♠ Clustering

- Clustering is applied to multivariate data
  - Gene $i, i = 1, \ldots, p$
  - Expression level $x_{i,j}$ on array j
- The data matrix

$$\mathbf{X} = [x_{ij}] = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{p,1} & x_{p,2} & \cdots & x_{p,n} \end{pmatrix}$$

- Columns are snapshots of gene expression
- Normalization of rows or columns (to make them comparable)
- Statistics used for similarity – comparing "distances" between genes
  - Euclidean distance between gene i and gene k

$$d(i,k) = \sqrt{\sum_{j=1}^{n} (x_{i,j} - x_{k,j})^2}$$

  - Correlation $r_{i,k}$ between two genes
- Similar ideas to compare samples/microarrays

♠ K-Means clustering

- Partitions the data into unrelated clusters
- Shuffles observations from cluster to cluster to improve similarity within clusters
- Fast and makes efficient use of computer memory
- The final clustering depends on the initial partition
- Number of clusters remains constant and must be specified

♠ Hierarchical clustering

- Conceptually, hierarchical clustering recursively partitions the data into a tree like structure
- As usually implemented, clusters are agglomerated pairwise
- Hierarchical agglomeration
  Single linkage
  Average linkage
  Complete linkage
- Number of clusters can be assessed retrospectively

♠ You can make your own clustering algorithm

- Example: consolidation of hierarchical clusters
  Perform hierarchical clustering using average linkage
  Run the k-means algorithm using hierarchical clusters as the initial state

♠ Clustering in general

- The higher the dimension of the data the more sparse it is
- Well defined clusters may not and probably do not exist in the data
- There is no end to variations on the basic K-means and hierarchical clustering algorithms
- Clustering is a technique well suited to data exploration
- Clusters summarise the data
- No sensible clustering method is right or wrong or better than any other method. The appropriate methods to use are data dependent.
- Try more than one clustering method on your data!