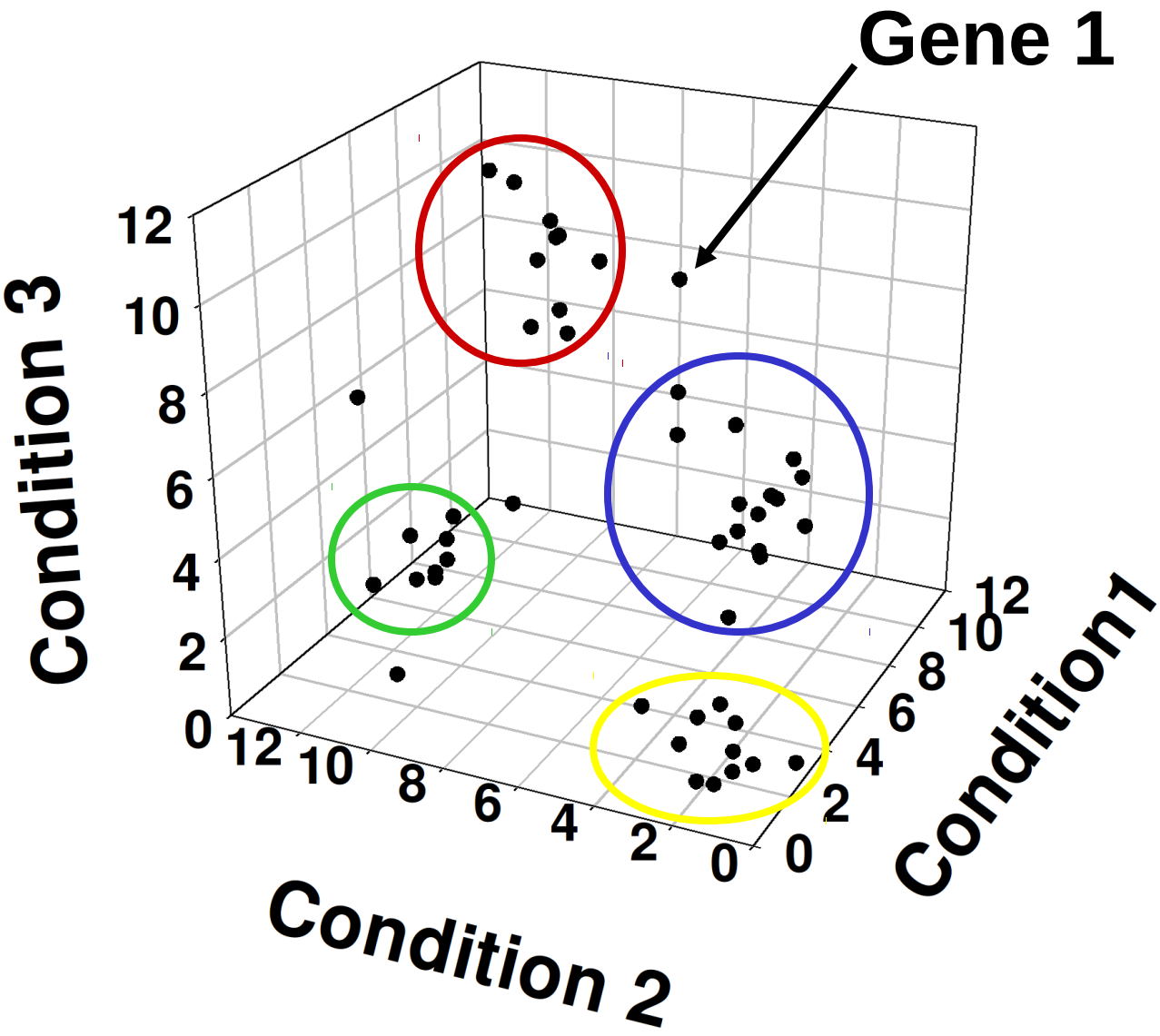
Microarrays 2

- Metrics for determining coexpression
- Unsupervised learning (clustering)

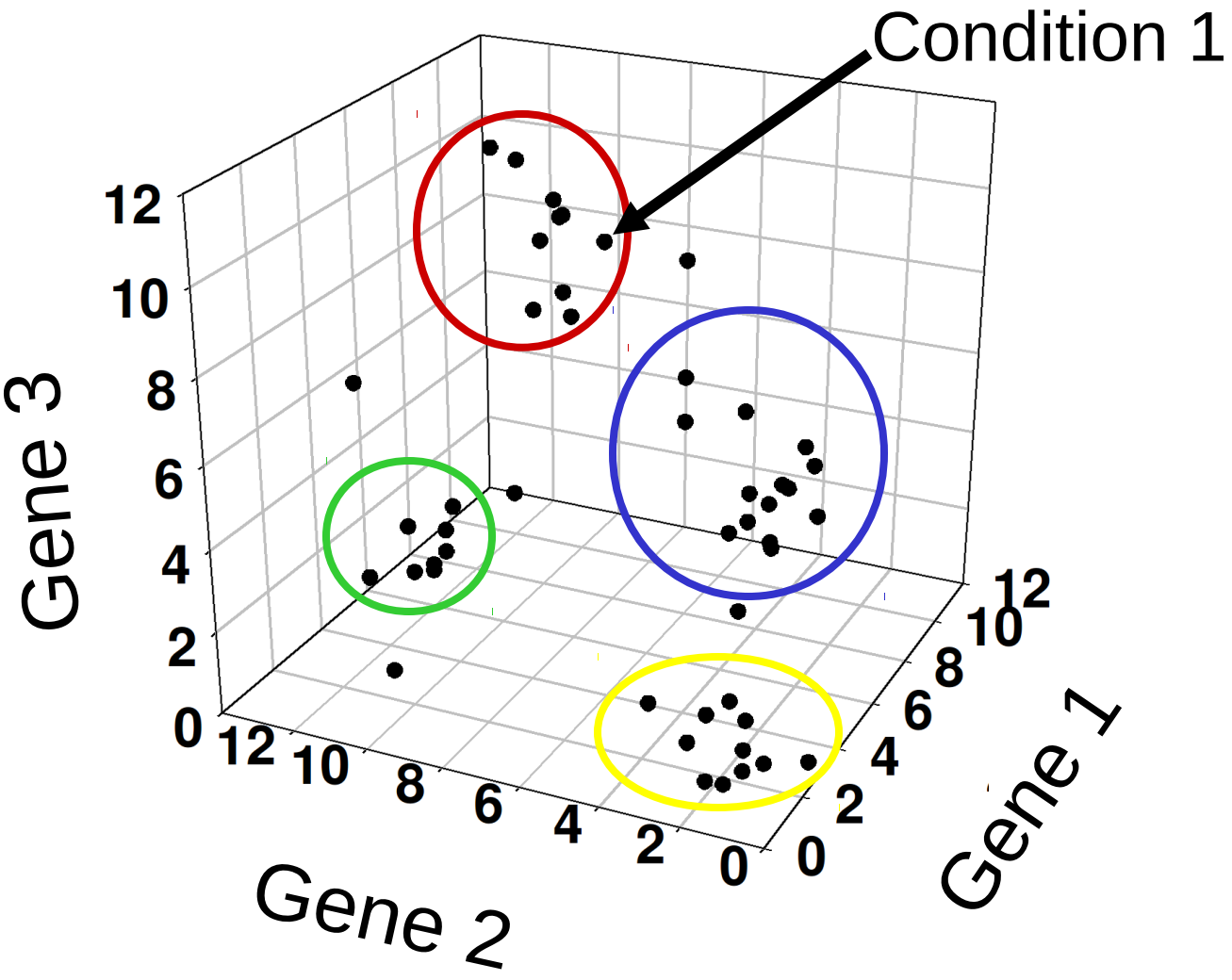
Comparing Gene Expression

- Compare the responses of two genes through different conditions (yes)
- Directly compare the expression of two genes in the same condition (no)
- Absolute versus relative changes

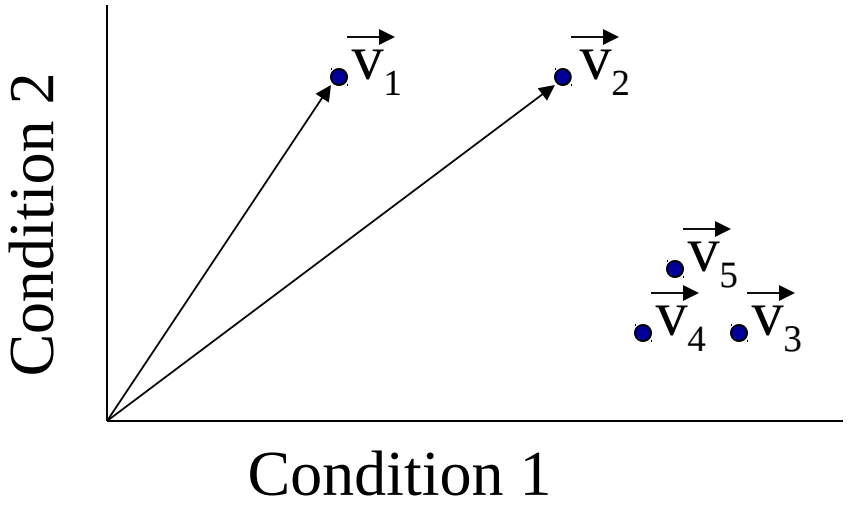
Gene Clustering



Condition Clustering



Clustering Gene Expression Data

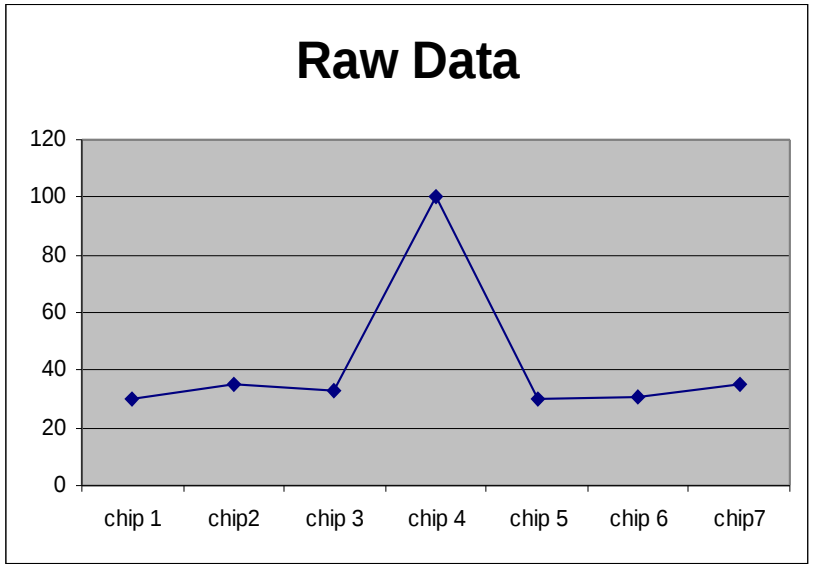


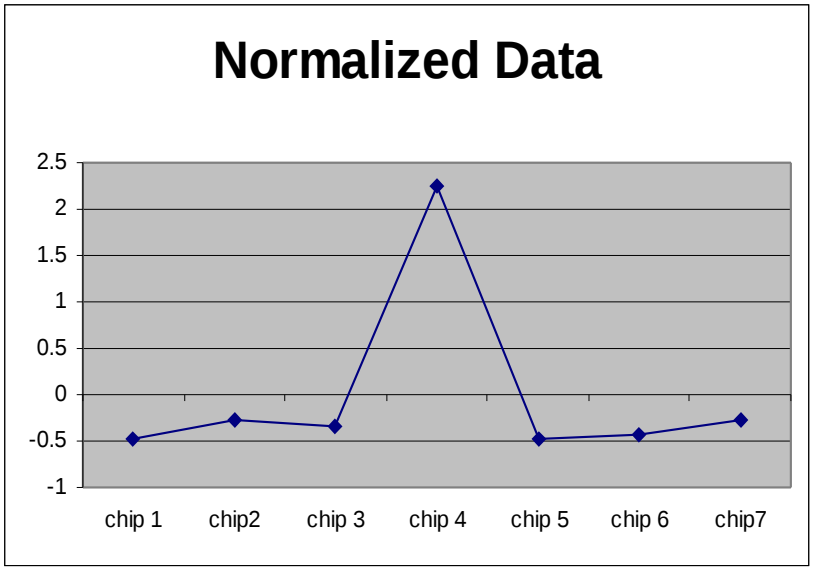
- Choose a distance metric
  - Pearson Correlation
  - Spearman Correlation
  - Euclidean Distance
  - Mutual Information
- Choose clustering algorithm
  - Hierarchical
  - Agglomerative
  - Principle Component Analysis
  - Super-paramagnetic and others

Normalization of Expression Profiles

$$x_i^{norm} = \frac{x_i - \bar{x}}{s}$$

	chip 1	chip2	chip 3	chip 4	chip 5	chip 6	chip7
Raw	30	35	33	100	30	31	35
Normalized	-0.46	-0.27	-0.35	2.25	-0.46	-0.42	-0.27





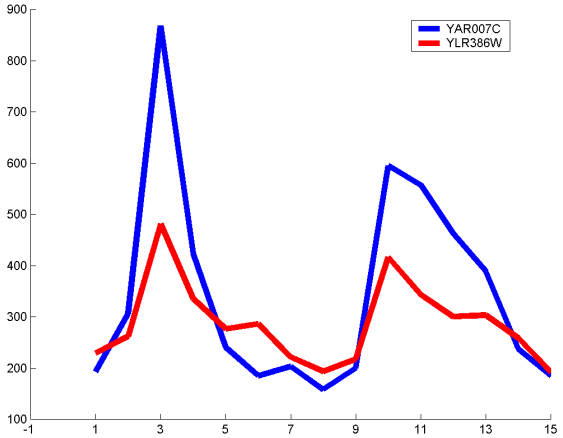
Pearson Correlation Coefficient

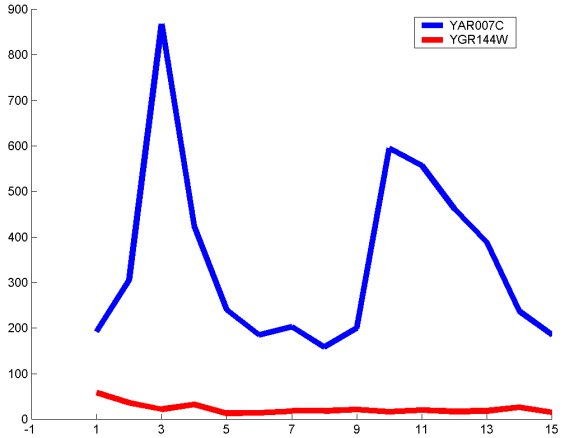
- Compares scaled profiles!
- Can detect inverse relationships
- Most commonly used
- Spearman rank correlation technically more correct

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

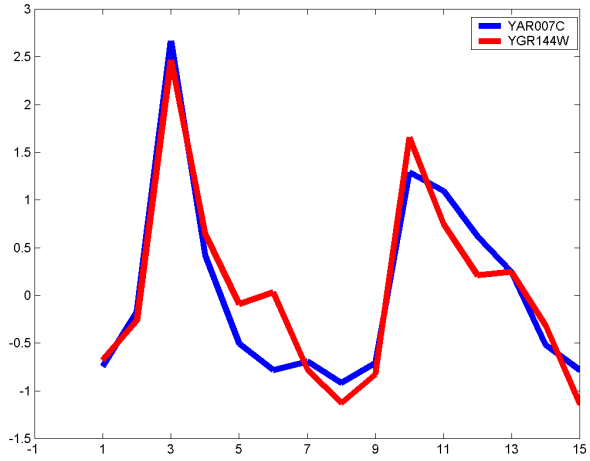
n=number of conditions
x=average expression of gene x in all n conditions
y=average expression of gene y in all n conditions
s_x=standard deviation of x
S_y=standard deviation of y

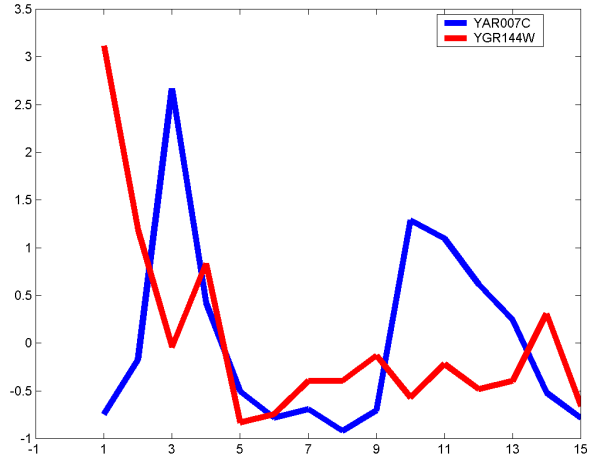
Correlation Examples





Raw Data



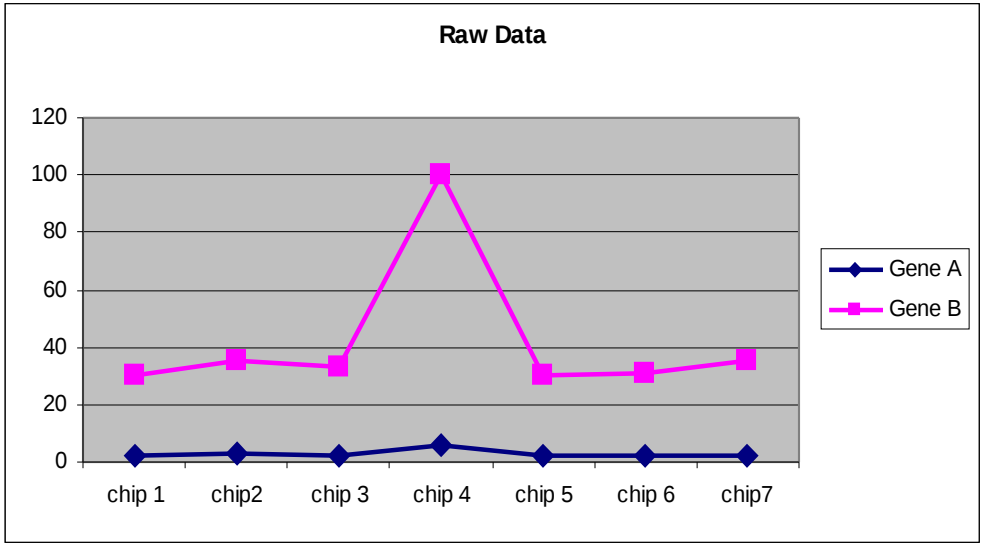


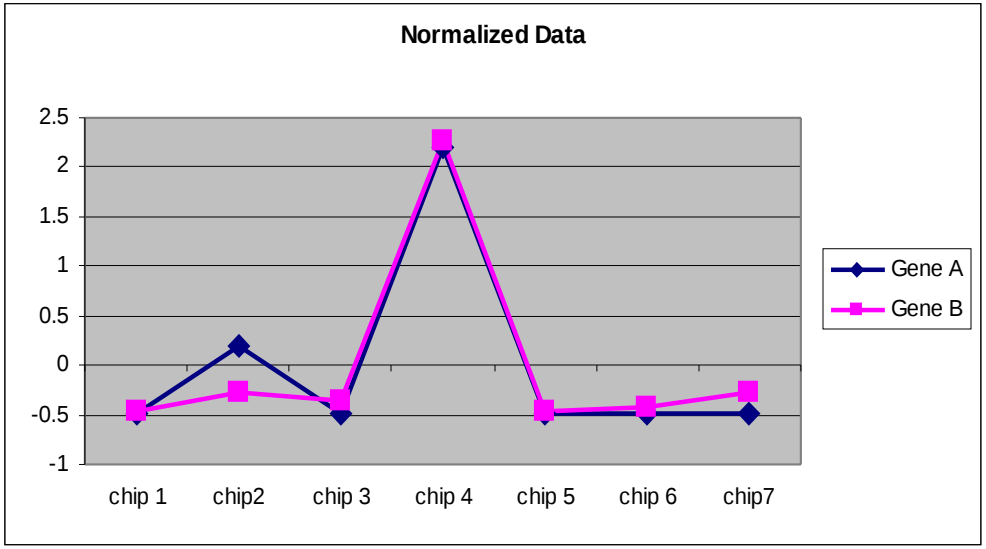
Normalized

Correlation = 0.94

Correlation = -0.087

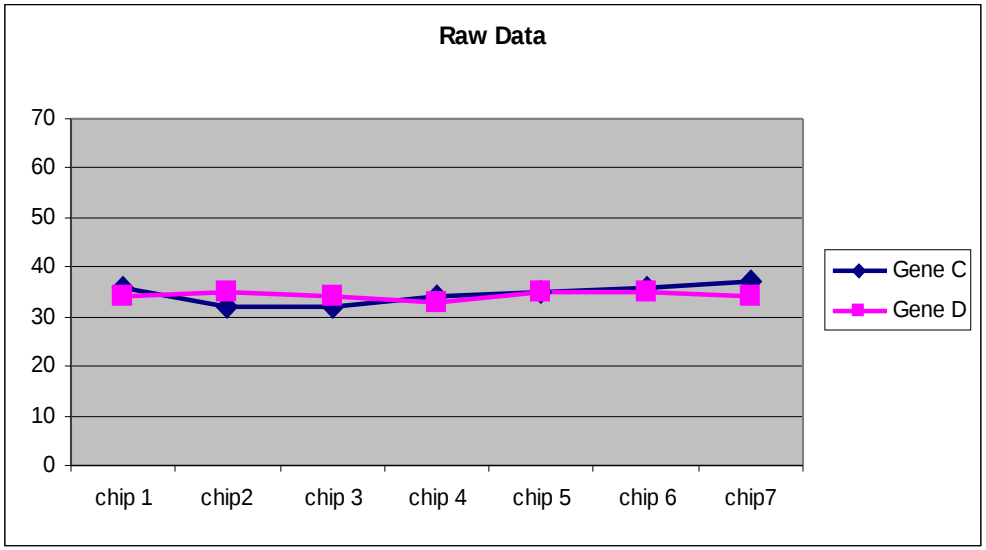
Correlation Pitfalls 1

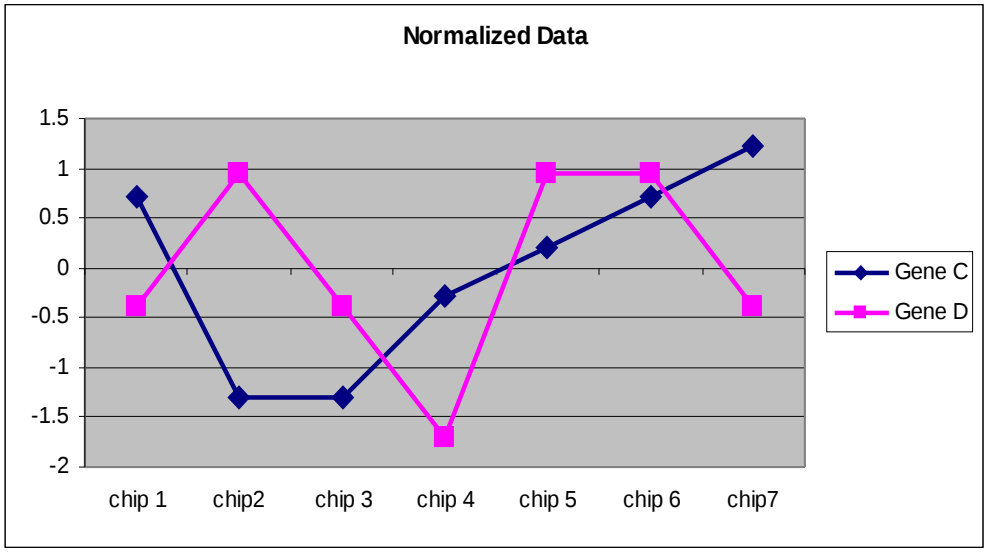




Correlation=0.97

Correlation Pitfalls 2





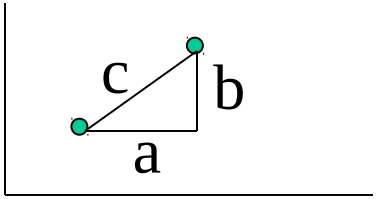
Correlation=-0.02

Avoid Pitfalls By Filtering The Data

- Remove Genes that do not reach some threshold level in at least one (or more) conditions
- For Affymetrix, remove genes whose stdev/mean ratio does not reach some threshold
- For spotted arrays, remove genes whose stdev does not reach some threshold

Euclidean Distance

- Based on Pythagoras
- Scaled versus unscaled
- Cannot detect inverse relation ships



$a^2 + b^2 = c^2$

For Gene $\vec{X}=(x_1, x_2,\ldots x_n)$ and Gene $\vec{Y}=(y_1, y_2,\ldots y_n)$

$$d(\vec{X},\vec{Y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots (x_n - y_n)^2}$$

Mutual Information

$$H(X) = -\sum_{i=1}^{I} P(X_i) \log P(X_i) \quad and \quad H(Y) = -\sum_{j=1}^{J} P(Y_j) \log P(Y_j)$$

$$H(X,Y) = -\sum_{i=1}^{I} \sum_{j=1}^{J} P(X_i, Y_j) \log P(X_i, Y_j)$$

$$M(X,Y) = H(X) + H(Y) - H(X,Y)$$

$$M(X,Y)_{norm} = M(X,Y) / \max[H(X), H(J)]$$

$$d(X,Y) = 1 - M(X,Y)_{norm}$$

- Must bin expression values (discreet).
- Can detect inverse and non-linear relationships (for example $y=x^3$)

Mutual Information Example

$\vec{X}$=(1,1,9,9,9,9,4,4)

$\vec{Y}$=(1,1,3,3,3,3,6,6)

$$H(X) = -\sum_{i=1}^{I} P(X_i)\log P(X_i) \quad and \quad H(Y) = -\sum_{j=1}^{J} P(Y_j)\log P(Y_j)$$

H(X)=H(Y)= -(.25log.25 + .5log.5 + .25log.25) = 1.5

$$H(X,Y) = -\sum_{i=1}^{I}\sum_{j=1}^{J} P(X_i,Y_j)\log P(X_i,Y_j)$$

H(X,Y)= -(.25log.25 + .5log.5 + .25log.25) = 1.5

$$M(X,Y) = H(X) + H(Y) - H(X,Y)$$

M(X,Y)= 1.5 + 1.5 – 1.5 =1.5

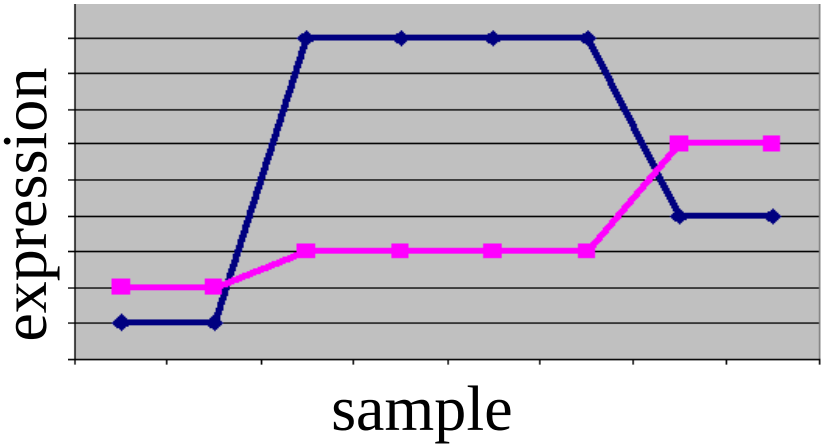
$$M(X,Y)_{norm} = M(X,Y) / \max[H(X), H(Y)]$$

$M(X,Y)_{norm}$= 1.5/1.5 = 1

$$d(X,Y) = 1 - M(X,Y)_{norm}$$

d(X,Y) = 1-1 = 0

Pearson correlation = -0.26

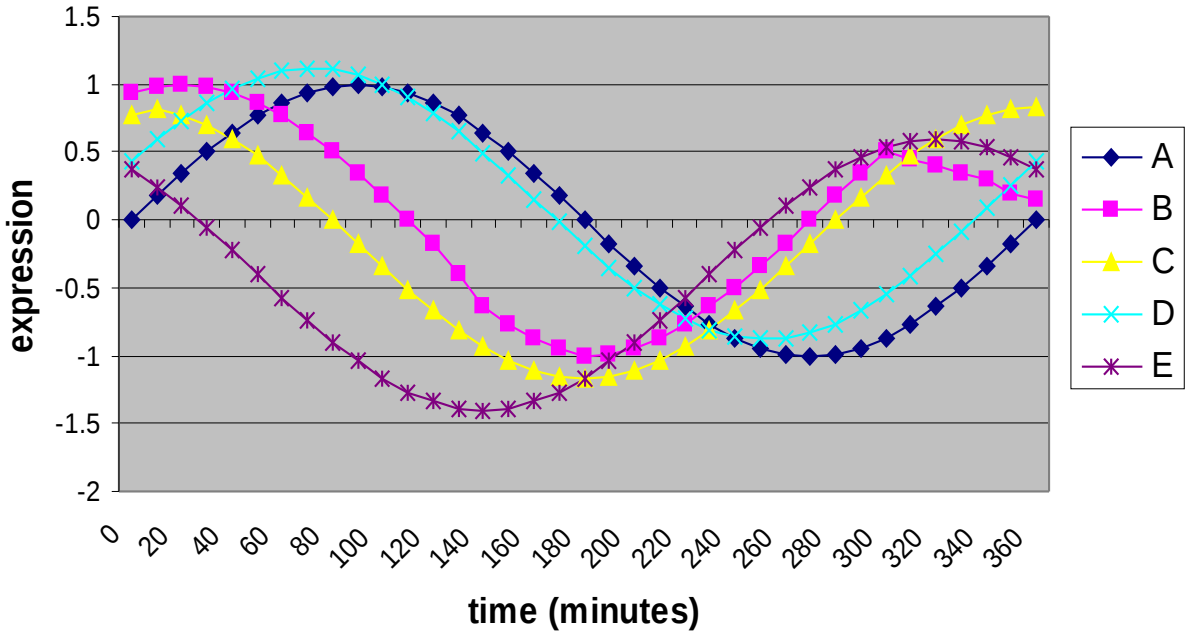


Clustering Algorithms

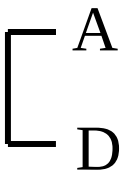
- Hierarchical
- Agglomerative
- Neural Networks
- Principal Component Analysis
- And many others

Clustering: Example 1, Step 1

Algorithm: Hierarchical, Distance Metric: Correlation

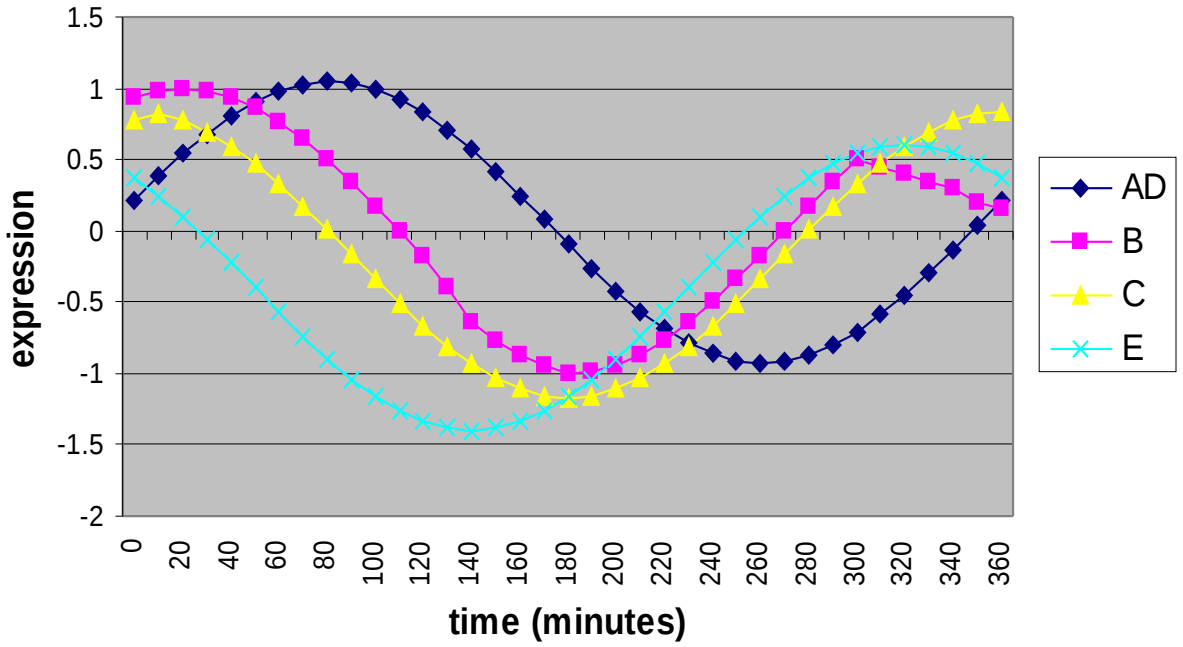


	A	B	C	D	E
A	-	0.23	0.00	0.95	-0.63
B	-	-	0.91	0.56	0.56
C	-	-	-	0.32	0.77
D	-	-	-	-	-0.36
E	-	-	-	-	-



Clustering: Example 1, Step 2

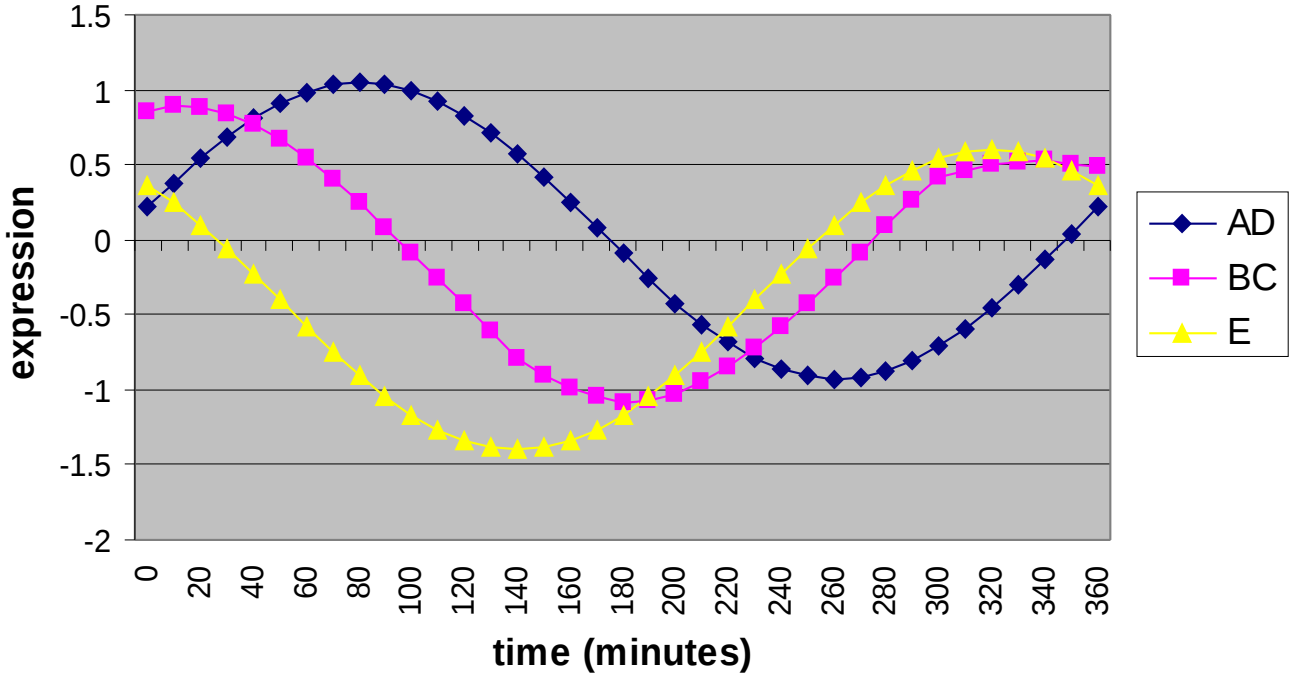
Algorithm: Hierarchical, Distance Metric: Correlation



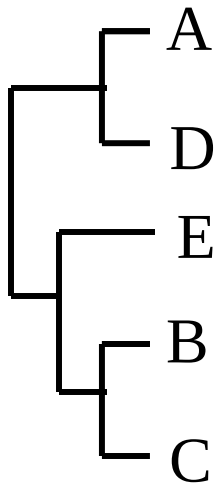
	AD	B	C	E
AD	-	0.37	0.16	-0.52
B	-	-	0.91	0.56
C	-	-	-	0.77
E	-	-	-	-

Clustering: Example 1, Step 3

Algorithm: Hierarchical, Distance Metric: Correlation

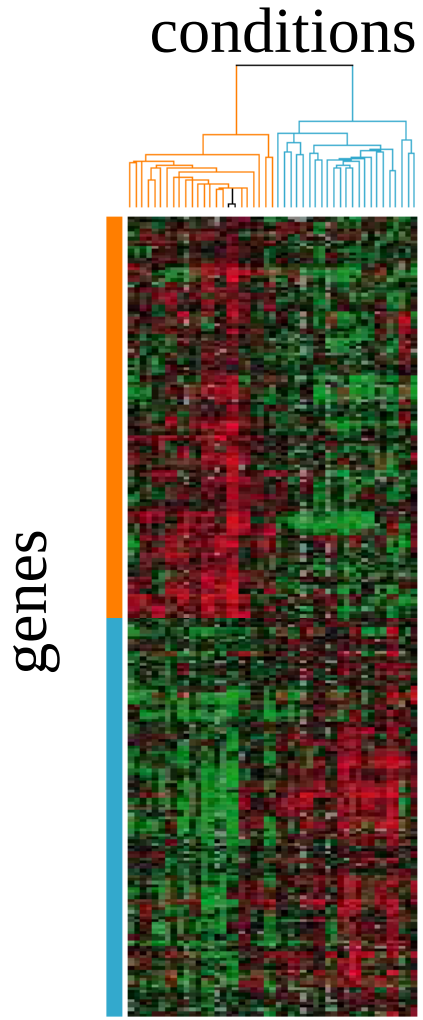


	AD	BC	E
AD	-	0.27	-0.52
BC	-	-	0.68
E	-	-	-



Tree View

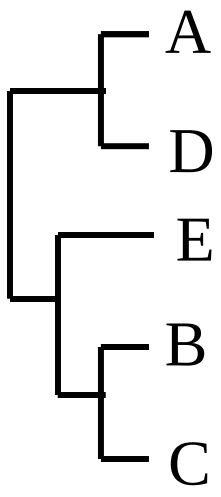
Eisen et al. (1998) PNAS 95: 14863-14868



Hierarchical Clustering Summary

Advantages

- Easy
- Very Visual
- Flexible (mean, median, etc.)

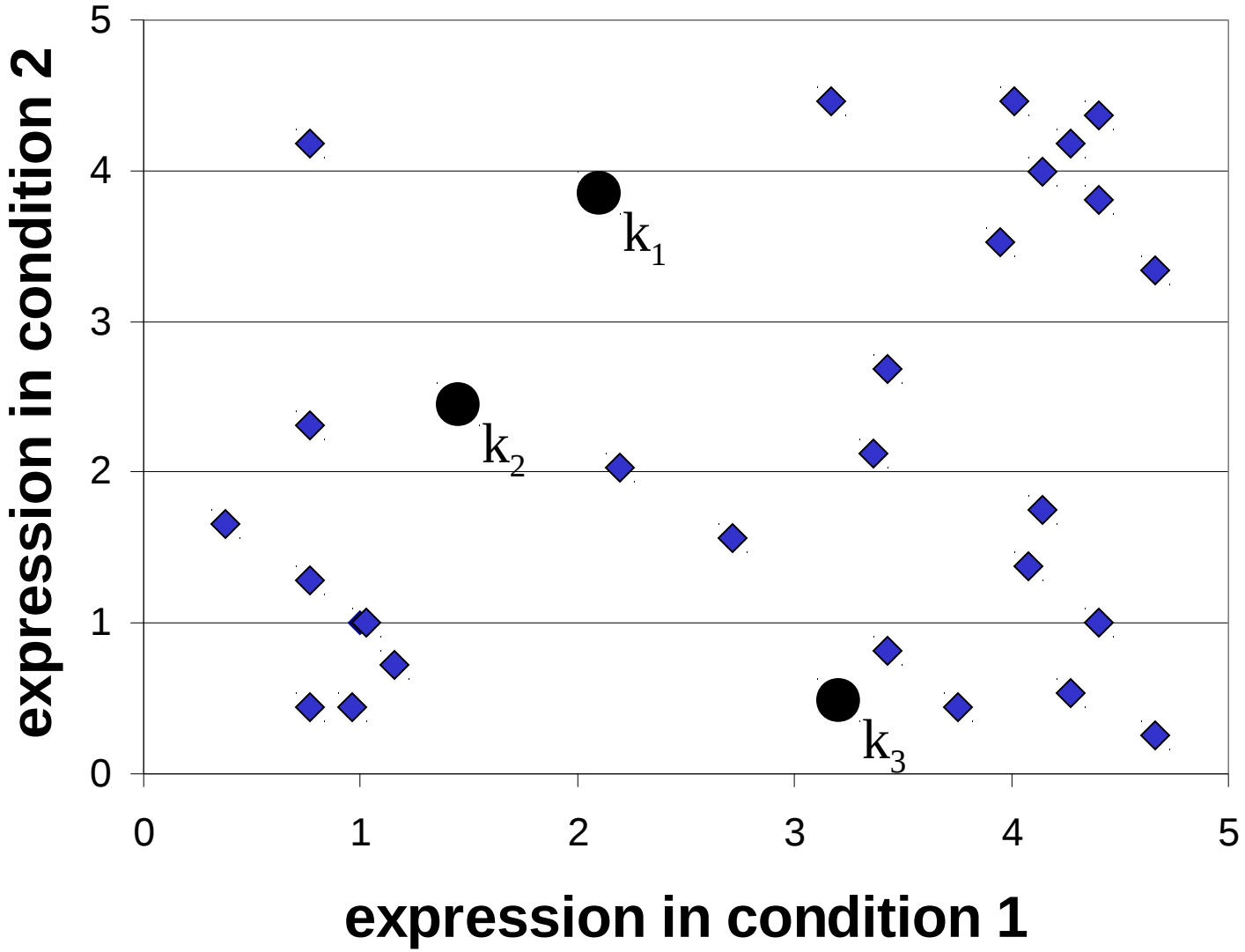


Disadvantages

- Unrelated Genes Are Eventually Joined
- Hard To Define Clusters
- Manual Interpretation Often Required

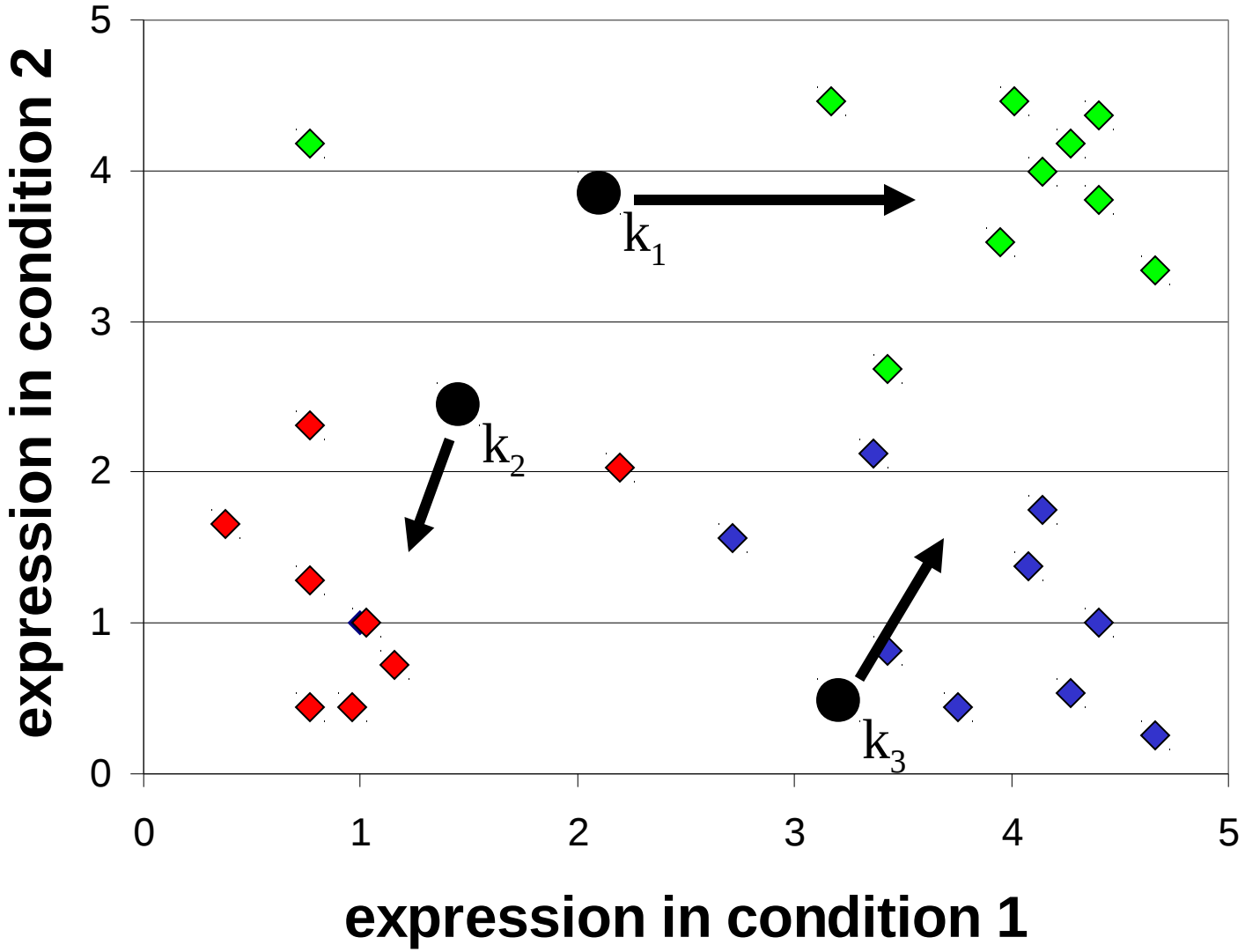
Clustering: Example 2, Step 1

Algorithm: k-means, Distance Metric: Euclidean Distance



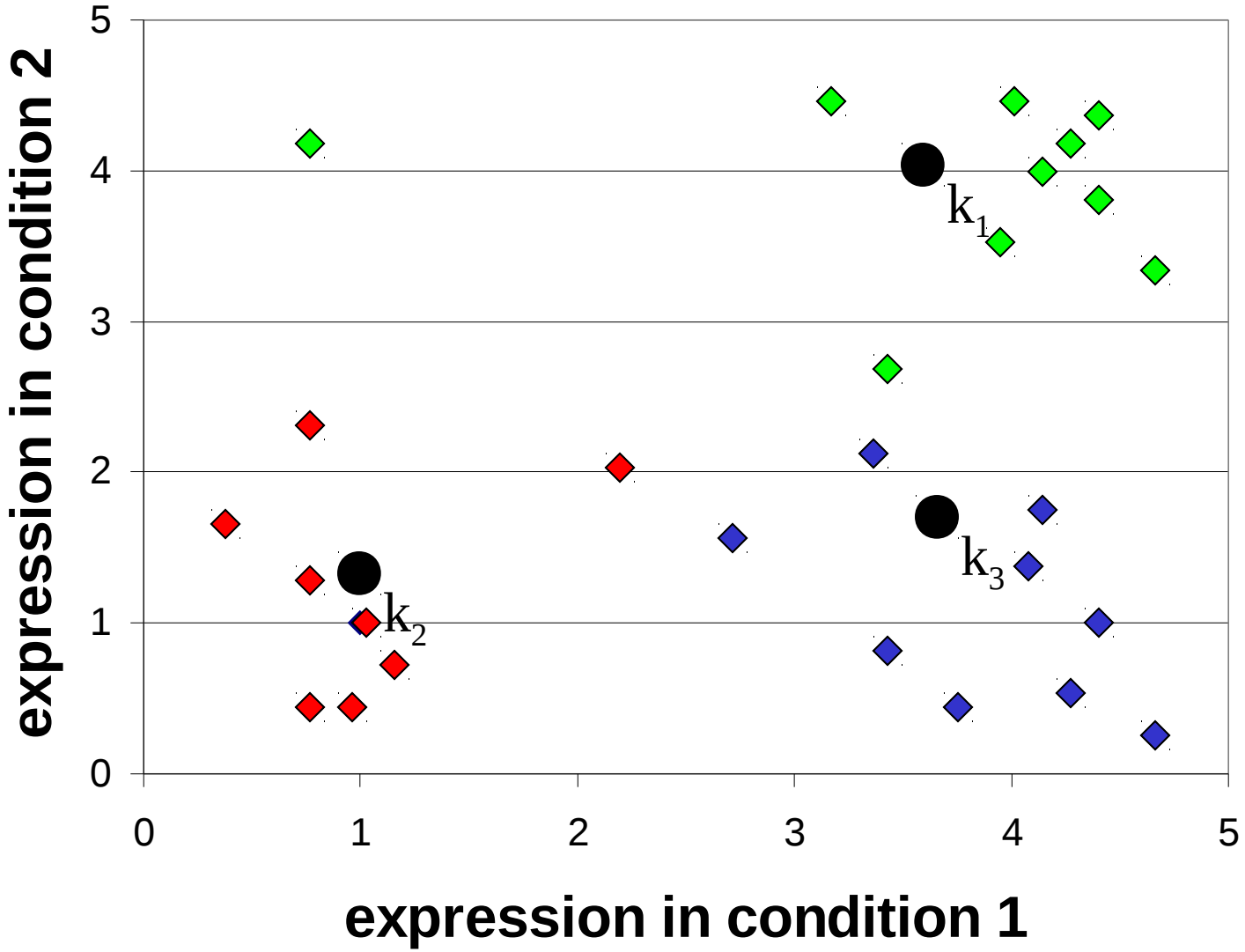
Clustering: Example 2, Step 2

Algorithm: k-means, Distance Metric: Euclidean Distance



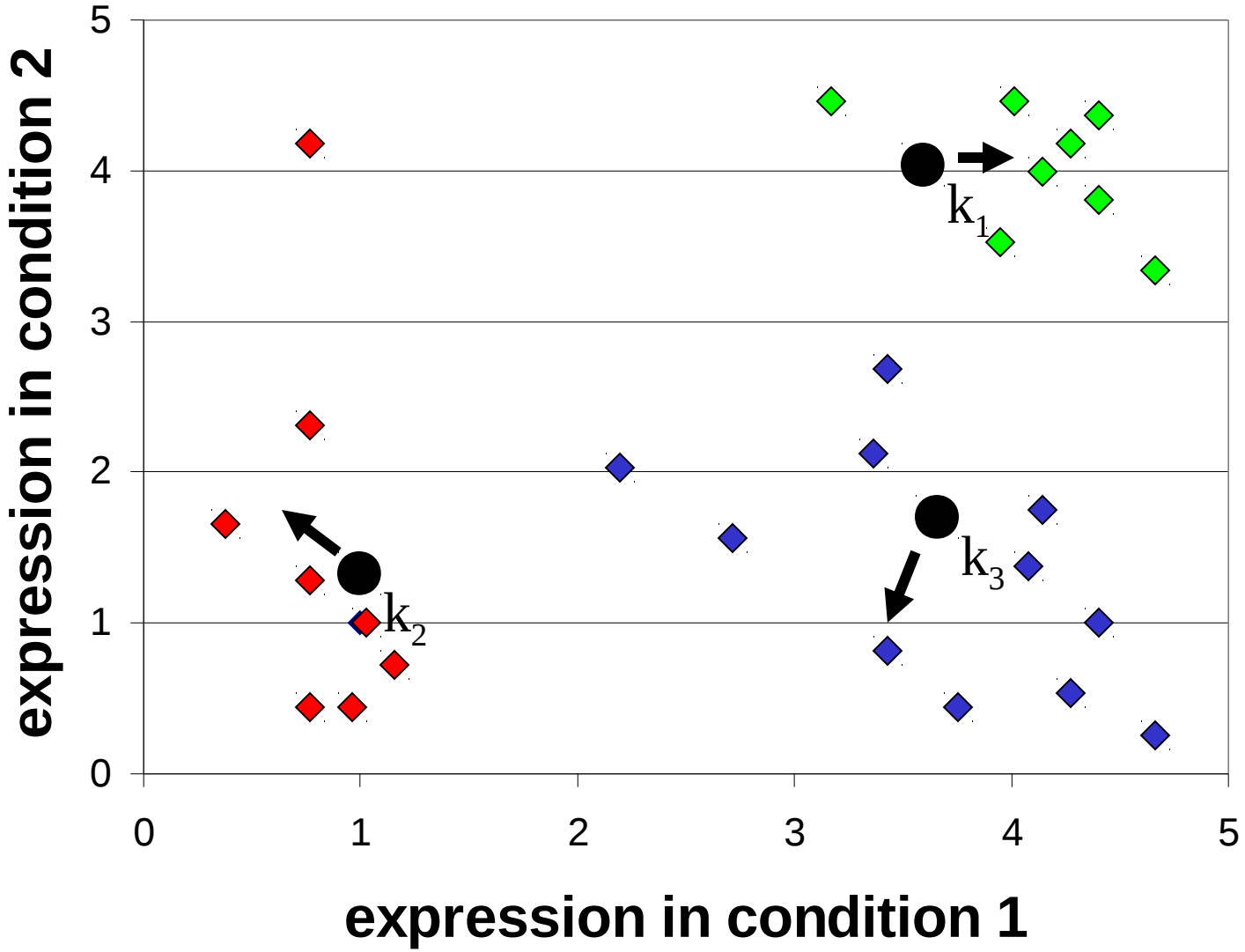
Clustering: Example 2, Step 3

Algorithm: k-means, Distance Metric: Euclidean Distance



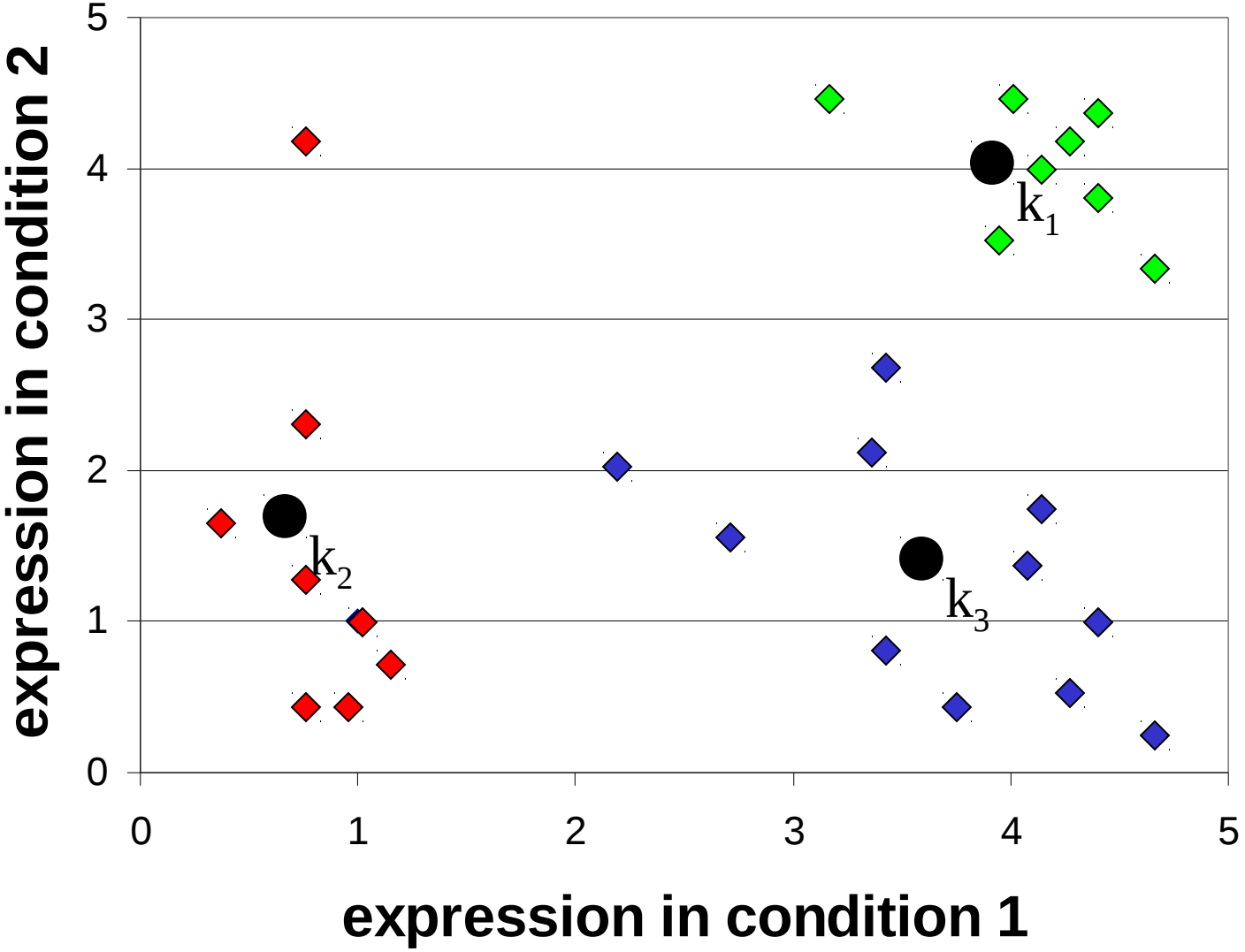
Clustering: Example 2, Step 4

Algorithm: k-means, Distance Metric: Euclidean Distance



Clustering: Example 2, Step 5

Algorithm: k-means, Distance Metric: Euclidean Distance



K-means algorithm

1) Pick a number (k) of cluster centers
2) Assign every gene to its nearest cluster center
3) Move each cluster center to the mean of its assigned genes
4) Repeat 2-3 until convergence

K-means clustering summary

Advantages

- Genes automatically assigned to clusters
- Can vary starting locations of cluster centers to determine initial condition dependence

Disadvantages

- Must pick number of clusters before hand
- All genes forced into a cluster

Which Clustering Method Should I Use?

- What is the biological question?
- Do I have a preconceived notion of how many clusters there should be?
- How strict do I want to be? Spilt or Join?
- Can a gene be in multiple clusters?
- Hard or soft boundaries between clusters

Keep in Mind.

- Clustering is NOT an analysis in itself.
- Clustering cannot NOT work.