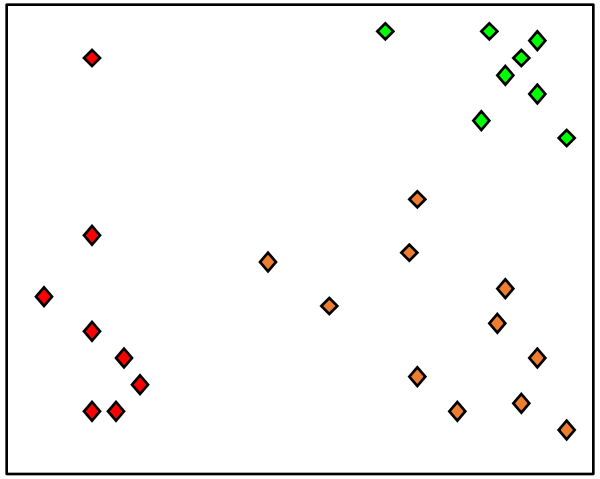
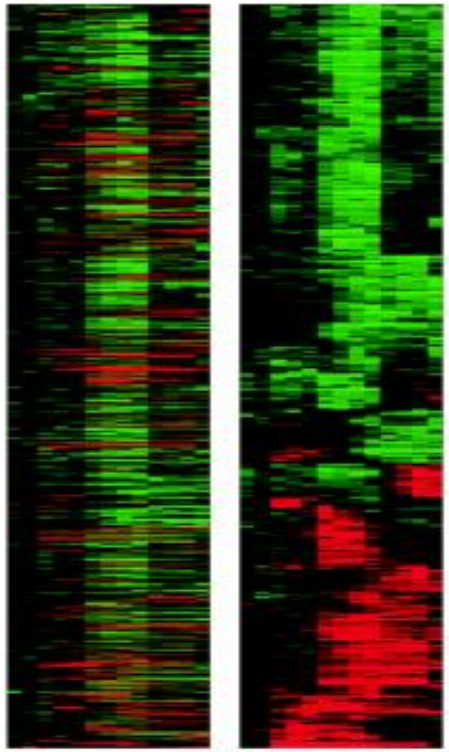
Clustering

What is clustering

- Organizing data into *clusters* such that there is
  - high intra-cluster similarity
  - low inter-cluster similarity
- Informally, finding natural groupings among objects.
- High dimensional data is often clustered for presentation
- Gene expression heatmap
  - Rows are genes
  - Columns are samples
  - Color is usually a Z-score (dimensionless)



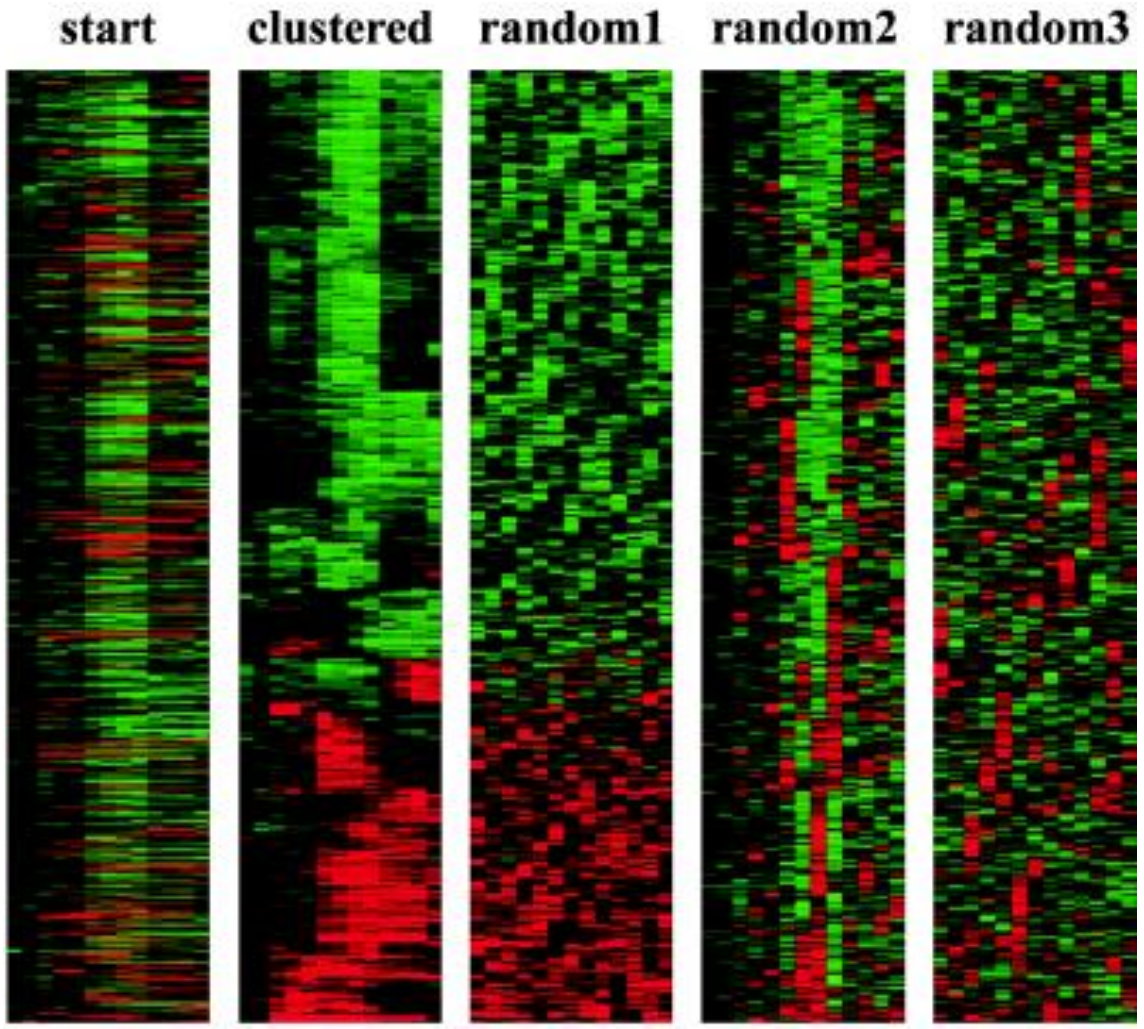


How significant is the clustering

Random 1 – randomized by rows.

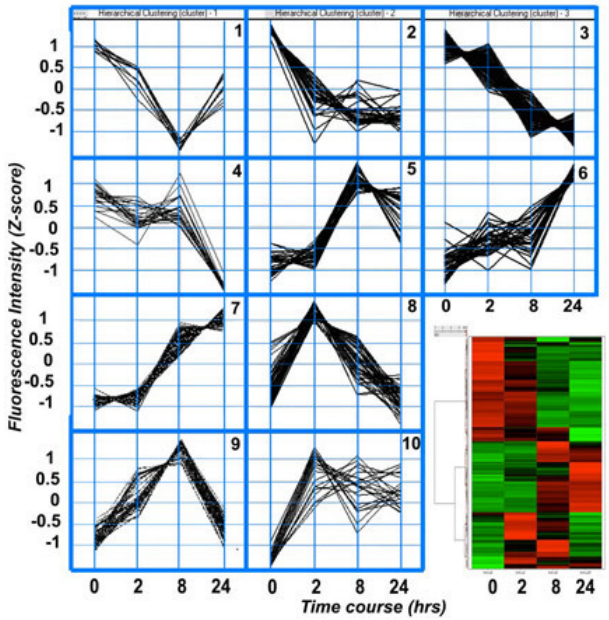
Random 2 – randomized by columns.

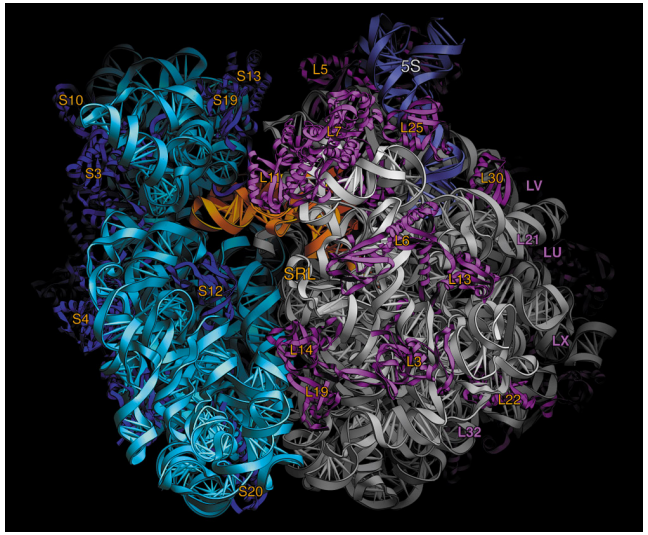
Random 3 – randomized by both rows and columns.



Gene clustering: biological interpretation

- Given a number of biologically distinct samples each gene has a specific patter of expression
- Gene expression is controlled by some upstream pathways that are shared across genes
- Examples: genes whose products are needed in stoichiometric quantities are highly correlated
  - Ribosome
  - Proteasome
  - ATP-synthase





Purpose of clustering

- Clustering across genes
  - Define functionally related gene classes
  - Can be used to infer functions of unknown genes
- Clustering across samples
  - Define groups of related samples
  - Example: infer the number of molecularly distinct cancer subtypes
- Both
  - Visualization
  - Dimensionality reduction:
    - Instead of saying genes X1,X2,X3… were up-regulated in samples Y1, Y2, Y3,…
    - Glycolysis genes were highly upregulated in the subtype A

Types of clustering

- Connectivity/bottom up/hierarchical clustering
- Centroid based: k-means
- Model based: mixture of Gaussians
- Dimensionality reduction techniques
  - PC, NMF

Hierarchical clustering

- Probably the most popular clustering algorithm in gene expression
- First presented in this context by Eisen in 1998

- Agglomerative **(**bottom-up)
- Algorithm:
  1. Initialize: each item a cluster
  2. Iterate:
     - select two most *similar* clusters
     - merge them
  3. Halt: when there is only one cluster left

. **Cluster analysis and display of genome-wide expression patterns.** *PNAS*

Hierarchical clustering

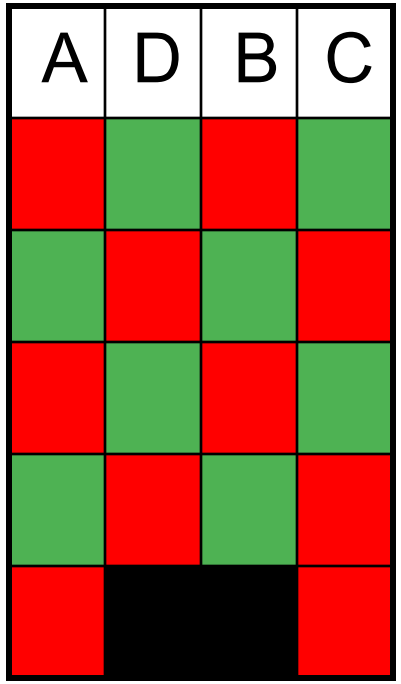
Calculate the Distance Matrix-euclidean distance

Gene	Chip1	Chip2
A	-2.0	1.0
B	-1.5	-0.5
C	1.0	0.25

→

	A	B	C
A	0.00	1.58	3.09
B	1.58	0.00	2.61
C	3.09	2.61	0.00

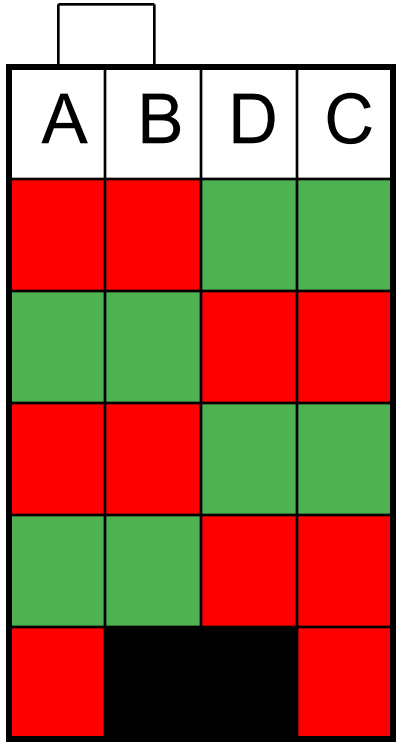
Hierarchical clustering



	A	**B**	**C**	**D**
A	0.00	1.58	3.09	4.74
B	1.58	0.00	2.61	5.00
C	3.09	2.61	0.00	2.70
D	4.74	5.00	2.70	0.00

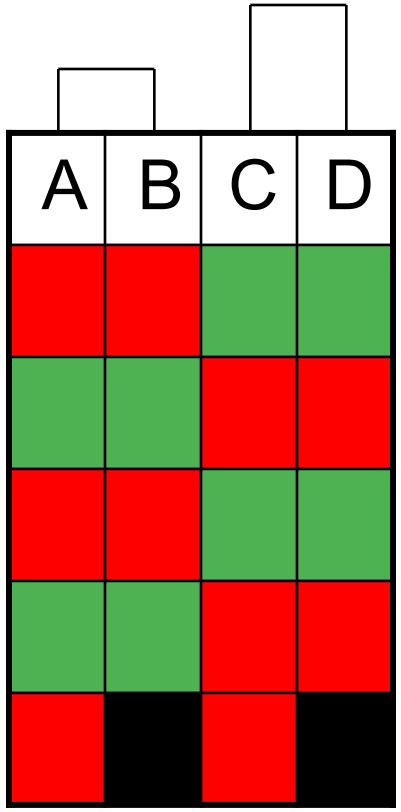
Hierarchical clustering

Average Linkage-update new distance with the average



	AB	**C**	**D**
AB	0.00	2.85	4.81
C	2.85	0.00	2.70
D	4.81	2.70	0.00

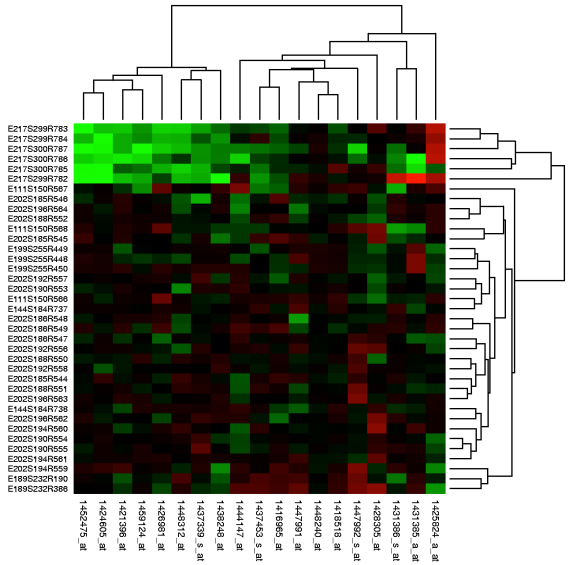
Hierarchical clustering



	AB	**CD**
AB	0.00	3.83
CD	3.83	0.00

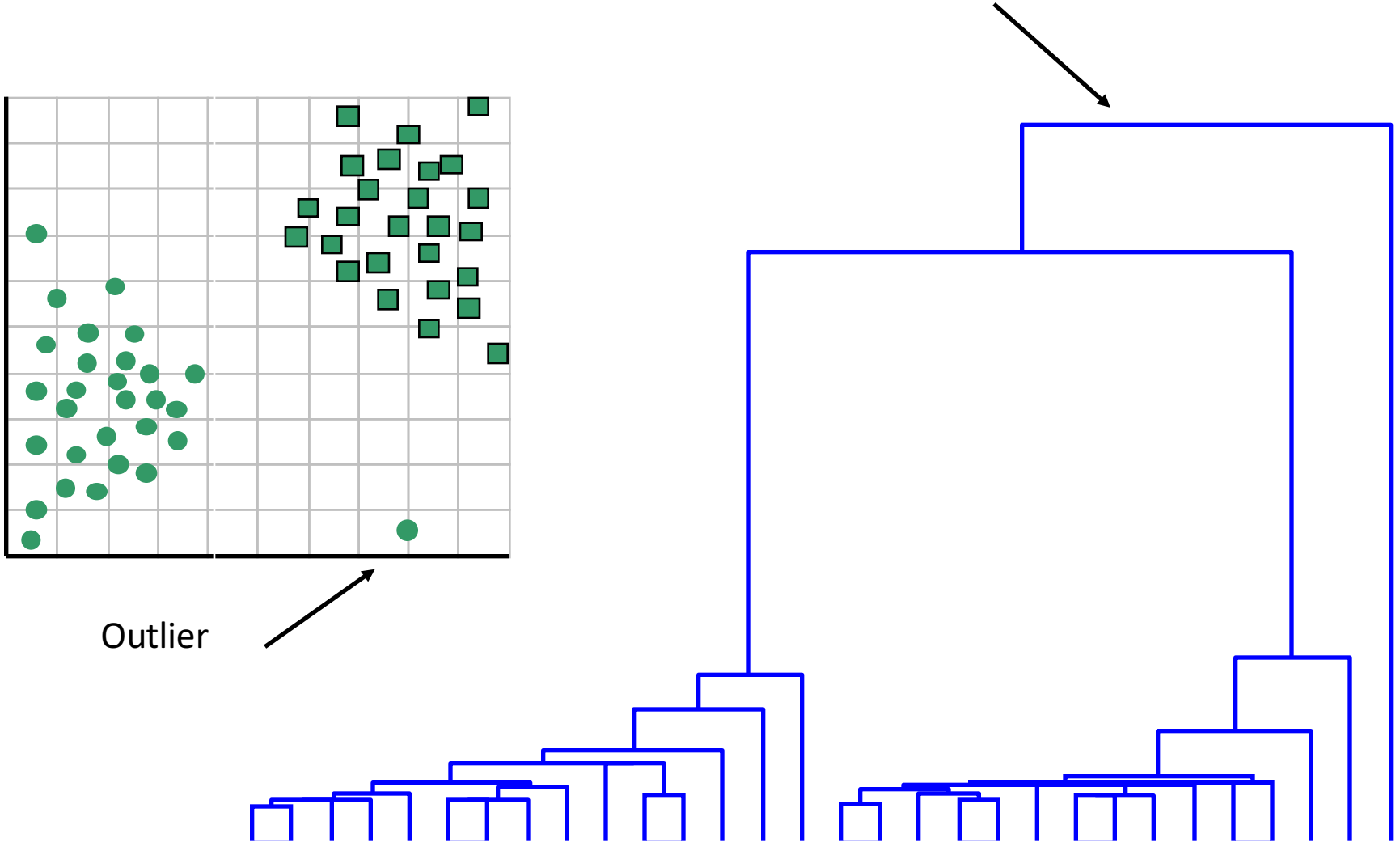
Details:

- Need a measure of similarity among two data vectors
  - Euclidian distance
  - 1-Correlation (raw and absolute value), rank(Spearman) correlation
- Also need a measure of similarity across clusters with multiple data vectors—how to we update our distance matrix when clusters are formed
  - Average linkage - midpoint.
  - Single linkage – smallest distance.
  - Complete linkage - largest distance.
- Very fast
- No need to specify number of clusters
- Important: final dendrogram is rotated arbitrarily: rotation freedom can be used to highlight or conceal different aspects of the data

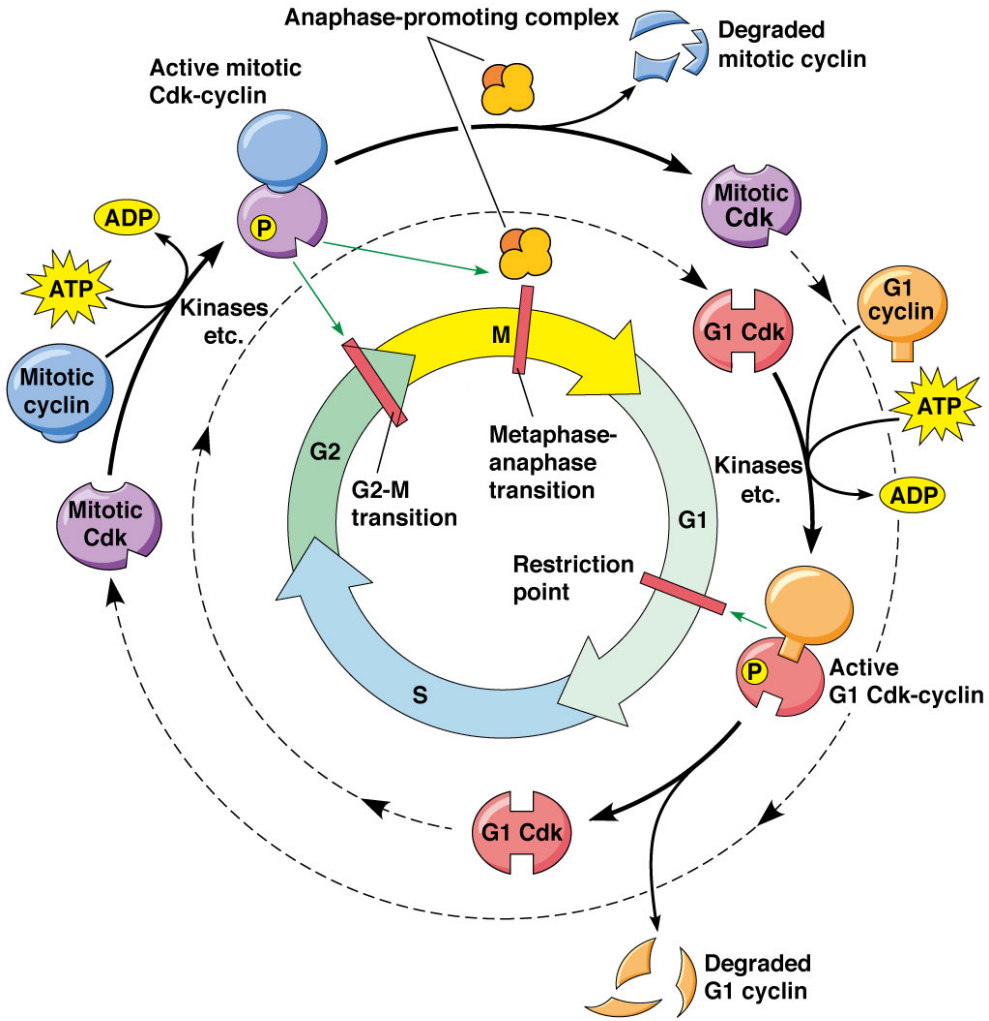


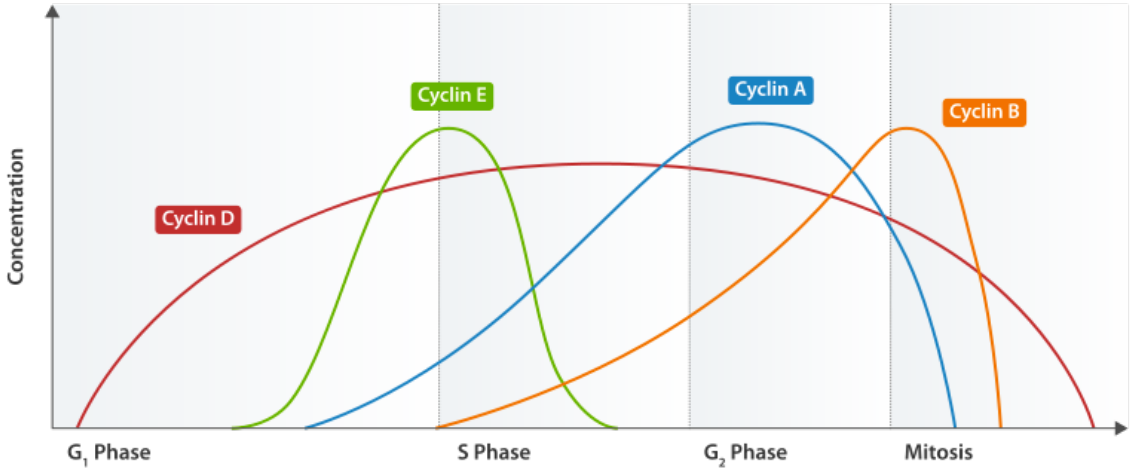
One potential use of a dendrogram is to detect outliers

The single isolated branch is suggestive of a data point that is very different to all others



Example I
Cell Cycle





- Cyclins are a family of proteins that control the progression of cells through the cell cycle by activating cyclin-dependent kinase (Cdk) enzymes
- Experimental design: use a microarray experiment to find out how genes are regulated with cell-cycle genome wide
- Spellman et al that was published in *Mol. Biol. Cell* 9, 3273-3297 (1998).

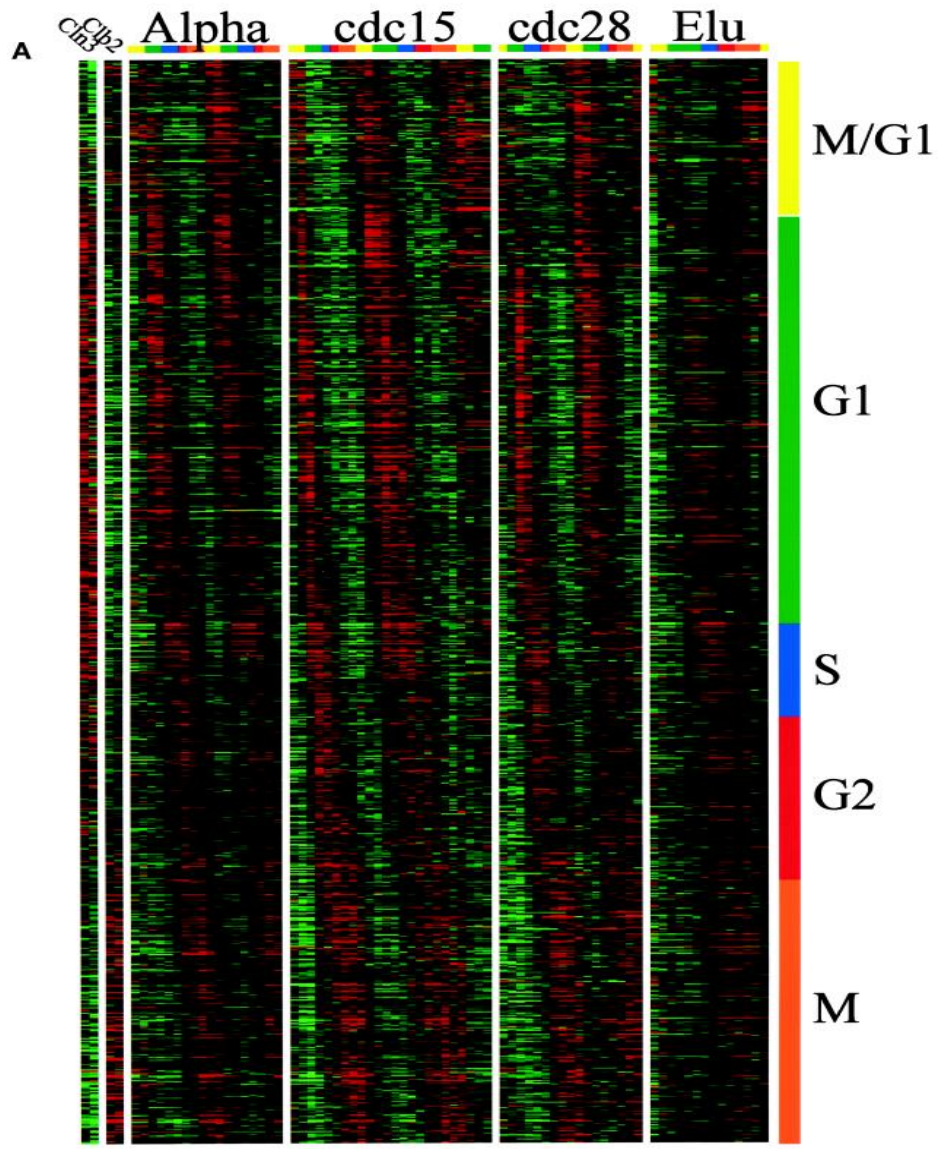
Example: cell -cycle

Methods

- DNA microarrays of the yeast genome
- Synchronization: cells must be in the same phase of the cell cycle
  - α factor.
  - Elutriation – size based.
  - Cdc15 – heat mutation.
- Manipulation-induction of cyclins
  - cln3p, clb2p deletion.
- Data from a previously published study (Cho et al. 1998)
- Control sample: asynchronous cultures.

Example: cell -cycle

Genes sorted by time of peak expression as calculated from the Fourier transform



Example: cell -cycle

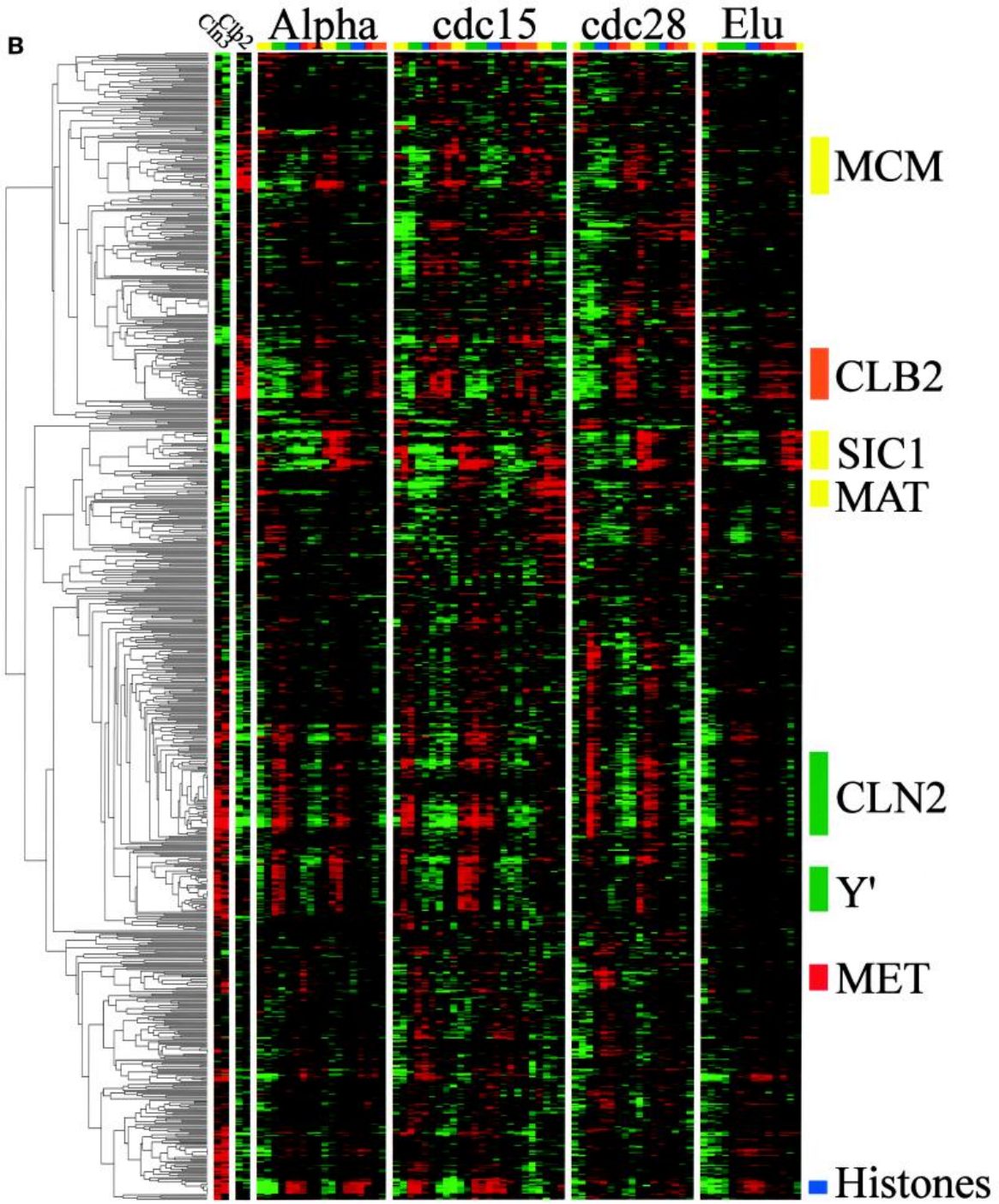
Clustering

Genes sorted by their hierarchical cluster relationships-- similarity of expression across the measurements:

Table 4.

Cluster summary

Cluster	No. of genes	Binding site	Regulator	Peak expression
CLN2	119	ACGCGT	MBF, SBF	G1
Y'	26	Unknown	Unknown	G1
FKS1	92	ACRMSAAA	SBF, (MBF?)	G1
Histone	10	ATGCGAAR	Unknown	S
MET	20	AAACTGTGG	Met31p, Met32p	S
CLB2	35	MCM1 + SFF	Mcm1p + SFF	M
MCM	34	MCM1	Mcm1p	M/G1
SIC1	27	RRCCAGCR	Swi5p/Ace2p	M/G1
Total	363			



Example: cell -cycle

- Cell-cycle timing was correlated with function for many genes with no obvious cell-cycle activity
- “The “MET” cluster was completely unexpected. It contains 10 genes involved in the biosynthesis of methionine.”

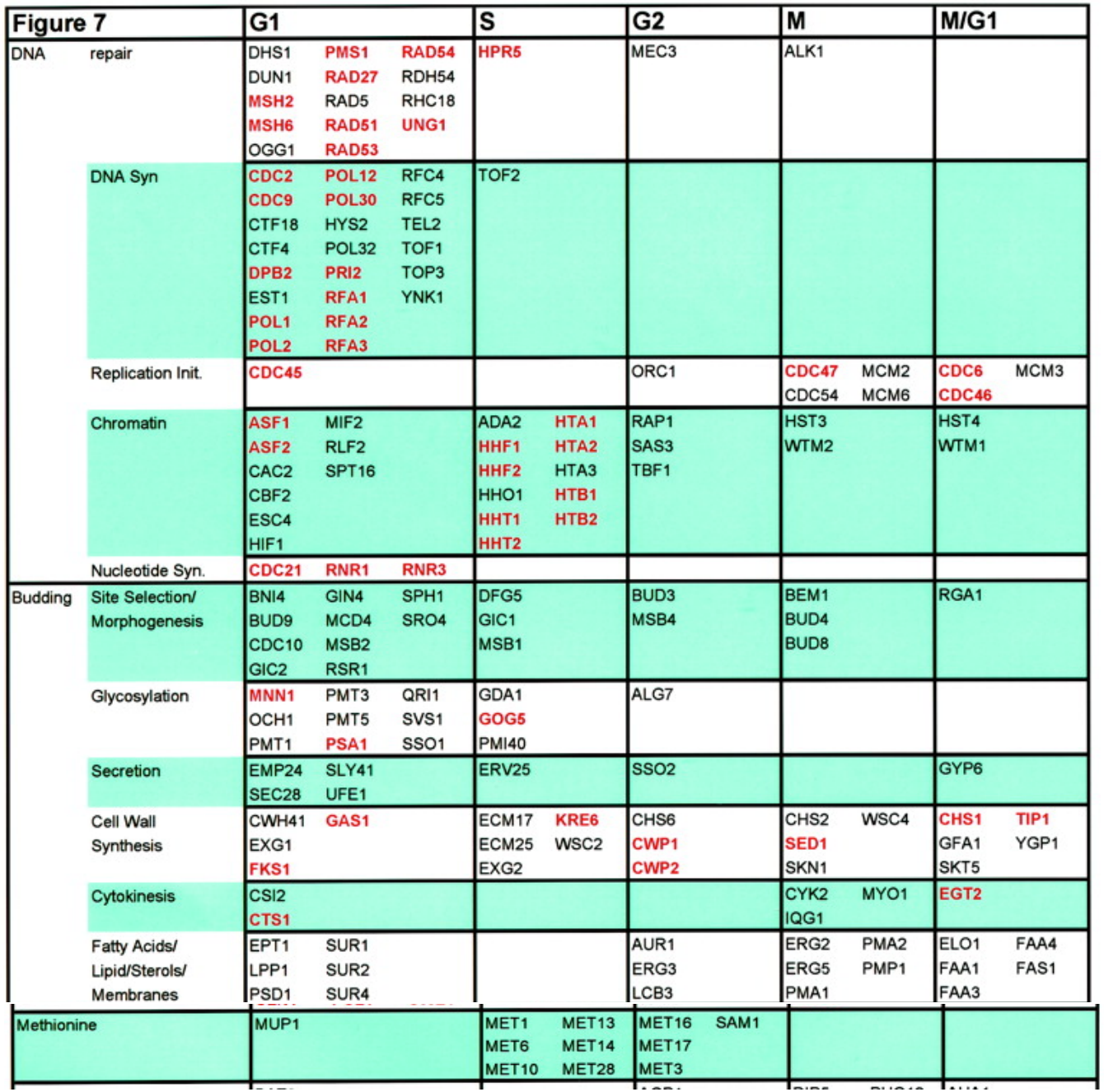


Figure 7		G1	S	G2	M	M/G1
DNA	repair	DHS1 PMS1 RAD54 DUN1 RAD27 RDH54 MSH2 RAD5 RHC18 MSH6 RAD51 UNG1 OGG1 RAD53	HPR5	MEC3	ALK1	
	DNA Syn	CDC2 POL12 RFC4 CDC9 POL30 RFC5 CTF18 HYS2 TEL2 CTF4 POL32 TOF1 DPB2 PRI2 TOP3 EST1 RFA1 YNK1 POL1 RFA2 POL2 RFA3	TOF2			
	Replication Init.	CDC45		ORC1	CDC47 MCM2 CDC54 MCM6	CDC6 MCM3 CDC46
	Chromatin	ASF1 MIF2 ASF2 RLF2 CAC2 SPT16 CBF2 ESC4 HIF1	ADA2 HTA1 HHF1 HTA2 HHF2 HTA3 HHO1 HTB1 HHT1 HTB2 HHT2	RAP1 SAS3 TBF1	HST3 WTM2	HST4 WTM1
	Nucleotide Syn.	CDC21 RNR1 RNR3				
Budding	Site Selection/ Morphogenesis	BNI4 GIN4 SPH1 BUD9 MCD4 SRO4 CDC10 MSB2 GIC2 RSR1	DFG5 GIC1 MSB1	BUD3 MSB4	BEM1 BUD4 BUD8	RGA1
	Glycosylation	MNN1 PMT3 QRI1 OCH1 PMT5 SVS1 PMT1 PSA1 SSO1	GDA1 GOG5 PMI40	ALG7		
	Secretion	EMP24 SLY41 SEC28 UFE1	ERV25	SSO2		GYP6
	Cell Wall Synthesis	CWH41 GAS1 EXG1 FKS1	ECM17 KRE6 ECM25 WSC2 EXG2	CHS6 CWP1 CWP2	CHS2 WSC4 SED1 SKN1	CHS1 TIP1 GFA1 YGP1 SKT5
	Cytokinesis	CSI2 CTS1			CYK2 MYO1 IQG1	EGT2
	Fatty Acids/ Lipid/Sterols/ Membranes	EPT1 SUR1 LPP1 SUR2 PSD1 SUR4		AUR1 ERG3 LCB3	ERG2 PMA2 ERG5 PMP1 PMA1	ELO1 FAA4 FAA1 FAS1 FAA3
Methionine		MUP1	MET1 MET13 MET6 MET14 MET10 MET28	MET16 SAM1 MET17 MET3		

Example: cell -cycle

- Genes in each cluster share DNA binding motifs
- Extract upstream promoter sequence for each genes
- Use Gibbs sampling (covered later) to identify overrepresented sequence for each cluster

- MCM1:T-T-A-C-C-N-A-A-T-T-N-G-G-T-A-A
- SFF: G-T-M-A-A-C-A-A
- New motif:

T-T-W-C-C-Y-A-A-W-N-N-G-G-W-A-A-W-W-N-R-T-A-A-A-Y-A-A

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27
T	23	23	17	3	2	13	2	6	16	13	7	0	1	7	3	1	11	11	8	3	25	3	2	0	15	0	4
A	5	5	11	0	0	5	25	26	18	8	17	2	0	21	32	33	21	22	7	8	6	24	33	34	0	34	26
C	4	4	0	32	33	14	2	1	1	7	3	0	0	4	0	1	0	0	7	4	2	8	0	0	20	0	1
G	3	3	7	0	1	3	6	2	0	7	8	33	34	3	0	0	3	2	5	20	2	0	0	1	0	1	4
	T	T	W	C	C	Y	A	A	W	N	N	G	G	W	A	A	W	W	N	R	T	A	A	A	Y	A	A

K-means clustering

- Partition clustering
- Nonhierarchical, each instance is placed in exactly one of K non-overlapping clusters.
- Since the output is only one set of clusters the user has to specify the desired number of clusters K.

K-means algorithm

1. Decide on a value for *K*, the number of clusters.

2. Initialize the *K* cluster centers (randomly, if necessary).

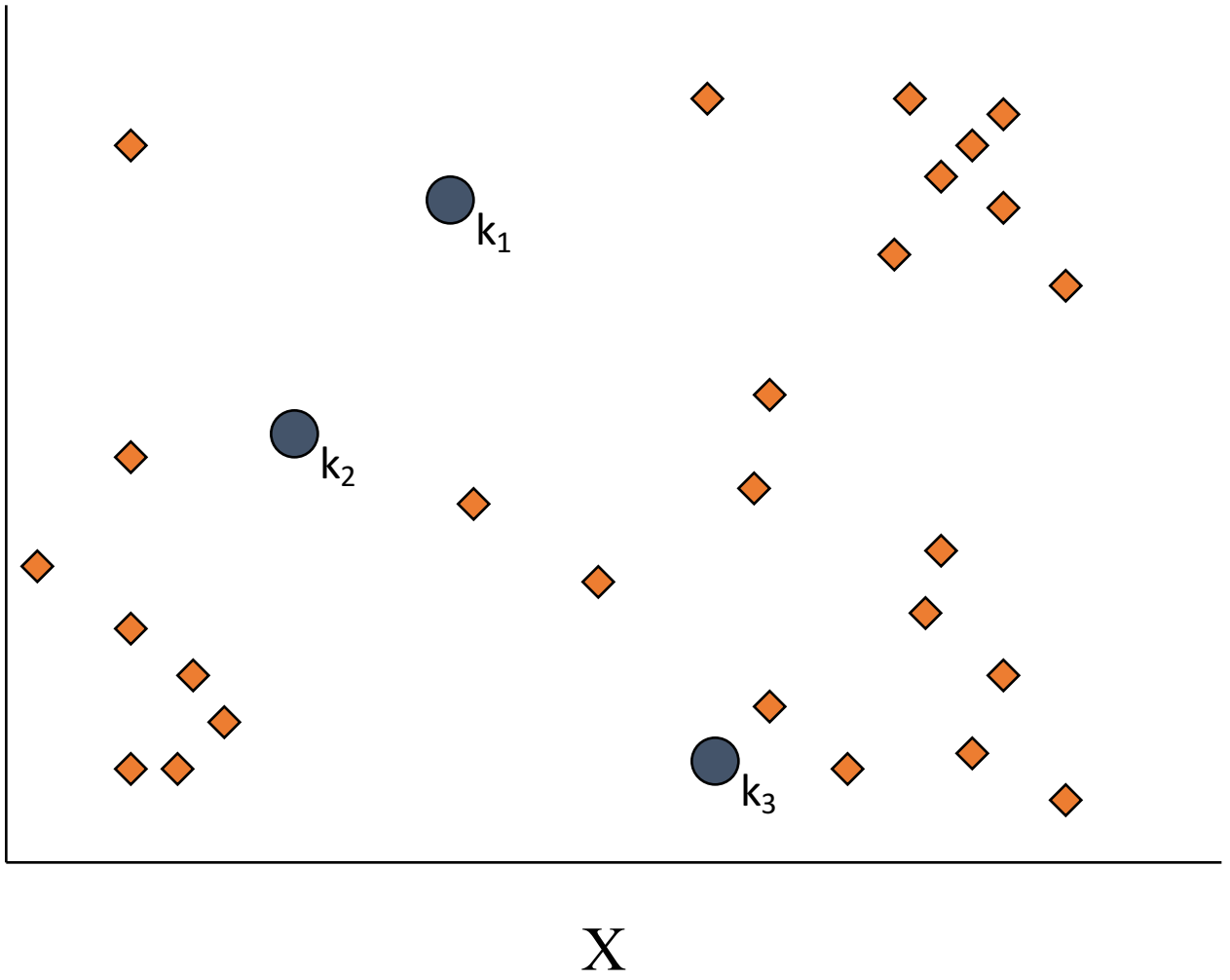
3. Decide the class memberships of the *N* objects by assigning them to the nearest cluster center.

4. Re-estimate the *K* cluster centers, by assuming the memberships found above are correct.

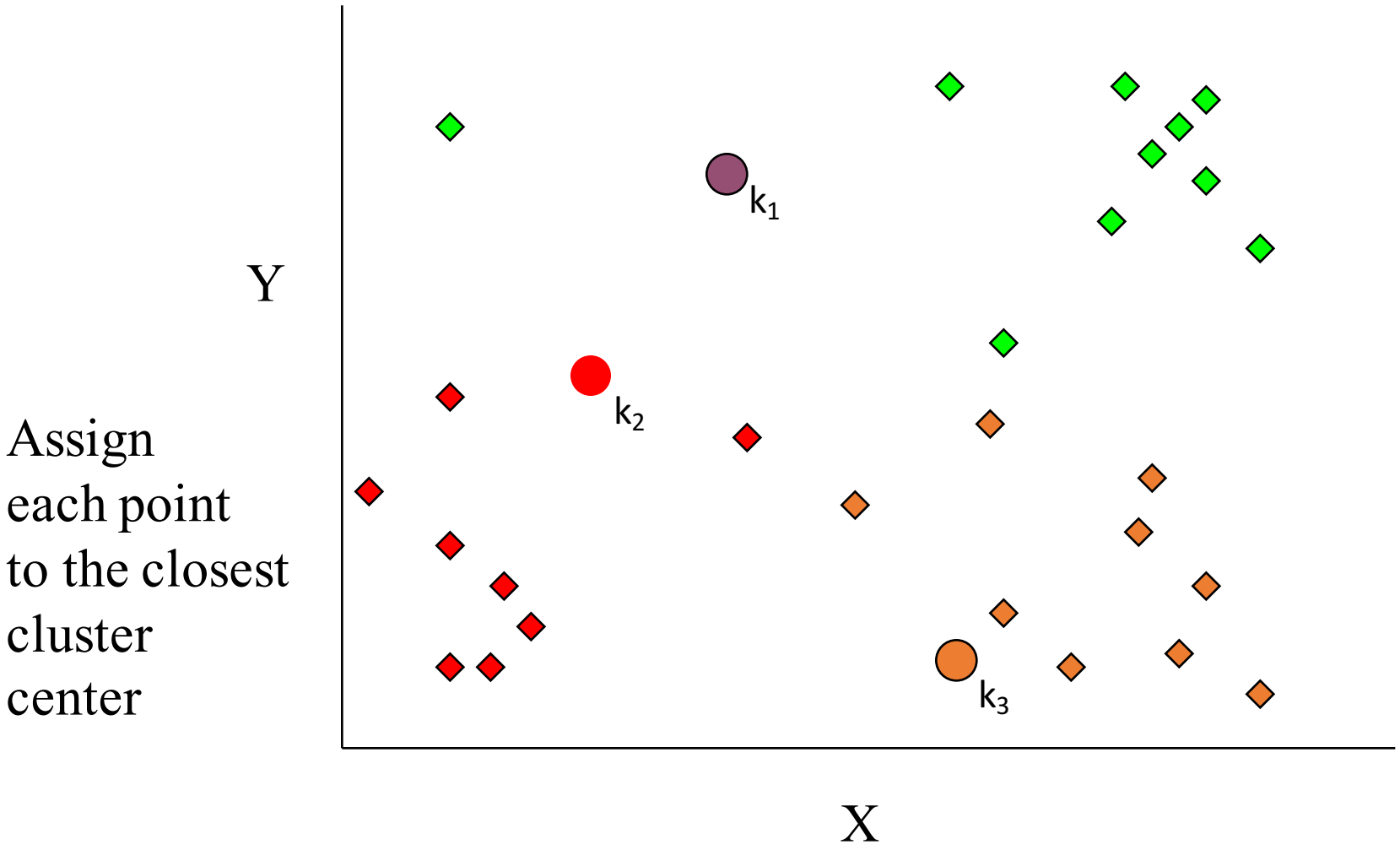
5. Repeat 3 and 4 until none of the *N* objects changed membership in the last iteration.

K-means example, step 1

Pick 3 initial cluster centers (randomly)

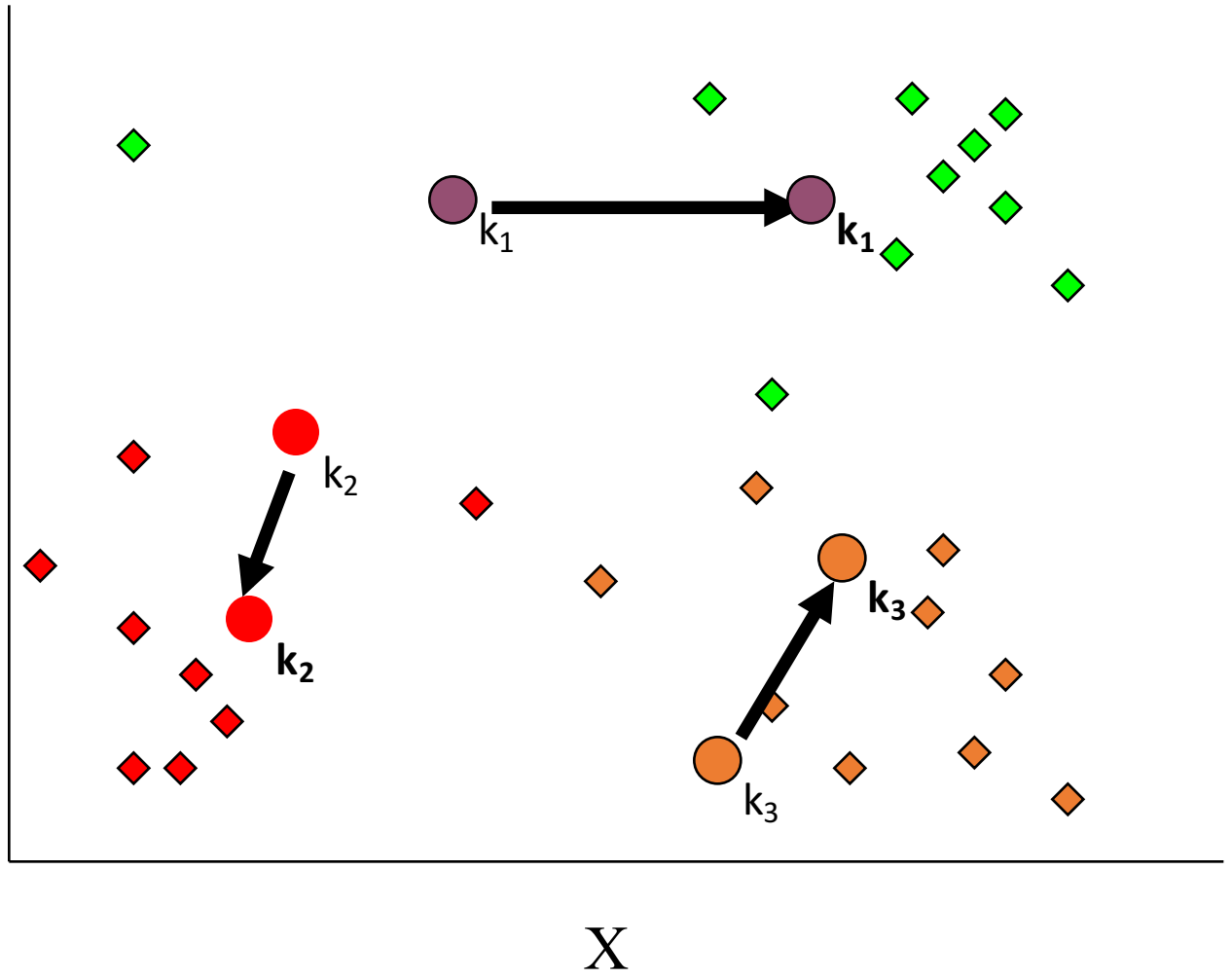


K-means example, step 2



K-means example, step 3

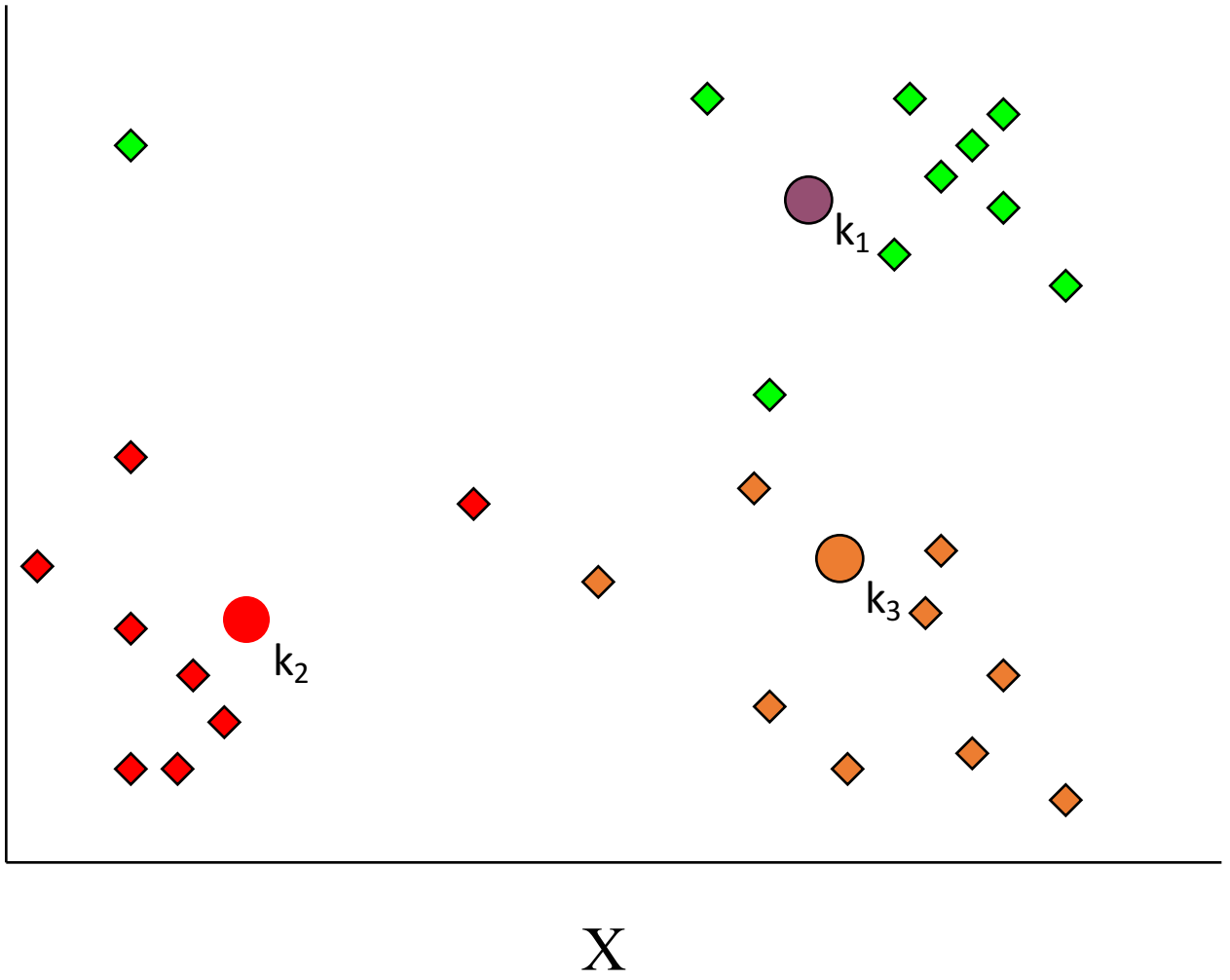
Move each cluster center to the mean of each cluster



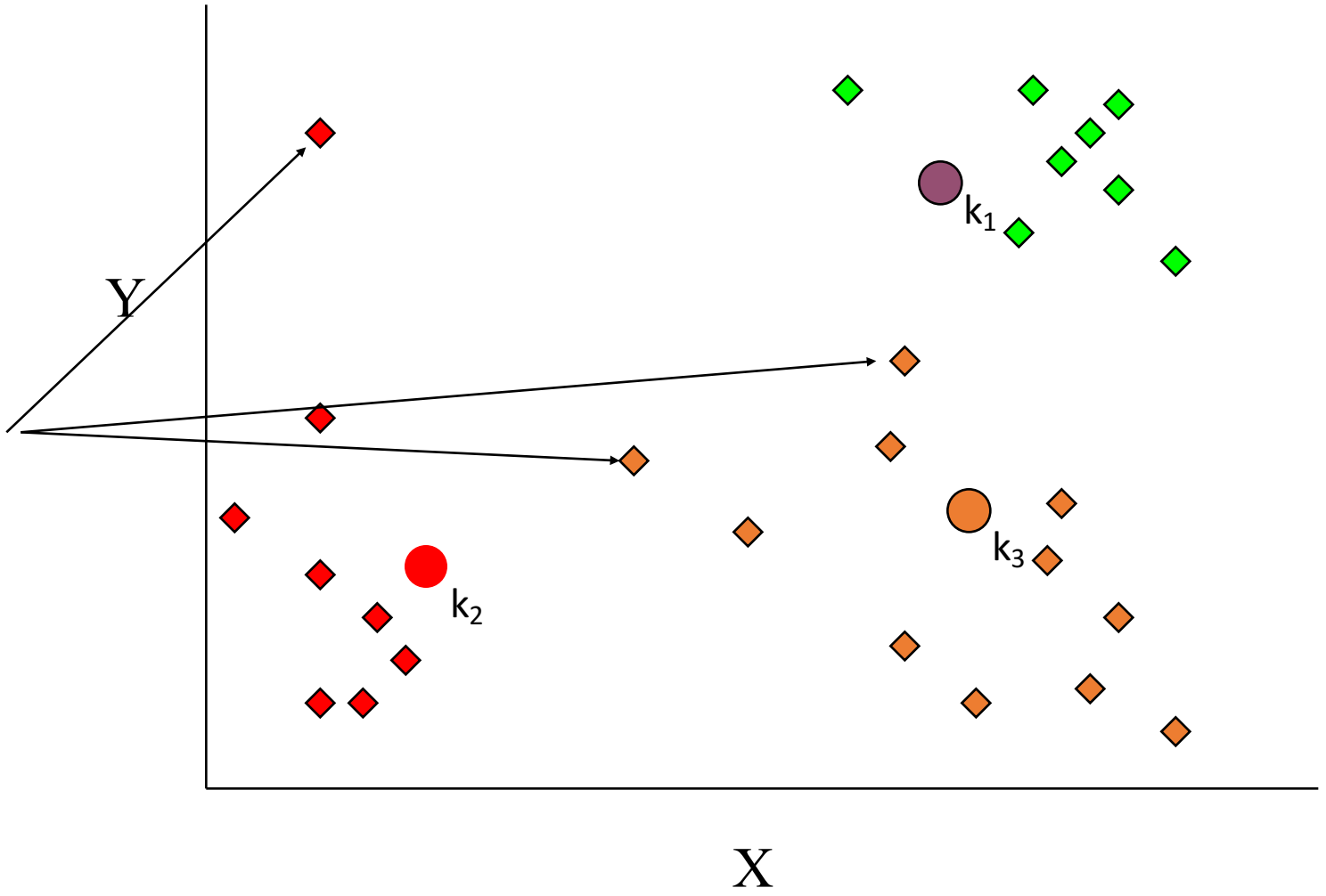
K-means example, step 4

Reassign points closest to a different new cluster center

Q: Which points are reassigned?

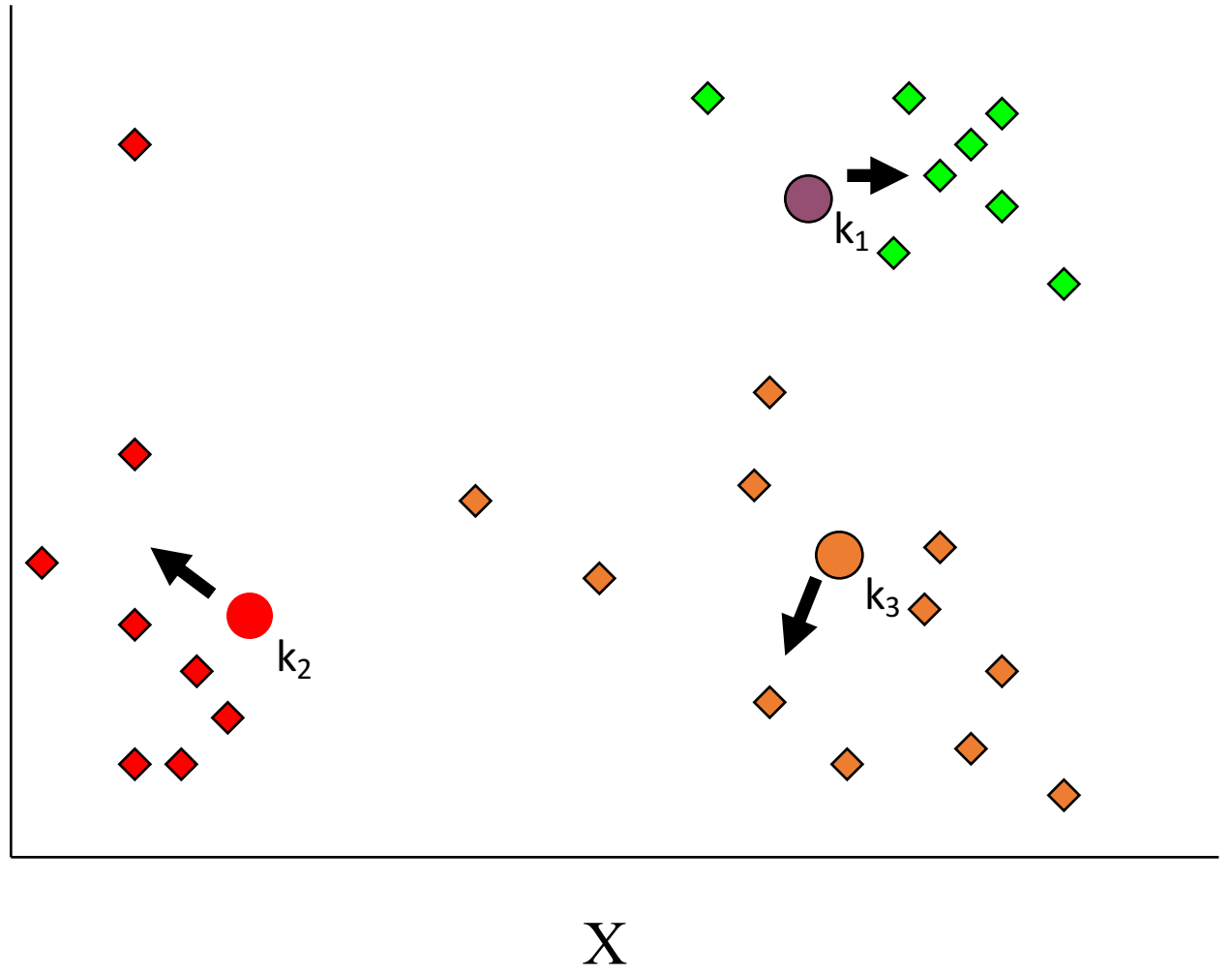


K-means example, step 4 …

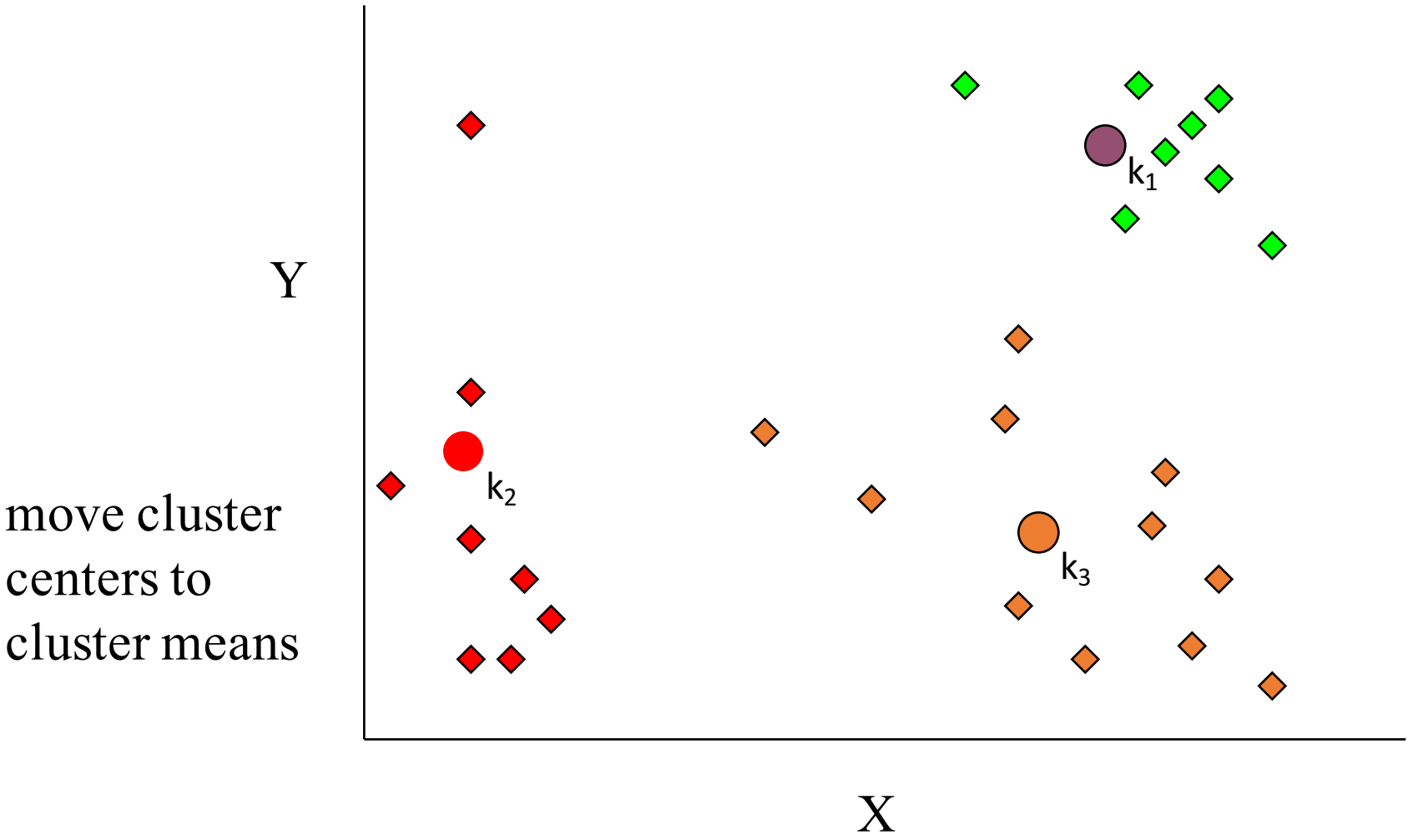


K-means example, step 4b

re-compute cluster means

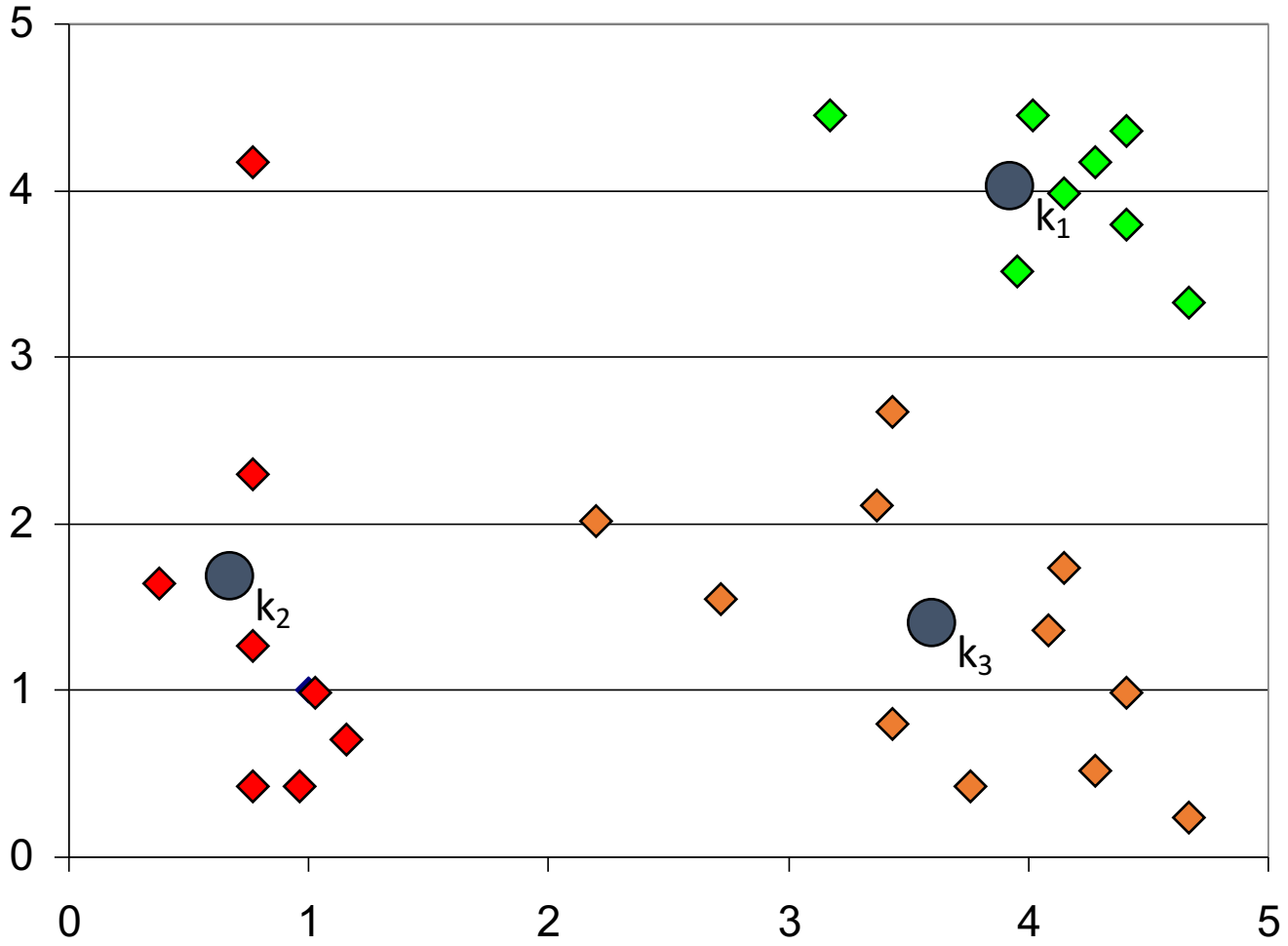


K-means example, step 5



K-means Clustering: Finished!

Re-assign and move centers, until …
no objects changed membership.



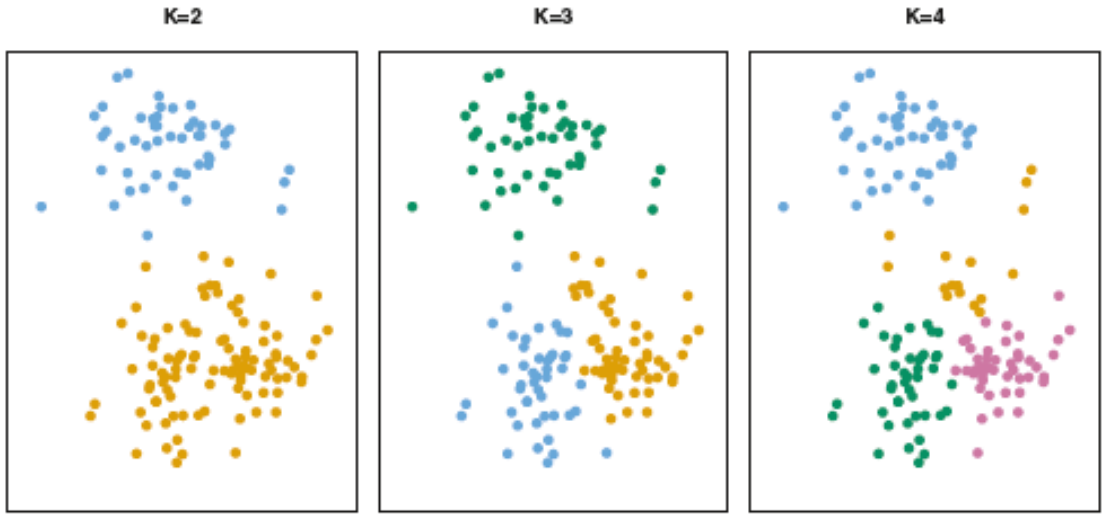
K-means details

- Formally: k-means minimize the within-cluster sum of squares (WCSS) objective function
- Global optimum is NP-hard: find local optimum – initialization matters
- Convergence of heuristic algorithm is guaranteed for a proper distance metric
- For gene expression we may wish to use a similarly measure that is not a distance metric such as rank correlation-works well in practice

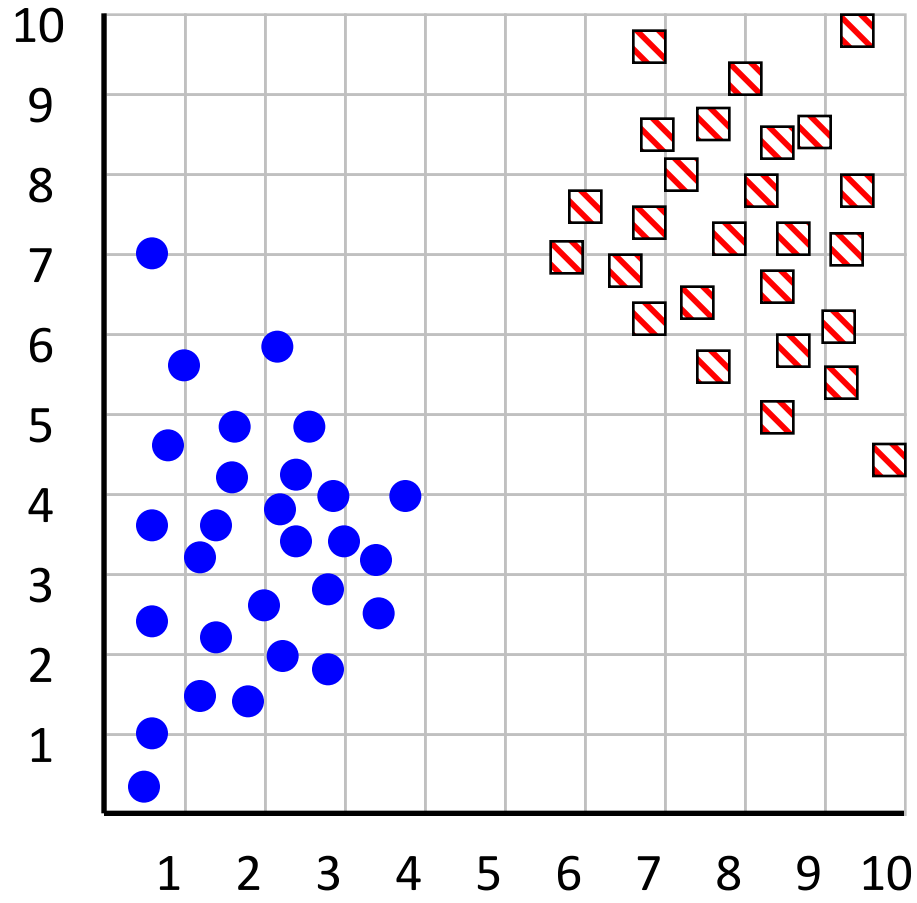
$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x} \in S_i} \|\mathbf{x} - \boldsymbol{\mu}_i\|^2$$

Optimal number of clusters

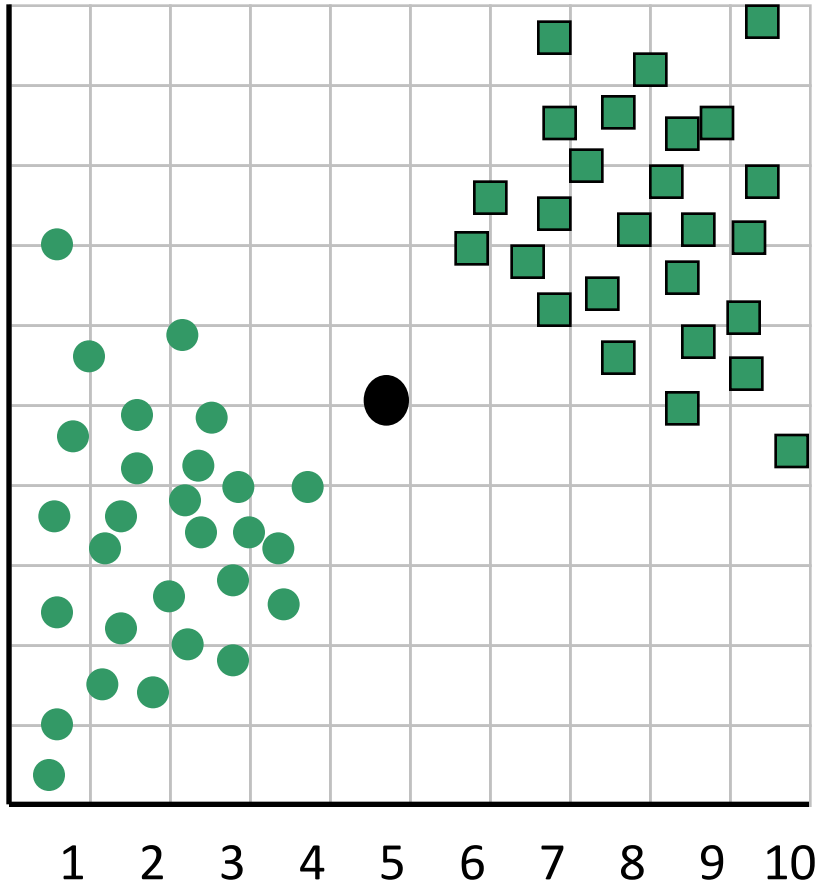
- No right answer
- Depends on assumptions/prior expectation about data distribution
  - What is the variance in a cluster
  - What shape should the clusters be?



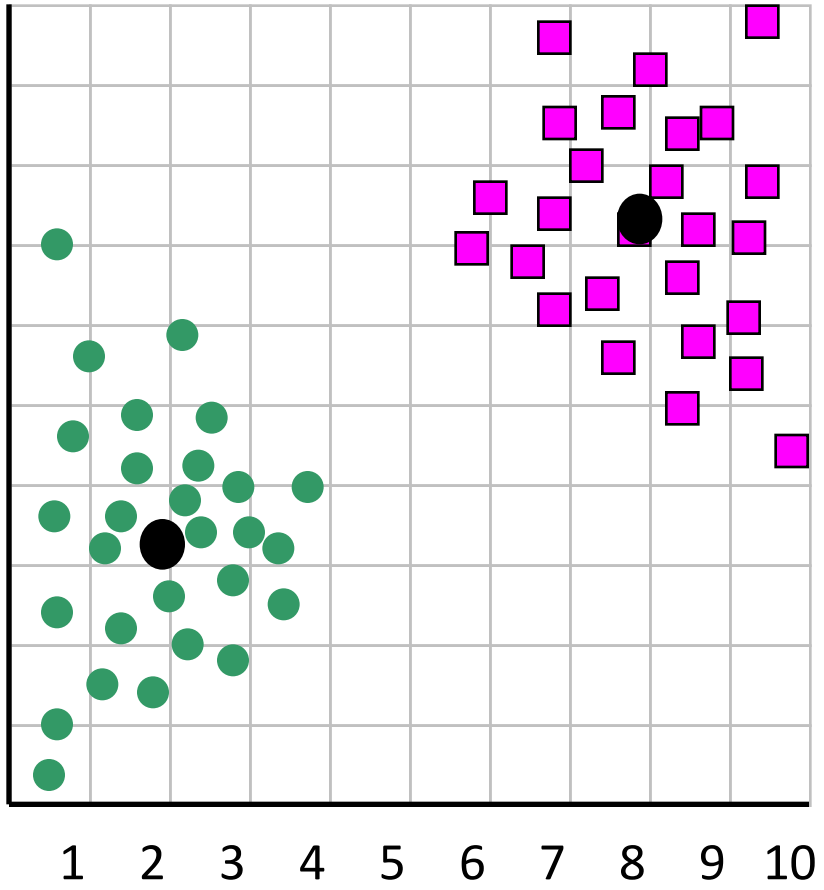
One option: How does the objective function change as we increase k



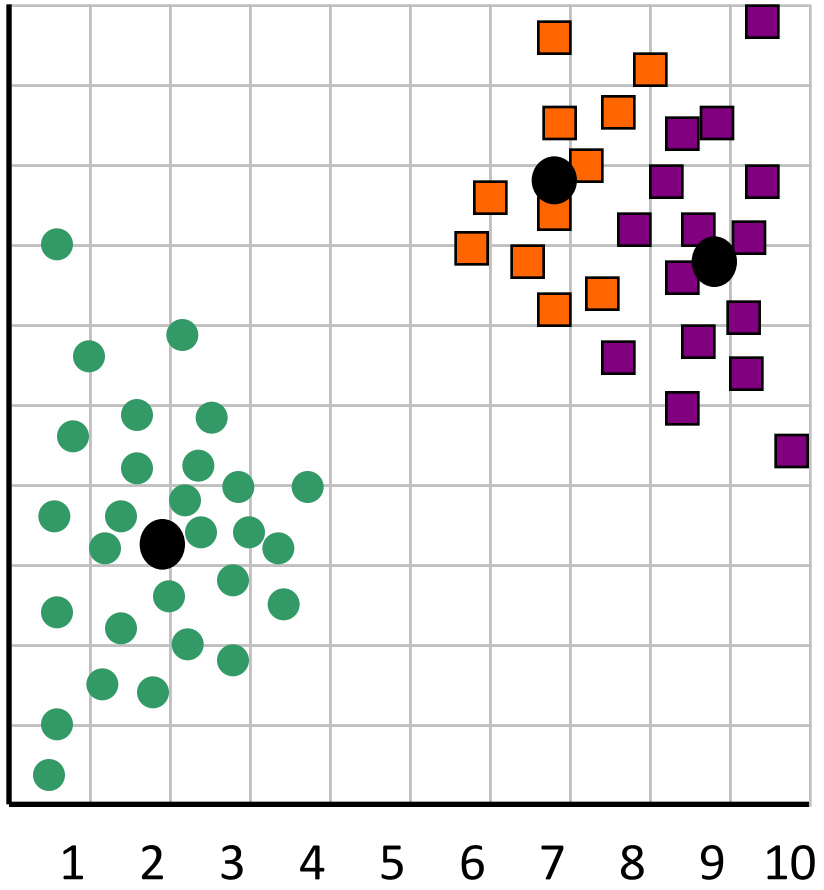
When k = 1, the objective function is 873.0



When k = 2, the objective function is 173.1

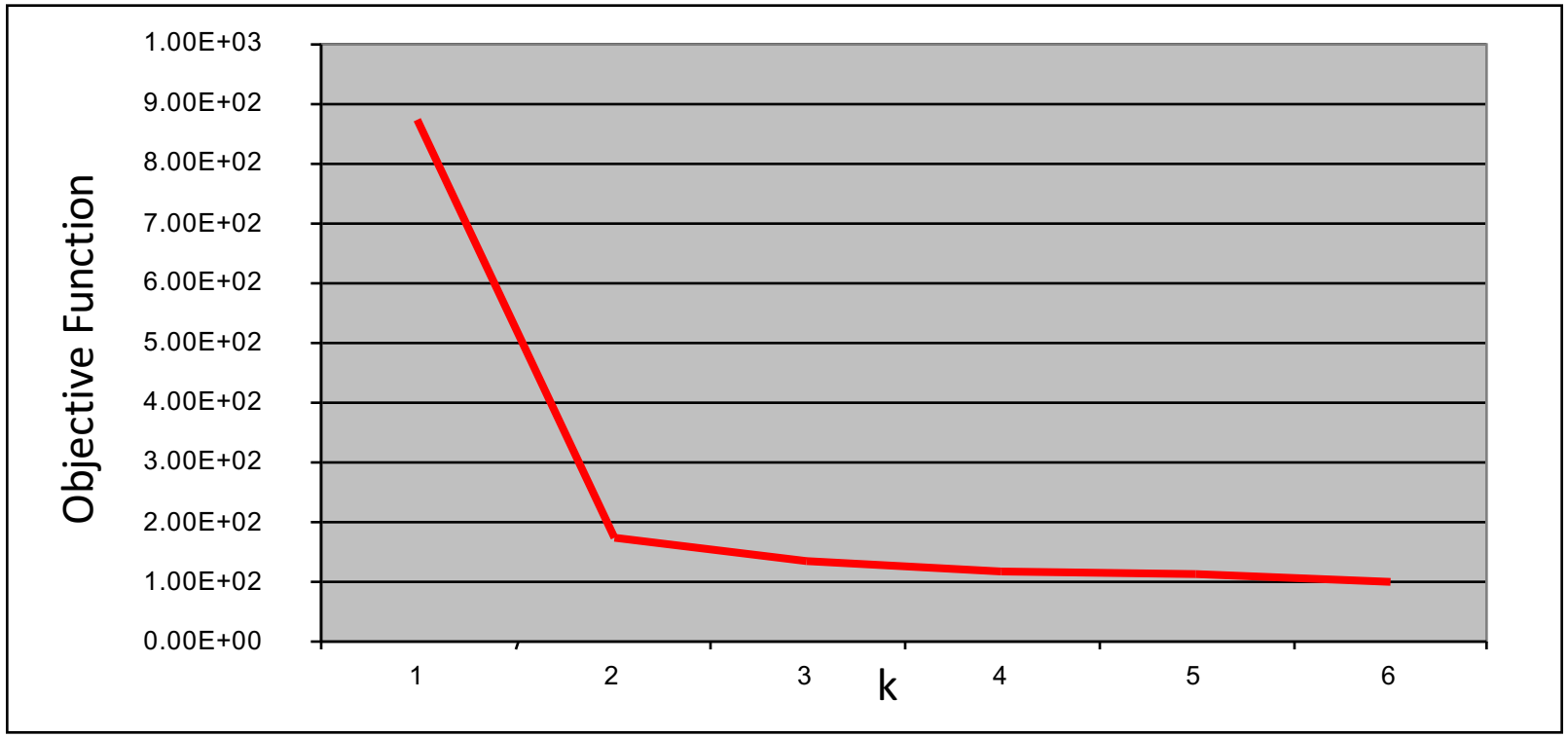


When k = 3, the objective function is 133.6



We can plot the objective function values for k equals 1 to 6…

The abrupt change at k = 2, is highly suggestive of two clusters in the data. This technique for determining the number of clusters is known as “knee finding” or “elbow finding”.



Note that the results are not always as clear cut as in this toy example

Model based clustering

- K-means assumes that the clusters are equal area “spheres”: we assume that assignment to the nearest cluster is correct.
- We can define a more general framework but we have to make more model assumptions
- Popular model based technique is gaussian mixture models

Gaussian Mixture Models

- Decide the number of clusters, K
- Initialize parameters (randomly)

$$\mathbf{x} \sim \mathcal{N}_k(\boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

with *k*-dimensional mean vector

$$\boldsymbol{\mu} = [\mathrm{E}[X_1], \mathrm{E}[X_2], \ldots, \mathrm{E}[X_k]]$$

and $k \times k$ covariance matrix

$$\boldsymbol{\Sigma} = [\mathrm{Cov}[X_i, X_j]], i = 1, 2, \ldots, k; j = 1, 2, \ldots, k$$

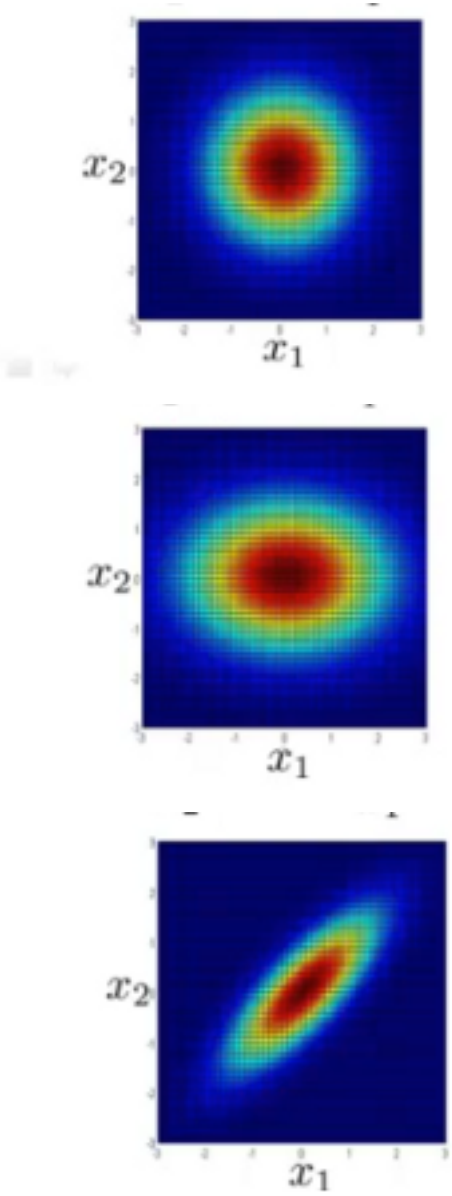
- E-step: assign *probabilistic* membership to all input samples

$$p(x) = \sum_{i=0}^{k} \pi_i N(x|\mu_k, \Sigma_k)$$

- M-step: re-estimate parameters based on *probabilistic* membership
- Repeat until change in parameters is smaller than a threshold

Number of parameters

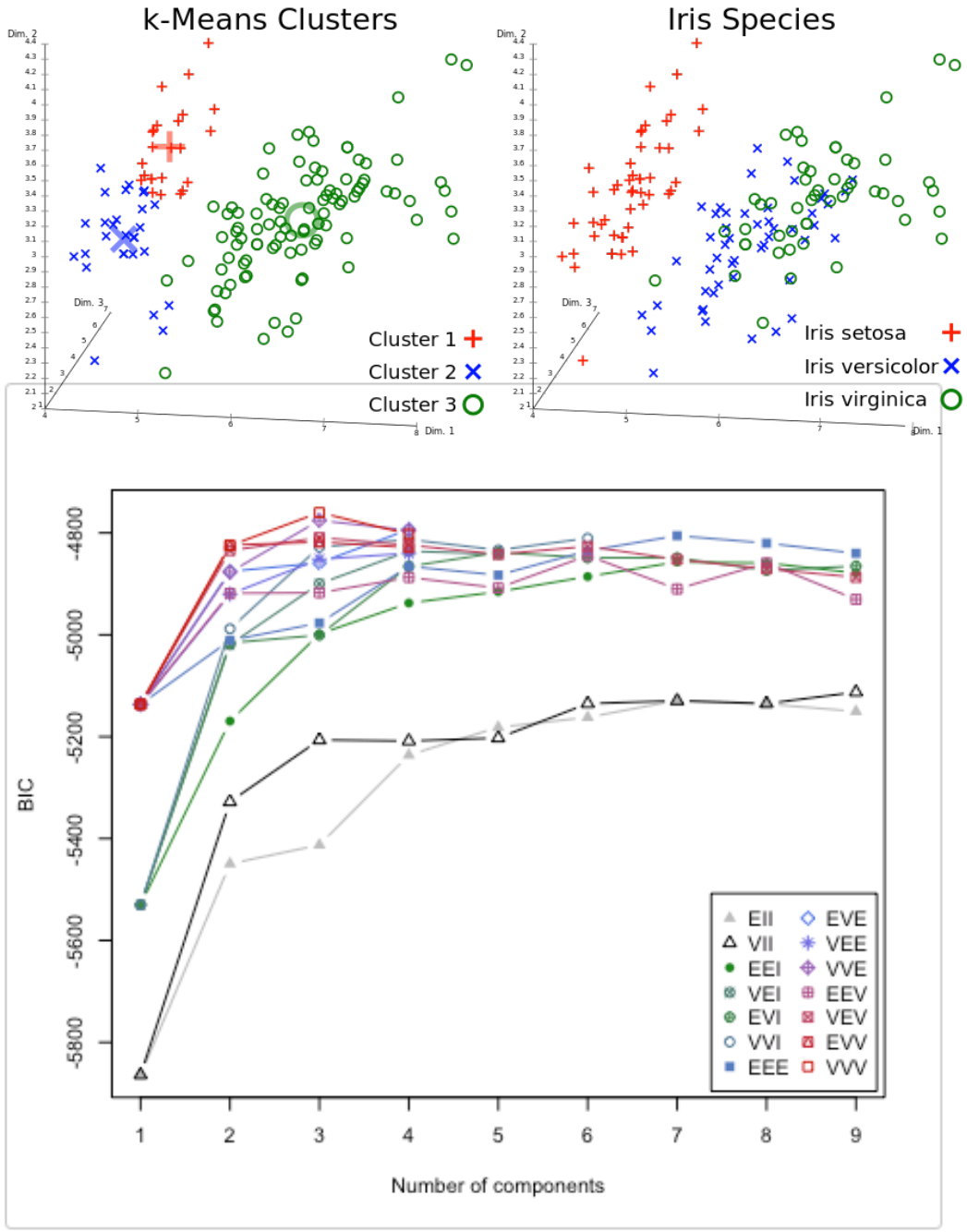
- We can specify relative flexibility for
  - We specify if clusters can have different covariance, if the diagonal entries are equal and if the non-diagonal entries are non-zero
    - "EII": spherical, equal volume
    - "VII": spherical, unequal volume
    - "EEI": diagonal, equal volume and shape
    - "VEI": diagonal, varying volume, equal shape
    - "EVI": diagonal, equal volume, varying shape
    - "VVI": diagonal, varying volume and shape
    - "EEE": ellipsoidal, equal volume, shape, and orientation
    - "EEV": ellipsoidal, equal volume and equal shape
    - "VEV": ellipsoidal, equal shape
    - "VVV": ellipsoidal, varying volume, shape, and orientation
- What about the number of clusters
  - We can compute the likelihood based on the model
  - Increasing the number of parameters always increases the likelihood of the data we need to define a parameter penalty



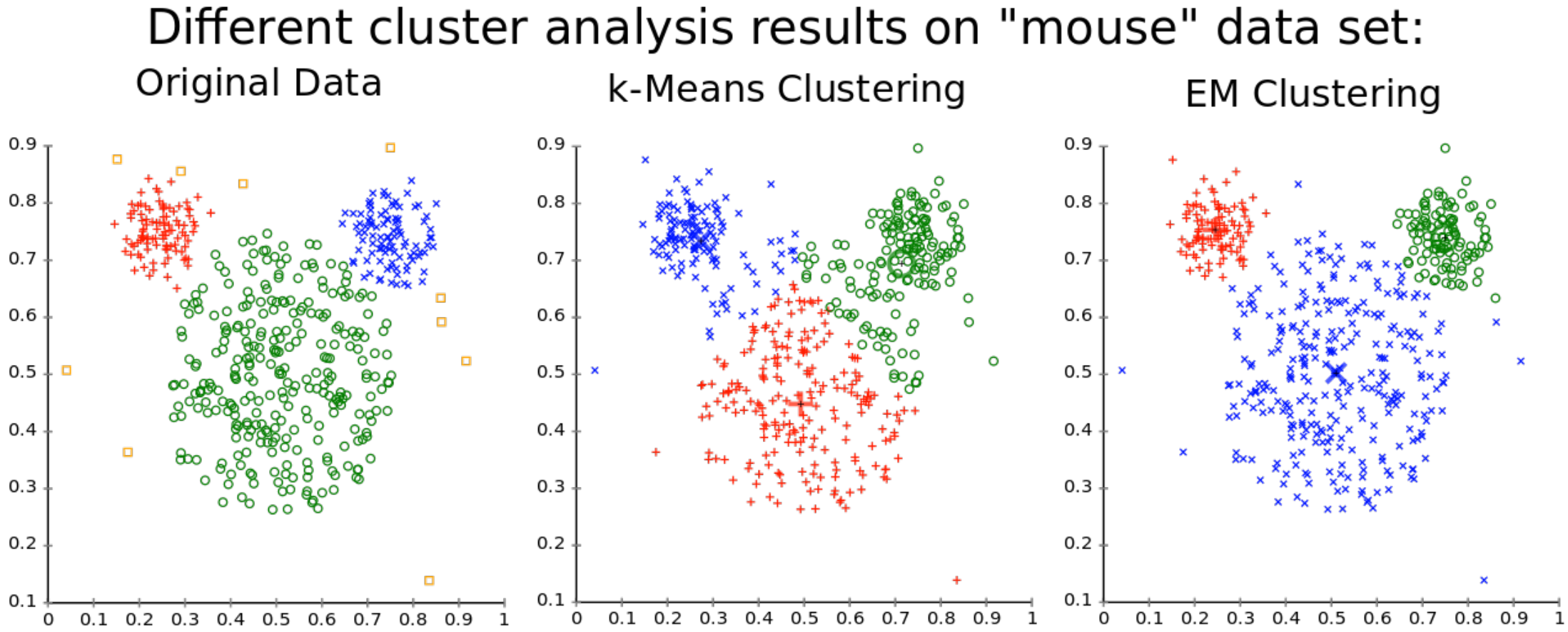
Example

$$\mathrm{BIC} = -2 \cdot \ln \hat{L} + k \cdot \ln(n).$$

- x = the observed data;
- θ = the parameters of the model;
- n = the number of data points in xx, the number of observations, or equivalently, the sample size;
- k = the number of free parameters to be estimated. If the model under consideration is a linear regression,
- L= the maximized value of the likelihood function of the model MM, i.e. Ł =p(x|θ ,M) where θ are the parameter values that maximize the likelihood function.



K-means vs model-based



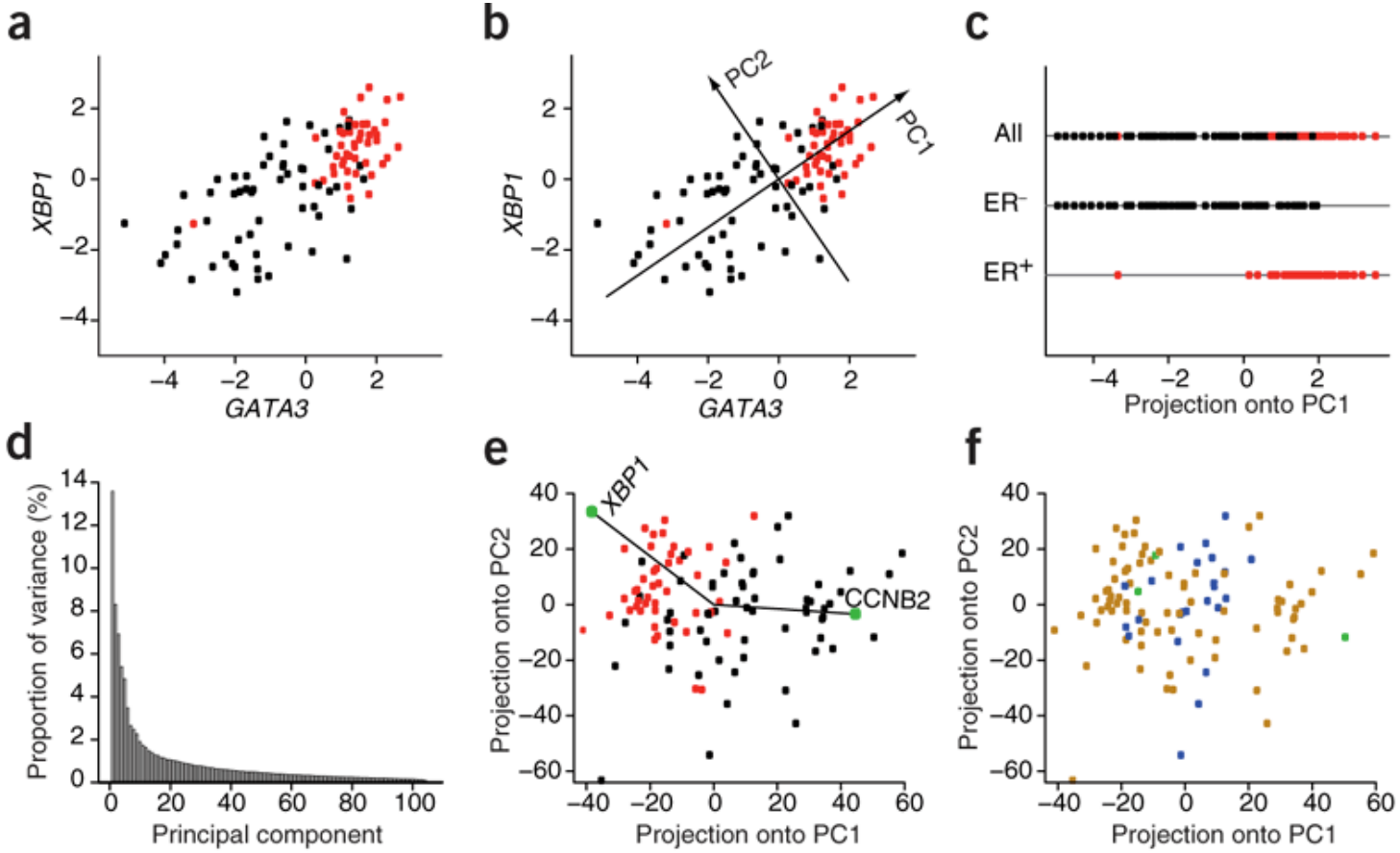
- K-means is preferred for gene-expression data
  - K-means is quite fast and relatively stable across initializations
  - Difficult to specify the correct model

PCA for gene expression

- Given a gene-by-sample matrix M we decompose (row a centered and scaled) M as UDV^T
  - We don't usually care about total expression level and the dynamic range which may be dependent on tehcnical factors
- U, V orthonormal
- D diagonal-elements are eigenvalues
  - Variance explained
- Columns of are
  - principle components
  - Eigengenes/metagenes
  - recurrent patters of gene expression
- Columns of U are the "loadings" is the correlation between it and the component
- Truncating U,V,D to the first k dimensions gives the best k-rank approximation of U

PCA for gene expression: example

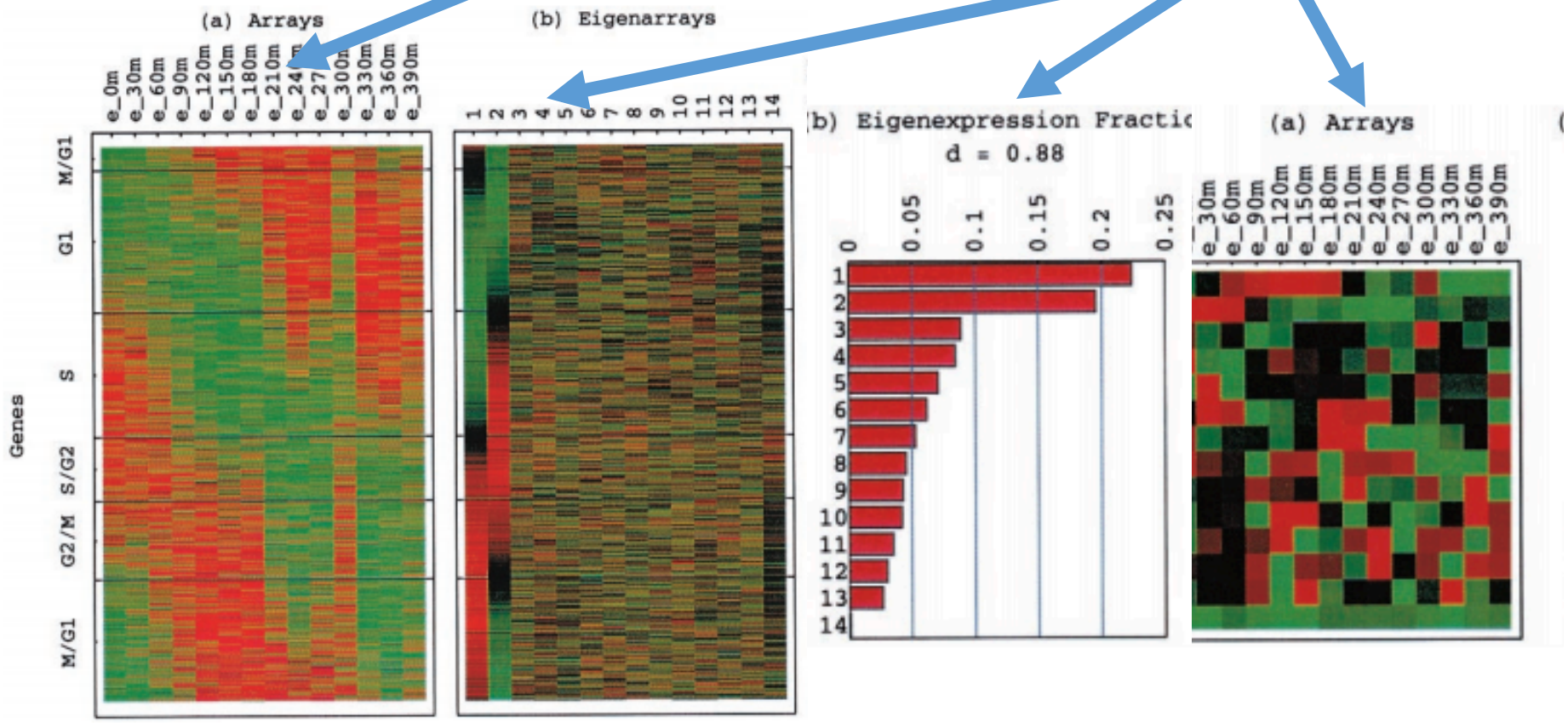
- Breast cancer samples
- Estrogen receptor (ER) status: ER^+, red; ER^-, black).
- (**b**) PCA identifies the two directions (PC1 and PC2) along which the data have the largest spread. (**c**) Samples plotted in one dimension using their projections onto the first principal component (PC1) for ER^+, ER^- and all samples separately.
- (**e**) PCA biplot with samples plotted in two dimensions using their projections onto the first two principal components, and two genes plotted using their weights for the components (green points).
- (**f**) Samples colored according to *ERBB2* status (blue, $ERBB2^+$; brown, $ERBB2^-$; green, unknown).



What is principal component analysis?
Nature Biotech

PCA applied to cell cycle data

$$\text{GeneExpression} = UDV^T$$



Singular value decomposition for genome-wide expression data processing and modeling. *PNAS*

Relationship with original data

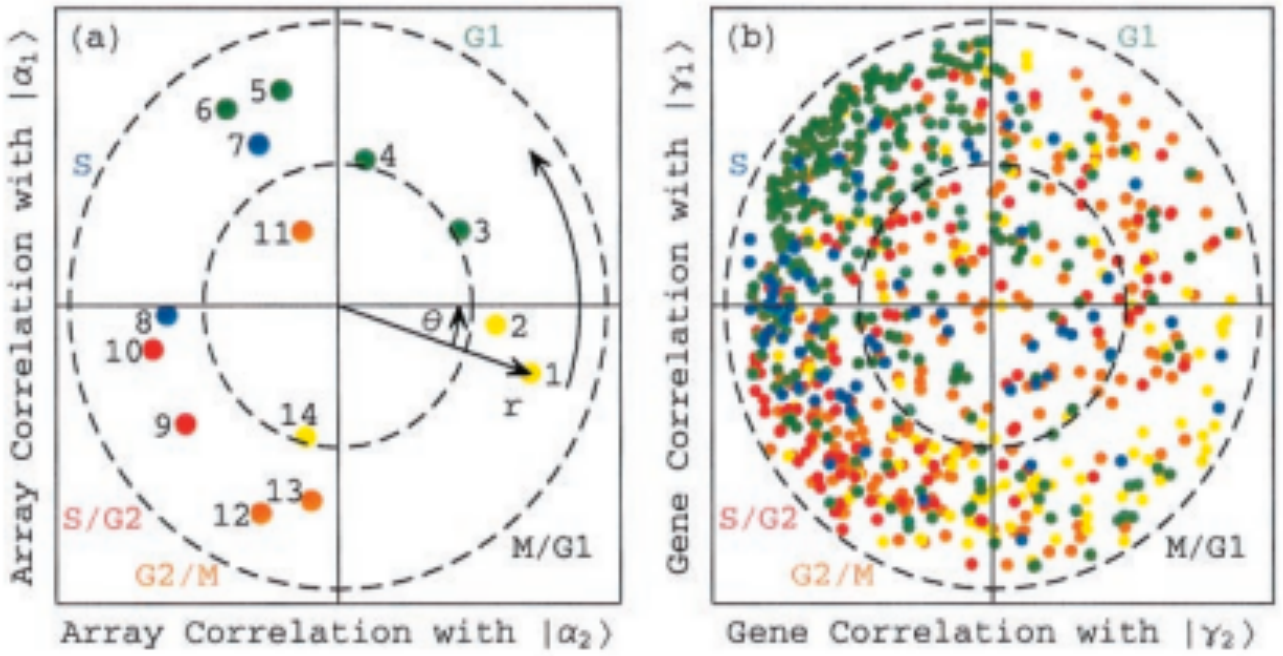


Fig. 2. Normalized elutriation expression in the subspace associated with the cell cycle. (*a*) Array correlation with $|\alpha_1\rangle_N$ along the *y*-axis vs. that with $|\alpha_2\rangle_N$ along the *x*-axis, color-coded according to the classification of the arrays into the five cell cycle stages, M/G_1 (yellow), G_1 (green), S (blue), S/G_2 (red), and G_2/M (orange). The dashed unit and half-unit circles outline 100% and 25% of overall normalized array expression in the $|\alpha_1\rangle_N$ and $|\alpha_2\rangle_N$ subspace. (*b*) Correlation of each gene with $|\gamma_1\rangle_N$ vs. that with $|\gamma_2\rangle_N$, for 784 cell cycle regulated genes, color-coded according to the classification by Spellman *et al.* (3).

Other decomposition techniques

- Non negative matrix factorization
- **A= WH** (A,W,H are non-negative)
- Many computational methods
  - Cost function |A-WH|
    - Squared error-aka Frobenius norm
    - Kullback–Leibler divergence to positive matrices
  - Optimization procedure
  - Most use stochastic initialization and the results don't always converge to the same answer
- H defined a meta-gene space: similar to eigengenes
- Classification can be done in the meta-gene space

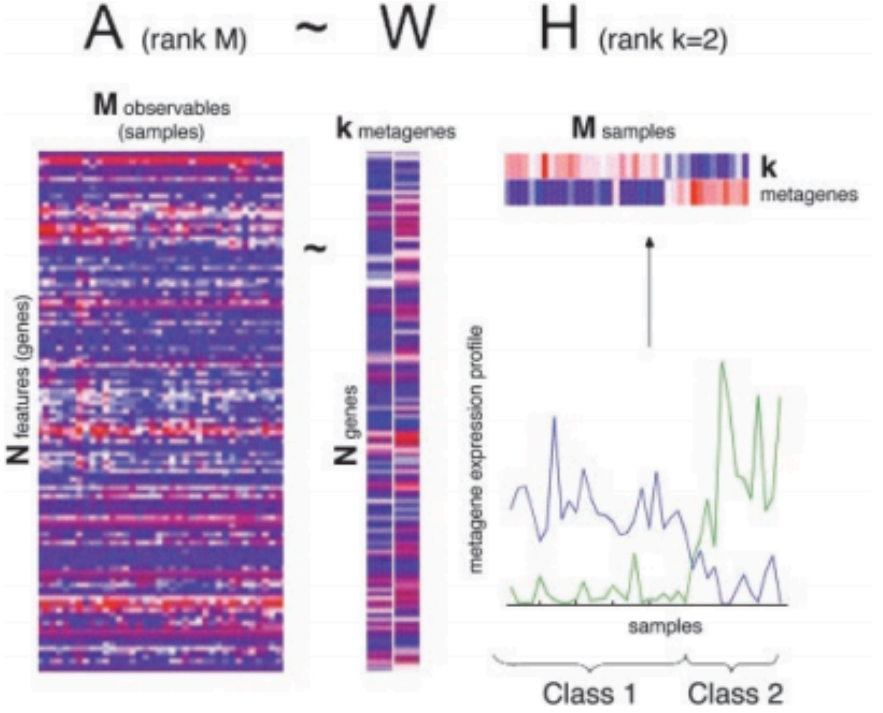
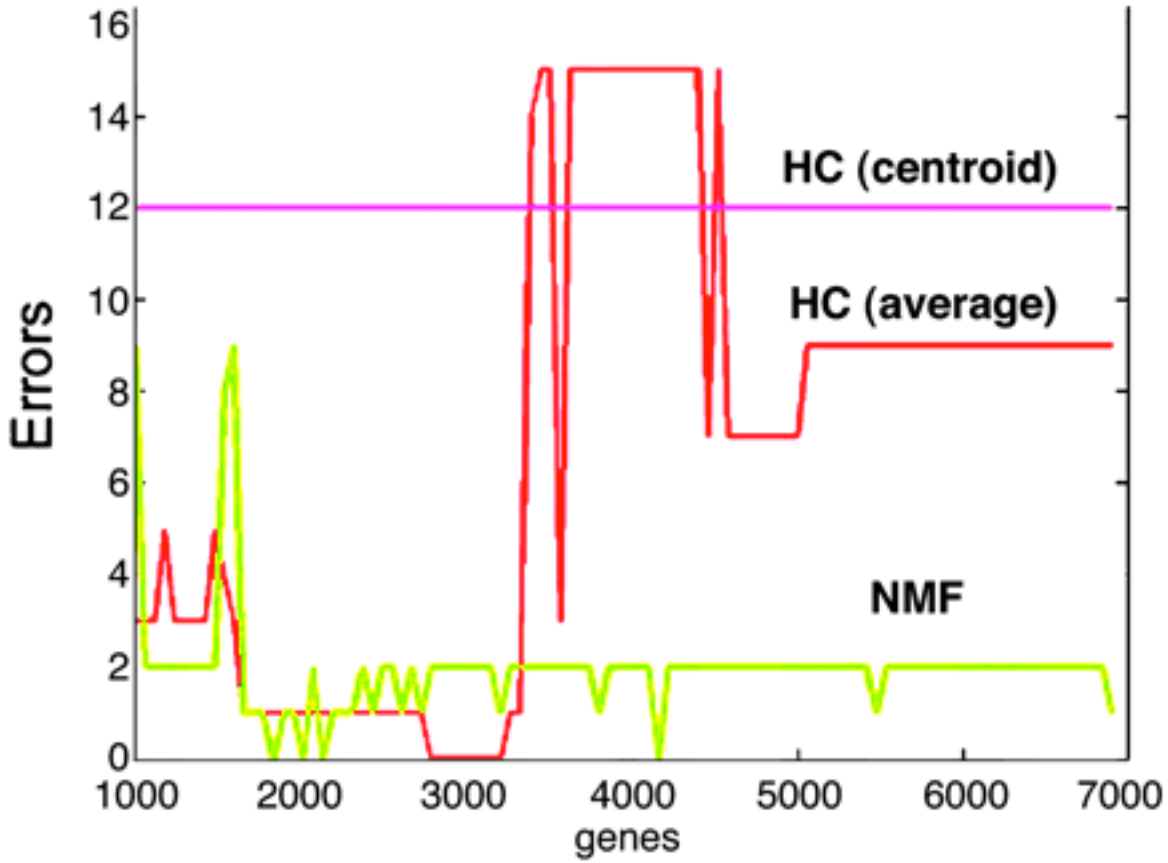


Fig. 1. A rank-2 reduction of a DNA microarray of *N* genes and *M* samples is obtained by NMF, *A* ~ *WH*. For better visibility, *H* and *W* are shown with exaggerated width compared with original data in *A*, and a white line separates the two columns of *W*. Metagene expression levels (rows of *H*) are color coded by using a heat color map, from dark blue (minimum) to dark red (maximum). The same data are shown as continuous profiles below. The relative amplitudes of the two metagenes determine two classes of samples, class 1 and class 2. Here, samples have been ordered to better expose the class distinction.

Metagenes and molecular pattern discovery using matrix factorization PNAS

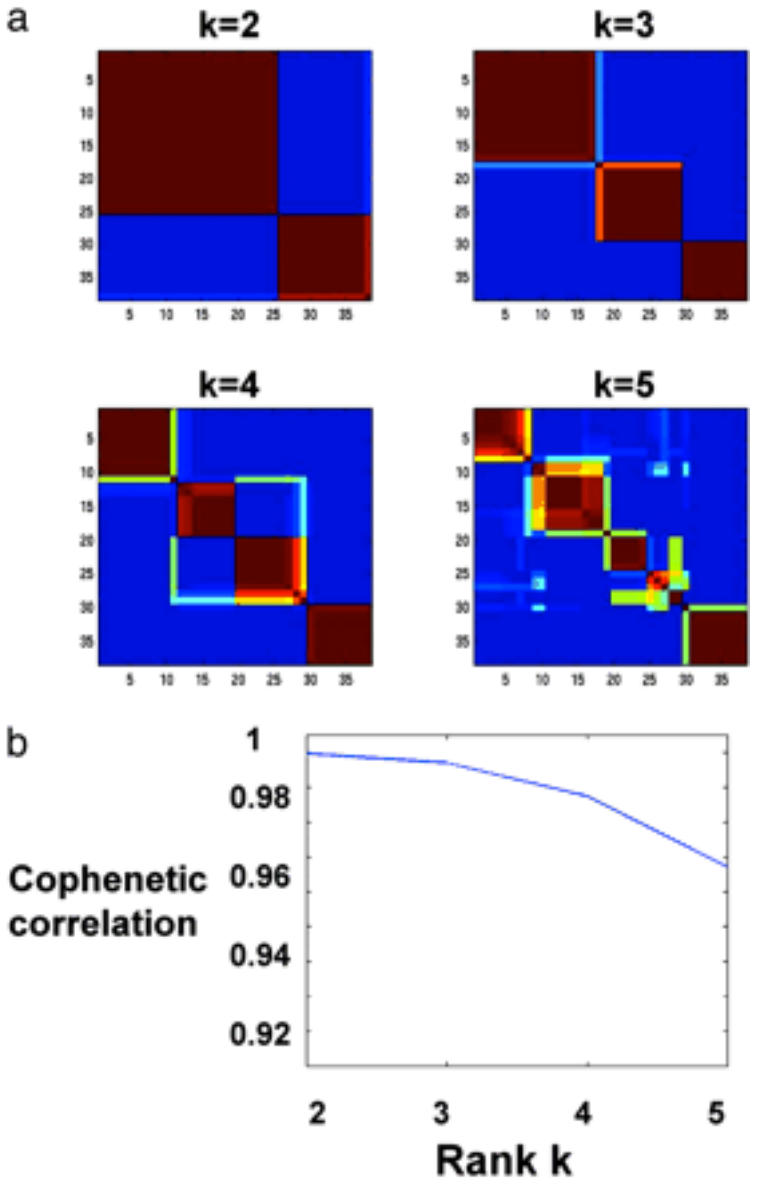
Example: classifying leukemia

- acute myelogenous leukemia (AML)
- acute lymphoblastic leukemia (ALL),
- Hierarchical clustering did no correctly recover the subtypes

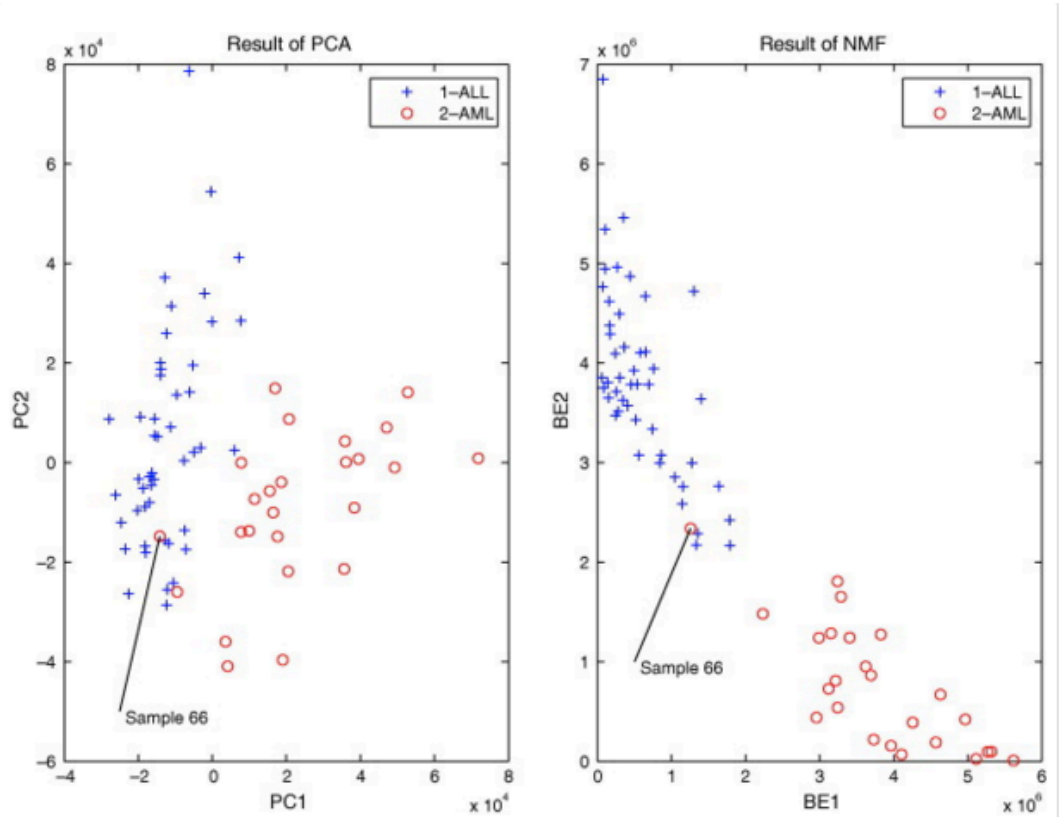


Determining k by cluster stability

- Consensus matrix C : what fraction of times two data points ended up in the same cluster
  - NMF with different initializations
  - For deterministic methods we can sue subsampling
- Cophenetic correlation
  - 1-C gives us a distance
  - We cluster based on this distance to get a clustering tree
  - Cophenetic correlation tells us how well the resulting tree distance reproduces the initial 1-C



NMF vs PCA



Reducing microarray data via nonnegative matrix factorization for visualization and clustering analysis

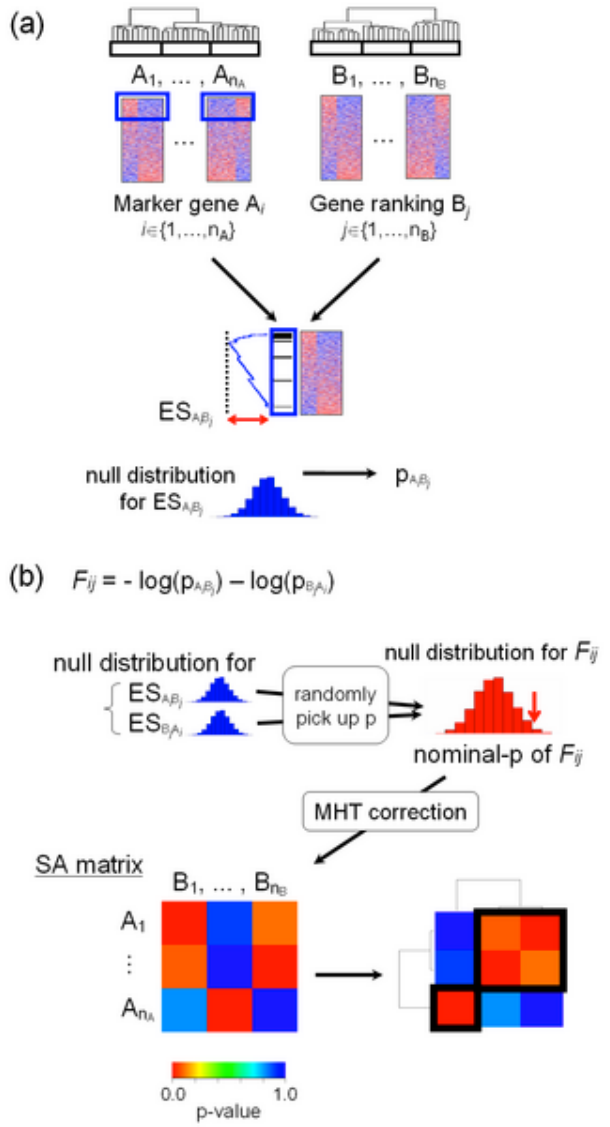
- NMF operates in the original non-negative measurement space
  - Highly expressed genes matter more
- Positivity constraint is advantageous: positive correlation among genes is more likely to be biologically meaningful
- NMF may more accurately capture the data generating process
- **A= WH** : Toy Biological interpretation
  - Assume k=2
  - We have 2 transcription factors that activate gene signatures W_1 and W_2
  - H represents the activity of each factor in each sample
  - TF effects are additive

Example: identifying subclasses of HCC

- Hepatocellular carcinoma: highly heterogeneous liver cancer
- Study design: identify reliable molecular sub-classes that are consistent across different datasets representing different patient cohorts
- Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Research*
- Subclass Mapping: Identifying Common Subtypes in Independent Disease Data Sets *PLoS One*

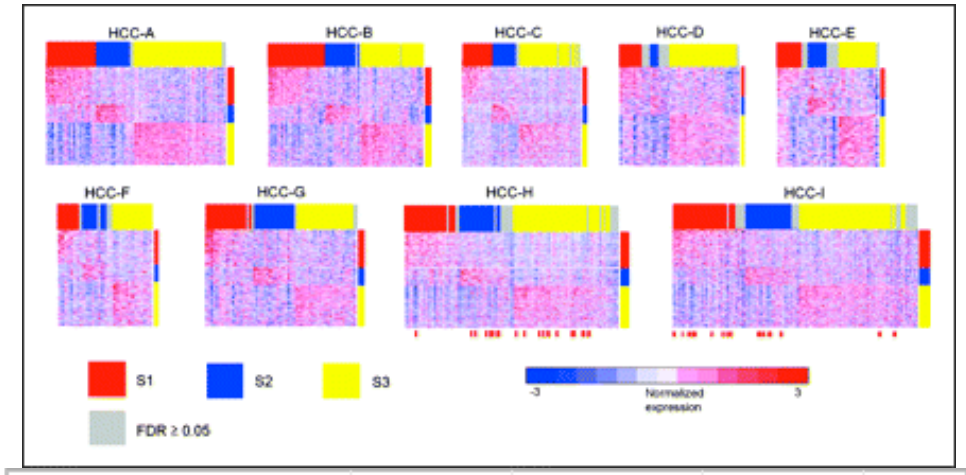
Identifying common clustering across datasets

- Two datasets A, B
- Cluster A into A_1, A_2 .. And B into B_1, B_2, …
  - Heuristic: number of subclasses is set to 4, subclasses have to contain at least 10% of the samples
- For each A_i define differentially expressed genes and use them as a gene-set for gene set enrichment in B
- Compute empirical p-values
- Matrix SA—mutual enrichment information for A_i B_j
- Cluster the SA matrix



Application to HCC

- Use a consensus of hierarchical, k-means, and NMF clustering (clustering on meta-genes)
- Identify 3 common clusters with consistent clinical phenotype, molecular markers and pathway activation



Variable	S1	S2	S3	*P*
Tumor size (cm) *	3.0 [2.0,4.5]	4.5 [2.5,7.0]	2.5 [1.8,4.3]	0.003
Tumor differentiation *				
Well	8 (16%)	4 (10%)	37 (44%)	
Moderate	27 (53%)	23 (59%)	45 (53%)	<0.001
Poor	16 (31%)	12 (31%)	3 (4%)	
AFP (ng/mL) †	50 [14,332]	171 [27,1,251]	13 [5,43]	<0.001
Hepatitis B virus infection ‡	39 (38%)	27 (36%)	39 (25%)	0.05
Hepatitis C virus infection ‡	55 (53%)	44 (58%)	109 (69%)	0.03