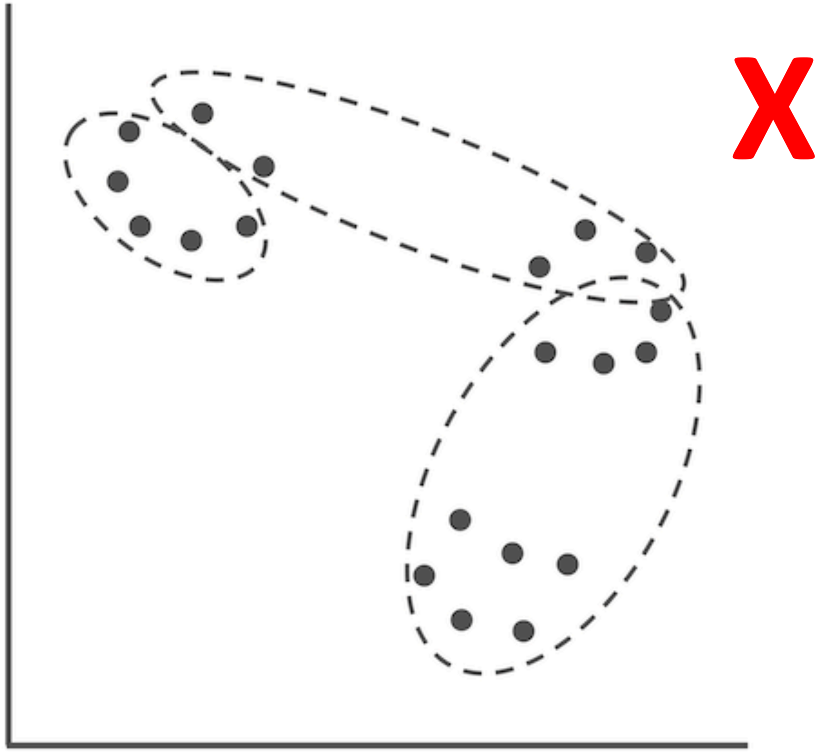
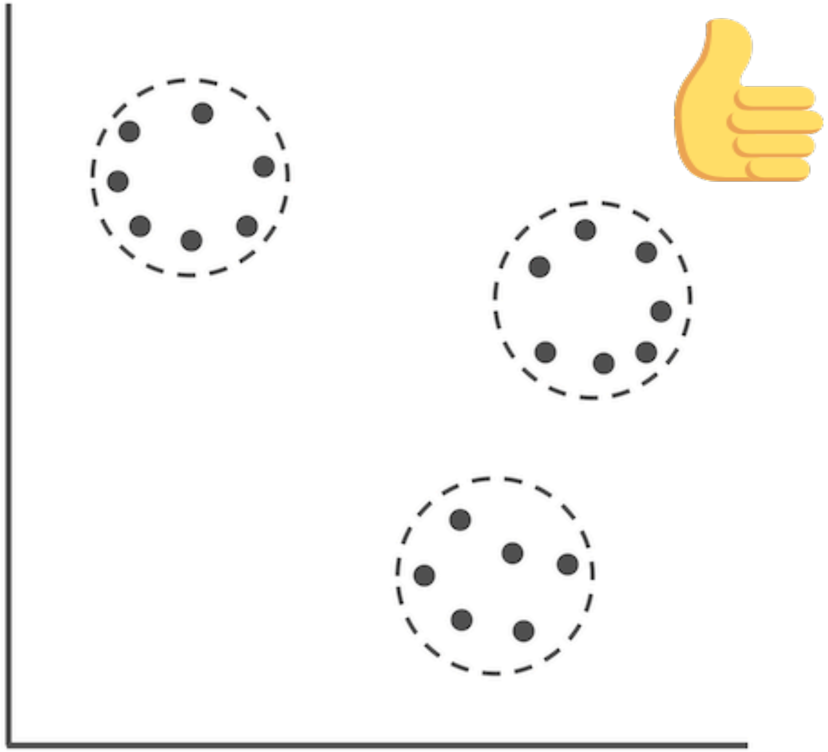
CMSC 423:
Data Clustering

Part 2

The Good Clustering Principle

- Homogeneity: All points in the cluster must be similar
- Separation: Points in different clusters are dissimilar





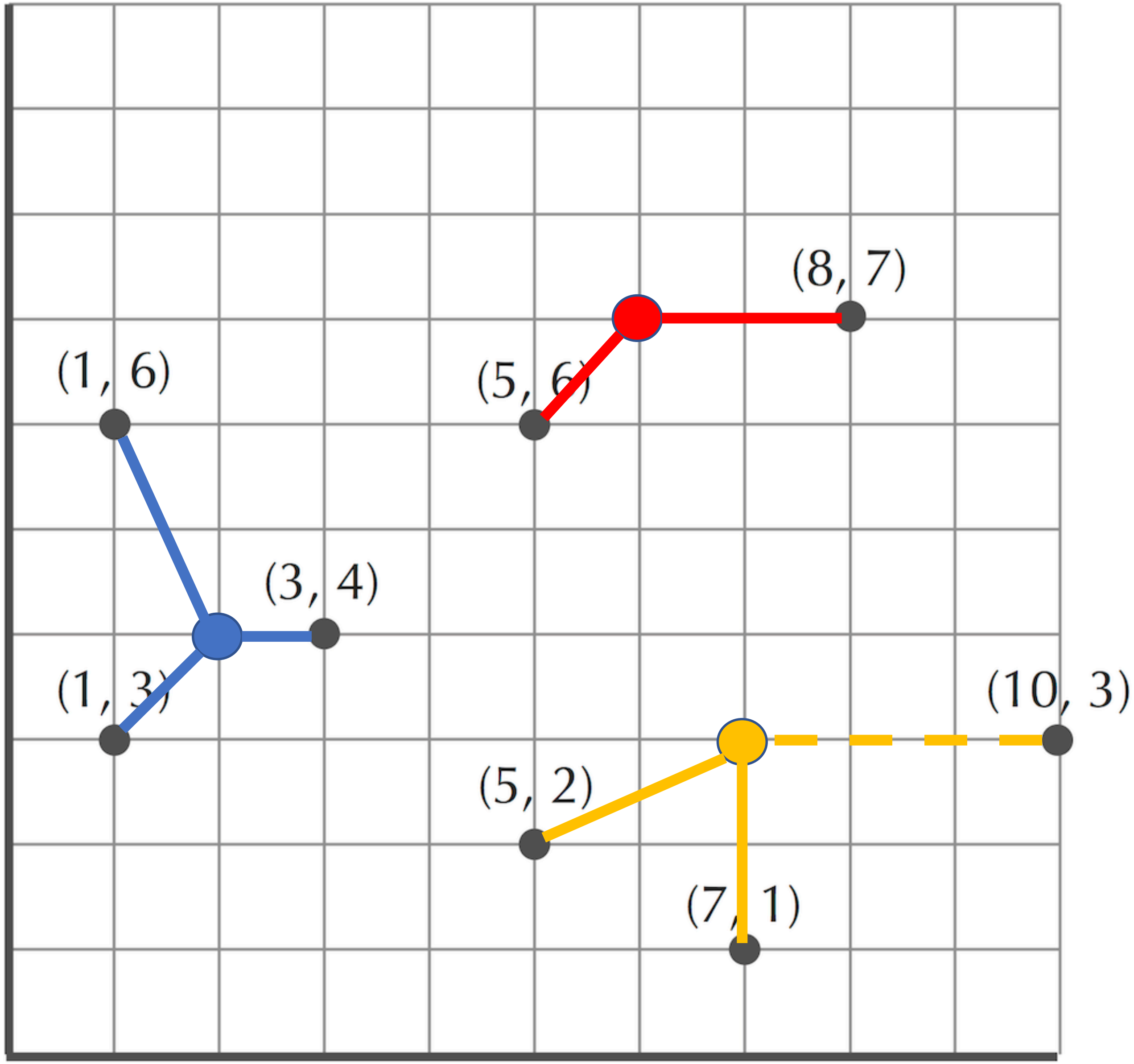
k-Center Clustering

- Pick *k* centers
- For each point, select the nearest center
- Find the set of *k* centers that minimizes the maximum distance between any point and its nearest center

Centers (2, 4), (6, 7), and (7, 3)

Euclidean distance

$$d(v, w) = \sqrt{\sum_{i=1}^{m} (v_i - w_i)^2}.$$



Properties of distance

- Distance used in previous example : Euclidean distance
- It is a metric– satisfies the triangle inequality theorem
- This property helps prove 2-approximation

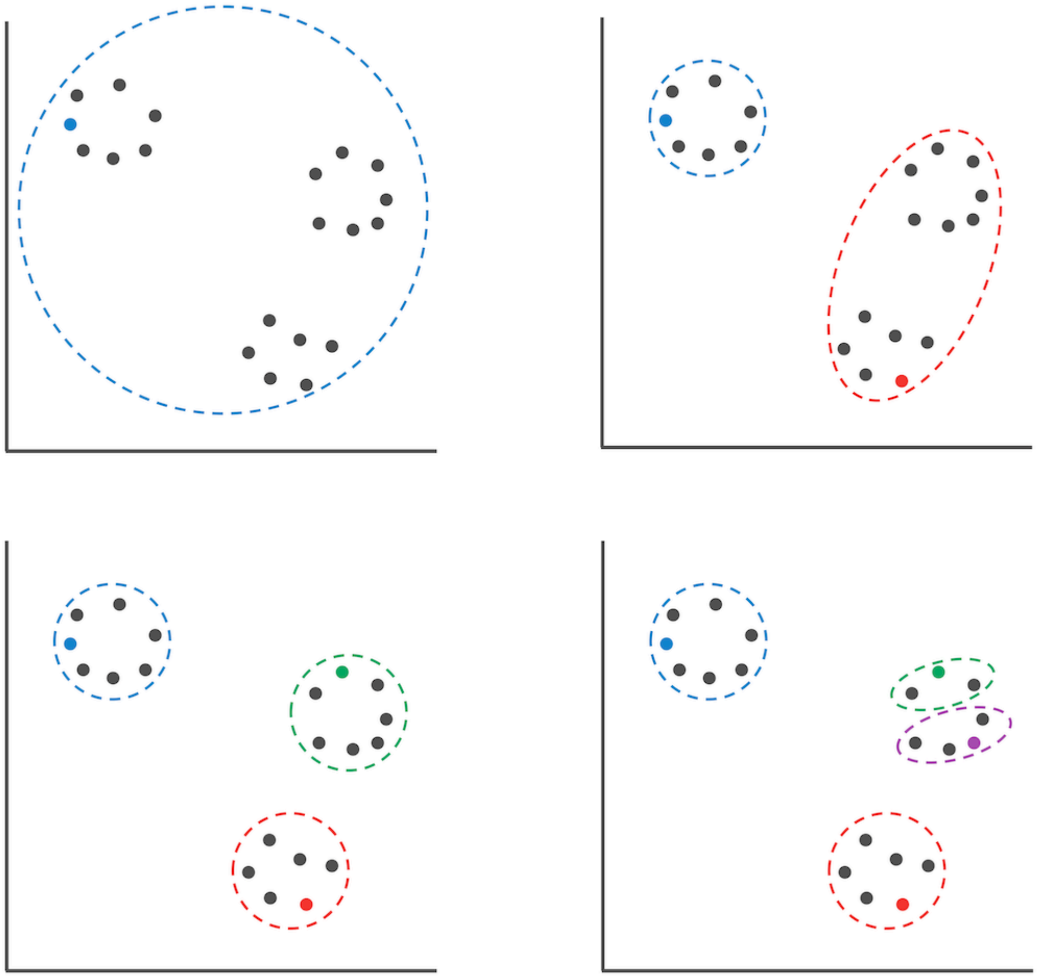
k-Center Clustering

- Pick *k* centers
- For each point, select the nearest center
- Find the set of *k* centers that minimizes the maximum distance between any point and its nearest center

- How many centers can there be?
- For k=1, how do you pick the center?

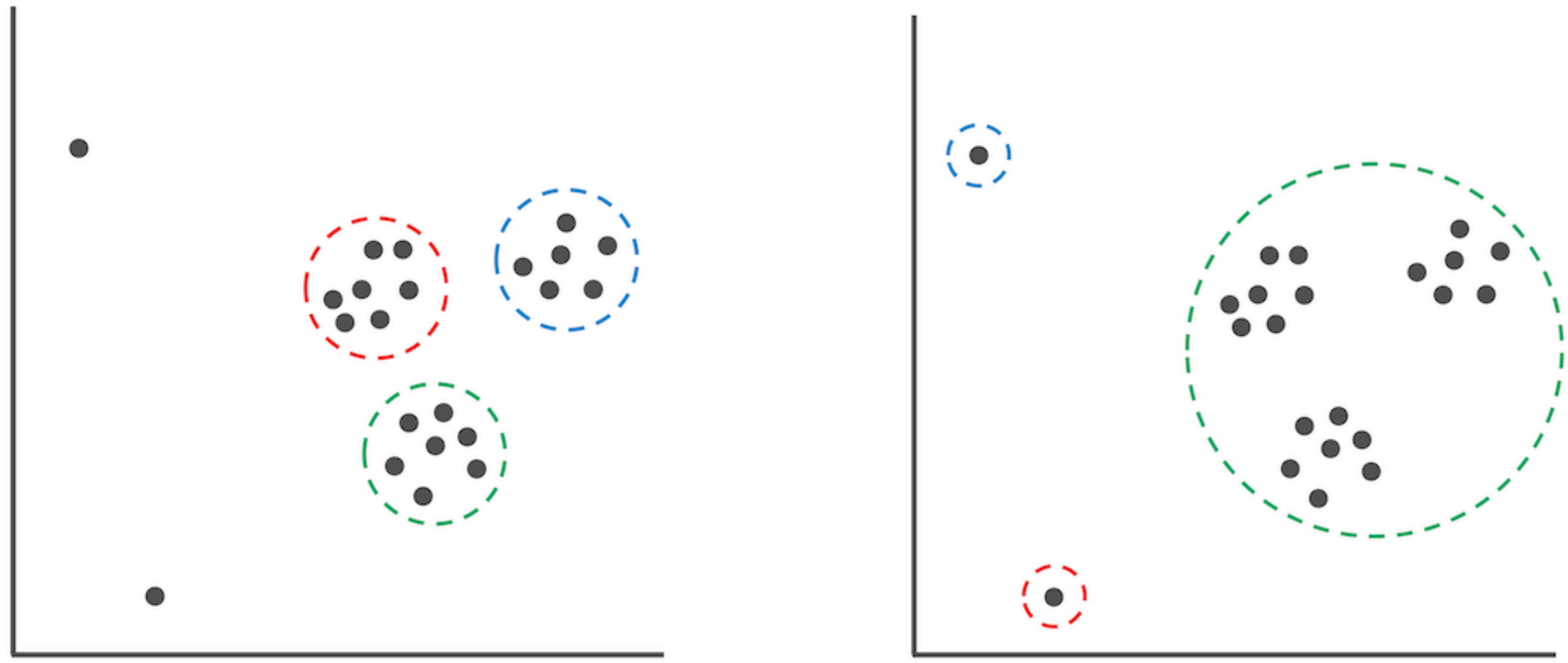
Farthest First Travel Heuristic

- Arbitrarily pick a point as the first center
- Second center is the point farthest from the first center
- Repeat until *k* centers are found



Note: Farthest distance works with any metric distance (not just Euclidean)

Is there an alternative scoring function that is more biologically appropriate?

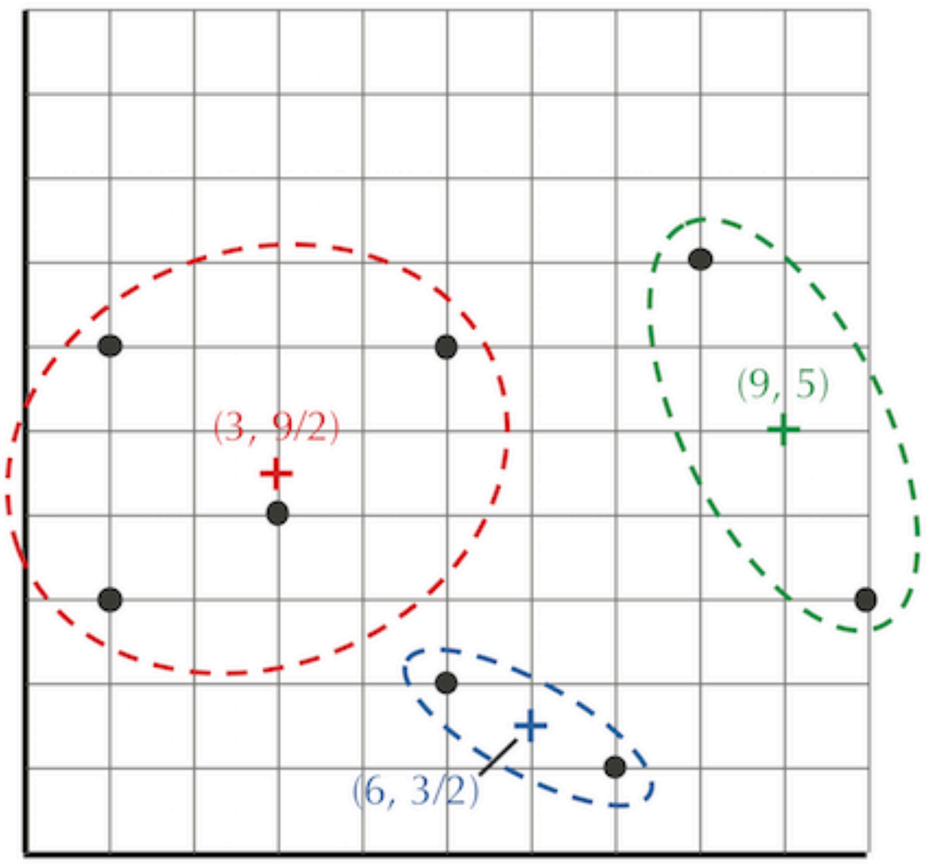


k-Means Clustering

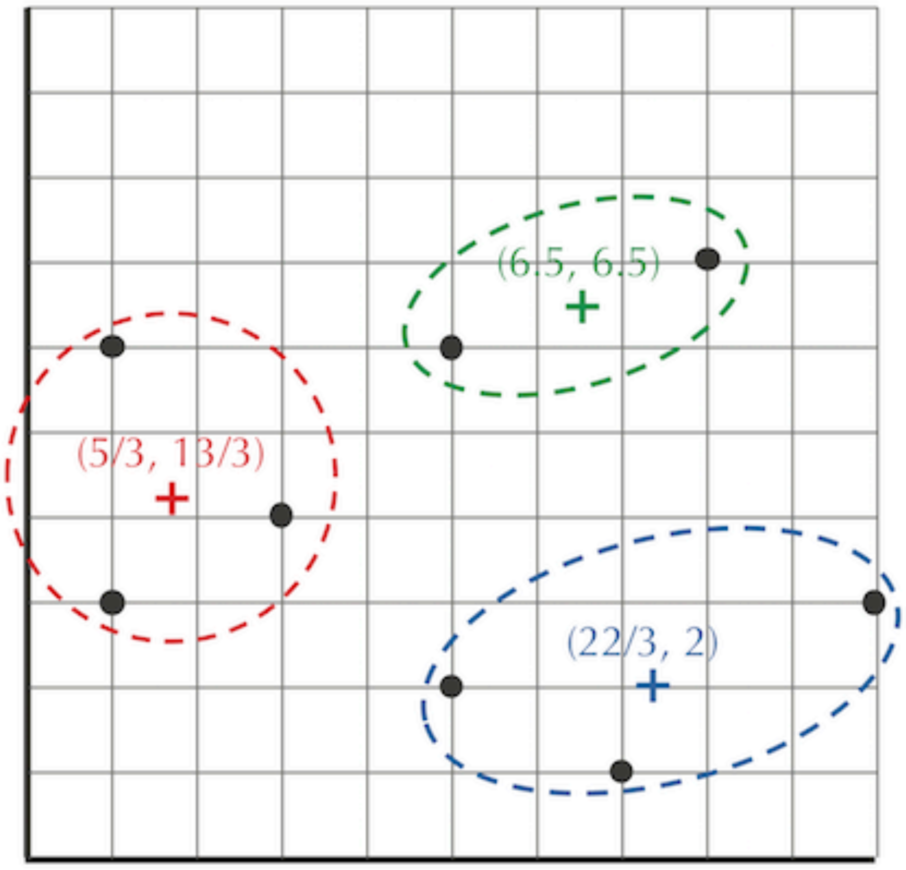
- Instead of min-max, use squared error distortion
- Squared error distortion- the average distance from points to the corresponding centers

$$Distortion(Data, Centers) = (1/n) \sum_{all\ points\ DataPoint\ in\ Data} d(DataPoint,\ Centers)^2\ .$$

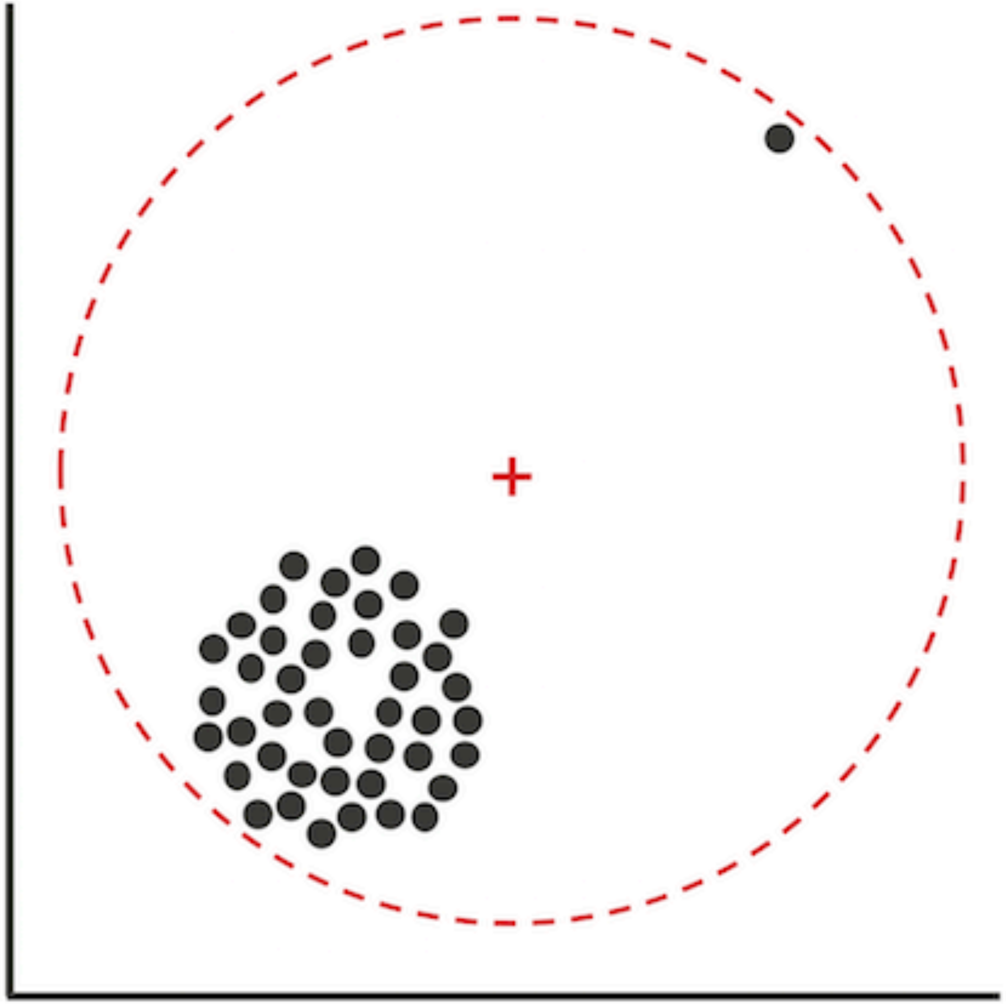
k-Centers Clustering



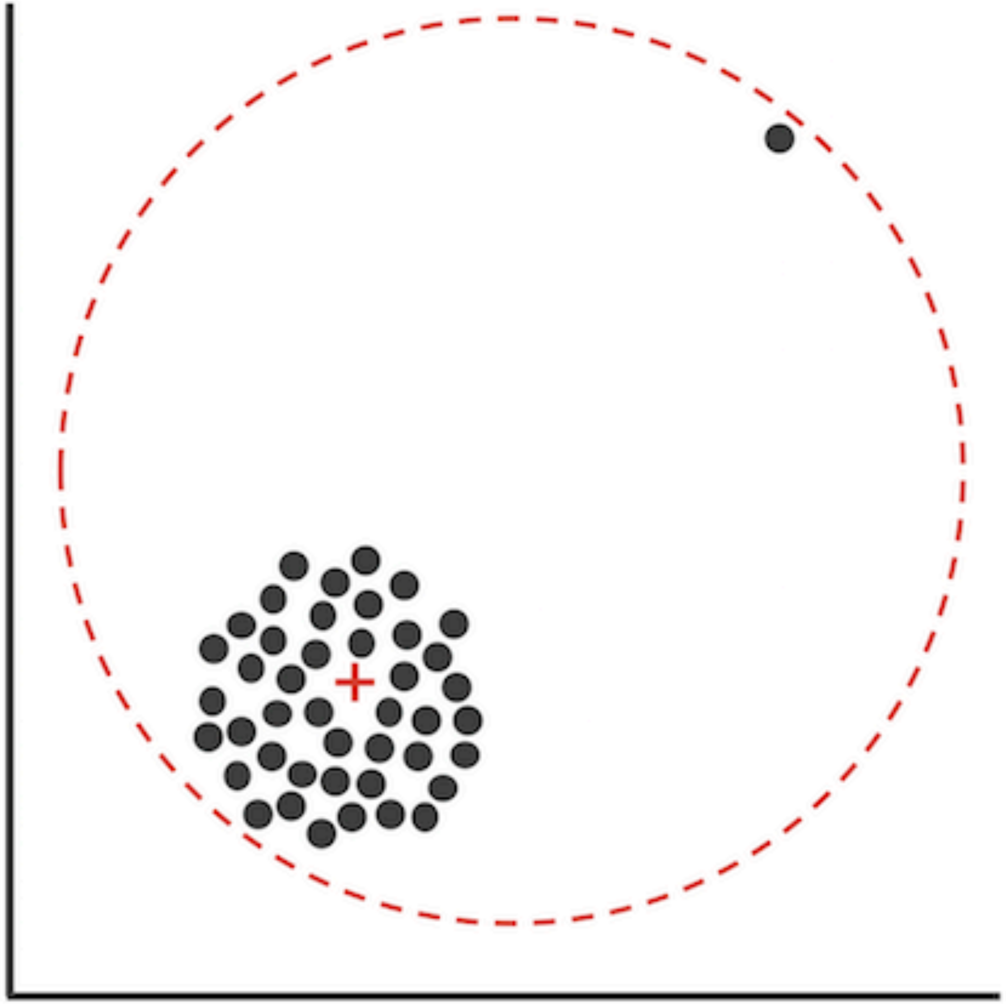
k-Means Clustering



k-Centers Clustering



k-Means Clustering



For k=1, how do you pick the center?

- **Center of gravity** is the point whose *i*-th coordinate is the average of the *i*-th coordinates of all data points
- For example, the center of gravity of the points (3, 8), (8, 0), and (7, 4) is

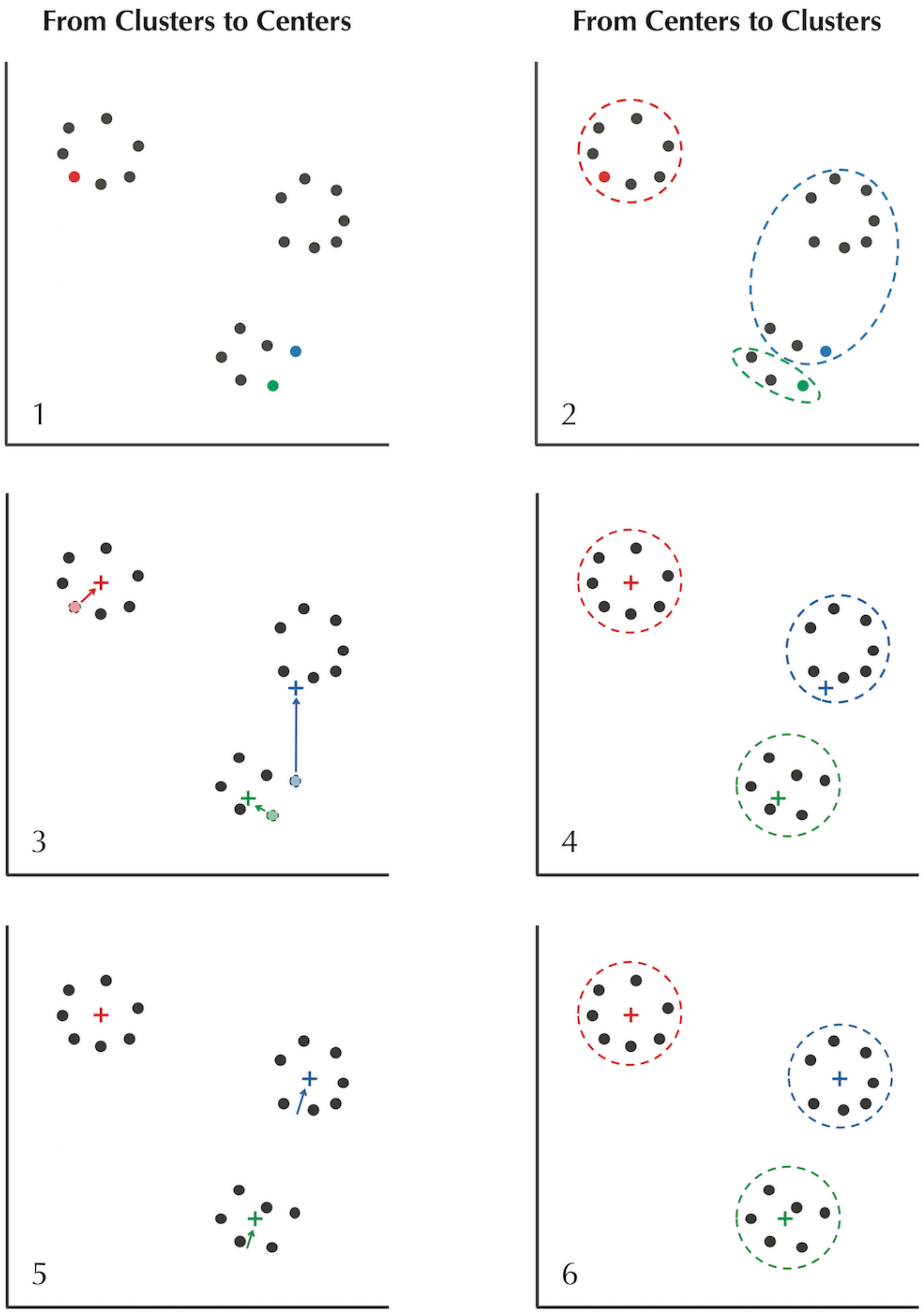
$$\left(\frac{3+8+7}{3}, \frac{8+0+4}{3}\right) = (6, 4).$$

- **Center of Gravity Theorem:** The center of gravity of a set of points Data is the unique point solving the *k*-Means Clustering Problem for *k* = 1

k-Means Clustering- Lloyd Algorithm

- Goal: split data into exactly *k* clusters
- Basic algorithm:
  - Create *k* arbitrary clusters- pick *k* points as cluster centers and assign each other point to the closest center
  - Re-compute the center of each cluster
  - Re-assign points to clusters
  - Repeat
- The algorithm has **converged** if the centers (and clusters) stop changing between iterations

k-Means Clustering- Lloyd Algorithm



Visualizing k-means clustering

- https://www.naftaliharris.com/blog/visualizing-k-means-clustering/