Coresets and their Applications

Dan Feldman

Elections Queries

- Input: $P = \left\{ \right.$  $\left. \right\}$

Elections Queries

- Input: $P = \{$  $\}$
- Query: $q \in \{$  ,  $\}$

Elections Queries

- Input: $P = \{$  $\}$
- Query: $q \in \{$  ,  $\}$
- Output: $f(P, q) =$

fraction of P that vote for q

ε-Coreset for Elections

Answer queries in sub-linear time.

ε-Coreset for Elections

Answer queries in sub-linear time.

Key Idea: Random Sampling $C \subseteq P$

ε-Coreset for Elections

Answer queries in sub-linear time.

Key Idea: Random Sampling $C \subseteq P$

Chernoff Inequality: If $|C| \geq 100/\varepsilon^2$,

$$\forall q : |f(P,q) - f(C,q)| \leq \varepsilon$$

with high probability.

Definition

C is an ε-coreset for $(P, \mathbb{Q}, f)$

if for every $q \in \mathbb{Q}$ we have

$$f(P, q) \sim f(C, q),$$

Definition

C is an ε-coreset for $(P, \mathbb{Q}, f)$

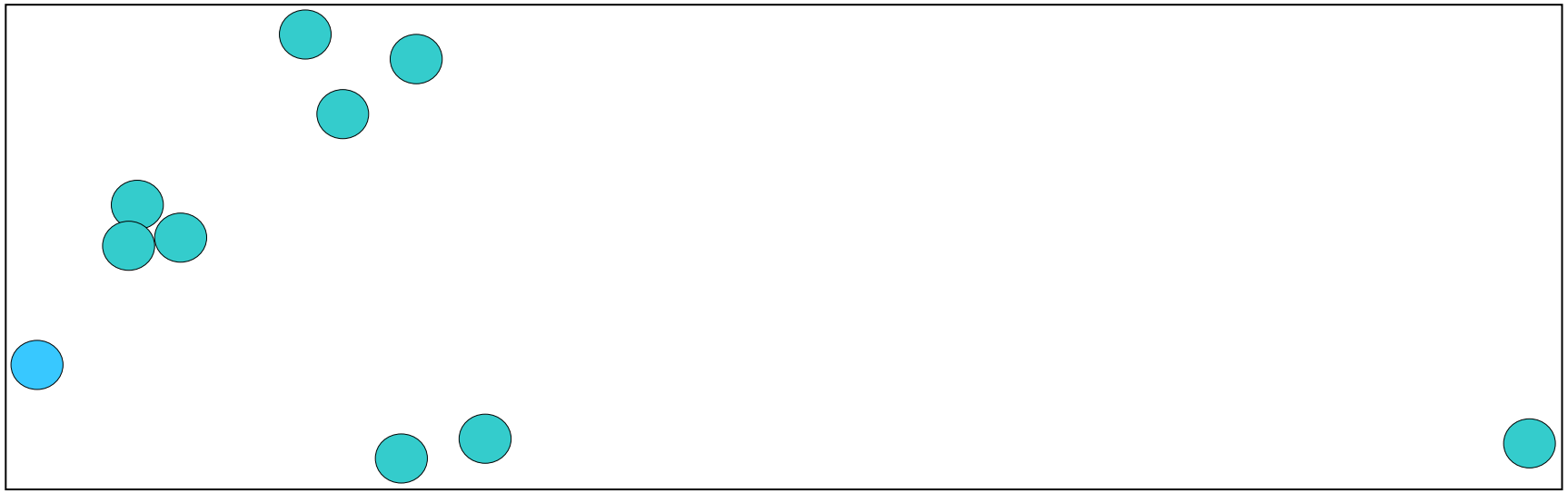
if for every $q \in \mathbb{Q}$ we have w.h.p

$$f(P, q) \sim f(C, q),$$

where the approximation depends on ε.

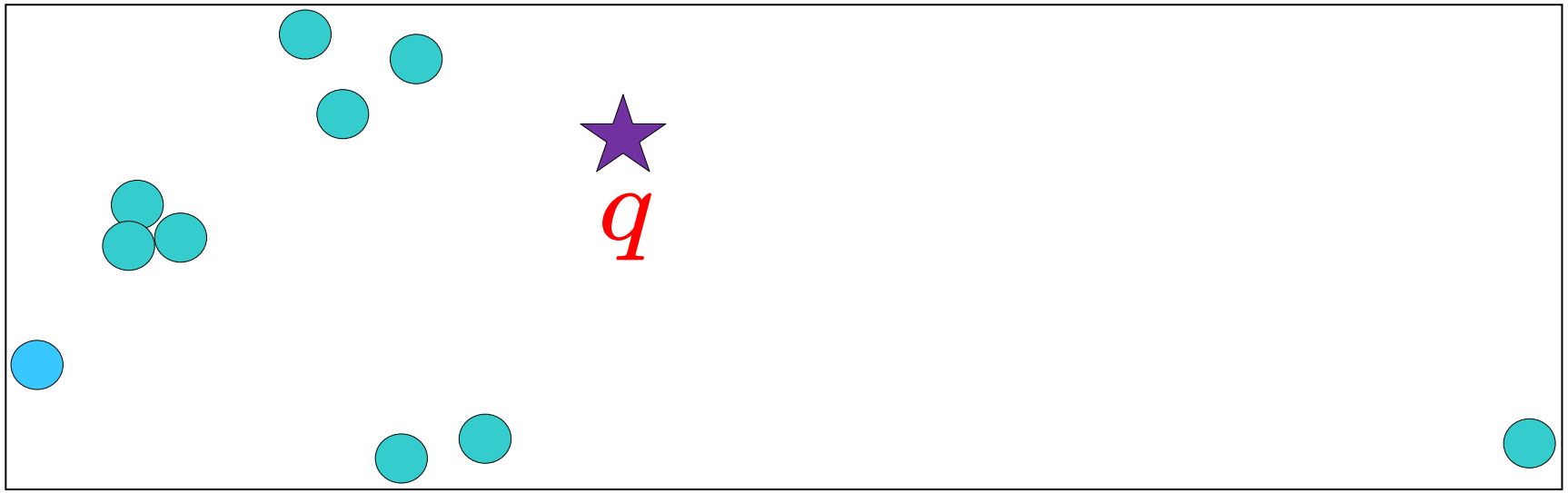
Mean Queries

- Input: P in $\mathbb{R}^d$



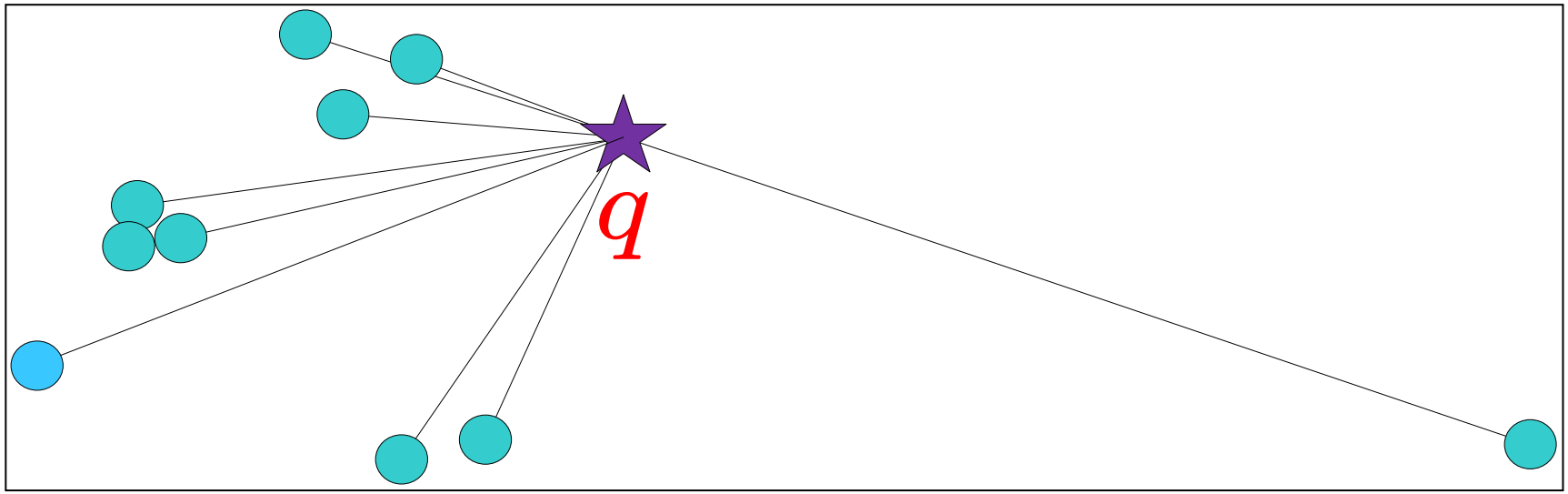
Mean Queries

- Input: P in $\mathbb{R}^d$
- Query: a point $q \in \mathbb{R}^d$



Mean Queries

- Input: P in $\mathbb{R}^d$
- Query: a point $q \in \mathbb{R}^d$
- Output: $f(P,q) = \sum_{p \in P} (\text{dist}(p,q))^2$



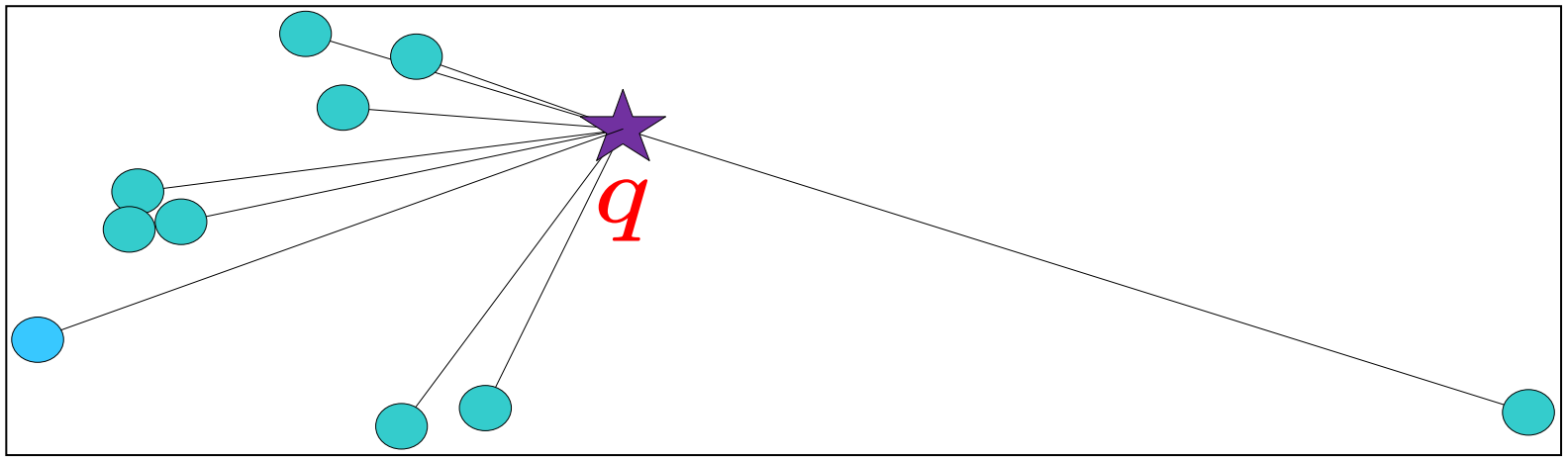
Coreset For Mean Queries

- Assume $\sum_{p \in P} = 0$.

  Oterwise, translate P.

- $\left(\mathrm{dist}(p,q)\right)^2 = \|p-q\|^2$

  $= \|p\|^2 + \|q\|^2 - 2p \cdot q$



Coreset For Mean Queries

- Assume $\sum_{p \in P} = 0$.

  Oterwise, translate P.

- $\left(\mathrm{dist}(p,q)\right)^2 = \|p-q\|^2$

$$= \|p\|^2 + \|q\|^2 - 2p \cdot q$$

- $\sum_{p \in P} \left(\mathrm{dist}(p,q)\right)^2 = \sum_{p \in P} \|p\|^2 + \sum_{p \in P} \|q\|^2 - 2q \cdot \sum_{p \in P} p$

$$= \sum_{p \in P} \|p\|^2 + |P| \, \|q\|^2$$

Coreset For Mean Queries

$$f(P,q) = \sum_{p \in P} \|p\|^2 + |P| \cdot \|q\|^2$$

$$= g(C,q)$$

where $C = \sum_{p \in P} p = 0$,

$$g(C,q) = |P| \cdot \mathrm{dist}(C,q) + \sum_{p \in P} \|p\|^2$$

Properties of Coresets

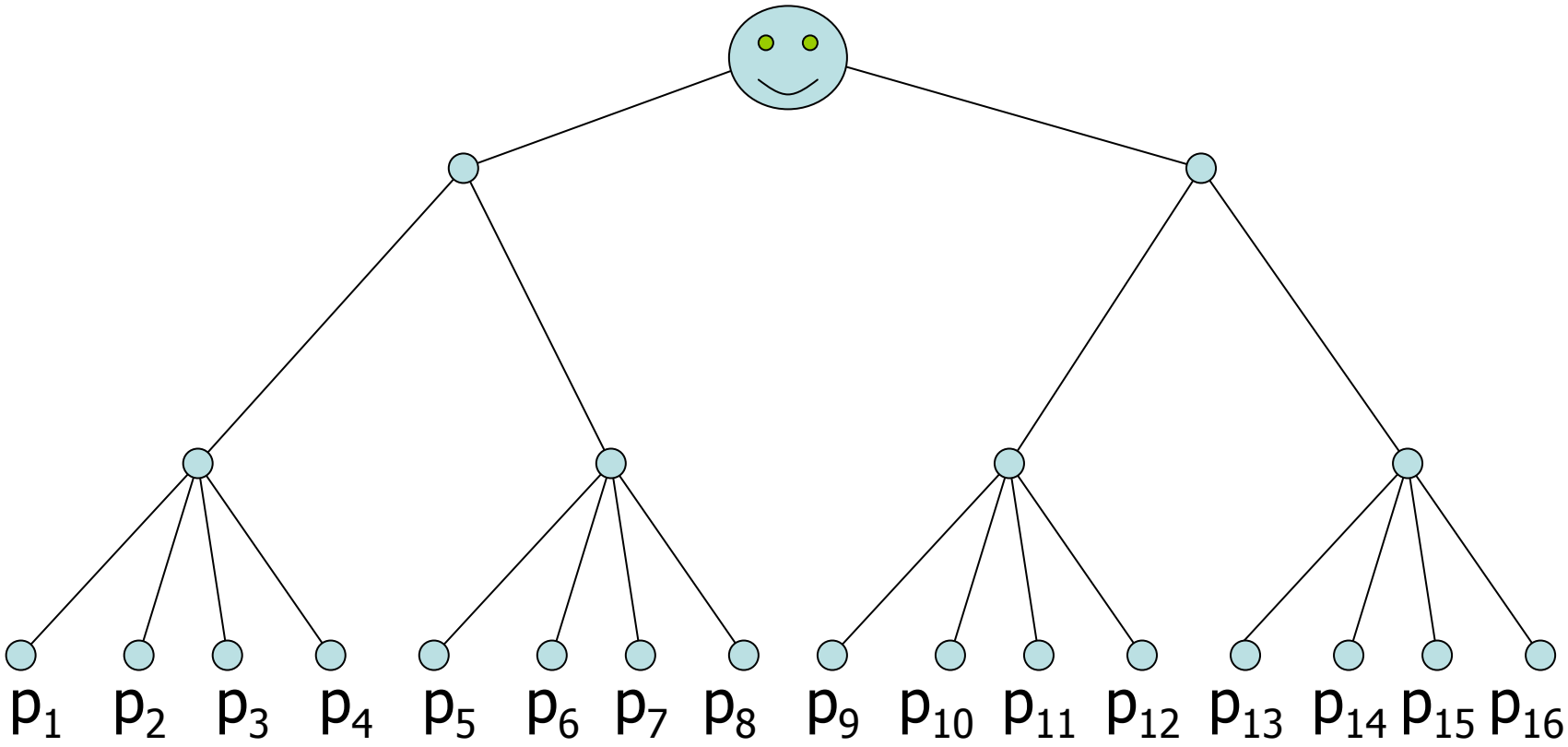
- Transitivity: A coreset of a coreset is a coreset
- Additivity: A union of coresets is a coreset

Streaming: The Model

- Single pass over the data: p_1, p_2, …,p_n
- $O(\log n)$ space
- Fast processing time per element

From a presentation by Piotr Indyk

Tree Computation



From a presentation by Piotr Indyk

Other Applications

- Distributed Computing
- GPUs
- Faster Heuristics
- Optimization with/without constraints

Other Coresets

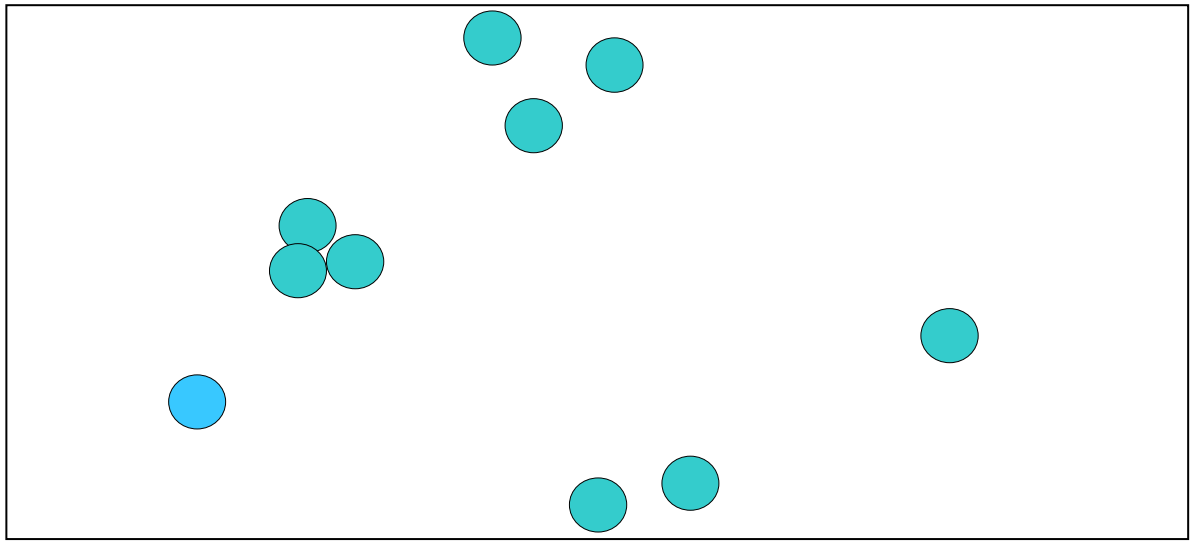
- k-Line median/means [with A. Fiat, M. Sharir]
- Private Coresets [with A. Fiat, H. Kaplan, K. Nissim]
- Regression/Matrix approximations [with C. Sohler]
- Compressed sensing [with B. Hassibi]
- Supported vector machines [with R. Condor]
- Mixture of Gaussians [with Andreas Krause]
- Dynamic coresets [with D. Golonov]
- k-Sparsest-cut [with M. Mahoney]
- A Unified Framework [with L. Schulman, M. Langberg]

Implementations

- Google's PageRank [with R. Bar-Yaets, Y. Koren]
- Image Compression [with M. Feige, N. Sochen]
- Text Mining [MyOgger Ltd.]

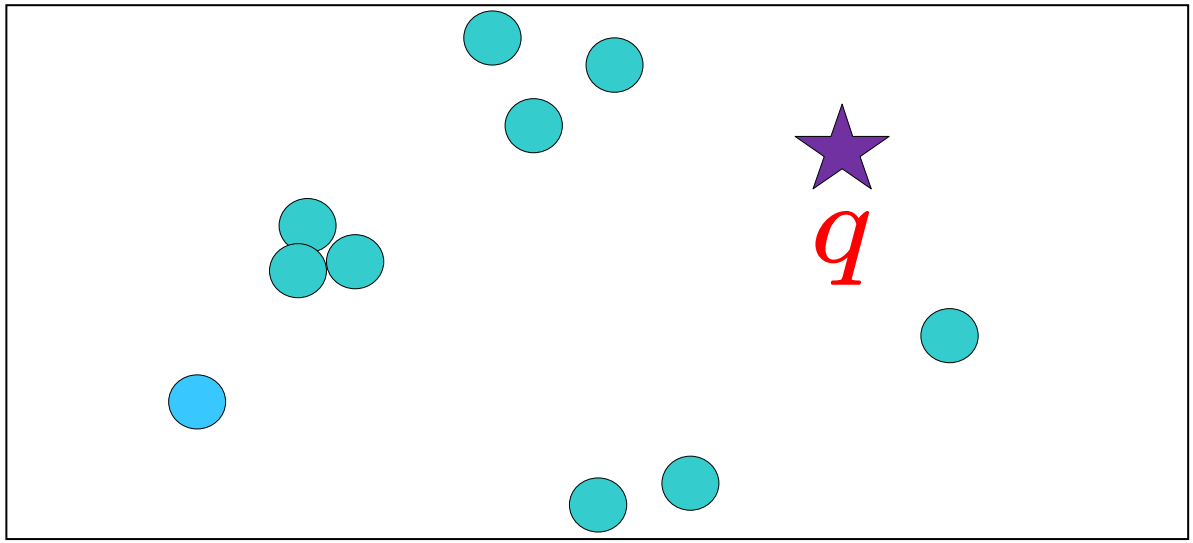
Minimum Enclosing Ball

- Input: P in $\mathbb{R}^d$



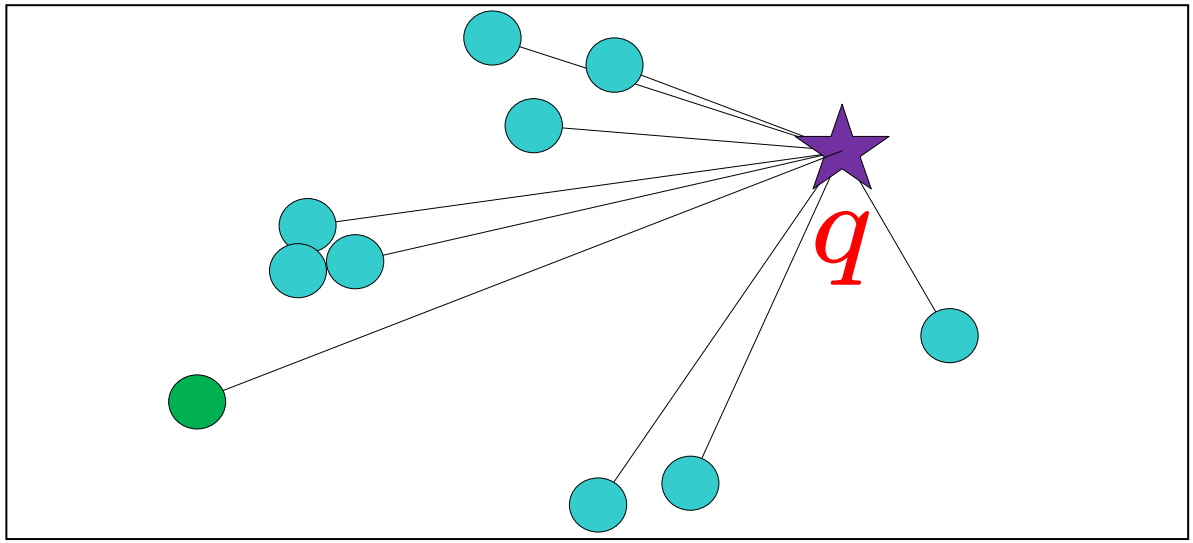
Minimum Enclosing Ball

- Input: P in $\mathbb{R}^d$
- Query: a point $q \in \mathbb{R}^d$

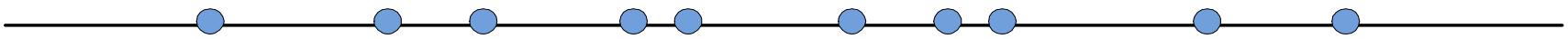


Minimum Enclosing Ball

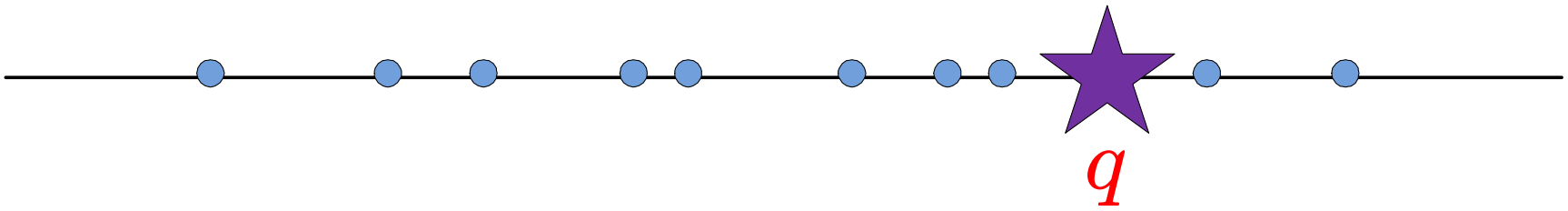
- Input: P in $\mathbb{R}^d$
- Query: a point $q \in \mathbb{R}^d$
- Output: $\mathsf{far}(P, q) = \max_{p \in P} \mathsf{dist}(p, q)$



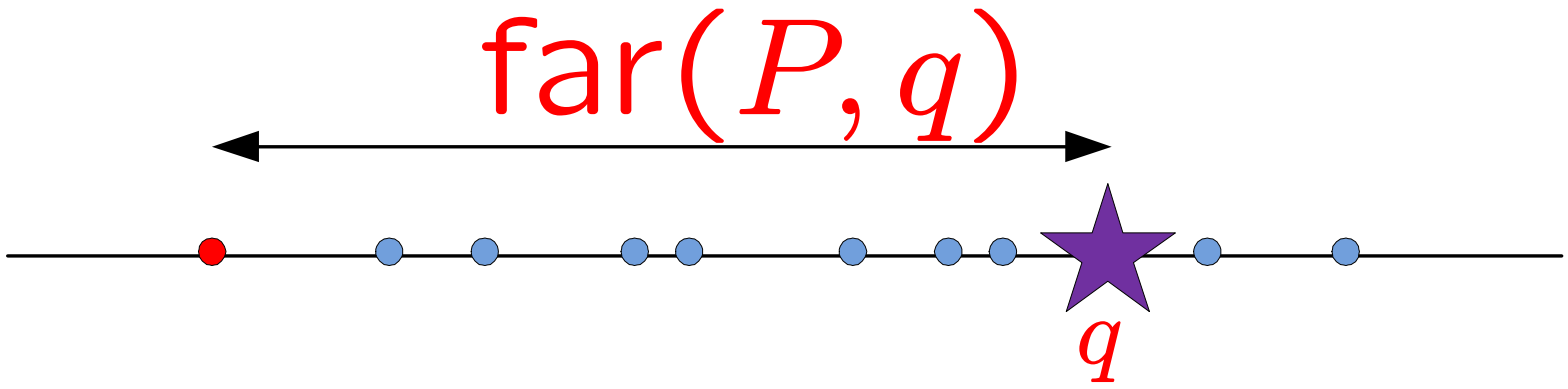
Coreset for Enclosing Balls $P \subseteq \mathbb{R}$



Coreset for Enclosing Balls $P \subseteq \mathbb{R}$



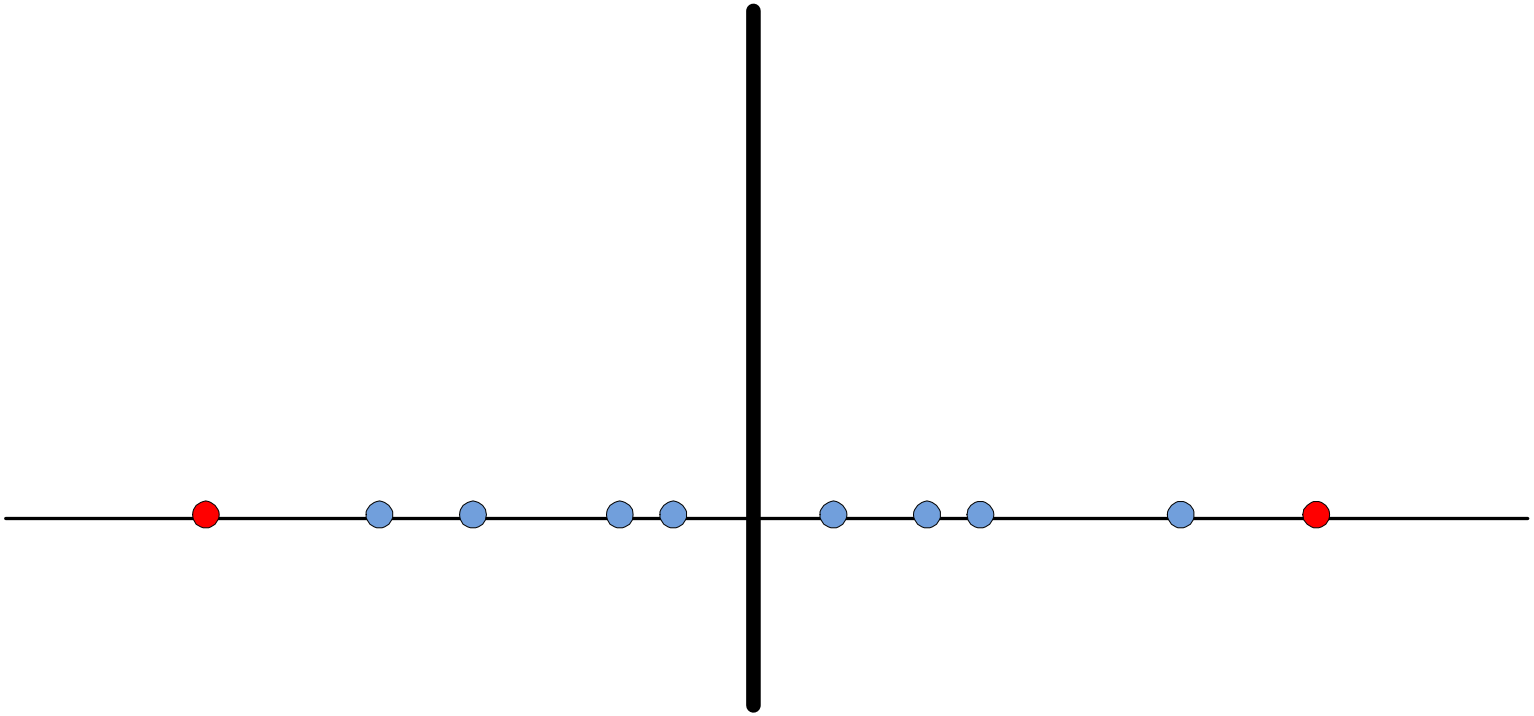
Coreset for Enclosing Balls $P \subseteq \mathbb{R}$



Coreset for Enclosing Balls $P \subseteq \mathbb{R}$

The fathest point from every query $q \in \mathbb{R}$

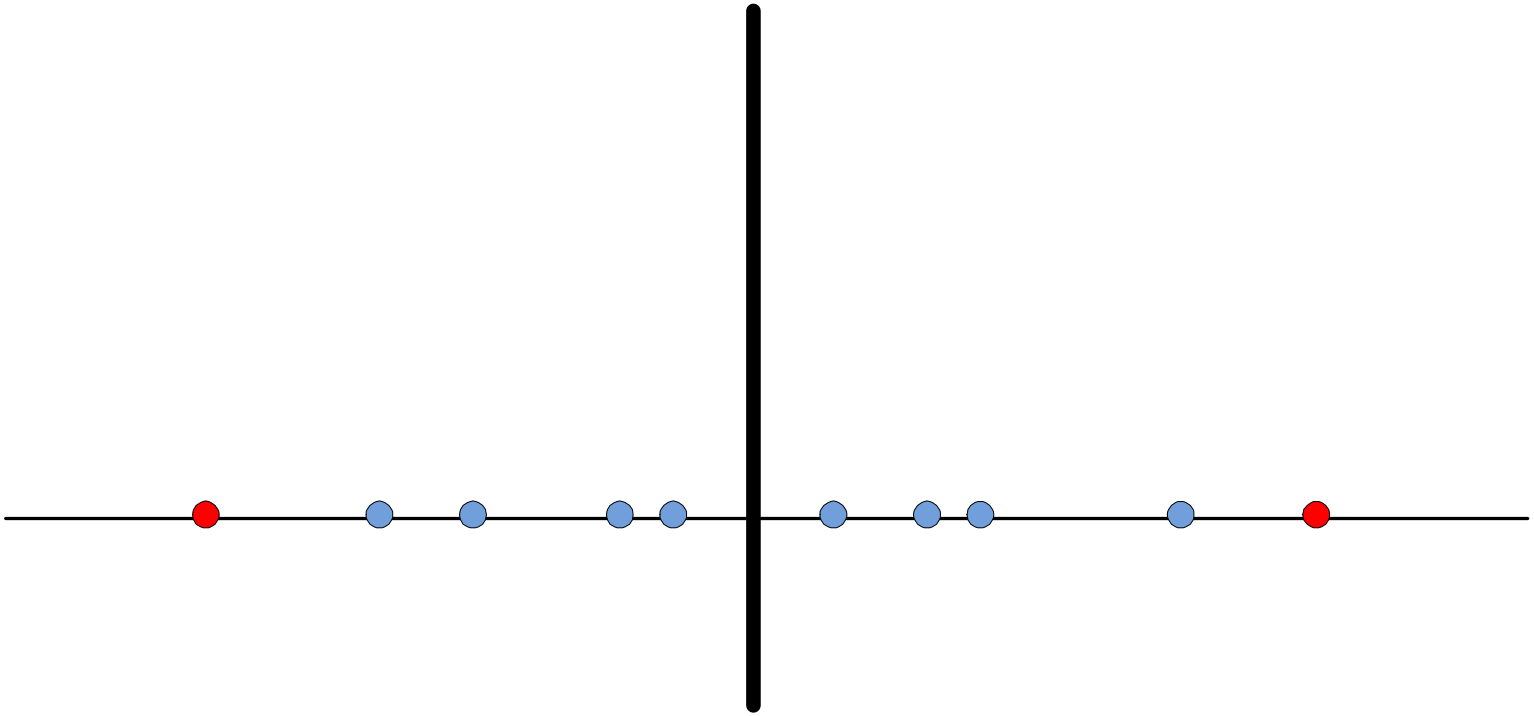
is a red point



Coreset for Enclosing Balls $P \subseteq \mathbb{R}$

The fathest point from every query $q \in \mathbb{R}^d$

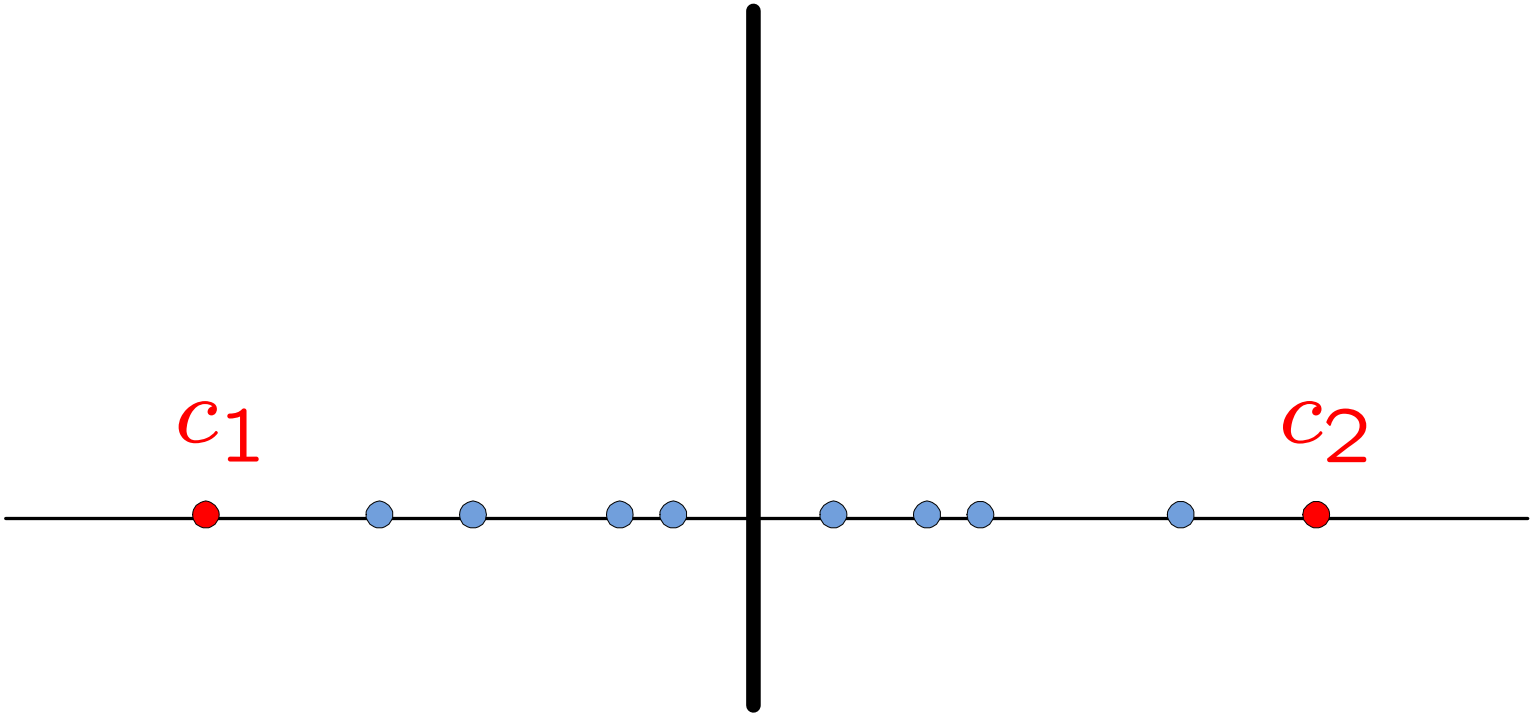
is a red point



Coreset for Enclosing Balls $P \subseteq \mathbb{R}$

The fathest point from every query $q \in \mathbb{R}^d$

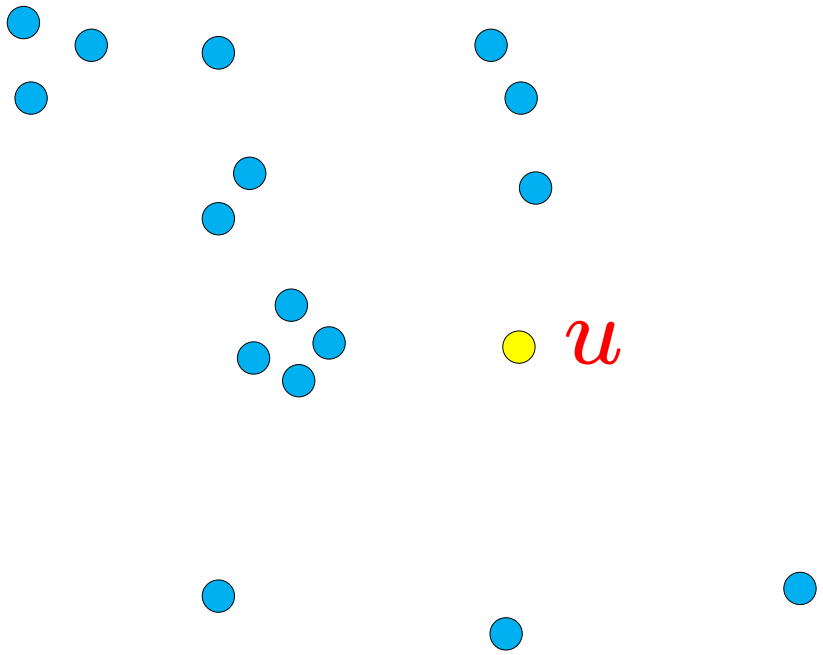
is a red point



$C := \{c_1, c_2\}$ is a coreset for P

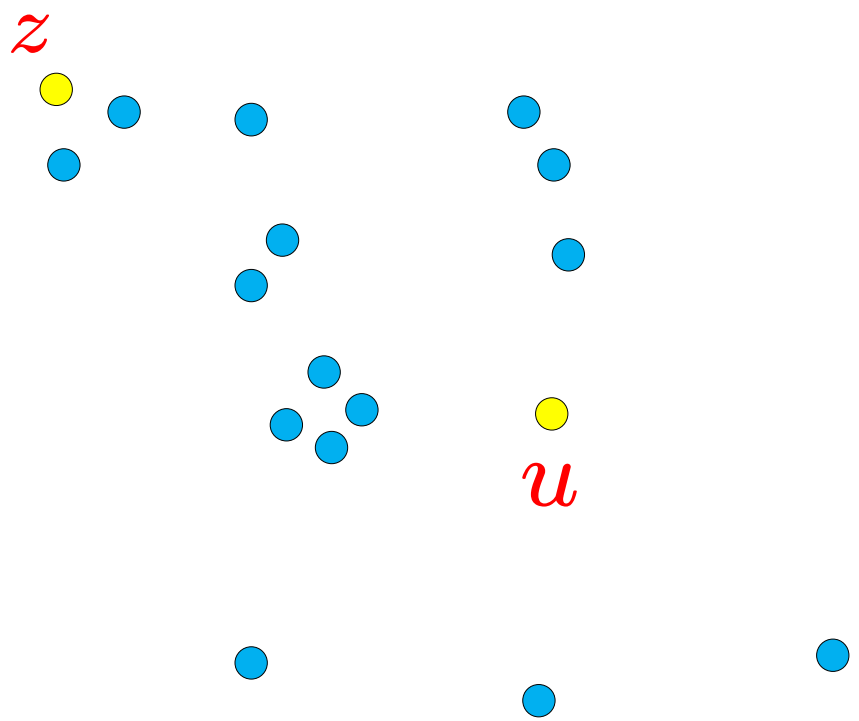
Coreset for Enclosing Balls

1) Choose an arbitrary point $u \in P$



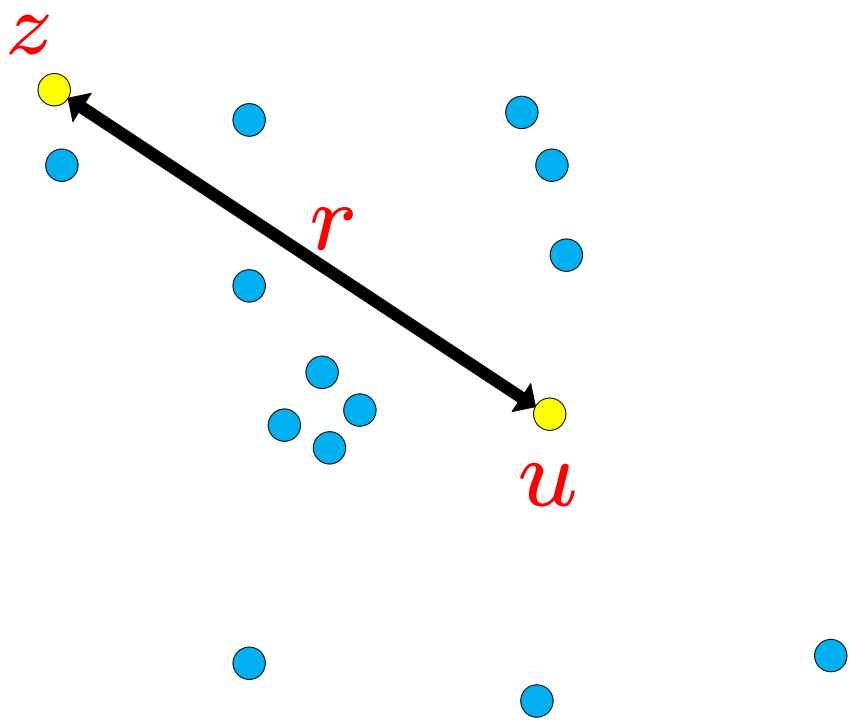
Coreset for Enclosing Balls

2) Find the farthest point $z \in P$ from u



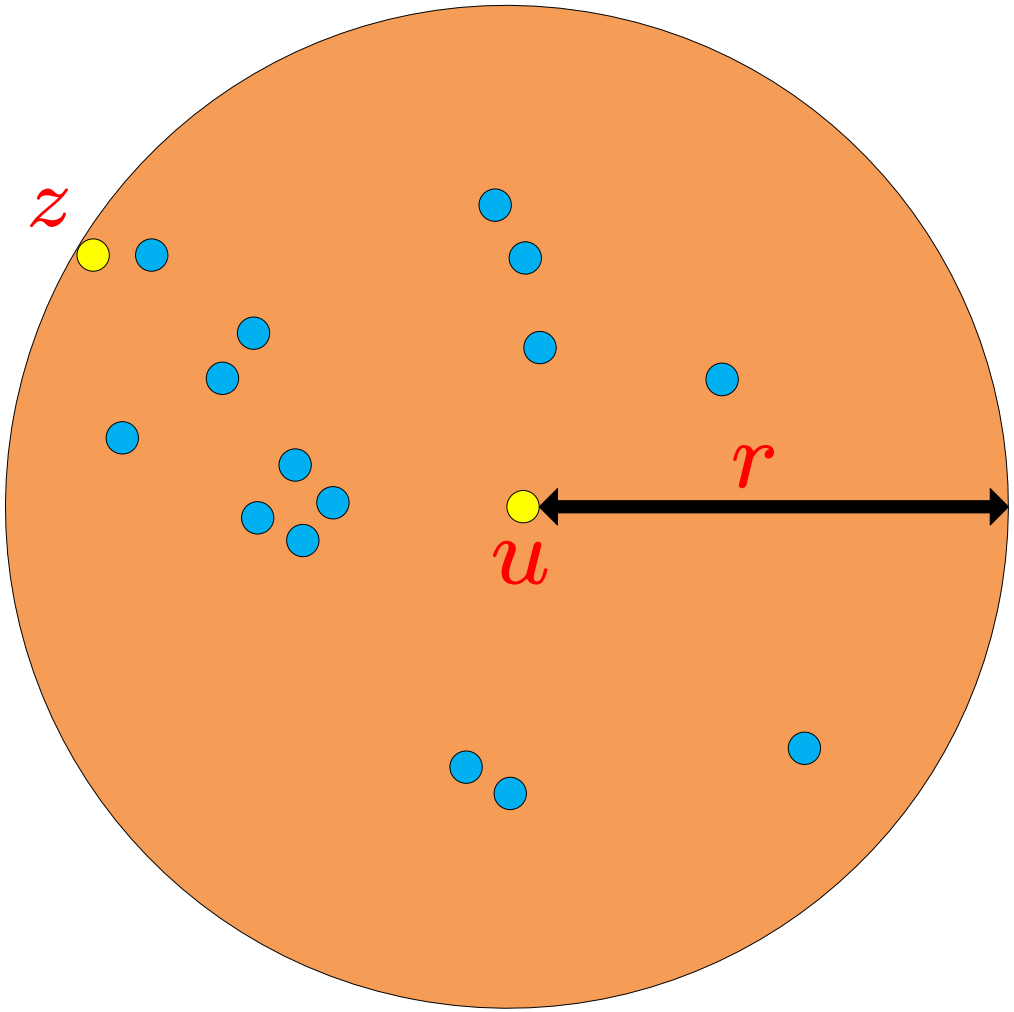
Coreset for Enclosing Balls

$r := \mathrm{dist}(u, z)$



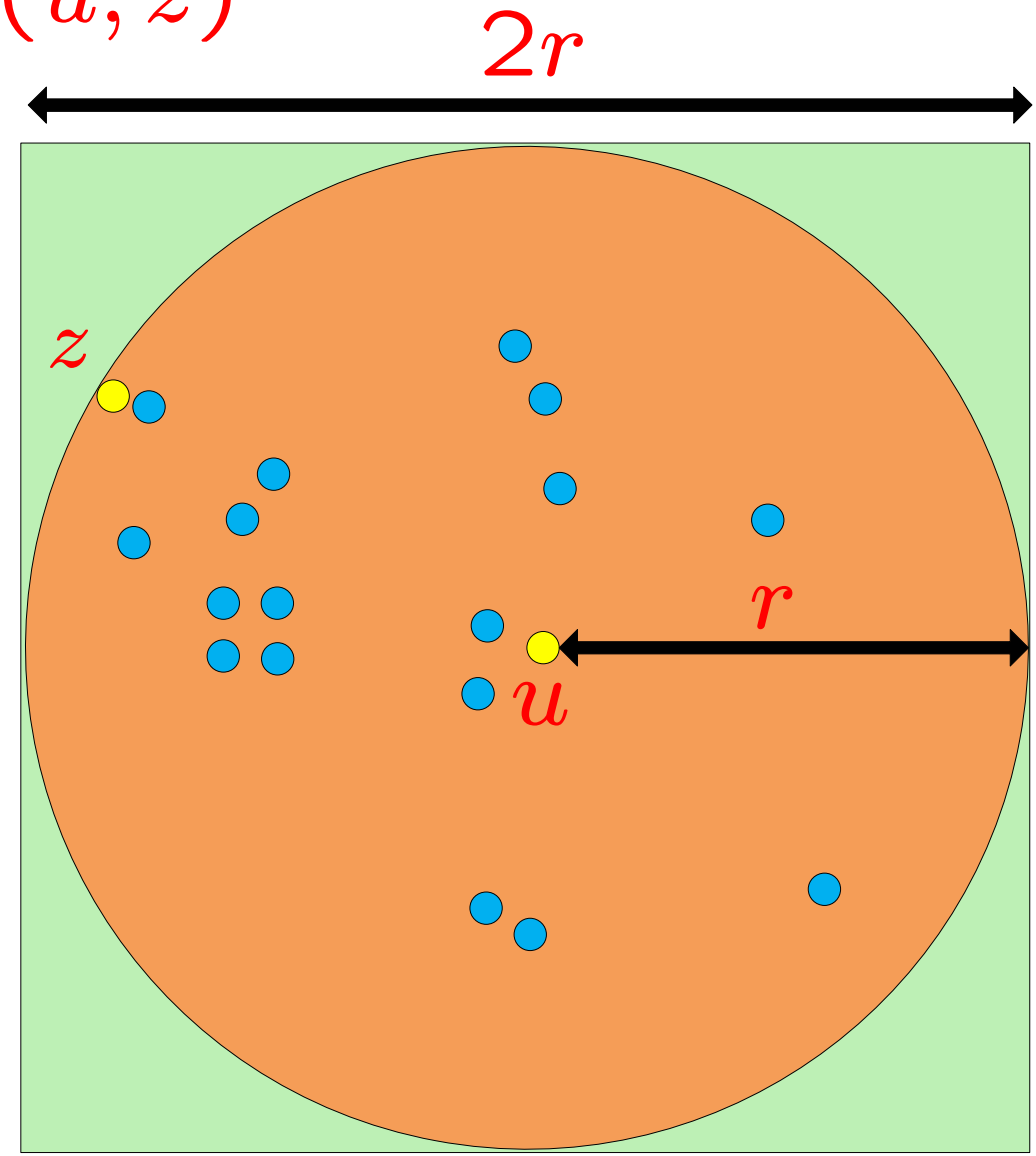
Coreset for Enclosing Balls

$r := \mathrm{dist}(u, z)$



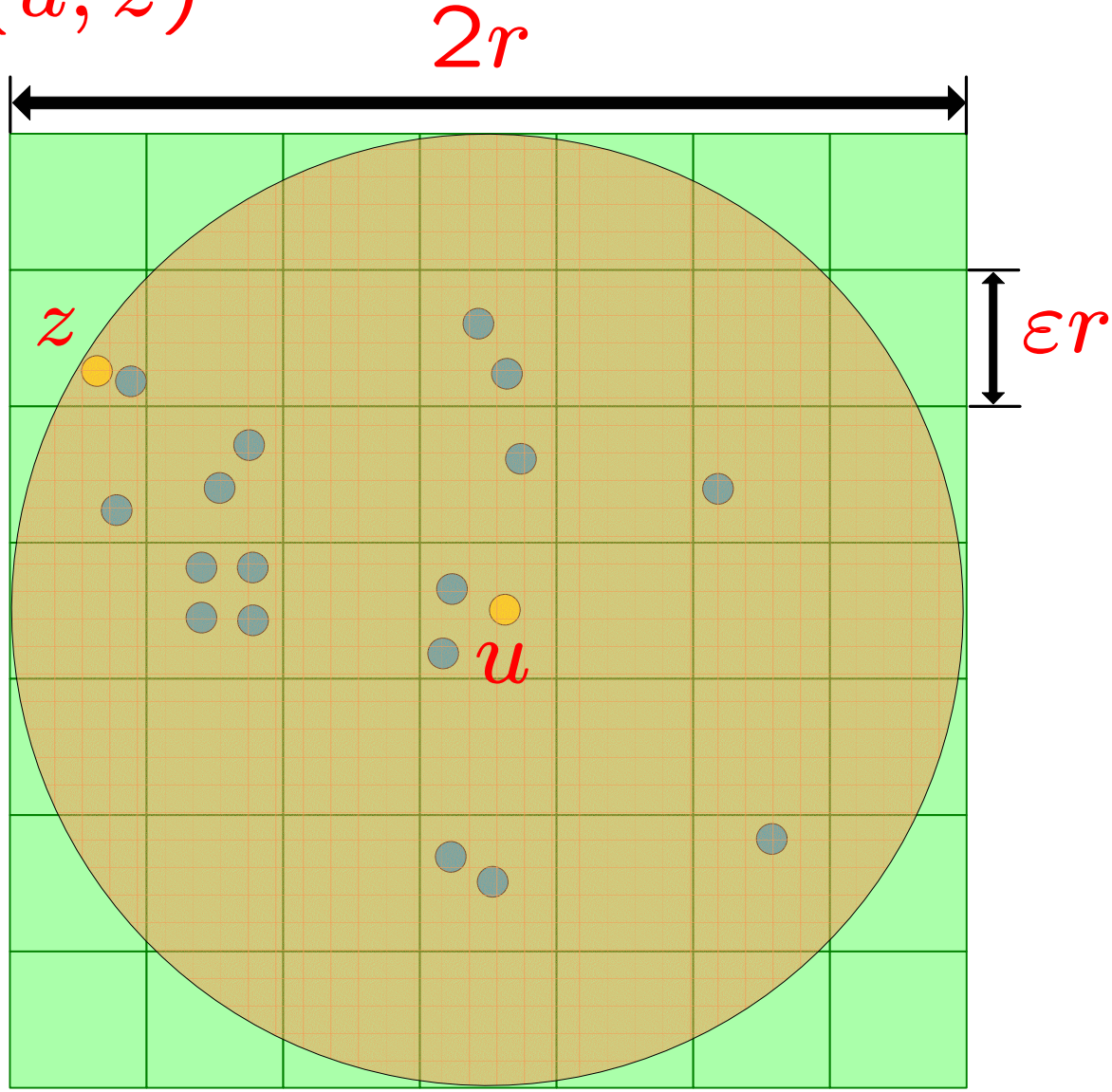
Coreset for Enclosing Balls

$r := \mathrm{dist}(u, z)$



Coreset for Enclosing Balls

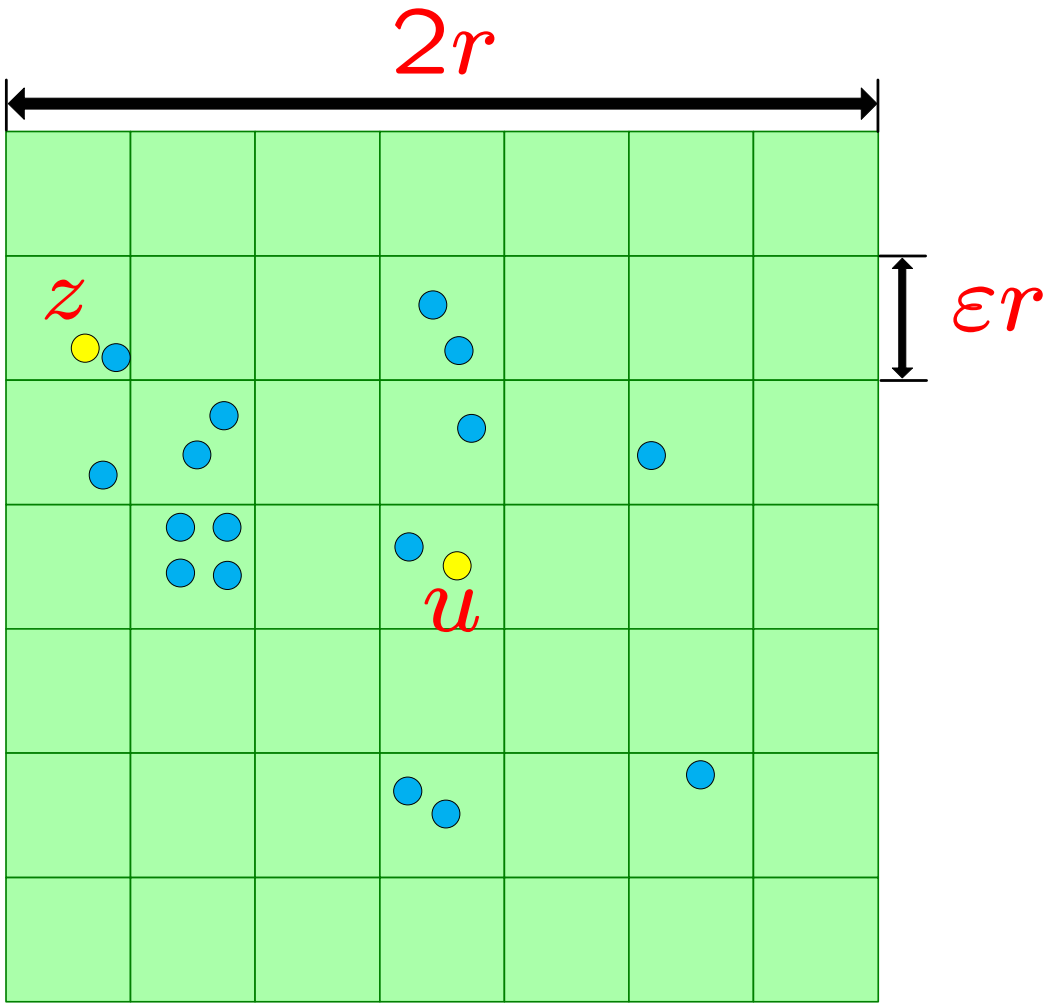
$r := \mathrm{dist}(u, z)$



Coreset for Enclosing Balls

3) Construct a grid of $\frac{2}{\varepsilon^2}$ cells of size $\varepsilon r \times \varepsilon r$, centered at u

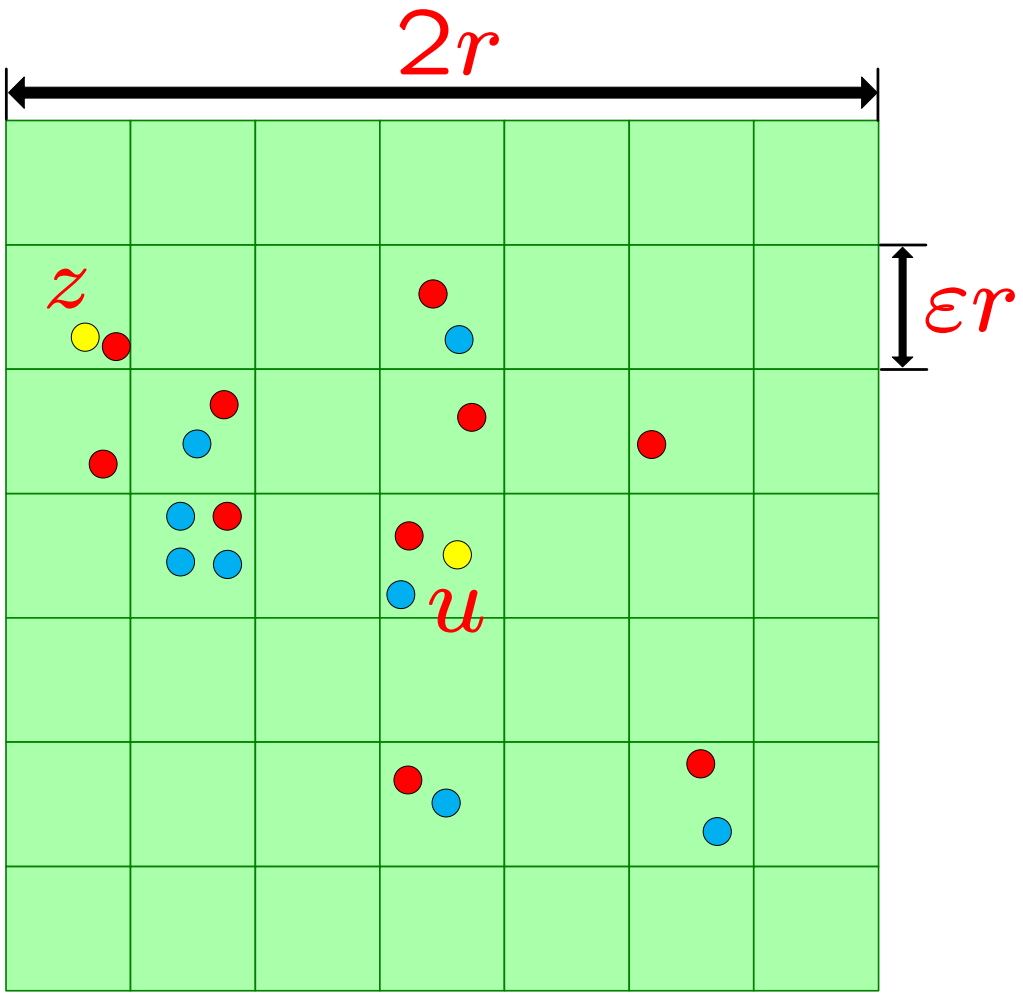
$r := \mathsf{dist}(u, z)$



Coreset for Enclosing Balls

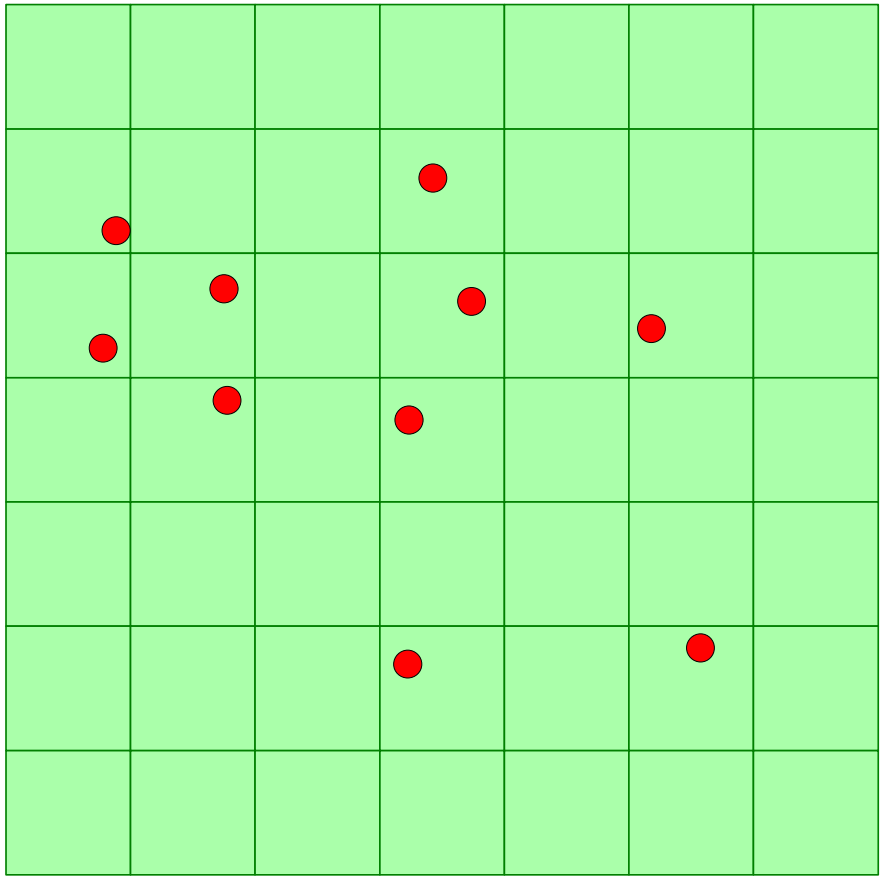
4) Pick a representative point from each non-empty cell

$r := \mathsf{dist}(u, z)$



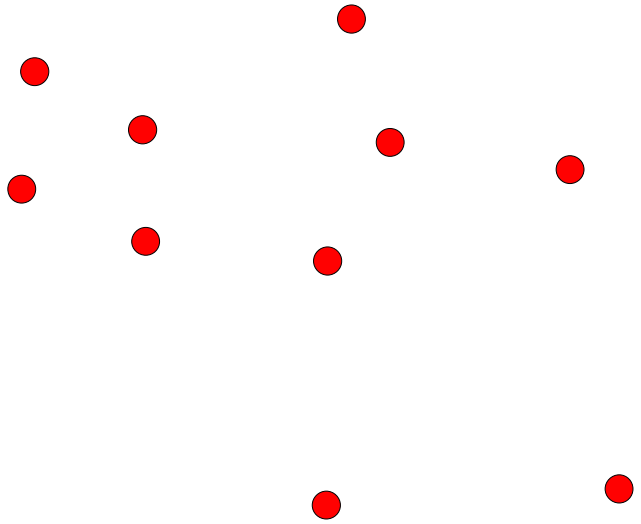
Coreset for Enclosing Balls

5) $C :=$ the set of the $O\left(\frac{1}{\varepsilon^2}\right)$ representatives



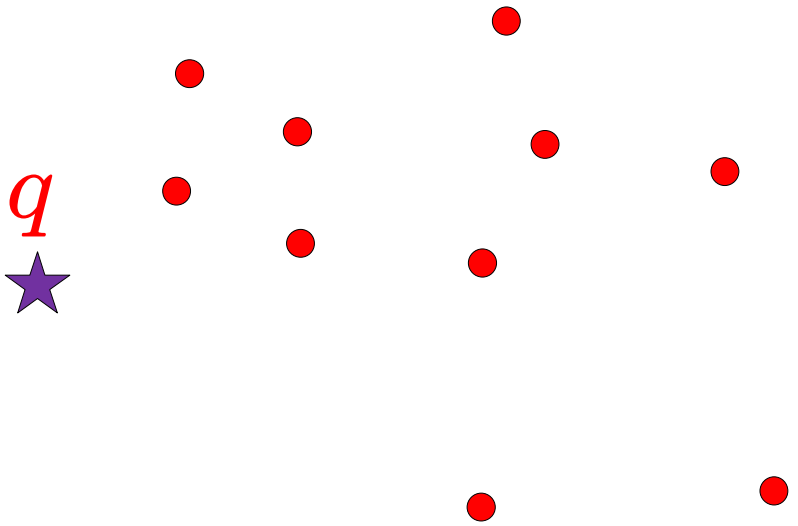
Coreset for Enclosing Balls

6) Return C



Proof of Correctness

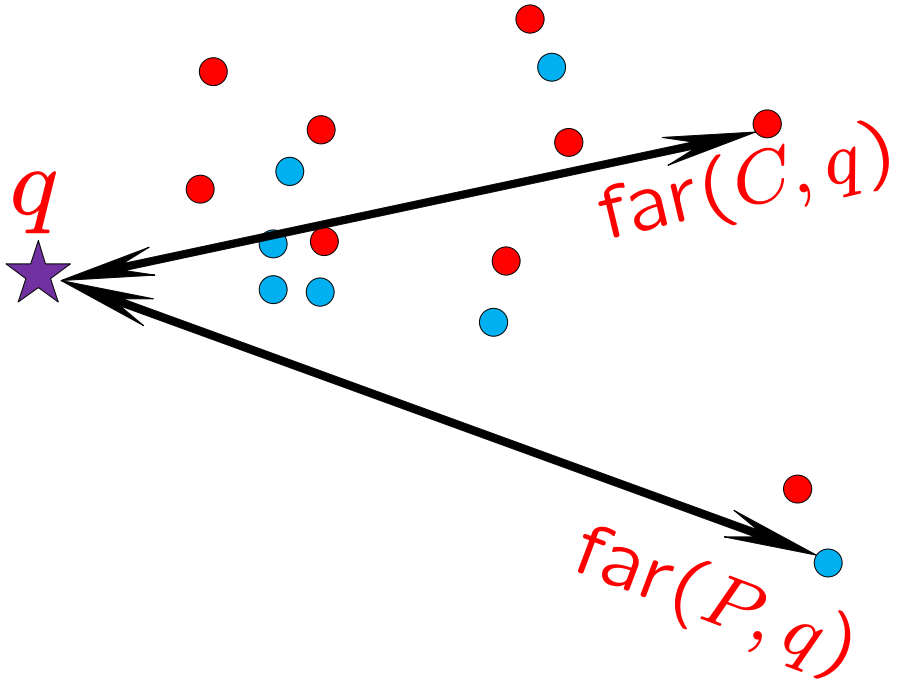
q := an arbitrary query point



Proof of Correctness

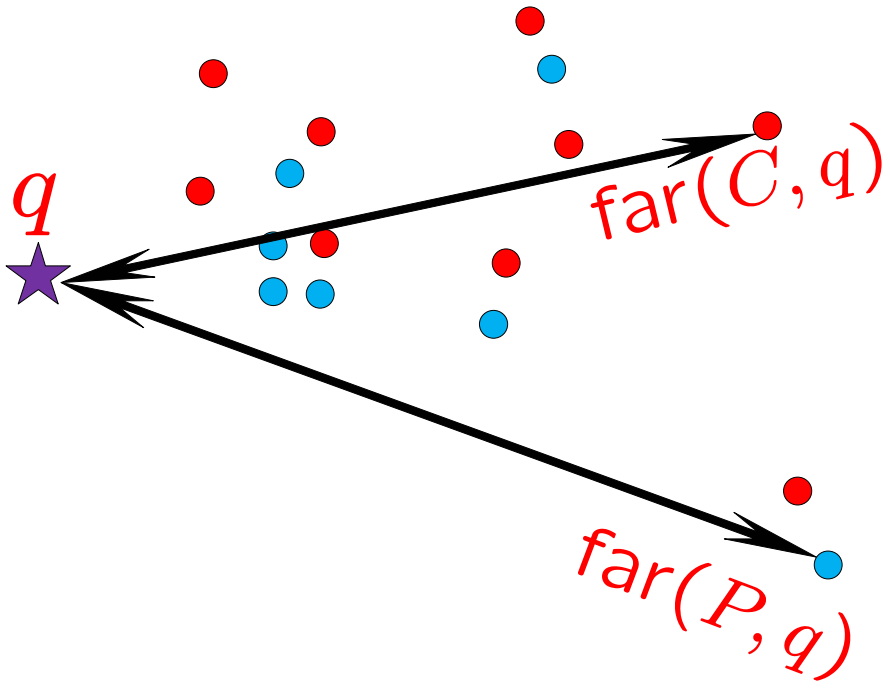
$$\mathrm{far}(P, q) = \max_{p \in P} \mathrm{dist}(p, q)$$

$$\mathrm{far}(C, q) = \max_{c \in C} \mathrm{dist}(c, q)$$



Proof of Correctness

$C \subseteq P \quad \Longrightarrow \quad \mathrm{far}(C, q) \leq \mathrm{far}(P, q)$

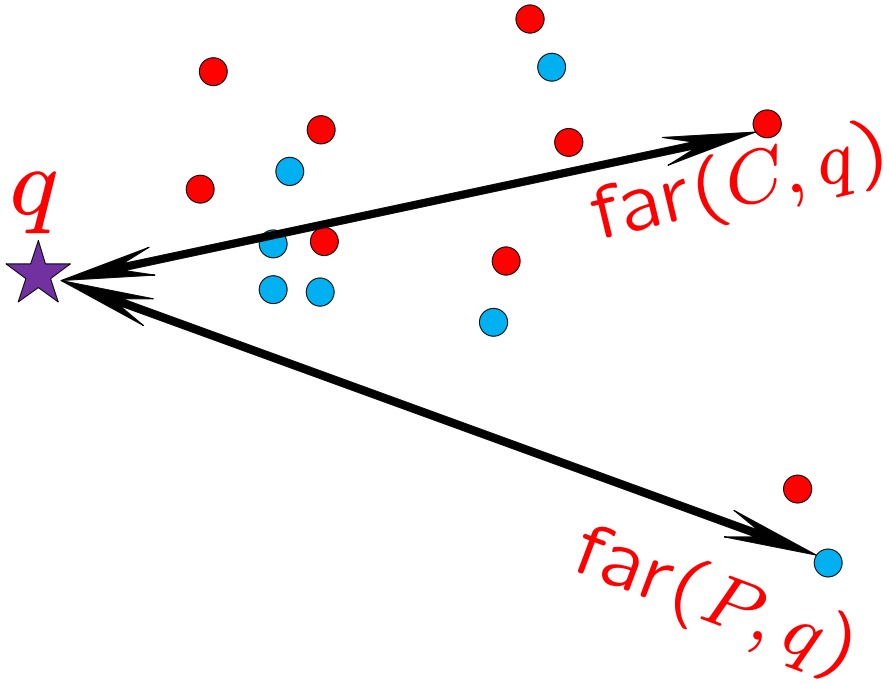


Proof of Correctness

$C \subseteq P \quad \Longrightarrow \quad \mathrm{far}(C,q) \leq \mathrm{far}(P,q)$

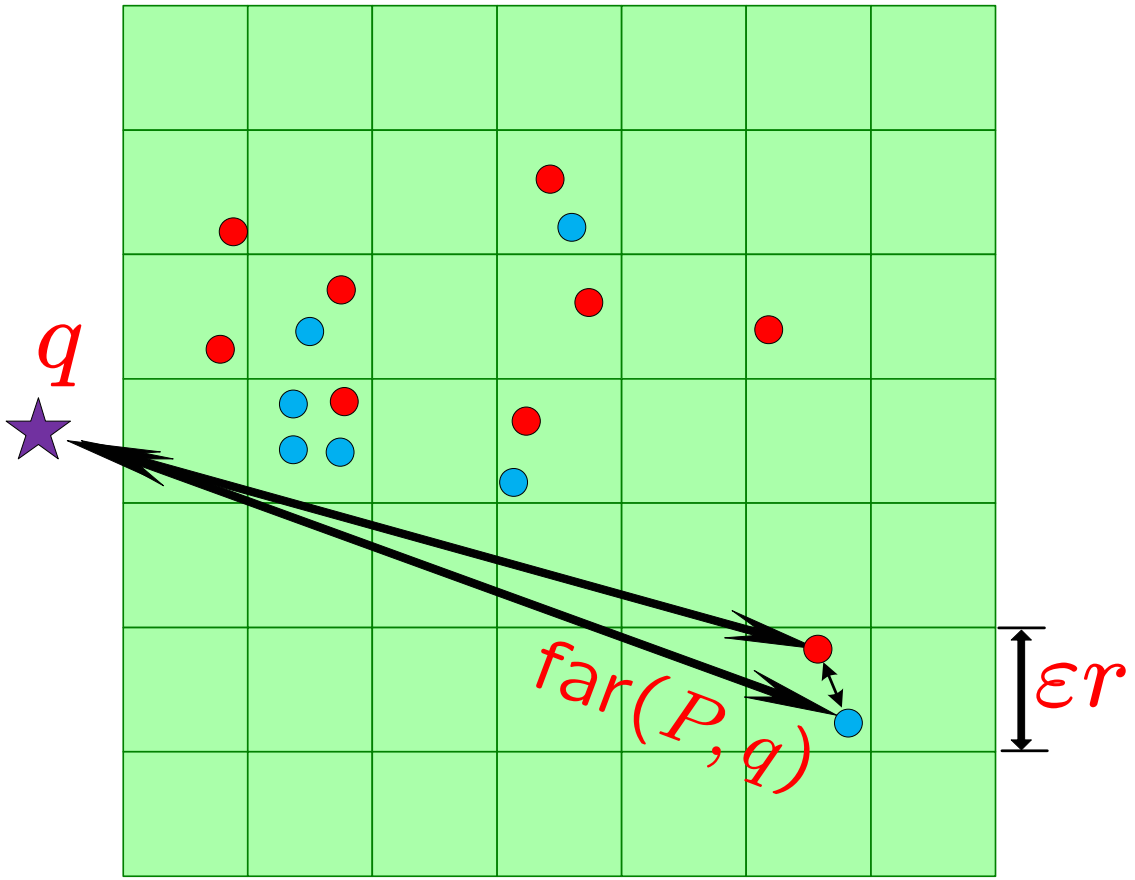
Need to prove:

$\mathrm{far}(P,q) - \mathrm{far}(C,q) \leq O(\varepsilon)\mathrm{far}(P,q)$



Proof of Correctness

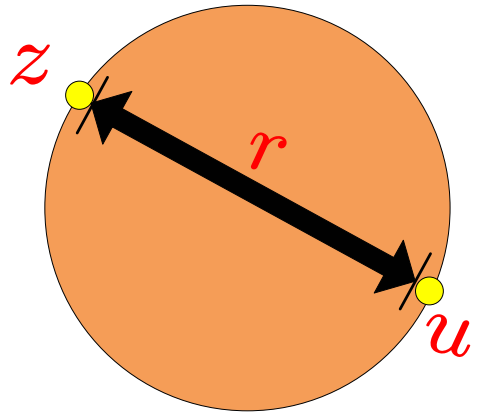
$$\mathsf{far}(P, q) \leq \mathsf{far}(C, q) + O(\varepsilon r)$$



Main Observation

Every ball that covers u and z,

has a diameter of at least r:

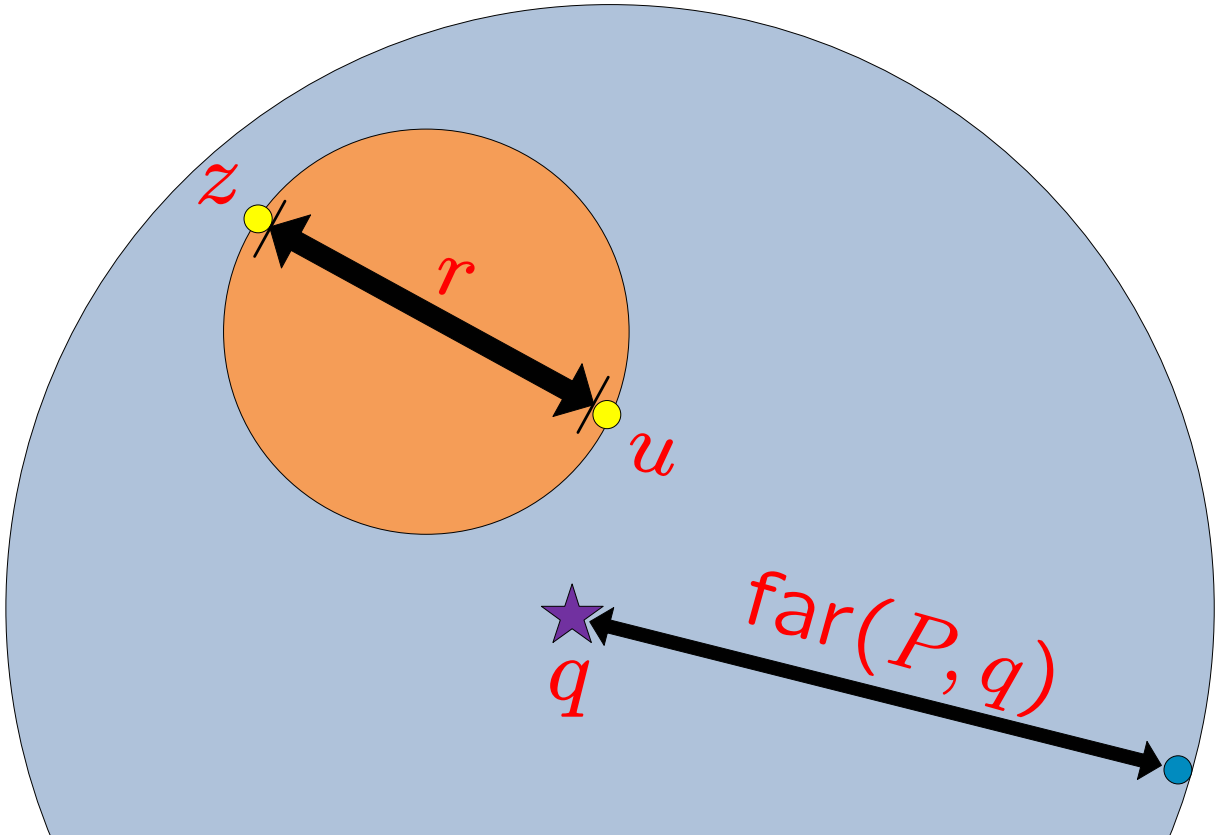


Minimum enclosing ball of $\{u, z\}$

Main Observation

Every ball that covers u and z,

has a diameter of at least r:

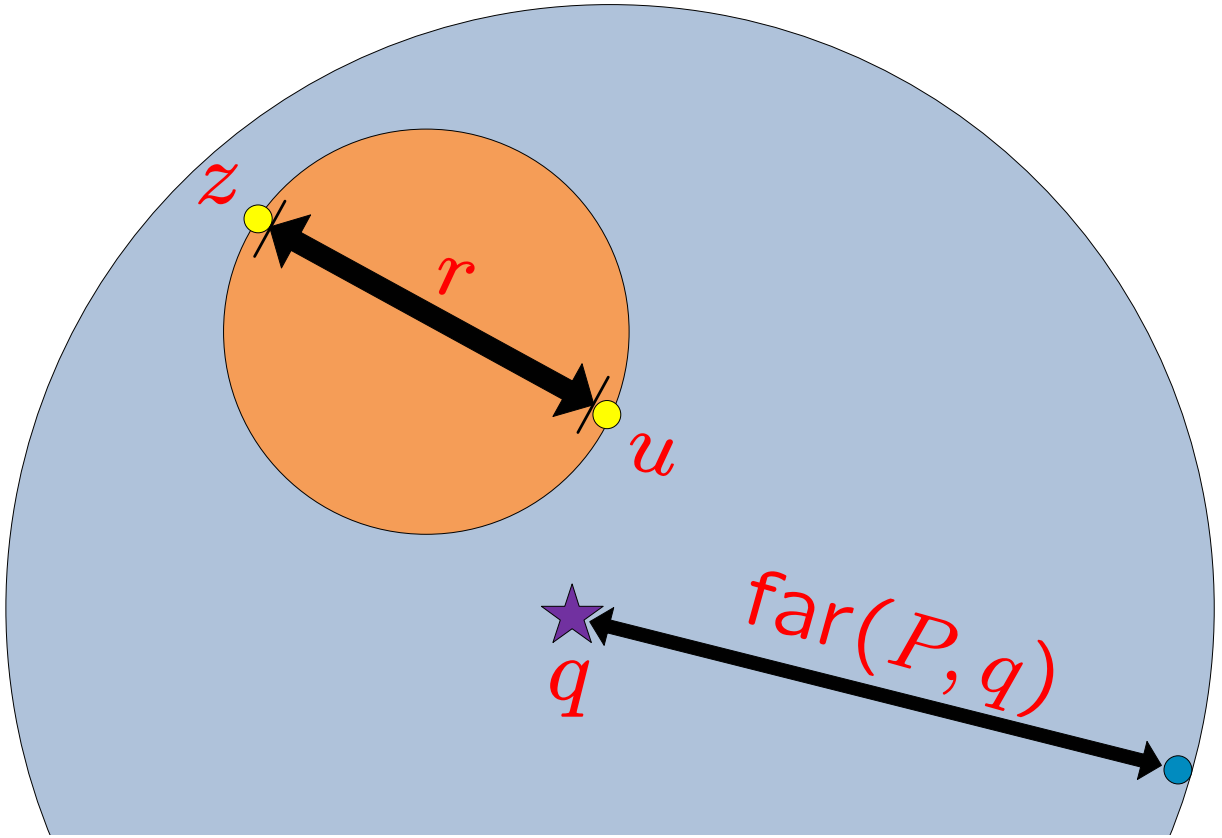


$r \leq 2\mathsf{far}(P,q)$

Main Observation

Every ball that covers u and z,

has a diameter of at least r:

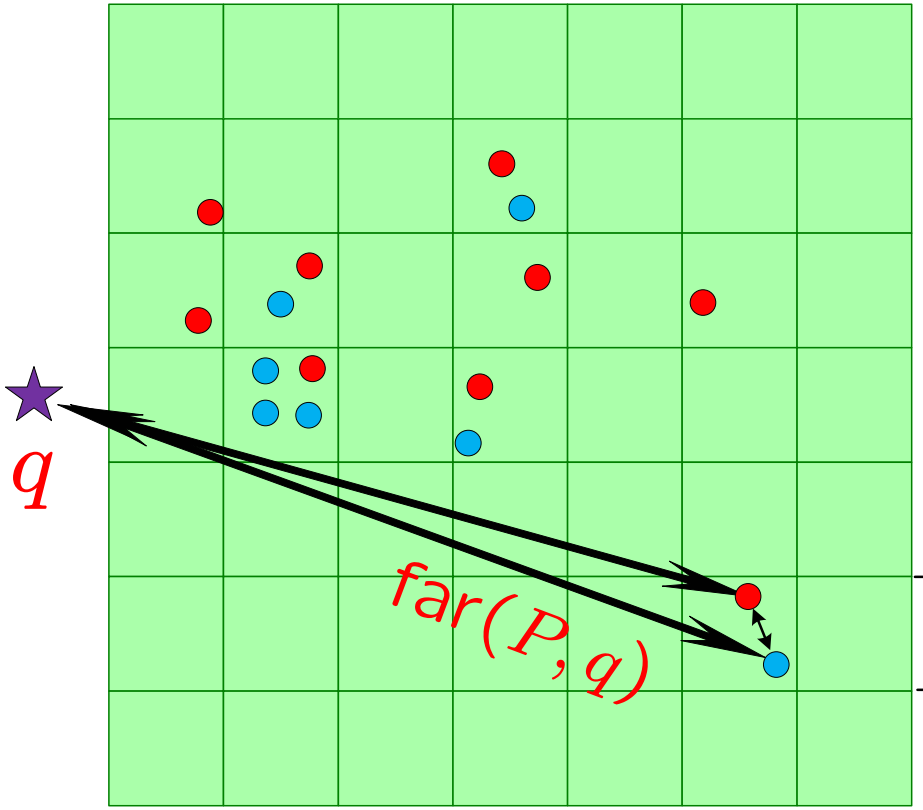


$$r \leq 2\mathsf{far}(P,q) \quad \Rightarrow \quad O(\varepsilon r) \leq O(\varepsilon)\mathsf{far}(P,q)$$

Proof of Correctness

$$\mathsf{far}(P,q) \leq \mathsf{far}(C,q) + O(\varepsilon r)$$

$$\leq \mathsf{far}(C,q) + O(\varepsilon)\mathsf{far}(P,q)$$



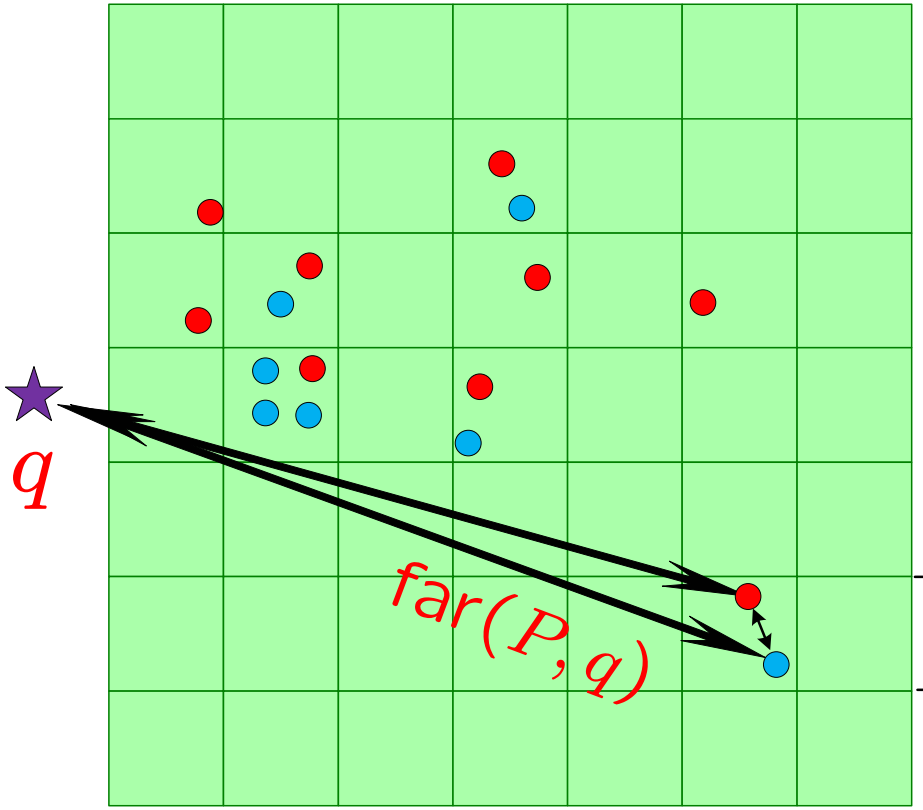
$$\varepsilon r \leq O(\varepsilon)\mathsf{far}(P,q)$$

Proof of Correctness

$$\mathsf{far}(P,q) \leq \mathsf{far}(C,q) + O(\varepsilon r)$$

$$\leq \mathsf{far}(C,q) + O(\varepsilon)\mathsf{far}(P,q)$$

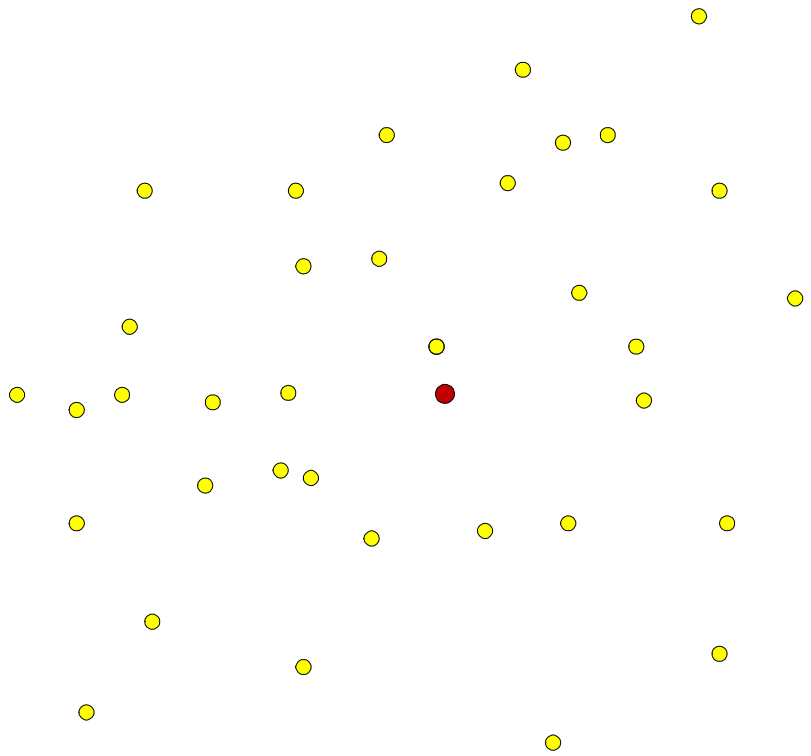
■



$\varepsilon r \leq O(\varepsilon)\mathsf{far}(P,q)$

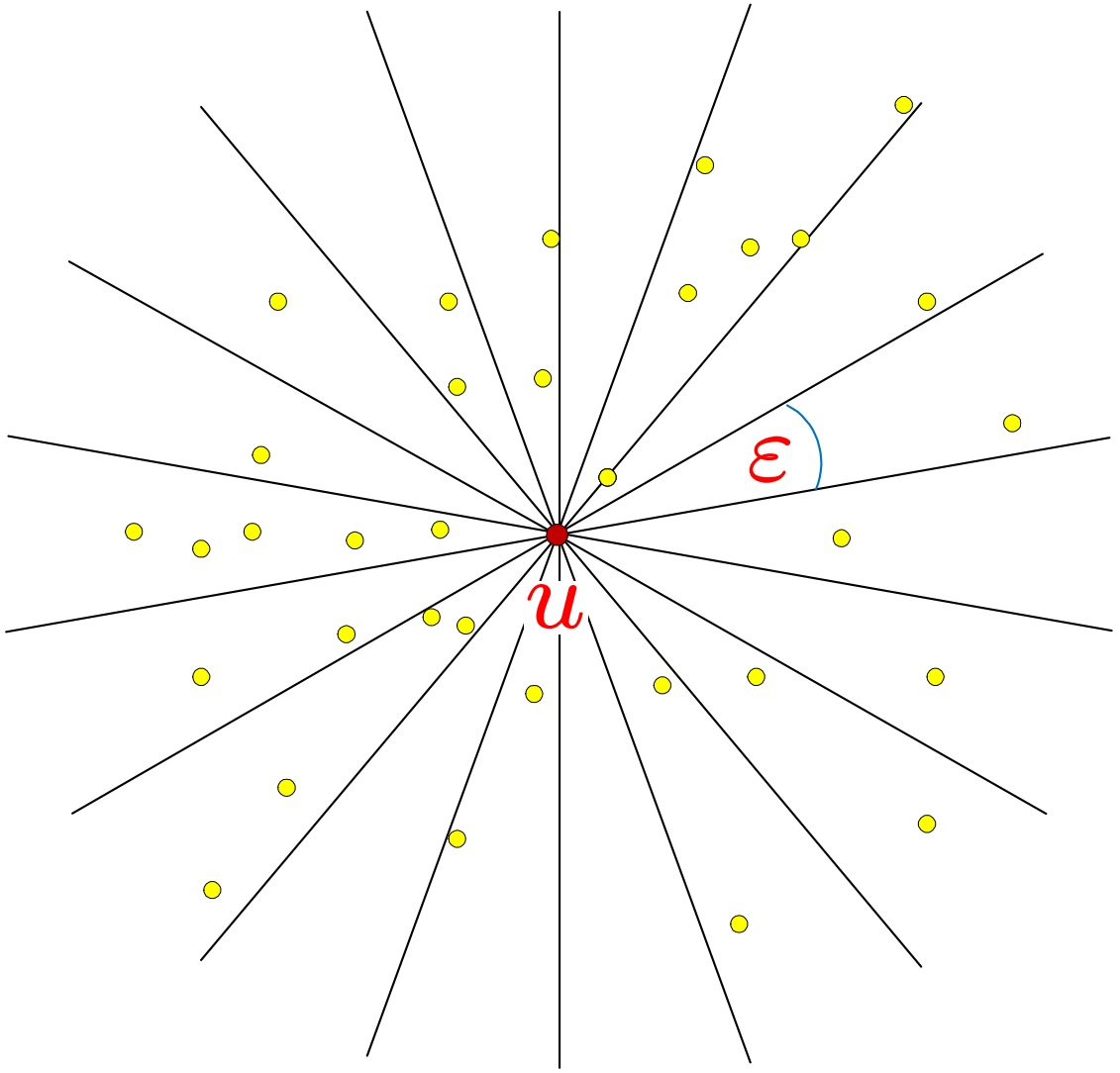
Smaller Coreset

1) Choose an arbitrary point $u \in P$



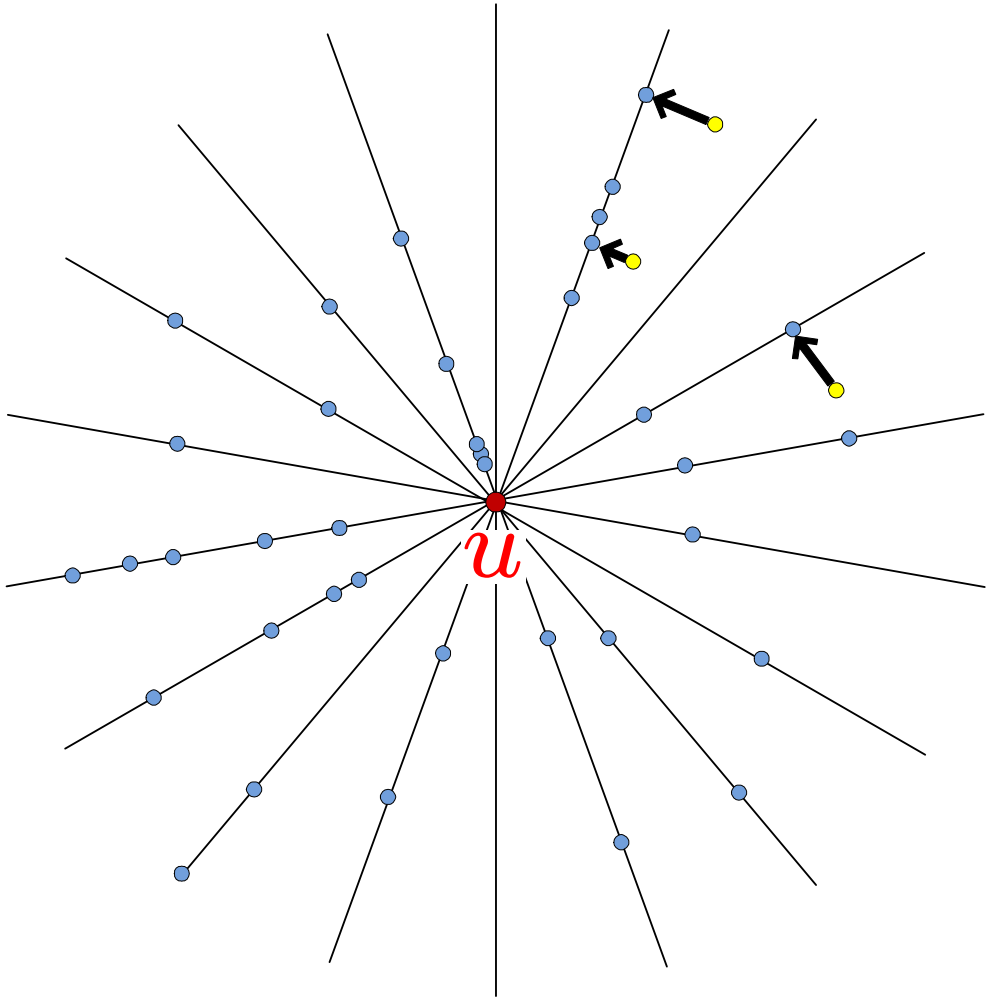
Smaller Coreset

2) Draw a "star" of $\frac{2\pi}{\varepsilon}$ lines around u



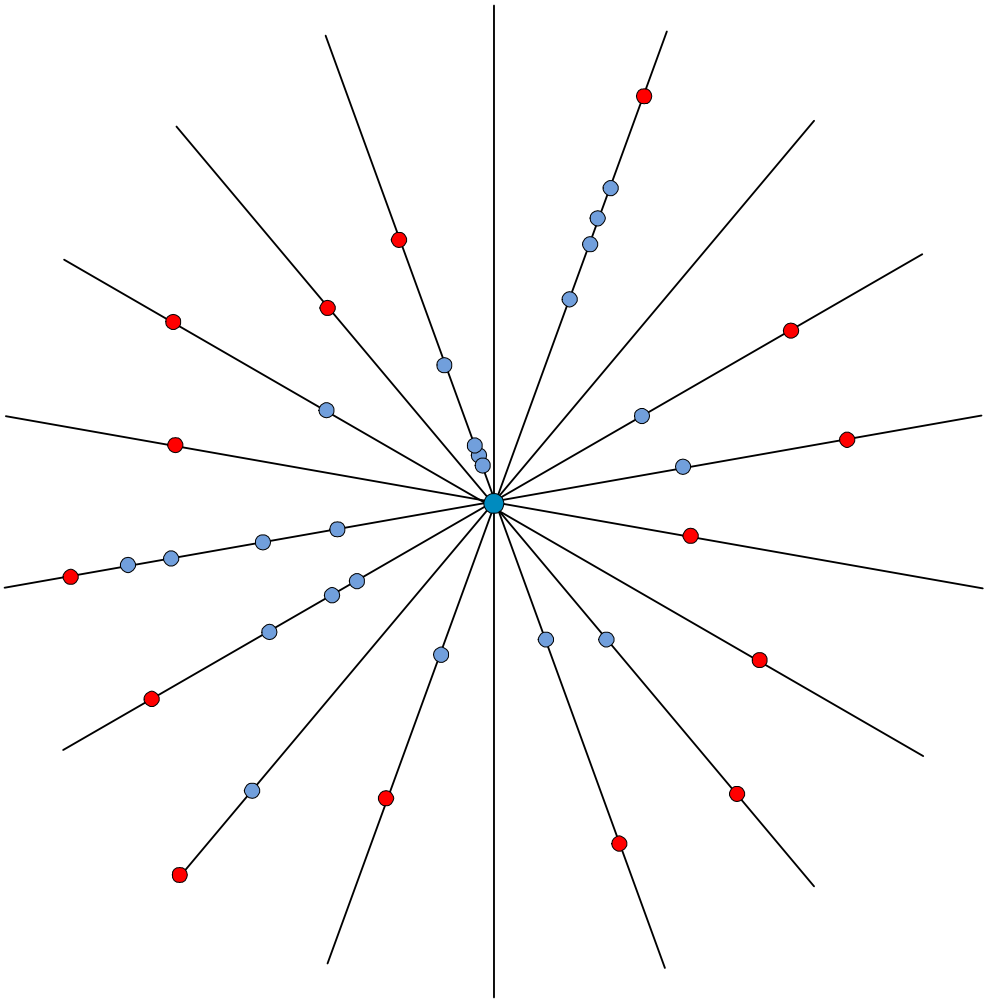
Smaller Coreset

3) $P' :=$ Projection of P onto the lines



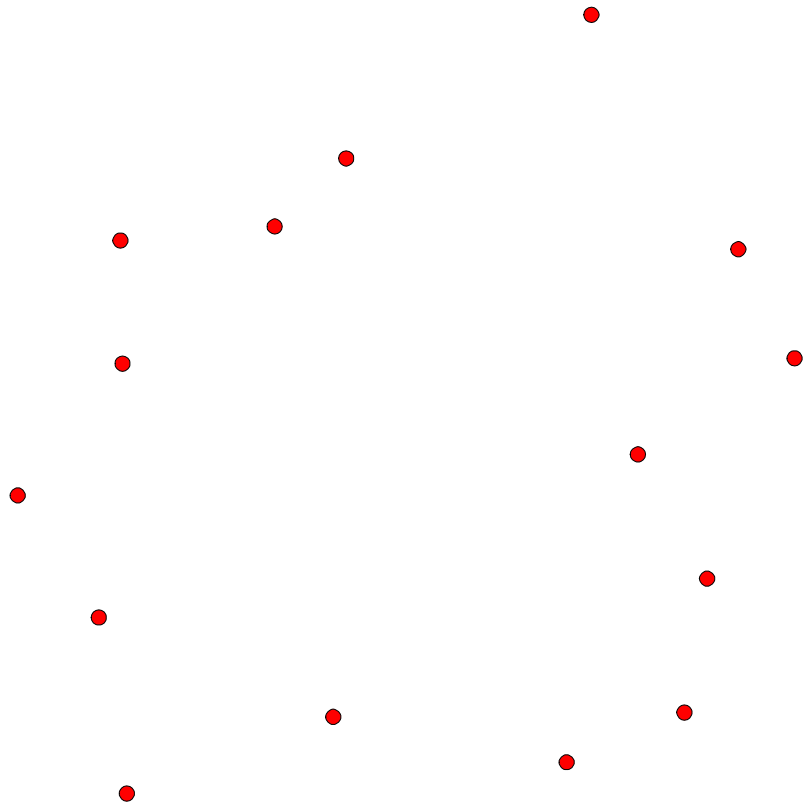
Smaller Coreset

4) $C :=$ union of endpoints on the lines



Smaller Coreset

5) Return C



Proof: Overview

C is a coreset for P'

Proof: Overview

C is a coreset for P'

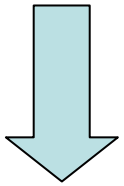
P' is a coreset for P

Proof: Overview

C is a coreset for P'

P' is a coreset for P

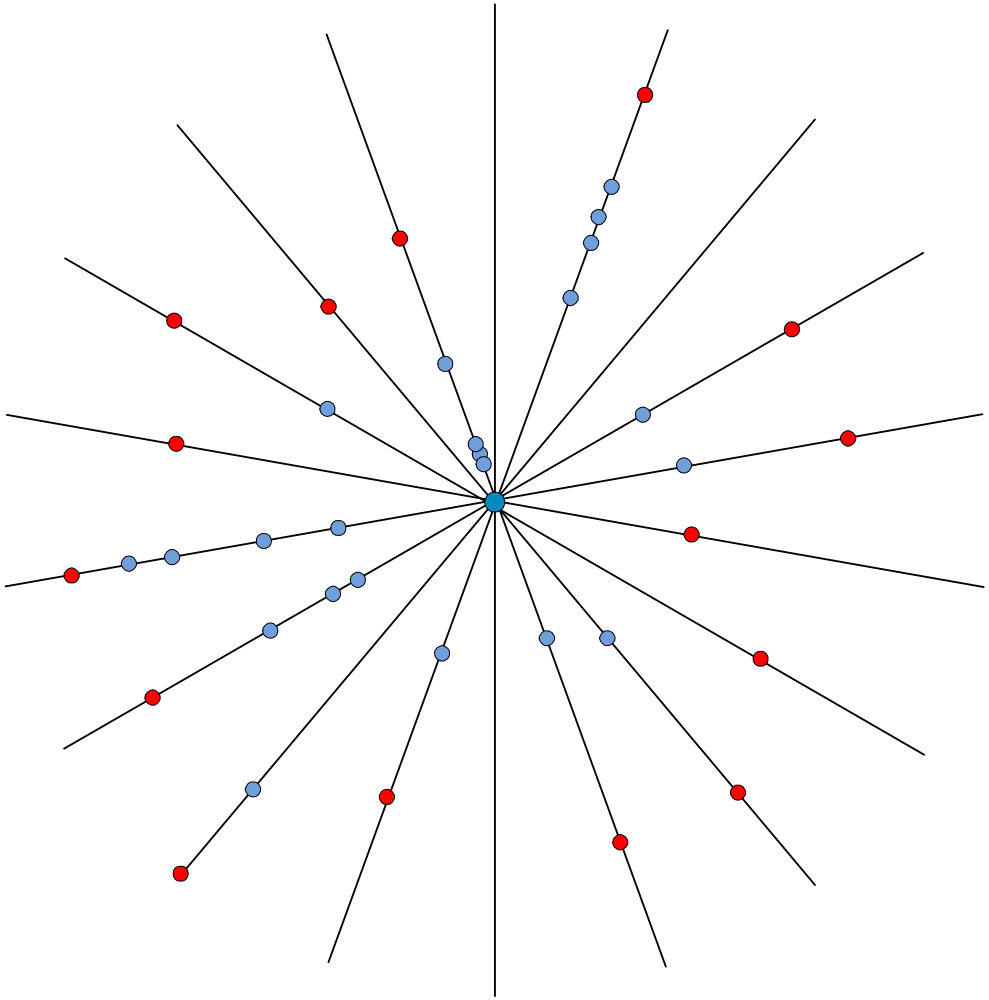
(Large coreset but on few lines)



C is a coreset for P

(Transitive Property)

Claim: C is a coreset for P'

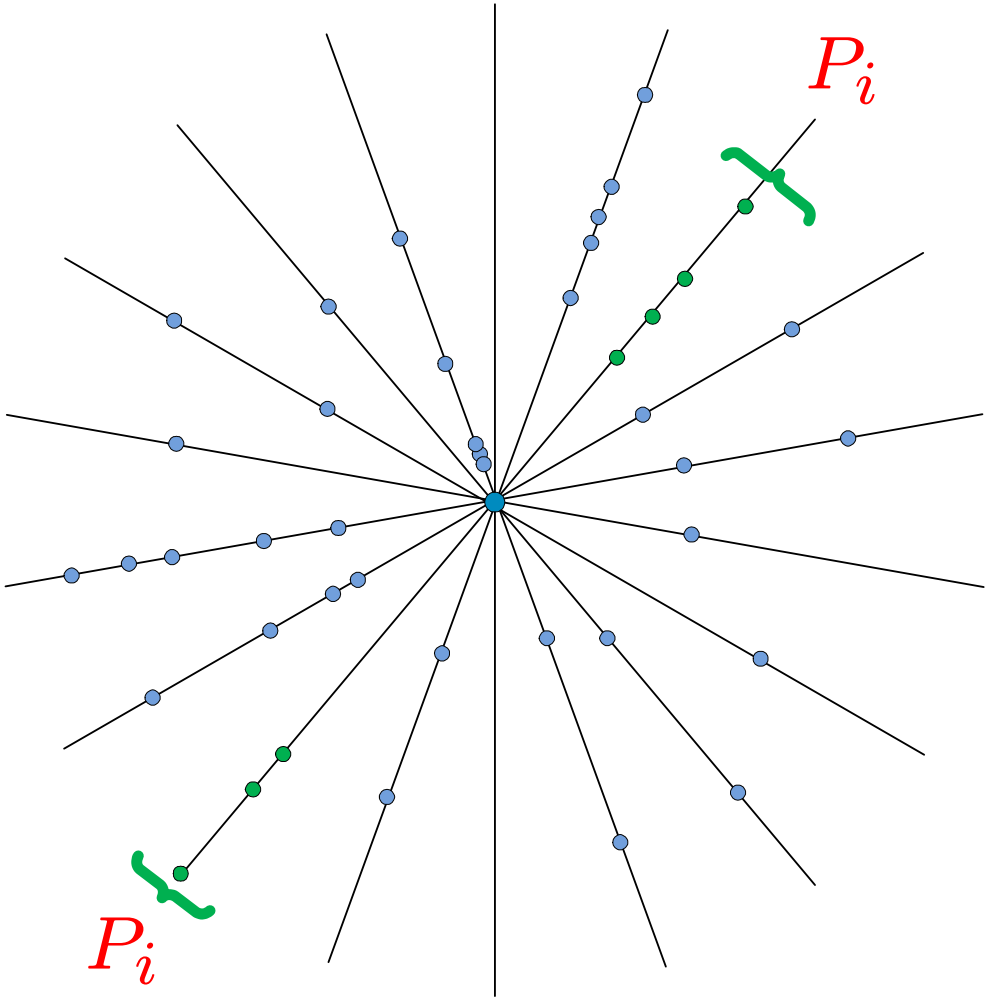


$P' :=$ ● ∪ ●

$C :=$ ●

Claim: C is a coreset for P'

$P_i :=$ intersection of P' with the ith line



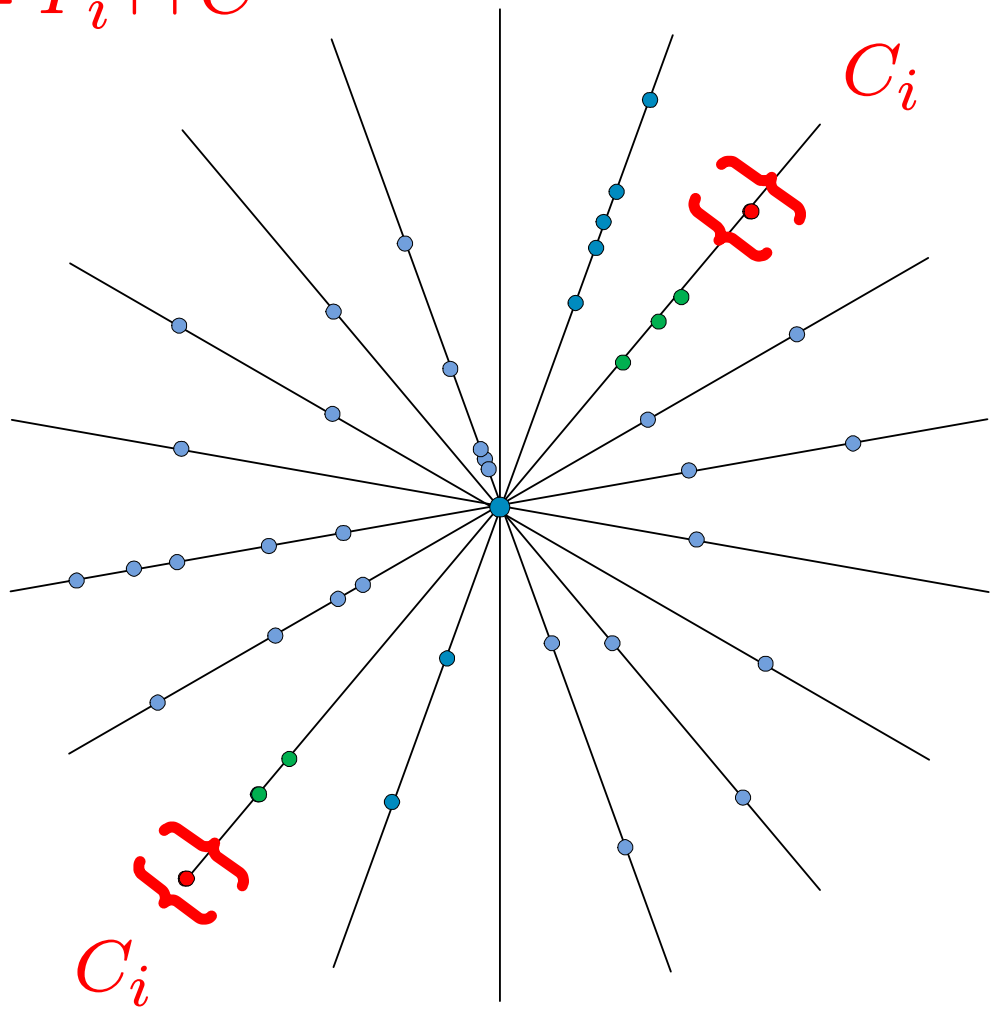
$P' :=$ ● ∪ ●

$P_i :=$ ●

Claim: C is a coreset for P'

$P_i :=$ intersection of P' with the ith line

$C_i := P_i \cap C$



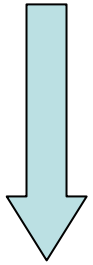
$P_i :=$ ● ∪ ●

$C_i :=$ ●

Claim: C is a coreset for P'

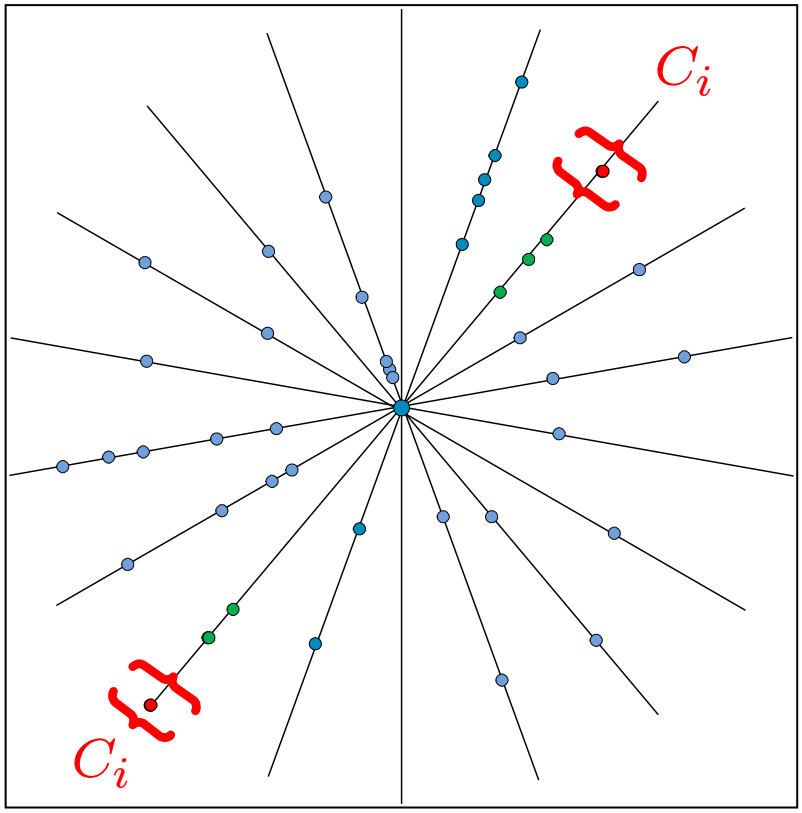
$P_i :=$ intersection of P' with the ith line

$C_i := P_i \cap C$



By proof for $d = 1$

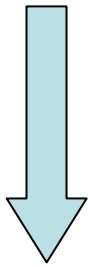
C_i is a coreset for P_i



Claim: C is a coreset for P'

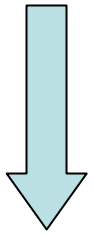
$P_i :=$ intersection of P' with the ith line

$C_i := P_i \cap C$

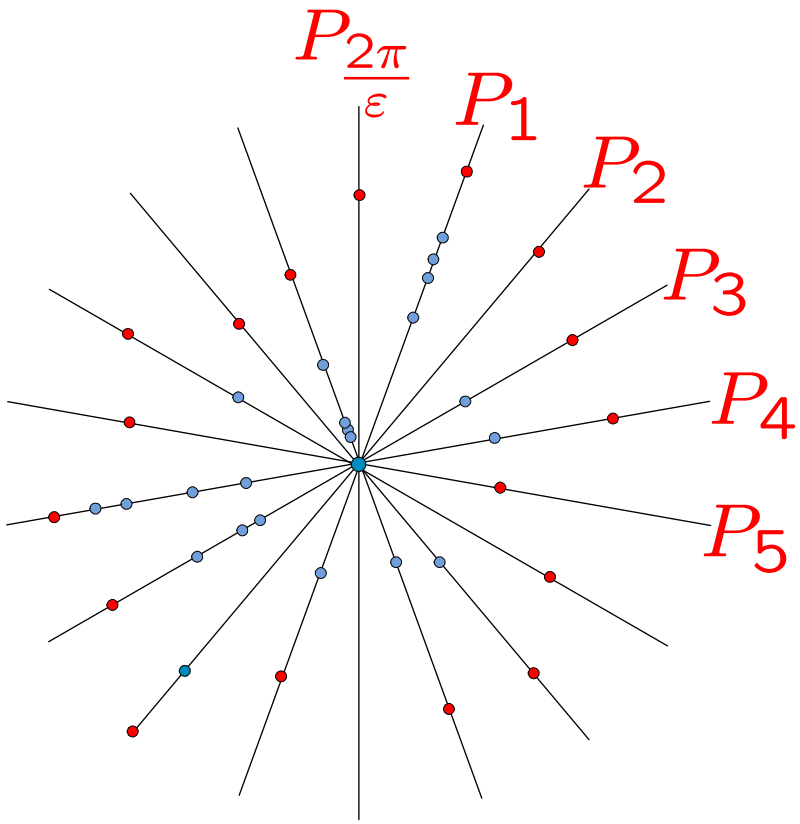


By proof for $d = 1$

C_i is a coreset for P_i



Union Rule

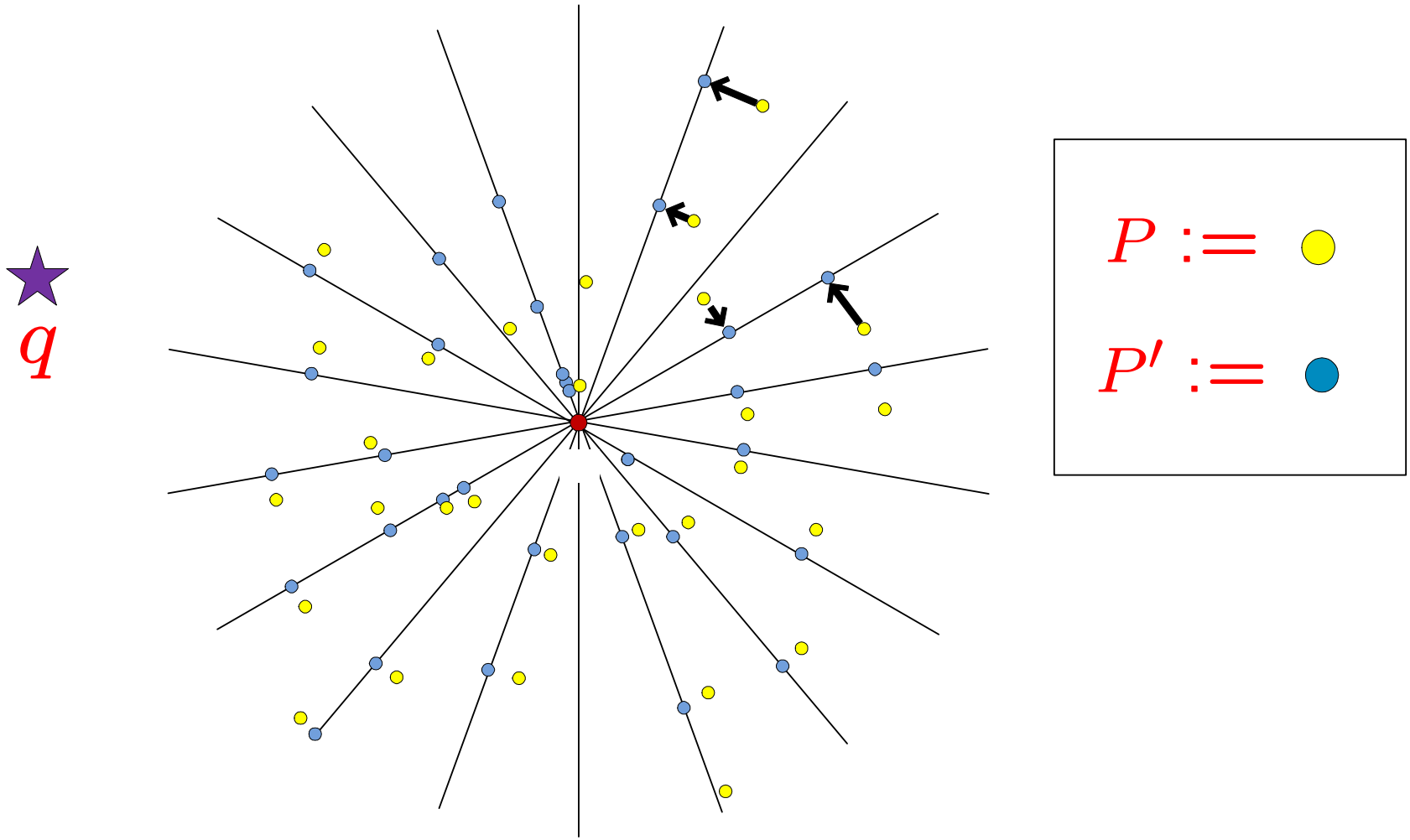


$C = \bigcup_i C_i$ is a coreset for $P' = \bigcup_i P_i$

■

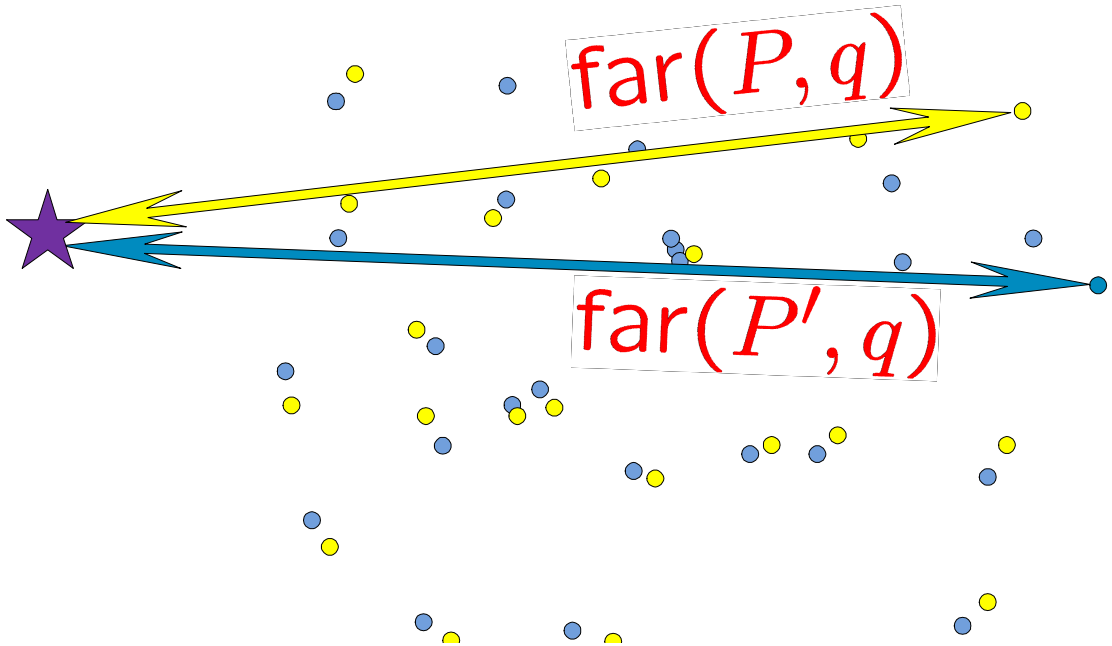
Claim: P' is a coreset for P

$q :=$ an arbitrary query point



Claim: P' is a coreset for P

$q :=$ an arbitrary query point



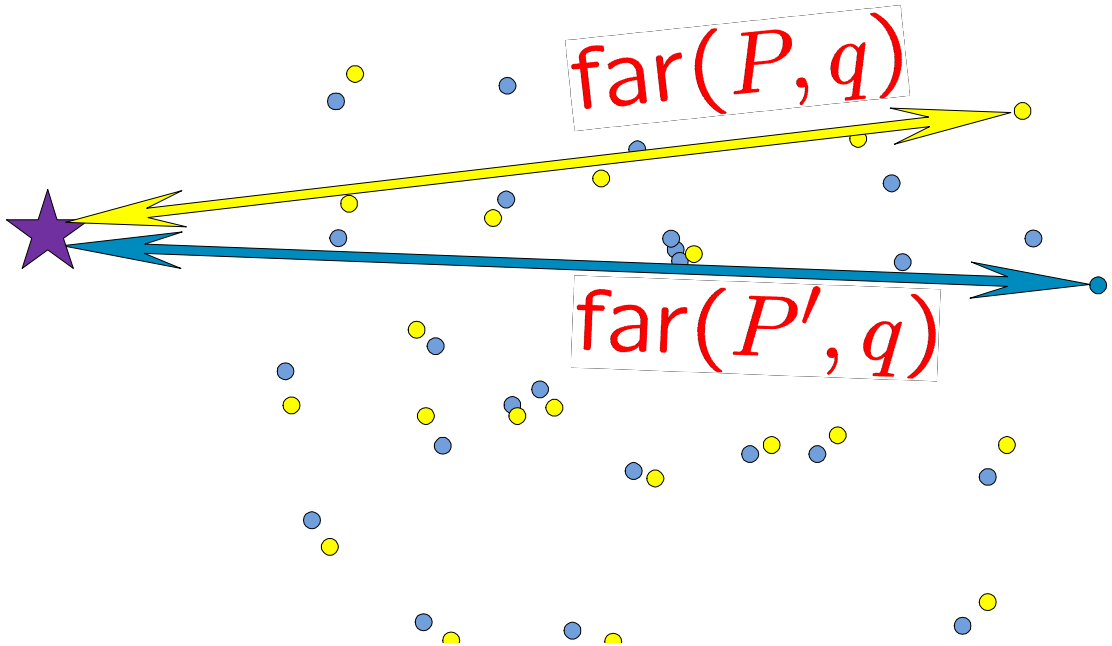
$P :=$

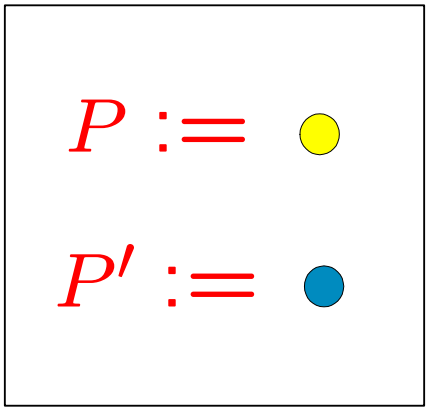
$P' :=$

Claim: P' is a coreset for P

$q :=$ an arbitrary query point

~~$P' \subseteq P \Rightarrow \mathrm{far}(P', q) \leq \mathrm{far}(P, q)$~~

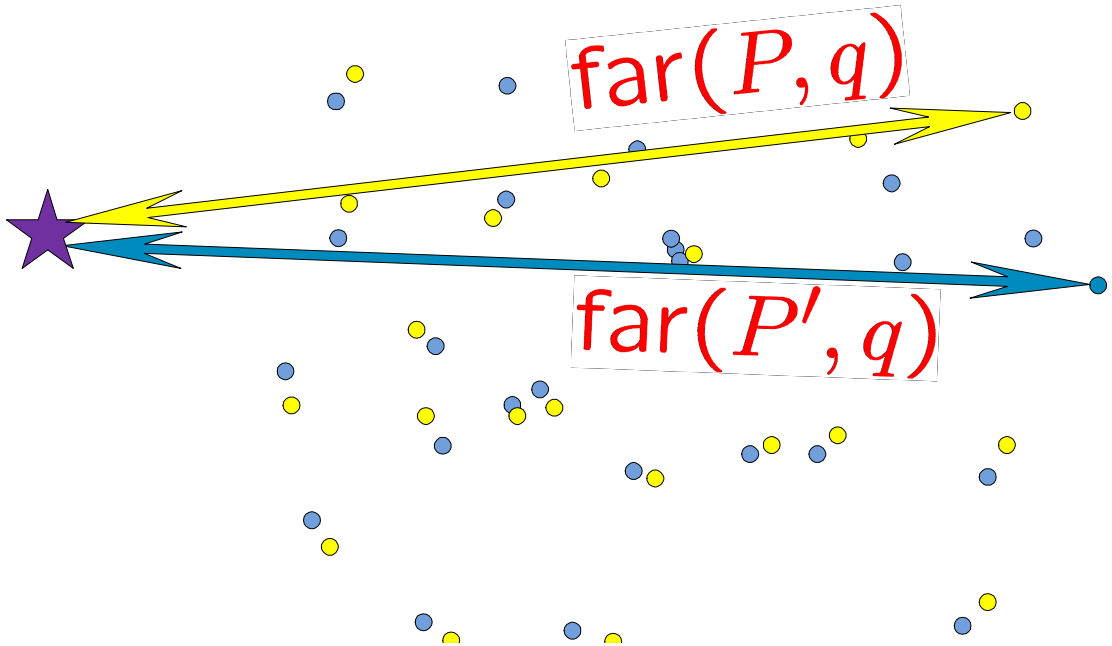




Claim: P' is a coreset for P

Need to prove:

$$\mathsf{far}(P,q) - \mathsf{far}(P',q) \leq \varepsilon \mathsf{far}(P,q)$$



$P :=$

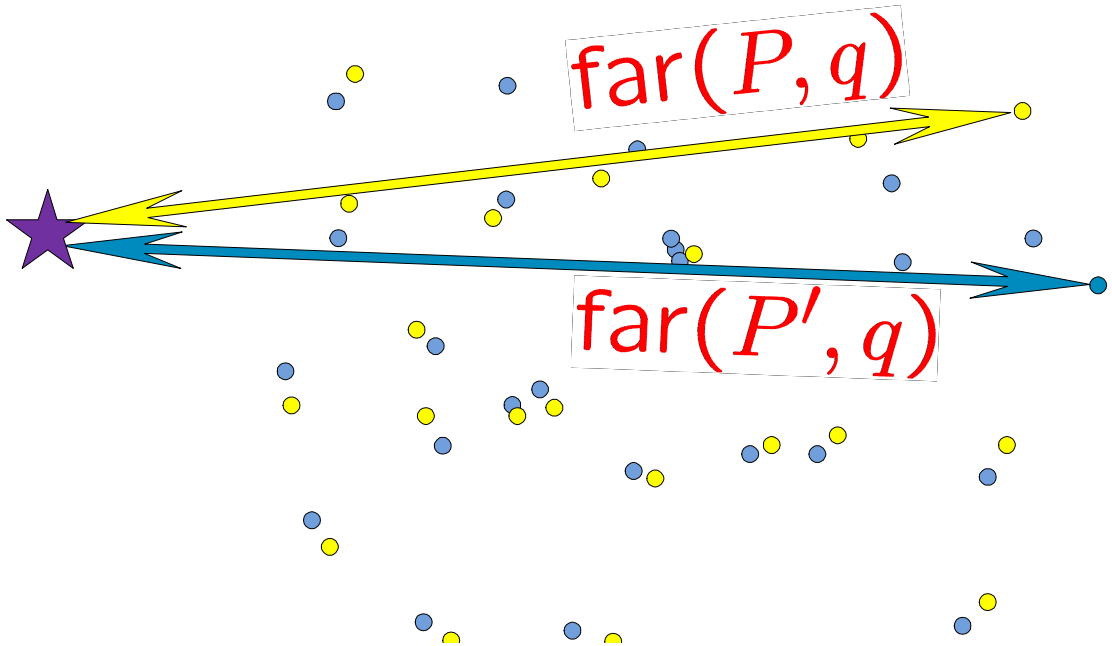
$P' :=$

Claim: P' is a coreset for P

Need to prove:

$$\mathsf{far}(P,q) - \mathsf{far}(P',q) \leq \varepsilon \mathsf{far}(P,q)$$

$$\mathsf{far}(P',q) - \mathsf{far}(P,q) \leq \varepsilon \mathsf{far}(P,q)$$



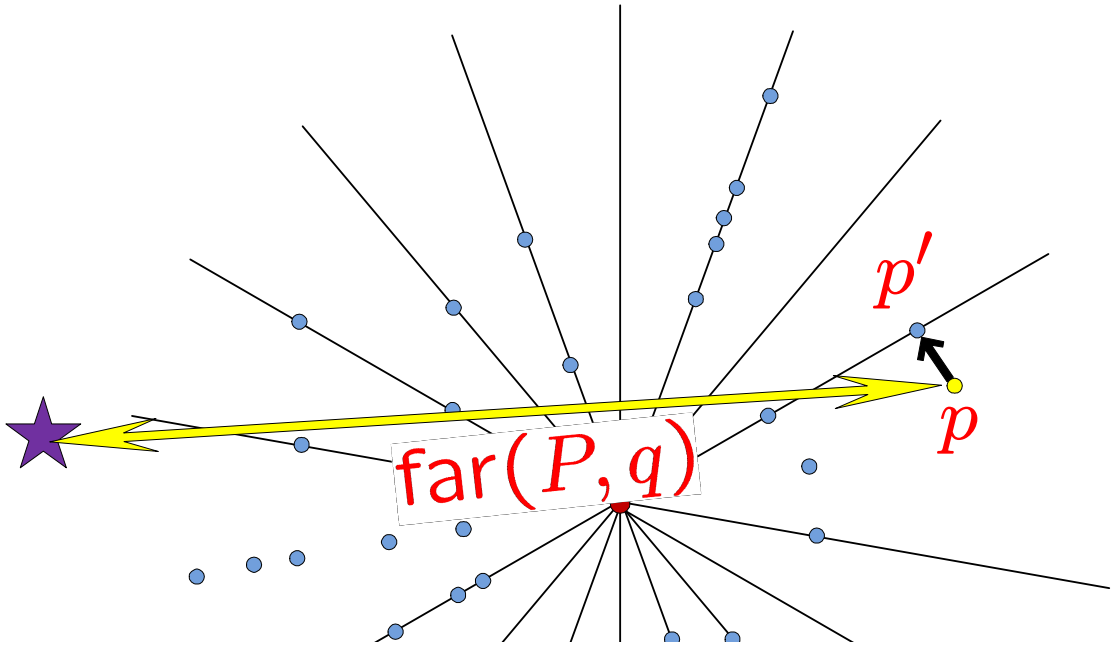
$P :=$ ●

$P' :=$ ●

Claim: P' is a coreset for P

Let $\mathsf{far}(P, q) = \mathsf{dist}(p, q)$

$p' :=$ the projection of p on the "star"



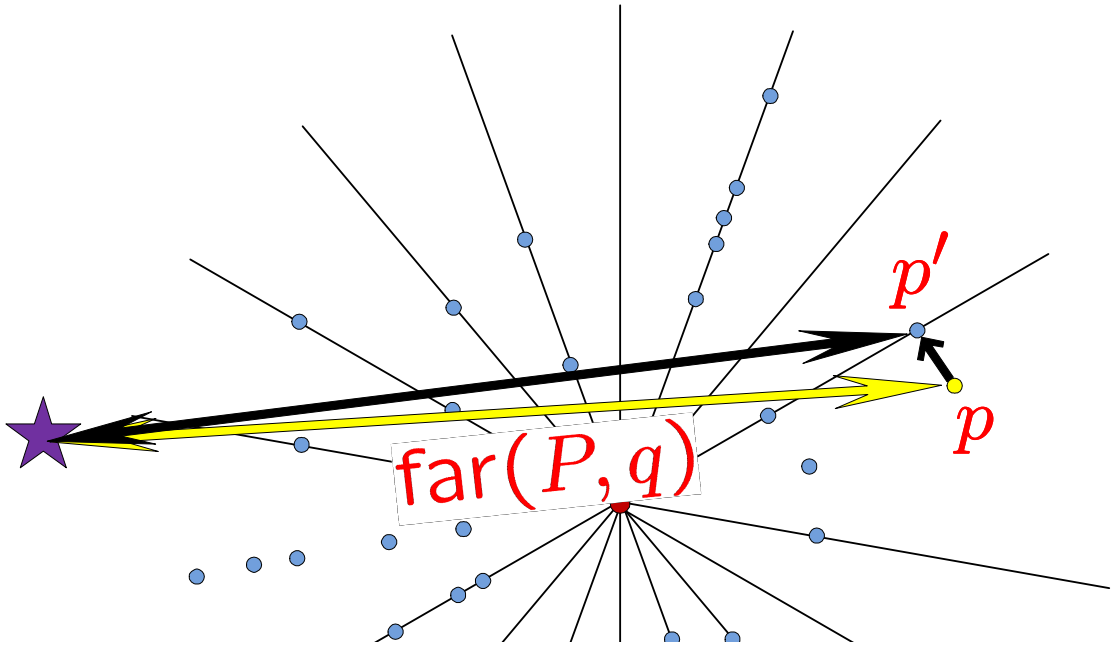
Claim: P' is a coreset for P

Let $\mathrm{far}(P, q) = \mathrm{dist}(p, q)$

$p' :=$ the projection of p on the "star"

⇩

$\mathrm{far}(P, q) - \mathrm{far}(P', q) \leq \mathrm{far}(P, q) - \mathrm{dist}(p', q)$



Claim: P' is a coreset for P

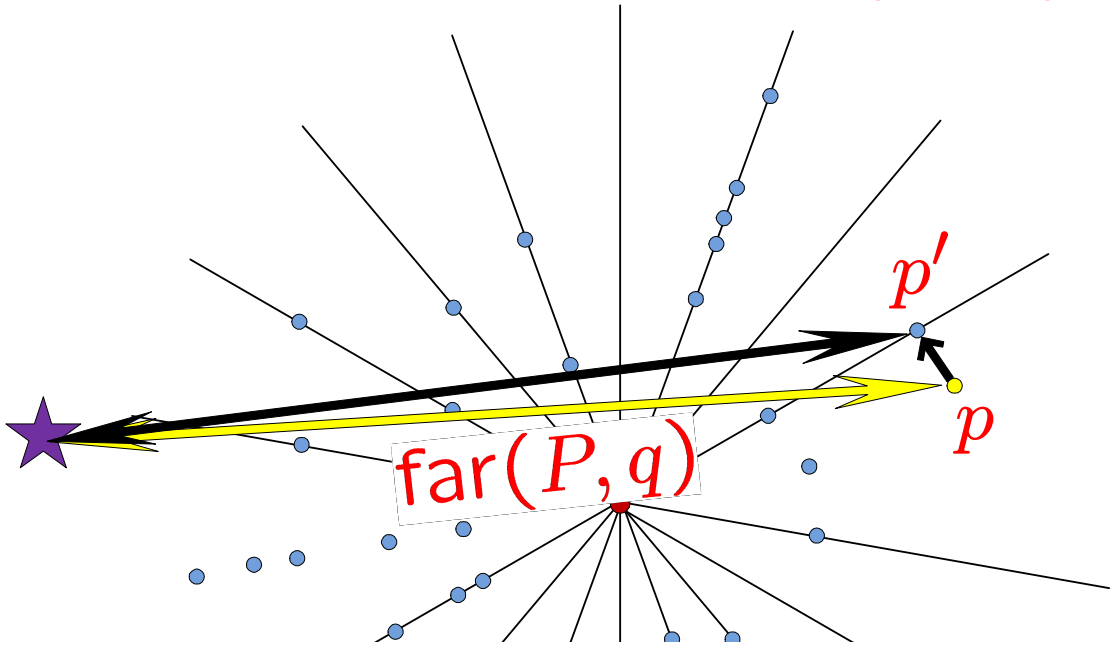
Let $\mathrm{far}(P, q) = \mathrm{dist}(p, q)$

$p' :=$ the projection of p on the "star"

⇩

$$\mathrm{far}(P, q) - \mathrm{far}(P', q) \leq \mathrm{far}(P, q) - \mathrm{dist}(p', q)$$

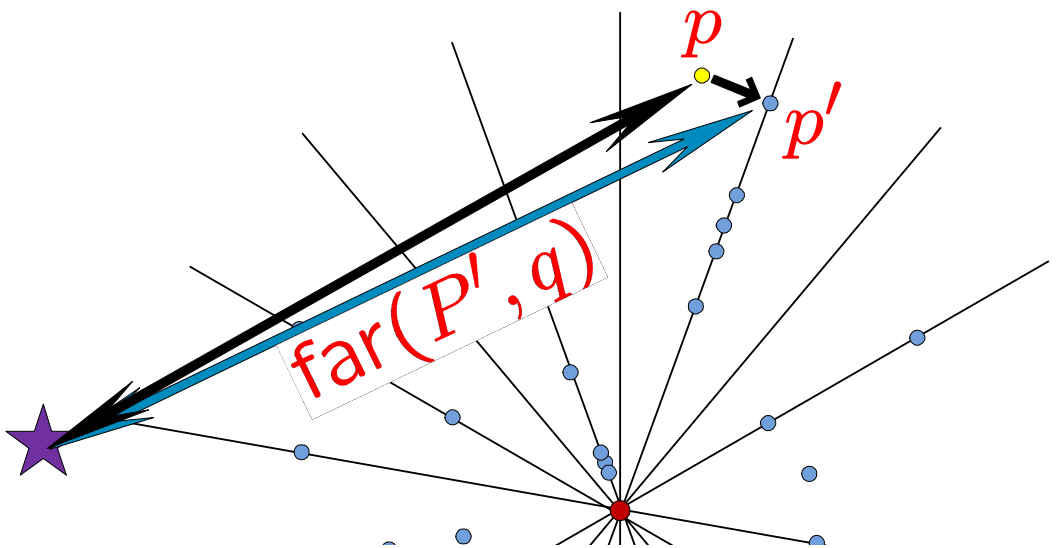
$$\leq \mathrm{dist}(p, p')$$



Claim: P' is a coreset for P

Let $\mathrm{far}(P', q) = \mathrm{dist}(p', q)$

$p :=$ the point whose projection is p'

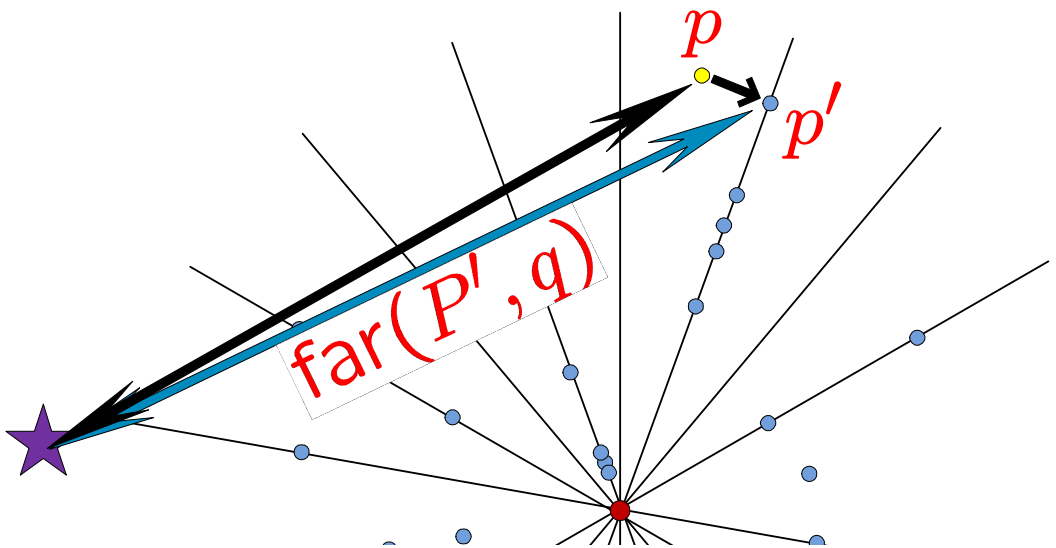


Claim: P' is a coreset for P

Let $\mathrm{far}(P', q) = \mathrm{dist}(p', q)$

$p :=$ the point whose projection is p'

$\mathrm{far}(P', q) - \mathrm{far}(P, q) \leq \mathrm{far}(P', q) - \mathrm{dist}(p, q)$



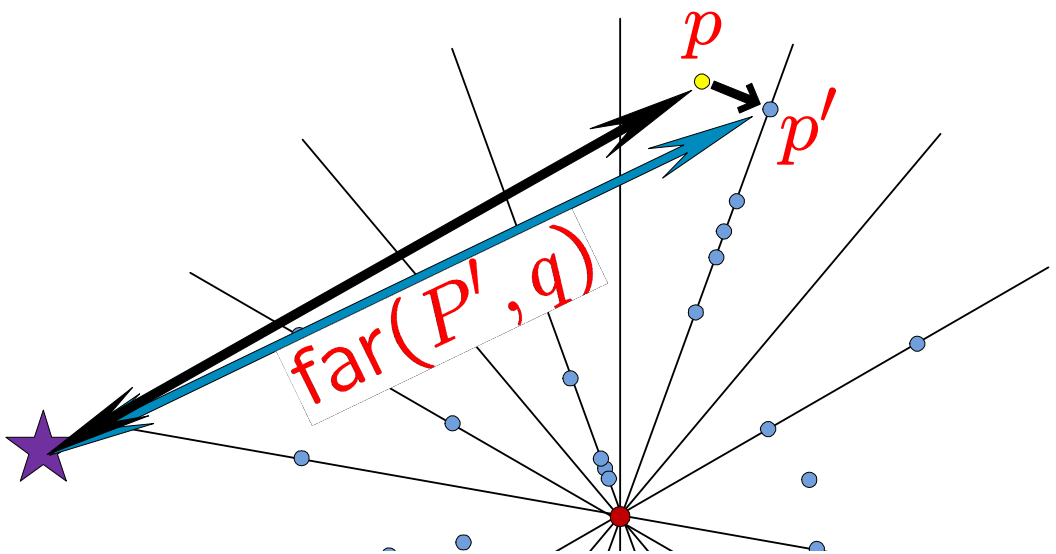
Claim: P' is a coreset for P

Let $\mathrm{far}(P', q) = \mathrm{dist}(p', q)$

$p :=$ the point whose projection is p'

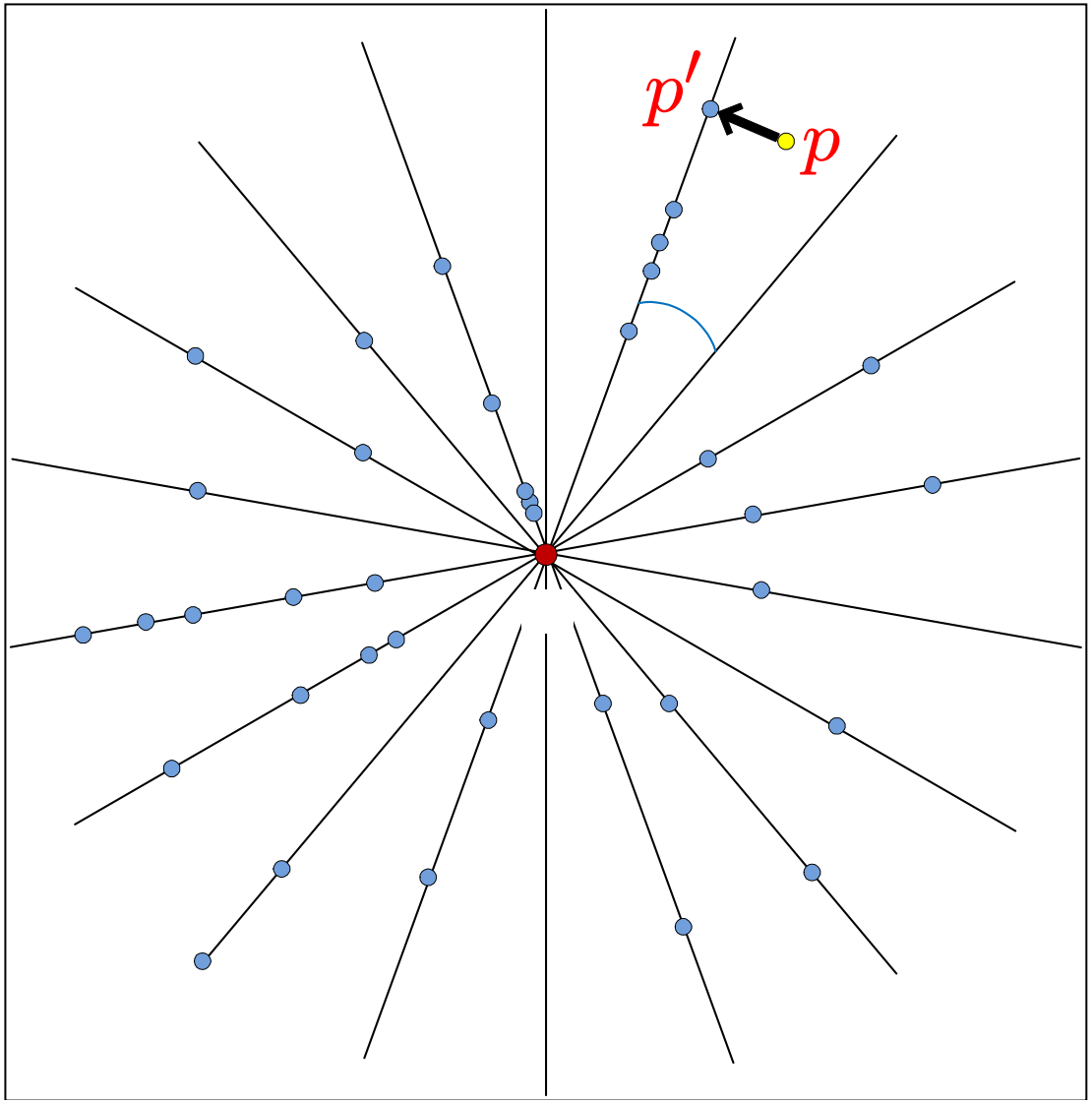
$$\mathrm{far}(P', q) - \mathrm{far}(P, q) \leq \mathrm{far}(P', q) - \mathrm{dist}(p, q)$$

$$\leq \mathrm{dist}(p, p')$$



Bounding $\mathrm{dist}(p, p')$

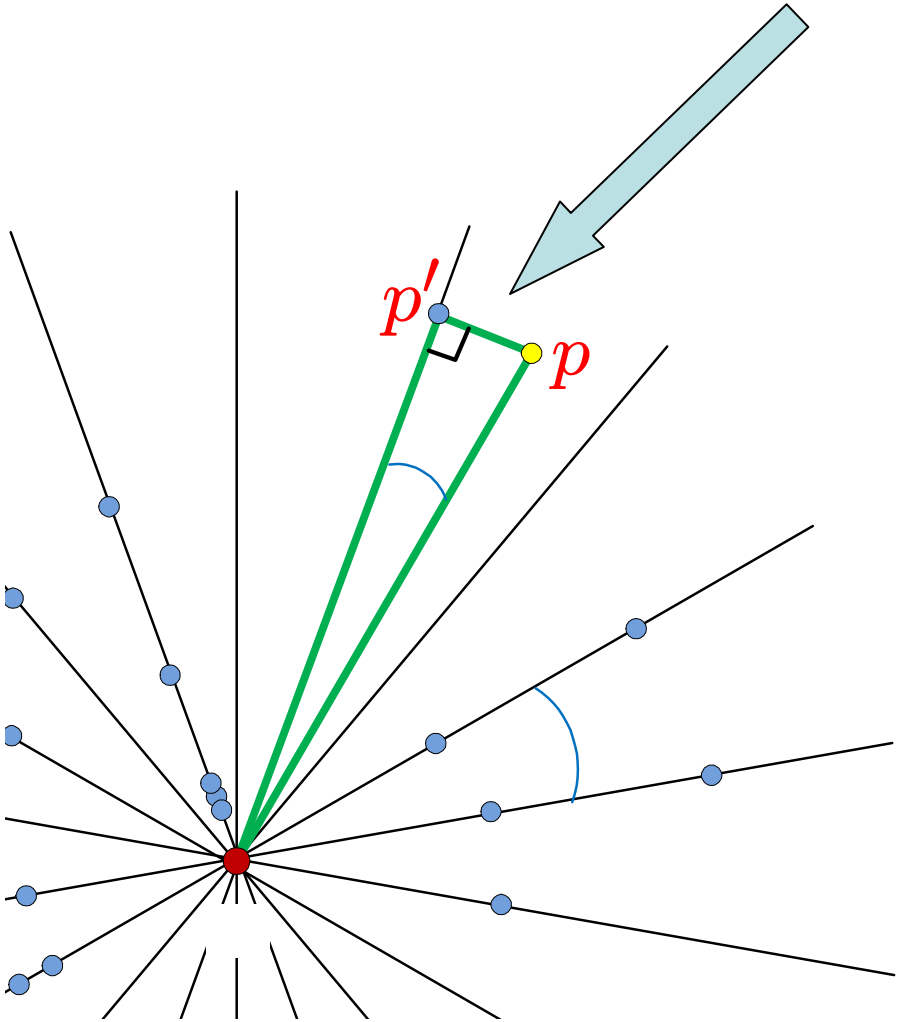
$p' :=$ the projection of $p \in P$



Bounding $\mathrm{dist}(p, p')$

$p' :=$ the projection of $p \in P$

$$\mathrm{dist}(p, p') = \sin\alpha \cdot \mathrm{dist}(u, p)$$

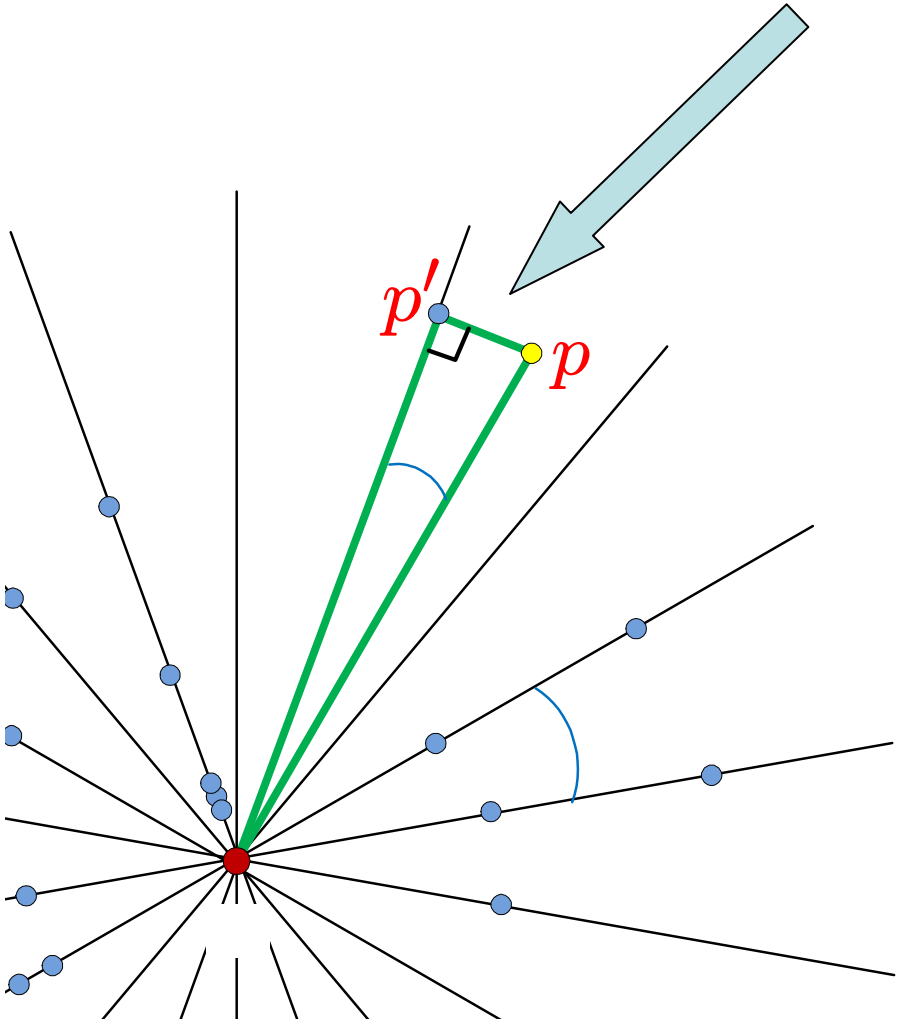


Bounding $\mathrm{dist}(p, p')$

$p' :=$ the projection of $p \in P$

$$\mathrm{dist}(p, p') = \sin\alpha \cdot \mathrm{dist}(u, p)$$

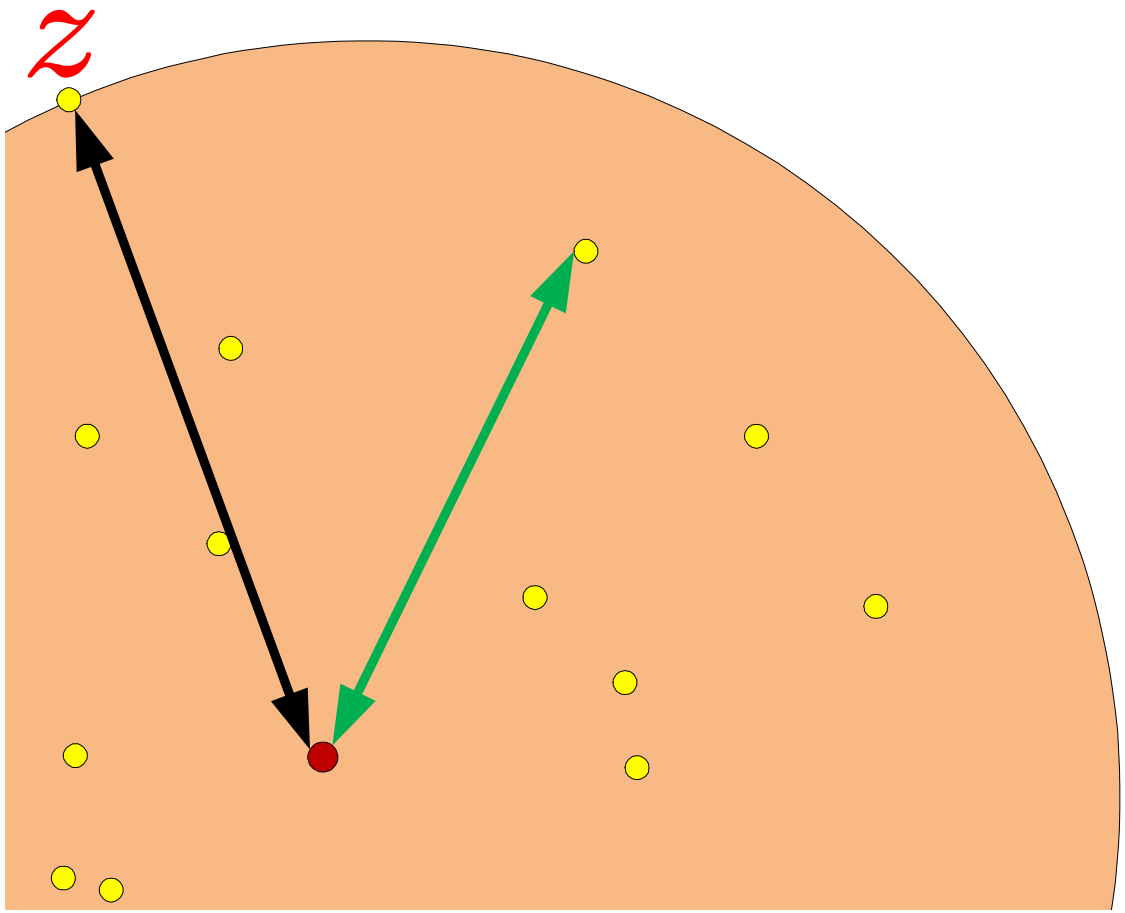
$$\leq O(\varepsilon) \cdot \mathrm{dist}(u, p)$$



Bounding $\mathsf{dist}(p, p')$

$p' :=$ the projection of $p \in P$

$$\mathsf{dist}(p, p') = \sin\alpha \cdot \mathsf{dist}(u, p)$$
$$\leq O(\varepsilon) \cdot \mathsf{dist}(u, p)$$
$$\leq O(\varepsilon) \cdot r$$

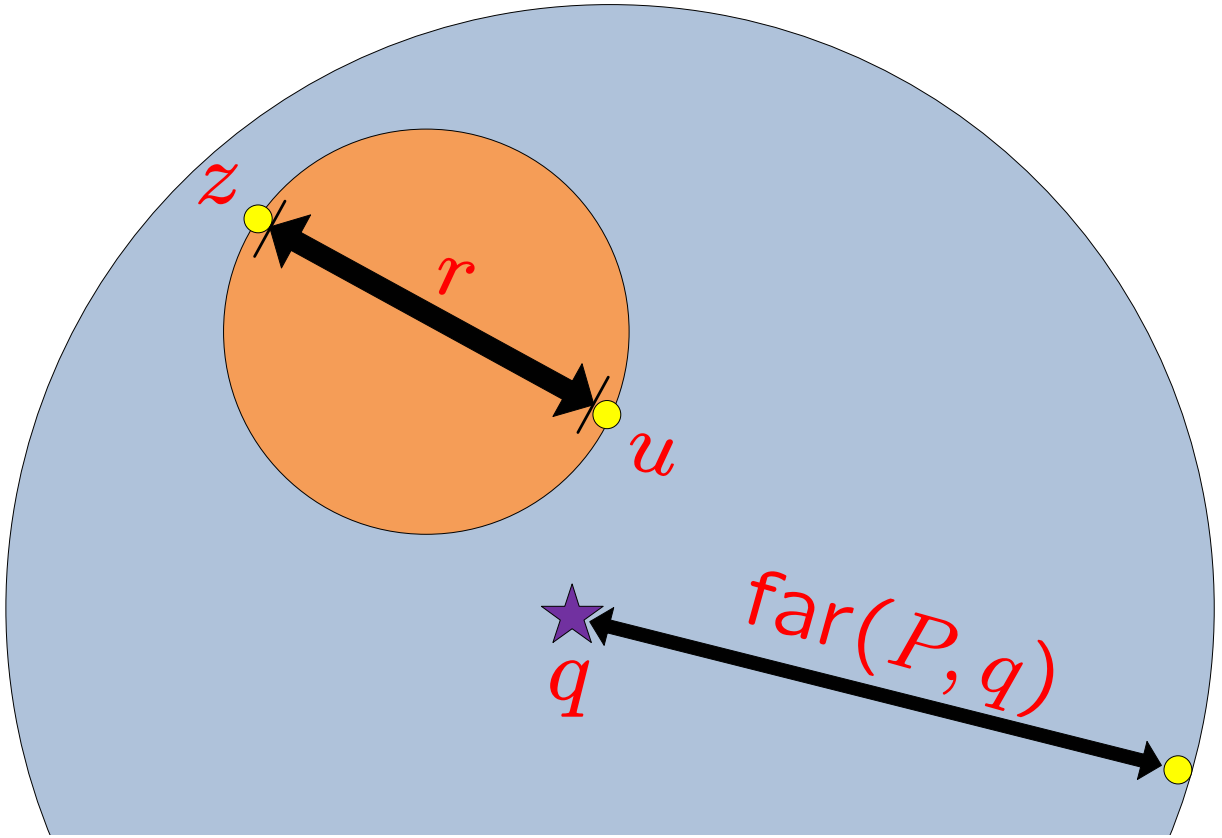


$$r := \max_{p \in P} \mathsf{dist}(u, p)$$

Main Observation

Every ball that covers u and z,

has a diameter of at least r:

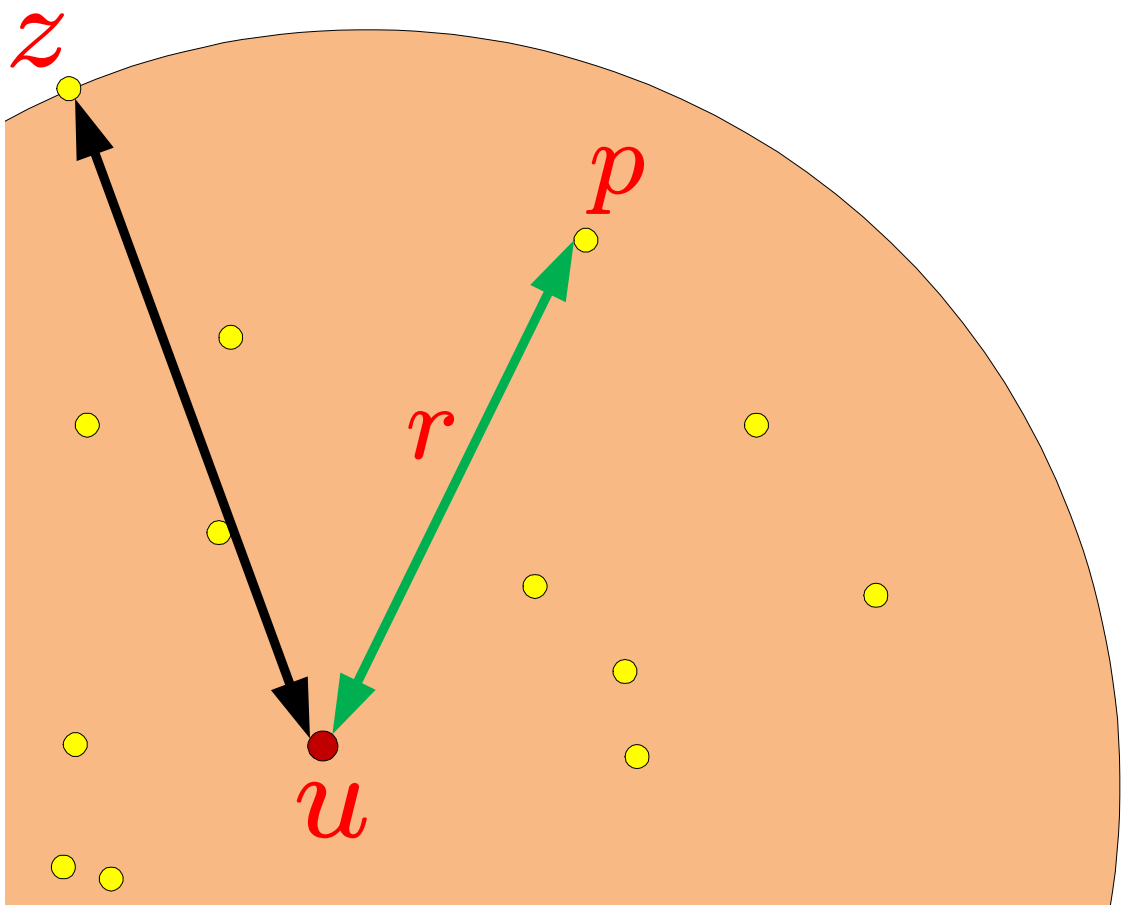


$$r \leq 2\mathsf{far}(P,q) \quad \Rightarrow \quad O(\varepsilon r) \leq O(\varepsilon)\mathsf{far}(P,q)$$

Bounding $\mathsf{dist}(p, p')$

$p' :=$ the projection of $p \in P$

$$\mathsf{dist}(p, p') = \sin\alpha \cdot \mathsf{dist}(u, p)$$
$$\leq O(\varepsilon) \cdot \mathsf{dist}(u, p)$$
$$\leq O(\varepsilon) \cdot r$$
$$\leq O(\varepsilon) \cdot \mathsf{far}(P, q)$$

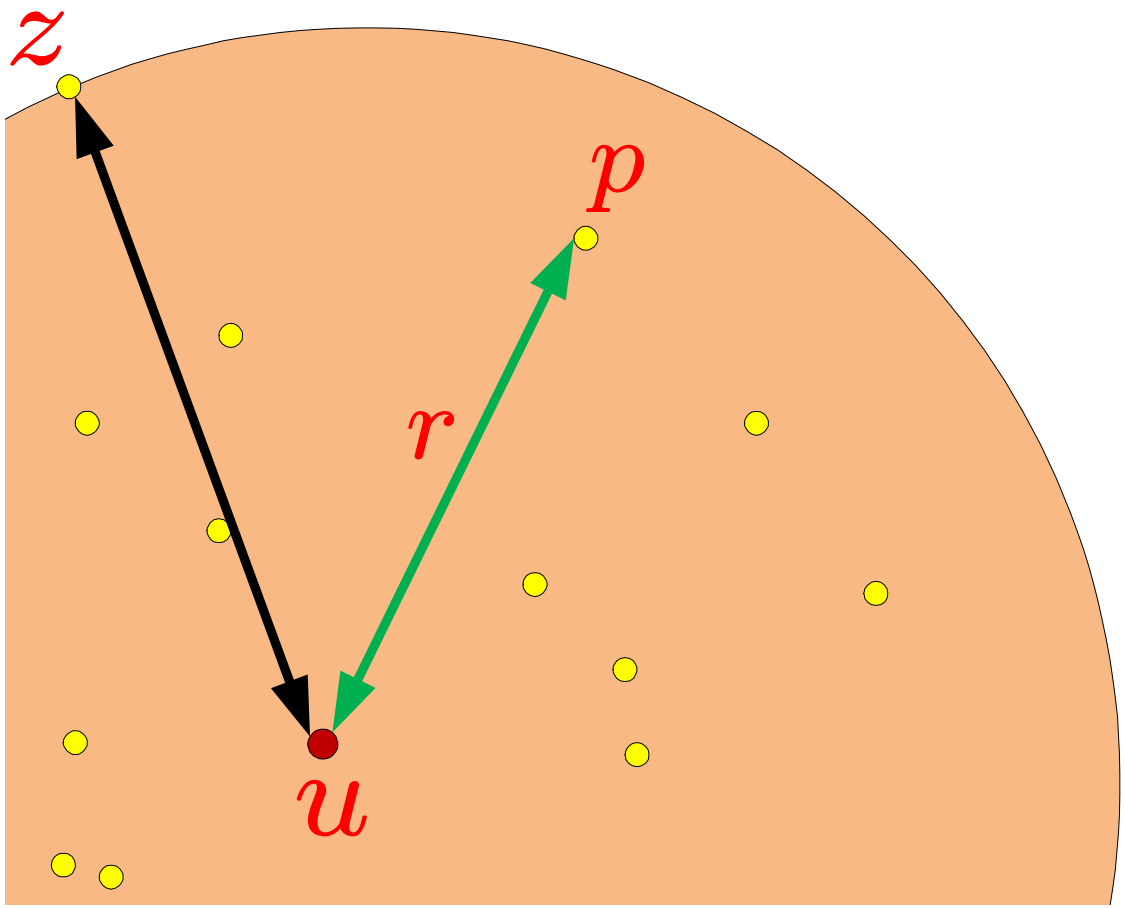


$$r := \max_{p \in P} \mathsf{dist}(u, p)$$

Bounding $\mathsf{dist}(p, p')$

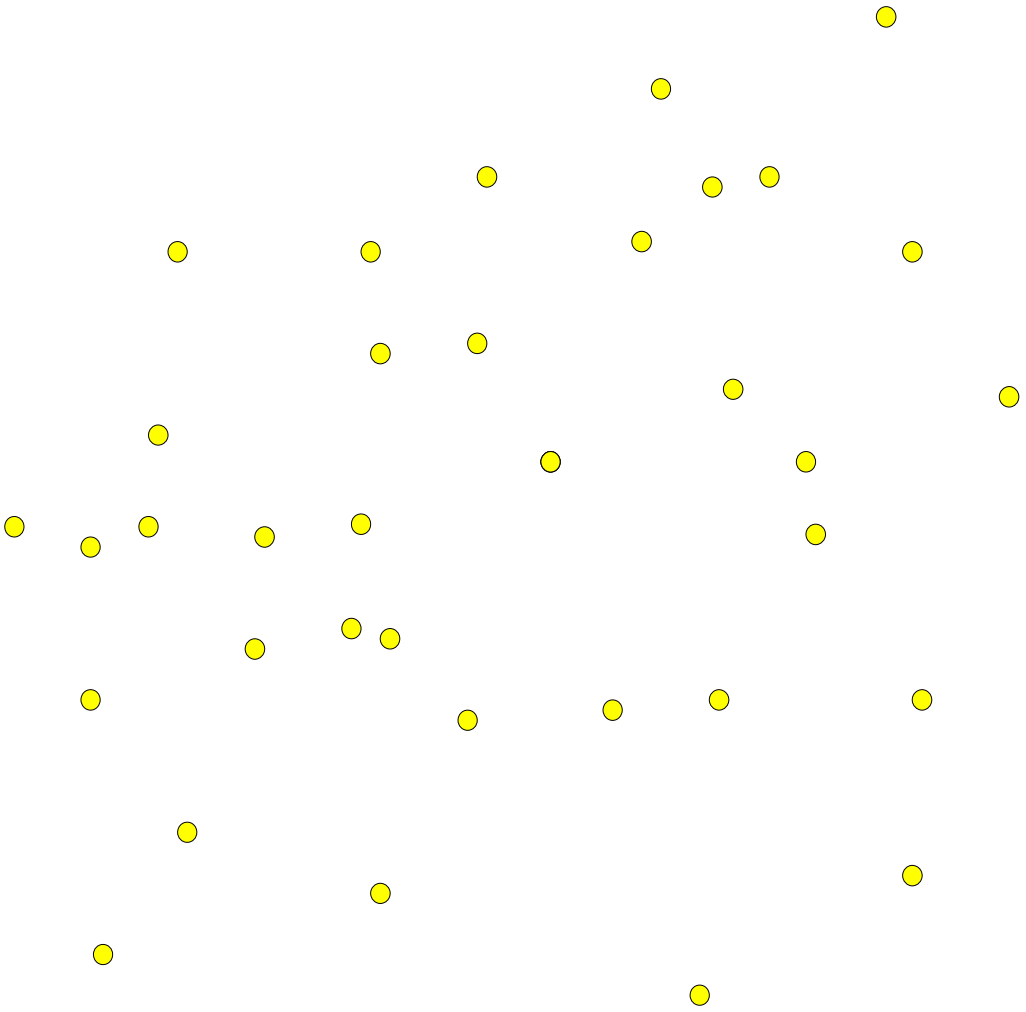
$p' :=$ the projection of $p \in P$

$$\mathsf{dist}(p, p') = \sin\alpha \cdot \mathsf{dist}(u, p)$$
$$\leq O(\varepsilon) \cdot \mathsf{dist}(u, p)$$
$$\leq O(\varepsilon) \cdot r$$
$$\leq O(\varepsilon) \cdot \mathsf{far}(P, q)$$



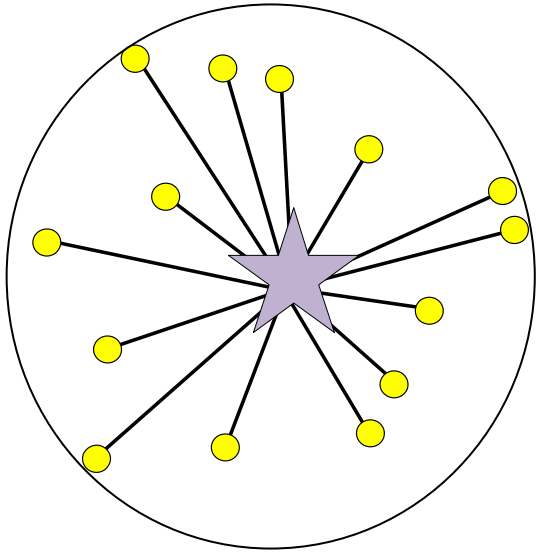
$$r := \max_{p \in P} \mathsf{dist}(u, p)$$

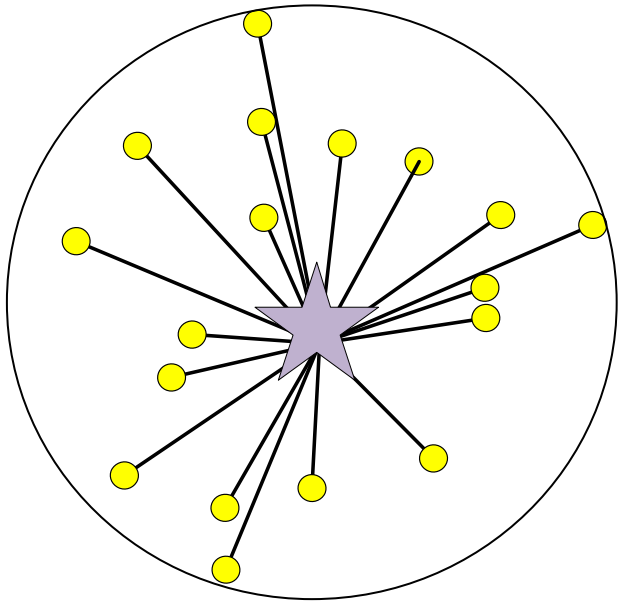
Generalization for k-Center



The k-Center of P

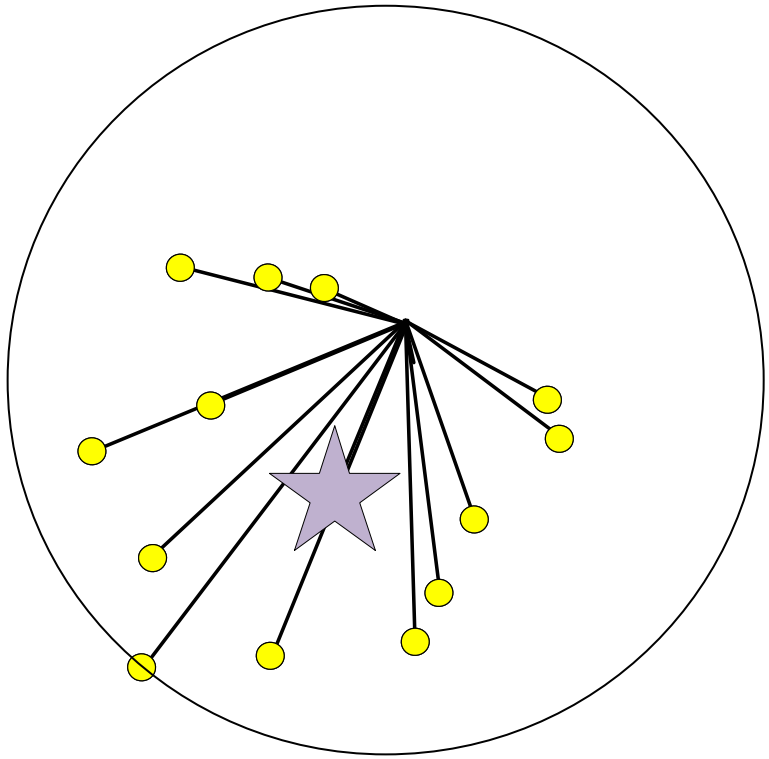
$$\text{opt} = \min_{|\text{OPT}|=k} \max_{p \in P} \text{dist}(p, \text{OPT})$$

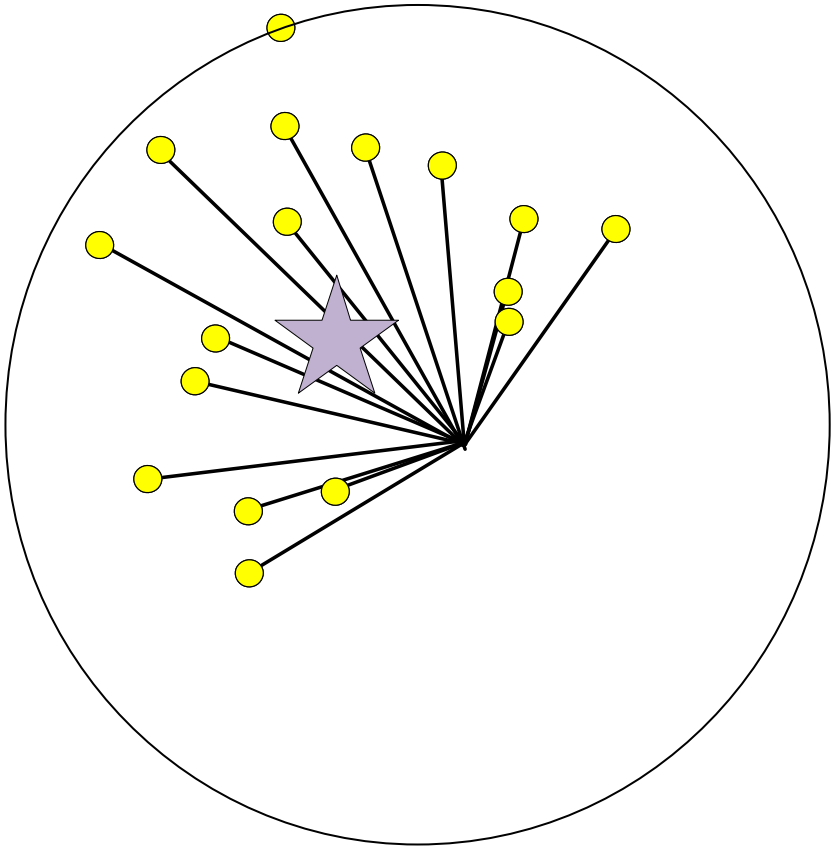




Constant Approximation

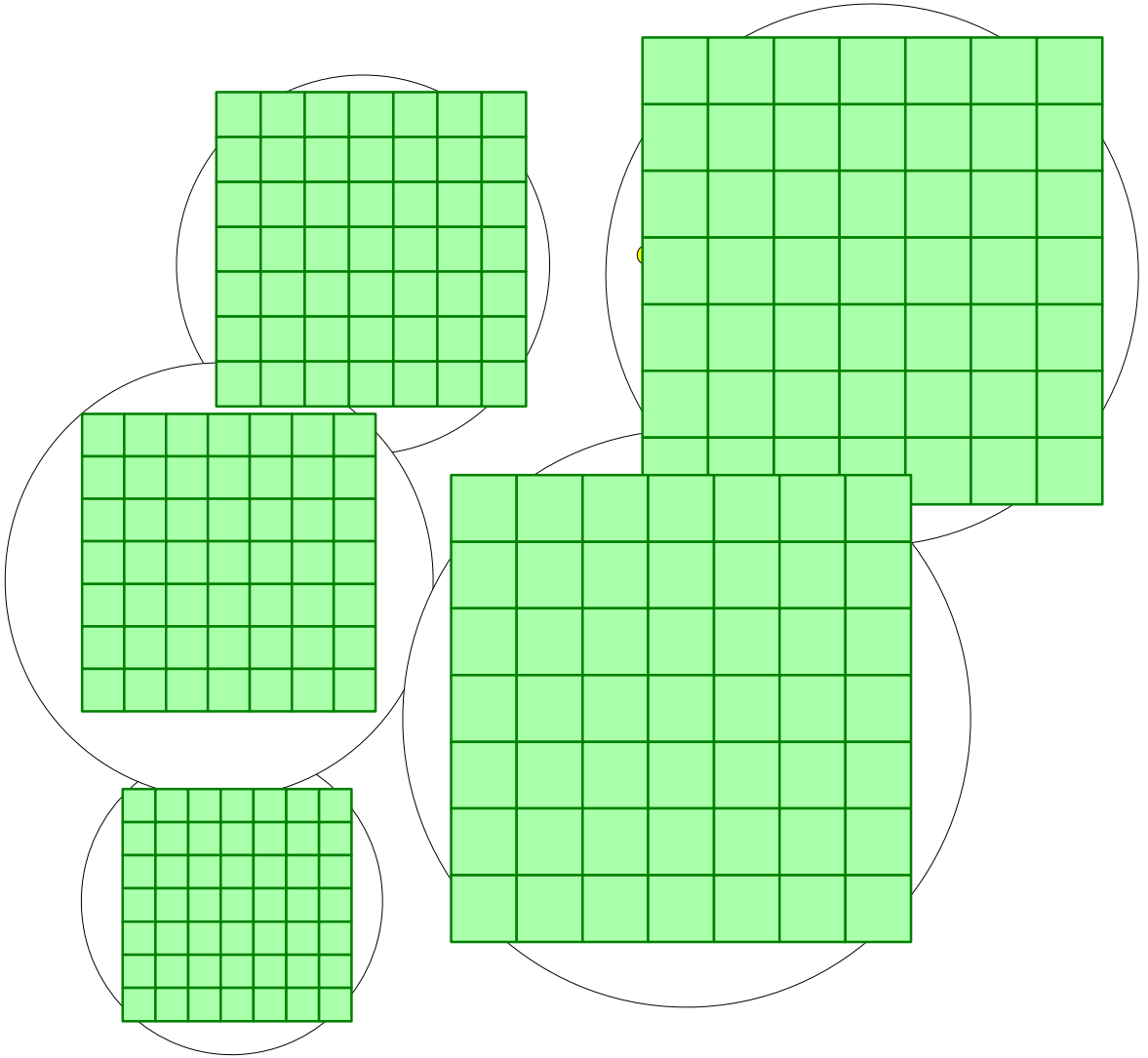
$|\widetilde{\mathrm{OPT}}| = k$ $\qquad$ $\max_{p \in P} \mathrm{dist}(p, \widetilde{\mathrm{OPT}}) \leq O(1) \cdot \mathrm{opt}$





For each cluster:

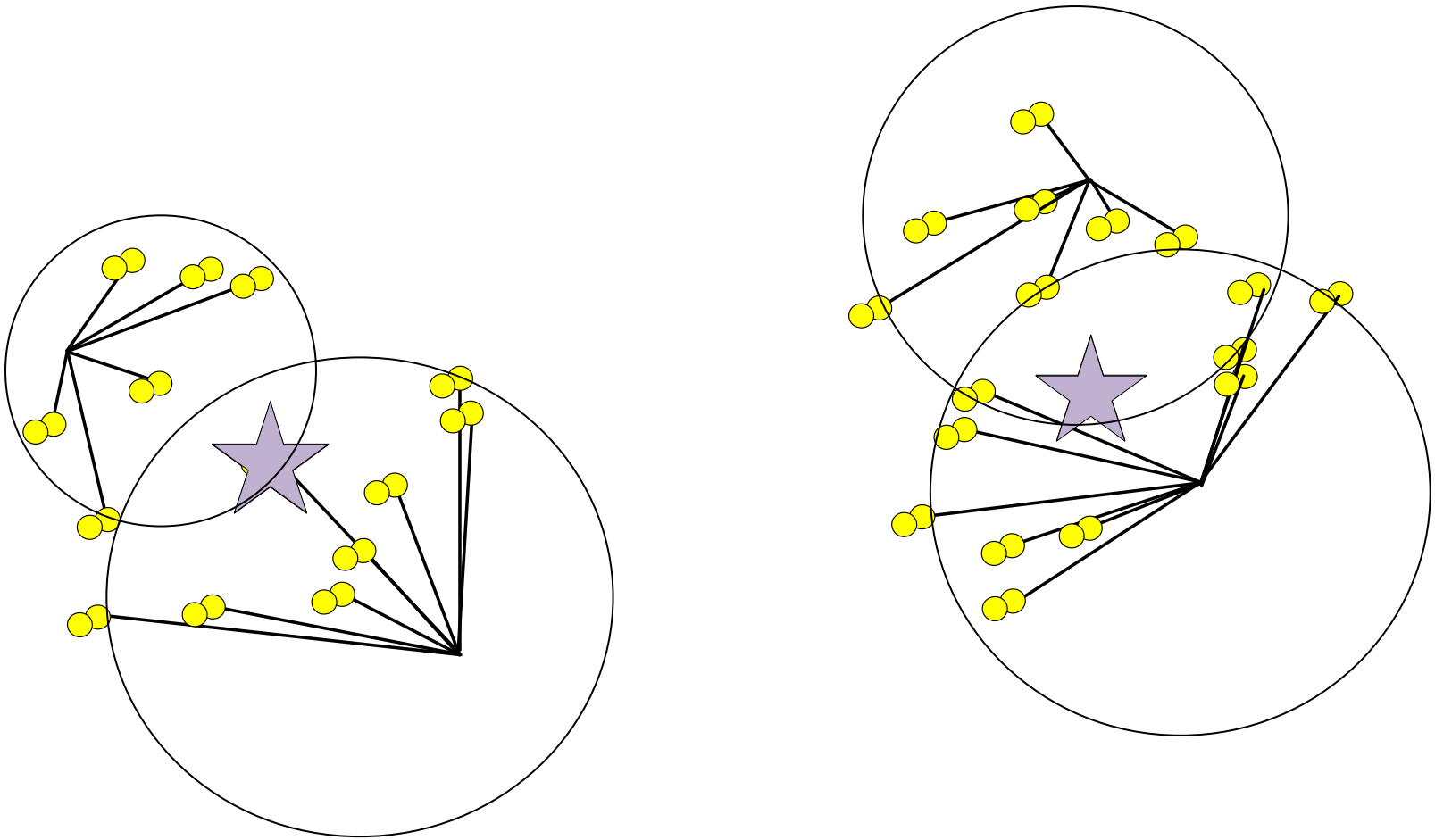
Apply construction for $k = 1$



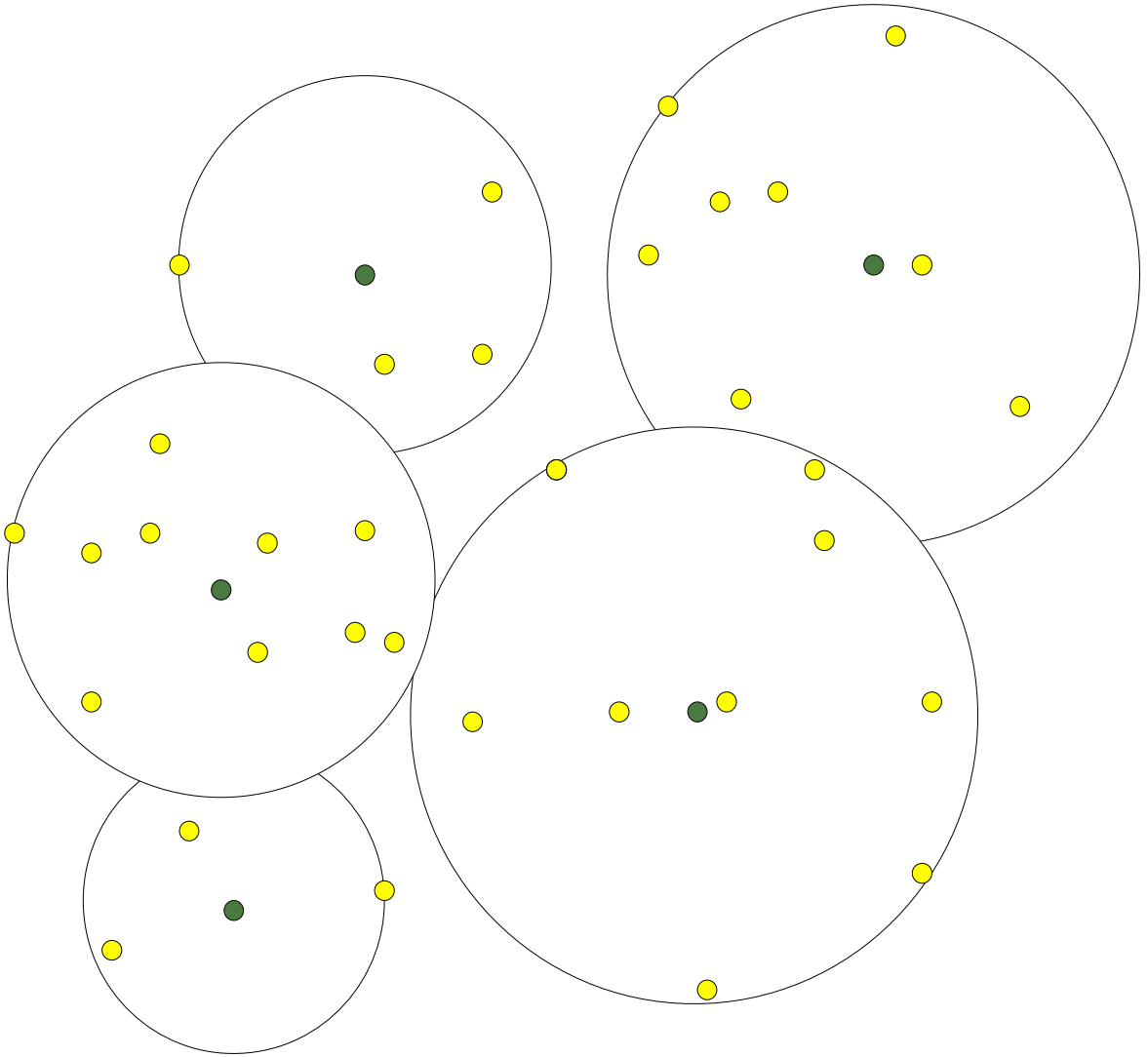
Bi-criteria Approximation

$|B| = O(k \log n)$

$\max_{p \in P} \mathrm{dist}(p, B) \leq O(1) \cdot \mathrm{opt}$

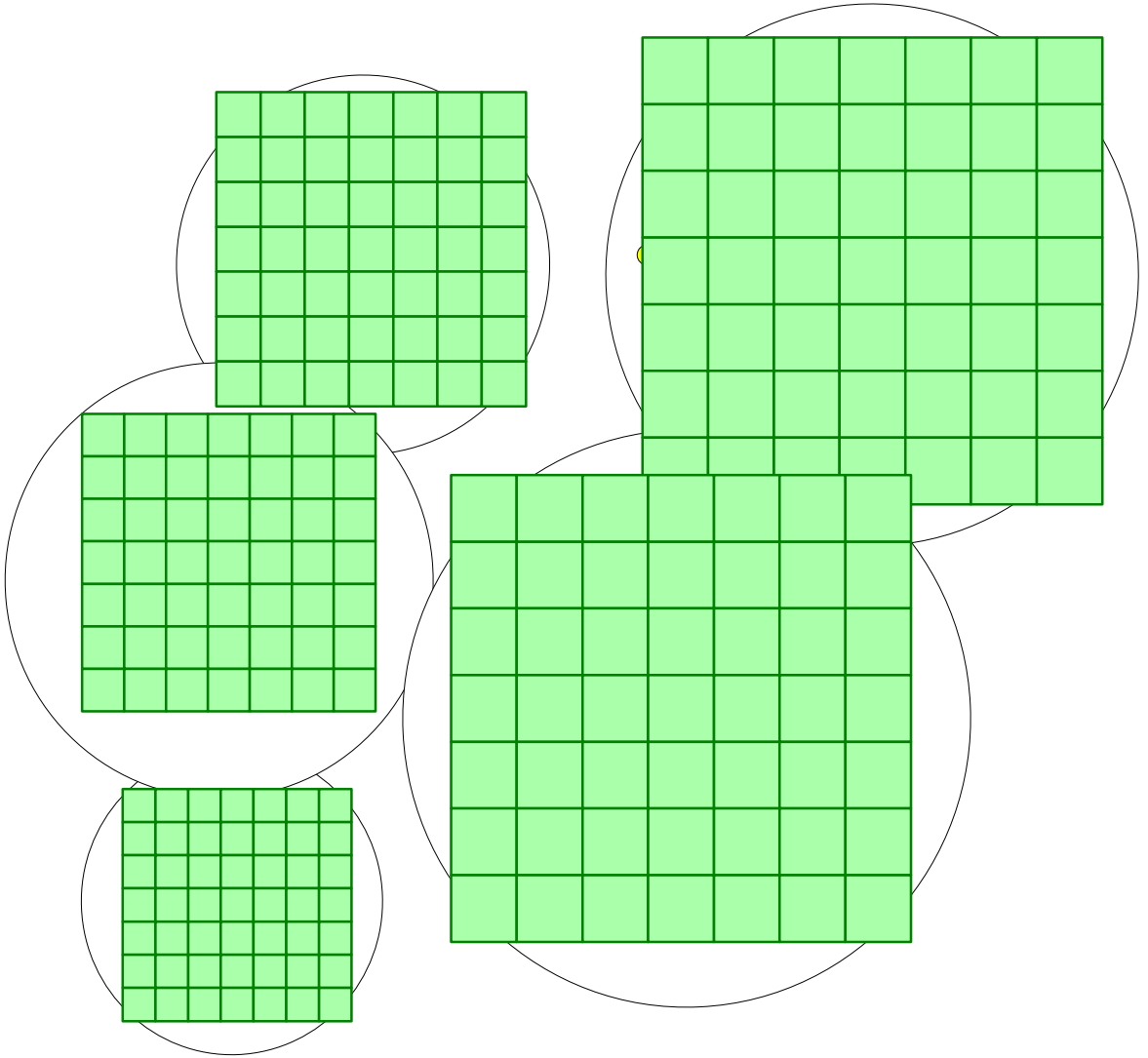


Compute Bi-criteria Approximation
Using ε-Approximations



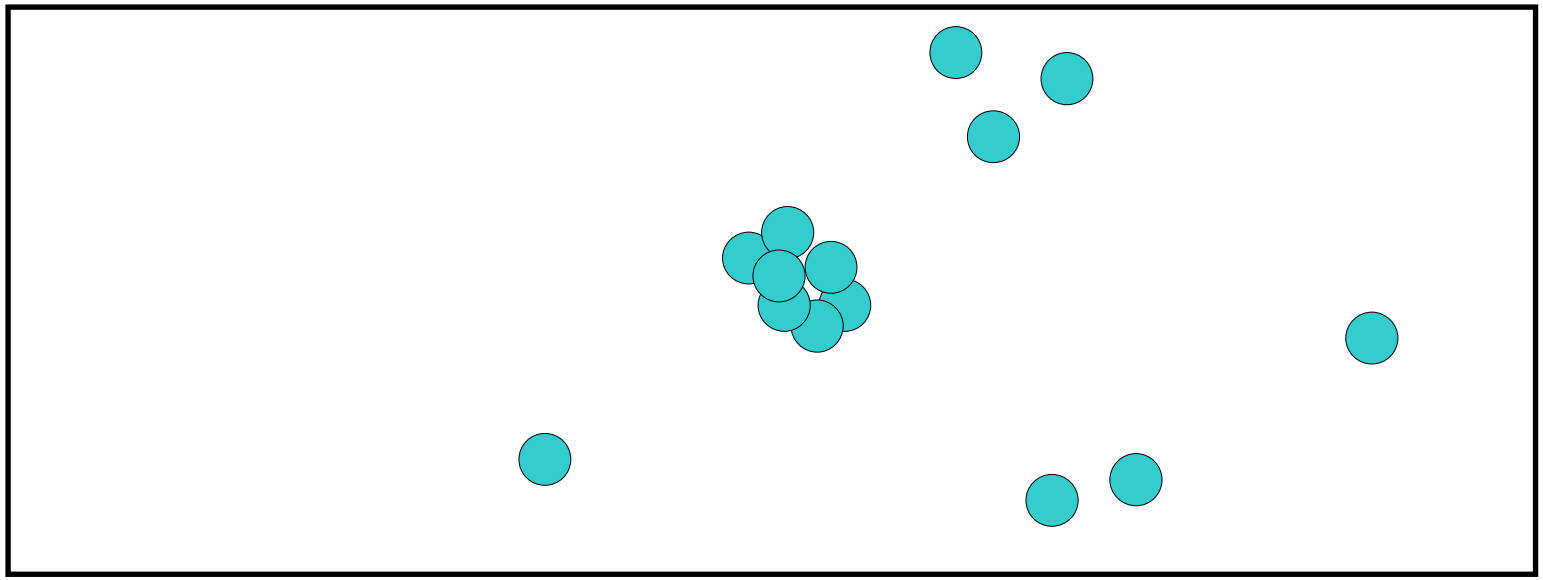
For each cluster:

Apply construction for $k = 1$



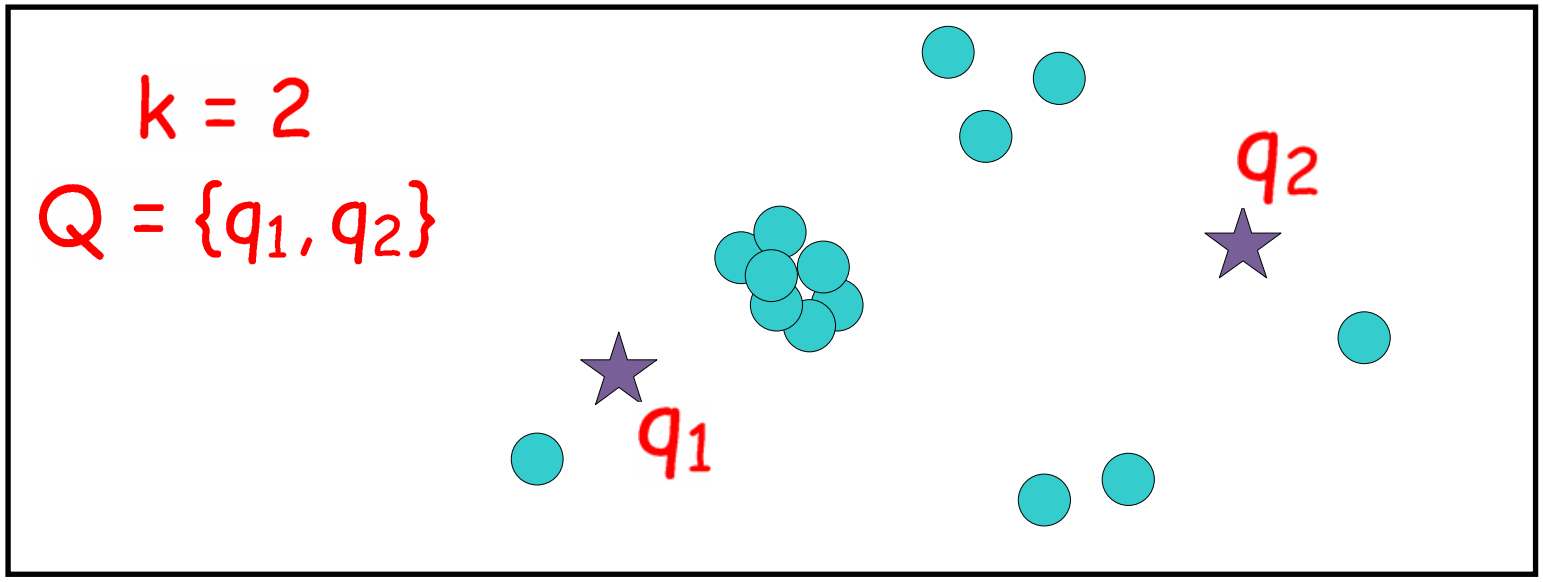
k-Median Queries

- Input: $P \subseteq \mathbb{R}^d$



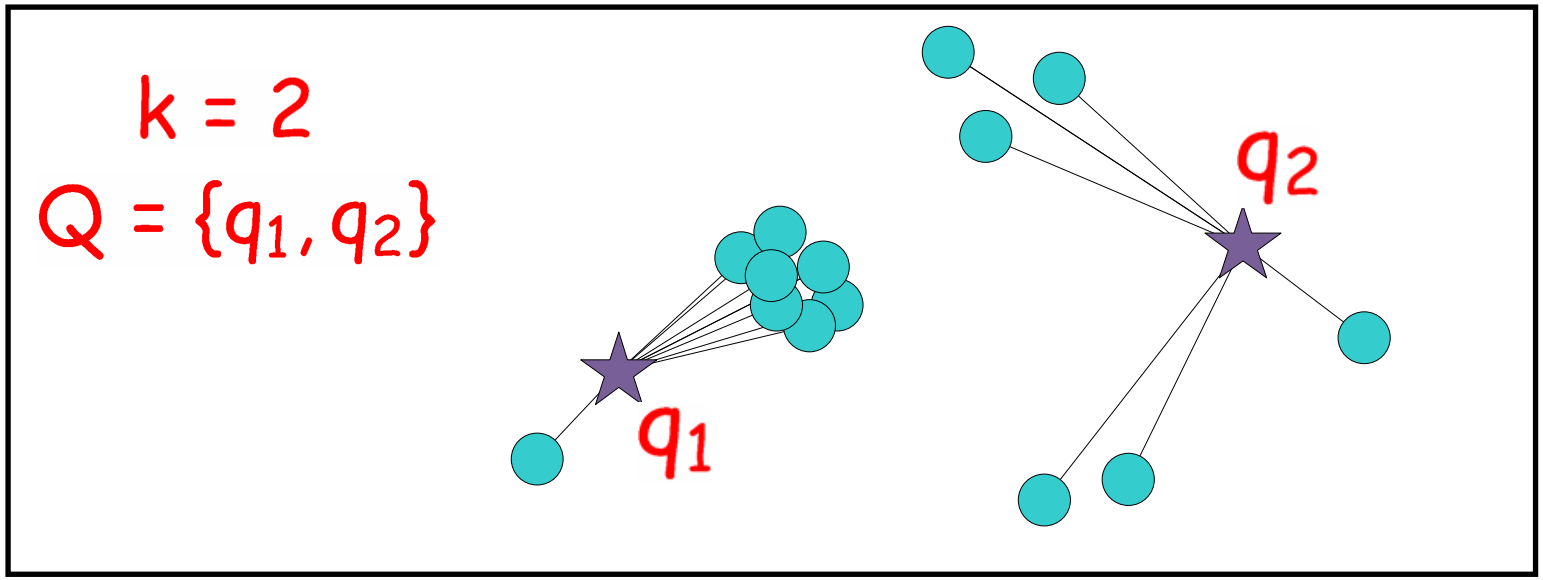
k-Median Queries

- Input: $P \subseteq \mathbb{R}^d$
- Query: A set Q of k points



k-Median Queries

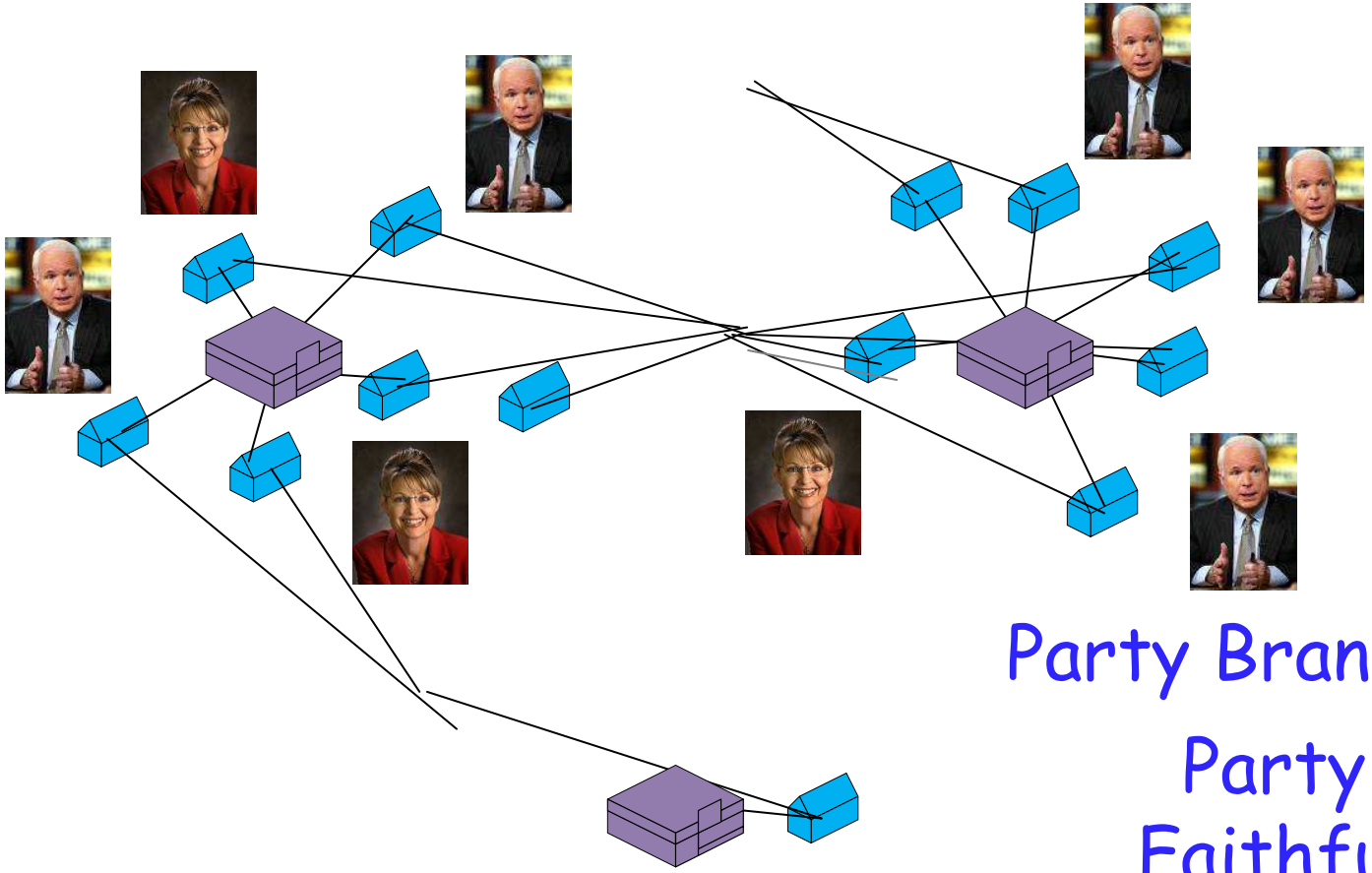
- Input: $P \subseteq \mathbb{R}^d$
- Query: A set Q of k points
- Output: $\sum_{p \in P} \text{dist}(p, Q) = \sum_{p \in P} \min_{q \in Q} \|p - q\|$



Motivation

Comparing alternatives:

How good is this placement?



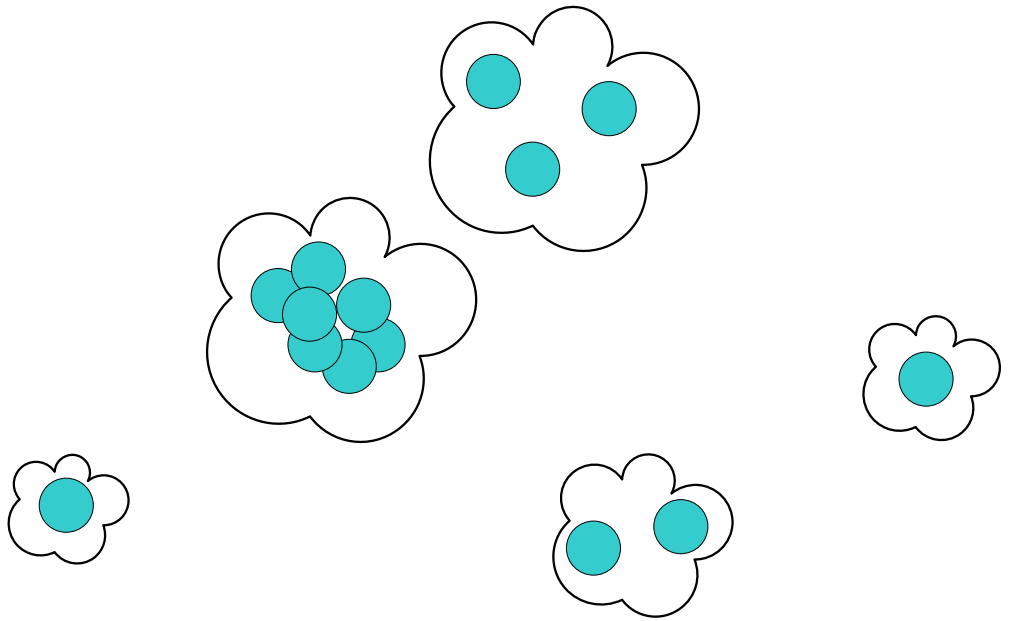
Party Branch

Party Faithful

(k, ε)-Median Coreset

Answer k-median queries in sub-linear time

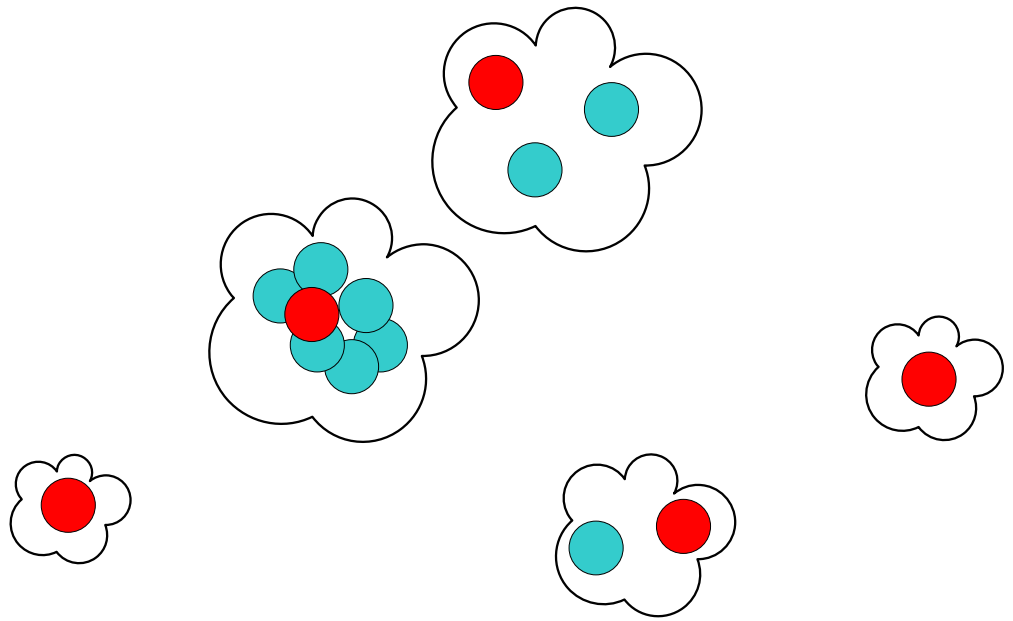
Key Idea: Replace many points by one weighted representative:



(k, ε)-Median Coreset

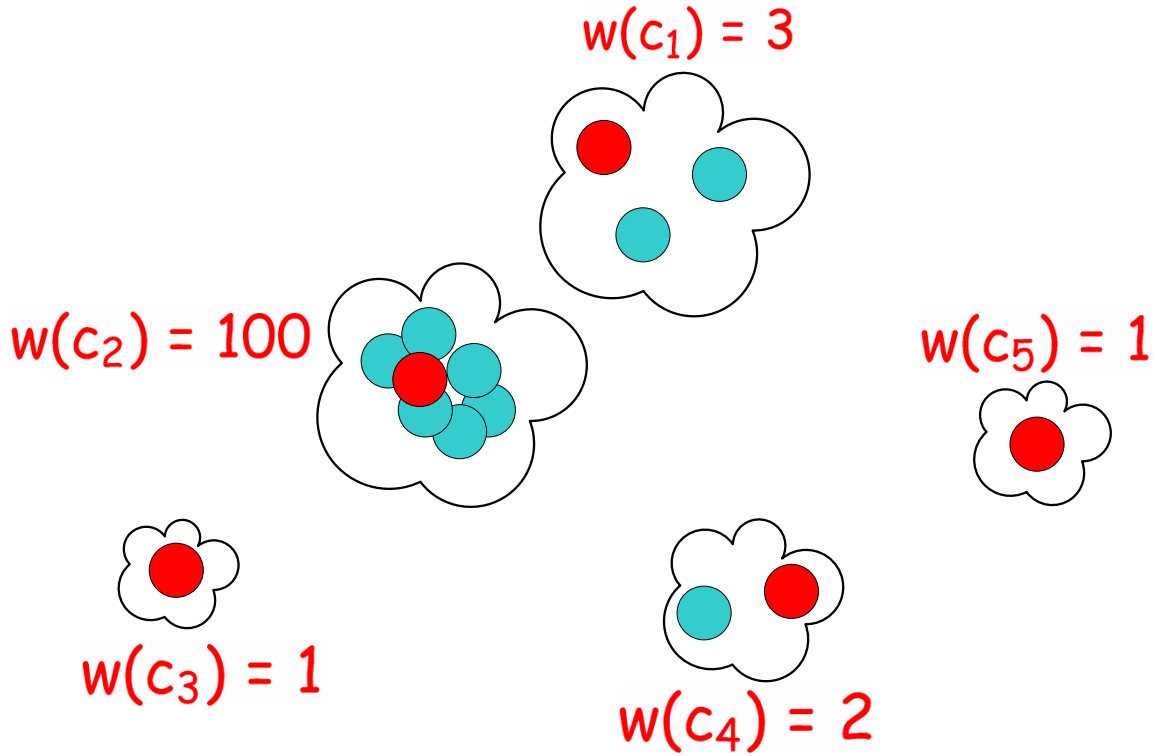
Answer k-median queries in sub-linear time

Key Idea: Replace many points by one weighted representative:



(k, ε)-Median Coreset

Key Idea: Replace many points by one weighted representative:

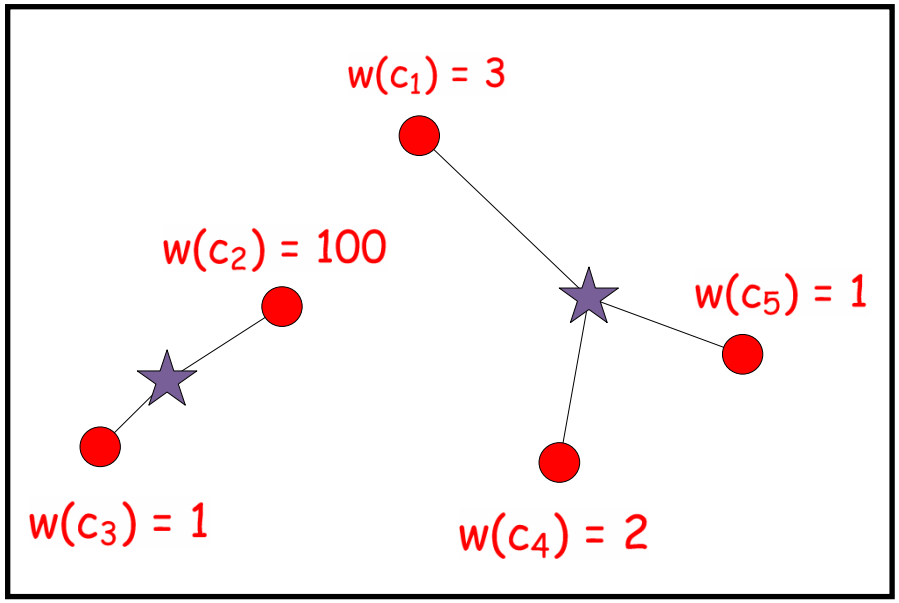


Definition

C is a (k, ε)-coreset for P, if $\forall Q, |Q| = k$:

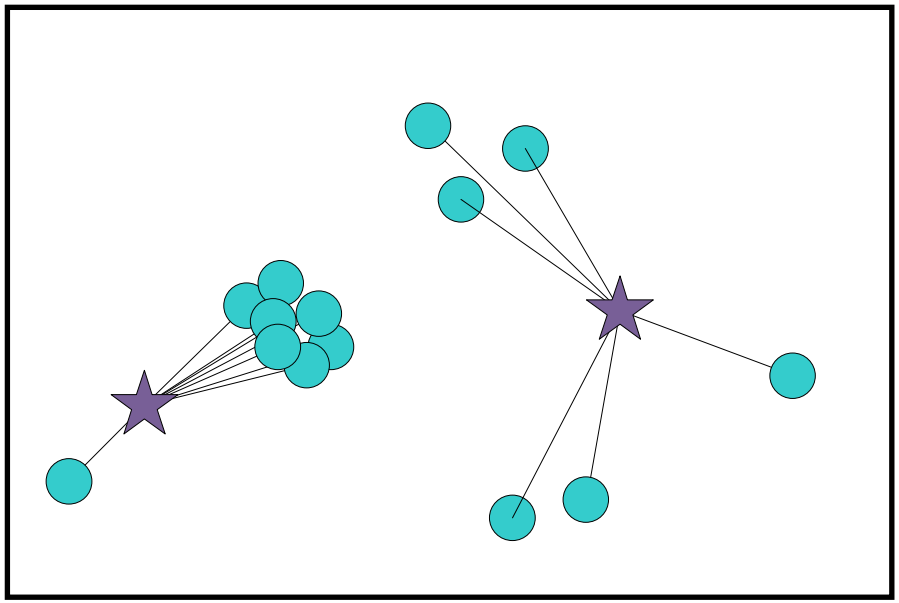
$$\sum_{p \in P} dist(p, Q) \sim \sum_{c \in C} w(c) \cdot dist(c, Q)$$

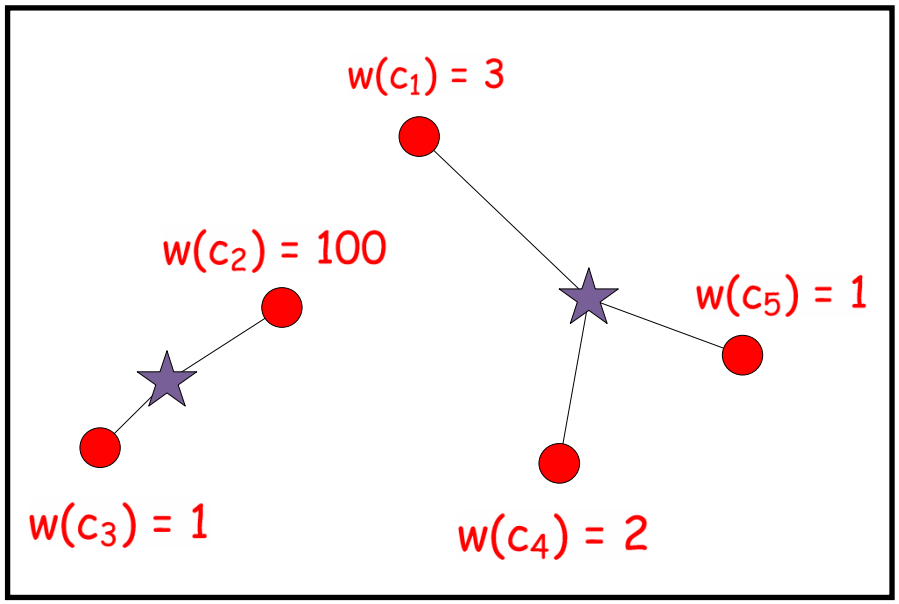
Multiplicative error $\leq 1 + \varepsilon$



(k, ε)-Median Coreset

$$\sum_{p \in P} \mathrm{dist}(p, Q) \quad \sim \quad \sum_{c \in C} w(c) \cdot \mathrm{dist}(c, Q)$$





Related Work

- Strong coreset (also for k-median/mean),

  Size: $(k/\varepsilon)^{d/\varepsilon}$, [Har-Peled, Kosharl'05]

  Size: $\text{poly}(kd \log n/\varepsilon)$, [Ke-Chen'06]

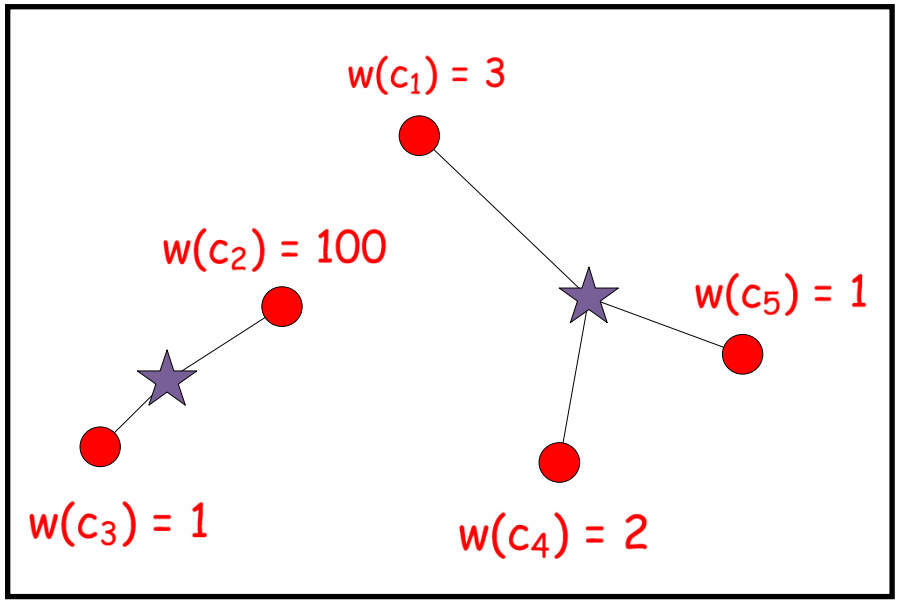
  Size: $\text{poly}(kd/\varepsilon)$, [Langberg, Schulman'10],

  [F, Langberg '10]

Definition

C is a weak (k,ε)-coreset for P if ~~$\forall Q, |Q| = k$~~ :

$$\min_{Q} \sum_{p \in P} \mathrm{dist}(p,Q) \sim \min_{Q} \sum_{c \in C} w(c) \cdot \mathrm{dist}(c,Q)$$



Related Work

- Strong coreset (also for k-median/mean),

  Size: $(k/\varepsilon)^{d/\varepsilon}$, [Har-Peled, Kosharl'05]

  Size: $\text{poly}(kd \log n/\varepsilon)$, [Ke-Chen'06]

  Size: $\text{poly}(kd/\varepsilon)$, [Langberg, Schulman'10],

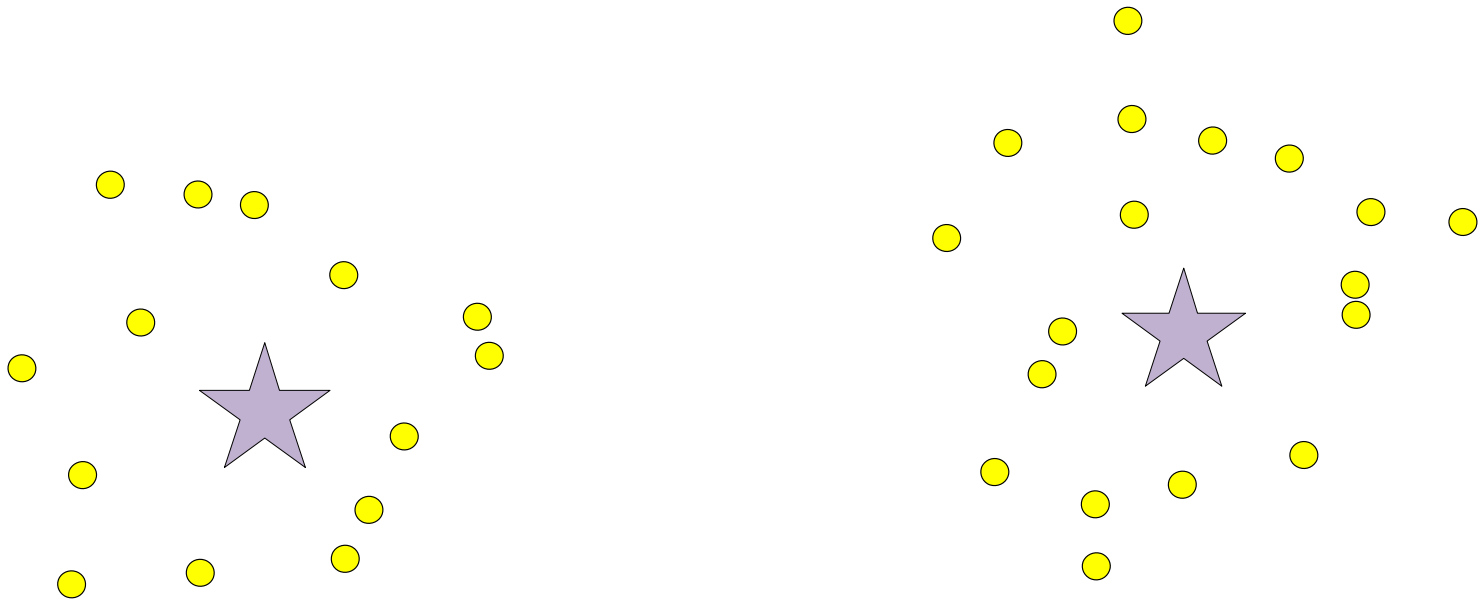
  [F, Langberg '10]

- Weak coresets, [F, Sohler, Monemizadeh'07]

  Size: $\text{poly}(k/\varepsilon)$,

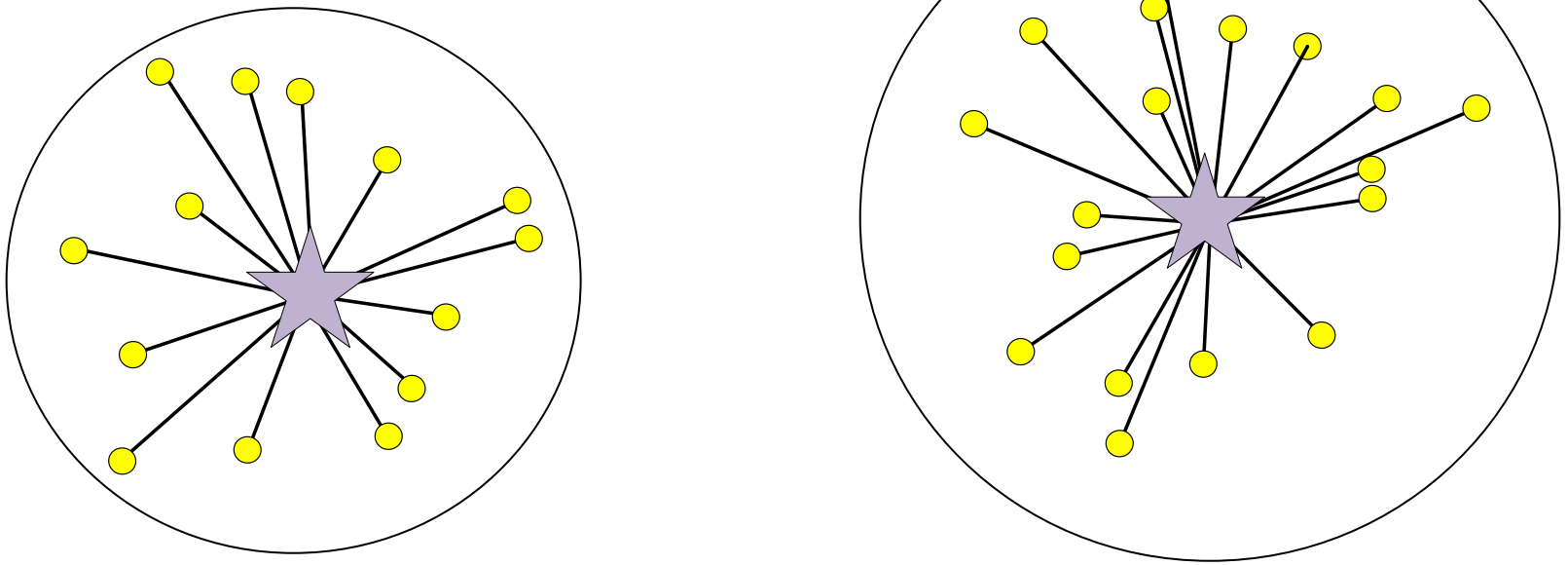
The k-Median of P

$$\text{opt} = \min_{|OPT|=k} \sum_{p \in P} \text{dist}(p, OPT)$$



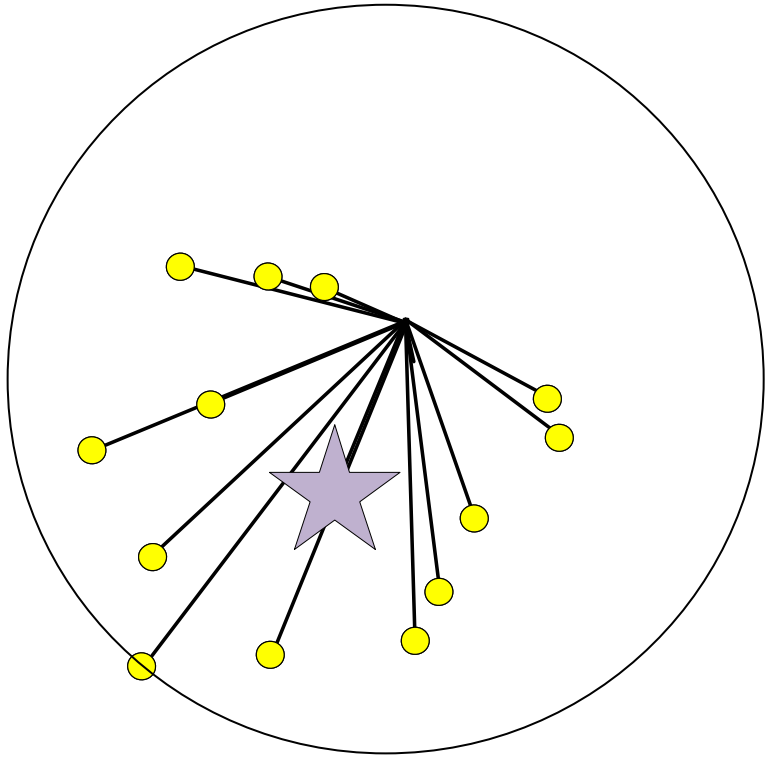
The k-Median of P

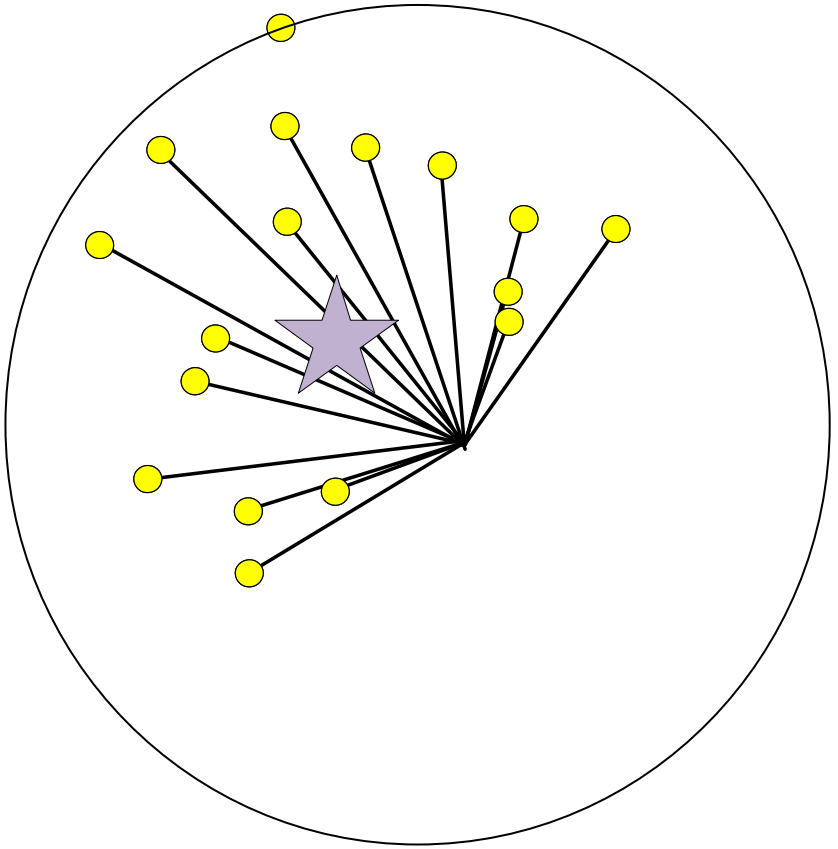
$$opt = \min_{|OPT|=k} \sum_{p \in P} dist(p, OPT)$$



Constant Approximation

$|\widetilde{OPT}| = k,$ $\quad \sum_{p \in P} \mathrm{dist}(p, \widetilde{OPT}) \le c \cdot \mathrm{opt}$

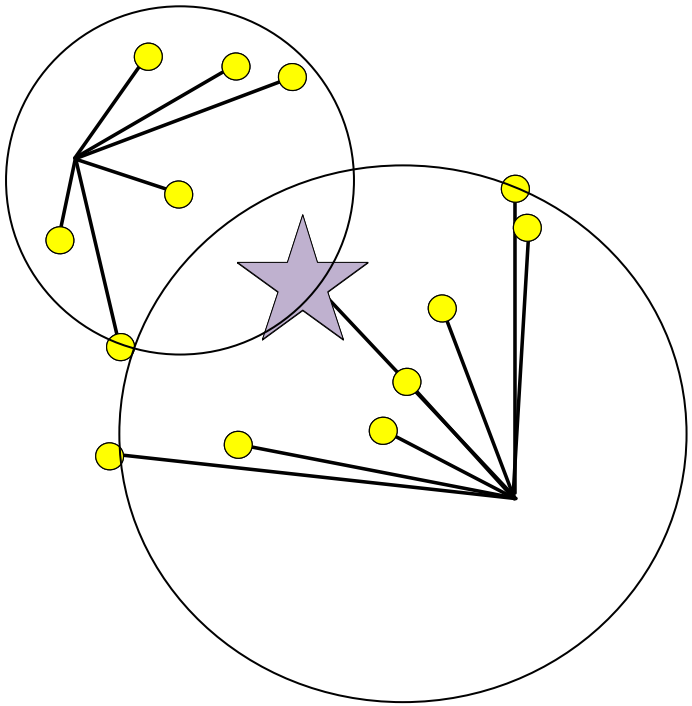


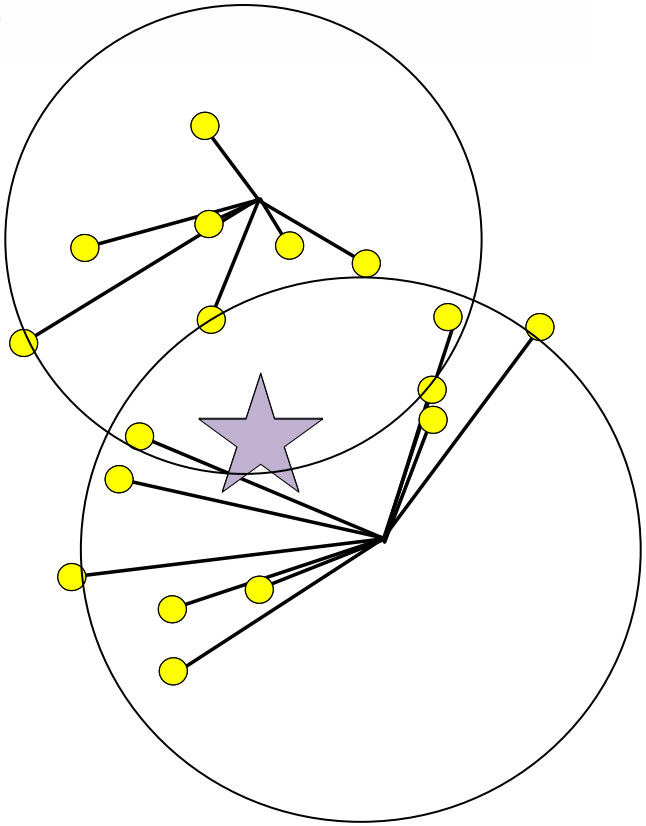


Bi-Criteria Approximation [FFS07]

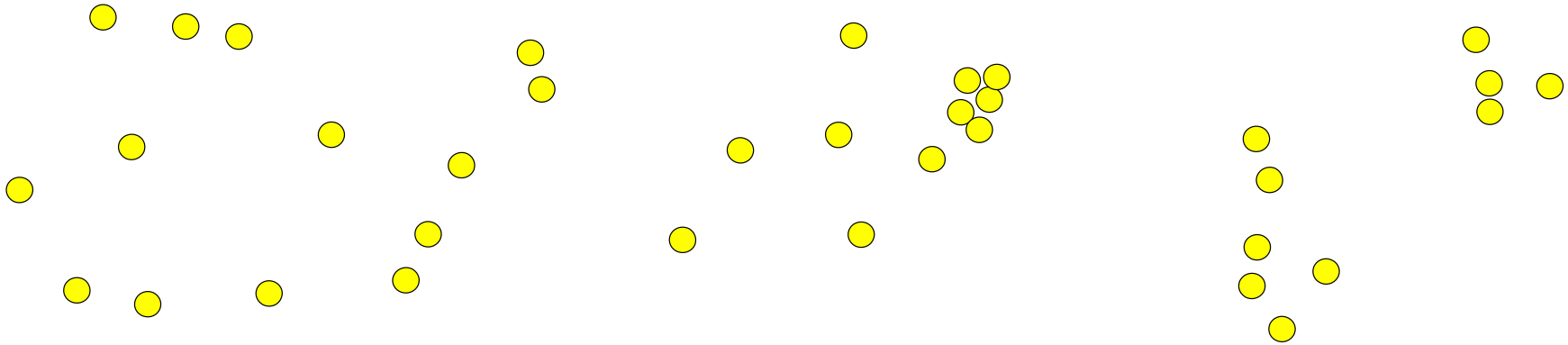
$|B| = O(k \log n)$,

$\sum_{p \in P} \text{dist}(p, B) \leq c \cdot \text{opt}$

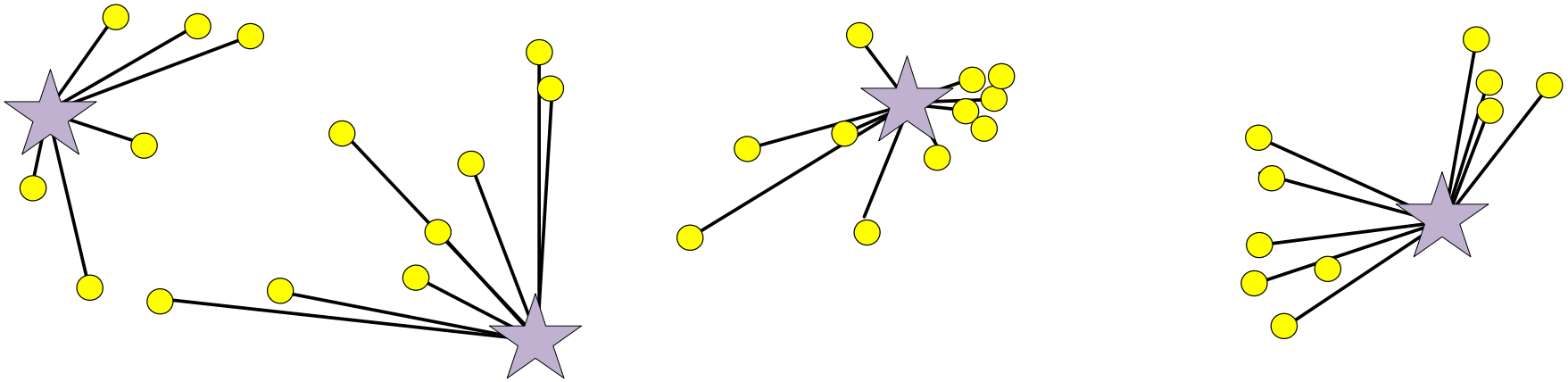




Coreset for 2-Median [FMS07]



Coreset for 2-Median

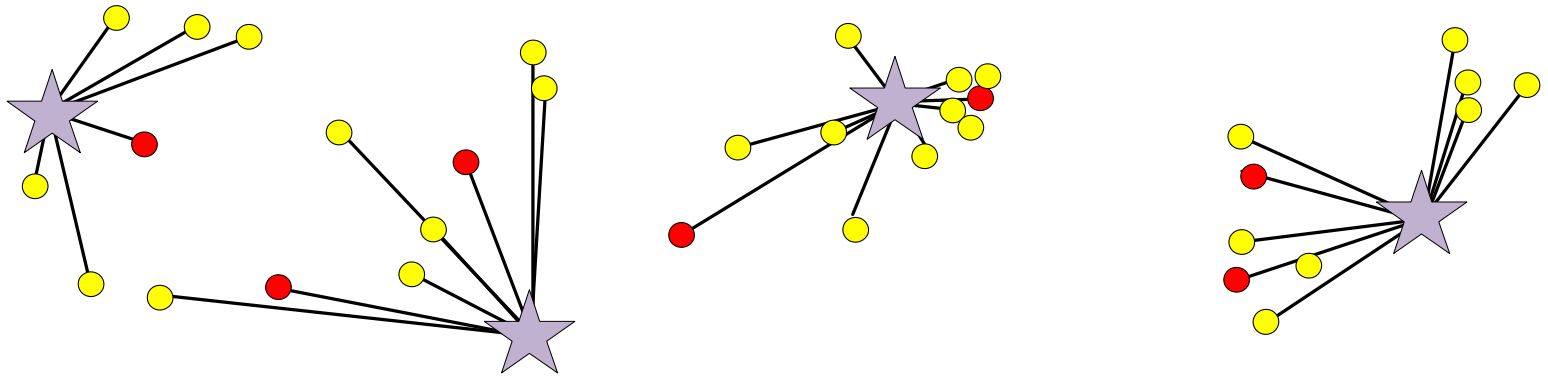


Using a bi-criteria approximation for 2-Median

Coreset for 2-Median

1. Pick a sample S of $d/\varepsilon^{O(1)}$ points, where

$$\mathbf{Pr}[p] = \frac{\mathsf{dist}(p, B)}{\sum_{p \in P} \mathsf{dist}(p, B)}$$

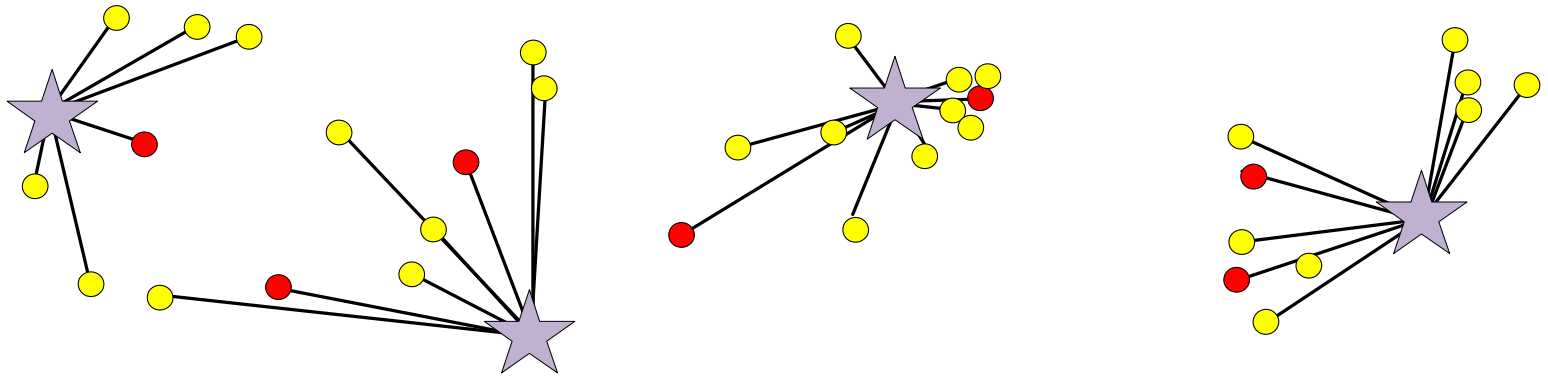


Coreset for 2-Median

1. Pick a sample S of $d/\varepsilon^{O(1)}$ points, where

$$\mathbf{Pr}[p] = \frac{\mathsf{dist}(p, B)}{\sum_{p \in P} \mathsf{dist}(p, B)}$$

2. $\forall p \in S : w(p) \leftarrow \frac{1}{|S| \cdot \mathbf{Pr}[p]}$



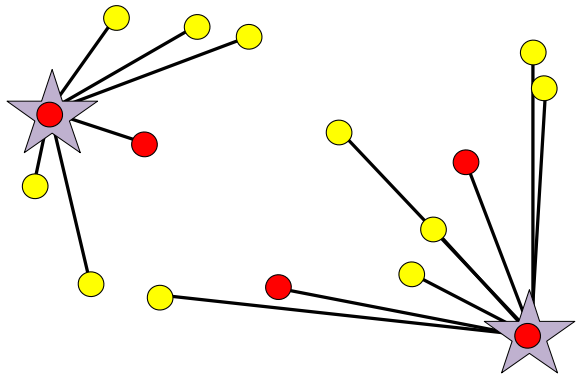
Coreset for 2-Median

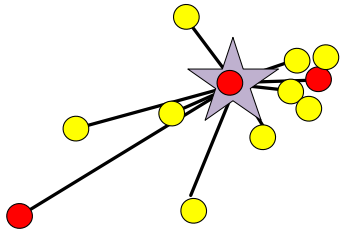
1. Pick a sample S of $d/\varepsilon^{O(1)}$ points, where

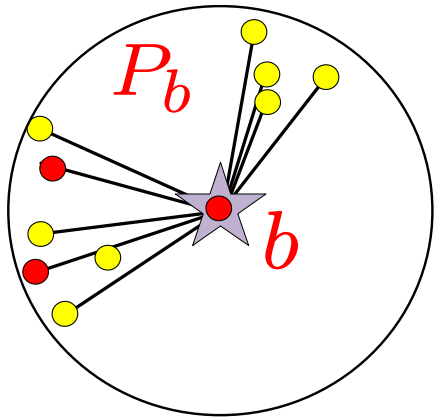
$$\mathbf{Pr}[p] = \frac{\mathsf{dist}(p, B)}{\sum_{p \in P} \mathsf{dist}(p, B)}$$

2. $\forall p \in S : w(p) \leftarrow \dfrac{1}{|S| \cdot \mathbf{Pr}[p]}$

3. $\forall b \in B : w(b) \leftarrow |P_b| - w(S \cap P_b)$







Coreset for 2-Median

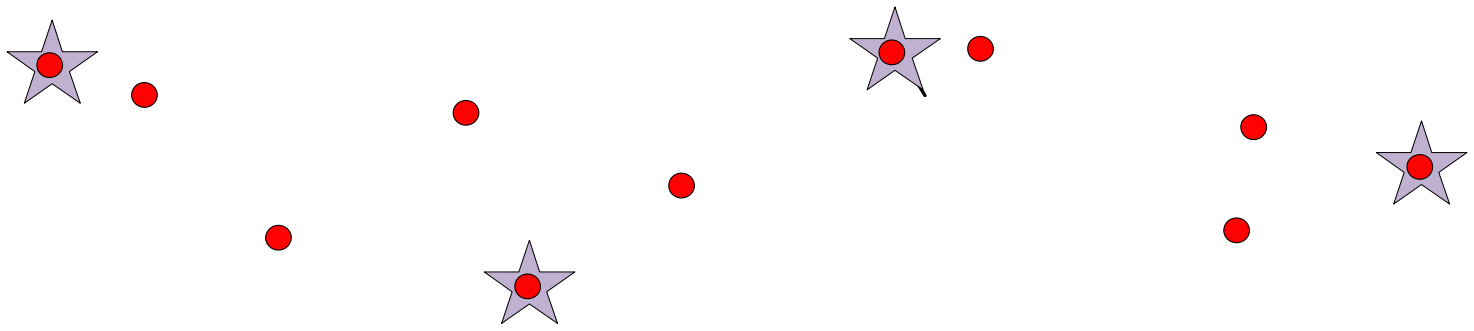
1. Pick a sample S of $d/\varepsilon^{O(1)}$ points, where

$$\mathbf{Pr}[p] \leftarrow \frac{\mathsf{dist}(p, B)}{\sum_{p \in P} \mathsf{dist}(p, B)}$$

2. $\forall p \in S : w(p) \leftarrow \dfrac{1}{|S| \cdot \mathbf{Pr}[p]}$

3. $\forall b \in B : w(b) \leftarrow |P_b| - w(S \cap P_b)$

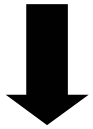
4. Return $S \cup B$



Correctness

We will prove that, with high probability,

$$\sum_{p\in P}\Big(\mathsf{dist}(p,Q)-\mathsf{dist}(b_p,Q)\Big)\sim\sum_{p\in S}w(p)\cdot\Big(\mathsf{dist}(p,Q)-\mathsf{dist}(b_p,Q)\Big)$$



$$\begin{aligned}\sum_{p\in P}\mathsf{dist}(p,Q)&=\sum_{p\in P}\mathsf{dist}(b_p,Q)+\sum_{p\in P_b}\Big(\mathsf{dist}(p,Q)-\mathsf{dist}(b_p,Q)\Big)\\&\sim\sum_{b\in B}|P_b|\cdot\mathsf{dist}(b,Q)+\sum_{p\in S}w(p)\Big(\mathsf{dist}(p,Q)-\mathsf{dist}(b_p,Q)\Big)\\&=\sum_{b\in B}\Big(|P_b|-w(S\cap P_b)\Big)\mathsf{dist}(b,Q)+\sum_{p\in S}w(p)\mathsf{dist}(p,Q)\quad.\end{aligned}$$

$$\sum_{p\in S} w(p)\cdot\Big(\mathsf{dist}(p,Q)-\mathsf{dist}(b_p,Q)\Big)\stackrel{?}{\sim}\sum_{p\in P}\Big(\mathsf{dist}(p,Q)-\mathsf{dist}(b_p,Q)\Big)$$

Let p be the ith sample in S, and

$$x_i := w(p)\cdot\Big(\mathsf{dist}(p,Q)-\mathsf{dist}(b_p,Q)\Big).$$

$$E\left[\sum_{i=1}^{|S|} x_i\right] = |S|\cdot E[x_1]$$

$$= |S|\sum_{p\in P}\mathbf{Pr}[p]\cdot w(p)\Big(\mathsf{dist}(p,Q)-\mathsf{dist}(b_p,Q)\Big)$$

$$= \sum_{p\in P}\Big(\mathsf{dist}(p,Q)-\mathsf{dist}(b_p,Q)\Big)$$

$$\sum_{i=1}^{|S|} x_i \stackrel{?}{\sim} E\left[\sum_{i=1}^{|S|} x_i\right]$$

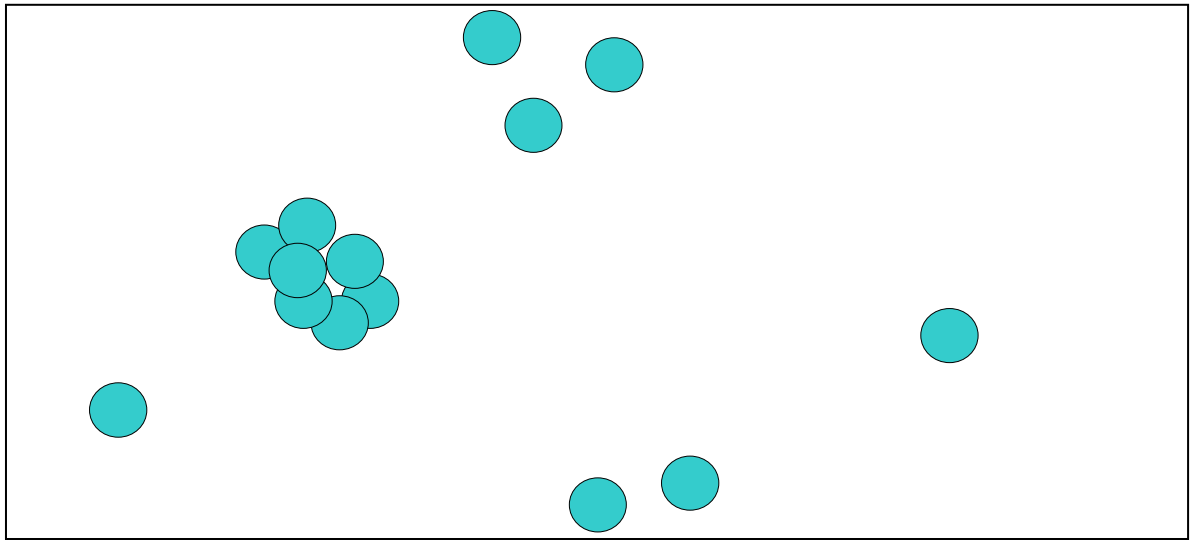
$$\begin{aligned}
|x_i| &\le w(p) \cdot \Big|\mathsf{dist}(p,Q) - \mathsf{dist}(b_p,Q)\Big| \\
&= \frac{\sum_{p\in P} \mathsf{dist}(p,B)}{|S|\mathsf{dist}(p,b_p)} \cdot \Big|\mathsf{dist}(p,Q) - \mathsf{dist}(b_p,Q)\Big| \\
&\le \frac{c\cdot \mathsf{opt}}{|S|}
\end{aligned}$$

By Chernoff inequality, with constant probability,

$$\left|\sum_{i=1}^{|S|} x_i - E\left[\sum_{i=1}^{|S|} x_i\right]\right| \le |S| \cdot \varepsilon \max_i |x_i| \le \varepsilon c \cdot \mathsf{opt}$$

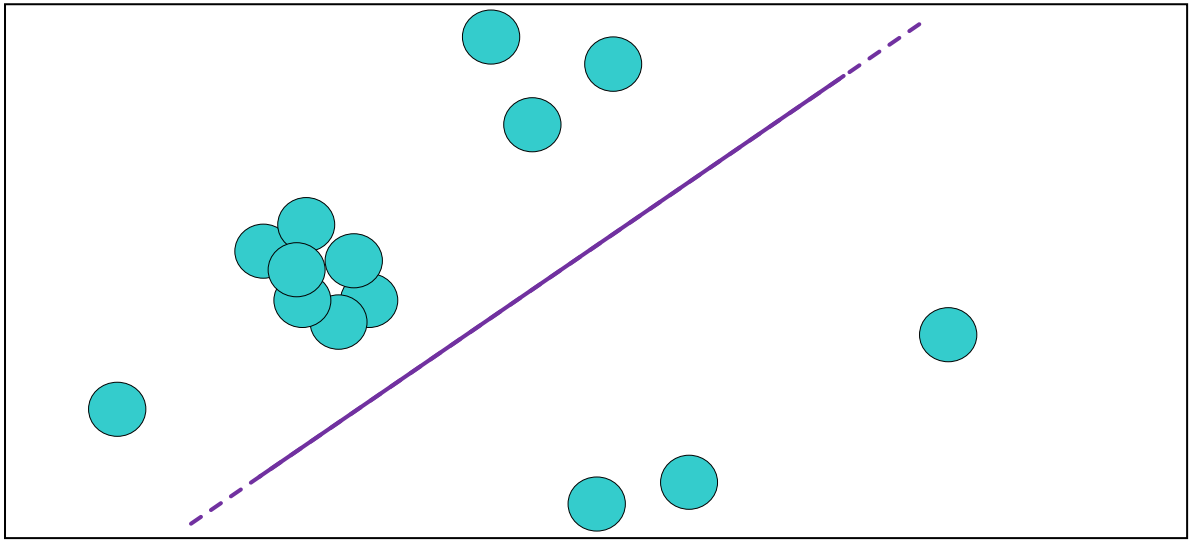
Line Queries

- Input: P in $\mathbb{R}^d$



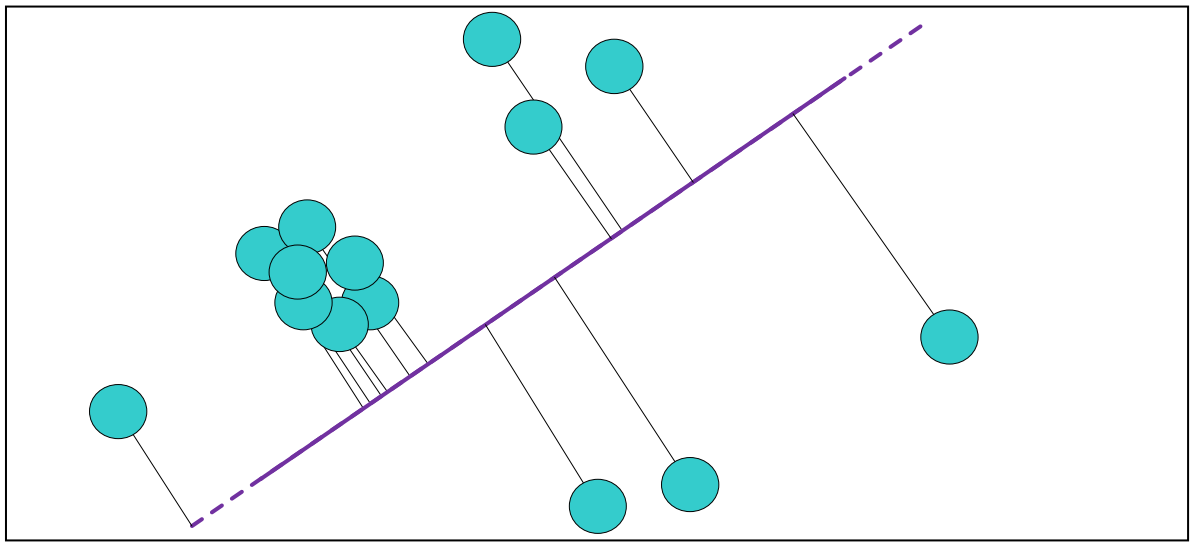
Line Queries

- Input: P in $\mathbb{R}^d$
- Query: a line Q



Line Queries

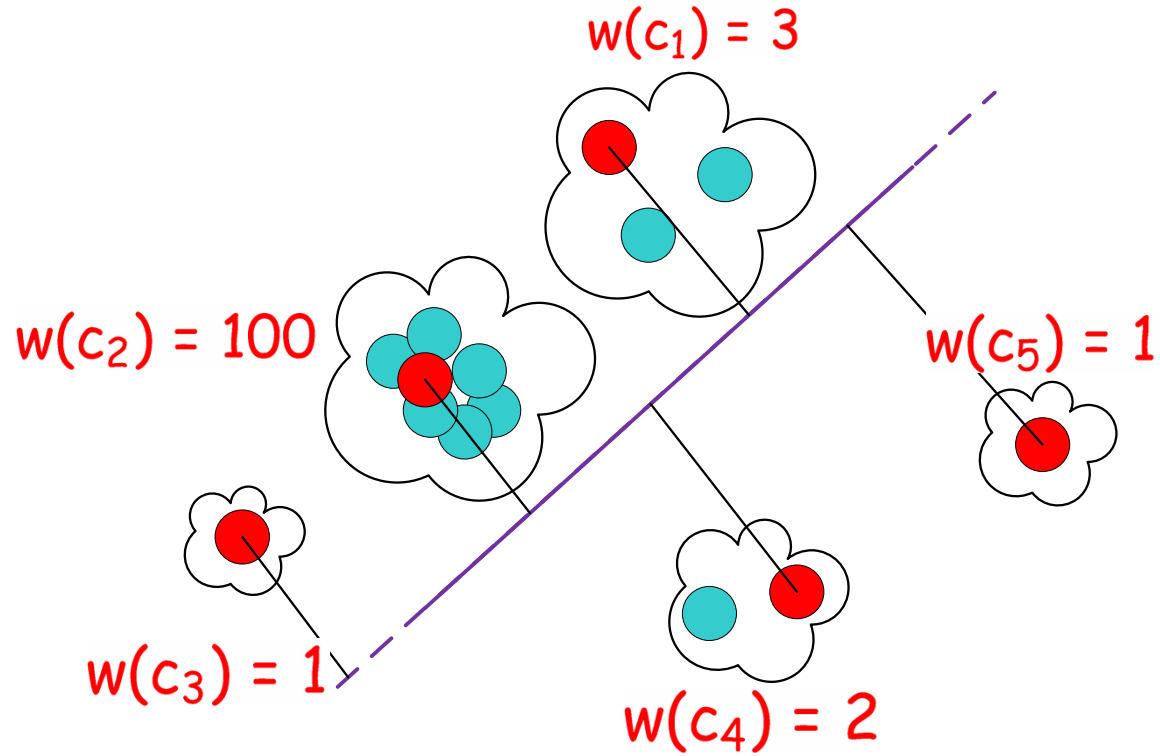
- Input: P in $\mathbb{R}^d$
- Query: a line Q
- Output: $\sum_{p \in P} \mathrm{dist}(p, Q)$



ε-Line Coreset

Answer line queries in sub-linear time

Key Idea: Replace many points by weighted representatives:

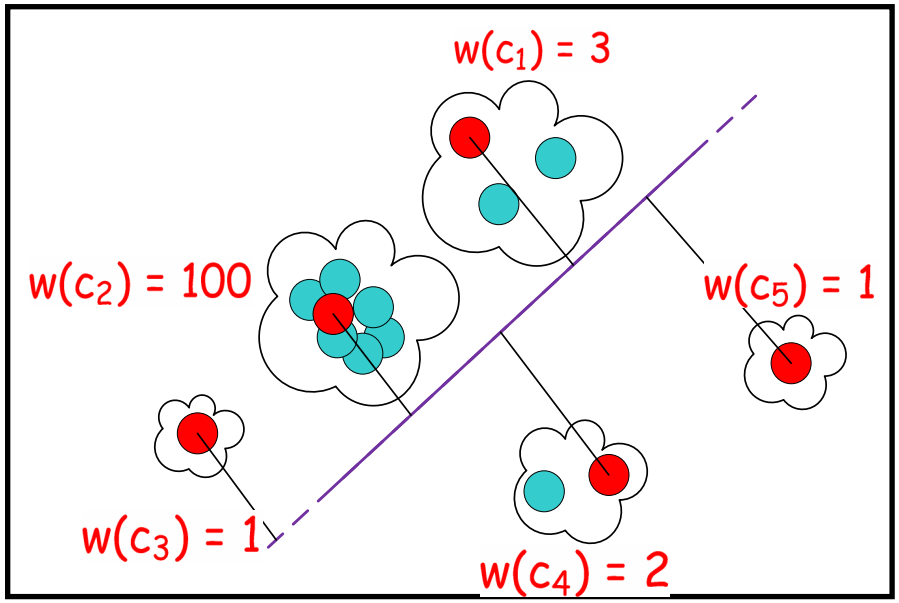


(j, ε)-Subspace Coreset

C is a (j, ε)-subspace coreset for P, if for every j-dimensional subspace Q,

$$\sum_{p \in P} \text{dist}(p, Q) \sim \sum_{c \in C} w(c) \cdot \text{dist}(c, Q)$$

Multiplicative error $\leq 1 + \varepsilon$



Approximtion Algorithms (PTAS)

- Mean:

  $O(nd^2)$ time, Exact (SVD) [Pearson, 1901]

  $nd \cdot \text{poly}(j, 1/\epsilon)$ time, PTAS
  [Deshpande et al.,][Sarlos]
  [Har-Peled] (2006)

- Median:

  $nd \cdot 2^{\text{poly}(j,1/\epsilon)}$ time, PTAS
  [Shyamalkumar & Varadarajan, 2007]

- Center:

  $nd \cdot \exp\left(\text{poly}(2^{(j^2)}, 1/\epsilon)\right)$ time, PTAS
  [Har-Peled & Varadarajan, 2004]

Coresets

- Strong (j, ε)-subspace coresets,

  Size: $\exp(d)$, [F, Fiat, Sharir'06]

- Weak coresets for j-subspace Q

  which minimizes $\max_{p \in P} \text{dist}(p, Q)$

  Size: $\text{poly}(j/\varepsilon)$, [Har-Peled, Varadarajan'04]

- Weak coresets for regression distances,

  Size: $\text{poly}(d/\varepsilon)$,

  [Dasgupta, Drineas, Harb, Kumar, Mahoney'08]

Recent Results [FMSW10]

- Weak (j, ε)-coresets,

  Size: $\text{poly}(1/\varepsilon)$ for a fixed j

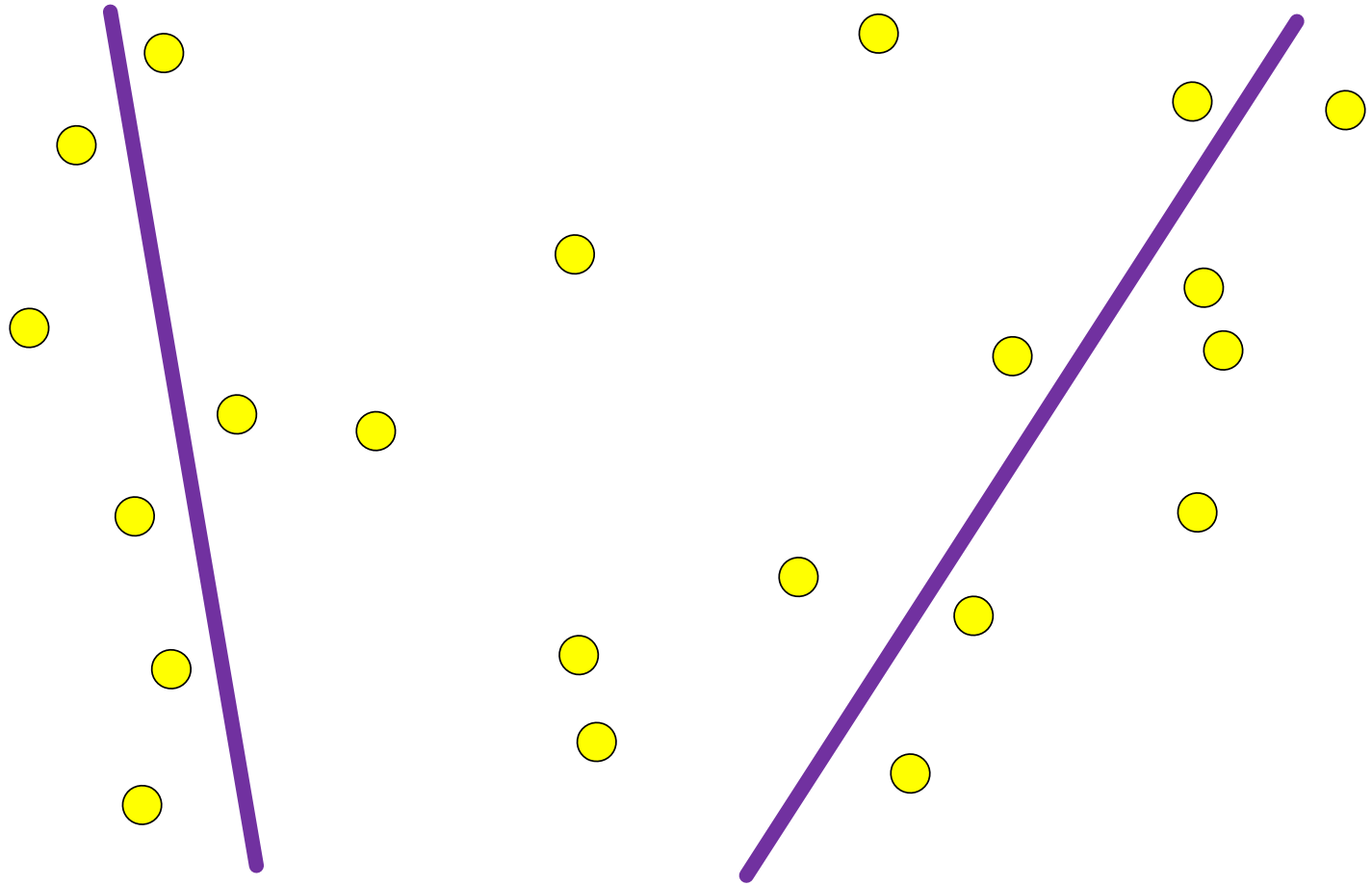
- Strong (j, ε)-coresets,

  Size: $\text{poly}(d/\varepsilon)$ for a fixed j

- Small sketches for streaming updates of coordinates of points

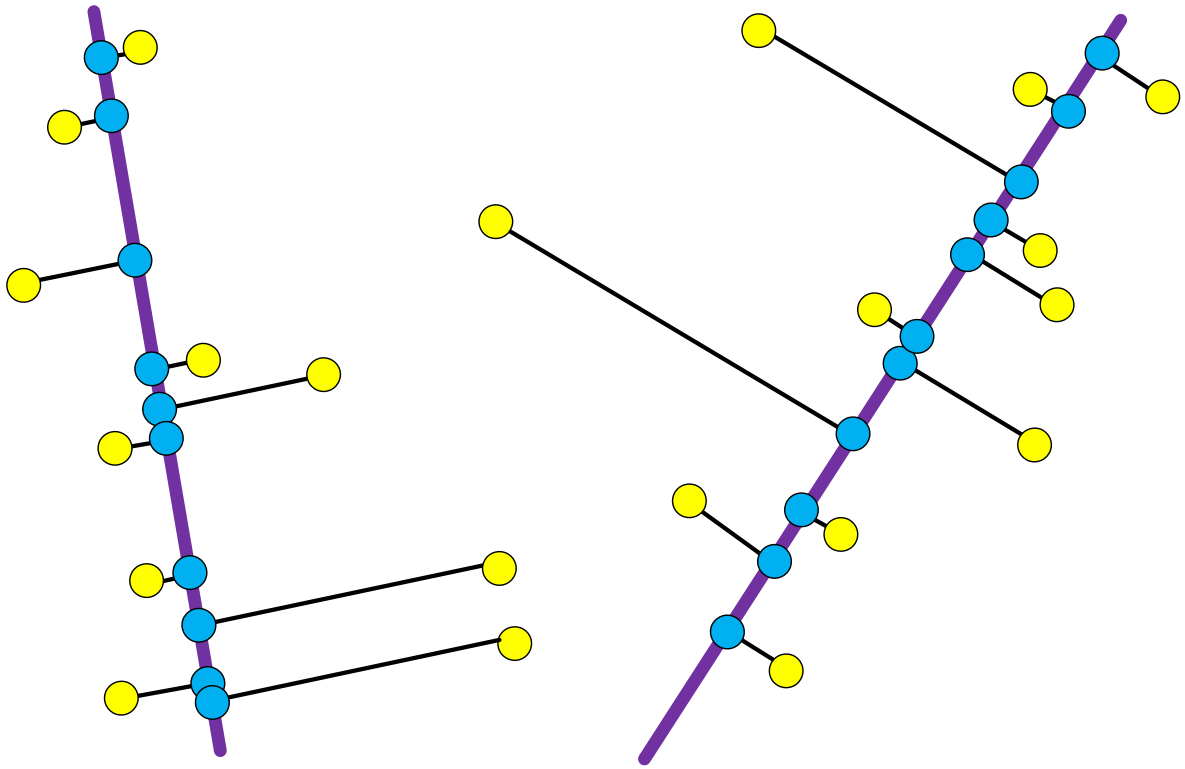
- General dimensional reduction technique for shape fitting

Coreset for 2-lines Median



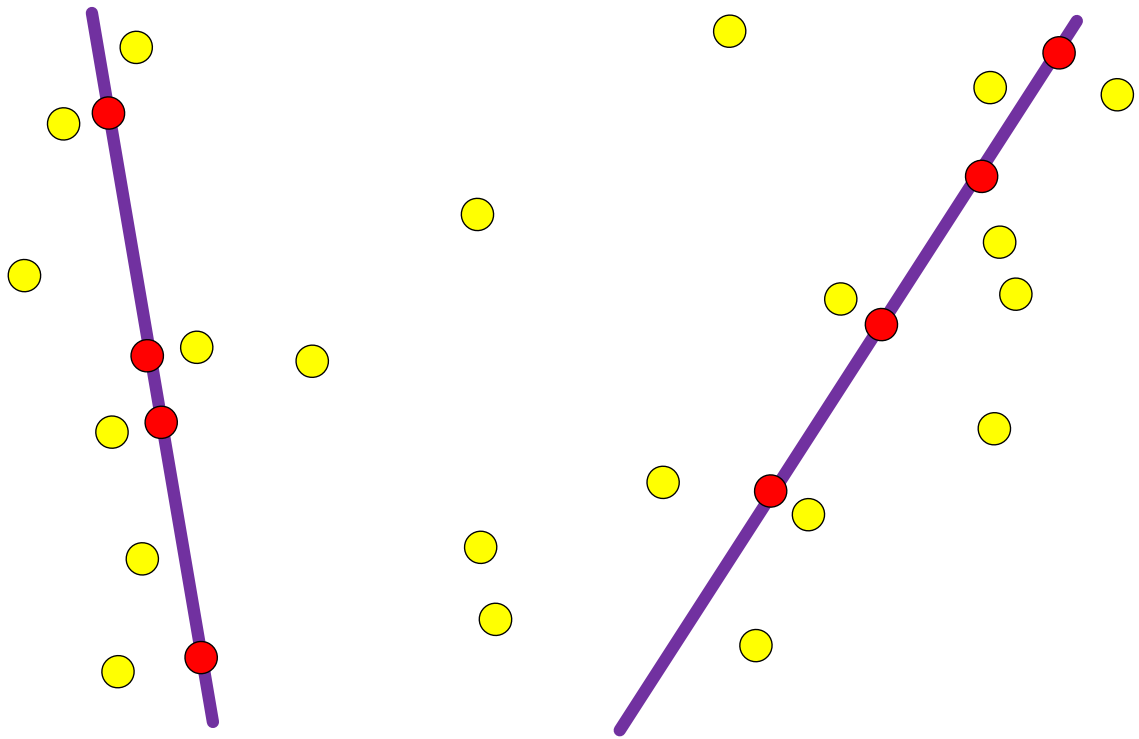
Coreset for 2-lines Median

1. $\forall b \in B$: $P_b \leftarrow$ projection of P onto b



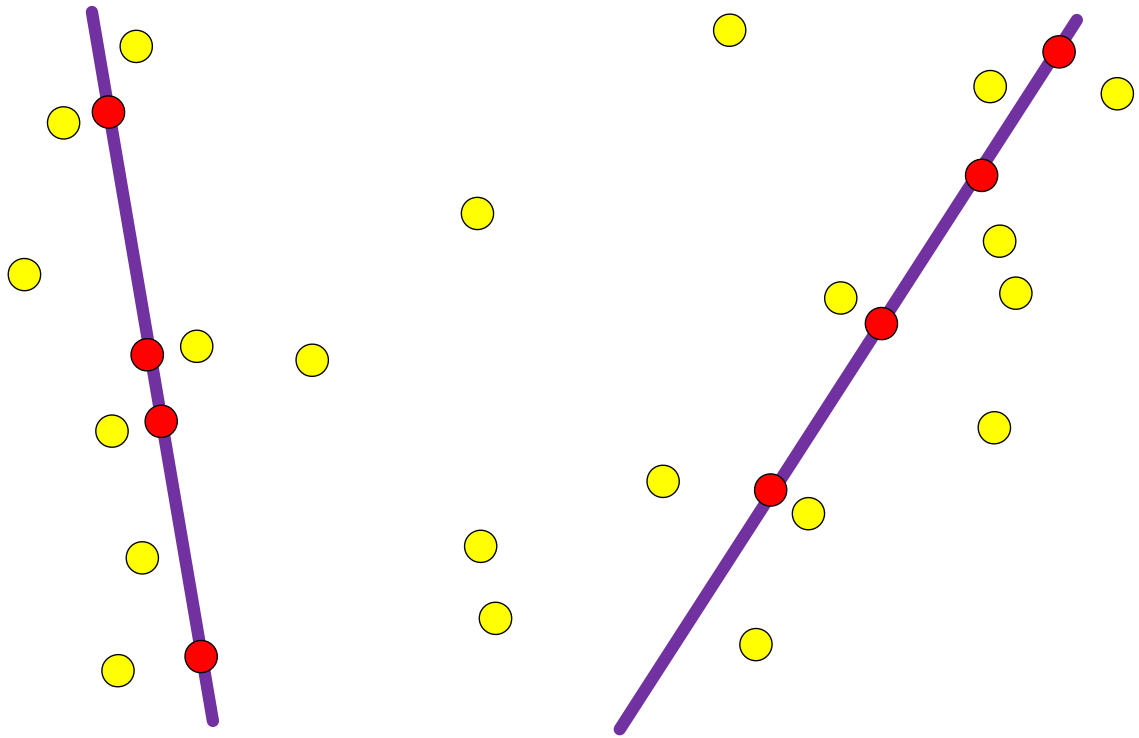
Coreset for 2-lines Median

1. $\forall b \in B : P_b \leftarrow$ projection of P onto b
2. $T_b \leftarrow$ coreset for P_b [FFS06]



Coreset for 2-lines Median

1. $\forall b \in B : P_b \leftarrow$ projection of P onto b
2. $T_b \leftarrow$ coreset for P_b [FFS06]
3. $T \leftarrow \bigcup_{b \in B} T_b$

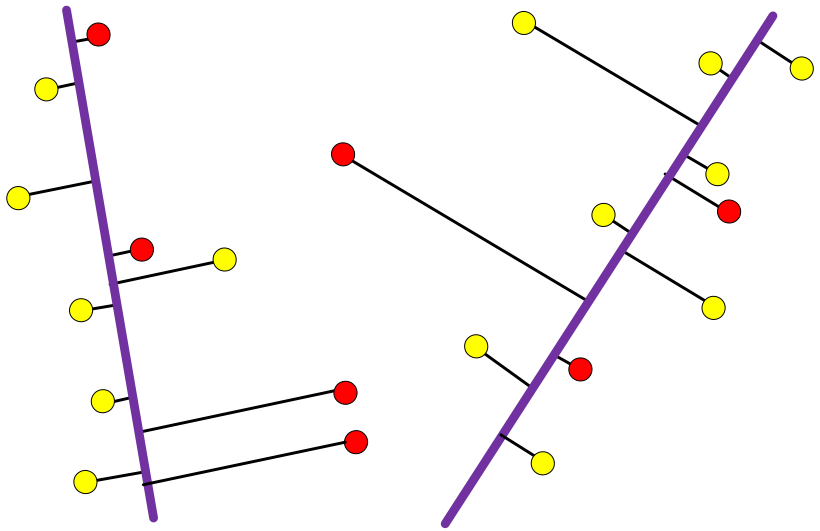


Coreset for 2-lines Median

4. Pick a sample S of $d/\varepsilon^{O(1)}$ points, where

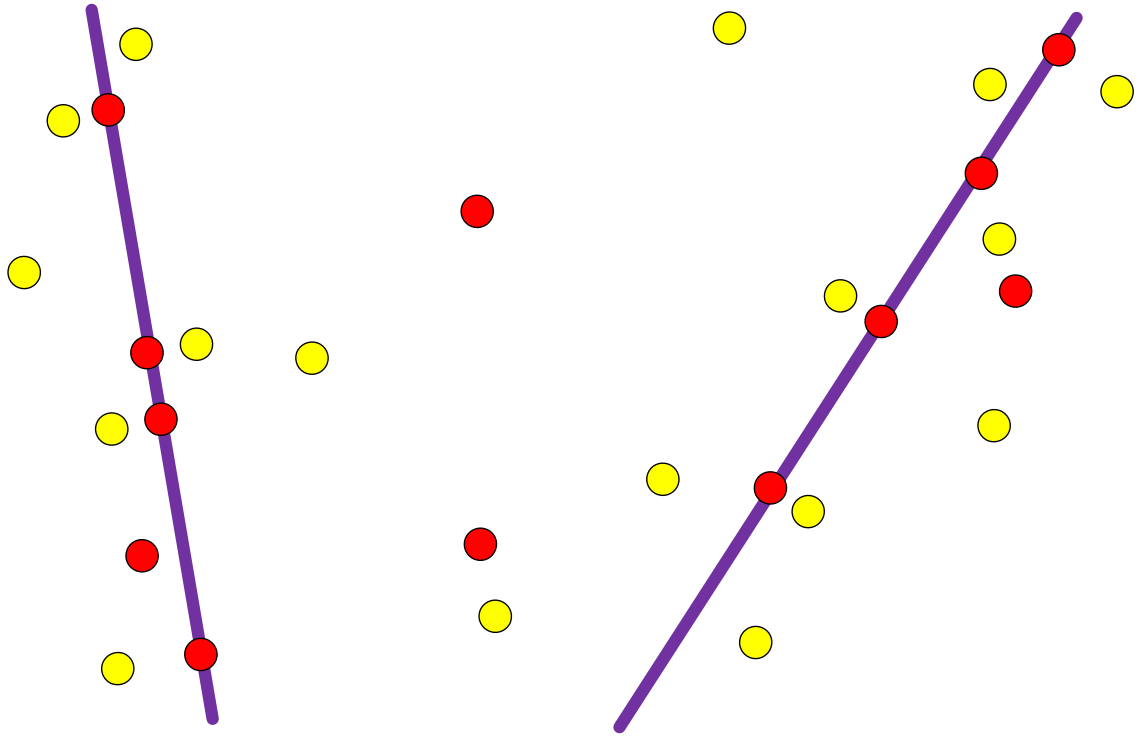
$$\mathbf{Pr}[p] = \frac{\mathrm{dist}(p, B)}{\sum_{p \in P} \mathrm{dist}(p, B)}$$

5. $\forall p \in S : w(p) \leftarrow \dfrac{1}{|S| \cdot \mathbf{Pr}[p]}$



Coreset for 2-lines Median

6. Return $T \cup S \cup \text{project}(S^-, B)$



Results for $j, k > 1$

PTAS that takes $d \cdot n^{\text{poly}(j,k,1/\epsilon)}$ time.

Mean: [Deshpande et al., 2006]

Median: [Shyamalkumar & Varadarajan, 2007]

Center: [Har-Peled & Varadarajan, 2002]

Thank You!