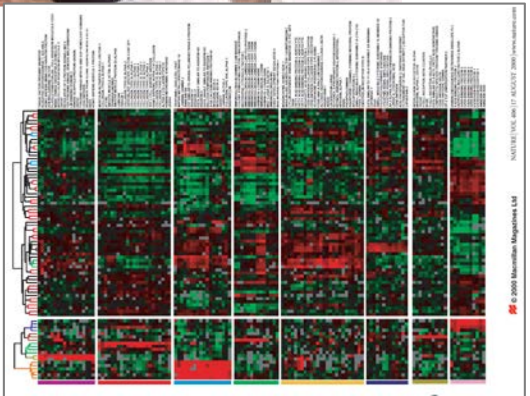
Classical Biology:
Driven by macroscopic observation

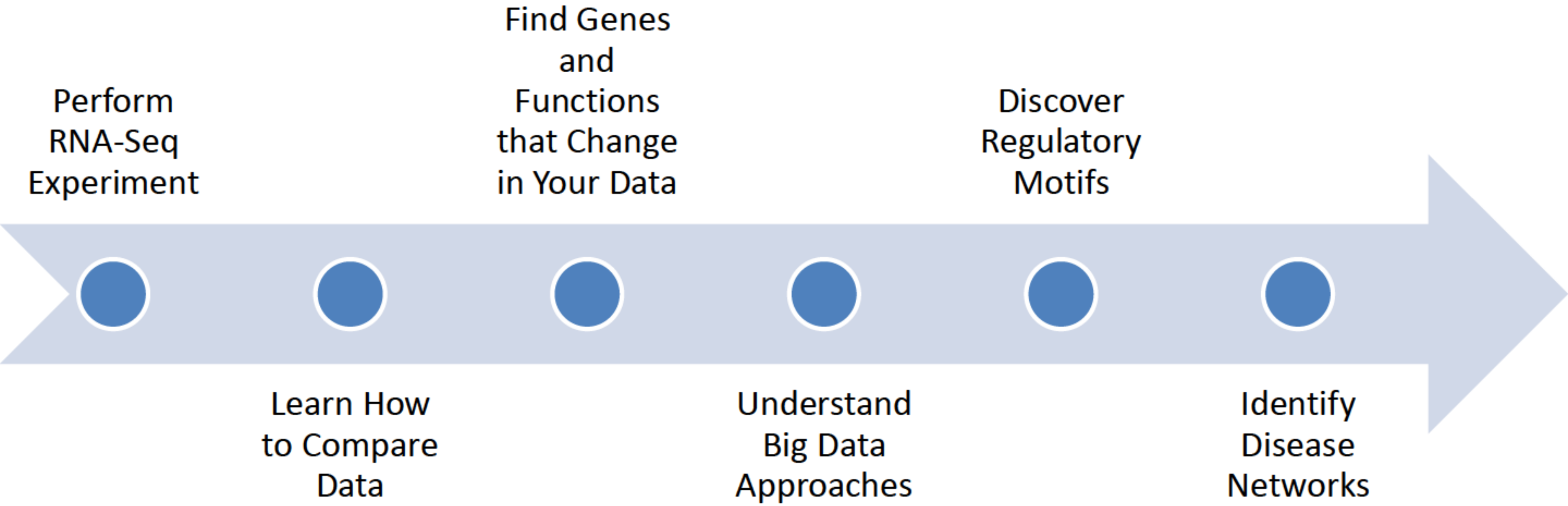
Molecular Biology:
Driven by simple hypotheses



Systems Biology:
Driven by molecular data

Learning Objectives

- Choose the right distance metric to compare the expression of two genes
- Describe why you would cluster expression by genes or experiments
- Manually cluster small vectors using hierarchical or k-means clustering
- Read a dendrogram
- Describe the results of Principal Component Analysis (PCA)



Perform RNA-Seq Experiment

Learn How to Compare Data

Find Genes and Functions that Change in Your Data

Understand Big Data Approaches

Discover Regulatory Motifs

Identify Disease Networks

Comparing
the Expression of Two Genes

Expression data as multidimensional vectors

In our timecourse:
X_A=(e_{A1}, e_{A2},… e_{AN})
X_b=(e_{B1}, e_{B2},… e_{BN})

- Euclidean distance provides an intuitive description:

$$d(X_A, X_B) = \sqrt{\sum_{i=1}^{N} (e_{Ai} - e_{Bi})^2}$$

Pearson Correlation

(one of several possible measures of correlation)

- To understand Pearson Correlation, we need to define a Z-score
- $X_{j,K}$= Expression of gene *j* in experiment *K*
- $Z_{j,K}$= z-score of gene *j* in experiment K:

Z-score

$$Z_{jK} = \frac{X_{jK} - \bar{X}_K}{\sigma}$$

Standard deviation

$$\sigma = \sqrt{\frac{\sum(X_{jK} - \bar{X}_K)}{N}}$$

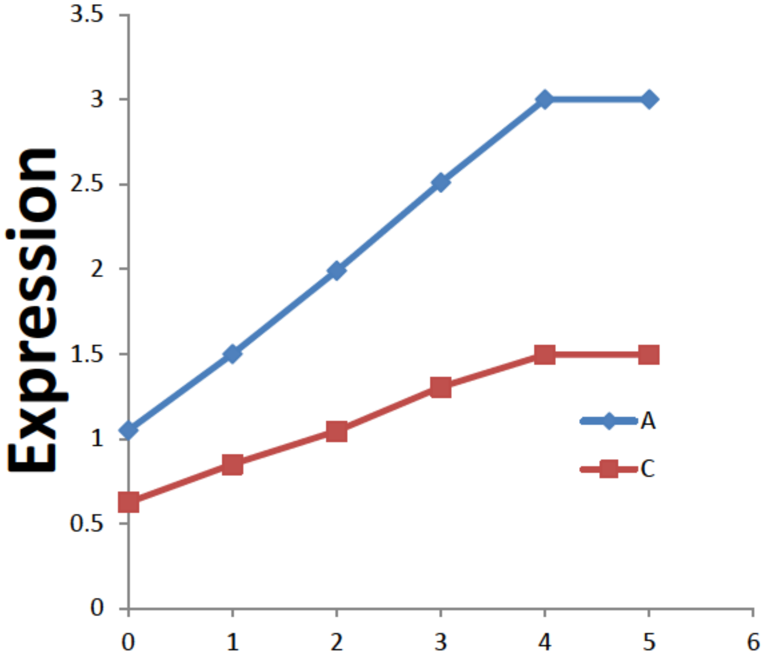
- Pearson correlation from +1 (perfect correlation) to -1 (anti-correlated)

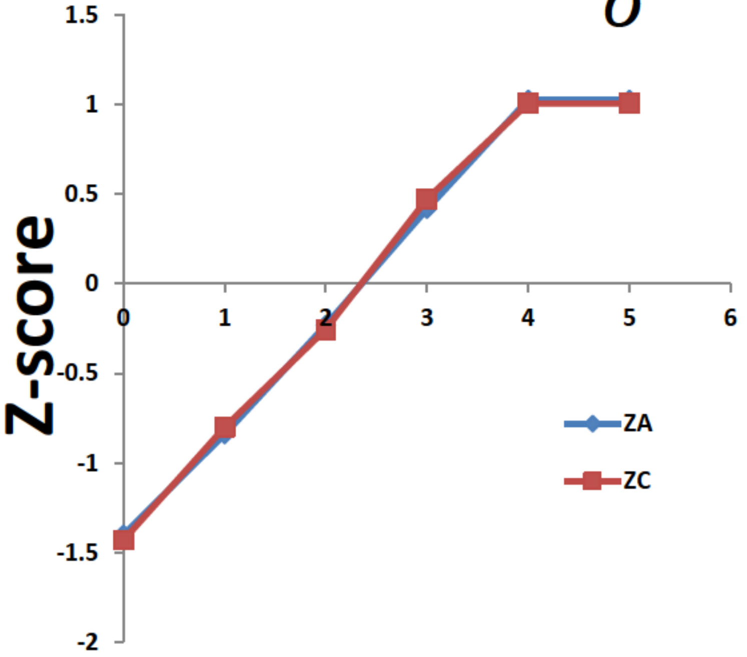
over all experiments

$$r_{A,B} = \frac{\sum_{k=1}^{N_{expt}} Z_{Ak} Z_{Bk}}{N}$$

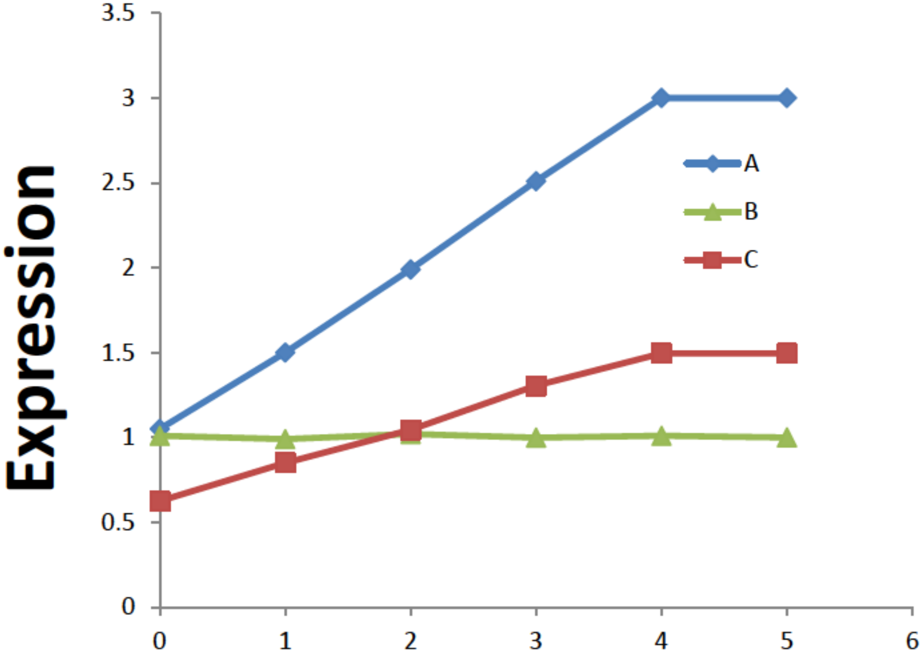
Distance = 1-$r_{A,B}$

$$Z_{Ki} = \frac{X_{Ki} - \bar{X}_K}{\sigma}$$

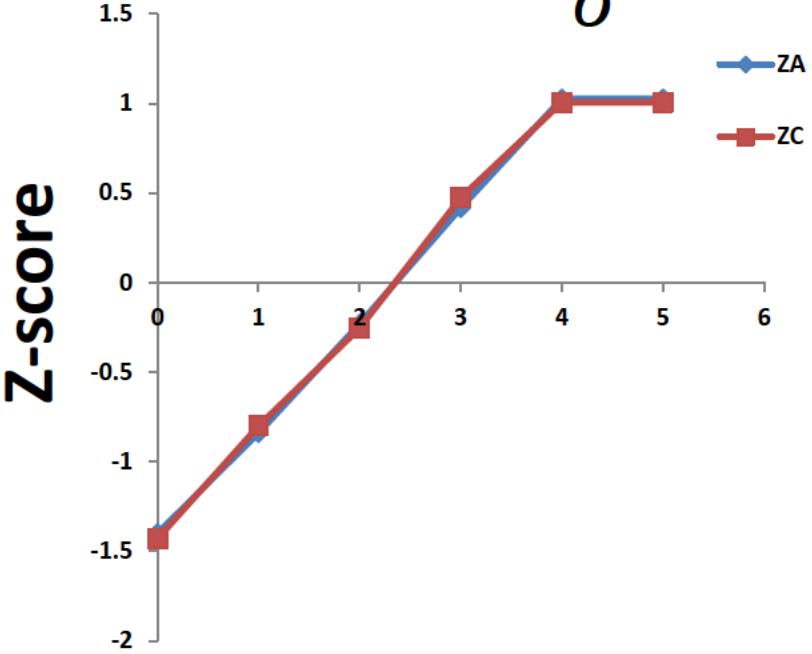




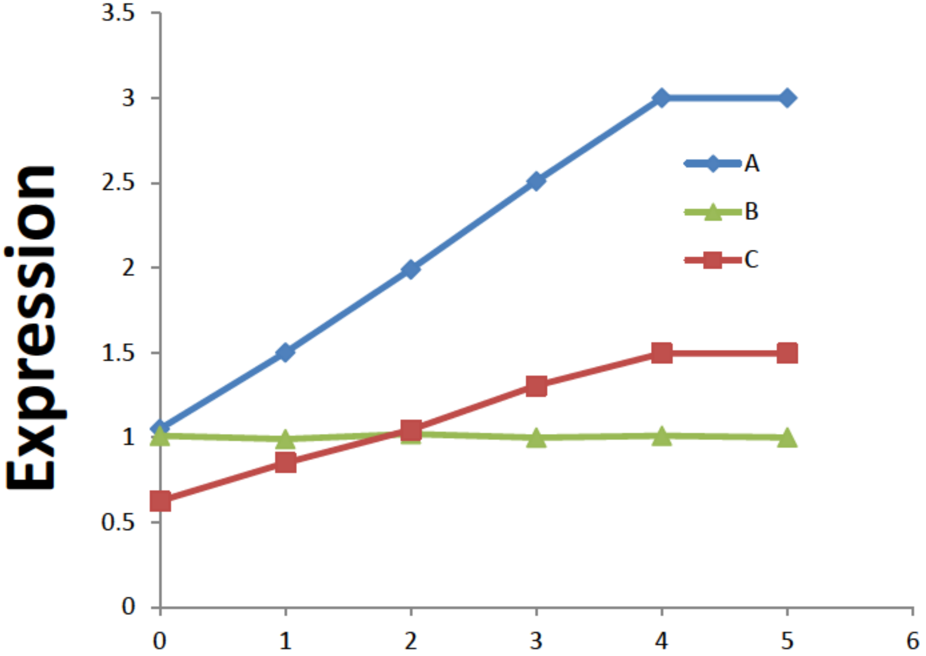
$$r_{A,B} = \frac{\sum_{k=1}^{N_{expt}} Z_{kA} Z_{kB}}{N}$$

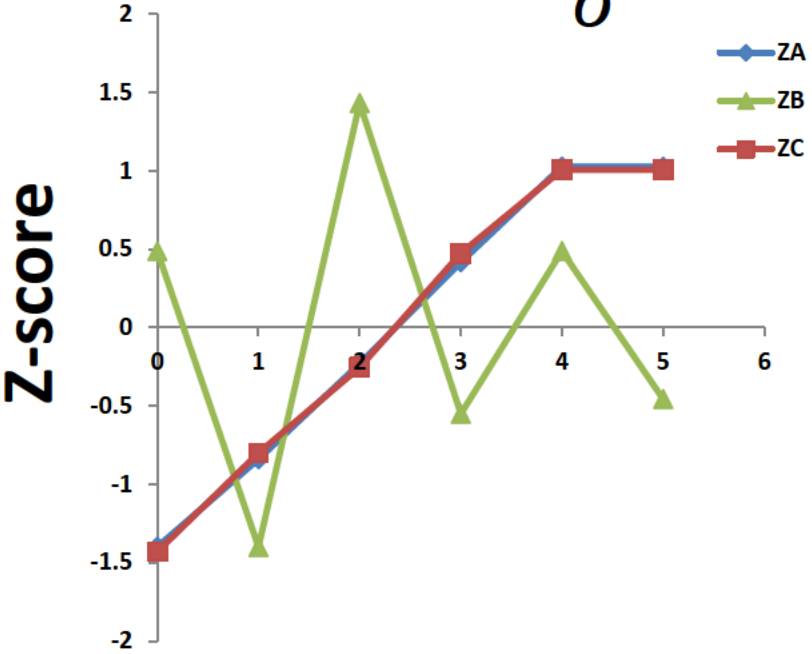


$$Z_{Ki} = \frac{X_{Ki} - \bar{X}_K}{\sigma}$$



$$Z_{Ki} = \frac{X_{Ki} - \bar{X}_K}{\sigma}$$





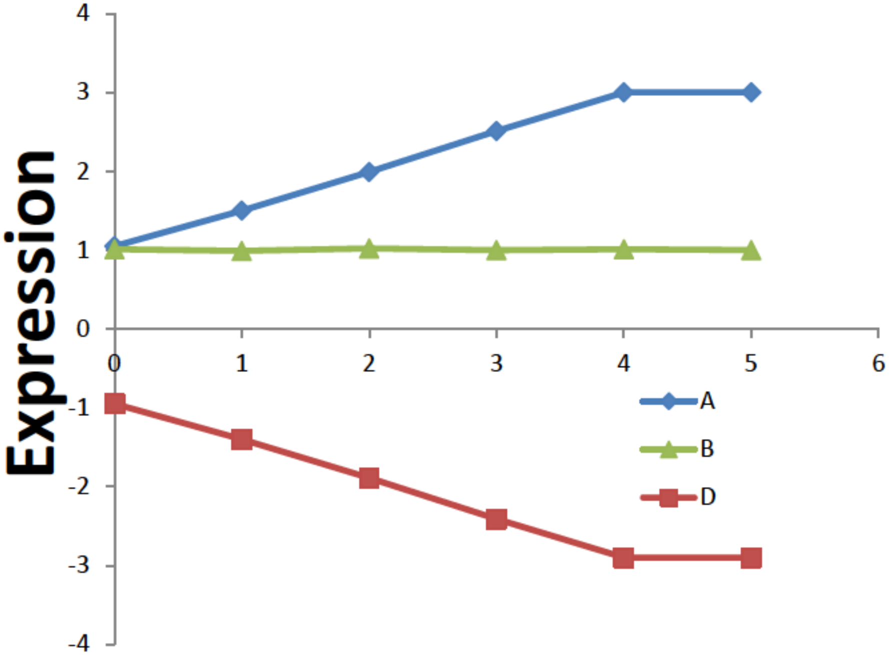
$r_{A,B}$= -0.01

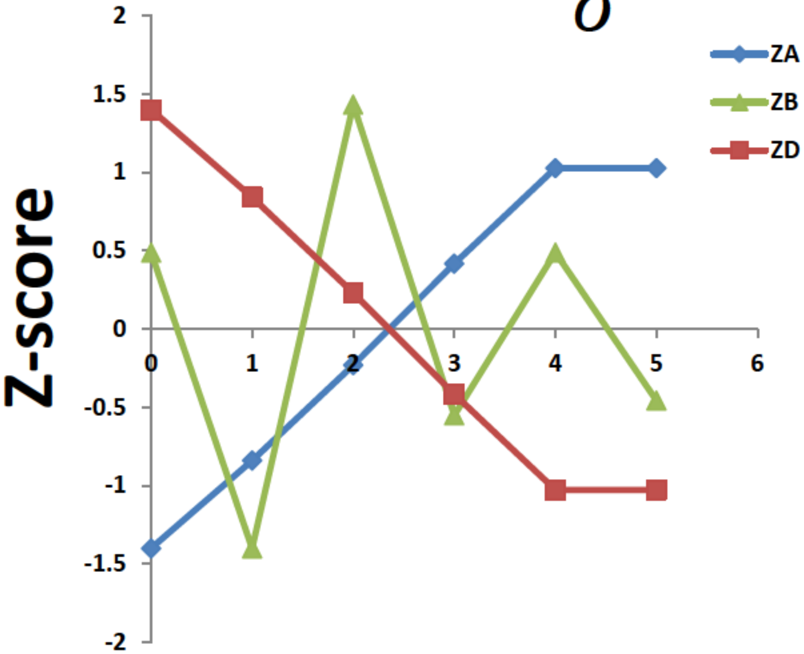
$r_{A,C}$= 0.999

$r_{B,C}$= -0.03

$$r_{A,B} = \frac{\sum_{k=1}^{N_{expt}} Z_{kA} Z_{kB}}{N}$$

$$Z_{Ki} = \frac{X_{Ki} - \bar{X}_K}{\sigma}$$





$r_{A,B}$= -0.01

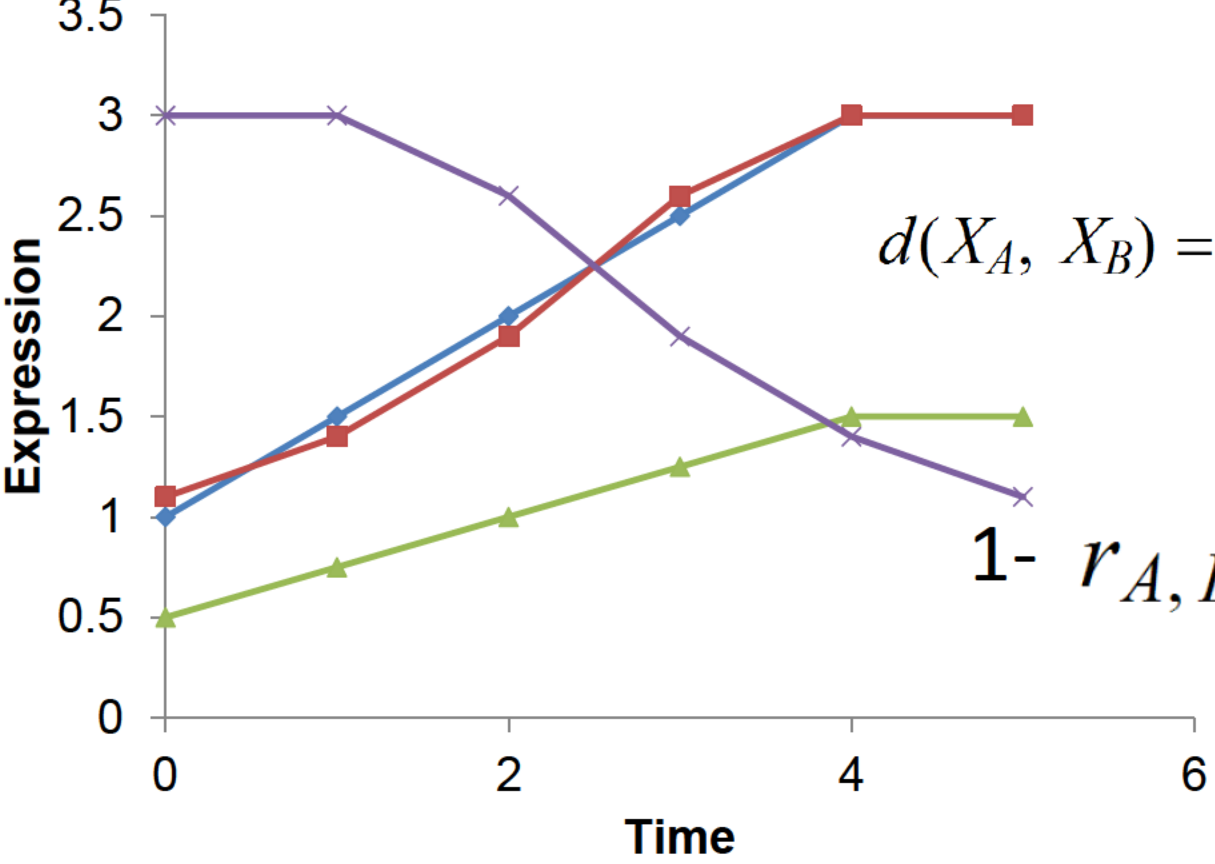
$r_{A,D}$= -1.0

$r_{B,D}$= 0.007

$$r_{A,B} = \frac{\sum_{k=1}^{N_{expt}} Z_{kA} Z_{kB}}{N}$$

Distance Metrics

Which would you use to find co-regulated genes?

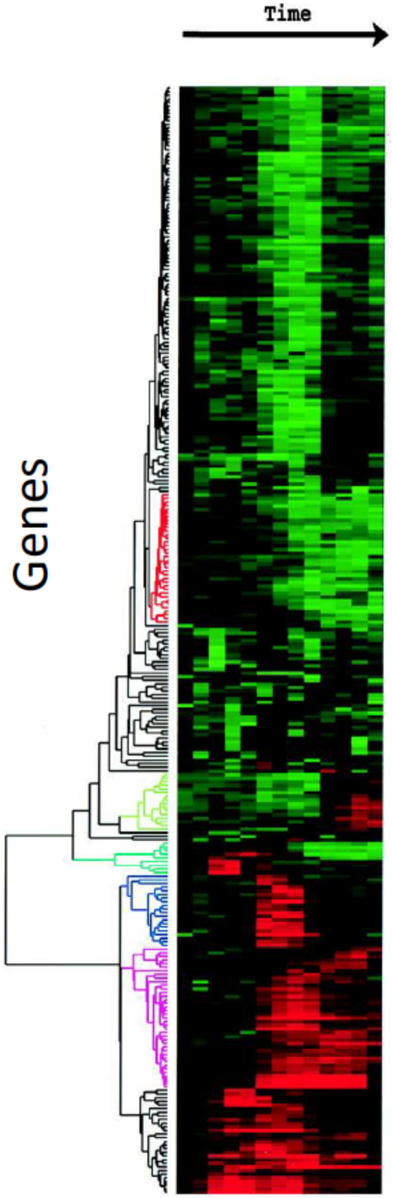


$$d(X_A,\ X_B) = \sqrt{\sum_{k=1}^{N} (X_{A,k} - X_{B,k})^2}$$

$$1\text{-}\ r_{A,B} = 1\text{-} \frac{\sum Z_A Z_B}{N}$$

Write on Board: Learning Objectives

- Choose the right distance metric to compare the expression of two genes
- Describe why you would cluster expression by genes or experiments
- Manually cluster small vectors using hierarchical or k-means clustering
- Read a dendrogram
- Describe the results of Principal Component Analysis (PCA)



Clustering 8600 human genes based on time course of expression following serum stimulation of fibroblasts

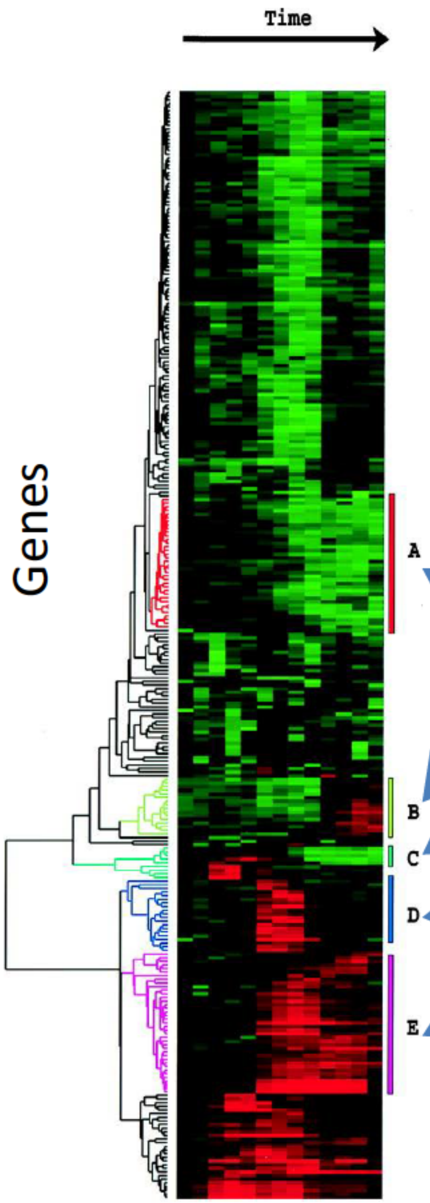
Key: Black = little change Green = down Red = up

(relative to initial time point)

What can you learn from the clustering?

Iyer et al. *Science* 1999

Clustering 8600 human genes based on time course of expression following serum stimulation of fibroblasts



Key: Black = little change Green = down Red = up

(relative to initial time point)

Why might you cluster experiments?

(A) cholesterol biosynthesis

(B) the cell cycle

(C) the immediate-early response

(D) signaling and angiogenesis

(E) wound healing and tissue remodeling

Iyer et al. *Science* 1999

How to cluster …

Why cluster?

- Cluster genes (rows)
  - Measure expression at multiple time-points, different conditions, etc.

Similar expression patterns may suggest similar functions of genes

- Cluster samples (columns)
  - e.g., expression levels of thousands of genes for each tumor sample

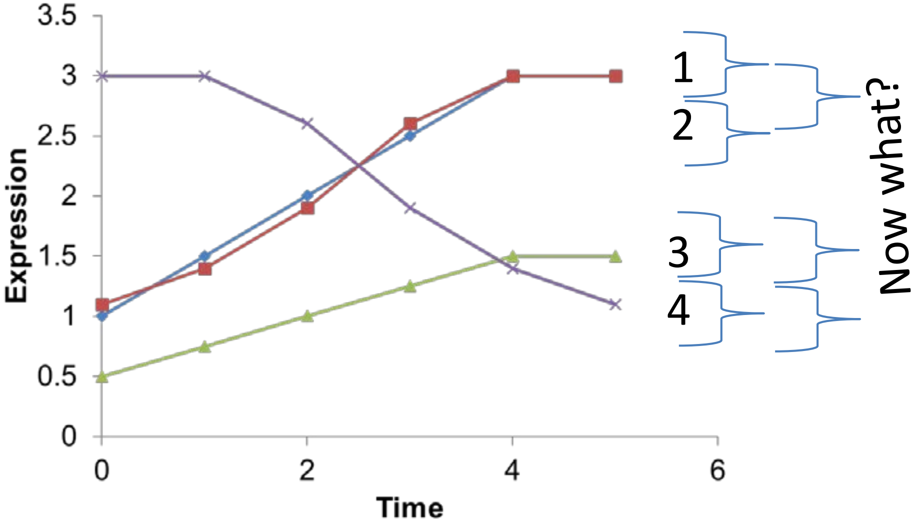
Similar expression patterns may suggest biological relationship among samples

Two types of approaches: Agglomerative & Divisive

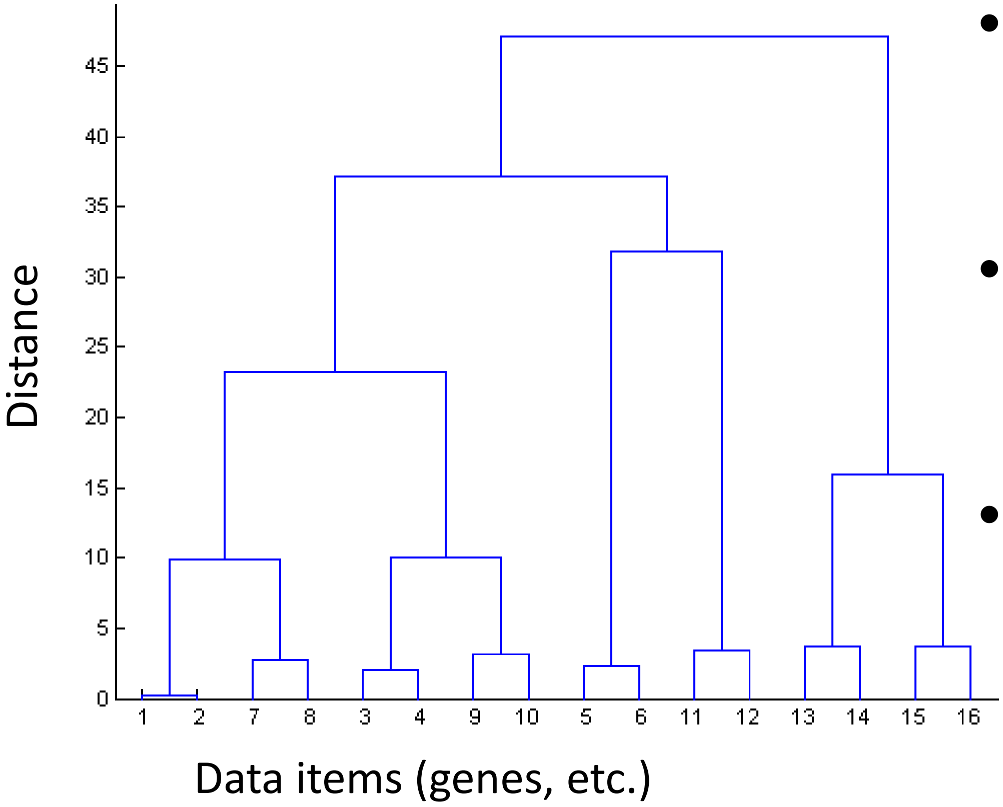
Agglomerative:

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

Distance is defined for a vector; how do I compare clusters? Several choices (min, max, average)



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

Write on Board: Learning Objectives

- Choose the right distance metric to compare the expression of two genes
- Describe why you would cluster expression by genes or experiments
- Manually cluster small vectors using hierarchical or k-means clustering
- Read a dendrogram
- Describe the results of Principal Component Analysis (PCA)

How to cluster with K-means

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:

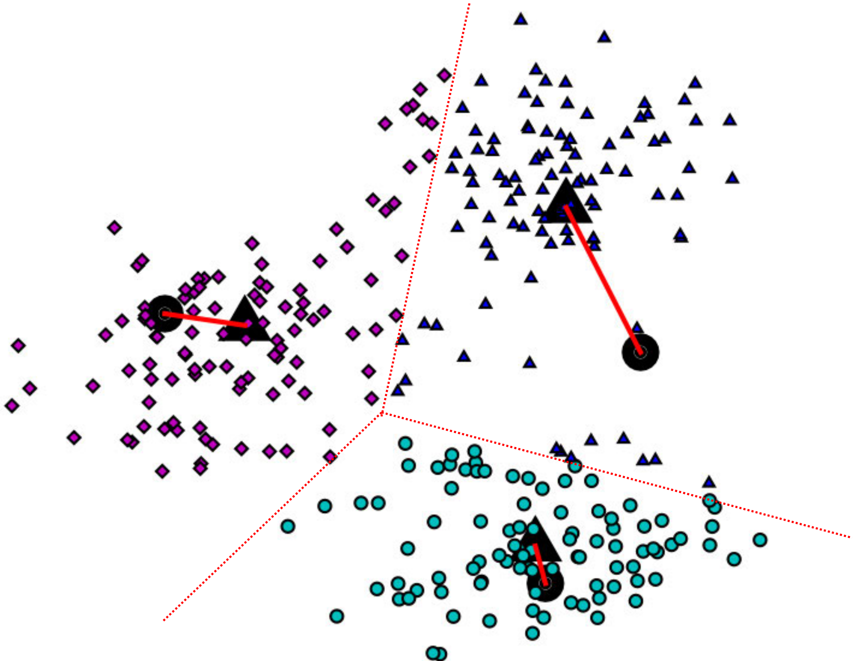
$$\text{centroid}_j = \hat{Y}_j = \frac{1}{N_{Y_j}} \sum_{i \in Y_j} X_i$$

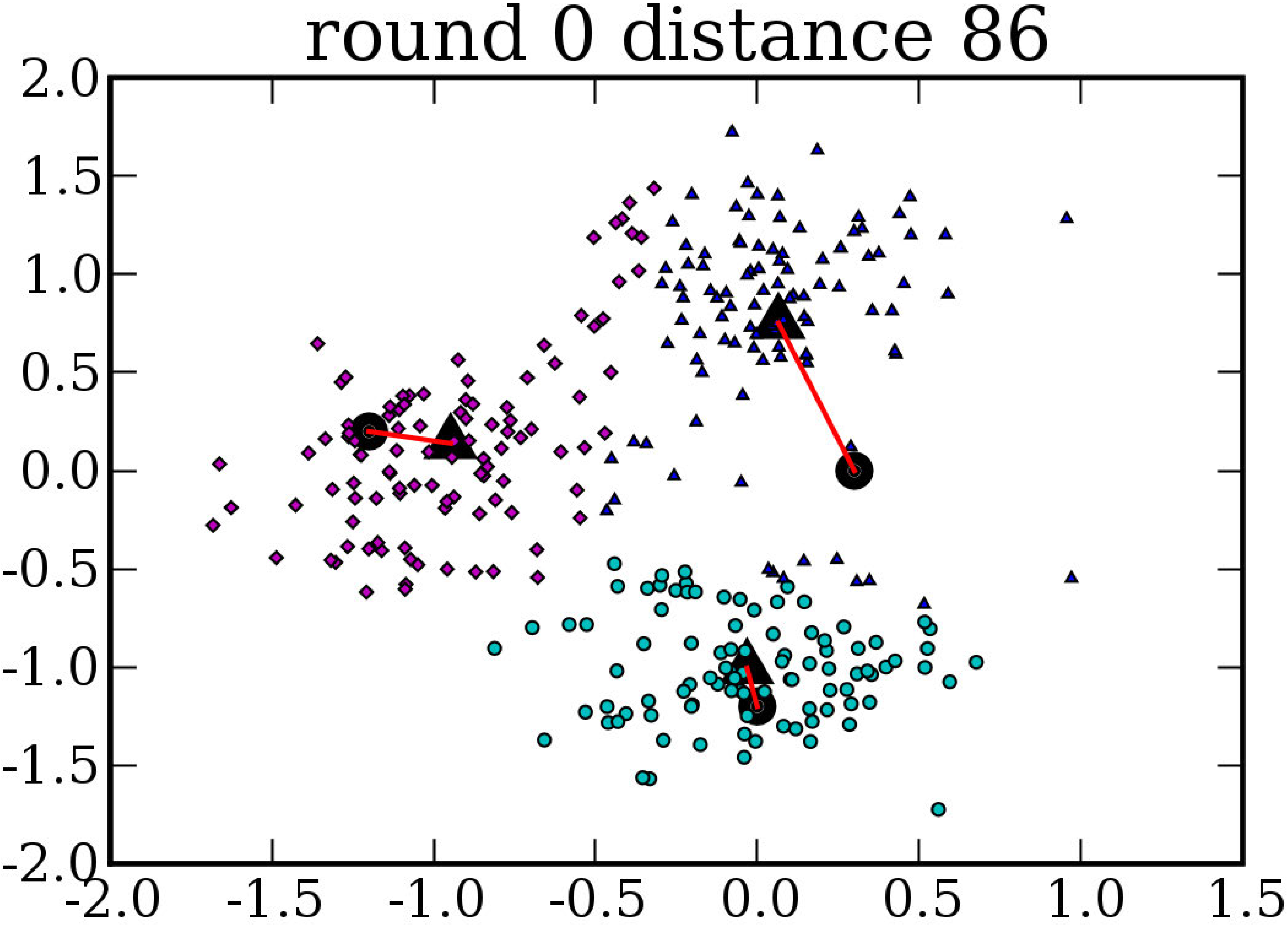
Euclidean
Vector Addition

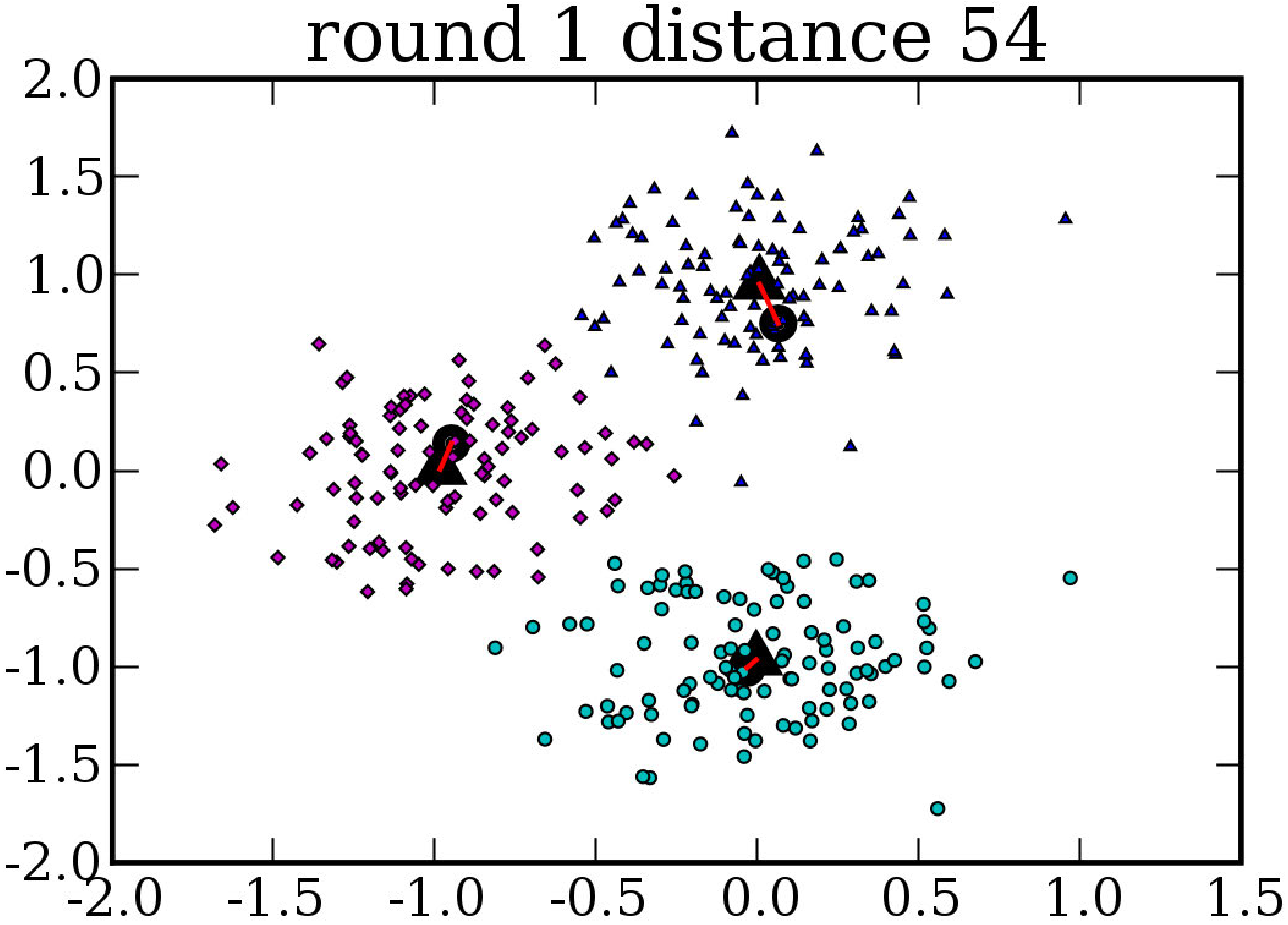
$$\underset{C}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in C(i)} \left| X_j - \hat{Y}_i \right|^2$$

K-means clustering algorithm

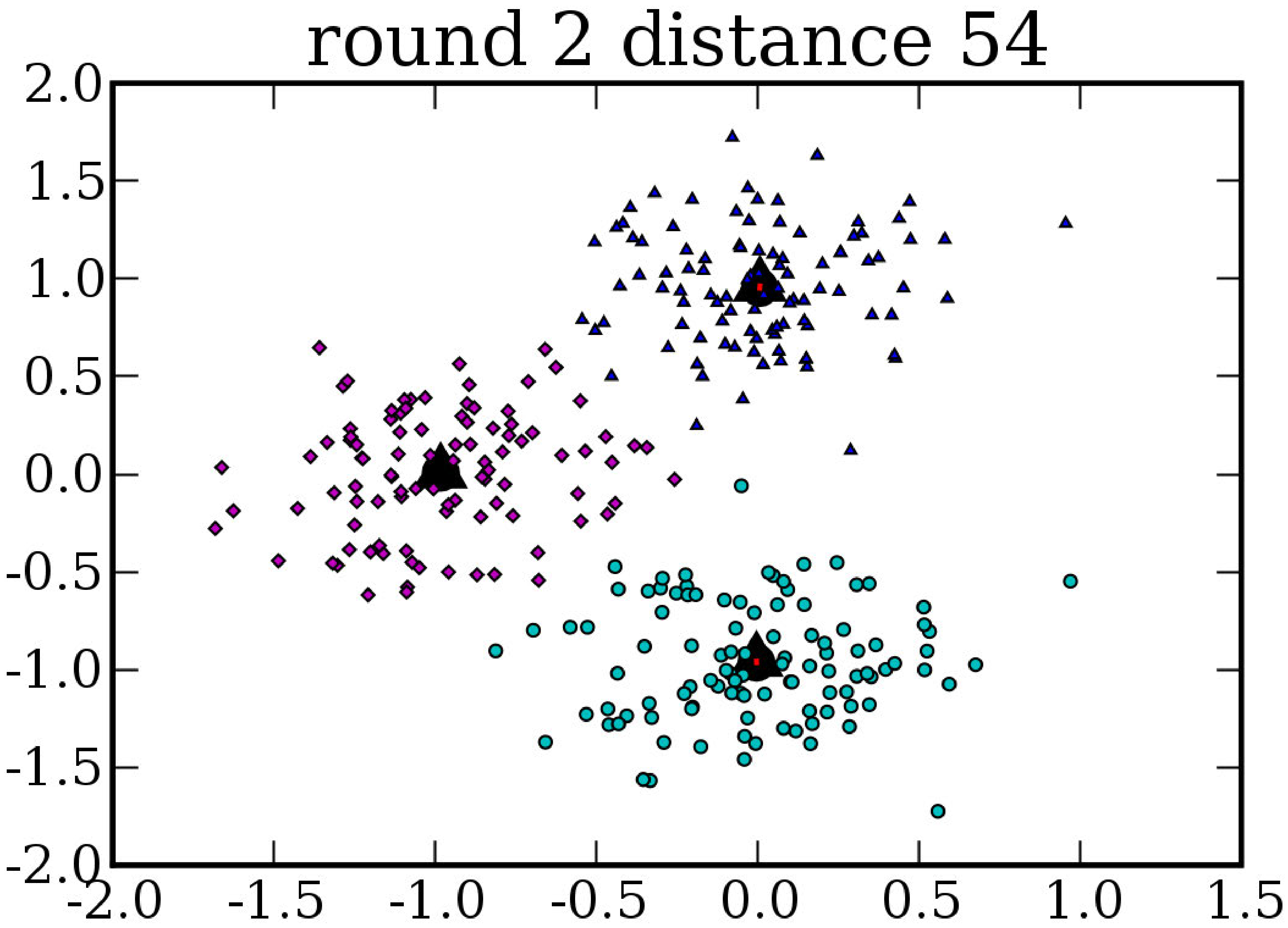
- Initialize: choose k points as cluster means
- Repeat until convergence:
  - Assignment: place each point X_i in the cluster with the closest mean.
  - Update: recalculate the mean for each cluster



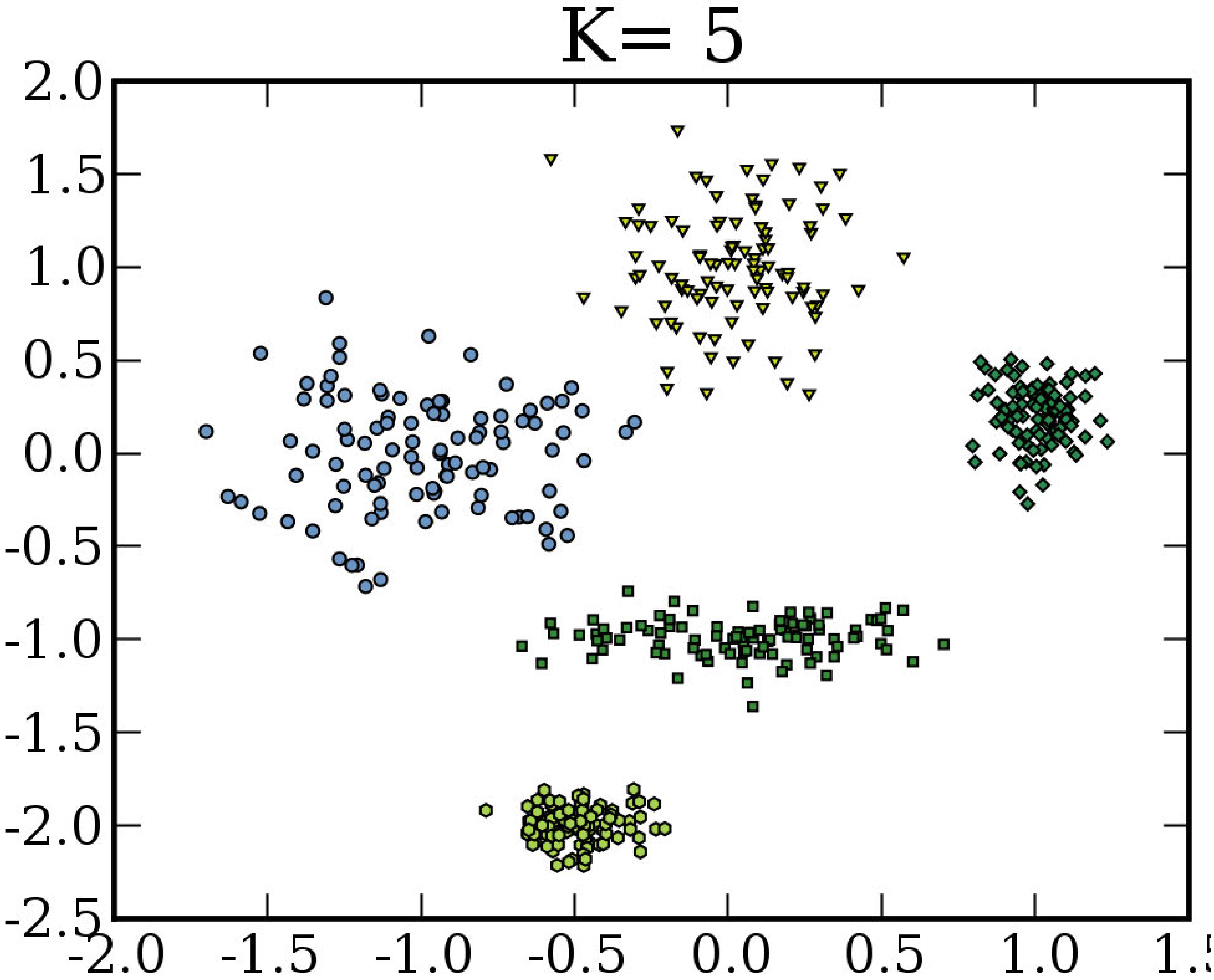

round 0 distance 86
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

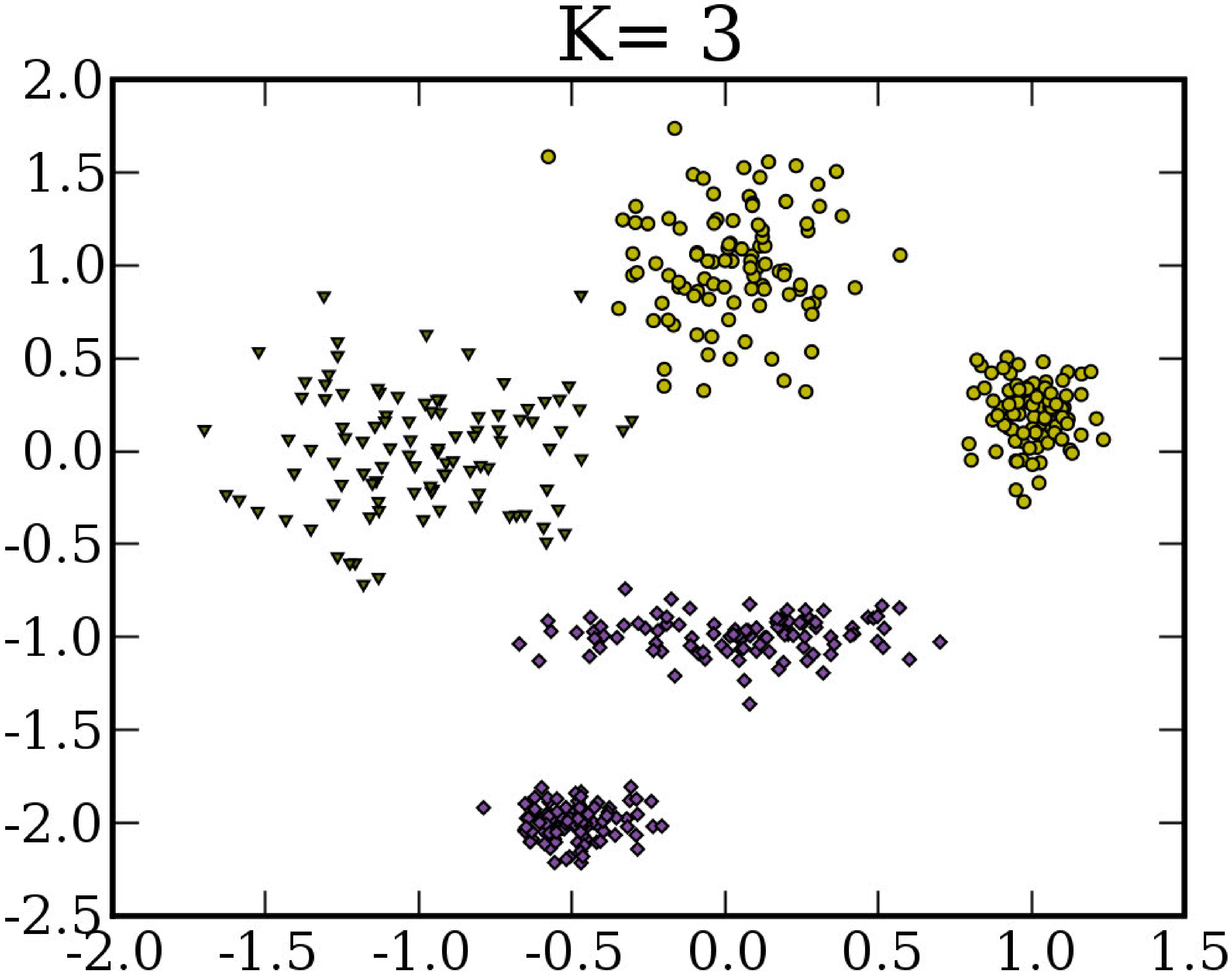


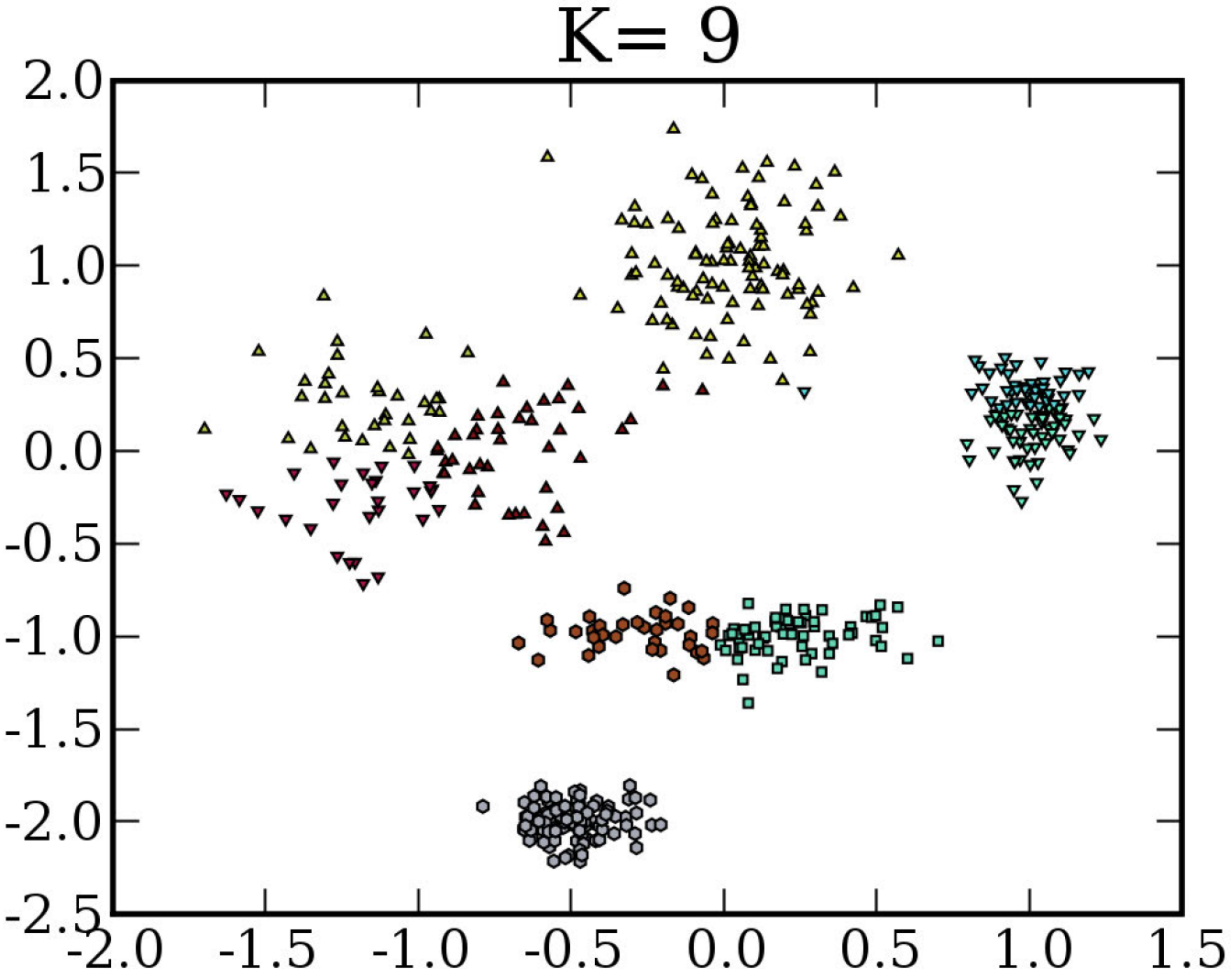
round 1 distance 54

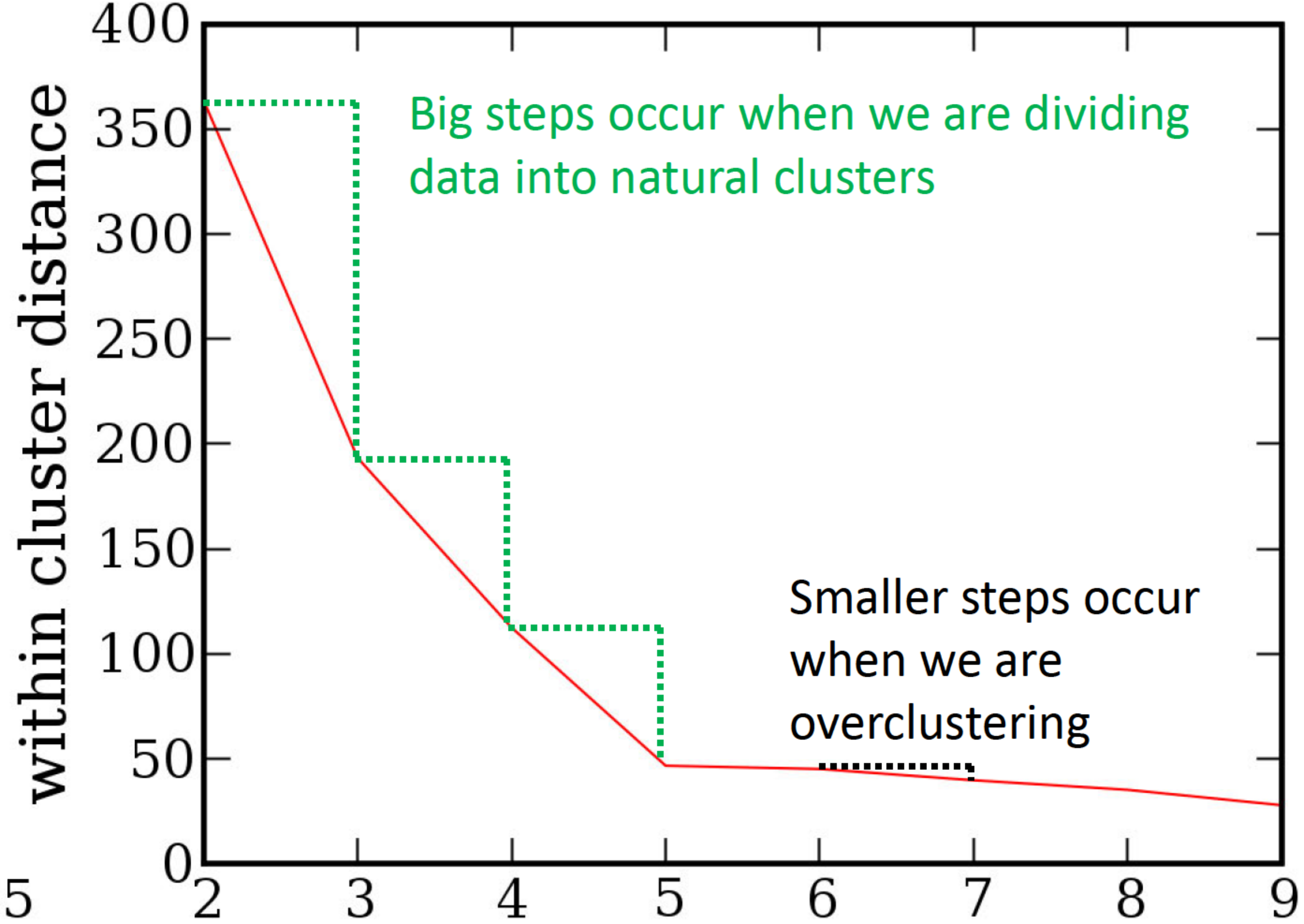

round 2 distance 54
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

What if you choose the wrong K?


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
1.5


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

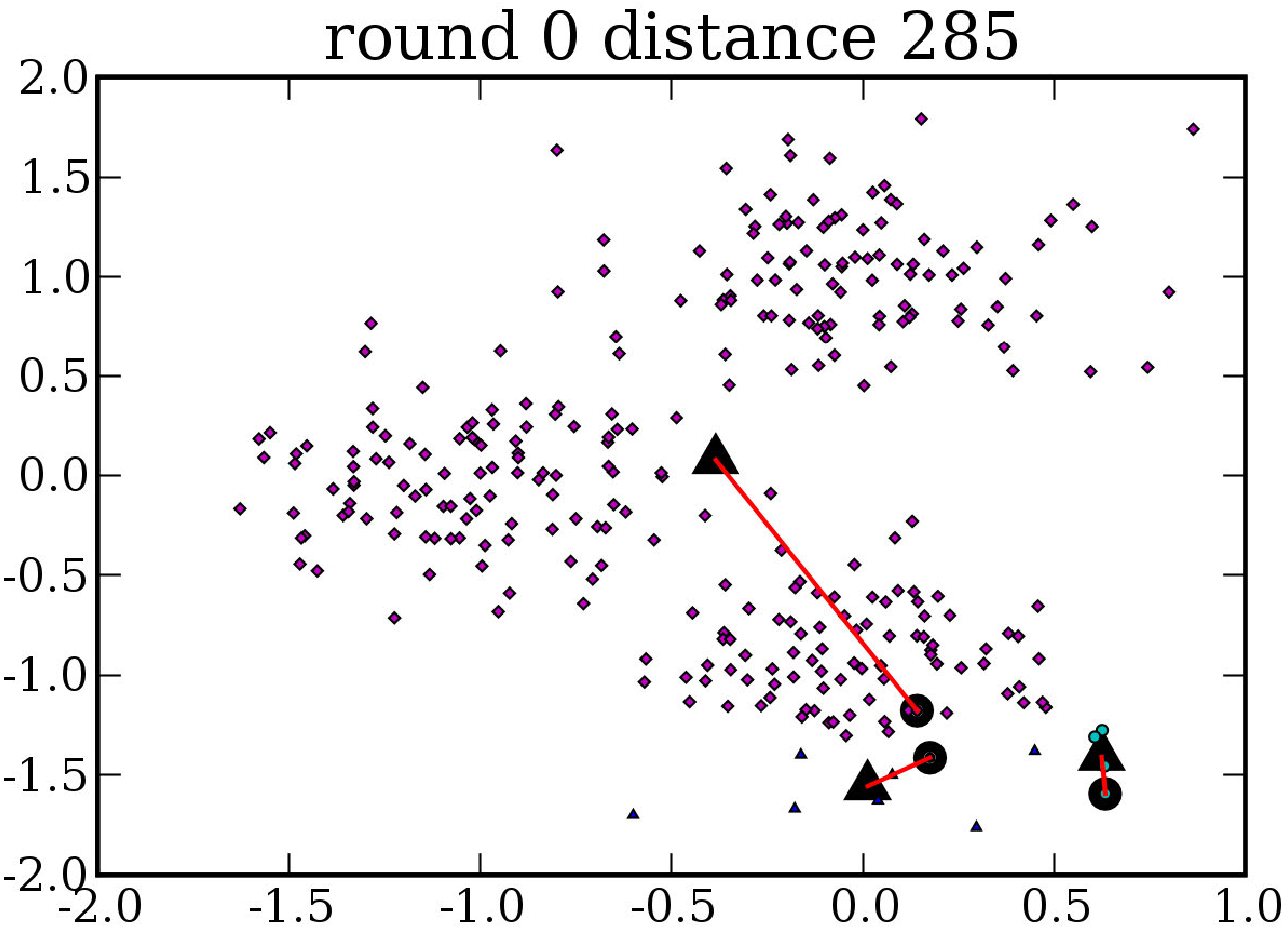

K= 9
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

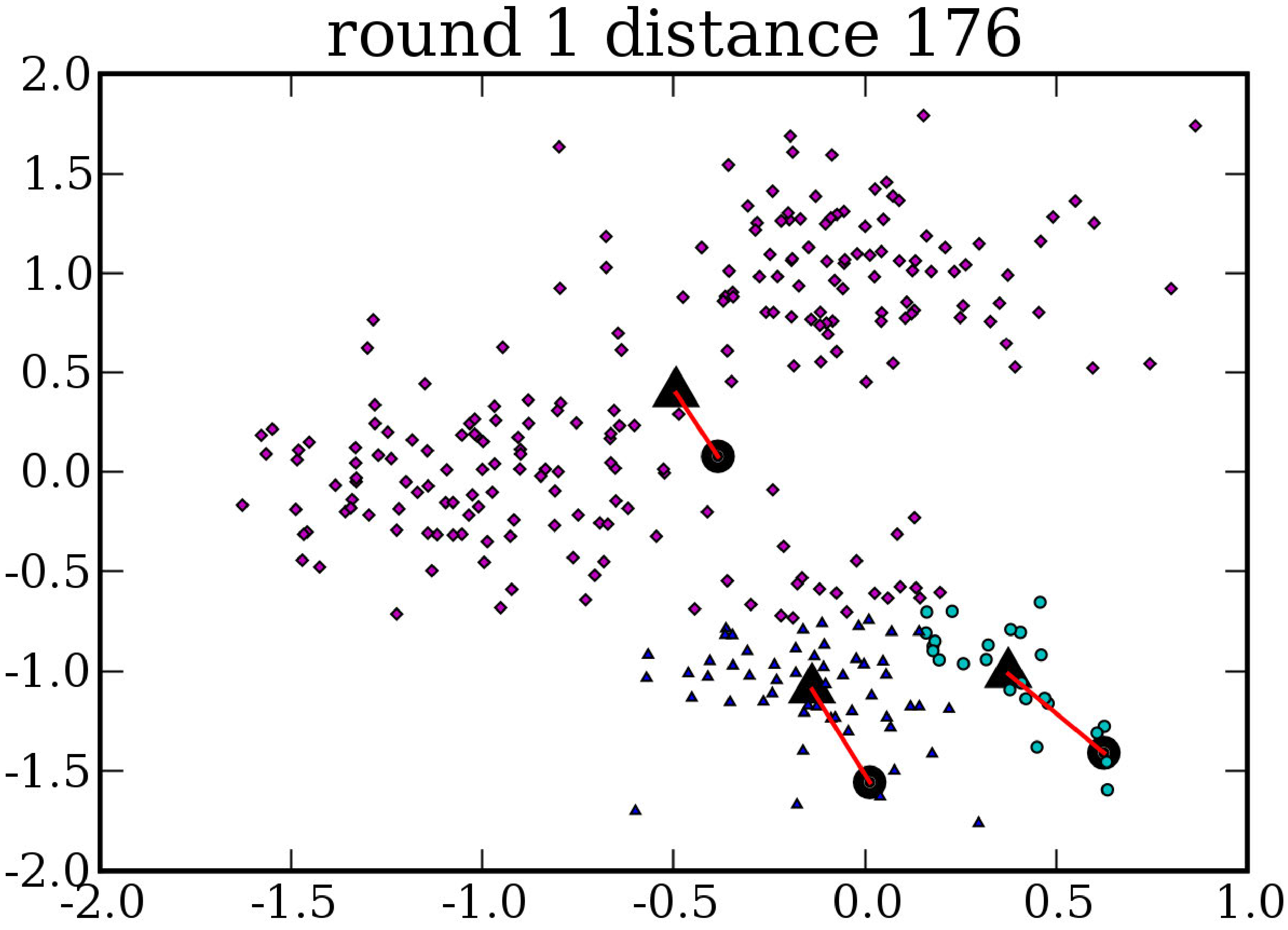


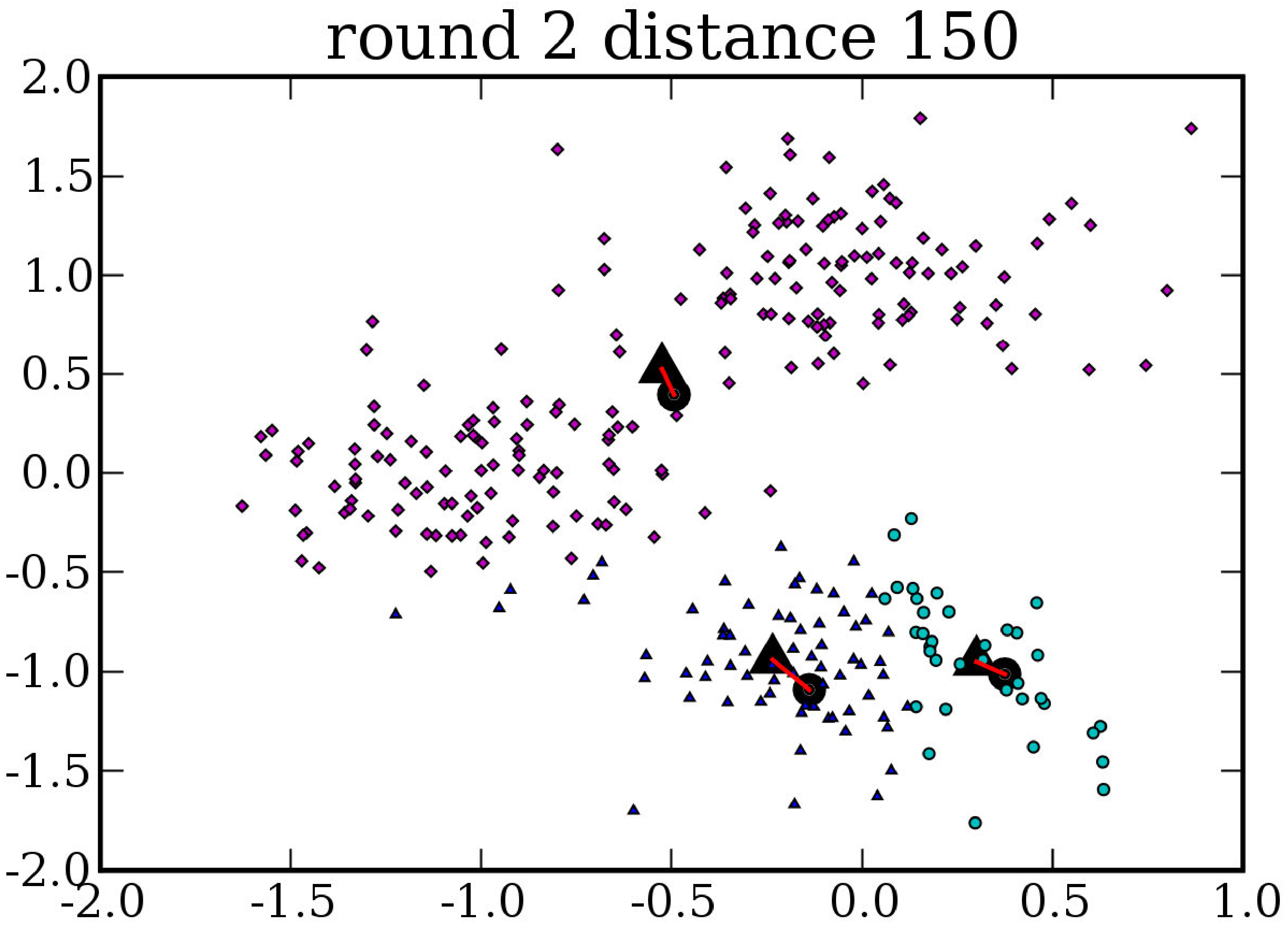
Big steps occur when we are dividing data into natural clusters

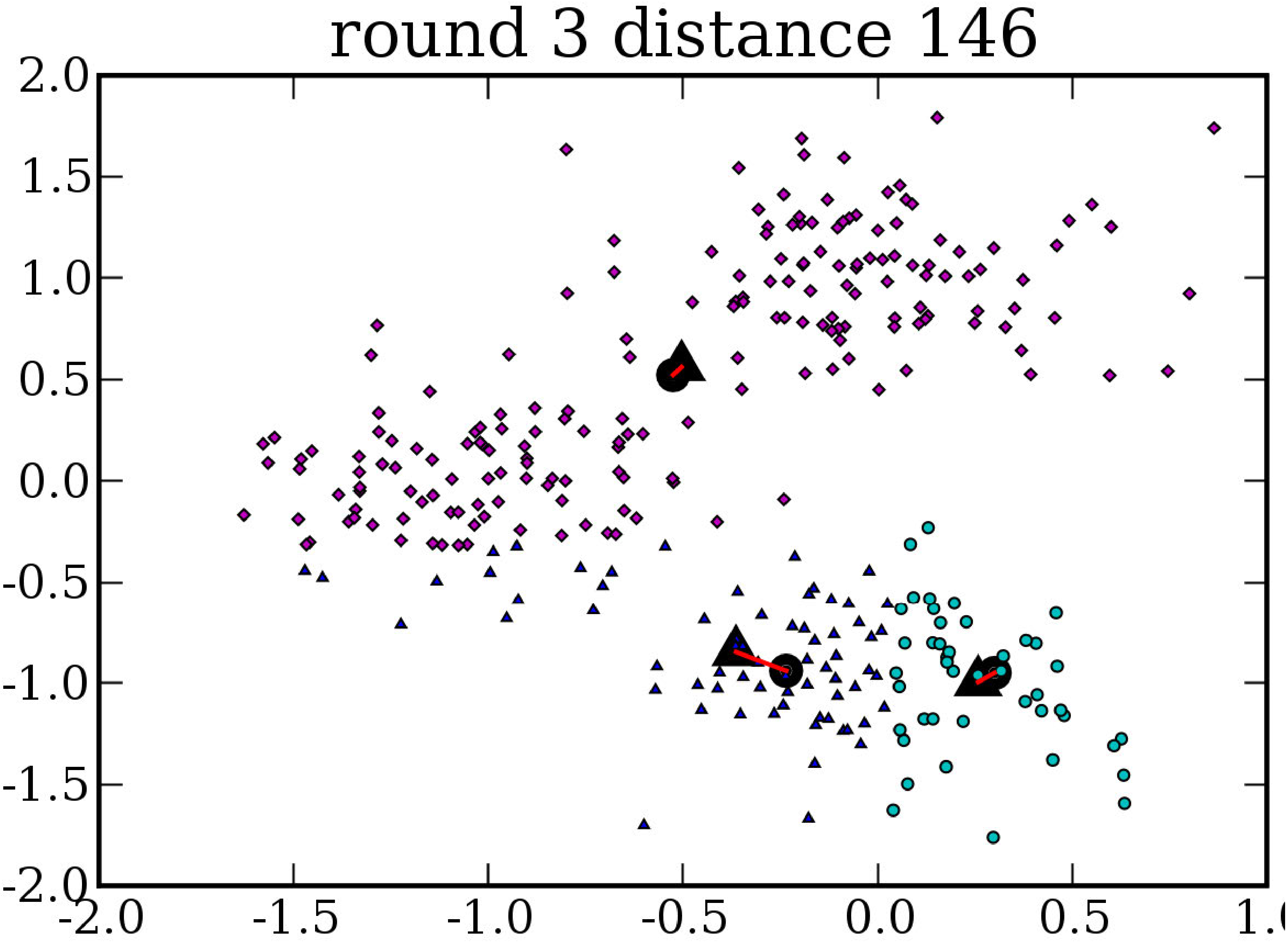
Smaller steps occur when we are overclustering

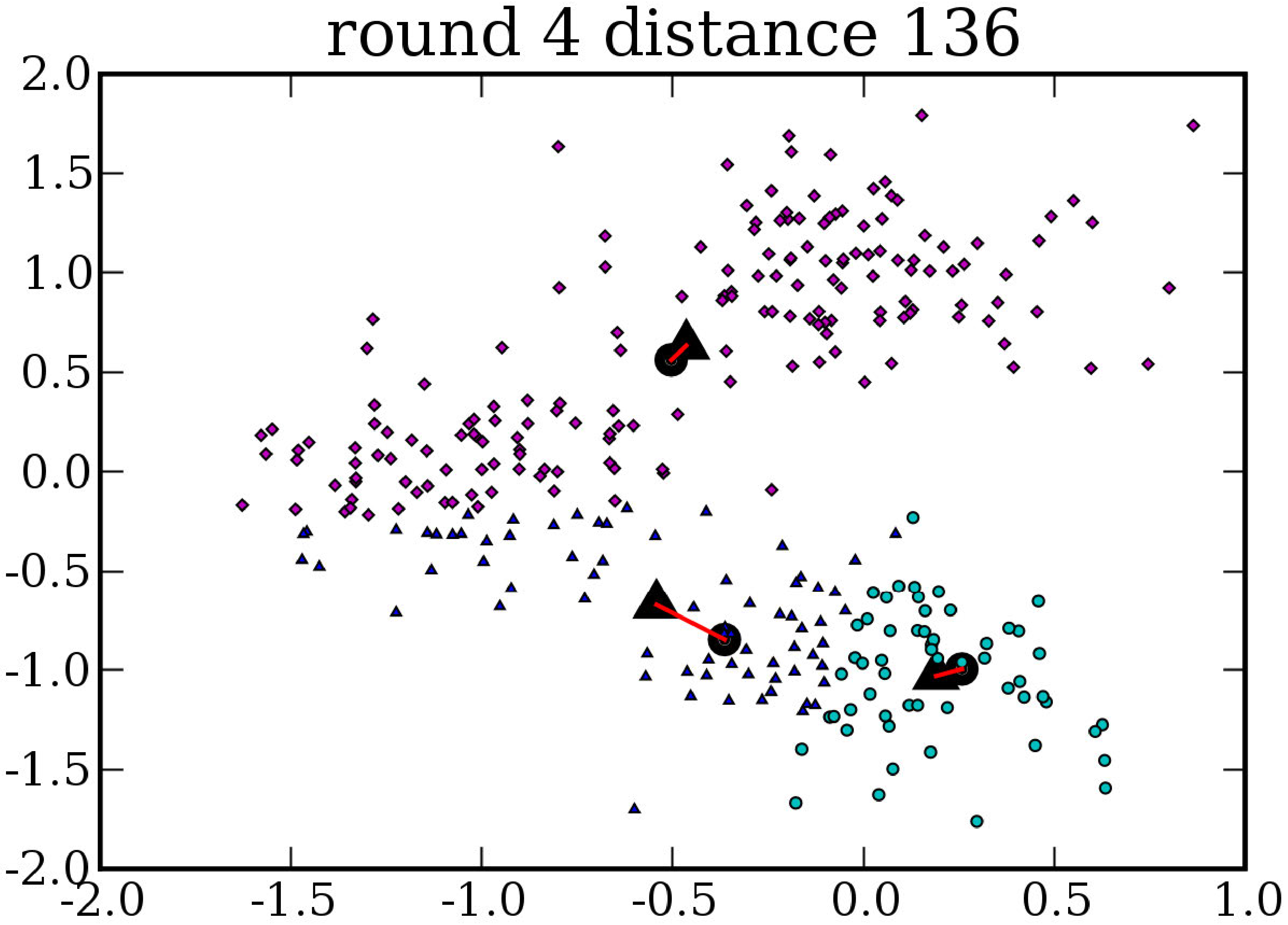
What if we choose pathologically bad initial positions?

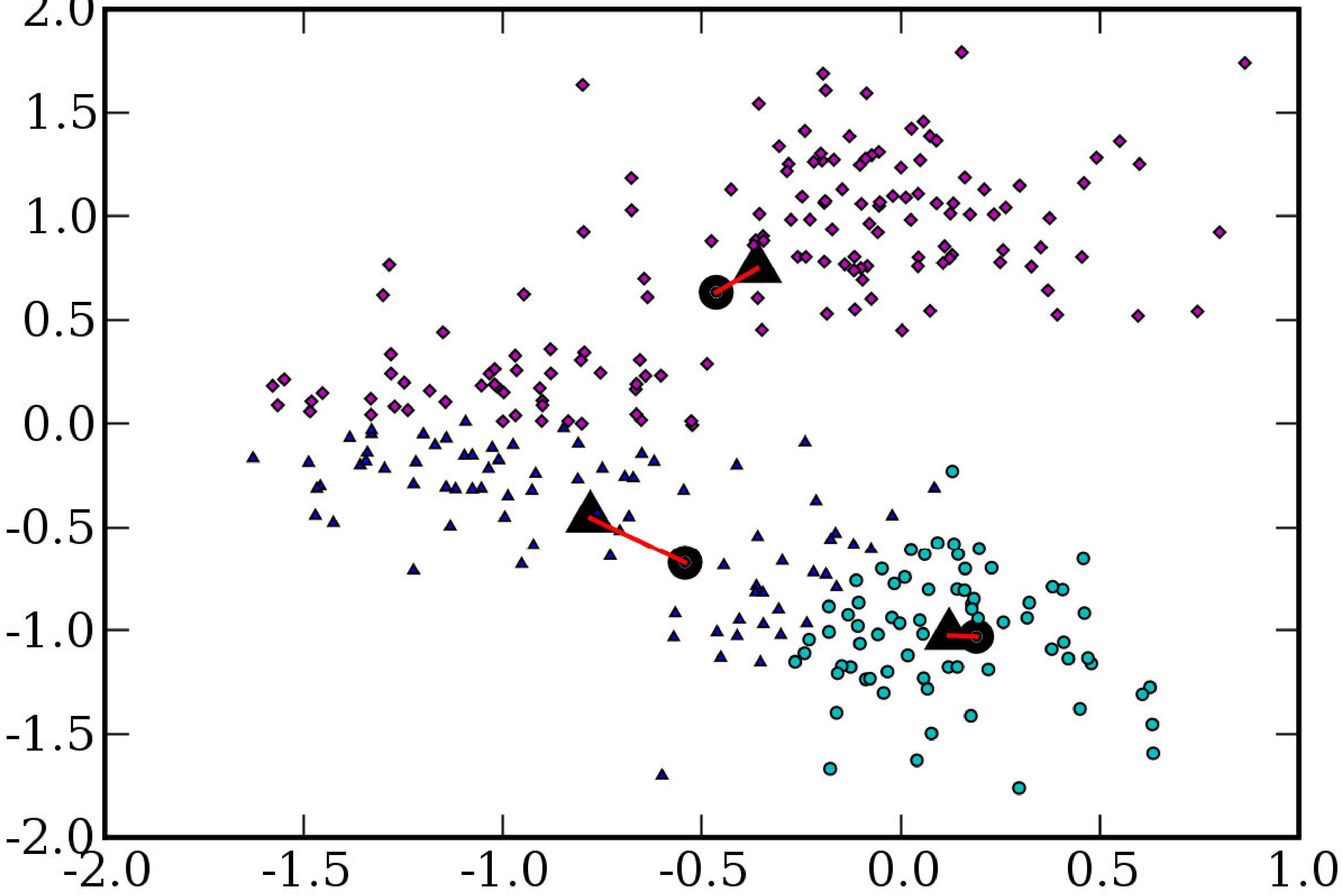

round 0 distance 285
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0


round 1 distance 176
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

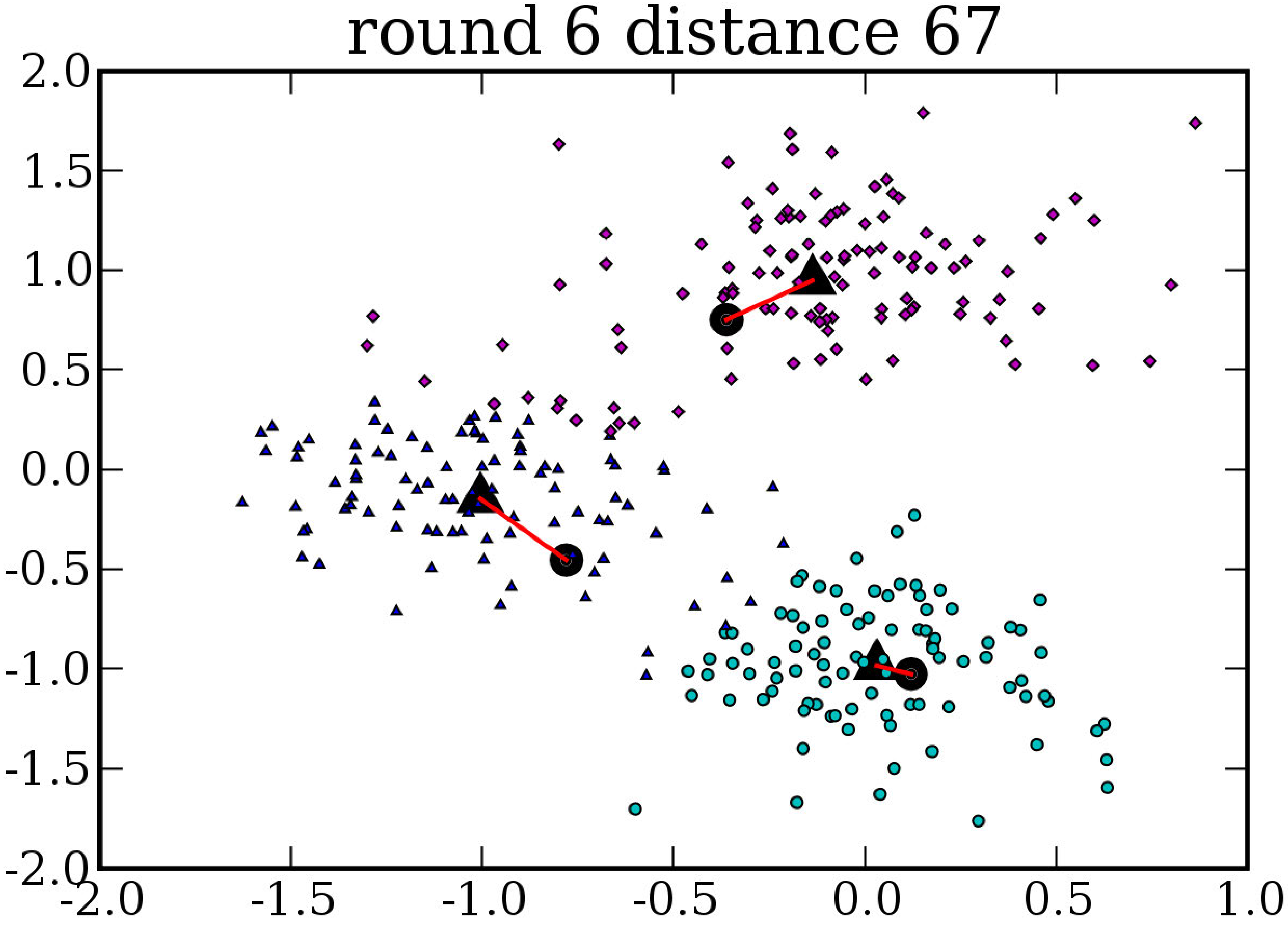

round 2 distance 150
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

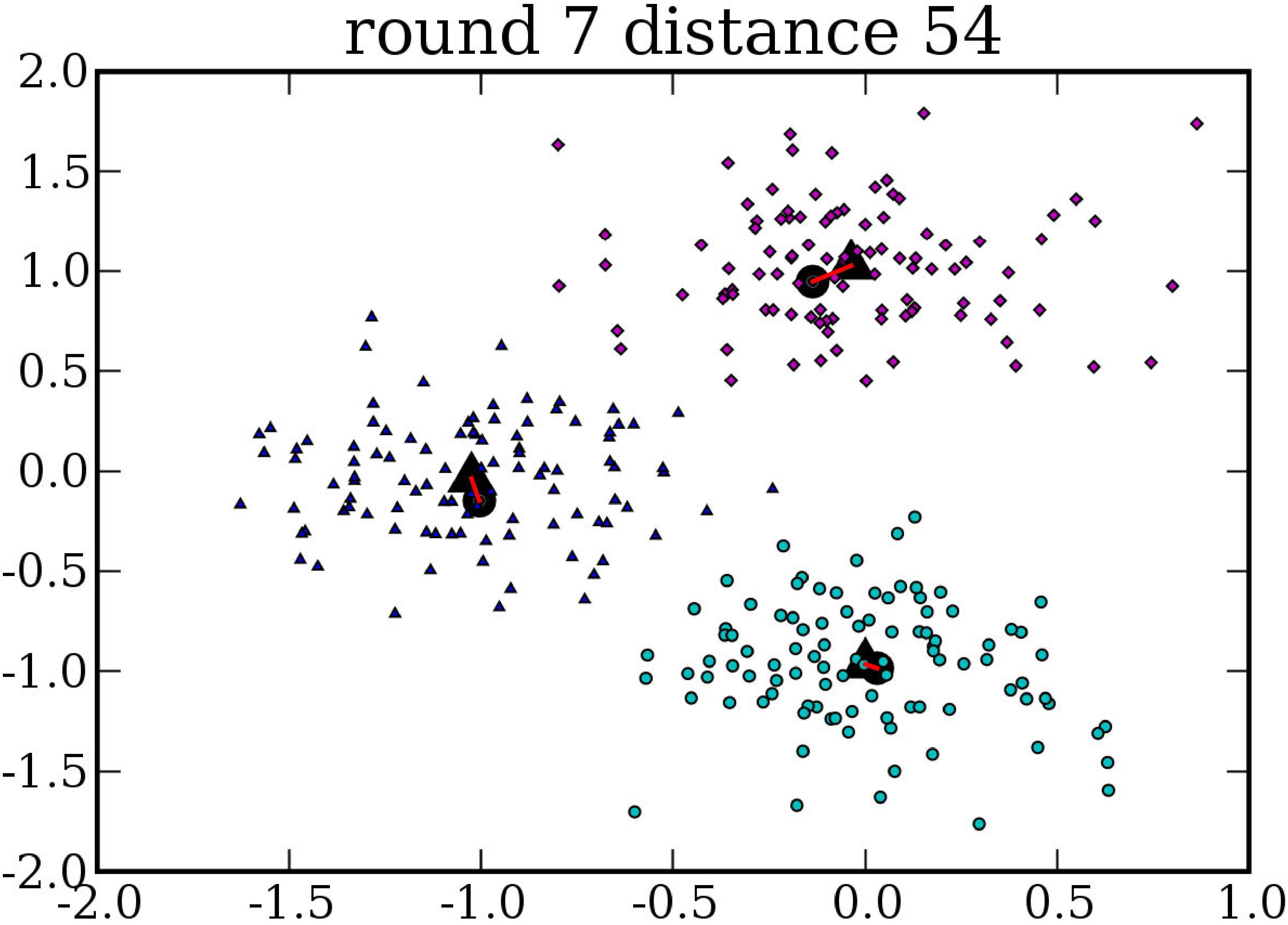

round 3 distance 146
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

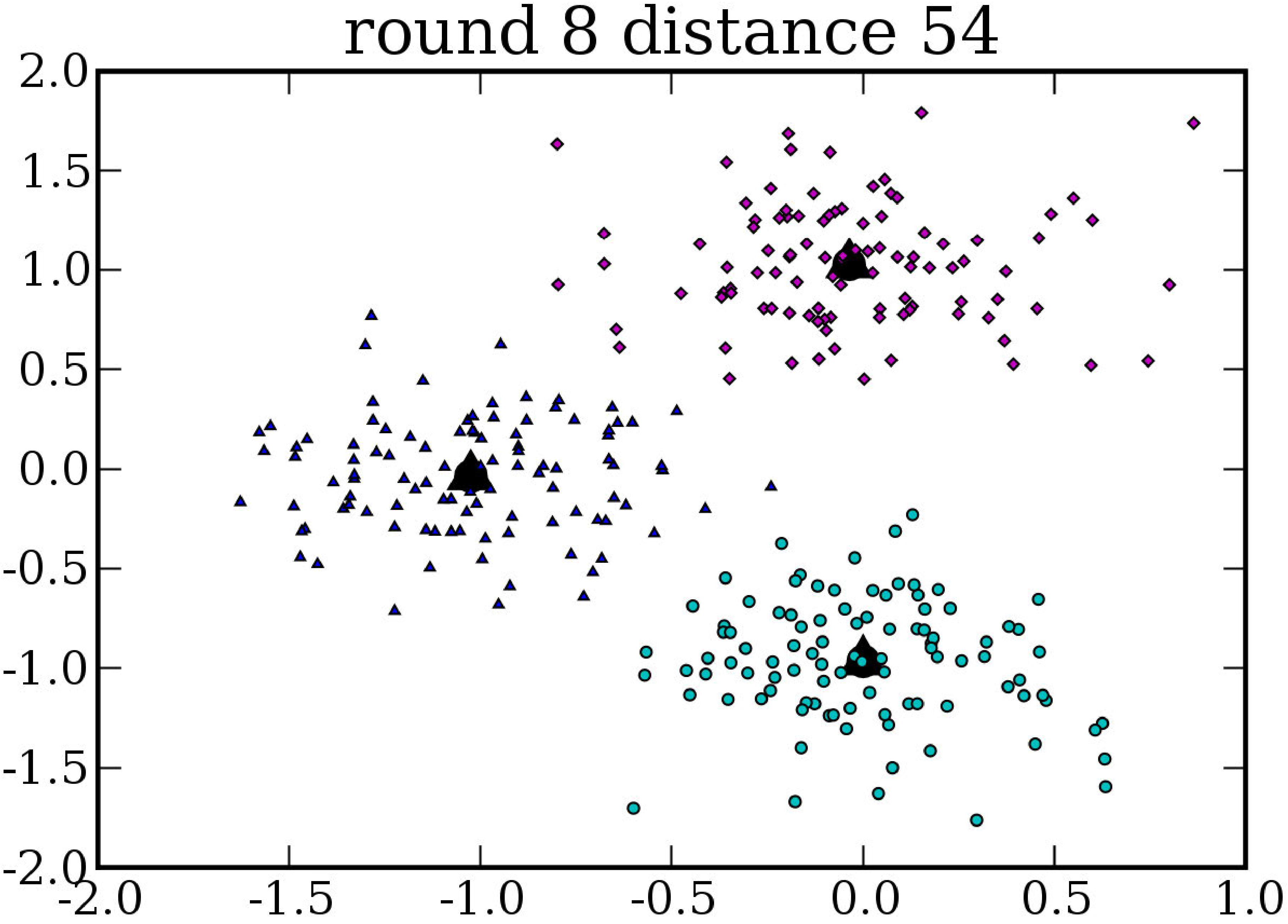

round 4 distance 136
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0



round 5 distance 114


round 6 distance 67
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

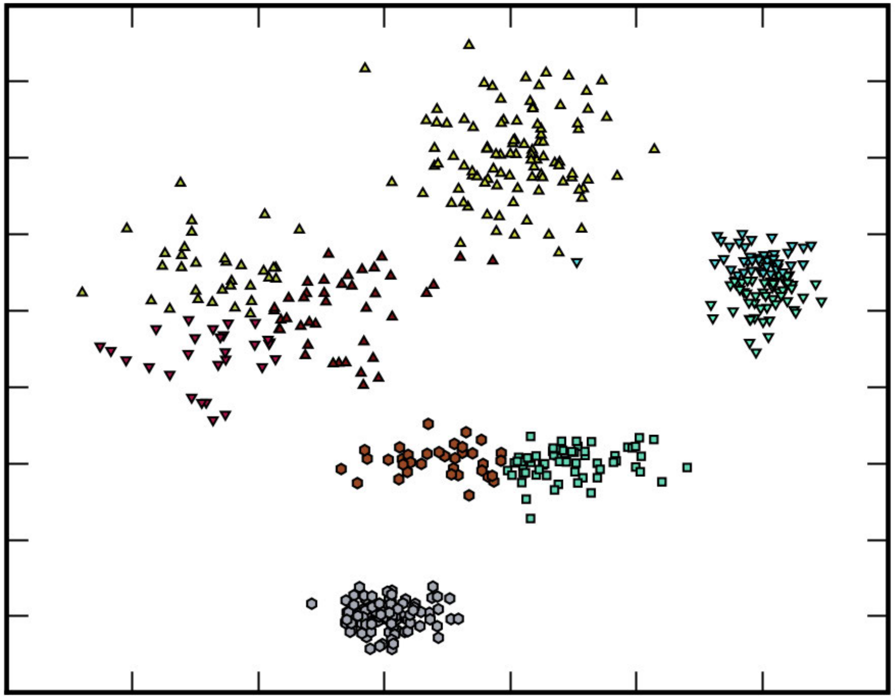

round 7 distance 54
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0


round 8 distance 54
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

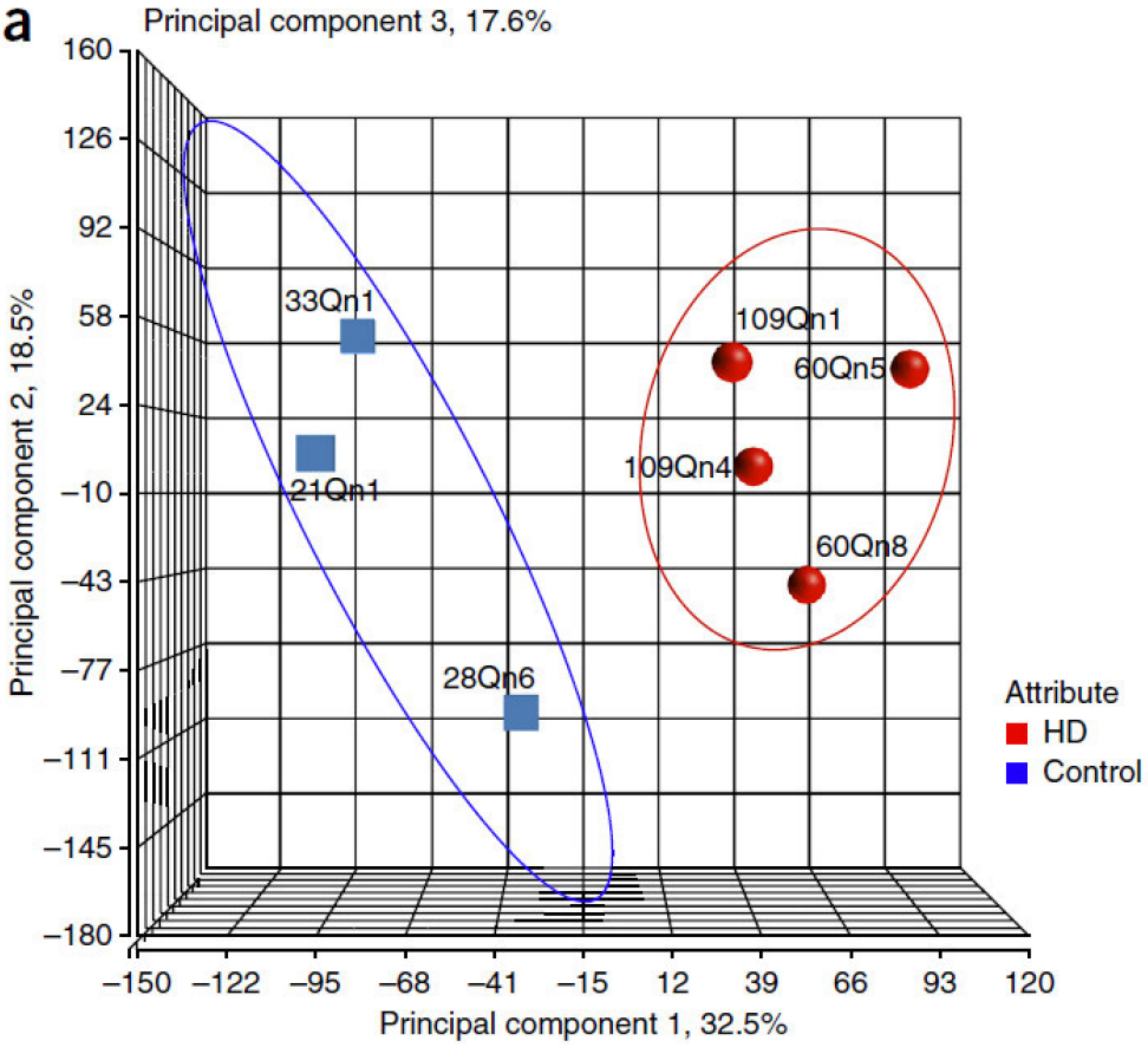
What if we choose pathologically bad initial positions?

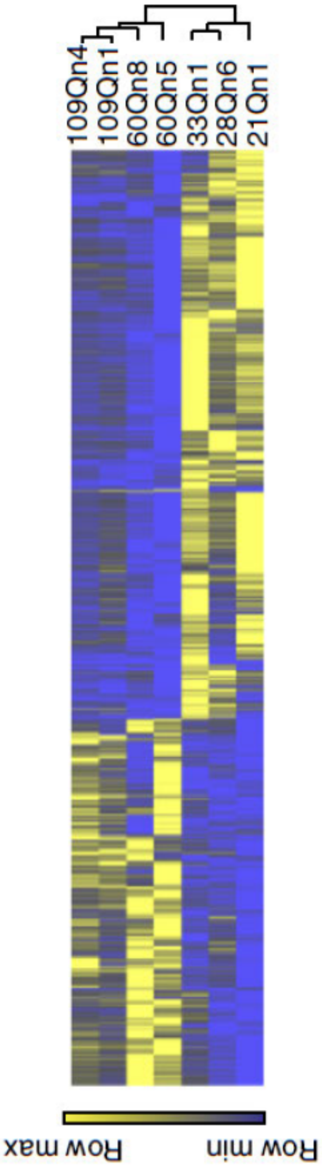
Often, the algorithm gets a reasonable answer, but not always!

How could you visualize clusters in 20,000D instead of 2D?



Principal Component Analysis



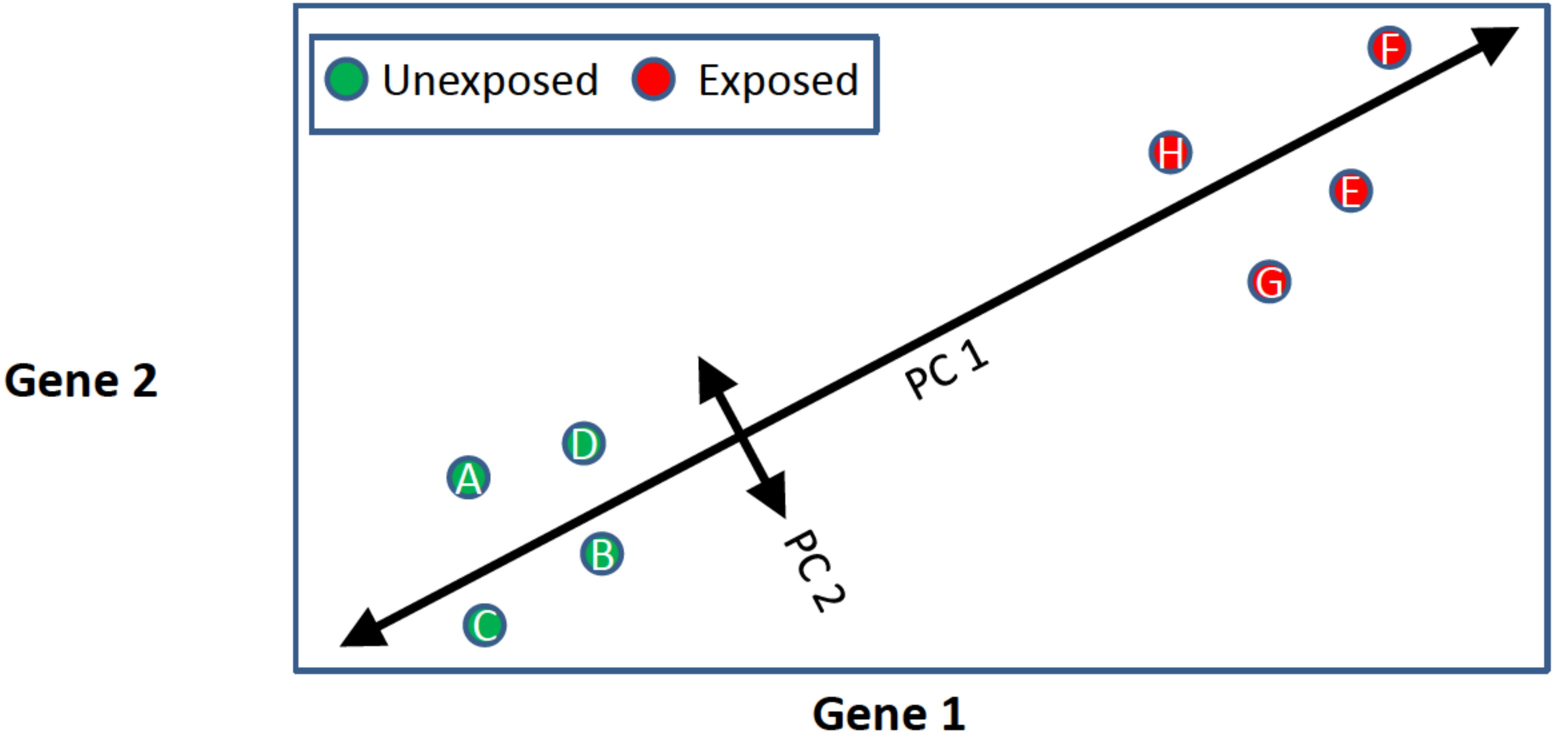


The basics of PCA

Principal Component Analysis

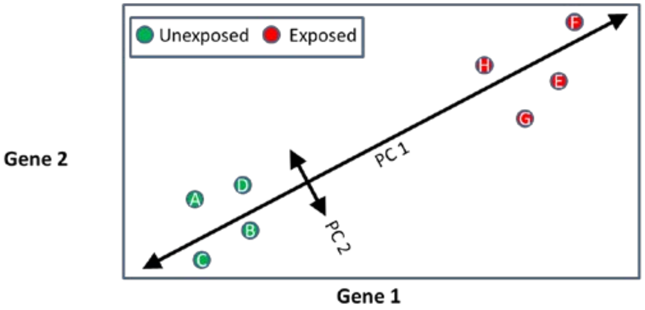
- Each sample is currently described by the expression of roughly 20,000 genes.
- Let’s imagine instead that we only had measured two genes
- If we wanted to compare samples, we could just plot the expression values like this:

Principal Component Analysis



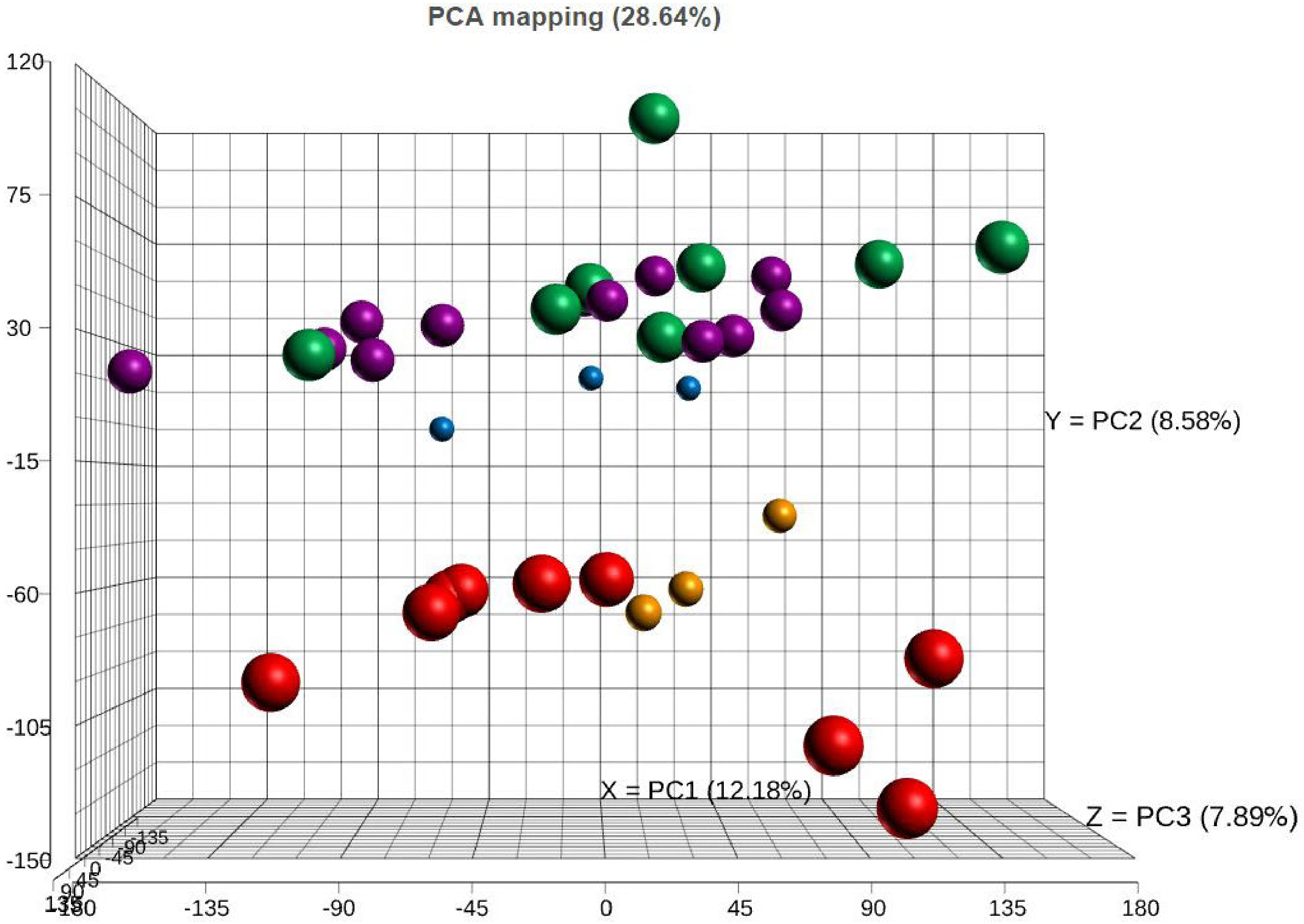
Goal: find a linear combination of the axes that captures most of the variation

Principal Component Analysis

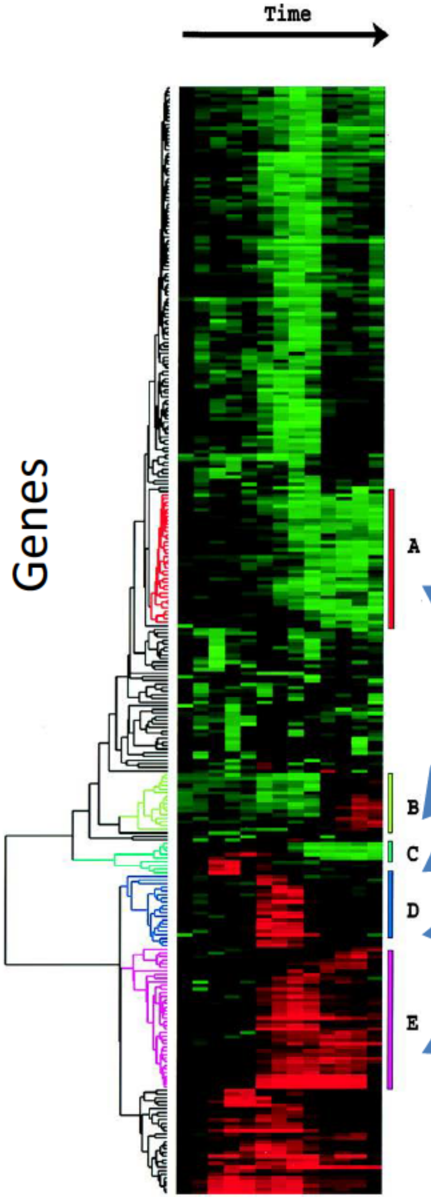


1. Here, Gene 1 and Gene 2 are equally good at explaining the variance in the data.
2. The big arrow indicates a linear combination of G1 and G2 that represents the direction of maximal variance. This is called PC 1 (Principal Component 1)
3. PCA lets us find such linear combinations even if there are thousands of variables.
4. There are as many PCs as there were dimensions in the original data.
5. The PCs are orthogonal.
6. Often, a few PCs will capture most of the variance. Here we can ignore PC2.

PCA can help spot patterns in the data



Next time: Interpreting your results



How did they figure out what the clusters of genes did?

GREAT SEMINAR TODAY

AT 4PM IN 32-141

TOWARD PERSONALIZED MEDICINE USING GUT MICROBIOME AND CLINICAL DATA

(A) cholesterol biosynthesis

(B) the cell cycle

(C) the immediate-early response

(D) signaling and angiogenesis

(E) wound healing and tissue remodeling

Iyer et al. *Science* 1999