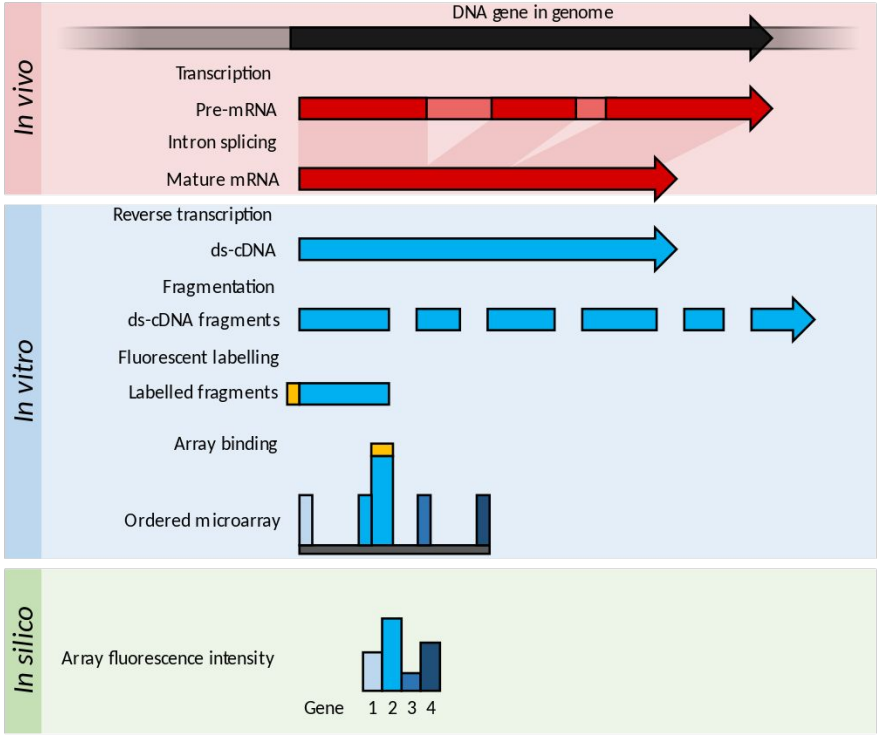
Lecture 14-15: Functional genomics

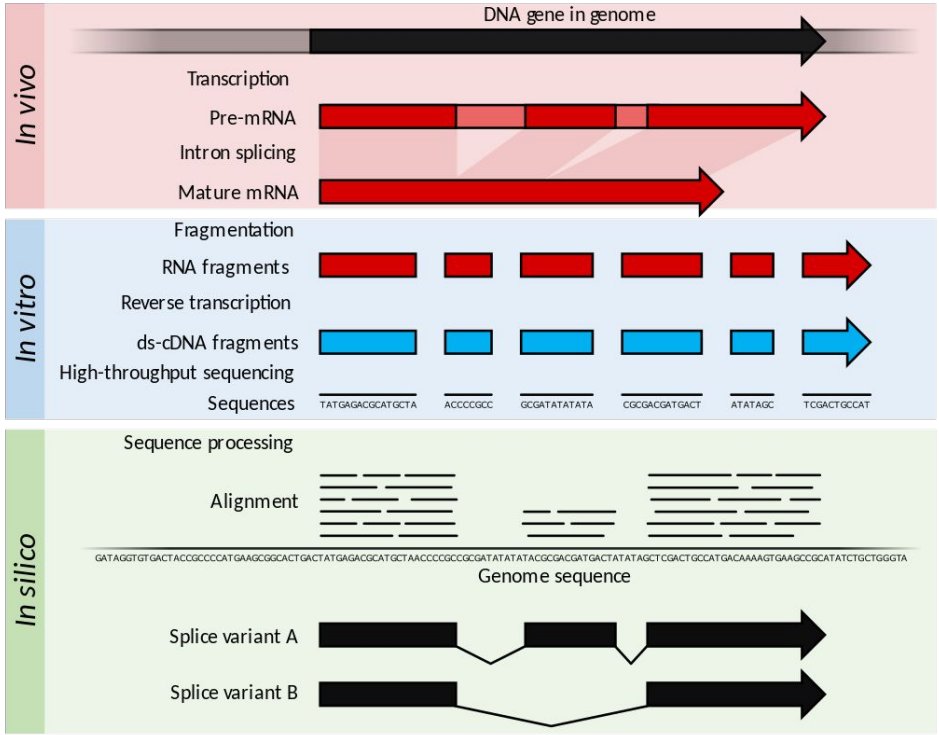
- Measuring gene-expression
- Distance measures
- Clustering & Dimension reduction
- Classification

Measuring gene-expression on a large-scale

DNA microarrays

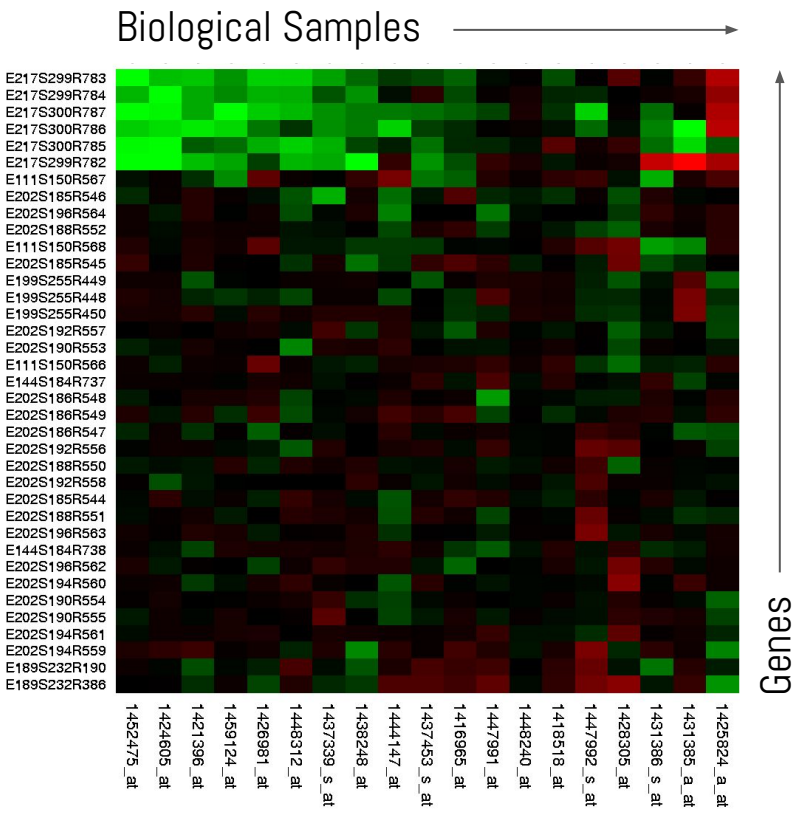


RNA-seq



Wikipedia

Measuring gene-expression on a large-scale



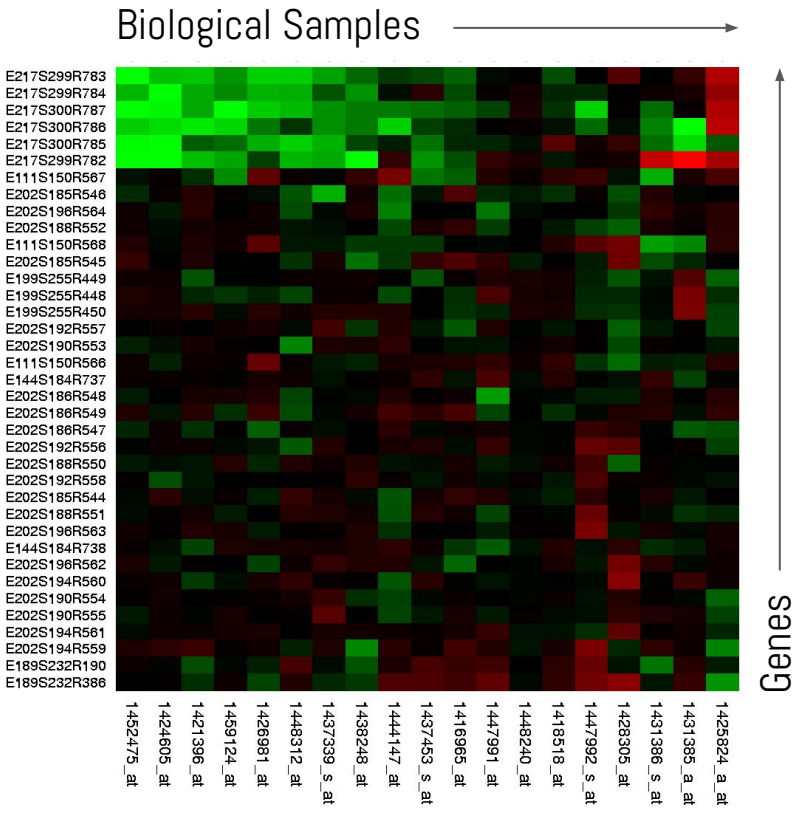
Gene-level Qs:

1. What's expressed (& by how much) in a given context/condition?
2. What's differentially expressed between two (or more) contexts/conditions?

Group-level Qs:

1. Are there groups of genes that respond similarly to changing contexts (across samples)?
2. Are there groups of samples that have very similar gene expression profiles?

Calculating “distance” between genes or samples



Variables

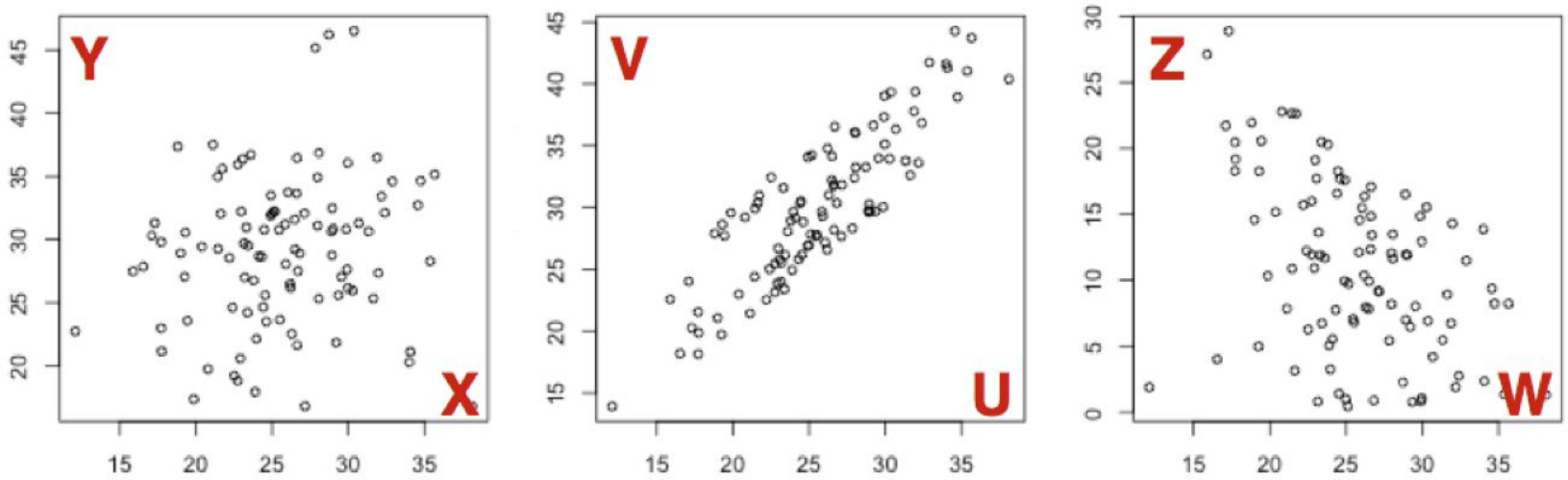
Attributes / Features

x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68

Calculating “distance” between genes or samples

Variables Attributes / Features

x	10	8	13	9	11	14	6	4	12	7	5
y	8.04	6.95	7.58	8.81	8.33	9.96	7.24	4.26	10.84	4.82	5.68



Distance measures

Pearson Correlation Coefficient

- Measures 'linear' relationship between variables.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

where:

- n is the sample size
- x_i, y_i are the single samples indexed with *i*
- $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (the sample mean); and analogously for $\bar{y}$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$

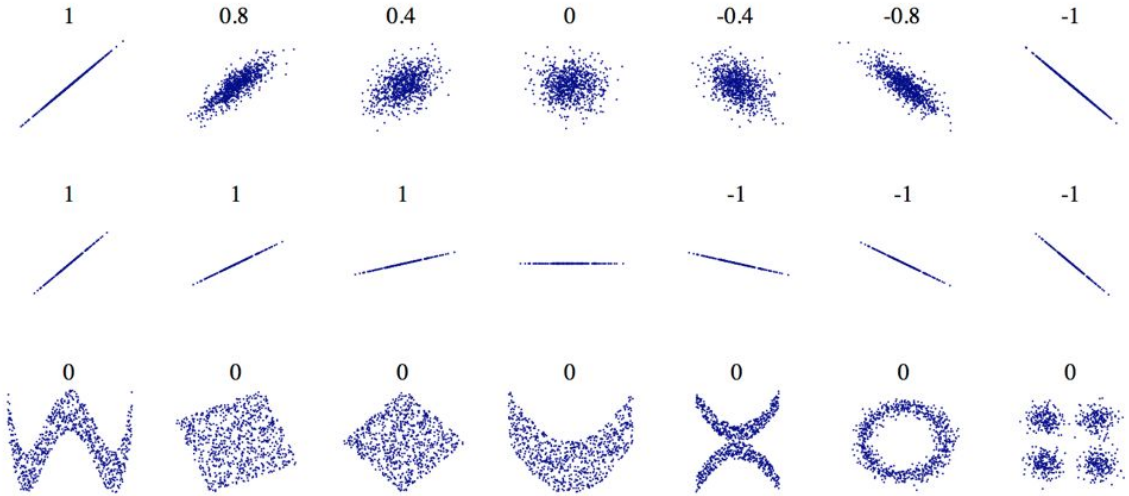
Wikipedia

Distance measures

Pearson Correlation Coefficient

- Measures 'linear' relationship between variables.

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right)$$



$-1 \leq r \leq +1$

−1 is total −ve correlation | 0 is no correlation | +1 is total +ve correlation

Wikipedia

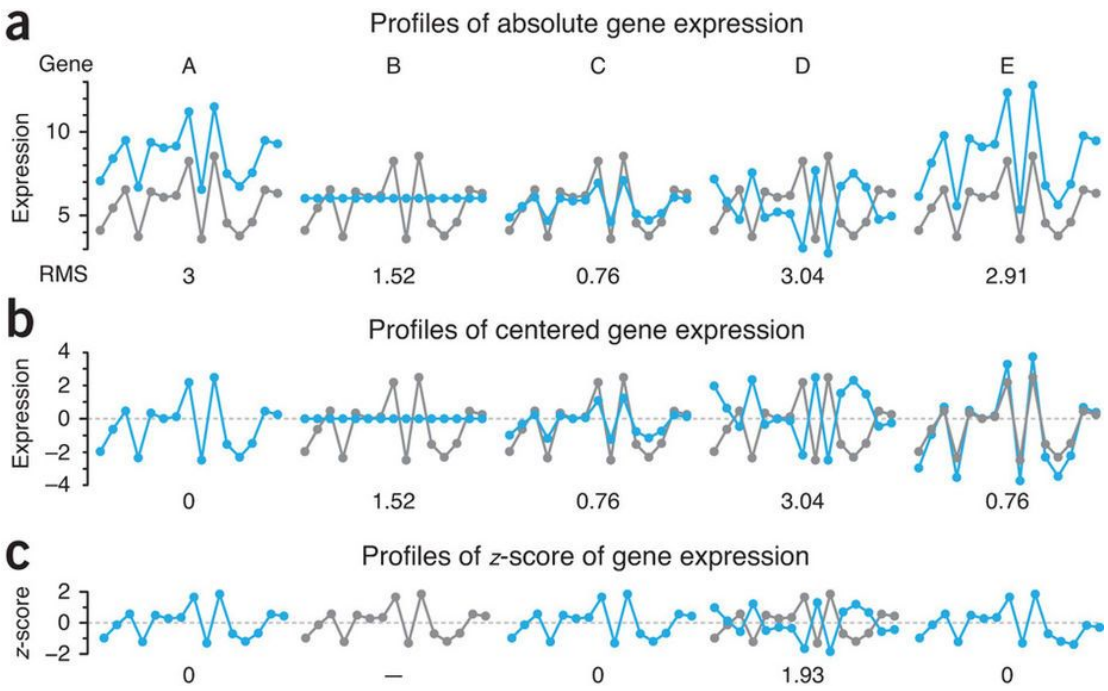
Distance measures

Pearson Correlation Coefficient

- Captures the relationship between 2 vectors after centering each vector by its mean and scaling by its standard deviation.

- The final quantities for each vector are called z-scores.

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s_x} \right) \left(\frac{y_i - \bar{y}}{s_y} \right)$$

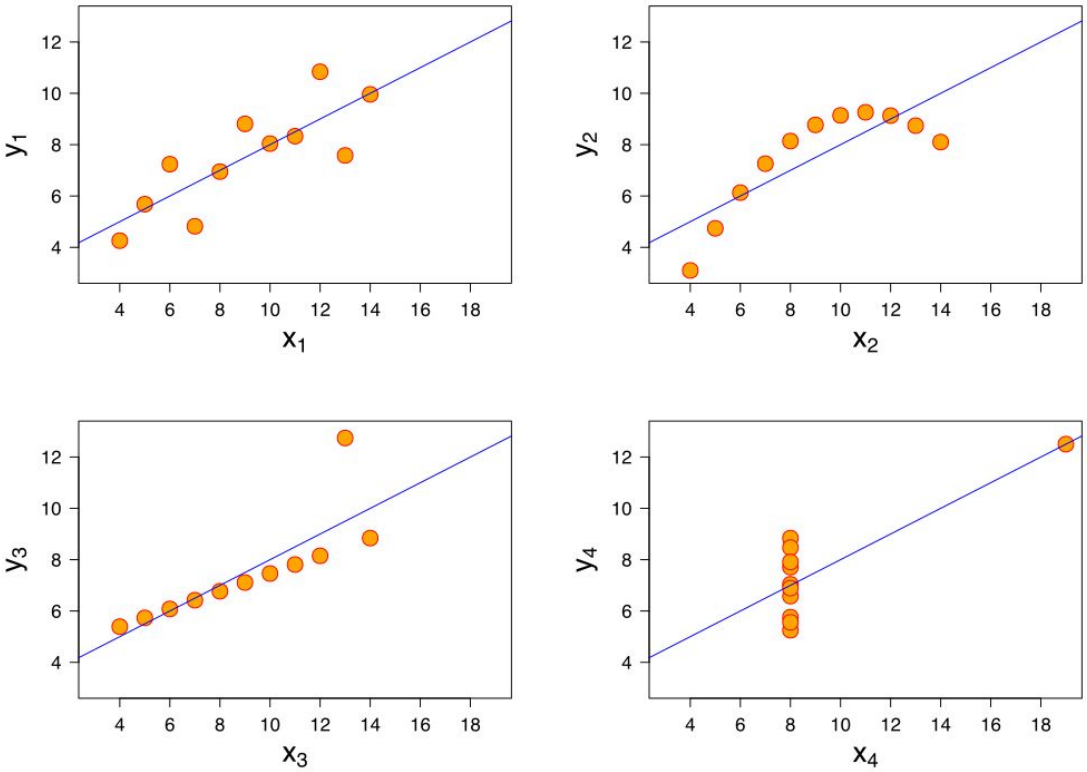


Altman & Krzywinski (2017) Nat. Meth.

Anscombe's quartet: "calculation are exact; graphs are rough!"

11 datapoints

- Mean (x) = 9
- Var (x) = 11
- Mean (y) = 7.50
- Var (y) ~ 4.12
- Cor (x, y) = 0.816
- Linear regression line:
  - y = 3.00 + 0.500x

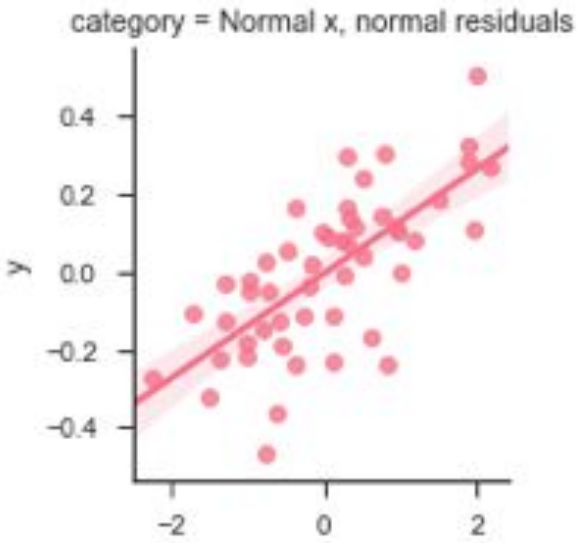


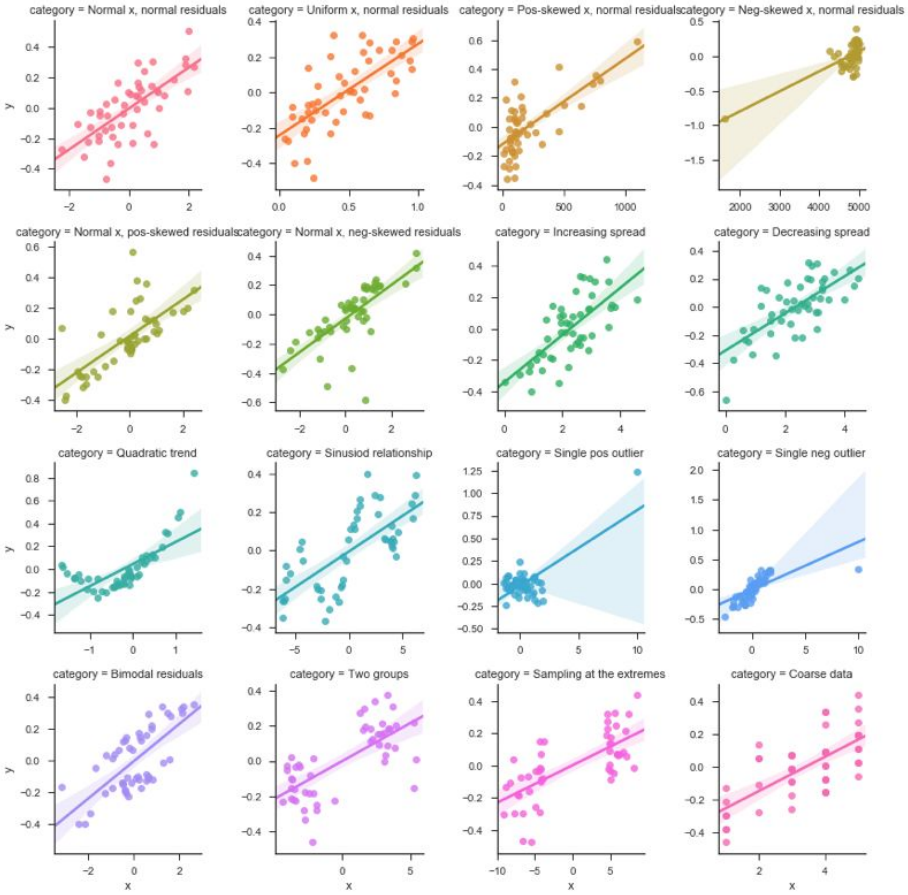
Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician 27 (1): 17–21.

Wikipedia

What does a correlation coefficient tell you about the data?

Correlation = 0.7

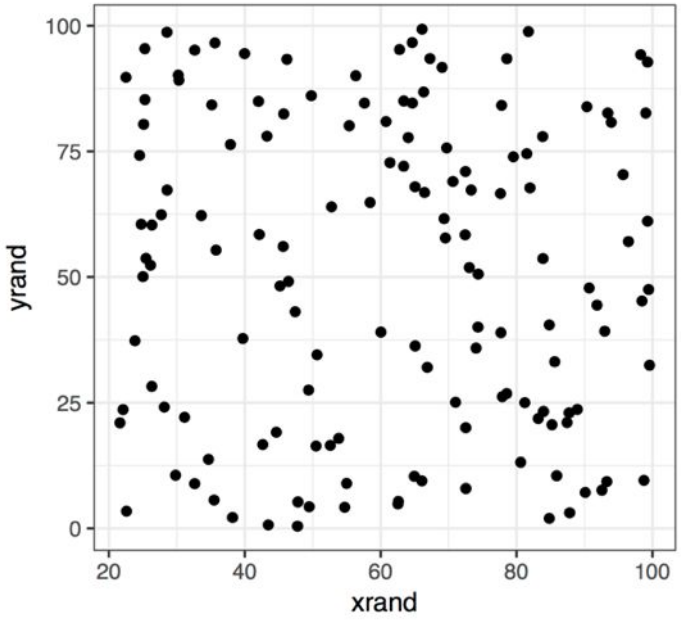




Wikipedia

What does a correlation coefficient tell you about the data?

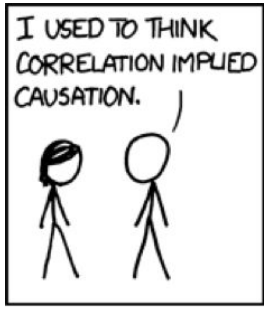
Correlation = -0.06

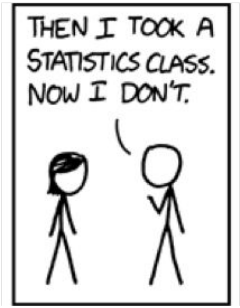


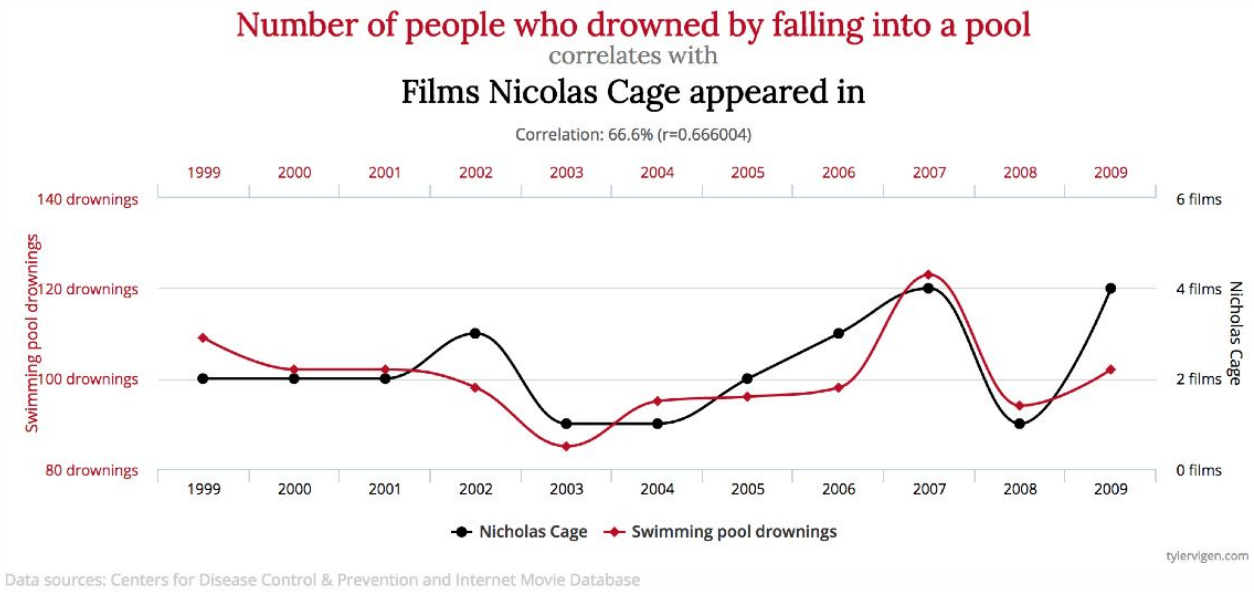
https://www.autodeskresearch.com/publications/samestats

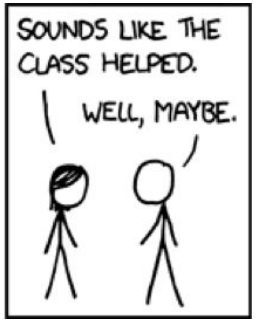
Spurious correlations

What does Nicholas Cage have to do with people drowning in swimming pools?









Checkout https://www.google.com/trends/correlate

xkcd.com/552

Many distance measures

Pearson Correlation Coefficient

Spearman Rank Correlation

Euclidean Distance

Mutual Information

...

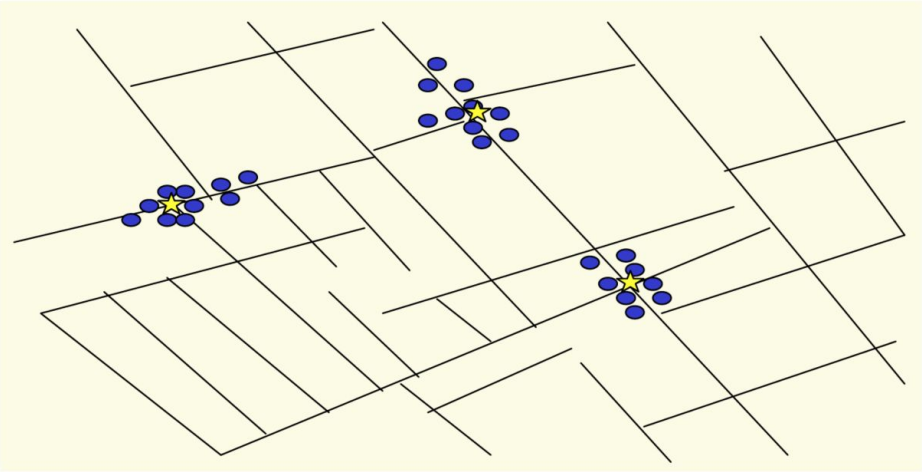
$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

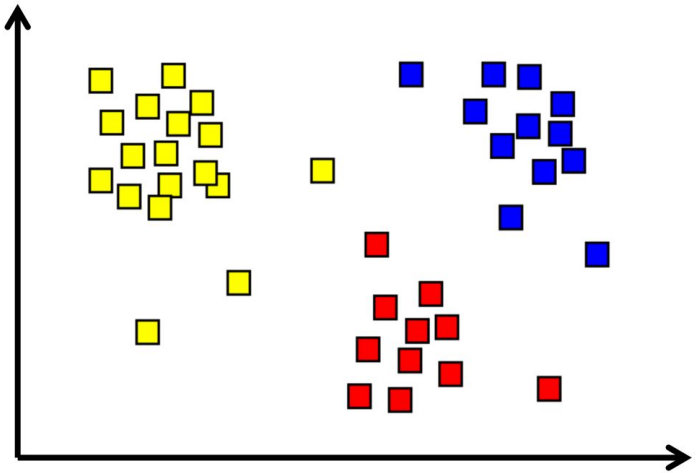
$$r = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x} \right) \left(\frac{y_i - \bar{y}}{\sigma_y} \right)$$

$$\rho = 1 - \frac{6 \sum_{i=1}^{n} [rank(x_i) - rank(y_i)]}{n(n^2 - 1)}$$

Wikipedia

Clustering

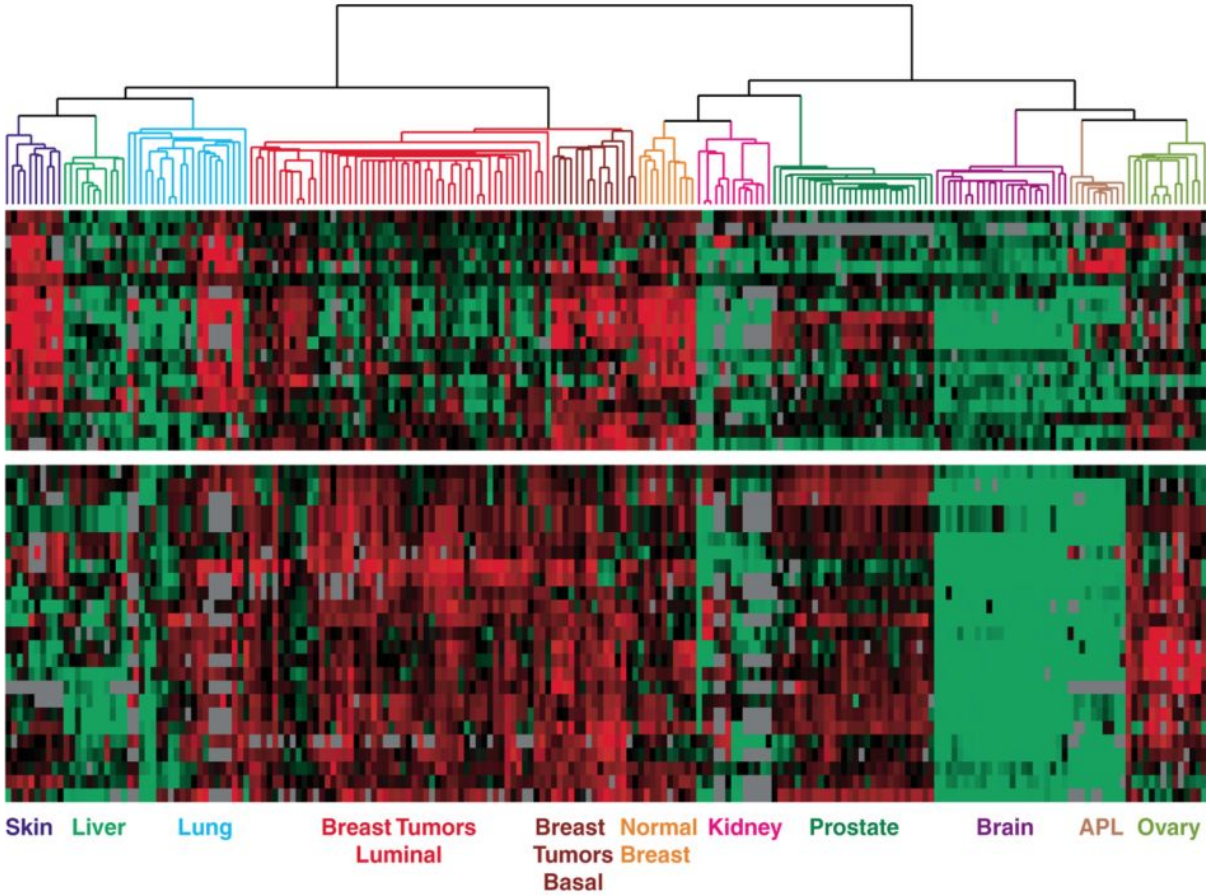




Group-level Qs:

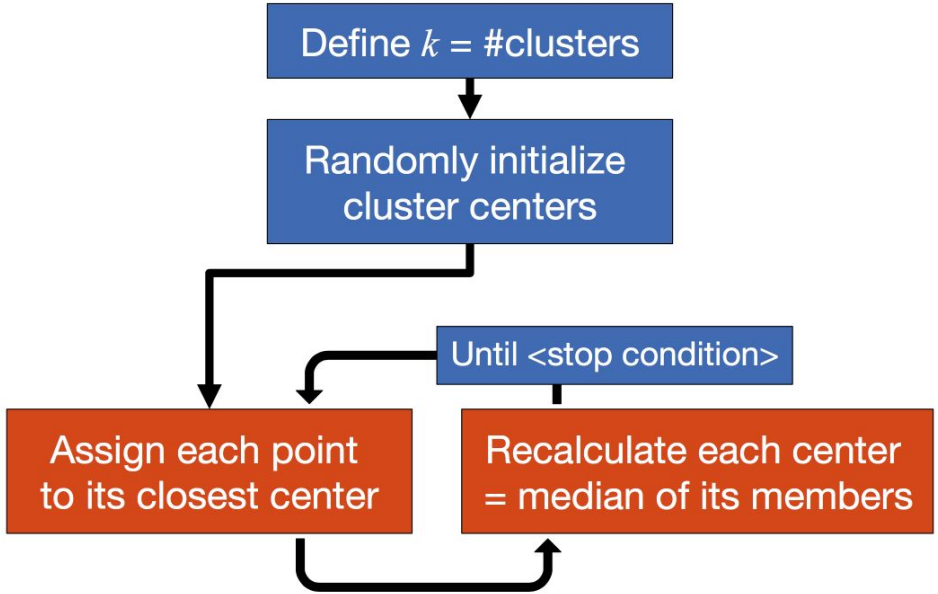
1. Are there groups of genes that respond similarly to changing contexts (across samples)?
2. Are there groups of samples that have very similar gene expression profiles?

Clustering



Wikipedia

K-means clustering



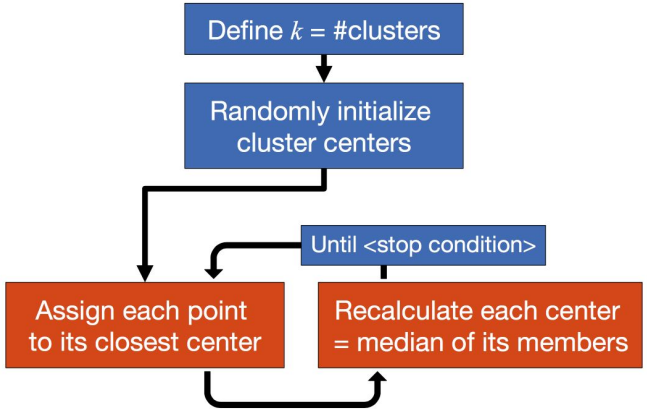
Conceptually similar to Expectation-Maximization, alternating between 2 two steps:

1. E step: Creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters.
2. M step: Computes parameters maximizing the expected log- likelihood found on the E step.

These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

Checkout http://www.naftaliharris.com/blog/visualizing-k-means-clustering/

K-means clustering



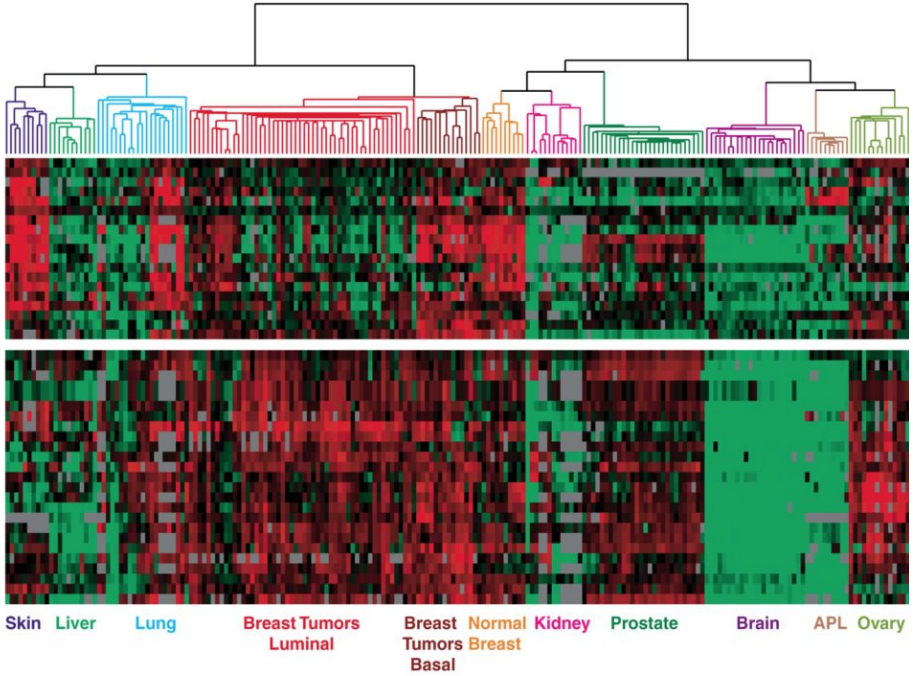
Stopping condition

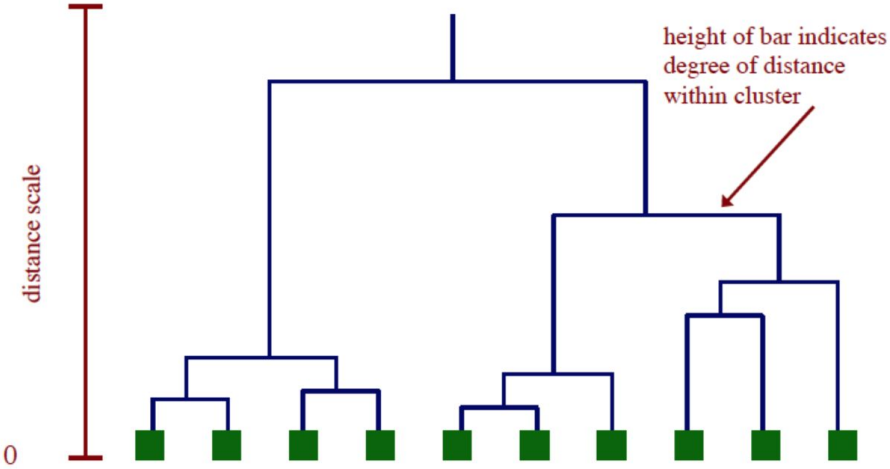
- Until the change in centers is less than <constant>.
- Until all genes get assigned to the same partition twice in a row.
- Until some minimal number of genes (e.g. 90%) get assigned to the same partition twice in a row.

Some issues

- Have to set k ahead of time.
- Works well if clusters of approx. similar sizes.
- Each gene only belongs to 1 cluster.
- Genes assigned to clusters on the basis of all experiments.

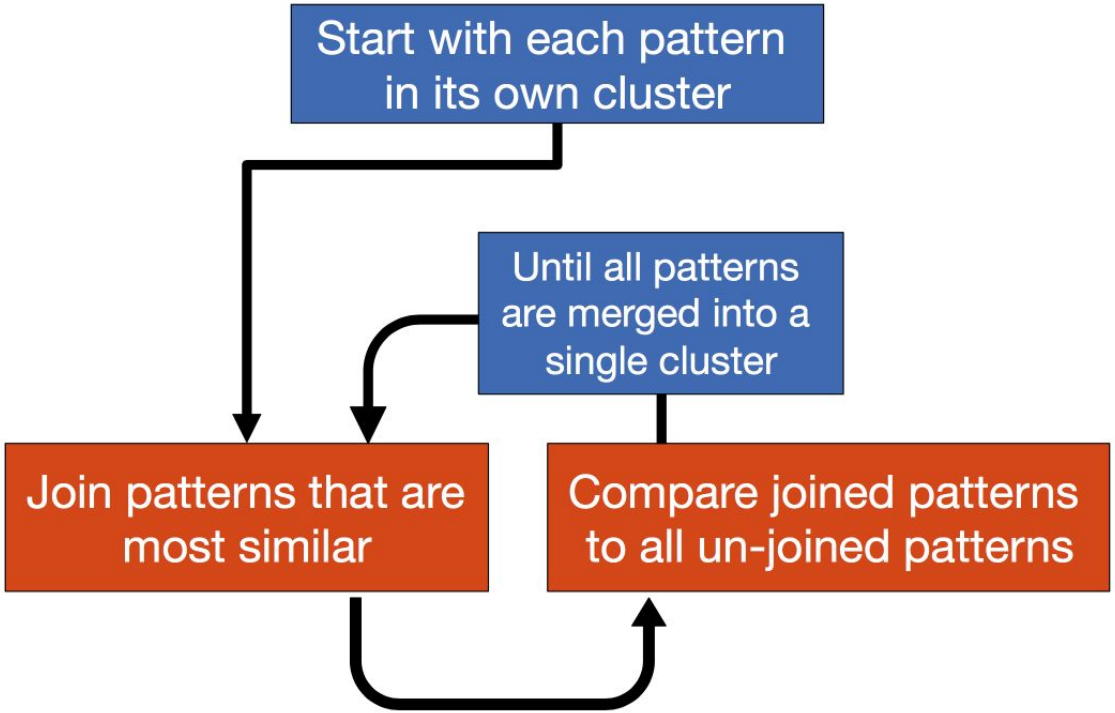
Hierarchical clustering



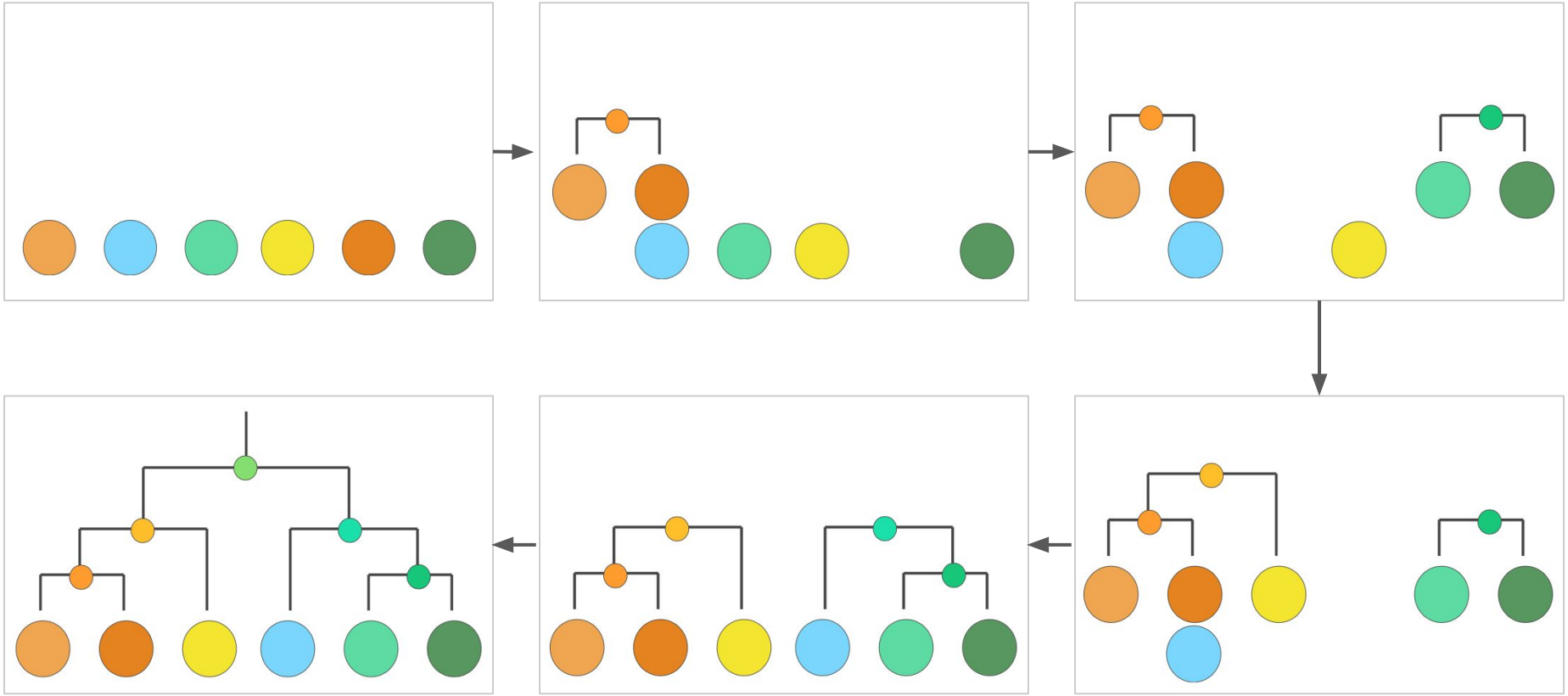


- Imposes hierarchical structure on all of the data.
- Easy visualization of similarities and differences between genes (experiments) and clusters of genes (experiments).

Hierarchical clustering



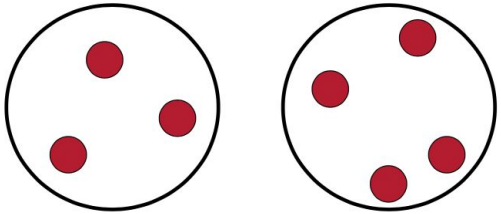
Hierarchical clustering



Hierarchical clustering

Linkage criteria:

- Single/Minimum linkage (nearest neighbors)
- Complete/Maximum linkage (farthest neighbors)
- Average linkage (average of all pairs)



Complete linkage clustering of 5 objects

1

	A	B	C	D	E
A	0				
B	5	0			
C	10	3	0		
D	15	6	7	0	
E	20	8	2	11	0

2

	A	B	CE	D
A	0			
B	5	0		
CE	20	8	0	
D	15	6	11	0

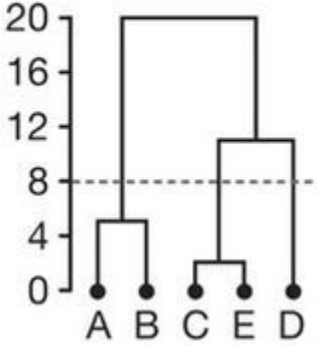
3

	AB	CE	D
AB	0		
CE	20	0	
D	15	11	0

4

	AB	CED
AB	0	
CED	20	0

Dendrogram

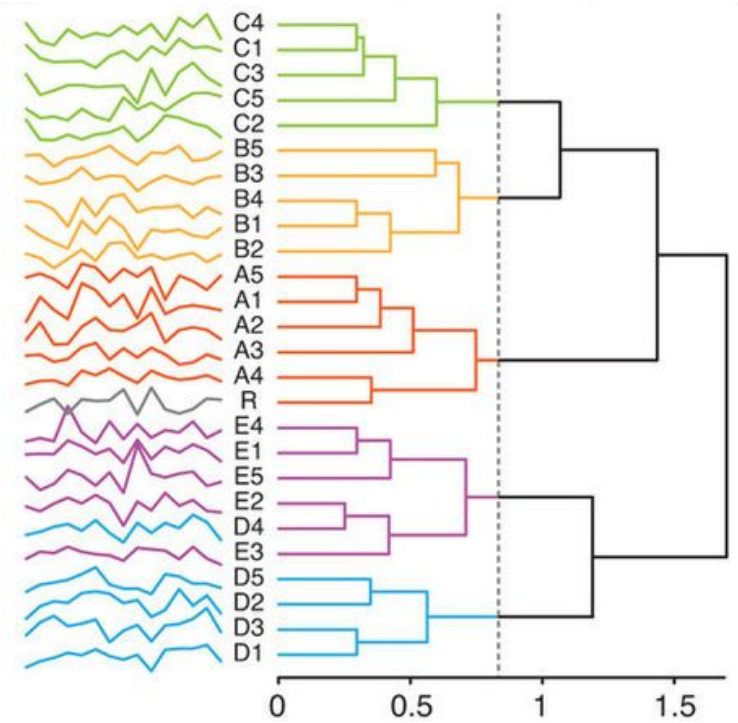


Altman & Krzywinski (2017) Nat. Meth.

Hierarchical clustering

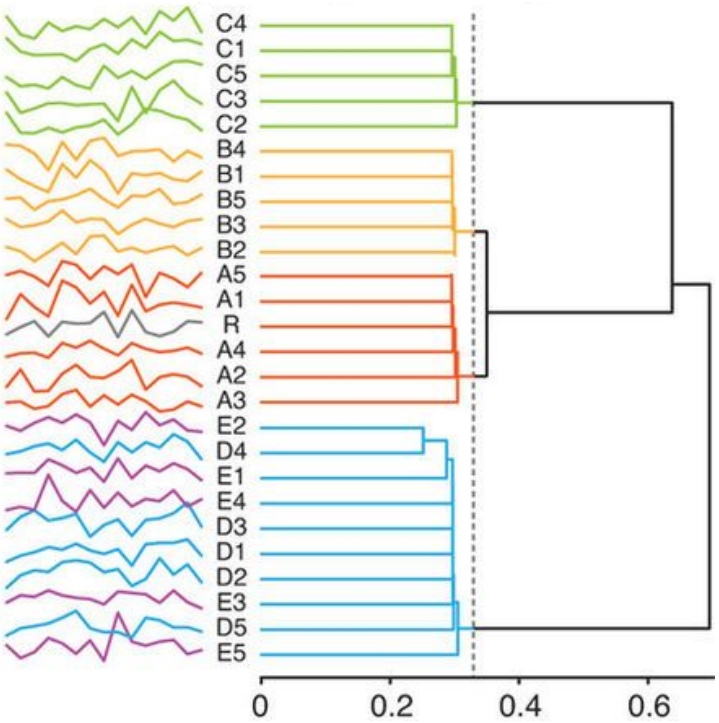
Complete linkage clustering

Tends to create balanced dendrograms by first clustering objects into small nodes and then clustering the nodes



Single linkage clustering

Tends to create stringy dendrograms by first creating a few nodes and then adding objects to them one at a time

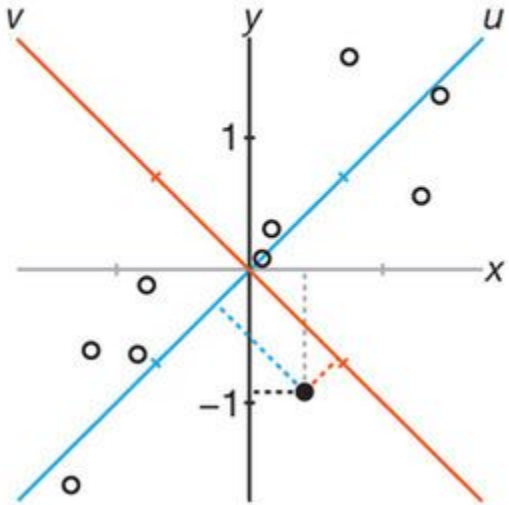


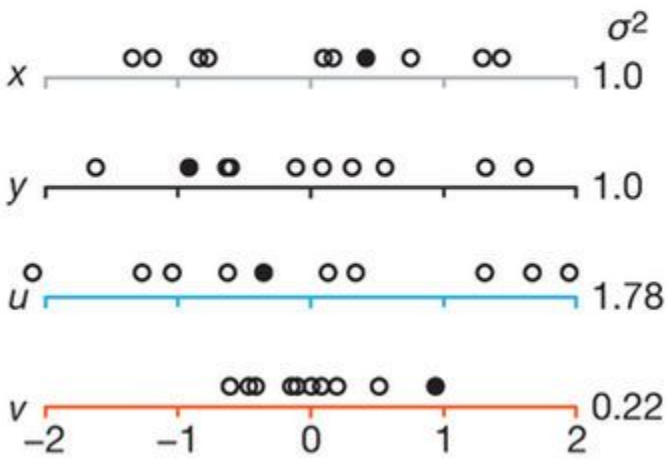
Altman & Krzywinski (2017) Nat. Meth.

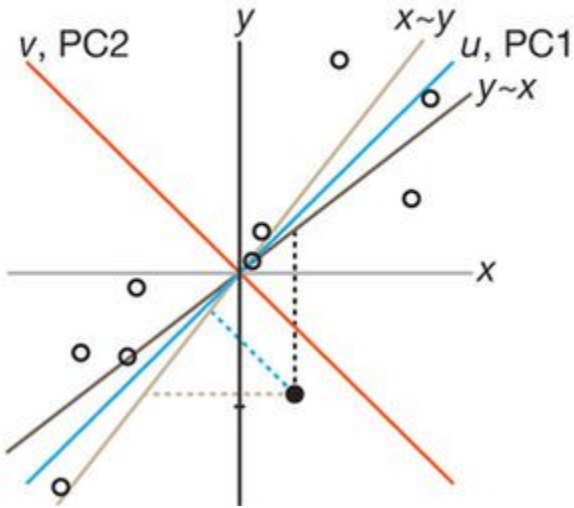
Dimensionality reduction by PCA

PCA geometrically projects data onto a lower-dimensional space

- Each lower dimension is a 'linear' combination of correlated original dimensions.
- The principal components (PCs) represent the directions of maximum variation.

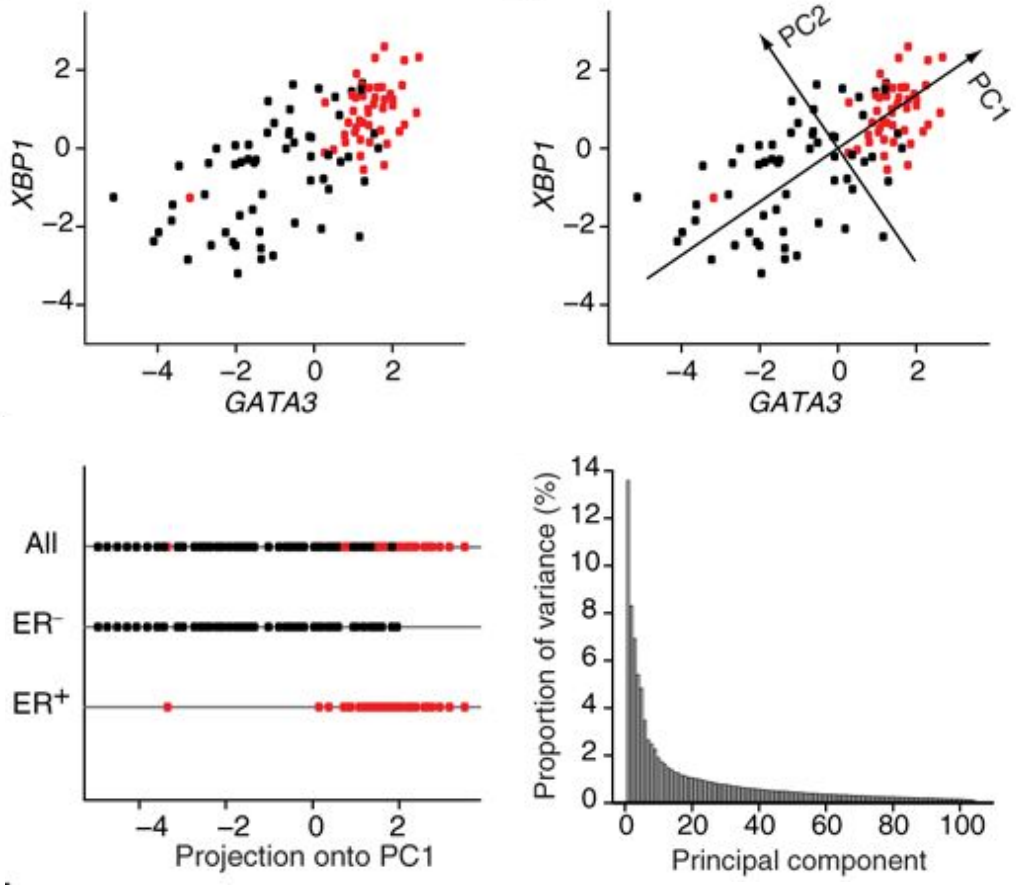






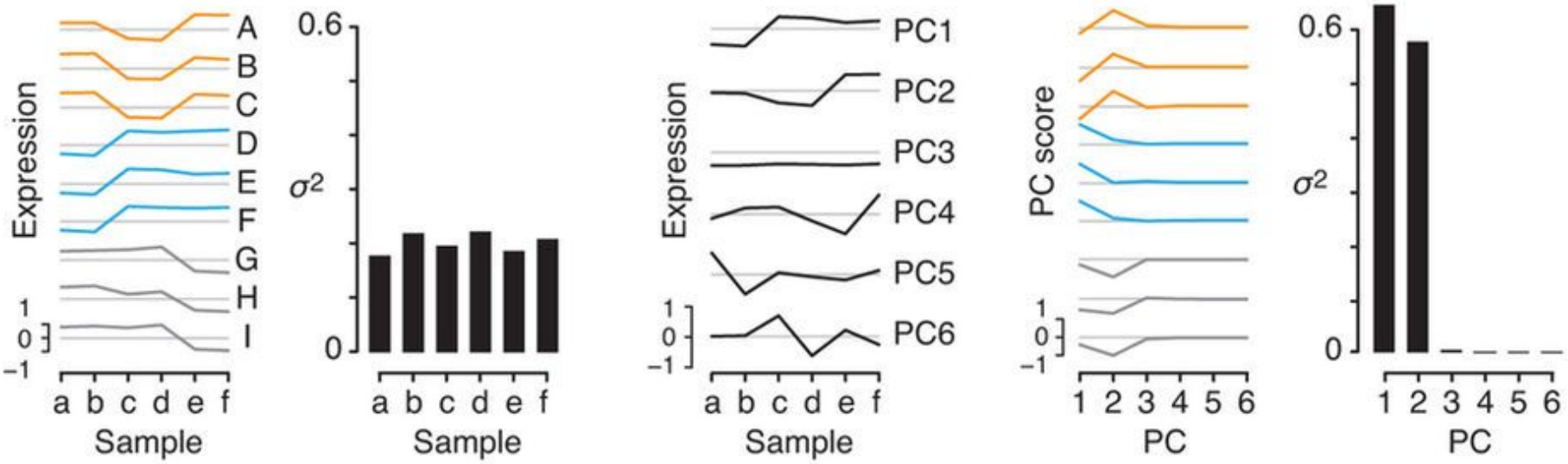
Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



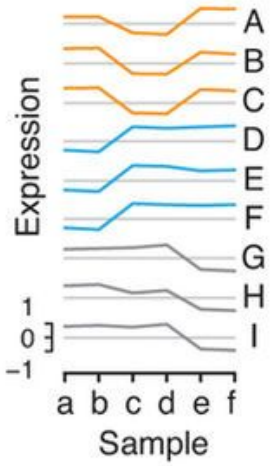
Ringnér (2008) Nat. Biotech

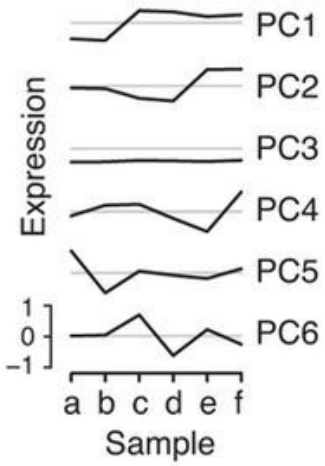
Dimensionality reduction by PCA

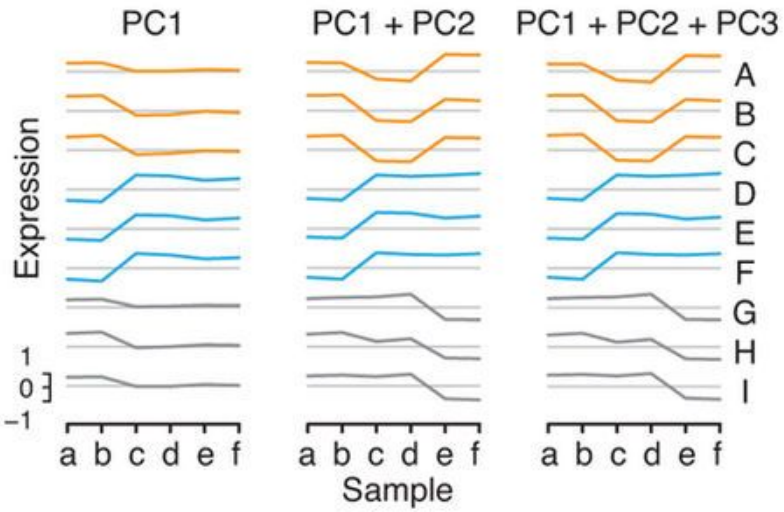


Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA

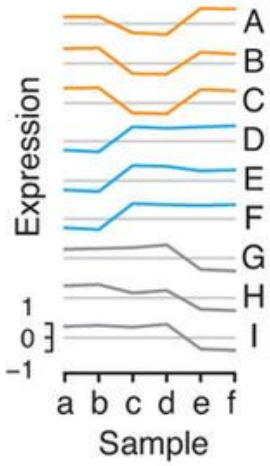


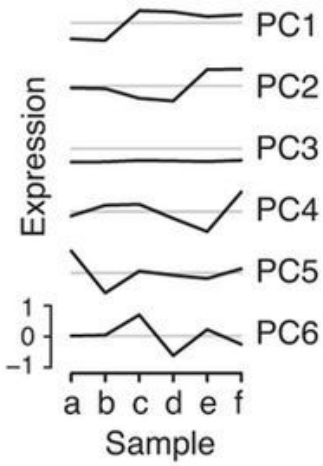


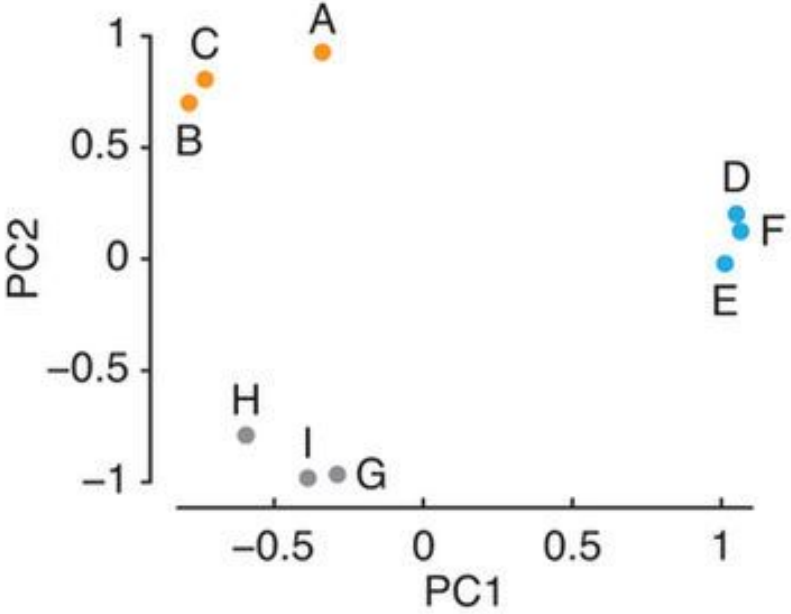


Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA

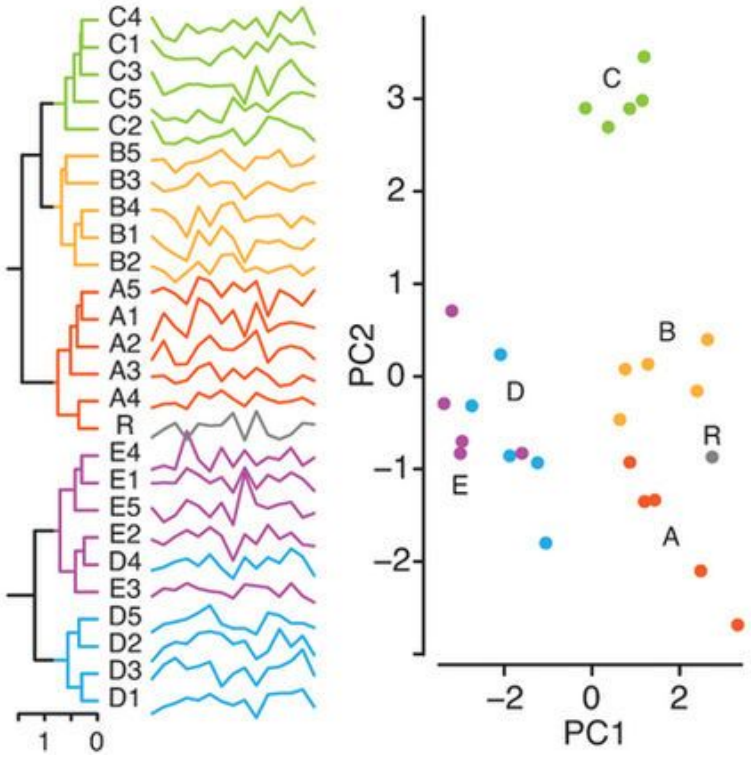






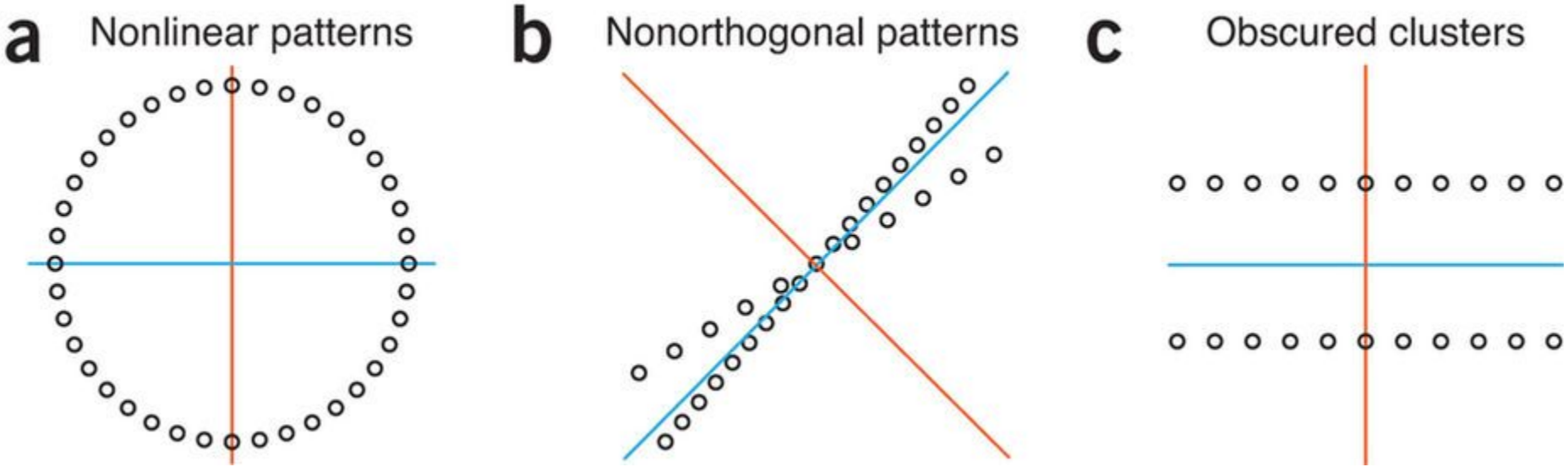
Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



Lever, Krzywinski, & Altman (2017) Nat. Meth.

Dimensionality reduction by PCA



Lever, Krzywinski, & Altman (2017) Nat. Meth.

Evaluating a clustering with external/prior knowledge

Actual \ Predicted	+	−
+	TP	FN Type II error
−	FP Type I error	TN

Sensitivity (recall)

TP/

False negative rate

FN/

False positive rate

FP/

Specificity

TN/

Precision

TP/

FDR

FP/

Accuracy

(TP + TN)/(+)

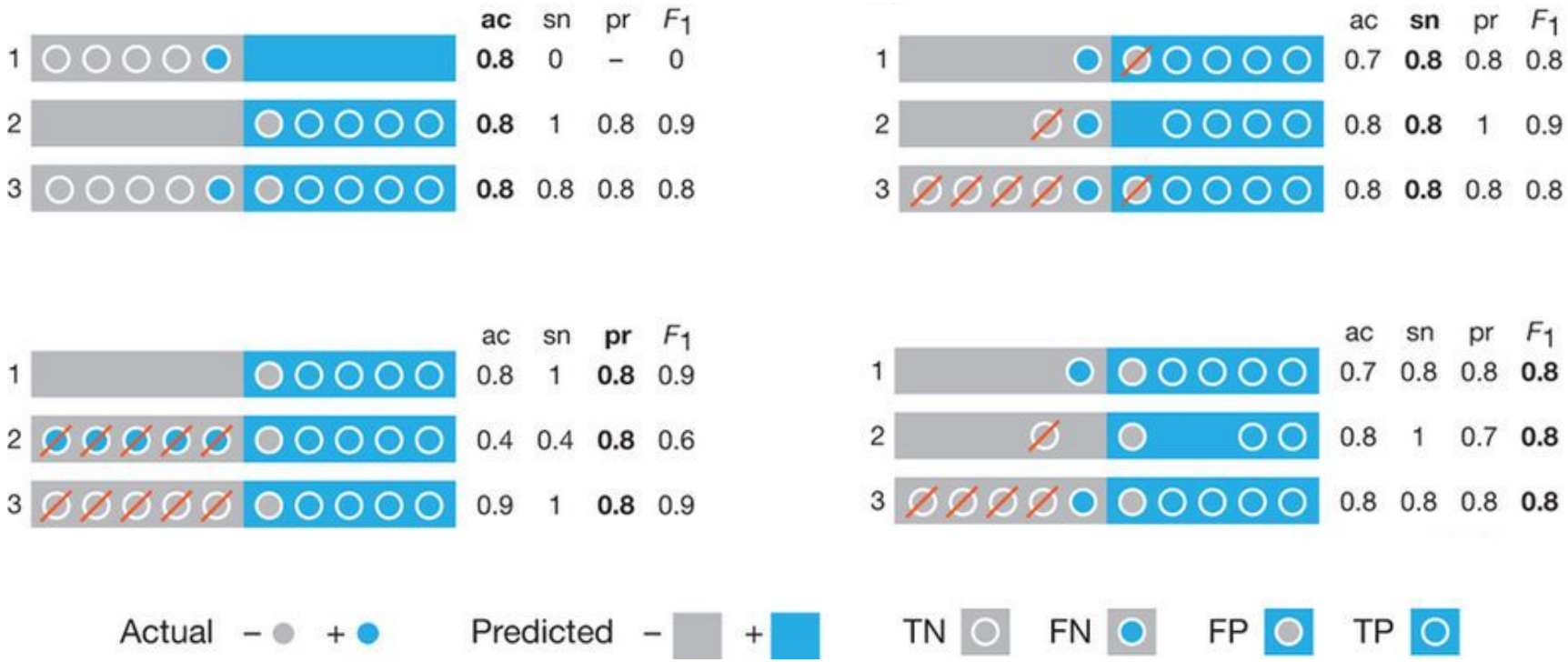
F_1 score

2TP/(2TP + FP + FN)

Lever, Krzywinski, & Altman (2016) Nat. Meth.

Evaluating a clustering with external/prior knowledge

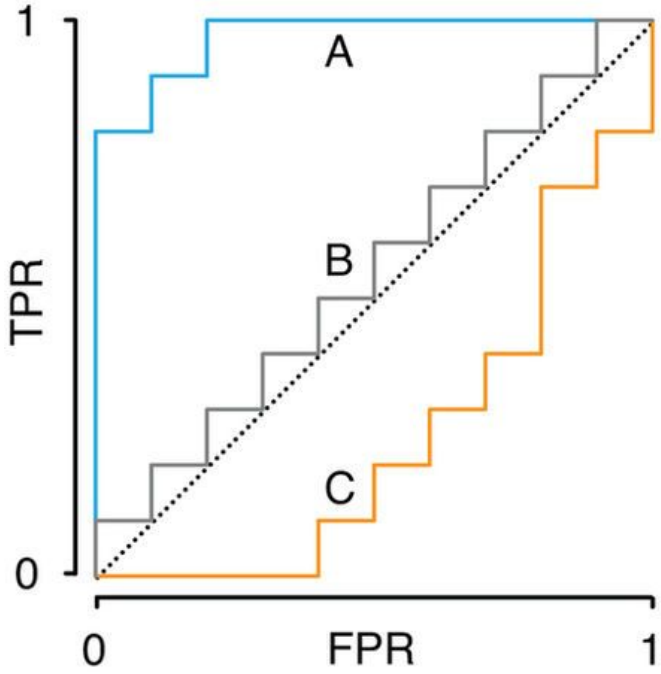
Four groups of 3 different classification scenarios that have the same value (0.8) for the metric in bold



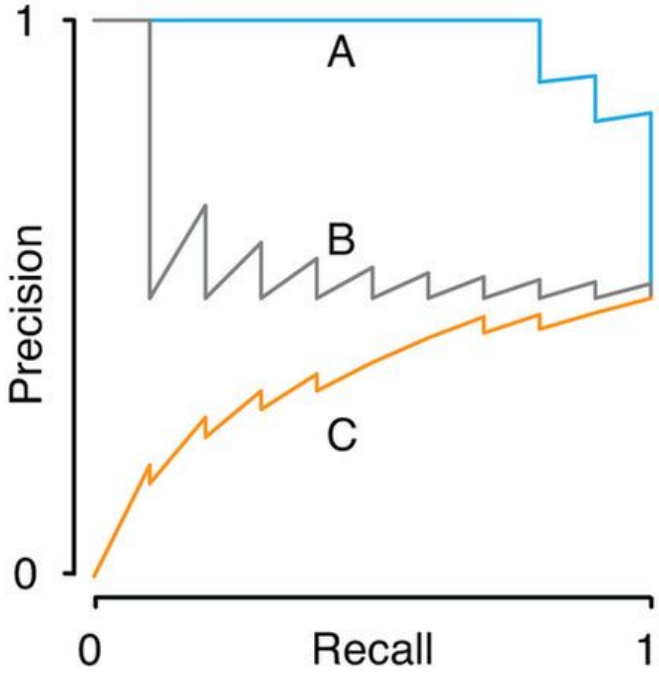
Lever, Krzywinski, & Altman (2016) Nat. Meth.

Evaluating a clustering with external/prior knowledge

ROC curve



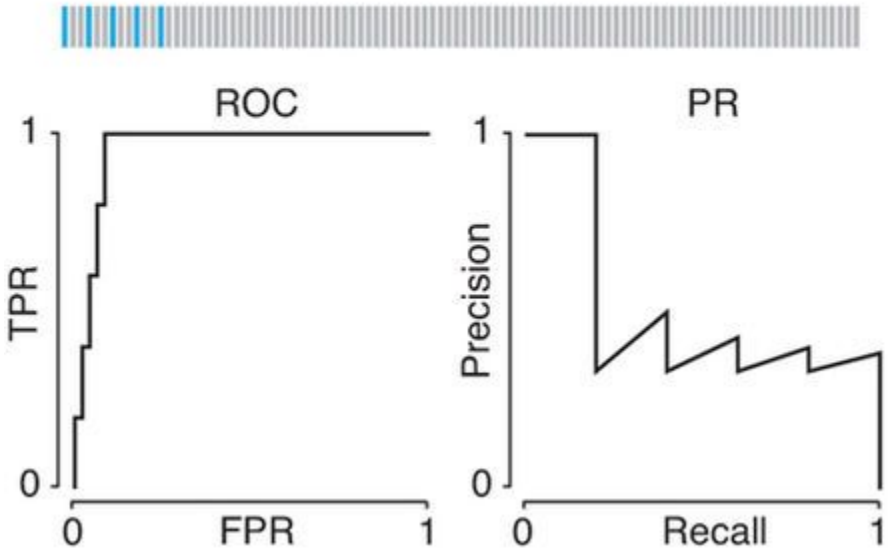
Precision recall curve



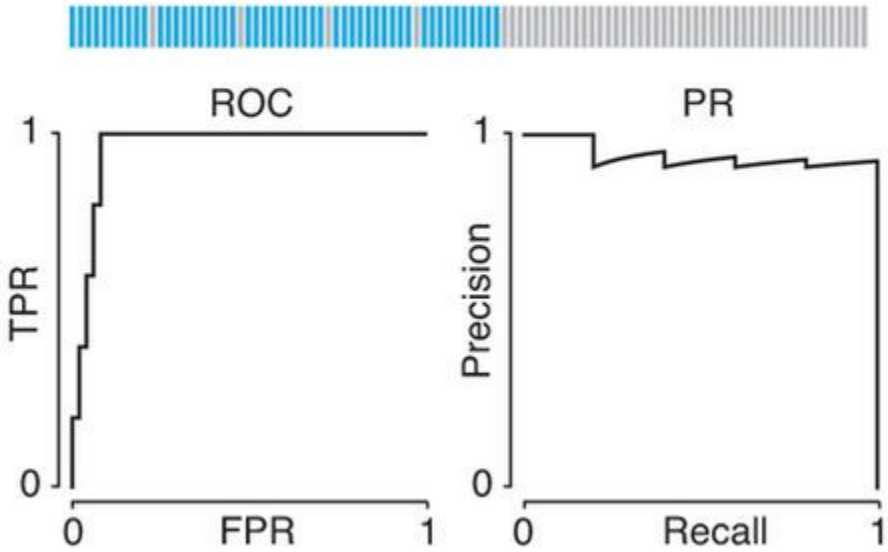
Lever, Krzywinski, & Altman (2016) Nat. Meth.

Evaluating a clustering with external/prior knowledge

5% positive



50% positive



Lever, Krzywinski, & Altman (2016) Nat. Meth.