Big Data Summer School

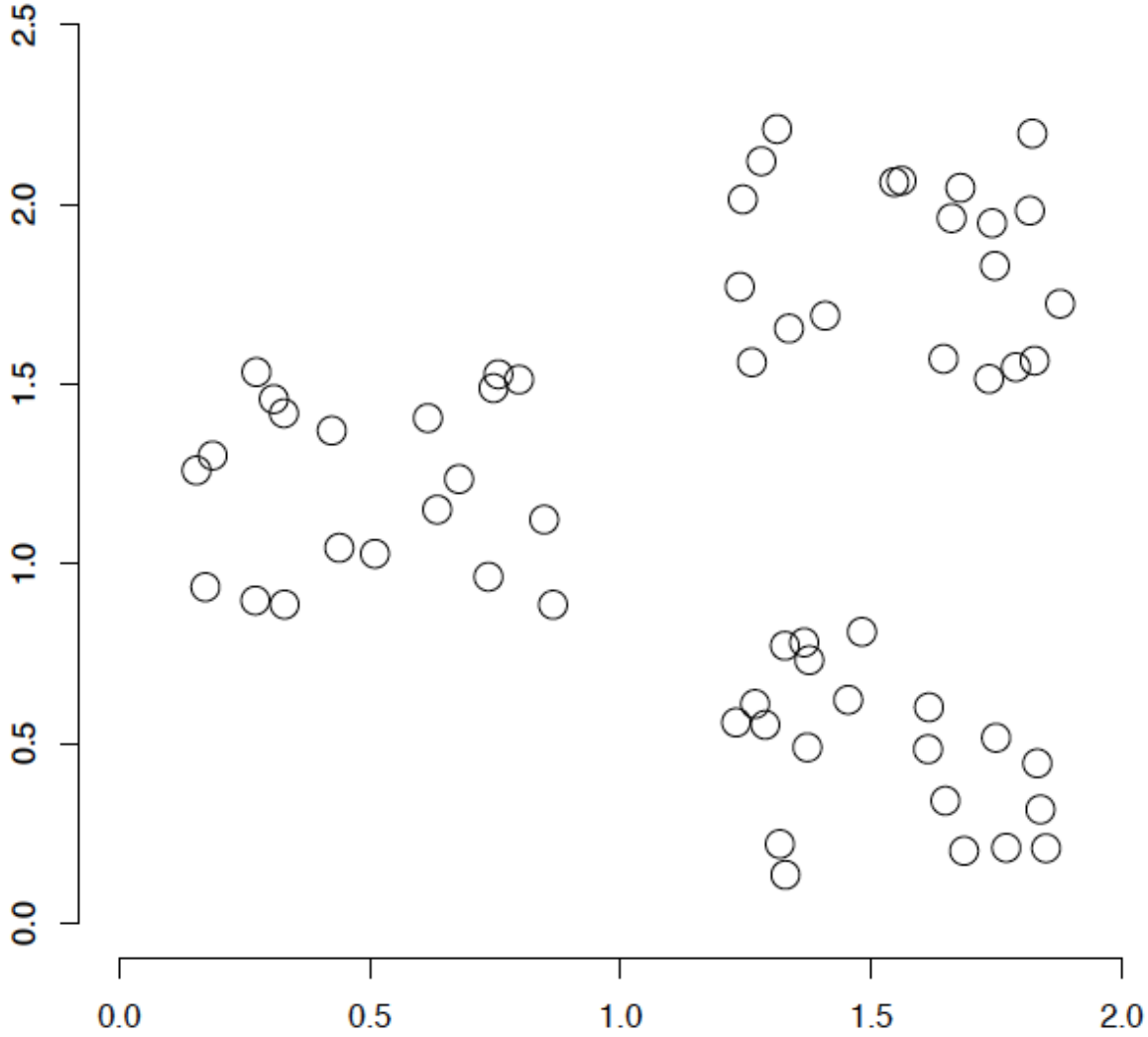
Clustering

the process of grouping a set of objects into classes of similar objects

Applications?

Order Prediction

A data set with clear cluster structure



What are some of the issues for clustering?

What clustering algorithms have you seen/used?

Issues for clustering

Representation for clustering

- How do we represent an example
  - features, etc.
- Similarity/distance between examples

Number of clusters

- Fixed a priori
- Data driven?

Hard vs. soft clustering

Hard clustering: Each example belongs to exactly one cluster

Soft clustering: An example can belong to more than one cluster (probabilistic)

- Makes more sense for applications like creating browsable hierarchies
- You may want to put a pair of sneakers in two clusters: (i) sports apparel and (ii) shoes

K-means

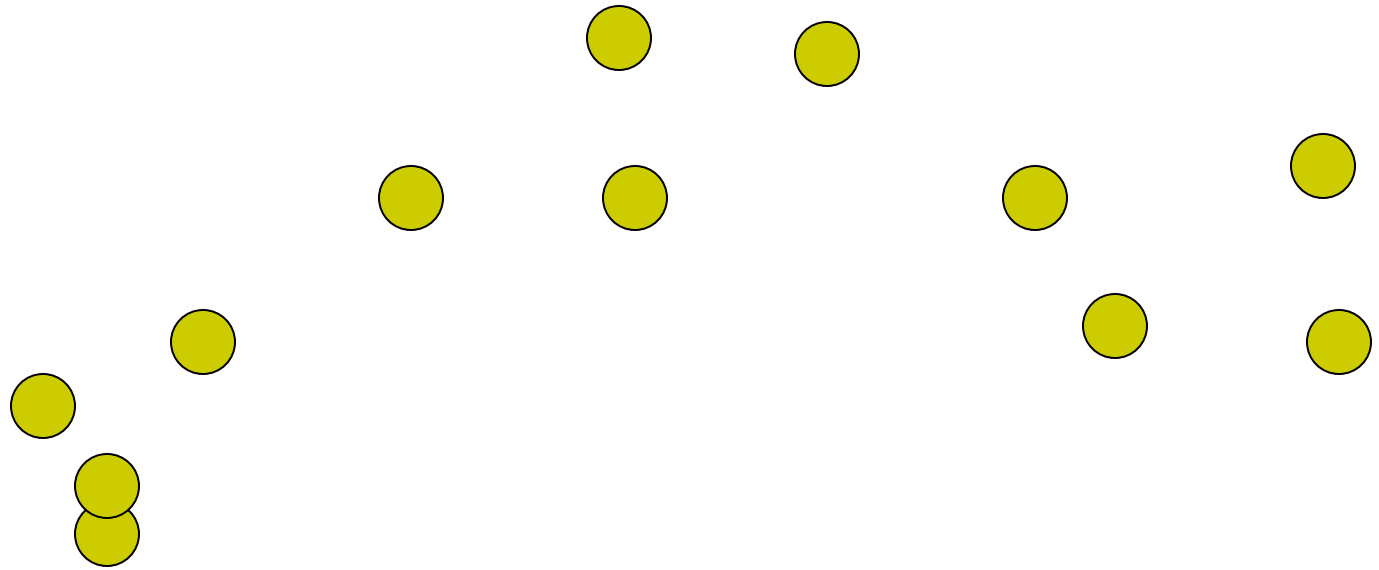
Most well-known and popular clustering algorithm:

Start with some initial cluster centers

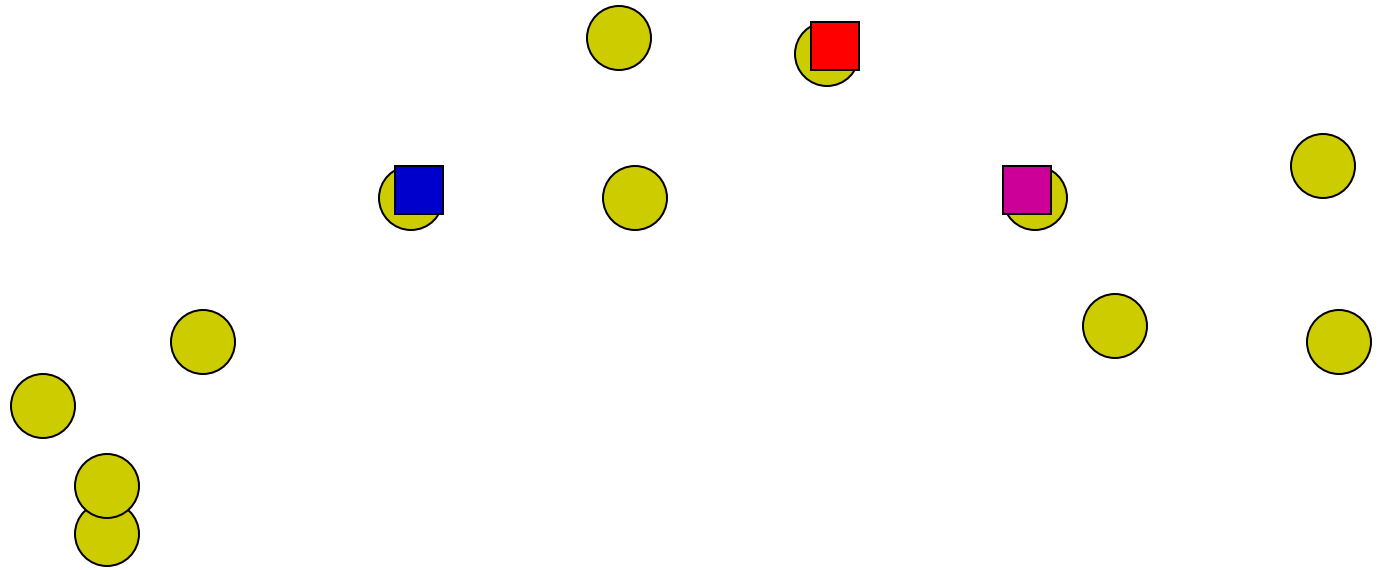
Iterate:

- Assign/cluster each example to closest center
- Recalculate centers as the mean of the points in a cluster

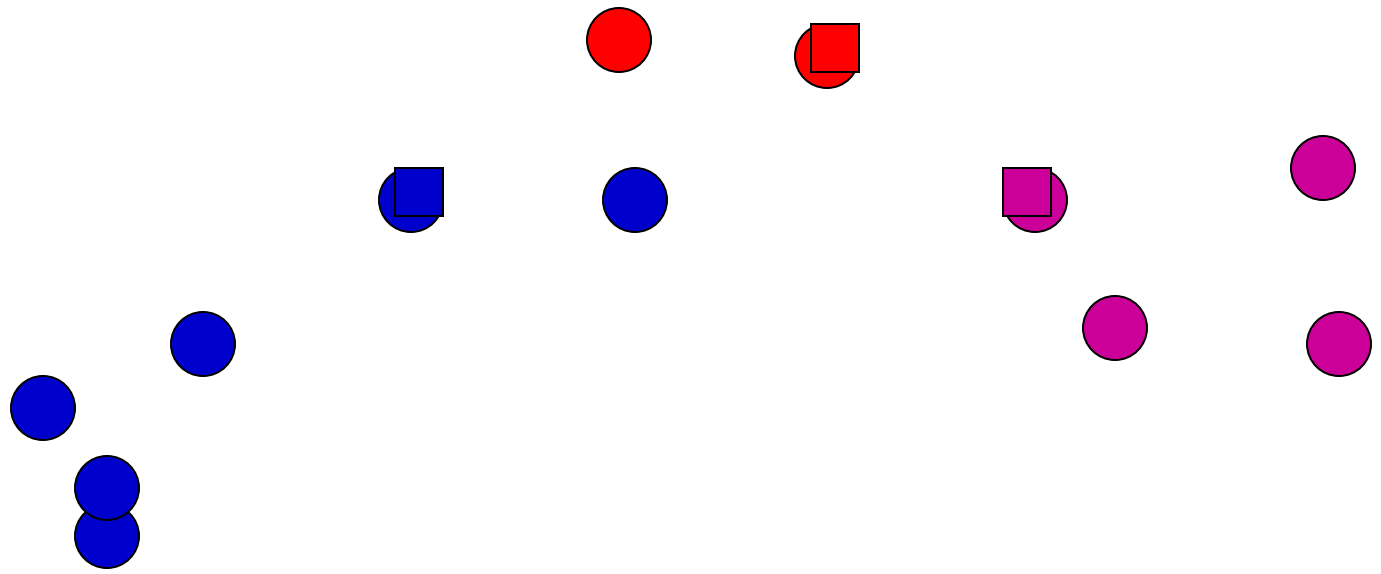
K-means: an example



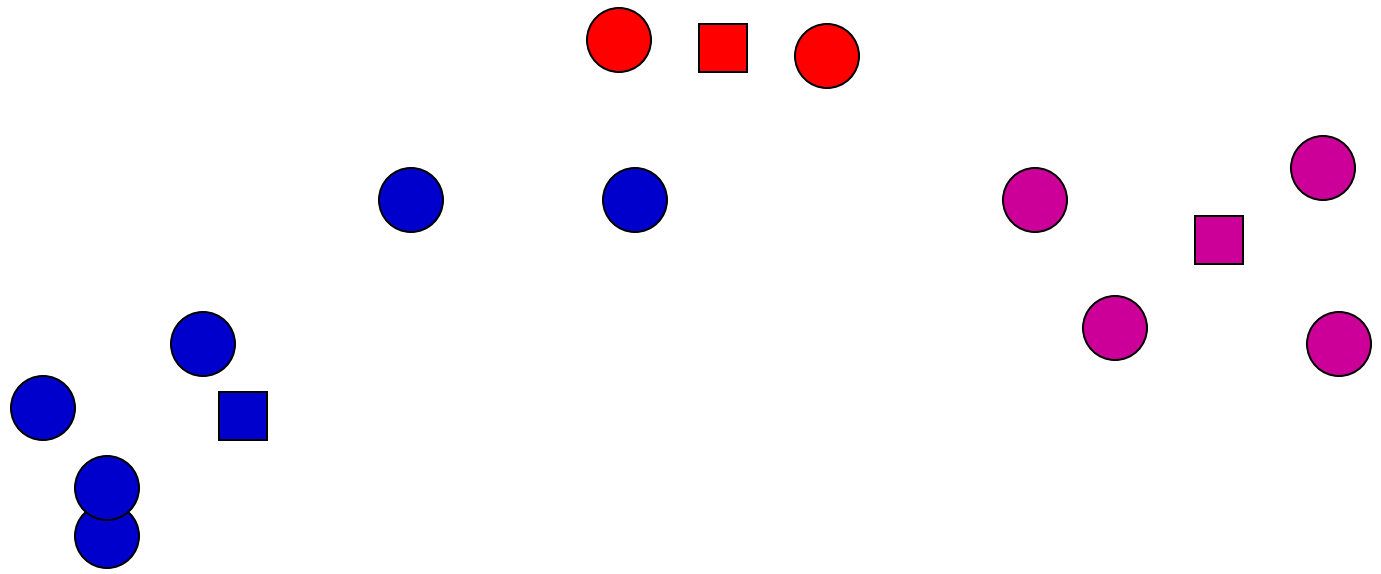
K-means: Initialize centers randomly



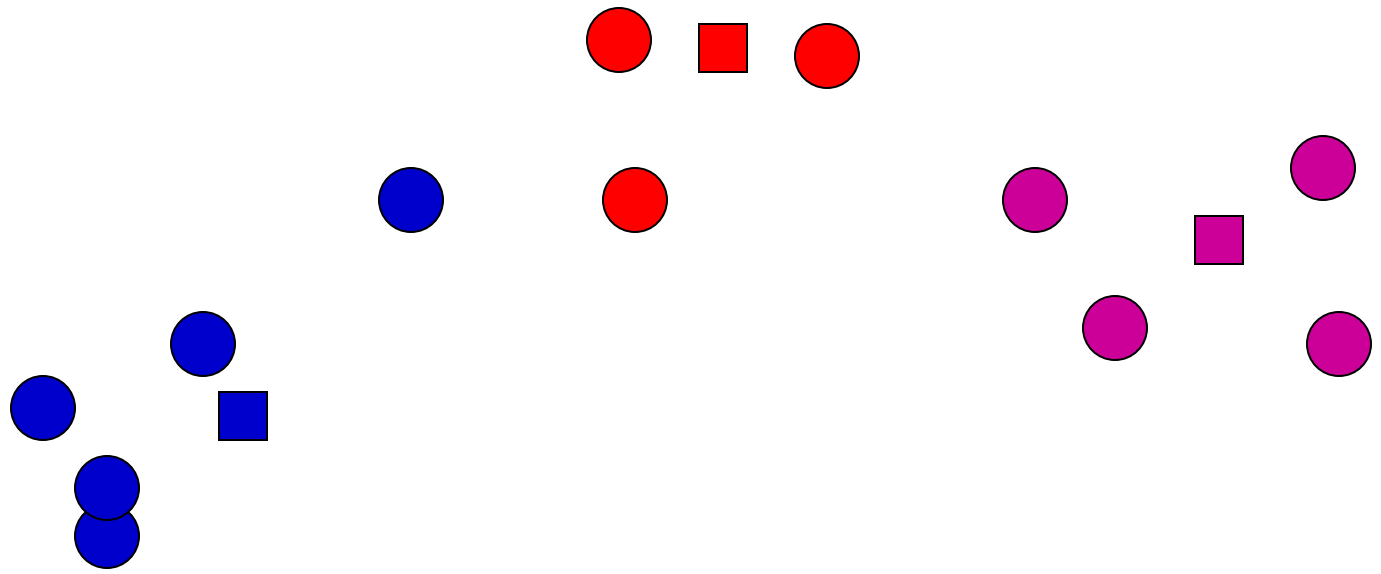
K-means: assign points to nearest center



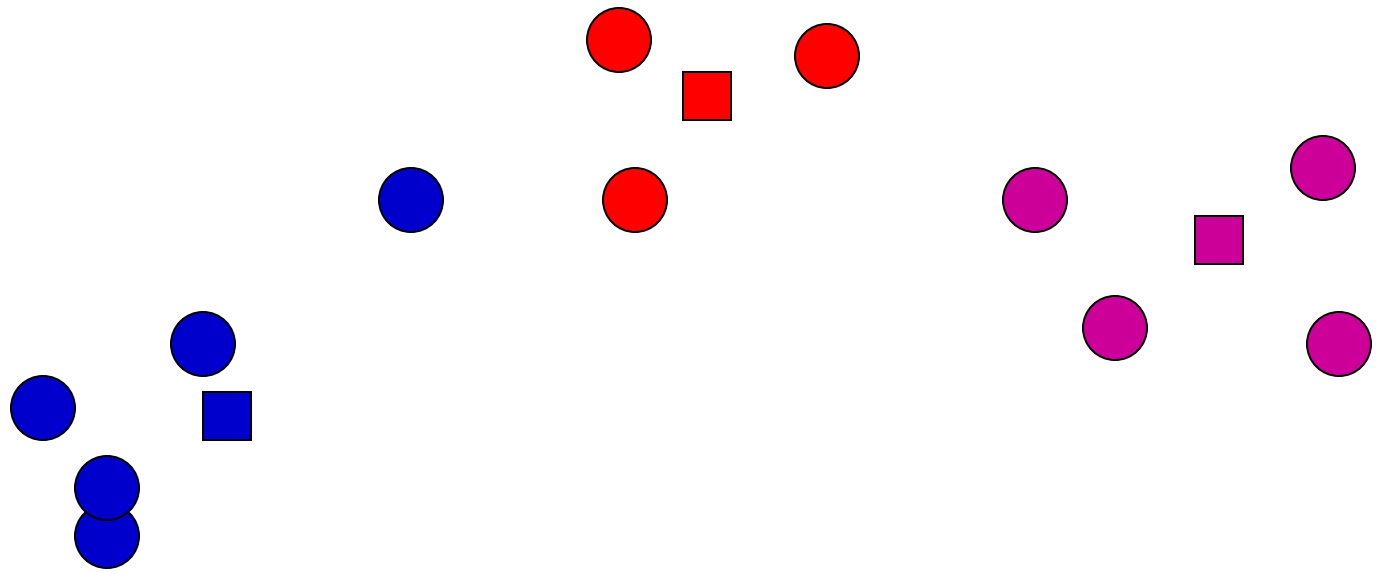
K-means: readjust centers



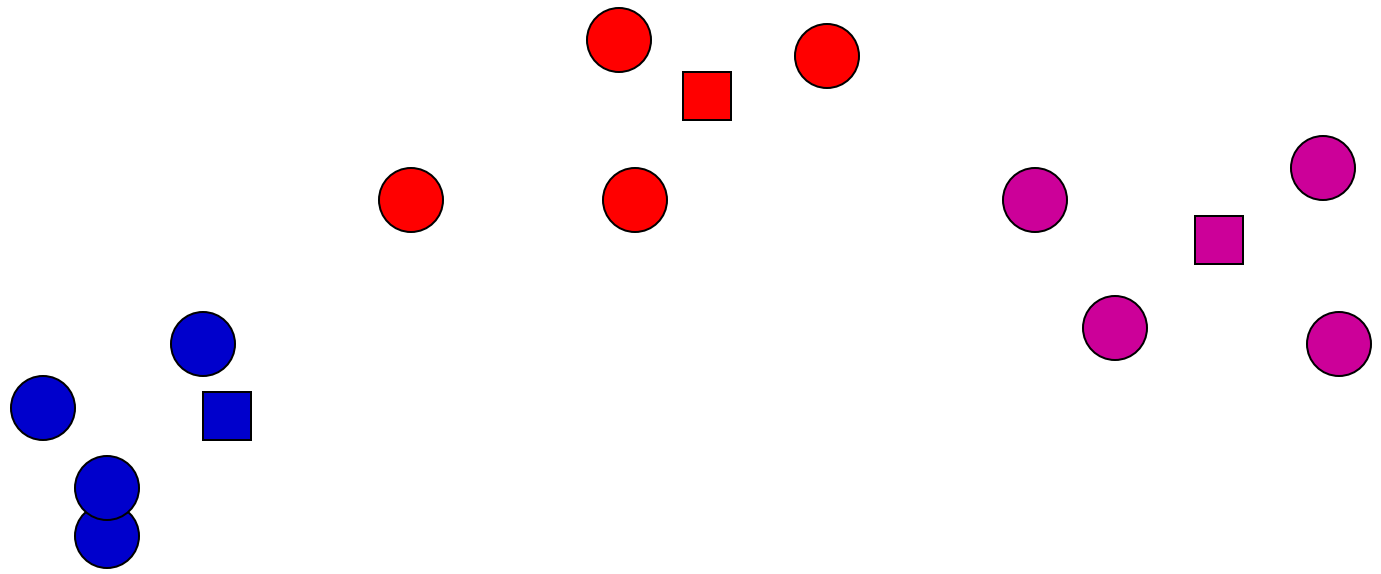
K-means: assign points to nearest center



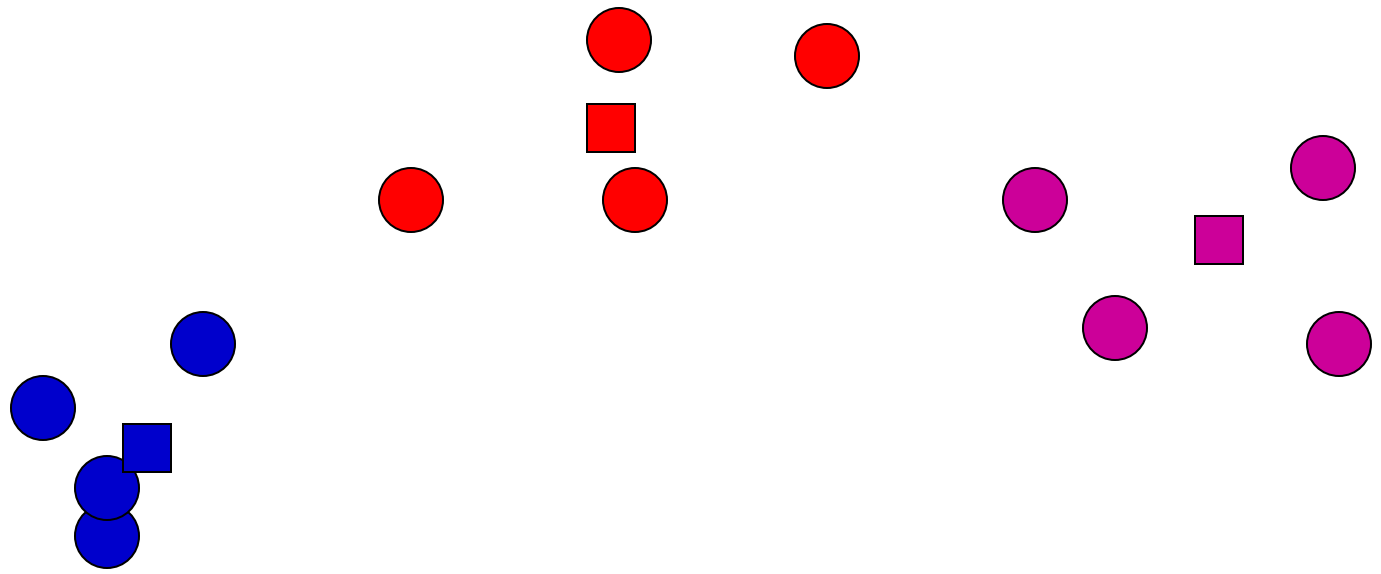
K-means: readjust centers



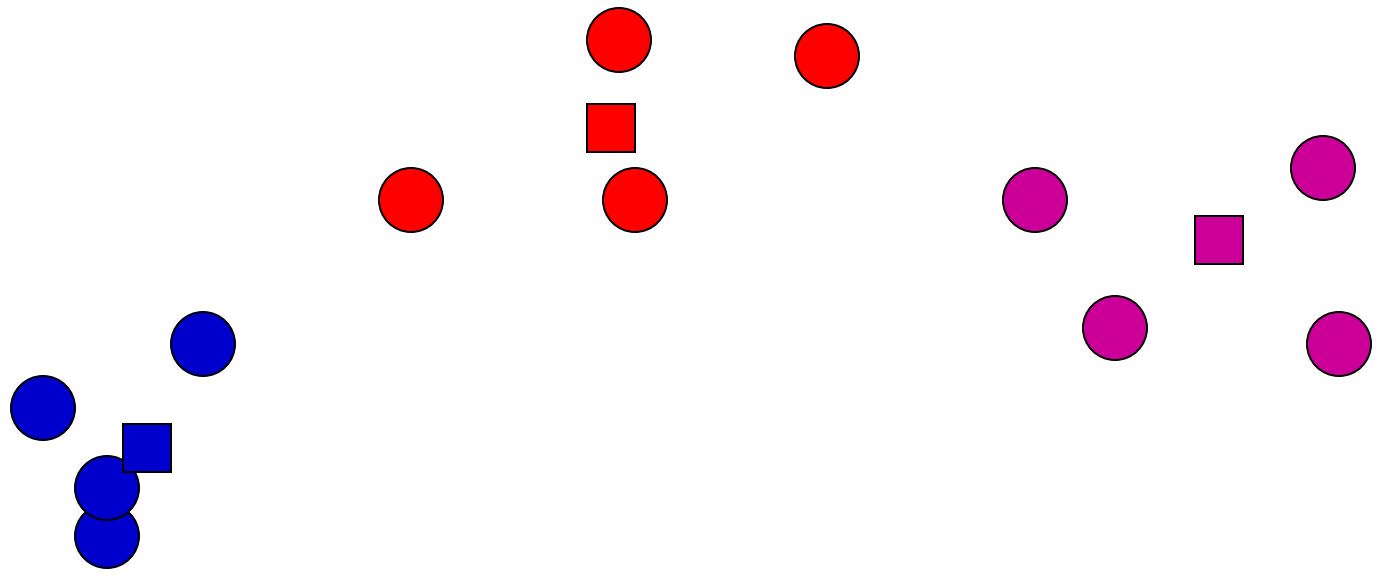
K-means: assign points to nearest center



K-means: readjust centers



K-means: assign points to nearest center

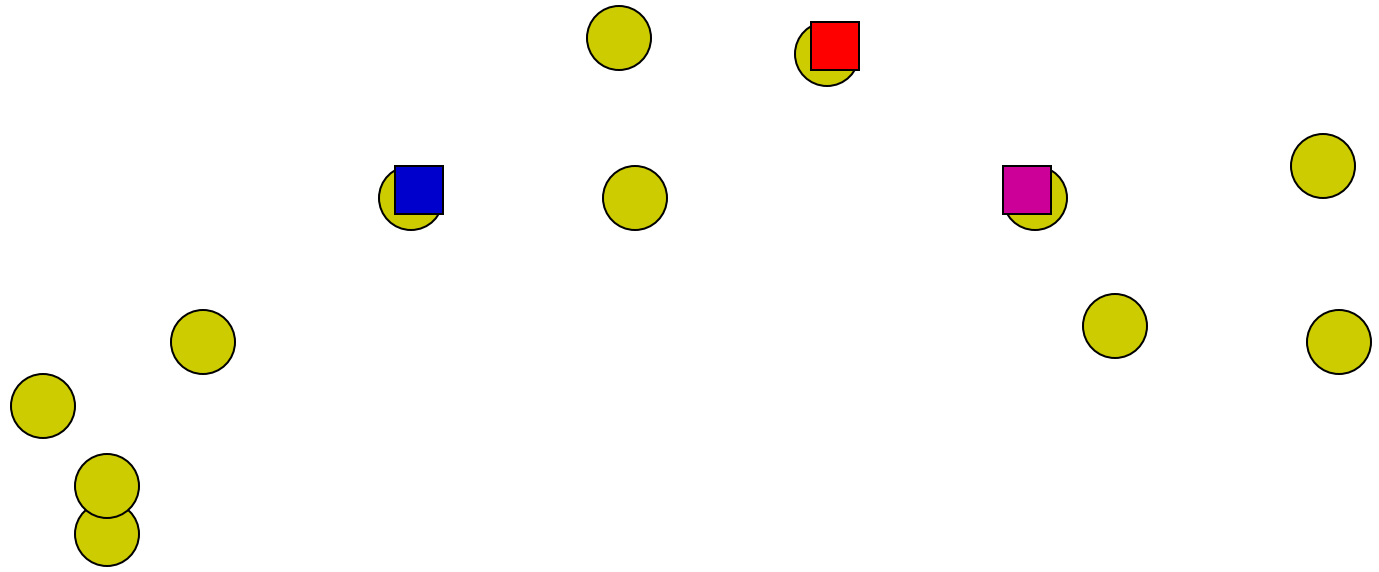


No changes: Done

K-means

Iterate:

- **Assign/cluster each example to closest center**
- Recalculate centers as the mean of the points in a cluster



How do we do this?

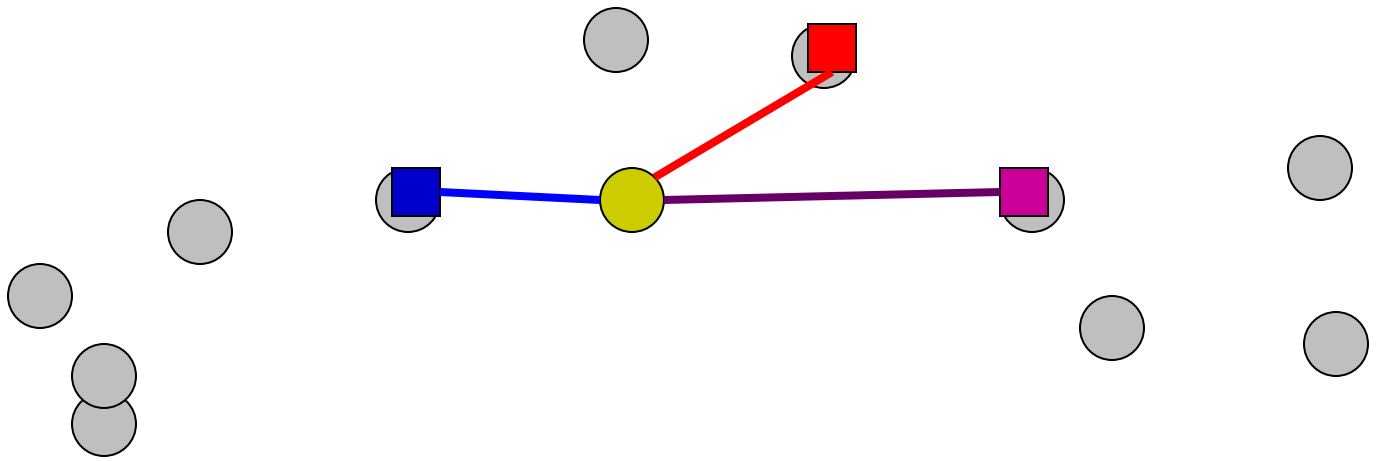
K-means

Iterate:

- **Assign/cluster each example to closest center**

  iterate over each point:
  - get distance to each cluster center
  - assign to closest center (hard cluster)

- **Recalculate centers as the mean of the points in a cluster**



K-means

Iterate:

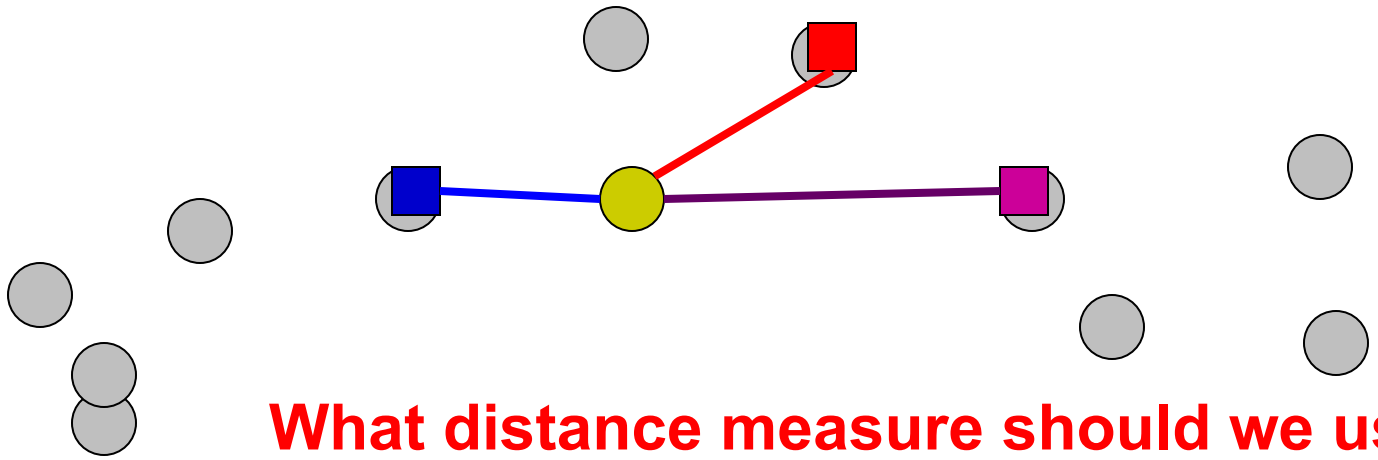
- **Assign/cluster each example to closest center**

  iterate over each point:

  - get distance to each cluster center

  - assign to closest center (hard cluster)

- **Recalculate centers as the mean of the points in a cluster**



What distance measure should we use?

Distance measures

Euclidean:

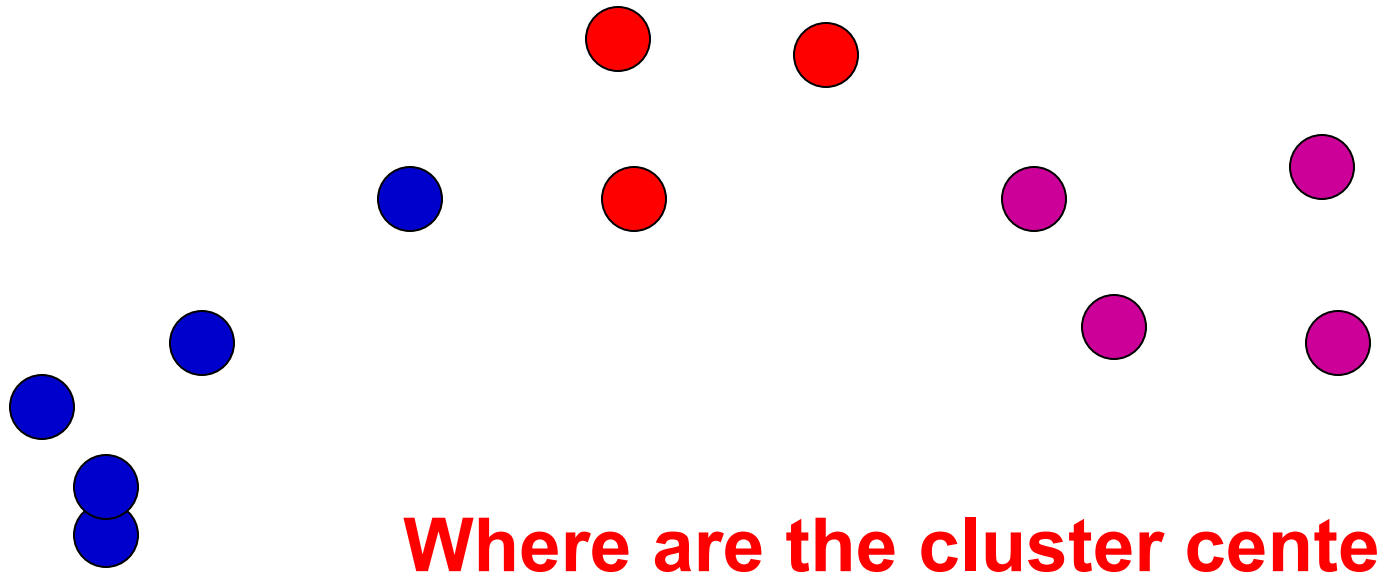
$$d(x,y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

good for spatial data

K-means

Iterate:

- Assign/cluster each example to closest center
- Recalculate centers as the mean of the points in a cluster

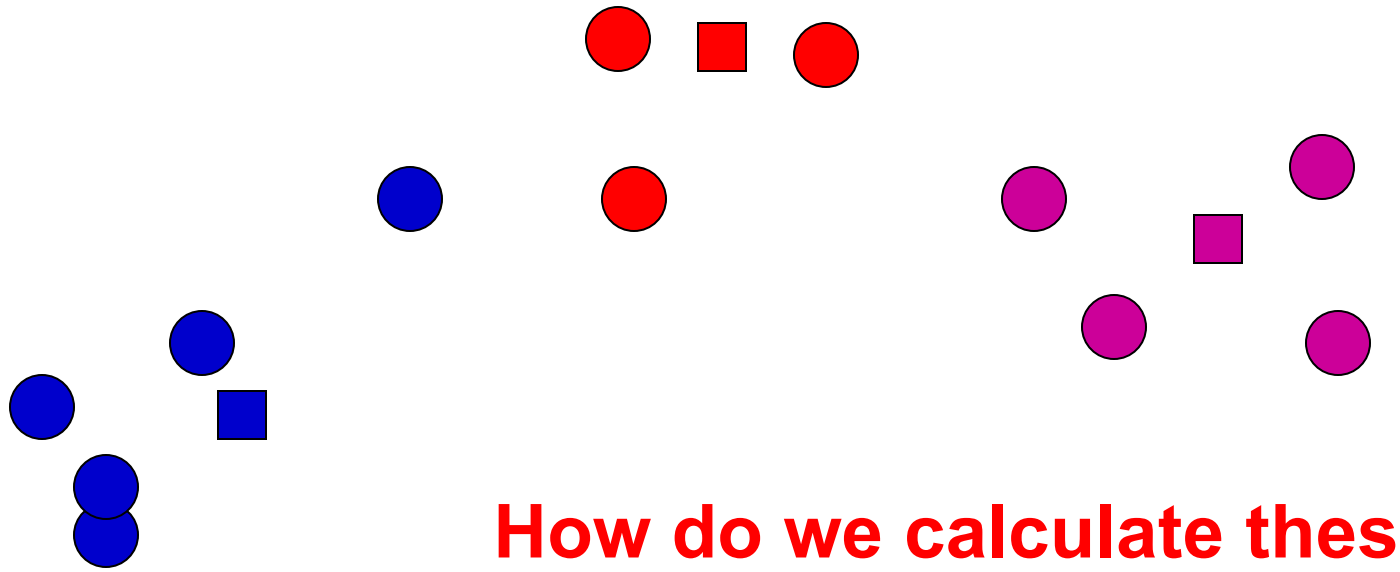


Where are the cluster centers?

K-means

Iterate:

- Assign/cluster each example to closest center
- Recalculate centers as the mean of the points in a cluster



How do we calculate these?

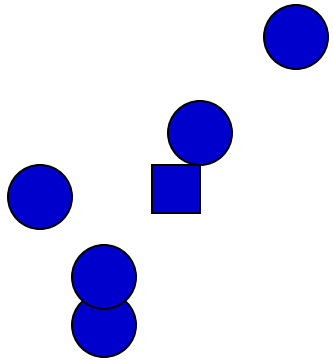
K-means

Iterate:

- Assign/cluster each example to closest center
- Recalculate centers as the mean of the points in a cluster

Mean of the points in the cluster:

$$\mu(\mathrm{C}) = \frac{1}{|C|}\sum_{x \in C} x$$



K-means loss function

K-means tries to minimize what is called the “k-means” loss function:

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2$$

S_i: class i
u_i: center of S_i

that is, minimize the sum of the squared distances from each point to the associated cluster center

K-means algorithm

1) Pick a number (k) of cluster centers
2) Assign every object to its nearest cluster center
3) Move each cluster center to the mean of its assigned objects
4) Repeat 2-3 until convergence

K-means variations/parameters

Start with some initial cluster centers

Iterate:

- **Assign/cluster each example to closest center**
- **Recalculate centers as the mean of the points in a cluster**

What are some other variations/parameters we haven't specified?

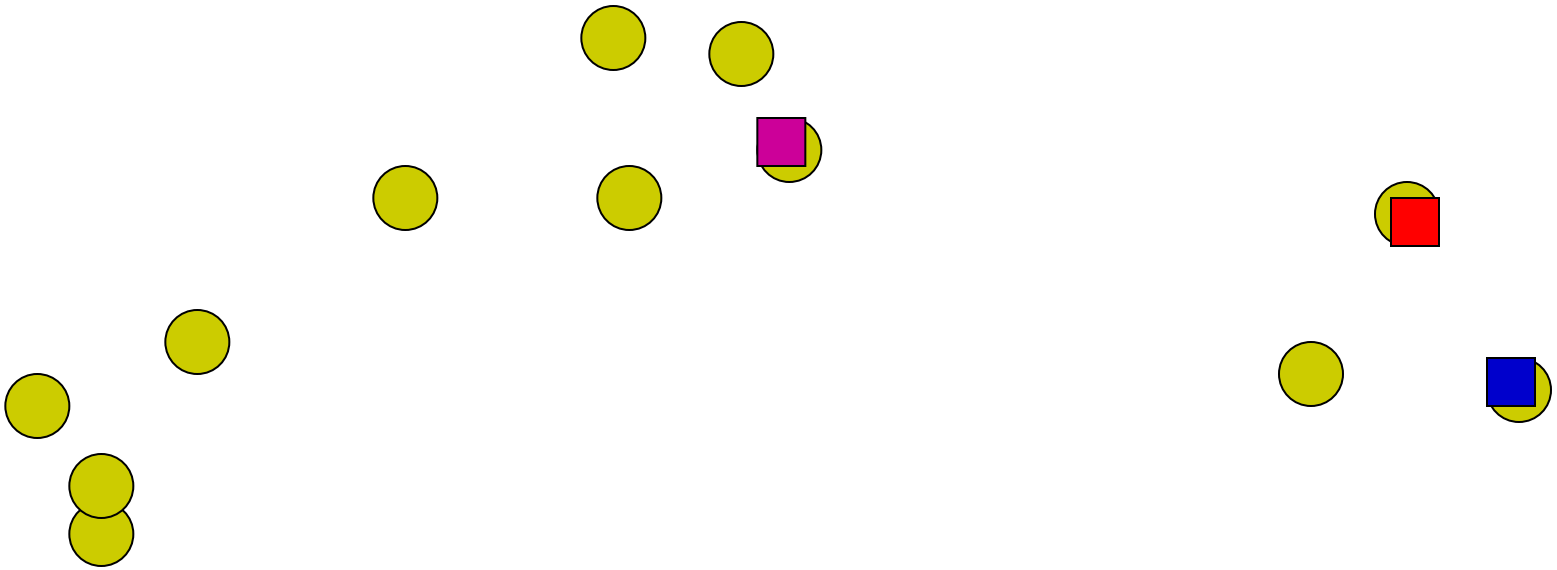
K-means variations/parameters

Initial (seed) cluster centers

Convergence

- A fixed number of iterations
- partitions unchanged
- Cluster centers don’t change

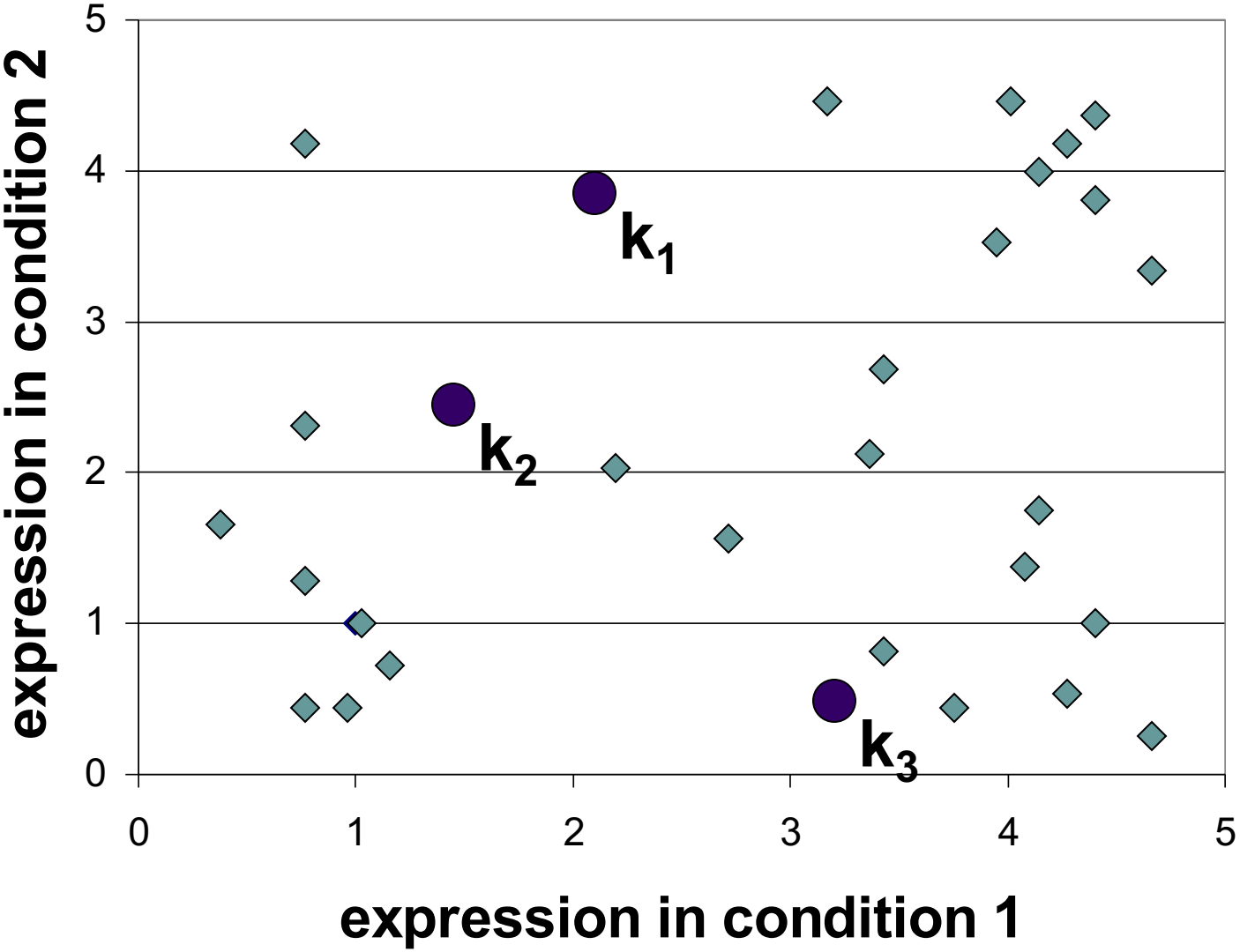
K-means: Initialize centers randomly



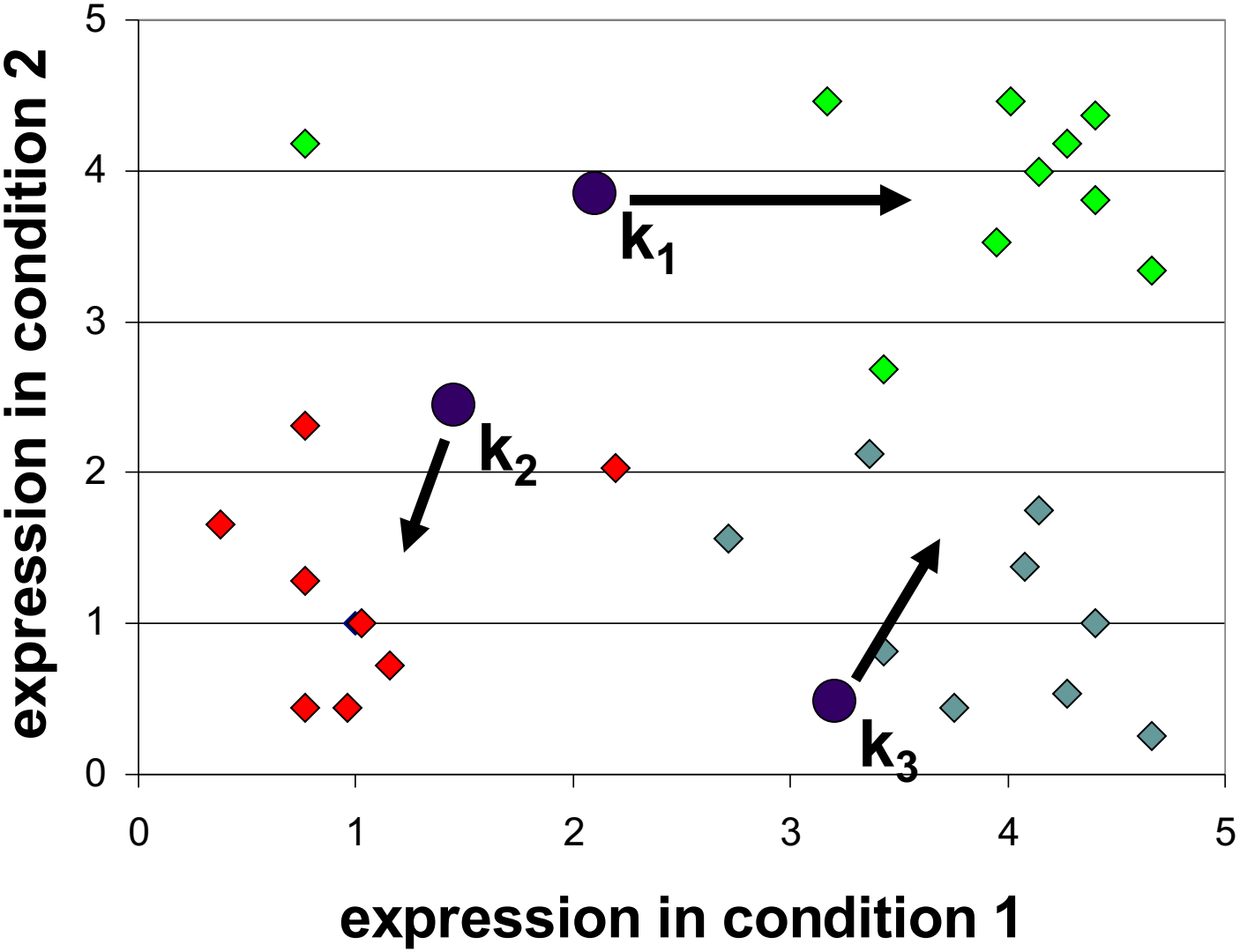
What would happen here?

Seed selection ideas?

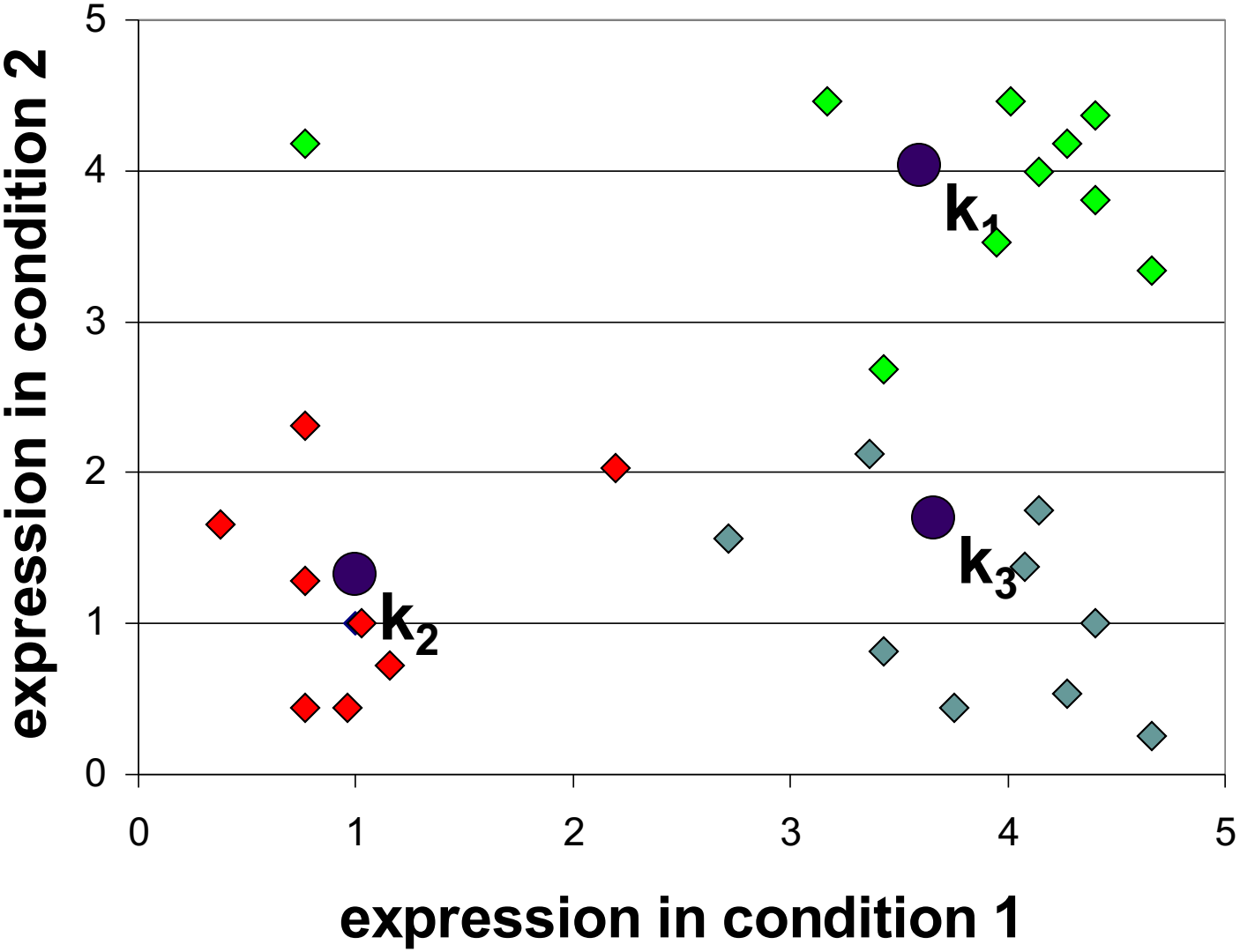
Clustering: Example 2, Step 1



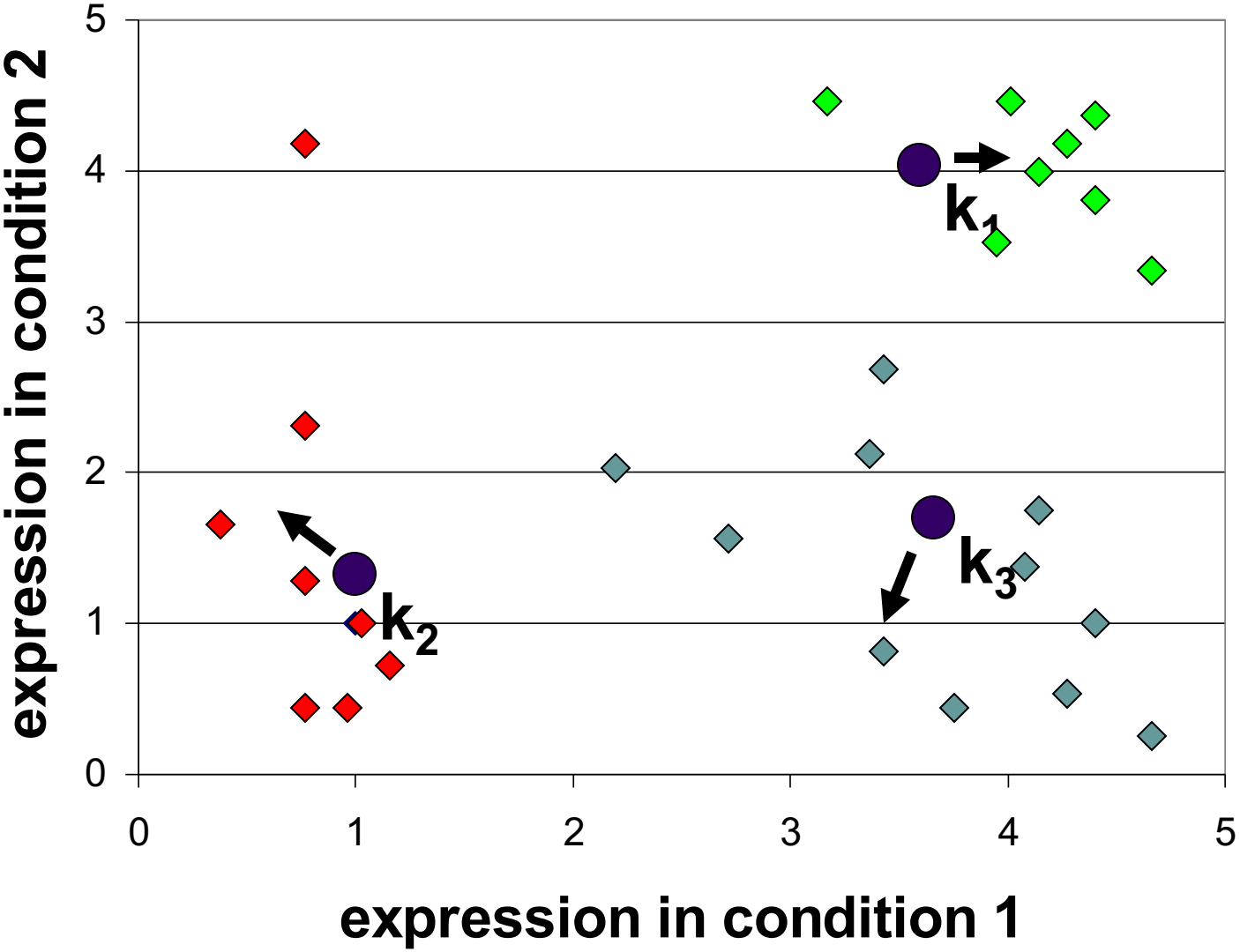
Clustering: Example 2, Step 2



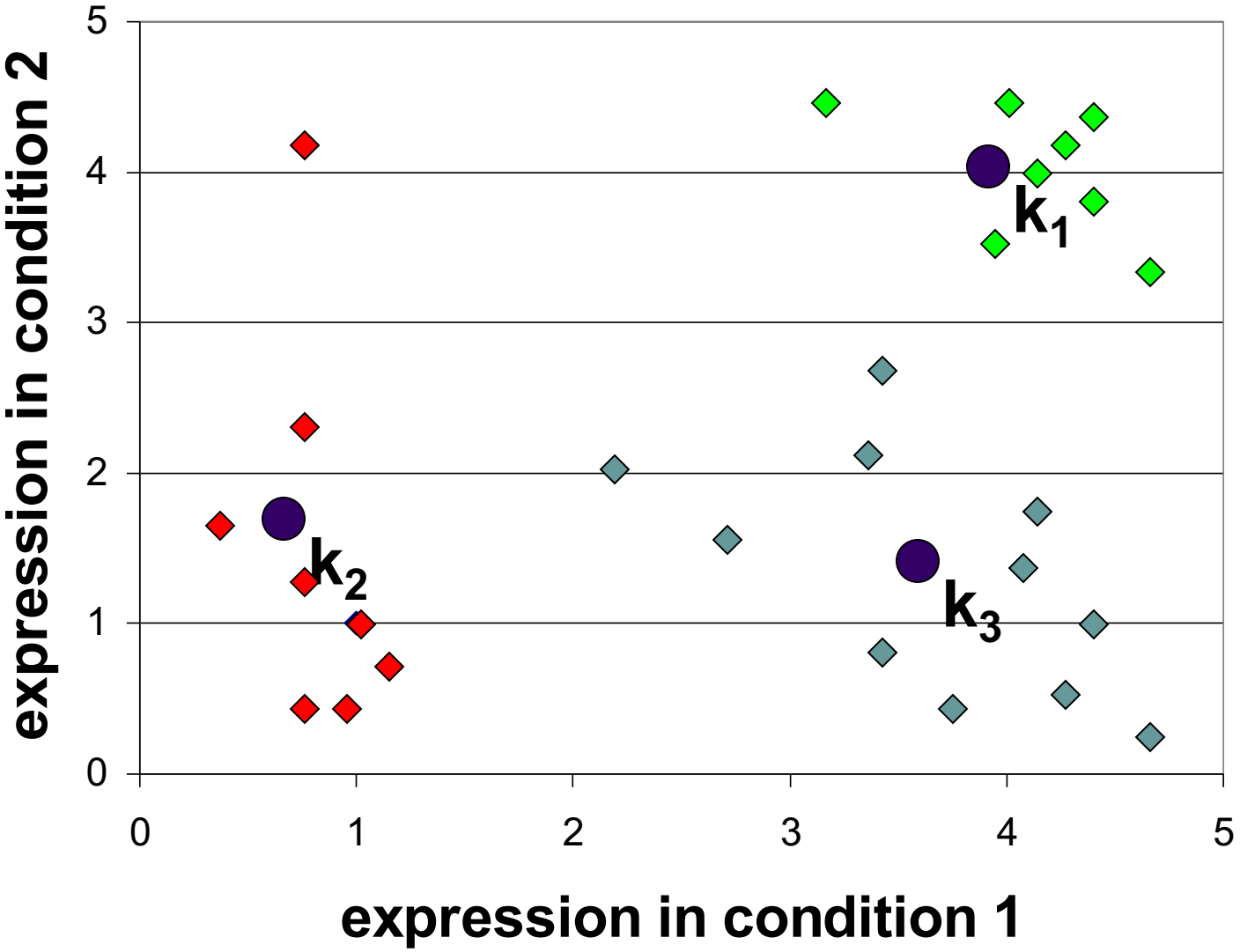
Clustering: Example 2, Step 3



Clustering: Example 2, Step 4



Clustering: Example 2, Step 5



Seed choice

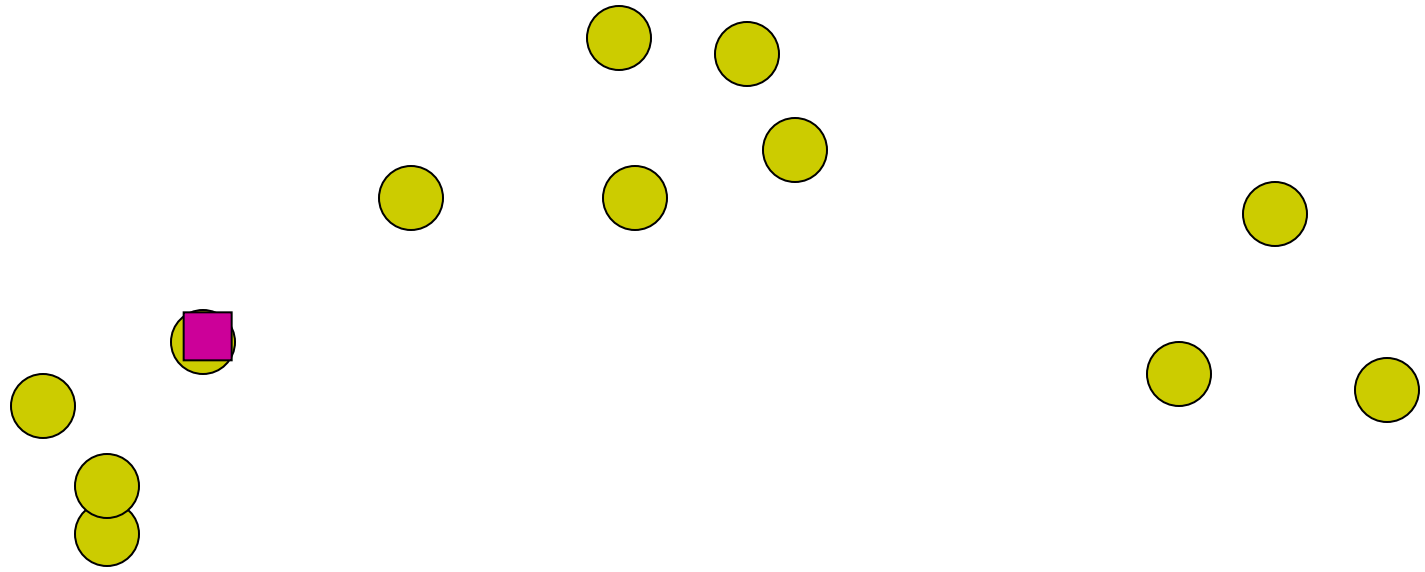
Results can vary drastically based on random seed selection

Some seeds can result in poor convergence rate, or convergence to sub-optimal clusterings

Common heuristics

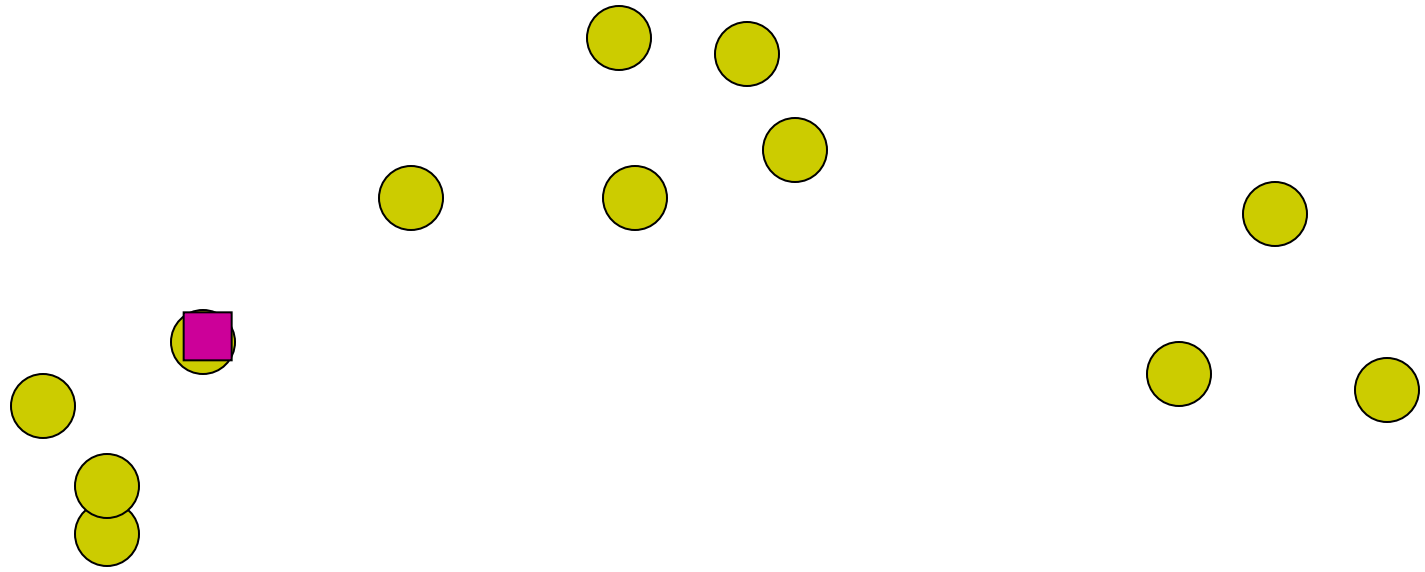
- Random centers in the space
- Randomly pick examples
- Points least similar to any existing center (furthest centers heuristic)
- **Try out multiple starting points**
- Initialize with the results of another clustering method

K-means: Initialize furthest from centers



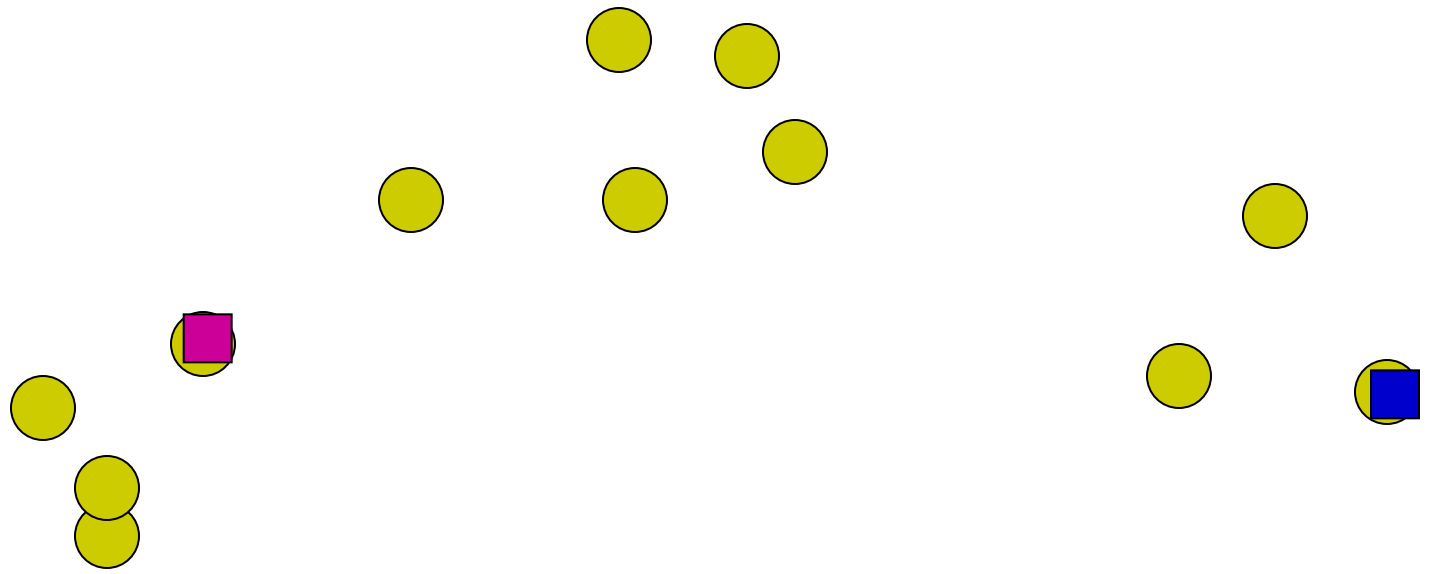
Pick a random point for the first center

K-means: Initialize furthest from centers



What point will be chosen next?

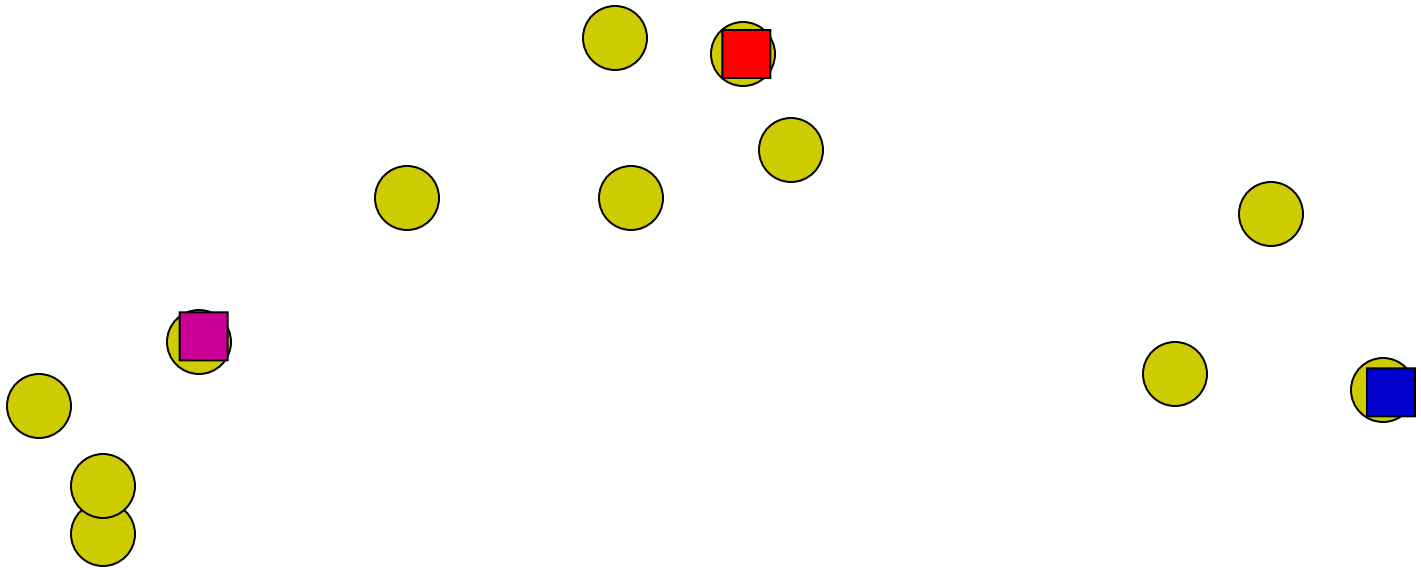
K-means: Initialize furthest from centers



Furthest point from center

What point will be chosen next?

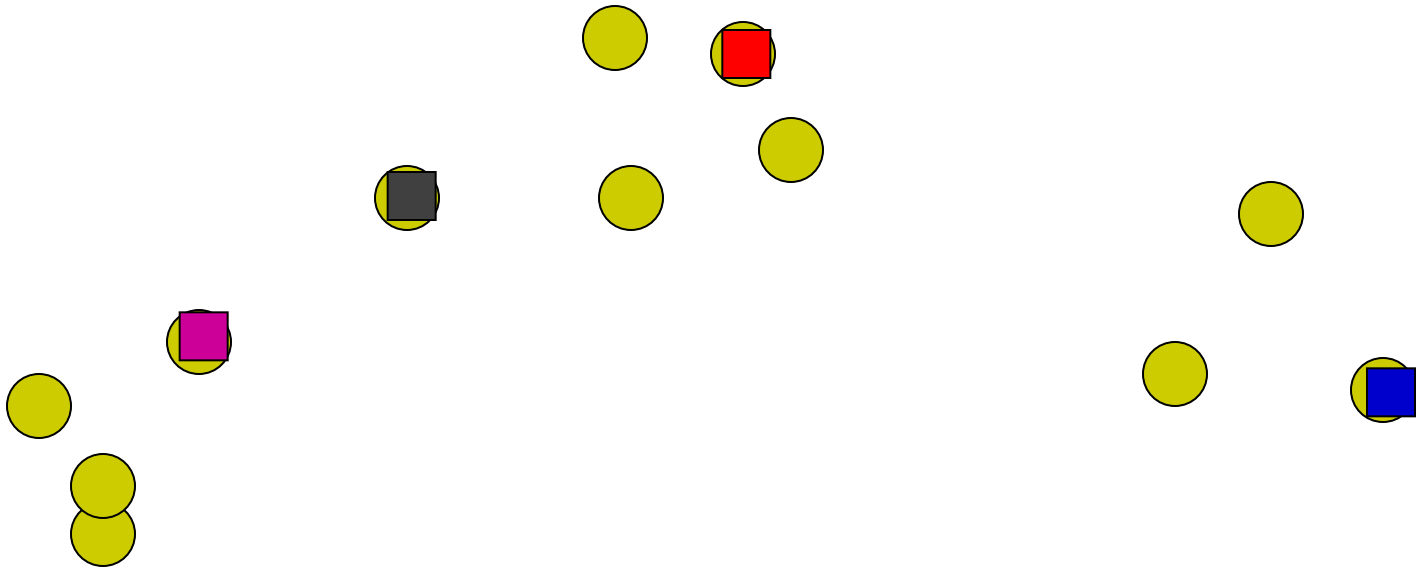
K-means: Initialize furthest from centers



Furthest point from center

What point will be chosen next?

K-means: Initialize furthest from centers



Furthest point from center

Any issues/concerns with this approach?