Machine Learning

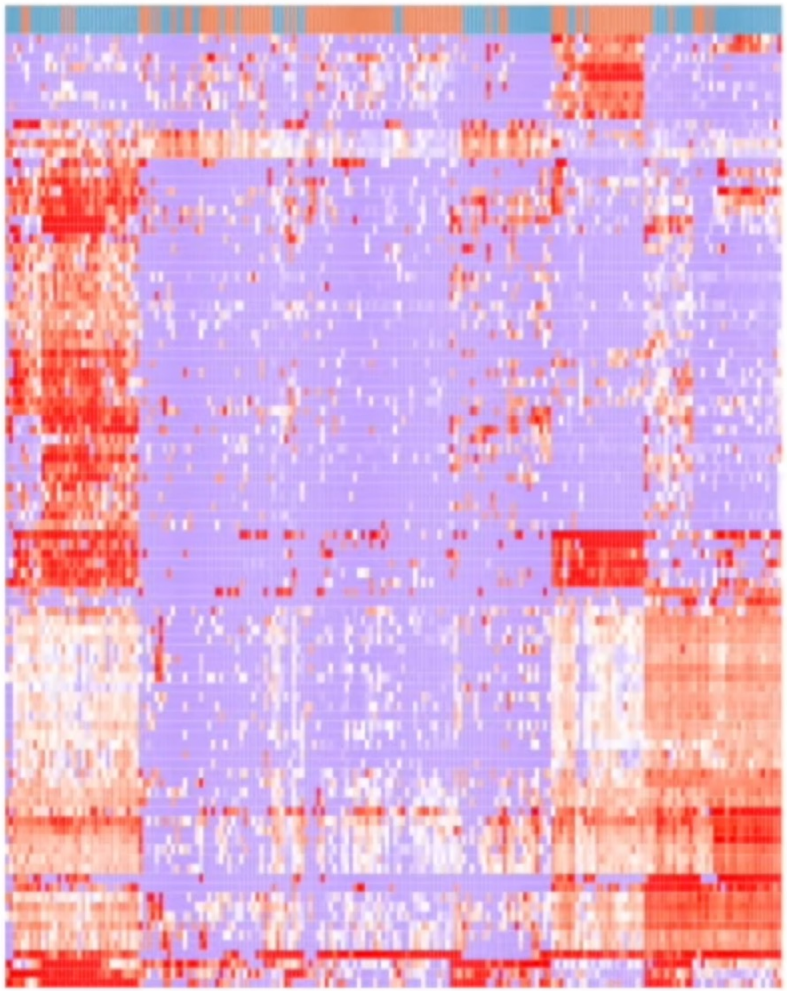
Hierarchical Clustering

Phd. César Astudillo | Facultad de Ingeniería

Introduction

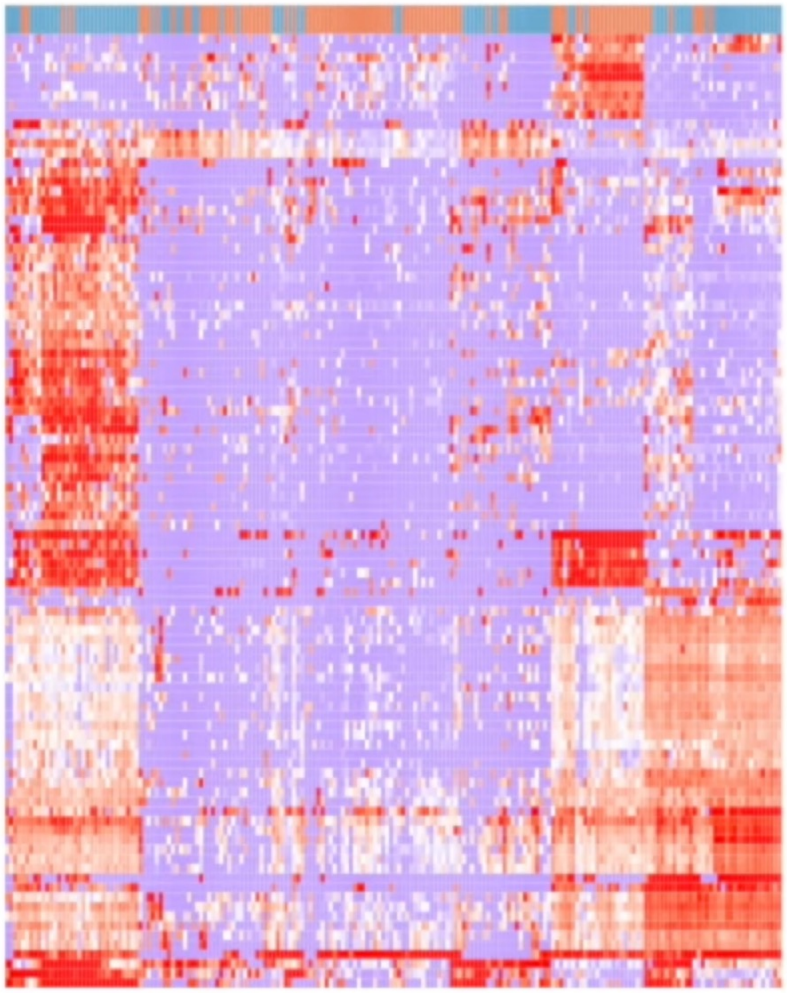
Hierarchical Clustering

The columns represent different samples.

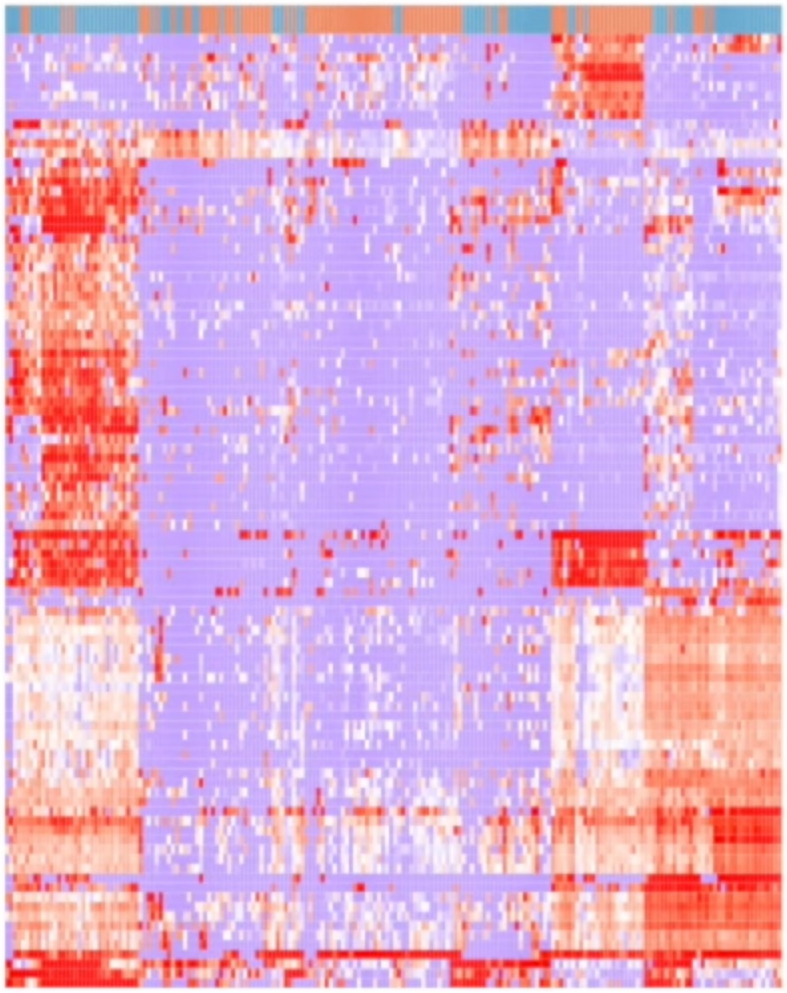


The columns represent different samples.

The rows represent measurements from different genes.



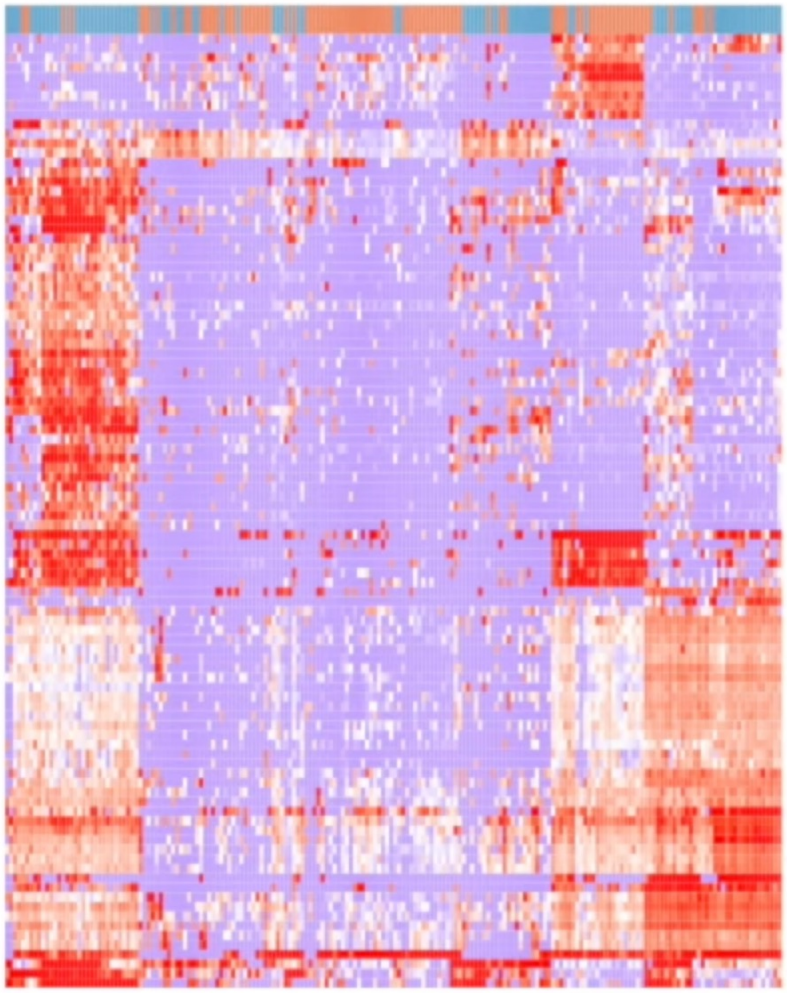
The columns represent different samples.



The rows represent measurements from different genes.

Hierarchical clustering orders the rows and/or the columns based on similarity.

The columns represent different samples.

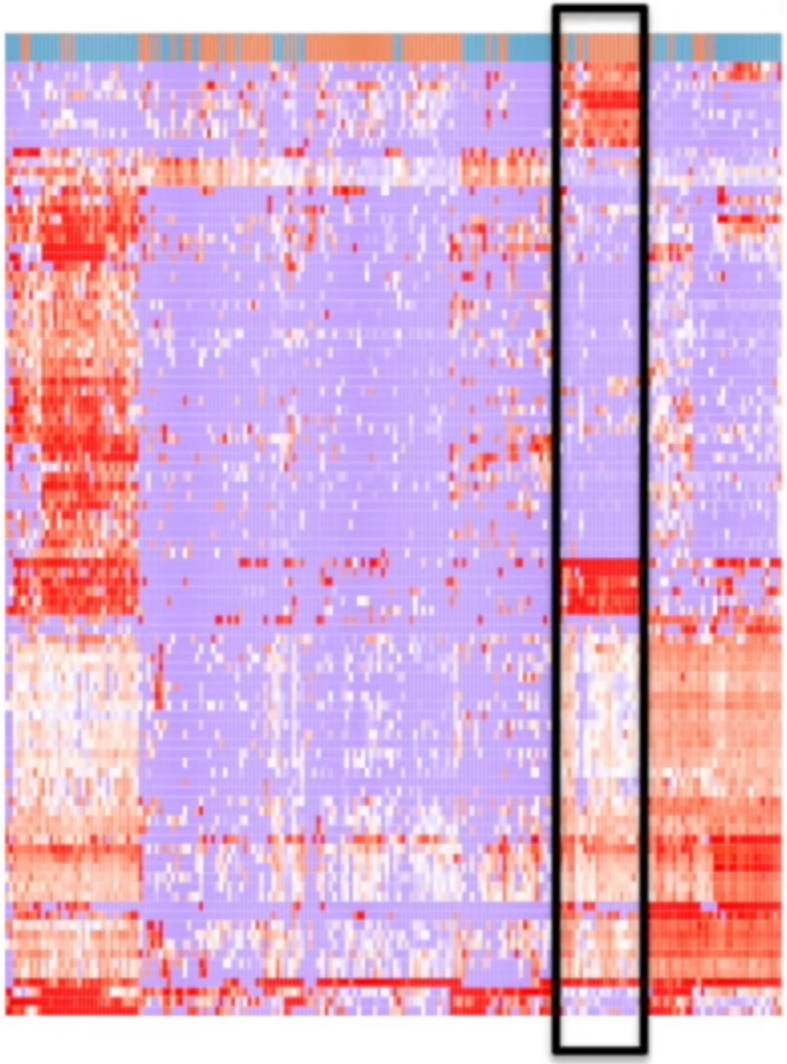


The rows represent measurements from different genes.

Hierarchical clustering orders the rows and/or the columns based on similarity.

This makes it easy to see correlations in the data.

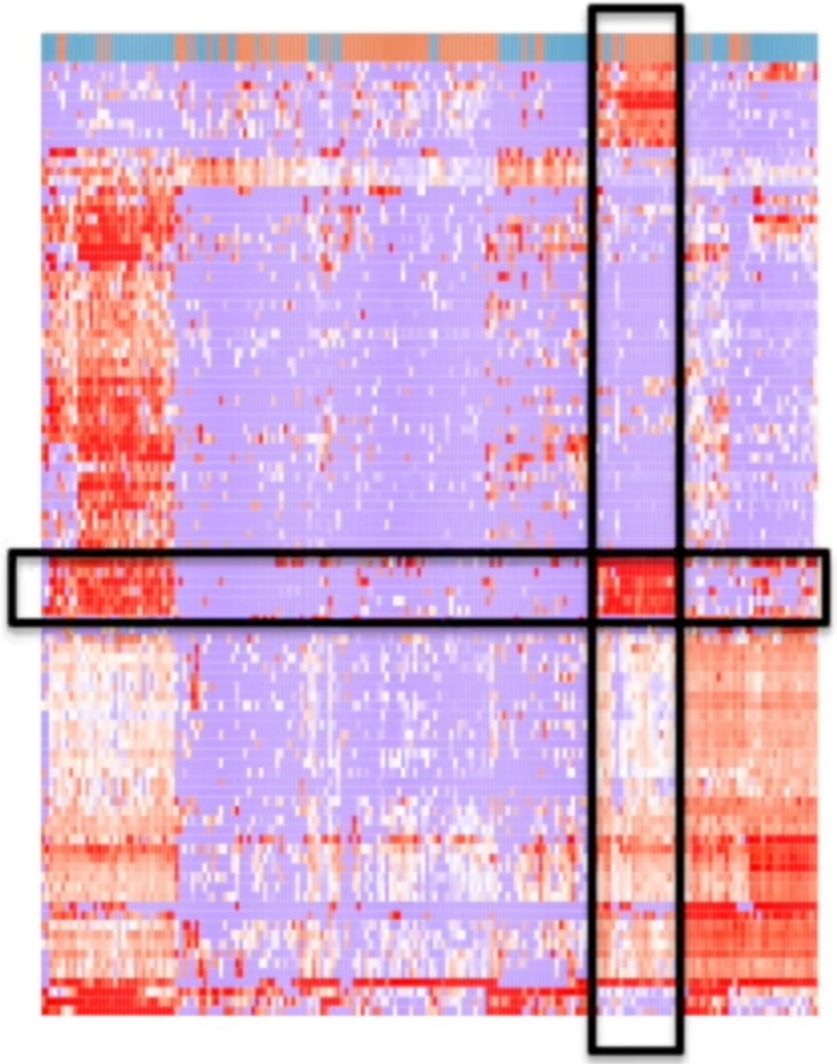
These samples express the same genes



Hierarchical clustering orders the rows and/or the columns based on similarity.

This makes it easy to see correlations in the data.

These samples express the same genes



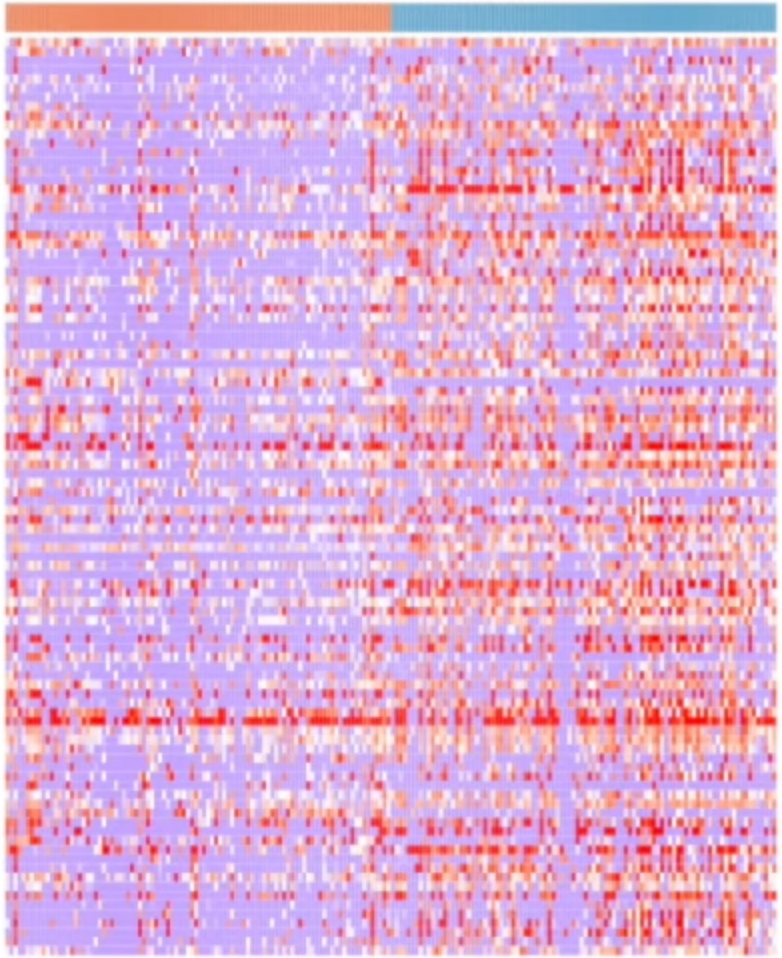
These genes behave the same.

Hierarchical clustering orders the rows and/or the columns based on similarity.

This makes it easy to see correlations in the data.

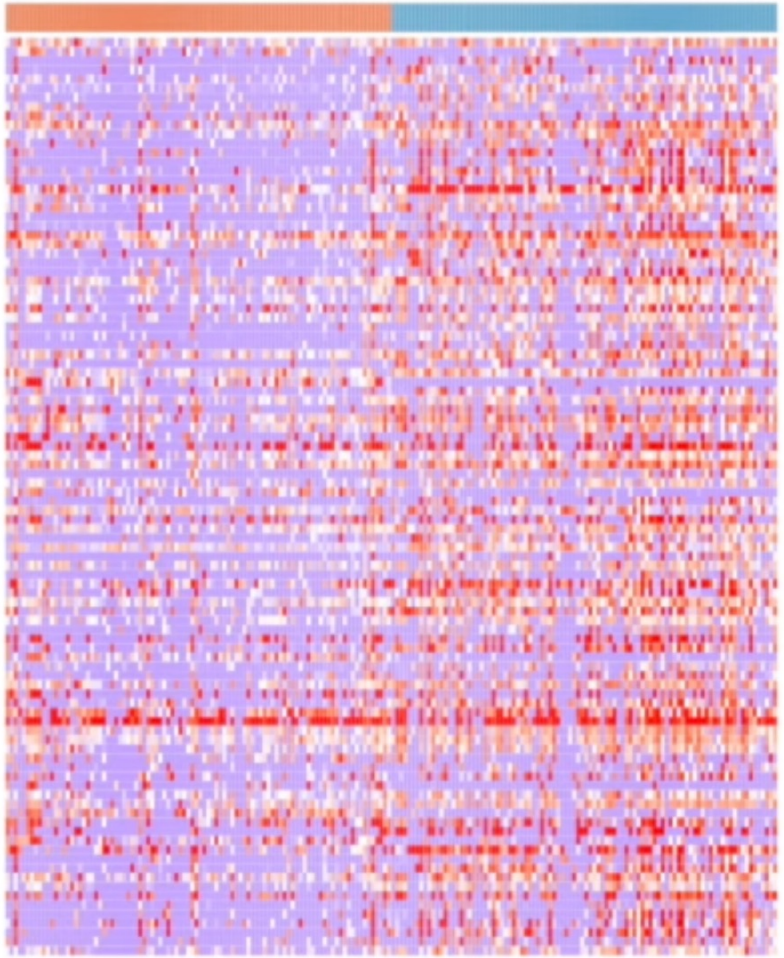
Hierarchical clustering is often associated with heatmaps.

Without hierarchical clustering…

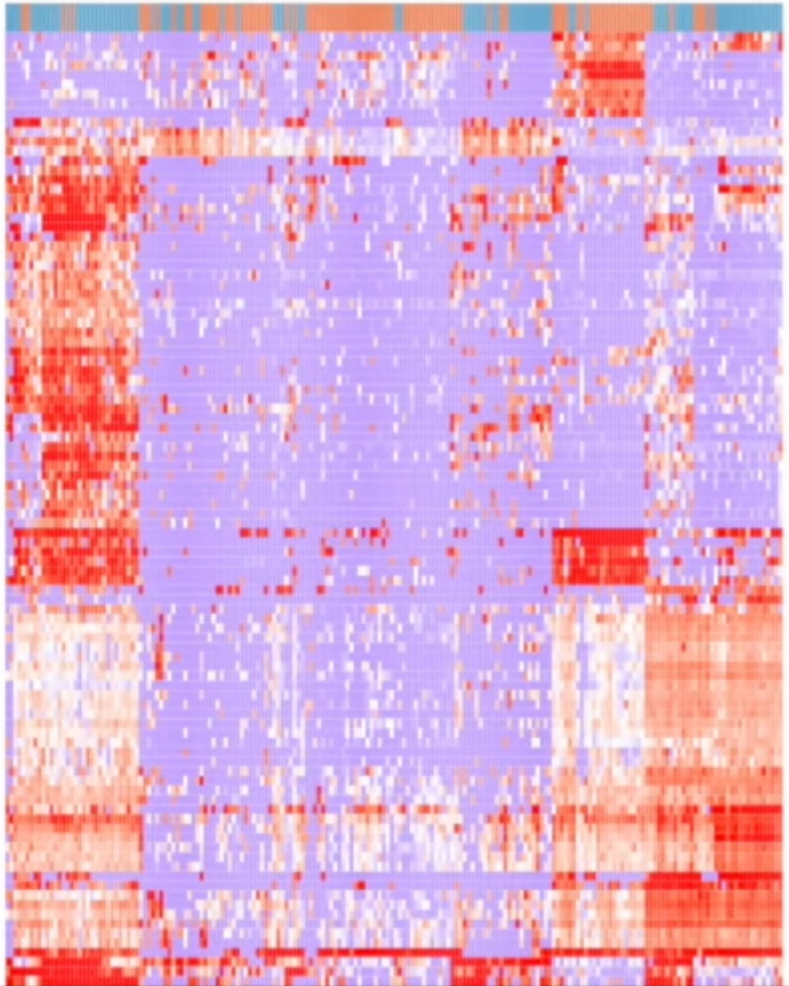


Hierarchical clustering is often associated with heatmaps.

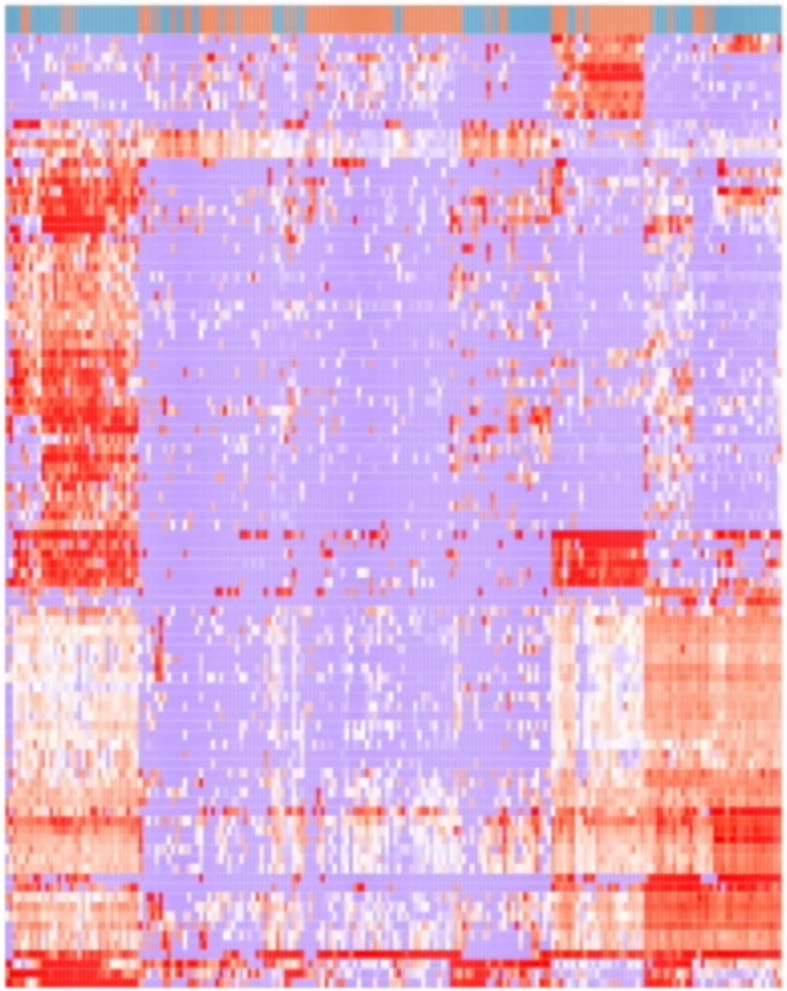
Without hierarchical clustering…

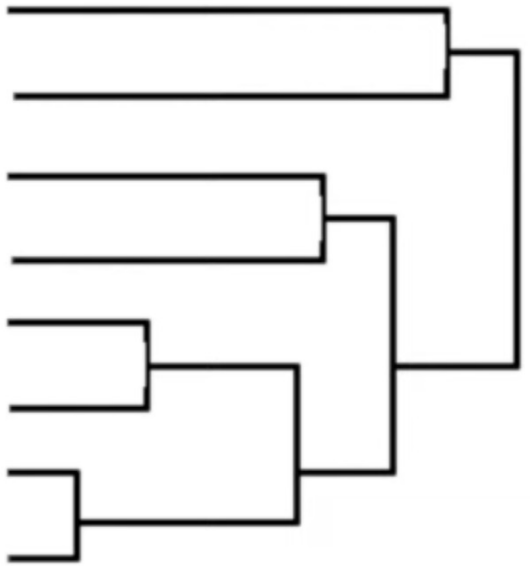


…with hierarchical clustering

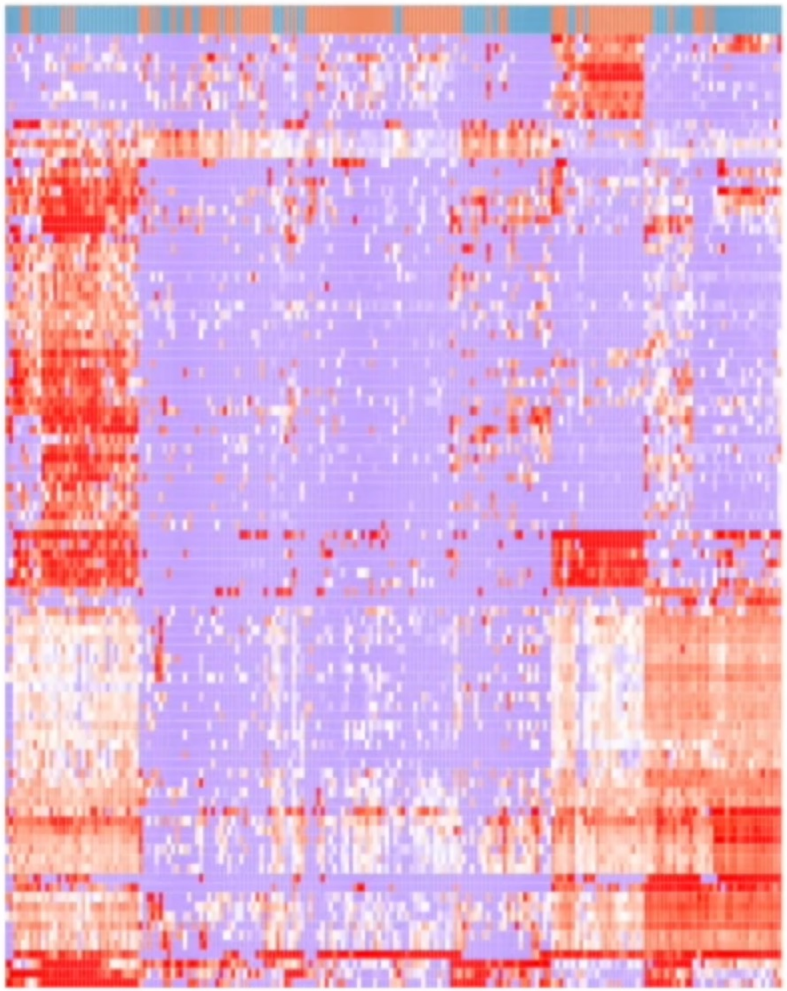


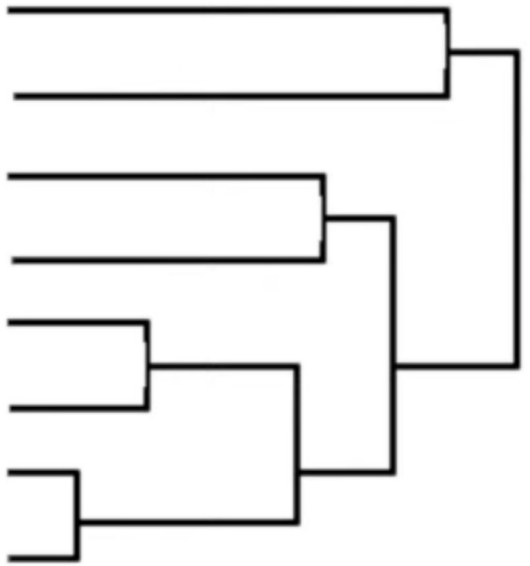
Heatmaps often come with dendrograms…





Heatmaps often come with dendrograms…

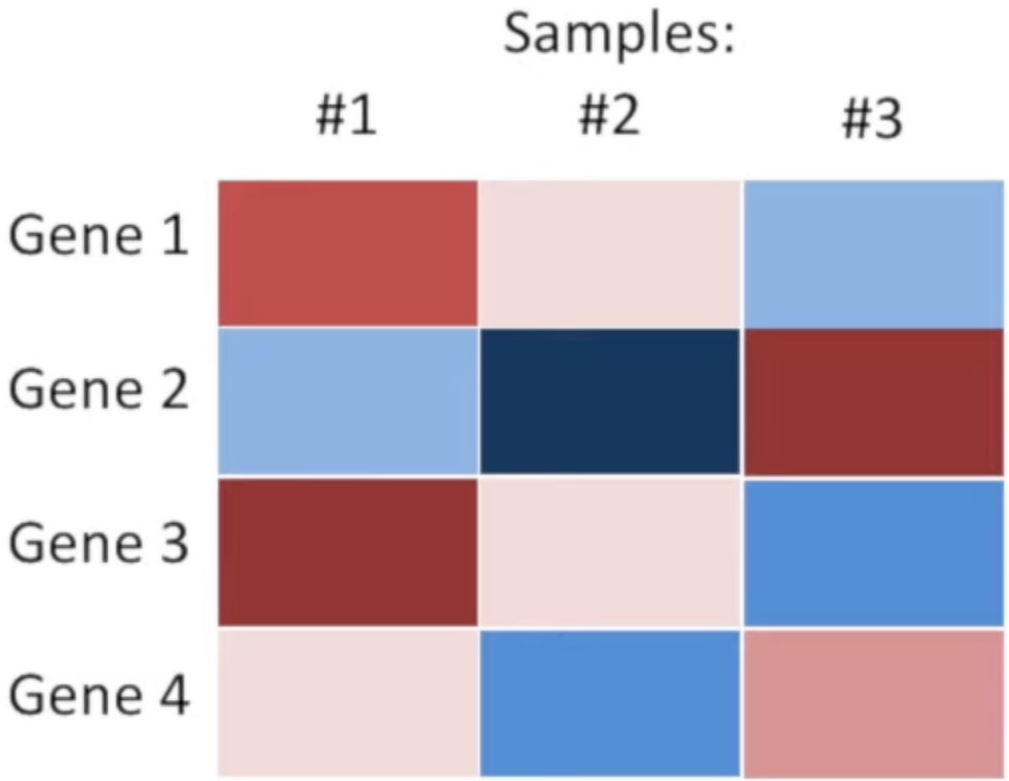


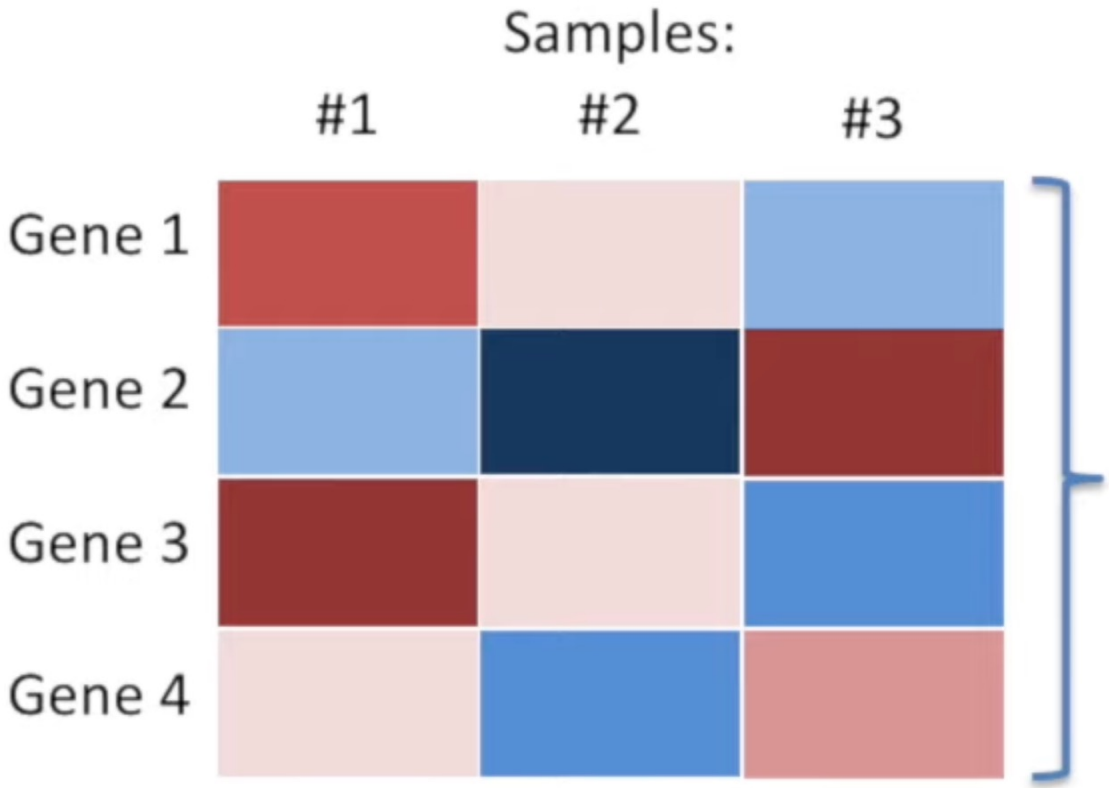


So we'll talk about those, too!

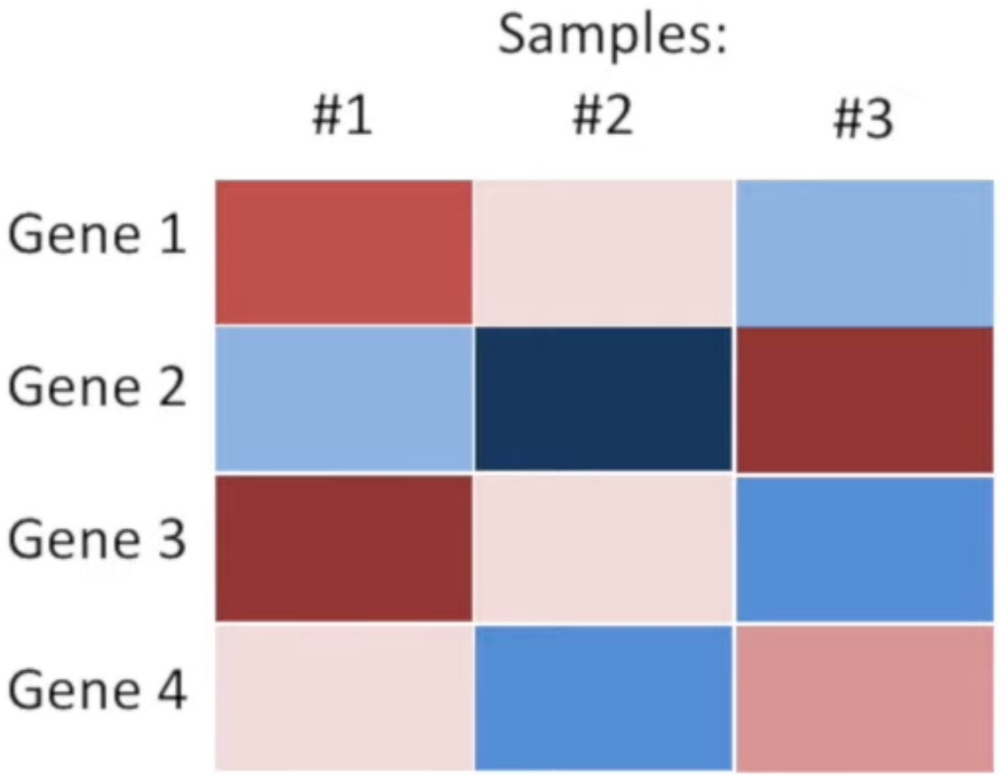
Heatmap Example

Hierarchical Clustering


Samples:
#1
#2
#3
Gene 1
Gene 2
Gene 3
Gene 4

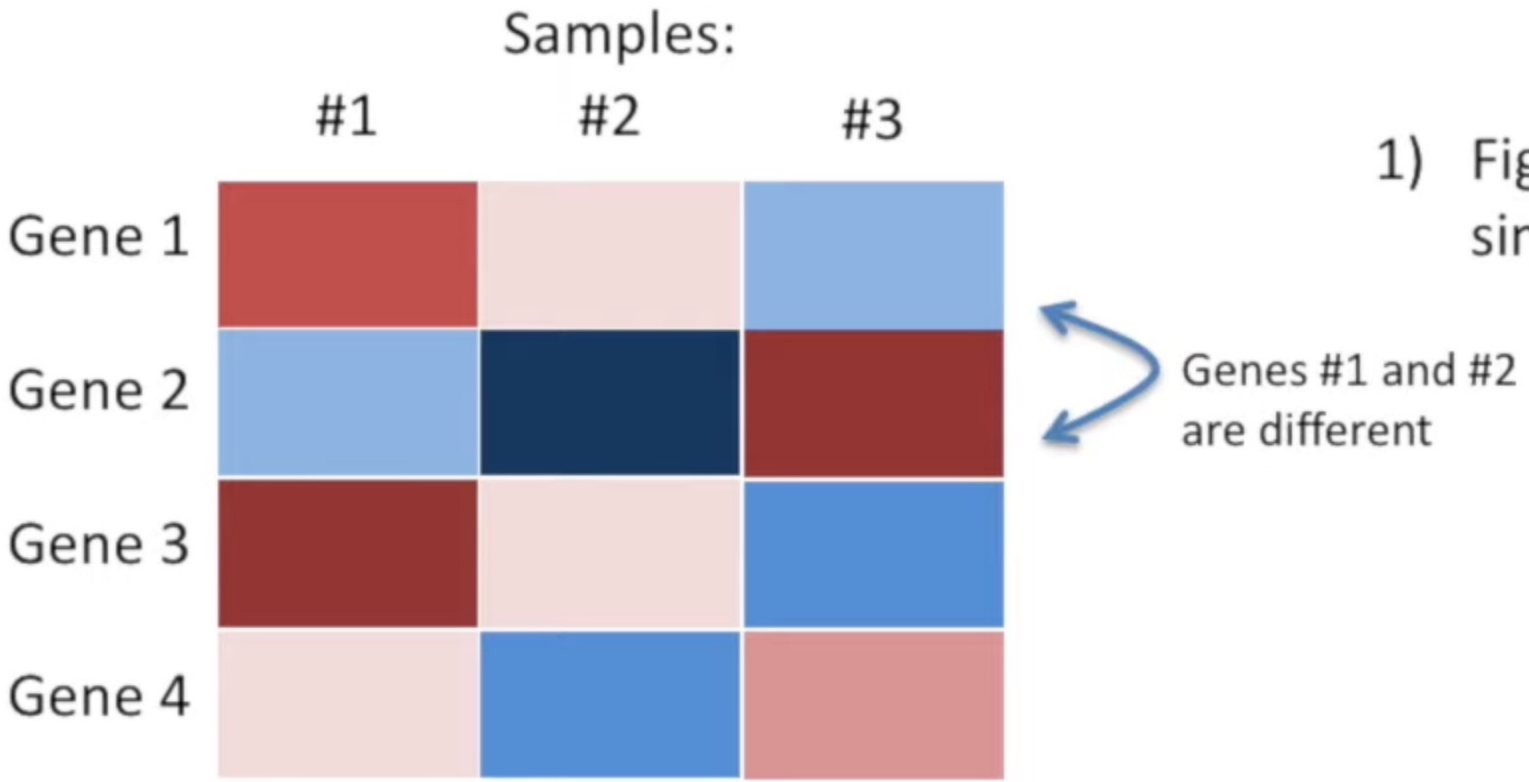


For this example, we are just going to cluster (reorder) the rows (genes).



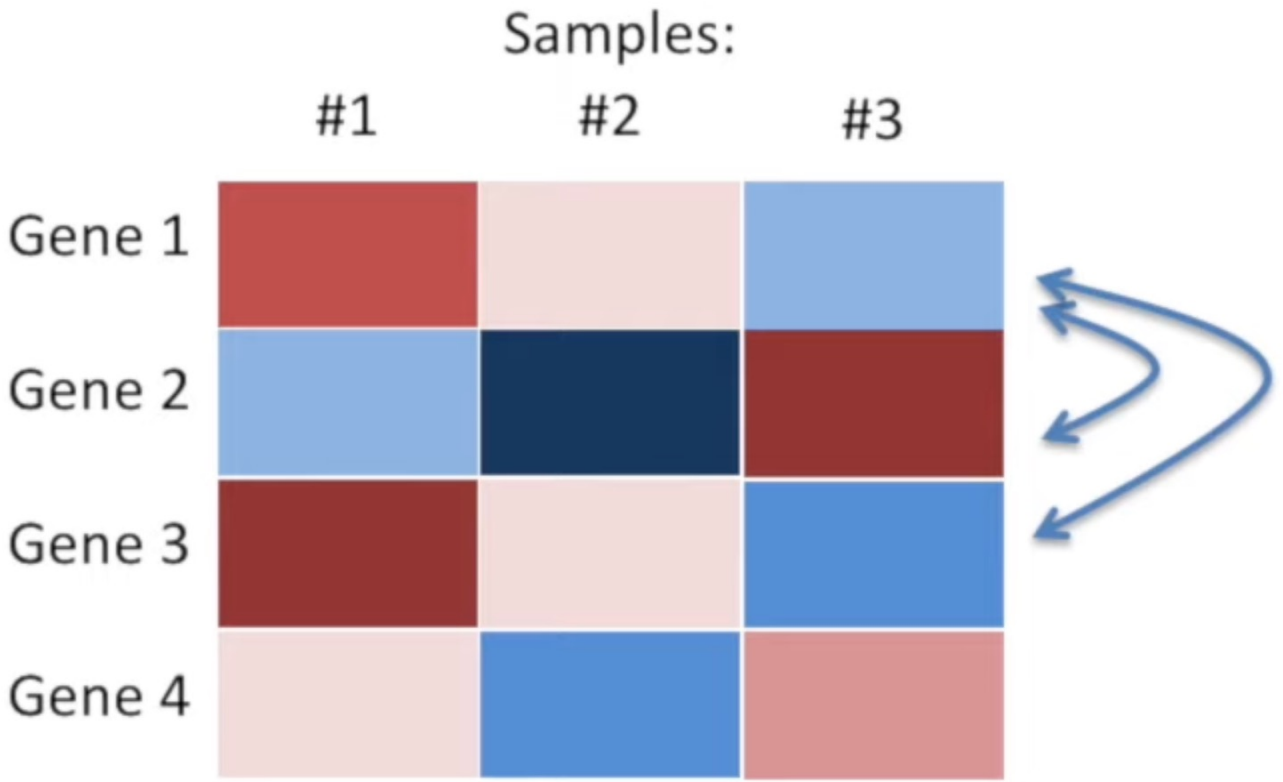
Conceptually…

1) Figure out which gene is most similar to gene #1.



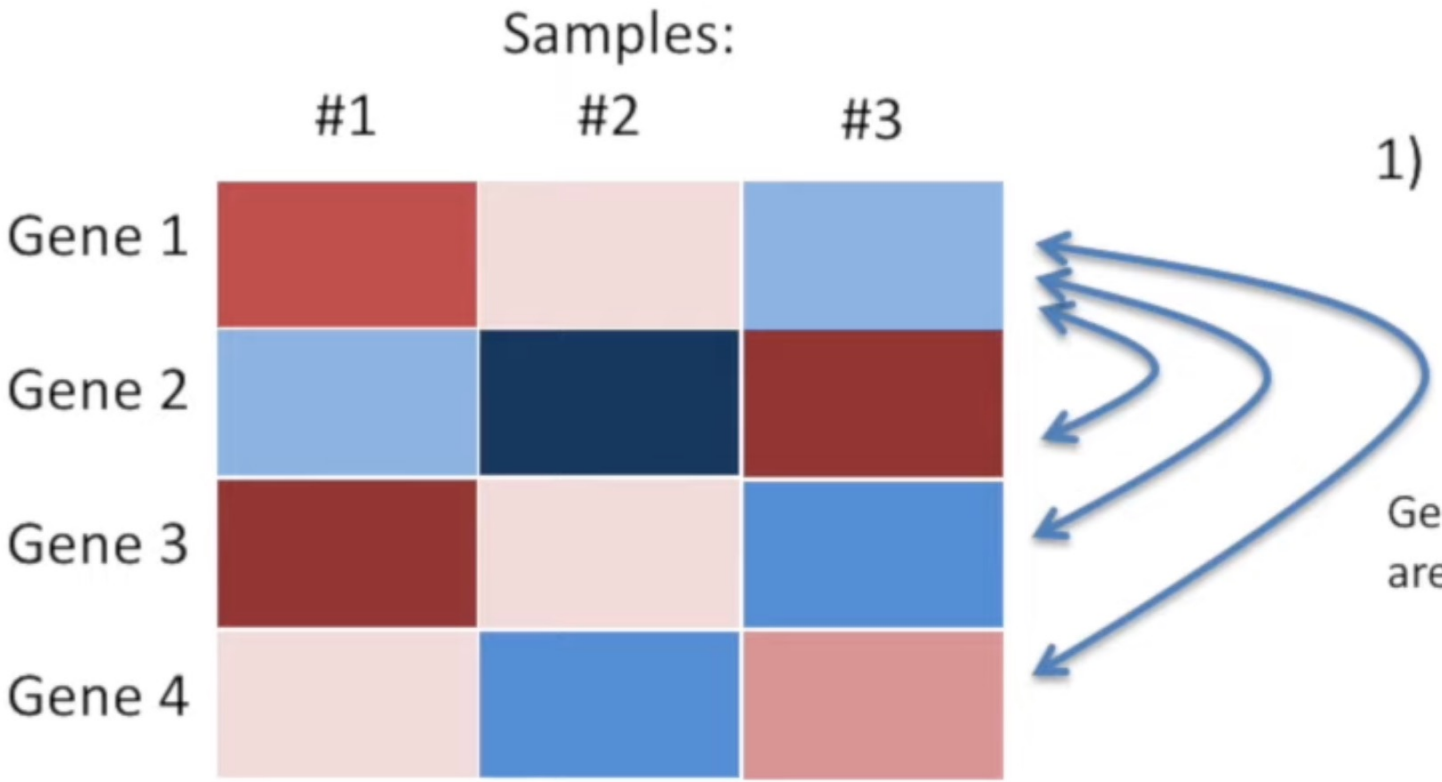
Conceptually…

1) Figure out which gene is most similar to gene #1.



Conceptually…

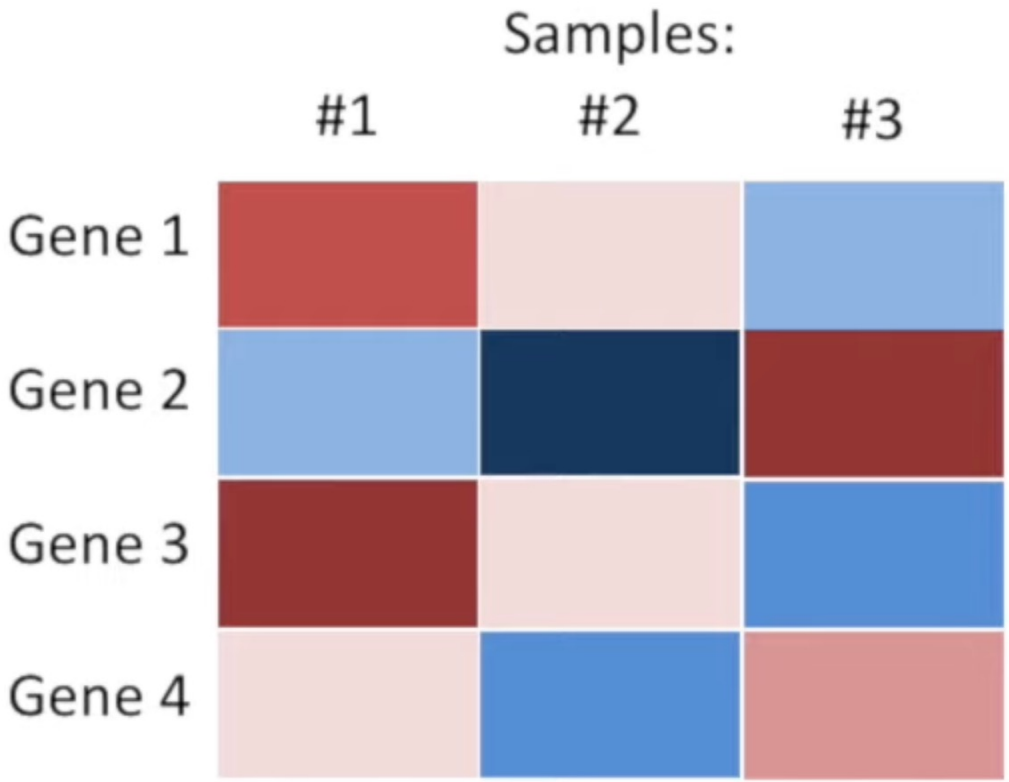
1) Figure out which gene is most similar to gene #1.



Conceptually…

1) Figure out which gene is most similar to gene #1.

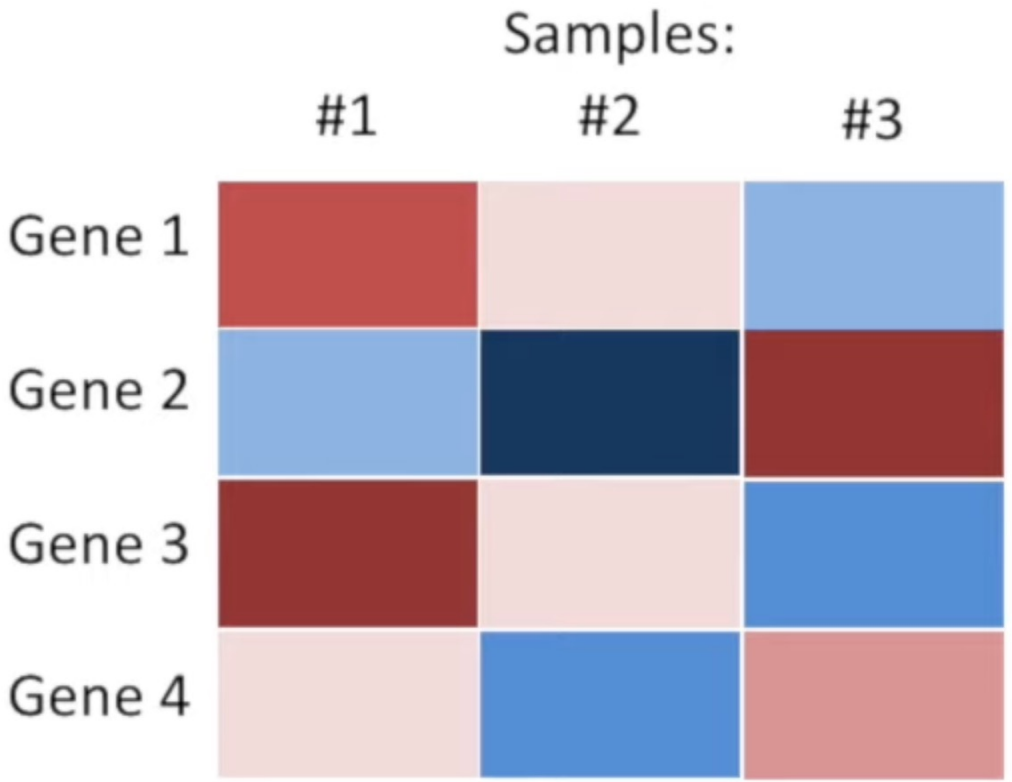
Genes #1 and #4 are similar.



Conceptually…

1) Figure out which gene is most similar to gene #1.

2) Figure out which genes is most similar to gene #2...

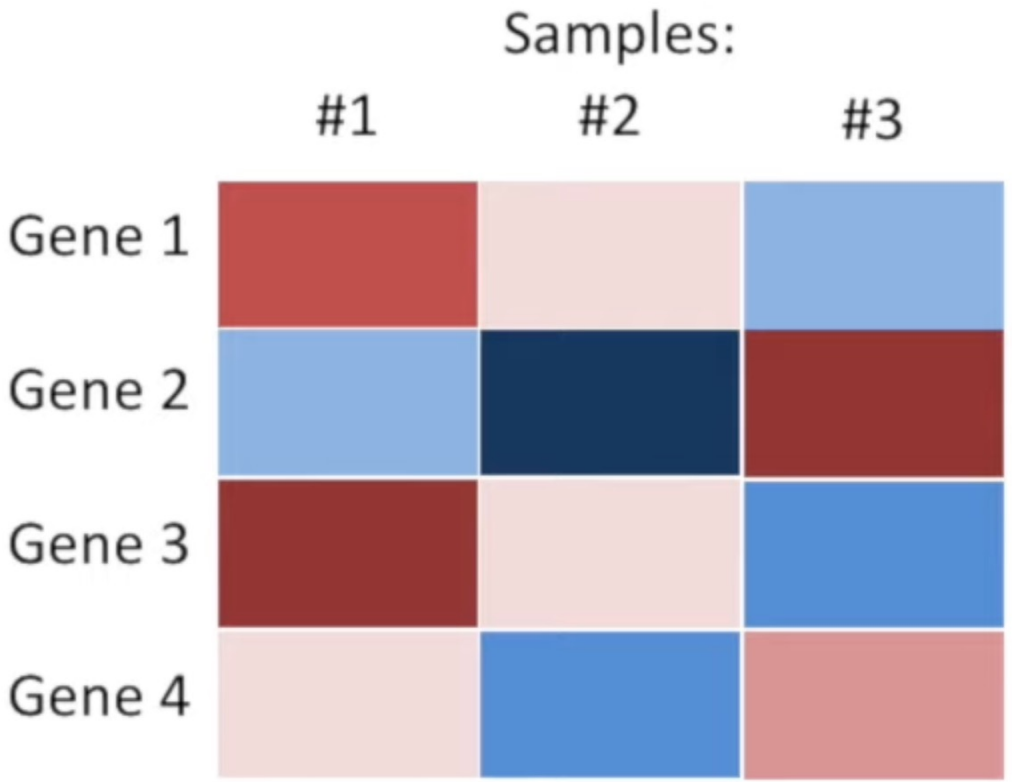


Conceptually…

1) Figure out which gene is most similar to gene #1.

2) Figure out which genes is most similar to gene #2...

Gene #2 is most similar to gene #4

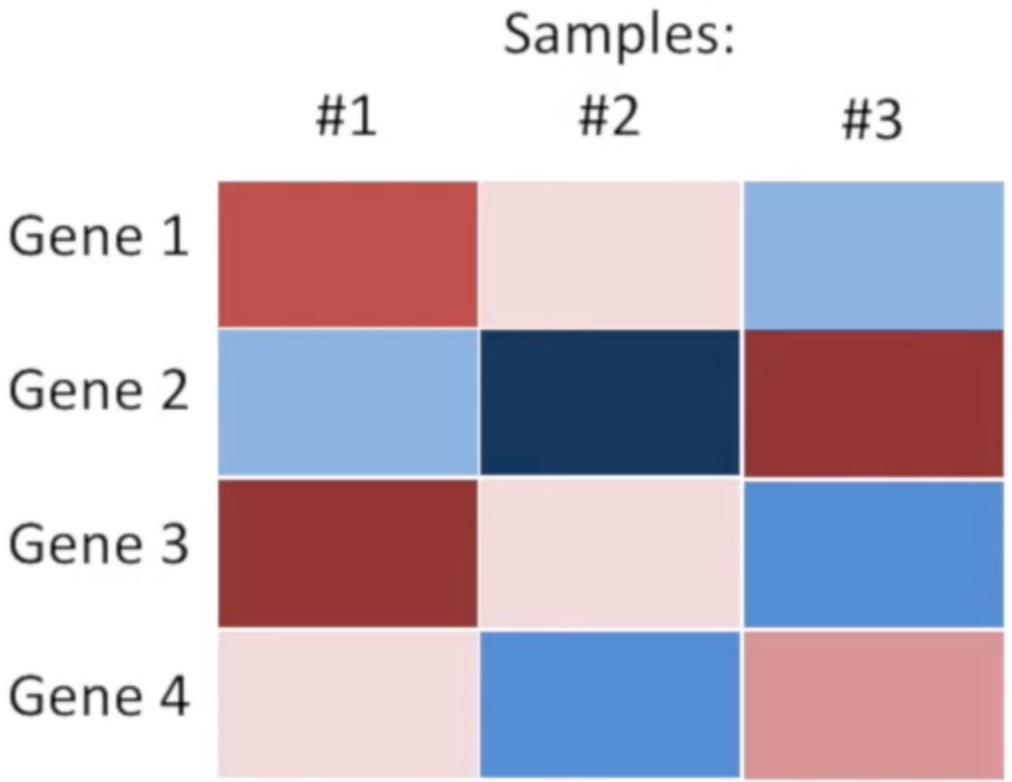


Conceptually…

1) Figure out which gene is most similar to gene #1.

2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).

Gene #2 is most similar to gene #4

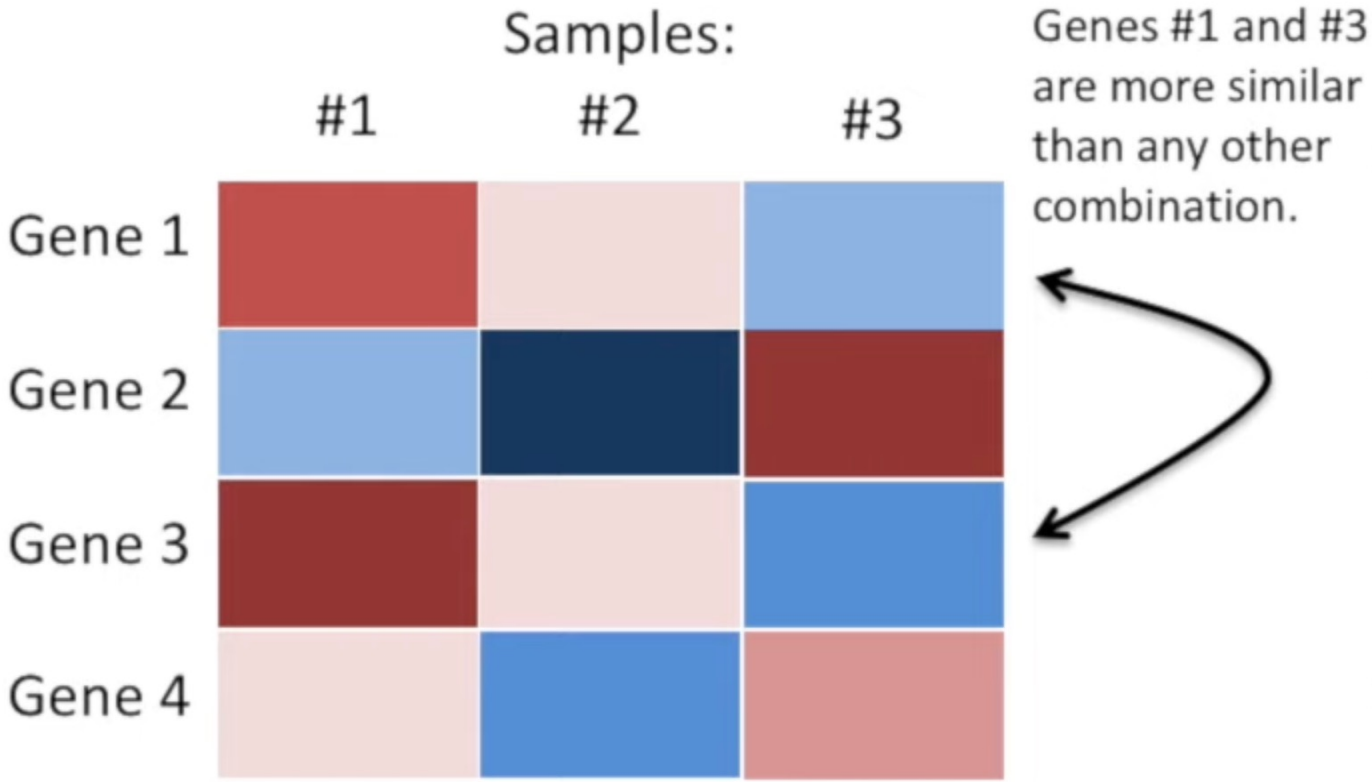


Conceptually…

1) Figure out which gene is most similar to gene #1.

2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).

3) Of the different combinations, figure out which two genes are the most similar. Merge them into a cluster.



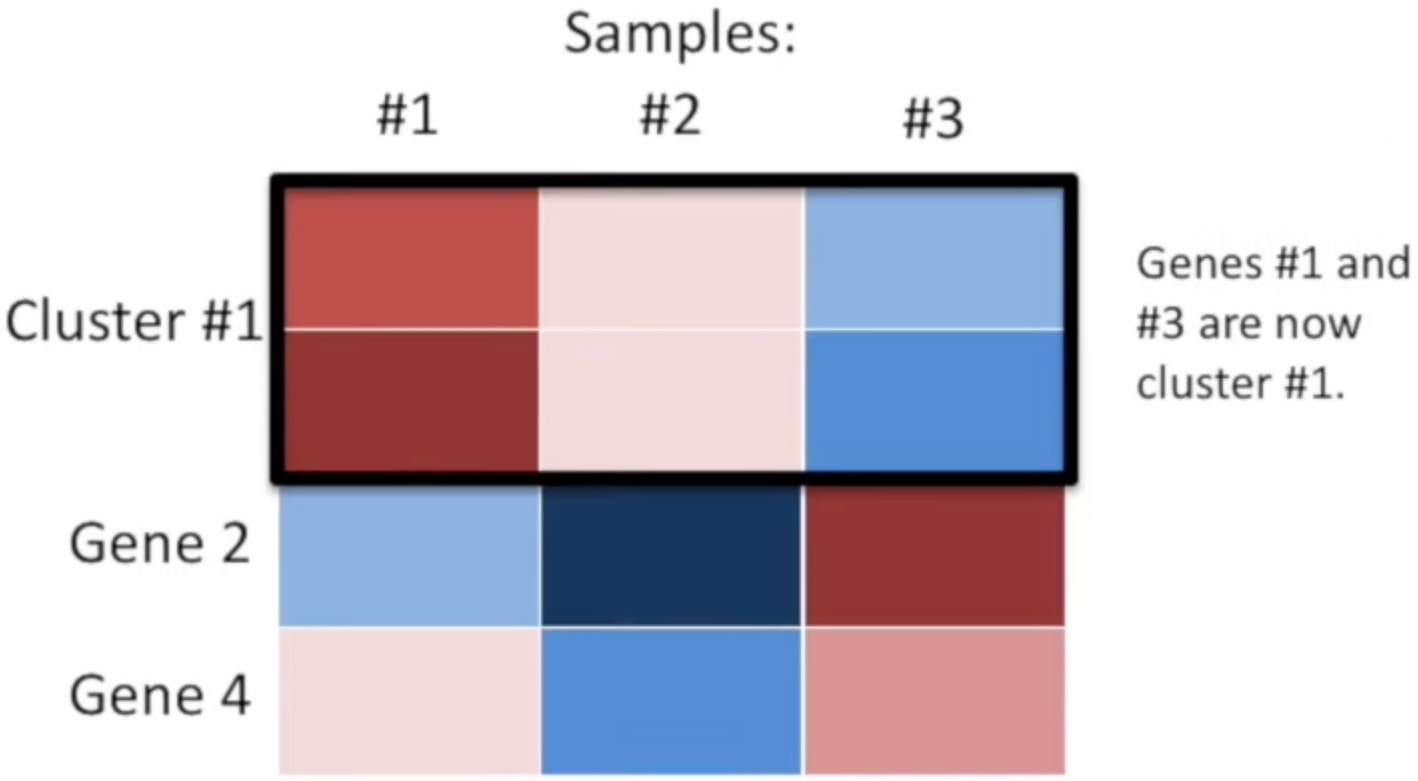
Genes #1 and #3 are more similar than any other combination.

Conceptually…

1) Figure out which gene is most similar to gene #1.

2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).

3) Of the different combinations, figure out which two genes are the most similar. Merge them into a cluster.



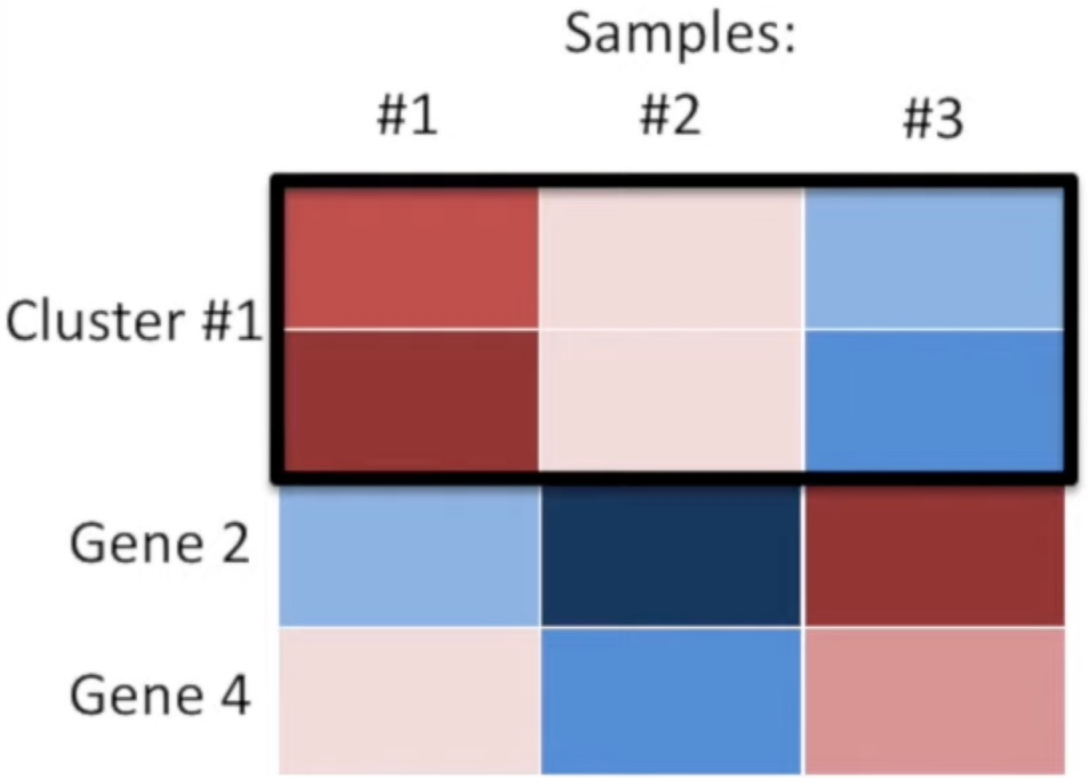
Genes #1 and #3 are now cluster #1.

Conceptually…

1) Figure out which gene is most similar to gene #1.

2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).

3) Of the different combinations, figure out which two genes are the most similar. Merge them into a cluster.



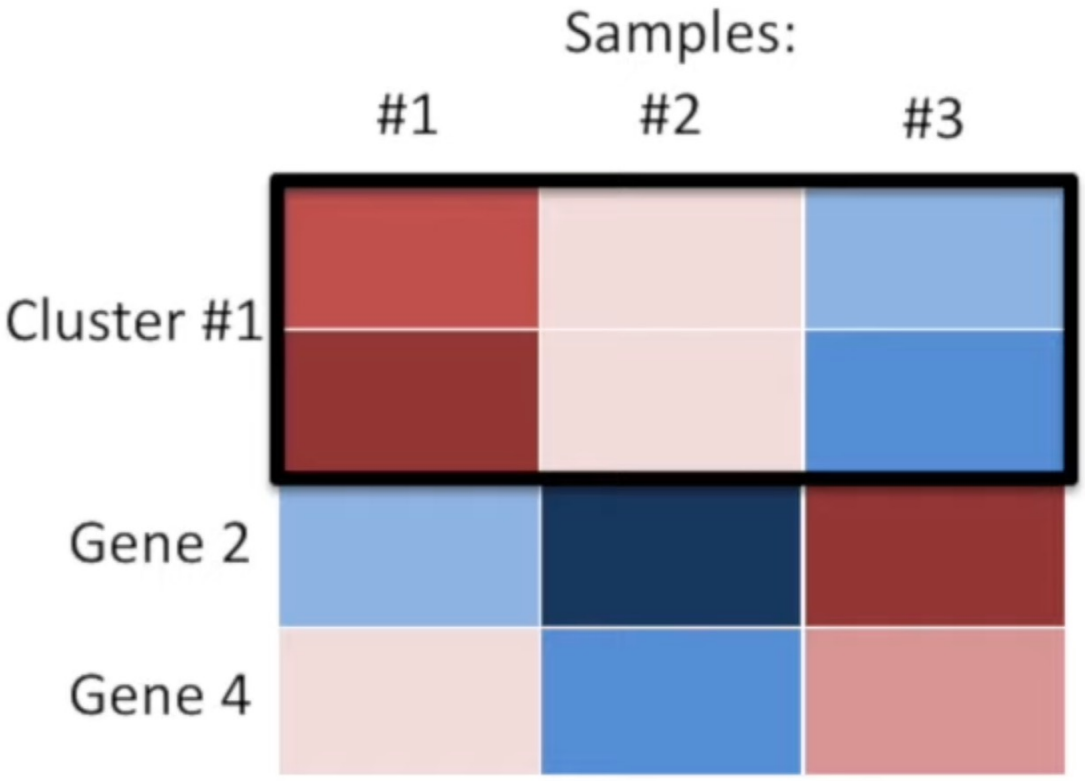
Conceptually…

1) Figure out which gene is most similar to gene #1.

2) Figure out which genes is most similar to gene #2… (and then #3 and then #4).

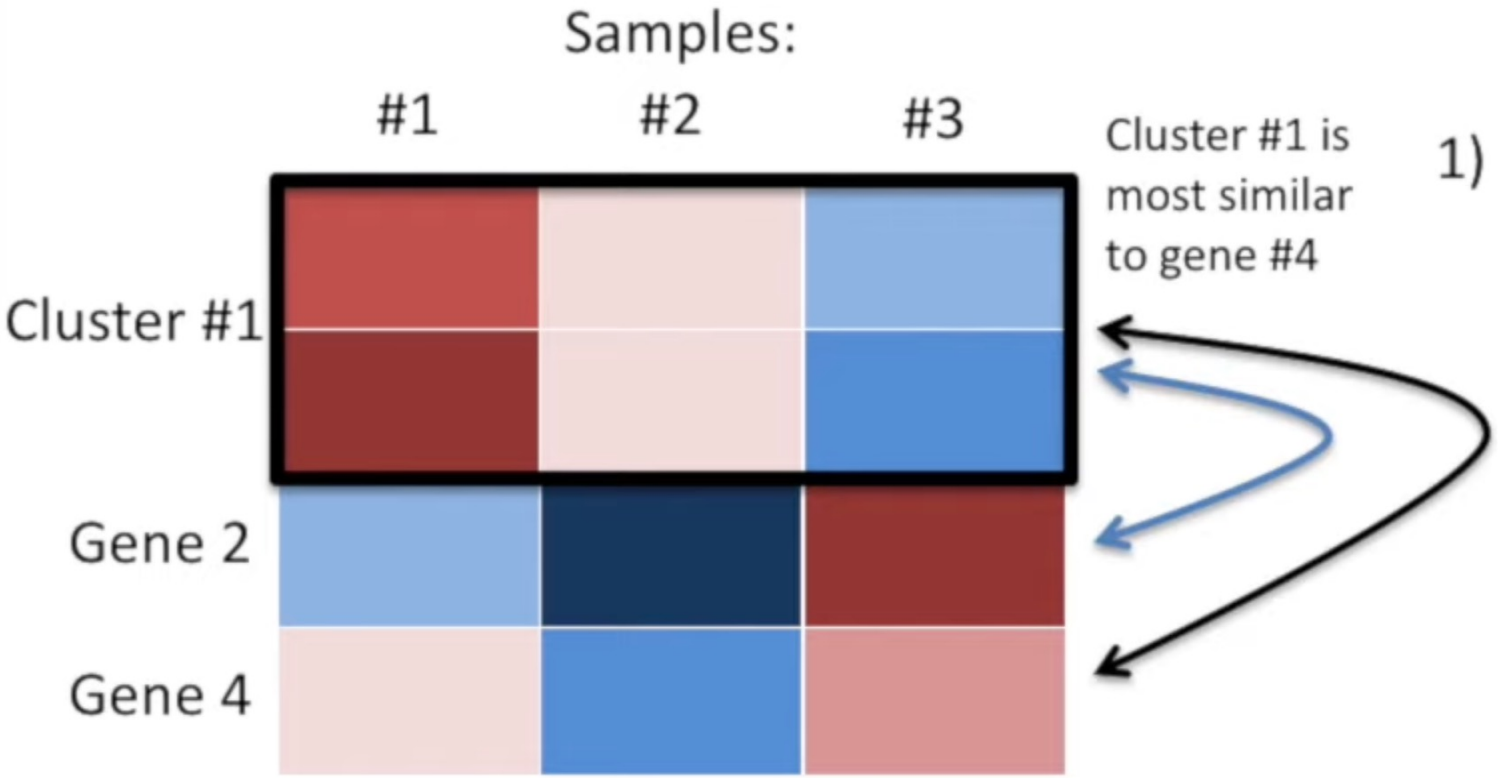
3) Of the different combinations, figure out which two genes are the most similar. Merge them into a cluster.

4) Go back to step 1, but now treat the new cluster like it's a single gene.



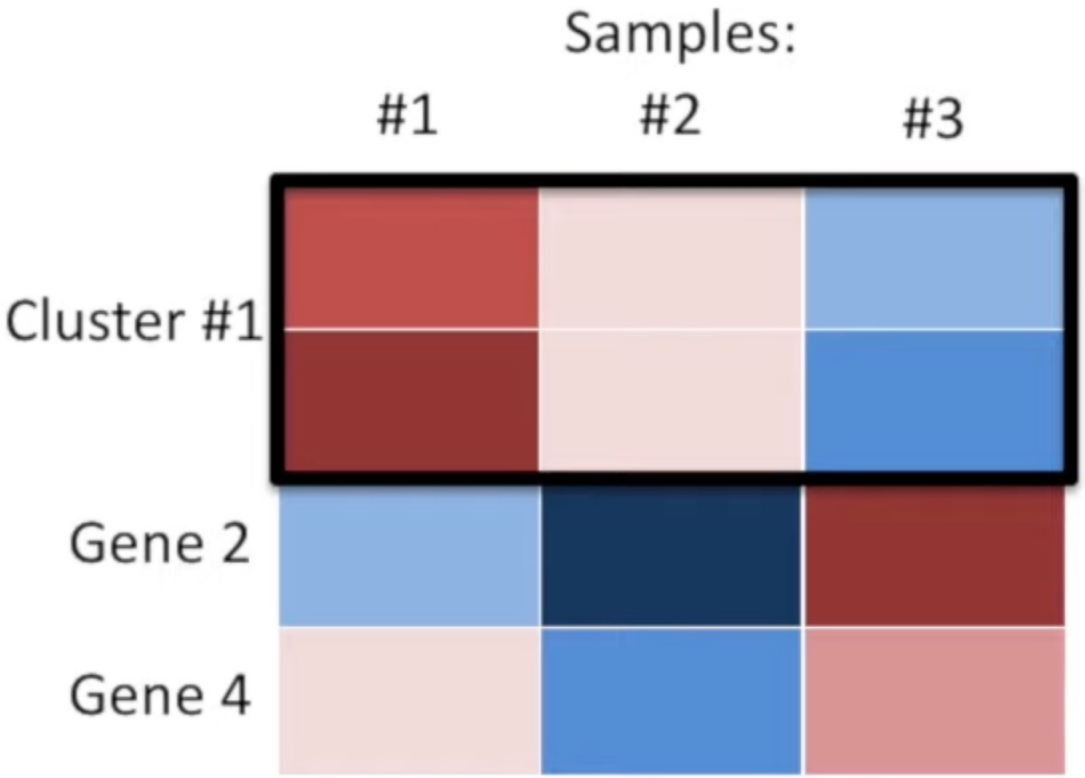
Conceptually…

1) Figure out which gene is most similar to cluster #1.



Conceptually…

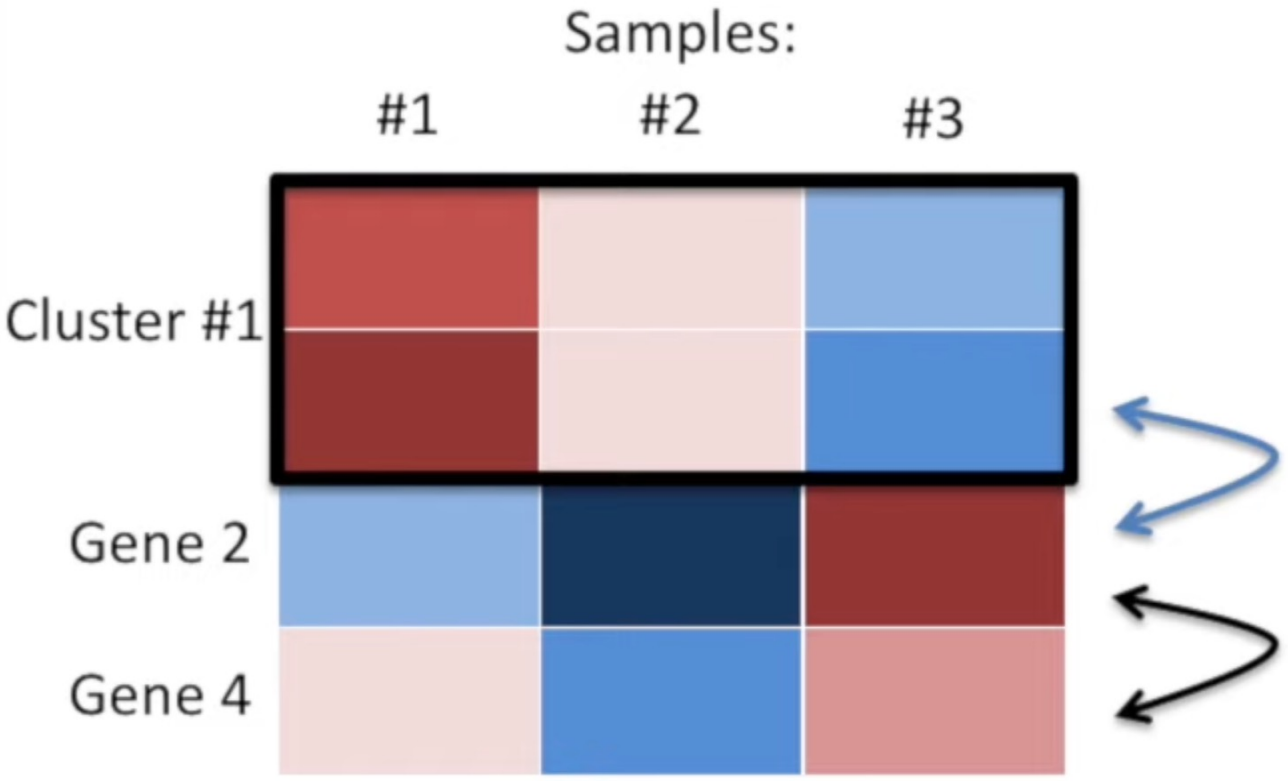
1) Figure out which gene is most similar to cluster #1.



Conceptually…

1) Figure out which gene is most similar to cluster #1.

2) Figure out which genes is most similar to gene #2...

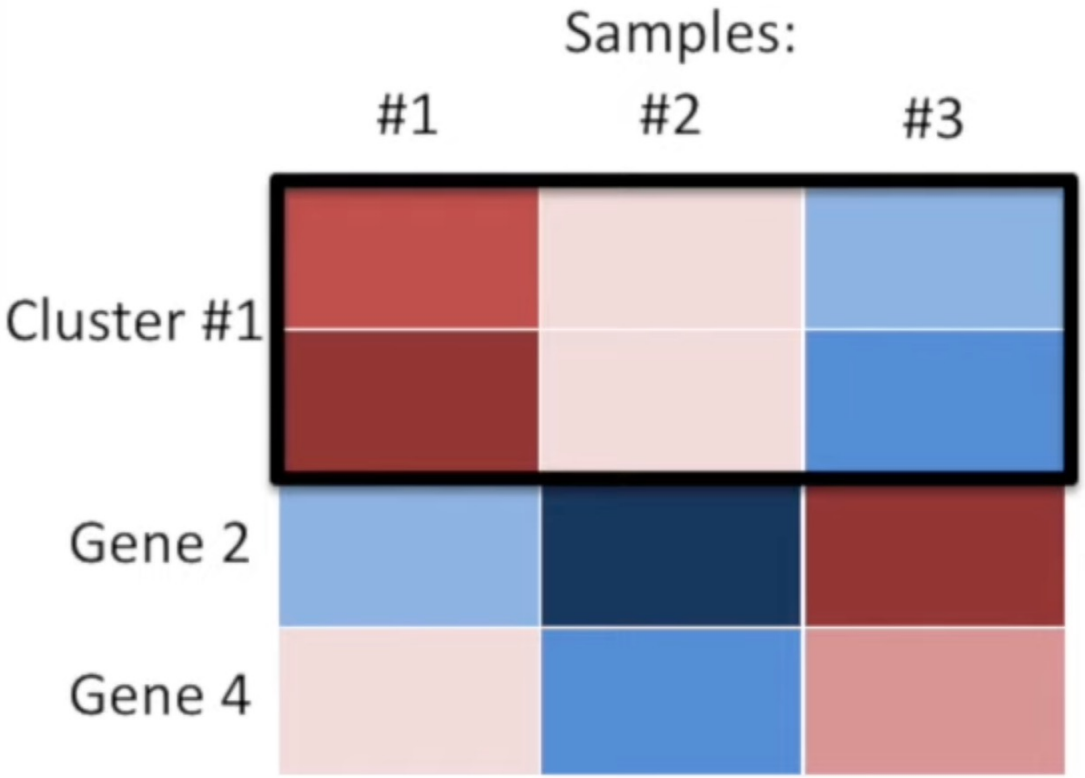


Gene #2 is most similar to gene #4, but notice that we compared gene #2 to cluster #1

Conceptually…

1) Figure out which gene is most similar to cluster #1.

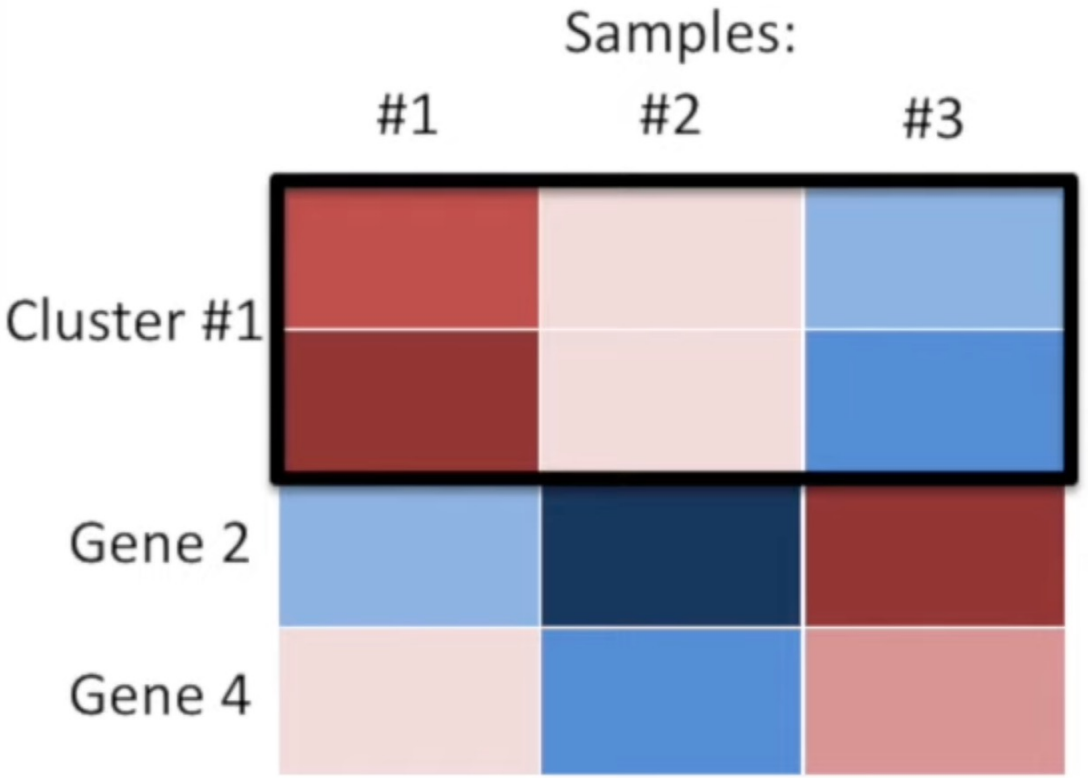
2) Figure out which genes is most similar to gene #2...



Conceptually…

1) Figure out which gene is most similar to cluster #1.

2) Figure out which genes is most similar to gene #2... (and then #4).

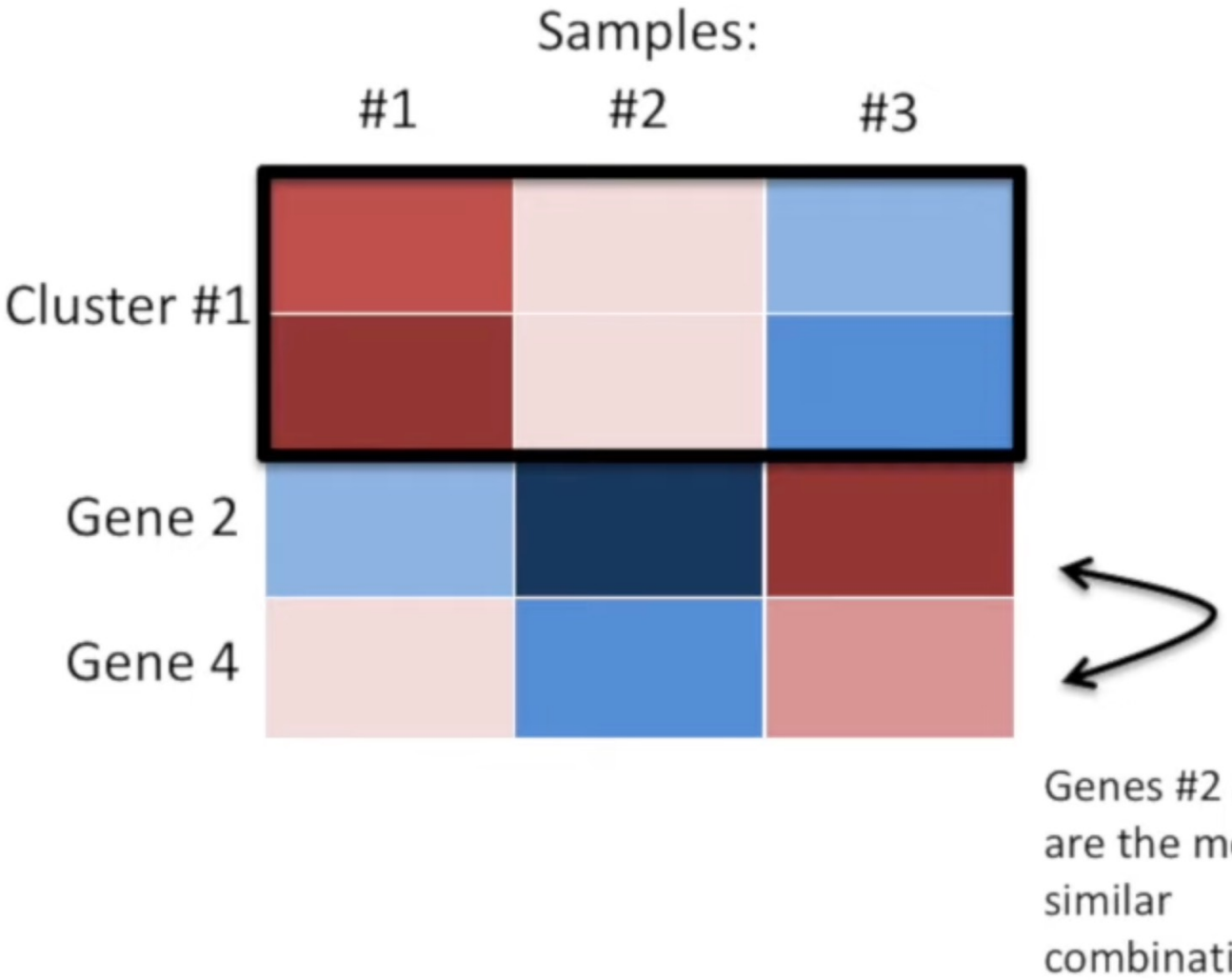


Conceptually…

1) Figure out which gene is most similar to cluster #1.

2) Figure out which genes is most similar to gene #2... (and then #4).

3) Of the different combinations, figure out which two genes are the most similar. Merge them into a cluster.



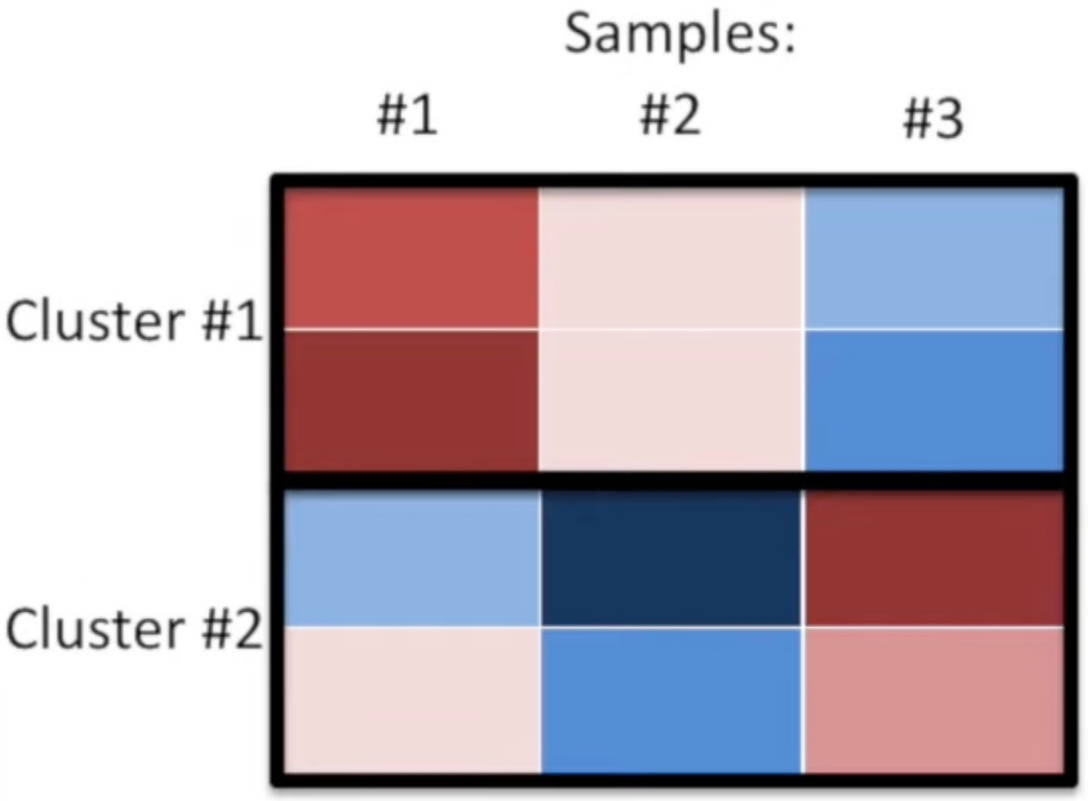
Genes #2 and #4 are the most similar combination.

Conceptually…

1) Figure out which gene is most similar to cluster #1.

2) Figure out which genes is most similar to gene #2… (and then #4).

3) Of the different combinations, figure out which two genes are the most similar. Merge them into a cluster.

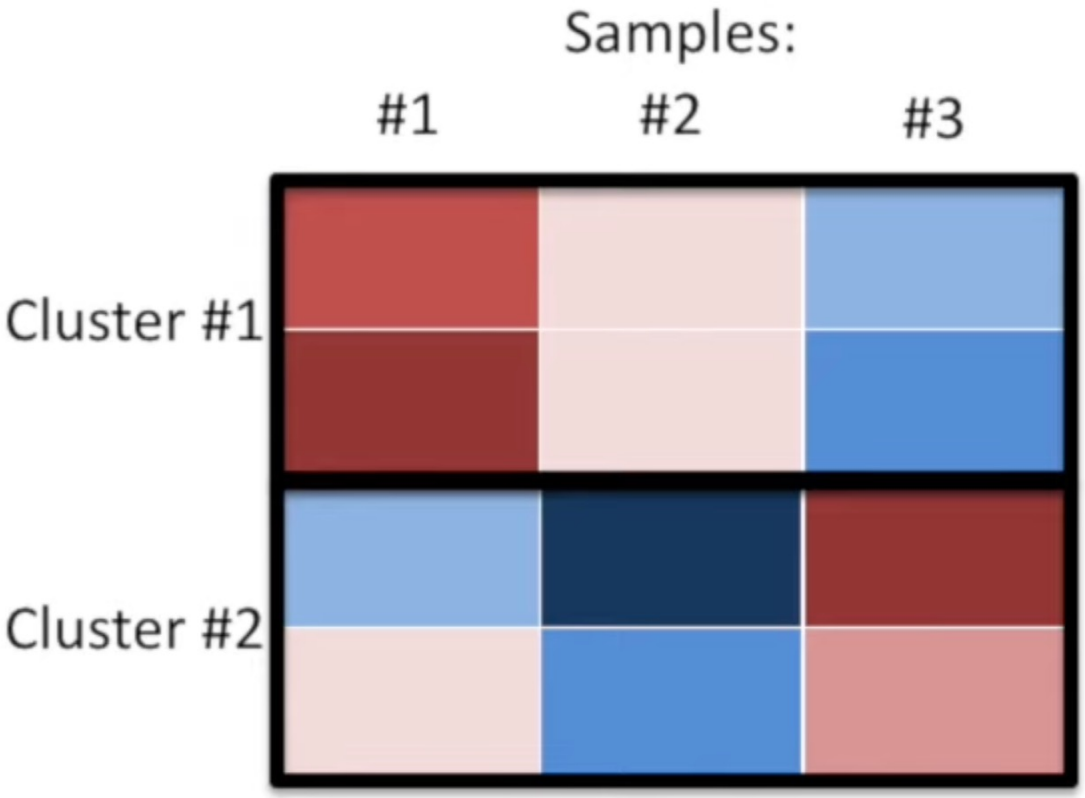


Conceptually…

1) Figure out which gene is most similar to cluster #1.

2) Figure out which genes is most similar to gene #2... (and then #4).

3) Of the different combinations, figure out which two genes are the most similar. Merge them into a cluster.



Conceptually…

1) Figure out which gene is most similar to cluster #1.

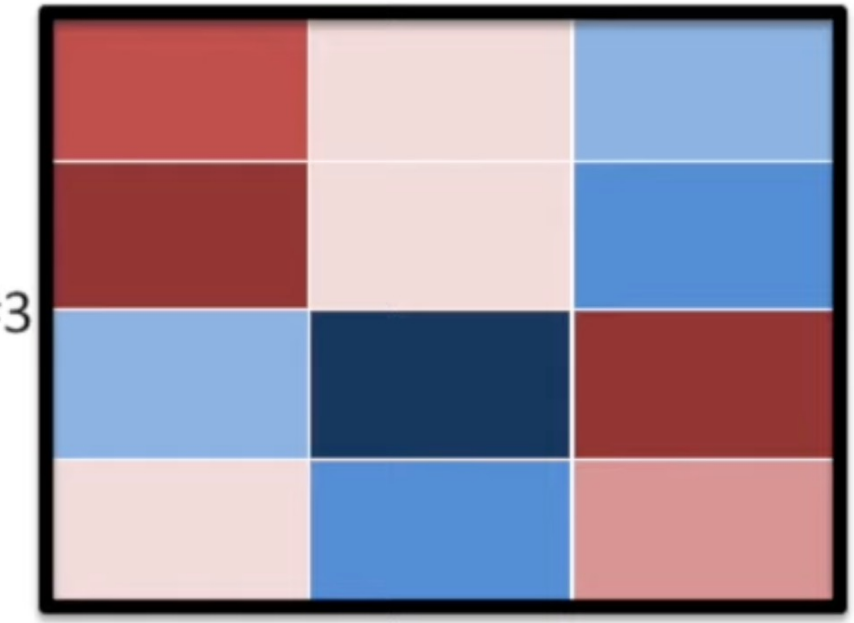
2) Figure out which genes is most similar to gene #2... (and then #4).

3) Of the different combinations, figure out which two genes are the most similar. Merge them into a cluster.

4) Go back to step 1.

Samples:

#1 #2 #3



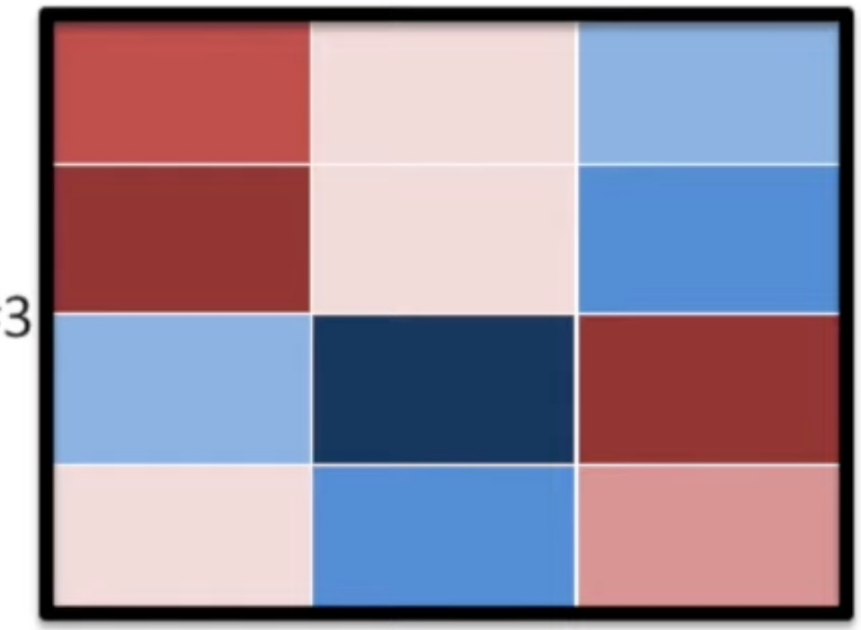
Cluster #3

Conceptually…

Since all we have left are 2 clusters, we merge them.

Samples:

#1 #2 #3



Cluster #3

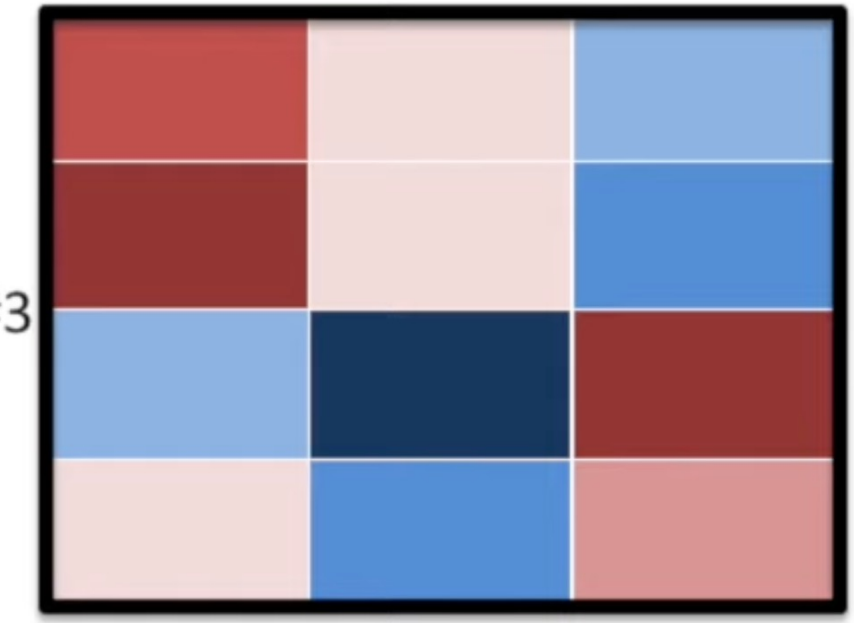
Conceptually…

Since all we have left are 2 clusters, we merge them.

BAM!!! We're all done!

Samples:

#1 #2 #3

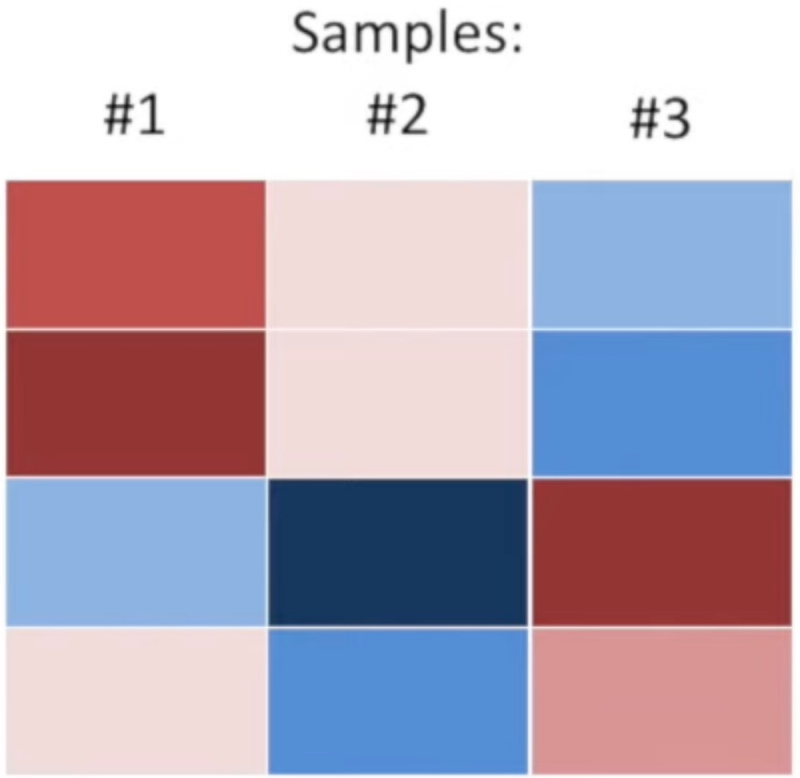


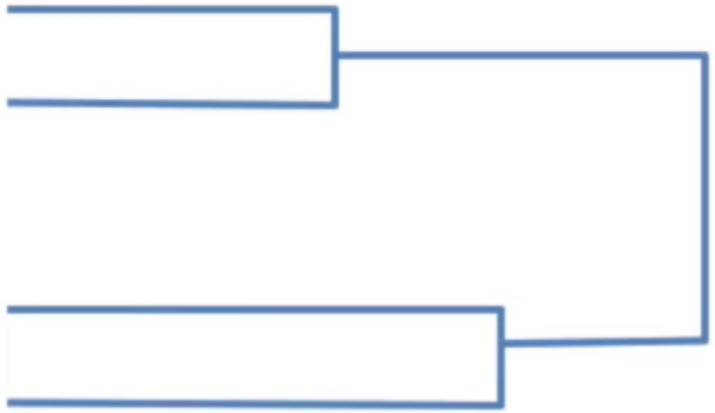
Cluster #3

Conceptually…

Since all we have left are 2 clusters, we merge them.

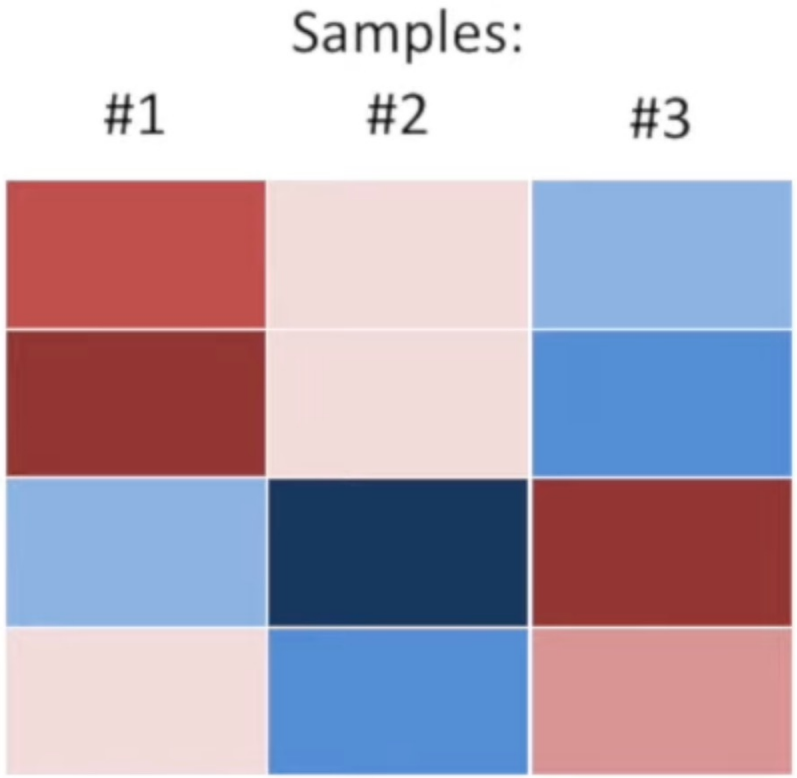
BAM!!! We’re all done!



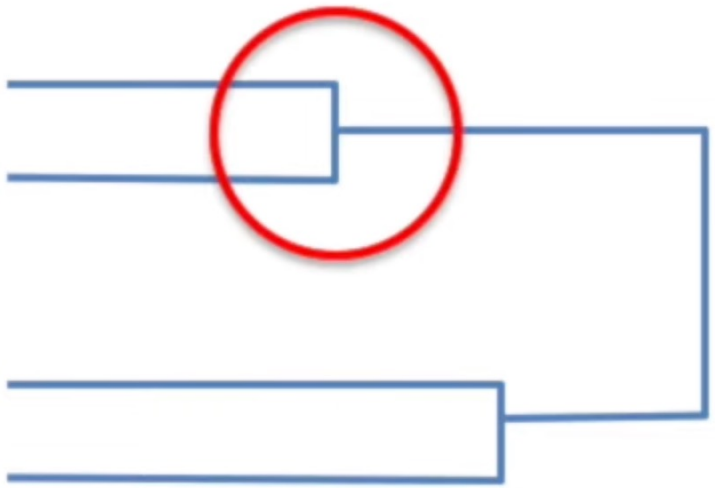


Hierarchical clustering is usually accompanied by a “dendrogram”.

It indicates both the similarity and the order that the clusters were formed.

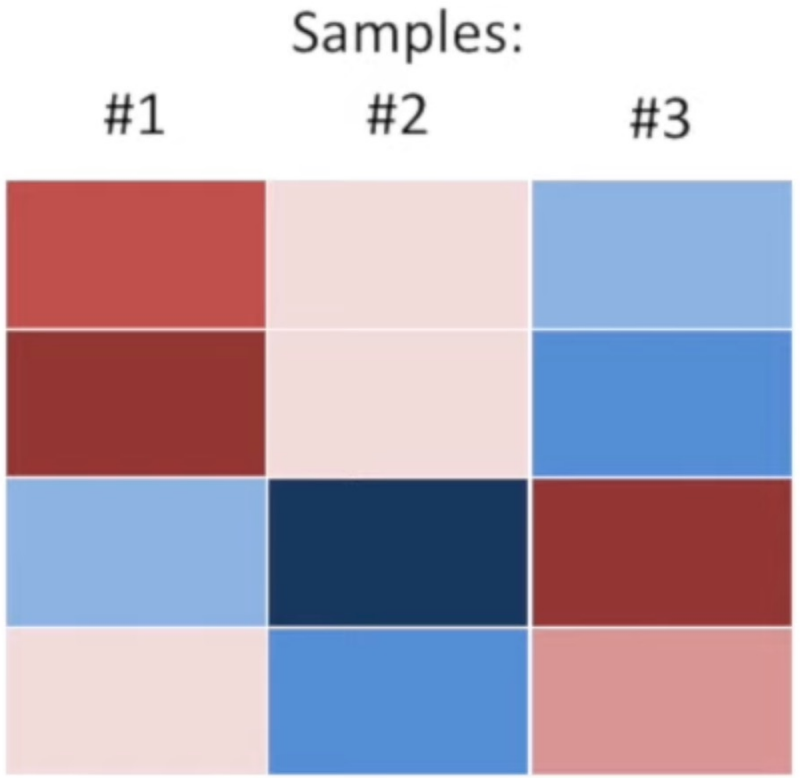


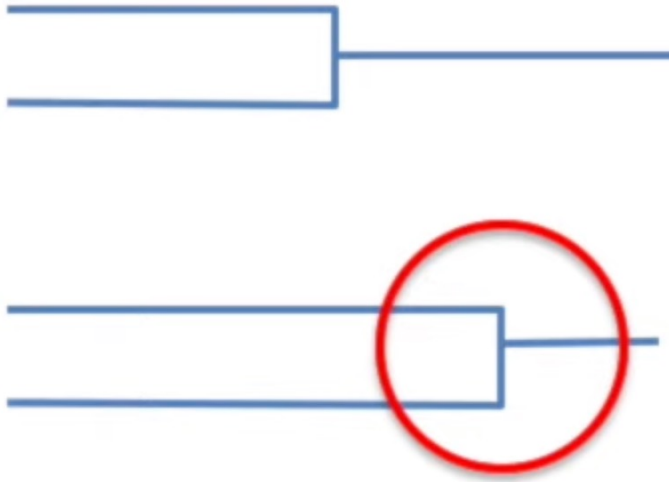
Cluster #1 was formed first and is most similar.
It has the shortest branch.



Hierarchical clustering is usually accompanied by a “dendrogram”.

It indicates both the similarity and the order that the clusters were formed.

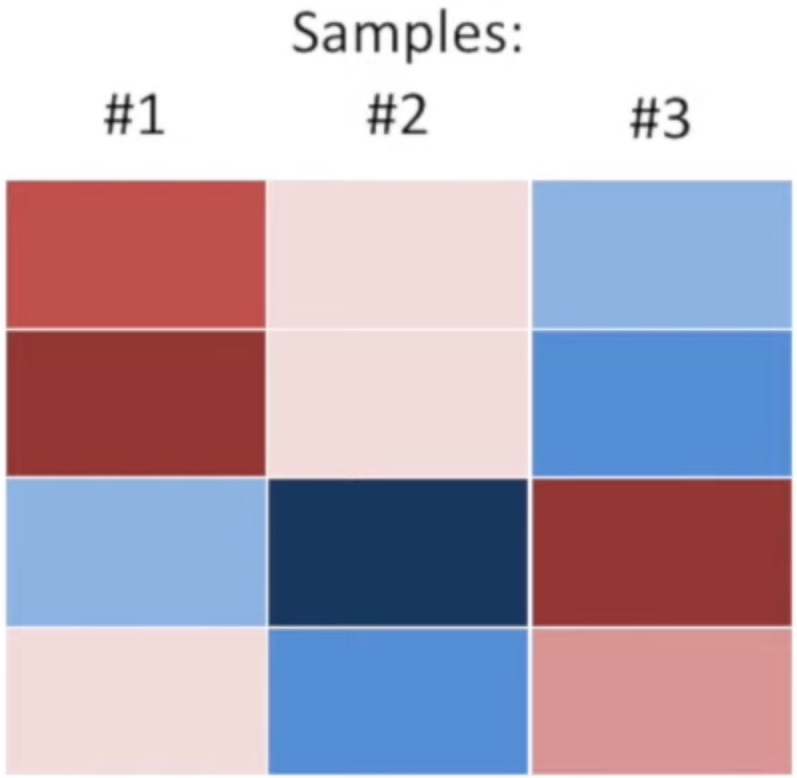




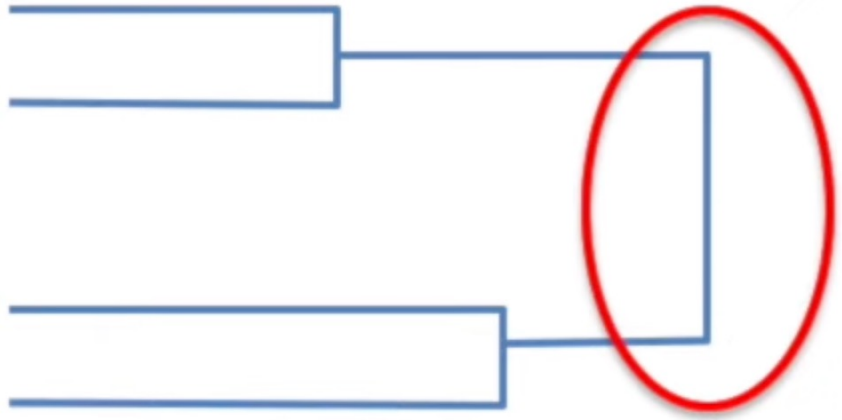
Cluster #2 was second and is the second most similar. It has the second shortest branch.

Hierarchical clustering is usually accompanied by a “dendrogram”.

It indicates both the similarity and the order that the clusters were formed.



Cluster #3, which contains all of the genes, was formed last. It has the longest branch.

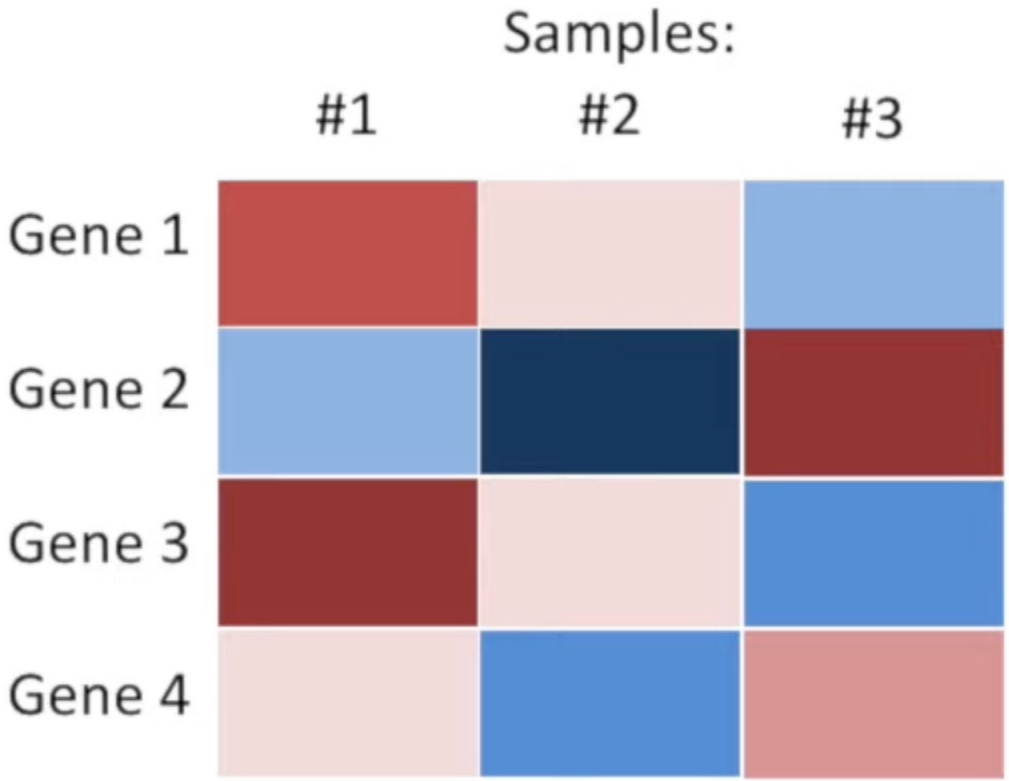


Hierarchical clustering is usually accompanied by a “dendrogram”.

It indicates both the similarity and the order that the clusters were formed.

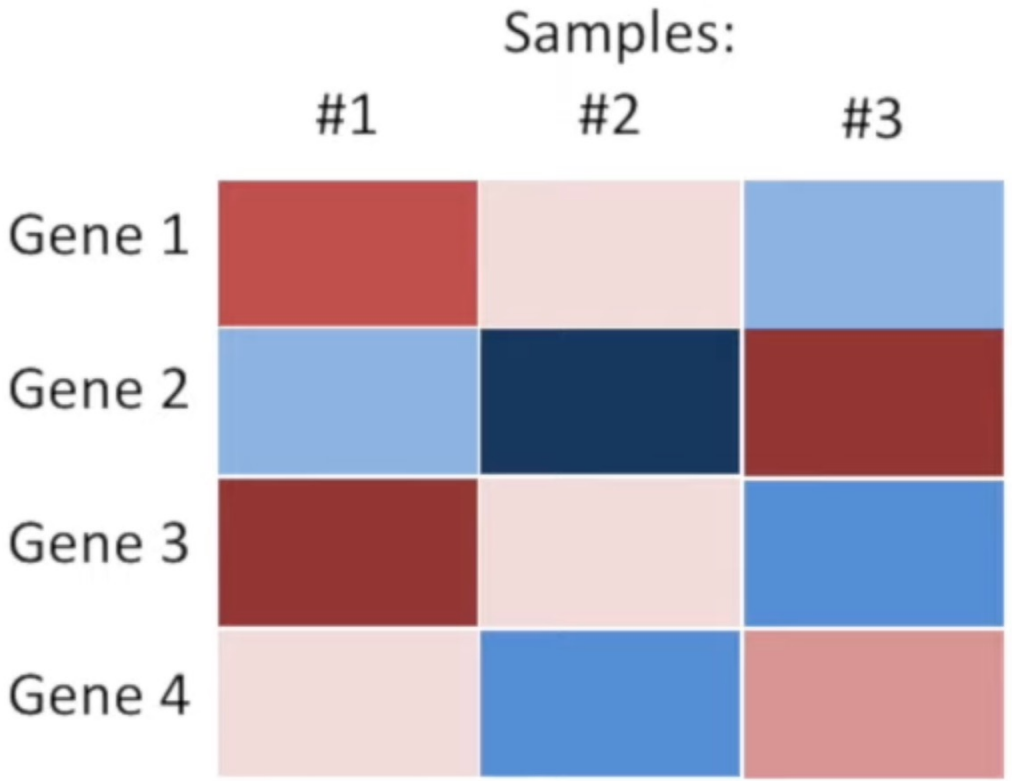
Measuring Similarity

Hierarchical Clustering



Remember the first step?

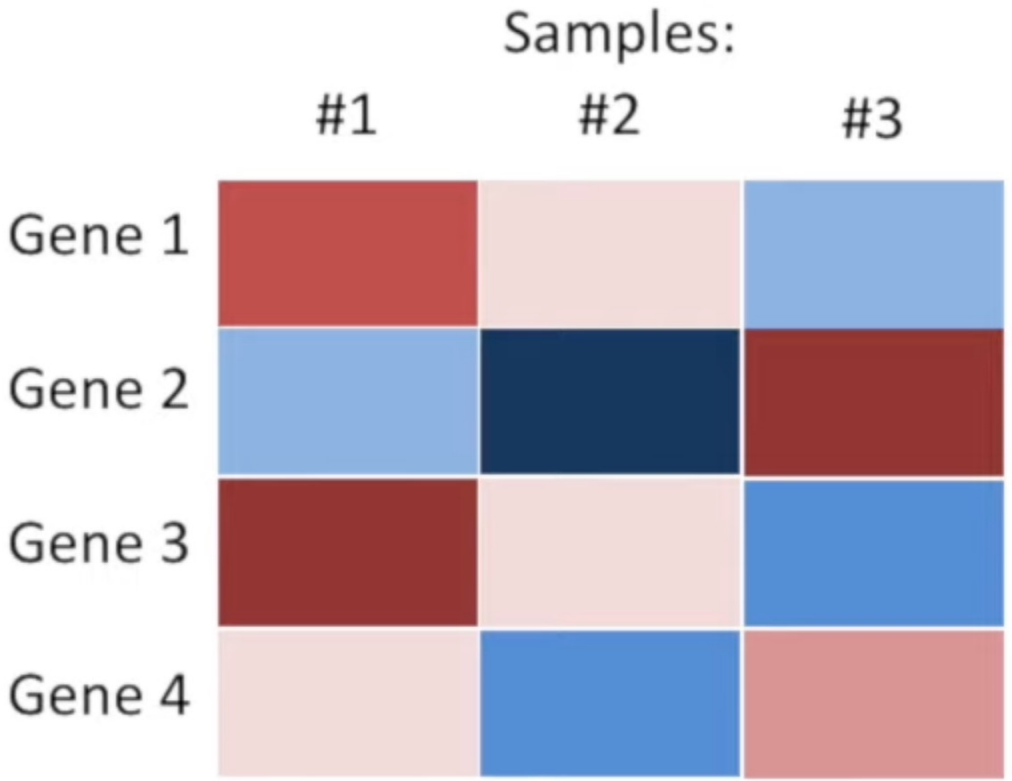
1) Figure out which gene is **most similar** to gene #1.



Remember the first step?

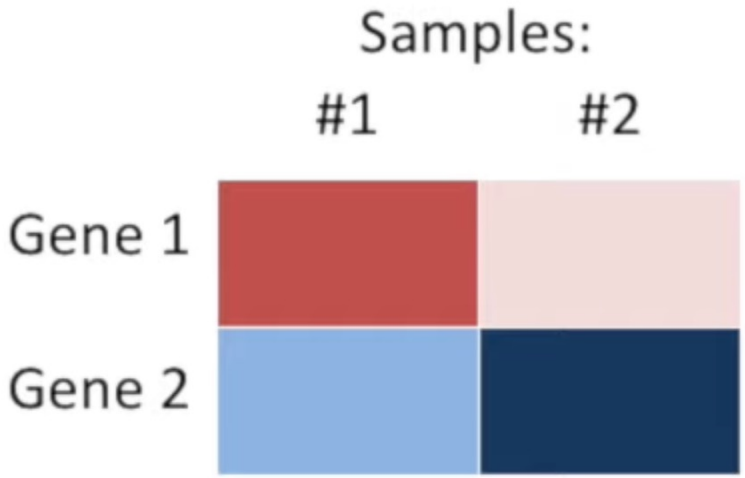
1) Figure out which gene is **most similar** to gene #1.

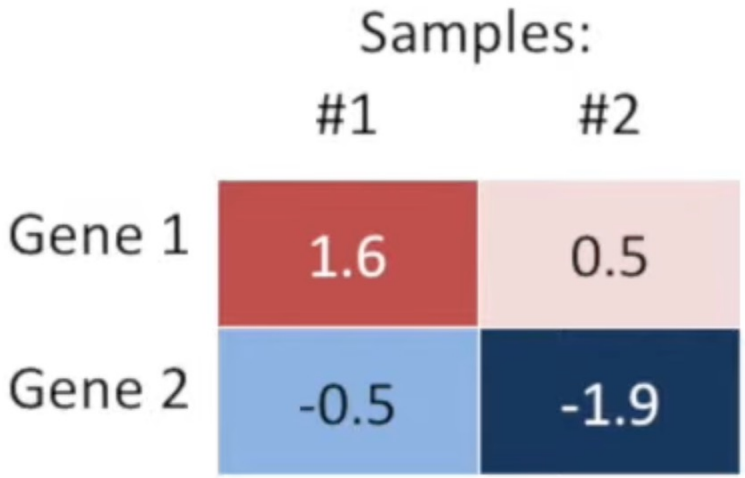
We have to define what "**most similar**" means!

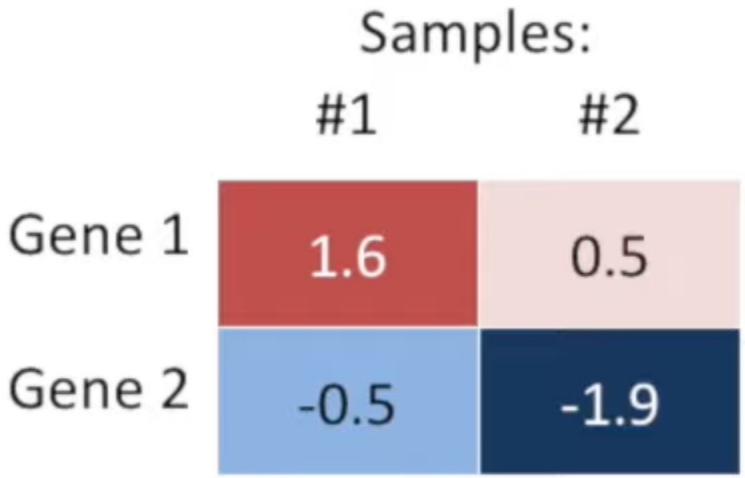


1) Figure out which gene is **most similar** to gene #1.

The method for determining similarity is arbitrarily chosen. However, the Euclidian distance between genes is used a lot.

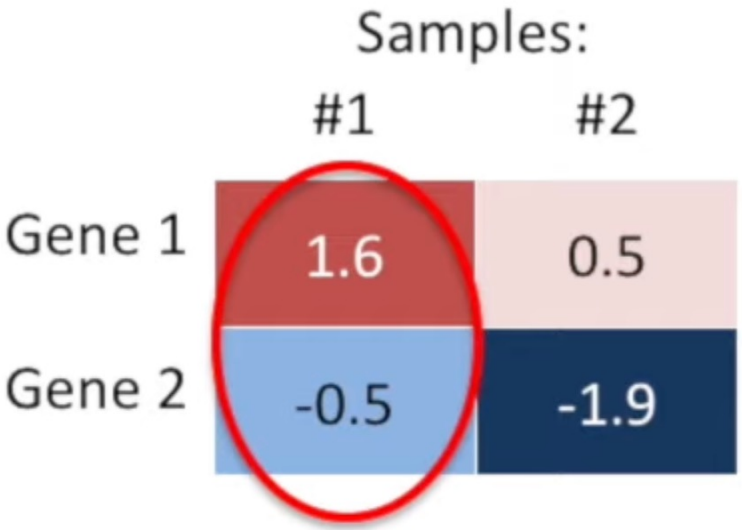

Samples:
#1
#2
Gene 1
Gene 2


Samples:
#1
#2
Gene 1
1.6
0.5
Gene 2
-0.5
-1.9

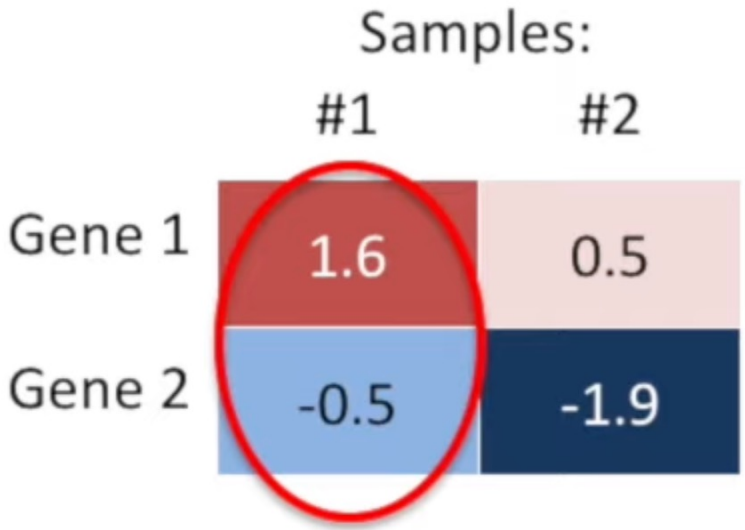


The Euclidean distance between Genes 1 and 2.

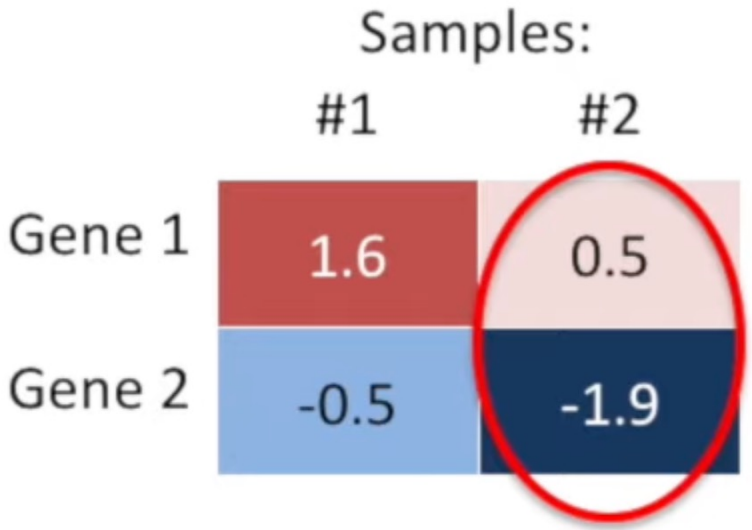
$$\sqrt{(\text{difference in sample \#1})^2 + (\text{difference in sample \#2})^2}$$



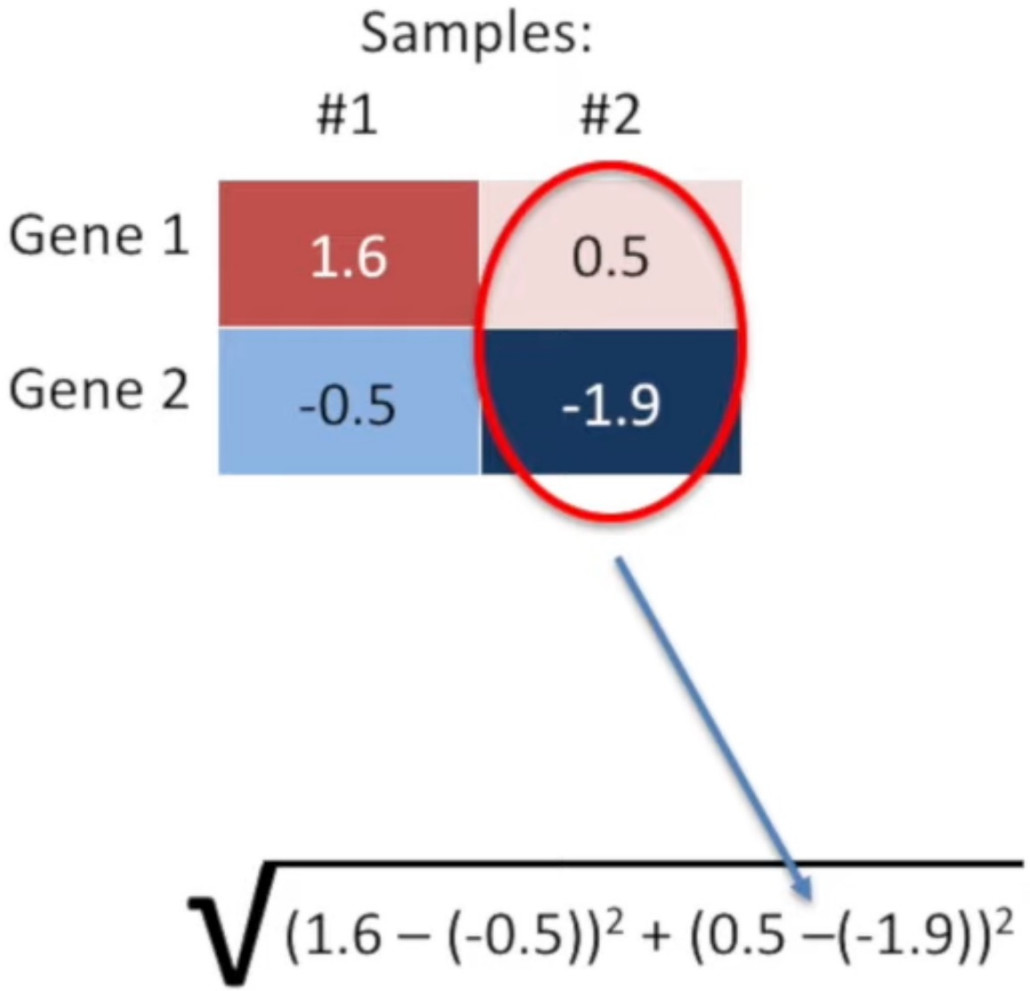
$$\sqrt{(\text{difference in sample \#1})^2 + (\text{difference in sample \#2})^2}$$

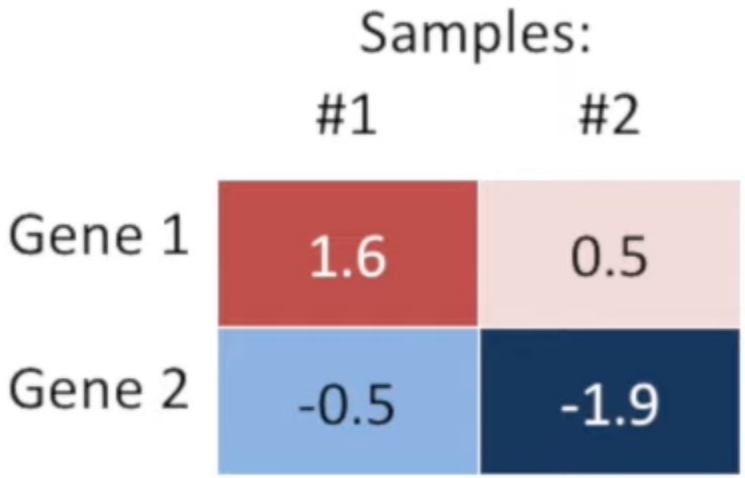


$$\sqrt{(1.6 - (-0.5))^2 + (\text{difference in sample \#2})^2}$$



$$\sqrt{(1.6 - (-0.5))^2 + (\text{difference in sample \#2})^2}$$


Samples:
#1
#2
Gene 1
Gene 2
1.6
0.5
-0.5
-1.9
$\sqrt{(1.6 - (-0.5))^2 + (0.5 - (-1.9))^2}$



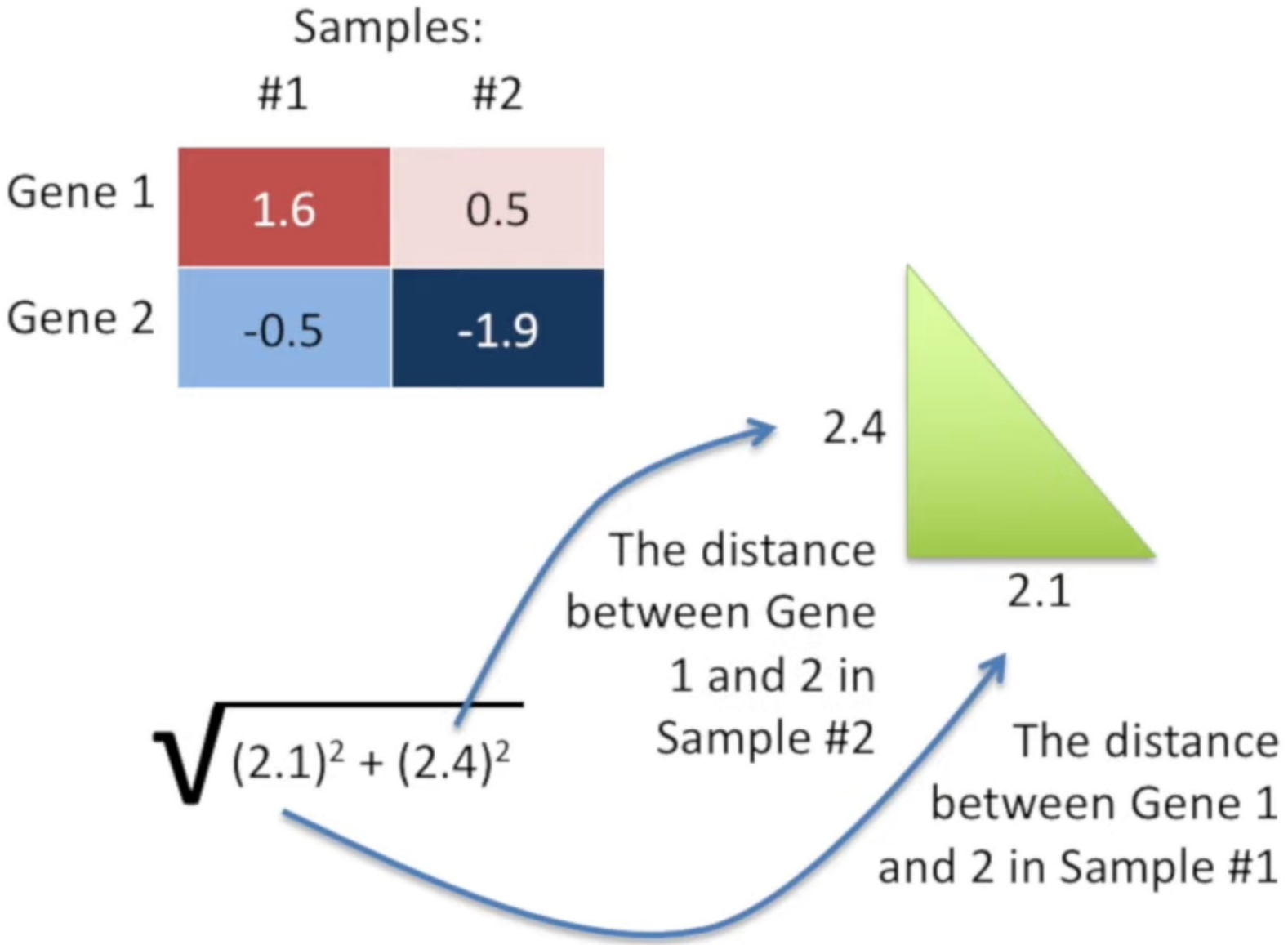
$$\sqrt{(2.1)^2 + (2.4)^2}$$

	Samples: #1	#2
Gene 1	1.6	0.5
Gene 2	-0.5	-1.9

2.1

$$\sqrt{(2.1)^2 + (2.4)^2}$$

The distance between Gene 1 and 2 in Sample #1


Samples:
#1
#2
Gene 1
1.6
0.5
Gene 2
-0.5
-1.9
2.4
2.1
The distance between Gene 1 and 2 in Sample #2
The distance between Gene 1 and 2 in Sample #1
$\sqrt{(2.1)^2 + (2.4)^2}$

	Samples: #1	#2
Gene 1	1.6	0.5
Gene 2	-0.5	-1.9

The hypotenuse is the total "distance" between genes #1 and #2.

$$\sqrt{(2.1)^2 + (2.4^2)}$$

	Samples: #1	#2
Gene 1	1.6	0.5
Gene 2	-0.5	-1.9

The Pythagorean theorem says that the hypotenuse = $\sqrt{x^2 + y^2}$

$$\sqrt{(2.1)^2 + (2.4)^2}$$

	Samples: #1	#2
Gene 1	1.6	0.5
Gene 2	-0.5	-1.9

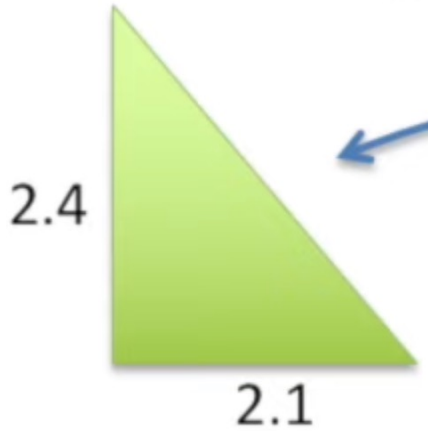
The Pythagorean theorem says that the hypotenuse = $\sqrt{x^2 + y^2}$

$= \sqrt{(2.1)^2 + (2.4)^2}$

$\sqrt{(2.1)^2 + (2.4)^2}$

	Samples: #1	#2
Gene 1	1.6	0.5
Gene 2	-0.5	-1.9

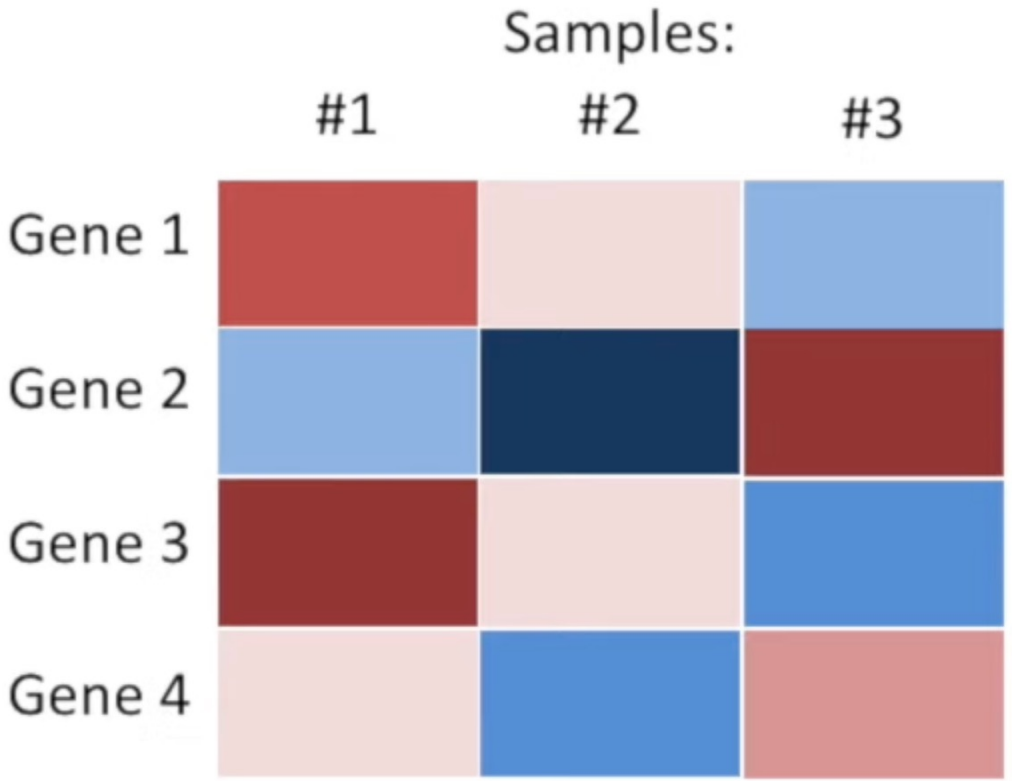
The Pythagorean theorem says that the hypotenuse = $\sqrt{x^2 + y^2}$



$= \sqrt{(2.1)^2 + (2.4)^2}$

$= 3.2$

$$\sqrt{(2.1)^2 + (2.4)^2}$$



When we have more samples, we just extend the equation…

$$\sqrt{(\text{difference in sample \#1})^2 + (\text{difference in sample \#2})^2 + (\text{difference in sample…})^2}$$

Distance metrics

- Euclidian distance is just one method… there are lots more, including:

Distance metrics

- Euclidian distance is just one method… there are lots more, including:
  - The Manhattan distance.

Distance metrics

- Euclidian distance is just one method… there are lots more, including:
  - The Manhattan distance.

    The Manhattan distance is just the absolute value of the differences….

    |difference in sample #1|+ |difference in sample #2| + |difference in gene …|

Distance metrics

- Euclidian distance is just one method… there are lots more, including:
  - The Manhattan distance.

    The Manhattan distance is just the absolute value of the differences….

    |difference in sample #1|+ |difference in sample #2| + |difference in gene …|

Distance metrics

- Euclidian distance is just one method… there are lots more, including:
  - The Manhattan distance.

The Manhattan distance is just the absolute value of the differences….

|difference in sample #1|+ |difference in sample #2| + |difference in gene …|

Distance metrics

- Euclidian distance is just one method… there are lots more, including:
  - The Manhattan distance.

The Manhattan distance is just the absolute value of the differences….

|difference in sample #1|+ |difference in sample #2| + |difference in gene …|

Distance metrics

- Euclidian distance is just one method… there are lots more, including:
  - The Manhattan distance.

The Manhattan distance is just the absolute value of the differences….

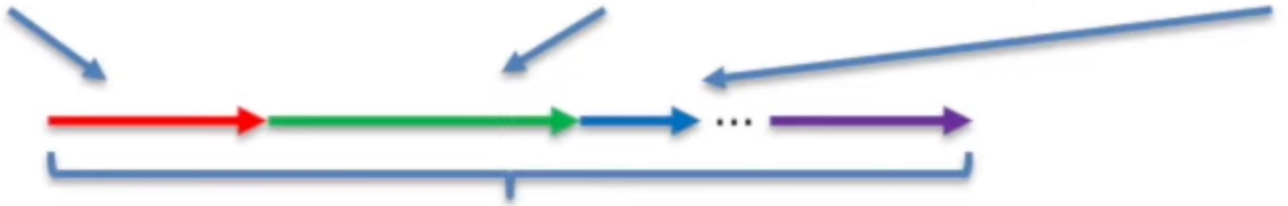
|difference in sample #1|+ |difference in sample #2| + |difference in gene …|

Distance metrics

- Euclidian distance is just one method… there are lots more, including:
  - The Manhattan distance.

The Manhattan distance is just the absolute value of the differences….

|difference in sample #1|+ |difference in sample #2| + |difference in gene …|



Total distance = Manhattan Distance.

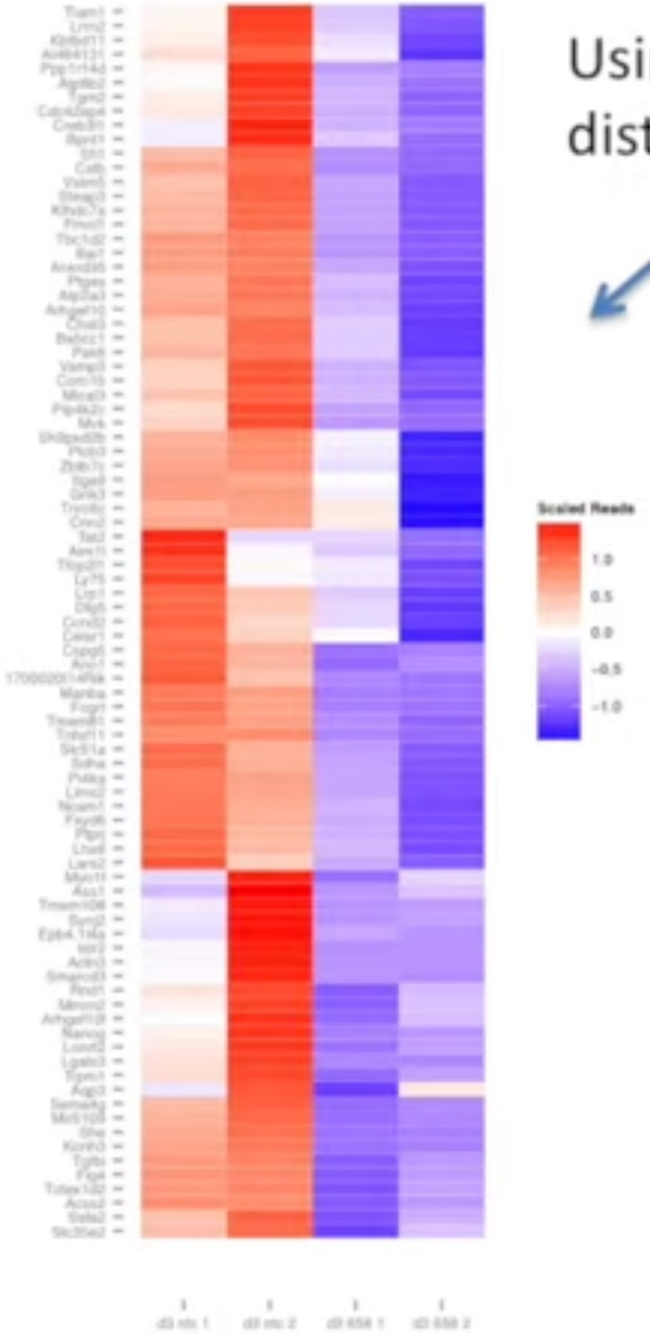
Distance metrics

- Euclidian distance is just one method… there are lots more, including:
  - The Manhattan distance.

    The Manhattan distance is just the absolute value of the differences….

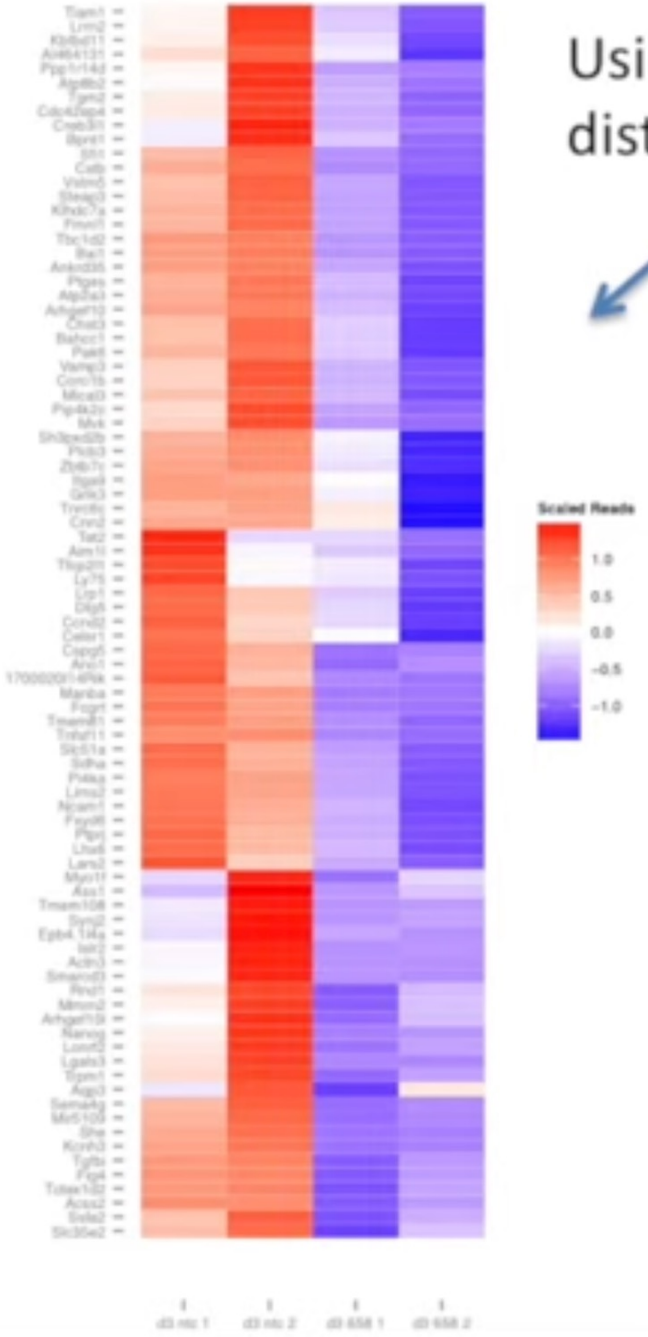
    |difference in sample #1|+ |difference in sample #2| + |difference in gene …|

- Yes, it makes a difference.

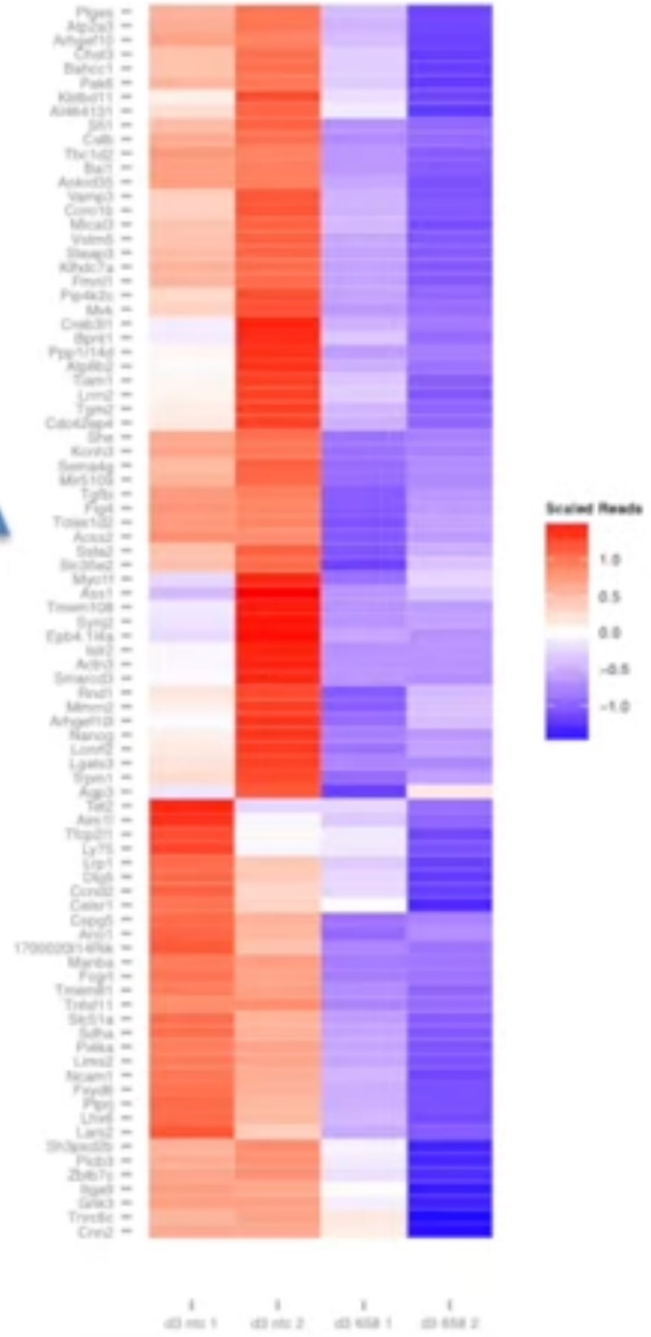


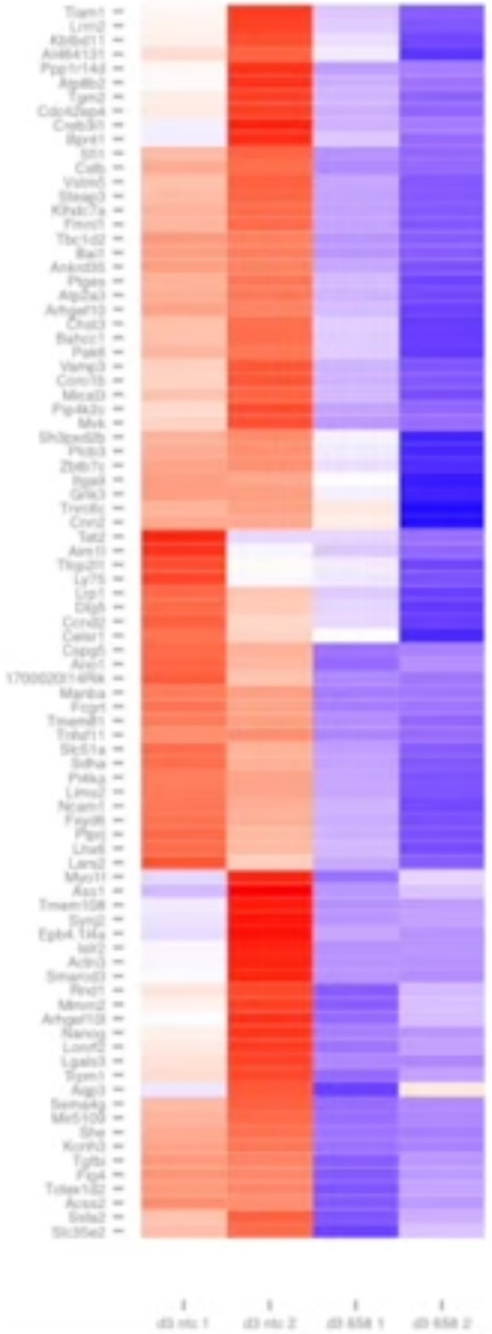
Using the “Euclidean” distance…

Using the “Euclidean” distance…



Using the “Manhattan” distance…

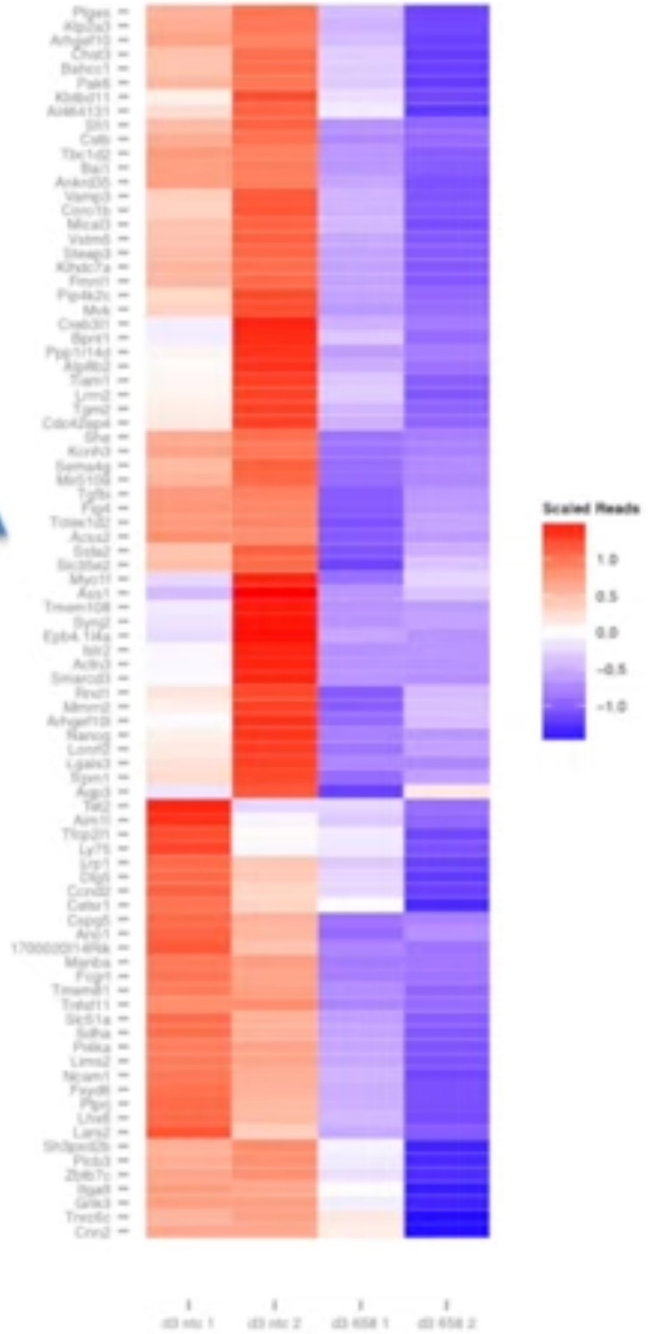


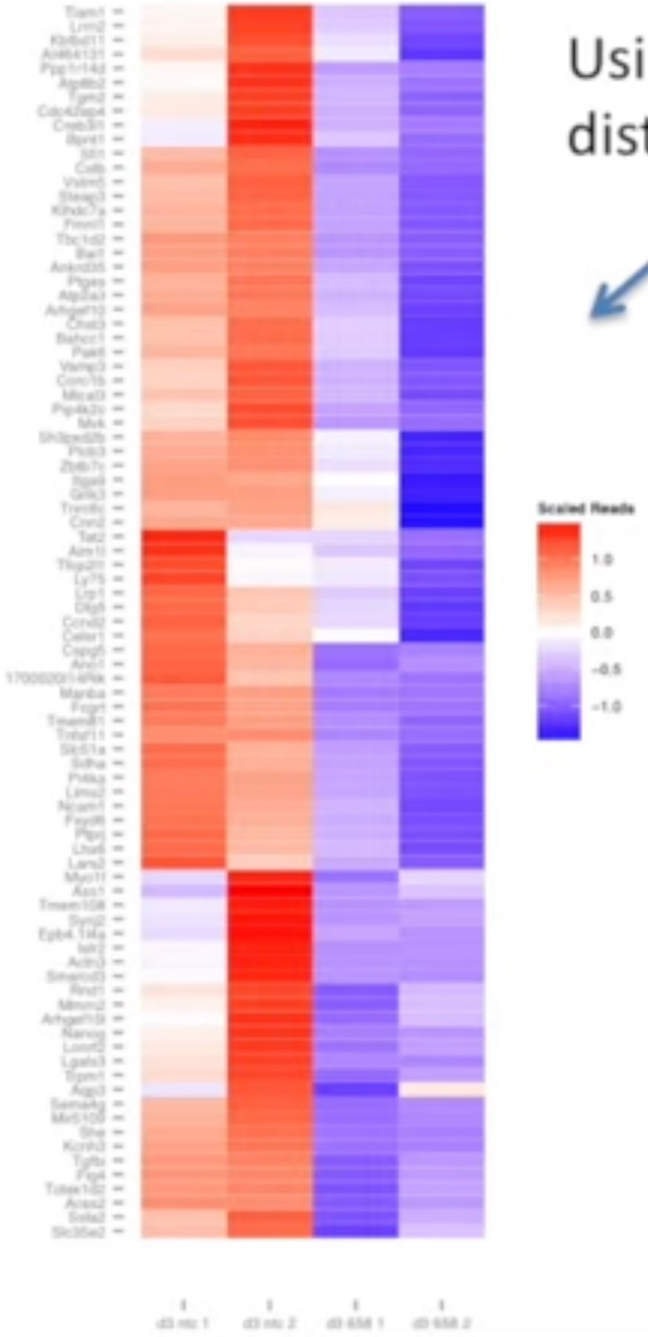


Using the “Euclidean” distance…

But the choice is arbitrary…

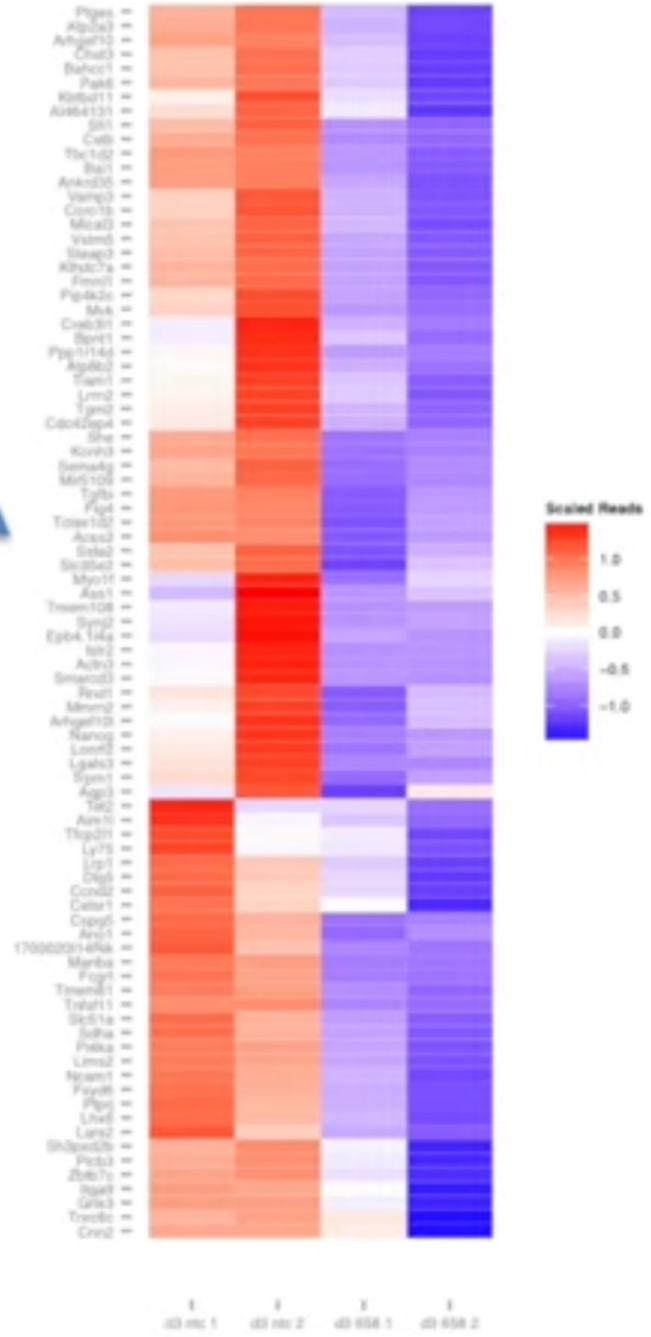
Using the “Manhattan” distance…





Using the “Euclidean” distance…

Using the “Manhattan” distance…



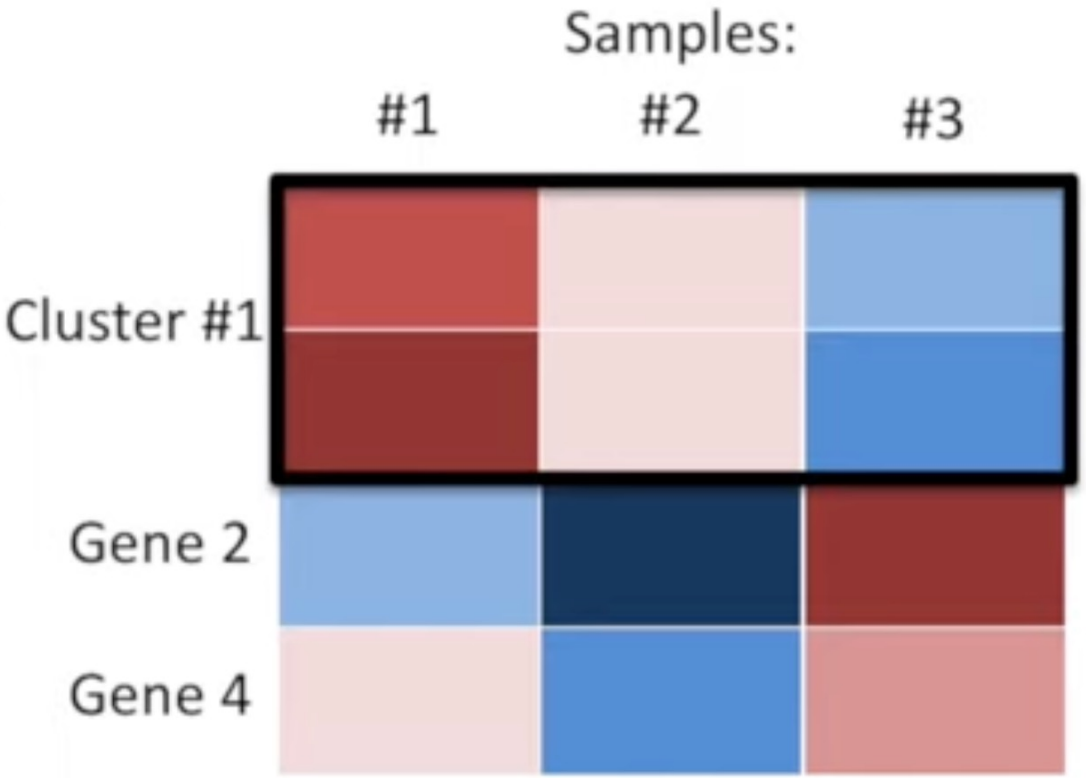
But the choice is arbitrary… :(

There’s no biological or physical reason to choose one and not the other.

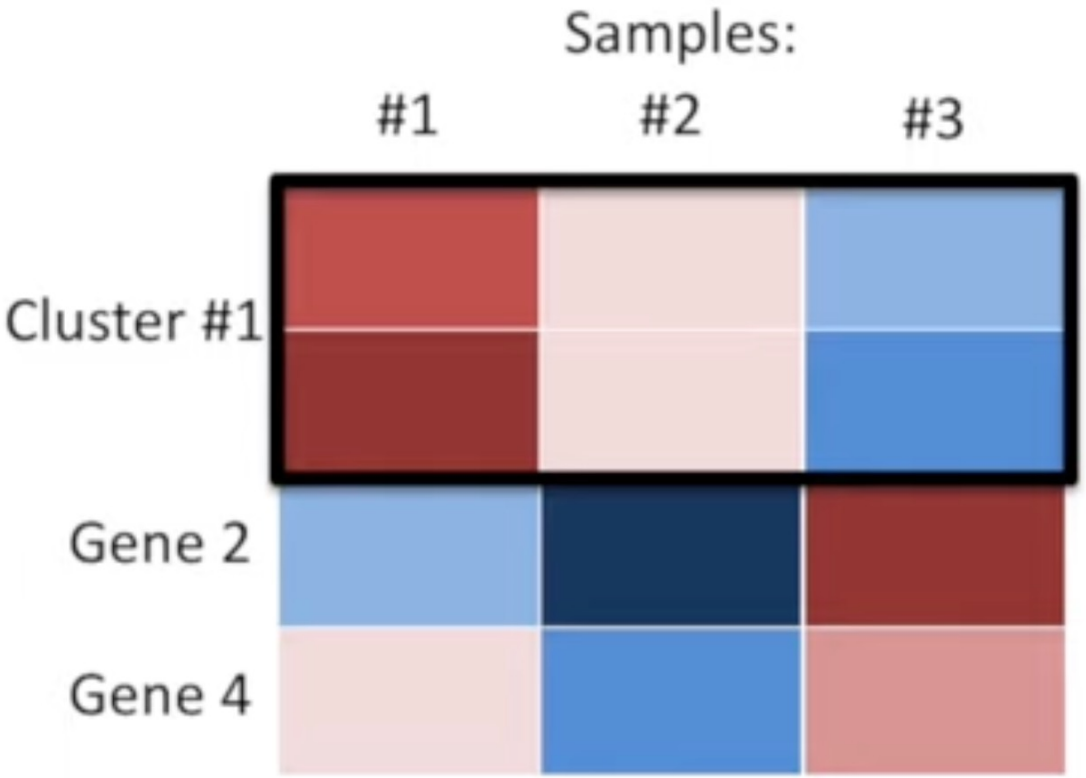
Pick the one that gives you more insight into your data.

Measuring Distance to Clusters

Hierarchical Clustering

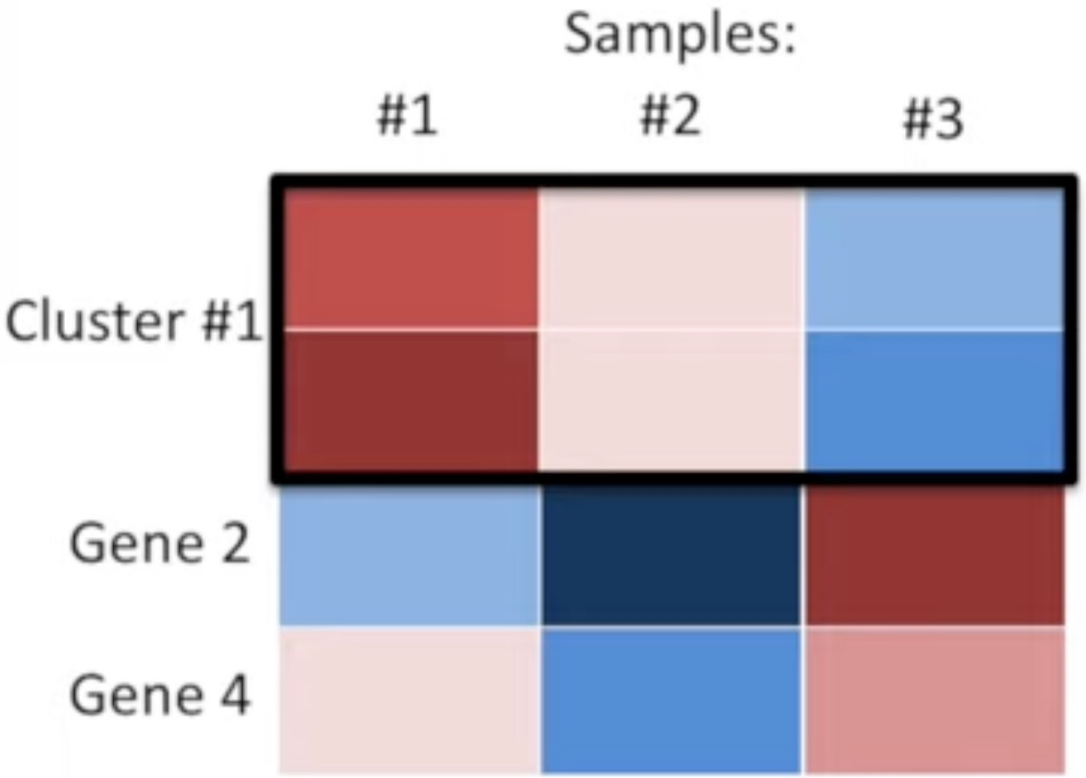


Do you remember how we merged genes #1 and #3 into cluster #1 and compared it to other genes?



Do you remember how we merged genes #1 and #3 into cluster #1 and compared it to other genes?

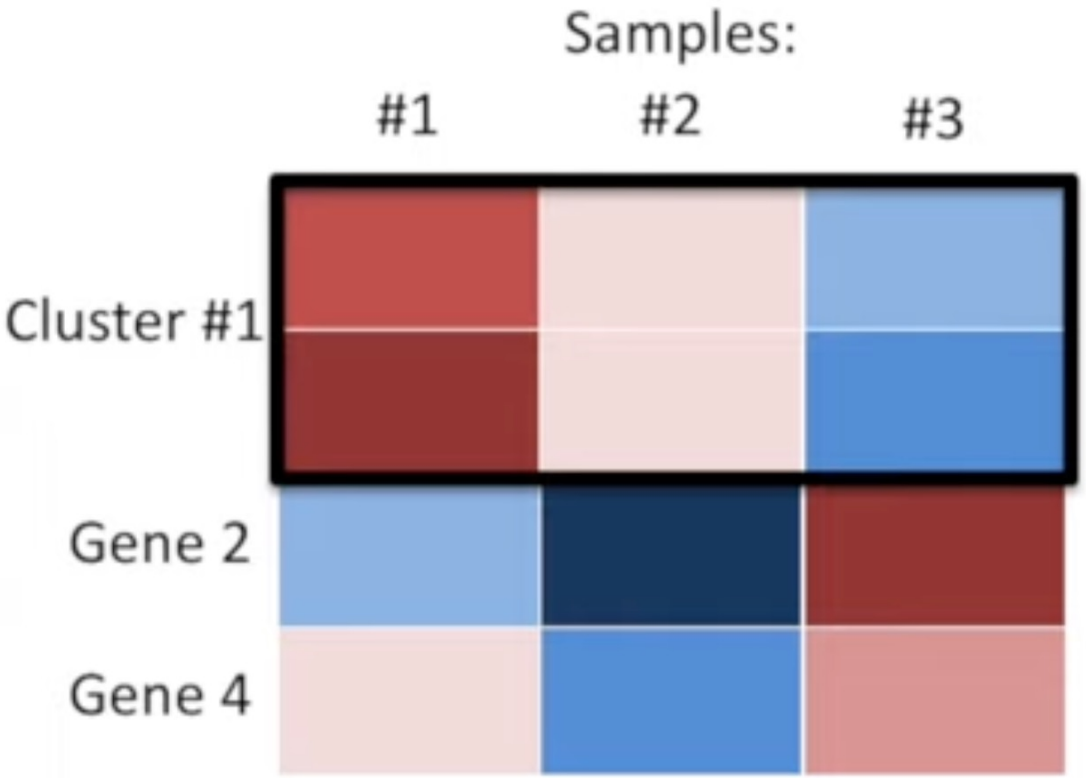
Well, there are different ways to compare clusters, too.



Do you remember how we merged genes #1 and #3 into cluster #1 and compared it to other genes?

Well, there are different ways to compare clusters, too.

One simple idea is to use the average of the measurements from each sample.



Do you remember how we merged genes #1 and #3 into cluster #1 and compared it to other genes?

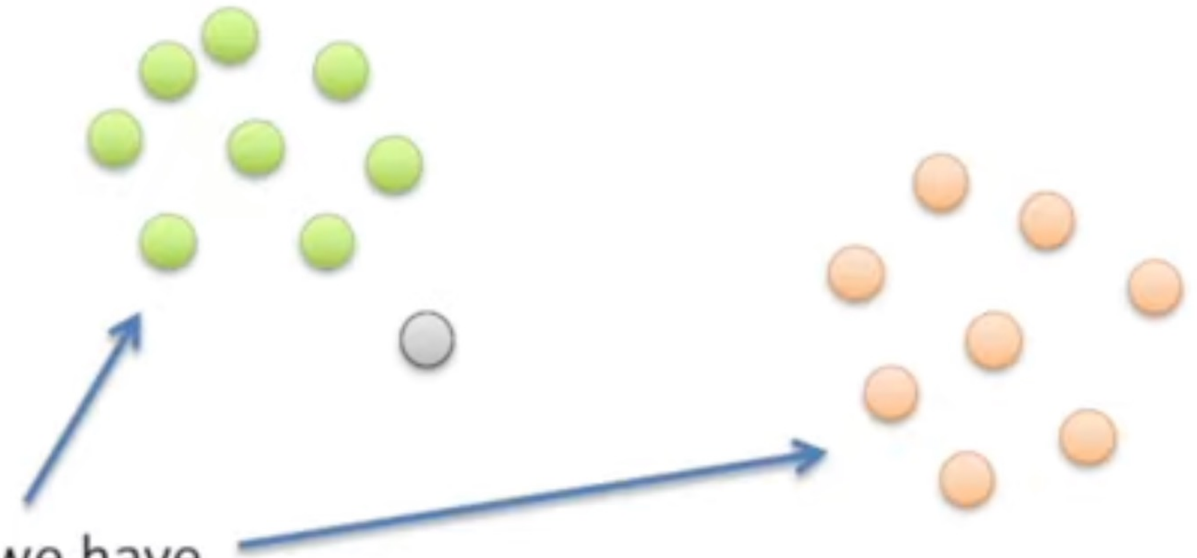
Well, there are different ways to compare clusters, too.

One simple idea is to use the average of the measurements from each sample.

But there are lots more.

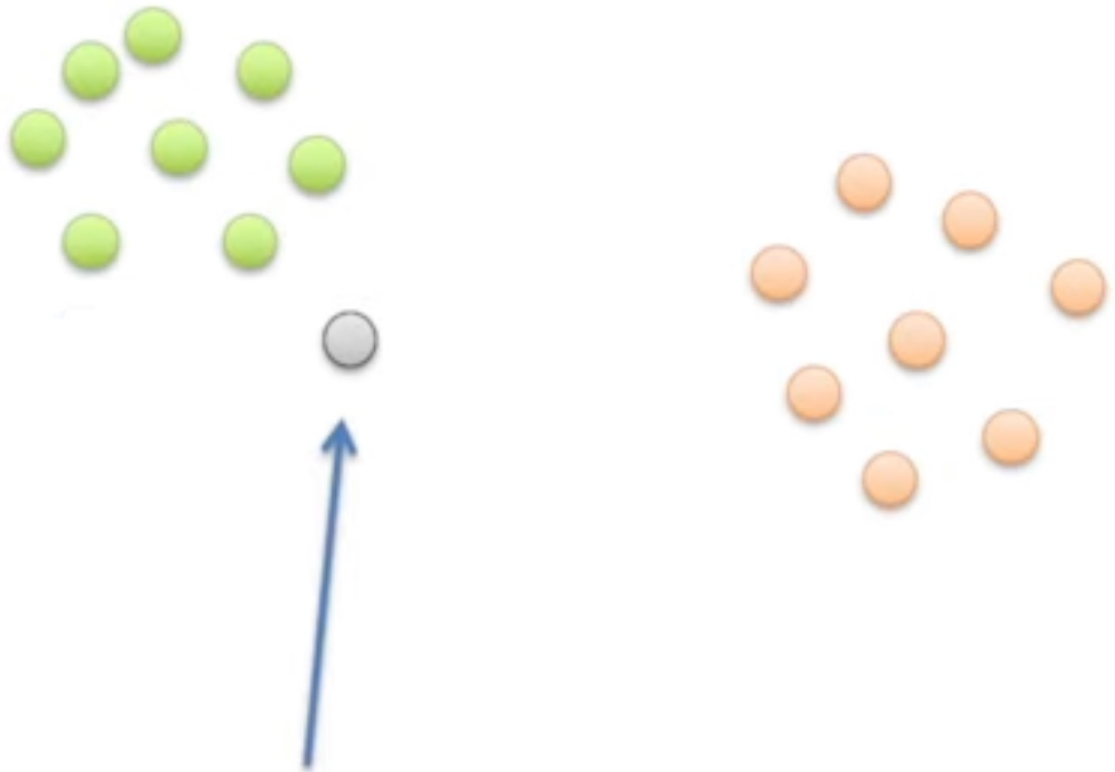
For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.

For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.



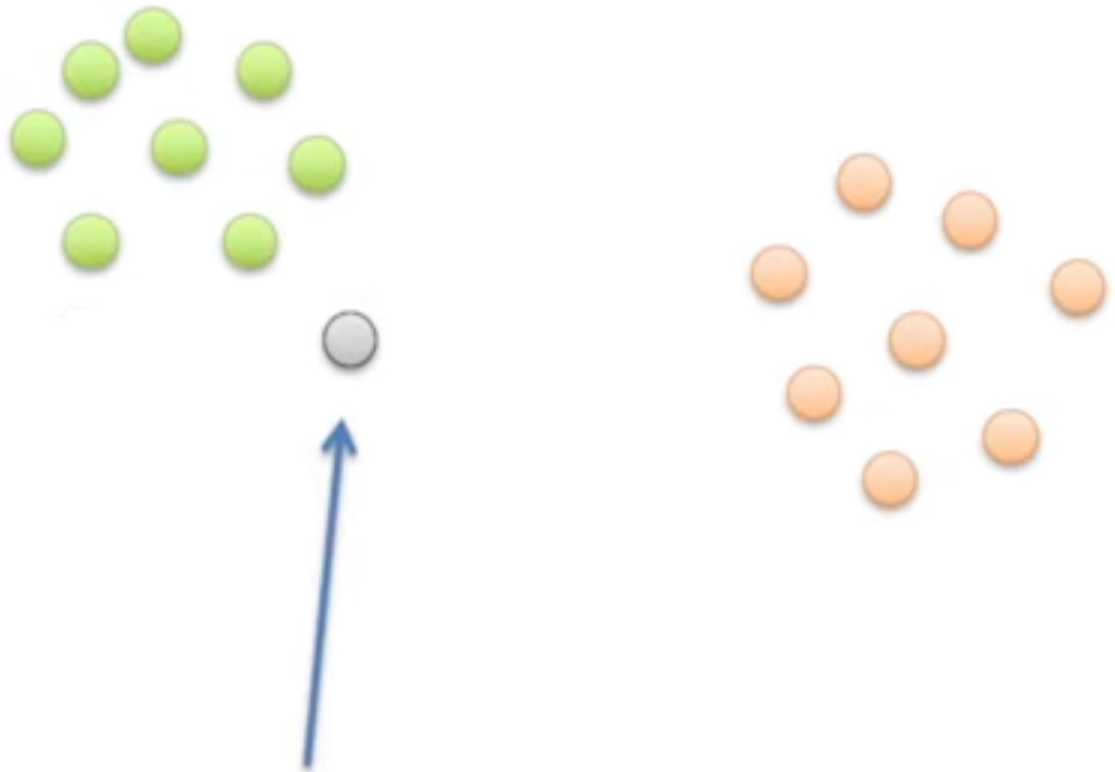
Now imagine that we have already formed these two clusters…

For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.



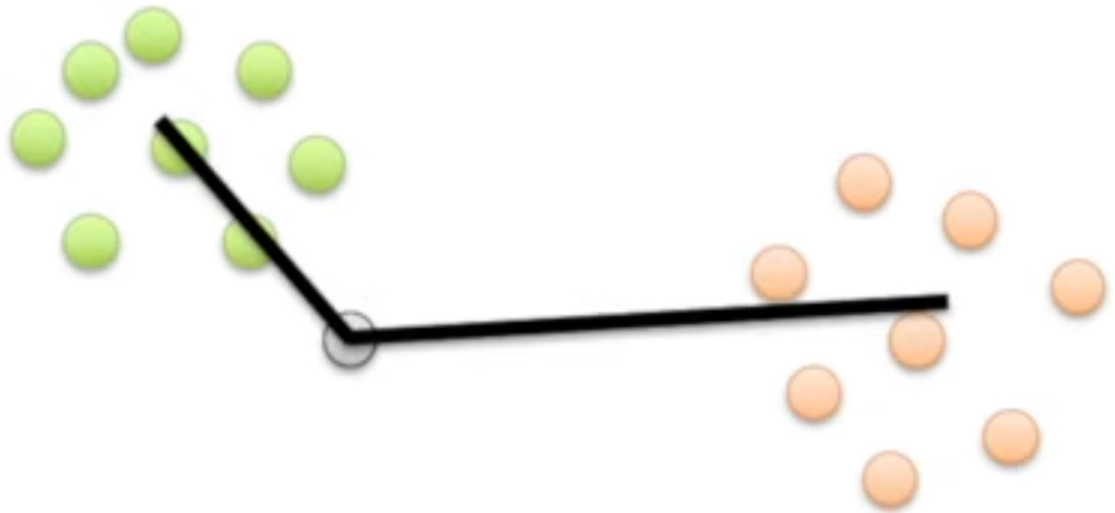
… and we just want to figure out which cluster this last point belongs to.

For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.



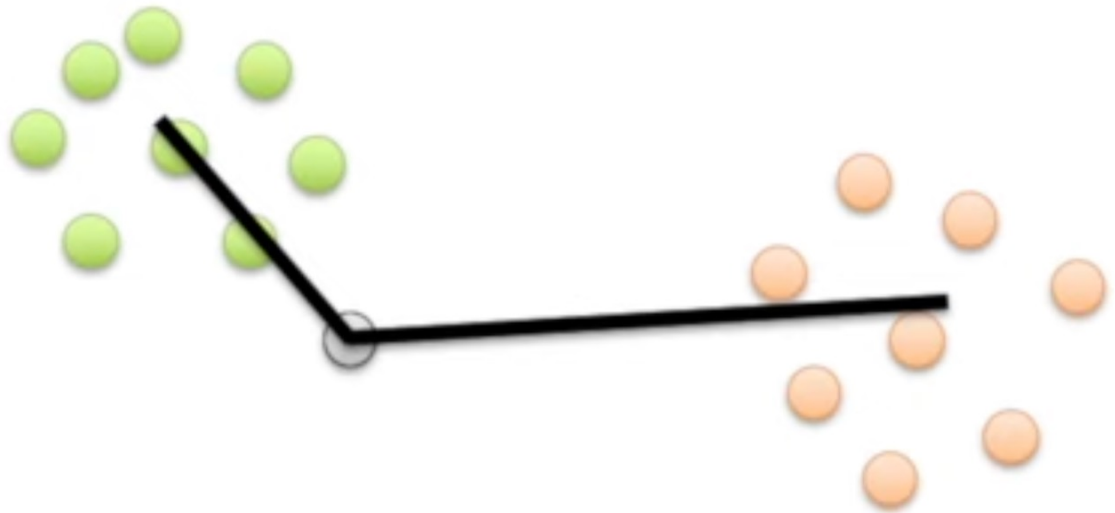
… and we just want to figure out which cluster this last point belongs to.

For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.



We can compare that point to…

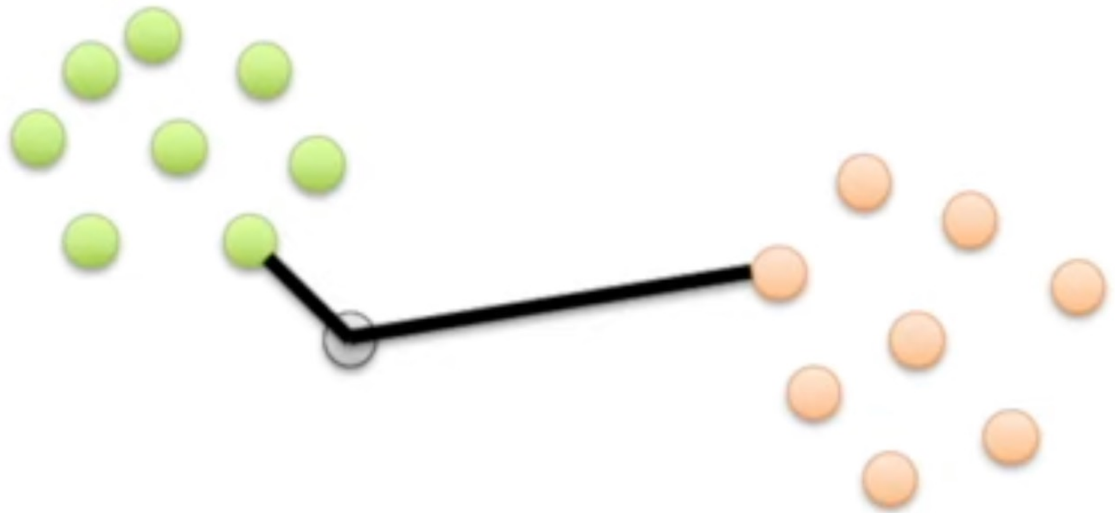
For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.



We can compare that point to…

1) The average of each cluster (this is called the “centroid”)

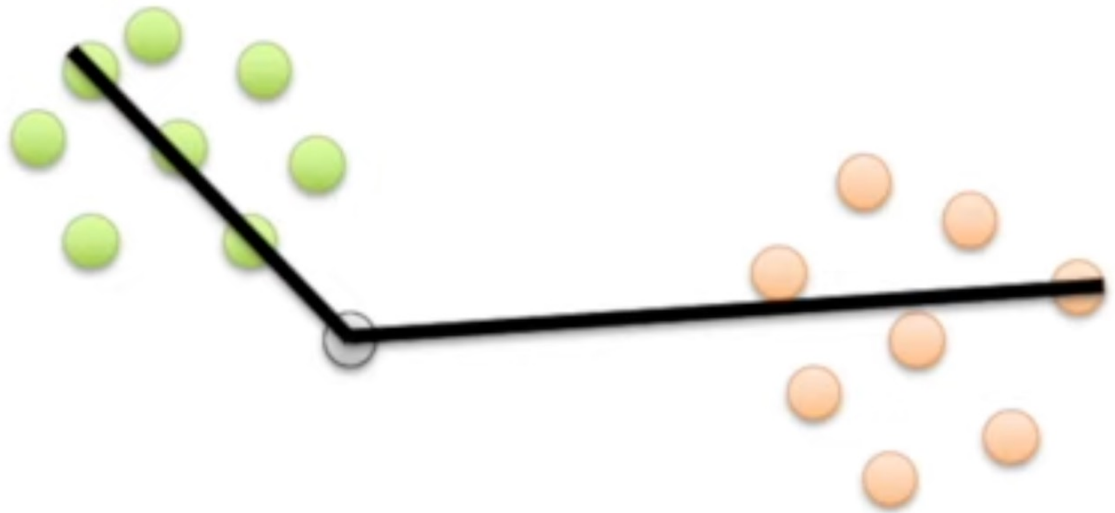
For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.



We can compare that point to…

1) The average of each cluster (this is called the “centroid”)
2) The closest point in each cluster (this is called “single-linkage”)

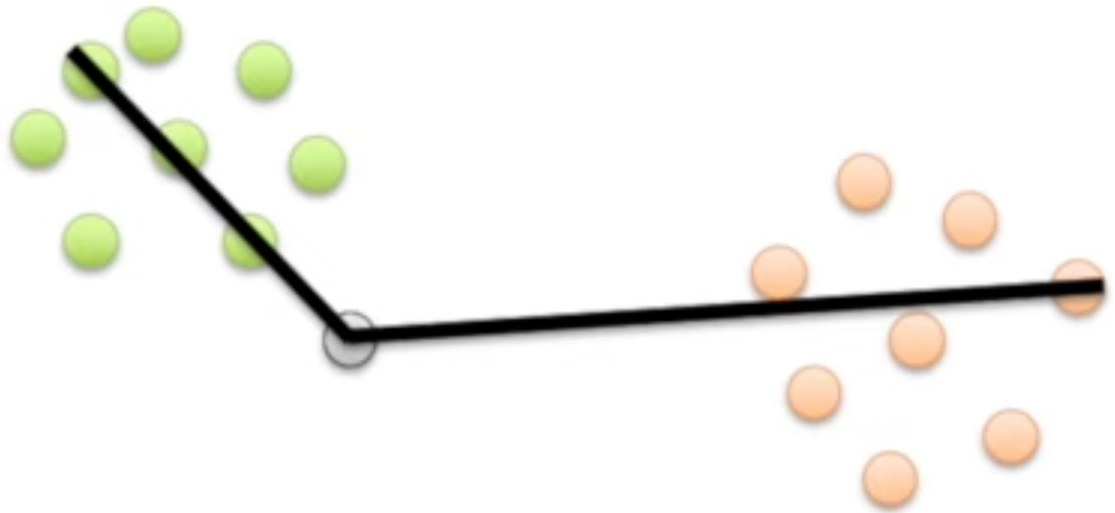
For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.



We can compare that point to…

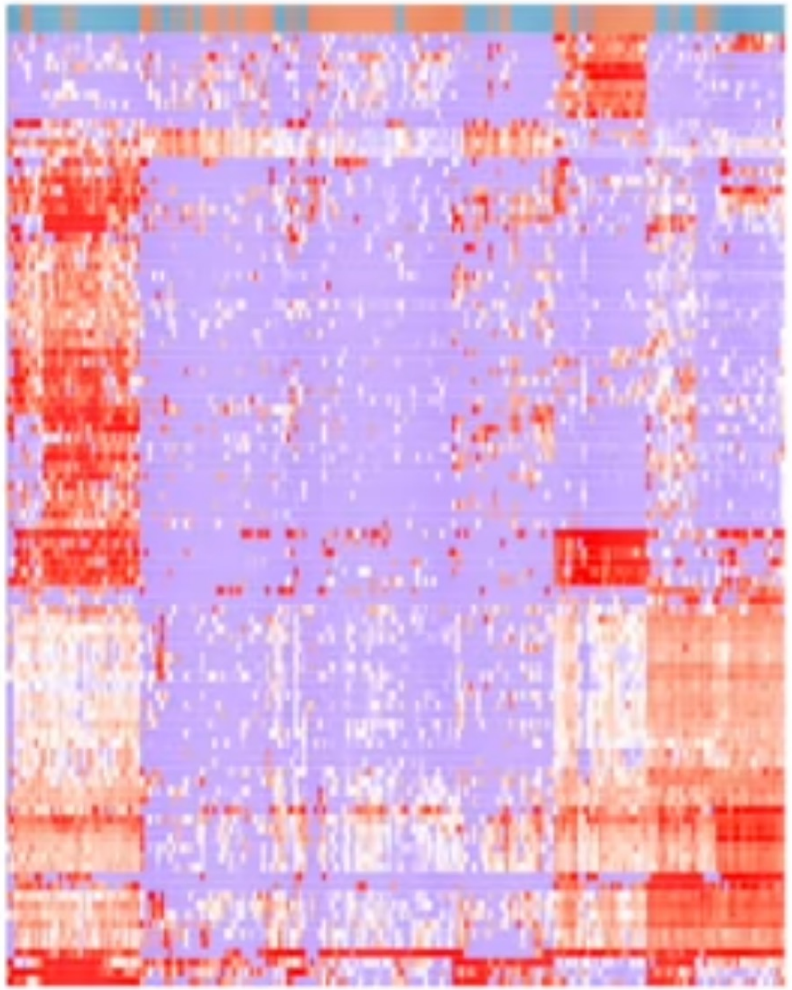
1) The average of each cluster (this is called the “centroid”)
2) The closest point in each cluster (this is called “single-linkage”)
3) The furthest point in each cluster (this is called “complete-linkage”)

For the sake of visualizing how the different methods work, imagine our data was spread out on an X-Y plane.

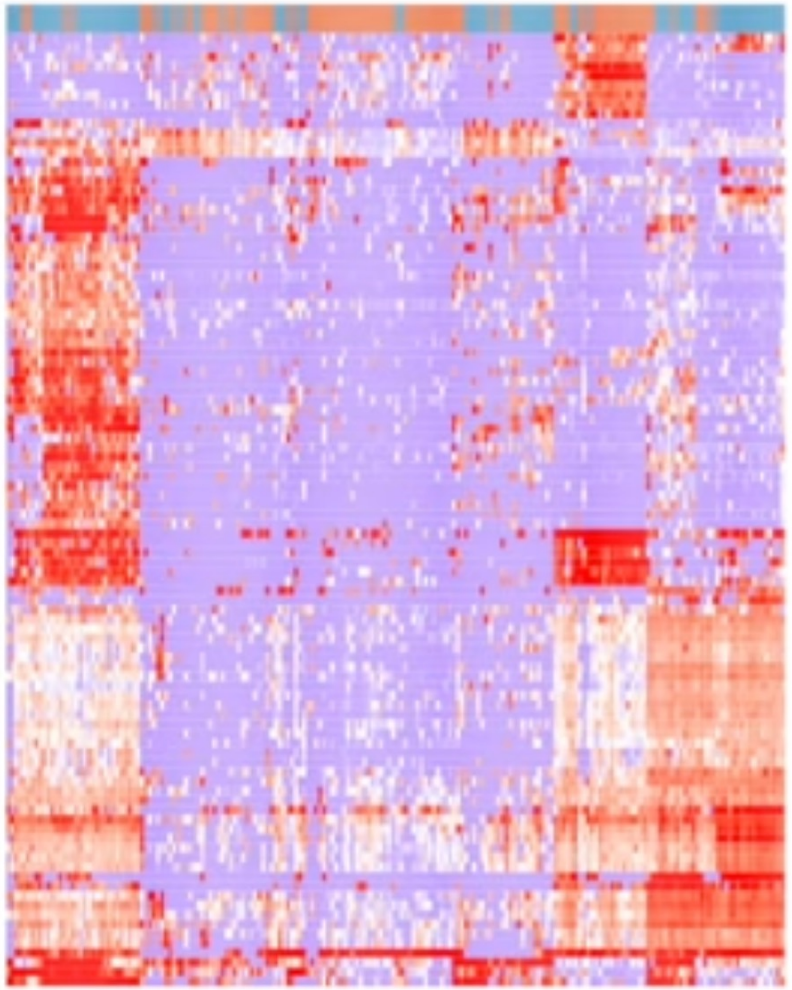


We can compare that point to…

1) The average of each cluster (this is called the “centroid”)
2) The closest point in each cluster (this is called “single-linkage”)
3) The furthest point in each cluster (this is called “complete-linkage”)
4) Etc.

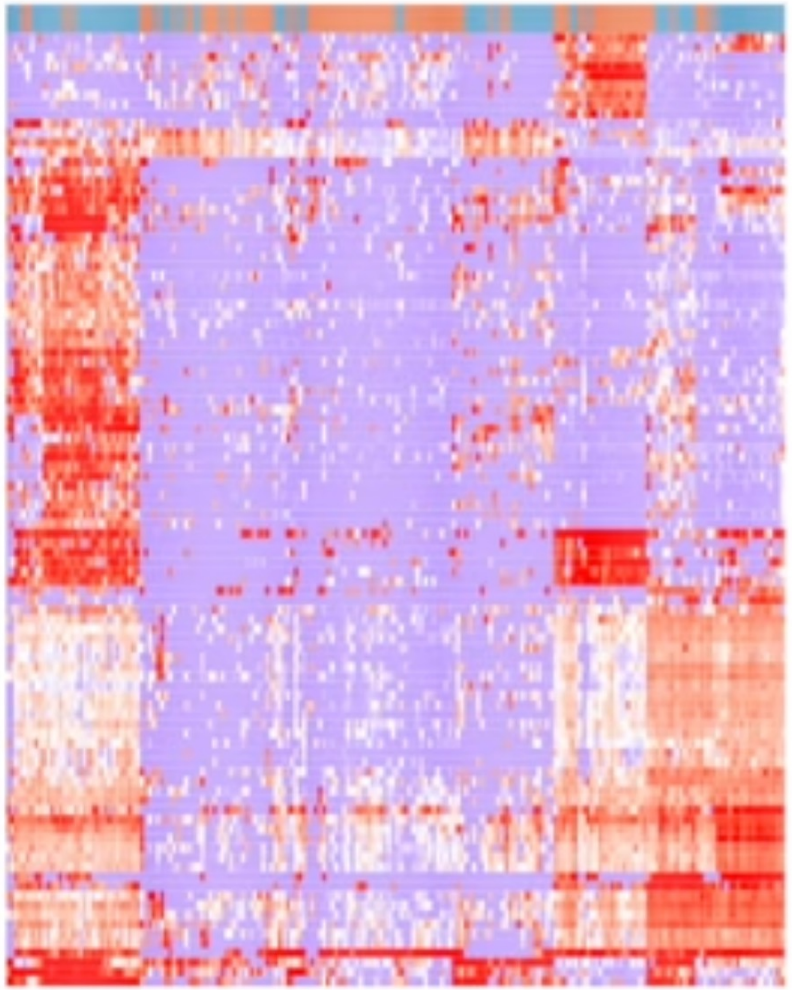


Here’s a heatmap that compares the **furthest** points in the clusters.

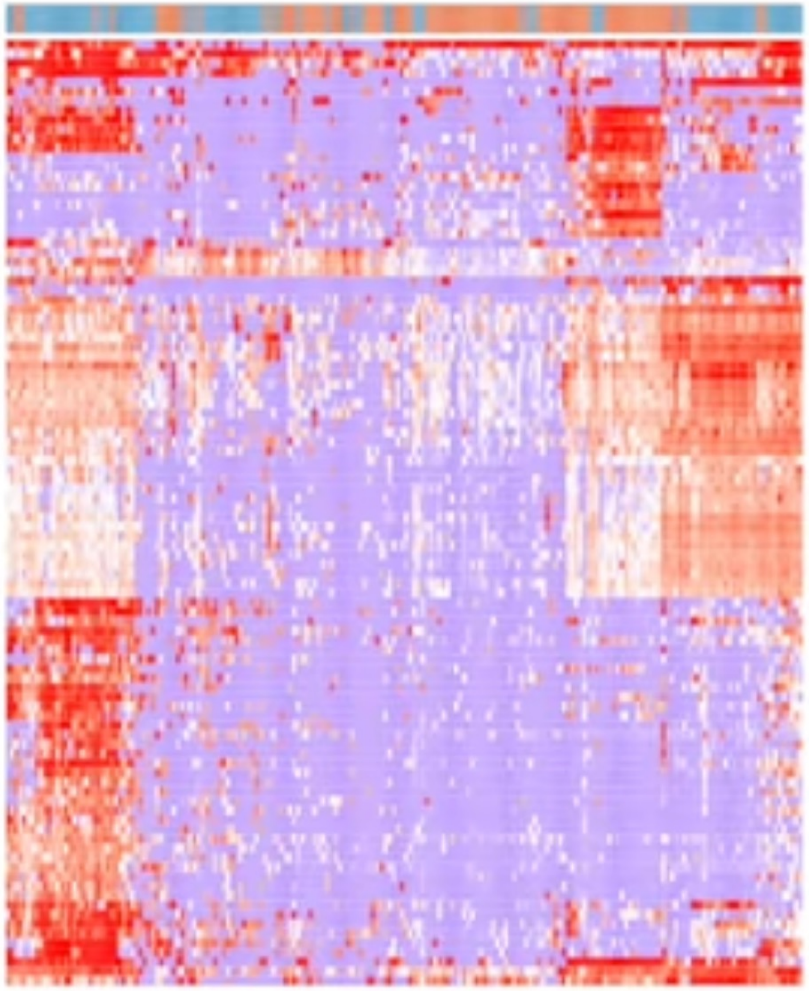


Here's a heatmap that compares the **furthest** points in the clusters.

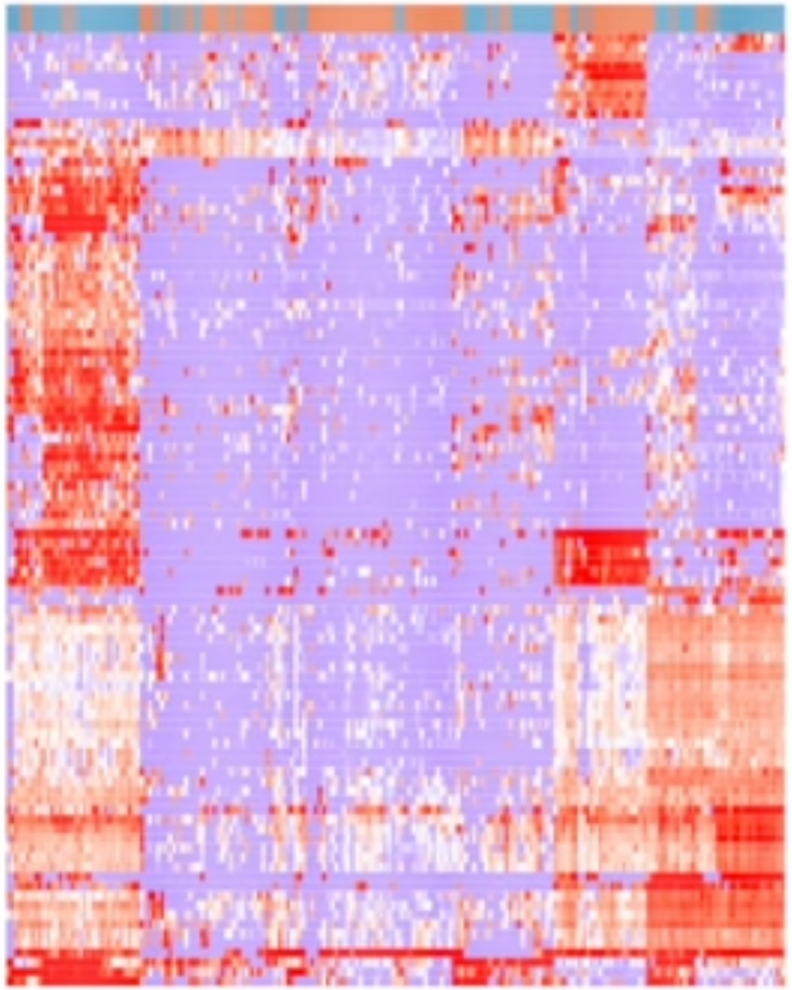
By the way, if you R, this is the default setting for the hclust() function.



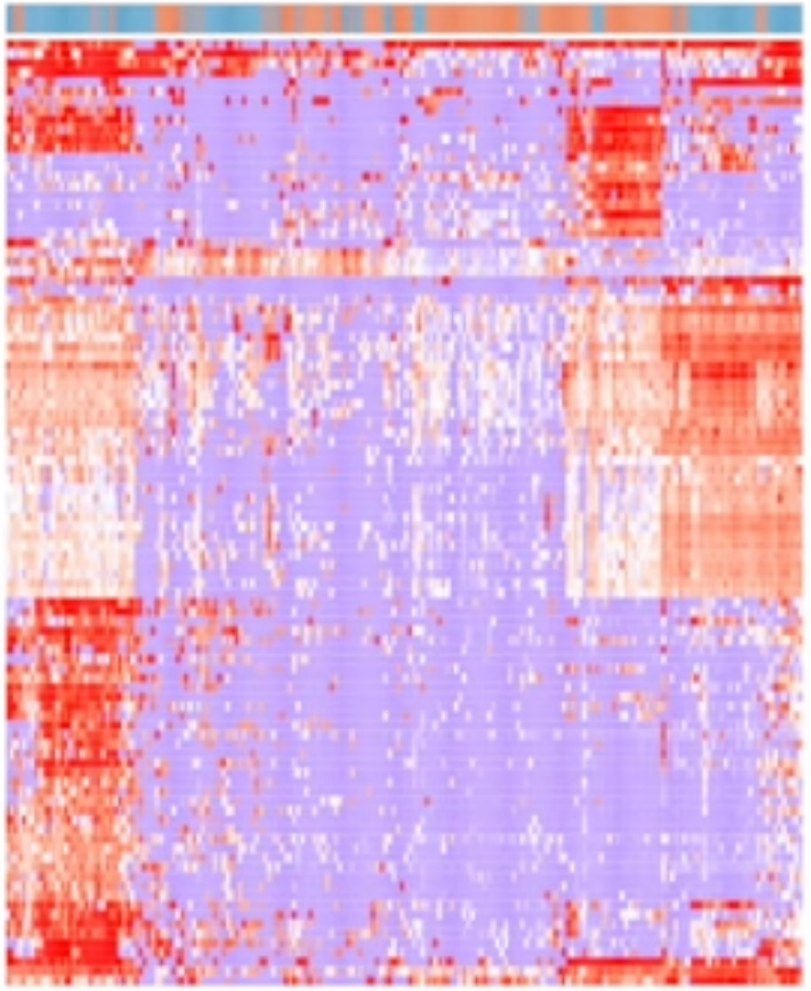
Here's a heatmap that compares the **furthest** points in the clusters.



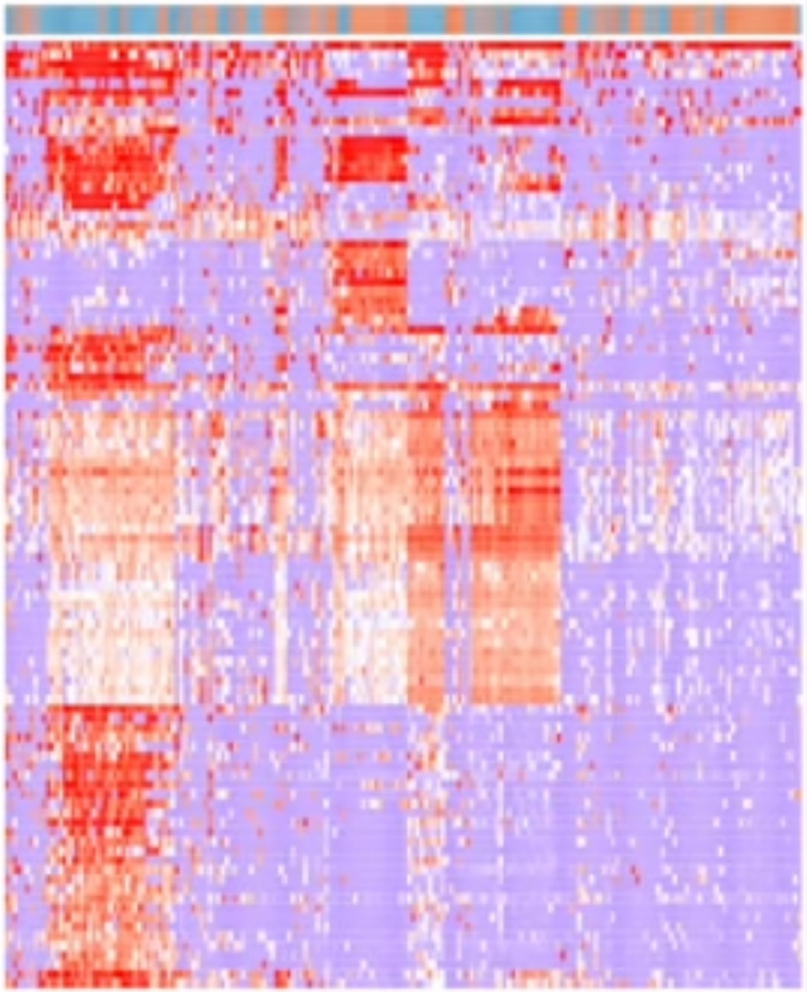
This one compares the **average** points in the clusters.



Here's a heatmap that compares the **furthest** points in the clusters.



This one compares the **average** points in the clusters.

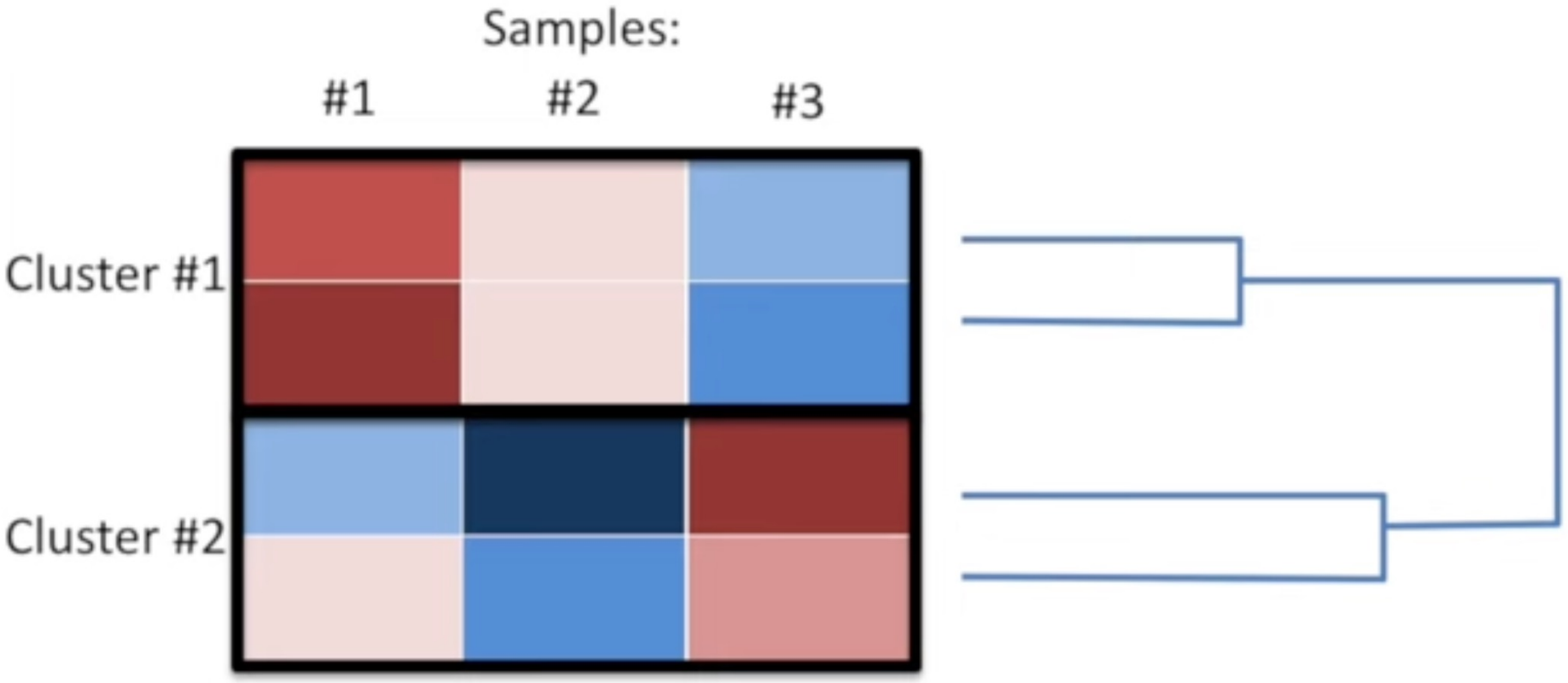


This one compares the **closest** points in the clusters.

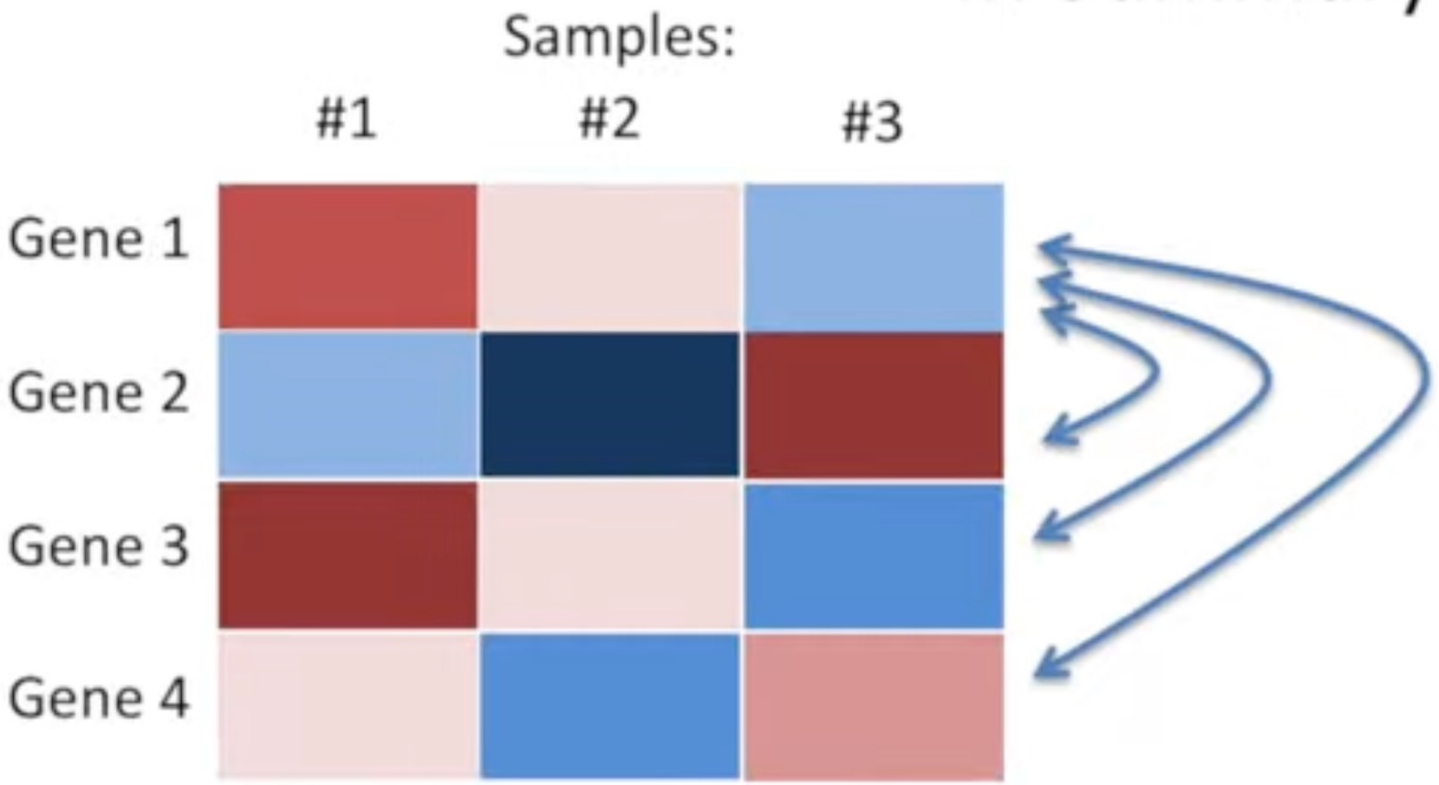
Summary

Hierarchical Clustering

In Summary

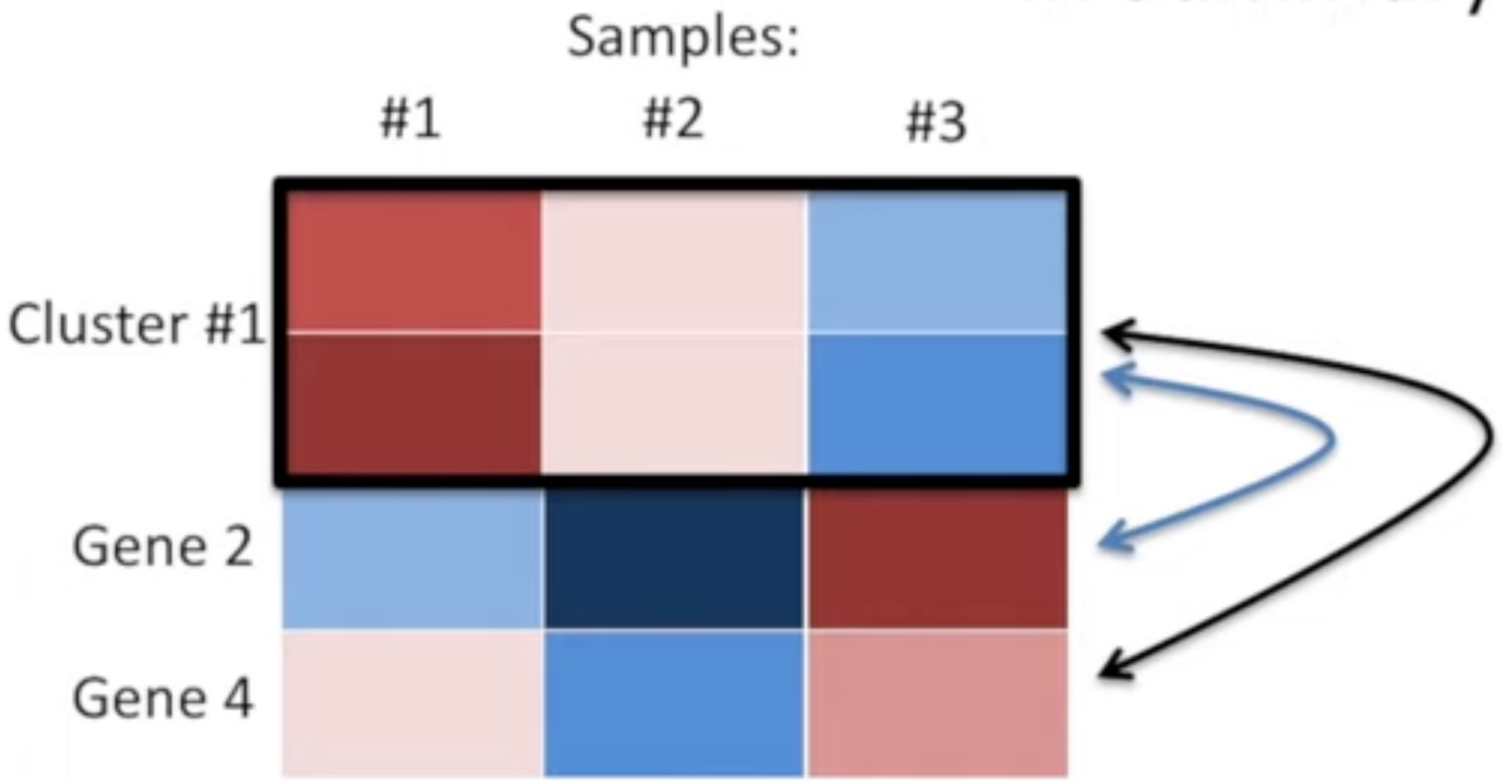


In Summary



1) Clusters are formed based on some notion of “similarity”.

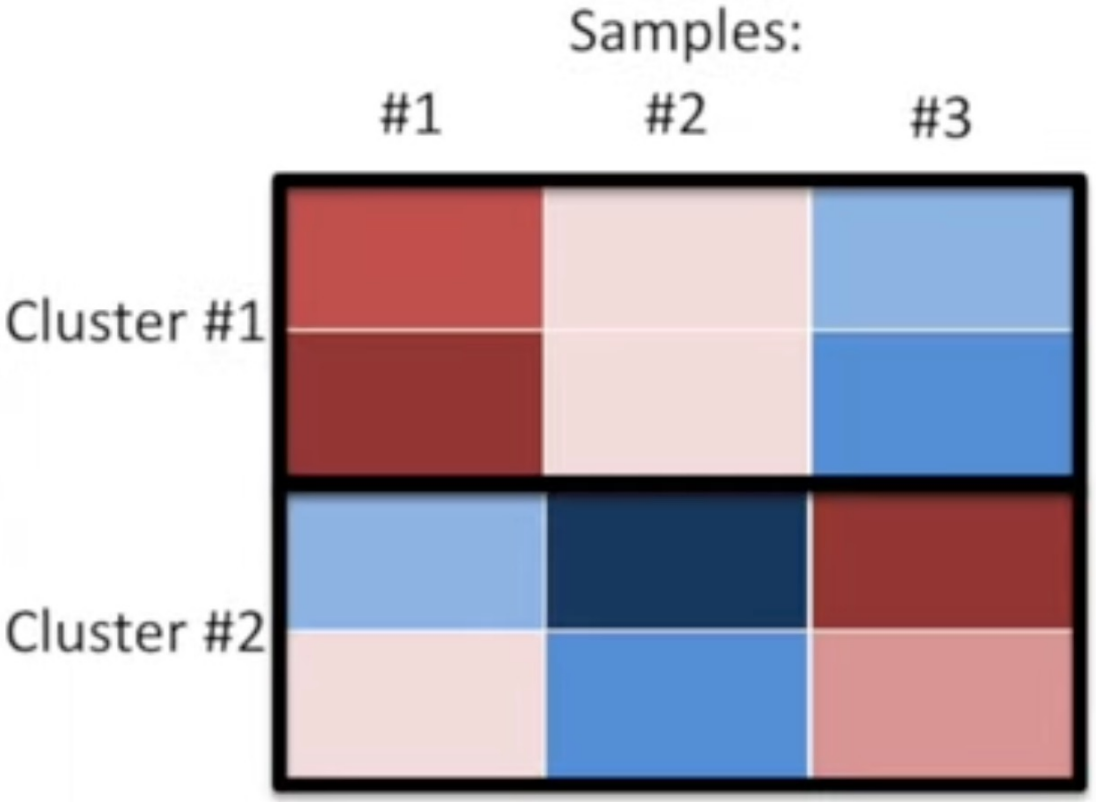
In Summary



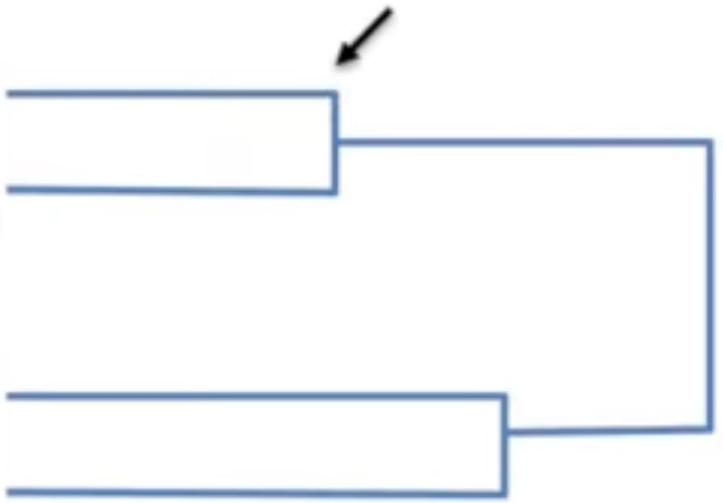
1) Clusters are formed based on some notion of “similarity”.

2) Once you have a “sub-cluster”, you have to decide how it should be compared to other rows/columns/sub-clusters/etc.

In Summary



And the height of the branches in the "dendrogram" shows you what is most similar…



Machine Learning

Hierarchical Clustering

Phd. César Astudillo | Facultad de Ingeniería