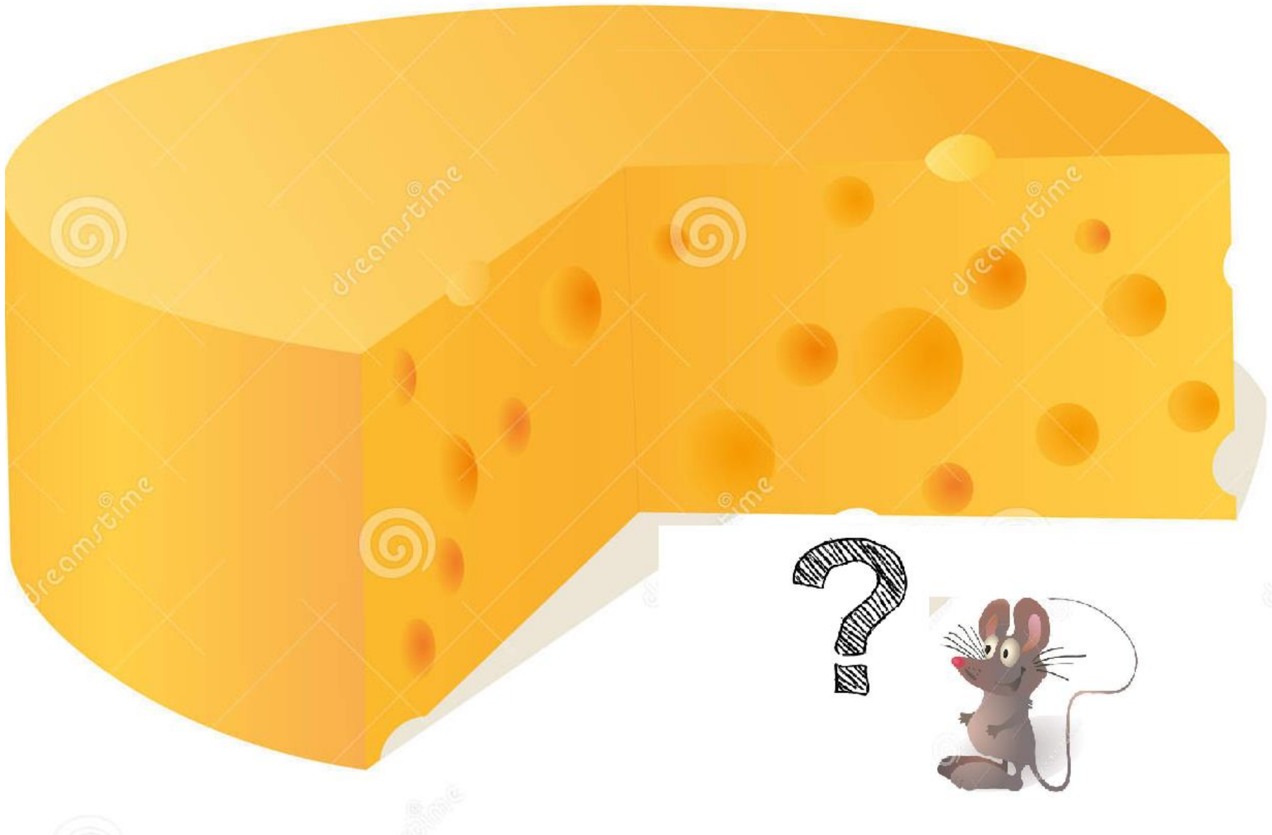
Big Data Class



LECTURER: DAN FELDMAN

TEACHING ASSISTANTS:

IBRAHIM JUBRAN

ALAA MAALOUF

So Far We Have Seen

- Query space – Examples: One mean, points to hyperplanes, one center.
- Exact coreset – Examples: One mean, points to hyperplanes, one center for 1D input points.
- ε-coreset for unweighted input– One center (2 different coresets), k-center.
- ε-coreset for weighted input – One center.
- Streaming tree
- ε-net

Problem: Coresets seen so far are very specific and problem dependent.
Solution: A general framework for coreset construction.

Unified Framework for Coreset Construction

Query space
(P, ω, Q, f)

Coreset

Coreset
$(C, \mu), C \subseteq P, |C| \ll |P|$

Sample a subset of "important" points.

Query space
(P, ω, Q, f)

Compute
"importance"/ "sensitivity"
for each point

Sensitivity
$S: P \times \omega \to [0, \infty)$ s.t.

$$S(p) \geq \max_{q \in Q} \frac{\omega(p) \cdot f(p, q)}{\sum_{p' \in P} \omega(p') \cdot f(p', q)}$$

Helps compute the importance for each input points by a reduction to a simpler problem

Query space
(P, ω, Q, f)

Rough approximation for the optimal solution
(α, β)-approximation

A candidate solution to the problem with provable guarantees.

We will now focus on this block

Definitions

Let $(P, X, dist)$ be a query space, where $dist: P \times X \to [0, \infty)$.
For every $p \in P$ and $Y \subseteq X$ define $dist(p, Y) = \min_{y \in Y} dist(p, y)$.

Definitions

Let $(P, X, dist)$ be a query space, where $dist: P \times X \to [0, \infty)$.
For every $p \in P$ and $Y \subseteq X$ define $dist(p, Y) = \min_{y \in Y} dist(p, y)$.

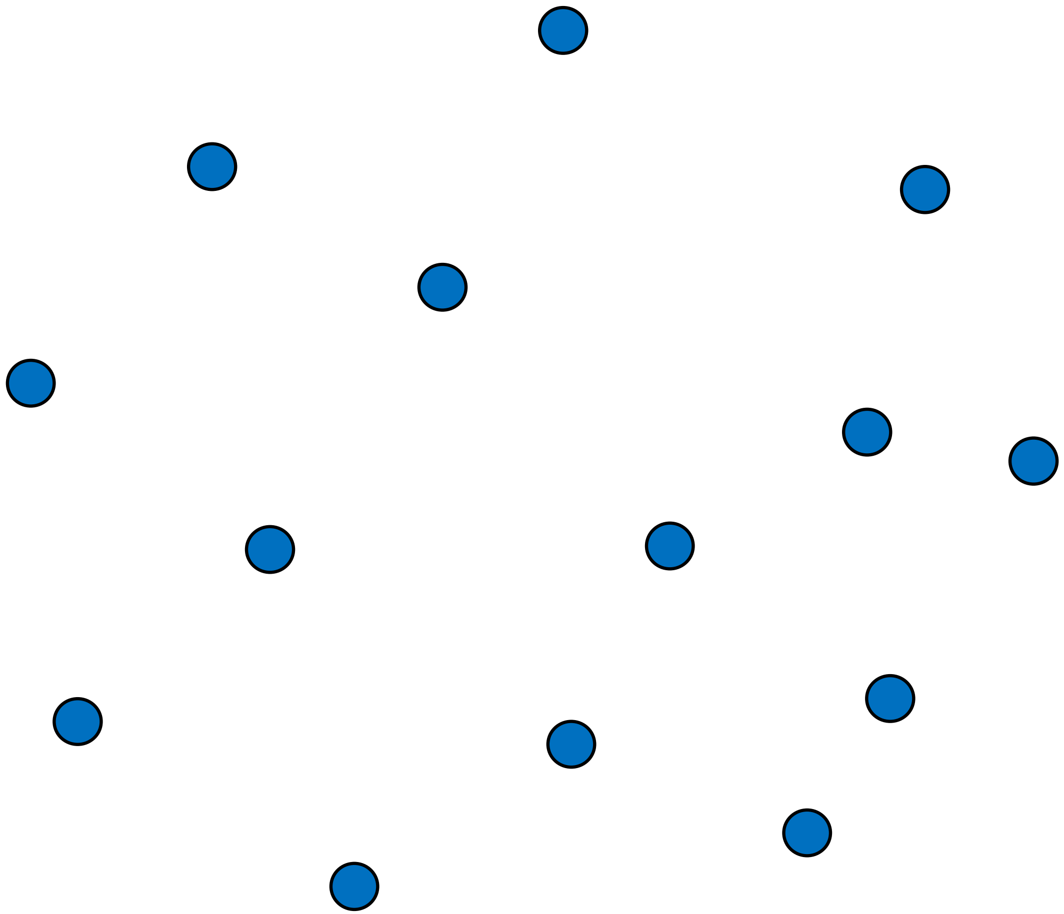
- $OPT = \min_{y \in X} \sum_{p \in P} dist(p, y)$.

OPT Illustration

Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

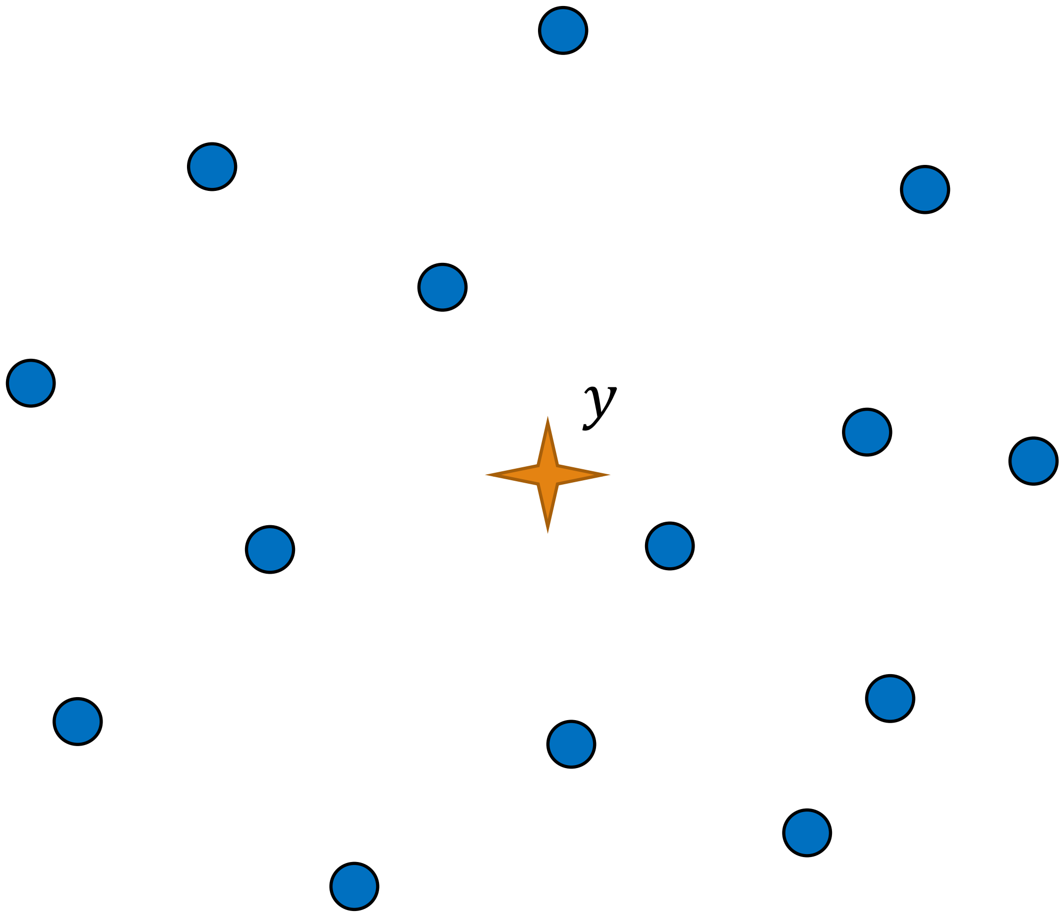


OPT Illustration

Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$



OPT Illustration

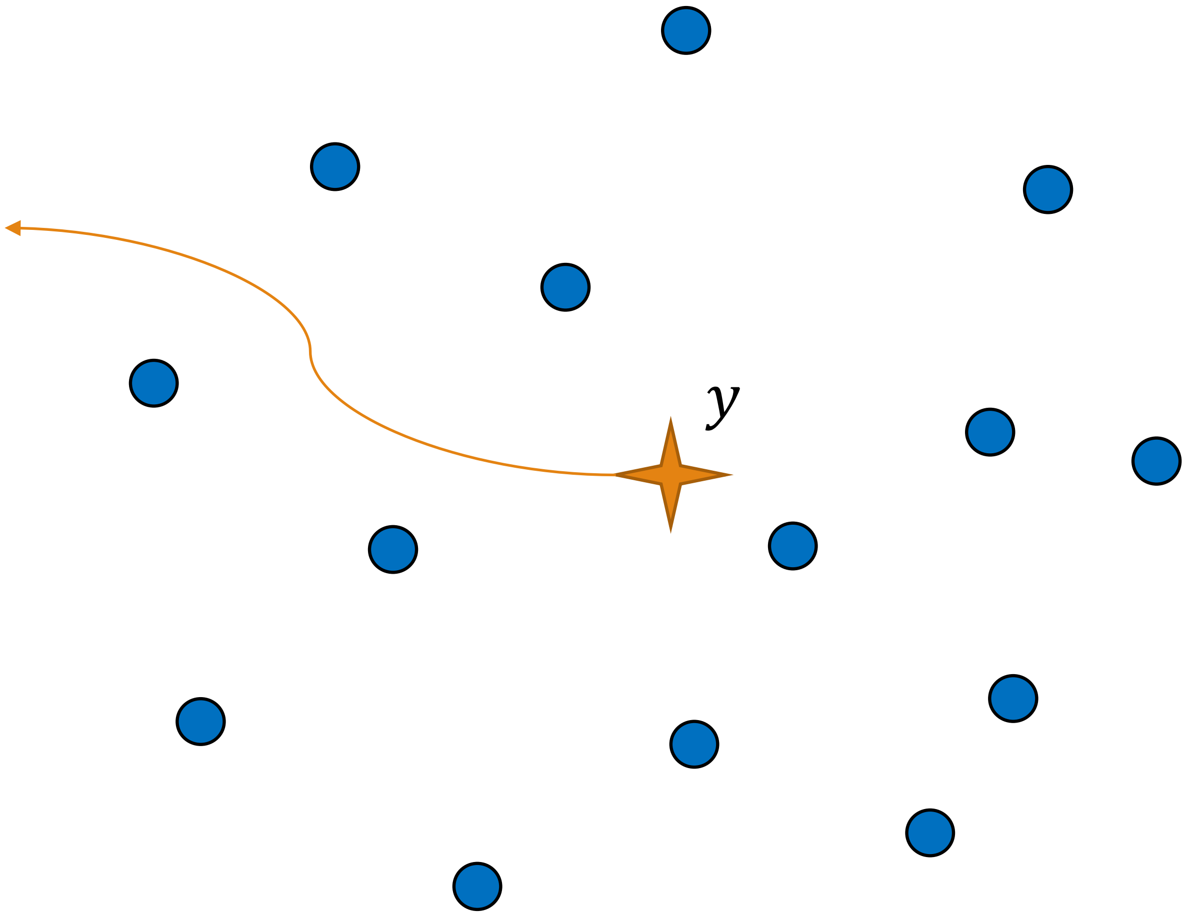
Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$$\boldsymbol{y \in arg\min_{y' \in X} \sum_{p \in P} dist(p, y')}$$

$$\boldsymbol{OPT = \sum_{p \in P} dist(p, y)}$$



Definitions

Let $(P, X, dist)$ be a query space, where $dist: P \times X \to [0, \infty)$.
For every $p \in P$ and $Y \subseteq X$ define $dist(p, Y) = \min_{y \in Y} dist(p, y)$.

- $OPT = \min_{y \in X} \sum_{p \in P} dist(p, y)$.

Definitions

Let $(P, X, dist)$ be a query space, where $dist: P \times X \to [0, \infty)$.
For every $p \in P$ and $Y \subseteq X$ define $dist(p, Y) = \min_{y \in Y} dist(p, y)$.

- $OPT = \min_{y \in X} \sum_{p \in P} dist(p, y)$.
- y' is an α-approximation if $\sum_{p \in P} dist(p, y') \leq \alpha \cdot OPT$.

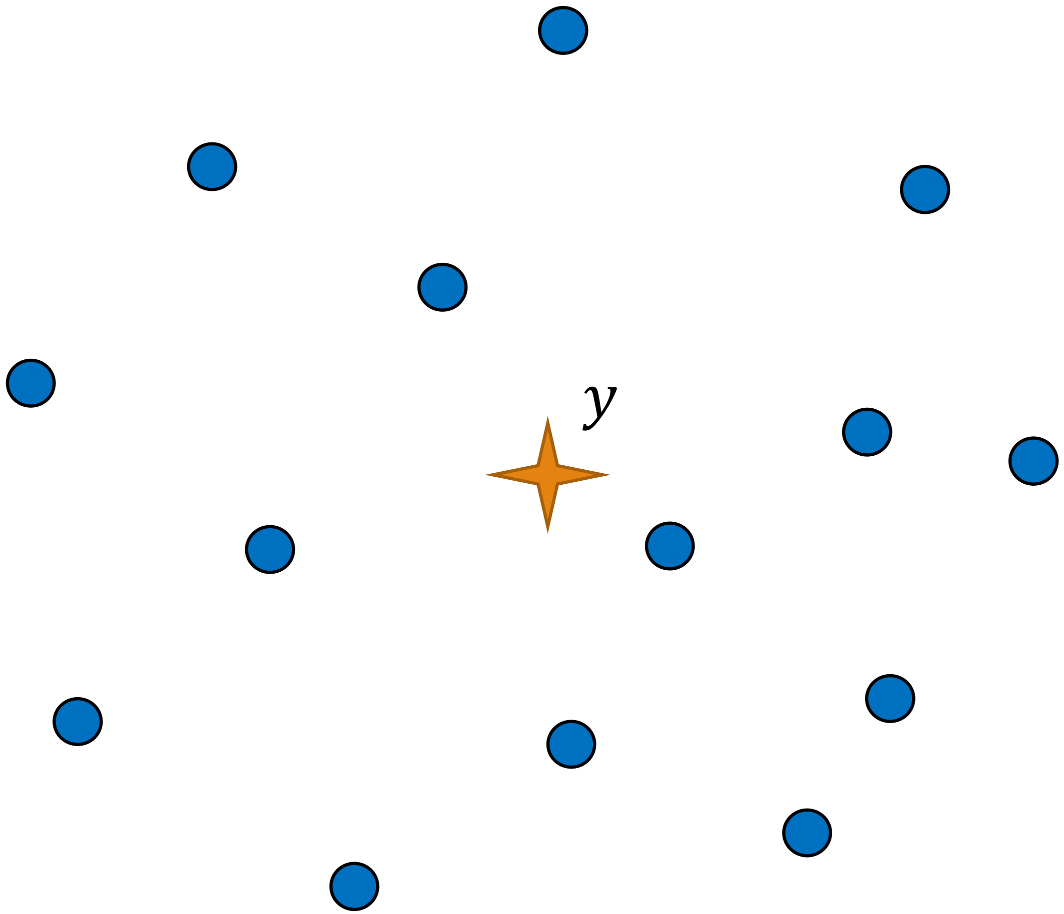
α-approximation Illustration

Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\alpha = 2$



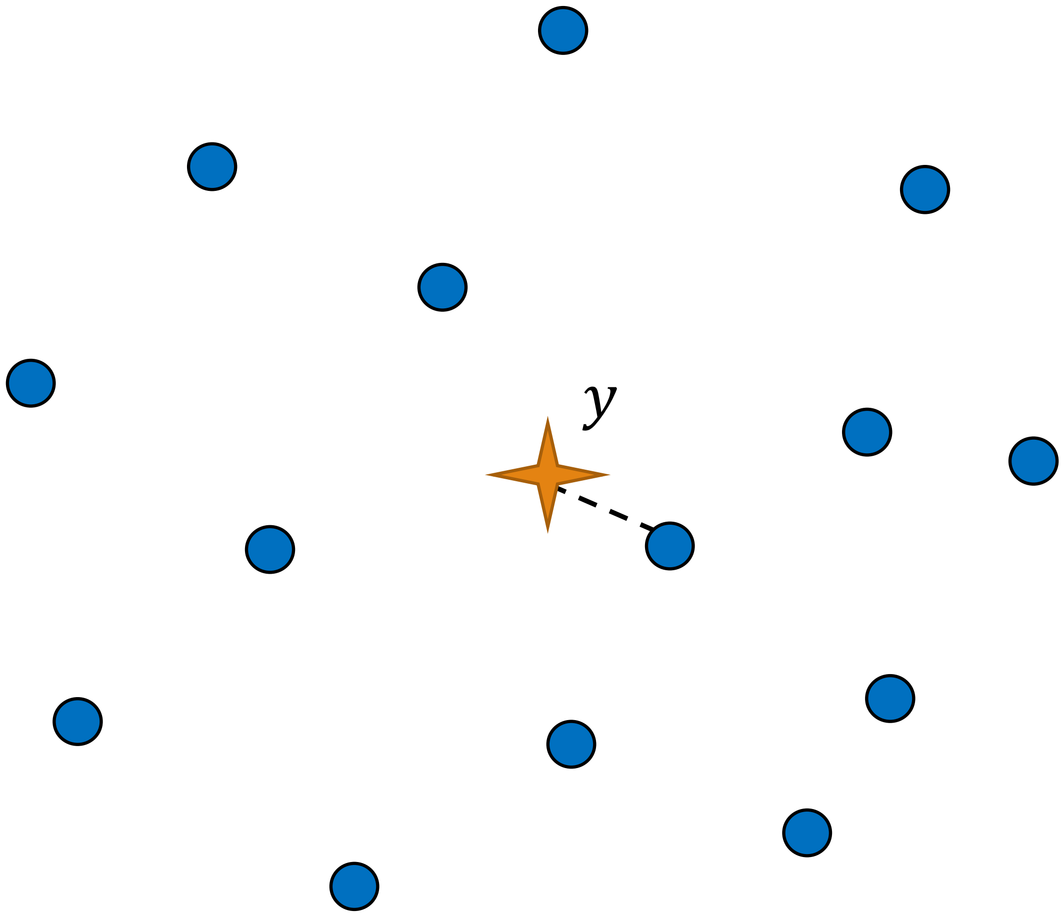
α-approximation Illustration

Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\alpha = 2$



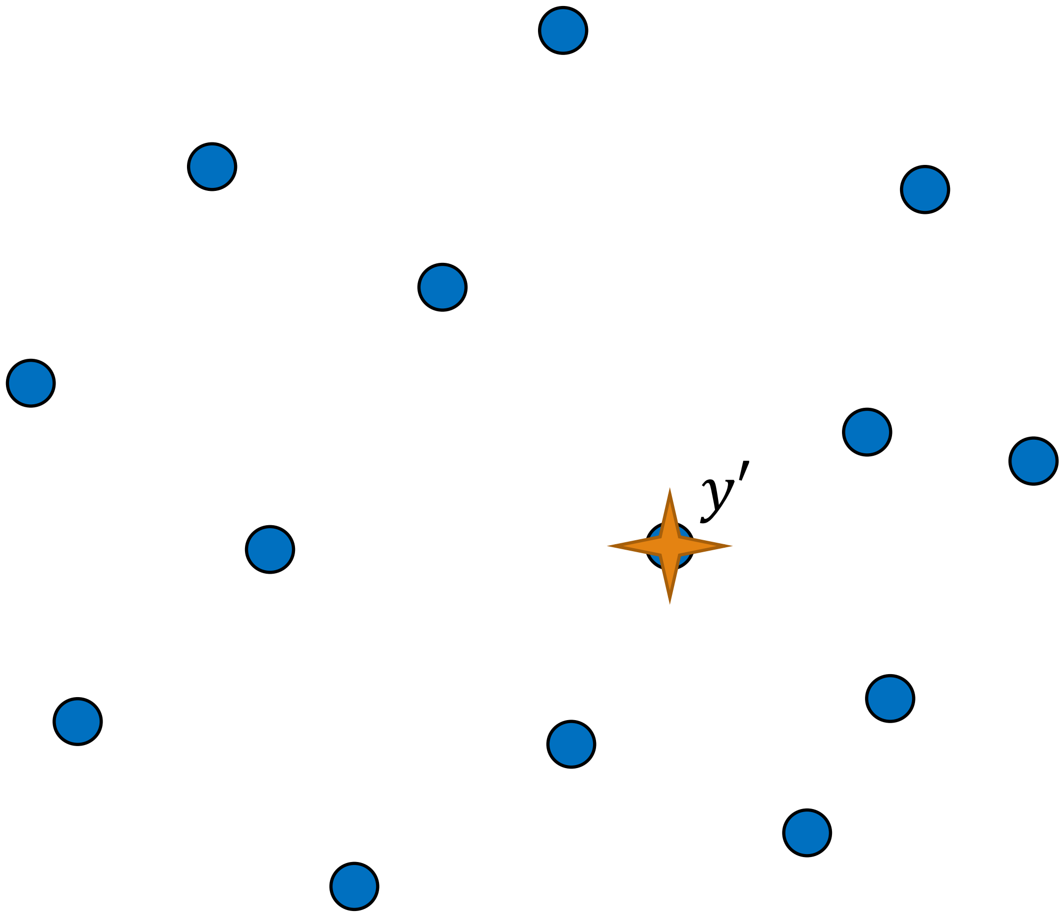
α-approximation Illustration

Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\alpha = 2$



α-approximation Illustration

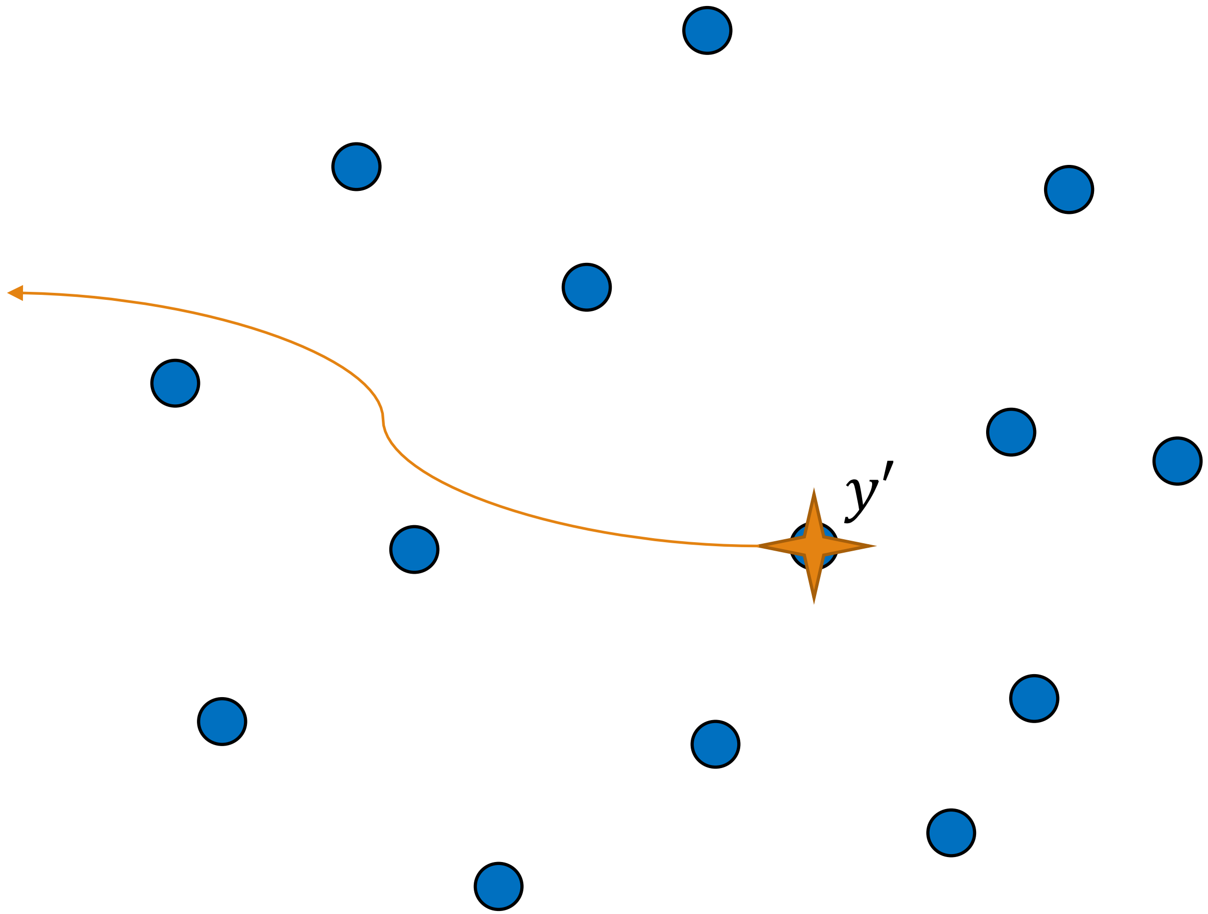
Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\alpha = 2$

$$\sum_{p \in P} dist(p, y') \leq 2 \cdot OPT$$



Definitions

Let $(P, X, dist)$ be a query space, where $dist: P \times X \to [0, \infty)$.
For every $p \in P$ and $Y \subseteq X$ define $dist(p, Y) = \min_{y \in Y} dist(p, y)$.

- $OPT = \min_{y \in X} \sum_{p \in P} dist(p, y)$.
- y' is an α-approximation if $\sum_{p \in P} dist(p, y') \leq \alpha \cdot OPT$.

Definitions

Let $(P, X, dist)$ be a query space, where $dist: P \times X \to [0, \infty)$.
For every $p \in P$ and $Y \subseteq X$ define $dist(p, Y) = \min_{y \in Y} dist(p, y)$.

- $OPT = \min_{y \in X} \sum_{p \in P} dist(p, y)$.
- y' is an α-approximation if $\sum_{p \in P} dist(p, y') \leq \alpha \cdot OPT$.
- $Y \subseteq X$ is a β-approximation if $|Y| = \beta$ and $\sum_{p \in P} dist(p, Y) \leq OPT$.

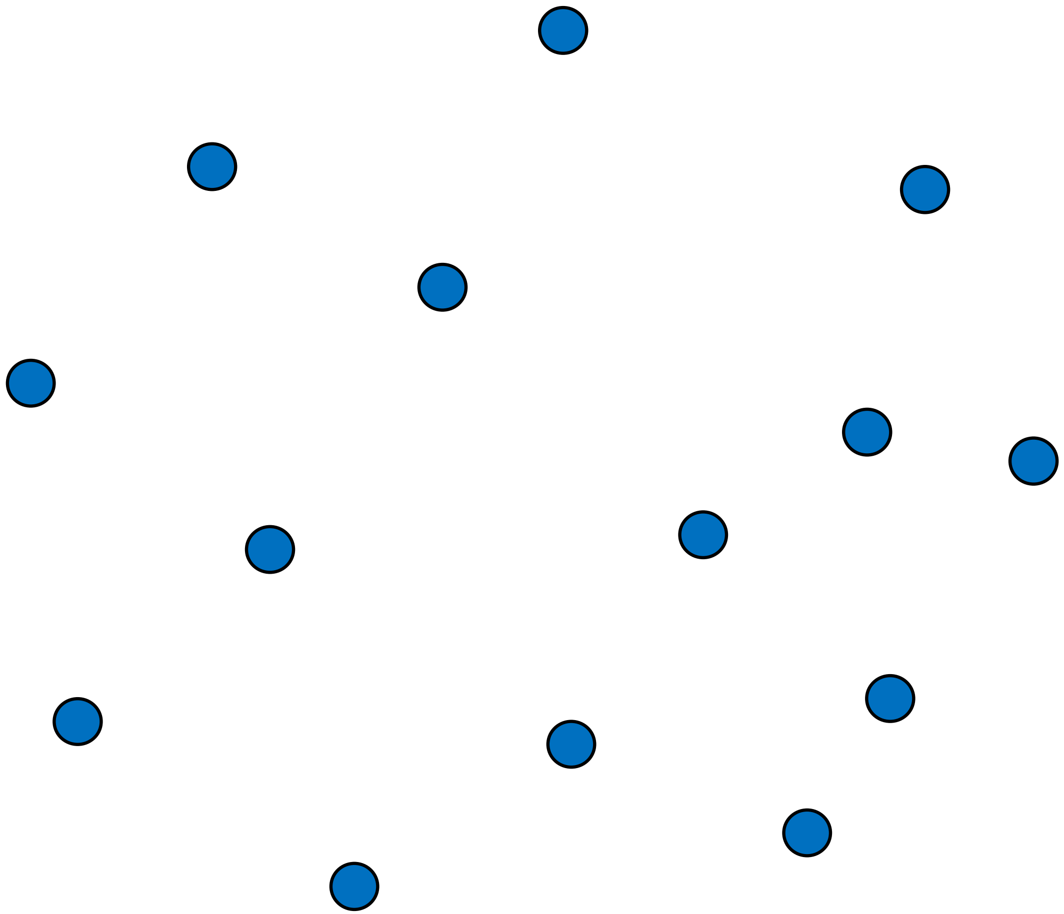
β-approximation Illustration

Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\beta = 2$



β-approximation Illustration

Example:

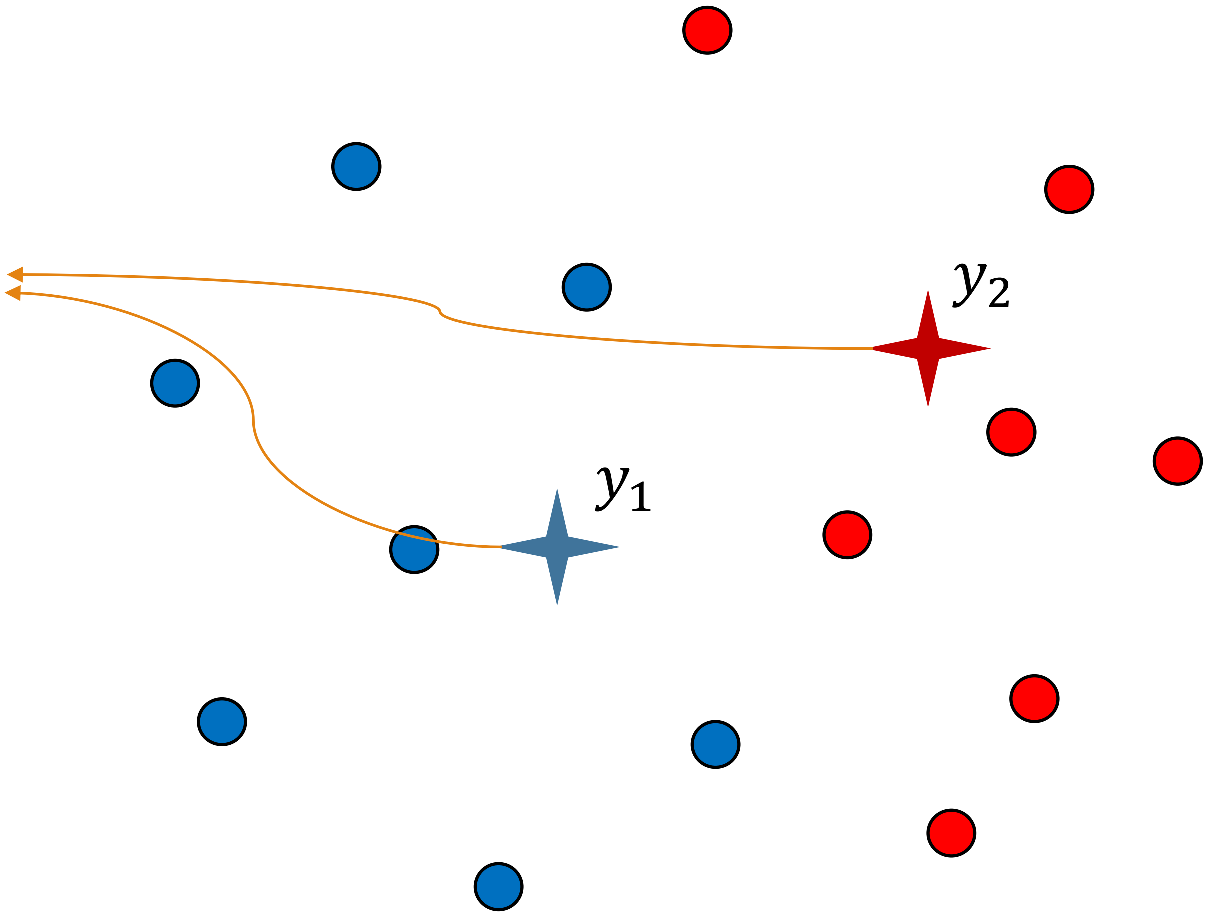
$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\beta = 2$

$$\sum_{p \in P} \boldsymbol{dist(p, Y) \leq OPT}$$

$\boldsymbol{Y = \{y_1, y_2\}, |Y| = \beta}$



Definitions

Let $(P, X, dist)$ be a query space, where $dist: P \times X \to [0, \infty)$.
For every $p \in P$ and $Y \subseteq X$ define $dist(p, Y) = \min_{y \in Y} dist(p, y)$.

- $OPT = \min_{y \in X} \sum_{p \in P} dist(p, y)$.
- y' is an α-approximation if $\sum_{p \in P} dist(p, y') \leq \alpha \cdot OPT$.
- $Y \subseteq X$ is a β-approximation if $|Y| = \beta$ and $\sum_{p \in P} dist(p, Y) \leq OPT$.

Definitions

Let $(P, X, dist)$ be a query space, where $dist: P \times X \to [0, \infty)$.
For every $p \in P$ and $Y \subseteq X$ define $dist(p, Y) = \min_{y \in Y} dist(p, y)$.

- $OPT = \min_{y \in X} \sum_{p \in P} dist(p, y)$.
- y' is an α-approximation if $\sum_{p \in P} dist(p, y') \leq \alpha \cdot OPT$.
- $Y \subseteq X$ is a β-approximation if $|Y| = \beta$ and $\sum_{p \in P} dist(p, Y) \leq OPT$.
- $Y' \subseteq X$ is an (α, β)-approximation if $|Y'| = \beta$ and $\sum_{p \in P} dist(p, Y) \leq \alpha \cdot OPT$

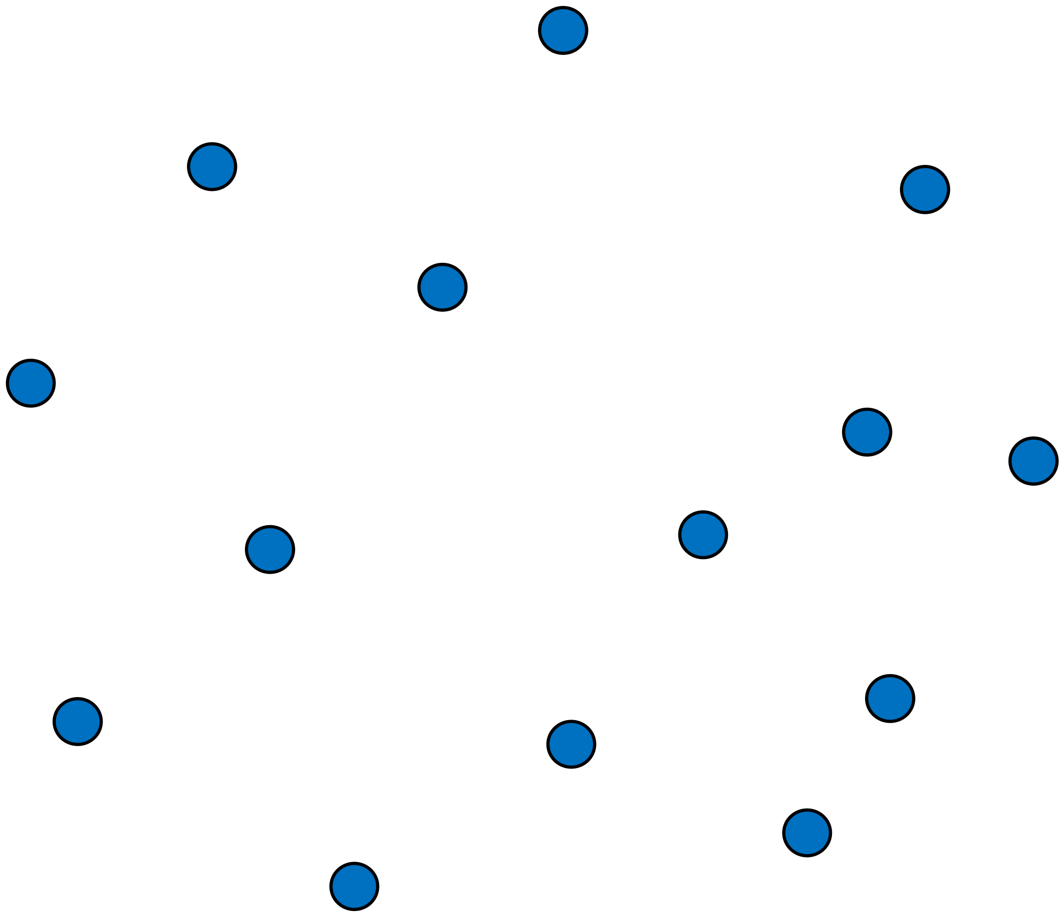
(α, β)-approximation Illustration

<u>**Example:**</u>

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\alpha = 2, \beta = 2$



(α, β)-approximation Illustration

Example:

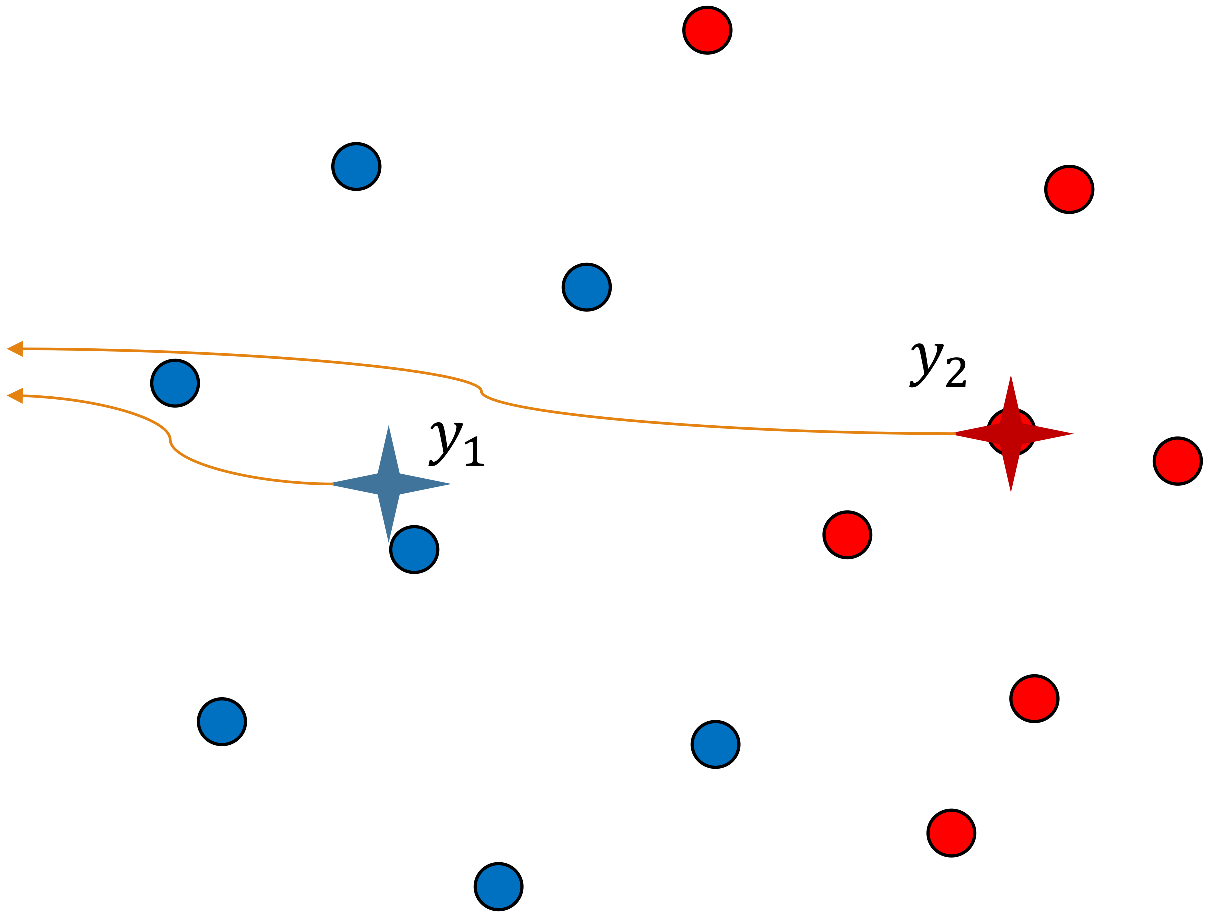
$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\alpha = 2, \beta = 2$

$$\sum_{p \in P} dist(p, Y) \leq 2OPT$$

$Y = \{y_1, y_2\}, |Y| = 2$



Definitions

Let P be an input set of n elements, X be a query space and $dist: P \times X \to [0, \infty)$.
For every $p \in P$ and $Y \subseteq X$ define $dist(p, Y) = \min_{y \in Y} dist(p, y)$.

- $OPT = \min_{y \in X} \sum_{p \in P} dist(p, y)$.
- y' is an α-approximation if $\sum_{p \in P} dist(p, y') \leq \alpha \cdot OPT$.
- $Y \subseteq X$ is a β-approximation if $|Y| = \beta$ and $\sum_{p \in P} dist(p, Y) \leq OPT$.
- $Y' \subseteq X$ is an (α, β)-approximation if $|Y'| = \beta$ and $\sum_{p \in P} dist(p, Y) \leq \alpha \cdot OPT$

Definitions

Let $(P, X, dist)$ be a query space, where $dist: P \times X \to [0, \infty)$.
For every $p \in P$ and $Y \subseteq X$ define $dist(p, Y) = \min_{y \in Y} dist(p, y)$.

- $OPT = \min_{y \in X} \sum_{p \in P} dist(p, y)$.
- y' is an α-approximation if $\sum_{p \in P} dist(p, y') \leq \alpha \cdot OPT$.
- $Y \subseteq X$ is a β-approximation if $|Y| = \beta$ and $\sum_{p \in P} dist(p, Y) \leq OPT$.
- $Y' \subseteq X$ is an (α, β)-approximation if $|Y'| = \beta$ and $\sum_{p \in P} dist(p, Y) \leq \alpha \cdot OPT$

Define $Closest(P, Y, \gamma)$ to be the $\lceil \gamma n \rceil$ points $p \in P$ with smallest value $dist(p, Y)$.

- $\gamma\text{-}Robust\text{-}OPT = \min_{y \in X} \sum_{p \in Closest(P, y, \gamma)} dist(p, y)$.

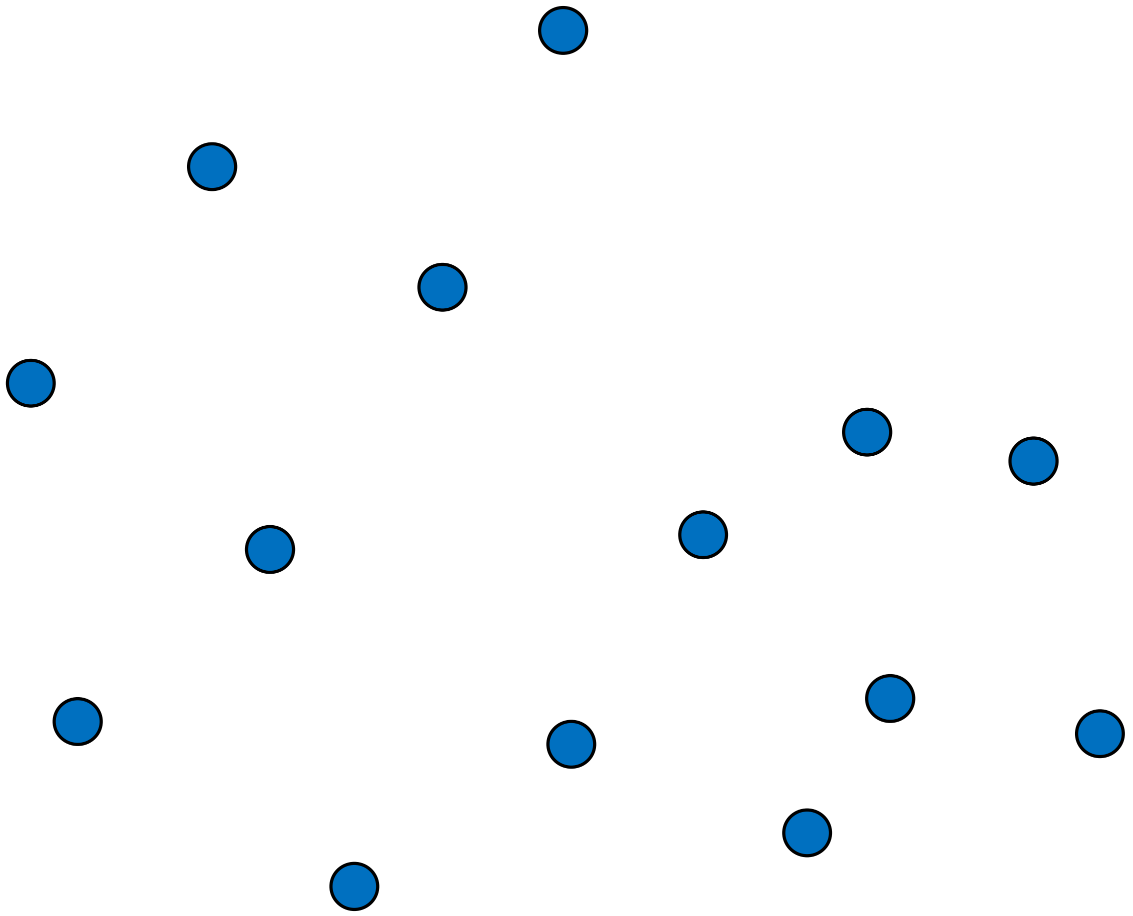
γ-$Robust$-OPT Illustration

Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\gamma = \frac{1}{2}, \lceil \gamma n \rceil = 7$



γ-*Robust-OPT* Illustration

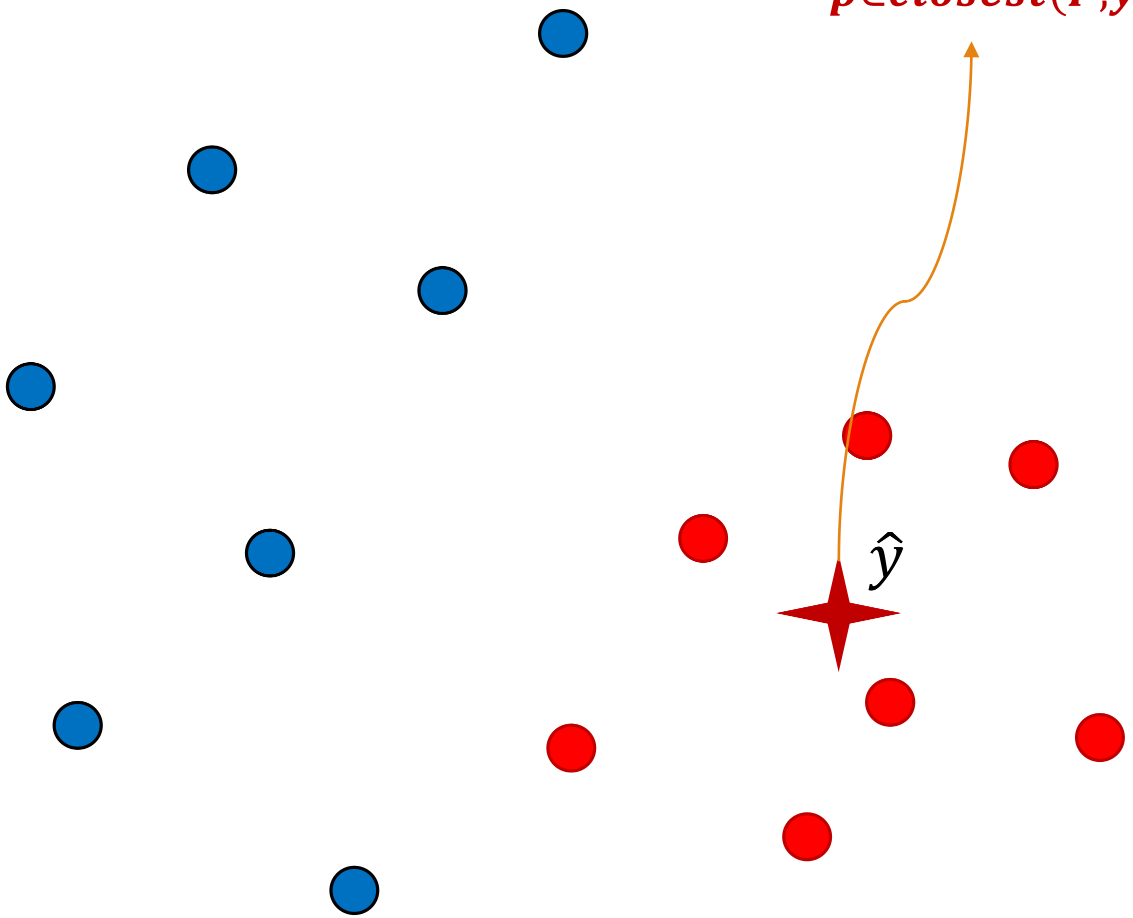
Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p,y) = \|p - y\|$

$\gamma = \frac{1}{2}, \lceil \gamma n \rceil = 7$

$$\hat{y} \in \arg\min_{y \in X} \sum_{p \in closest(P,y,\frac{1}{2})} dist(p,y)$$



γ-$Robust$-OPT Illustration

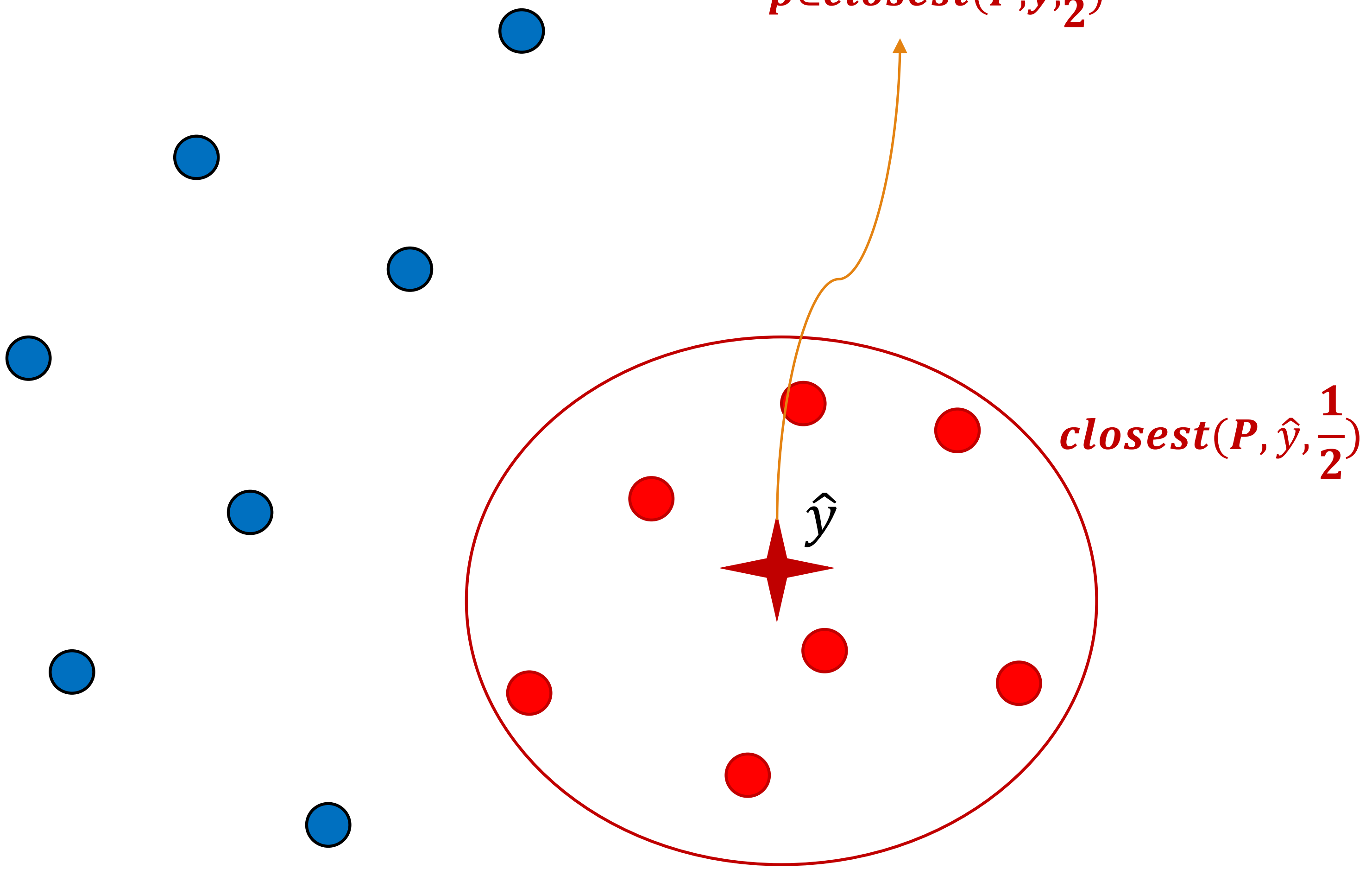
Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\gamma = \frac{1}{2}, \lceil \gamma n \rceil = 7$

$$\hat{y} \in \arg\min_{y \in X} \sum_{p \in closest(P, y, \frac{1}{2})} dist(p, y)$$



Definitions

Let $(P, X, dist)$ be a query space, where $dist: P \times X \to [0, \infty)$.
For every $p \in P$ and $Y \subseteq X$ define $dist(p, Y) = \min_{y \in Y} dist(p, y)$.

- $OPT = \min_{y \in X} \sum_{p \in P} dist(p, y)$.
- y' is an α-approximation if $\sum_{p \in P} dist(p, y') \leq \alpha \cdot OPT$.
- $Y \subseteq X$ is a β-approximation if $|Y| = \beta$ and $\sum_{p \in P} dist(p, Y) \leq OPT$.
- $Y' \subseteq X$ is an (α, β)-approximation if $|Y'| = \beta$ and $\sum_{p \in P} dist(p, Y) \leq \alpha \cdot OPT$

Define $Closest(P, Y, \gamma)$ to be the $\lceil \gamma n \rceil$ points $p \in P$ with smallest value $dist(p, Y)$.

- $\gamma\text{-}Robust\text{-}OPT = \min_{y \in X} \sum_{p \in Closest(P, y, \gamma)} dist(p, y)$.

Definitions

Define $Closest(P, Y, \gamma)$ to be the $\lceil \gamma n \rceil$ points $p \in P$ with smallest value $dist(p, Y)$.

- $\gamma\text{-}Robust\text{-}OPT = \min_{y \in X} \sum_{p \in Closest(P,y,\gamma)} dist(p, y)$.

Definitions

Define $Closest(P, Y, \gamma)$ to be the $\lceil \gamma n \rceil$ points $p \in P$ with smallest value $dist(p, Y)$.

- $\gamma\text{-}Robust\text{-}OPT = \min_{y \in X} \sum_{p \in Closest(P,y,\gamma)} dist(p, y)$.
- $Y' \subseteq X$ is a (γ, α, β)-approximation if $|Y'| = \beta$ and

$$\sum_{p \in Closest(P,Y',\gamma)} dist(p, Y') \leq \alpha \cdot (\gamma - Robust - OPT)$$

(γ, α, β)-approximation Illustration

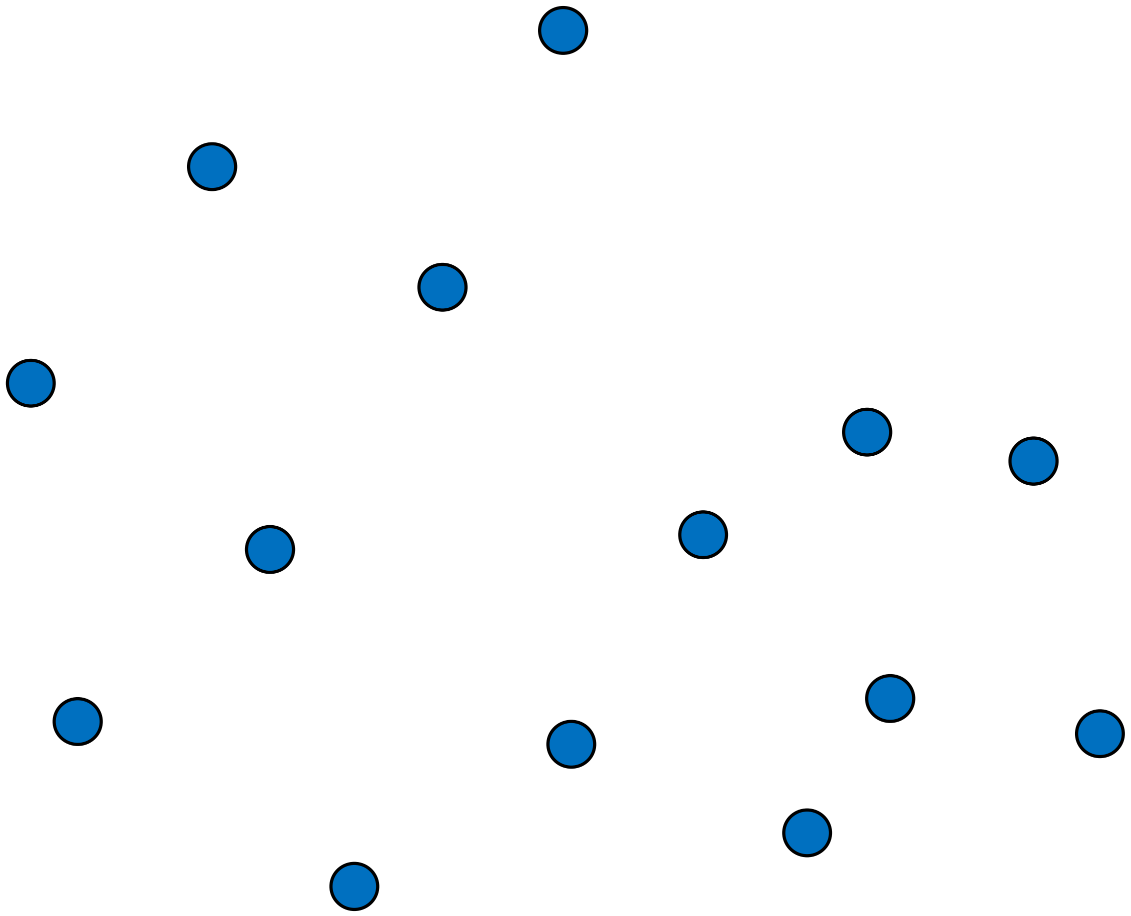
<u>Example:</u>

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\gamma = \frac{1}{2}, \lceil \gamma n \rceil = 7$

$\alpha = 1, \beta = 2$



(γ, α, β)-approximation Illustration

$$\sum_{p \in closest(P,Y,\gamma)} dist(p, Y) \leq \alpha \cdot (\gamma - \text{Robust} - \text{OPT})$$

$$Y = \{y_1, y_2\}$$

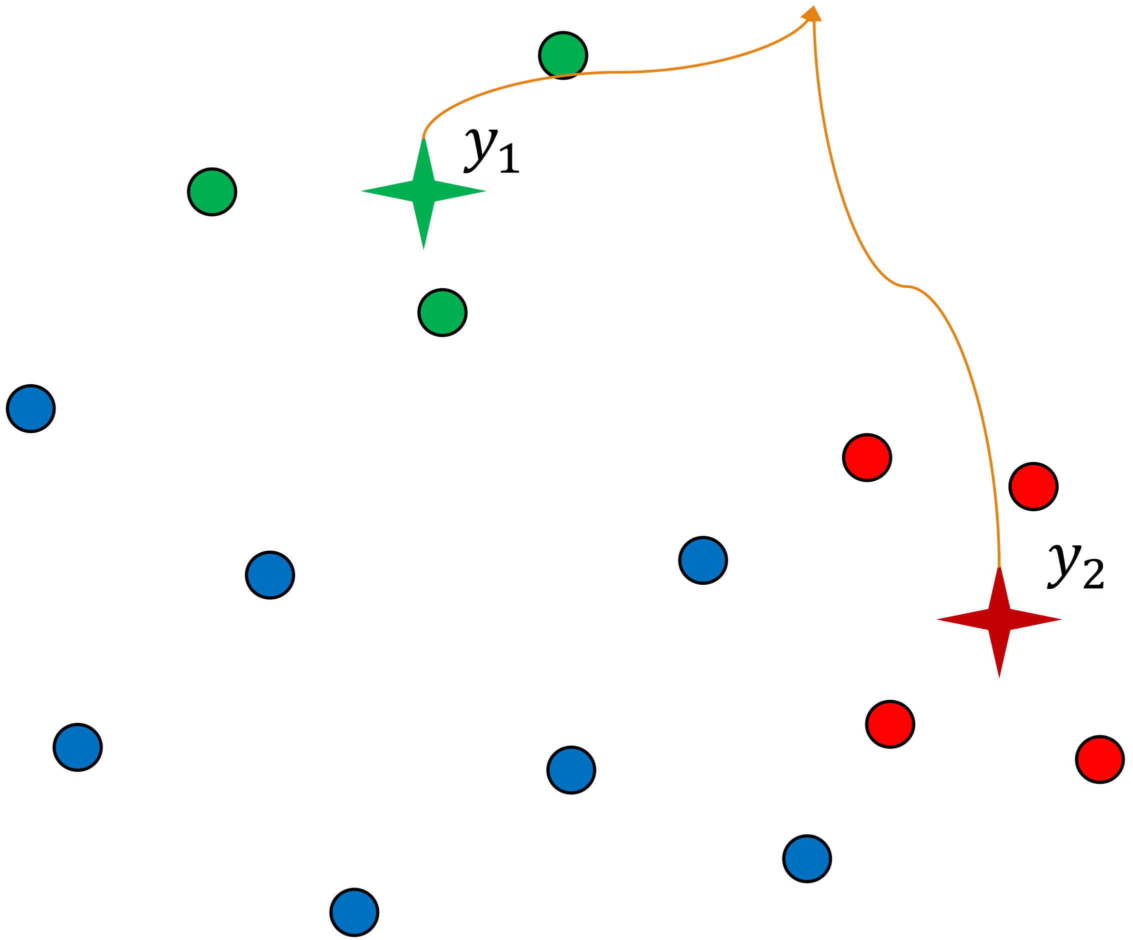
Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\gamma = \frac{1}{2}, \lceil \gamma n \rceil = 7$

$\alpha = 1, \beta = 2$



(γ, α, β)-approximation Illustration

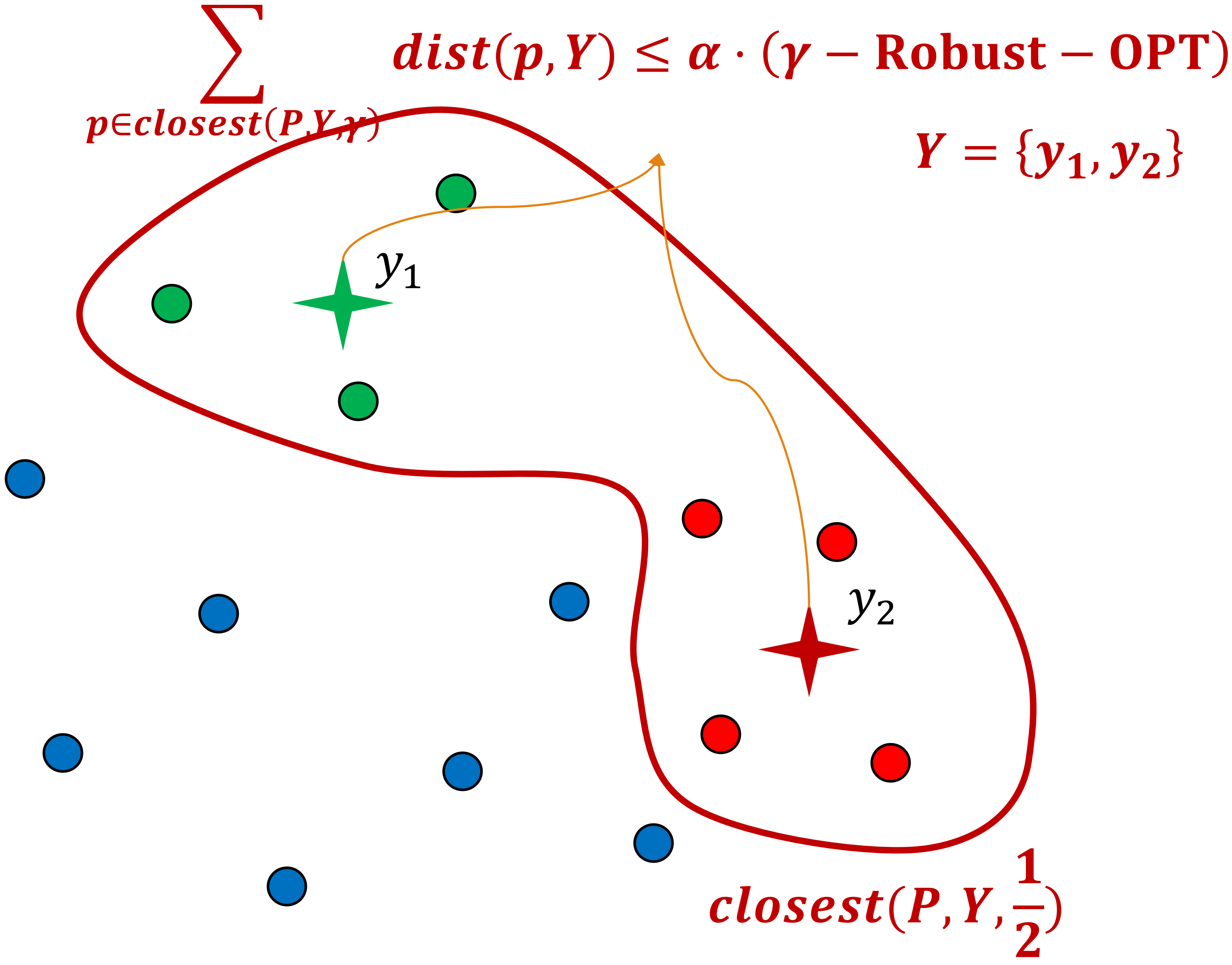
Example:

$P \subseteq R^d, X \subseteq R^d$

$dist(p, y) = \|p - y\|$

$\gamma = \frac{1}{2}, \lceil \gamma n \rceil = 7$

$\alpha = 1, \beta = 2$



Definitions

Define $Closest(P, Y, \gamma)$ to be the $\lceil \gamma n \rceil$ points $p \in P$ with smallest value $dist(p, Y)$.

- $\gamma\text{-}Robust\text{-}OPT = \min_{y \in X} \sum_{p \in Closest(P,y,\gamma)} dist(p, y)$.
- $Y' \subseteq X$ is a (γ, α, β)-approximation if $|Y'| = \beta$ and

$$\sum_{p \in Closest(P,Y',\gamma)} dist(p, Y') \leq \alpha \cdot (\gamma{-}Robust{-}OPT)$$

Definitions

Define $Closest(P, Y, \gamma)$ to be the $\lceil \gamma n \rceil$ points $p \in P$ with smallest value $dist(p, Y)$.

- $\gamma\text{-}Robust\text{-}OPT = \min_{y \in X} \sum_{p \in Closest(P,y,\gamma)} dist(p, y)$.
- $Y' \subseteq X$ is a (γ, α, β)-approximation if $|Y'| = \beta$ and

$$\sum_{p \in Closest(P,Y',\gamma)} dist(p, Y') \leq \alpha \cdot (\gamma - Robust - OPT)$$

- $Y' \subseteq X$ is a $(\gamma, \epsilon, \alpha, \beta)$-approximation if $|Y'| = \beta$ and

$$\sum_{p \in Closest(P,Y',(1-\epsilon)\gamma)} dist(p, Y') \leq \alpha \cdot (\gamma - Robust - OPT)$$

Definitions (k-centers)

Let P be an input set of n elements, X be a query space and $dist: P \times X \to [0, \infty)$.
For every $p \in P$ and $Y \subseteq X$ define $dist(p, Y) = \min_{y \in Y} dist(p, y)$.

- $OPT_k = \min_{Y \subseteq X, |Y| = k} \sum_{p \in P} dist(p, Y)$.
- Y' is an α_k-approximation if $|Y'| = k$ and $\sum_{p \in P} dist(p, Y') \leq \alpha \cdot OPT_k$.
- $Y \subseteq X$ is a β_k-approximation if $|Y| = \beta k$ and $\sum_{p \in P} dist(p, Y) \leq OPT_k$.
- $Y' \subseteq X$ is an $(\alpha, \beta)_k$-approximation if $|Y'| = \beta k$ and $\sum_{p \in P} dist(p, Y) \leq \alpha \cdot OPT_k$

Define $Closest(P, Y, \gamma)$ to be the $\lceil (1 - \gamma) n \rceil$ points $p \in P$ with smallest value $dist(p, Y)$.

- $\gamma\text{-}Robust\text{-}OPT_k = \min_{Y \subseteq X, |Y| = k} \sum_{p \in Closest(P, Y, \gamma)} dist(p, Y)$.

Definitions (k-centers)

Define $Closest(P, Y, \gamma)$ to be the $\lceil(1-\gamma)n\rceil$ points $p \in P$ with smallest value $dist(p, Y)$.

- $\gamma\text{-}Robust\text{-}OPT_k = \min_{Y \subseteq X, |Y|=k} \sum_{p \in Closest(P,Y,\gamma)} dist(p, Y)$.
- $Y' \subseteq X$ is a $(\gamma, \alpha, \beta)_k$-approximation if $|Y'| = \beta k$ and

$$\sum_{p \in Closest(P,Y',\gamma)} dist(p, Y') \leq \alpha \cdot (\gamma{-}Robust{-}OPT_k)$$

- $Y' \subseteq X$ is a $(\gamma, \epsilon, \alpha, \beta)_k$-approximation if $|Y'| = \beta k$ and

$$\sum_{p \in Closest(P,Y',(1-\epsilon)\gamma)} dist(p, Y') \leq \alpha \cdot (\gamma{-}Robust{-}OPT_k)$$

Bi-Criteria / (α, β)-Approximation for k-squares

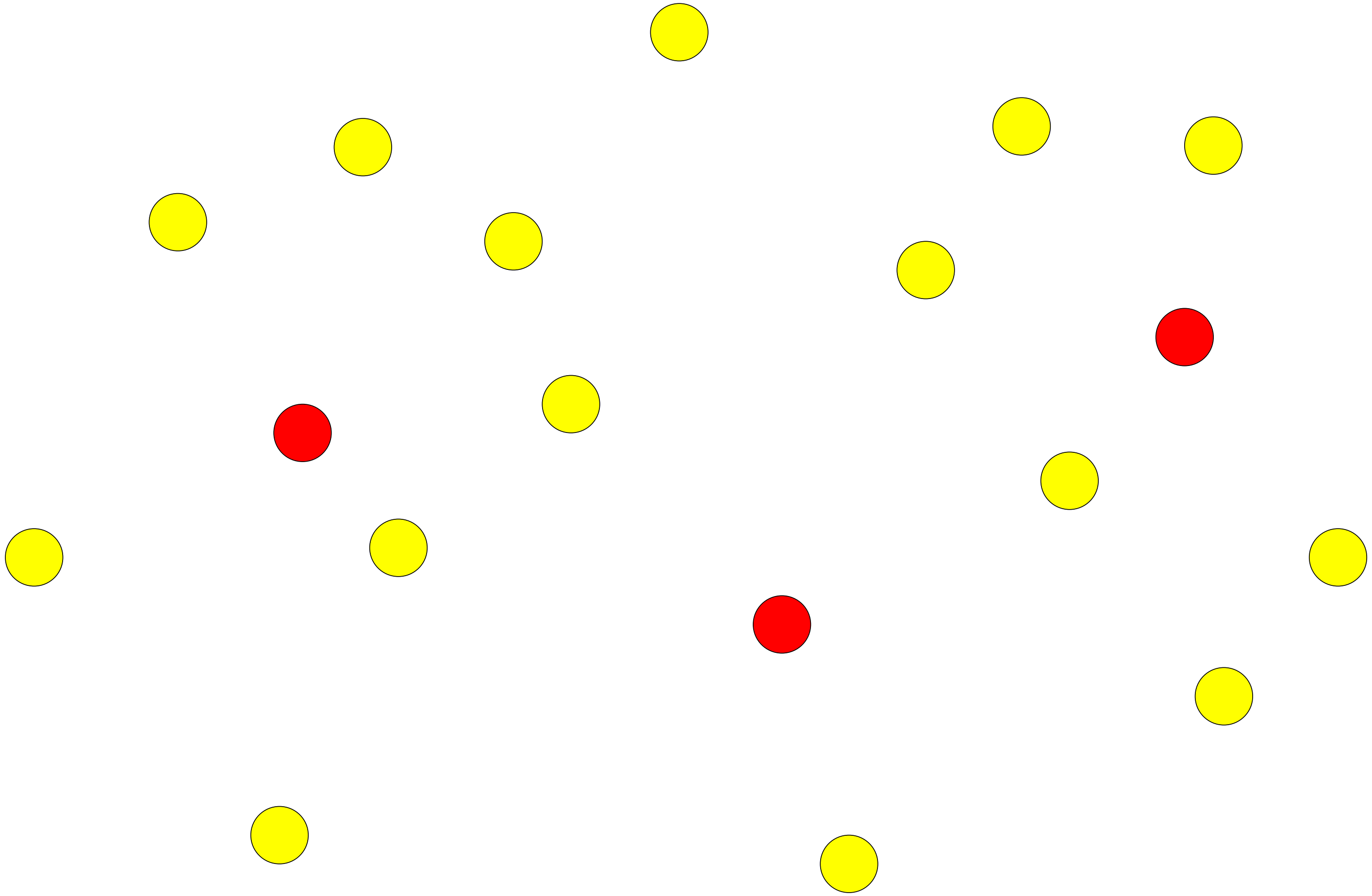
Example:

Initialization

1) $t \leftarrow 1$ (iterations counter)

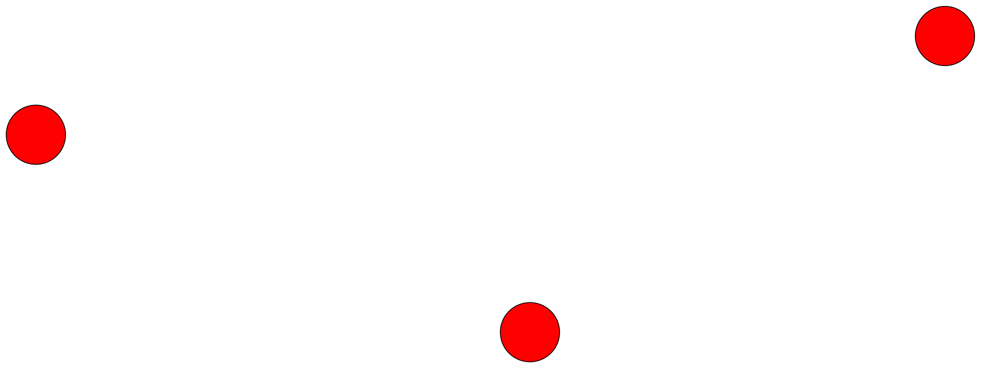
2) $C \leftarrow \emptyset$ (Output)

3) Construct an $C_t = \epsilon$-net for P



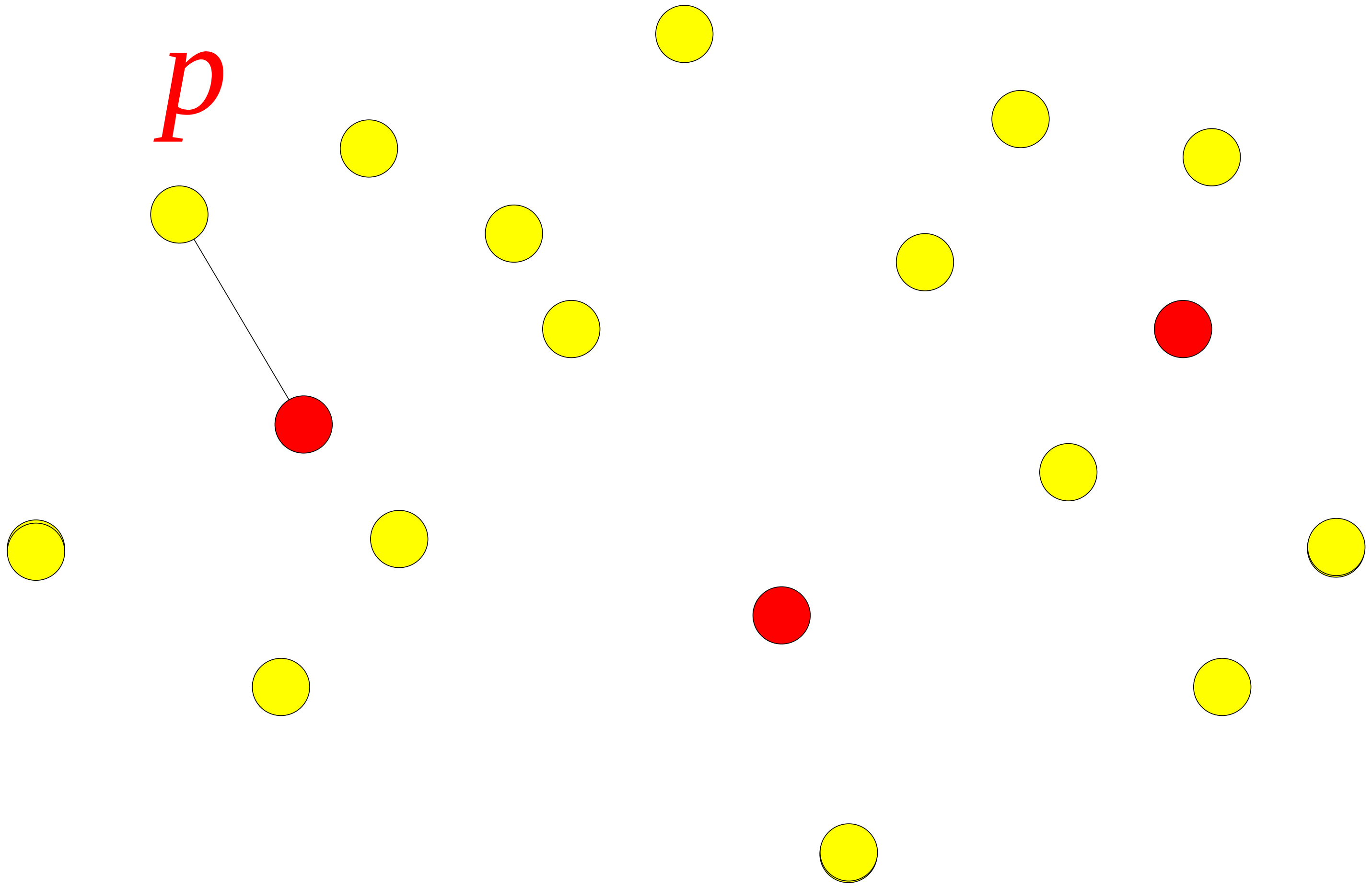
$t = 1$

4) $C \leftarrow C \cup C_t$



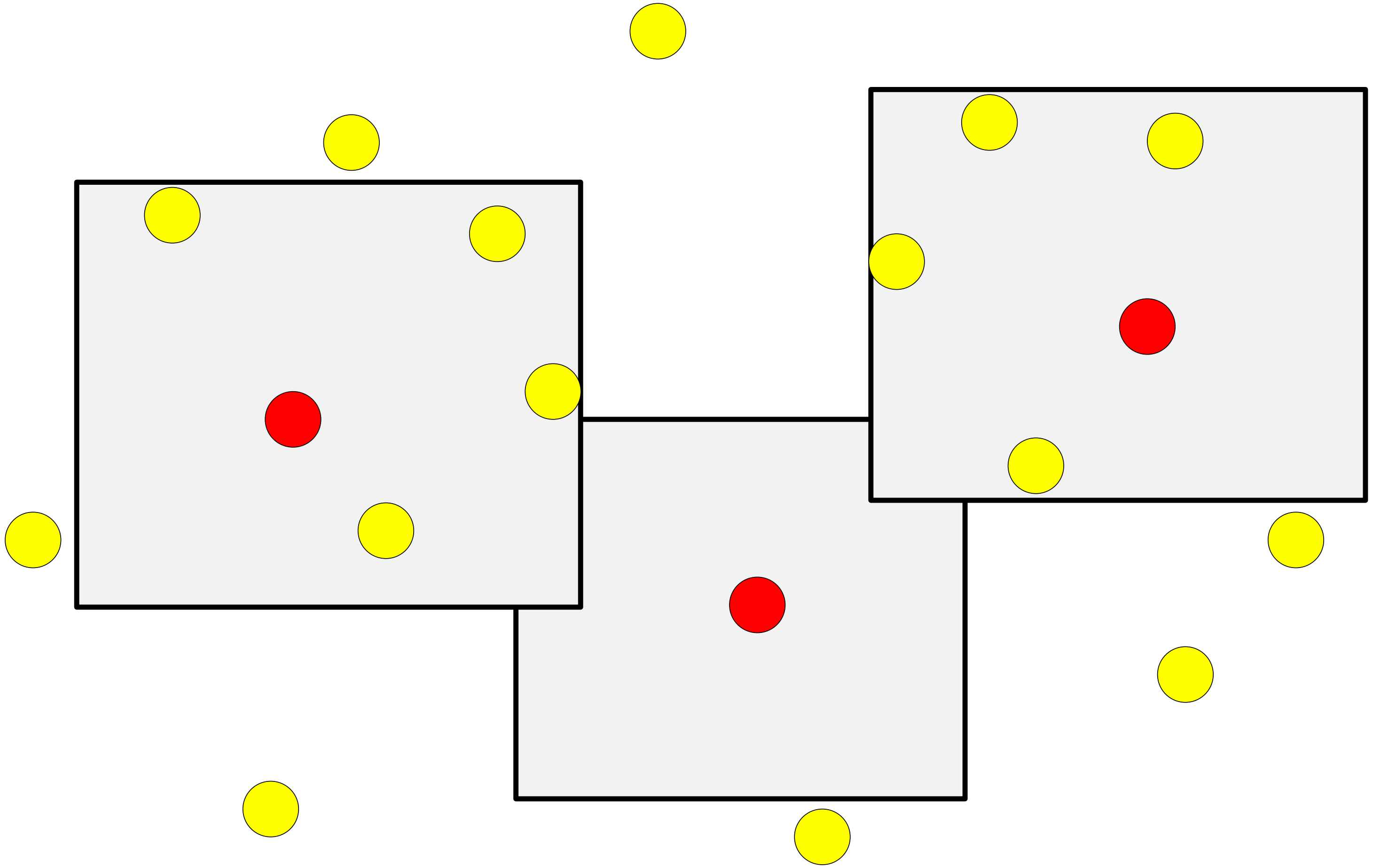
$t = 1$

5) $\forall p$ Compute $far_{\infty}(p, C_t)$



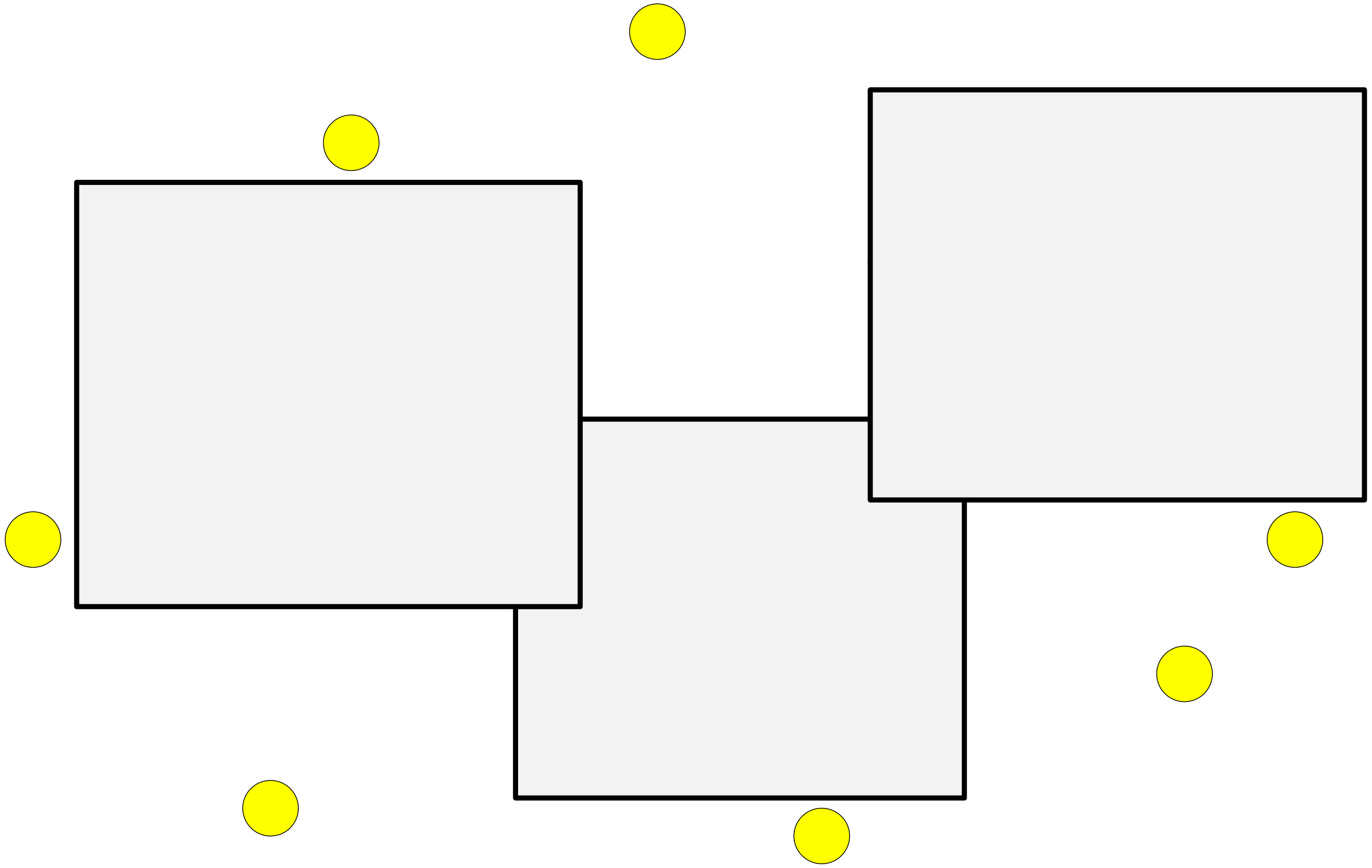
$t = 1$

6) Remove P_t: the points of P that are covered by the optimal squares centered at C_t. $|P_t| > (1-\epsilon)n$.



$t = 1$

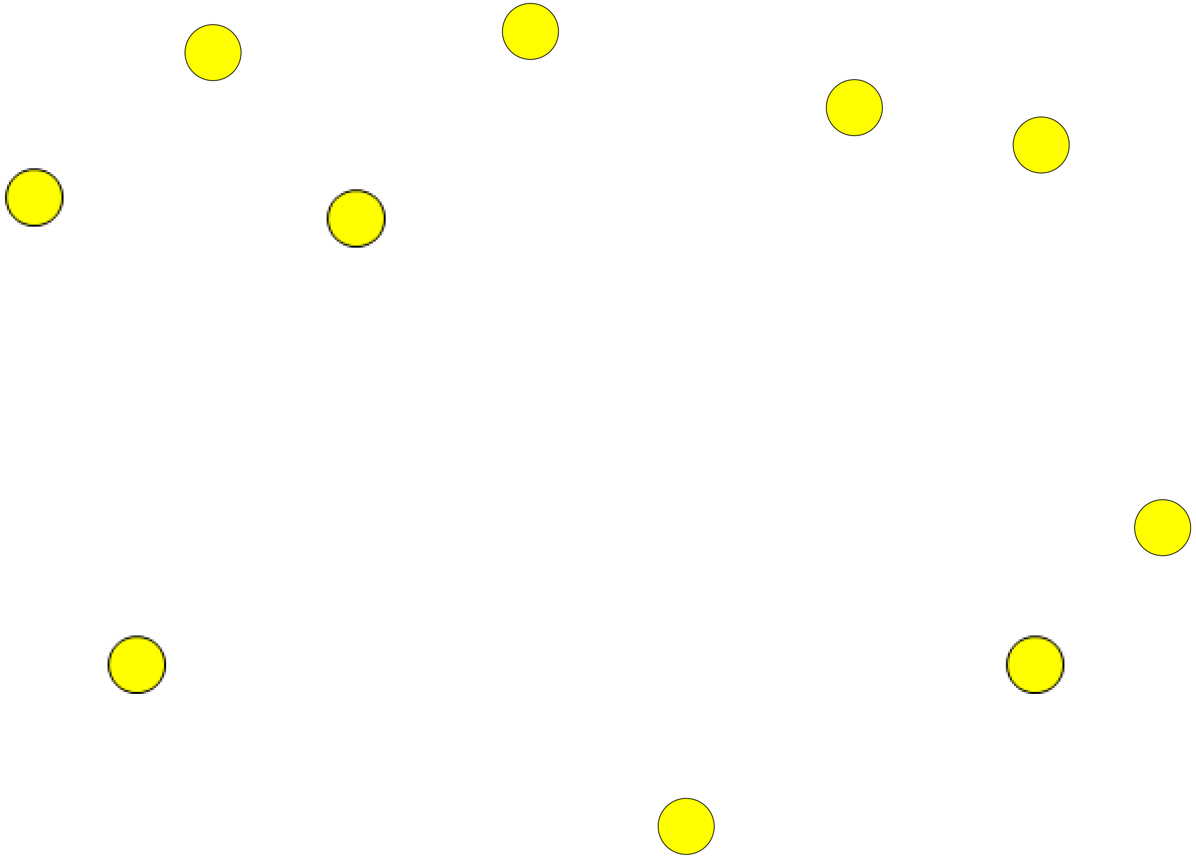
6) Remove P_t: the points of P that are covered by the optimal squares centered at C_t. $|P_t| > (1-\epsilon)n$.



$t = 1$

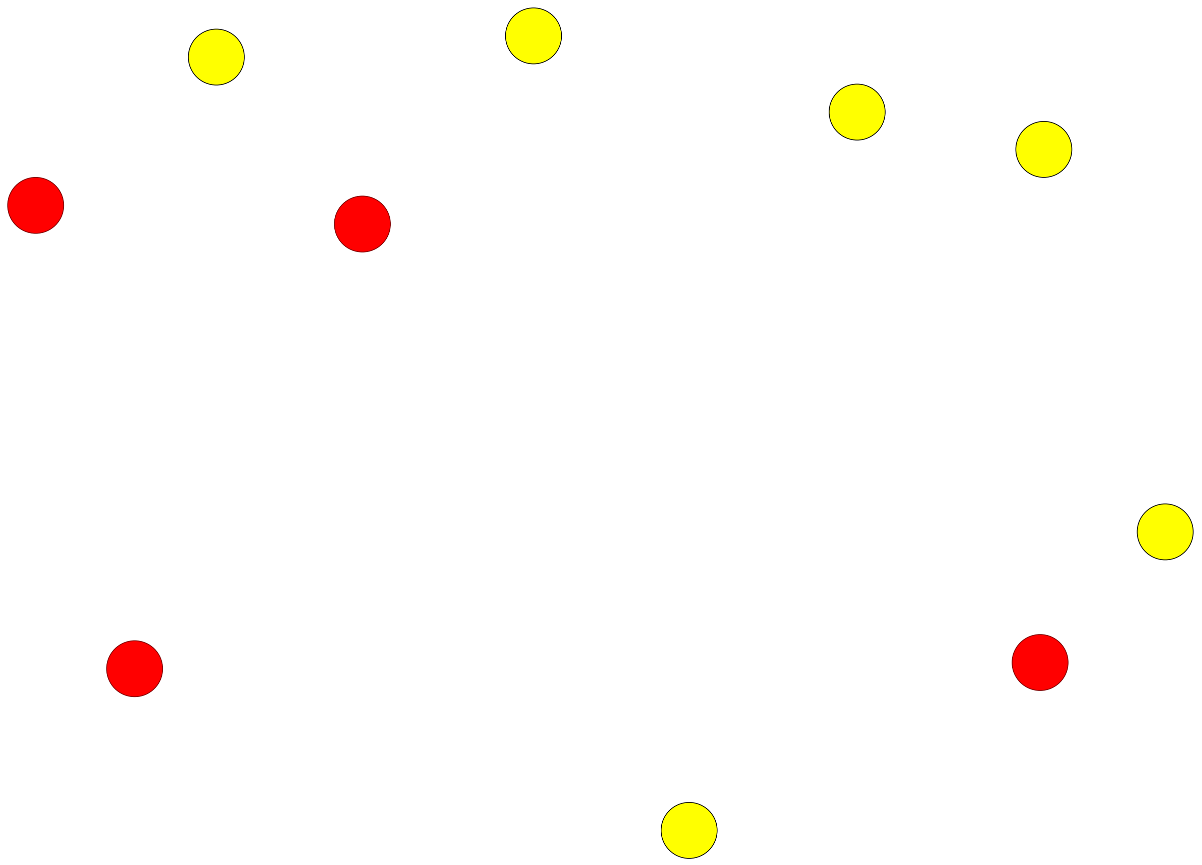
7) $t \leftarrow t + 1$

8) Repeat steps 3 to 6



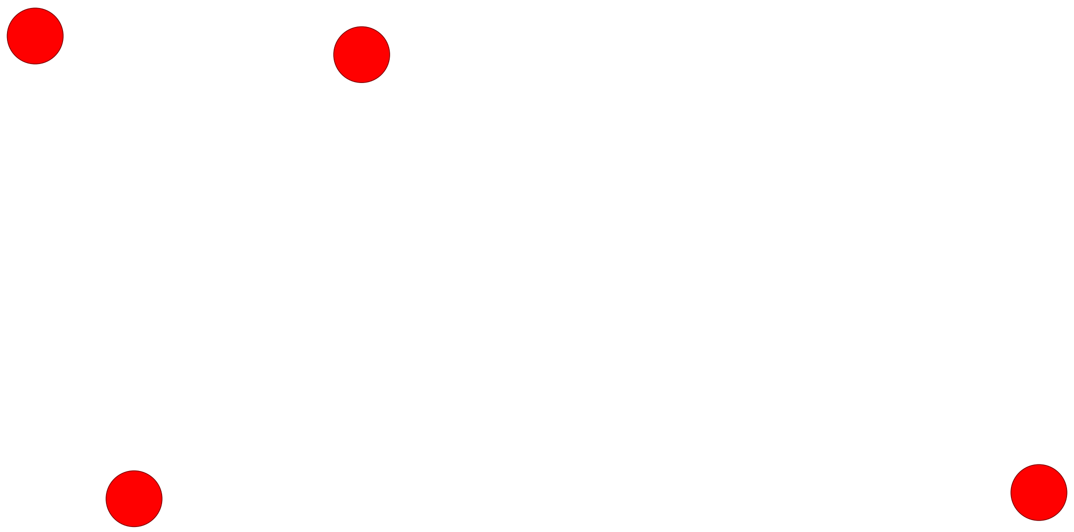
$t = 1$

3) Construct an $C_t = \epsilon$-sample for P



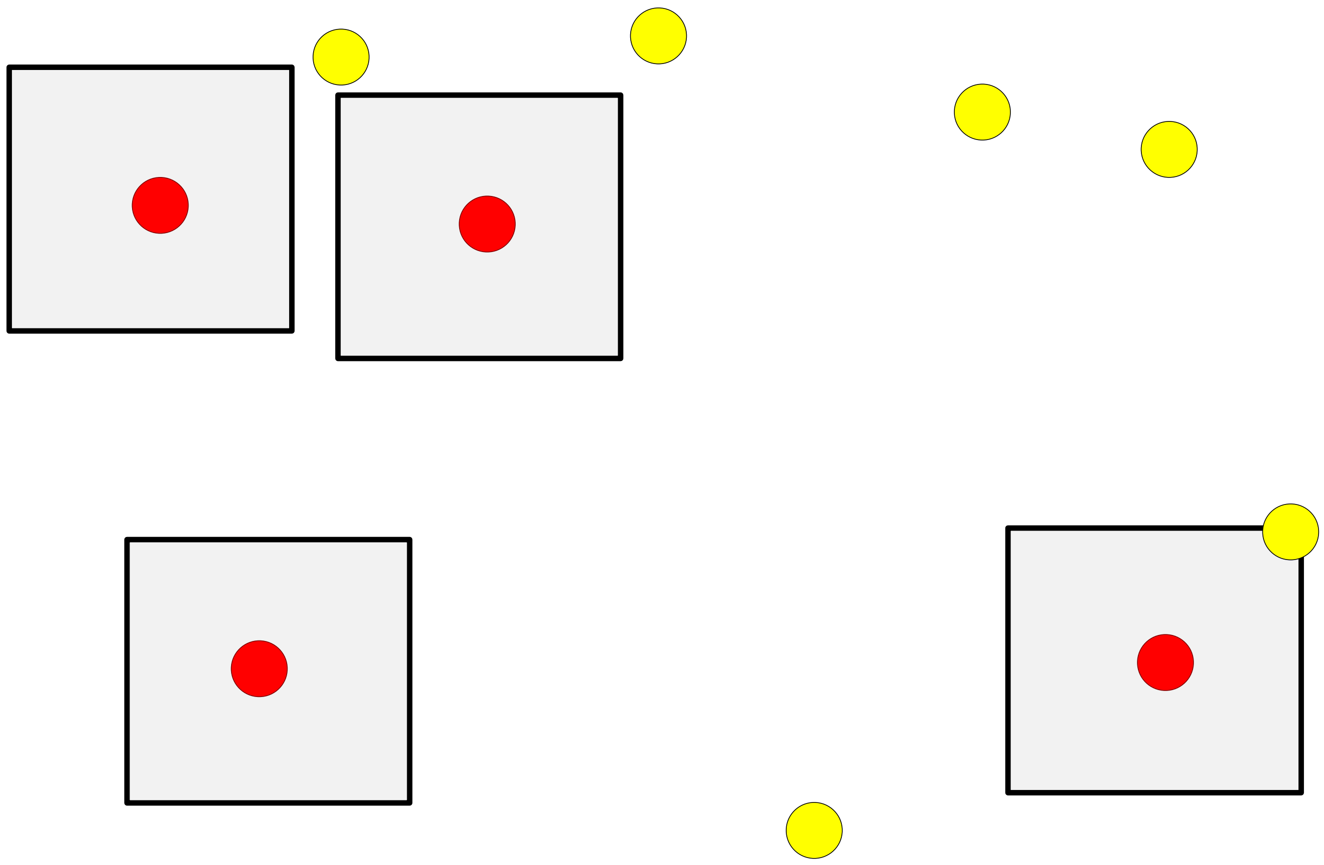
$t = 2$

4) $C \leftarrow C \cup C_t$

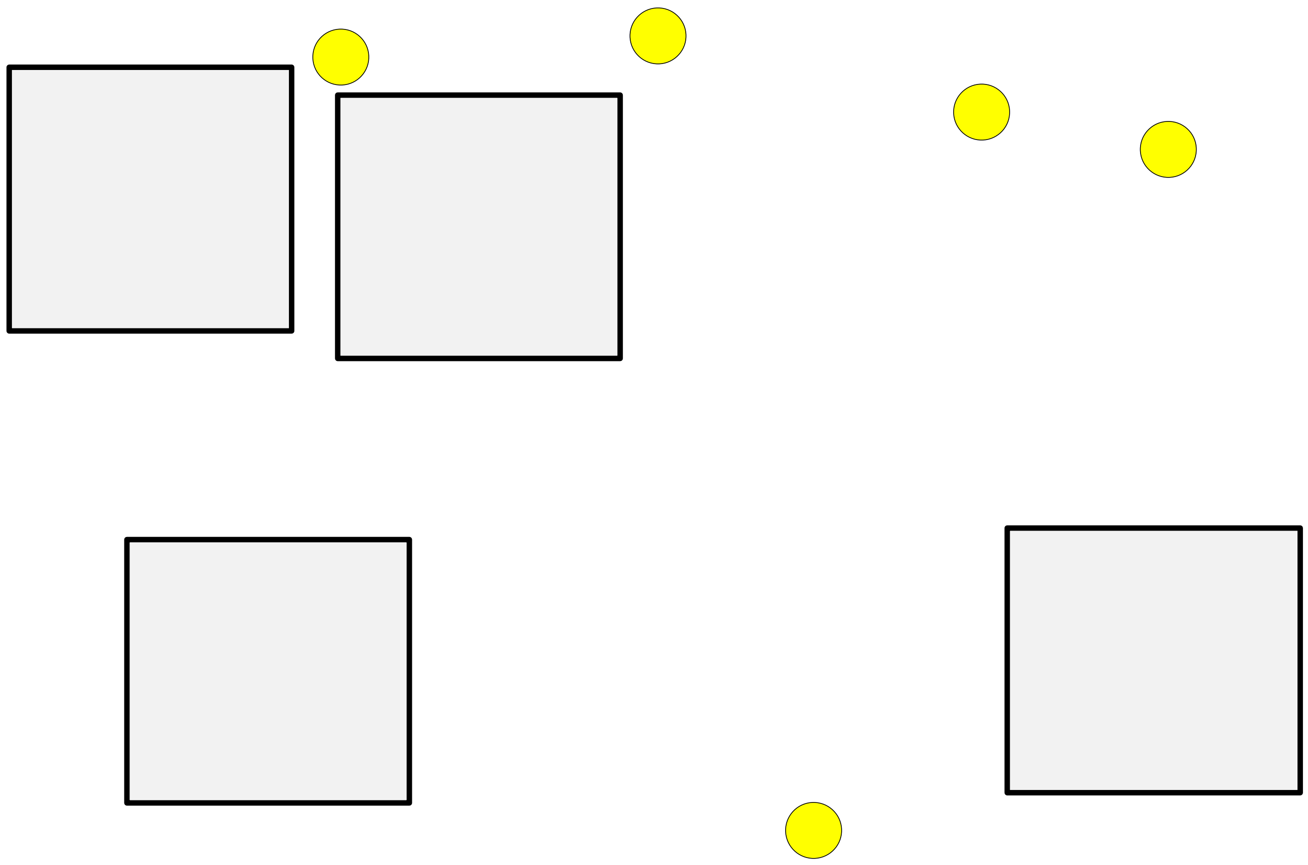


$t = 2$

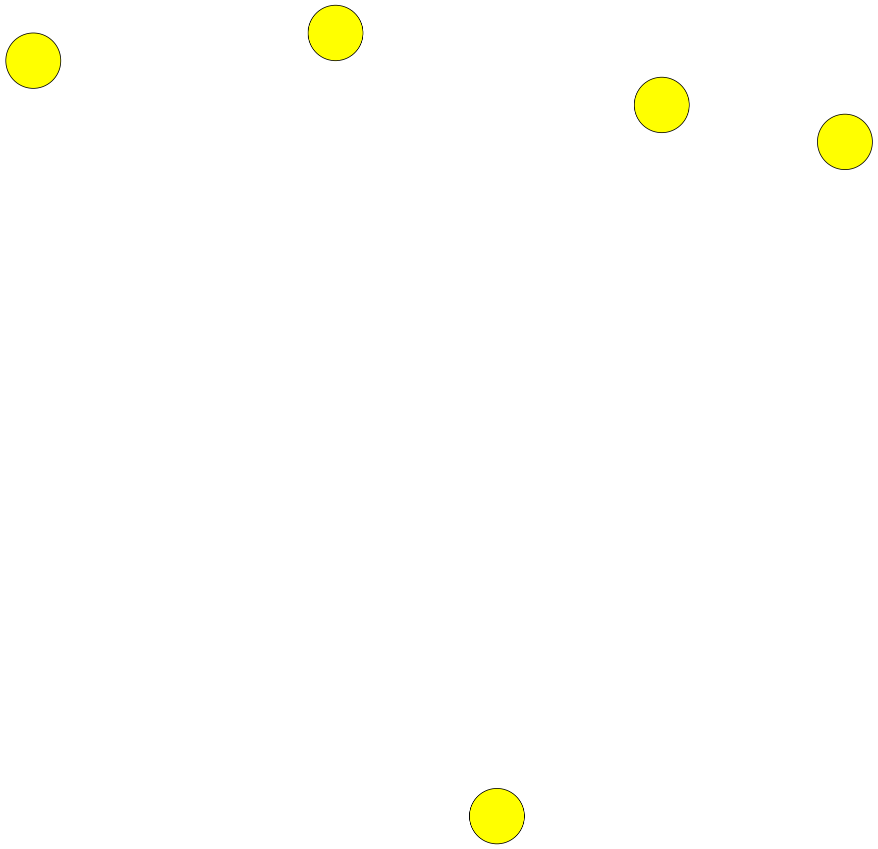
6) Remove P_t: the points of P that are covered by the optimal squares centered at C_t. $|P_t| > (1 - \epsilon)n$.



6) Remove P_t: the points of P that are covered by the optimal squares centered at C_t. $|P_t| > (1 - \epsilon)n$.



6) Remove P_t: the points of P that are covered by the optimal squares centered at C_t. $|P_t| > (1-\epsilon)n$.

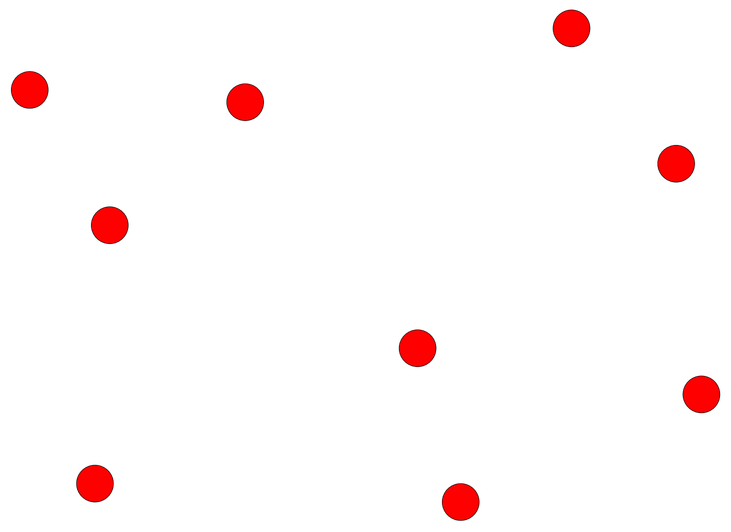


$t = 2$

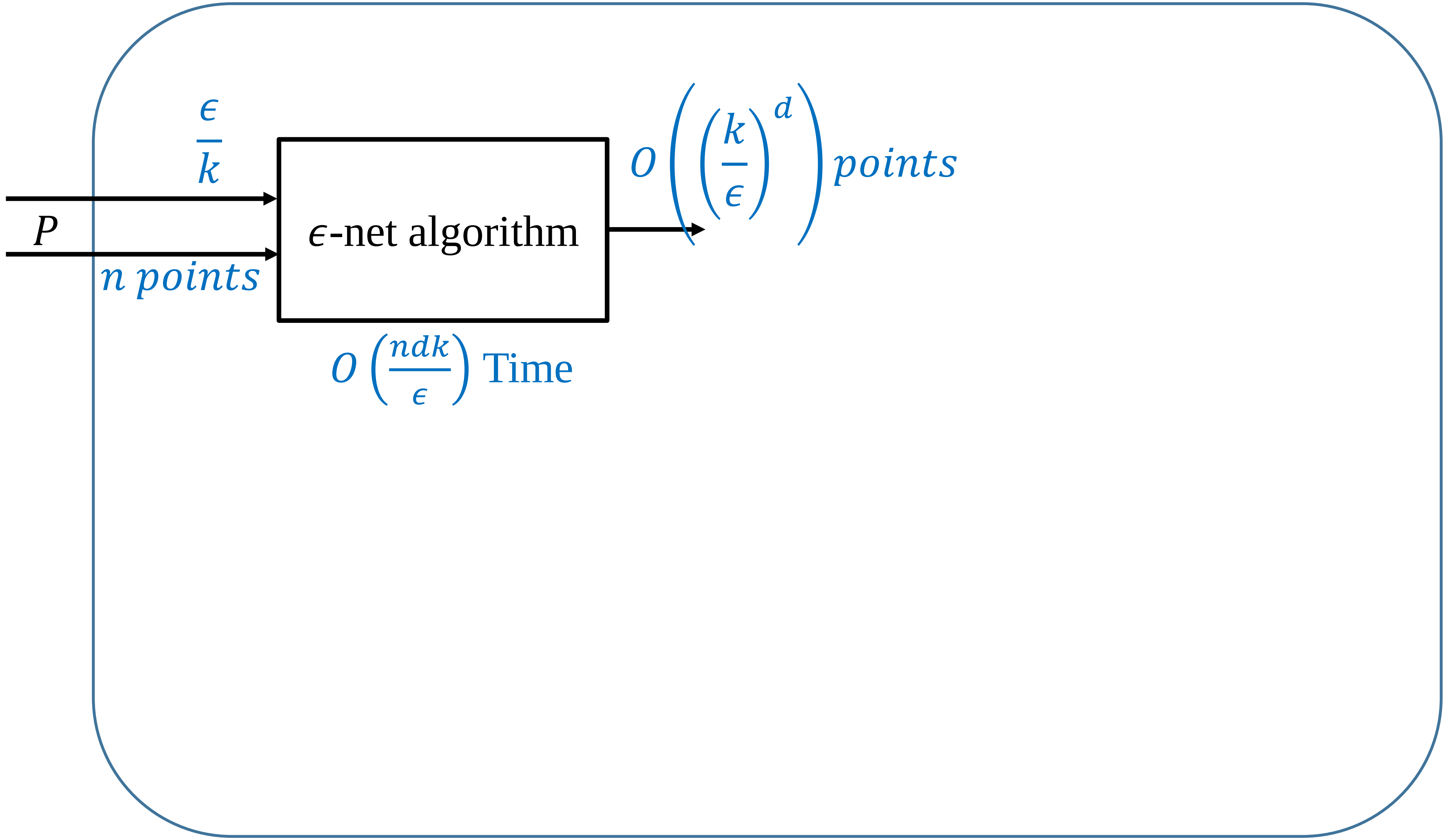
7) $t \leftarrow t + 1$

8) Repeat steps 3 to 6 till there are no more input points

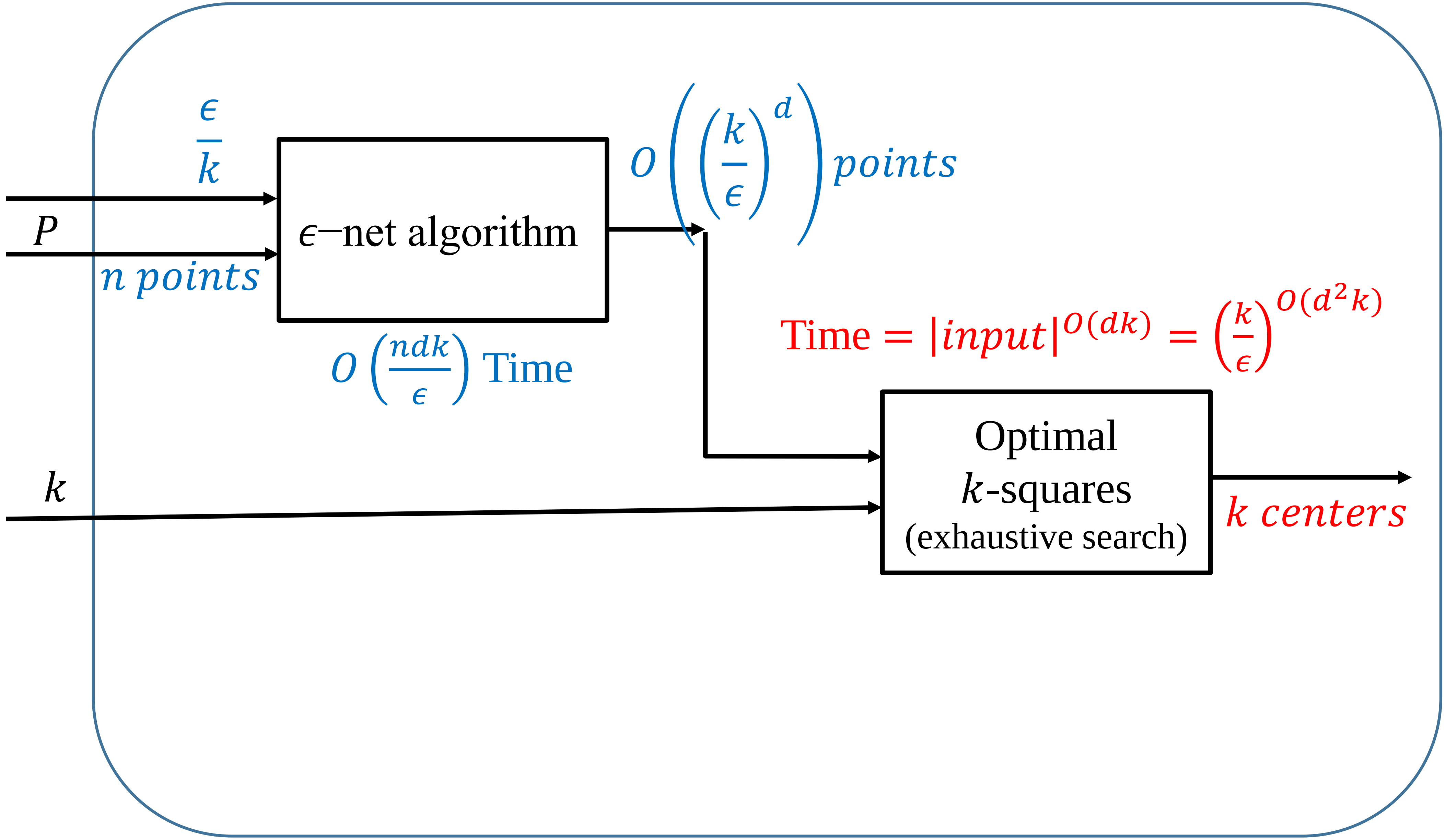
9) Return C



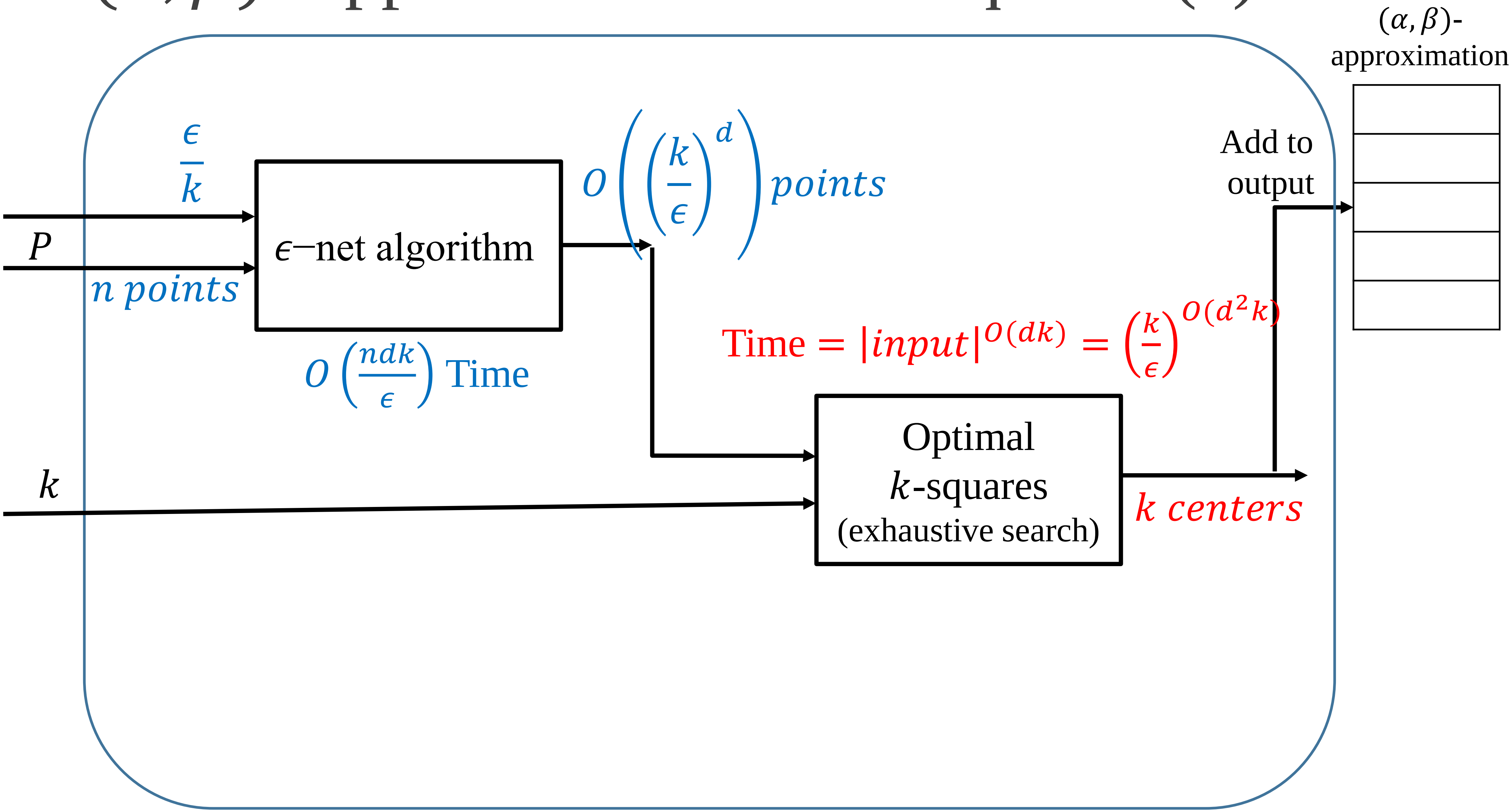
(α, β)-Approximation for k-Squares (1)



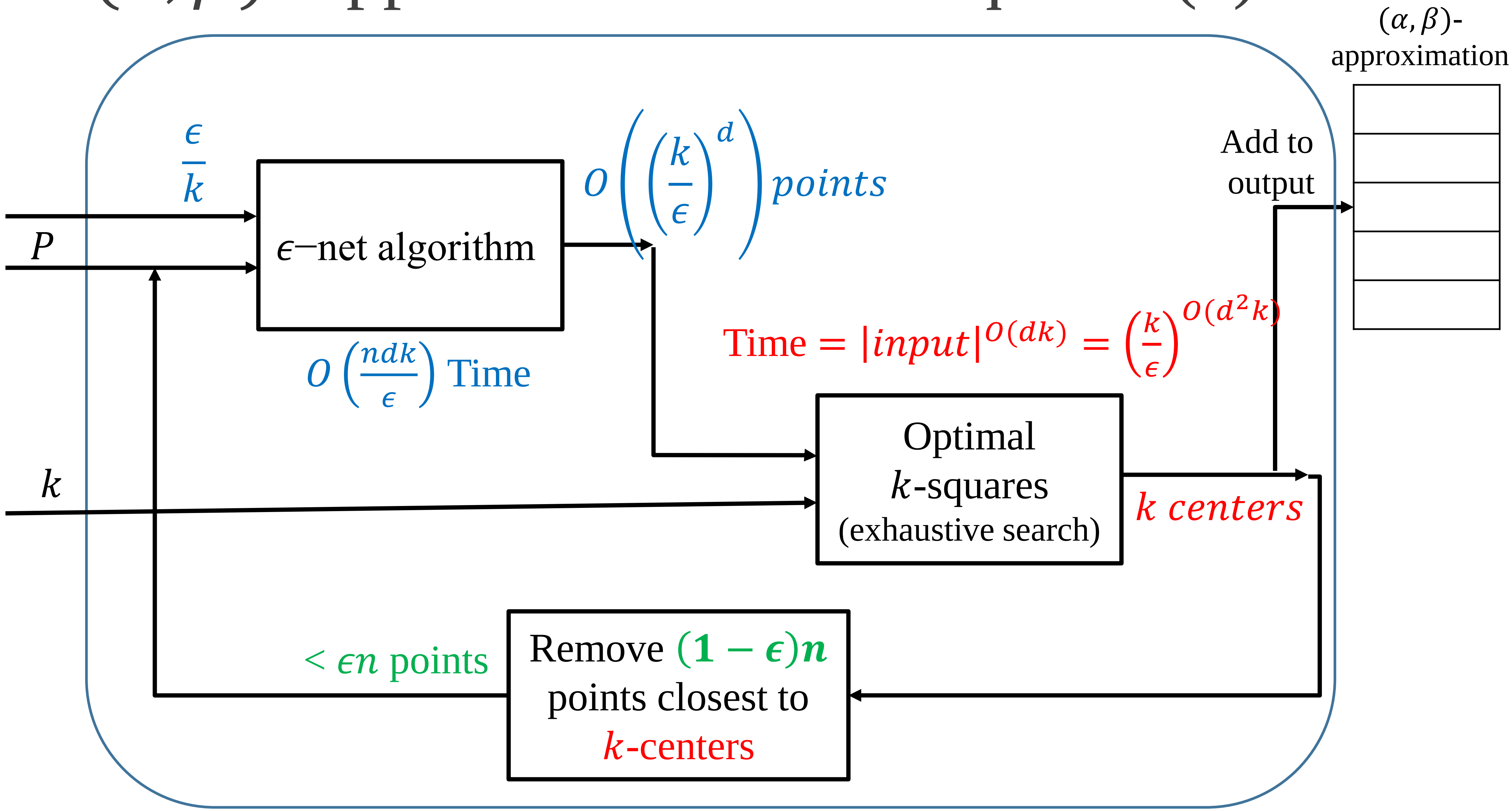
(α, β)-Approximation for k-Squares (1)



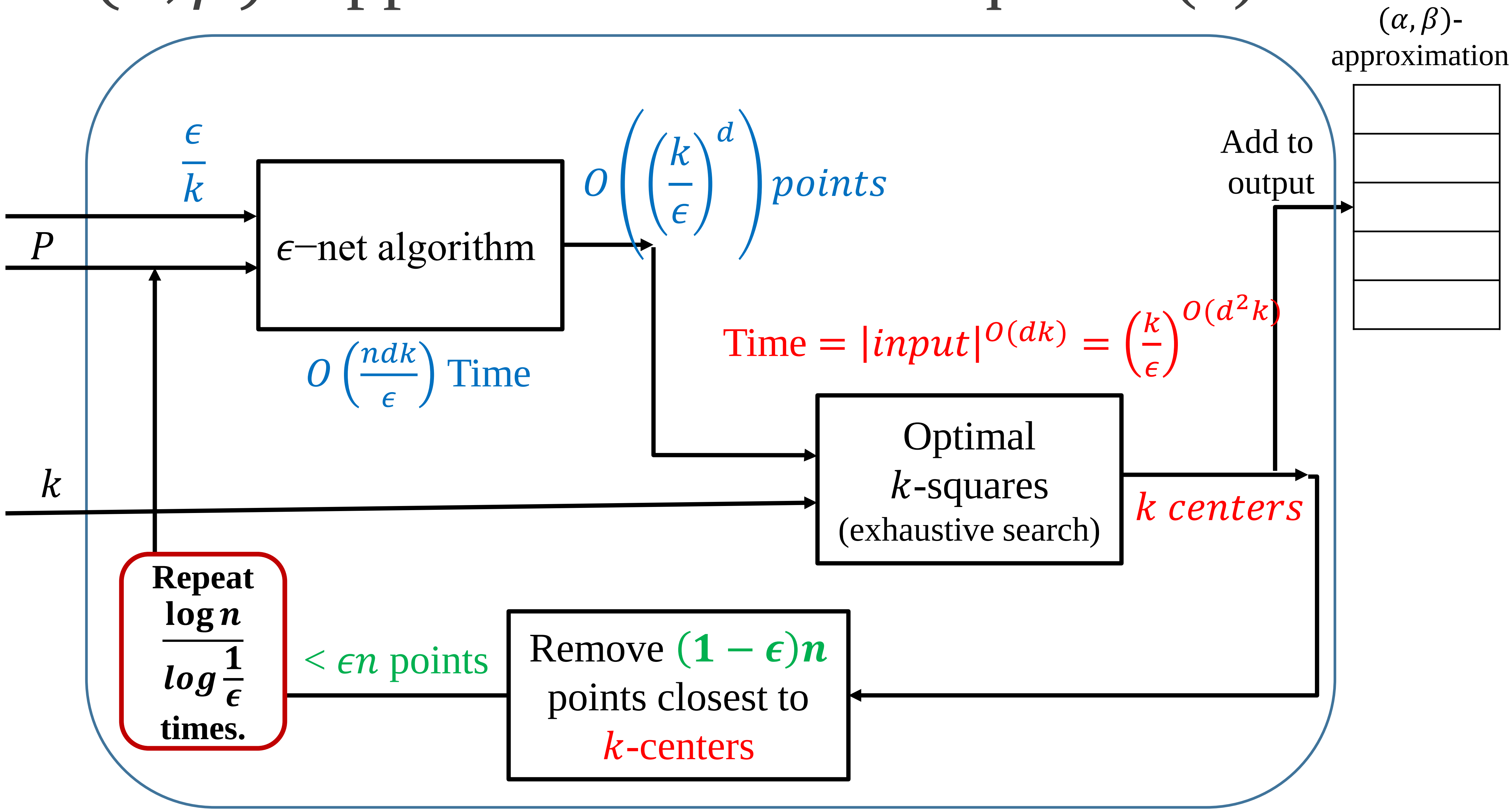
(α, β)-Approximation for k-Squares (1)



(α, β)-Approximation for k-Squares (1)



(α, β)-Approximation for k-Squares (1)



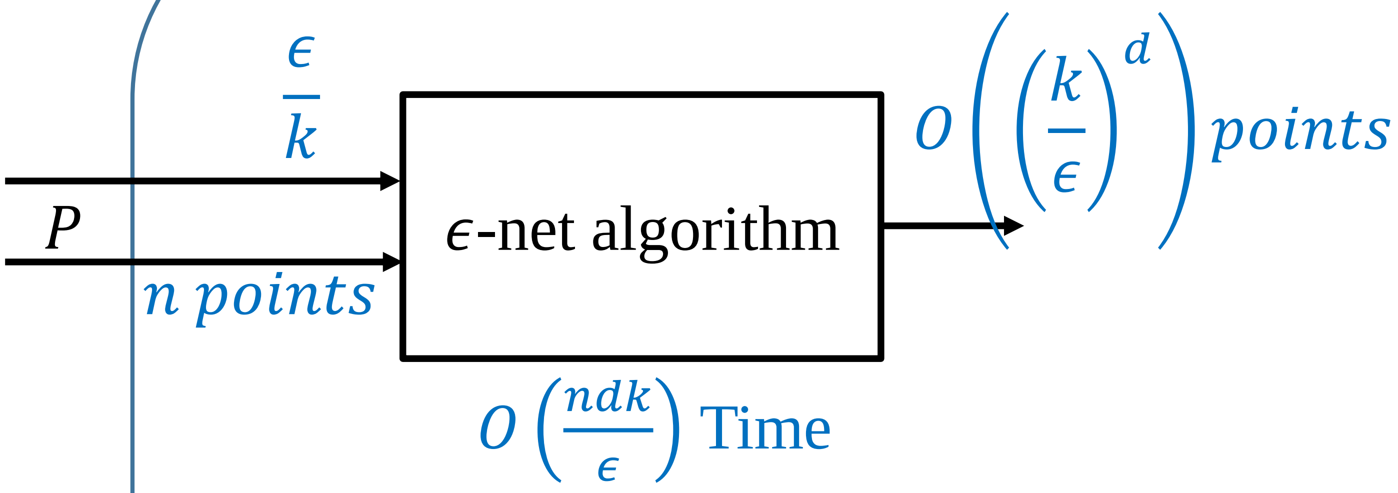
(α, β)-Approximation for k-Squares (1)

First approach:

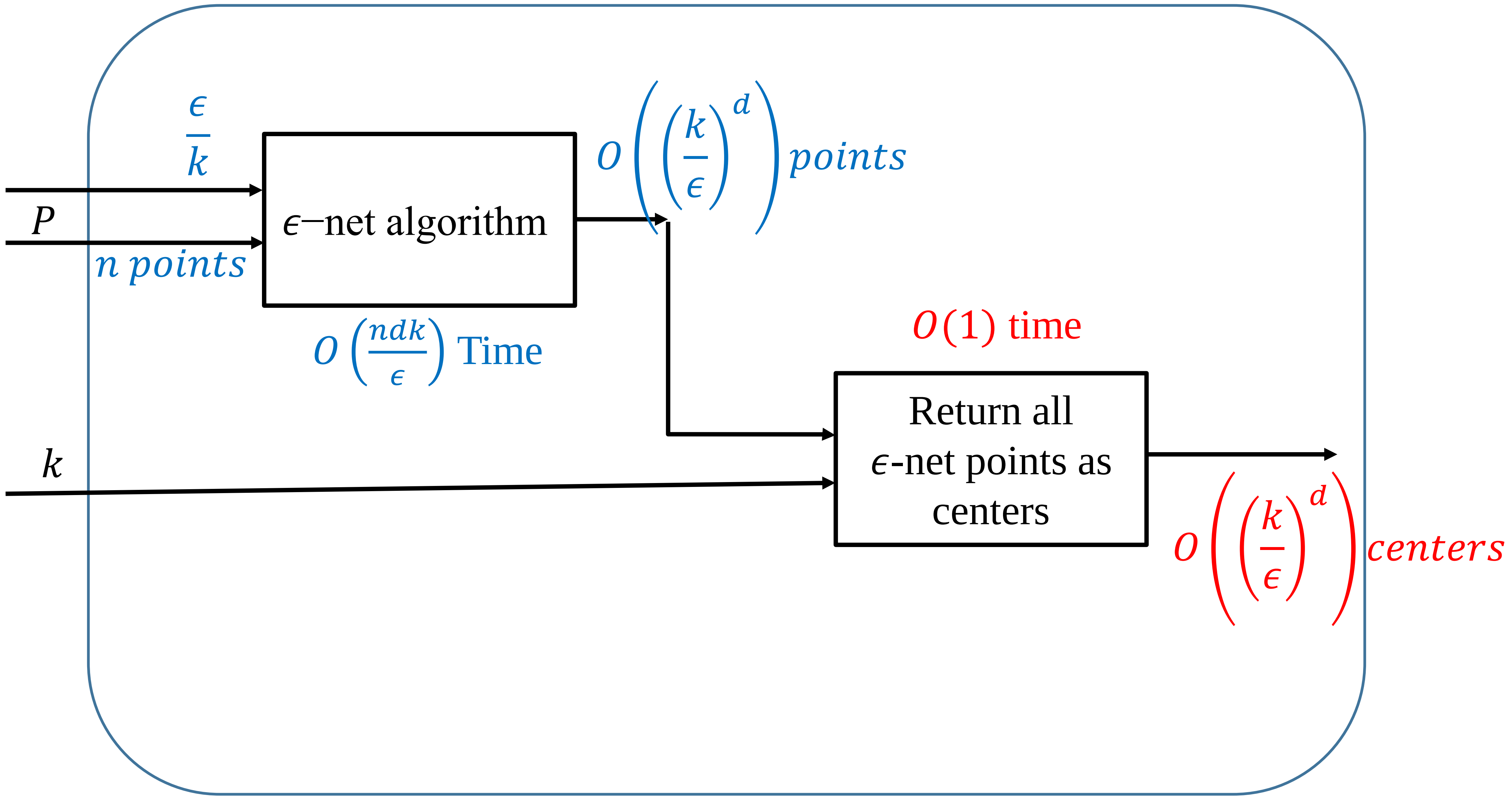
- $\beta = \frac{\log n}{\log \frac{1}{\epsilon}}$.
- $\alpha = 1$ since optimal squares were computed on a subset of the data.
- Time $< \frac{\log n}{\log \frac{1}{\epsilon}} \cdot \left(\frac{ndk}{\epsilon} + \left(\frac{k}{\epsilon}\right)^{O(d^2 k)} \right)$
- Time $< \frac{ndk}{\epsilon} + \frac{\frac{n}{2}dk}{\epsilon} + \cdots + \frac{dk}{\epsilon} + \frac{\log n}{\log \frac{1}{\epsilon}} \cdot \left(\frac{k}{\epsilon}\right)^{O(d^2 k)}$

$$= \frac{2ndk}{\epsilon} + \frac{\log n}{\log \frac{1}{\epsilon}} \cdot \left(\frac{k}{\epsilon}\right)^{O(d^2 k)}$$

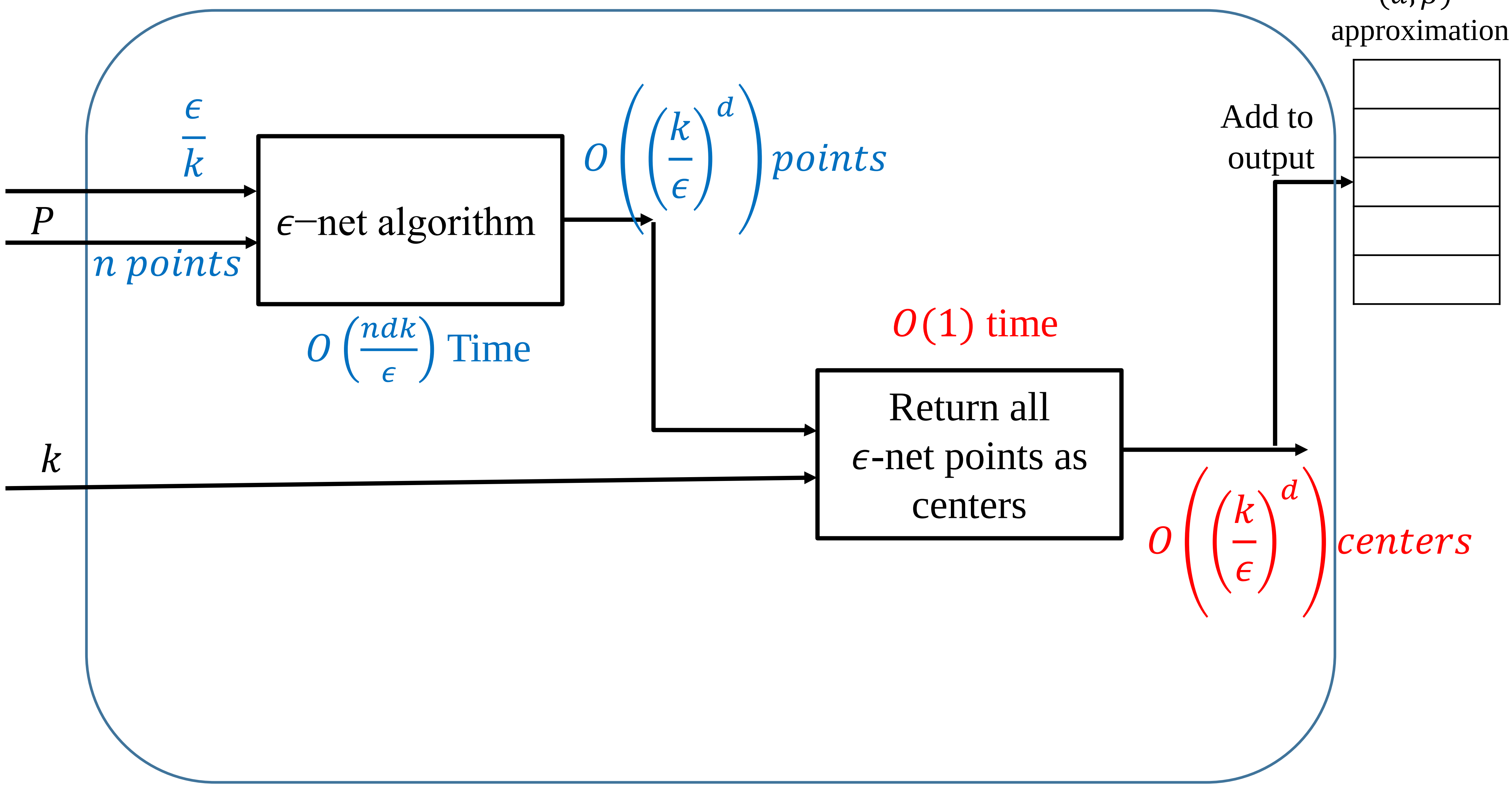
(α, β)-Approximation for k-Squares (2)



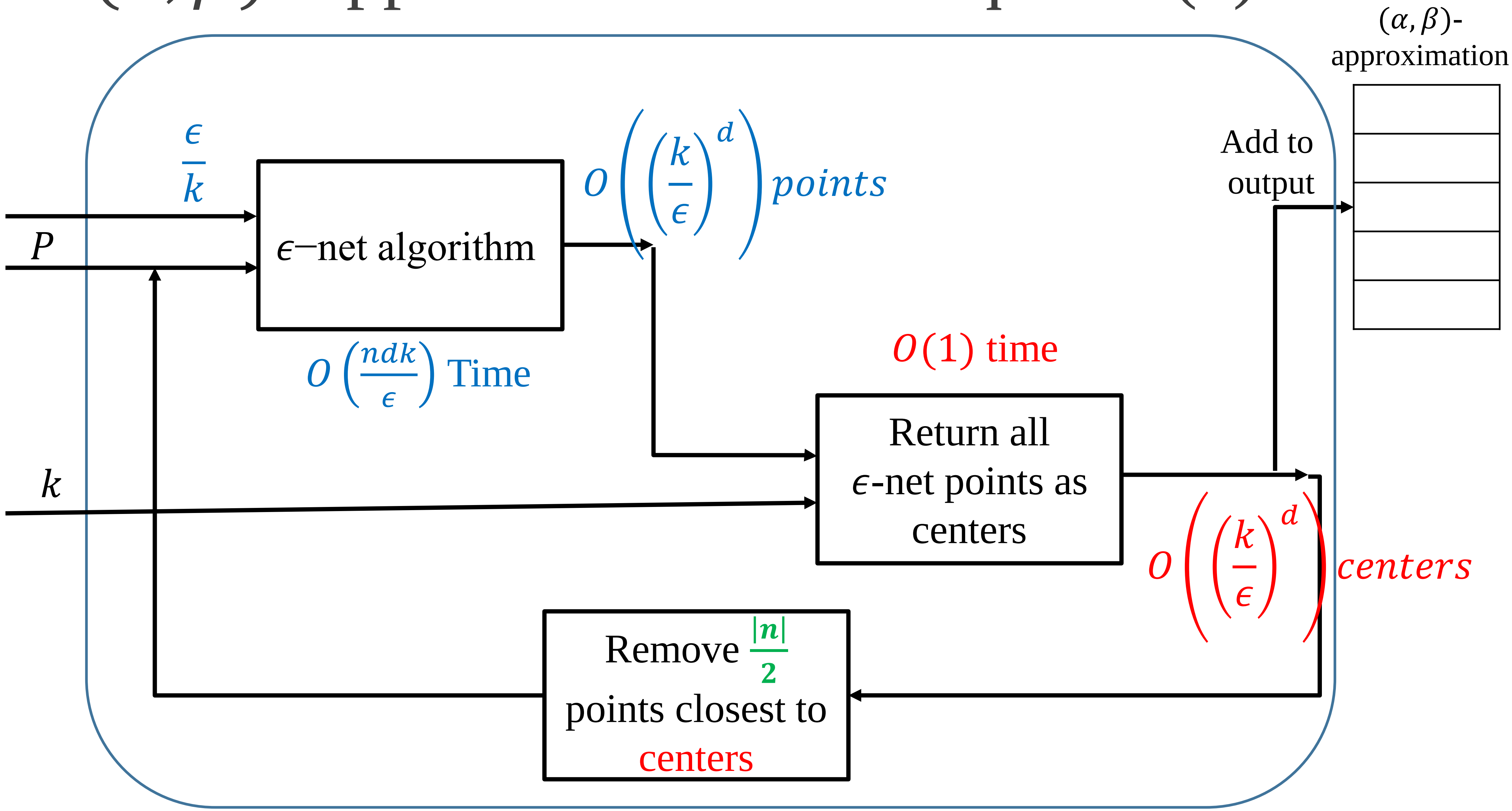
(α, β)-Approximation for k-Squares (2)



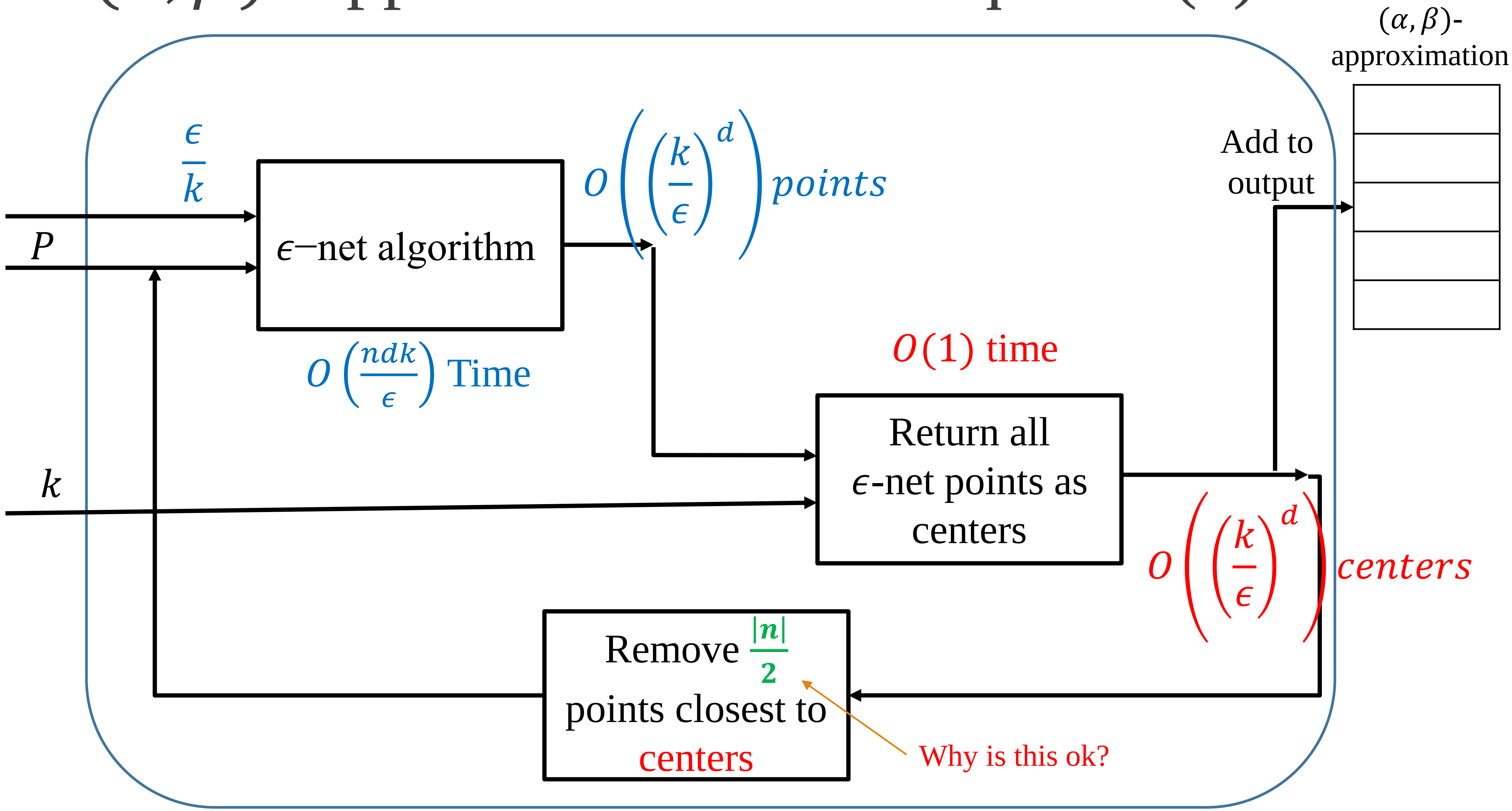
(α, β)-Approximation for k-Squares (2)



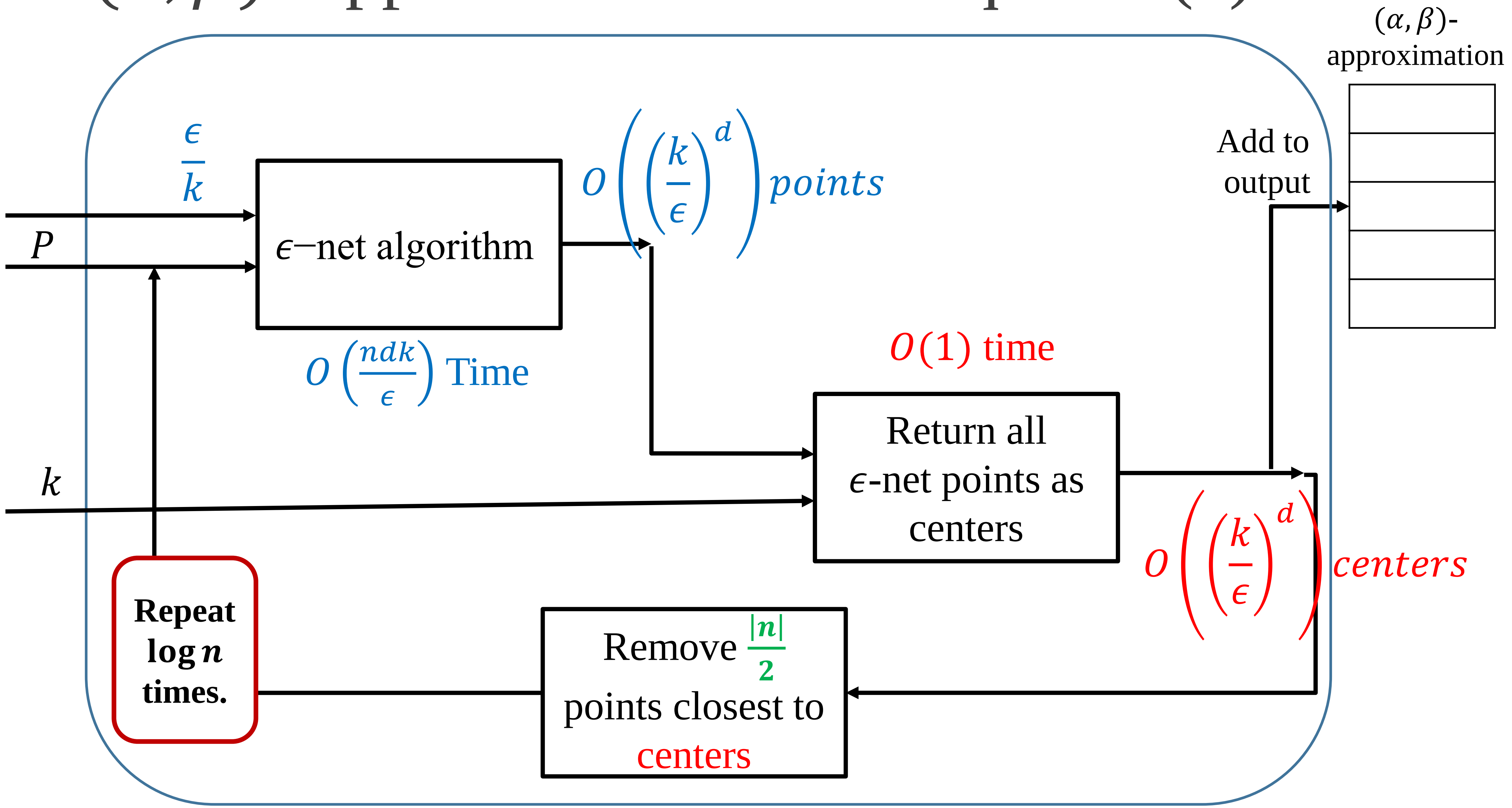
(α, β)-Approximation for k-Squares (2)



(α, β)-Approximation for k-Squares (2)



(α, β)-Approximation for k-Squares (2)



(α, β)-Approximation for k-Squares (2)

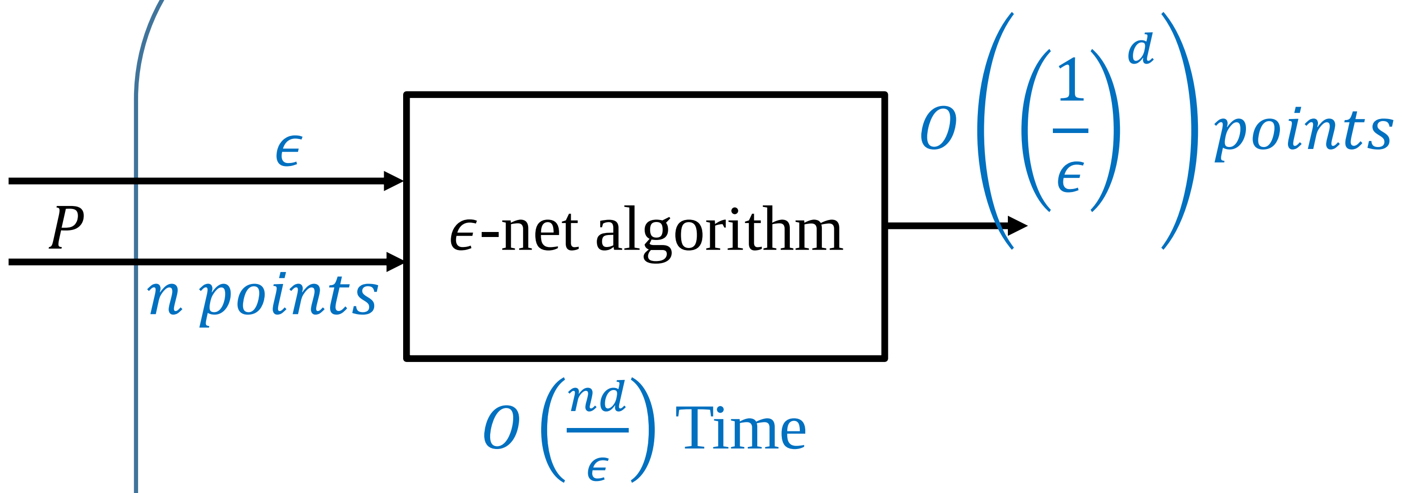
First approach:

- $\beta = O\left(\left(\frac{k}{\epsilon}\right)^d \log n\right)$.

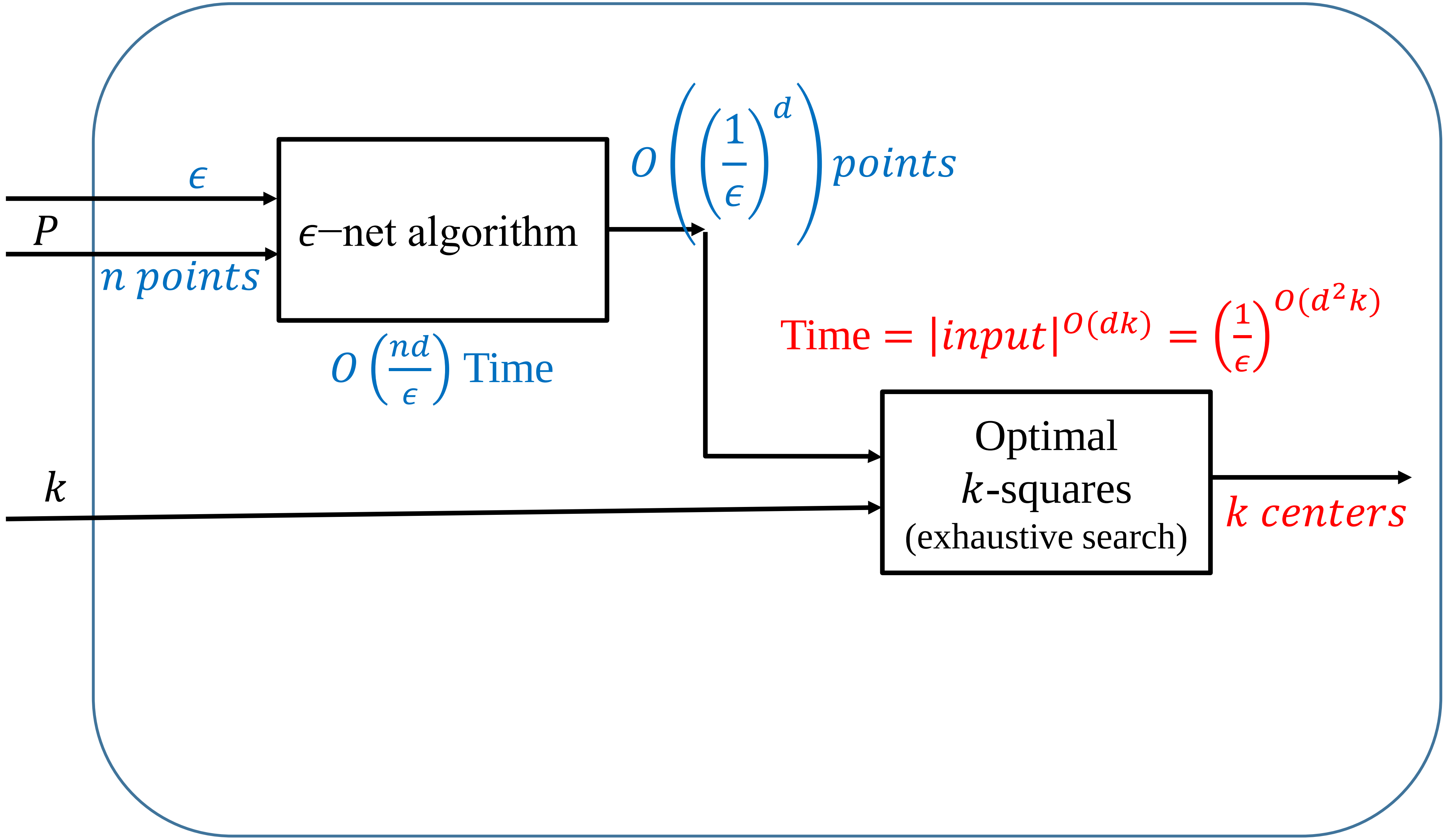
- $\alpha = 1$ since optimal squares were computed on a subset of the data.

- Time $< \frac{ndk}{\epsilon} + \frac{\frac{n}{2}dk}{\epsilon} + \cdots + \frac{dk}{\epsilon} = O\left(\frac{ndk}{\epsilon}\right)$

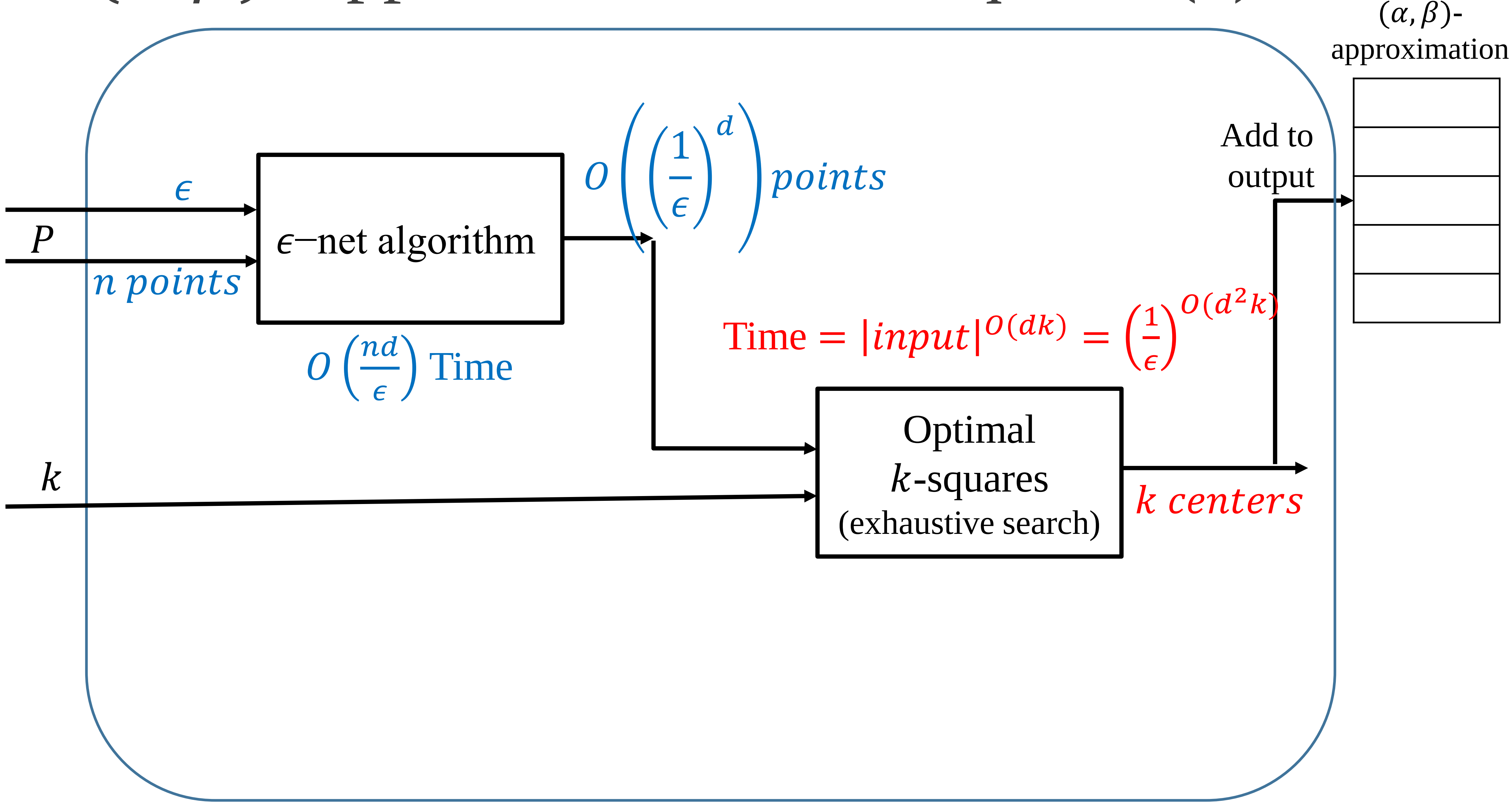
(α, β)-Approximation for k-Squares (3)



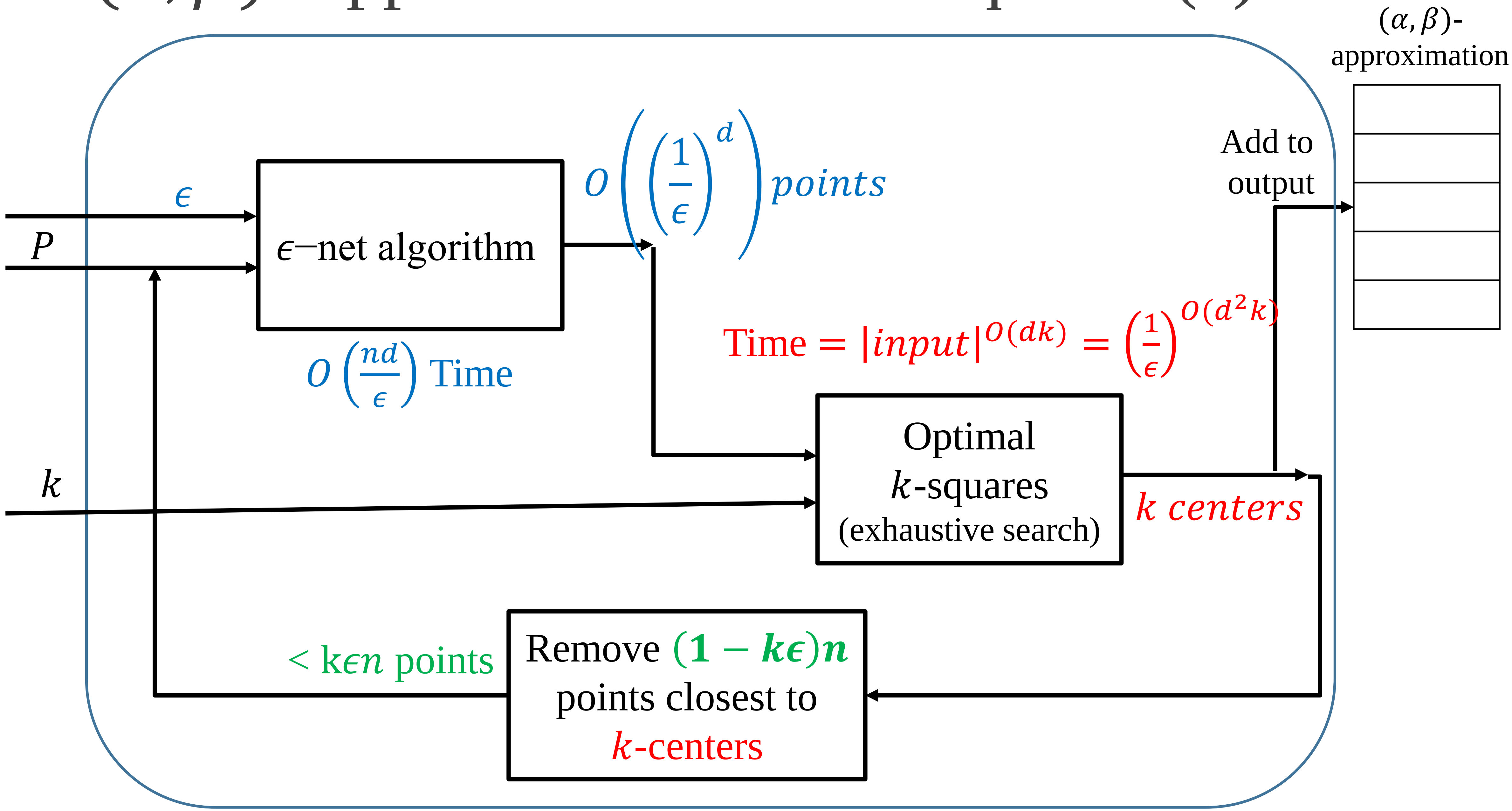
(α, β)-Approximation for k-Squares (3)



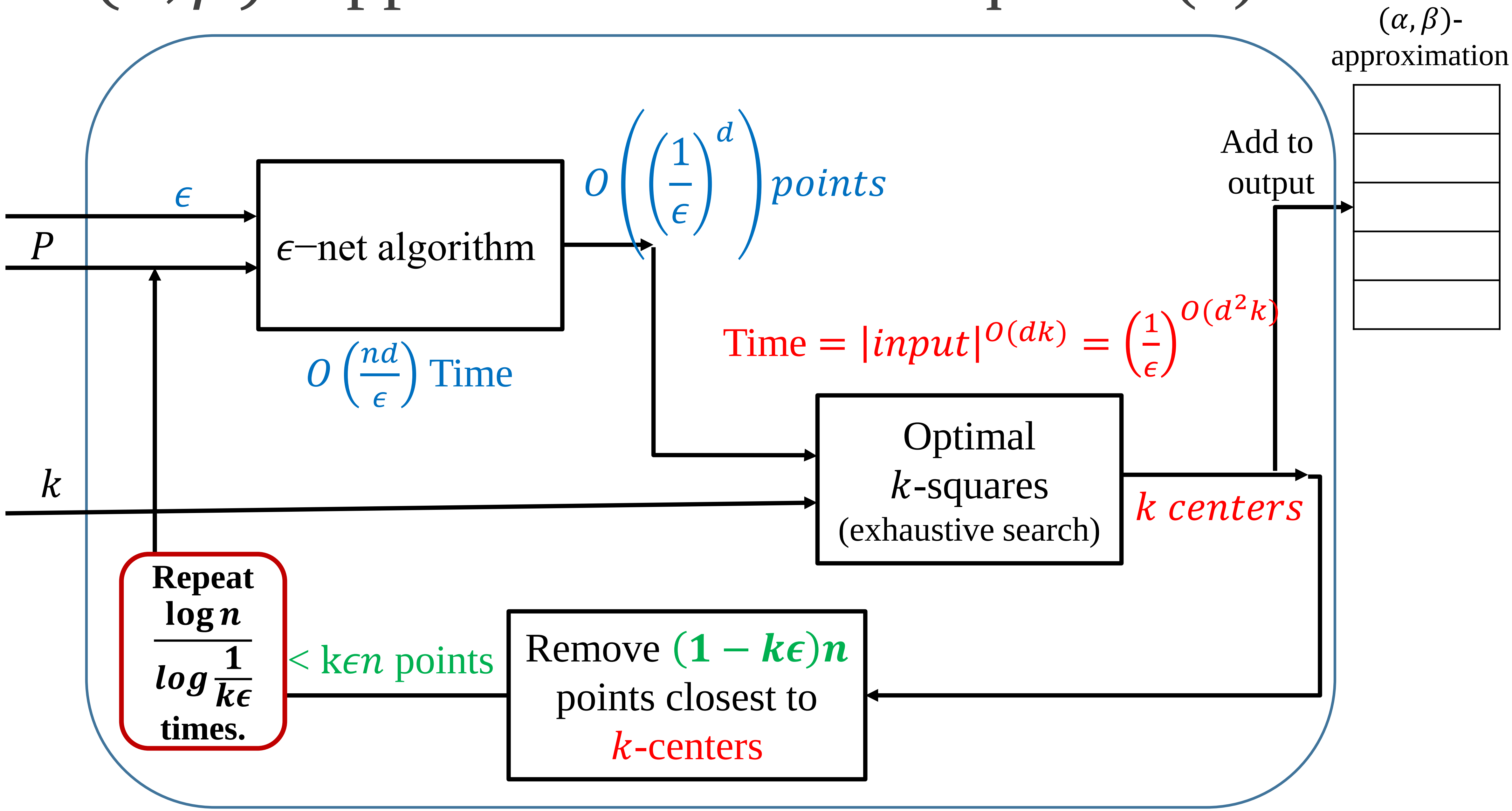
(α,β)-Approximation for k-Squares (3)



(α, β)-Approximation for k-Squares (3)



(α, β)-Approximation for k-Squares (3)



(α, β)-Approximation for k-Squares (3)

First approach:

- $\beta = \log n$.

- $\alpha = 1$ since optimal squares were computed on a subset of the data.

- Time $< \frac{\log n}{\log \frac{1}{k\epsilon}} \cdot \left(\frac{nd}{\epsilon} + \left(\frac{1}{\epsilon}\right)^{O(d^2 k)} \right)$