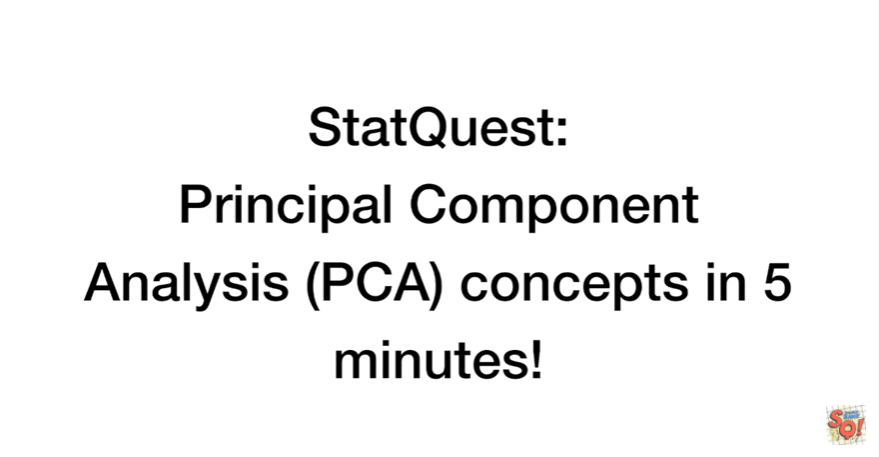
<https://www.youtube.com/watch?v=HMOI_lkzW08&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=25>

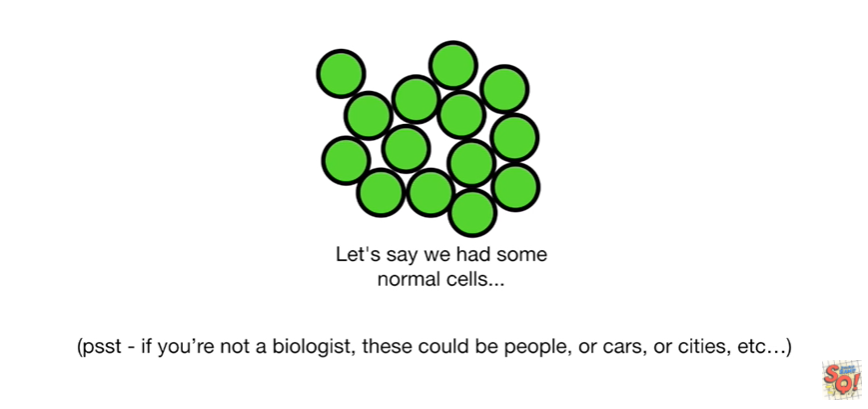


Today we're gonna be talking about the main ideas behind principle component analysis and we're going to cover those concepts in five minutes !

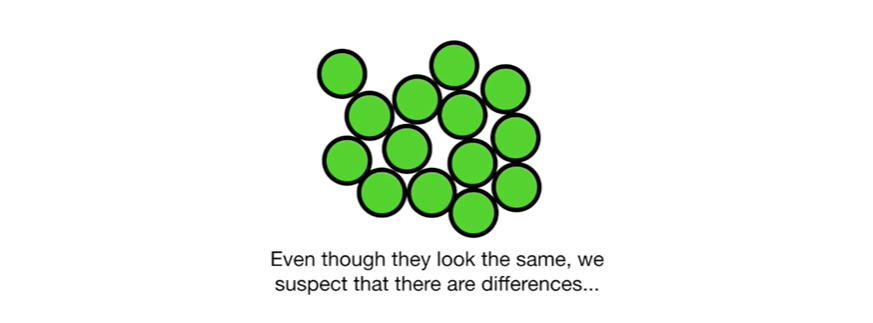
If you want more details than you get here be sure to check out my other PCA video.



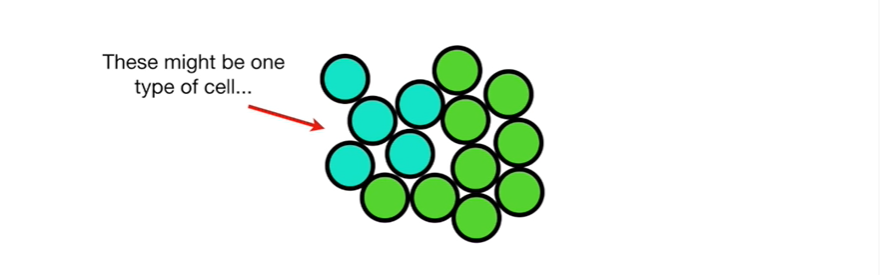
Let's say we had some normal cells.



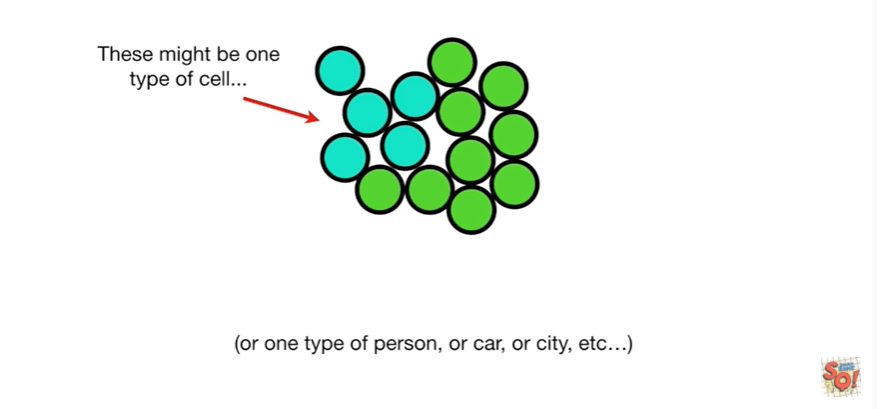
Psst : if you're not a biologist imagine that these could be people or cars or cities or etc… They could be anything.



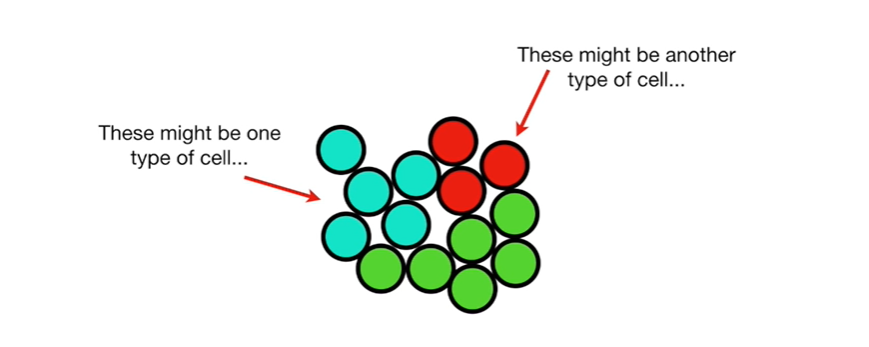
Even though they look the same we suspect that there are differences.



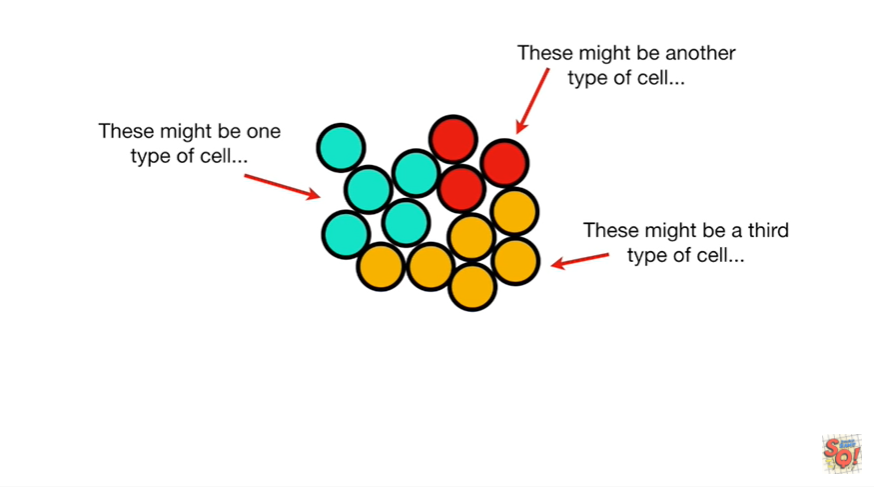
These might be one type of cell



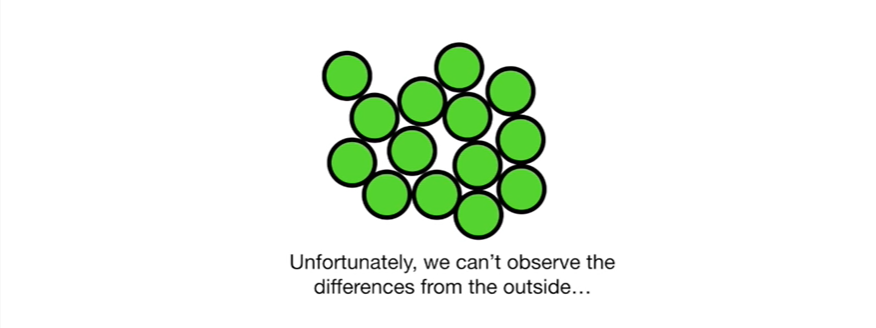
or one type of person or car or city etc…



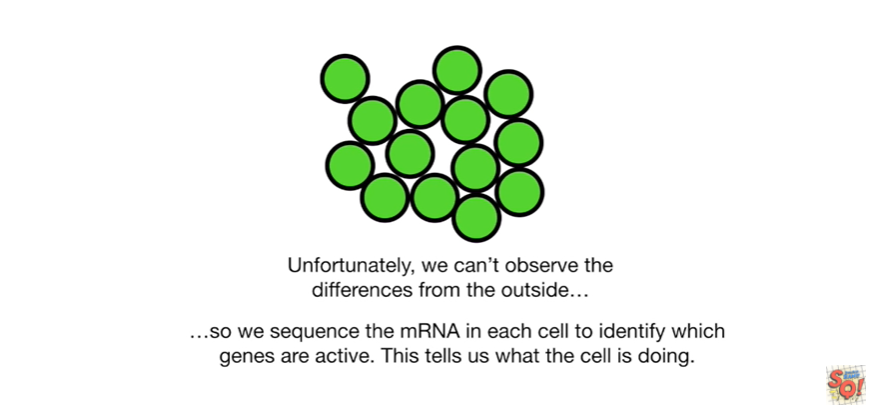
These might be another type of cell



and lastly these might be a third type of cell.

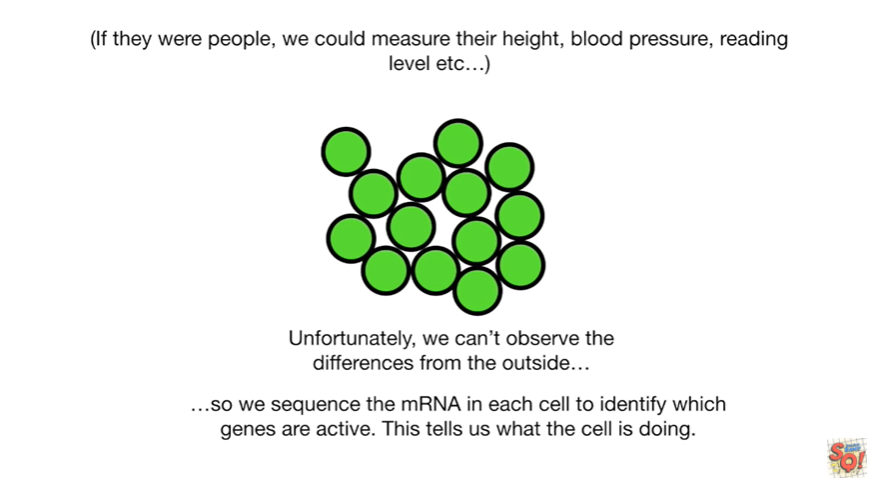


Unfortunately we can't observe differences from the outside

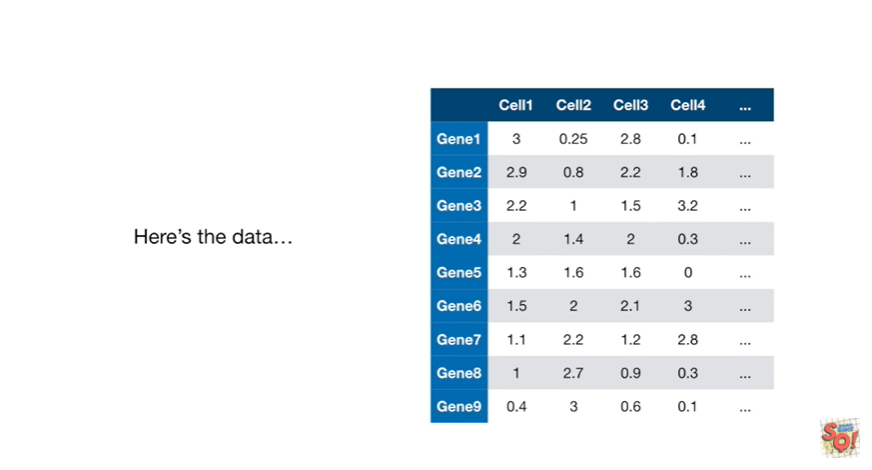


so we sequence the messenger RNA in each cell2 identify which genes are active.

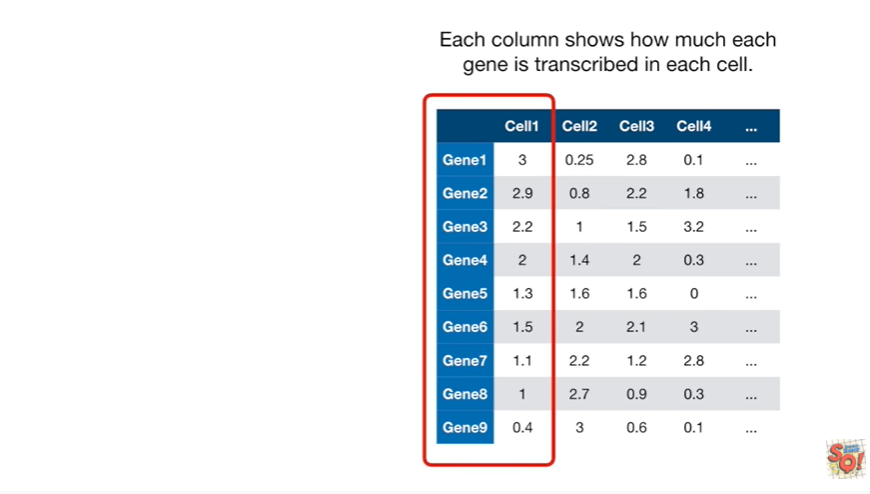
This tells us what the cell is doing.

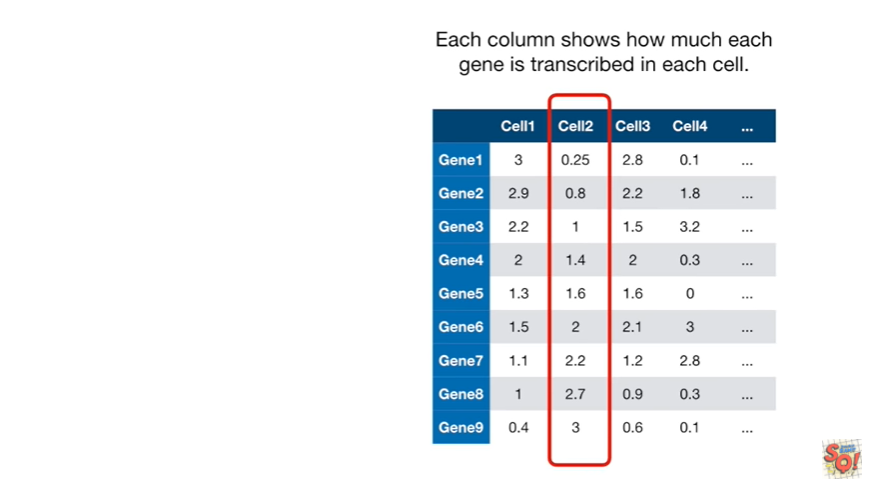


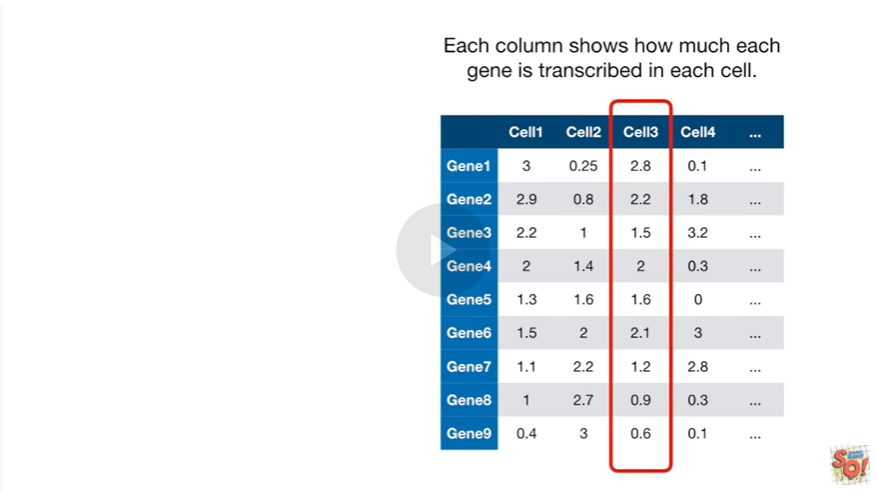
If they were people we could measure their weight blood pressure reading level etc…

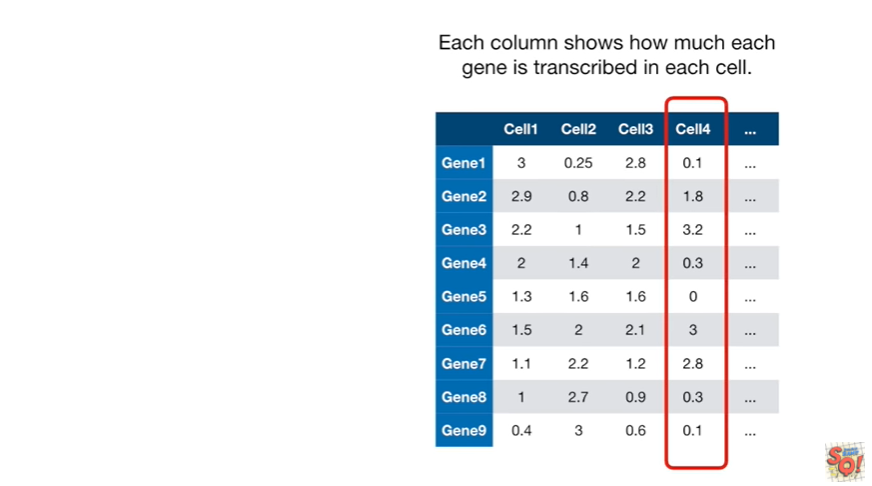


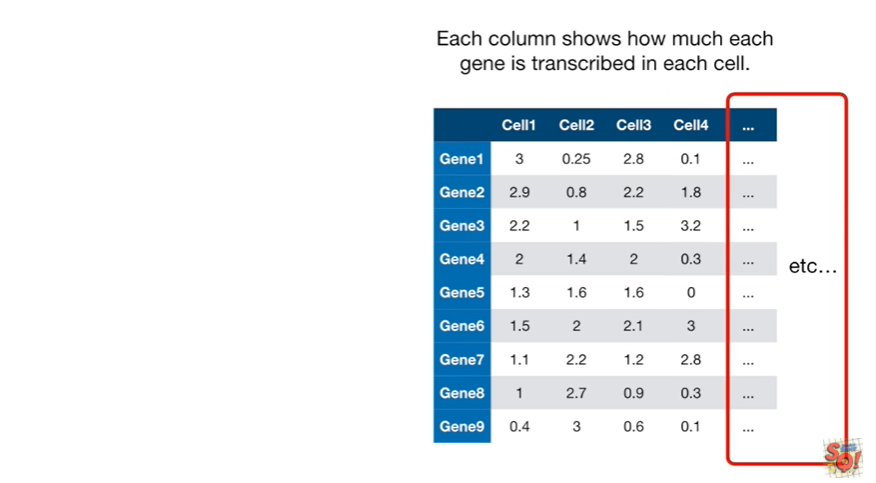
Okay here's the data.



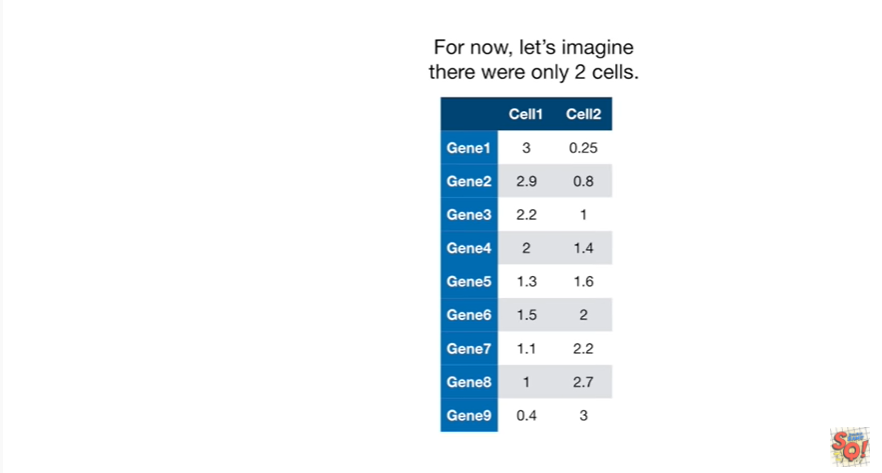




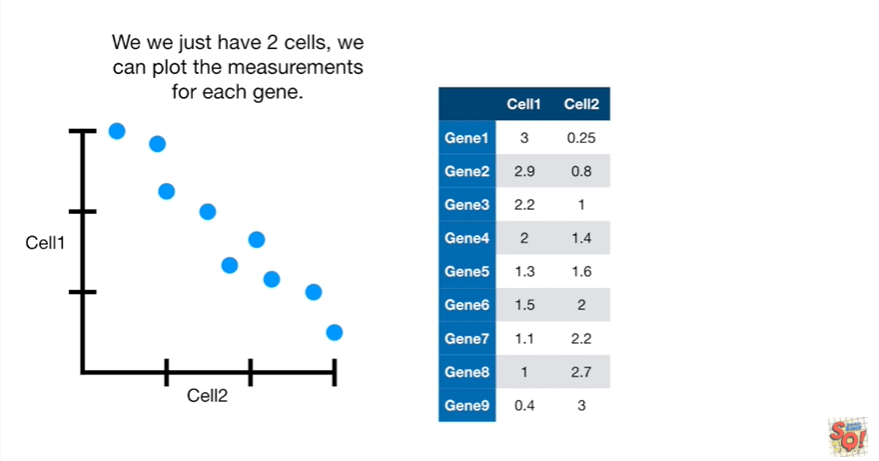




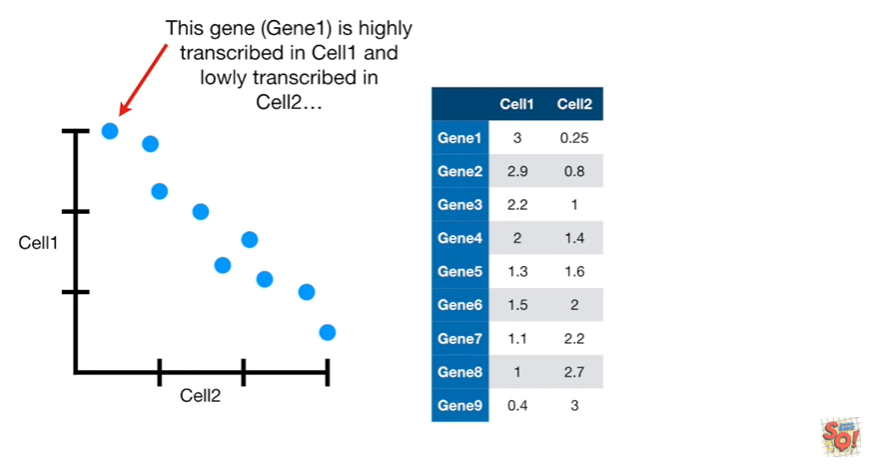
Each column shows how much each gene is transcribed in each cell.



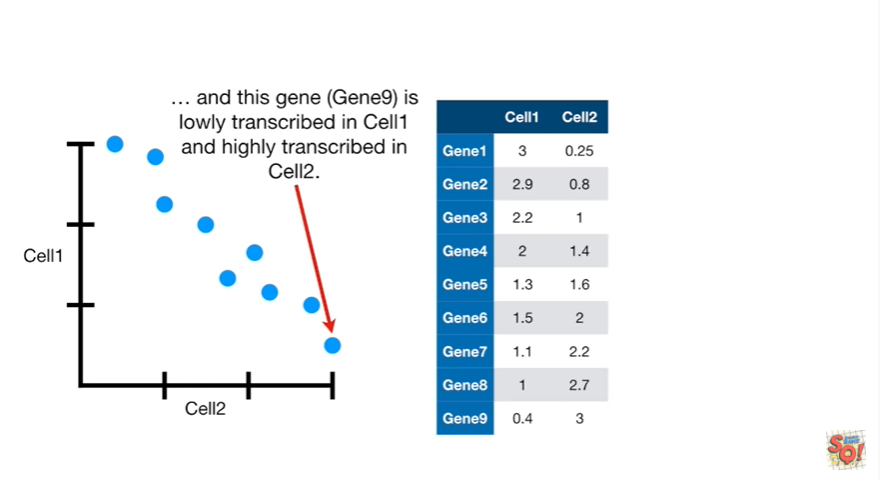
For now let's imagine there are only two cells.



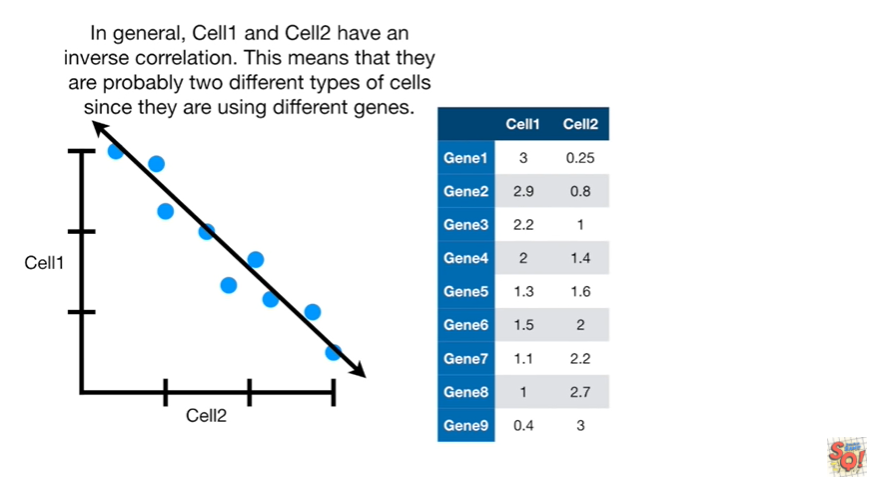
If we just have two cells then we can plot the measurements for each gene.



This gene gene one is highly transcribed in cell1 and lowly transcribed in cell2

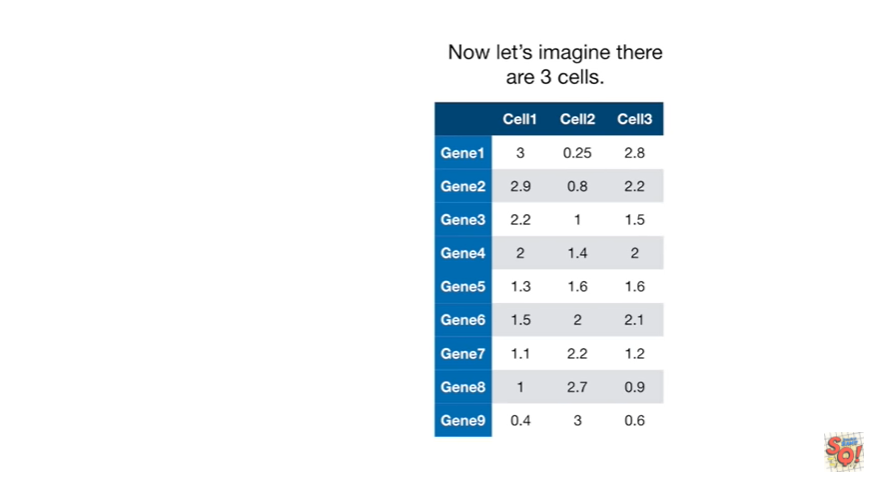


and this gene gene 9 is lowly transcribed in cell 1 and highly transcribed in cell 2.

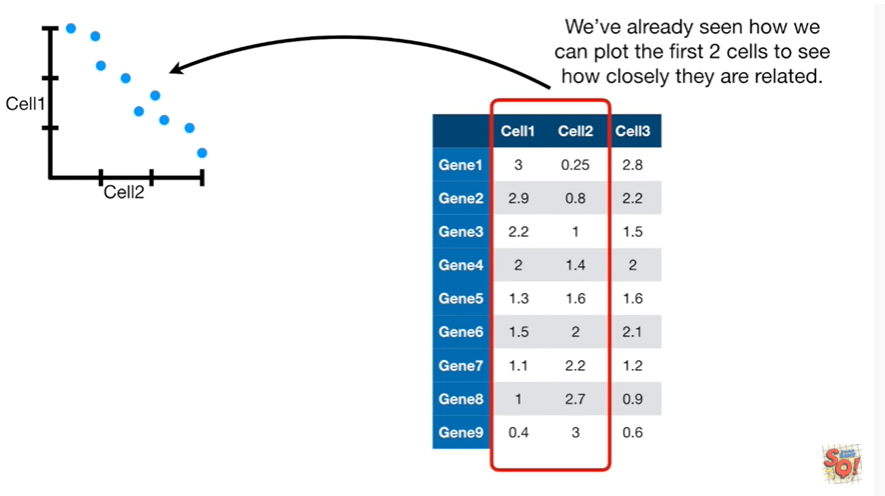


In general cell1 and cell2 have an inverse correlation.

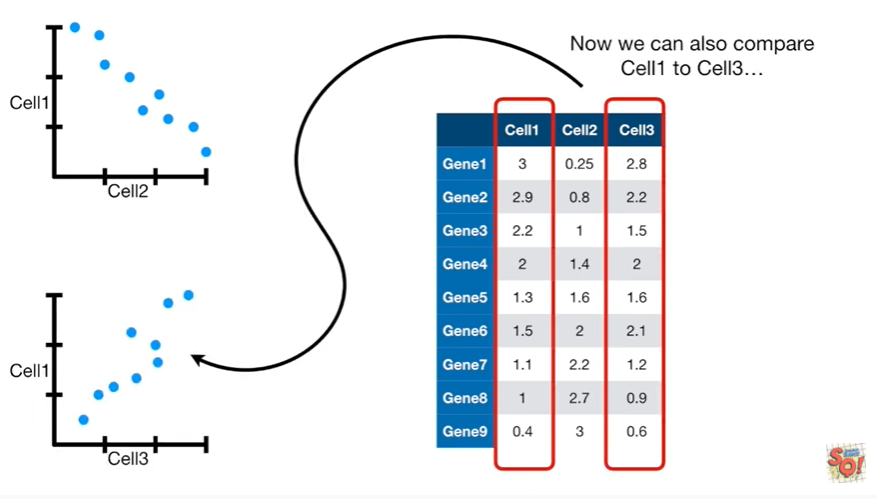
This means that they are probably two different types of cells since they are using different genes.



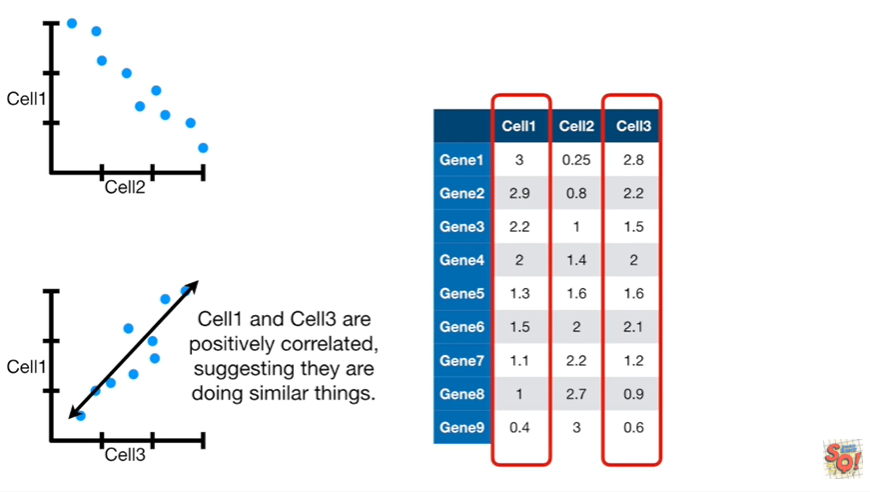
Now let's imagine there are three cells.



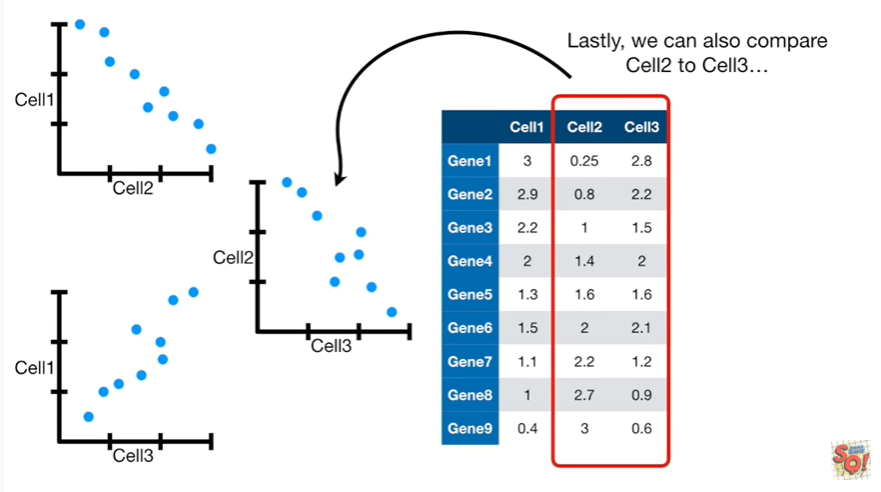
We've already seen how we can plot the first two cells to see how closely they are related.



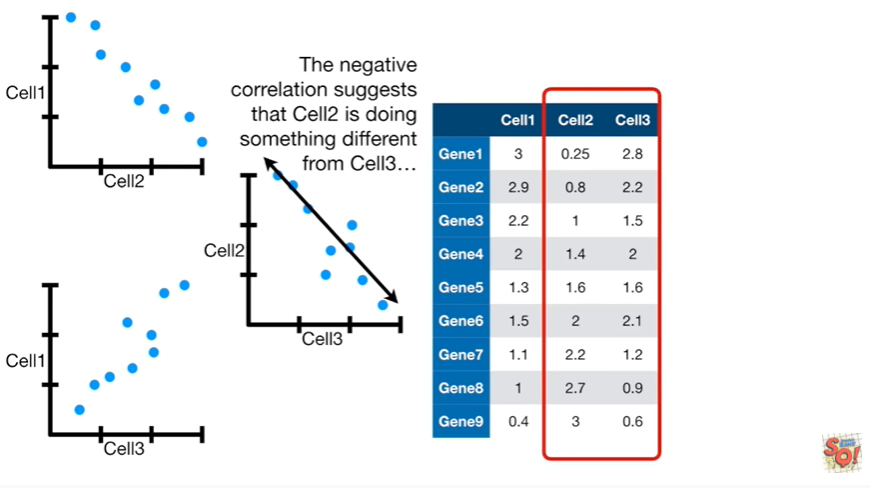
Now we can also compare cell1 to sell three



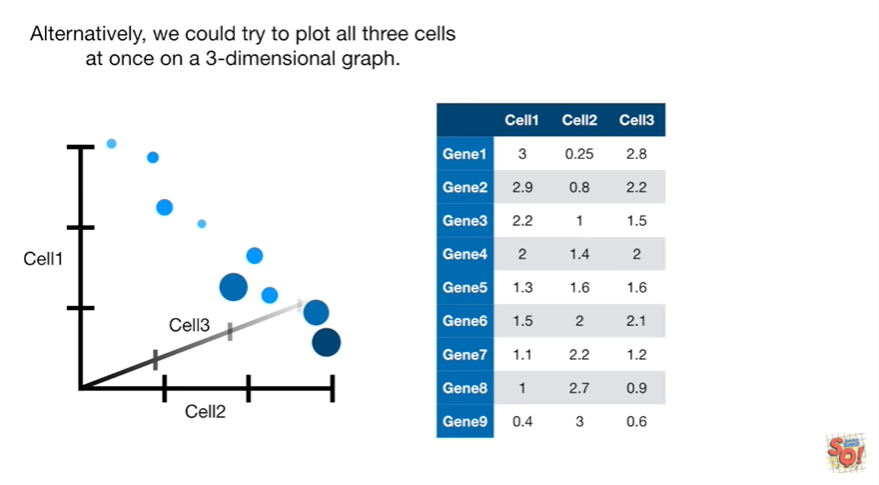
cell1 and cell3 are positively correlated suggesting they are doing similar things.



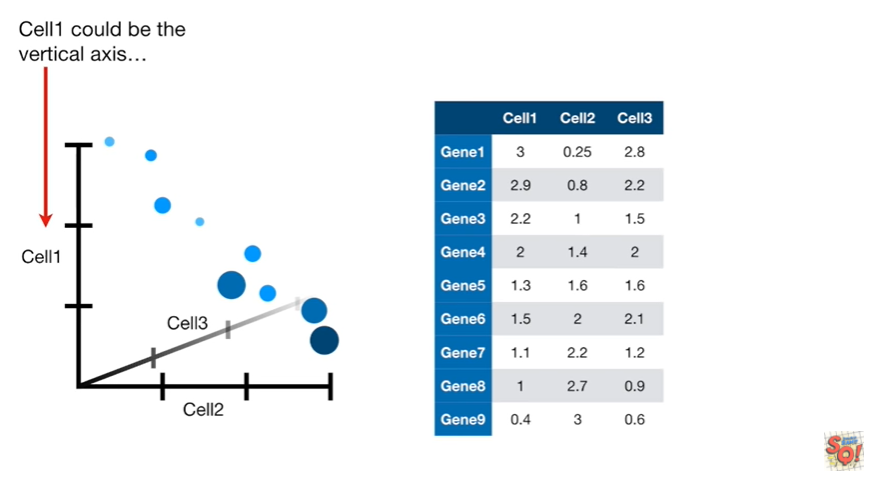
Lastly we can also compare cell2 to cell3.



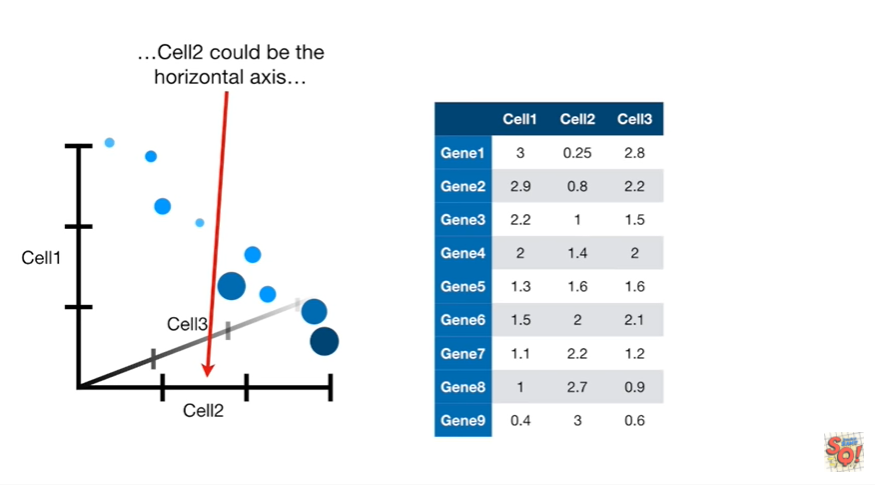
The negative correlation suggests that cell2 is doing something different from cell 3.



Alternatively we could try to plot all three cells at once on a three dimensional graph.



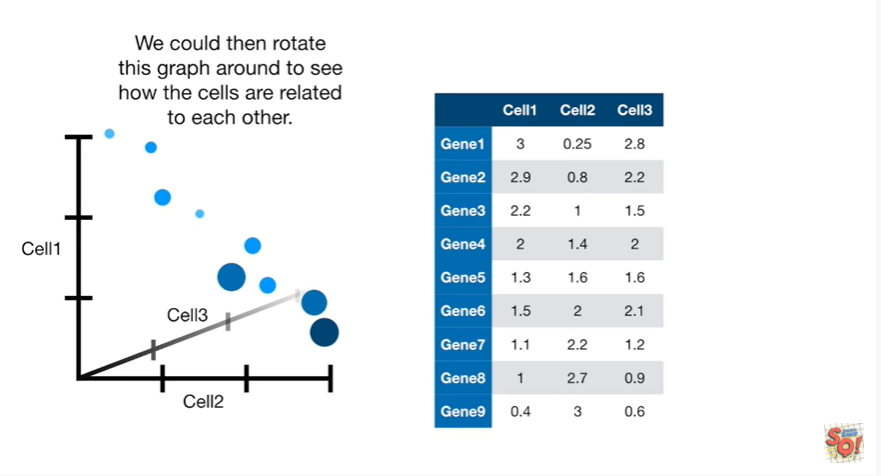
Cell1 could be the vertical axis



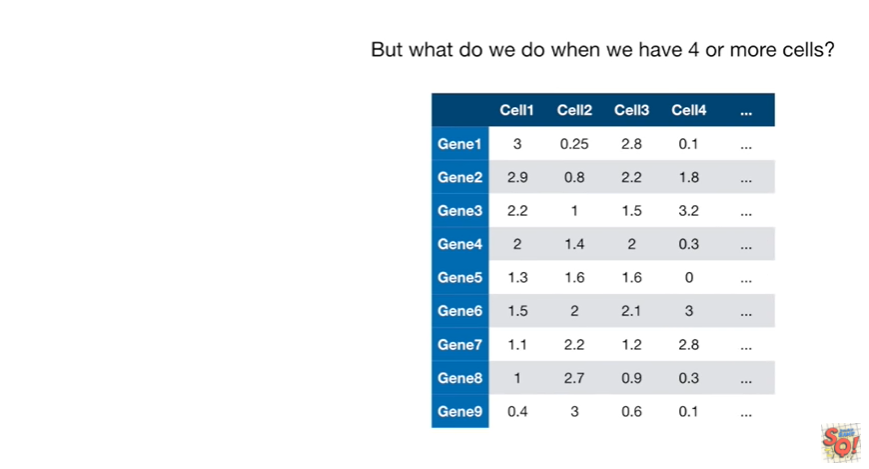
cell2 could be the horizontal axis



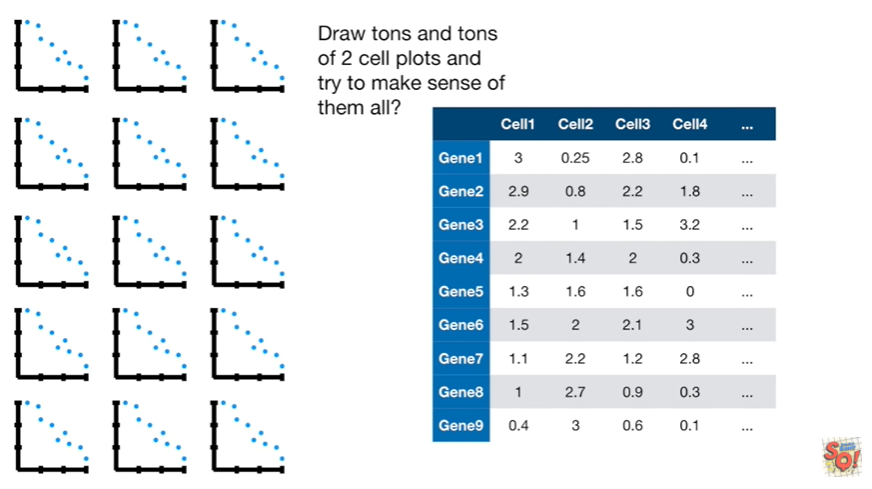
and cell3 could be depth.



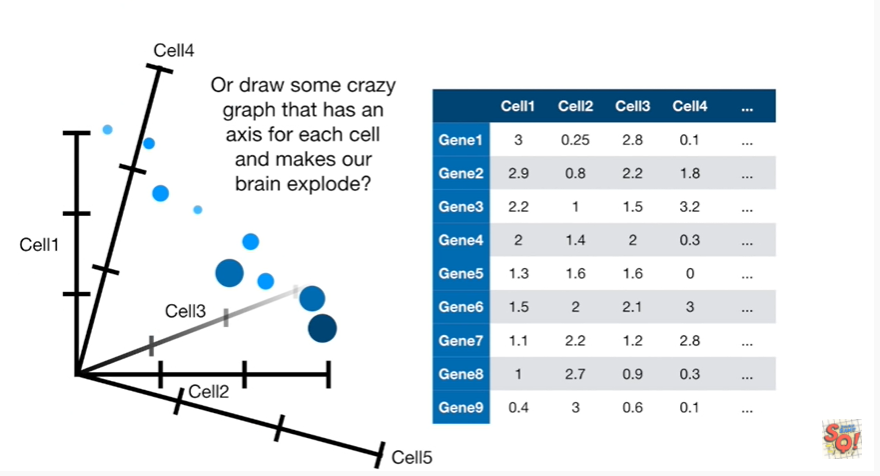
We could then rotate this graph around to see how the cells are related to each other.



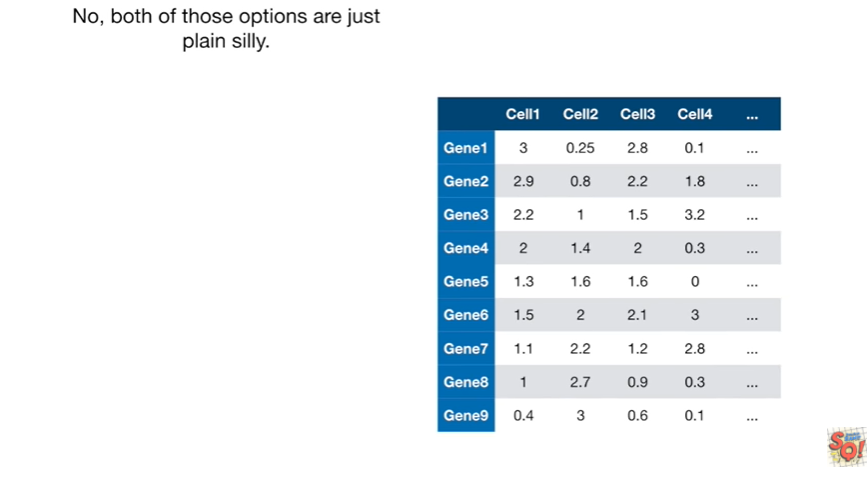
But what do we do when we have four or more cells ?



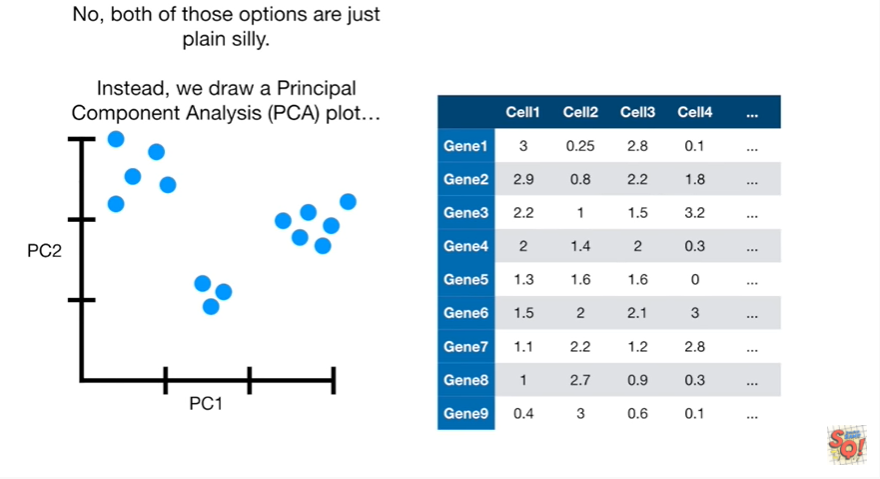
Draw tons and tons of to sell plots and try to make sense of them all ?



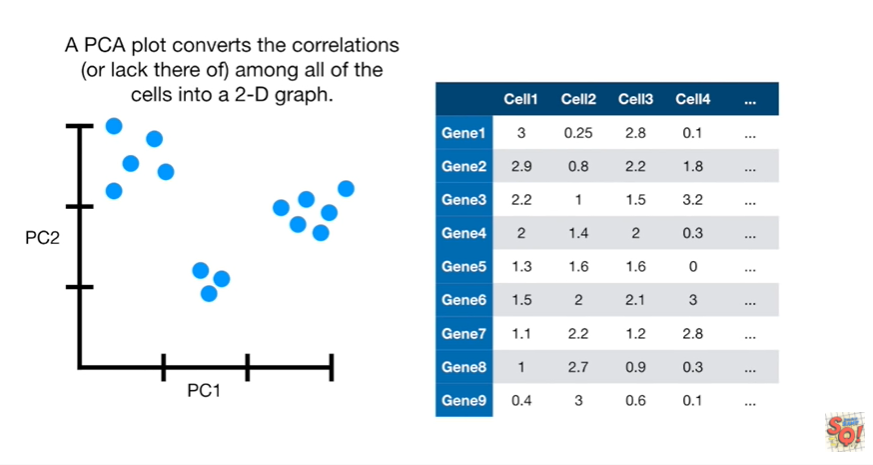
Or draw some crazy graph that has an axis for each cell and makes our brain explode ?



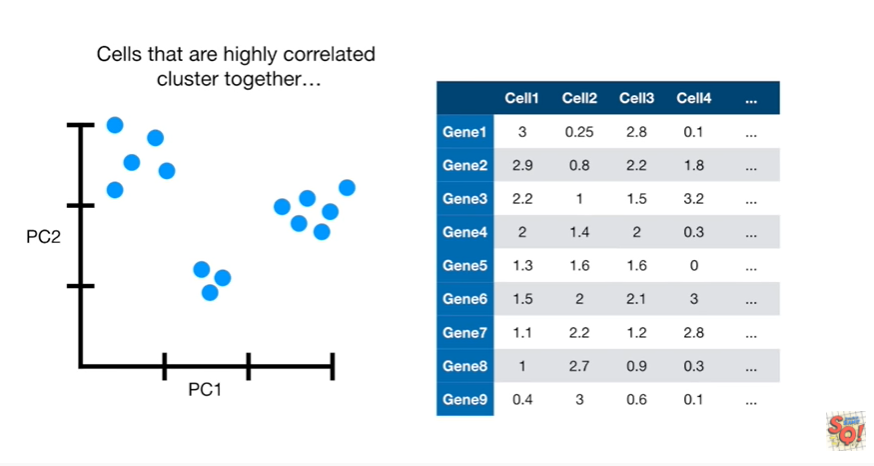
No both of those options are just plain silly.



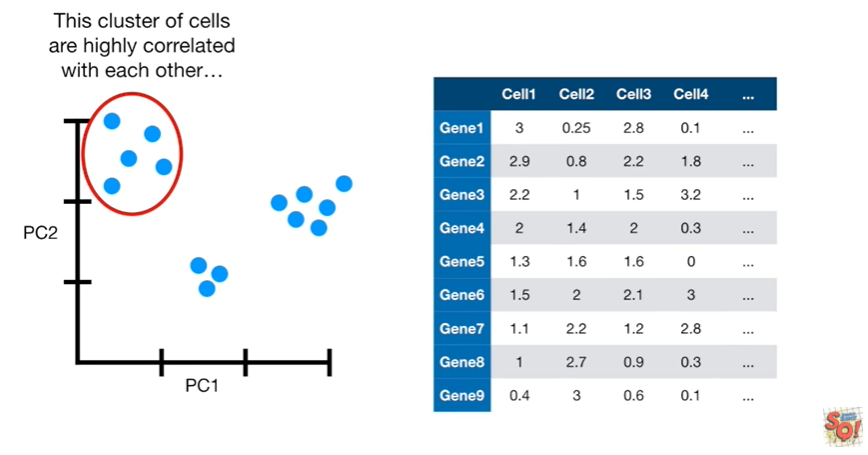
Instead we draw a principal component analysis or PCA plot.



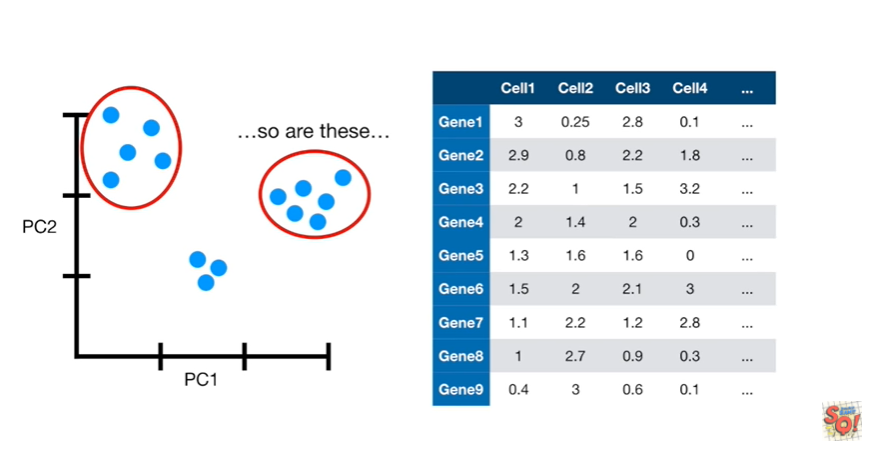
A PCA plot converts the correlations or lack thereof among the cells into a 2d graph.



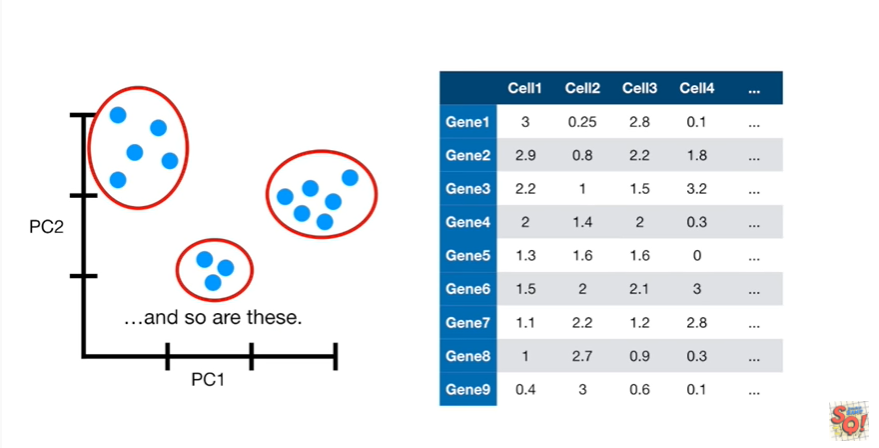
Cells that are highly correlated cluster together.



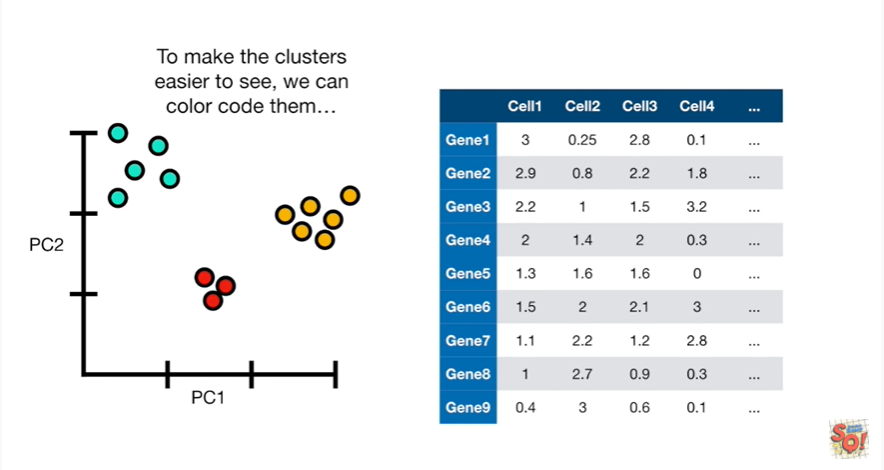
This cluster of cells are highly correlated with each other



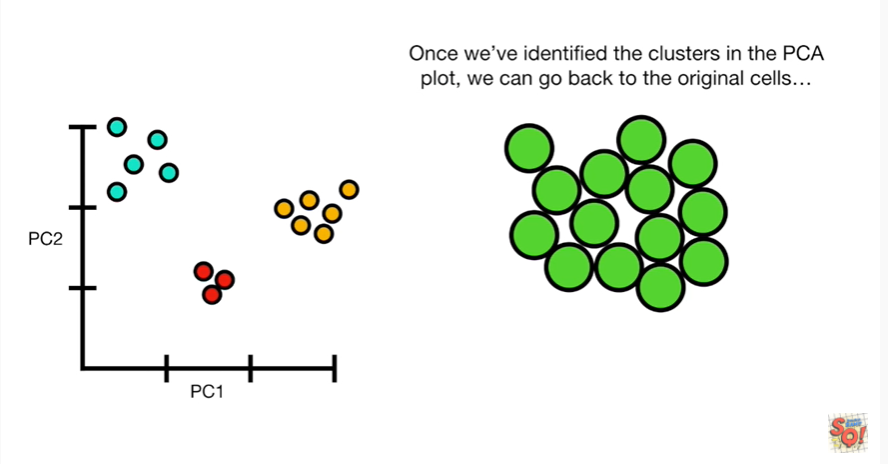
so are these



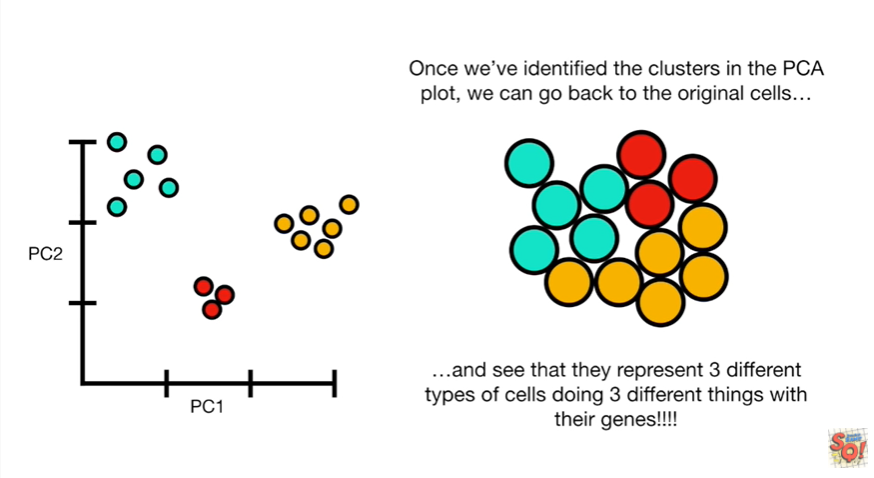
and so are these.



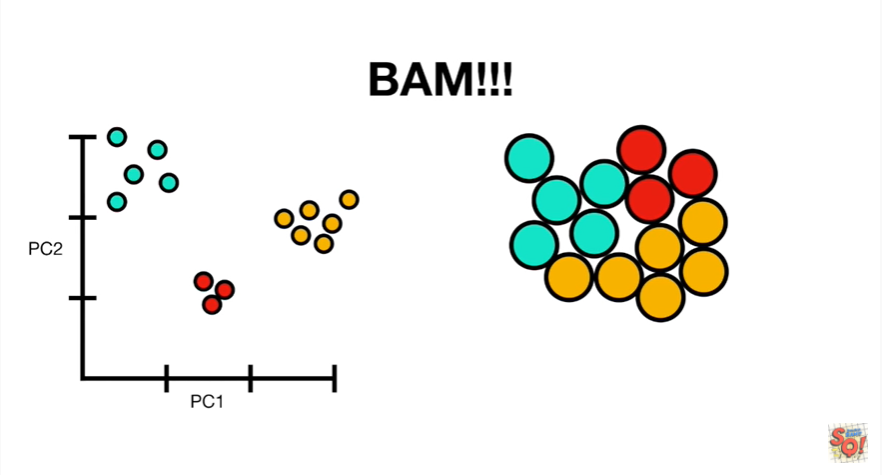
To make the clusters easier to see we can color-code them.



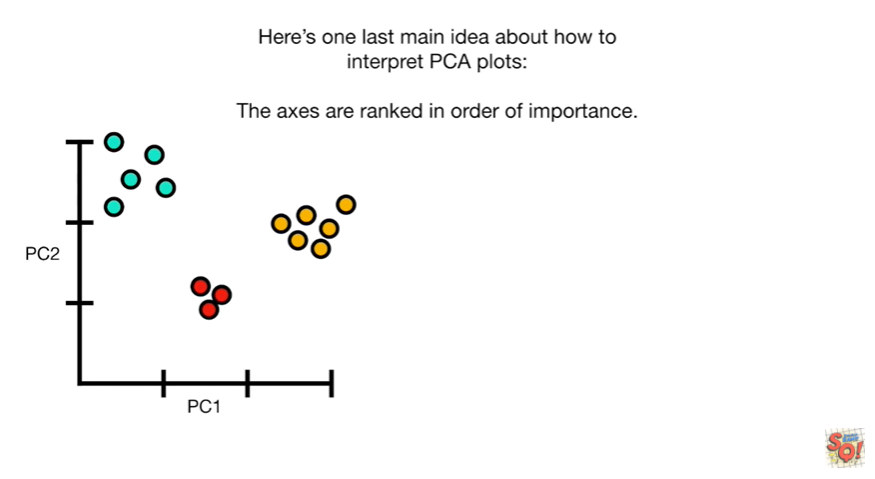
Once we've identified the clusters in the PCA plot we can go back to the original cells



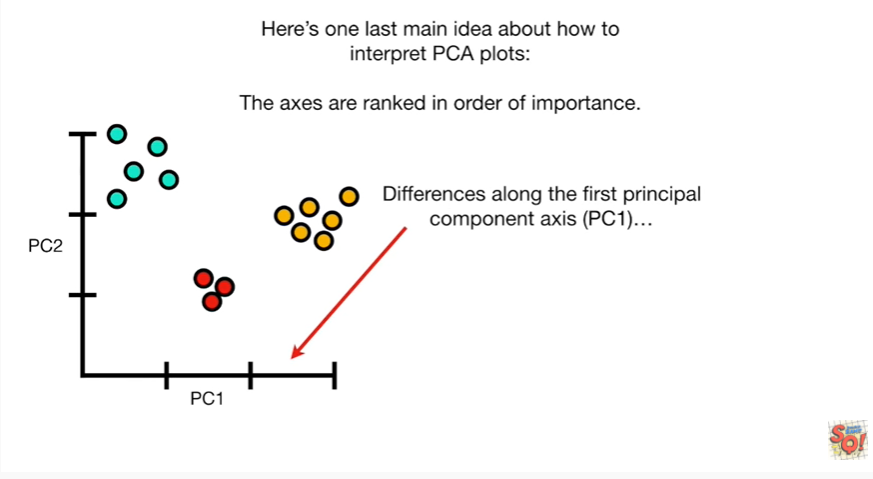
and see that they represent three different types of cells doing three different types of things with their genes !!!!



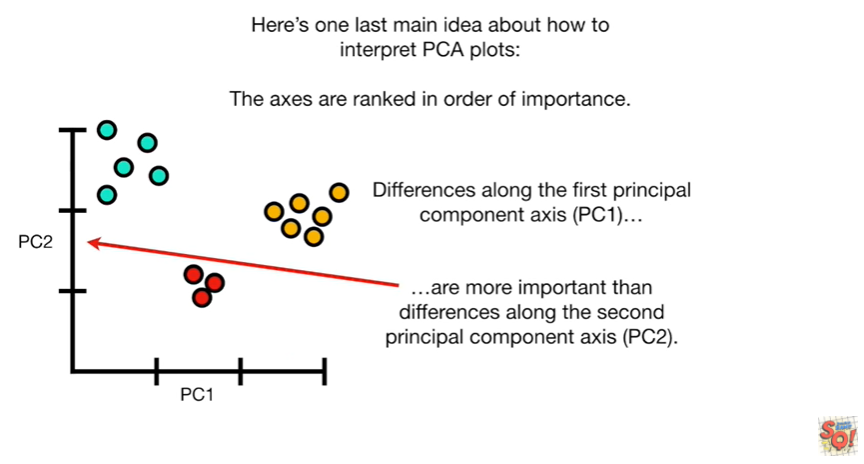
BAM !!!



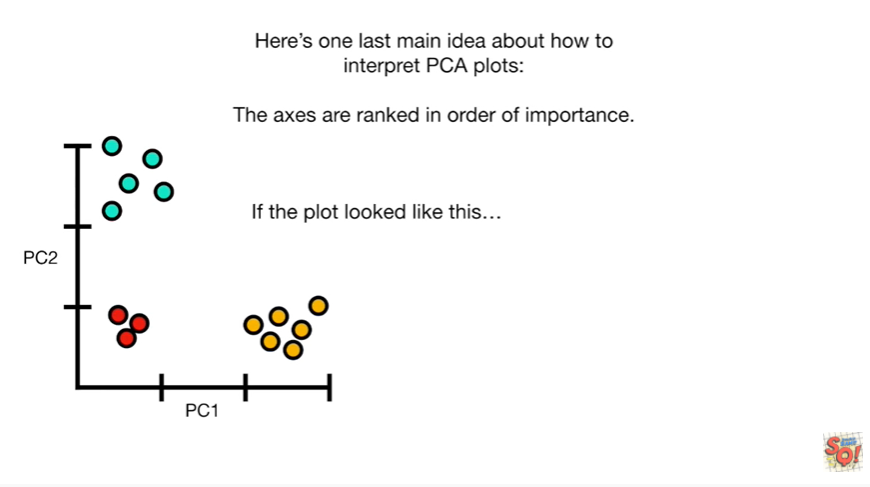
Here's one last main idea about how to interpret PCA plots : the axes are ranked in order of importance.



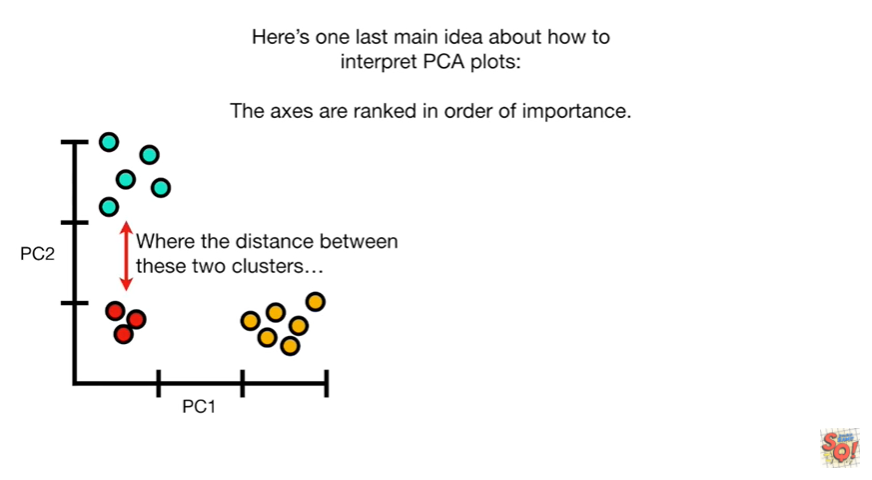
Differences among the first principal component access (PC1)



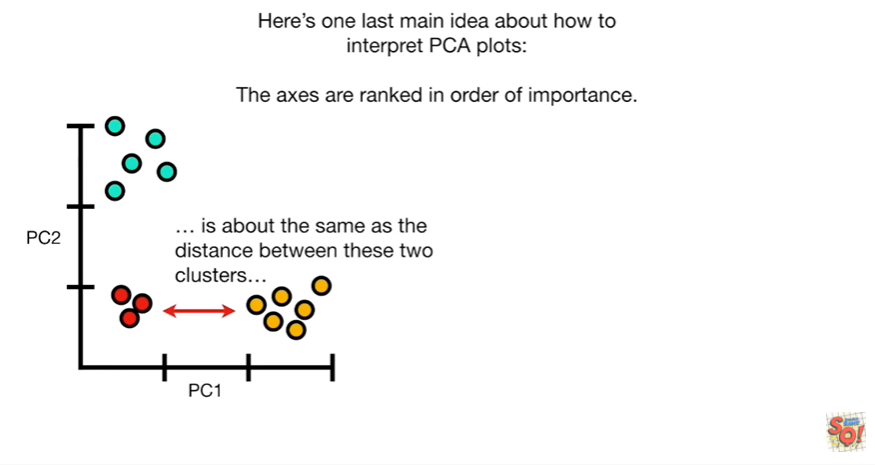
are more important than differences along the second principal component access PC2.



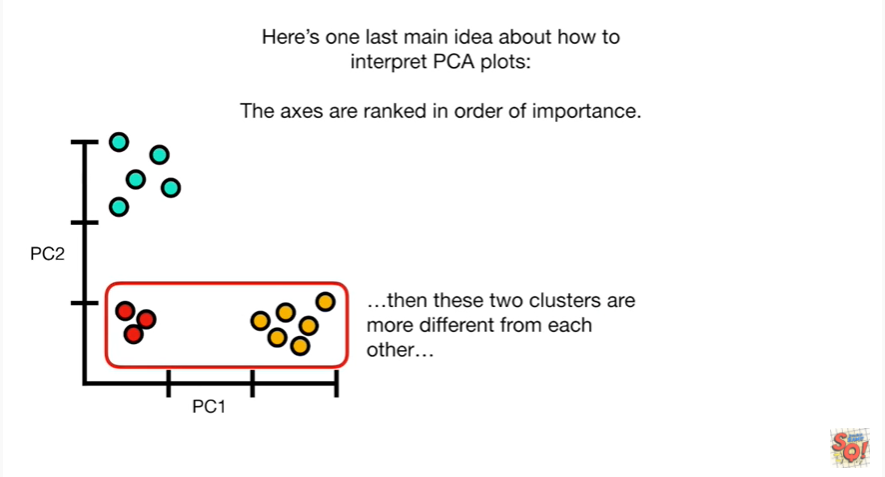
If the plot looked like this



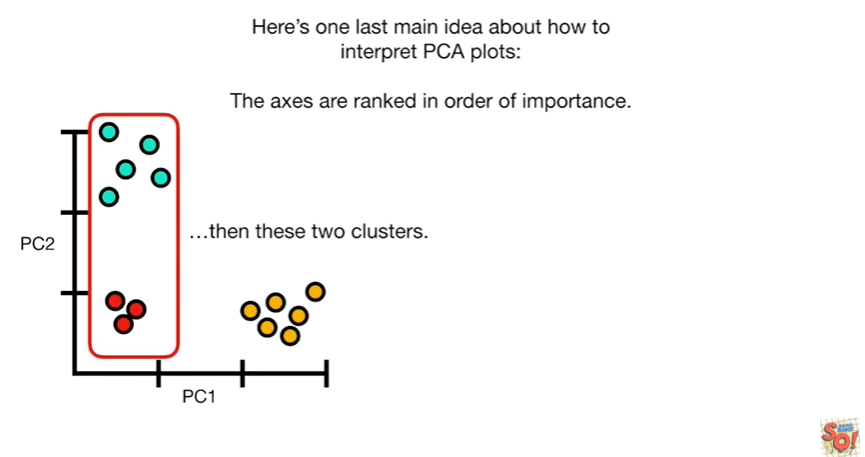
where the distance between these two clusters



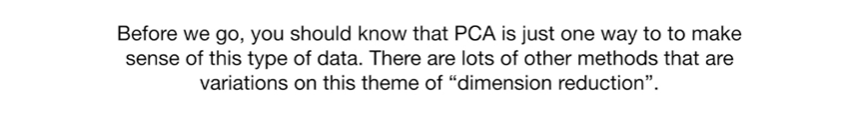
is about the same as the distance between these two clusters



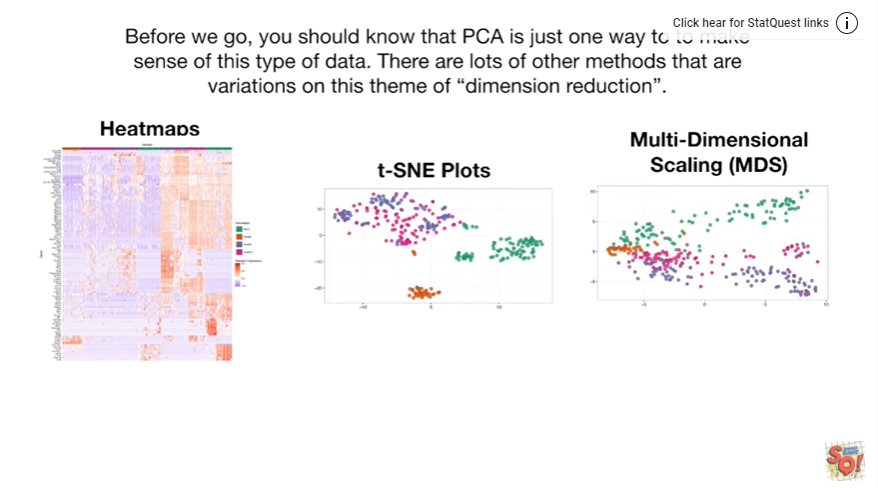
then these two clusters are more different from each other



than these two clusters.

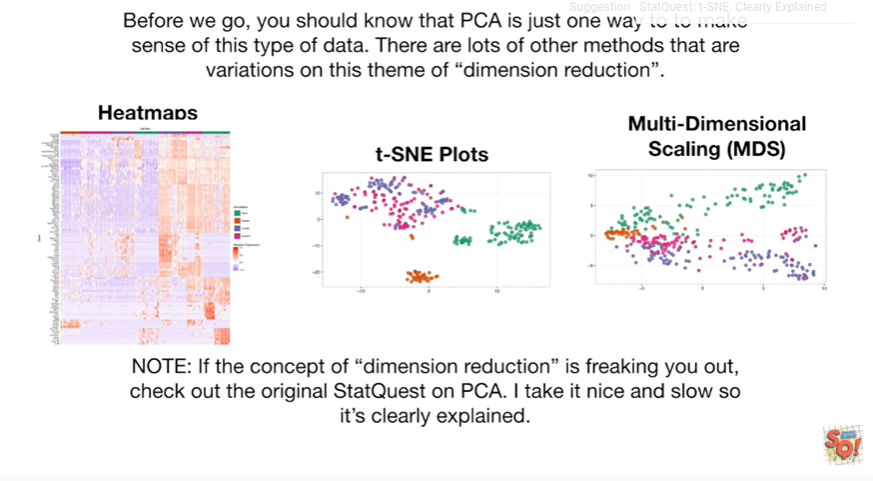


Before we go you should know that PCA is just one way to make sense of this type of data there are lots of other methods that are variations on this theme of dimension reduction.



These methods include heat maps, t-SNE plots and multiple dimension scaling (MDS) plots.

The good news is that I've got stat quests for all of these so you can check those out if you want to learn more.



Note : if the concept of dimension reduction is freaking you out check out the original stat quest on PCA.