

# K-means clustering algorithm

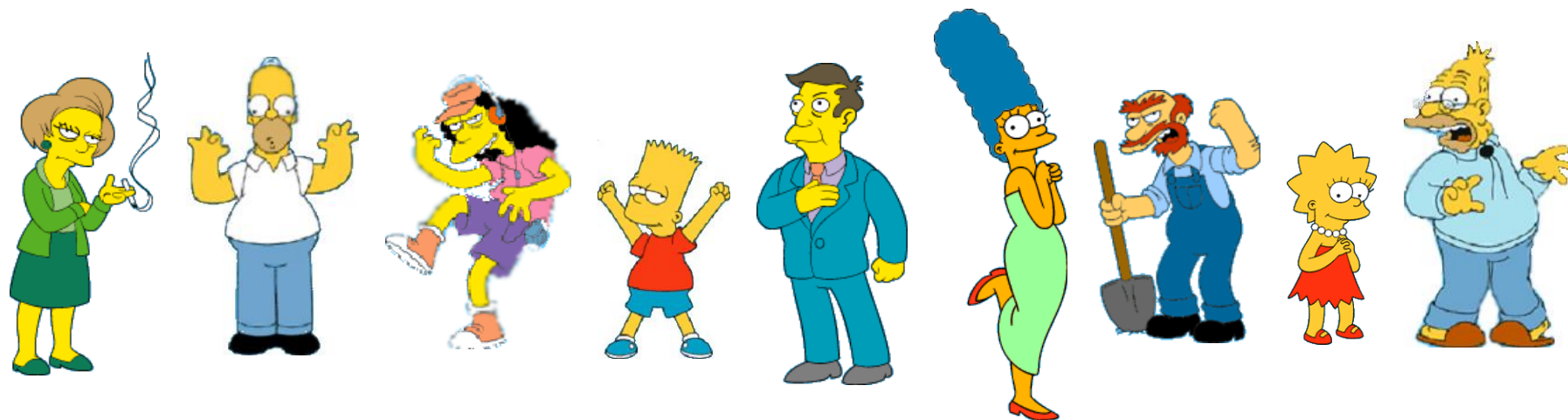


Kasun Ranga Wijeweera  
(S/06/483)

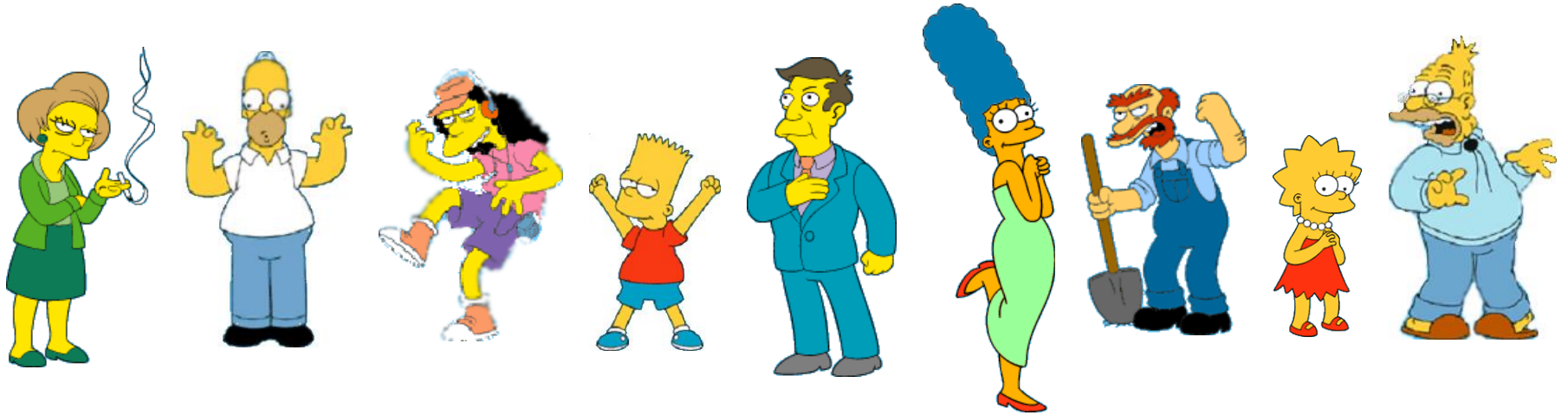
# What is Clustering?

- Organizing data into classes such that there is
  - high intra-class similarity
  - low inter-class similarity
- Finding the class labels and the number of classes directly from the data (in contrast to classification).
- More informally, finding natural groupings among objects.

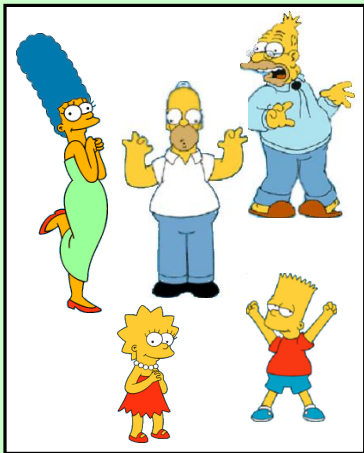
# What is a natural grouping among these objects?



# What is a natural grouping among these objects?



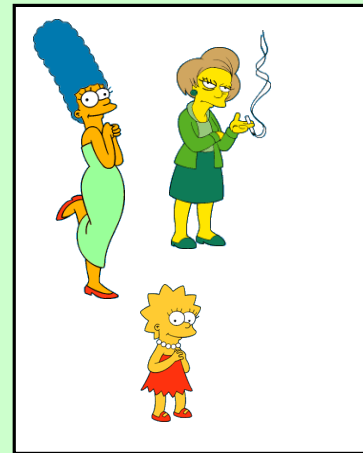
## Clustering is subjective



Simpson's Family



School Employees



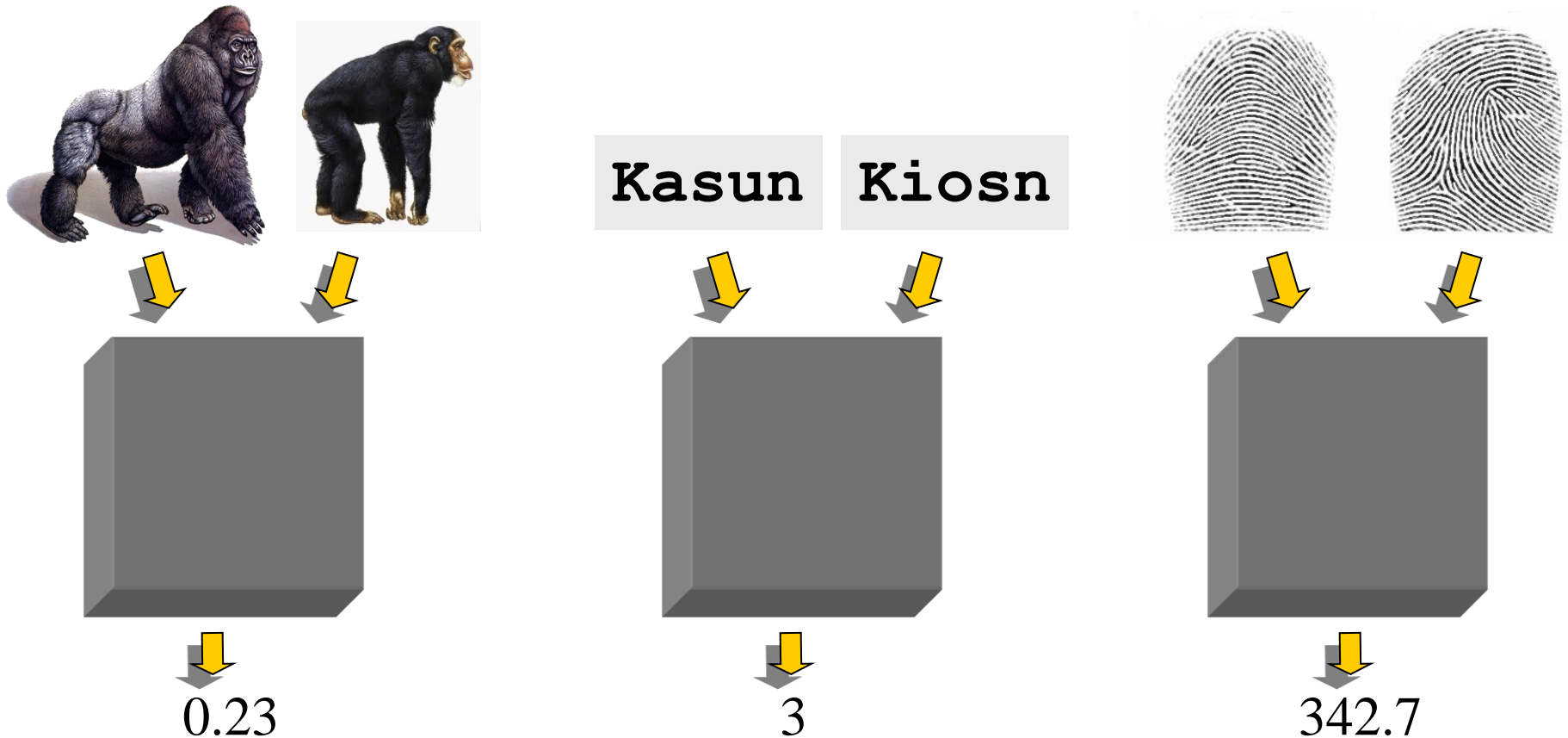
Females



Males

# Defining Distance Measures

**Definition:** Let  $O_1$  and  $O_2$  be two objects from the universe of possible objects. The distance (dissimilarity) between  $O_1$  and  $O_2$  is a real number denoted by  $D(O_1, O_2)$



Consider a Set of Data Points,

$$X = \{x_1, x_2, \dots, x_n\} \quad x_i = \begin{pmatrix} x_{i1} \\ x_{i2} \\ \vdots \\ x_{id} \end{pmatrix}_{d \times 1}$$

And a Set of Clusters,

$$C = \{c_1, c_2, \dots, c_K\}$$

# The Goal,

$\mu_k = \text{mean of the cluster } c_k$

*The squared error,*

$$J(c_k) := \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$$

*The sum of squared errors,*

$$J(C) := \sum_{k=1}^K \sum_{x_i \in c_k} \|x_i - \mu_k\|^2$$

# Algorithm *k-means*

1. Randomly choose  $K$  data items from  $X$  as initial centroids.

2. Repeat

- Assign each data point to the cluster which has the closest centroid.
- Calculate new cluster centroids.

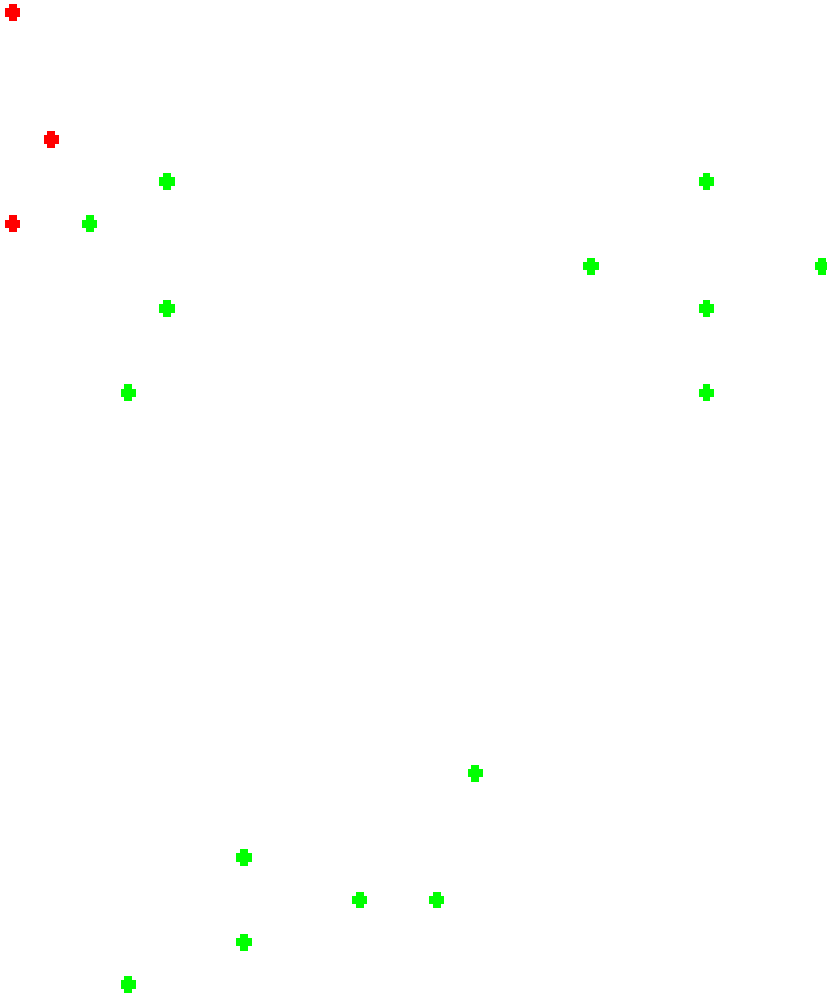
Until the convergence criteria is met.



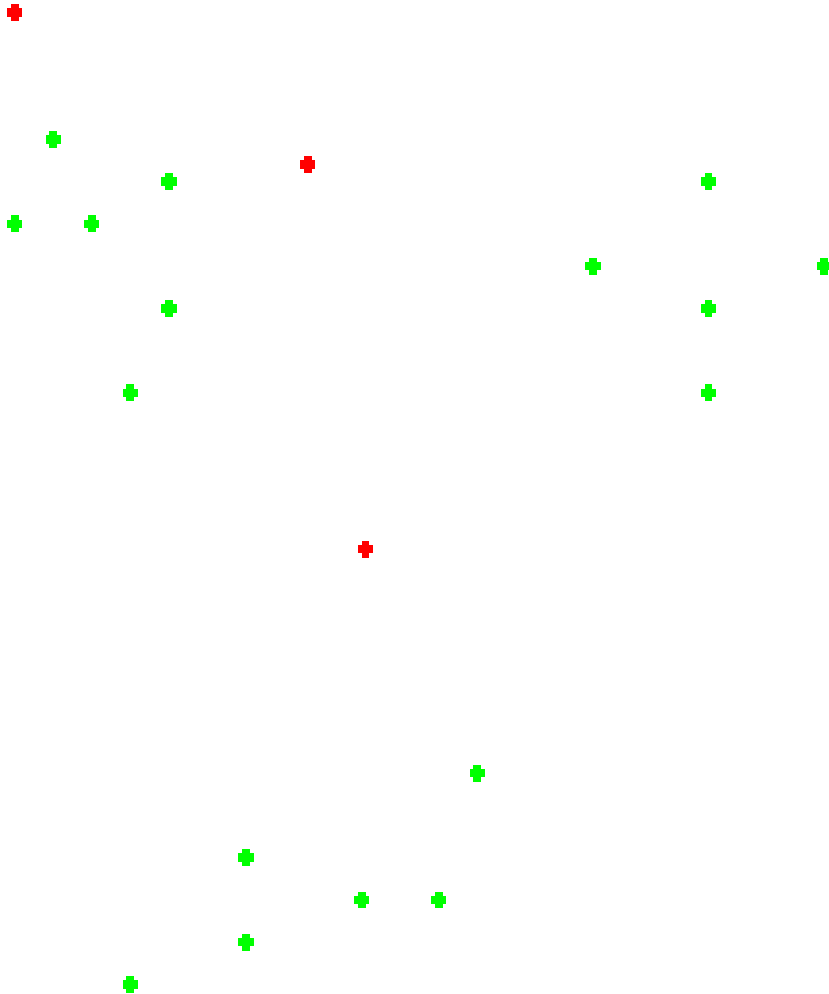
# The data points



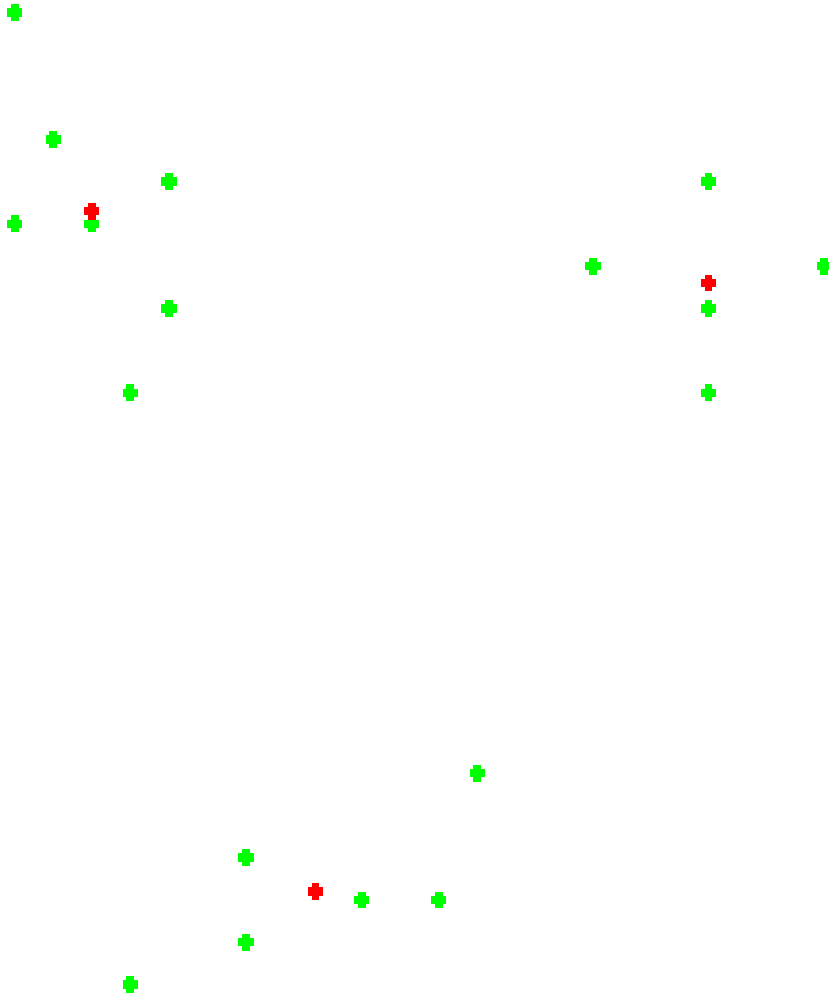
# Initialization



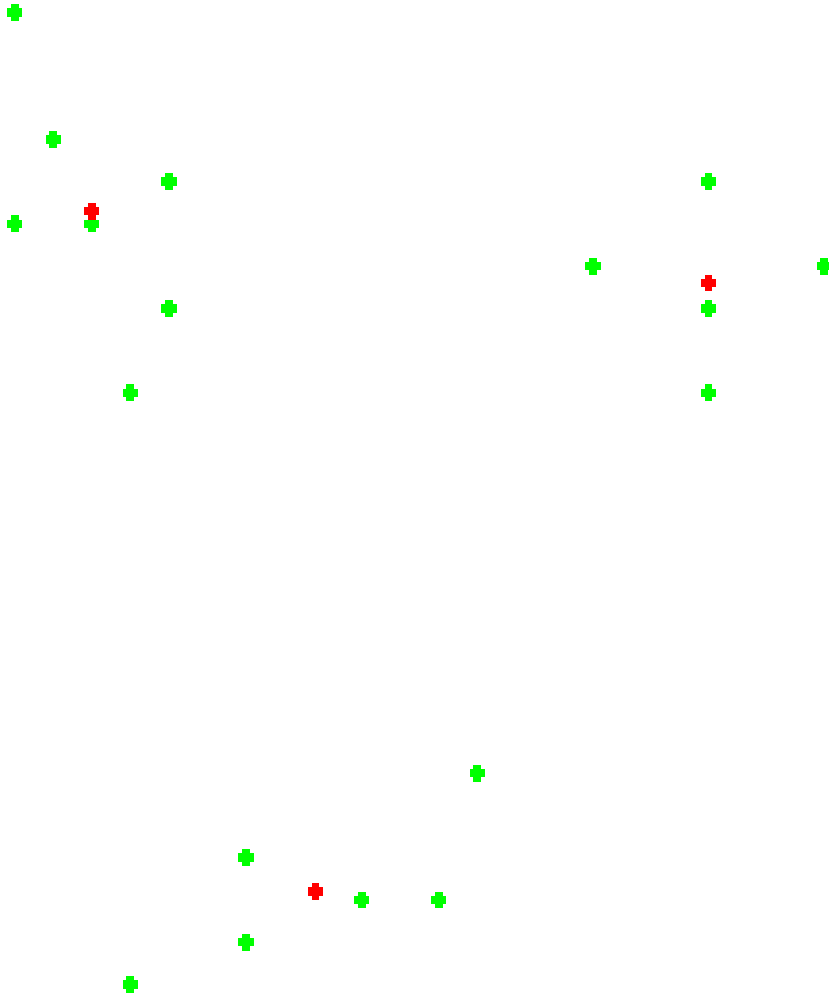
#Runs = 1



#Runs = 2

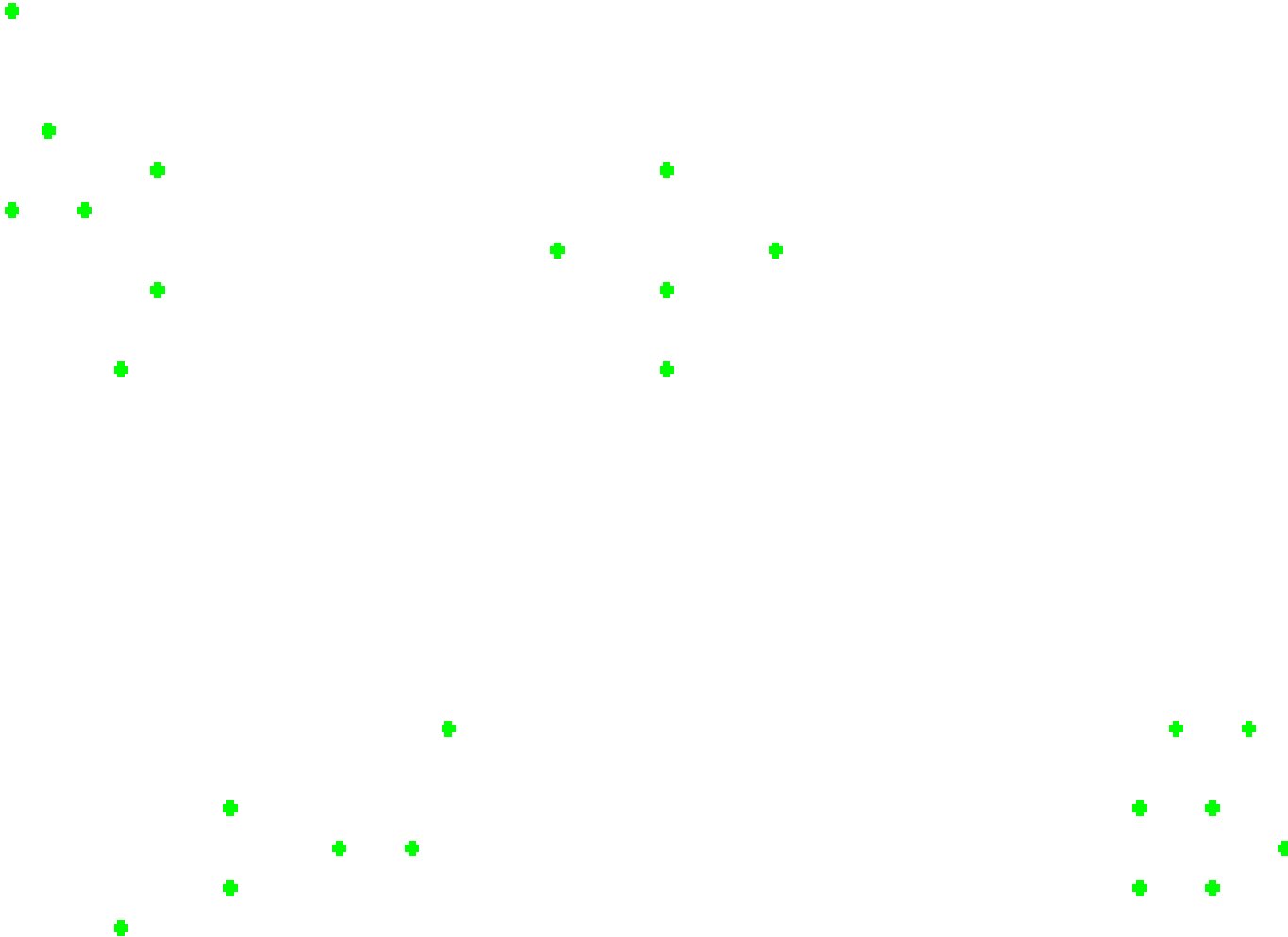


#Runs = 3

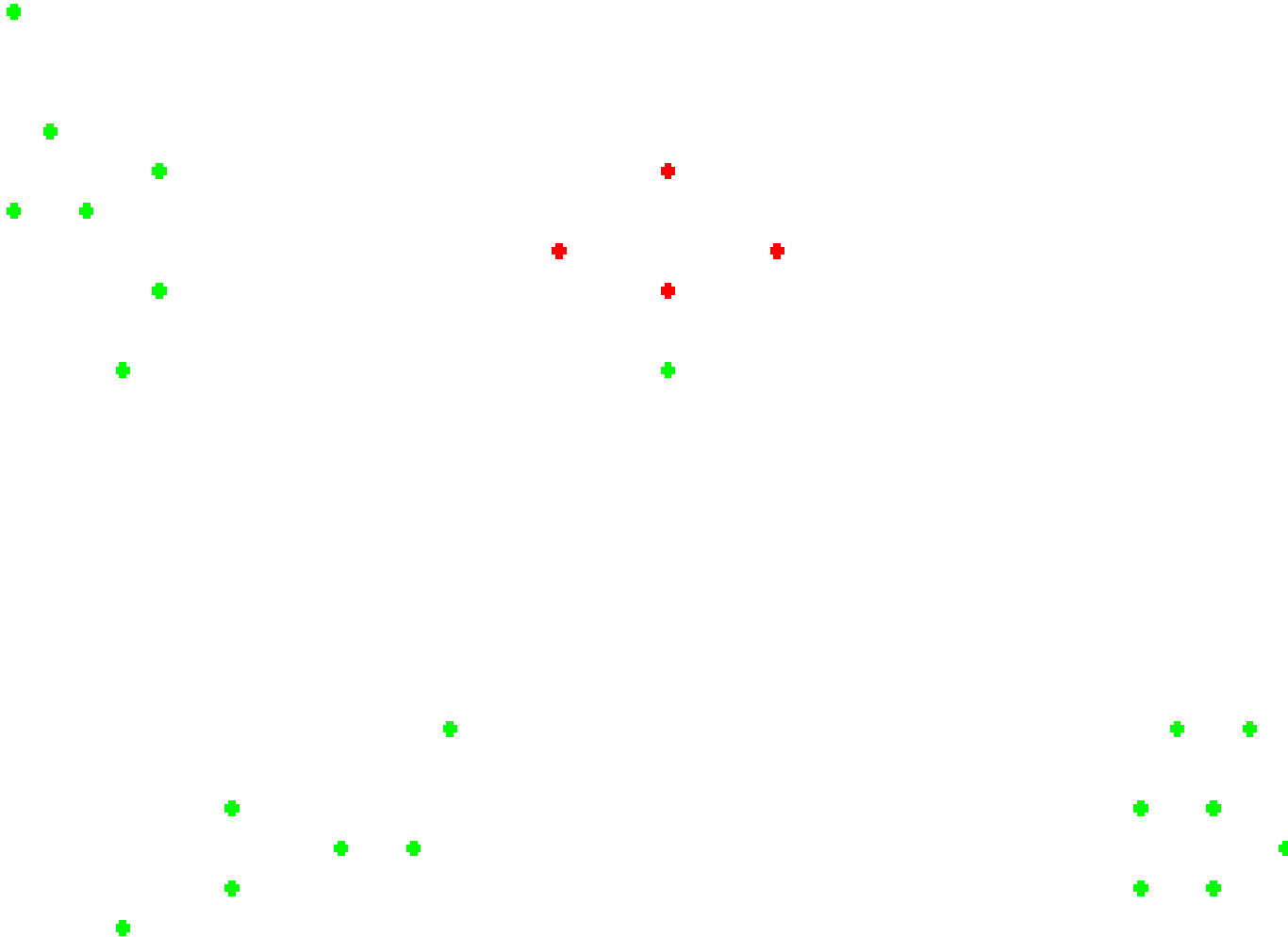


K-means gets stuck in a local  
optima

# The data points

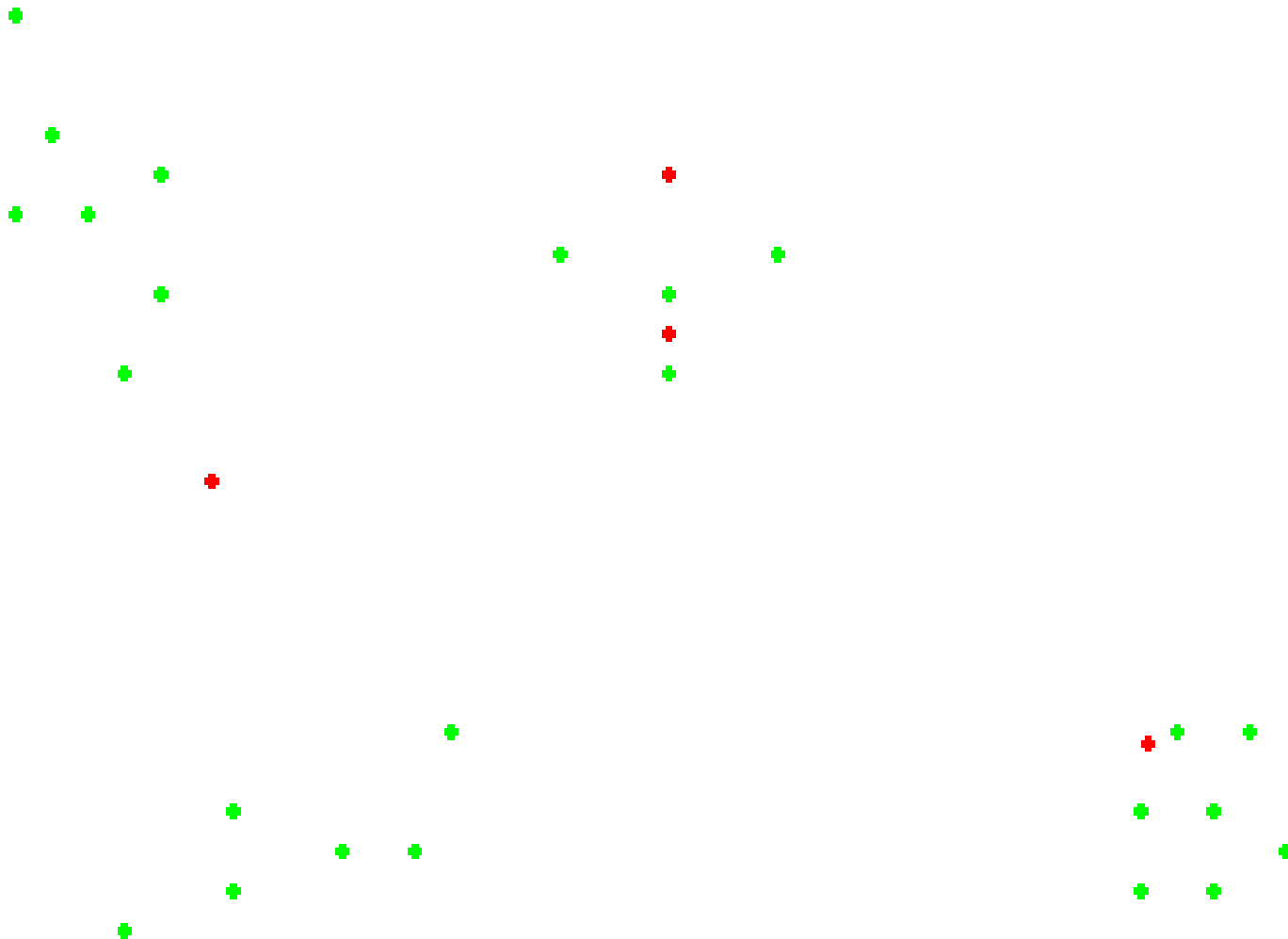


# Initialization

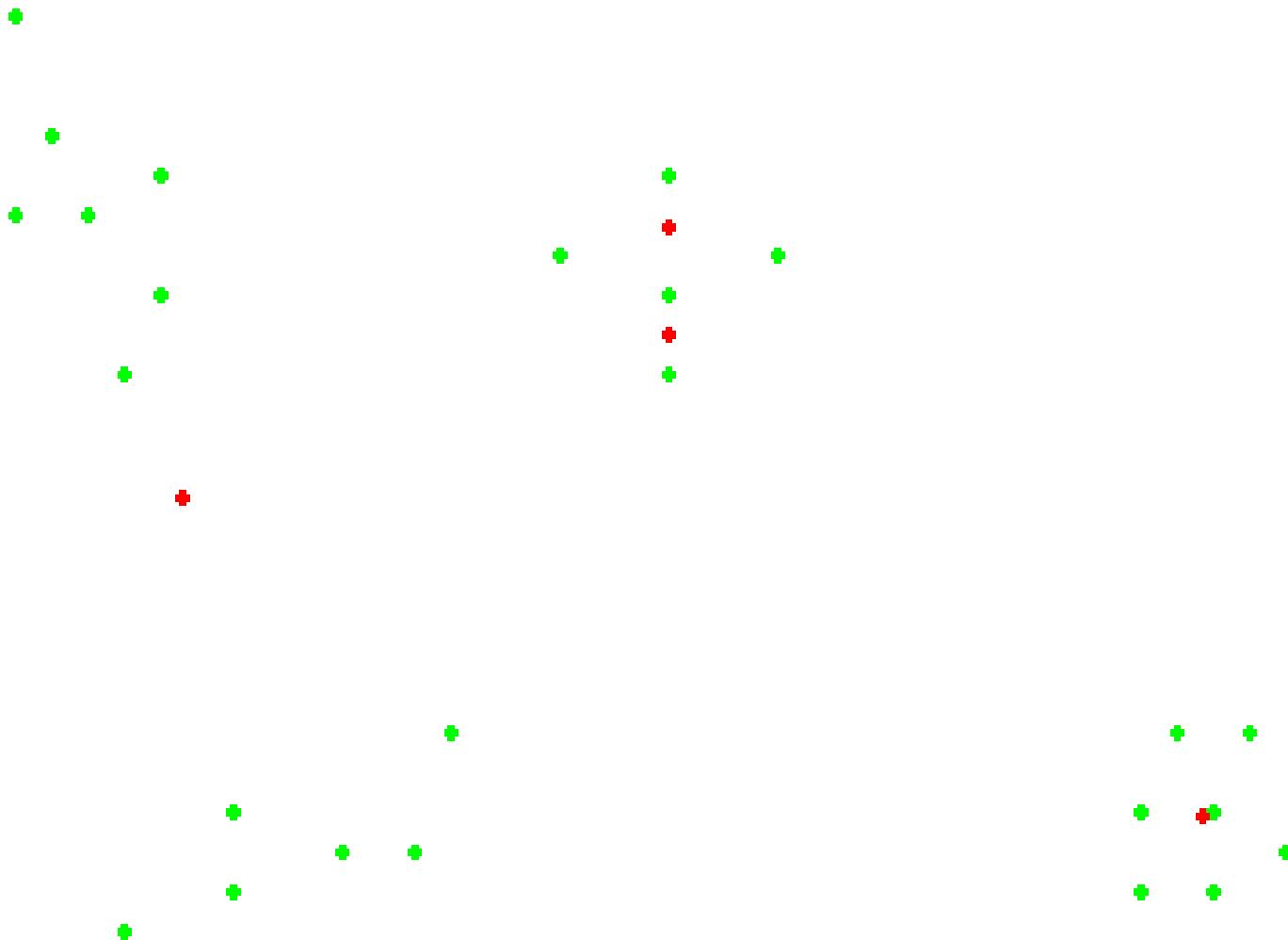




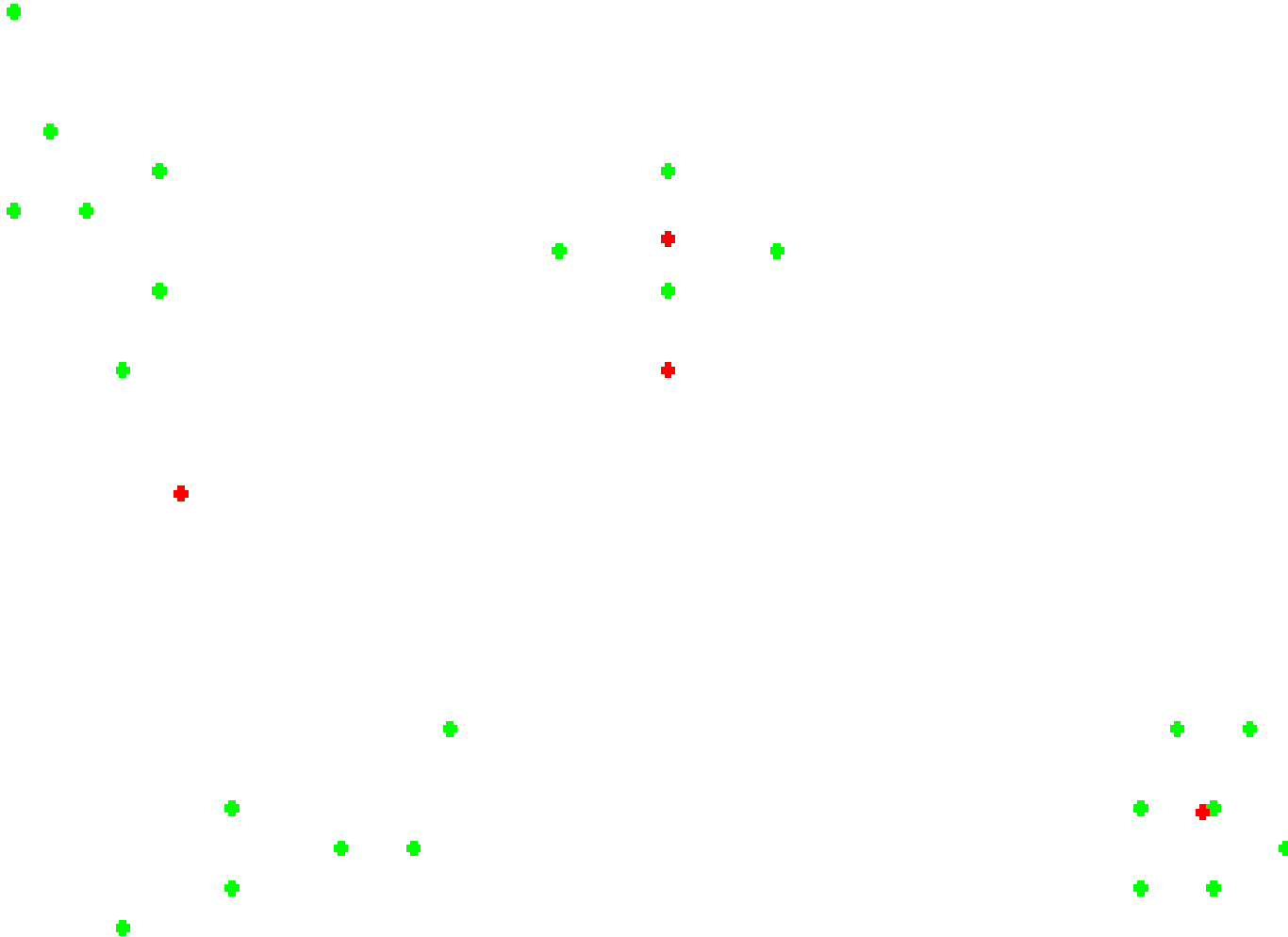
#Runs = 1



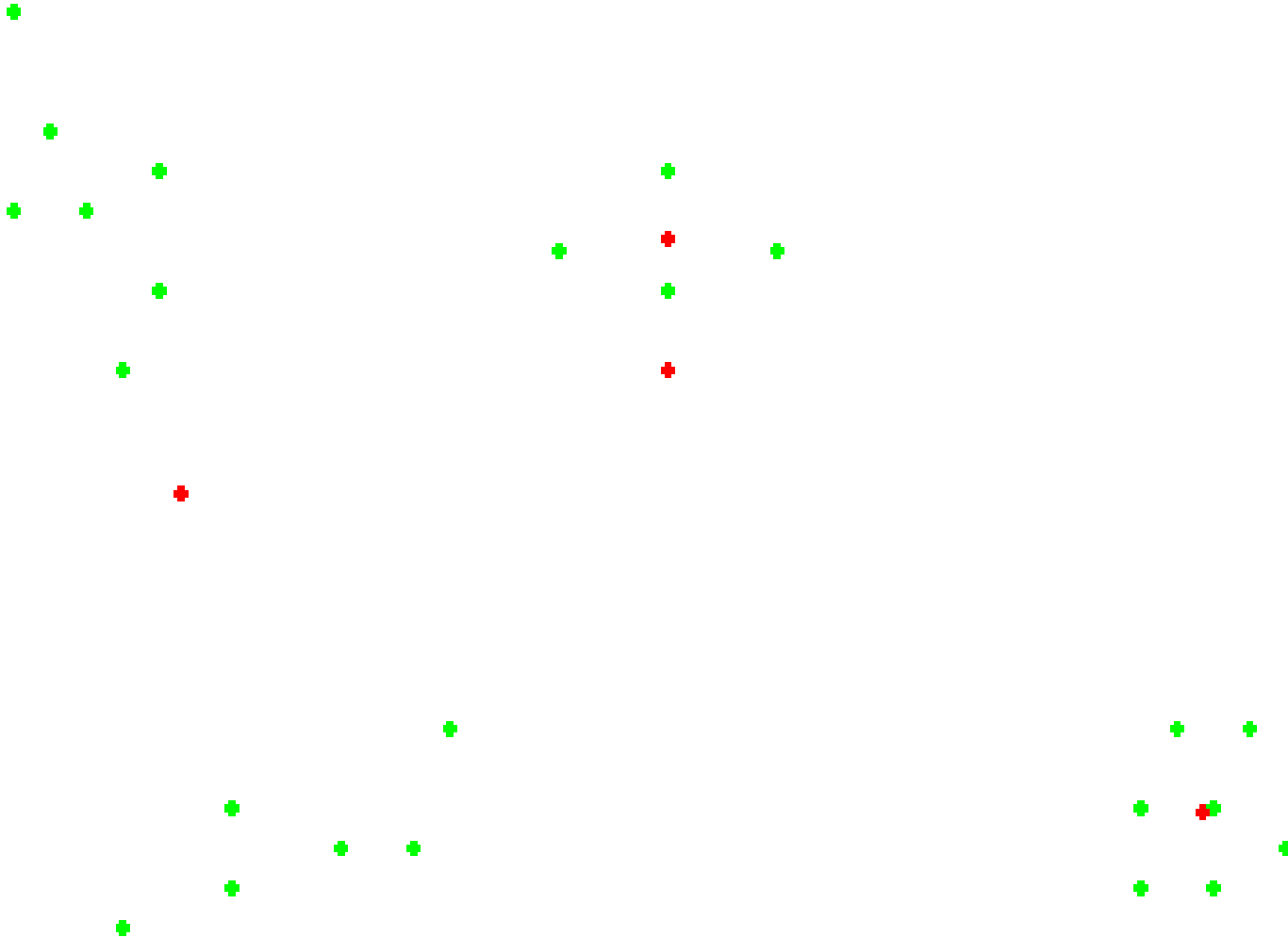
#Runs = 2



#Runs = 3



#Runs = 4



# Applications of K-means Method

- Optical Character Recognition
- Biometrics
- Diagnostic Systems
- Military Applications

# Comments on the *K-Means* Method

- Strength

- *Relatively efficient:  $O(tkn)$* , where  $n$  is # objects,  $k$  is # clusters, and  $t$  is # iterations. Normally,  $k, t \ll n$ .
- Often terminates at a *local optimum*. The *global optimum* may be found using techniques such as: *deterministic annealing* and *genetic algorithms*

- Weakness

- Applicable only when *mean* is defined, then what about categorical data?
- Need to specify  $k$ , the *number* of clusters, in advance
- Unable to handle noisy data and *outliers*
- Not suitable to discover clusters with *non-convex shapes*

Any Questions ?

Thanks for your attention !