CSCE 689: Special Topics in Modern Algorithms for Data Science

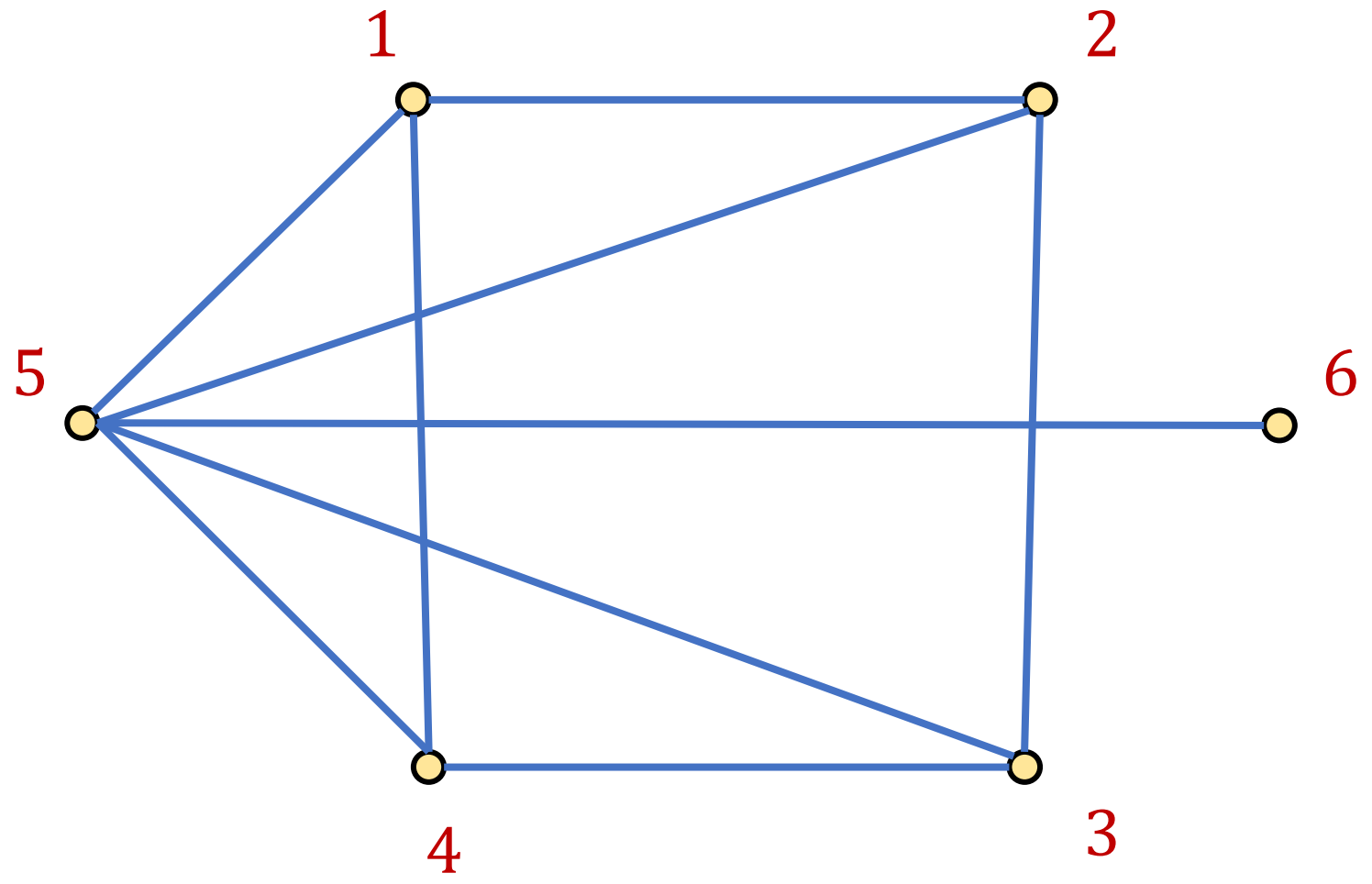
Lecture 22

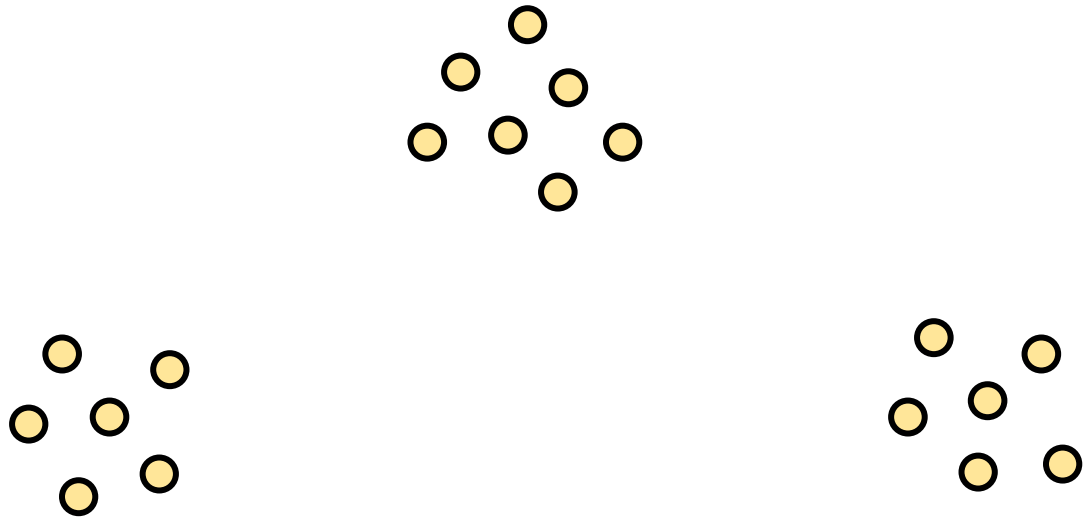
Samson Zhou

Presentation Schedule

- November 27: Chunkai, Jung, Galaxy AI
- November 29: STMI, Anmol, Jason
- December 1: Bokun, Ayesha, Dawei, Lipai

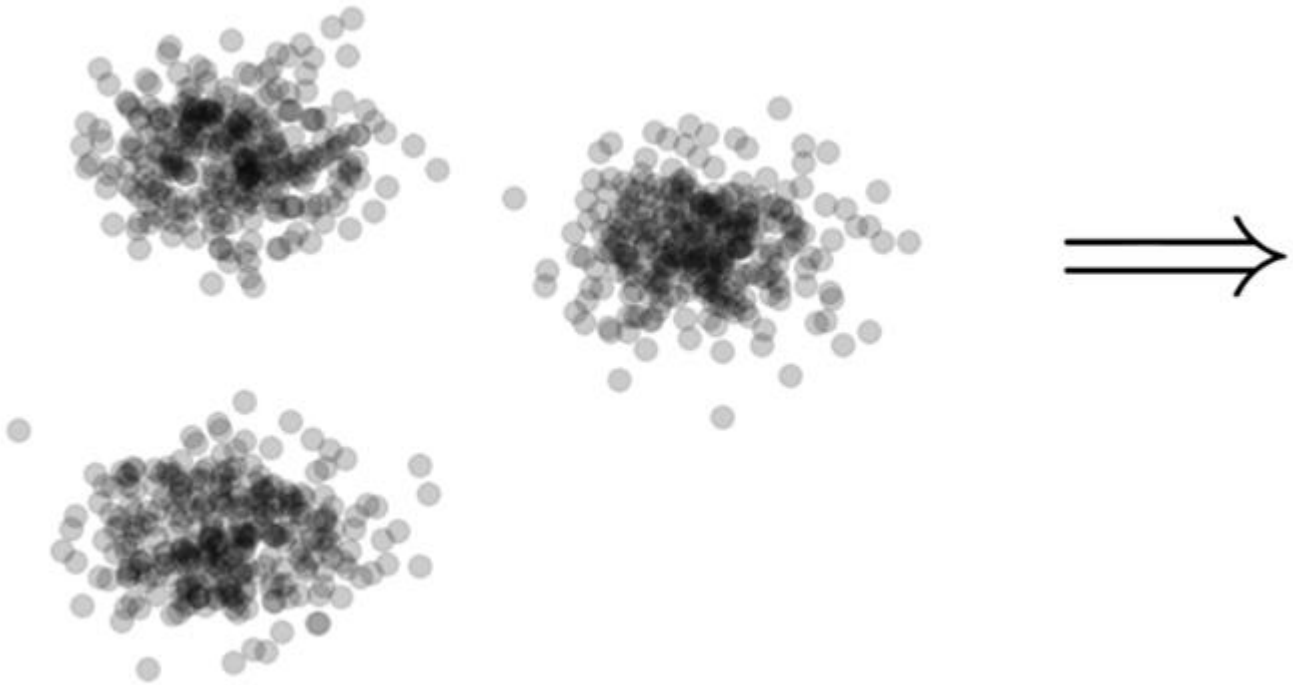
47 72 81 10 14 33 51 29 54 9 36 46 10


1
2
5
6
4
3



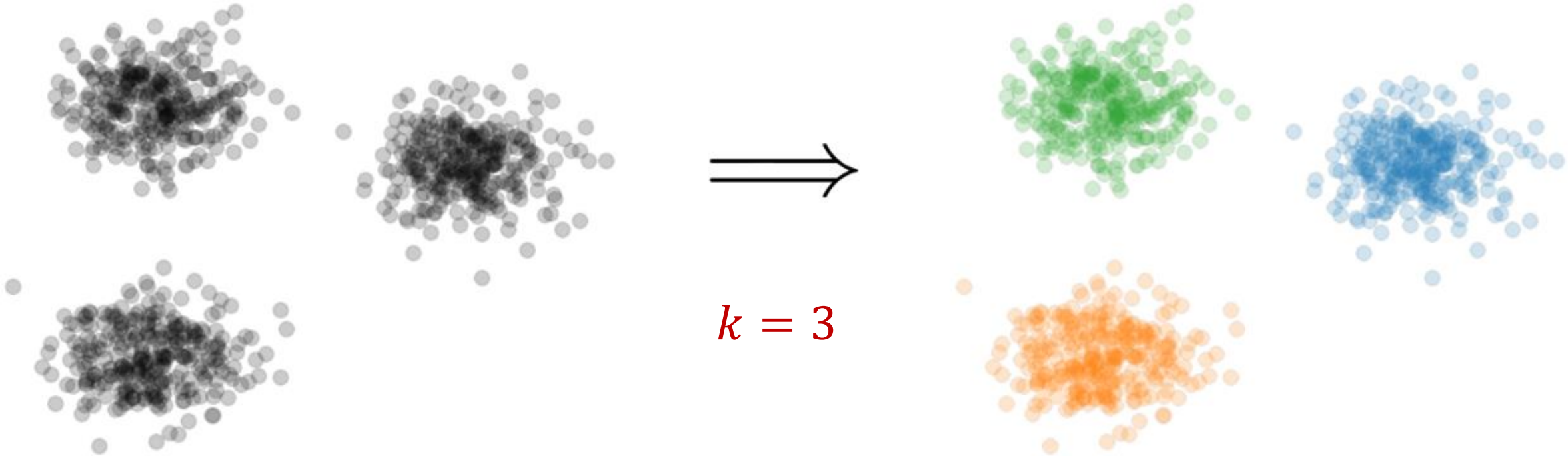
Clustering

- Goal: Given input dataset X, partition X so that “similar” points are in the same cluster and “different” points are in different clusters



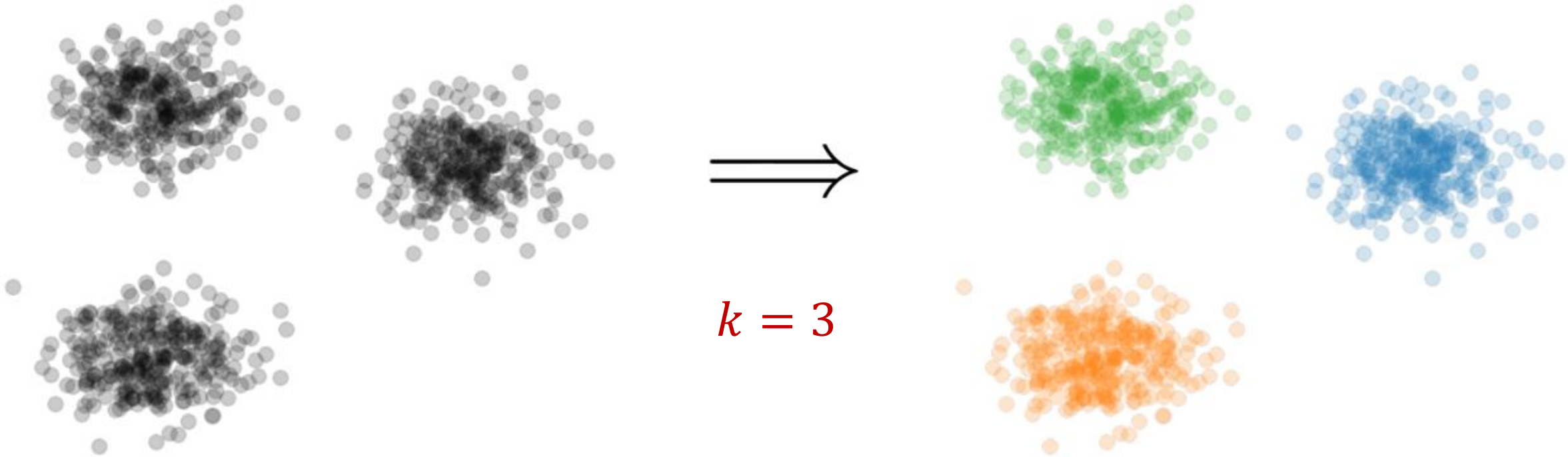
k-Clustering

- Goal: Given input dataset X, partition X so that "similar" points are in the same cluster and "different" points are in different clusters
- There can be at most k different clusters



k-Clustering

- Question: How do we measure the “quality” of each clustering?



k-Clustering

- Question: How do we measure the “quality” of each clustering?
- Assign a “center” c_i to each cluster
- Have a cost function induced by c_i for all of the points P_i assigned to cluster i

k-Clustering

- Question: How do we measure the "quality" of each clustering?
- Assign a "center" c_i to each cluster
- Have a cost function induced by c_i for all of the points P_i assigned to cluster i
  - Assume points are in metric space with distance function $\mathrm{dist}(\cdot,\cdot)$
  - Define $\mathrm{Cost}(P_i, c_i)$ to be a function of $\{\mathrm{dist}(x, c_i)\}_{x \in P_i}$

k-Clustering

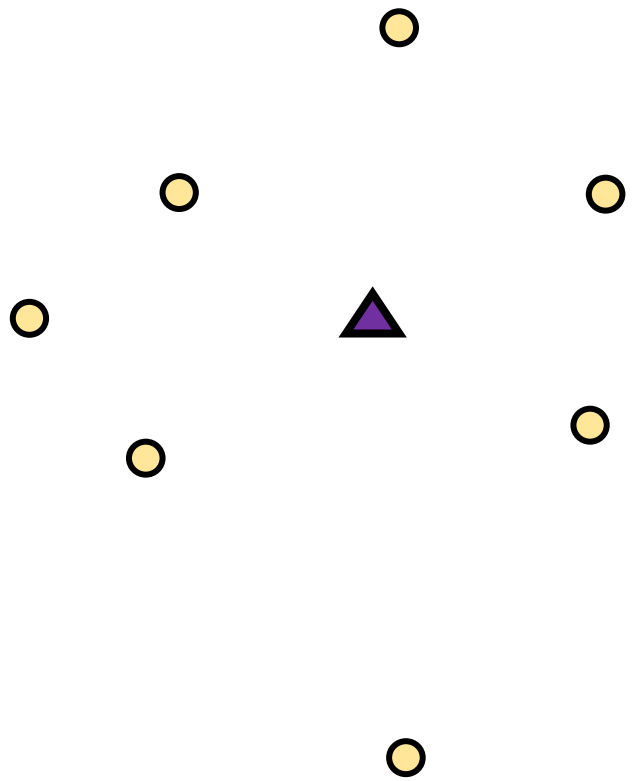
- Question: How do we measure the "quality" of each clustering?

- Have a cost function induced by c_i for all of the points P_i assigned to cluster i
  - Define $\mathrm{Cost}(P_i, c_i)$ to be a function of $\{\mathrm{dist}(x, c_i)\}_{x \in P_i}$

- Suppose the set of centers is $C = \{c_1, \ldots, c_k\}$
  - Define clustering cost $\mathrm{Cost}(X, C)$ to be a function of $\{\mathrm{dist}(x, C)\}_{x \in C}$

k-Clustering

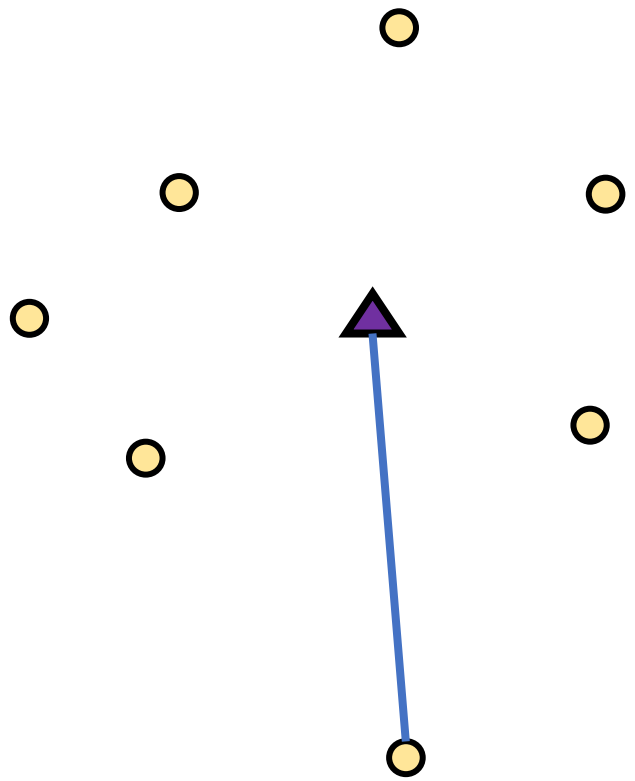
- Define clustering cost $\mathrm{Cost}(X, C)$ to be a function of $\{\mathrm{dist}(x, C)\}_{x \in C}$



k-Clustering

- Define clustering cost $\mathrm{Cost}(X, C)$ to be a function of $\{\mathrm{dist}(x, C)\}_{x \in C}$

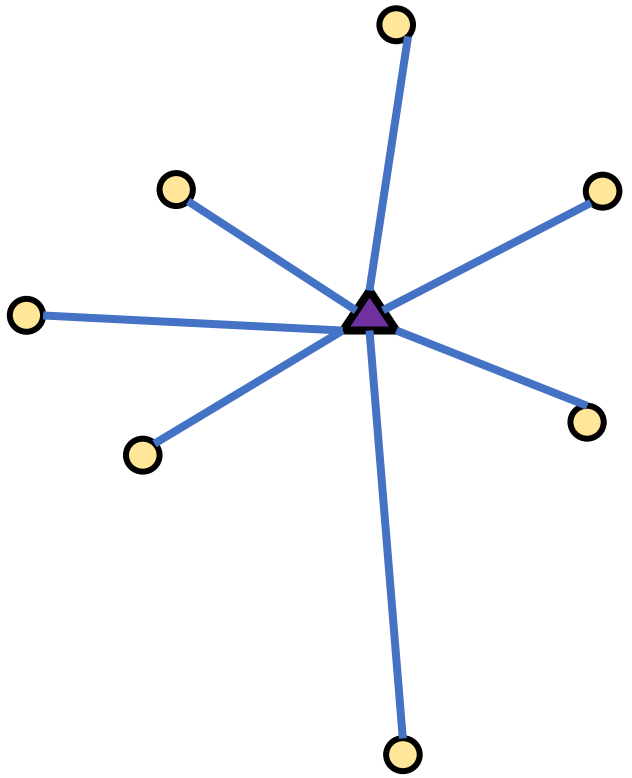
- k-center: $\mathrm{Cost}(X, C) = \max_{x \in X} \mathrm{dist}(x, C)$



k-Clustering

- Define clustering cost $\mathrm{Cost}(X, C)$ to be a function of $\{\mathrm{dist}(x, C)\}_{x \in C}$

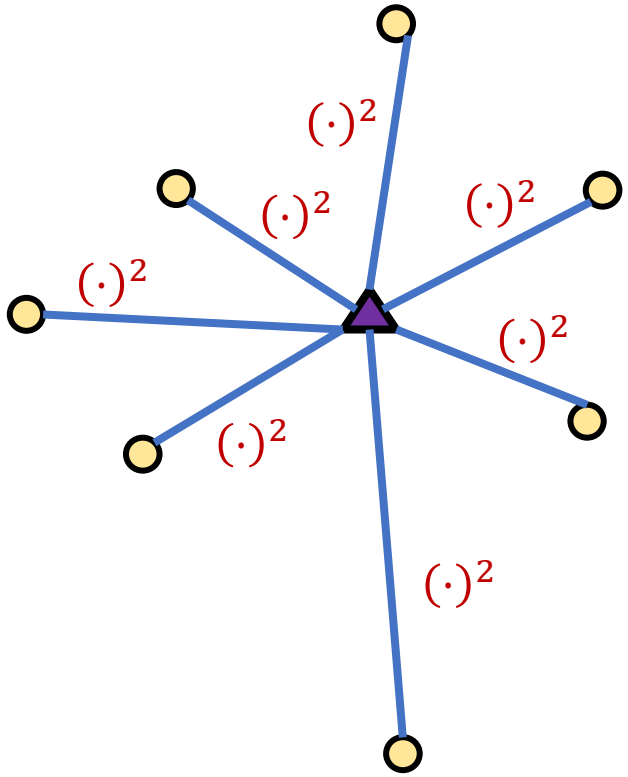
- k-center: $\mathrm{Cost}(X, C) = \max_{x \in X} \mathrm{dist}(x, C)$
- k-median: $\mathrm{Cost}(X, C) = \sum_{x \in X} \mathrm{dist}(x, C)$



k-Clustering

- Define clustering cost $\mathrm{Cost}(X, C)$ to be a function of $\{\mathrm{dist}(x, C)\}_{x \in C}$

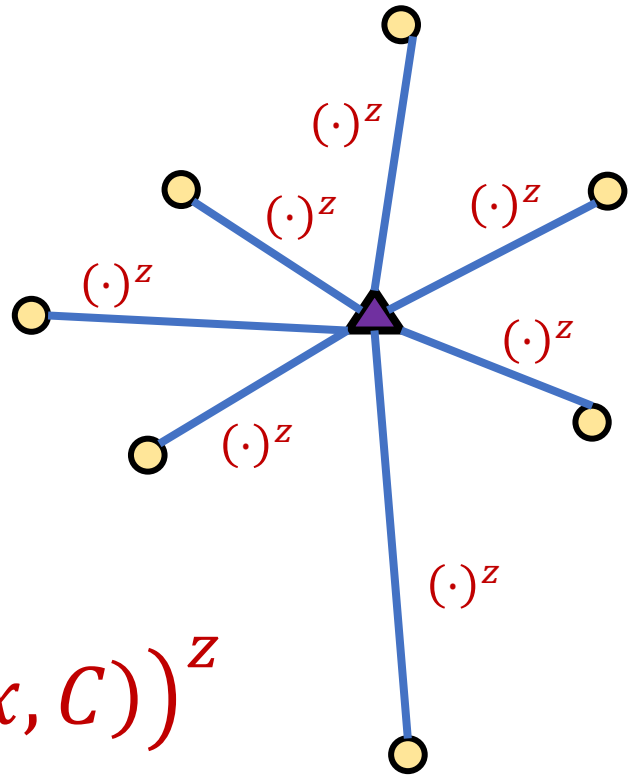
- k-center: $\mathrm{Cost}(X, C) = \max_{x \in X} \mathrm{dist}(x, C)$
- k-median: $\mathrm{Cost}(X, C) = \sum_{x \in X} \mathrm{dist}(x, C)$
- k-means: $\mathrm{Cost}(X, C) = \sum_{x \in X} \big(\mathrm{dist}(x, C)\big)^2$



k-Clustering

- Define clustering cost $\mathrm{Cost}(X, C)$ to be a function of $\{\mathrm{dist}(x, C)\}_{x \in C}$

- k-center: $\mathrm{Cost}(X, C) = \max_{x \in X} \mathrm{dist}(x, C)$
- k-median: $\mathrm{Cost}(X, C) = \sum_{x \in X} \mathrm{dist}(x, C)$
- k-means: $\mathrm{Cost}(X, C) = \sum_{x \in X} \big(\mathrm{dist}(x, C)\big)^2$
- (k, z)-clustering: $\mathrm{Cost}(X, C) = \sum_{x \in X} \big(\mathrm{dist}(x, C)\big)^z$



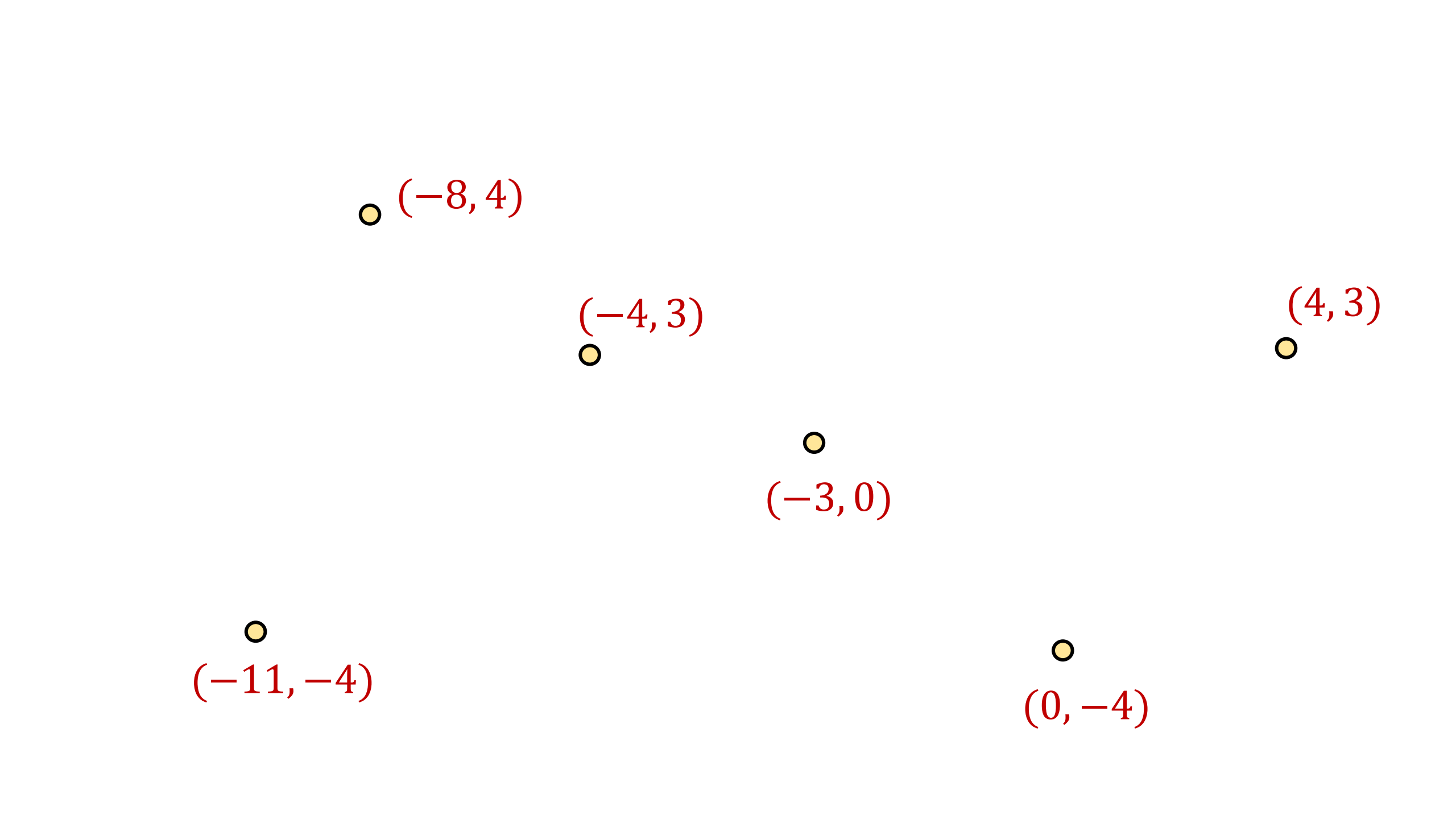
Euclidean k-Clustering

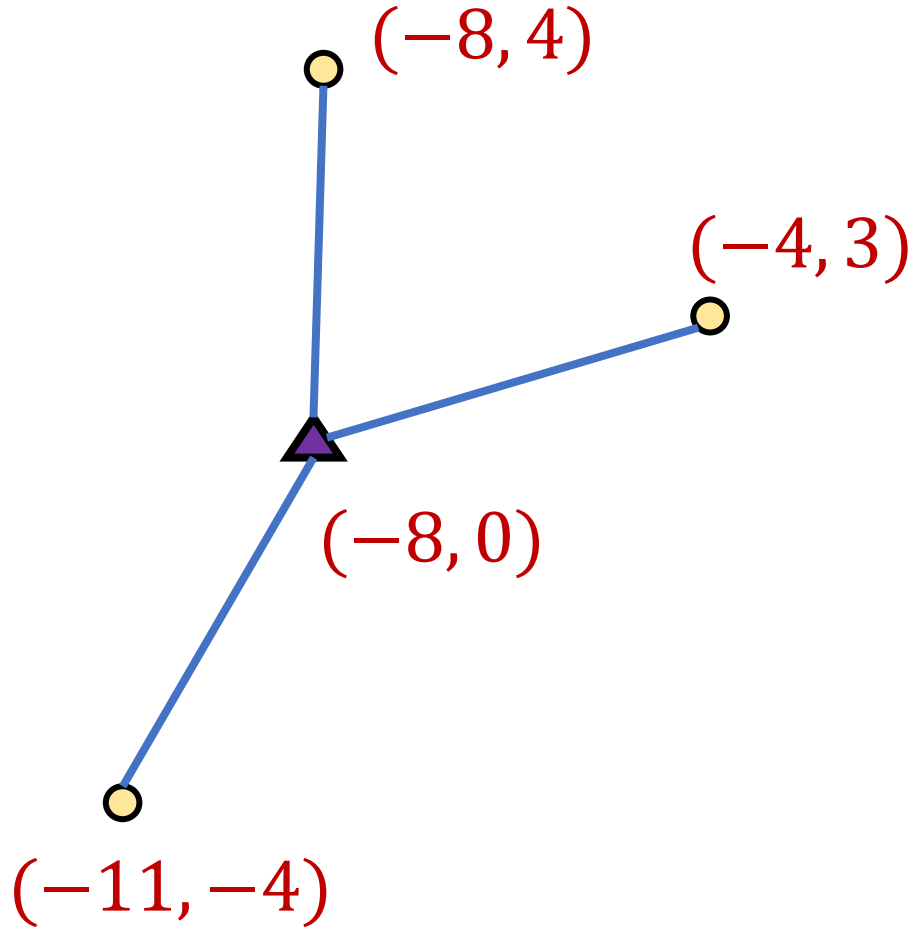
- For Euclidean k-clustering, input points $X = x_1, \dots, x_n$ are in $\mathbb{R}^d$ (for us, they will be in $[\Delta]^d := \{1,2, \dots, \Delta\}^d$)

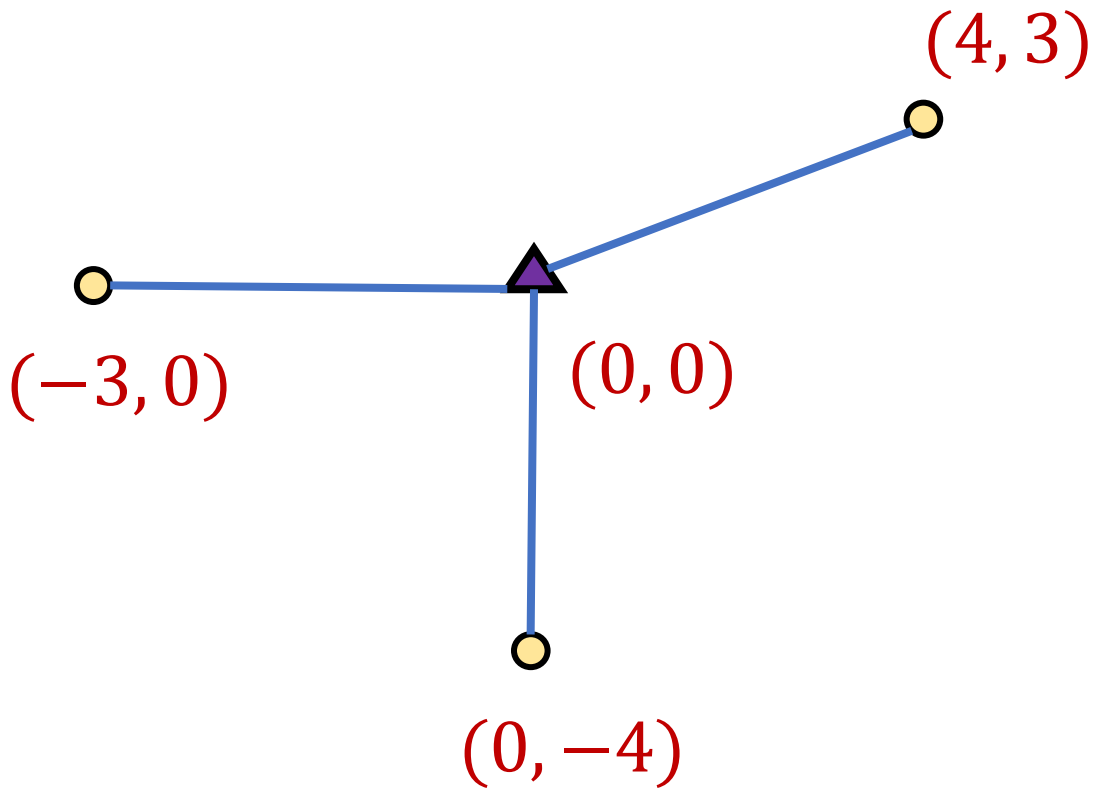
- $\mathrm{dist}(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_d - y_d)^2}$ is the Euclidean distance

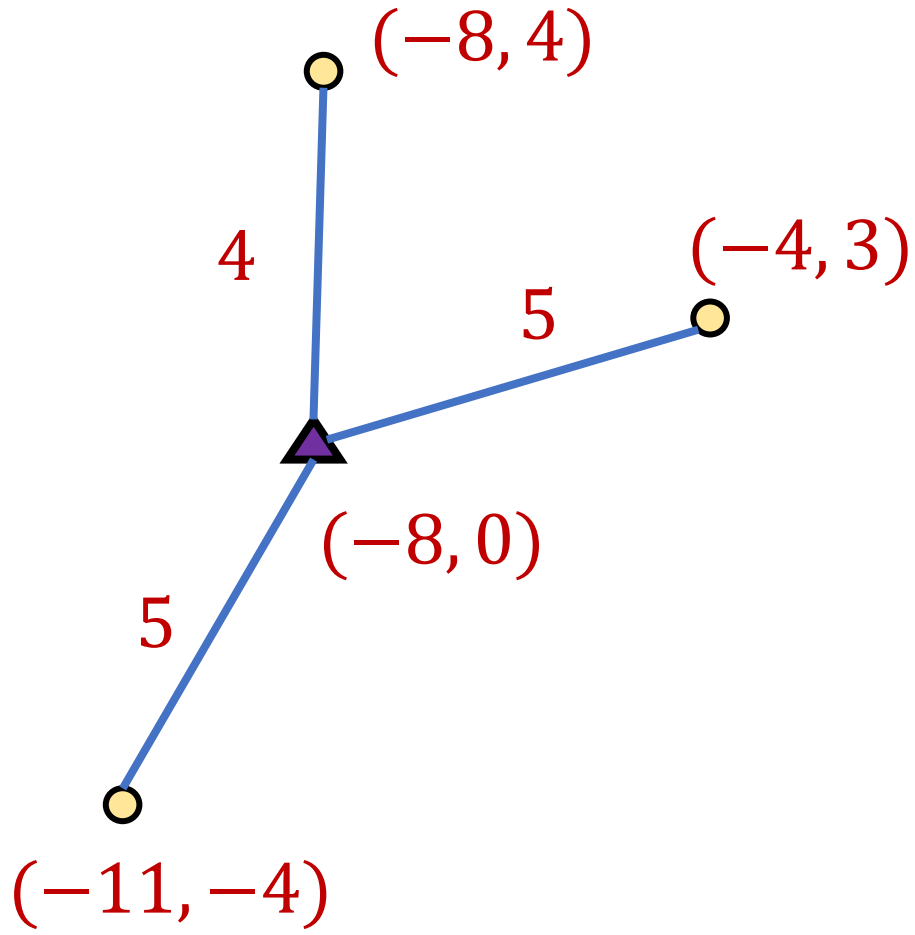
- (k, z)-clustering problem:

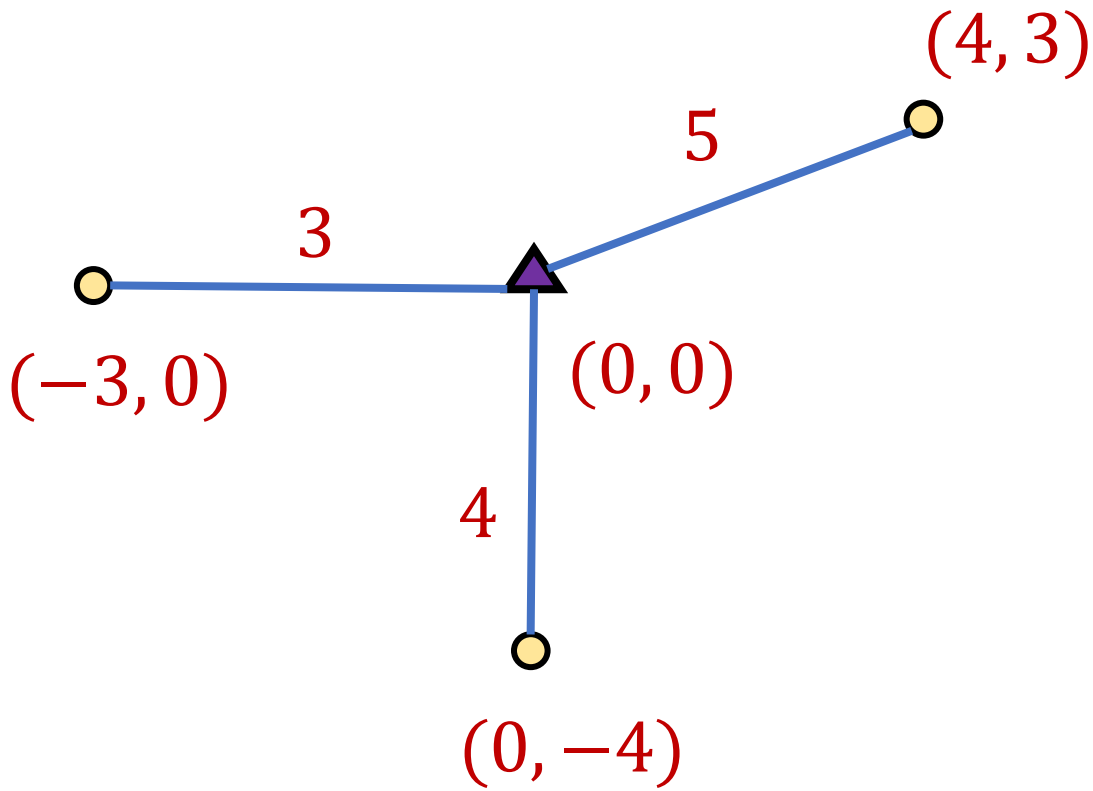
$$\min_{C:|C|\leq k} \mathrm{Cost}(X, C) = \min_{C:|C|\leq k} \Sigma_{x\in X} \big(\mathrm{dist}(x, C)\big)^z$$


$(-8, 4)$
$(-4, 3)$
$(4, 3)$
$(-3, 0)$
$(-11, -4)$
$(0, -4)$

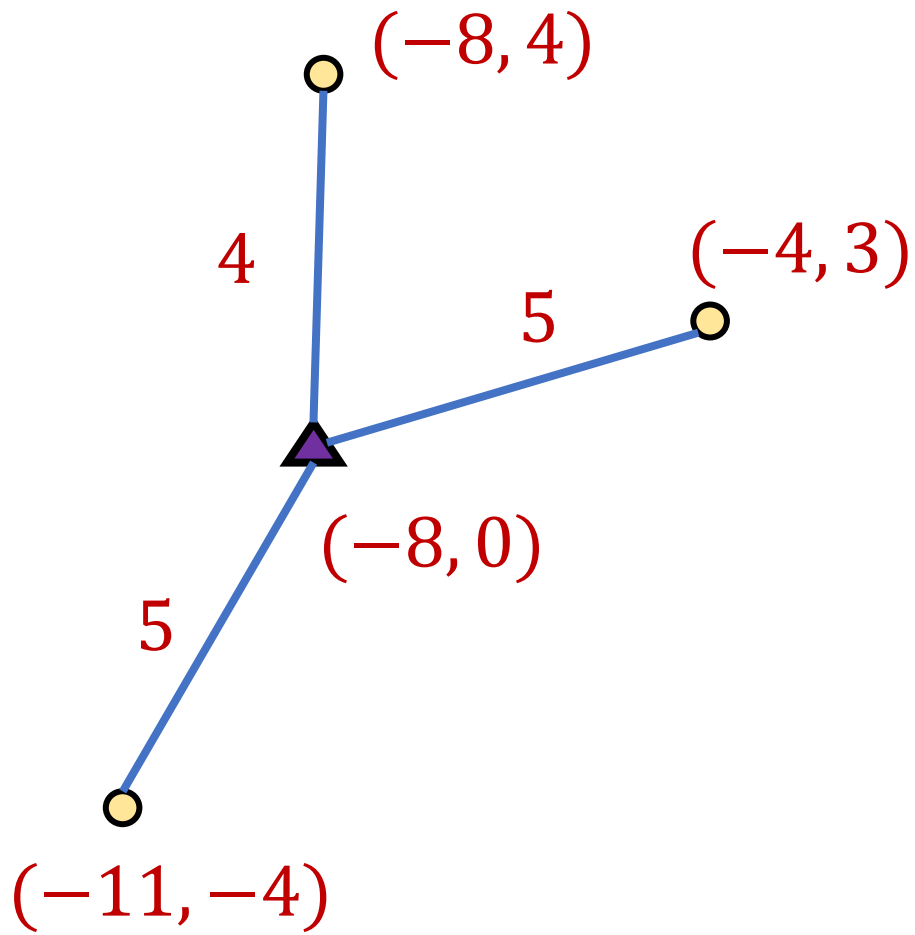

(−8, 4)
(−4, 3)
(−8, 0)
(−11, −4)

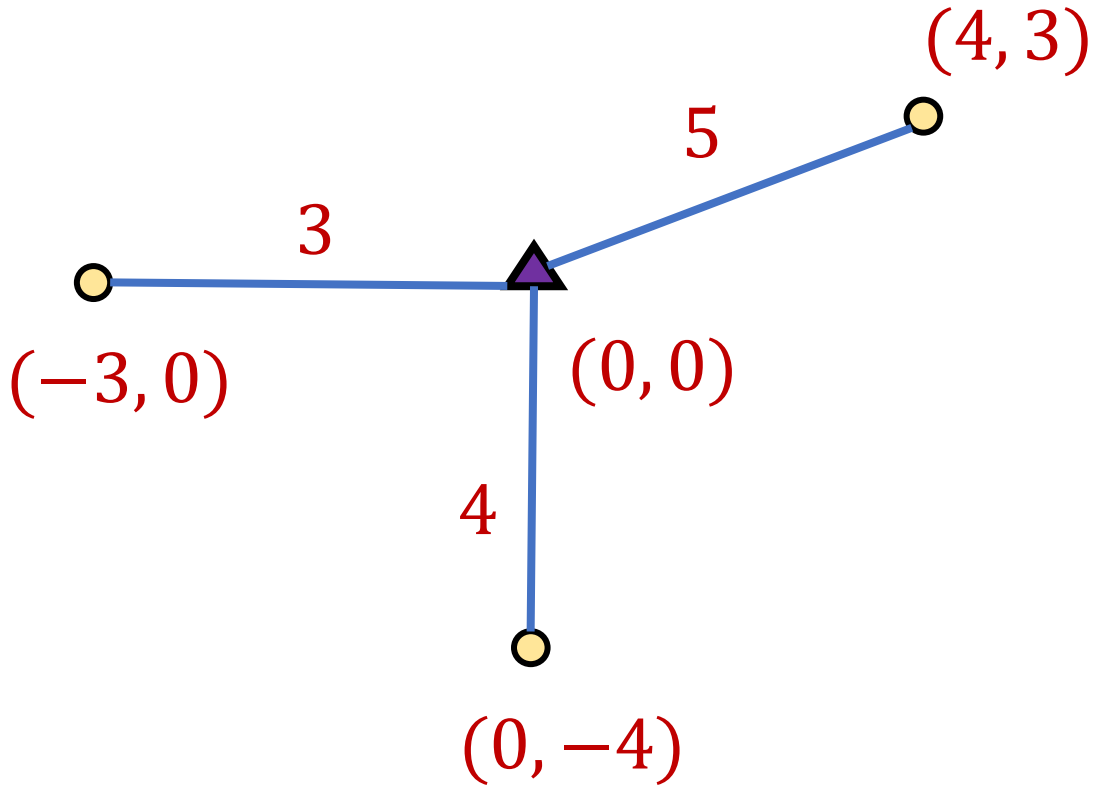

(4, 3)
(−3, 0)
(0, 0)
(0, −4)


(−8, 4)
4
(−4, 3)
5
(−8, 0)
5
(−11, −4)

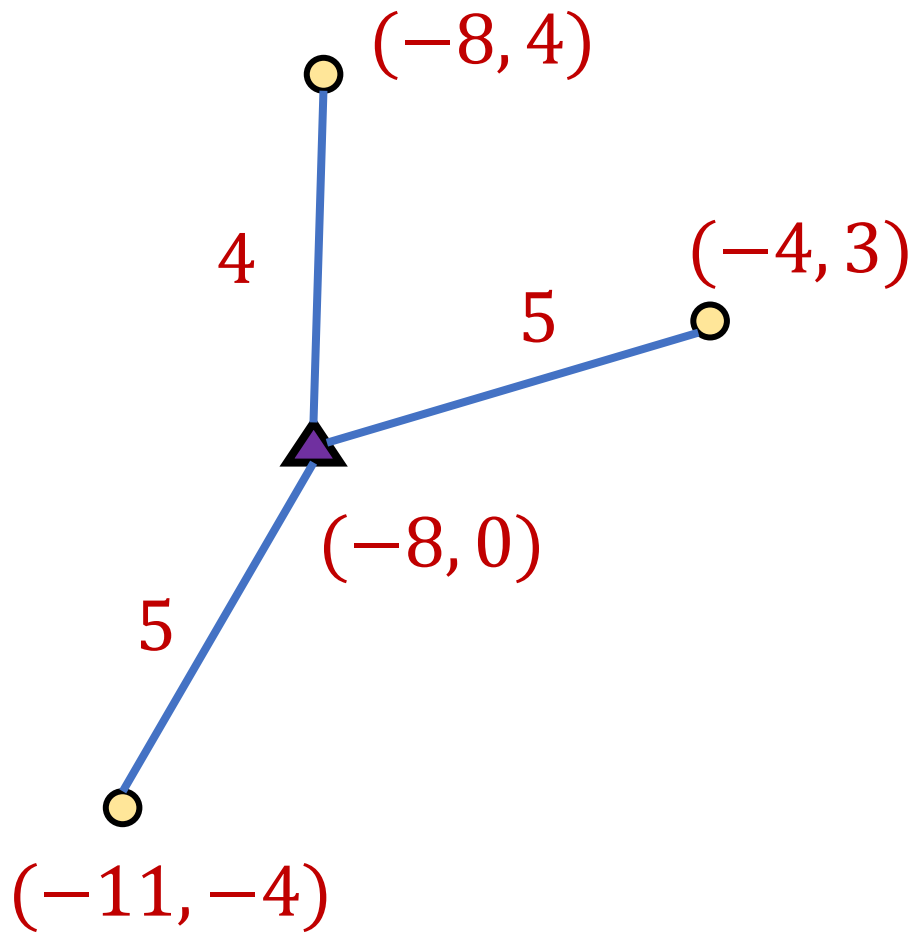

(4, 3)
5
3
(−3, 0)
(0, 0)
4
(0, −4)

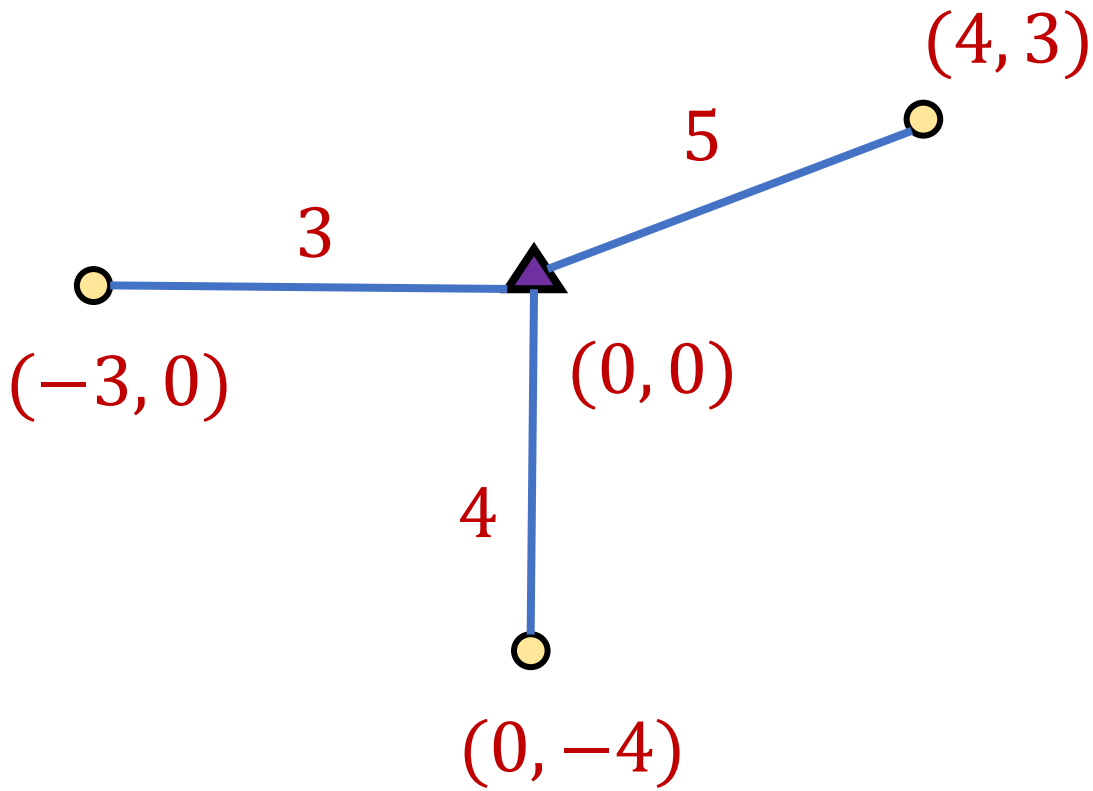
k-center: $\mathrm{Cost}(X, C) = \max_{x \in X} \mathrm{dist}(x, C) = 5$





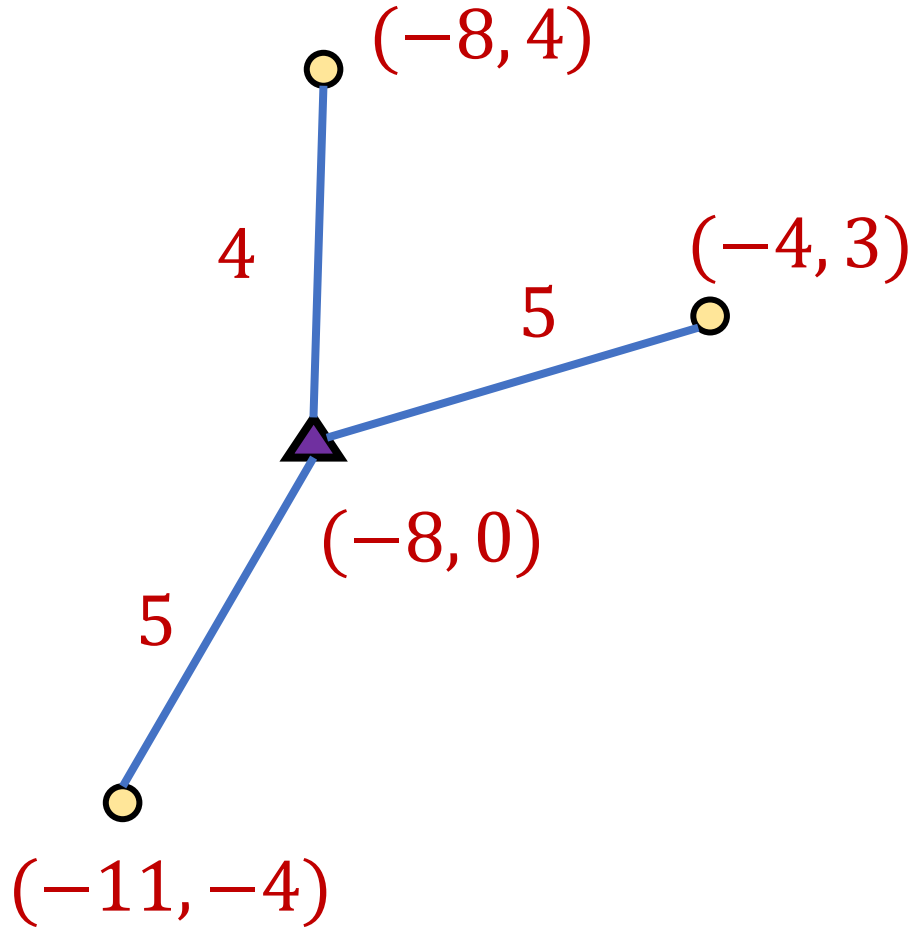
k-median: $\mathrm{Cost}(X, C) = \sum_{x \in X} \mathrm{dist}(x, C) = 4 + 5 + 5 + 3 + 4 + 5 = 26$

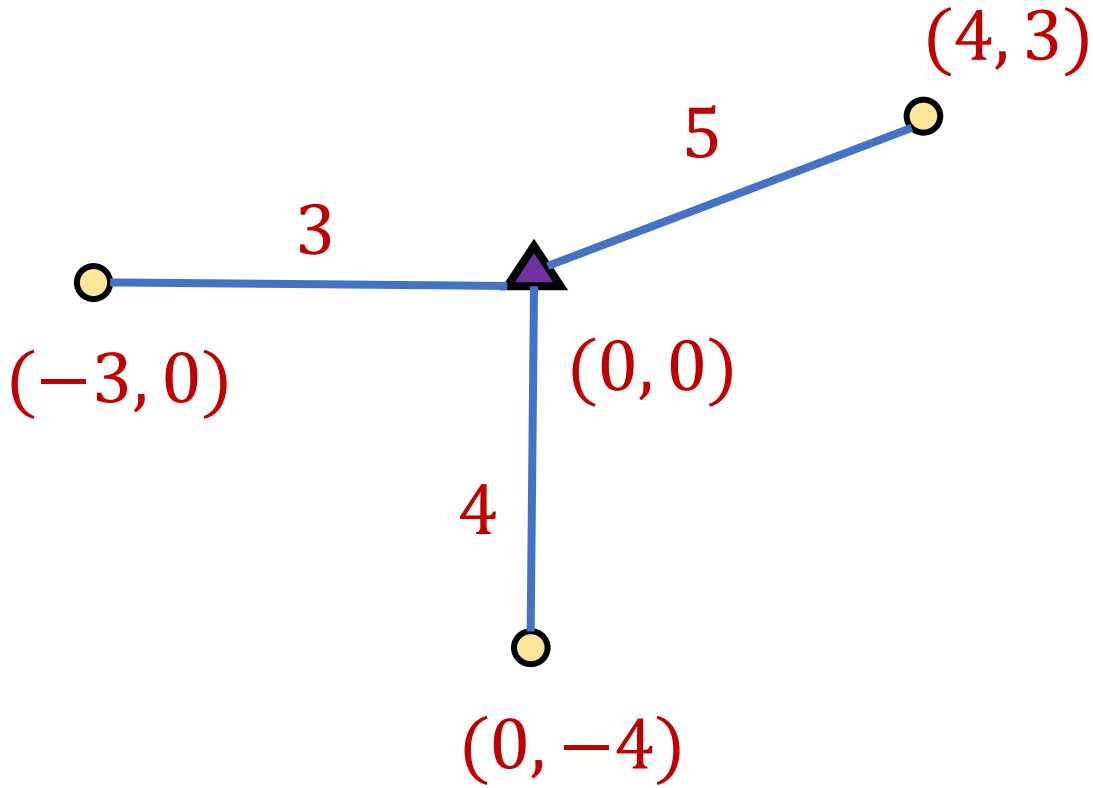




k-means: $\text{Cost}(X, C) = \sum_{x \in X} \big(\text{dist}(x, C)\big)^2 = 16 + 25 + 25 + 9 + 16 + 25$

$= 116$





Coreset

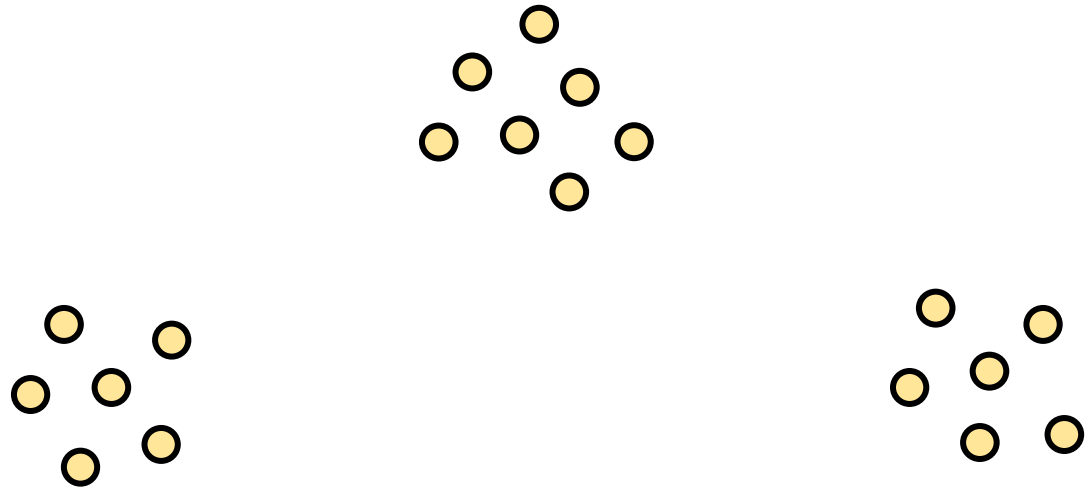
- Subset X' of representative points of X for a specific clustering objective

- $\mathrm{Cost}(X, C) \approx \mathrm{Cost}(X', C)$ for all sets C with $|C| = k$

Coreset

- Subset X' of representative points of X for a specific clustering objective

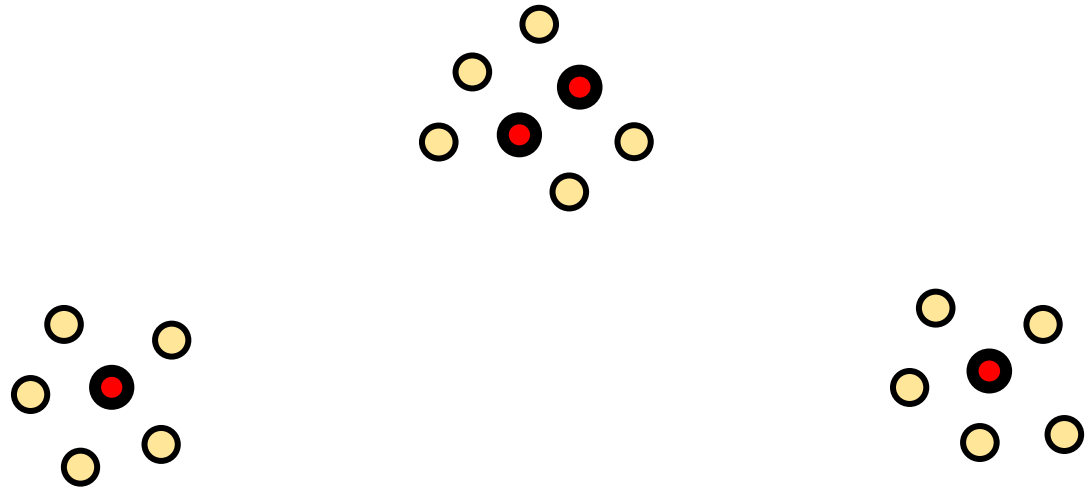
- $\mathrm{Cost}(X, C) \approx \mathrm{Cost}(X', C)$ for all sets C with $|C| = k$



Coreset

- Subset X' of representative points of X for a specific clustering objective

- $\mathrm{Cost}(X, C) \approx \mathrm{Cost}(X', C)$ for all sets C with $|C| = k$



Coreset (Formal Definition)

- Given a set X and an accuracy parameter $\varepsilon > 0$, we say a set X' with weight function w is an $(1+\varepsilon)$-*multiplicative coreset* for a cost function Cost, if for all queries C with $|C| \leq k$, we have

$$(1-\varepsilon)\mathrm{Cost}(X,C) \leq \mathrm{Cost}(X',C,w) \leq (1+\varepsilon)\mathrm{Cost}(X,C)$$

(k,z)-clustering: $\mathrm{Cost}(X',C,w) = \sum_{x\in X'} w(x)\cdot\big(\mathrm{dist}(x,C)\big)^z$

(k, z)-Clustering in the Streaming Model

- Merge-and-reduce framework

- Suppose there exists a $(1+\varepsilon)$-coreset construction for (k, z)-clustering that uses $f\left(k, \frac{1}{\varepsilon}\right)$ weighted input points

$$\tilde{O}\left(\frac{k^2}{\varepsilon^2}\right)$$

- Partition the stream into blocks containing $f\left(k, \frac{\log n}{\varepsilon}\right)$ points

(k, z)-Clustering in the Streaming Model

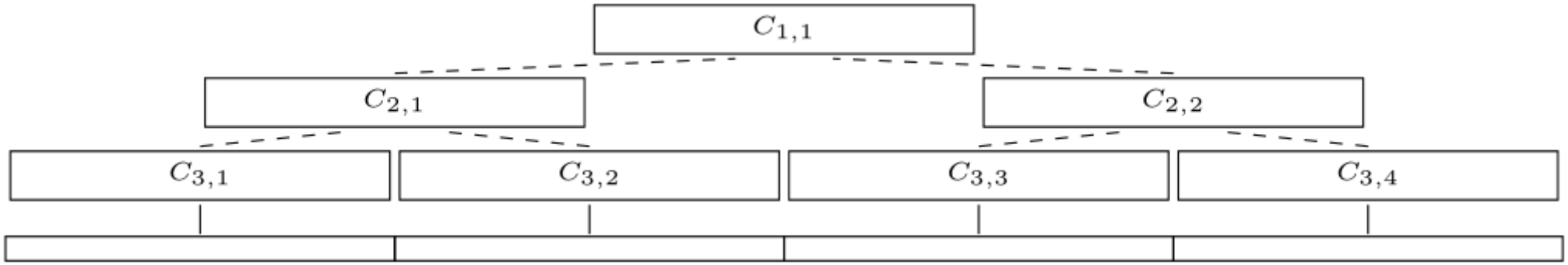
- Partition the stream into blocks containing $f\left(k, \frac{\log n}{\varepsilon}\right)$ points
- Create a $\left(1 + \frac{\varepsilon}{\log n}\right)$-coreset for each block
- Create a $\left(1 + \frac{\varepsilon}{\log n}\right)$-coreset for the set of points formed by the union of two coresets for each block

Reduce

Merge

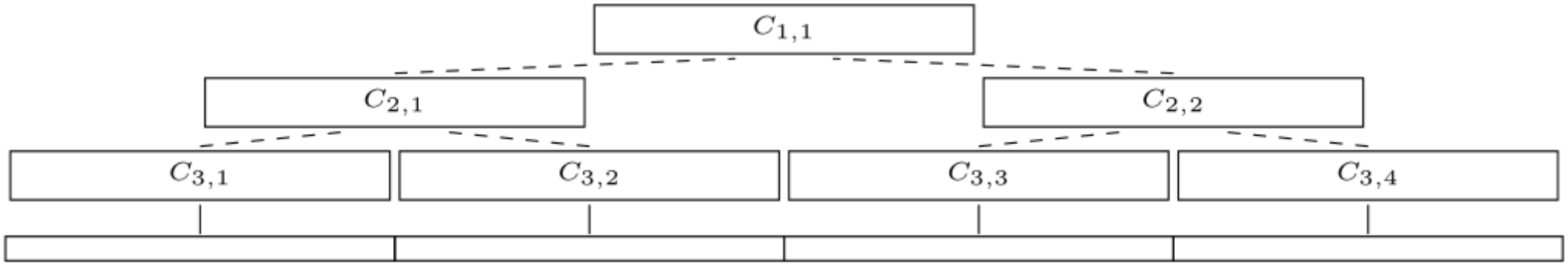
(k, z)-Clustering in the Streaming Model

- Partition the stream into blocks containing $f\left(k, \frac{\log n}{\varepsilon}\right)$ points
- Create a $\left(1 + \frac{\varepsilon}{\log n}\right)$-coreset for each block
- Create a $\left(1 + \frac{\varepsilon}{\log n}\right)$-coreset for the set of points formed by the union of two coresets for each block



(k, z)-Clustering in the Streaming Model

- There are $O(\log n)$ levels
- Each coreset is a $\left(1 + \frac{\varepsilon}{\log n}\right)$-coreset of two coresets
- Total approximation is $\left(1 + \frac{\varepsilon}{\log n}\right)^{\log n} = (1 + O(\varepsilon))$



(k, z)-Clustering in the Streaming Model

- Suppose there exists a $(1+\varepsilon)$-coreset construction for (k, z)-clustering that uses $f\left(k, \frac{1}{\varepsilon}\right)$ weighted input points
- Partition the stream into blocks containing $f\left(k, \frac{\log n}{\varepsilon}\right)$ points
- Total space is $f\left(k, \frac{\log n}{\varepsilon}\right) \cdot O(\log n)$ points

For k-means clustering, this is $\tilde{O}\left(\frac{k^2}{\varepsilon^2} \cdot \log^3 n\right)$ points