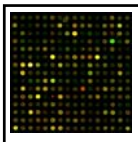
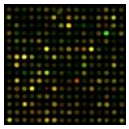
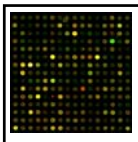


DNA Microarrays

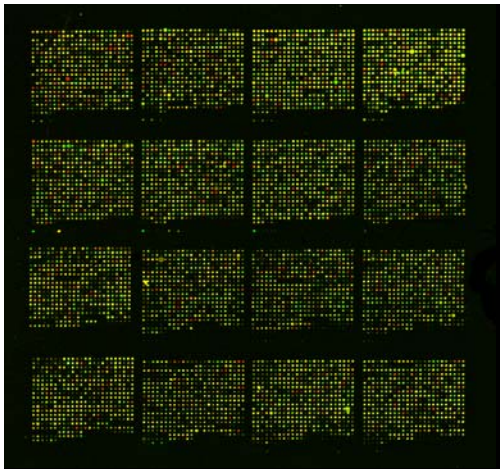


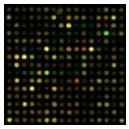
Microarrays: What are they good for?

Microarrays offer the ability to measure simultaneously the expression level of thousands of genes in a single experiment!

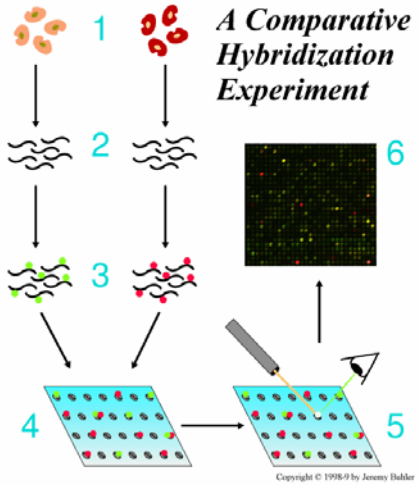


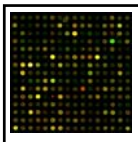
Yeast Genome Microarray



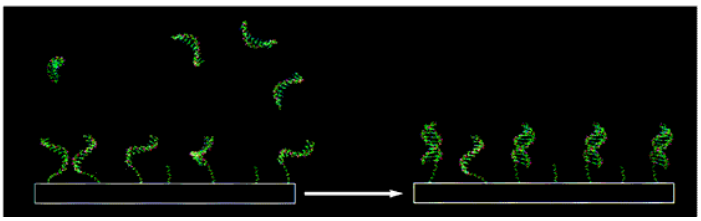


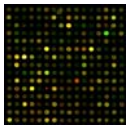
Comparative Hybridization





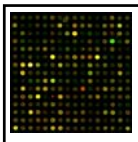
Individual Spot





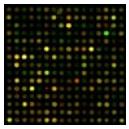
What are we "comparing"?

- Cell cycle variations
- Environmental response of cells
- Genetically heterogeneous diseases (cancers, heart disease, multiple sclerosis, diabetes, etc.)

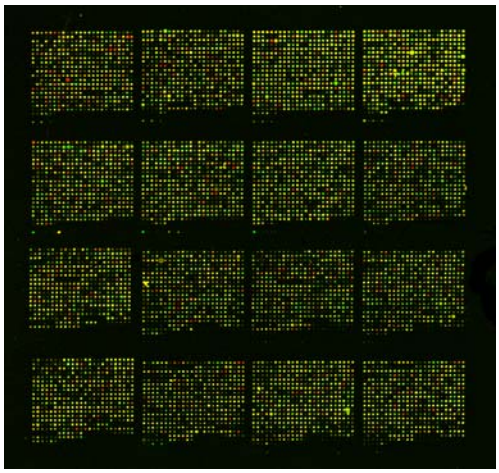


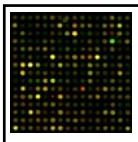
Microarray Limitations

- Gene expression may not be indicative of protein expression
- Error and variability in results
  - Not all mRNA is reverse transcribed to cDNA with the same efficiency
  - The number of flours which label each cDNA depends on its length and its sequence composition
  - Different cDNAs hybridize with different affinities
  - Quantifying array spot intensities is subject to noise



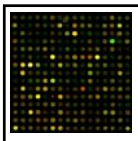
What are the results of a microarray experiment?





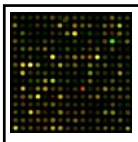
Data… and lots of it!

	Experiment 1	Experiment 2	Experiment 3	Experiment 4	…	Experiment *m*-1	Experiment *m*
Gene 1	0.6	4.4	1.3	1.0	…	3.1	2.2
Gene 2	1.5	2.6	5.2	0.8	…	2.8	2.9
Gene 3	0.7	3.7	2.4	1.9	…	1.5	1.6
Gene 4	0.3	0.7	0.2	1.3	…	4.9	3.0
Gene 5	3.1	3.0	2.1	1.4	…	4.2	0.9
…	…	…	…	…	…	…	…
Gene *n*-1	1.8	2.5	1.8	0.7	…	2.7	3.1
Gene *n*	0.5	3.4	3.0	0.5	…	1.8	2.5

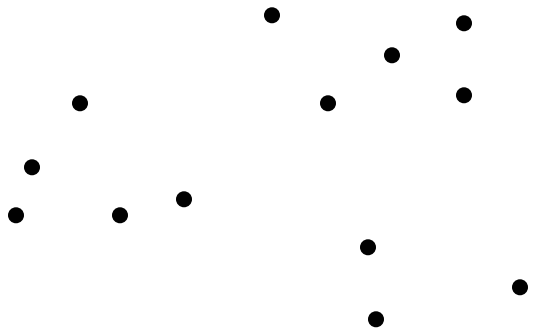


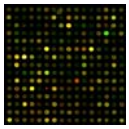
Finding Similarly Expressed Genes

- It may be useful to partition the *n* genes into groups of similarly expressed genes
- Clustering is the art of finding groups of genes, such that genes in the same group are as similar to each other as possible and as dissimilar to genes in other groups as possible



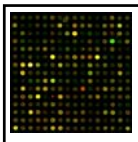
Clustering Example





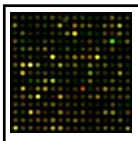
Example with 2 Experiments

	Experiment 1	Experiment 2
Gene 1	0.6	4.4
Gene 2	1.5	2.6
Gene 3	0.7	3.7
Gene 4	0.3	0.7
Gene 5	3.1	3.0
…	…	…
Gene *n*-1	1.8	2.5
Gene *n*	0.5	3.4

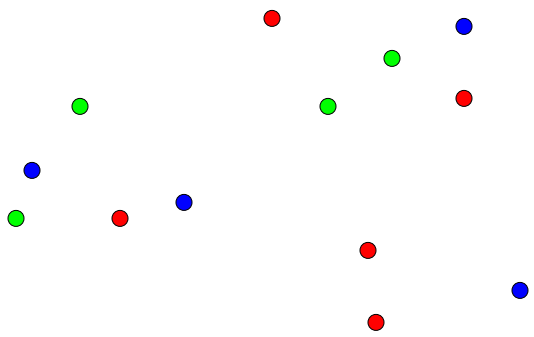


k-means Clustering Algorithm

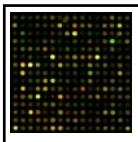
- Given a set of data points (genes) as input
- Randomly assign each point (gene) to one of the *k* clusters
- Repeat until convergence
  - Calculate center of each of the *k* clusters
  - Assign each point (gene) to the cluster with the closest center



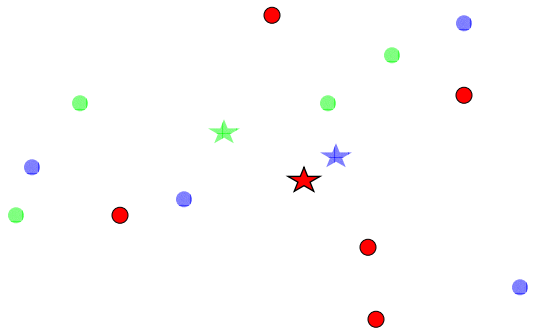
k-means Clustering Example



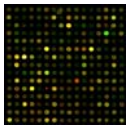
Randomly assign each point to one of the clusters



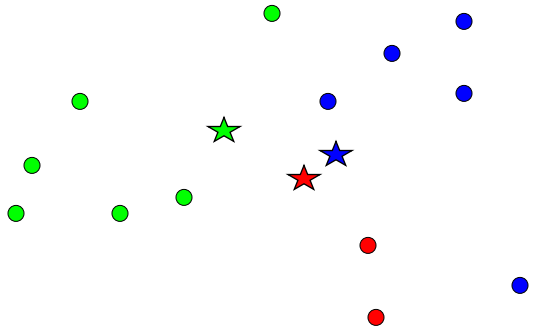
k-means Clustering Example



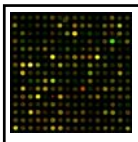
Calculate center of each cluster



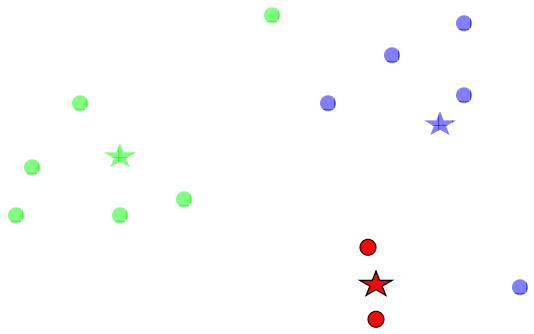
k-means Clustering Example



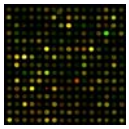
Assign each point to closest cluster center



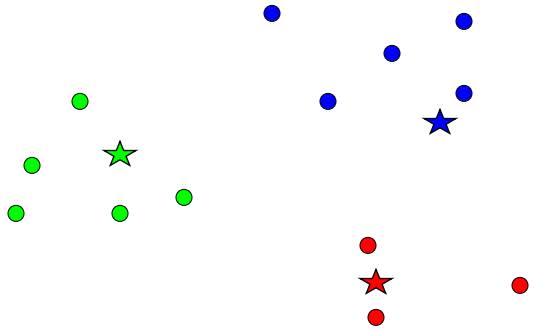
k-means Clustering Example



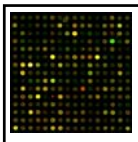
Calculate center of each cluster



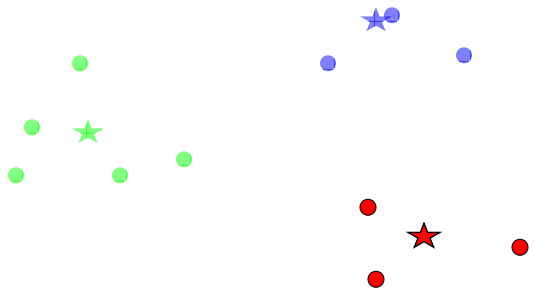
k-means Clustering Example



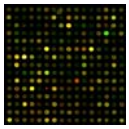
Assign each point to closest cluster center



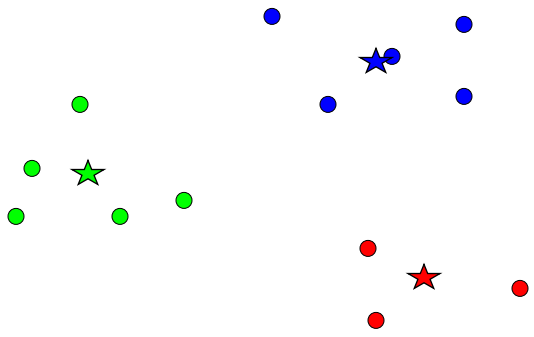
k-means Clustering Example



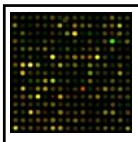
Calculate center of each cluster



k-means Clustering Example

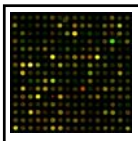


Convergence

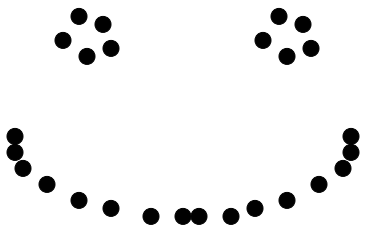


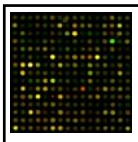
Clustering Problem

- *Clustering Problem*: Partition *n* data points into *k* clusters such that the total distance from each point to its cluster center is minimized.
- Clustering is an NP-complete problem



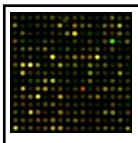
Does *k*-means always work?



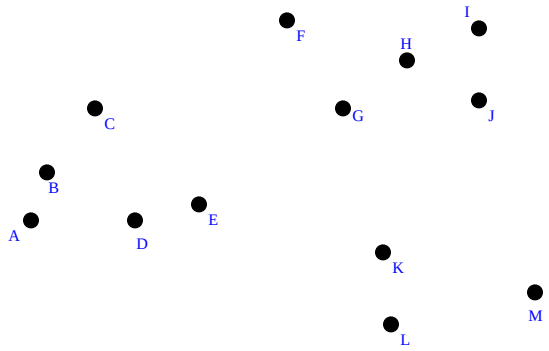


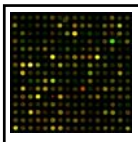
Hierarchical Clustering

- Assume each point is its own cluster
- Repeat the following step
  - Merge together the two closest clusters

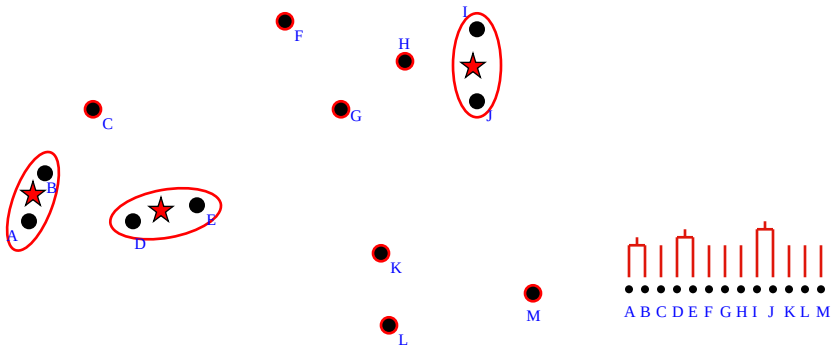


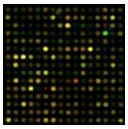
Hierarchical Clustering



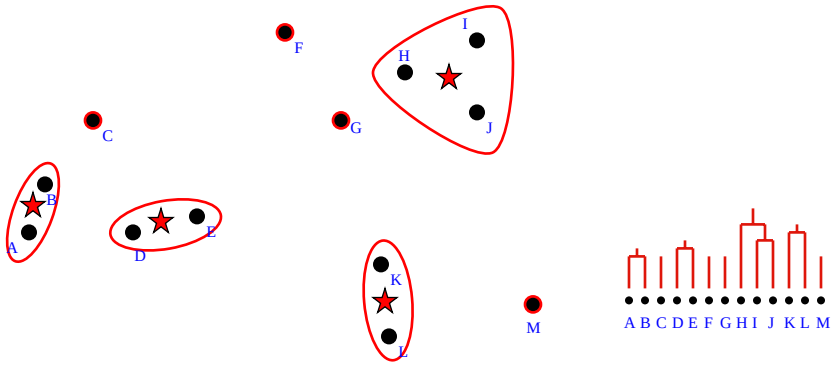


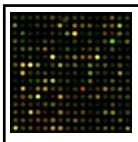
Hierarchical Clustering



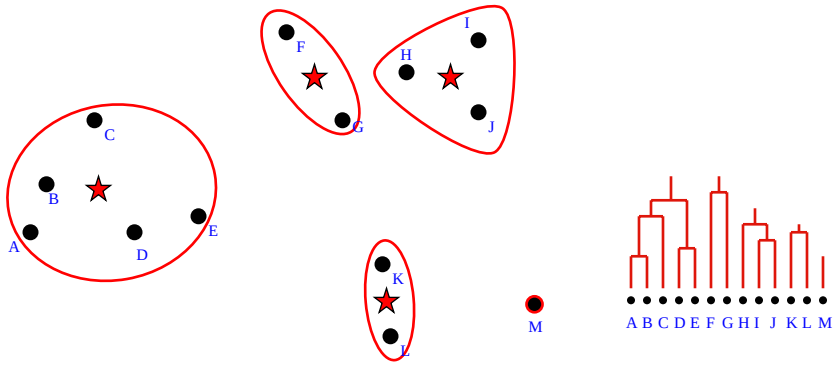


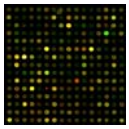
Hierarchical Clustering



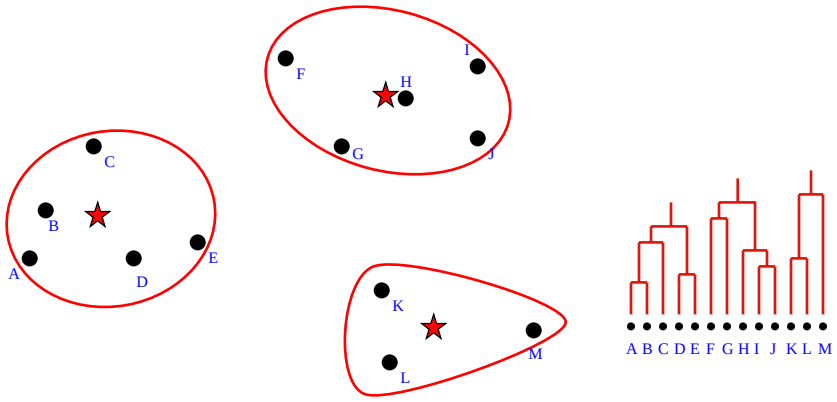


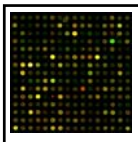
Hierarchical Clustering



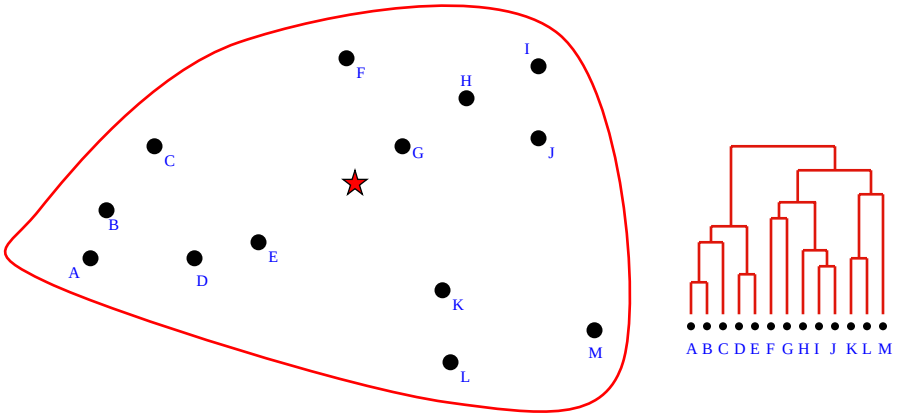


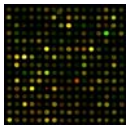
Hierarchical Clustering





Hierarchical Clustering





Clustering Application