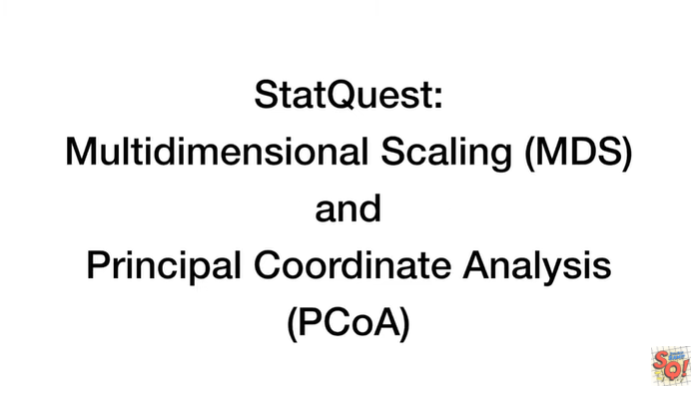
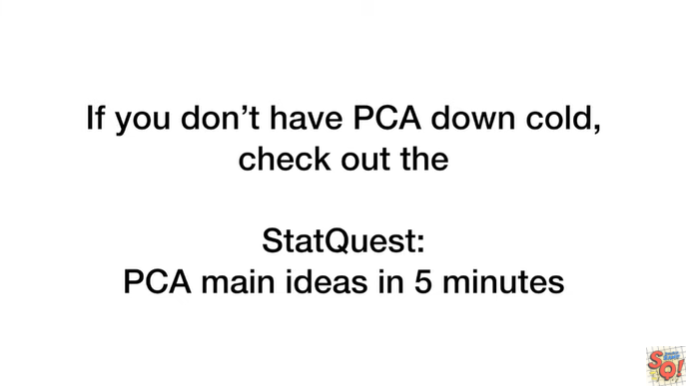
<https://www.youtube.com/watch?v=GEn-_dAyYME&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=30>



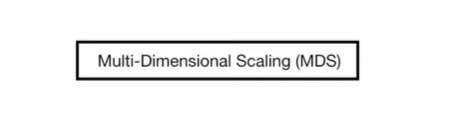
Today we're going to be talking about multi-dimensional scaling MDS and principle coordinate analysis PCoA.



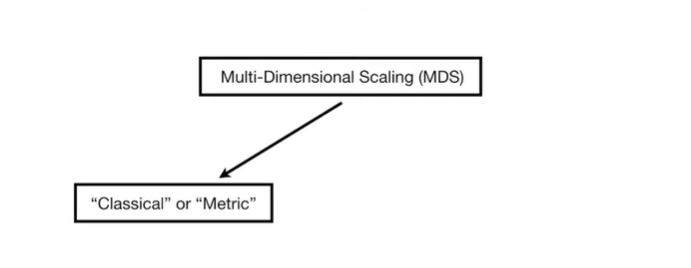
First of all if you don't have principal component analysis PCA down cold, check out the stat quest PCA main ideas in five minutes.

Principal component analysis and multi-dimensional scaling are both very very similar.

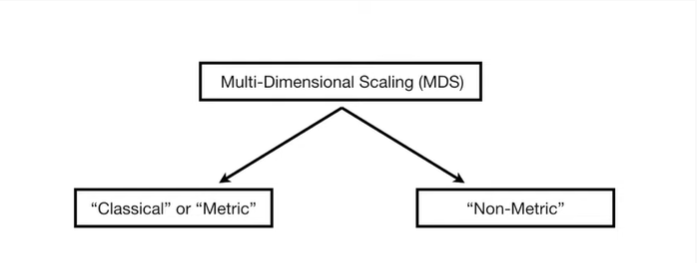
So I want you to be able to understand PCA before we move on with this one.



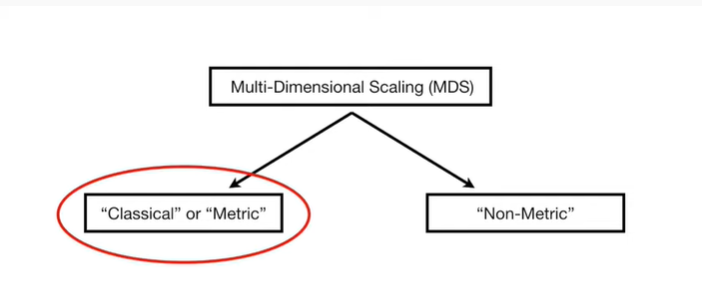
Also to be clear about what we will cover in this stat quest there are two types of multi-dimensional scaling.



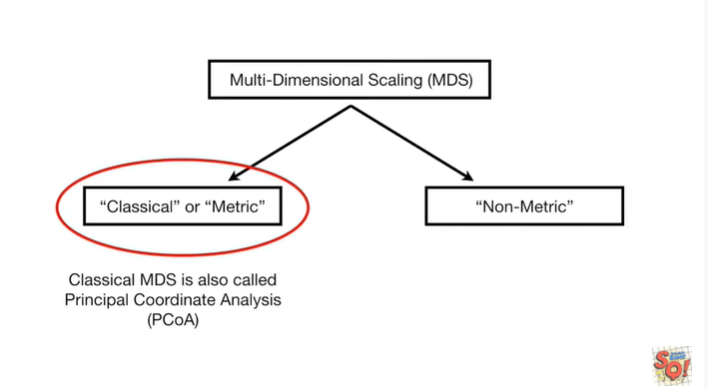
There's classical or metric multi-dimensional scaling



versus non metric multi-dimensional scaling.



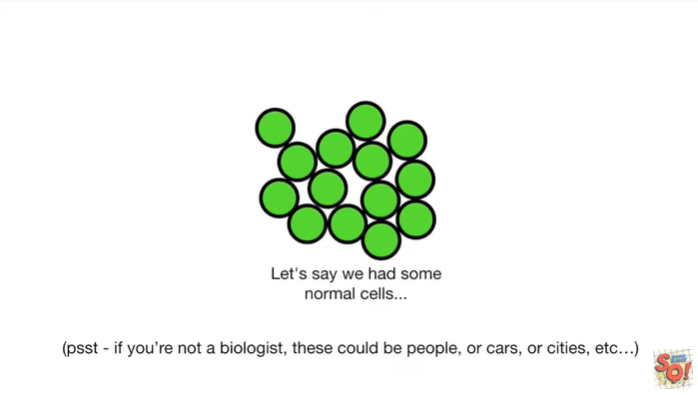
I'm only going to talk about classical multi-dimensional scaling in this stat quest.



And classical multi-dimensional scaling is the exact same thing as principle coordinate analysis (PCoA).

Okay enough preliminary stuff.

Let's dive in.

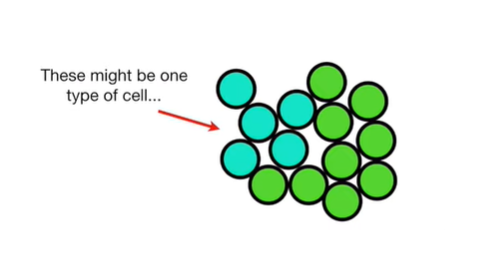


Let's say we had some normal cells.

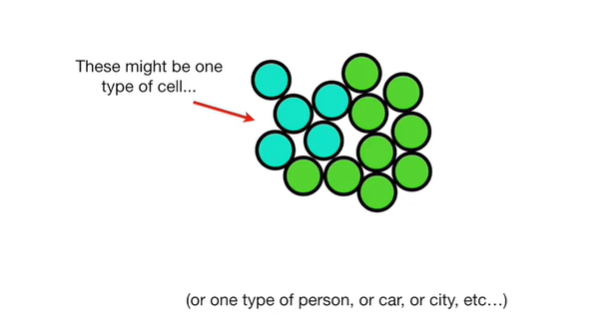
Pssst, if you're not a biologist these could be people, or cars, or cities etc…



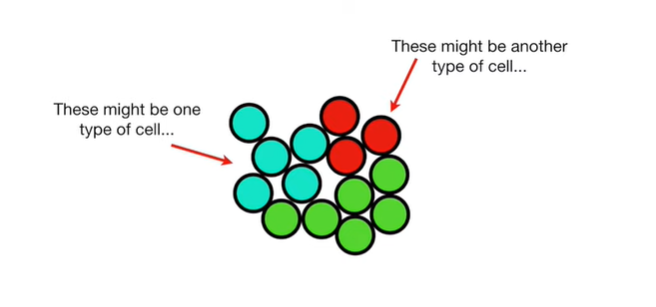
Even though they look the same we suspect that there are differences.



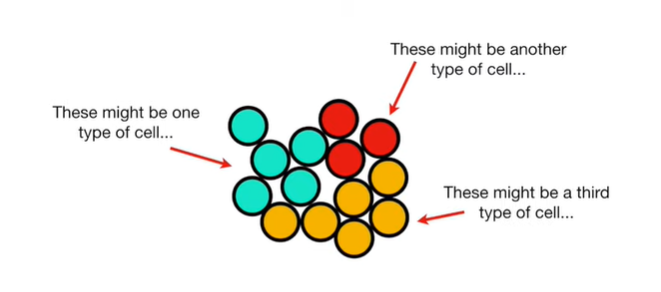
These might be one type of cell



or one type of person or car or city etc…



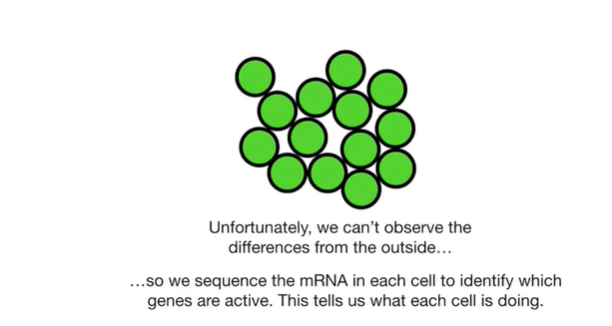
These might be another type of cell



and these might be a third type of cell.

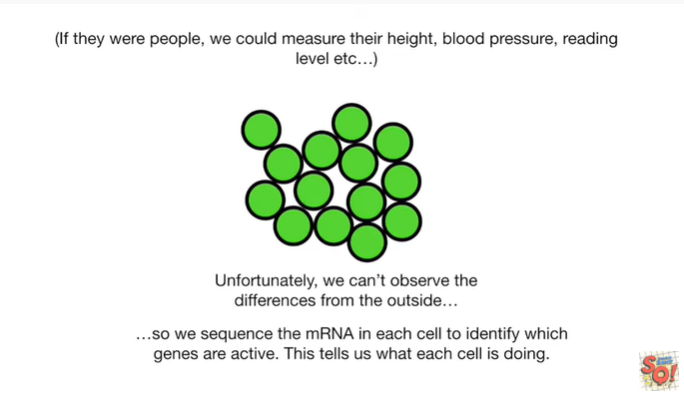


Unfortunately, we can't observe the differences from the outside.

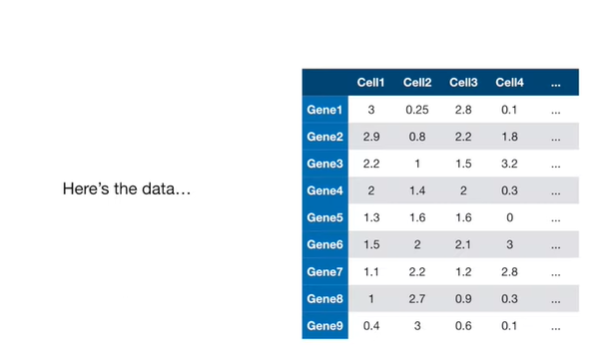


So we sequence the messenger RNA in each cell to identify which genes are active.

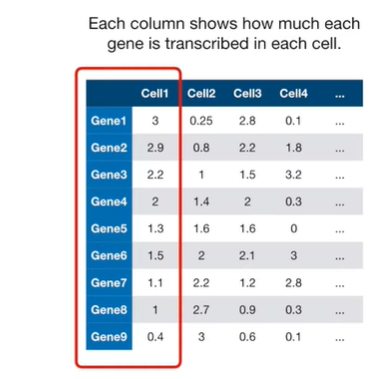
This tells us what each cell is doing.

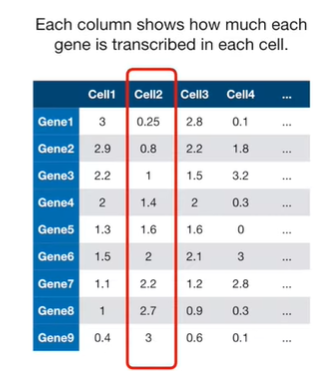


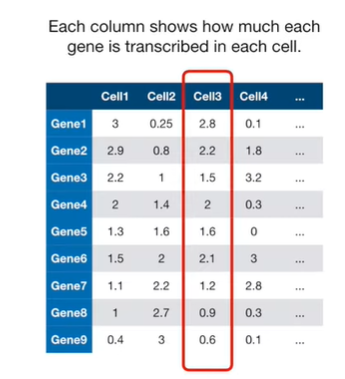
Alternatively if they were people we could measure their height, blood pressure, reading level, etc…

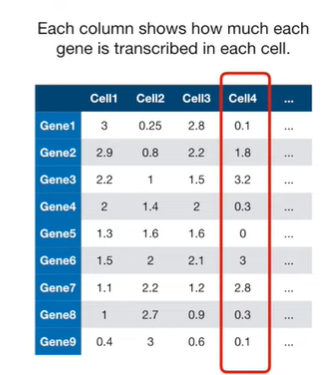


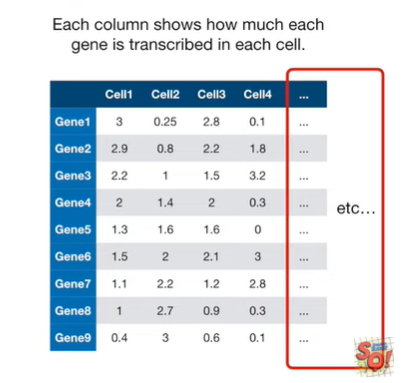
Here's the data.



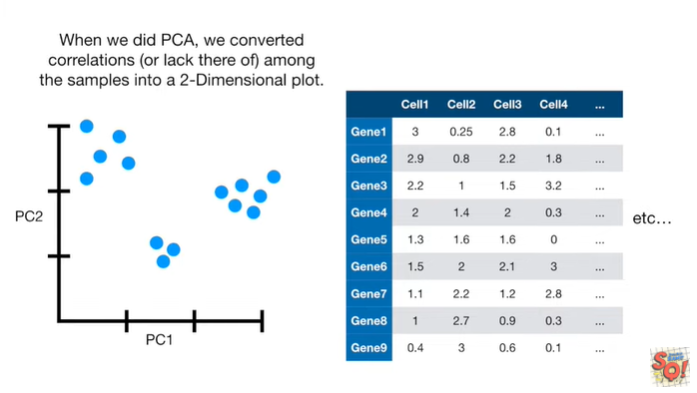




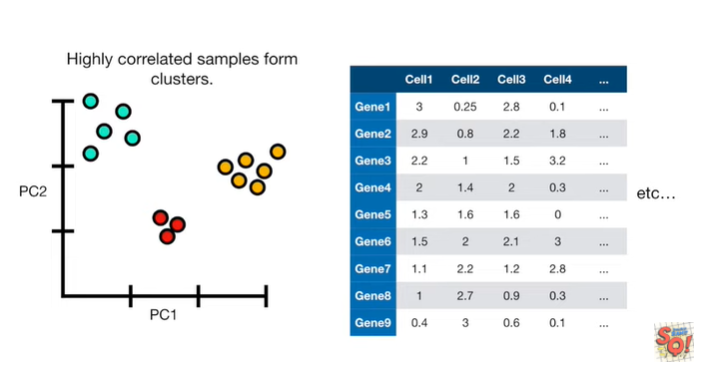




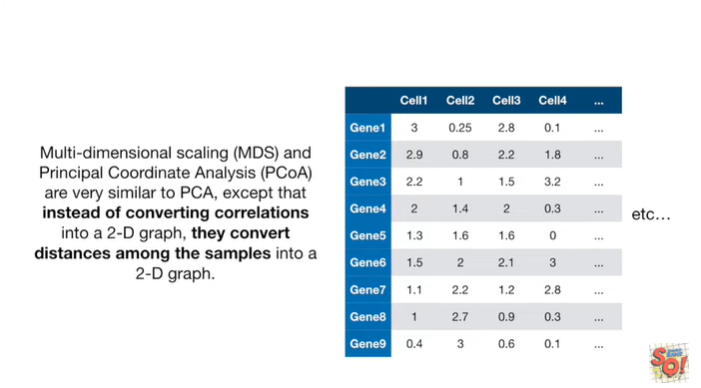
Each column shows how much each gene is transcribed in each cell.



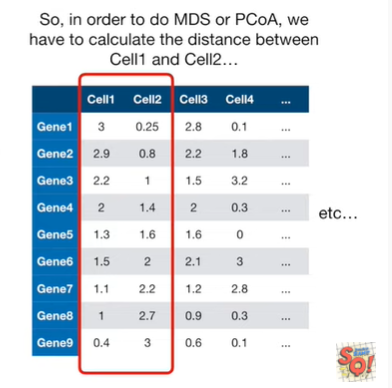
When we did PCA, in the stat quest that explains the main ideas of PCA in five minutes, we converted the correlations, or lack there of, among the samples into a 2d principal component analysis plot.



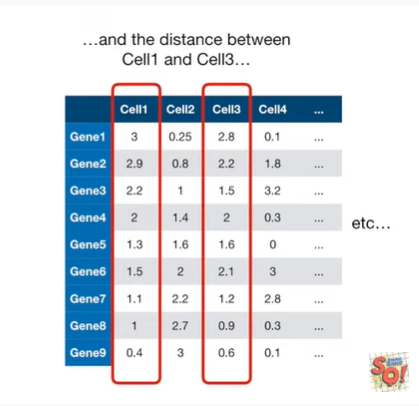
And we saw that highly correlated samples form clusters.



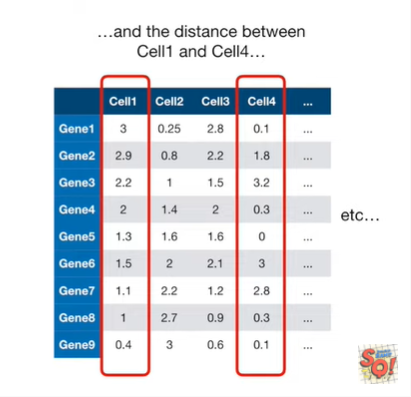
Multi-dimensional scaling MDS and principle coordinate analysis PCoA are very similar to PCA except that instead of converting correlations into a 2d graph, they convert distances among the samples into a 2d graph.



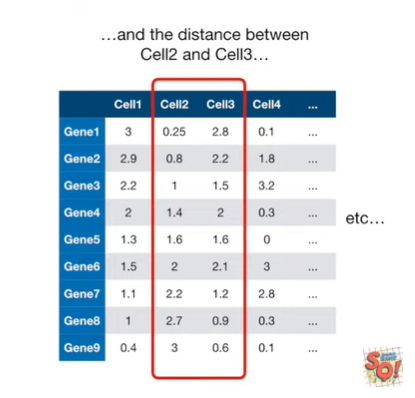
So in order to do MDS or PCoA, we have to calculate the distance between cell one and cell two



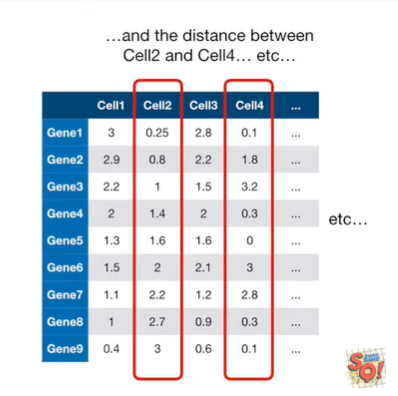
and the distance between cell one and cell three



and the distance between cell one and cell four

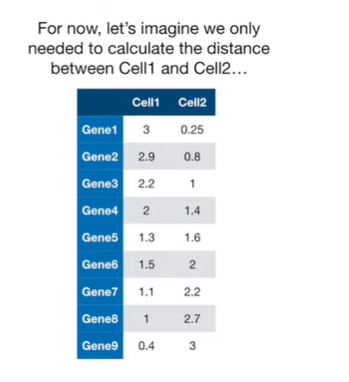


and the distance between cell two and cell three.

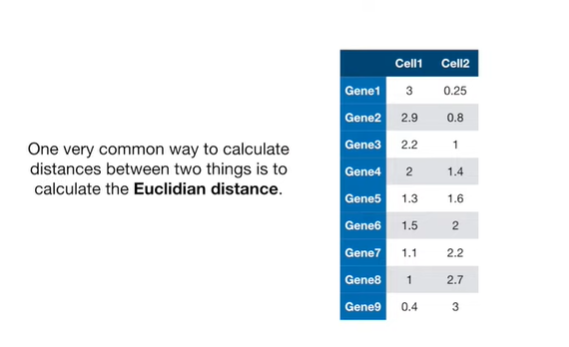


All right you get the idea, we calculate the distance between every pair of cells.

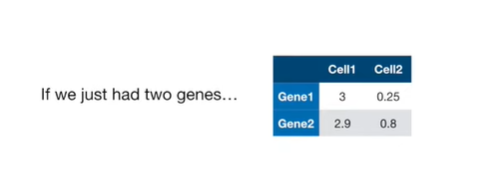
Now let's just talk about how to calculate distances.



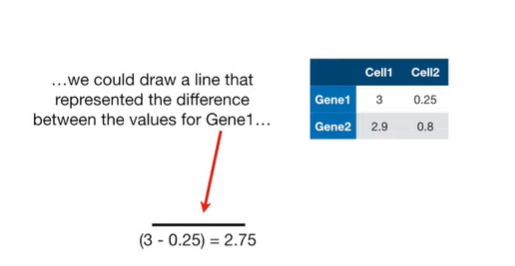
For now, let's imagine we only needed to calculate the distance between cell one and cell two.



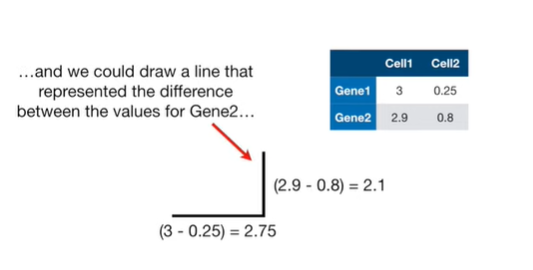
One very common way to calculate distances between two things is to calculate the Euclidean distance.



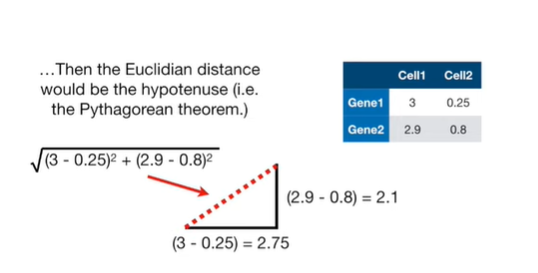
If we just had two genes



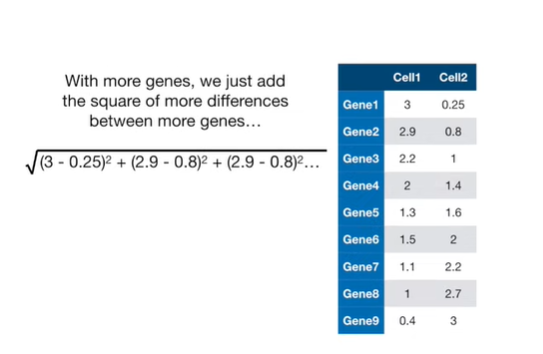
we could draw a line that represented the difference between the values for gene 1



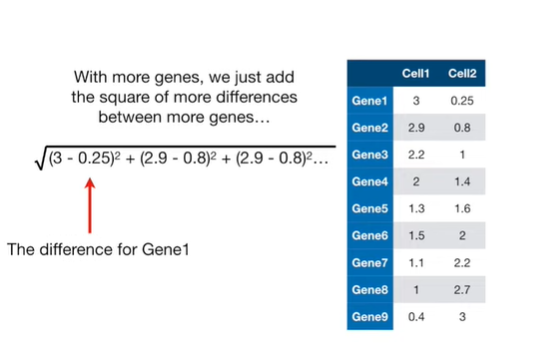
and we could draw a line that represented the difference between the values for gene 2.



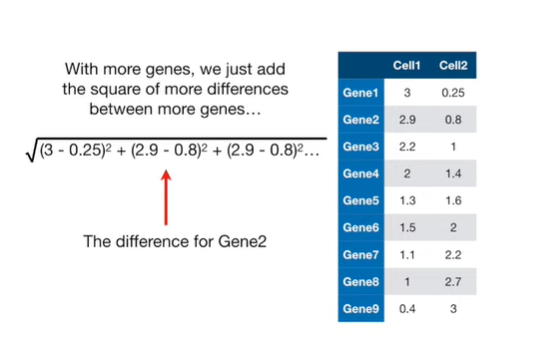
Then the Euclidean distance between cell one and cell 2 would be the hypotenuse ie the Pythagorean theorem.



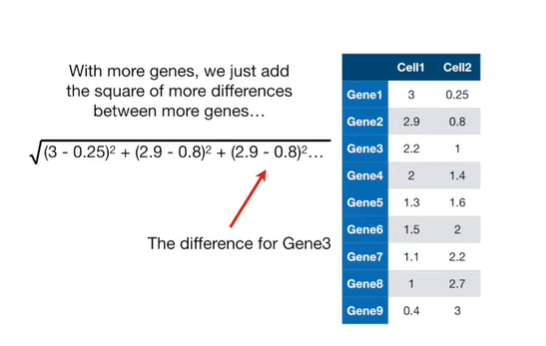
With more genes we just add the square of more differences between more genes.



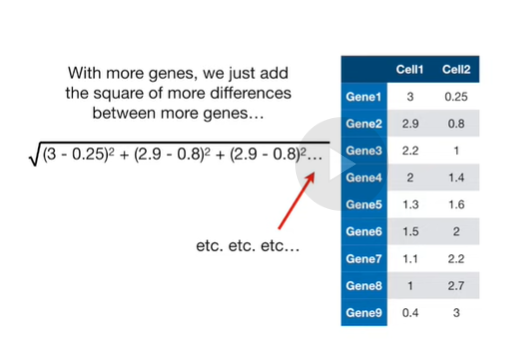
This is the difference for gene one this is.



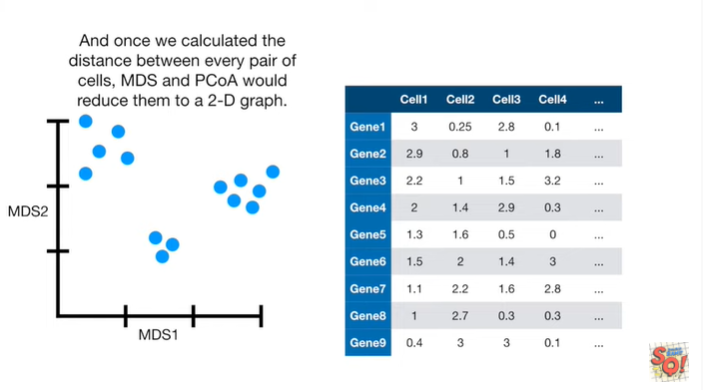
This is the difference for gene 2.



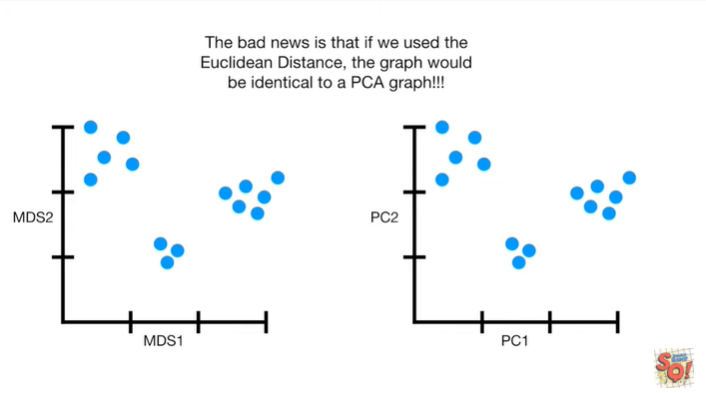
And this is the difference for gene three



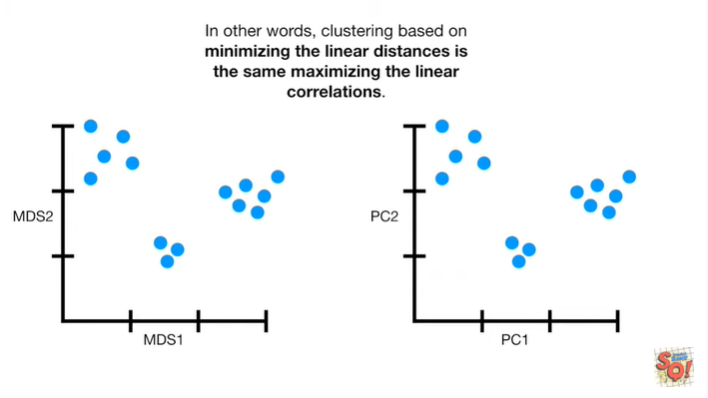
etc etc etc.



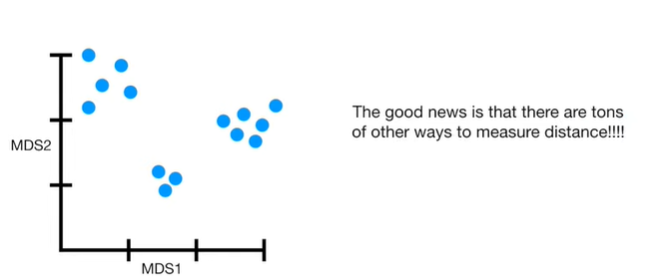
And once we calculated the distance between every pair of cells, MDS and P CoA would reduce them to a two-dimensional graph.



The bad news is that if we use the Euclidean distance the graph would be identical to a PCA graph.

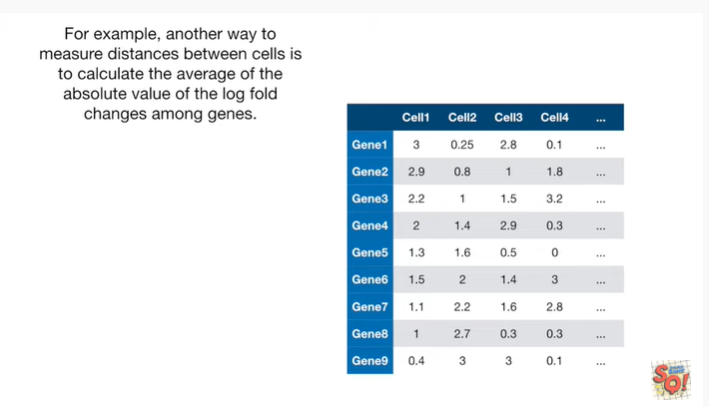


In other words clustering based on minimizing the linear distances is the same as maximizing the linear correlations.

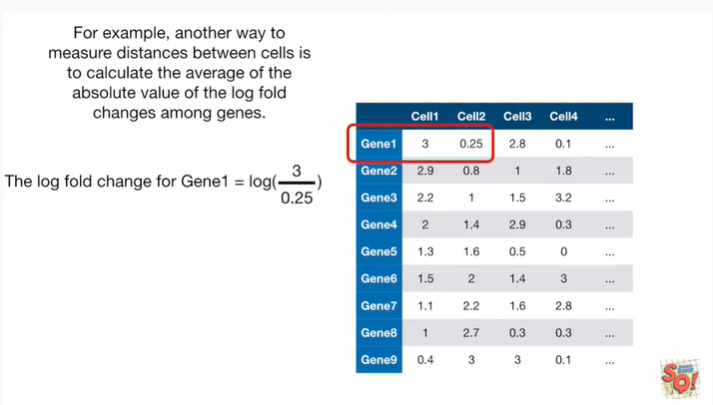


The good news is that there are tons of other ways to measure distance, we don't have to use the Euclidean distance.

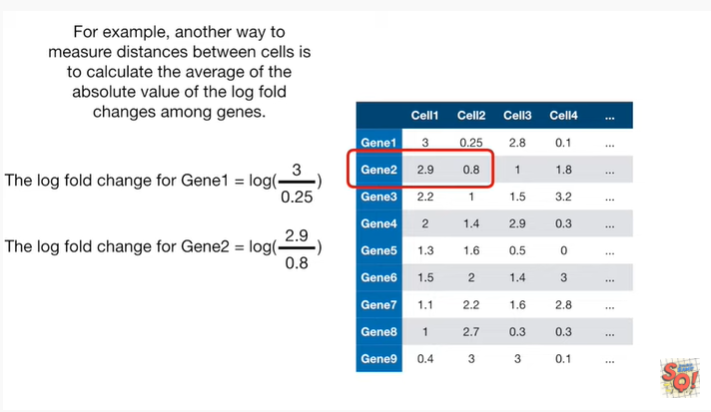
Although sometimes people choose to use it anyways.



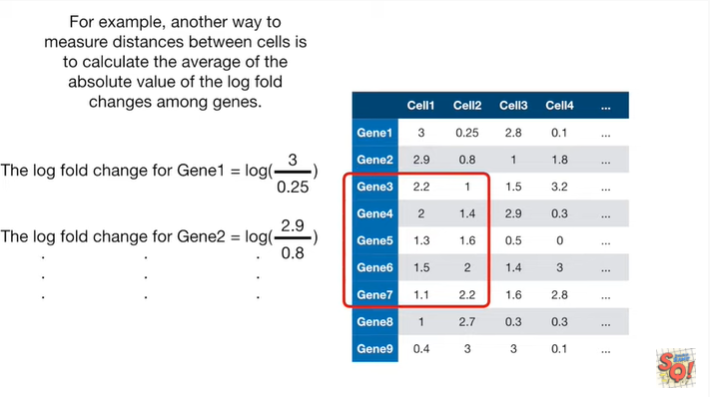
For example, another way to measure distances between cells is to calculate the average of the absolute value of the log fold changes among the genes.



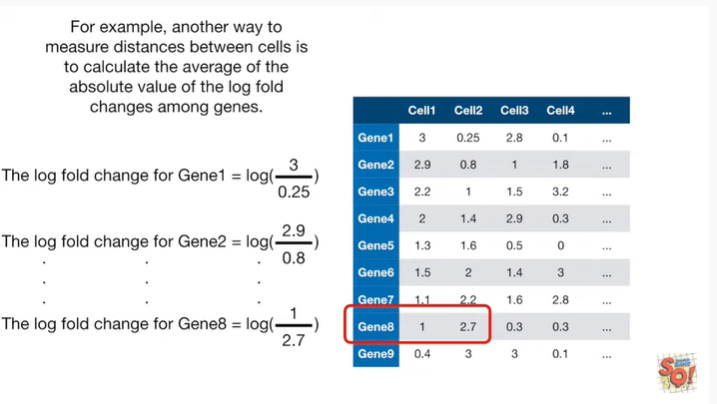
Using the data from cells 1 & 2, the log fold change for gene 1 is the log of 3 divided by 0.25.



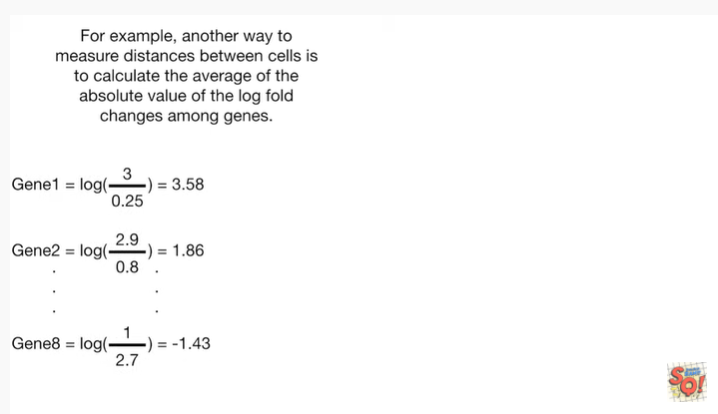
And the log fold change for gene 2 is the log of 2.9 divided by 0.8.



And we just keep going calculating log fold changes for each gene for cells 1 & 2.

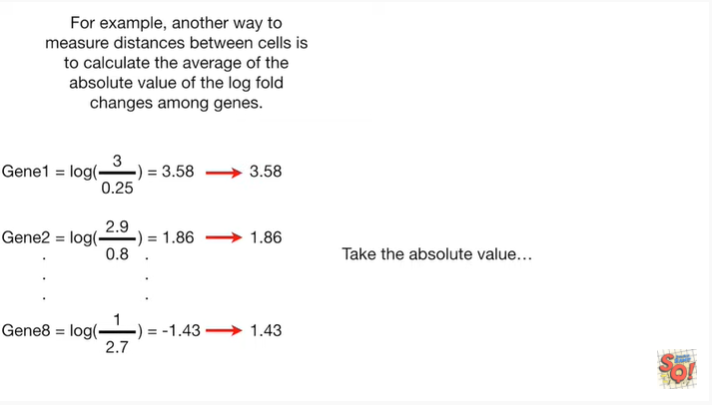


For example the log fold change for gene 8 equals the log of 1 divided by 2 point 7.



Here are the actual log two values of the ratios I've just talked about.

We've got three point five eight, one point eight six, dot dot dot and then negative one point four three.

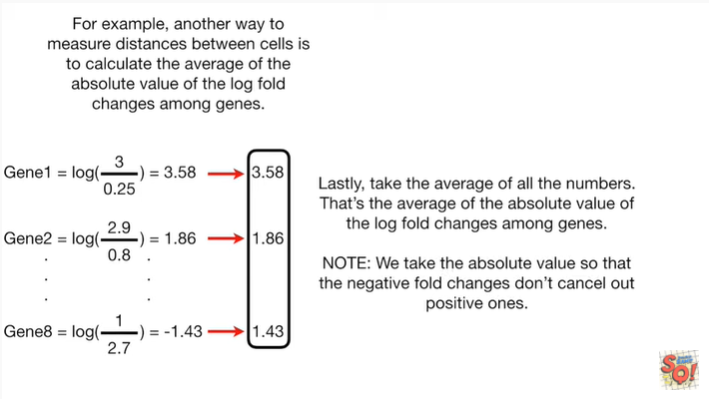


Now we take the absolute value.

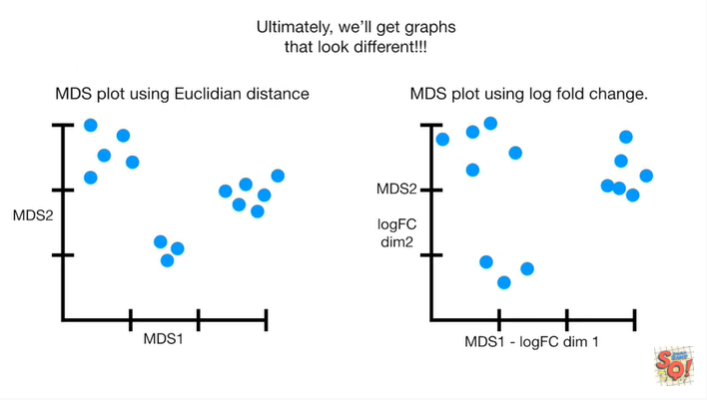


Lastly, we take the average of all the numbers.

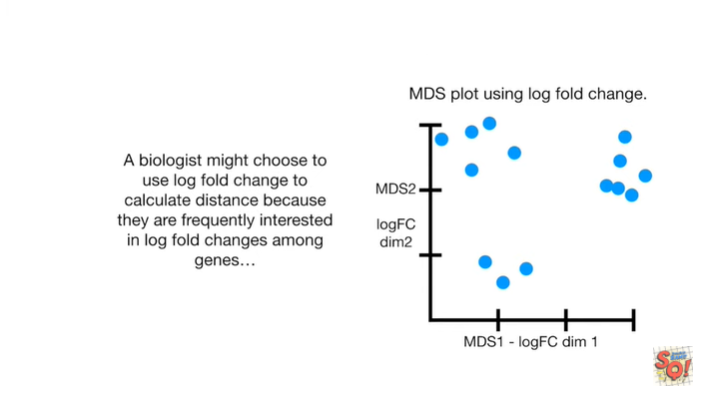
That's the average of the absolute value of the log fold change among the genes.



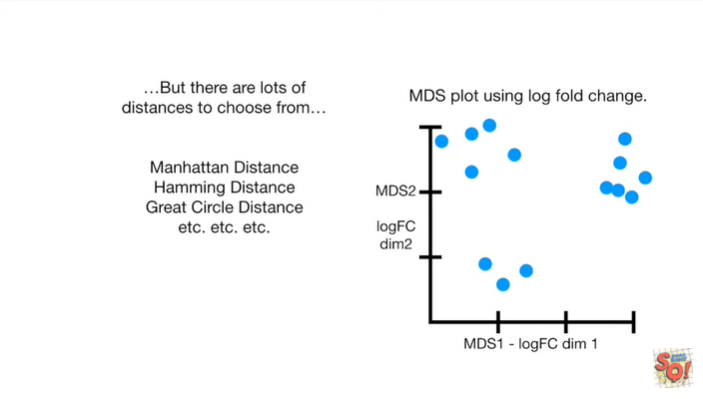
Note : we take the absolute value so that the negative fold changes don't cancel out the positive ones.



Ultimately, we’ll get graphs that look different.

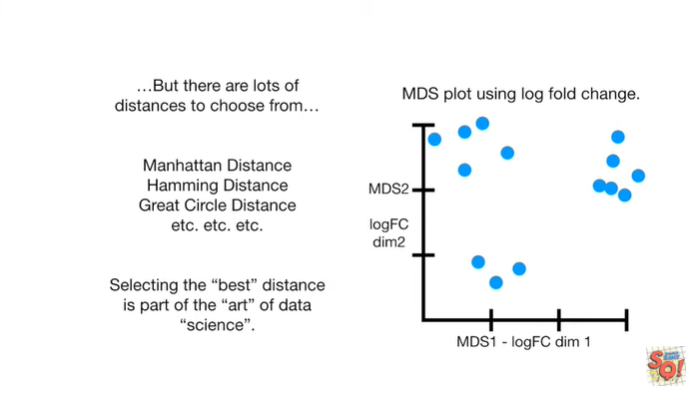


A biologist might choose to use log fold change to calculate distance, because they are frequently interested in log fold changes among genes.

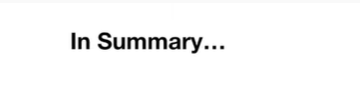


But there are lots of distances to choose from : the Manhattan distance, Hamming distance, Great Circle distance, etc etc etc.

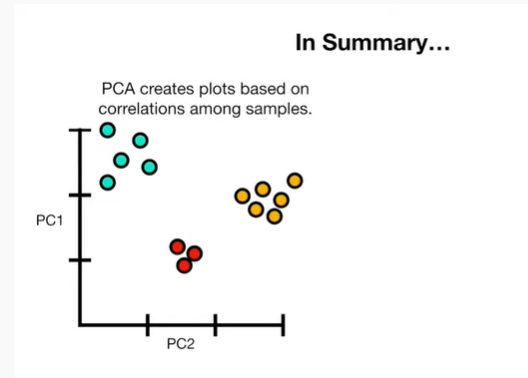
You can look them up on the web.



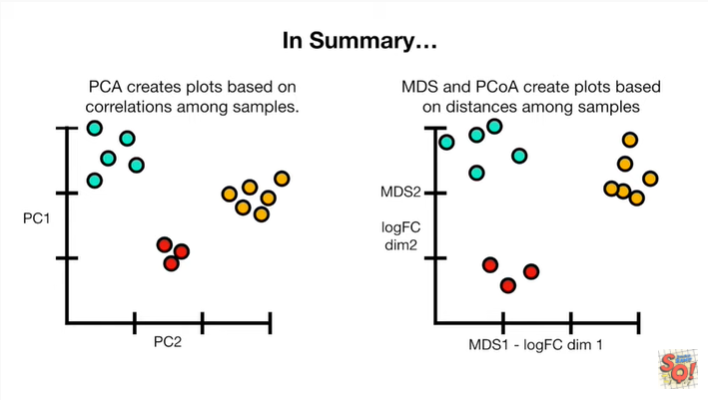
Selecting the best distance is part of the art of data science.



In summary :

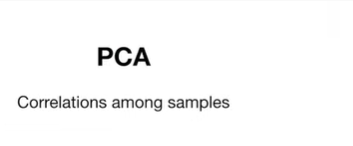


PCA creates plots based on correlations among samples.

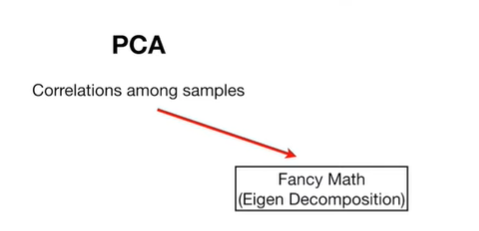


And MDS and principle coordinate analysis create plots based on distances among samples.

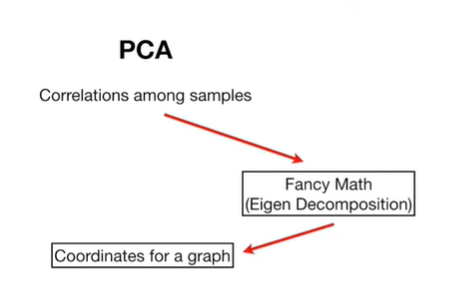
The closer samples are to each other, the more tightly they cluster in the final plot.



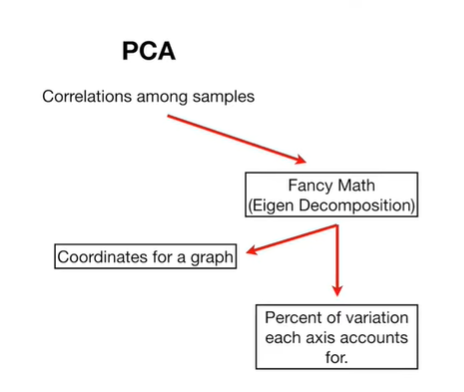
In other words PCA starts by calculating the correlations among the samples



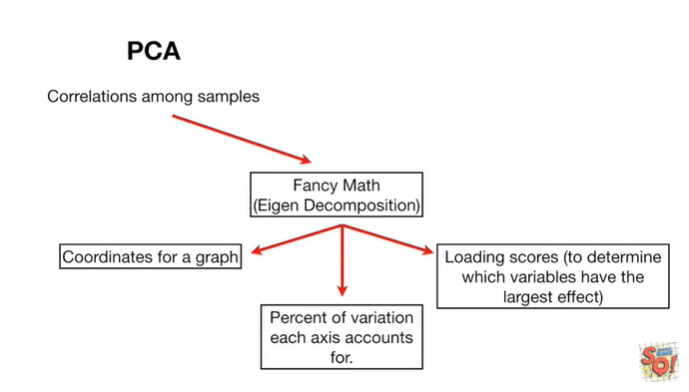
then there's some fancy math specifically eigen decomposition



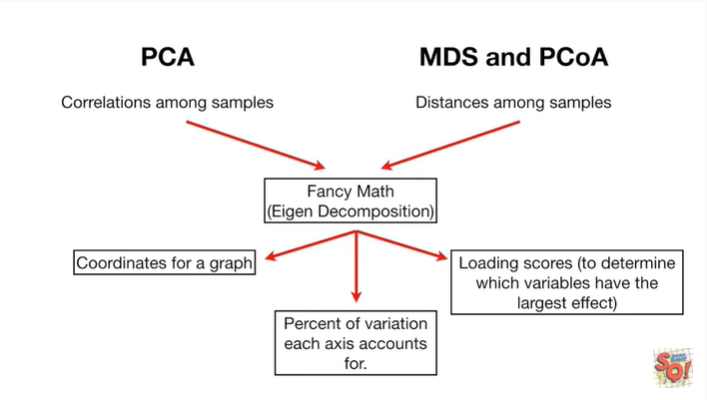
and out of that we get coordinates for a graph



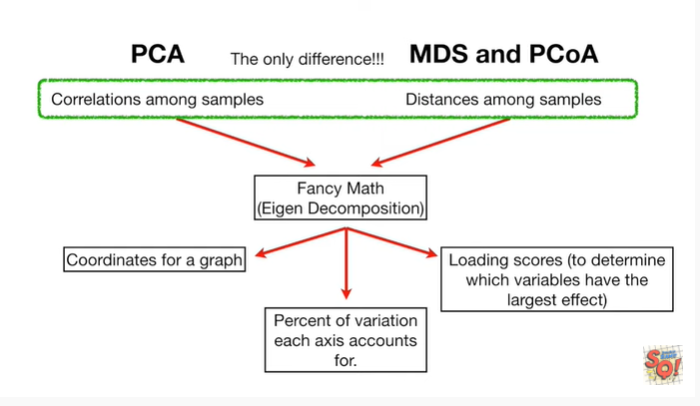
We get the percent of variation each access accounts for.



And we get loading scores to determine which variables have the largest effect.



In contrast MDS and principle coordinate analysis start by calculating distances among the samples.



However that's the only difference.

The fancy math is the exact same and the output is similar.

You get coordinates for a graph the percent of variation that each access accounts for.

And loading scores to determine which variables have the largest effect.