A Quick Introduction to Machine Learning (K-means Clustering)

Lecturer: John Guttag

K-means Clustering

Given a set of points *X*, and a positive integer *k*, partition *X* into *k* clusters such that it approximately minimizes the objective function

$$\sum_{c=1}^{k}\sum_{x\in c}\left\|x-\mu_c\right\|^2$$

Minimizing the sum of the mean square differences

clusters

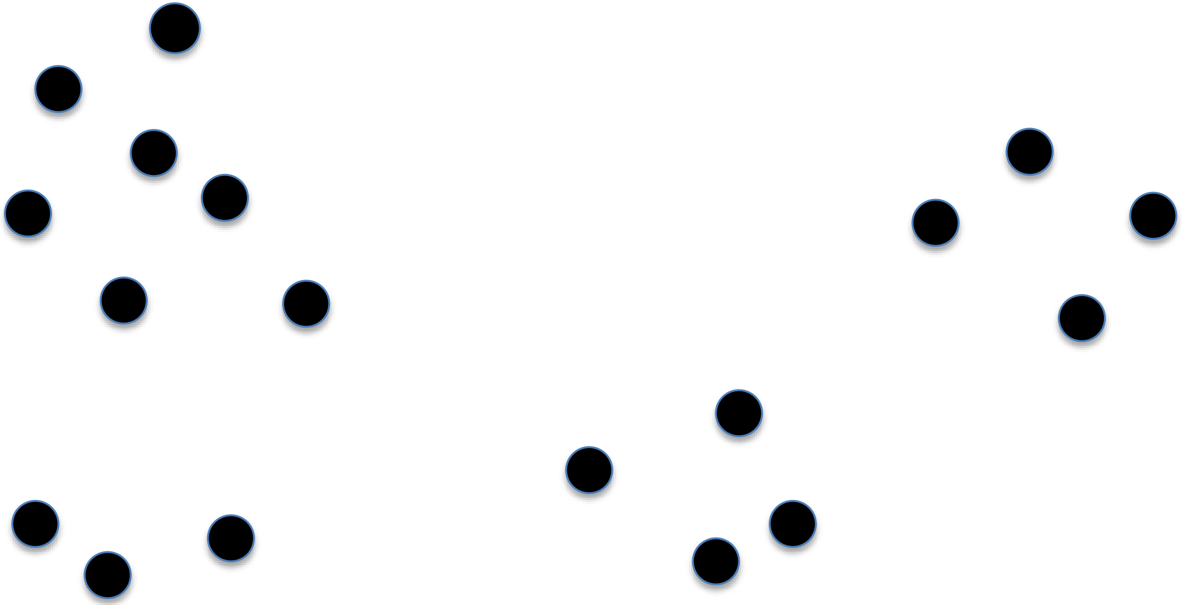
points

distance from point to centroid of cluster

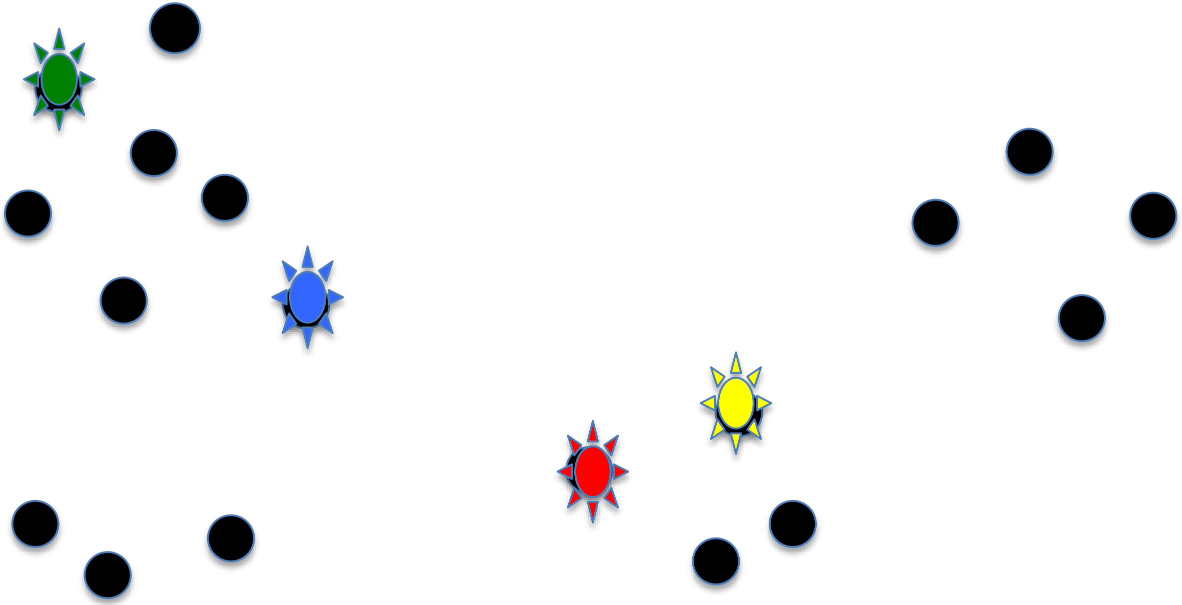
K-means Algorithm

```
randomly choose k examples as centroids
while true:
   create k clusters by assigning each
       example to closest centroid
   compute k new centroids by averaging
       examples in each cluster
   if centroids don’t change:
       break
```

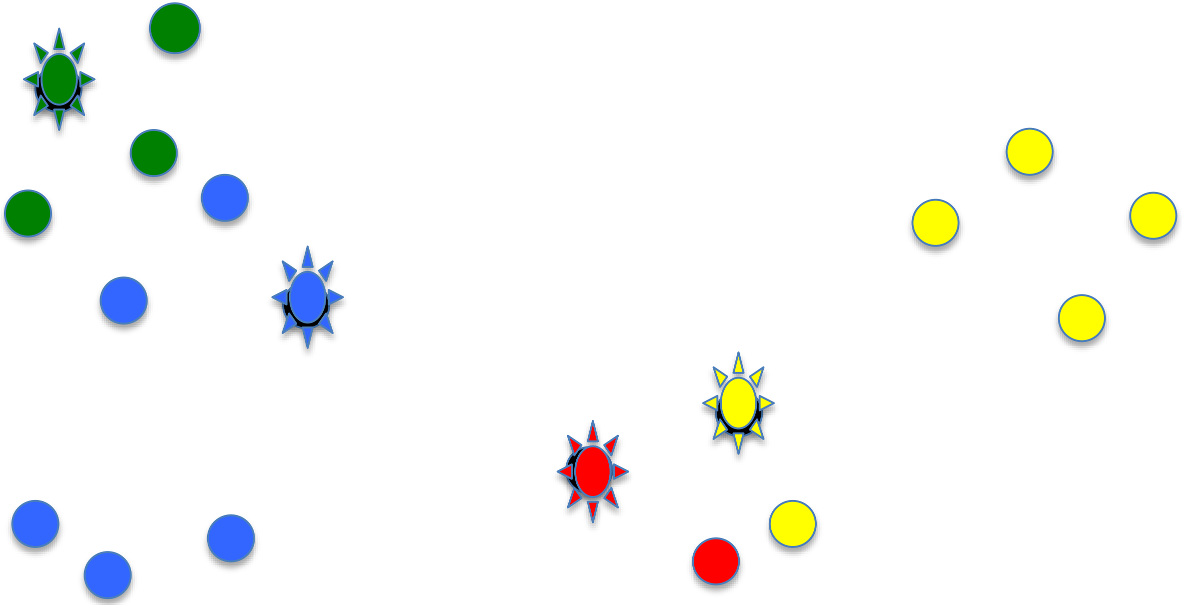
Example



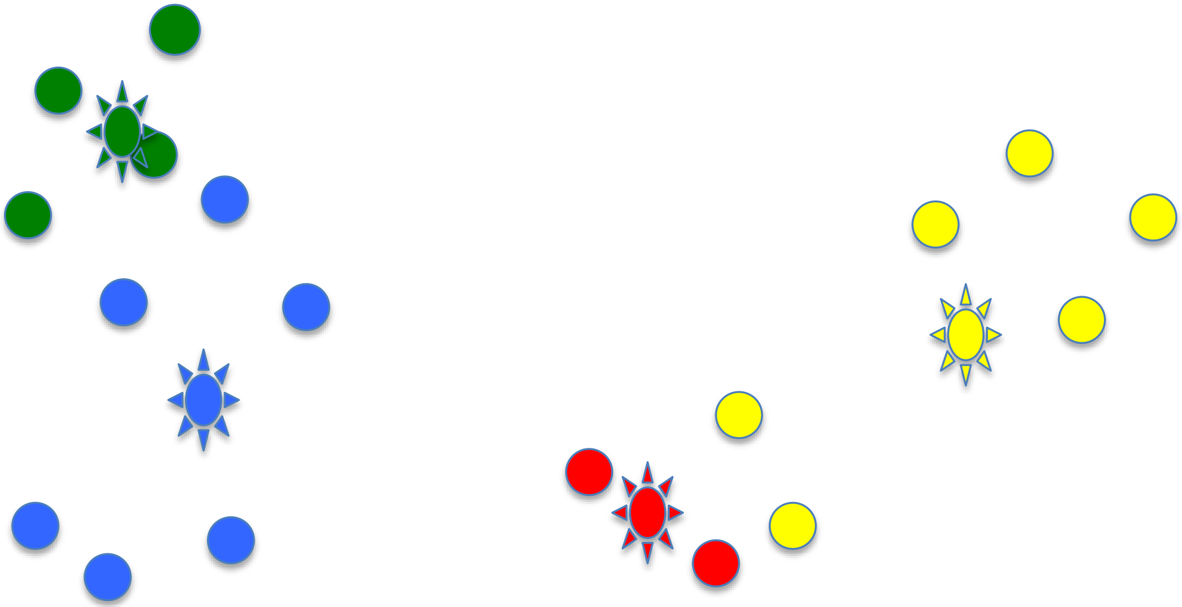
Choose Initial Centroids (k = 4)



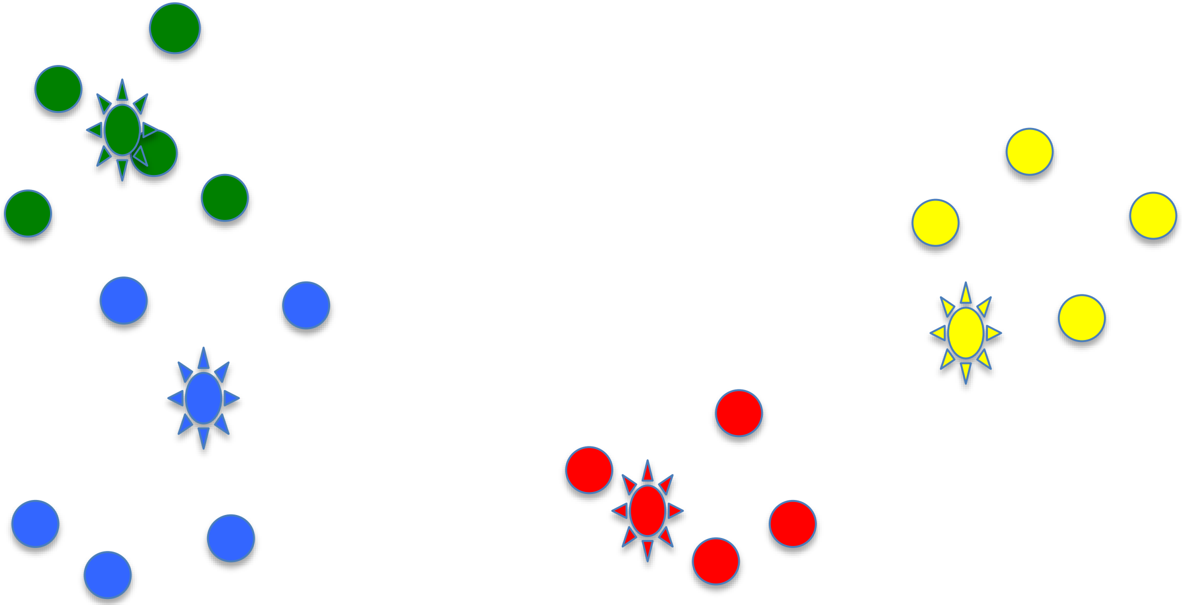
Assign Points to Clusters



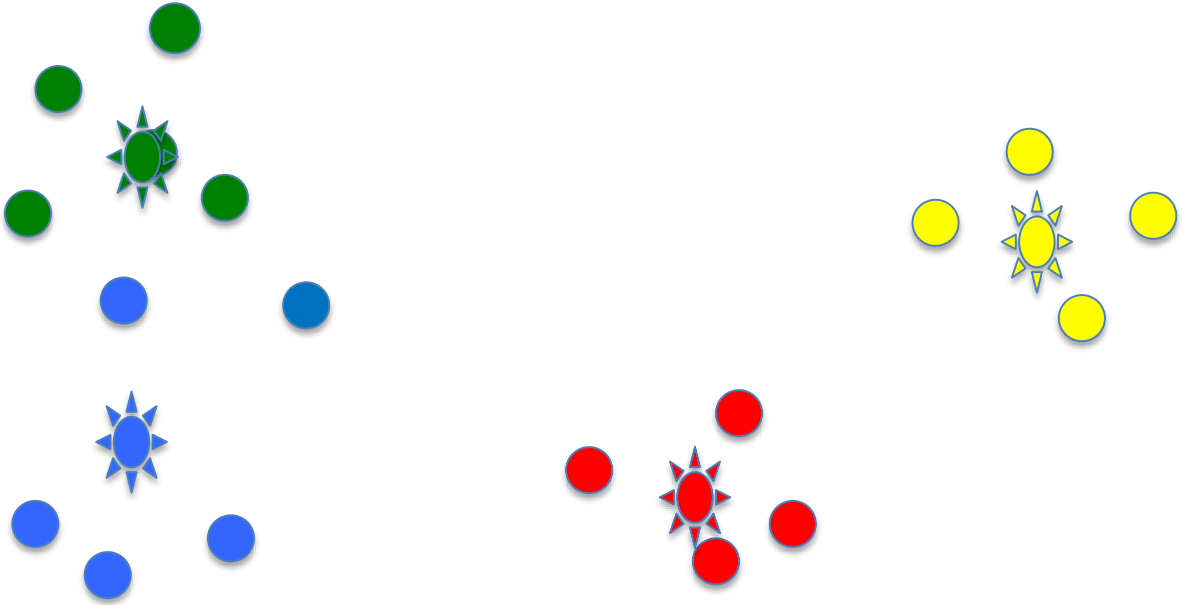
Compute New Centroids



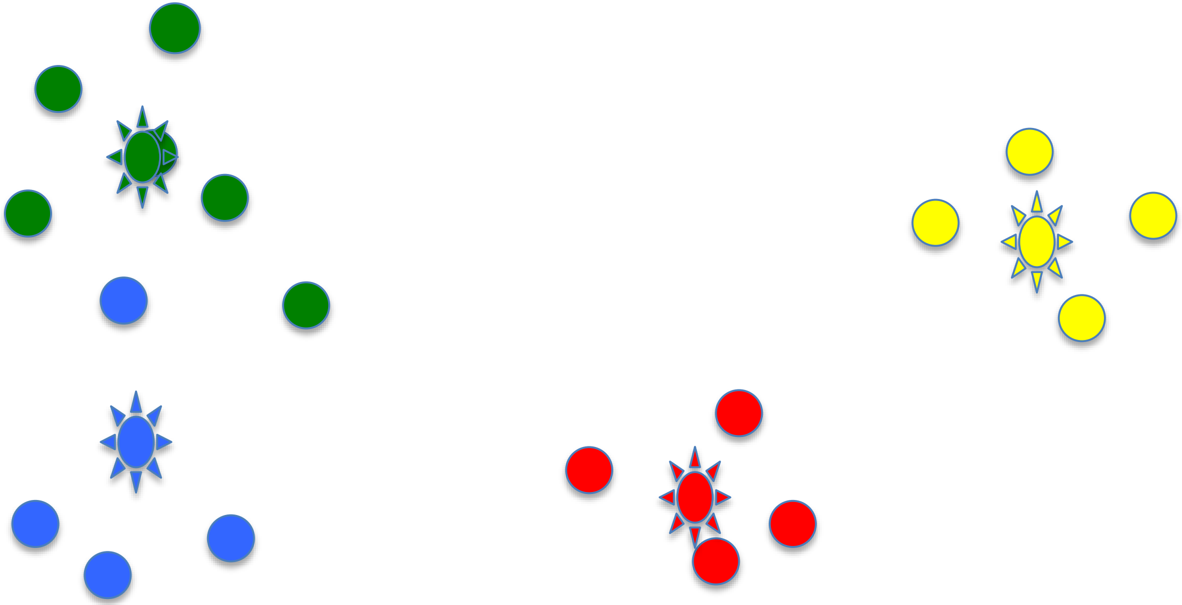
Reassign Points to Clusters



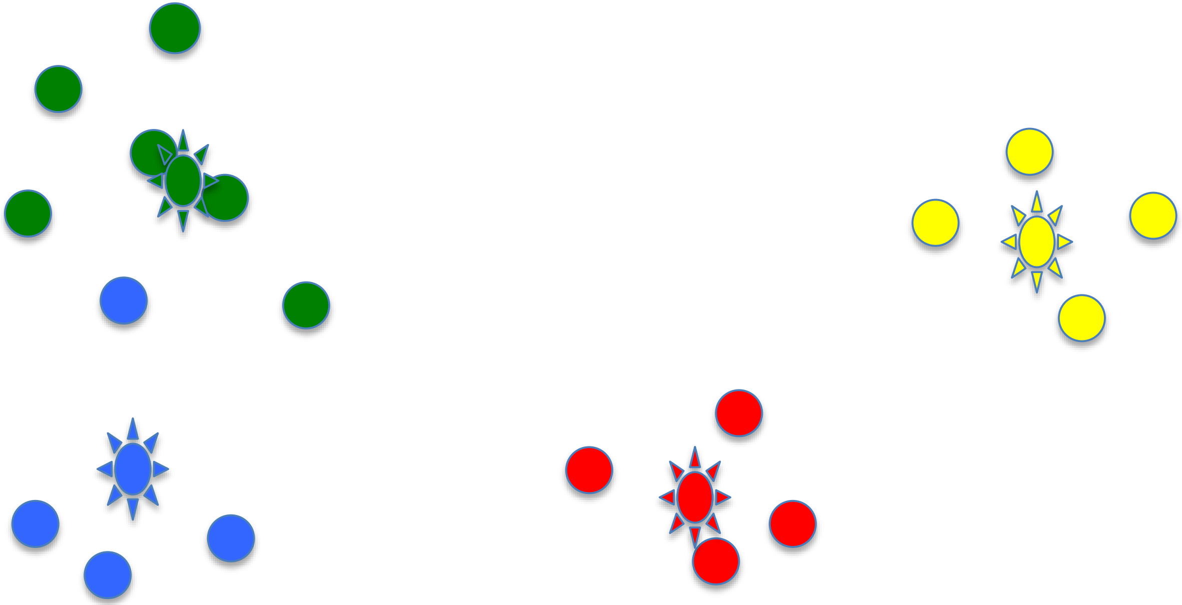
Compute New Centroids



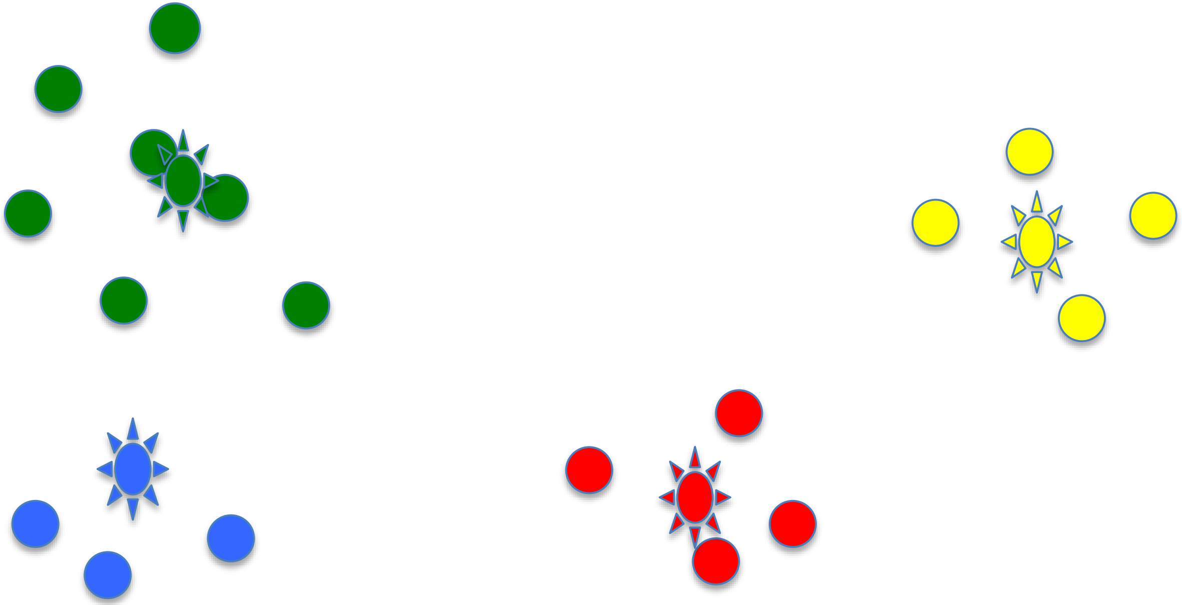
Reassign Points



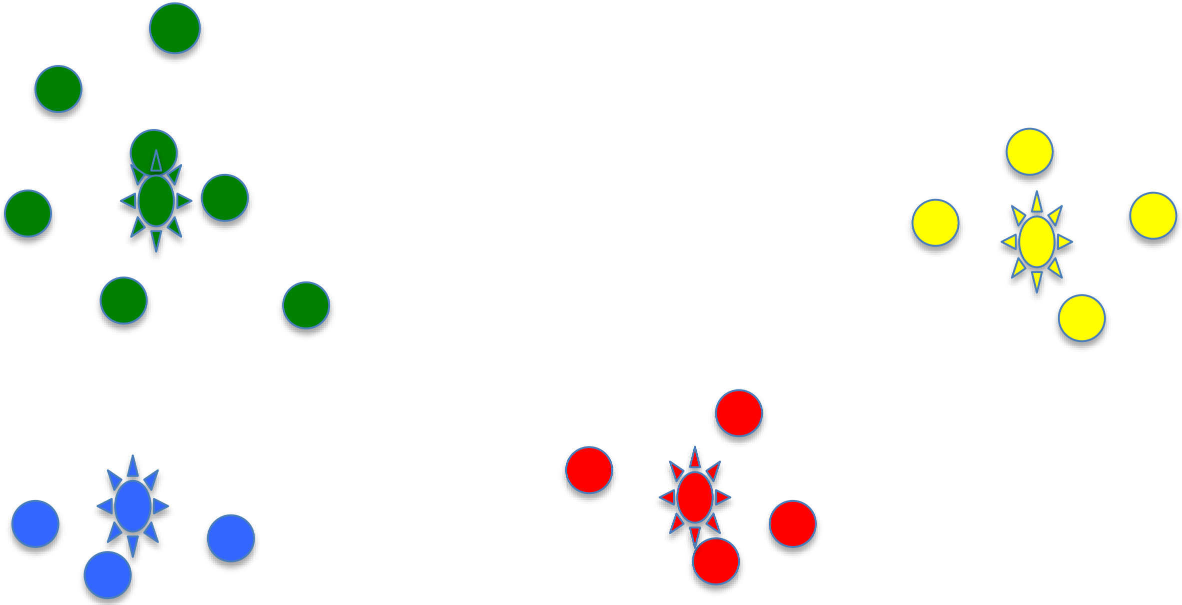
Compute New Centroids



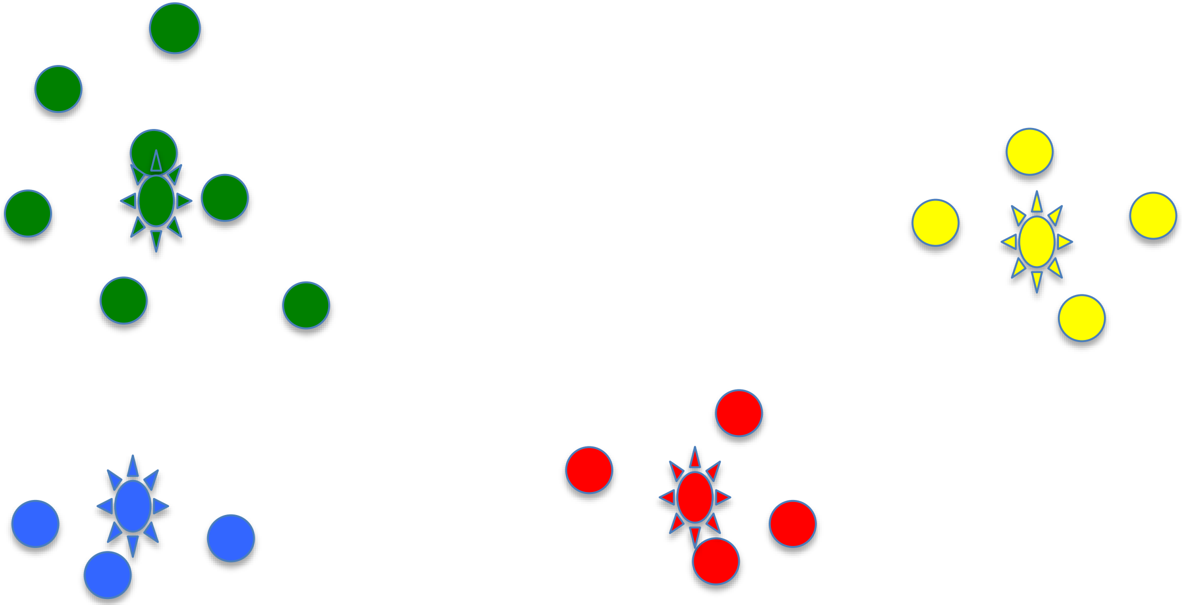
Reassign Points



Compute New Centroids



No Points Move

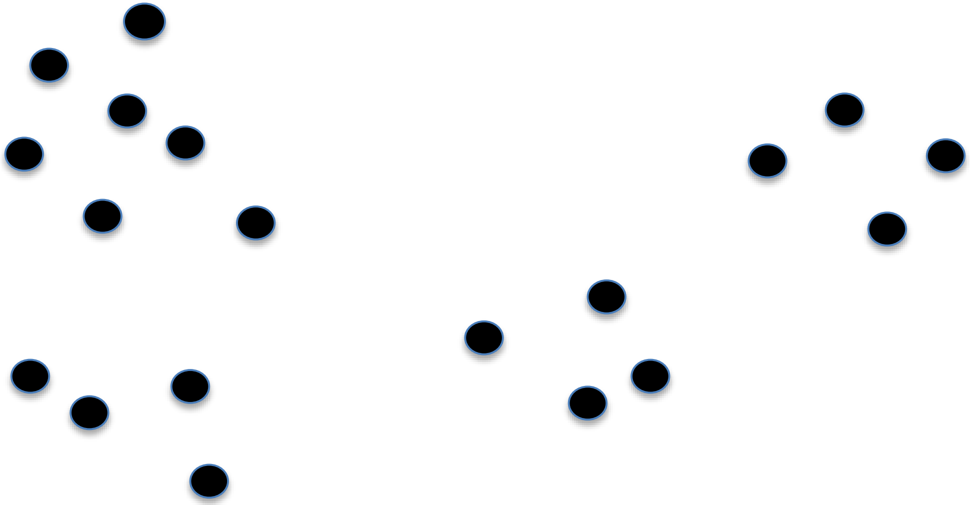


Issues with K-means

Final result can depend upon initial centroids

- **Greedy algorithm can find different local optima**

Choosing the "wrong" k can lead to nonsense



Choosing K

A priori knowledge about application domain

There are five different kinds of bacteria: k = 5

There are two kinds of people in the world: k = 2

Search for a good k

Try different values of k, and evaluate quality of results

Choosing Centroids

Try multiple random choices and choose best

Finding the “Best” Solution

```
best = kMeans(points)
for t in range(numTrials):
    C = kMeans(points)
    if badness(C) < badness(best):
        best = C
```

$$V(c) = \sum_{x \in c} (mean(c) - x)^2 \qquad badness(C) = \sum_{c \in C} V(c)$$

Hierarchical vs. K-means

Hierarchical looks at different numbers of clusters

From 1 to n

K-means looks at many ways of creating k clusters

Hierarchical is slow

K-means is fast

Hierarchical is deterministic

K-means is non-deterministic

CC-BY Image Courtesy of MrGuilt