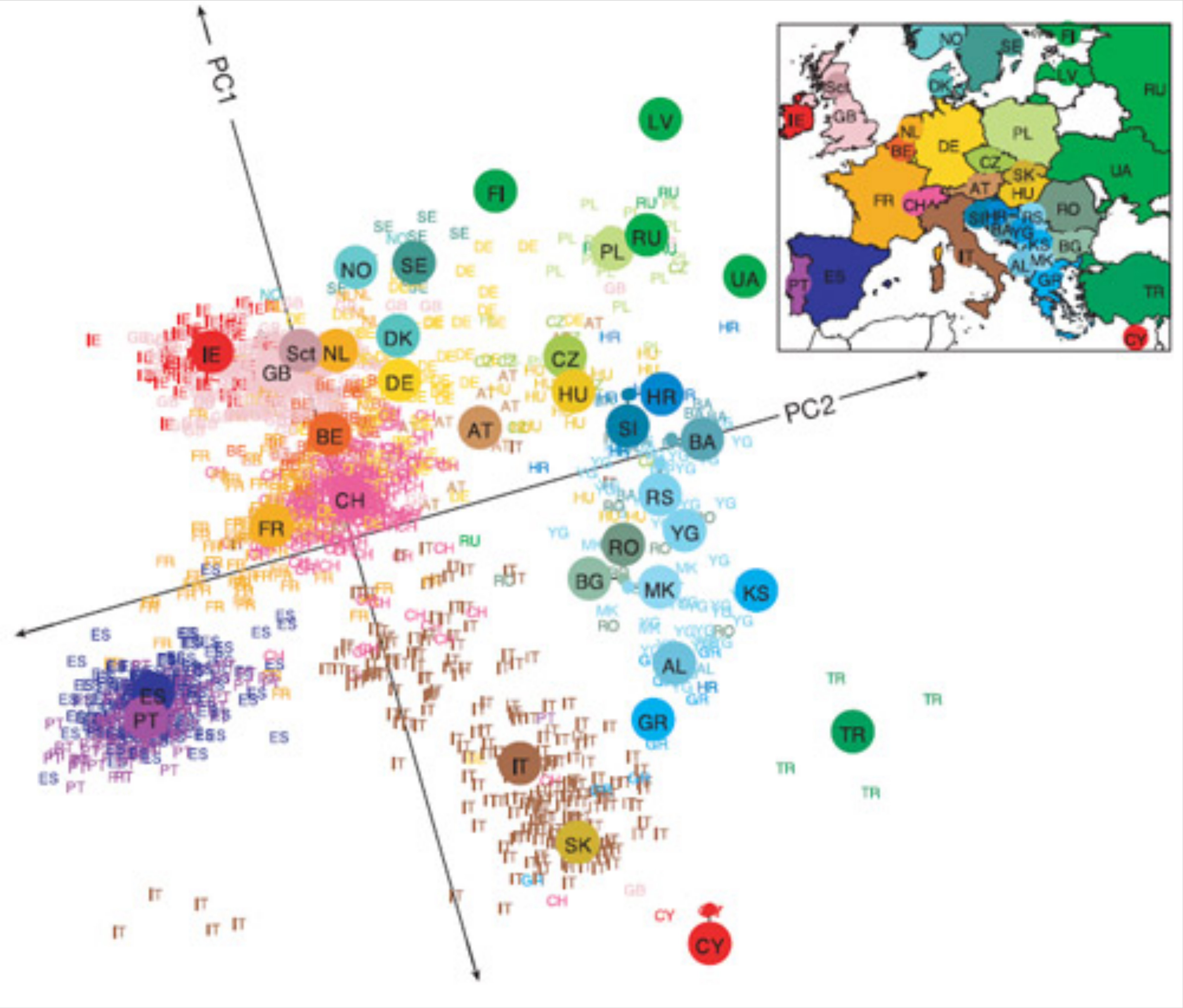
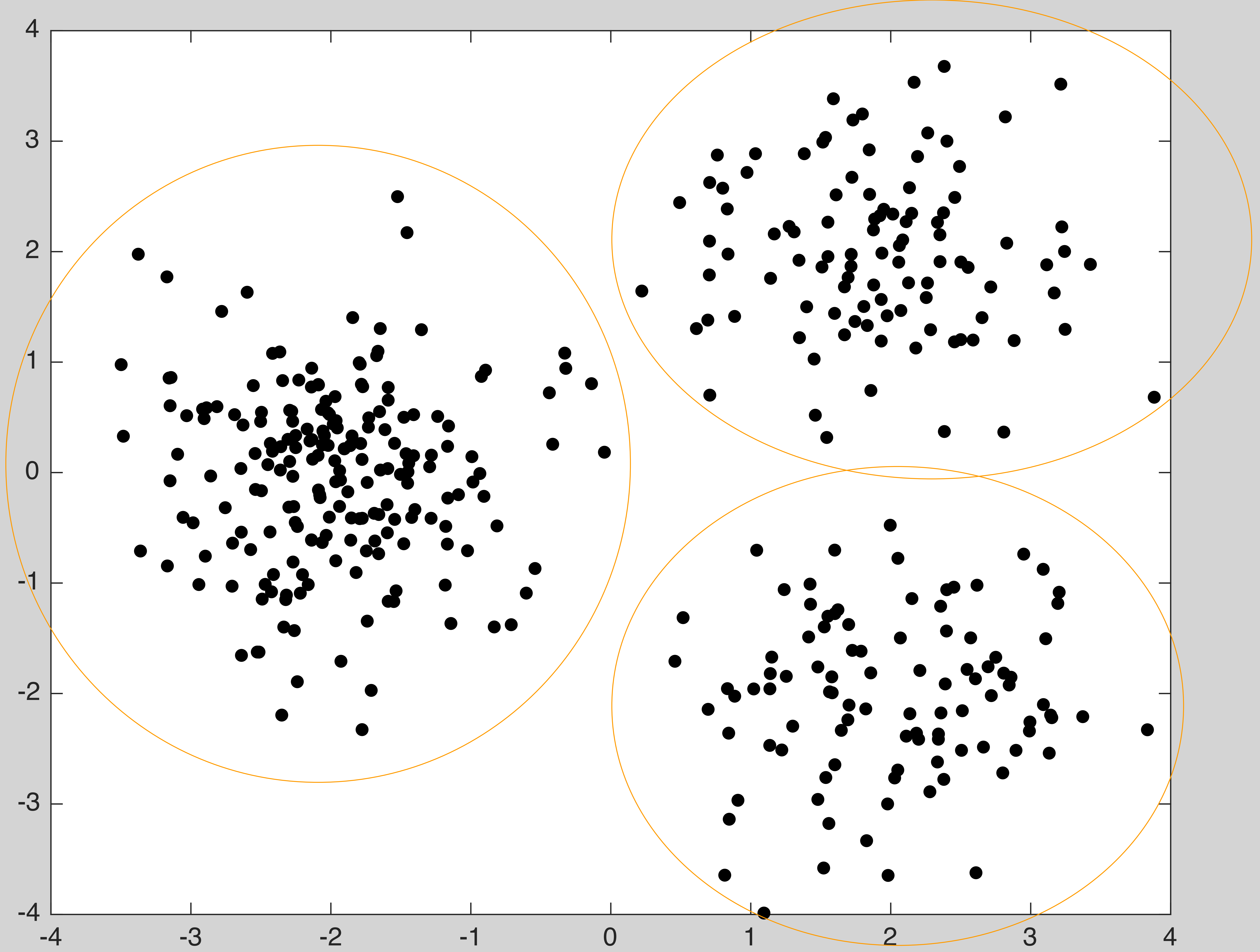
EE16B
Designing Information Devices and Systems II

Lecture 10A
k-means Clustering


PC1
PC2


4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4

k-means

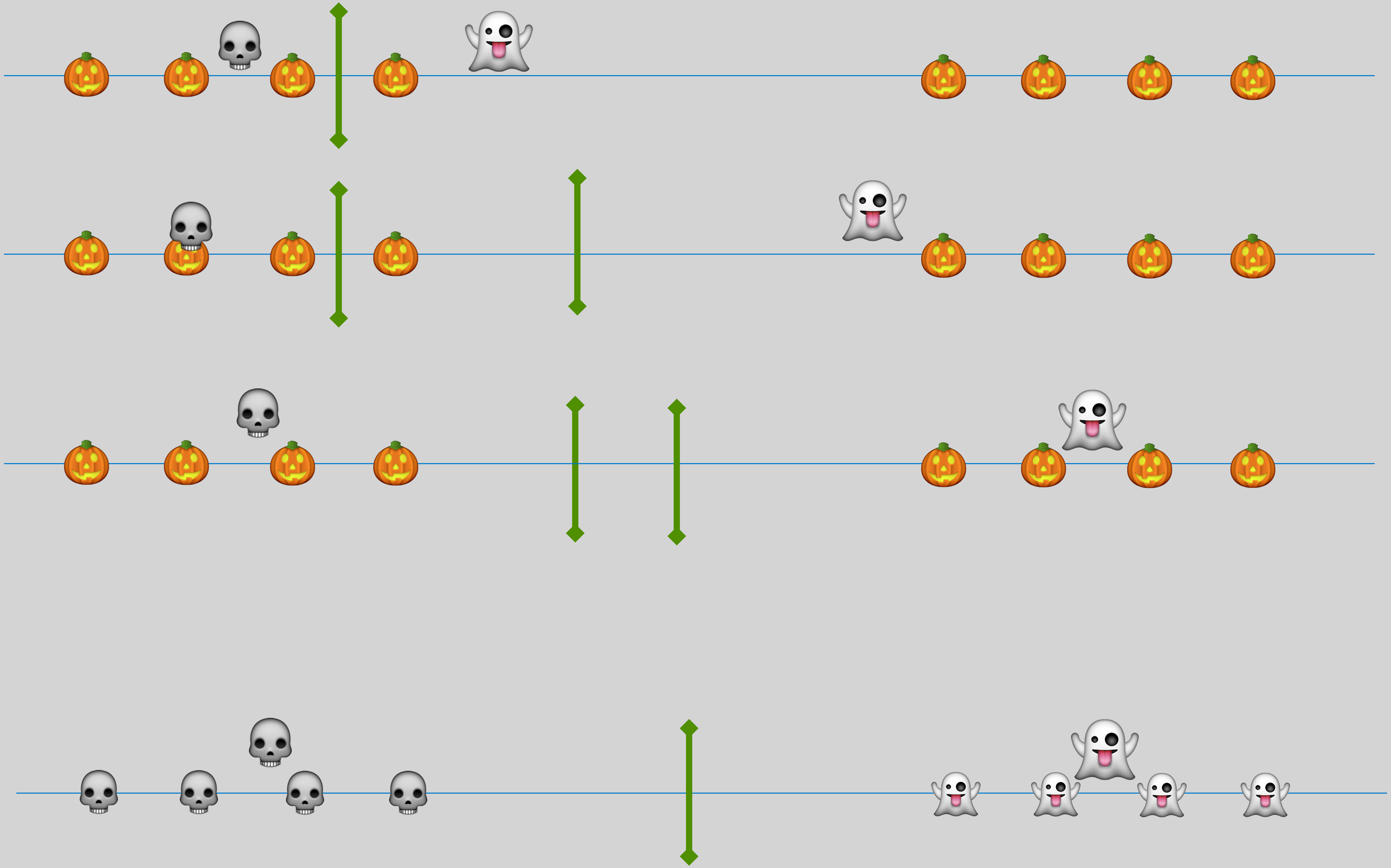
Given: $\vec{x}_1, \vec{x}_2, \cdots, \vec{x}_m \in \mathrm{R}^n$

Partition them into k << m groups

0) Guess cluster centers to initialize
1) Group points around nearest center
2) Update cluster centers by averaging within group
3) If centers have changed, repeat 1-3

k-means 1D example

$$n = 1, m = 8, k = 2$$



General k-means Algorithm

0) Initialize k cluster centers $\vec{m}_1, \vec{m}_2, \cdots, \vec{m}_k$

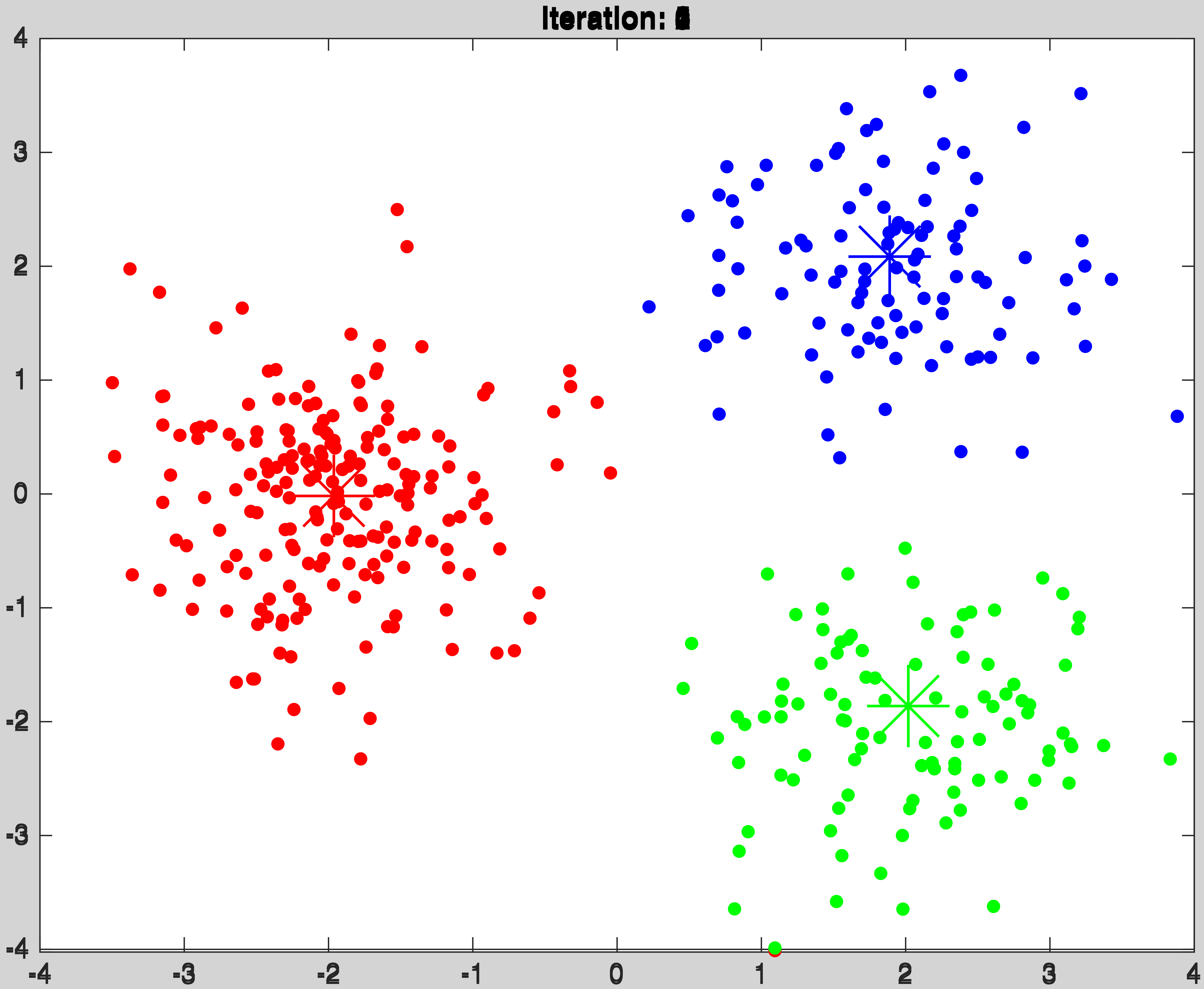
1) Assign points to cluster: point $\vec{x}$ goes to cluster i if,

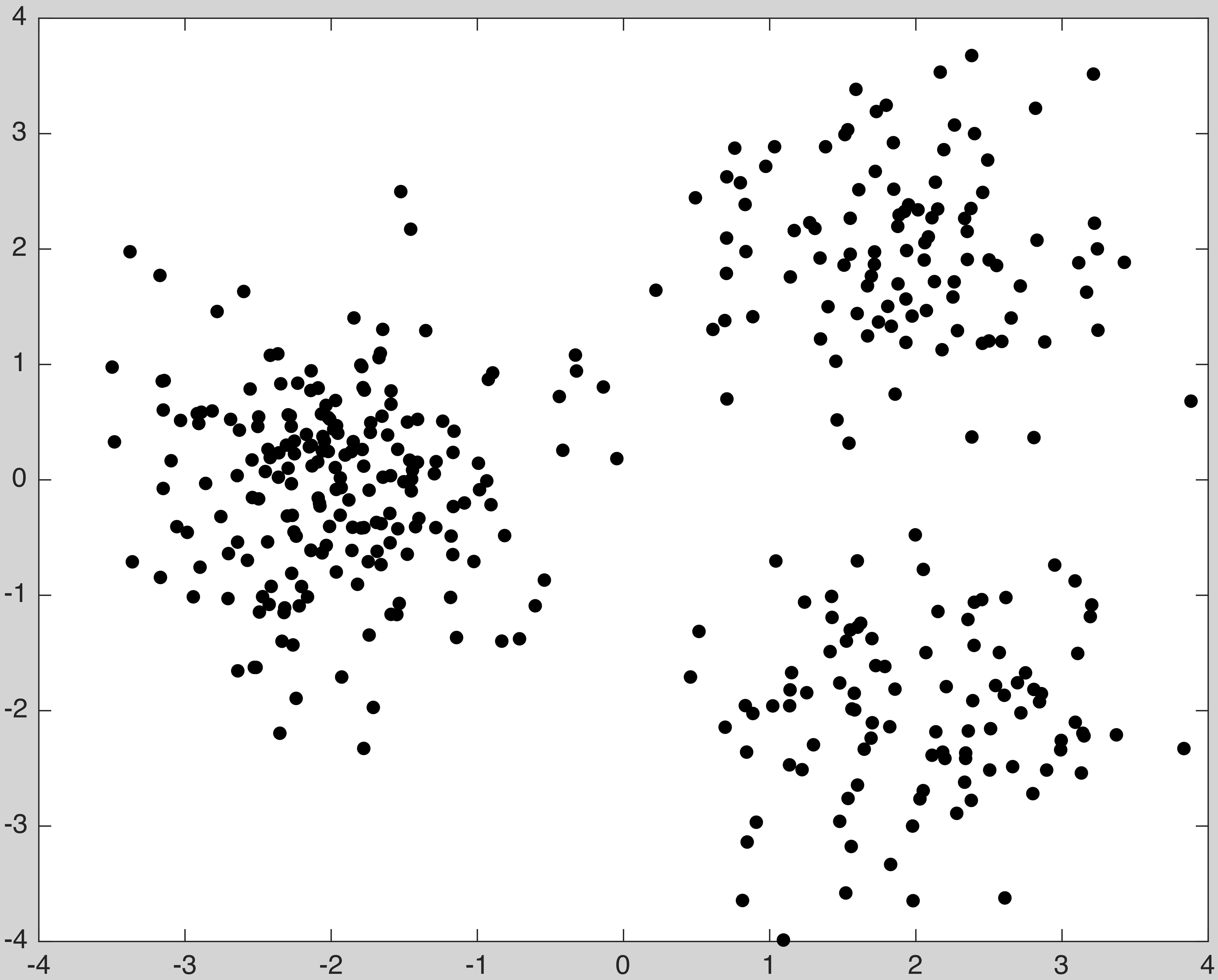
$$||\vec{x} - \vec{m}_i|| < ||\vec{x} - \vec{m}_j|| \quad \forall j \neq i$$

2) Let S_i be the set of samples in cluster i recompute cluster centers:

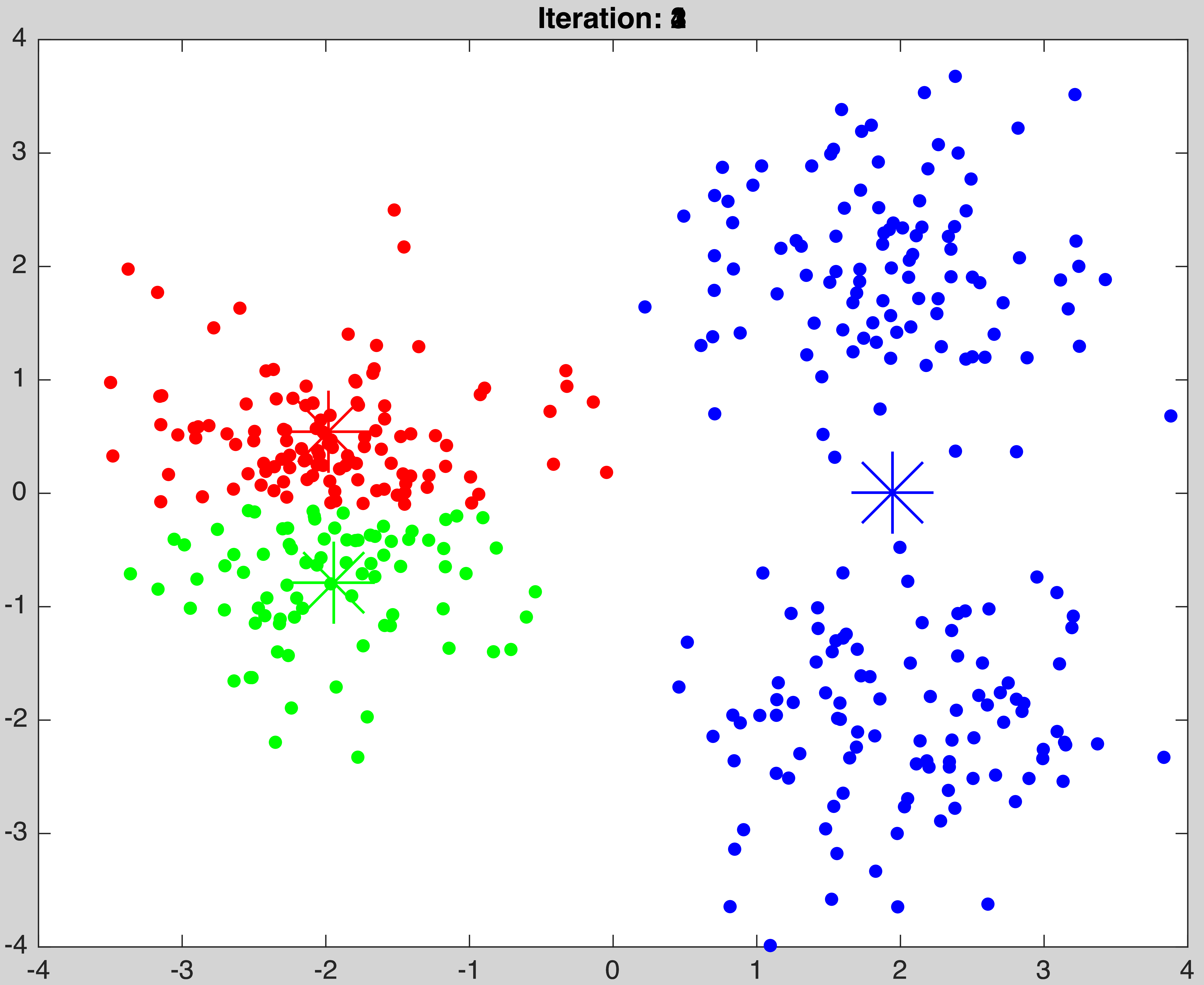
$$\vec{m}_i = \frac{1}{|S_i|} \sum_{\vec{x} \in S_i} \vec{x}$$

3) If any m_i has changed, repeat 1-3


Iteration:
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4



4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4


Iteration: 2
4
3
2
1
0
-1
-2
-3
-4
-4
-3
-2
-1
0
1
2
3
4

Objective Function

Find the clustering of $\vec{x}_1, \cdots, \vec{x}_m$ into sets $S_1, \cdots, S_k$ which minimizes:

$$D = \sum_{i=1}^{k} \sum_{\vec{x} \in S_i} ||\vec{x} - \mu_i||$$

$$\mu_i = \frac{1}{|S_i|} \sum_{x \in S_i} \vec{x}$$

While the algorithm decreases the objective, the objective is non-convex and can be stuck on local mimima.

General problem is N-P Complete

Management of intersections with multi-modal high-resolution data ☆,☆☆

Ajith Muralidharan [1], Samuel Coogan [2], Christopher Flores, Pravin Varaiya *

Sensys Networks, Inc, Berkeley, CA 94710, United States

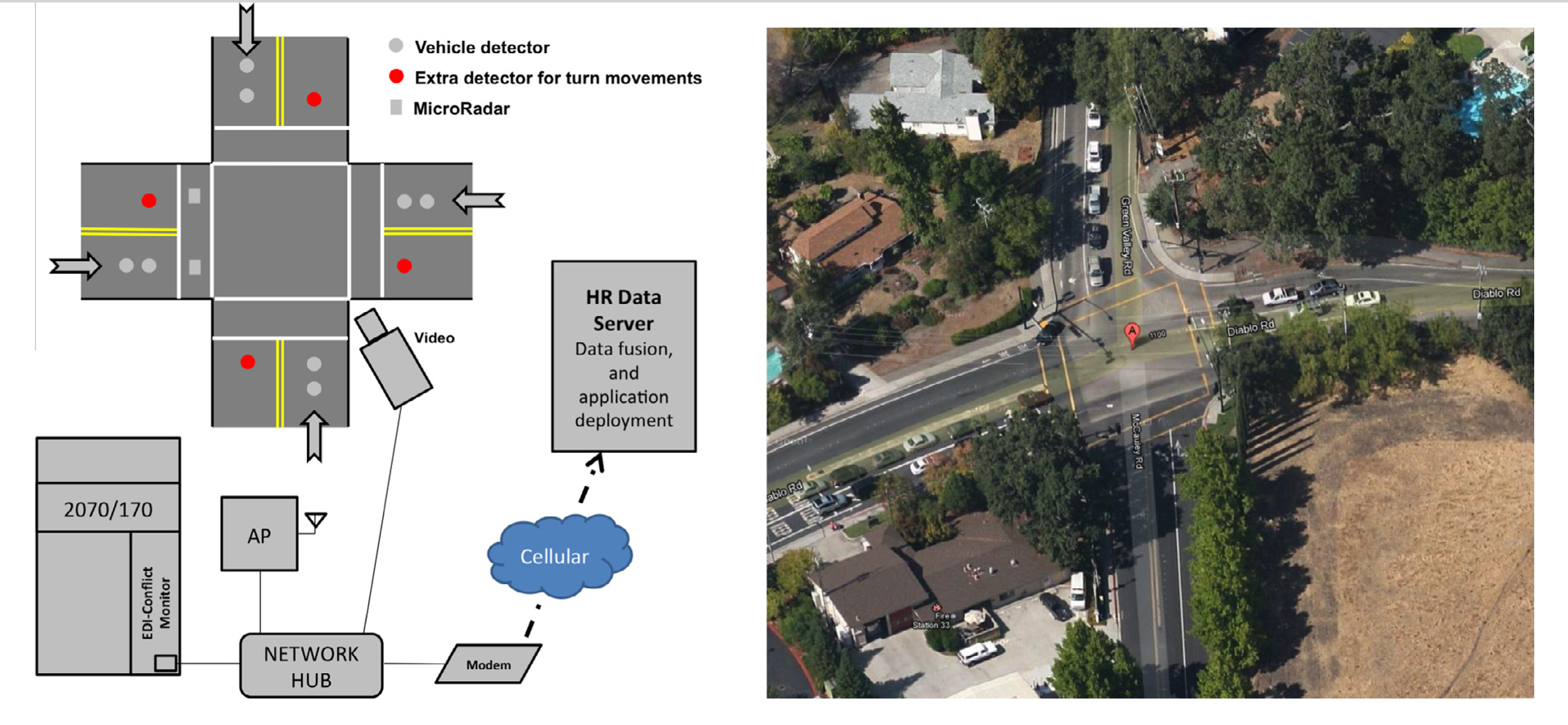
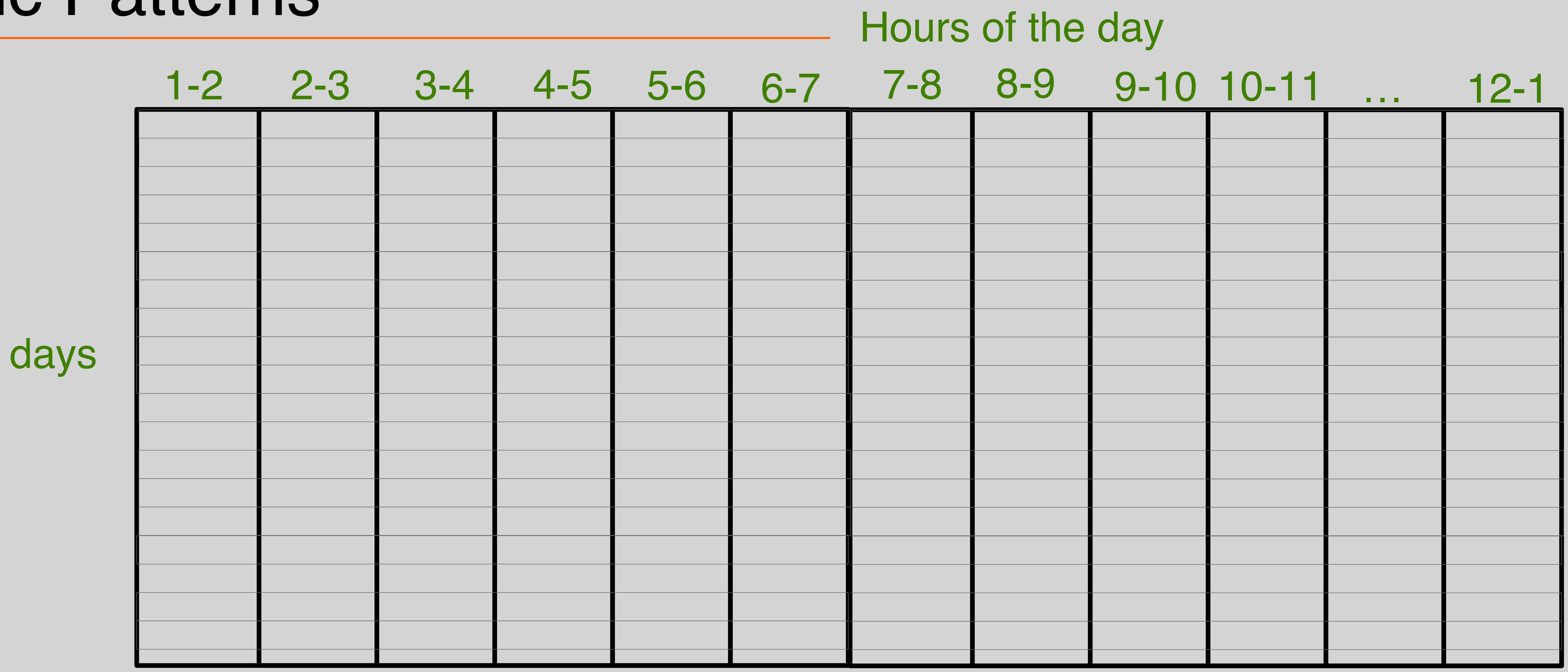


Fig. 1. HR system in Danville, CA.

Traffic Patterns



What would k-means cluster to?

K = 2?

K = 4?

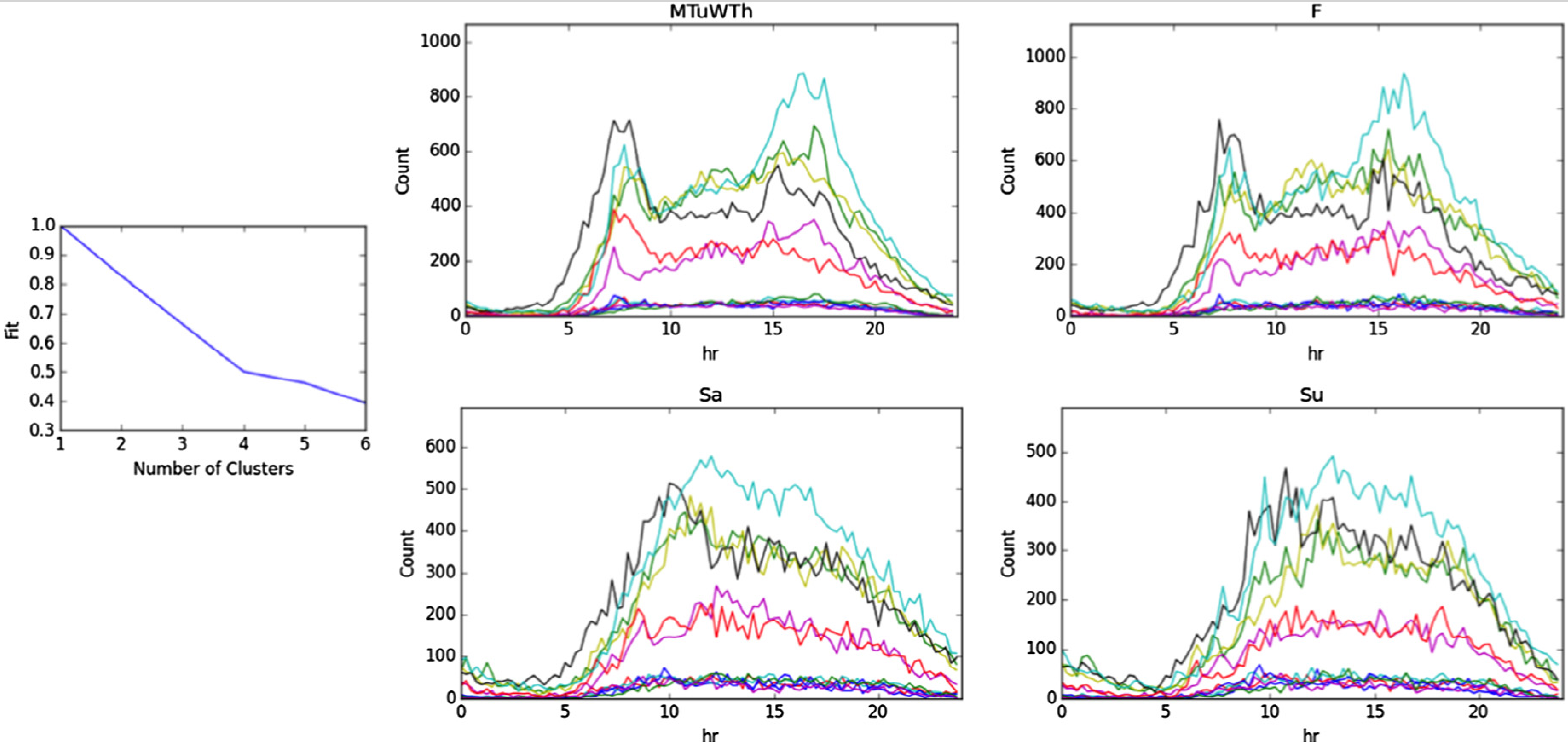


Fig. 5. Clustering of daily data for Dec 2014 to May 2015 in an intersection in Beaufort, SC.

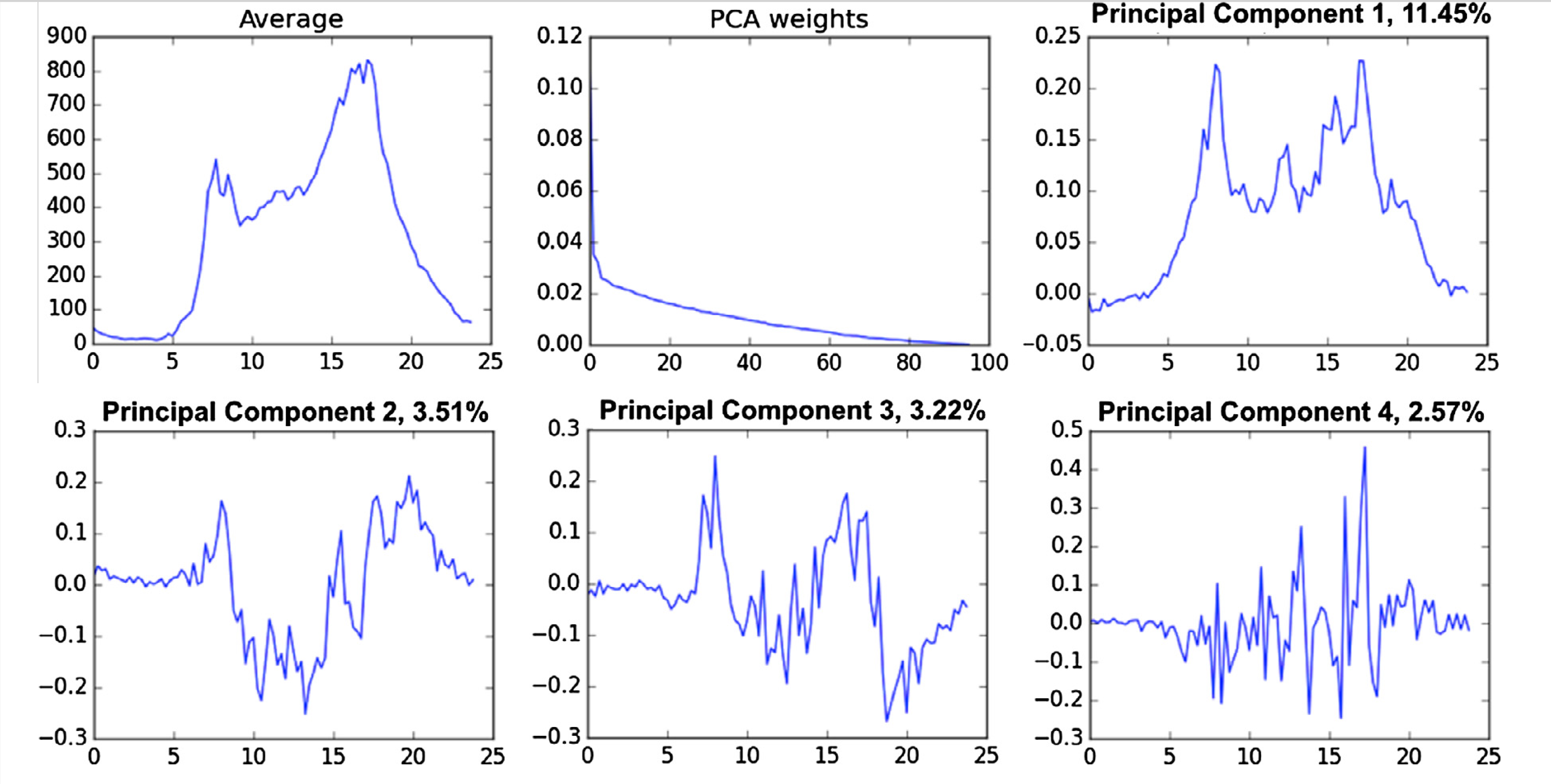


Fig. 8. Four PCA components of the North–South through movement and the average. The *x*-axis is hours.