The problem in Genomics:

Too many dimensions (data points)

- 35.000 genes X 20 microarray experiments = 700.000 dimensions
- But we can only perceive 3 dimensions…

So, we must decrease the dimensions of the data set

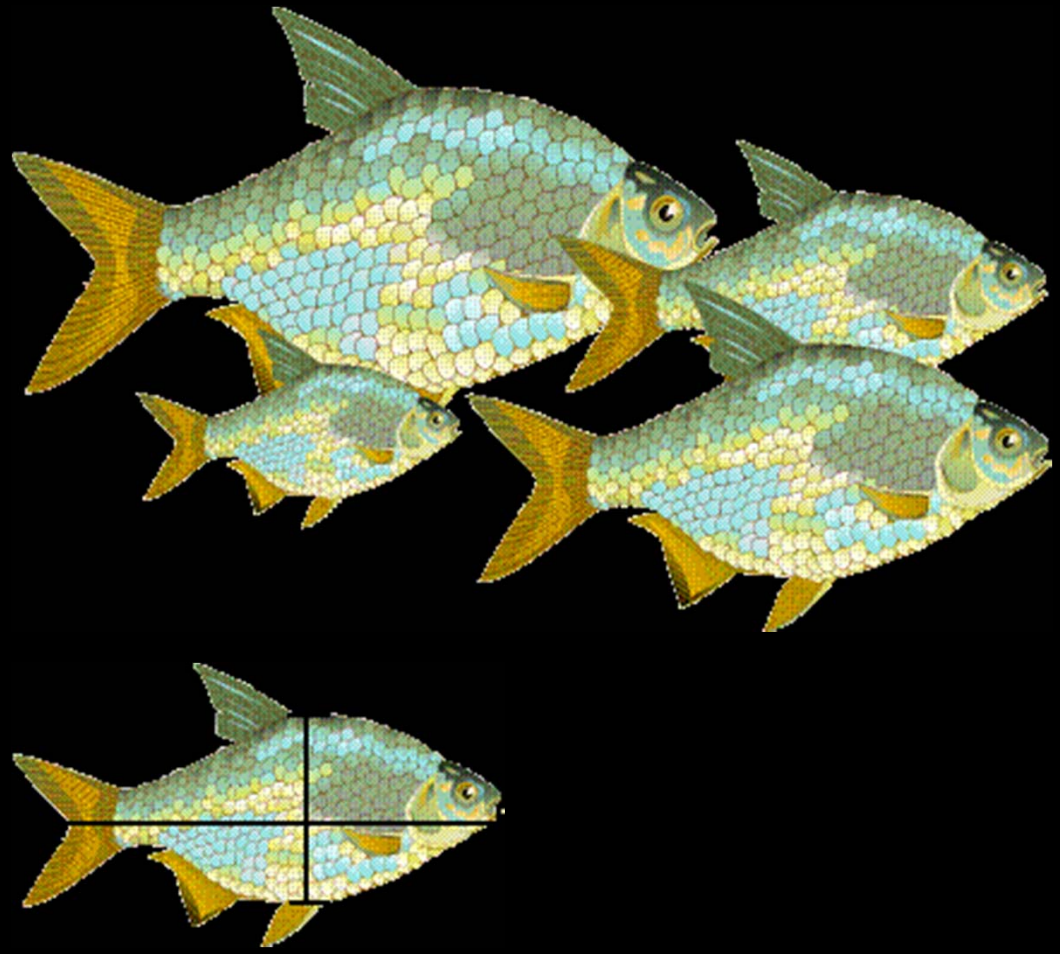
in order to perceive the data

Projecting data onto fewer dimensions may sound like science fiction but you are all familiar with it.

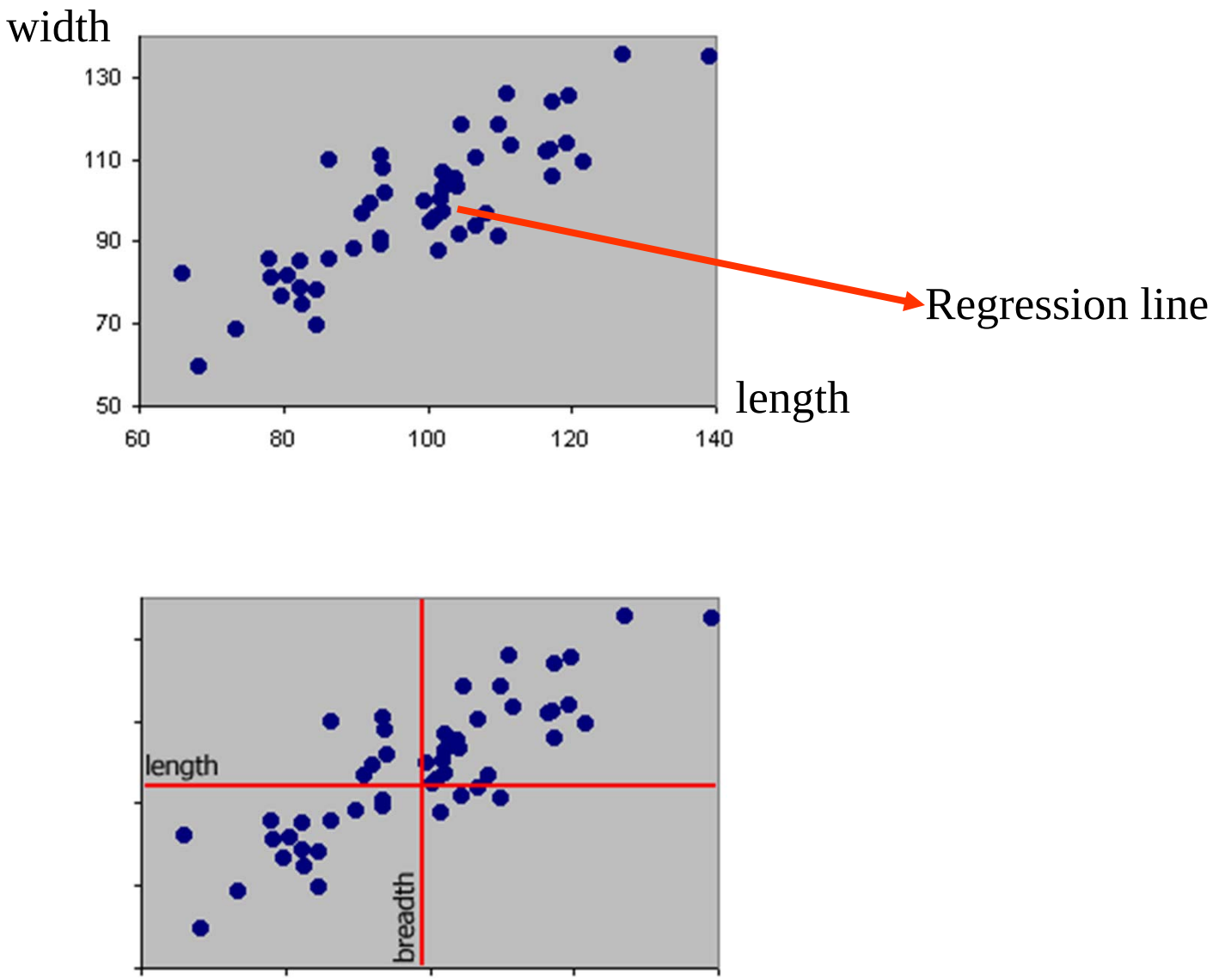
That's an eagle!

But: how do you know? It is a 2D picture… Not the real 3D bird

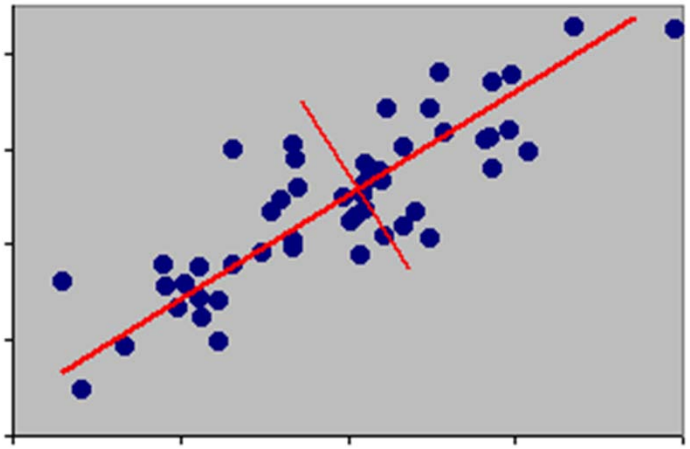
The truth is: **fewer dimensions can still retain much of the original information**

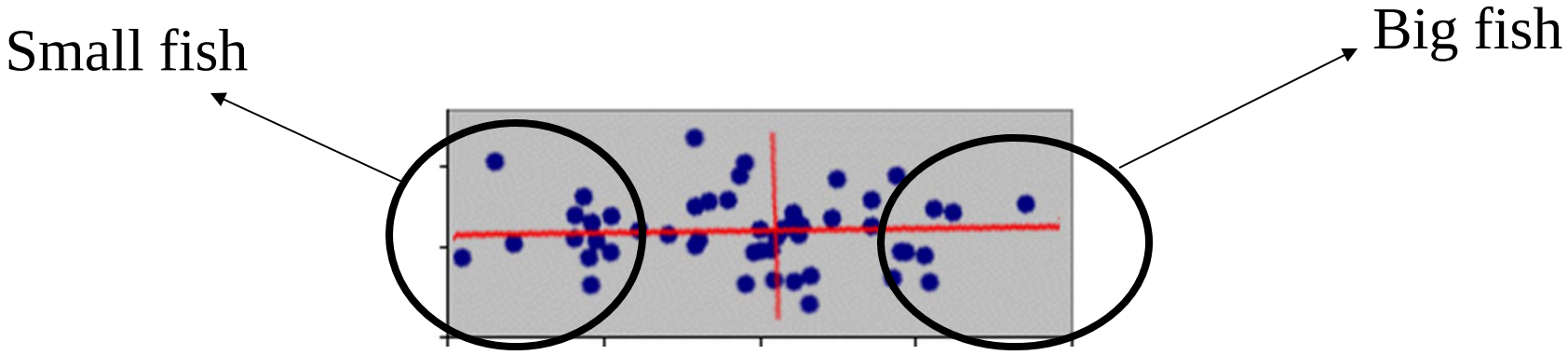


Suppose that 50 fish were measured for their length and width…


width
130
110
90
70
50
60
80
100
120
140
length
Regression line
length
breadth

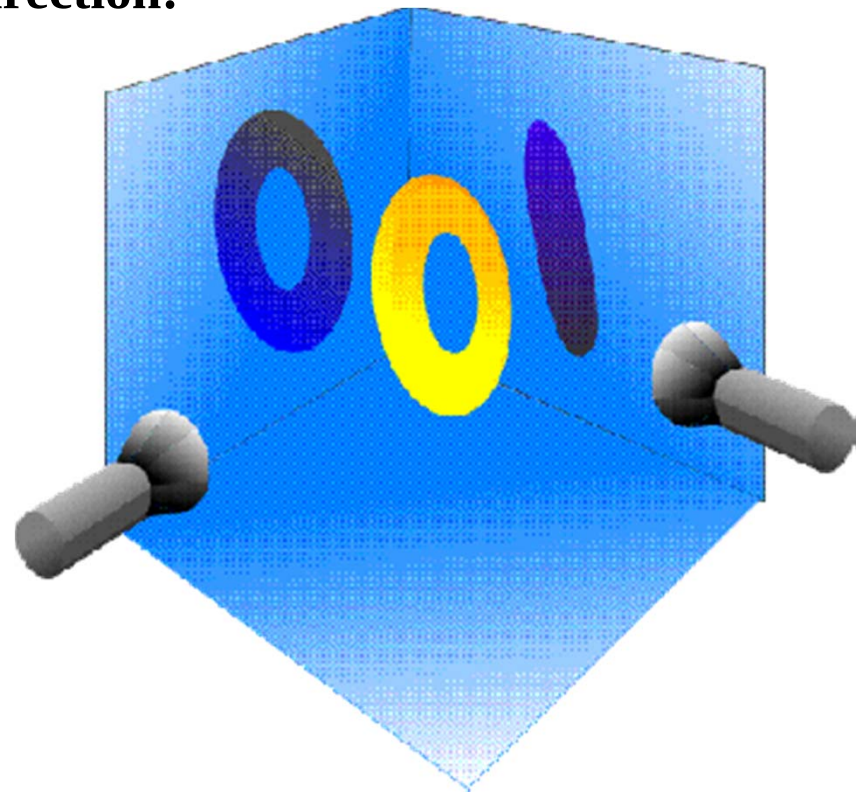
Try to rotate the axis:





Length? Width? after all, you mean **size**…?

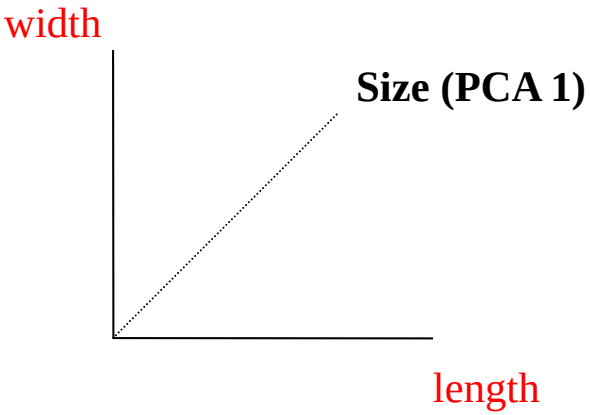
Your data have direction:

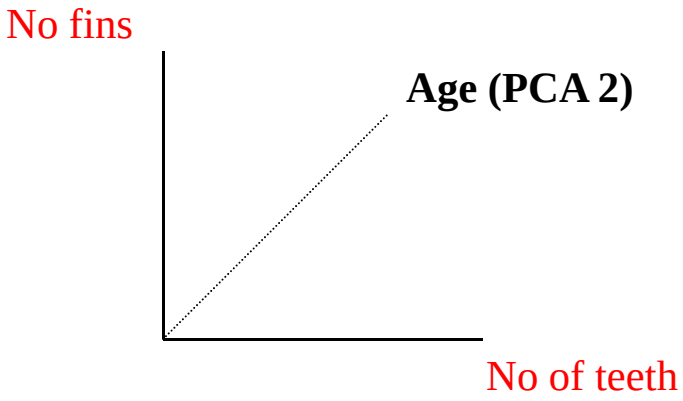


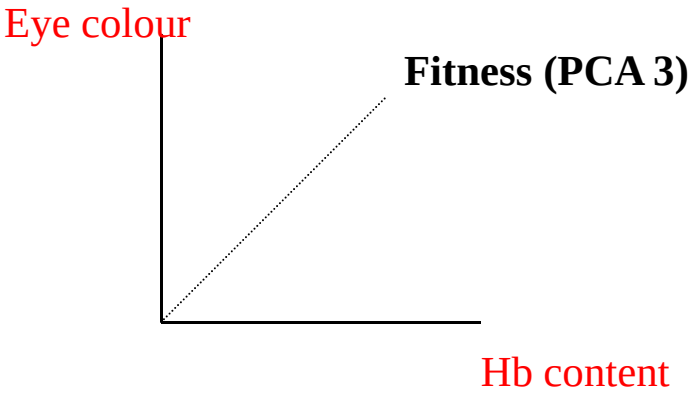
Shine a light onto this doughnut from two different directions.

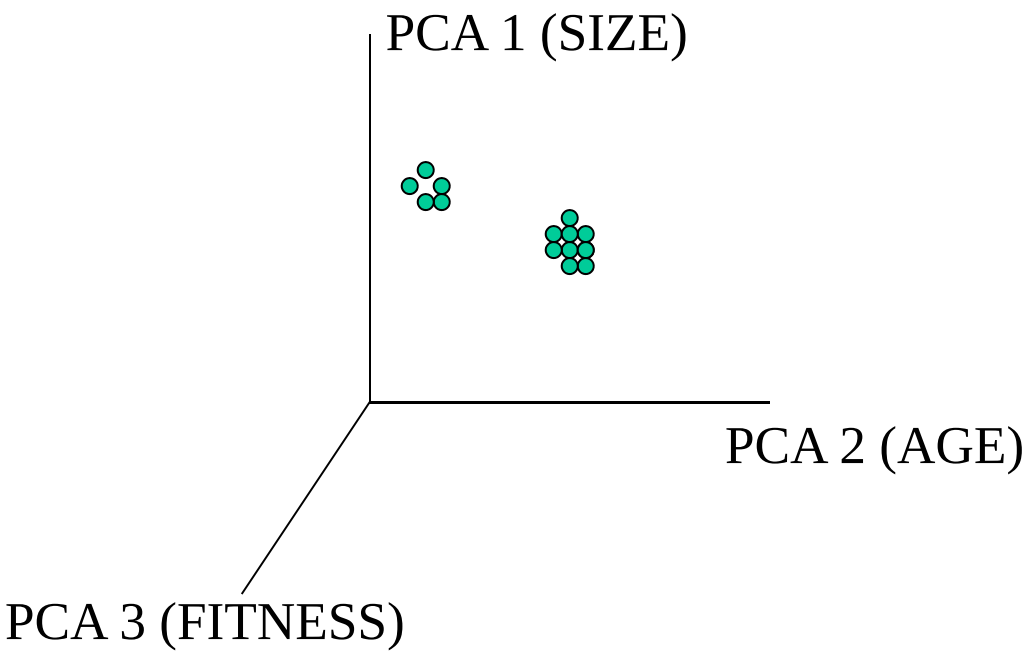
These lights cast shadows onto two 'screens'. The nature of the shadow is dependent on the position of the torch.

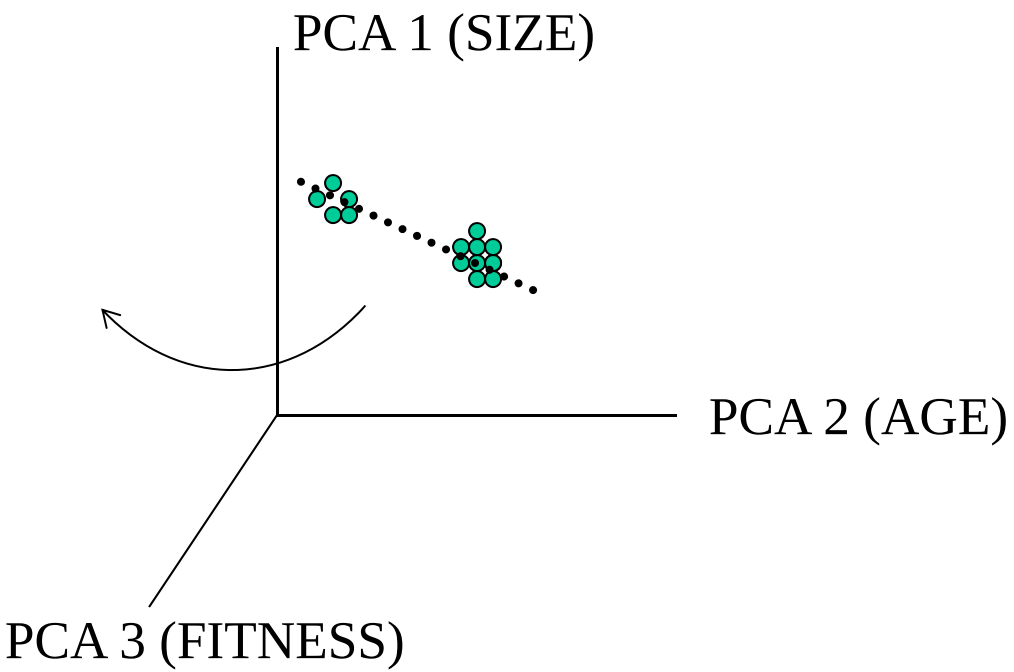
1. move the doughnut and keep the torches stationary
2. keep the doughnut stationary and move the torches.

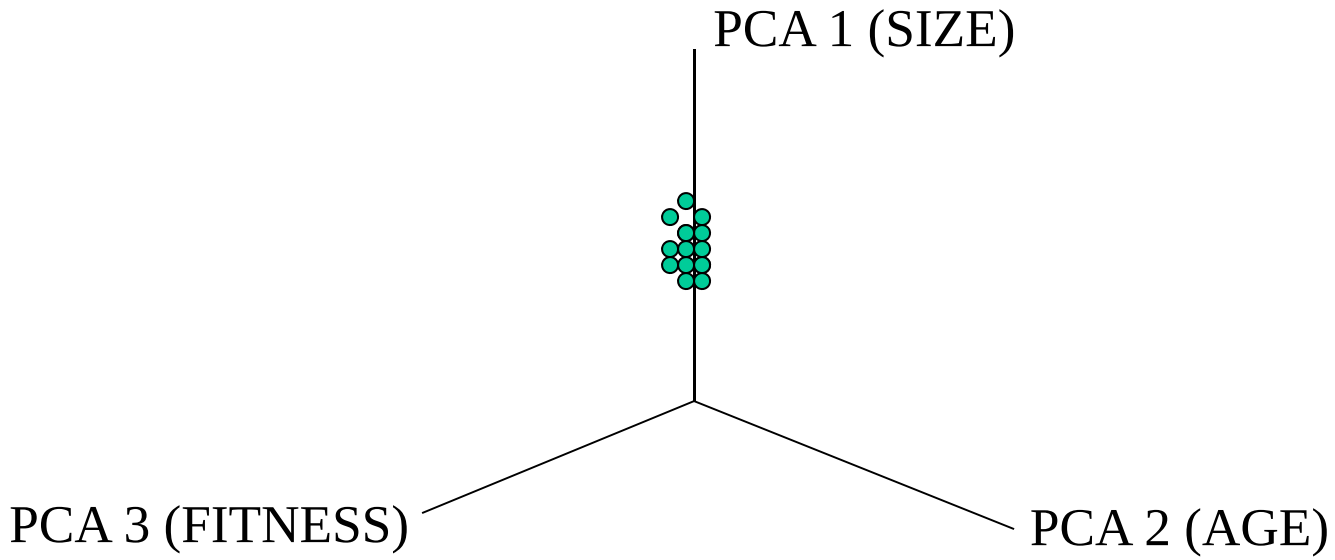

width
Size (PCA 1)
length

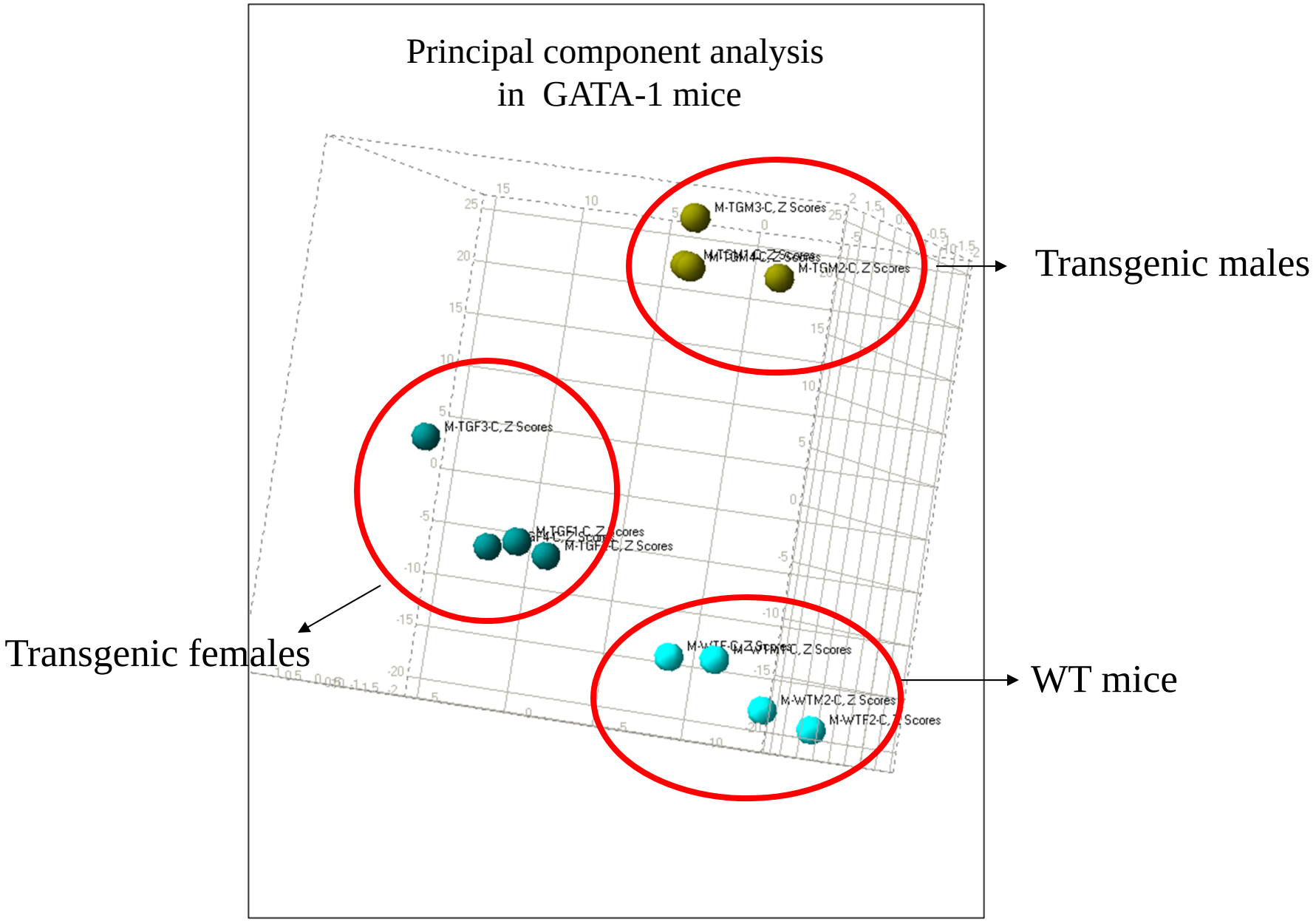

No fins
Age (PCA 2)
No of teeth


Eye colour
Fitness (PCA 3)
Hb content


PCA 1 (SIZE)
PCA 2 (AGE)
PCA 3 (FITNESS)

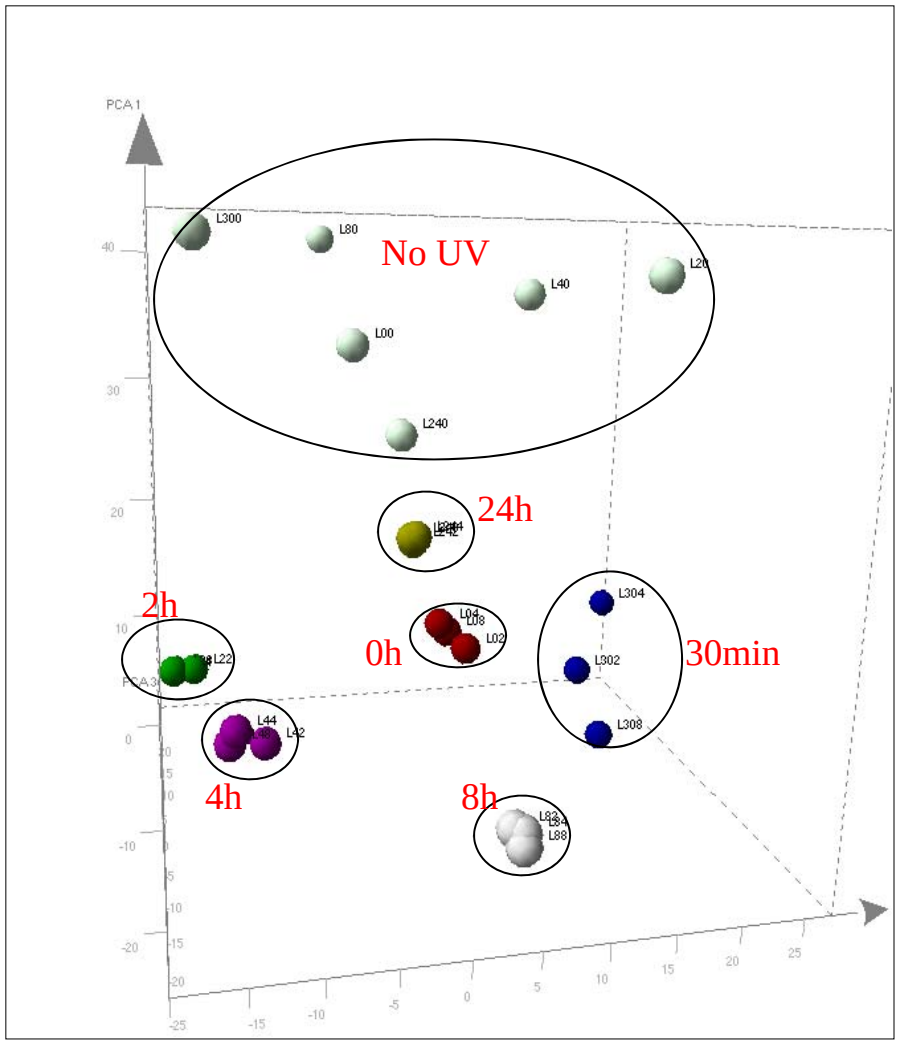

PCA 1 (SIZE)
PCA 2 (AGE)
PCA 3 (FITNESS)

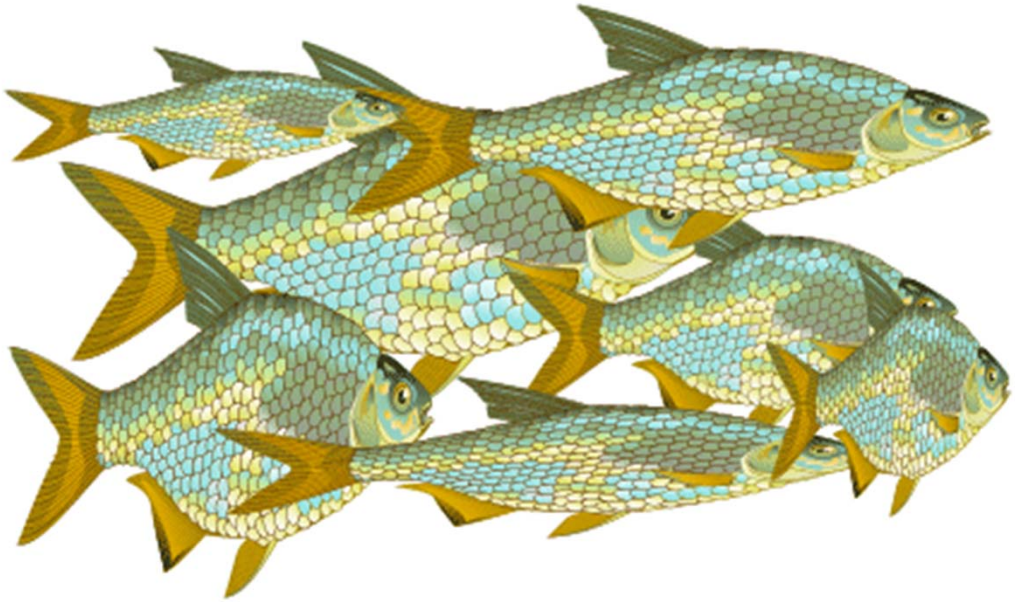

PCA 1 (SIZE)
PCA 3 (FITNESS)
PCA 2 (AGE)


Principal component analysis
in GATA-1 mice
M-TGM3-C, Z Scores
M-TGM2-C, Z Scores
M-TGF3-C, Z Scores
M-TGF2-C, Z Scores
M-WTM2-C, Z Scores
M-WTF2-C, Z Scores
Transgenic males
Transgenic females
WT mice

Grouping the time?

PCA





Could you represent these fish by a size variable alone? If not, why not?

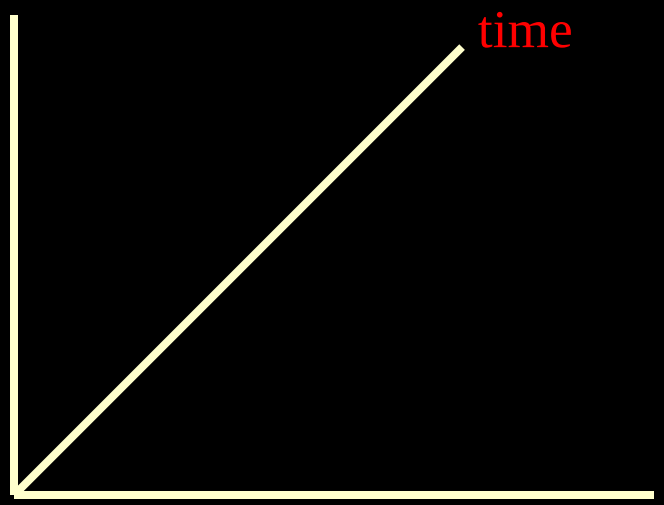
The obvious answer is that they are different sizes *and* shapes

probably because these fish are of a different species…

What is the principal component of AGING process?

variables

Grey hair
Weight
Heart failure incidence
Cancer incidence
Bone alterations



Probably…time is one of the principle components

Questions:
What are the variables that project in the PC of time?
But: which are the other 2 principle components?

Other ways to decrease dimensions of our data sets

is to…

GROUP THE VARIABLES

Various ways to group or to…cluster:

Hierarchical clusters

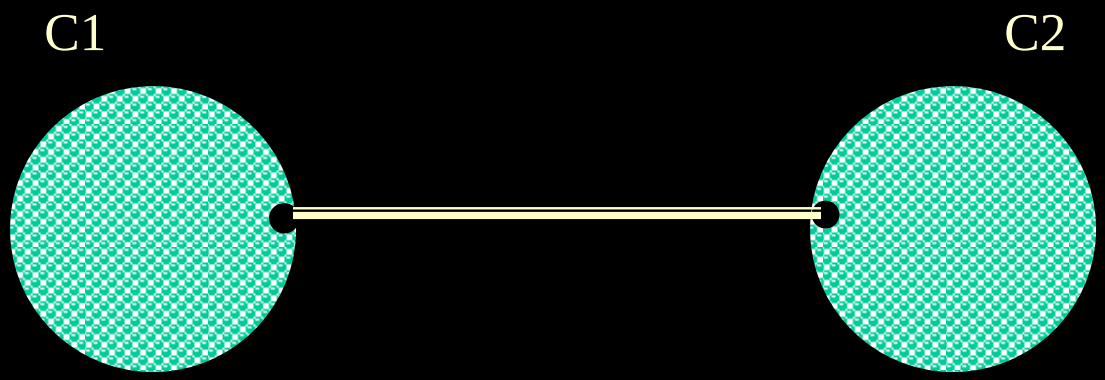
-Single linkage
-Average linkage
-Complete linkage

Partitioning clusters

-K-cluster
-Artificial neural networks

SINGLE LINKAGE

In *single-linkage* clustering, we consider the distance between one cluster and another cluster to be equal to the shortest distance from any member of one cluster to any member of the other cluster.

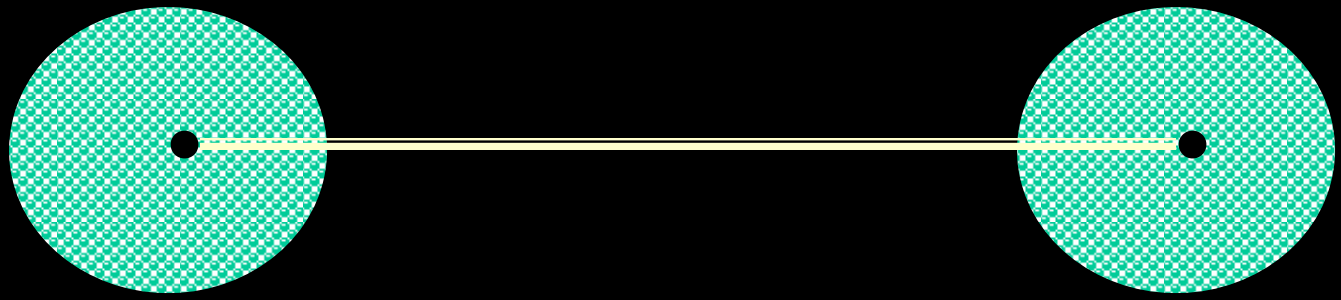


Complete linkage

In *complete-linkage* clustering we consider the distance between one cluster and another cluster to be equal to the longest distance from any member of one cluster to any member of the other cluster.

Average linkage

In *average-linkage* clustering, we consider the distance between one cluster and another cluster to be equal to the average distance from any member of one cluster to any member of the other cluster.



Hierarchical clusters

Given a set of N cities to be clustered, and an NxN distance (km) table, the basic process of hierarchical clustering is this:

	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0

After merging BOS with NY:

	BOS/NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY	**0**	**223**	**1308**	**802**	**2815**	**2934**	**2786**	**1771**
DC	**223**	**0**	**1075**	**671**	**2684**	**2799**	**2631**	**1616**
MIA	**1308**	**1075**	**0**	**1329**	**3273**	**3053**	**2687**	**2037**
CHI	**802**	**671**	**1329**	**0**	**2013**	**2142**	**2054**	**996**
SEA	**2815**	**2684**	**3273**	**2013**	**0**	**808**	**1131**	**1307**
SF	**2934**	**2799**	**3053**	**2142**	**808**	**0**	**379**	**1235**
LA	**2786**	**2631**	**2687**	**2054**	**1131**	**379**	**0**	**1059**
DEN	**1771**	**1616**	**2037**	**996**	**1307**	**1235**	**1059**	**0**

Then we compute the distance from this new cluster to all other clusters, to get a new distance matrix:

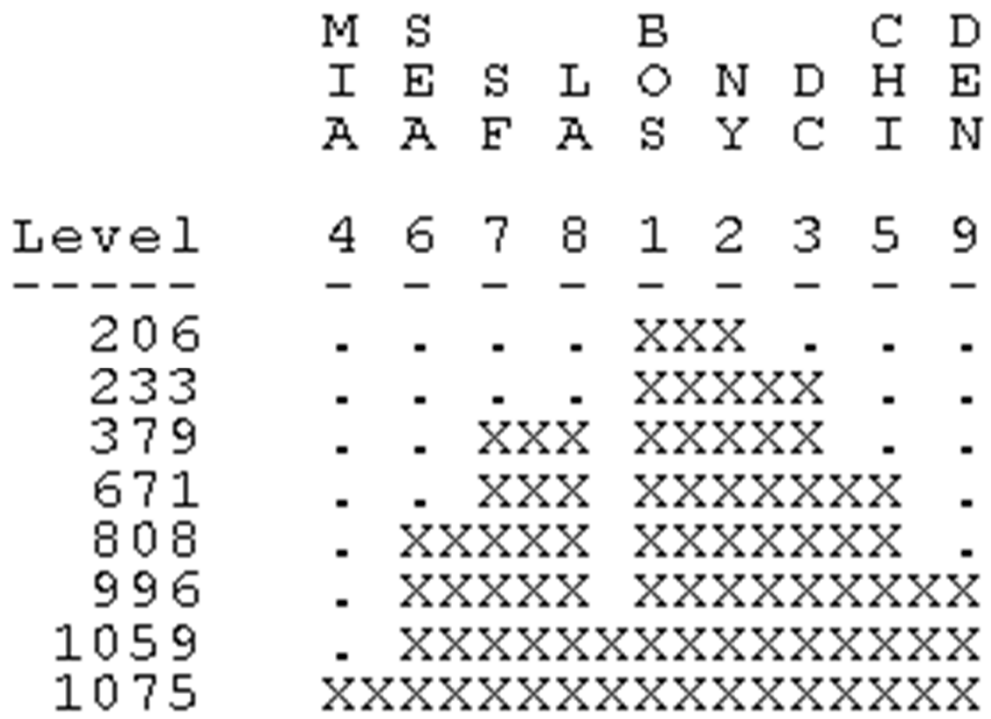
After merging DC with BOS-NY:

	BOS/NY/DC	MIA	CHI	SEA	SF	LA	DEN
BOS/NY/DC	0	1075	671	2684	2799	2631	1616
MIA	1075	0	1329	3273	3053	2687	2037
CHI	671	1329	0	2013	2142	2054	996
SEA	2684	3273	2013	0	808	1131	1307
SF	2799	3053	2142	808	0	**379**	1235
LA	2631	2687	2054	1131	379	0	1059
DEN	1616	2037	996	1307	1235	1059	0

Now, the nearest pair of objects is SF and LA, at distance 379. These are merged into a single cluster called "SF/LA". Then we compute the distance from this new cluster to all other objects, to get a new distance matrix:

	BOS/ NY/DC	MIA	CHI	SEA	SF/LA	DEN
BOS/NY/DC	0	1075	671	2684	2631	1616
MIA	1075	0	1329	3273	2687	2037
CHI	671	1329	0	2013	2054	996
SEA	2684	3273	2013	0	808	1307
SF/LA	2631	2687	2054	808	0	1059
DEN	1616	2037	996	1307	1059	0

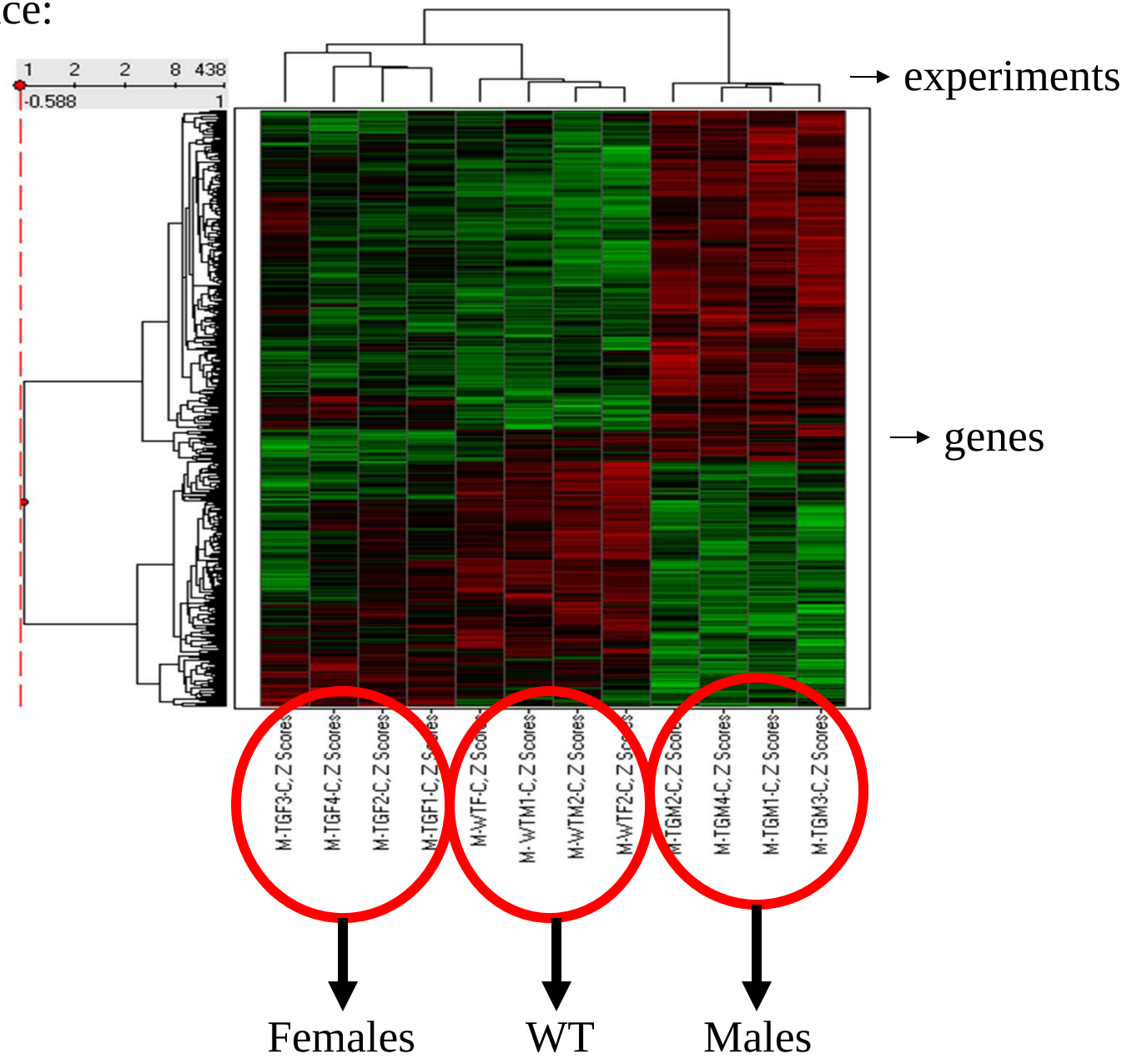
The whole process is then summarized:



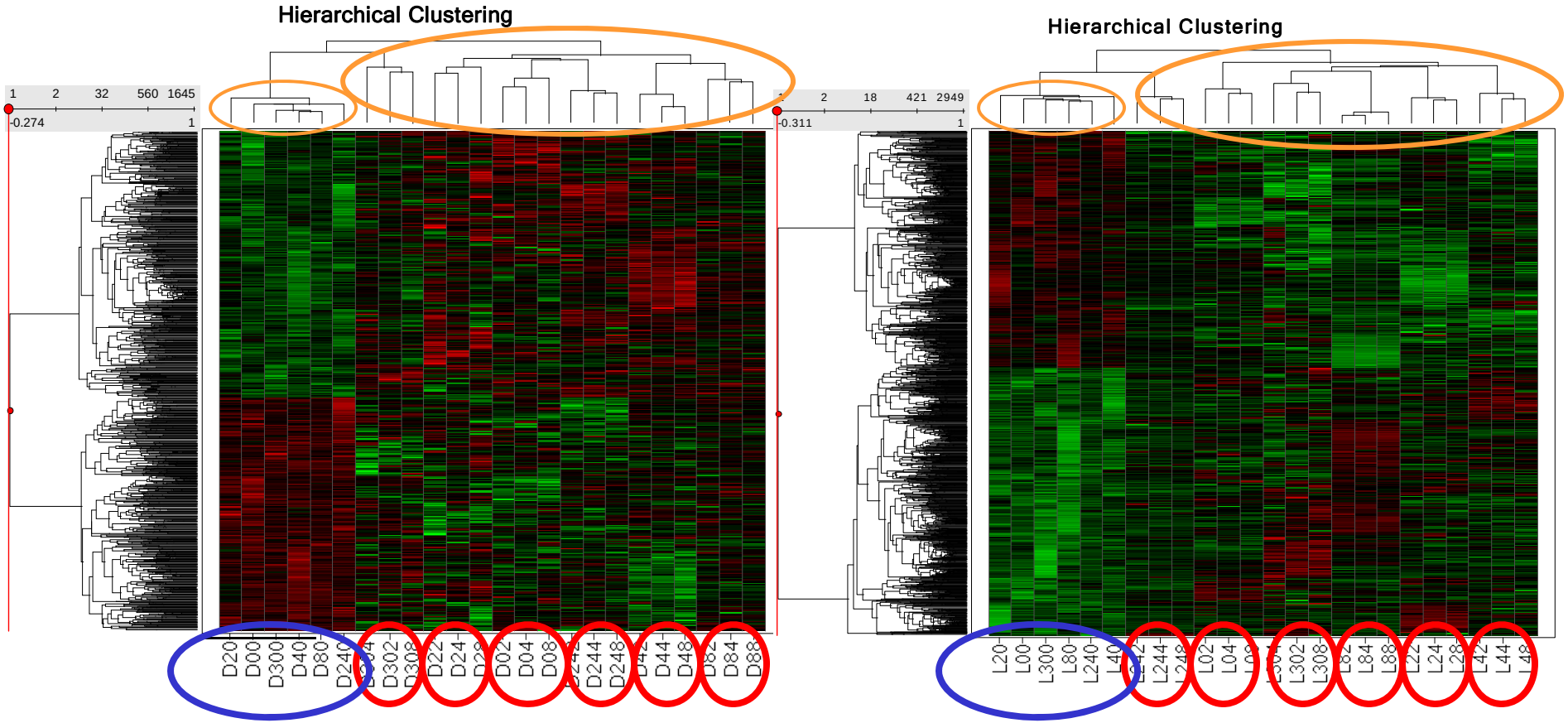
Level	4	6	7	8	1	2	3	5	9
	M	S			B			C	D
	I	E	S	L	O	N	D	H	E
	A	A	F	A	S	Y	C	I	N
206	.	.	.	.	XXX	.	.	.	
233	.	.	.	.	XXXXX	.	.		
379	.	.	XXX		XXXXX	.	.		
671	.	.	XXX		XXXXXXX	.			
808	.	XXXXX			XXXXXXX	.			
996	.	XXXXX			XXXXXXXXX				
1059	.	XXXXXXXXXXXXXXX							
1075	XXXXXXXXXXXXXXXXX								

In the diagram, the columns are associated with the items and the rows are associated with levels (stages) of clustering. An 'X' is placed between two columns in a given row if the corresponding items are merged at that stage in the clustering.

The Gata-1 mice:



UV-irradiated cells in time:



Partitioning clusters

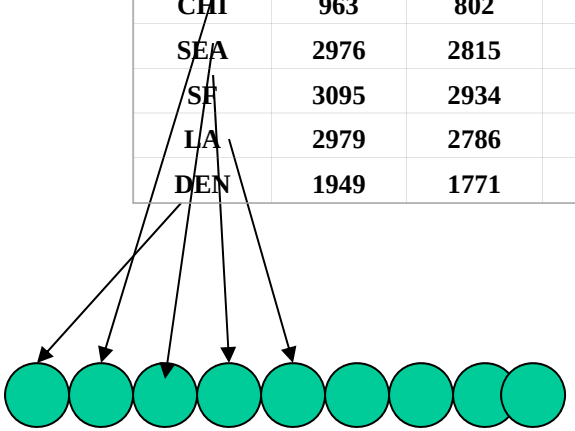
The difference between partitioning clusters from hierarchical clusters is:

-Here there is no summary of the clustering process

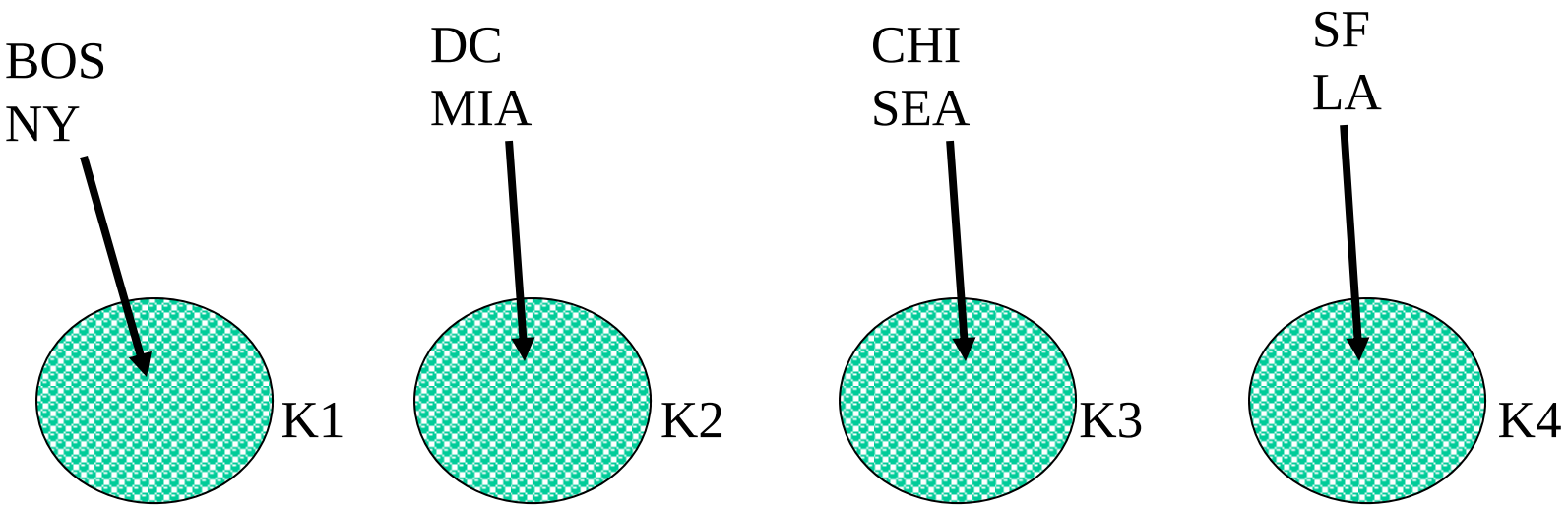
-We define the number of clusters and the threshold of their greatest possible distinction

K-cluster

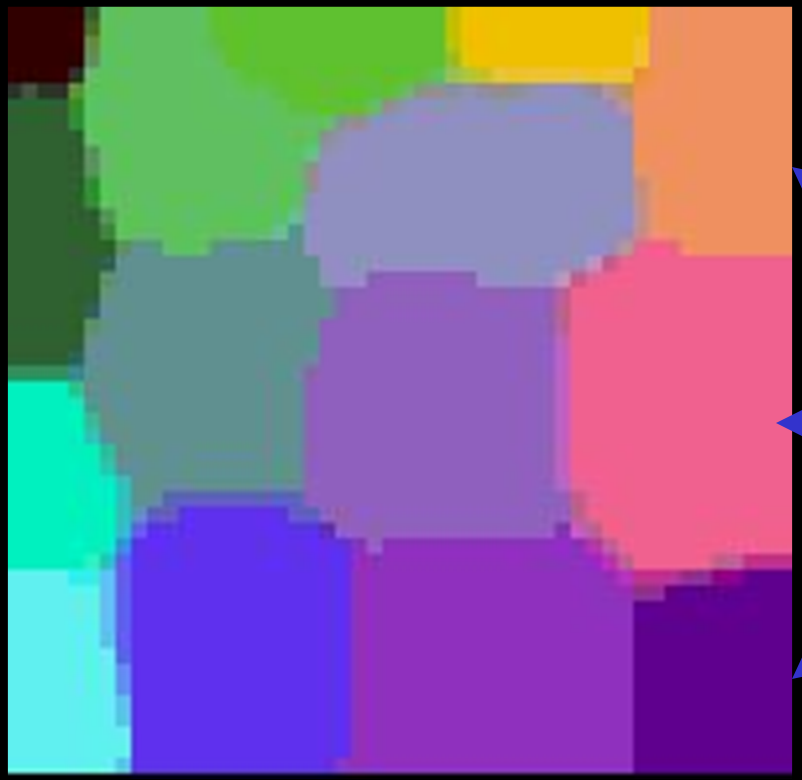
	BOS	NY	DC	MIA	CHI	SEA	SF	LA	DEN
BOS	0	206	429	1504	963	2976	3095	2979	1949
NY	206	0	233	1308	802	2815	2934	2786	1771
DC	429	233	0	1075	671	2684	2799	2631	1616
MIA	1504	1308	1075	0	1329	3273	3053	2687	2037
CHI	963	802	671	1329	0	2013	2142	2054	996
SEA	2976	2815	2684	3273	2013	0	808	1131	1307
SF	3095	2934	2799	3053	2142	808	0	379	1235
LA	2979	2786	2631	2687	2054	1131	379	0	1059
DEN	1949	1771	1616	2037	996	1307	1235	1059	0



…..K clusters where K= No of groups of cities



Self organizing maps

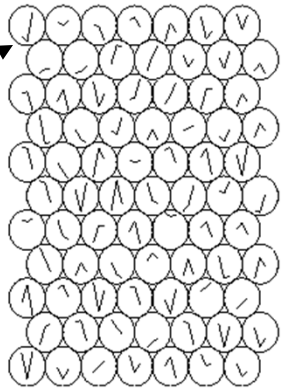


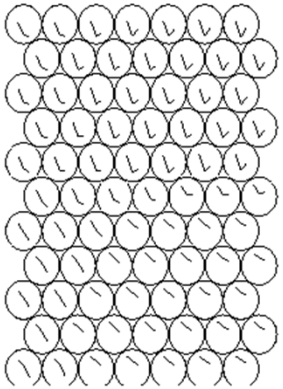
Self organizing maps

Random Ordered

Fold change

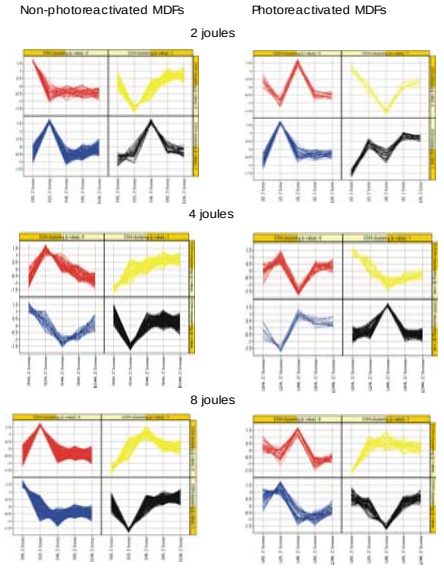
time





K-CLUSTER

You can group your genes by means of their expression profiles



Up

	Csb-Xpa	Ercc1	
Hamp	13.98	22.98	ercc1
Hamp	10.59	18.26	ercc1 ; csb-xpa
Afp	9.98	7.85	ercc1 ; csb-xpa
Afp	6.30	5.06	ercc1 ; csb-xpa
Akr1b7	6.30	4.70	ercc1 ; csb-xpa
Gtl2	3.97	4.44	ercc1 ; csb-xpa
Nope	5.81	4.22	ercc1 ; csb-xpa
Apoa4	5.24	4.19	ercc1
Peg3	3.23	4.10	ercc1 ; csb-xpa
Afp	4.69	3.86	ercc1
Rex3	4.42	3.44	ercc1 ; csb-xpa
Stfa1	6.18	3.08	csb-xpa
Apoa4	3.70	2.87	ercc1
4931408D1	3.38	2.87	ercc1 ; csb-xpa
Peg3	3.38	2.86	ercc1 ; csb-xpa
Bex2	3.66	2.83	ercc1 ; csb-xpa
Tnfrsf12a	5.05	2.60	ercc1 ; csb-xpa
Serpine1	3.79	2.36	csb-xpa
Nope	3.75	2.16	ercc1 ; csb-xpa
Ubd	4.12	2.15	ercc1 ; csb-xpa
Marco	4.84	1.98	ercc1 ; csb-xpa
Gp49b	3.70	1.97	ercc1 ; csb-xpa
Asns	4.36	1.95	ercc1 ; csb-xpa
Spink3	4.75	1.87	ercc1 ; csb-xpa
Hip1r	3.35	1.68	csb-xpa
Akr1c18	4.03	1.49	csb-xpa
LOC26888	6.46	1.12	csb-xpa
Ier3	3.43	1.01	csb-xpa

Down

	Csb-Xpa	Ercc1	
Thrsp	-7.05	-10.09	ercc1
Gck	-6.69	-6.99	ercc1 ; csb-xpa
Gck	-4.79	-6.82	ercc1 ; csb-xpa
Thrsp	-5.55	-5.11	ercc1
2300002F06Rik	-3.36	-3.51	ercc1 ; csb-xpa
2300002F06Rik	-3.43	-3.19	ercc1 ; csb-xpa
Cyp4f14	-3.57	-2.91	ercc1 ; csb-xpa
Cyp2j5	-3.08	-2.40	ercc1 ; csb-xpa
Slco2b1	-2.79	-2.29	ercc1 ; csb-xpa
Es31	-3.45	-2.25	ercc1 ; csb-xpa
Gpt1	-3.13	-2.16	ercc1 ; csb-xpa
Lpin1	-2.79	-2.11	csb-xpa
Cml1	-3.12	-1.88	ercc1 ; csb-xpa
Lpin1	-2.68	-1.82	csb-xpa
Serpina3k	-4.03	-1.74	ercc1 ; csb-xpa
Mmd2	-2.95	-1.71	csb-xpa
Cyp2b20	-2.90	-1.69	csb-xpa
Cyp2b10	-2.81	-1.56	csb-xpa
Car5a	-2.96	-1.54	csb-xpa
Fmo2	-4.50	-1.45	csb-xpa
Fmo2	-2.80	-1.38	csb-xpa
Fmo2	-3.23	-1.37	csb-xpa
Cmah	-3.22	-1.34	csb-xpa
Npm3	-3.16	-1.16	csb-xpa
Slc26a1	-3.32	-1.09	csb-xpa
C730021L23	-3.48	1.06	csb-xpa
Xpa	-2.85	1.08	csb-xpa
Cflar	-3.44	1.37	csb-xpa

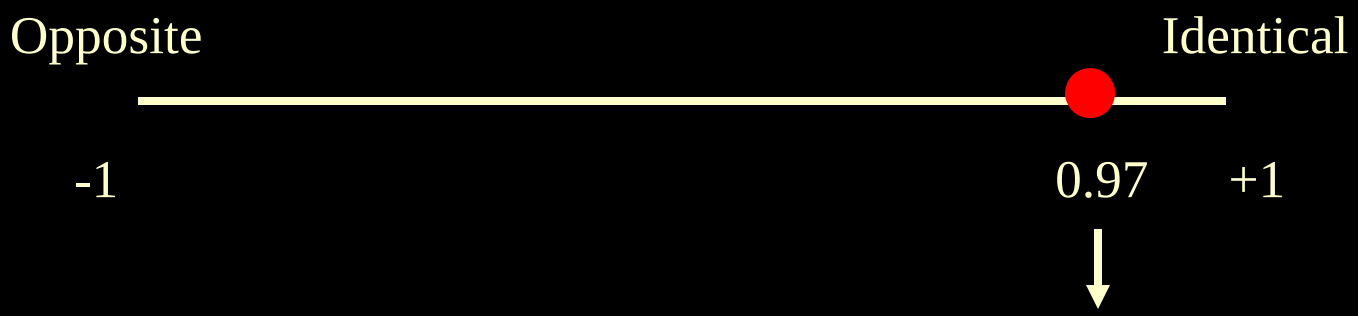
Is a *p-value* and a *fold change* enough

to distinguish

the relevant genes from the rest

and show how *similar* or *different* are these mice?

Ercc1 and Csb-Xpa mice are quite similar…



Pearson's correlation metrics
on all genes (regardless the *fold change*
and the *p-value*).

…but neither the *p-value* nor the *fold change* or both
could display this similarity.

Group genes by biological information: