

# Clustering algorithms

2IMW30 - Foundations of data mining  
TU Eindhoven, Quartile 3, 2016-2016

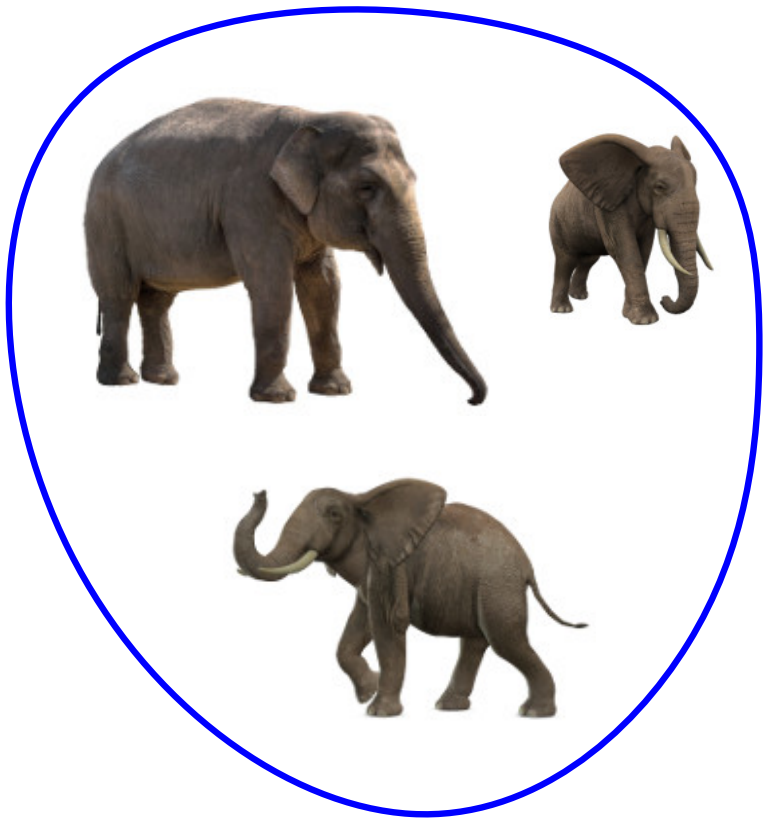
Anne Driemel

# Overview of this lecture

- Clustering
- Facility Location
- Gonzales' algorithm
- Lloyd's algorithm (k-means)
- k-means++ algorithm
- Clustering in graphs

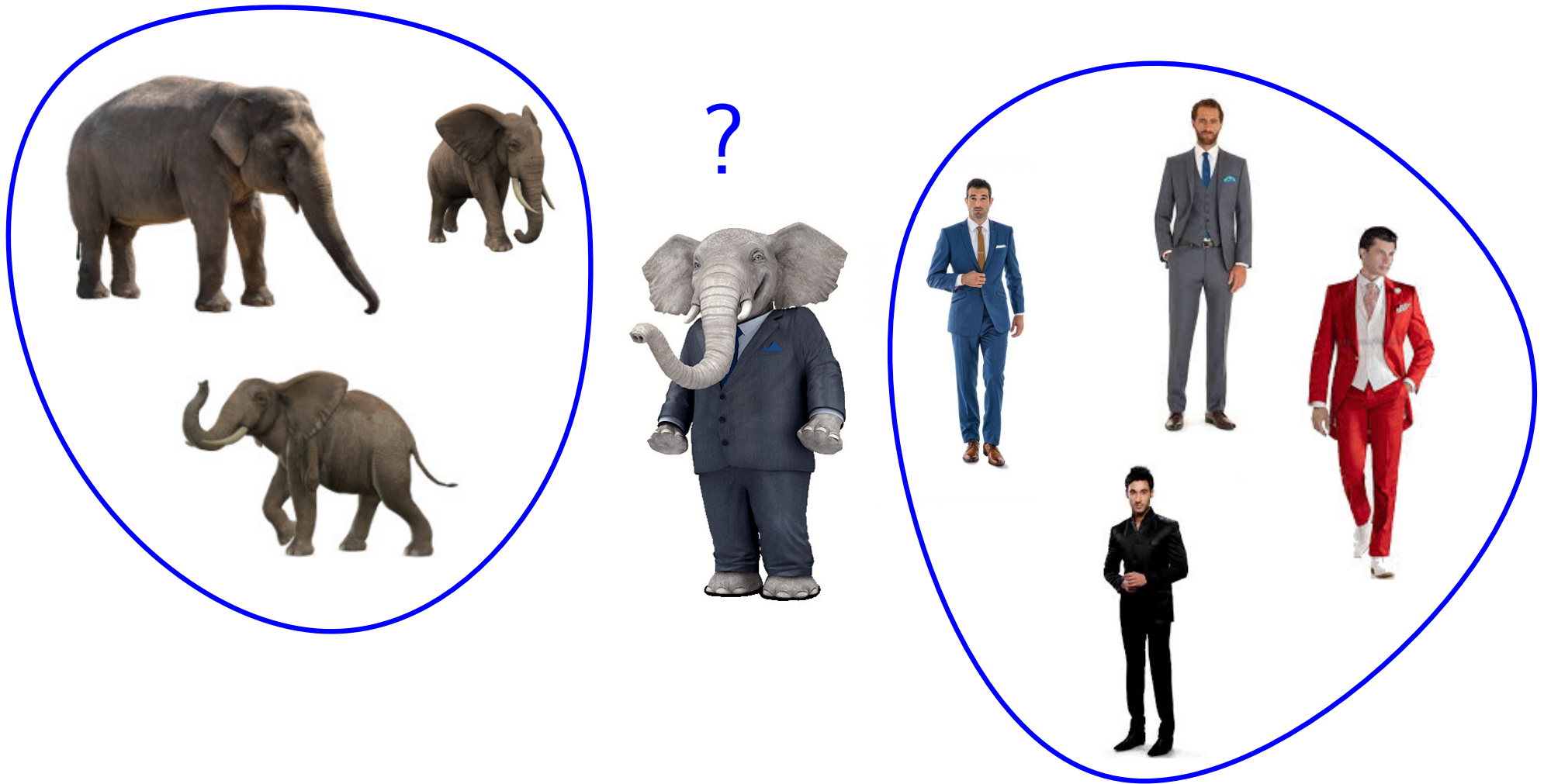
# What is Clustering?

Clustering is the task of grouping similar objects into clusters



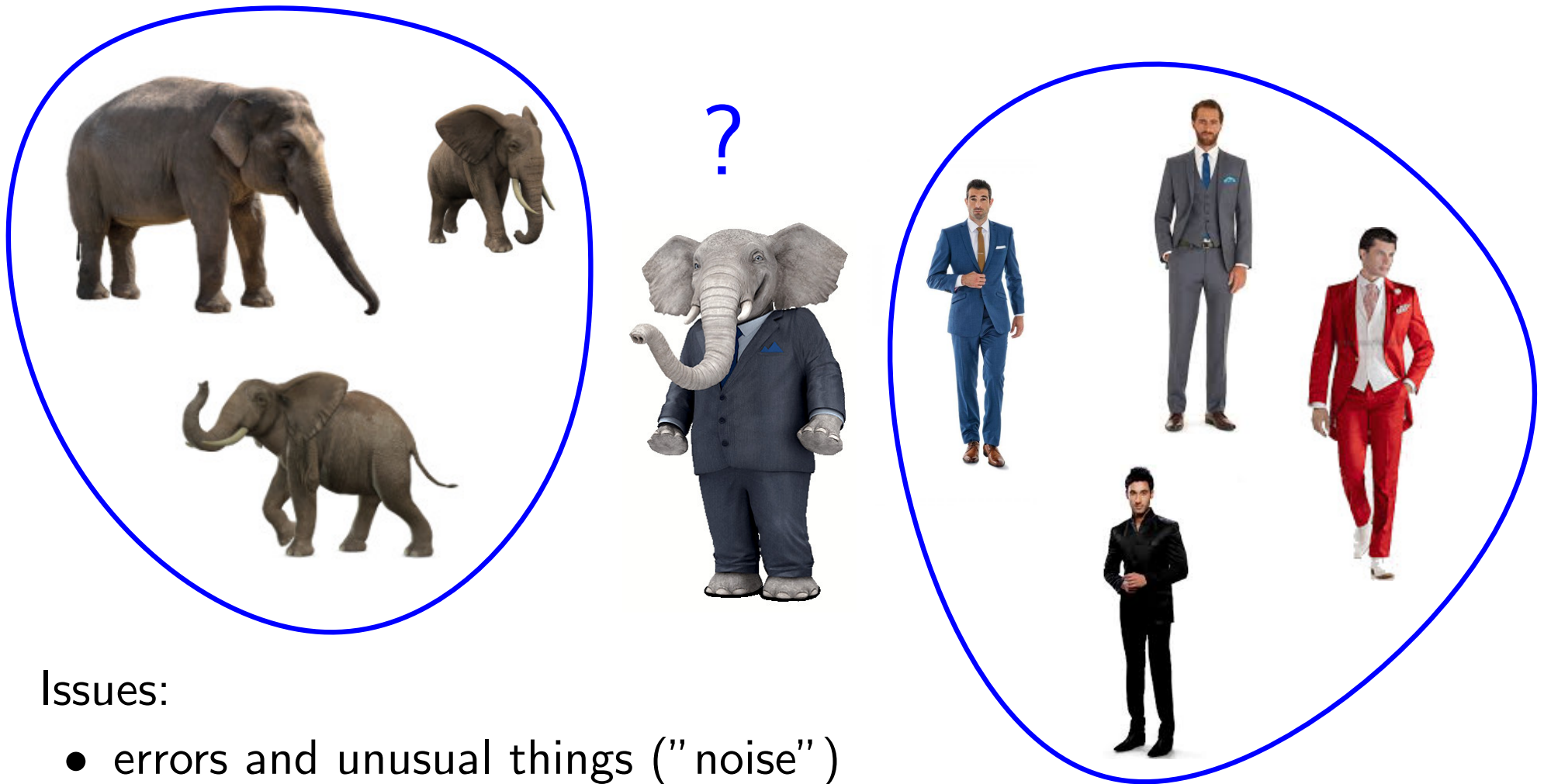
# What is Clustering?

Clustering is the task of grouping similar objects into clusters



# What is Clustering?

Clustering is the task of grouping similar objects into clusters

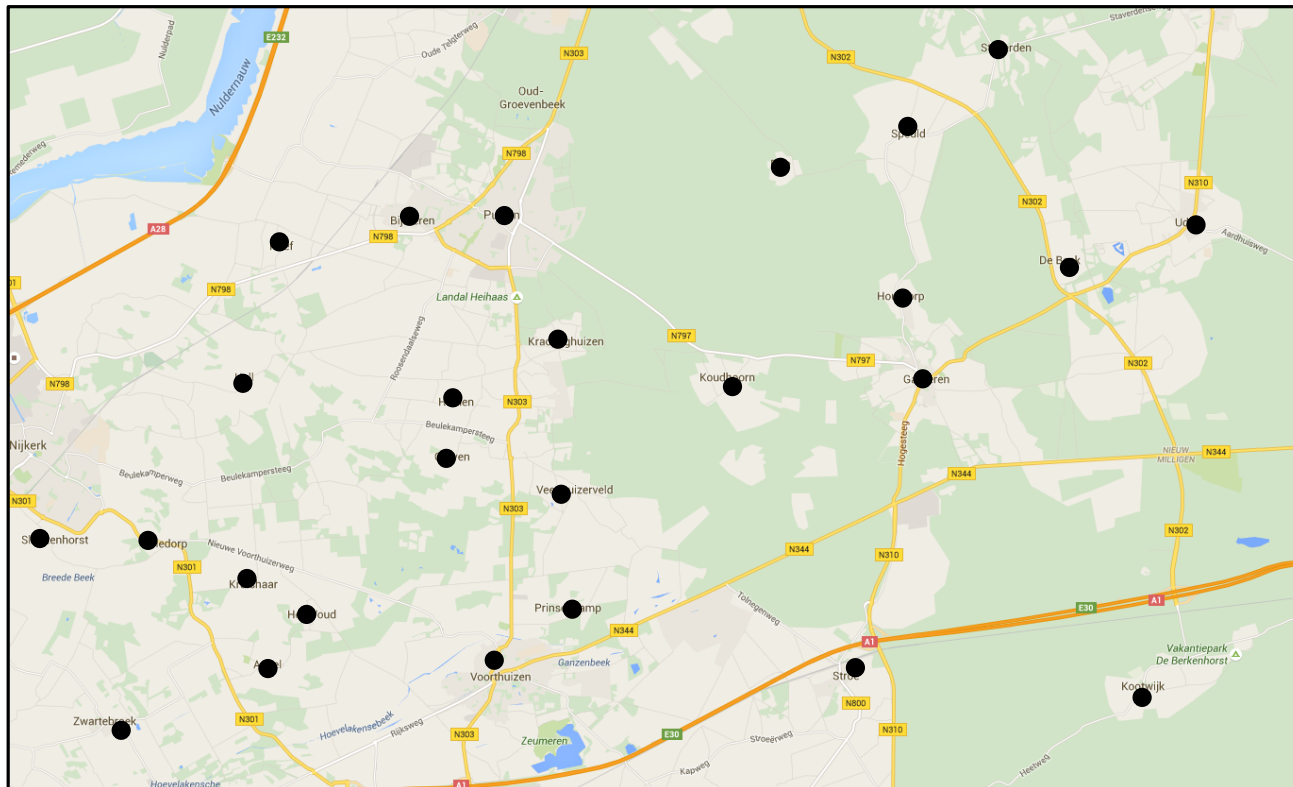


Issues:

- errors and unusual things ("noise")
- what is the "right" clustering?

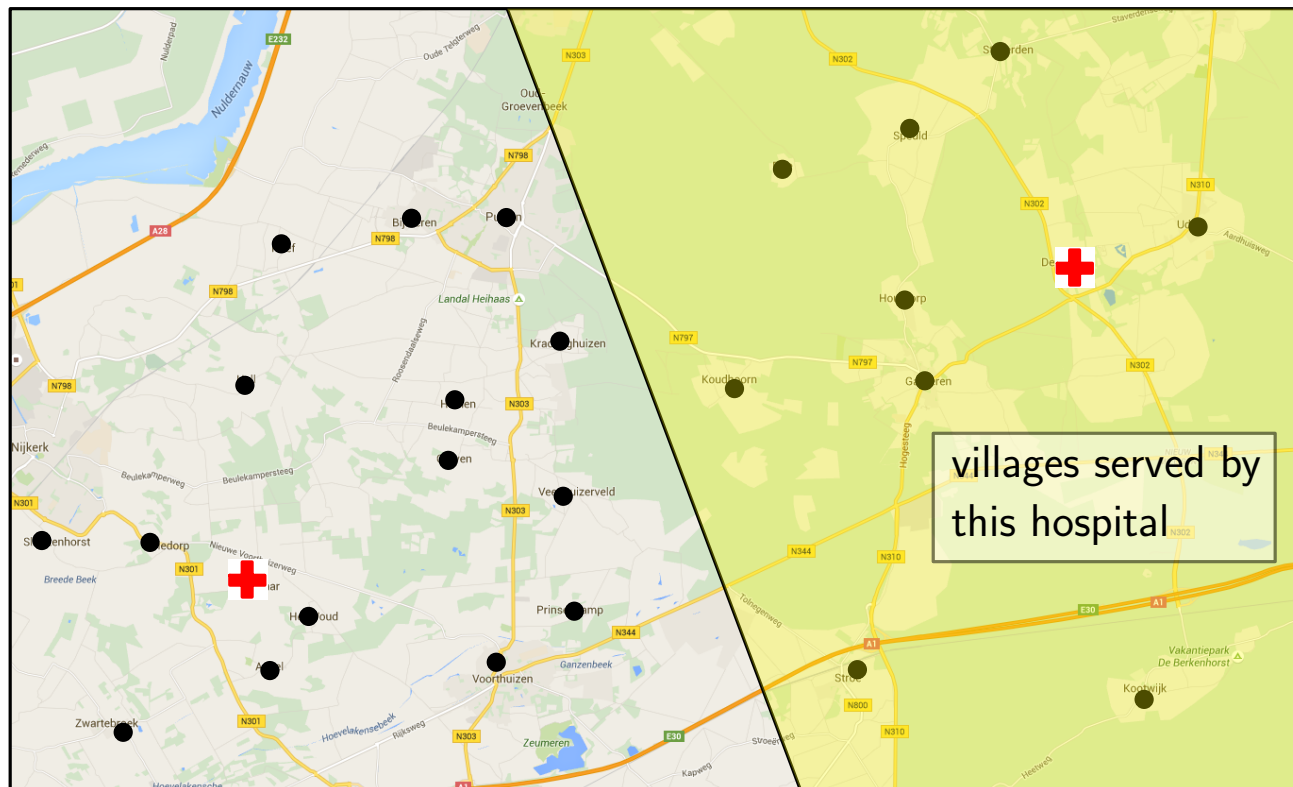
# Facility Location

You may build two hospitals in two different villages serving the surrounding villages. Where do you place them to minimize the maximal distance from any village to its serving hospital?



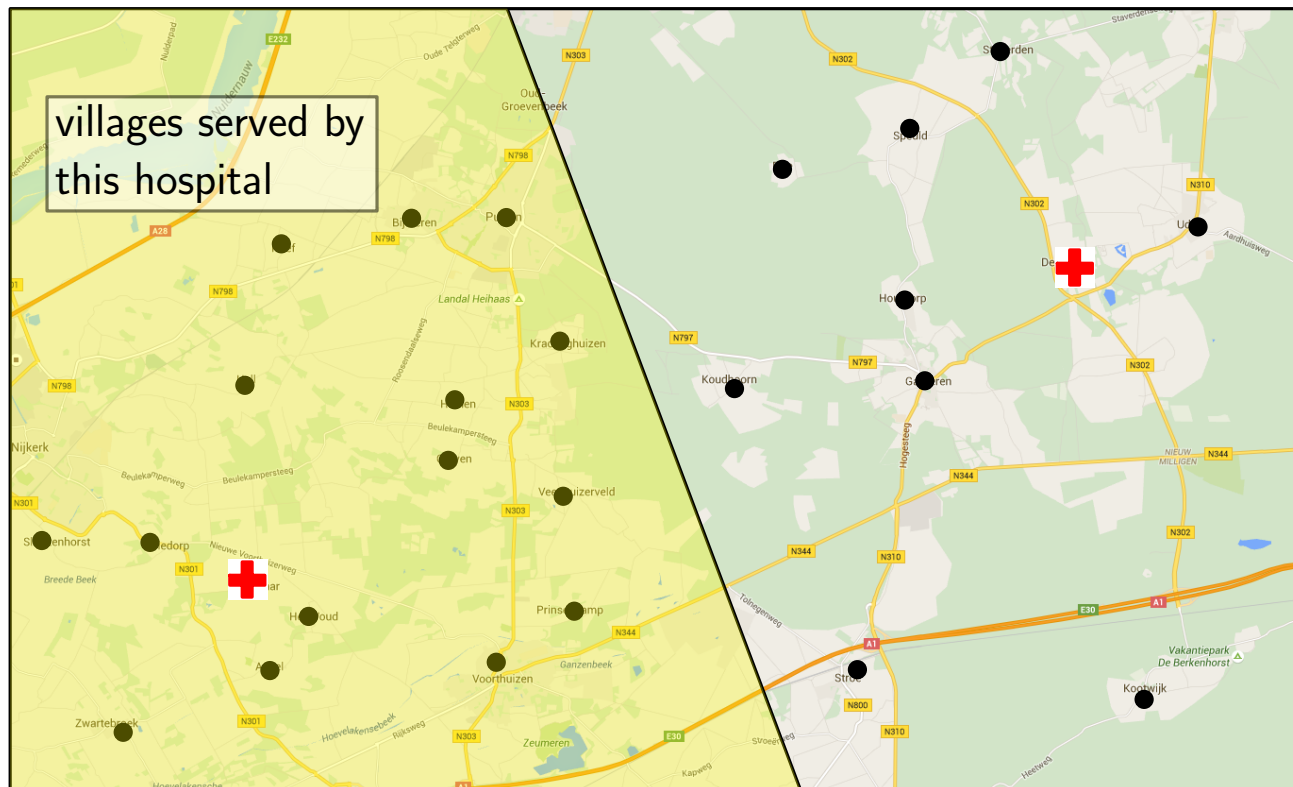
# Facility Location

You may build two hospitals in two different villages serving the surrounding villages. Where do you place them to minimize the maximal distance from any village to its serving hospital?



# Facility Location

You may build two hospitals in two different villages serving the surrounding villages. Where do you place them to minimize the maximal distance from any village to its serving hospital?





# $k$ -center clustering

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq P$

---

## Problem:

- each  $p_i \in P$  is "served by" its closest center

$$\operatorname{argmin}_{c_j \in C} \|p_i - c_j\|$$

- all points served by a center  $c_j$  together form a "cluster"
- we want to choose  $\{c_1, \dots, c_k\}$  to minimize the cost function

$$\phi(P, C) = \max_{p_i \in P} \left\| p_i - \operatorname{argmin}_{c_j \in C} \|p_i - c_j\| \right\|$$

# Gonzales' algorithm for $k$ -center

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq P$

---

## Algorithm:

- choose an arbitrary point  $p_i \in P$  and set  $c_1 = p_i$
- for  $t = 2, \dots, k$  set

$$c_t = \operatorname{argmax}_{p_i \in P} \left\| p_i - \operatorname{argmin}_{c_j \in \{c_1, \dots, c_{t-1}\}} \|p_i - c_j\| \right\|$$

# Gonzales' algorithm for $k$ -center

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq P$

---

## Algorithm:

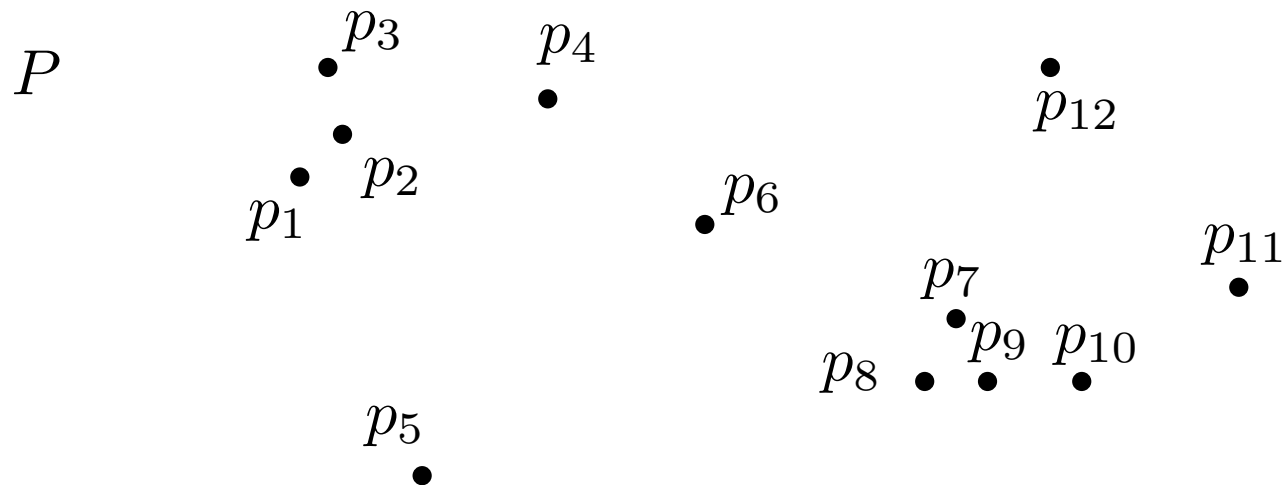
- choose an arbitrary point  $p_i \in P$  and set  $c_1 = p_i$
- for  $t = 2, \dots, k$  set

$$c_t = \operatorname{argmax}_{p_i \in P} \left\| p_i - \operatorname{argmin}_{c_j \in \{c_1, \dots, c_{t-1}\}} \|p_i - c_j\| \right\|$$

"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$

# Gonzales' algorithm for $k$ -center

$\phi(P, \{c_1, \dots, c_t\})$  corresponds to the smallest radius such that all points are covered

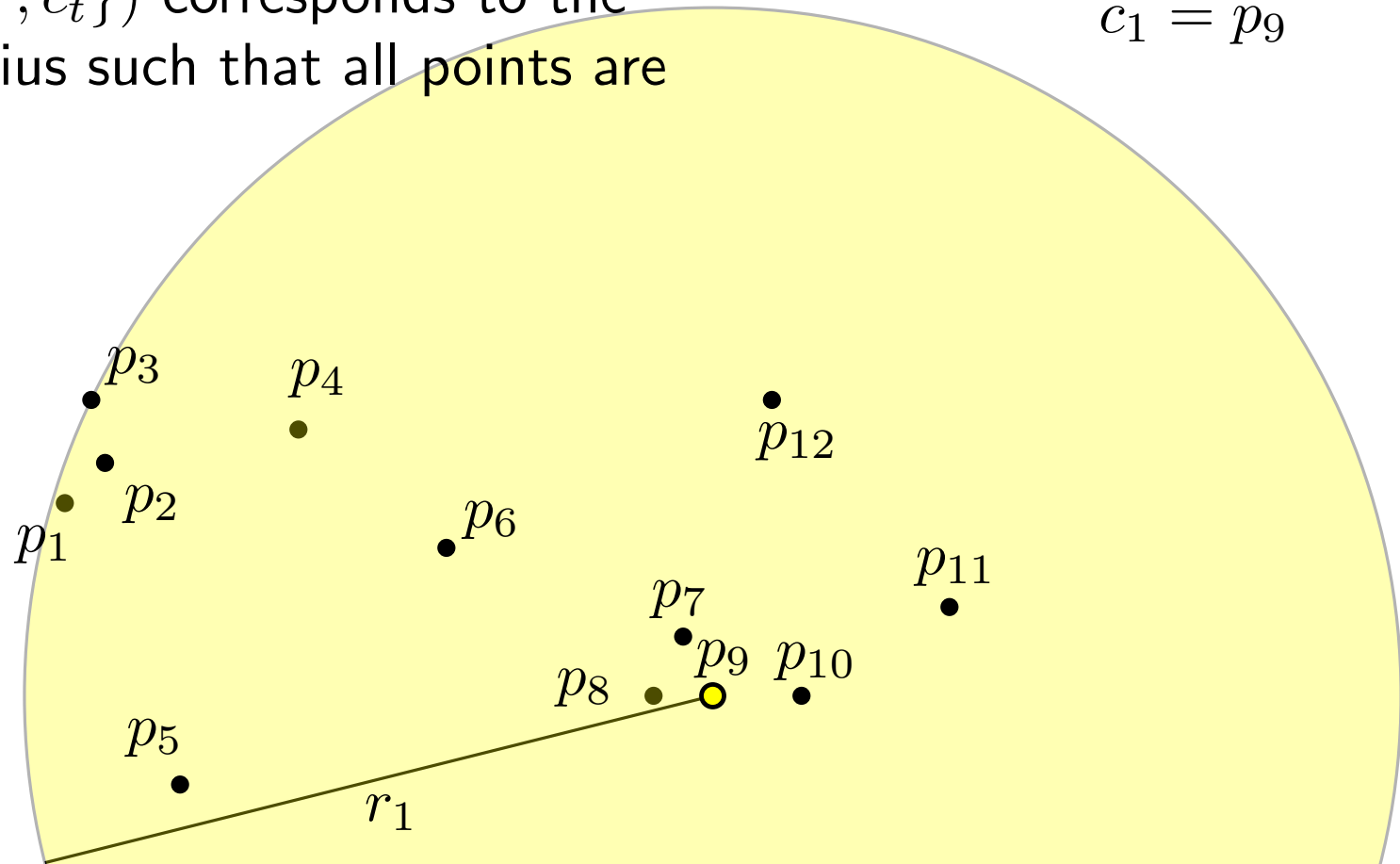


"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$

# Gonzales' algorithm for $k$ -center

$\phi(P, \{c_1, \dots, c_t\})$  corresponds to the smallest radius such that all points are covered

$$c_1 = p_9$$



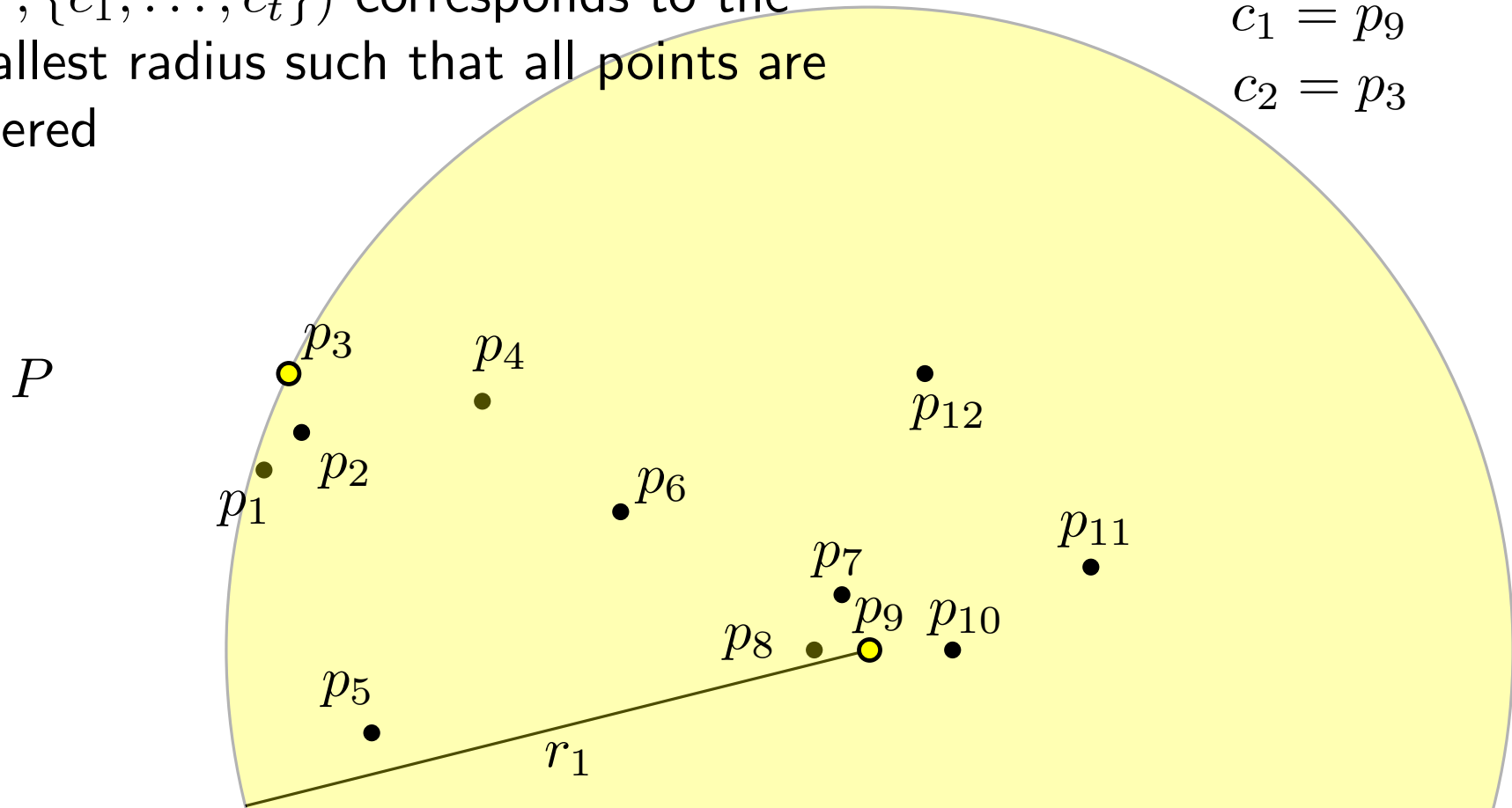
"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$

# Gonzales' algorithm for $k$ -center

$\phi(P, \{c_1, \dots, c_t\})$  corresponds to the smallest radius such that all points are covered

$$c_1 = p_9$$

$$c_2 = p_3$$



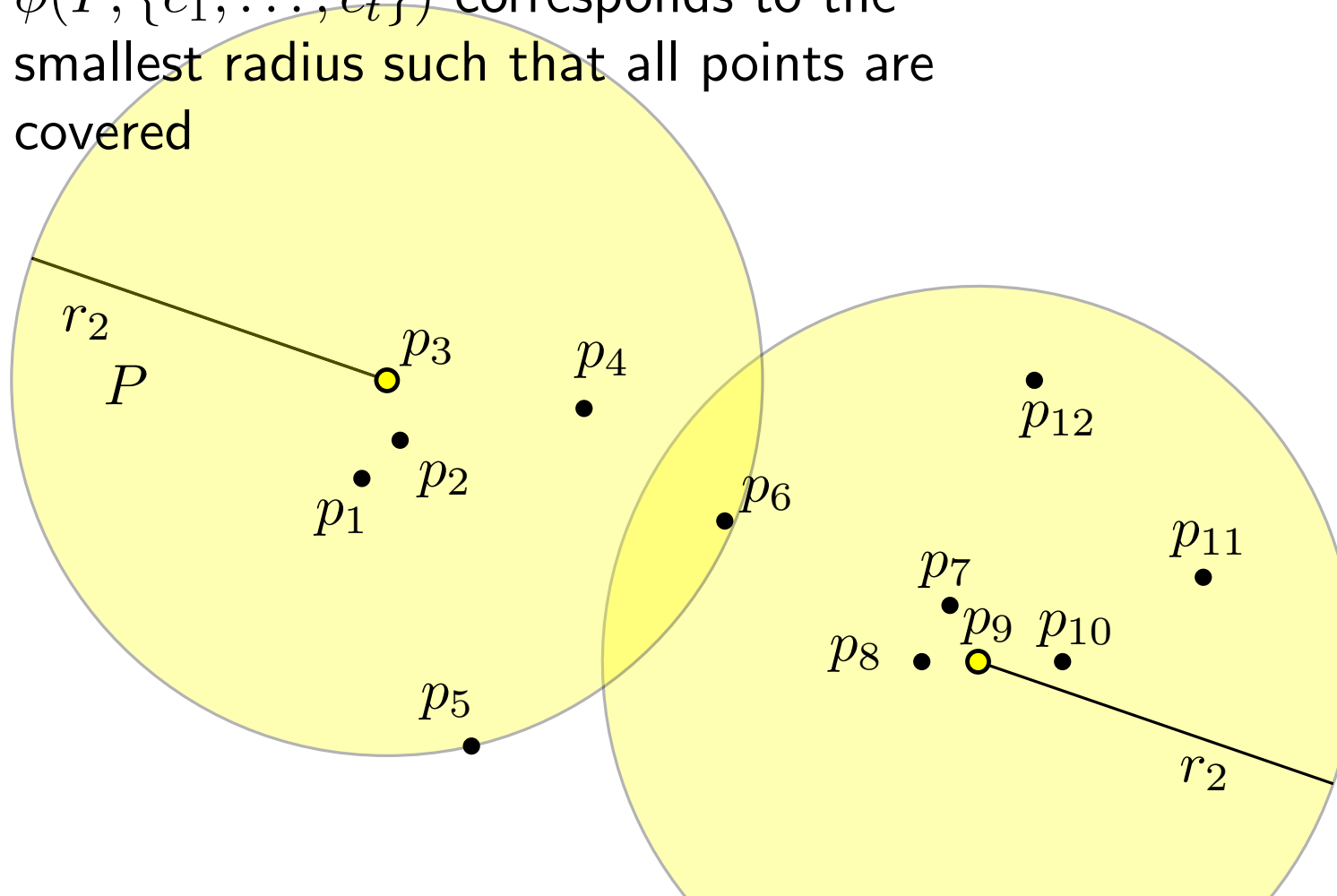
"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$

# Gonzales' algorithm for $k$ -center

$\phi(P, \{c_1, \dots, c_t\})$  corresponds to the smallest radius such that all points are covered

$$c_1 = p_9$$

$$c_2 = p_3$$



"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$

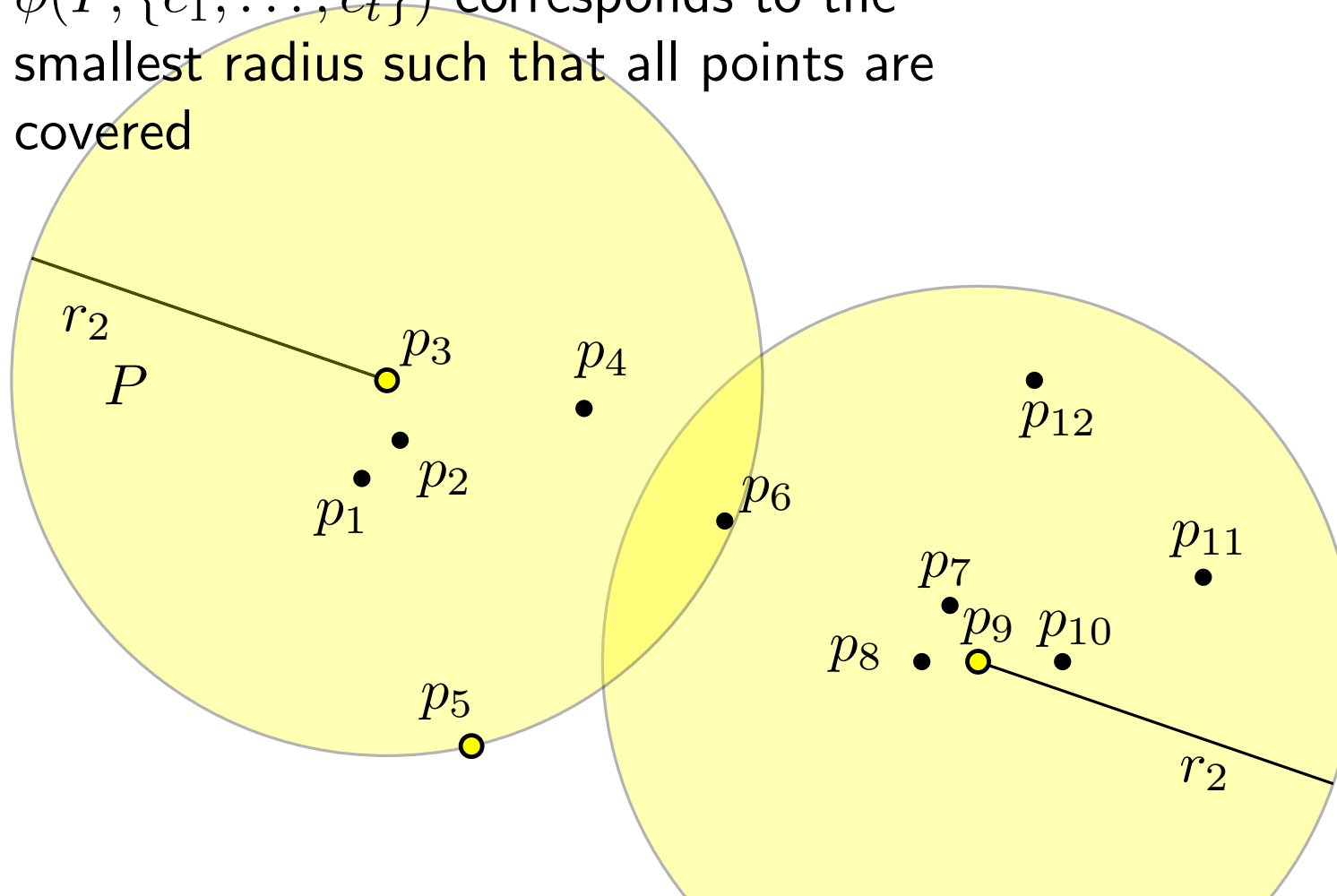
# Gonzales' algorithm for $k$ -center

$\phi(P, \{c_1, \dots, c_t\})$  corresponds to the smallest radius such that all points are covered

$$c_1 = p_9$$

$$c_2 = p_3$$

$$c_3 = p_5$$



"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$



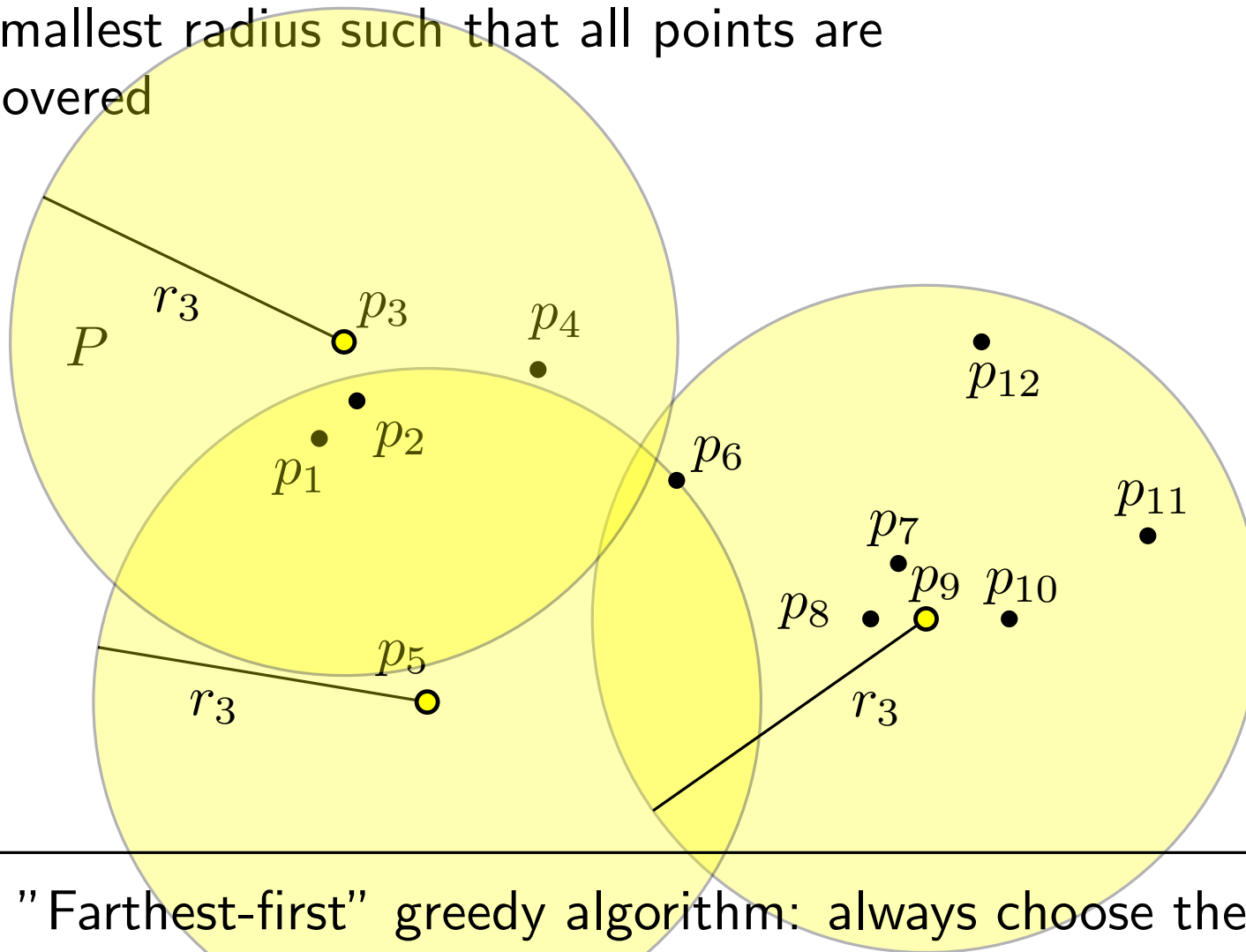
# Gonzales' algorithm for $k$ -center

$\phi(P, \{c_1, \dots, c_t\})$  corresponds to the smallest radius such that all points are covered

$$c_1 = p_9$$

$$c_2 = p_3$$

$$c_3 = p_5$$



"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$

# Gonzales' algorithm for $k$ -center

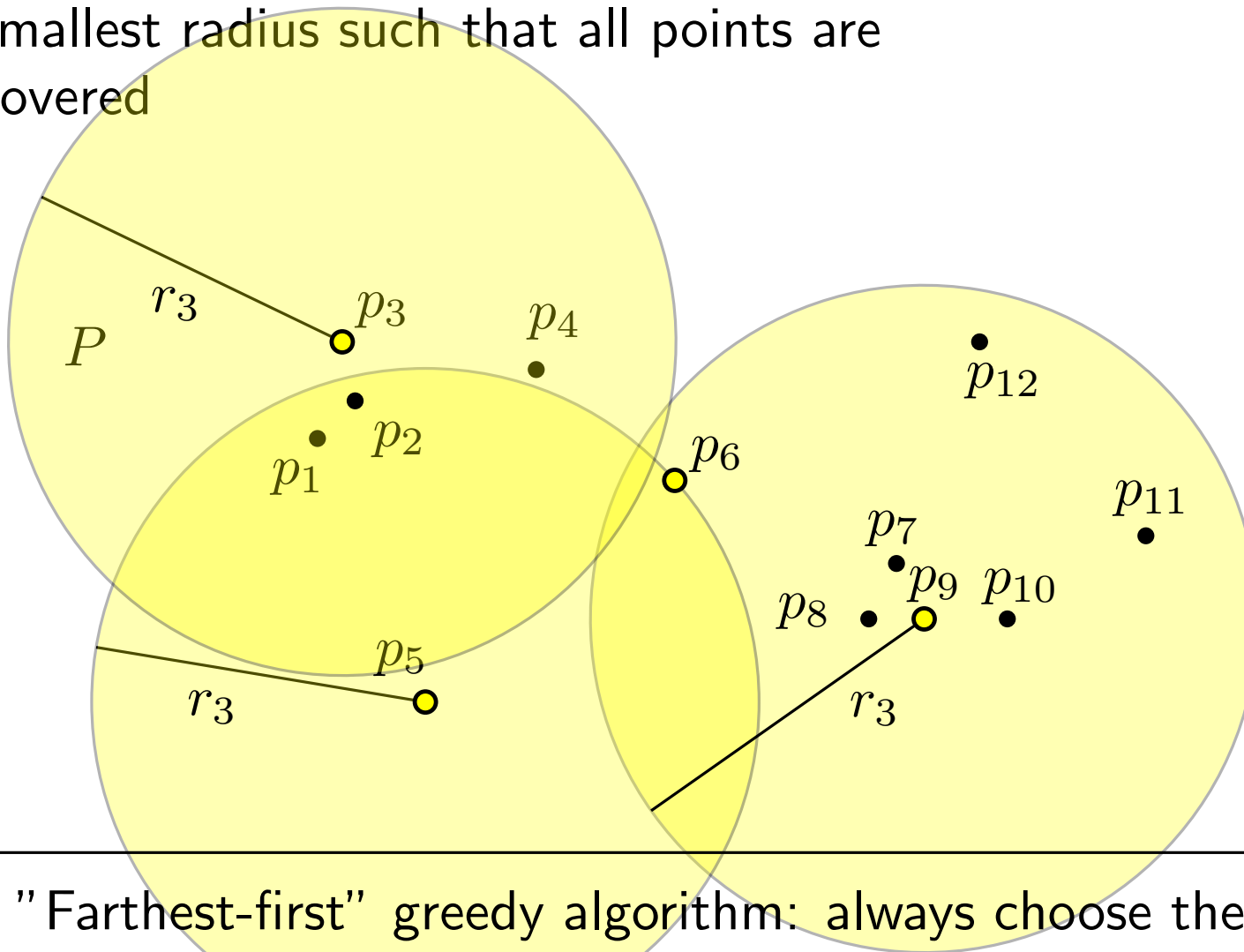
$\phi(P, \{c_1, \dots, c_t\})$  corresponds to the smallest radius such that all points are covered

$$c_1 = p_9$$

$$c_2 = p_3$$

$$c_3 = p_5$$

$$c_4 = p_6$$



"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$

# Gonzales' algorithm for $k$ -center

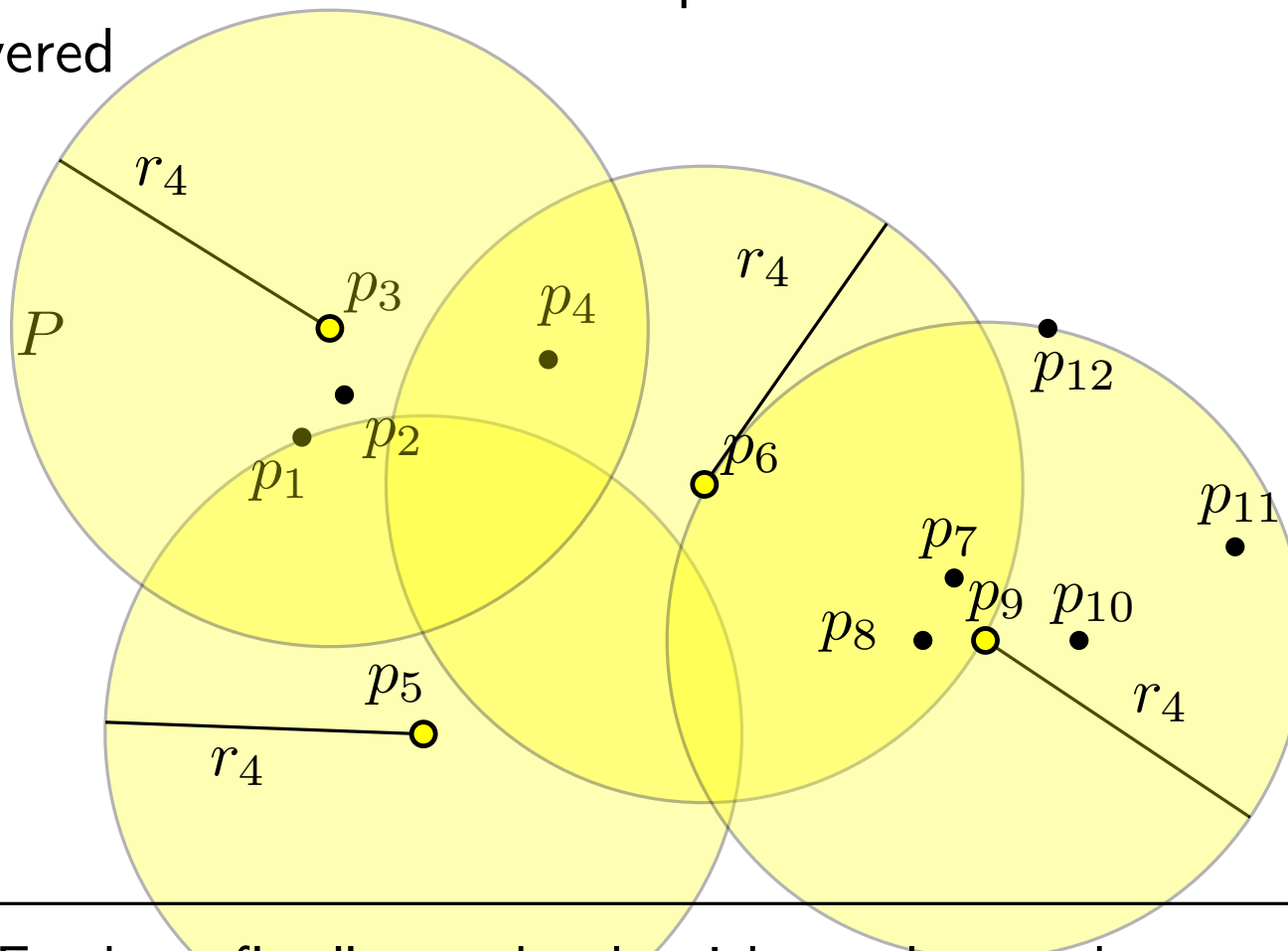
$\phi(P, \{c_1, \dots, c_t\})$  corresponds to the smallest radius such that all points are covered

$$c_1 = p_9$$

$$c_2 = p_3$$

$$c_3 = p_5$$

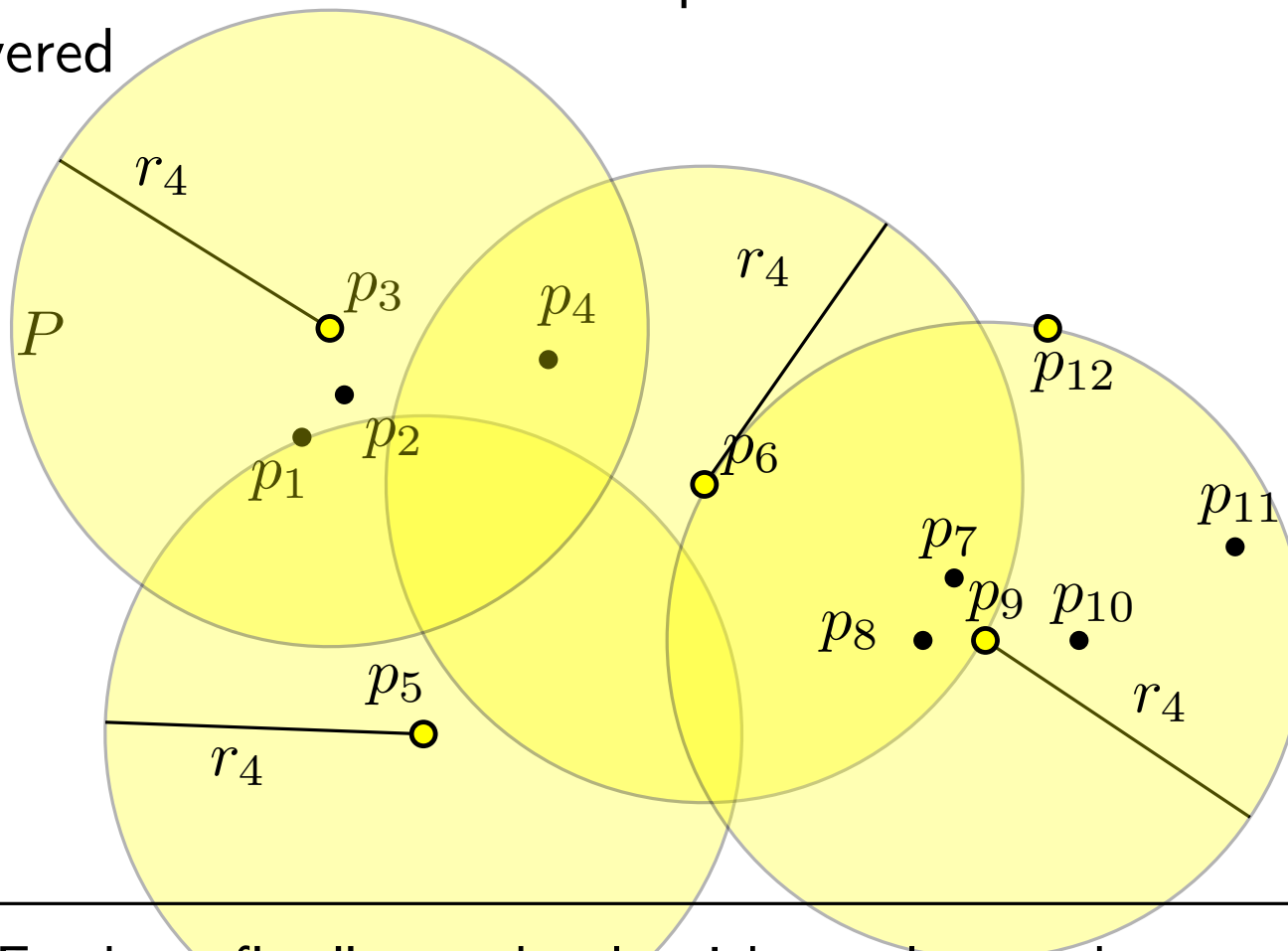
$$c_4 = p_6$$



"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$

# Gonzales' algorithm for $k$ -center

$\phi(P, \{c_1, \dots, c_t\})$  corresponds to the smallest radius such that all points are covered

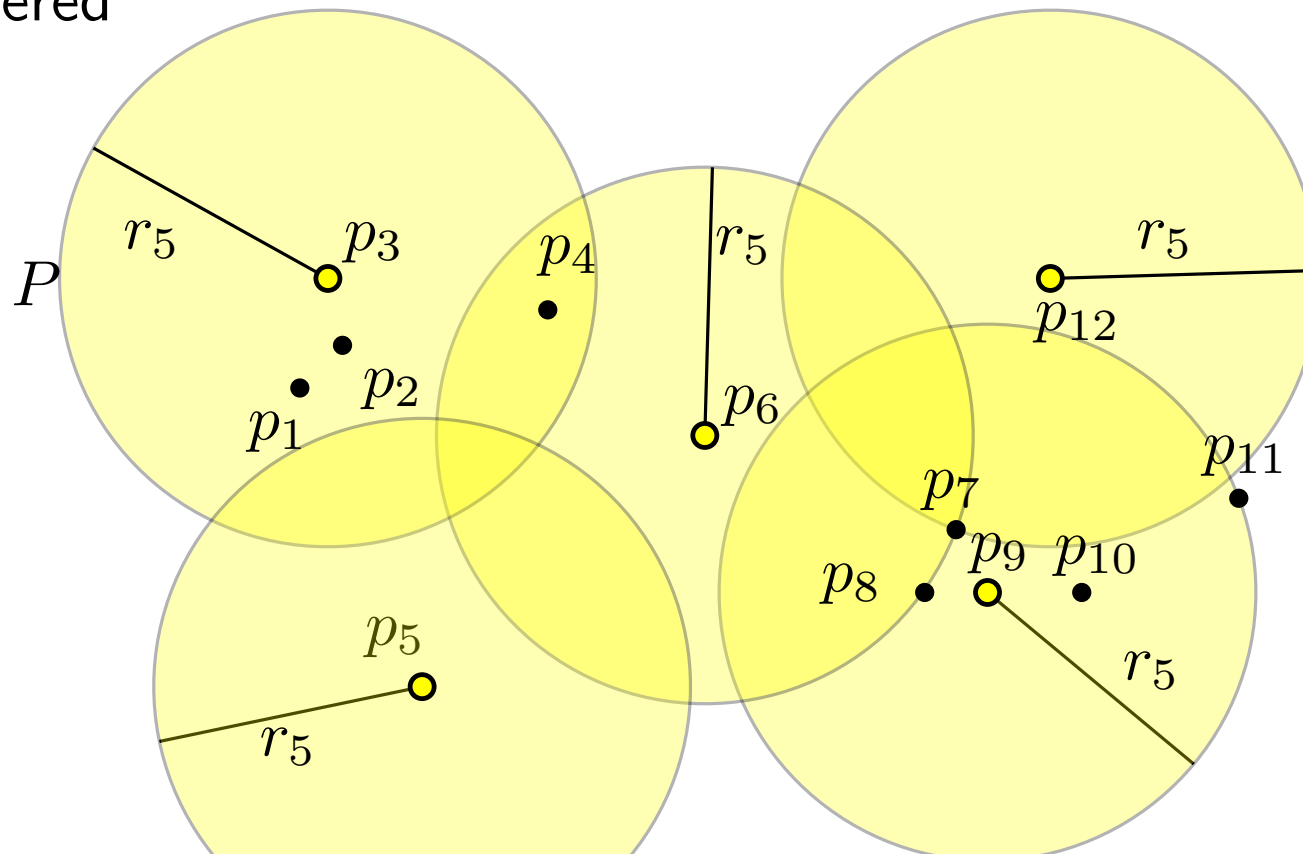


$c_1 = p_9$   
 $c_2 = p_3$   
 $c_3 = p_5$   
 $c_4 = p_6$   
 $c_5 = p_{12}$

"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$

# Gonzales' algorithm for $k$ -center

$\phi(P, \{c_1, \dots, c_t\})$  corresponds to the smallest radius such that all points are covered

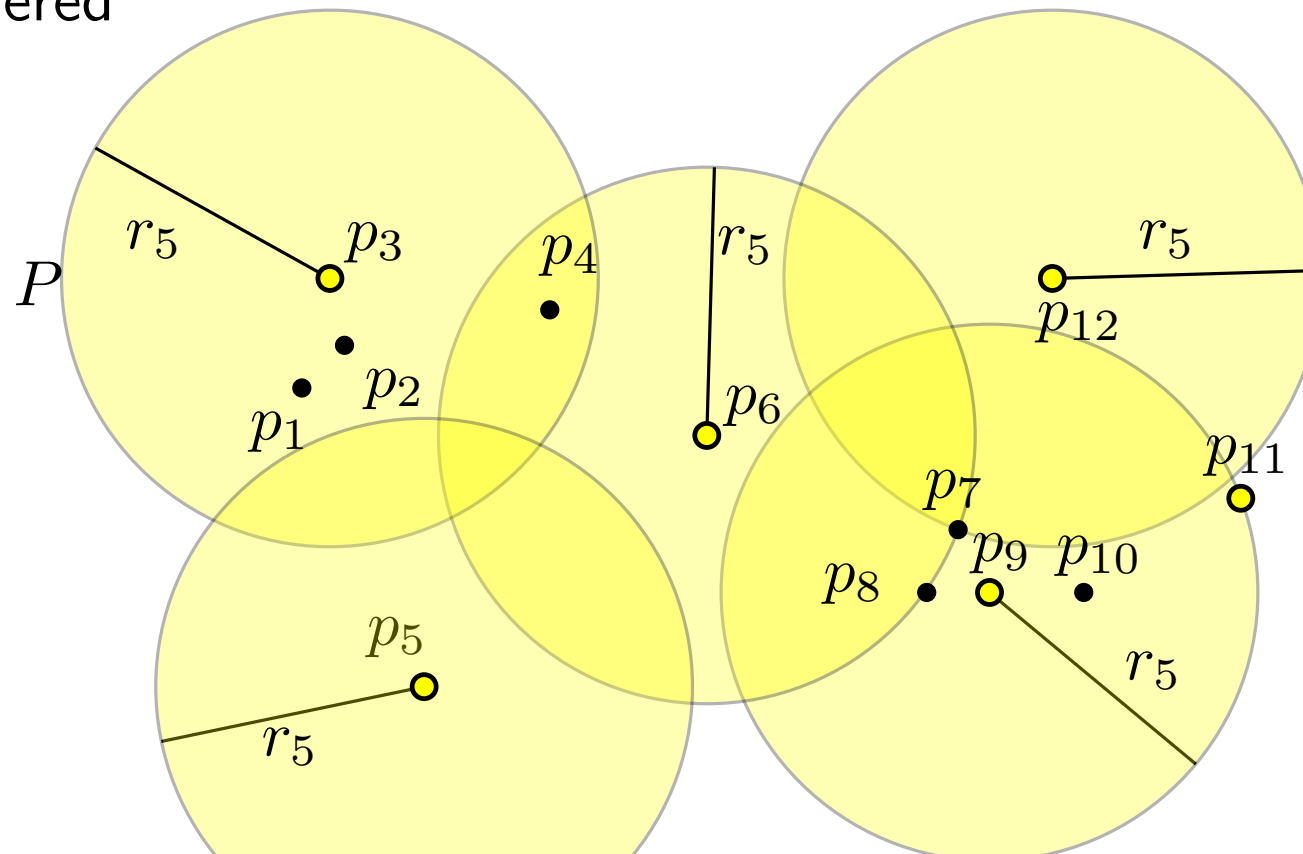


$c_1 = p_9$   
 $c_2 = p_3$   
 $c_3 = p_5$   
 $c_4 = p_6$   
 $c_5 = p_{12}$

"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$

# Gonzales' algorithm for $k$ -center

$\phi(P, \{c_1, \dots, c_t\})$  corresponds to the smallest radius such that all points are covered

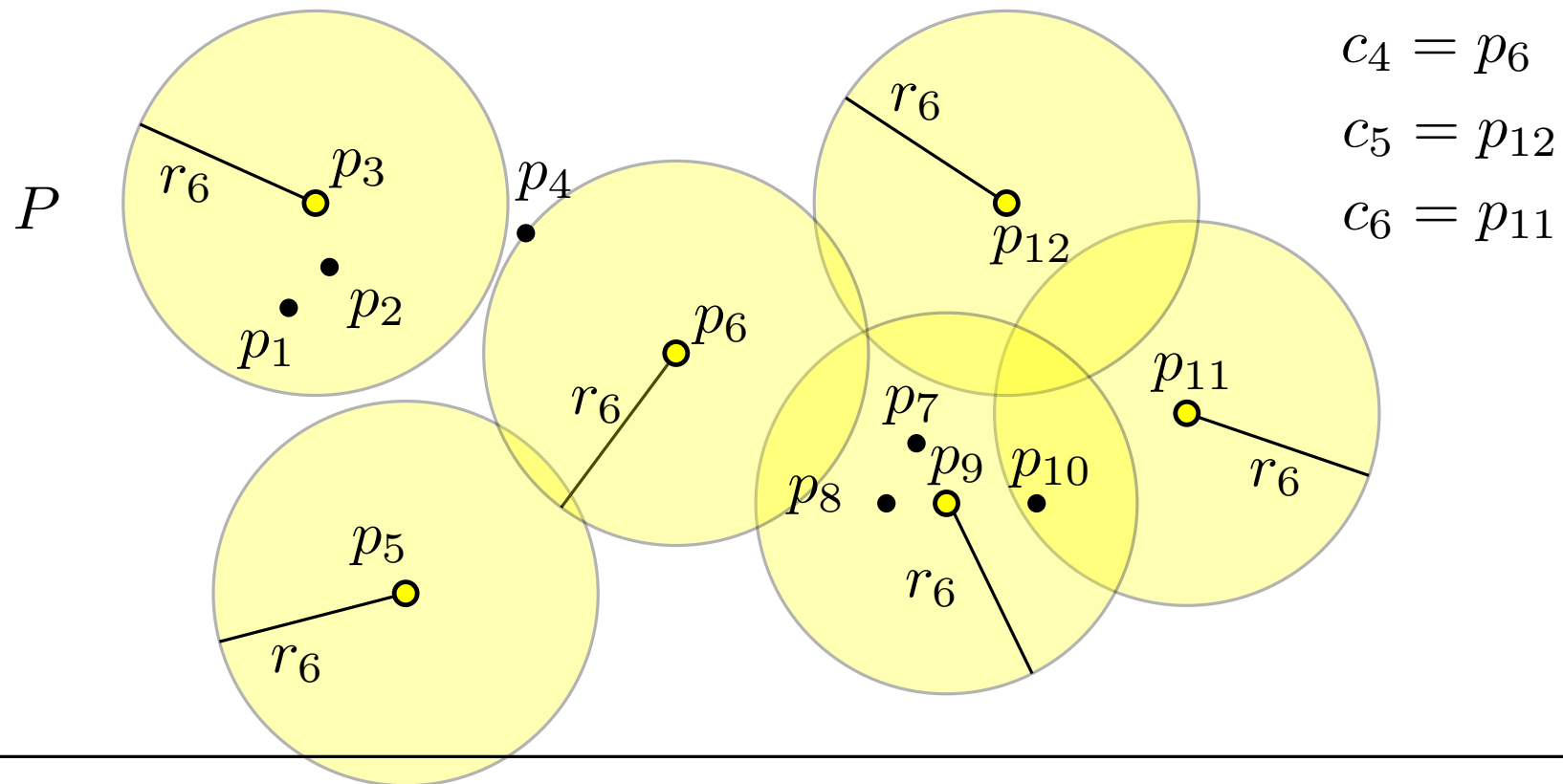


$c_1 = p_9$   
 $c_2 = p_3$   
 $c_3 = p_5$   
 $c_4 = p_6$   
 $c_5 = p_{12}$   
 $c_6 = p_{11}$

"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$

# Gonzales' algorithm for $k$ -center

$\phi(P, \{c_1, \dots, c_t\})$  corresponds to the smallest radius such that all points are covered

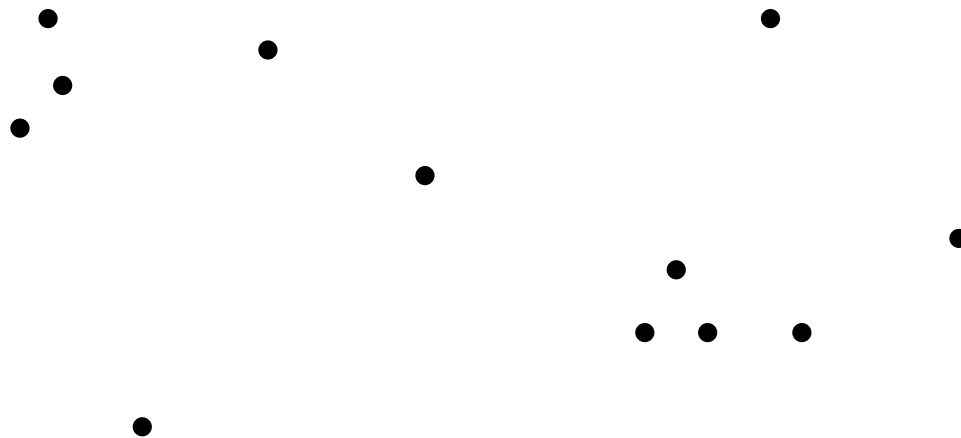


"Farthest-first" greedy algorithm: always choose the point that maximizes the current cost  $\phi(P, \{c_1, \dots, c_t\})$

# Gonzales' algorithm (Analysis)

**Claim:**  $r_k \leq 2\phi(P, C^*)$  (where  $C^*$  is an optimal solution)

**Proof:**



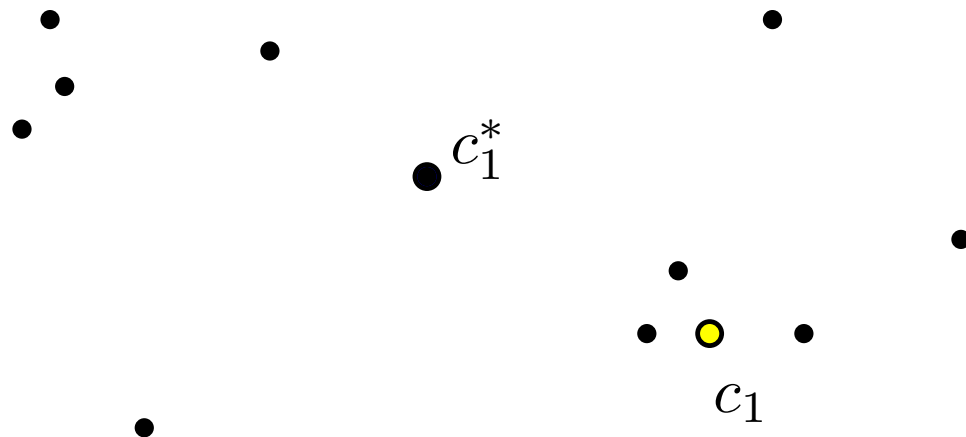


# Gonzales' algorithm (Analysis)

**Claim:**  $r_k \leq 2\phi(P, C^*)$  (where  $C^*$  is an optimal solution)

**Proof:**

( $k = 1$ )

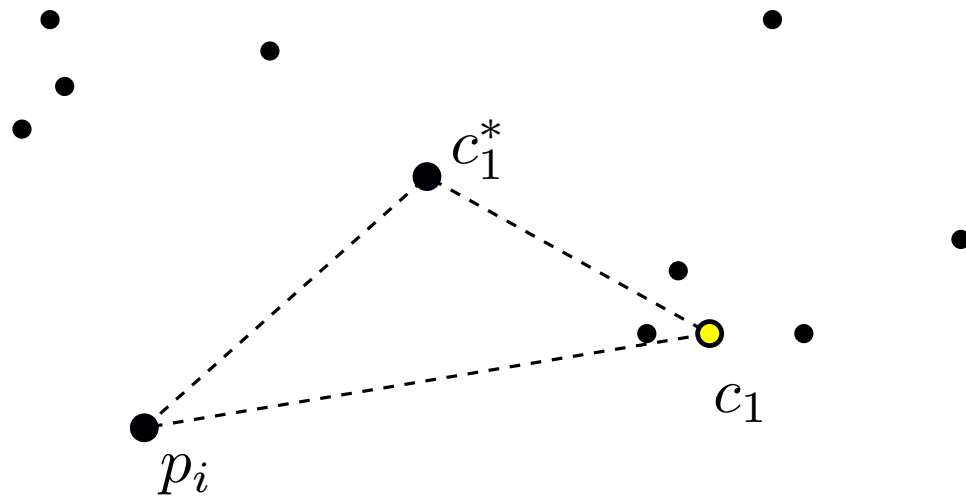


# Gonzales' algorithm (Analysis)

**Claim:**  $r_k \leq 2\phi(P, C^*)$  (where  $C^*$  is an optimal solution)

**Proof:**

( $k = 1$ )



(Triangle inequality)

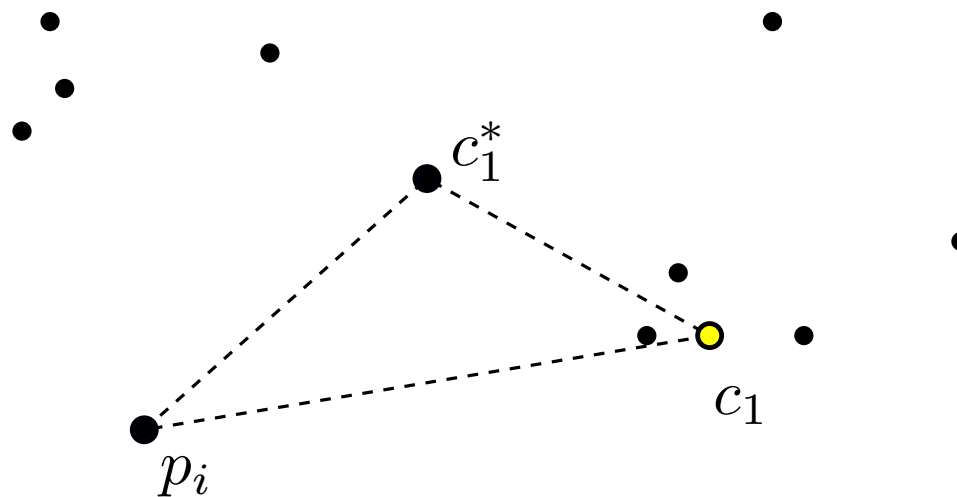
$$\forall p_i \in P : \quad \|p_i - c_1\| \leq \|p_i - c_1^*\| + \|c_1^* - c_1\|$$

# Gonzales' algorithm (Analysis)

**Claim:**  $r_k \leq 2\phi(P, C^*)$  (where  $C^*$  is an optimal solution)

**Proof:**

( $k = 1$ )



(Triangle inequality)

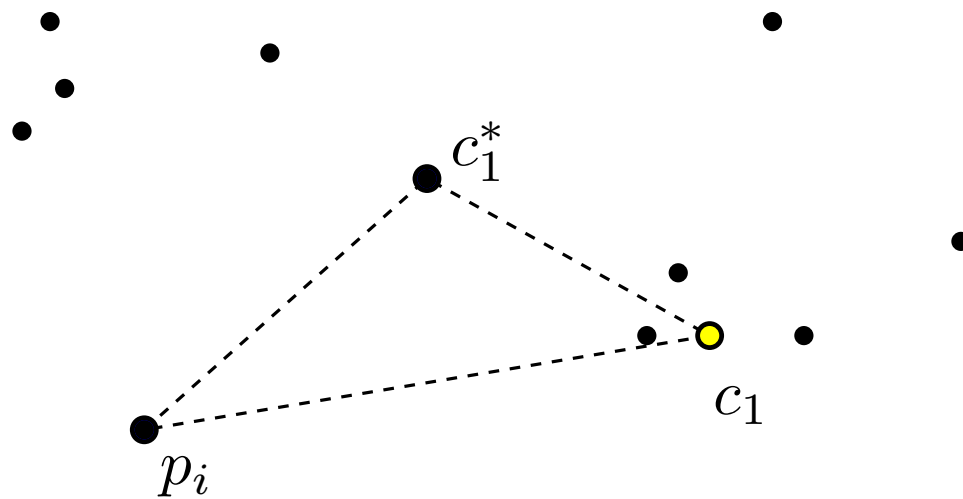
$$\forall p_i \in P : \quad \|p_i - c_1\| \leq \underbrace{\|p_i - c_1^*\|}_{\leq \phi(P, C^*)} + \underbrace{\|c_1^* - c_1\|}_{\leq \phi(P, C^*)}$$

# Gonzales' algorithm (Analysis)

**Claim:**  $r_k \leq 2\phi(P, C^*)$  (where  $C^*$  is an optimal solution)

**Proof:**

( $k = 1$ )



(Triangle inequality)

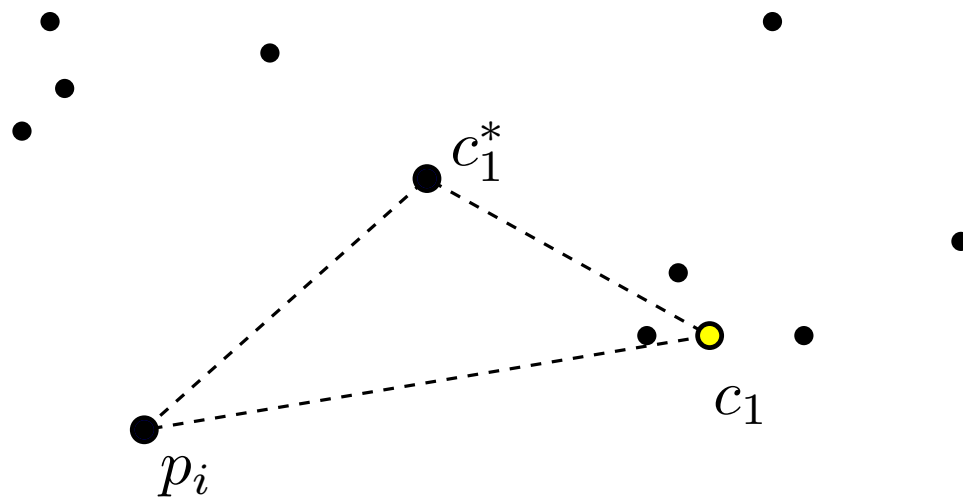
$$\forall p_i \in P : \quad \|p_i - c_1\| \leq \underbrace{\|p_i - c_1^*\|}_{\leq \phi(P, C^*)} + \underbrace{\|c_1^* - c_1\|}_{\leq \phi(P, C^*)} \leq 2\phi(P, C^*)$$

# Gonzales' algorithm (Analysis)

**Claim:**  $r_k \leq 2\phi(P, C^*)$  (where  $C^*$  is an optimal solution)

**Proof:**

( $k = 1$ )



(Triangle inequality)

$$\forall p_i \in P : \quad \|p_i - c_1\| \leq \underbrace{\|p_i - c_1^*\|}_{\leq \phi(P, C^*)} + \underbrace{\|c_1^* - c_1\|}_{\leq \phi(P, C^*)} \leq 2\phi(P, C^*)$$

$$\Rightarrow r_1 \leq 2\phi(P, C^*)$$

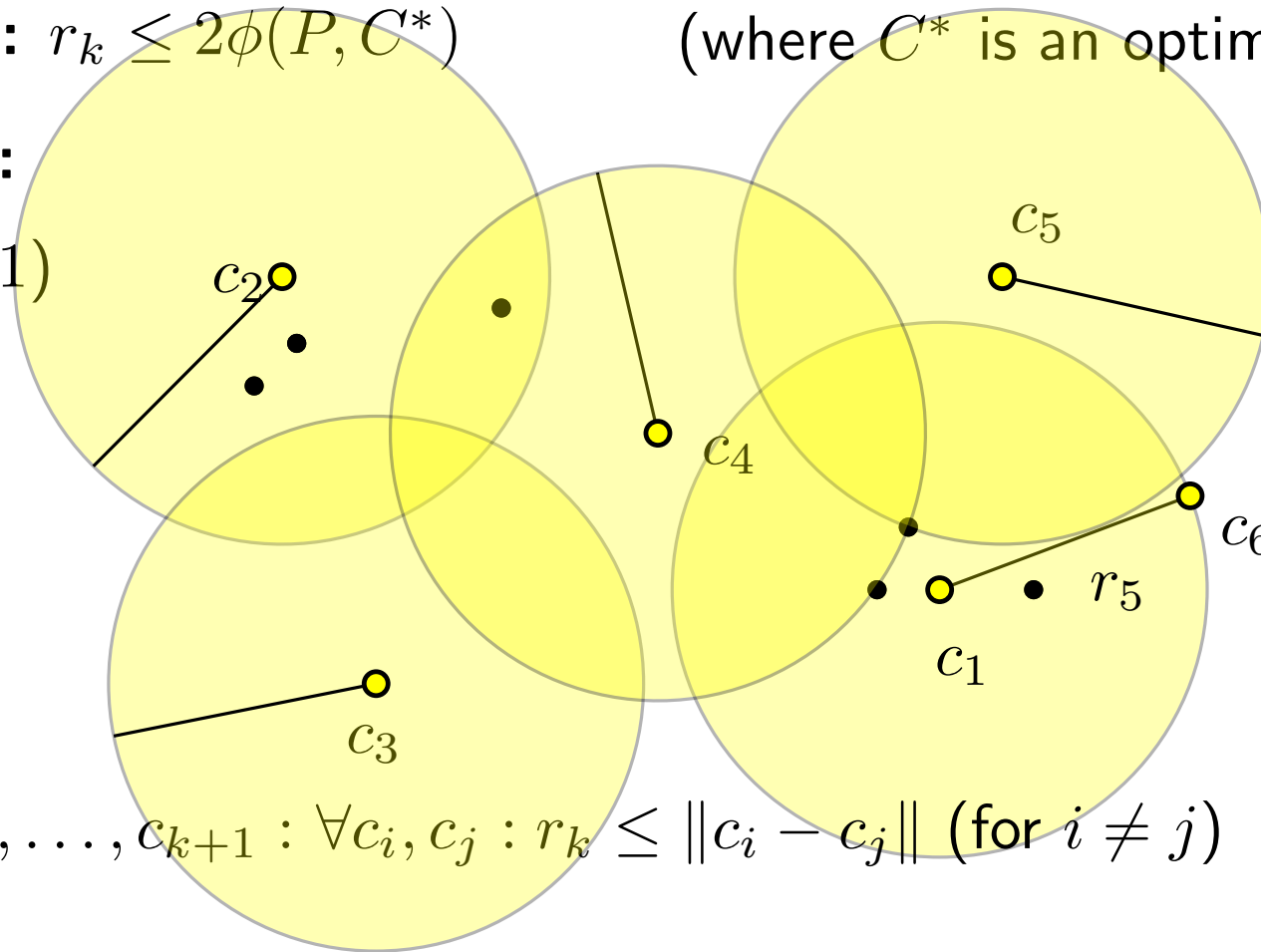
# Gonzales' algorithm (Analysis)

**Claim:**  $r_k \leq 2\phi(P, C^*)$

(where  $C^*$  is an optimal solution)

**Proof:**

$(k \geq 1)$



For  $c_1, \dots, c_{k+1} : \forall c_i, c_j : r_k \leq \|c_i - c_j\|$  (for  $i \neq j$ )

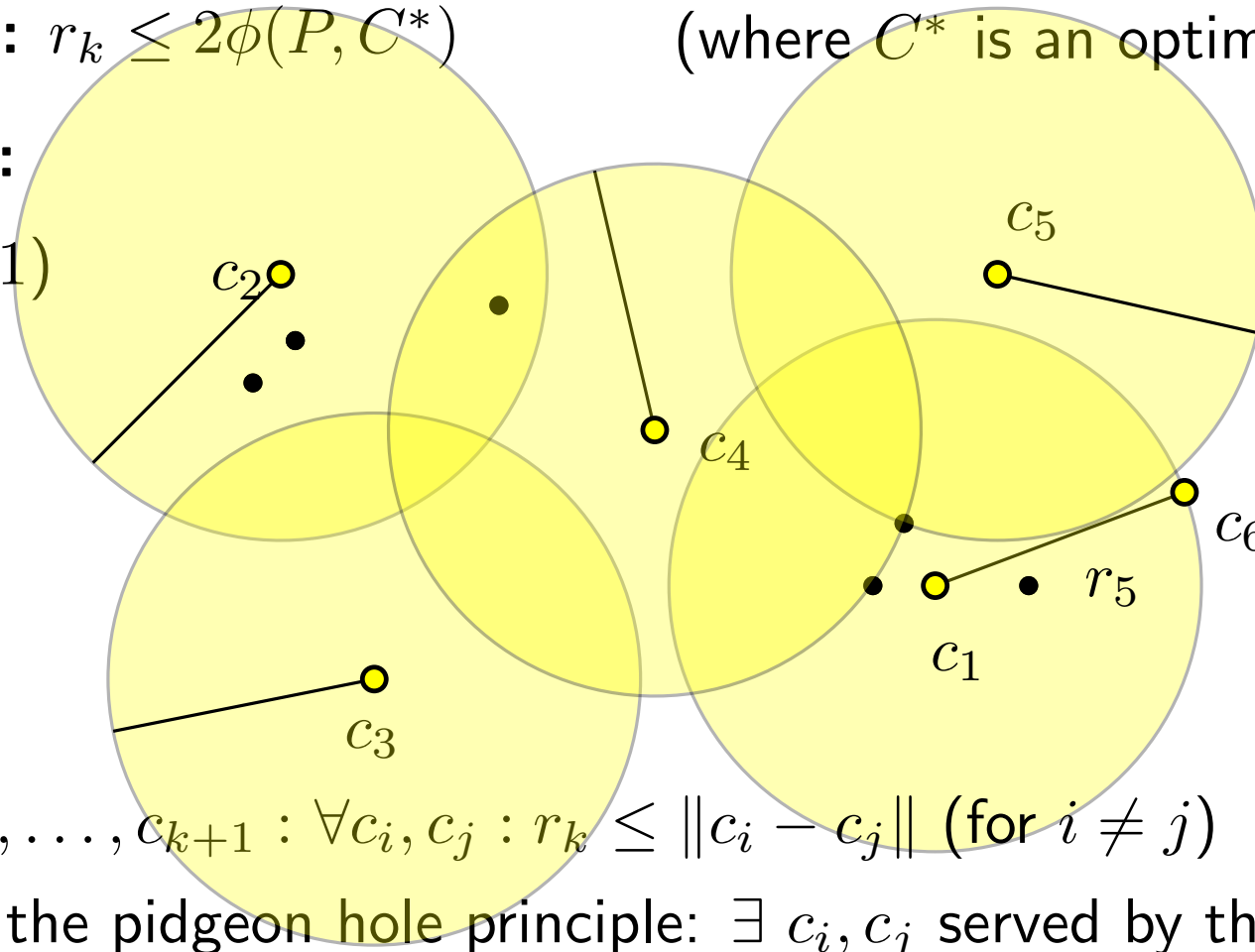
# Gonzales' algorithm (Analysis)

**Claim:**  $r_k \leq 2\phi(P, C^*)$

(where  $C^*$  is an optimal solution)

**Proof:**

$(k \geq 1)$



For  $c_1, \dots, c_{k+1} : \forall c_i, c_j : r_k \leq \|c_i - c_j\|$  (for  $i \neq j$ )

$\Rightarrow$  by the pigeon hole principle:  $\exists c_i, c_j$  served by the same  $c_s^*$

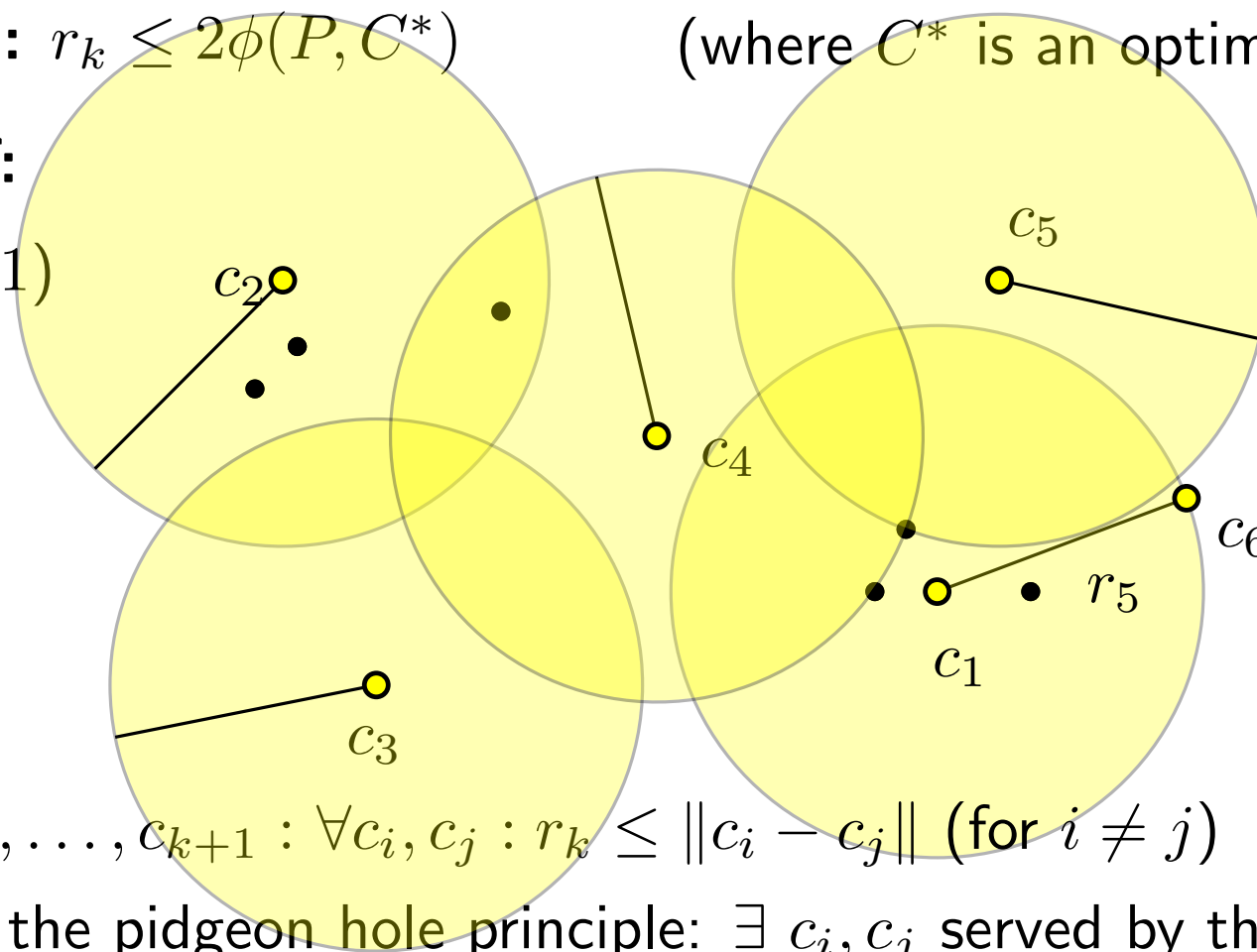
# Gonzales' algorithm (Analysis)

**Claim:**  $r_k \leq 2\phi(P, C^*)$

(where  $C^*$  is an optimal solution)

**Proof:**

( $k \geq 1$ )



For  $c_1, \dots, c_{k+1} : \forall c_i, c_j : r_k \leq \|c_i - c_j\|$  (for  $i \neq j$ )

$\Rightarrow$  by the pigeon hole principle:  $\exists c_i, c_j$  served by the same  $c_s^*$

$\Rightarrow$  by the triangle inequality for  $c_i, c_j, c_s^*$ :

$$\|c_i - c_j\| \leq \|c_i - c_s^*\| + \|c_s^* - c_j\| \leq 2\phi(P, C^*)$$



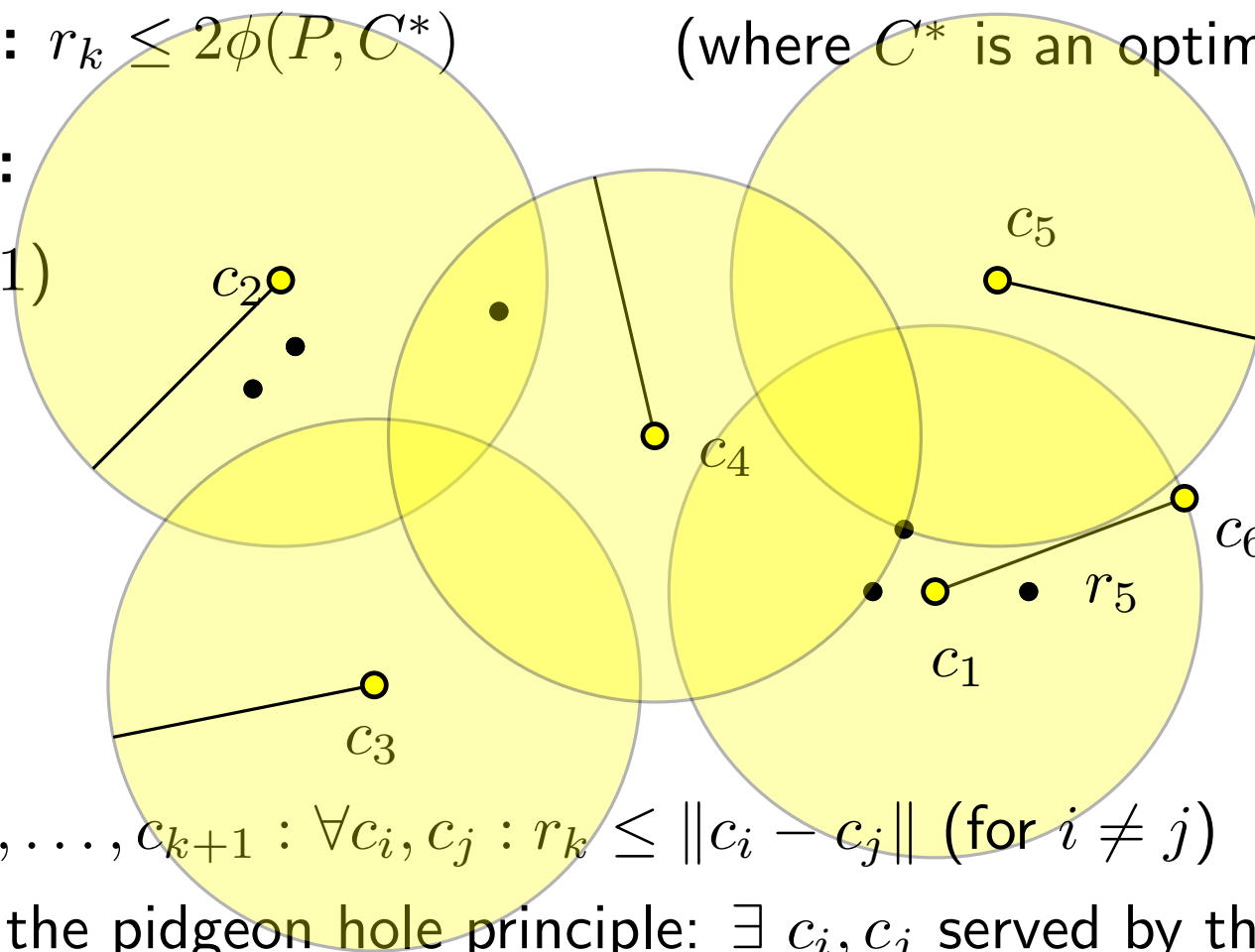
# Gonzales' algorithm (Analysis)

**Claim:**  $r_k \leq 2\phi(P, C^*)$

(where  $C^*$  is an optimal solution)

**Proof:**

( $k \geq 1$ )



For  $c_1, \dots, c_{k+1} : \forall c_i, c_j : r_k \leq \|c_i - c_j\|$  (for  $i \neq j$ )

$\Rightarrow$  by the pigeon hole principle:  $\exists c_i, c_j$  served by the same  $c_s^*$

$\Rightarrow$  by the triangle inequality for  $c_i, c_j, c_s^*$ :

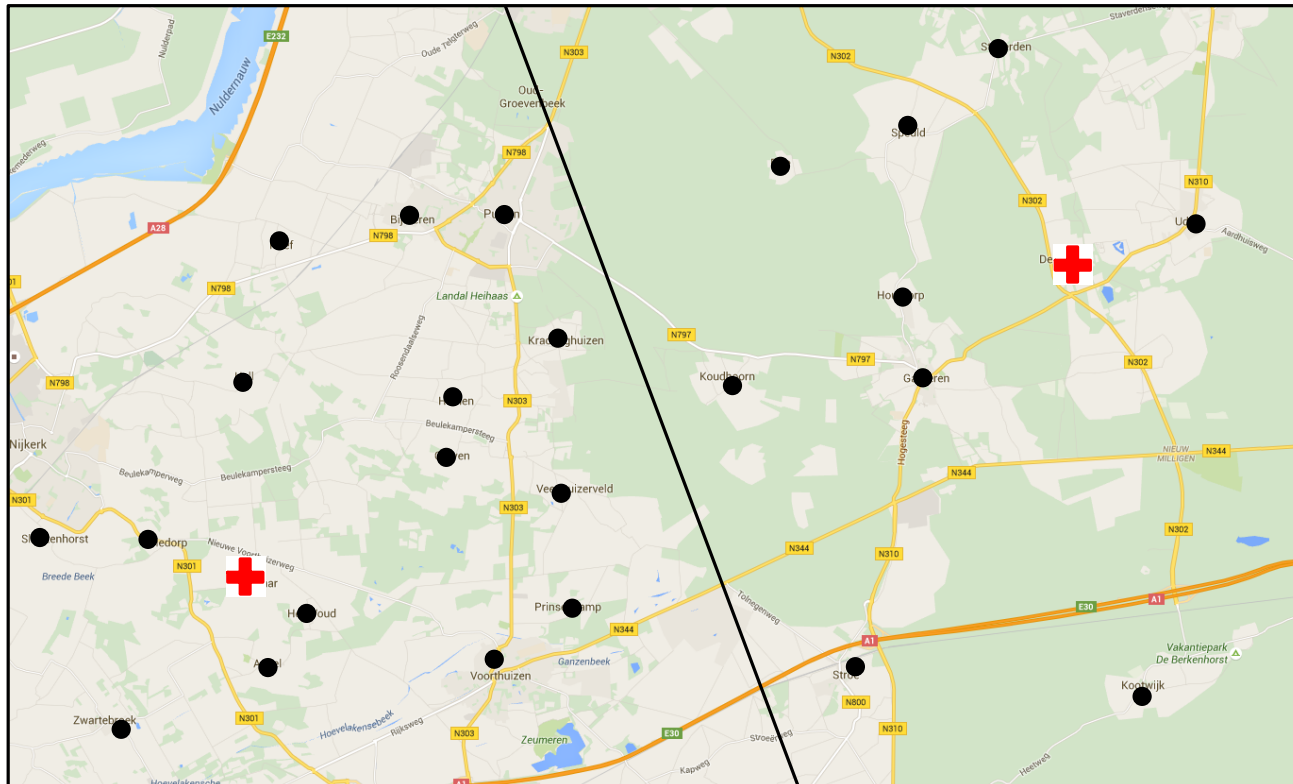
$$\|c_i - c_j\| \leq \|c_i - c_s^*\| + \|c_s^* - c_j\| \leq 2\phi(P, C^*)$$

So we have  $r_k \leq 2\phi(P, C^*)$

## Facility Location (Variant)

You may build two hospitals in two different villages serving the surrounding villages. Where do you place them to minimize the maximal distance from any village to its serving hospital?

**Variant:** • minimize the (squared) average distance

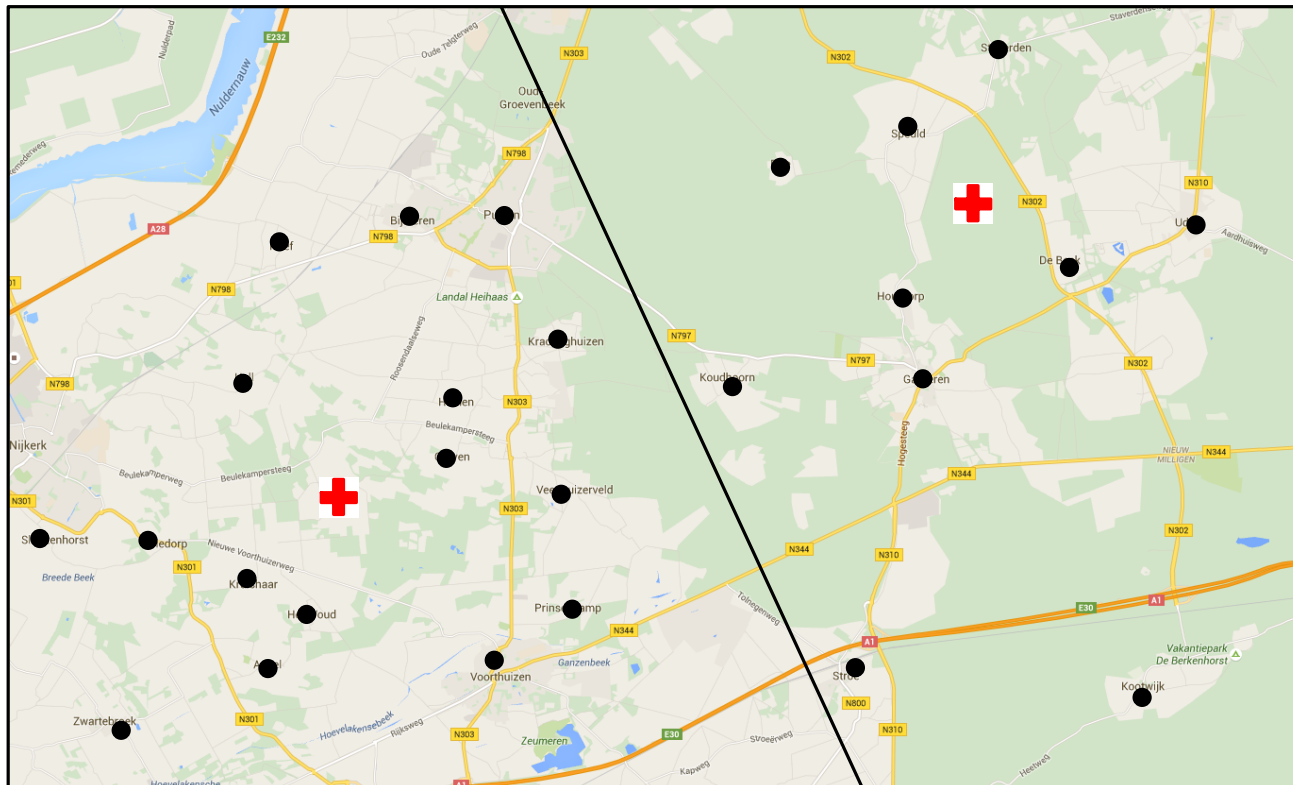


## Facility Location (Variant)

You may build two hospitals in two different villages serving the surrounding villages. Where do you place them to minimize the maximal distance from any village to its serving hospital?

**Variant:**

- minimize the (squared) average distance
- hospitals may be built "in the middle of nowhere"



# k-means clustering

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^d$

## Problem:

centers may be in the middle of nowhere

- each  $p_i \in P$  is associated with its closest center

$$\operatorname{argmin}_{c_j \in C} \|p_i - c_j\|$$

- points associated with a center  $c_j$  together form a "cluster".
- we want to choose  $\{c_1, \dots, c_k\}$  to minimize the cost function

average instead of maximum

$$\phi(P, C) = \sum_{p_i \in P} \left\| p_i - \operatorname{argmin}_{c_j \in C} \|p_i - c_j\| \right\|^2$$

(squared distance)

# Lloyd's algorithm for $k$ -means

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^d$

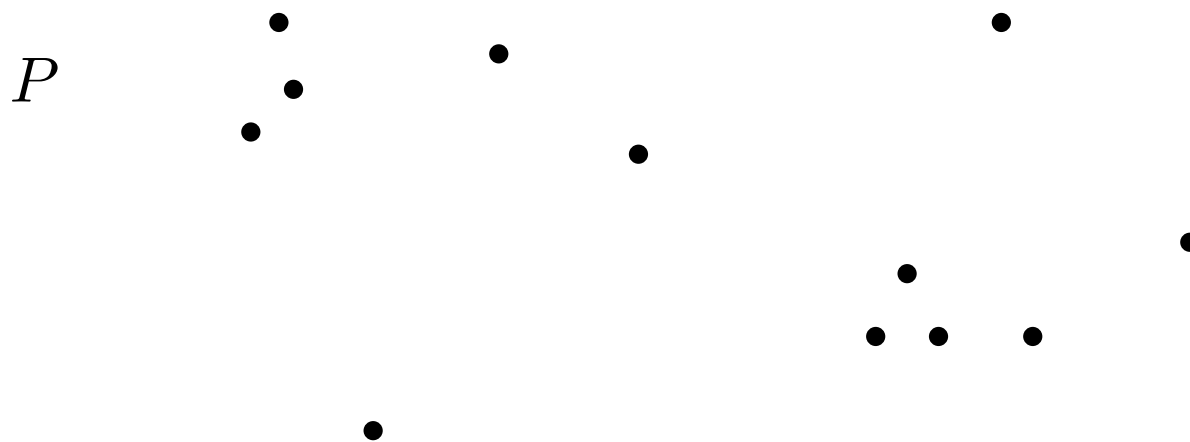
---

## Algorithm:

- choose initial centers arbitrarily  $\{c_1, \dots, c_k\}$  from  $P$
- until  $\{c_1, \dots, c_k\}$  does not change anymore:
  - (1) assign each  $p_i \in P$  to its closest center
  - (2) update center for each cluster  $\Theta_j$ :

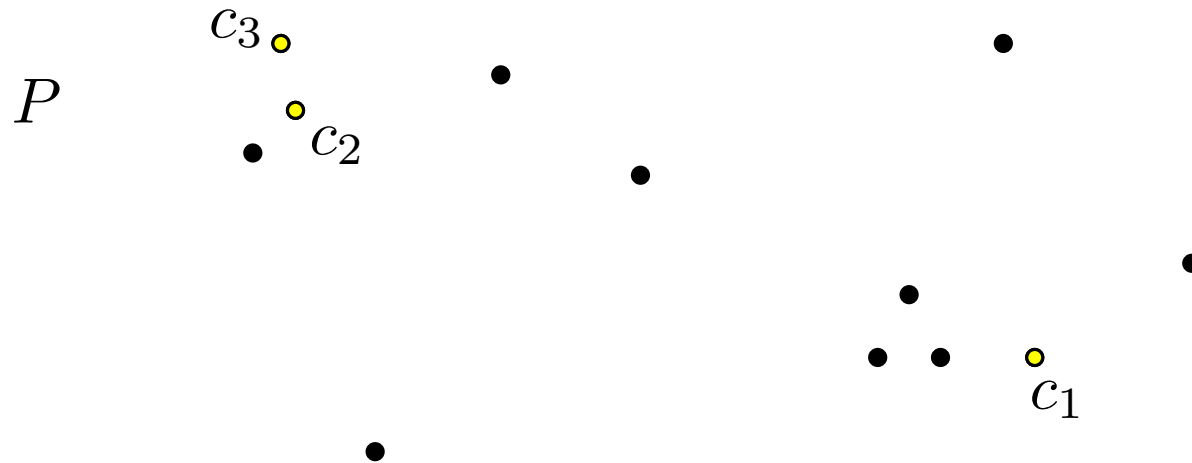
$$c_j := \frac{1}{m} \sum_{p_i \in \Theta_j} p_i$$

# Lloyd's algorithm for $k$ -means



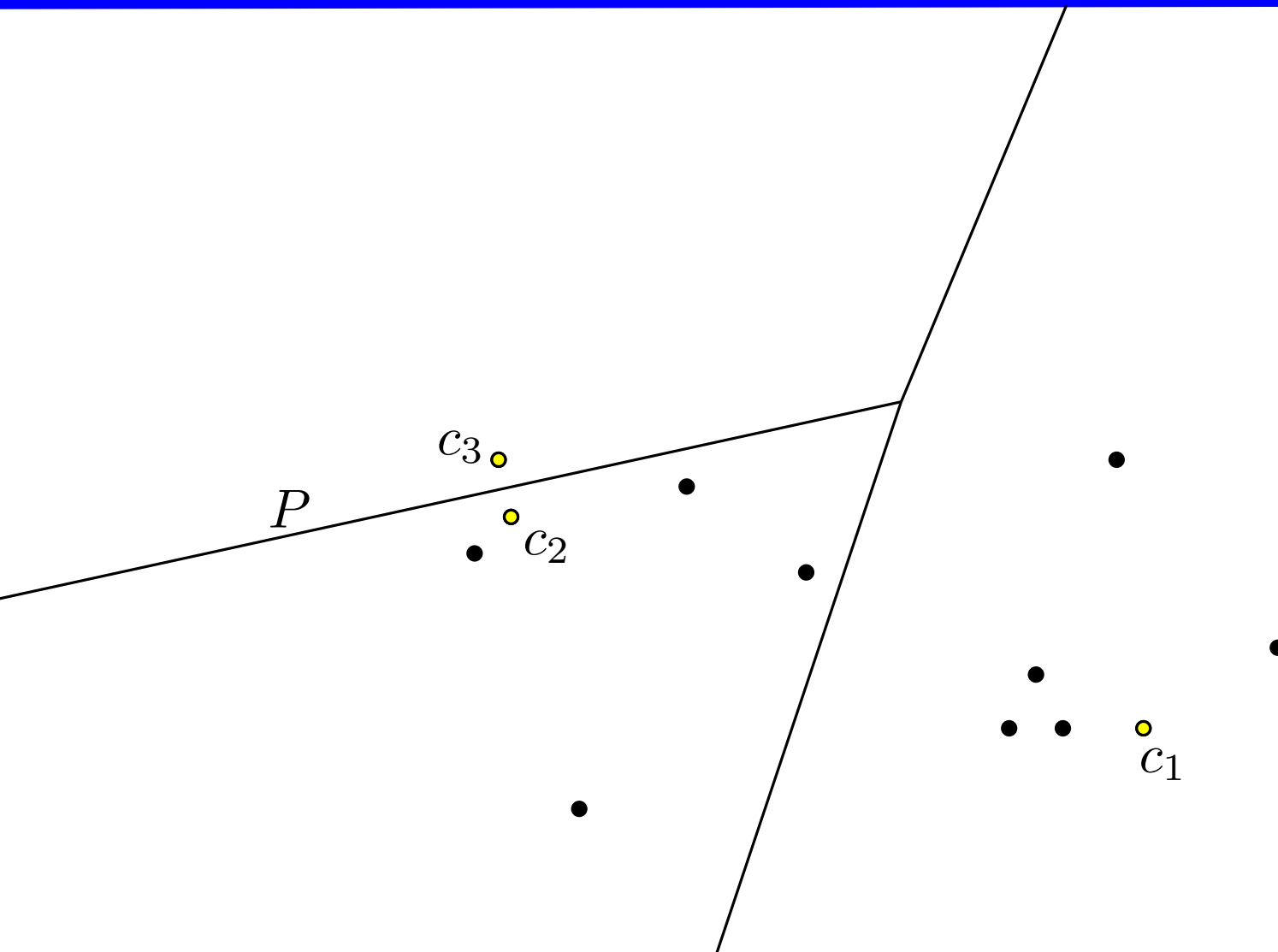
Repeat two steps until convergence:  
(step 1) update assignment (step 2) update centers

# Lloyd's algorithm for $k$ -means



Repeat two steps until convergence:  
(step 1) update assignment (step 2) update centers

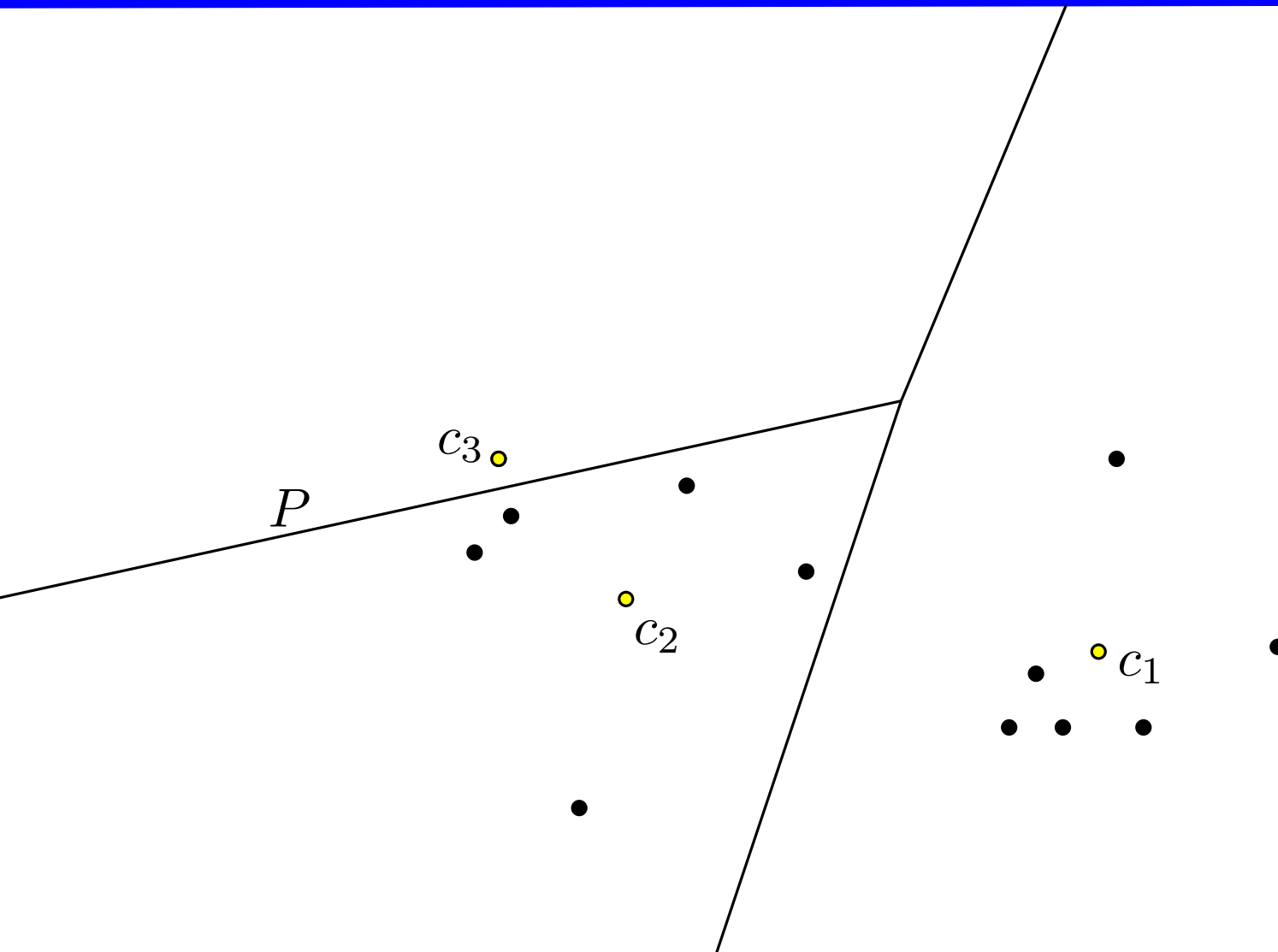
# Lloyd's algorithm for $k$ -means



Repeat two steps until convergence:  
(step 1) update assignment (step 2) update centers

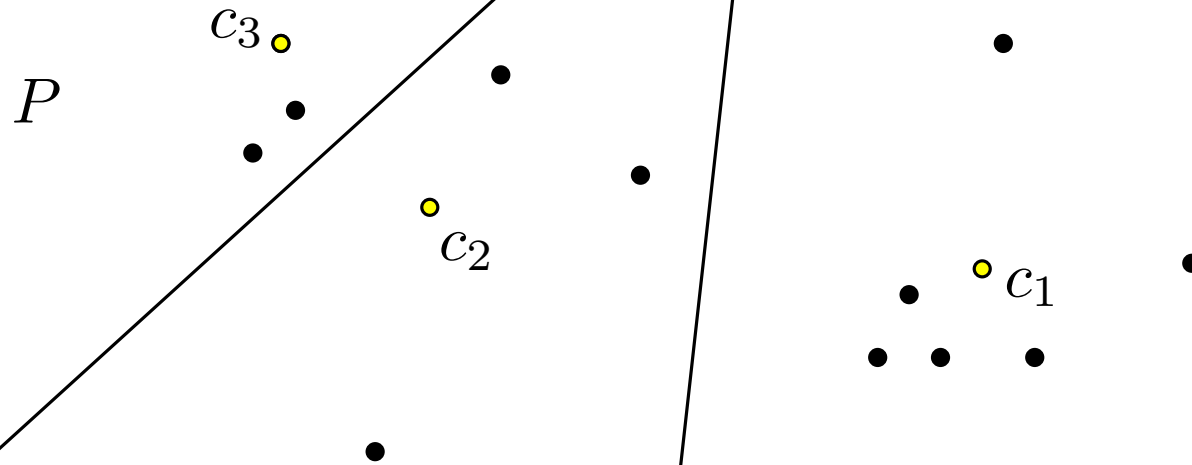


# Lloyd's algorithm for $k$ -means



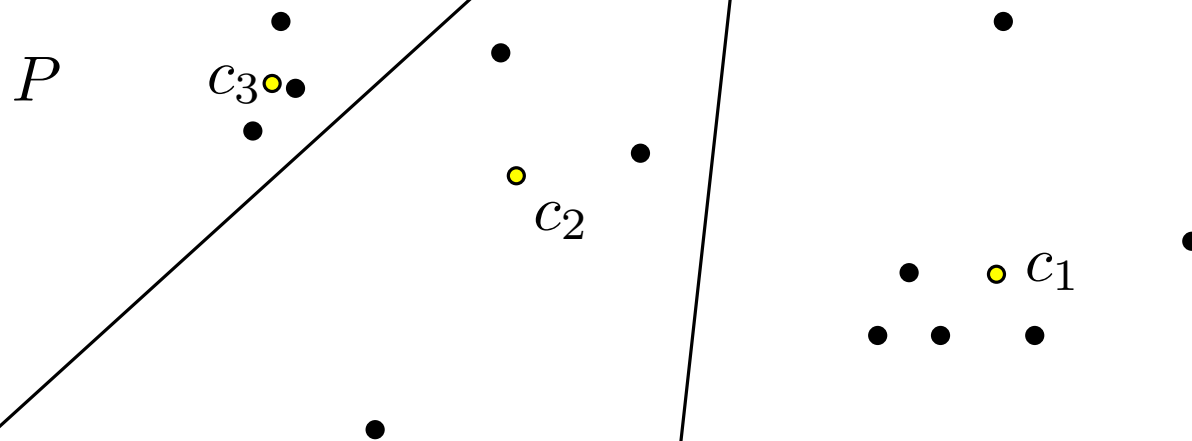
Repeat two steps until convergence:  
(step 1) update assignment (step 2) update centers

# Lloyd's algorithm for $k$ -means



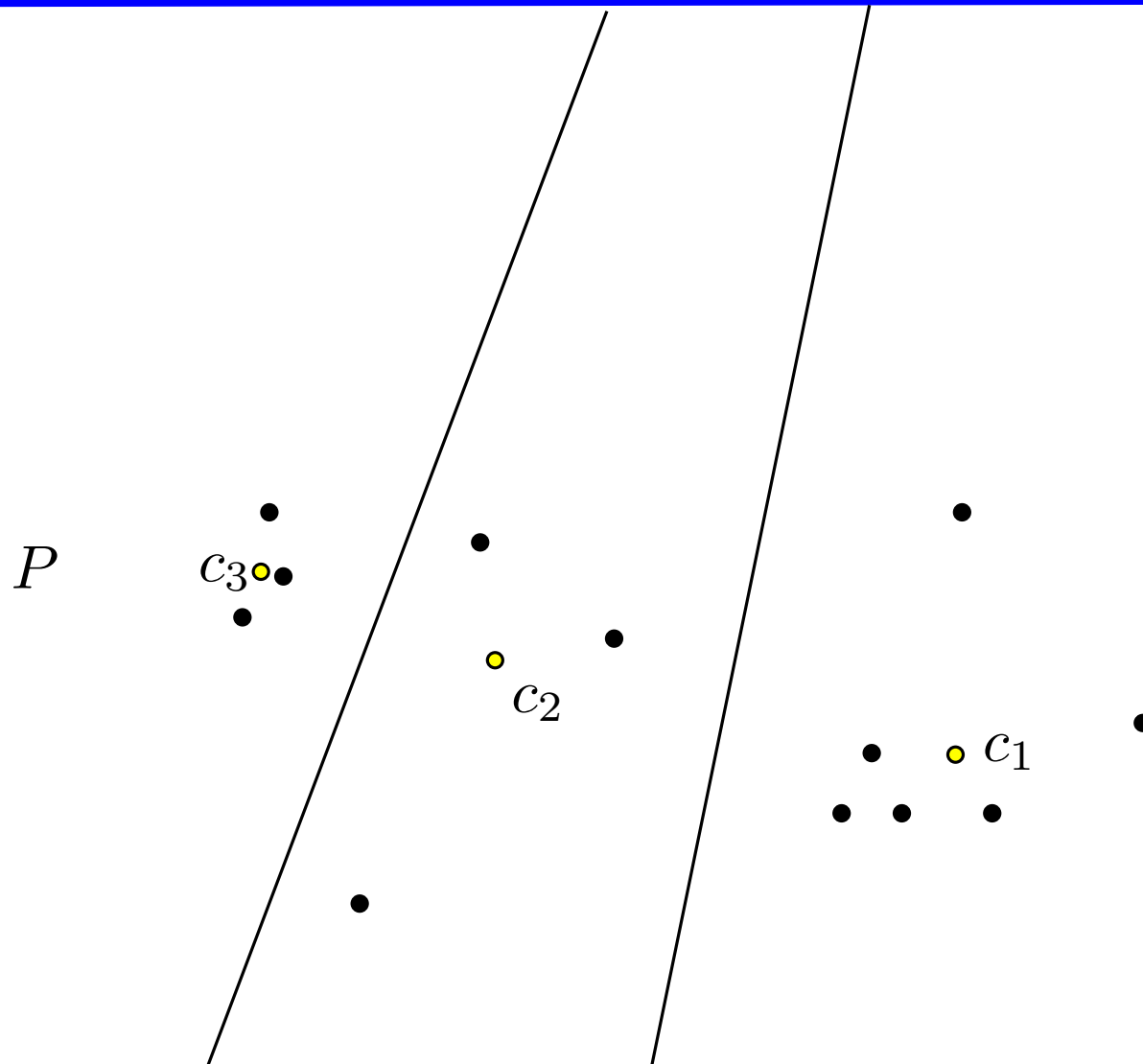
Repeat two steps until convergence:  
(step 1) update assignment (step 2) update centers

# Lloyd's algorithm for $k$ -means



Repeat two steps until convergence:  
(step 1) update assignment (step 2) update centers

# Lloyd's algorithm for $k$ -means



Repeat two steps until convergence:  
(step 1) update assignment (step 2) update centers

# Lloyd's algorithm (Analysis for $k = 1$ )

**Claim:**

$$\operatorname{argmin}_{c \in \mathbb{R}^d} \sum_{p_i \in P} \|p_i - c\|^2 = \frac{1}{n} \sum_{i=1}^n p_i$$

# Lloyd's algorithm (Analysis for $k = 1$ )

**Claim:**

$$\operatorname{argmin}_{c \in \mathbb{R}^d} \sum_{p_i \in P} \|p_i - c\|^2 = \frac{1}{n} \sum_{i=1}^n p_i$$

**Proof:** Let  $\bar{p} := \frac{1}{n} \sum_{i=1}^n p_i$

$$\sum_{i=1}^n \|p_i - c\|^2 = \sum_{i=1}^n \|p_i - \bar{p} + \bar{p} - c\|^2$$

# Lloyd's algorithm (Analysis for $k = 1$ )

**Claim:**

$$\operatorname{argmin}_{c \in \mathbb{R}^d} \sum_{p_i \in P} \|p_i - c\|^2 = \frac{1}{n} \sum_{i=1}^n p_i$$

**Proof:** Let  $\bar{p} := \frac{1}{n} \sum_{i=1}^n p_i$

$$\begin{aligned} \sum_{i=1}^n \|p_i - c\|^2 &= \sum_{i=1}^n \|p_i - \bar{p} + \bar{p} - c\|^2 \\ &= \sum_{i=1}^n \sum_{j=1}^d (p_{ij} - \bar{p}_j + \bar{p}_j - c_j)^2 \end{aligned}$$

# Lloyd's algorithm (Analysis for $k = 1$ )

**Claim:**

$$\operatorname{argmin}_{c \in \mathbb{R}^d} \sum_{p_i \in P} \|p_i - c\|^2 = \frac{1}{n} \sum_{i=1}^n p_i$$

**Proof:**

$$\text{Let } \bar{p} := \frac{1}{n} \sum_{i=1}^n p_i$$

$$\begin{aligned} \sum_{i=1}^n \|p_i - c\|^2 &= \sum_{i=1}^n \|p_i - \bar{p} + \bar{p} - c\|^2 \\ &= \sum_{i=1}^n \sum_{j=1}^d (p_{ij} - \bar{p}_j + \bar{p}_j - c_j)^2 \\ &= \sum_{i=1}^n \sum_{j=1}^d ((p_{ij} - \bar{p}_j)^2 + 2(p_{ij} - \bar{p}_j)(\bar{p}_j - c_j) + (\bar{p}_j - c_j)^2) \end{aligned}$$



# Lloyd's algorithm (Analysis for $k = 1$ )

**Claim:**

$$\operatorname{argmin}_{c \in \mathbb{R}^d} \sum_{p_i \in P} \|p_i - c\|^2 = \frac{1}{n} \sum_{i=1}^n p_i$$

**Proof:** Let  $\bar{p} := \frac{1}{n} \sum_{i=1}^n p_i$

$$\begin{aligned} \sum_{i=1}^n \|p_i - c\|^2 &= \sum_{i=1}^n \|p_i - \bar{p} + \bar{p} - c\|^2 \\ &= \sum_{i=1}^n \sum_{j=1}^d (p_{ij} - \bar{p}_j + \bar{p}_j - c_j)^2 \\ &= \sum_{i=1}^n \sum_{j=1}^d ((p_{ij} - \bar{p}_j)^2 + 2(p_{ij} - \bar{p}_j)(\bar{p}_j - c_j) + (\bar{p}_j - c_j)^2) \\ &= \sum_{i=1}^n \|p_i - \bar{p}\|^2 + 2 \sum_{i=1}^n \langle p_i - \bar{p}, \bar{p} - c \rangle + n \cdot \|\bar{p} - c\|^2 \end{aligned}$$

# Lloyd's algorithm (Analysis for $k = 1$ )

**Claim:**

$$\operatorname{argmin}_{c \in \mathbb{R}^d} \sum_{p_i \in P} \|p_i - c\|^2 = \frac{1}{n} \sum_{i=1}^n p_i$$

**Proof:**

$$\text{Let } \bar{p} := \frac{1}{n} \sum_{i=1}^n p_i$$

$$\begin{aligned} \sum_{i=1}^n \|p_i - c\|^2 &= \sum_{i=1}^n \|p_i - \bar{p} + \bar{p} - c\|^2 \\ &= \sum_{i=1}^n \sum_{j=1}^d (p_{ij} - \bar{p}_j + \bar{p}_j - c_j)^2 \\ &= \sum_{i=1}^n \sum_{j=1}^d ((p_{ij} - \bar{p}_j)^2 + 2(p_{ij} - \bar{p}_j)(\bar{p}_j - c_j) + (\bar{p}_j - c_j)^2) \\ &= \sum_{i=1}^n \|p_i - \bar{p}\|^2 + 2 \sum_{i=1}^n \langle p_i - \bar{p}, \bar{p} - c \rangle + n \cdot \|\bar{p} - c\|^2 \\ &= \sum_{i=1}^n \|p_i - \bar{p}\|^2 + 2 \left\langle \sum_{i=1}^n p_i - n \cdot \bar{p}, \bar{p} - c \right\rangle + n \cdot \|\bar{p} - c\|^2 \end{aligned}$$

# Lloyd's algorithm (Analysis for $k = 1$ )

**Claim:**

$$\operatorname{argmin}_{c \in \mathbb{R}^d} \sum_{p_i \in P} \|p_i - c\|^2 = \frac{1}{n} \sum_{i=1}^n p_i$$

**Proof:**

$$\text{Let } \bar{p} := \frac{1}{n} \sum_{i=1}^n p_i$$

$$\begin{aligned} \sum_{i=1}^n \|p_i - c\|^2 &= \sum_{i=1}^n \|p_i - \bar{p} + \bar{p} - c\|^2 \\ &= \sum_{i=1}^n \sum_{j=1}^d (p_{ij} - \bar{p}_j + \bar{p}_j - c_j)^2 \\ &= \sum_{i=1}^n \sum_{j=1}^d ((p_{ij} - \bar{p}_j)^2 + 2(p_{ij} - \bar{p}_j)(\bar{p}_j - c_j) + (\bar{p}_j - c_j)^2) \\ &= \sum_{i=1}^n \|p_i - \bar{p}\|^2 + 2 \sum_{i=1}^n \langle p_i - \bar{p}, \bar{p} - c \rangle + n \cdot \|\bar{p} - c\|^2 \\ &= \sum_{i=1}^n \|p_i - \bar{p}\|^2 + 2 \left\langle \underbrace{\sum_{i=1}^n p_i - n \cdot \bar{p}}_0, \bar{p} - c \right\rangle + n \cdot \|\bar{p} - c\|^2 \end{aligned}$$

# Lloyd's algorithm (Analysis for $k = 1$ )

**Claim:**

$$\operatorname{argmin}_{c \in \mathbb{R}^d} \sum_{p_i \in P} \|p_i - c\|^2 = \frac{1}{n} \sum_{i=1}^n p_i$$

**Proof:** (continued)

$$\sum_{i=1}^n \|p_i - c\|^2 = \sum_{i=1}^n \|p_i - \bar{p}\|^2 + n \cdot \|\bar{p} - c\|^2$$

# Lloyd's algorithm (Analysis for $k = 1$ )

**Claim:**

$$\operatorname{argmin}_{c \in \mathbb{R}^d} \sum_{p_i \in P} \|p_i - c\|^2 = \frac{1}{n} \sum_{i=1}^n p_i$$

**Proof:** (continued)

$$\sum_{i=1}^n \|p_i - c\|^2 = \sum_{i=1}^n \|p_i - \bar{p}\|^2 + n \cdot \|\bar{p} - c\|^2$$

right hand side is minimized for  $c = \bar{p}$



# Lloyd's algorithm (Analysis)

Does the algorithm always terminate?

# Lloyd's algorithm (Analysis)

Does the algorithm always terminate?

Yes, because:

- each step decreases the cost  $\phi(P, C)$
- there exists only a finite number of cluster assignments

# Lloyd's algorithm (Analysis)

Does the algorithm always terminate?

Yes, because:

- each step decreases the cost  $\phi(P, C)$
- there exists only a finite number of cluster assignments

Does it always converge to an optimal solution?



# Lloyd's algorithm (Analysis)

Does the algorithm always terminate?

Yes, because:

- each step decreases the cost  $\phi(P, C)$
- there exists only a finite number of cluster assignments

Does it always converge to an optimal solution?

No, can get stuck in local minima.

$p_1$  •  
 $p_2$  •

•  $p_3$   
•  $p_4$

# Lloyd's algorithm (Analysis)

Does the algorithm always terminate?

Yes, because:

- each step decreases the cost  $\phi(P, C)$
- there exists only a finite number of cluster assignments

Does it always converge to an optimal solution?

No, can get stuck in local minima.

$p_1$   $c_1$

$p_3$

$p_2$   $c_2$

$p_4$

# Lloyd's algorithm (Analysis)

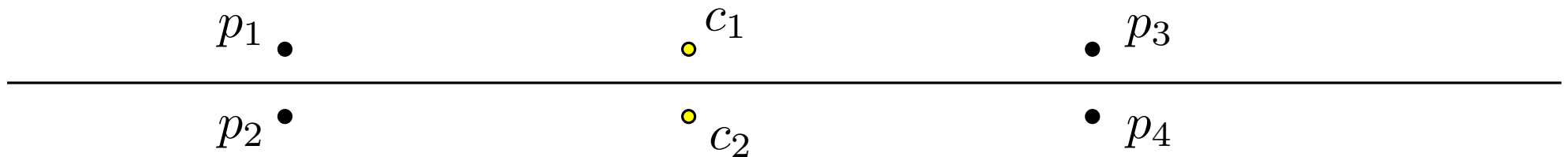
Does the algorithm always terminate?

Yes, because:

- each step decreases the cost  $\phi(P, C)$
- there exists only a finite number of cluster assignments

Does it always converge to an optimal solution?

No, can get stuck in local minima.



# k-means++ Algorithm

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^2$

---

## Algorithm:

- choose  $c_1$  uniformly at random from  $P$
- for  $t = 2, \dots, k$ : choose  $c_t = p_i$  with probability  $\alpha_i$

$$\alpha_i := \frac{\left\| p_i - \operatorname{argmin}_{c_j \in \{c_1, \dots, c_{t-1}\}} \|p_i - c_j\| \right\|^2}{\phi(P, \{c_1, \dots, c_{t-1}\})}$$

new

# k-means++ Algorithm

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^2$

## Algorithm:

new

- choose  $c_1$  uniformly at random from  $P$
- for  $t = 2, \dots, k$ : choose  $c_t = p_i$  with probability  $\alpha_i$

$$\alpha_i := \frac{\left\| p_i - \operatorname{argmin}_{c_j \in \{c_1, \dots, c_{t-1}\}} \|p_i - c_j\| \right\|^2}{\phi(P, \{c_1, \dots, c_{t-1}\})}$$

as before

- until  $\{c_1, \dots, c_k\}$  does not change anymore:
  - (1) assign each  $p_i \in P$  to its closest center
  - (2) update center for each cluster  $\Theta_j$ :

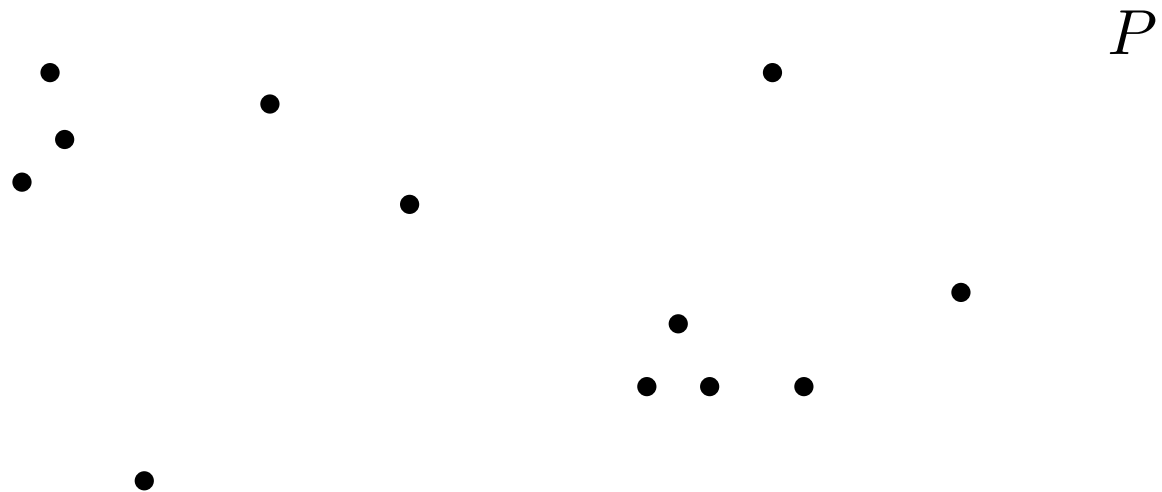
$$c_j := \frac{1}{m} \sum_{p_i \in \Theta_j} p_i$$

# k-means++ Algorithm

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^2$

---

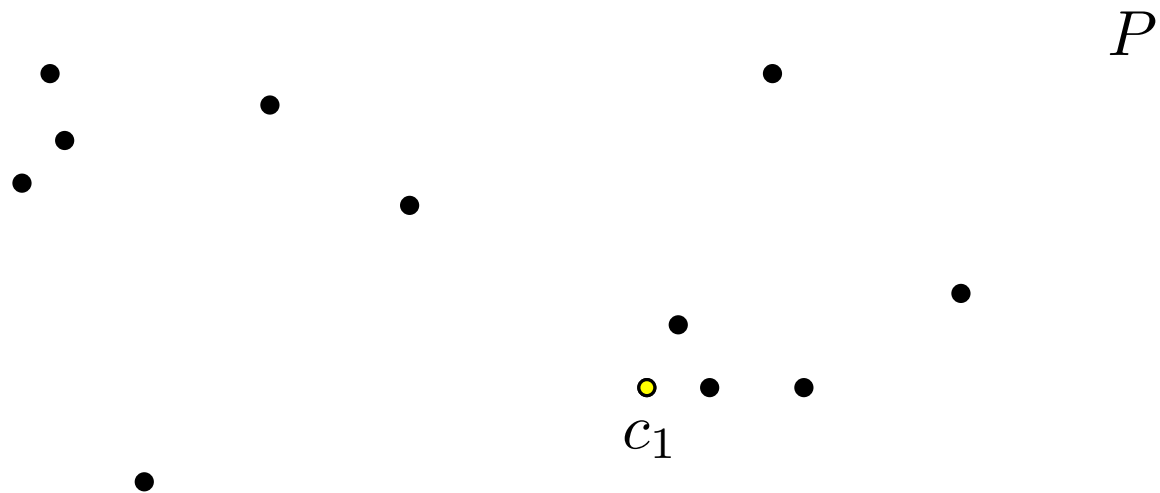


# k-means++ Algorithm

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^2$

---

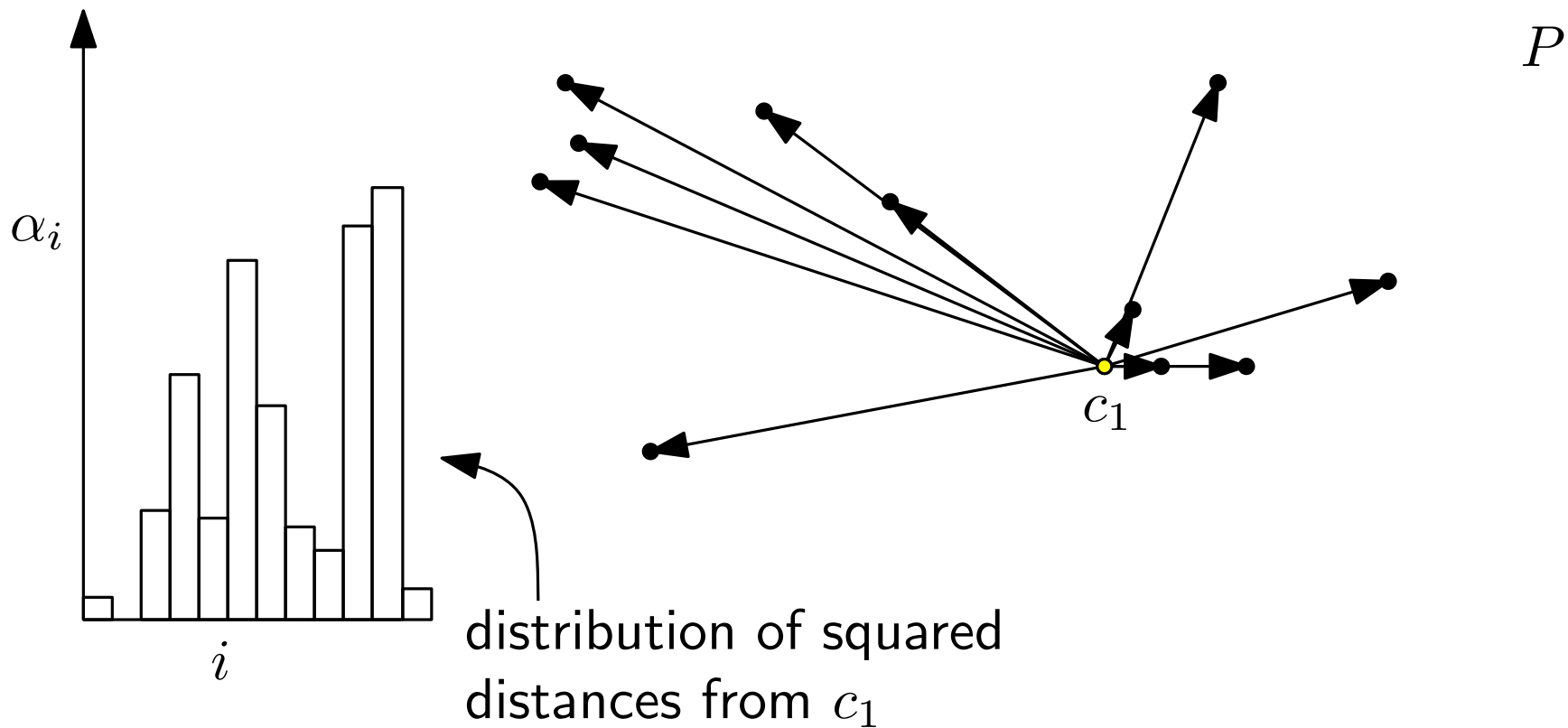


# k-means++ Algorithm

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^2$

---



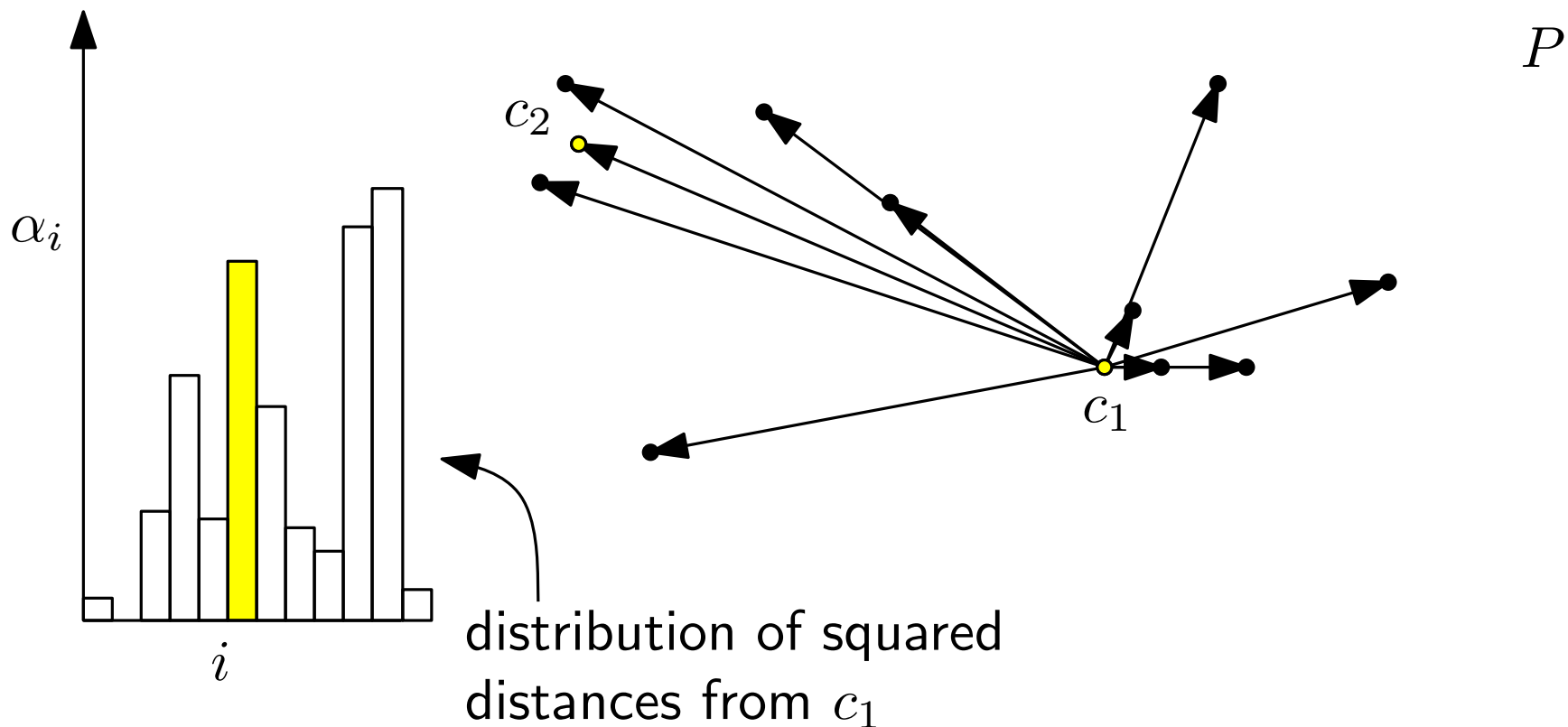


# k-means++ Algorithm

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^2$

random sample

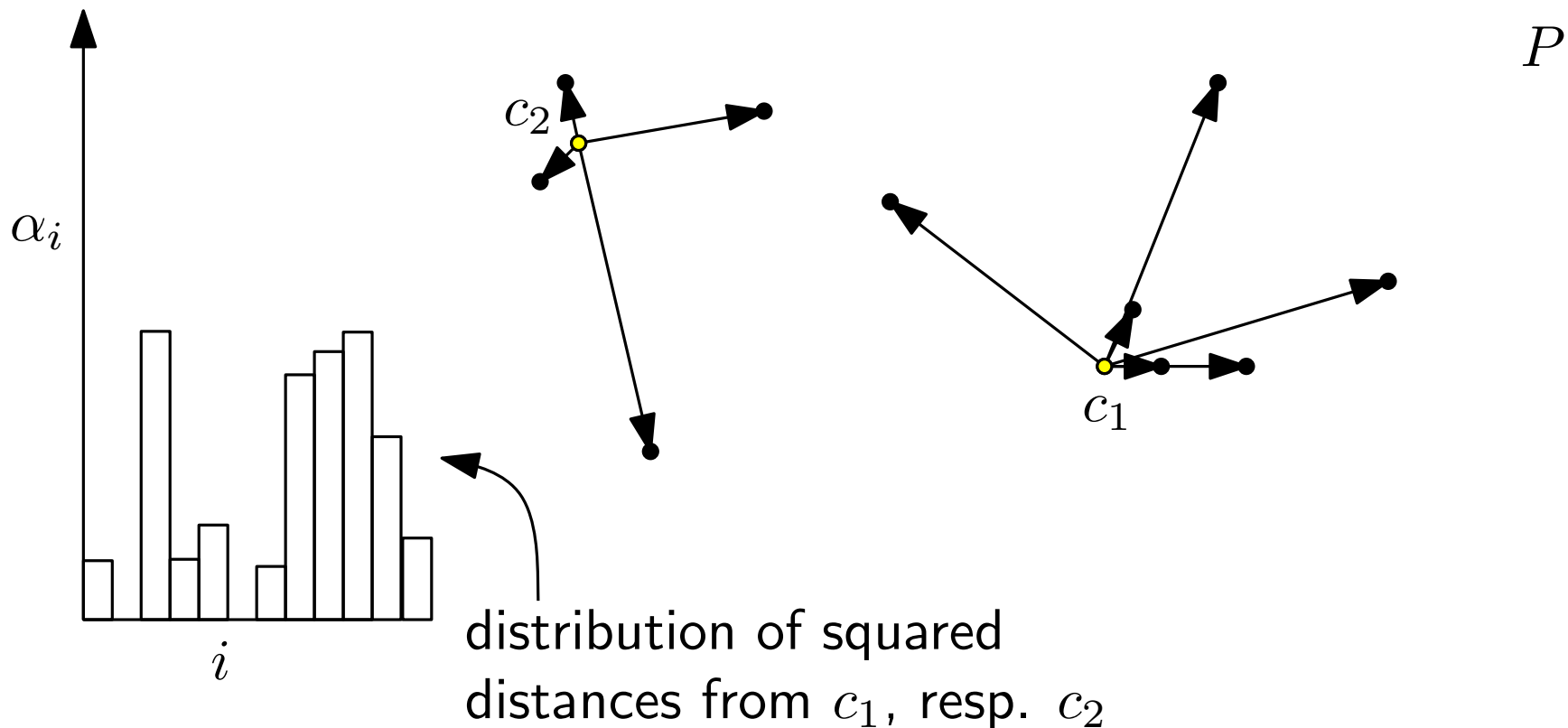


# k-means++ Algorithm

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^2$

---

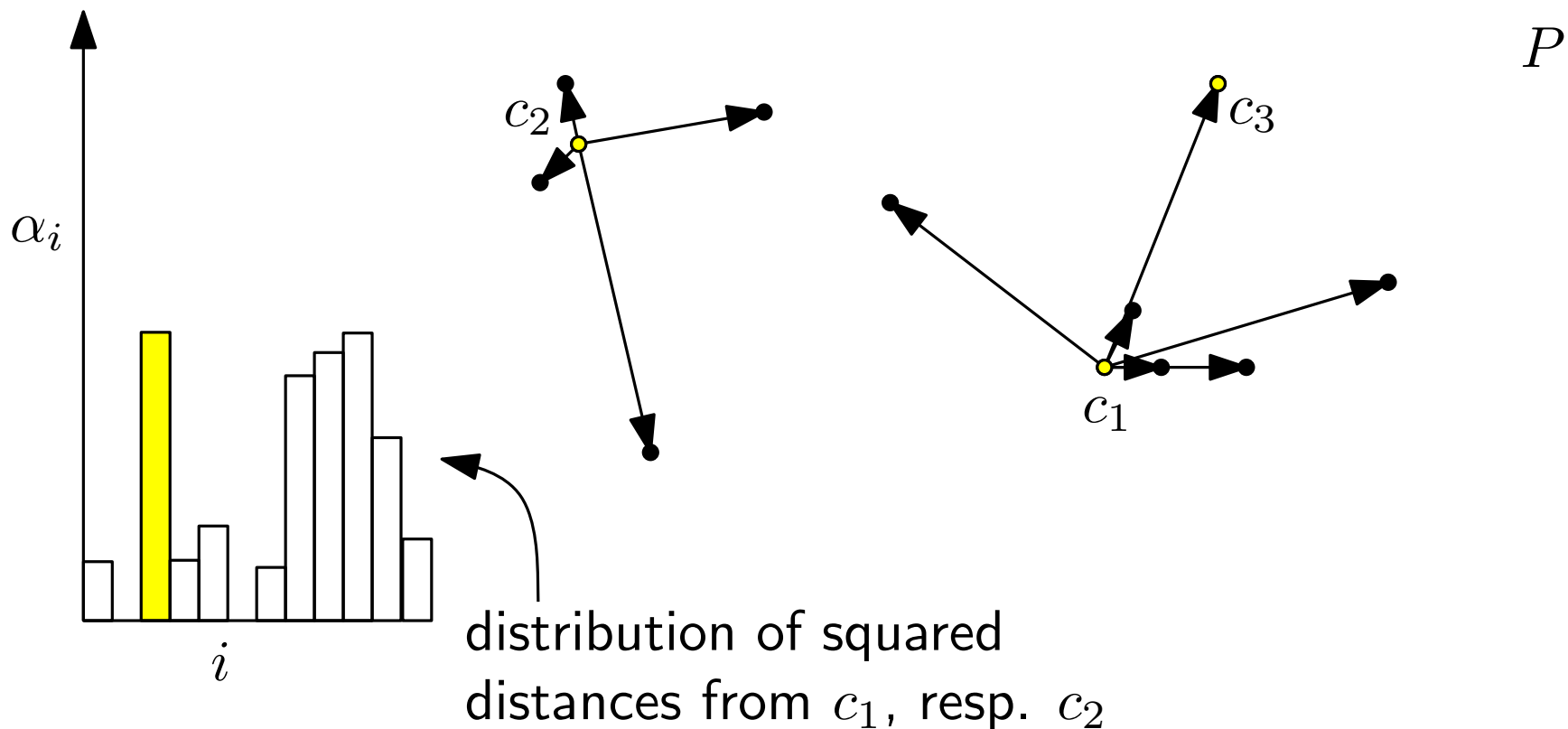


# k-means++ Algorithm

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^2$

random sample

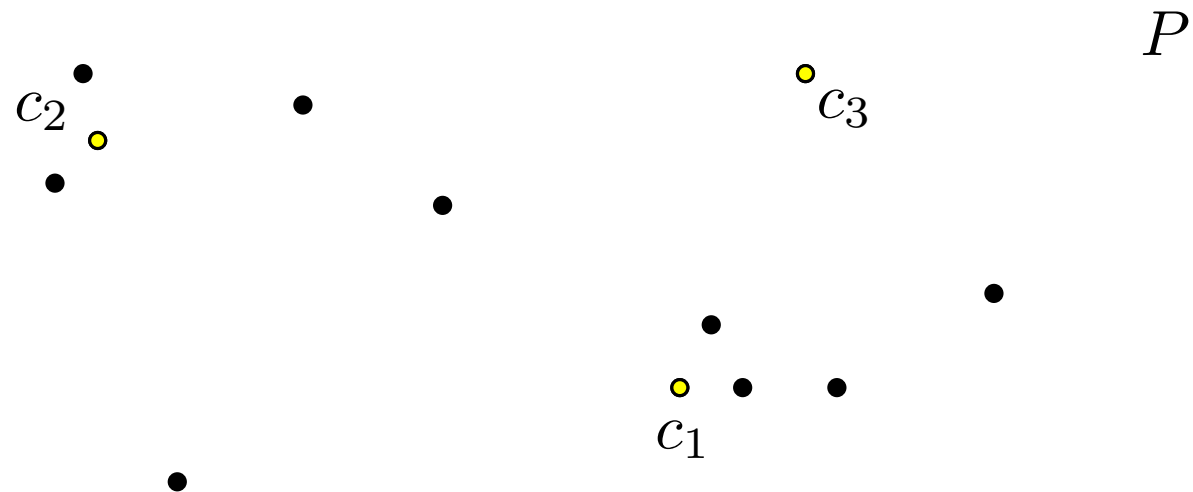


# k-means++ Algorithm

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^2$

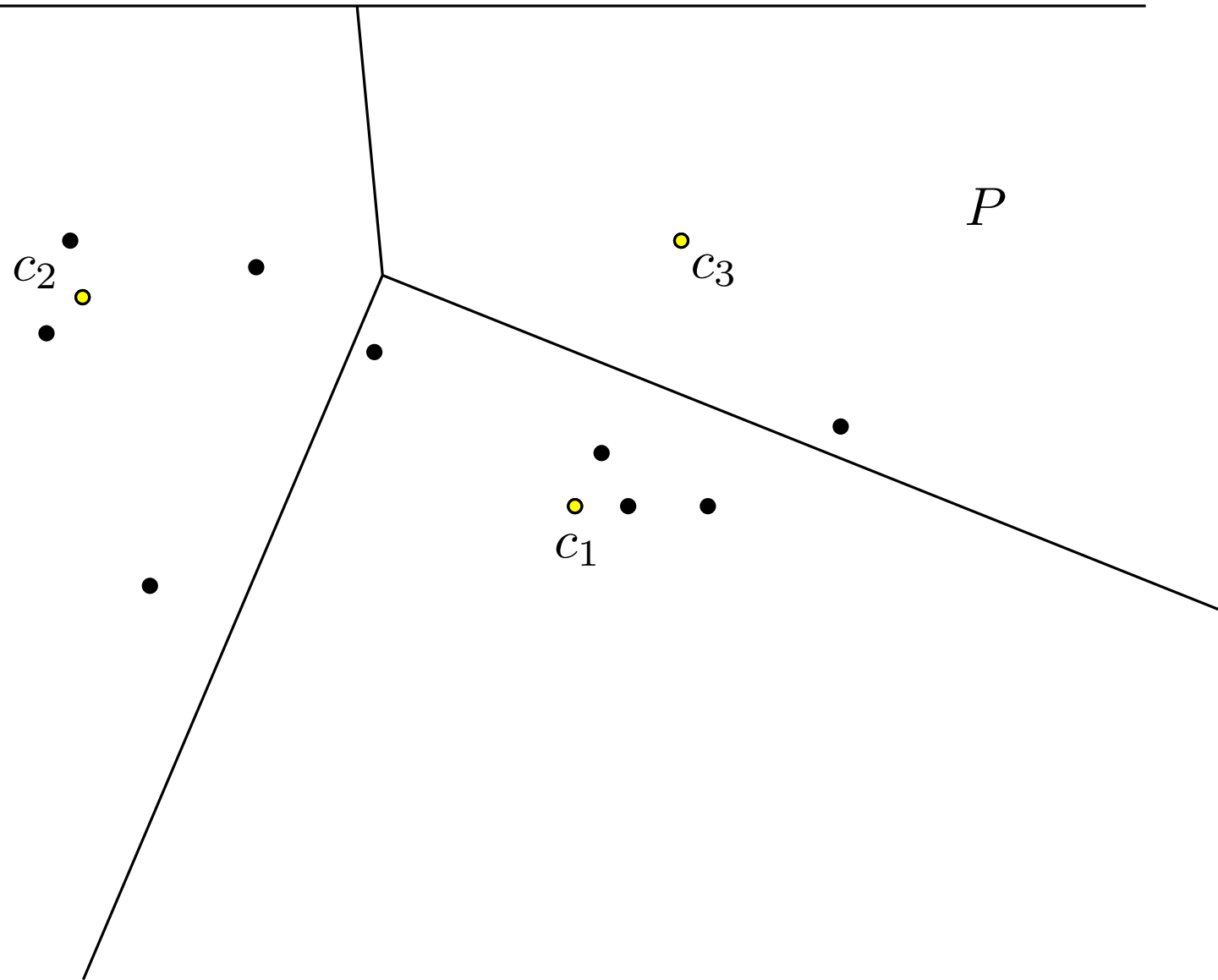
---



# k-means++ Algorithm

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

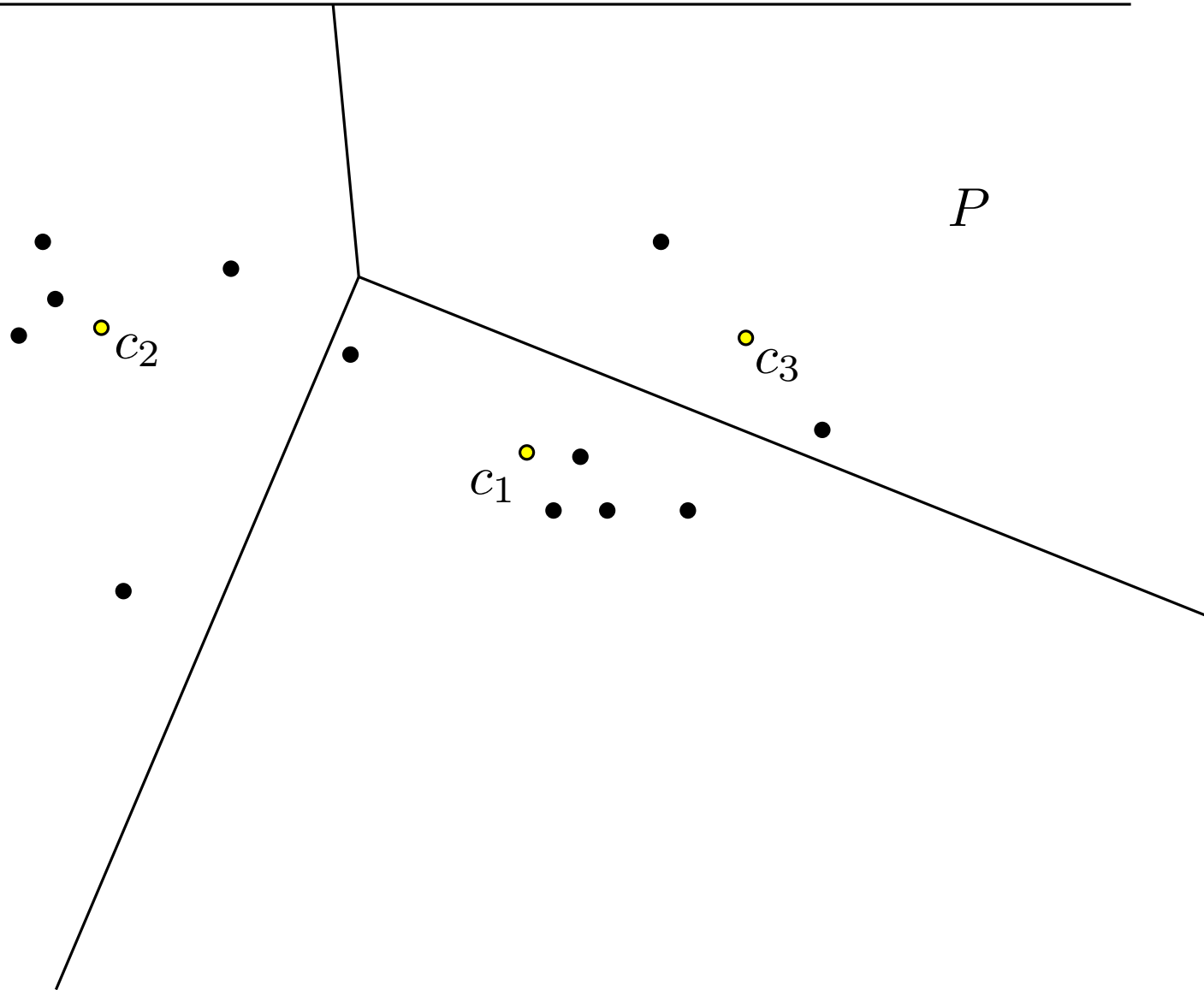
**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^2$



# k-means++ Algorithm

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^2$

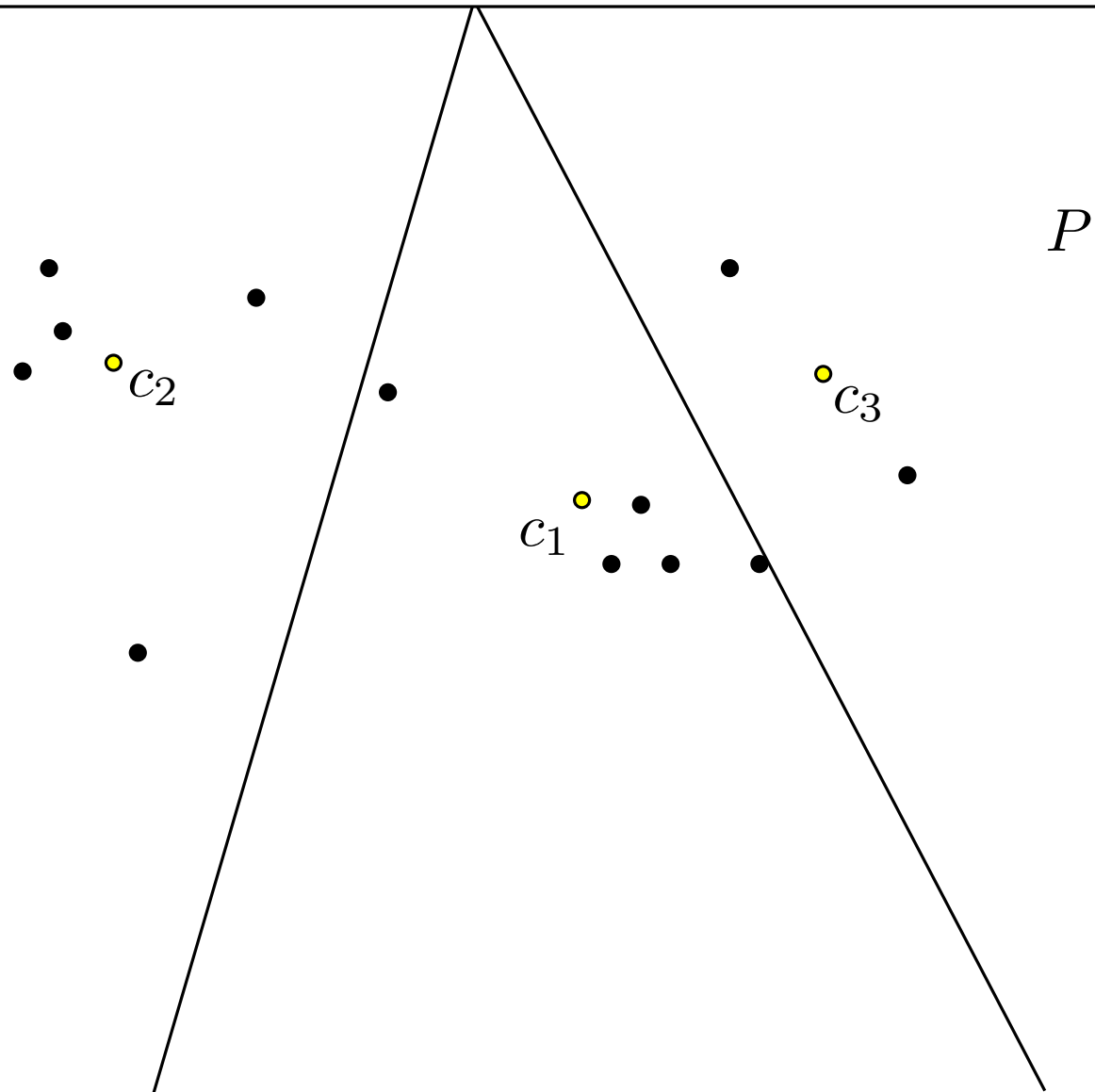


# k-means++ Algorithm

**Input:** set of points  $P = \{p_1, \dots, p_n\} \subseteq \mathbb{R}^d$ , value of  $k$

**Output:** set of centers  $C = \{c_1, \dots, c_k\} \subseteq \mathbb{R}^2$

---



# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$



# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:**

$$E[\phi(P)] = \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_1 = p_i \wedge c_2 = p_j]$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:**

$$\begin{aligned} E[\phi(P)] &= \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_1 = p_i \wedge c_2 = p_j] \\ &= \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_2 = p_j | c_1 = p_i] \cdot \Pr[c_1 = p_i] \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:**

$$\begin{aligned} E[\phi(P)] &= \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_1 = p_i \wedge c_2 = p_j] \\ &= \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_2 = p_j | c_1 = p_i] \cdot \Pr[c_1 = p_i] \\ &= \sum_{p_i \in P} \Pr[c_1 = p_i] \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_2 = p_j | c_1 = p_i] \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:**

$$\begin{aligned} E[\phi(P)] &= \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_1 = p_i \wedge c_2 = p_j] \\ &= \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_2 = p_j | c_1 = p_i] \cdot \Pr[c_1 = p_i] \\ &= \sum_{p_i \in P} \Pr[c_1 = p_i] \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_2 = p_j | c_1 = p_i] \\ &= \sum_{p_i \in P} \frac{1}{n} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_2 = p_j | c_1 = p_i] \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:**

$$\begin{aligned} E[\phi(P)] &= \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_1 = p_i \wedge c_2 = p_j] \\ &= \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_2 = p_j | c_1 = p_i] \cdot \Pr[c_1 = p_i] \\ &= \sum_{p_i \in P} \Pr[c_1 = p_i] \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_2 = p_j | c_1 = p_i] \\ &= \sum_{p_i \in P} \frac{1}{n} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_2 = p_j | c_1 = p_i] \\ &= \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \Pr[c_2 = p_j | c_1 = p_i] \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:**

$$\begin{aligned} E[\phi(P)] &= \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_1 = p_i \wedge c_2 = p_j] \\ &= \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_2 = p_j | c_1 = p_i] \cdot \Pr[c_1 = p_i] \\ &= \sum_{p_i \in P} \Pr[c_1 = p_i] \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_2 = p_j | c_1 = p_i] \\ &= \sum_{p_i \in P} \frac{1}{n} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \cdot \Pr[c_2 = p_j | c_1 = p_i] \\ &= \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \underbrace{\Pr[c_2 = p_j | c_1 = p_i]}_{\alpha_j} \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$



# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$

$$\begin{aligned} \leq \frac{1}{n} & \left( \sum_{p_i \in A_1} \sum_{p_j \in A_1} \phi(P, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \right. \\ & \left. + \sum_{p_i \in A_2} \sum_{p_j \in A_1} \phi(P, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_2} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \right) \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$

$$\leq \frac{1}{n} \left( \sum_{p_i \in A_1} \sum_{p_j \in A_1} \phi(P, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \right. \\ \left. + \sum_{p_i \in A_2} \sum_{p_j \in A_1} \phi(P, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_2} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \right)$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\sum_{p_i \in A_1} \sum_{p_j \in A_1} \phi(P, \{p_i, p_j\}) \alpha_j$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\sum_{p_i \in A_1} \sum_{p_j \in A_1} \underbrace{\phi(P, \{p_i, p_j\})}_{\leq \phi(P, \{p_i\})} \alpha_j$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\sum_{p_i \in A_1} \sum_{p_j \in A_1} \underbrace{\phi(P, \{p_i, p_j\})}_{\leq \phi(P, \{p_i\})} \alpha_j \leq \sum_{p_i \in A_1} \phi(P, \{p_i\}) \sum_{p_j \in A_1} \frac{\|p_i - p_j\|^2}{\phi(P, \{p_i\})}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} \sum_{p_i \in A_1} \sum_{p_j \in A_1} \underbrace{\phi(P, \{p_i, p_j\})}_{\leq \phi(P, \{p_i\})} \alpha_j &\leq \sum_{p_i \in A_1} \phi(P, \{p_i\}) \sum_{p_j \in A_1} \frac{\|p_i - p_j\|^2}{\phi(P, \{p_i\})} \\ &\leq \sum_{p_i \in A_1} \sum_{p_j \in A_1} \|p_i - p_j\|^2 \quad \text{pairwise distances within } A_1 \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} \sum_{p_i \in A_1} \sum_{p_j \in A_1} \underbrace{\phi(P, \{p_i, p_j\})}_{\leq \phi(P, \{p_i\})} \alpha_j &\leq \sum_{p_i \in A_1} \phi(P, \{p_i\}) \sum_{p_j \in A_1} \frac{\|p_i - p_j\|^2}{\phi(P, \{p_i\})} \\ &\leq \sum_{p_i \in A_1} \sum_{p_j \in A_1} \|p_i - p_j\|^2 \quad \text{pairwise distances within } A_1 \\ &\leq \sum_{p_i \in A_1} \left( \sum_{p_j \in A_1} \|p_j - \bar{a}_1\|^2 + |A_1| \cdot \|p_i - \bar{a}_1\|^2 \right) \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} \sum_{p_i \in A_1} \sum_{p_j \in A_1} \underbrace{\phi(P, \{p_i, p_j\})}_{\leq \phi(P, \{p_i\})} \alpha_j &\leq \sum_{p_i \in A_1} \phi(P, \{p_i\}) \sum_{p_j \in A_1} \frac{\|p_i - p_j\|^2}{\phi(P, \{p_i\})} \\ &\leq \sum_{p_i \in A_1} \sum_{p_j \in A_1} \|p_i - p_j\|^2 \quad \text{pairwise distances within } A_1 \\ &\leq \sum_{p_i \in A_1} \left( \sum_{p_j \in A_1} \|p_j - \bar{a}_1\|^2 + |A_1| \cdot \|p_i - \bar{a}_1\|^2 \right) \\ &\leq 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 \end{aligned}$$



# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$

$$\leq \frac{1}{n} \left( \sum_{p_i \in A_1} \sum_{p_j \in A_1} \phi(P, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \right. \\ \left. + \sum_{p_i \in A_2} \sum_{p_j \in A_1} \phi(P, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_2} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \right)$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j$$

$$= \sum_{p_i \in A_1} \sum_{p_j \in A_2} (\phi(A_1, \{p_i, p_j\}) + \phi(A_2, \{p_i, p_j\})) \alpha_j$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} & \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \\ &= \sum_{p_i \in A_1} \sum_{p_j \in A_2} (\phi(A_1, \{p_i, p_j\}) + \phi(A_2, \{p_i, p_j\})) \alpha_j \\ &= \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(A_1, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(A_2, \{p_i, p_j\}) \alpha_j \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j$$

$$= \sum_{p_i \in A_1} \sum_{p_j \in A_2} (\phi(A_1, \{p_i, p_j\}) + \phi(A_2, \{p_i, p_j\})) \alpha_j$$

$$= \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(A_1, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(A_2, \{p_i, p_j\}) \alpha_j$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(A_1, \{p_i, p_j\}) \alpha_j$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\sum_{p_i \in A_1} \sum_{p_j \in A_2} \underbrace{\phi(A_1, \{p_i, p_j\})}_{\leq \phi(A_1, \{p_i\})} \alpha_j$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\sum_{p_i \in A_1} \sum_{p_j \in A_2} \underbrace{\phi(A_1, \{p_i, p_j\})}_{\leq \phi(A_1, \{p_i\})} \alpha_j \leq \sum_{p_i \in A_1} \phi(A_1, \{p_i\}) \sum_{p_j \in A_2} \alpha_j$$



# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \underbrace{\phi(A_1, \{p_i, p_j\})}_{\leq \phi(A_1, \{p_i\})} \alpha_j &\leq \sum_{p_i \in A_1} \phi(A_1, \{p_i\}) \sum_{p_j \in A_2} \alpha_j \\ &\leq \sum_{p_i \in A_1} \sum_{p \in A_1} \|p - p_i\|^2 \sum_{p_j \in A_2} \alpha_j \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \underbrace{\phi(A_1, \{p_i, p_j\})}_{\leq \phi(A_1, \{p_i\})} \alpha_j &\leq \sum_{p_i \in A_1} \phi(A_1, \{p_i\}) \sum_{p_j \in A_2} \alpha_j \\ &\leq \sum_{p_i \in A_1} \sum_{p \in A_1} \|p - p_i\|^2 \underbrace{\sum_{p_j \in A_2} \alpha_j}_{\leq 1} \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \underbrace{\phi(A_1, \{p_i, p_j\})}_{\leq \phi(A_1, \{p_i\})} \alpha_j &\leq \sum_{p_i \in A_1} \phi(A_1, \{p_i\}) \sum_{p_j \in A_2} \alpha_j \\ &\leq \sum_{p_i \in A_1} \sum_{p \in A_1} \|p - p_i\|^2 \underbrace{\sum_{p_j \in A_2} \alpha_j}_{\leq 1} \\ &\leq \sum_{p_i \in A_1} \sum_{p \in A_1} \|p - p_i\|^2 \text{ pairwise distances within } A_1 \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \underbrace{\phi(A_1, \{p_i, p_j\})}_{\leq \phi(A_1, \{p_i\})} \alpha_j &\leq \sum_{p_i \in A_1} \phi(A_1, \{p_i\}) \sum_{p_j \in A_2} \alpha_j \\ &\leq \sum_{p_i \in A_1} \sum_{p \in A_1} \|p - p_i\|^2 \underbrace{\sum_{p_j \in A_2} \alpha_j}_{\leq 1} \\ &\leq \sum_{p_i \in A_1} \sum_{p \in A_1} \|p - p_i\|^2 \text{ pairwise distances within } A_1 \\ &\leq 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} & \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \\ &= \sum_{p_i \in A_1} \sum_{p_j \in A_2} (\phi(A_1, \{p_i, p_j\}) + \phi(A_2, \{p_i, p_j\})) \alpha_j \\ &= \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(A_1, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(A_2, \{p_i, p_j\}) \alpha_j \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(A_2, \{p_i, p_j\}) \alpha_j \quad (\text{this is the crucial case})$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} & \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(A_2, \{p_i, p_j\}) \alpha_j \quad (\text{this is the crucial case}) \\ &= \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \alpha_j \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} & \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(A_2, \{p_i, p_j\}) \alpha_j \quad (\text{this is the crucial case}) \\ &= \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \alpha_j \\ &= \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\|p_i - p_j\|^2}{\sum_{p \in P} \|p - p_i\|^2} \end{aligned}$$



# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} & \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(A_2, \{p_i, p_j\}) \alpha_j \quad (\text{this is the crucial case}) \\ &= \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \alpha_j \\ &= \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\|p_i - p_j\|^2}{\sum_{p \in P} \|p - p_i\|^2} \\ &\leq \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\frac{2}{|A_2|} \sum_{p \in A_2} (\|p_i - p\|^2 + \|p - p_j\|^2)}{\sum_{p \in P} \|p - p_i\|^2} \end{aligned}$$

(using power-mean inequality)

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} &= \frac{2}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\sum_{p \in A_2} \|p_i - p\|^2}{\sum_{p \in P} \|p - p_i\|^2} \\ &\quad + \frac{2}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\sum_{p \in A_2} \|p - p_j\|^2}{\sum_{p \in P} \|p - p_i\|^2} \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned} & \leq \sum_{p \in A_2} \|p - p_j\|^2 \quad \leq 1 \\ &= \frac{2}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\sum_{p \in A_2} \|p_i - p\|^2}{\sum_{p \in P} \|p - p_i\|^2} \\ & \quad + \frac{2}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\sum_{p \in A_2} \|p - p_j\|^2}{\sum_{p \in P} \|p - p_i\|^2} \\ & \quad \leq \sum_{p \in A_2} \|p - p_i\|^2 \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned}
 & \leq \sum_{p \in A_2} \|p - p_j\|^2 \quad \leq 1 \\
 = & \frac{2}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\sum_{p \in A_2} \|p_i - p\|^2}{\sum_{p \in P} \|p - p_i\|^2} \\
 & + \frac{2}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\sum_{p \in A_2} \|p - p_j\|^2}{\sum_{p \in P} \|p - p_i\|^2} \\
 & \leq \sum_{p \in A_2} \|p - p_i\|^2
 \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned}
 & \leq \sum_{p \in A_2} \|p - p_j\|^2 \quad \leq 1 \\
 & = \frac{2}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\sum_{p \in A_2} \|p_i - p\|^2}{\sum_{p \in P} \|p - p_i\|^2} \\
 & \quad + \frac{2}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\sum_{p \in A_2} \|p - p_j\|^2}{\sum_{p \in P} \|p - p_i\|^2} \\
 & \quad \leq \sum_{p \in A_2} \|p - p_i\|^2 \\
 & \leq \frac{4}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \sum_{p \in A_2} \|p - p_j\|^2
 \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned}
 & \leq \sum_{p \in A_2} \|p - p_j\|^2 \quad \leq 1 \\
 & = \frac{2}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\sum_{p \in A_2} \|p_i - p\|^2}{\sum_{p \in P} \|p - p_i\|^2} \\
 & \quad + \frac{2}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\sum_{p \in A_2} \|p - p_j\|^2}{\sum_{p \in P} \|p - p_i\|^2} \\
 & \quad \leq \sum_{p \in A_2} \|p - p_i\|^2 \\
 & \leq \frac{4}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \sum_{p \in A_2} \|p - p_j\|^2 \\
 & \leq \frac{4n}{|A_2|} \sum_{p_j \in A_2} \sum_{p \in A_2} \|p - p_j\|^2
 \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$\begin{aligned}
 & \leq \sum_{p \in A_2} \|p - p_j\|^2 \quad \leq 1 \\
 & = \frac{2}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\sum_{p \in A_2} \|p_i - p\|^2}{\sum_{p \in P} \|p - p_i\|^2} \\
 & \quad + \frac{2}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \left( \sum_{p \in A_2} \min(\|p - p_i\|^2, \|p - p_j\|^2) \right) \frac{\sum_{p \in A_2} \|p - p_j\|^2}{\sum_{p \in P} \|p - p_i\|^2} \\
 & \quad \leq \sum_{p \in A_2} \|p - p_i\|^2 \\
 & \leq \frac{4}{|A_2|} \sum_{p_i \in A_1} \sum_{p_j \in A_2} \sum_{p \in A_2} \|p - p_j\|^2 \\
 & \leq \frac{4n}{|A_2|} \sum_{p_j \in A_2} \sum_{p \in A_2} \|p - p_j\|^2 \leq 8n \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2
 \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$

$$\begin{aligned} \leq \frac{1}{n} & \left( \sum_{p_i \in A_1} \sum_{p_j \in A_1} \phi(P, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \right. \\ & \left. + \sum_{p_i \in A_2} \sum_{p_j \in A_1} \phi(P, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_2} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \right) \end{aligned}$$



# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$

$$\leq \frac{1}{n} \left( 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + \sum_{p_i \in A_1} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \right. \\ \left. + \sum_{p_i \in A_2} \sum_{p_j \in A_1} \phi(P, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_2} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \right)$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$

$$\leq \frac{1}{n} \left( \begin{aligned} &2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 2|A_2| \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 + 8n \sum_{p_i \in A_2} \|p_i - \bar{a}_1\|^2 \\ &+ \sum_{p_i \in A_2} \sum_{p_j \in A_1} \phi(P, \{p_i, p_j\}) \alpha_j + \sum_{p_i \in A_2} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \end{aligned} \right)$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$

$$\leq \frac{1}{n} \left( \begin{aligned} &2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 8n \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 \\ &2|A_2| \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 + 8n \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + \sum_{p_i \in A_2} \sum_{p_j \in A_2} \phi(P, \{p_i, p_j\}) \alpha_j \end{aligned} \right)$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$

$$\leq \frac{1}{n} \left( \begin{aligned} &2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 8n \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 \\ &2|A_2| \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 + 8n \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 2|A_2| \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 \end{aligned} \right)$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$

optimal cost within cluster  $A_1$

$$\leq \frac{1}{n} \left( \begin{aligned} & 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_2\|^2 + 8n \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 \\ & 2|A_2| \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 + 8n \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 2|A_2| \sum_{p_i \in A_2} \|p_i - \bar{a}_1\|^2 \end{aligned} \right)$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$

optimal cost within cluster  $A_1$

$$\leq \frac{1}{n} \left( \begin{aligned} & 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 8n \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 \\ & 2|A_2| \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 + 8n \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 2|A_2| \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 \end{aligned} \right)$$

optimal cost within cluster  $A_2$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$

$$\begin{aligned} &\leq \frac{1}{n} \left( 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 8n \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 \right. \\ &\quad \left. 2|A_2| \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 + 8n \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 2|A_2| \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 \right) \\ &\leq 12 \left( \sum_{p \in A_1} \|p - \bar{a}_1\|^2 + \sum_{p \in A_2} \|p - \bar{a}_2\|^2 \right) \end{aligned}$$

# k-means++ (Analysis for $k = 2$ )

**Claim:** For  $k = 2$ :  $E[\phi(P)] \leq O(\phi(P, C^*))$

**Proof:** (continued)

$$E[\phi(P)] = \frac{1}{n} \sum_{p_i \in P} \sum_{p_j \in P} \phi(P, \{p_i, p_j\}) \alpha_j$$

Let  $A_1 \subset P$  and  $A_2 \subset P$  be two clusters of an optimal 2-means clustering of  $P$

$$\begin{aligned} &\leq \frac{1}{n} \left( 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 2|A_1| \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 8n \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 \right. \\ &\quad \left. 2|A_2| \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 + 8n \sum_{p_i \in A_1} \|p_i - \bar{a}_1\|^2 + 2|A_2| \sum_{p_i \in A_2} \|p_i - \bar{a}_2\|^2 \right) \\ &\leq 12 \left( \sum_{p \in A_1} \|p - \bar{a}_1\|^2 + \sum_{p \in A_2} \|p - \bar{a}_2\|^2 \right) \leq 12\phi(P, C^*) \end{aligned}$$

□



# k-means++ Algorithm

For general  $k$ , the solution obtained by  $k$ -means++ will, in expectation, be at most a factor  $O(\log k)$  worse than the optimal solution.

# k-means++ Algorithm

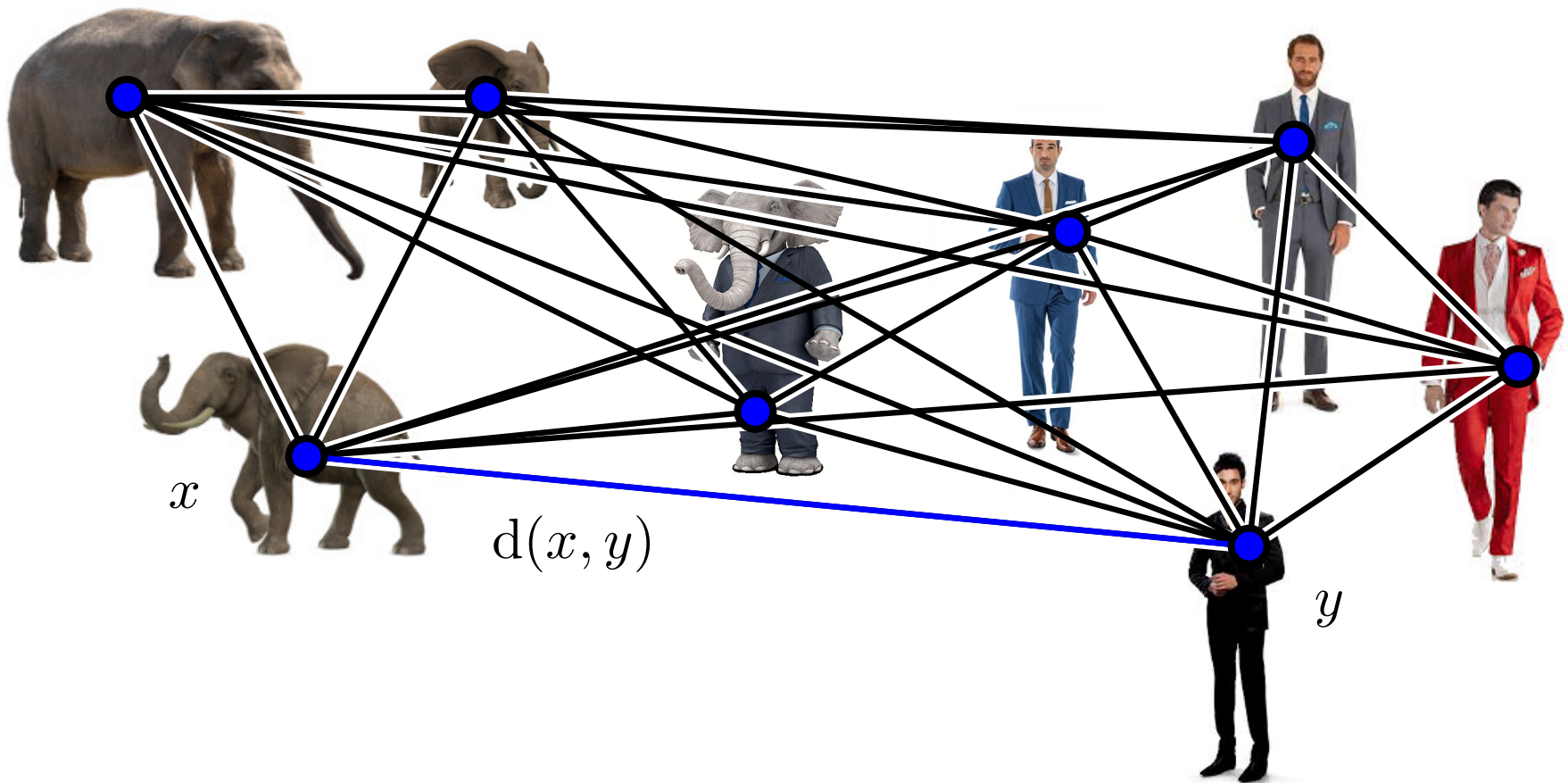
For general  $k$ , the solution obtained by  $k$ -means++ will, in expectation, be at most a factor  $O(\log k)$  worse than the optimal solution.

Gonzales' algorithm and  $k$ -means++:

- **Gonzales**: choose the next center from  $P$  as the point that maximizes the current cost
- **k-means++**: choose the next center from  $P$  with probability relative to the contribution to the current cost

# Clustering in Graphs

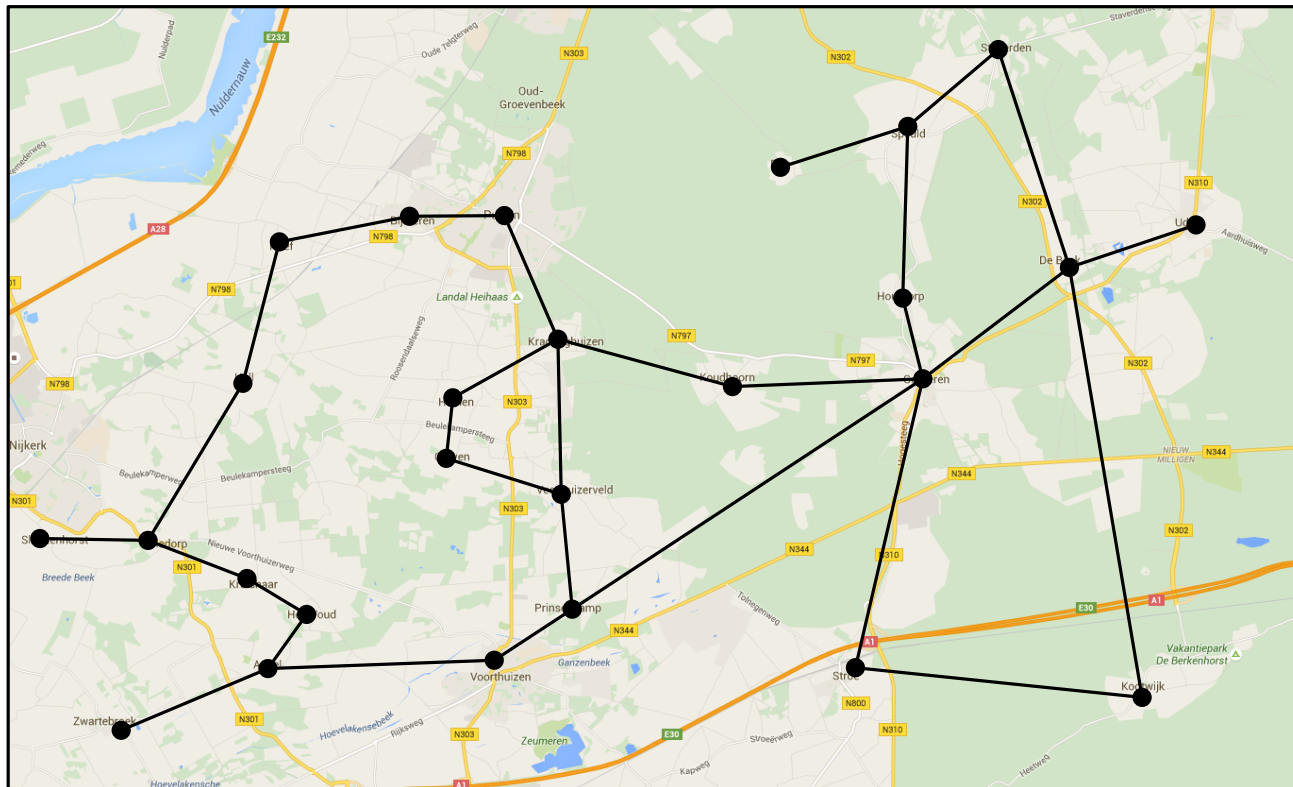
- Vertices of the graph represent the objects to be clustered
- Distance is measured by shortest path



# Facility Location with Road Network

You may build two hospitals in two different villages serving the surrounding villages. Where do you place them to minimize the maximal distance from any village to its serving hospital?

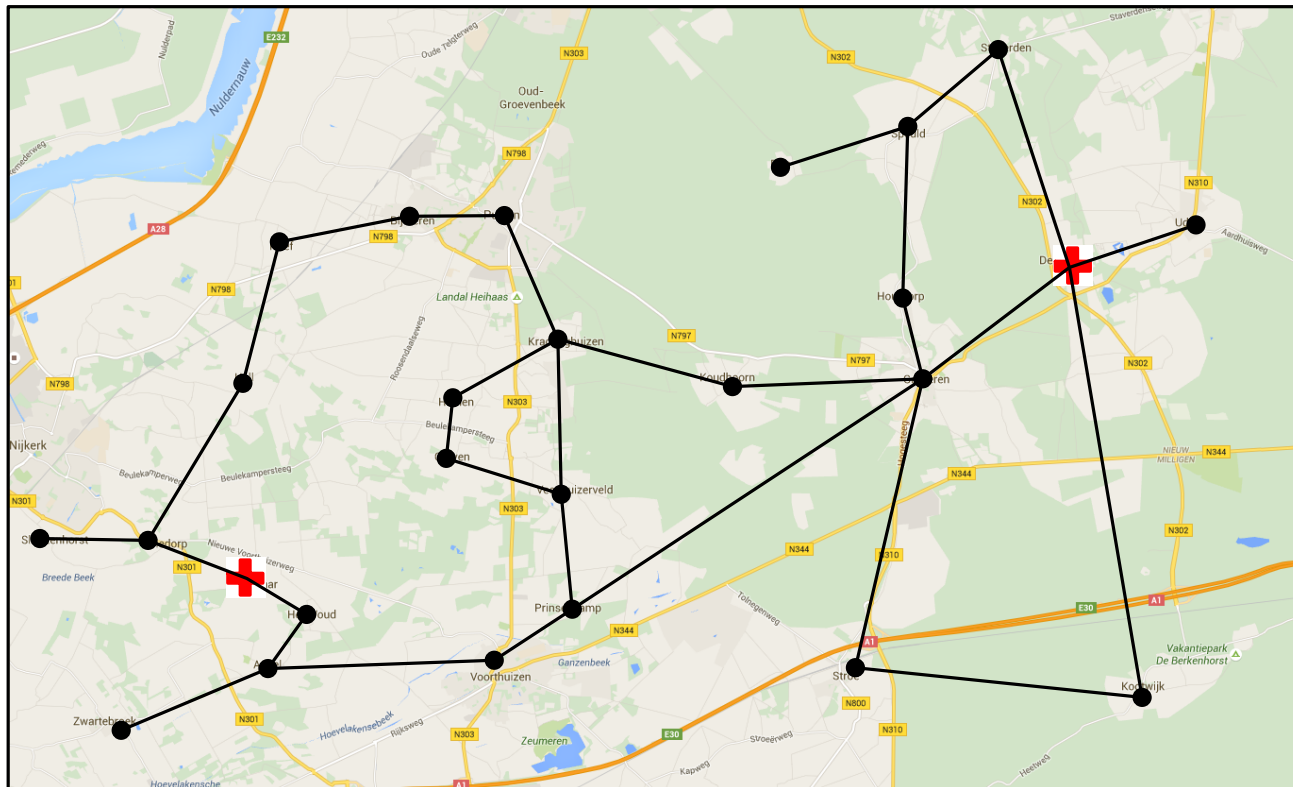
**Variant:** Measure distance along the road network (travel distance).



# Facility Location with Road Network

You may build two hospitals in two different villages serving the surrounding villages. Where do you place them to minimize the maximal distance from any village to its serving hospital?

**Variant:** Measure distance along the road network (travel distance).



# Summary

- Clustering
- Facility Location
- Gonzales' algorithm
- Lloyd's algorithm (k-means)
- k-means++ algorithm
- Clustering in graphs

# References

- Avrim Blum, John Hopcroft, Ravindran Khannan: *Foundations of Data Science*
- Sarel Har-Peled: *Geometric Approximation Algorithms*
- Arthur, D. and Vassilvitskii, S. (2007). "k-means++: the advantages of careful seeding" (PDF). Proc. 18th ACM-SIAM Symposium on Discrete Algorithms. pp. 1027-1035.