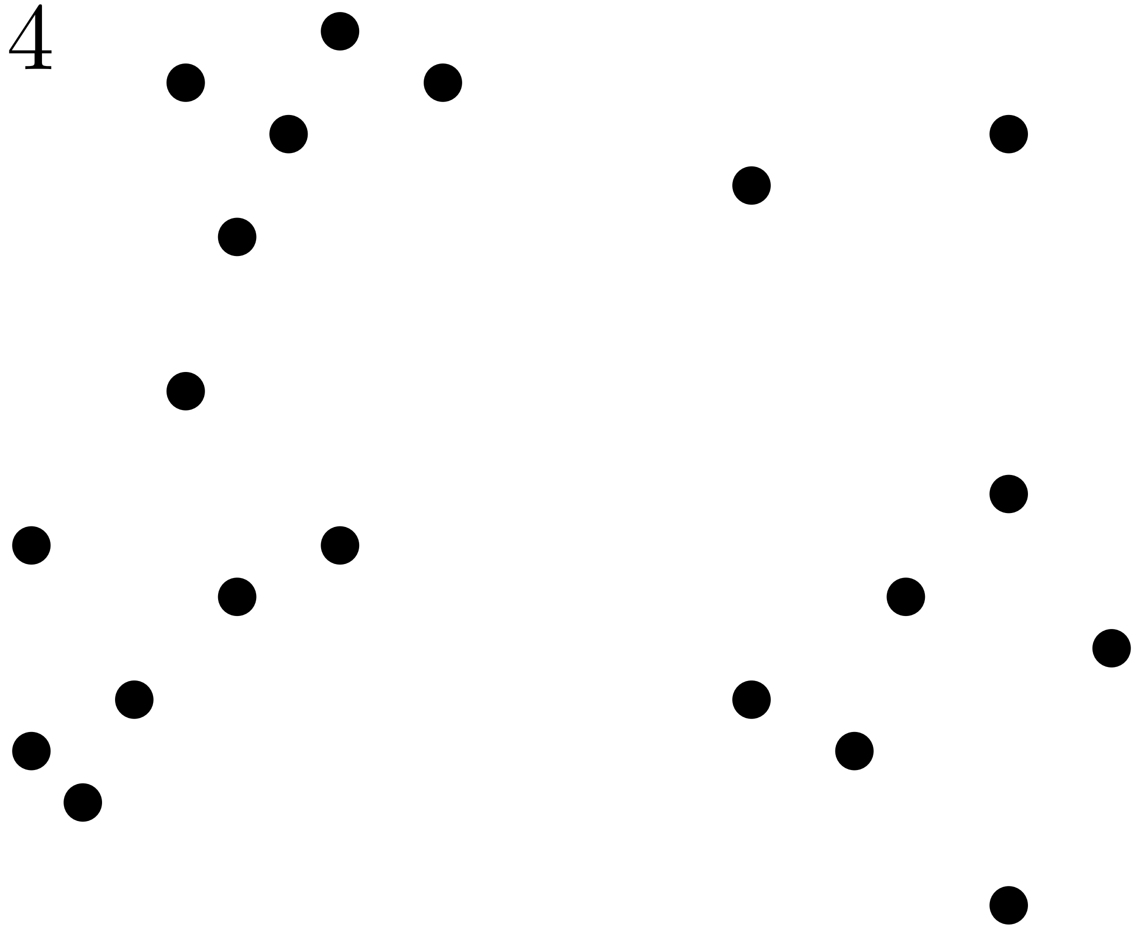
Center-based Clustering

2-approximation for k-center clustering
5-approximation for k-median clustering
k-means clustering

Center-based clustering – intuition

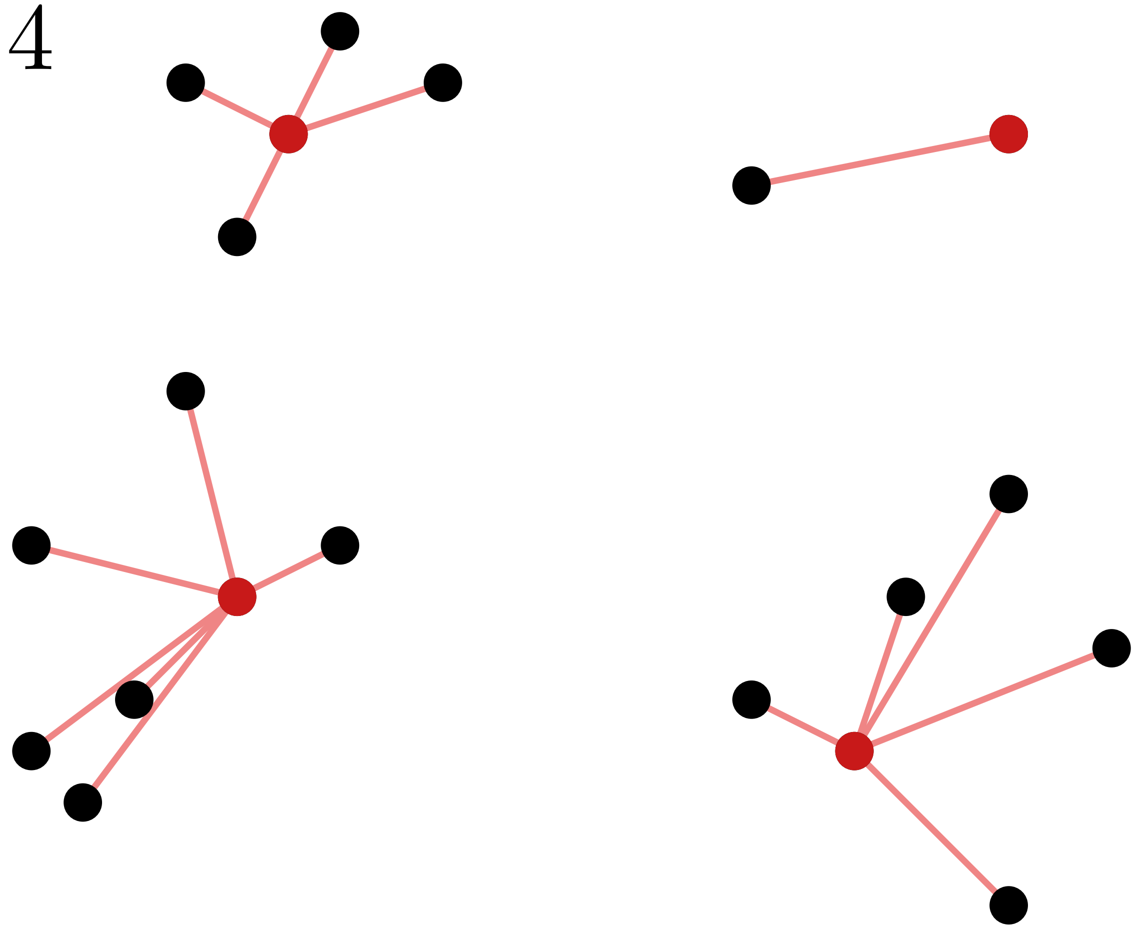
Given: integer k, point set P

$k = 4$



Center-based clustering – intuition

$k = 4$

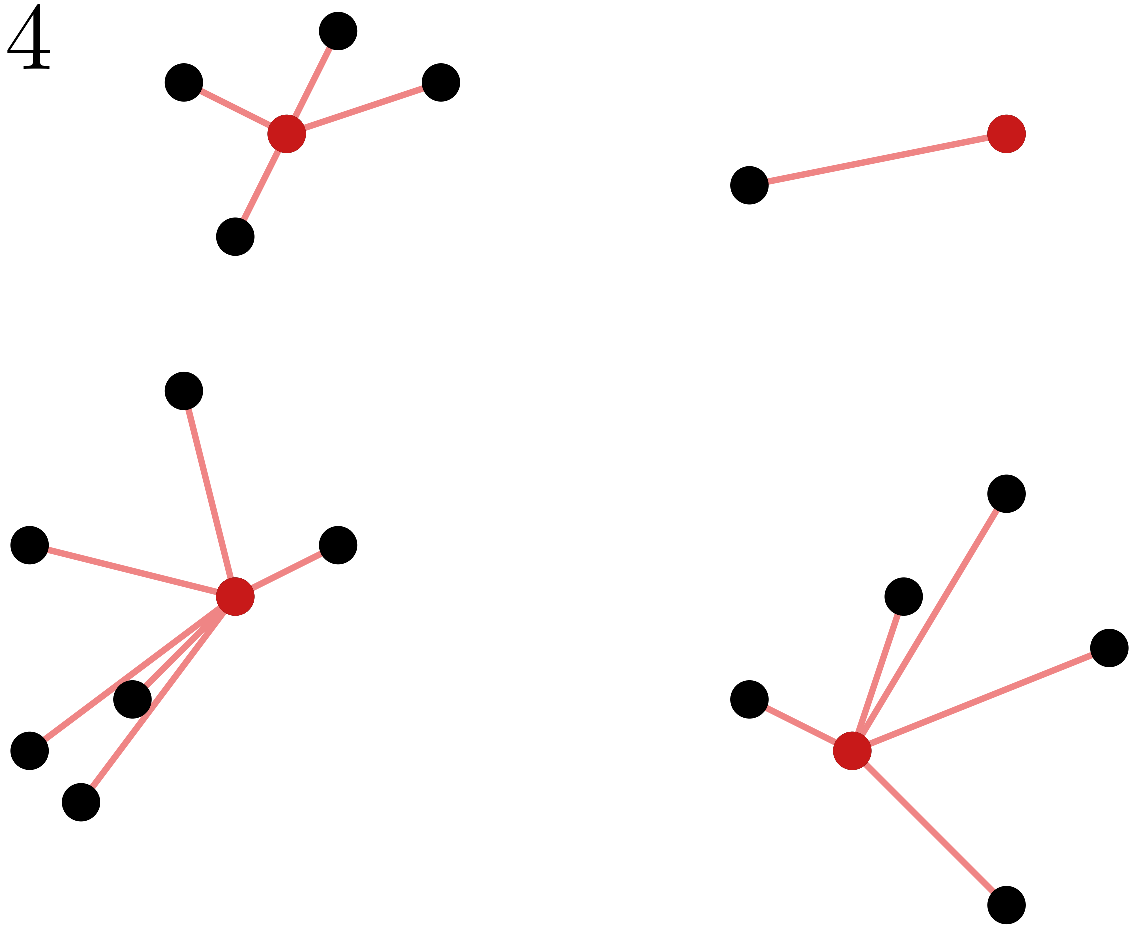


Given: integer k, point set P

Goal: point set C, of size k such that every point in P is close to a point in C

Center-based clustering – intuition

$k = 4$



Given: integer k, point set P

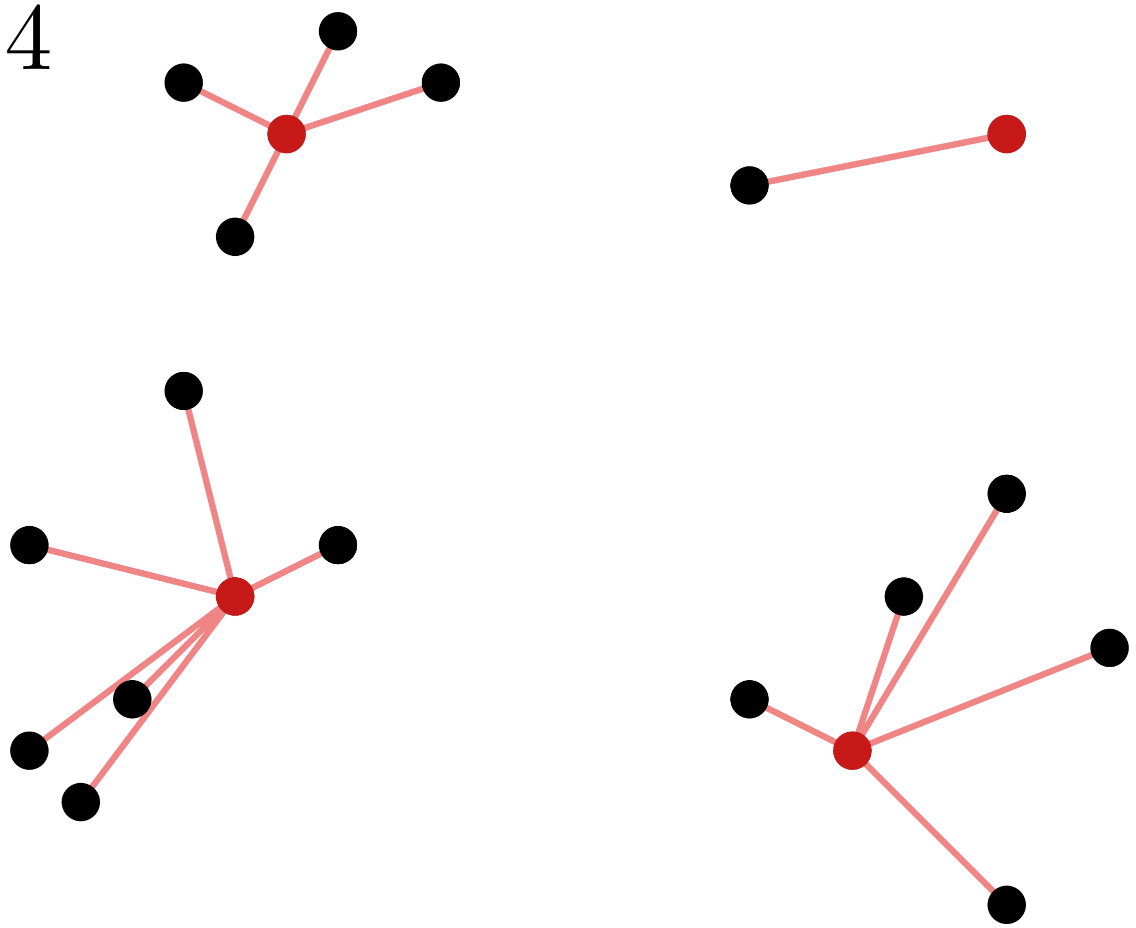
Goal: point set C, of size k such that every point in P is close to a point in C

Motivation:

- placing facilities, e.g., hospitals
- finding groups of nearby points

Center-based clustering – intuition

$k = 4$



Given: integer k, point set P in metric space

Goal: point set C, of size k such that every point in P is close to a point in C

Motivation:

- placing facilities, e.g., hospitals
- finding groups of nearby points

Preliminaries

metric space: pair (X, d) with X a set, and $d\colon X \times X \to [0, \infty)$ satisfying

Preliminaries

metric space: pair (X, d) with X a set, and $d \colon X \times X \to [0, \infty)$ satisfying

$d(x, y) = 0$ if and only if $x = y$,

Preliminaries

metric space: pair (X, d) with X a set, and $d\colon X \times X \to [0, \infty)$ satisfying

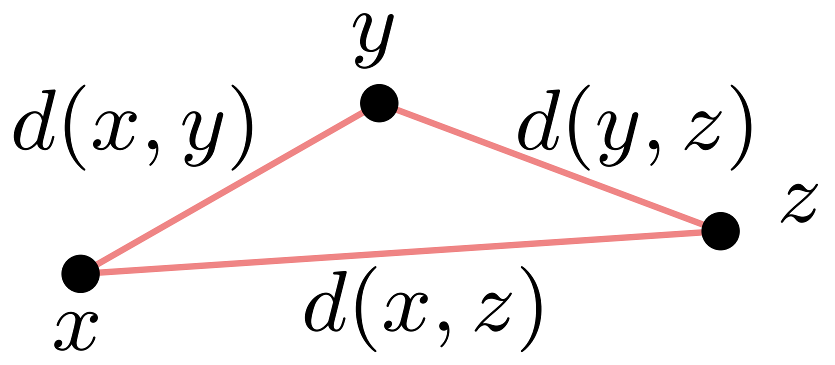
$d(x, y) = 0$ if and only if $x = y$,

$d(x, y) = d(y, x)$,

Preliminaries

metric space: pair (X, d) with X a set, and $d\colon X \times X \to [0, \infty)$ satisfying

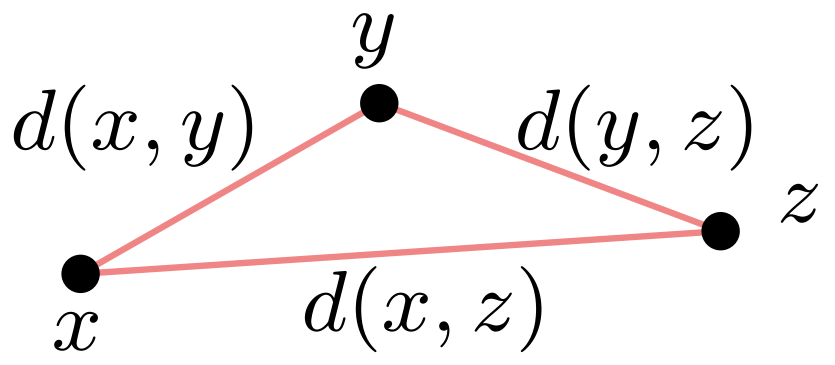
$$d(x, y) = 0 \text{ if and only if } x = y,$$
$$d(x, y) = d(y, x),$$
$$d(x, z) \leq d(x, y) + d(y, z).$$ (triangle inequality)



Preliminaries

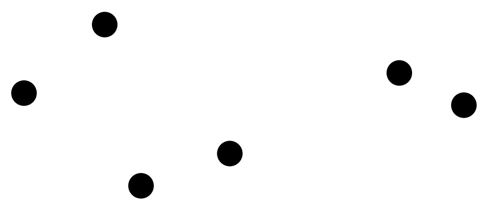
metric space: pair (X, d) with X a set, and $d\colon X \times X \to [0, \infty)$ satisfying

$$d(x,y) = 0 \text{ if and only if } x = y,$$
$$d(x,y) = d(y,x),$$
$$d(x,z) \leq d(x,y) + d(y,z). \quad \text{(triangle inequality)}$$



examples:

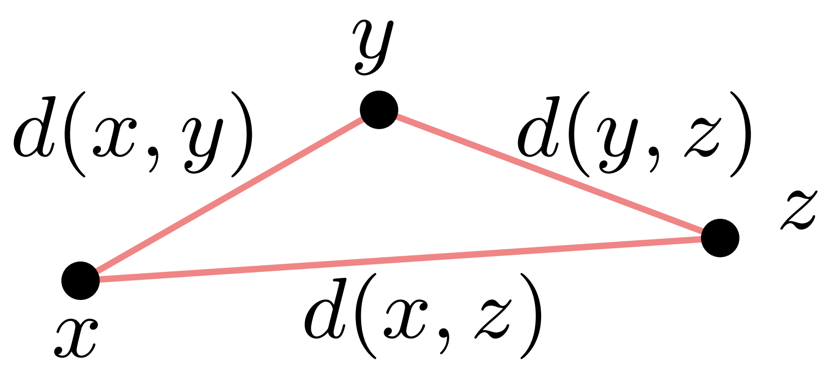
R^2 with Euclidean distance



Preliminaries

metric space: pair (X, d) with X a set, and $d\colon X \times X \to [0, \infty)$ satisfying

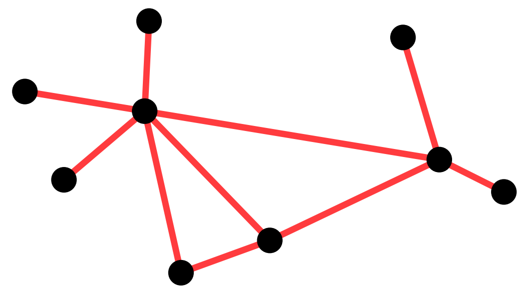
$$d(x, y) = 0 \text{ if and only if } x = y,$$
$$d(x, y) = d(y, x),$$
$$d(x, z) \leq d(x, y) + d(y, z).$$ (triangle inequality)



examples:

R^2 with Euclidean distance

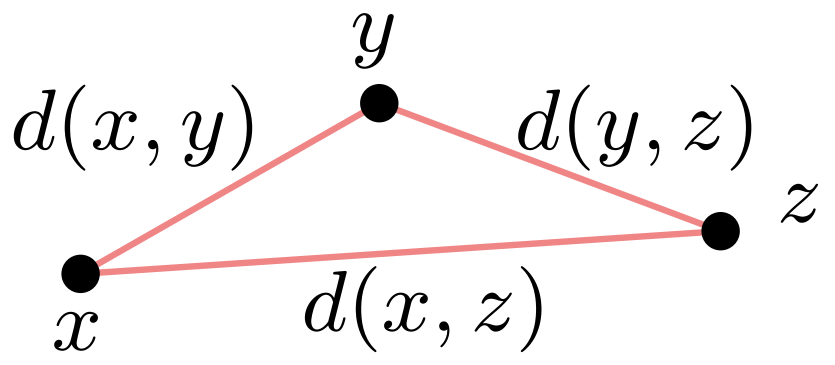
Graph with shortest-path distance



Preliminaries

metric space: pair (X, d) with X a set, and $d\colon X \times X \to [0, \infty)$ satisfying

$d(x, y) = 0$ if and only if $x = y$,
$d(x, y) = d(y, x)$,
$d(x, z) \leq d(x, y) + d(y, z)$. (triangle inequality)

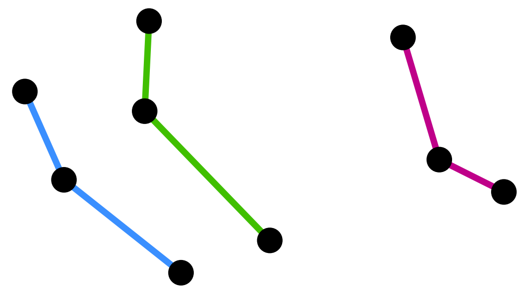


examples:

R^2 with Euclidean distance

Graph with shortest-path distance

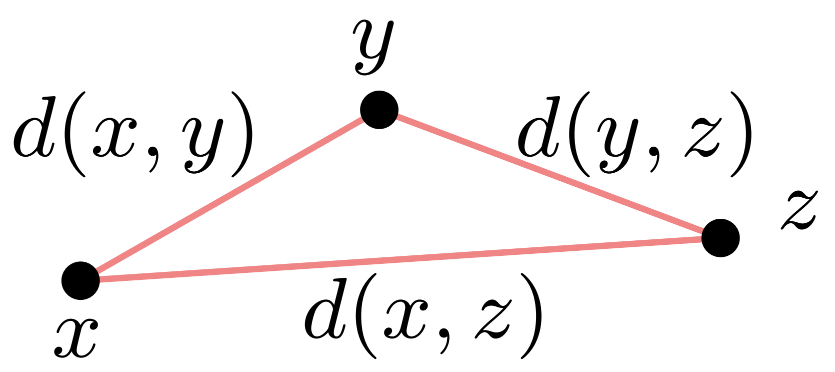
curves with Fréchet distance



Preliminaries

metric space: pair (X, d) with X a set, and $d\colon X \times X \to [0, \infty)$ satisfying

$d(x, y) = 0$ if and only if $x = y$,
$d(x, y) = d(y, x)$,
$d(x, z) \leq d(x, y) + d(y, z)$. (triangle inequality)

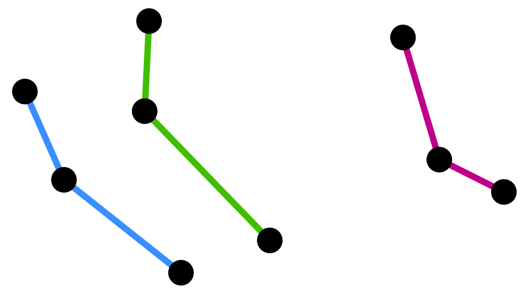


examples:

R^2 with Euclidean distance

Graph with shortest-path distance

curves with Fréchet distance



notation: $d(p, C) := \min_{q \in C} d(p, q)$

k-center clustering in metric space (X, d)

Given: $P \subset X$ and integer k

Goal: Find $C \subset X$ of size k such that

$$\max_{p \in P} d(p, C)$$

is minimized.

$k = 2$

k-center clustering in metric space (X, d)

Given: $P \subset X$ and integer k

Goal: Find $C \subset X$ of size k such that

$$\max_{p \in P} d(p, C)$$

is minimized.

$k = 2$

k-center clustering in metric space (X, d)

$k = 2$

Given: $P \subset X$ and integer k

Goal: Find $C \subset X$ of size k such that

$$\max_{p \in P} d(p, C)$$

is minimized.

discrete k-center problem: $C \subset P$

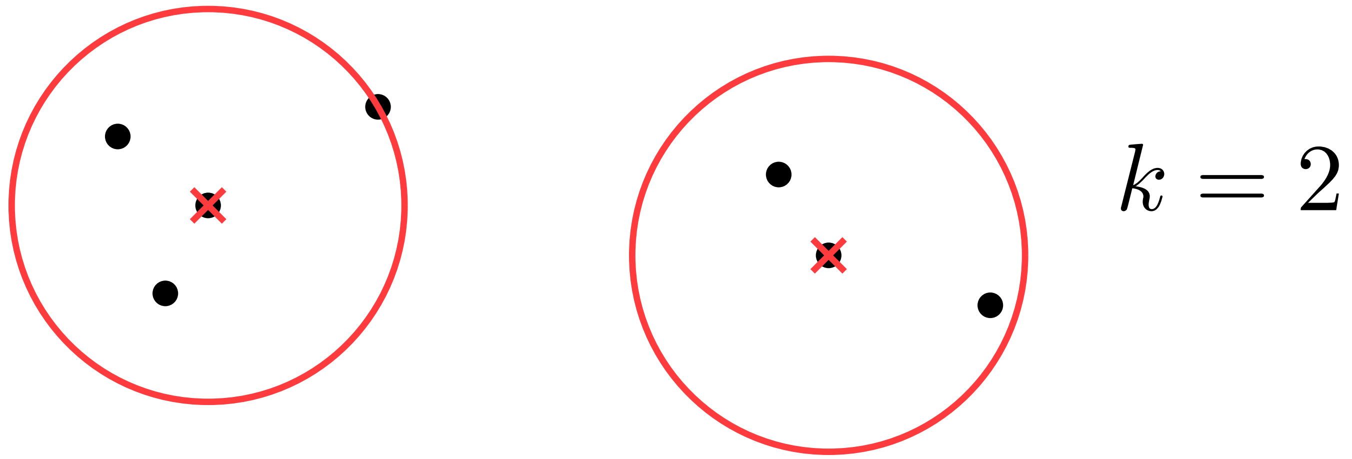
k-center clustering in metric space (X, d)

Given: $P \subset X$ and integer k

Goal: Find $C \subset X$ of size k such that

$$\max_{p \in P} d(p, C)$$

is minimized.



discrete k-center problem: $C \subset P$

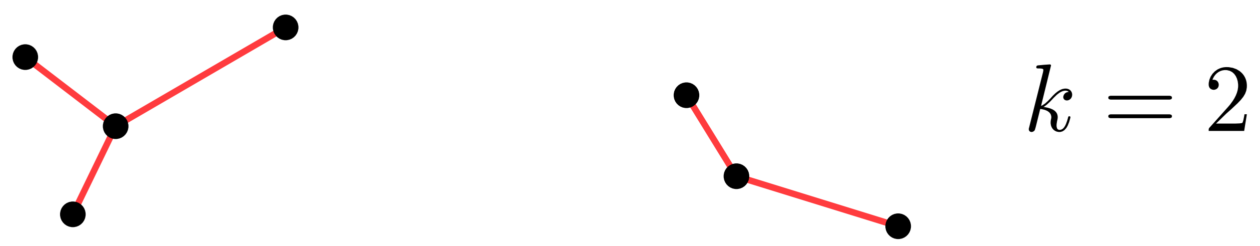
k-center clustering in metric space (X, d)

Given: $P \subset X$ and integer k

Goal: Find $C \subset X$ of size k such that

$$\max_{p \in P} d(p, C)$$

is minimized.



discrete k-center problem: $C \subset P$

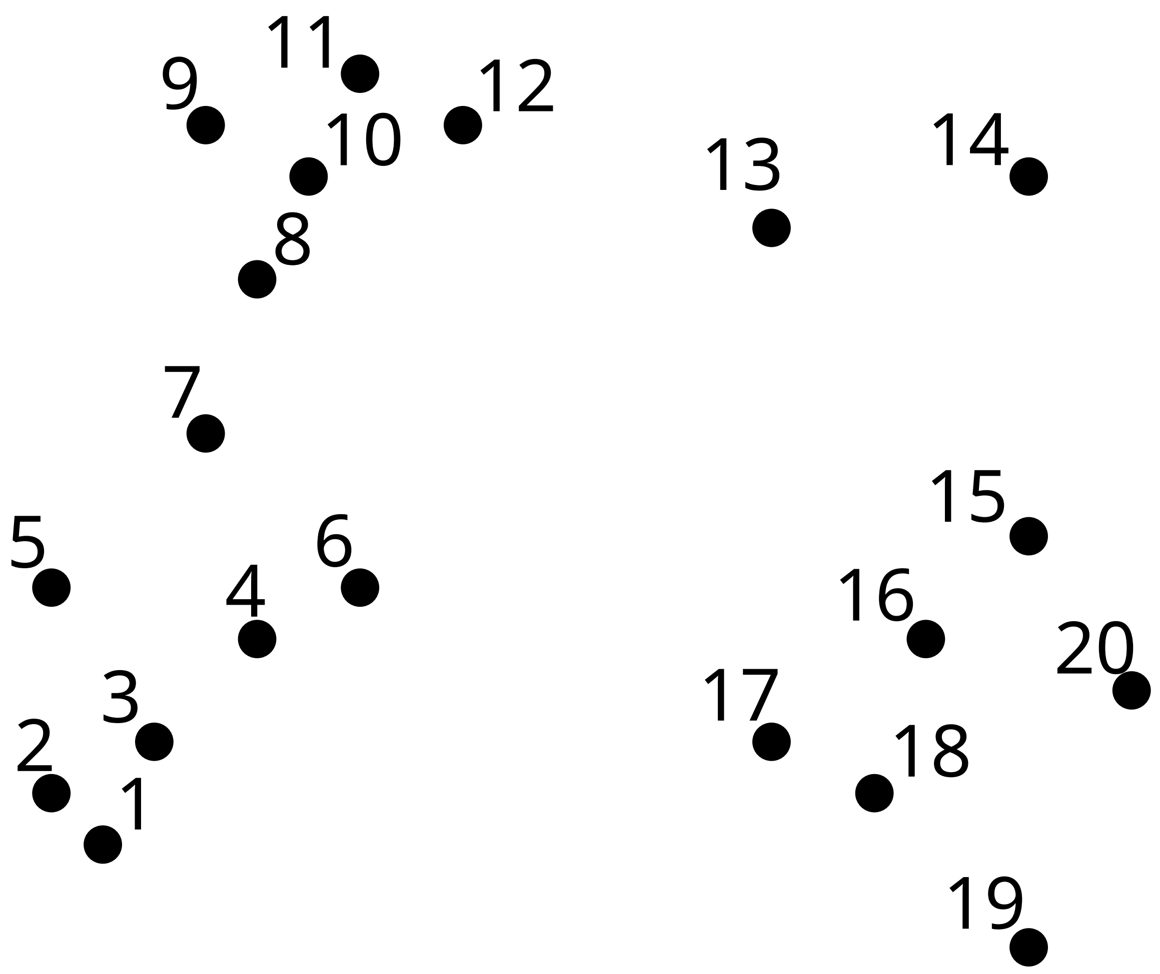
later:

(discrete) k-median problem: sum instead of $\max$

k-means: sum of squares

Quiz

Which of the following is an optimal set of centers for $k = 3$?



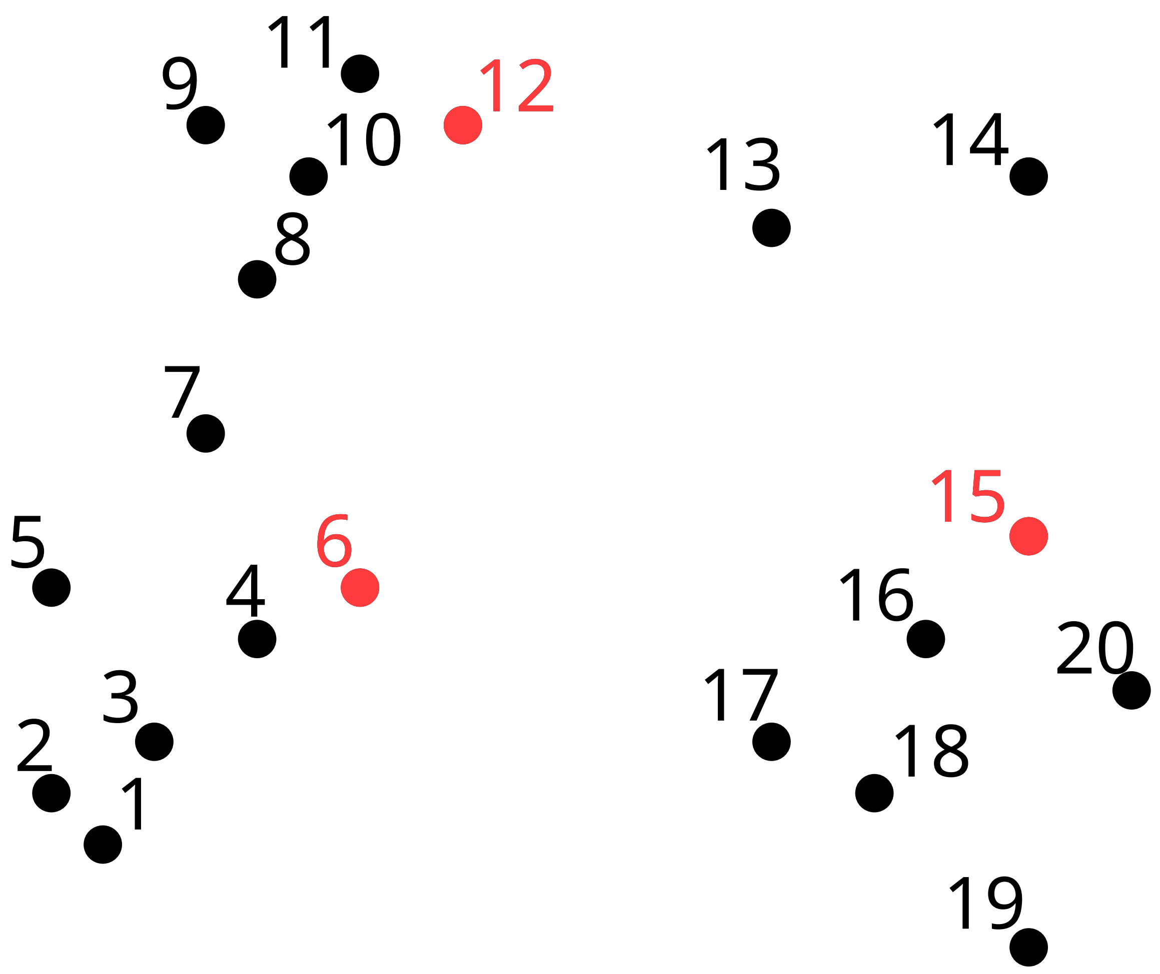
A 4, 10, 16

B 6, 12, 15

C 7, 13, 16

Quiz

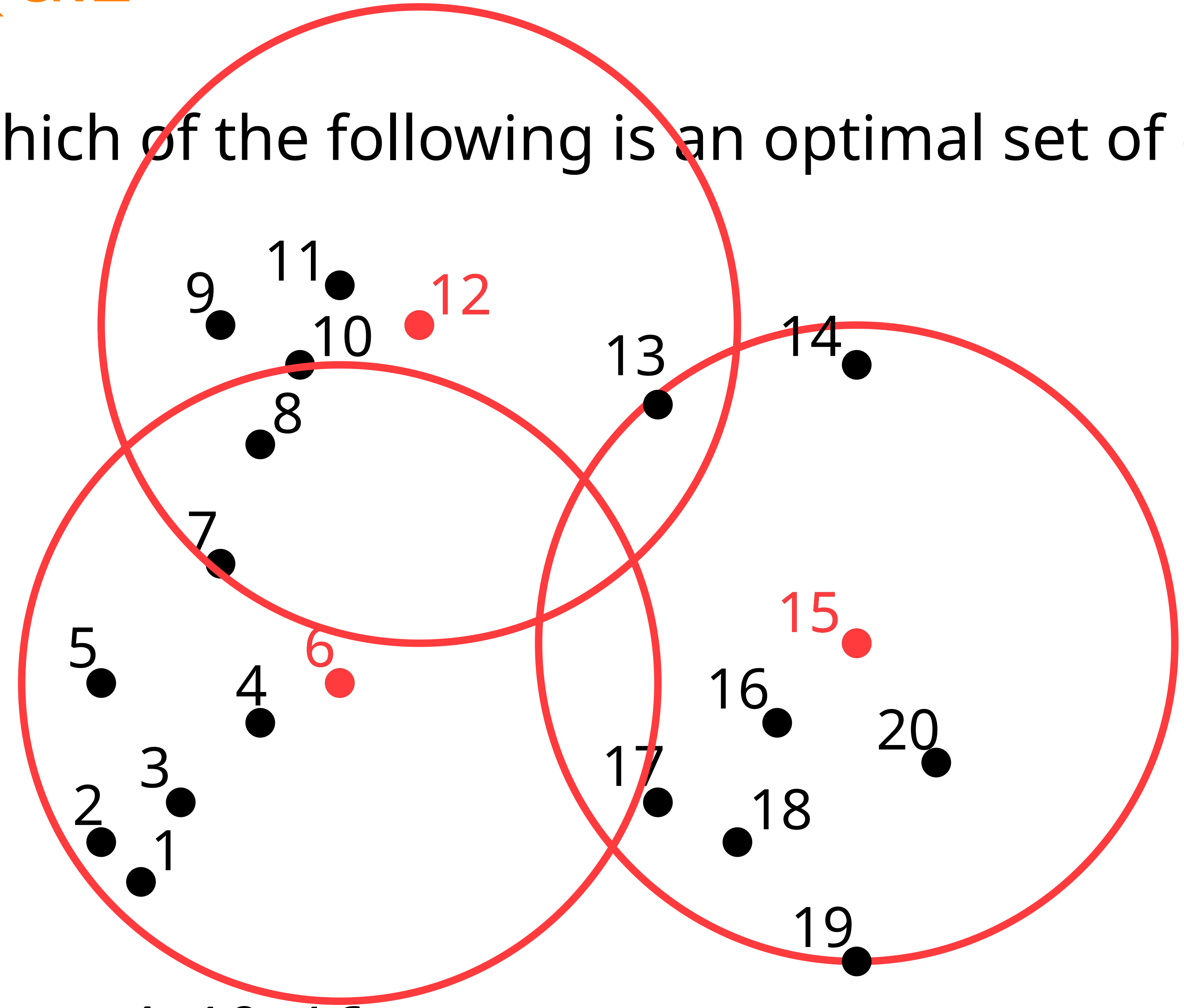
Which of the following is an optimal set of centers for $k = 3$?



A	4, 10, 16
B	6, 12, 15
C	7, 13, 16

Quiz

Which of the following is an optimal set of centers for $k = 3$?



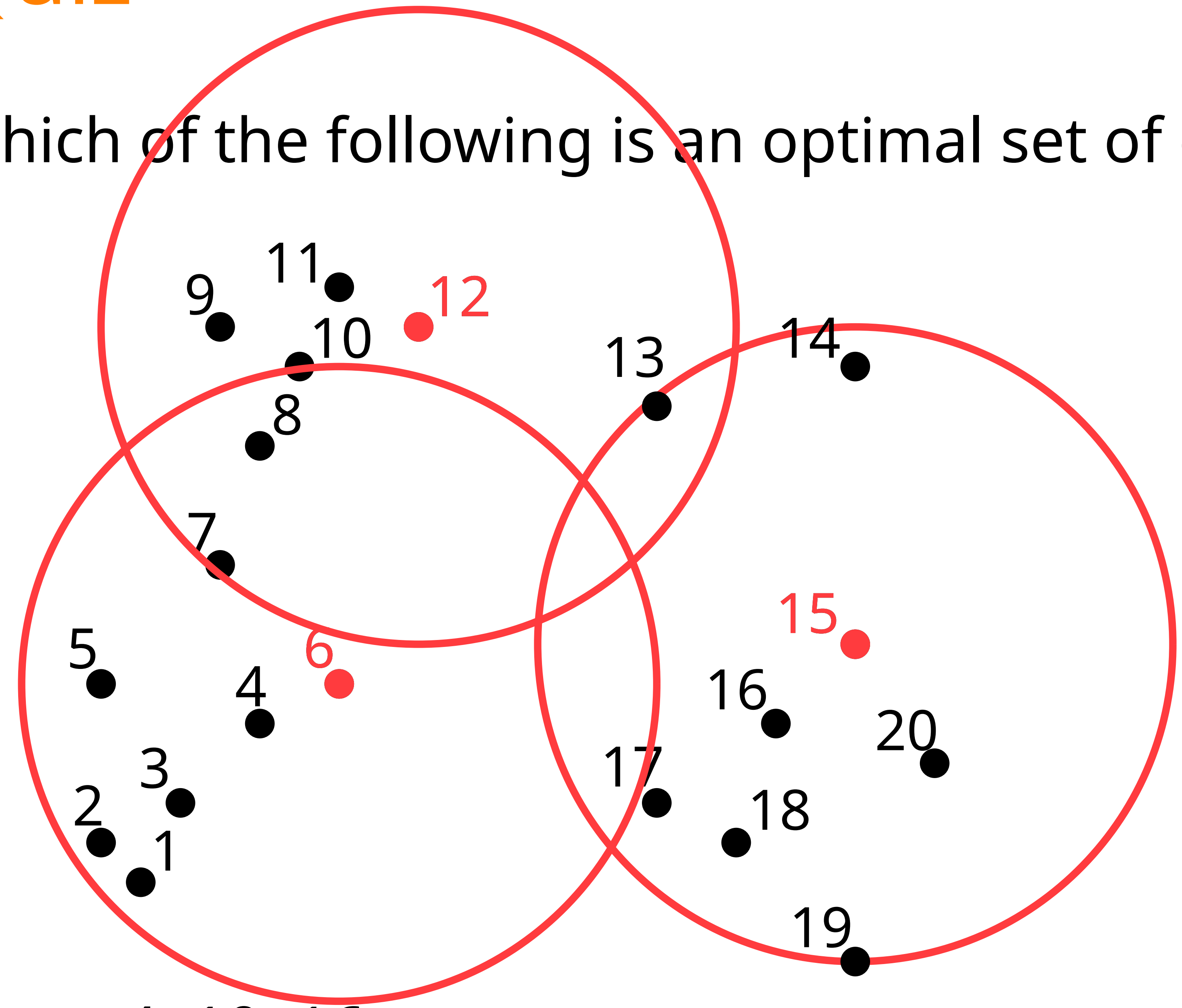
A 4, 10, 16

B 6, 12, 15

C 7, 13, 16

Quiz

Which of the following is an optimal set of centers for $k = 3$?



This problem is NP-hard

A 4, 10, 16

B 6, 12, 15

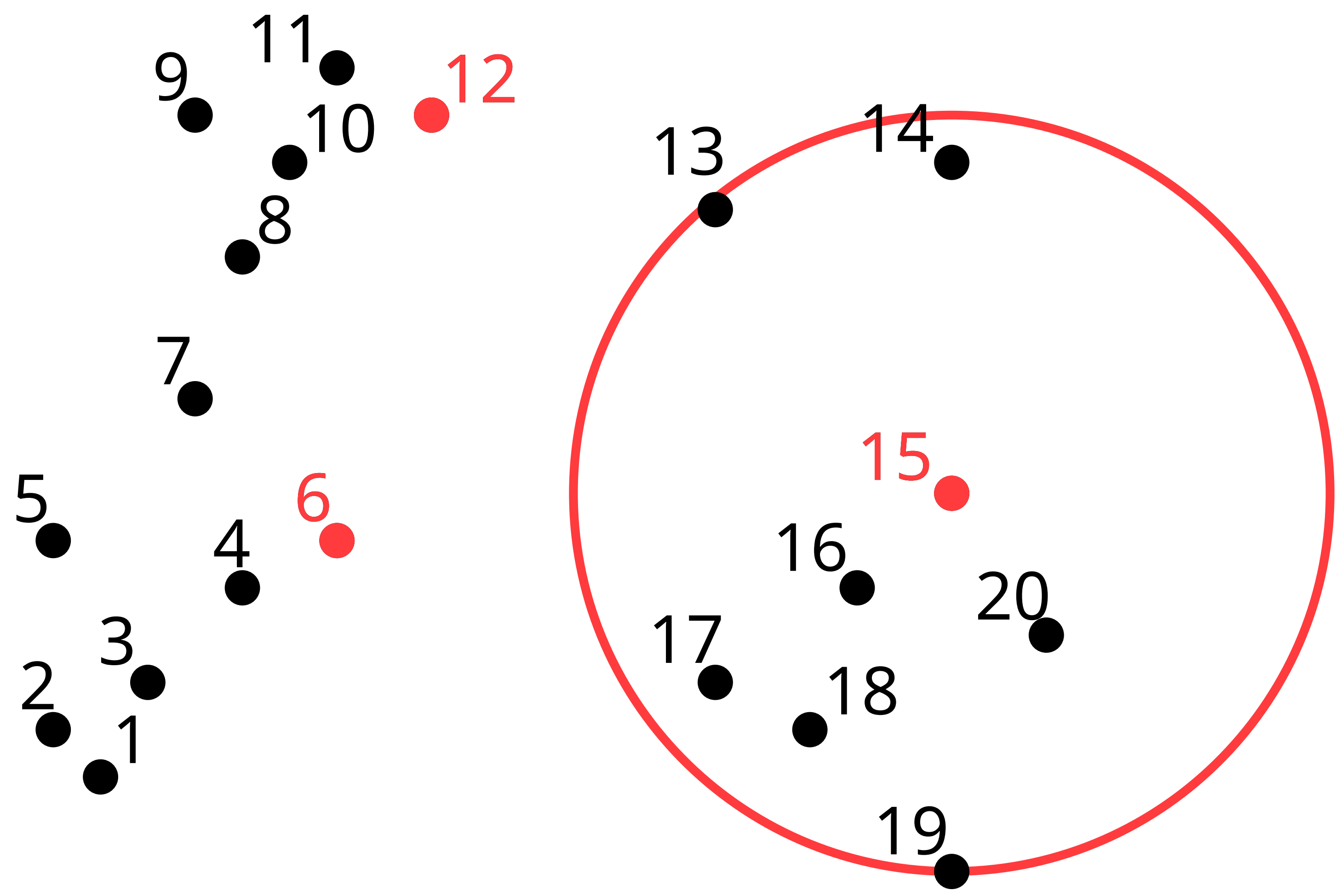
C 7, 13, 16

k-center clustering

approximation algorithm

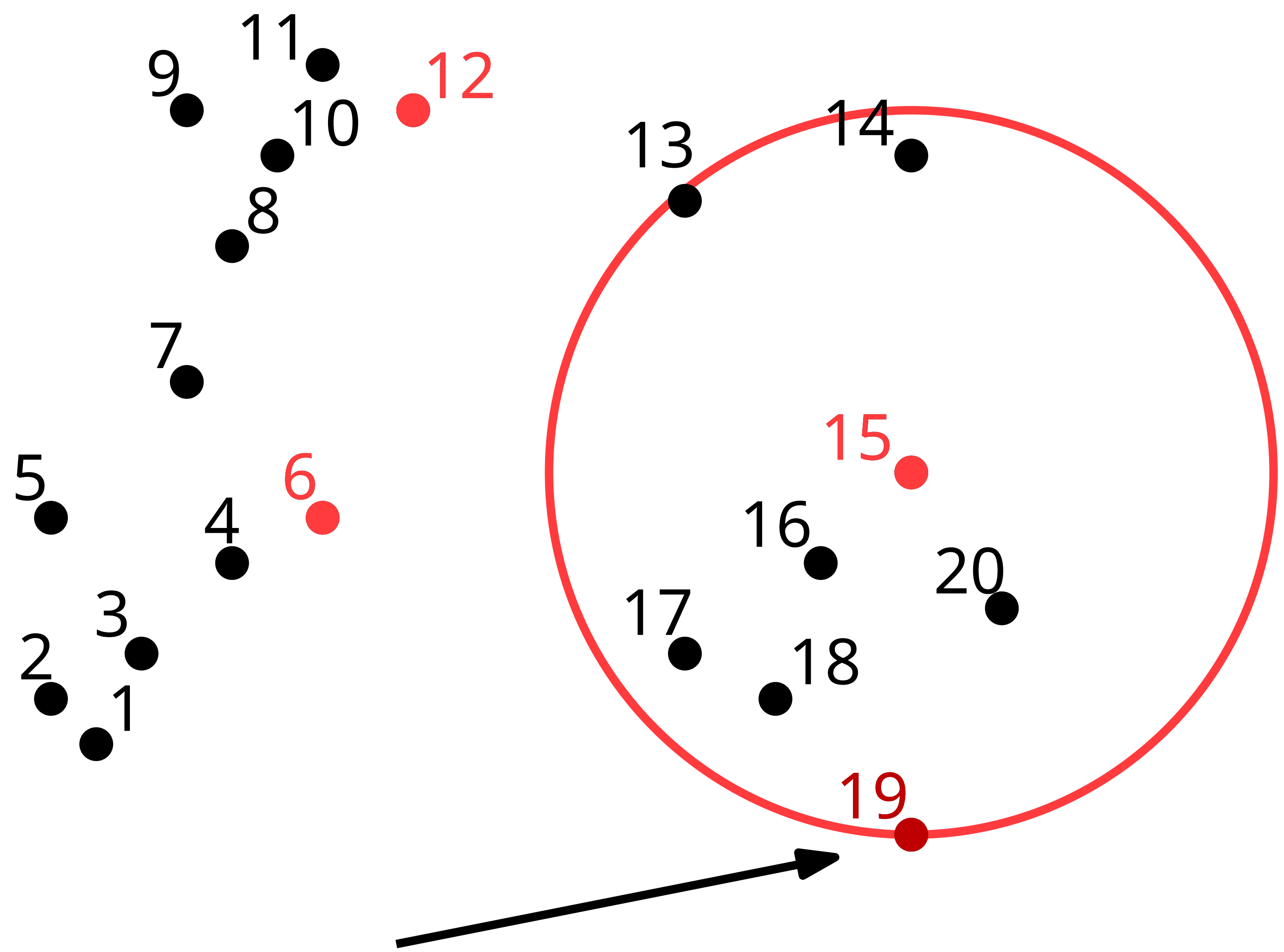
Algorithm GreedyKCenter(P, k)

Incrementally add points to C. How can we guarantee to reduce the maximum?



Algorithm GreedyKCenter(P, k)

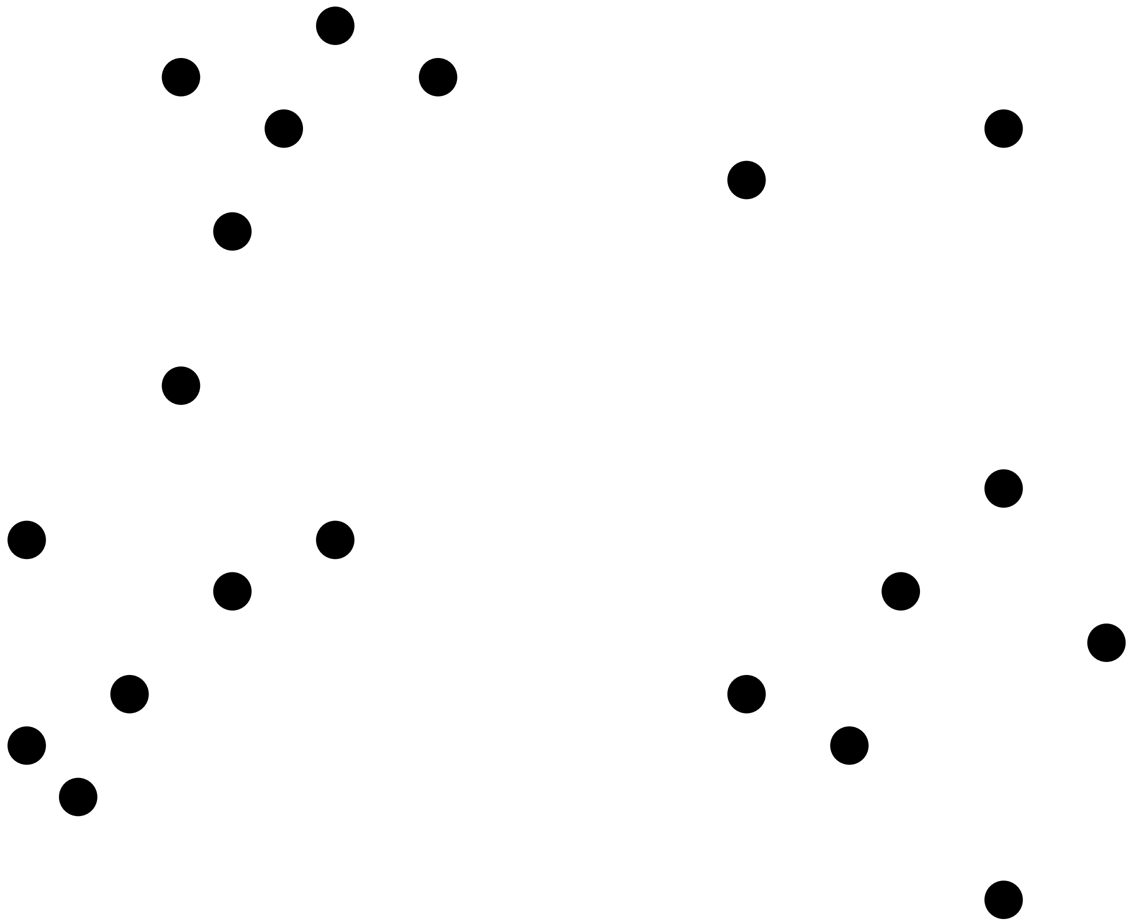
Incrementally add points to C. How can we guarantee to reduce the maximum?



Add the point p with maximum $d(p, C)$!

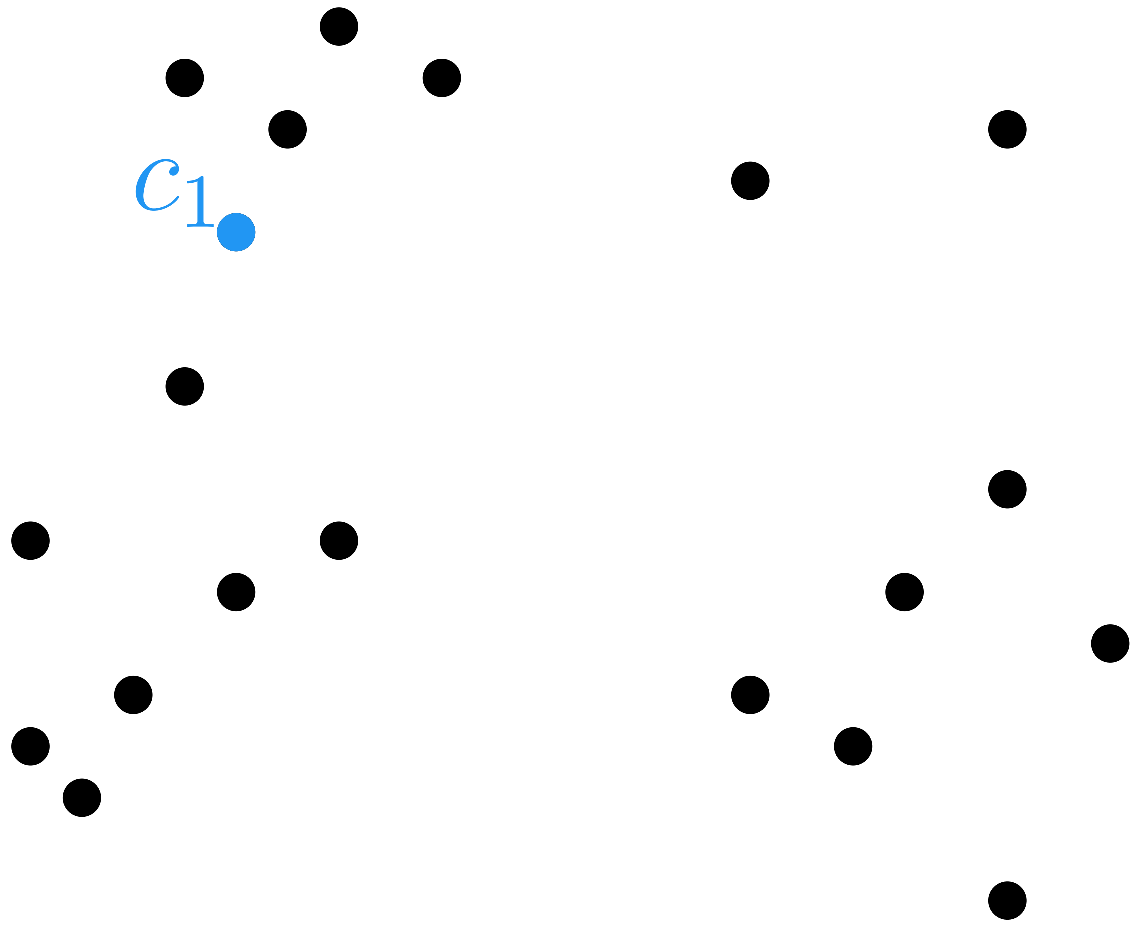
Algorithm GreedyKCenter(P, k)

$c_1 \leftarrow$ arbitrary point of P
$C_1 \leftarrow \{c_1\}$
for $i = 2, 3, \ldots, k$:
    Let $c_i \in P$ be the point such that $d(c_i, C_{i-1})$ is maximal
    $C_i \leftarrow C_{i-1} \cup s_i$
return C_k



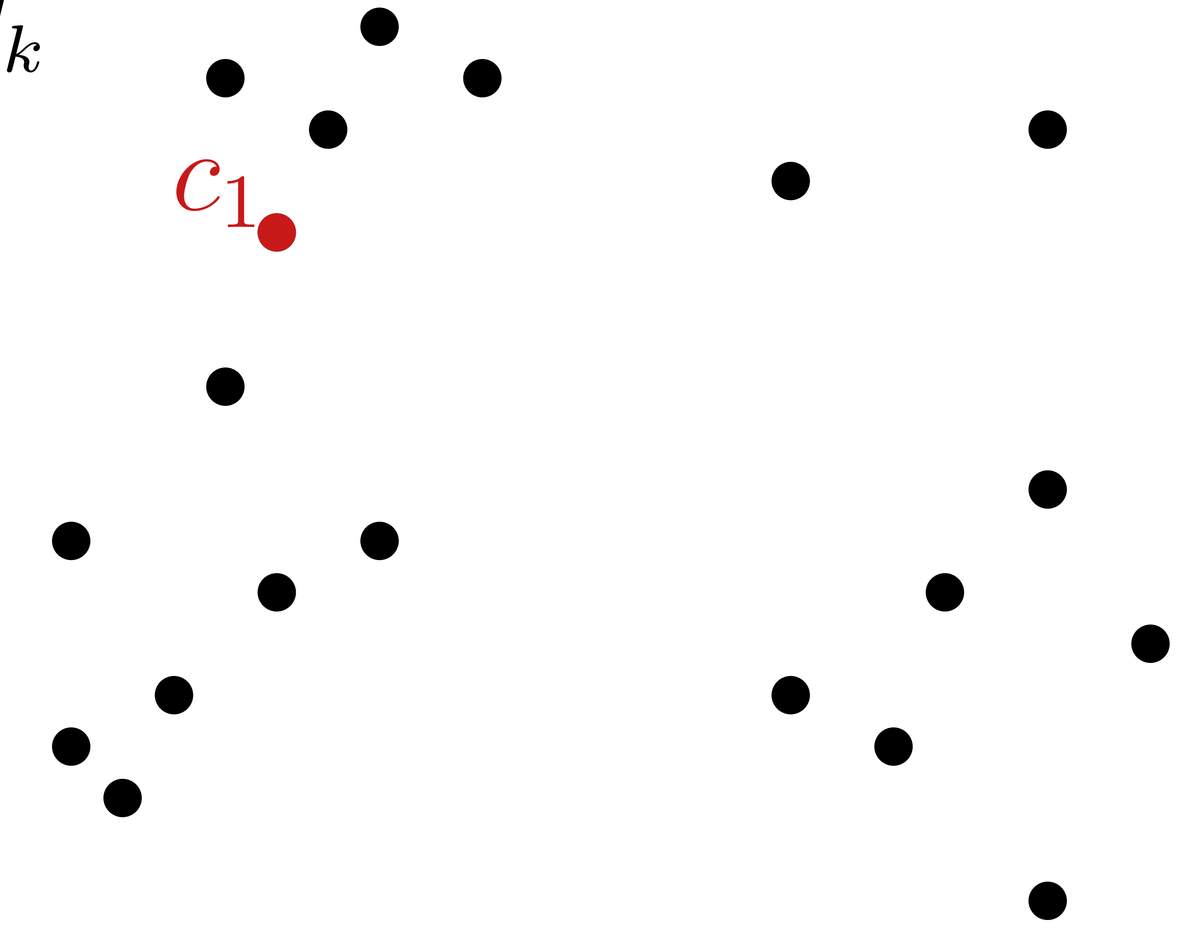
Algorithm GreedyKCenter(P, k)

$c_1 \leftarrow$ arbitrary point of P
$C_1 \leftarrow \{c_1\}$
for $i = 2, 3, \ldots, k$:
    Let $c_i \in P$ be the point such that $d(c_i, C_{i-1})$ is maximal
    $C_i \leftarrow C_{i-1} \cup s_i$
return C_k



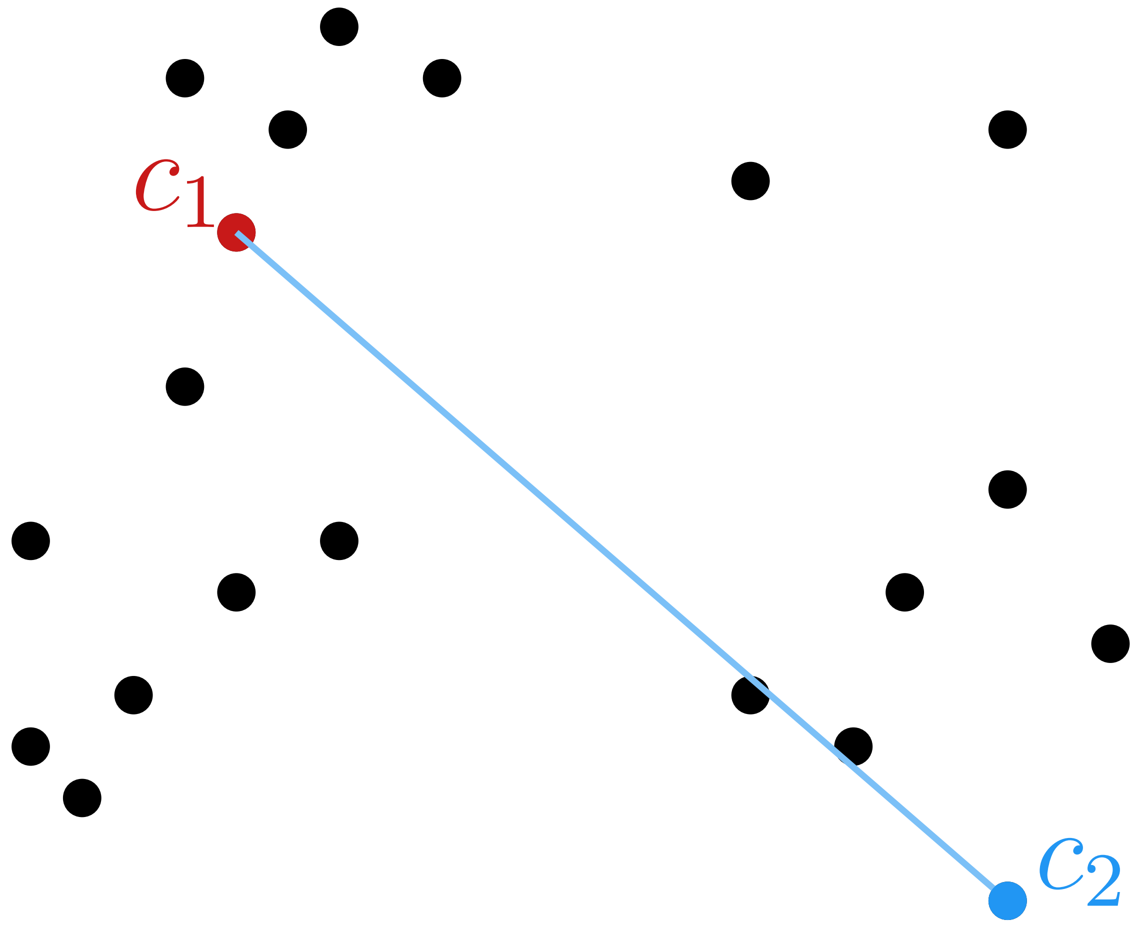
Algorithm GreedyKCenter(P, k)

$c_1 \leftarrow$ arbitrary point of P
$C_1 \leftarrow \{c_1\}$
for $i = 2, 3, \ldots, k$:
    Let $c_i \in P$ be the point such that $d(c_i, C_{i-1})$ is maximal
    $C_i \leftarrow C_{i-1} \cup s_i$
return C_k



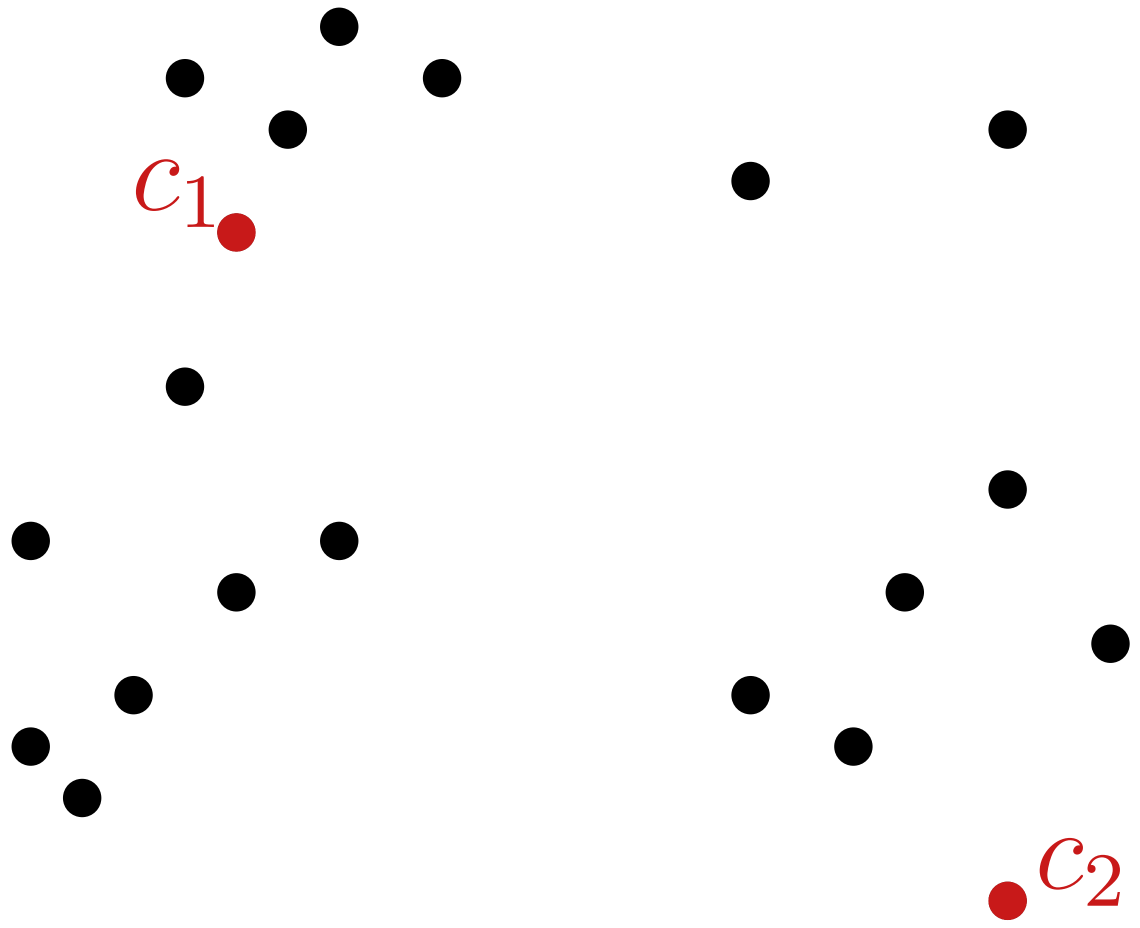
Algorithm GreedyKCenter(P, k)

```
c_1 ← arbitrary point of P
C_1 ← {c_1}
for i = 2, 3, ..., k:
    Let c_i ∈ P be the point such that d(c_i, C_{i-1}) is maximal
    C_i ← C_{i-1} ∪ s_i
return C_k
```



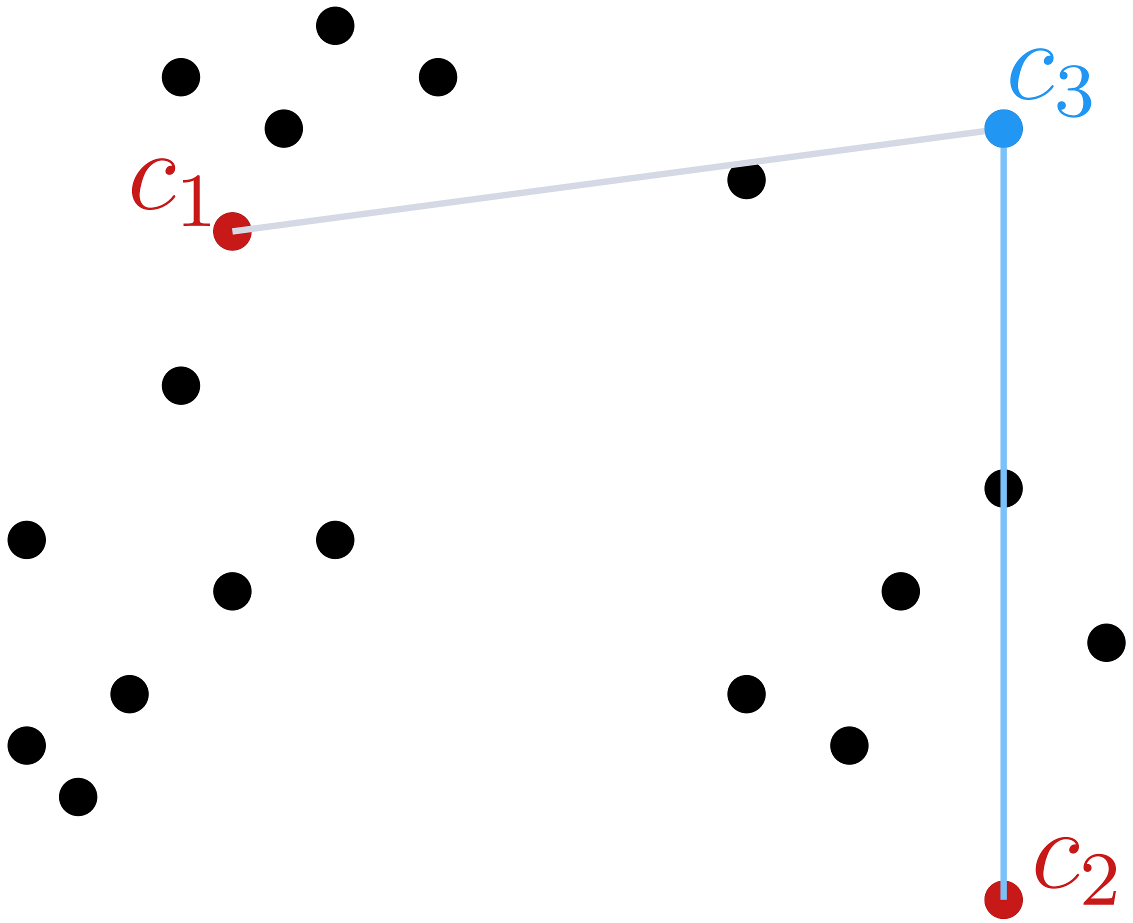
Algorithm GreedyKCenter(P, k)

```
c_1 ← arbitrary point of P
C_1 ← {c_1}
for i = 2, 3, ..., k:
    Let c_i ∈ P be the point such that d(c_i, C_{i−1}) is maximal
    C_i ← C_{i−1} ∪ s_i
return C_k
```



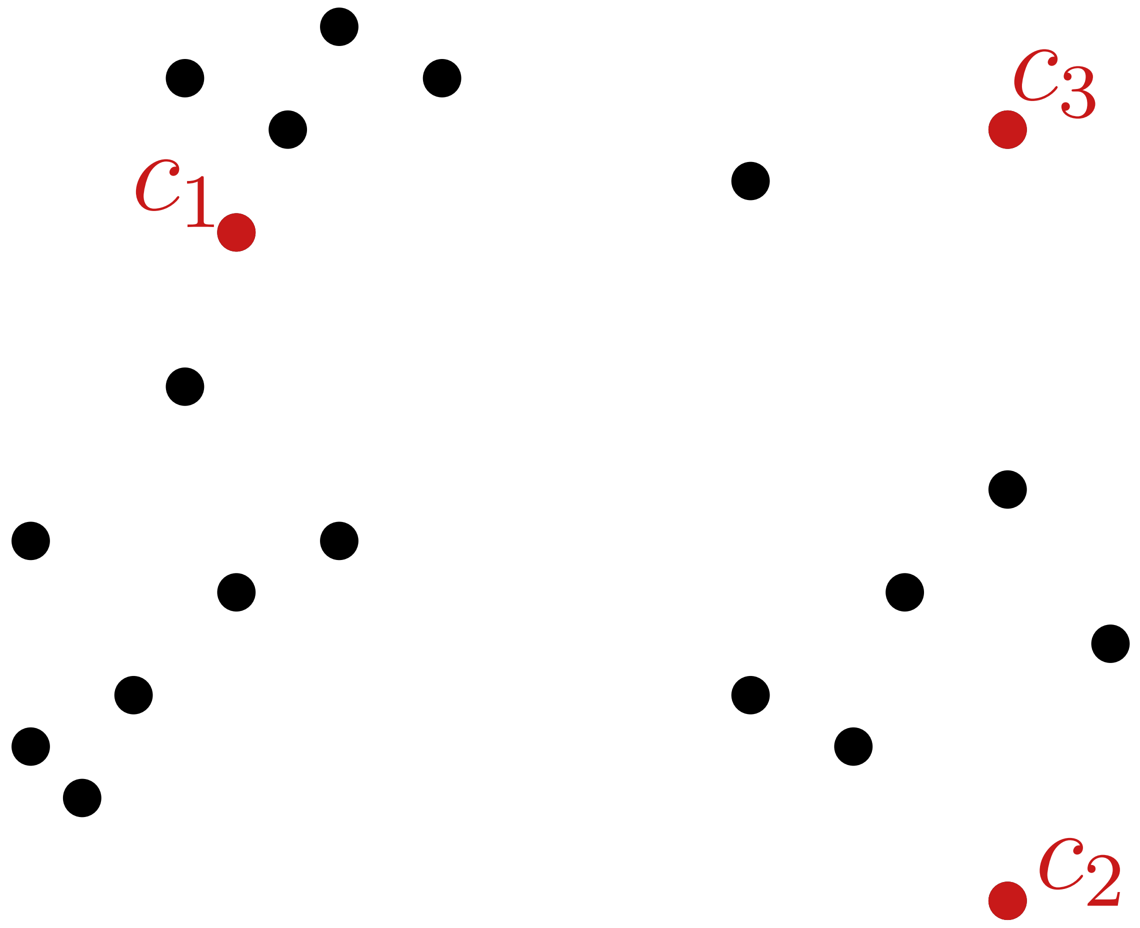
Algorithm GreedyKCenter(P, k)

$c_1 \leftarrow$ arbitrary point of P
$C_1 \leftarrow \{c_1\}$
for $i = 2, 3, \ldots, k$:
    Let $c_i \in P$ be the point such that $d(c_i, C_{i-1})$ is maximal
    $C_i \leftarrow C_{i-1} \cup s_i$
return C_k



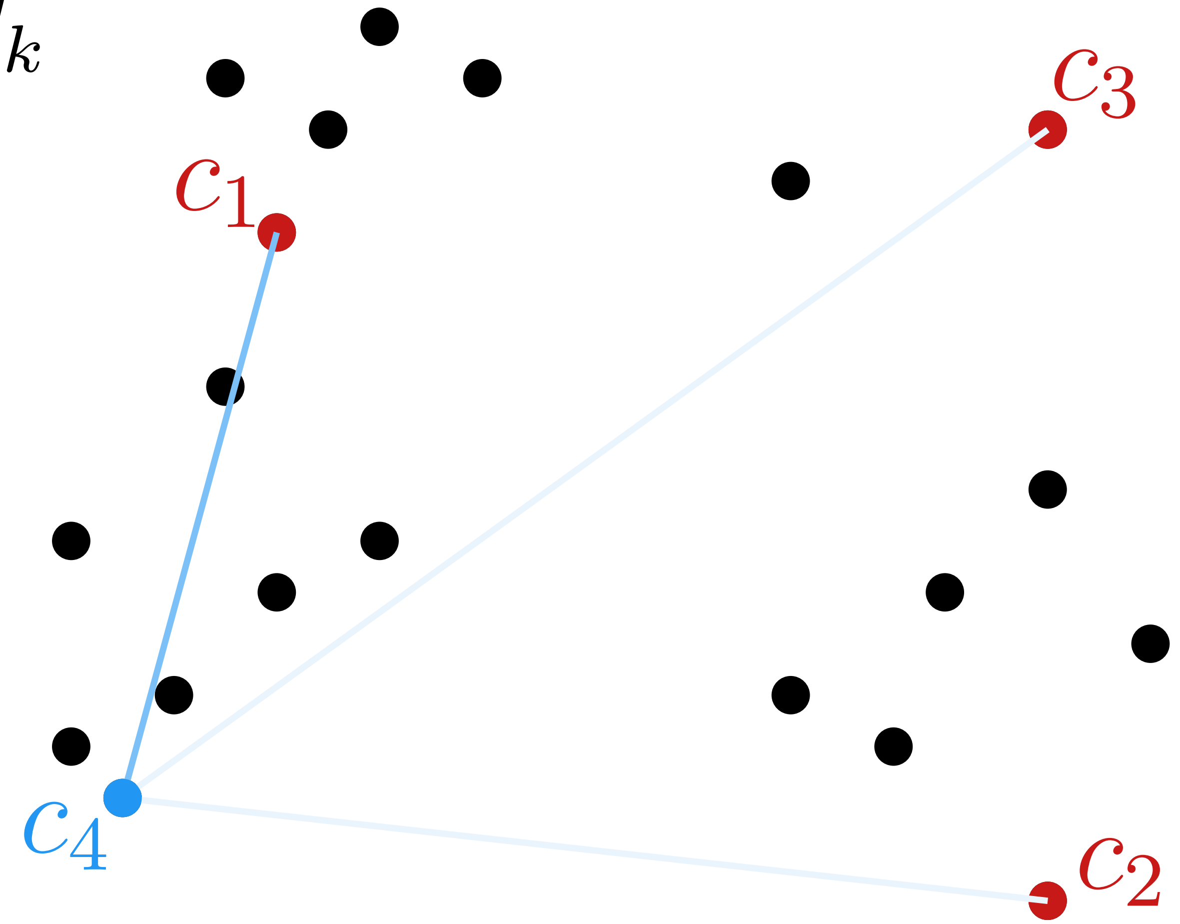
Algorithm GreedyKCenter(P, k)

$c_1 \leftarrow$ arbitrary point of P
$C_1 \leftarrow \{c_1\}$
for $i = 2, 3, \ldots, k$:
    Let $c_i \in P$ be the point such that $d(c_i, C_{i-1})$ is maximal
    $C_i \leftarrow C_{i-1} \cup s_i$
return C_k



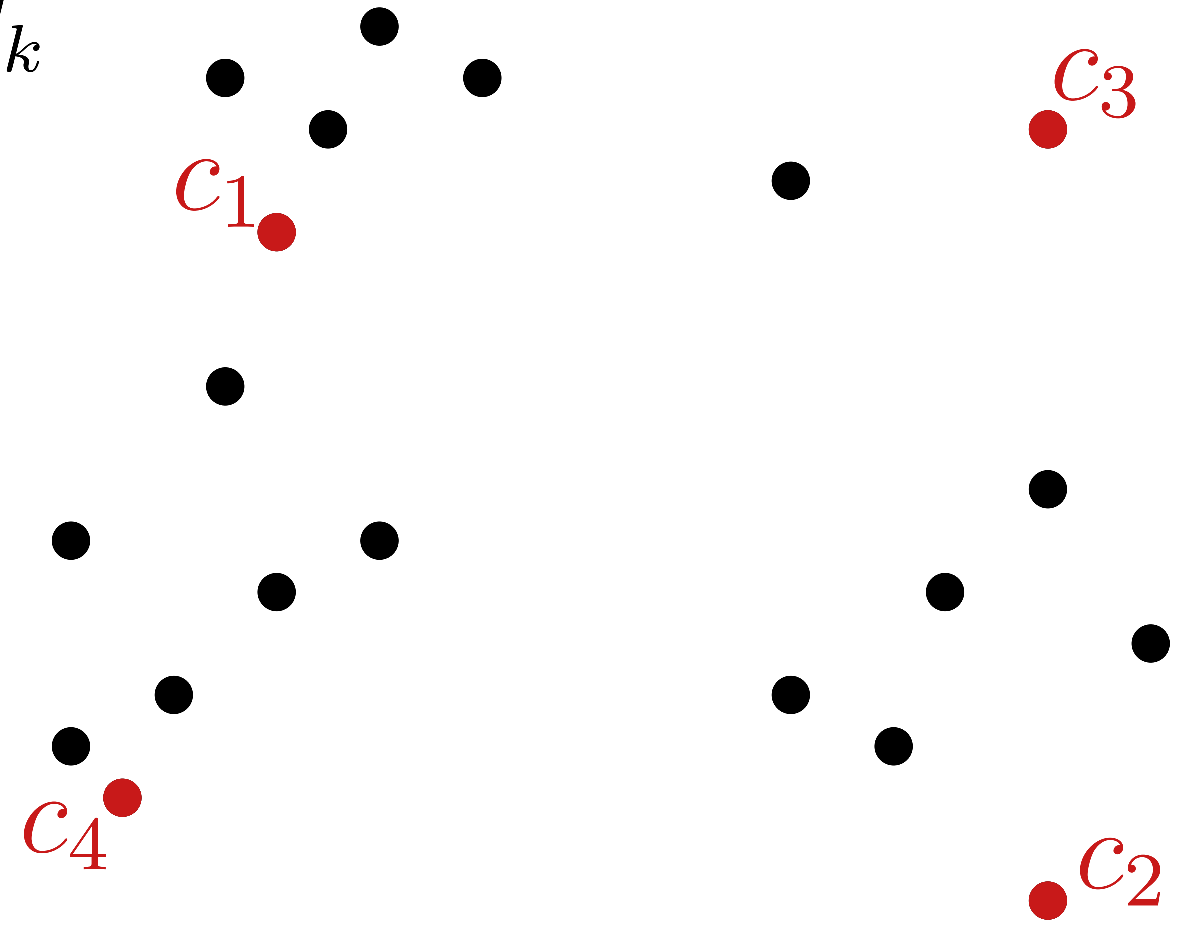
Algorithm GreedyKCenter(P, k)

```
c_1 ← arbitrary point of P
C_1 ← {c_1}
for i = 2, 3, ..., k:
    Let c_i ∈ P be the point such that d(c_i, C_{i−1}) is maximal
    C_i ← C_{i−1} ∪ s_i
return C_k
```



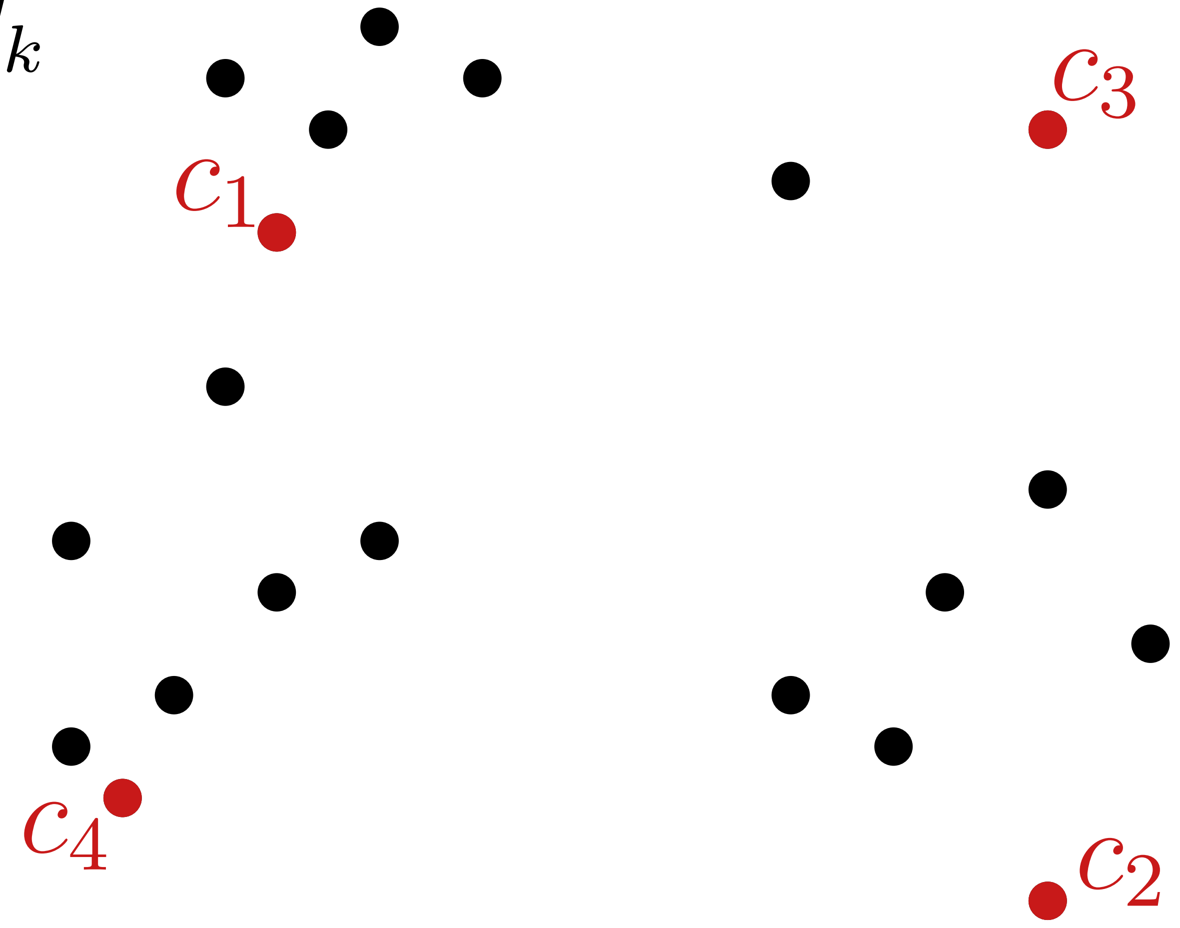
Algorithm GreedyKCenter(P, k)

$c_1 \leftarrow$ arbitrary point of P
$C_1 \leftarrow \{c_1\}$
for $i = 2, 3, \ldots, k$:
    Let $c_i \in P$ be the point such that $d(c_i, C_{i-1})$ is maximal
    $C_i \leftarrow C_{i-1} \cup s_i$
return C_k



Algorithm GreedyKCenter(P, k)

$c_1 \leftarrow$ arbitrary point of P
$C_1 \leftarrow \{c_1\}$
for $i = 2, 3, \ldots, k$:
    Let $c_i \in P$ be the point such that $d(c_i, C_{i-1})$ is maximal
    $C_i \leftarrow C_{i-1} \cup s_i$
return C_k



Approximation factor

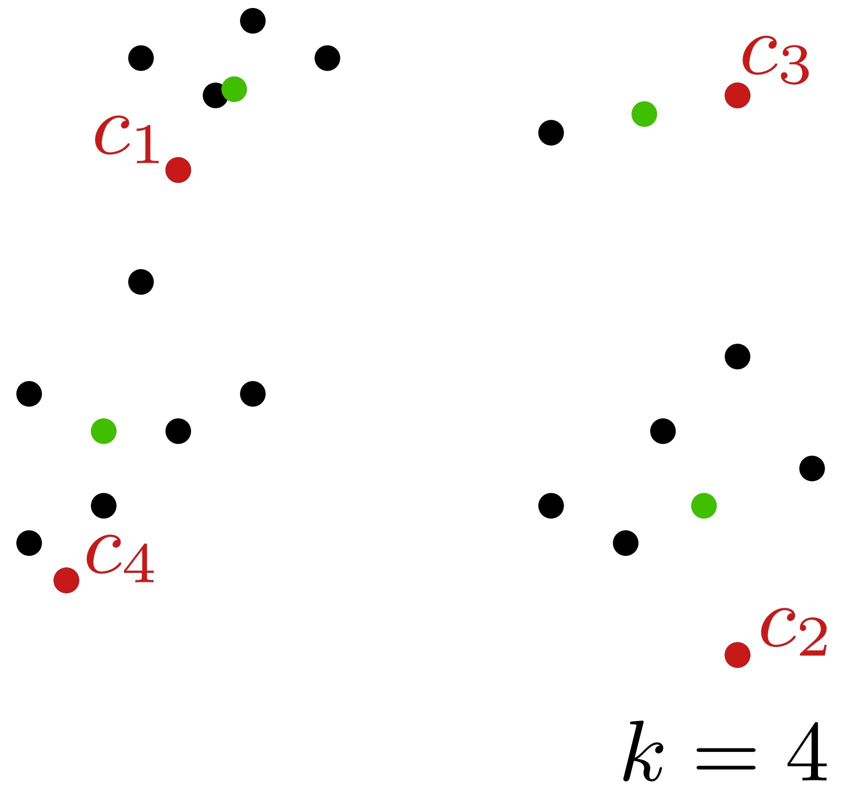
GreedyKCenter(P, k) computes a 2-approximation for k-center clustering.

Approximation factor

GreedyKCenter(P, k) computes a 2-approximation for k-center clustering.

C^*: an optimal solution with $OPT := \max_{p \in P} d(p, C^*)$

$C_k = \{c_1, \ldots, c_k\}$ computed solution



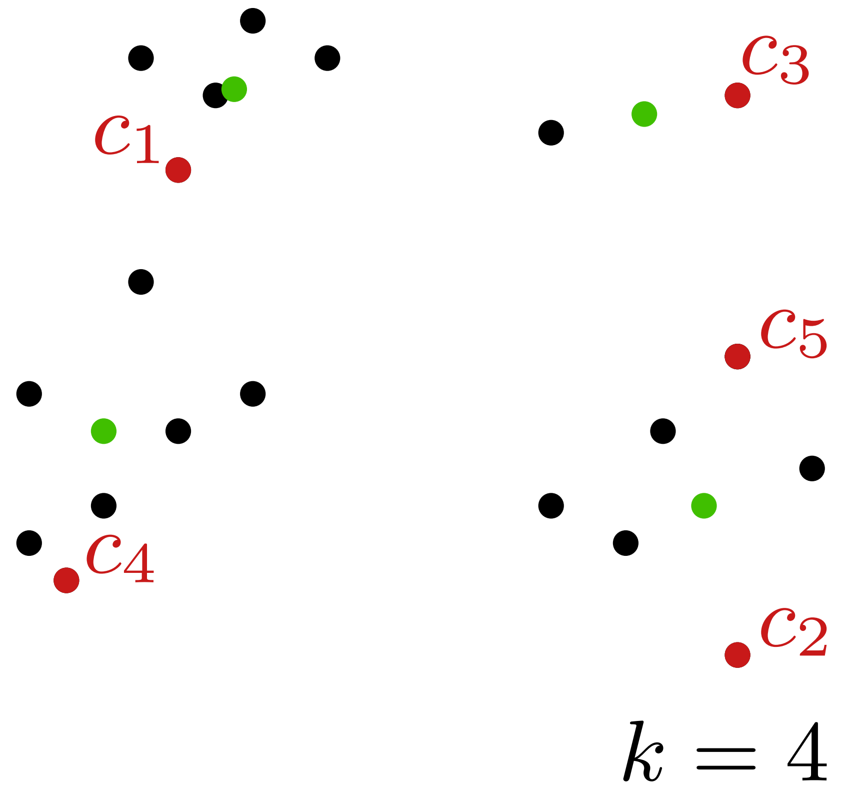
Approximation factor

GreedyKCenter(P, k) computes a 2-approximation for k-center clustering.

C^*: an optimal solution with $OPT := \max_{p \in P} d(p, C^*)$

$C_k = \{c_1, \ldots, c_k\}$ computed solution

c_{k+1}: point maximizing $d(c_{k+1}, C_k) =: r$



Approximation factor

GreedyKCenter(P, k) computes a 2-approximation for k-center clustering.

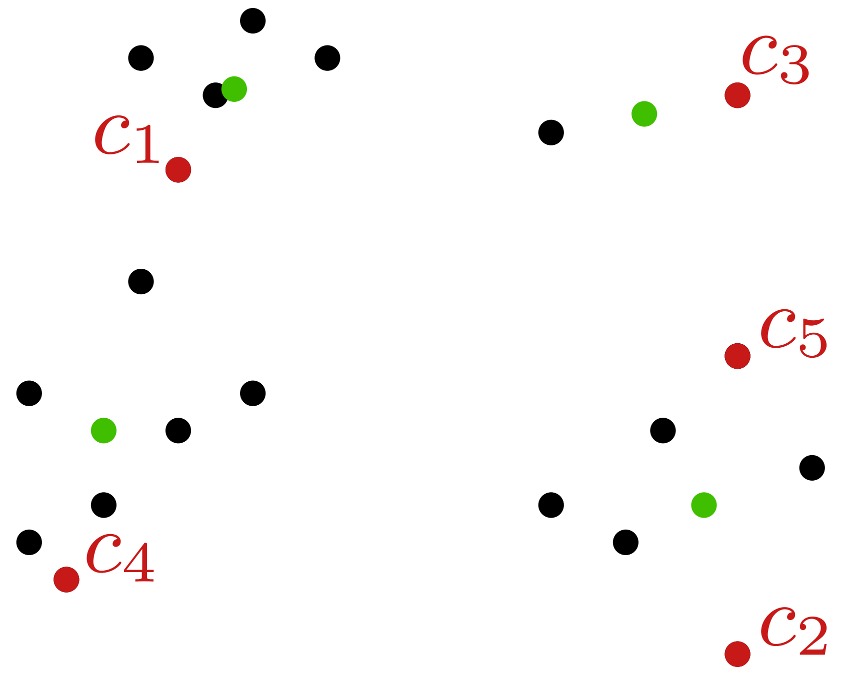
C^*: an optimal solution with $OPT := \max_{p \in P} d(p, C^*)$

$C_k = \{c_1, \ldots, c_k\}$ computed solution

c_{k+1}: point maximizing $d(c_{k+1}, C_k) =: r$

for $i < j$:

$d(c_j, c_i) \geq d(c_j, C_{j-1}) \geq d(c_{k+1}, C_{j-1}) \geq d(c_{k+1}, C_k) = r$



$k = 4$

Approximation factor

GreedyKCenter(P, k) computes a 2-approximation for k-center clustering.

C^*: an optimal solution with $OPT := \max_{p \in P} d(p, C^*)$

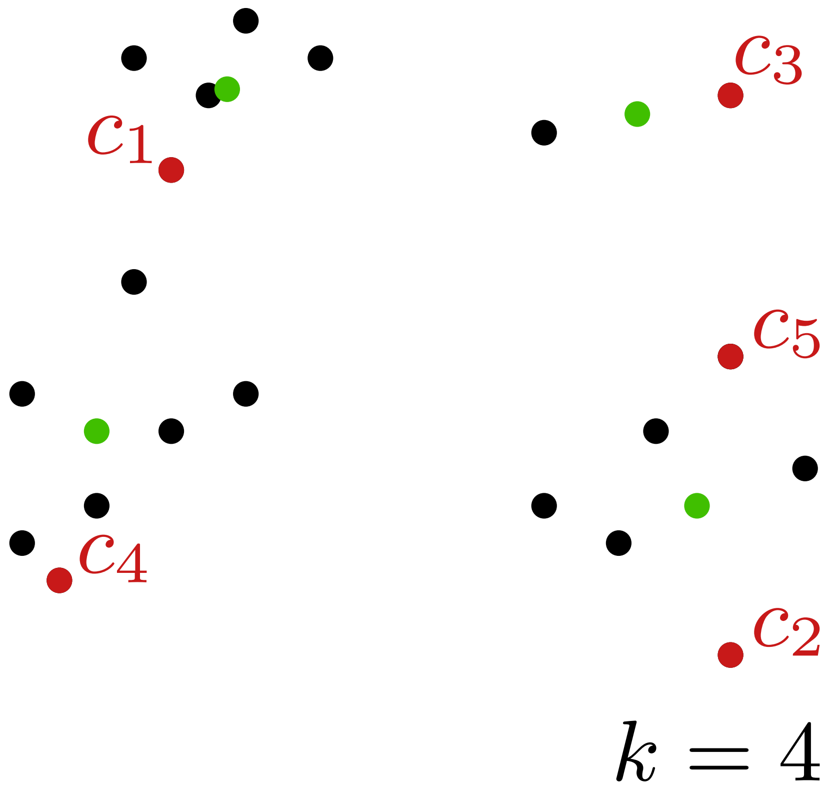
$C_k = \{c_1, \ldots, c_k\}$ computed solution

c_{k+1}: point maximizing $d(c_{k+1}, C_k) =: r$

for $i < j$:

$$d(c_j, c_i) \geq d(c_j, C_{j-1}) \geq d(c_{k+1}, C_{j-1}) \geq d(c_{k+1}, C_k) = r$$

$c_i \in C_{j-1}$



Approximation factor

GreedyKCenter(P, k) computes a 2-approximation for k-center clustering.

C^*: an optimal solution with $OPT := \max_{p \in P} d(p, C^*)$

$C_k = \{c_1, \ldots, c_k\}$ computed solution

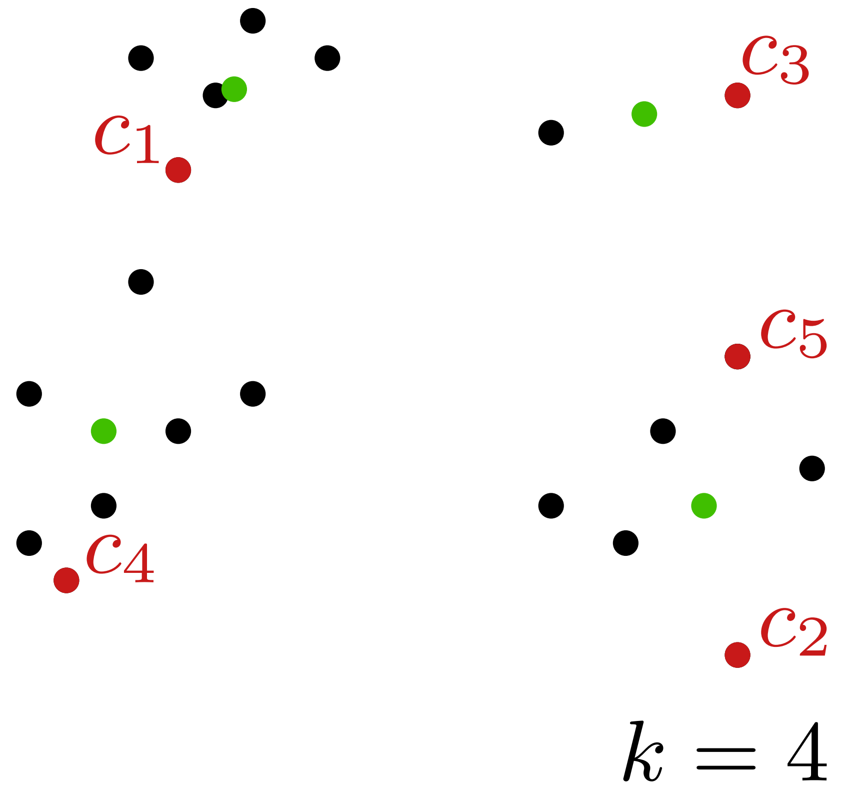
c_{k+1}: point maximizing $d(c_{k+1}, C_k) =: r$

for $i < j$:

$$d(c_j, c_i) \geq d(c_j, C_{j-1}) \geq d(c_{k+1}, C_{j-1}) \geq d(c_{k+1}, C_k) = r$$

$c_i \in C_{j-1}$

c_j had max distance in iteration j



Approximation factor

GreedyKCenter(P, k) computes a 2-approximation for k-center clustering.

C^*: an optimal solution with $OPT := \max_{p \in P} d(p, C^*)$

$C_k = \{c_1, \ldots, c_k\}$ computed solution

c_{k+1}: point maximizing $d(c_{k+1}, C_k) =: r$

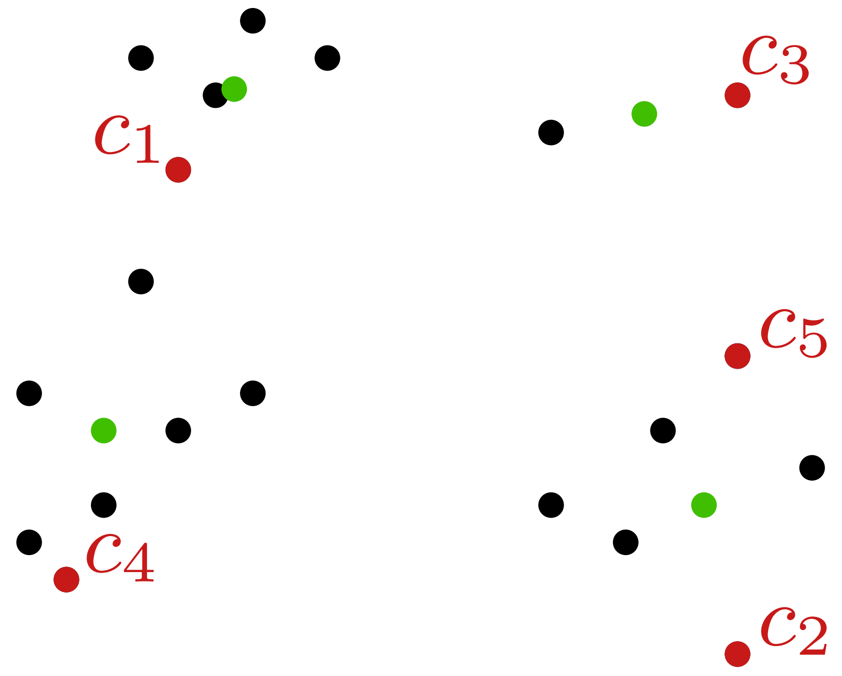
for $i < j$:

$$d(c_j, c_i) \geq d(c_j, C_{j-1}) \geq d(c_{k+1}, C_{j-1}) \geq d(c_{k+1}, C_k) = r$$

$c_i \in C_{j-1}$

c_j had max distance in iteration j

$C_{j-1} \subset C_k$



$k = 4$

Approximation factor

GreedyKCenter(P, k) computes a 2-approximation for k-center clustering.

C^*: an optimal solution with $OPT := \max_{p \in P} d(p, C^*)$

$C_k = \{c_1, \ldots, c_k\}$ computed solution

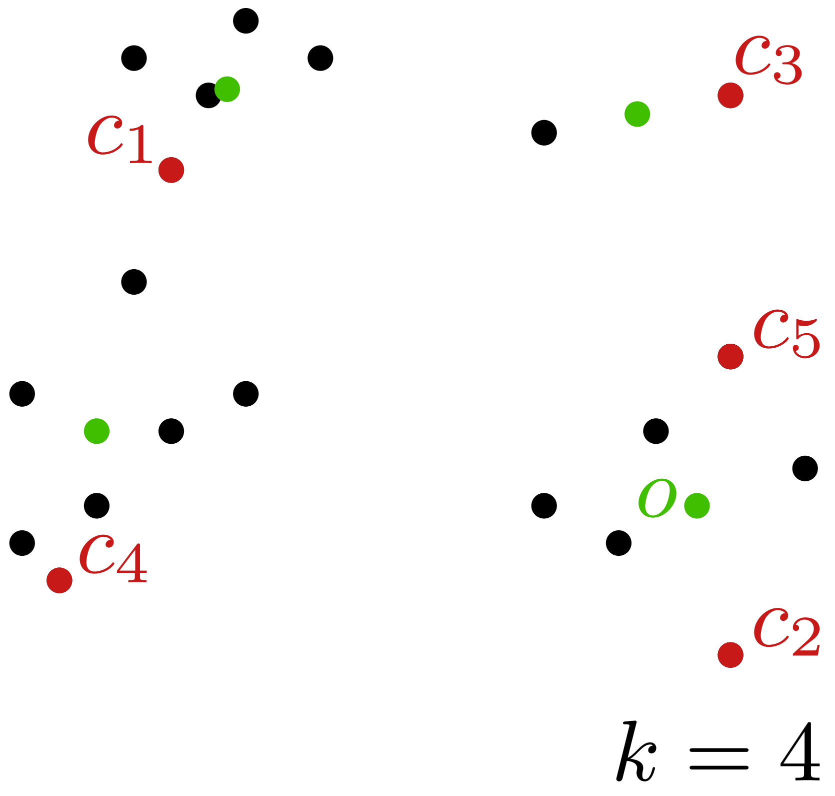
c_{k+1}: point maximizing $d(c_{k+1}, C_k) =: r$

for $i < j$:

$d(c_j, c_i) \geq d(c_j, C_{j-1}) \geq d(c_{k+1}, C_{j-1}) \geq d(c_{k+1}, C_k) = r$

pigeonhole principle:

$\exists c_i, c_j$ in the same cluster of C^*; $o :=$ corresponding center



Approximation factor

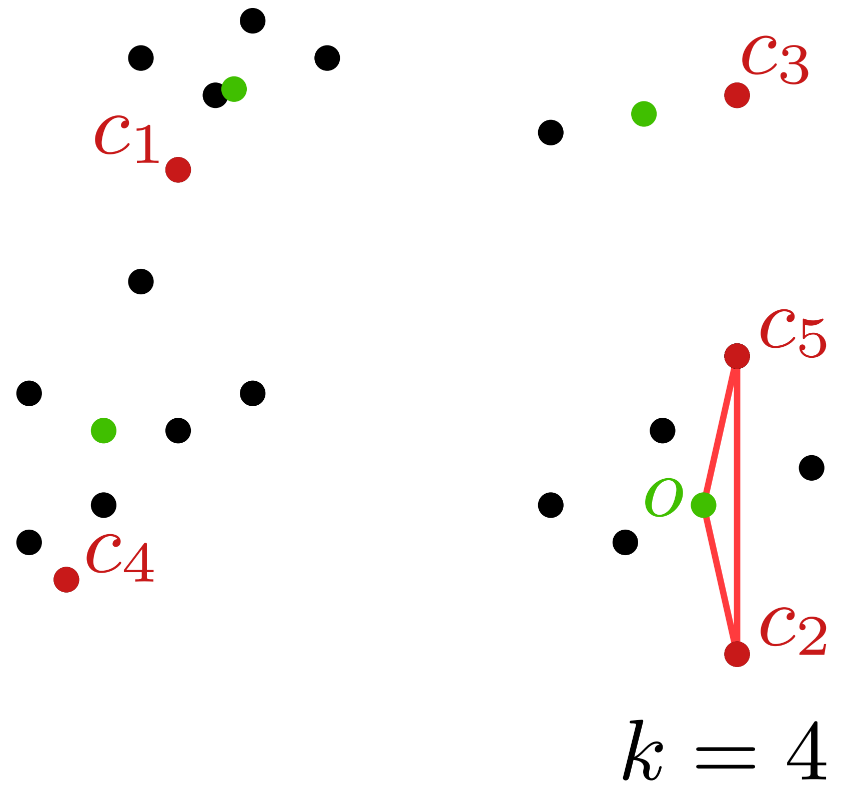
GreedyKCenter(P, k) computes a 2-approximation for k-center clustering.

C^*: an optimal solution with $OPT := \max_{p \in P} d(p, C^*)$

$C_k = \{c_1, \ldots, c_k\}$ computed solution

c_{k+1}: point maximizing $d(c_{k+1}, C_k) =: r$

for $i < j$:
$d(c_j, c_i) \geq d(c_j, C_{j-1}) \geq d(c_{k+1}, C_{j-1}) \geq d(c_{k+1}, C_k) = r$



pigeonhole principle:

$\exists c_i, c_j$ in the same cluster of C^*; $o :=$ corresponding center

triangle inequality:

$r \leq d(c_j, c_i) \leq d(c_j, o) + d(o, c_i) \leq 2OPT$

Quiz

The proof that GreedyKCenter gives a 2-approximation works . . .

A only in R^2 with Euclidean distance

B in R^d but only with Euclidean distance

C in any metric space

Quiz

The proof that GreedyKCenter gives a 2-approximation works . . .

A only in R^2 with Euclidean distance

B in R^d but only with Euclidean distance

C in any metric space

since it only uses the triangle inequality

Quiz

The proof that GreedyKCenter gives a 2-approximation works . . .

A only in R^2 with Euclidean distance

B in R^d but only with Euclidean distance

C in any metric space

since it only uses the triangle inequality

When k is part of the input, the k-center problem is NP-hard to approximate within a factor

$2 - \varepsilon$ for general metric spaces

Quiz

The proof that GreedyKCenter gives a 2-approximation works . . .

A only in R^2 with Euclidean distance

B in R^d but only with Euclidean distance

C in any metric space

since it only uses the triangle inequality

When k is part of the input, the k-center problem is NP-hard to approximate within a factor

$2 - \varepsilon$ for general metric spaces

1.82 for R^2 with Euclidean distance

Quiz

The proof that GreedyKCenter gives a 2-approximation works . . .

A only in R^2 with Euclidean distance

B in R^d but only with Euclidean distance

C in any metric space

since it only uses the triangle inequality

When k is part of the input, the k-center problem is NP-hard to approximate within a factor

- $2 - \varepsilon$ for general metric spaces
- 1.82 for R^2 with Euclidean distance
- $2 - \varepsilon$ for R^2 with L_1- or L_∞- distance

discrete k-median clustering

approximation algorithm

discrete k-median clustering in metric space (X, d)

Given: $P \subset X$ and integer k

Goal: Find $C \subset P$ of size k such that

$$\sum_{p \in P} d(p, C)$$

is minimized.

$k = 2$

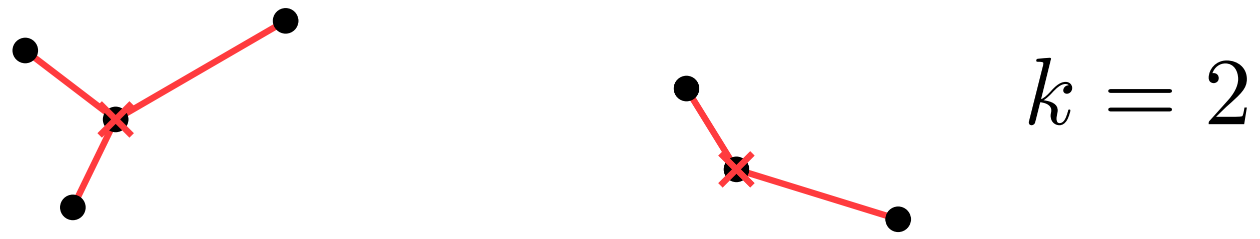
discrete k-median clustering in metric space (X, d)

Given: $P \subset X$ and integer k

Goal: Find $C \subset P$ of size k such that

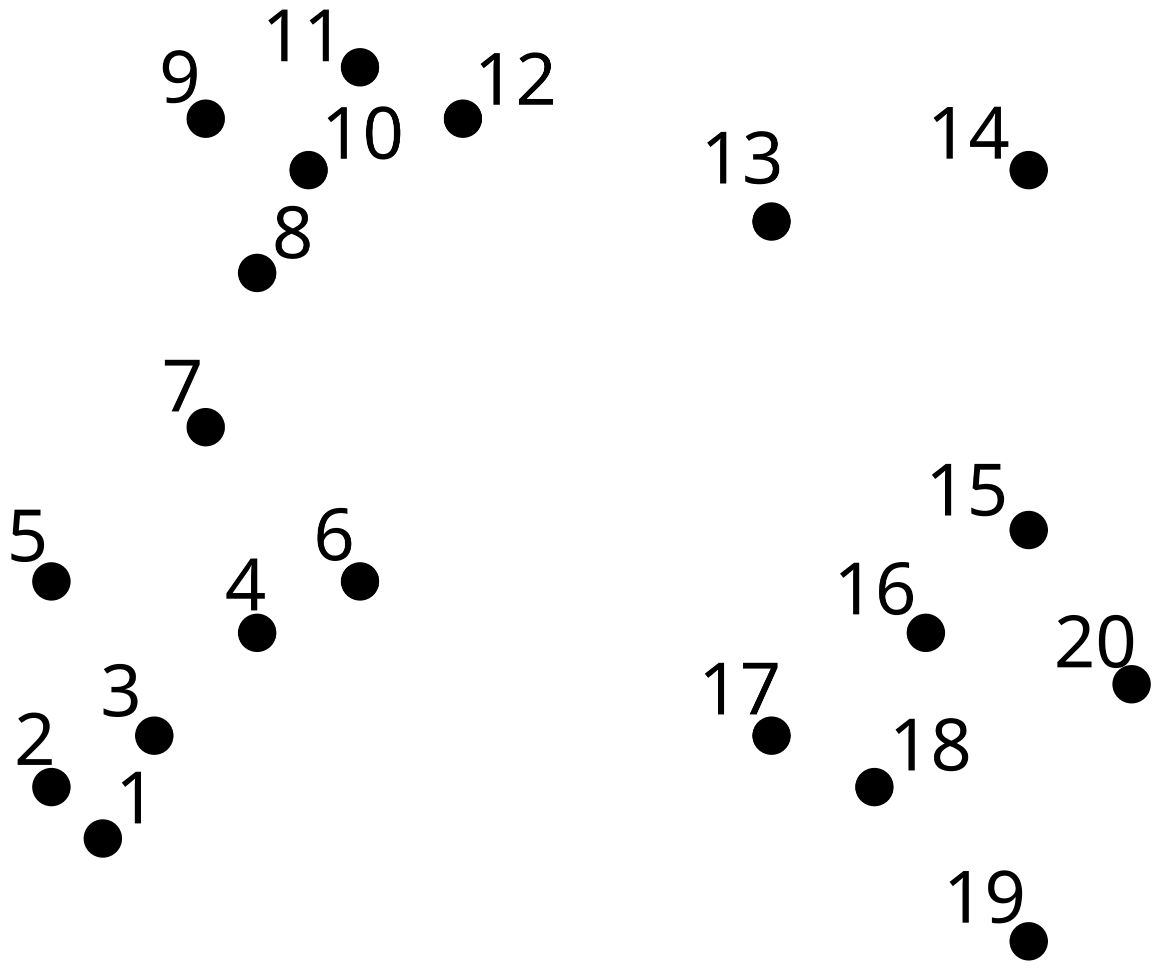
$$\sum_{p \in P} d(p, C)$$

is minimized.



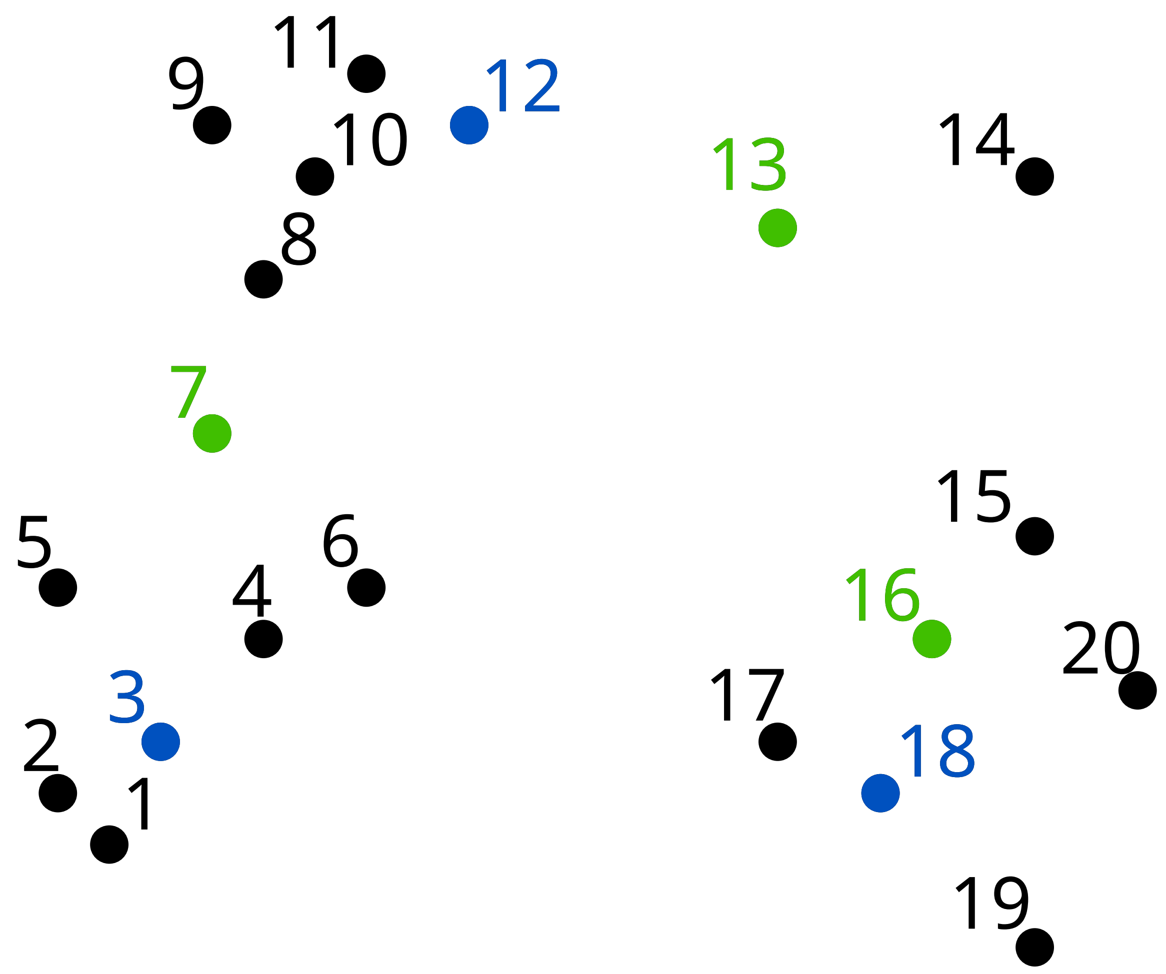
Question ($k = 3$)

Which set C of 3 points minimizes $\sum_{p \in P} d(p, C)$?



Question ($k = 3$)

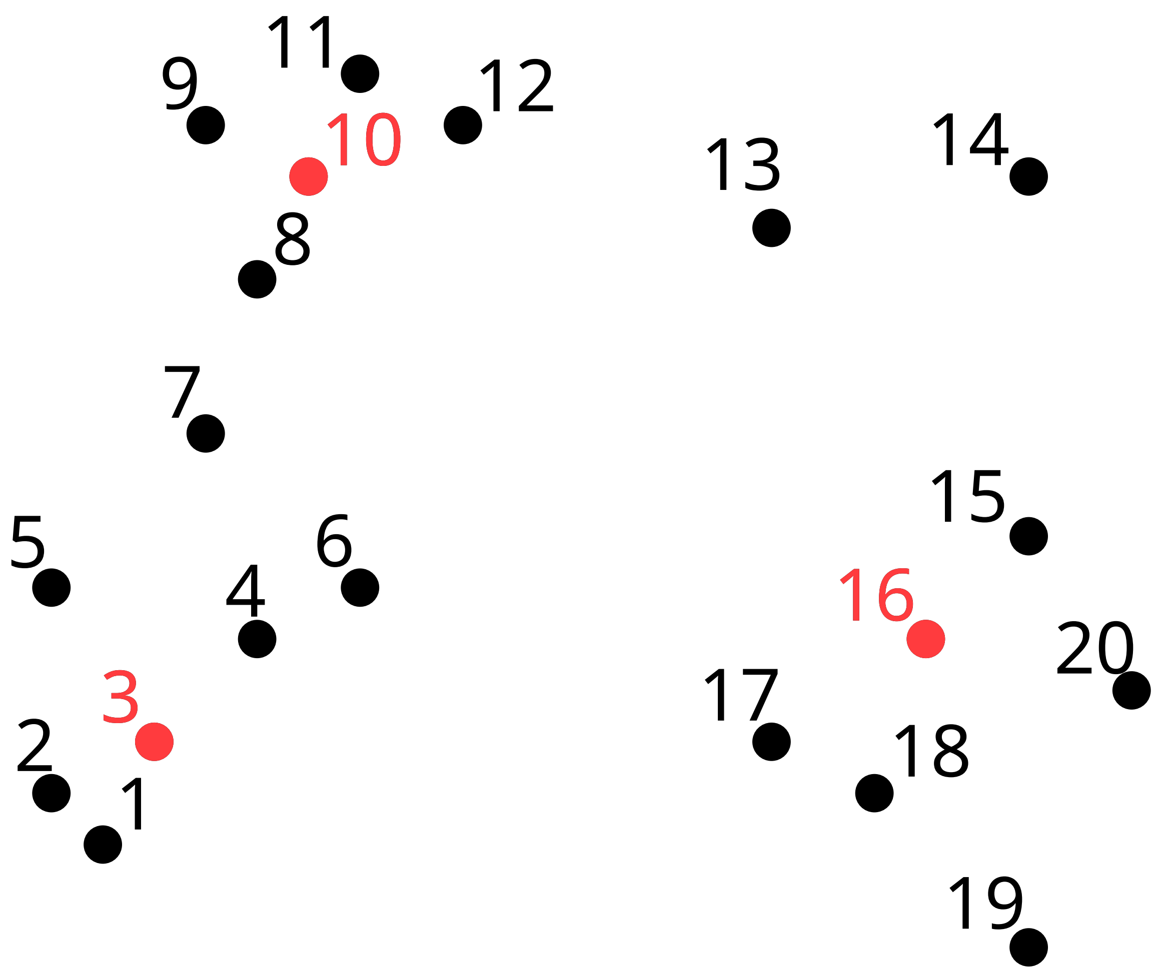
Which set C of 3 points minimizes $\sum_{p \in P} d(p, C)$?



good? $\{3, 12, 18\}$, $\{7, 13, 16\}$

Question ($k = 3$)

Which set C of 3 points minimizes $\sum_{p \in P} d(p, C)$?



good? $\{3, 12, 18\}$, $\{7, 13, 16\}$

optimal: $\{3, 10, 16\}$

GreedyKCenter for k-median?

Use 2-approximation for k-center clustering (?) on n points

$$\max_{p \in P} d(p, C) \leq \sum_{p \in P} d(p, C) \leq \sum_{p \in P} \max_{p \in P} = n \cdot \max_{p \in P}$$

GreedyKCenter for k-median?

Use 2-approximation for k-center clustering (?) on n points

$$\max_{p \in P} d(p, C) \leq \sum_{p \in P} d(p, C) \leq \sum_{p \in P} \max_{p \in P} = n \cdot \max_{p \in P}$$

GreedyKCenter for k-median?

Use 2-approximation for k-center clustering (?) on n points

$$\max_{p\in P} d(p,C) \leq \sum_{p\in P} d(p,C) \leq \sum_{p\in P} \max_{p\in P} = n \cdot \max_{p\in P}$$

This means:

optimal solution to k-center clustering is n-approximation for k-median

GreedyKCenter for k-median?

Use 2-approximation for k-center clustering (?) on n points

$$\max_{p \in P} d(p, C) \leq \sum_{p \in P} d(p, C) \leq \sum_{p \in P} \max_{p \in P} = n \cdot \max_{p \in P}$$

This means:

- optimal solution to k-center clustering is n-approximation for k-median
- 2-approximation for k-center clustering is $2n$-approximation for k-median

GreedyKCenter for k-median?

Use 2-approximation for k-center clustering (?) on n points

$$\max_{p\in P} d(p,C) \leq \sum_{p\in P} d(p,C) \leq \sum_{p\in P} \max_{p\in P} = n \cdot \max_{p\in P}$$

This means:

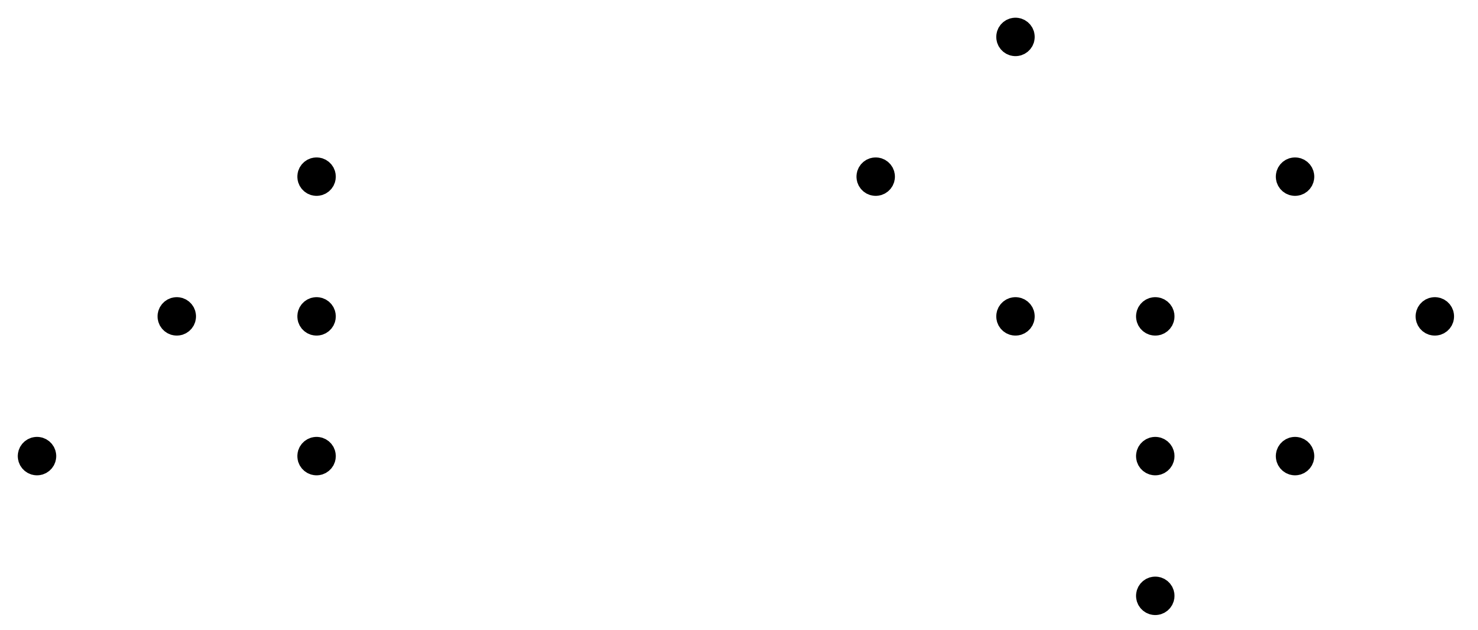
optimal solution to k-center clustering is n-approximation for k-median

2-approximation for k-center clustering is $2n$-approximation for k-median

We can do better with local search!

LocalSearchKMedian(P, k)

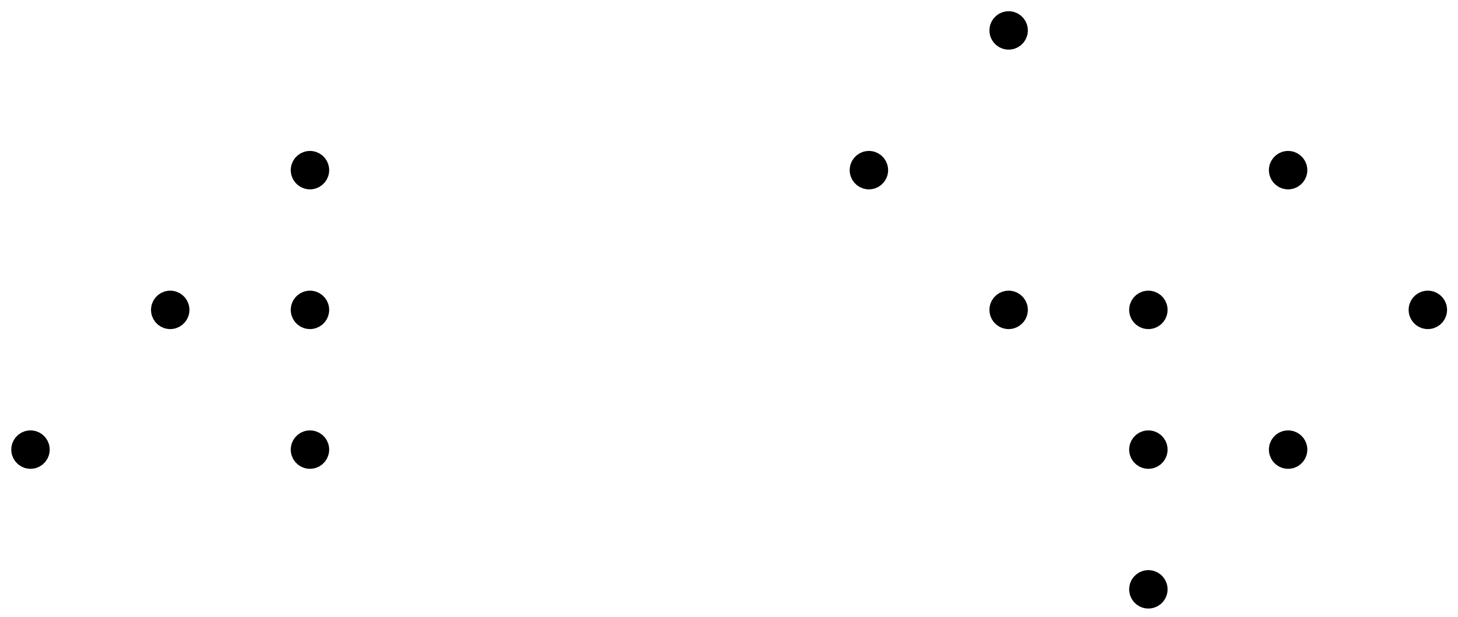
```
C ← GreedyKCenter(P, k)
while ∃p ∈ P \ C, c ∈ C s.t. replacing c by p in C reduces ∑_{p∈P} d(p, C)
by factor 1 − τ:
    C ← C ∪ {p} \ {c}
return C
```



LocalSearchKMedian(P, k)

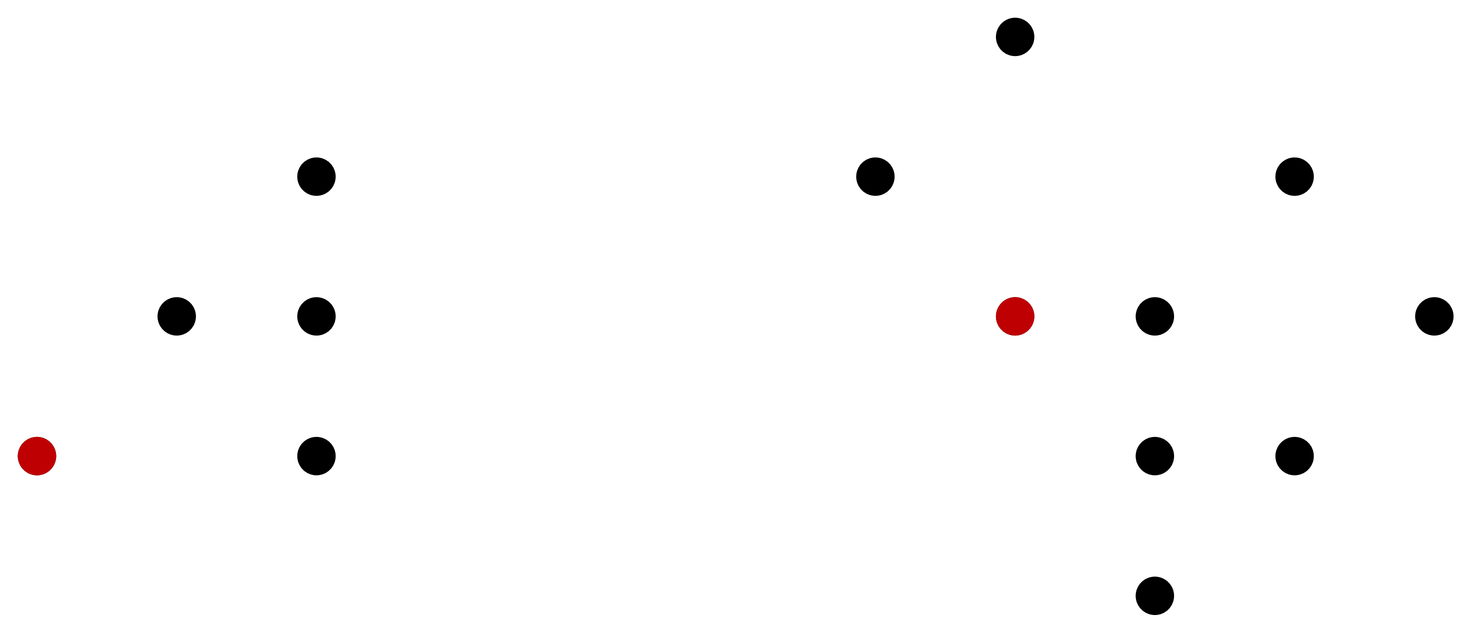
$C \leftarrow$ GreedyKCenter(P, k)
while $\exists p \in P \setminus C, c \in C$ s.t. replacing c by p in C reduces $\sum_{p \in P} d(p, C)$ by factor $1 - \tau$:
    $C \leftarrow C \cup \{p\} \setminus \{c\}$
return C

we choose τ later.



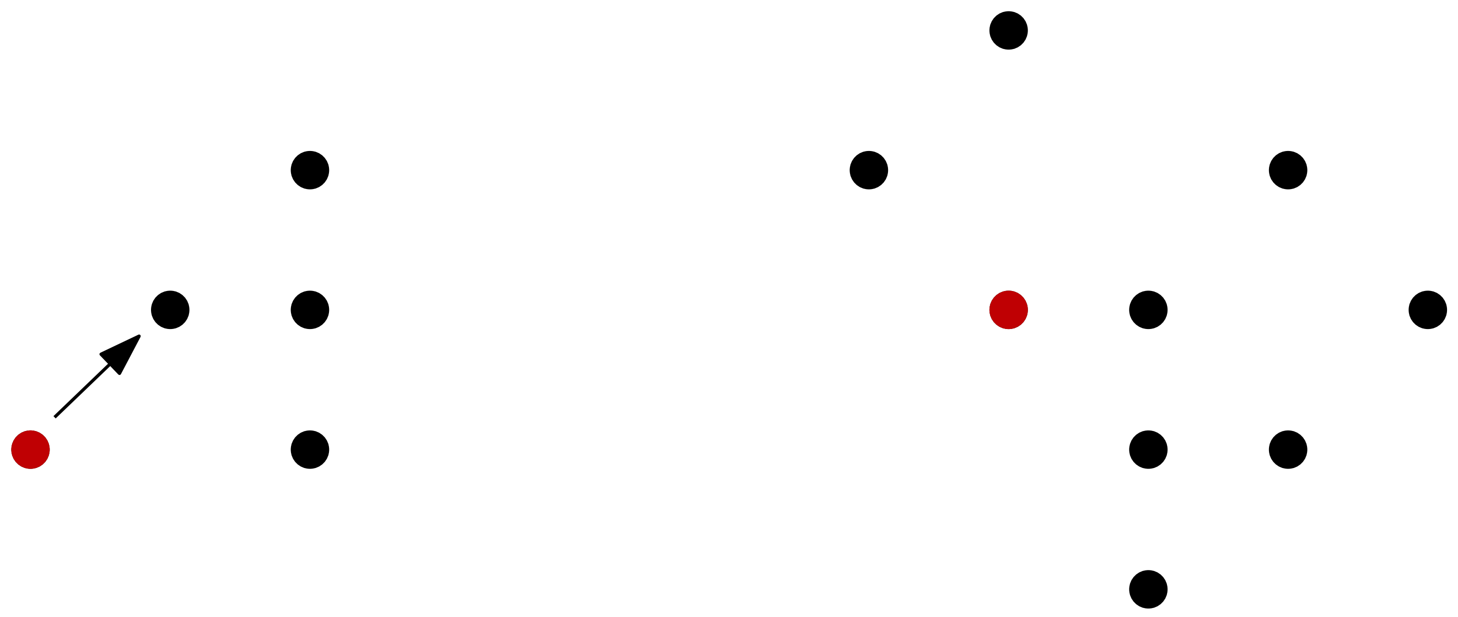
LocalSearchKMedian(P, k)

$C \leftarrow$ GreedyKCenter(P, k)
while $\exists p \in P \setminus C, c \in C$ s.t. replacing c by p in C reduces $\sum_{p \in P} d(p, C)$ by factor $1 - \tau$:
    $C \leftarrow C \cup \{p\} \setminus \{c\}$
return C



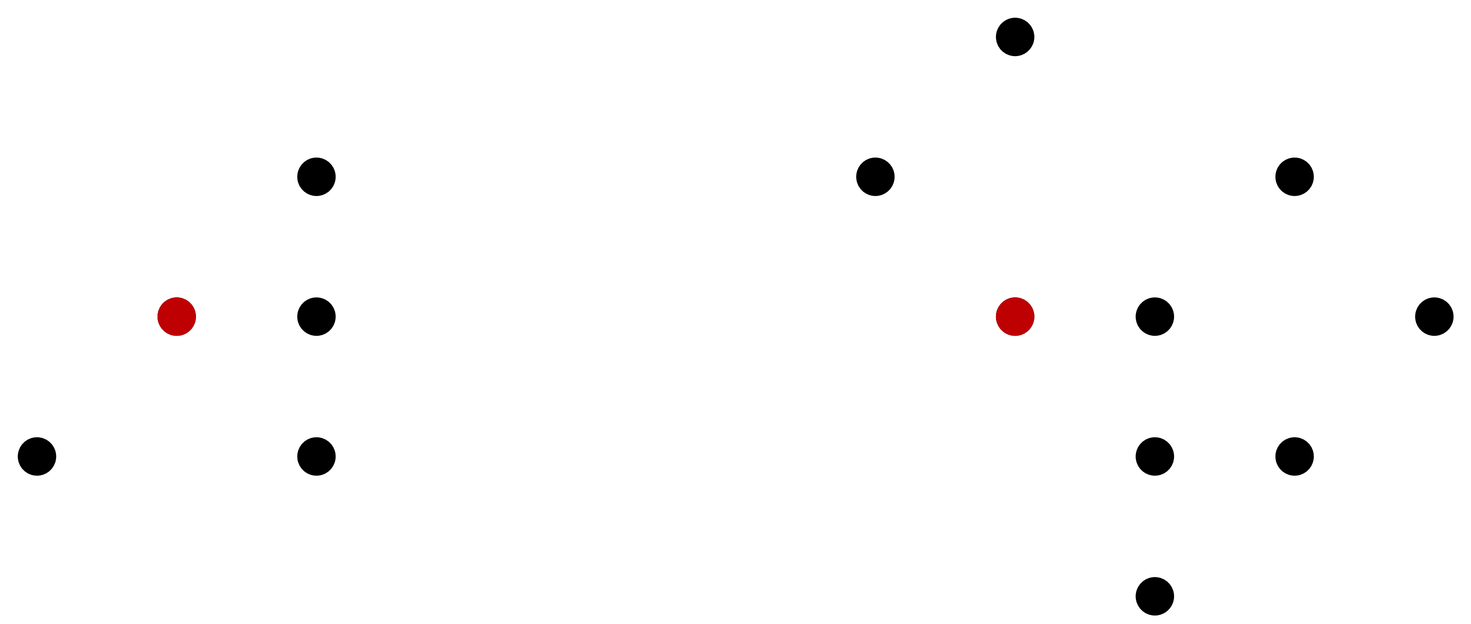
LocalSearchKMedian(P, k)

$C \leftarrow$ GreedyKCenter(P, k)
while $\exists p \in P \setminus C, c \in C$ s.t. replacing c by p in C reduces $\sum_{p \in P} d(p, C)$ by factor $1 - \tau$:
    $C \leftarrow C \cup \{p\} \setminus \{c\}$
return C



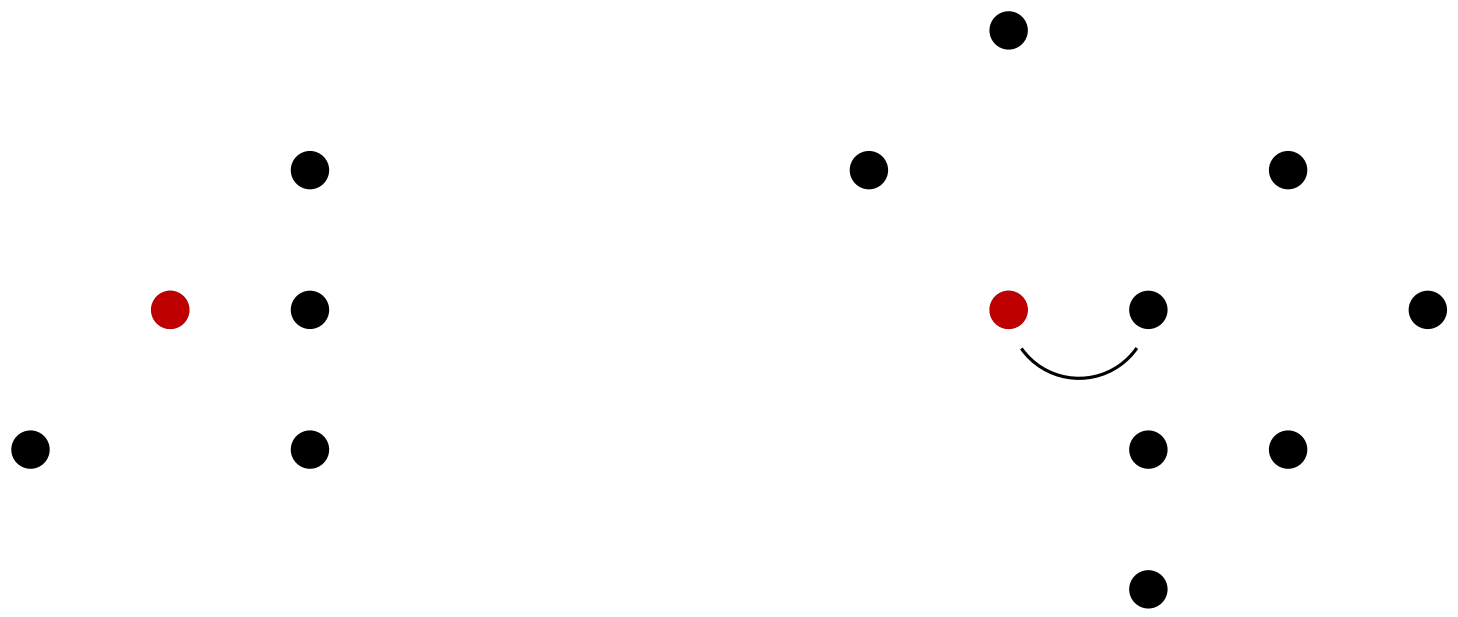
LocalSearchKMedian(P, k)

$C \leftarrow$ GreedyKCenter(P, k)
while $\exists p \in P \setminus C, c \in C$ s.t. replacing c by p in C reduces $\sum_{p \in P} d(p, C)$ by factor $1 - \tau$:
    $C \leftarrow C \cup \{p\} \setminus \{c\}$
return C



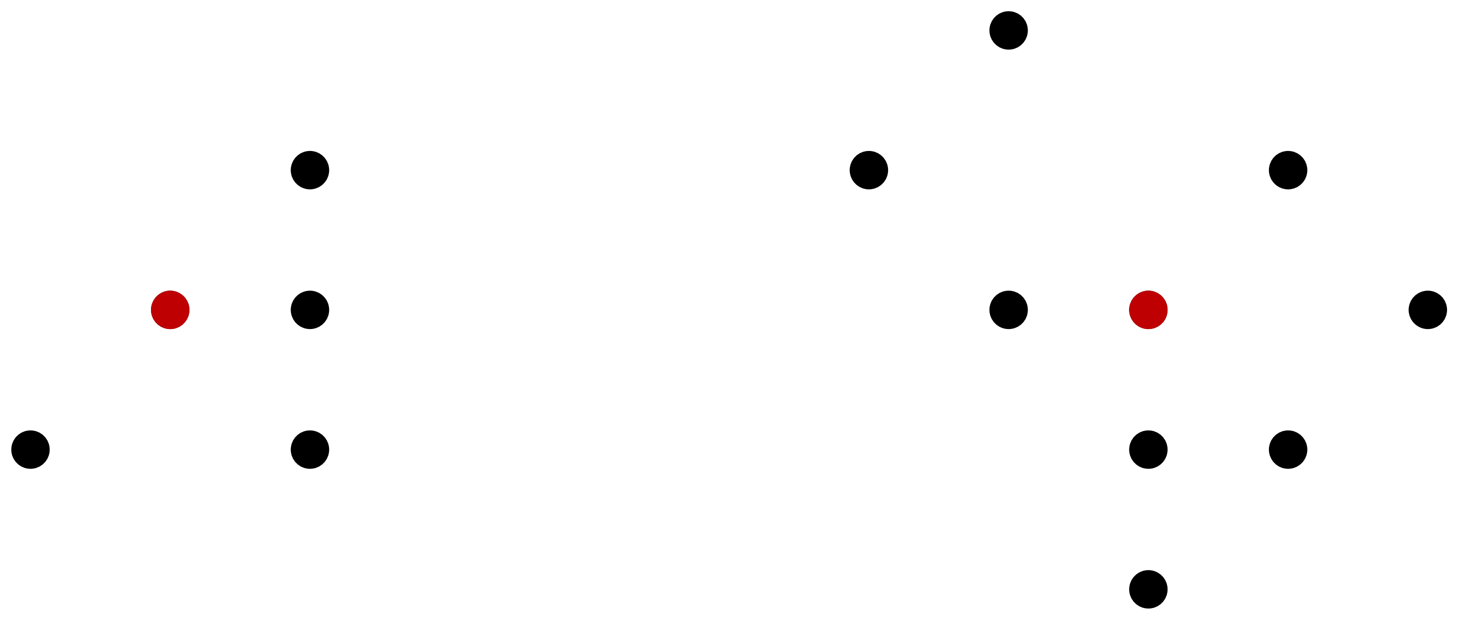
LocalSearchKMedian(P, k)

$C \leftarrow$ GreedyKCenter(P, k)
while $\exists p \in P \setminus C, c \in C$ s.t. replacing c by p in C reduces $\sum_{p \in P} d(p, C)$ by factor $1 - \tau$:
    $C \leftarrow C \cup \{p\} \setminus \{c\}$
return C



LocalSearchKMedian(P, k)

```
C ← GreedyKCenter(P, k)
while ∃p ∈ P \ C, c ∈ C s.t. replacing c by p in C reduces ∑_{p∈P} d(p, C)
      by factor 1 − τ:
    C ← C ∪ {p} \ {c}
return C
```



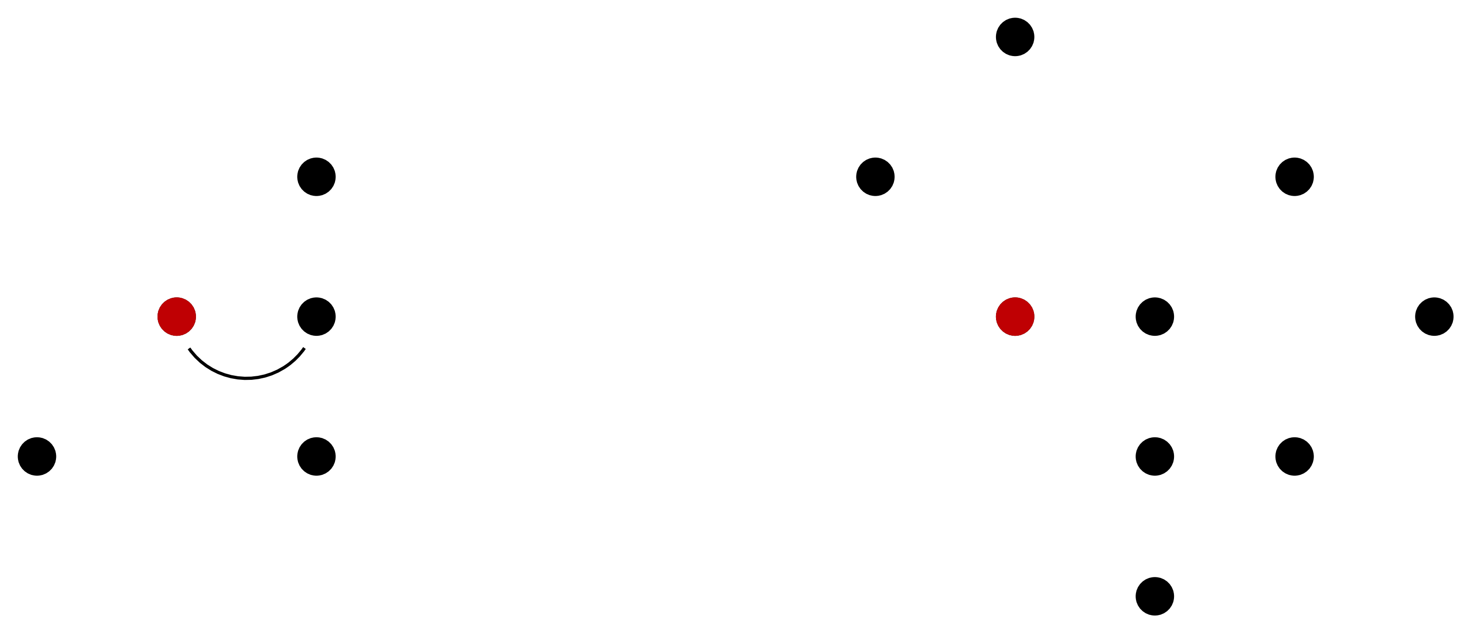
LocalSearchKMedian(P, k)

$C \leftarrow$ GreedyKCenter(P, k)

while $\exists p \in P \setminus C, c \in C$ s.t. replacing c by p in C reduces $\sum_{p \in P} d(p, C)$ by factor $1 - \tau$:

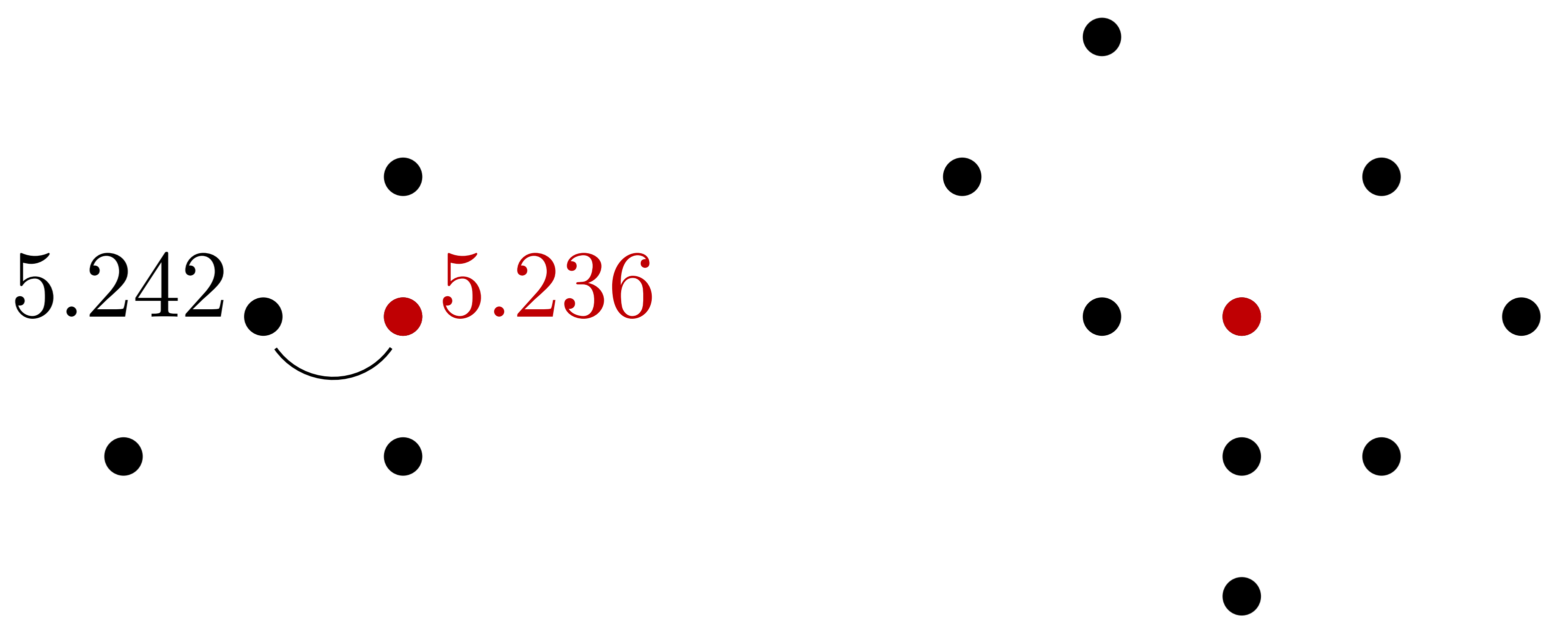
    $C \leftarrow C \cup \{p\} \setminus \{c\}$

return C



LocalSearchKMedian(P, k)

$C \leftarrow$ GreedyKCenter(P, k)
while $\exists p \in P \setminus C, c \in C$ s.t. replacing c by p in C reduces $\sum_{p \in P} d(p, C)$ by factor $1 - \tau$:
    $C \leftarrow C \cup \{p\} \setminus \{c\}$
return C



Running time

Try swapping every $p \in P \setminus c$ with every $c \in C$:

Running time

Try swapping every $p \in P \setminus c$ with every $c \in C$:

$O(nk)$ possibile swaps

Running time

Try swapping every $p \in P \setminus c$ with every $c \in C$:

- $O(nk)$ possibile swaps
- computing $\sum_{p \in P} d(p, C \cup \{p\} \setminus \{c\}$: $O(nk)$ time

Running time

Try swapping every $p \in P \setminus c$ with every $c \in C$:

$O(nk)$ possibile swaps

computing $\sum_{p \in P} d(p, C \cup \{p\} \setminus \{c\}$: $O(nk)$ time

time per iteration of **while**-loop: $O((nk)^2)$

Running time

Try swapping every $p \in P \setminus c$ with every $c \in C$:

$O(nk)$ possibile swaps

computing $\sum_{p \in P} d(p, C \cup \{p\} \setminus \{c\}$: $O(nk)$ time

time per iteration of **while**-loop: $O((nk)^2)$

number of iterations: $\log_{1/(1-\tau)} \frac{\text{initialCost}}{\text{optimalCost}}$

Running time

Try swapping every $p \in P \setminus c$ with every $c \in C$:

$O(nk)$ possibile swaps

computing $\sum_{p \in P} d(p, C \cup \{p\} \setminus \{c\}$: $O(nk)$ time

time per iteration of **while**-loop: $O((nk)^2)$

number of iterations: $\log_{1/(1-\tau)} \frac{\text{initialCost}}{\text{optimalCost}} \leq \log_{1/(1-\tau)} 2n$ (from $2n$-approx.)

Running time

Try swapping every $p \in P \setminus c$ with every $c \in C$:

$O(nk)$ possibile swaps

computing $\sum_{p \in P} d(p, C \cup \{p\} \setminus \{c\}$: $O(nk)$ time

time per iteration of **while**-loop: $O((nk)^2)$

number of iterations: $\log_{1/(1-\tau)} \frac{\text{initialCost}}{\text{optimalCost}} \leq \log_{1/(1-\tau)} 2n$ (from $2n$-approx.)

Can be simplified to $O(\frac{\log n}{\tau})$ [without proof but elementary maths]

Approximation factor

LocalSearchKMedian(P, k): $(5 + \varepsilon)$ - approximation for discrete k-median

Approximation factor

LocalSearchKMedian(P, k): $(5 + \varepsilon)$ - approximation for discrete k-median

Warning: proof tedious (but fun (?) and insightful)

Approximation factor

LocalSearchKMedian(P, k): $(5 + \varepsilon)$ - approximation for discrete k-median

Warning: proof tedious (but fun (?) and insightful)

I will sketch the core ideas

Approximation factor

LocalSearchKMedian(P, k): $(5 + \varepsilon)$ - approximation for discrete k-median

Warning: proof tedious (but fun (?) and insightful)

I will sketch the core ideas

I will show: if we replace until no improvement (aka: ignore τ),
we get 5-approximation

Approximation factor

LocalSearchKMedian(P, k): $(5+\varepsilon)$ - approximation for discrete k-median

Notation:

C: computed centers, C^* opt. centers

$A_p := d(p, C)$, $O_p := d(p, C^*)$

$\gamma(p) =$ center of $p \in C$, $\gamma^*(p)$ same in C^*

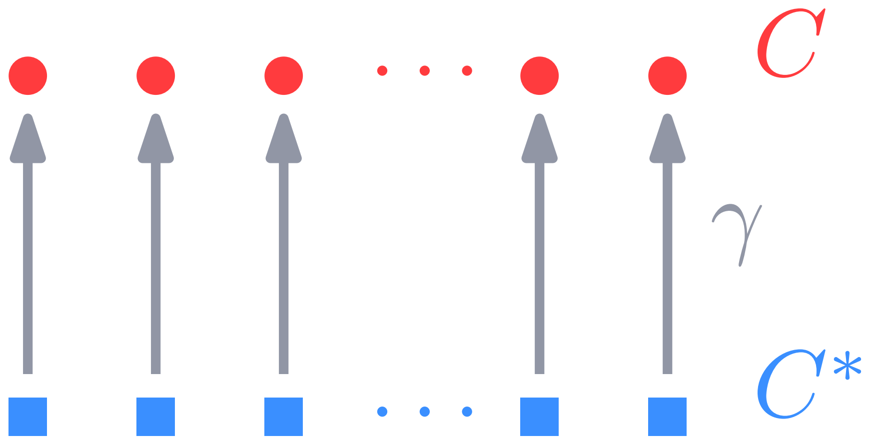
$N(c)$: cluster of $c \in C$, $N^*(c^*)$ likewise

Approximation factor

LocalSearchKMedian(P, k): $(5+\varepsilon)$ - approximation for discrete k-median

simple case: for all $o, o' \in C^*$:

$o \neq o' \rightarrow \gamma(o) \neq \gamma(o')$



Notation:

C: computed centers, C^* opt. centers

$A_p := d(p, C)$, $O_p := d(p, C^*)$

$\gamma(p) =$ center of $p \in C$, $\gamma^*(p)$ same in C^*

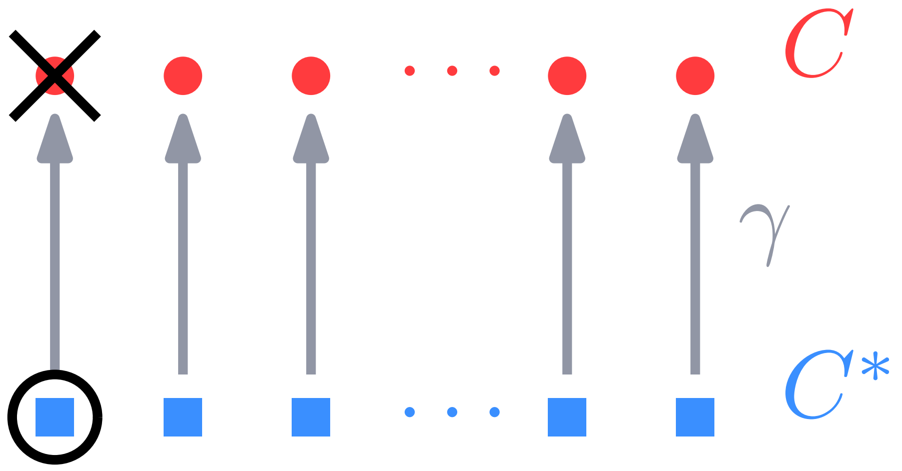
$N(c)$: cluster of $c \in C$, $N^*(c^*)$ likewise

Approximation factor

LocalSearchKMedian(P, k): $(5+\varepsilon)$ - approximation for discrete k-median

simple case: for all $o, o' \in C^*$:

$o \neq o' \rightarrow \gamma(o) \neq \gamma(o')$



Notation:

C: computed centers, C^* opt. centers

$A_p := d(p, C)$, $O_p := d(p, C^*)$

$\gamma(p) =$ center of $p \in C$, $\gamma^*(p)$ same in C^*

$N(c)$: cluster of $c \in C$, $N^*(c^*)$ likewise

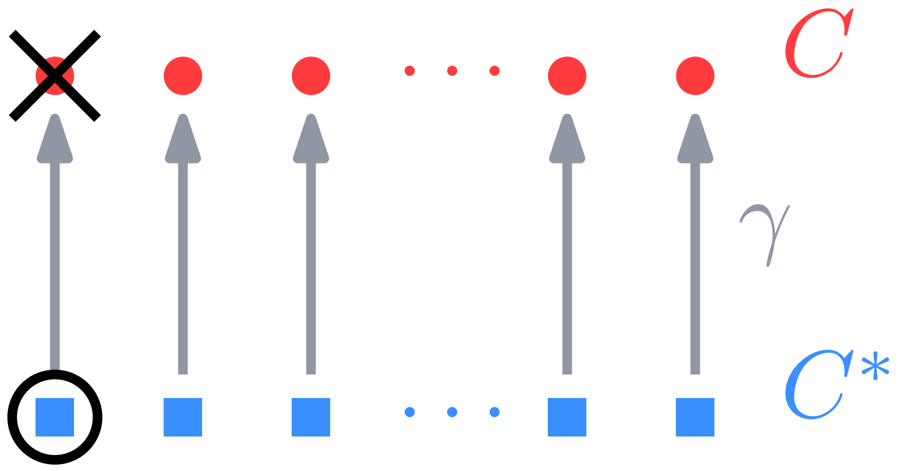
Idea: for $o \in C^*$ consider $C' := C + o - \gamma(o)$

Approximation factor

LocalSearchKMedian(P, k): $(5+\varepsilon)$ - approximation for discrete k-median

simple case: for all $o, o' \in C^*$:

$o \neq o' \rightarrow \gamma(o) \neq \gamma(o')$



Notation:
C: computed centers, C^* opt. centers
$A_p := d(p, C)$, $O_p := d(p, C^*)$
$\gamma(p) =$ center of $p \in C$, $\gamma^*(p)$ same in C^*
$N(c)$: cluster of $c \in C$, $N^*(c^*)$ likewise

Idea: for $o \in C^*$ consider $C' := C + o - \gamma(o)$

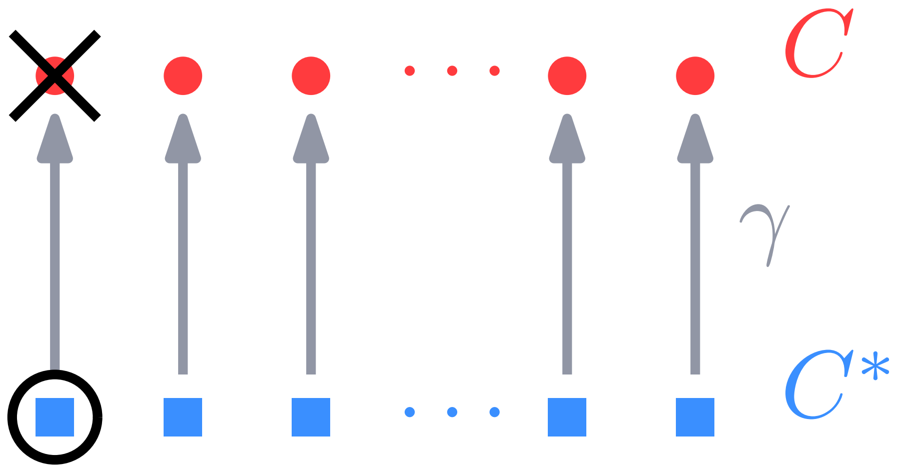
$$0 \leq cost(C + o - \gamma(o)) - cost(C)$$

Approximation factor

LocalSearchKMedian(P, k): $(5+\varepsilon)$ - approximation for discrete k-median

simple case: for all $o, o' \in C^*$:

$o \neq o' \rightarrow \gamma(o) \neq \gamma(o')$



Notation:

C: computed centers, C^* opt. centers

$A_p := d(p, C)$, $O_p := d(p, C^*)$

$\gamma(p)$ = center of $p \in C$, $\gamma^*(p)$ same in C^*

$N(c)$: cluster of $c \in C$, $N^*(c^*)$ likewise

Idea: for $o \in C^*$ consider $C' := C + o - \gamma(o)$

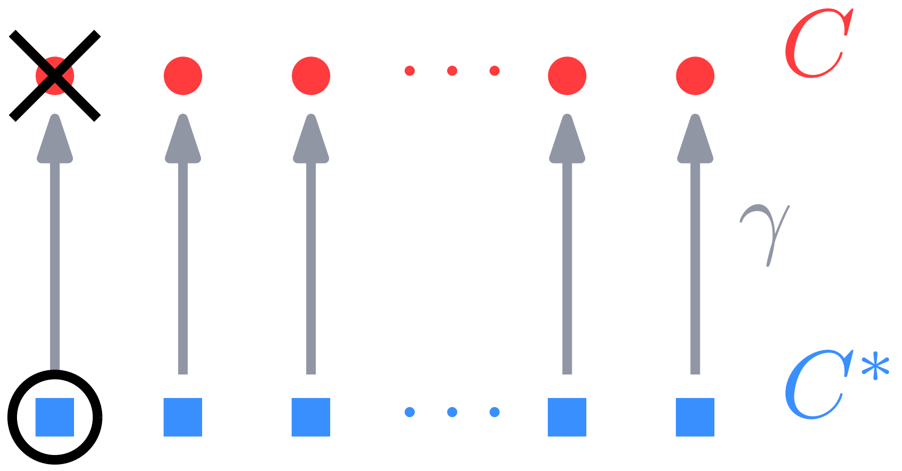
$$0 \leq cost(C + o - \gamma(o)) - cost(C)$$
$$\leq \sum_{p \in N^*(o)} (O_p - A_p) + \sum_{q \in N(\gamma(o))} (d(q, \gamma(\gamma^*(q))) - A_q)$$

Approximation factor

LocalSearchKMedian(P, k): $(5+\varepsilon)$ - approximation for discrete k-median

simple case: for all $o, o' \in C^*$:

$o \neq o' \rightarrow \gamma(o) \neq \gamma(o')$



Notation:

C: computed centers, C^* opt. centers

$A_p := d(p, C)$, $O_p := d(p, C^*)$

$\gamma(p) =$ center of $p \in C$, $\gamma^*(p)$ same in C^*

$N(c)$: cluster of $c \in C$, $N^*(c^*)$ likewise

Idea: for $o \in C^*$ consider $C' := C + o - \gamma(o)$

$$0 \leq cost(C + o - \gamma(o)) - cost(C)$$
$$\leq \sum_{p \in N^*(o)} (O_p - A_p) + \sum_{q \in N(\gamma(o))} (d(q, \gamma(\gamma^*(q))) - A_q)$$

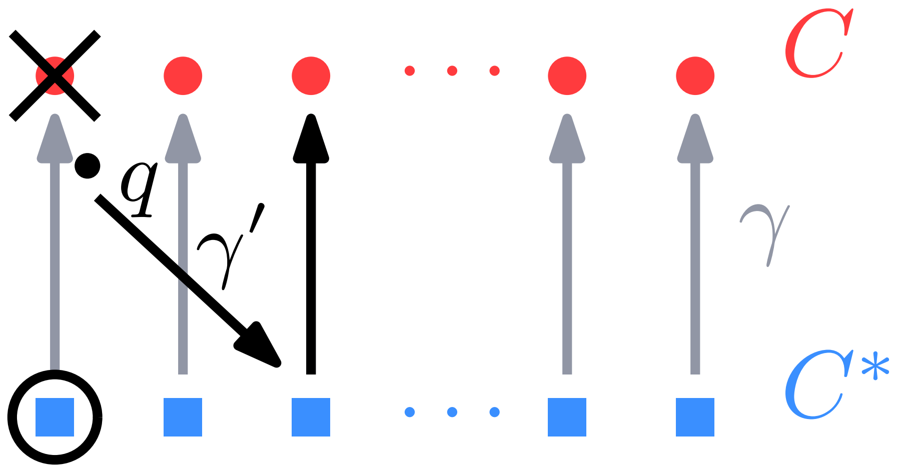
$d(p, C') \leq$
$d(p, o) = O_p$

Approximation factor

LocalSearchKMedian(P, k): $(5+\varepsilon)$ - approximation for discrete k-median

simple case: for all $o, o' \in C^*$:

$o \neq o' \rightarrow \gamma(o) \neq \gamma(o')$



Notation:
C: computed centers, C^* opt. centers
$A_p := d(p, C)$, $O_p := d(p, C^*)$
$\gamma(p) =$ center of $p \in C$, $\gamma^*(p)$ same in C^*
$N(c)$: cluster of $c \in C$, $N^*(c^*)$ likewise

Idea: for $o \in C^*$ consider $C' := C + o - \gamma(o)$

$$0 \leq cost(C + o - \gamma(o)) - cost(C)$$
$$\leq \sum_{p \in N^*(o)} (O_p - A_p) + \sum_{q \in N(\gamma(o))} (d(q, \gamma(\gamma^*(q))) - A_q)$$

$d(p, C') \leq$
$d(p, o) = O_p$

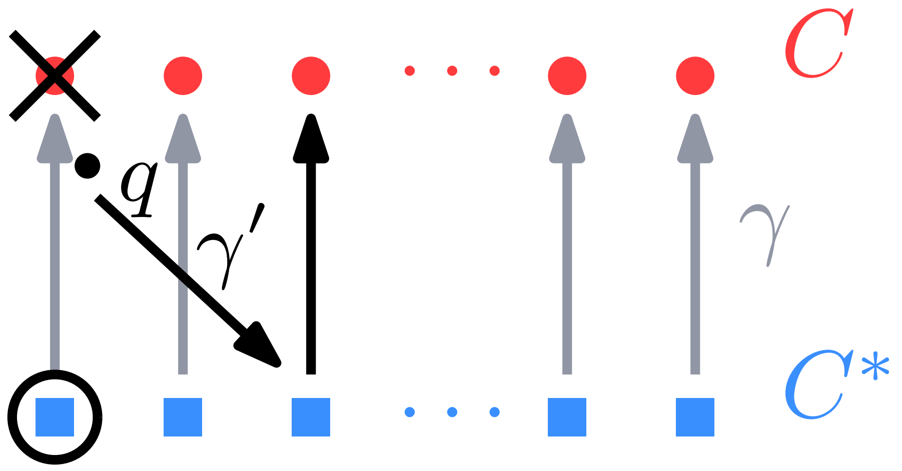
bound cost for $q \in N(\gamma(o)) \setminus N^*(o)$
by taking $d(q, \gamma(\gamma^*(q)))$

Approximation factor

LocalSearchKMedian(P, k): $(5+\varepsilon)$ - approximation for discrete k-median

simple case: for all $o, o' \in C^*$:

$o \neq o' \rightarrow \gamma(o) \neq \gamma(o')$



Notation:

C: computed centers, C^* opt. centers

$A_p := d(p, C)$, $O_p := d(p, C^*)$

$\gamma(p) =$ center of $p \in C$, $\gamma^*(p)$ same in C^*

$N(c)$: cluster of $c \in C$, $N^*(c^*)$ likewise

Idea: for $o \in C^*$ consider $C' := C + o - \gamma(o)$

$$0 \leq cost(C + o - \gamma(o)) - cost(C)$$
$$\leq \sum_{p \in N^*(o)} (O_p - A_p) + \sum_{q \in N(\gamma(o))} (d(q, \gamma(\gamma^*(q))) - A_q)$$

by triangle ineq. (proof later): $\leq \sum_{q \in N(\gamma(o))} 2O_q$

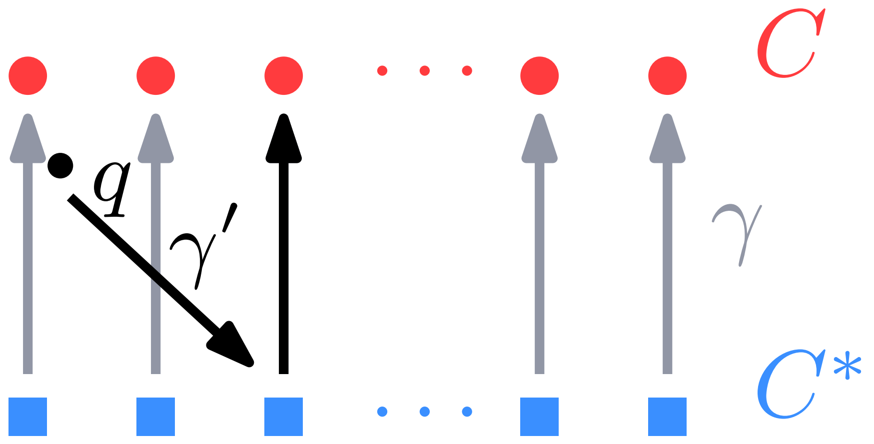
By doing this for all $o \in C^*$ and summing: $\sum A_p \leq 3 \sum O_p$

Approximation factor

LocalSearchKMedian(P, k): $(5+\varepsilon)$ - approximation for discrete k-median

simple case: for all $o, o' \in C^*$:

$o \neq o' \rightarrow \gamma(o) \neq \gamma(o')$



proof of $d(q, \gamma(\gamma^*(q))) - A_q \leq 2O_q$:

Notation:

C: computed centers, C^* opt. centers

$A_p := d(p, C)$, $O_p := d(p, C^*)$

$\gamma(p) =$ center of $p \in C$, $\gamma^*(p)$ same in C^*

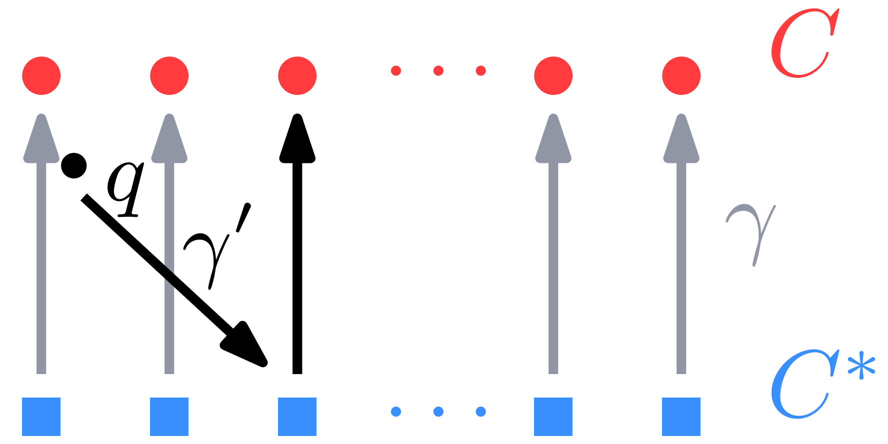
$N(c)$: cluster of $c \in C$, $N^*(c^*)$ likewise

Approximation factor

LocalSearchKMedian(P, k): $(5+\varepsilon)$ - approximation for discrete k-median

simple case: for all $o, o' \in C^*$:

$o \neq o' \rightarrow \gamma(o) \neq \gamma(o')$



Notation:

C: computed centers, C^* opt. centers

$A_p := d(p, C)$, $O_p := d(p, C^*)$

$\gamma(p) =$ center of $p \in C$, $\gamma^*(p)$ same in C^*

$N(c)$: cluster of $c \in C$, $N^*(c^*)$ likewise

proof of $d(q, \gamma(\gamma^*(q))) - A_q \leq 2O_q$:

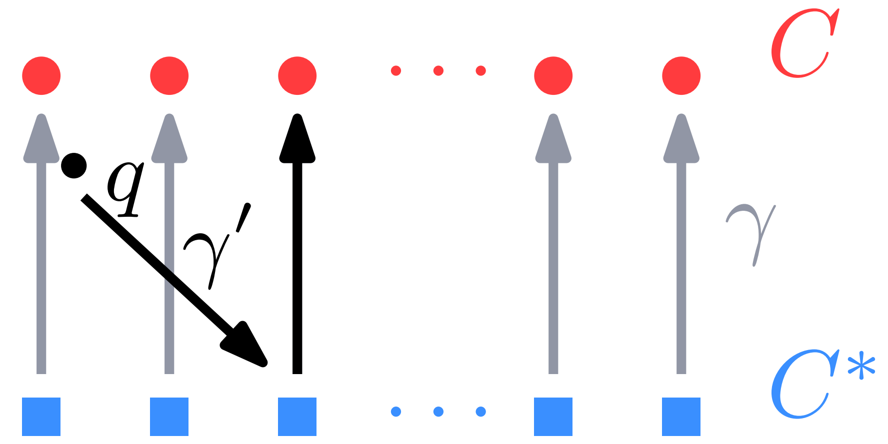
$$d(q, \gamma(\gamma^*(q))) \leq d(q, \gamma^*(q)) + d(\gamma^*(q), \gamma(\gamma^*(q)))$$

Approximation factor

LocalSearchKMedian(P, k): $(5+\varepsilon)$ - approximation for discrete k-median

simple case: for all $o, o' \in C^*$:

$o \neq o' \rightarrow \gamma(o) \neq \gamma(o')$



Notation:

C: computed centers, C^* opt. centers

$A_p := d(p, C)$, $O_p := d(p, C^*)$

$\gamma(p) =$ center of $p \in C$, $\gamma^*(p)$ same in C^*

$N(c)$: cluster of $c \in C$, $N^*(c^*)$ likewise

proof of $d(q, \gamma(\gamma^*(q))) - A_q \leq 2O_q$:

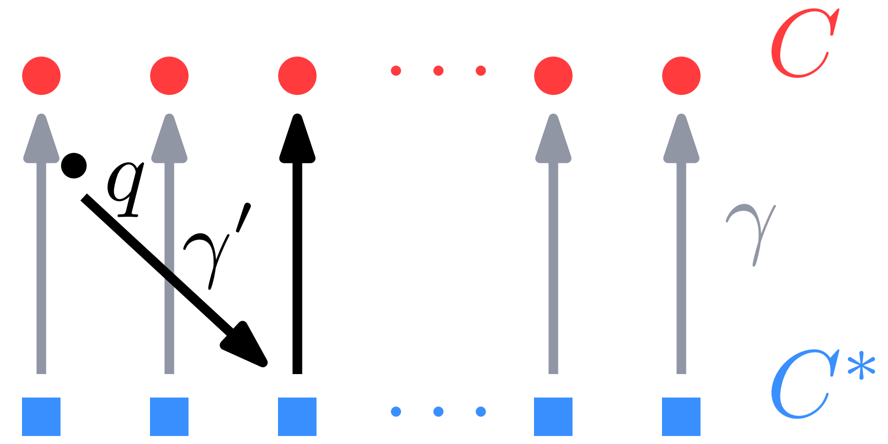
$$d(q, \gamma(\gamma^*(q))) \leq d(q, \gamma^*(q)) + d(\gamma^*(q), \gamma(\gamma^*(q)))$$
$$\leq O_q + d(\gamma^*(q), \gamma(\gamma^*(q)))$$

Approximation factor

LocalSearchKMedian(P, k): $(5+\varepsilon)$ - approximation for discrete k-median

simple case: for all $o, o' \in C^*$:

$o \neq o' \rightarrow \gamma(o) \neq \gamma(o')$



Notation:

C: computed centers, C^* opt. centers

$A_p := d(p, C)$, $O_p := d(p, C^*)$

$\gamma(p) =$ center of $p \in C$, $\gamma^*(p)$ same in C^*

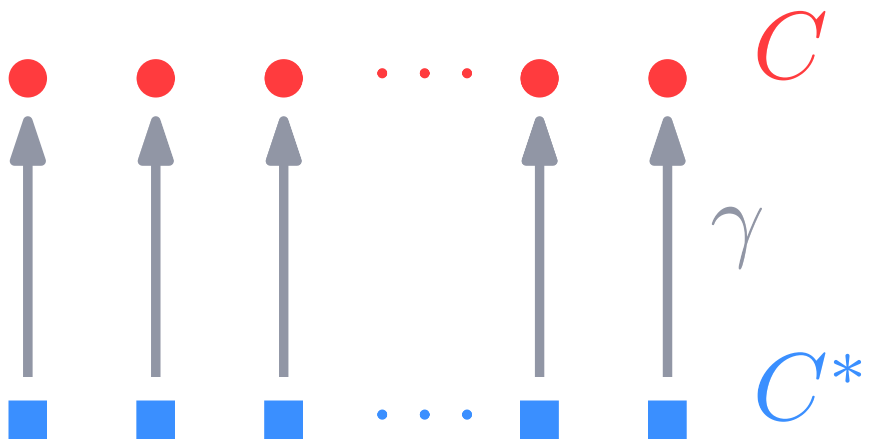
$N(c)$: cluster of $c \in C$, $N^*(c^*)$ likewise

proof of $d(q, \gamma(\gamma^*(q))) - A_q \leq 2O_q$:

$$\begin{aligned} d(q, \gamma(\gamma^*(q))) &\leq d(q, \gamma^*(q)) + d(\gamma^*(q), \gamma(\gamma^*(q))) \\ &\leq O_q + d(\gamma^*(q), \gamma(\gamma^*(q))) \\ &\leq O_q + d(\gamma^*(q), \gamma(q)) \\ &\leq O_q + d(\gamma^*(q), q) + d(q, \gamma(q)) \\ &= O_q + O_q + A_q \end{aligned}$$

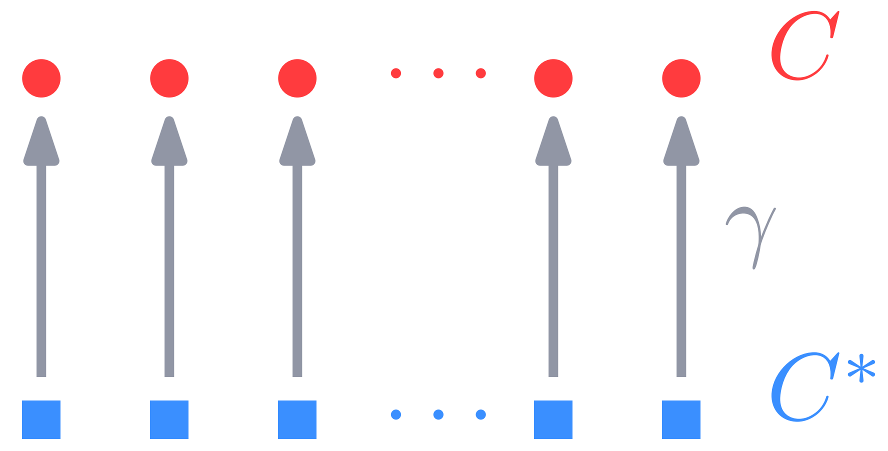
Approximation factor (general case)

so far

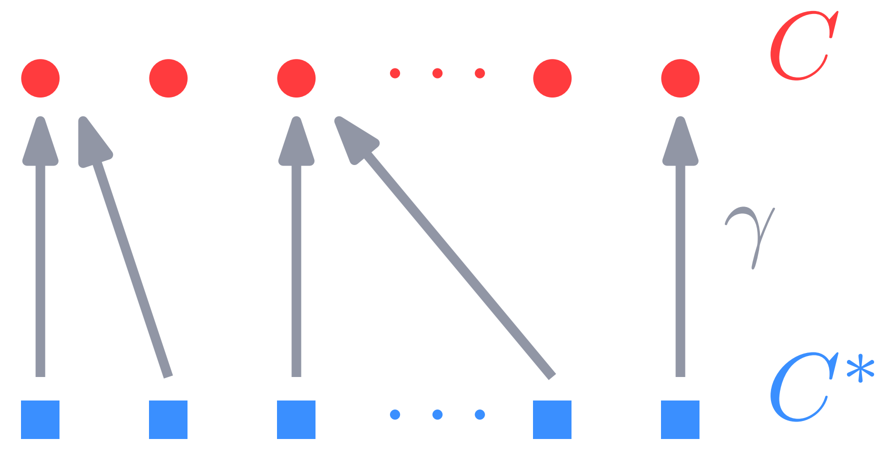


Approximation factor (general case)

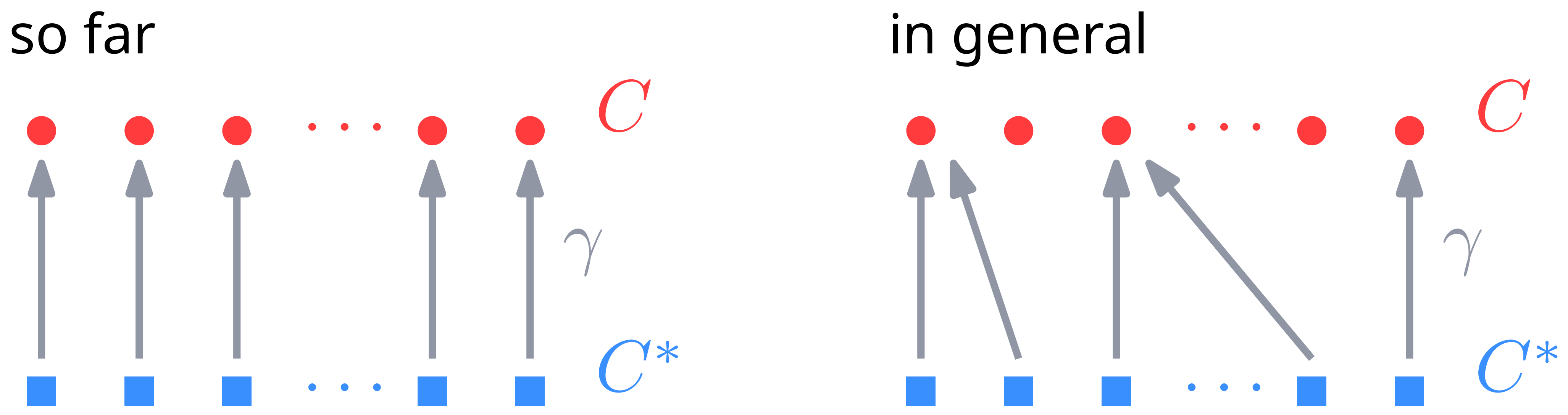
so far



in general

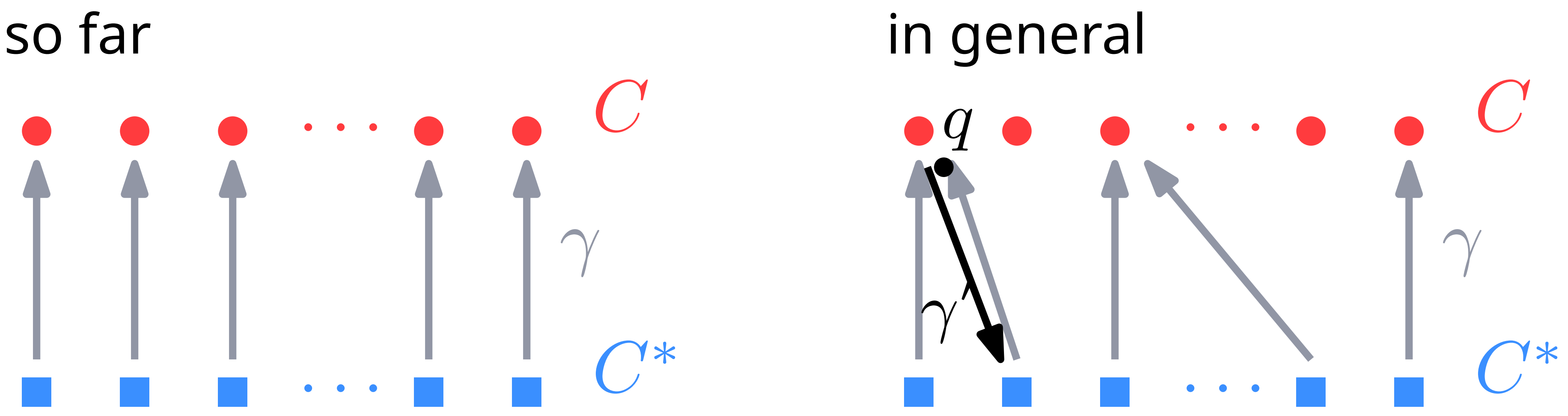


Approximation factor (general case)



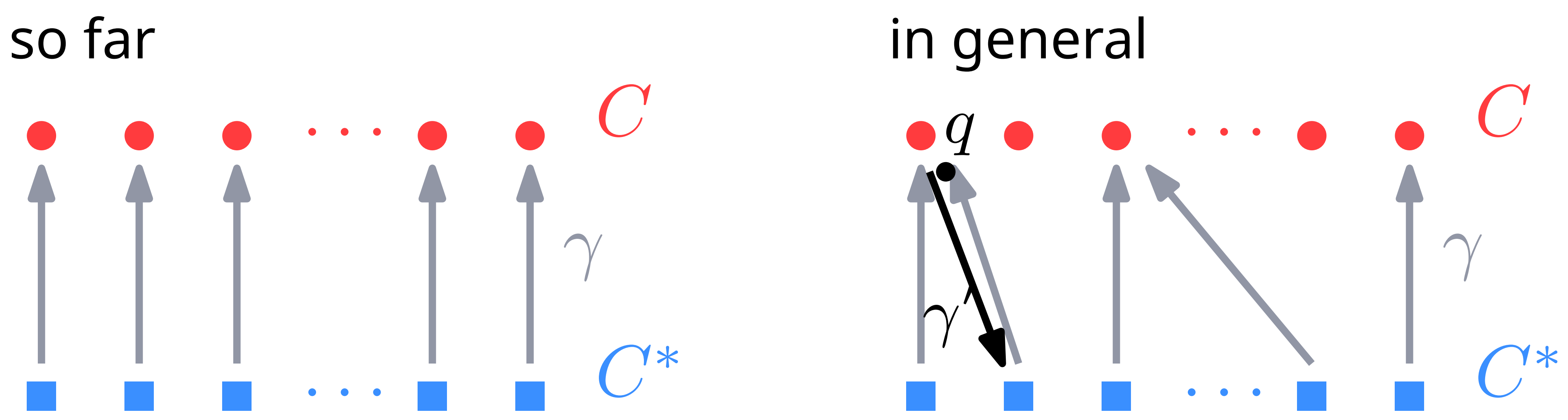
problem: if we swap o with $c := \gamma(o) = \gamma(o')$, we can't reassign $q \in N(c) \cap N^*(o')$

Approximation factor (general case)



problem: if we swap o with $c := \gamma(o) = \gamma(o')$, we can't reassign $q \in N(c) \cap N^*(o')$

Approximation factor (general case)

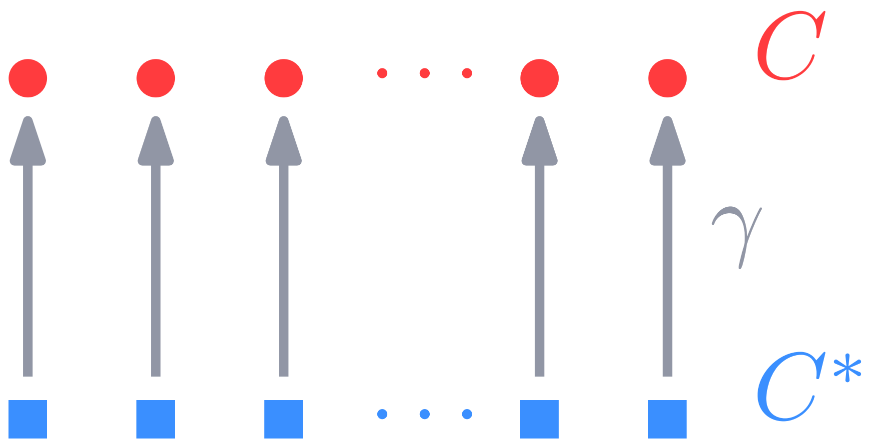


problem: if we swap o with $c := \gamma(o) = \gamma(o')$, we can't reassign $q \in N(c) \cap N^*(o')$

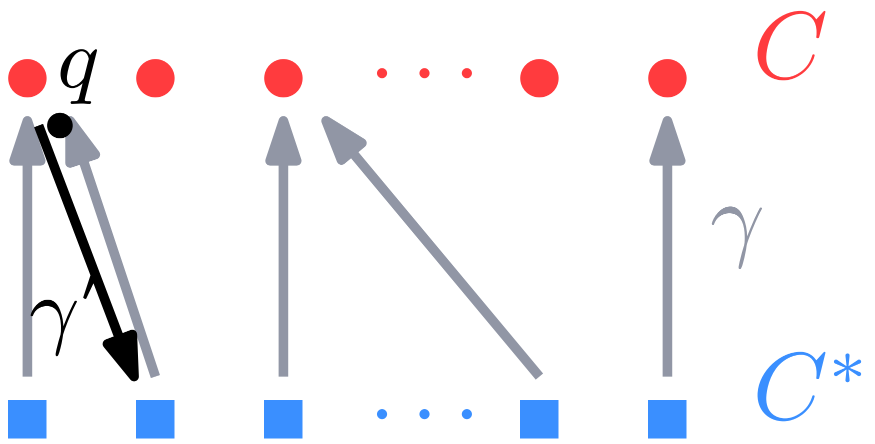
solution: swap $o \in C^*$ with $\eta(o)$ chosen s.t.

Approximation factor (general case)

so far



in general

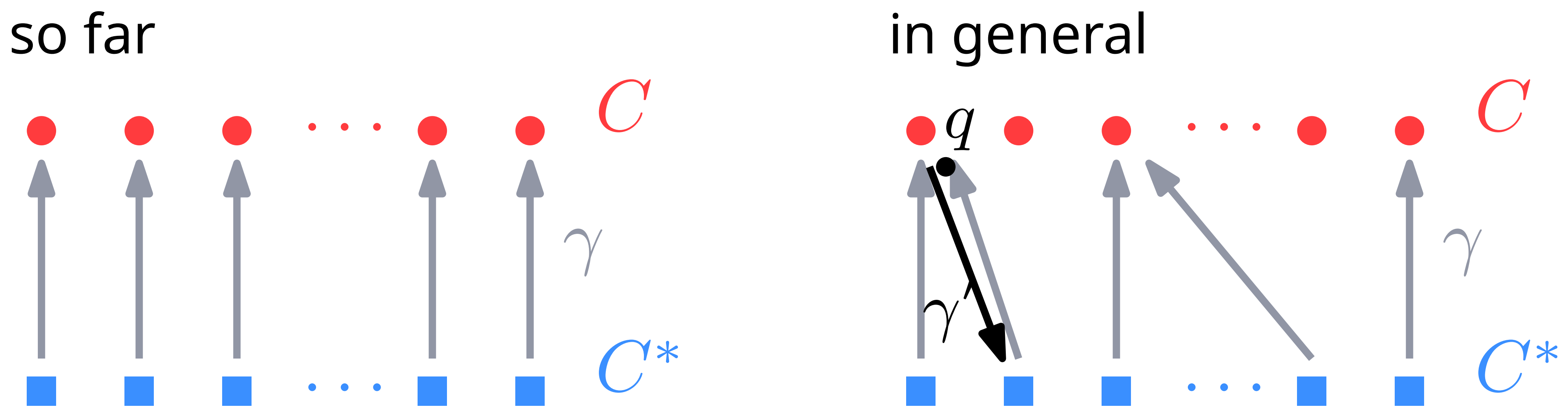


problem: if we swap o with $c := \gamma(o) = \gamma(o')$, we can't reassign $q \in N(c) \cap N^*(o')$

solution: swap $o \in C^*$ with $\eta(o)$ chosen s.t.

$$\eta(o) = \gamma(o) \text{ if } \gamma(o) \neq \gamma(o') \text{ for } o \neq o' \in C^*$$

Approximation factor (general case)



problem: if we swap o with $c := \gamma(o) = \gamma(o')$, we can't reassign $q \in N(c) \cap N^*(o')$

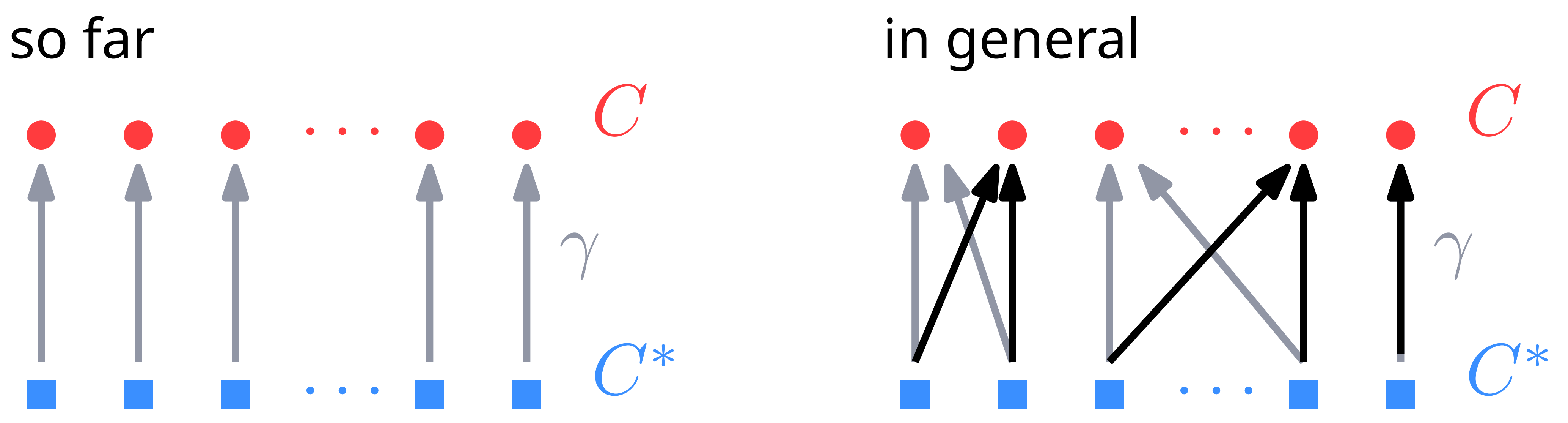
solution: swap $o \in C^*$ with $\eta(o)$ chosen s.t.

$\eta(o) = \gamma(o)$ if $\gamma(o) \neq \gamma(o')$ for $o \neq o' \in C^*$

$\eta(o) \neq \gamma(o')$ for all $o' \in C^*$ and

$\eta(o) = \eta(o')$ for at most one other o'

Approximation factor (general case)



problem: if we swap o with $c := \gamma(o) = \gamma(o')$, we can't reassign $q \in N(c) \cap N^*(o')$

solution: swap $o \in C^*$ with $\eta(o)$ chosen s.t.

$\eta(o) = \gamma(o)$ if $\gamma(o) \neq \gamma(o')$ for $o \neq o' \in C^*$

$\eta(o) \neq \gamma(o')$ for all $o' \in C^*$ and
$\eta(o) = \eta(o')$ for at most one other o'

Same argument works, but since we swap out each $c \in C$
up to 2 times, we get $\sum A_p \leq \sum O_p + 2 \cdot 2O_p$

summary + discrete k-means + open problems

k-center: 2-approximation by greedy algorithm

discrete k-median: $(5+\varepsilon)$-approximation by local search

summary + discrete k-means + open problems

k-center: 2-approximation by greedy algorithm

discrete k-median: $(5 + \varepsilon)$-approximation by local search

discrete k-means: minimize $\sum_{p \in P} d(p, C)^2$

summary + discrete k-means + open problems

k-center: 2-approximation by greedy algorithm

discrete k-median: $(5+\varepsilon)$-approximation by local search

discrete k-means: minimize $\sum_{p \in P} d(p, C)^2$

open: α-approximation for k-center in R^d with Euclidean distance and $1.82 < \alpha < 2$?

summary + discrete k-means + open problems

k-center: 2-approximation by greedy algorithm

discrete k-median: $(5 + \varepsilon)$-approximation by local search

discrete k-means: minimize $\sum_{p \in P} d(p, C)^2$

open: α-approximation for k-center in R^d with Euclidean distance and $1.82 < \alpha < 2$?

in my research: geometric spaces beyond points, in particular, clustering curves