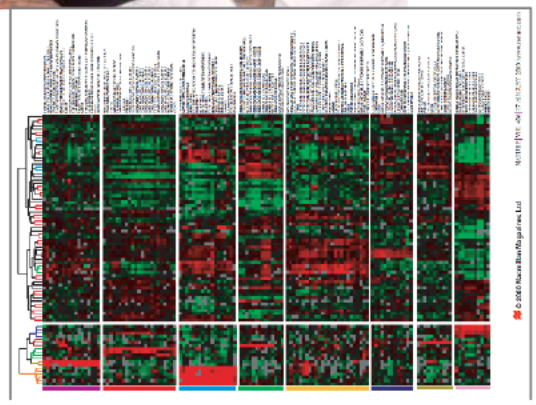
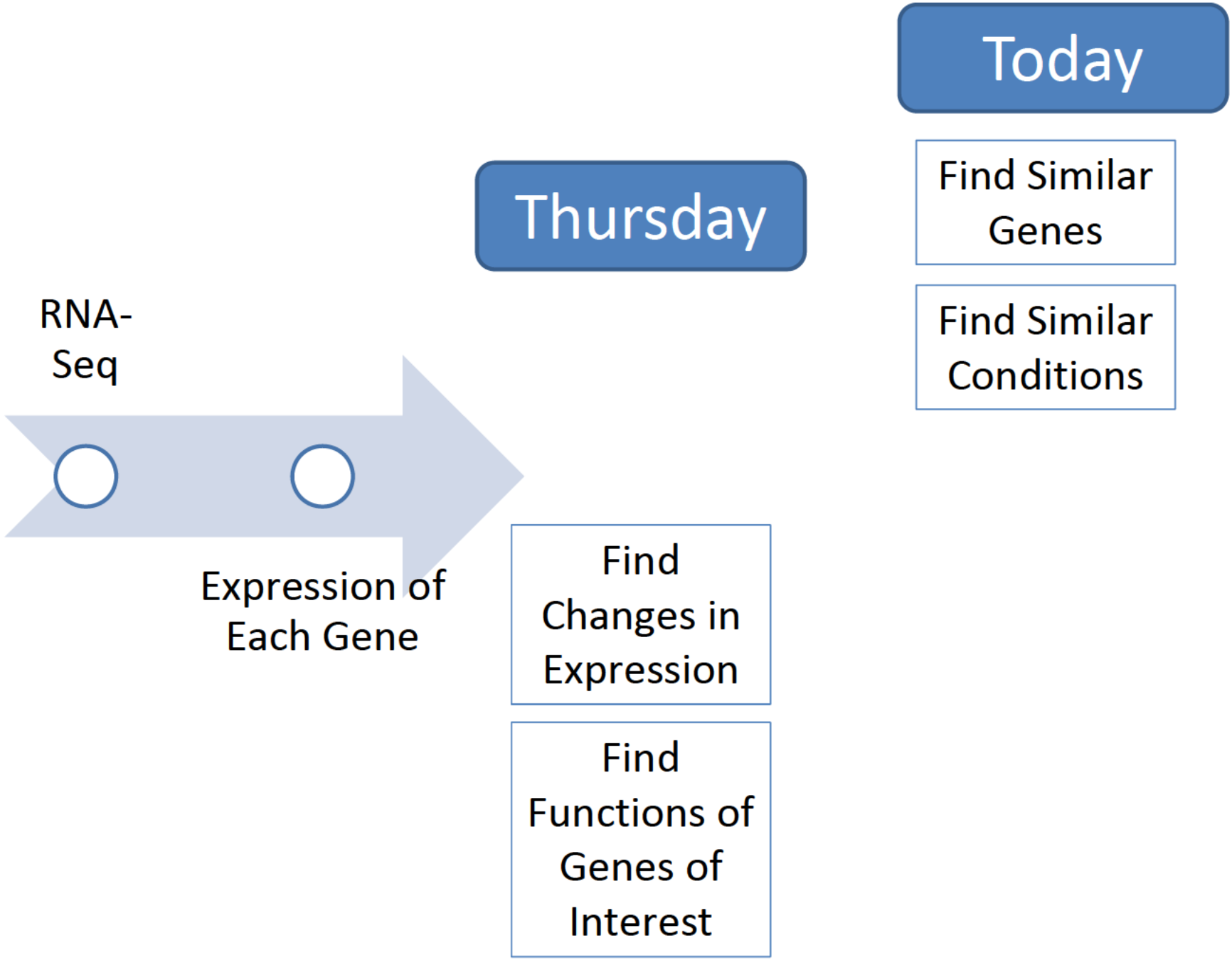
Classical Biology:
Driven by macroscopic observation

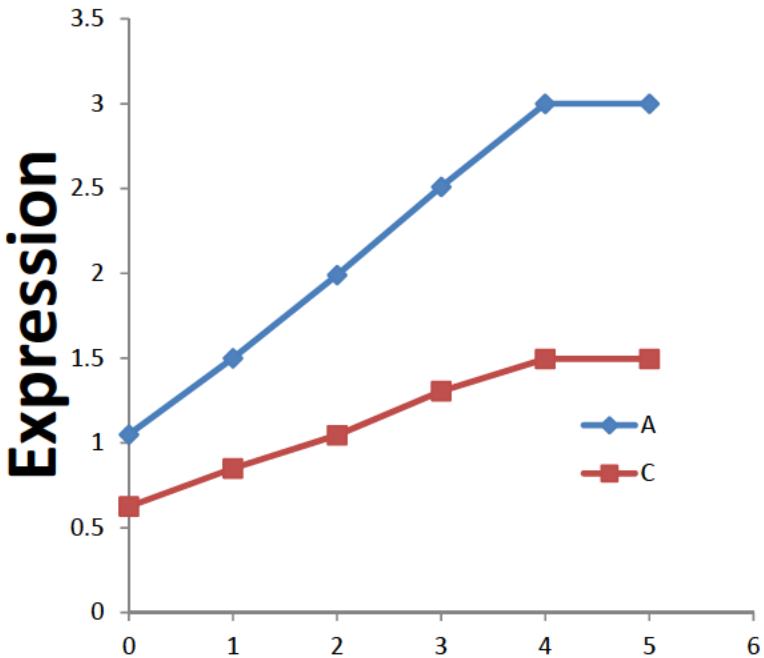
Molecular Biology:
Driven by simple hypotheses

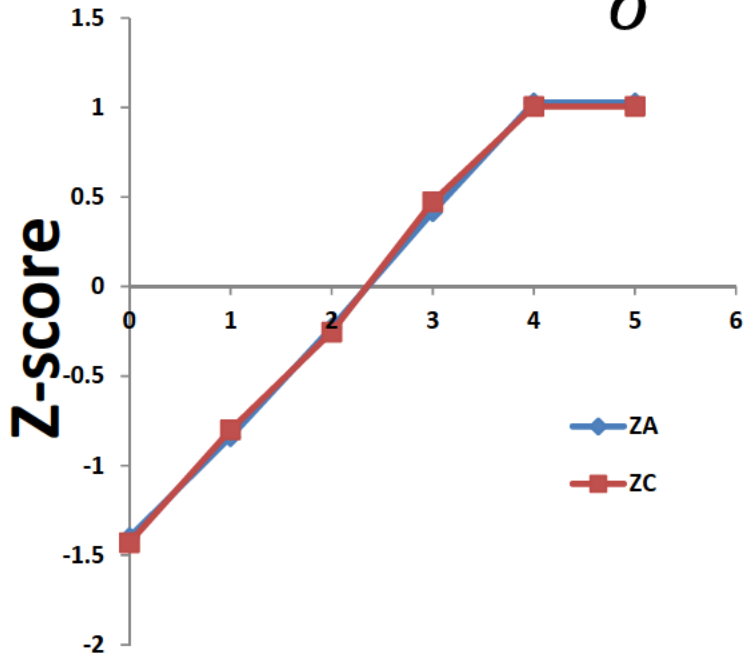


Systems Biology:
Driven by molecular data


Today
Find Similar Genes
Find Similar Conditions
Thursday
RNA-Seq
Expression of Each Gene
Find Changes in Expression
Find Functions of Genes of Interest

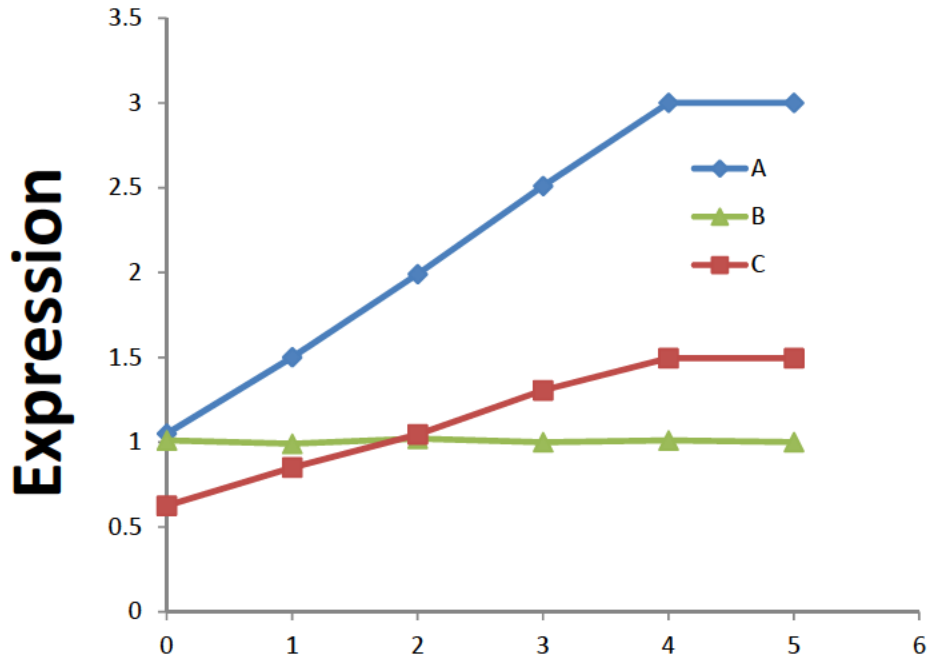
$$Z_{Ki} = \frac{X_{Ki} - \bar{X}_K}{\sigma}$$

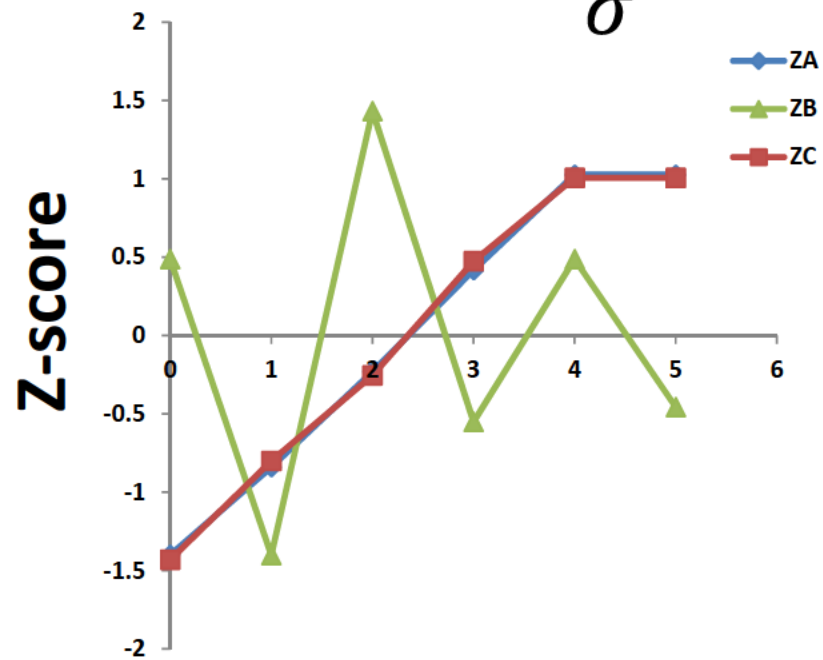




$$r_{A,B} = \frac{\sum_{k=1}^{N_{expt}} Z_{kA} Z_{kB}}{N}$$

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





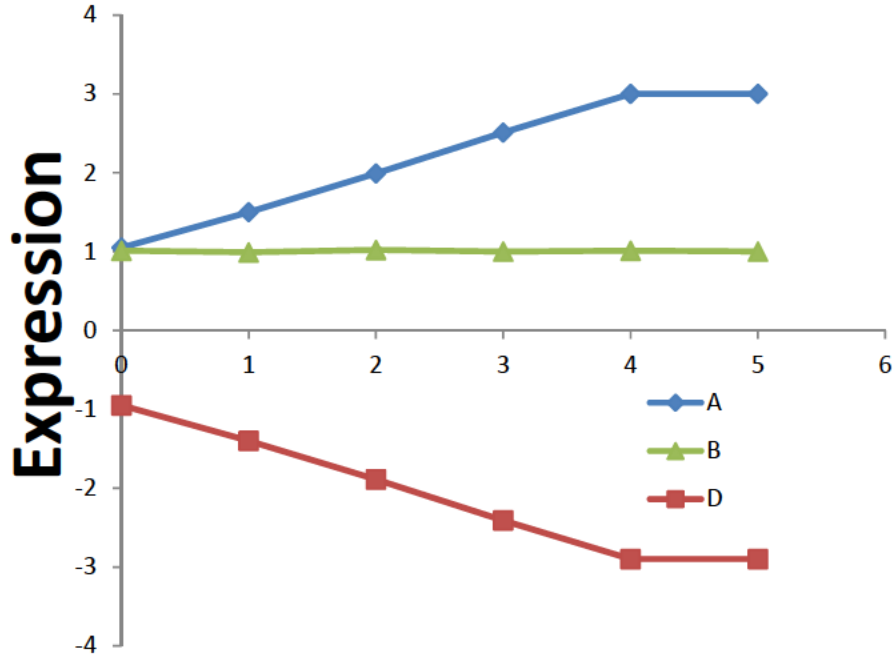
$r_{A,B}$= -0.01

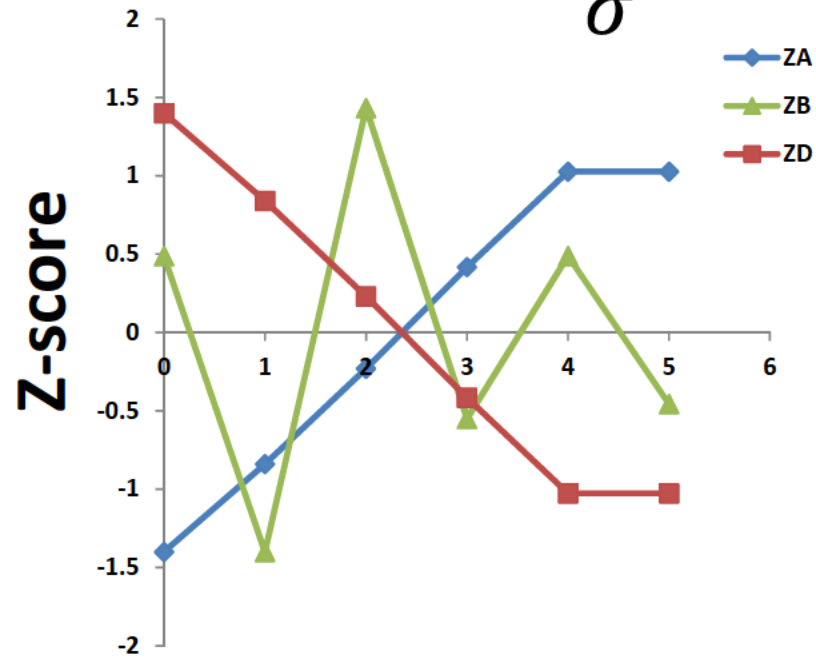
$r_{A,C}$= 0.999

$r_{B,C}$= -0.03

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

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





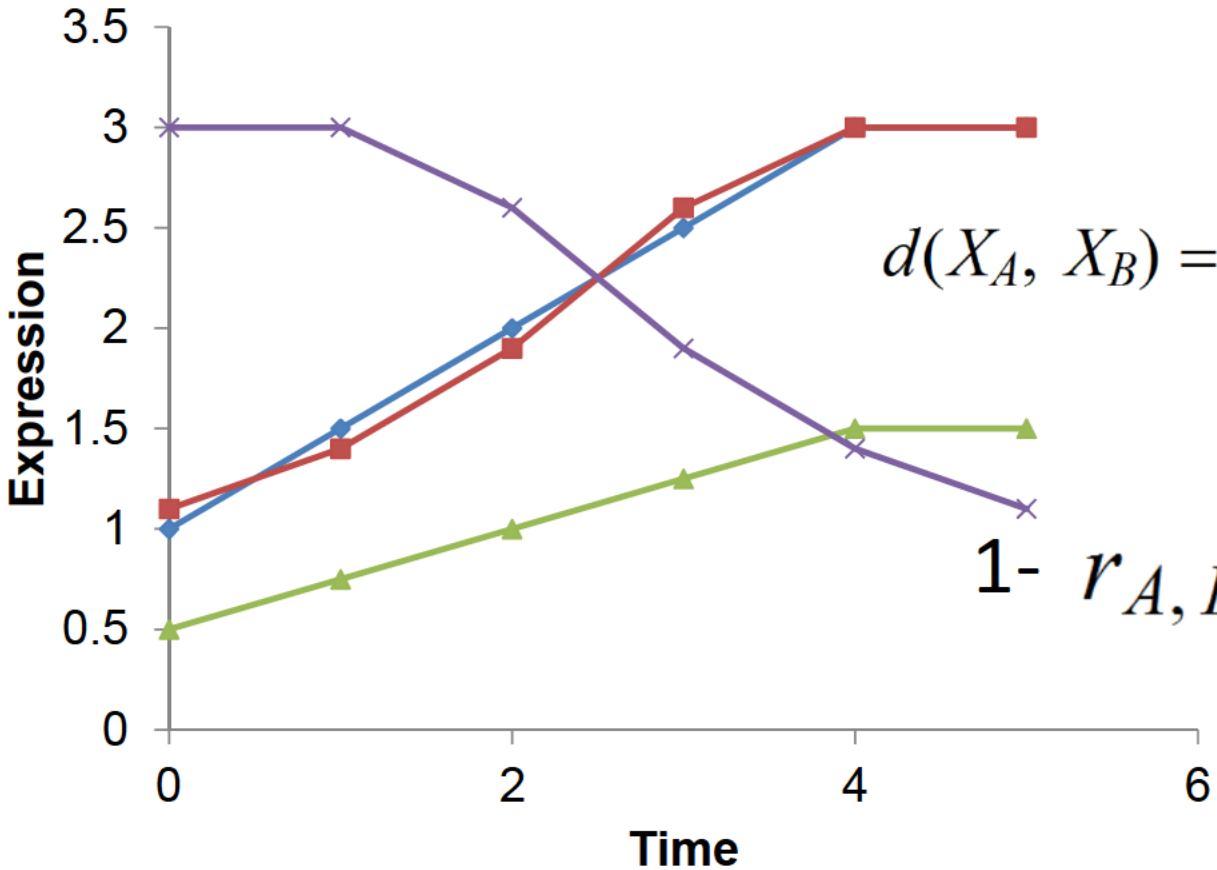
$r_{A,B}$= -0.01

$r_{A,D}$= -1.0

$r_{B,D}$= 0.007

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

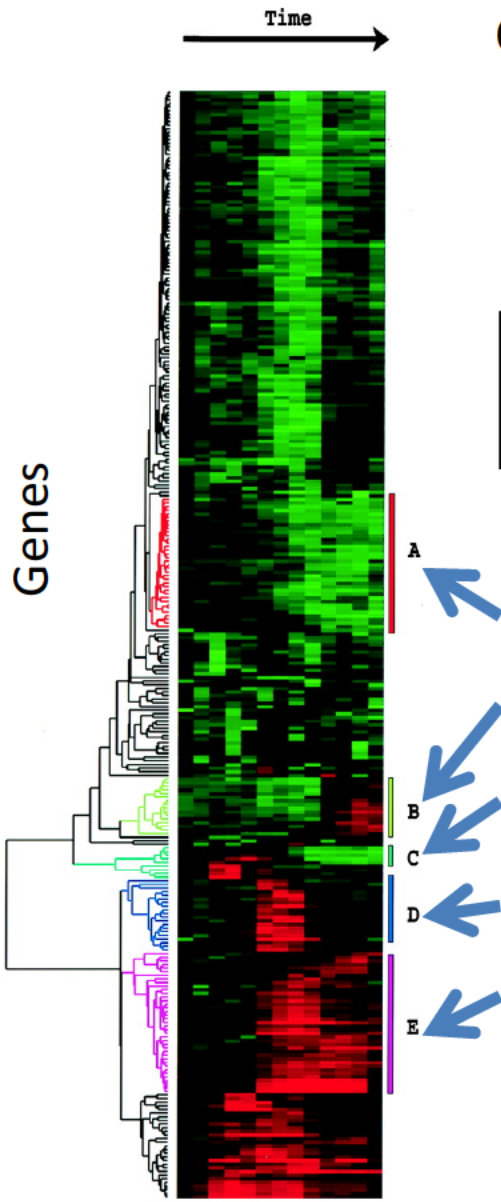
Distance Metrics



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

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

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)

(A) cholesterol biosynthesis

(B) the cell cycle

(C) the immediate-early response

(D) signaling and angiogenesis

(E) wound healing and tissue remodeling

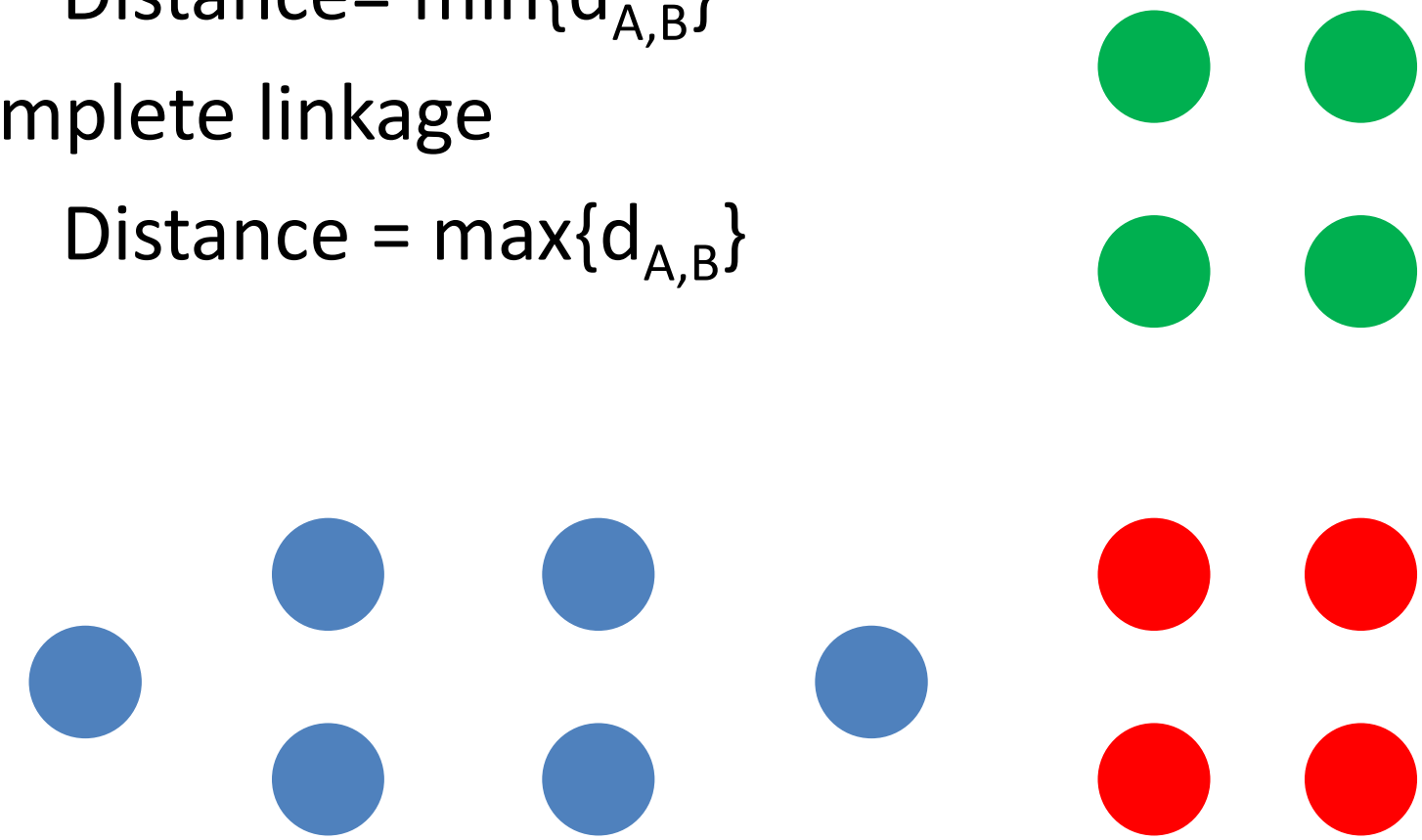
Iyer et al. *Science* 1999

- Single linkage

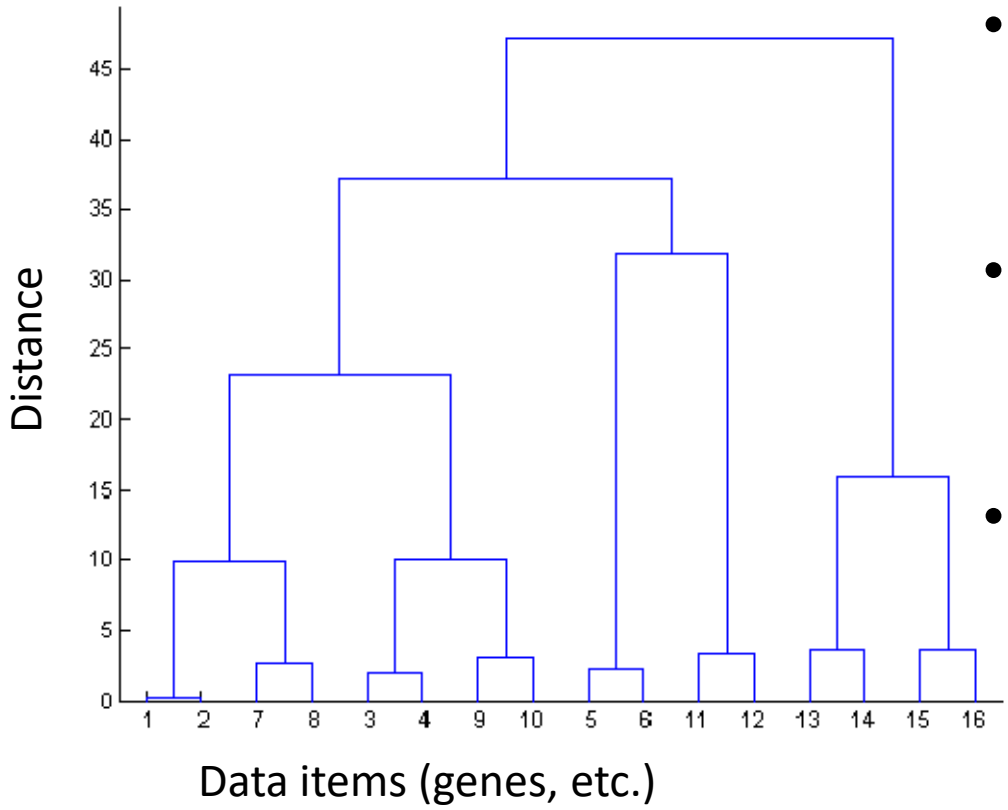
  Distance= $\min\{d_{A,B}\}$

- Complete linkage

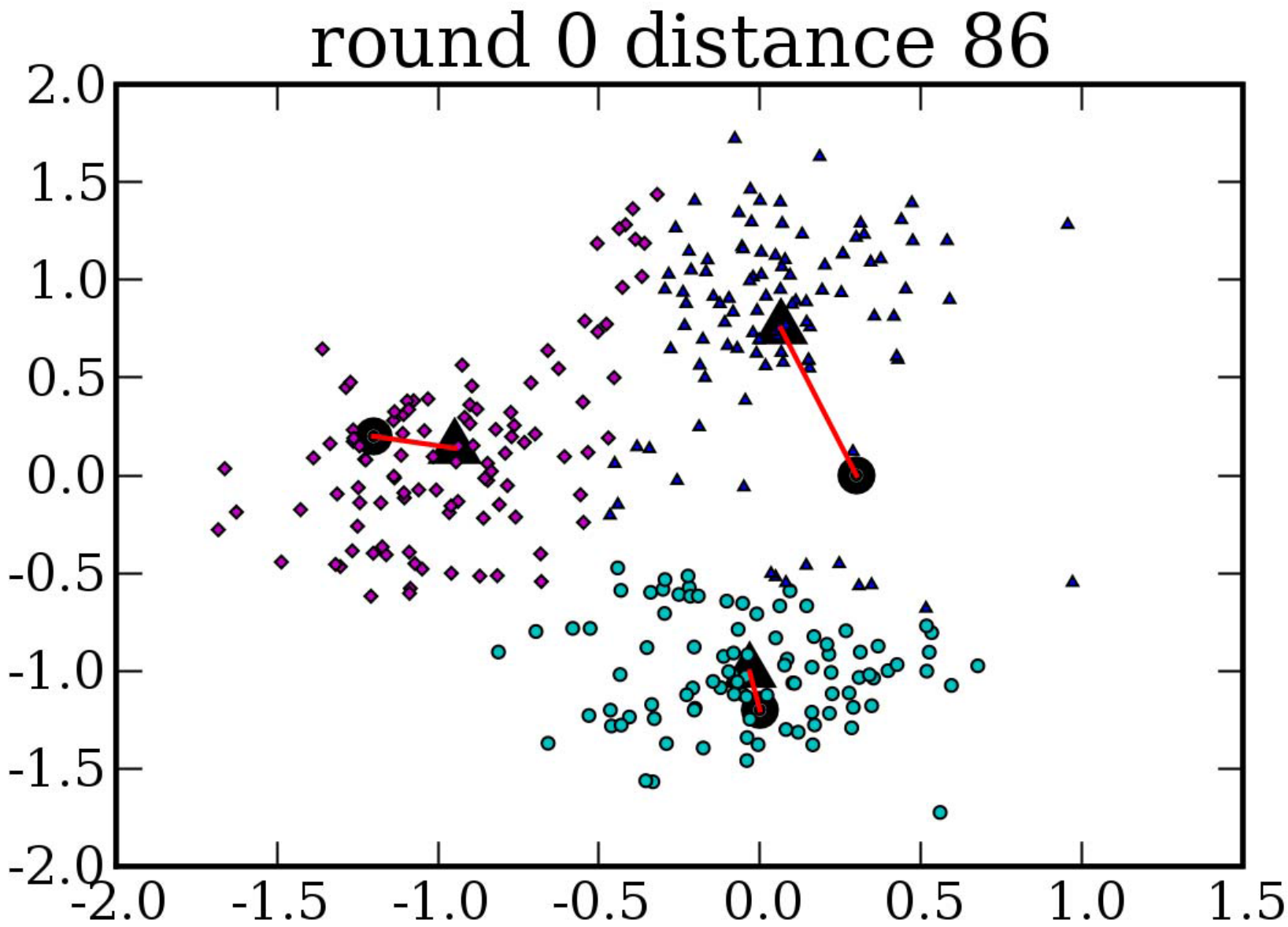
  Distance = $\max\{d_{A,B}\}$

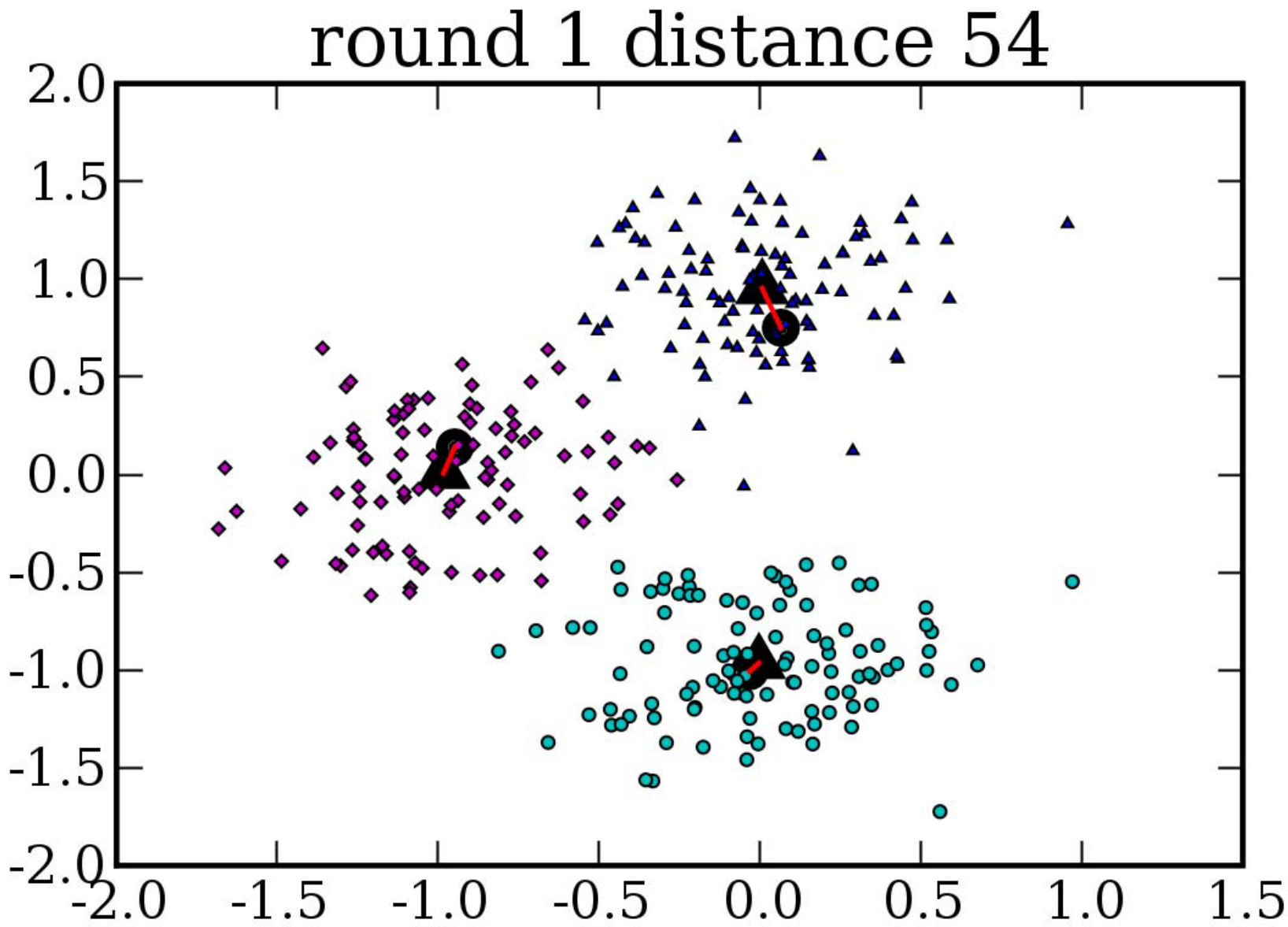


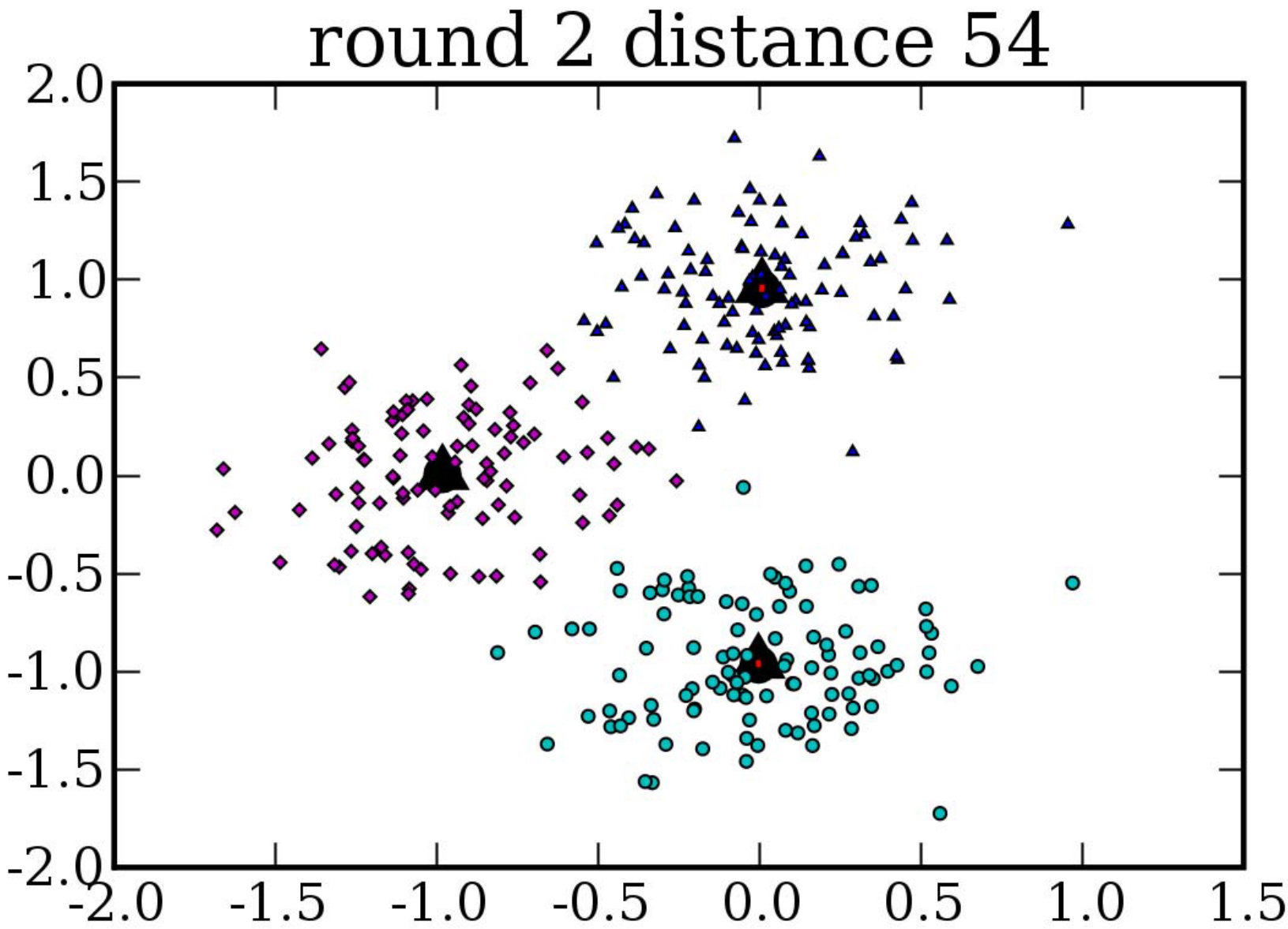
Dendograms



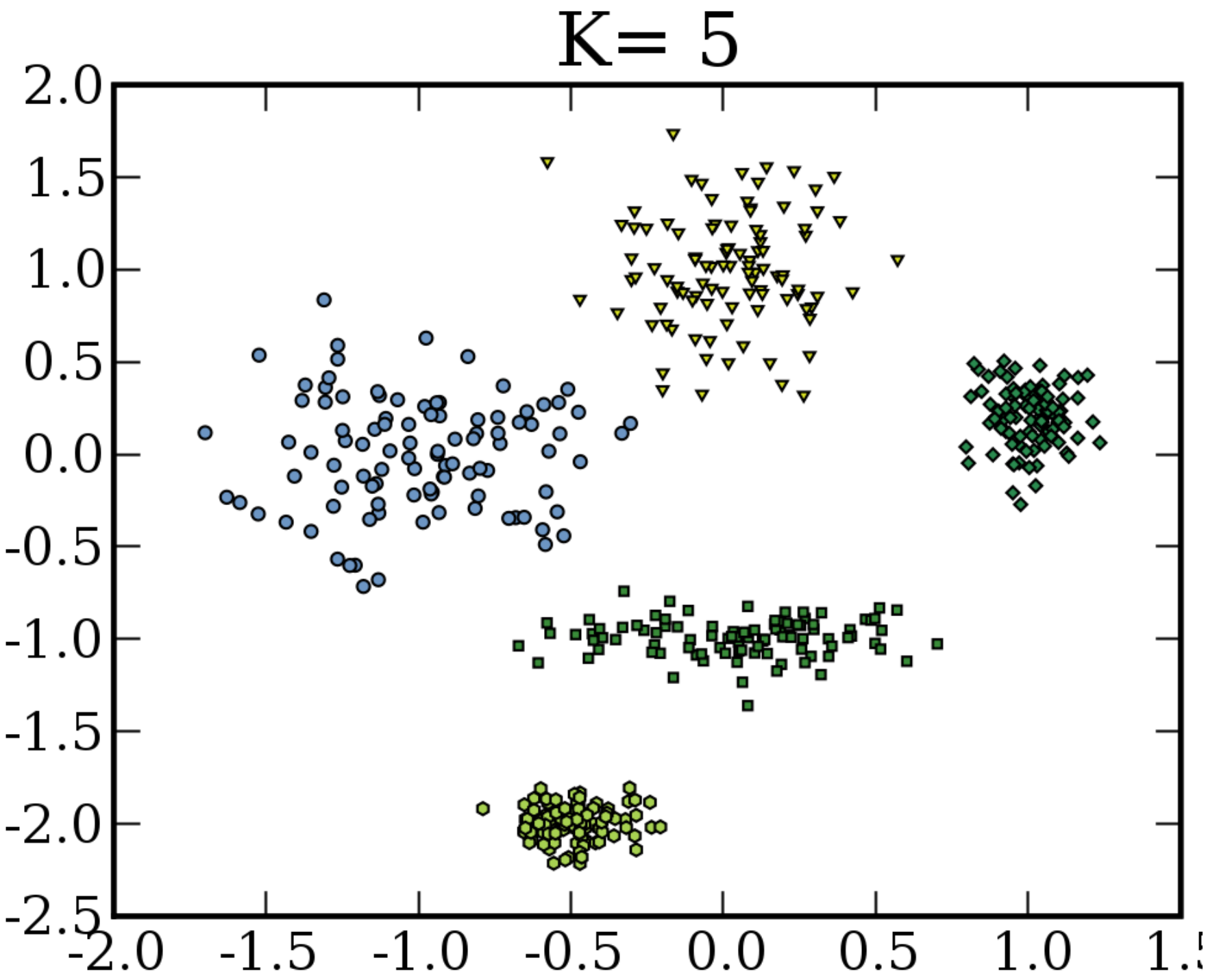
- The final cluster is the root and each data item is a leaf
- The heights of the bars indicate how close the items are
- Can ‘slice’ the tree at any distance cutoff to produce discrete clusters
- The results will always be hierarchical, even if the data are not.
- The order of the leaf nodes is not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

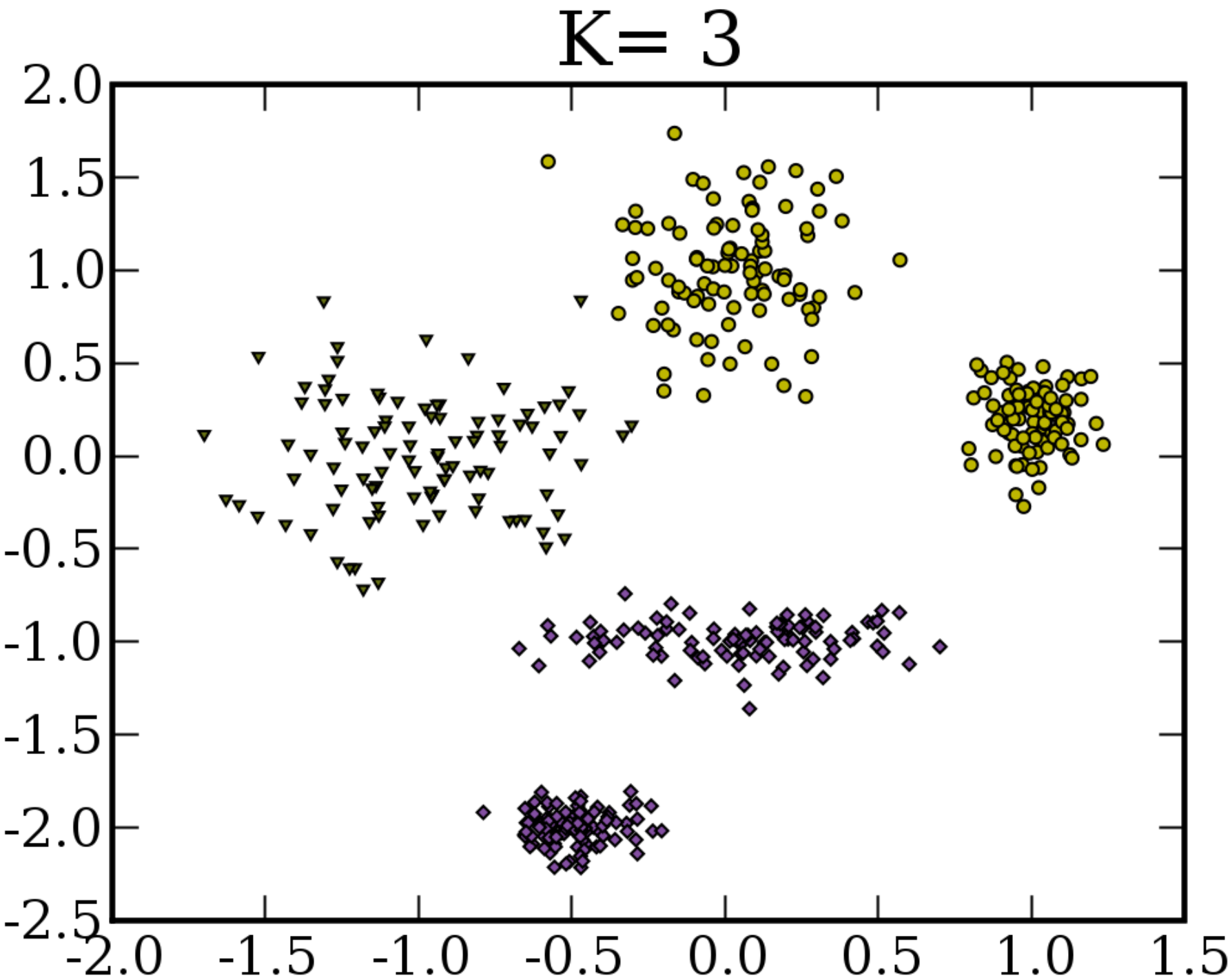

round 0 distance 86
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

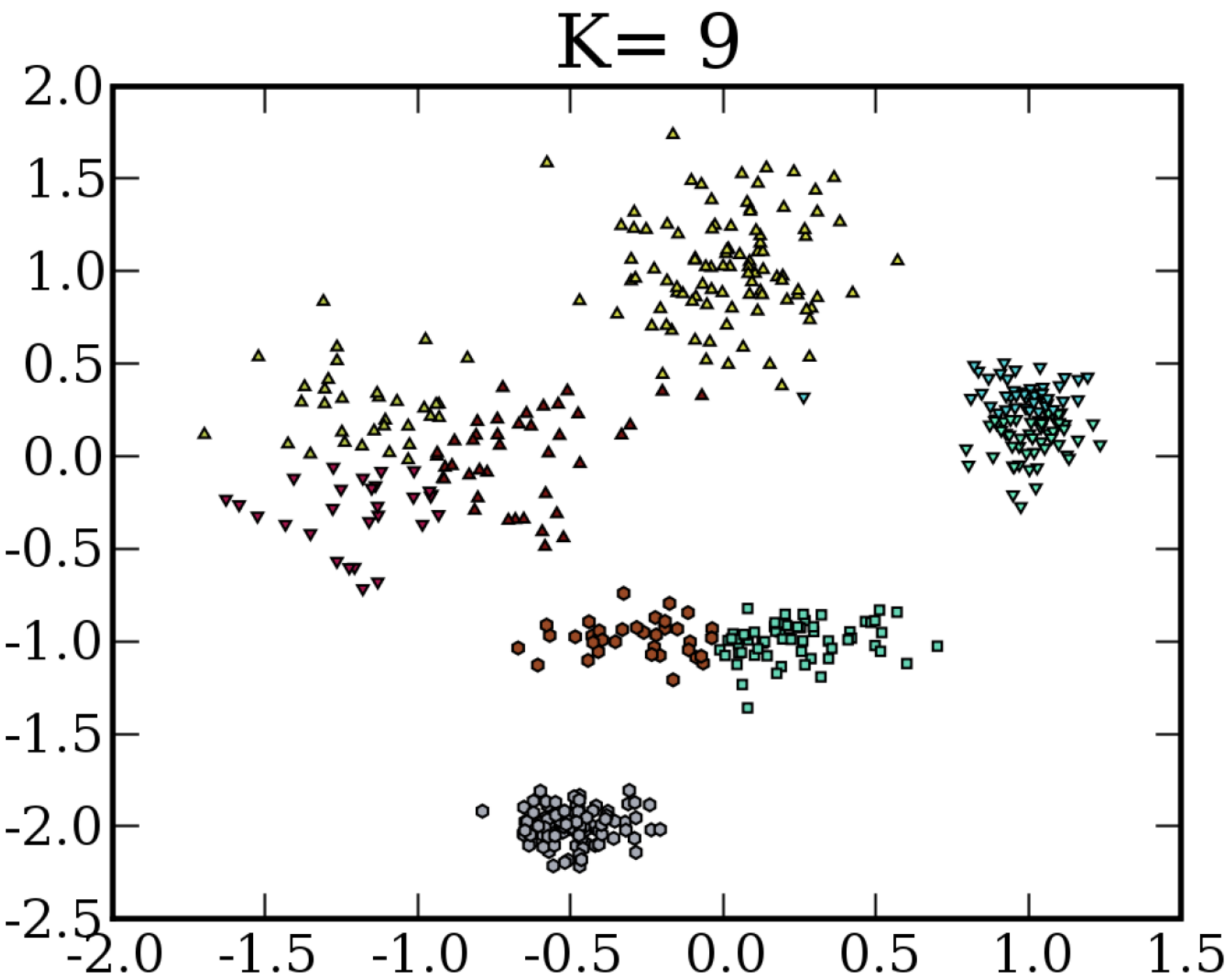

round 1 distance 54
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

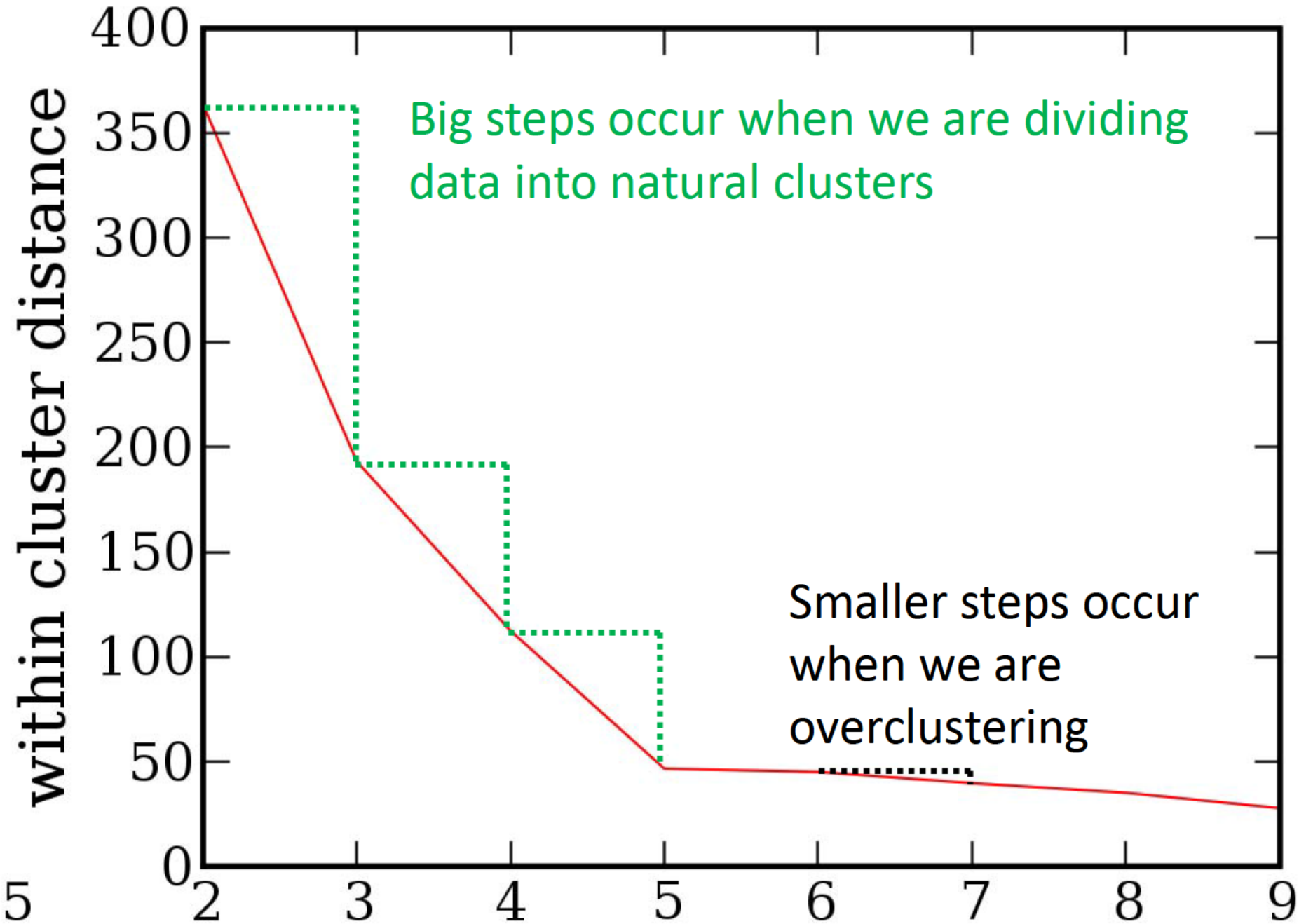

round 2 distance 54
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

What if you choose the wrong K?

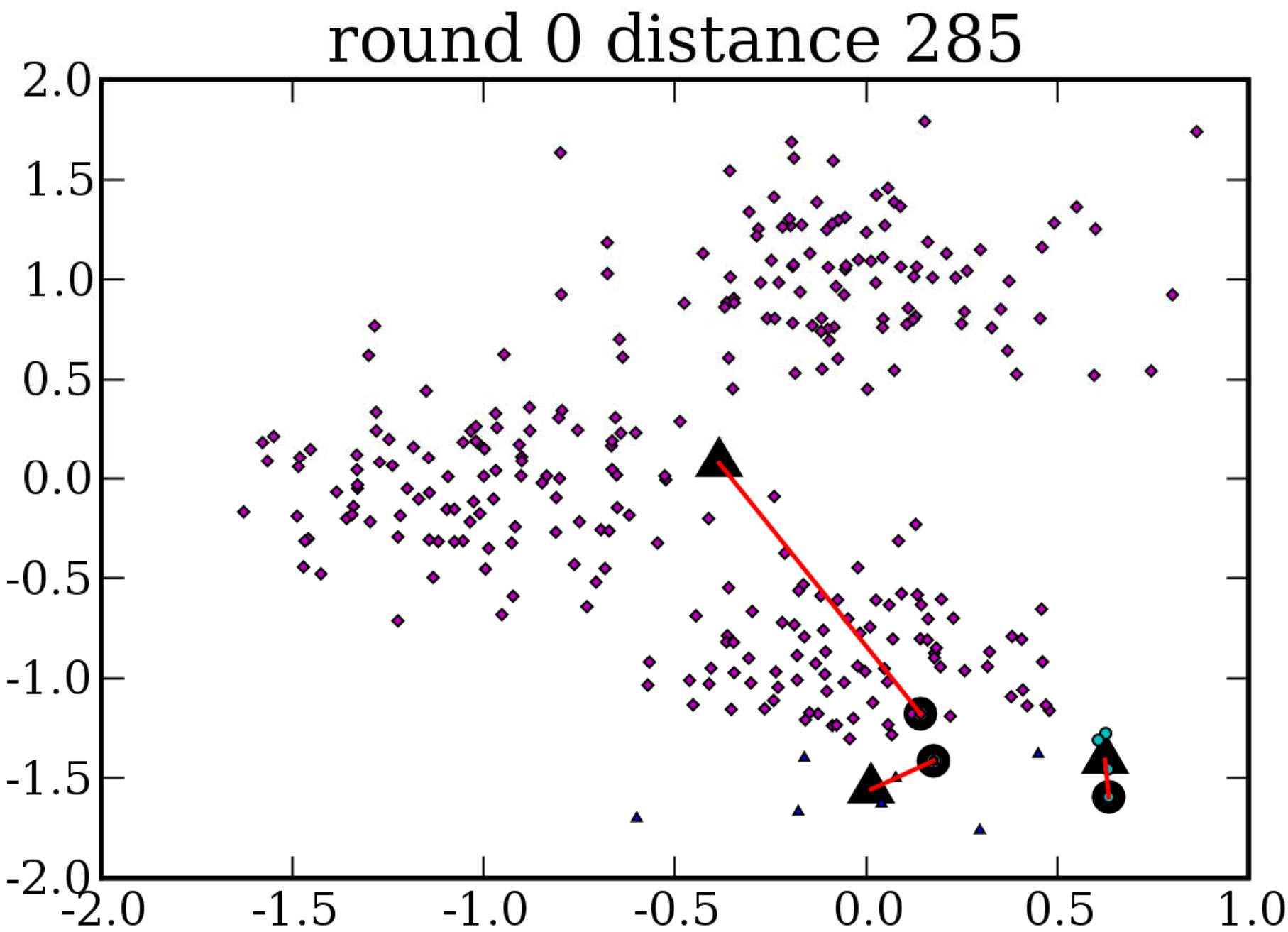

K= 5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

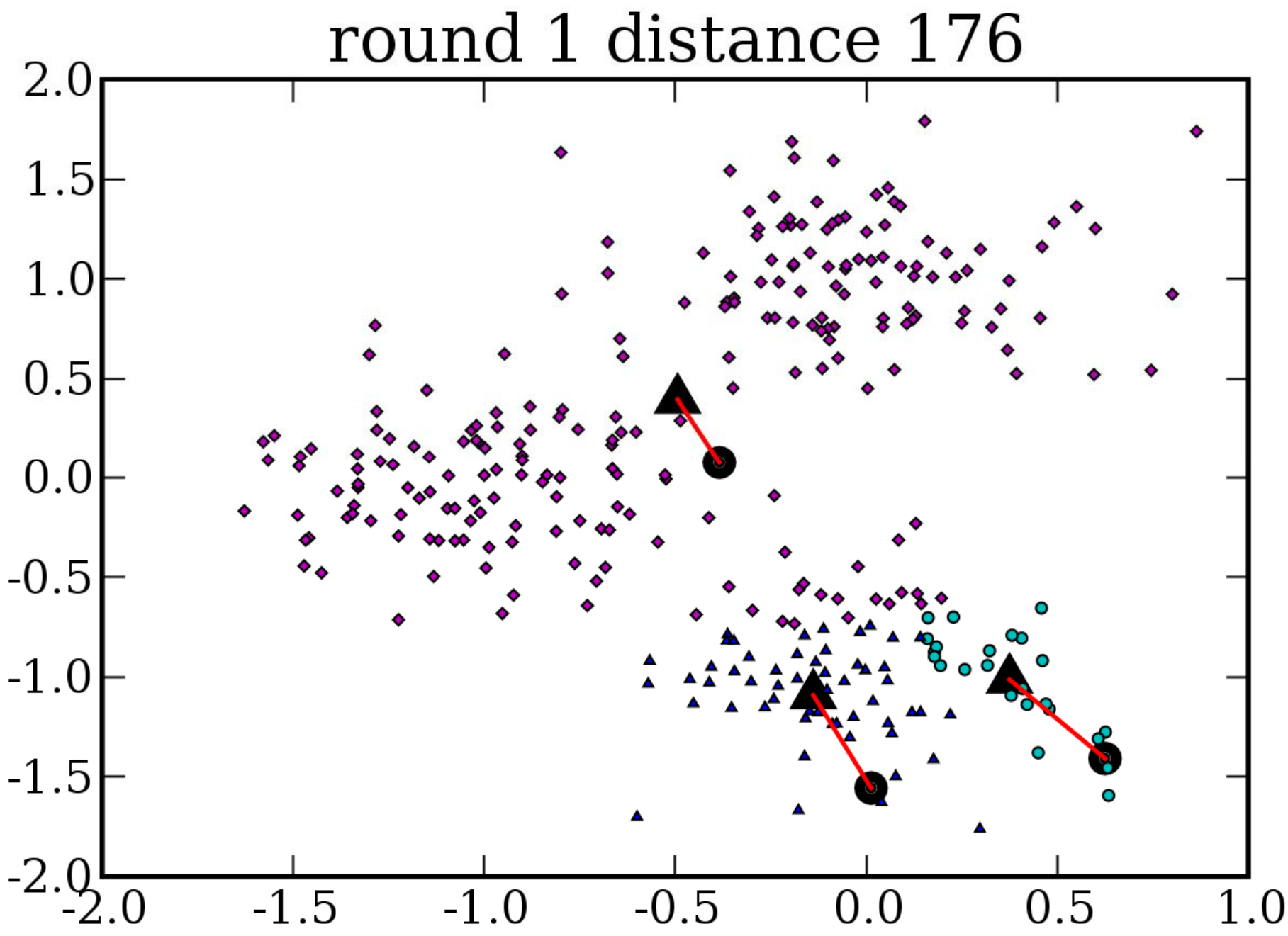

K= 3
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

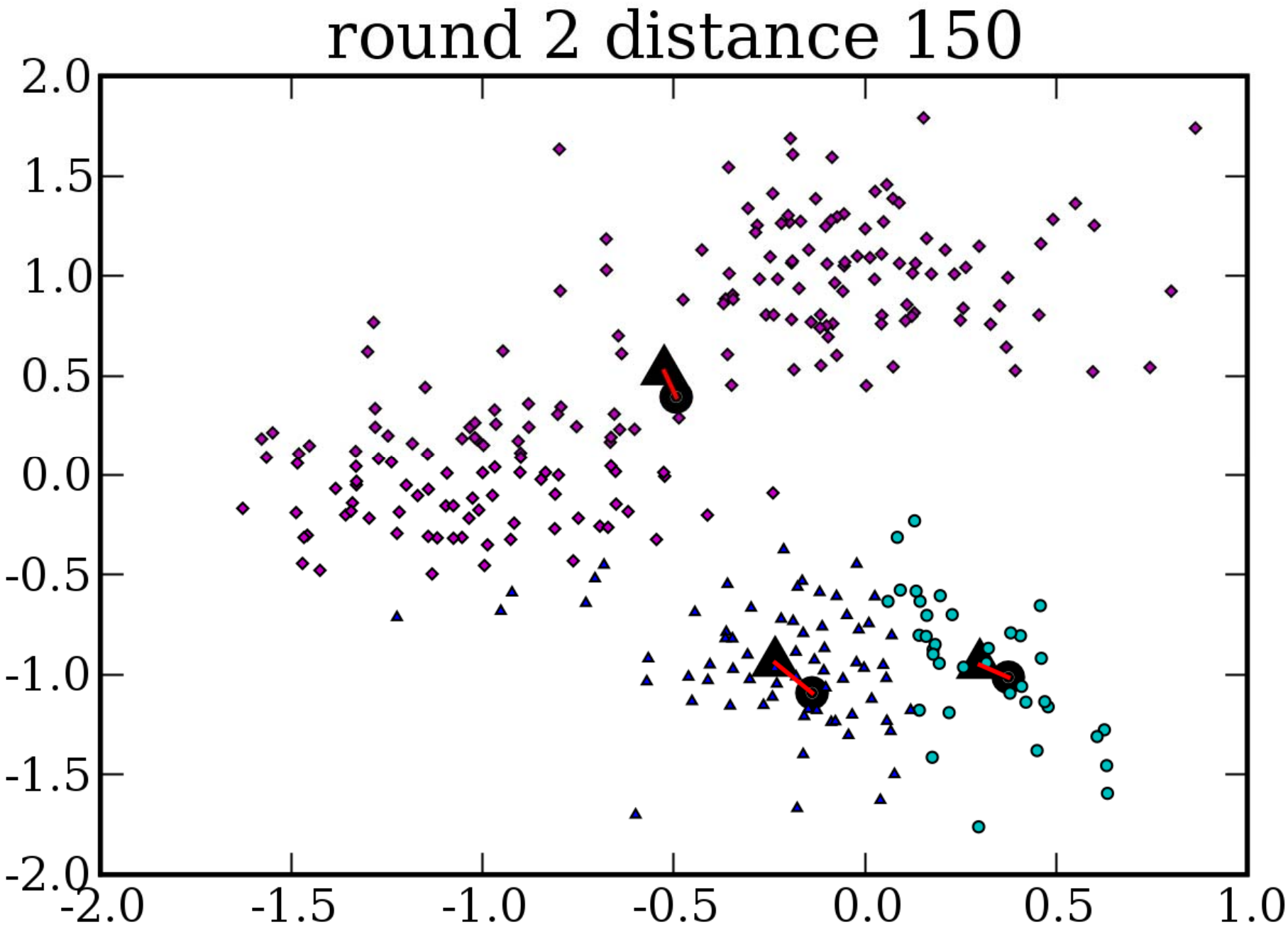

K= 9
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.5
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0
1.5

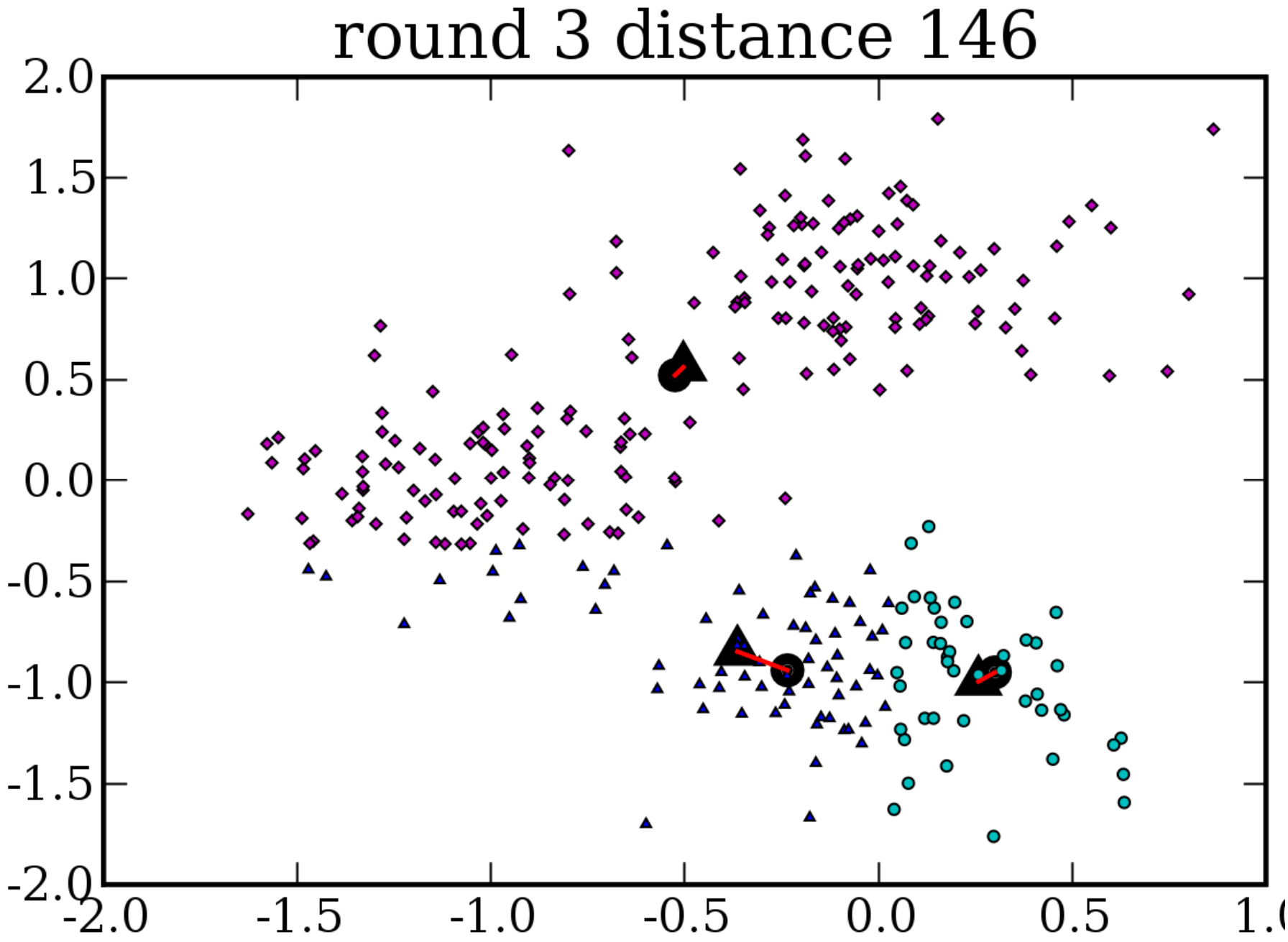

within cluster distance
400
350
300
250
200
150
100
50
0
2
3
4
5
6
7
8
9
Big steps occur when we are dividing data into natural clusters
Smaller steps occur when we are overclustering

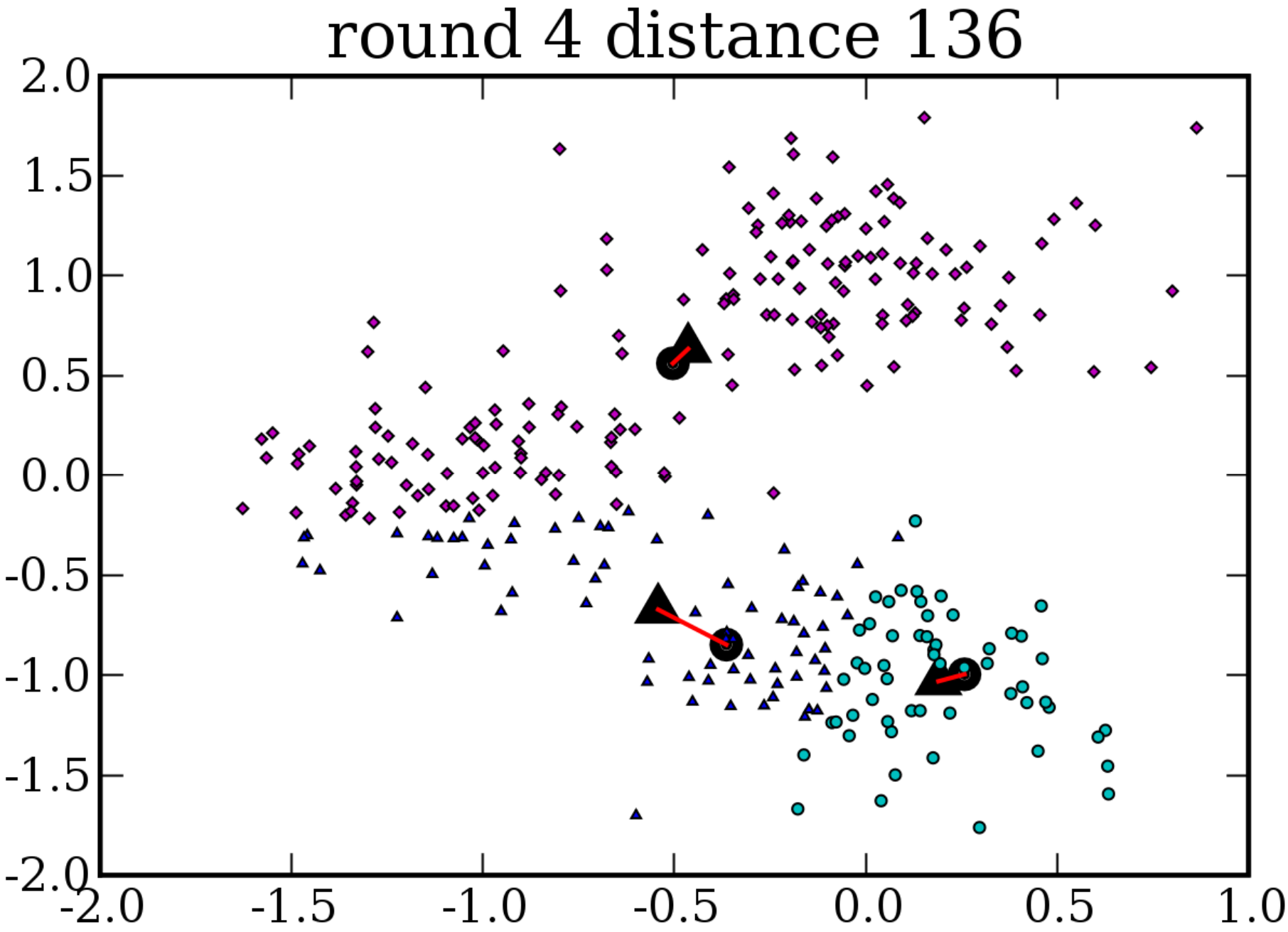
What if we choose pathologically bad initial positions?

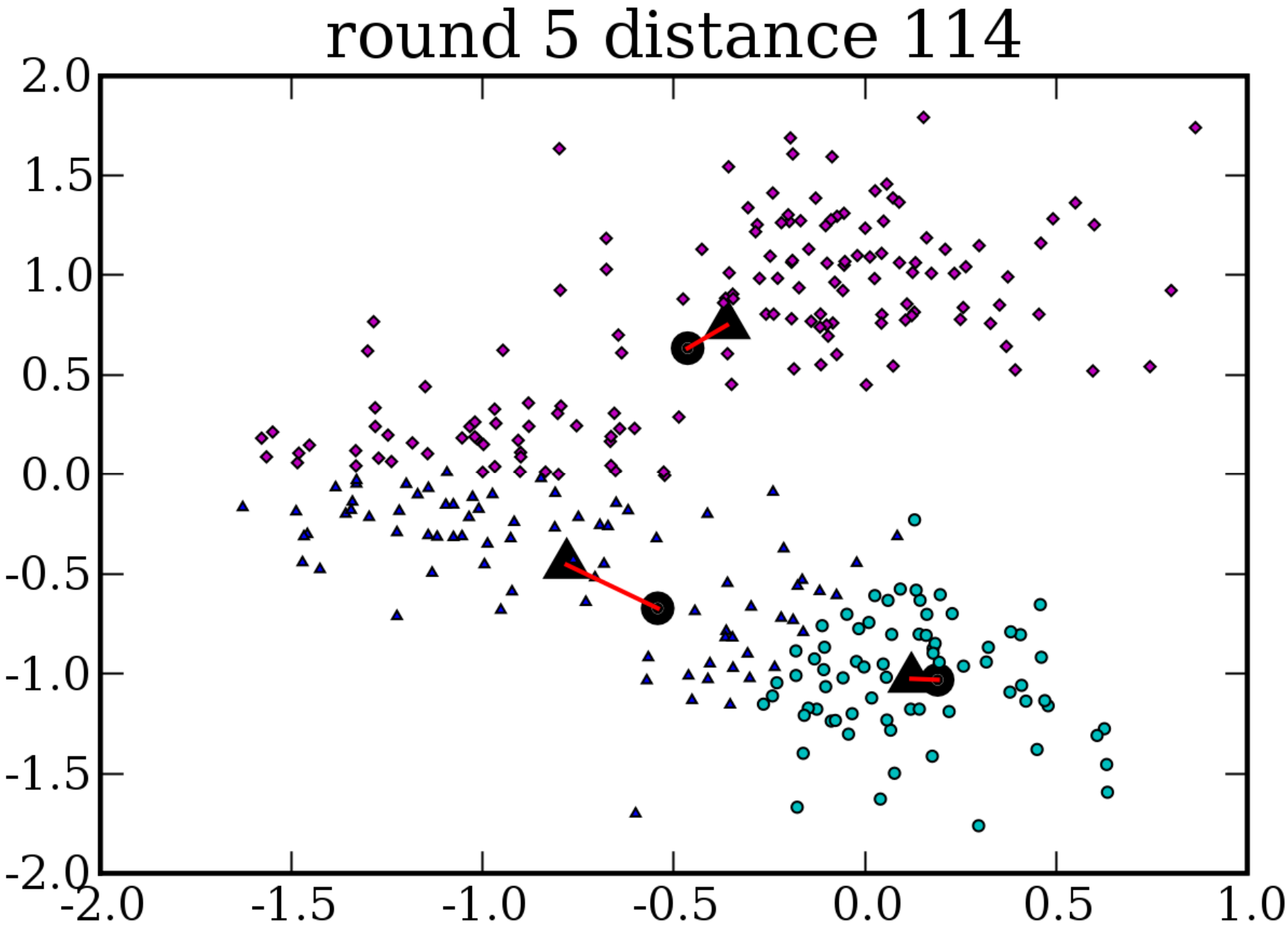

round 0 distance 285
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

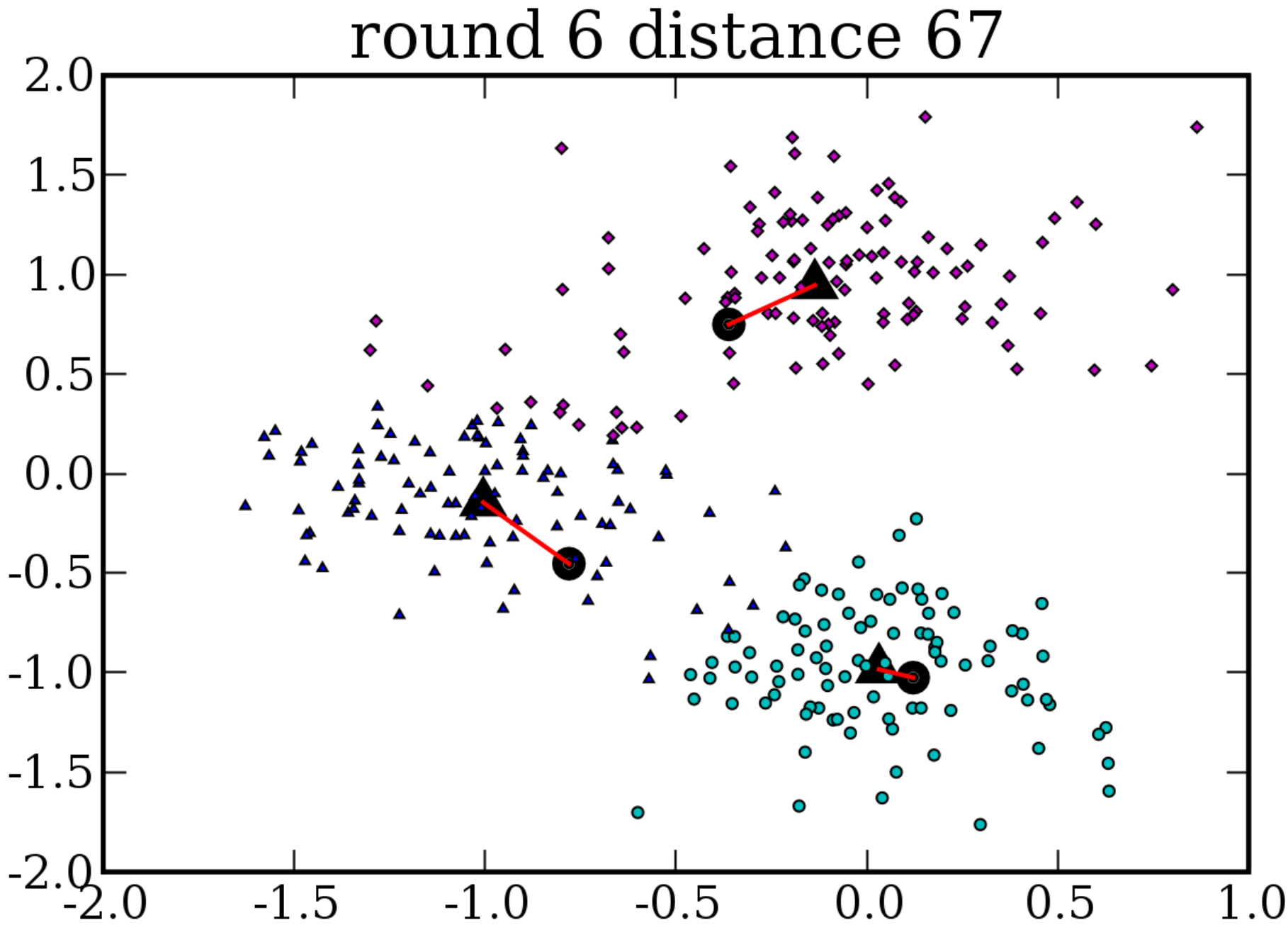

round 1 distance 176
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

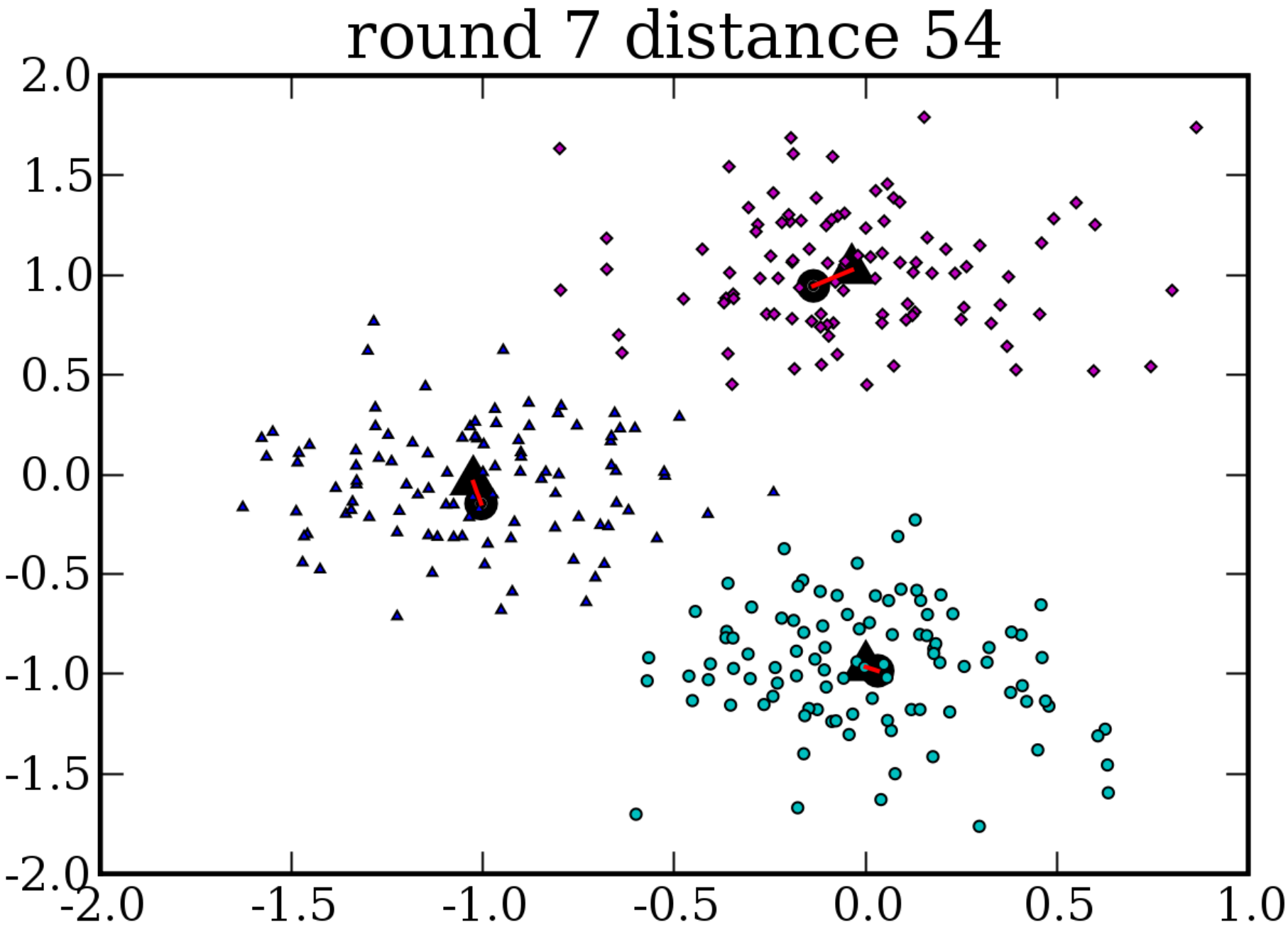

round 2 distance 150
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

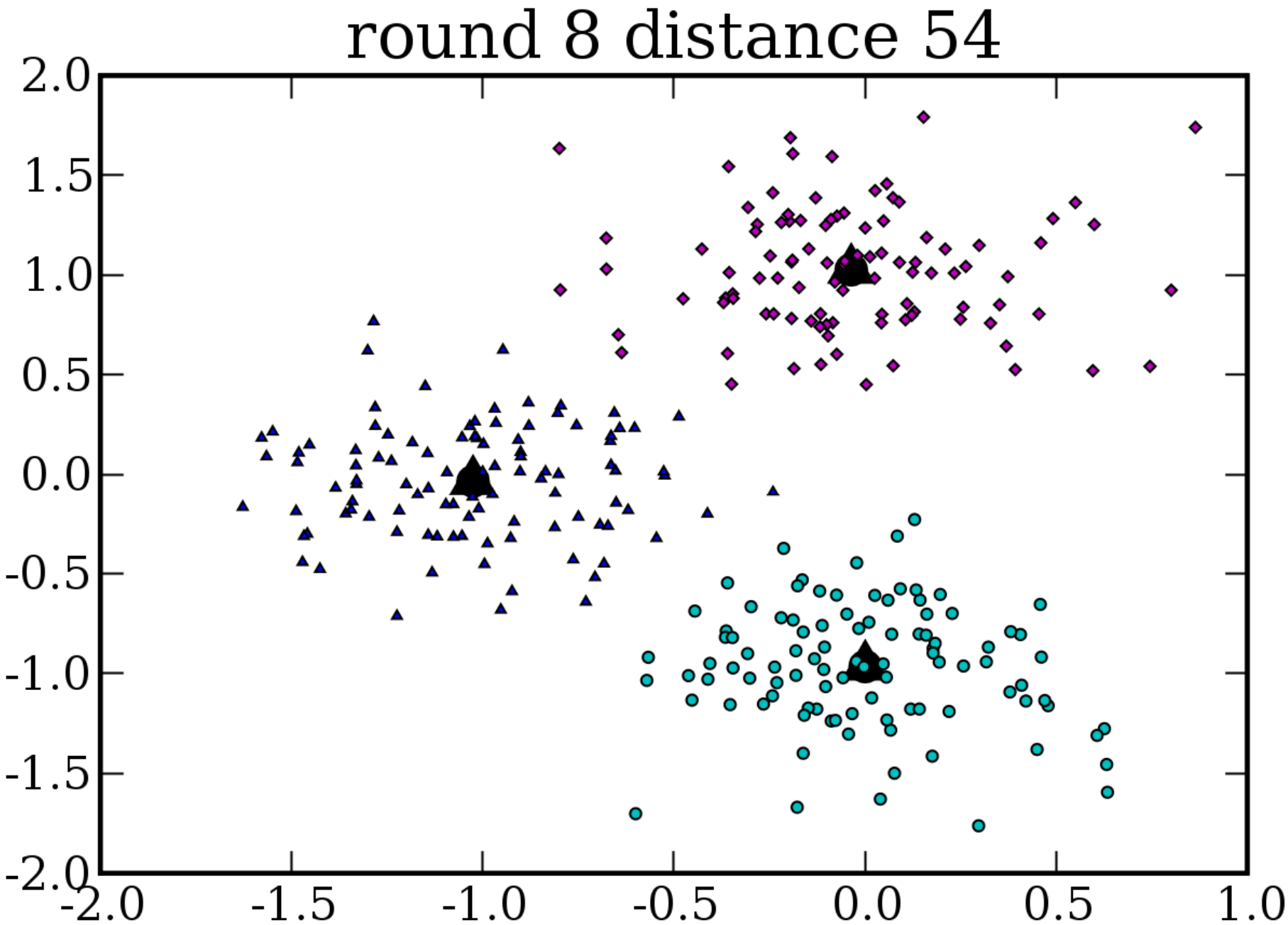

round 3 distance 146
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5


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
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 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

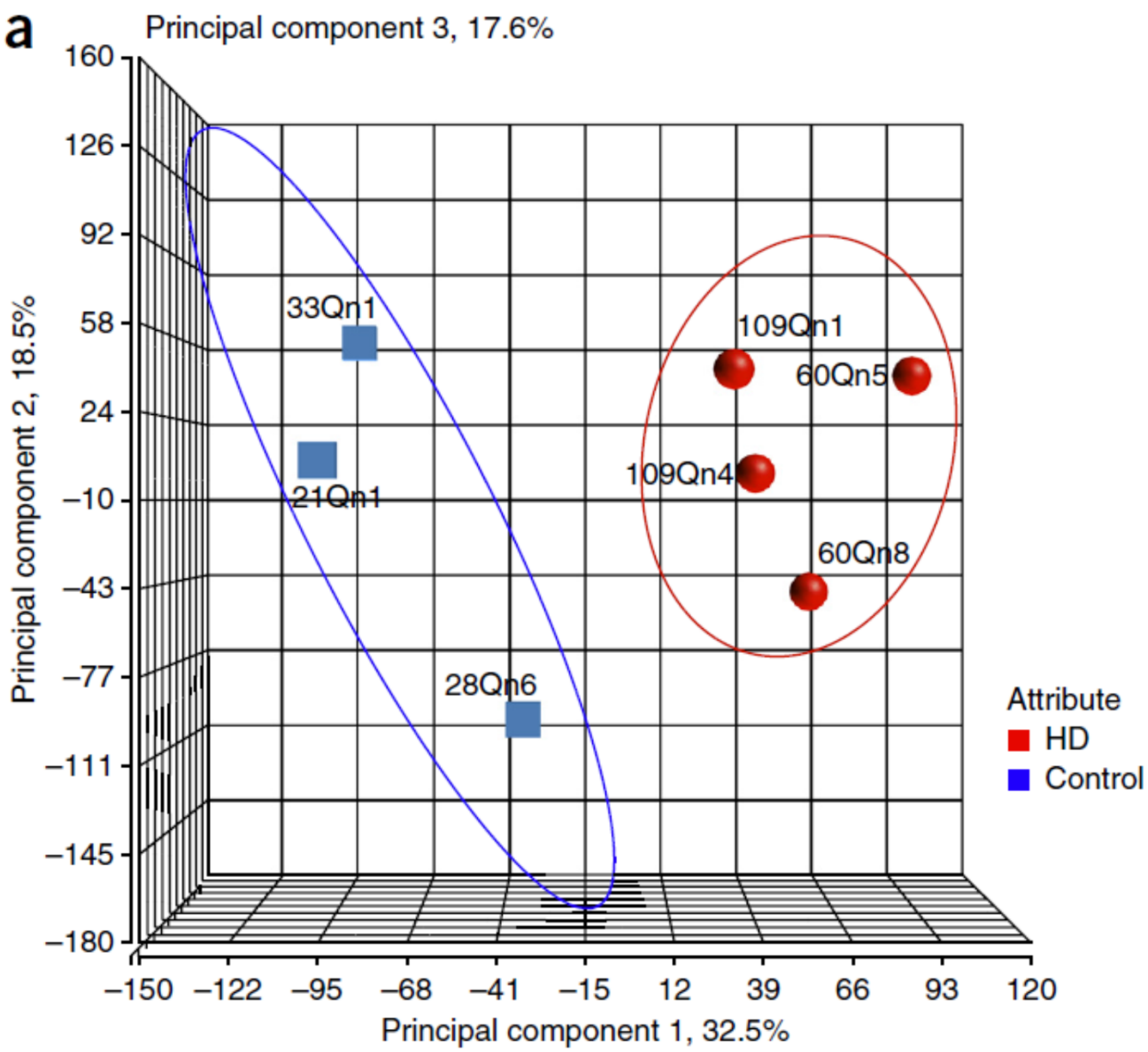

round 7 distance 54
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

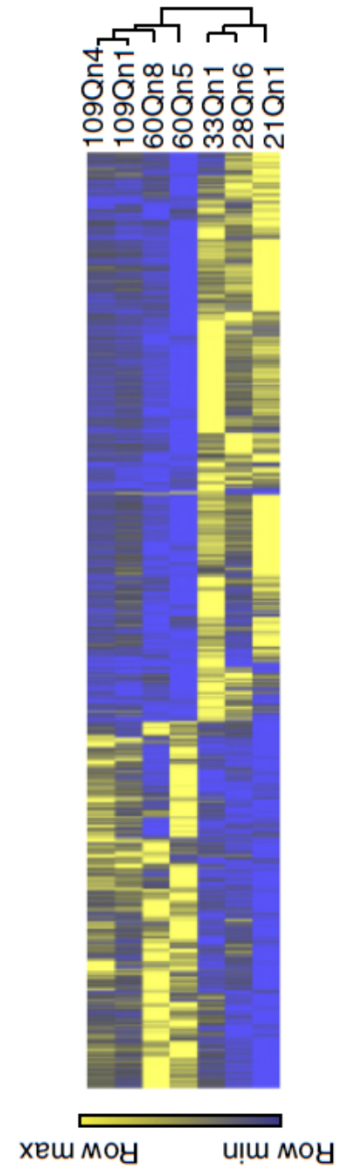

round 8 distance 54
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
-2.0
-1.5
-1.0
-0.5
0.0
0.5
1.0

What if we choose pathologically bad initial positions?

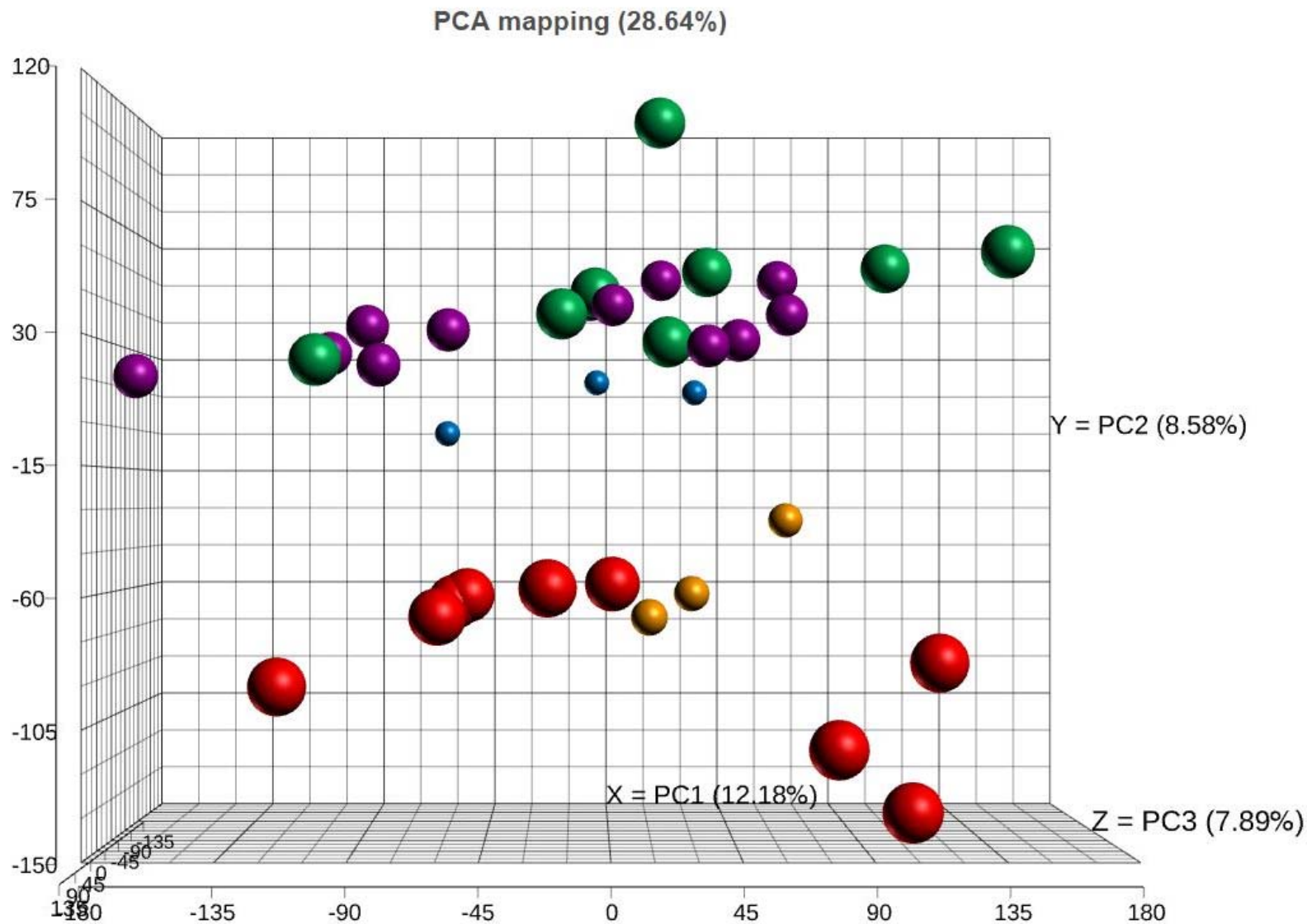
Often, the algorithm gets a reasonable answer, but not always!

Principal Component Analysis



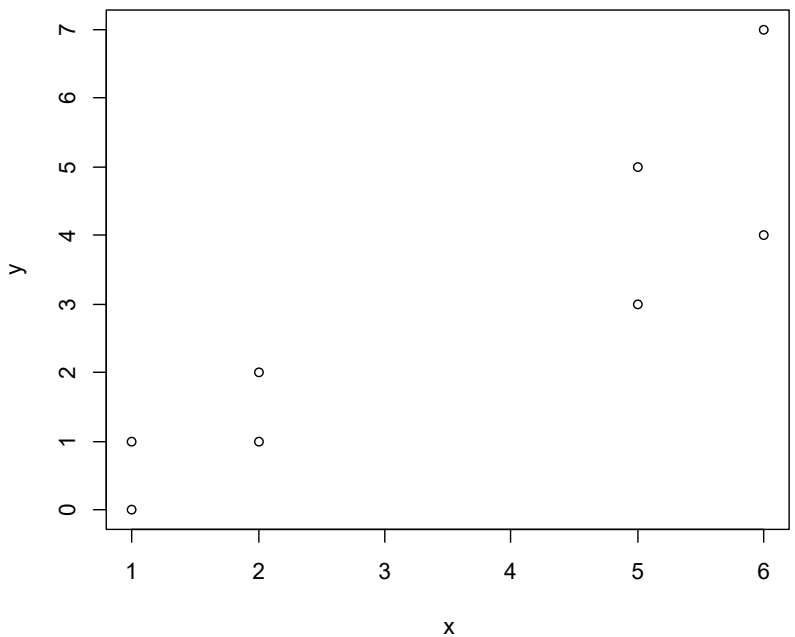


PCA can help spot anomalies in the data

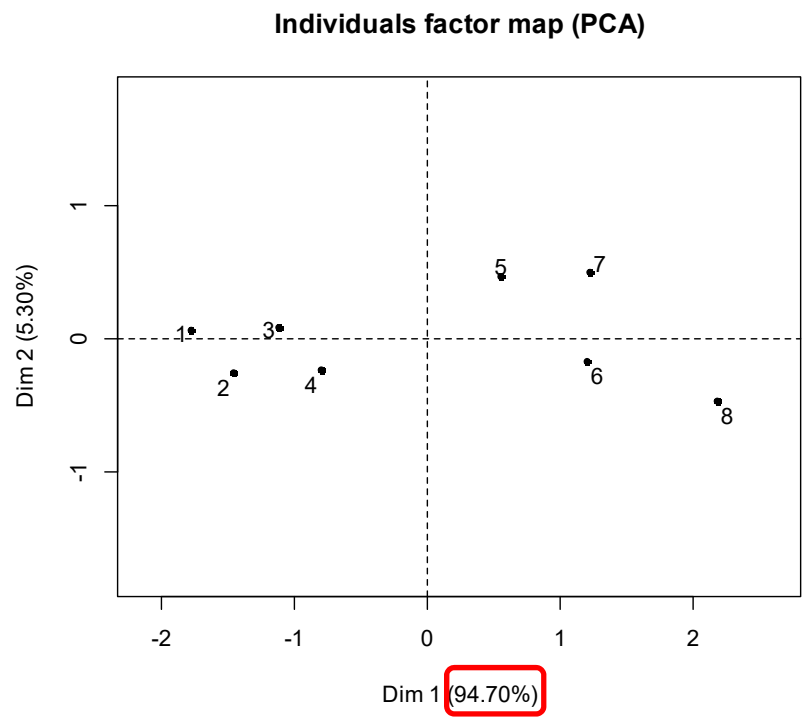


The types of plots you will make

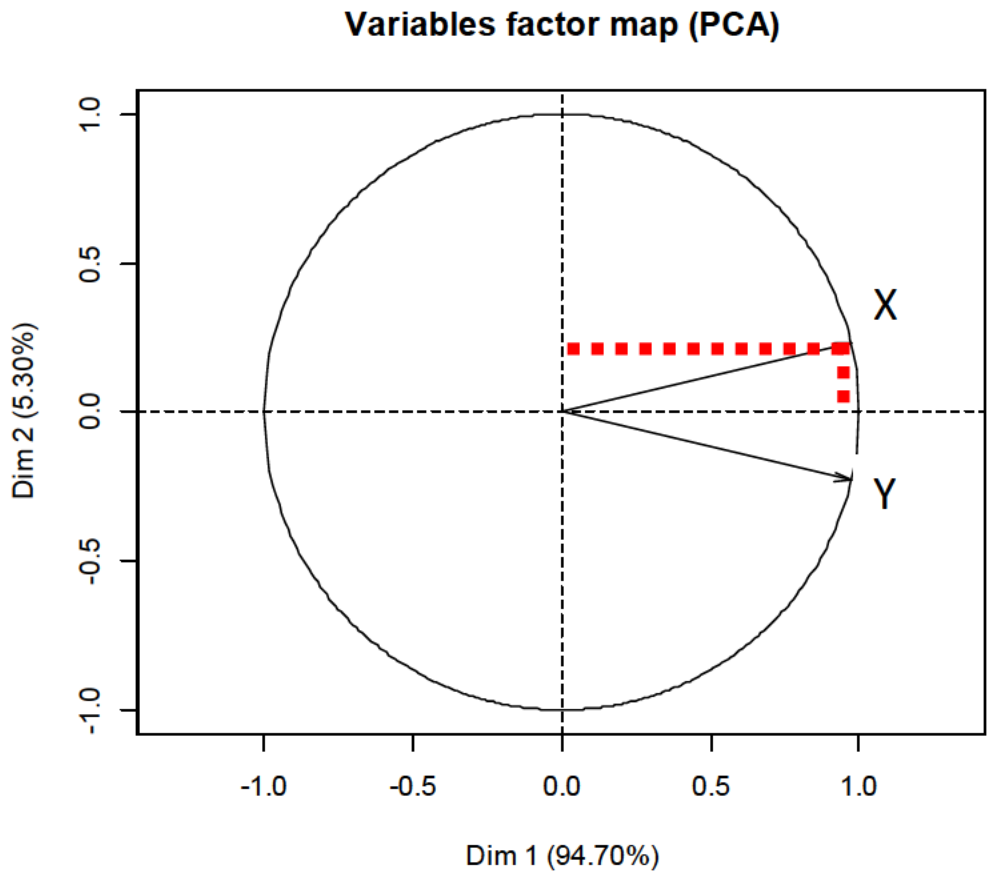
Original Data



PCA space



Note: 95% of the variance is explained by the first PCA dimension



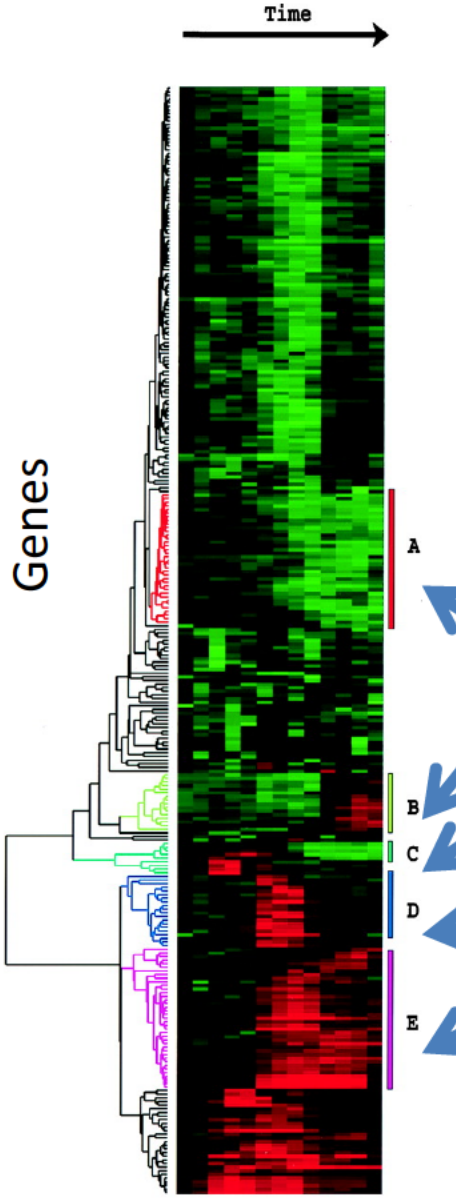
This plot shows us how much each of the original variables contributed to the PCA dimensions.

Here,

Dim1 = .97x + .97y

Dim2 = .23x -.23y

Next time: Interpreting your results



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

(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