Related Topic: Clustering

Input: n **sites:** $S = \{s_1, s_2, ..., s_n\}$

Output: **Locations of** k **centers:** $C = \{c_1, c_2, ..., c_k\}$

Objective: **Minimize the total distance from each site to the nearest center.**

2

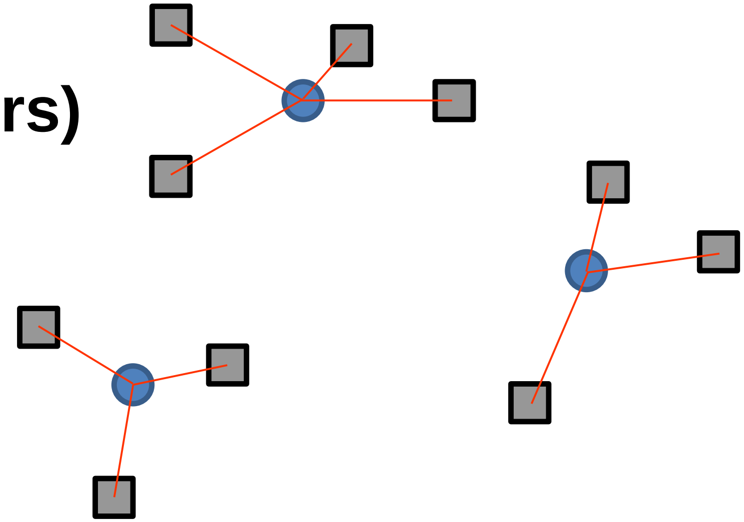
$$\text{Minimize} \sum_{s \in S} dist(s, C)^2$$

$dist(s, C)$: Distance from s to the nearest center.

$$dist(s, C) = \operatorname*{Min}_{c \in C} \{dist(s, c)\}$$

Reformulation

■ **site (*n* sites)**
● **center (*k* centers)**



(1) Divide the *n* sites into *k* clusters based on the nearest center.

$$S = \{s_1, s_2, \dots, s_n\}$$

$$\Downarrow$$

$$S = S_1 \cup S_2 \cup \cdots \cup S_k$$

(2) Reformulate the objective function as follows:

$$Minimize \sum_{j=1}^{k} \sum_{s \in S_j} dist(s - c_j)^2$$

Related Topic: Clustering

$$Minimize \sum_{j=1}^{k} \sum_{s \in S_j} dist(s - c_j)^2$$

$$S = \{s_1, s_2, \dots, s_n\}$$

$$\Downarrow$$

$$S = S_1 \cup S_2 \cup \cdots \cup S_k$$

s and c_j: Points in the 2D space.

k-means Algorithm: Iterate the following two steps from a random partition of *S* into *k* subsets: S_1, S_2, …, S_k

(i) $c_j = \dfrac{1}{|S_j|} \sum_{s \in S_j} s$

(ii) $S_j = \{s \,|\, dist\left(s, c_j\right) = \min_{l=1,2,\dots,k} dist(s, c_l)\}, \quad j = 1, 2, \dots, k$

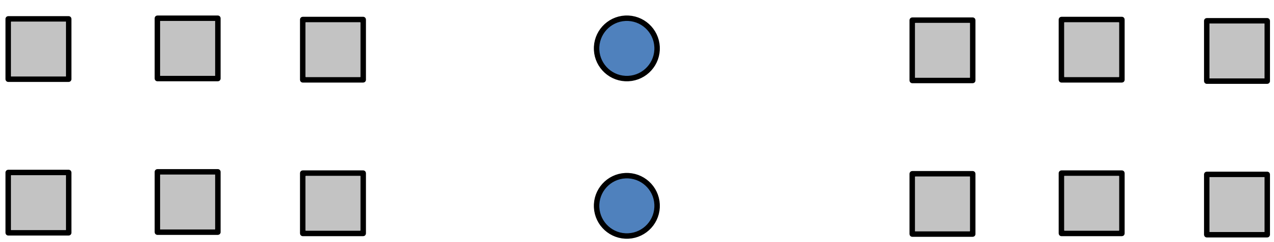
<u>Exercise 4-1</u>:

In the k-means algorithm, we can start with (i) using an initial partition $\{S_1, S_2, \dots, S_k\}$ or with (ii) using initial centers $\{c_1, c_2, \dots, c_k\}$. Design a good initialization method for k-means algorithm for (i) and also for (ii).

Example (12 sites and 2 centers)

If we start with the following partitions:

If we start with the following centers:

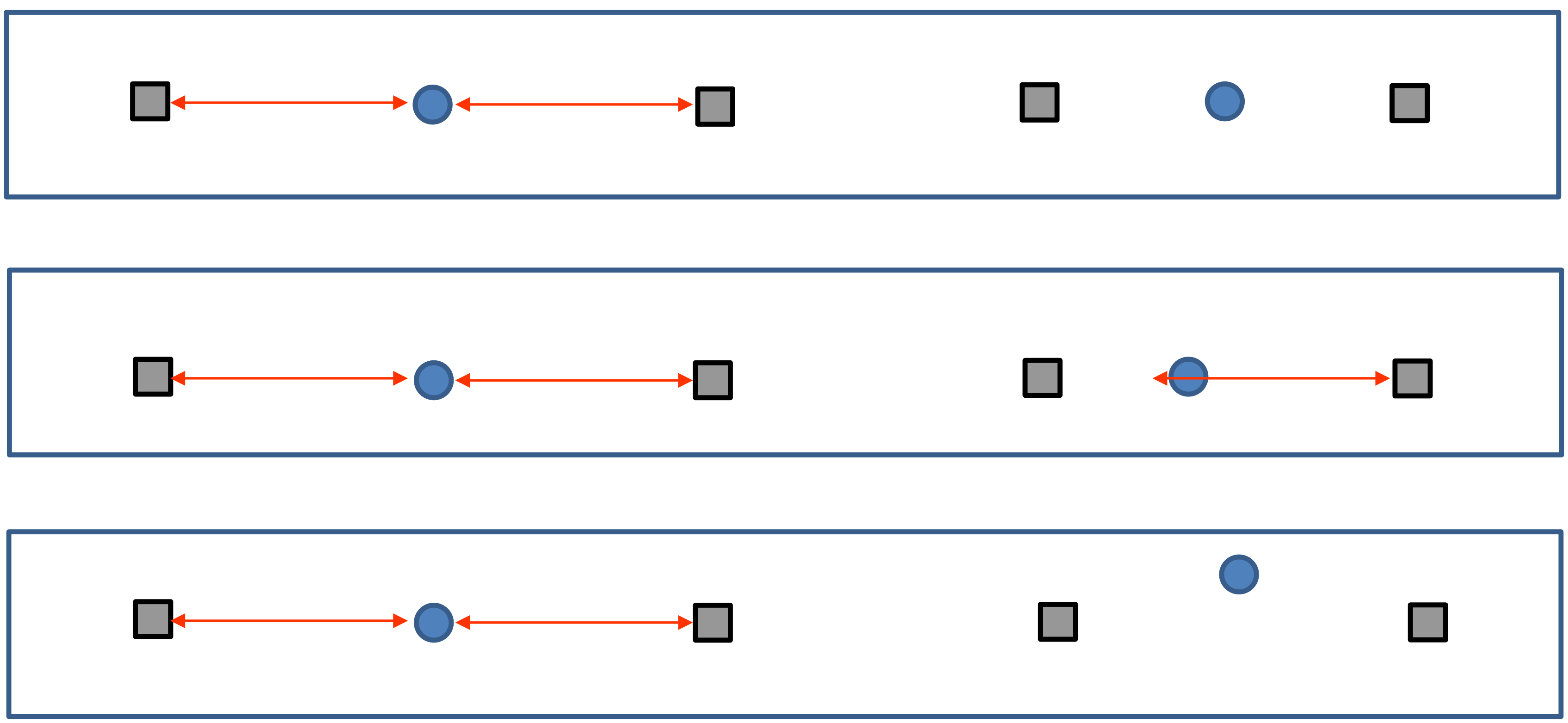


Difficulty of "Min-Max" objective function:
("minimize the worst case" objective function)

Minimization of the maximum distance from each site to the nearest center.

$$\text{Minimize } \underset{s \in S}{\text{Max}}\ dist(s, C)$$

All the following solutions are optimal (for *k* = 2).



Comparison of Problems:

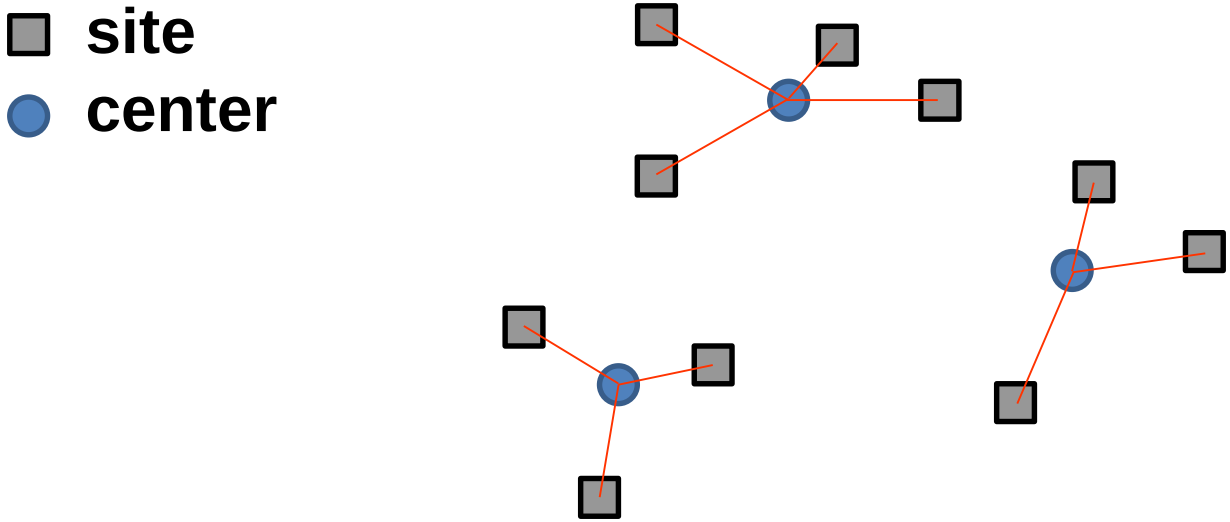
(1) Minimization of the maximum distance from each site to the nearest center.

$$\text{Minimize } \max_{s \in S} dist(s, C)$$

(2) Minimization of the total distance from each site to the nearest center

$$\text{Minimize } \sum_{s \in S} dist(s, C)^2$$

Q. Which is a better problem formulation?

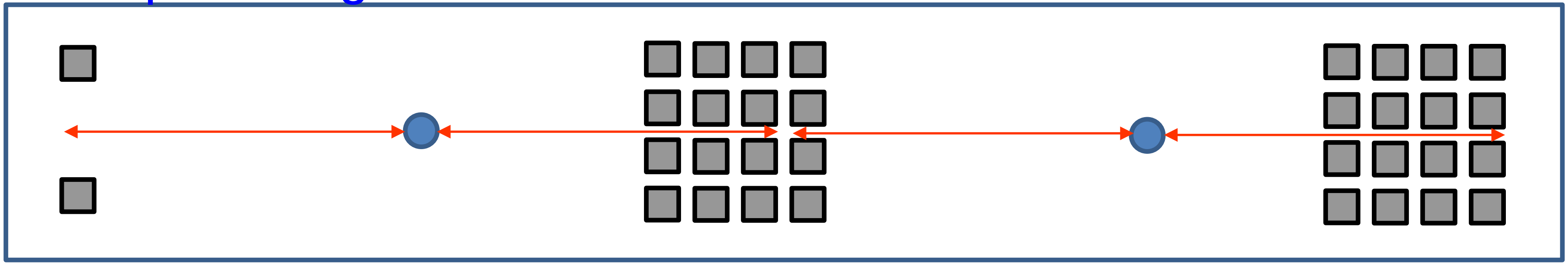


Comparison of Problems:

(1) Minimization of the maximum distance from each site to the nearest center.

$$\text{Minimize } \max_{s \in S} dist(s, C)$$

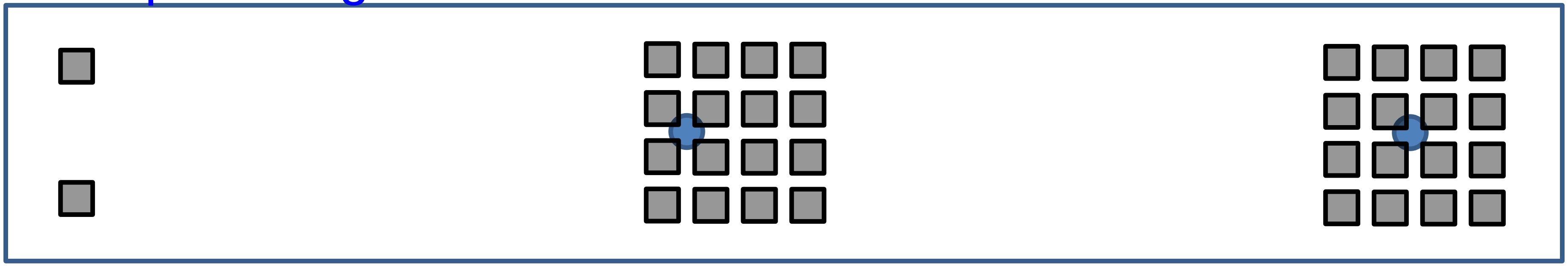
Example of a good solution



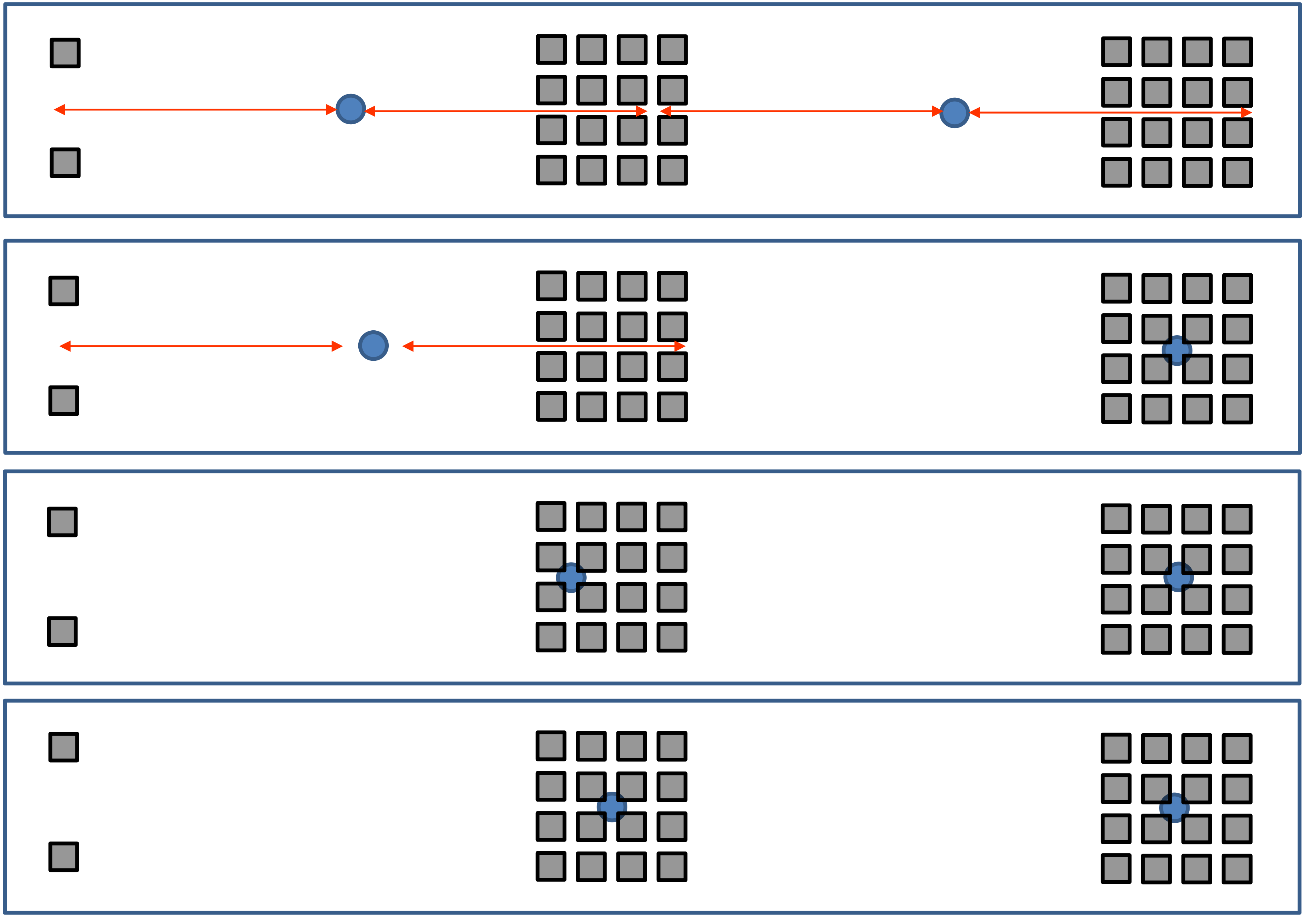
(2) **Minimization of the total distance** from each site to the nearest center

$$\text{Minimize } \sum_{s \in S} dist(s, C)^2$$

Example of a good solution



Which is the best solution?



Comparison of Algorithms:

(1) Minimization of the maximum distance from each site to the nearest center.

$$\text{Minimize } \underset{s \in S}{\text{Max}}\, dist(s, C)$$

Center Selection Algorithm:

Simple heuristics (a greedy algorithm)

2-Approximation algorithm

(2) Minimization of the total distance from each site to the nearest center

$$\text{Minimize } \sum_{s \in S} dist(s, C)^2$$

K-means Algorithm

Iterative adjustment algorithm

(iterations of two greedy algorithms)

Not an exact optimization algorithm