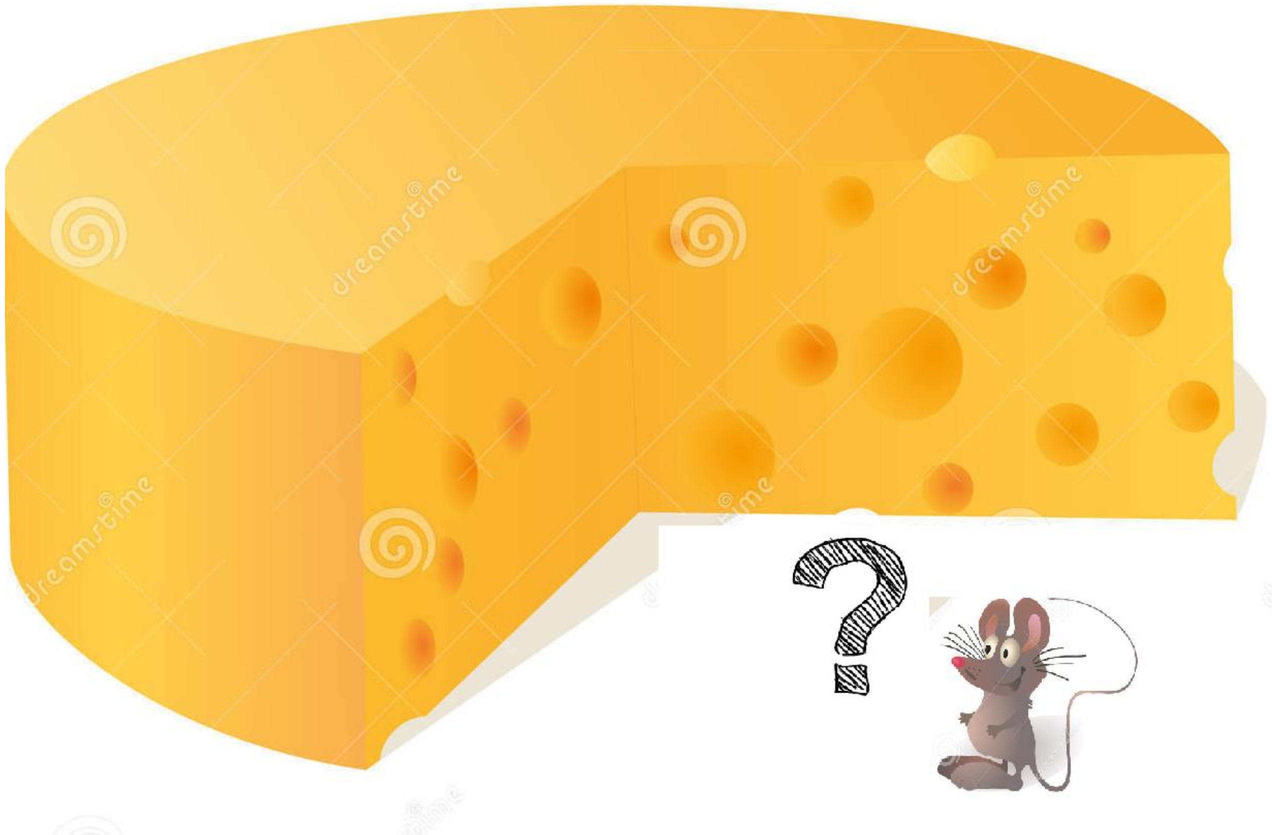
Big Data Class



LECTURER: DAN FELDMAN

TEACHING ASSISTANTS:

IBRAHIM JUBRAN

SOLIMAN NASSER

Weighted Input

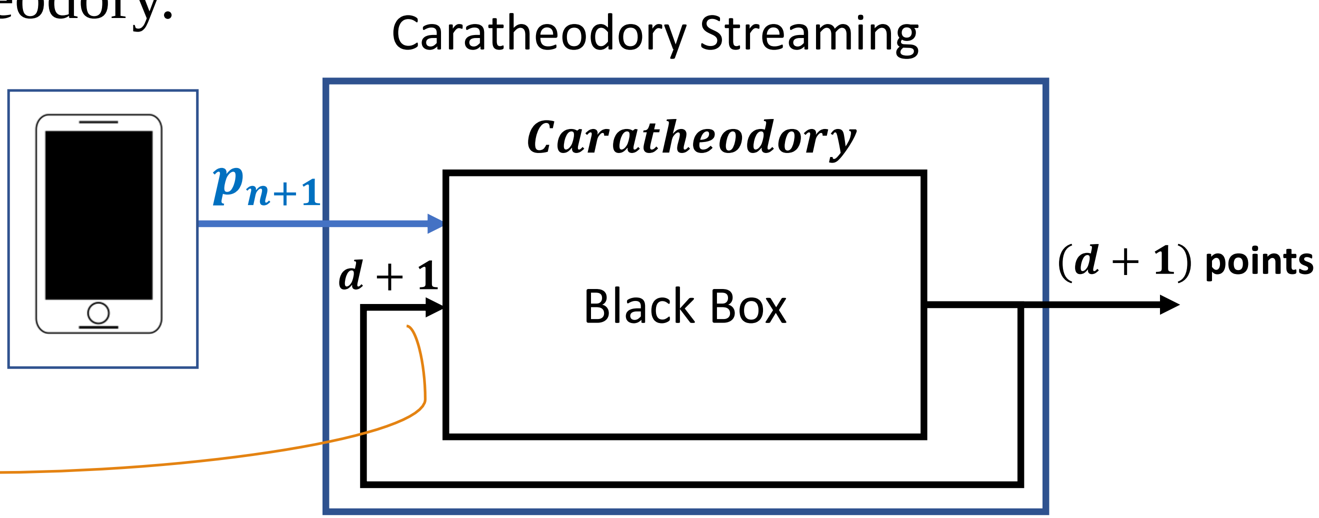
Why / When ?

- Coreset for coreset
  - Streaming model: we want to update the coreset every time a new data point arrives. Thus we have a weighted input (the old coreset) and we want to compute a new coreset.

Weighted Input

Why / When ?

- Coreset for coreset
  - Streaming model:
    Reminder: **Streaming** Caratheodory.



Weighted input (C, ω) from prev. iteration s.t.

$$\sum_{p\in P} \|p-x\|^2 = \sum_{c\in C} \boldsymbol{\omega}(\boldsymbol{c}) \|c-x\|^2$$

Weighted Input

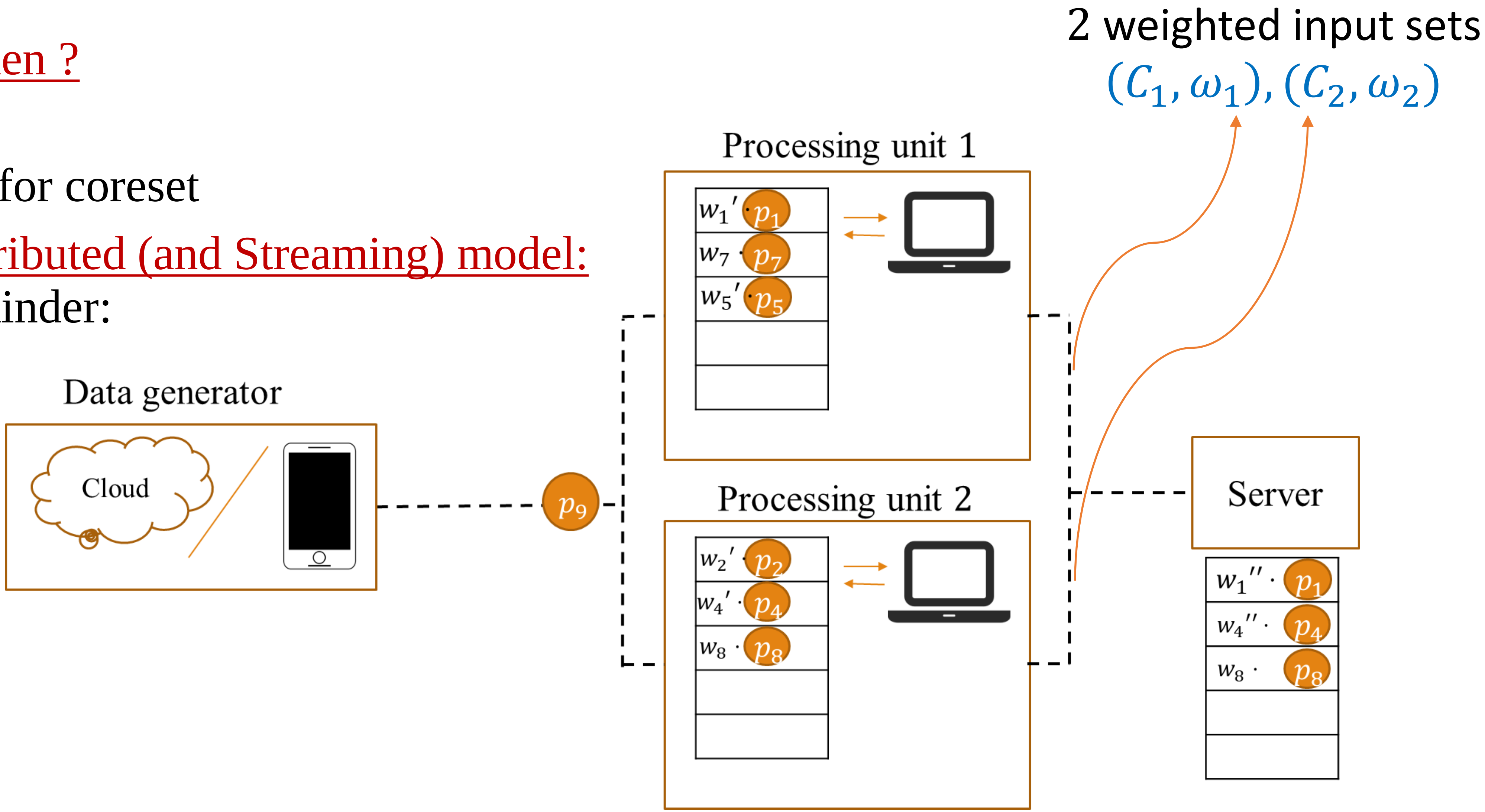
Why / When ?

- Coreset for coreset
  - Distributed (and streaming) model: the output of each (distributed) machine is a coreset. Thus to compute the final coreset we have multiple weighted sets of input.

Weighted Input

Why / When ?

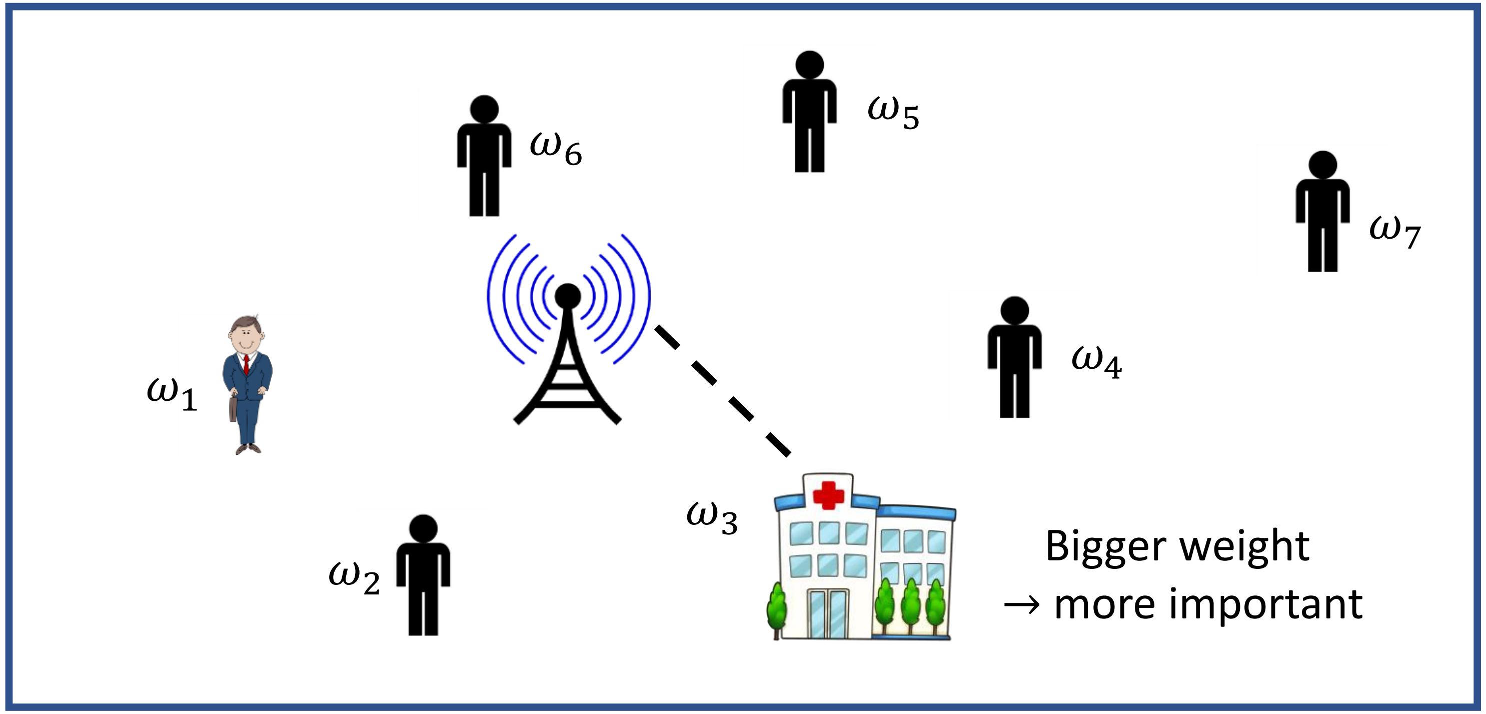
- Coreset for coreset
  - Distributed (and Streaming) model:
    Reminder:



Weighted Input

<u>Why / When ?</u>

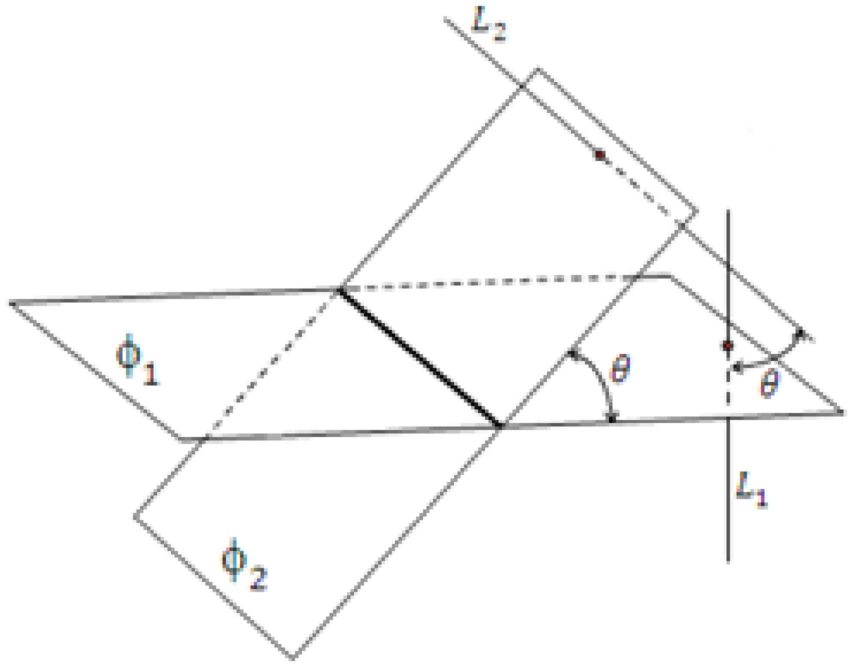
- Importance sampling (sensitivity)
  - Unlike uninform random sampling, where the probability is equal for all data points, we want to be able to give different probabilities (importance).

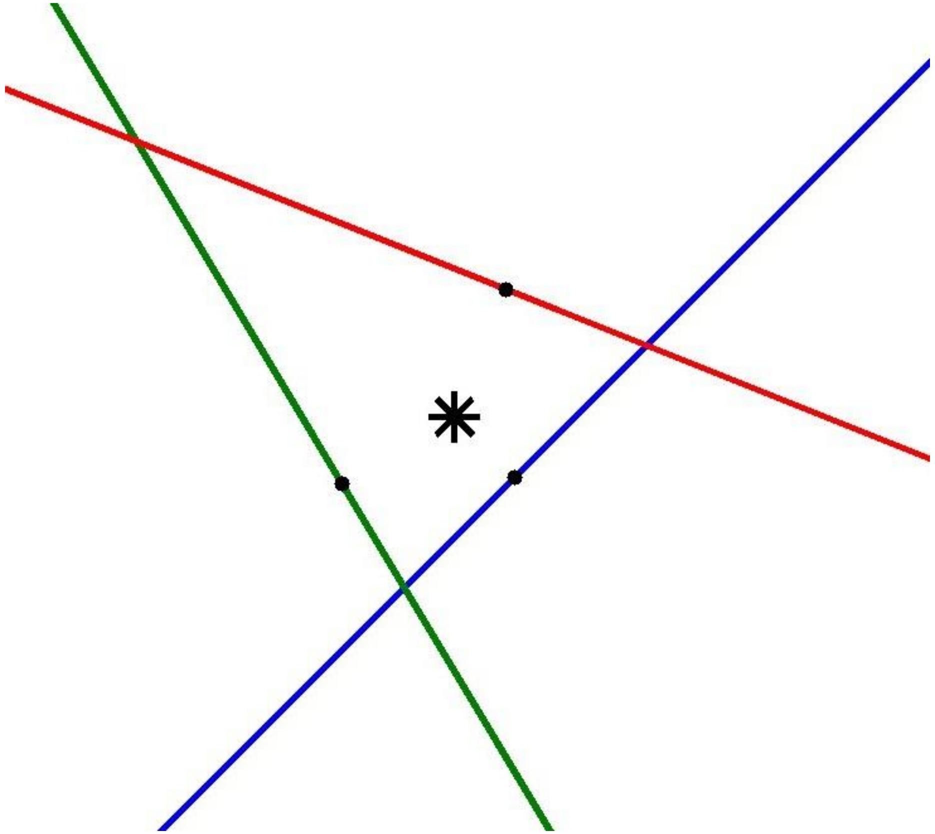


Weighted Input

Why / When ?

- Non-points data input (e.g. lines/planes)
    - Example: 1-mean/center for lines/planes

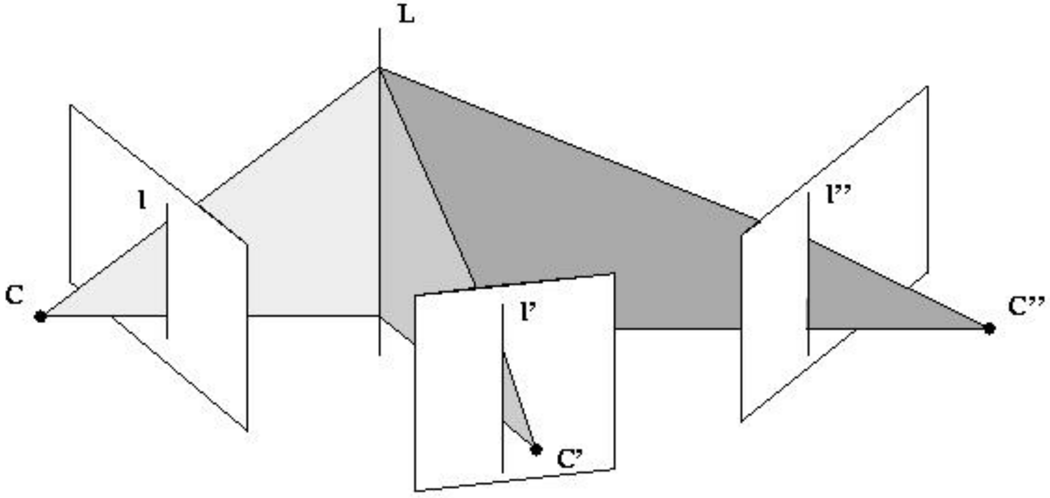


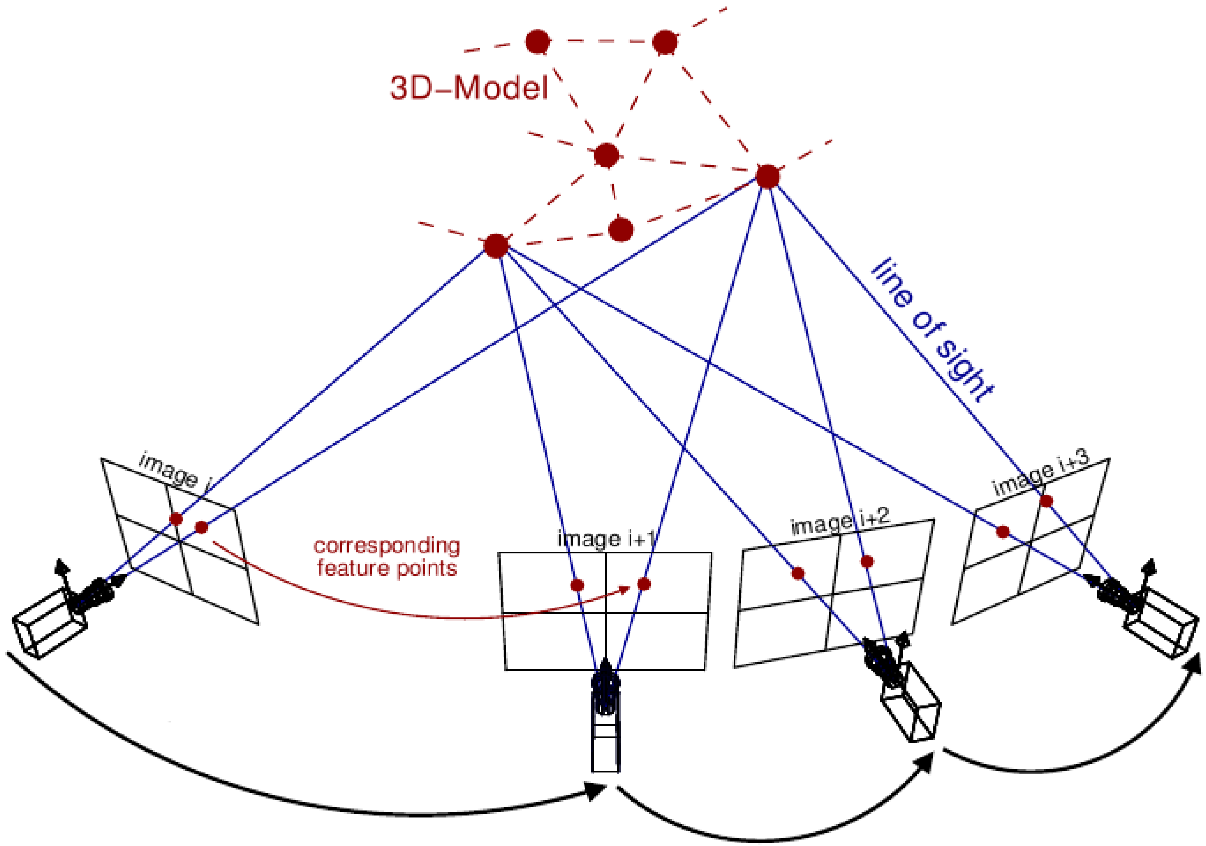


Weighted Input

Why / When ?

- Non-points data input (e.g. lines)
    - Motivation: **Computer Vision**

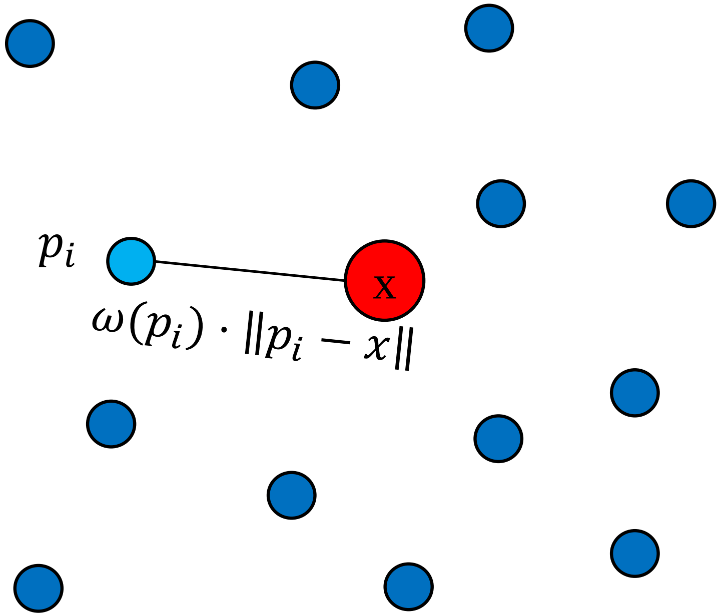




1-Center for Weighted Input

Given (P, ω) where $P \subseteq R^d$ and $\omega : P \to R$ such that $\sum_{p \in P} \omega(p) = 1$, find the point $x \in R^d$ that minimizes:

$$\boldsymbol{far(P, \omega, x) = \max_{p \in P} \omega(p) \|p - x\|}$$

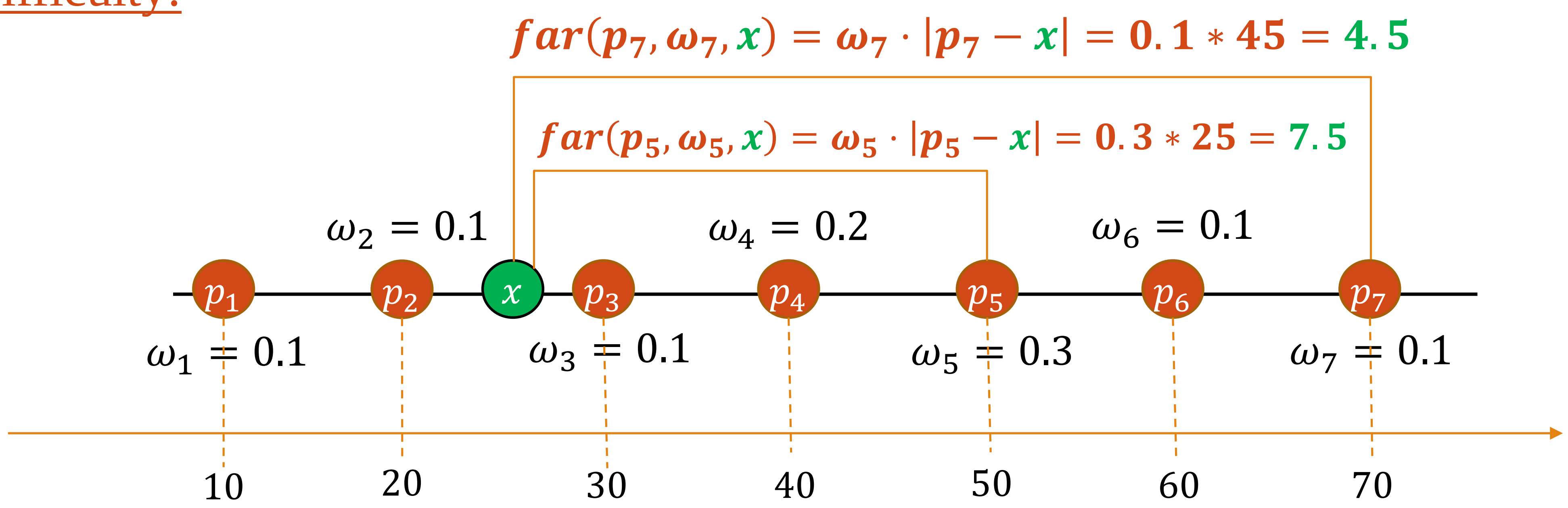


1-Center Queries for Weighted Input

- Input: (P, ω, X, far) where $P \subseteq R^d, \omega: P \to R$ and $\sum \omega(p) = 1, X \subseteq R^d$,
  $far(P, \omega, x) = \max_{p \in P} \omega(p) \cdot \|p - x\|$

Farthest point from x is not necessarily one of the edge points!

- Difficulty:



1-Center for Weighted Input

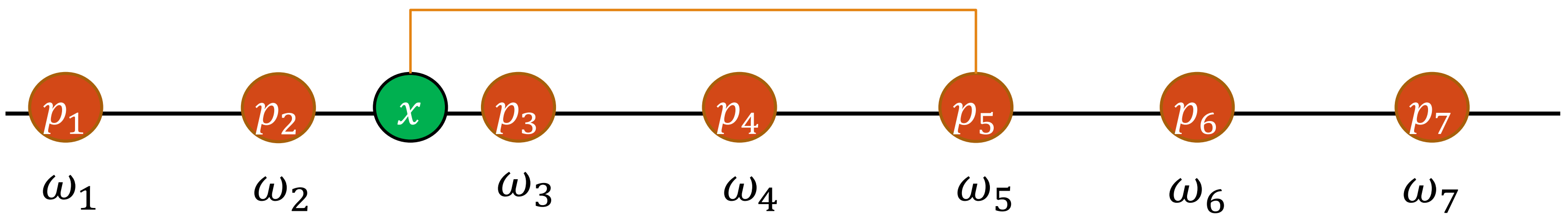
Weighted 1-Center

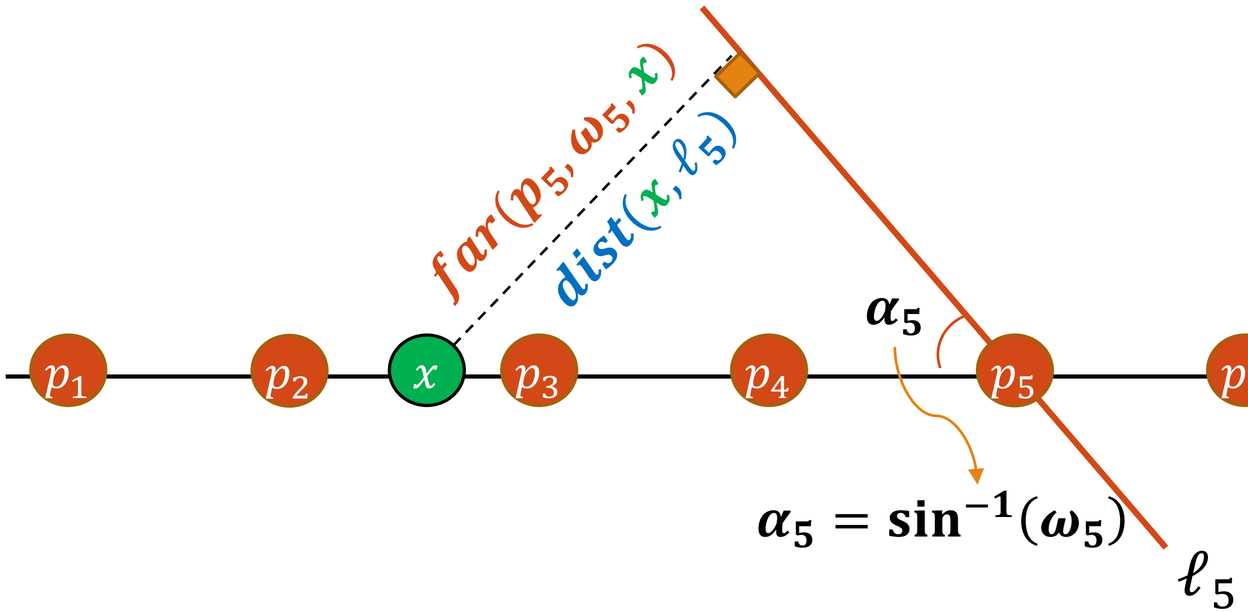
$\Updownarrow$

Distance to lines in R^2

- Connection to lines in $d = 2$:

$$far(p_5, \omega_5, x) = \omega_5 \cdot |p_5 - x|$$



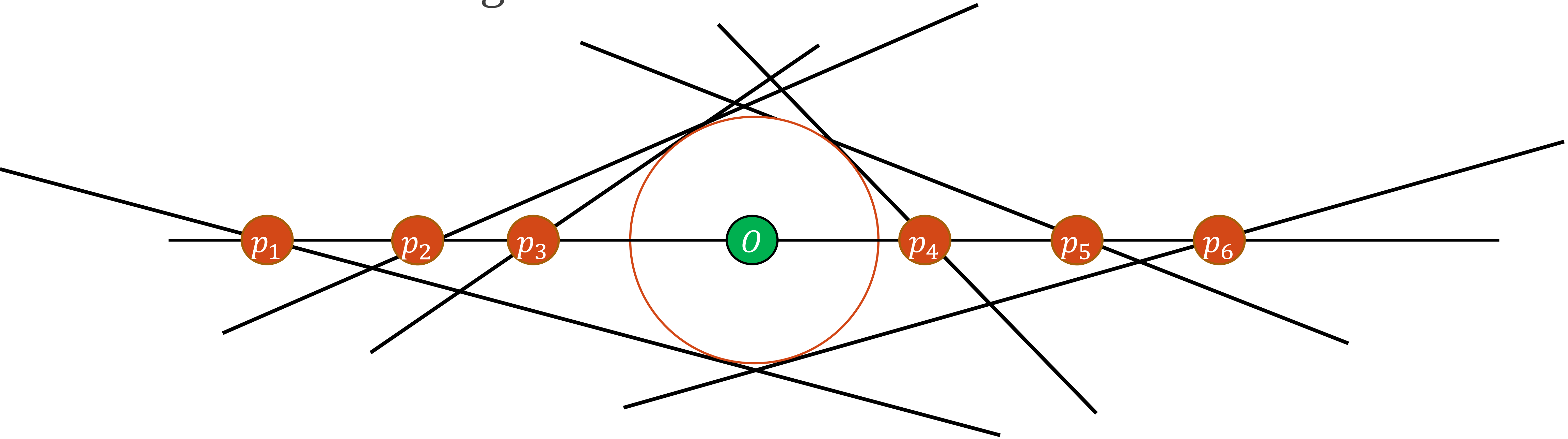


$$\begin{aligned} dist(x, \ell_5) &= \sin \alpha_5 \cdot dist(p_5, x) \\ &= \sin(\sin^{-1} \omega_5) \cdot |p_5 - x| \\ &= \omega_5 \cdot |p_5 - x| \\ &= far(p_5, \omega_5, x) \end{aligned}$$

1-Center for Weighted Input

- <u>Observation:</u>

All points might have the same weighted distance $\omega(p_i)\|p_i\| = 1$ to the origin:
All lines are tangent to the unit circle.



1-Center for Weighted Input

- <u>Claim:</u>

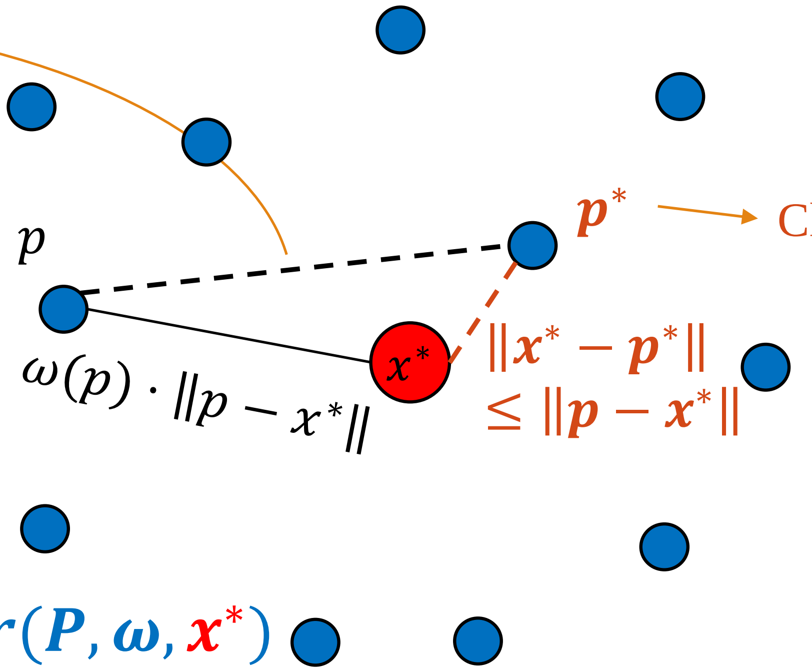
There is an input point $p^* \in P$ which is a factor 2 approximation to the optimal 1-center x^* of the weighted set (P, ω):

$$\boldsymbol{far(P, \omega, p^*) \leq 2 \cdot far(P, \omega, x^*)}$$

$\boldsymbol{far(P, \omega, x^*) \leq far(P, \omega, x)}$
for every $\boldsymbol{x \in Q}$

Triangle inequality

$$\begin{aligned} &\|\boldsymbol{p} - \boldsymbol{p}^*\| \\ &\leq \|\boldsymbol{p} - \boldsymbol{x}^*\| + \|\boldsymbol{x}^* - \boldsymbol{p}^*\| \\ &\leq 2 \cdot \|\boldsymbol{p} - \boldsymbol{x}^*\| \end{aligned}$$



Closest point to x^*

$$\rightarrow \boldsymbol{\omega(p) \cdot \|p - p^*\| \leq 2 \cdot \omega(p) \cdot \|p - x^*\|}$$

for every $\boldsymbol{p \in P}$

$$\rightarrow \boldsymbol{far(P, \omega, p^*) \leq 2 \cdot far(P, \omega, x^*)}$$

Coreset for 1-Center with Equally Weighted Input

- Observation:

If all the data points have the same weight, i.e. for ever $p \in P, \omega(p) = \Delta$, then a coreset for 1-center with non-weighted input (P) is also a coreset for 1-center with weighted input (P, ω).

- Proof:

Let C be a coreset for the non-weighted data P. Then for every q in the query space Q:

$$|far(P,q) - far(C,q)| = far(P,q) - far(C,q) \leq O(\epsilon) \cdot far(P,q)$$

Therefore, it also holds that:

$$\Delta \cdot far(P,q) - \Delta \cdot far(C,q) \leq \Delta \cdot O(\epsilon) \cdot far(P,q)$$

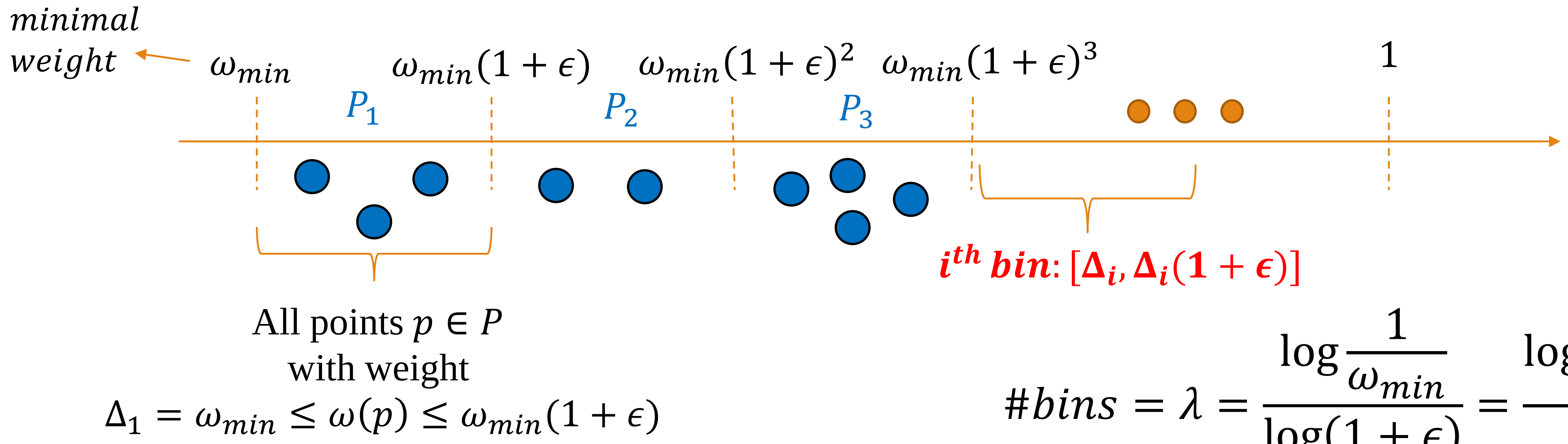
$$\rightarrow \boldsymbol{far(P,\Delta,q) - far(C,\Delta,q) \leq O(\epsilon) \cdot far(P,\Delta,q)}$$

Coreset for 1-Center with Weighted Input

- Input: (P, ω, X, far) where $P \subseteq R^d, \omega: P \to R$ and $\sum \omega(p) = 1$,

  $X \subseteq R^d$, $far(P, \omega, x) = \max_{p \in P} \omega(p) \cdot \|p - x\|$

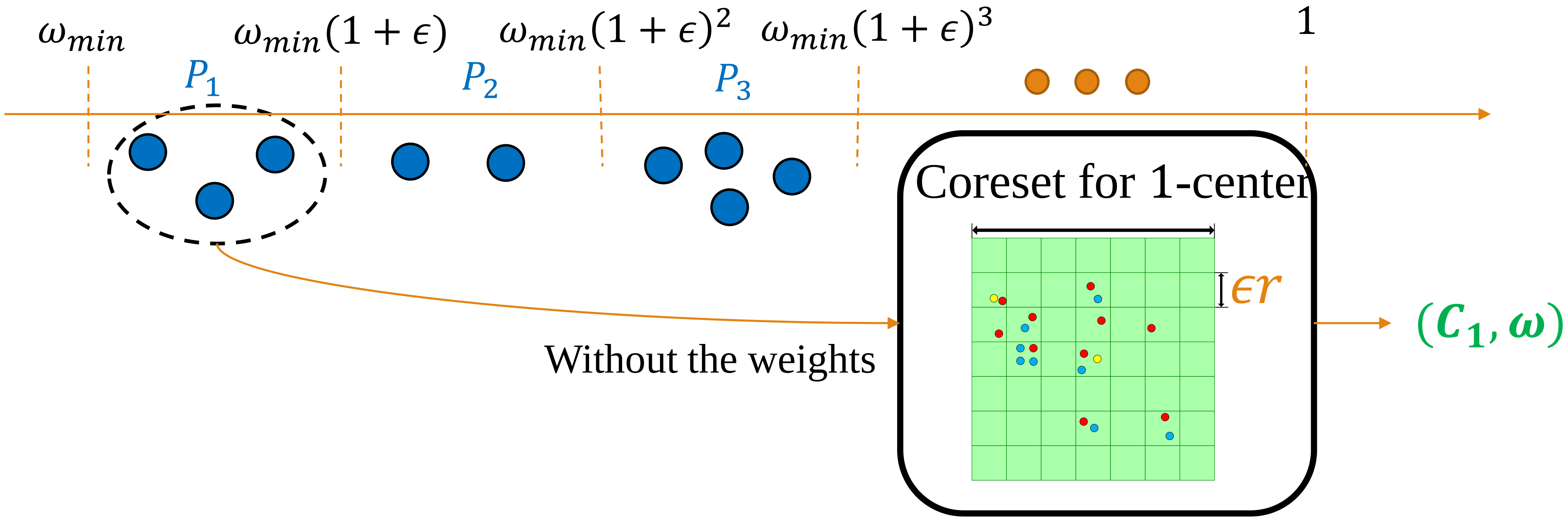
- Output: $C \subseteq P$ $s.t.$ $|far(P, \omega, x) - far(C, \omega, x)| \leq O(\epsilon) \cdot far(P, \omega, x)$



$$\#bins = \lambda = \frac{\log \frac{1}{\omega_{min}}}{\log(1+\epsilon)} = \frac{\log \frac{1}{\omega_{min}}}{\epsilon}$$

Coreset for 1-Center with Weighted Input

- Input: (P, ω, X, far) where $P \subseteq R^d, \omega: P \to R$ and $\sum \omega(p) = 1$,
  $X \subseteq R^d$, $far(P, \omega, x) = \max_{p \in P} \omega(p) \cdot \|p - x\|$
- Output: $\boldsymbol{C} \subseteq P\ s.t. |far(P, \omega, x) - far(\boldsymbol{C}, \omega, x)| \leq O(\epsilon) \cdot far(P, \omega, x)$

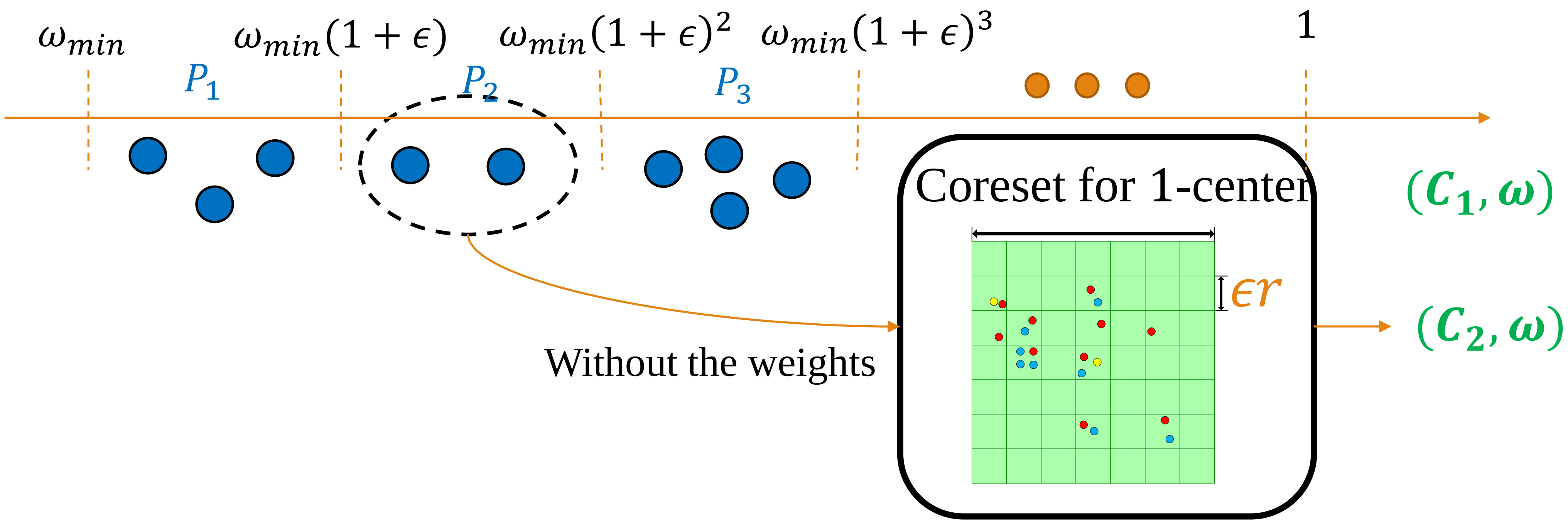


Coreset for 1-Center with Weighted Input

- Input: (P, ω, X, far) where $P \subseteq R^d, \omega: P \to R$ and $\sum \omega(p) = 1$,

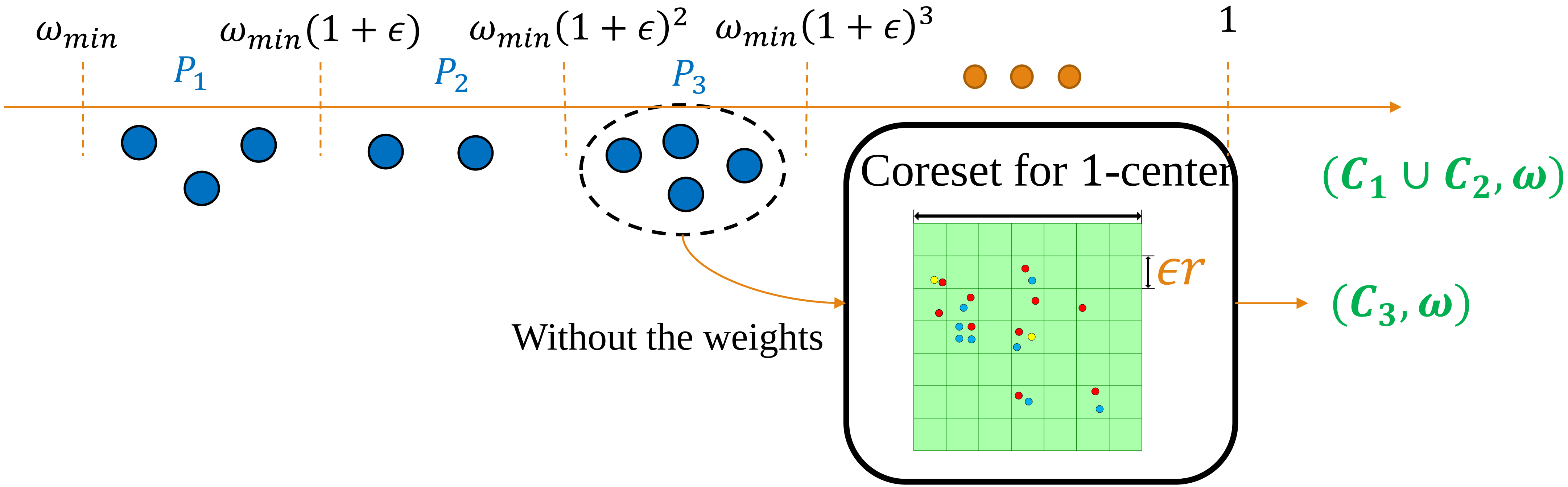
  $X \subseteq R^d$, $far(P, \omega, x) = \max_{p \in P} \omega(p) \cdot \|p - x\|$

- Output: $C \subseteq P\ s.t. |far(P, \omega, x) - far(C, \omega, x)| \leq O(\epsilon) \cdot far(P, \omega, x)$



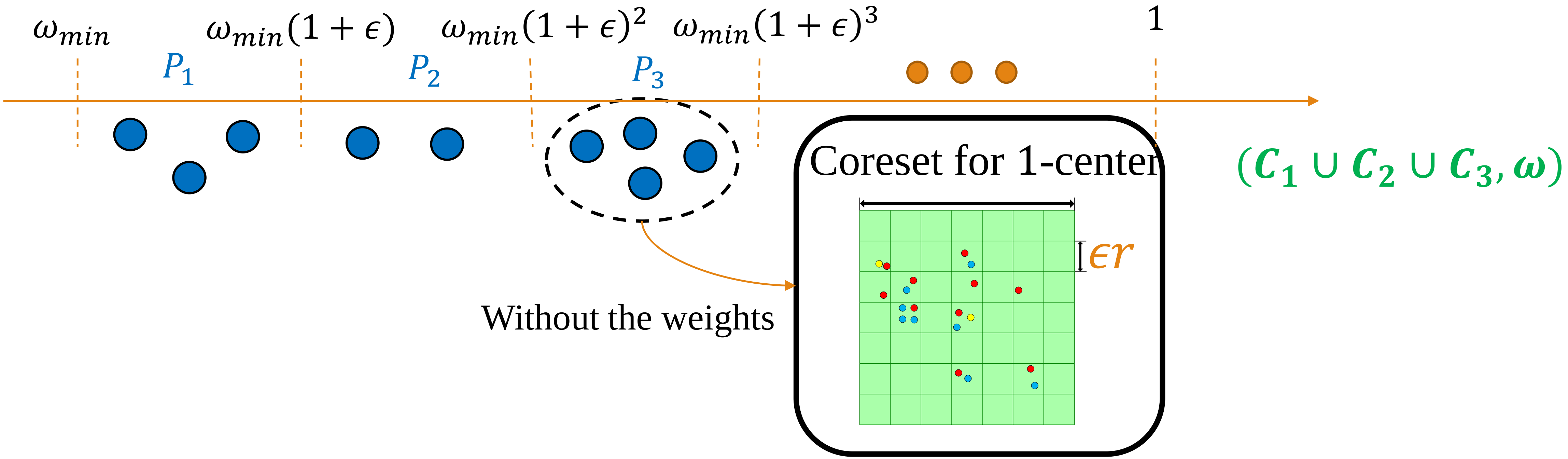
Coreset for 1-Center with Weighted Input

- Input: (P, ω, X, far) where $P \subseteq R^d, \omega: P \to R$ and $\sum \omega(p) = 1$,
  $X \subseteq R^d$, $far(P, \omega, x) = \max_{p \in P} \omega(p) \cdot \|p - x\|$
- Output: $\boldsymbol{C} \subseteq P\ s.t. |far(P, \omega, x) - far(\boldsymbol{C}, \omega, x)| \leq O(\epsilon) \cdot far(P, \omega, x)$



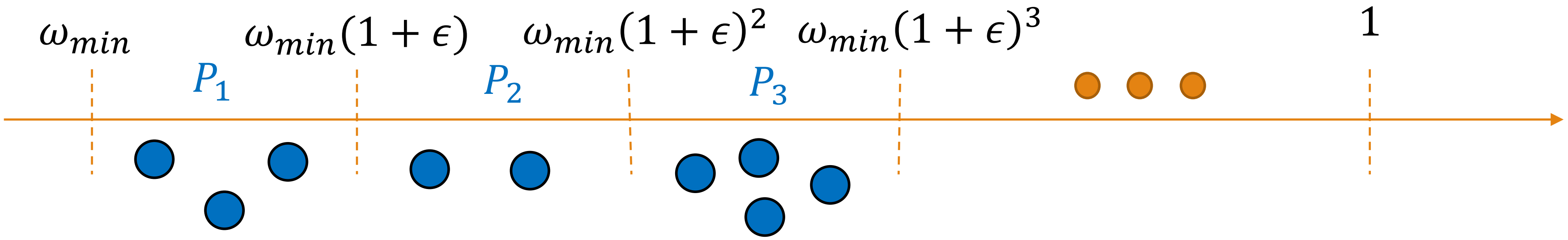
Coreset for 1-Center with Weighted Input

- Input: (P, ω, X, far) where $P \subseteq R^d, \omega: P \to R$ and $\sum \omega(p) = 1$,
  $X \subseteq R^d$, $far(P, \omega, x) = \max_{p \in P} \omega(p) \cdot \|p - x\|$
- Output: $\boldsymbol{C} \subseteq P\ s.t. \left|far(P, \omega, x) - far(\boldsymbol{C}, \omega, x)\right| \leq O(\epsilon) \cdot far(P, \omega, x)$



Coreset for 1-Center with Weighted Input

- Input: (P, ω, X, far) where $P \subseteq R^d, \omega: P \to R$ and $\sum \omega(p) = 1$,
  $X \subseteq R^d$, $far(P, \omega, x) = \max_{p \in P} \omega(p) \cdot \|p - x\|$
- Output: $\boldsymbol{C} \subseteq P \ s.t. \left|far(P, \omega, x) - far(\boldsymbol{C}, \omega, x)\right| \leq O(\epsilon) \cdot far(P, \omega, x)$

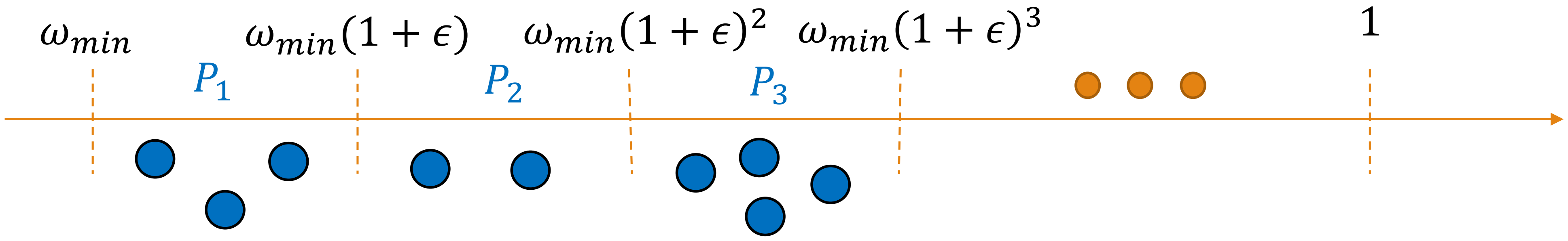


$$\mathbf{C} = \boldsymbol{C_1} \cup \boldsymbol{C_2} \cup \cdots \cup \boldsymbol{C_\lambda}$$

$$|\boldsymbol{C}| = |\boldsymbol{C_1} \cup \boldsymbol{C_2} \cup \cdots \cup \boldsymbol{C_\lambda}| = \lambda \cdot \left(\frac{1}{\epsilon}\right)^{O(d)}$$

Coreset for 1-Center with Weighted Input

- Input: (P, ω, X, far) where $P \subseteq R^d, \omega: P \rightarrow R$ and $\sum \omega(p) = 1$,
  $X \subseteq R^d$, $far(P, \omega, x) = \max_{p \in P} \omega(p) \cdot \|p - x\|$
- Output: $\boldsymbol{C} \subseteq P\ s.t. |far(P, \omega, x) - far(\boldsymbol{C}, \omega, x)| \leq O(\epsilon) \cdot far(P, \omega, x)$



- Left to prove that:
  For every $i \in \{1, \dots, \lambda\}$ and every $x \in X$:
  $far(P_i, \omega, x) - far(\boldsymbol{C_i}, \omega, x) \leq O(\epsilon) \cdot far(P_i, \omega, x)$

Coreset for 1-Center with Weighted Input

- Left to prove that:
  For every $i \in \{1, \dots, \lambda\}$ and every $q \in Q$:
  $far(P_i, \omega, x) - far(\boldsymbol{C_i}, \omega, x) \le O(\epsilon) \cdot far(P_i, \omega, x)$
- We know that:
  For every $i \in \{1, \dots, \lambda\}$ and every $q \in Q$:
  $far(P_i, x) - far(\boldsymbol{C_i}, x) \le O(\epsilon) \cdot far(P_i, x)$

we proved this in previous slides $\rightarrow$ $far(P_i, \Delta_\mathrm{i}, x) - far(\boldsymbol{C_i}, \Delta_\mathrm{i}, x) \le O(\epsilon) \cdot far(P_i, \Delta_i, x$

$far(P_i, \omega, x) = \omega(p^*)\|p^* - x\|.$
$far(P_i, \Delta_\mathrm{i}, x) = \omega'(p^{*\prime})\|p^{*\prime} - x\|.$

$$\frac{far(P_i, \omega, x)}{far(P_i, \Delta_\mathrm{i}, x)} = \frac{\omega(p^*)\|p^* - x\|}{\omega'(p^{*\prime})\|p^{*\prime} - x\|} \le \frac{\omega(p^*)\|p^* - x\|}{\omega'(p^{*\prime})\|p^* - x\|} = \frac{\omega(p^*)}{\omega'(p^{*\prime})} \le (1+\epsilon)$$

$$\boldsymbol{far(P_i, \omega, x) - far(C_i, \omega, x)}$$

$\Delta_i \le \omega(p_i)$ for every $p_i \in P_i$

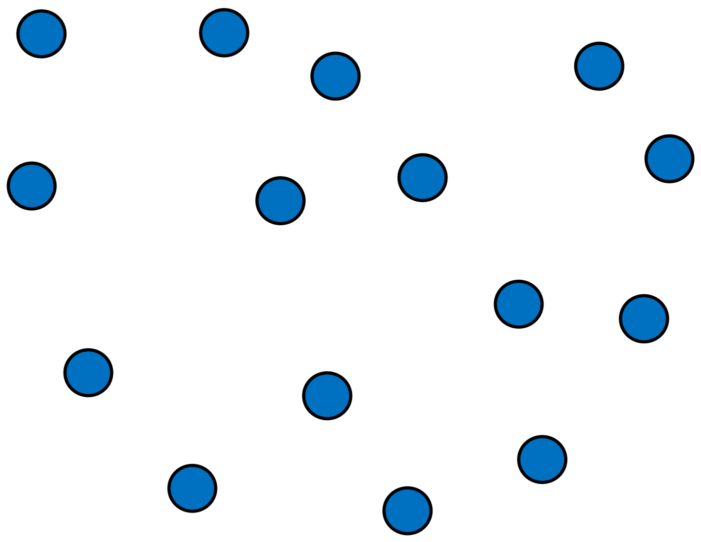
$$\le (1+\epsilon) \cdot far(P_i, \Delta_\mathrm{i}, x) - far(\boldsymbol{C_i}, \Delta_\mathrm{i}, x)$$

$$= \epsilon \cdot far(P_i, \Delta_\mathrm{i}, x) + far(P_i, \Delta_\mathrm{i}, x) - far(C_i, \Delta_\mathrm{i}, x) \le 2\epsilon \cdot far(P_i, \Delta_\mathrm{i}, x)$$

$$\le 2\epsilon \cdot far(P_i, \omega, x) = \boldsymbol{O(\epsilon) \cdot far(P_i, \omega, x)}$$

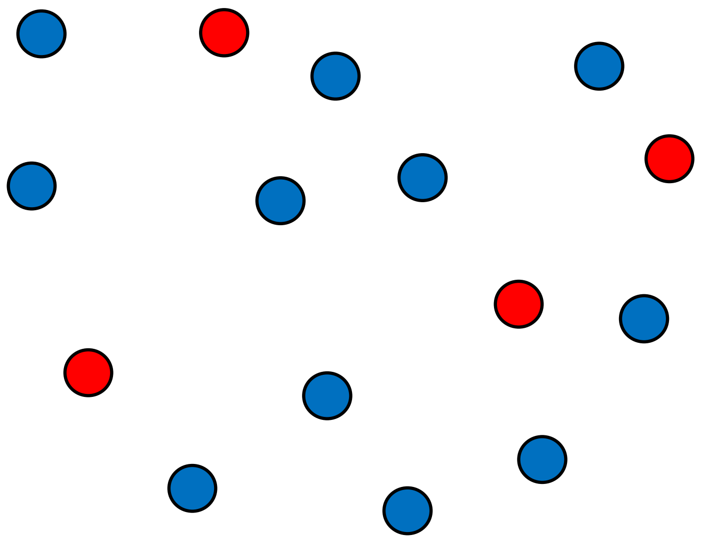
Coreset for 1-Line in R^2

- Input: $P \subseteq R^2, Q = \{\ell \mid \ell \text{ is a line in } R^2\}, dist(p,\ell) = \min_{x \in \ell} \|p - x\|_2$
- Output: $\boldsymbol{C} \subseteq P \ s.t. \ \forall \ell \in Q: \max_{p \in P} dist(p,\ell) - \max_{c \in C} dist(c,\ell) \leq \epsilon \cdot \max_{p \in P} dist(p,\ell)$



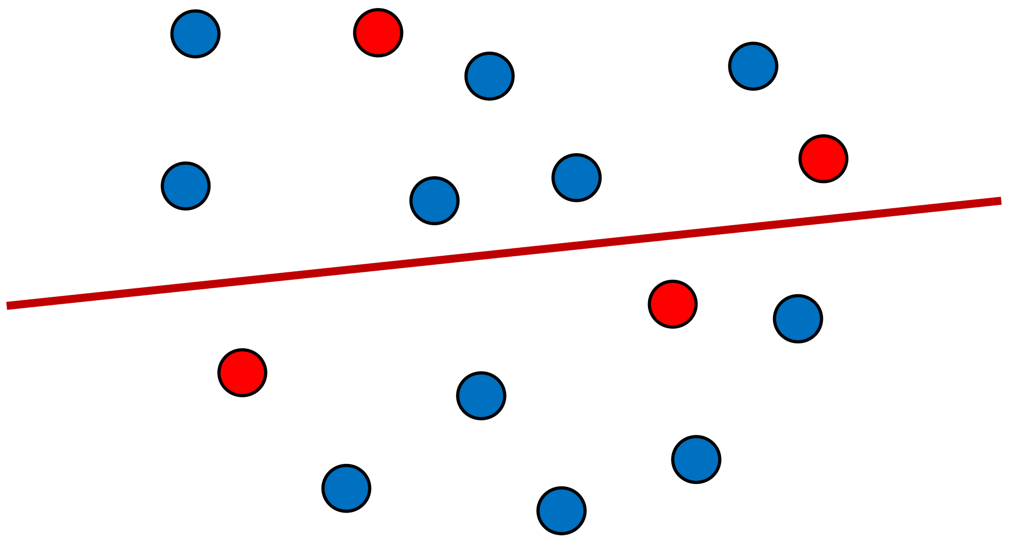
Coreset for 1-Line in R^2

- Input: $P \subseteq R^2, Q = \{\ell \mid \ell \text{ is a line in } R^2\}, dist(p,\ell) = \min_{x\in\ell} \|p - x\|_2$
- Output: $\boldsymbol{C} \subseteq P \text{ s.t. } \forall \ell \in Q: \max_{p\in P} dist(p,\ell) - \max_{c\in C} dist(c,\ell) \leq \epsilon \cdot \max_{p\in P} dist(p,\ell)$



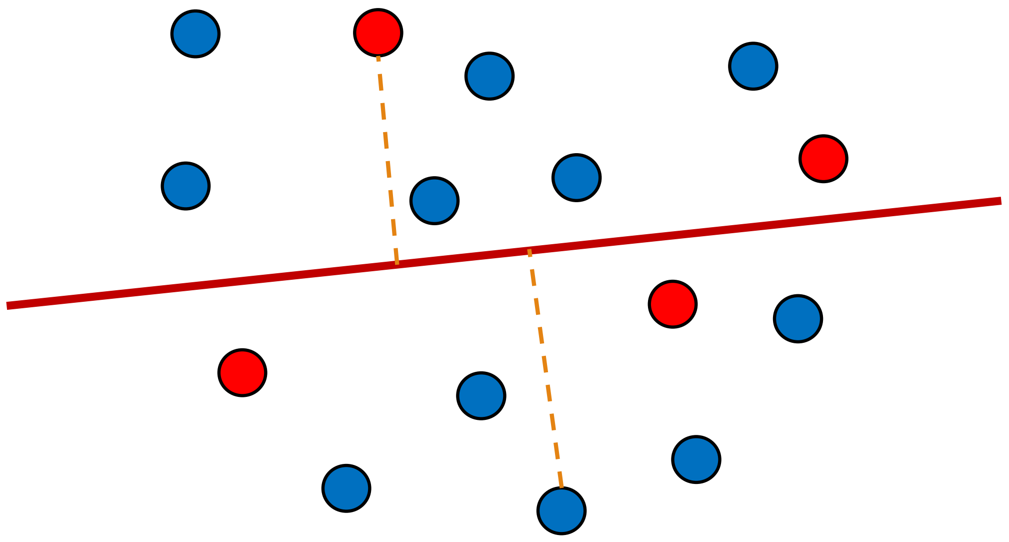
Coreset for 1-Line in R^2

- <u>Input:</u> $P \subseteq R^2, Q = \{\ell \mid \ell \textit{ is a line in } R^2\}, dist(p,\ell) = \min_{x\in\ell} \|p - x\|_2$
- <u>Output:</u> $\boldsymbol{C} \subseteq P \; s.t. \; \forall \ell \in Q: \max_{p\in P} dist(p,\ell) - \max_{c\in C} dist(c,\ell) \le \epsilon \cdot \max_{p\in P} dist(p,\ell)$

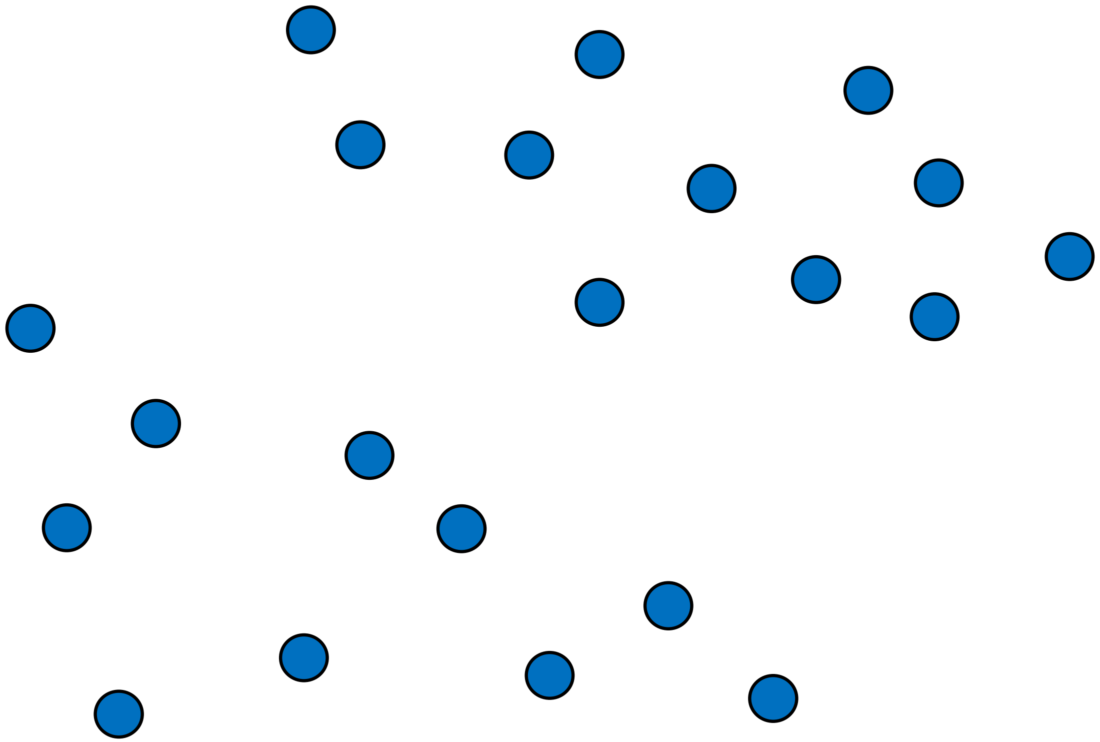


Coreset for 1-Line in R^2

- Input: $P \subseteq R^2, Q = \{\ell \mid \ell \textit{ is a line in } R^2\}, dist(p,\ell) = \min_{x\in\ell} \|p - x\|_2$
- Output: $\boldsymbol{C} \subseteq P \; s.t.\; \forall \ell \in Q: \max_{p\in P} dist(p,\ell) - \max_{c\in C} dist(c,\ell) \leq \epsilon \cdot \max_{p\in P} dist(p,\ell)$



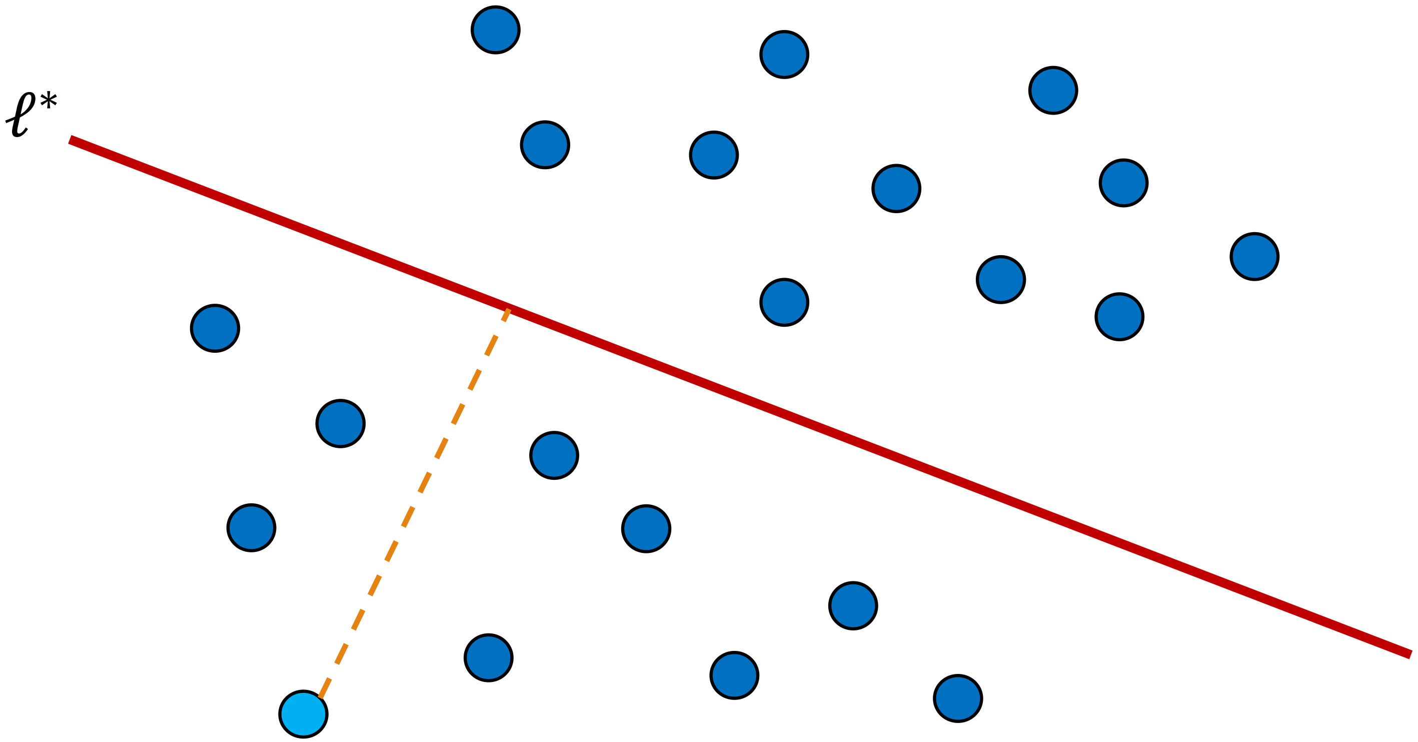
Coreset for 1-Line in R^2



Coreset for 1-Line in R^2

ℓ^* is the line that minimizes

$$\max_{p \in P} dist(p, \ell)$$

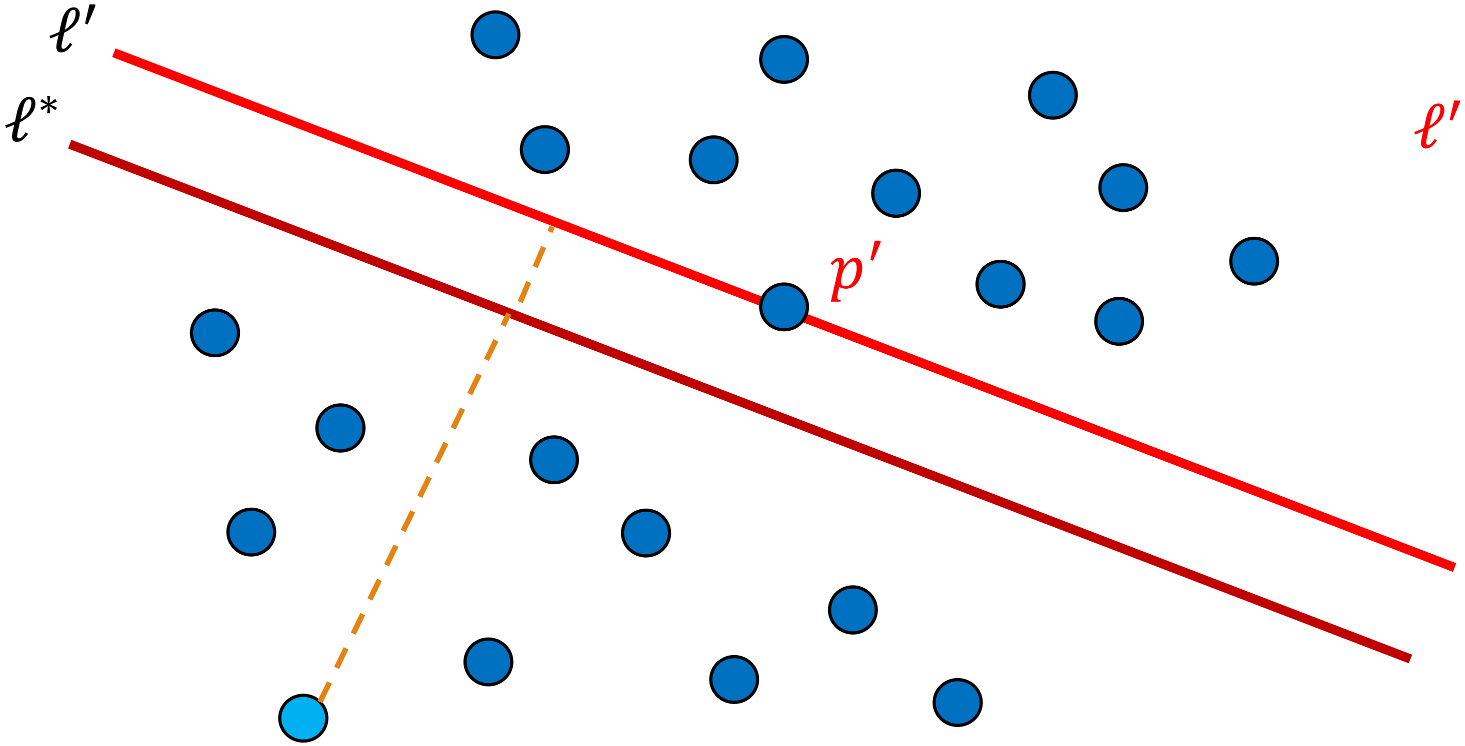


$$p^* = arg \max_{p \in P} dist(p, \ell^*)$$

Coreset for 1-Line in R^2

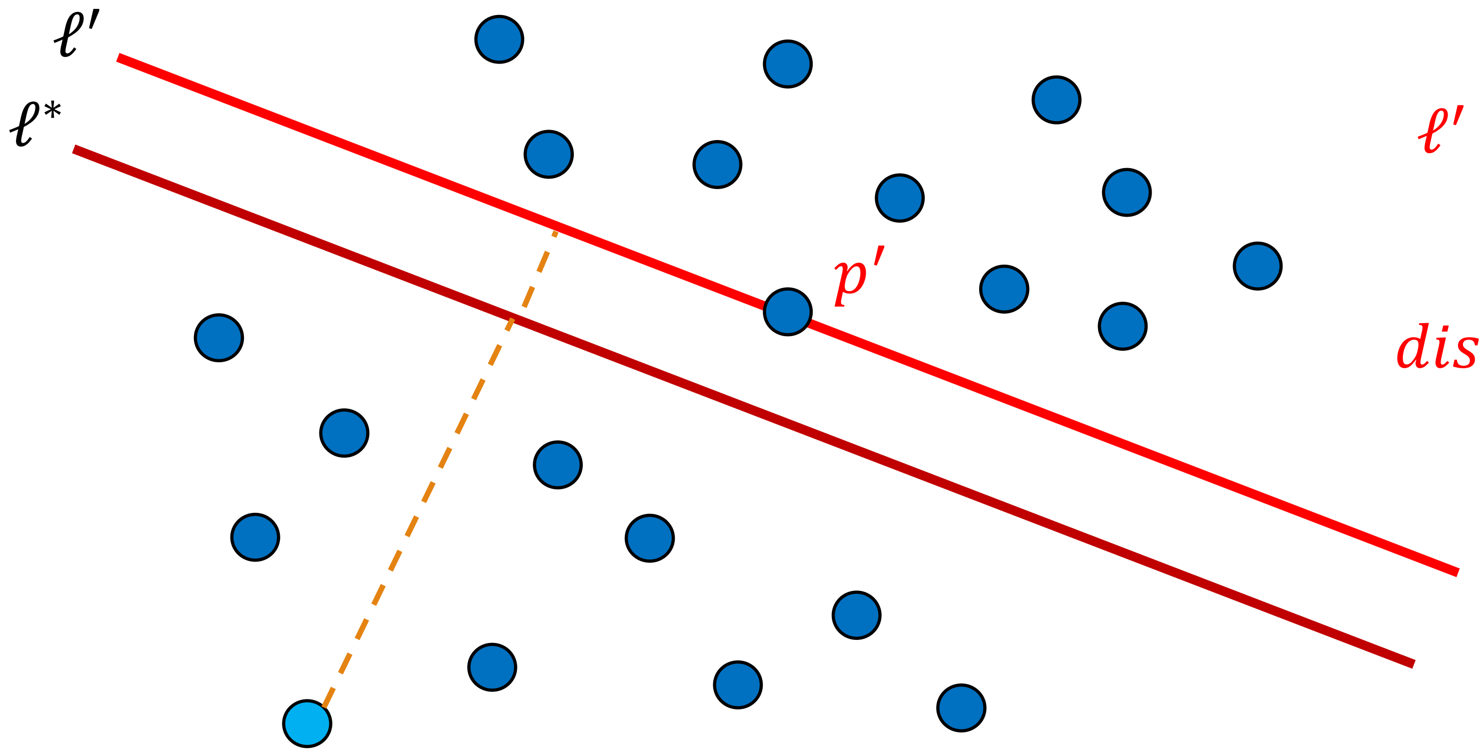
ℓ^* is the line that minimizes
$$\max_{p \in P} dist(p, \ell)$$

ℓ' is the translation of ℓ^* to $\ell^{*\prime}s$ closest point p'



$$p^* = arg \max_{p \in P} dist(p, \ell^*)$$

Coreset for 1-Line in R^2



$p^* = arg \max_{p \in P} dist(p, \ell^*)$

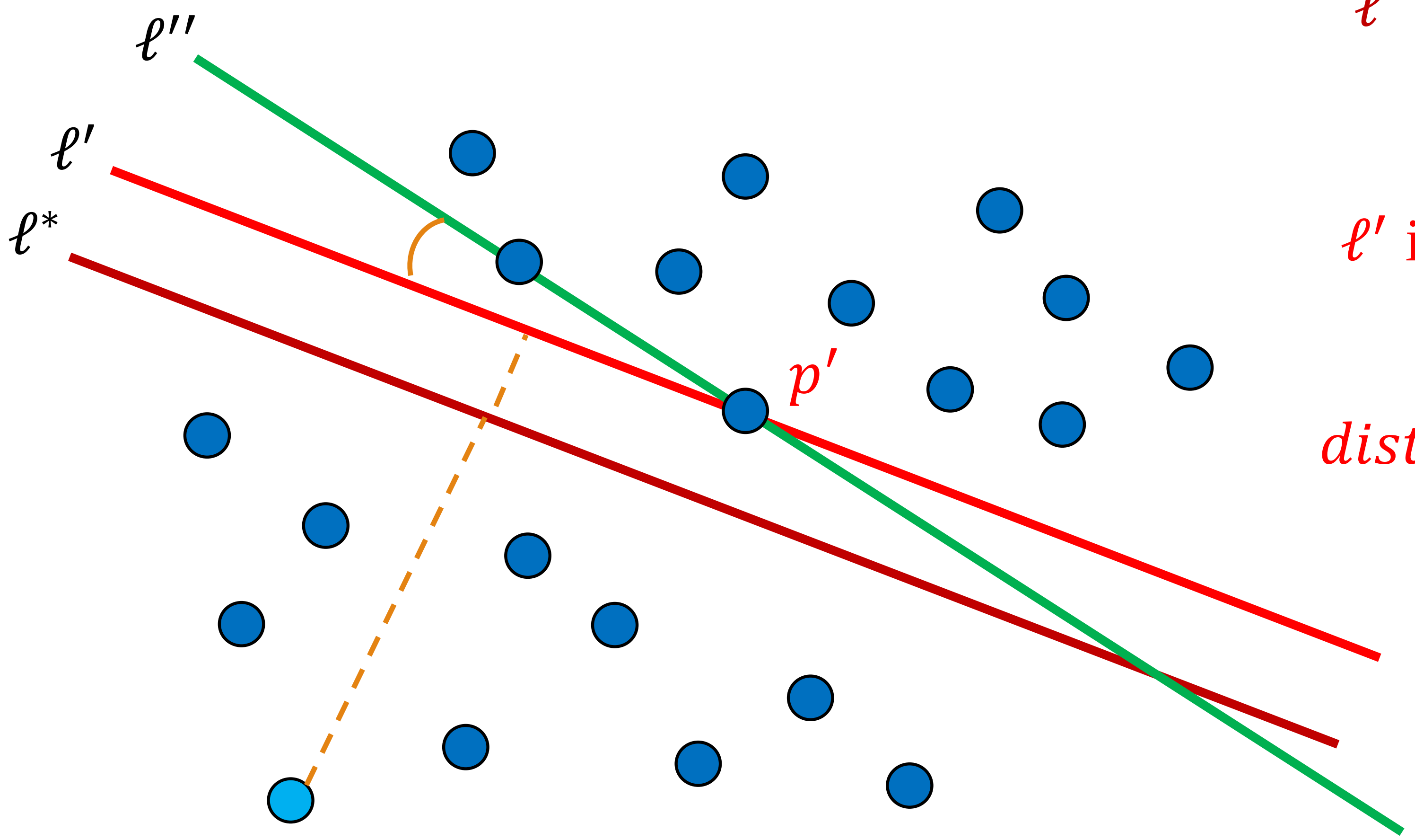
ℓ^* is the line that minimizes

$\max_{p \in P} dist(p, \ell)$

ℓ' is the translation of ℓ^* to $\ell^{*\prime}s$ closest point p'

$dist(p, \ell') \leq 2 \cdot dist(p, \ell^*)$

Coreset for 1-Line in R^2



$p^* = arg \max_{p \in P} dist(p, \ell^*)$

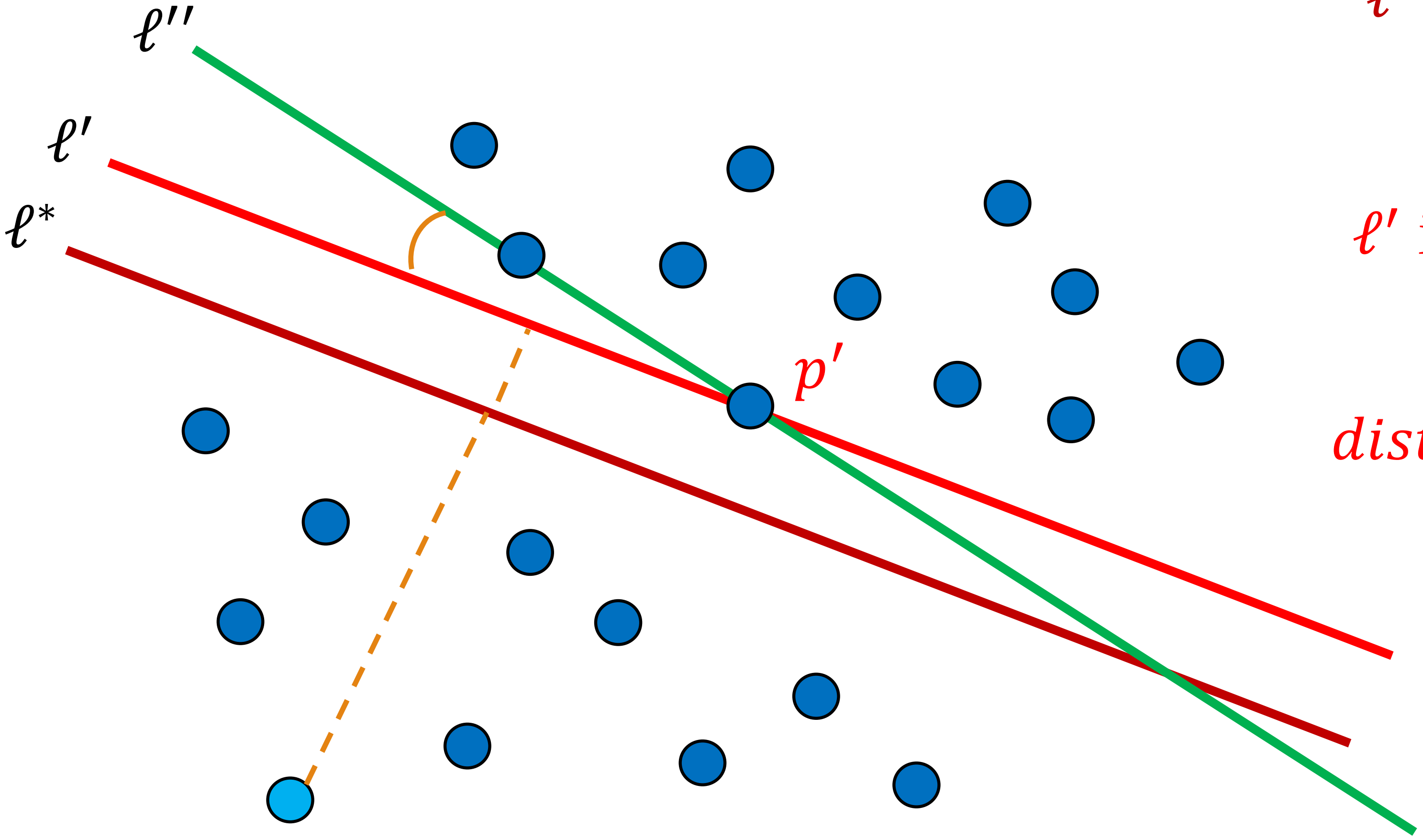
ℓ^* is the line that minimizes $\max_{p \in P} dist(p, \ell)$

ℓ' is the translation of ℓ^* to $\ell^{*\prime}s$ closest point p'

$dist(p, \ell') \leq 2 \cdot dist(p, \ell^*)$

ℓ'' is the rotation of ℓ' around p' to $\ell''s$ closest point

Coreset for 1-Line in R^2



ℓ^* is the line that minimizes $\max_{p \in P} dist(p, \ell)$

ℓ' is the translation of ℓ^* to $\ell^{*\prime}s$ closest point p'

$dist(p, \ell') \leq 2 \cdot dist(p, \ell^*)$

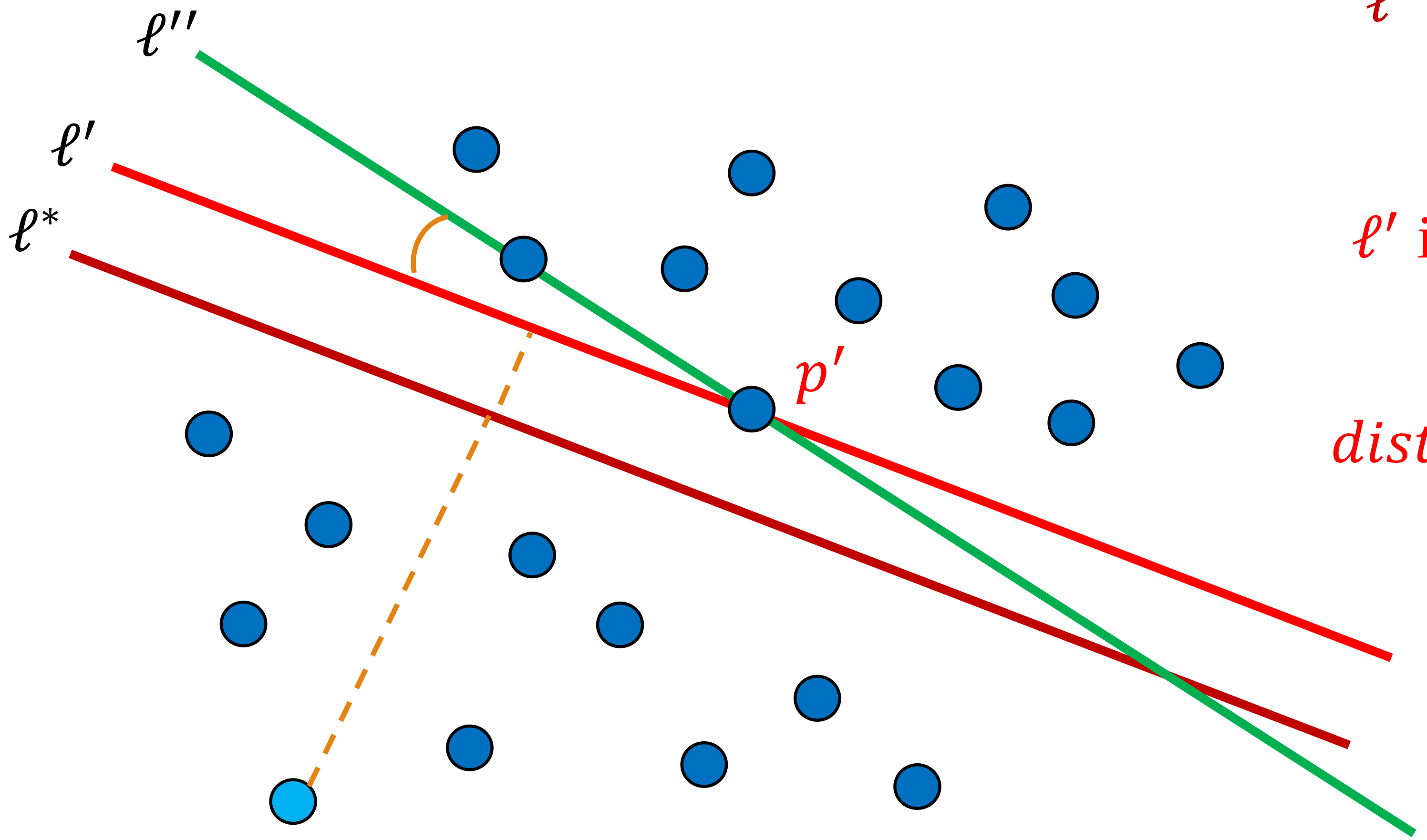
ℓ'' is the rotation of ℓ' around p' to $\ell''s$ closest point

$dist(p, \ell'') \leq 2 \cdot dist(p, \ell')$

$p^* = arg \max_{p \in P} dist(p, \ell^*)$

Coreset for 1-Line in R^2

$$\boldsymbol{dist(p, \ell'') \le 4 \cdot dist(p, \ell^*)}$$



$$p^* = arg \max_{p \in P} dist(p, \ell^*)$$

ℓ^* is the line that minimizes $\max_{p \in P} dist(p, \ell)$

ℓ' is the translation of ℓ^* to $\ell^{*\prime}s$ closest point p'

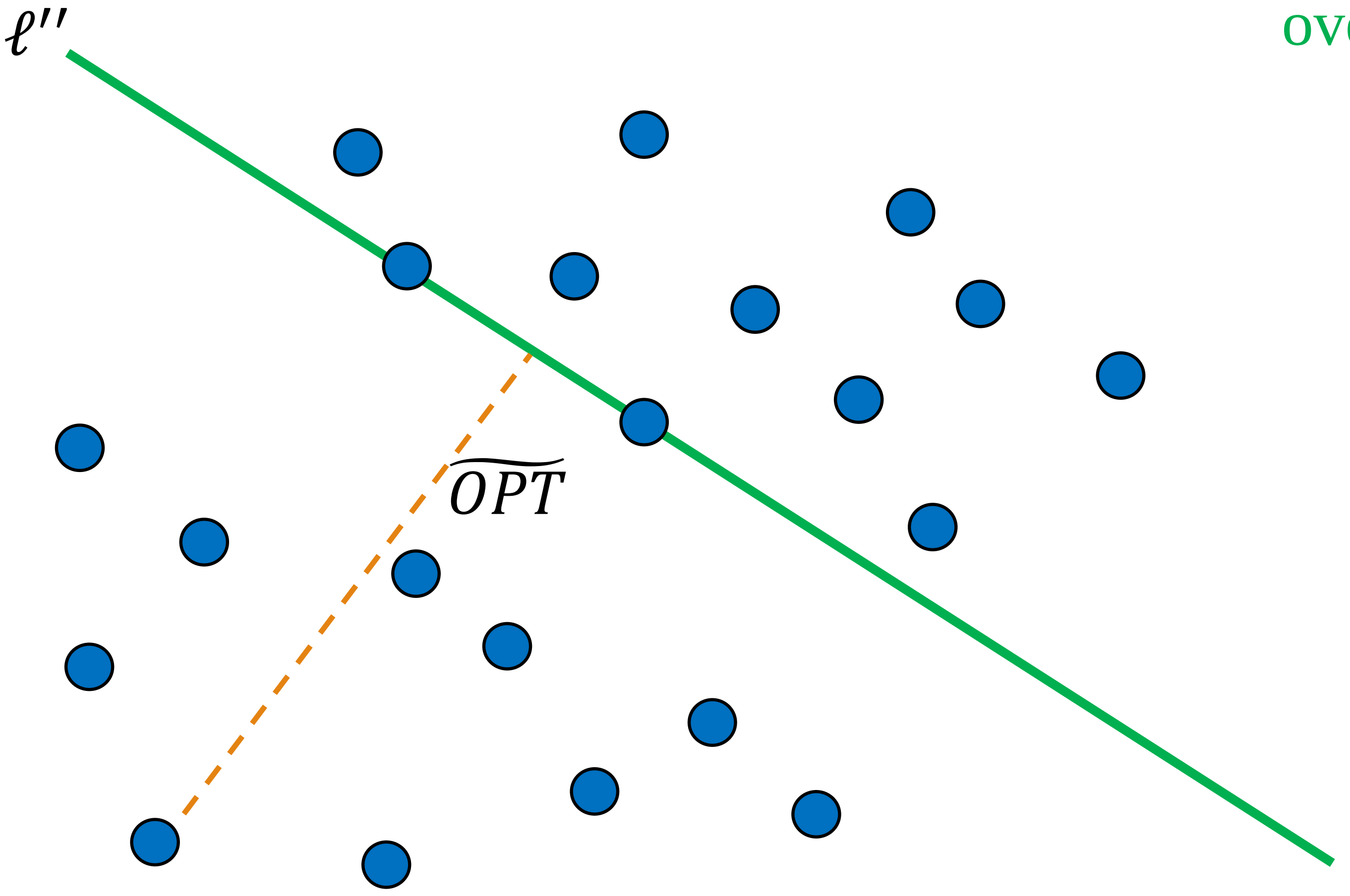
$$dist(p, \ell') \le 2 \cdot dist(p, \ell^*)$$

ℓ'' is the rotation of ℓ' around p' to $\ell''{}'s$ closest point

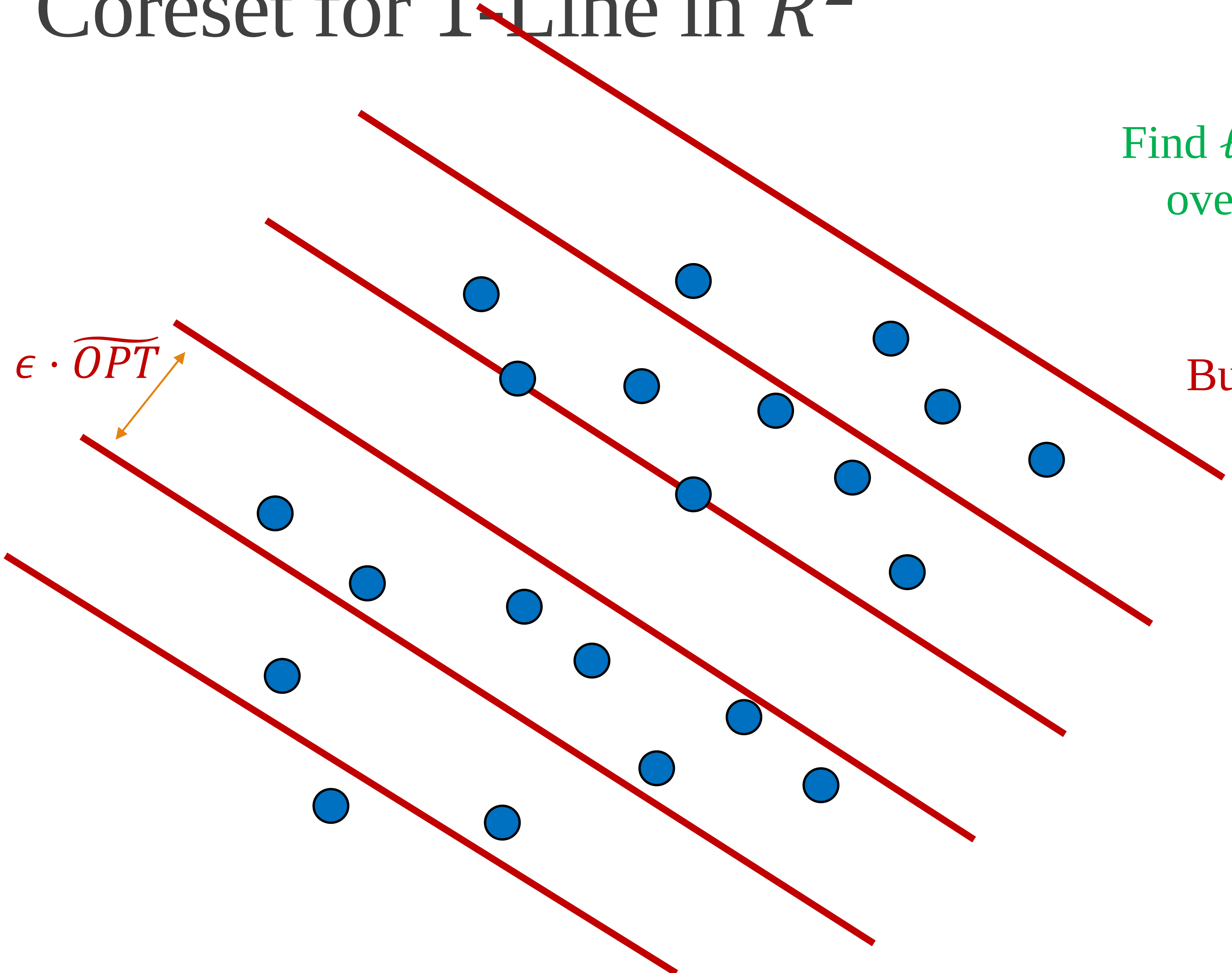
$$dist(p, \ell'') \le 2 \cdot dist(p, \ell')$$

Coreset for 1-Line in R^2

Find ℓ'' by exhaustive search over every pair of points.
$\boldsymbol{O(n^3)}$



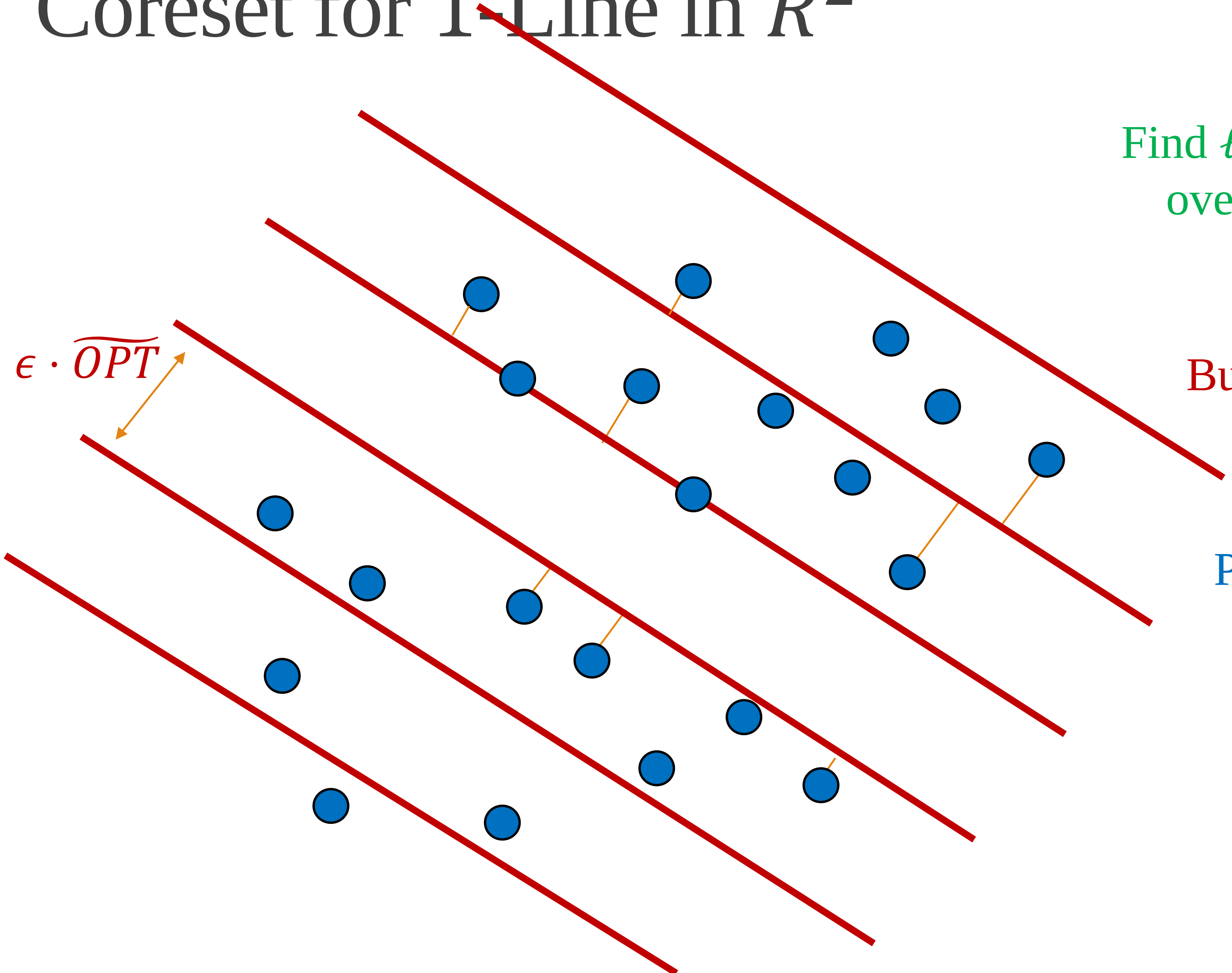
Coreset for 1-Line in R^2



Find ℓ'' by exhaustive search over every pair of points.
$\boldsymbol{O(n^3)}$

Build a grid of lines with $\epsilon \cdot \widetilde{OPT}$ distance

Coreset for 1-Line in R^2

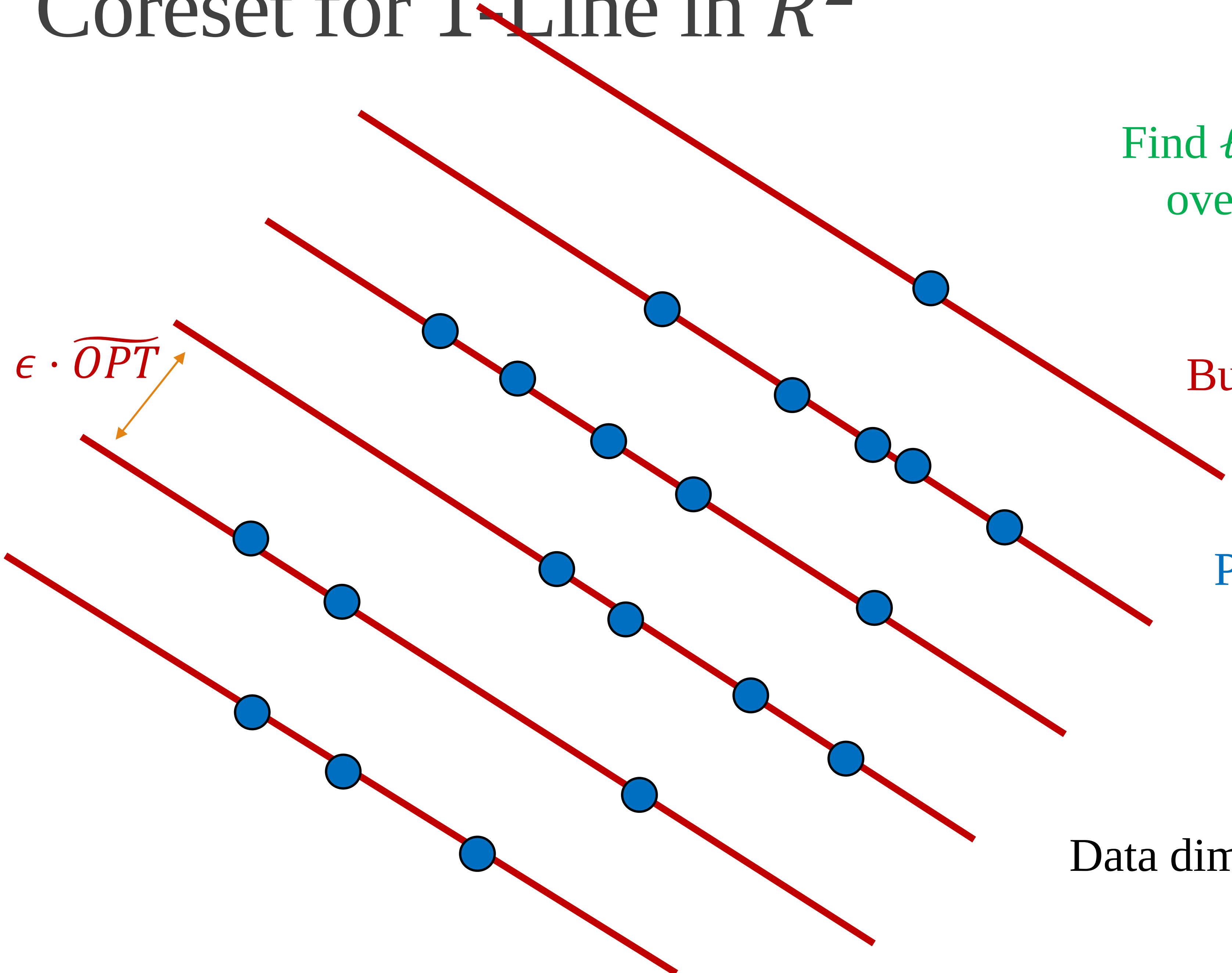


Find ℓ'' by exhaustive search over every pair of points. $\boldsymbol{O(n^3)}$

Build a grid of lines with $\epsilon \cdot \widetilde{OPT}$ distance

Project each point onto it's closest line

Coreset for 1-Line in R^2



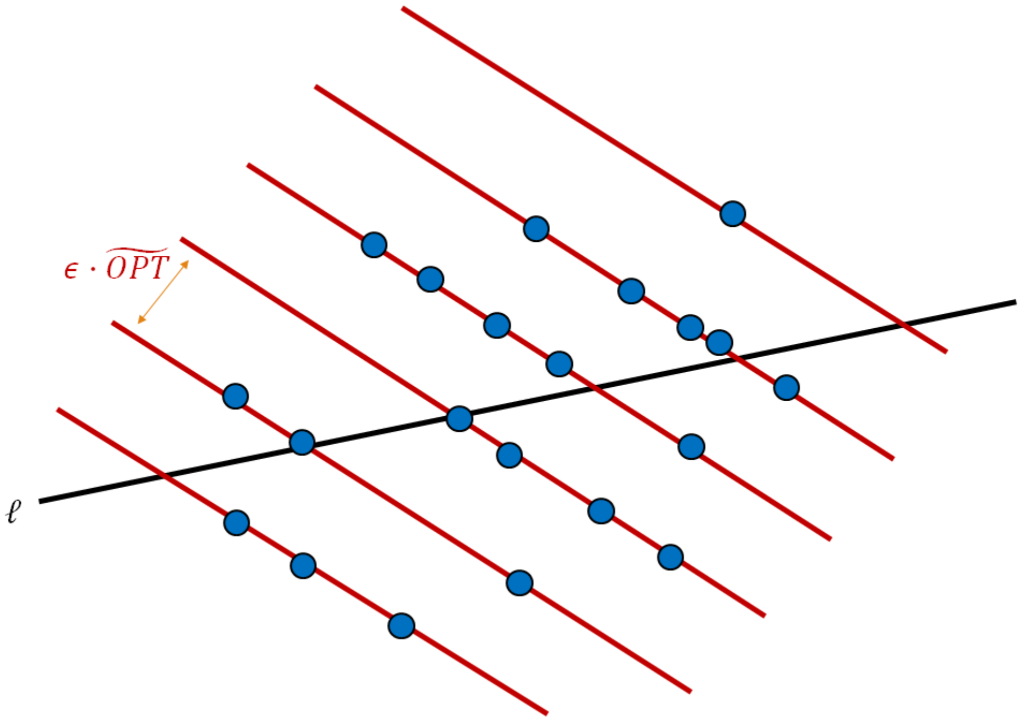
Find ℓ'' by exhaustive search over every pair of points.
$\boldsymbol{O(n^2)}$

Build a grid of lines with $\epsilon \cdot \widetilde{OPT}$ distance

Project each point onto it's closest line

Data dimension is now reduced.

Coreset for 1-Line in R^2



Claim: The projected n points P'
are a "coreset" (not part of the input data)
for any line query:

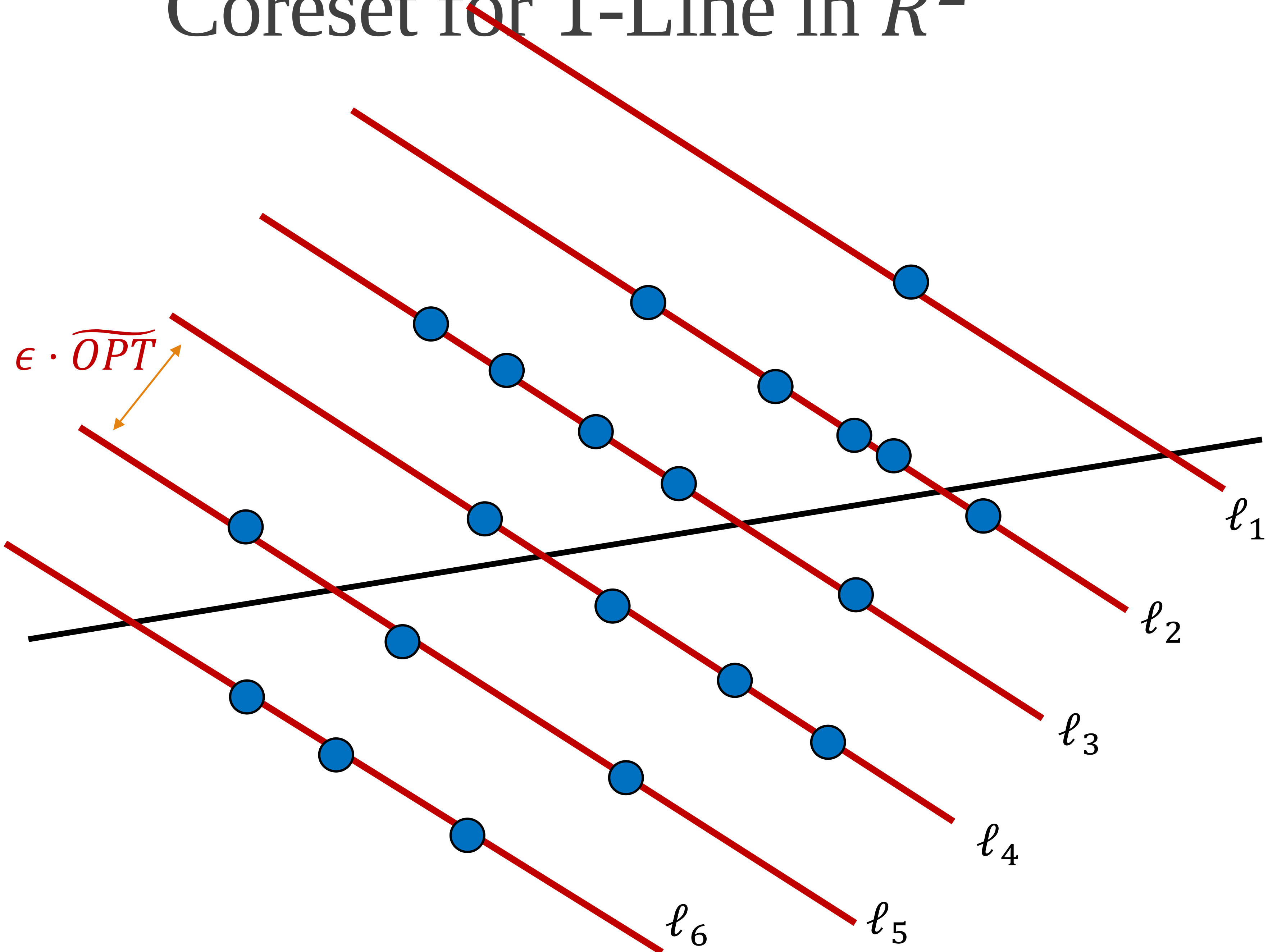
$$\max_{p \in P} dist(p, \ell) - \max_{p \in P'} dist(p, \ell) \leq \epsilon \cdot \widetilde{OPT}$$

$$\leq 4\epsilon \cdot OPT$$

$$\leq 4\epsilon \cdot \max_{p \in P} dist(p, \ell)$$

$\rightarrow$ Run with $\epsilon' = \frac{\epsilon}{4}$

Coreset for 1-Line in R^2

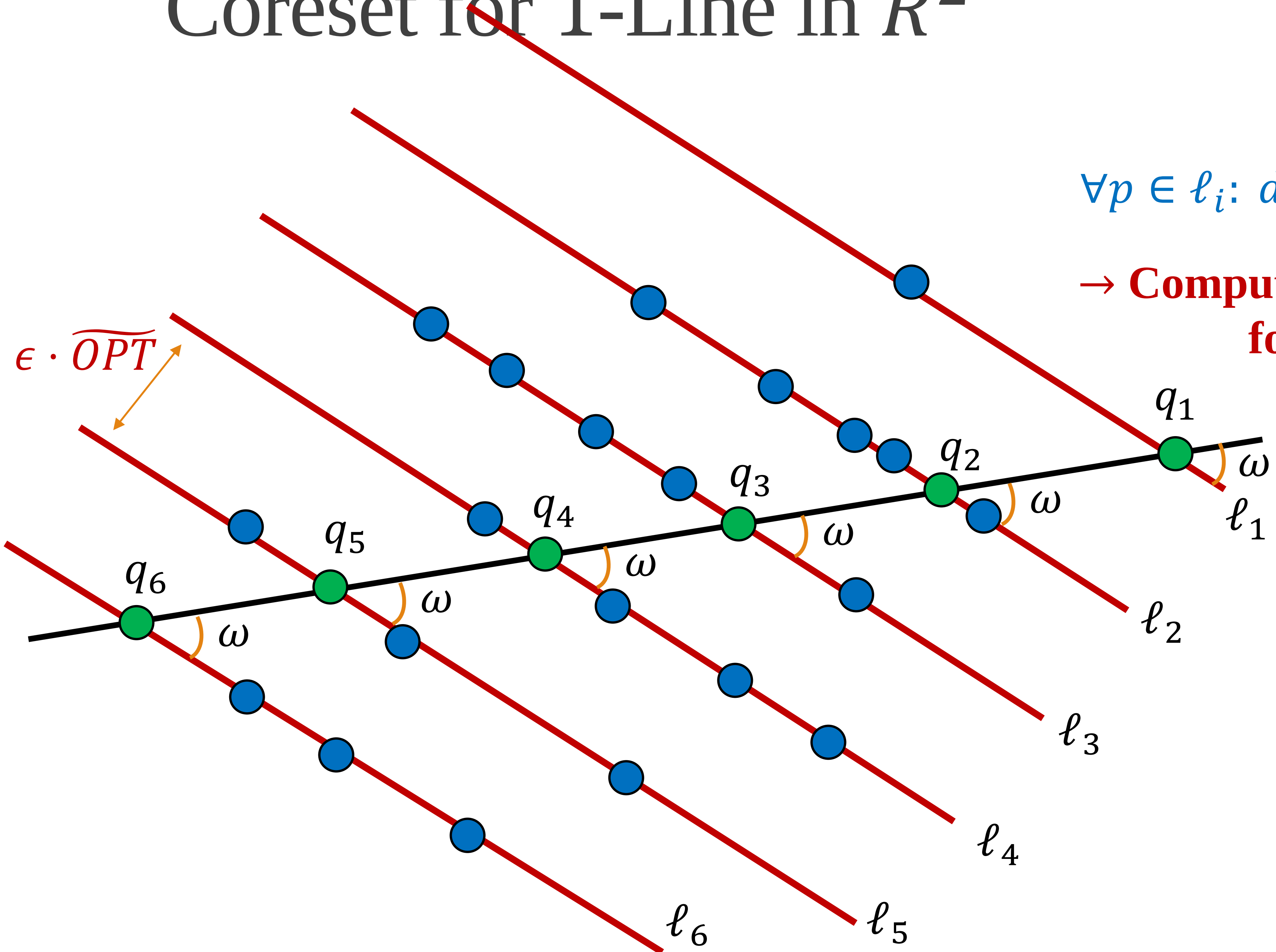


Coreset for 1-Line in R^2

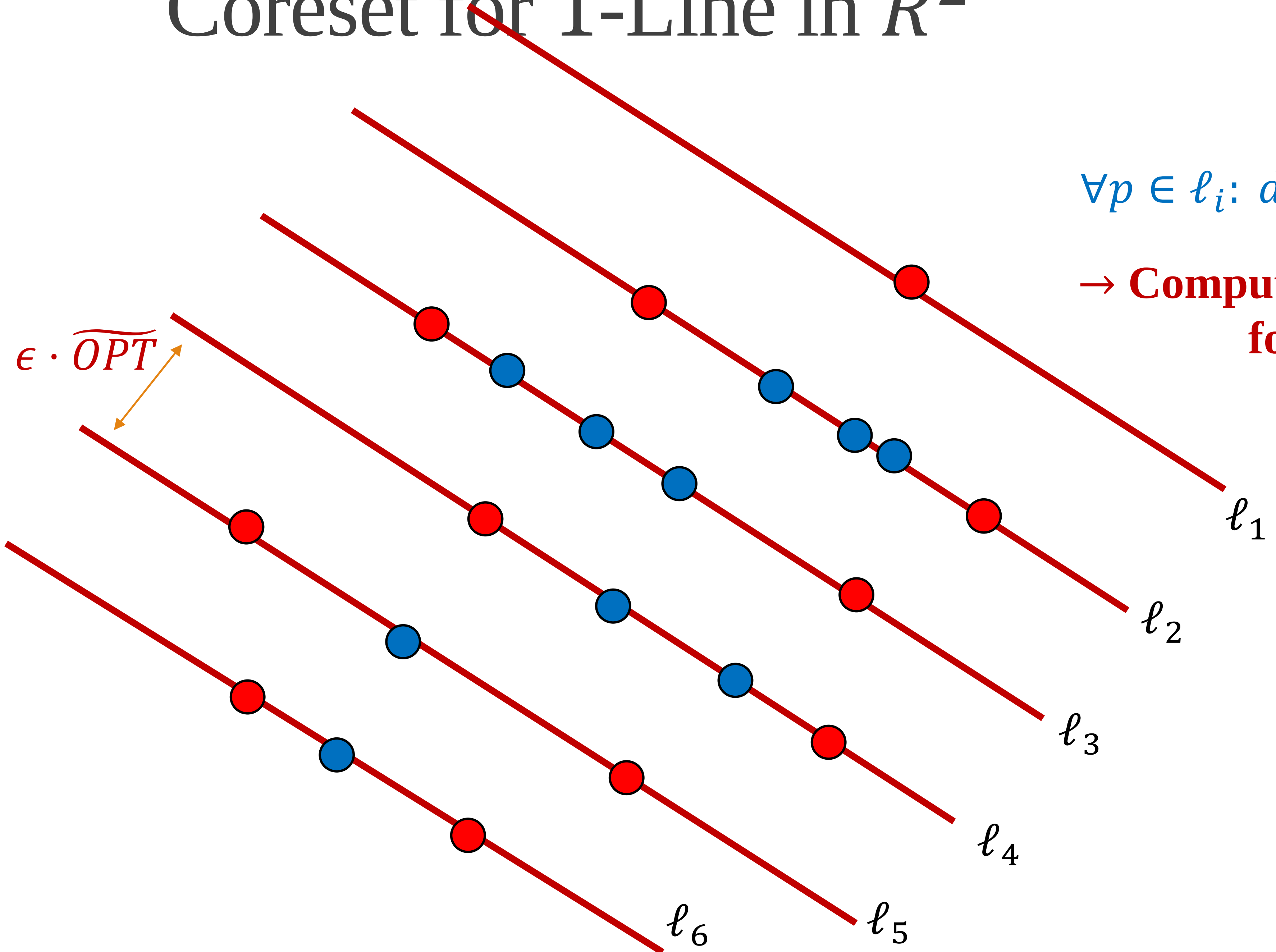
Has no effect since it is the same weight for all points

$\forall p \in \ell_i$: $dist(p, \ell) = \omega \cdot dist(p, q_i)$

→ Compute a 1-Center coreset C_i for each line ℓ_i!



Coreset for 1-Line in R^2



Has no effect since it is the same weight for all points

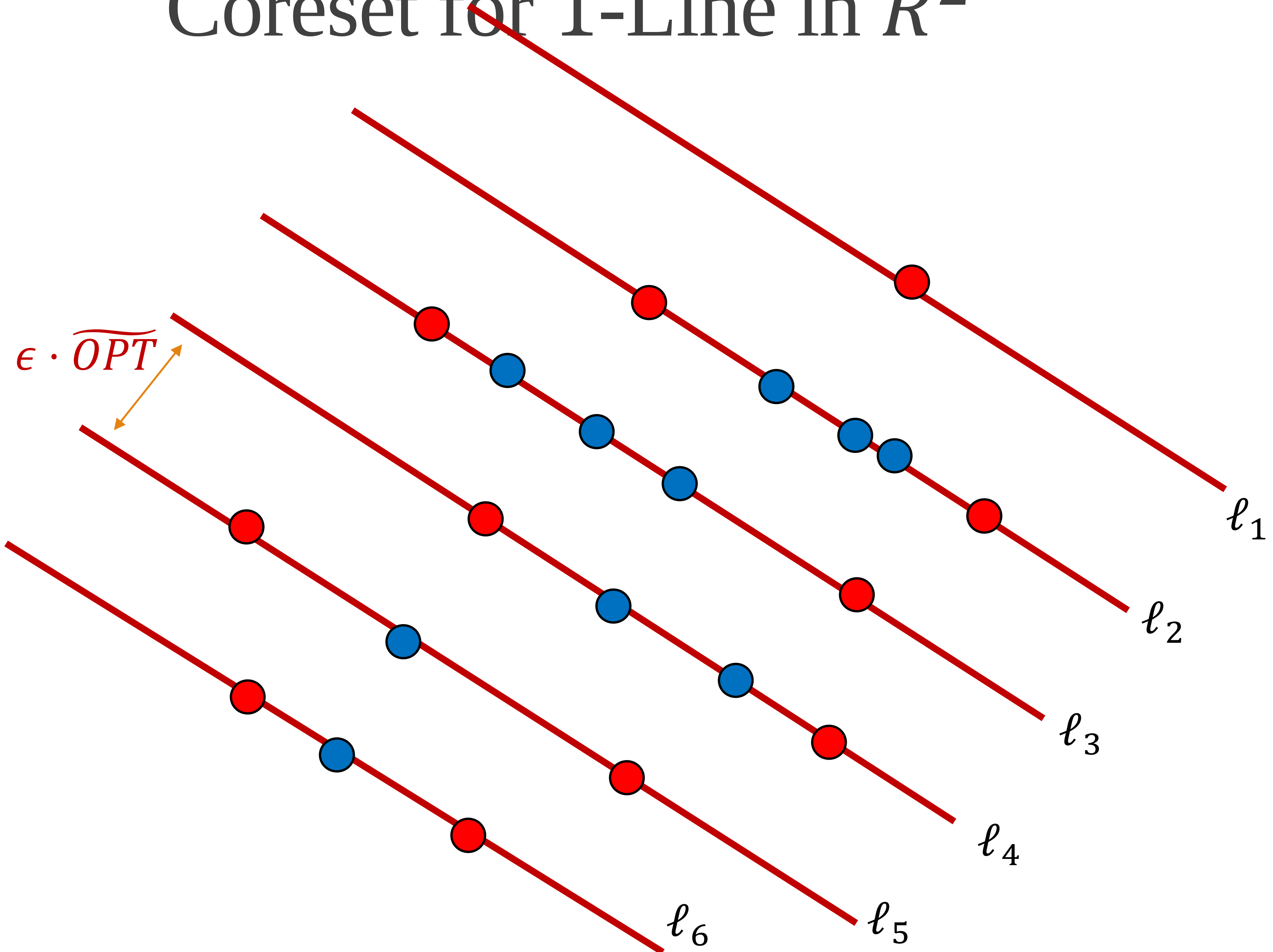
$\forall p \in \ell_i$: $dist(p, \ell) = \omega \cdot dist(p, q_i)$

→ Compute a 1-Center coreset C_i for each line ℓ_i!

$$C = \bigcup C_i$$

since a union of two coresets is a coreset.

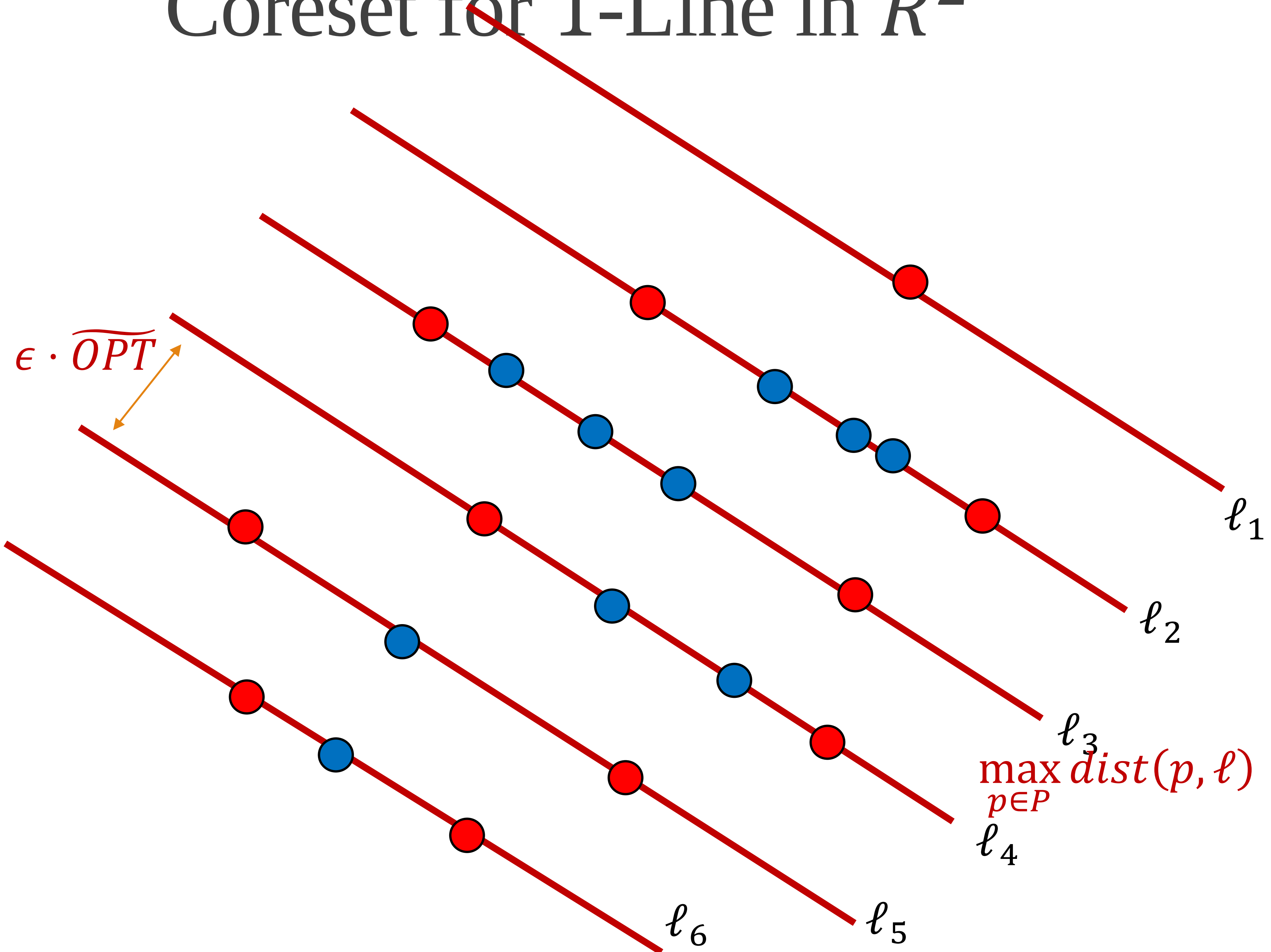
Coreset for 1-Line in R^2



<u>Problem:</u>
The coreset is not part of the input data.

<u>Solution:</u>
Pick the closest points in the input data to the points of C.

Coreset for 1-Line in R^2



<u>Problem:</u>
The coreset is not part of the input data.

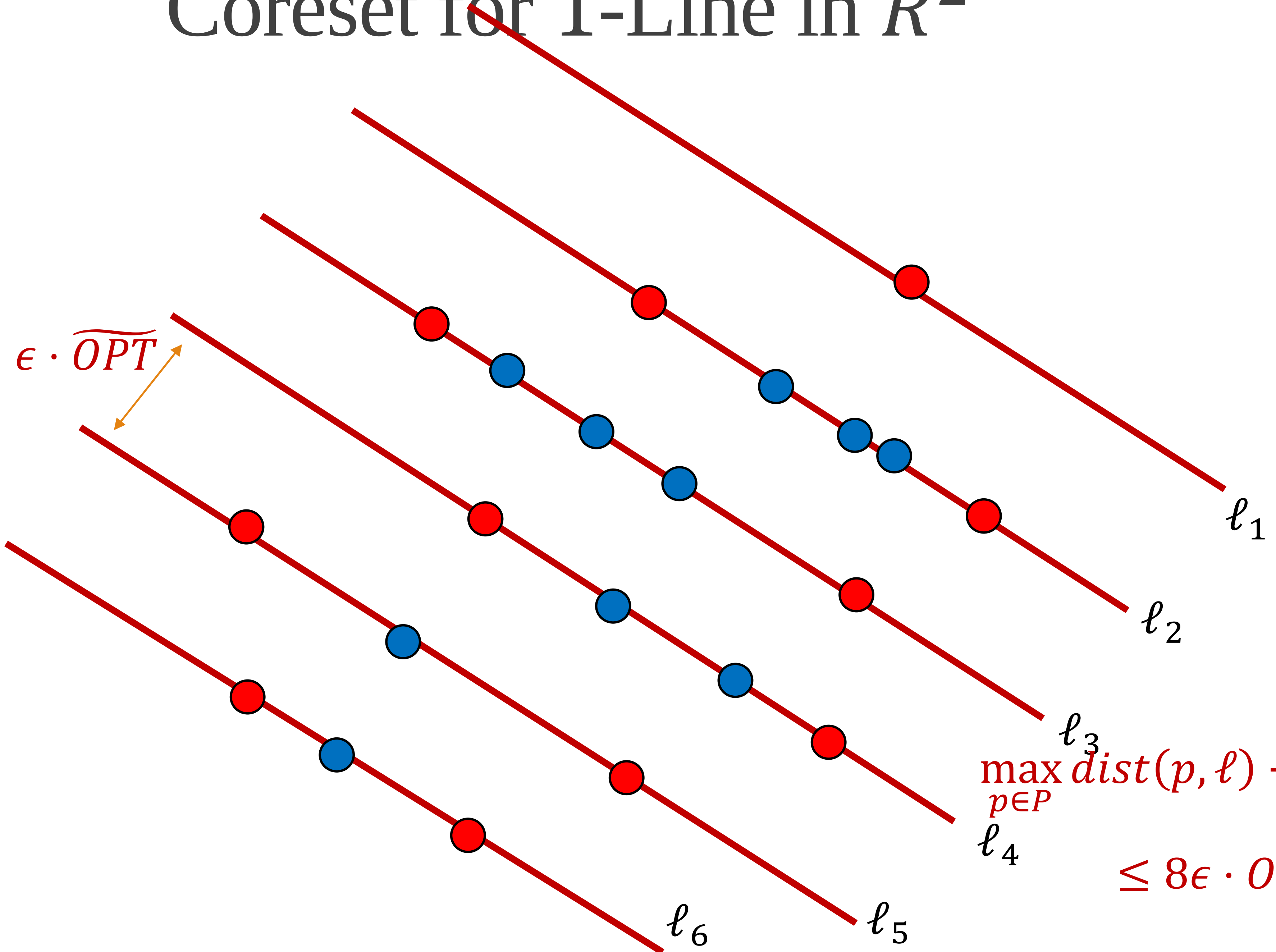
<u>Solution:</u>
Pick the closest points in the input data to the points of C.

→ **This adds another error of $\boldsymbol{\epsilon \cdot \widetilde{OPT}}$**

$$\max_{p \in P} dist(p, \ell) \leq \max_{p \in P'} dist(p, \ell) + 2\epsilon \cdot \widetilde{OPT}$$

$$\leq (1 + 8\epsilon) \cdot \max_{p \in P'} dist(p, \ell)$$

Coreset for 1-Line in R^2



<u>Problem:</u>
The coreset is not part of the input data.

<u>Solution:</u>
Pick the closest points in the input data to the points of C.

→ **This adds another error of $\boldsymbol{\epsilon \cdot \widetilde{OPT}}$**

$$\max_{p \in P} dist(p, \ell) - \max_{p \in P'} dist(p, \ell) \leq 2\epsilon \cdot \widetilde{OPT}$$

$$\leq 8\epsilon \cdot OPT \leq 8\epsilon \cdot \max_{p \in P} dist(p, \ell)$$

Coreset for 1-Line in R^2

<u>Total time:</u>

$O(n^3)$.

<u>Coreset size:</u>

$$|C| \leq 2 \cdot \#lines = 2 \cdot \frac{2}{\epsilon} = \frac{4}{\epsilon}.$$

Coreset for 1-Line in R^2

Total time:

$O(n^3)$.

Coreset size:

$$|C| \leq 2 \cdot \#lines = 2 \cdot \frac{2}{\epsilon} = \frac{4}{\epsilon}.$$

Improvement:

Run the above algorithm using the streaming tree.

Run on batches of size $2 \cdot |C| = \frac{8}{\epsilon}$.

Total time:

$$\mathrm{O}(\mathrm{n} \cdot TimeForBatch) = \mathrm{O}\left(n \cdot \left(\frac{8}{\epsilon}\right)^3\right).$$

Error for streaming tree:

The error increases to $(1+\epsilon)^{\log n} \sim (1+\epsilon \log n)$

→ Run with $\epsilon' = \frac{\epsilon}{\log n}$.

Off-line Coreset Construction

1) **(Reduce):** C is a $1+\epsilon$ - (core) set for P if:

$$\forall q \in Q, |f(P,q) - f(C,q)| \leq \epsilon f(P,q)$$

2) **(Merge):** If C_1 is a coreset for P_1 and C_2 is a coreset for P_2, then:

$$|f(P_1 \cup P_2) - f(C_1 \cup C_2)| \leq \epsilon f(P_1 \cup P_2)$$

$\boldsymbol{P, |P| = m}$

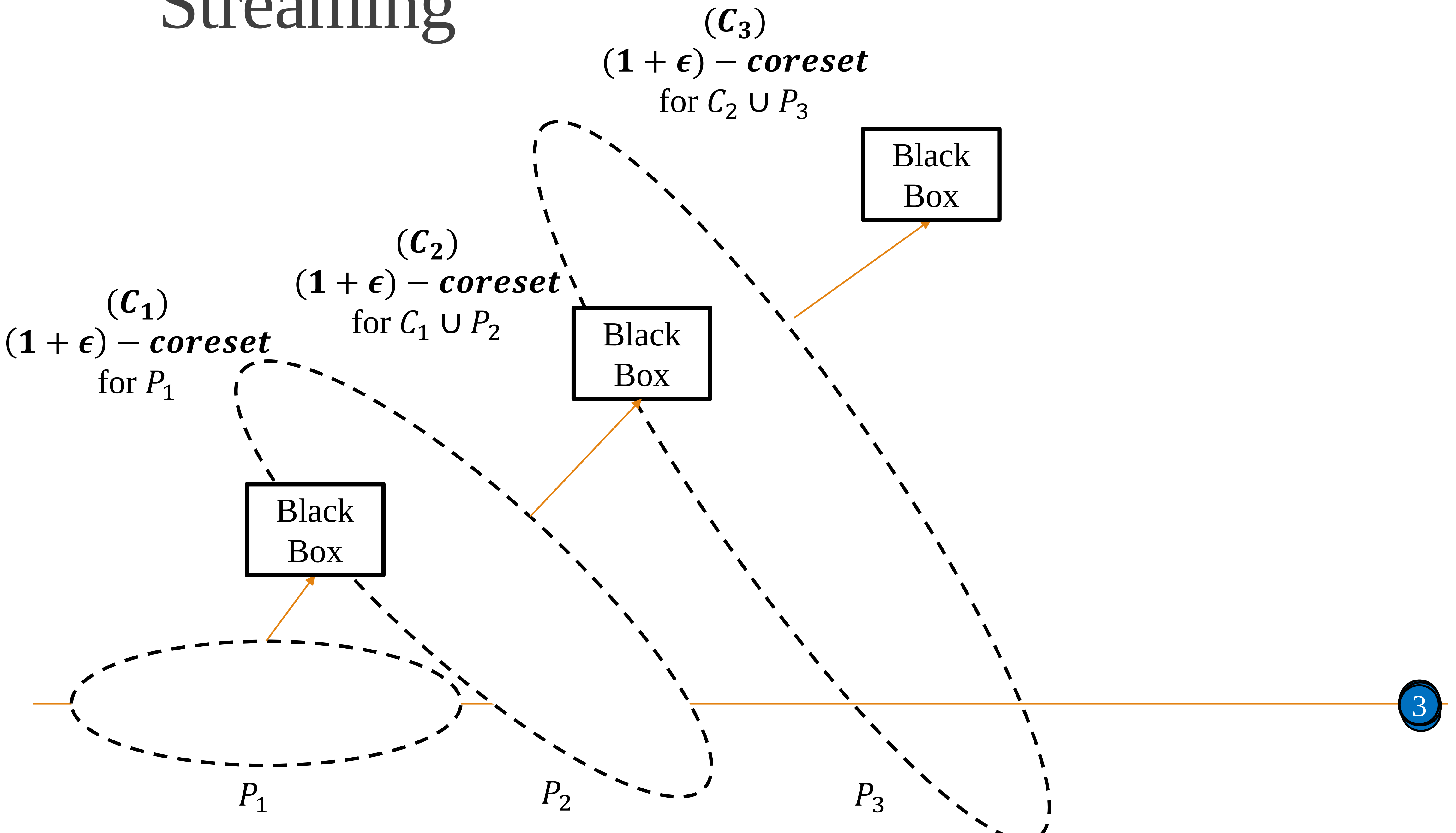
Input: **A query space $\boldsymbol{(P, Q, f)}$**

Black Box

Output: **A (core) set $\boldsymbol{C}$ for P, $\boldsymbol{|C| \leq \frac{|P|}{2}}$**

$\boldsymbol{C, |C| = \frac{m}{2}}$

Streaming



(C_3)
$(1+\epsilon)-coreset$
for $C_2 \cup P_3$
Black Box
(C_2)
$(1+\epsilon)-coreset$
for $C_1 \cup P_2$
Black Box
(C_1)
$(1+\epsilon)-coreset$
for P_1
Black Box
P_1
P_2
P_3

Proof

$C_1 = P_1$

C_2 is a coreset for $P_1 \cup P_2$

C_i is a coreset for $C_{i-1} \cup P_i$

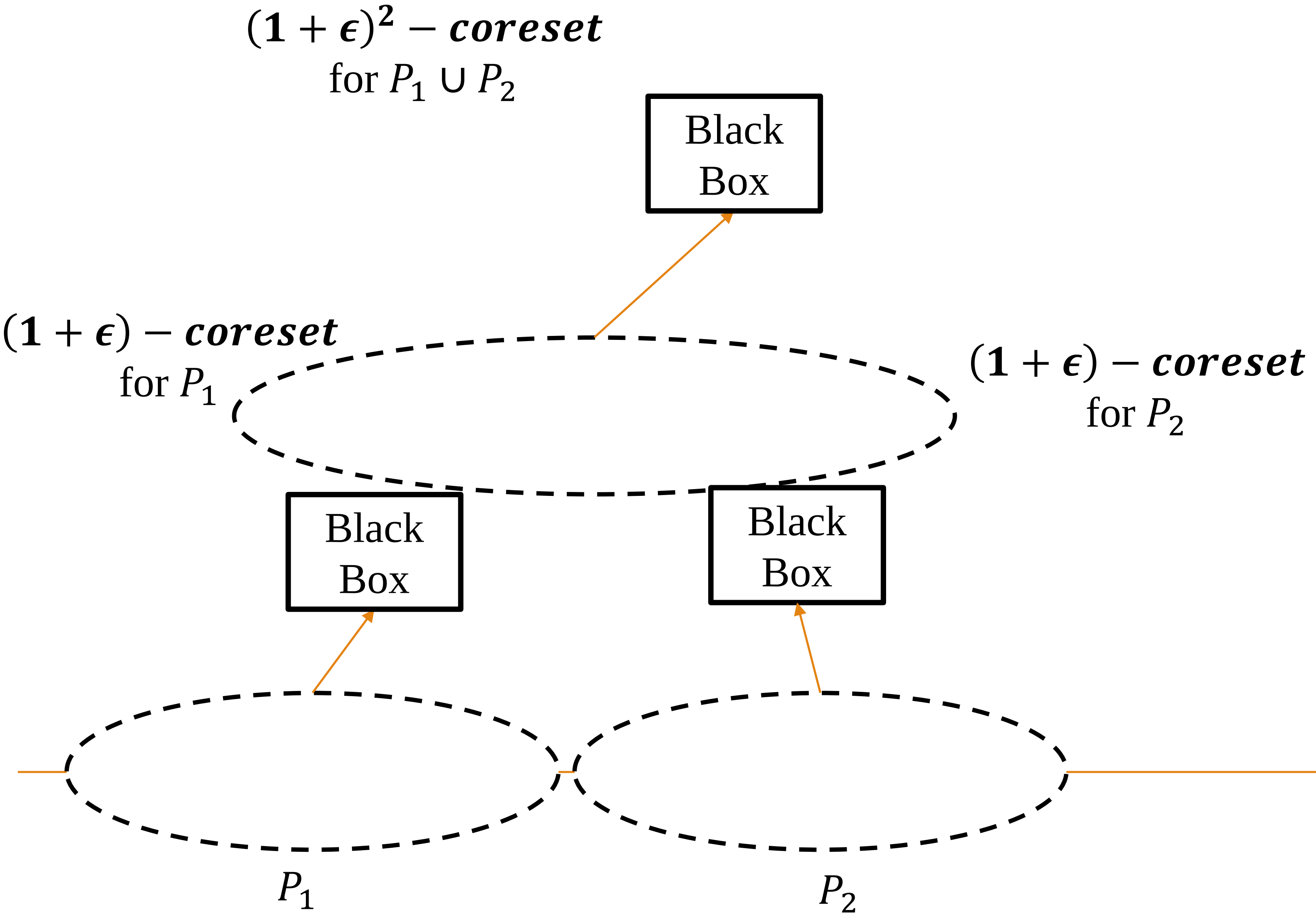
$|f(P_1 \cup P_2) - f(C_2)| \le \epsilon f(P_1 \cup P_2)$

$|f(C_2 \cup P_3) - f(C_3)| \le \epsilon f(C_2 \cup P_3)$

Need to prove that: $|f(P_1 \cup P_2 \cup P_3) - f(C_3)| \le \epsilon f(P_1 \cup P_2 \cup P_3)$

$|f(P_1 \cup P_2 \cup P_3) - f(C_3)| \le |f(P_1 \cup P_2) + f(P_3) - f(C_3)| = |f(P_1 \cup P_2) + f(C_2) - f(C_2) + f(P_3) - f(C_3)| \le |f(P_1 \cup P_2) - f(C_2)| + |f(C_2) + f(P_3) - f(C_3)| \le \epsilon f(P_1 \cup P_2) + |f(C_2 \cup P_3) - f(C_3)| \le \epsilon f(P_1 \cup P_2) + \epsilon f(C_2 \cup P_3) \le \epsilon\big(f(P_1 \cup P_2 \cup P_3) + f(C_2)\big) \le \epsilon\big(f(P_1 \cup P_2 \cup P_3) + 2f(P_1 \cup P_2 \cup P_3)\big) \le O(\epsilon) f(P_1 \cup P_2 \cup P_3)$

Streaming



$(\mathbf{1}+\boldsymbol{\epsilon})^{\mathbf{2}} - \boldsymbol{coreset}$
for $P_1 \cup P_2$
Black Box
$(\mathbf{1}+\boldsymbol{\epsilon}) - \boldsymbol{coreset}$
for P_1
$(\mathbf{1}+\boldsymbol{\epsilon}) - \boldsymbol{coreset}$
for P_2
Black Box
Black Box
P_1
P_2