Lecture Slides for Tuesday March 31st

11:05 AM EDT by Zoom

https://mit.zoom.us/j/348659452

For audio you can use your computer or call:

US : +1 646 558 8656 or +1 669 900 6833

Meeting ID: 348 659 452

International Numbers:
https://mit.zoom.us/u/adLEbsadSS

Note: class will be recorded and posted for later viewing.

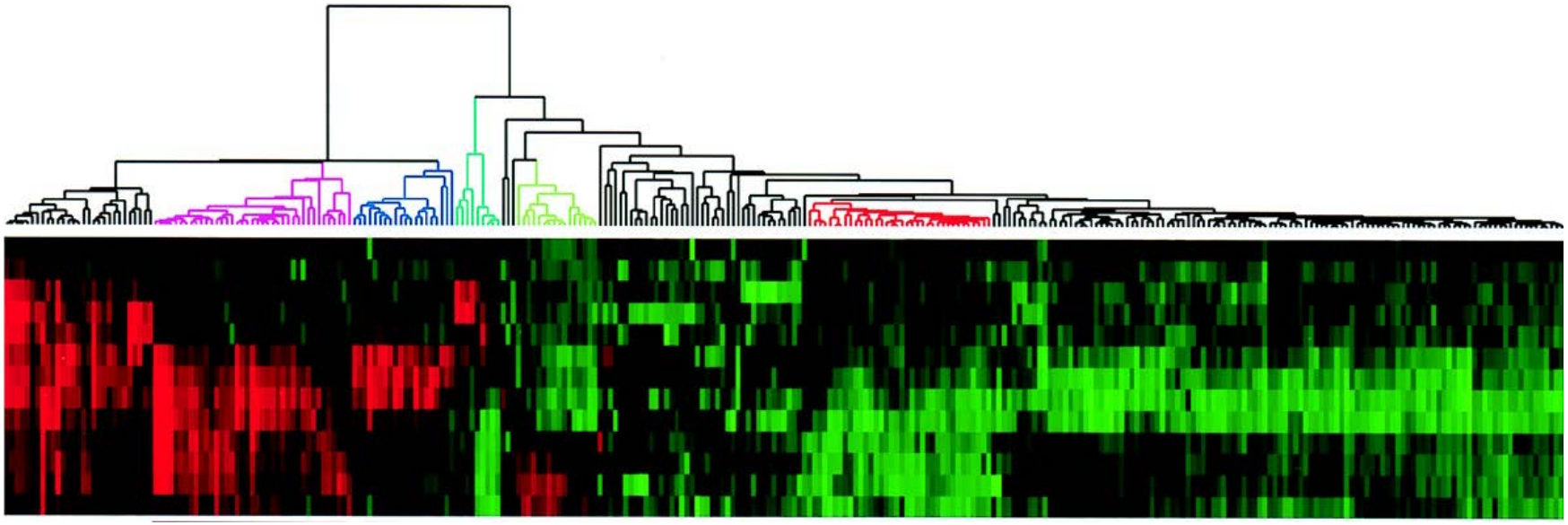
My Revised Lecture Schedule

Date	Topic
March 31st	Cluster, PCA
April 2nd	RNA-Seq
April 7th	Transcriptional Regulation

Learning Objectives

- Manually cluster small vectors using k-means clustering
- Describe the results of Principal Component Analysis (PCA)

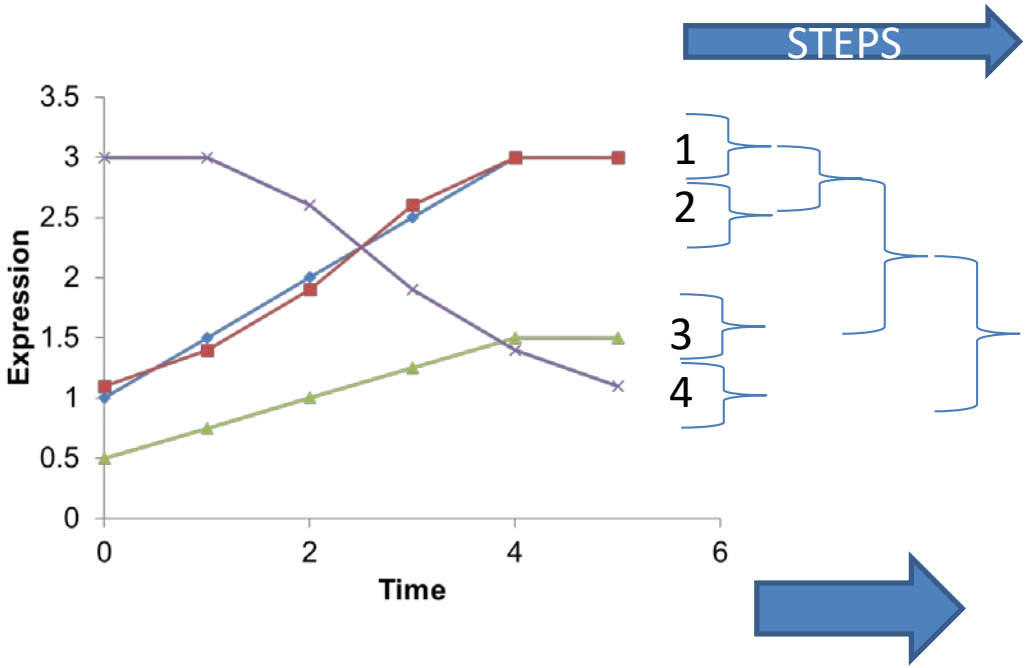
Lightning Review of Hierarchical Clustering



Two types of approaches: Agglomerative & Divisive

Agglomerative:

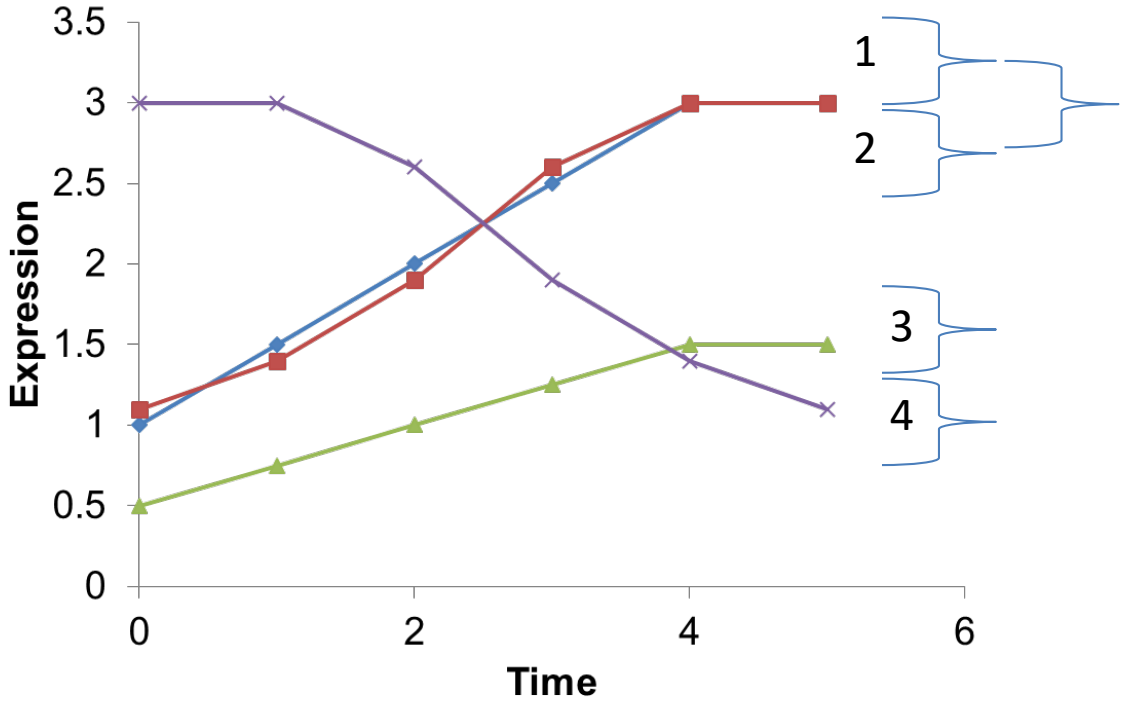
- Initialize: Each vector is in its own cluster
- Repeat until there is only one cluster:
  - Merge the two most similar clusters.



Step 1: each gene is its own cluster
Step 2: combine the two most similar genes
Step 3: find the two most similar clusters

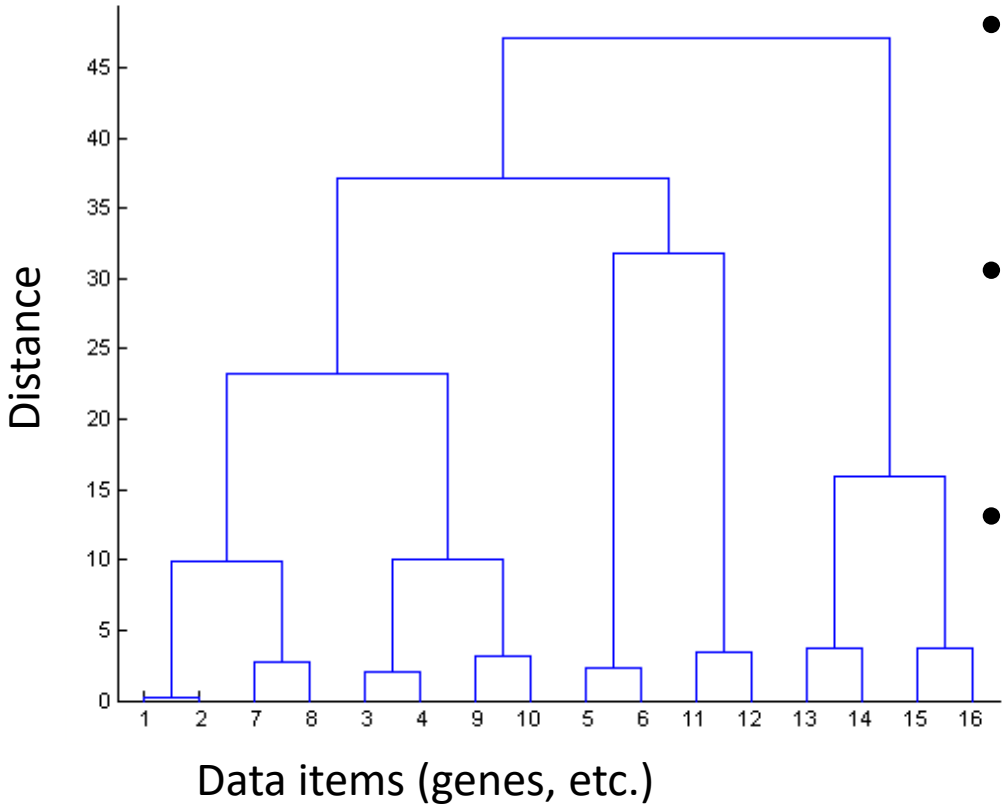
Several options:

minimum distance between members of cluster A,B
maximum distance between members of cluster A,B
average distance between members of cluster A,B



… but I have not told you how to compute distance between the two genes in the new cluster with individual genes

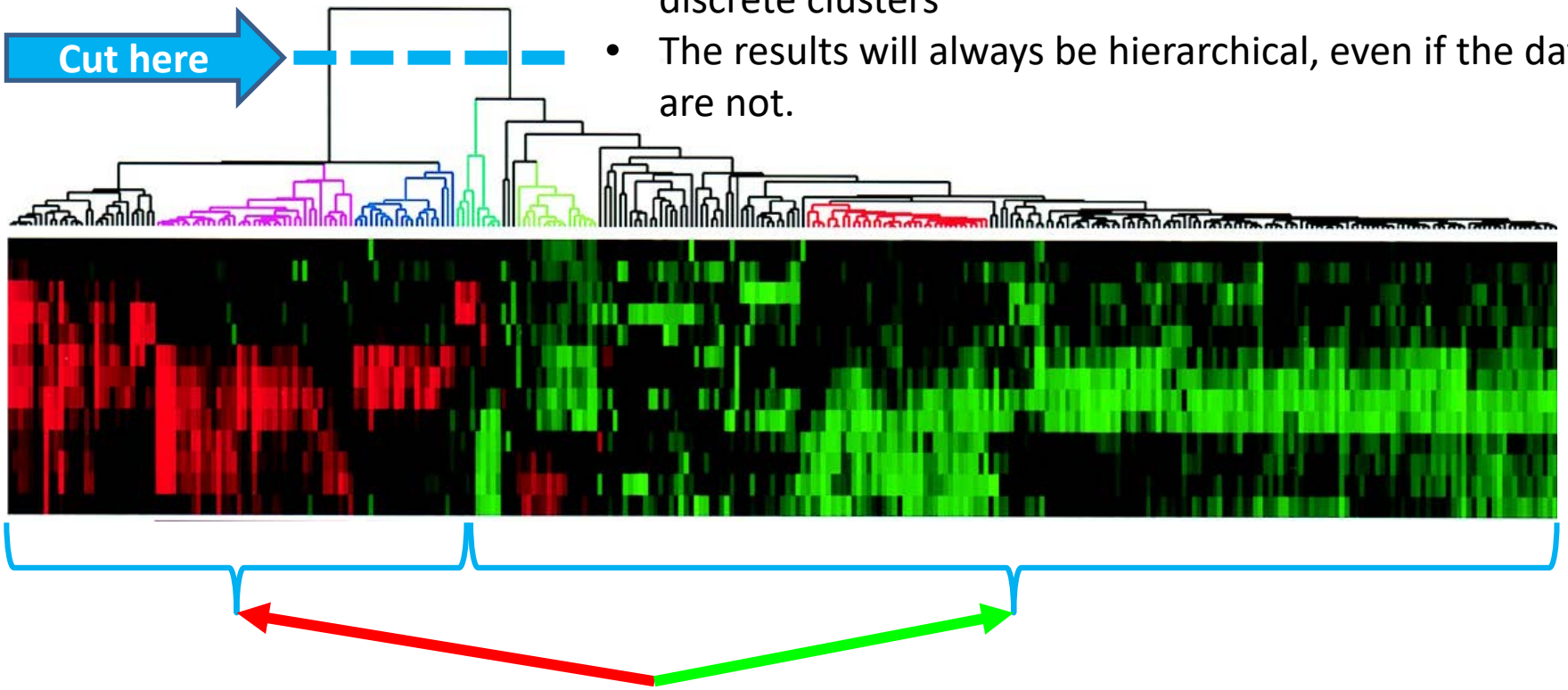
Dendrograms



- The final cluster is the root and each data item is a leaf
- The heights of the bars indicate how close the items are
- Can ‘slice’ the tree at any distance cutoff to produce discrete clusters
- The results will always be hierarchical, even if the data are not.
- The order of the leaf nodes is not meaningful

Hierarchical Clustering Links All Samples

- The heights of the bars indicate how close the items are
- Can ‘slice’ the tree at any distance cutoff to produce discrete clusters
- The results will always be hierarchical, even if the data are not.

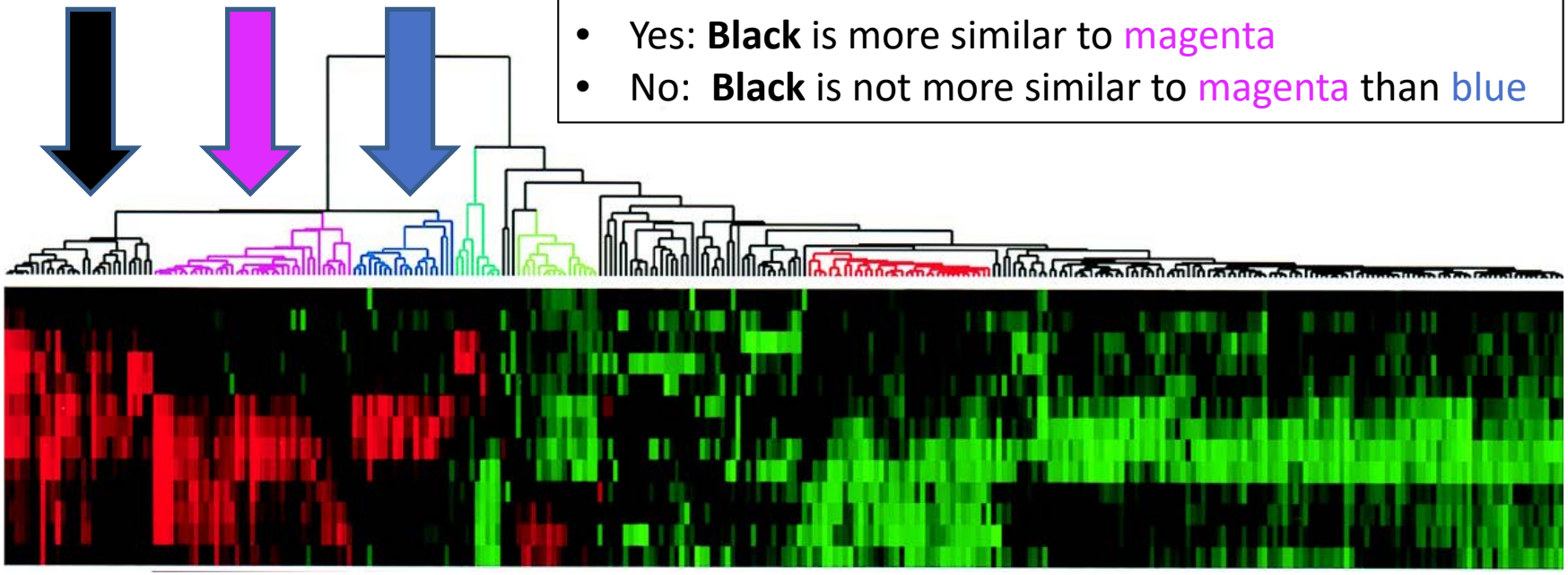


These genes are not very similar!

Quick Questions

Is the **black** cluster more similar to the magenta than the blue?

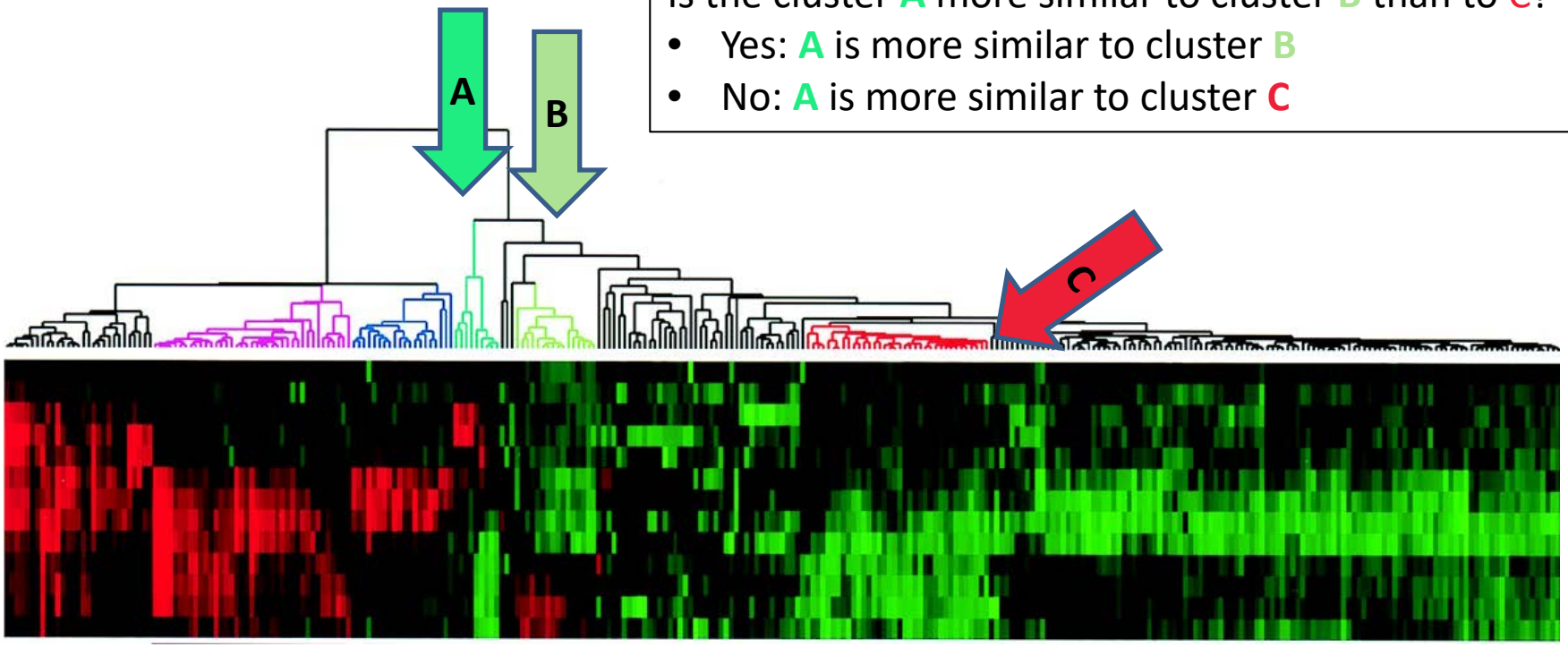
- Yes: **Black** is more similar to magenta
- No: **Black** is not more similar to magenta than blue



Quick Questions

Is the cluster A more similar to cluster B than to C?

- Yes: A is more similar to cluster B
- No: A is more similar to cluster C



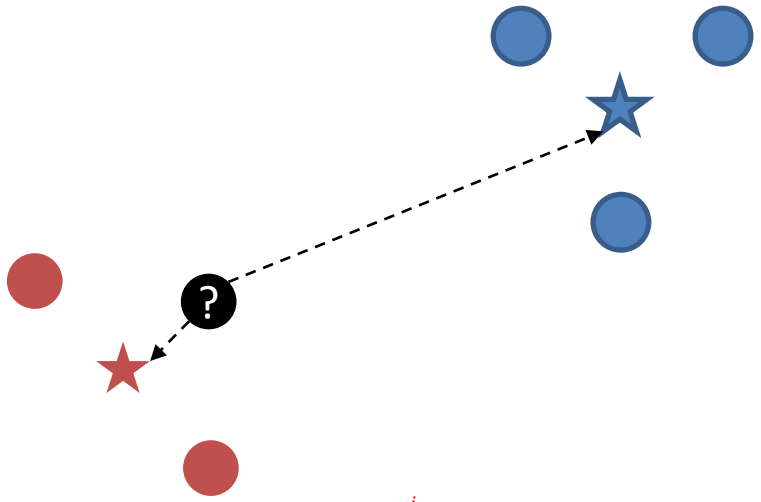
K-means clustering

- Advantage: gives sharp partitions of the data
- Disadvantage: need to specify the number of clusters (K).
- Goal: find a set of k clusters that minimizes the distances of each point in the cluster to the cluster mean:

How to cluster with K-means

K-means clustering algorithm

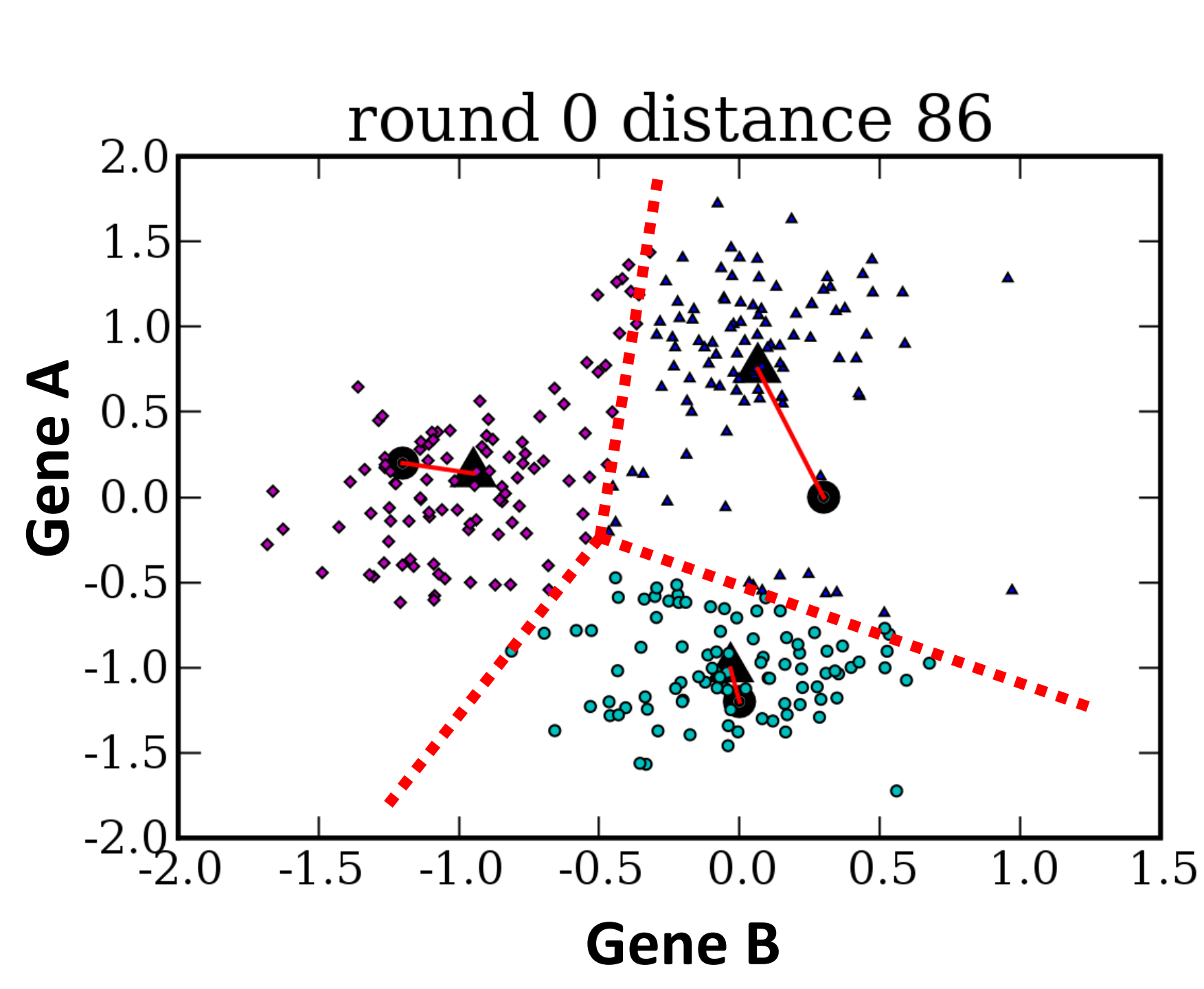
- Initialize: choose k points as cluster means
- Repeat until convergence:
  - Assignment: place each point X_i in the cluster with the closest mean.
  - Update: recalculate the mean for each cluster

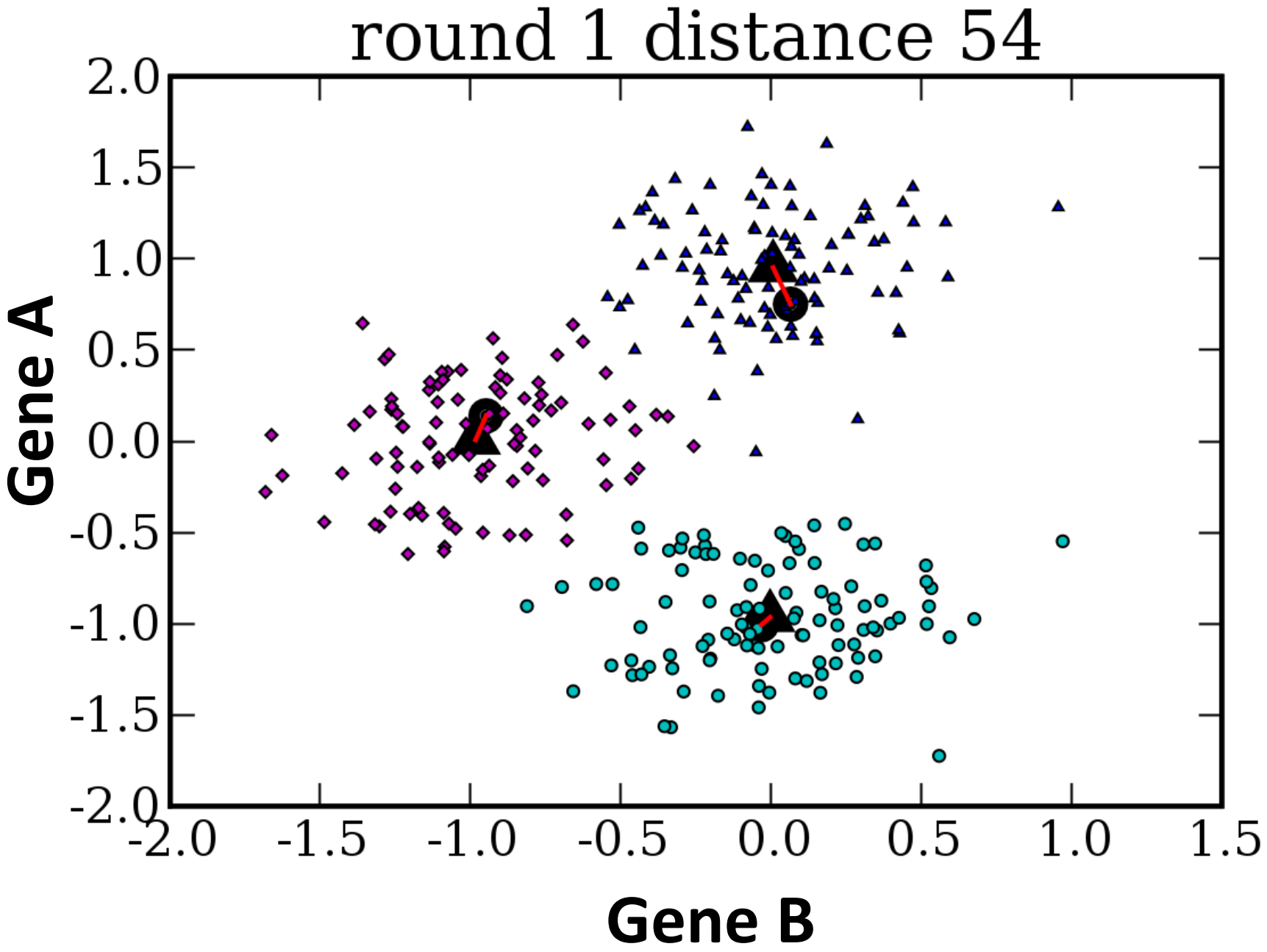


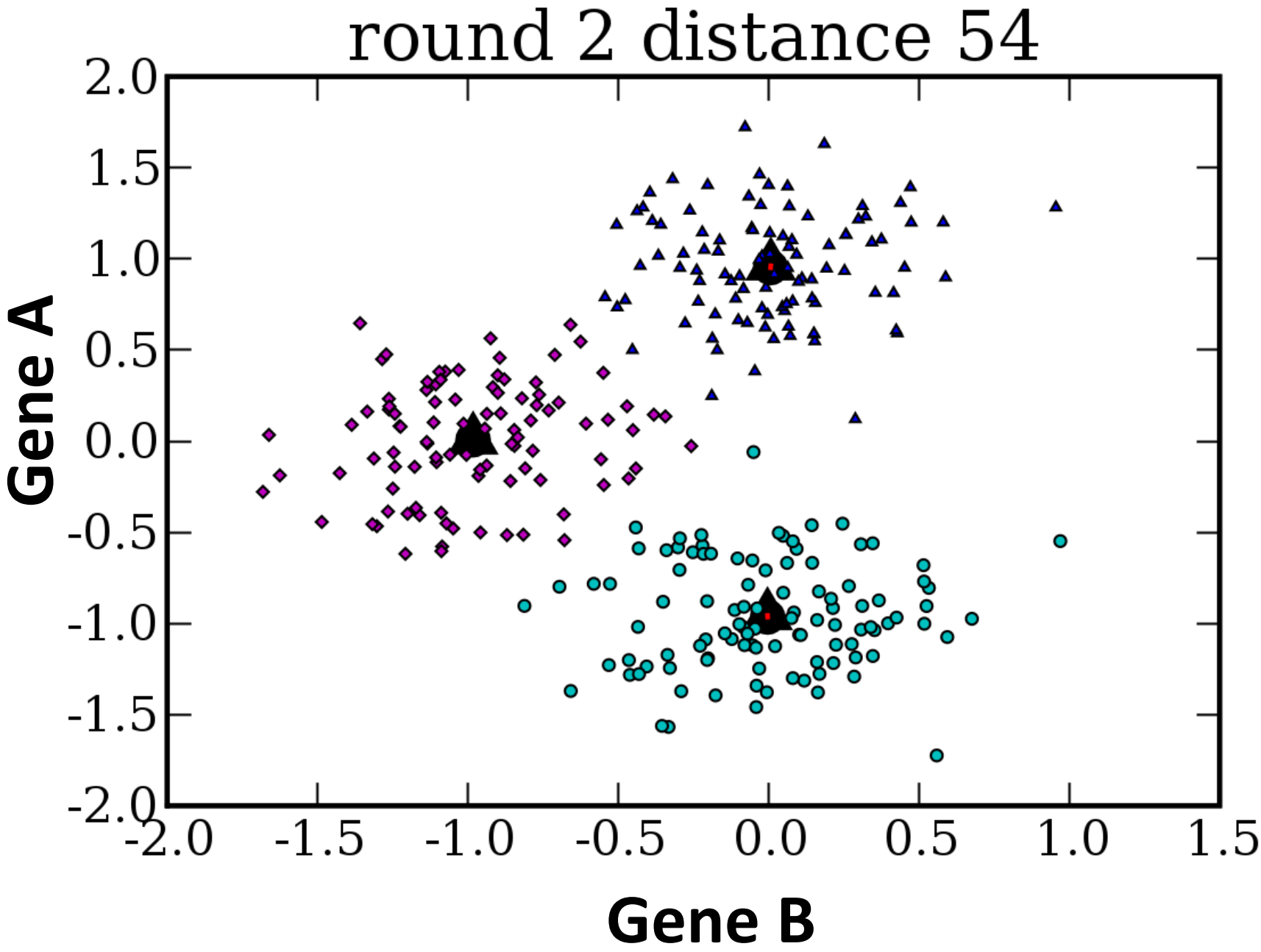
Mean=

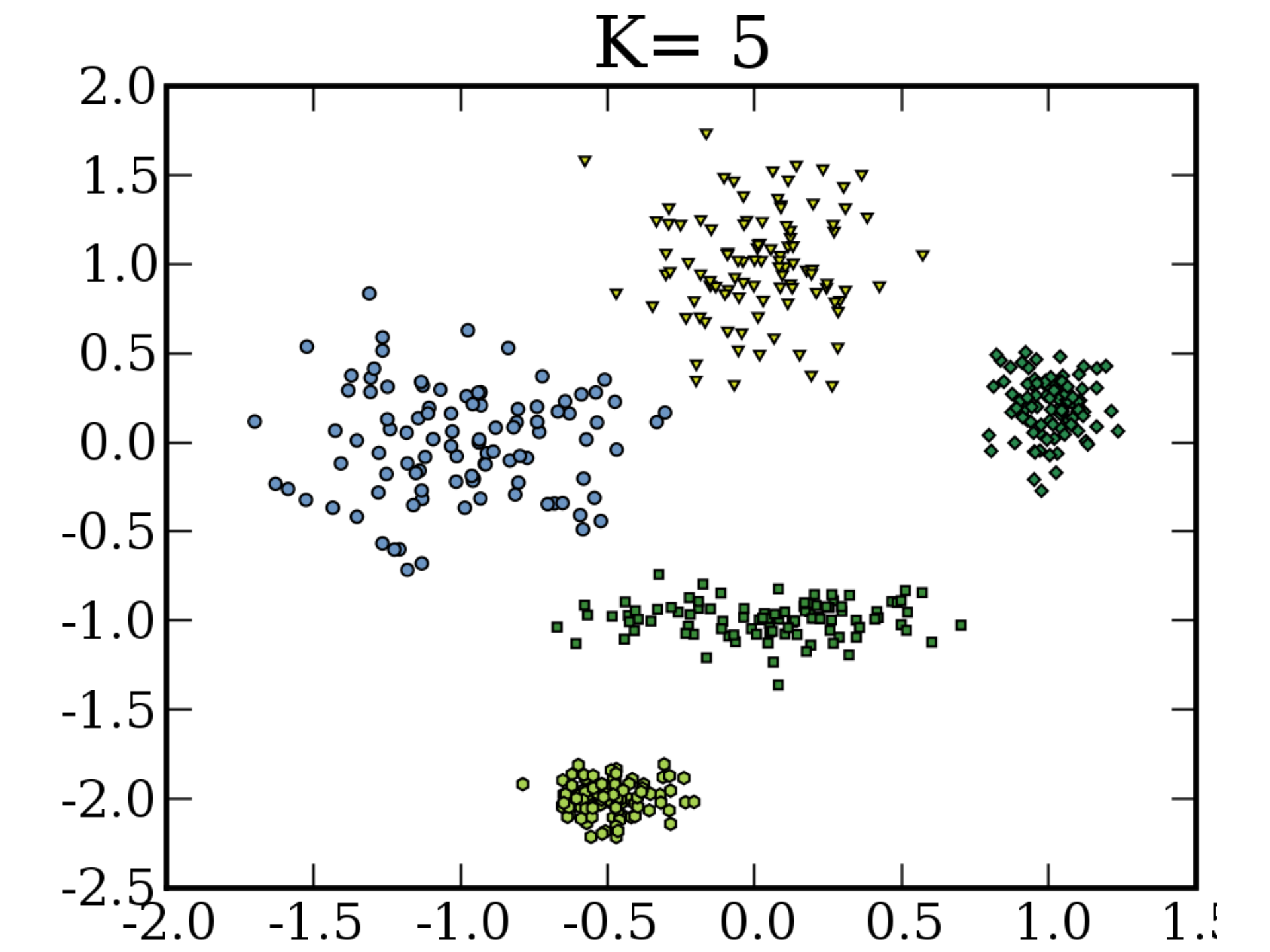
$$\text{centroid}_j = \hat{Y}_j = \frac{1}{N_{Y_j}} \sum_{i \in Y_j} X_i$$

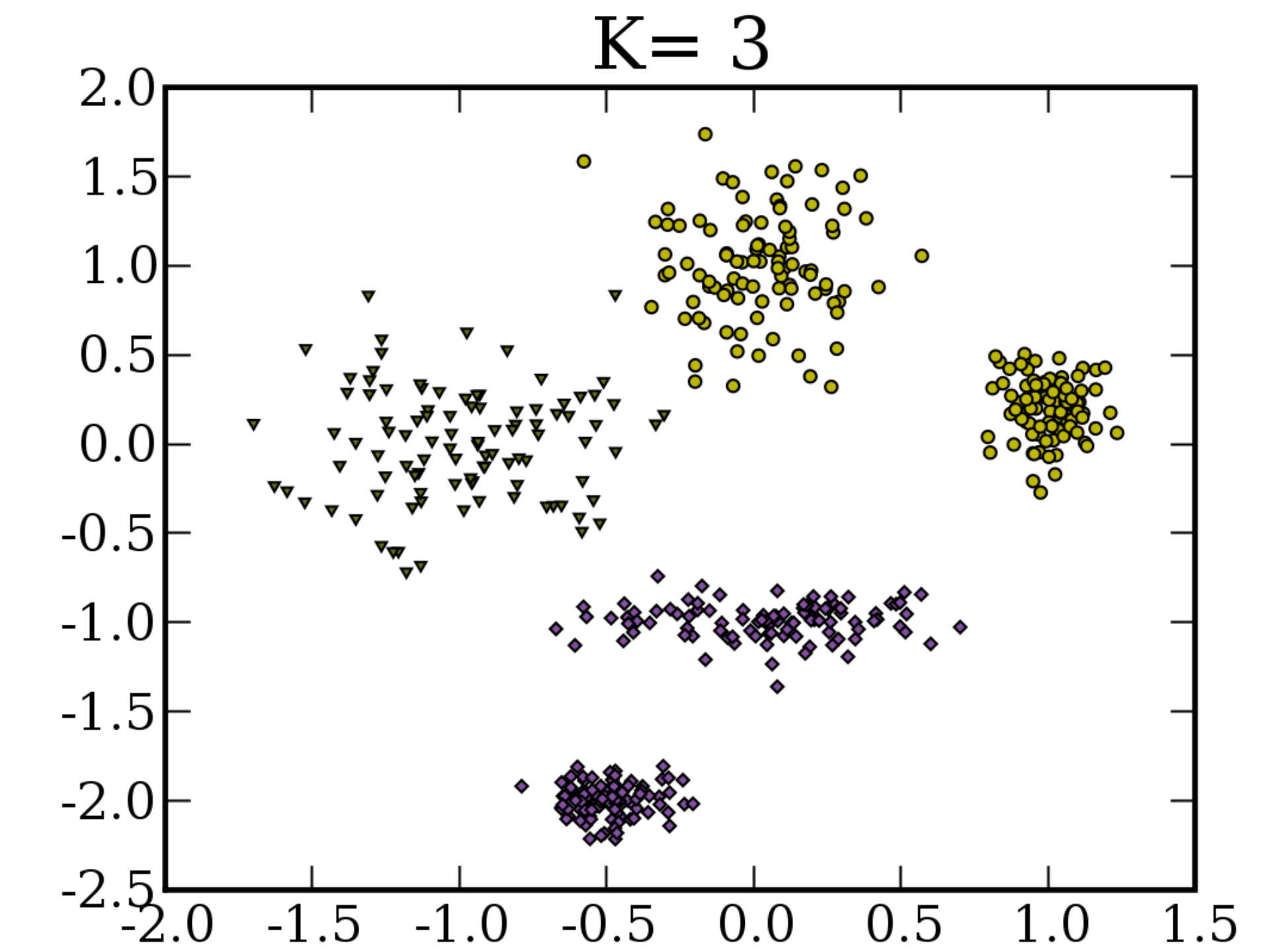
$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in C(i)} \left| X_j - \hat{Y}_i \right|^2$$

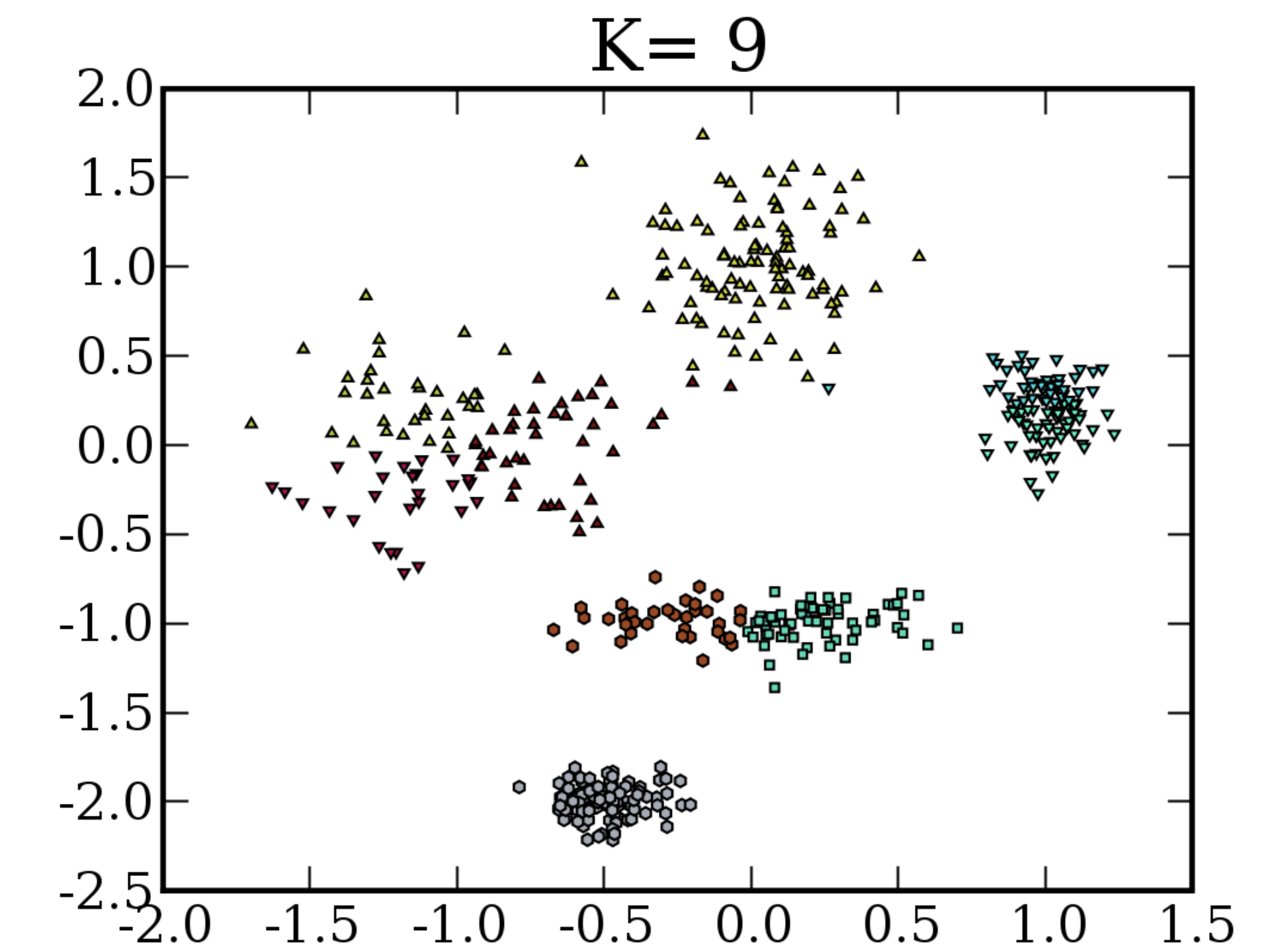

round 0 distance 86
Gene A
Gene B
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5


round 1 distance 54
Gene A
Gene B
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

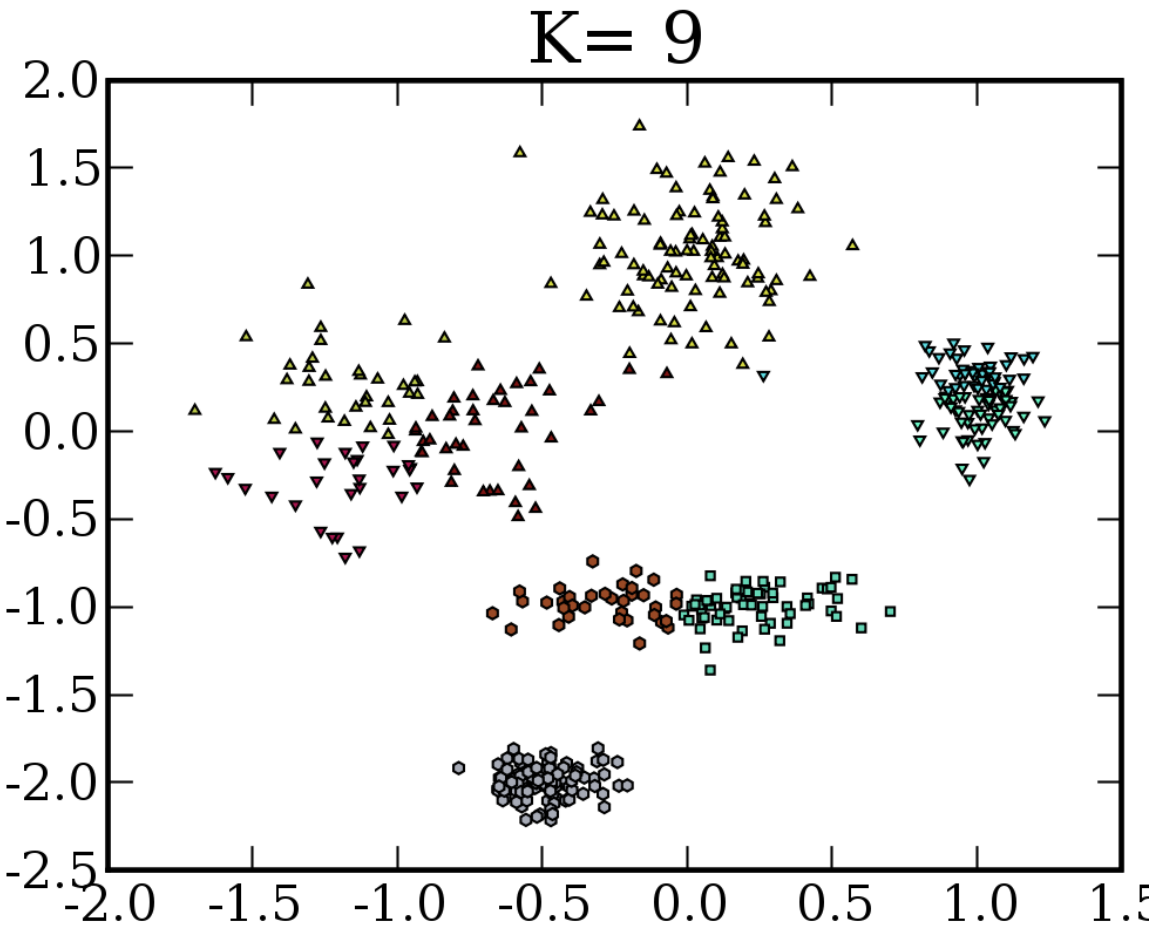

round 2 distance 54
Gene A
Gene B
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

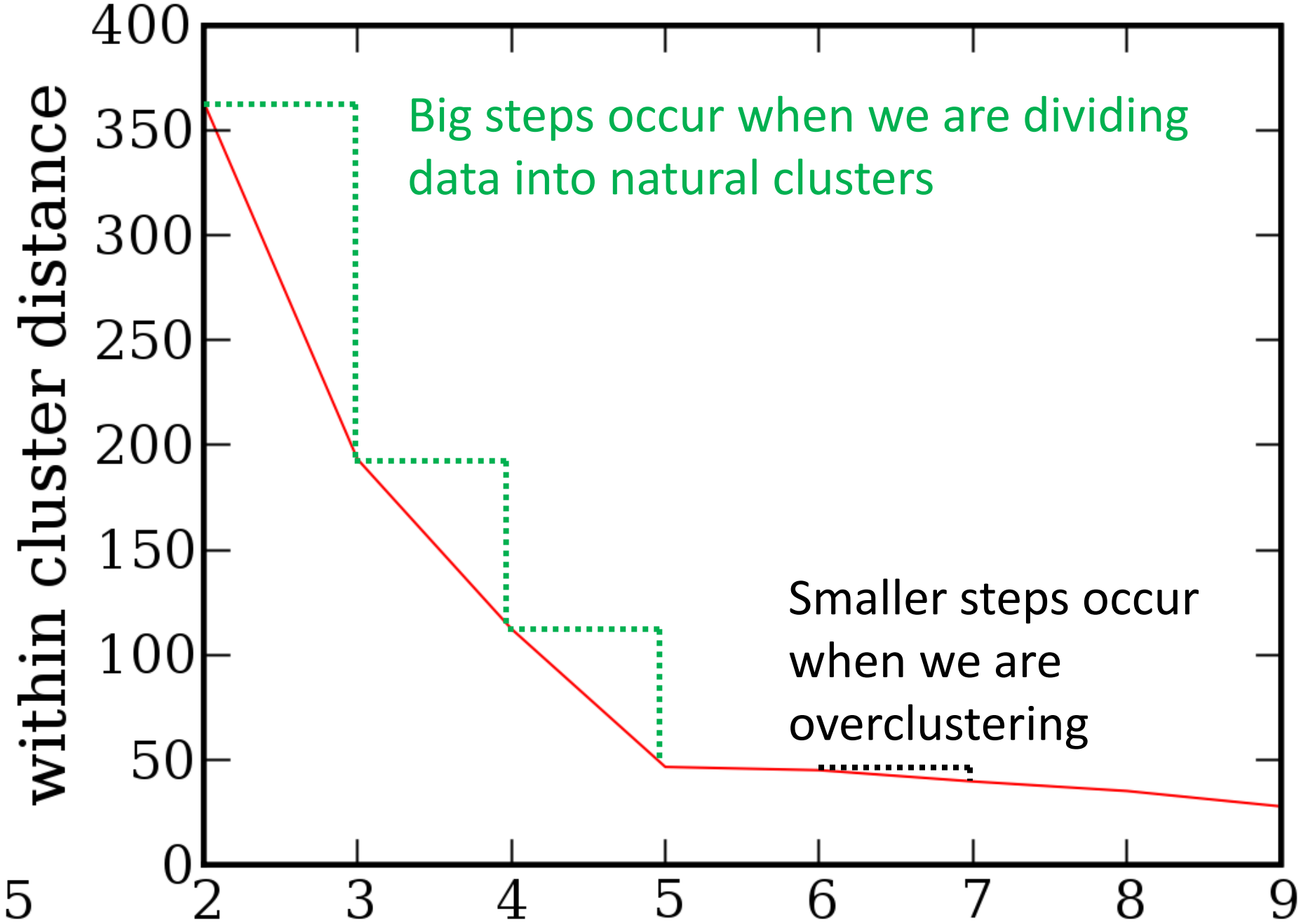

K= 5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0


K= 3
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

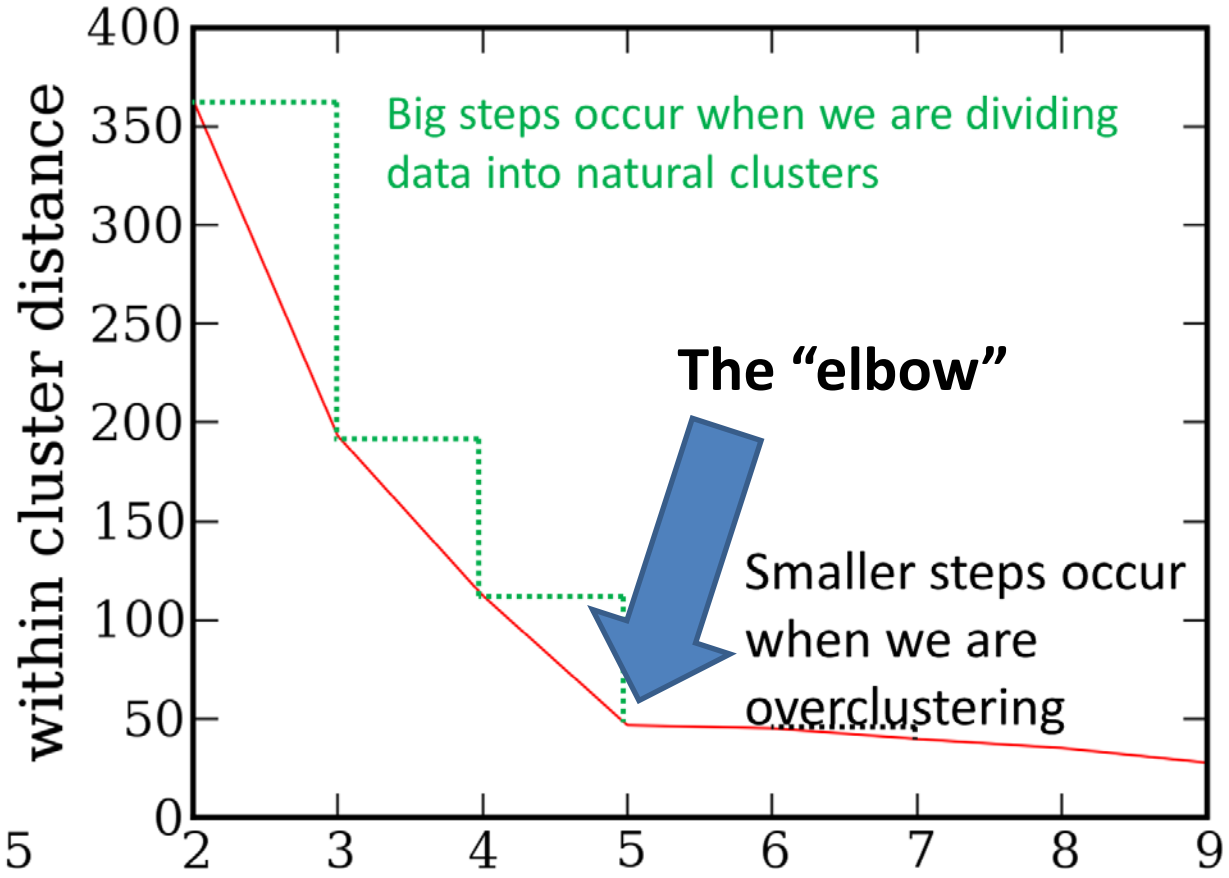

K= 9
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

So how do you choose K?




within cluster distance
400
350
300
250
200
150
100
50
0
2
3
4
5
6
7
8
9
Big steps occur when we are dividing data into natural clusters
Smaller steps occur when we are overclustering

This “elbow” plot can help find the right value of K

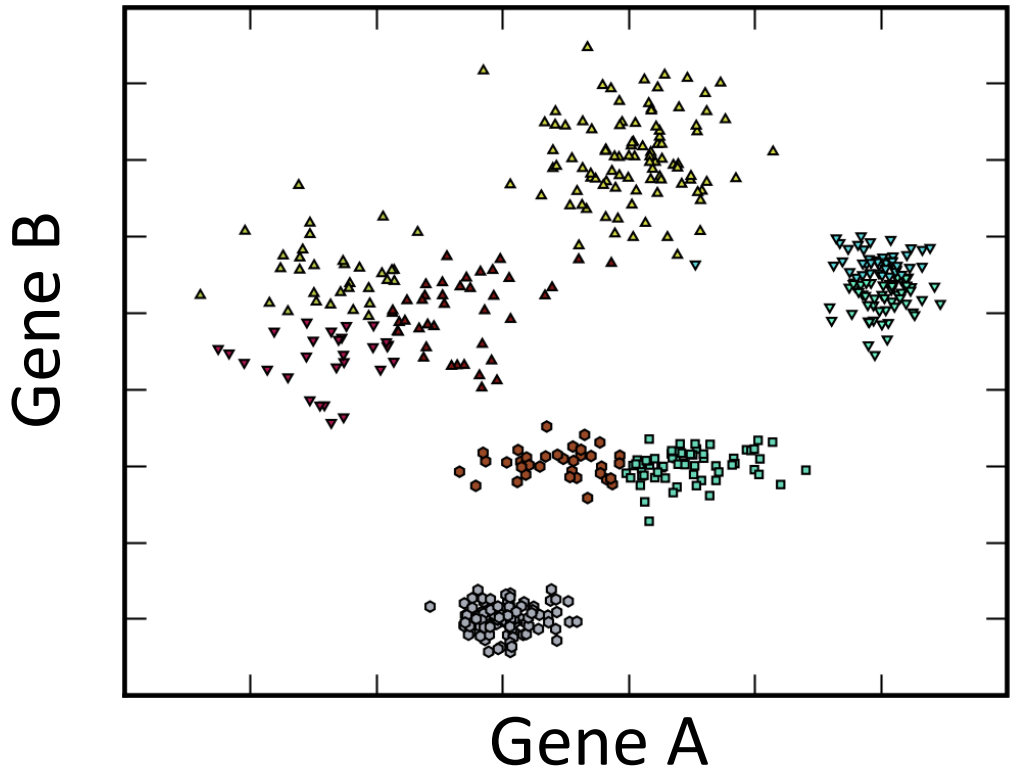


K-means clustering works with vectors of any length, but it's hard to visualize

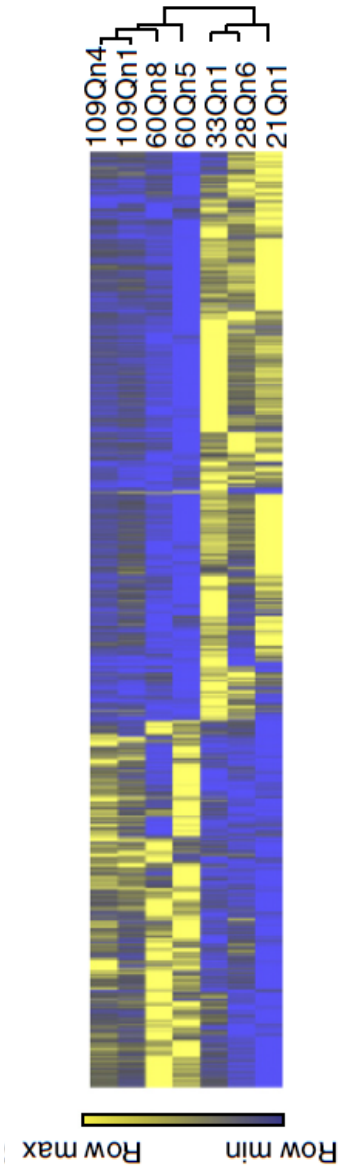
Dimension =

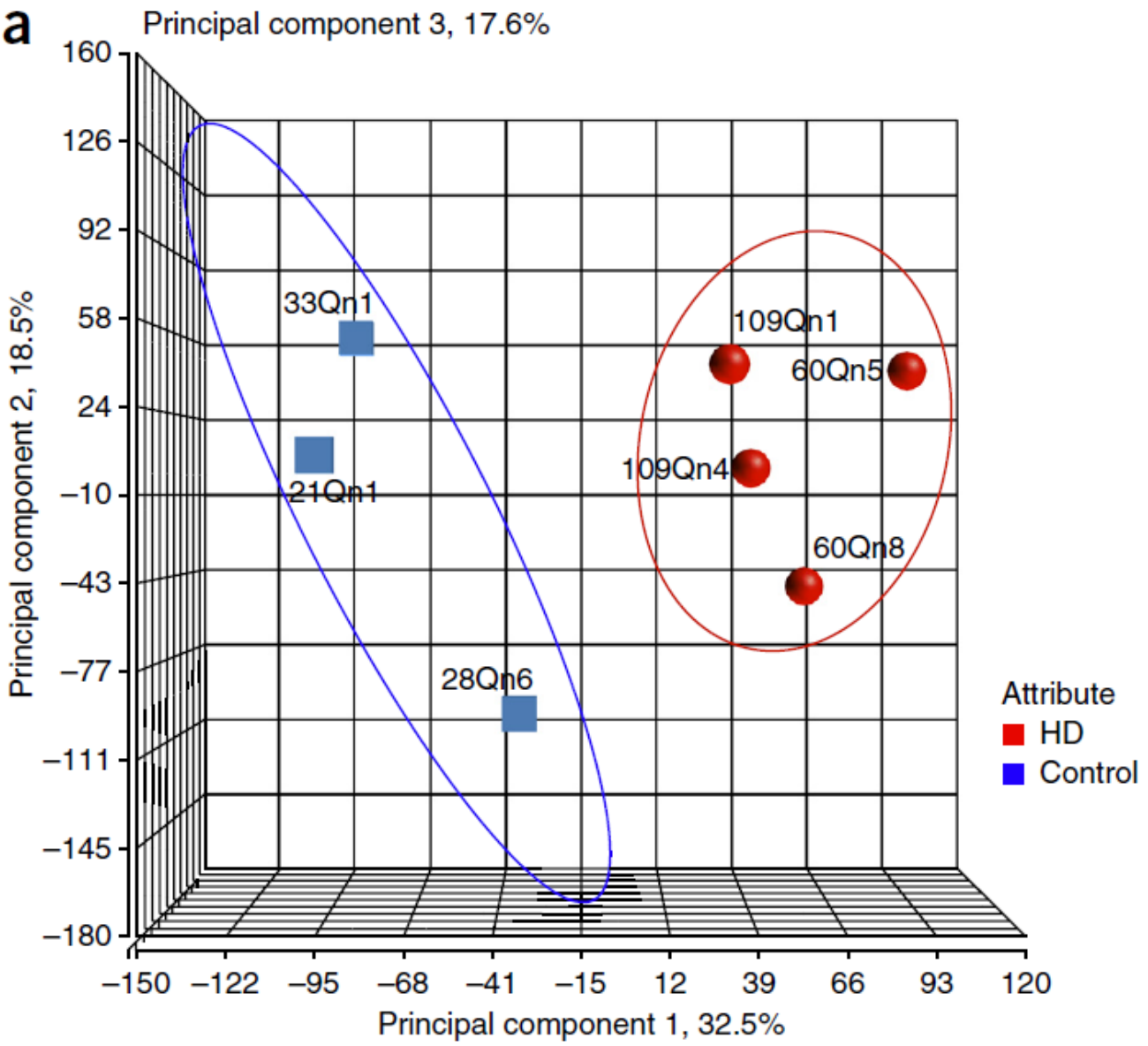
10 if I cluster genes by their expression in 10 conditions

20,000 if I cluster conditions based on the expression of each gene



The basics of PCA





Principal Component Analysis

- Each sample is currently described by the expression of roughly 20,000 genes.
- Our goal:
  to find a 2-D or 3-D way to present the data that captures the greatest variance
  - Obviously, I could select any two genes, but they might be the wrong ones.
  - Can we find “interesting” linear combinations of genes?

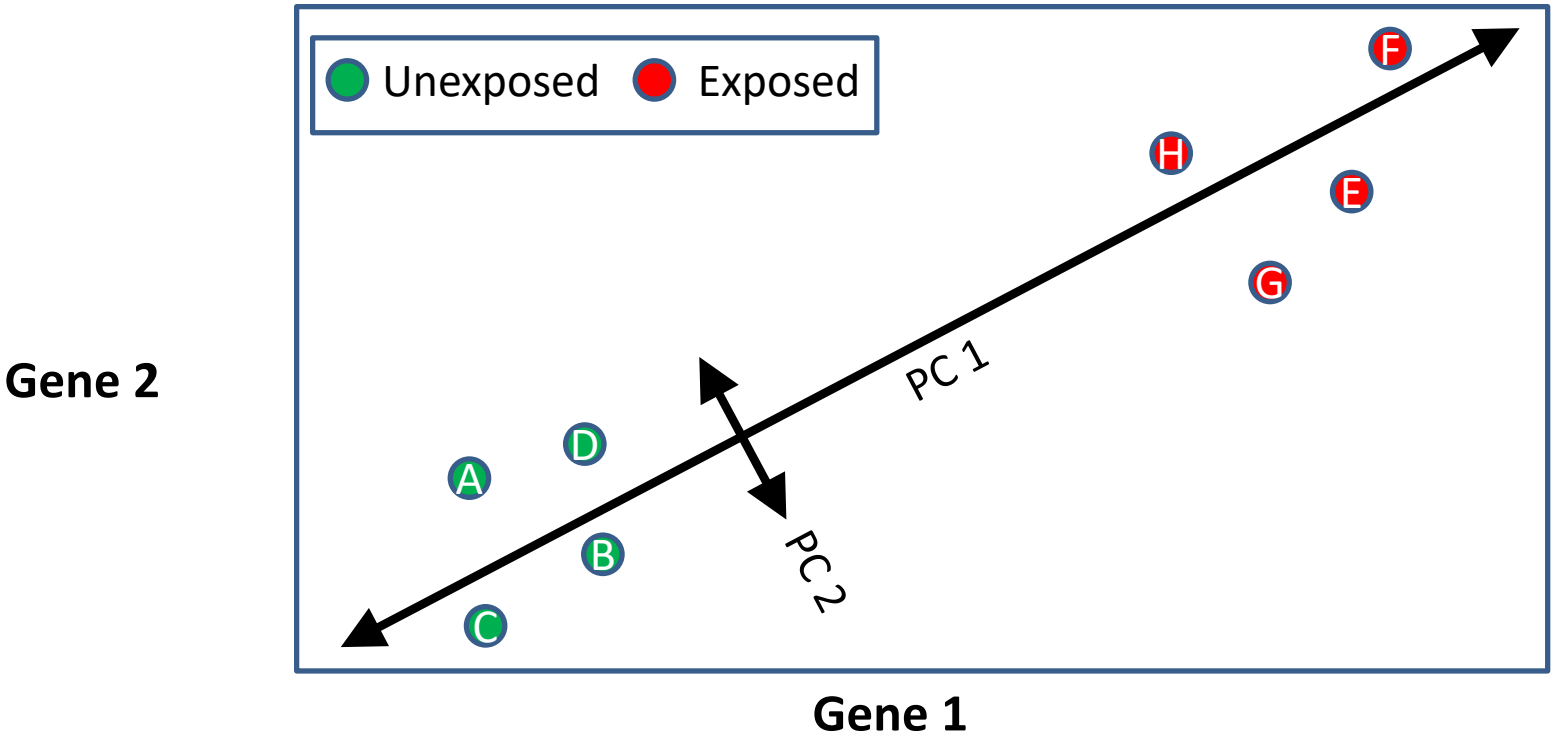
What do we mean by capturing the most variance?

- Recall the definition of variance for a scalar x:

$$var(x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

- We can compute the variance along any individual of our original dimensions.
- We can also transform the data to compute new axes that are linear combinations of the old ones. For N dimensional data, I can write a new axis: $y = \sum_{j=1}^{N} a_j x_j$, where a_j is a scalar.
- PCA finds such linear combinations that have maximum variance.

Principal Component Analysis



Here's an example where one dimension is almost as good as two.

We can generalize this approach so that one dimension is almost as good as N, where N is large.

1. PCA finds useful linear combinations of thousands of variables.
2. There are as many PCs as there were dimensions in the original data.
3. The PCs are orthogonal.
4. Often, a few PCs will capture most of the variance.