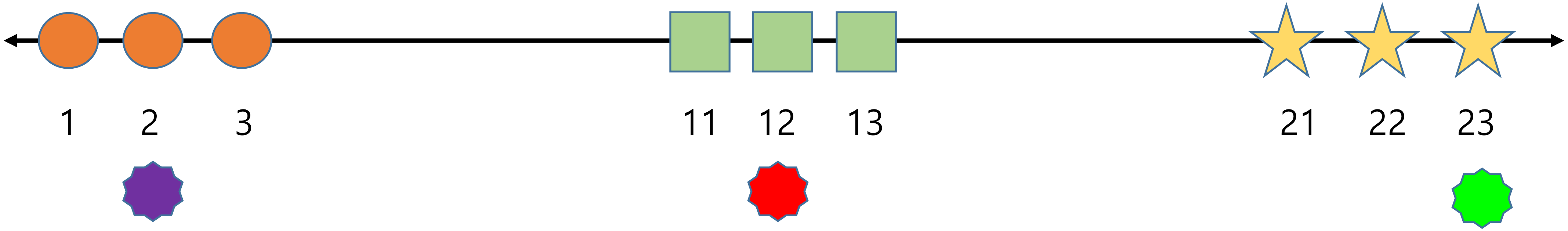


2.2 The k-means++ algorithm

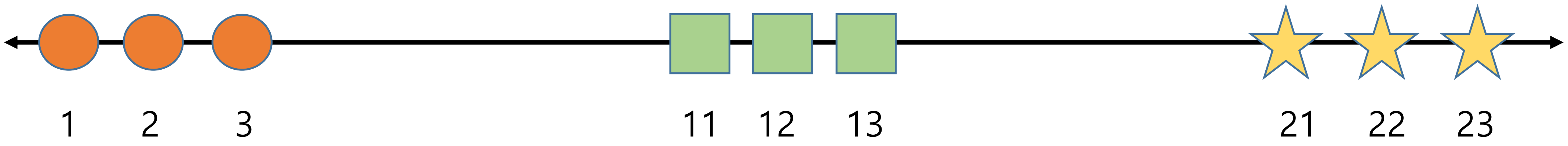
We propose a specific way of choosing centers for the **k-means** algorithm. In particular, let $D(x)$ denote the shortest distance from a data point to the closest center we have already chosen. Then, we define the following algorithm, which we call **k-means++**.

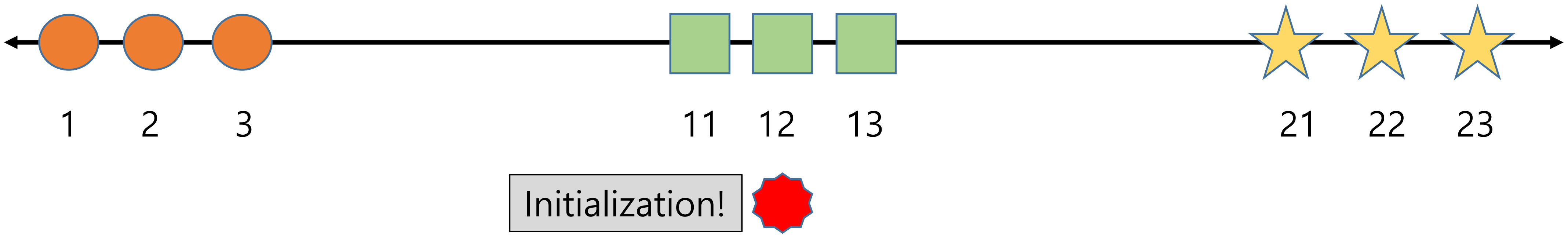
1a. Take one center c_1, chosen uniformly at random from $\mathcal{X}$.

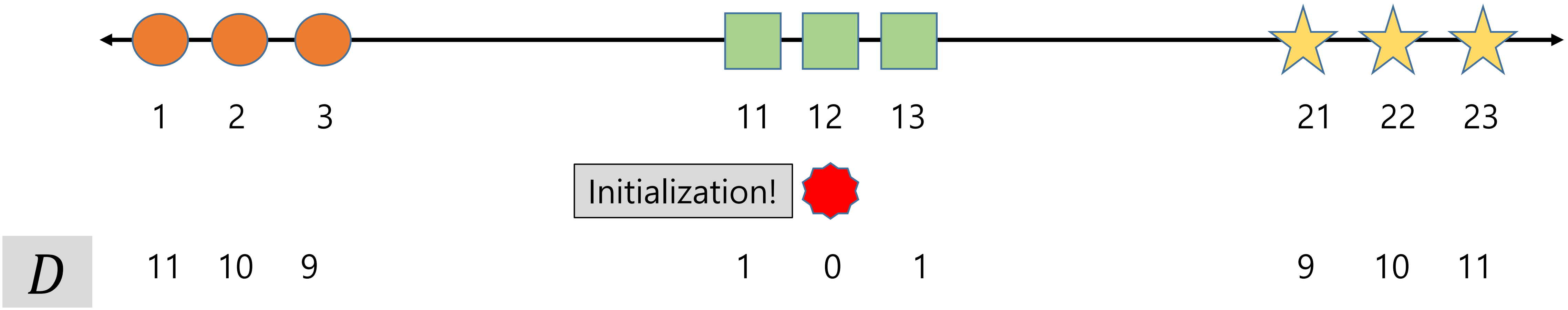
1b. Take a new center c_i, choosing $x \in \mathcal{X}$ with probability $\frac{D(x)^2}{\sum_{x \in \mathcal{X}} D(x)^2}$.

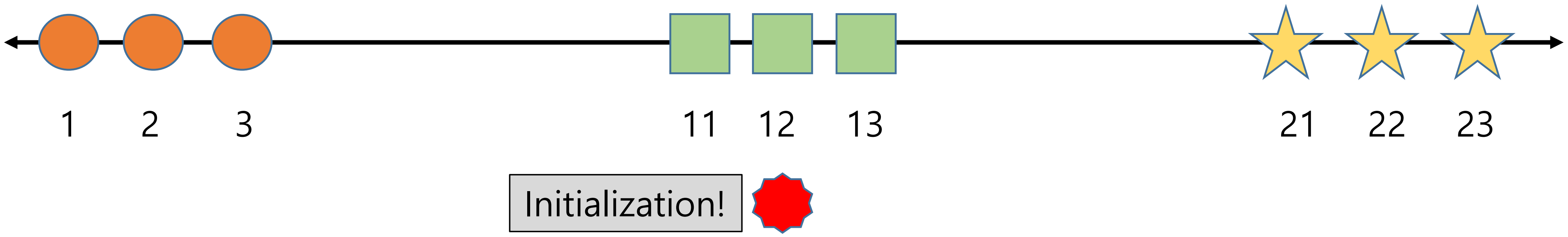
1c. Repeat Step 1b. until we have taken k centers altogether.

2-4. Proceed as with the standard **k-means** algorithm.

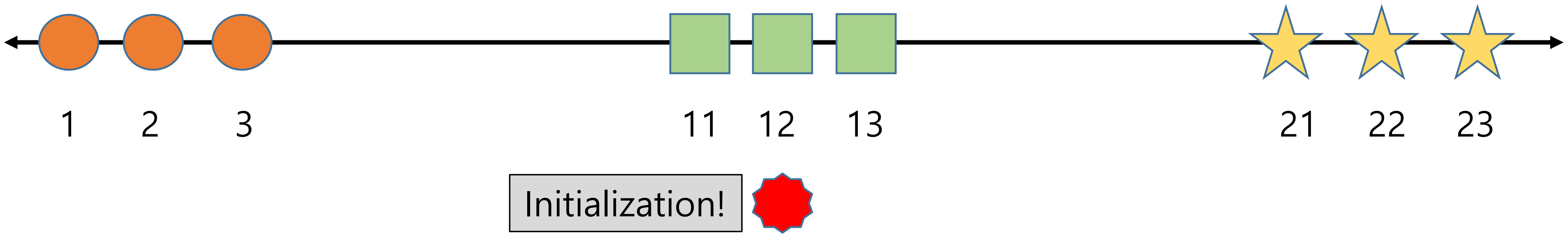

1
2
3
11
12
13
21
22
23


1
2
3
11
12
13
21
22
23
Initialization!


1 2 3 11 12 13 21 22 23
Initialization!
D
11 10 9 1 0 1 9 10 11



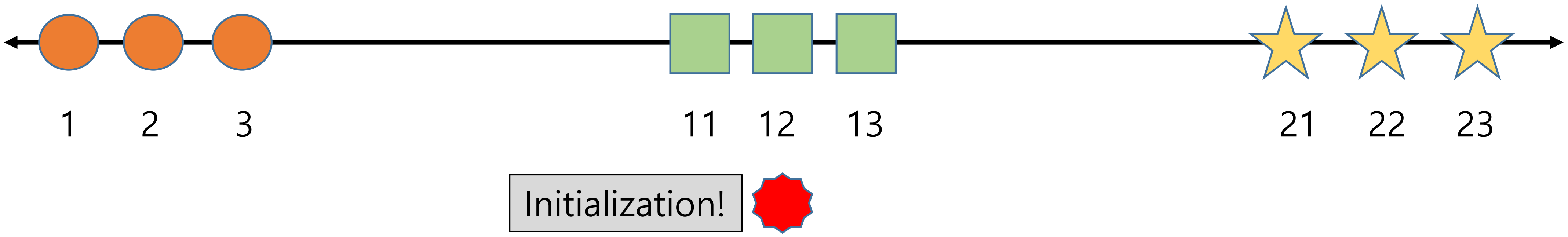
D	11	10	9	1	0	1	9	10	11
D^2	121	100	81	1	0	1	81	100	121



D	11	10	9	1	0	1	9	10	11
D^2	121	100	81	1	0	1	81	100	121

$$sum(D^2) = 606$$

$$sum\left(\frac{D^2}{sum(D^2)}\right) = 1$$



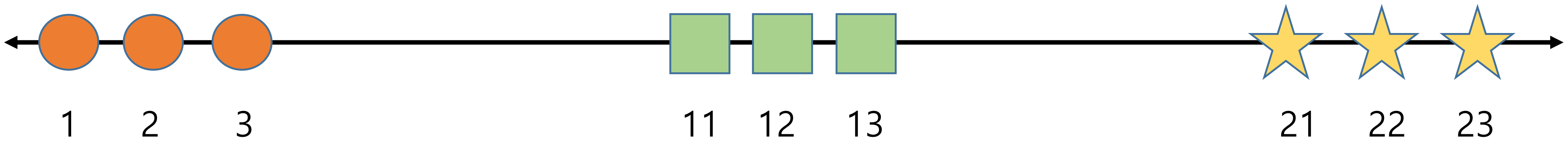
D	11	10	9	1	0	1	9	10	11
D^2	121	100	81	1	0	1	81	100	121

$$sum(D^2) = 606$$

$$sum\left(\frac{D^2}{sum(D^2)}\right) = 1$$

$$D^2/sum(D^2)$$

121/606	100/606	81/606	1/606	0/606	1/606	81/606	100/606	121/606



Initialization!

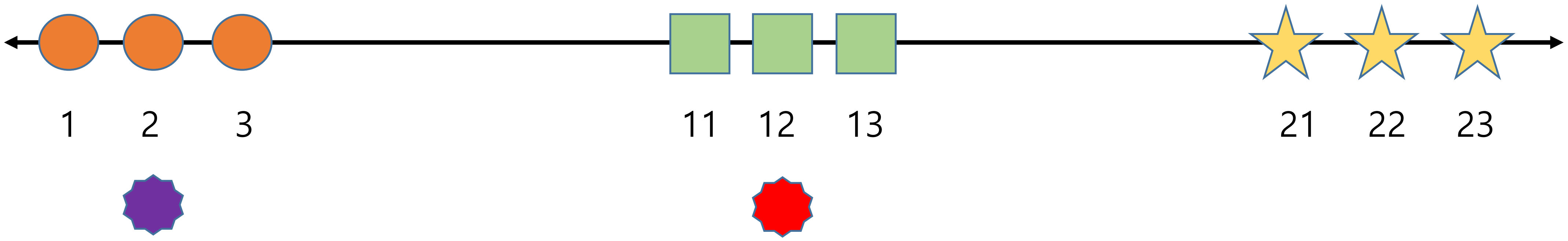
D	11	10	9	1	0	1	9	10	11
D^2	121	100	81	1	0	1	81	100	121

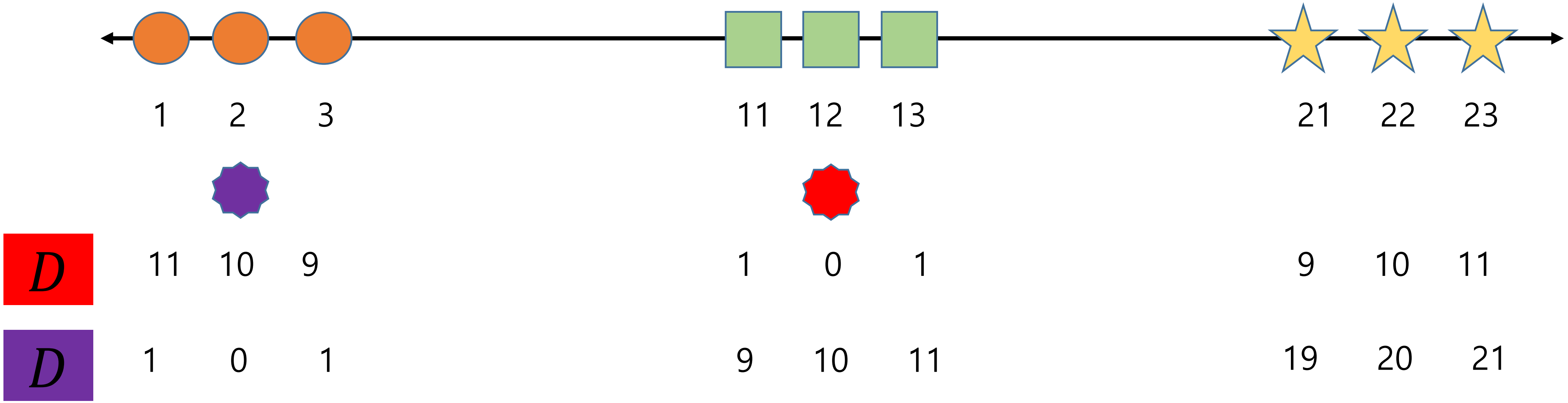
$$sum(D^2) = 606$$

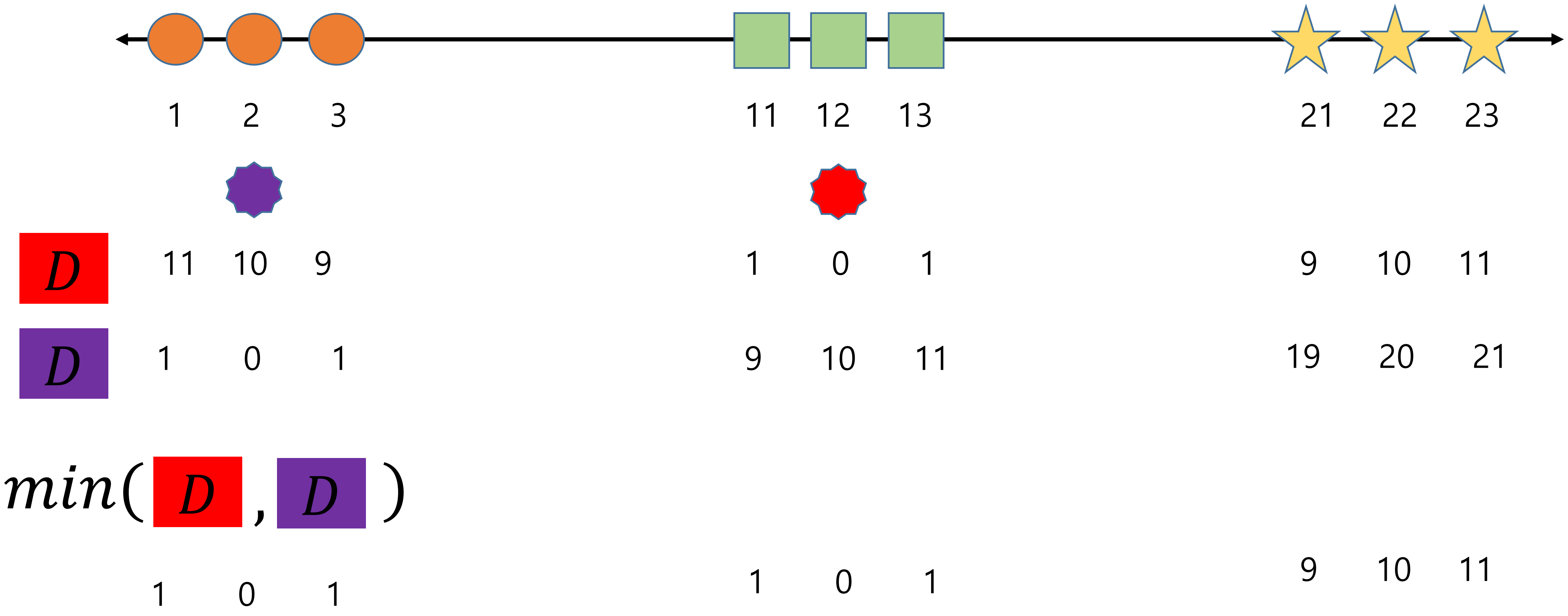
$$sum\left(\frac{D^2}{sum(D^2)}\right) = 1$$

$$D^2/sum(D^2)$$

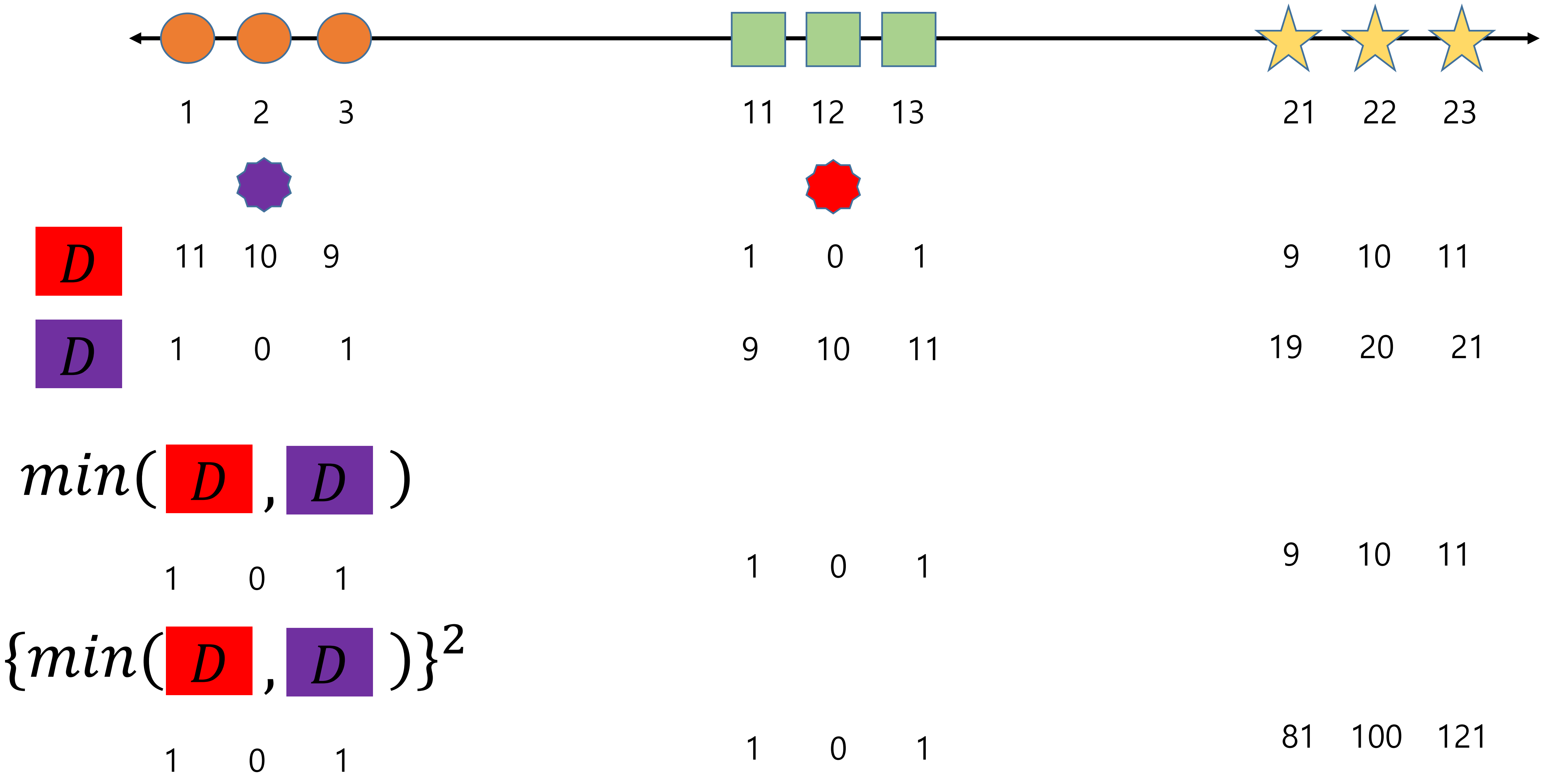
121/606 100/606 81/606 1/606 0/606 1/606 81/606 100/606 121/606

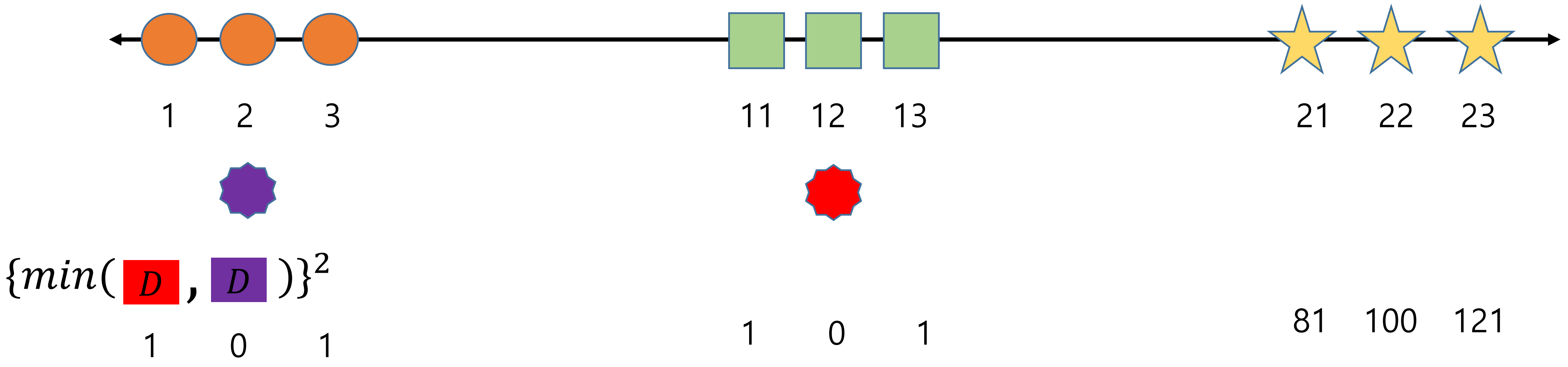

1
2
3
11
12
13
21
22
23

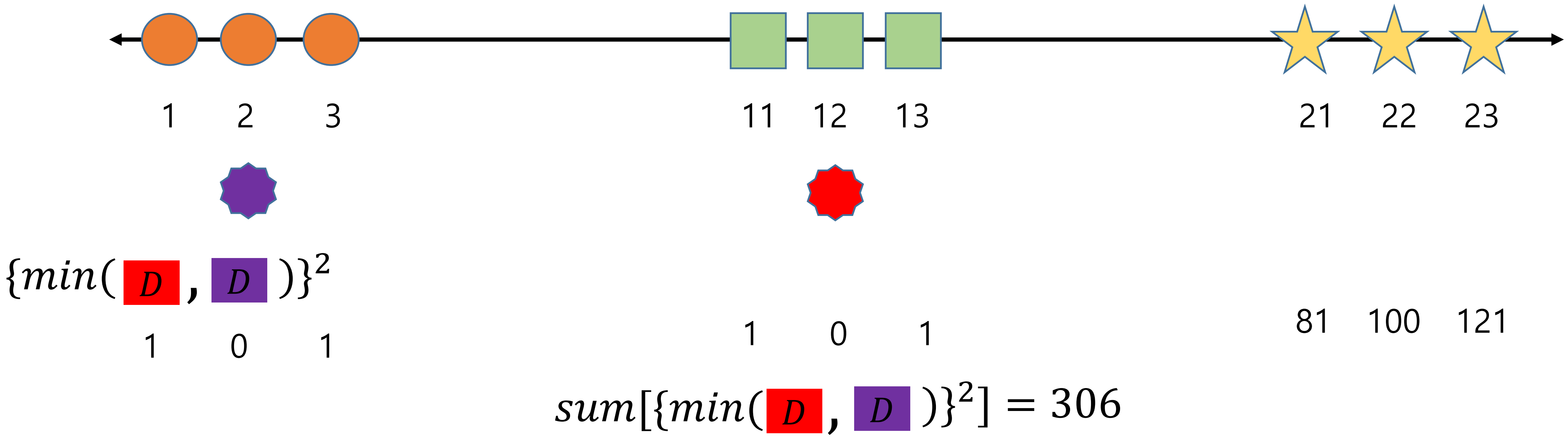

1 2 3 11 12 13 21 22 23
D 11 10 9 1 0 1 9 10 11
D 1 0 1 9 10 11 19 20 21

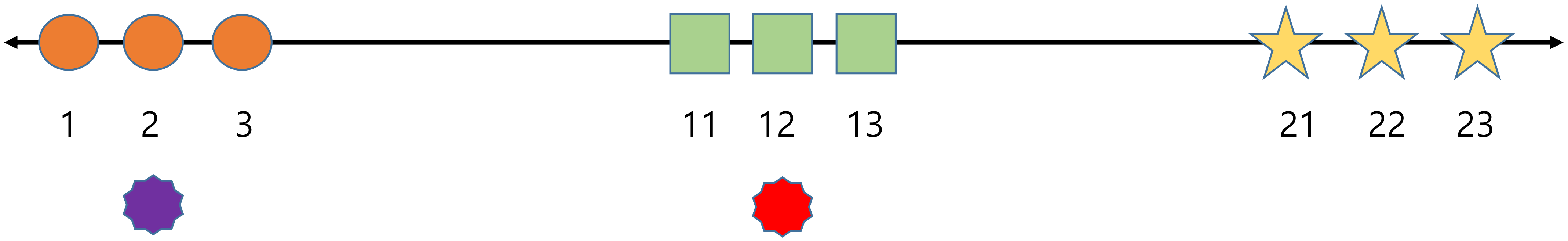


1 2 3
11 12 13
21 22 23
D 11 10 9 1 0 1 9 10 11
D 1 0 1 9 10 11 19 20 21
min(D , D)
1 0 1 1 0 1 9 10 11


1 2 3 11 12 13 21 22 23
D 11 10 9 1 0 1 9 10 11
D 1 0 1 9 10 11 19 20 21
min(D , D)
1 0 1 1 0 1 9 10 11
{min(D , D)}²
1 0 1 1 0 1 81 100 121


1
2
3
11
12
13
21
22
23
{min(D, D)}^2
1
0
1
1
0
1
81
100
121


1 2 3
11 12 13
21 22 23
{min(D , D)}²
1 0 1
1 0 1
81 100 121
sum[{min(D , D)}²] = 306



$\{min(D, D)\}^2$

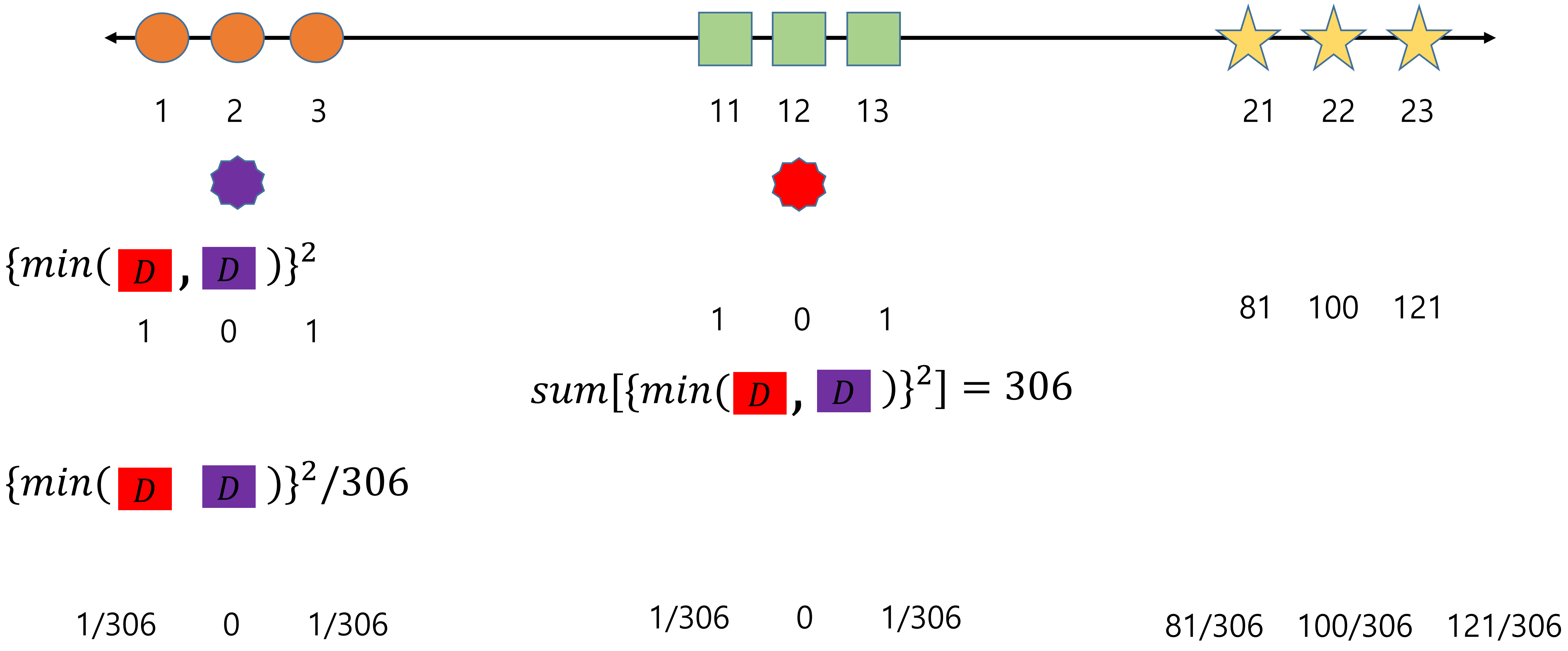
1 0 1

1 0 1

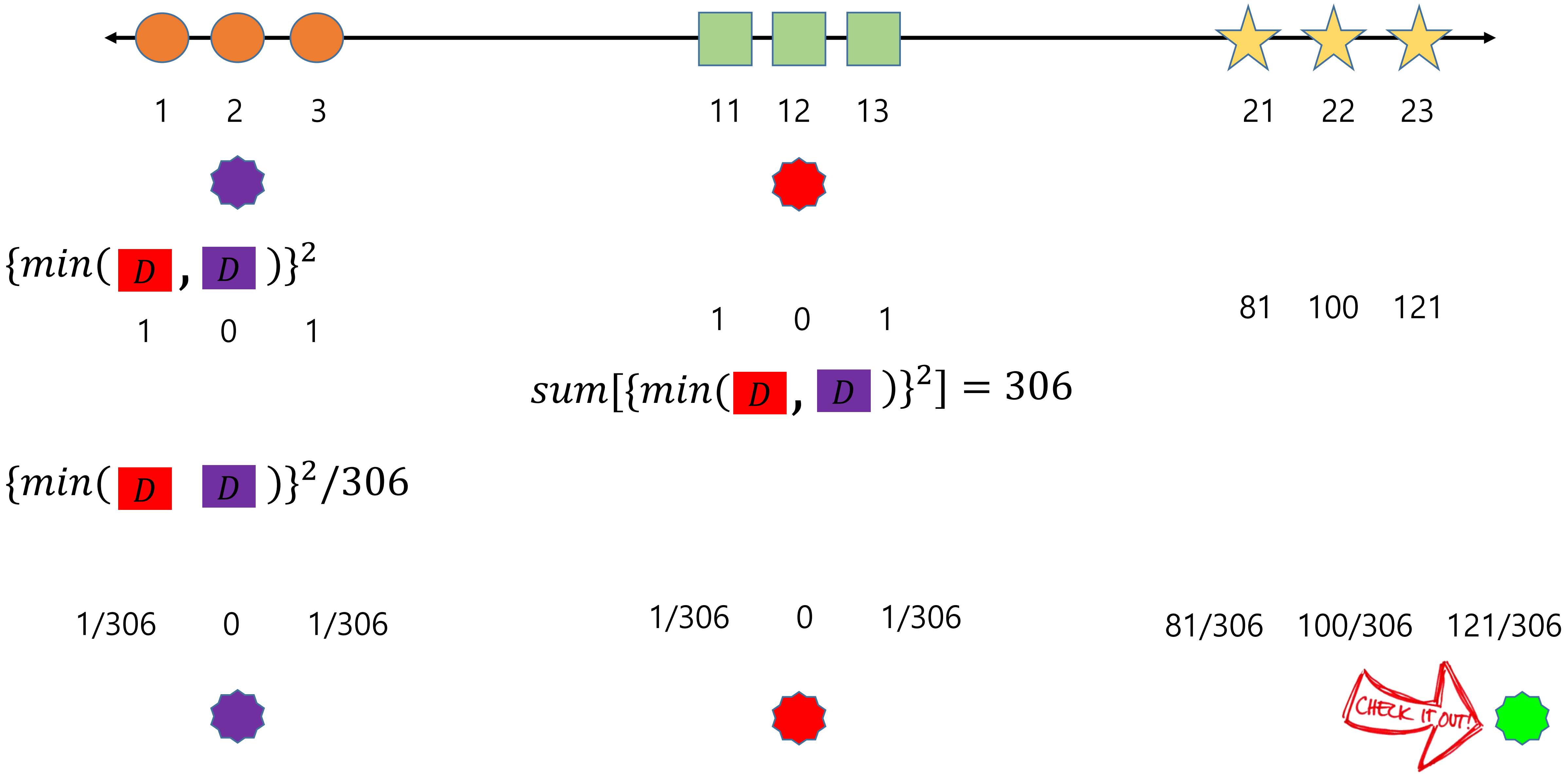
81 100 121

$$sum[\{min(D, D)\}^2] = 306$$

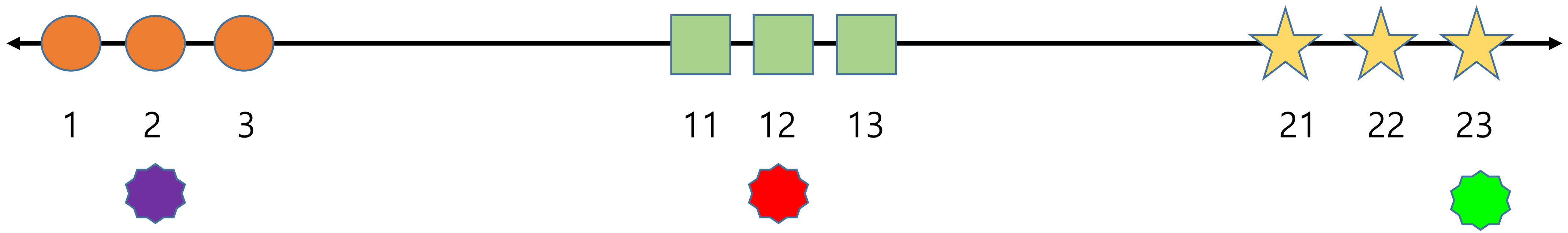
$\{min(D \quad D)\}^2/306$

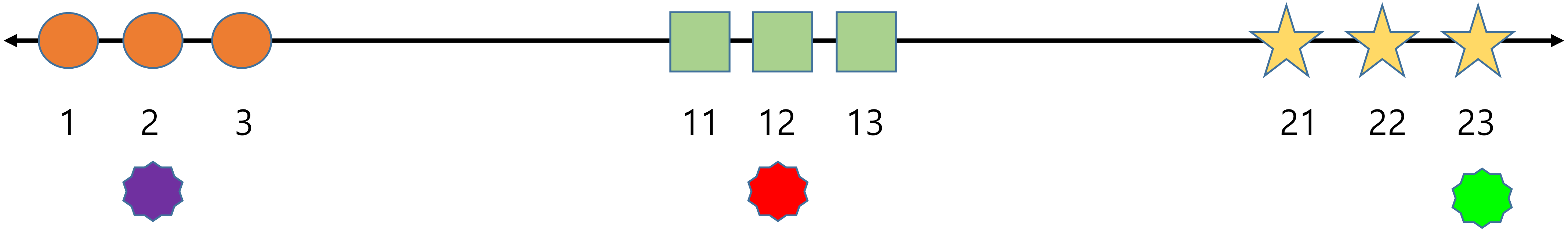


1 2 3
11 12 13
21 22 23
$\{min(\boxed{D}, \boxed{D})\}^2$
1 0 1
1 0 1
81 100 121
$sum[\{min(\boxed{D}, \boxed{D})\}^2] = 306$
$\{min(\boxed{D}\ \boxed{D})\}^2/306$
1/306 0 1/306
1/306 0 1/306
81/306 100/306 121/306



1 2 3
11 12 13
21 22 23
{min(D , D)}²
1 0 1
1 0 1
81 100 121
sum[{min(D , D)}²] = 306
{min(D D)}²/306
1/306 0 1/306
1/306 0 1/306
81/306 100/306 121/306
CHECK IT OUT!


1
2
3
11
12
13
21
22
23



2.2 The k-means++ algorithm

We propose a specific way of choosing centers for the `k-means` algorithm. In particular, let $D(x)$ denote the shortest distance from a data point to the closest center we have already chosen. Then, we define the following algorithm, which we call `k-means++`.

1a. Take one center c_1, chosen uniformly at random from $\mathcal{X}$.

1b. Take a new center c_i, choosing $x \in \mathcal{X}$ with probability $\frac{D(x)^2}{\sum_{x \in \mathcal{X}} D(x)^2}$.

1c. Repeat Step 1b. until we have taken k centers altogether.

2-4. Proceed as with the standard `k-means` algorithm.