k-Means algorithm

Procheta Sen

UNIVERSITY OF LIVERPOOL

Representative-based algorithms

- Let k be the number of clusters
- Let $D = \{\overline{X}_1, \ldots, \overline{X}_n\}$ be the dataset
- The goal is to determine k representatives $\overline{Y}_1, \ldots, \overline{Y}_k$ that minimise the following objective function

$$\sum_{i=1}^{n} \left[\min_{j} d(\overline{X}_i, \overline{Y}_j) \right]$$

i.e. the sum of the distances of the objects to their closest representatives needs to be minimised.

Representative-based algorithms

We obtain specific algorithms by specifying

- the way of choosing representatives, and
- the distance function $d(\cdot, \cdot)$

In general representatives do not necessarily belong to the dataset.

General k-representatives approach

- **Initialise**: pick initial k representatives
- **Iteratively refine**:
  - (**assign step**) Assign each object to its closest representative using distance function $d(\cdot\,,\cdot)$. Denote the corresponding clusters $C_1, \ldots, C_k$
  - (**optimise step**) Determine the optimal representative $\overline{Y}_j$ for each cluster C_j that minimises its local objective function $\sum_{\overline{X}_i \in C_j} \left[d\left(\overline{X}_i, \overline{Y}_j\right) \right]$

k-Means algorithm

- Representatives are chosen **not necessarily** from the dataset
- The distance function is **squared Euclidean distance**

k-Means algorithm

The objective function $$\min_{\overline{Y}_1,\ldots,\overline{Y}_k} \sum_{i=1}^{k} \sum_{\overline{X}\in C_i} \|\overline{X} - \overline{Y}_i\|^2$$

where C_i consists of the objects that are closest to $\overline{Y}_i$.

We want to minimise the total **squared Euclidean distance** between data objects and their cluster representatives $\overline{Y}_1, \ldots, \overline{Y}_k$

This objective function is called the *within cluster sum of squares* (WCSS) objective

k-Means algorithm

Assume that the clusters $C_1, C_2, \ldots, C_k$ are fixed.

Find the set $\overline{Y}_1, \overline{Y}_2, \ldots, \overline{Y}_k$ of representatives such that

$$f_{C_1,\ldots,C_k}(\overline{Y}_1, \ldots, \overline{Y}_k) = \sum_{i=1}^{k} \sum_{\overline{X} \in C_i} \|\overline{X} - \overline{Y}_i\|^2 \text{ is minimised.}$$

$$\frac{\partial f_{C_1,\ldots,C_k}(\overline{Y}_1, \ldots, \overline{Y}_k)}{\partial \overline{Y}_i} = -\sum_{\overline{X} \in C_i} 2(\overline{X} - \overline{Y}_i) = 0 \qquad \overline{Y}_i = \frac{1}{|C_i|} \sum_{\overline{X} \in C_i} \overline{X}$$

Just compute the centroid (mean) of each cluster and that will give you the new **optimal** cluster representatives

k-Means algorithm

k-MeansClustering (Number of clusters: k, Dataset: $\{\overline{X}_1, \ldots, \overline{X}_n\}$)

1. Initialisation phase

Choose k cluster representatives $\overline{Y}_1, \ldots, \overline{Y}_k$ from the dataset randomly

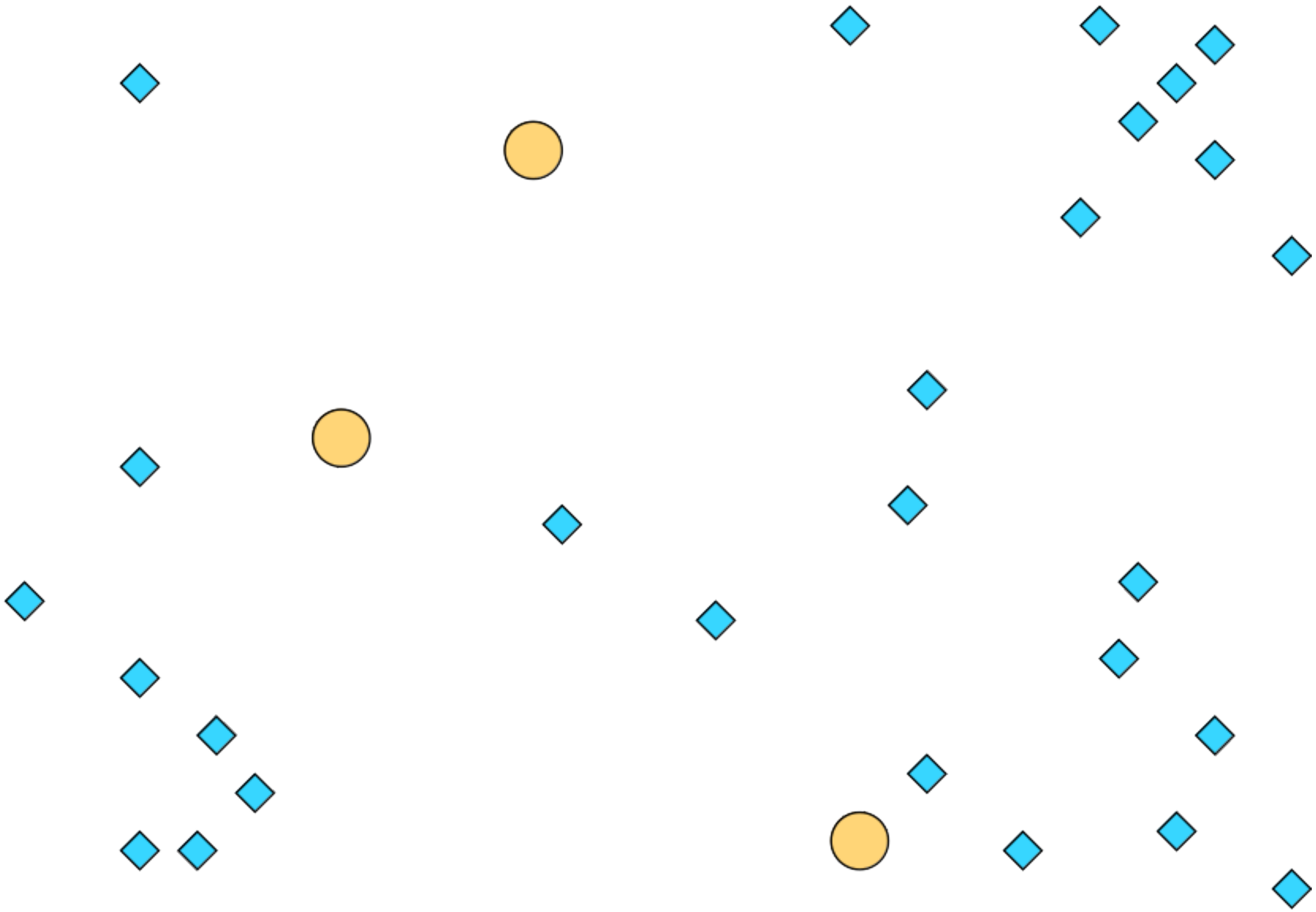
2. Assignment phase

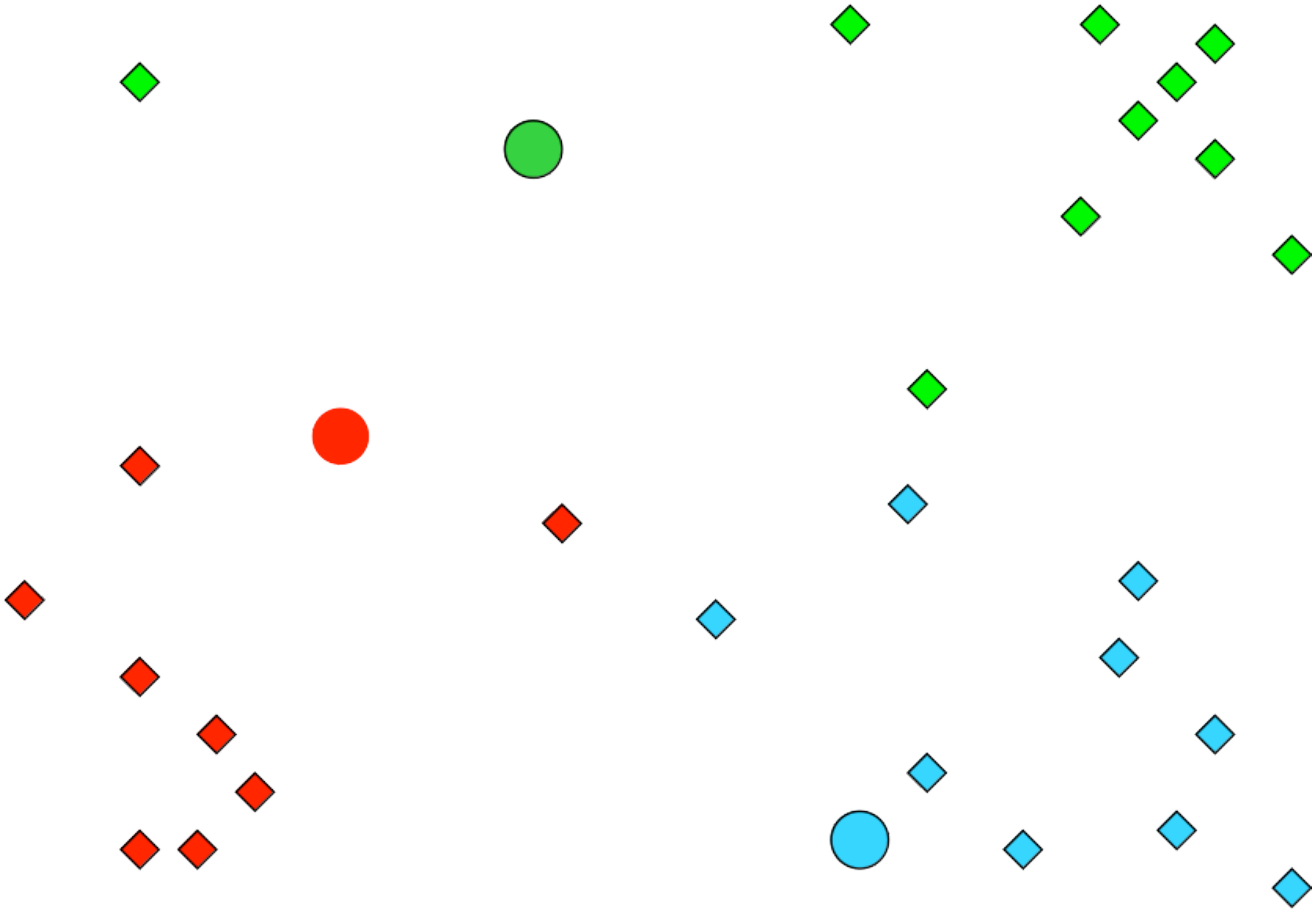
Assign all objects in the dataset to the closest representative with respect to the (squared) Euclidean distance. The resulting clusters: $C_1, \ldots, C_k$

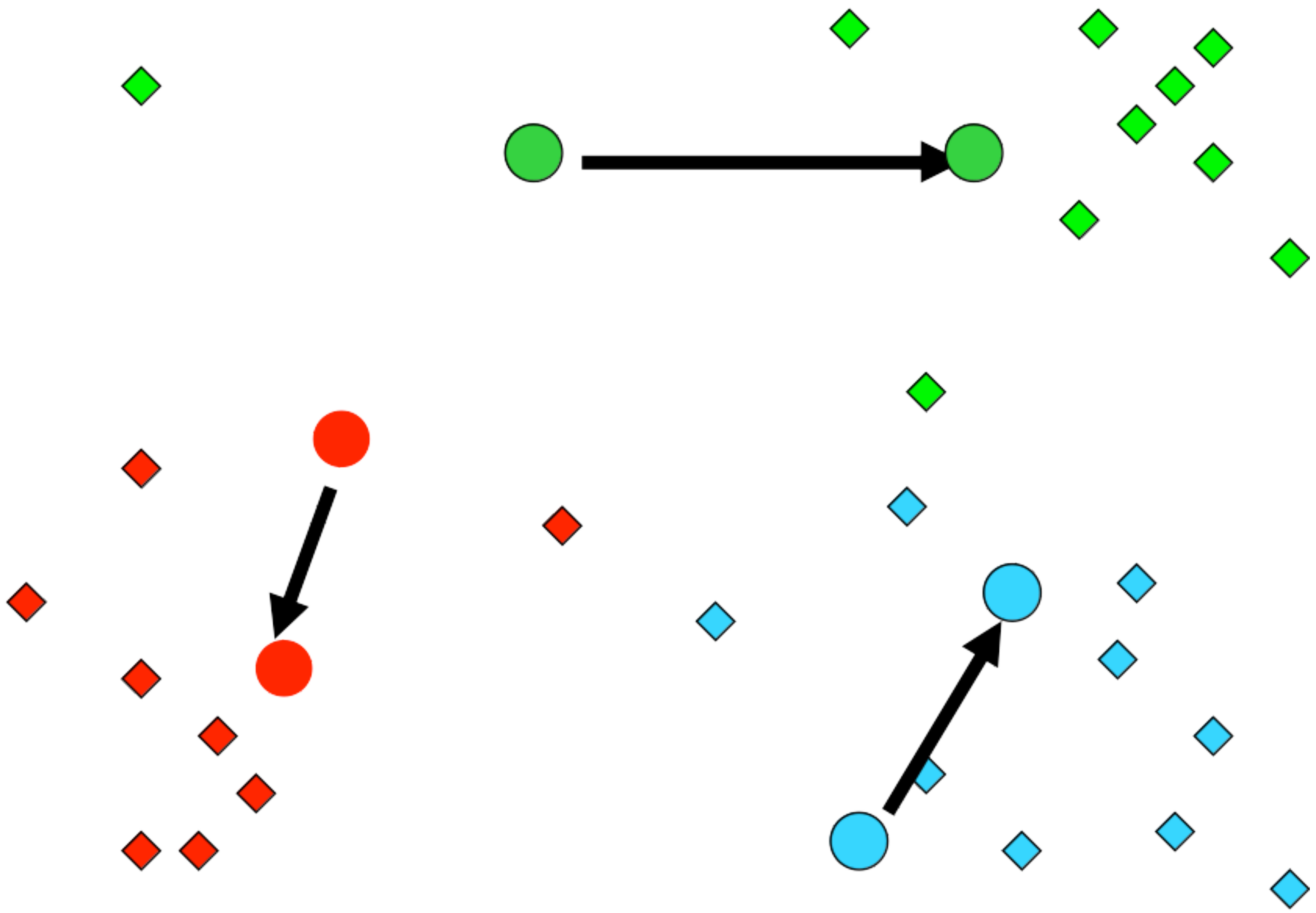
3. Optimisation phase

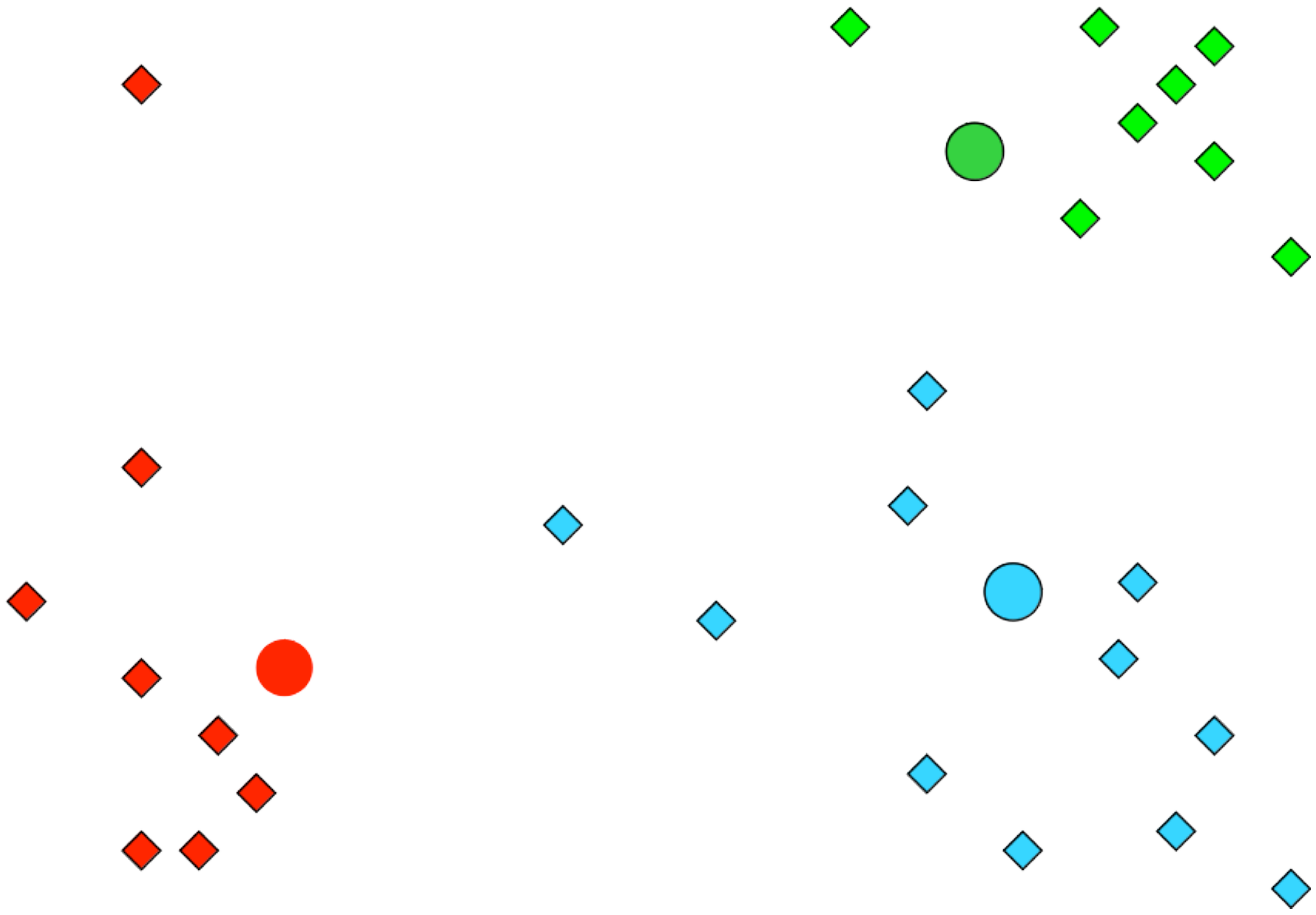
Compute the new representatives $\overline{Y}_1, \ldots, \overline{Y}_k$ as the means of the current clusters $C_1, \ldots, C_k$

Repeat Phases 2 and 3 until convergence. (Convergence is either “no objects have moved among clusters” or “fixed number of iterations specified by user”)









Issues with k-Means algorithm

- Results can vary depending on initial random choices
- Can get trapped in a local minimum that isn't the global optimal solution
  - Repeat the clustering procedure multiple times with different initialisations and select the *best* final clustering
- Outliers have a larger effect on the mean value, hence cluster centre and the cluster
- Cluster centres (means) are not actual instances in the cluster
- Euclidean distance used in the algorithm is inappropriate for categorical features