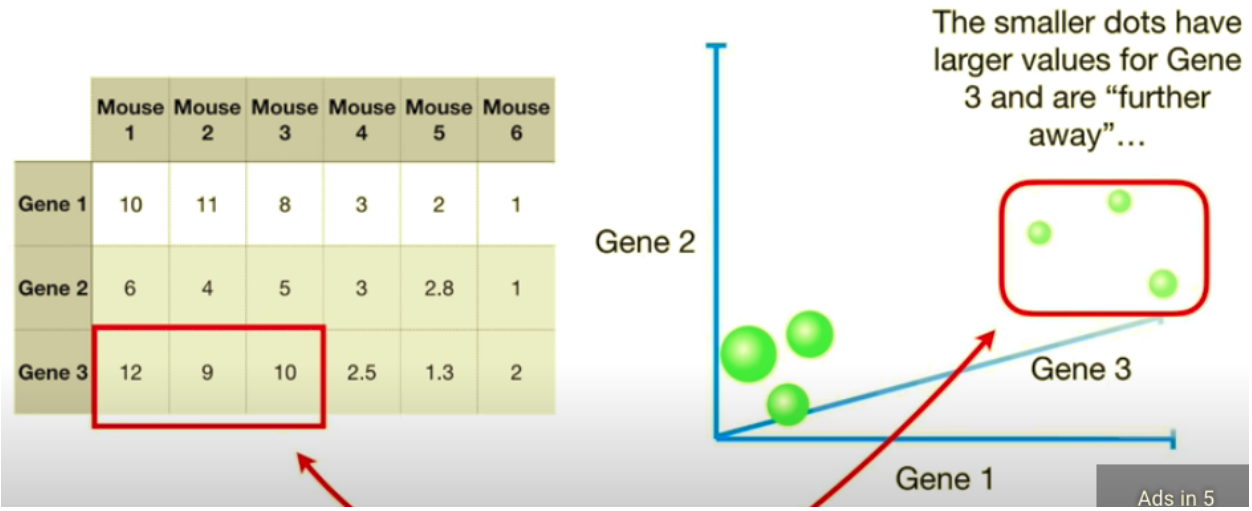
PCA Step by Step

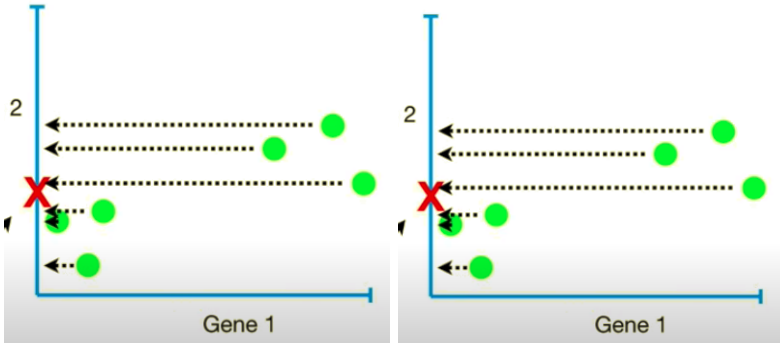
- PCA becomes necessary to bring down the dimensionality of a dataset. Below are the max points that the human eye can visually see - a 3 dimensional set.
- ** Disclaimer - It is important to normalize data before performing PCA. The new axis calculated are based on the standard deviation of your variables. So a variable with a high standard deviation will have a higher weight for the calculation of axis than a variable with a low standard deviation.

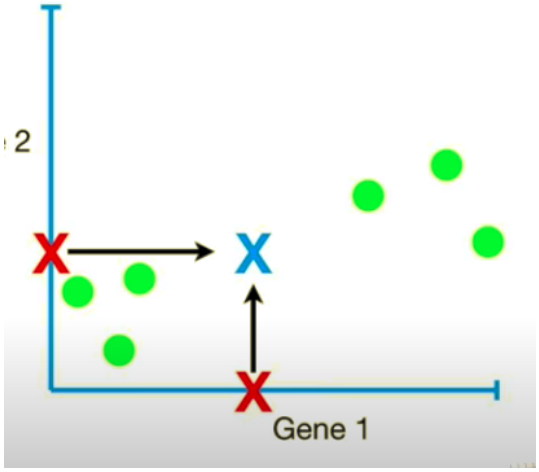


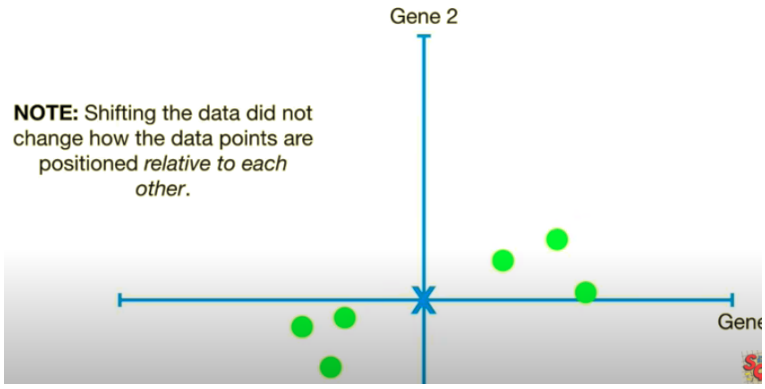
- If we measured 4 variables in the plot, we can take 4 dimensions of data and make a 2d plot. The plot will show us similarities, and which variable is the most valuable for clustering the data.
- PCA can also tell us the accuracy of a 2d graph.

PCA ANALYSIS - 2 VARIABLES

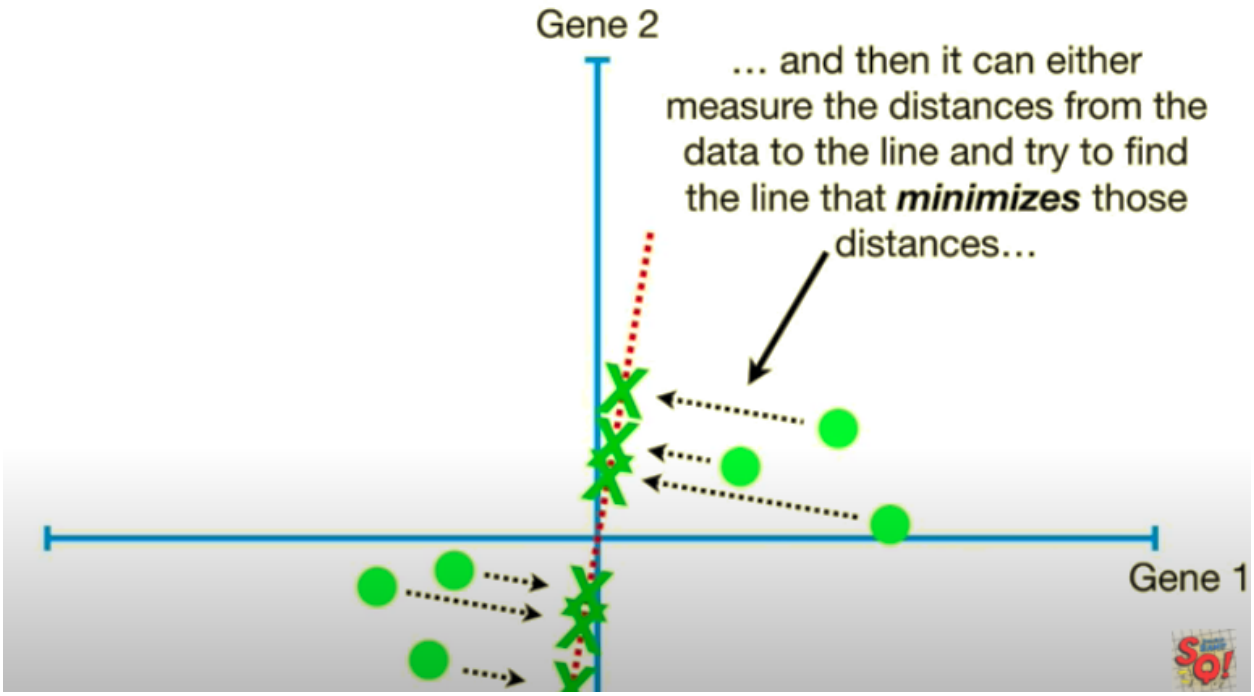
- First thing that needs to be done is to calculate the average measurement for gene 1 and gene 2
- Once we have the average we shift the data so the center is on top of the origin. From this point the original data is no longer needed.



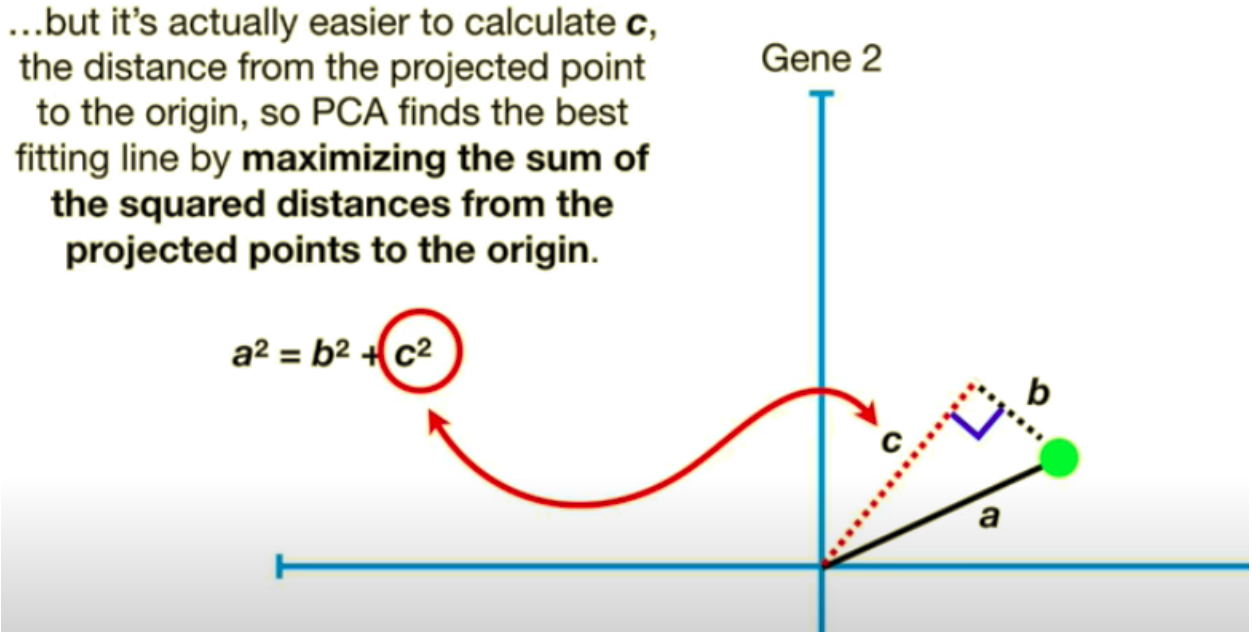




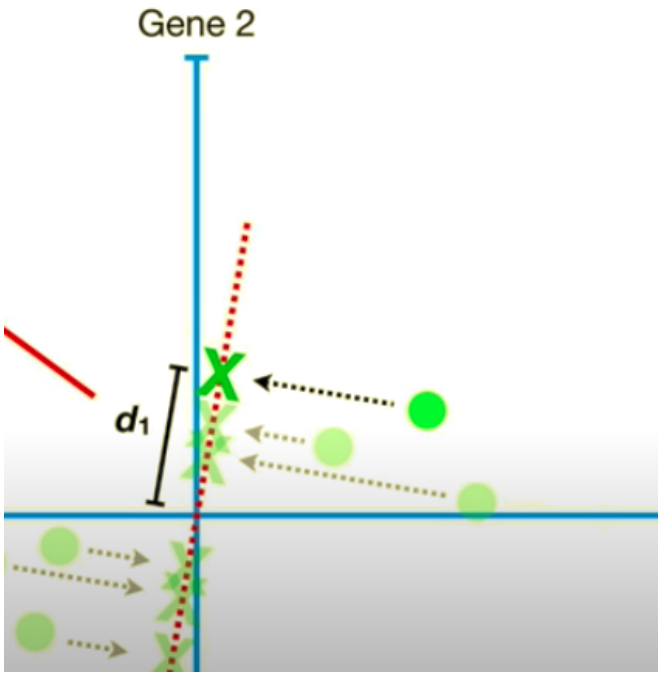
- Once the data is centered on the origin we fit a line to it. The line must go through the origin. We rotate the line so the data fits as best as possible. PCA decides if a line is a good fit or not by measuring the distances from the data to the line and trying to find that minimizes those distances.
- Or it can try and find the line that maximizes the distances from the projected points to the origin.

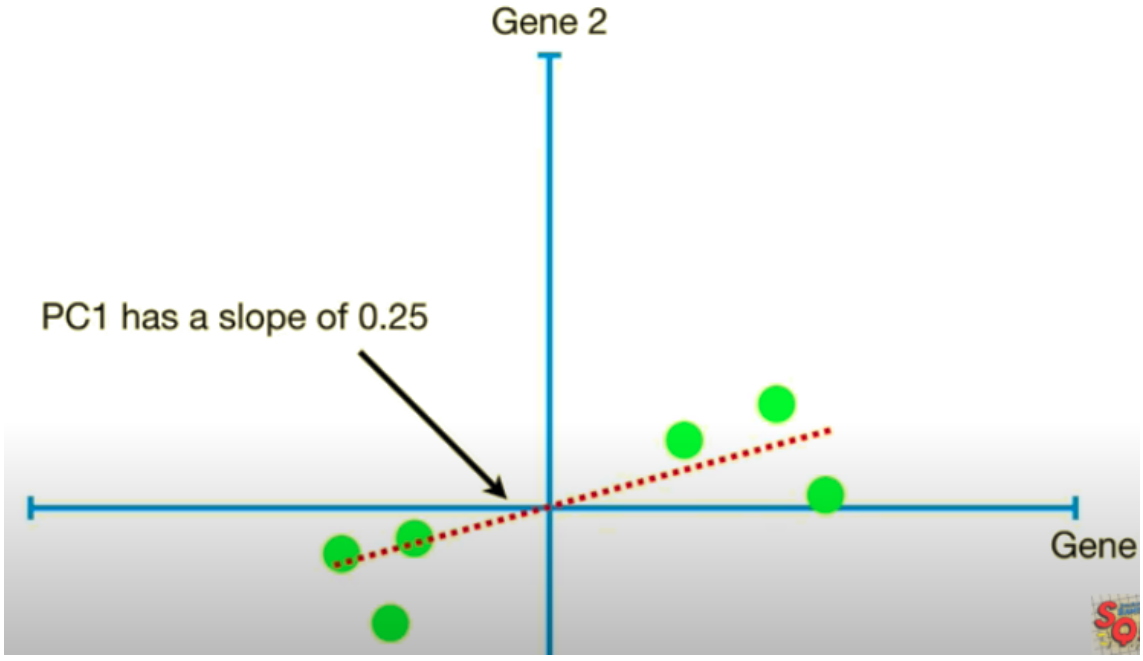


- Considering one data point. The position is fixed from the point to the origin. The distance does not change when the dotted red line rotates.

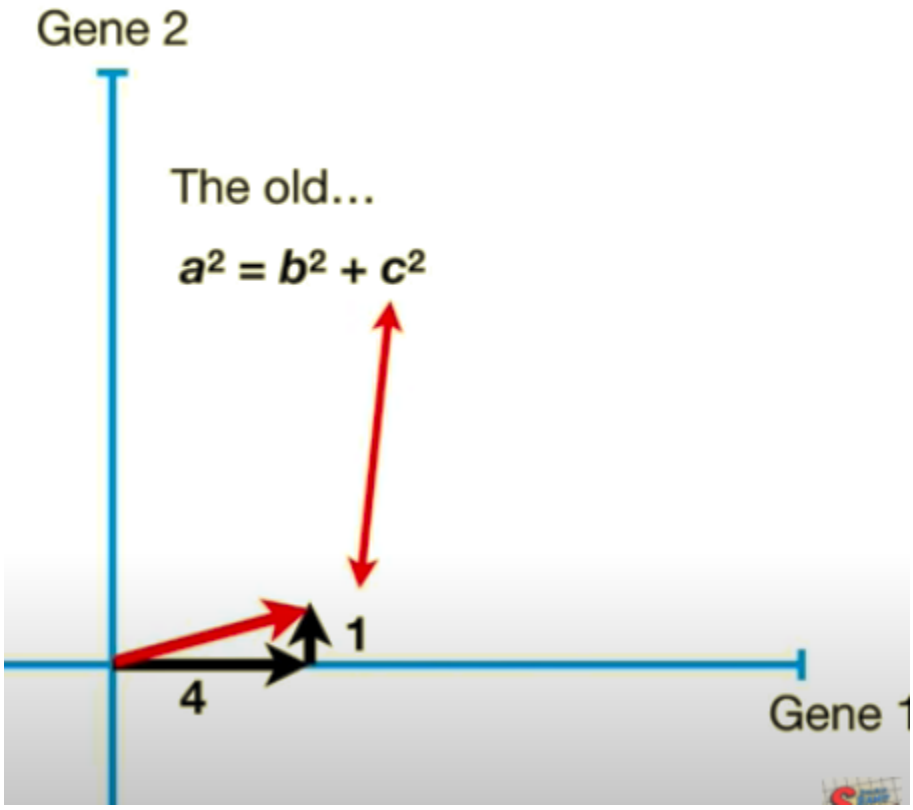


- PCA finds the best fitting line by maximizing the sum of the squared differences from the projected points to the origin. A visualization can be seen below. After the differences for all points are found the points are squared and summed. (Squared to offset negative values).



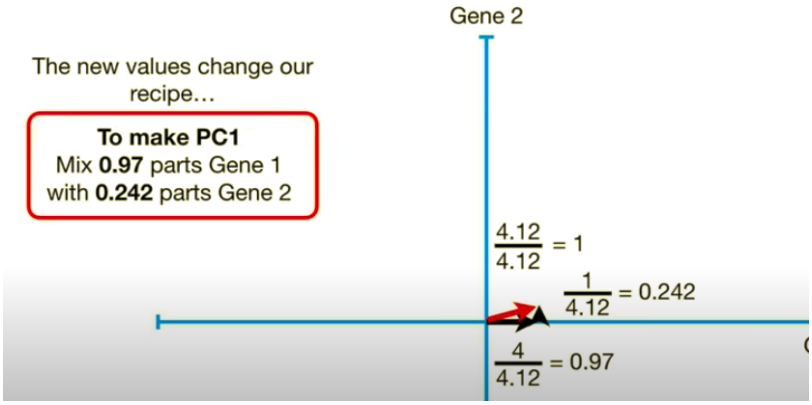


- The best fitting line for PC1 is shown above. The line has a slope of 0.25
- In other words for every 4 units moved on the x axis we move up 1 unit on the y axis. That means the data is mostly spread out upon the x axis, and less on the y.
- Given this information of gene one on the x axis - gene one is more important when it comes to describing how the data is spread out.
- PCA 1 is a linear combination of variables.

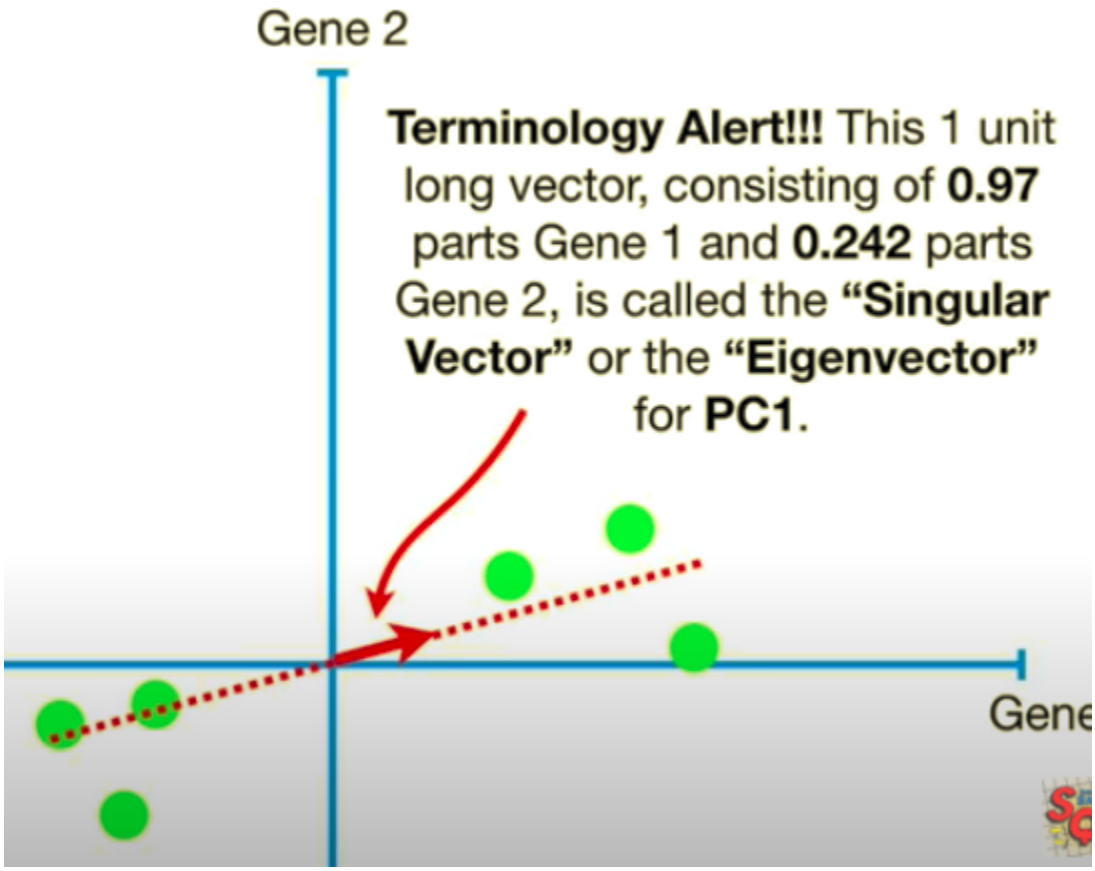


- To find the length of the red line we have the pythagorean theorem.

  4^2 + 1^2 = sqrt17 or 4.12
- The length of the red line is 4.12, but when you do PCA the recipe is scaled so that the red line is 1. In order to rework we divide all numbers by 4.12 scaled values below…

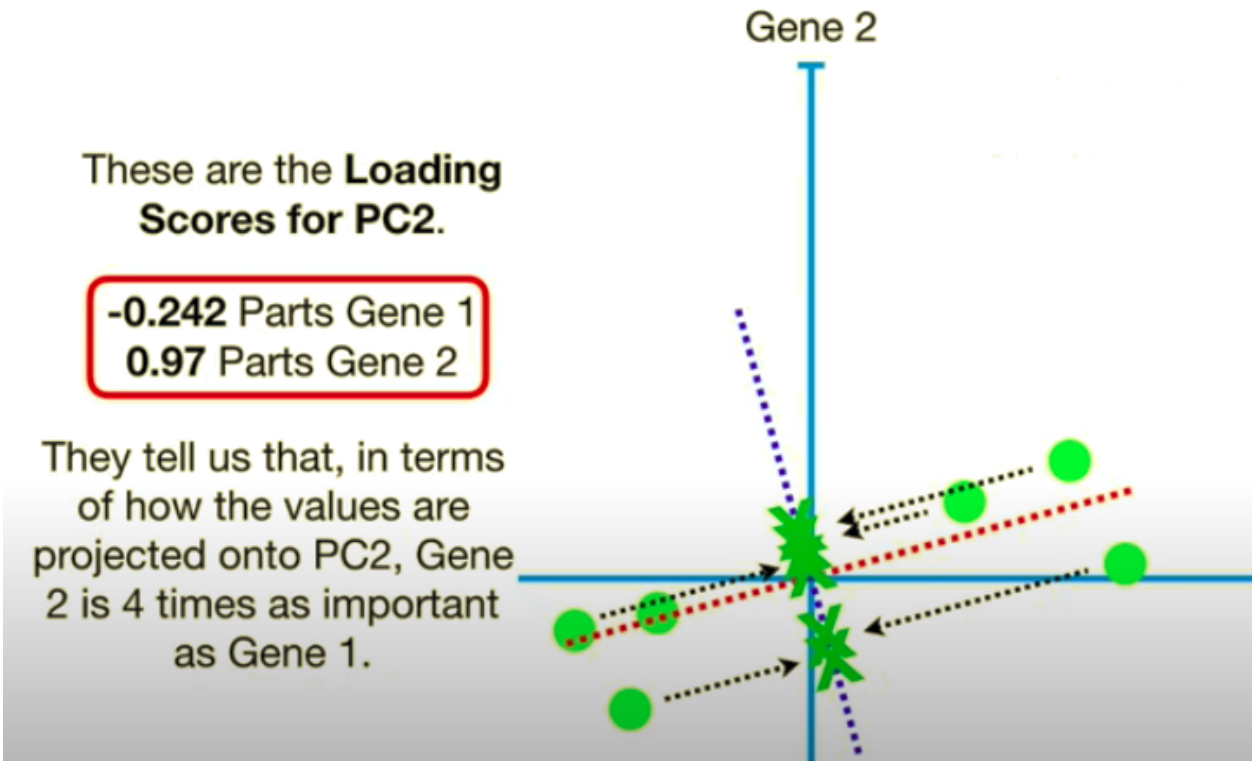


- The new values change the recipe but the ratio is the same, still 4 times as much.

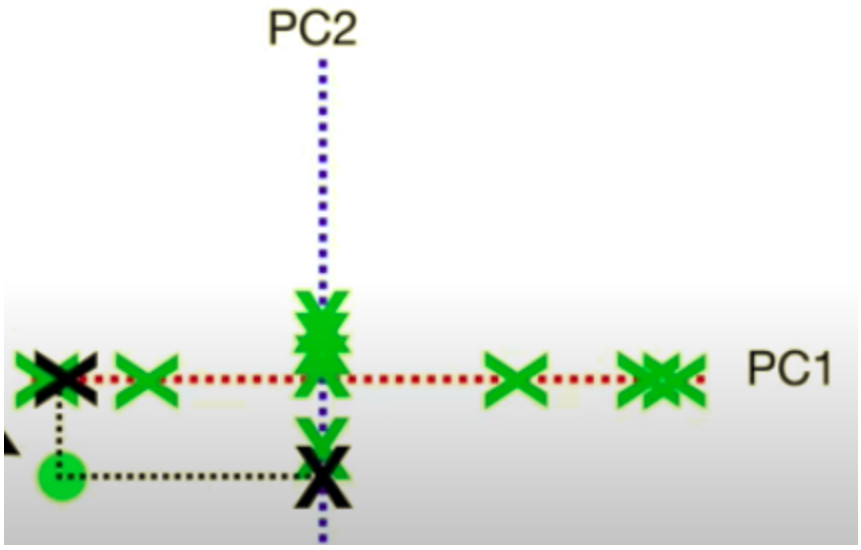


- The proportions of each gene are called loading scores - .97 parts gene 1, and .242 parts gene 2.

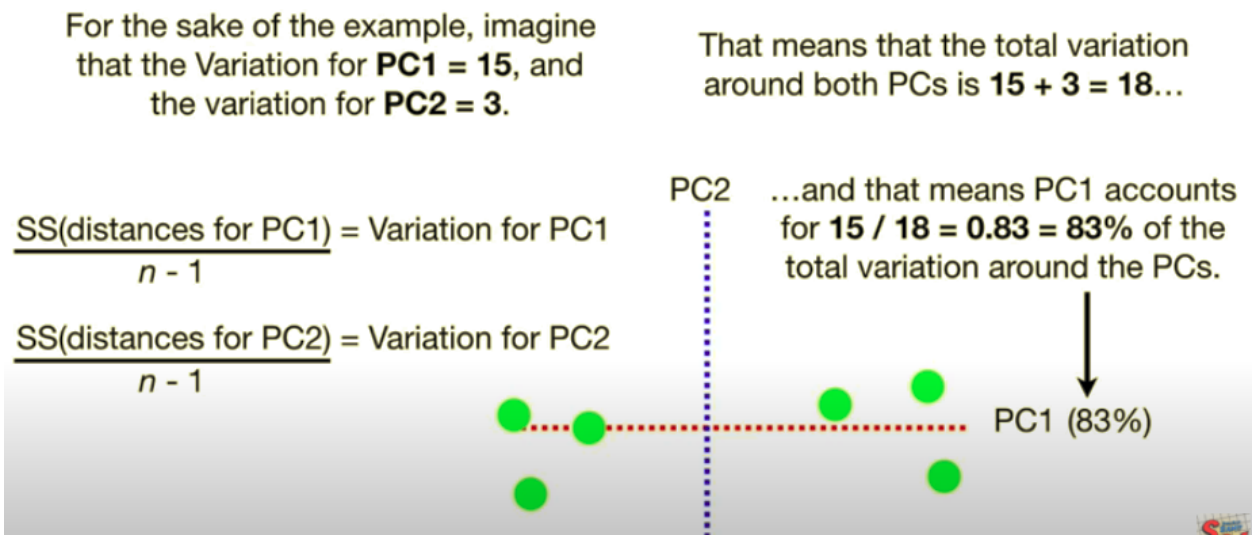
- Also PCA calls the sum of squared differences for the best fit line the eigenvalue. The square root of the eigenvalue for pc1 is called the singular value.
- Because it is only a 2-d graph, PC2 is simply the line through the origin that is perpendicular to PC1, without any further optimization that has to be done.
- This means the recipe is -1 parts gene 1, to 4 parts gene 2
- The loading scores = -0.242 parts gene 1, and 0.97 gene 2, this provides us with the eigenvector for PC2.

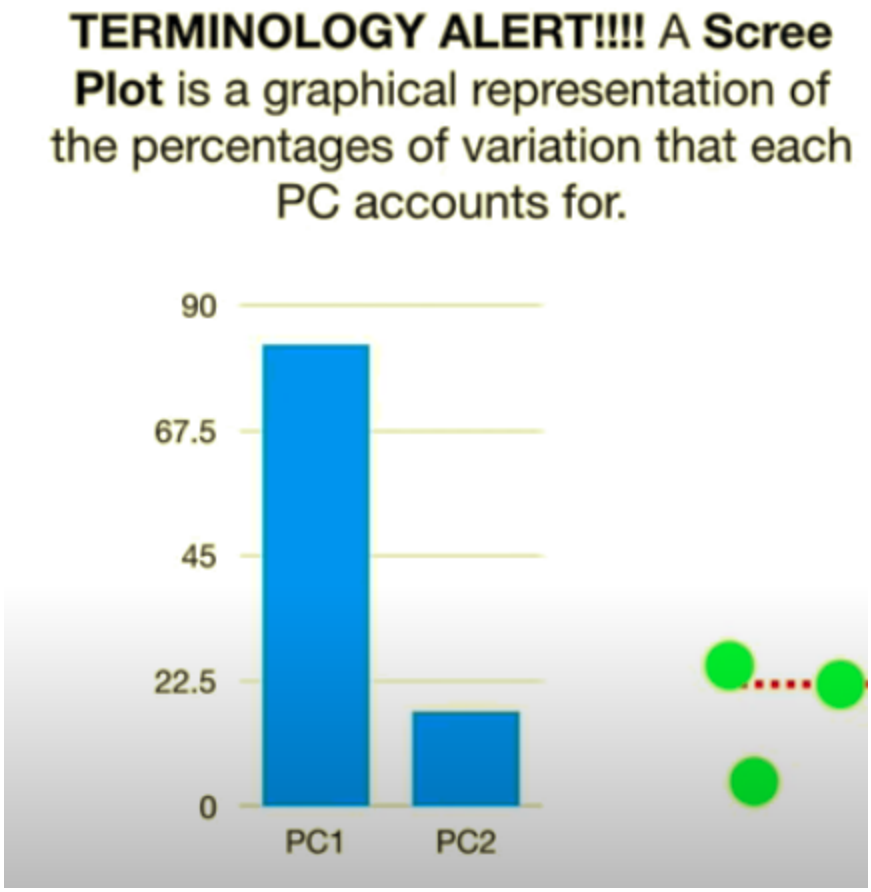


- The eigenvalue for PC2 is the sum of squares between the projected points and the origin.
- The plot is then pivoted to where the PCA origin is straight up and down then the points are placed back accordingly.



VARIATION OF THE PRINCIPAL COMPONENTS

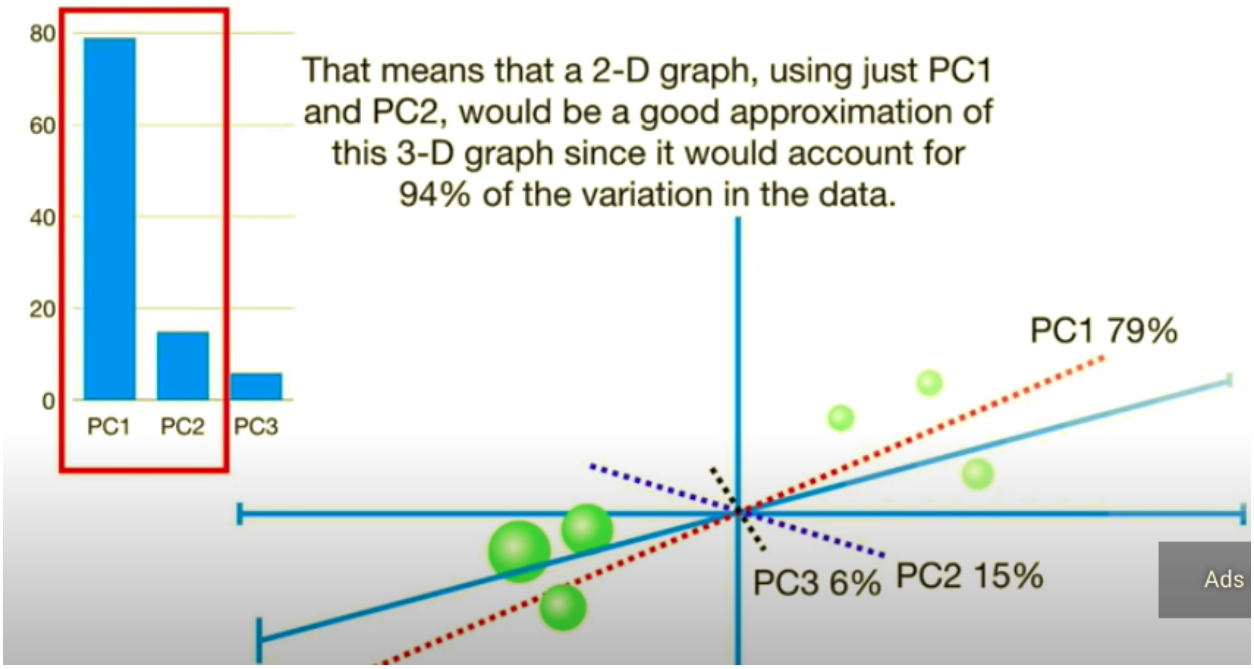




- PC1 accounts for 83 percent of the total variation around the PCs

  PC2 accounts for 17 percent of the total variation around the PCs

PCA 3 DIMENSIONAL DATA



- For 3 dimensional data the process continues. Find the best fitting line that goes through the origin, and is perpendicular to the other PCAs. The process would continue to go on with each PCA added.
- Given PCA1 and PCA2 account for 94% of the variation that means a 2D graph would be a good representation of the 3d graph.