STATISTICS AND BIG DATA '25-'26

Hierarchical Clustering

Dr. Niccolò Salvini

Adjunct Professor @UCSC campus Rome,
Sr. Data Scientist

Hierarchical clustering concepts

1 clusters with heatmap

2 hierarchical clust

pseudocode in R

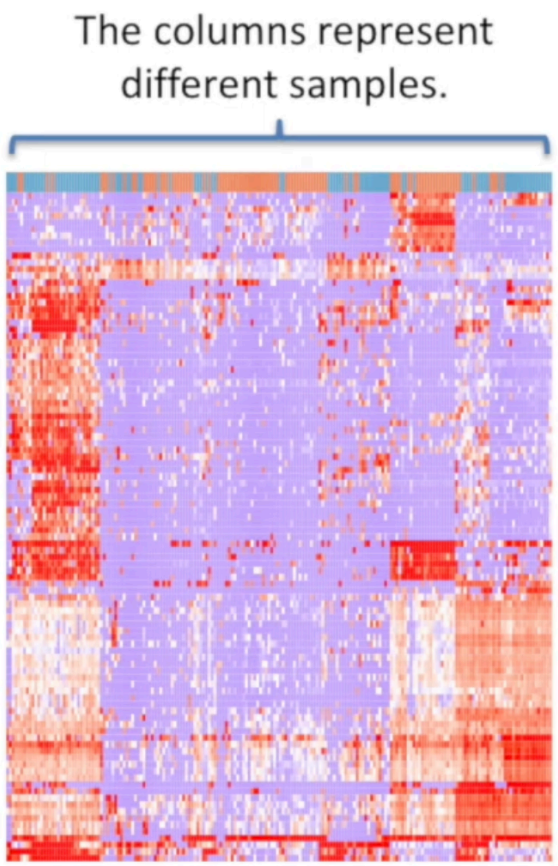
3

live coding session!

Section 1

Principles

cluster with **heatmaps**

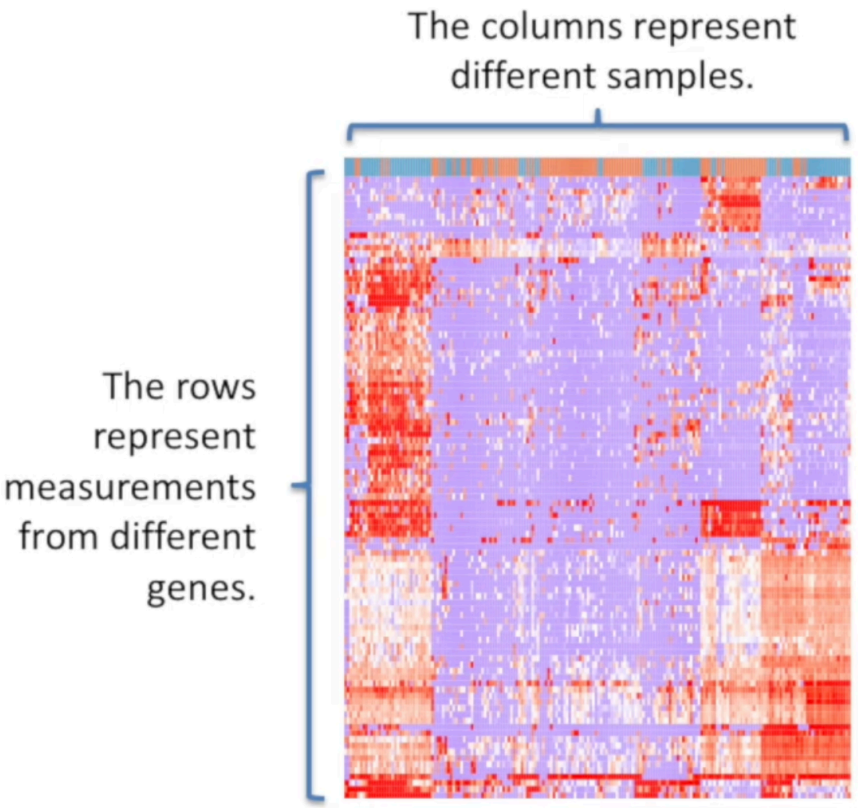


heatmaps

Have you ever seen an heatmap?

"A heatmap is a graphical representation of data that uses color-coded cells to represent different values. Heatmaps are typically used to visualize the distribution of data across two or more variables, with the color of each cell representing the value of a particular variable. Heatmaps can also be used to show correlations"

cluster with **heatmaps**

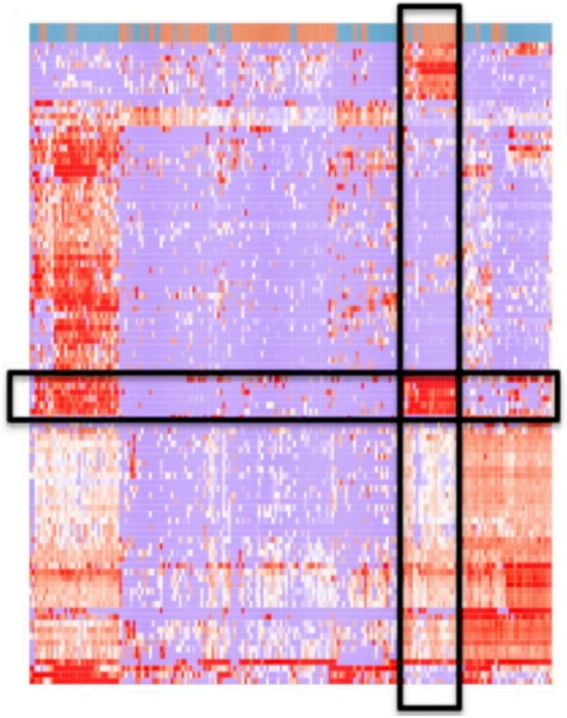


how to read this?

columns are different samples (data points). columnas are measuramente for different genes (the categorical variable)

cluster with **heatmaps**

These samples express the same genes



e genes
ıave the
same.

Hierarchical clustering or the rows and/or the colu based on similarity.

This makes it easy to see correlations in the data.

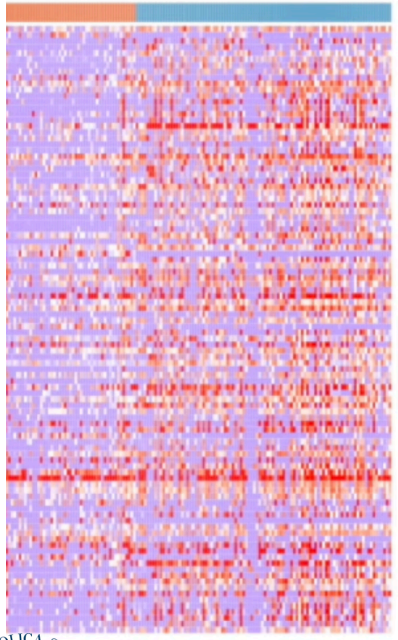
the cluster?

cluster are aggregation of data point. These aggregations may be random or based on some chars.

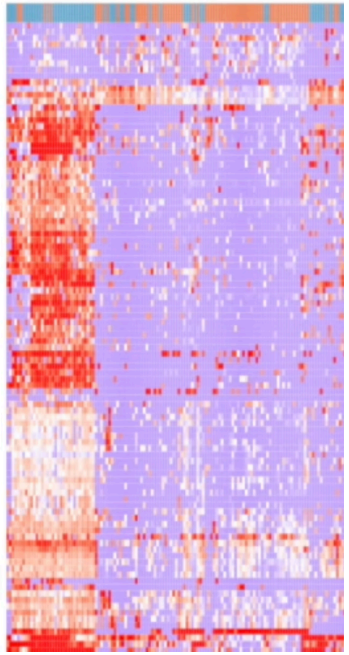
cluster with **heatmaps**

Hierarchical clustering is often associated with heatmaps.

ıt hierarchical clustering...

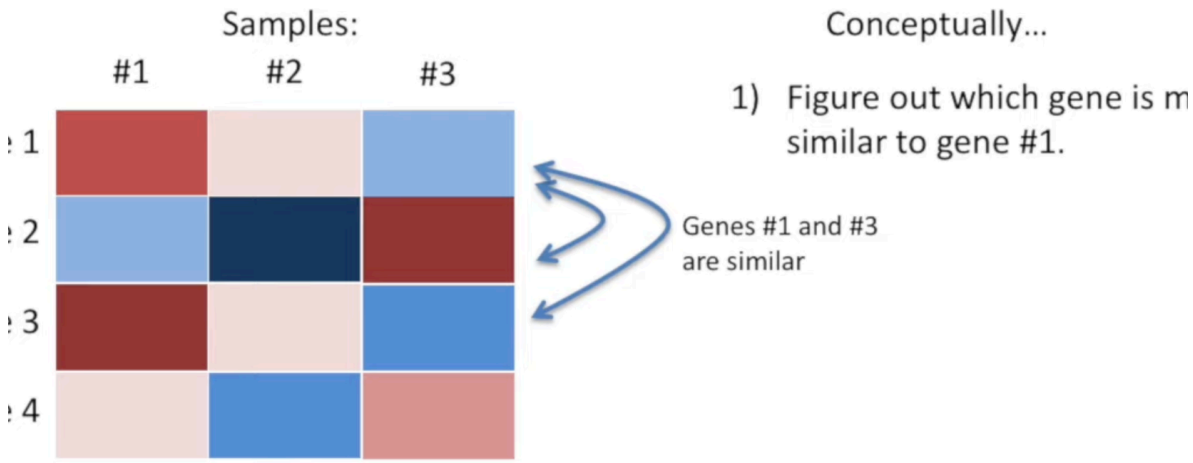


...with hierarchical cl



let's draw hier clust

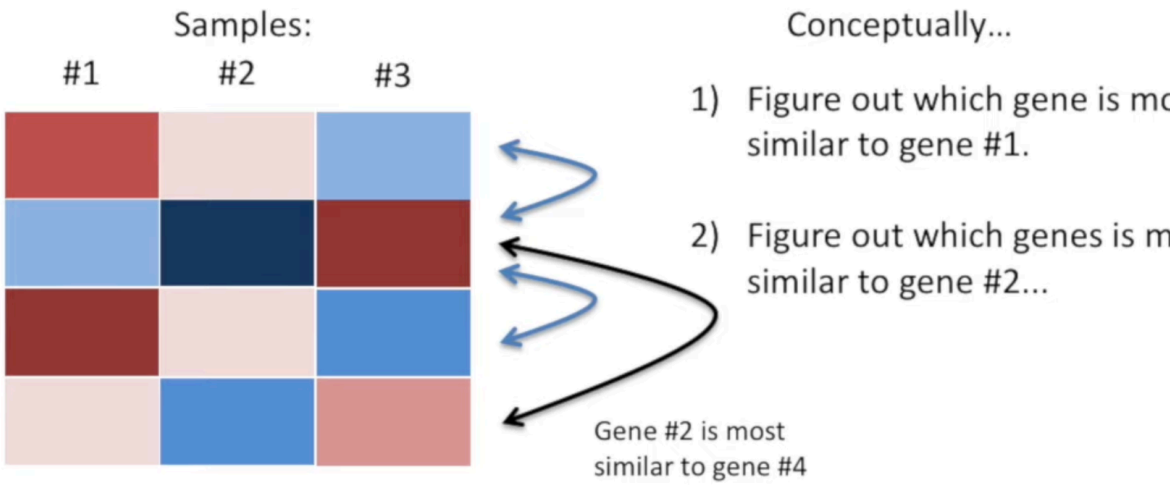
hierarchical clustering



give me data

these are 3 samples per 4 genes specifies.

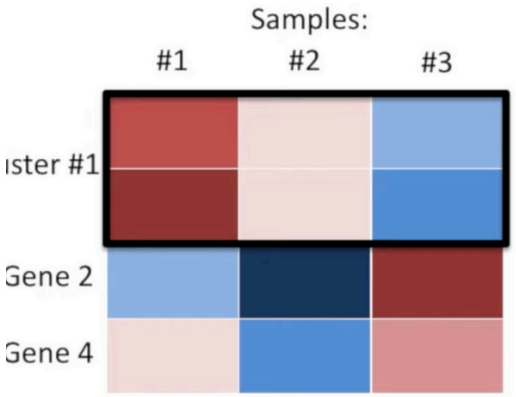
hierarchical clustering



HWY group it?

you do that pairwise and you select the pair ob obseration that are closer to each other.

hierarchical clustering



Conceptually...

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

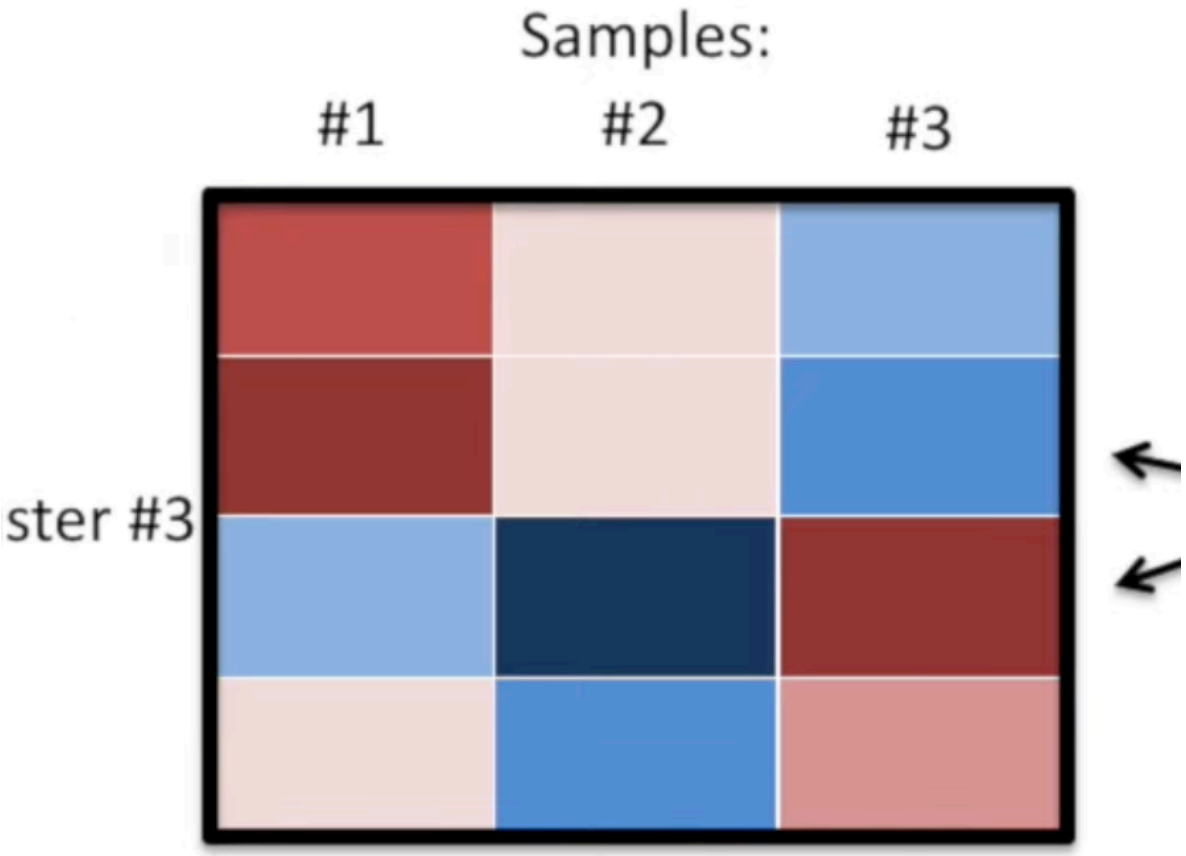
2) Figure out which genes is most similar to gene #2... (and then #3 and then #4).

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

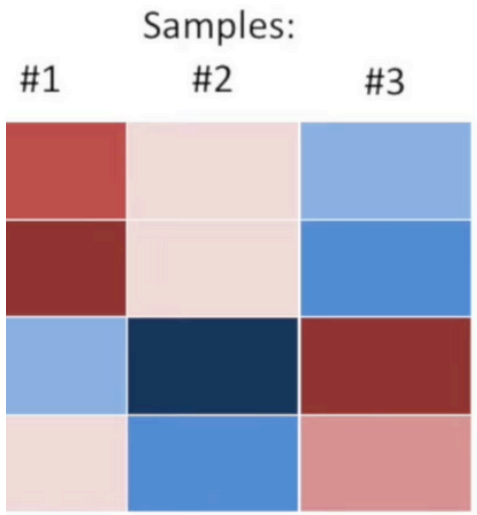
4) Go back to step 1, but now treat th new cluster like it's a single gene.

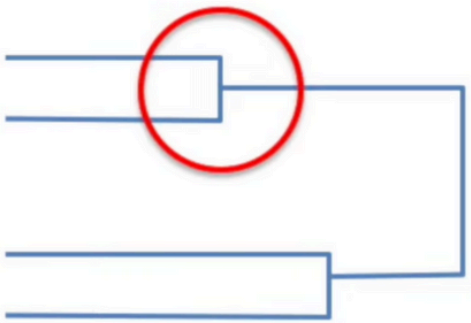
gene #1 and #3 are the closest

hierarchical clustering



hierarchical clustering



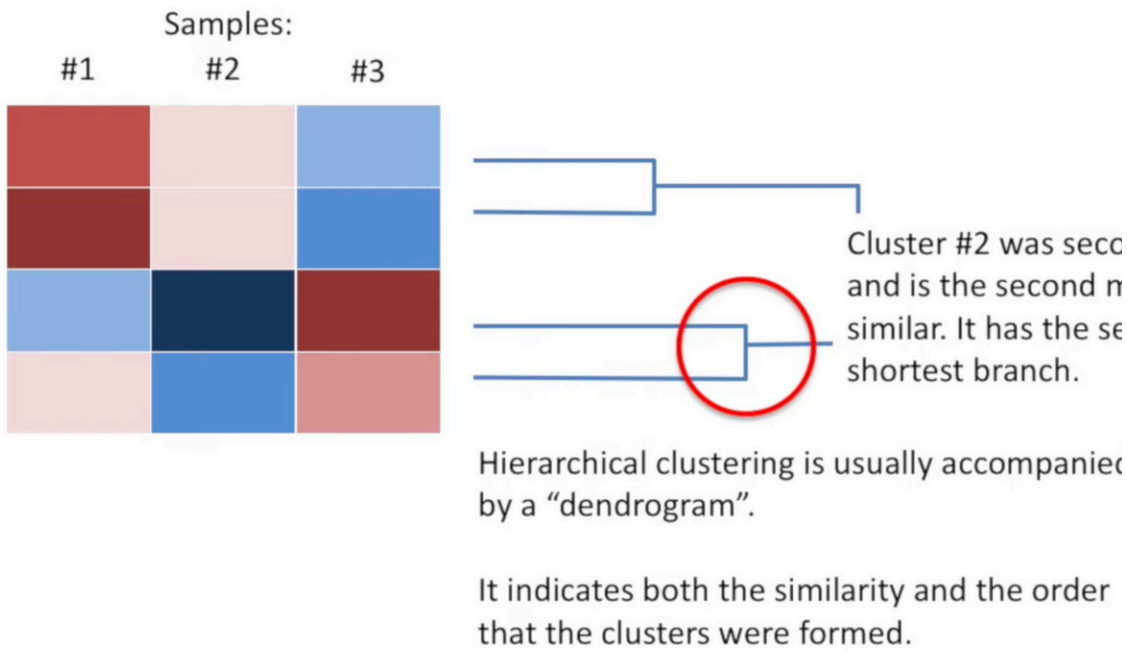


Cluster #1 was formed first and is
It has the sh

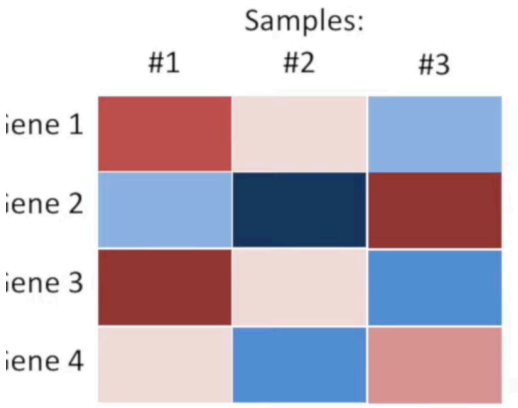
Hierarchical clustering is usually accomp
by a “dendrogram”.

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

hierarchical clustering



hierarchical clustering



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.

hierarchical clustering

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

The Euclidean distance between Genes 1 and 2.

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

hierarchical clustering

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

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

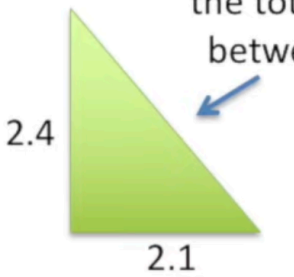
hierarchical clustering

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

$$\sqrt{(1.6-(-0.5))^2+(0.5-(-1.9))^2}$$

hierarchical clustering

	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.

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

Section 2

Pseudocode in R

```r
# Finding distance matrix
distance_mat <- dist(mtcars, method = 'euclidean')

set.seed(28)  # Setting seed
mtcats_hiercluster <- hclust(distance_mat, method = "complete")

# Plotting dendrogram
plot(mtcats_hiercluster)

# prune tree (to the best 3 clusters)
sub_grps <- cutree(mtcats_hiercluster, k = 3)
rect.hclust(mtcats_hiercluster, k = 3, border = 2:5)
```

Section 4

Live coding session!

JUMP TO RSTUDIO!