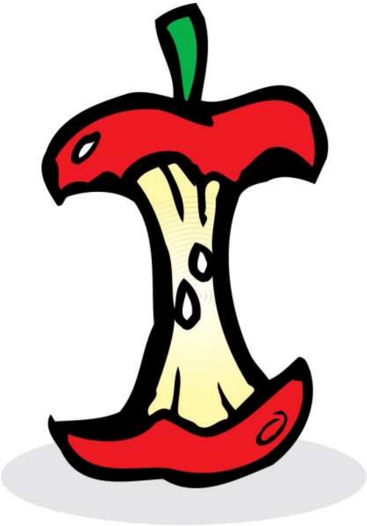
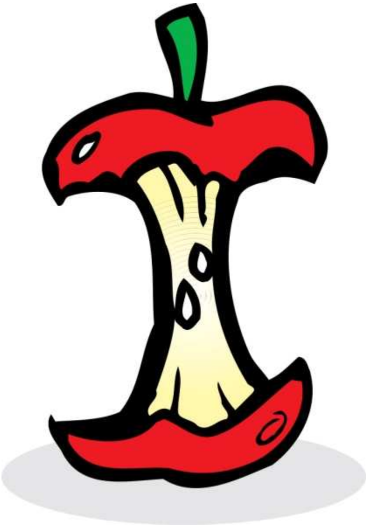


Private Coresets



Danny Feldman, Amos Fiat,
Haim Kaplan, Kobbi Nissim

Tor Vergata June 2009

Maybe, ½ hour from now you can insert the following in your small talk:

- Differential Privacy
- Coresets
- Private Coresets New
- Private Data New Structures (not only coresets) New
- Private ε nets New
- Private Bi Criteria

Why Privacy?

- $r_i \in \{0, 1\}$: indicator variable $= 1$ if i Republican

Getty

Why Privacy?

Indicator variable:

$$r_i = 1$$

$$r_i = 0$$

$$r = r_1 + \cdots + r_n$$

Problems

- If everyone has known political opinion but for voter n :

$$r_n = r - \sum_{i=1}^{n-1} r_i$$

Differential Privacy [DMNS06]

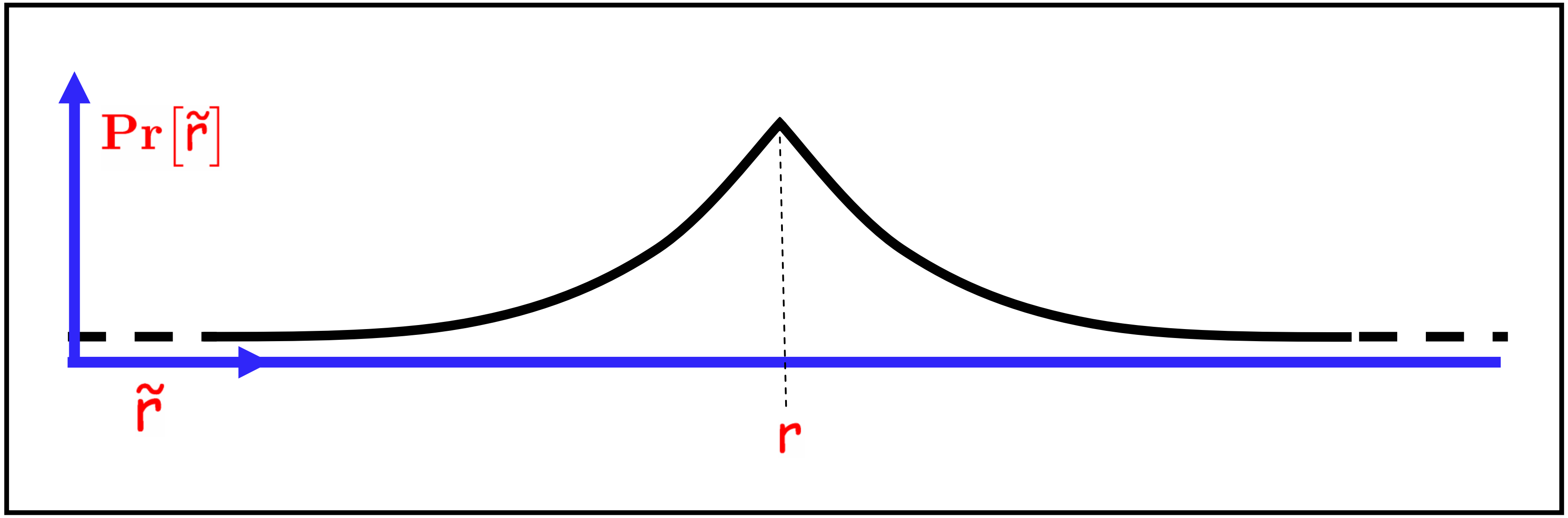
Algorithm A is α-*differentially private* if:

- for every two sets P and P' that differ by a single item:
- for every set S of possible outputs:

$$\frac{\Pr[A(P) \in S]}{\Pr[A(P') \in S]} \leq e^{\alpha} \approx 1 + \alpha$$

Private Counting

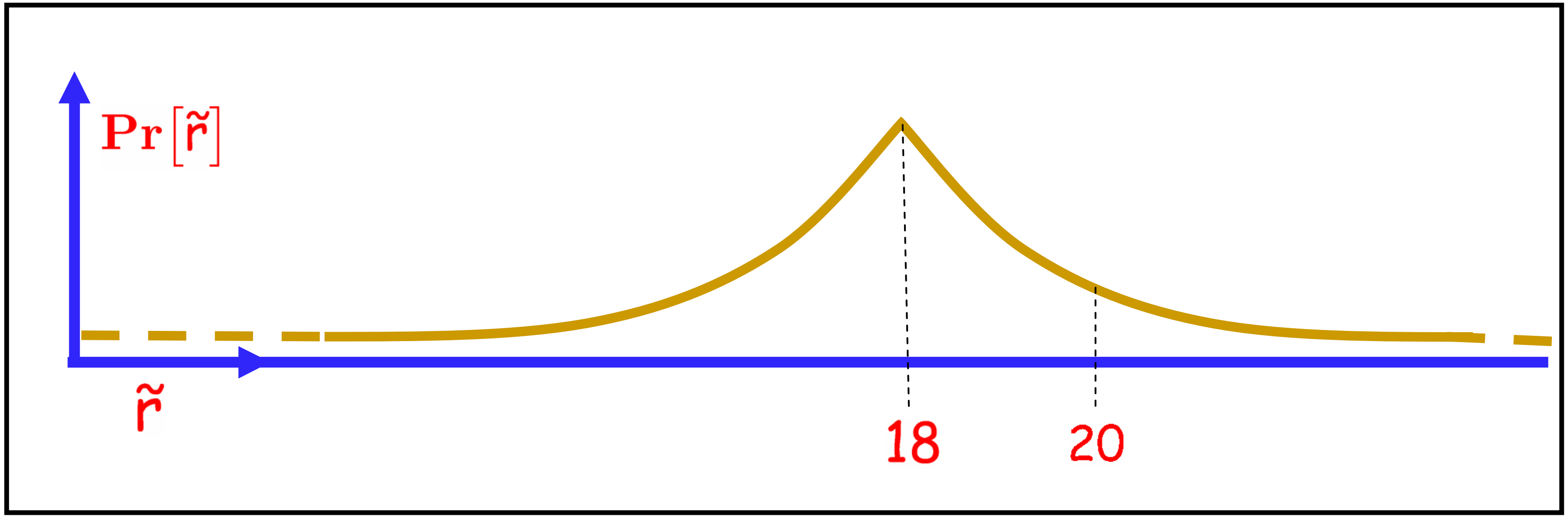
We publish r̃ = r + Noise



$$\mathbf{Pr}\left[\tilde{r} \in r + \text{Noise} \pm \epsilon\right] \approx \epsilon \frac{\alpha}{2} \cdot e^{-\alpha|\text{Noise}|}$$

Example: r = 18

We publish r̃ = 18 + Noise

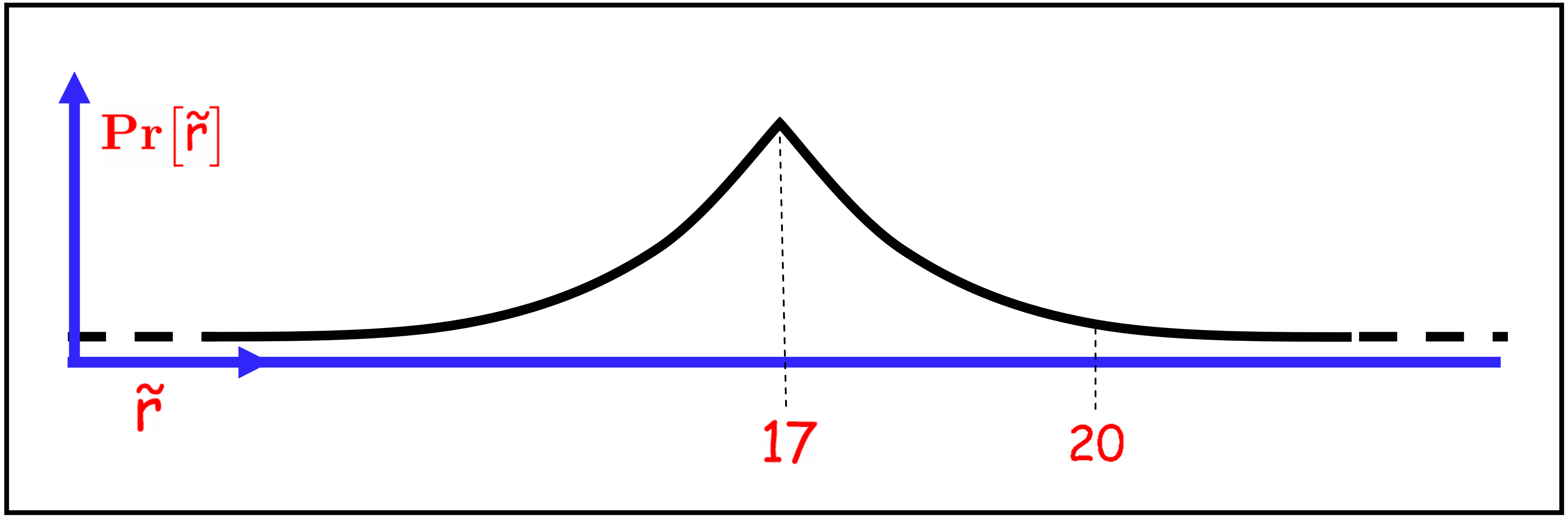


$$\mathbf{Pr}\left[\tilde{r} \in 20 \pm \epsilon\right] \approx \epsilon \frac{\alpha}{2} \cdot e^{-2\alpha}$$

(Noise = 2)

Example: r = 17

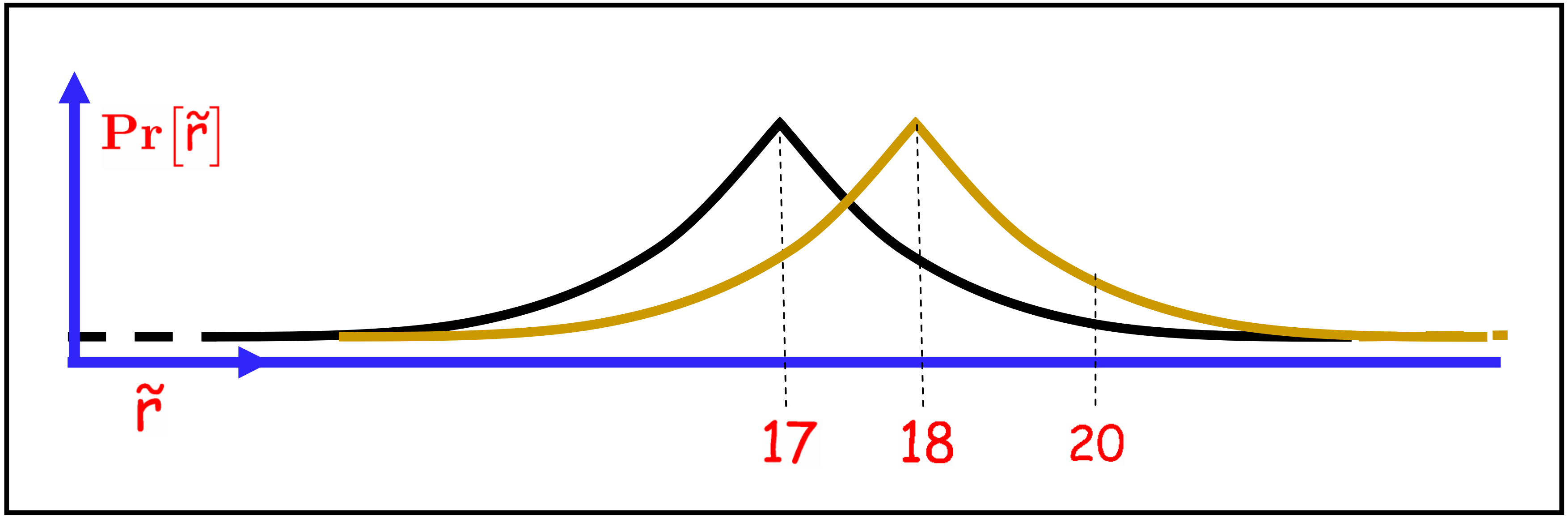
We publish $\tilde{r}$ = 17 + Noise



$$\mathbf{Pr}\left[\tilde{r} \in 20 \pm \epsilon\right] \approx \epsilon\frac{\alpha}{2} \cdot e^{-3\alpha}$$

(Noise = 3)

Private Counting



$$\frac{\mathbf{Pr}\left[\tilde{r} \in 20 \pm \epsilon \mid r = 18\right]}{\mathbf{Pr}\left[\tilde{r} \in 20 \pm \epsilon \mid r = 17\right]} = \frac{e^{-2\alpha}}{e^{-3\alpha}} = e^{\alpha} \approx 1 + \alpha$$

$\tilde{r}$ = r + Noise is α-differentially private

Strong Notion of Privacy

The attacker learns little "useful"
Prior Information does not help

Because of ε leakage:
Cannot be used to answer many queries

Strong Notion of Privacy

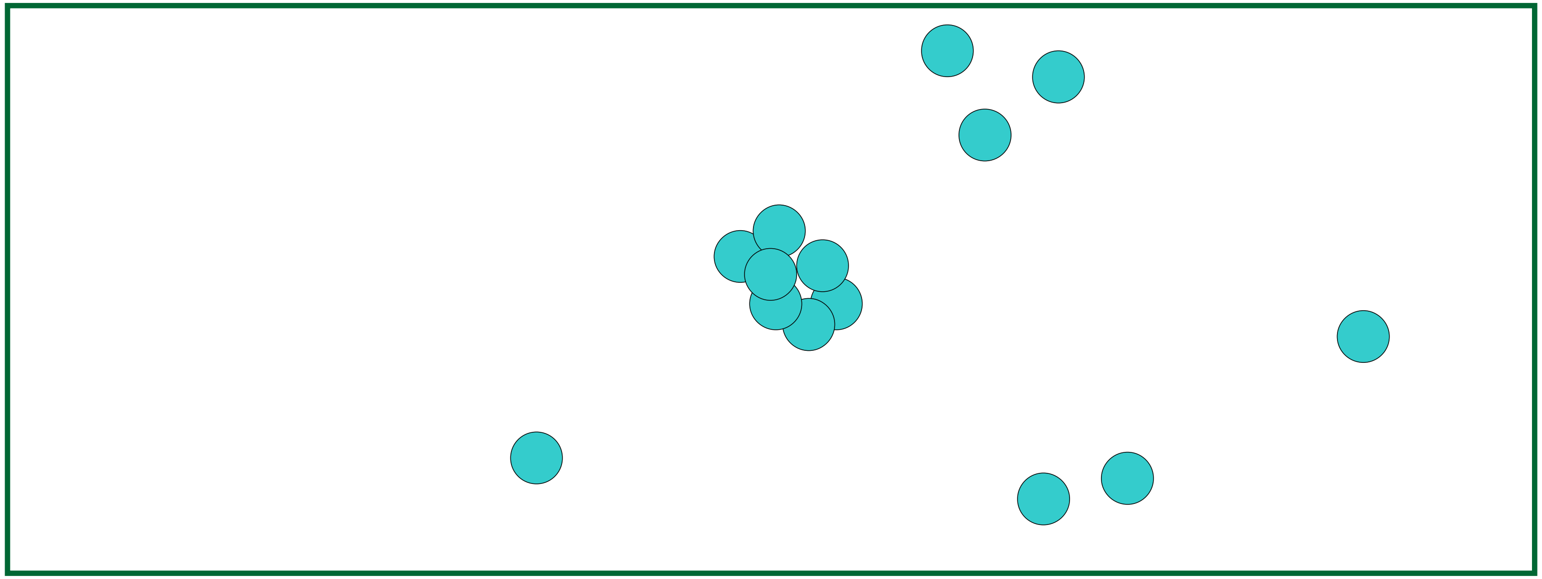
Want to answer not one query privately
But many queries privately

Leak ε only once, create Sanitized Data Set/Data Structure

k-Median Queries

No privacy

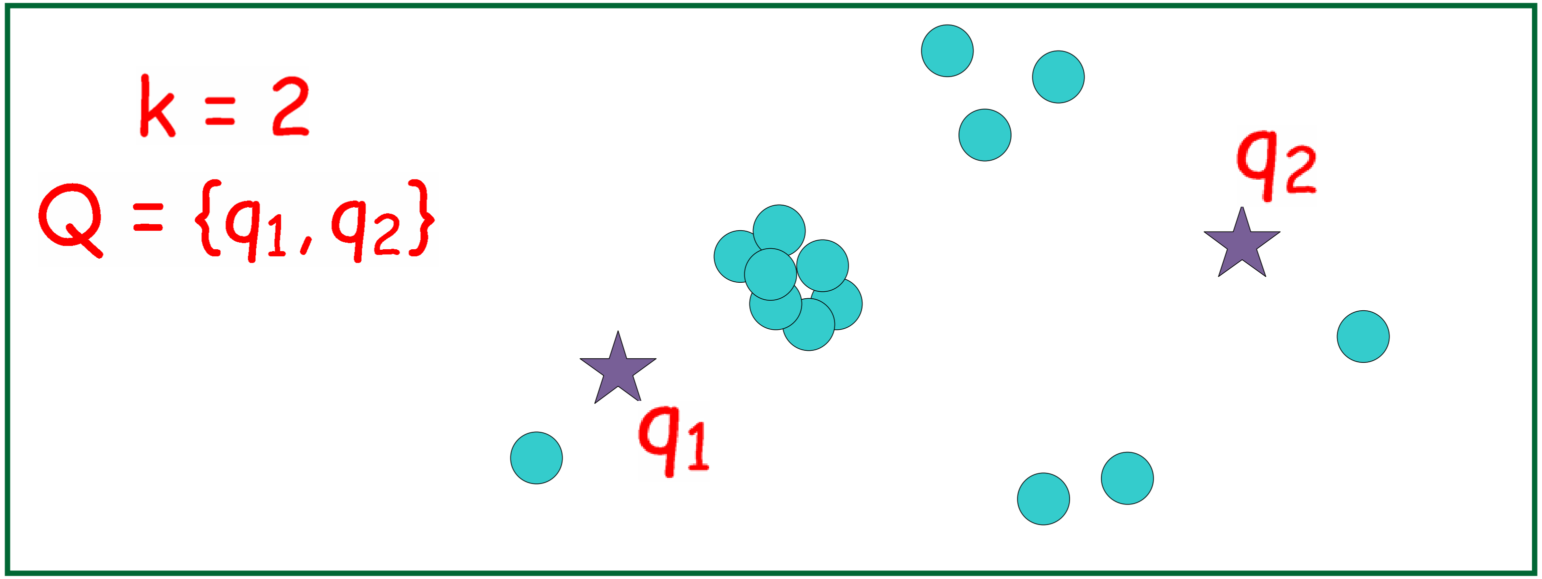
- Input: $P \subseteq [0,1]^d$



k-Median Queries

No privacy

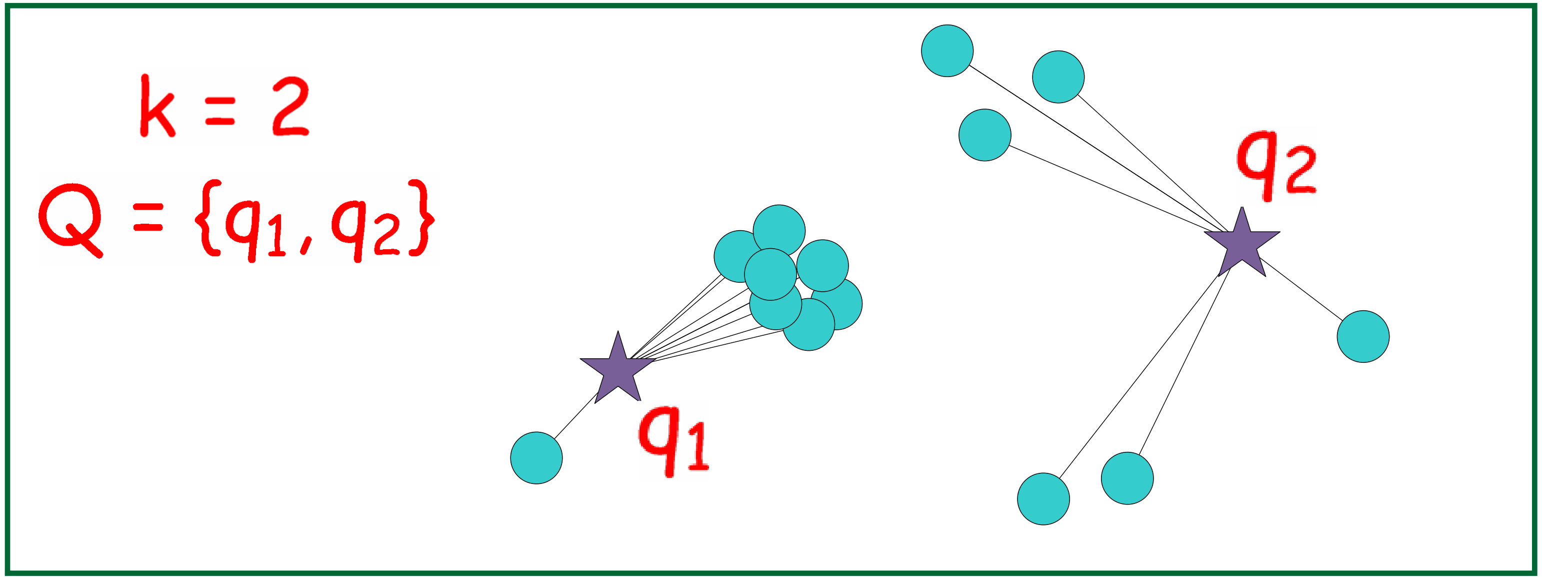
- Input: $P \subseteq [0,1]^d$
- Query: A set Q of k points



k-Median Queries

No privacy

- Input: $P \subseteq [0,1]^d$
- Query: A set Q of k points
- Output: $\sum_{p \in P} \text{dist}(p, Q) = \sum_{p \in P} \min_{q \in Q} \|p - q\|$

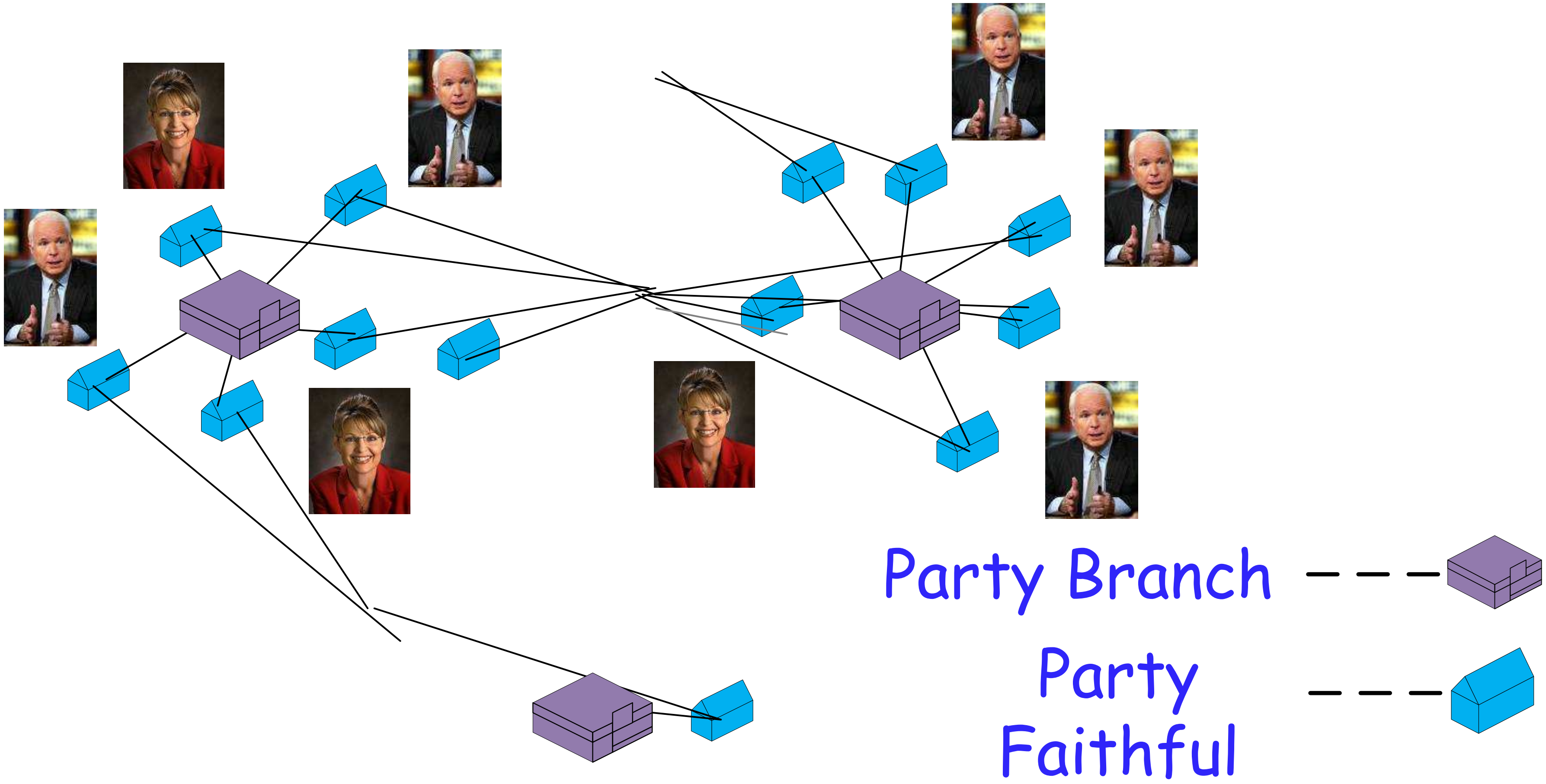


Motivation

No privacy

Comparing alternatives:

How good is this placement?



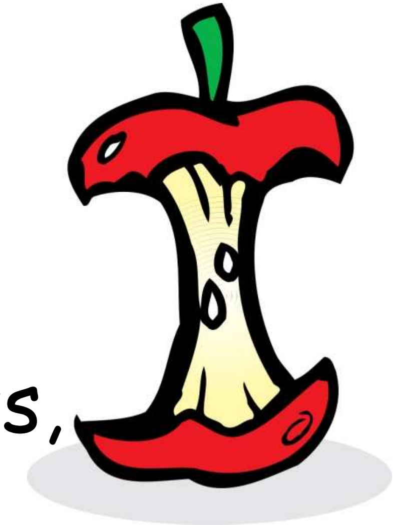
Party Branch ---

Party Faithful ---

Coresets

No privacy

- Coresets: “Clever Sample”
- Answer approximate queries from reduced representation (Coreset)
- Often leads to PTAS, FPTAS
- Many, many, papers, surveys
- Many problems: median, mean, flats, projective clustering, regression
- Intuition: Coresets give privacy on average

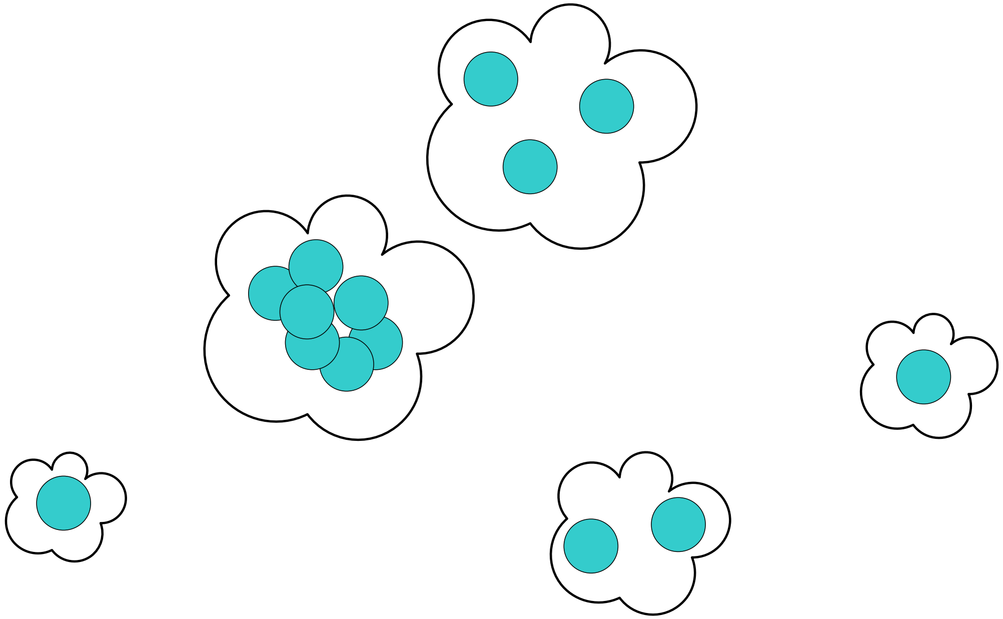


(k, ε)-Median Coreset

No privacy

Answer k-median queries in sub-linear time

Key Idea: Replace many points by one weighted representative:

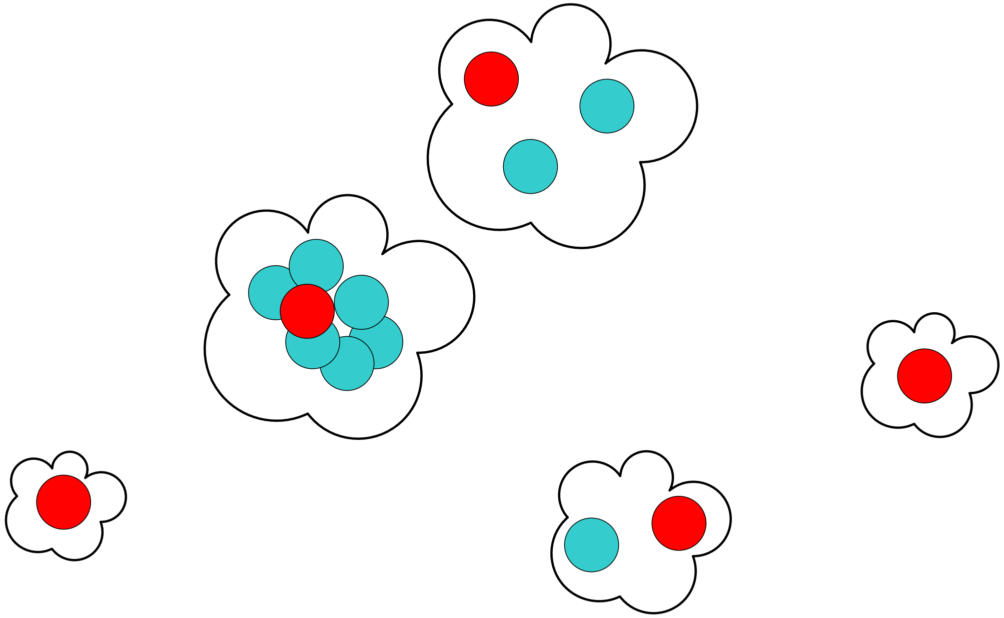


(k, ε)-Median Coreset

No privacy

Answer k-median queries in sub-linear time

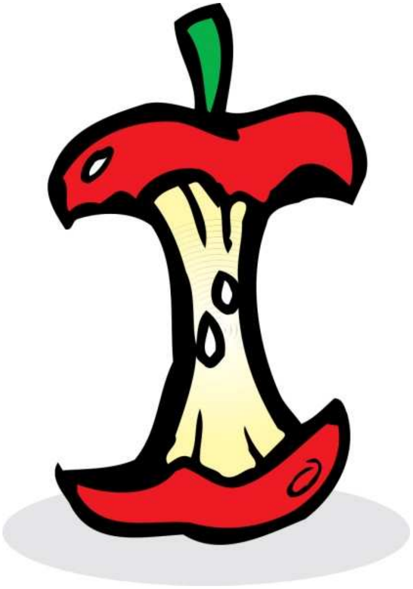
Key Idea: Replace many points by one weighted representative:

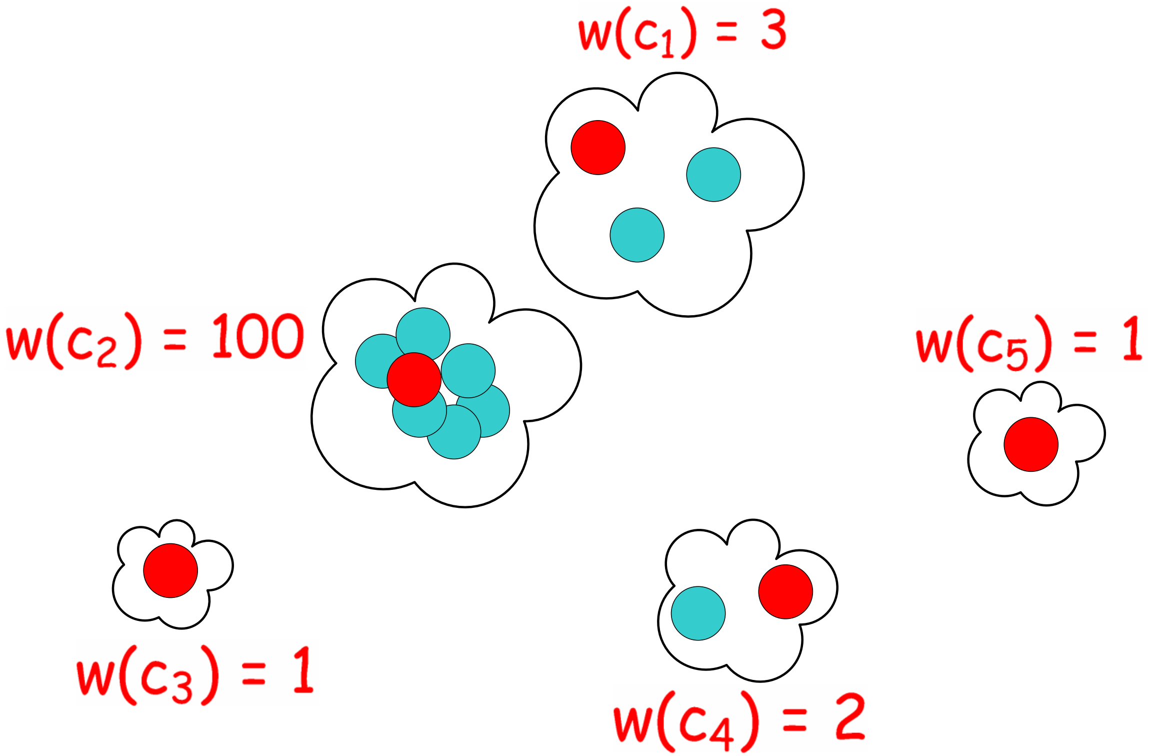


(k, ε)-Median Coreset

No privacy

Key Idea: Replace many points by one weighted representative:

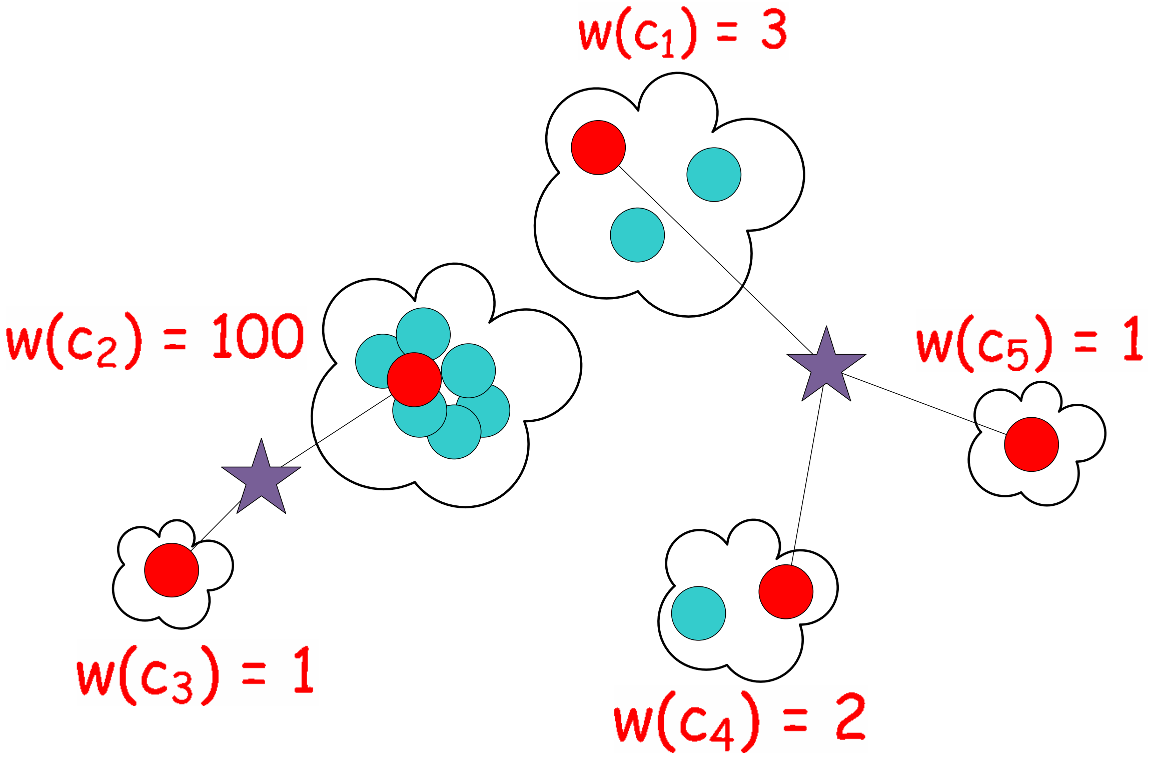




(k, ε)-Median Coreset

No privacy

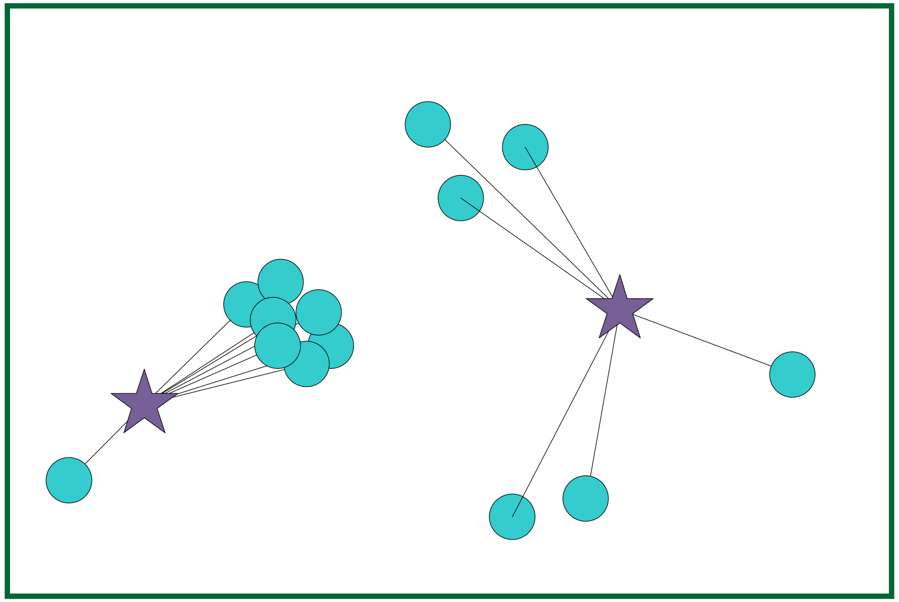
Key Idea: Replace many points by one weighted representative:

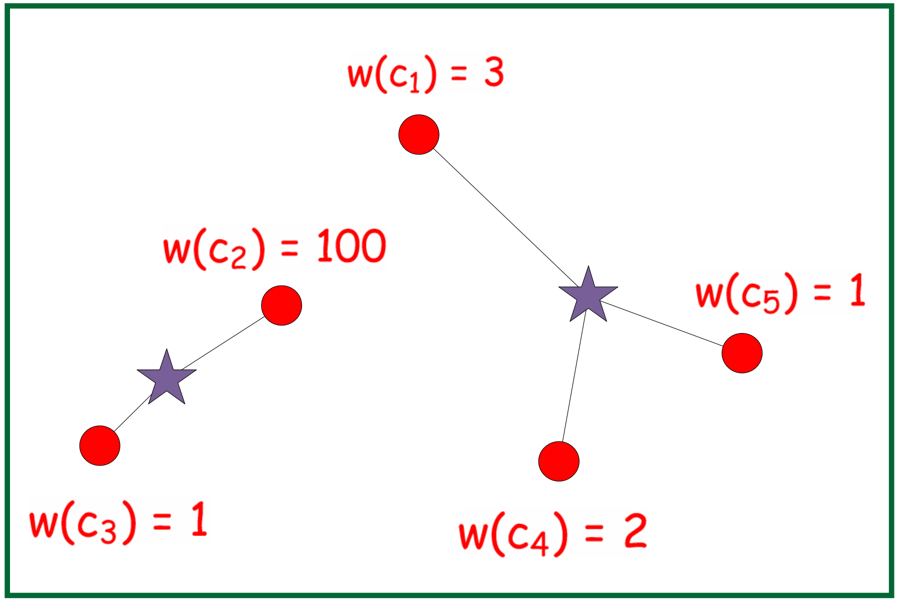


(k, ε)-Median Coreset

No privacy

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





Definition

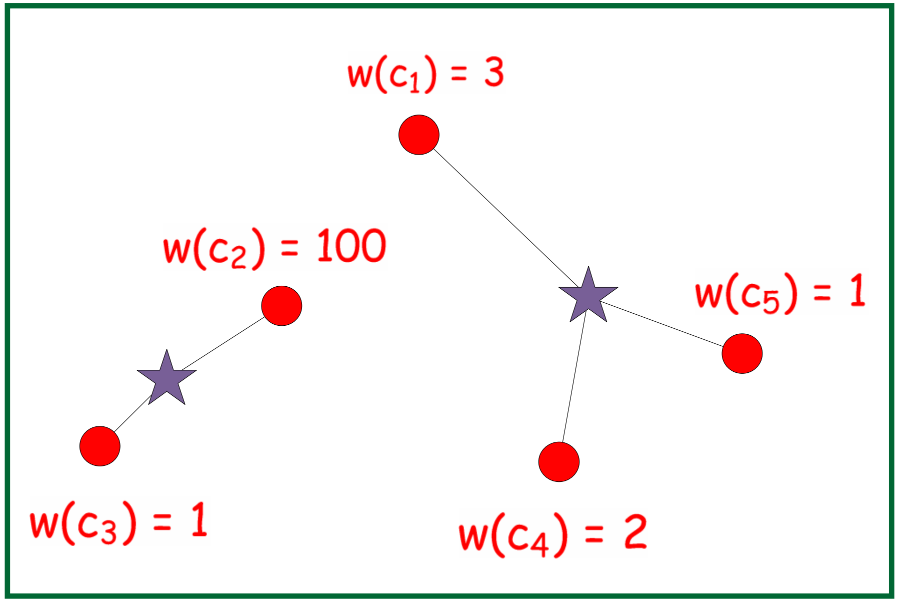
No privacy

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

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

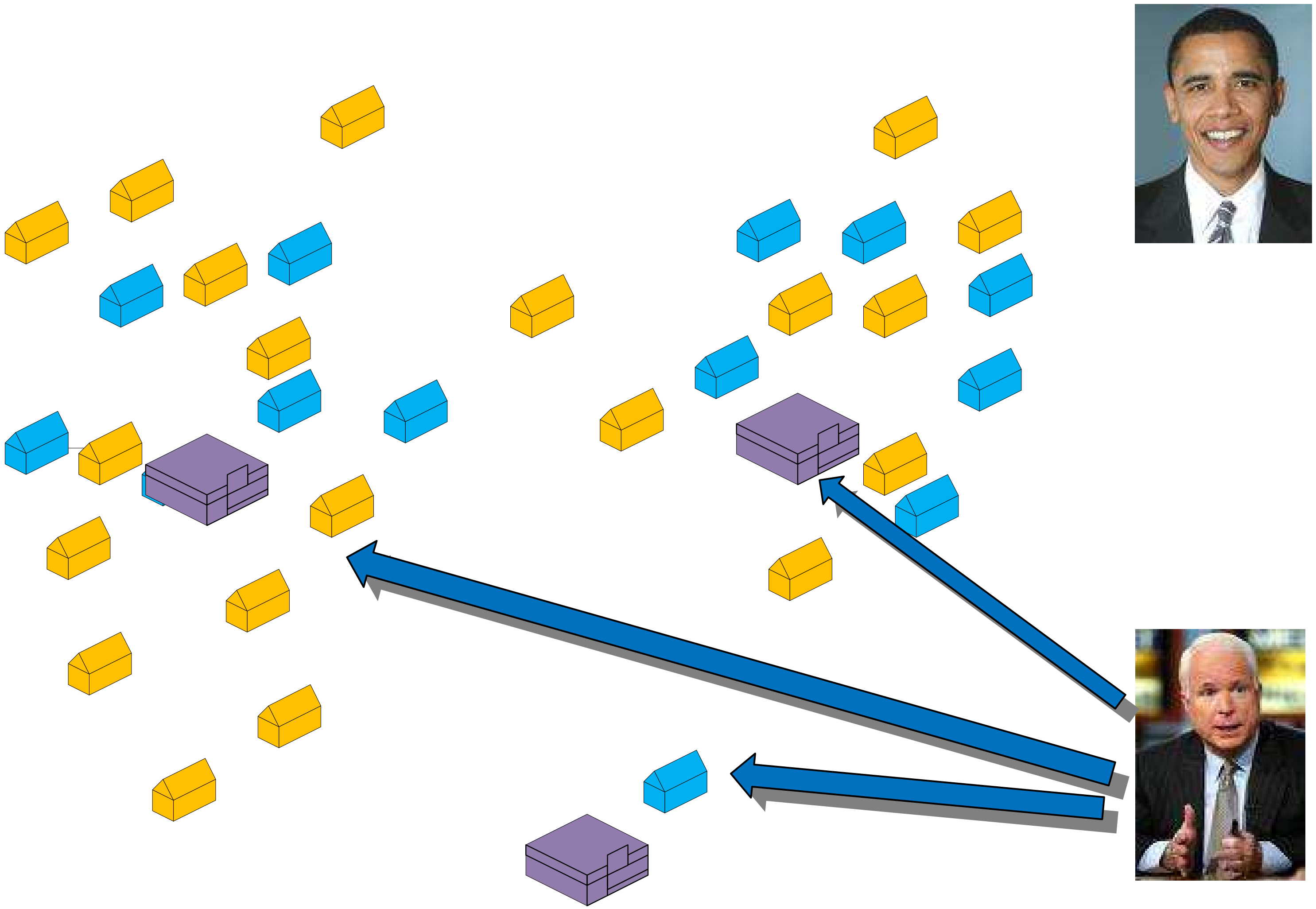
Multiplicative error $\leq 1 + \varepsilon$

Additive error $\leq \frac{1}{\varepsilon}$



Locating Branch Offices

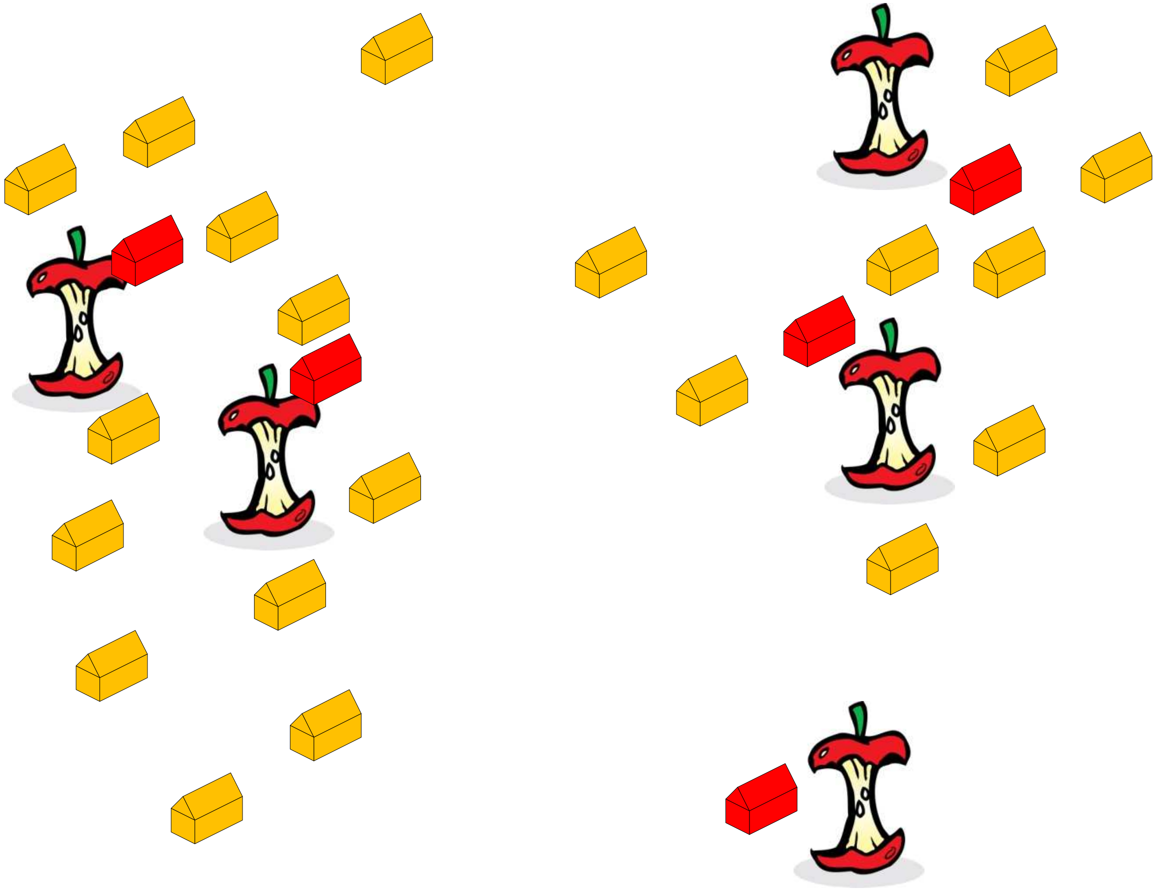
No privacy



Private (Republican) Coresets

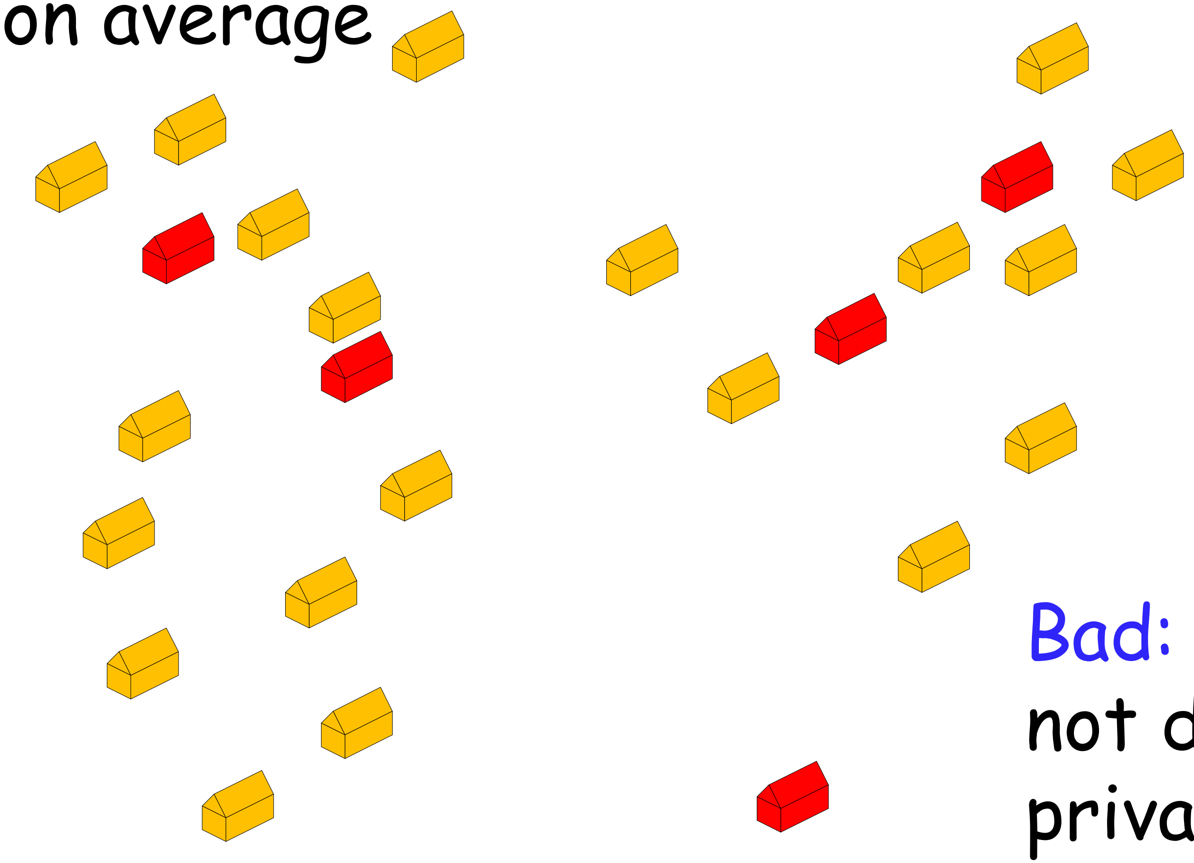
No privacy

Intuition: Coresets reveal little information



Coresets & Privacy

Good: Coresets reveal little information on average



Bad: Coresets are not differential private

Private Coreset Scheme

An algorithm that:

- is α-differentially private.
- for $P \subseteq [0,1]^d$, outputs a (k, ε)-coreset, w.h.p.

Our Contributions

1. [Simple, non-constructive]:

k-median coreset → Private k-median coreset

k-mean coreset → Private k-mean coreset

. . .

Using Exp. Mechanism of [MT07]

Our Contributions

2. [Constructive, linear time]:

    - Private k-median coreset
    - Private k-mean coreset

Our Contributions

2. [Constructive, linear time]:

   - Private k-median coreset
   - Private k-mean coreset

3. Lower bound tradeoffs on multiplicative-additive approximation for private coresets

Applications

- Private k-median clustering
- Comparing alternatives privately
- Private streaming algorithms
- Approximately truthful mechanisms [MT07]

Related Work

- Sanitized Database [BLR08]
- (Non-private) coresets for k-median [HM04][HK05][FS05][Chen06][[FMS07]
- Private clustering [BDMN05][NRS07]

Overview

- Private coreset for 1-median, P on line .

Overview

- Private coreset for 1-median, P on line .

↓

- Private coreset for 1-median, P in $[0,1]^d$

Overview

- Private coreset for 1-median, P on line .
- Private coreset for 1-median, P in $[0,1]^d$
- Private bi-criteria approximation for k-median

Overview

- Private coreset for 1-median, P on line .

↓

- Private coreset for 1-median, P in $[0,1]^d$

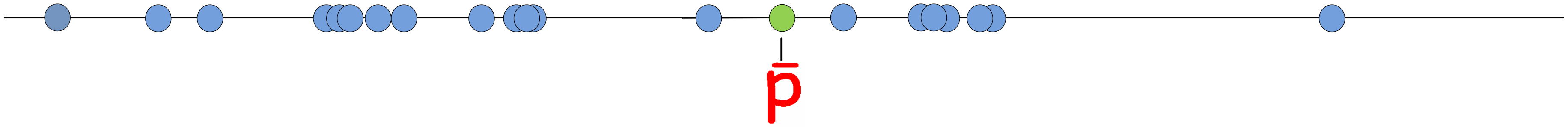
+

- Private bi-criteria approximation for k-median

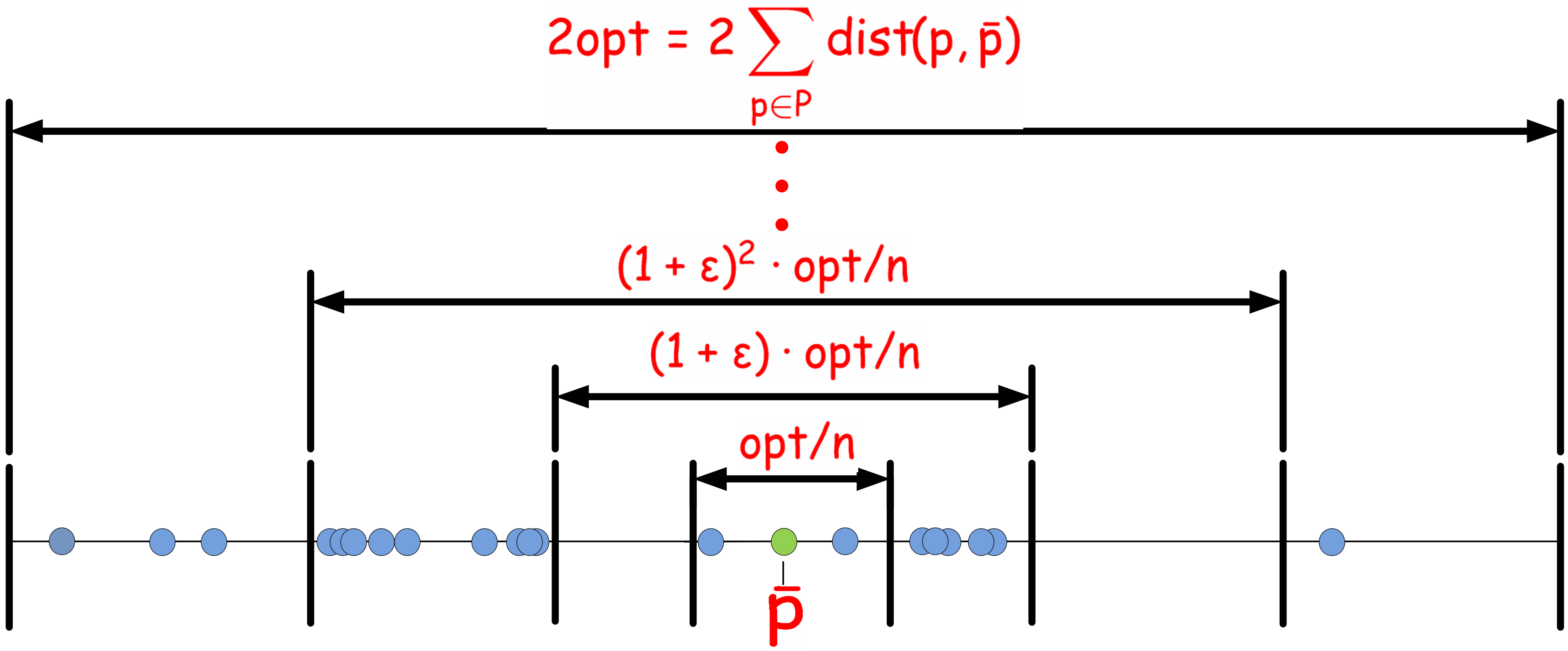
=

- Private coresets for k-median, $P \subseteq [0,1]^d$

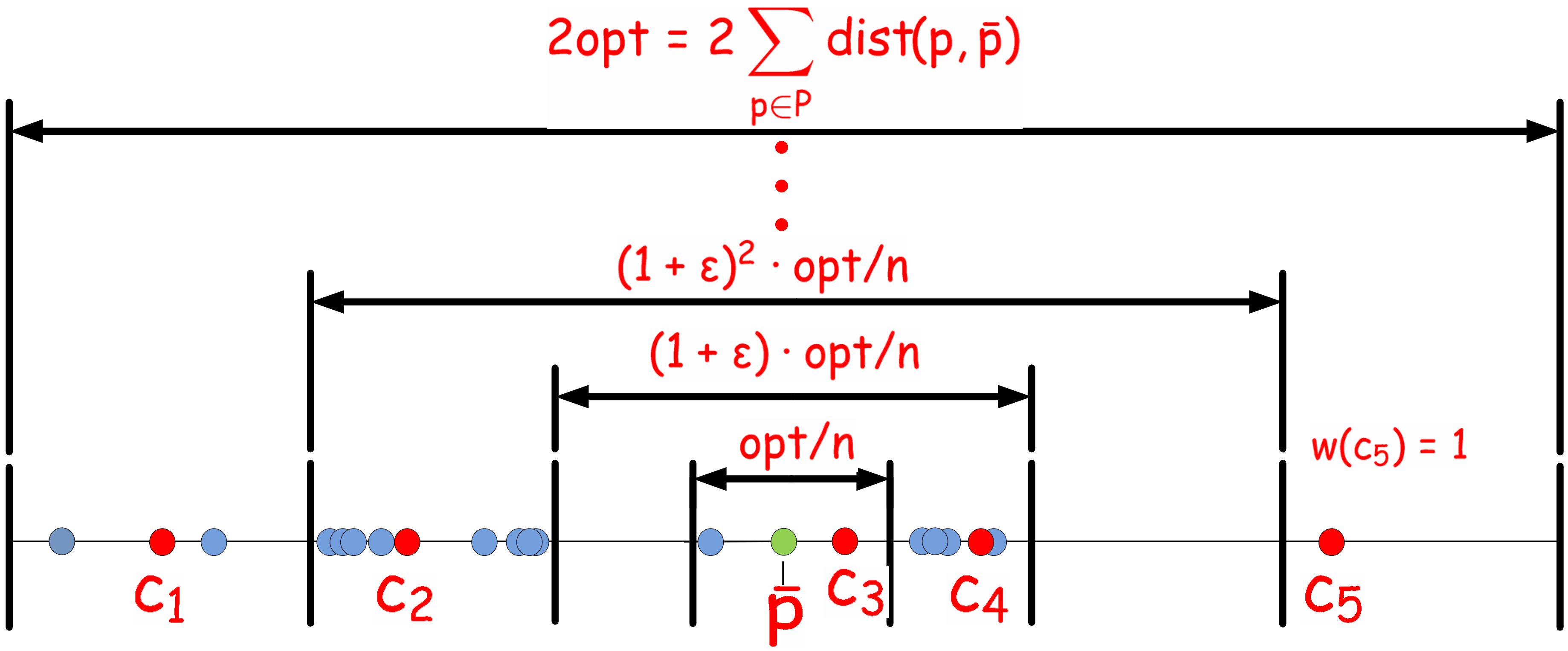
Coreset for $P \subseteq [0,1]$, $k = 1$ [HM04]



Coreset for $P \subseteq [0,1]$, $k = 1$ [HM04]


$2\text{opt} = 2\sum_{p\in P} \text{dist}(p,\bar{p})$
$(1+\varepsilon)^2 \cdot \text{opt}/n$
$(1+\varepsilon) \cdot \text{opt}/n$
opt/n
$\bar{p}$

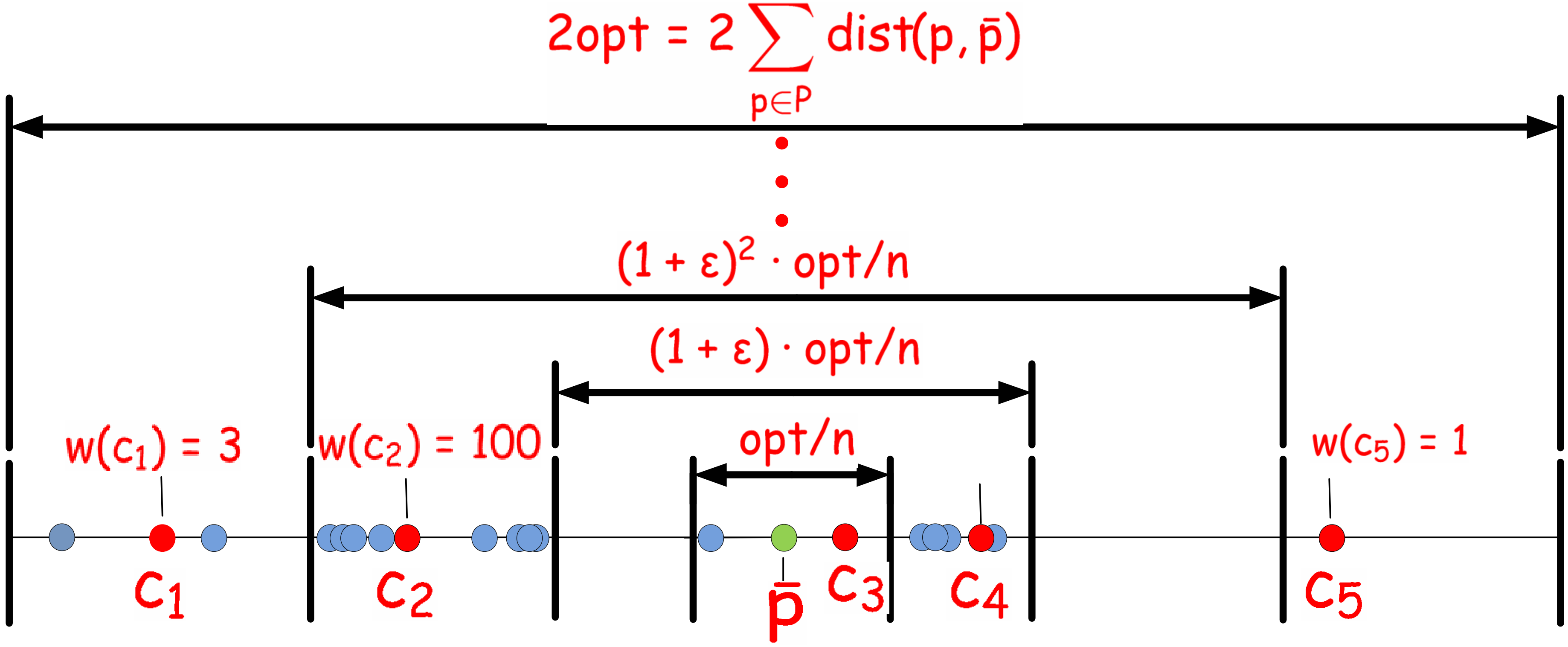
Coreset for $P \subseteq [0,1]$, $k = 1$ [HM04]


$2\text{opt} = 2\sum_{p \in P} \text{dist}(p, \bar{p})$
$(1+\varepsilon)^2 \cdot \text{opt}/n$
$(1+\varepsilon) \cdot \text{opt}/n$
opt/n
$w(c_5) = 1$
c_1
c_2
$\bar{p}$
c_3
c_4
c_5

Coreset for $P \subseteq [0,1]$, $k = 1$ [HM04]

For each interval I:

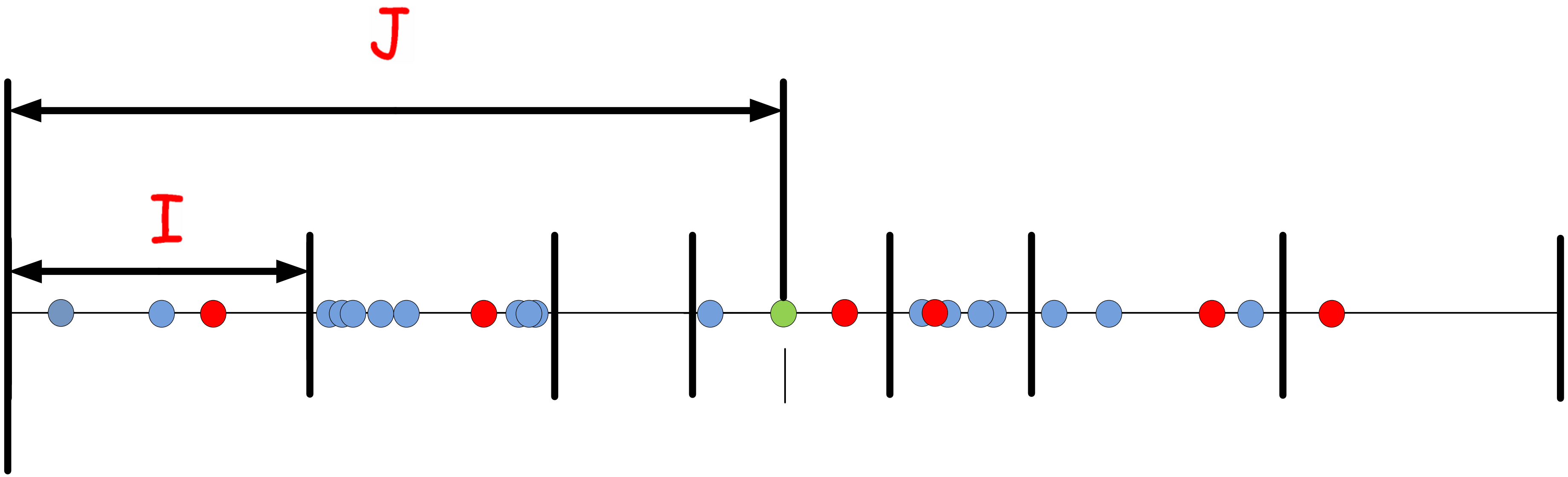
- Choose an arbitrary representative $c \in P \cap I$
- $w(c) \leftarrow |P \cap I|$



$2\text{opt} = 2\sum_{p \in P} \text{dist}(p, \bar{p})$
$(1+\varepsilon)^2 \cdot \text{opt}/n$
$(1+\varepsilon) \cdot \text{opt}/n$
opt/n
$w(c_1) = 3$
$w(c_2) = 100$
$w(c_5) = 1$
c_1
c_2
$\bar{p}$
c_3
c_4
c_5

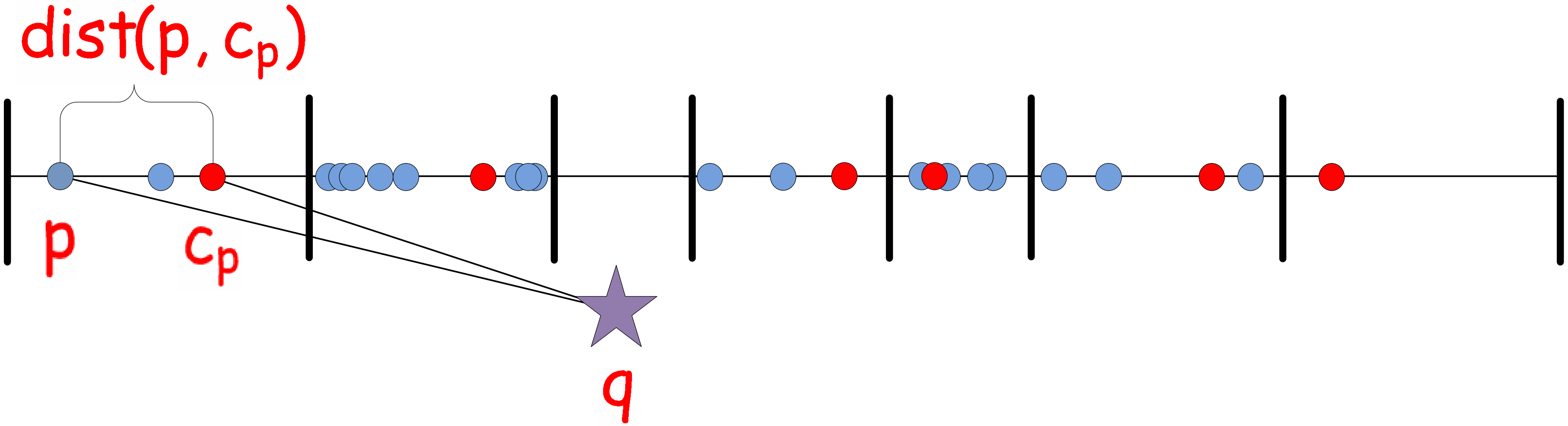
Main Observation: $|I| \leq \varepsilon|J|$

Because the size of the intervals forms a geometric sequence of ratio $(1 + \varepsilon)$

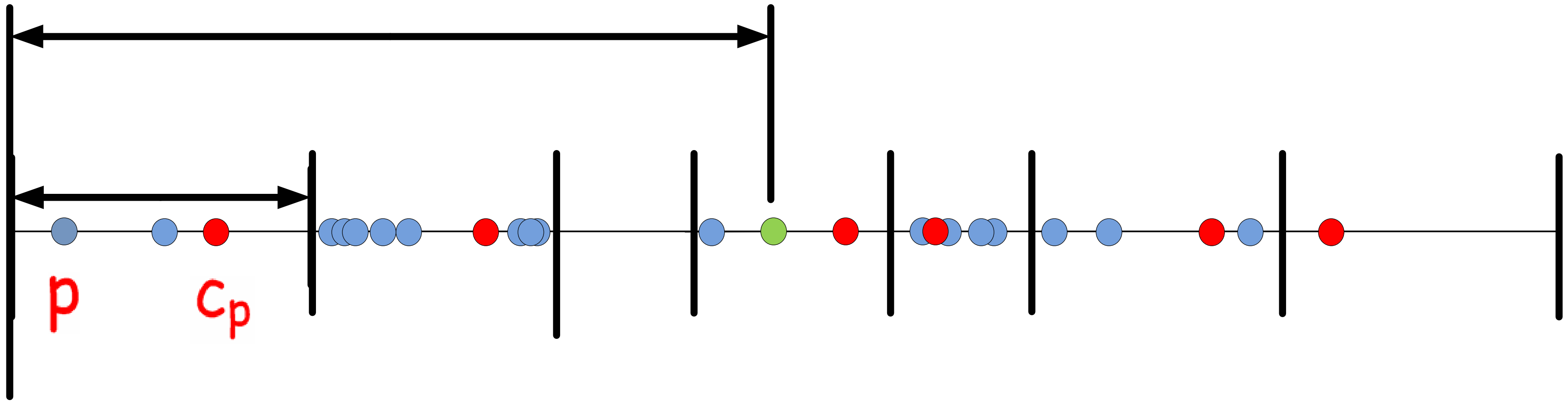


$$\mathrm{error}(q) = \left| \sum_{p \in P} \mathrm{dist}(p, q) - \sum_{p \in P} \mathrm{dist}(c_p, q) \right|$$

$$\le \sum_{p \in P} \mathrm{dist}(p, c_p)$$



$$
\begin{aligned}
\mathrm{error}(q) &= \left| \sum_{p \in P} \mathrm{dist}(p, q) - \sum_{p \in P} \mathrm{dist}(c_p, q) \right| \\
&\leq \sum_{p \in P} \mathrm{dist}(p, c_p) \leq \sum_{p \in P} \varepsilon \cdot \mathrm{dist}(p, \bar{p})
\end{aligned}
$$

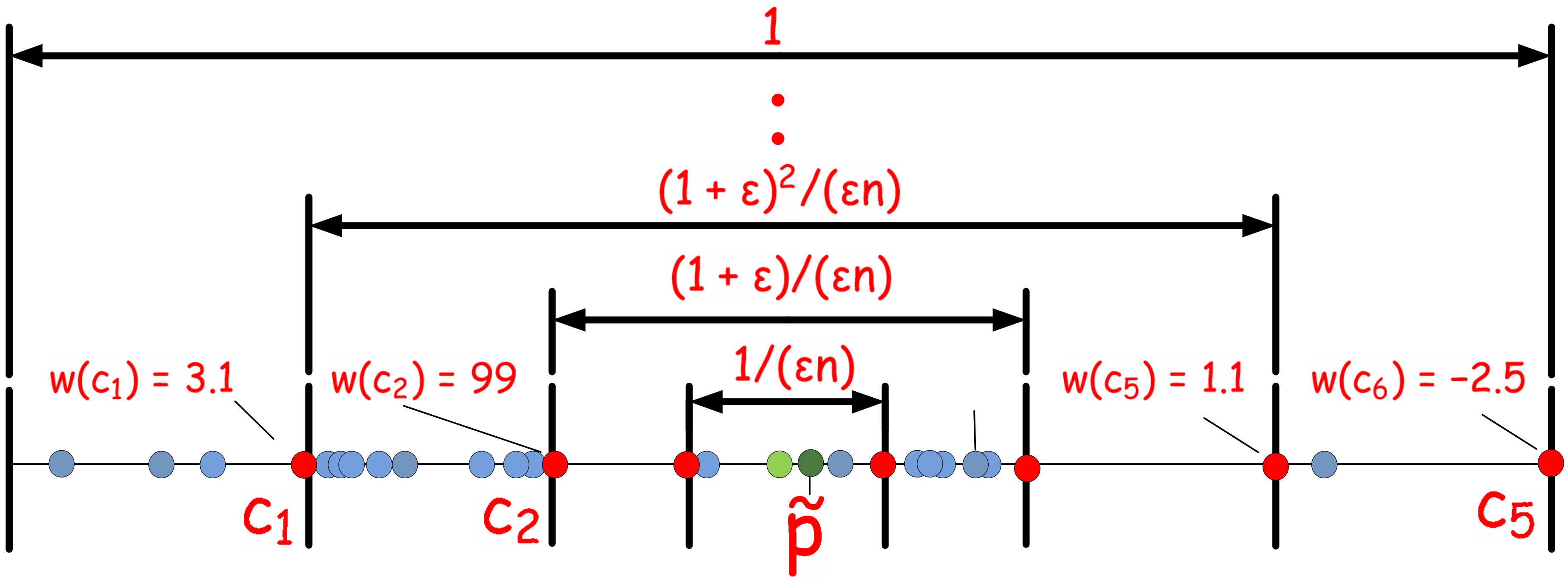


$$
\begin{aligned}
\mathrm{error} &= \left| \sum_{p\in P} \mathrm{dist}(p,q) - \sum_{p\in P} \mathrm{dist}(c_p,q) \right| \\
&\le \sum_{p\in P} \mathrm{dist}(p,c_p) \le \sum_{p\in P} \varepsilon \cdot \mathrm{dist}(p,\bar{p}) \\
&\le 2\varepsilon \cdot \mathrm{opt} \\
&\le 2\varepsilon \sum_{p\in P} \mathrm{dist}(p,q)
\end{aligned}
$$

New: Private Coreset

For each interval I:

- Choose the rightmost point $c \in I$
- $\tilde{w}(c) \leftarrow |P \cap I| + \text{Noise}$



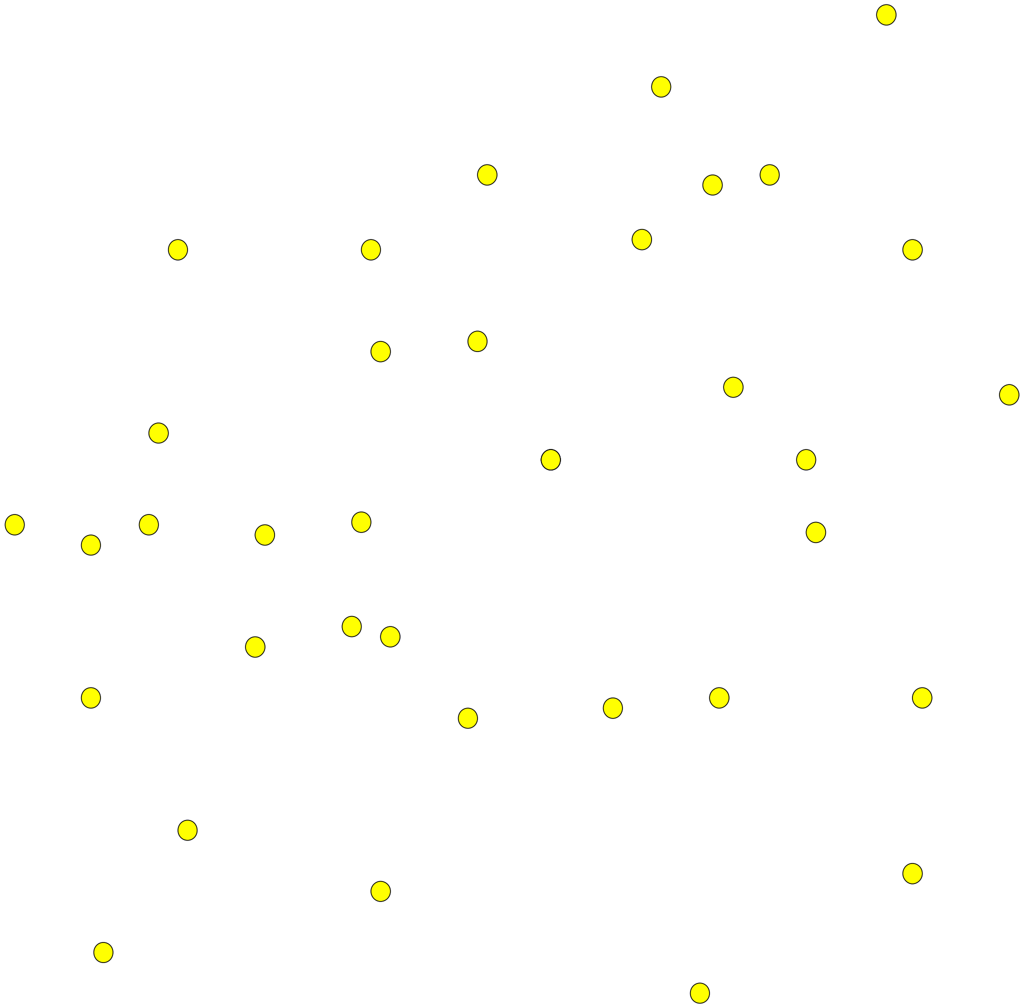
Coreset for $P \subseteq [0,1]$, $k = 1$ [HM04]

$$\left| \sum_{p \in P} \mathrm{dist}(p,q) - \sum_{c \in C} w(c) \cdot \mathrm{dist}(c,q) \right| \leq \varepsilon \sum_{p \in P} \mathrm{dist}(p,q)$$

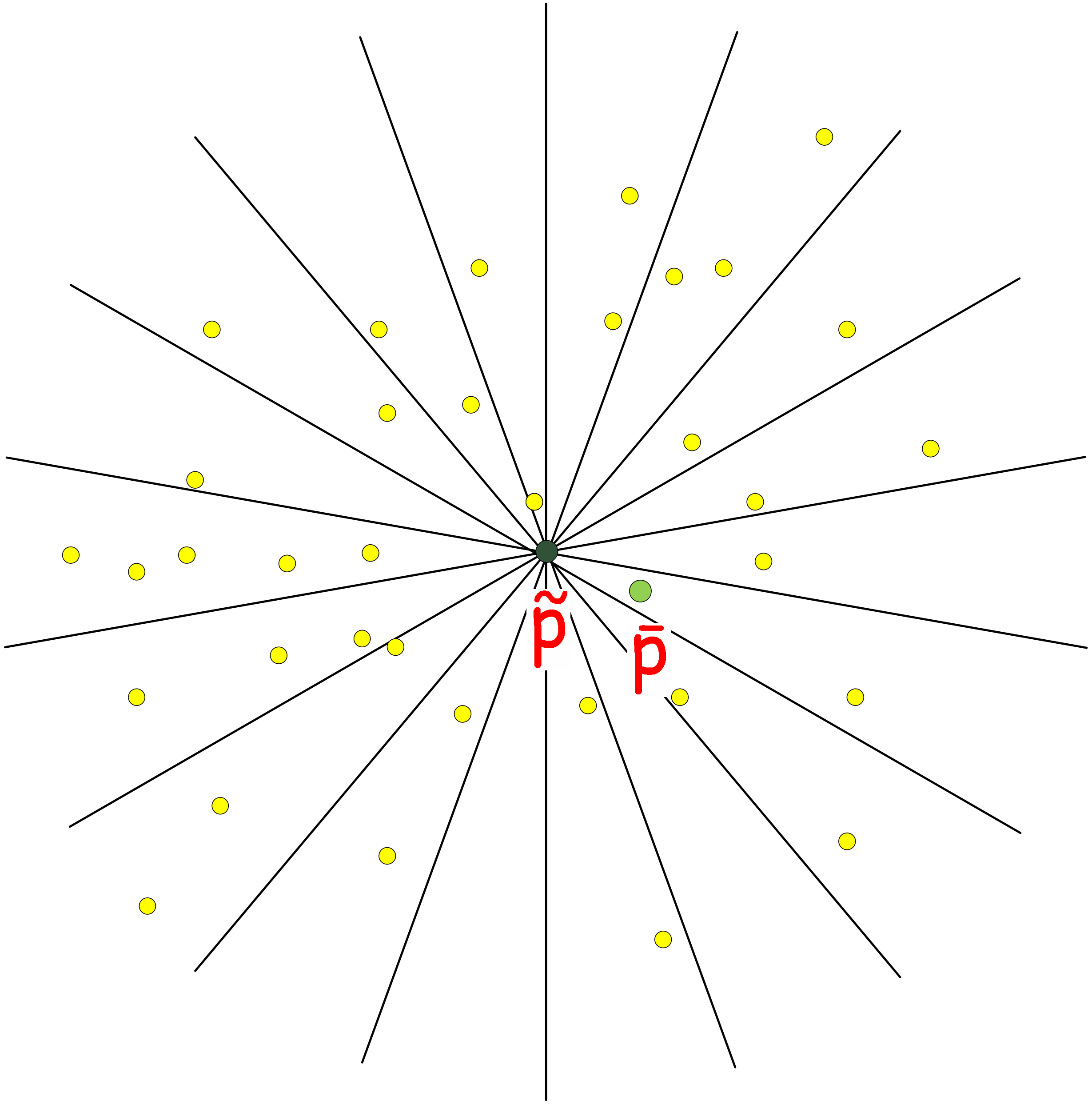
New: Private Coreset

$$\left| \sum_{p \in P} \mathrm{dist}(p,q) - \sum_{c \in C} w(c) \cdot \mathrm{dist}(c,q) \right| \leq \varepsilon \sum_{p \in P} \mathrm{dist}(p,q) + O\left(\frac{1}{\varepsilon}\right)$$

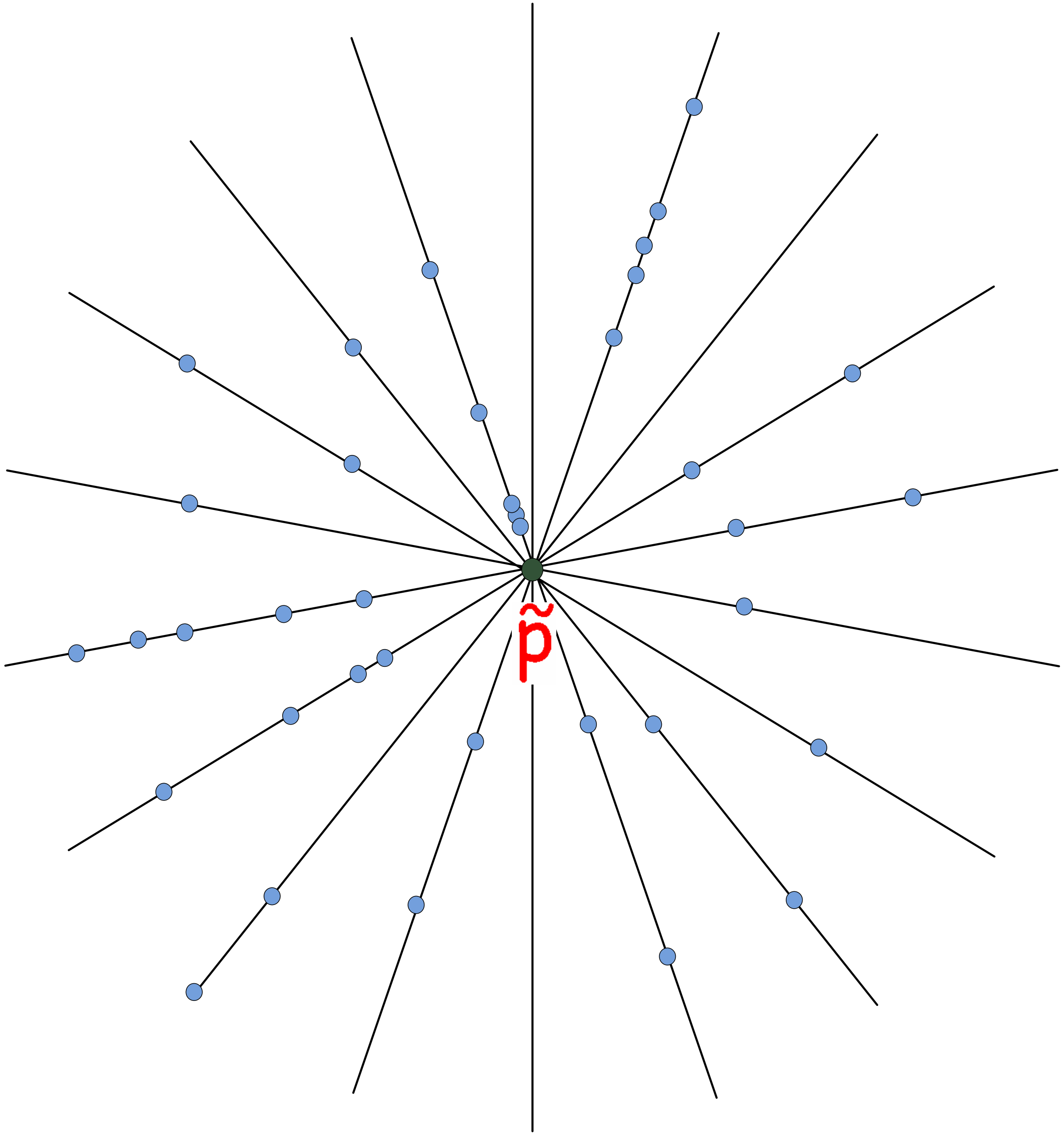
Generalization for $P \subseteq [0,1]^d$



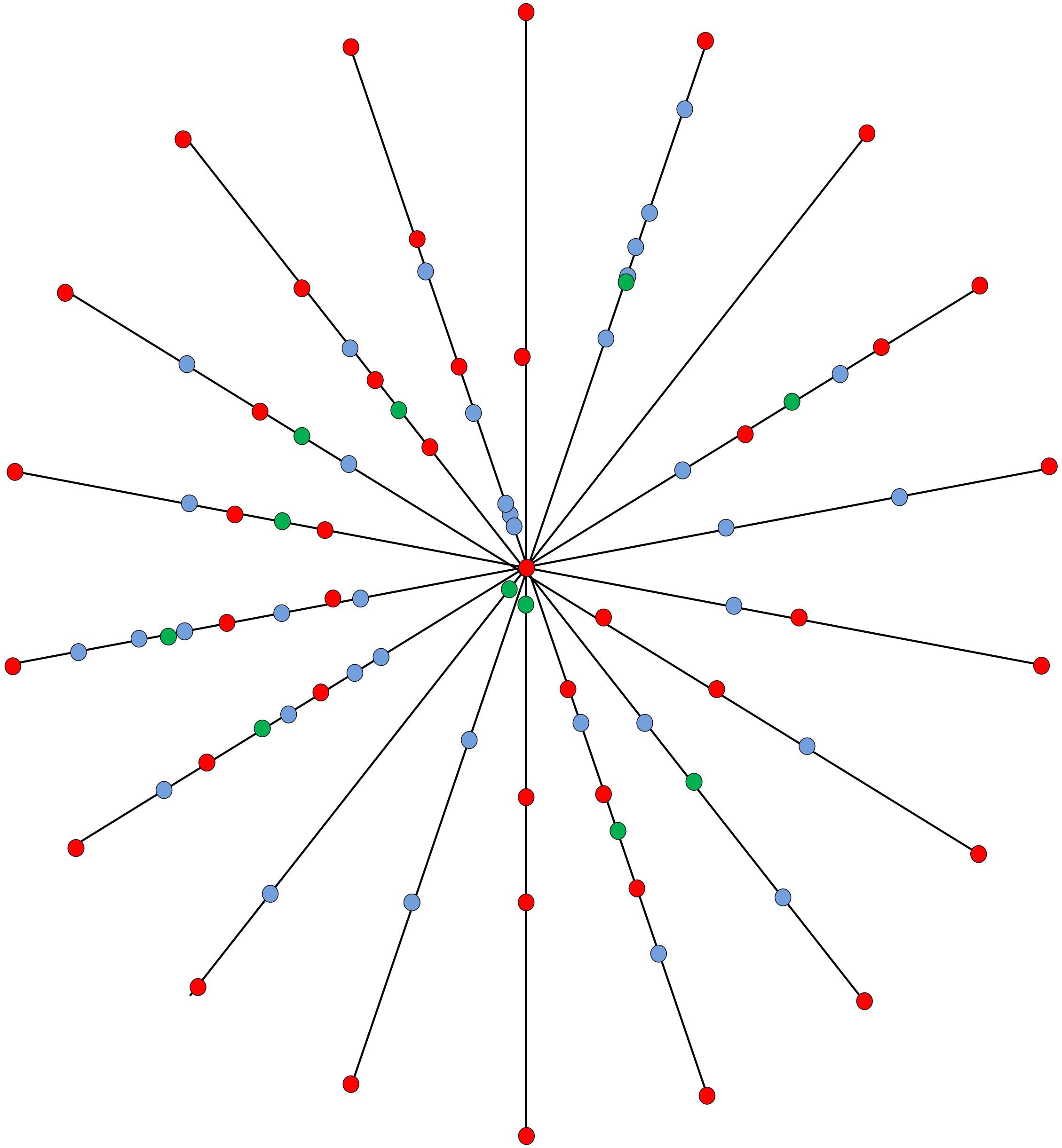
Generalization for $P \subseteq [0,1]^d$



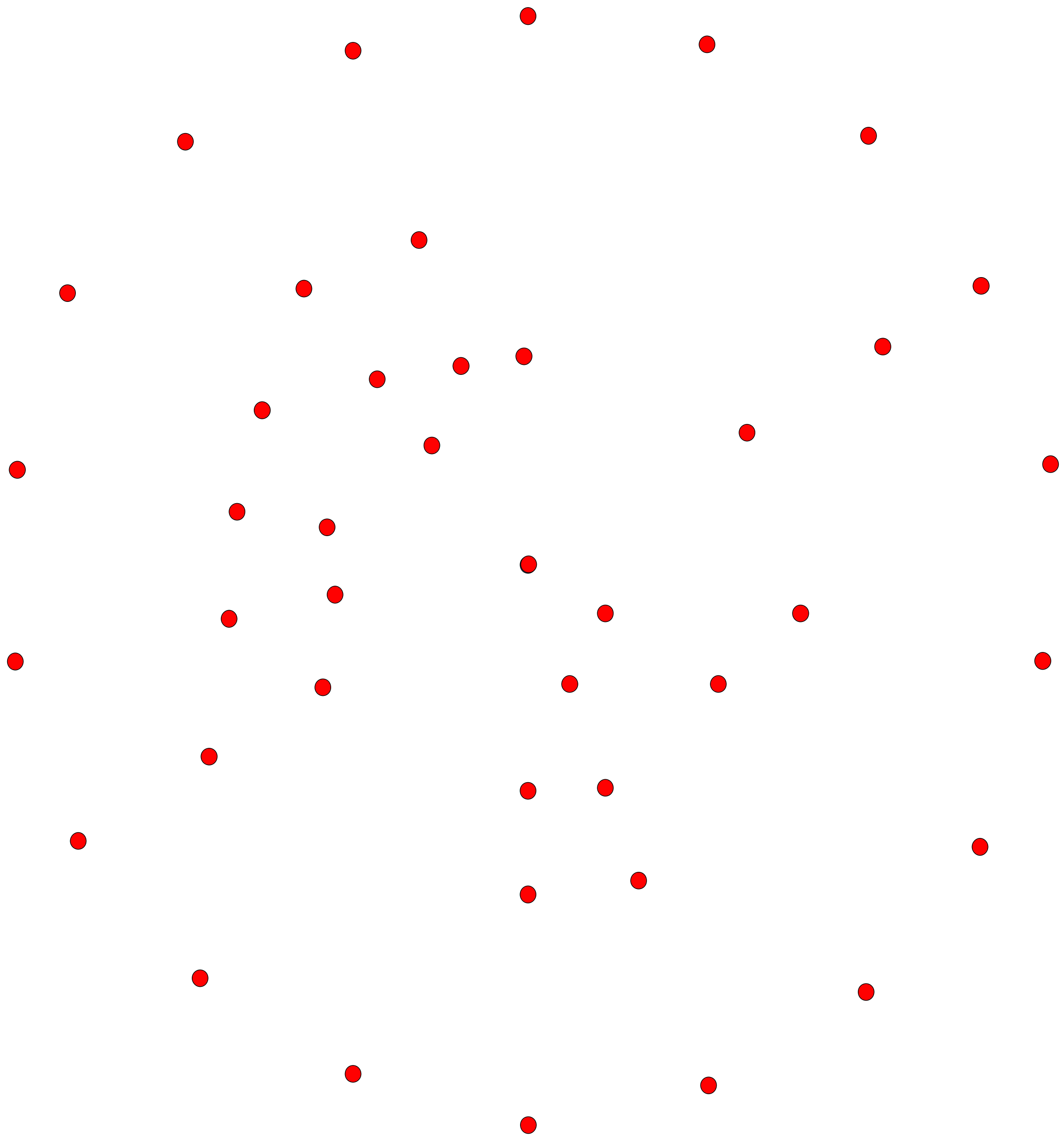
Generalization for $P \subseteq [0,1]^d$



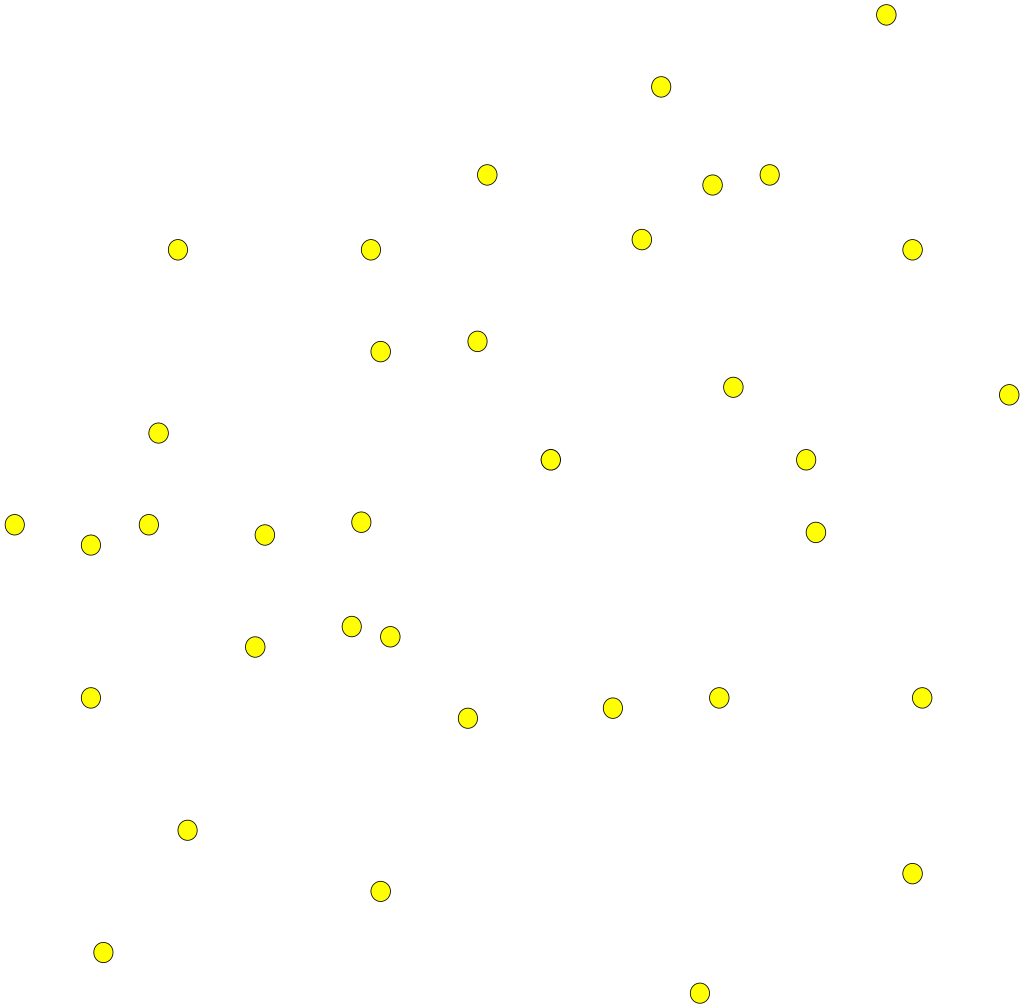
Generalization for $P \subseteq [0,1]^d$



Generalization for $P \subseteq [0,1]^d$

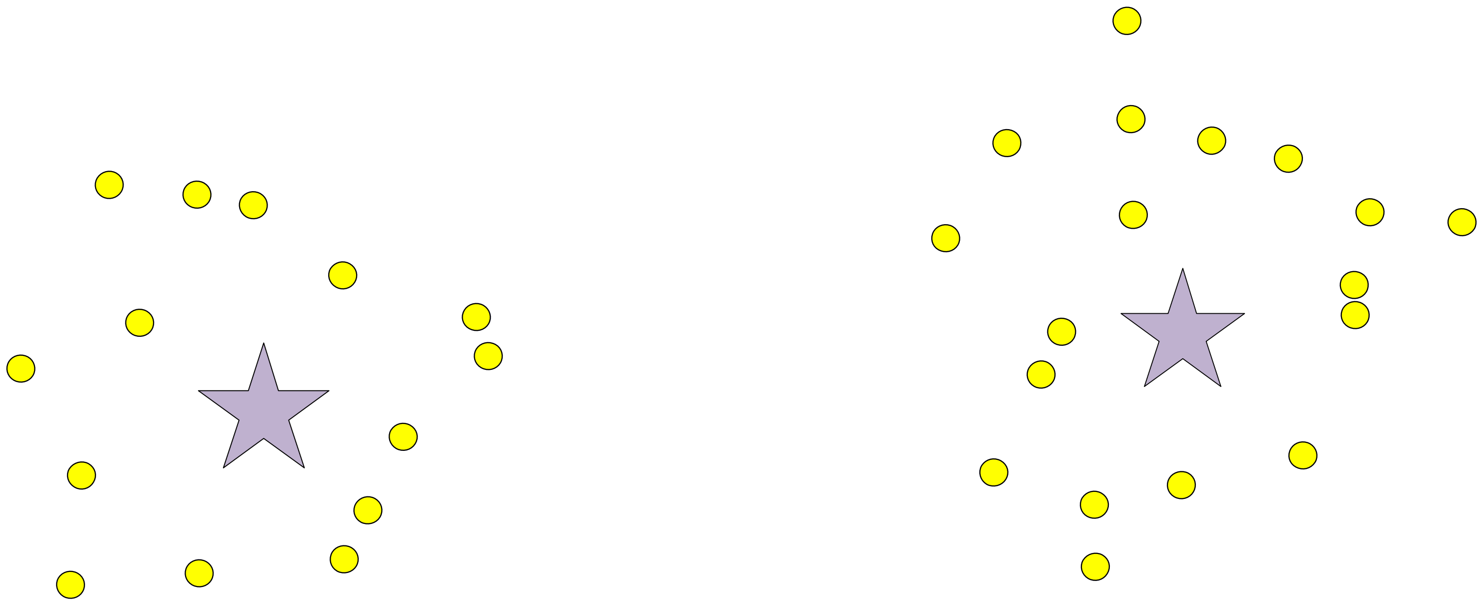


Generalization for k > 1



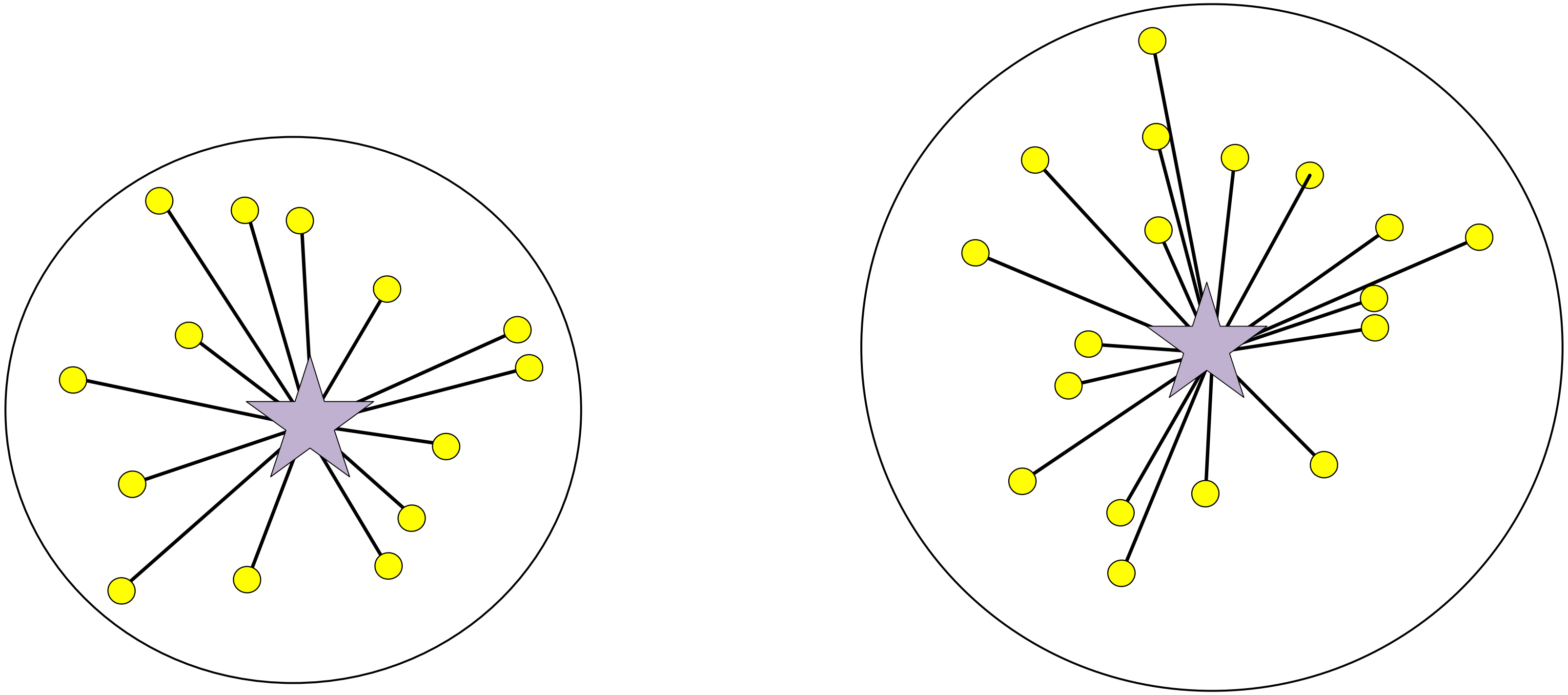
The k-Median of P

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



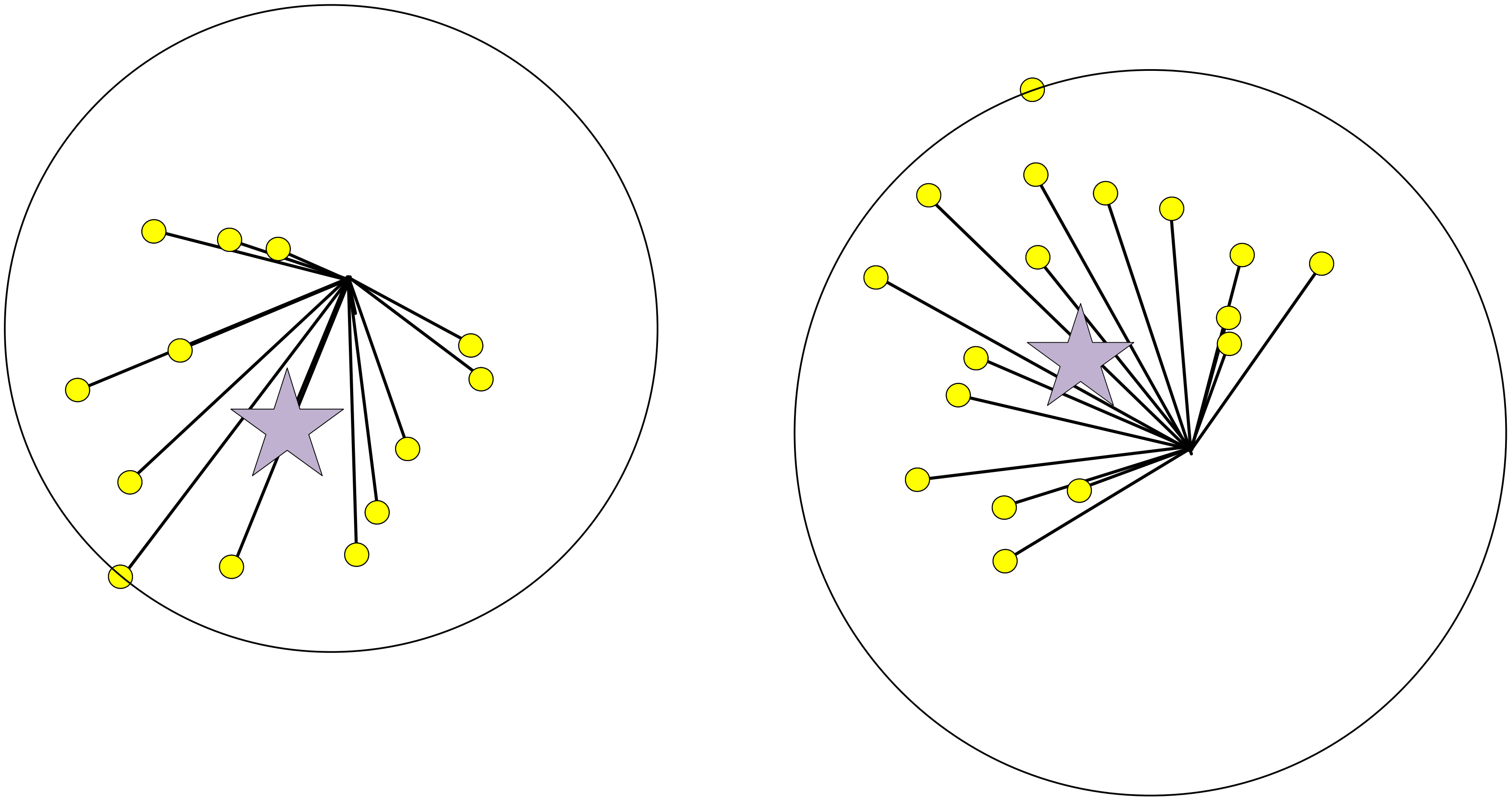
The k-Median of P

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



Constant Approximation

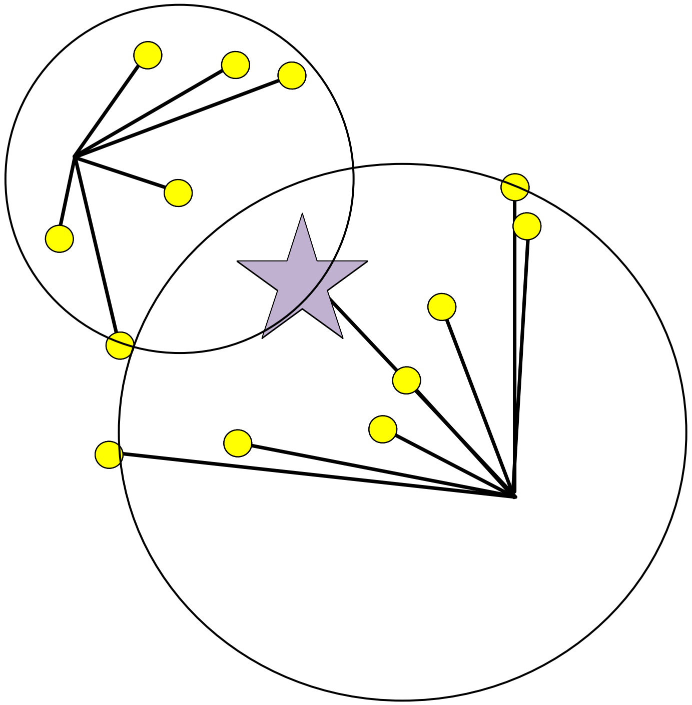
$|\widetilde{OPT}| = k,$ $\sum_{p \in P} \mathrm{dist}(p, \widetilde{OPT}) \leq c \cdot \mathrm{opt}$

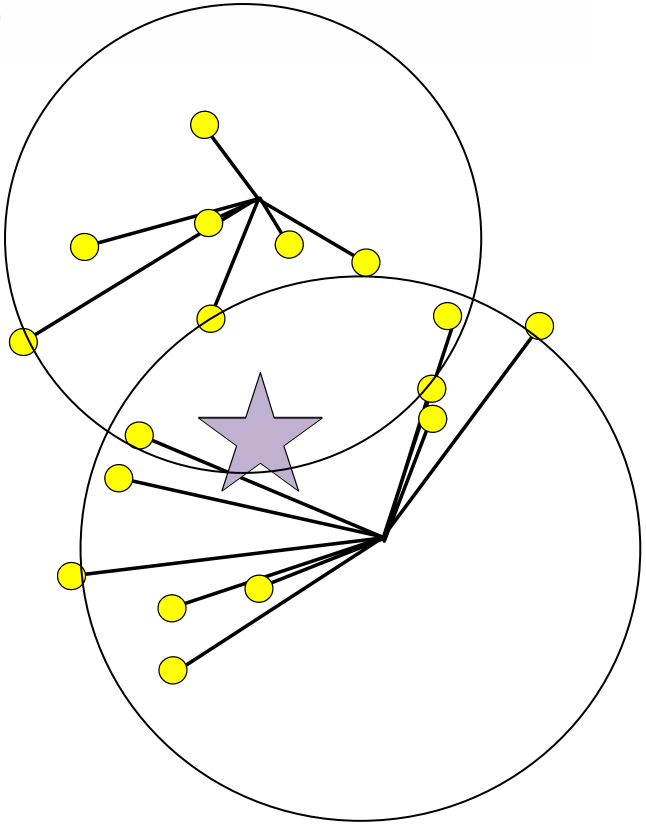


Bi-Criteria Approximation

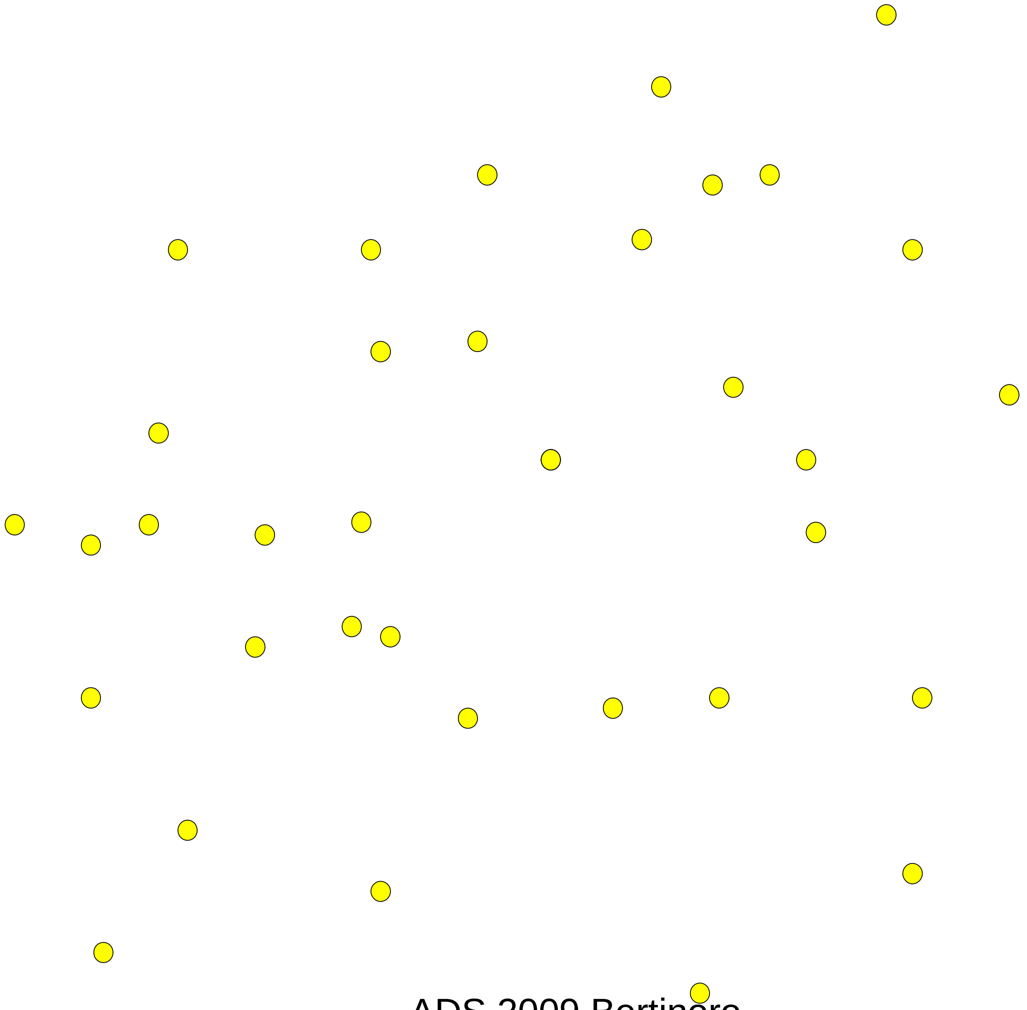
$|B| = O(k \log n),$

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



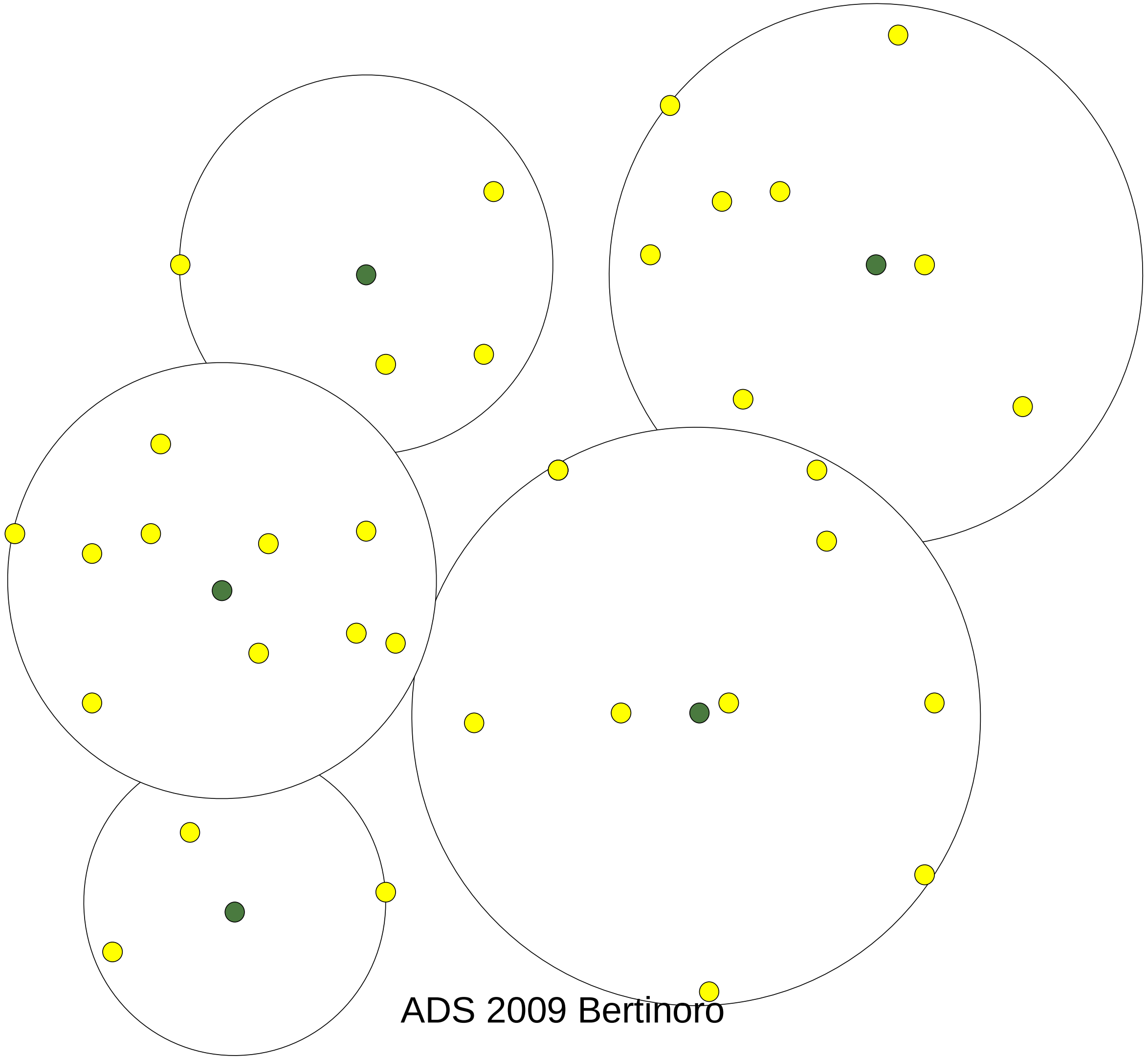


Generalization for k > 1



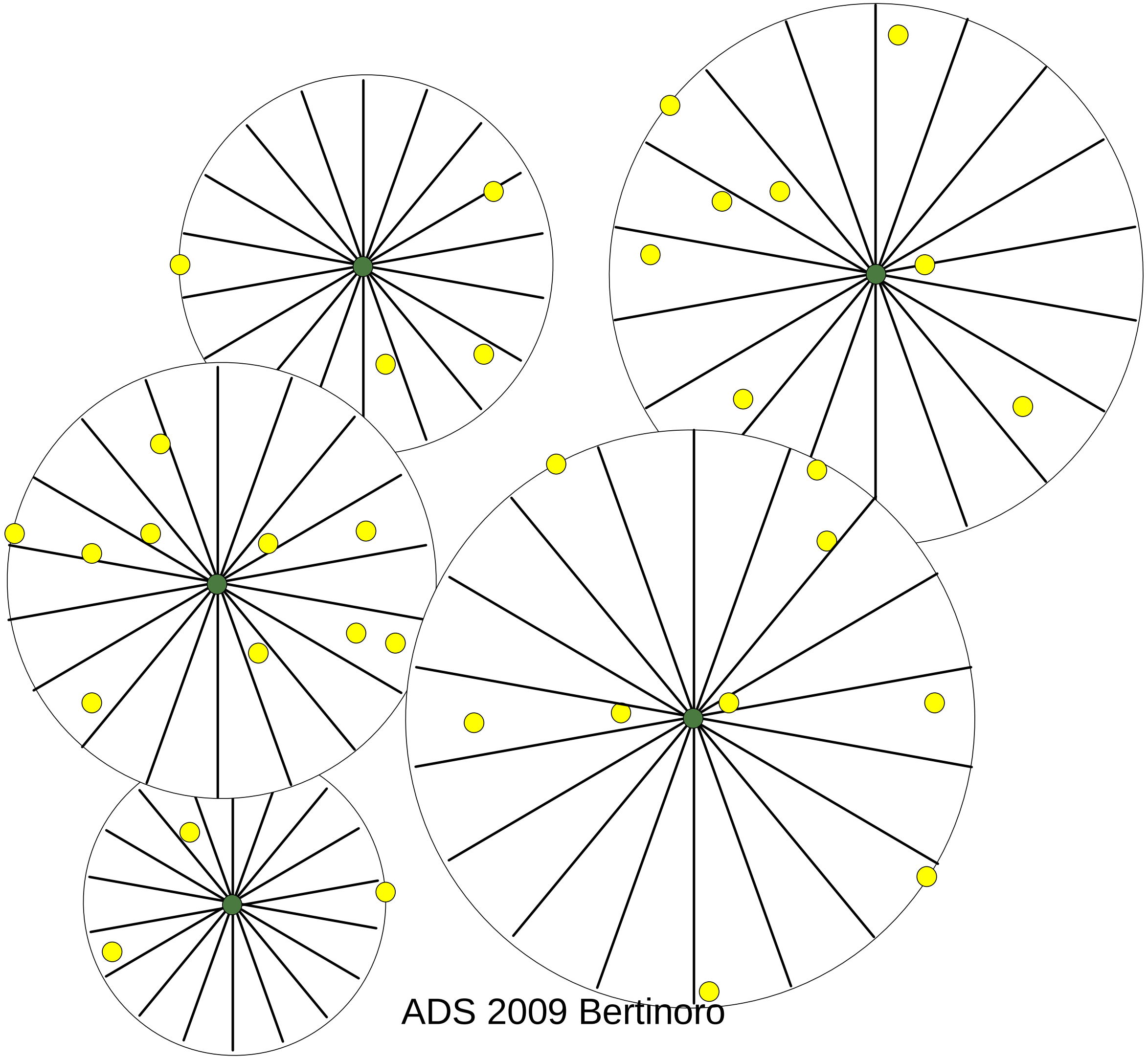
Compute Private Bi-Criteria Approx.

Based on[FFSS07]

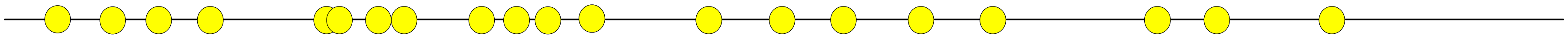


On Each Cluster:

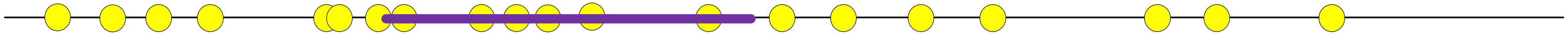
Apply construction for k = 1



Weak ε-Net N for P ⊆ [0,1]



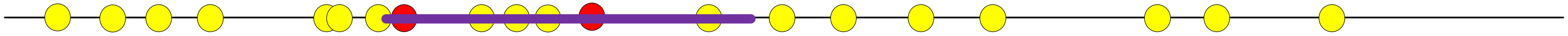
Weak ε-Net N for $P \subseteq [0,1]$



For every interval I:

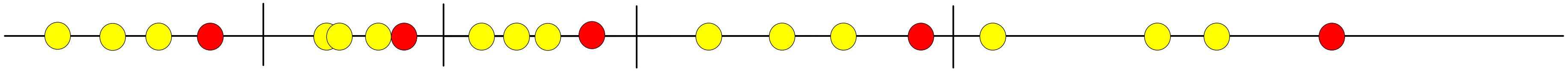
$|I \cap P| \geq \varepsilon n \Rightarrow |I \cap N| \geq 1$

Weak $\frac{1}{4}$-Net N for $P \subseteq [0,1]$

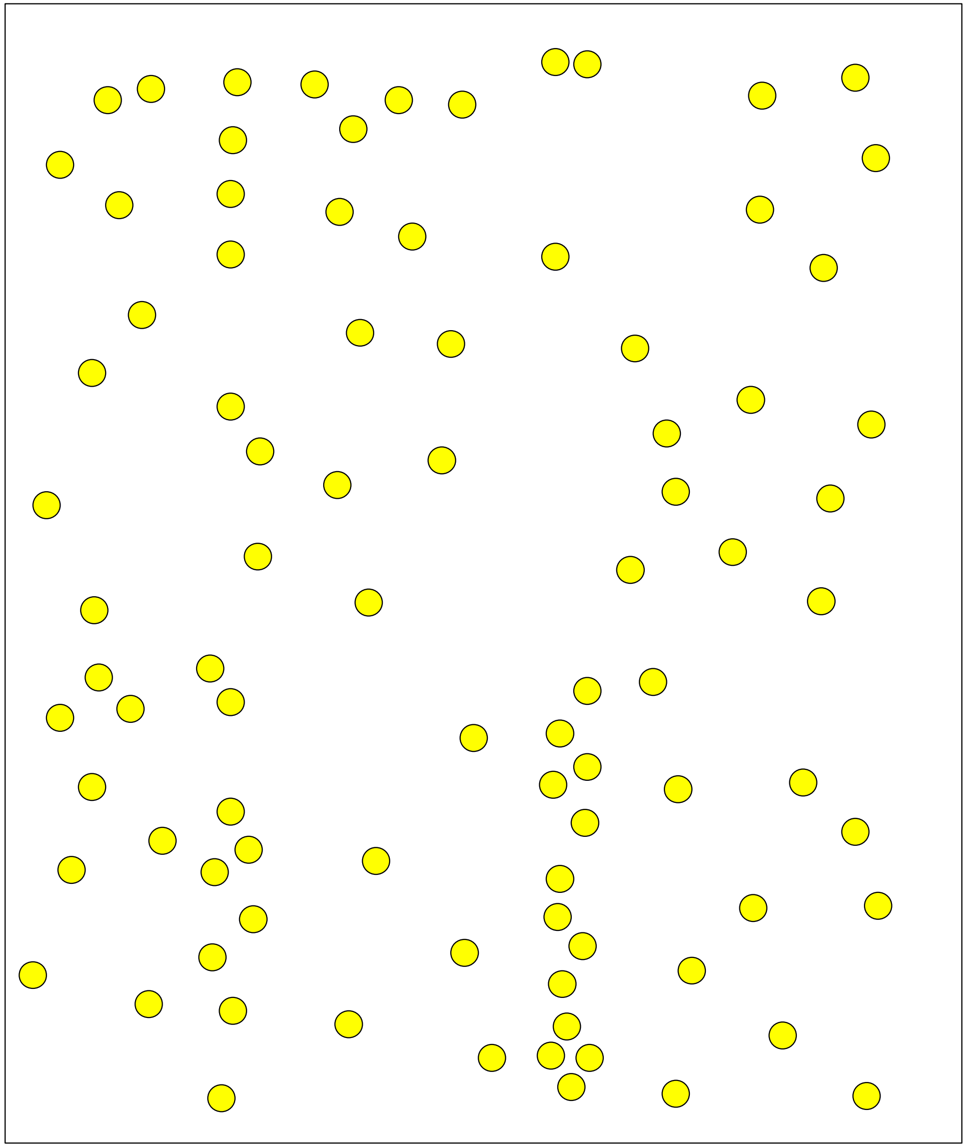


For every interval I:

$|I \cap P| \geq n/4 \Rightarrow |I \cap N| \geq 1$



Weak ε-Net for $P \subseteq [0,1]^d$

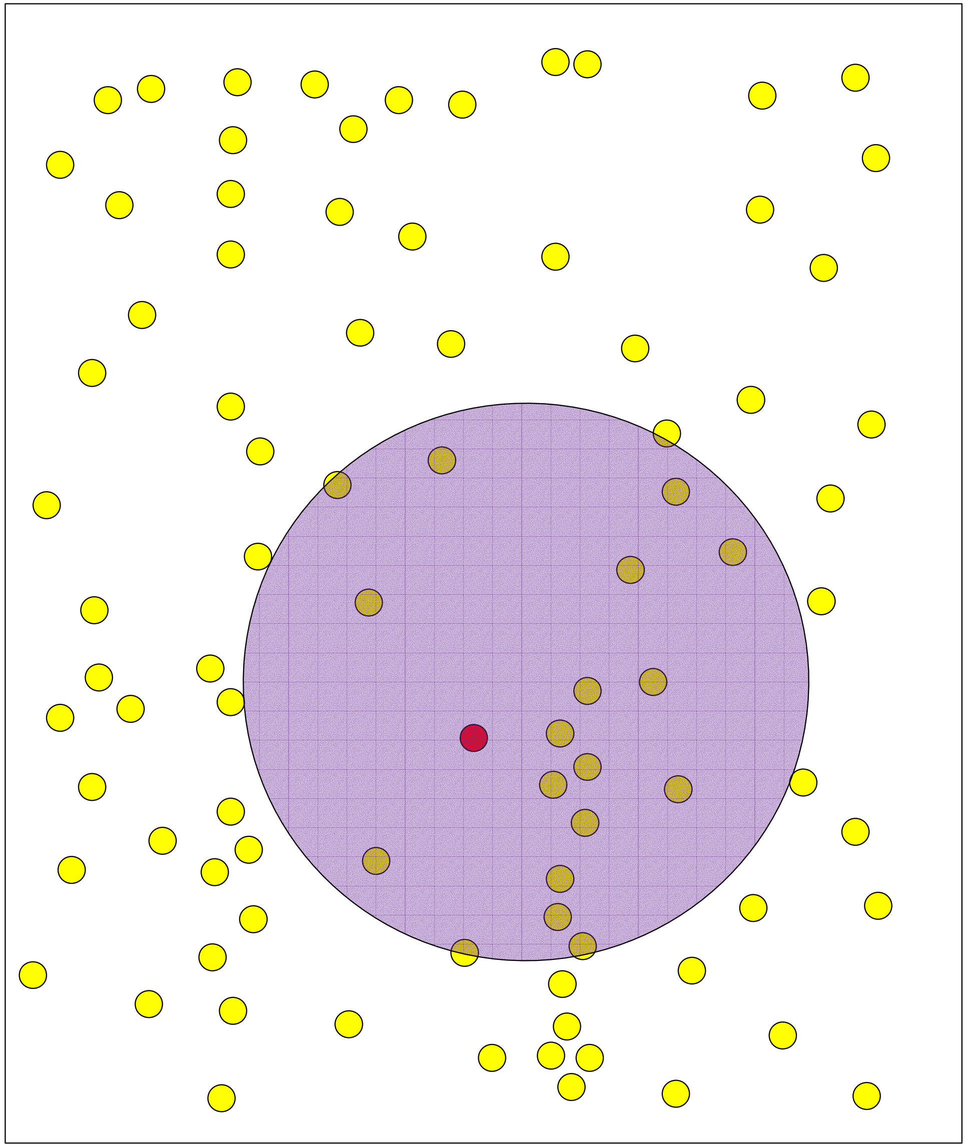


Weak ε-Net for $P \subseteq [0,1]^d$

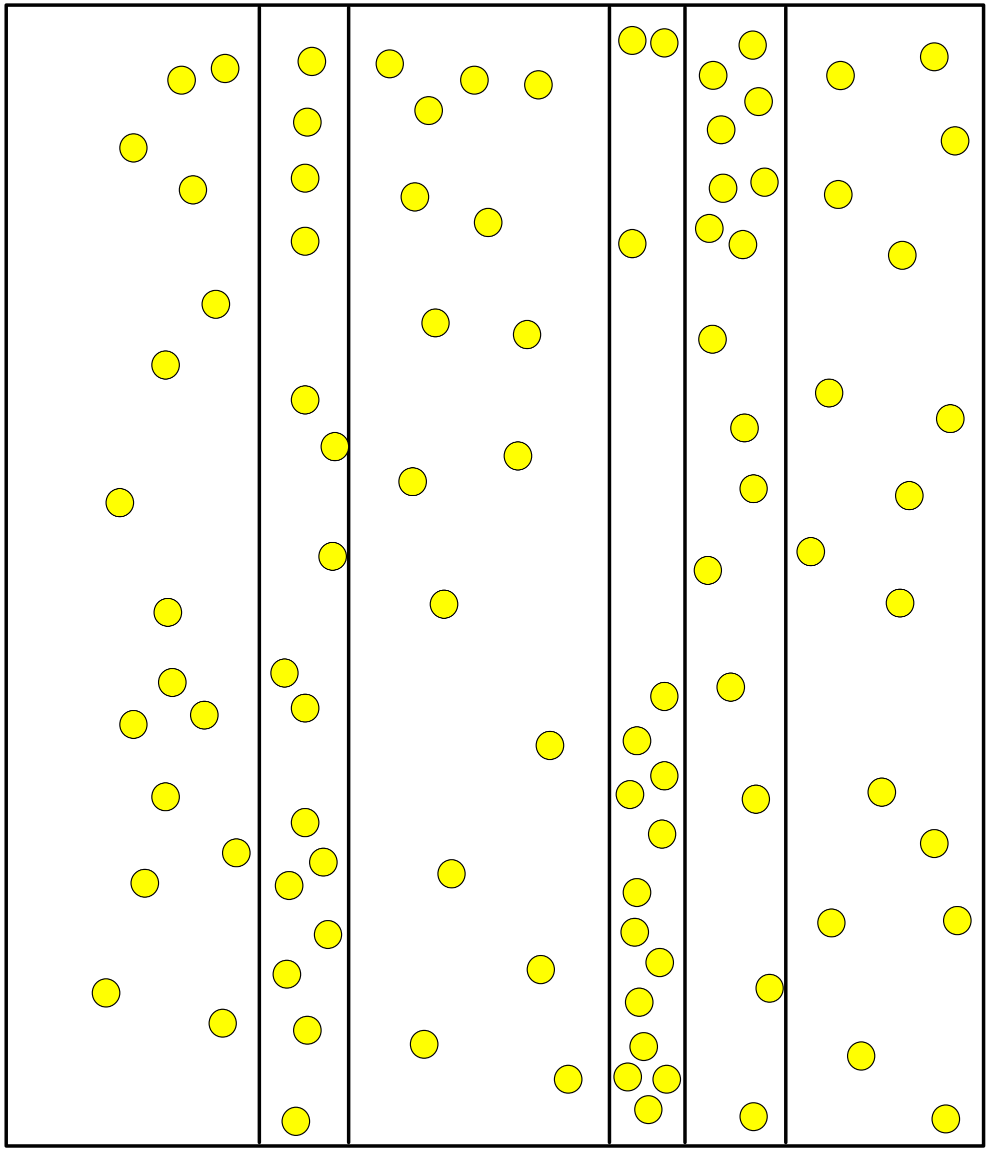
$$|B \cap P| \geq \varepsilon n$$

$$\Downarrow$$

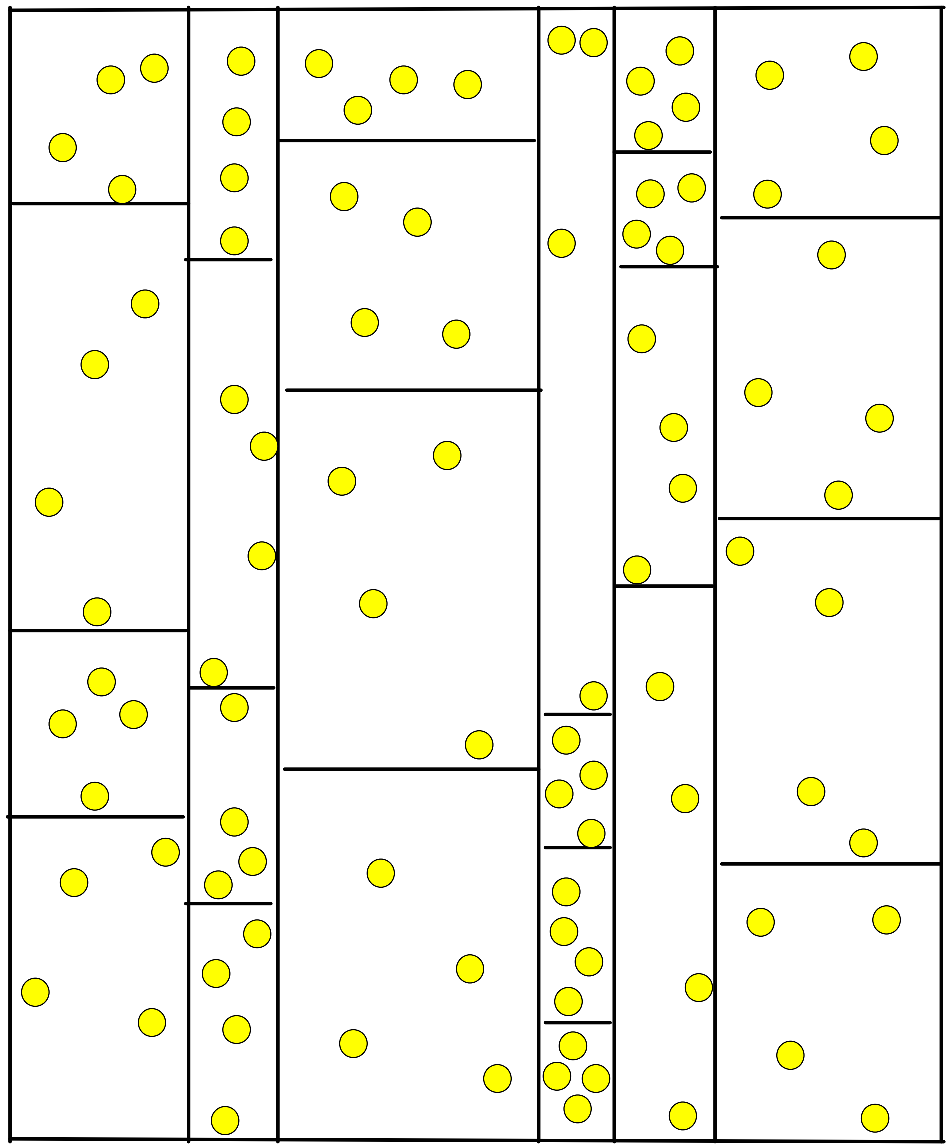
$$|B \cap N| \geq 1$$



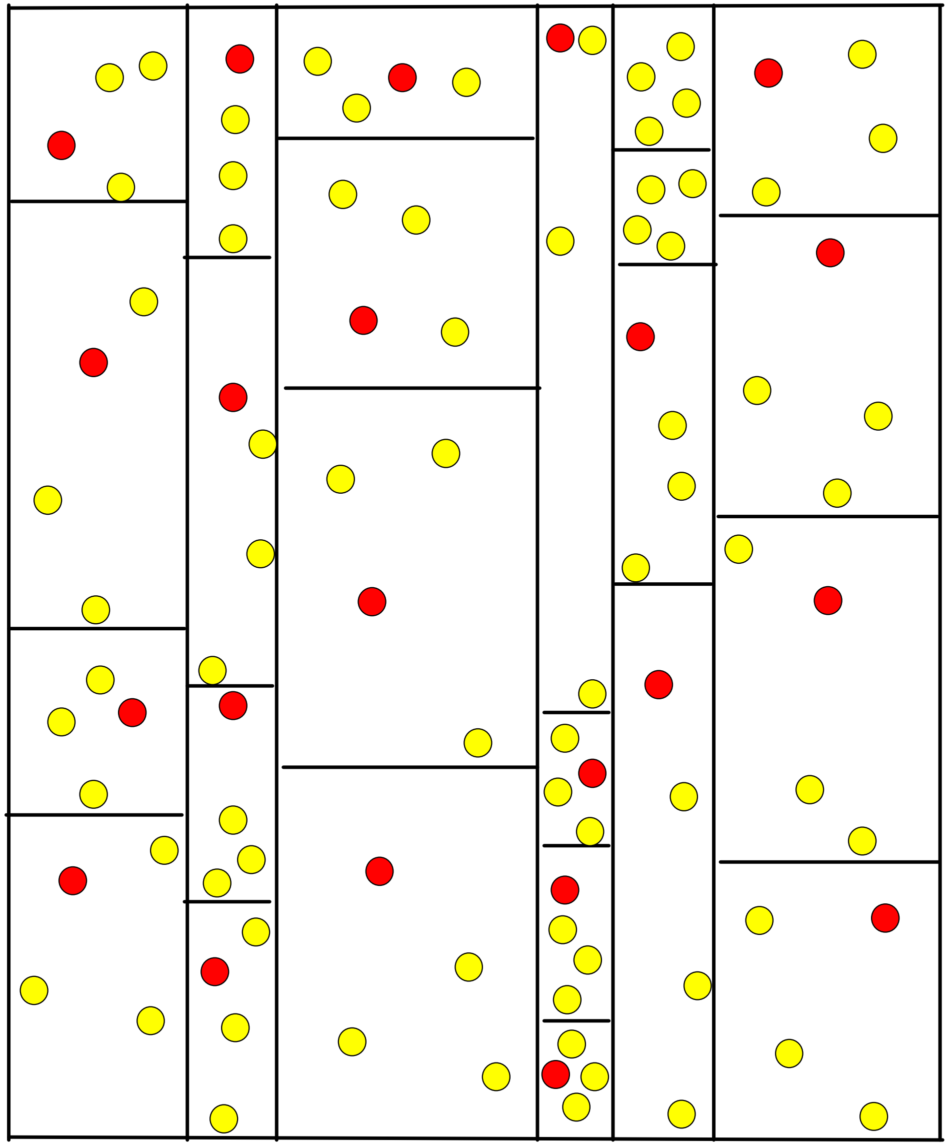
Weak ε-Net for $P \subseteq [0,1]^d$



Weak ε-Net for $P \subseteq [0,1]^d$

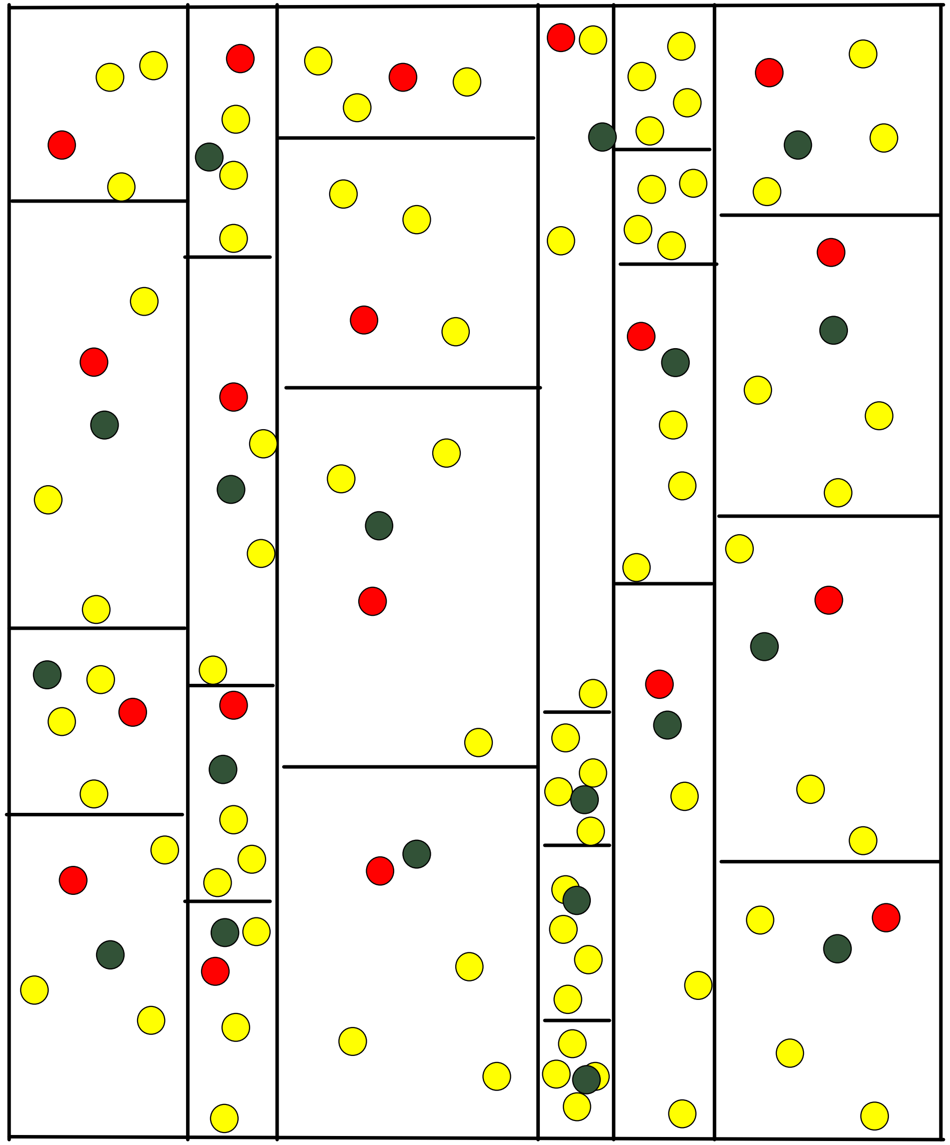


Weak ε-Net for $P \subseteq [0,1]^d$



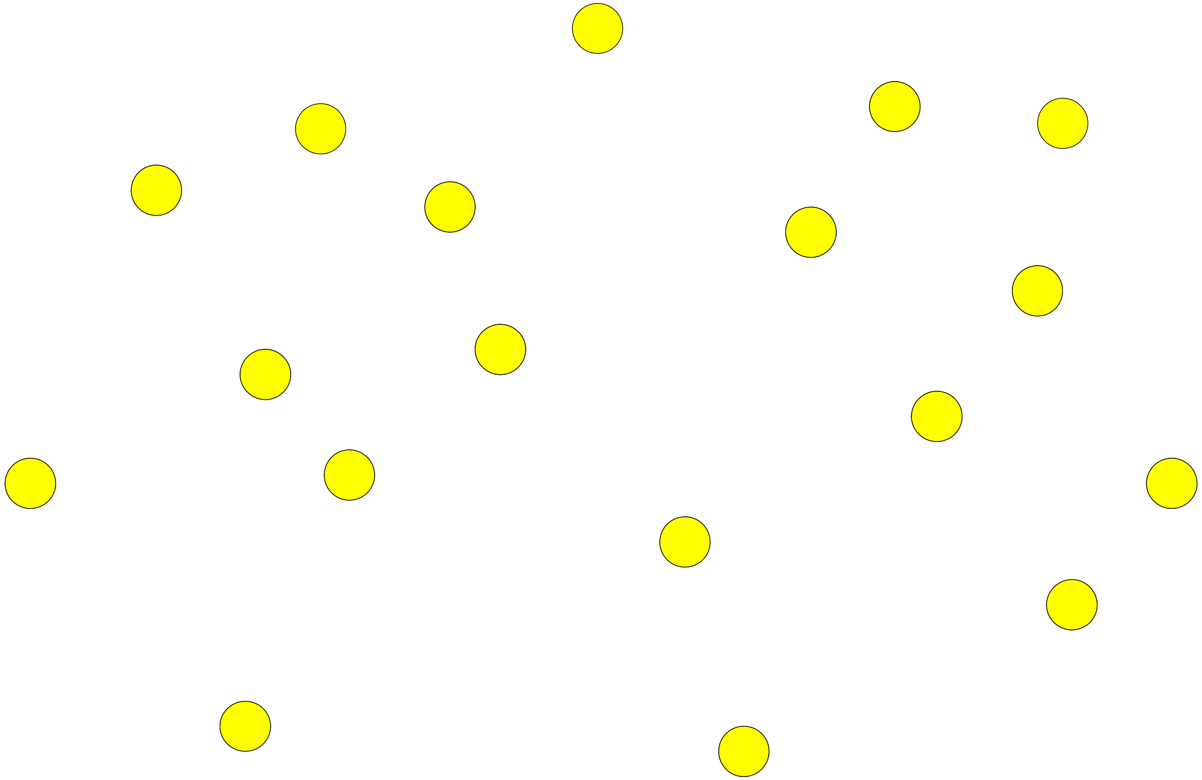
Private ε-Net for $P \subseteq [0,1]^d$

Add noise to each representative



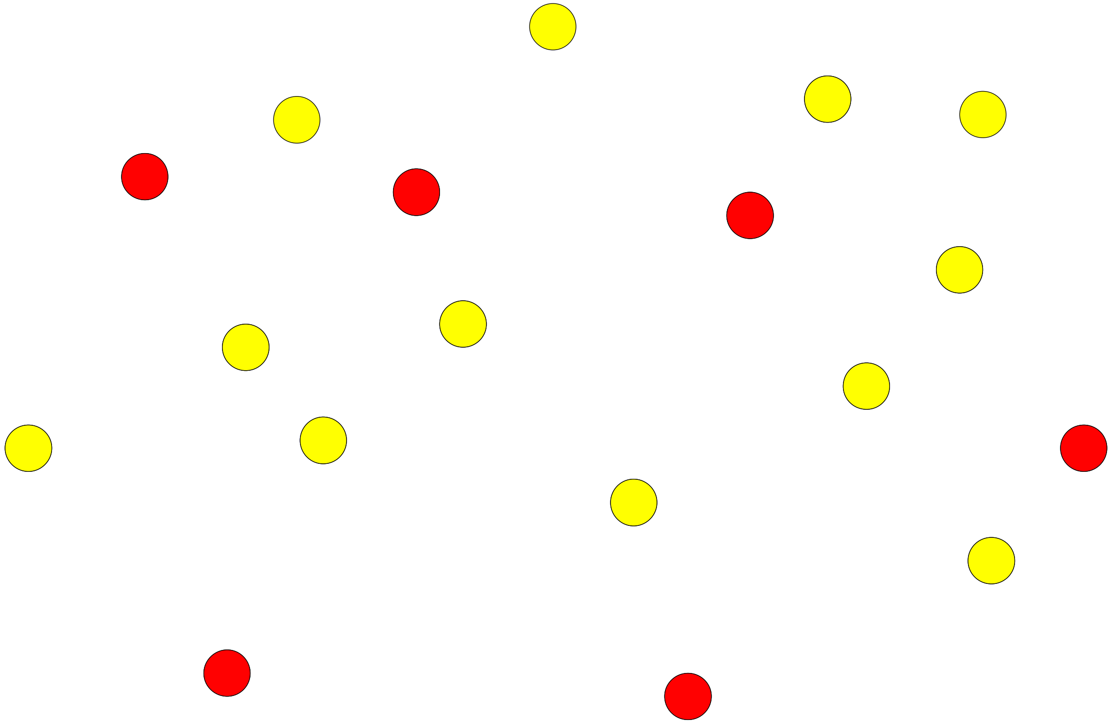
Input

A set of n points $P \subset \mathbb{R}^d$, $k \geq 1$.



Output

N : a small bicriteria approximation to the k median of P

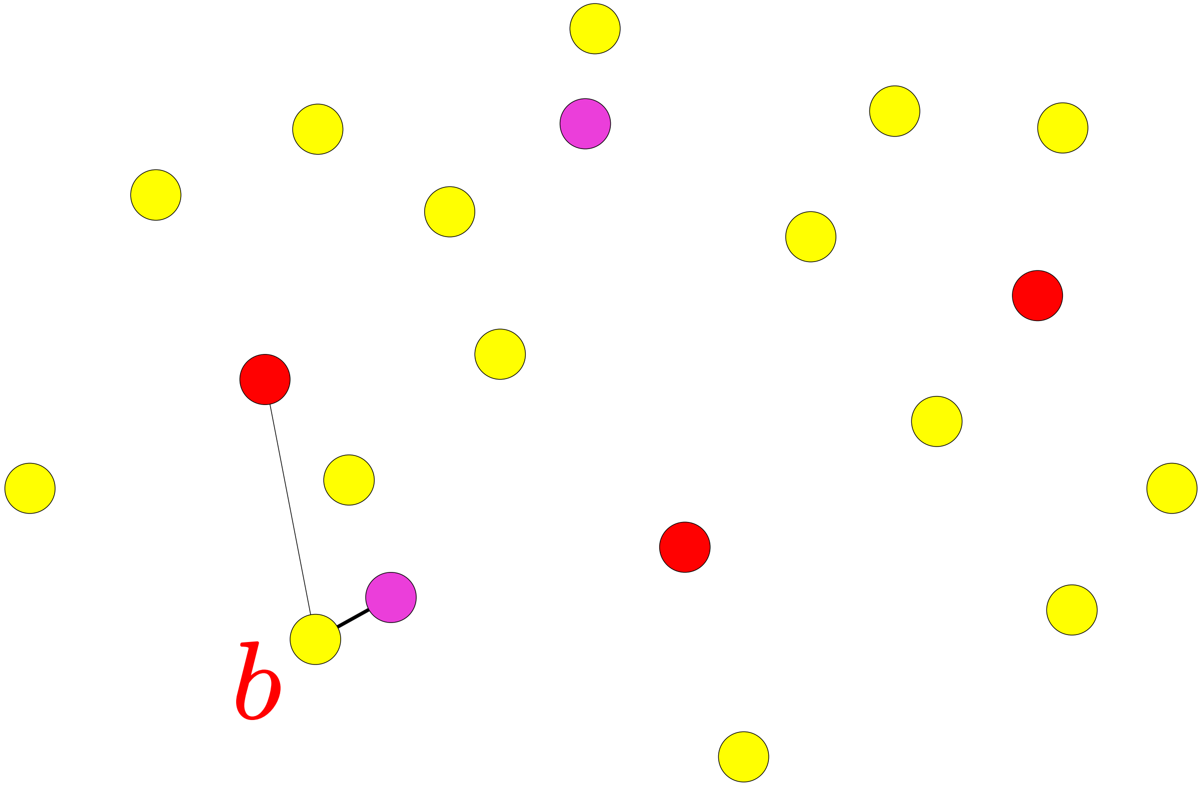


The Bicriteria Algorithm

```
1) $t \leftarrow 1$
2) $N \leftarrow \emptyset$
3) Construct a weak $(\frac{1}{8k})$-net $N_t$ for $P$
4) $N \leftarrow N \cup N_t$
5) Discard $P_t$: $P/2$ pts closer to $N_t$
6) $t \leftarrow t + 1$
7) Repeat steps 3 to 6 until no more pts
8) Return $N$
```

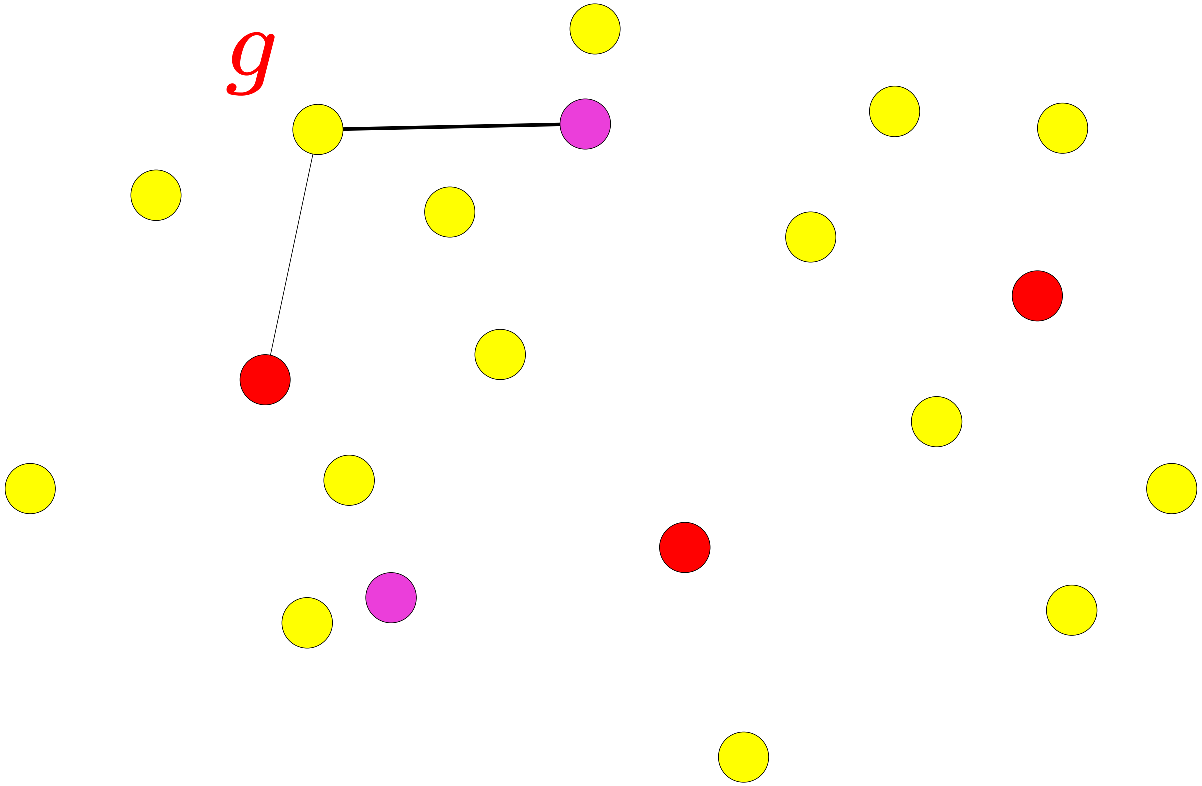
Thank You!
ZWANI.COM

A point $b \in P$ is bad for N_t, if:



$$\mathsf{dist}(b, N_t) > 2\,\mathsf{dist}(b, N^*)$$

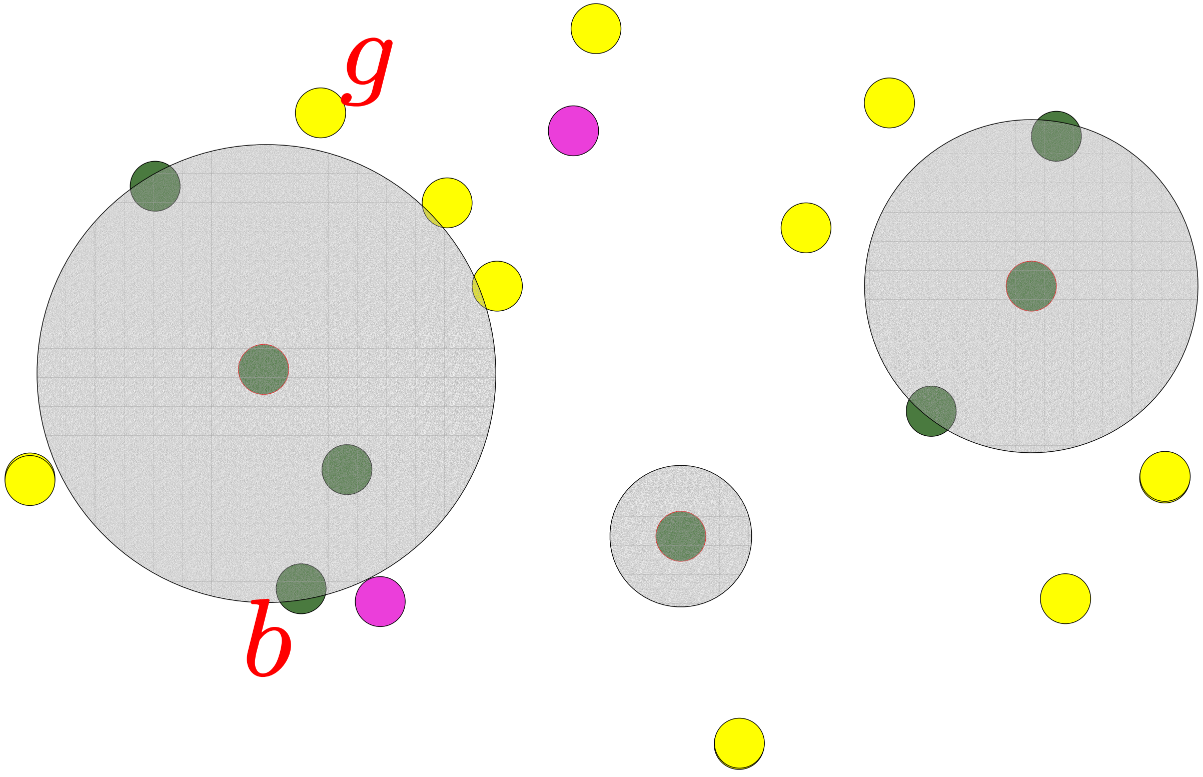
A point $g \in P$ is good for N_t otherwise:



$$\mathrm{dist}(g, N_t) \leq 2\,\mathrm{dist}(g, N^*)$$

Main Technical Theorem

We can map every bad point $b \in P_t$ to a distinct good point $g \in P_{t+1}$.



$\mathrm{dist}(b, N) \leq \mathrm{dist}(b, N_t)$, because $N \supseteq N_t$.

Since $b \in P_t$ and $g \in P_{t+1}$:

$$\mathrm{dist}(b, N_t) \leq \mathrm{dist}(g, N_t)$$

Since g is good for N_t:

$$\mathrm{dist}(g, N_t) \leq 2\,\mathrm{dist}(g, N^*)$$

$\boxed{\mathrm{dist}(b, N)} \leq \mathrm{dist}(b, N_t)$, because $N \supseteq N_t$.

Since $b \in P_t$ and $g \in P_{t+1}$:

$$\mathrm{dist}(b, N_t) \leq \mathrm{dist}(g, N_t)$$

Since g is good for N_t:

$$\mathrm{dist}(g, N_t) \leq \boxed{2\,\mathrm{dist}(g, N^*)}$$

⬇

$$\mathrm{dist}(b, N) \leq 2\,\mathrm{dist}(g, N^*)$$

Bi-Criteria for *k*-Median

$$\begin{aligned}\sum_{p\in P}\mathsf{dist}(p,N) &= \sum_{g}\mathsf{dist}(g,N) + \sum_{b}\mathsf{dist}(b,N)\\ &\le \sum_{g} 2\,\mathsf{dist}(g,N^*) + \sum_{g} 2\,\mathsf{dist}(g,N^*)\\ &\le 4\sum_{p\in P}\mathsf{dist}(p,N^*)\end{aligned}$$

Open Questions

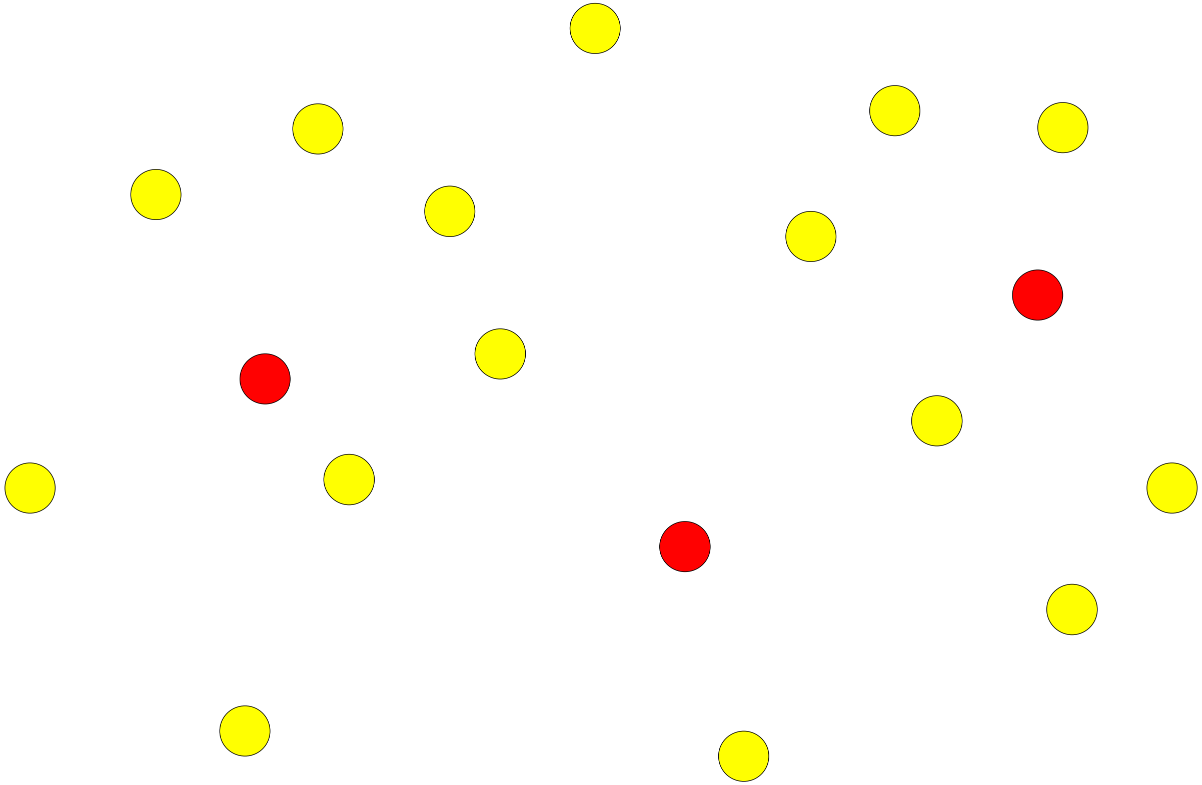
- Private coresets for k-median in high dimensional spaces
- Private coresets for k subspaces of $\mathbb{R}^d$
- Private coresets for other shapes.
- Private dynamic Coresets

Bi-Criteria Approximation Algorithm [FFS07]

Initialization

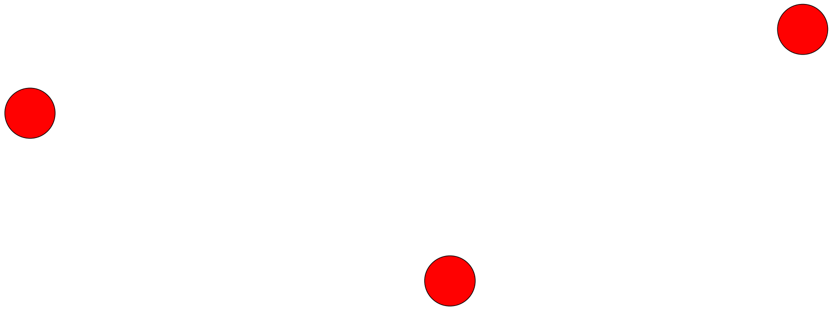
```
1) $t \leftarrow 1$
        $\triangleright$ Counter for iterations

2) $F \leftarrow \emptyset$
        $\triangleright$ The output set of $j$-flats
```

3) Construct a weak $(\frac{1}{8k})$-net N_t for P



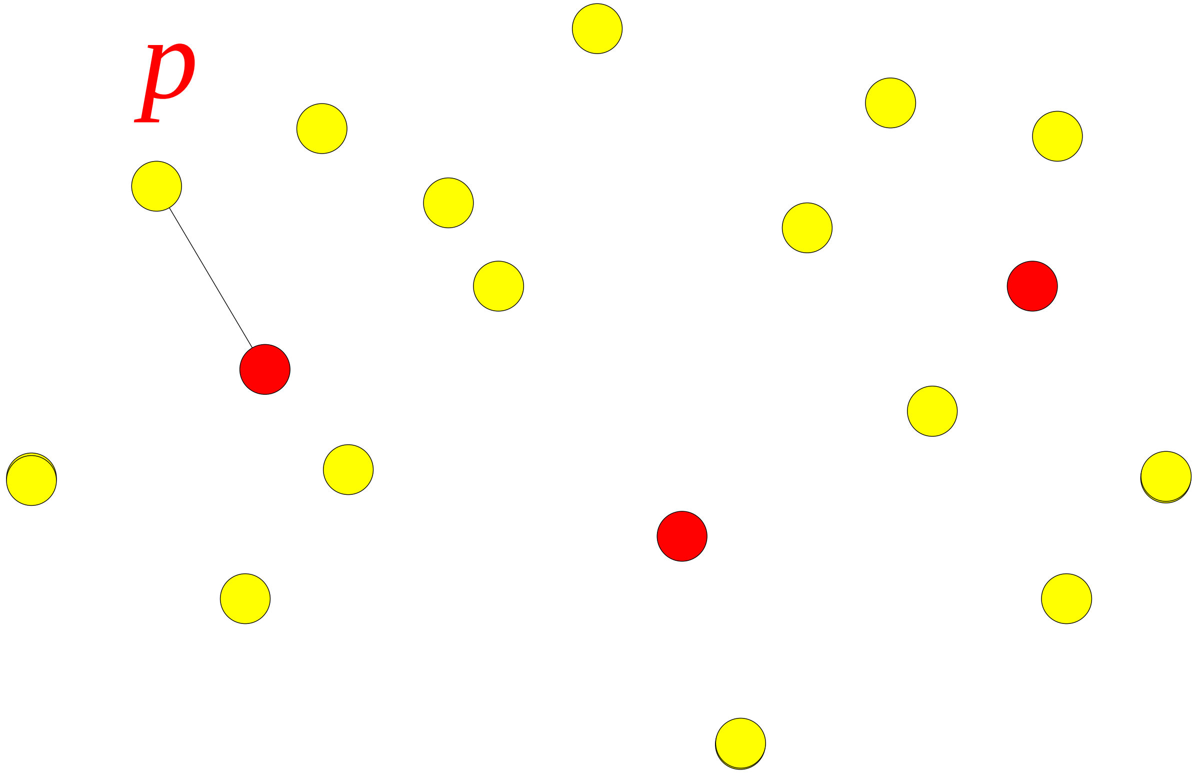
$t = 1$

4) $N \leftarrow N \cup N_t$



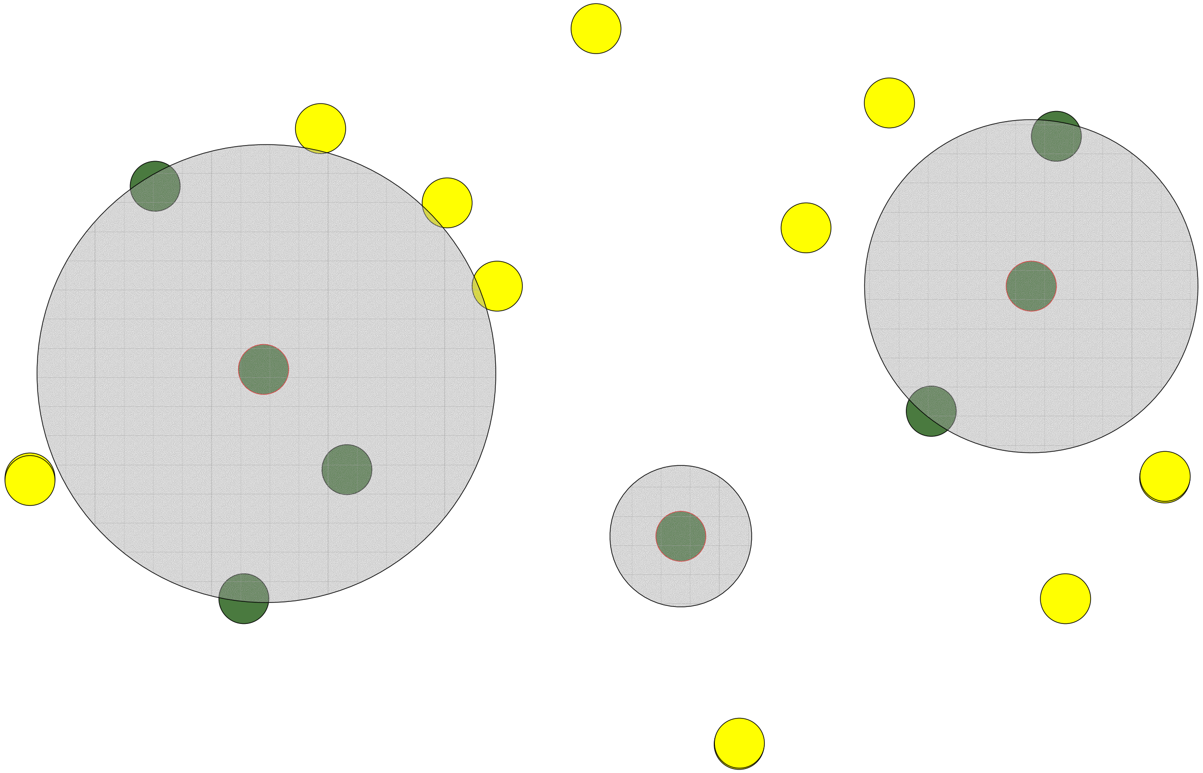
$(t = 1)$

5) $\forall p$: Compute $\mathrm{dist}(p, N_t)$



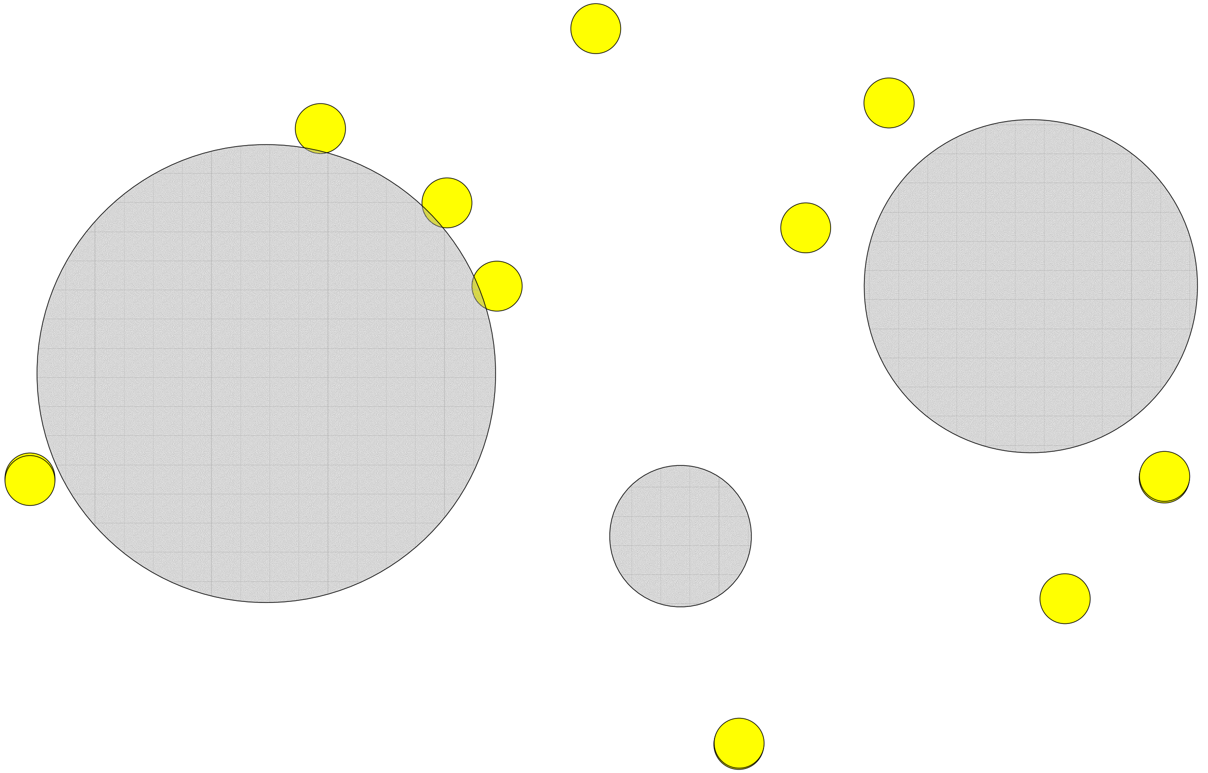
$(t = 1)$

6) Remove P_t: the half of P that is closer to N_t



$(t = 1)$

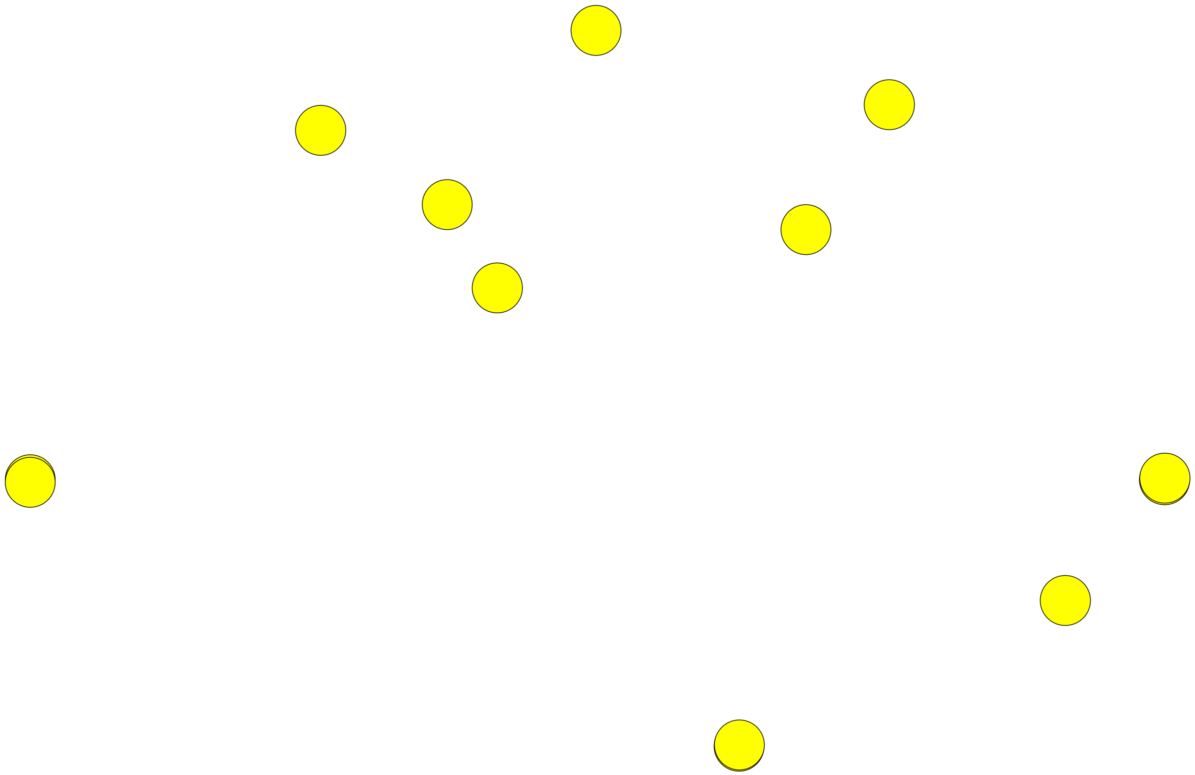
6) Remove P_t: the half of P that is closer to N_t



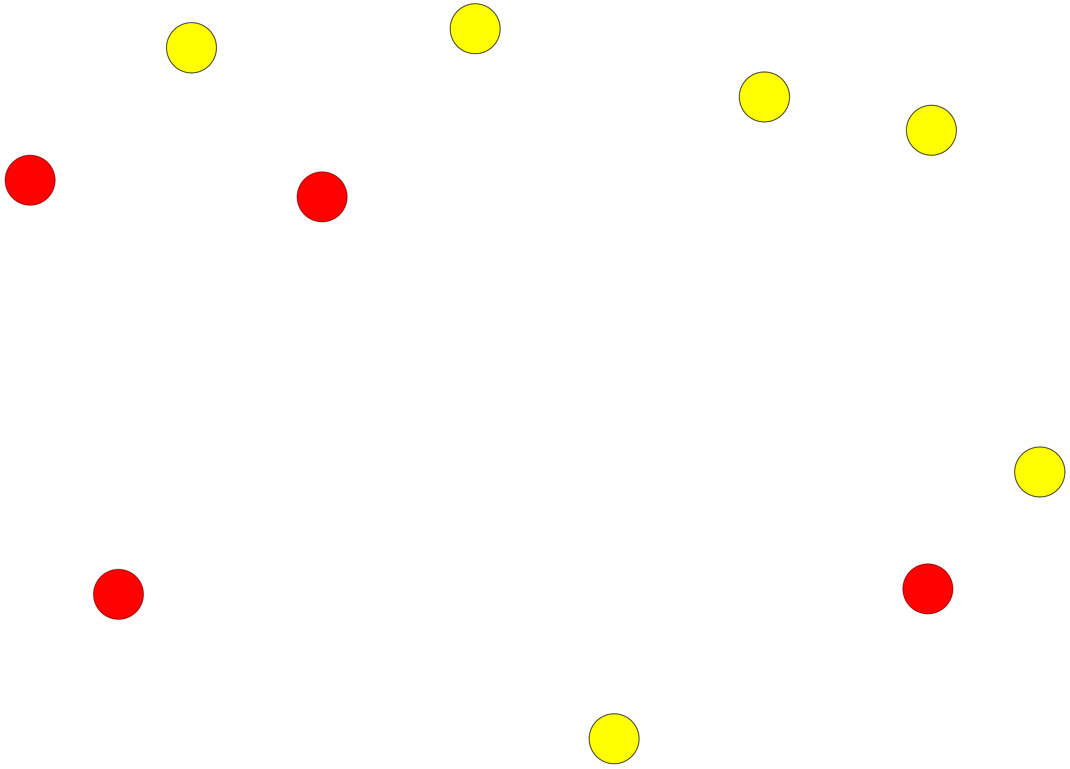
$(t = 1)$

7) $t \leftarrow t + 1$

8) Repeat steps 3 to 6:

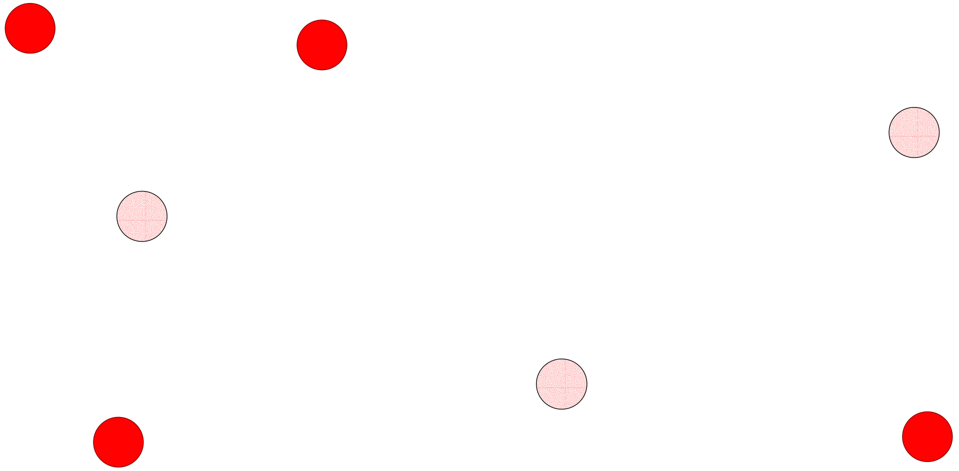


3) Construct a weak $(1/k)$-net N_t for P



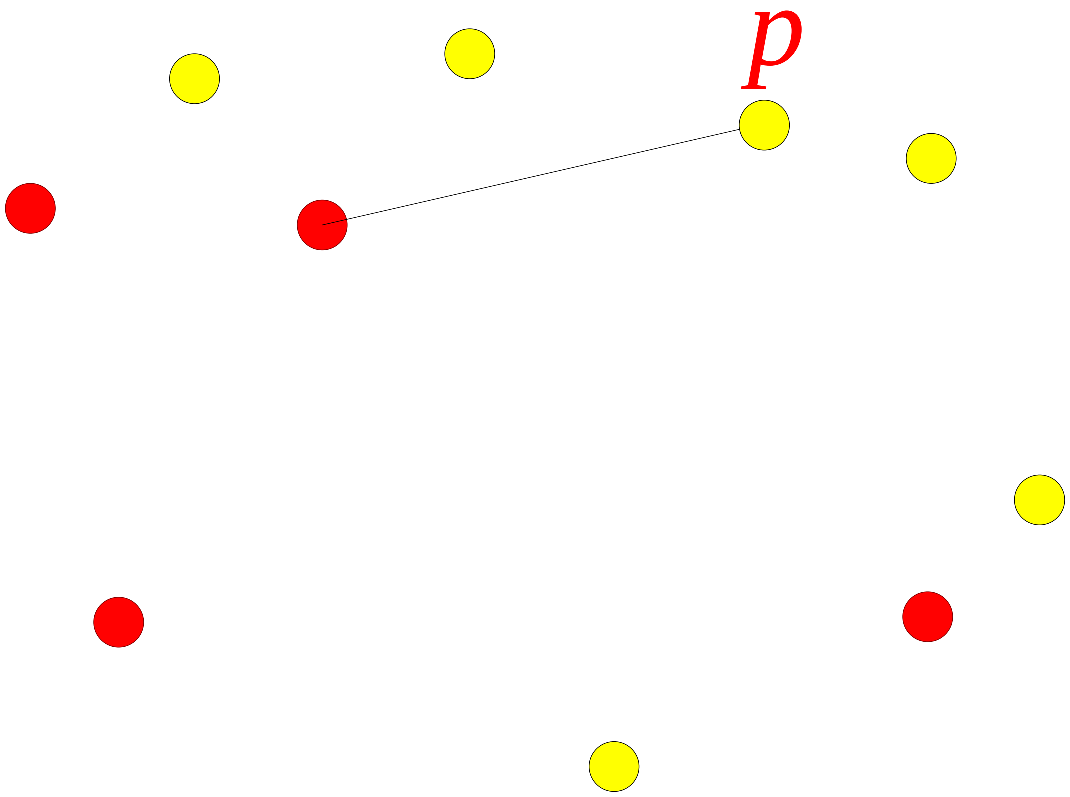
$(t = 2)$

4) $N \leftarrow N \cup N_t$



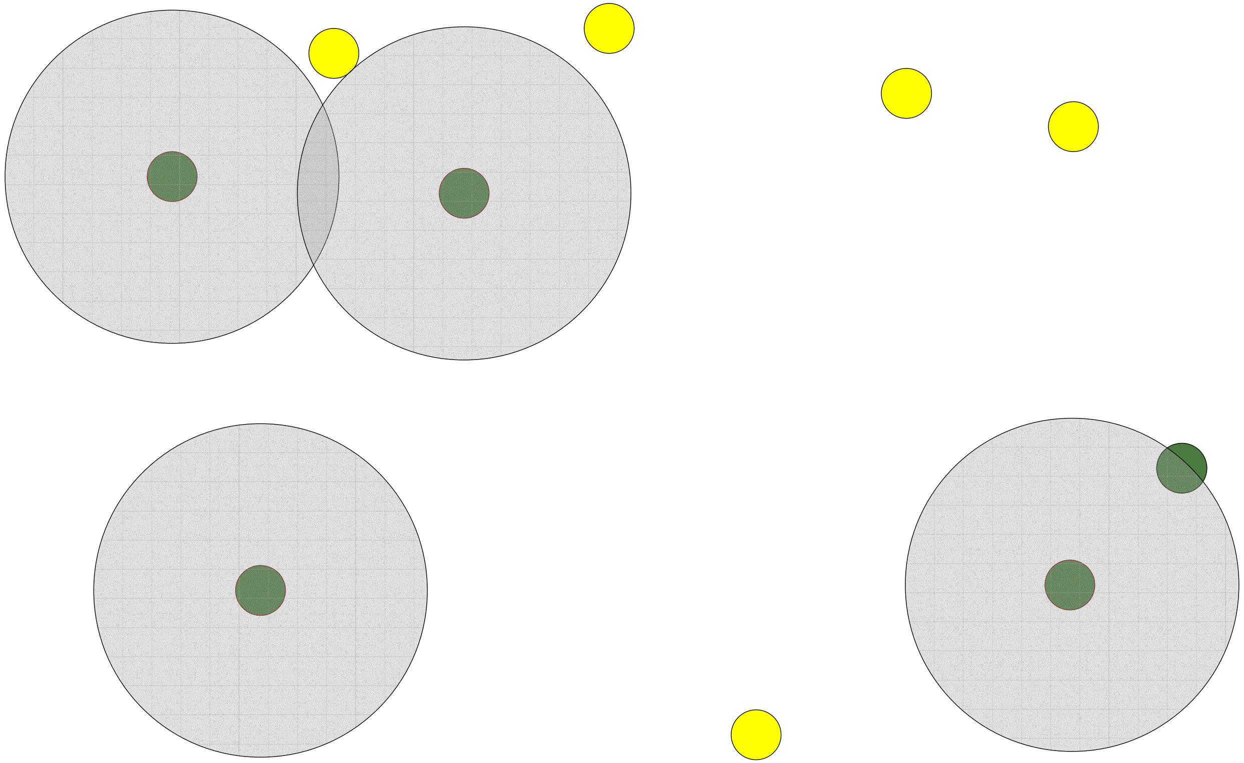
$(t = 2)$

5) $\forall p$: Compute $\mathrm{dist}(p, N_t)$



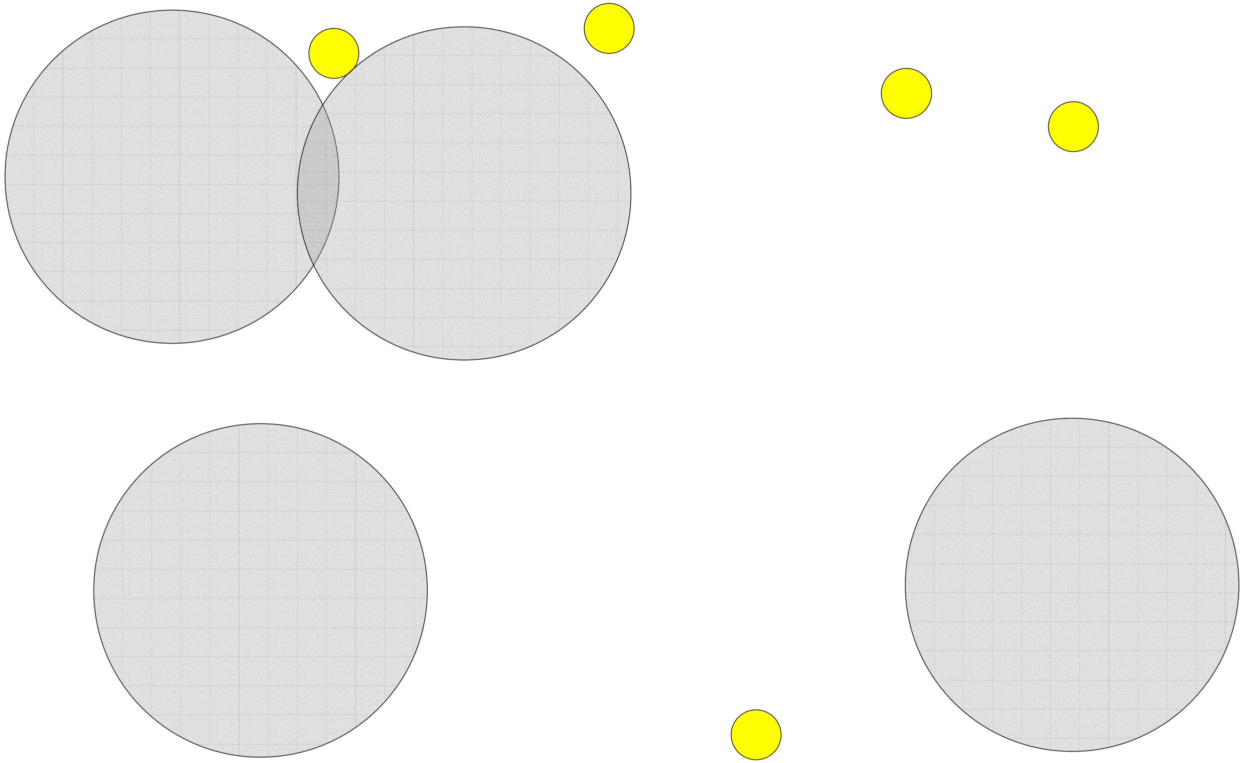
$(t = 2)$

6) Remove P_t: the half of P that is closer to N_t



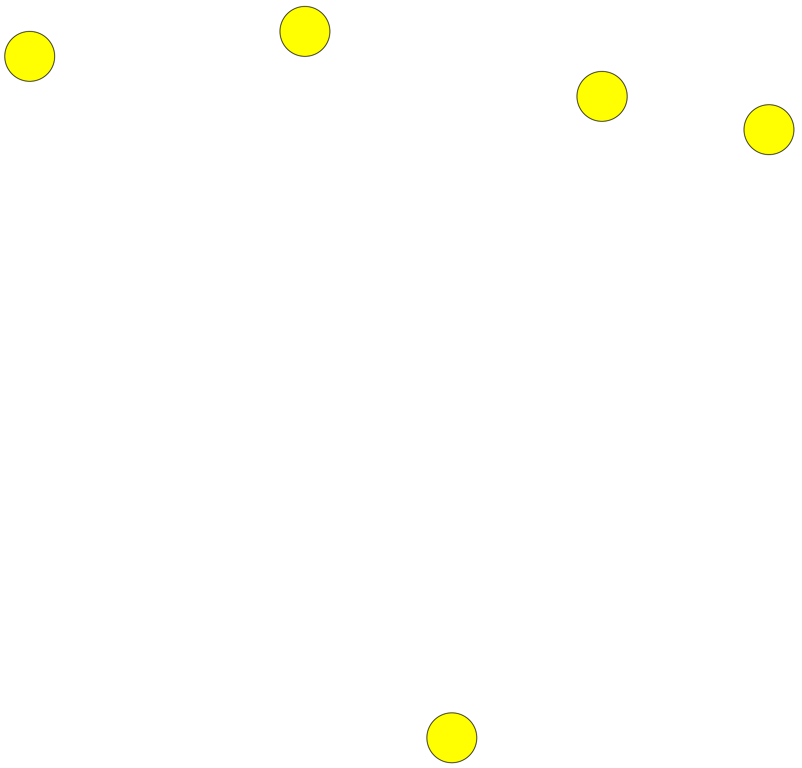
$(t = 2)$

6) Remove P_t: the half of P that is closer to N_t



$(t = 2)$

6) Remove P_t: the half of P that is closer to N_t



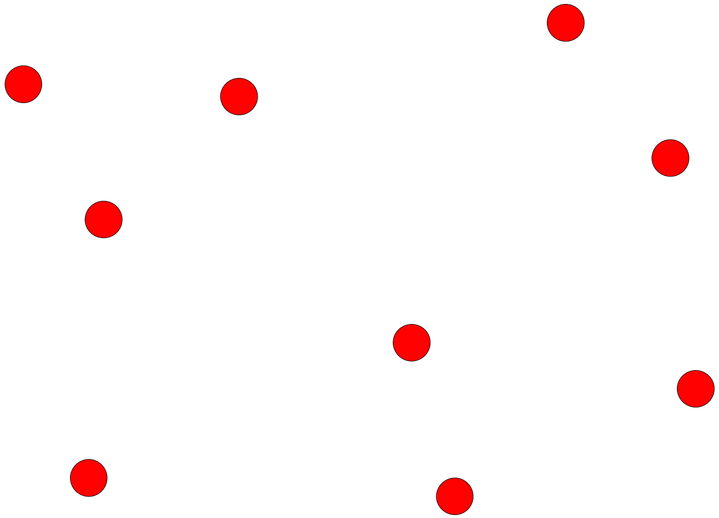
$(t = 2)$

7) $t \leftarrow t + 1$

8) Repeat steps 3 to 6

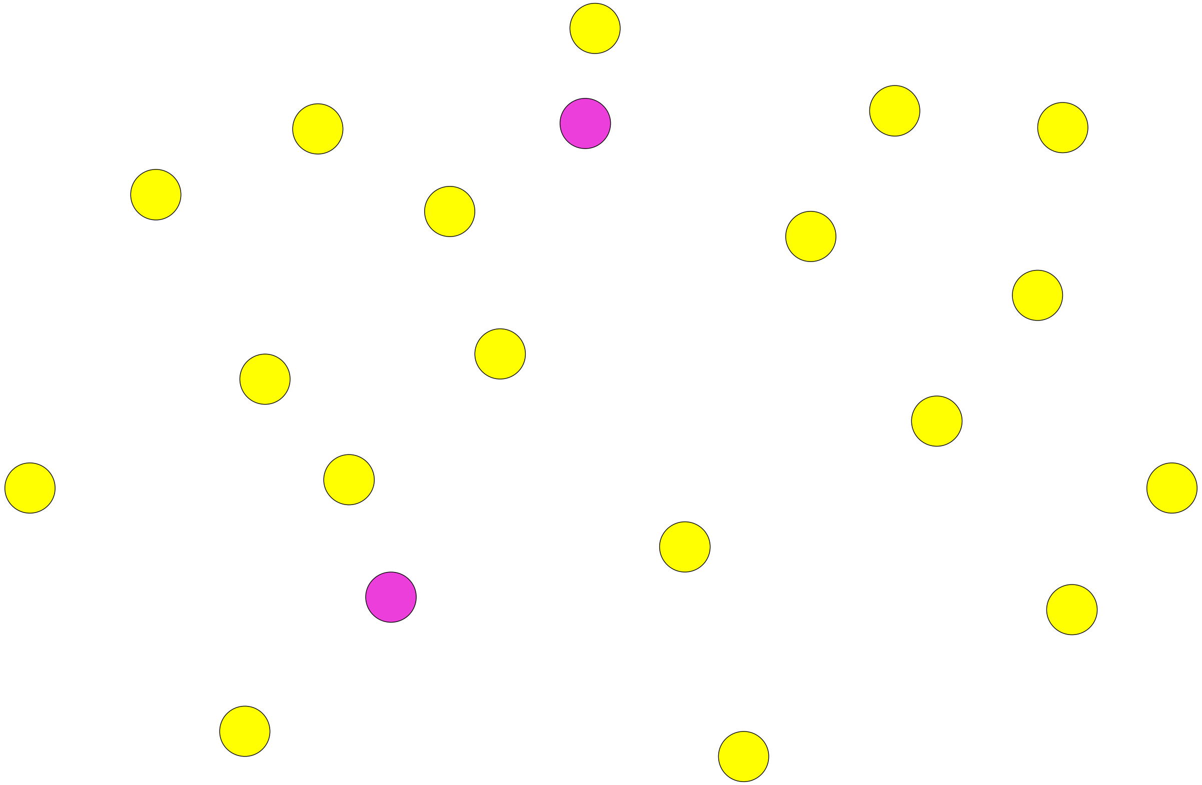
till there are no more input points.

9) Return N :

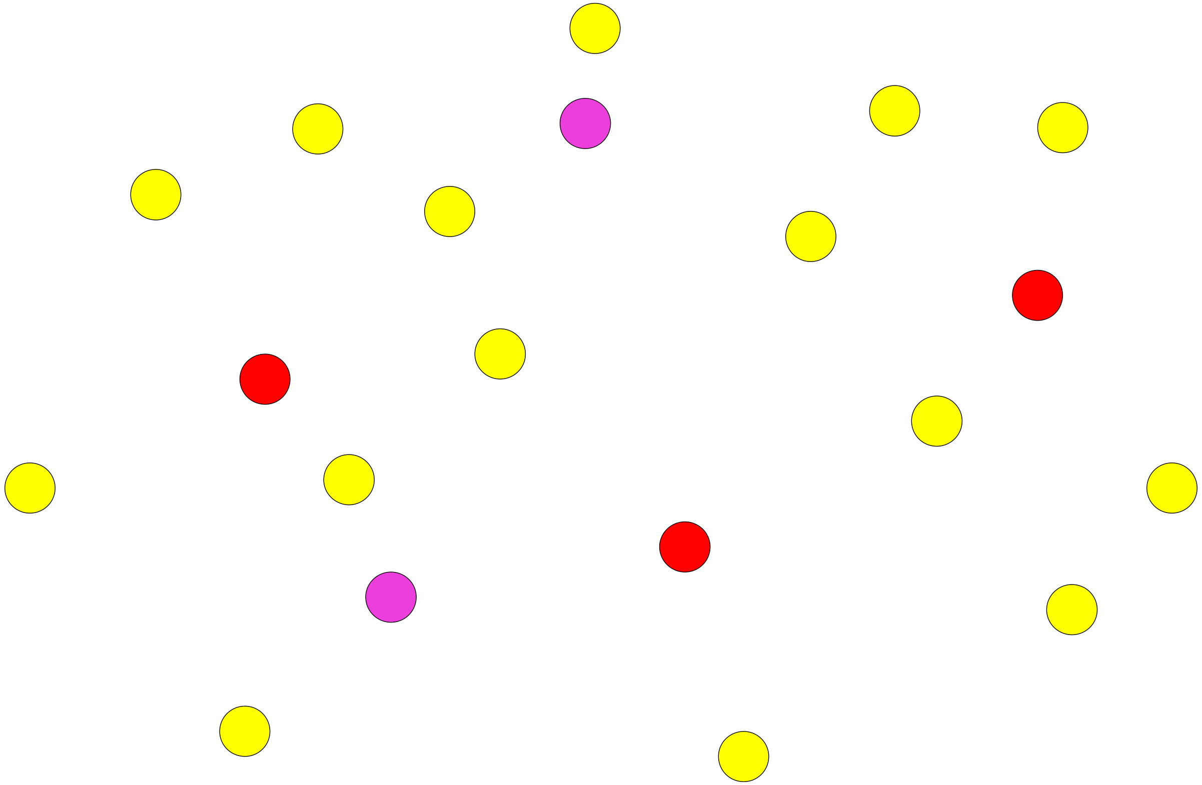


Let N^* be any set of k points in $\mathbb{R}^d$.

Let N^* be any set of k points in $\mathbb{R}^d$.

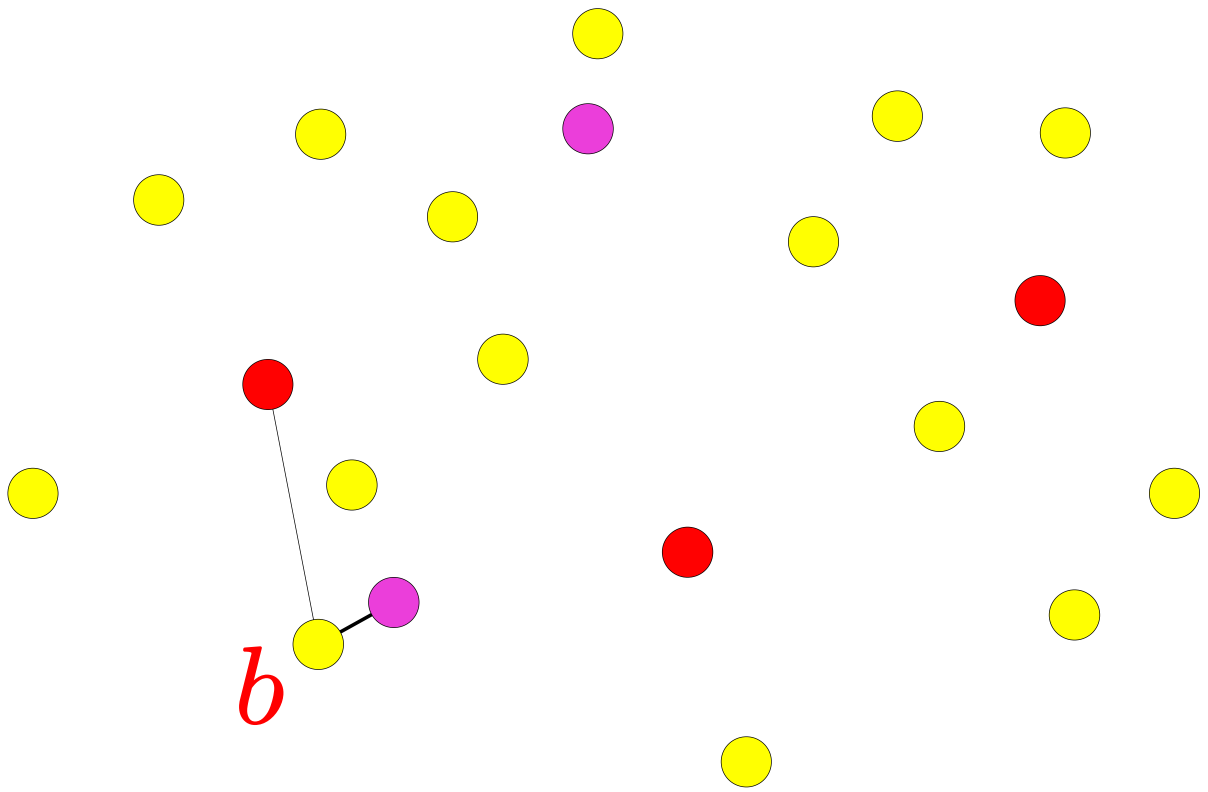


Let N^* be any set of k points in $\mathbb{R}^d$.



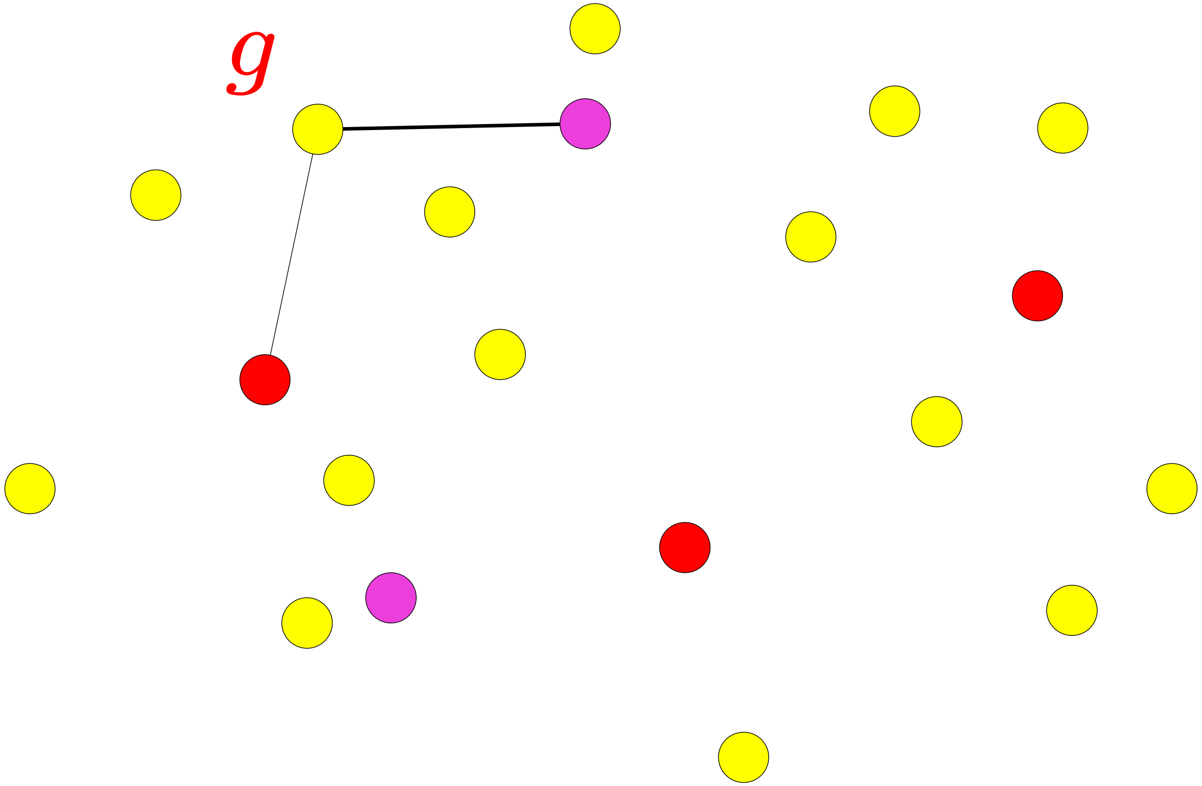
Consider N_t that is constructed during the t^{th} iteration.

A point $b \in P$ is bad for N_t, if:



$$\mathrm{dist}(b, N_t) > 2\,\mathrm{dist}(b, N^*)$$

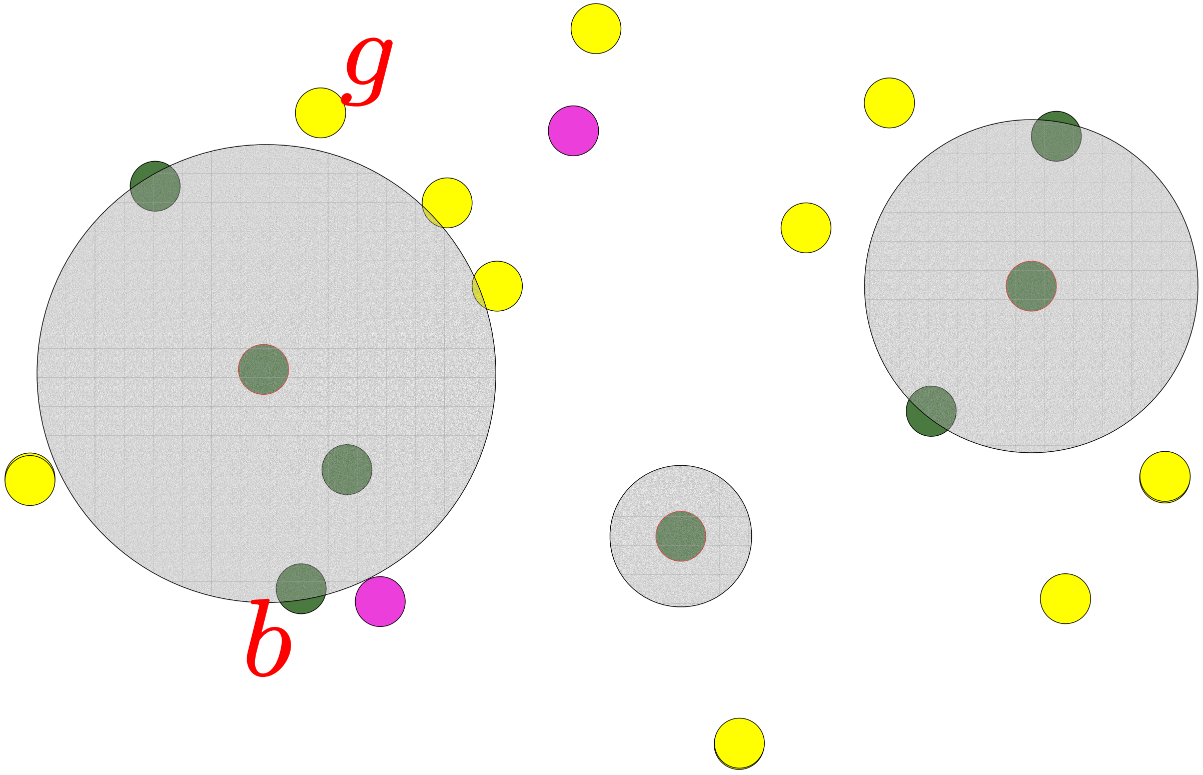
A point $g \in P$ is good for N_t otherwise:



$$\mathrm{dist}(g, N_t) \leq 2\,\mathrm{dist}(g, N^*)$$

Main Technical Theorem

We can map every bad point $b \in P_t$ to a distinct good point $g \in P_{t+1}$.



$\mathsf{dist}(b, N) \le \mathsf{dist}(b, N_t)$, because $N \supseteq N_t$.

Since $b \in P_t$ and $g \in P_{t+1}$:

$$\mathsf{dist}(b, N_t) \le \mathsf{dist}(g, N_t)$$

Since g is good for N_t:

$$\mathsf{dist}(g, N_t) \le 2\,\mathsf{dist}(g, N^*)$$

$\boxed{\text{dist}(b, N)} \leq \text{dist}(b, N_t)$, because $N \supseteq N_t$.

Since $b \in P_t$ and $g \in P_{t+1}$:

$$\text{dist}(b, N_t) \leq \text{dist}(g, N_t)$$

Since g is good for N_t:

$$\text{dist}(g, N_t) \leq \boxed{2\,\text{dist}(g, N^*)}$$

$$\Downarrow$$

$$\text{dist}(b, N) \leq 2\,\text{dist}(g, N^*)$$

Bi-Criteria for *k*-Median

$$
\begin{aligned}
\sum_{p \in P} \mathsf{dist}(p, N) &= \sum_{g} \mathsf{dist}(g, N) \qquad + \sum_{b} \mathsf{dist}(b, N) \\
&\leq \sum_{g} 2\,\mathsf{dist}(g, N^*) + \sum_{g} 2\,\mathsf{dist}(g, N^*) \\
&\leq 4 \sum_{p \in P} \mathsf{dist}(p, N^*)
\end{aligned}
$$

Proof of the Technical Theorem

- The number of bad points is at most

$$|B| = \frac{|P_t|}{8}$$

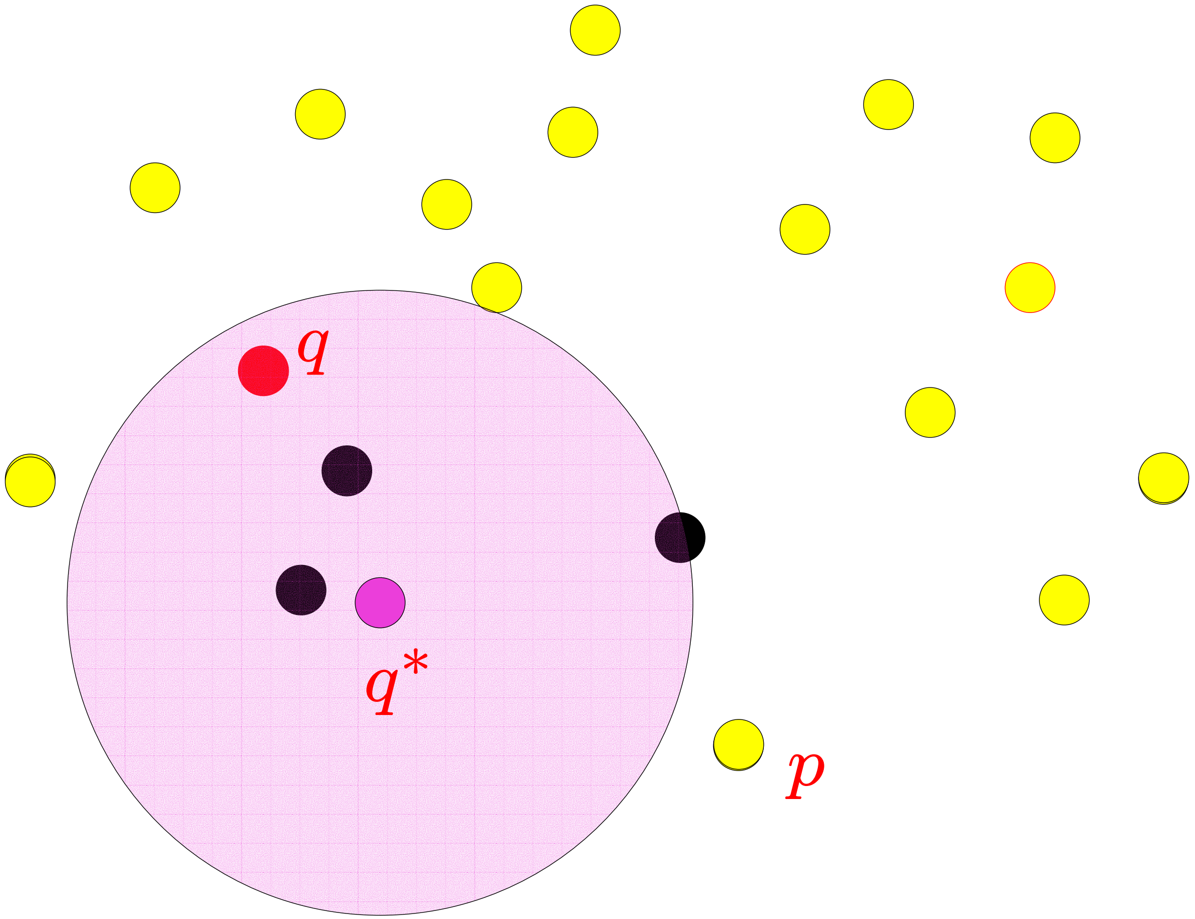
- $$\left|P_{t+1}\right| = \frac{|P_t|}{2}$$

⬇

The number of good points in P_{t+1} is at least

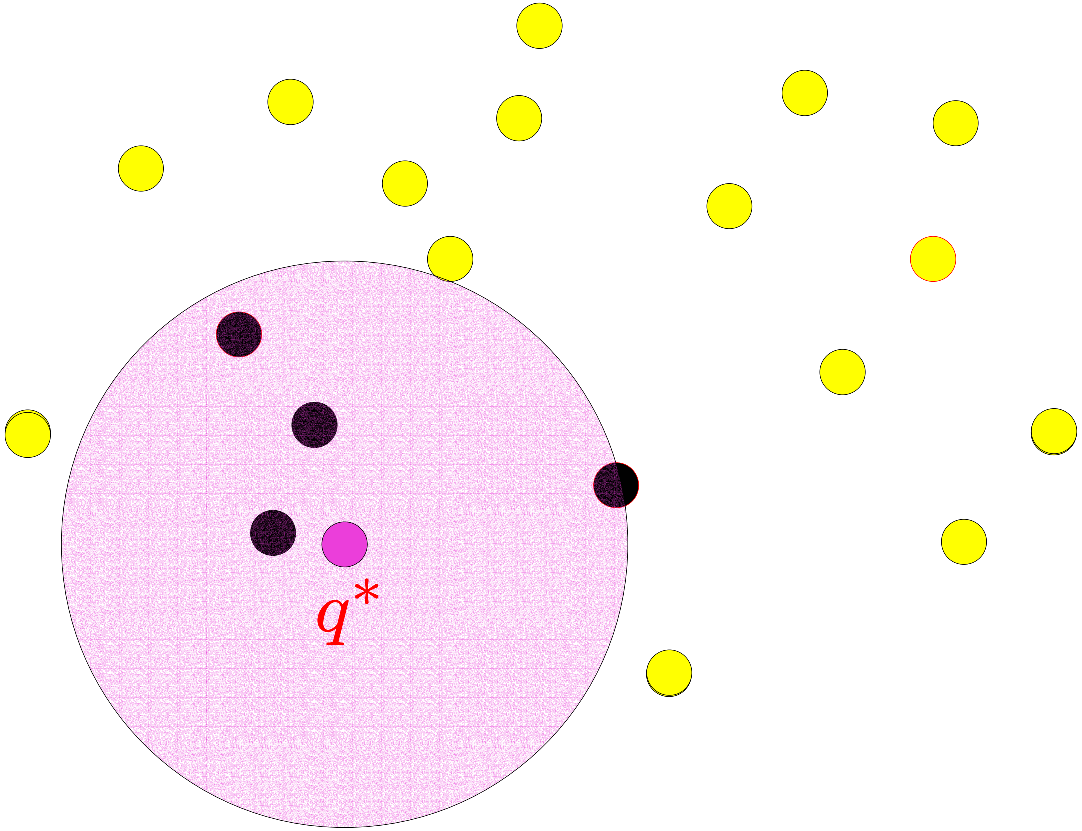
$$\left|P_{t+1}\right| - |B| \geq \frac{|P_t|}{2} - \frac{|P_t|}{8} \geq |B|$$

Claim: Only $B_0 = \frac{|P_t|}{8k}$ points are bad for $q \in N_t$

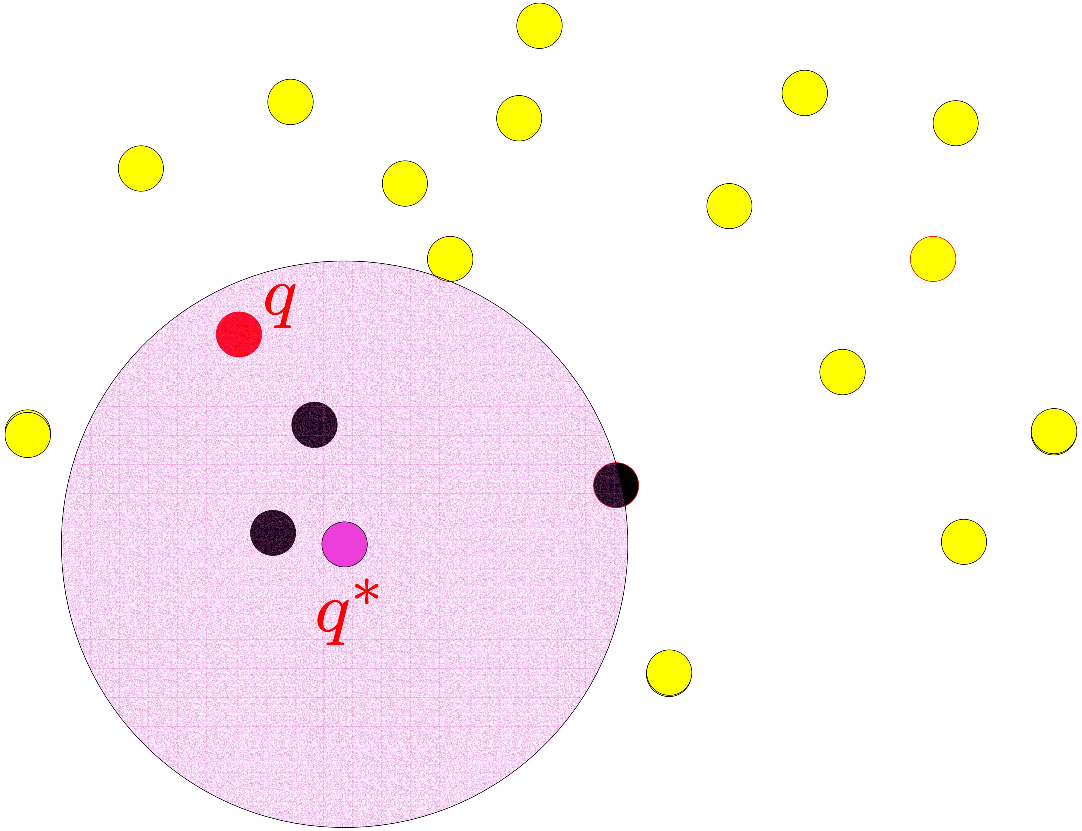


$$\mathsf{dist}(p, q) \leq 2\,\mathsf{dist}(p, q^*)$$

B_0: the $\frac{|P_t|}{8k}$ closest points to q^*

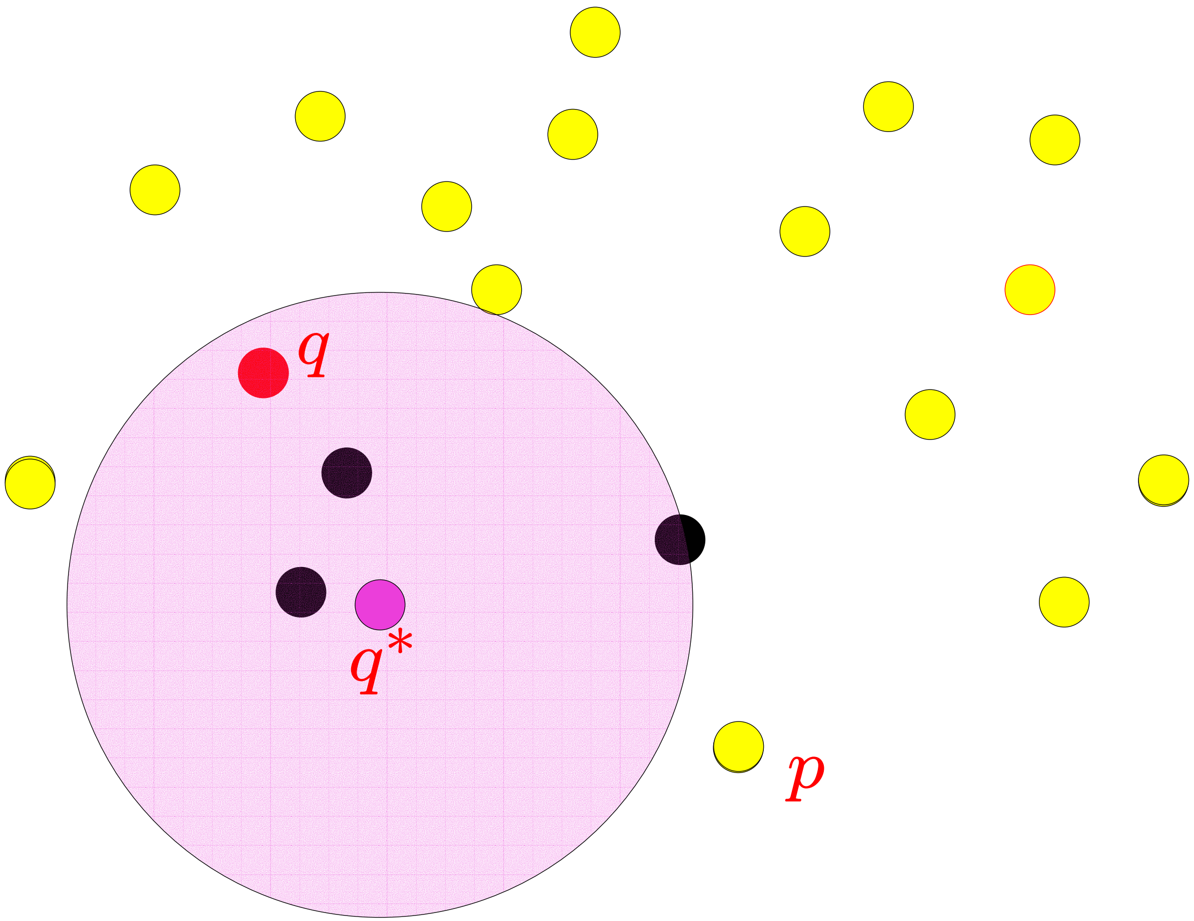


B_0: the $\frac{|P_t|}{8k}$ closest points to q^*



B_0 contains $q \in N_t$ $\left(\frac{1}{8k}\text{-net}\right)$

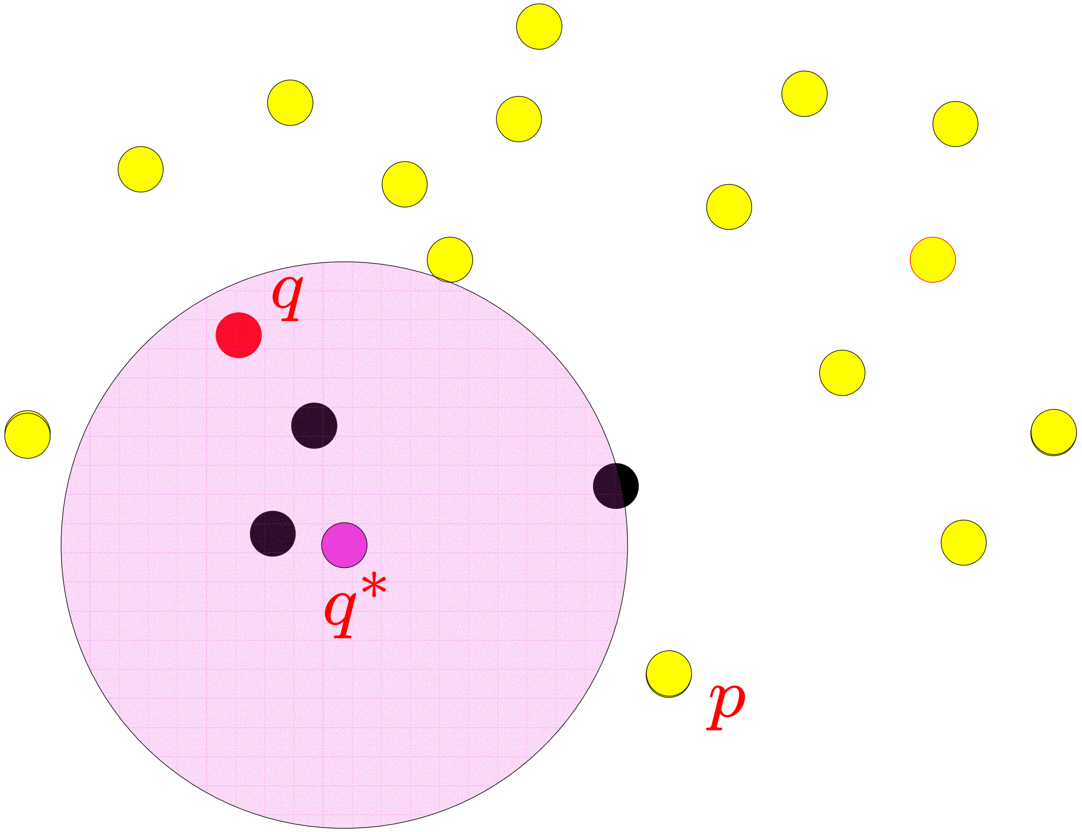
For every yellow point $p \in P \setminus B_0$:



$$\mathrm{dist}(p,q) \le \mathrm{dist}(p,q^*) + \mathrm{dist}(q^*,q)$$

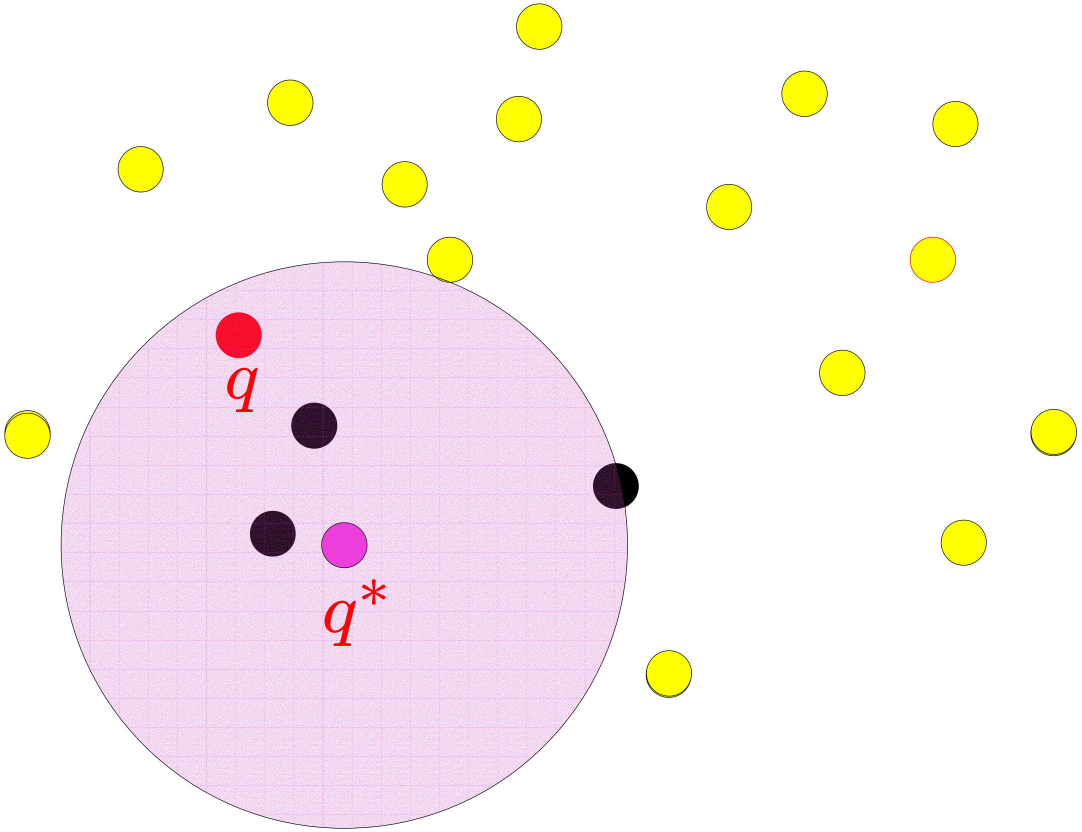
$$\le 2\,\mathrm{dist}(p,q^*)$$

All the yellow points are good for N_t



$$\mathrm{dist}(p,q) \leq 2\,\mathrm{dist}(p,q^*)$$

Only the black points B_0 are bad for N_t



$$|B_0| = \frac{|P_t|}{8}$$