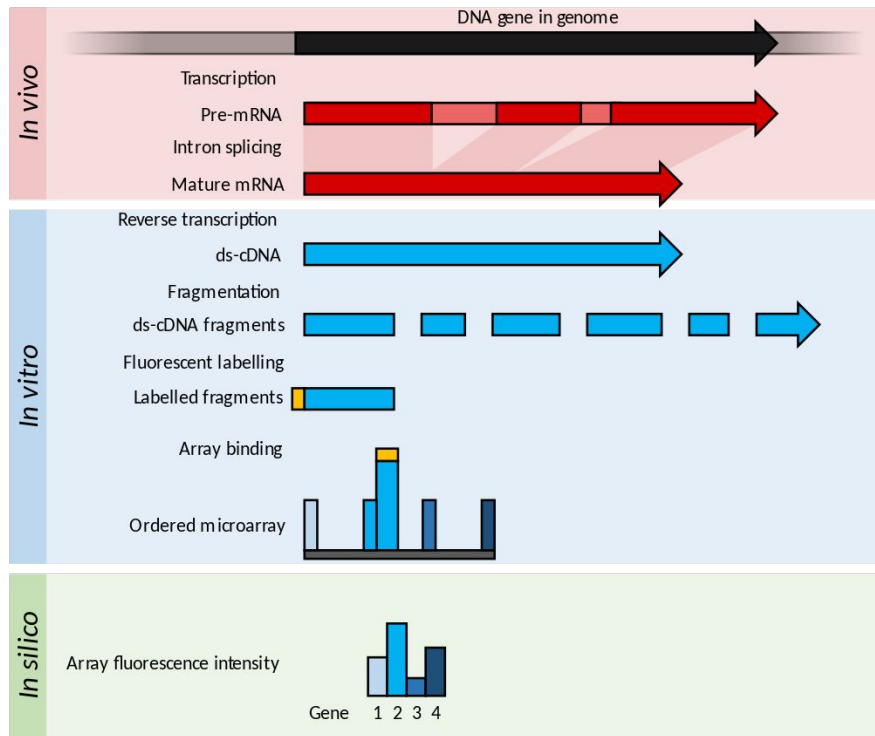


# Lecture 8: Functional genomics

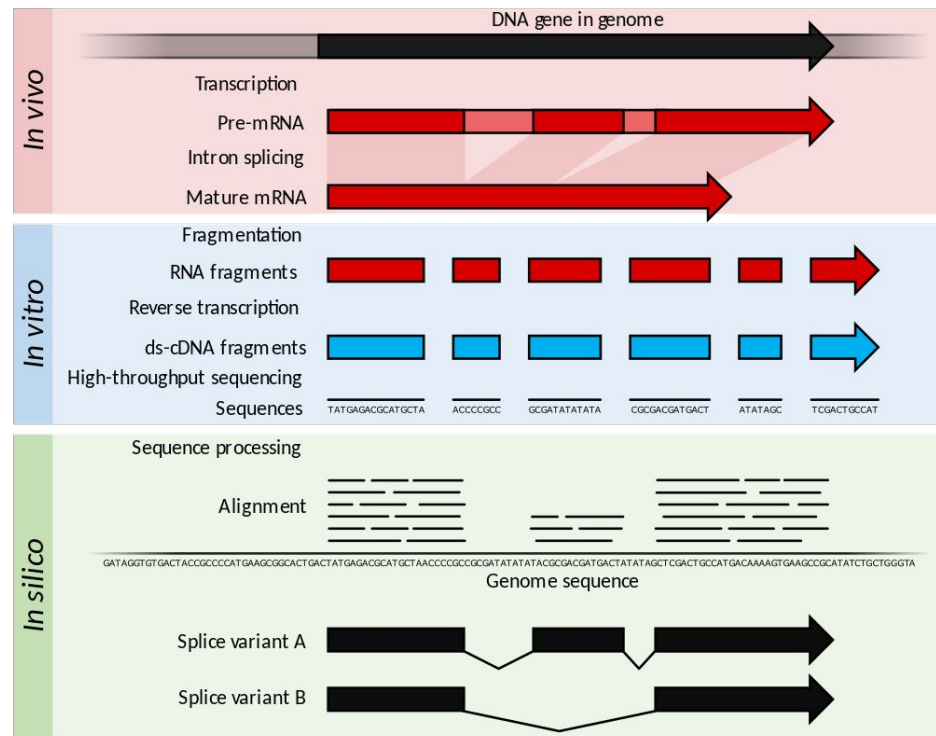
- Measuring gene-expression
  - Microarrays & RNA-seq
- Distance measures
- Clustering & Dimension reduction
- Classification

# Measuring gene-expression on a large-scale

## DNA microarrays

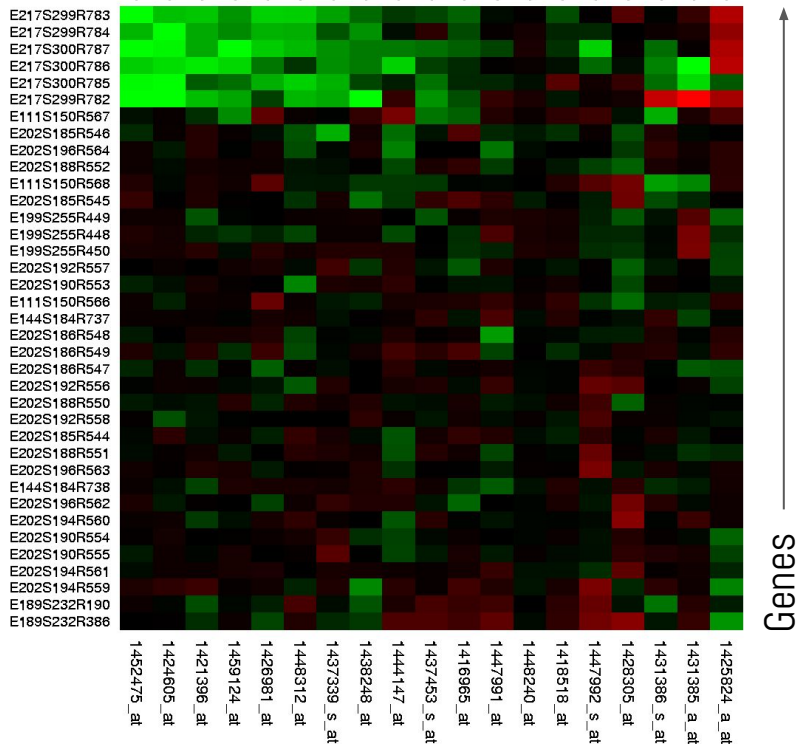


## RNA-seq



# Measuring gene-expression on a large-scale

Biological Samples →



Gene-level Qs:

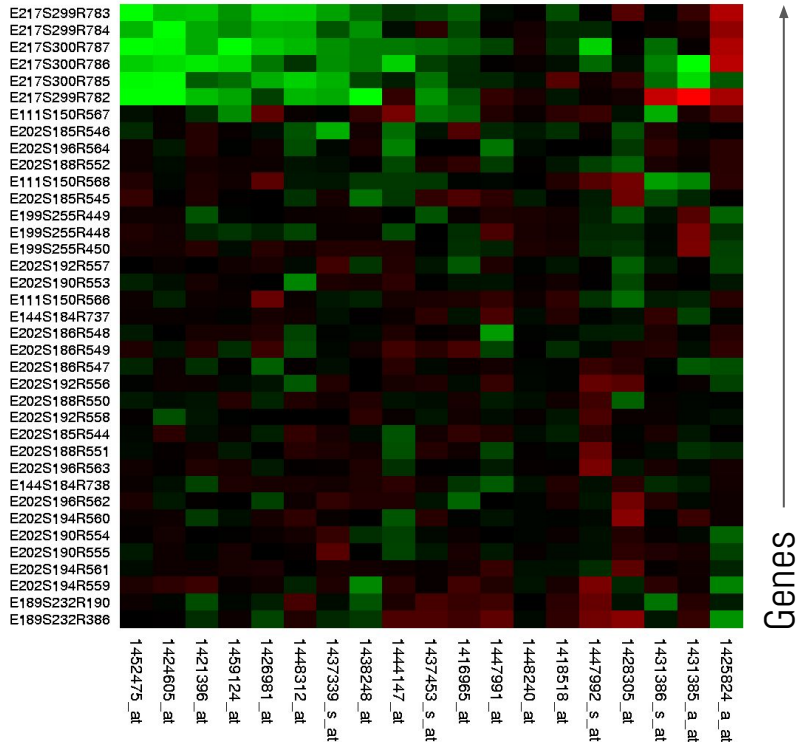
1. What's expressed (& by how much) in a given context/condition?
2. What's differentially expressed between two (or more) contexts/conditions?

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

Biological 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

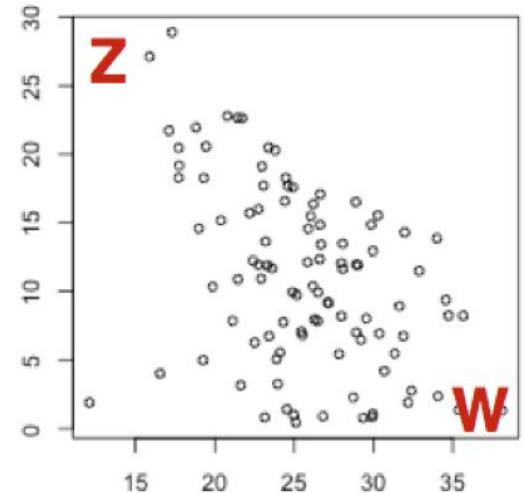
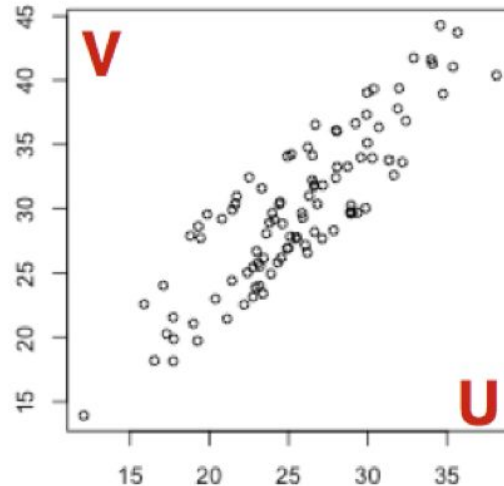
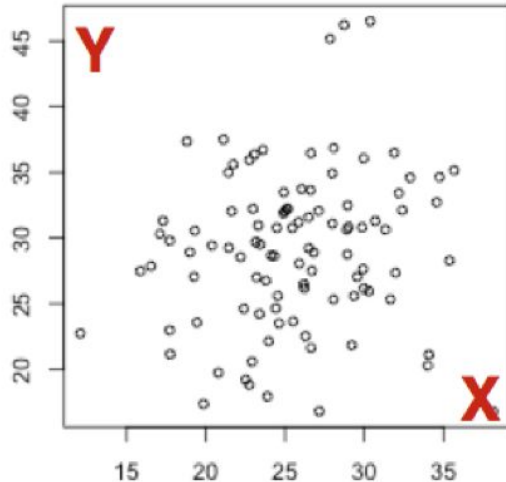
# Calculating “distance” between genes or samples

Variables

Attributes / Features



<b>x</b>	10	8	13	9	11	14	6	4	12	7	5
<b>y</b>	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  $\bar{y}$

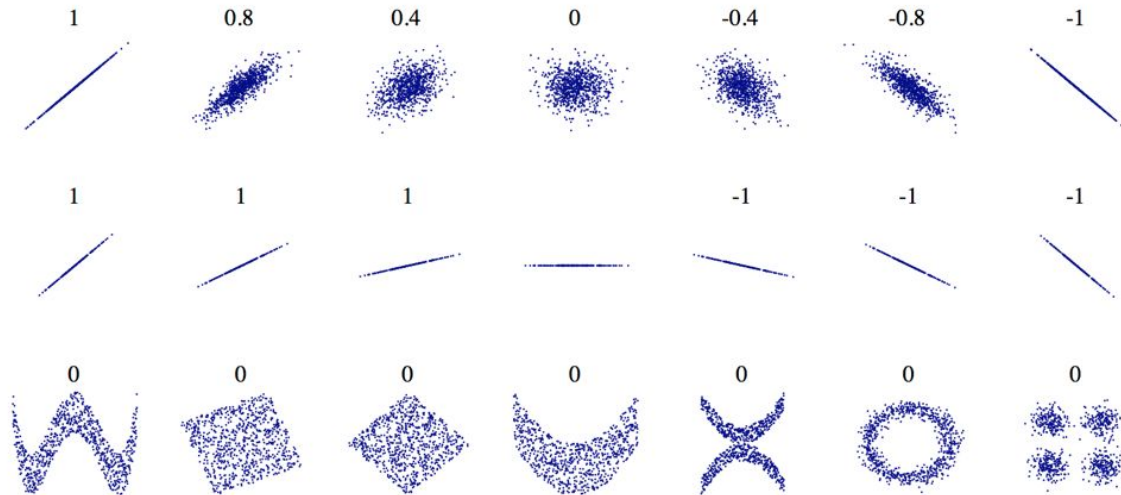
$$r = \frac{1}{n-1} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right)$$

# Distance measures

## Pearson Correlation Coefficient

- Measures 'linear' relationship between variables.

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



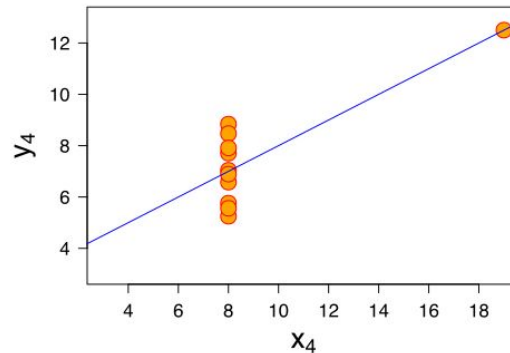
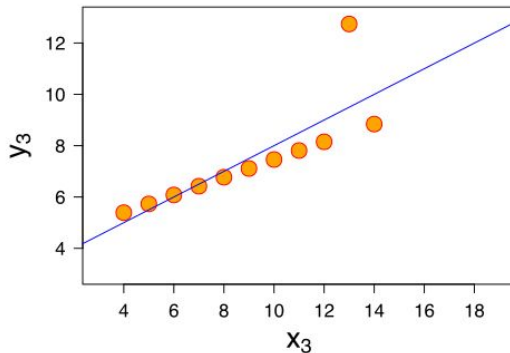
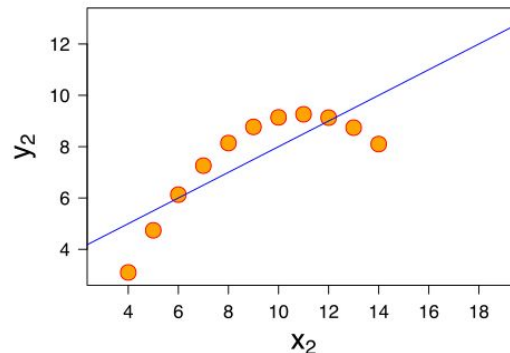
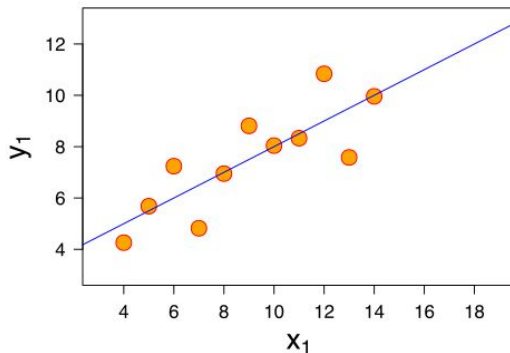
$$-1 \leq r \leq +1$$

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

# 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:
  - $y = 3.00 + 0.500x$

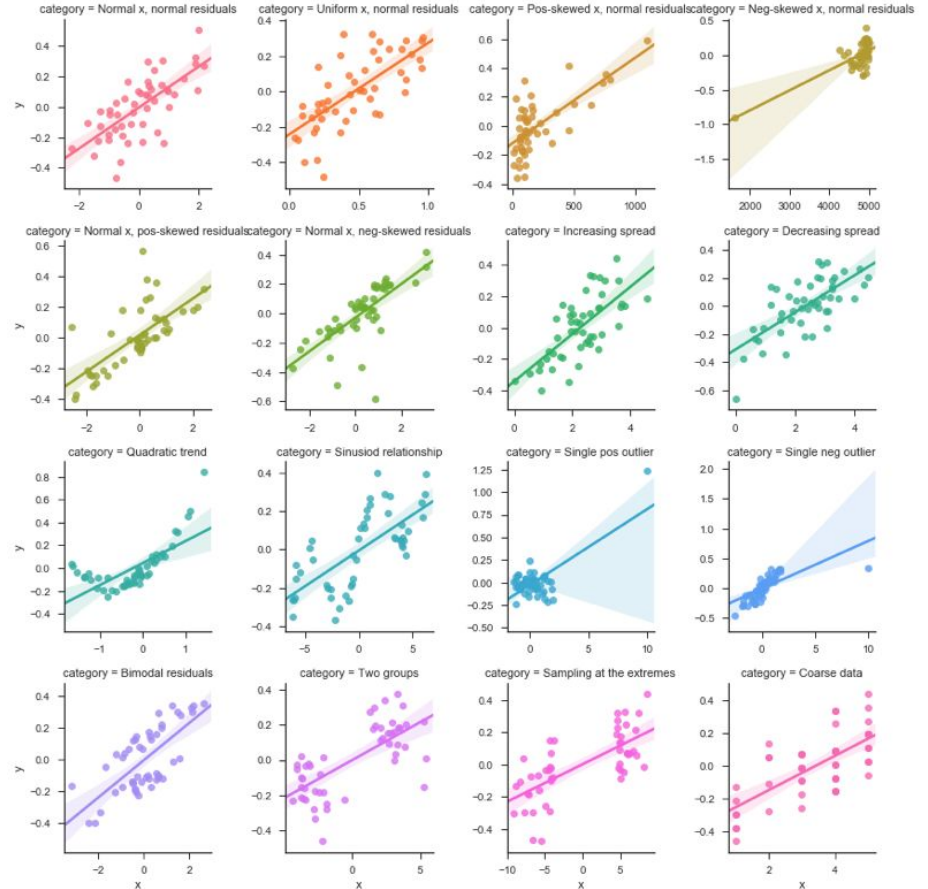
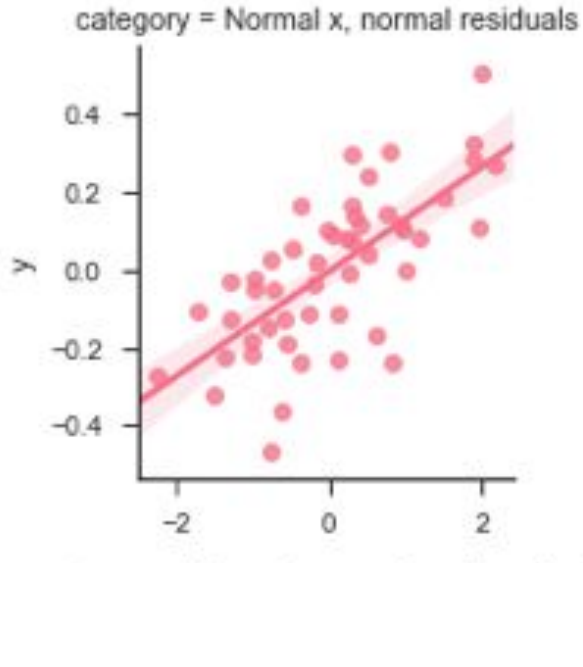


Anscombe, F. J. (1973). "Graphs in Statistical Analysis". American Statistician 27 (1): 17–21.



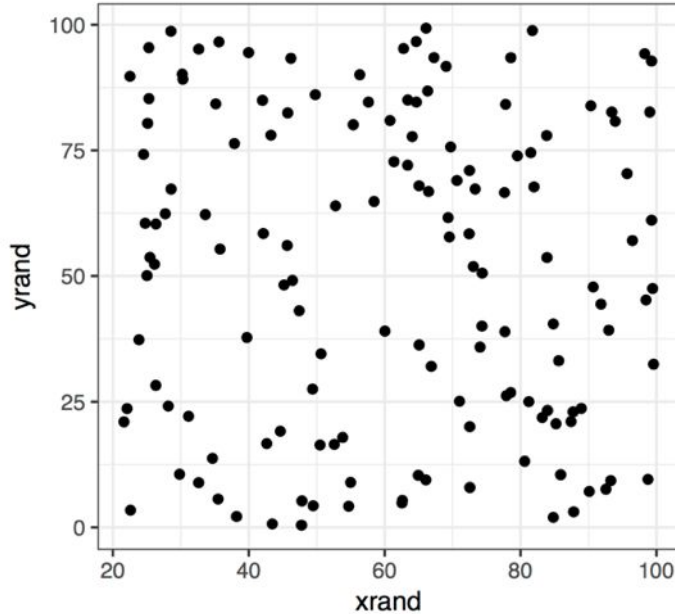
# What does a correlation coefficient tell you about the data?

Correlation = 0.7



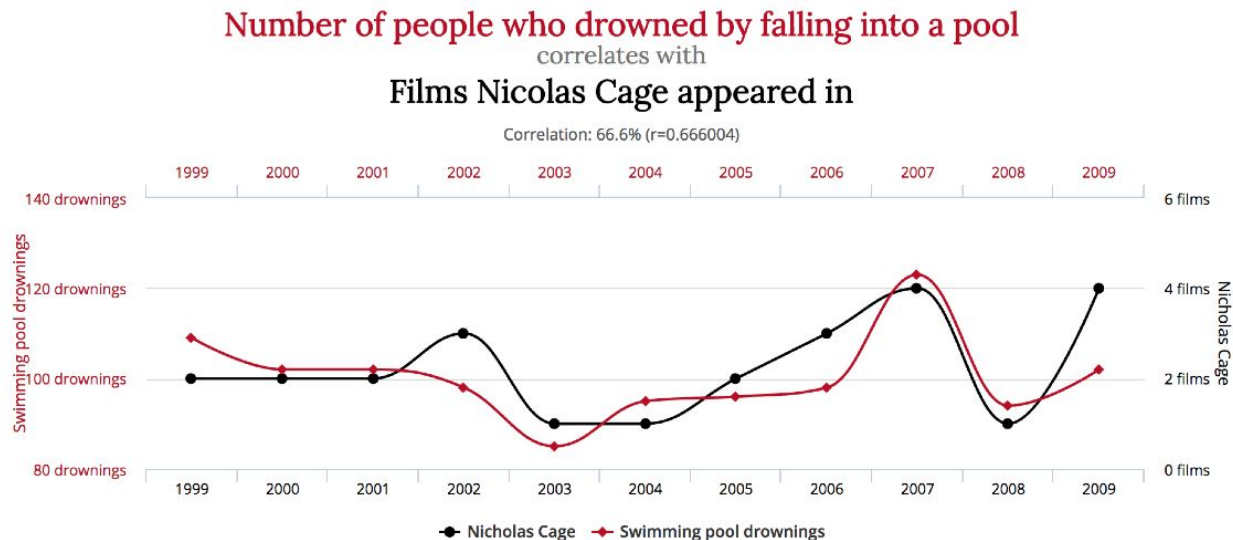
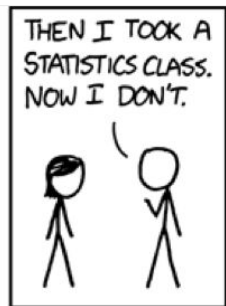
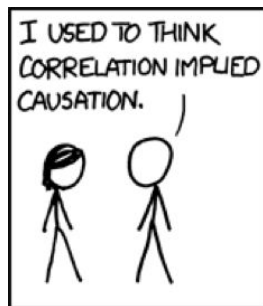
# What does a correlation coefficient tell you about the data?

Correlation = -0.06



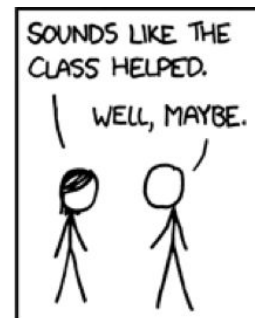
# Spurious correlations

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



Data sources: Centers for Disease Control & Prevention and Internet Movie Database

tylervigen.com



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

# Many distance measures

Pearson Correlation Coefficient

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

Spearman Rank Correlation

Euclidean Distance

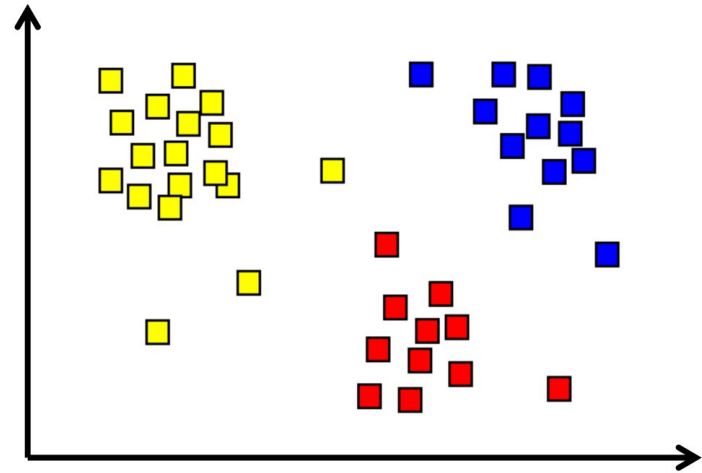
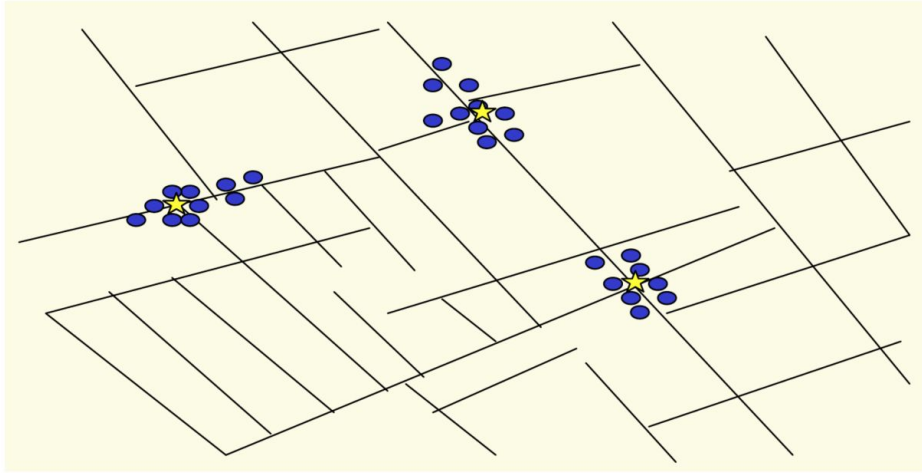
$$r = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{\sigma_x} \right) \left( \frac{y_i - \bar{y}}{\sigma_y} \right)$$

Mutual Information

...

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

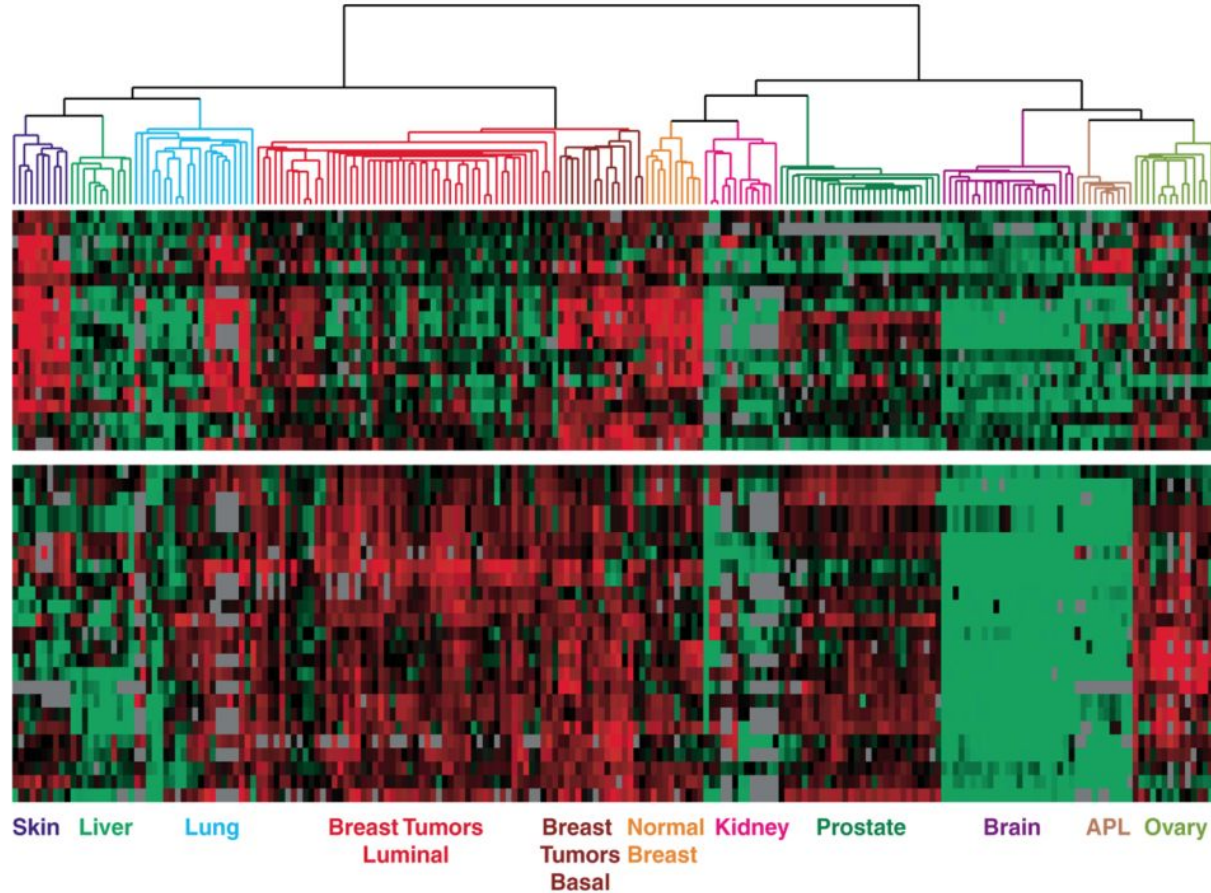
# Clustering



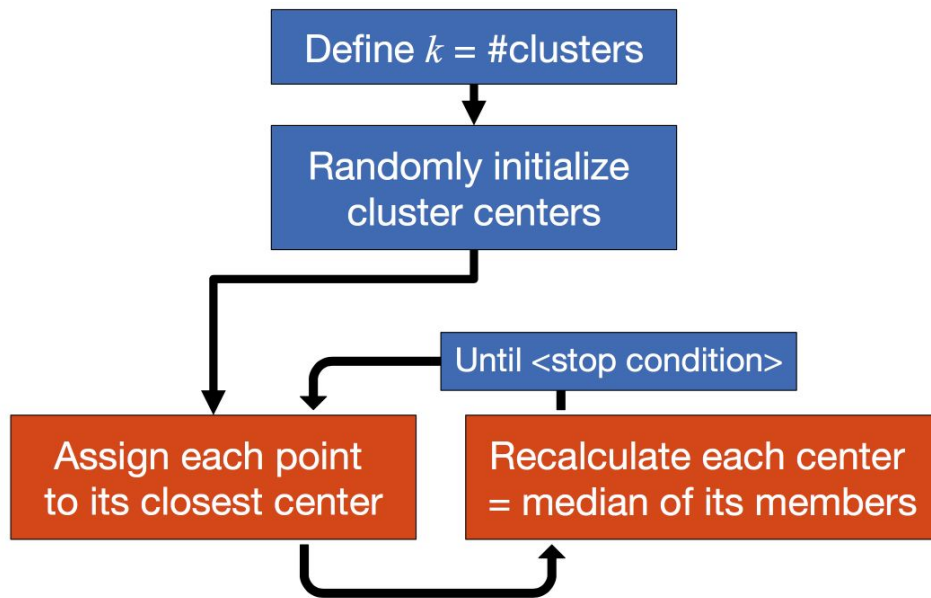
Group-level Qs:

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



# K-means clustering

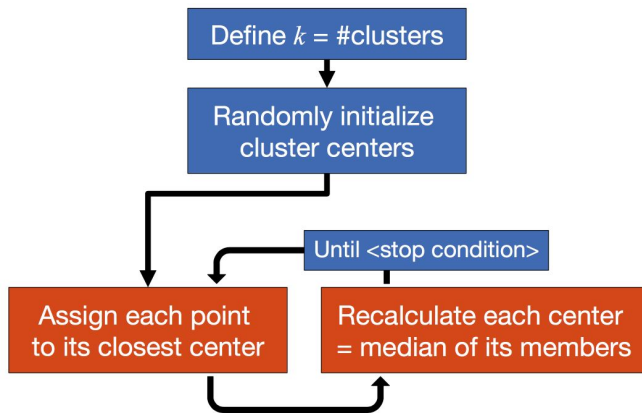


Conceptually similar to Expectation-Maximization, alternating between 2 two steps:

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.

# K-means clustering



## Stopping condition

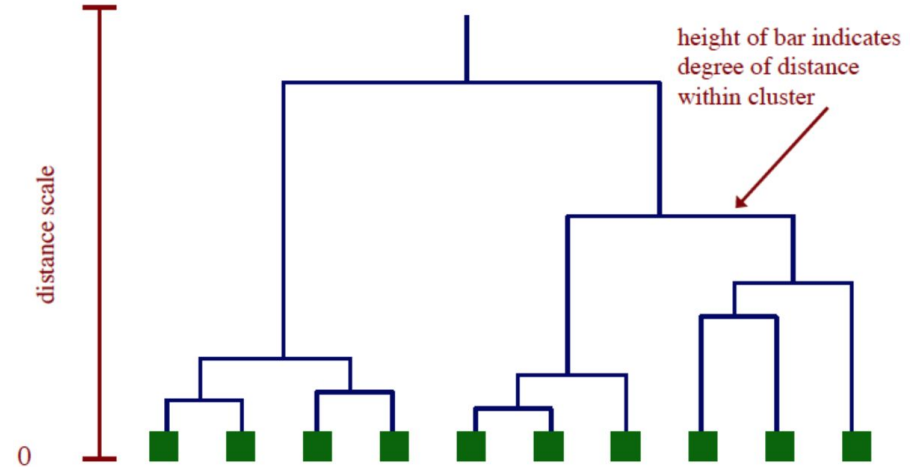
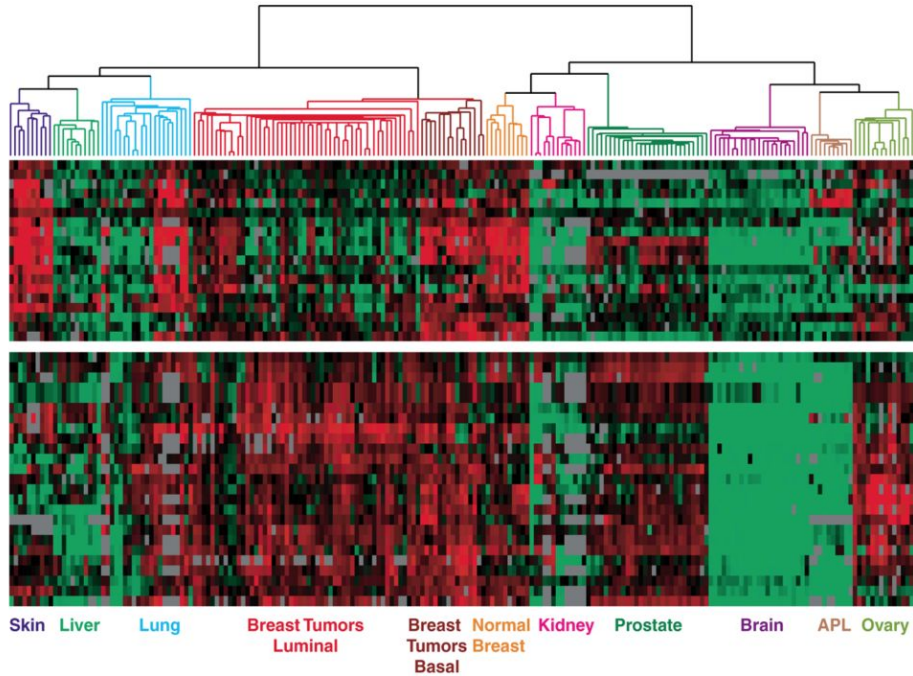
- Until the change in centers is less than <constant>.
- Until all genes get assigned to the same partition twice in a row.
- Until some minimal number of genes (e.g. 90%) get assigned to the same partition twice in a row.

## 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.

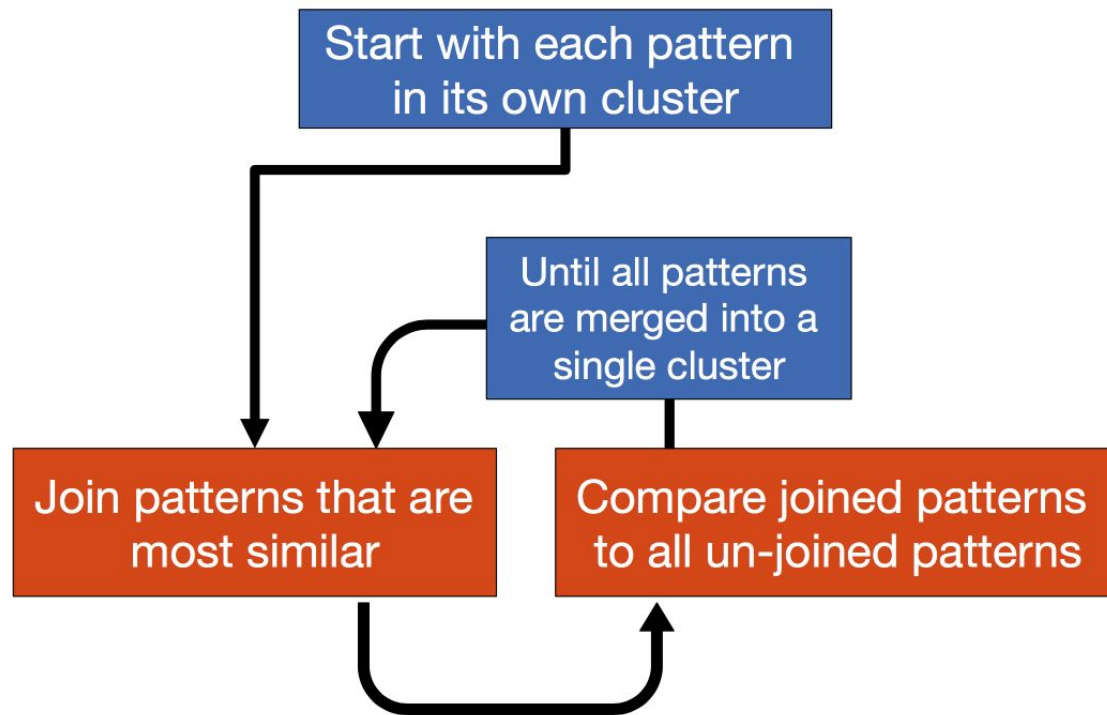


# Hierarchical clustering



- 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

