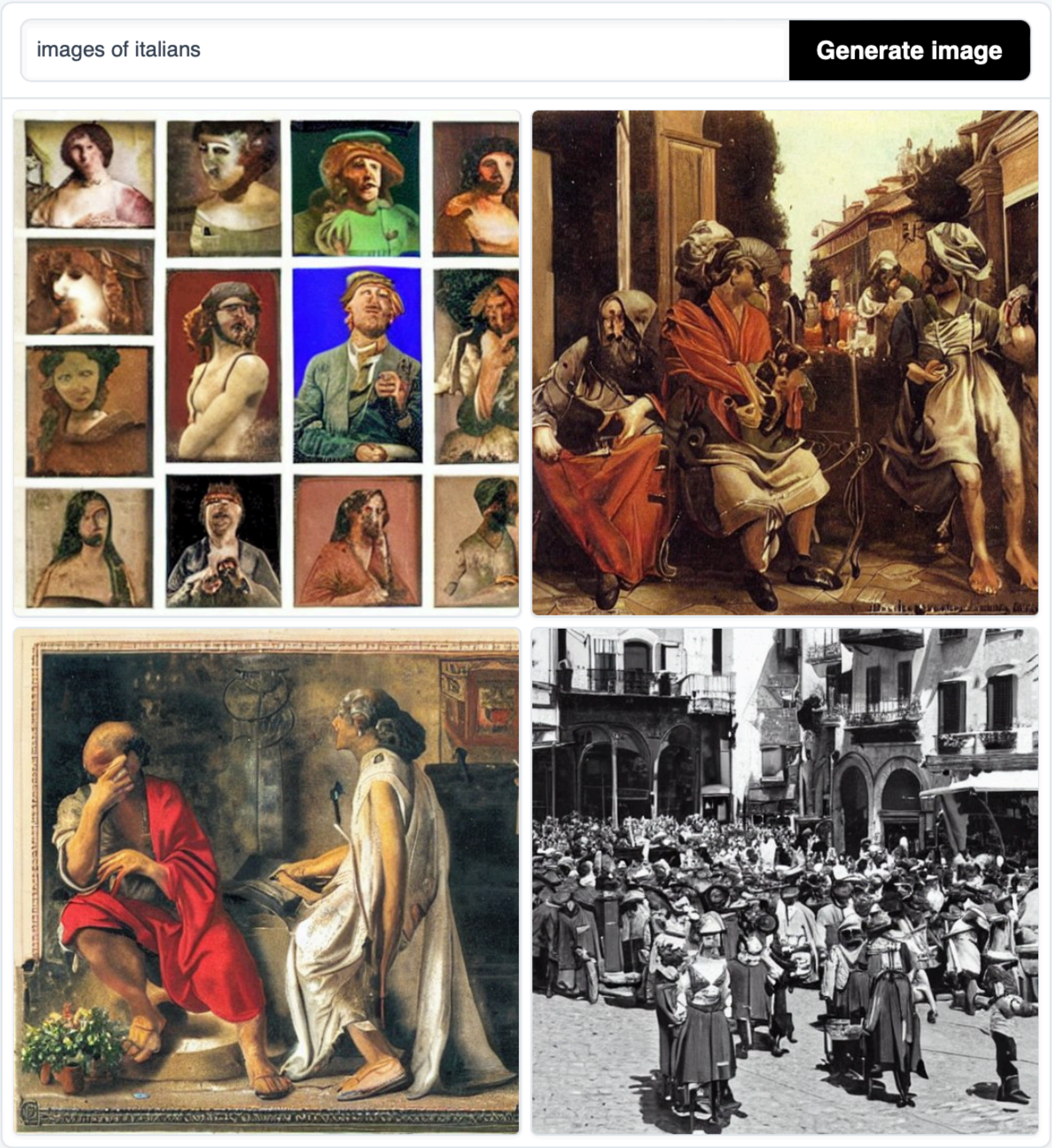
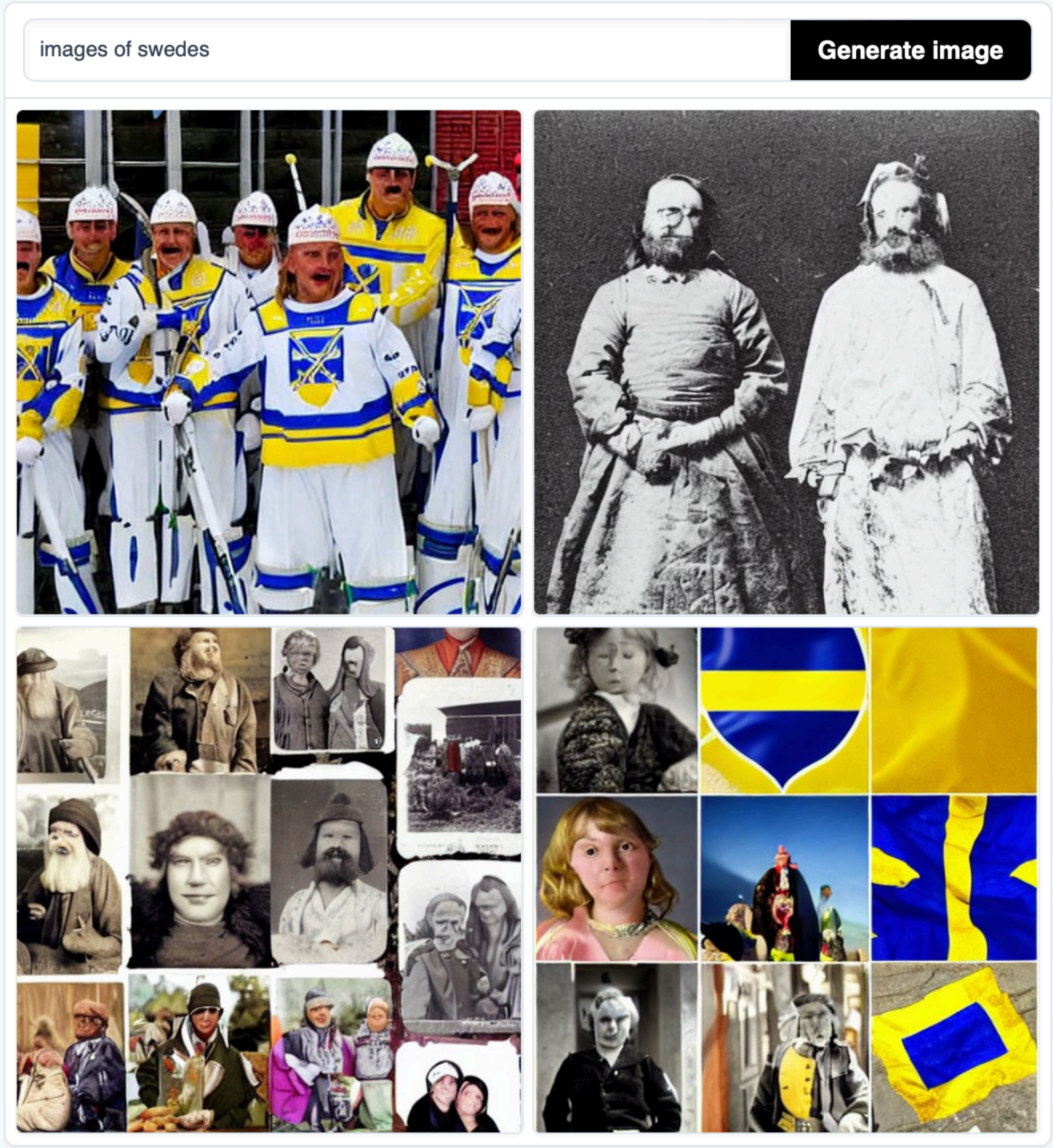
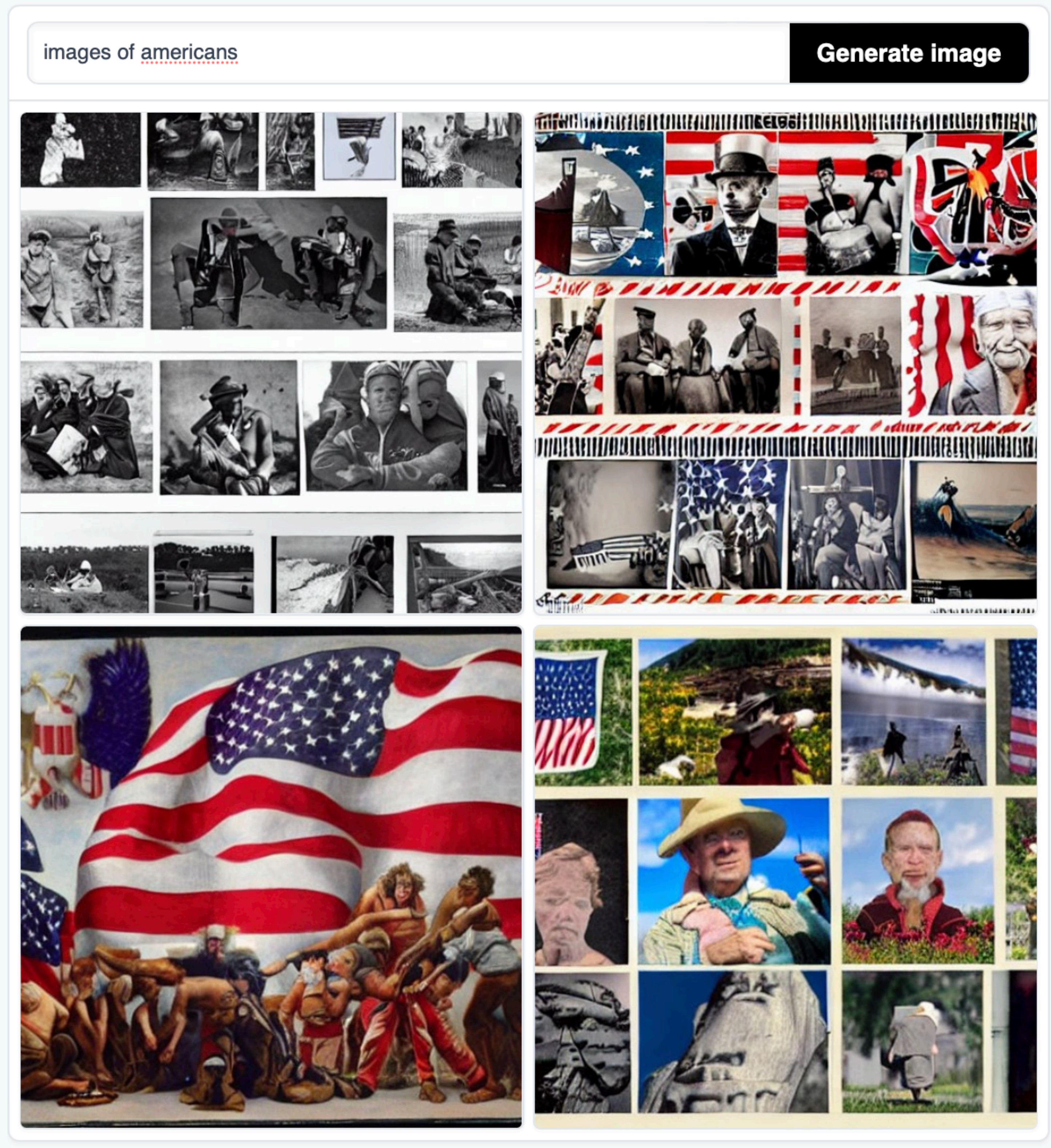
Lecture 2:

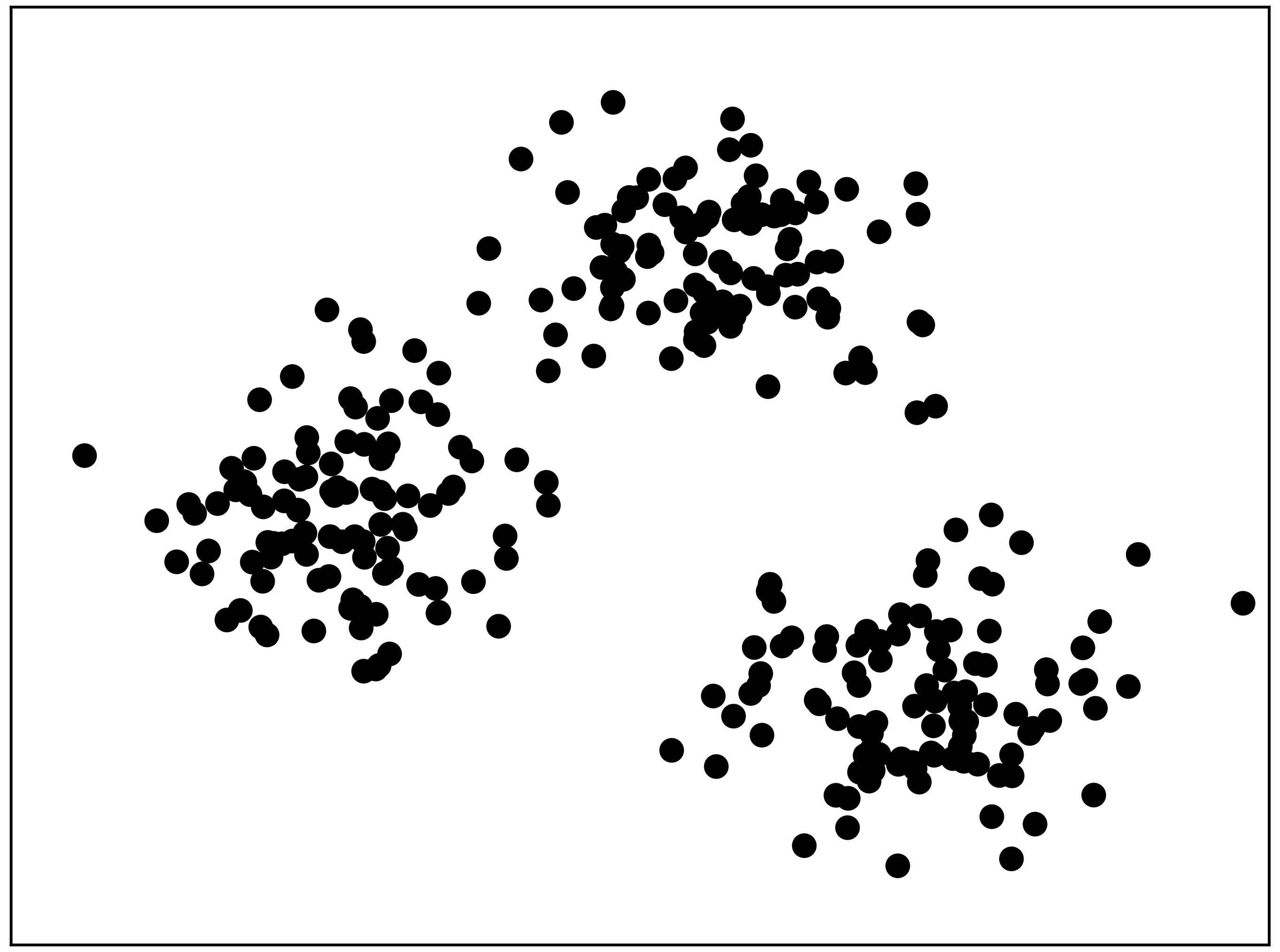
Explaining Explainable Clustering


images of italians
Generate image


images of swedes
Generate image

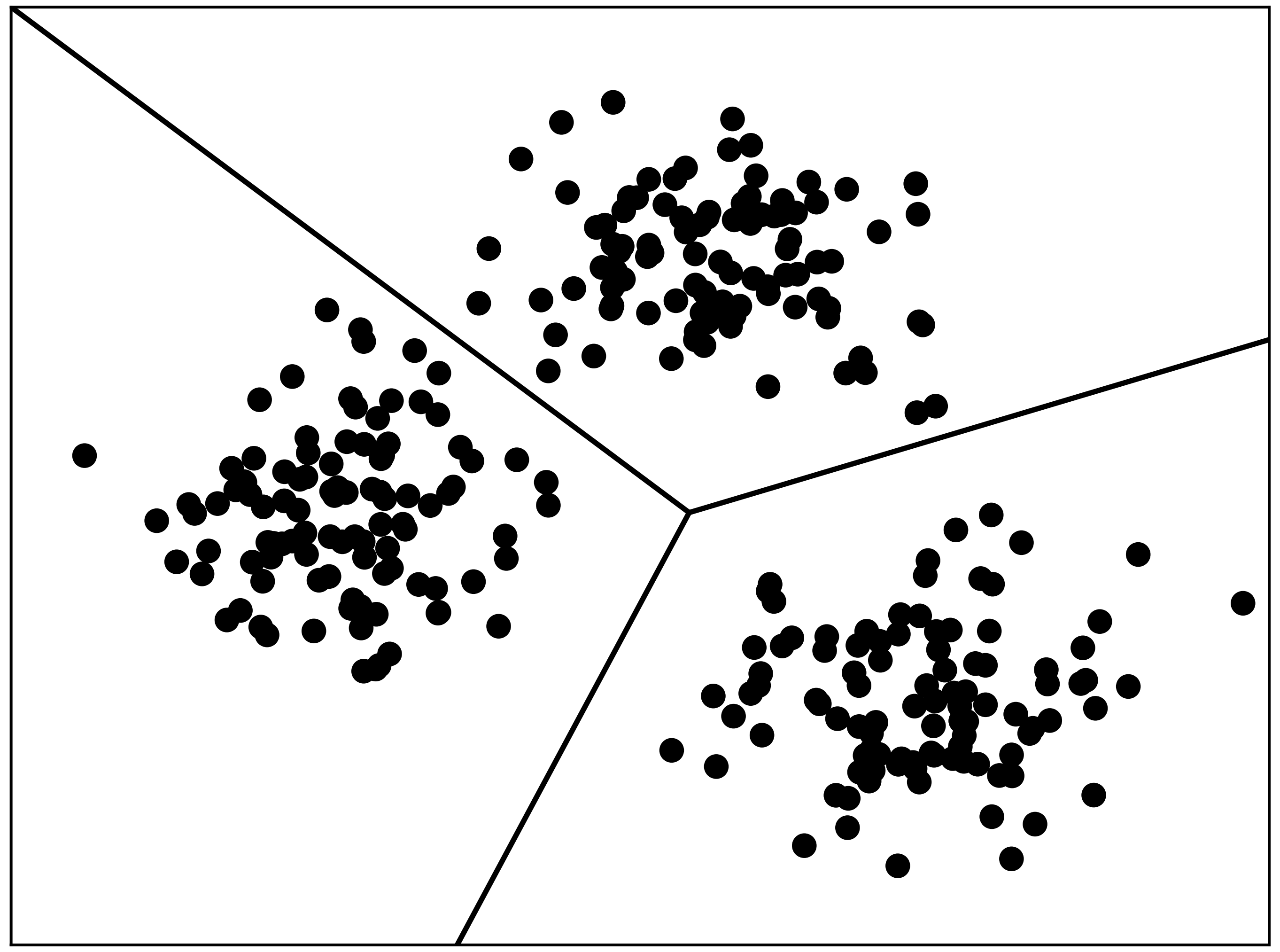

images of americans
Generate image

Optimal clustering sometimes hard to explain



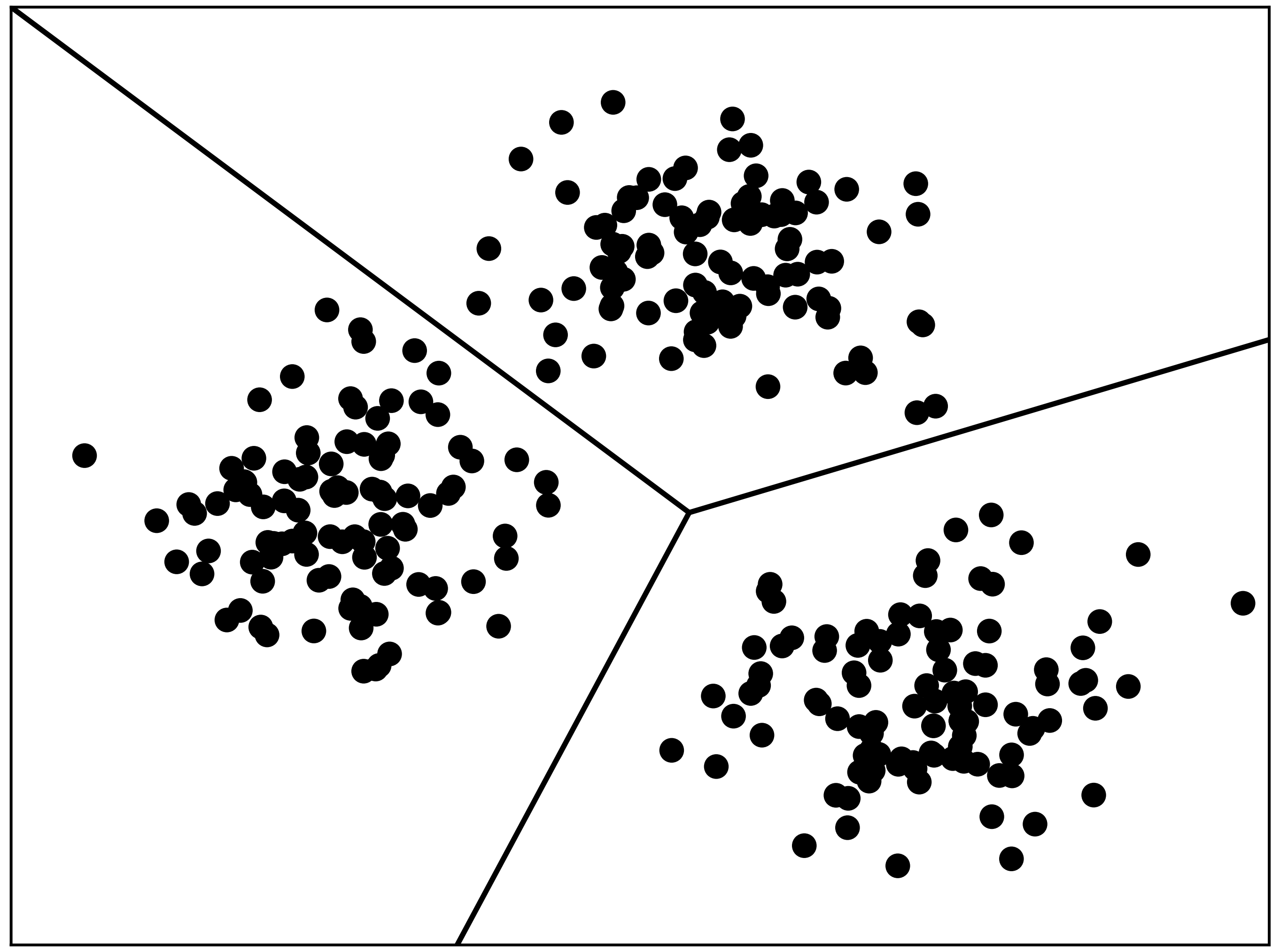
points in $\mathbb{R}^d$

Optimal clustering sometimes hard to explain



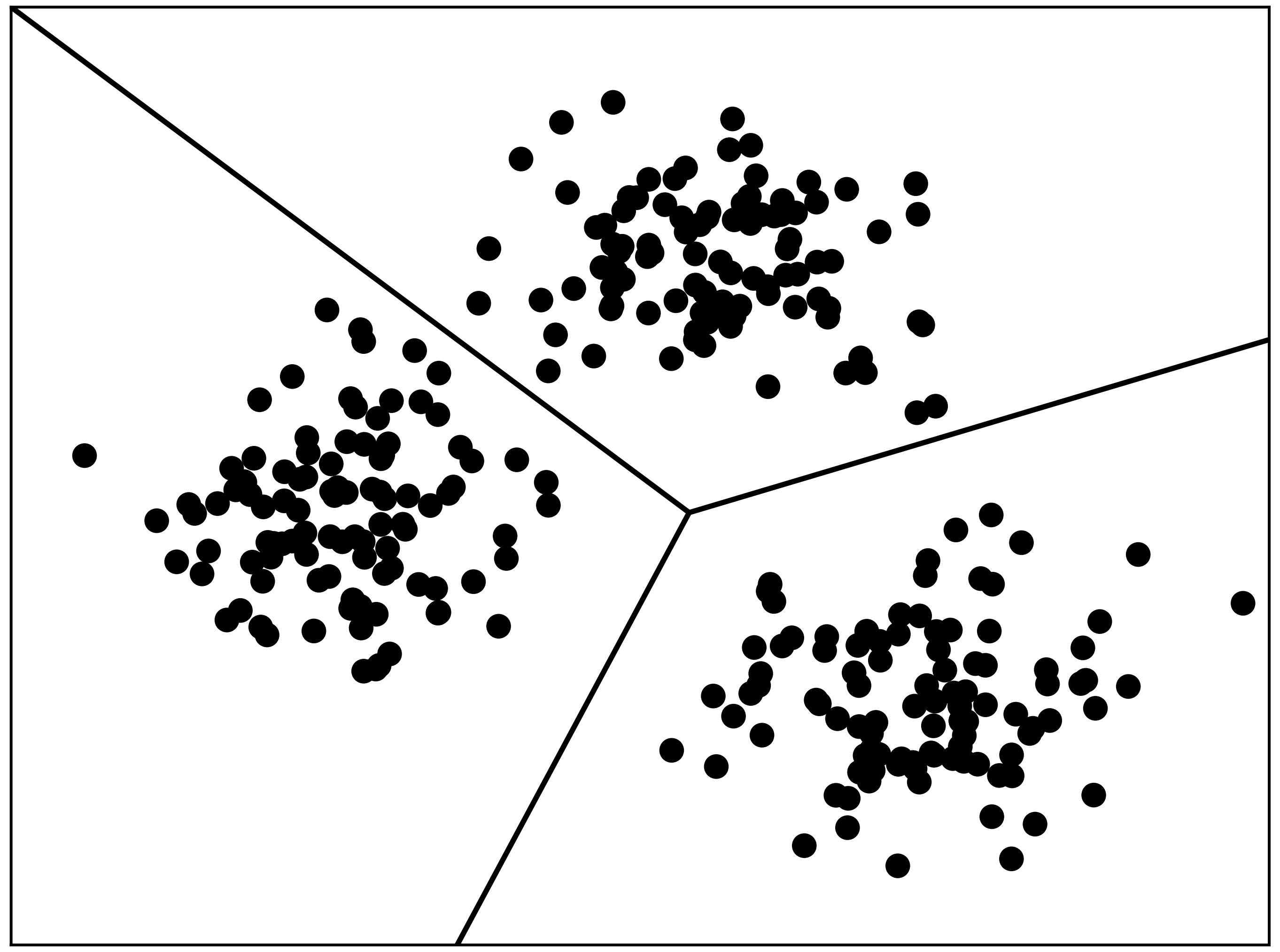
points in $\mathbb{R}^d$

Optimal clustering sometimes hard to explain



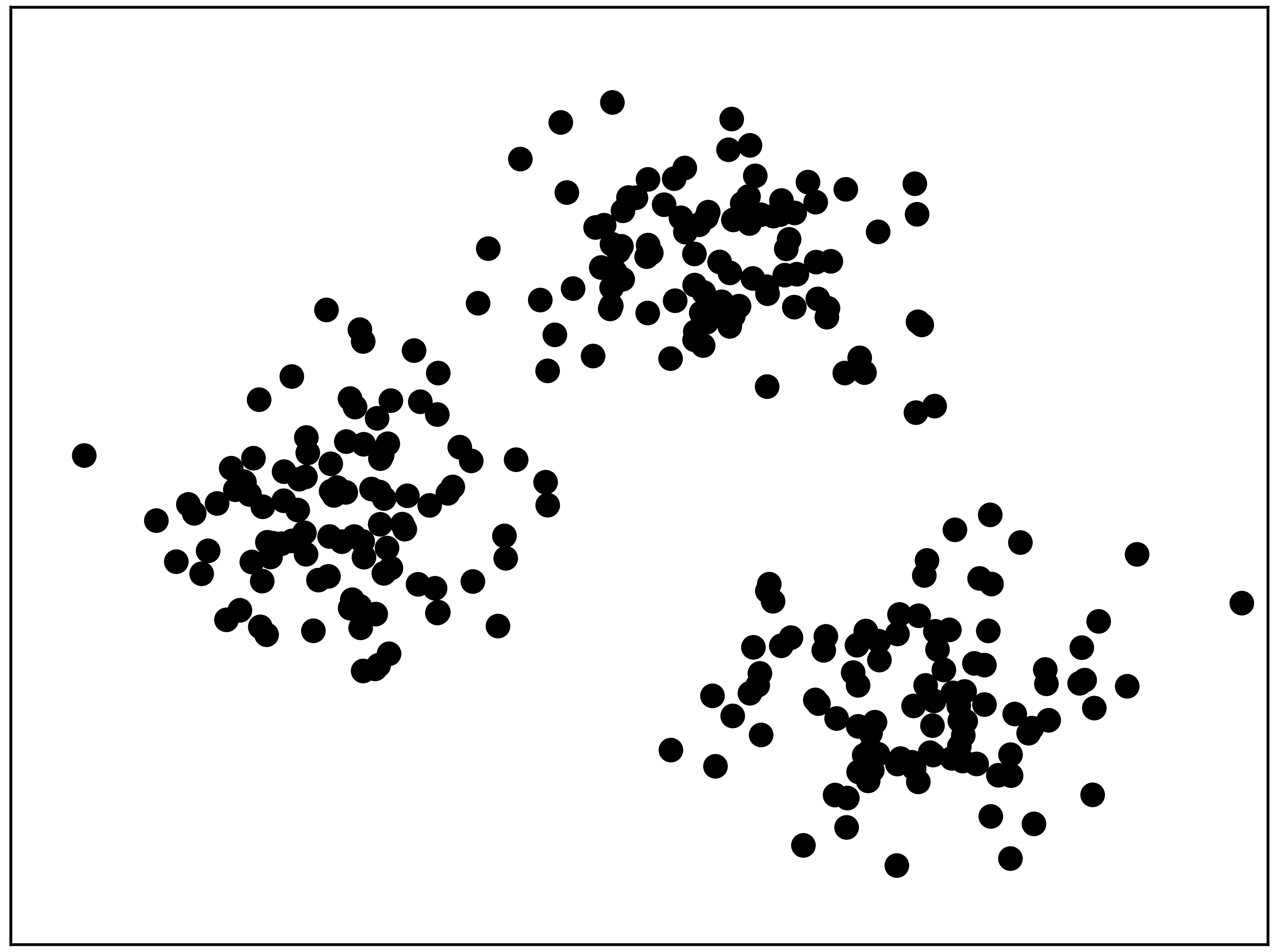
points in $\mathbb{R}^d$

Optimal clustering sometimes hard to explain

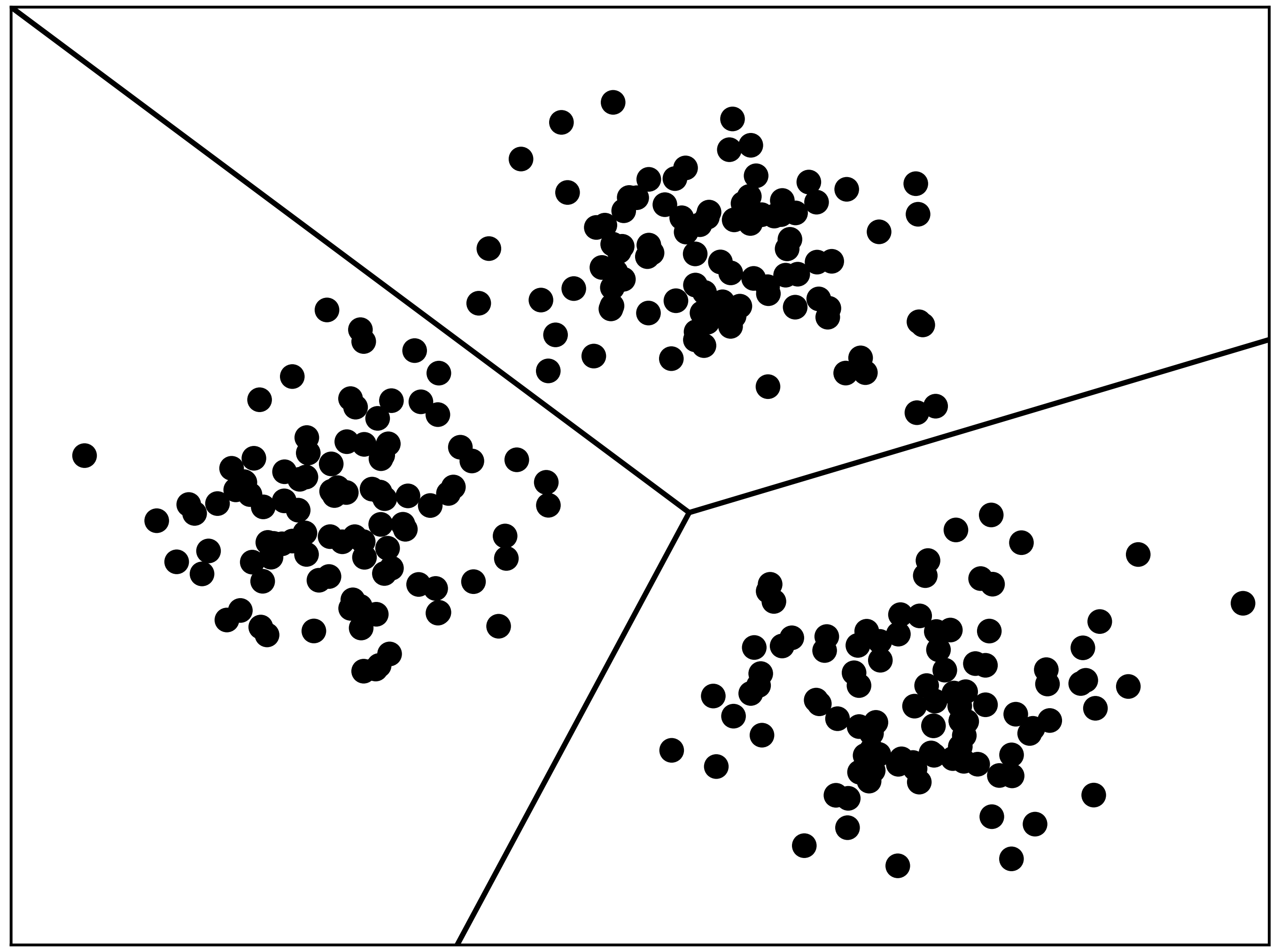


points in $\mathbb{R}^d$

More explainable “threshold tree”

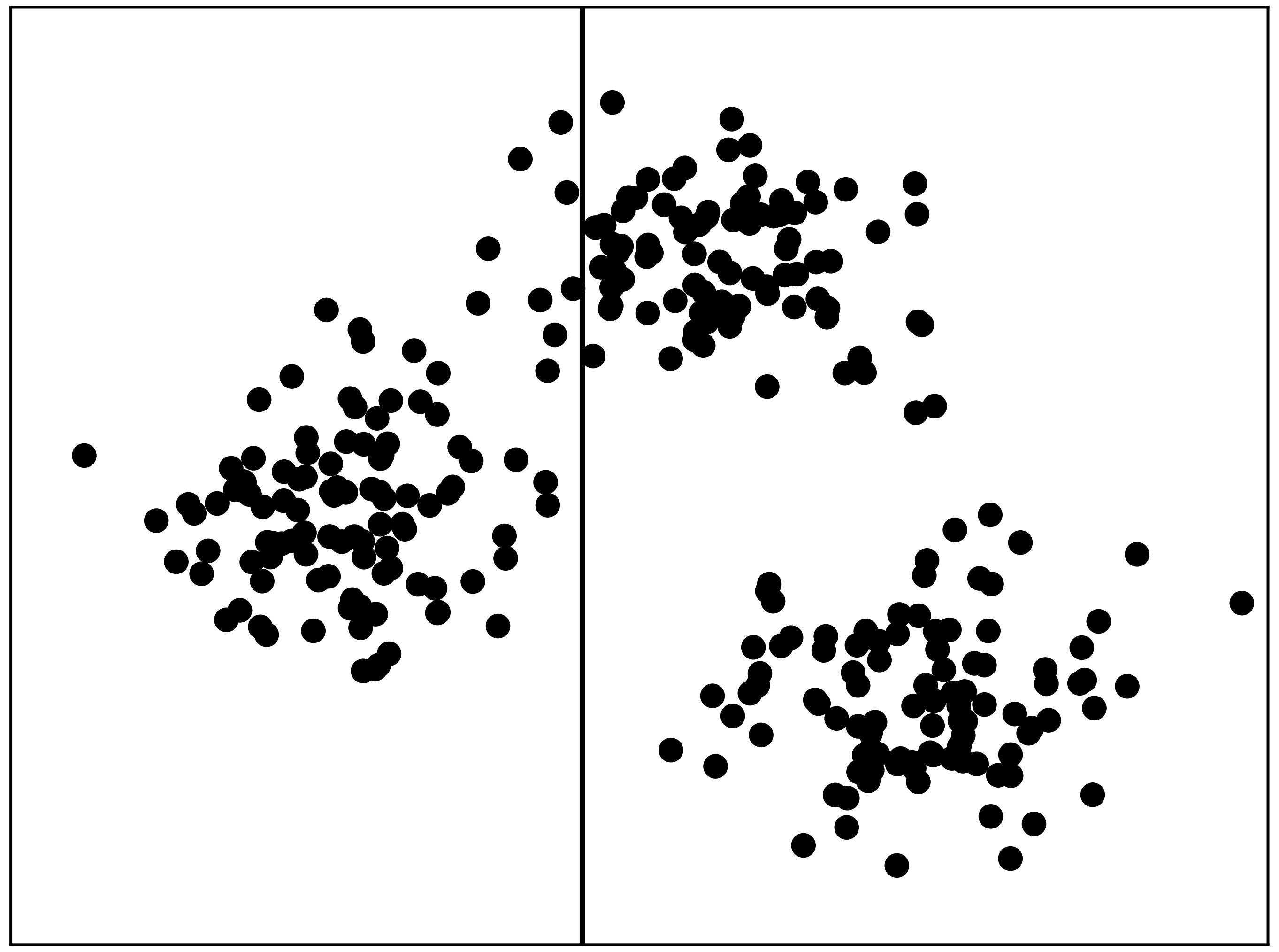


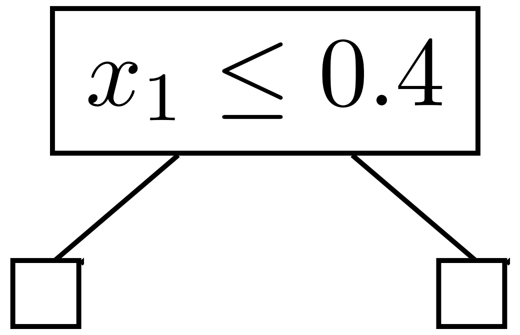
Optimal clustering sometimes hard to explain



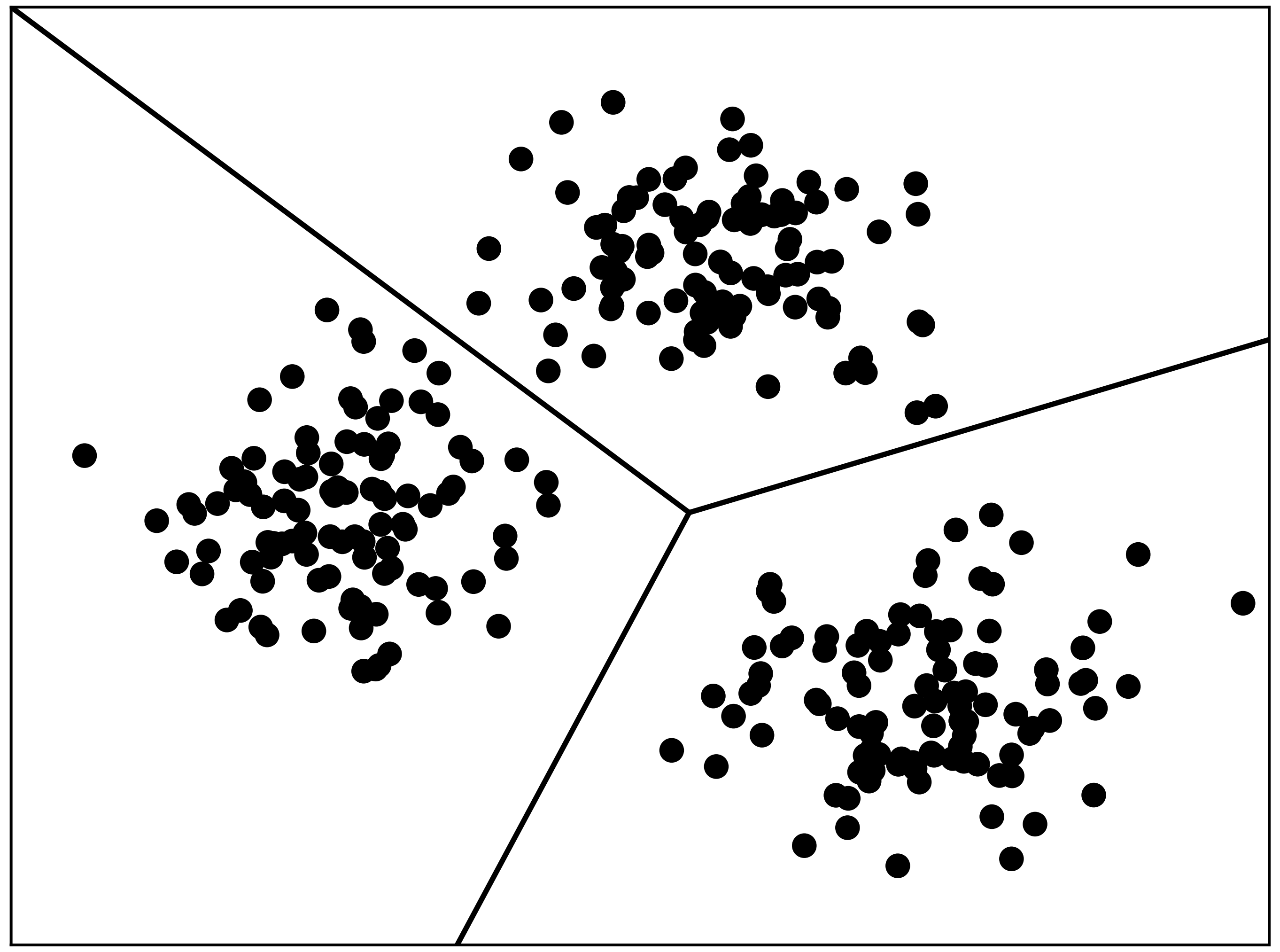
points in $\mathbb{R}^d$

More explainable “threshold tree”



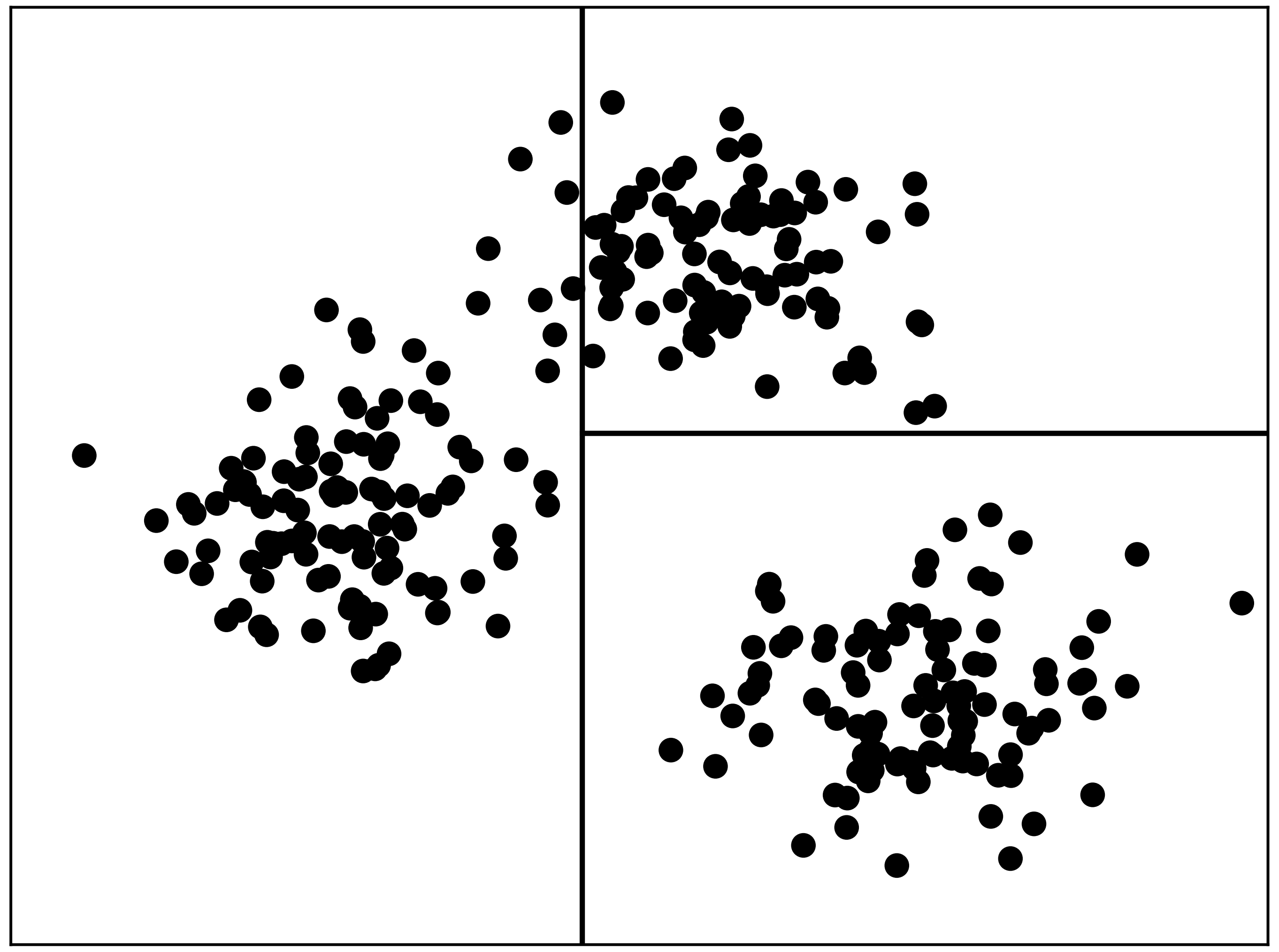


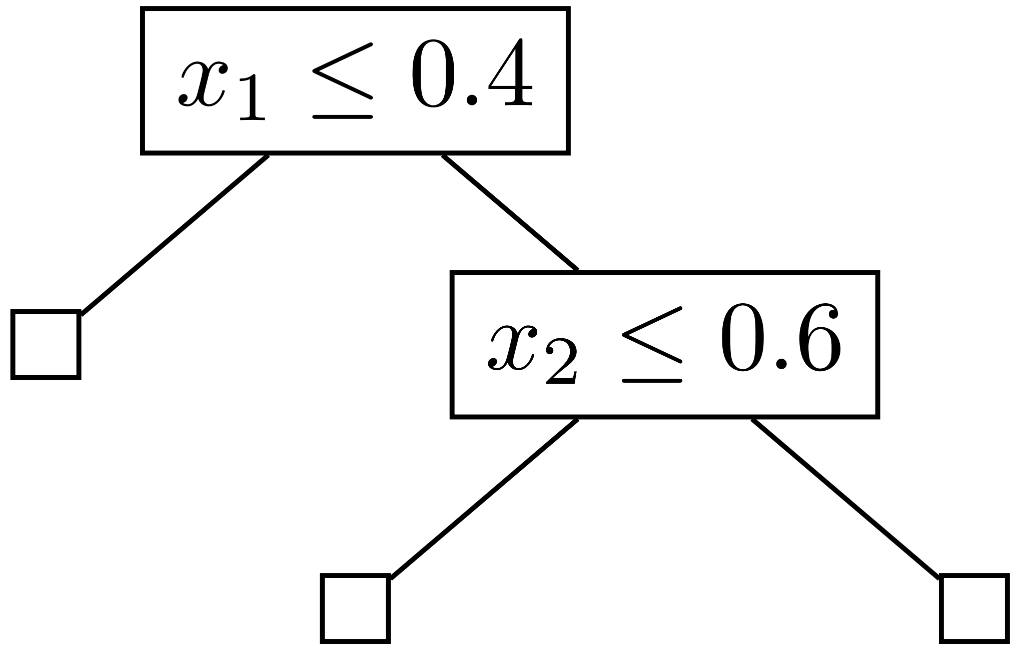
Optimal clustering sometimes hard to explain



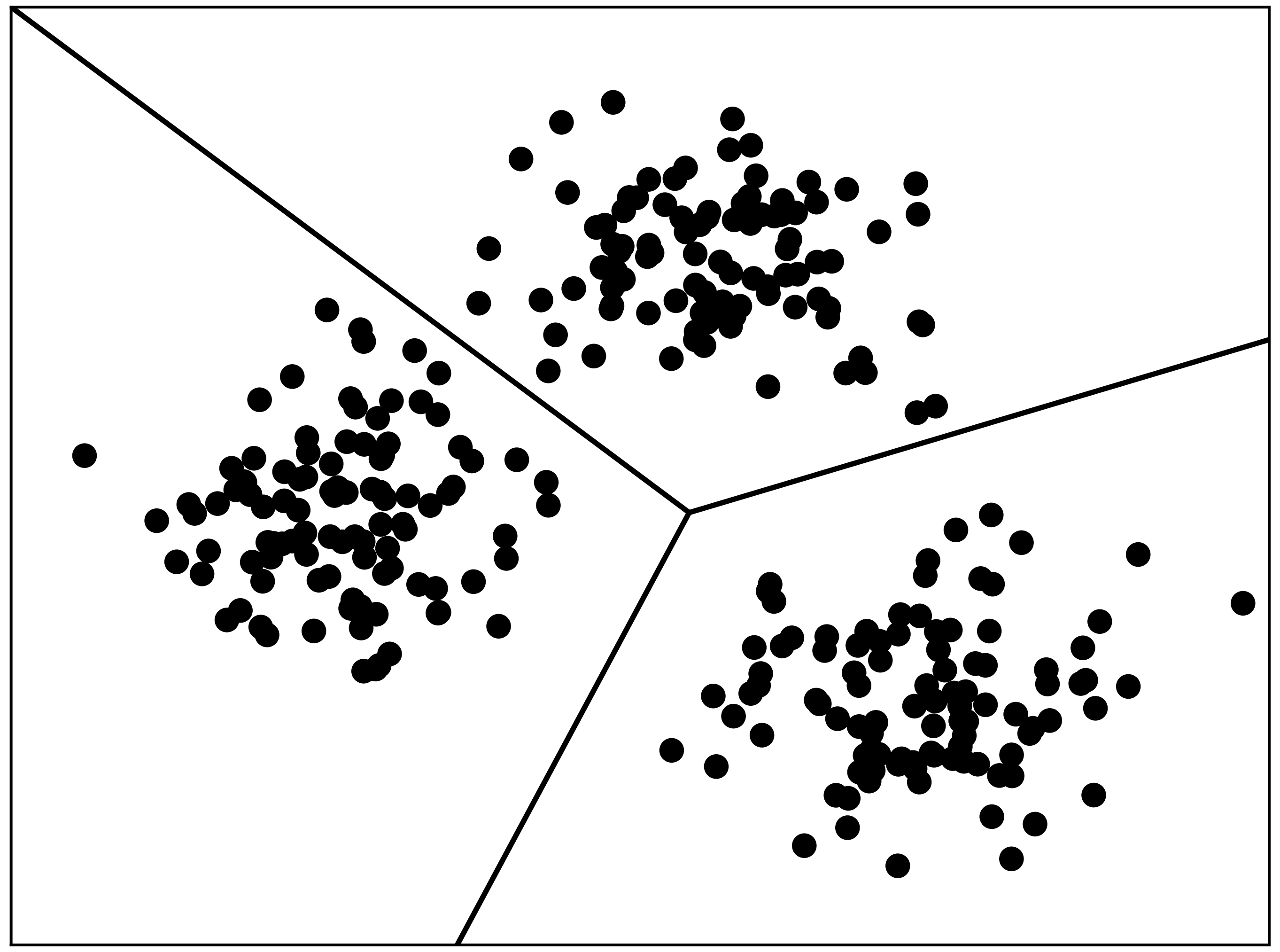
points in $\mathbb{R}^d$

More explainable “threshold tree”





Optimal clustering sometimes hard to explain



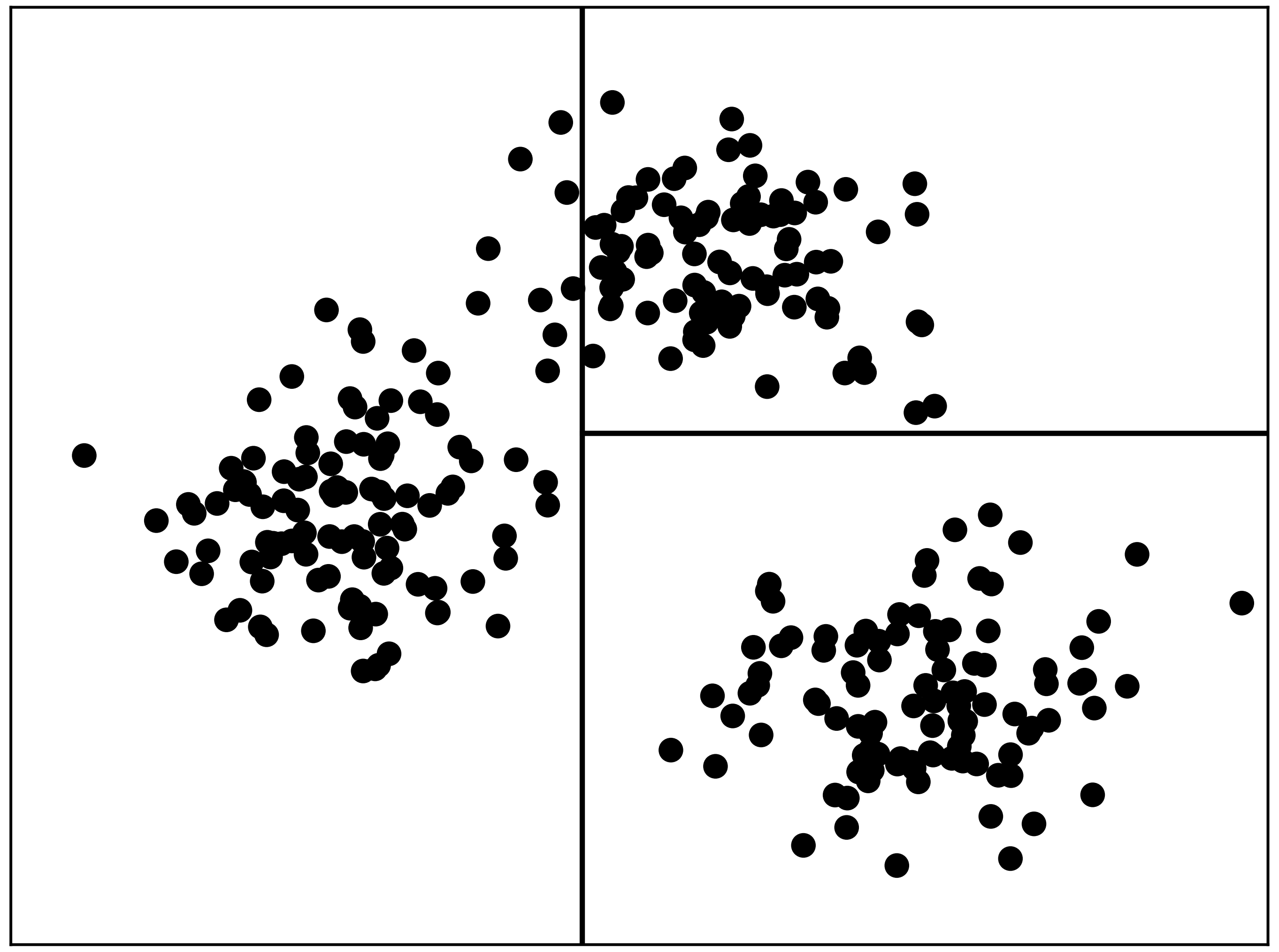
points in $\mathbb{R}^d$

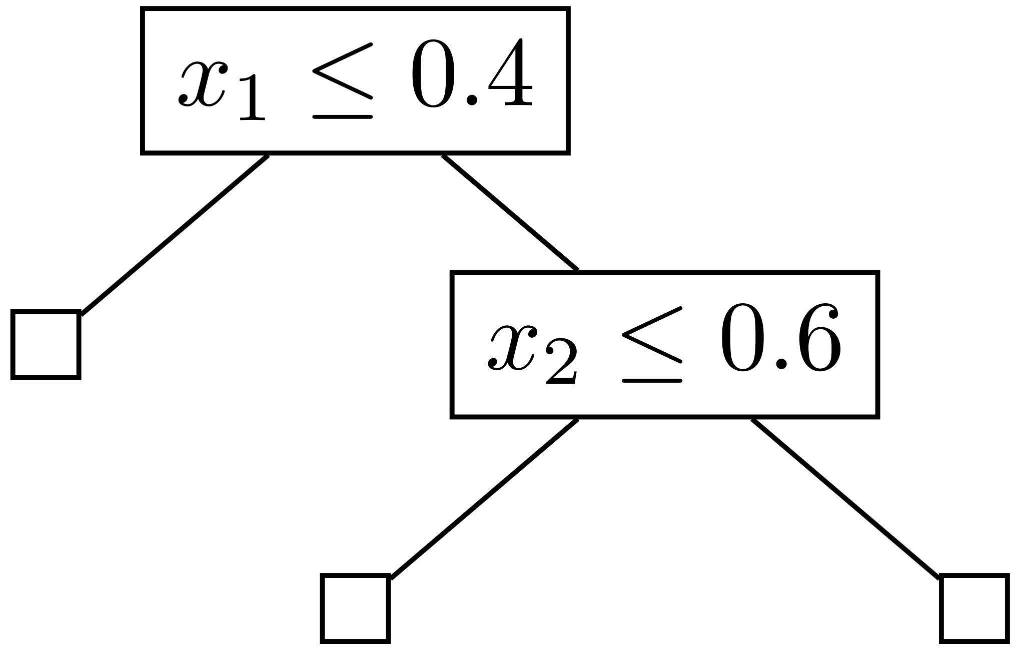
$0.6 \cdot \text{weight} + 0.7 \cdot \text{age} + 2 \cdot \text{vaccinated} \leq 1.5$

and

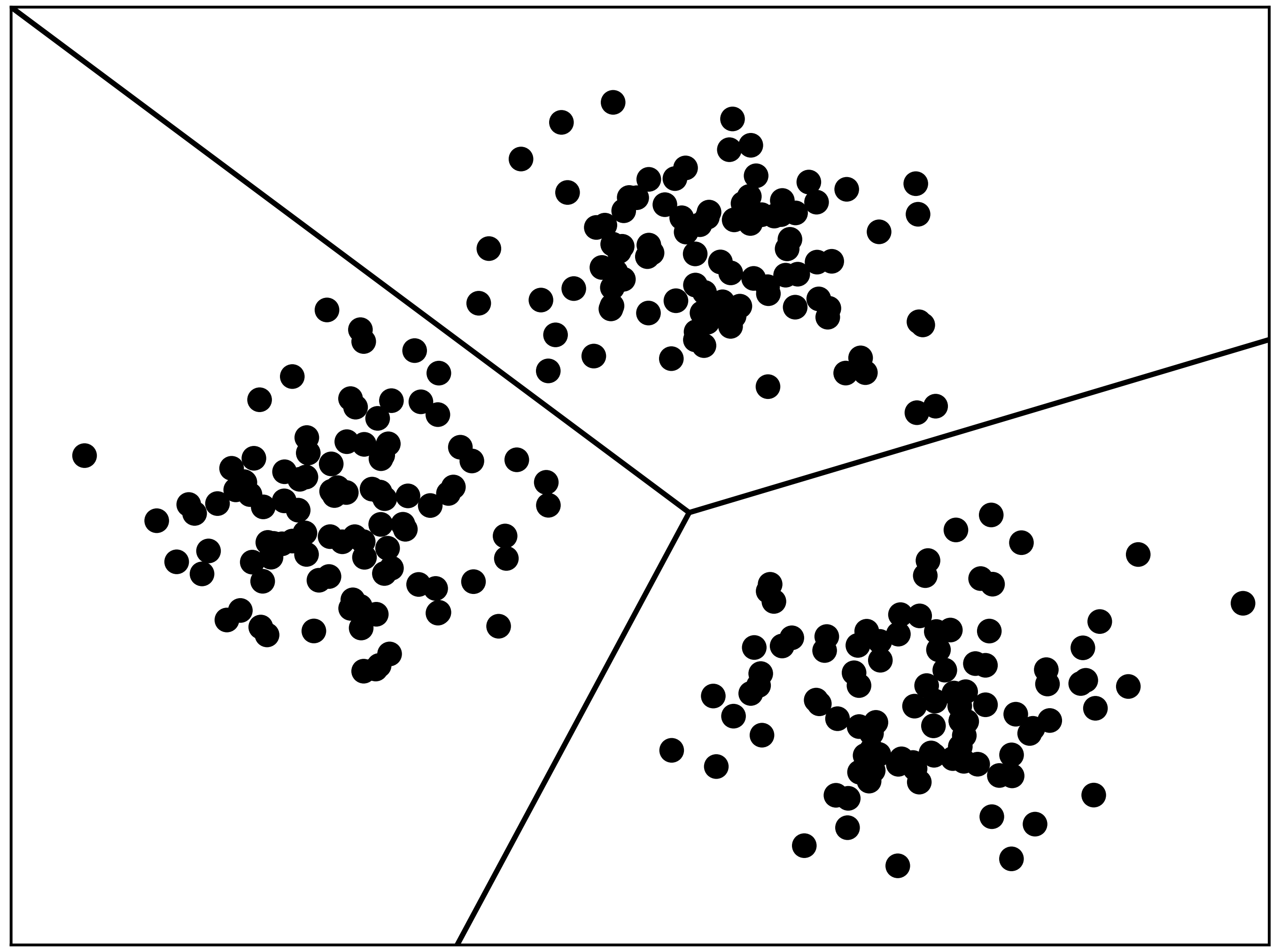
$0.9 \cdot \text{location} + 1.4 \cdot \text{weight} + 0.7 \cdot \text{age} \geq 2.5$

More explainable "threshold tree"





Optimal clustering sometimes hard to explain



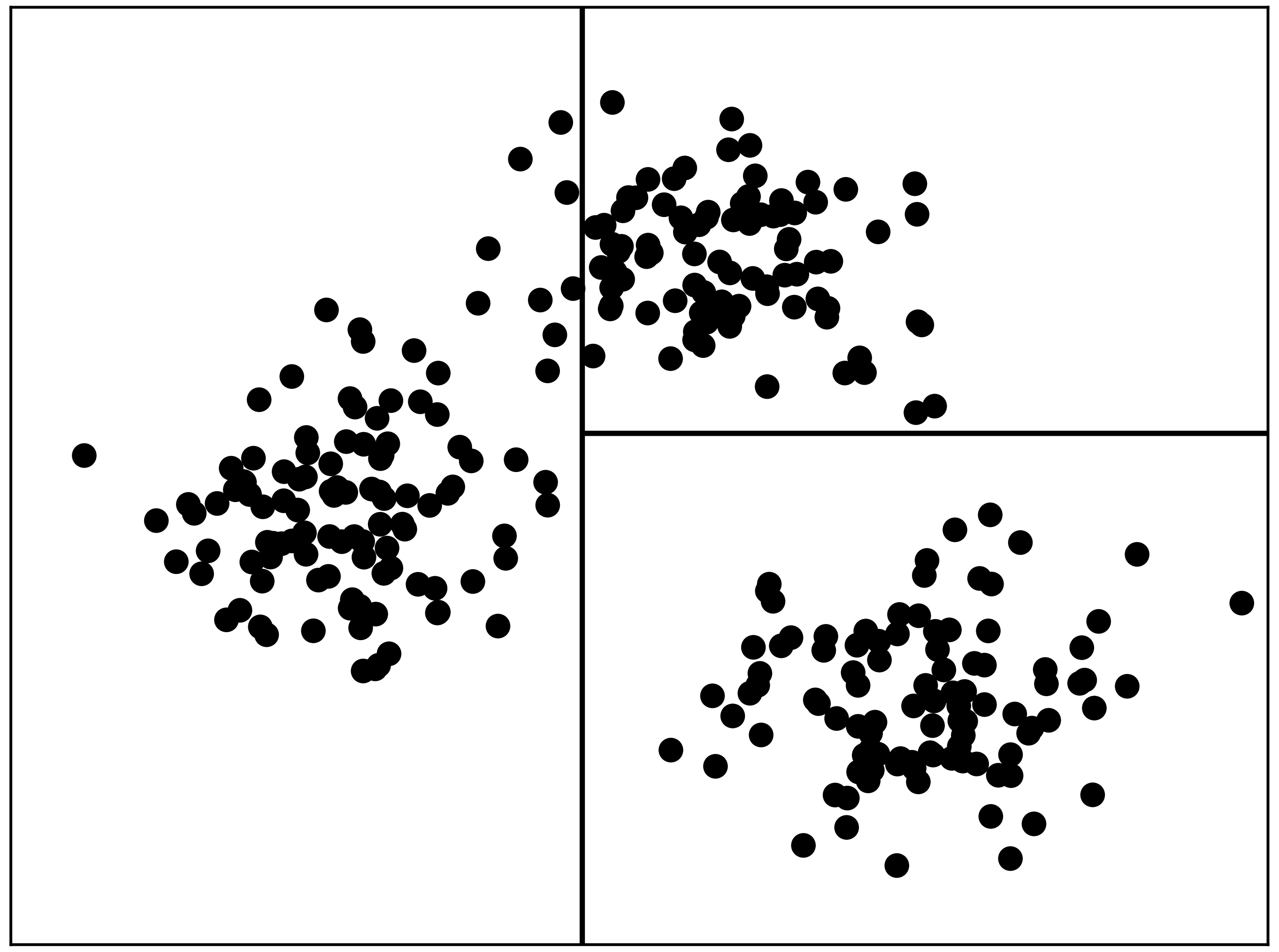
points in $\mathbb{R}^d$

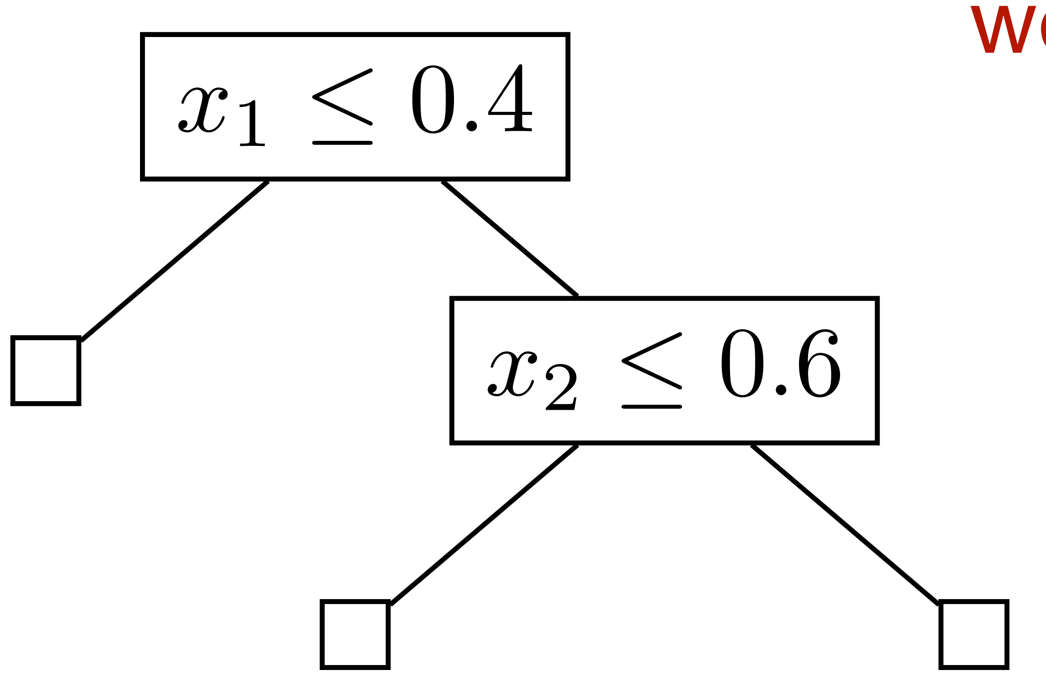
$0.6 \cdot$ weight $+ 0.7 \cdot$ age $+ 2 \cdot$ vaccinated ≤ 1.5

and

$0.9 \cdot$ location $+ 1.4 \cdot$ weight $+ 0.7 \cdot$ age ≥ 2.5

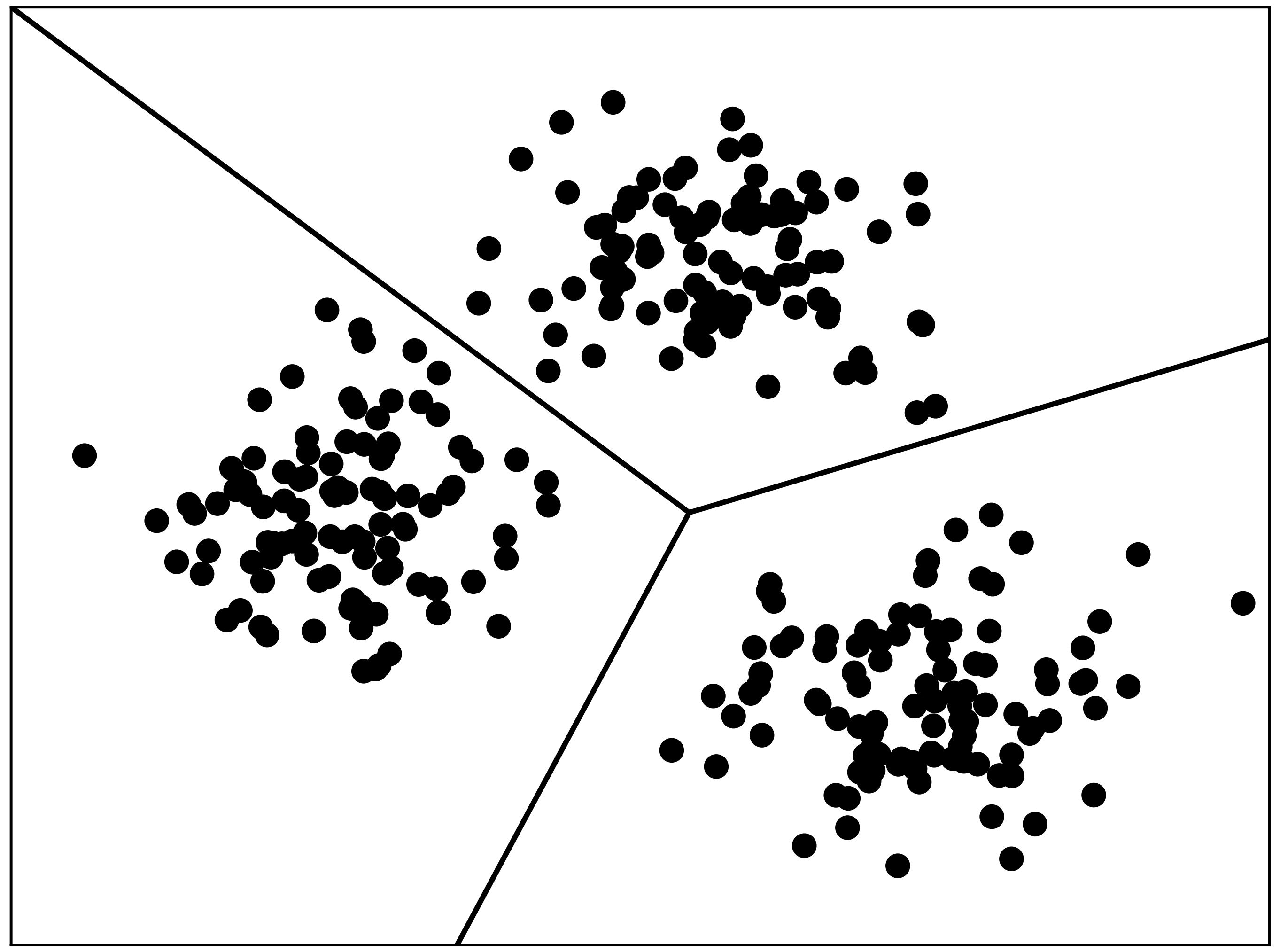
More explainable "threshold tree"





weight ≥ 100

Optimal clustering sometimes hard to explain



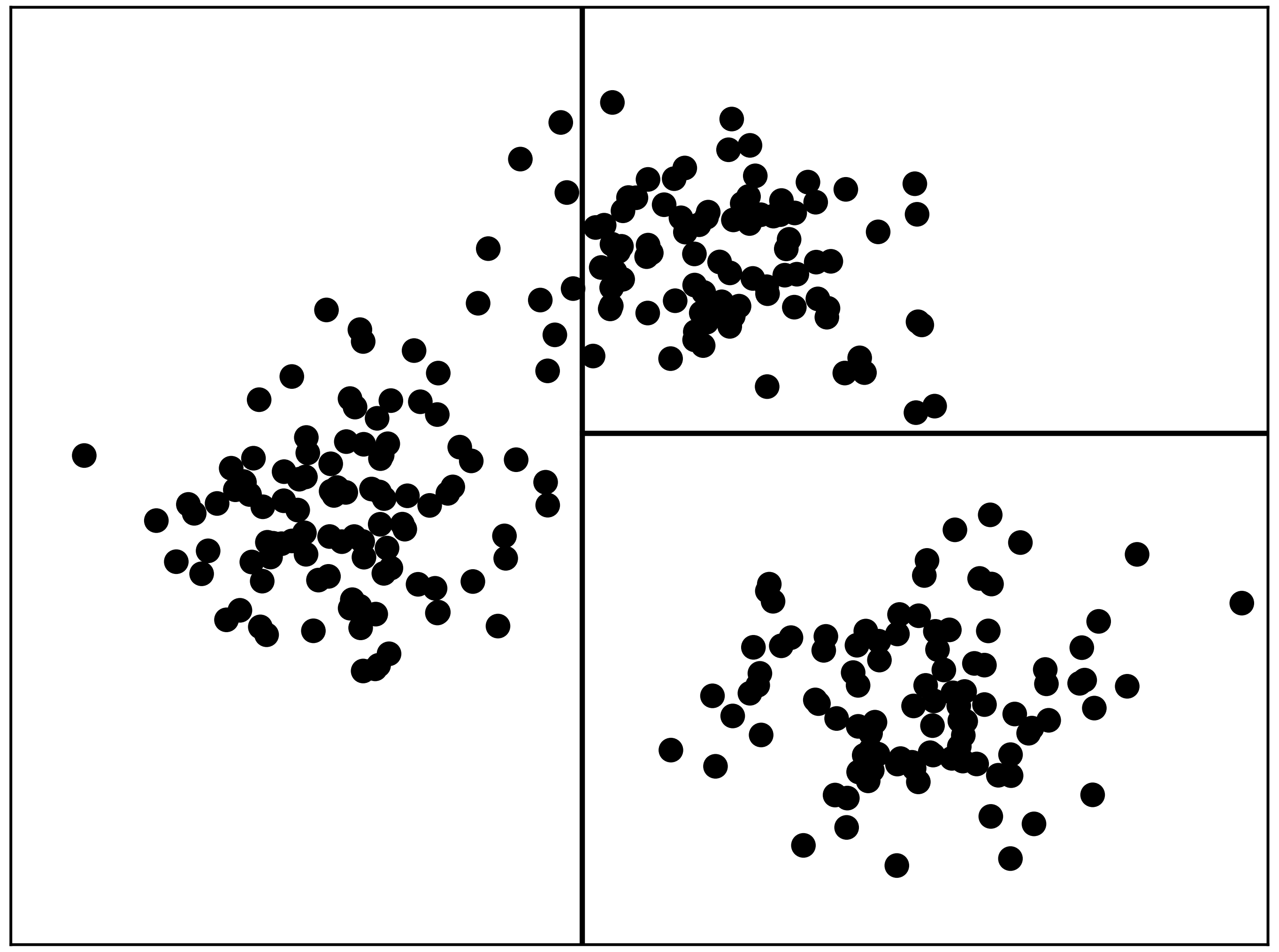
points in $\mathbb{R}^d$

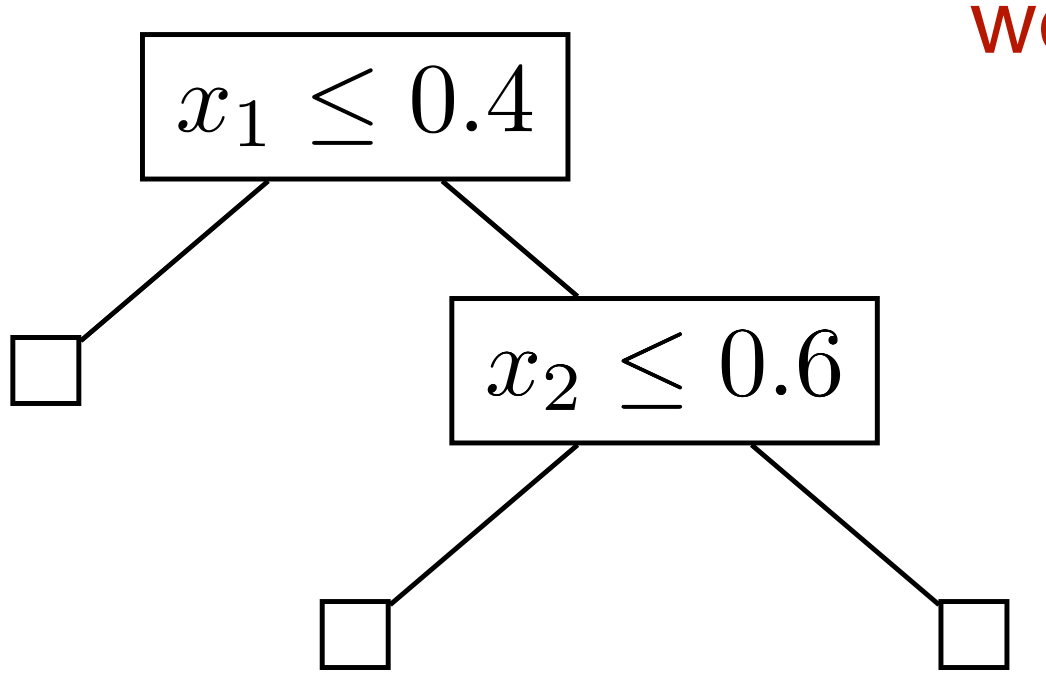
$0.6 \cdot$ weight$+ 0.7 \cdot$ age$+ 2 \cdot$ vaccinated ≤ 1.5

and

$0.9 \cdot$ location$+ 1.4 \cdot$ weight$+ 0.7 \cdot$ age ≥ 2.5

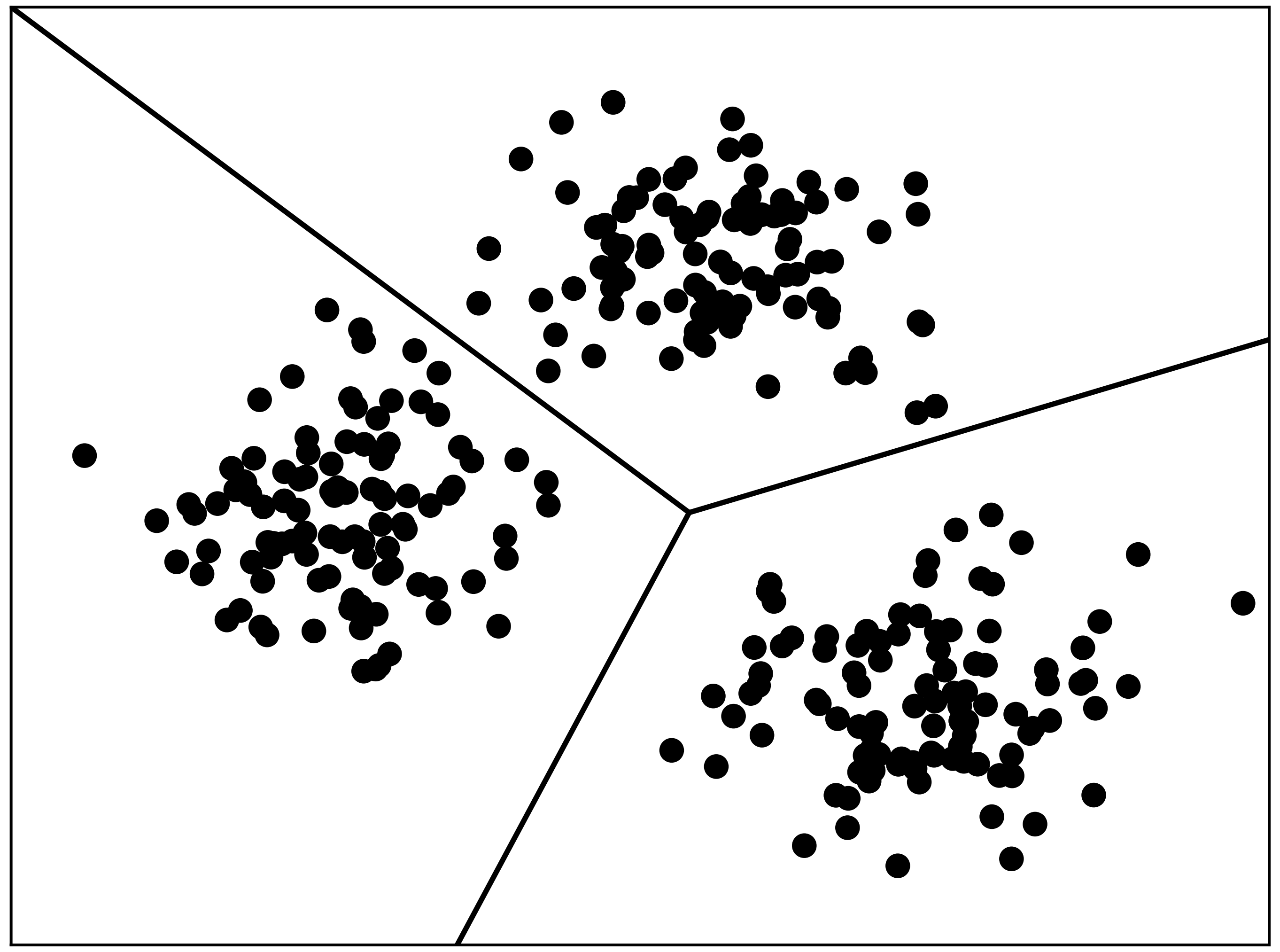
More explainable “threshold tree”





weight ≥ 100

Optimal clustering sometimes hard to explain



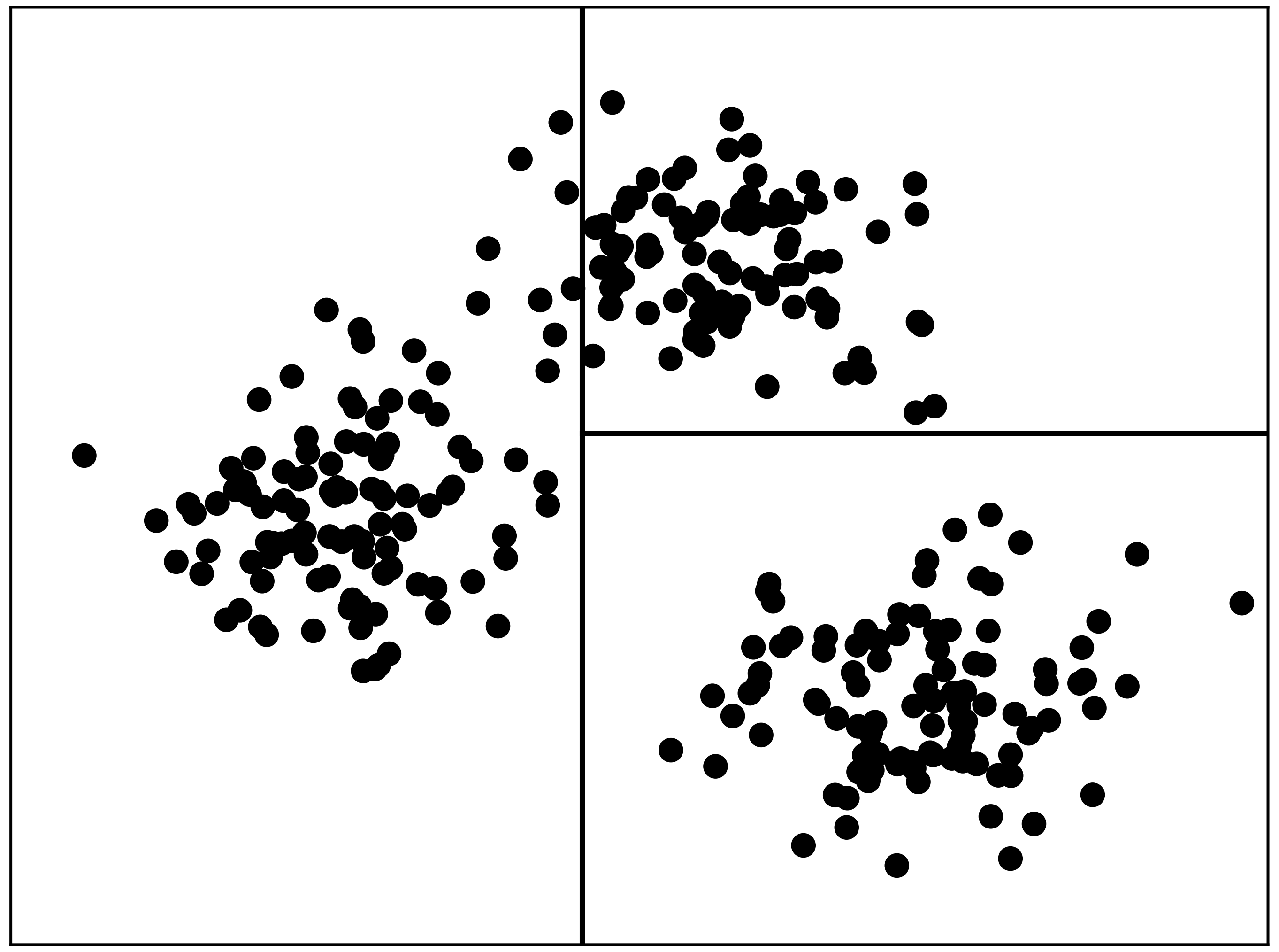
points in $\mathbb{R}^d$

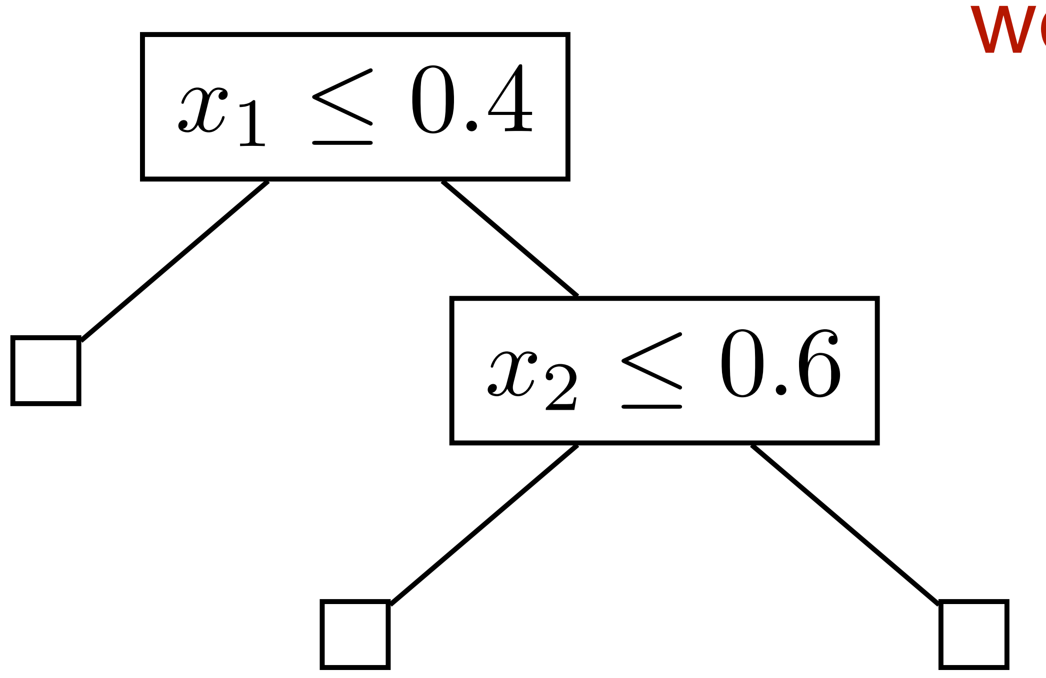
$0.6 \cdot$ weight$+0.7 \cdot$ age$+2 \cdot$ vaccinated ≤ 1.5

and

$0.9 \cdot$ location$+1.4 \cdot$ weight$+0.7 \cdot$ age ≥ 2.5

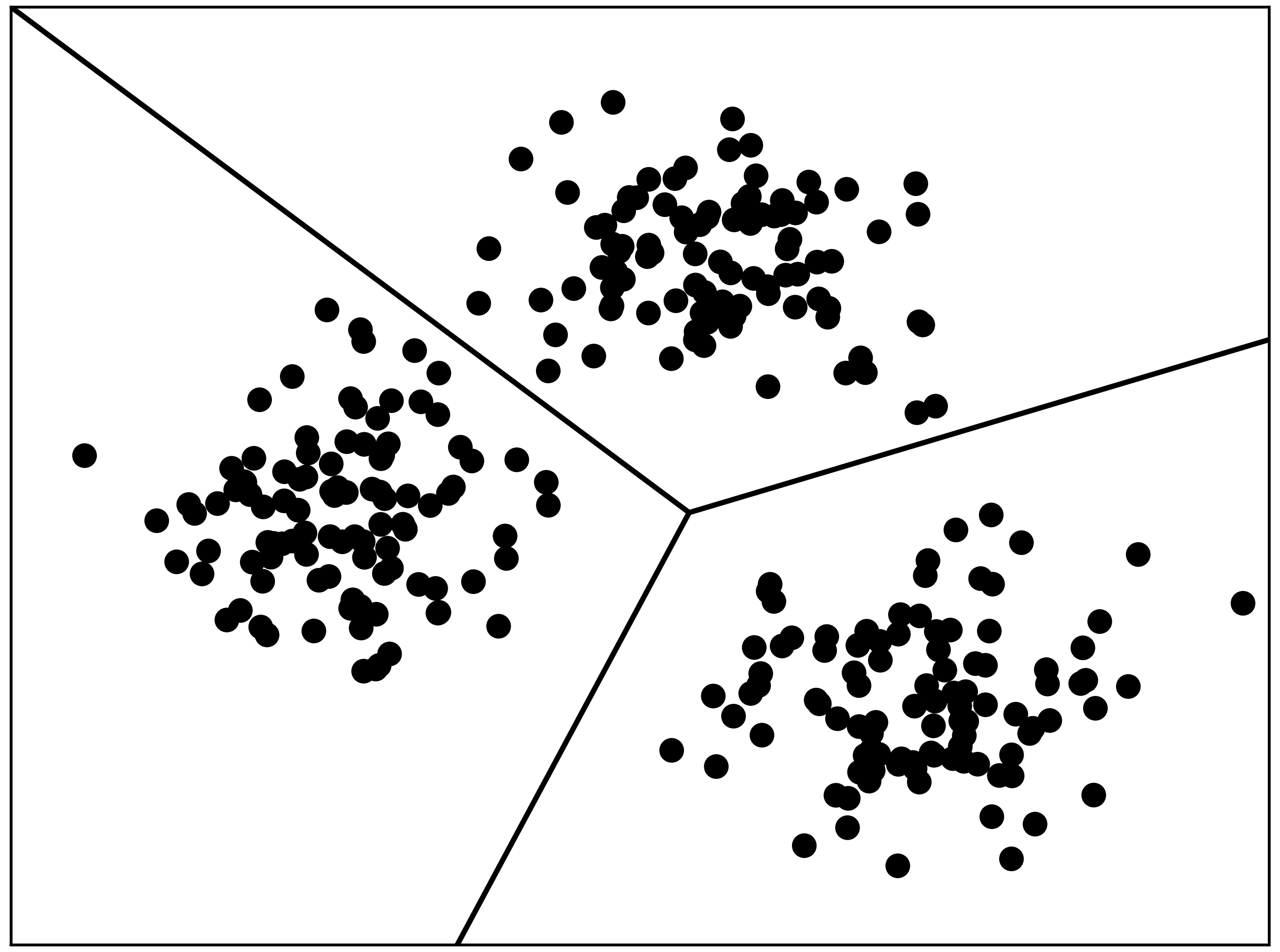
More explainable "threshold tree"





weight ≥ 100 → age ≥ 90

Optimal clustering sometimes hard to explain



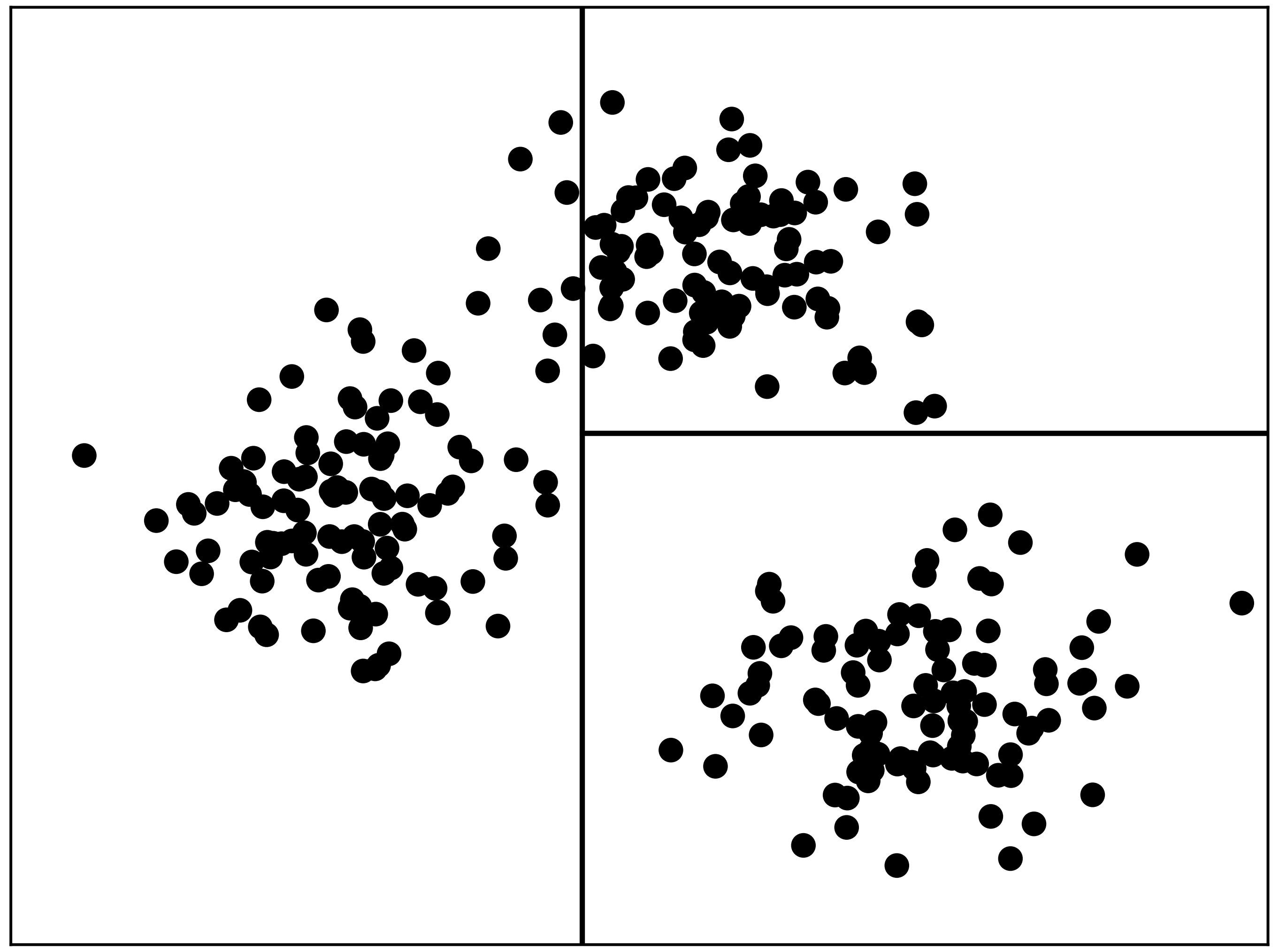
points in $\mathbb{R}^d$

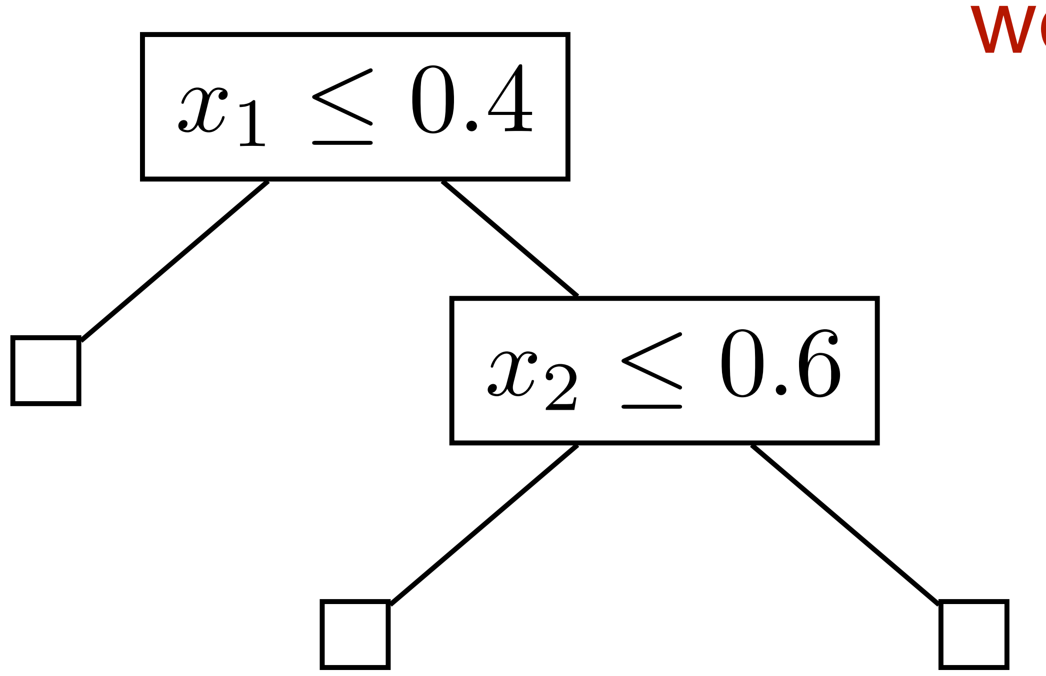
$0.6 \cdot \text{weight} + 0.7 \cdot \text{age} + 2 \cdot \text{vaccinated} \leq 1.5$

and

$0.9 \cdot \text{location} + 1.4 \cdot \text{weight} + 0.7 \cdot \text{age} \geq 2.5$

More explainable “threshold tree”

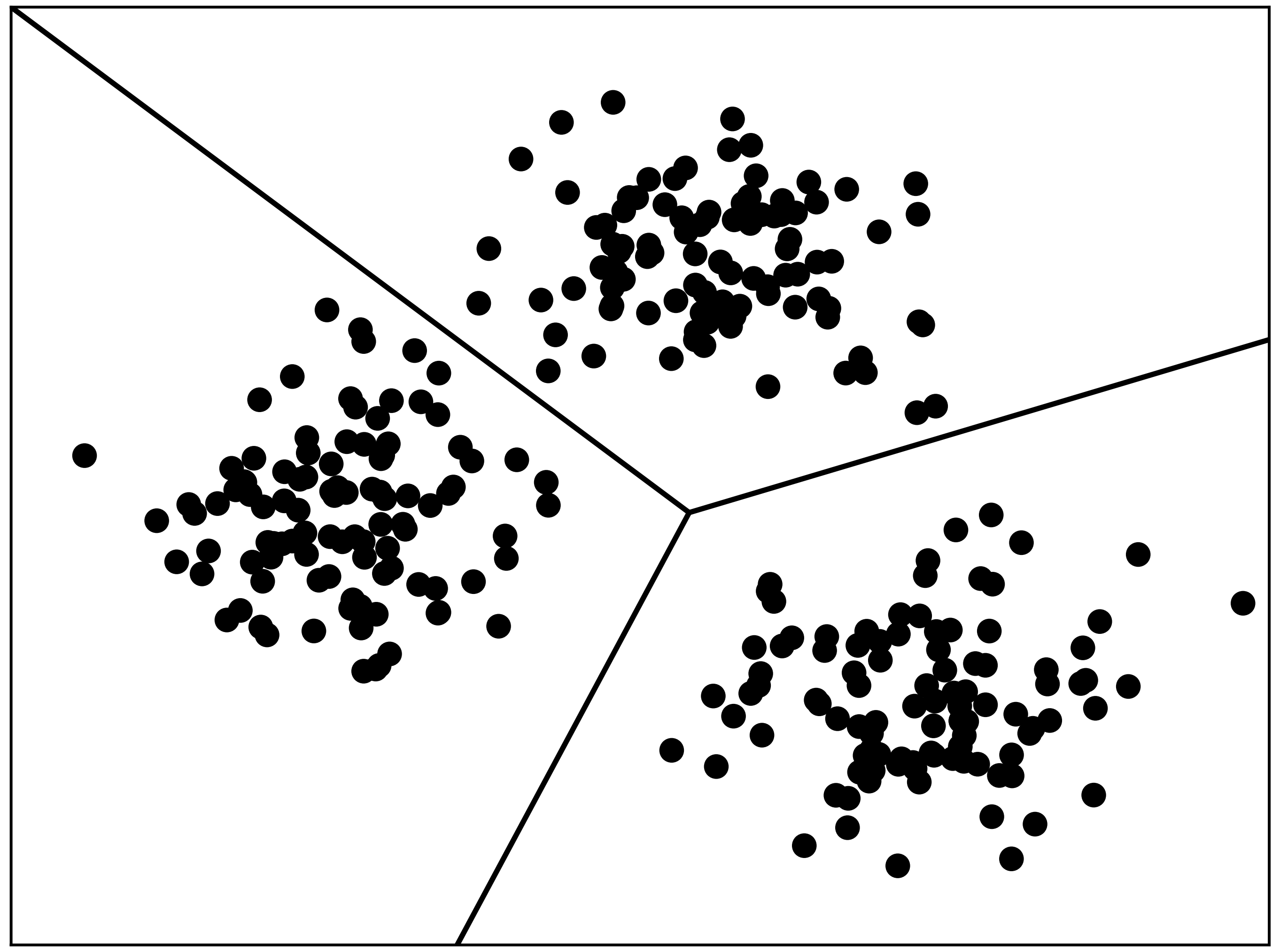




weight ≥ 100

age ≥ 90

Optimal clustering sometimes hard to explain



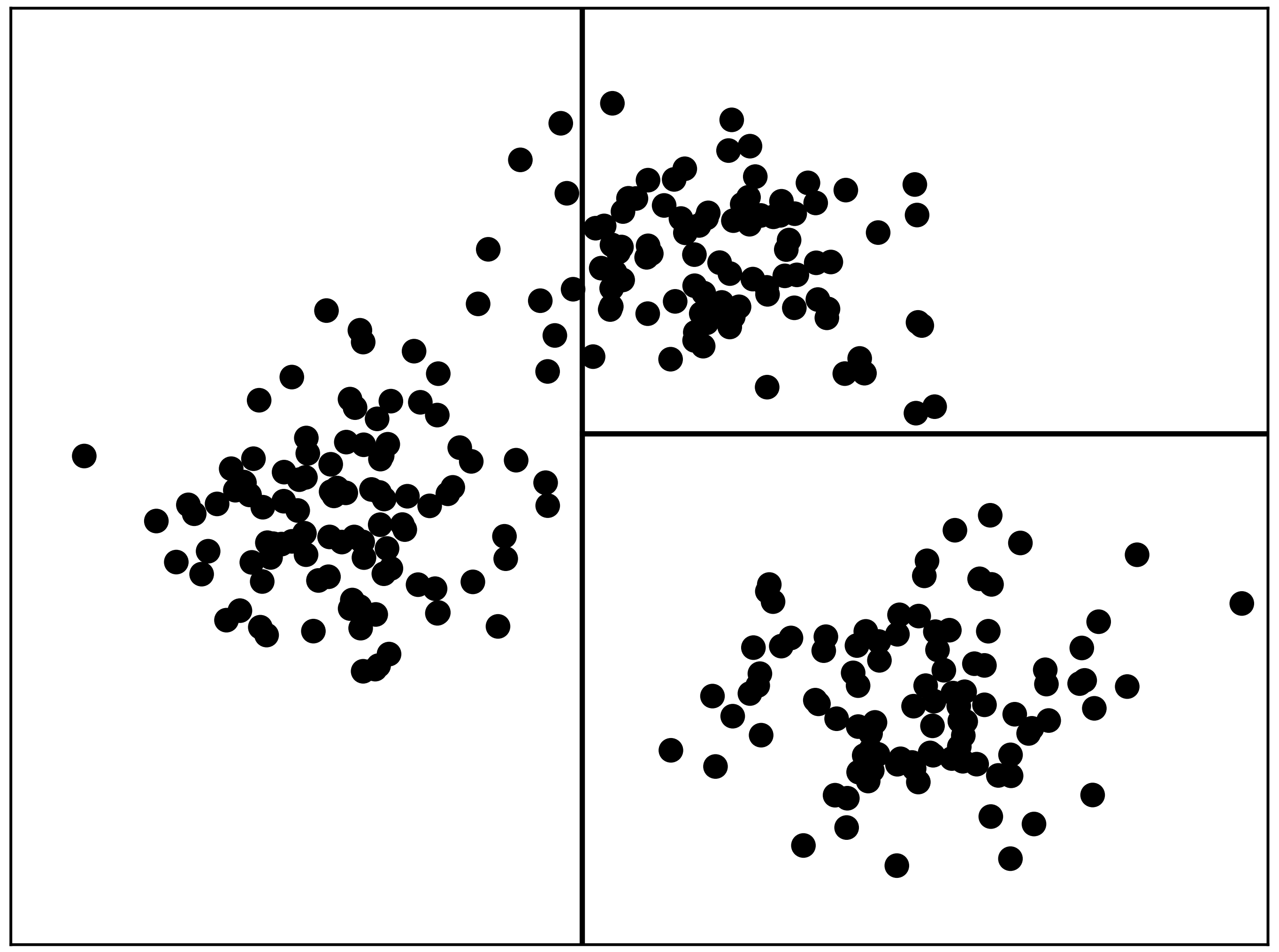
points in $\mathbb{R}^d$

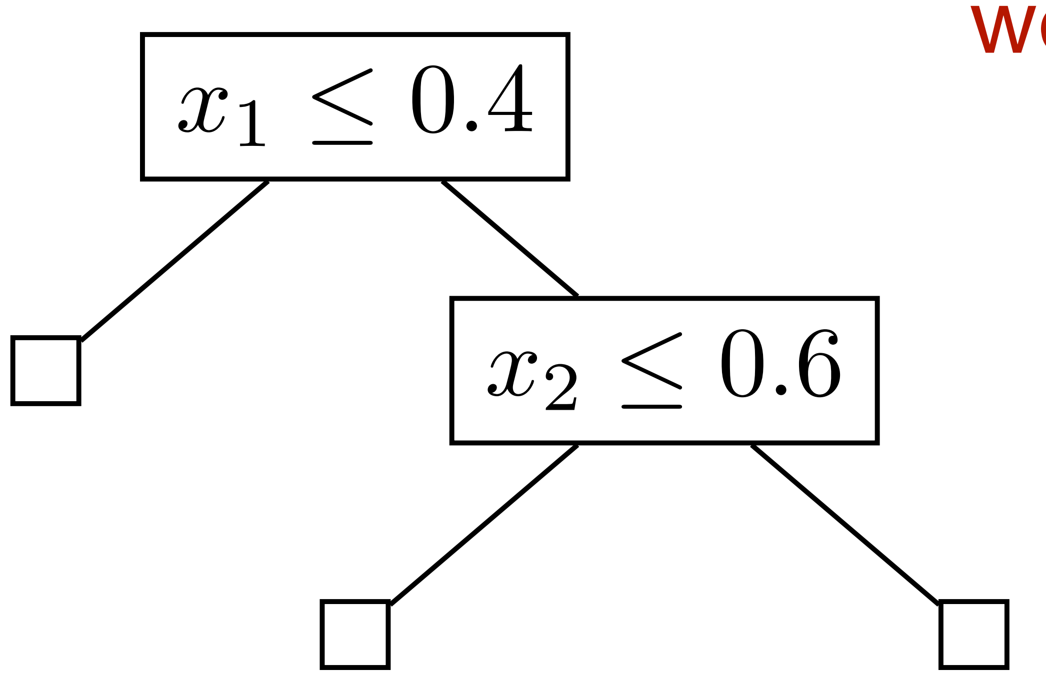
$0.6 \cdot \text{weight} + 0.7 \cdot \text{age} + 2 \cdot \text{vaccinated} \leq 1.5$

and

$0.9 \cdot \text{location} + 1.4 \cdot \text{weight} + 0.7 \cdot \text{age} \geq 2.5$

More explainable "threshold tree"



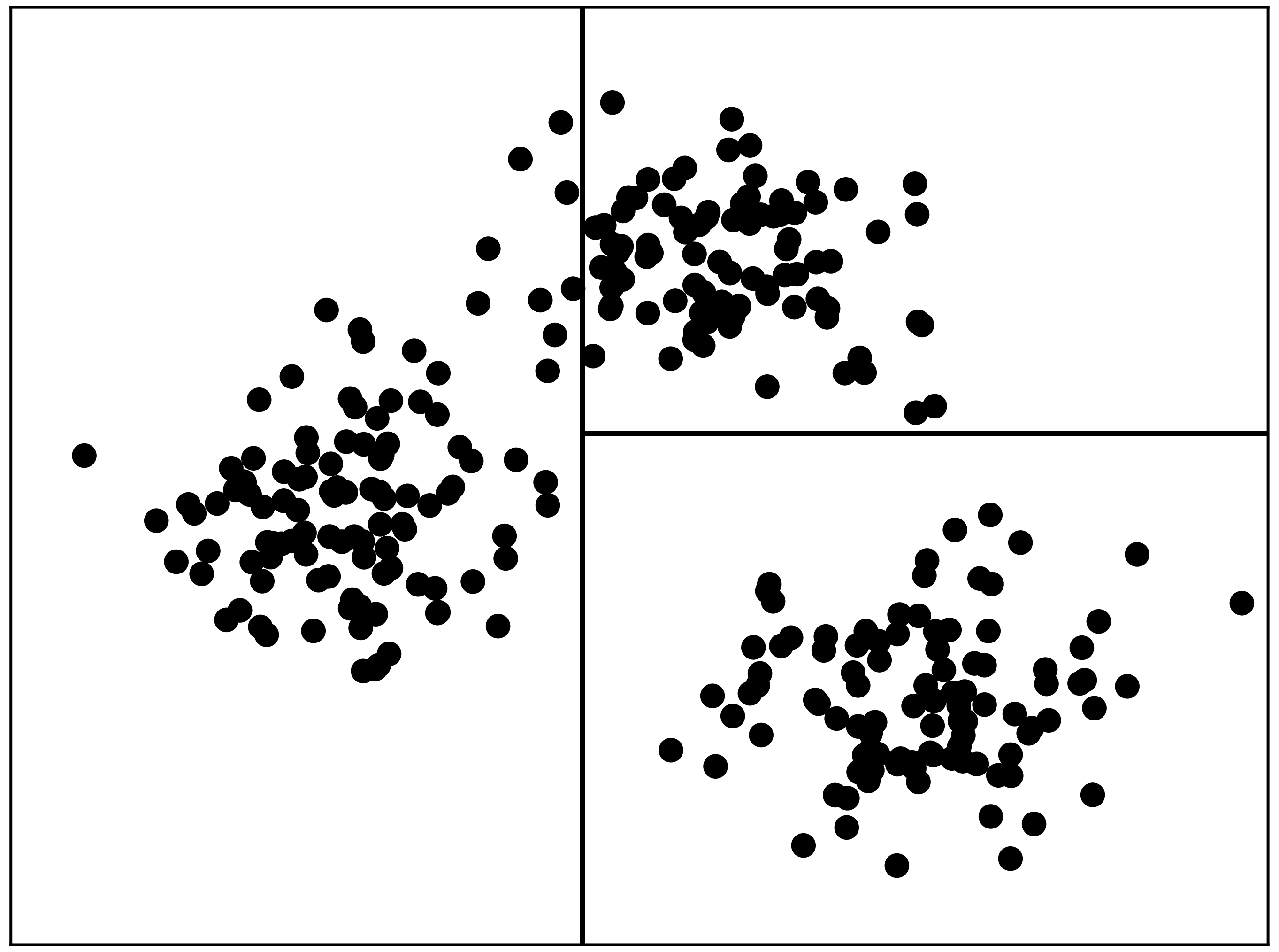


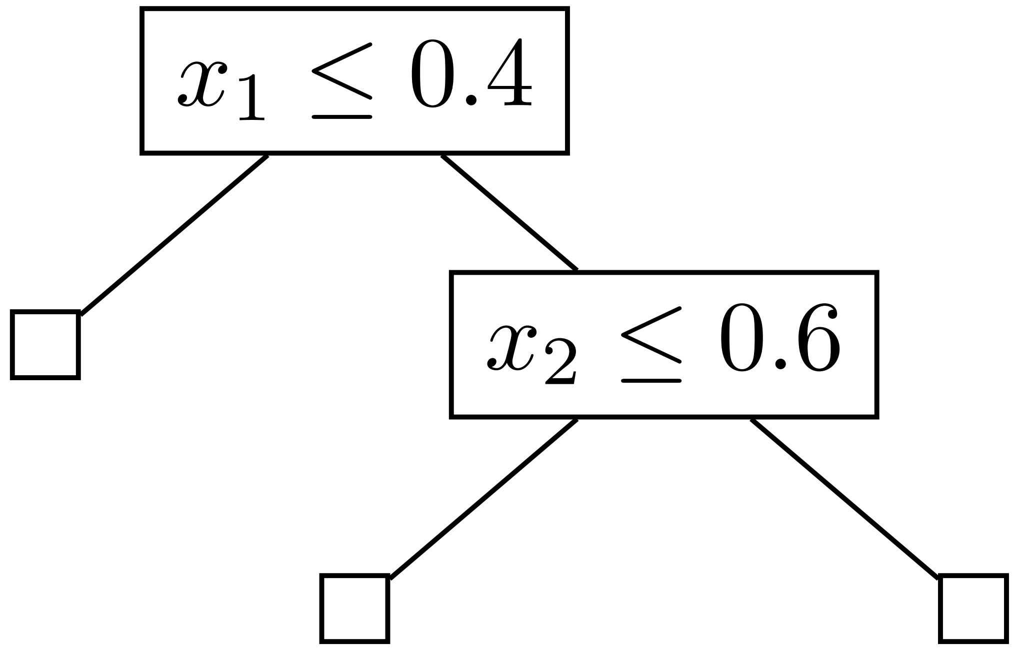
weight ≥ 100

age ≥ 90

unvaccina

Explainable clustering





- A *threshold tree* is a binary tree where each non-leaf node is an axis-aligned threshold cut
- An explainable *k*-clustering is one formed by a threshold tree with *k* leaves

Price of explainability

Price of explainability

- How much more expensive is an optimal *explainable* clustering?

Price of explainability

- How much more expensive is an optimal *explainable* clustering?

- Can we find a good explainable clustering efficiently?

Price of explainability

- How much more expensive is an optimal *explainable* clustering?
- Can we find a good explainable clustering efficiently?
- Can we (approximately) find the best possible explainable clustering?

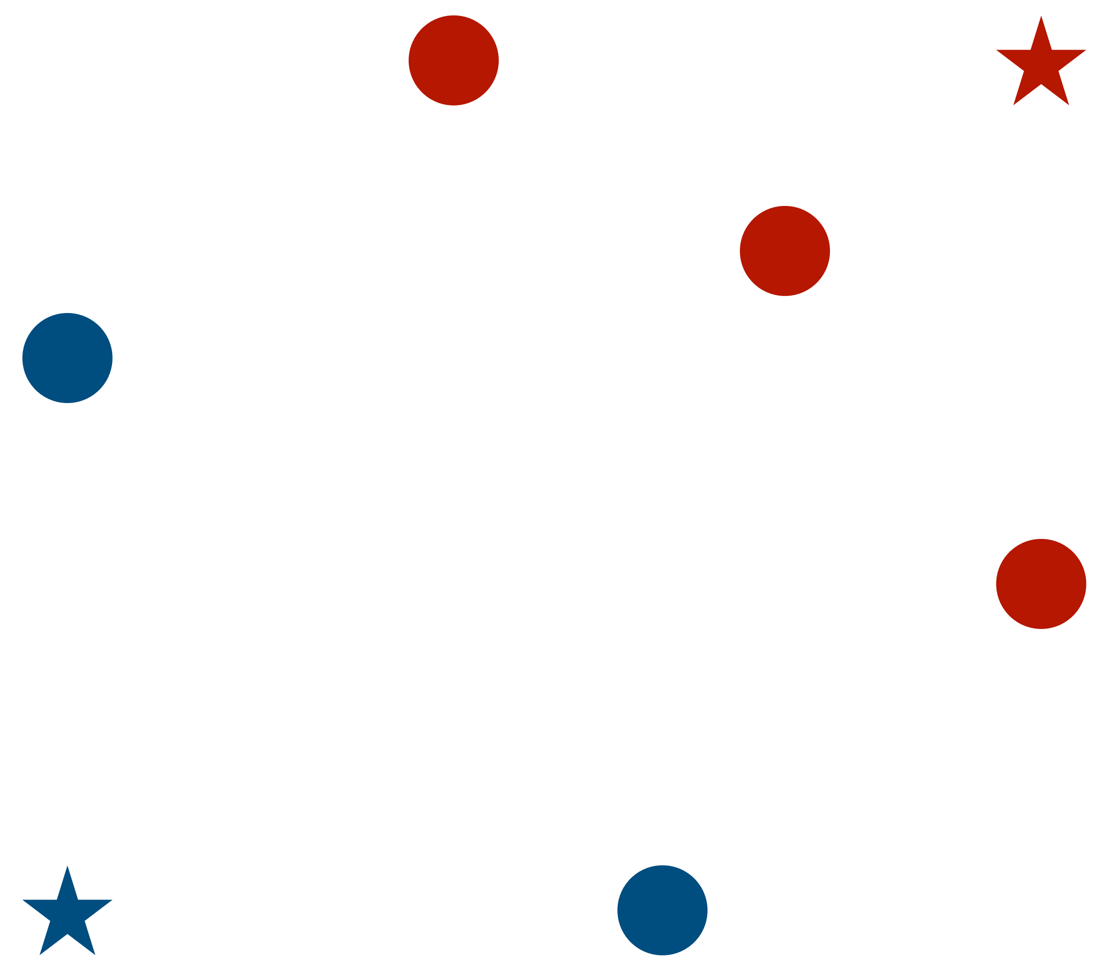
Price of explainability

- How much more expensive is an optimal *explainable* clustering?
- Can we find a good explainable clustering efficiently?
- Can we (approximately) find the best possible explainable clustering?
- First introduced and studied theoretically by Moshkovitz, Dasgupta, Rashtchian, and Frost (ICML'20)

Explaining explainable clustering in four steps

- General Approach of Moshkovitz, Dasgupta, Rashtchian, and Frost
- *TCS-Algorithm*
- Ideas of analysis
- State-of-the-art and open questions

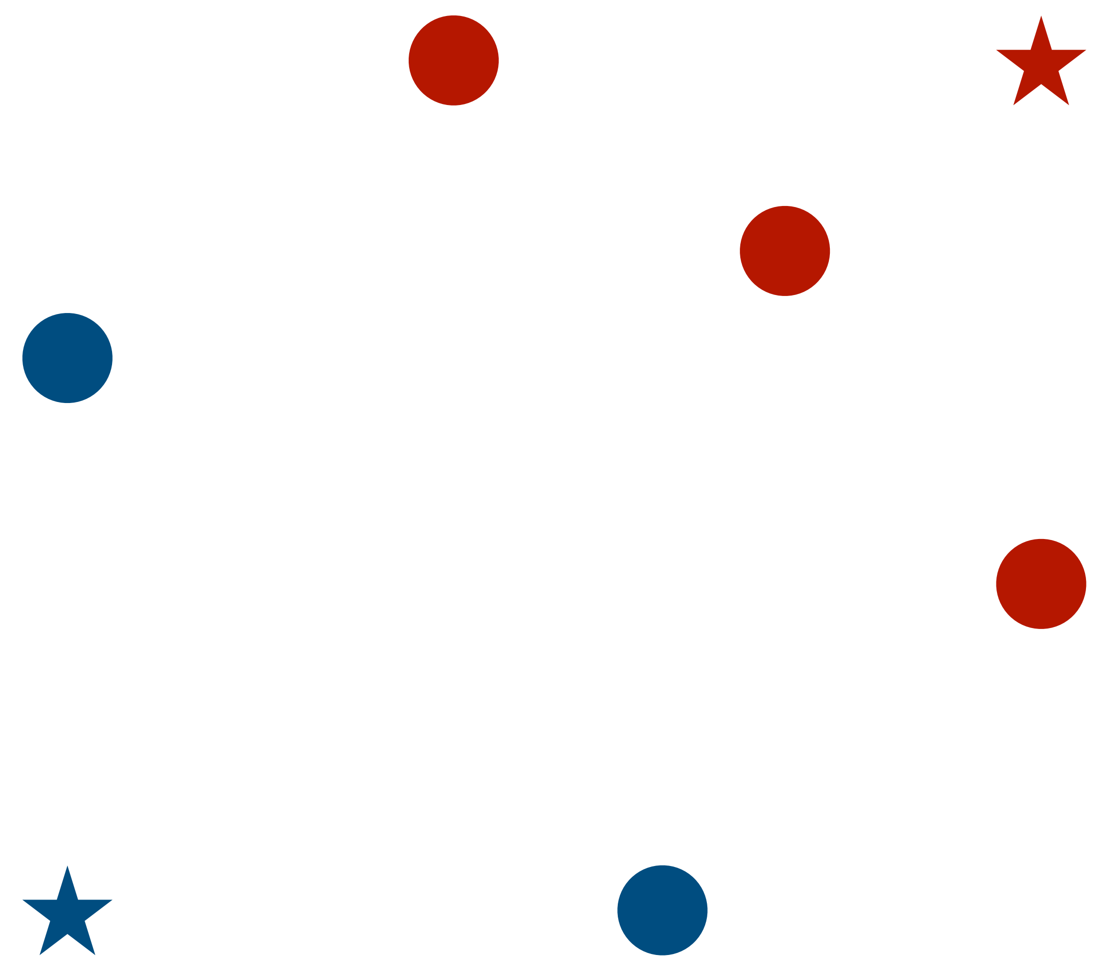
Focus on k-median



Focus on k-median

- Points X in $\mathbb{R}^d$
- Distance ℓ_1-norm. That is

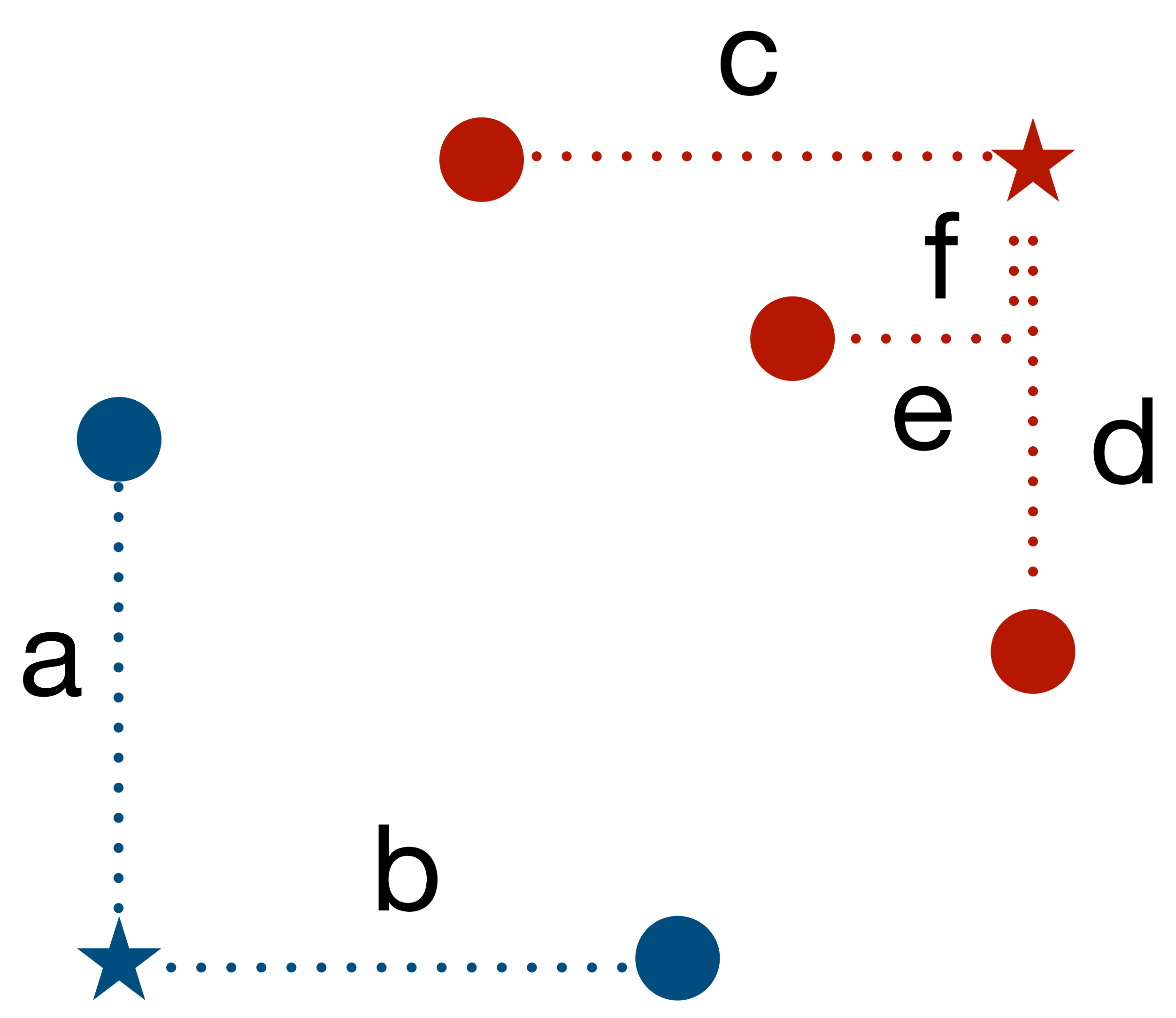
$$dist(x, y) = \sum_{i=1}^{d} |x_i - y_i|$$



Focus on k-median

- Points X in $\mathbb{R}^d$
- Distance ℓ_1-norm. That is

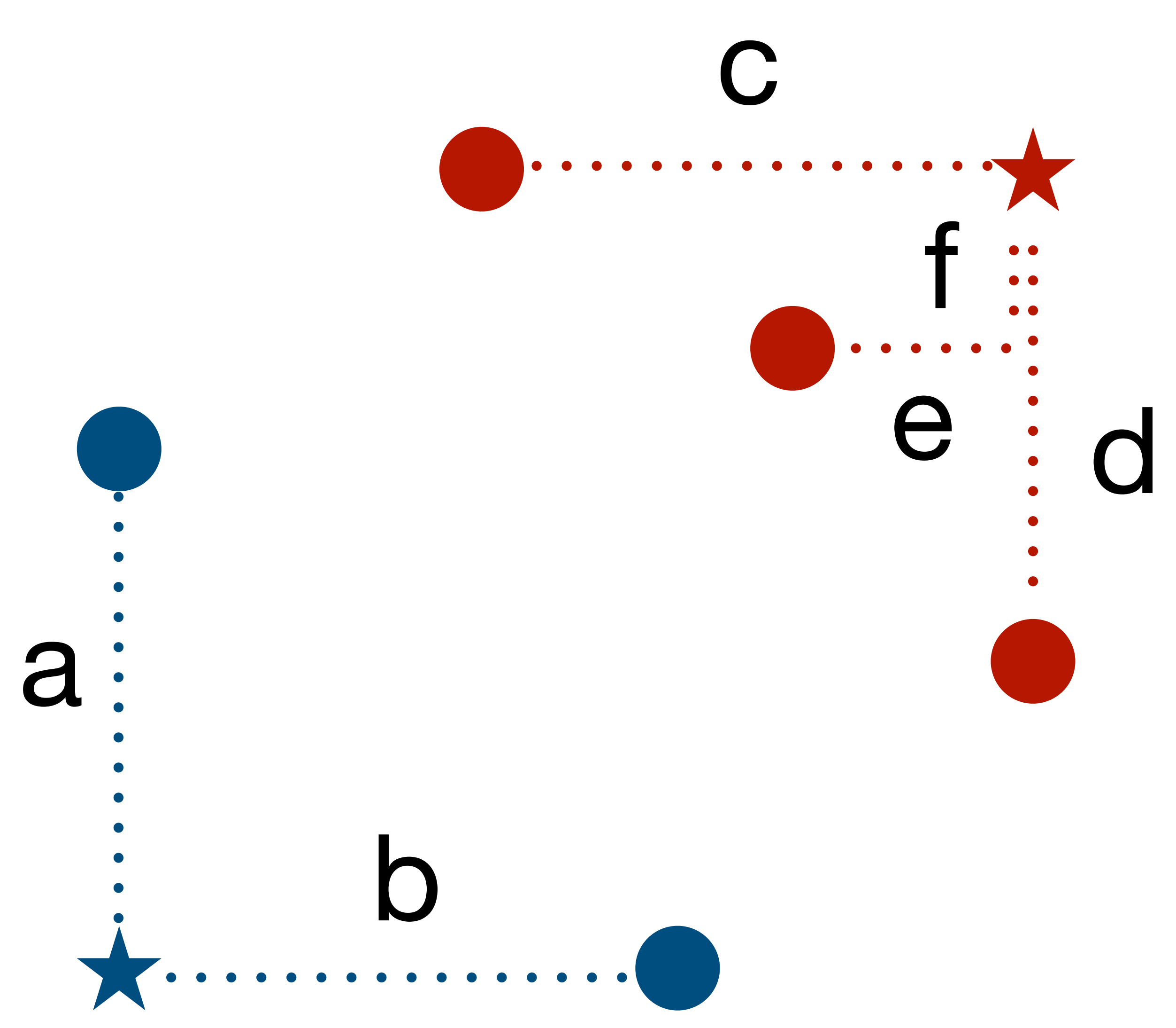
$$dist(x, y) = \sum_{i=1}^{d} |x_i - y_i|$$



Focus on k-median

- Points X in $\mathbb{R}^d$
- Distance ℓ_1-norm. That is

$$dist(x, y) = \sum_{i=1}^{d} |x_i - y_i|$$

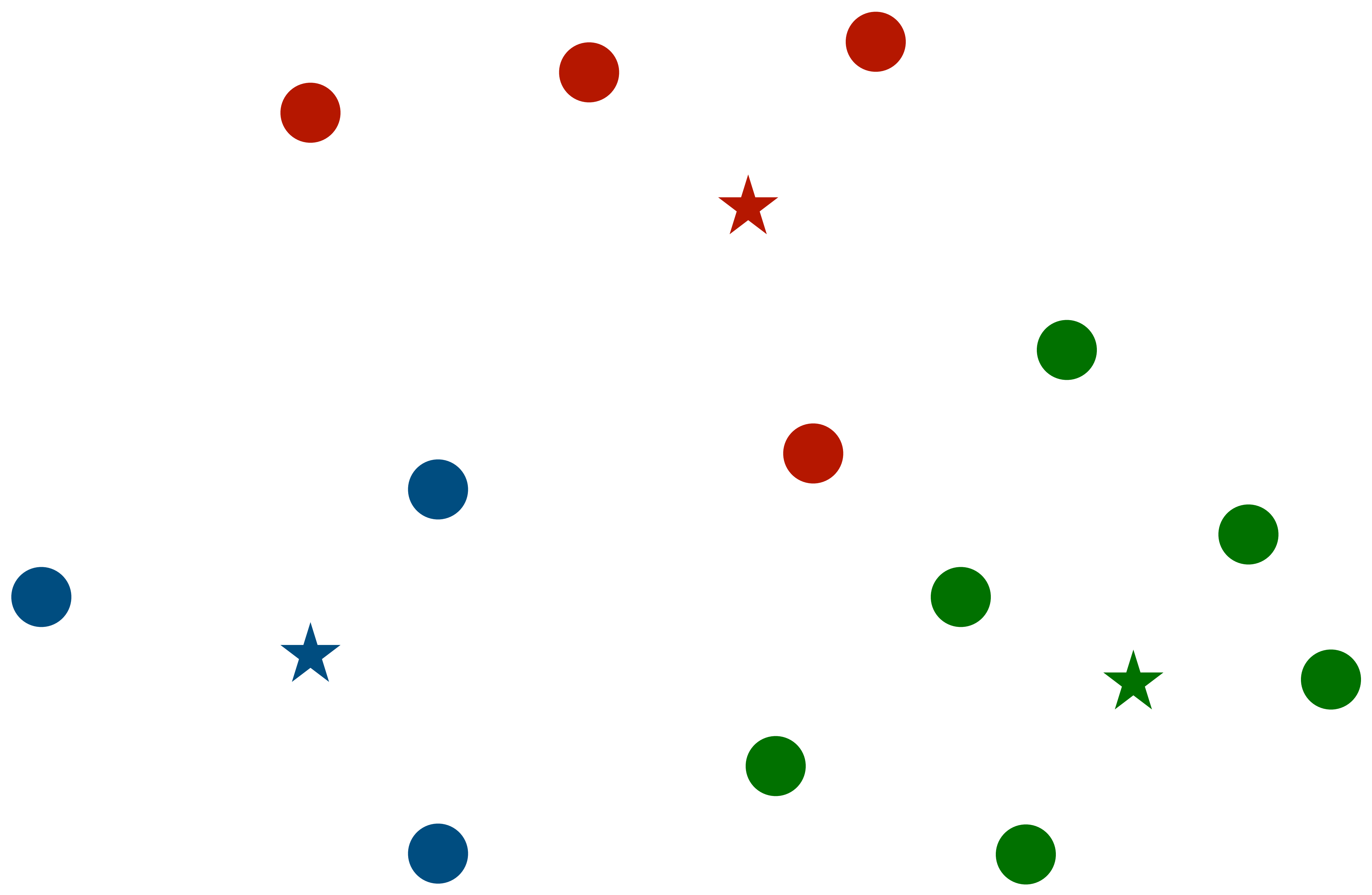


- Cost of optimal unconstrained clustering equals sum of dotted edges

$$OPT = a + b + c + d + e + f$$

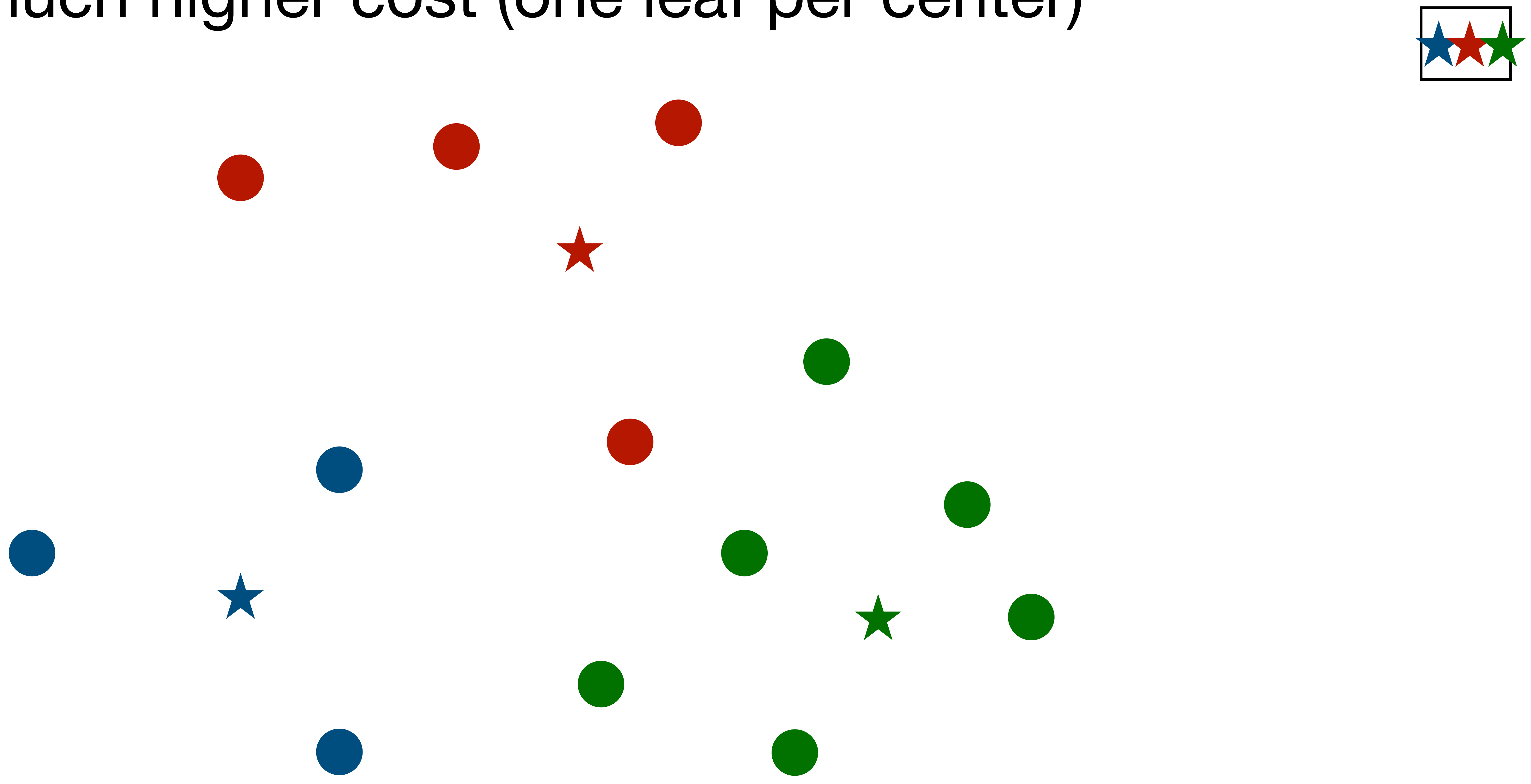
General approach

- Transform given reference clustering of cost OPT to an explainable clustering with not much higher cost (one leaf per center)



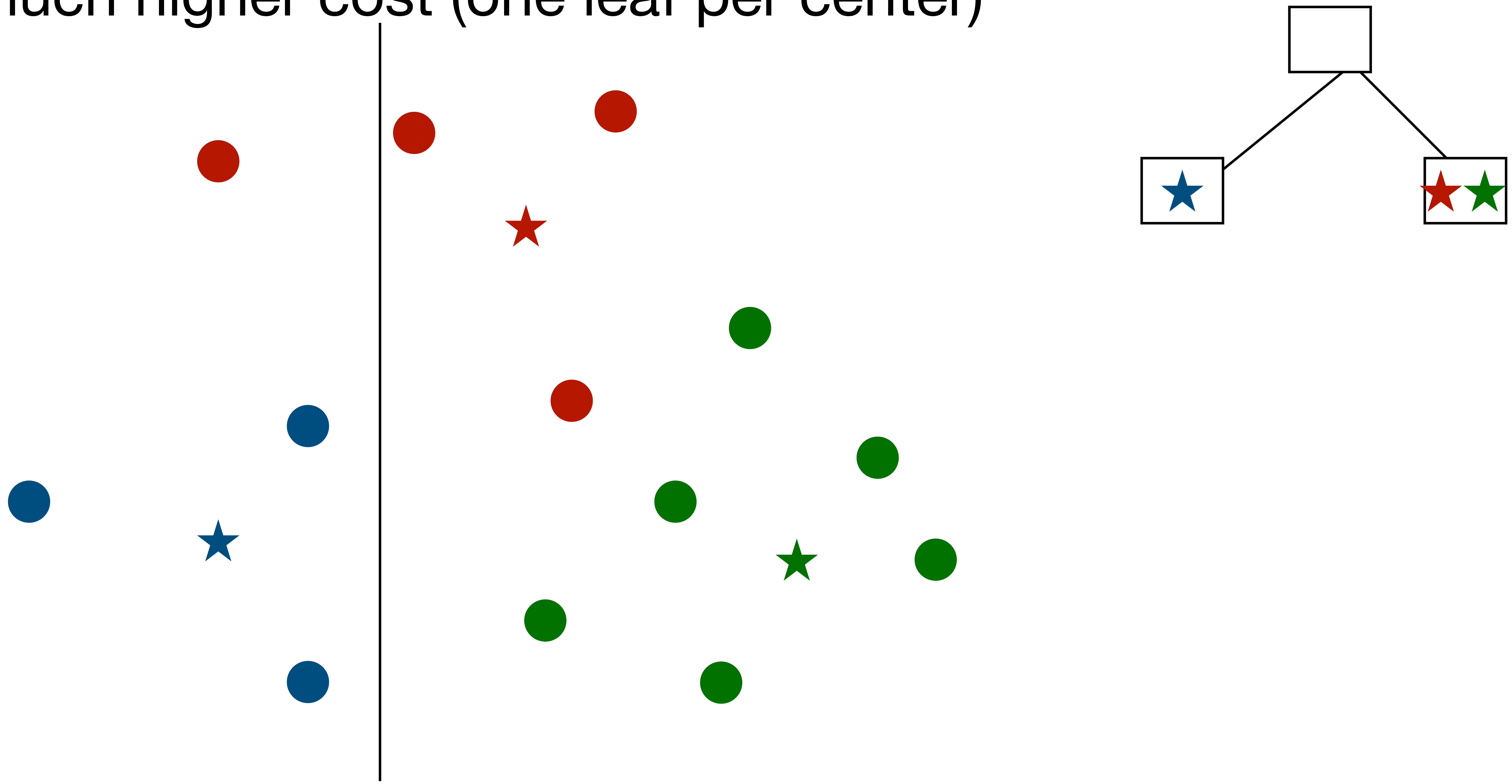
General approach

- Transform given reference clustering of cost OPT to an explainable clustering with not much higher cost (one leaf per center)



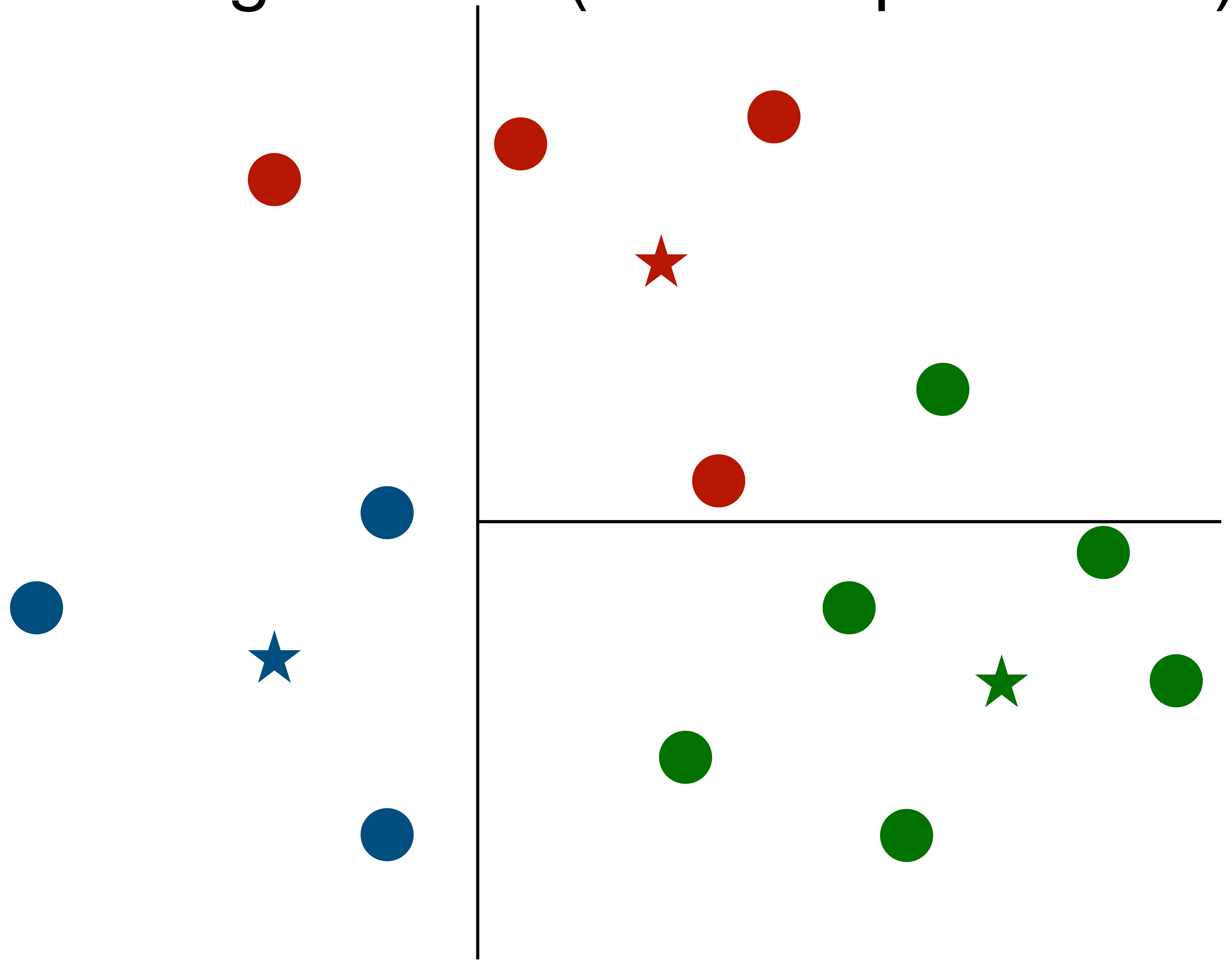
General approach

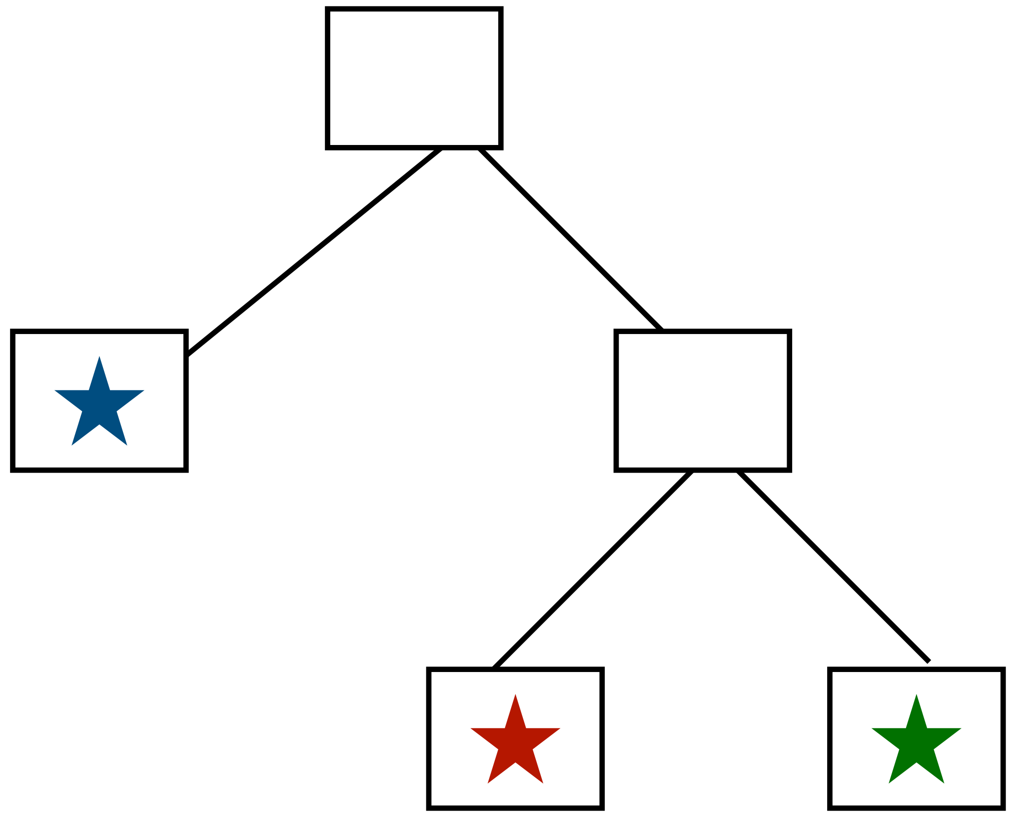
- Transform given reference clustering of cost OPT to an explainable clustering with not much higher cost (one leaf per center)



General approach

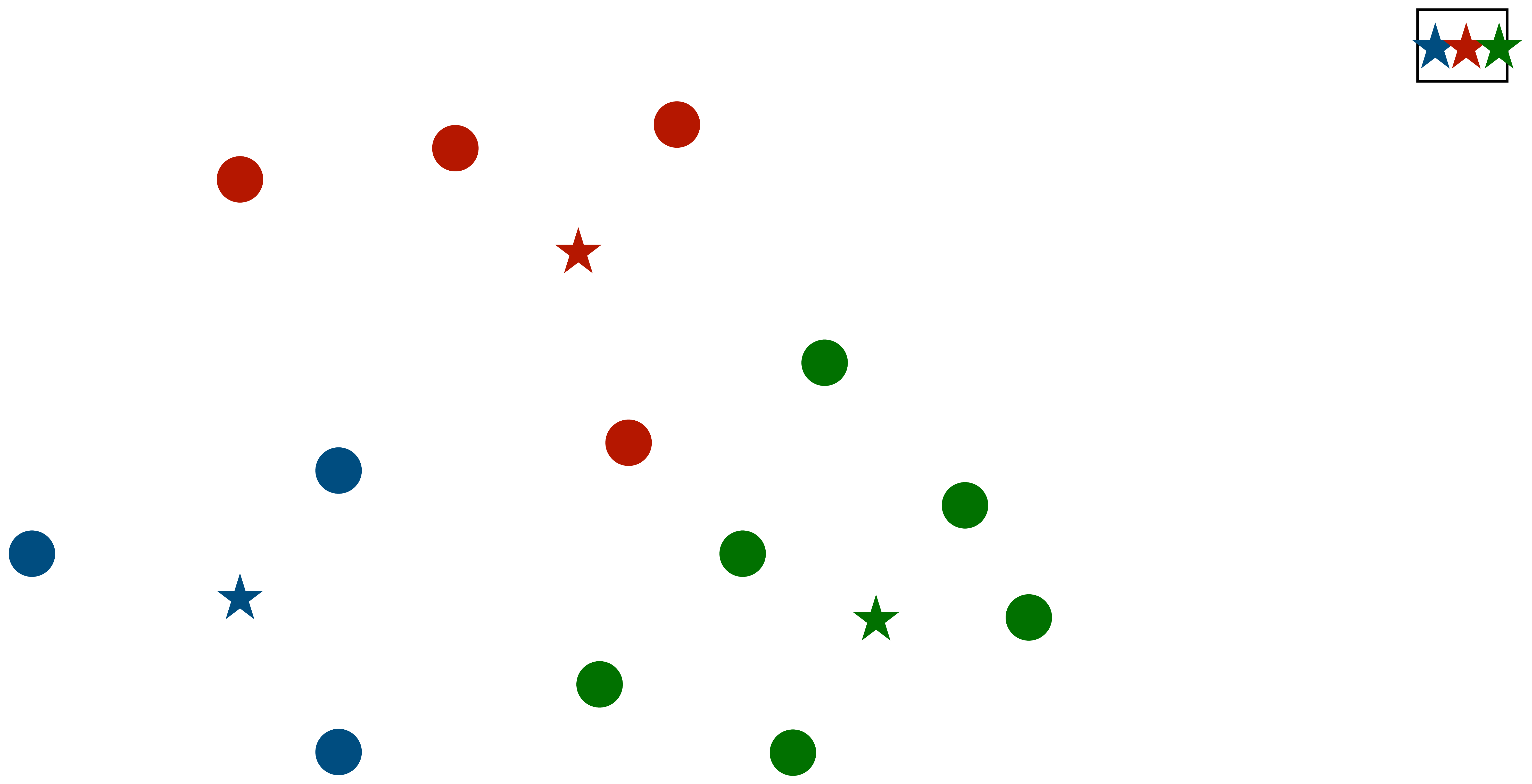
- Transform given reference clustering of cost OPT to an explainable clustering with not much higher cost (one leaf per center)





The algorithm of MDRF

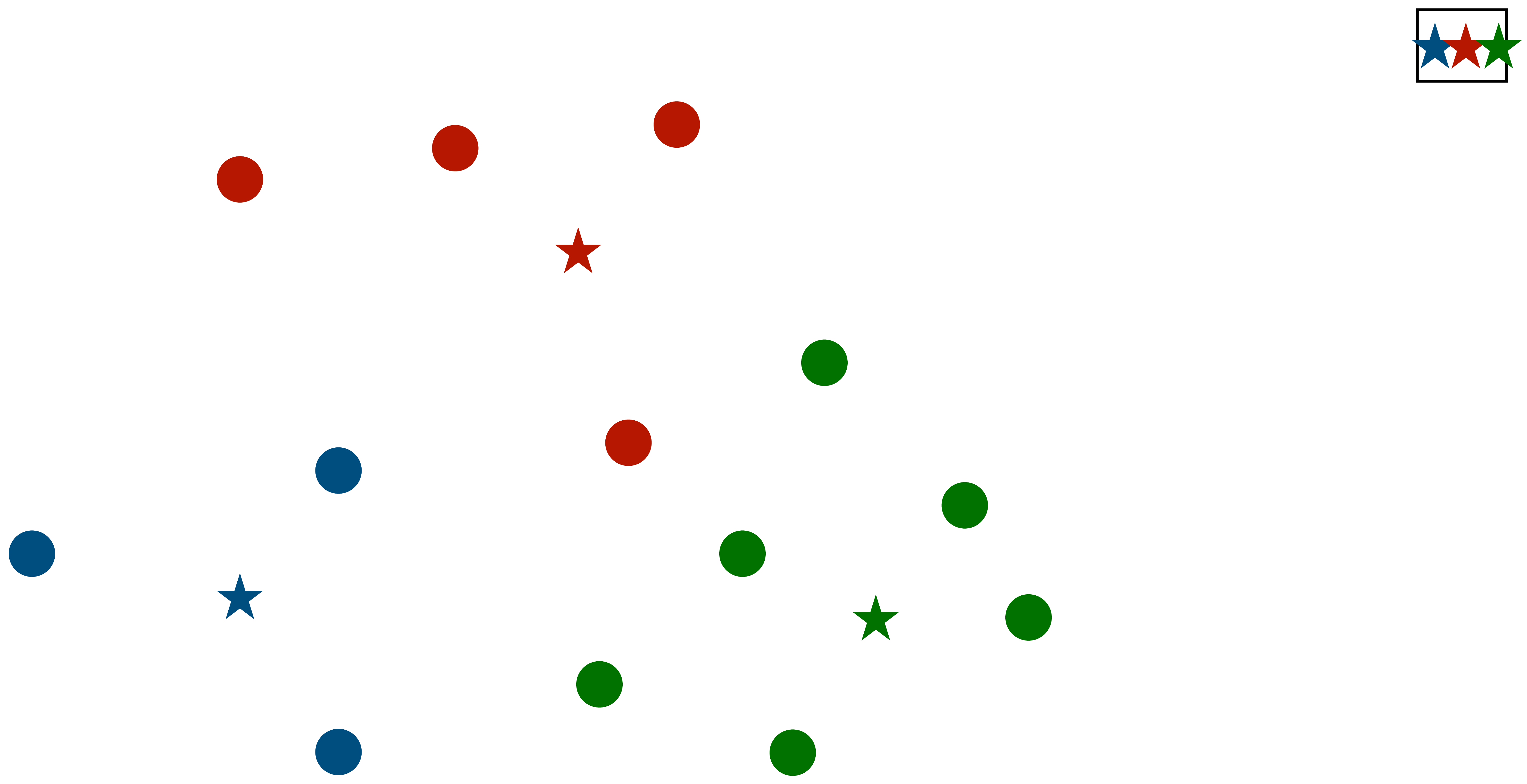
- While there is a leaf with more than one center, **select a min-cut**



The algorithm of MDRF

- While there is a leaf with more than one center, **select a min-cut**

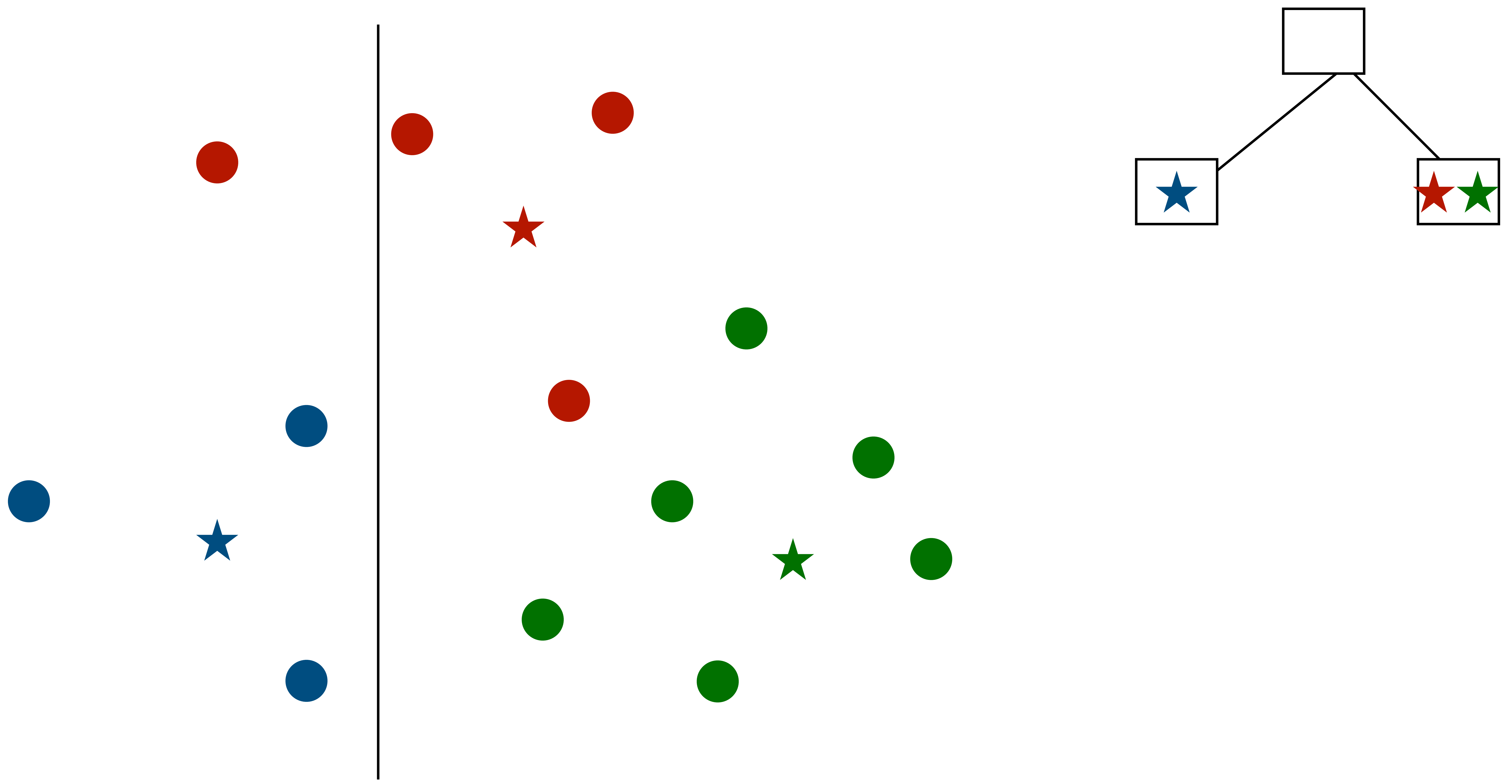
cut that separates fewest number of points from closest center



The algorithm of MDRF

- While there is a leaf with more than one center, **select a min-cut**

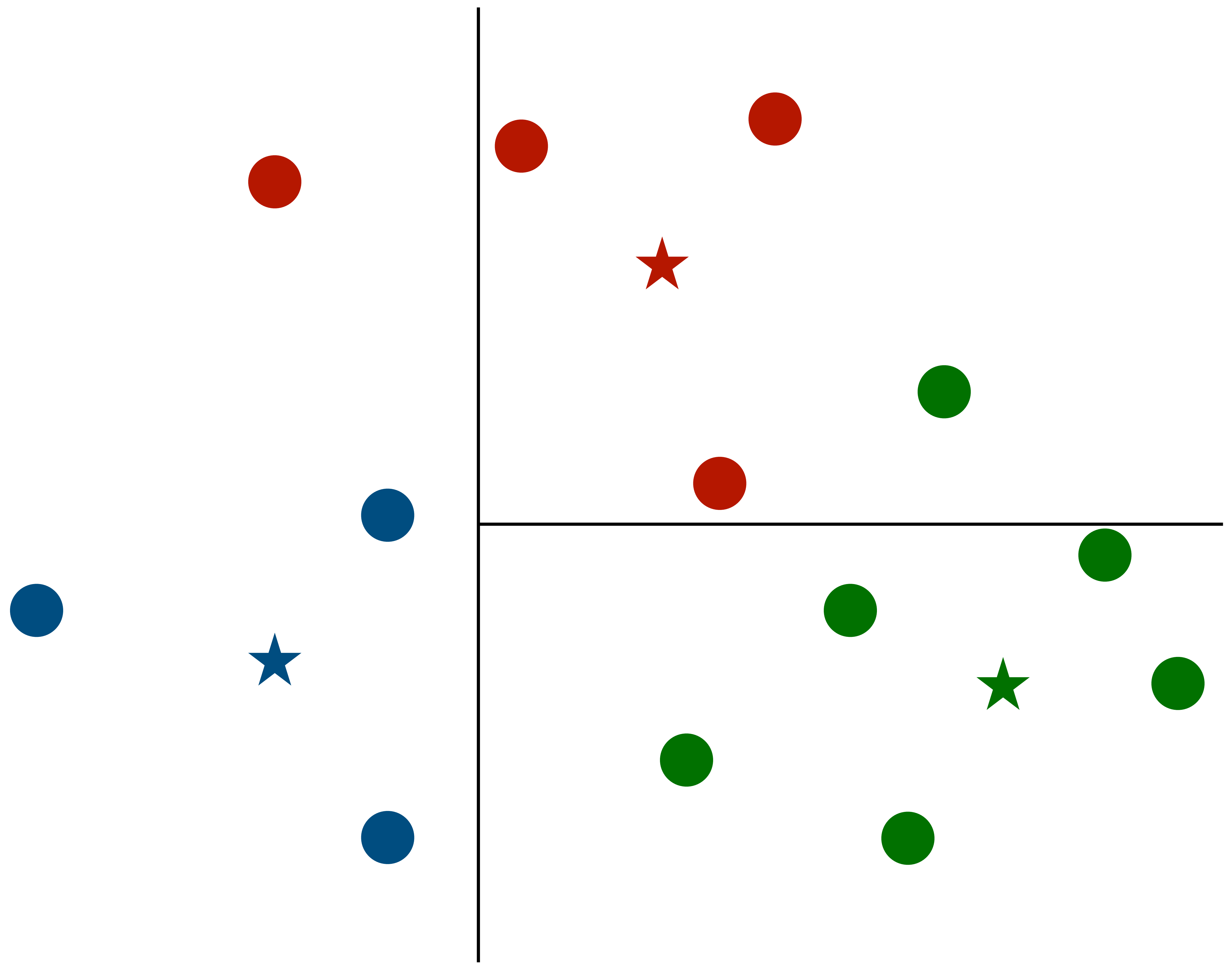
cut that separates fewest number of points from closest center

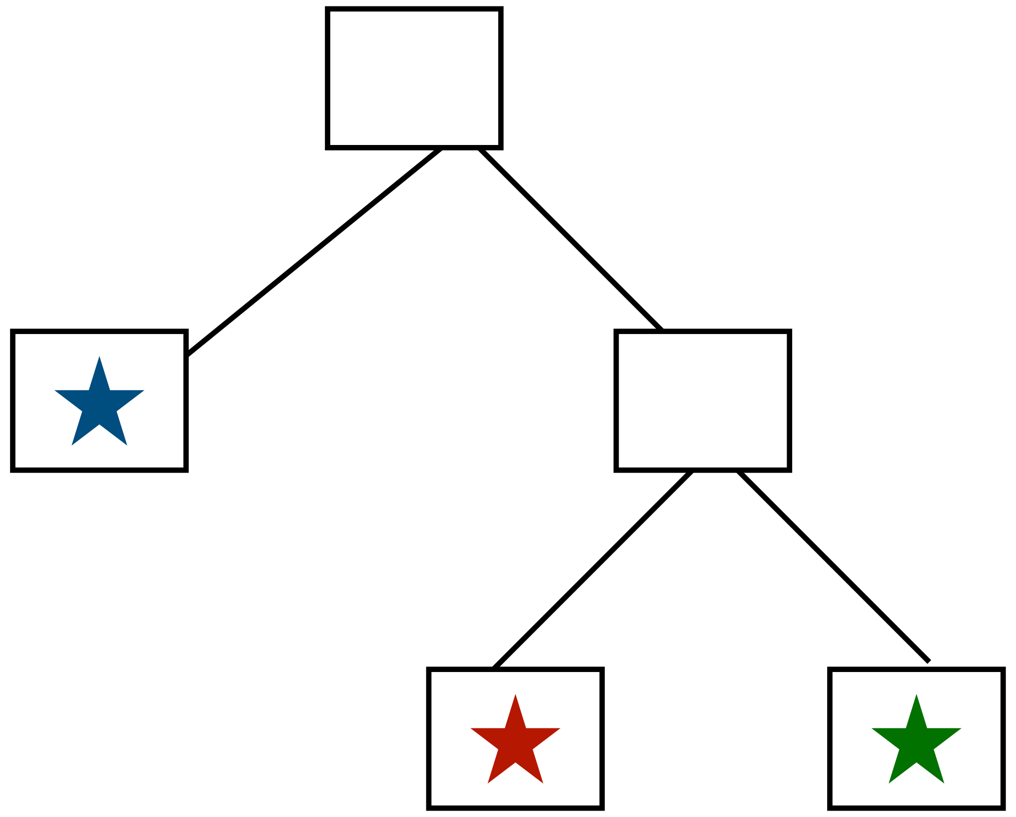


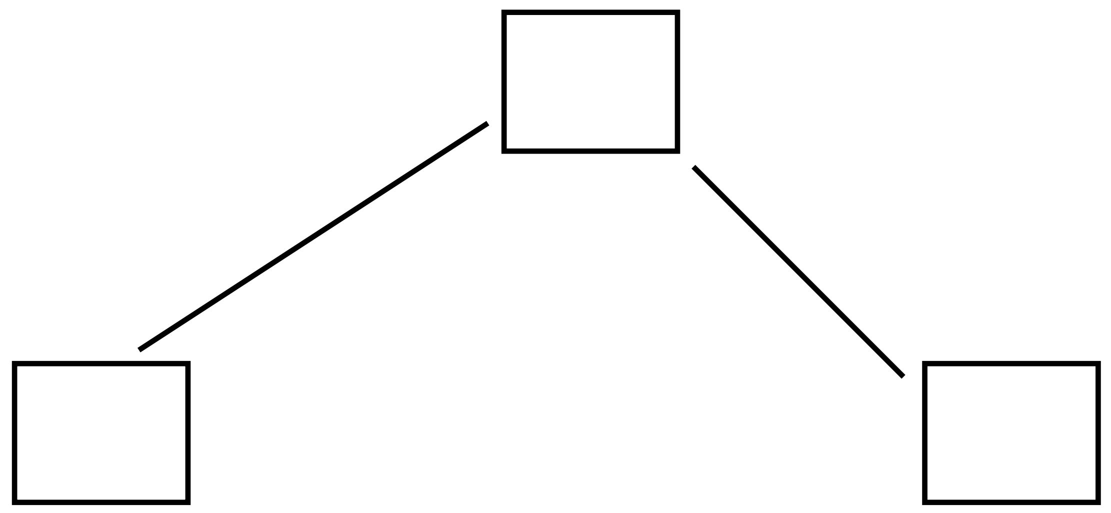
The algorithm of MDRF

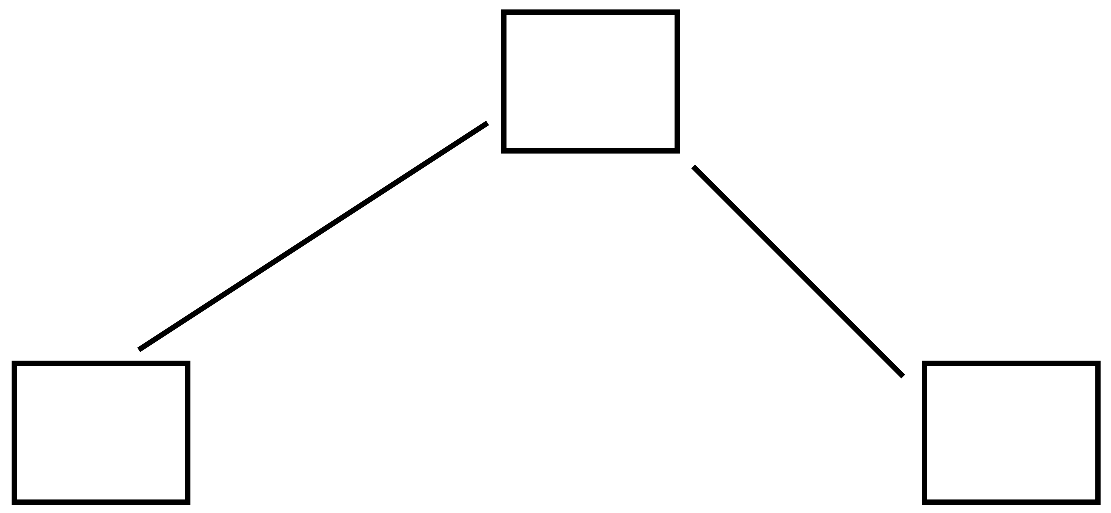
- While there is a leaf with more than one center, **select a min-cut**

cut that separates fewest number of points from closest center

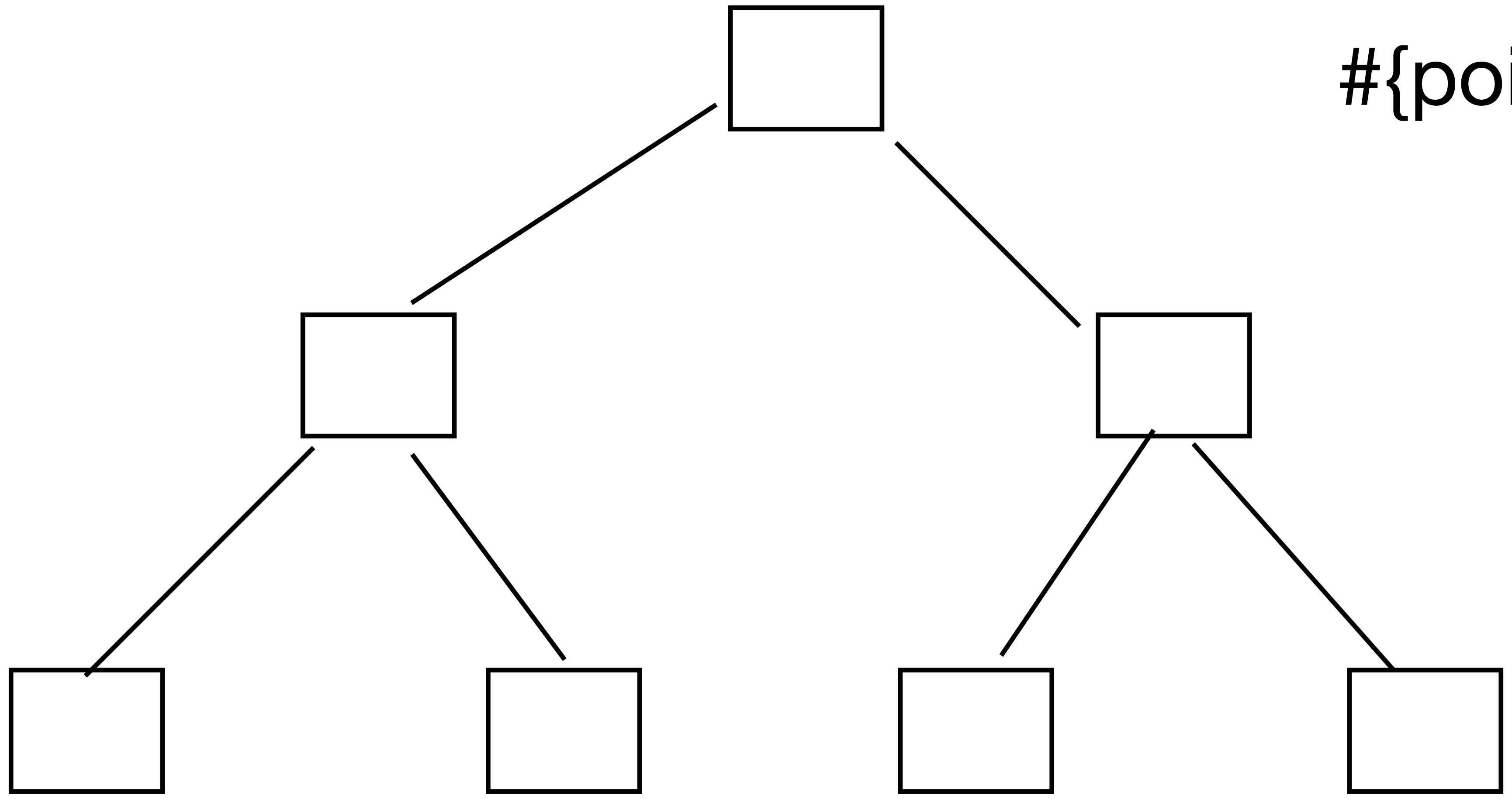






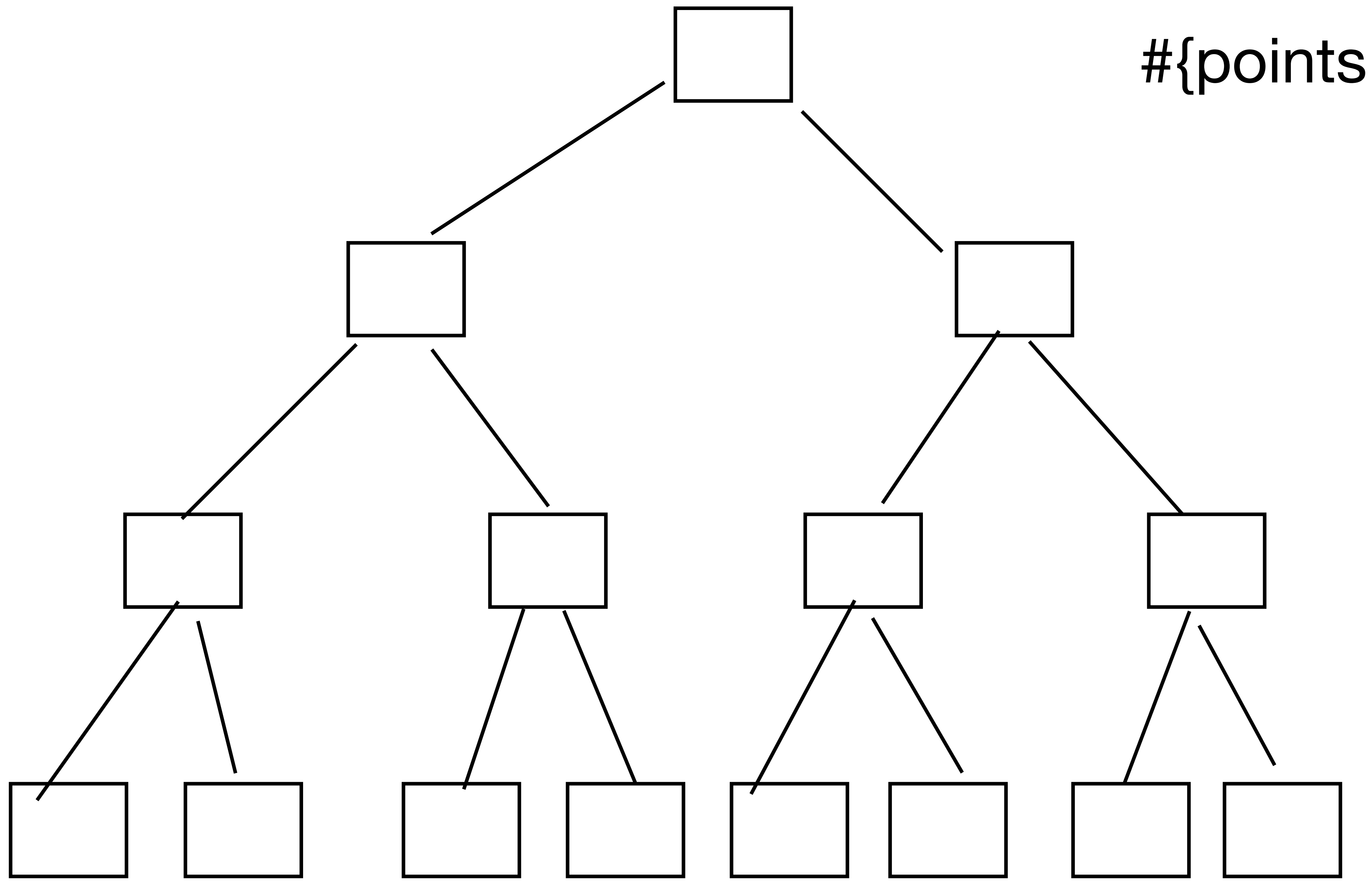


#{points separated by min-cut} * {distance to farthest away centre} $\leq$ OPT



#{points separated by min-cut} * {distance to farthest away centre} $\leq$ OPT

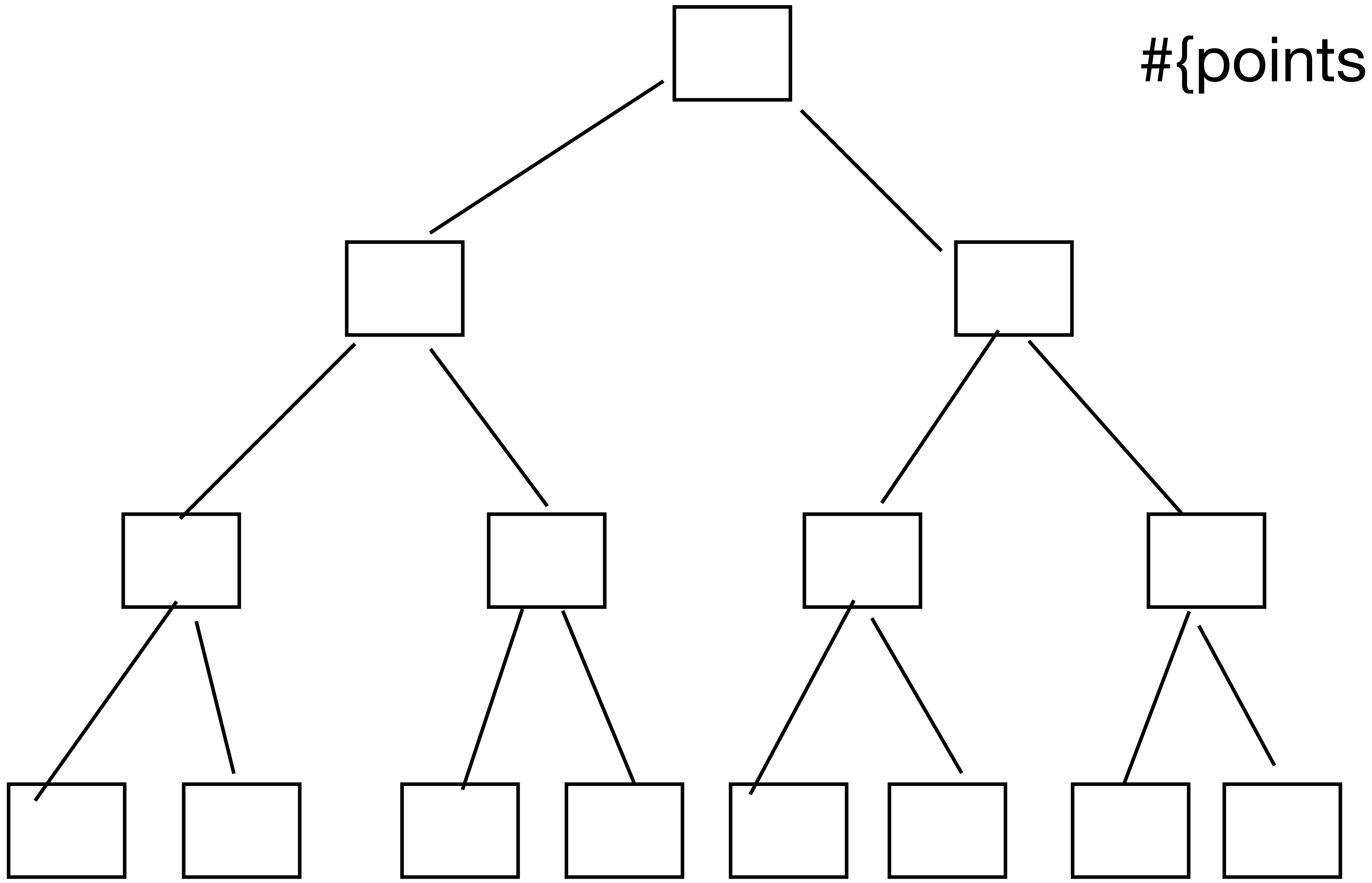
OPT(Left) + OPT(Right) $\leq$ OPT



#{points separated by min-cut} * {distance to farthest away centre} $\leq$ OPT

OPT(Left) + OPT(Right) $\leq$ OPT

$\leq$ OPT

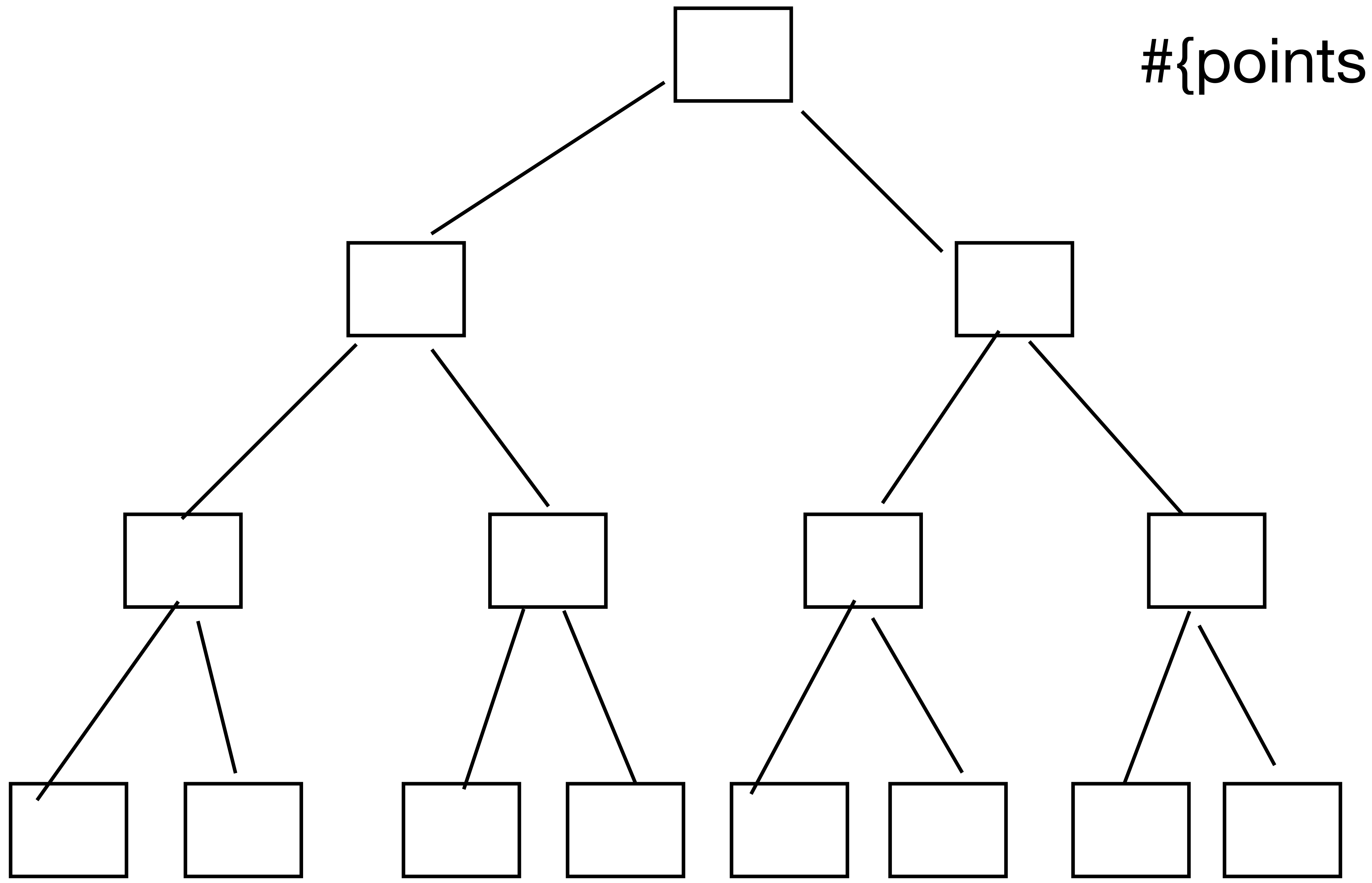


#{points separated by min-cut} * {distance to farthest away centre} $\leq$ OPT

OPT(Left) + OPT(Right) $\leq$ OPT

$\leq$ OPT

Cost increase at each level is at most OPT



#{points separated by min-cut} * {distance to farthest away centre} $\leq$ OPT

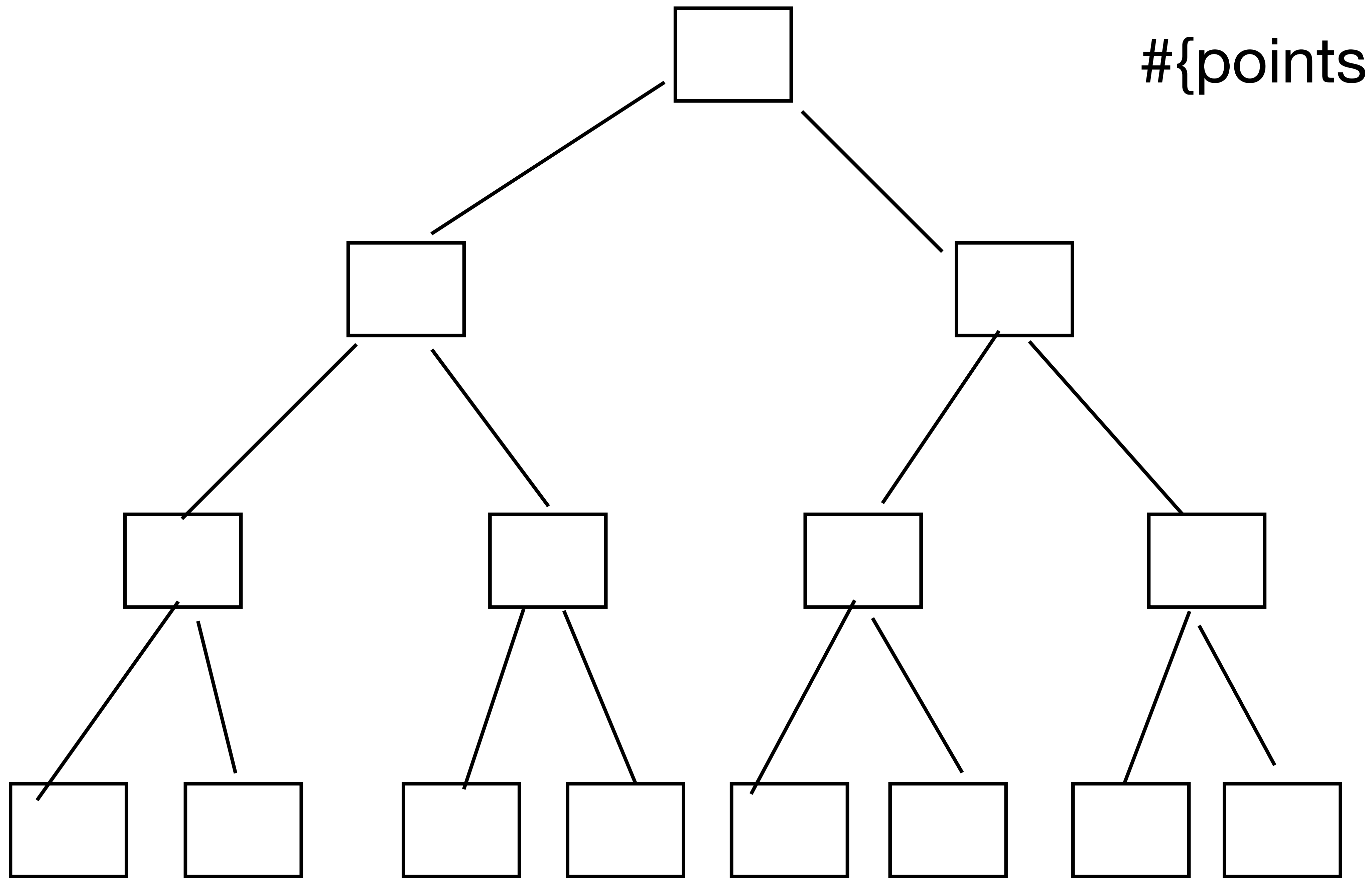
OPT(Left) + OPT(Right) $\leq$ OPT

$\leq$ OPT

Cost increase at each level is at most OPT

MDRF'20

Price of explainability is at most the height of tree and hence at most O(k)



#{points separated by min-cut} * {distance to farthest away centre} $\leq$ OPT

OPT(Left) + OPT(Right) $\leq$ OPT

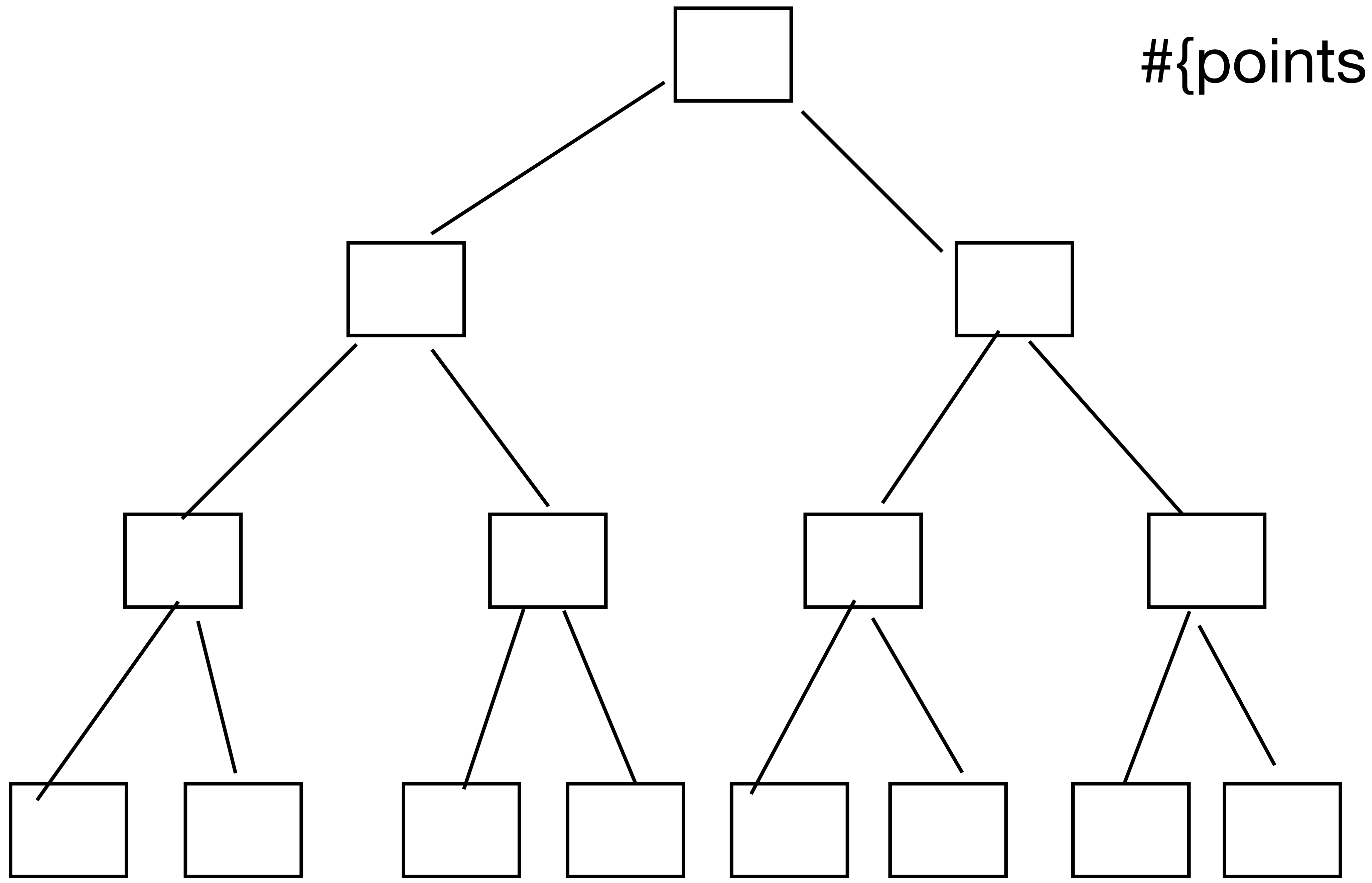
$\leq$ OPT

Cost increase at each level is at most OPT

MDRF'20

Price of explainability is at most the height of tree and hence at most O(k)

Price of explainability of k-means is $O(k^2)$



#{points separated by min-cut} * {distance to farthest away centre} $\leq$ OPT

OPT(Left) + OPT(Right) $\leq$ OPT

$\leq$ OPT

Cost increase at each level is at most OPT

MDRF'20

Price of explainability is at most the height of tree and hence at most O(k)

Price of explainability of k-means is $O(k^2)$

MDRF'20

There are instances where the price of explainability is $\Omega(\log k)$


In 2021, Makarychev and Shan
proposed *TCS-Algorithm*


In 2021, Makarychev and Shan proposed *TCS-Algorithm*
In 2021, Gamlath, Jia, Polak, Svensson proposed *TCS-Algorithm*


In 2021, Gamlath, Jia, Polak, Svensson proposed *TCS-Algorithm*
In 2021, Makarychev and Shan proposed *TCS-Algorithm*
In 2021, Esfandiari, Mirrokni, Narayanan proposed *TCS-Algorithm*

How can three different groups independently come up with ≈ same algorithm?

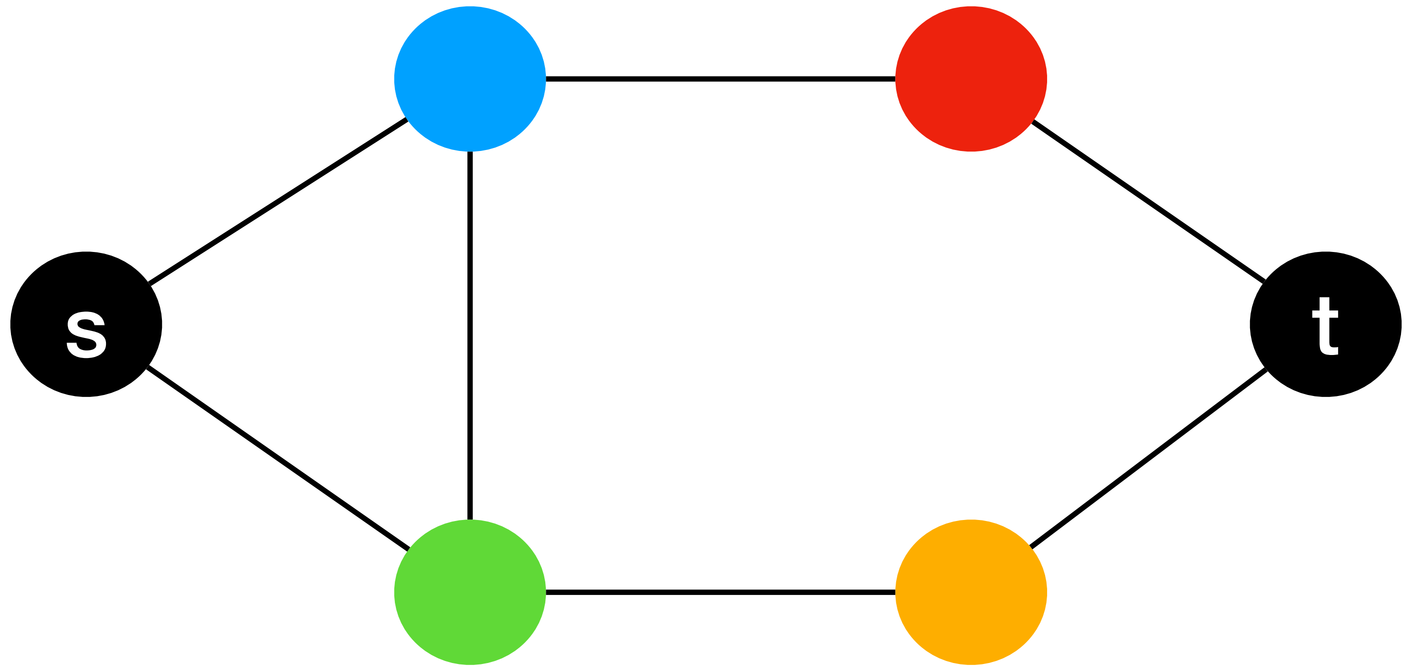
Well, it’s not very complicated

5 Consider an undirected graph $G = (V, E)$ and let $s \neq t \in V$. Show that there is an s,t-cut of value at most the optimal value of the following linear program

$$\begin{array}{ll} \textbf{minimize} & \displaystyle\sum_{\{u,v\}\in E} |x_u - x_v| \\ \textbf{subject to} & x_s = 0, x_t = 1, \text{ and } x_v \in [0,1] \text{ for every } v \in V \end{array}$$

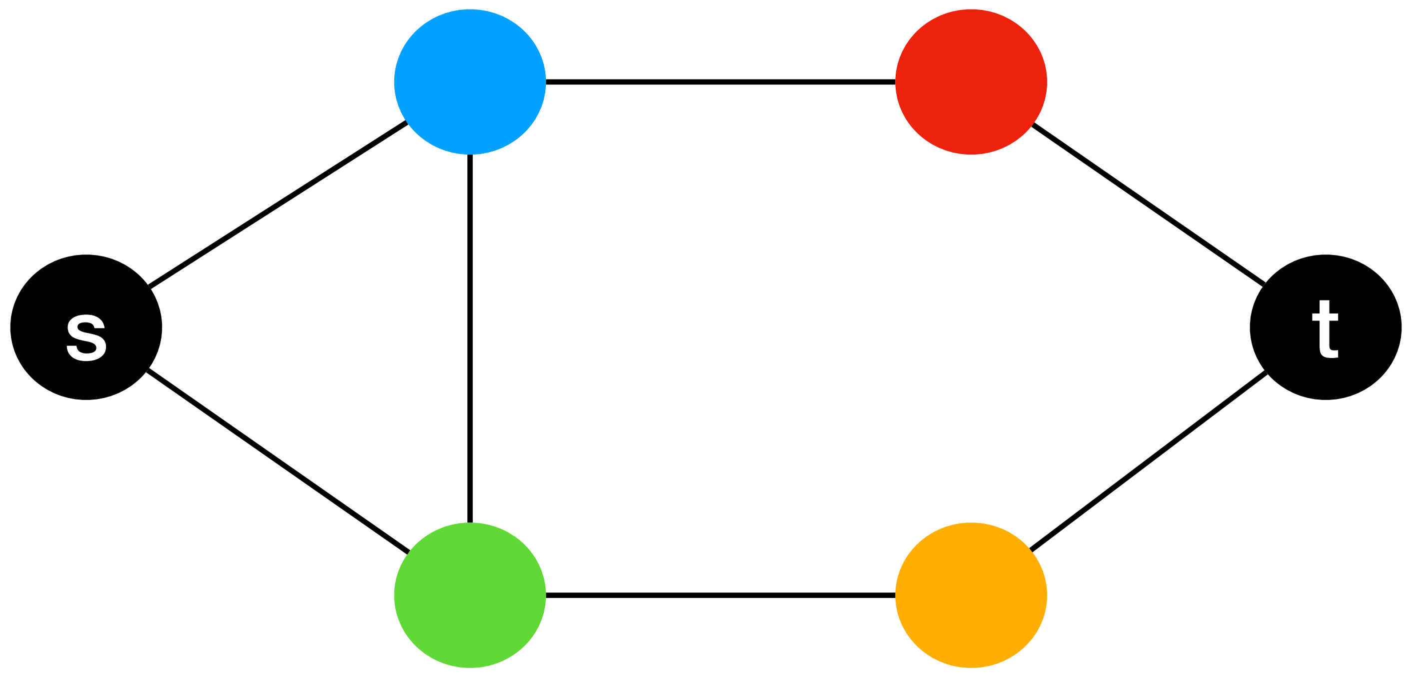
5 Consider an undirected graph $G = (V, E)$ and let $s \neq t \in V$. Show that there is an s, t-cut of value at most the optimal value of the following linear program

$$\begin{aligned} \textbf{minimize} \quad & \sum_{\{u,v\} \in E} |x_u - x_v| \\ \textbf{subject to} \quad & x_s = 0, x_t = 1, \text{ and } x_v \in [0,1] \text{ for every } v \in V \end{aligned}$$



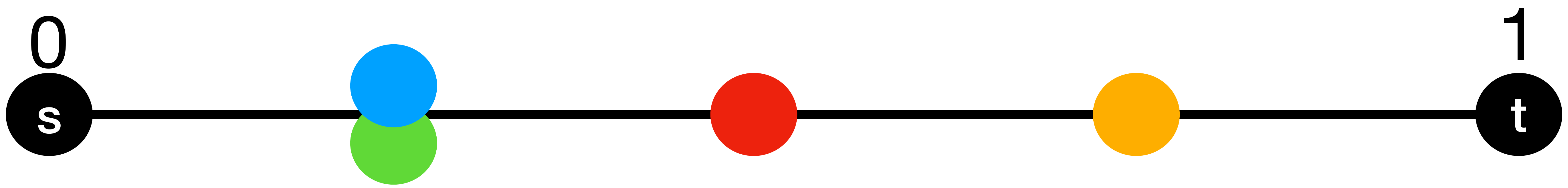
5 Consider an undirected graph $G = (V, E)$ and let $s \neq t \in V$. Show that there is an s,t-cut of value at most the optimal value of the following linear program

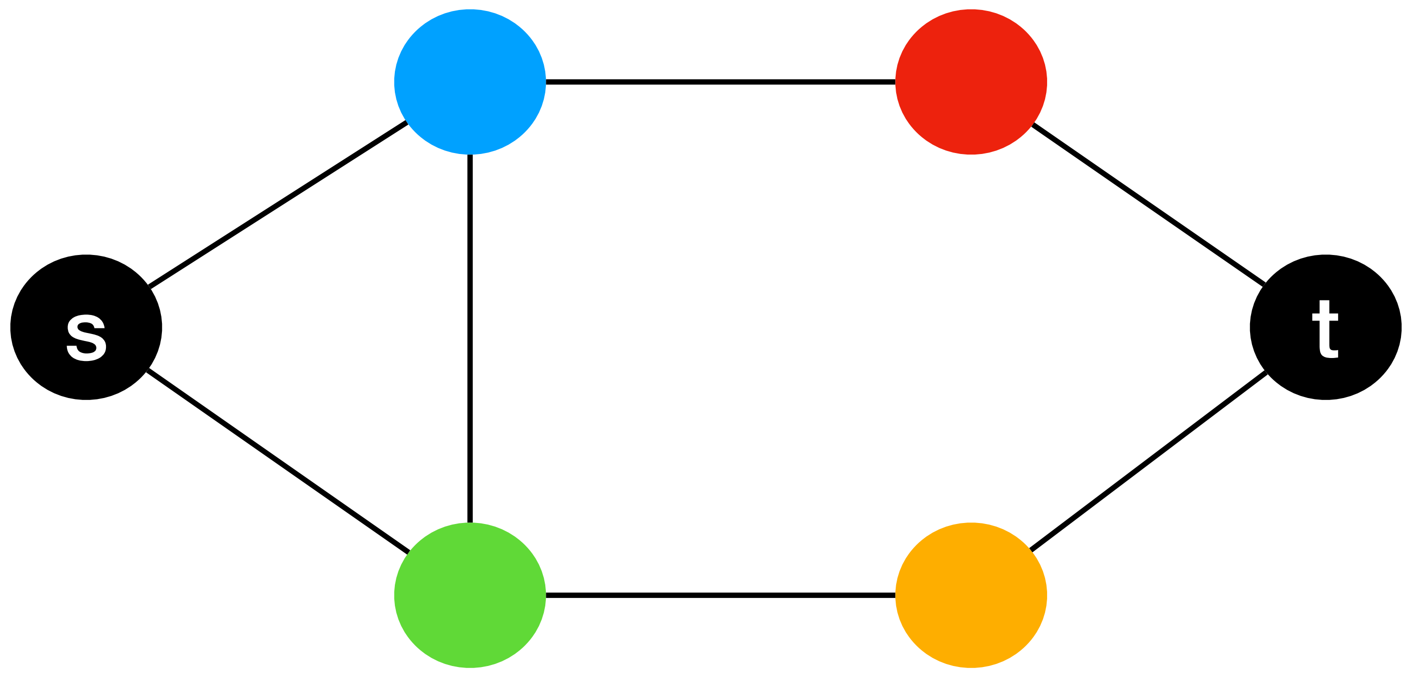
$$\begin{aligned} \textbf{minimize} \quad & \sum_{\{u,v\} \in E} |x_u - x_v| \\ \textbf{subject to} \quad & x_s = 0, x_t = 1, \text{ and } x_v \in [0,1] \text{ for every } v \in V \end{aligned}$$



5 Consider an undirected graph $G = (V, E)$ and let $s \neq t \in V$. Show that there is an s,t-cut of value at most the optimal value of the following linear program

$$\begin{aligned} \textbf{minimize} \quad & \sum_{\{u,v\}\in E} |x_u - x_v| \\ \textbf{subject to} \quad & x_s = 0, x_t = 1, \text{ and } x_v \in [0,1] \text{ for every } v \in V \end{aligned}$$

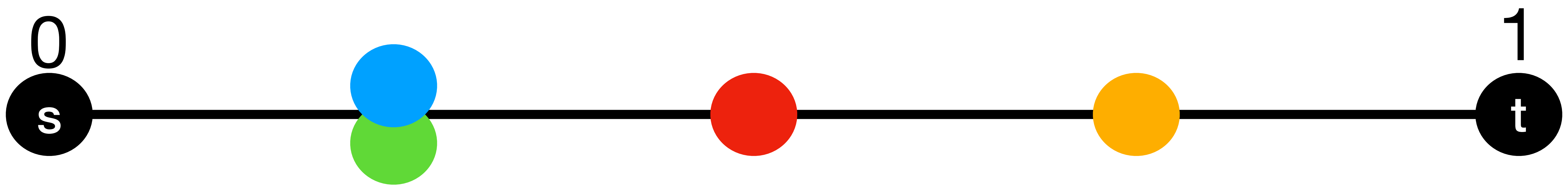


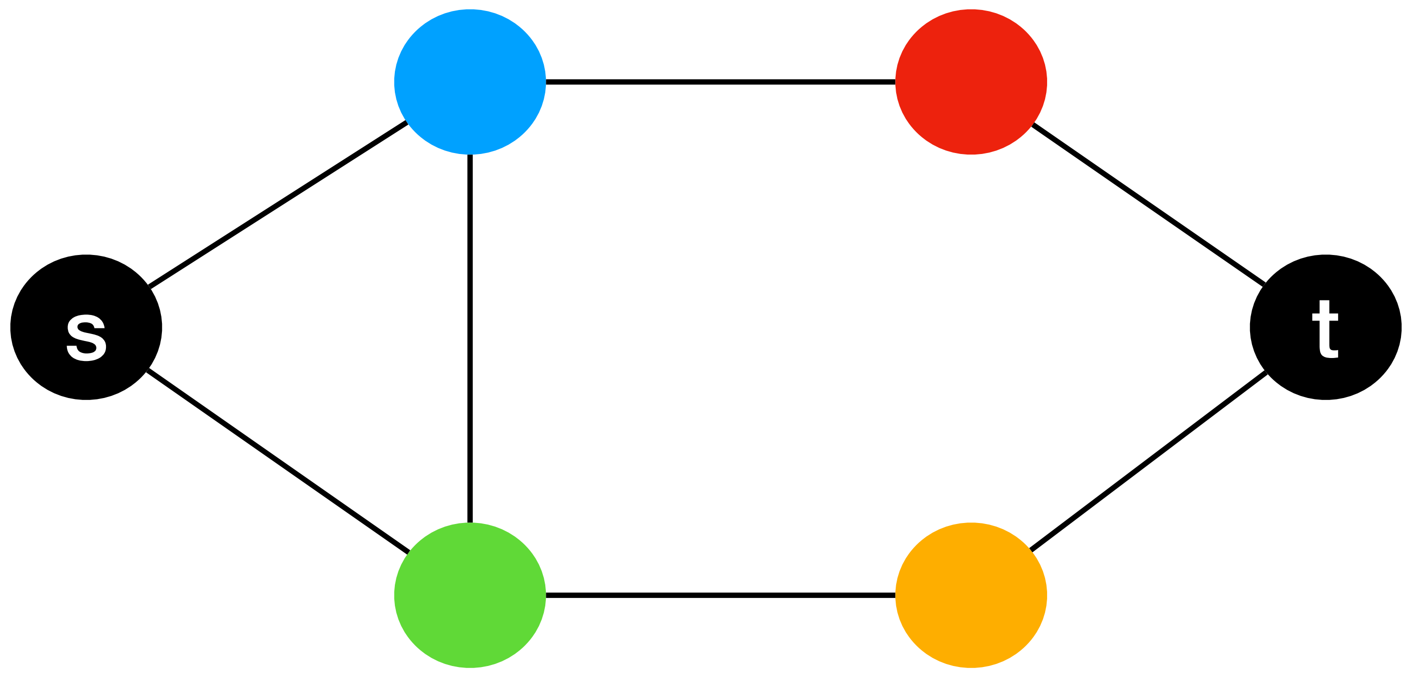


5 Consider an undirected graph $G = (V, E)$ and let $s \neq t \in V$. Show that there is an s, t-cut of value at most the optimal value of the following linear program

$$\textbf{minimize} \quad \sum_{\{u,v\} \in E} |x_u - x_v|$$

$$\textbf{subject to} \quad x_s = 0, x_t = 1, \text{ and } x_v \in [0,1] \text{ for every } v \in V$$

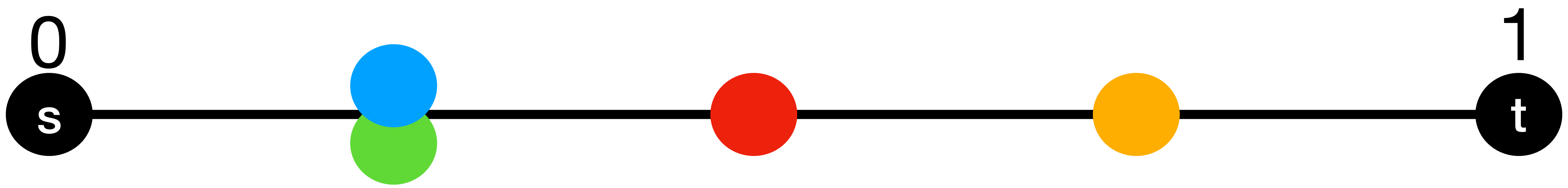


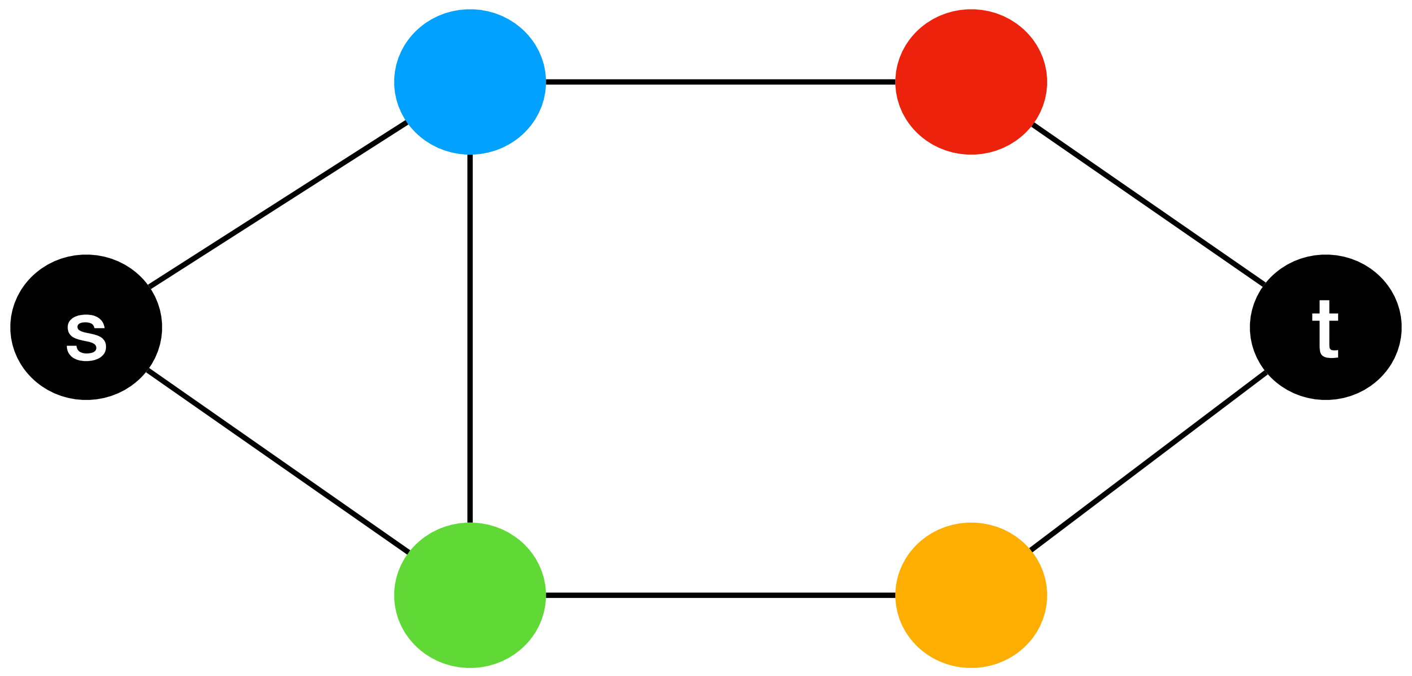


- Select $\Theta \sim [0,1]$ uniformly at random, output cut $S = \{u \in V : x_u \leq \theta\}$

5 Consider an undirected graph $G = (V, E)$ and let $s \neq t \in V$. Show that there is an s, t-cut of value at most the optimal value of the following linear program

$$\begin{array}{ll} \textbf{minimize} & \sum_{\{u,v\} \in E} |x_u - x_v| \\ \textbf{subject to} & x_s = 0, x_t = 1, \text{ and } x_v \in [0,1] \text{ for every } v \in V \end{array}$$

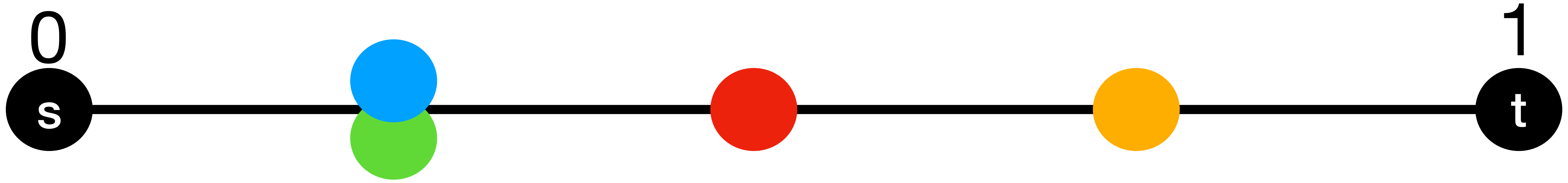


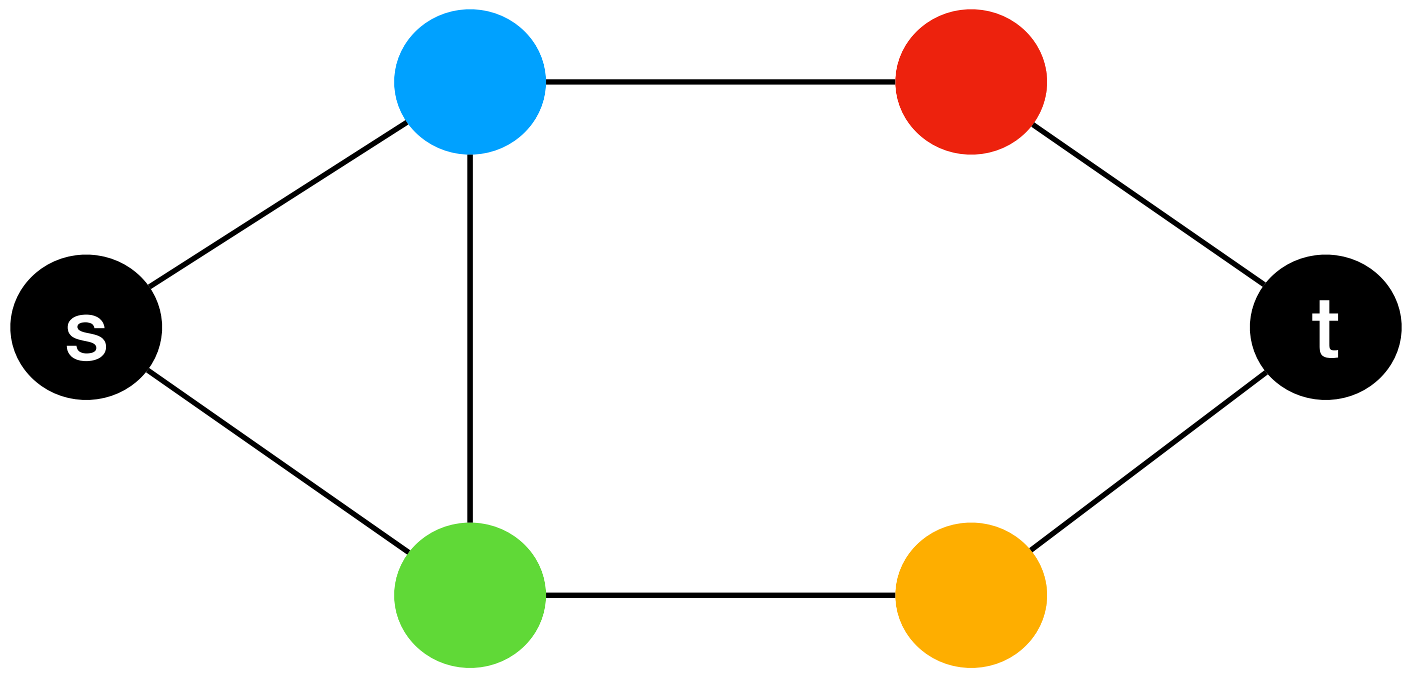


- Select $\Theta \sim [0,1]$ uniformly at random, output cut $S = \{u \in V : x_u \leq \theta\}$
- Probability edge (●, ●) cut equals |x(●) - x(●)|

5 Consider an undirected graph $G = (V, E)$ and let $s \neq t \in V$. Show that there is an s, t-cut of value at most the optimal value of the following linear program

$$\begin{aligned} \textbf{minimize} \quad & \sum_{\{u,v\} \in E} |x_u - x_v| \\ \textbf{subject to} \quad & x_s = 0, x_t = 1, \text{ and } x_v \in [0,1] \text{ for every } v \in V \end{aligned}$$

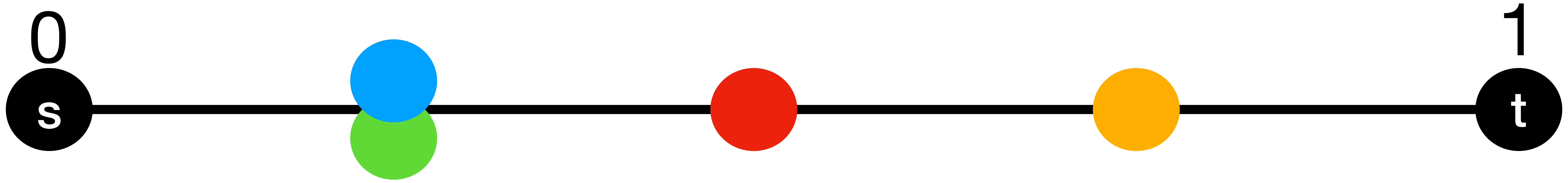


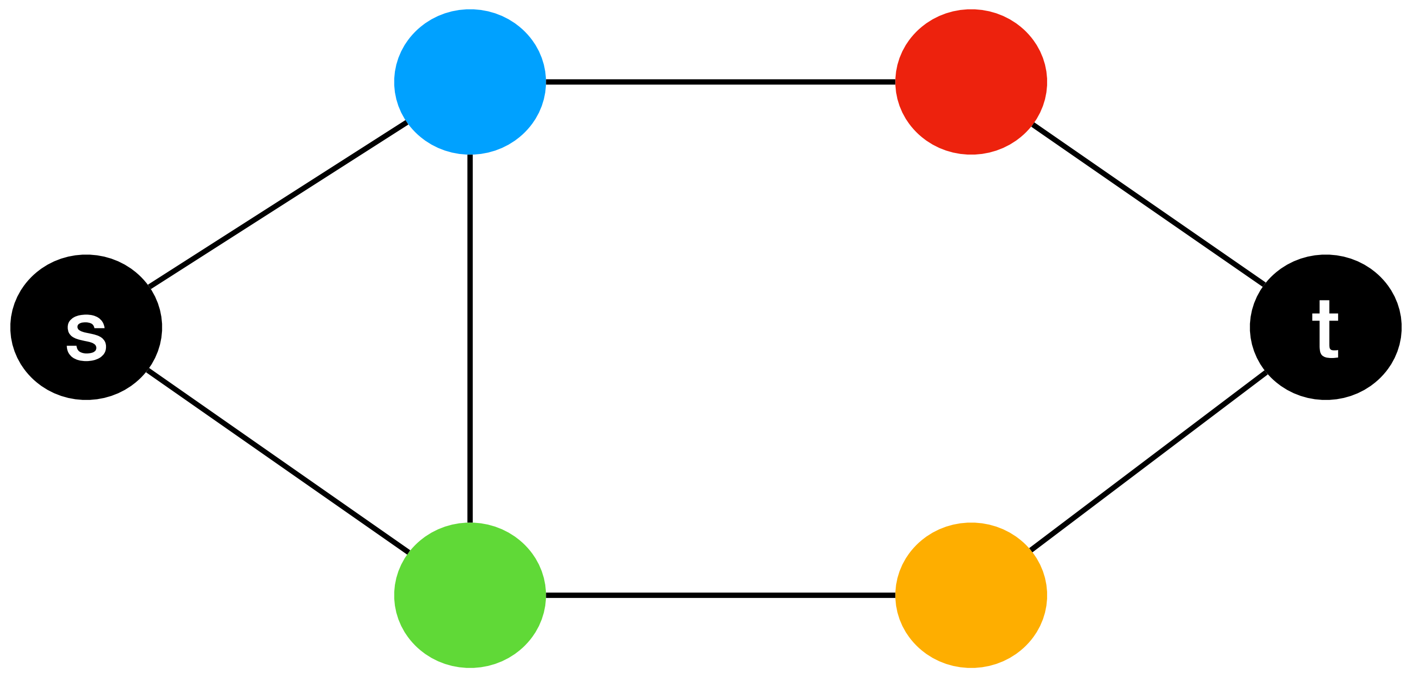


- Select $\Theta \sim [0,1]$ uniformly at random, output cut $S = \{u \in V : x_u \leq \theta\}$
- Probability edge (●, ●) cut equals |x(●) - x(●)|
- Expected value of cut = value of LP solution

5 Consider an undirected graph $G = (V, E)$ and let $s \neq t \in V$. Show that there is an s,t-cut of value at most the optimal value of the following linear program

$$\begin{aligned} \textbf{minimize} \quad & \sum_{\{u,v\} \in E} |x_u - x_v| \\ \textbf{subject to} \quad & x_s = 0, x_t = 1, \text{ and } x_v \in [0,1] \text{ for every } v \in V \end{aligned}$$





- Select $\Theta \sim [0,1]$ **uniformly at random**, output cut $S = \{u \in V : x_u \leq \theta\}$
- Probability edge (●, ●) cut equals |x(●) - x(●)|
- Expected value of cut = value of LP solution

The algorithm of MDRF

- While there is a leaf with more than one center, **select a min-cut**

TCS-Algorithm

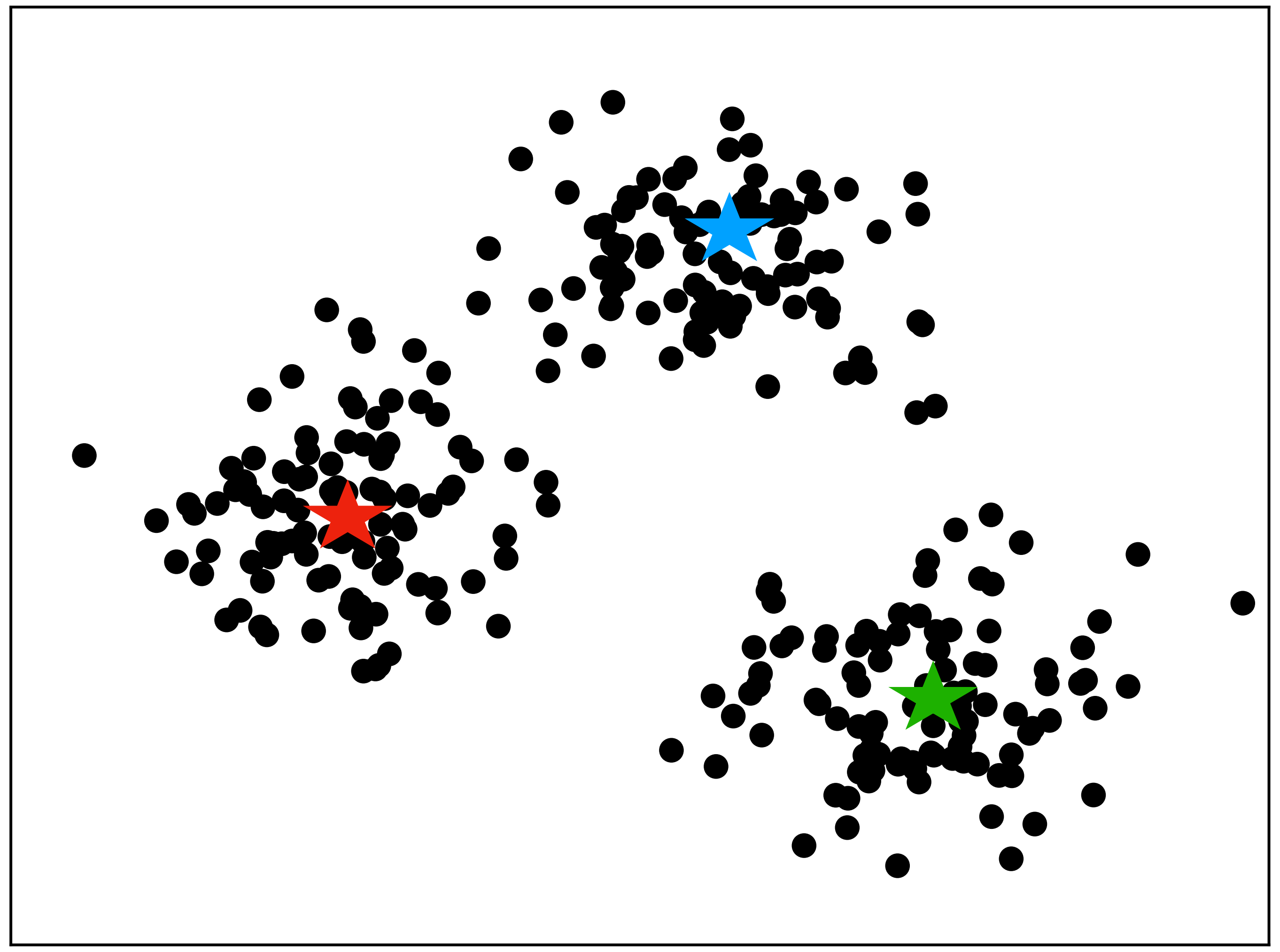
- While there is a leaf with more than one center, ~~**select a min-cut**~~

  select a uniformly at random cut

TCS-Algorithm

- While there is a leaf with more than one center, ~~**select a min-cut**~~

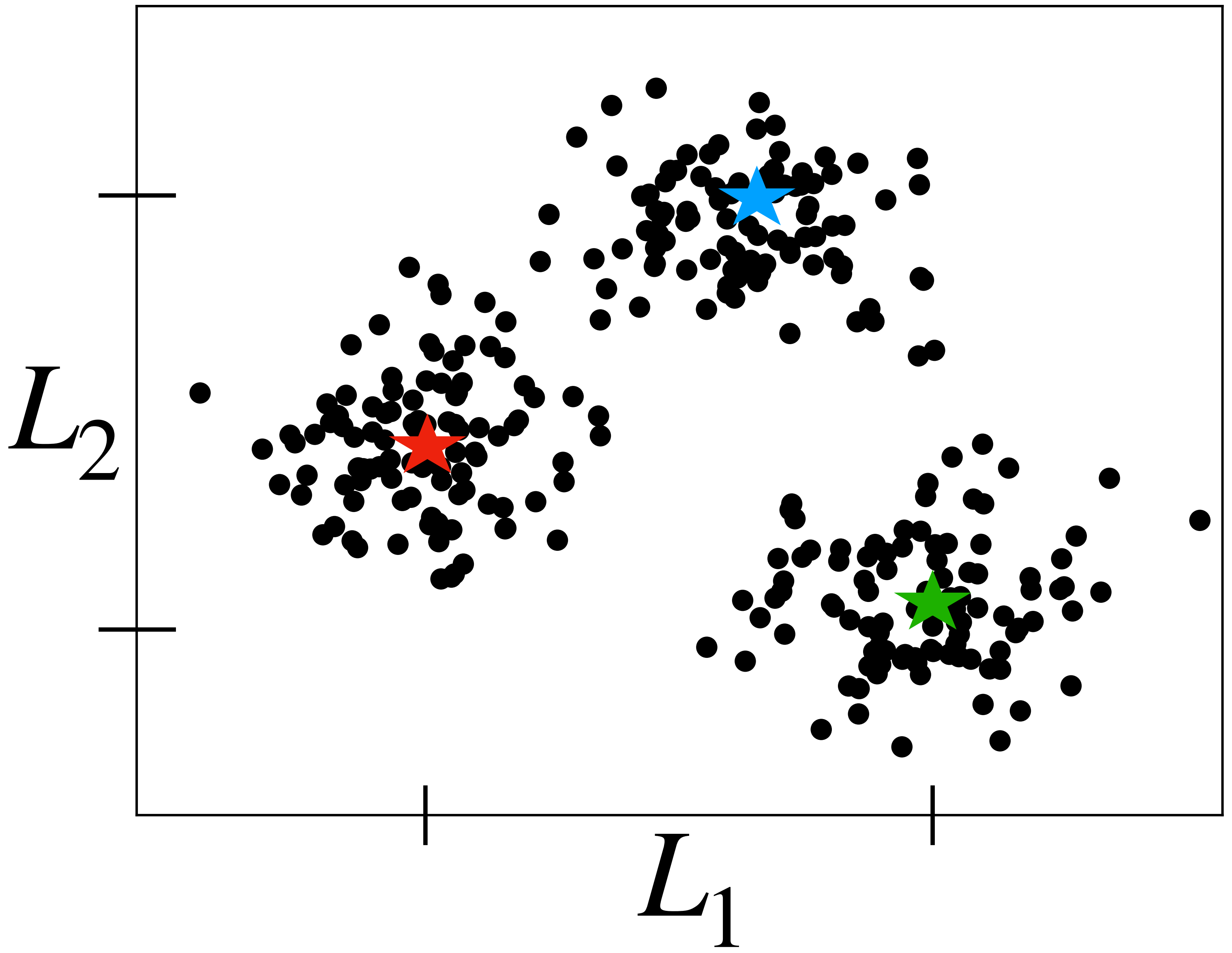
select a uniformly at random cut



TCS-Algorithm

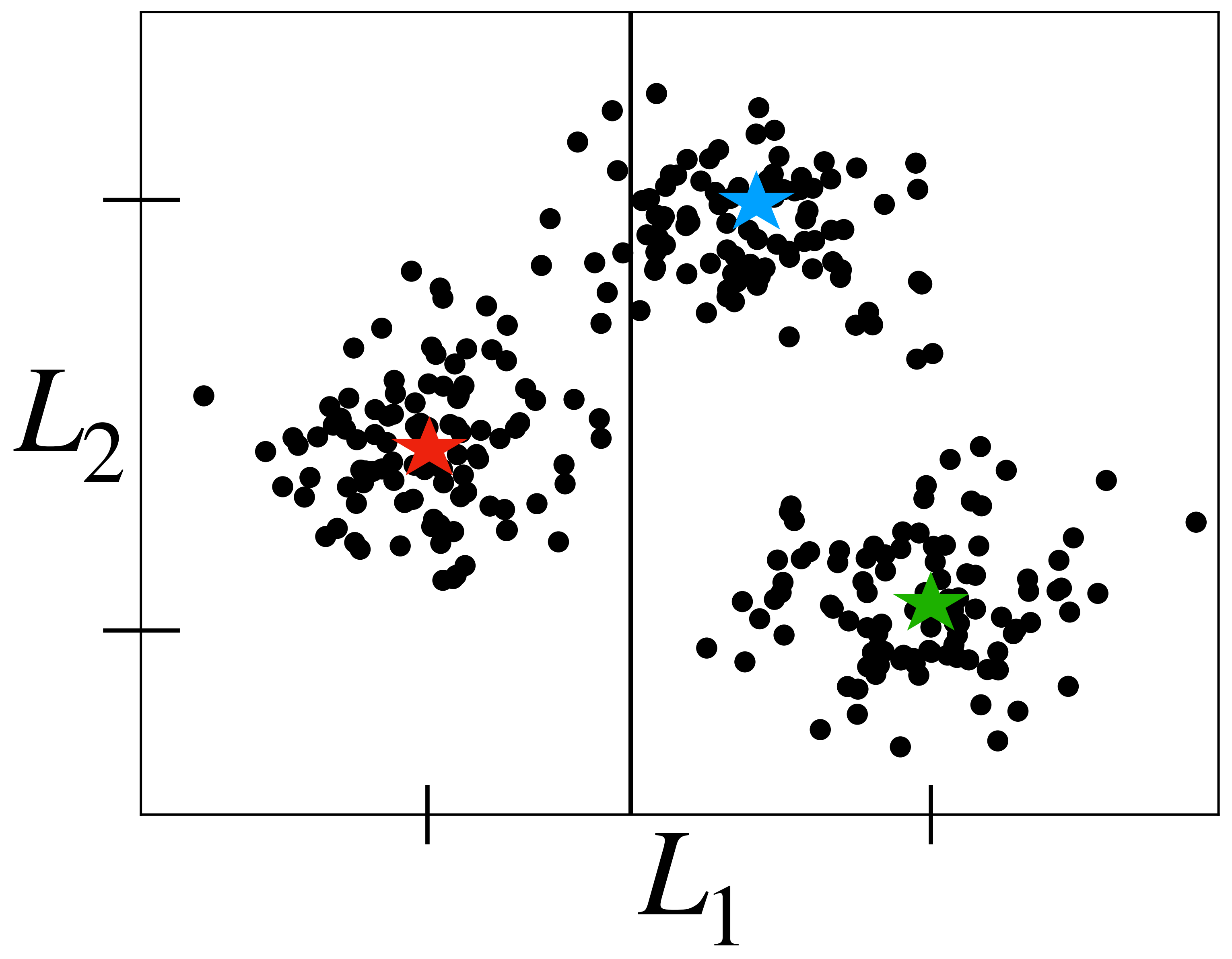
- While there is a leaf with more than one center, ~~**select a min-cut**~~

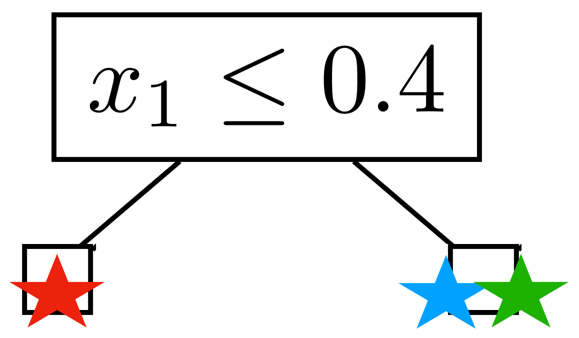
select a uniformly at random cut



TCS-Algorithm

- While there is a leaf with more than one center, ~~**select a min-cut**~~ **select a uniformly at random cut**

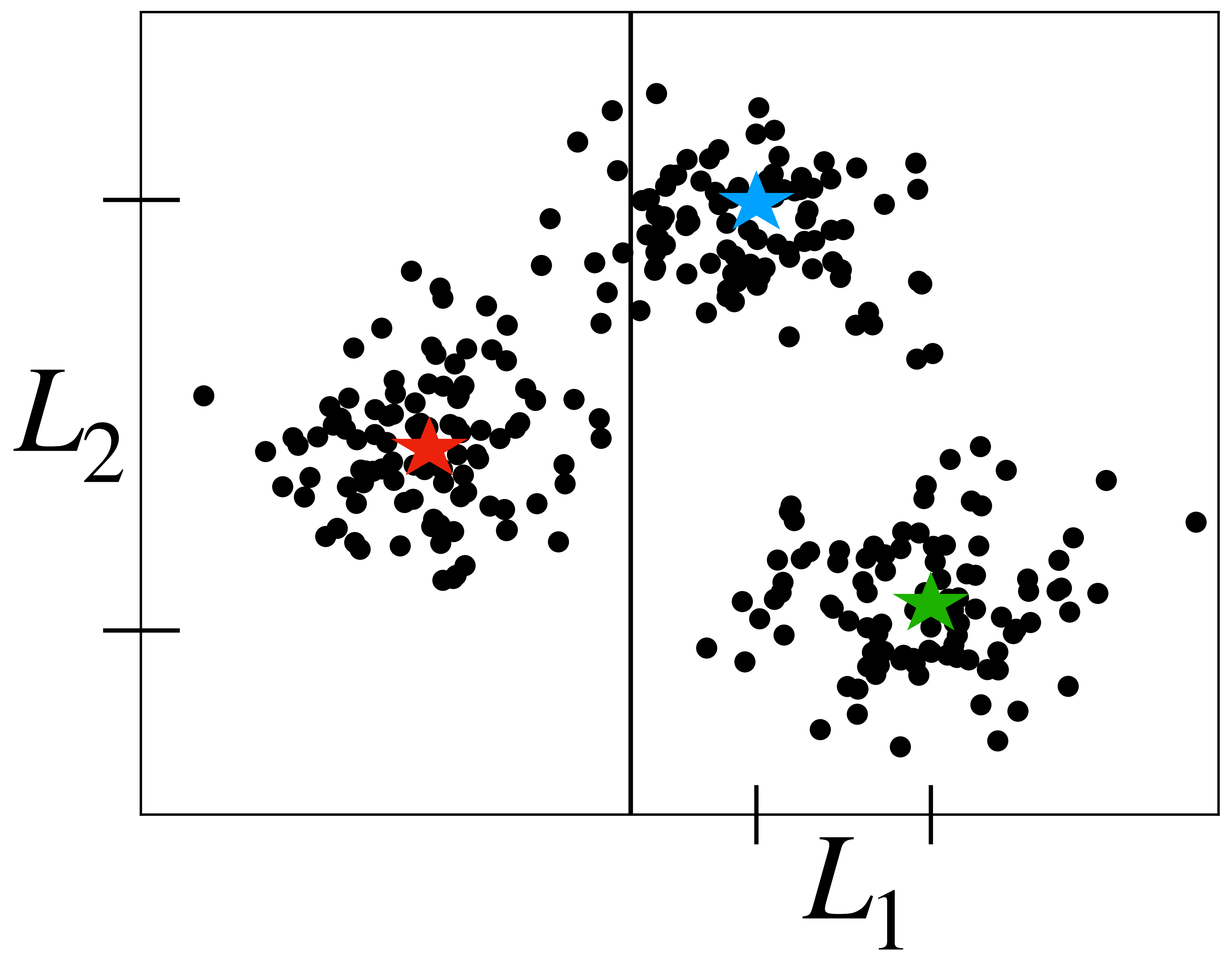


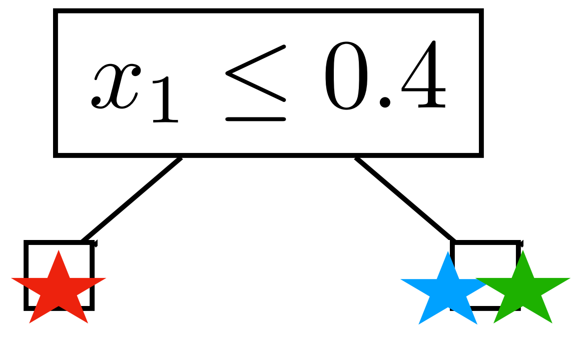


TCS-Algorithm

- While there is a leaf with more than one center, ~~**select a min-cut**~~

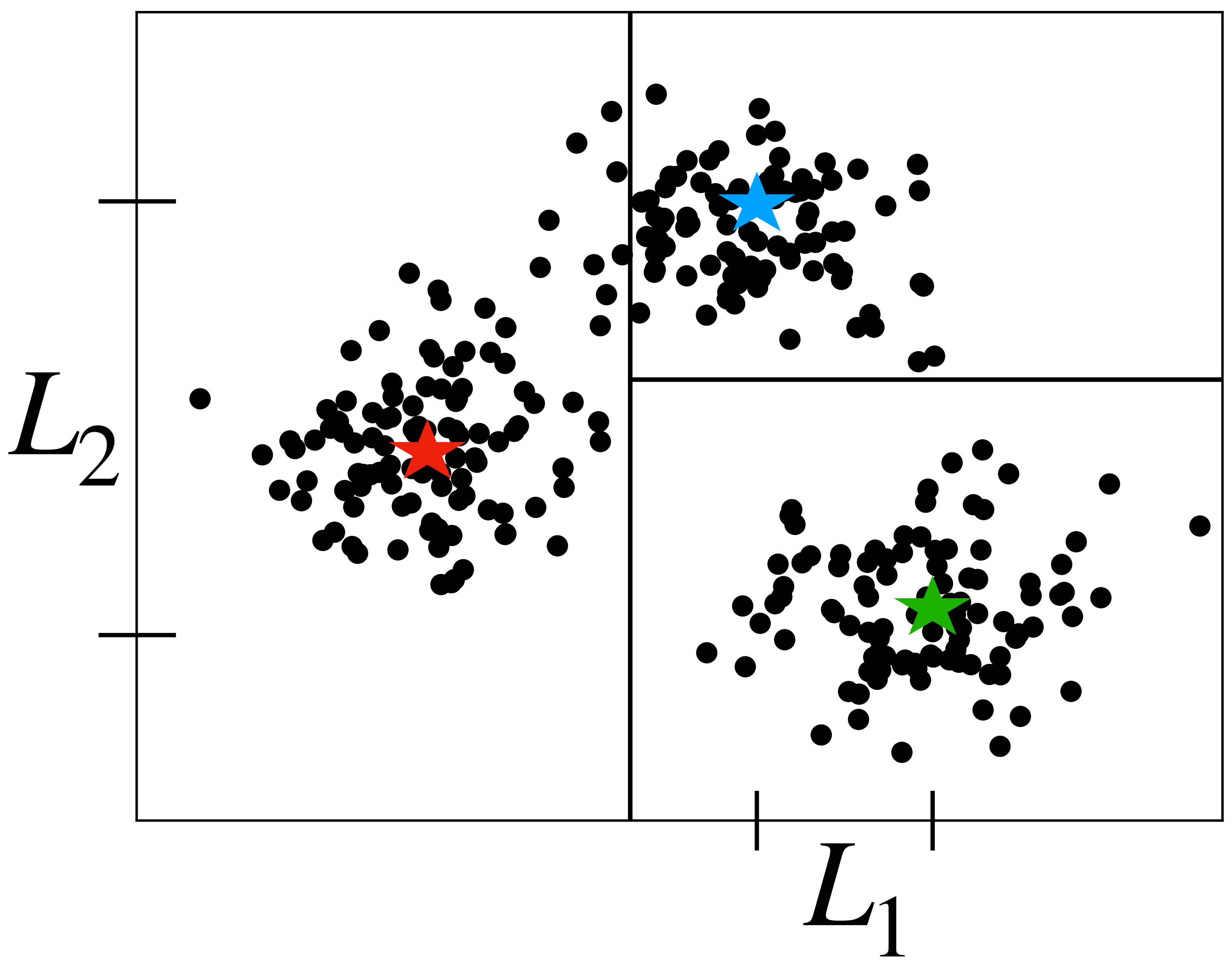
select a uniformly at random cut

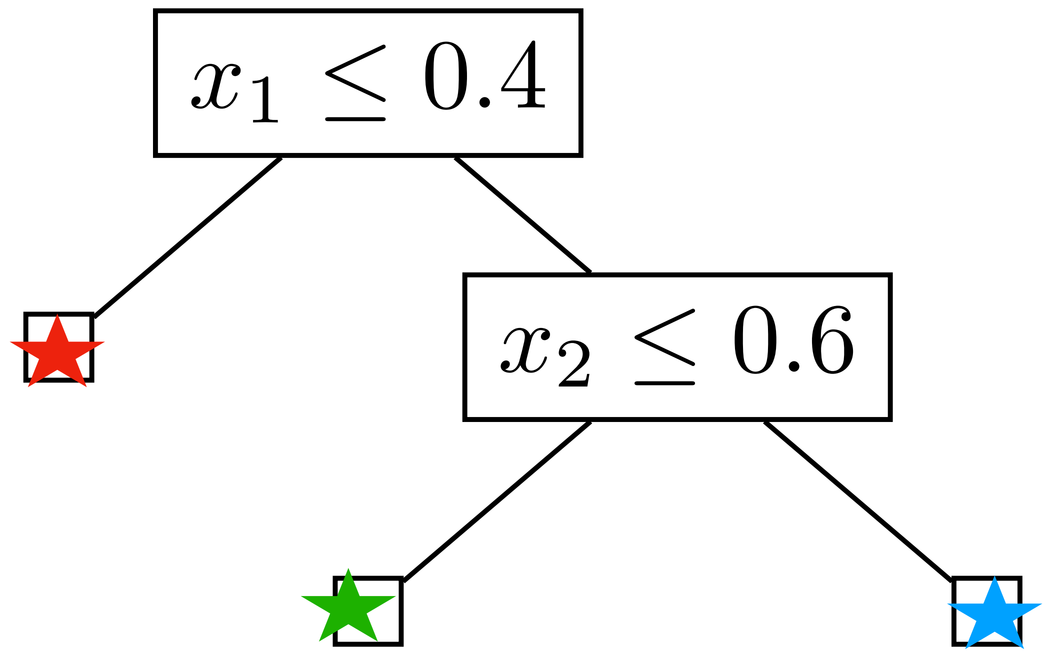




TCS-Algorithm

- While there is a leaf with more than one center, ~~select a min-cut~~ **select a uniformly at random cut**





The independent works in 2021

- Makarychev and Shan:
  - $O(\log k \log \log k)$
- Gamlath, Jia, Polak, Svensson:
  - $O(\log^2 k)$
- Esfandiari, Mirrokni, Narayanan:
  - $O(\min(\log k \log \log k, d \log^2 d))$

- Gupta, Pitty, Svensson, Yuan'23:

Theorem: The price of explainability given by *TCS-Algorithm* is $1 + H_{k-1}$

- Gupta, Pitty, Svensson, Yuan'23:

Theorem: The price of explainability given by *TCS-Algorithm* is $1 + H_{k-1}$

Theorem: The price of explainability is at least $(1 - \epsilon)\ln(k)$ for any $\epsilon > 0$

- Gupta, Pitty, Svensson, Yuan'23:

Theorem: The price of explainability given by *TCS-Algorithm* is $1 + H_{k-1}$

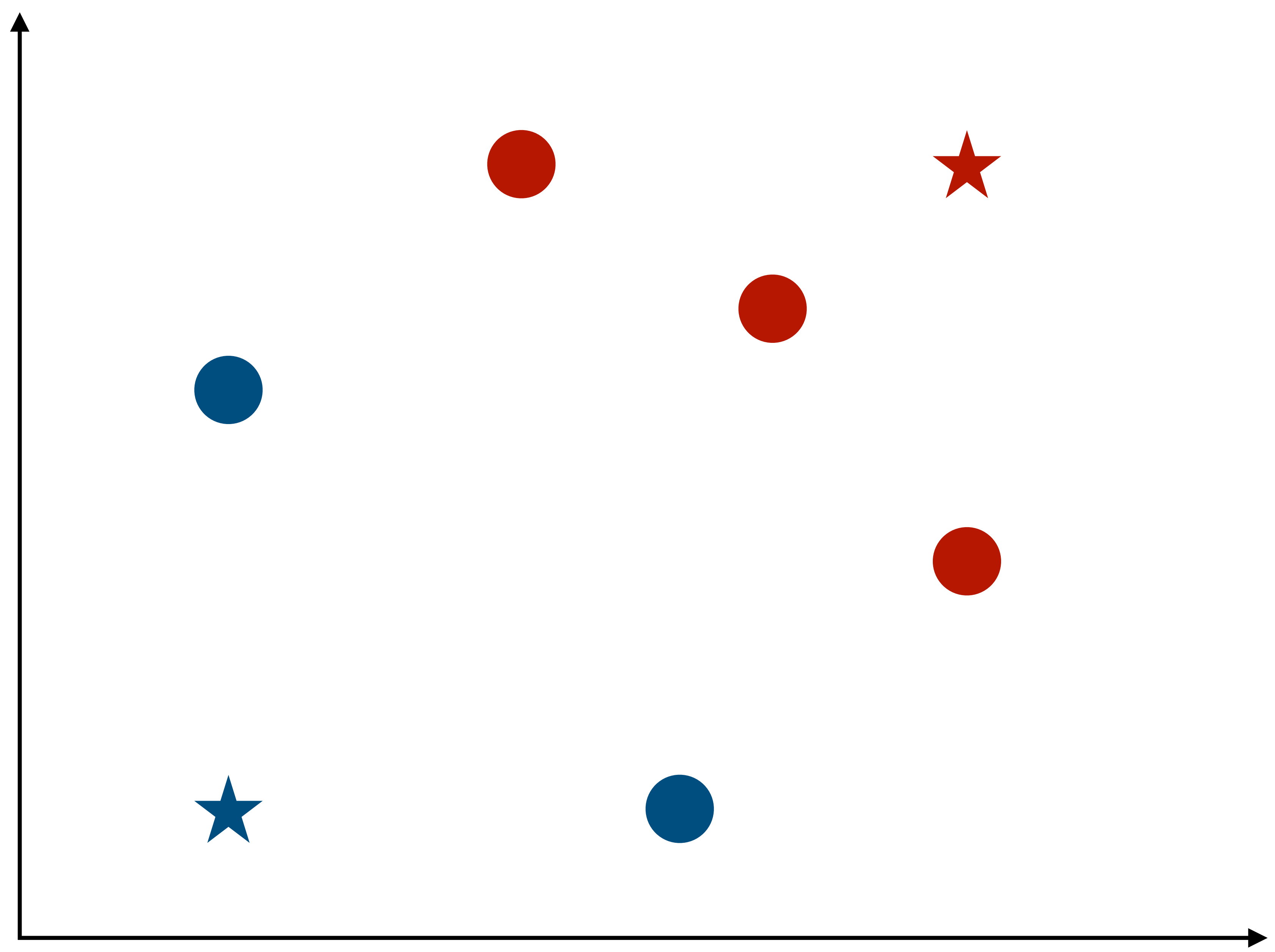
Theorem: The price of explainability is at least $(1 - \epsilon)\ln(k)$ for any $\epsilon > 0$

+ It is NP-hard to approximate explainable k-median better than $O(\ln k)$

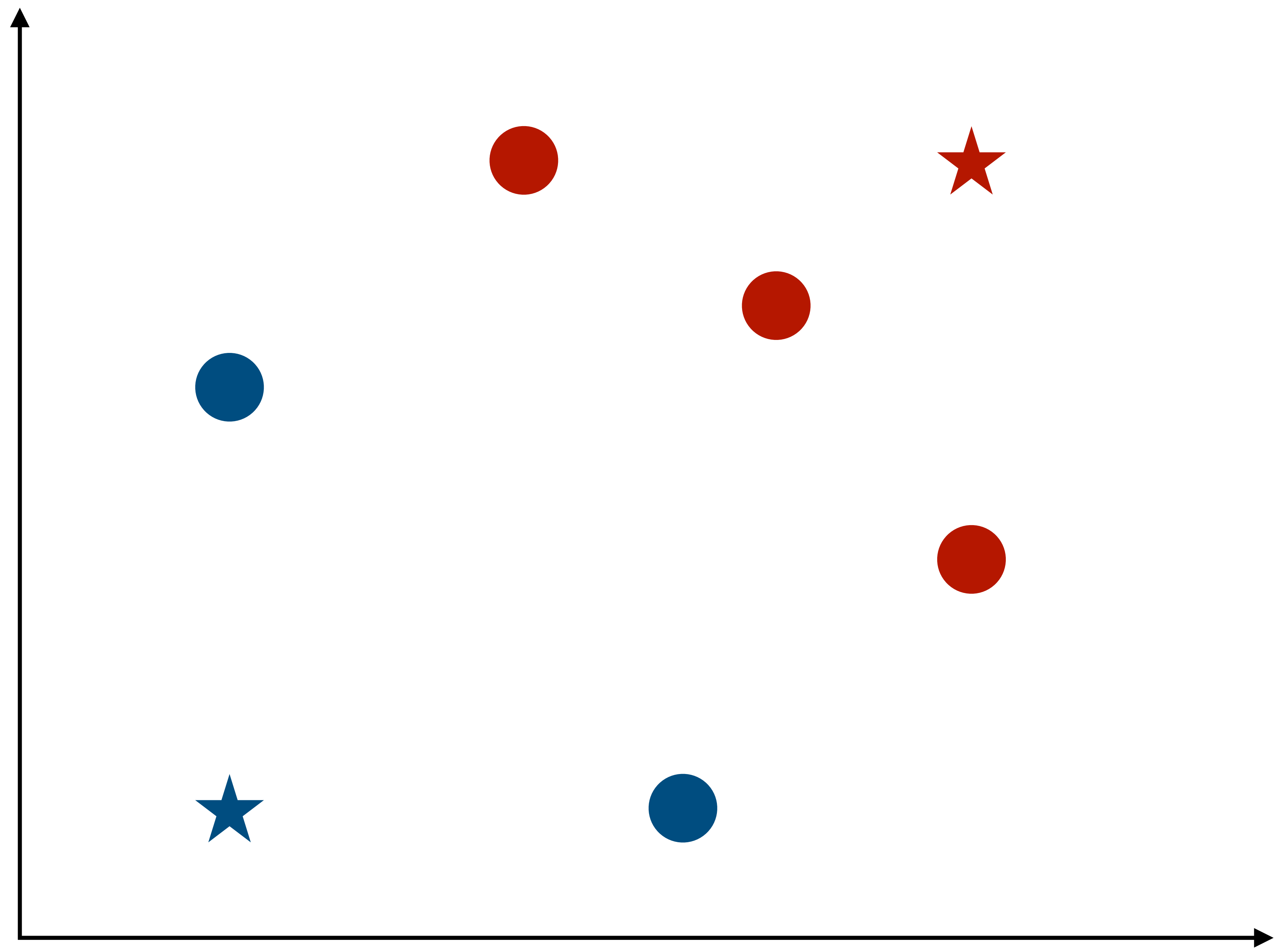
Ideas of analysis

- Enough to analyze the cost increase of a single point (by linearity of expectation)
- By translation, we may assume the point is located at the origin

Two clusters in two dimensions

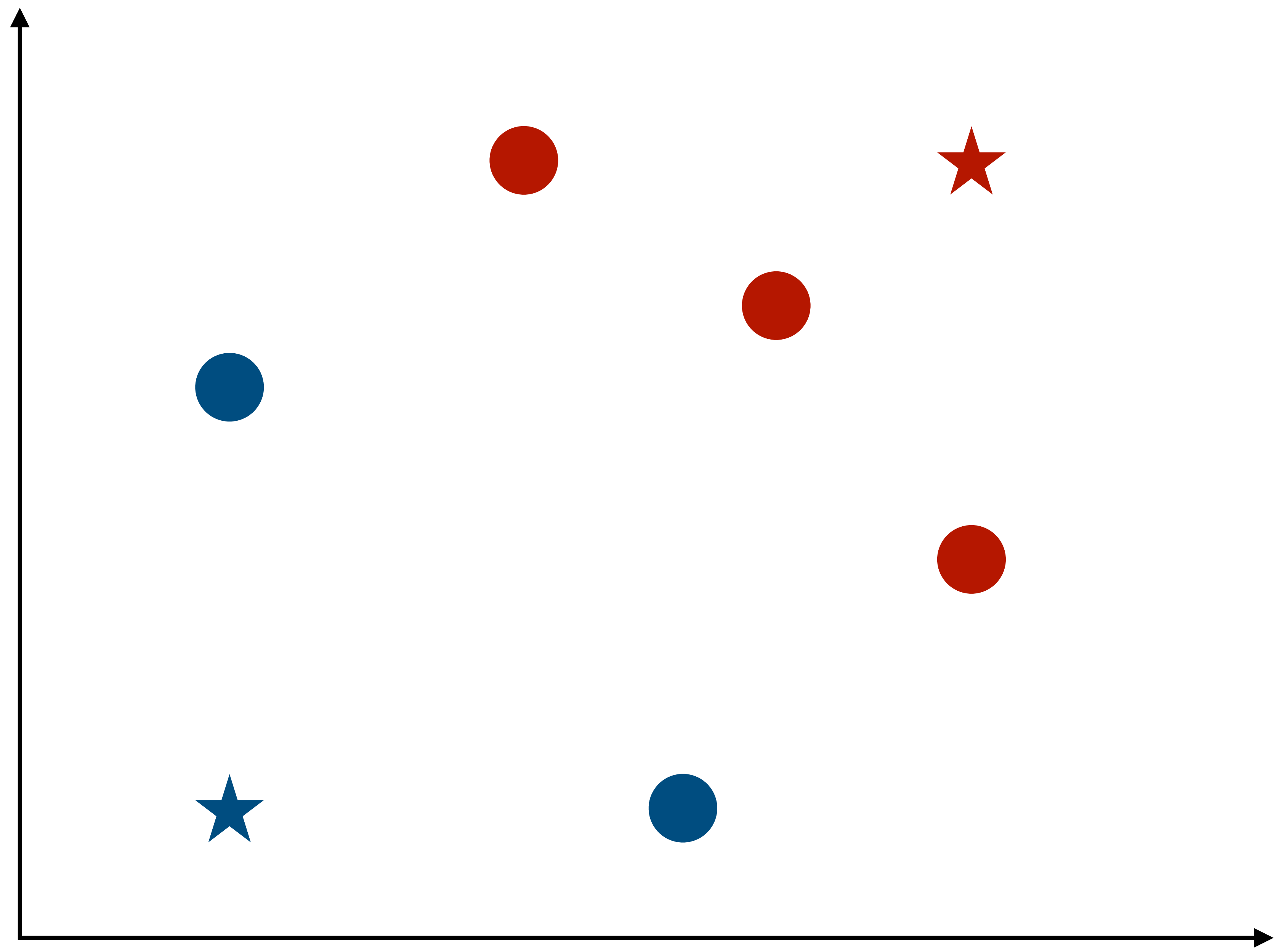


Two clusters in two dimensions

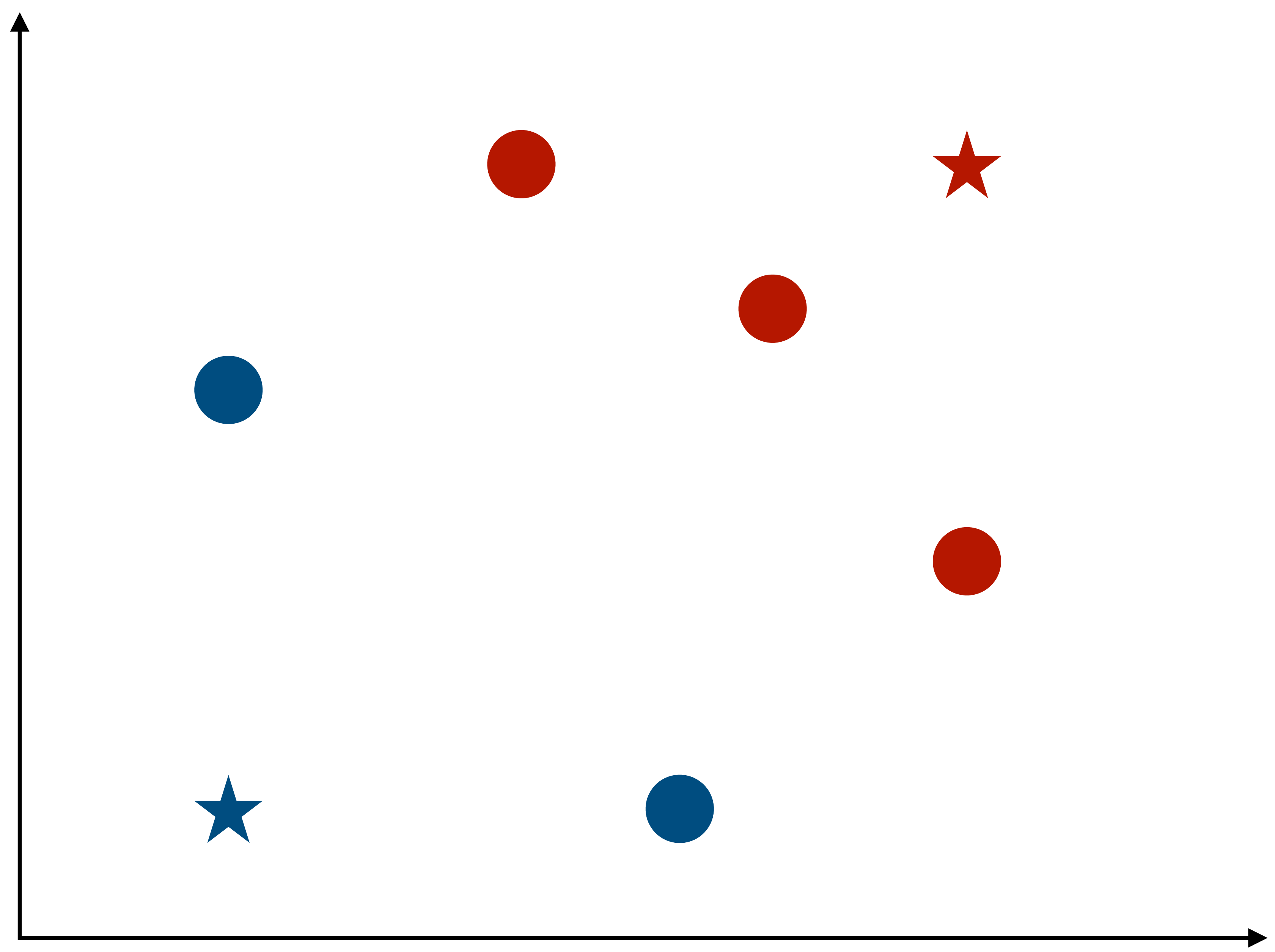


There is a price for explainability even in this simple case

Two clusters in two dimensions

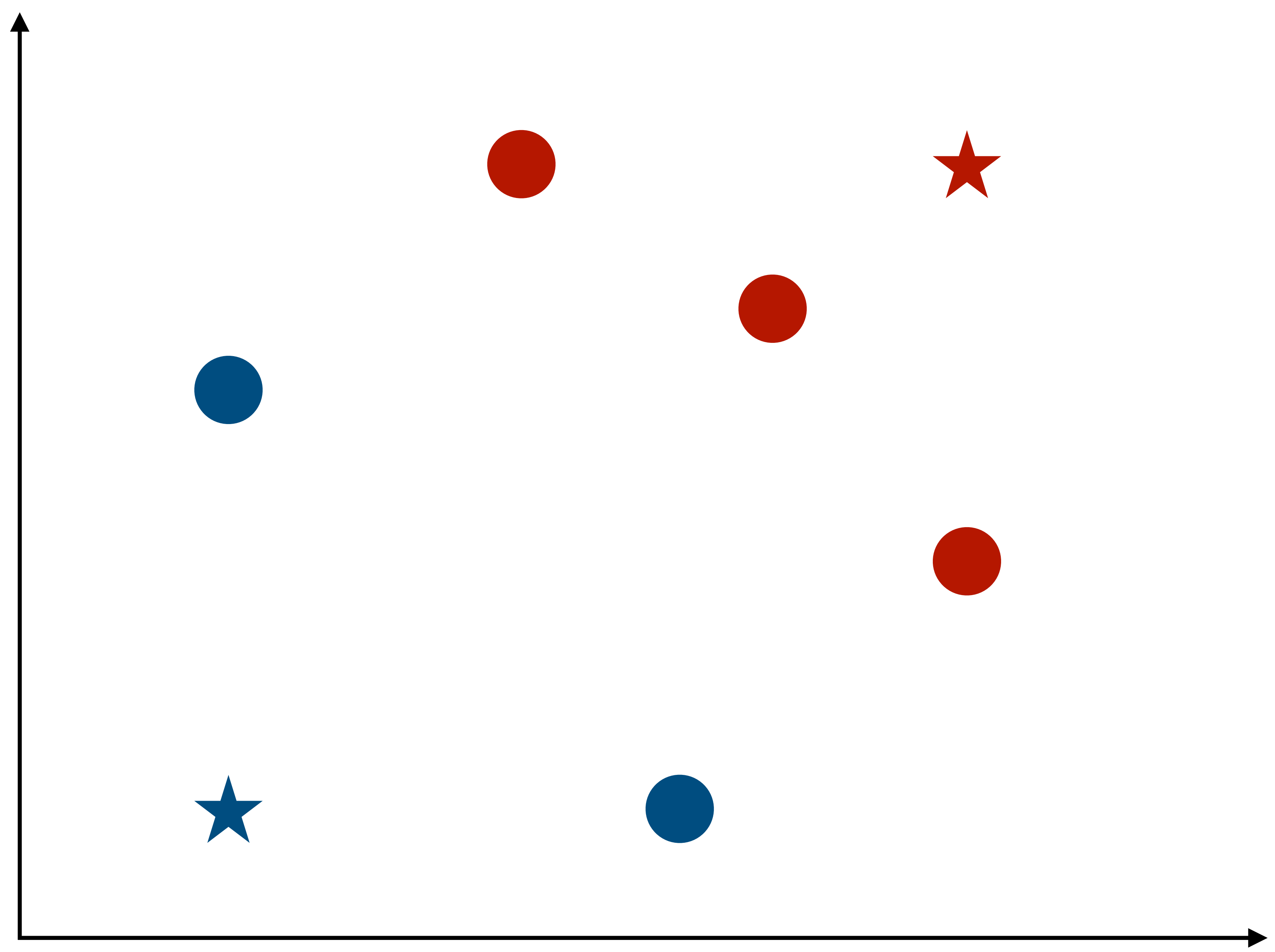


Two clusters in two dimensions

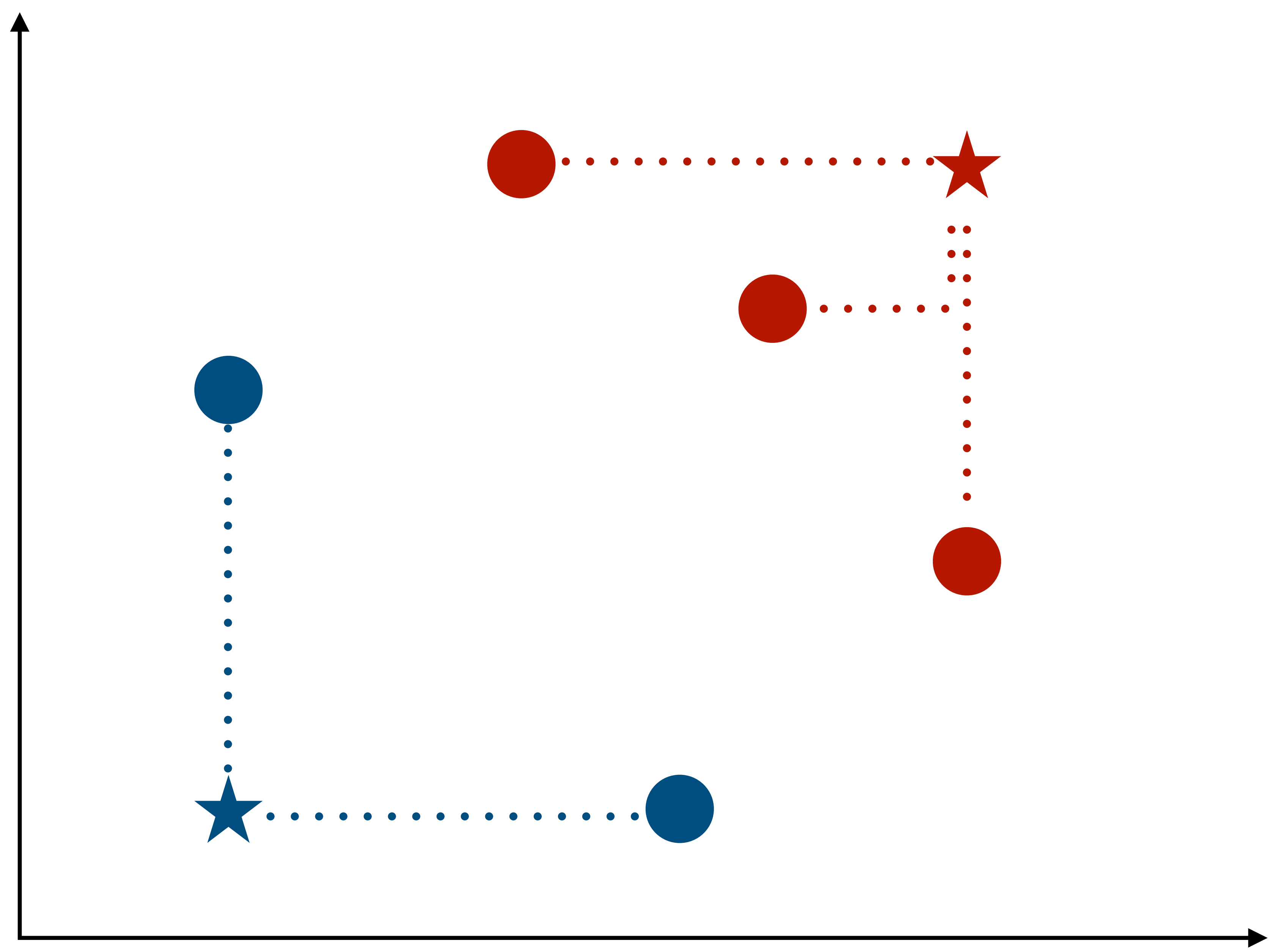


For two clusters, there is an explainable clustering of cost $\leq 2 \cdot OPT$

Two clusters in two dimensions

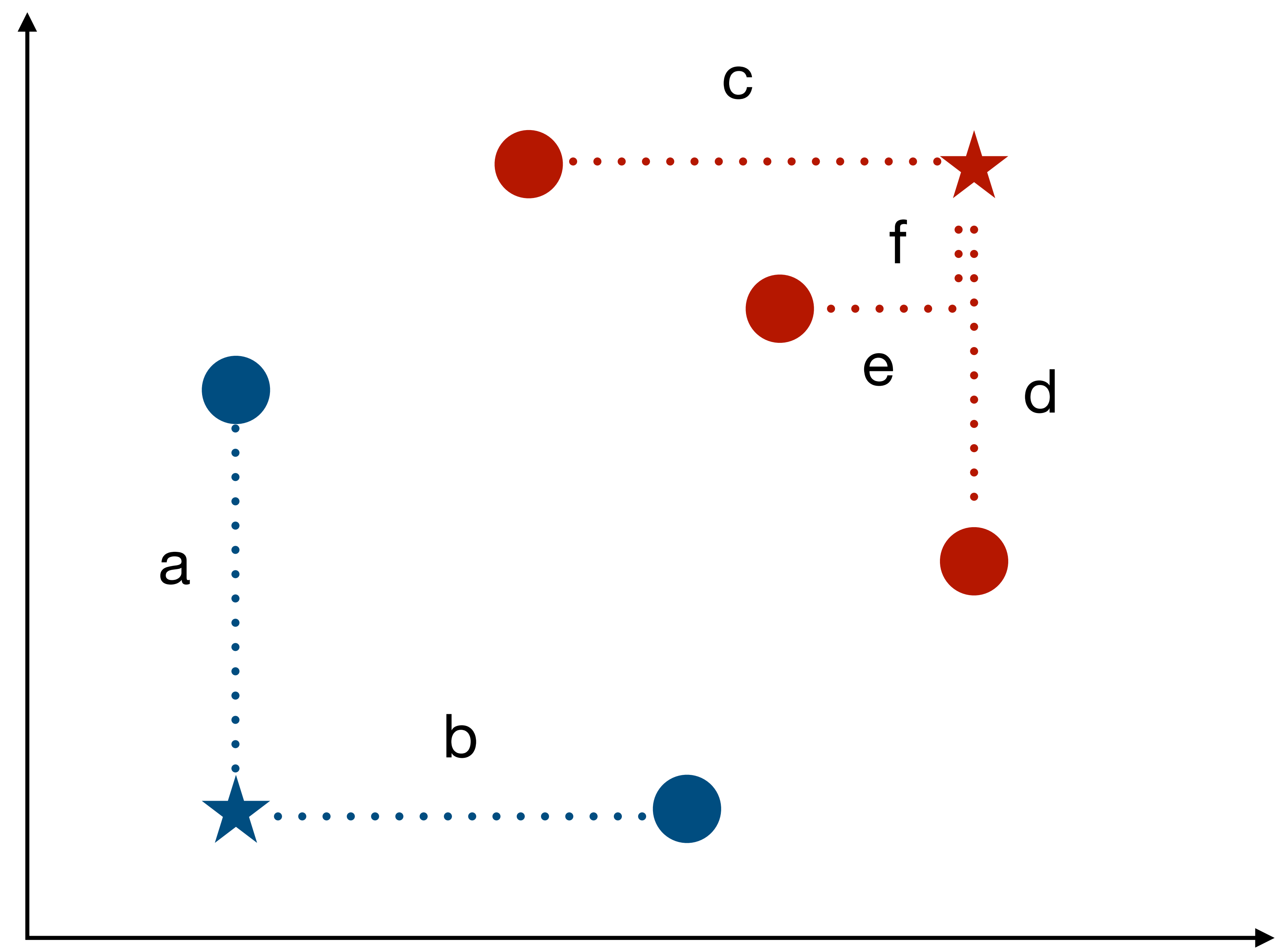


Two clusters in two dimensions



- Cost of optimal unconstrained clustering equals sum of dotted edges

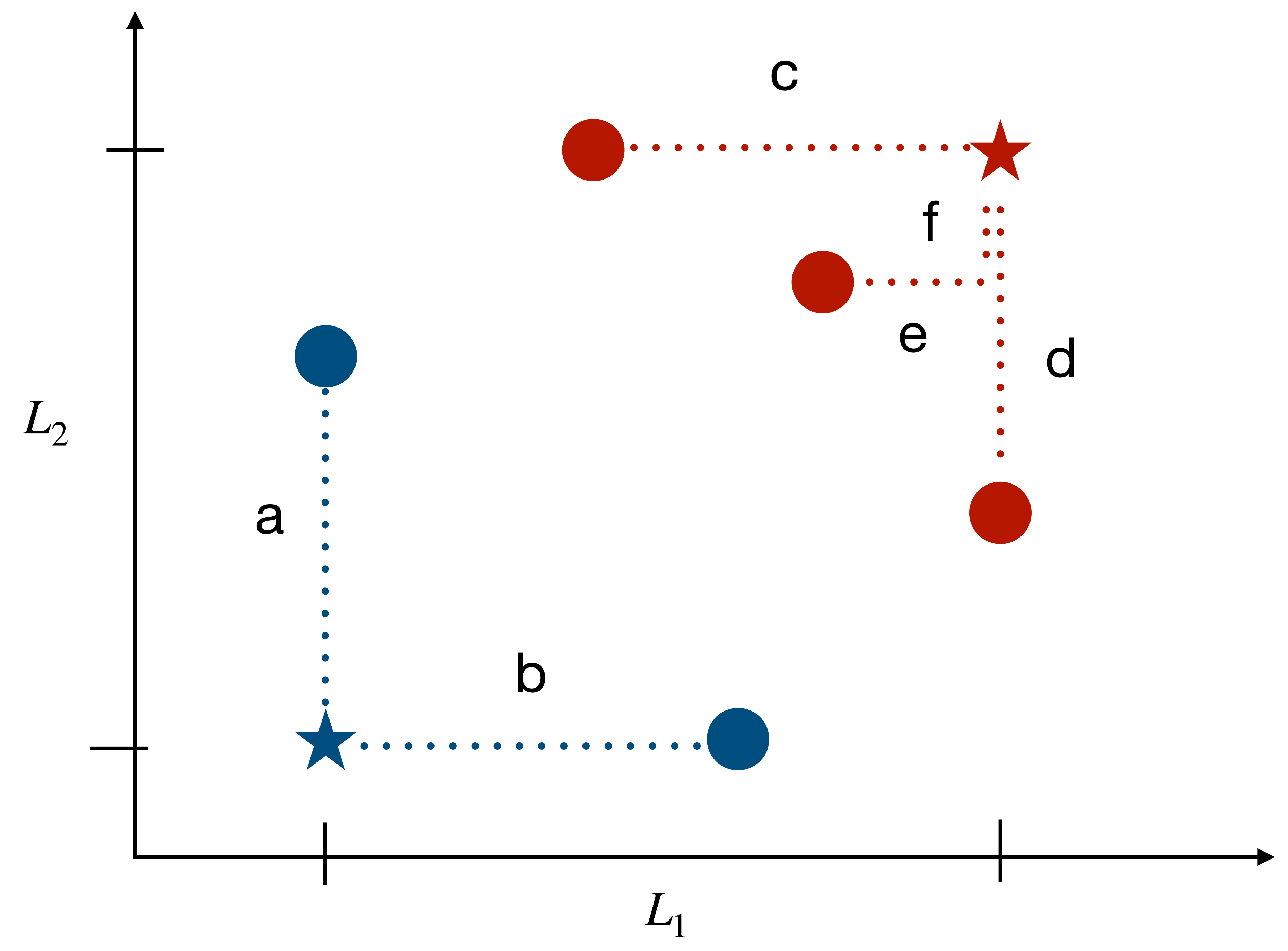
Two clusters in two dimensions



- Cost of optimal unconstrained clustering equals sum of dotted edges

$$OPT = a + b + c + d + e + f = \frac{a + b + c}{(L_1}$$

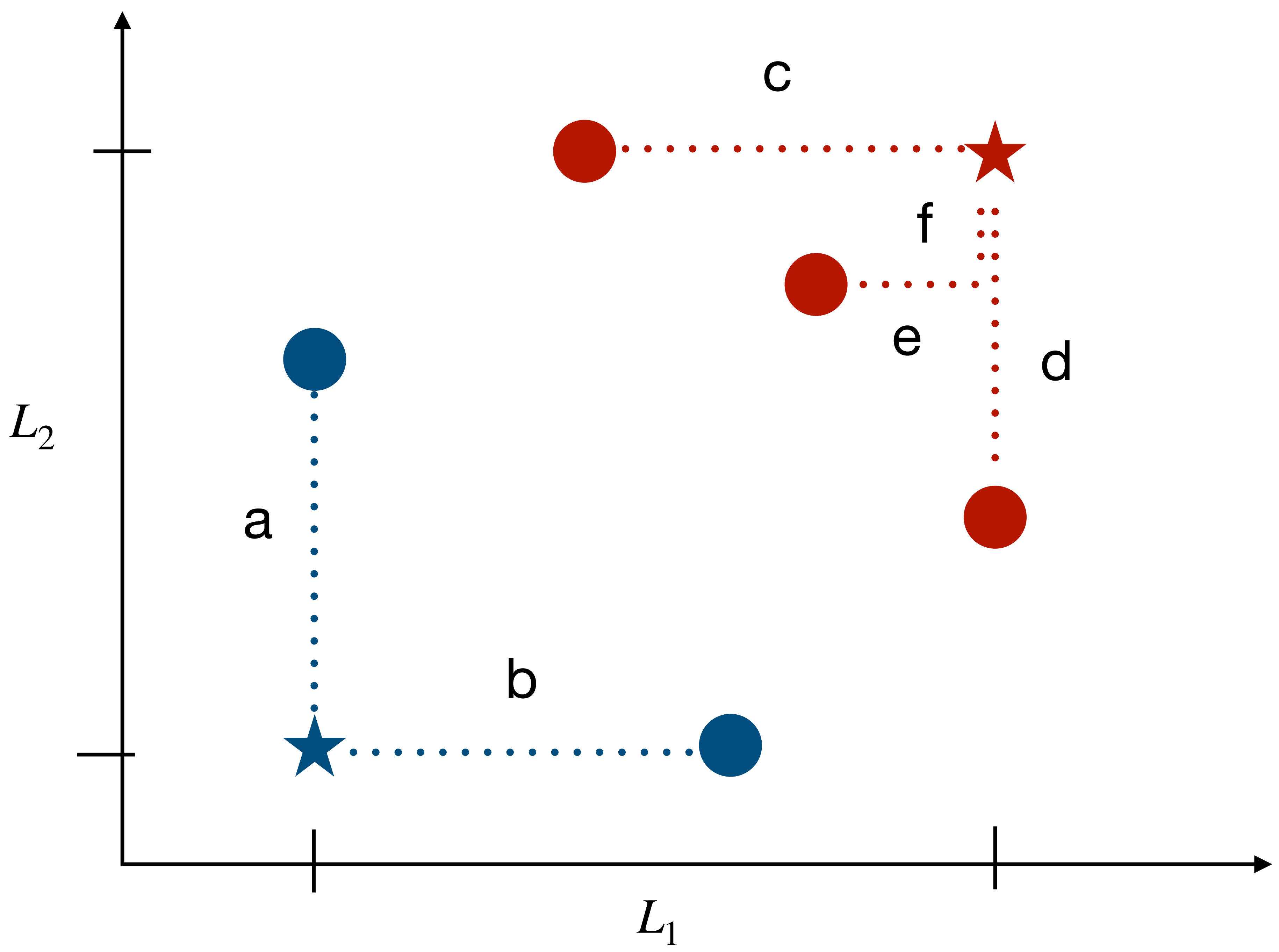
Two clusters in two dimensions



- Cost of optimal unconstrained clustering equals sum of dotted edges

$$OPT = a + b + c + d + e + f = \frac{a + b + c}{(L_1}$$

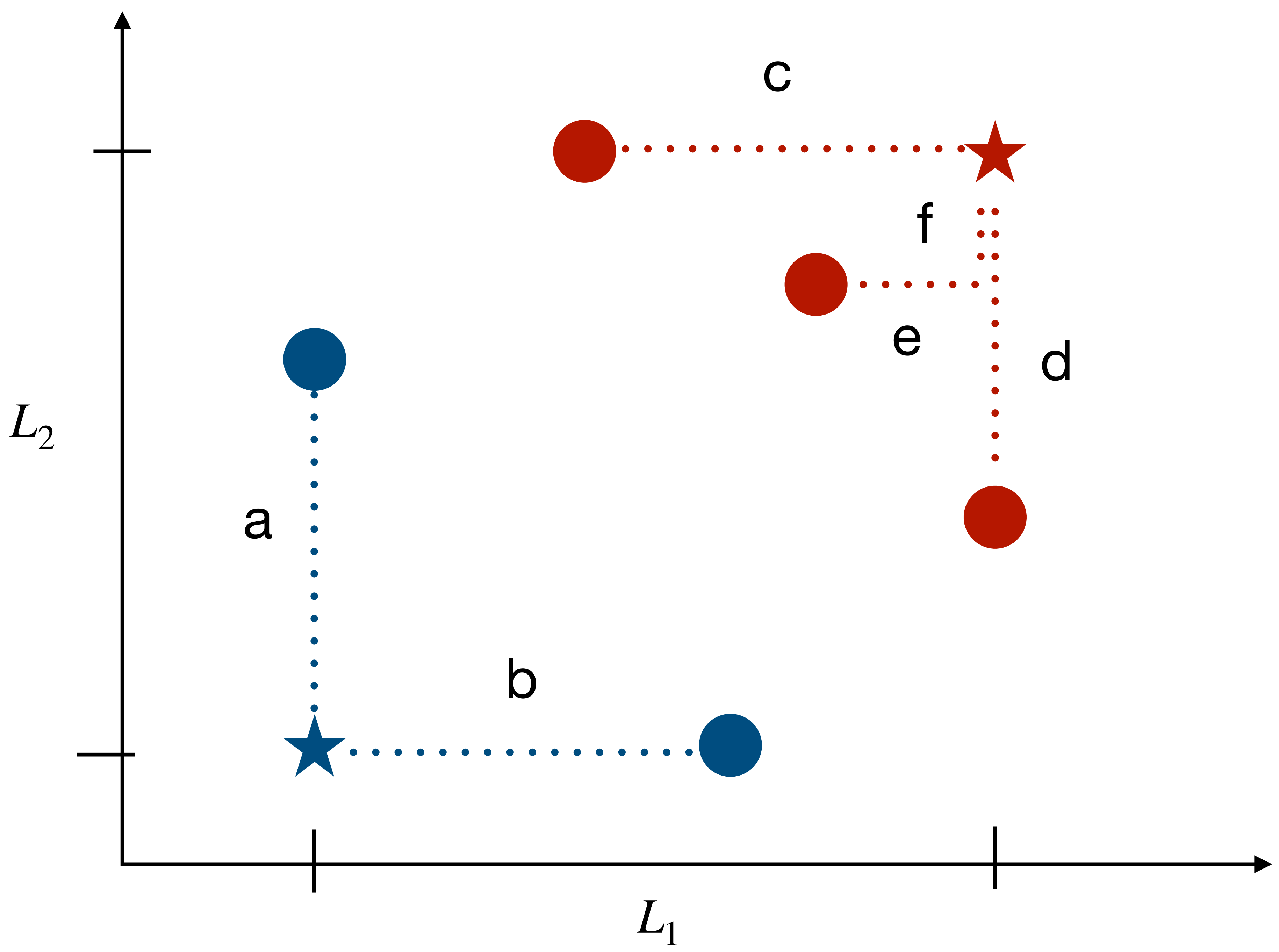
Two clusters in two dimensions



- Cost of optimal unconstrained clustering equals sum of dotted edges

$$OPT = a + b + c + d + e + f = \frac{a + b + c + d + e + f}{(L_1 + L_2)} \cdot (L_1 + L_2)$$

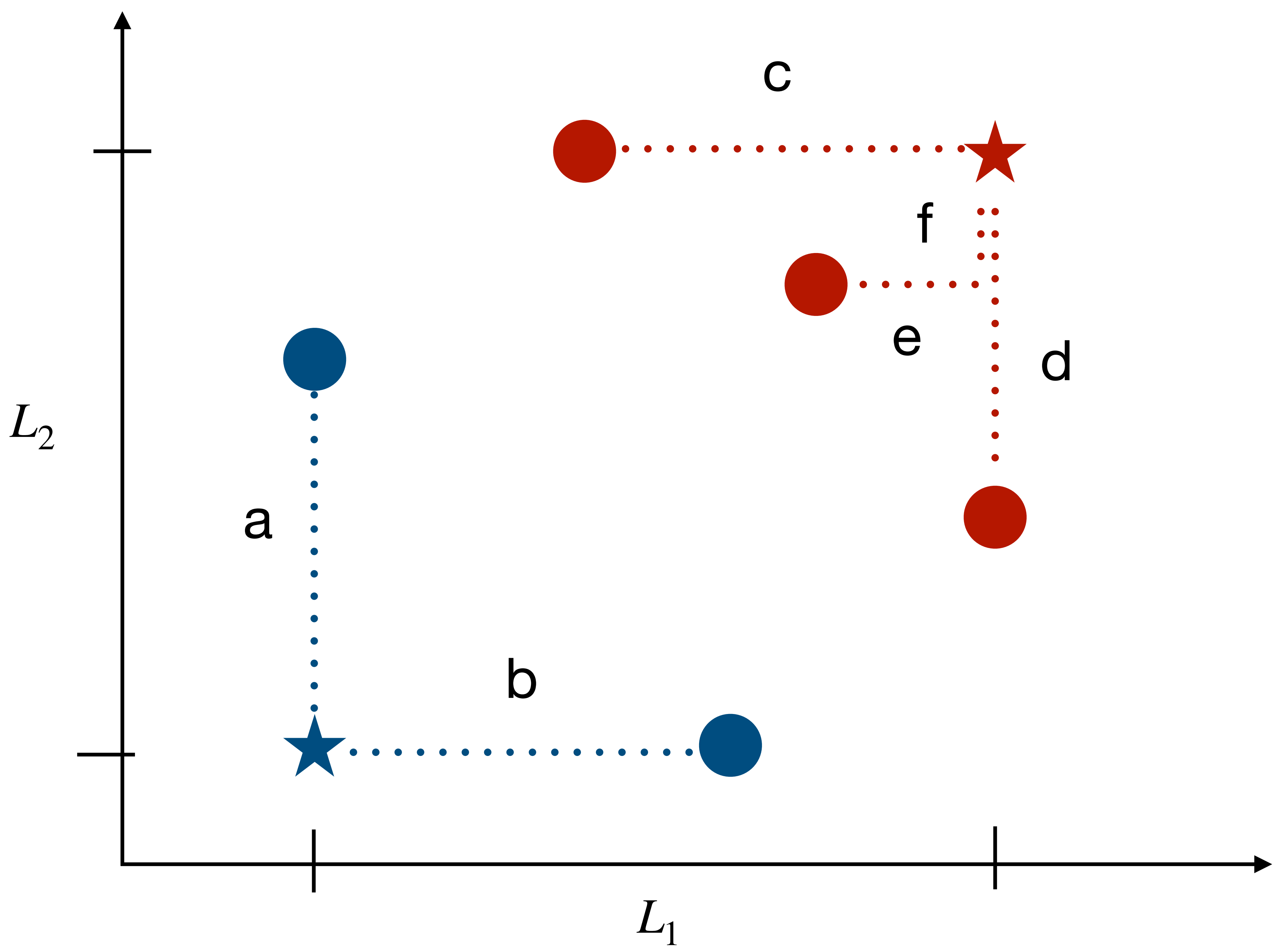
Two clusters in two dimensions



- Cost of optimal unconstrained clustering equals sum of dotted edges

$$OPT = a + b + c + d + e + f = \frac{a + b + c + d + e + f}{(L_1 + L_2)} \cdot (L_1 + L_2)$$

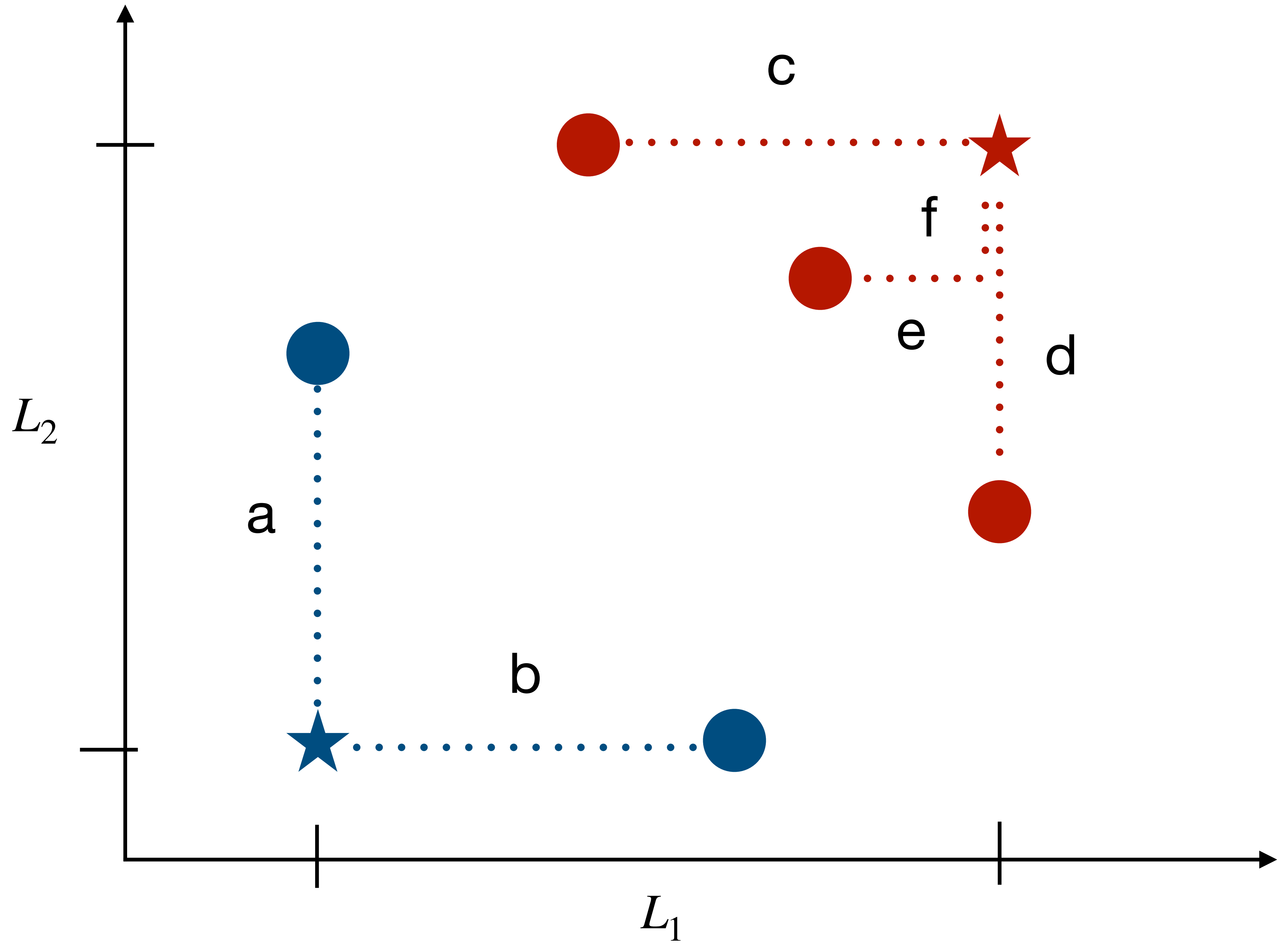
Two clusters in two dimensions



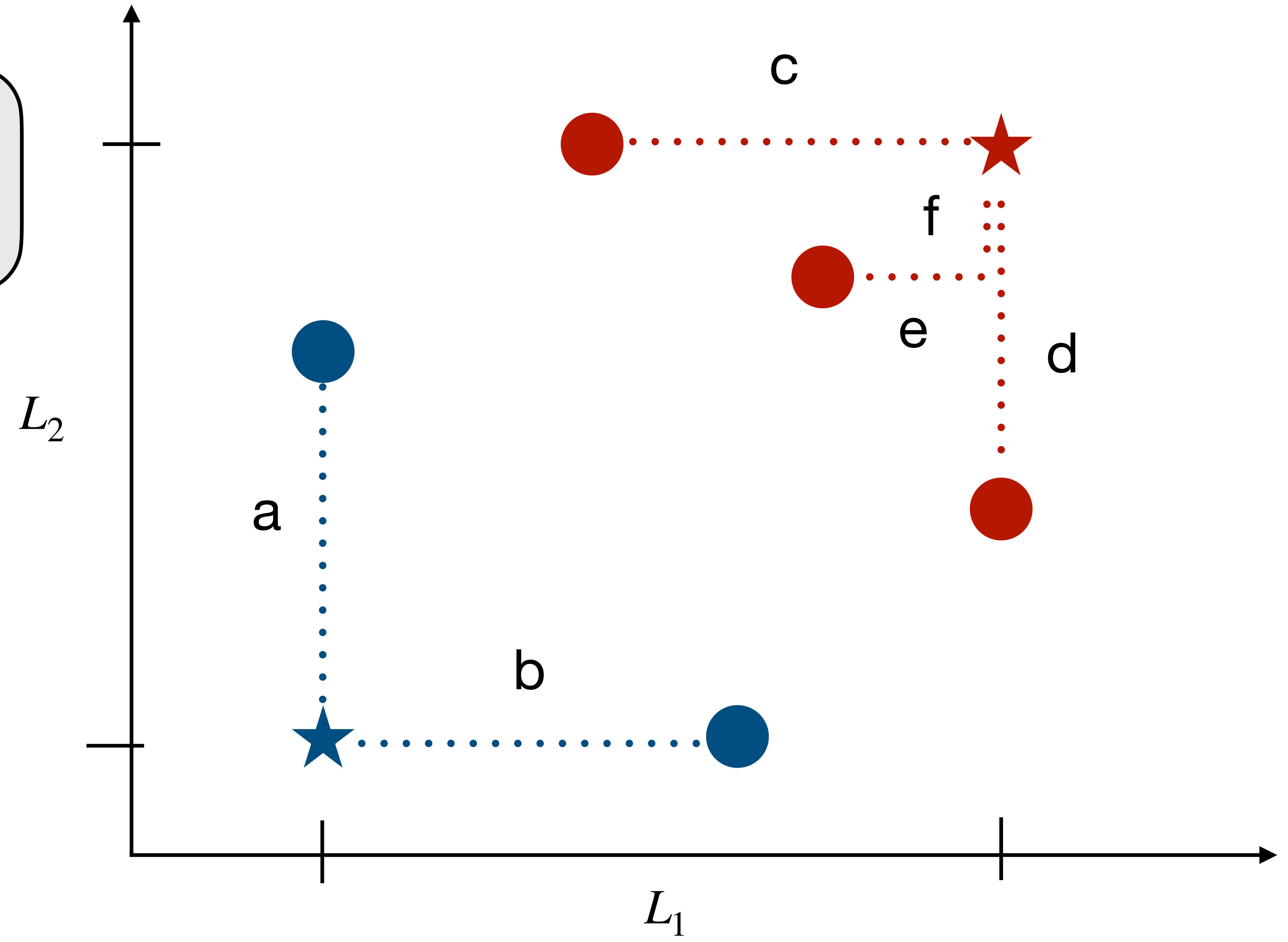
- Cost of optimal unconstrained clustering equals sum of dotted edges

$$OPT = a + b + c + d + e + f = \frac{a + b + c + d + e + f}{(L_1 + L_2)} \cdot (L_1 + L_2)$$

If we take a separating cut uniformly at random then this is at most the expected number of clients separated from their closest center

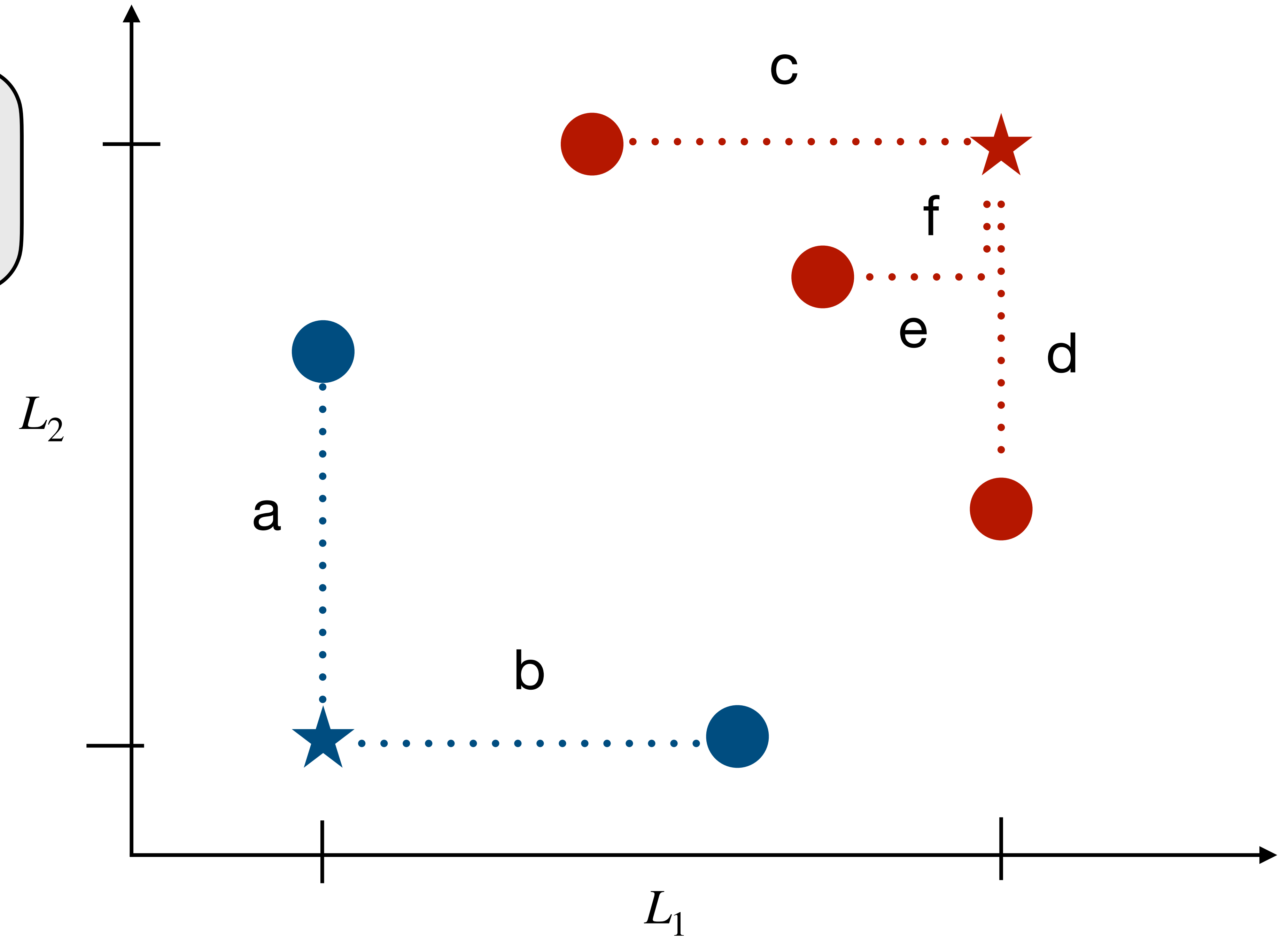

c
f
e
d
a
b
L_2
L_1

$\mathbb{E}[\text{number separated clients}] \leq OPT/(L_1 + L_2)$



$\mathbb{E}$[number separated clients] $\leq OPT/(L_1 + L_2)$

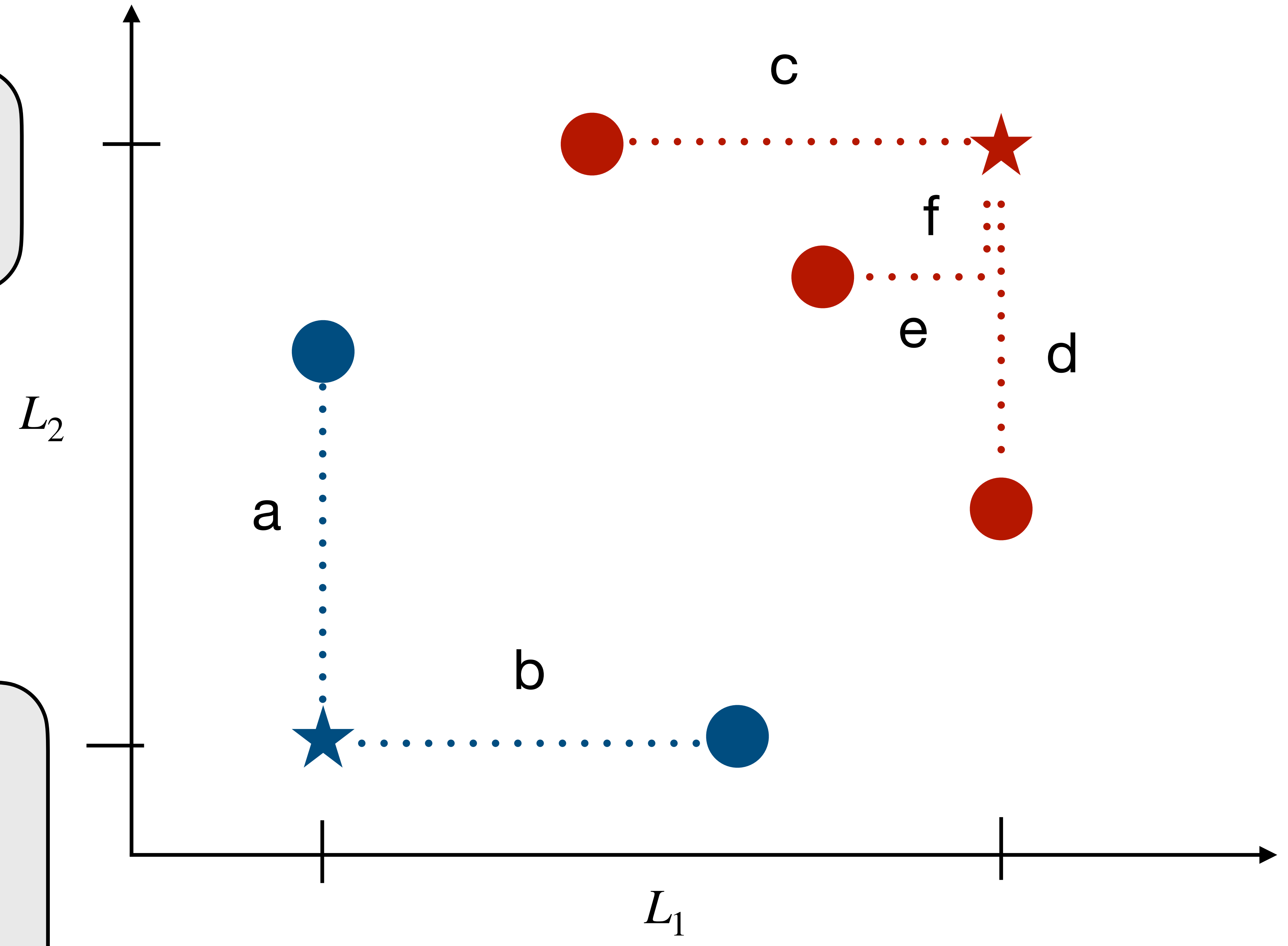
+



$\mathbb{E}$[number separated clients] $\leq OPT/(L_1 + L_2)$

+

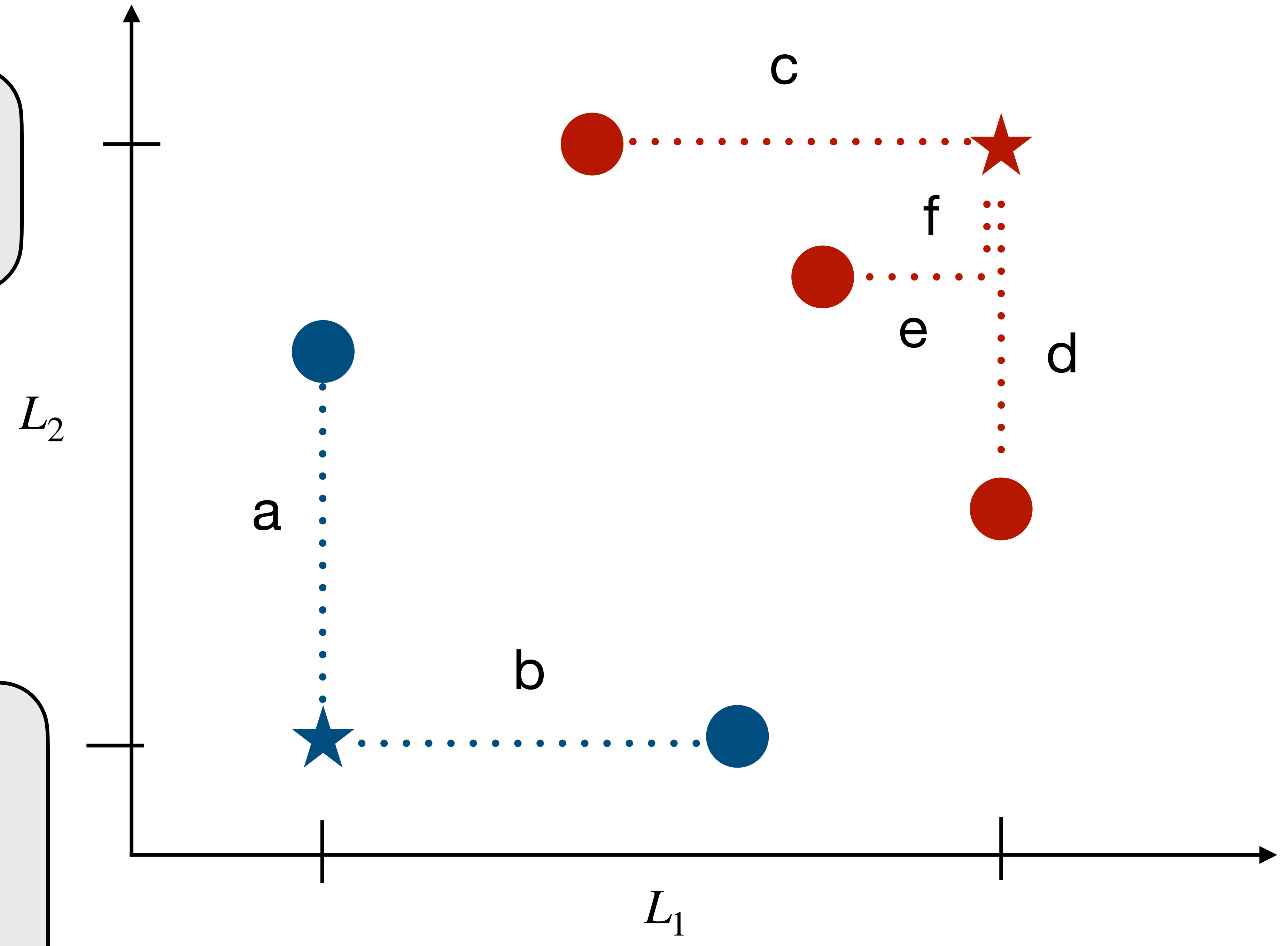
If a client is separated, it increases its cost by at most the maximum distance between centers which is at most $L_1 + L_2$



$\mathbb{E}[\text{number separated clients}] \leq OPT/(L_1 + L_2)$

+

If a client is separated, it increases its cost by at most the maximum distance between centers which is at most $L_1 + L_2$



- A uniformly random cut that separates the two centers increases the cost by at most

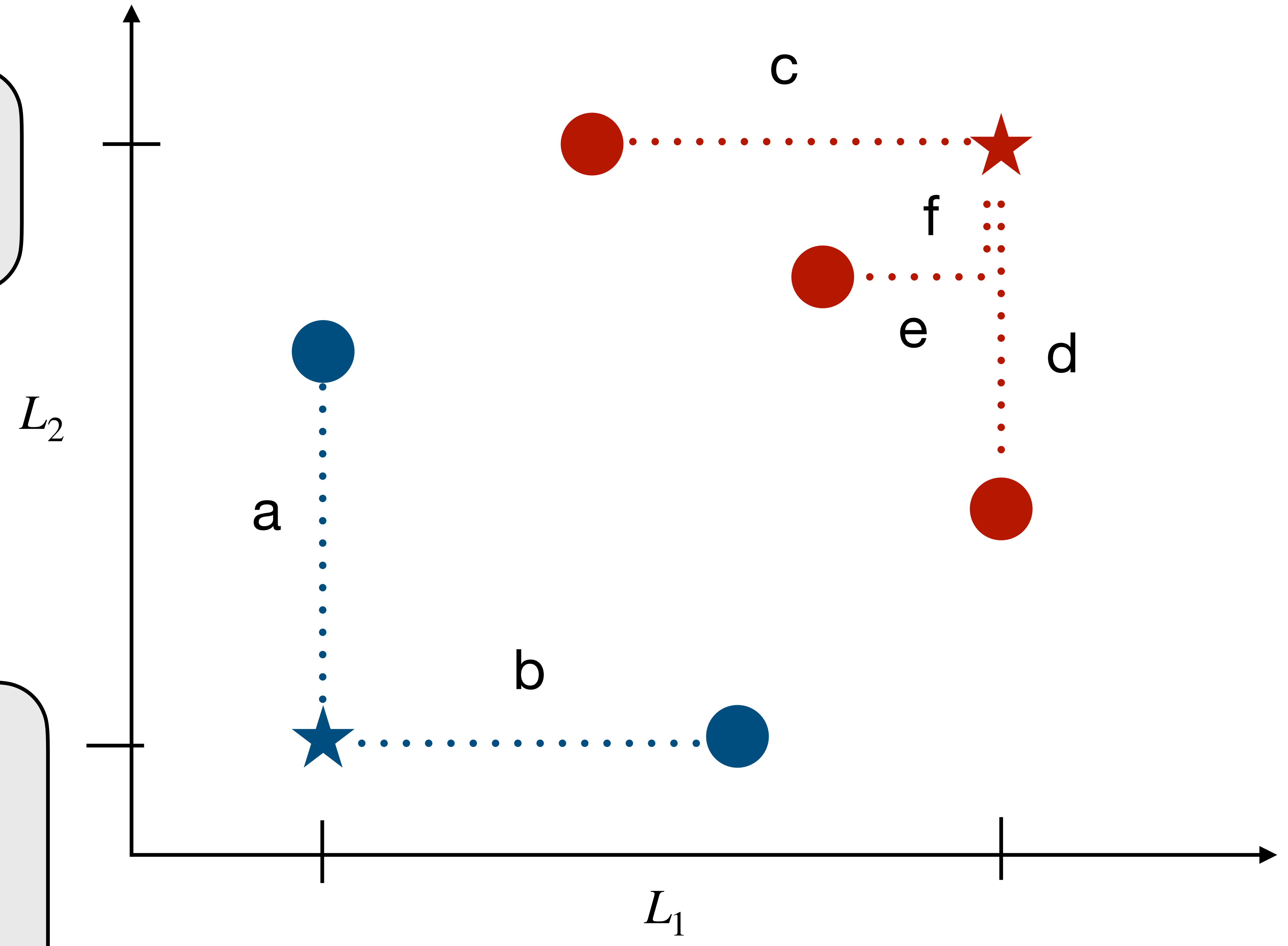
$$\mathbb{E}[\text{number separated clients}] \cdot (L_1 + L_2) \leq OPT$$

- It follows that there is an explainable clustering of cost at most $2 \cdot OPT$

$\mathbb{E}$[number separated clients] $\leq OPT/(L_1 + L_2)$

+

If a client is separated, it increases its cost by at most the maximum distance between centers which is at most $L_1 + L_2$



- A uniformly random cut that separates the two centers increases the cost by at most

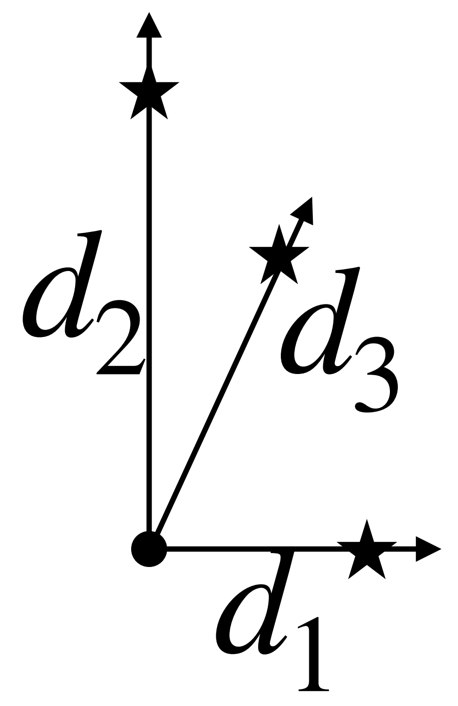
$$\mathbb{E}[\text{number separated clients}] \cdot (L_1 + L_2) \leq OPT$$

- It follows that there is an explainable clustering of cost at most $2 \cdot OPT$

This analysis works if you take the cut that separates the fewest points, which is the approach of Moshkovitz et al.

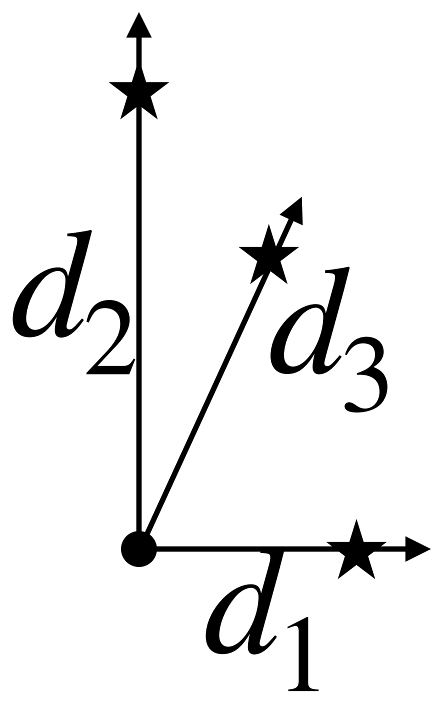
Special case

- A single point at origin
- Centers at distances $d_1 \leq d_2 \leq \ldots \leq d_k$ *each along unique dimension*
- Cost of unconstrained clustering thus equals d_1



Special case

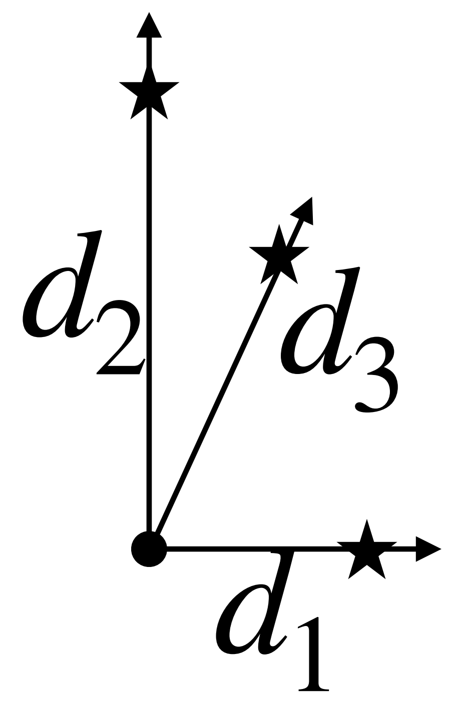
- A single point at origin
- Centers at distances $d_1 \leq d_2 \leq \ldots \leq d_k$ *each along unique dimension*
- Cost of unconstrained clustering thus equals d_1



Expected cost of explainable clustering determined by following process

Special case

- A single point at origin
- Centers at distances $d_1 \leq d_2 \leq \ldots \leq d_k$ *each along unique dimension*
- Cost of unconstrained clustering thus equals d_1

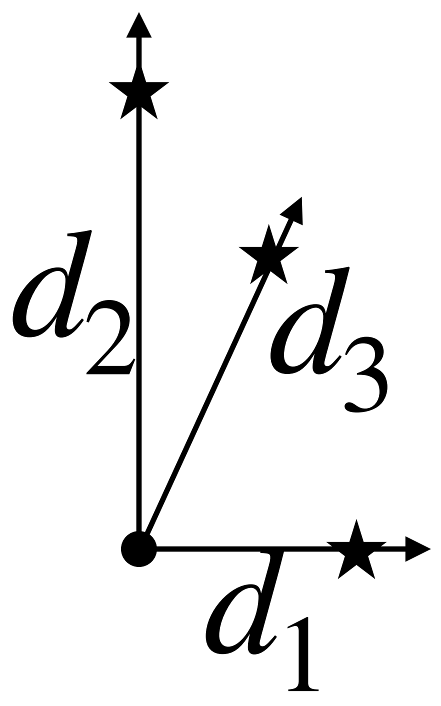


Expected cost of explainable clustering determined by following process

- While there are more than one center
  - Remove a center i with probability proportional to its distance d_i

Special case

- A single point at origin
- Centers at distances $d_1 \leq d_2 \leq \ldots \leq d_k$ *each along unique dimension*
- Cost of unconstrained clustering thus equals d_1



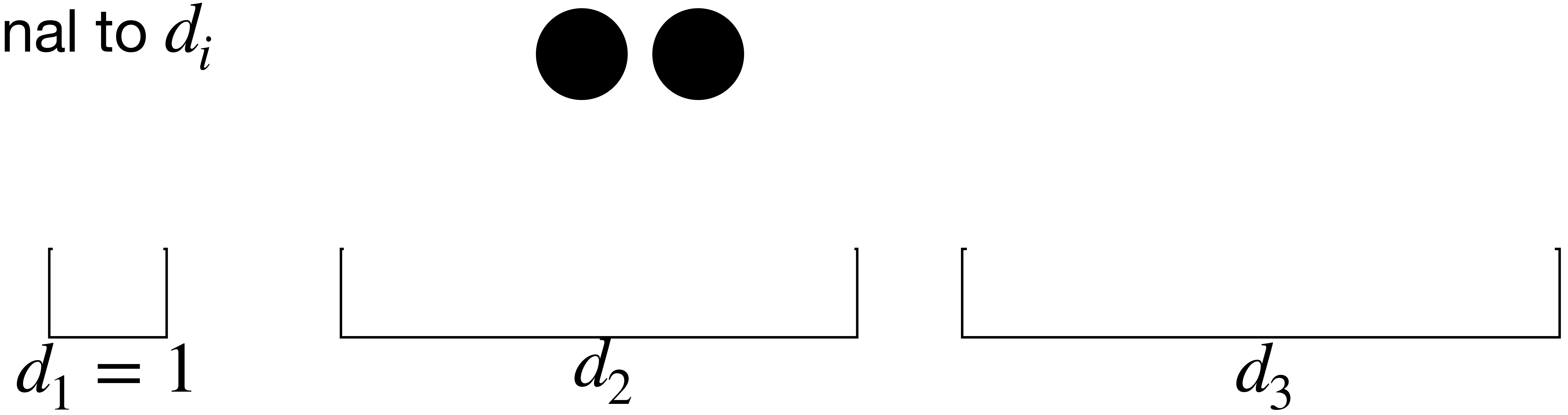
Expected cost of explainable clustering determined by following process

- While there are more than one center
  - Remove a center i with probability proportional to its distance d_i
- What is the expected distance to the last remaining center?

Balls-and-bin perspective

Equivalent to special case

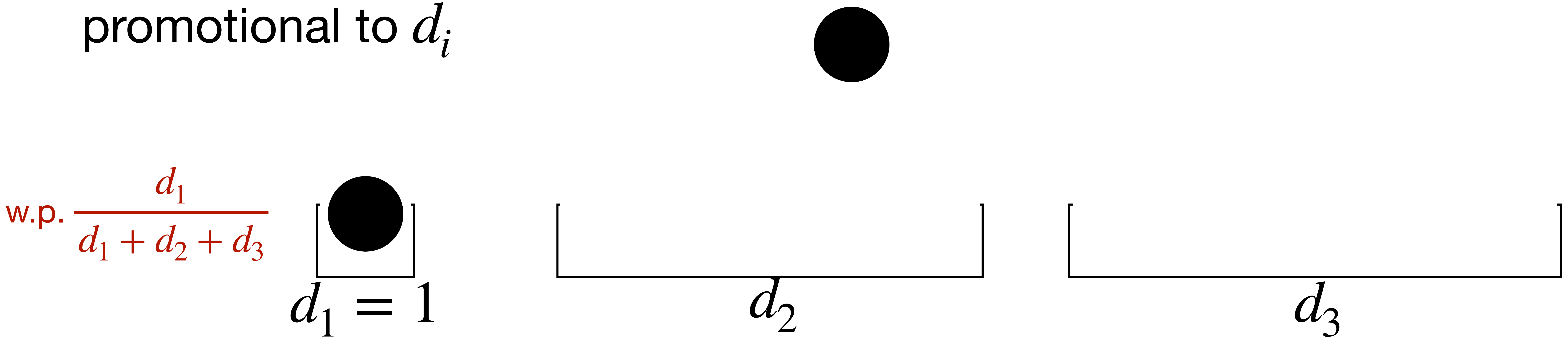
- k bins of different width $1 = d_1 \leq d_2 \leq \ldots \leq d_k$
- For k-1 steps, random ball hits one of the remaining bins with probability promotional to d_i



Balls-and-bin perspective

Equivalent to special case

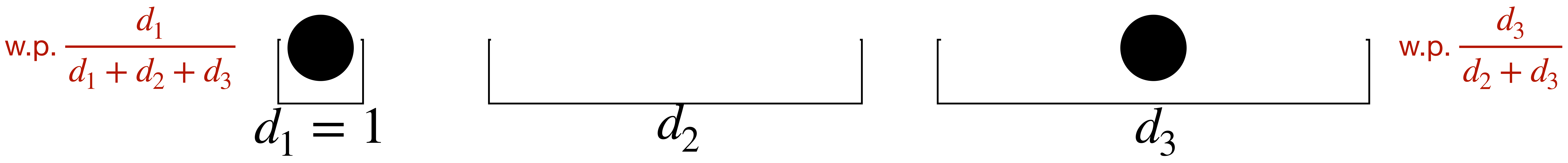
- k bins of different width $1 = d_1 \leq d_2 \leq \ldots \leq d_k$
- For k-1 steps, random ball hits one of the remaining bins with probability promotional to d_i



Balls-and-bin perspective

Equivalent to special case

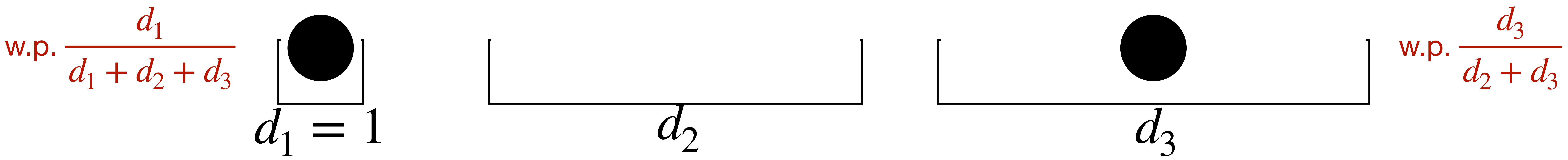
- k bins of different width $1 = d_1 \leq d_2 \leq \ldots \leq d_k$
- For k-1 steps, random ball hits one of the remaining bins with probability promotional to d_i



Balls-and-bin perspective

Equivalent to special case

- k bins of different width $1 = d_1 \leq d_2 \leq \ldots \leq d_k$
- For k-1 steps, random ball hits one of the remaining bins with probability promotional to d_i



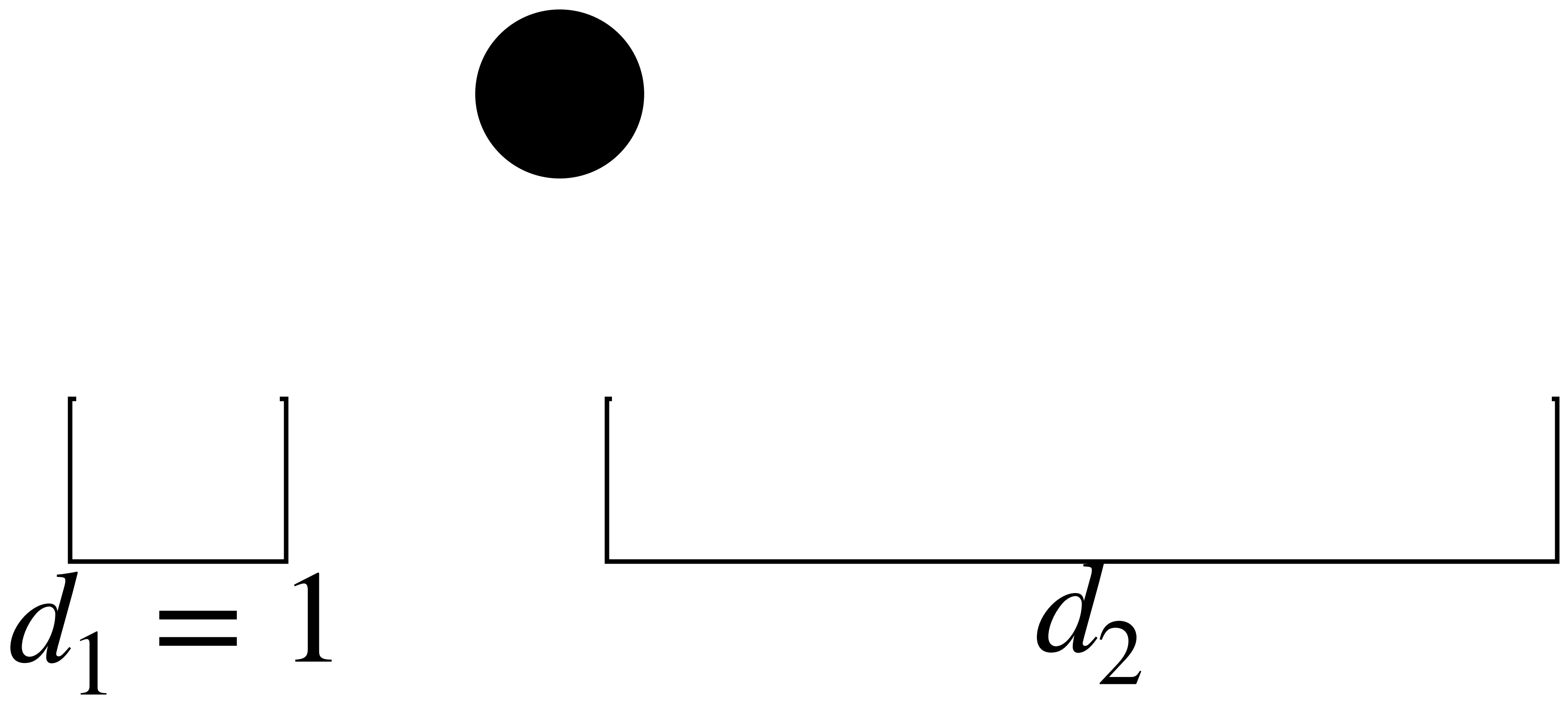
- **The expected width of remaining bin = price of explainability in special case**

One bin

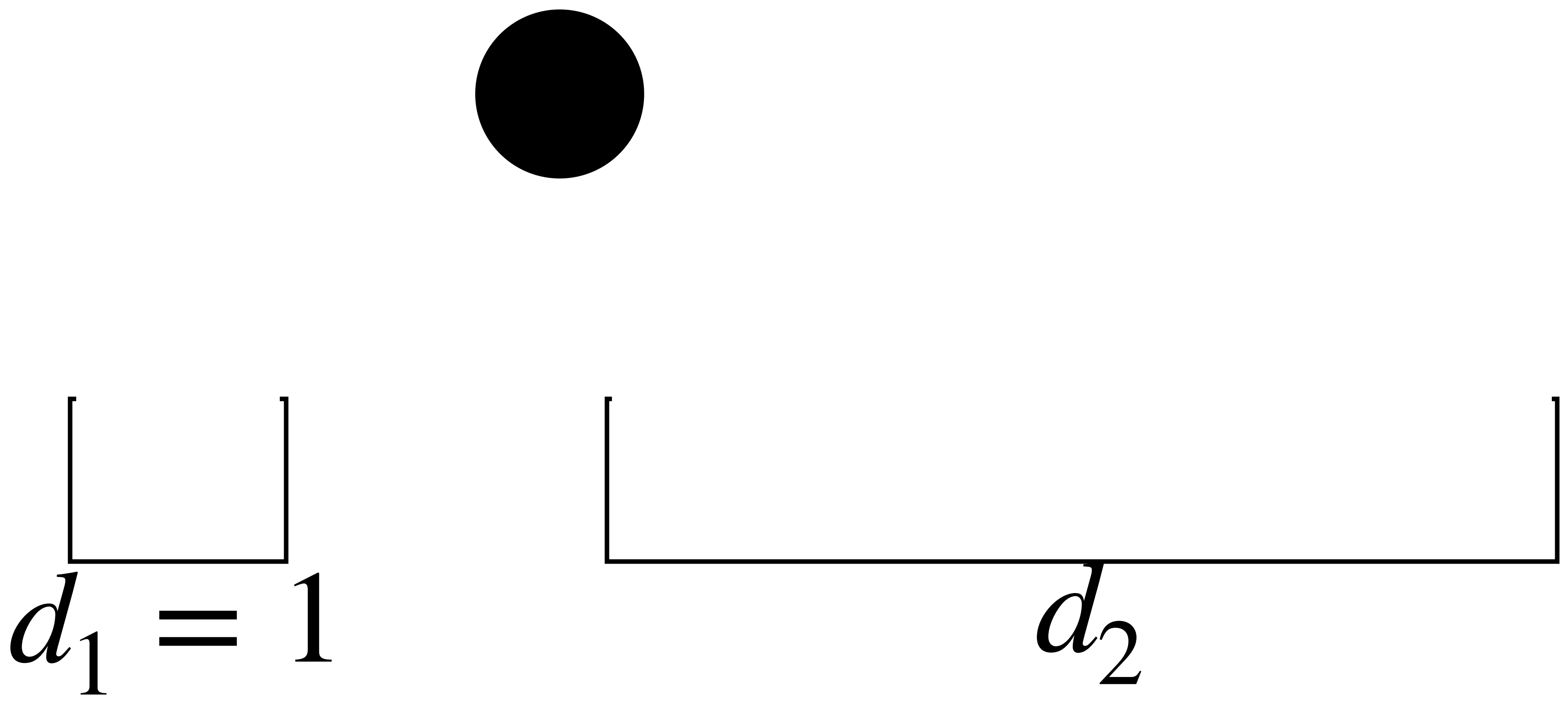
$$d_1 = 1$$

- Remaining bin is of width 1
- Price of explainability with one center is 1…

Two bins

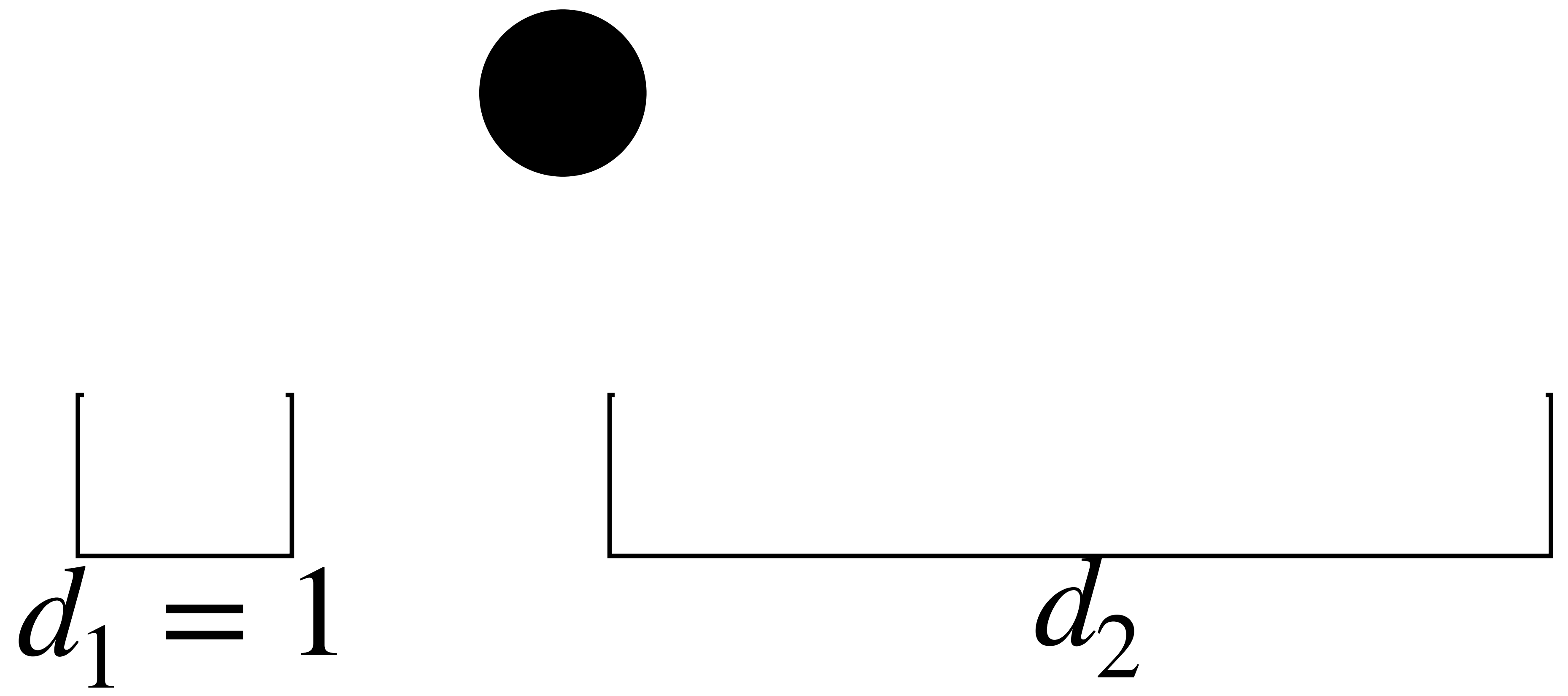


Two bins



$$\Pr[d_1 \textbf{ remains}] \cdot d_1 + \Pr[d_2 \textbf{ remains}] \cdot d_2$$

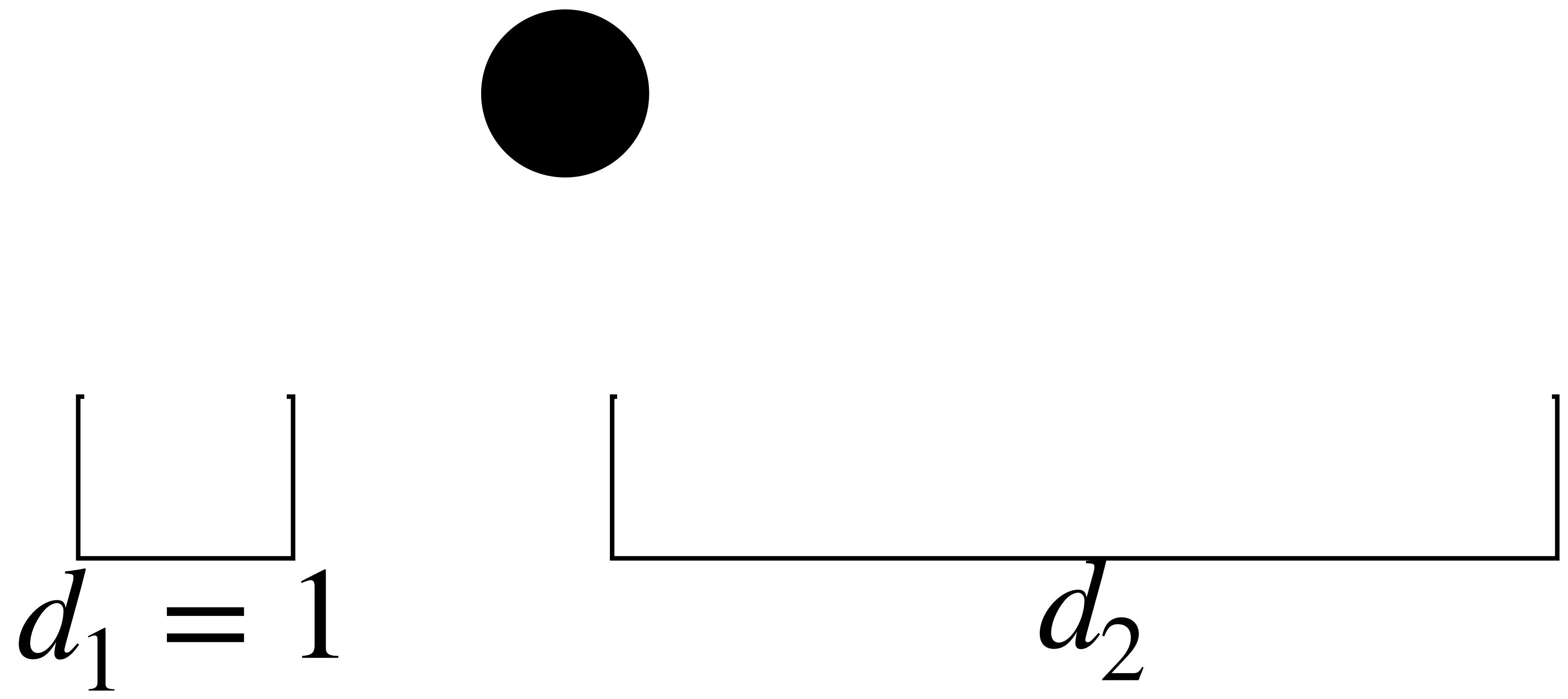
Two bins



$$\Pr[d_1 \textbf{ remains}] \cdot d_1 + \Pr[d_2 \textbf{ remains}] \cdot d_2$$

$$= \frac{d_2}{d_1 + d_2} \cdot d_1 + \frac{d_1}{d_1 + d_2} \cdot d_2$$

Two bins

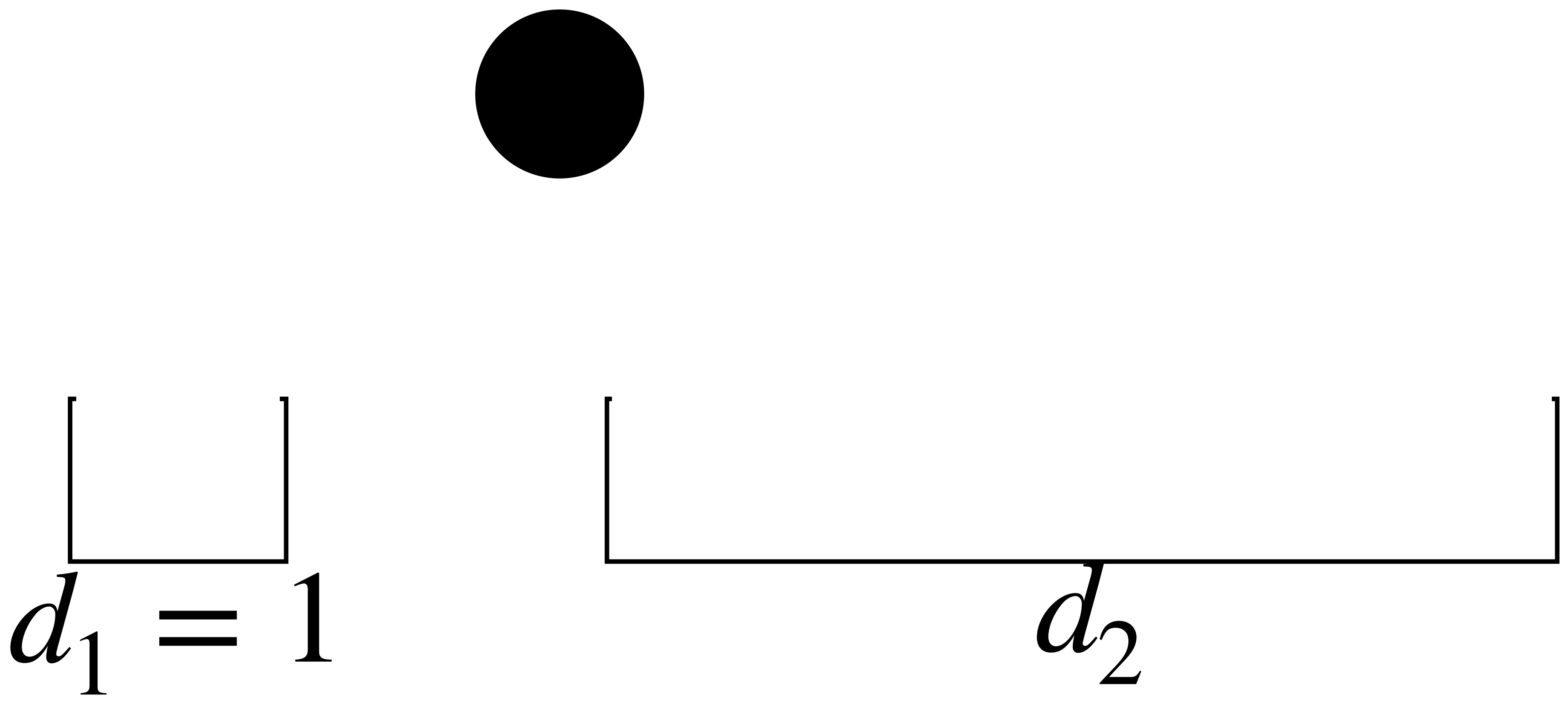


$$\Pr[d_1 \textbf{ remains}] \cdot d_1 + \Pr[d_2 \textbf{ remains}] \cdot d_2$$

$$= \frac{d_2}{d_1 + d_2} \cdot d_1 + \frac{d_1}{d_1 + d_2} \cdot d_2$$

$$\leq 2d_1 = 2$$

Two bins



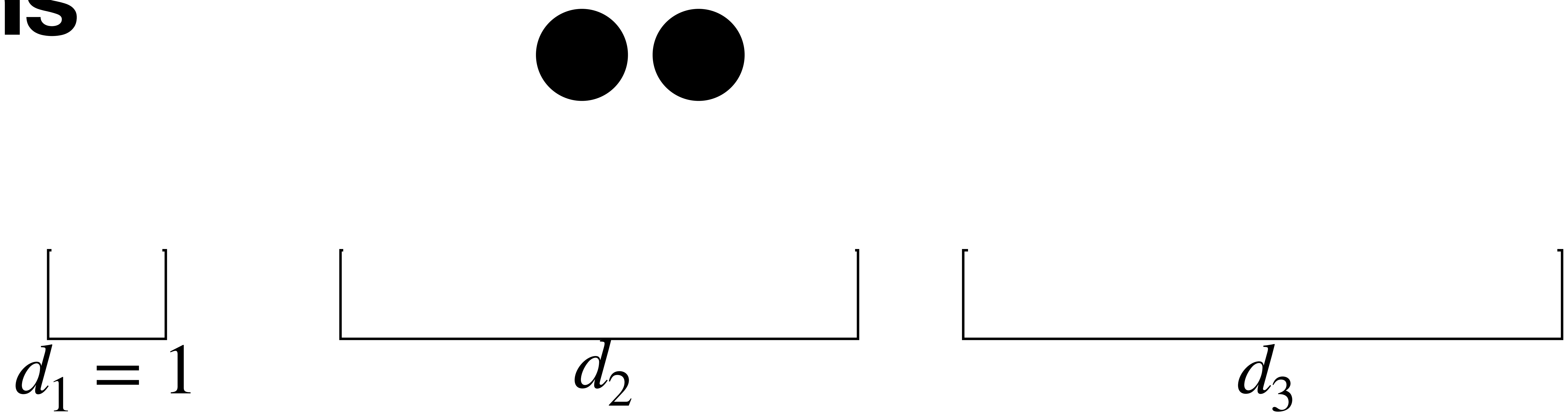
$$\Pr[d_1 \textbf{ remains}] \cdot d_1 + \Pr[d_2 \textbf{ remains}] \cdot d_2$$

$$= \frac{d_2}{d_1 + d_2} \cdot d_1 + \frac{d_1}{d_1 + d_2} \cdot d_2$$

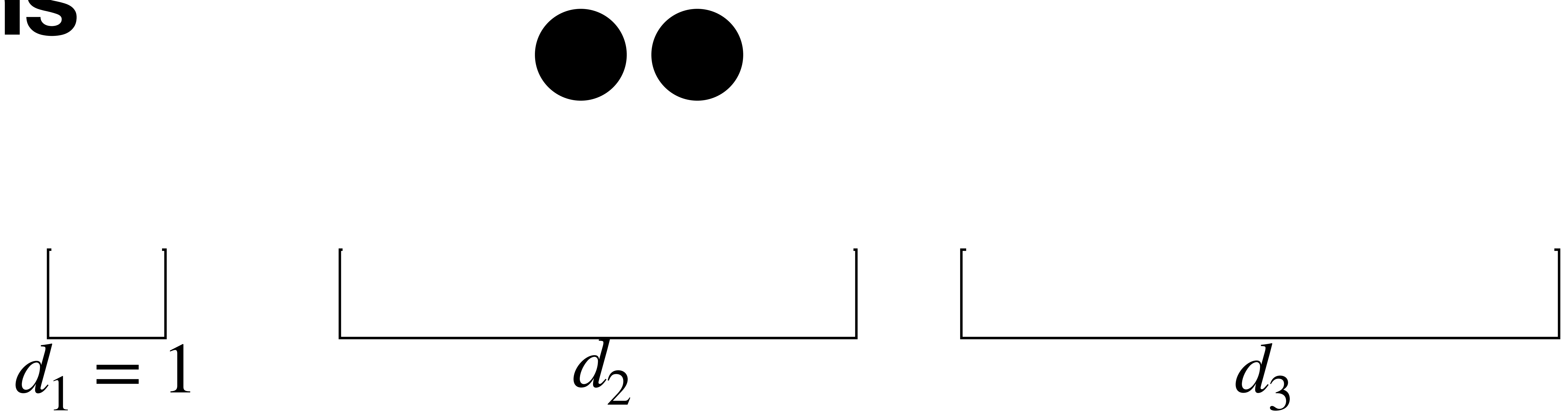
$$\leq 2d_1 = 2$$

- Price of explainability in special case with two centers is 2…

Three bins

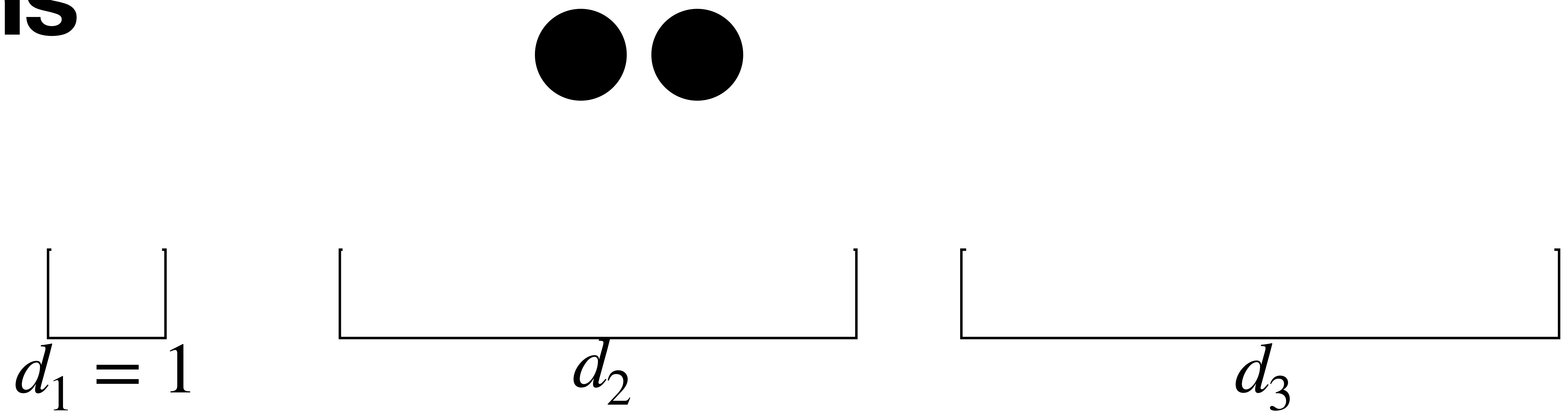


Three bins



$$\Pr[d_1 \textbf{ remains}] \cdot d_1 + \Pr[d_2 \textbf{ remains}] \cdot d_2 + \Pr[d_3 \textbf{ remains}] \cdot d_3$$

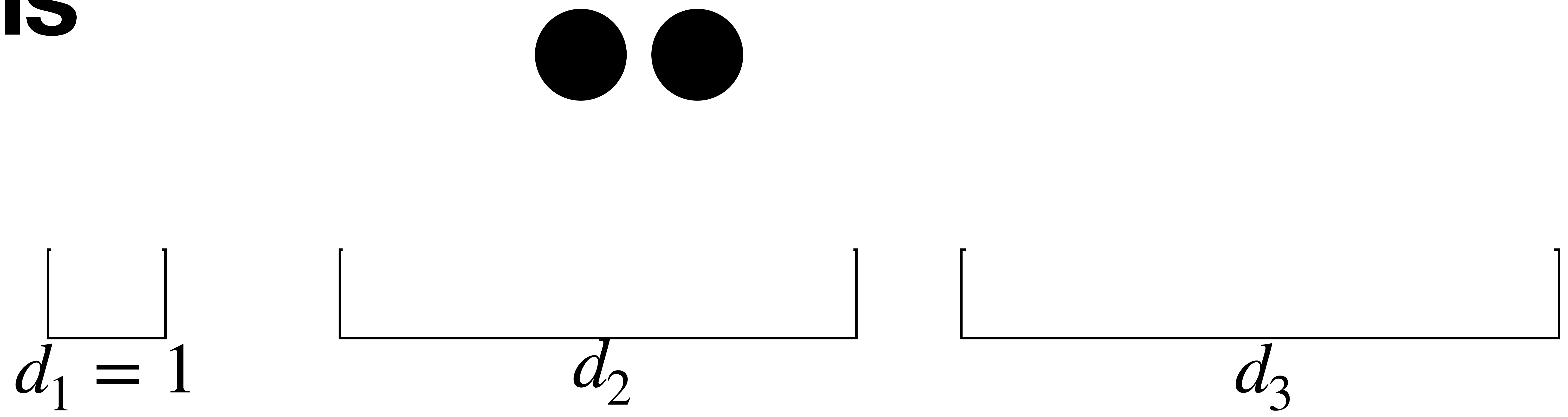
Three bins



$$\Pr[d_1 \textbf{ remains}] \cdot d_1 + \Pr[d_2 \textbf{ remains}] \cdot d_2 + \Pr[d_3 \textbf{ remains}] \cdot d_3$$

$$= \ldots$$

Three bins

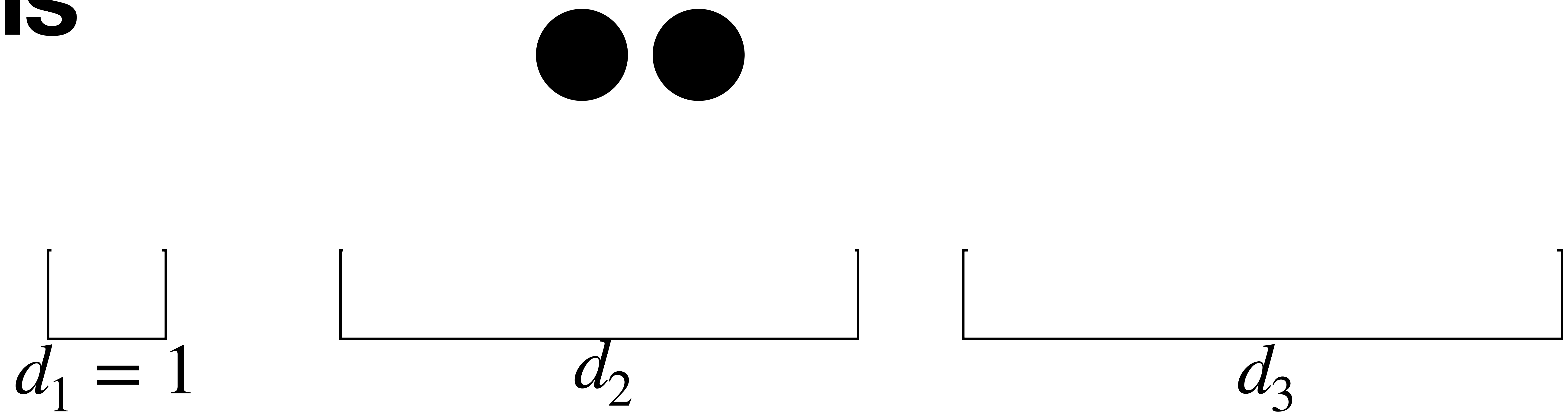


$$\Pr[d_1 \textbf{ remains}] \cdot d_1 + \Pr[d_2 \textbf{ remains}] \cdot d_2 + \Pr[d_3 \textbf{ remains}] \cdot d_3$$

$$= \ldots$$

$$\leq (1 + 1/1 + 1/2)d_1$$

Three bins



$$\Pr[d_1 \textbf{ remains}] \cdot d_1 + \Pr[d_2 \textbf{ remains}] \cdot d_2 + \Pr[d_3 \textbf{ remains}] \cdot d_3$$

$$= \ldots$$

$$\leq (1 + 1/1 + 1/2)d_1$$

- Price of explainability in special case with two centers is 1+1/1+1/2…

k-Bins

$1 = d_1 \leq d_2 \leq \ldots \leq d_k$

k-Bins

$1 = d_1 \leq d_2 \leq \ldots \leq d_k$

- The expected width of remaining bin is $1 + H_{k-1}$

k-Bins

$1 = d_1 \leq d_2 \leq \ldots \leq d_k$

- The expected width of remaining bin is $1 + H_{k-1}$
- In special case, price of explainability is $1 + H_{k-1}$

k-Bins

$1 = d_1 \leq d_2 \leq \ldots \leq d_k$

- The expected width of remaining bin is $1 + H_{k-1}$
- In special case, price of explainability is $1 + H_{k-1}$

- Here, $H_{k-1} = 1/1 + 1/2 + \ldots + 1/(k-1) \approx \ln k$

Proof by induction on k

We already verified base cases and here is the inductive step

Proof by induction on k

We already verified base cases and here is the inductive step

- Let $D = d_1 + d_2 + \ldots + d_k$
- Expected width of remaining bin is

Proof by induction on k

We already verified base cases and here is the inductive step

- Let $D = d_1 + d_2 + \ldots + d_k$
- Expected width of remaining bin is

$$\Pr[\text{1st ball lands in } d_1] \cdot \mathbb{E}[\text{width of remaining bin in instance } (d_2, \ldots d_k)]$$

Proof by induction on k

We already verified base cases and here is the inductive step

- Let $D = d_1 + d_2 + \ldots + d_k$
- Expected width of remaining bin is

$\Pr[\text{1st ball lands in } d_1] \cdot \mathbb{E}[\text{width of remaining bin in instance } (d_2, \ldots d_k)]$

$+$

Proof by induction on k

We already verified base cases and here is the inductive step

- Let $D = d_1 + d_2 + \ldots + d_k$
- Expected width of remaining bin is

$$\Pr[\text{1st ball lands in } d_1] \cdot \mathbb{E}[\text{width of remaining bin in instance } (d_2, \ldots d_k)]$$

$$+$$

$$\sum_{i=2}^{k} \Pr[\text{1st ball lands in } d_i] \cdot \mathbb{E}[\text{width of remaining bin in instance without } d_i]$$

Proof by induction on k

We already verified base cases and here is the inductive step

- Let $D = d_1 + d_2 + \ldots + d_k$
- Expected width of remaining bin is

$\Pr[\text{1st ball lands in } d_1] \cdot \mathbb{E}[\text{width of remaining bin in instance } (d_2, \ldots d_k)]$ $\leq \frac{d_1}{D} \cdot \frac{(d_2 + \ldots + d_k)}{k-1} \leq \frac{d_1}{k-1}$

$+$

$\sum_{i=2}^{k} \Pr[\text{1st ball lands in } d_i] \cdot \mathbb{E}[\text{width of remaining bin in instance without } d_i]$

Proof by induction on k

We already verified base cases and here is the inductive step

- Let $D = d_1 + d_2 + \ldots + d_k$
- Expected width of remaining bin is

$\Pr[\text{1st ball lands in } d_1] \cdot \mathbb{E}[\text{width of remaining bin in instance } (d_2, \ldots d_k)]$ $\leq \frac{d_1}{D} \cdot \frac{(d_2 + \ldots + d_k)}{k-1} \leq \frac{d_1}{k-1}$

$+$

$\sum_{i=2}^{k} \Pr[\text{1st ball lands in } d_i] \cdot \mathbb{E}[\text{width of remaining bin in instance without } d_i]$ $\leq (1 + H_{k-2})d_1$

Proof by induction on k

We already verified base cases and here is the inductive step

- Let $D = d_1 + d_2 + \ldots + d_k$
- Expected width of remaining bin is

$\Pr[$1st ball lands in $d_1] \cdot \mathbb{E}[$width of remaining bin in instance $(d_2, \ldots d_k)]$ $\quad \leq \frac{d_1}{D} \cdot \frac{(d_2 + \ldots + d_k)}{k-1} \leq \frac{d_1}{k-1}$

$+$

$\sum_{i=2}^{k} \Pr[$1st ball lands in $d_i] \cdot \mathbb{E}[$width of remaining bin in instance without $d_i]$ $\quad \leq (1 + H_{k-2})d_1$

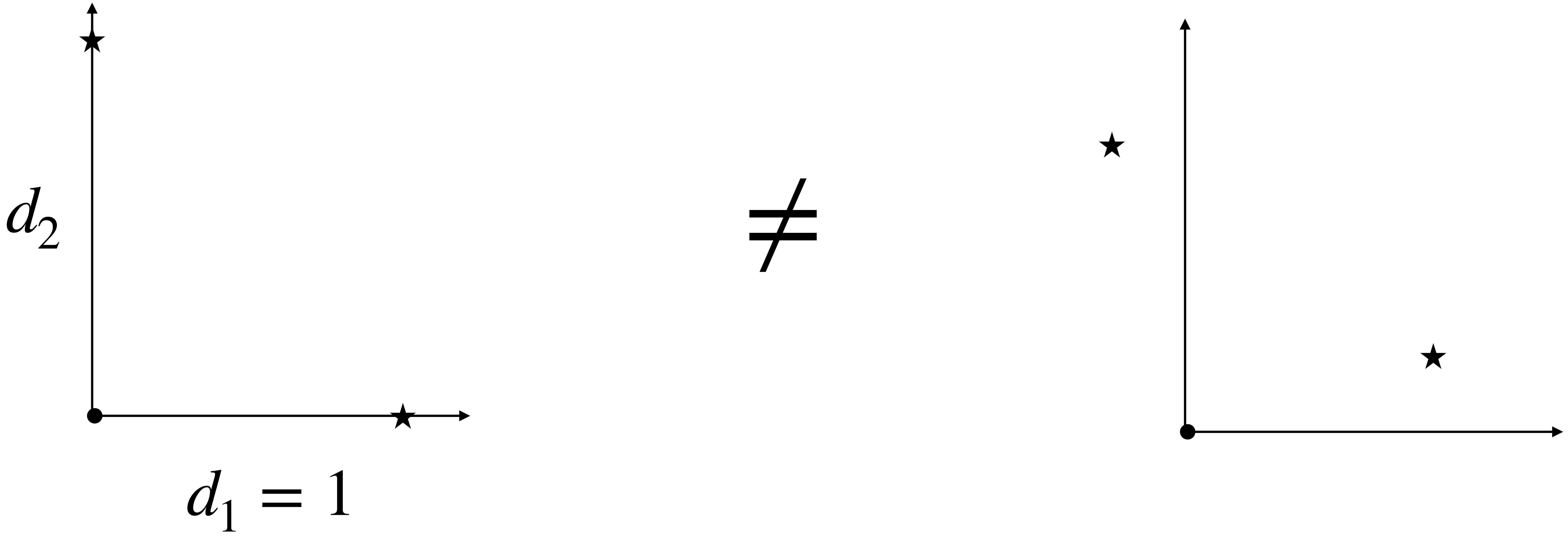
$$\leq (1 + H_{k-1})d_1 = 1 + H_{k-1}$$

Conjecture: The expected cost of the explainable clustering by TCS-Algorithm is at most $(1 + H_{k-1})$ times the cost of the input unconstrained clustering

$1 + H_{k-1}$ is tight even in the axis aligned case

Conjecture: The expected cost of the explainable clustering by TCS-Algorithm is at most $(1 + H_{k-1})$ times the cost of the input unconstrained clustering

$1 + H_{k-1}$ is tight even in the axis aligned case



State-of-the-art and open questions

The independent works in 2021

- Makarychev and Shan:
  - $O(\log k \log \log k)$
- Gamlath, Jia, Polak, Svensson:
  - $O(\log^2 k)$
- Esfandiari, Mirrokni, Narayanan:
  - $O(\min(\log k \log \log k, d \log^2 d))$

- Gupta, Pitty, Svensson, Yuan'23:

Theorem: The price of explainability given by *TCS-Algorithm* is $1 + H_{k-1}$

- Gupta, Pitty, Svensson, Yuan'23:

Theorem: The price of explainability given by *TCS-Algorithm* is $1 + H_{k-1}$

Theorem: The price of explainability is at least $(1 - \epsilon)\ln(k)$ for any $\epsilon > 0$

Open problems

Open problems

What if more than one dimension in threshold cuts?

Open problems

What if more than one dimension in threshold cuts?

Conjecture: price of explainability for k-means is $\Theta(k)$

Upper bound of O($k \log k$) [Esfandiari, Mirrokni, Narayanan'21], see also [Charikar and Hu'21]

Lower bound of $\Omega(k)$

Open problems

What if more than one dimension in threshold cuts?

Conjecture: price of explainability for k-means is $\Theta(k)$

Upper bound of $O(k \log k)$ [Esfandiari, Mirrokni, Narayanan'21], see also [Charikar and Hu'21]
Lower bound of $\Omega(k)$

What is the price of explaining clustering using k-dimensions?

Related to "feature selection" [Boutsidis, Mahoney, Drineas'09]

Open problems

What if more than one dimension in threshold cuts?

Conjecture: price of explainability for k-means is $\Theta(k)$

Upper bound of O($k \log k$) [Esfandiari, Mirrokni, Narayanan'21], see also [Charikar and Hu'21]
Lower bound of $\Omega(k)$

What is the price of explaining clustering using k-dimensions?

Related to "feature selection" [Boutsidis, Mahoney, Drineas'09]

What's the approximability of explainable clustering?

Resolved for k-median, it is ln(k) [GPSY'23], interesting for k-means

Thank you for your attention!

The details

- Gupta, Pitty, Svensson, Yuan'23:

Theorem: The price of explainability given by *TCS-Algorithm* is $1 + H_{k-1}$

Theorem: The price of explainability is at least $(1 - \epsilon)\ln(k)$ for any $\epsilon > 0$

Lower bounds via reduction from the Hitting Set Problem

s-uniform Hitting set problem

- INPUT: A set system $([d], T = \{S_1, S_2, \ldots, S_k\})$ where $|S_i| = s$
- OUTPUT: A subset $H \subseteq [d]$ of minimum cardinality that hits every S_i, i.e., $S_i \cap H \neq \emptyset$

- Integrality gap: Exist instances so that any hitting set has size $\frac{d}{s} \ln k$
- Feige: it is hard to approximate better than $(1 - \varepsilon)\ln k$ for any $\epsilon > 0$. Between friends hard to distinguish between size $\frac{d}{s}$ and $\frac{d}{s} \ln k$

Reduction

Construct the following explainable instance

- The reference clustering $\mathcal{U} = \{\mu_0, \mu_1, \ldots, \mu_k\}$ where
  - μ_0 is at the origin and μ_i is the characteristic vector of the set S_i
- Infinitely many points at each center in $\mathcal{U}$ => Any reasonable clustering must contain one leaf per center
- Plus one point at the location e_i for $i \in [d]$

- Observation 1: the cost of reference clustering is d
- Observation 2: we must separate μ_0 from all other centers and thus these selected threshold cuts form a hitting set. Each such cut increases the cost of a point from 1 to s
- Hence cost of optimal explainable clustering is $\approx h \cdot s + (d - s)$ where h is the size of optimal hitting set

Plugging in known results

- Integrality gap: $h \geq \frac{d}{s} \ln k$ leads to $h \cdot s + (d - s) \geq d \ln k$
- Hardness of approx: Hard to distinguish between $\leq 2d$ and $\geq d \ln k$
- Same results hold for k-means: stronger results known for price of explainability but not for approximability

Analysis via exponential clocks

The setting

- By linearity of expectation, enough to analyze single point which by translation is at the origin.
- At any point we take a cut S with probability proportional to z_S
- The distance to center i is thus $d_i = \sum_{S: i \in S} z_S$
- We assume by scaling that $d_1 = 1$ and for simplicity that $z_{\{1\}} = 1$

Exponential clocks

- Nice properties
  - Suppose that $X_i \sim \exp(\lambda_i)$ then X_j takes min value with probability $\frac{\lambda_j}{\lambda_1 + \ldots +}$ moreover the minimum is distributed as $\exp(\sum \lambda_i)$
  - Memorylessness: Suppose $X \sim \exp(\lambda)$ then $\Pr[X \geq s + t \mid X \geq t] = \Pr[X \geq s]$
  - The pdf $f_X(x) = \lambda e^{-\lambda x}$

Exponential clocks

- We can equivalently think of the process of selecting random cuts as using exponentially random variables
- First sample $x_i \sim \exp(d_i)$ for every S
- Then inspect the cuts in increasing order of their values.
- This is the same process as probability that i is next cut is proportional w.r.t d_i and remaining cuts

When do we pay d_i

- i is last among faraway centers and $X_1 \leq X_i$ where X_i
- Let E_i be the event that i is last among faraway centers
- Then the payment of d_i is at most d_i times
- $\Pr[X_1 \leq X_i \wedge E_i]$ which by the law of total probability equals
- $\int_0^\infty \Pr[X_1 \leq t \wedge E_i \mid X_i = t] f_{X_i}(t) = \int_0^\infty \Pr[X_1 \leq t] \cdot \Pr[E_i \mid X_i = t] f_{X_i}(t)$

- $\int_0^\infty \Pr[X_1 \le t] \cdot \Pr[E_i \mid X_i = t] f_{X_i}(t)$
- Let $p_i = \Pr[E_i]$ then the above expression is maximized when $E_i = 1$ for large values of t
- That is, $p_i = \int_0^\infty \Pr[E_i \mid X_i = t] f_{X_i}(t) = \int_{a_i}^\infty f_{X_i}(t)$
- Therefore the above expression is upper bounded by
- $\int_{a_i}^\infty \Pr[X_1 \le t] f_{X_i}(t)$
- Plugging in the cdf and pdf and doing the calculations give us that the total contribution to the cost of center i is at most $p_i + p_i \ln(1/p_i)$
- Summing up over all far away centers we get that their total contribution to the cost is at most
- $\sum_{i=2}^{k} p_i + p_i \ln(1/p_i) \le 1 + \ln(k-1)$
- Plus the cost of the closest center gives an upper bound of $2 + \ln(k)$