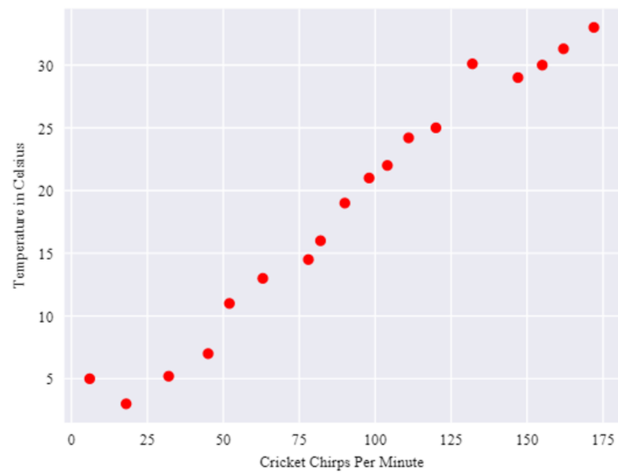
Introduction to Cluster Analysis

Lecture 09

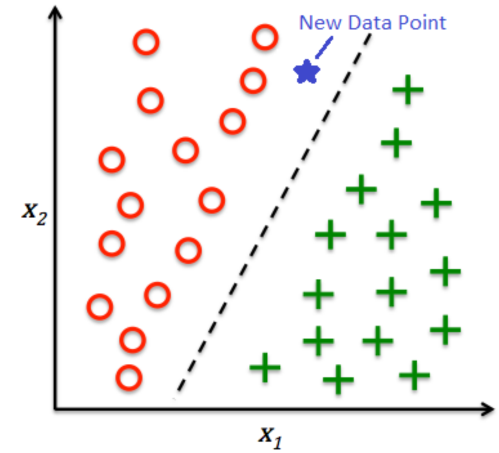
by Marina Barsky

Types of learning tasks

Supervised learning

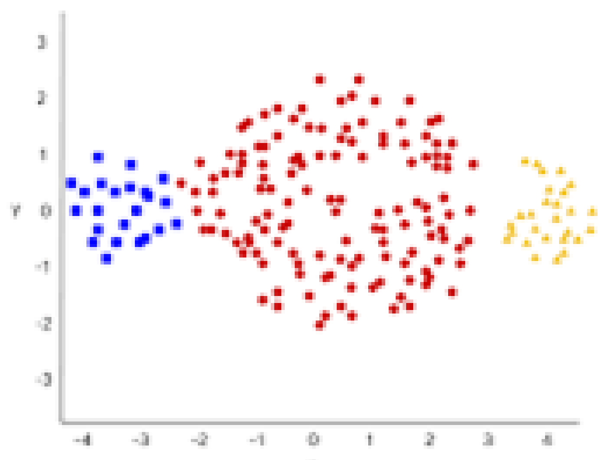


Prediction



Classification

Unsupervised learning



Clustering

TransactionId	Items
1	{A,C,D}
2	{B,C,D}
3	{A,B,C,D}
4	{B,D}
5	{A,B,C,D}

Associations

What is Cluster Analysis?

Finding groups of objects such that the objects in each group are similar (or related) to one another and different from (or unrelated to) the objects in other groups

Labeling objects with group label

- Humans are skilled at dividing objects into groups (clustering) and assigning new objects to one of the groups (classification)

- *Classes* – conceptually meaningful groups of objects that share common characteristics

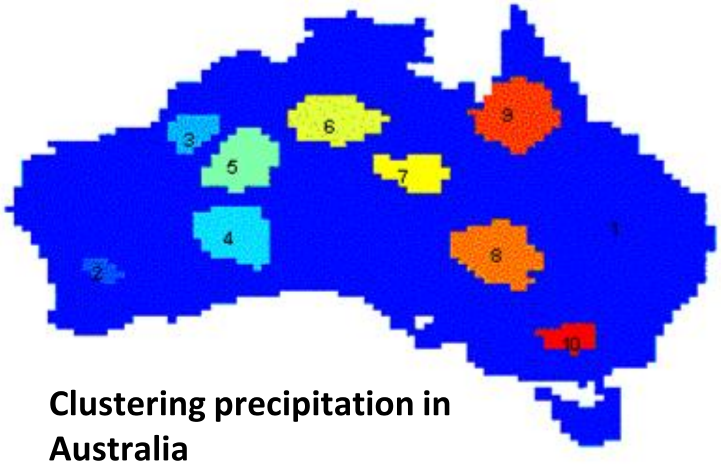
- Clusters are *potential classes*, and cluster analysis it a technique for automatically discovering classes from unlabeled data

Motivation for Cluster Analysis

- **Clustering for Understanding**
  - Group related documents for browsing
  - Group genes and proteins that have similar functionality
  - Group stocks with similar price fluctuations
  - Segment customers into a small number of groups for additional analysis and marketing activities.

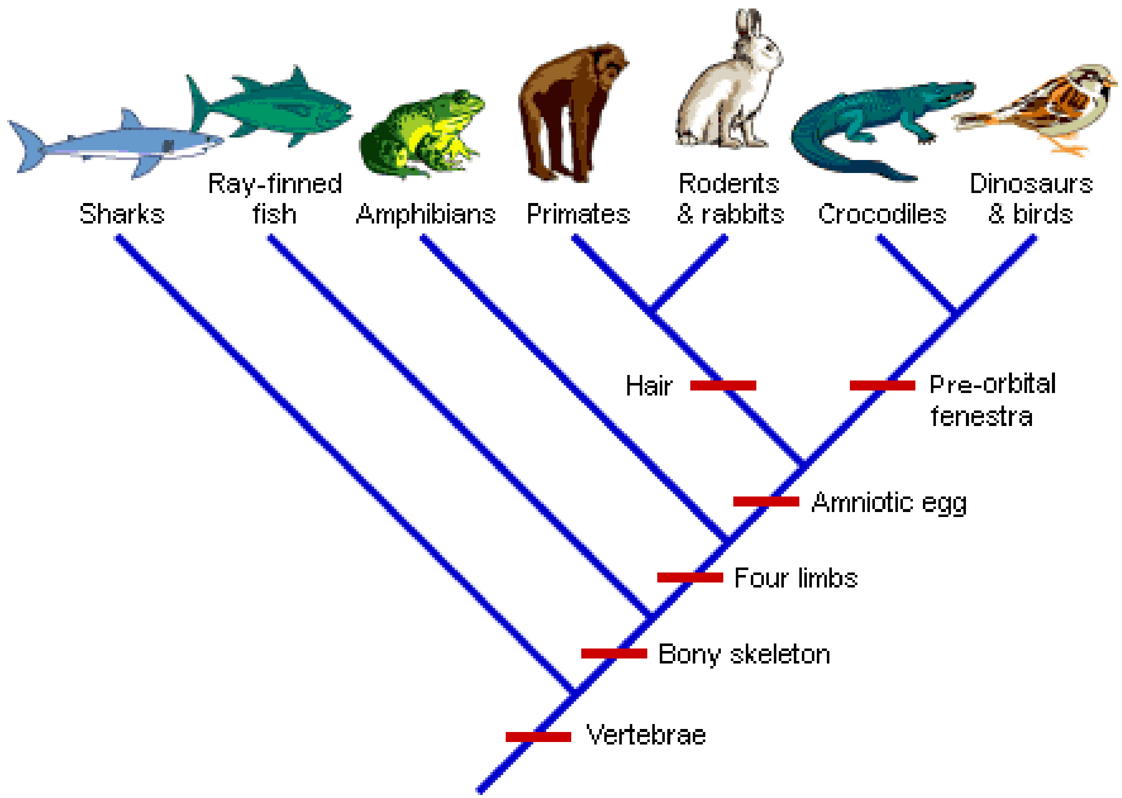
- **Clustering for Summarization**
  - Reduce the size of large data sets

	Discovered Clusters	*Industry Group*
1	Applied-Matl-DOWN,Bay-Network-Down,3-COM-DOWN, Cabletron-Sys-DOWN,CISCO-DOWN,HP-DOWN, DSC-Comm-DOWN,INTEL-DOWN,LSI-Logic-DOWN, Micron-Tech-DOWN,Texas-Inst-Down,Tellabs-Inc-Down, Natl-Semiconduct-DOWN,Oracl-DOWN,SGI-DOWN, Sun-DOWN	Technology1-DOWN
2	Apple-Comp-DOWN,Autodesk-DOWN,DEC-DOWN, ADV-Micro-Device-DOWN,Andrew-Corp-DOWN, Computer-Assoc-DOWN,Circuit-City-DOWN, Compaq-DOWN, EMC-Corp-DOWN, Gen-Inst-DOWN, Motorola-DOWN,Microsoft-DOWN,Scientific-Atl-DOWN	Technology2-DOWN
3	Fannie-Mae-DOWN,Fed-Home-Loan-DOWN, MBNA-Corp-DOWN,Morgan-Stanley-DOWN	Financial-DOWN
4	Baker-Hughes-UP,Dresser-Inds-UP,Halliburton-HLD-UP, Louisiana-Land-UP,Phillips-Petro-UP,Unocal-UP, Schlumberger-UP	Oil-UP



Clustering precipitation in Australia

Grouping animals into clusters: Biological Systematics



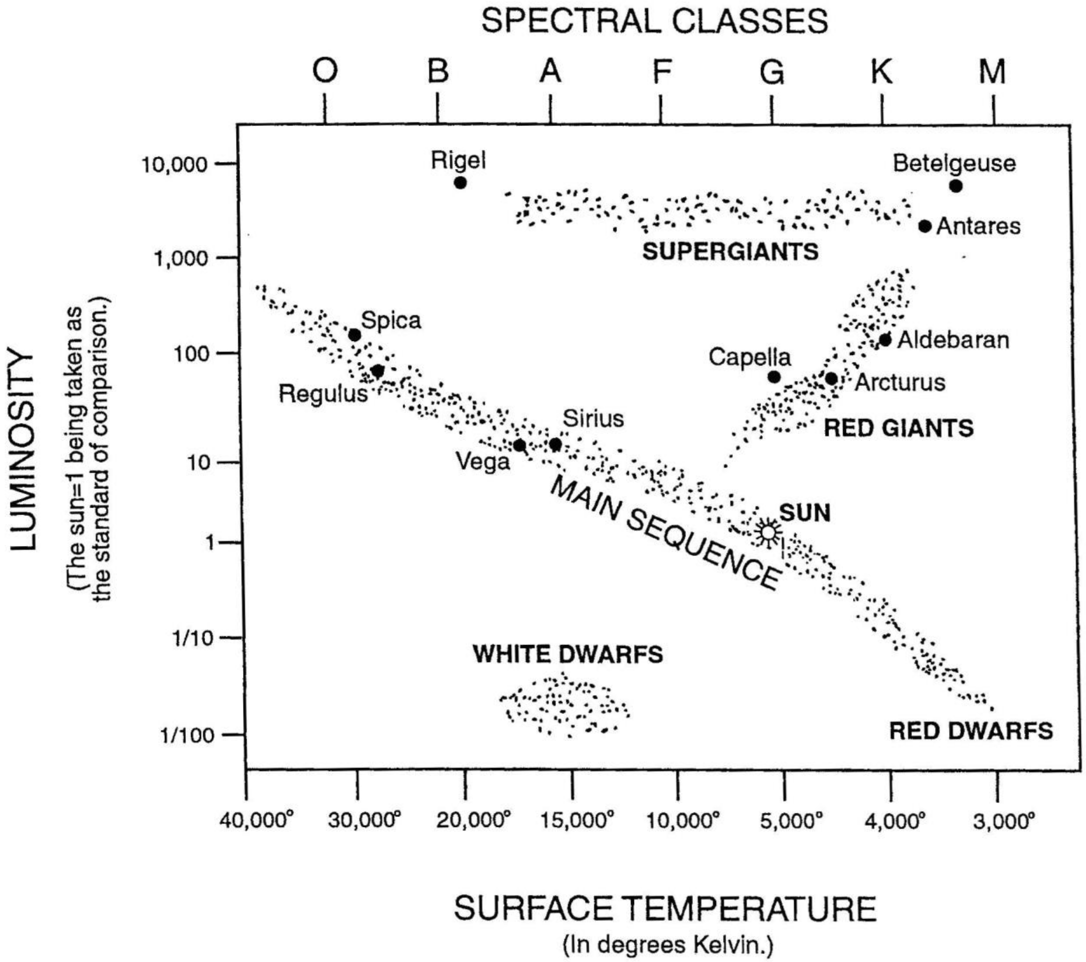
We cluster animals into hierarchical groups to better understand evolution

Grouping documents into clusters: information retrieval

- Grouping WEB query results into small number of clusters, each capturing a particular aspect of a query

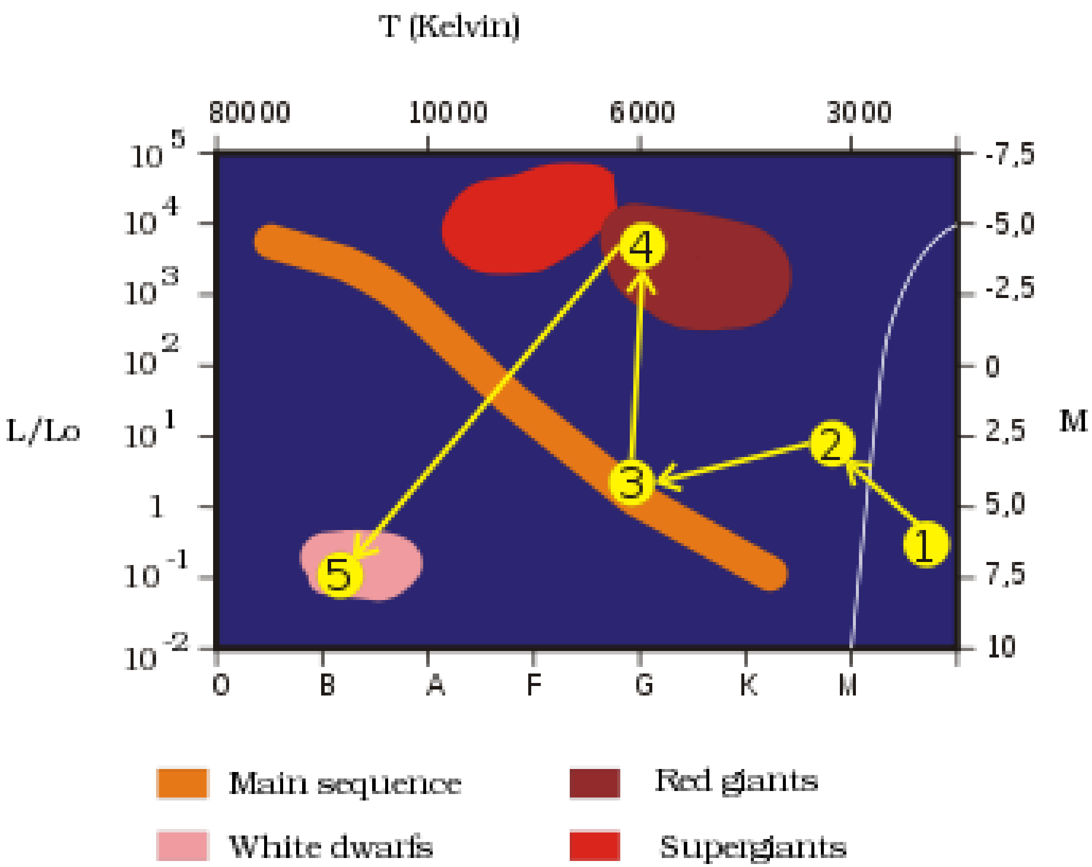
Search for *tiger*

Discovery:
Galaxies in 2 dimensions



The Hertzsprung-Russel diagram clusters stars by temperature and luminosity

Clustering → discoveries:



Galaxies evolution

Main sequence stars generate energy by fusing Hydrogen to Helium

When the hydrogen is used up, Helium fusion occurs, the star expands → **red giant**

The outer layer of gases is stripped away, the star cools down → **white dwarf**

Machine Learning task: automated clustering

- Discovering groups (classes) of objects from unlabeled data
- *Unsupervised learning*

Formalizing the task

We need to convey to a machine:

1. What do we mean when we say that two objects are similar (dissimilar): preferably *define similarity as a numeric value*

2. What to look for: *define a notion of a cluster*

3. Prescribe a **precise algorithm** for finding these clusters

1. DEFINE SIMILARITY/DISTANCE

Numeric *proximity* (similarity or distance) between two data points

- Each attribute is a separate and independent dimension of the data

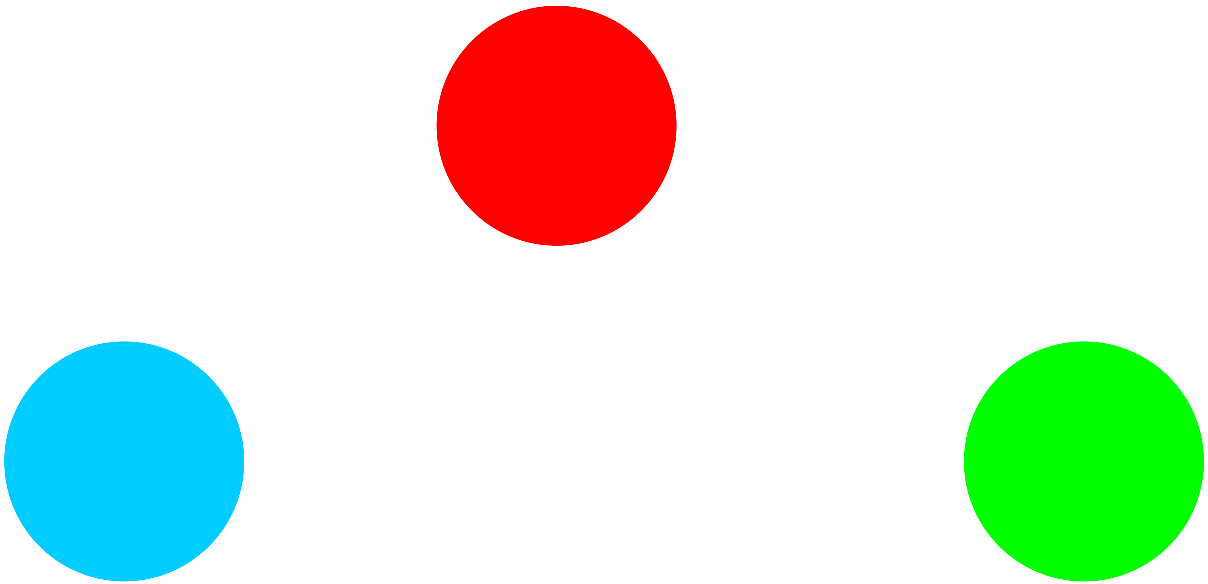
- Compute distance across each dimension and combine to an overall distance between objects

See K-NN lecture 06!

2. THE DEFINITION OF A CLUSTER

Types of Clusters 1/4: Well-Separated

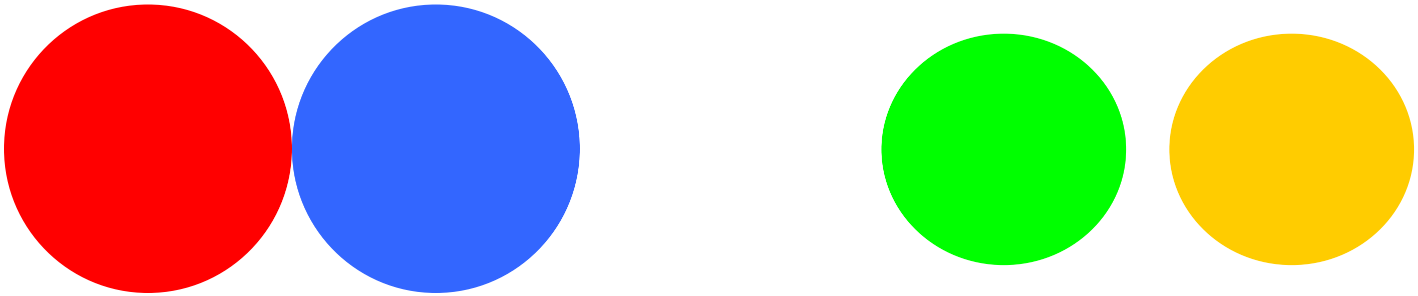
- Any point in a cluster is closer (or more similar) to every other point in the cluster than to any point not in the cluster.



3 well-separated clusters

Types of Clusters 2/4: Center-Based

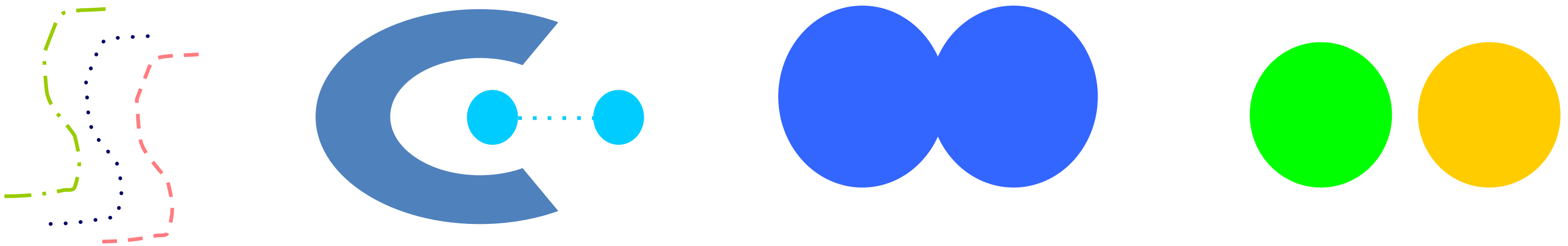
- An object in a cluster is closer (more similar) to the “center” of a cluster, than to the center of any other cluster
- The center of a cluster can be a centroid (the average of all the points in the cluster) or a medoid (the most “representative” point of a cluster)



4 center-based clusters

Types of Clusters 3/4: Contiguity-Based

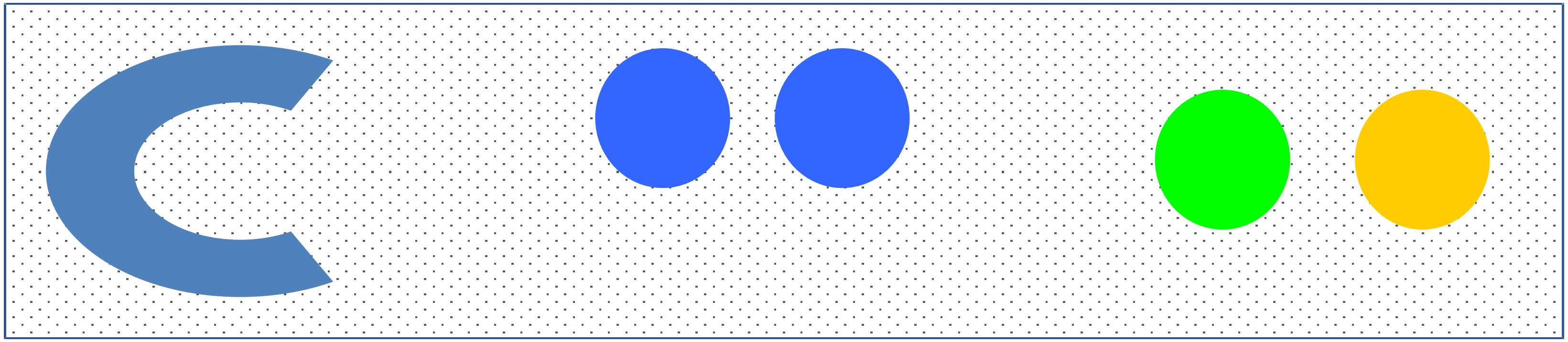
- Contiguous Cluster (Nearest-neighbor or Transitive)
- A point in a cluster is closer to at least one point in the cluster than to any point not in the cluster. The group of objects that are connected to one another.



8 contiguous clusters

Types of Clusters 4/4: Density-Based

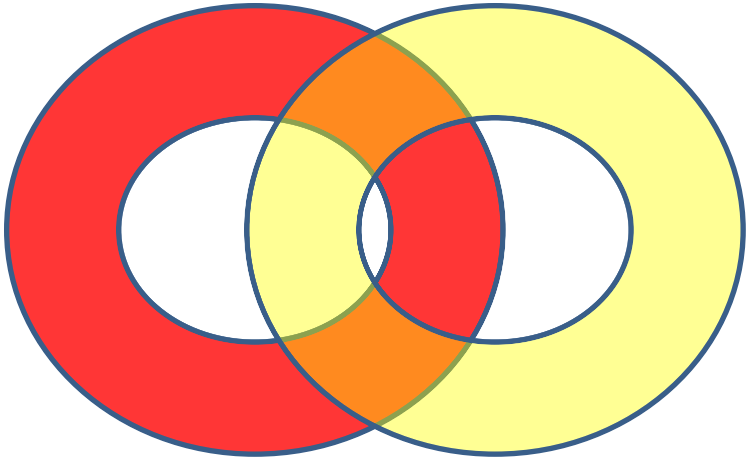
- A cluster is a dense region of points, which is separated by low-density regions, from other regions of high density.
- Used when the clusters are irregular or intertwined, and when noise and outliers are present.



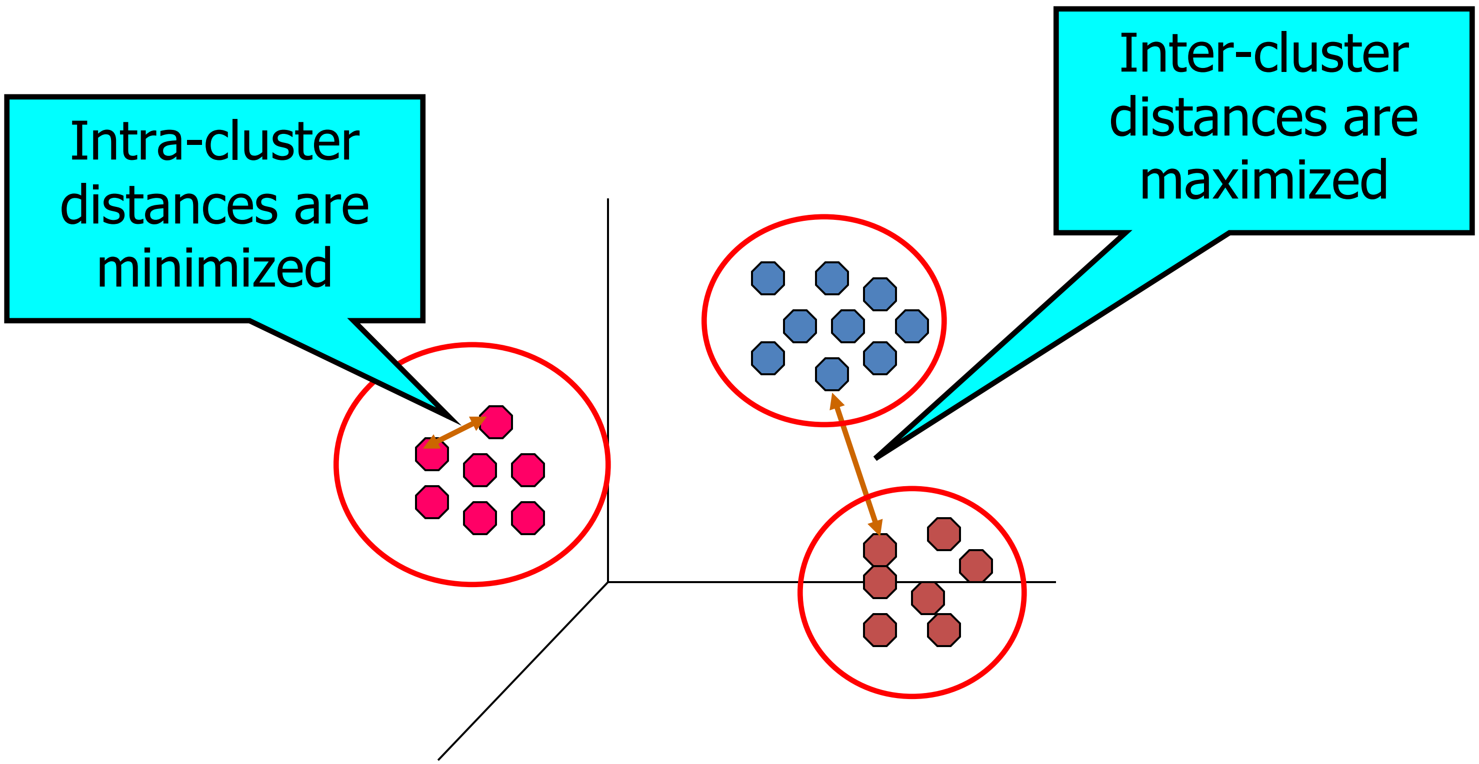
4 density-based clusters

General definition: +conceptual clusters

- *Cluster* is a set of objects that share some property. This includes all previous cluster types
- In addition it includes clusters defined by a *concept.* Such clusters are used in pattern recognition. To discover such clusters automatically, the concept should be defined first.



Clustering algorithm: goal



Clustering algorithms

- *K*-means clustering
- Agglomerative hierarchical clustering
- Density-based clustering

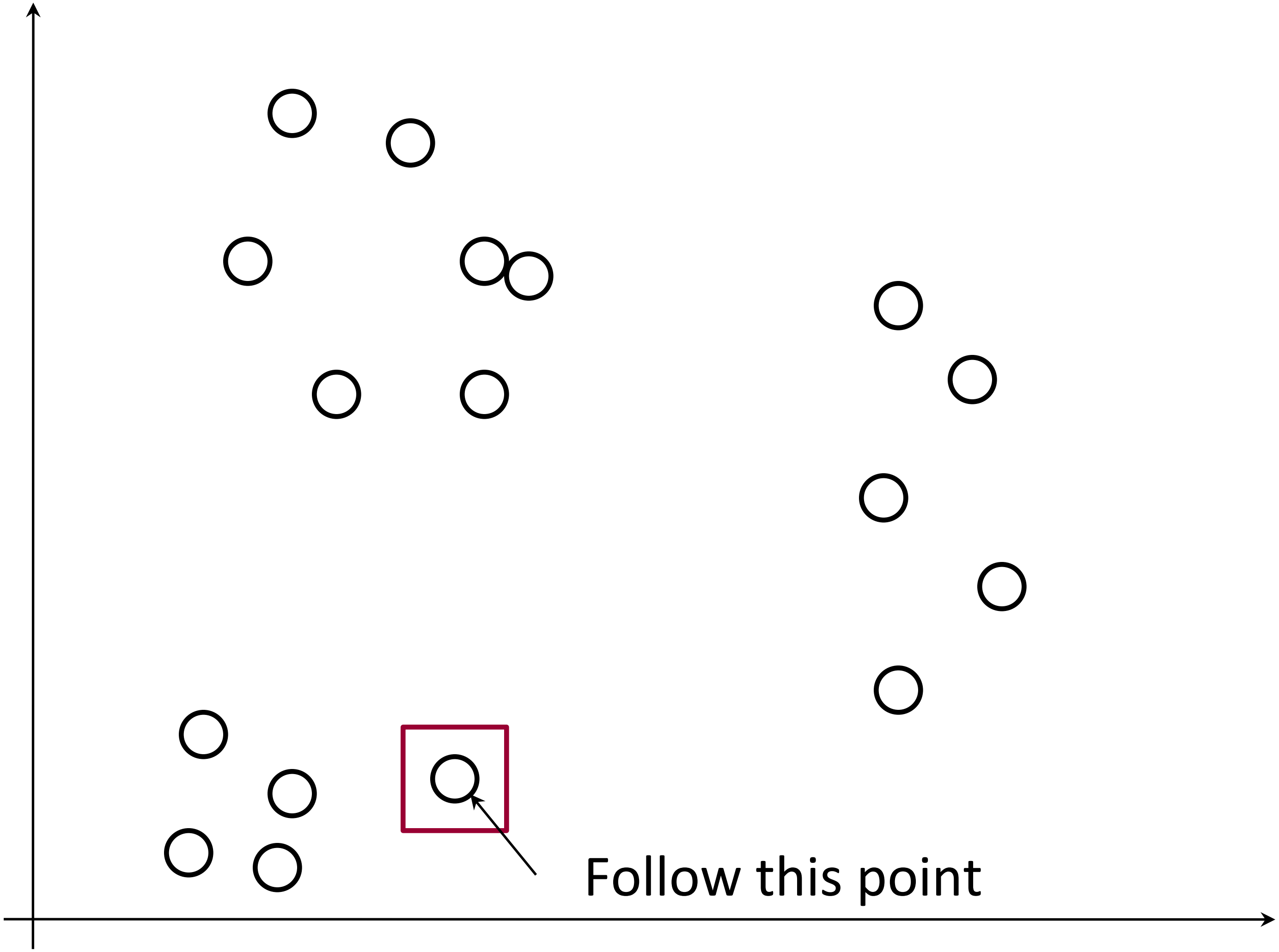
Iterative solution: *K*-means clustering algorithm

Select *K* random seeds
Do
    Assign each record to the closest seed
    Calculate centroid of each cluster
        (take average value for each dimension
            of all records in the cluster)
    Set these centroids as new seeds
Until coordinates of seeds *do not change*

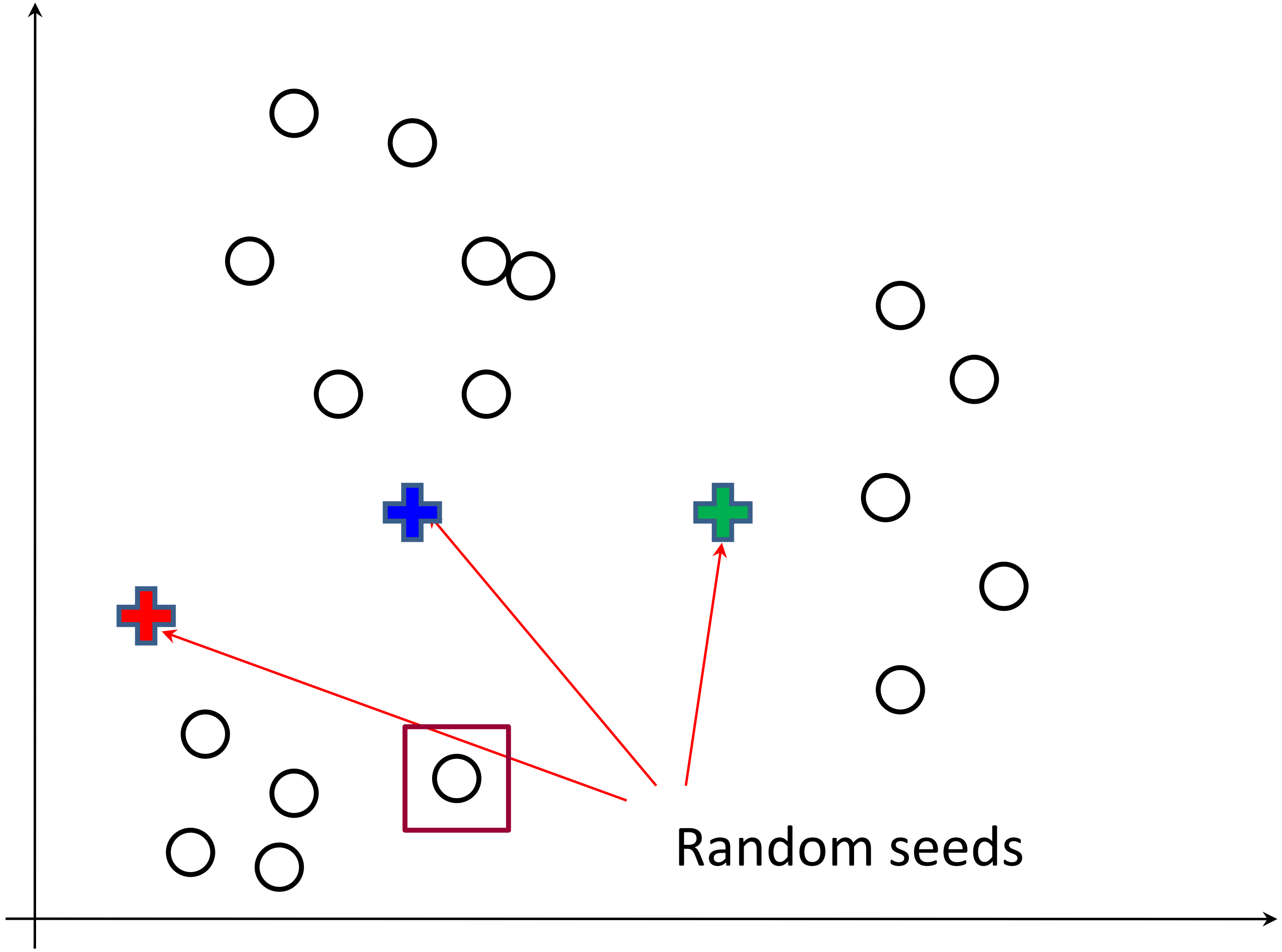
This algorithm in each iteration makes assignment of points such that intra-cluster distances are decreasing.

Local optimization technique – moves into the direction of local minimum, might miss the best solution

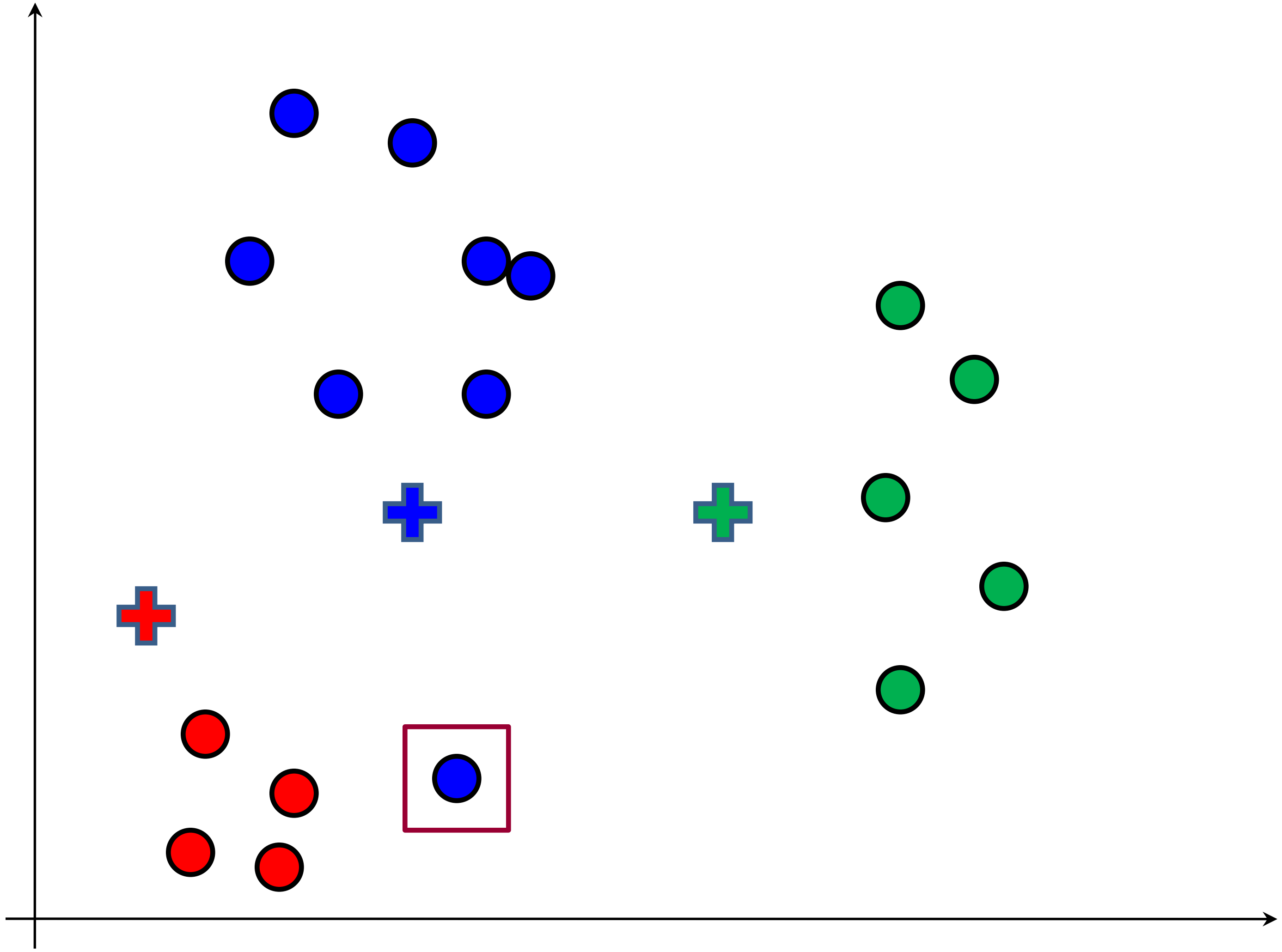
Example 1: K=3



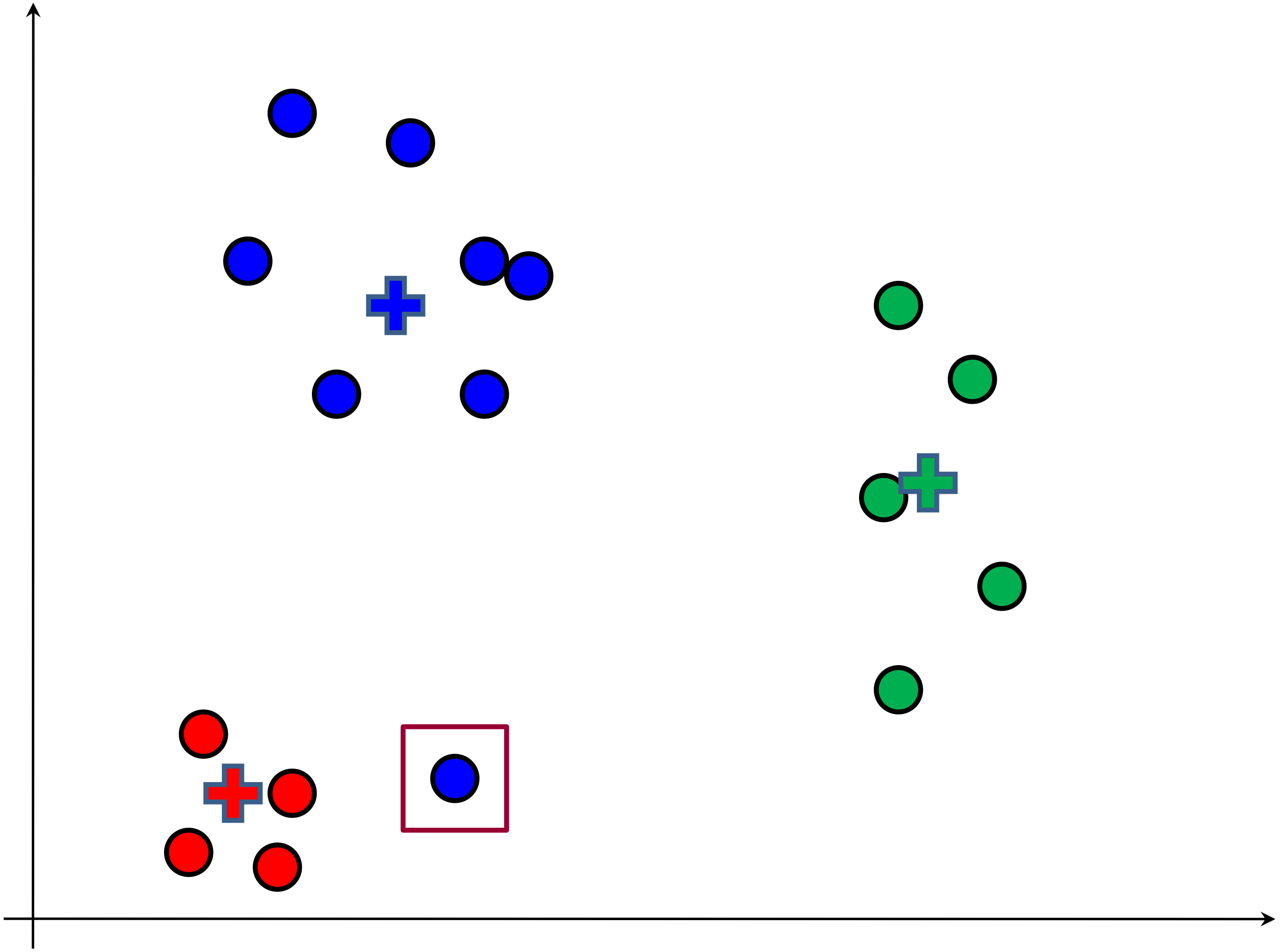
Example 1: initialization



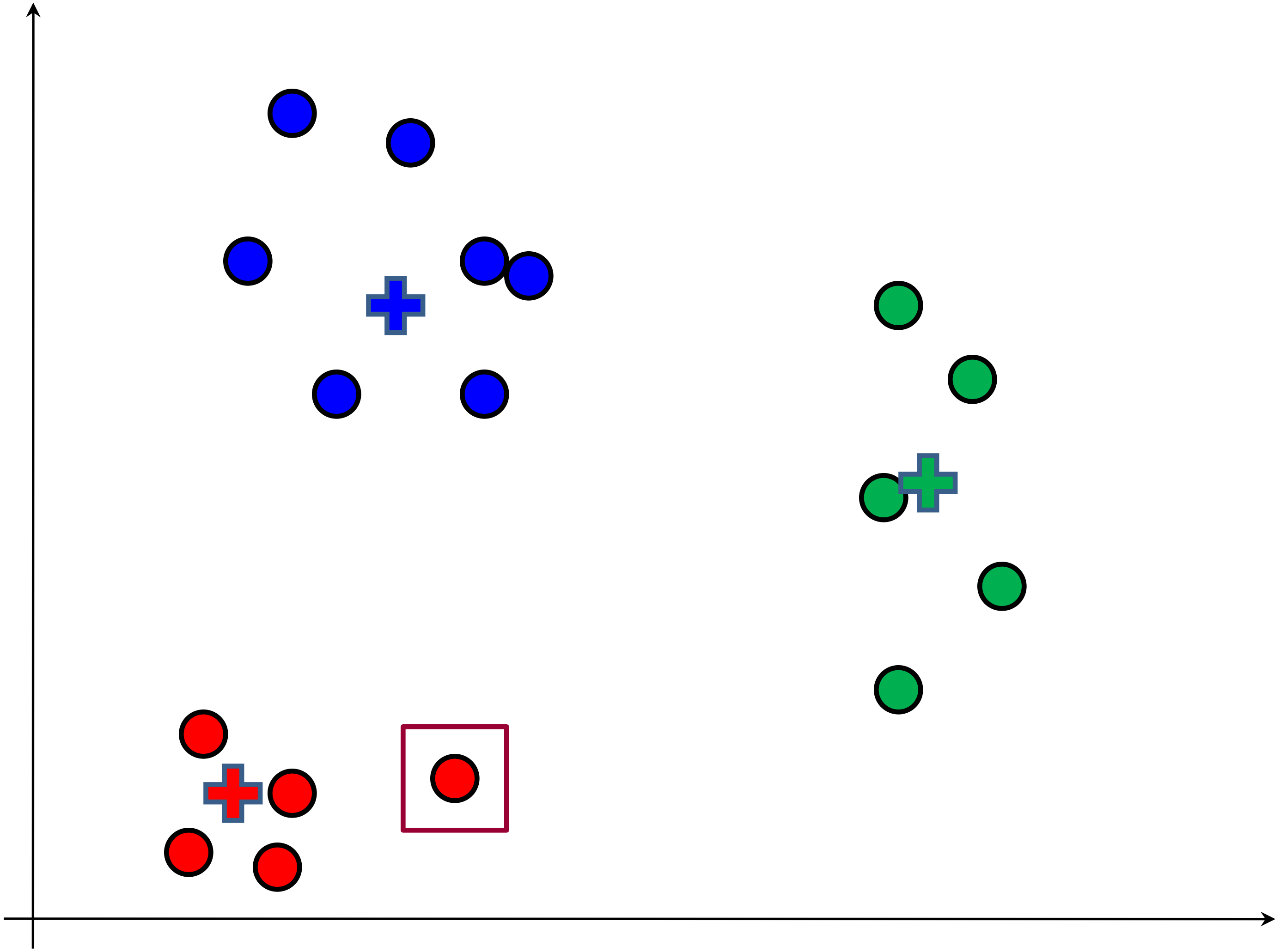
Example 1: iteration 1.
Assign each point to the closest seed



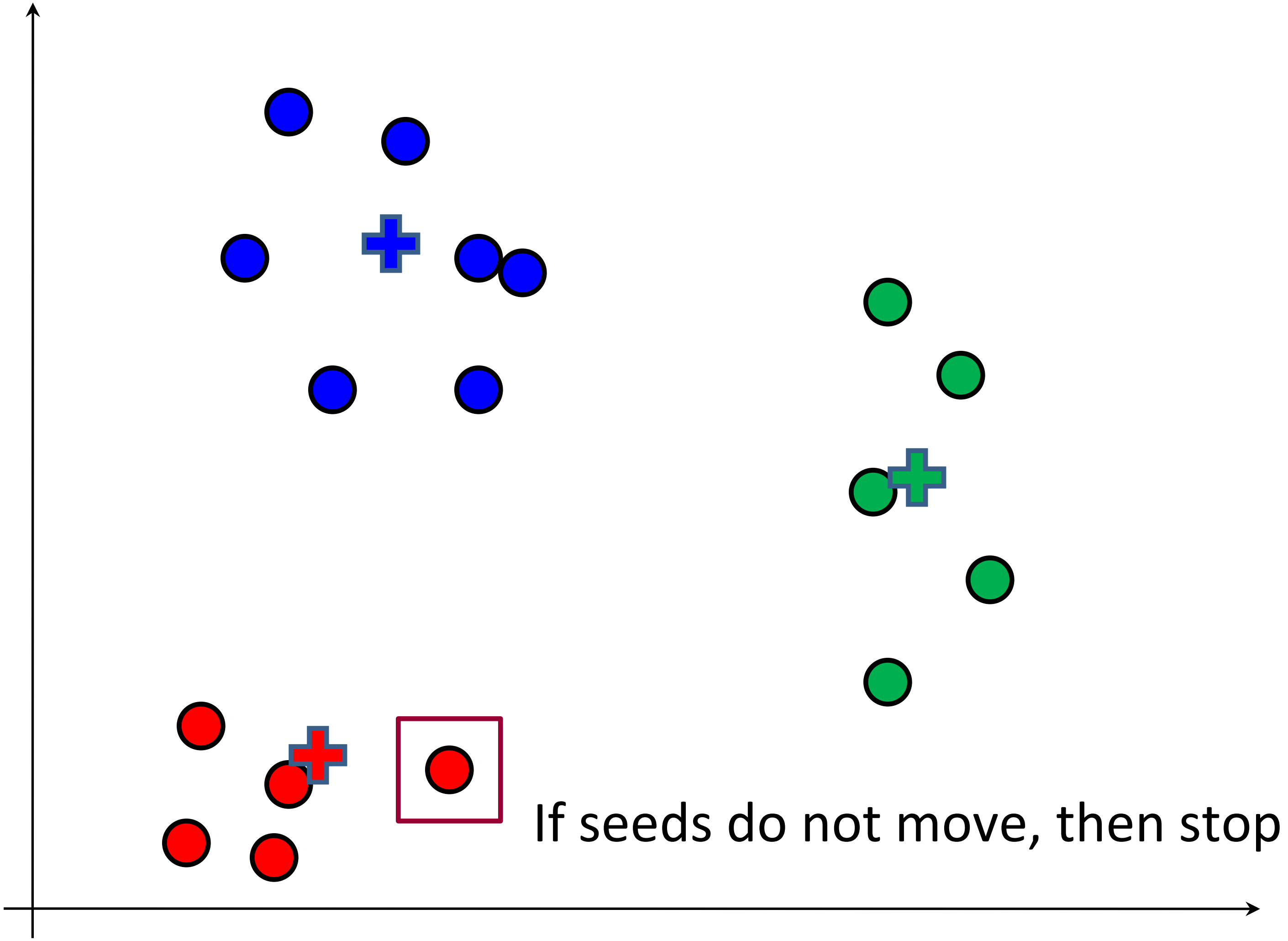
Example 1: iteration 1. Recalculate centroids



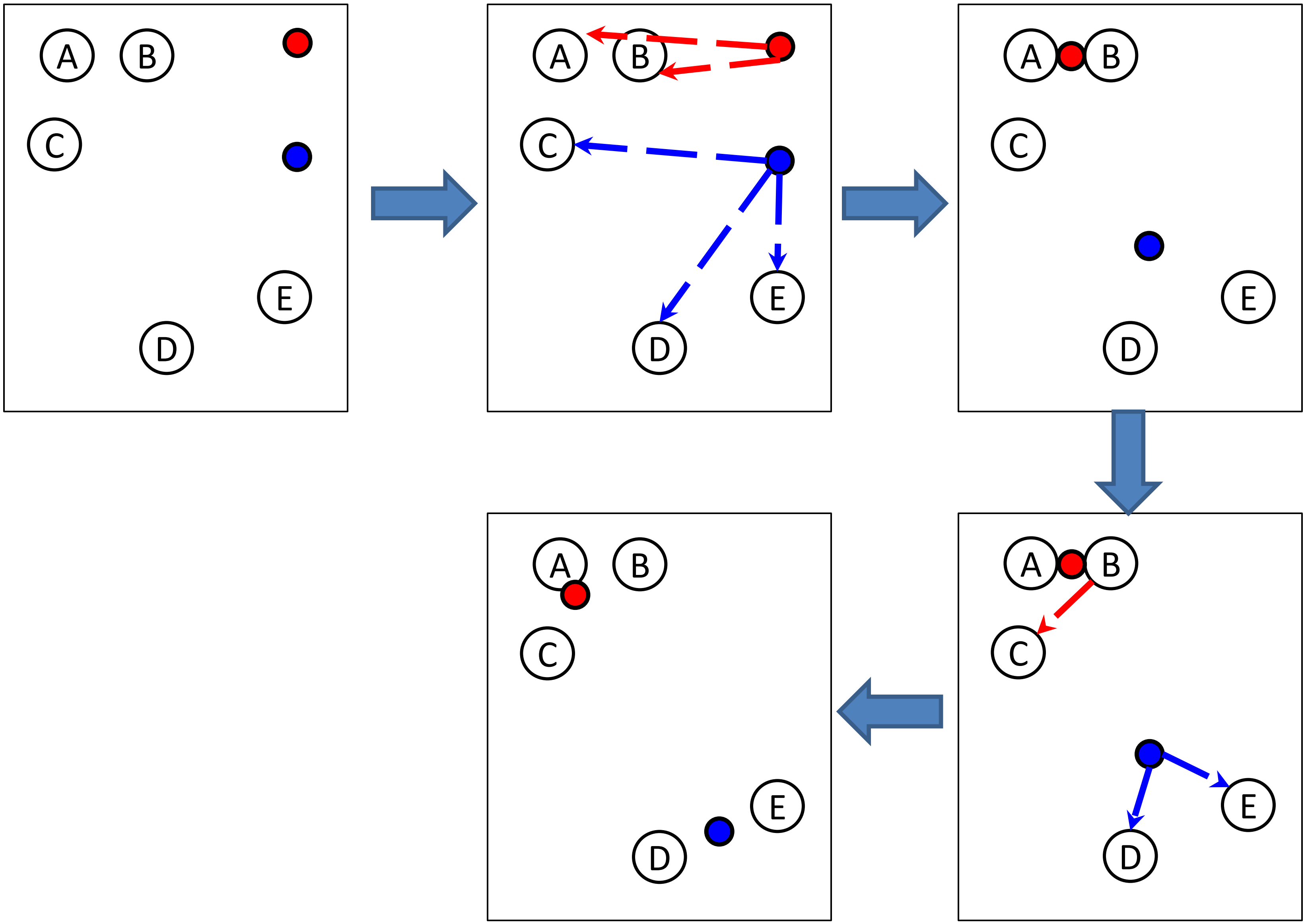
Example 1: iteration 2. Assign each point to the closest seed



Example 1: iteration 2. recalculate centroids – new seeds



Example 2: K=2



Evaluating K-means Clusters

- Most common measure is Sum of Squared Error (SSE)
  - For each point, the error is the distance to the nearest cluster centroid
  - To get SSE, we square these errors and sum them up.

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} [dist(m_i, x)]^2$$

x is a data point in cluster C_i and

m_i is the representative point for cluster C_i (in our case, centroid)

Centroid that minimizes an overall SSE of each cluster is its mean

At each iteration, we decrease total SSE, but with respect to a given set of centroids and point assignments

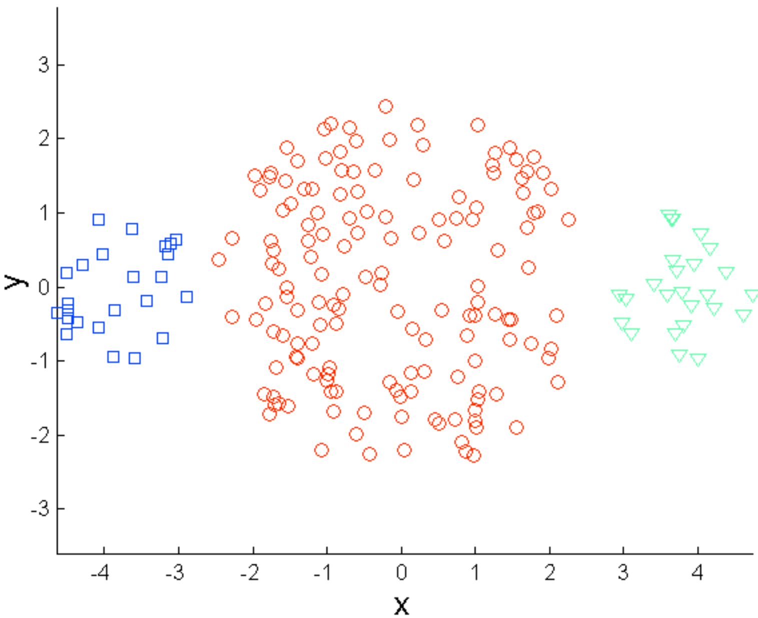
K-means Clustering – Details

- Initial centroids may be chosen randomly.
  - Produced clusters vary from one run to another.
- Most of the convergence happens in the first few iterations.
  - Often the stopping condition is changed to 'Until relatively few points change clusters'
- Complexity is $O(I * K * n * d)$
  - n = number of points, K = number of clusters,
    I = number of iterations, d = number of attributes

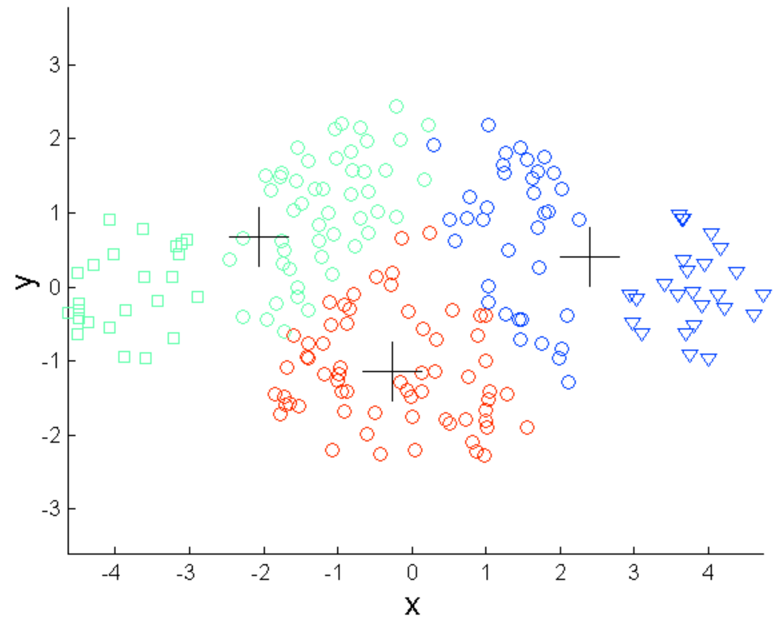
Limitations of K-means

- K-means has problems when clusters are of
  - Differing Sizes
  - Differing Densities
  - Non-globular shapes

Limitations of K-means: Differing Sizes

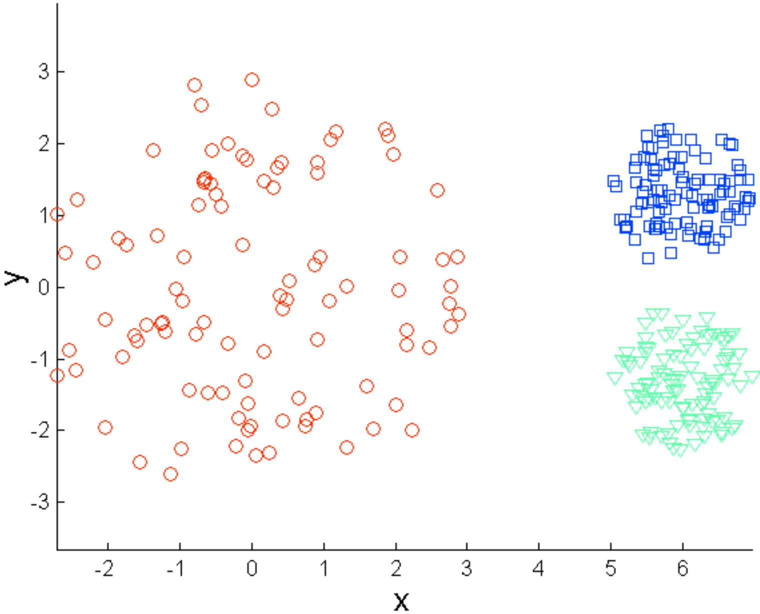


Original Points

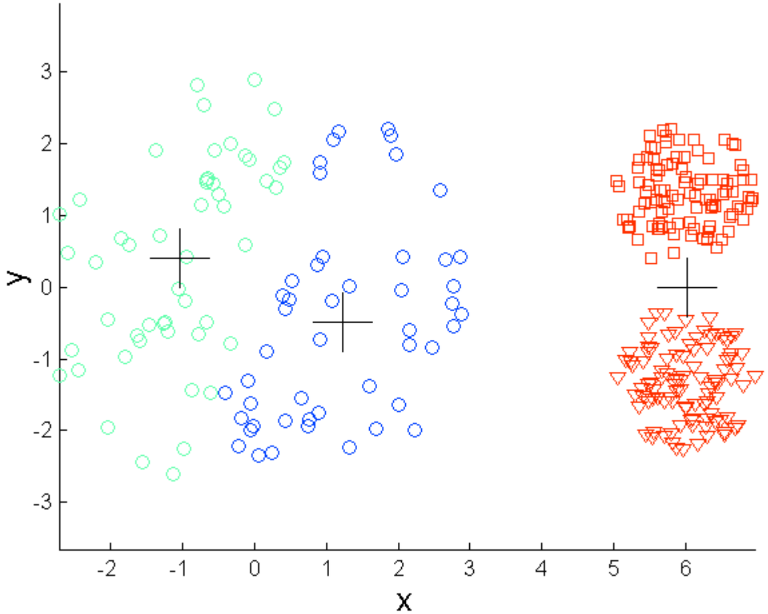


K-means (3 Clusters)

Limitations of K-means: Differing Density

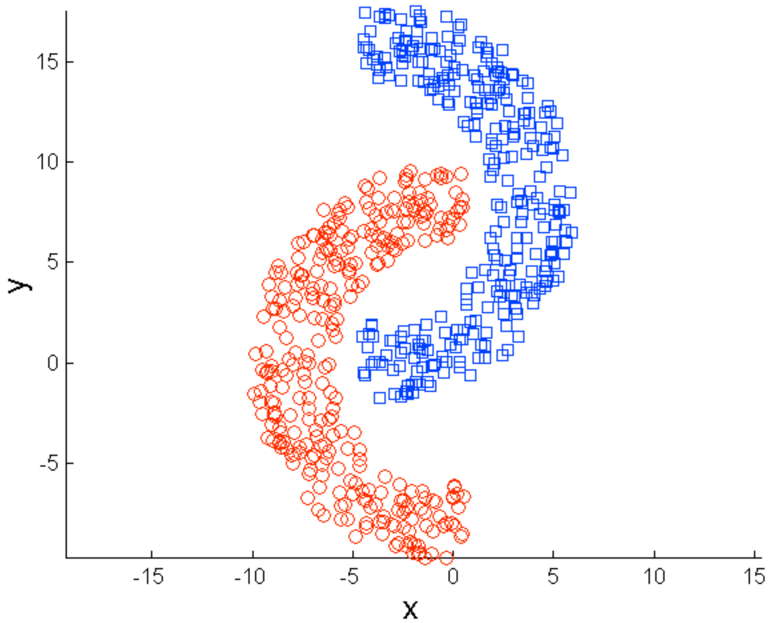


Original Points

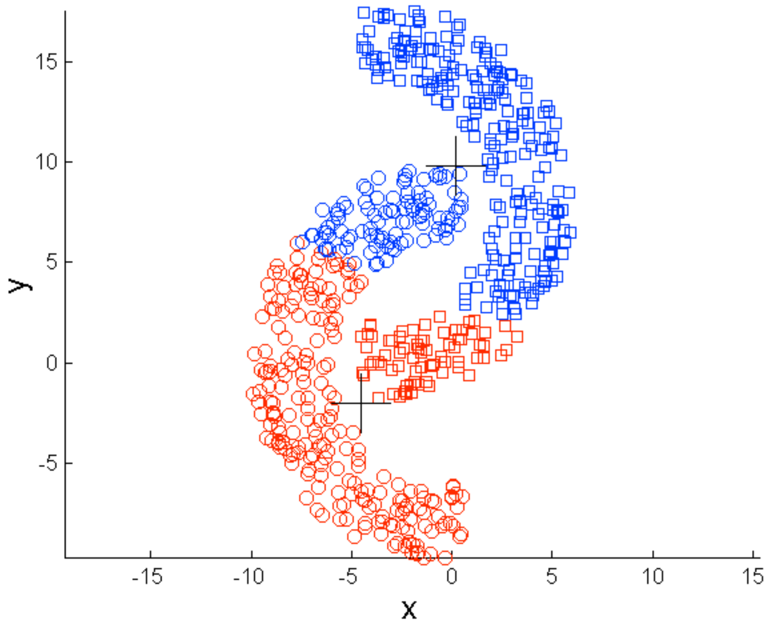


K-means (3 Clusters)

Limitations of K-means: Non-globular Shapes



Original Points



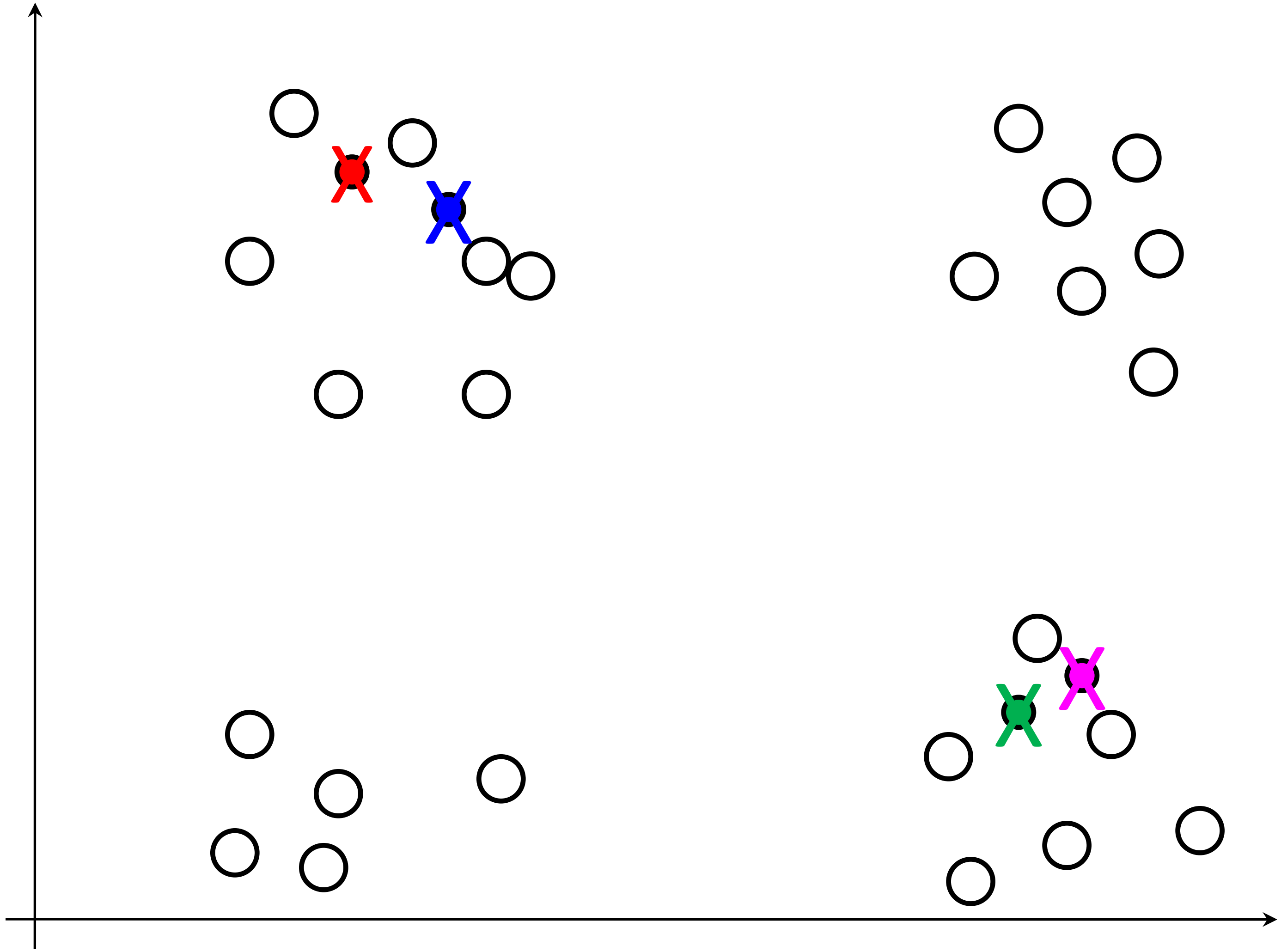
K-means (2 Clusters)

Limitations of K-means

- K-means has problems when clusters are of
  - Differing Sizes
  - Differing Densities
  - Non-globular shapes

- But even for globular clusters, the choice of initial centroids influences the quality of clustering

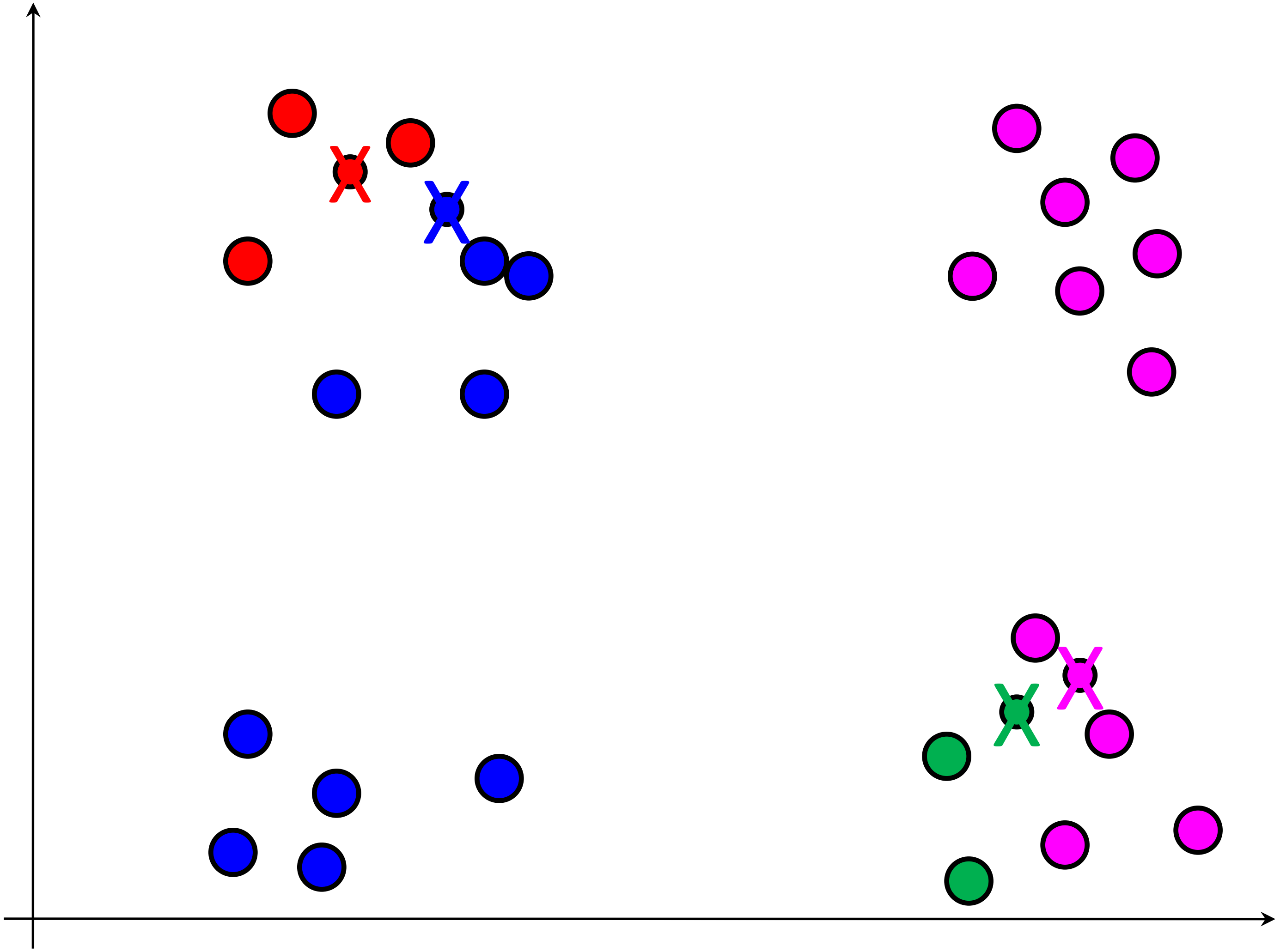
1. Importance of choosing initial centroids: K=4



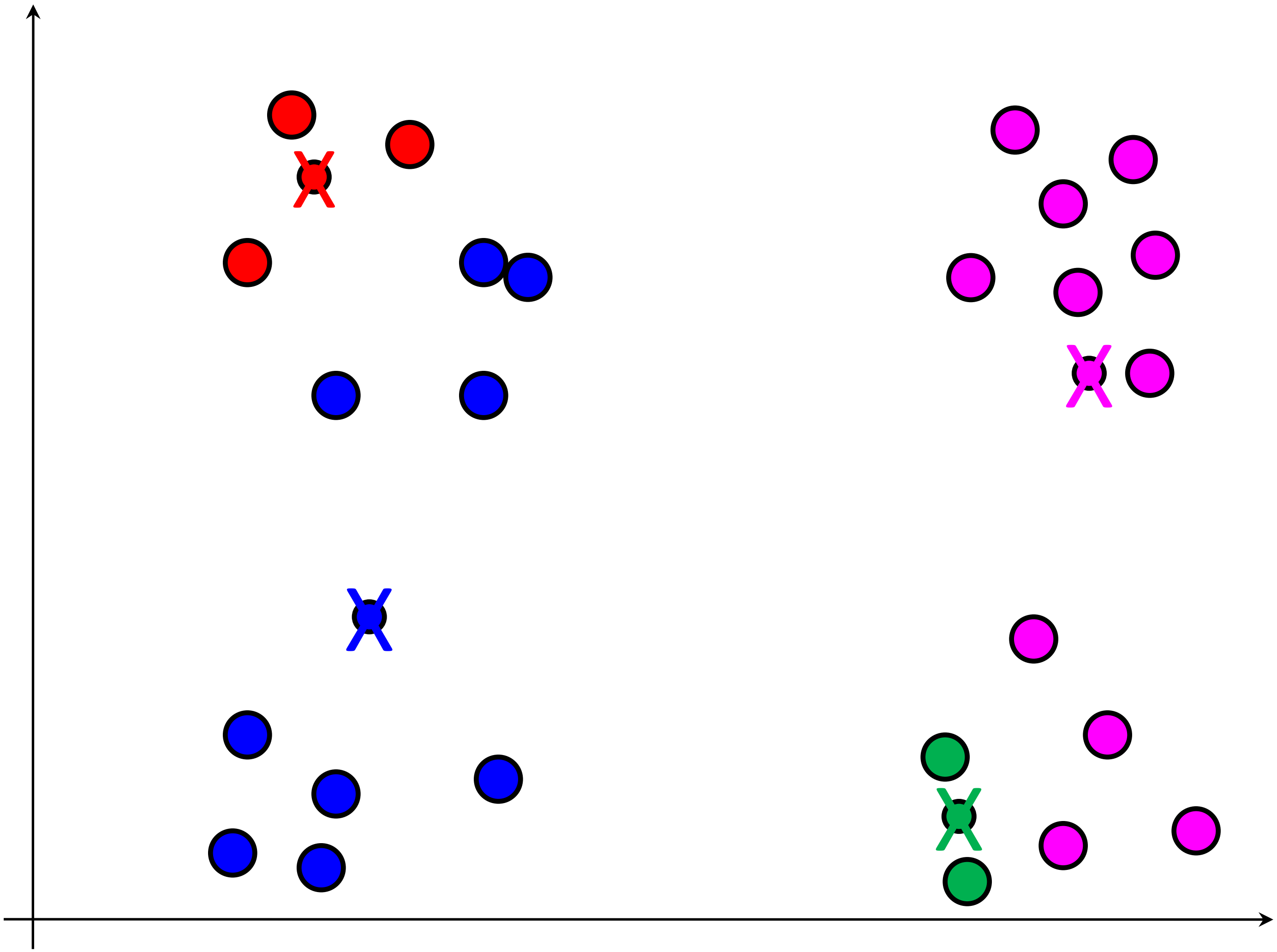
2 pairs of clusters

Initial seeds are chosen randomly: 2 seeds per each pair

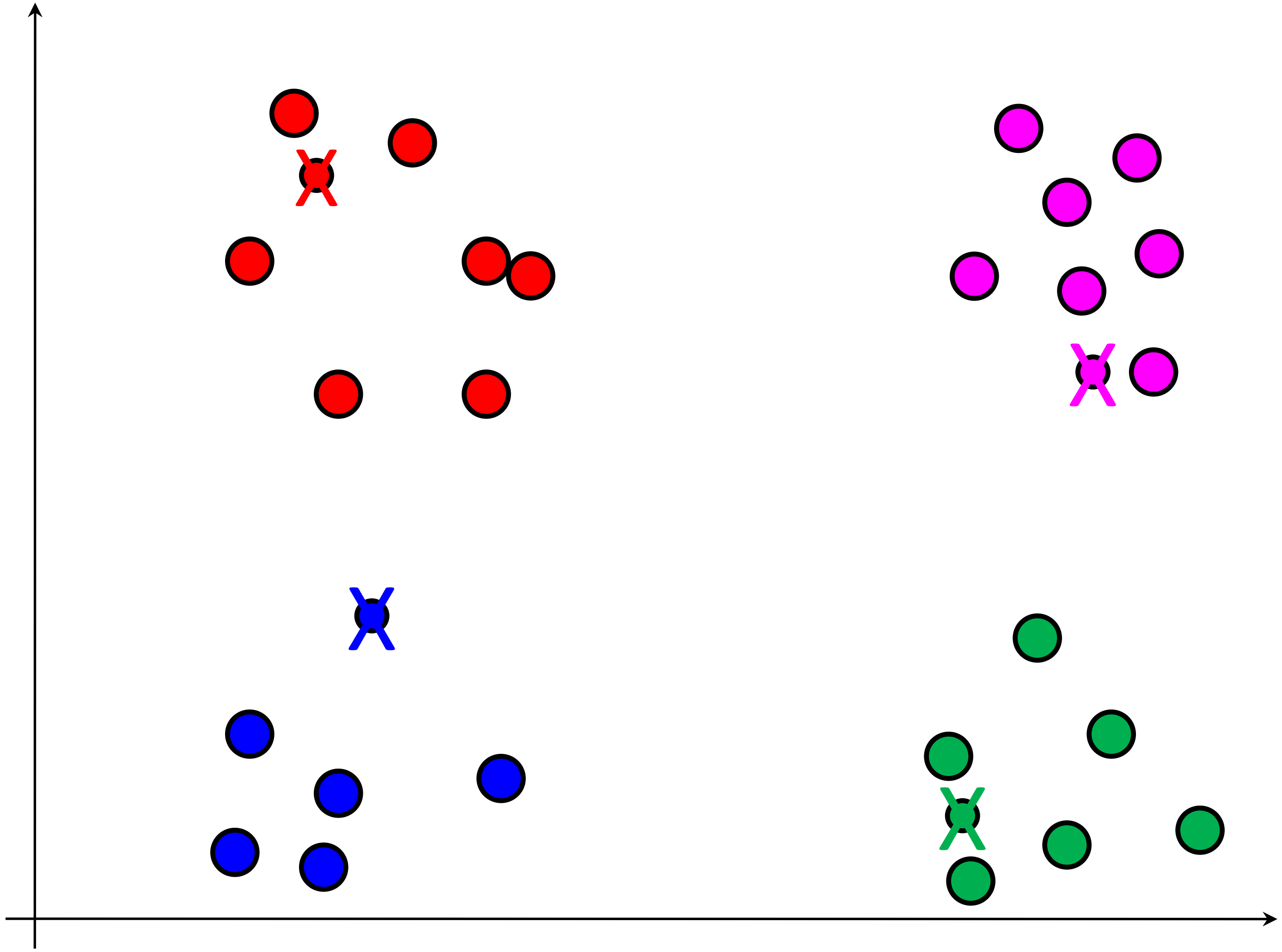
1. Importance of choosing initial centroids: point assignments



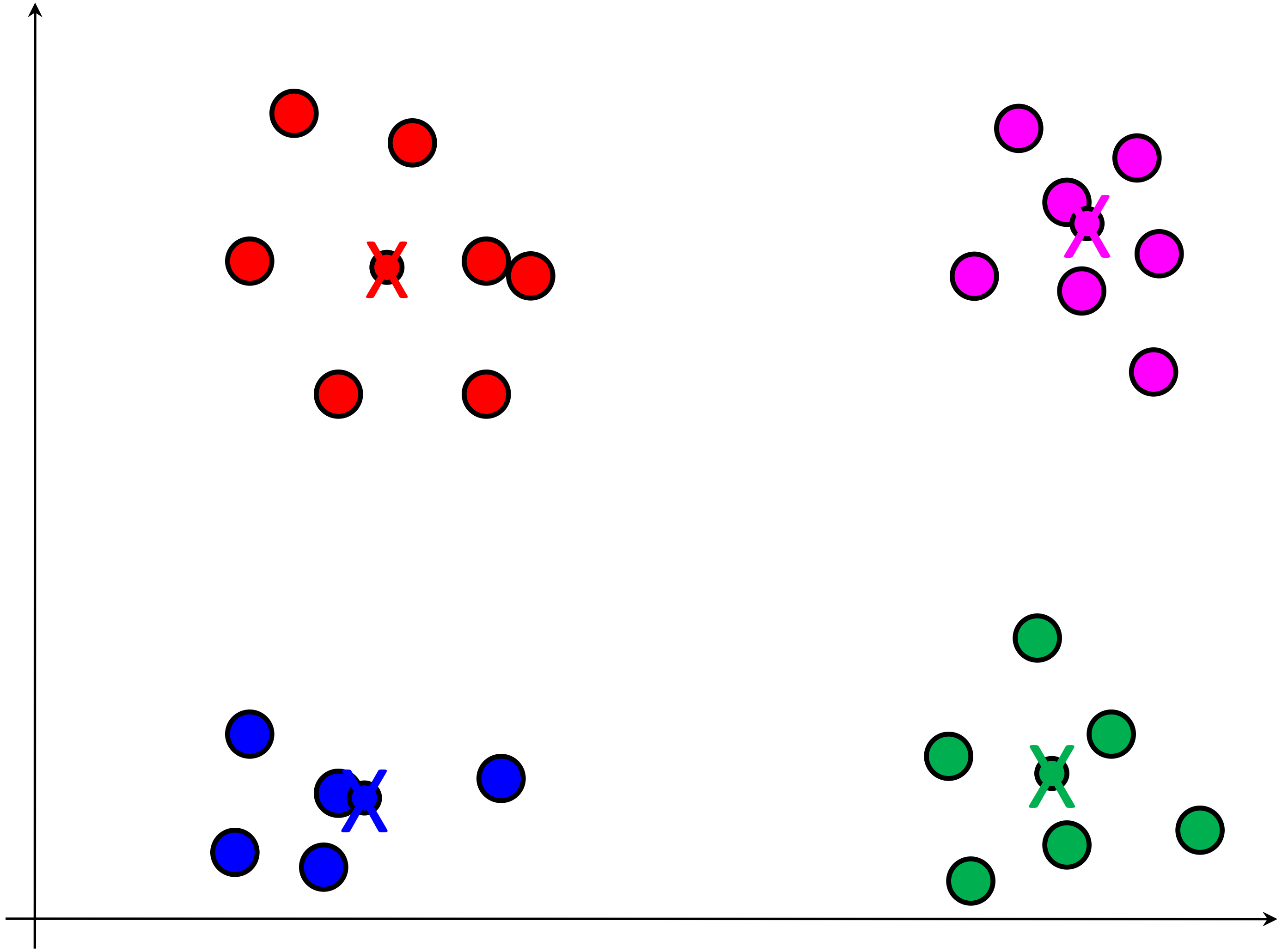
1. Importance of choosing initial centroids: recalculate centroids



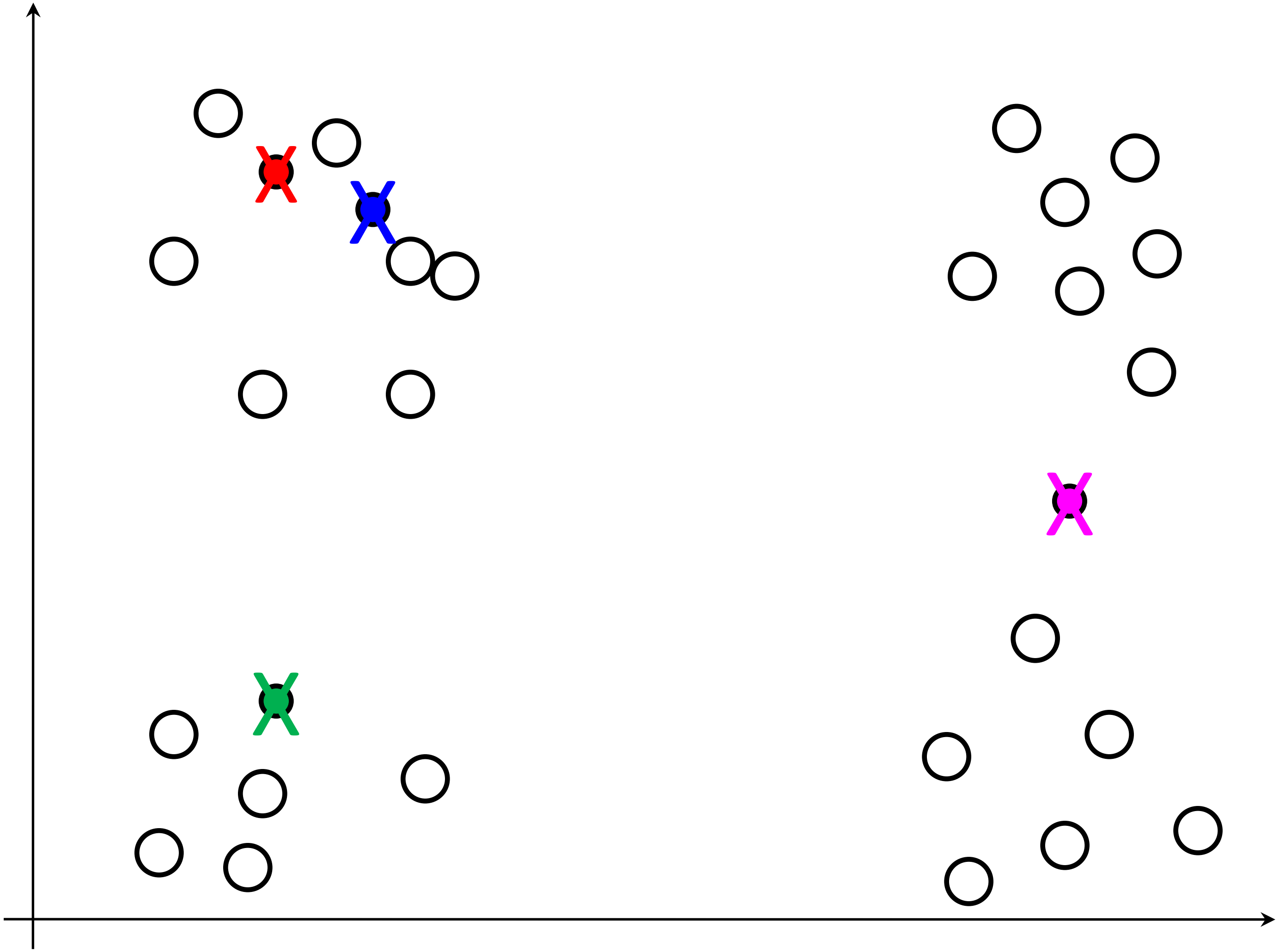
1. Importance of choosing initial centroids: points re-assignments



1. Importance of choosing initial centroids: success – correct clusters



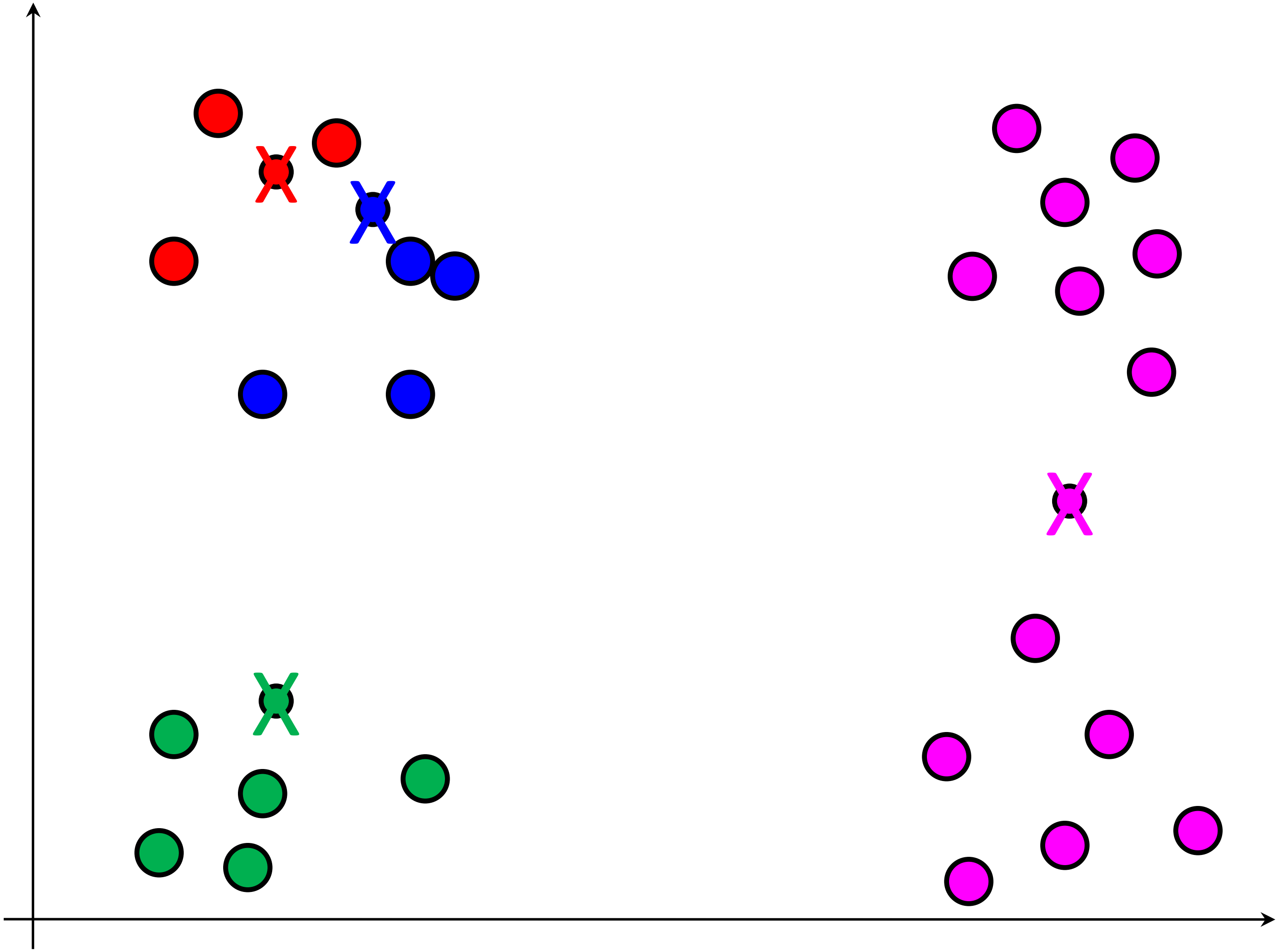
2. Importance of choosing initial centroids: K=4



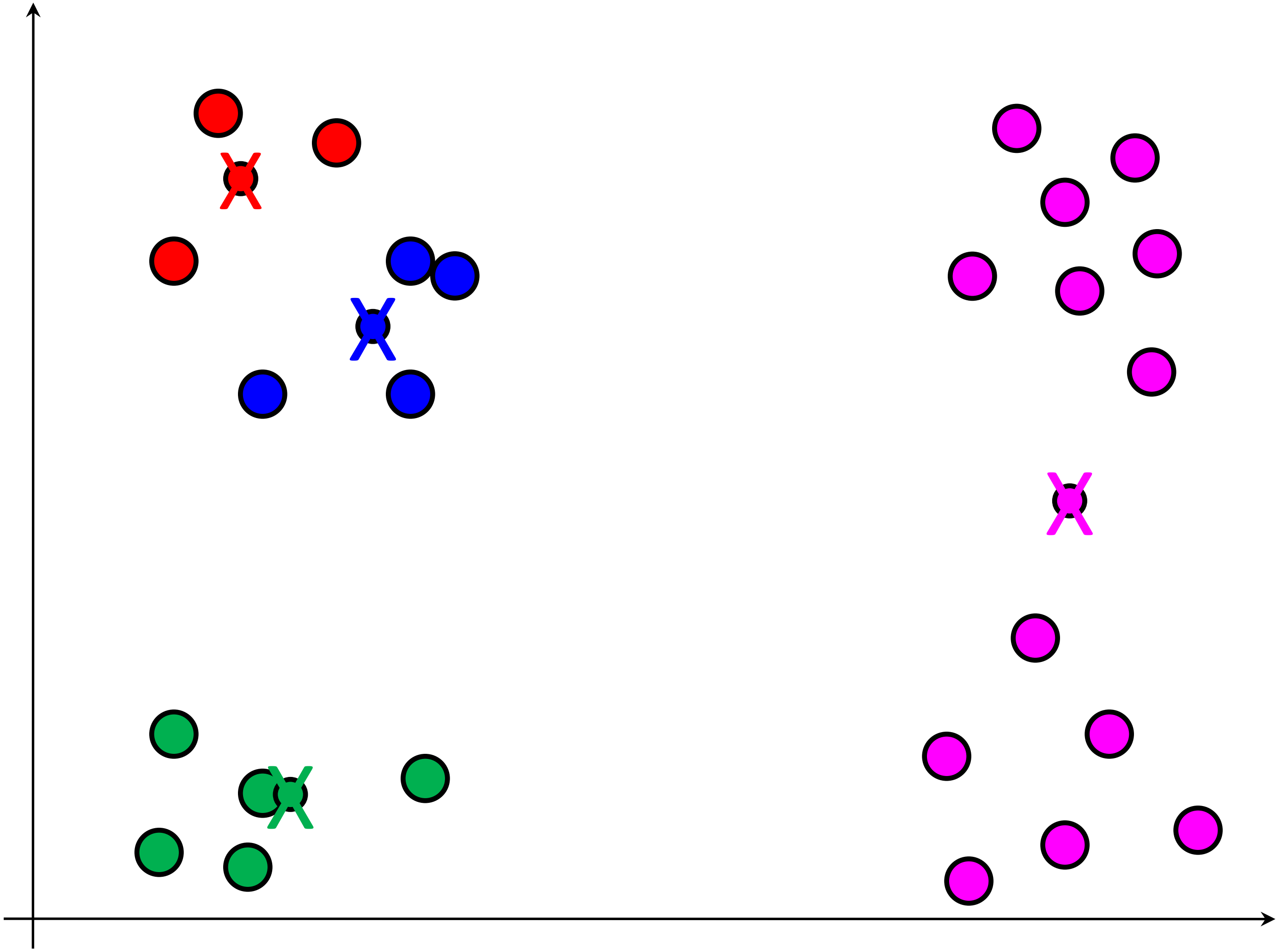
2 pairs of clusters

Initial seeds are chosen at random: 3 seeds in one pair

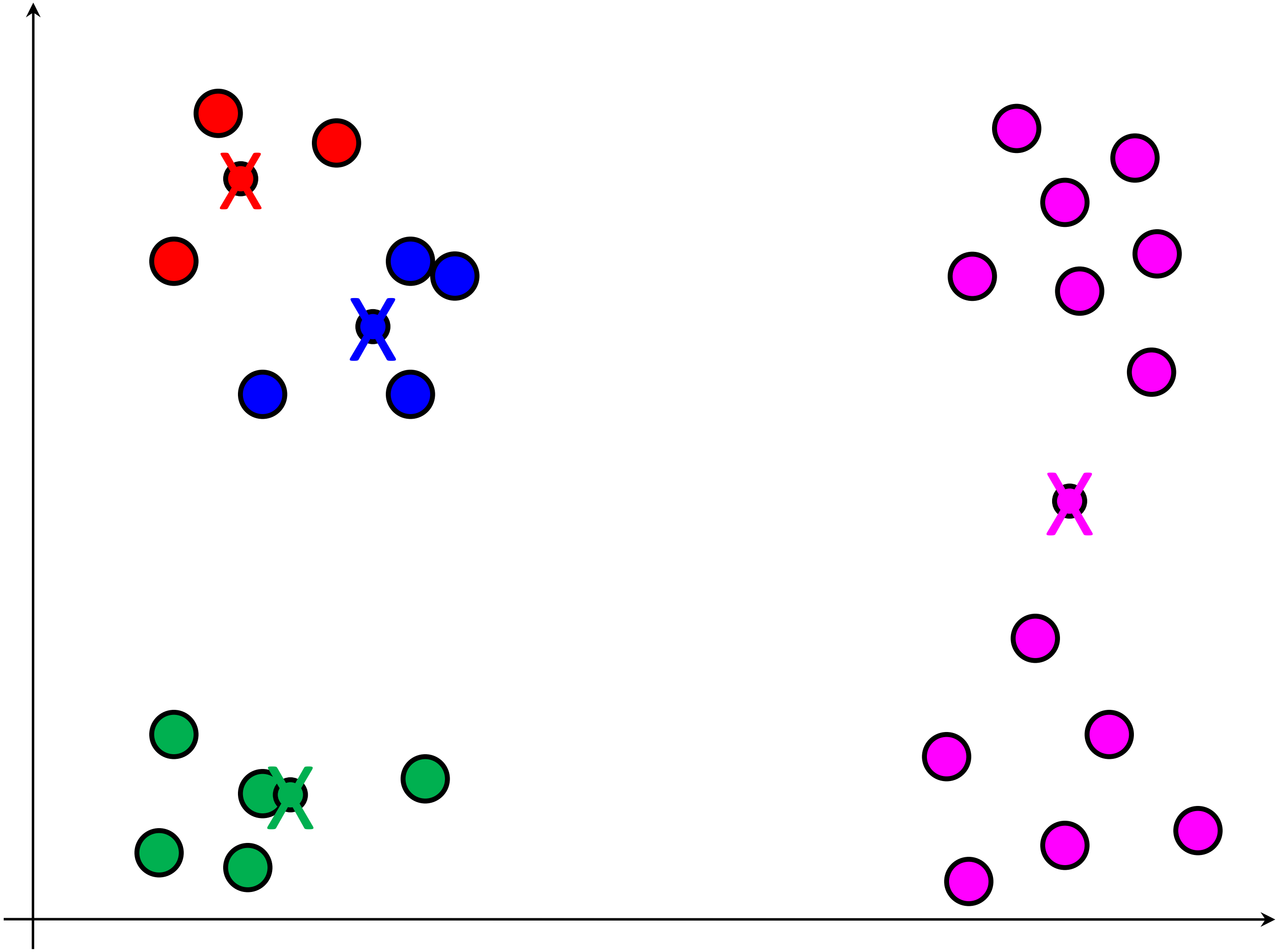
2. Importance of choosing initial centroids: assign points



2. Importance of choosing initial centroids: re-compute centroids



2. Importance of choosing initial centroids: found 4 clusters - incorrect!



Problem: selecting Initial Centroids

- Of course, the ideal would be to choose initial centroids, one from each true cluster.
- However, if there are *K* 'real' clusters then the chance of selecting one centroid from each cluster is extremely small.
  - Chance is relatively small when *K* is large
  - If clusters are the same size, *n*, then:

$$P = \frac{\text{number of ways to select one centroid from each cluster}}{\text{number of ways to select } K \text{ centroids}} = \frac{K!n^K}{(Kn)^K} = \frac{K!}{K^K}$$

- For example, if $K = 10$, then *probability* = $10!/10^{10} = 0.00036$
- Sometimes the initial centroids re-adjust themselves in the 'right' way, and sometimes they don't.

Solutions to Initial Centroids Problem

- Multiple runs
  - Helps, but probability is not on your side
- Bisecting K-means
  - Not as susceptible to initialization issues

Bisecting *K*-means

- Straightforward extension of the basic *K*-means algorithm
- Simple idea:

  To obtain *K* clusters, split the set of points into two clusters, select one of these clusters to split, and so on, until *K* clusters have been produced

Bisecting K-means

Initialize the list of clusters with one cluster consisting of all points.

Do

Select a cluster with the highest SSE from the list of clusters

Perform several "trial" bisections of the chosen cluster:

for i = 1 **to** number of trials **do**

Bisect the selected cluster using basic K-means (i.e. 2-means).

end for

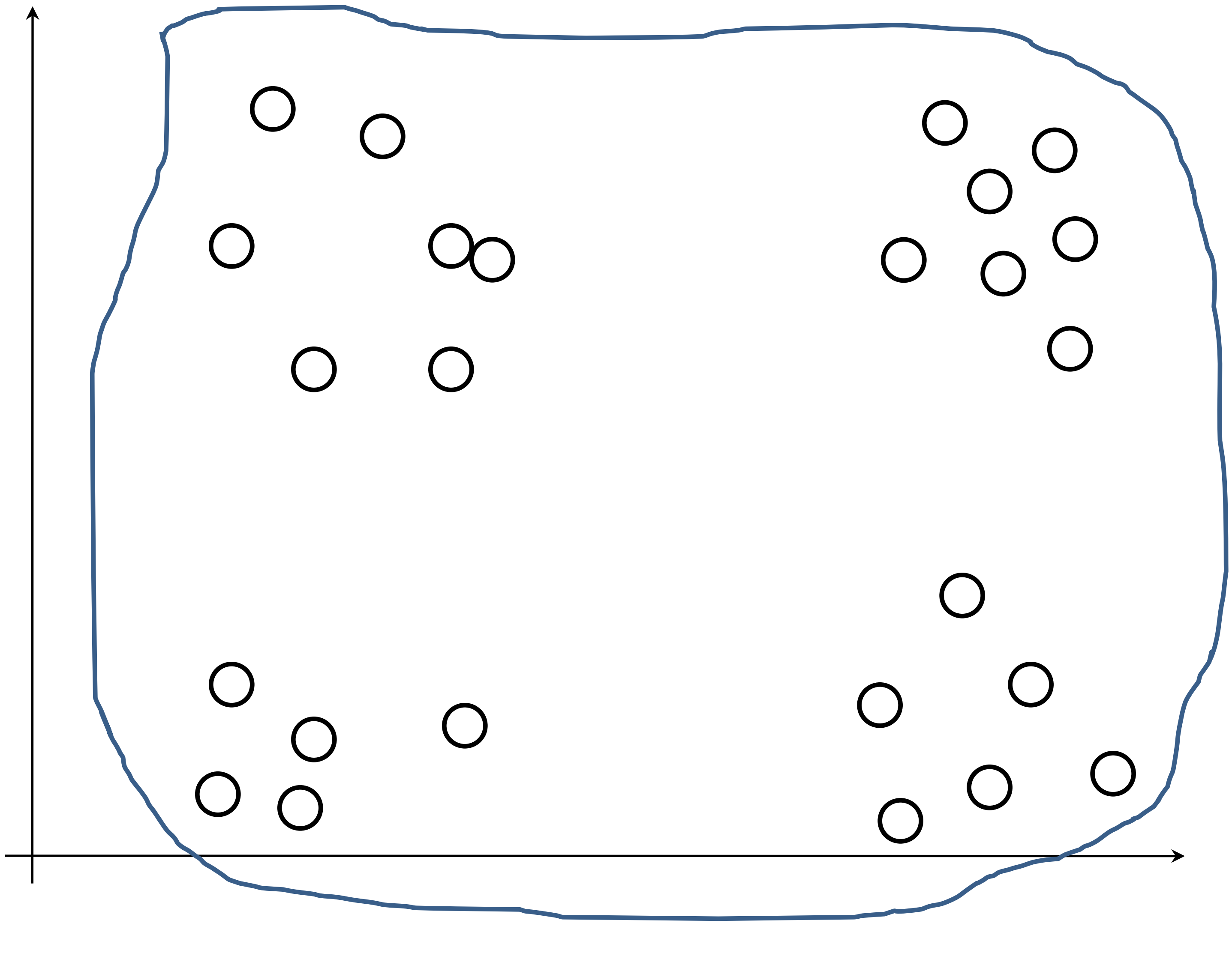
Select the two clusters from the bisection

with the lowest intra-cluster distances (SSE)

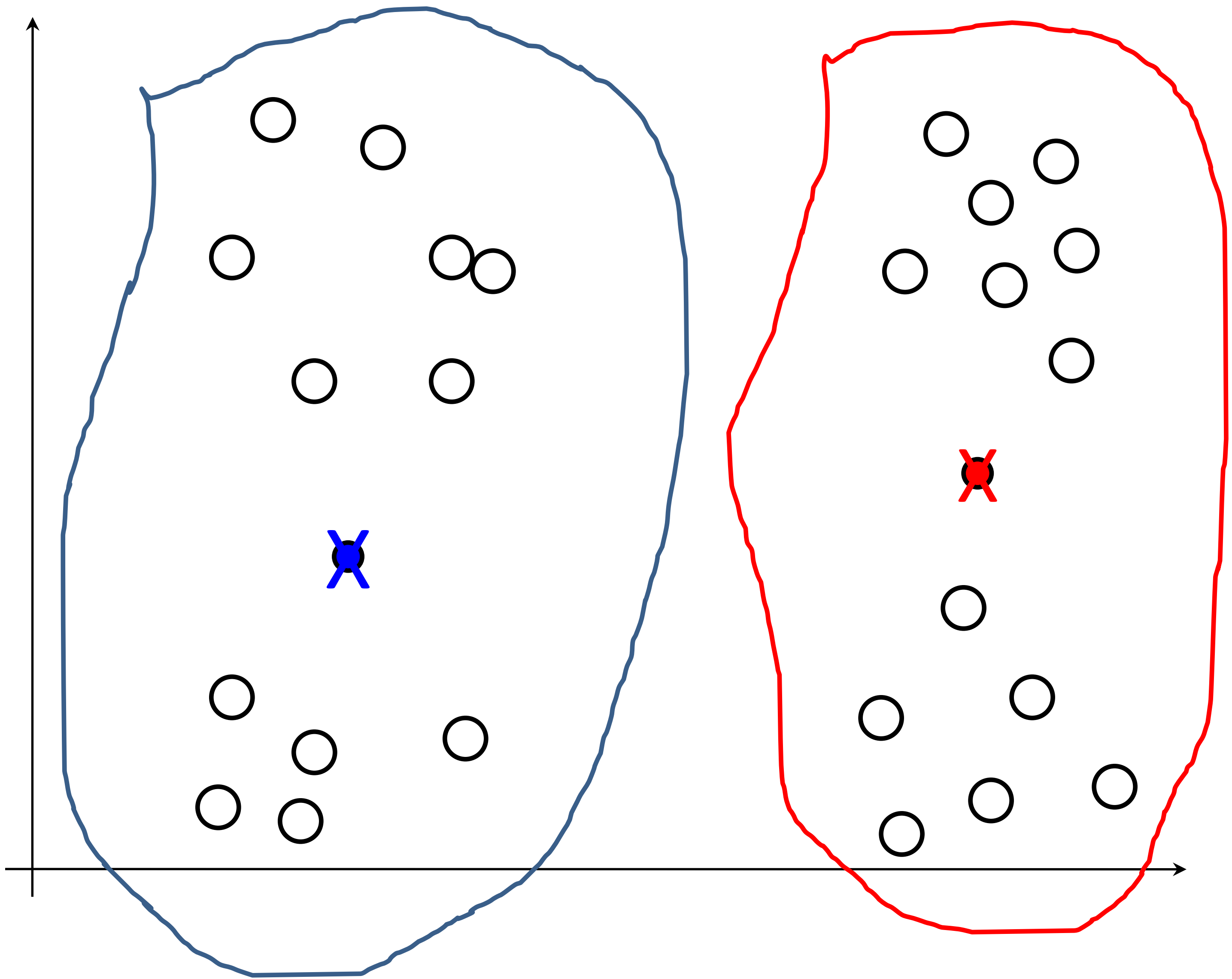
Add these two clusters to the list of clusters

Until the list of clusters contains K clusters.

Bisecting K-means example: one initial cluster



Bisecting K-means example: bisecting initial cluster

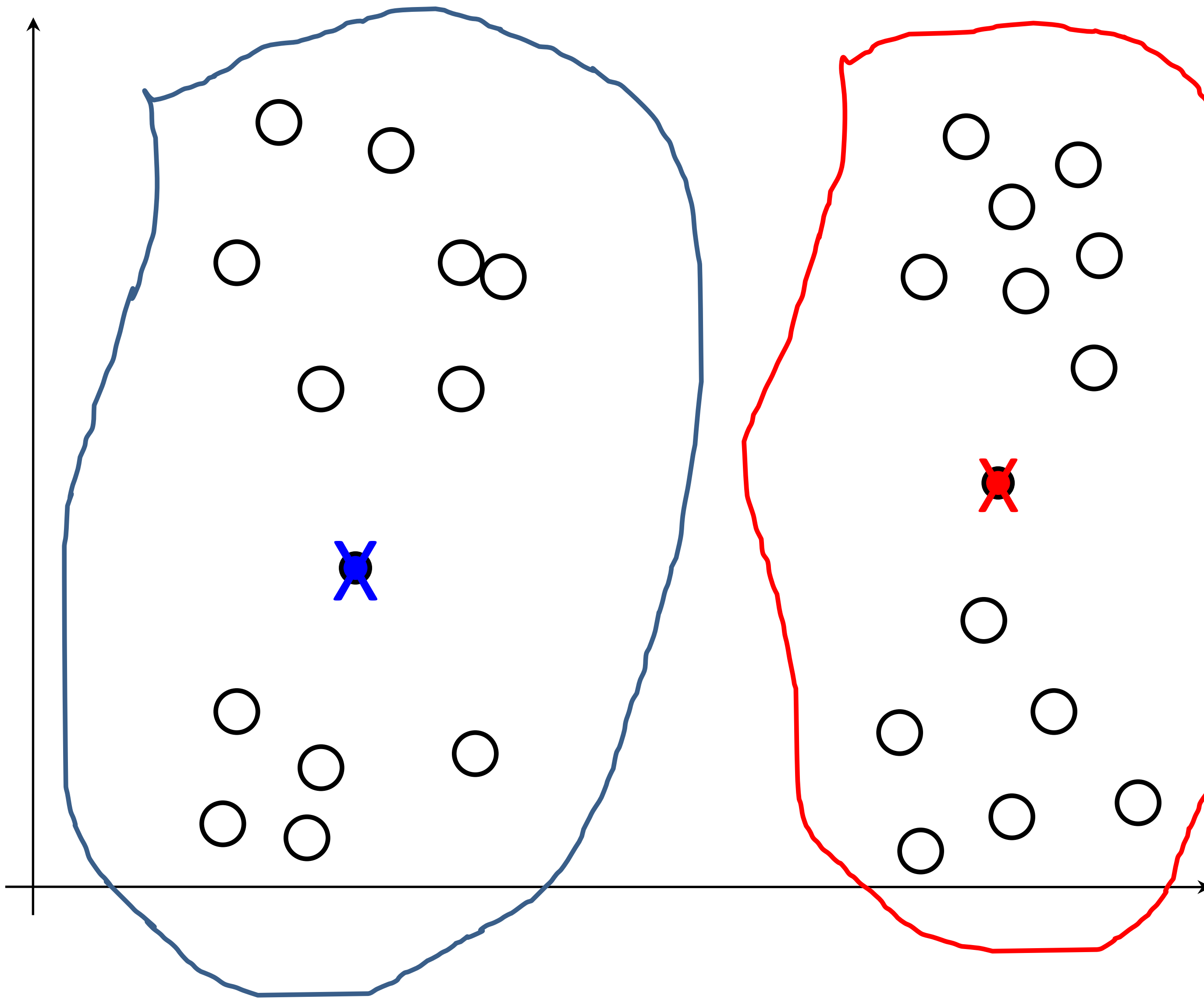


Perform K-means algorithm for K=2

Discovered 2 clusters

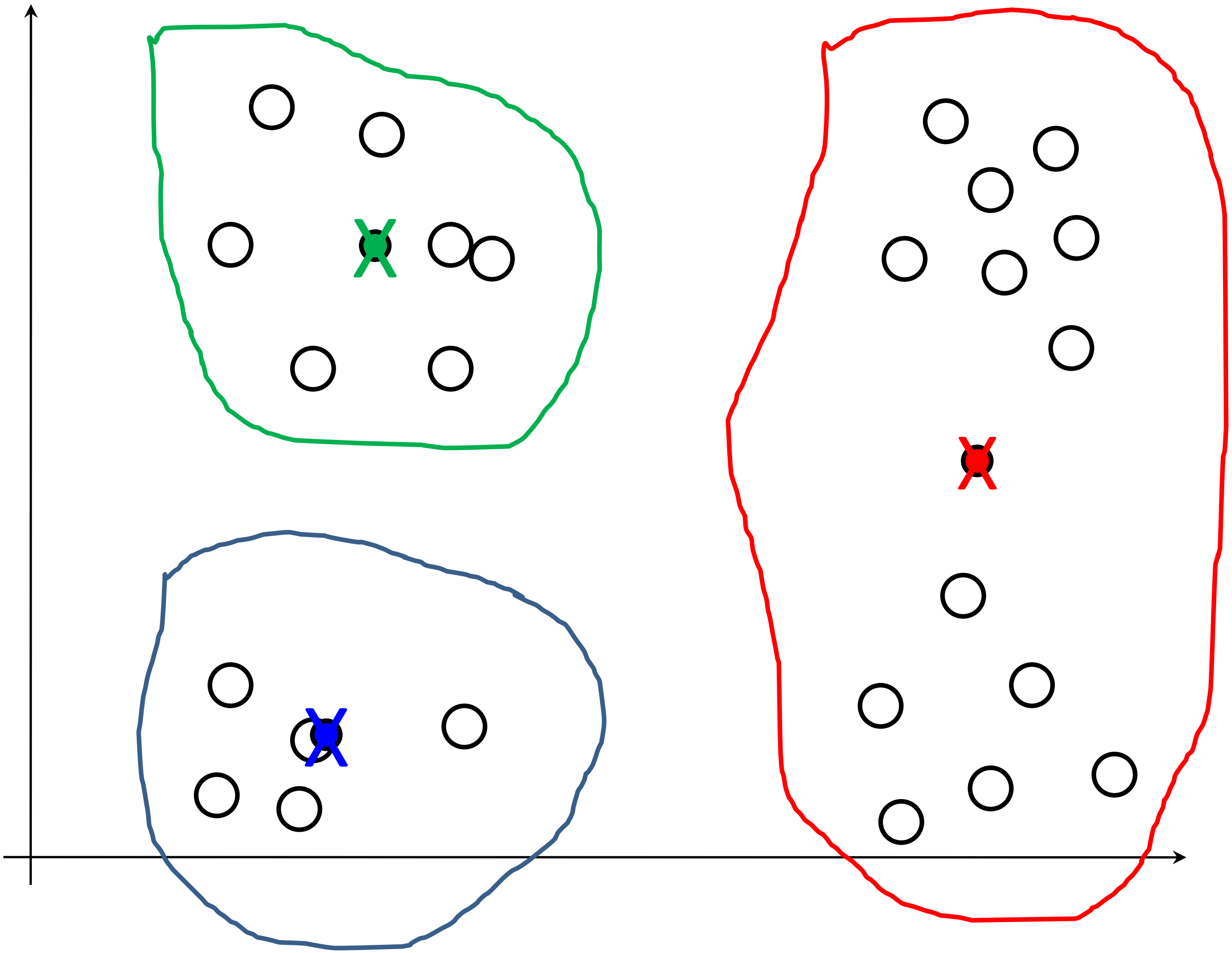
Blue cluster has larger SSE

Bisecting K-means example: bisecting blue cluster



Perform K-means algorithm for K=2 on a blue cluster

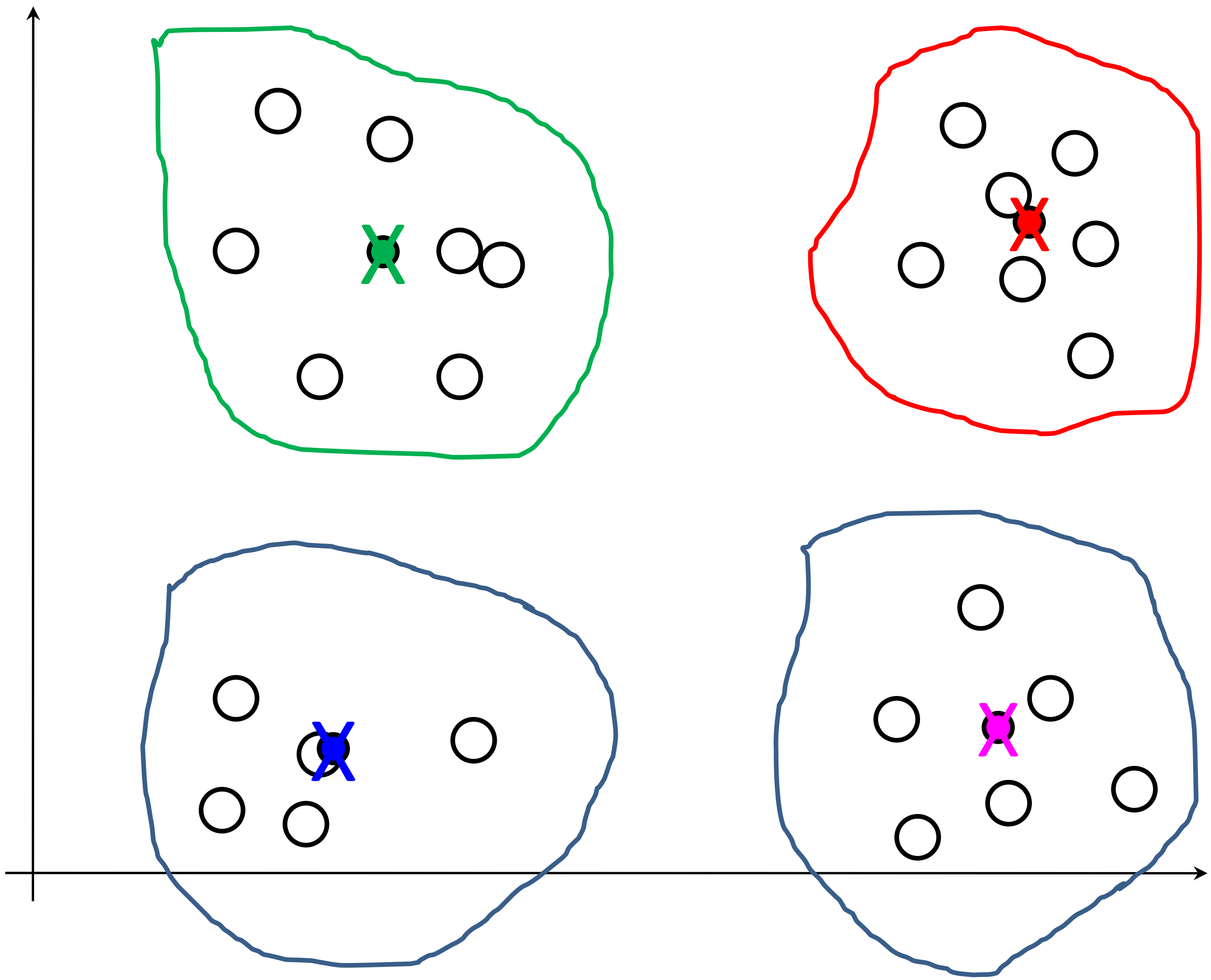
Bisecting K-means example: 3 clusters



Found 2 clusters.

Now red cluster has the largest SSE.

Bisecting K-means example: bisecting red cluster



Process red cluster.

Found 4 clusters.

Stop.

Bisecting K-means Example