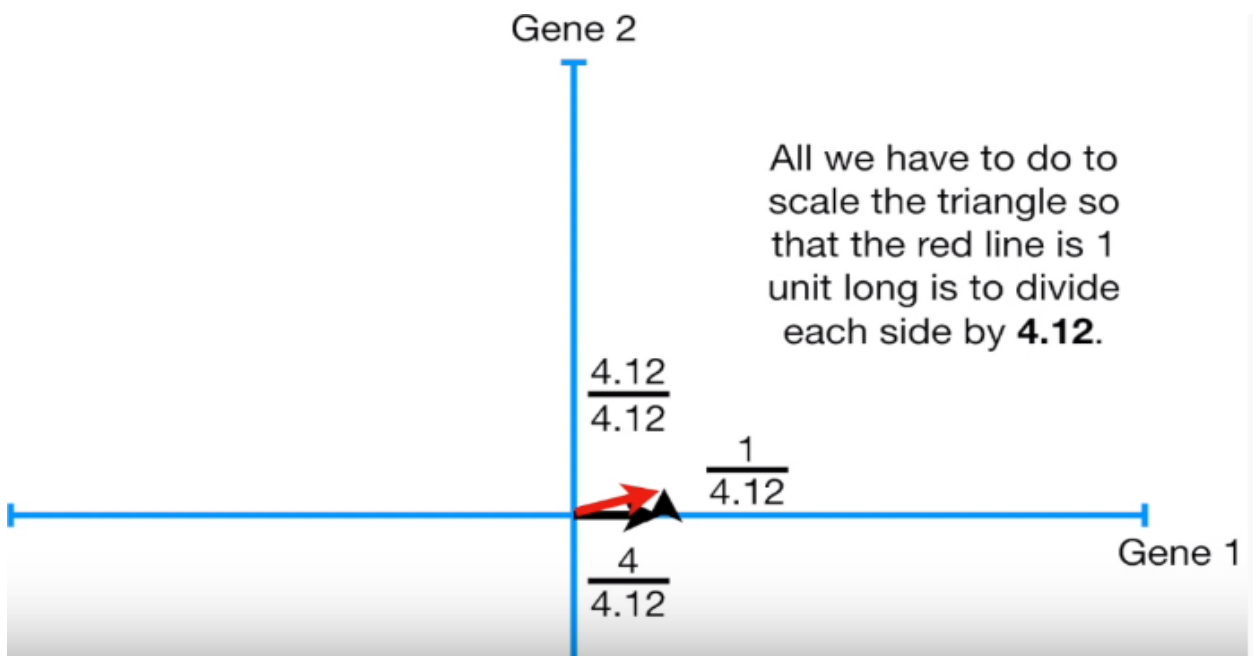
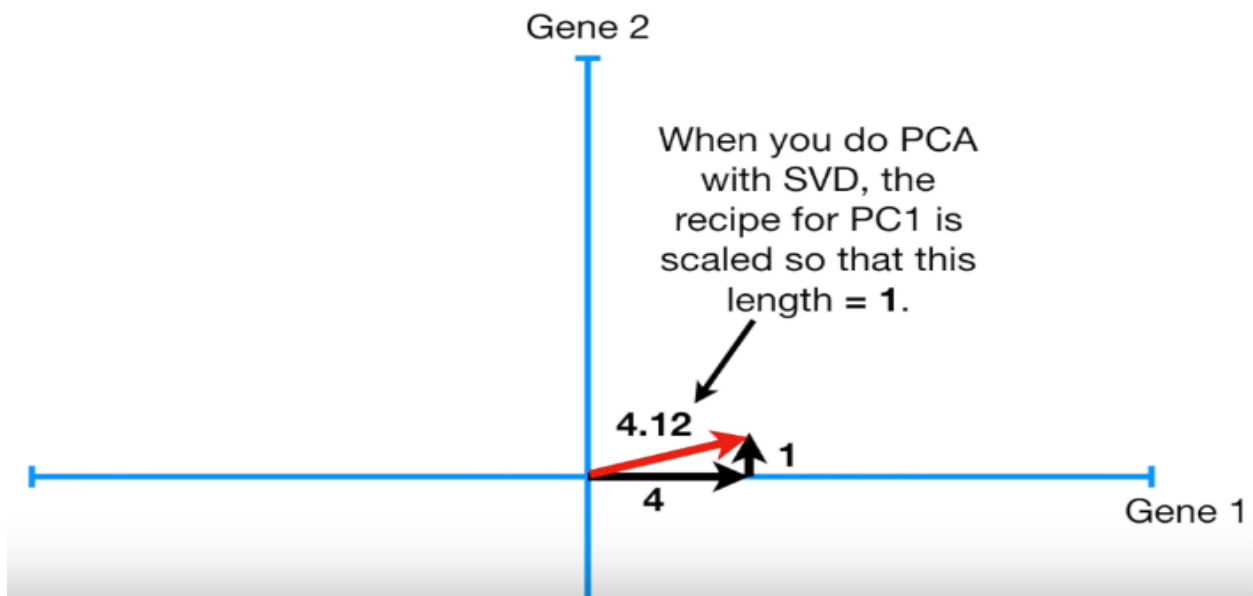


To make PC1
Mix **4** parts Gene 1
with **1** part Gene 2

The recipe for PC1, going
over 4 and up 1, gets us
to this point...



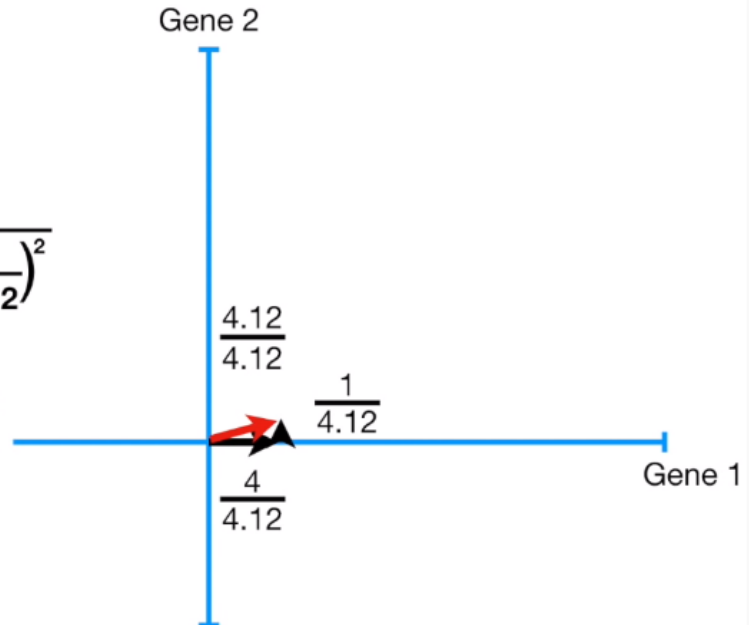
Projected line as a unit vector (eigen vector)



$$\frac{4.12}{4.12} = \frac{\sqrt{4^2 + 1^2}}{4.12} = \sqrt{\left(\frac{4^2 + 1^2}{4.12^2}\right)}$$

$$= \sqrt{\left(\frac{4}{4.12}\right)^2 + \left(\frac{1}{4.12}\right)^2}$$

For those of you keeping score, here's the math worked out that shows that all we need to do is divide all 3 sides by **4.12**.



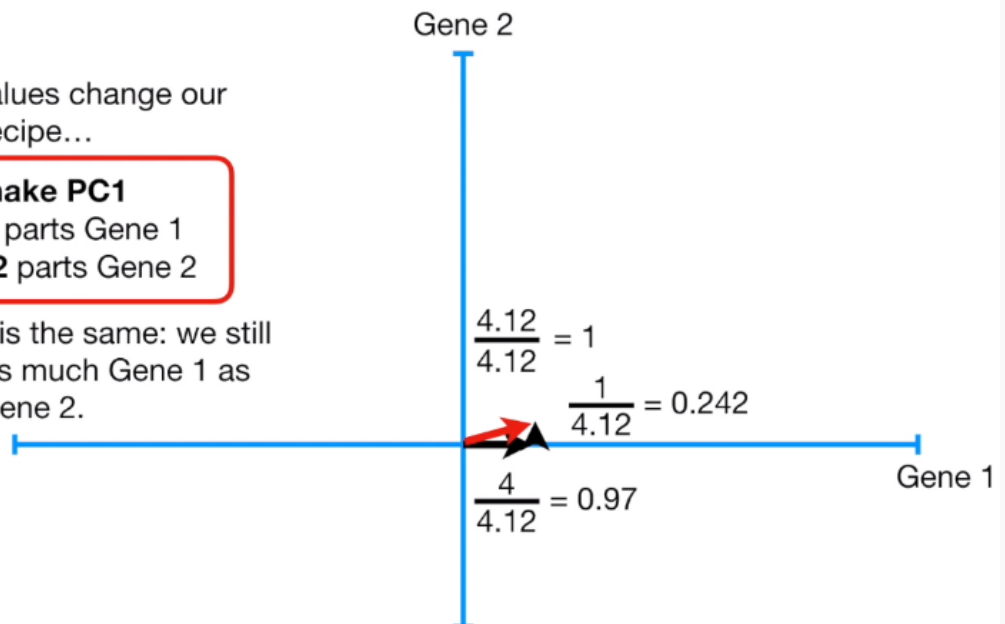
Percentage of variance explained for two samples in the direction of unit vector (line or eigen vector)

The new values change our recipe...

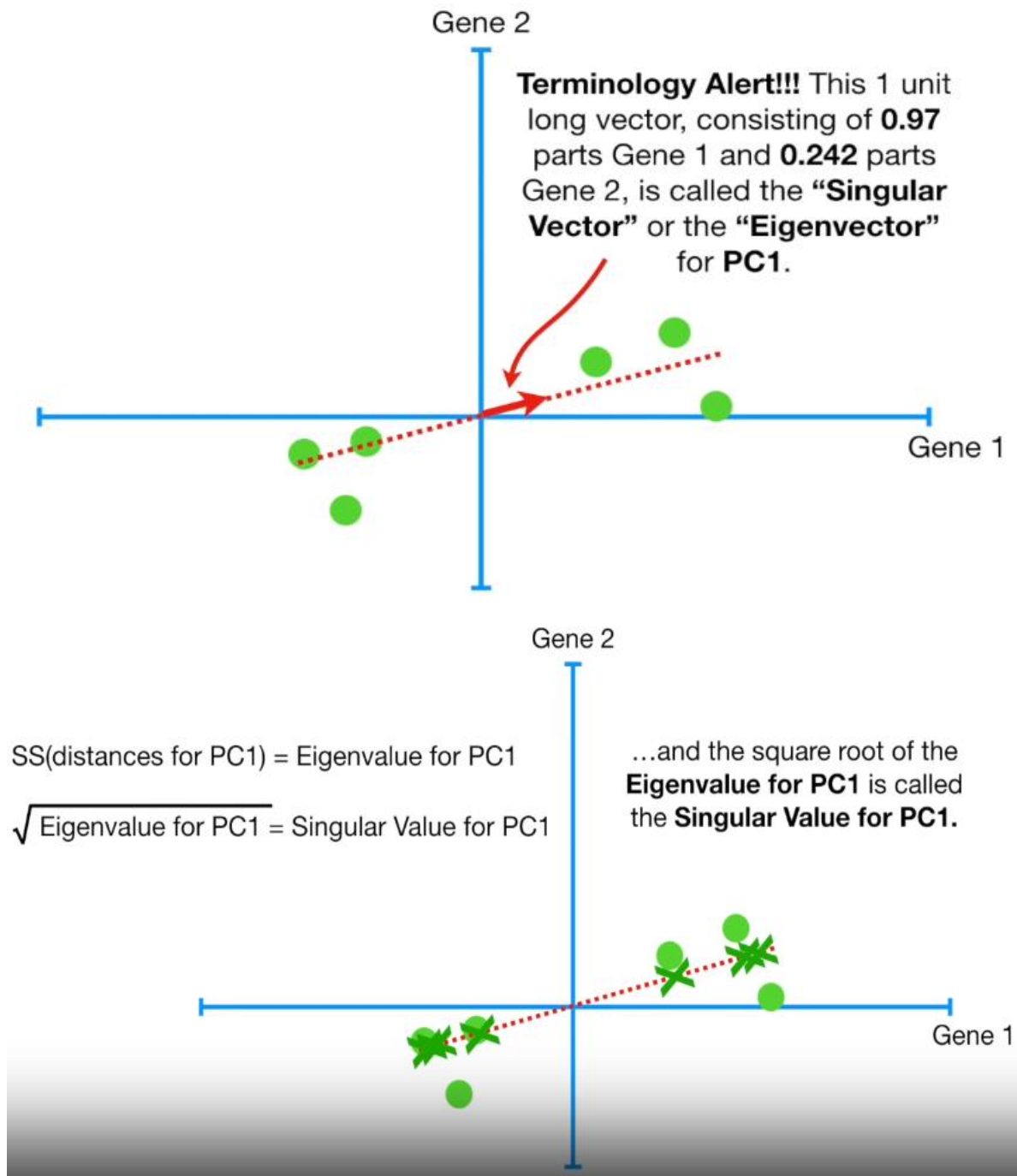
To make PC1

Mix **0.97** parts Gene 1
with **0.242** parts Gene 2

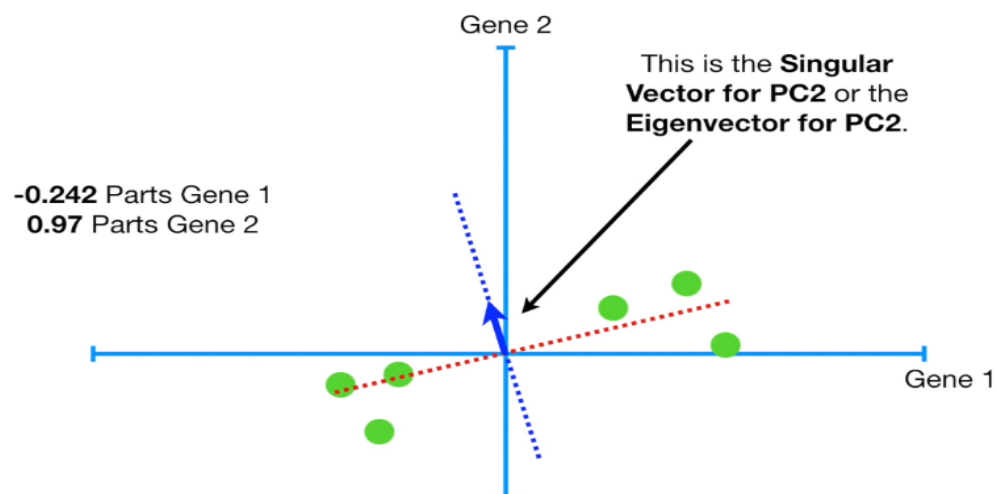
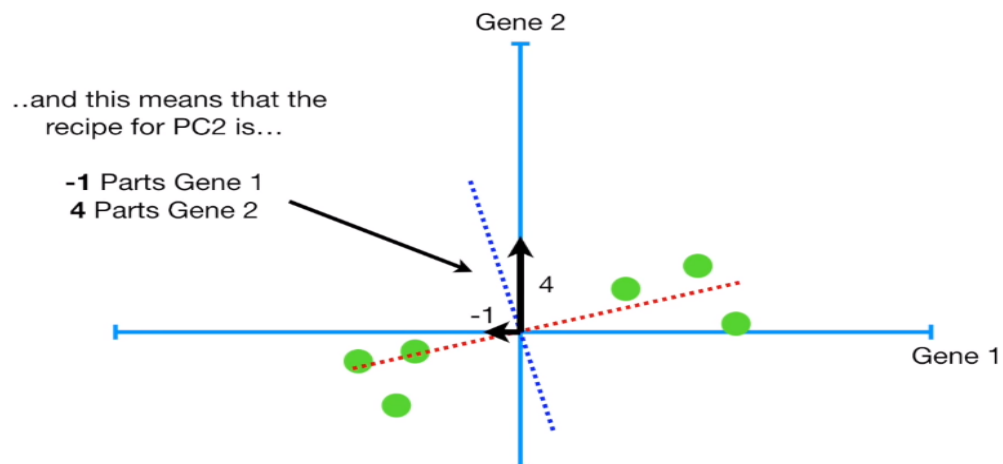
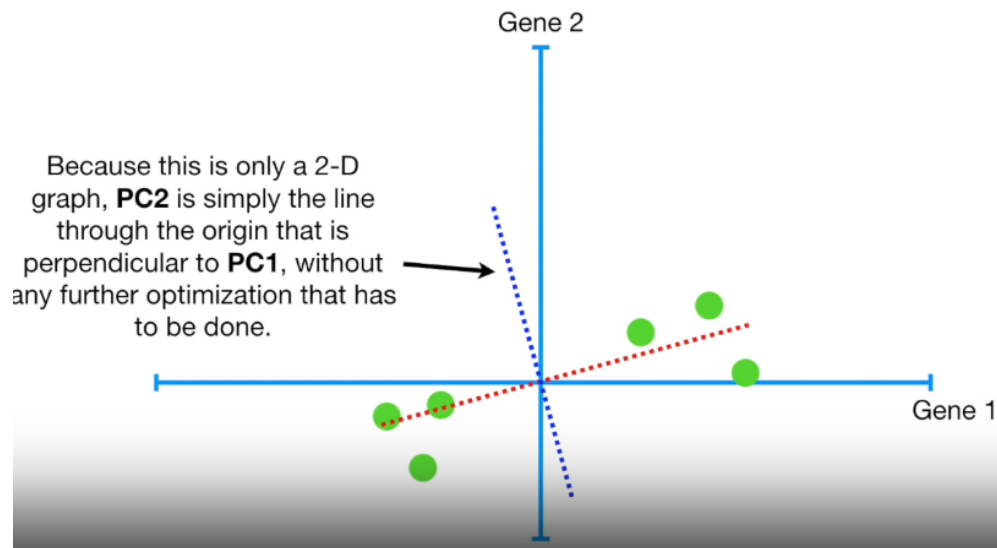
...but the ratio is the same: we still use 4 times as much Gene 1 as Gene 2.



Eigen value interpretation



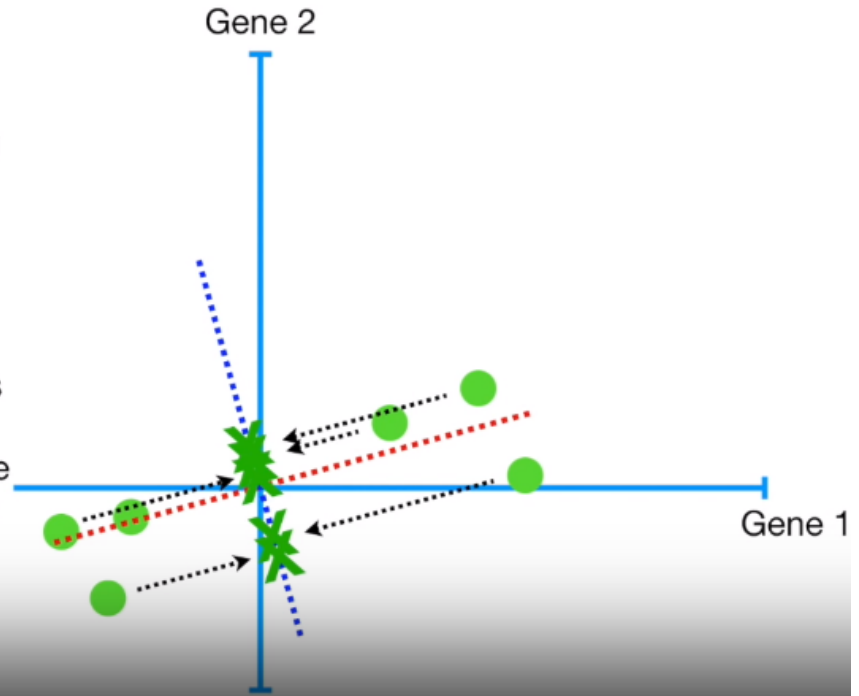
Deriving second principle components: -



These are the **Loading Scores for PC2.**

-0.242 Parts Gene 1
0.97 Parts Gene 2

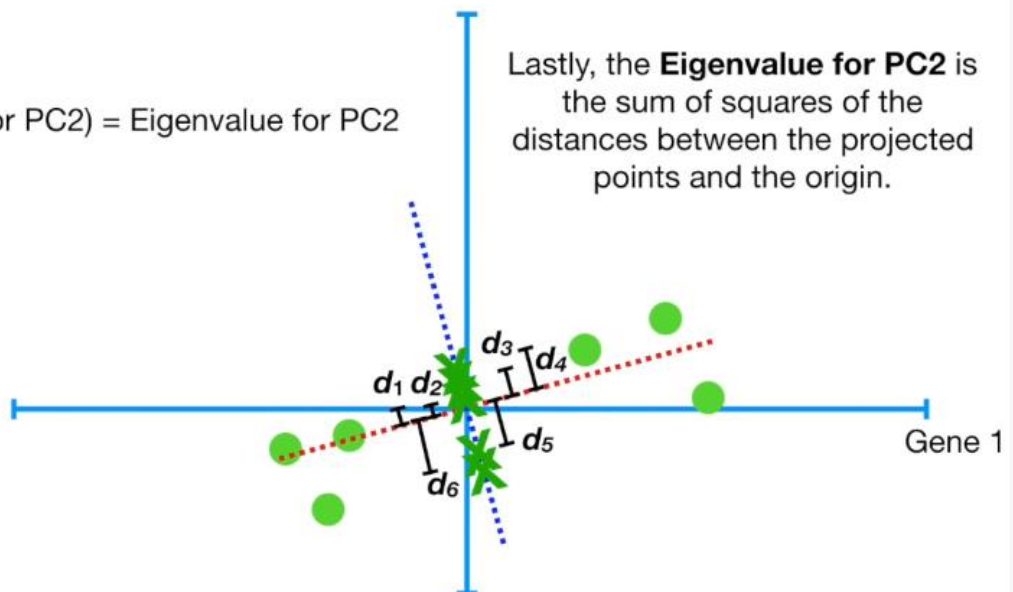
They tell us that, in terms of how the values are projected onto PC2, Gene 2 is 4 times as important as Gene 1.



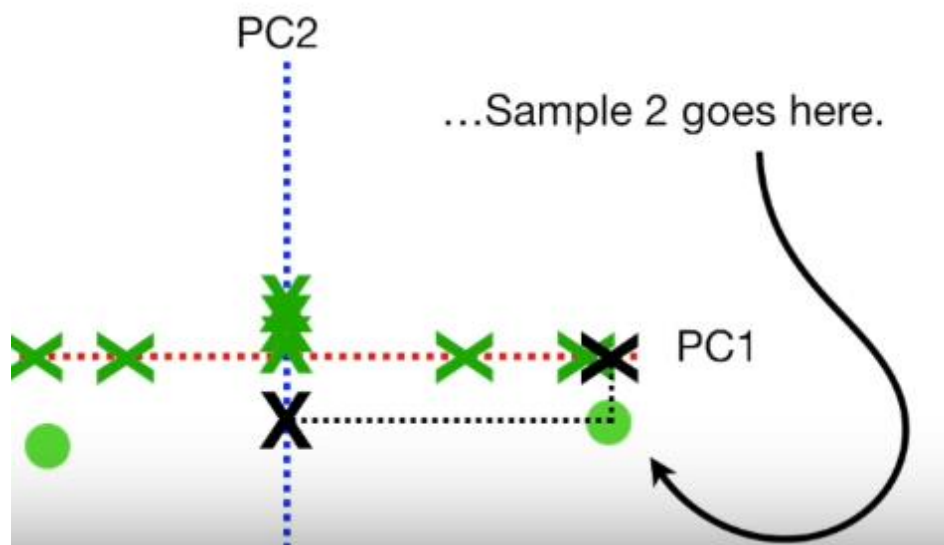
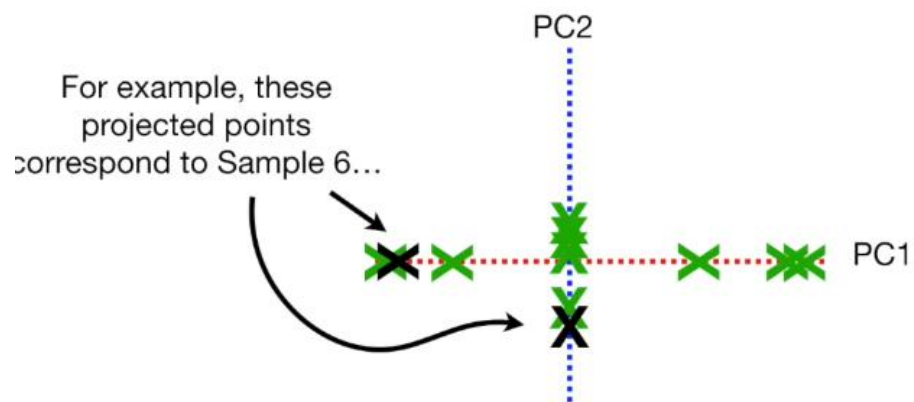
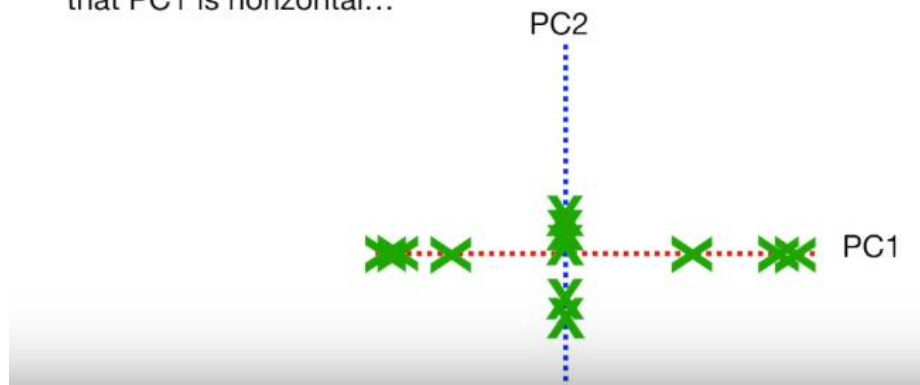
$$d_1^2 + d_2^2 + d_3^2 + d_4^2 + d_5^2 + d_6^2 = \text{sum of squared distances} = \text{SS}(\text{distances})$$

SS(distances for PC2) = Eigenvalue for PC2

Lastly, the **Eigenvalue for PC2** is the sum of squares of the distances between the projected points and the origin.



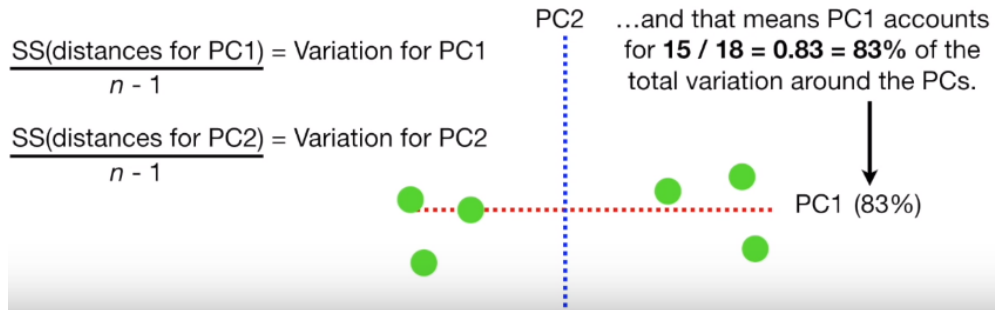
We simply rotate everything so
that PC1 is horizontal...



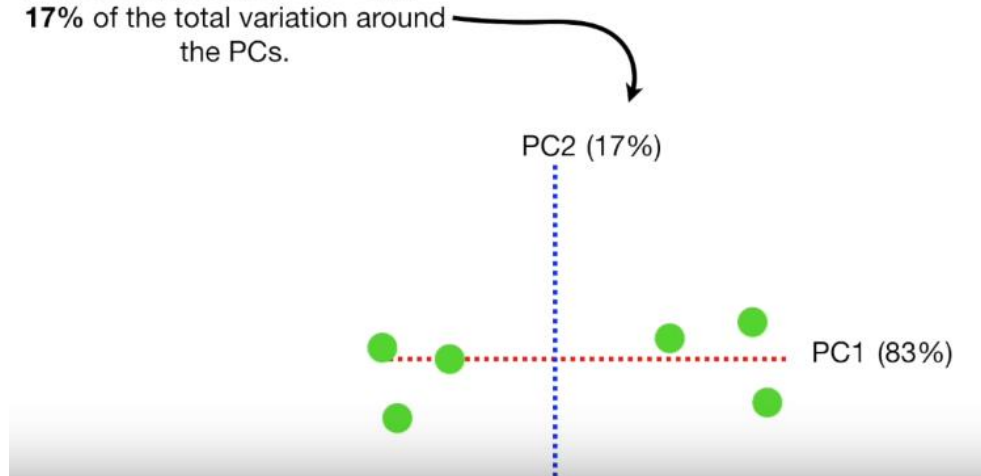
Variance explained in the direction of projected line (eigen vector)

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

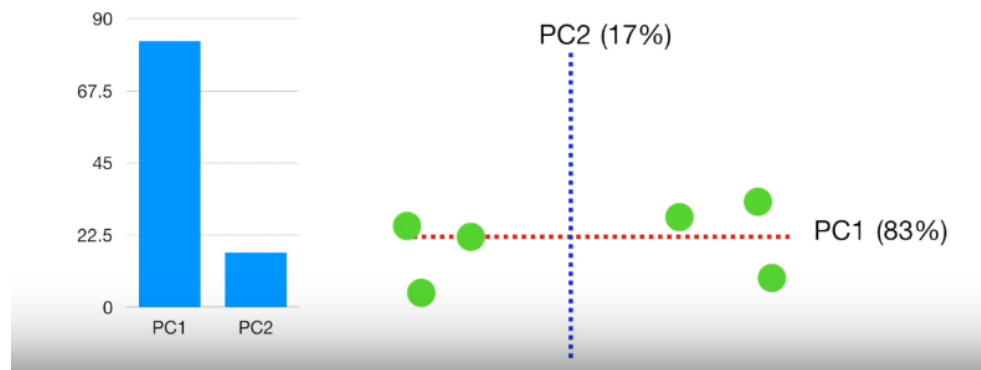
That means that the total variation around both PCs is **15 + 3 = 18...**



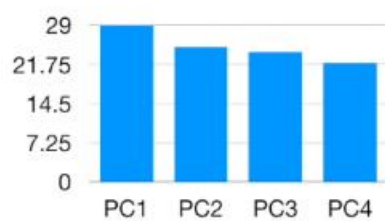
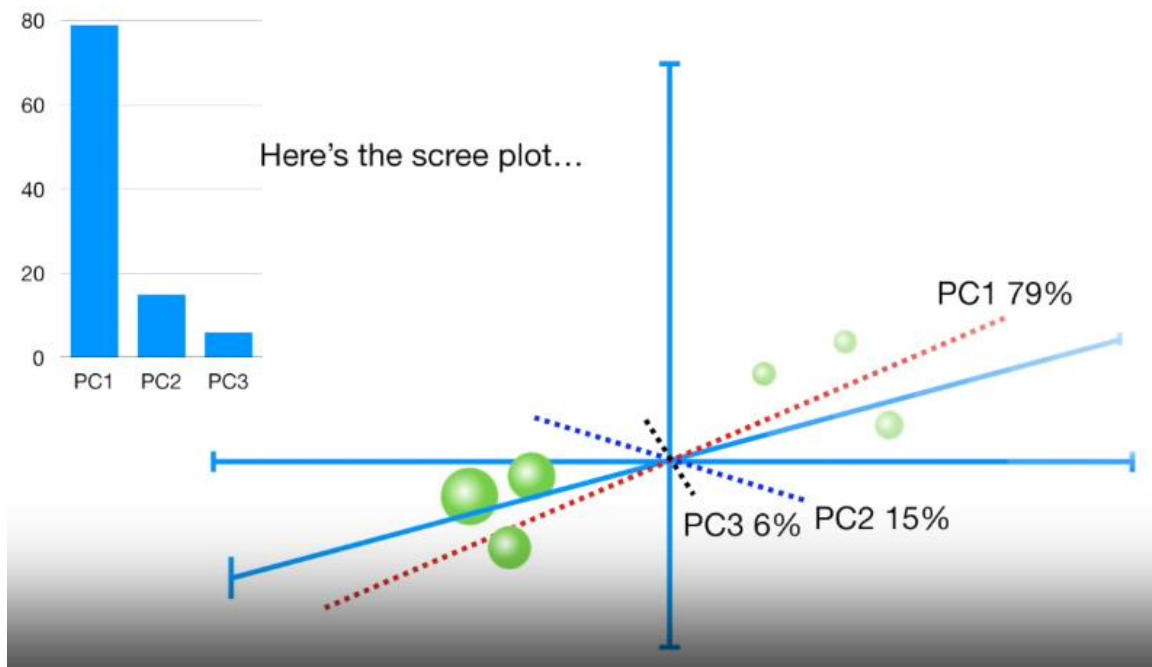
PC2 accounts for $3 / 18 = 0.17 = 17\%$ of the total variation around the PCs.



TERMINOLOGY ALERT!!!! A **Scree Plot** is a graphical representation of the percentages of variation that each PC accounts for.



Principle components and variance explained for more than two features: -



NOTE: If the scree plot looked like this, where PC3 and PC4 account for a substantial amount of variation, then just using the first 2 PCs would not create a very accurate representation of the data.

