StatQuest…

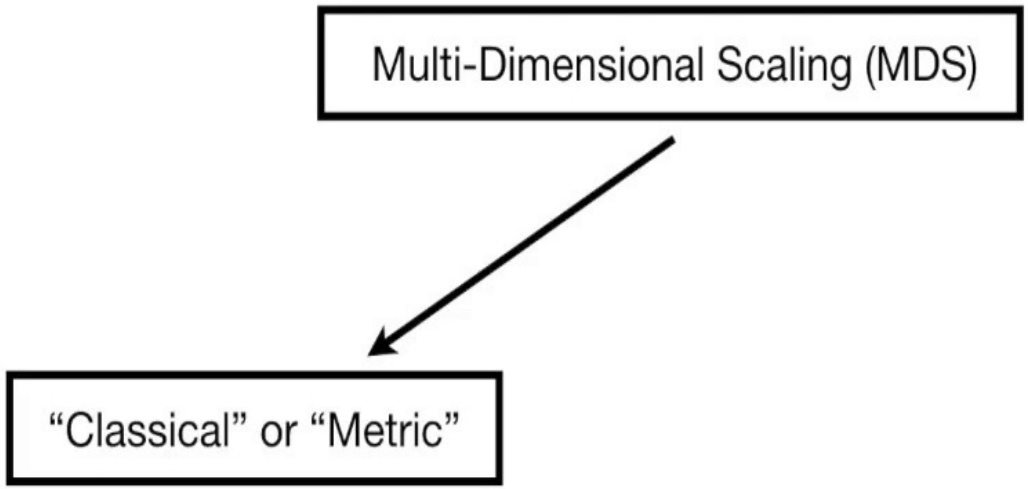
StatQuest:
Multidimensional Scaling (MDS)
and
Principal Coordinate Analysis
(PCoA)

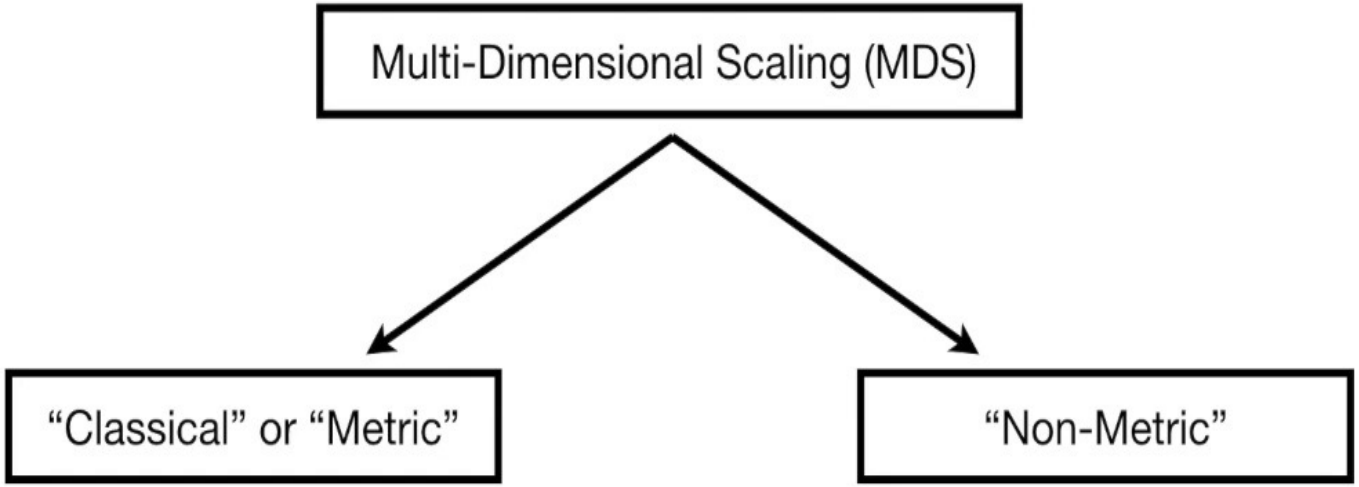
StatQuest:
Multidimensional Scaling (MDS)
and
Principal Coordinate Analysis
(PCoA)

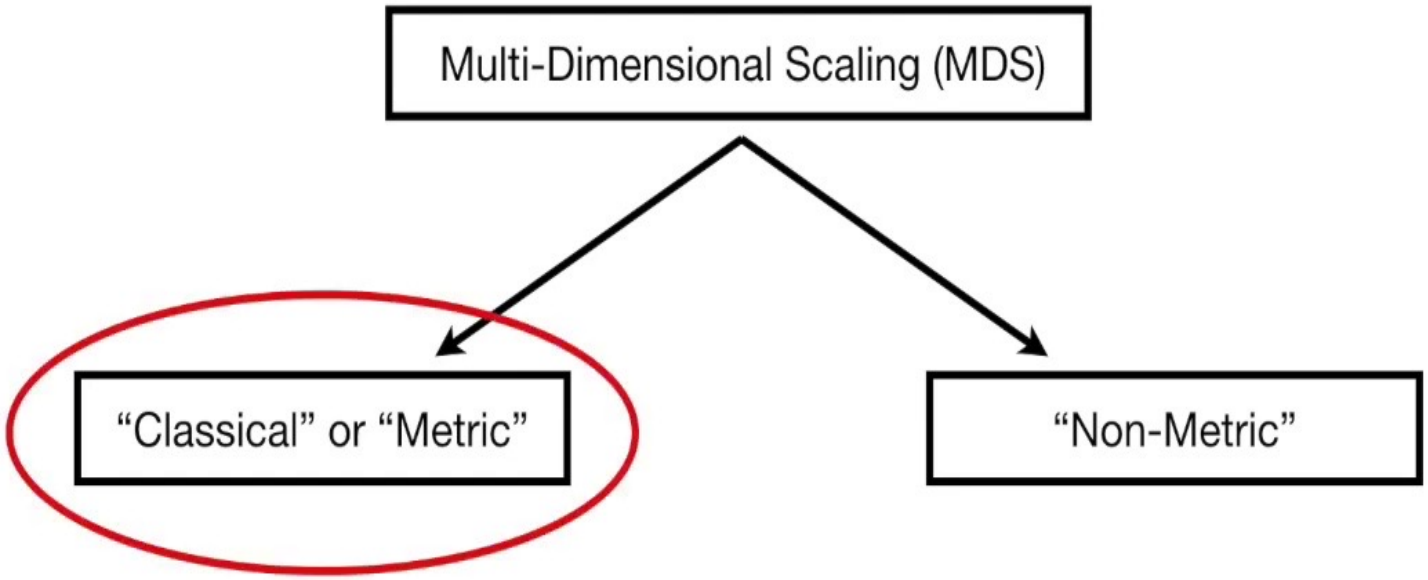
If you don't have PCA down cold,
check out the

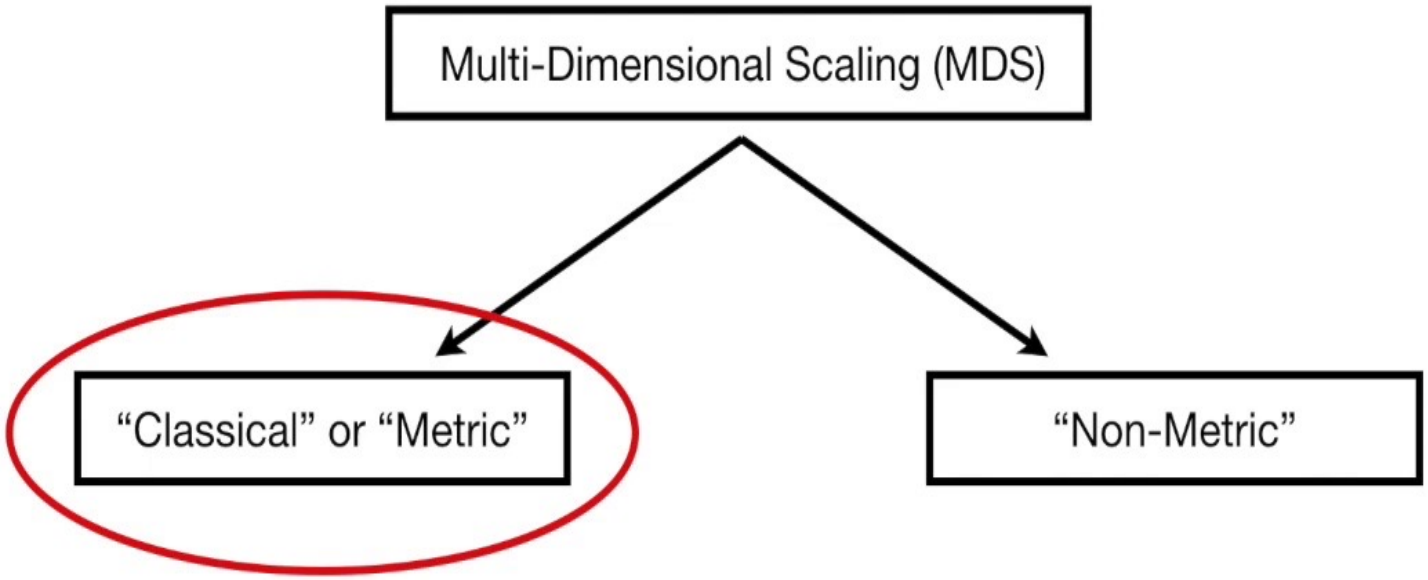
StatQuest:
PCA main ideas in 5 minutes

Multi-Dimensional Scaling (MDS)

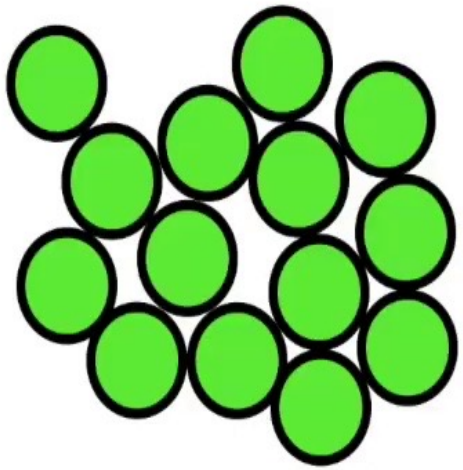

Multi-Dimensional Scaling (MDS)
"Classical" or "Metric"


Multi-Dimensional Scaling (MDS)
“Classical” or “Metric”
“Non-Metric”

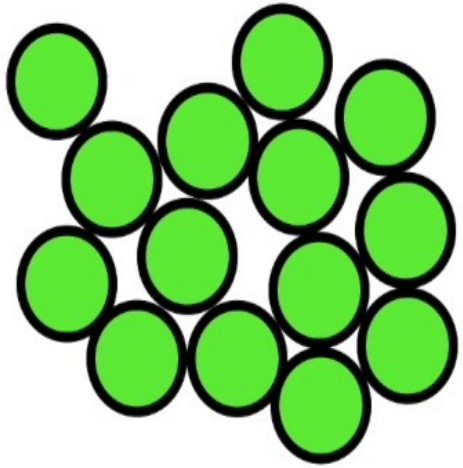

Multi-Dimensional Scaling (MDS)
“Classical” or “Metric”
“Non-Metric”



Classical MDS is also called Principal Coordinate Analysis (PCoA)

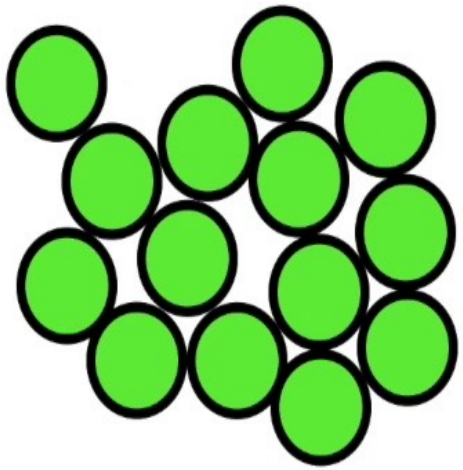


Let's say we had some
normal cells...



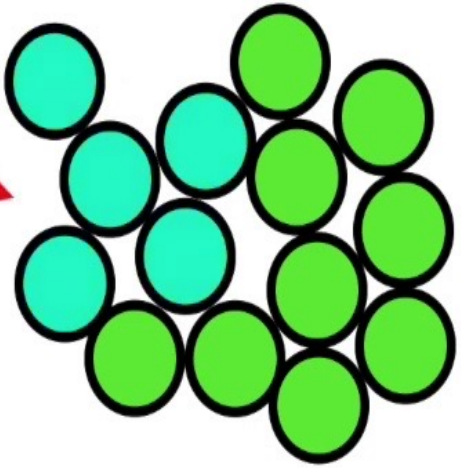
Let's say we had some
normal cells...

(psst - if you're not a biologist, these could be people, or cars, or cities, etc…)

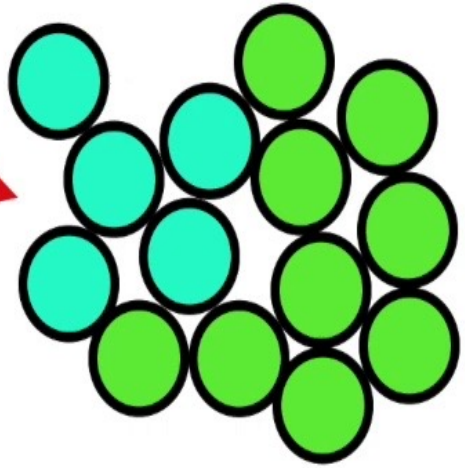


Even though they look the same, we suspect that there are differences...

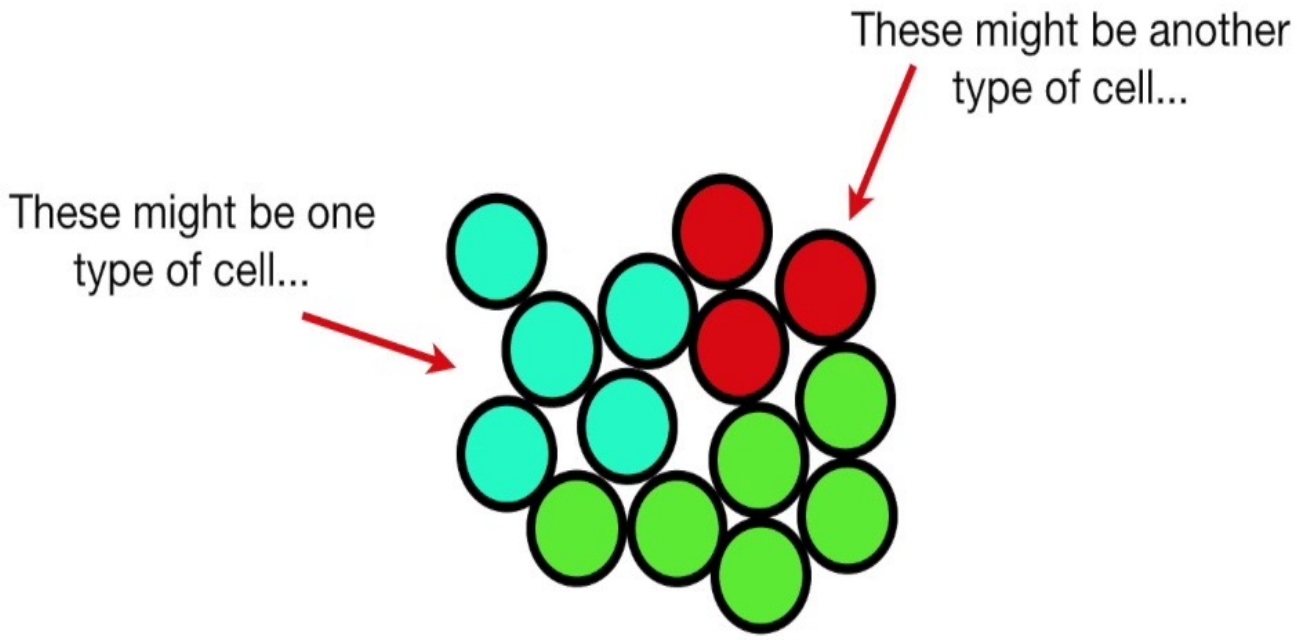
These might be one type of cell...

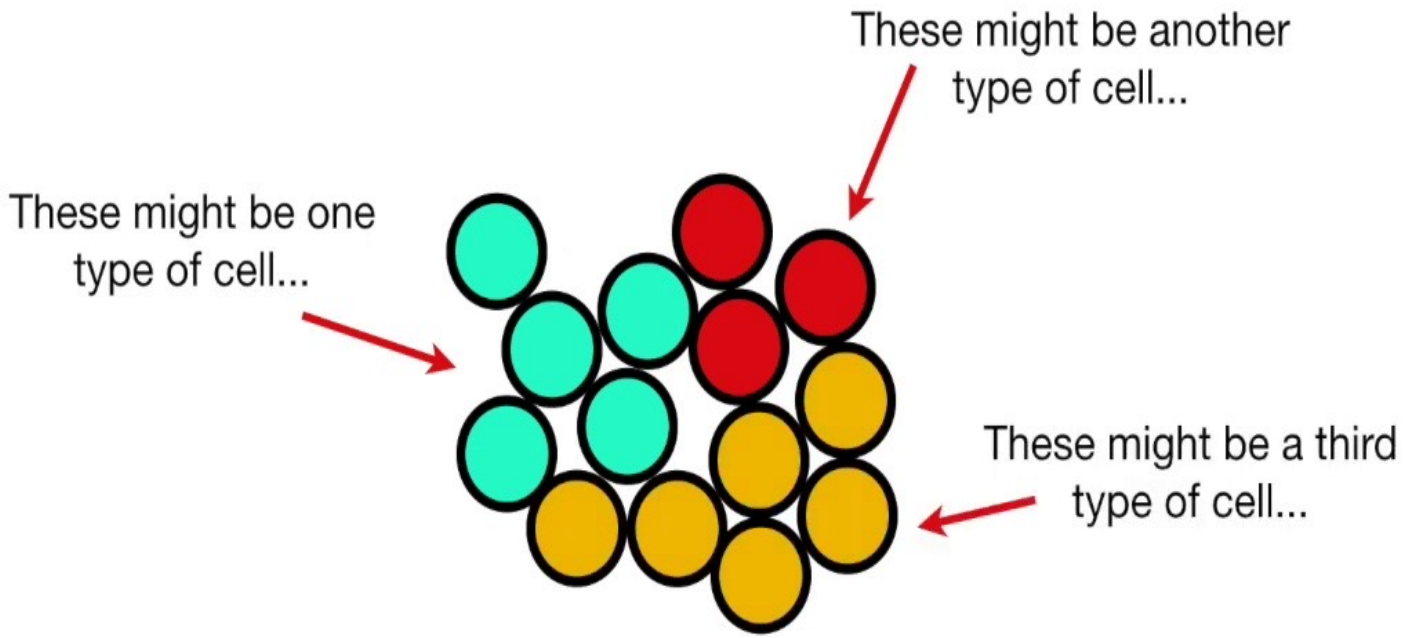


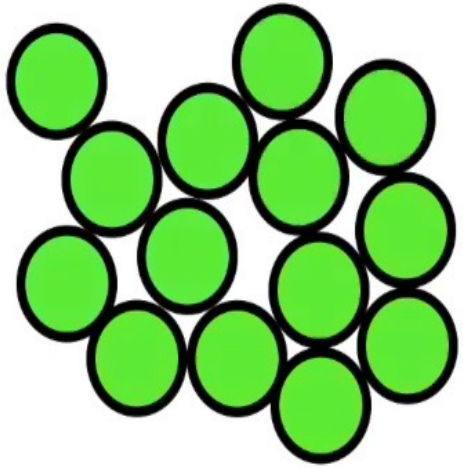
These might be one type of cell...



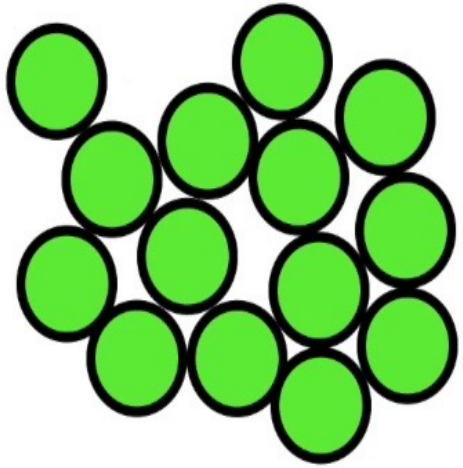
(or one type of person, or car, or city, etc…)


These might be another type of cell...
These might be one type of cell...


These might be another type of cell...
These might be one type of cell...
These might be a third type of cell...



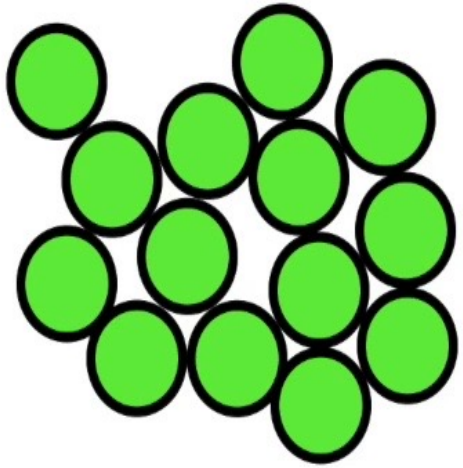
Unfortunately, we can't observe the differences from the outside…



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

…so we sequence the mRNA in each cell to identify which genes are active. This tells us what each cell is doing.

(If they were people, we could measure their height, blood pressure, reading level etc…)



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

…so we sequence the mRNA in each cell to identify which genes are active. This tells us what each cell is doing.

Here's the data…

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

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

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

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

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

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

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

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

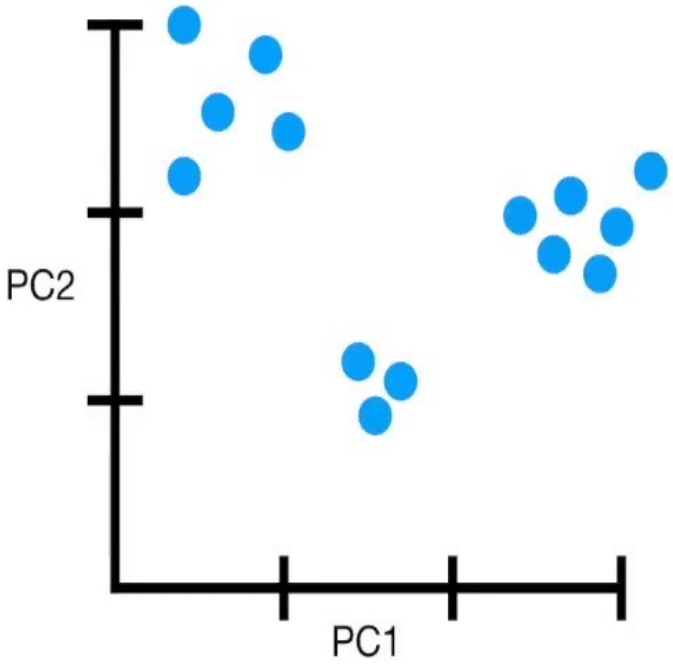
	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

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

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

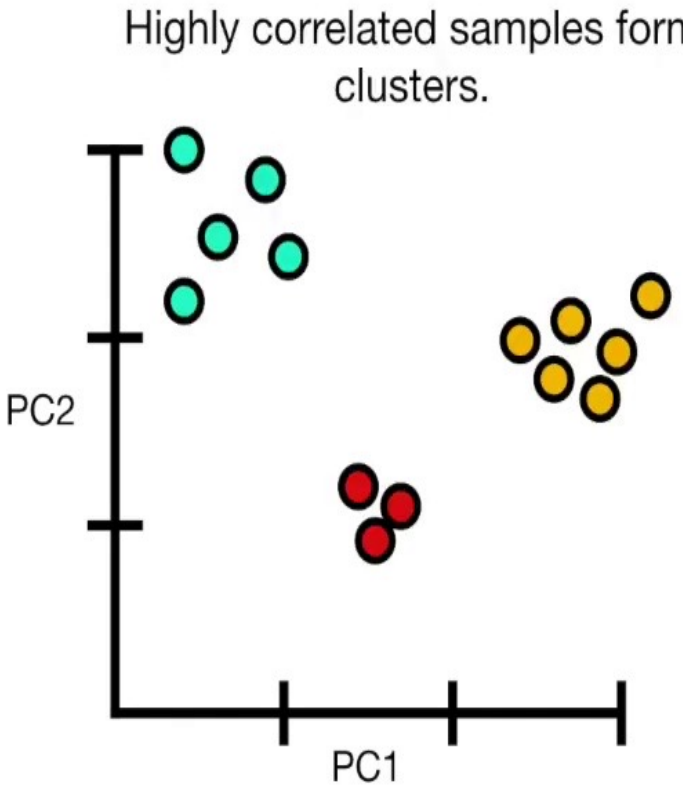
etc…

When we did PCA, we converted correlations (or lack there of) among the samples into a 2-Dimensional plot.



	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

etc…



	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

etc…

Multi-dimensional scaling (MDS) and Principal Coordinate Analysis (PCoA) are very similar to PCA, except that **instead of converting correlations** into a 2-D graph, **they convert distances among the samples** into a 2-D graph.

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

etc…

So, in order to do MDS or PCoA, we have to calculate the distance between Cell1 and Cell2…

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

etc…

…and the distance between
Cell1 and Cell3…

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

etc…

…and the distance between
Cell1 and Cell4…

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

etc…

…and the distance between
Cell2 and Cell3…

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

etc…

…and the distance between
Cell2 and Cell4… etc…

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	2.2	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2	0.3	...
Gene5	1.3	1.6	1.6	0	...
Gene6	1.5	2	2.1	3	...
Gene7	1.1	2.2	1.2	2.8	...
Gene8	1	2.7	0.9	0.3	...
Gene9	0.4	3	0.6	0.1	...

etc…

For now, let's imagine we only needed to calculate the distance between Cell1 and Cell2...

	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8
Gene3	2.2	1
Gene4	2	1.4
Gene5	1.3	1.6
Gene6	1.5	2
Gene7	1.1	2.2
Gene8	1	2.7
Gene9	0.4	3

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

	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8
Gene3	2.2	1
Gene4	2	1.4
Gene5	1.3	1.6
Gene6	1.5	2
Gene7	1.1	2.2
Gene8	1	2.7
Gene9	0.4	3

If we just had two genes…

	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8

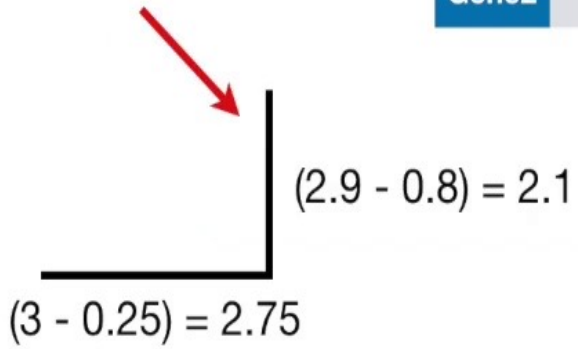
…we could draw a line that represented the difference between the values for Gene1…

(3 - 0.25) = 2.75

	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8

…and we could draw a line that represented the difference between the values for Gene2…

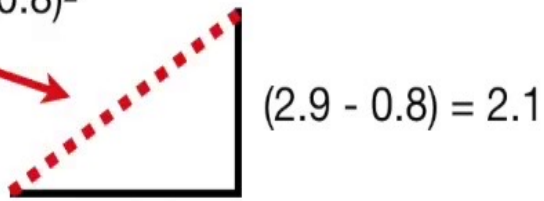
	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8



…Then the Euclidian distance would be the hypotenuse (i.e. the Pythagorean theorem.)

	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8

$$\sqrt{(3 - 0.25)^2 + (2.9 - 0.8)^2}$$



$(2.9 - 0.8) = 2.1$

$(3 - 0.25) = 2.75$

With more genes, we just add the square of more differences between more genes…

$$\sqrt{(3 - 0.25)^2 + (2.9 - 0.8)^2 + (2.9 - 0.8)^2 \ldots}$$

	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8
Gene3	2.2	1
Gene4	2	1.4
Gene5	1.3	1.6
Gene6	1.5	2
Gene7	1.1	2.2
Gene8	1	2.7
Gene9	0.4	3

With more genes, we just add the square of more differences between more genes…

$$\sqrt{(3 - 0.25)^2 + (2.9 - 0.8)^2 + (2.9 - 0.8)^2 \ldots}$$

The difference for Gene1

	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8
Gene3	2.2	1
Gene4	2	1.4
Gene5	1.3	1.6
Gene6	1.5	2
Gene7	1.1	2.2
Gene8	1	2.7
Gene9	0.4	3

With more genes, we just add the square of more differences between more genes…

$$\sqrt{(3 - 0.25)^2 + (2.9 - 0.8)^2 + (2.9 - 0.8)^2 \ldots}$$

The difference for Gene2

	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8
Gene3	2.2	1
Gene4	2	1.4
Gene5	1.3	1.6
Gene6	1.5	2
Gene7	1.1	2.2
Gene8	1	2.7
Gene9	0.4	3

With more genes, we just add the square of more differences between more genes…

$$\sqrt{(3 - 0.25)^2 + (2.9 - 0.8)^2 + (2.9 - 0.8)^2 \ldots}$$

The difference for Gene3

	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8
Gene3	2.2	1
Gene4	2	1.4
Gene5	1.3	1.6
Gene6	1.5	2
Gene7	1.1	2.2
Gene8	1	2.7
Gene9	0.4	3

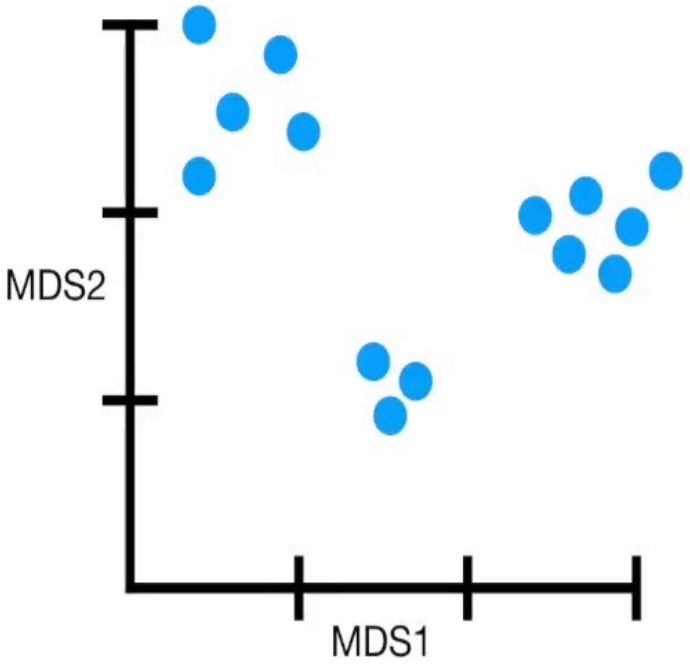
With more genes, we just add the square of more differences between more genes…

$$\sqrt{(3 - 0.25)^2 + (2.9 - 0.8)^2 + (2.9 - 0.8)^2 \ldots}$$

etc. etc. etc…

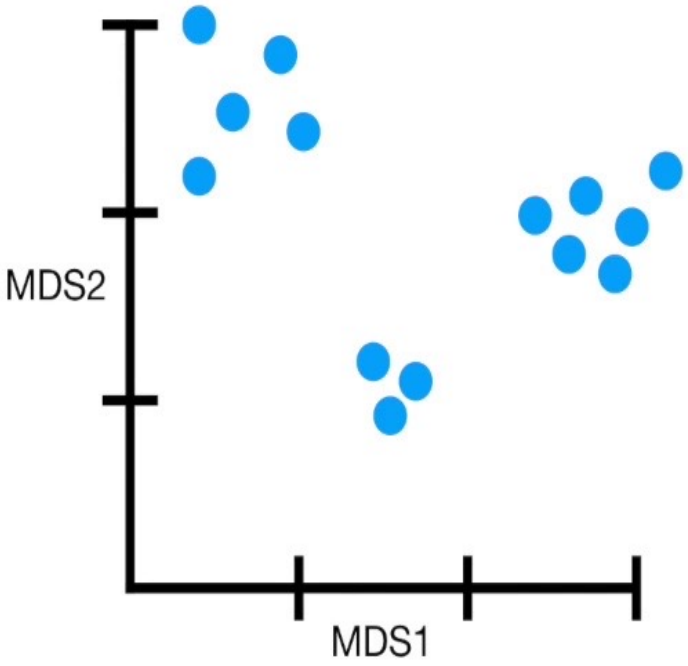
	Cell1	Cell2
Gene1	3	0.25
Gene2	2.9	0.8
Gene3	2.2	1
Gene4	2	1.4
Gene5	1.3	1.6
Gene6	1.5	2
Gene7	1.1	2.2
Gene8	1	2.7
Gene9	0.4	3

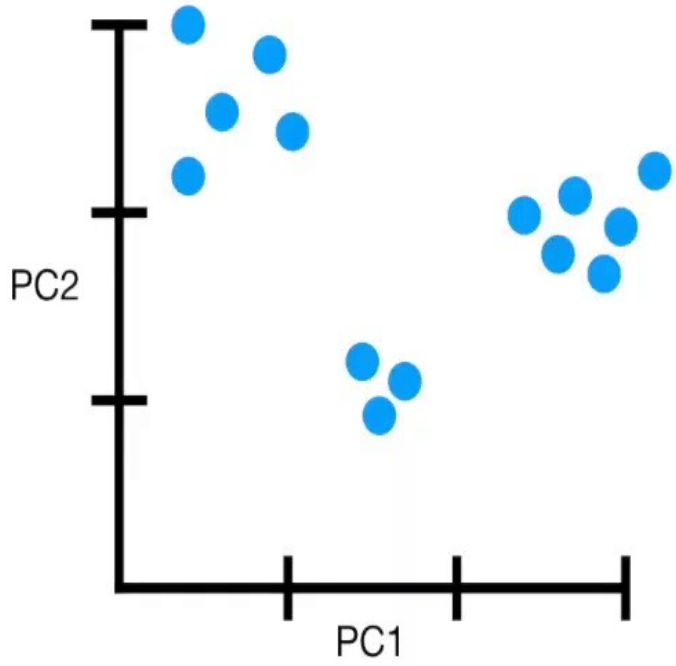
And once we calculated the distance between every pair of cells, MDS and PCoA would reduce them to a 2-D graph.



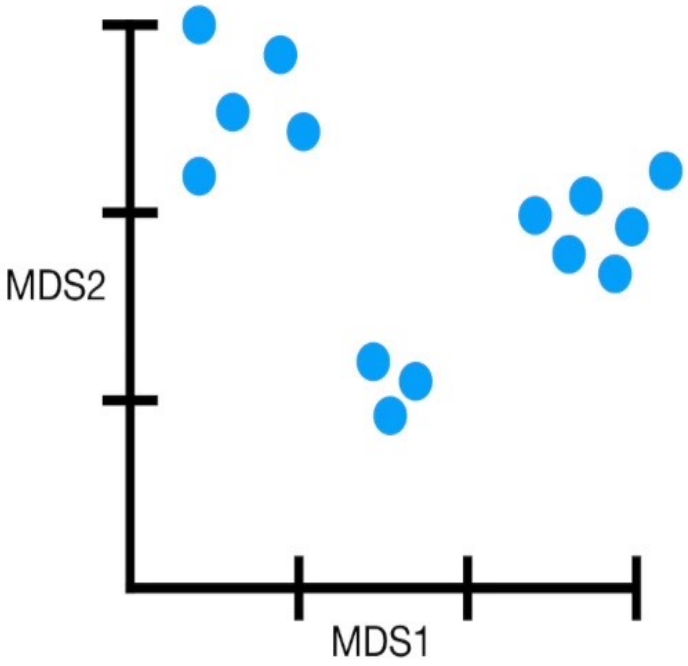
	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	1	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2.9	0.3	...
Gene5	1.3	1.6	0.5	0	...
Gene6	1.5	2	1.4	3	...
Gene7	1.1	2.2	1.6	2.8	...
Gene8	1	2.7	0.3	0.3	...
Gene9	0.4	3	3	0.1	...

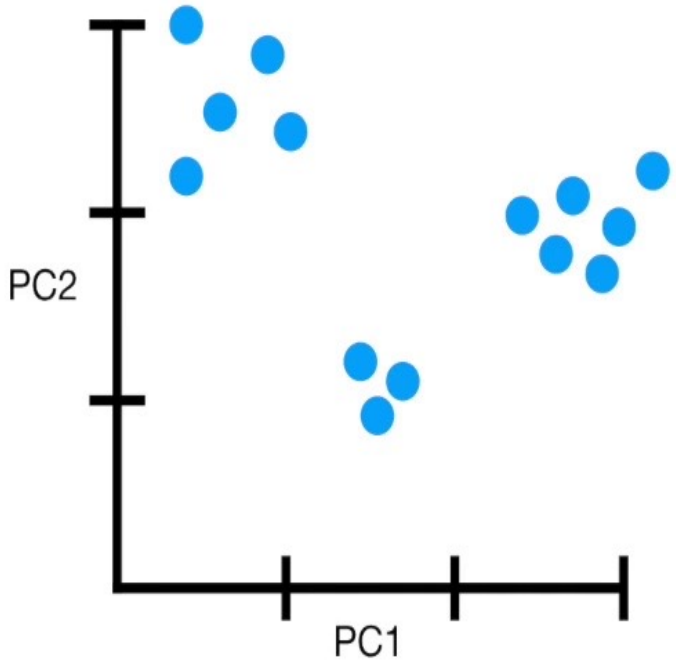
The bad news is that if we used the Euclidean Distance, the graph would be identical to a PCA graph!!!

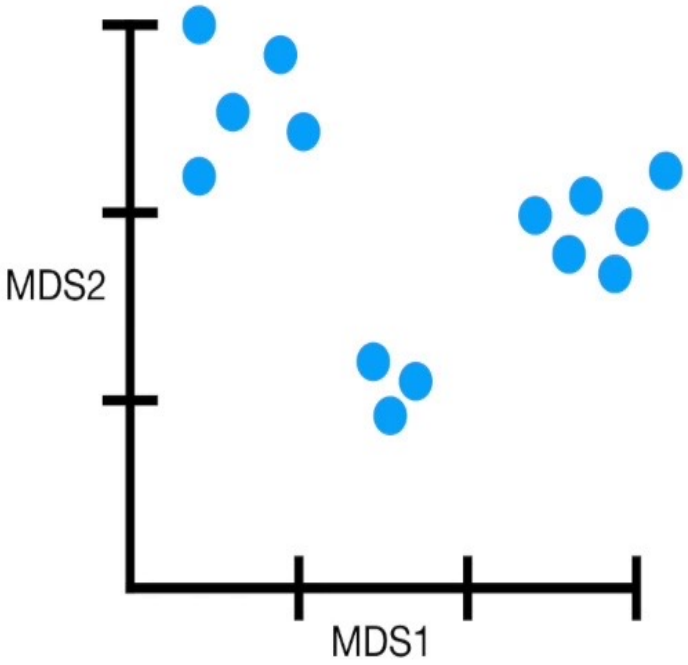




In other words, clustering based on **minimizing the linear distances is the same maximizing the linear correlations**.







The good news is that there are tons of other ways to measure distance!!!!

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

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	1	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2.9	0.3	...
Gene5	1.3	1.6	0.5	0	...
Gene6	1.5	2	1.4	3	...
Gene7	1.1	2.2	1.6	2.8	...
Gene8	1	2.7	0.3	0.3	...
Gene9	0.4	3	3	0.1	...

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

The log fold change for Gene1 = $\log(\frac{3}{0.25})$

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	1	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2.9	0.3	...
Gene5	1.3	1.6	0.5	0	...
Gene6	1.5	2	1.4	3	...
Gene7	1.1	2.2	1.6	2.8	...
Gene8	1	2.7	0.3	0.3	...
Gene9	0.4	3	3	0.1	...

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

The log fold change for Gene1 = $\log(\frac{3}{0.25})$

The log fold change for Gene2 = $\log(\frac{2.9}{0.8})$

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	1	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2.9	0.3	...
Gene5	1.3	1.6	0.5	0	...
Gene6	1.5	2	1.4	3	...
Gene7	1.1	2.2	1.6	2.8	...
Gene8	1	2.7	0.3	0.3	...
Gene9	0.4	3	3	0.1	...

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

$$\text{The log fold change for Gene1} = \log\left(\frac{3}{0.25}\right)$$

$$\text{The log fold change for Gene2} = \log\left(\frac{2.9}{0.8}\right)$$

. . .
. . .
. . .

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	1	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2.9	0.3	...
Gene5	1.3	1.6	0.5	0	...
Gene6	1.5	2	1.4	3	...
Gene7	1.1	2.2	1.6	2.8	...
Gene8	1	2.7	0.3	0.3	...
Gene9	0.4	3	3	0.1	...

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

$$\text{The log fold change for Gene1} = \log\left(\frac{3}{0.25}\right)$$

$$\text{The log fold change for Gene2} = \log\left(\frac{2.9}{0.8}\right)$$

. . .

. . .

. . .

$$\text{The log fold change for Gene8} = \log\left(\frac{1}{2.7}\right)$$

	Cell1	Cell2	Cell3	Cell4	...
Gene1	3	0.25	2.8	0.1	...
Gene2	2.9	0.8	1	1.8	...
Gene3	2.2	1	1.5	3.2	...
Gene4	2	1.4	2.9	0.3	...
Gene5	1.3	1.6	0.5	0	...
Gene6	1.5	2	1.4	3	...
Gene7	1.1	2.2	1.6	2.8	...
Gene8	1	2.7	0.3	0.3	...
Gene9	0.4	3	3	0.1	...

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

$$\text{Gene1} = \log\left(\frac{3}{0.25}\right) = 3.58$$

$$\text{Gene2} = \log\left(\frac{2.9}{0.8}\right) = 1.86$$

. . .

$$\text{Gene8} = \log\left(\frac{1}{2.7}\right) = -1.43$$

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

$$\text{Gene1} = \log\left(\frac{3}{0.25}\right) = 3.58 \rightarrow 3.58$$

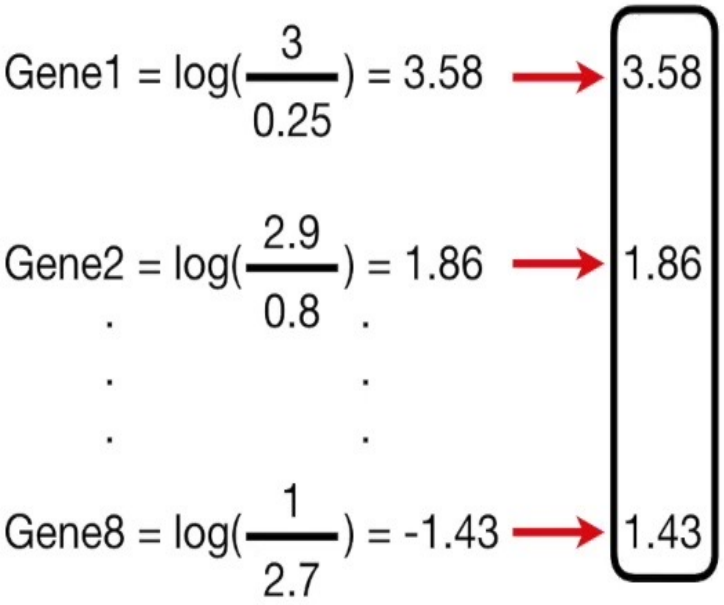
$$\text{Gene2} = \log\left(\frac{2.9}{0.8}\right) = 1.86 \rightarrow 1.86$$

. .
. .
. .

$$\text{Gene8} = \log\left(\frac{1}{2.7}\right) = -1.43 \rightarrow 1.43$$

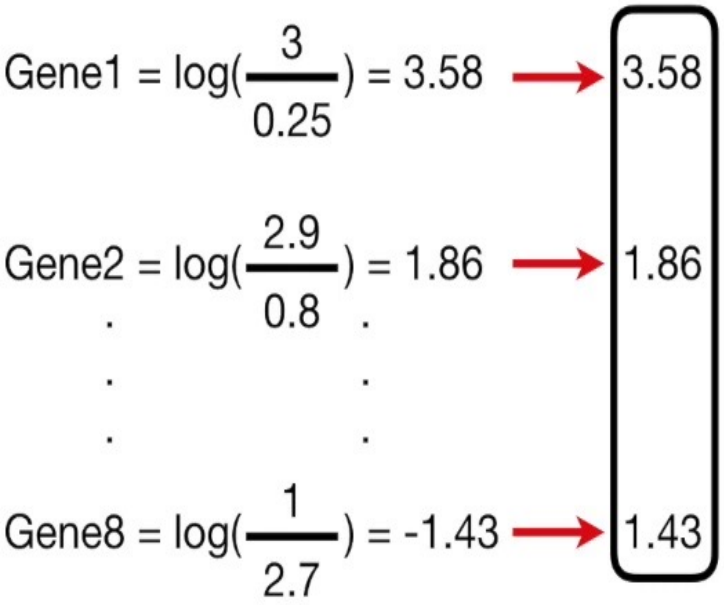
Take the absolute value…

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



Lastly, take the average of all the numbers. That's the average of the absolute value of the log fold changes among genes.

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

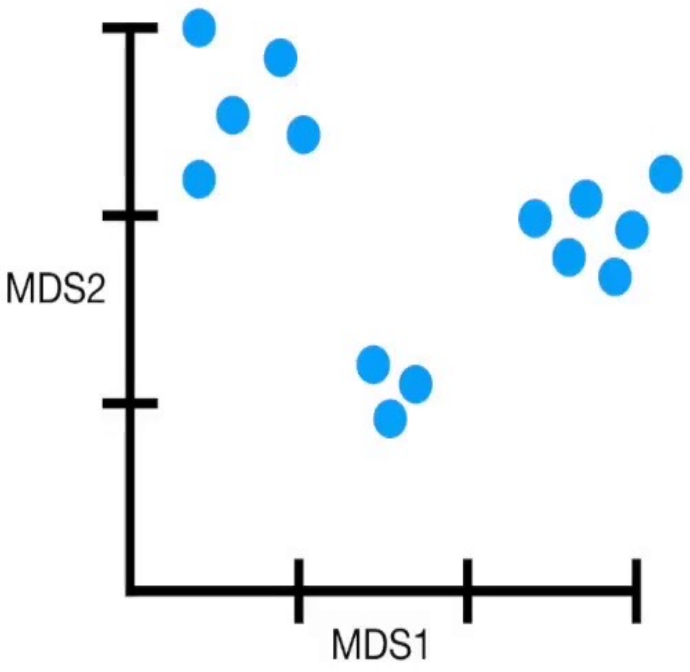


Lastly, take the average of all the numbers. That's the average of the absolute value of the log fold changes among genes.

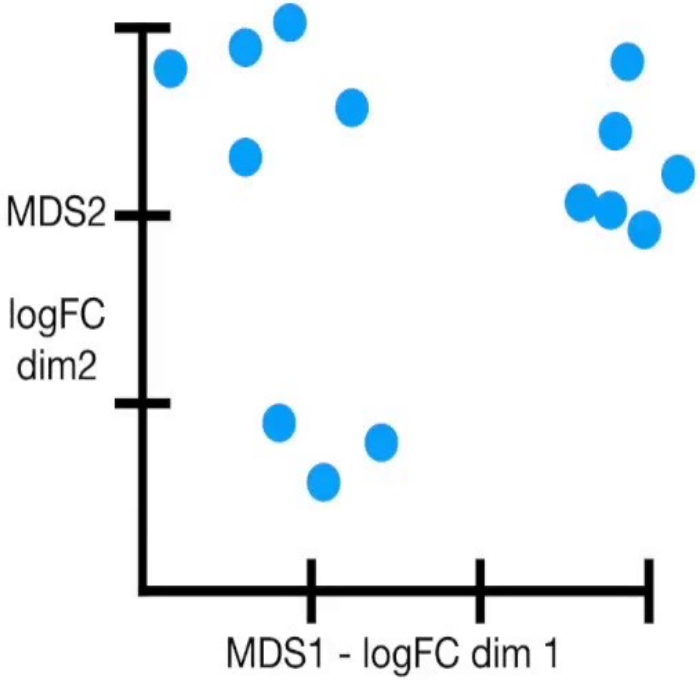
NOTE: We take the absolute value so that the negative fold changes don't cancel out positive ones.

Ultimately, we'll get graphs
that look different!!!

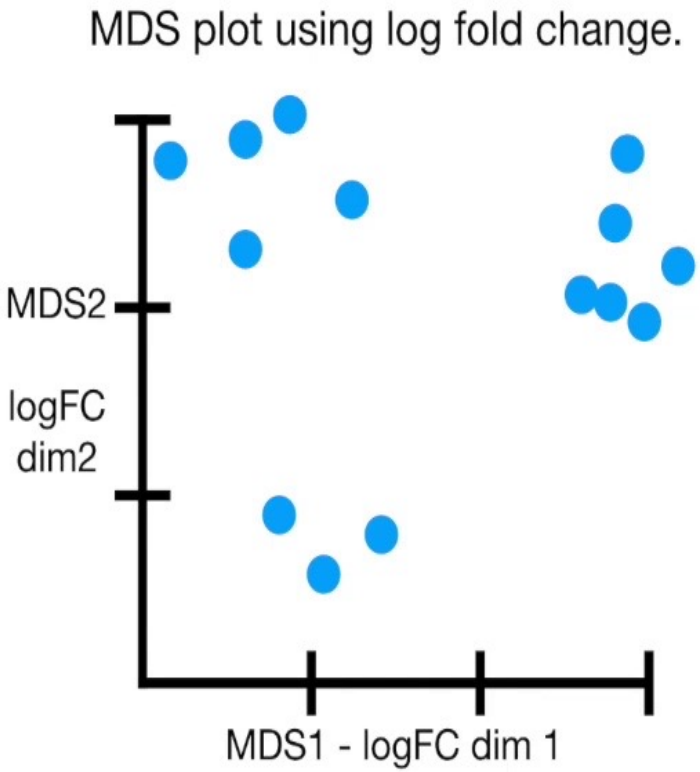
MDS plot using Euclidian distance



MDS plot using log fold change.



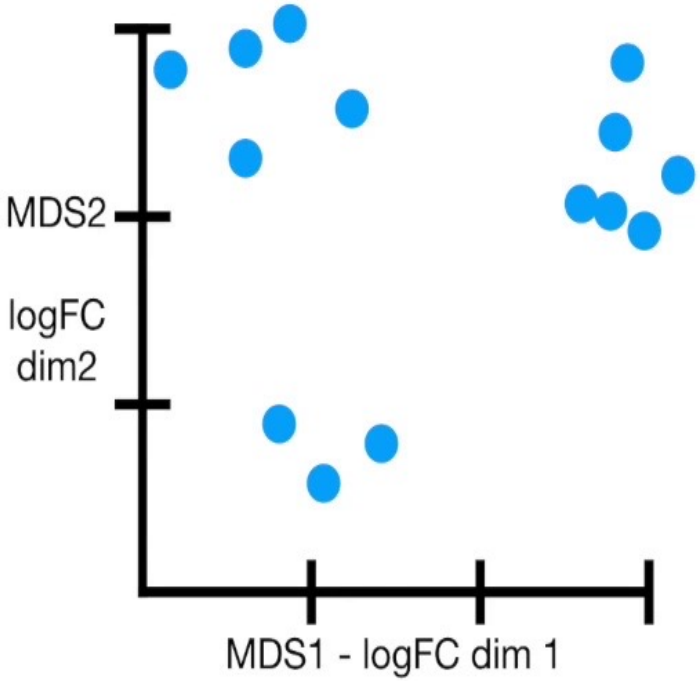
A biologist might choose to use log fold change to calculate distance because they are frequently interested in log fold changes among genes…



…But there are lots of distances to choose from…

Manhattan Distance
Hamming Distance
Great Circle Distance
etc. etc. etc.

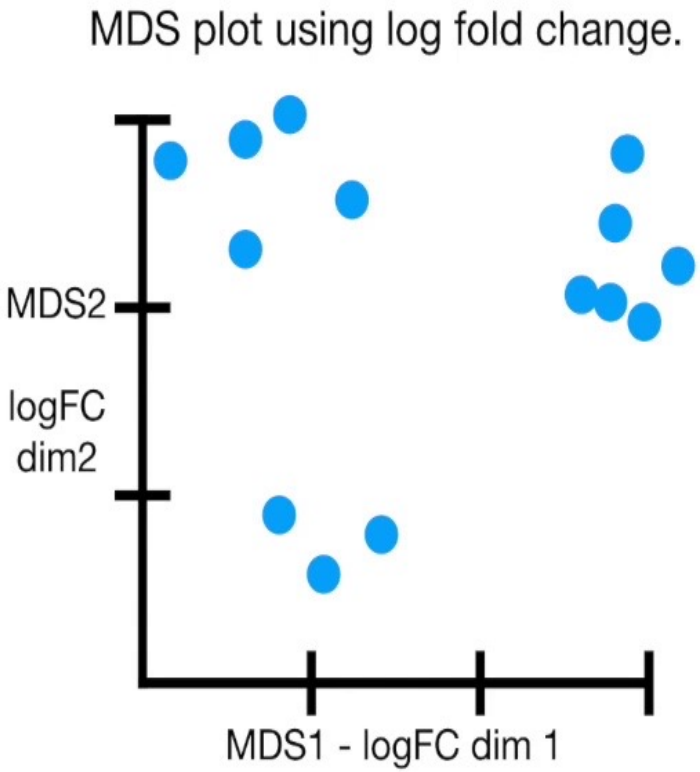
MDS plot using log fold change.



…But there are lots of distances to choose from…

Manhattan Distance
Hamming Distance
Great Circle Distance
etc. etc. etc.

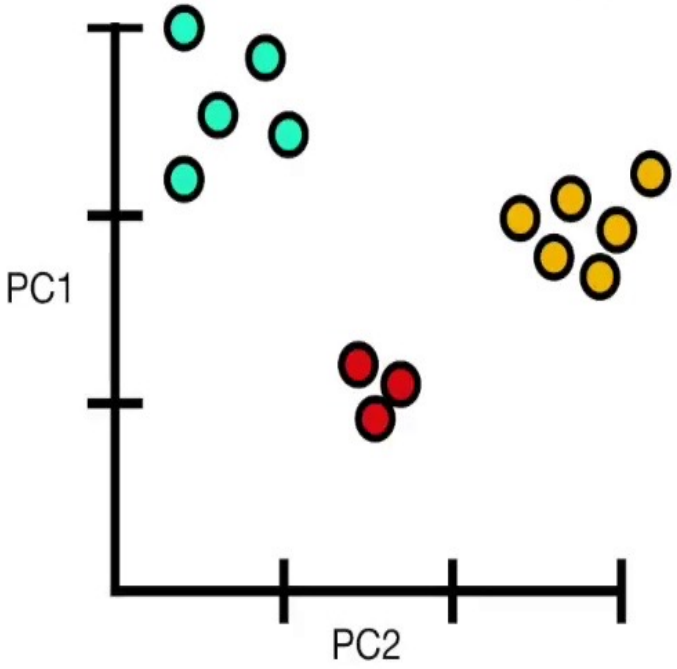
Selecting the “best” distance is part of the “art” of data “science”.



In Summary…

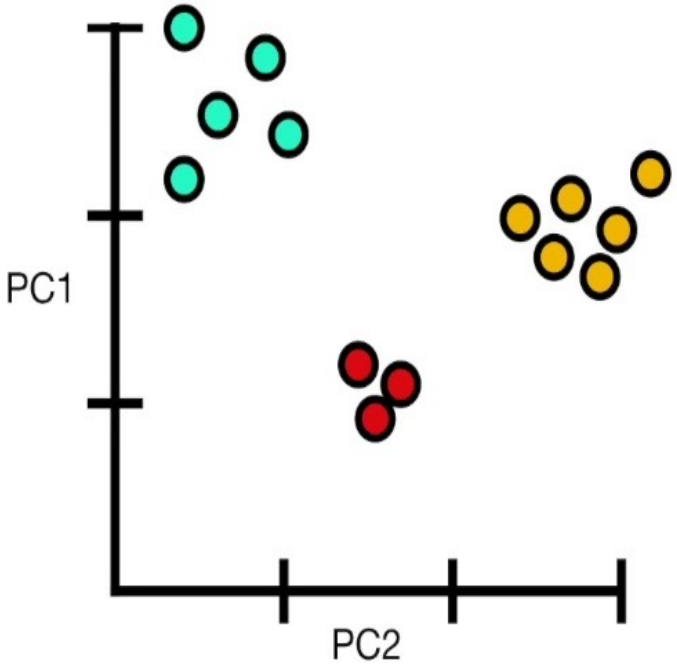
In Summary…

PCA creates plots based on correlations among samples.

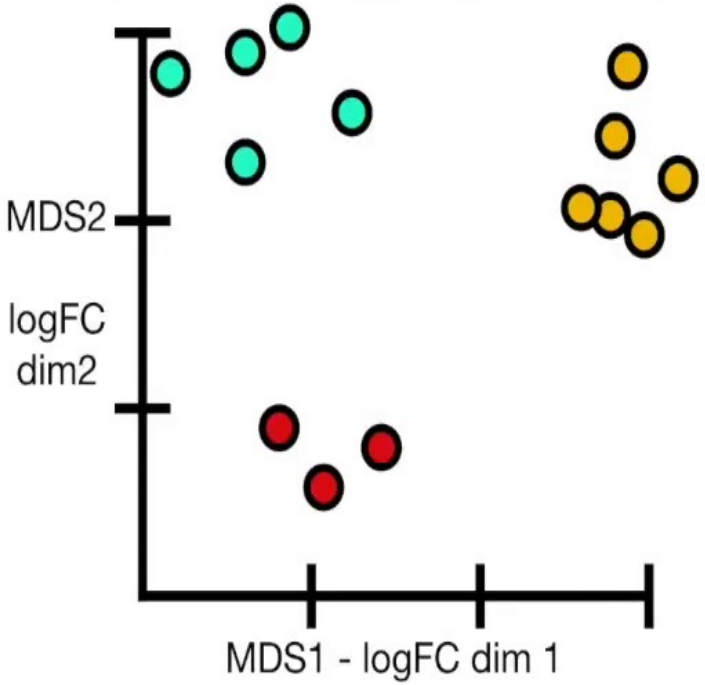


In Summary…

PCA creates plots based on correlations among samples.



MDS and PCoA create plots based on distances among samples

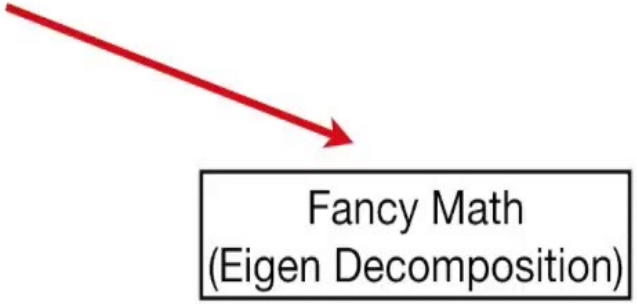


PCA

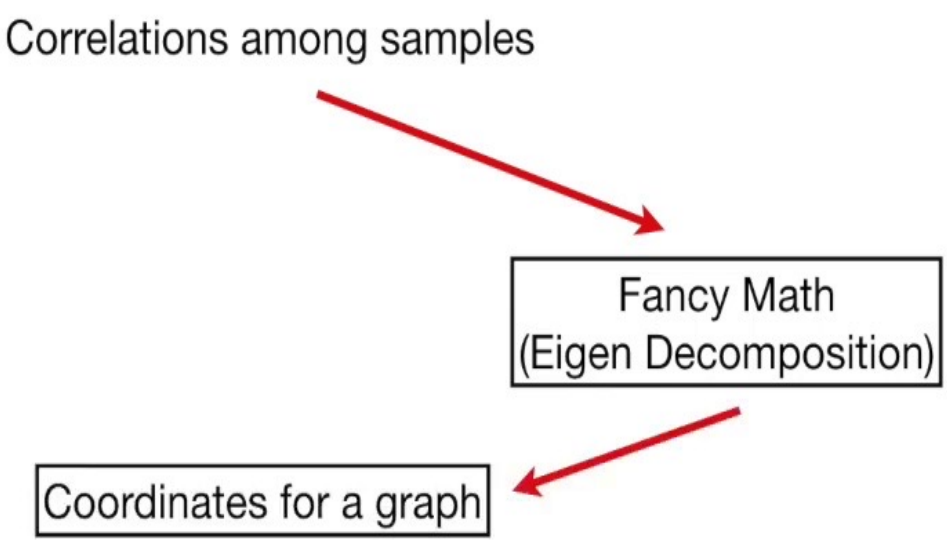
Correlations among samples

PCA

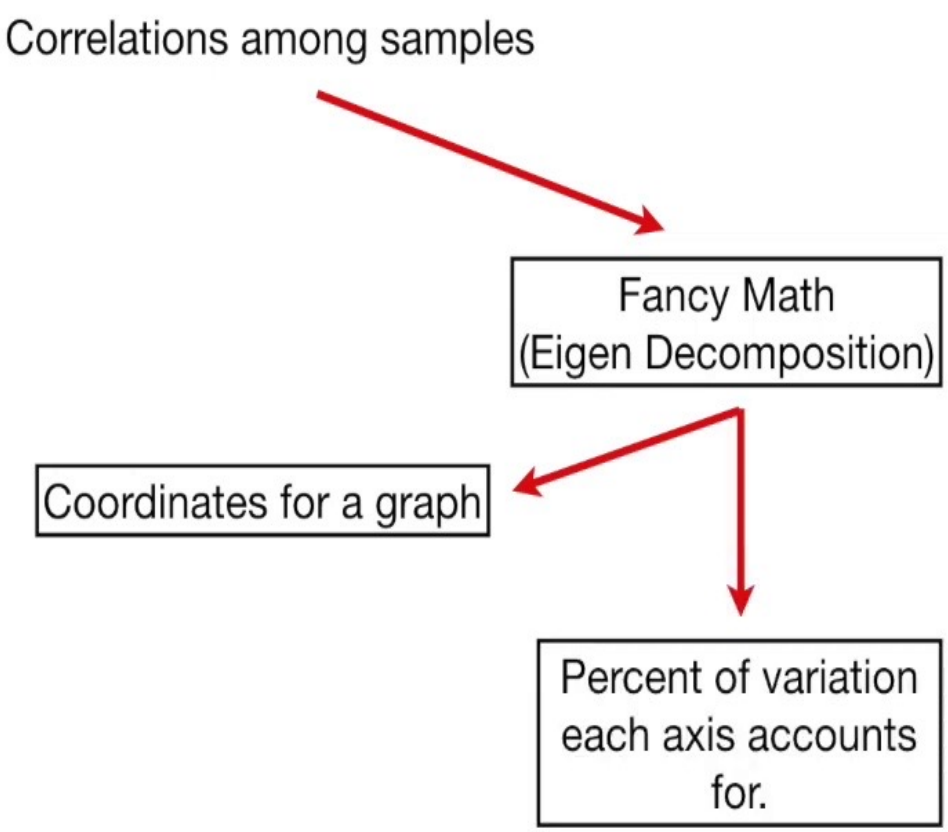
Correlations among samples



PCA

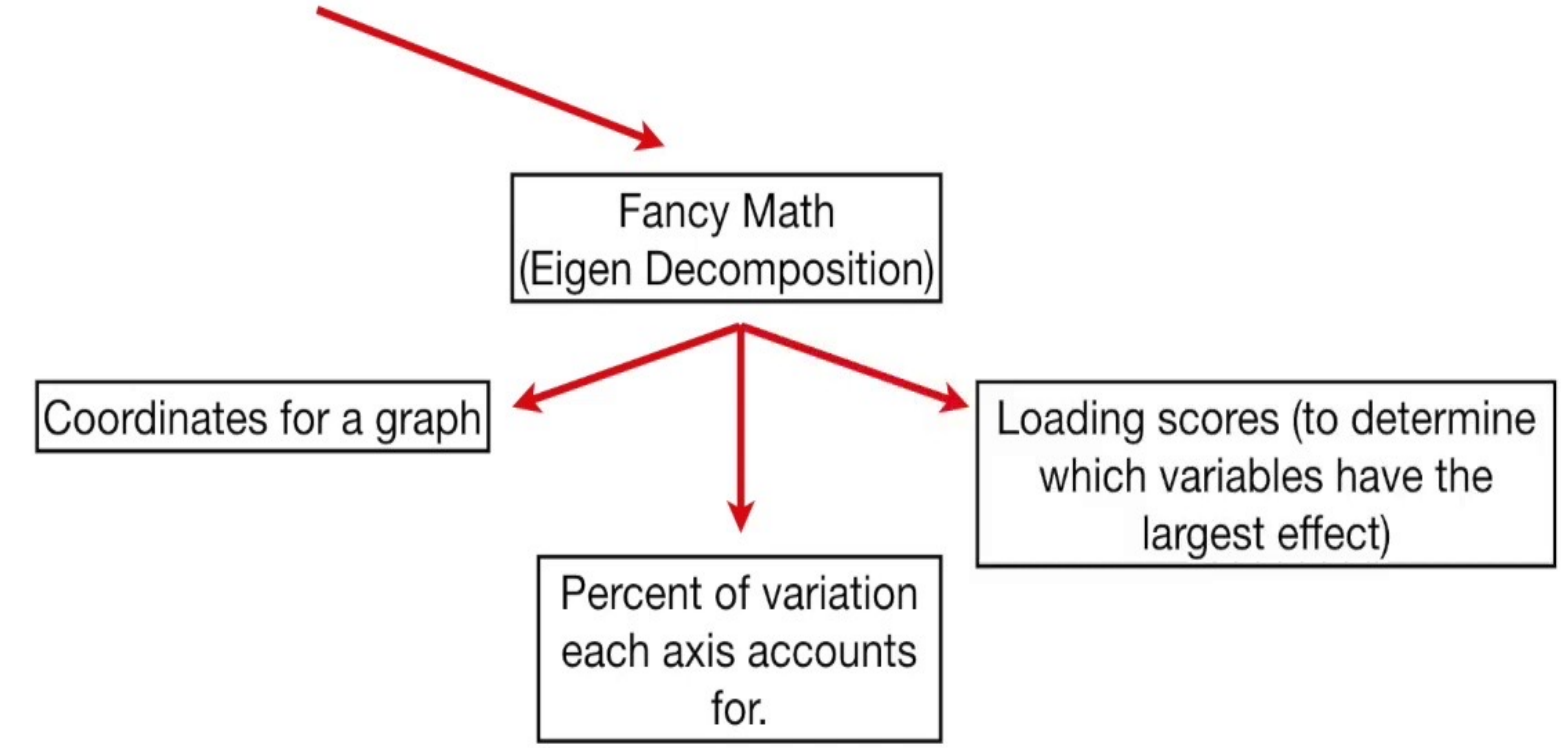

Correlations among samples
Fancy Math
(Eigen Decomposition)
Coordinates for a graph

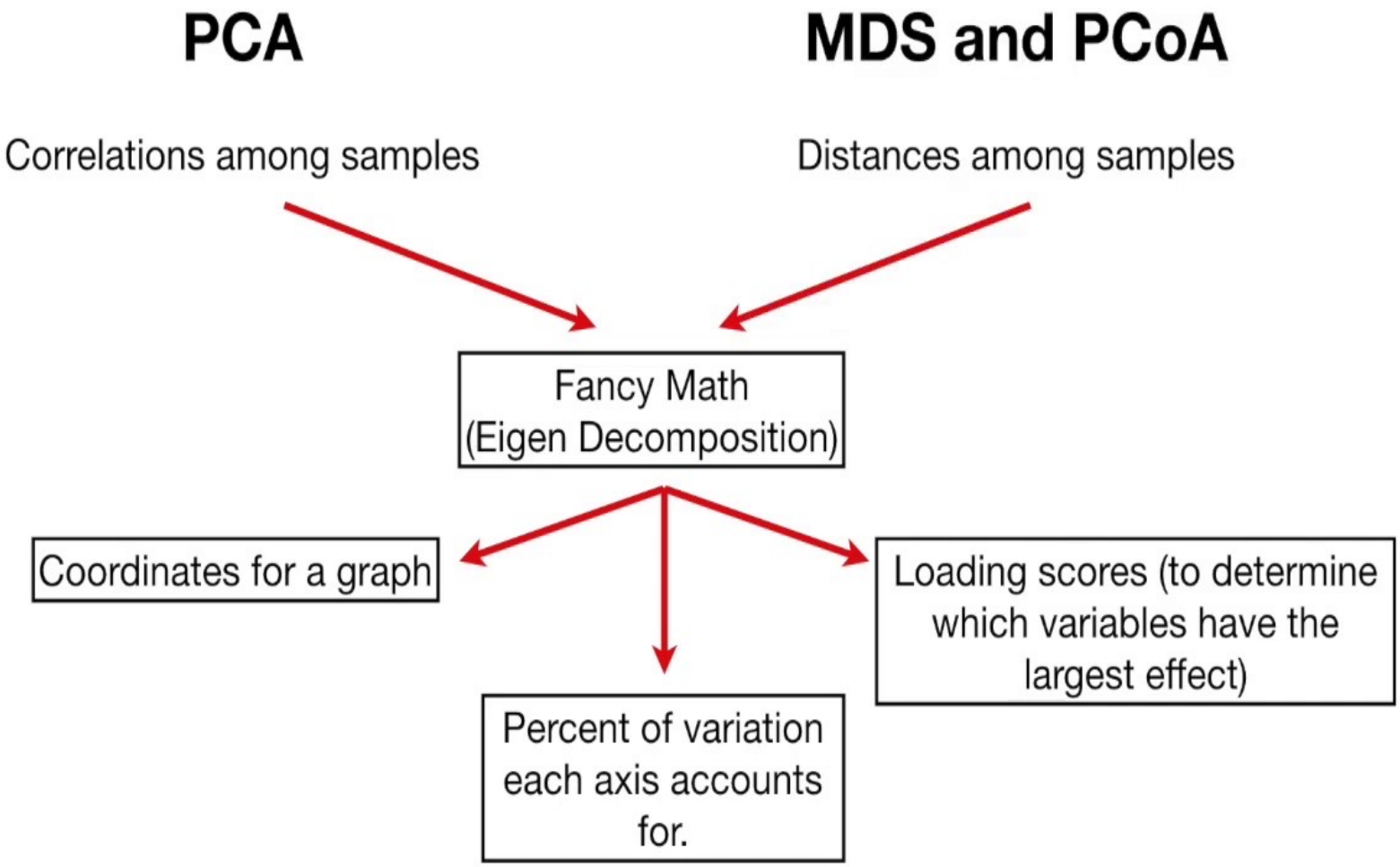
PCA

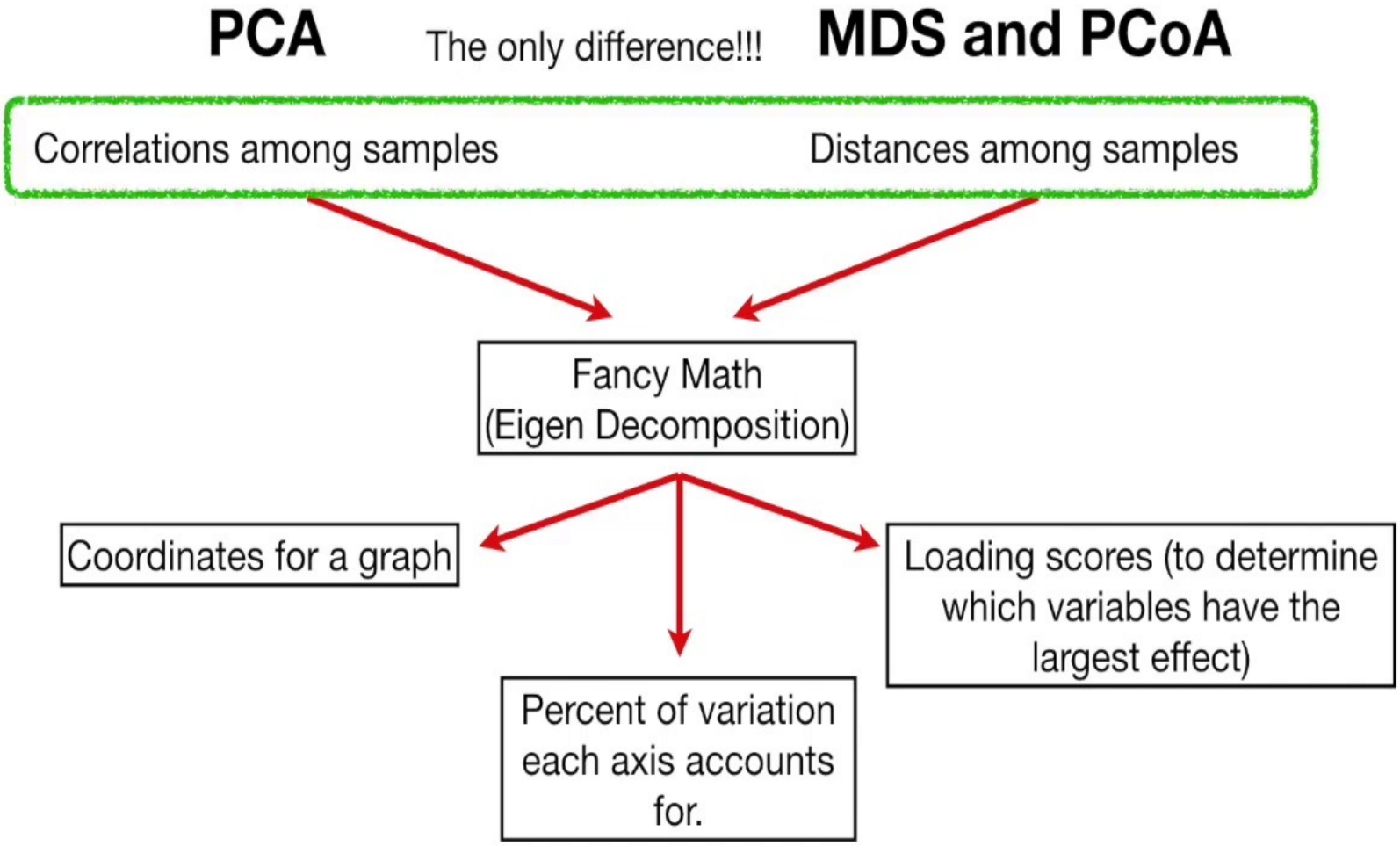

Correlations among samples
Fancy Math
(Eigen Decomposition)
Coordinates for a graph
Percent of variation each axis accounts for.

PCA

Correlations among samples




PCA
MDS and PCoA
Correlations among samples
Distances among samples
Fancy Math
(Eigen Decomposition)
Coordinates for a graph
Loading scores (to determine which variables have the largest effect)
Percent of variation each axis accounts for.


PCA
The only difference!!!
MDS and PCoA
Correlations among samples
Distances among samples
Fancy Math
(Eigen Decomposition)
Coordinates for a graph
Loading scores (to determine which variables have the largest effect)
Percent of variation each axis accounts for.

The End!!!