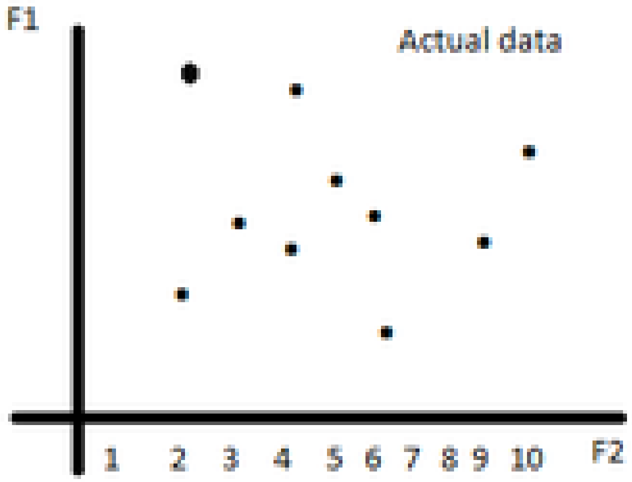
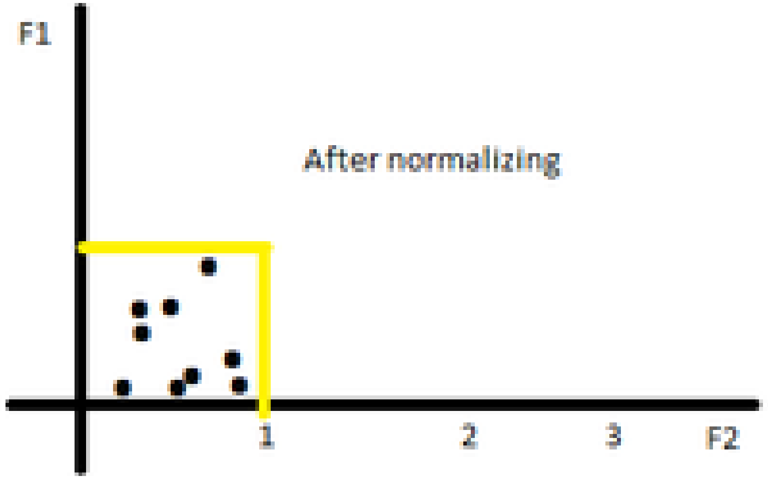
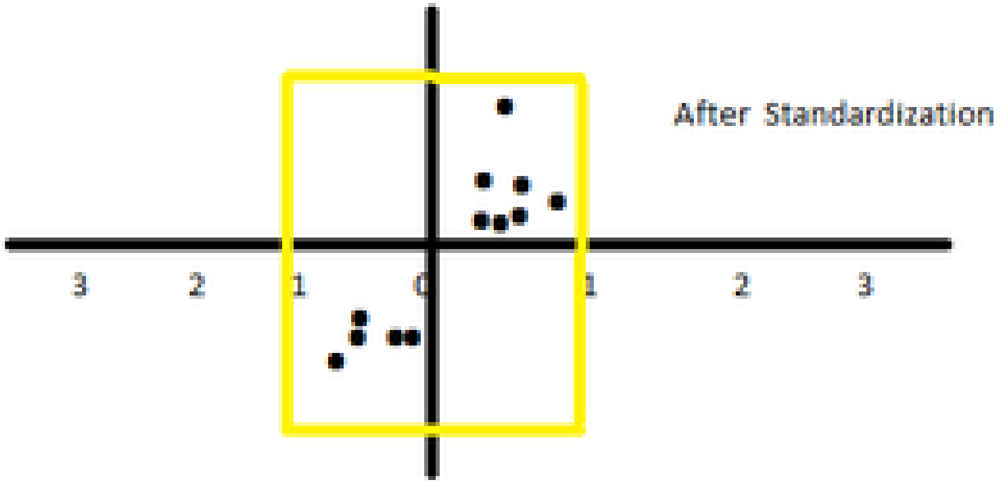
Statistics support

Mert Tanyur- Ipek Evren

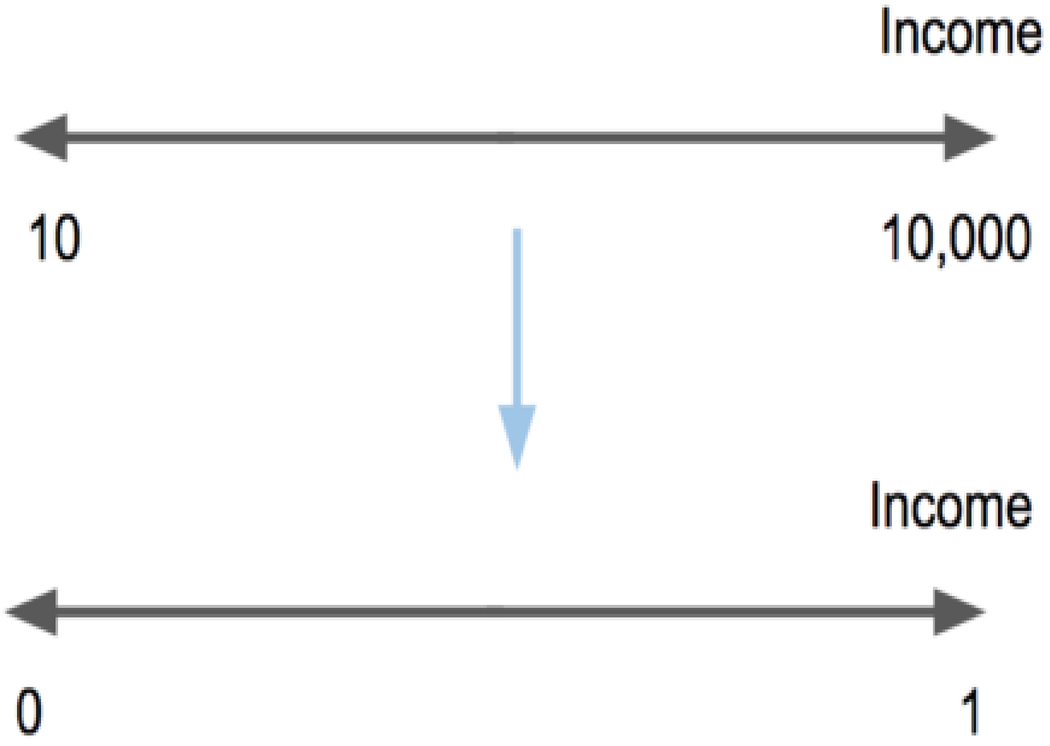
Standardization and normalization are best comrades of statistics


F1
Actual data
1 2 3 4 5 6 7 8 9 10
F2


F1
After normalizing
1
2
3
F2


After Standardization
3 2 1 0 1 2 3

Normalize



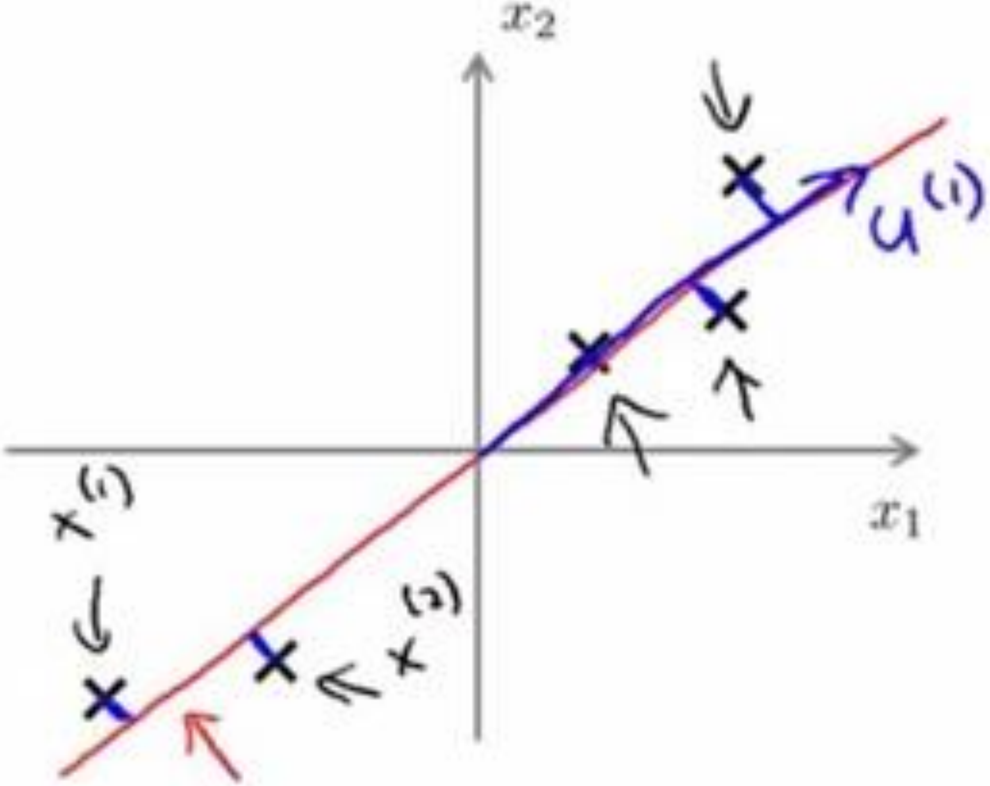
Standardize

$$z = \frac{x - \mu}{\sigma}$$

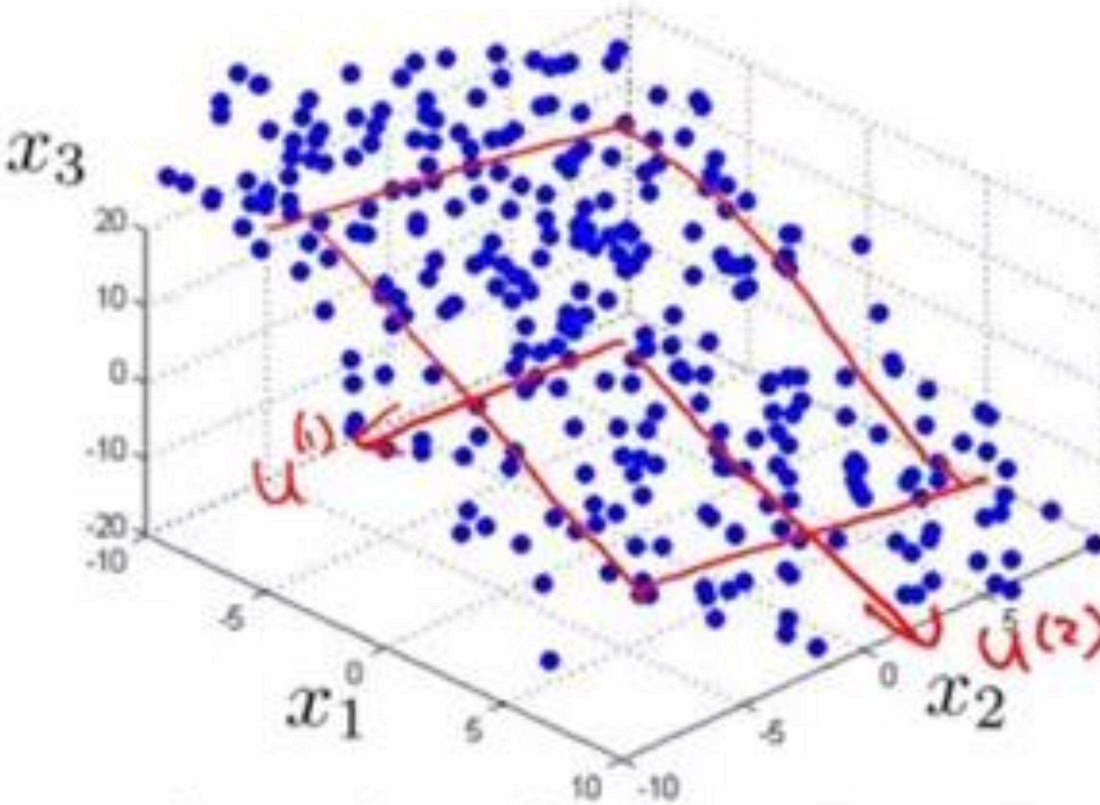
$\mu =$ Mean

$\sigma =$ Standard Deviation

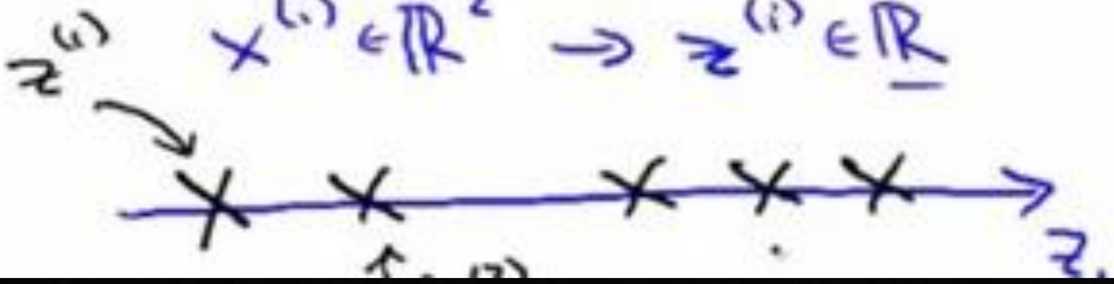
Principal Component Analysis (PCA) algorithm

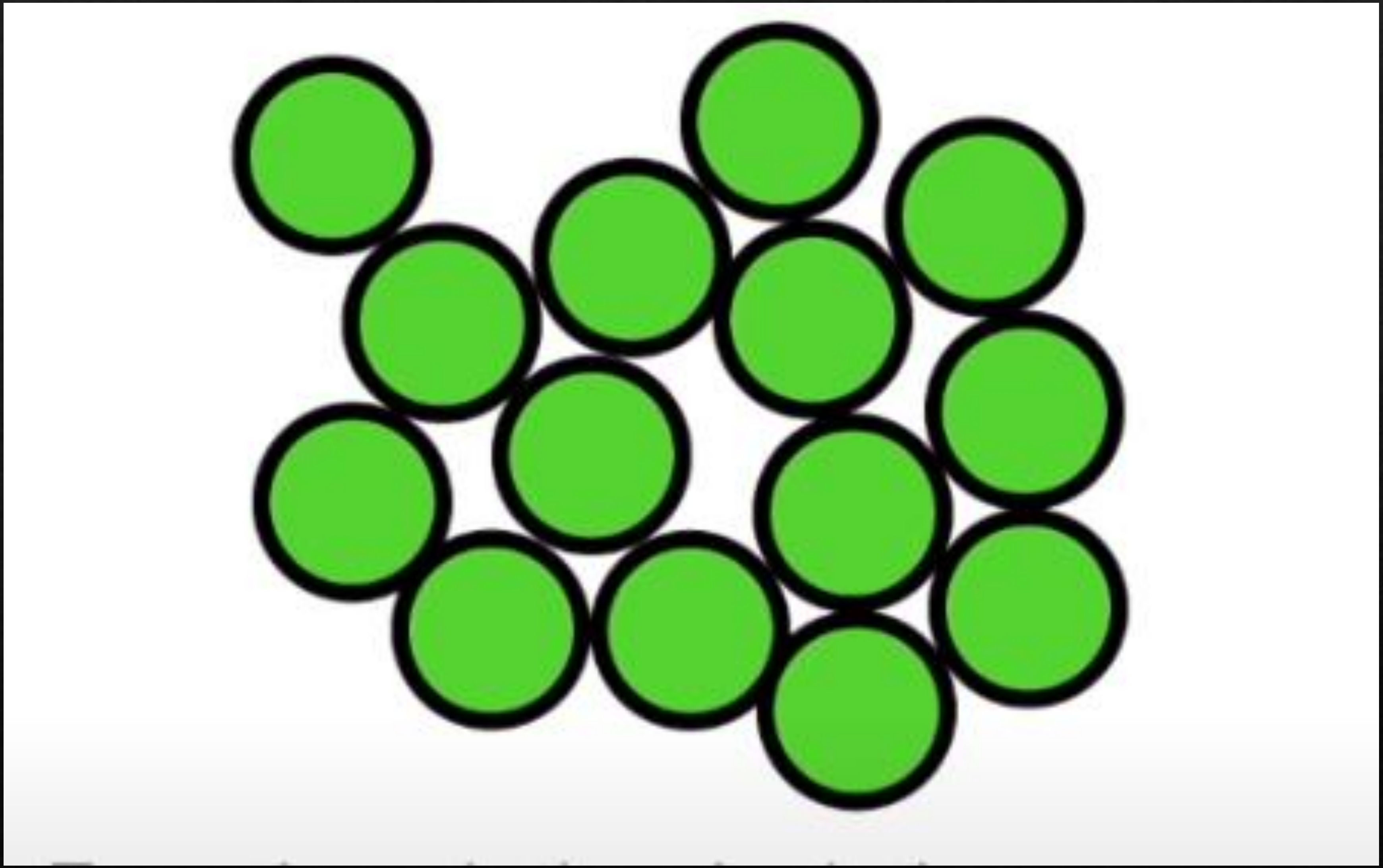


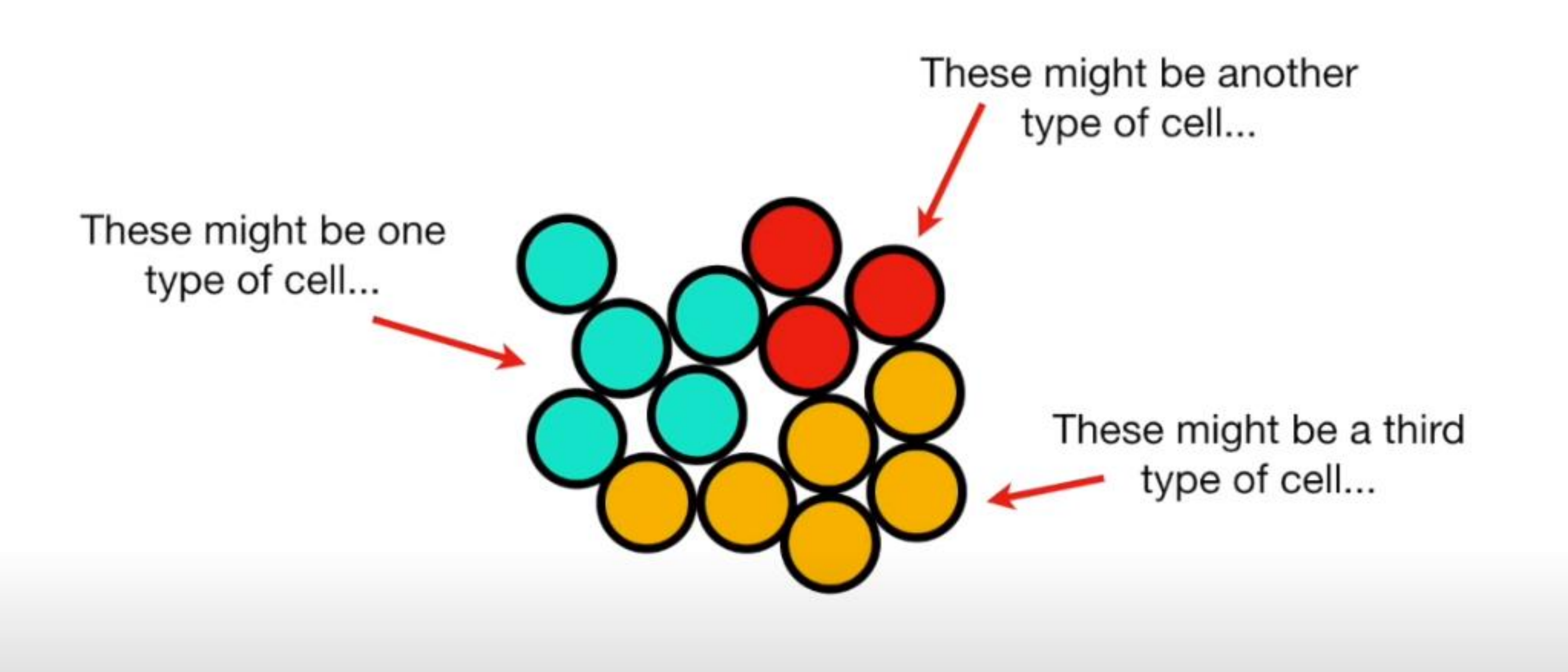
Reduce data from 2D to 1D

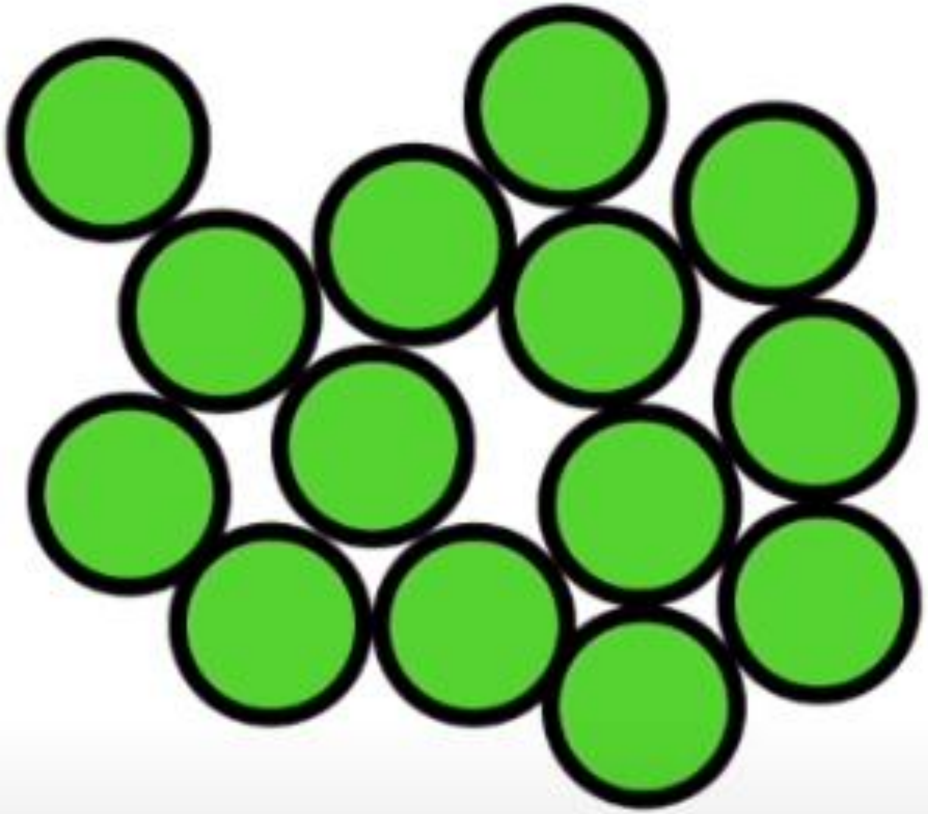


Reduce data from 3D to 2D



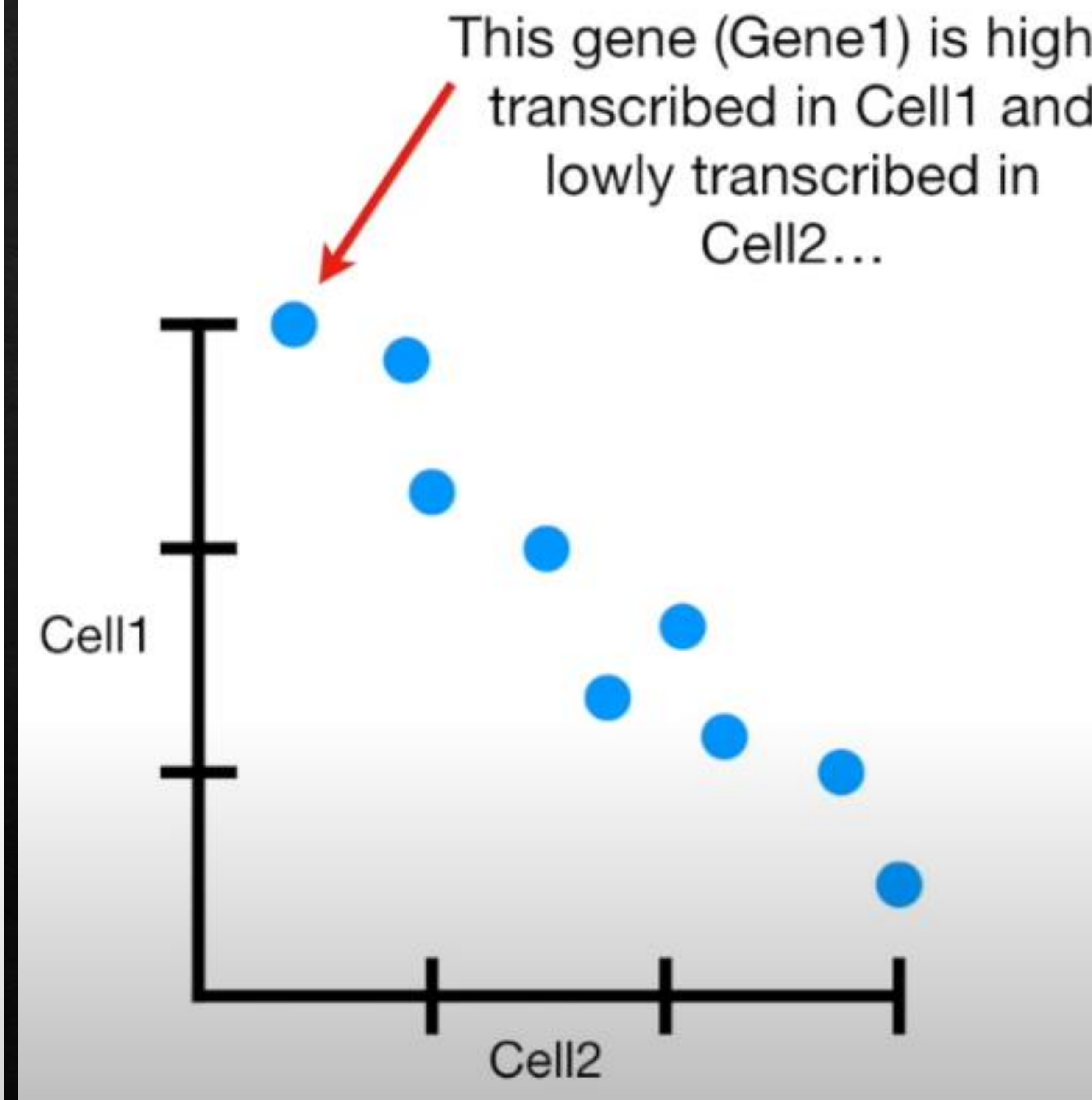


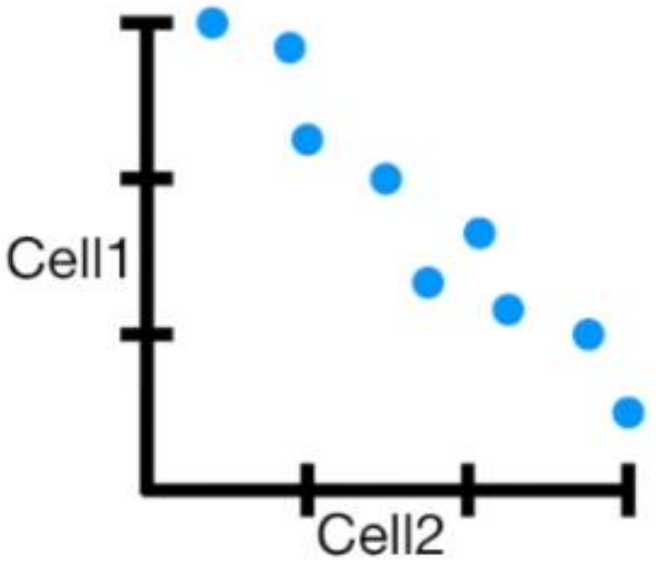

These might be another type of cell...
These might be one type of cell...
These might be a third type of cell...

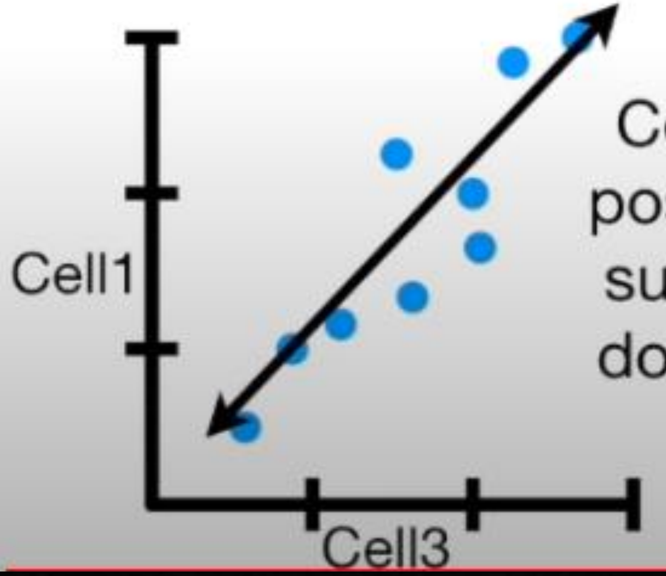


Unfortunately, we can't observe the differences from the outside…

	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
[illegible]	[illegible]	[illegible]



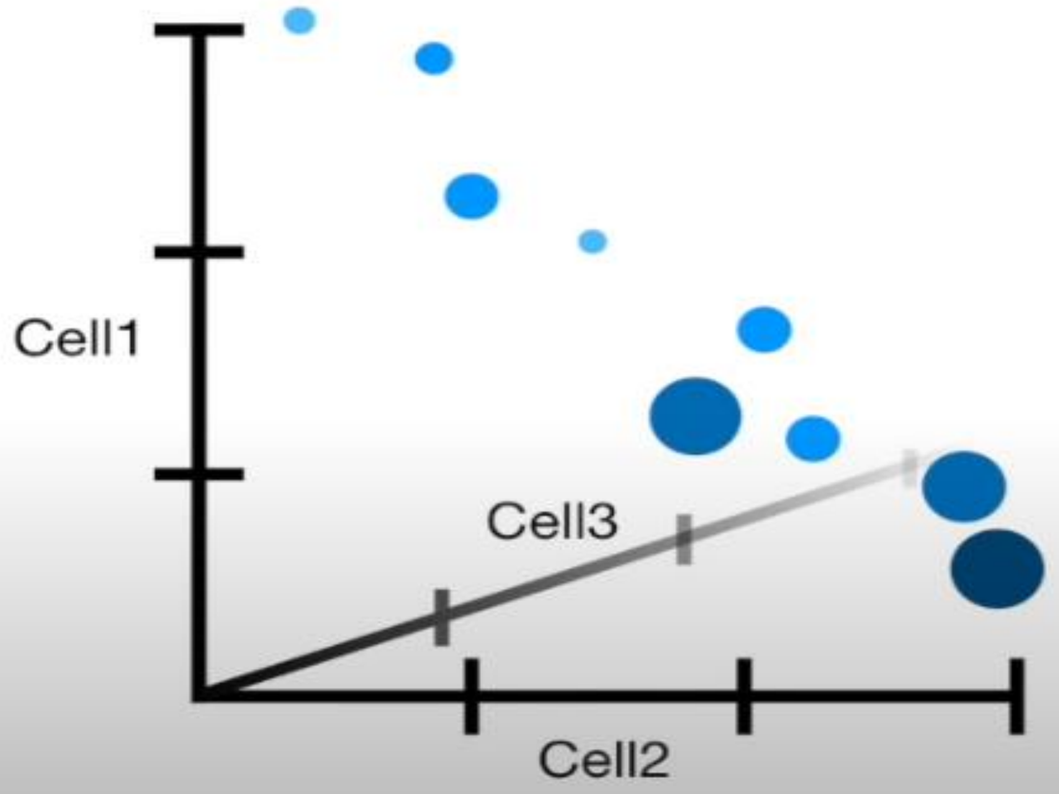




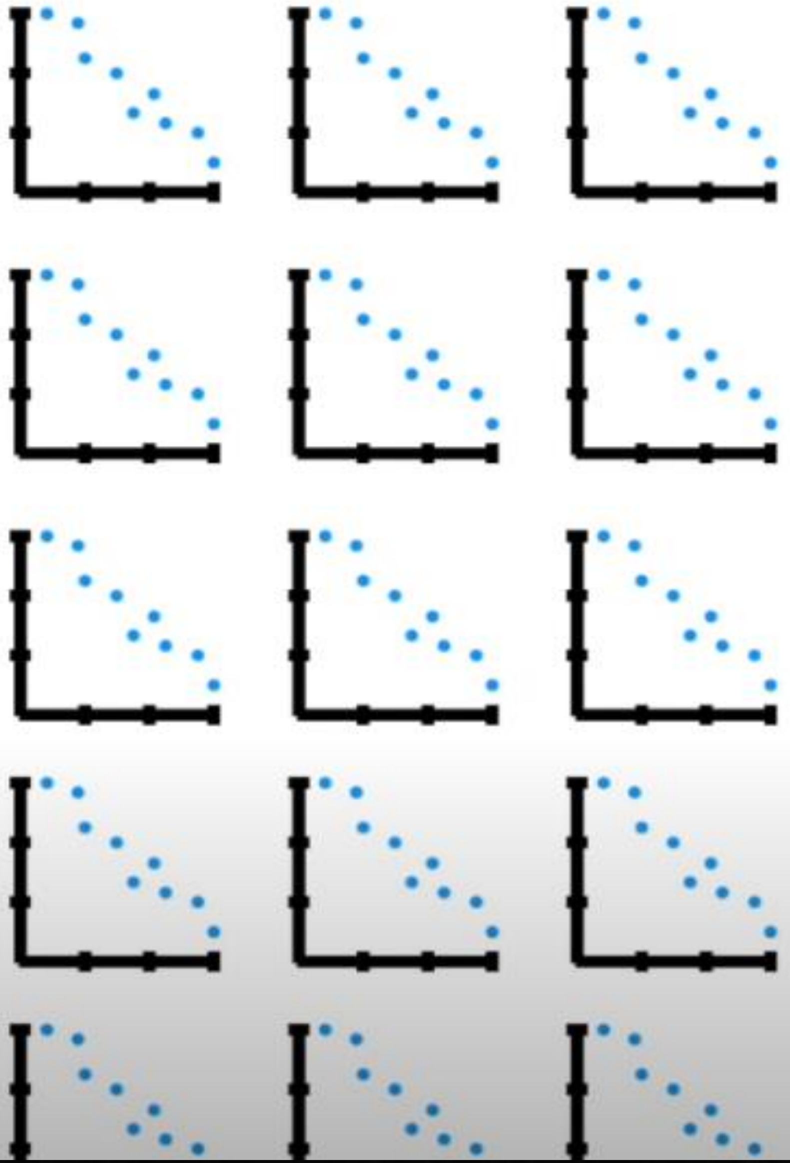
Cell1 and Cell3 are positively correlated, suggesting they are doing similar things.

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

Alternatively, we could try to plot all three cells at once on a 3-dimensional graph.



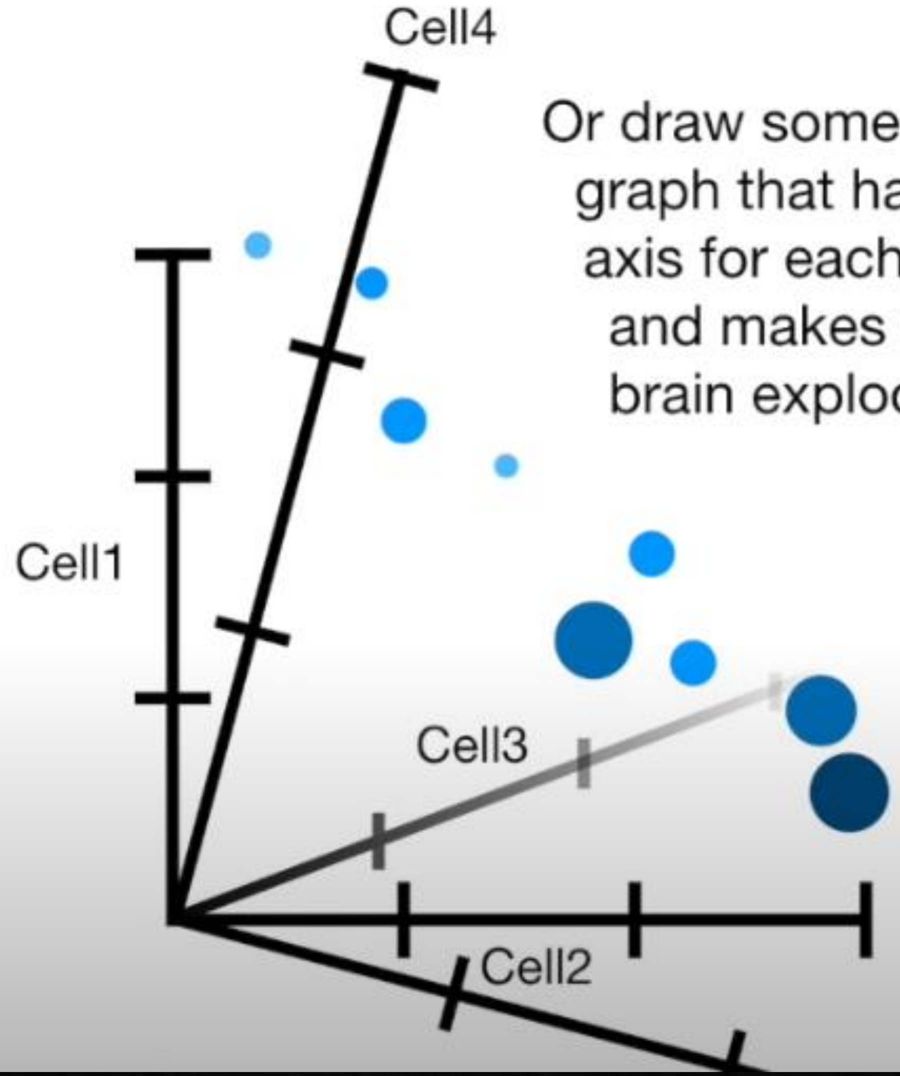
	Cell1	Cell2	Cell3
Gene1	3	0.25	2.8
Gene2	2.9	0.8	2.2
Gene3	2.2	1	1.5
Gene4	2	1.4	2
Gene5	1.3	1.6	1.6
Gene6	1.5	2	2.1
Gene7	1.1	2.2	1.2
Gene8	1	2.7	0.9
Gene9	0.4	3	0.6



Draw tons and tons of 2 cell plots and try to make sense of them all?

It would be Trivial !!

	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	...

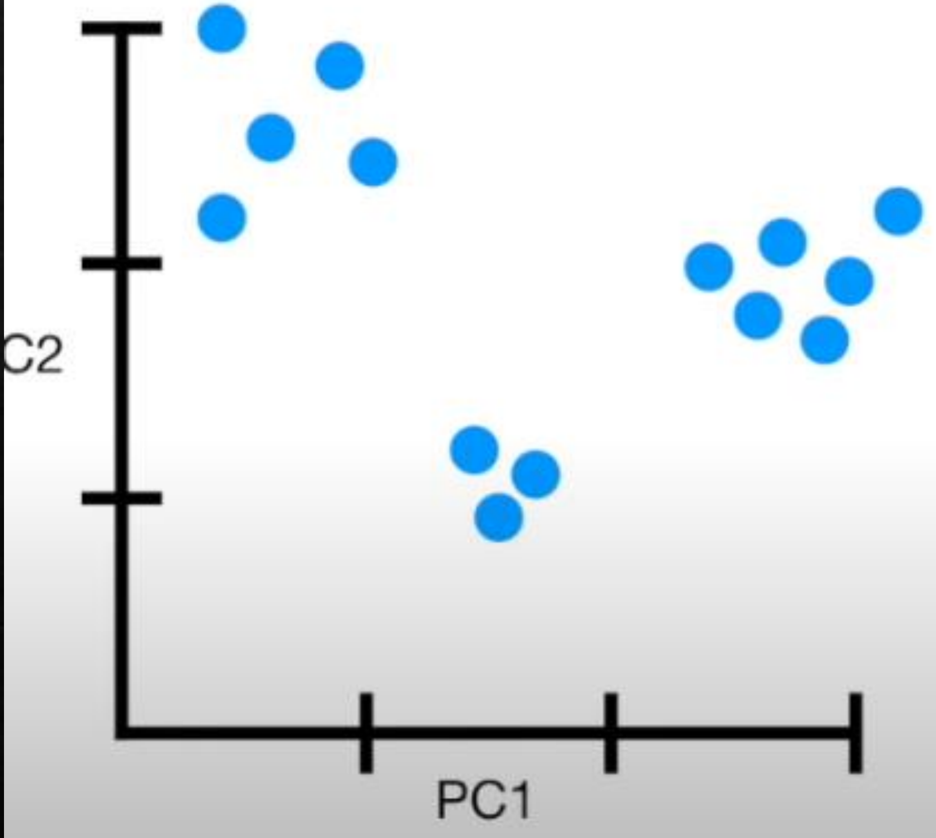


Or draw some crazy graph that has an axis for each cell and makes our brain explode?

	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	...

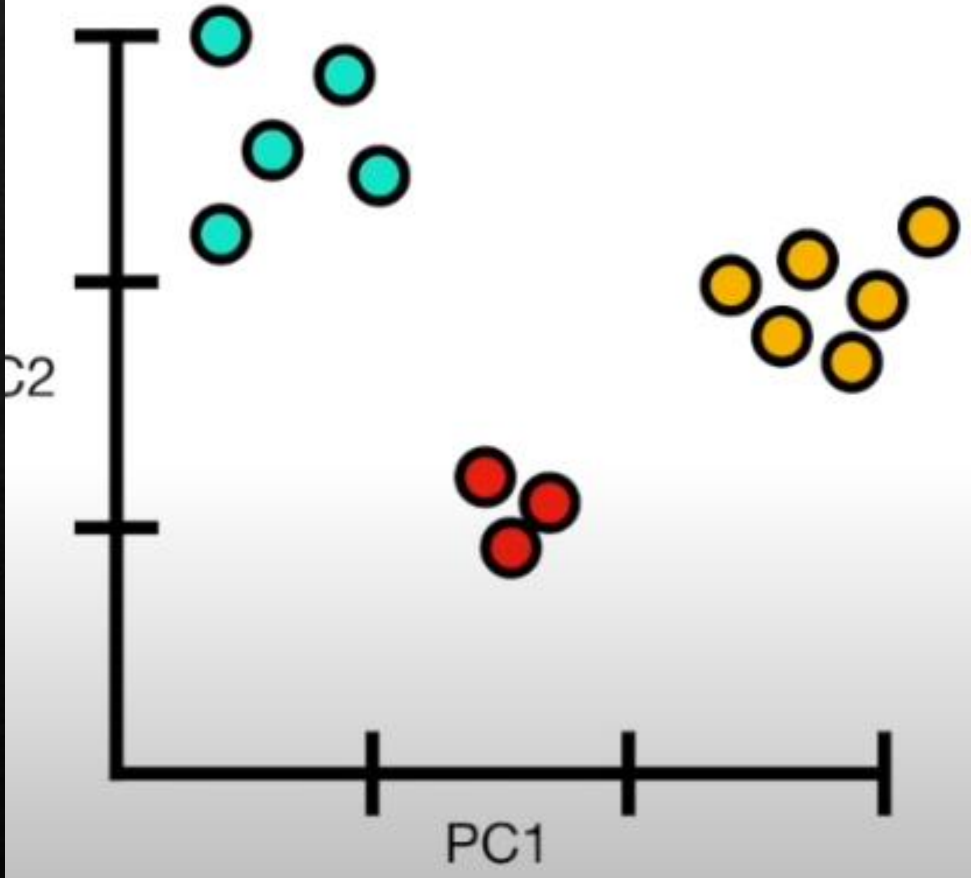
No, both of those options are just plain silly.

Instead, we draw a Principal Component Analysis (PCA) 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	...

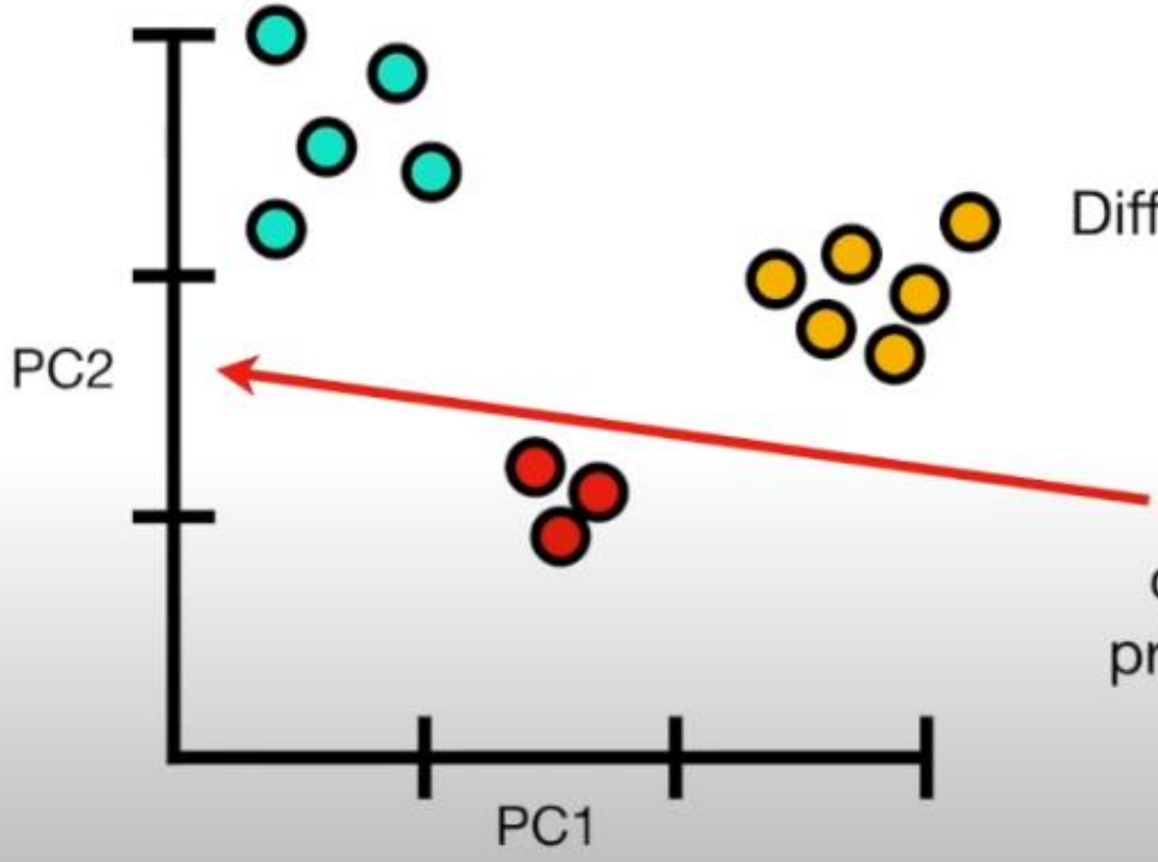
To make the clusters easier to see, we can color code them…



	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	...

Here’s one last main idea about how to interpret PCA plots:

The axes are ranked in order of importance.



Differences along the first principal component axis (PC1)…

…are more important than differences along the second principal component axis (PC2).

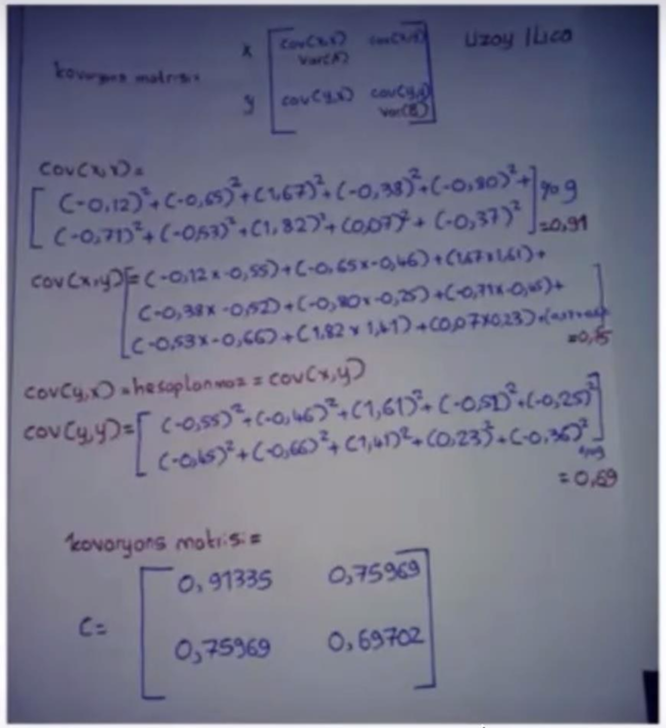
SORU

Örnekler	X	Y
1	0.72	0.13
2	0.18	0.23
3	2.50	2.30
4	0.45	0.16
5	0.04	0.44
6	0.13	0.24
7	0.30	0.03
8	2.65	2.10
9	0.91	0.91
10	0.46	0.32
Ortalama	0.83	0.69

Ortalamadan çıkardık

X	Y
0.72 - 0.83 = -0.12	0.13 - 0.69 = - 0.55
0.18 - 0.83 = -0.65	0.23 - 0.69 = - 0.46
2.50 - 0.83 = 1.67	2.30 - 0.69 = 1.61
0.45 - 0.83 = -0.38	0.16 - 0.69 = - 0.52
0.04 - 0.83 = -0.80	0.44 - 0.69 = - 0.25
0.13 - 0.83 = -0.71	0.24 - 0.69 = - 0.45
0.30 - 0.83 = -0.53	0.03 - 0.69 = - 0.66
2.65 - 0.83 = 1.82	2.10 - 0.69 = 1.41
0.91 - 0.83 = 0.07	0.91 - 0.69 = 0.23
0.46 - 0.83 = -0.37	0.32 - 0.69 = -0.36

Kovaryans matrisini buldук



Uzay Ilıca

kovaryans matrisi

$$\begin{array}{c|cc} & & \\ x & cov(x,x)\ var(A) & cov(x,y) \\ y & cov(y,x) & cov(y,y)\ var(B) \end{array}$$

cov(x,x) =

$$\left[(-0{,}12)^2+(-0{,}65)^2+(1{,}67)^2+(-0{,}38)^2+(-0{,}80)^2+(-0{,}71)^2+(-0{,}53)^2+(1{,}82)^2+(0{,}07)^2+(-0{,}37)^2\right] / 9 = 0{,}91$$

$$cov(x,y) = [(-0{,}12 \times -0{,}55)+(-0{,}65 \times -0{,}46)+(1{,}67 \times 1{,}61)+(-0{,}38 \times -0{,}52)+(-0{,}80 \times -0{,}25)+(-0{,}71 \times -0{,}45)+(-0{,}53 \times -0{,}66)+(1{,}82 \times 1{,}41)+(0{,}07 \times 0{,}23)+(\ldots)] = 0{,}85$$

cov(y,x) = hesaplanmaz = cov(x,y)

$$cov(y,y) = \left[(-0{,}55)^2+(-0{,}46)^2+(1{,}61)^2+(-0{,}52)^2+(-0{,}25)^2+(-0{,}45)^2+(-0{,}66)^2+(1{,}41)^2+(0{,}23)^2+(-0{,}36)^2\right] / 9 = 0{,}69$$

kovaryans matrisi =

$$C = \begin{bmatrix} 0{,}91335 & 0{,}75969 \\ 0{,}75969 & 0{,}69702 \end{bmatrix}$$

3.)Bu adımda elimizdeki bu 2×2 matrisin Özdeğerlerini (Eigenvalues) ve Özvektörlerini (Eigenvectors) buluyoruz.

2. Bu adımda aşağıdaki formülü kullanarak bu verinin kovaryans matrisini buluyoruz.

Örnekler	X	Y
1	0.72	0.13
2	0.18	0.23
3	2.50	2.30
4	0.45	0.16
5	0.04	0.44
6	0.13	0.24
7	0.30	0.03
8	2.65	2.10
9	0.91	0.91
10	0.46	0.32
Ortalama	0.83	0.69

X	Y
0.72 - 0.83 = -0.12	0.13 - 0.69 = - 0.55
0.18 - 0.83 = -0.65	0.23 - 0.69 = - 0.46
2.50 - 0.83 = 1.67	2.30 - 0.69 = 1.61
0.45 - 0.83 = -0.38	0.16 - 0.69 = - 0.52
0.04 - 0.83 = -0.80	0.44 - 0.69 = - 0.25
0.13 - 0.83 = -0.71	0.24 - 0.69 = - 0.45
0.30 - 0.83 = -0.53	0.03 - 0.69 = - 0.66
2.65 - 0.83 = 1.82	2.10 - 0.69 = 1.41
0.91 - 0.83 = 0.07	0.91 - 0.69 = 0.23
0.46 - 0.83 = -0.37	0.32 - 0.69 = -0.36

$$cov\,(x,y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$cov\,(Y,Y) = \frac{((-0.55)^2 + (-0.46)^2 + (1.61)^2 + (-0.52)^2 + (-0.25)^2 + (-0.45)^2 + (-0.66)^2 + (1.41)^2 + (0.23)^2 + (-0.36)^2)}{10-1}$$

$$cov\,(Y,Y) = 0.697029$$

$$\text{Covariance matrix} = C = \begin{bmatrix} 0.91335 & 0.75969 \\ 0.75969 & 0.69702 \end{bmatrix}$$

Uzay Ilıca

kovaryans matrisi =
$$\begin{array}{c c} x & \\ y & \end{array}\begin{bmatrix} cov(x,x) \; Var(A) & cov(x,y) \\ cov(y,x) & cov(y,y) \; Var(B) \end{bmatrix}$$

$cov(x,x) =$
$$\left[\begin{array}{l}(-0,12)^2 + (-0,65)^2 + (1,67)^2 + (-0,38)^2 + (-0,80)^2 + \\ (-0,71)^2 + (-0,53)^2 + (1,82)^2 + (0,07)^2 + (-0,37)^2\end{array}\right] \% 9 = 0,91$$

$$cov(x,y) = \left[\begin{array}{l}(-0,12 \times -0,55) + (-0,65 \times -0,46) + (1,67 \times 1,61) + \\ (-0,38 \times -0,52) + (-0,80 \times -0,25) + (-0,71 \times -0,45) + \\ (-0,53 \times -0,66) + (1,82 \times 1,41) + (0,07 \times 0,23) + (0,37 \times 0,36)\end{array}\right] = 0,75$$

$cov(y,x) =$ hesaplanmaz $= cov(x,y)$

$$cov(y,y) = \left[\begin{array}{l}(-0,55)^2 + (-0,46)^2 + (1,61)^2 + (-0,52)^2 + (-0,25)^2 \\ (-0,45)^2 + (-0,66)^2 + (1,41)^2 + (0,23)^2 + (-0,36)^2\end{array}\right] \% 9 = 0,69$$

kovaryans matrisi =

$$C = \begin{bmatrix} 0,91335 & 0,75969 \\ 0,75969 & 0,69702 \end{bmatrix}$$