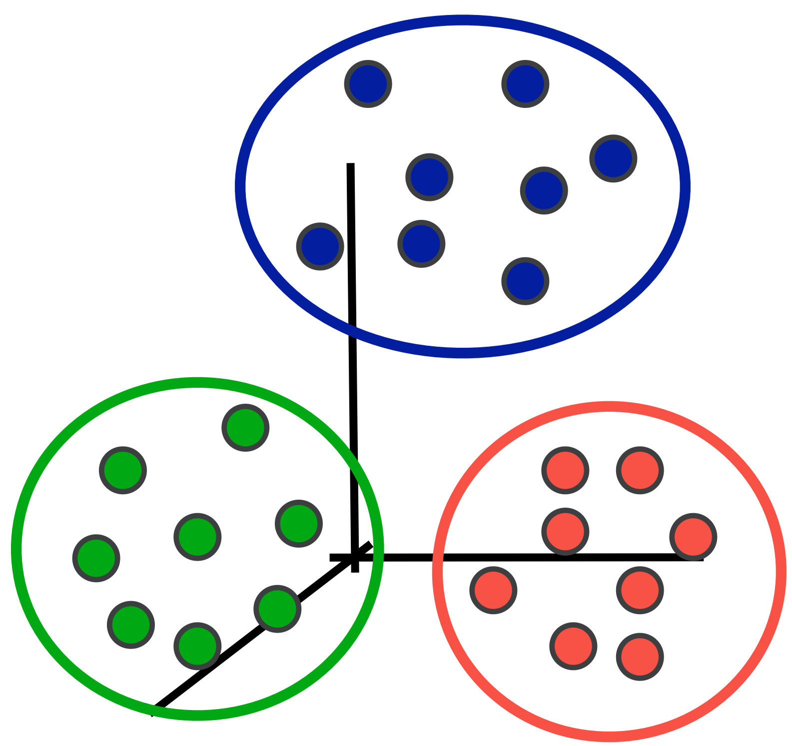
Clustering, k-means, k-means++ and the advantages of careful seeding

- David Arthur, Sergei Vassilvitskii. *k-means++: The Advantages of Careful Seeding*. In SODA 2007

What is clustering?

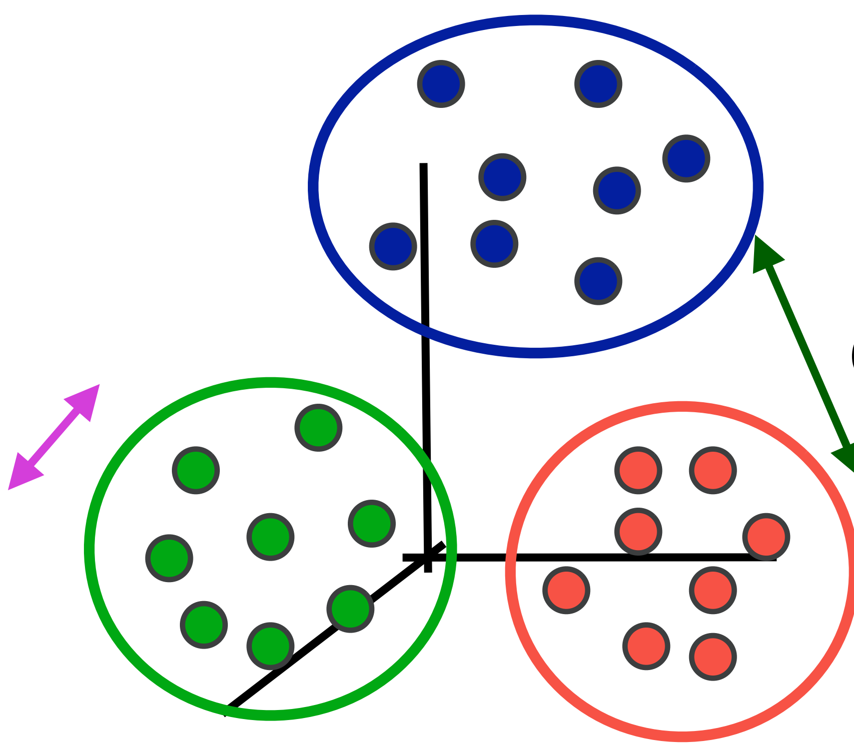
- a grouping of data objects such that the objects within a group are similar (or near) to one another and dissimilar (or far) from the objects in other groups



How to capture this objective?

a grouping of data objects such that the objects within a group are similar (or near) to one another and dissimilar (or far) from the objects in other groups

minimize
intra-cluster
distances



maximize
inter-cluster
distances

The clustering problem

- Given a collection of data objects
- Find a grouping so that
  - similar objects are in the same cluster
  - dissimilar objects are in different clusters

- Why we care ?

- stand-alone tool to gain insight into the data
  - visualization
- preprocessing step for other algorithms
  - indexing or compression often relies on clustering

Applications of clustering

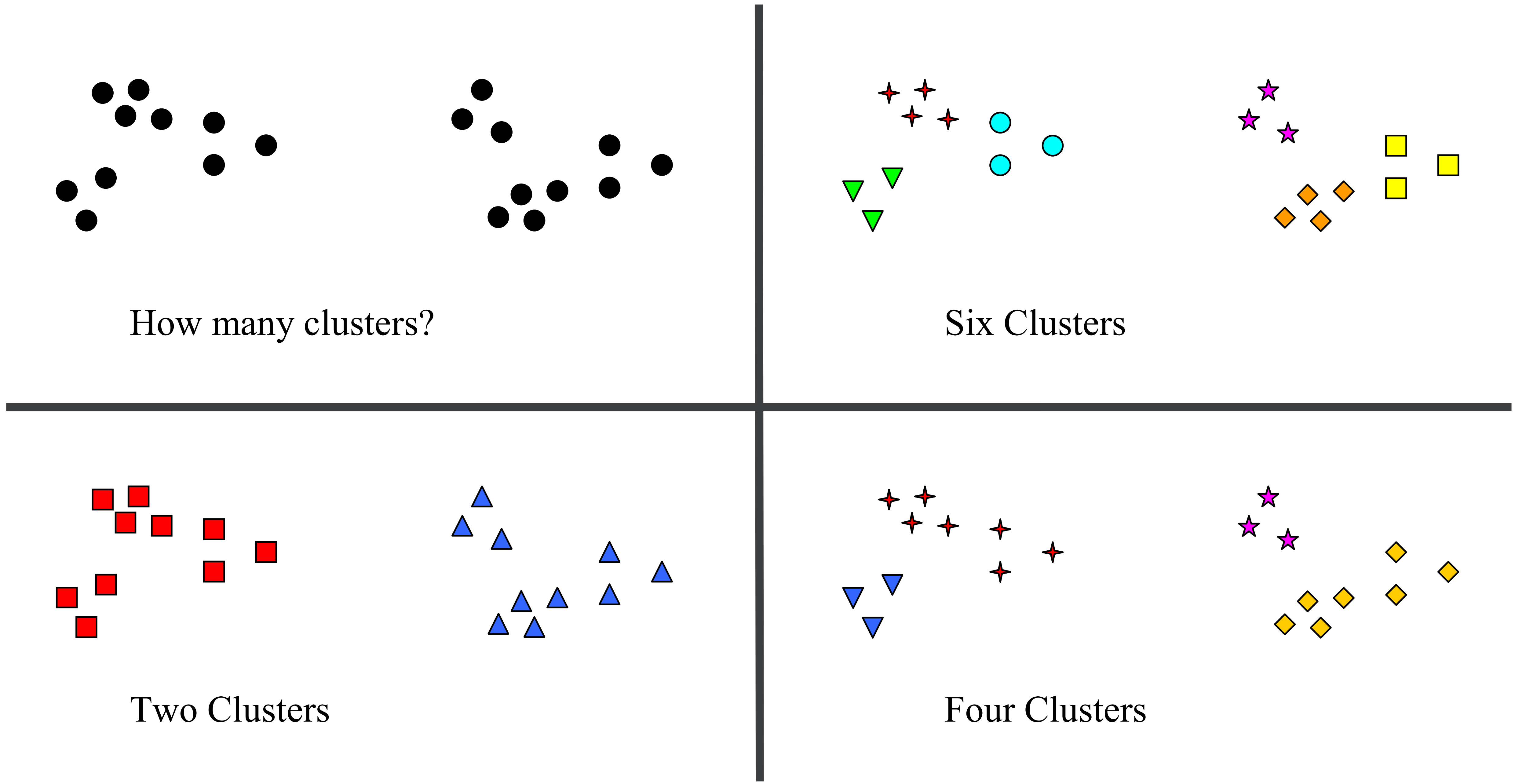
- image processing
  - cluster images based on their visual content
- web mining
  - cluster groups of users based on their access patterns on webpages
  - cluster webpages based on their content
- bioinformatics
  - cluster similar proteins together (similarity wrt chemical structure and/or functionality etc)
- many more...

The clustering problem

- Given a collection of data objects
- Find a grouping so that
  - similar objects are in the same cluster
  - dissimilar objects are in different clusters

- Basic questions:
  - what does similar mean?
  - what is a good partition of the objects?
    i.e., how is the quality of a solution measured?
  - how to find a good partition?

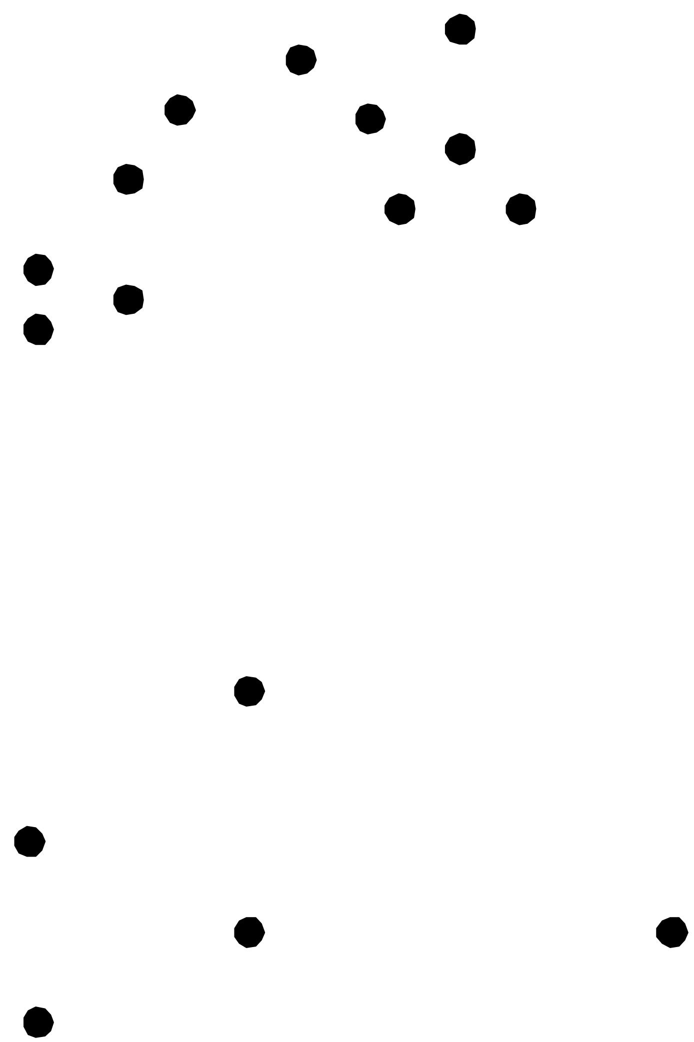
Notion of a cluster can be ambiguous



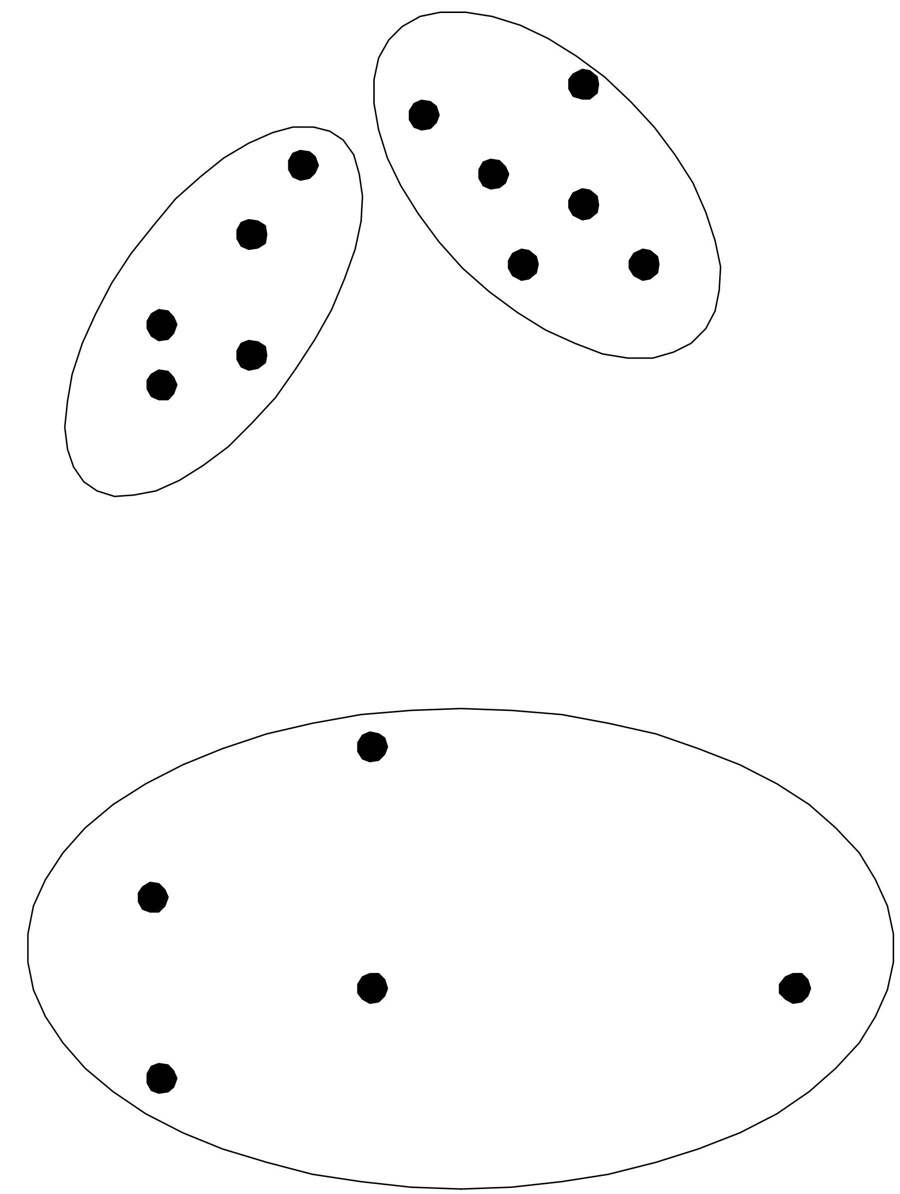
Types of clusterings

- Partitional
  - each object belongs in exactly one cluster
- Hierarchical
  - a set of nested clusters organized in a tree

Partitional clustering



Original Points



A Partitional Clustering

Partitional algorithms

- partition the n objects into k clusters
  - each object belongs to exactly one cluster
  - the number of clusters k is given in advance

The k-means problem

- consider set $X=\{x_1,\ldots,x_n\}$ of n points in R^d
- assume that the number k is given
- problem:
  - find k points $c_1,\ldots,c_k$ (named centers or means)

    so that the cost

$$\sum_{i=1}^{n} \min_{j} \left\{ L_2^2(x_i, c_j) \right\} = \sum_{i=1}^{n} \min_{j} ||x_i - c_j||_2^2$$

    is minimized

The k-means problem

- consider set $X=\{x_1,...,x_n\}$ of n points in R^d
- assume that the number k is given
- problem:
  - find k points $c_1,...,c_k$ (named centers or means)
  - and partition X into $\{X_1,...,X_k\}$ by assigning each point x_i in X to its nearest cluster center,
  - so that the cost

$$\sum_{i=1}^{n} \min_{j} ||x_i - c_j||_2^2 = \sum_{j=1}^{k} \sum_{x \in X_j} ||x - c_j||_2^2$$

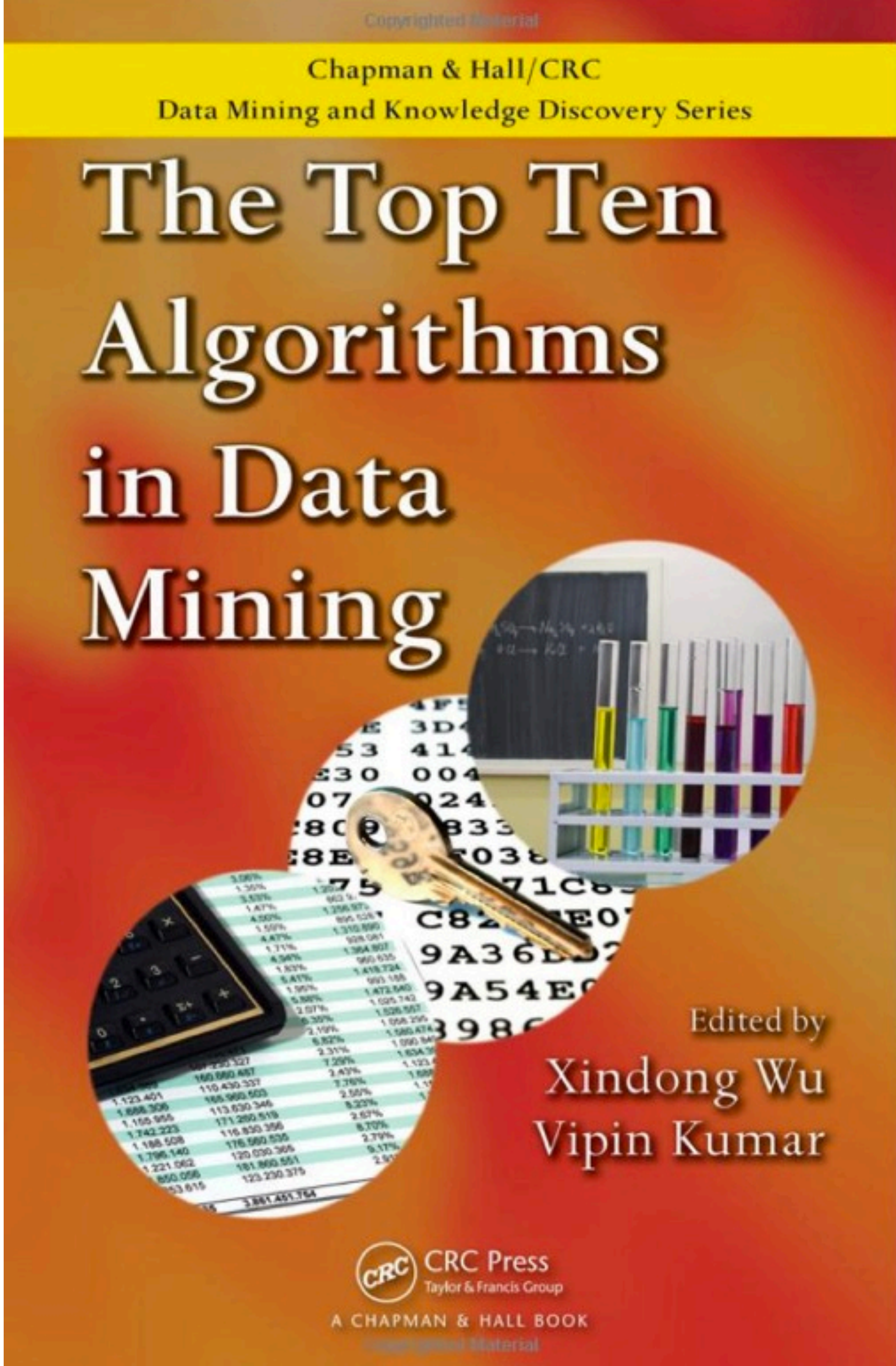
is minimized

The k-means problem

- $k=1$ and $k=n$ are easy special cases (why?)
- an NP-hard problem if the dimension of the data is at least 2 ($d \geq 2$)
- in practice, a simple iterative algorithm works quite well

The k-means algorithm

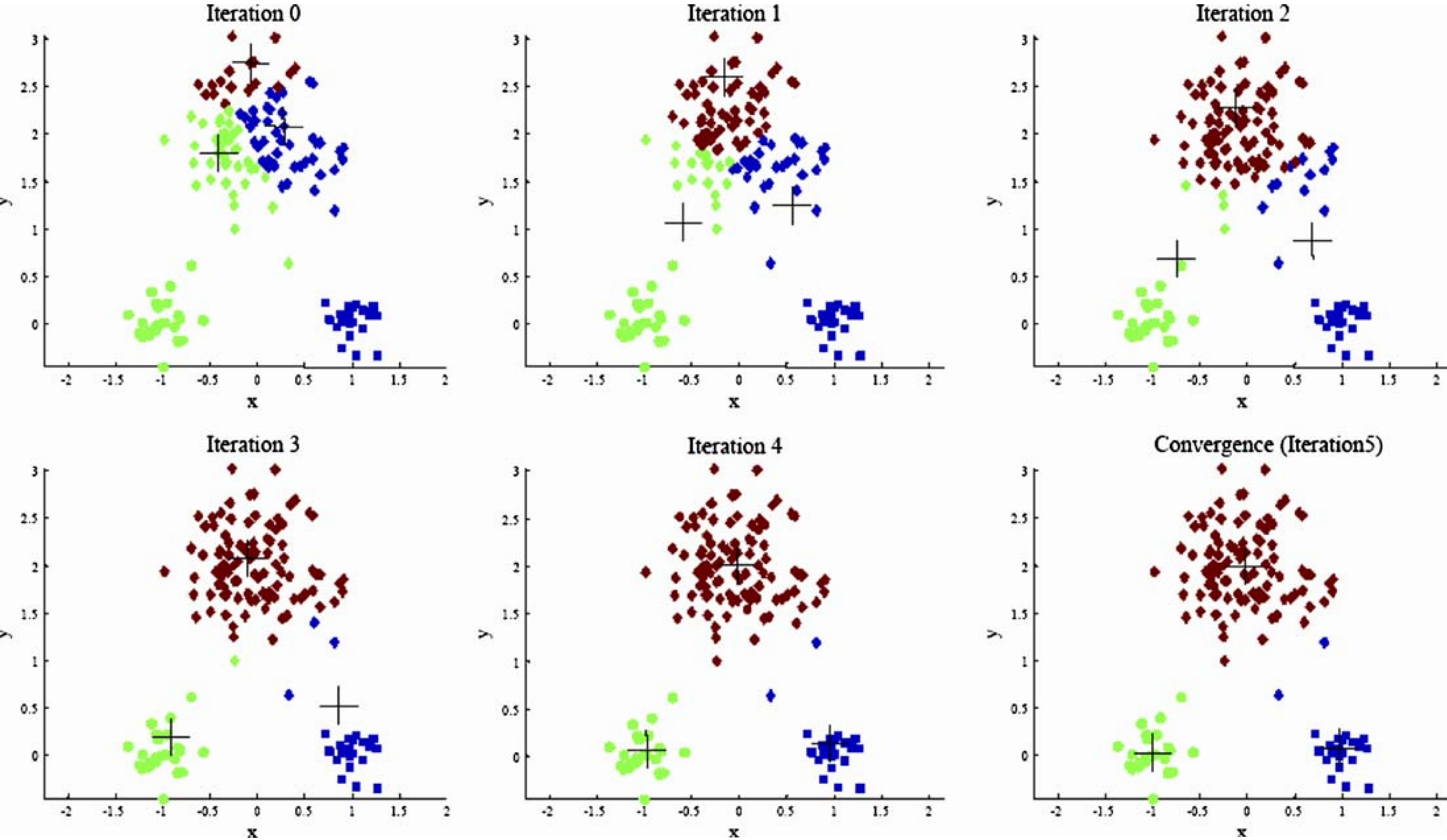
- voted among the top-10 algorithms in data mining
- one way of solving the k-means problem



The k-means algorithm

1. randomly (or with another method) pick k cluster centers $\{c_1,\ldots,c_k\}$
2. for each j, set the cluster X_j to be the set of points in X that are the closest to center c_j
3. for each j let c_j be the center of cluster X_j (mean of the vectors in X_j)
4. repeat (go to step 2) until convergence

Sample execution

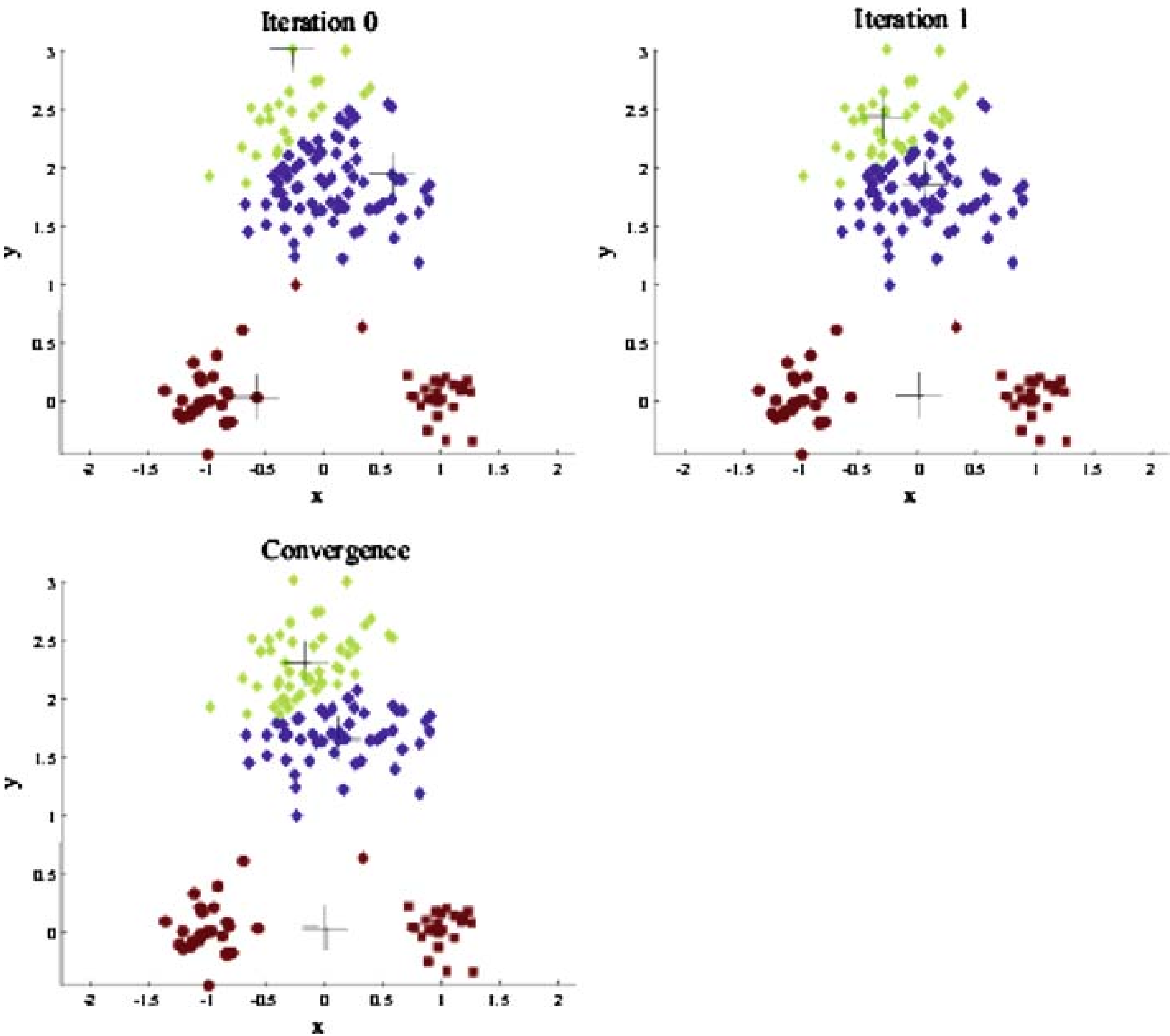


Properties of the k-means algorithm

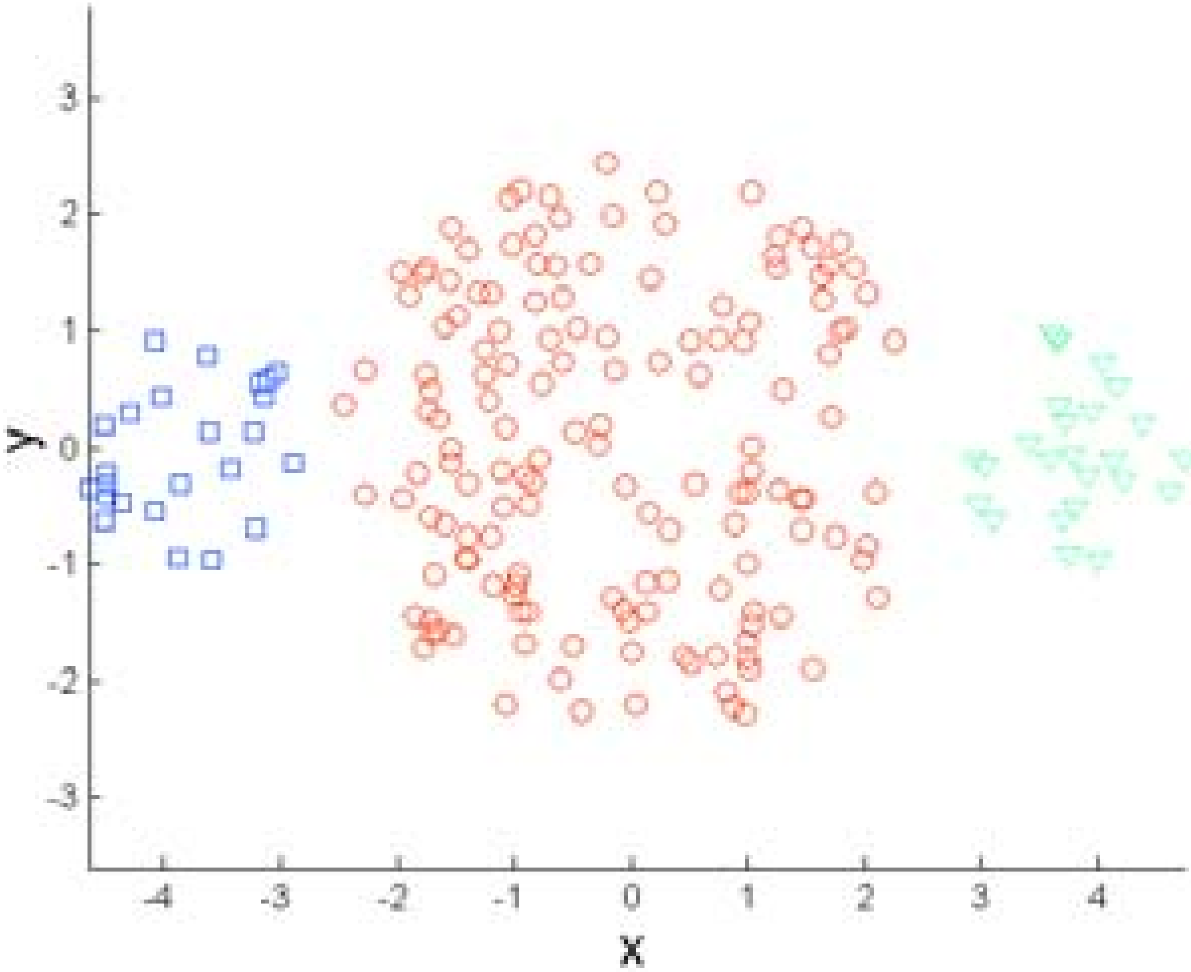
- finds a local optimum
- often converges quickly

  but not always
- the choice of initial points can have large influence in the result

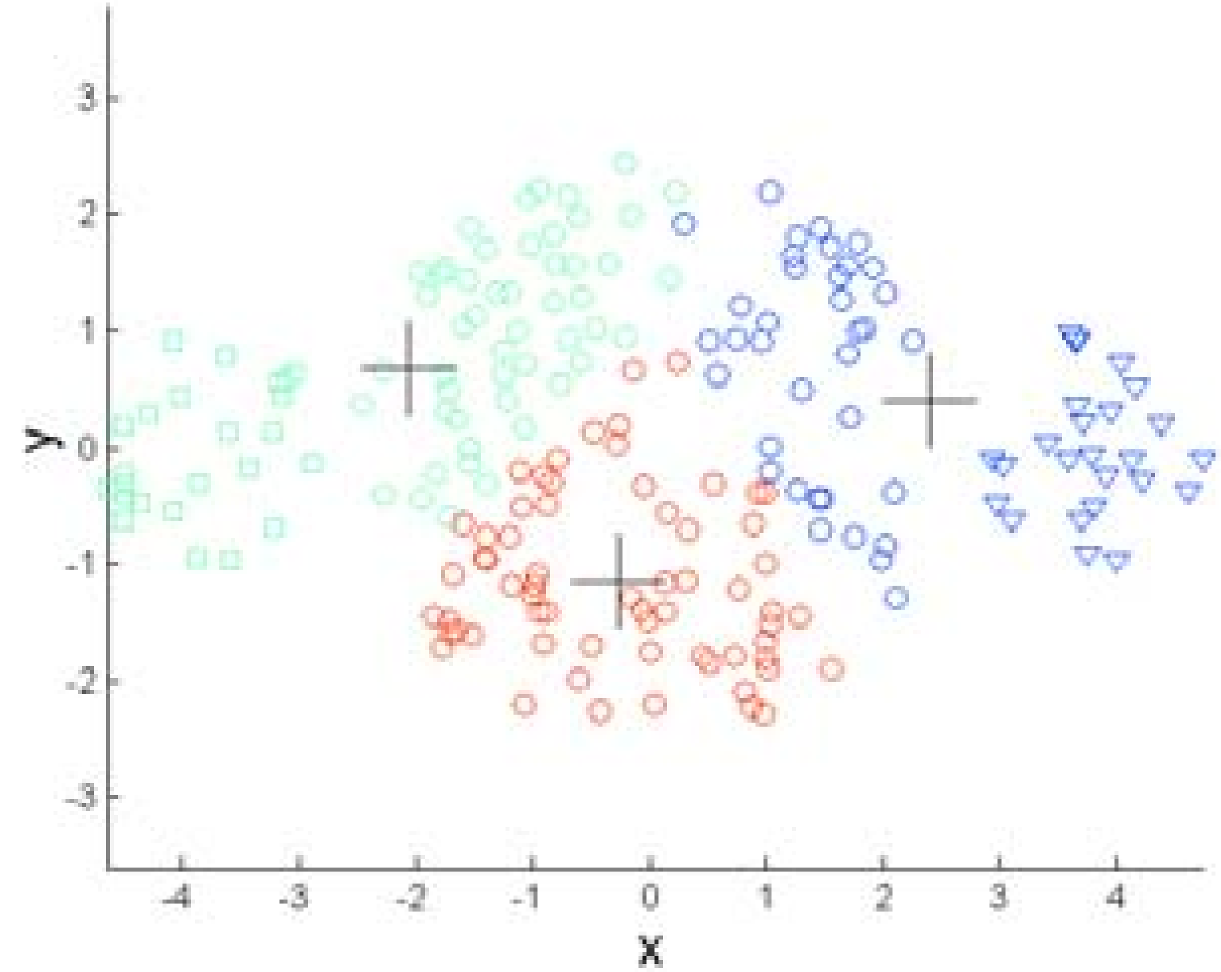
Effects of bad initialization



Limitations of k-means: different sizes

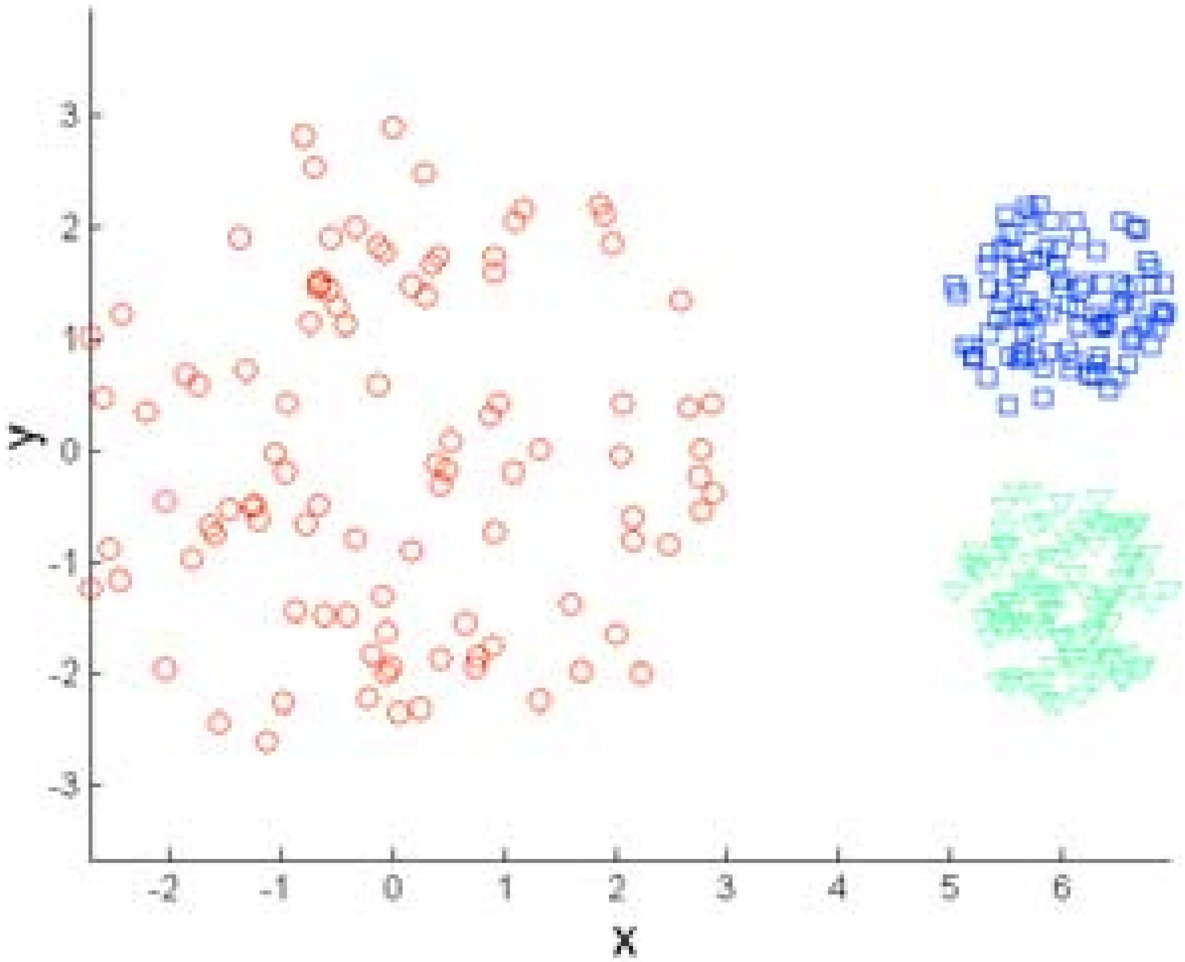


Original Points

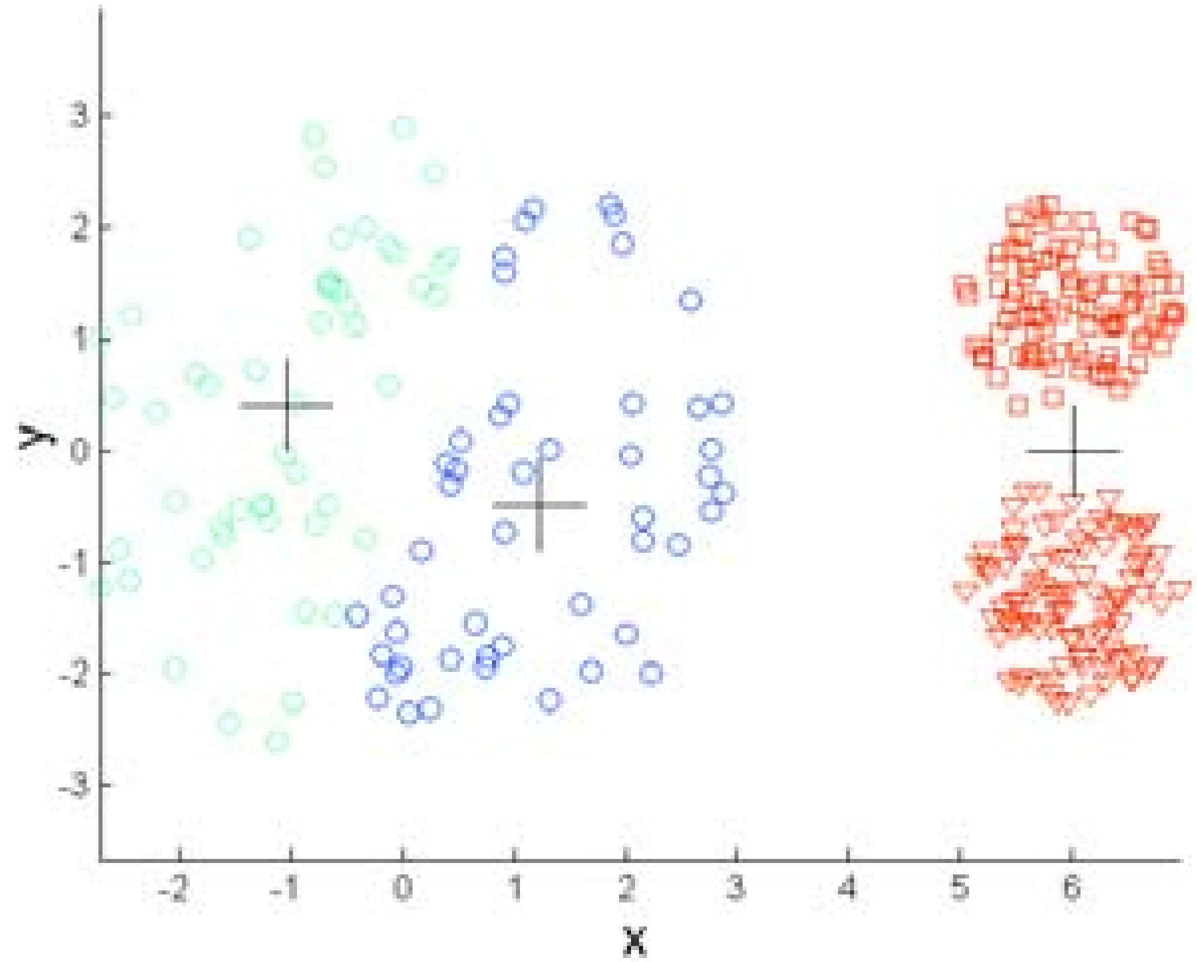


K-means (3 Clusters)

Limitations of k-means: different density

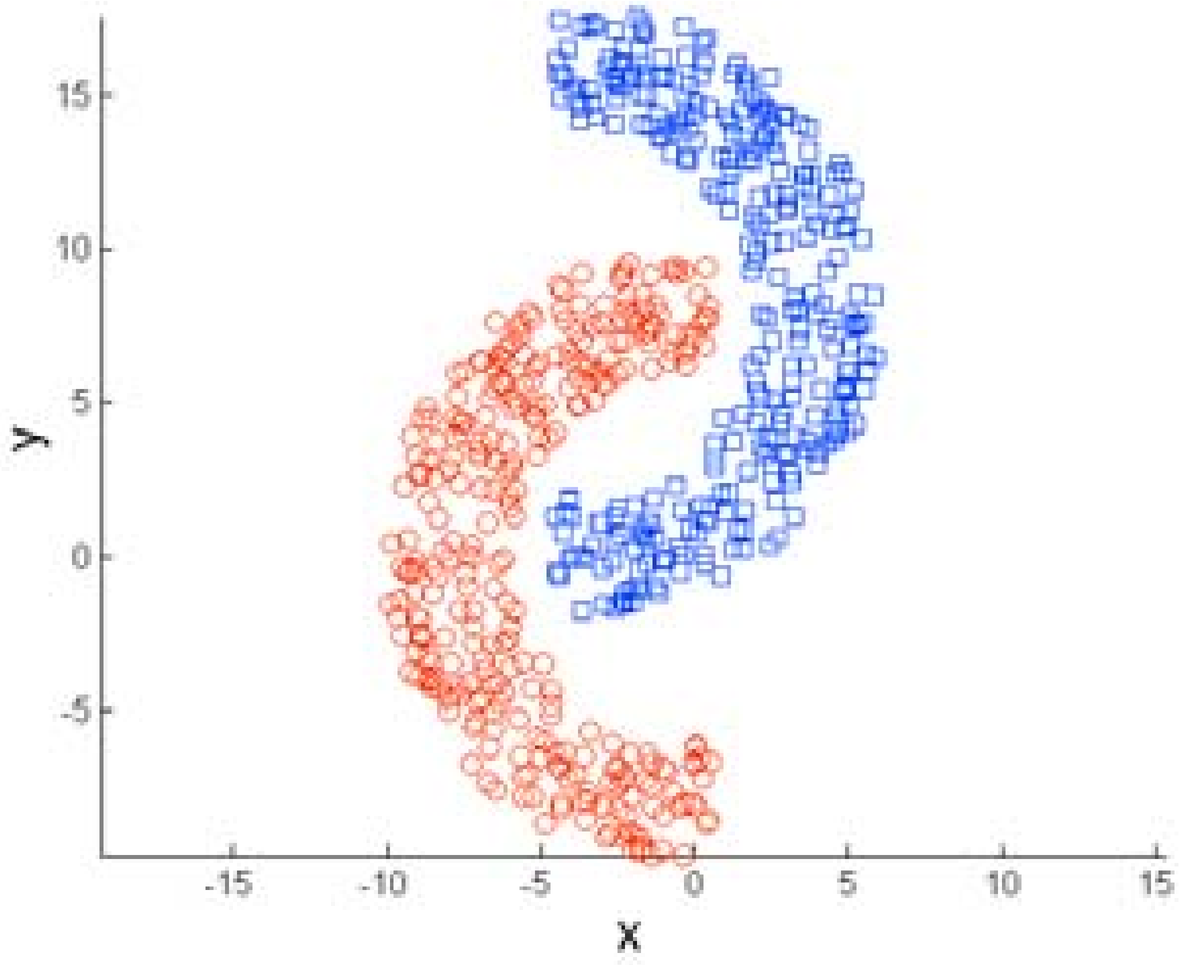


Original Points

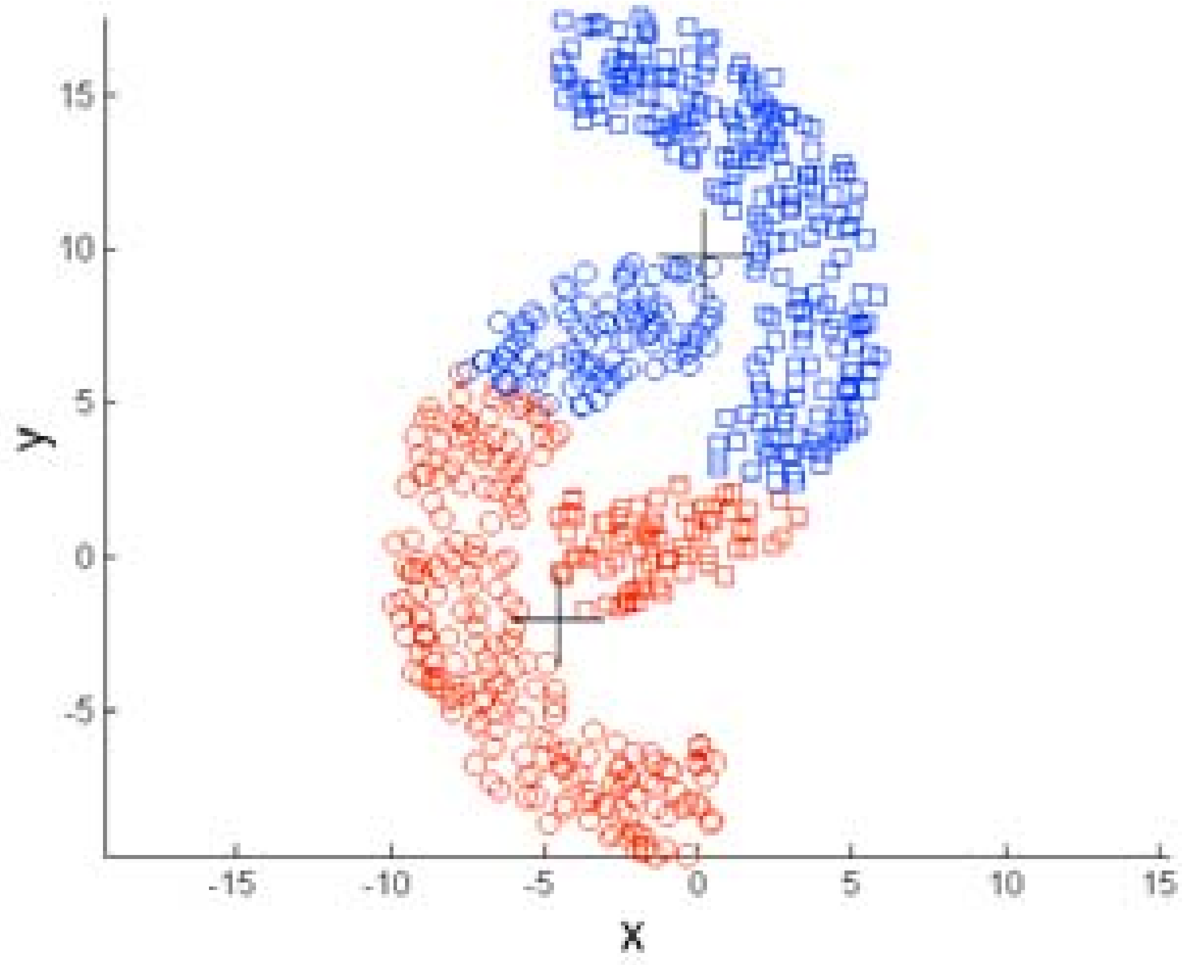


K-means (3 Clusters)

Limitations of k-means: non-spherical shapes



Original Points



K-means (2 Clusters)

Discussion on the k-means algorithm

- finds a local optimum
- often converges quickly
  - but not always
- the choice of initial points can have large influence in the result
- tends to find spherical clusters
- outliers can cause a problem
- different densities may cause a problem

Initialization

- random initialization
- random, but repeat many times and take the best solution
  - helps, but solution can still be bad
- pick points that are distant to each other
  - k-means++
  - provable guarantees

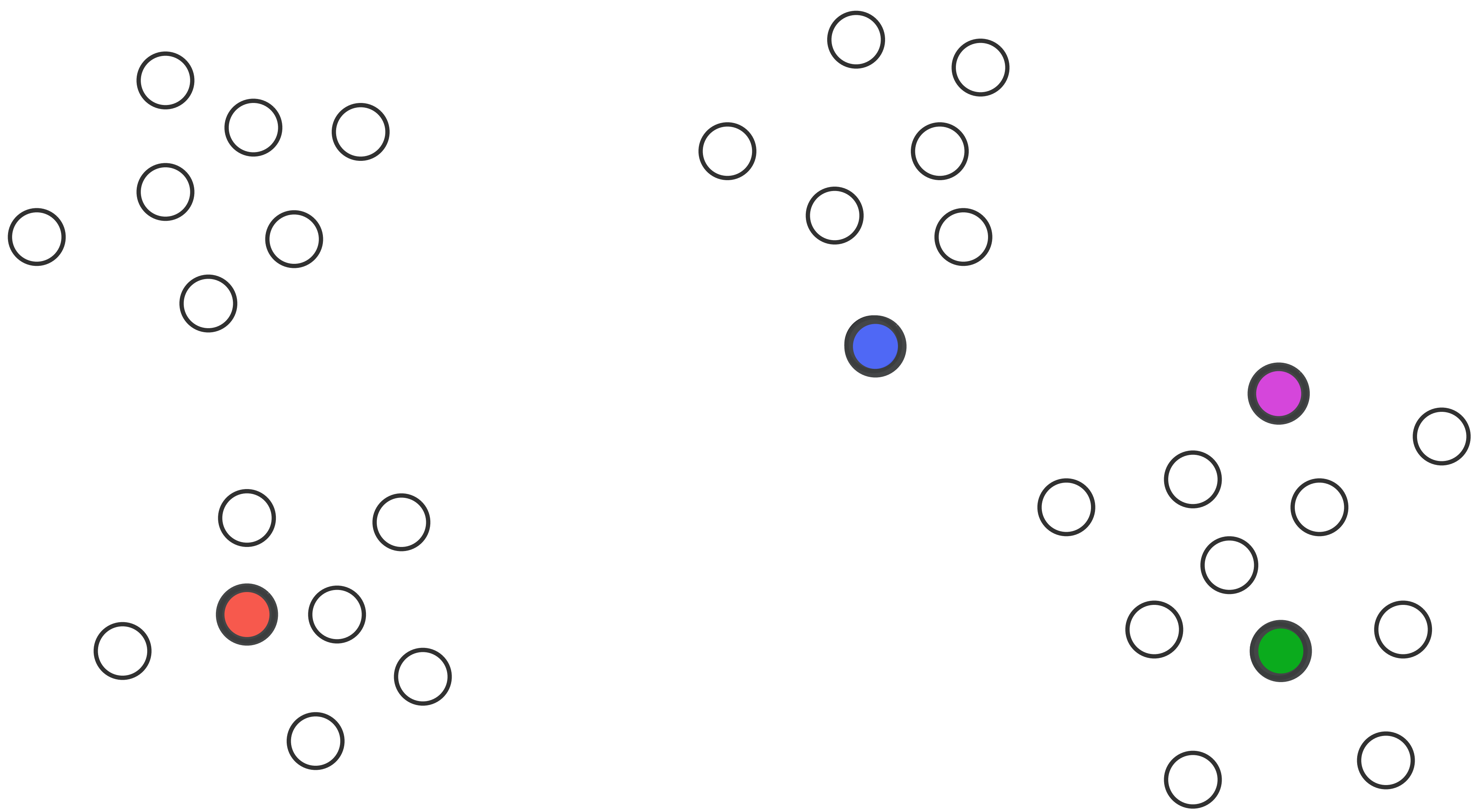
k-means++

David Arthur and Sergei Vassilvitskii

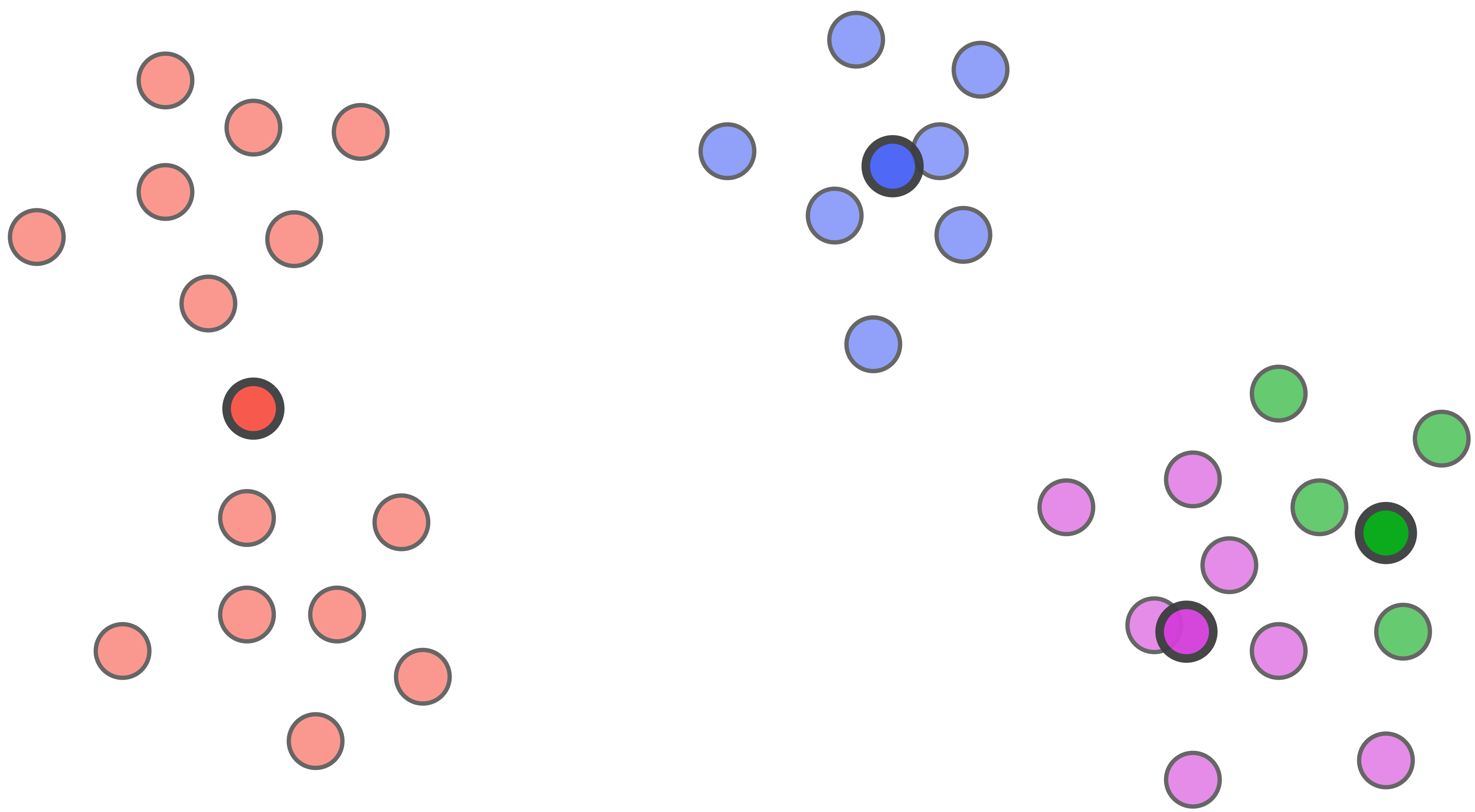
k-means++: The advantages of careful seeding

SODA 2007

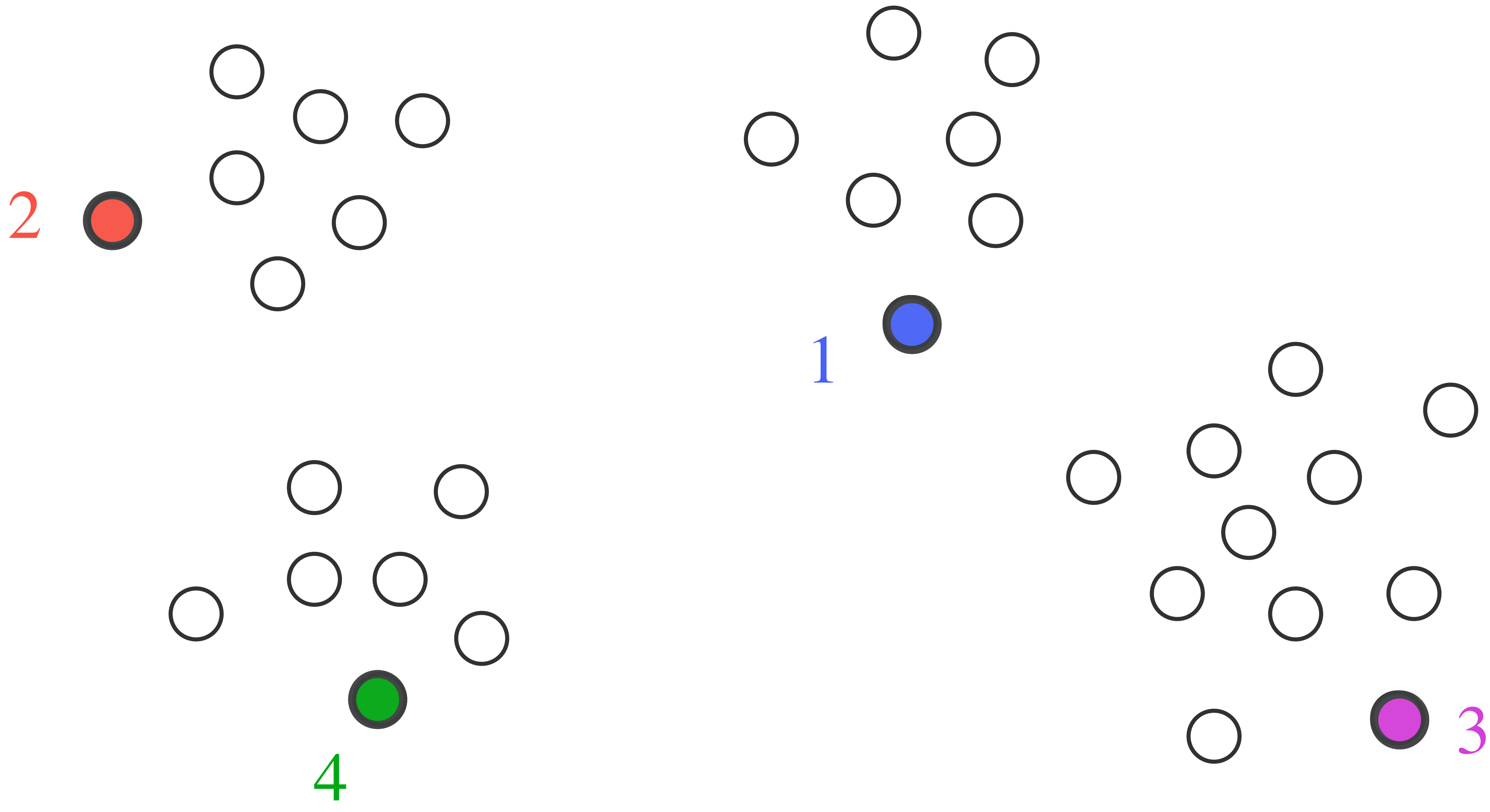
k-means algorithm: random initialization



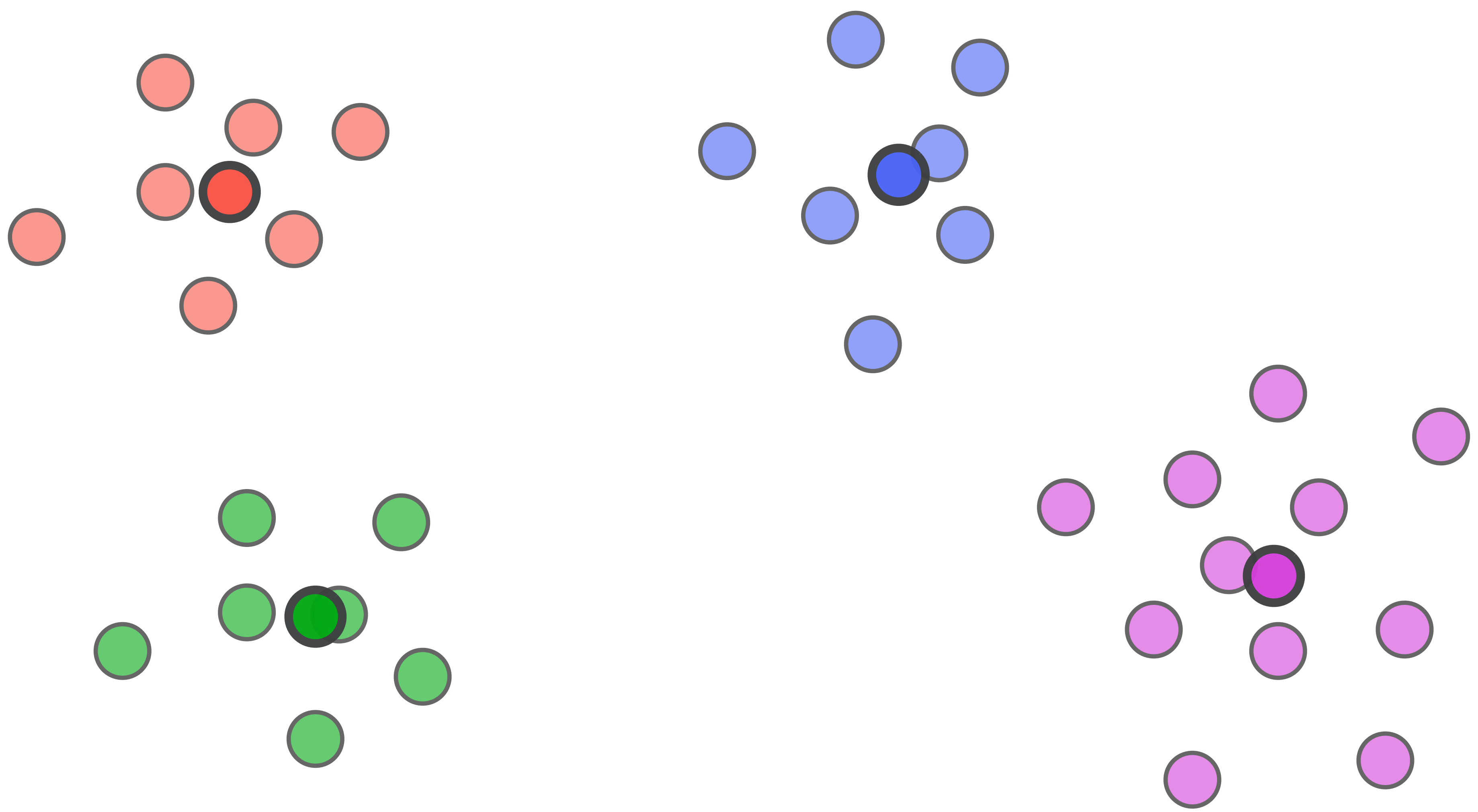
k-means algorithm: random initialization



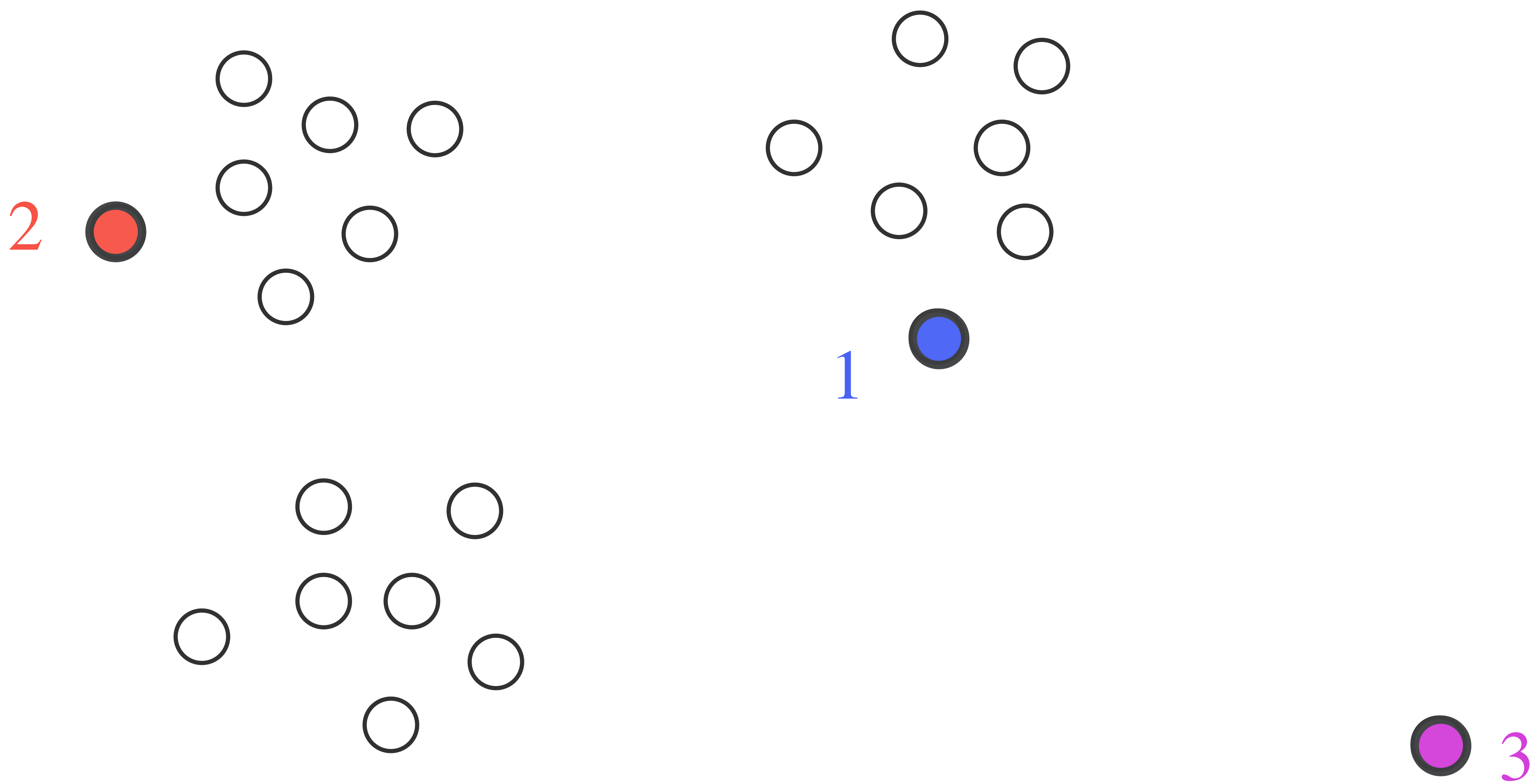
k-means algorithm: initialization with further-first traversal



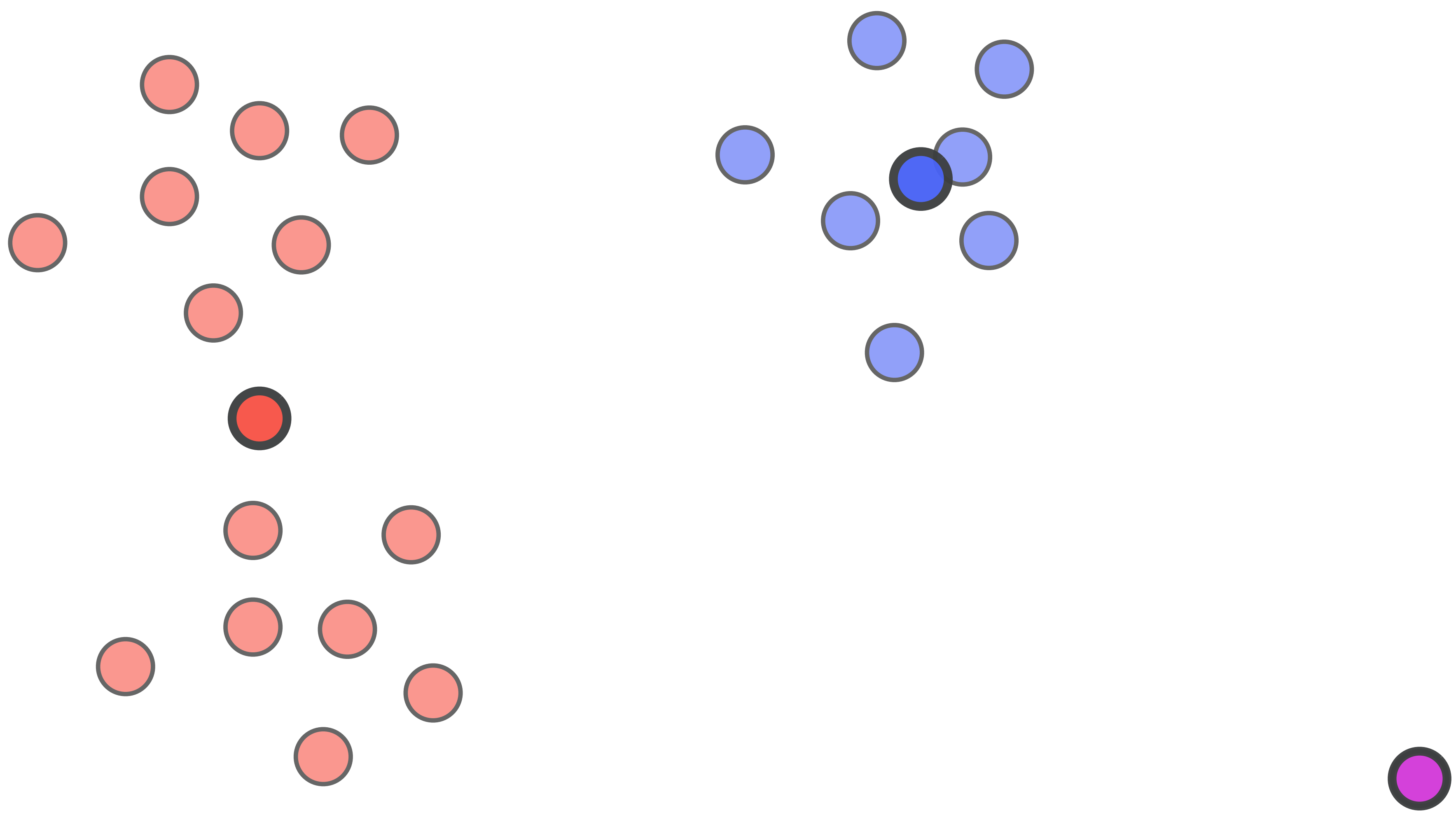
k-means algorithm: initialization with further-first traversal



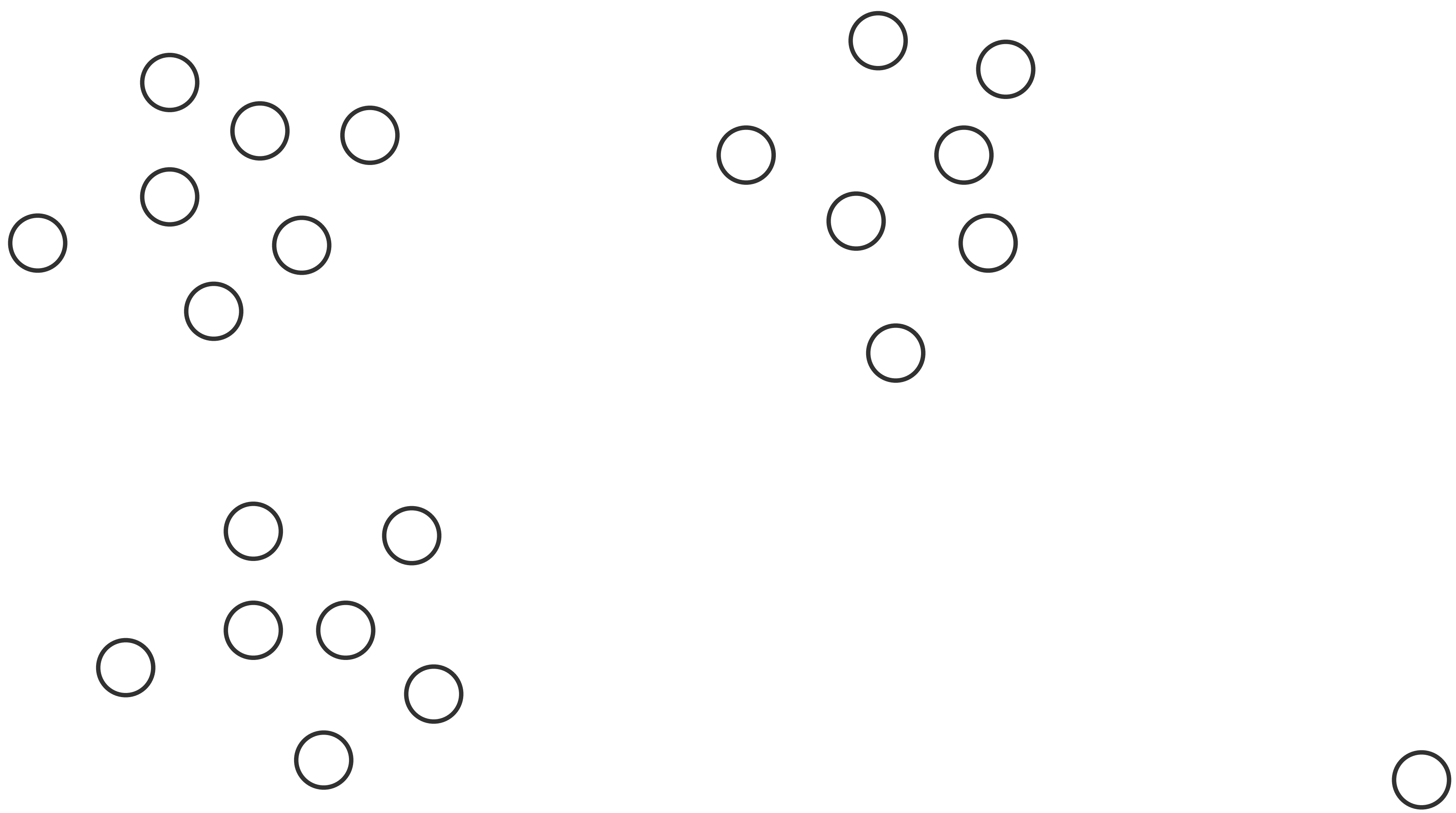
but... sensitive to outliers



but... sensitive to outliers



Here random may work well



k-means++ algorithm

- interpolate between the two methods
- let $D(x)$ be the distance between x and the nearest center selected so far
- choose next center with probability proportional to

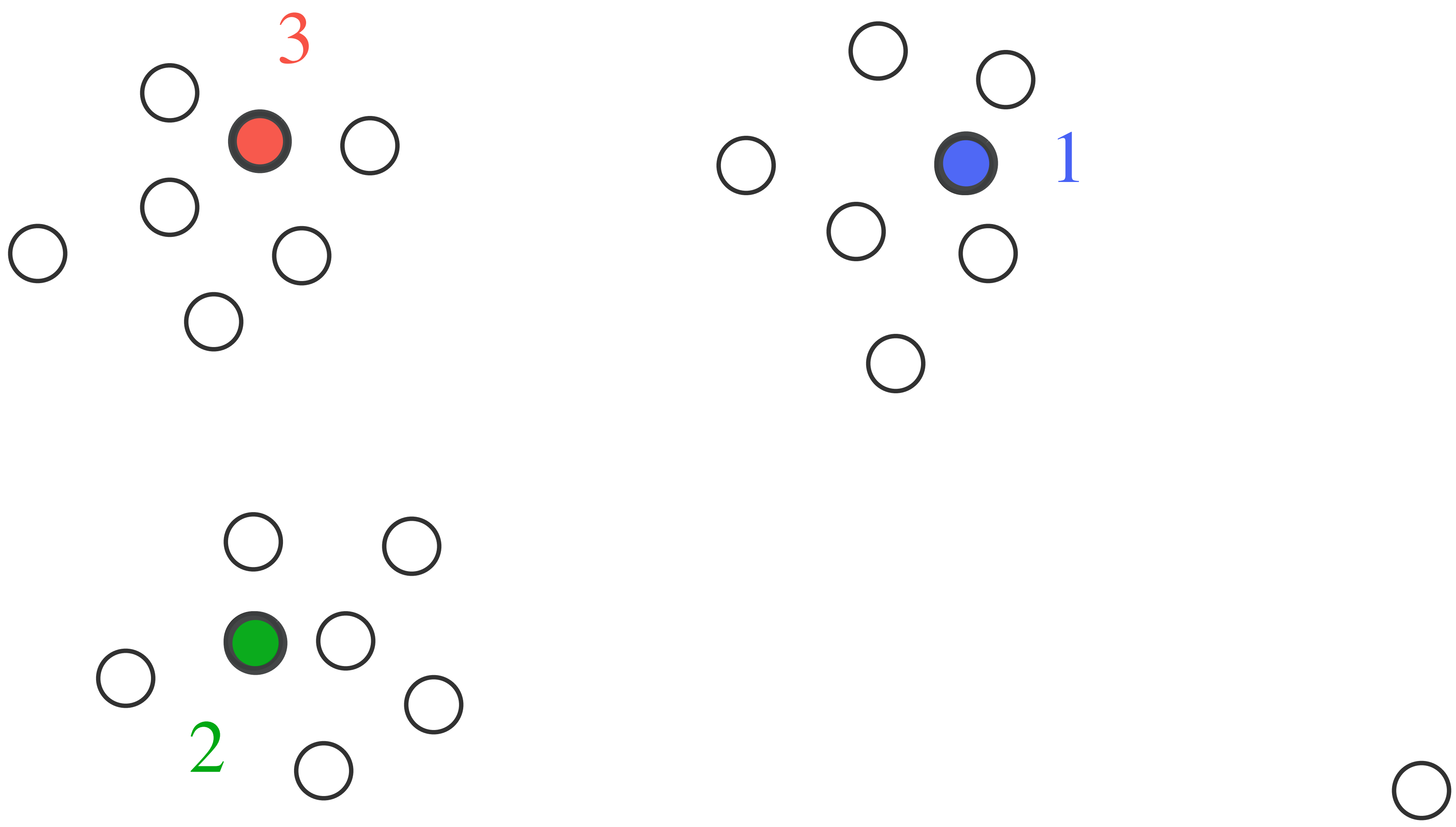
$$(D(x))^a = D^a(x)$$

- $a = 0$ random initialization
- $a = \infty$ furthest-first traversal
- $a = 2$ k-means++

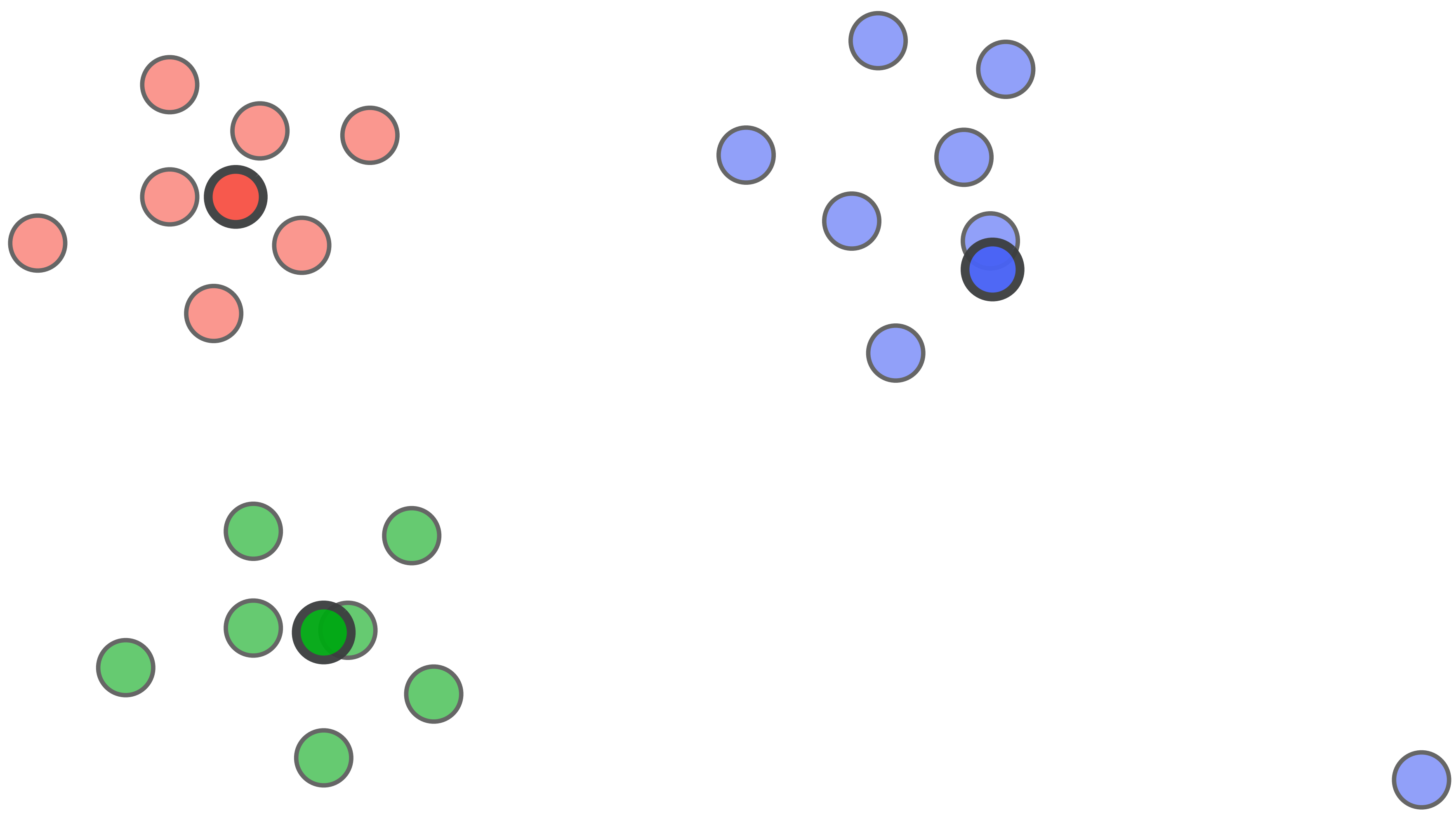
k-means++ algorithm

- initialization phase:
  - choose the first center uniformly at random
  - choose next center with probability proportional to $D^2(x)$
- iteration phase:
  - iterate as in the k-means algorithm until convergence

k-means++ initialization



k-means++ result



k-means++ provable guarantee

- approximation guarantee comes just from the first iteration (initialization)
- subsequent iterations can only improve cost

Lesson learned

- no reason to use k-means and not k-means++

- k-means++ :
  - easy to implement
  - provable guarantee
  - works well in practice