

Comparing more than two observations

CLUSTER ANALYSIS IN R



Dmitriy Gorenshteyn

Lead Data Scientist, Memorial Sloan
Kettering Cancer Center

The closest observation to a pair

	1	2	3
2	11.7		
3	16.8	18.0	
4	10.0	20.6	15.8

- Is 2 is closest to group 1,4?
- Is 3 is closest to group 1,4?

Linkage criteria: complete

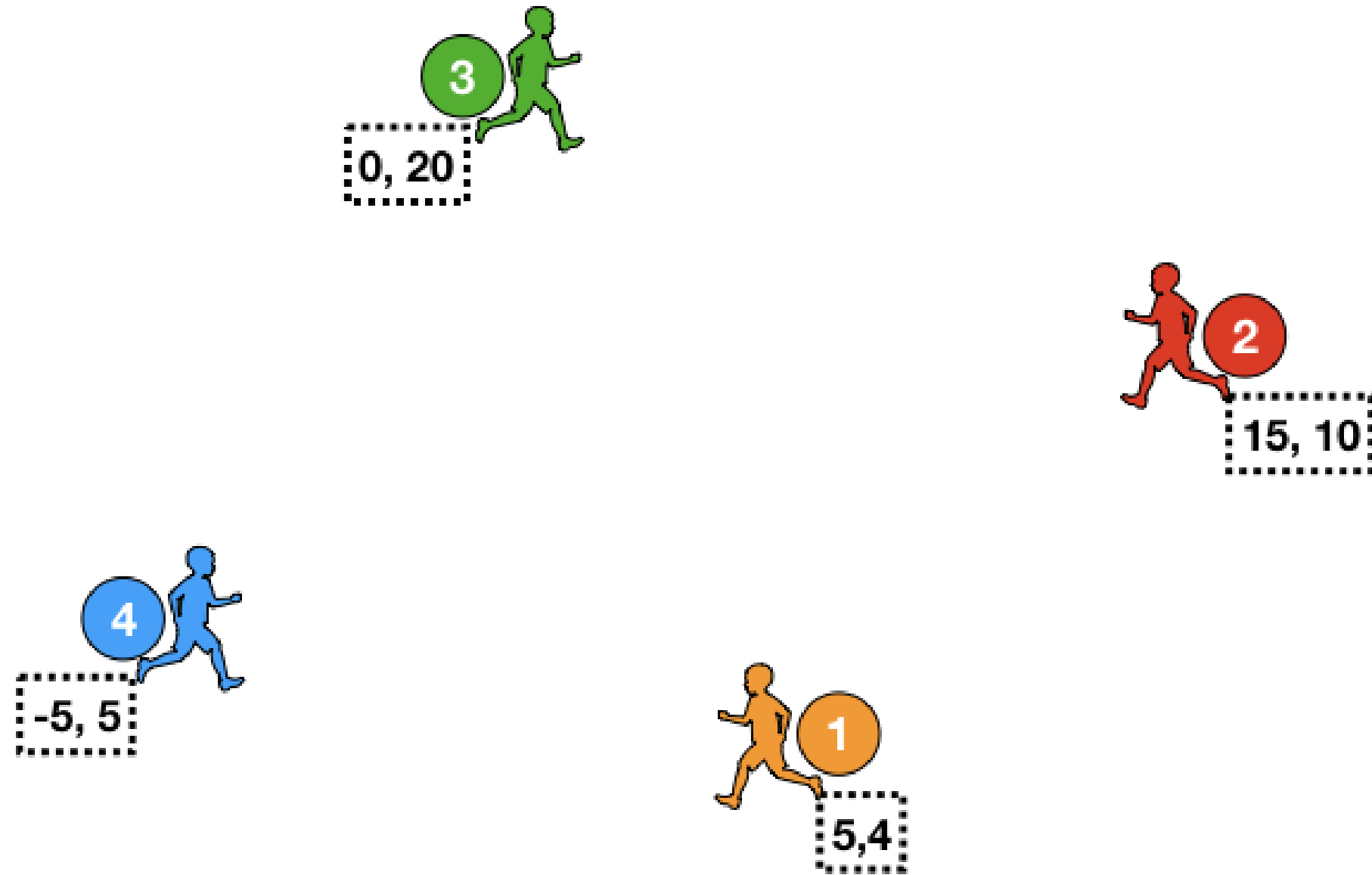
	1	2	3
2	11.7		
3	16.8	18.0	
4	10.0	20.6	15.8

- Is 2 is closest to group 1,4?
 - $\max(D(2,1), D(2,4)) = \mathbf{20.6}$
- Is 3 is closest to group 1,4?
 - $\max(D(3,1), D(3,4)) = \mathbf{16.8}$

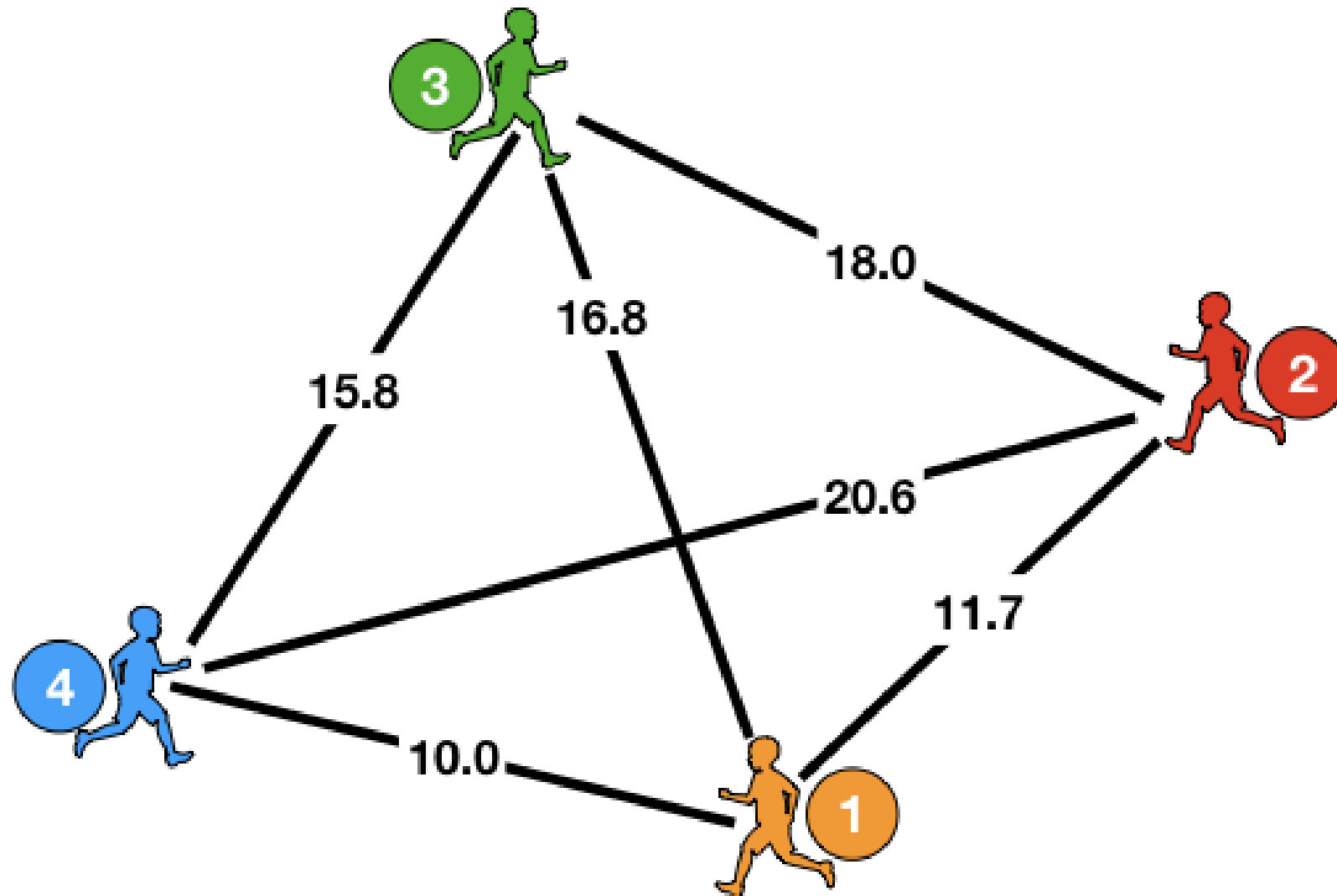
Hierarchical clustering

Complete Linkage: maximum distance between two sets

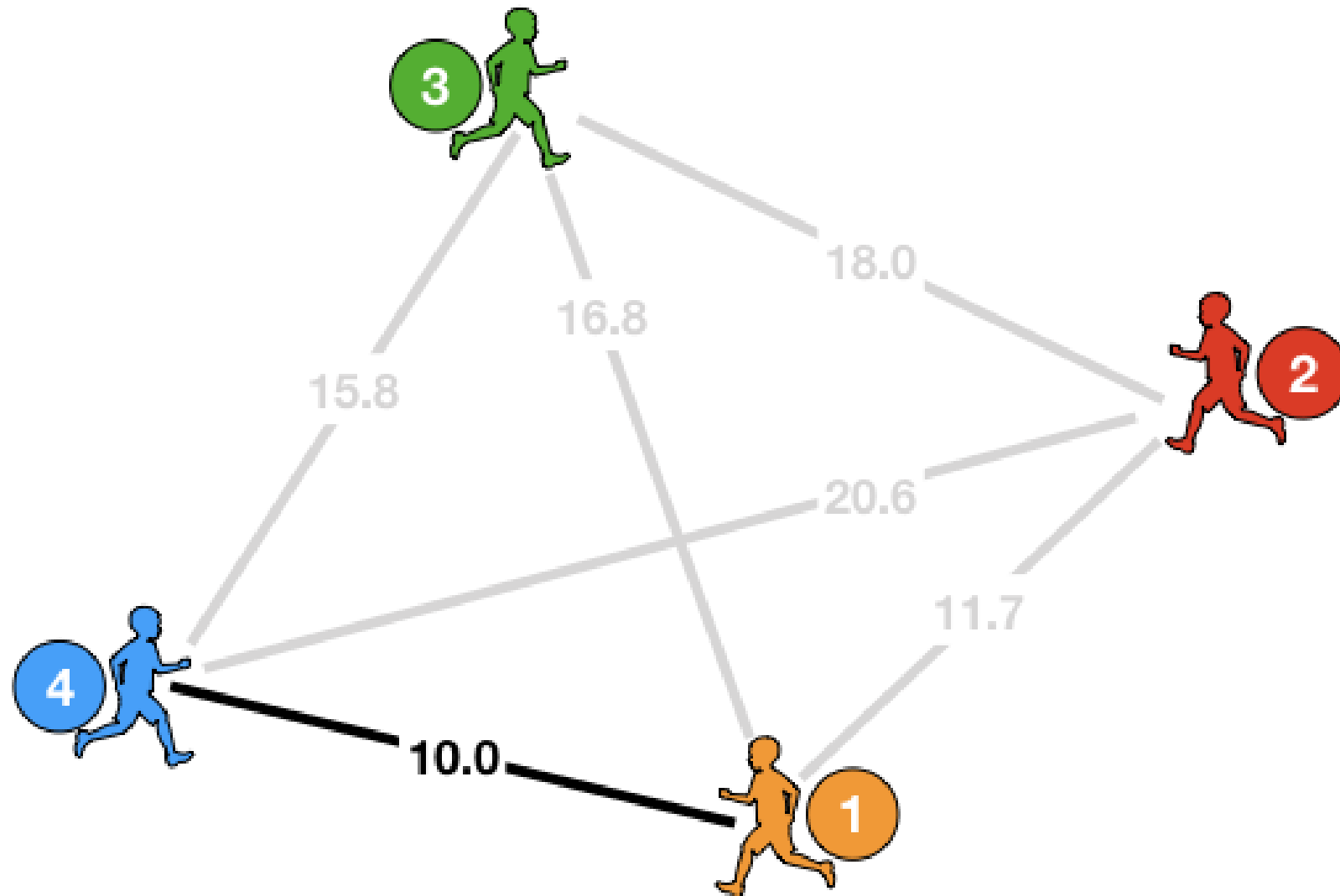
Grouping with linkage & distance



Grouping with linkage & distance



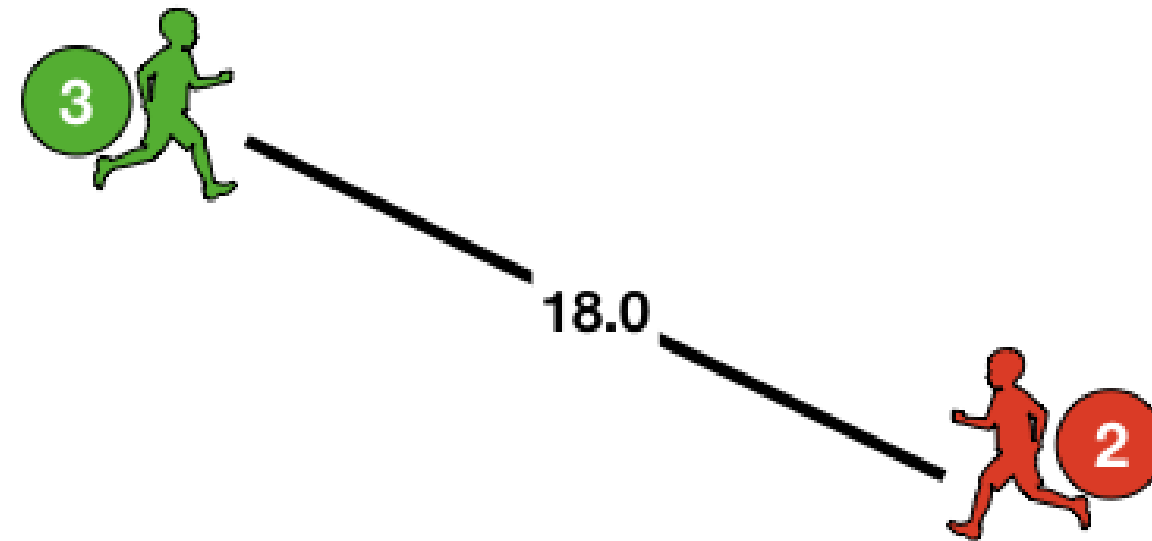
Grouping with linkage & distance



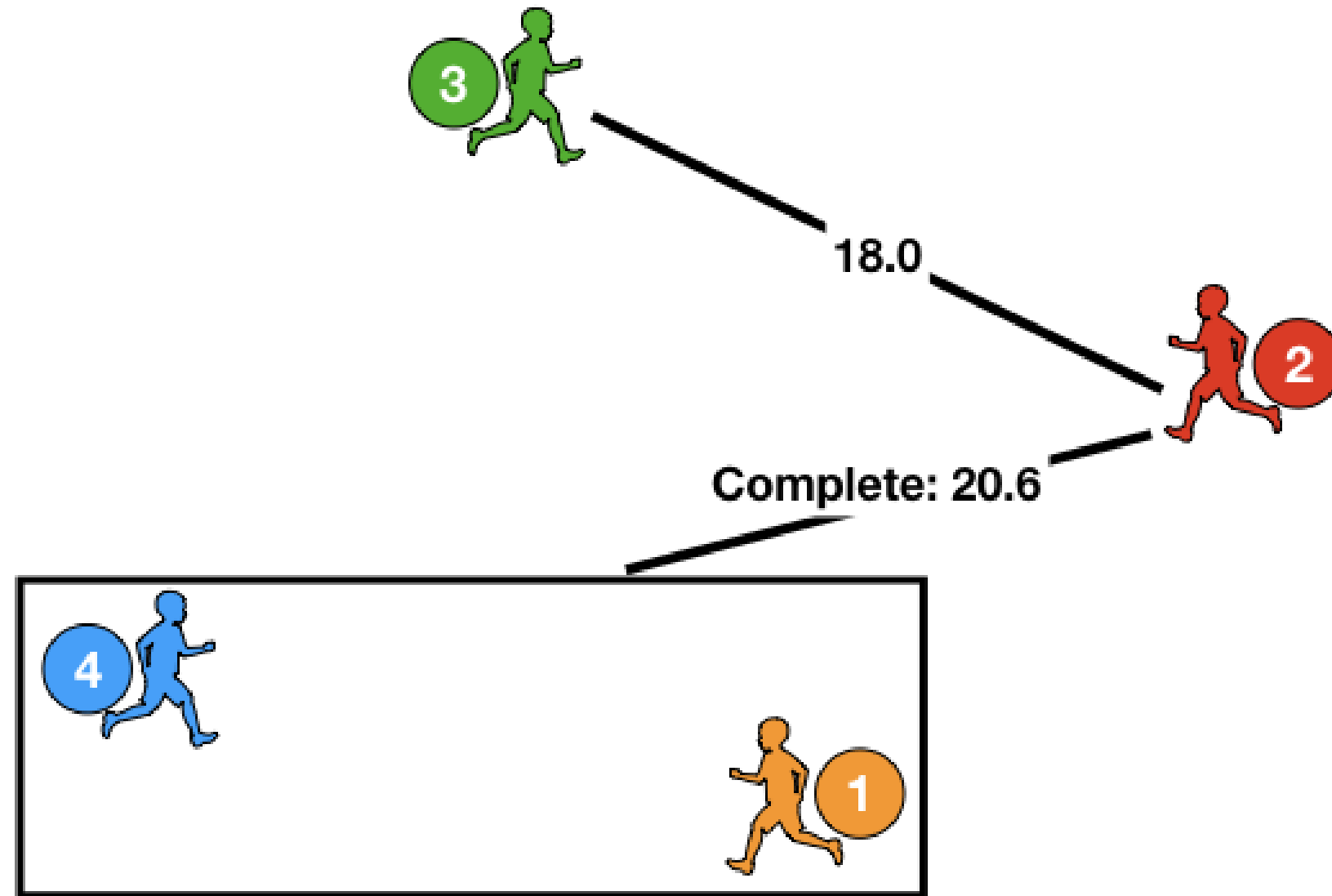
Grouping with linkage & distance



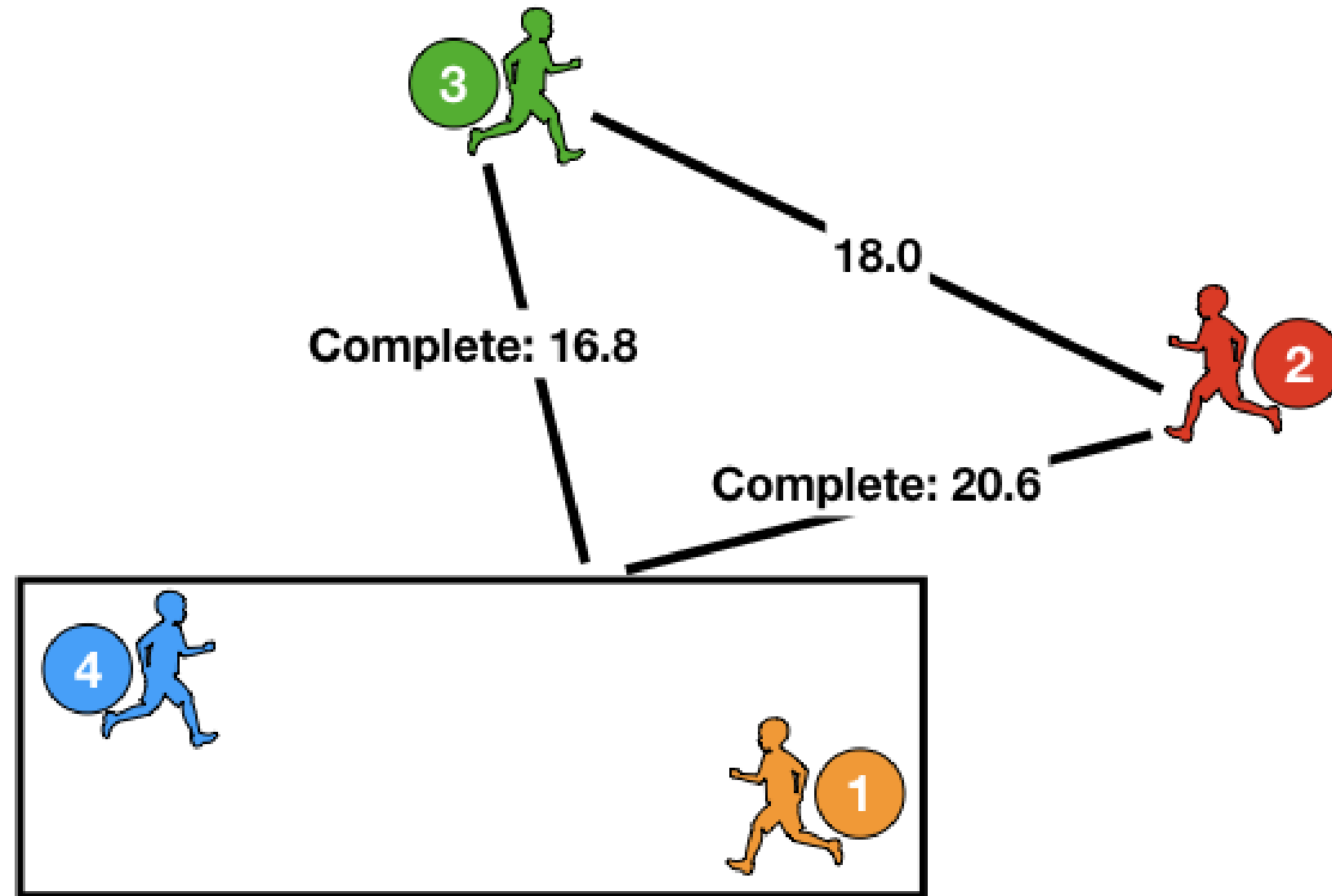
Grouping with linkage & distance



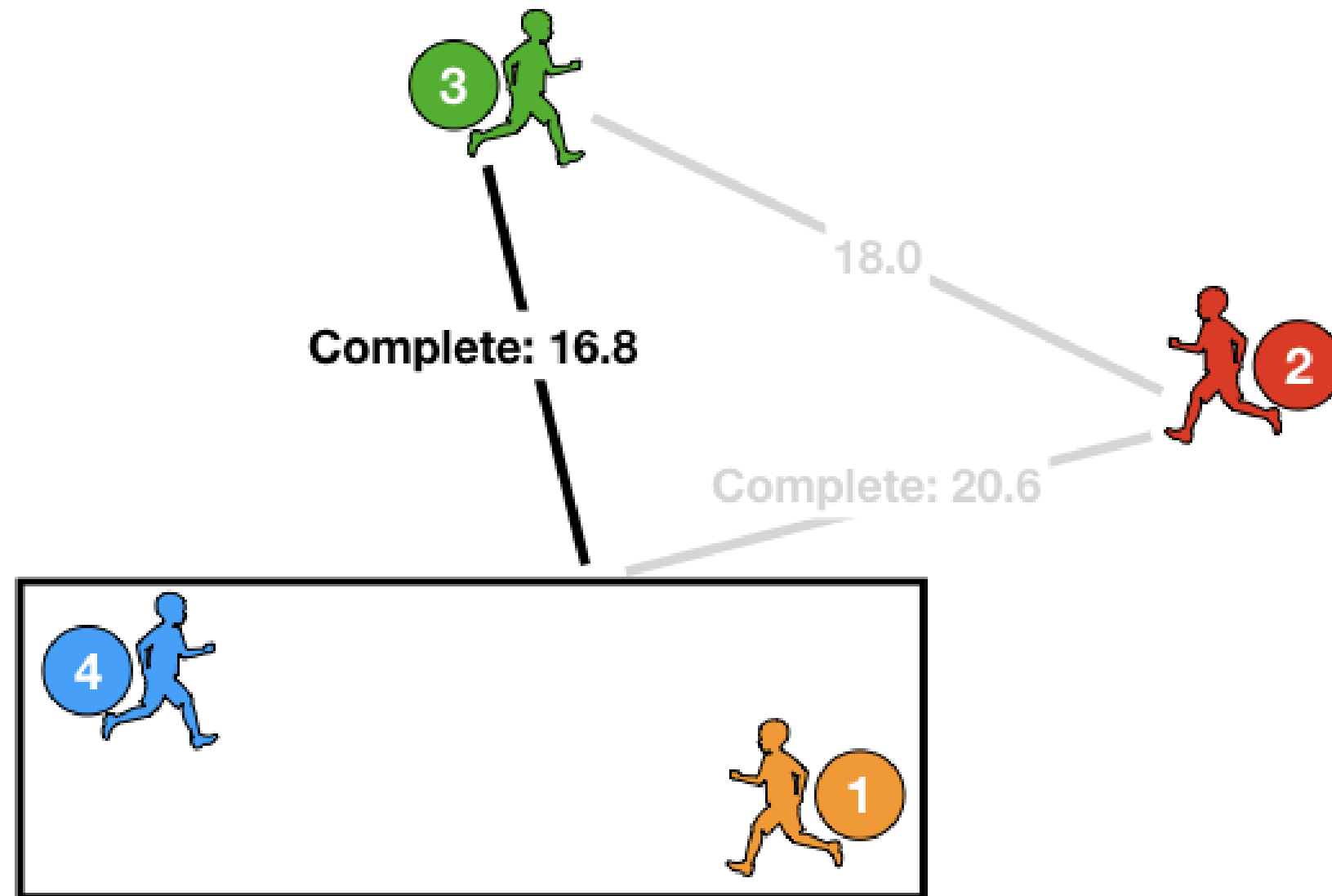
Grouping with linkage & distance



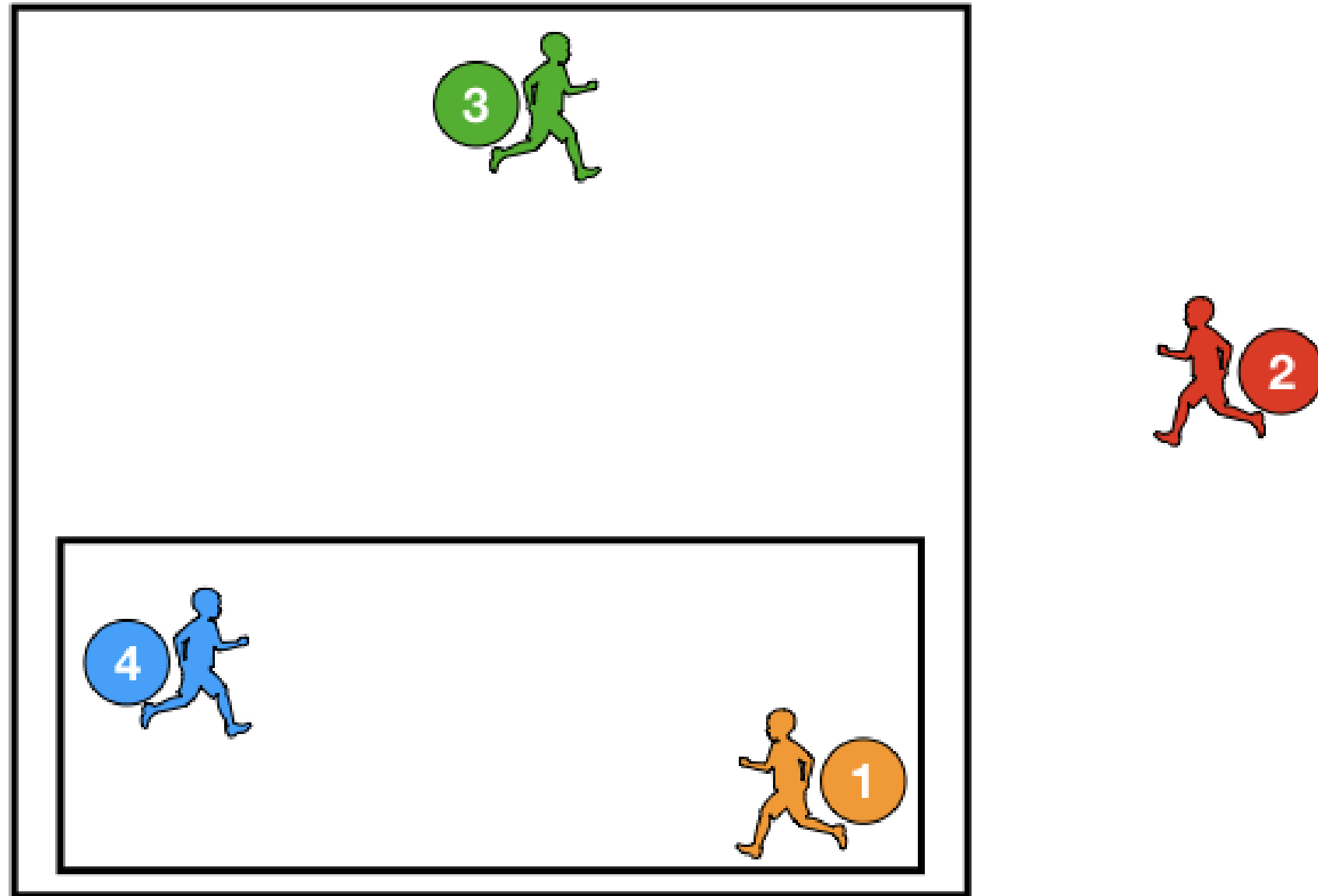
Grouping with linkage & distance



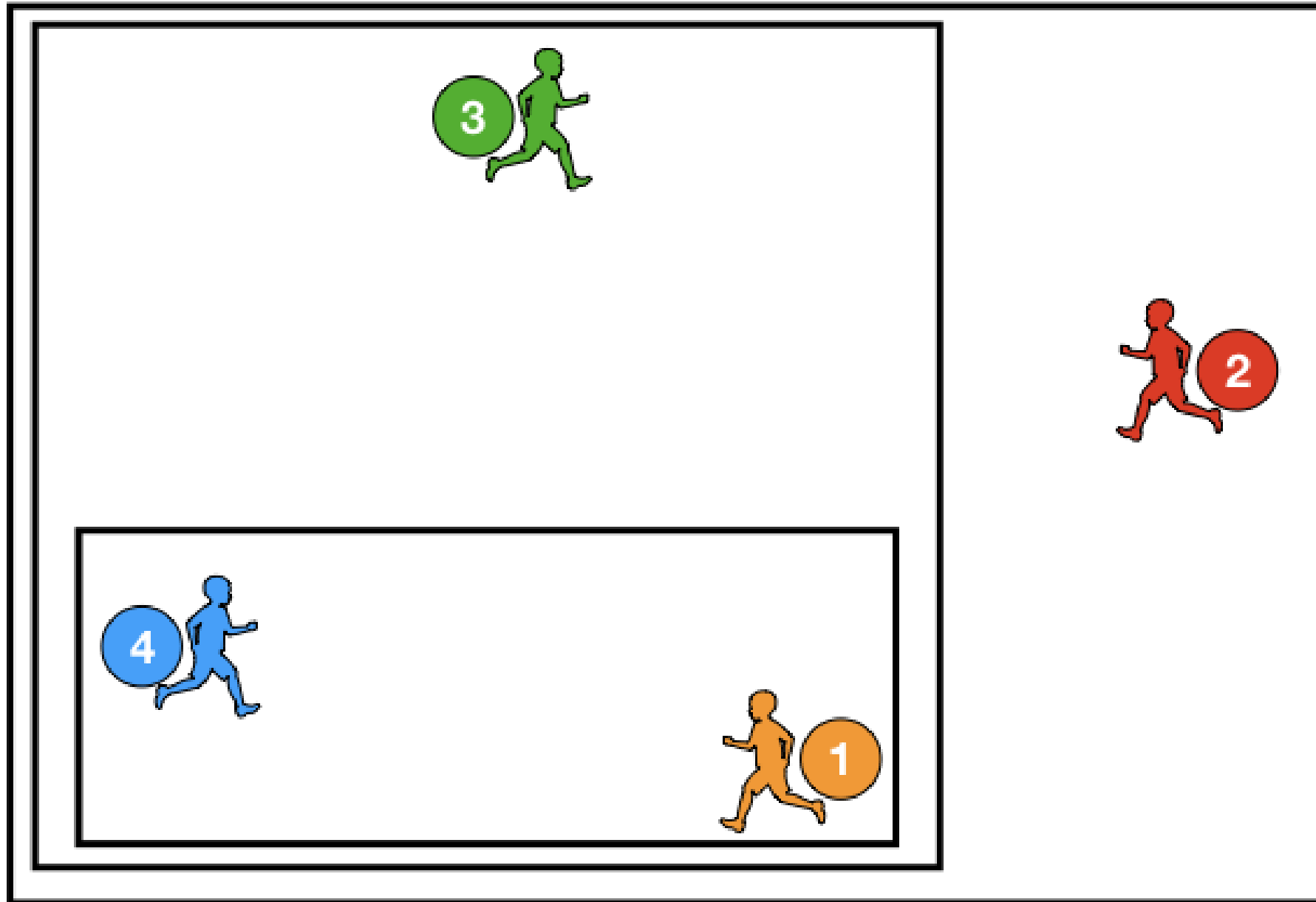
Grouping with linkage & distance



Grouping with linkage & distance



Grouping with linkage & distance



Linkage criteria

Complete Linkage: maximum distance between two sets

Single Linkage: minimum distance between two sets

Average Linkage: average distance between two sets

Let's practice!
CLUSTER ANALYSIS IN R

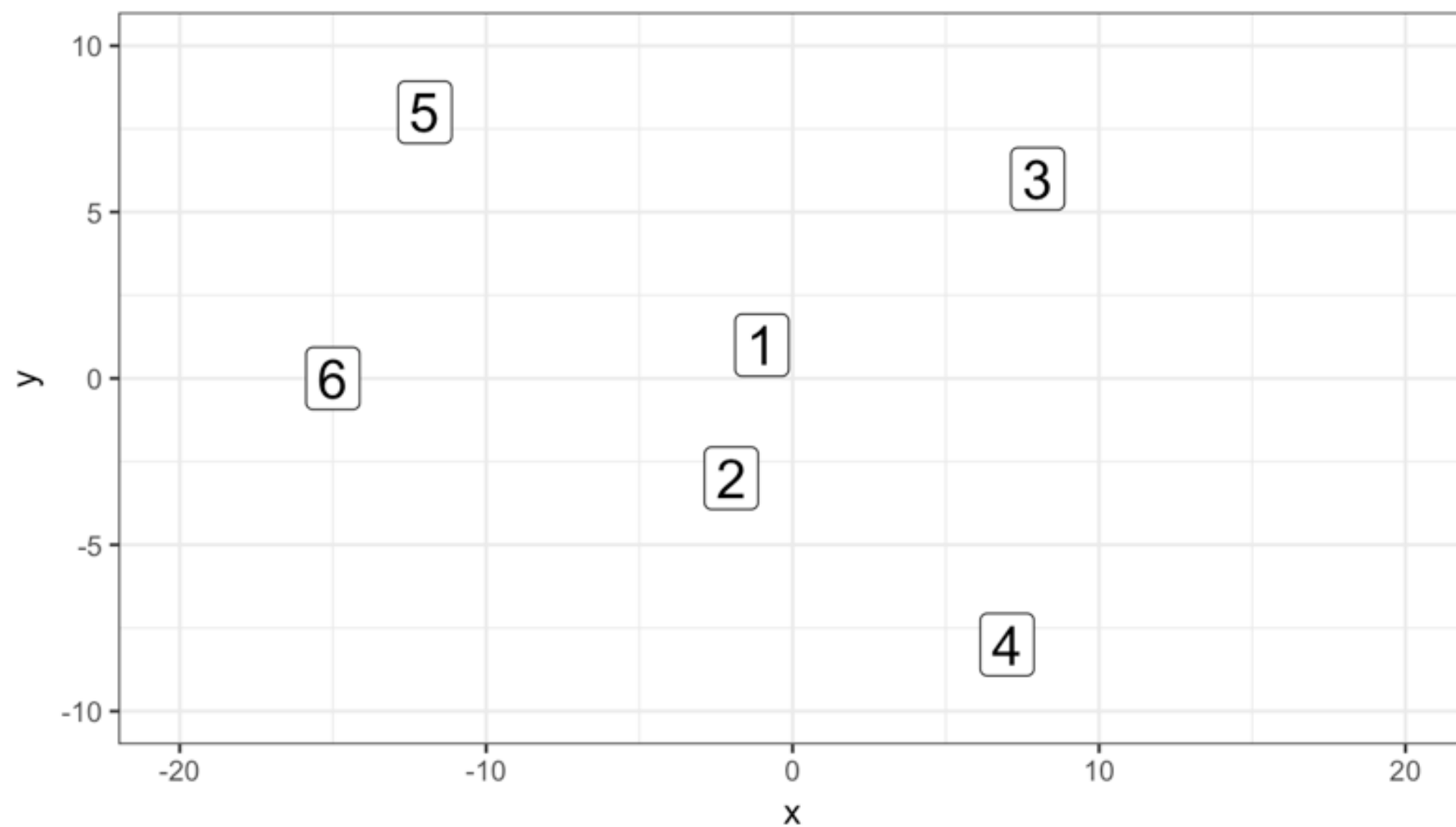
Capturing K clusters

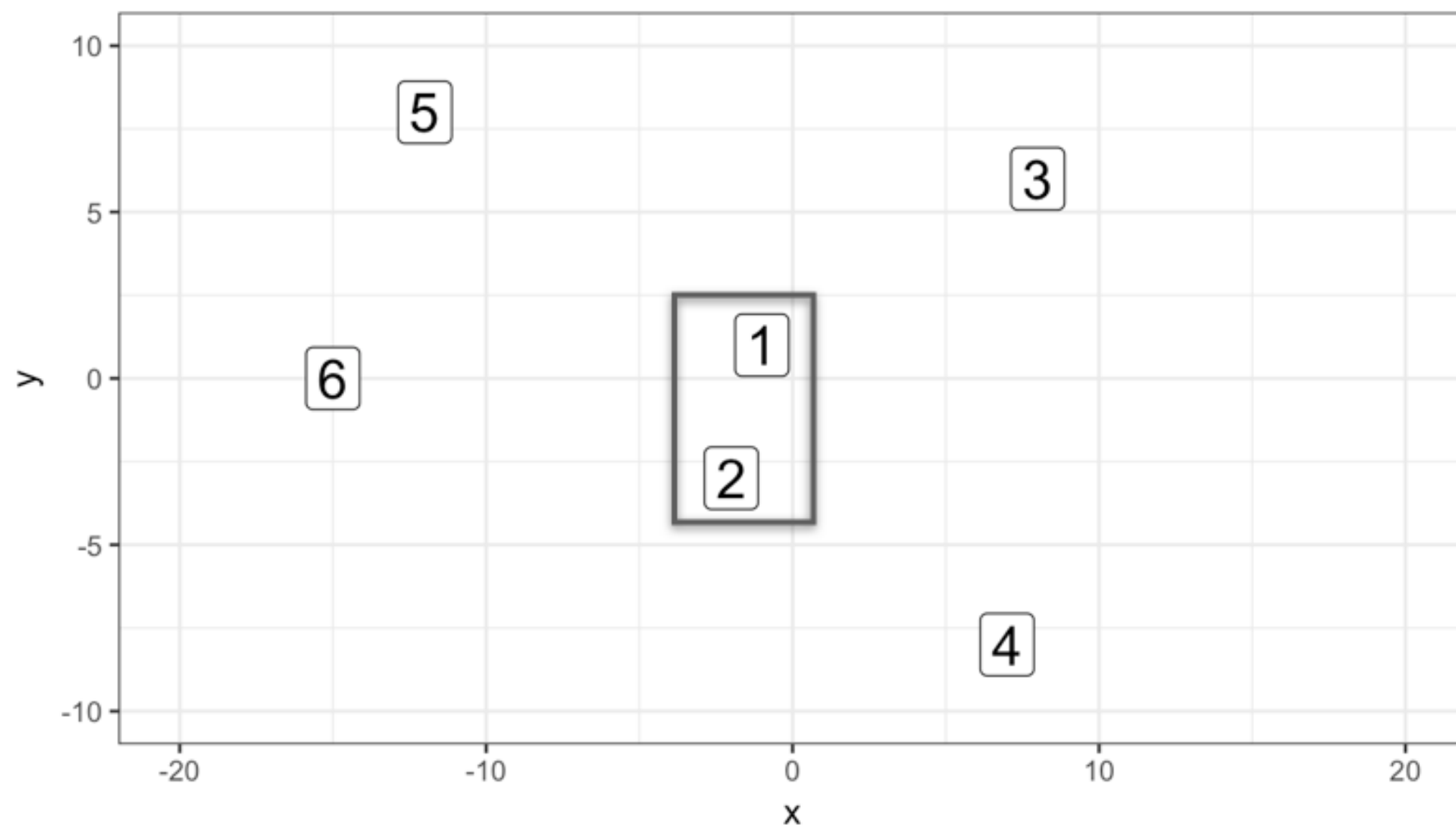
CLUSTER ANALYSIS IN R

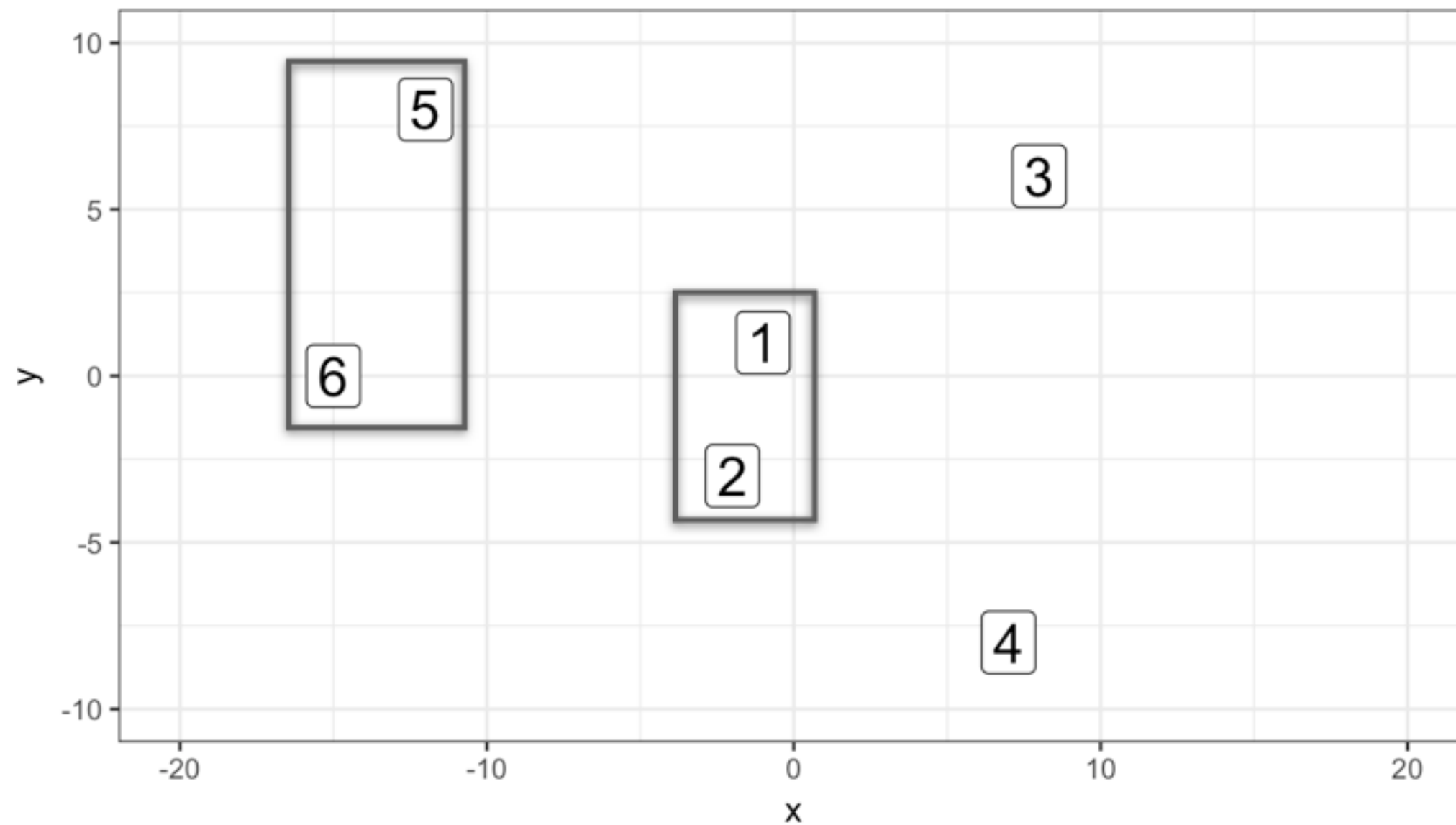


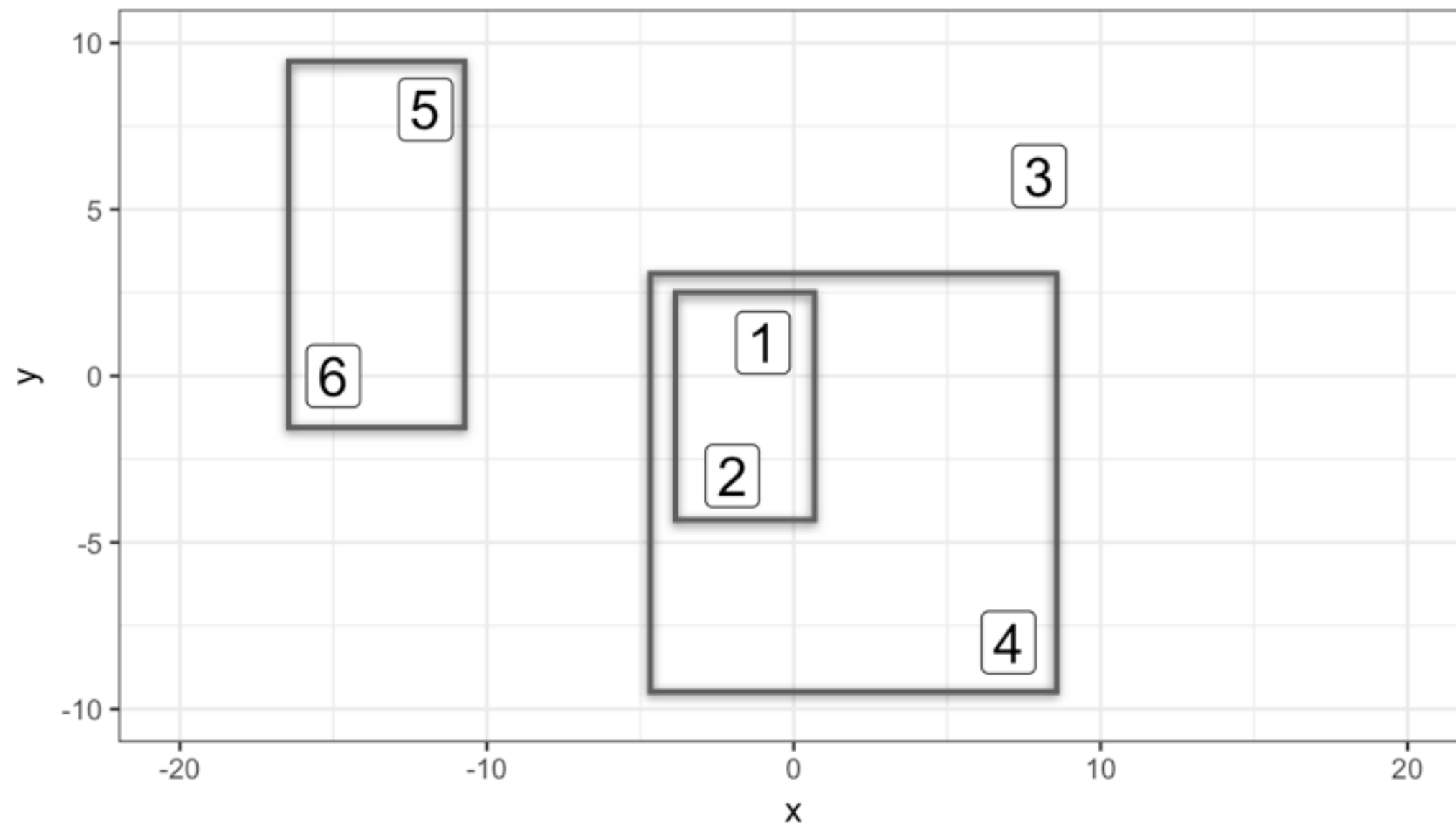
Dmitriy Gorenshteyn

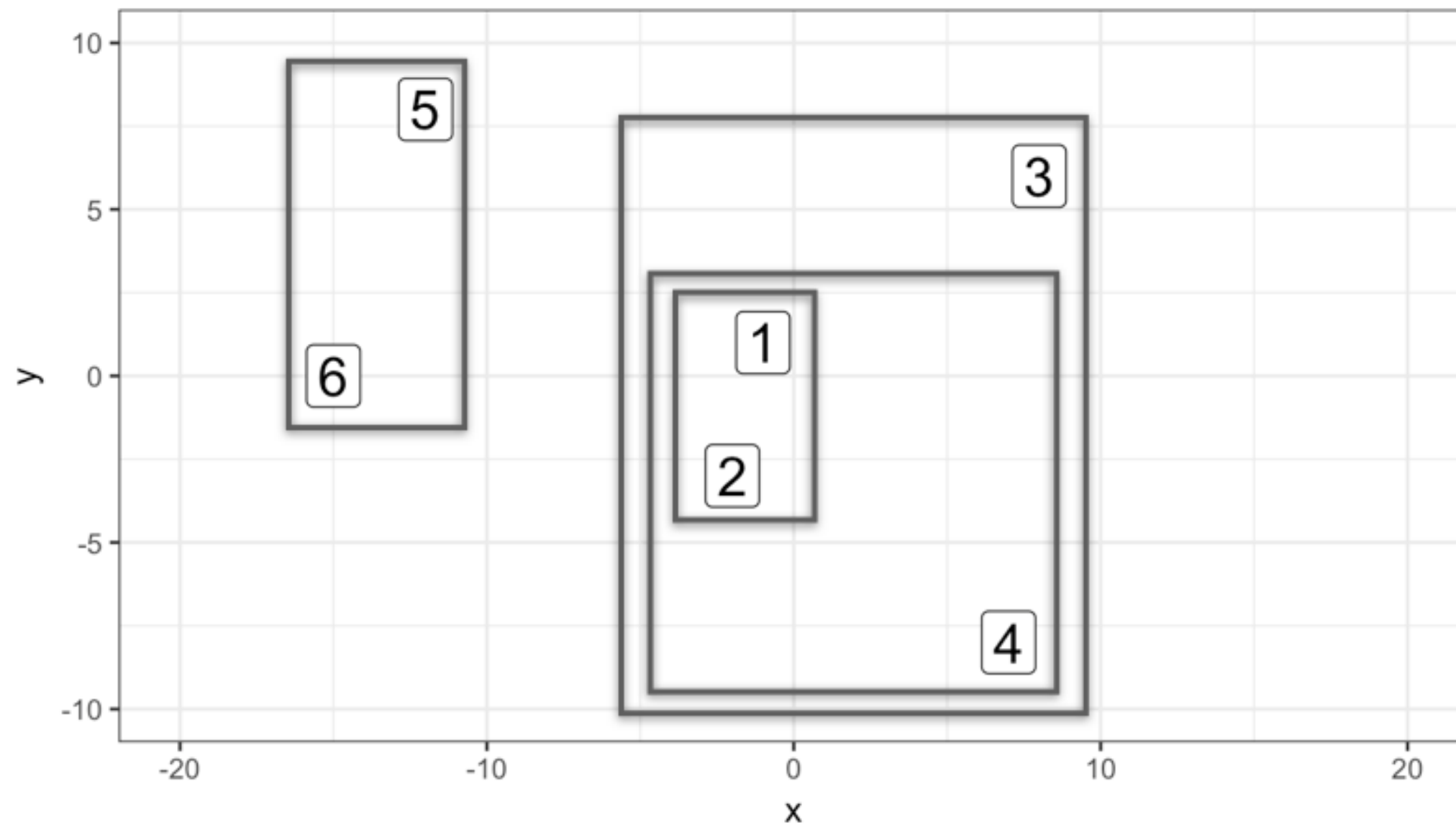
Lead Data Scientist, Memorial Sloan
Kettering Cancer Center

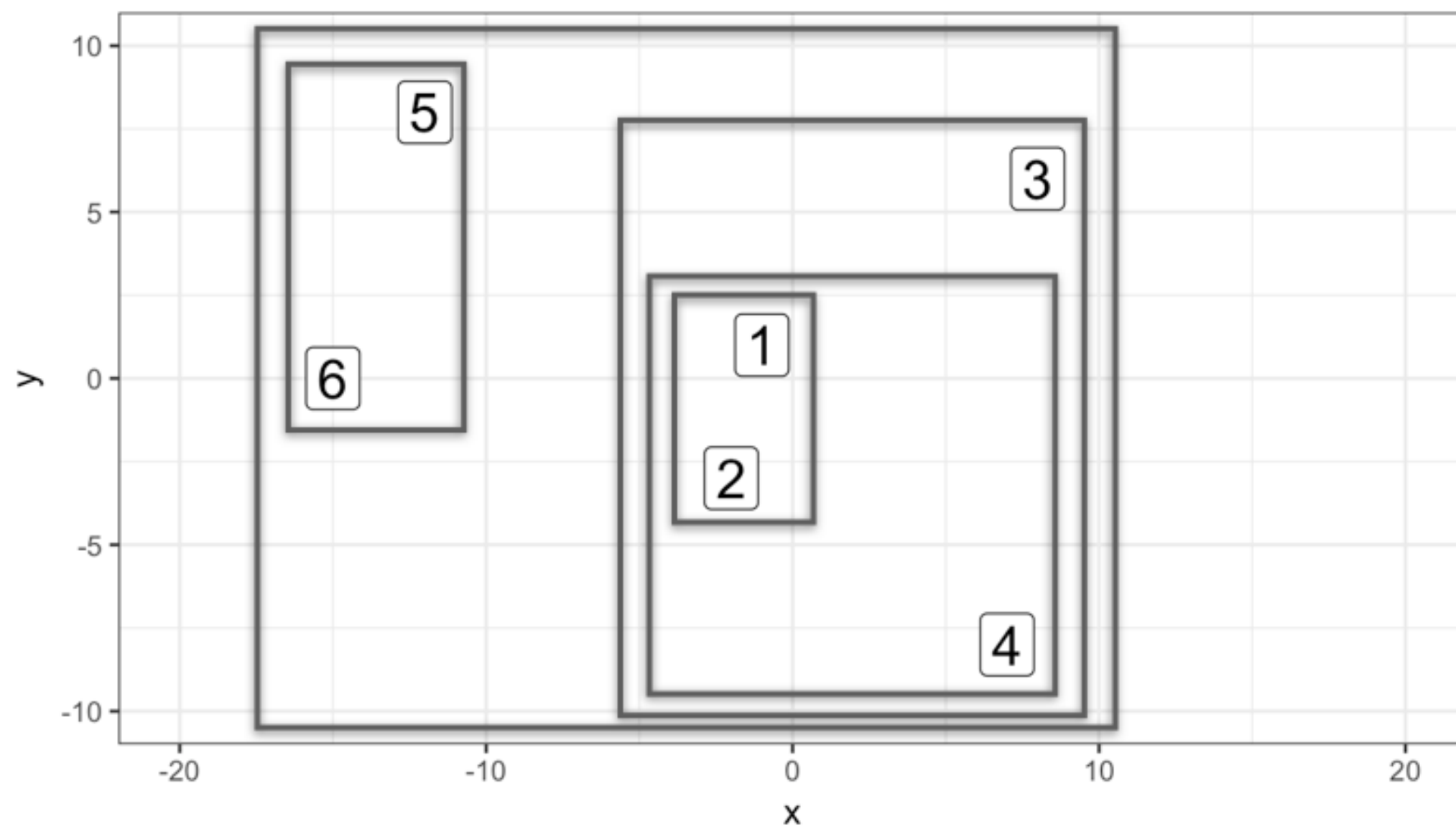


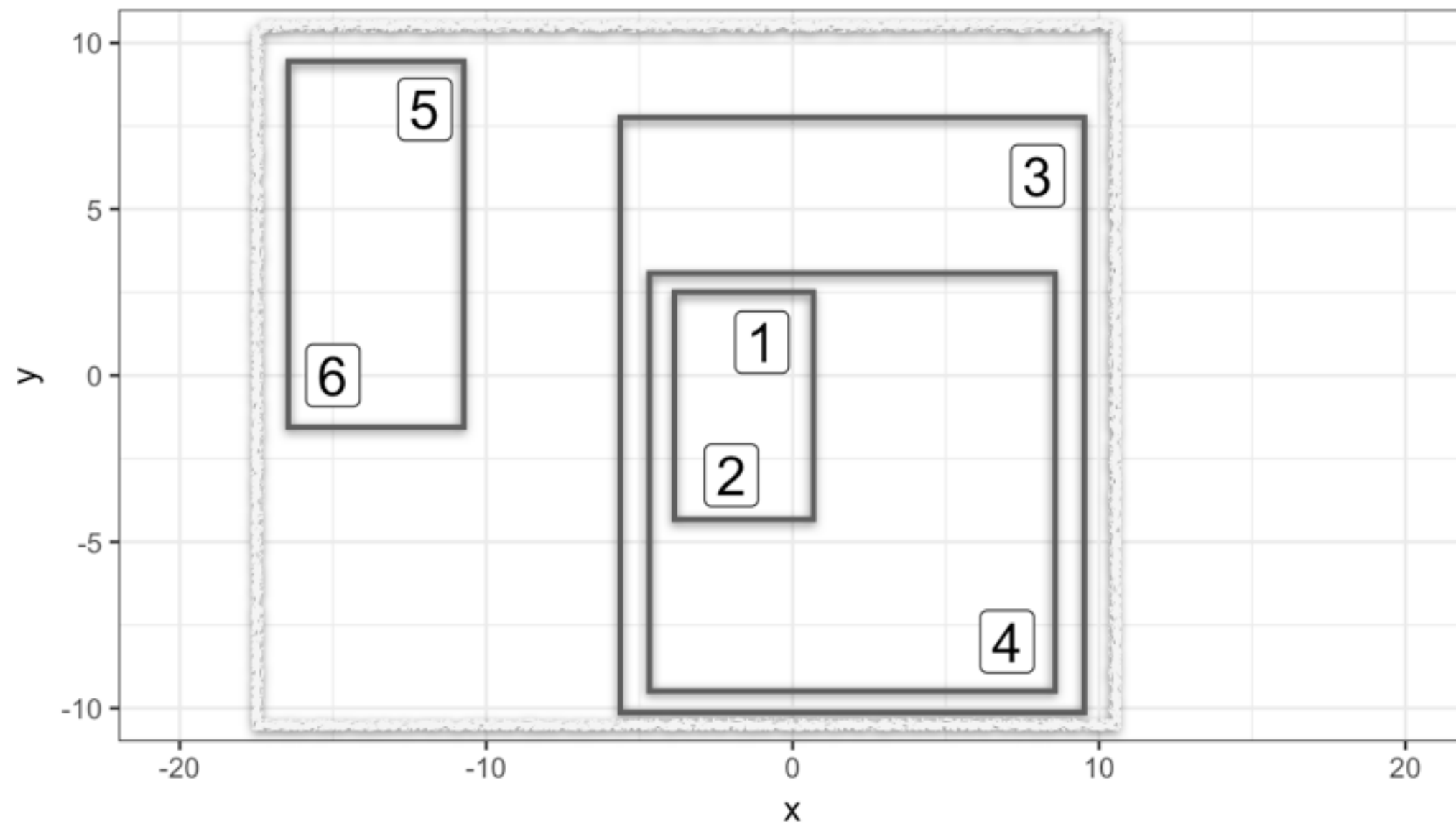


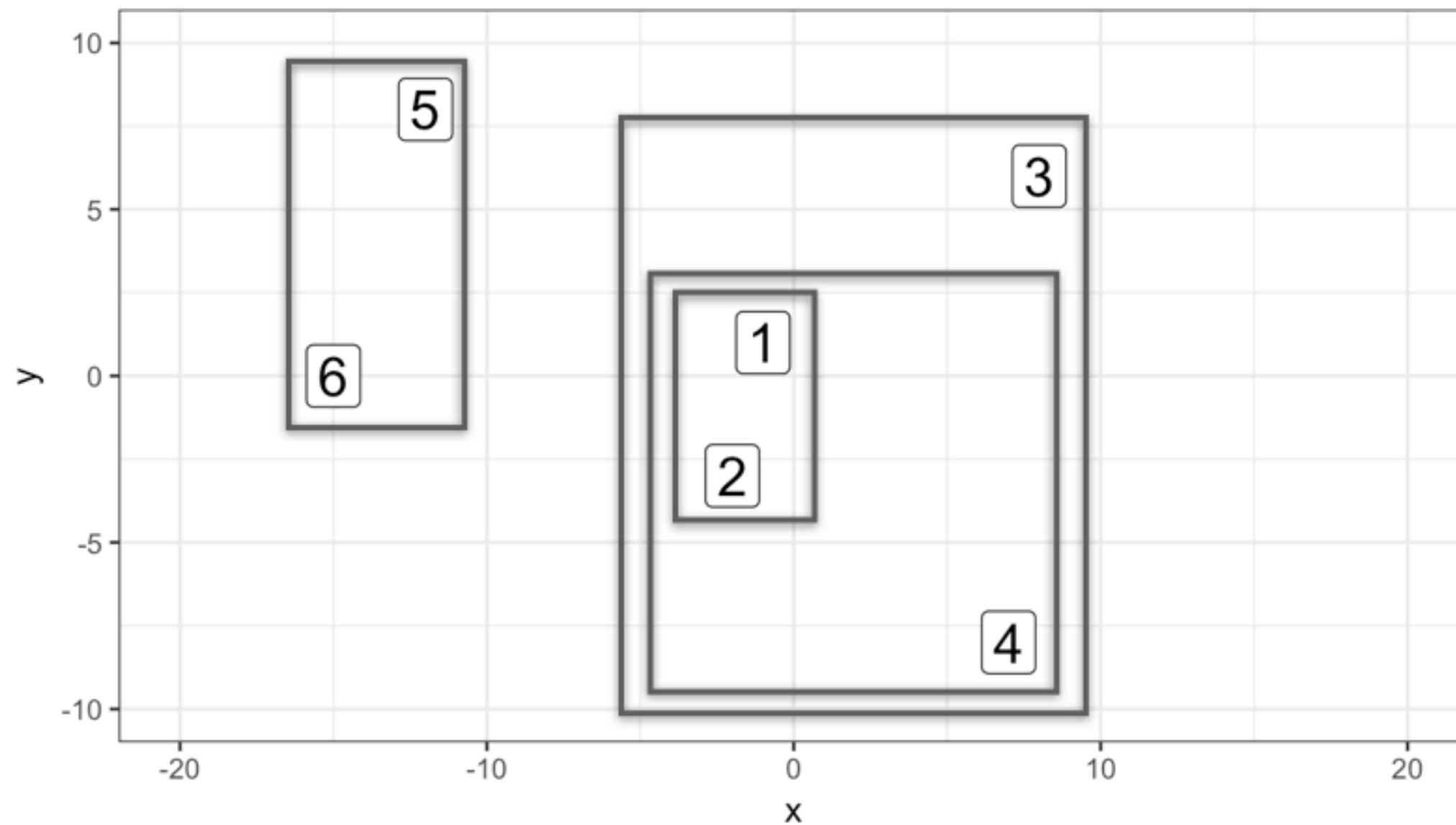


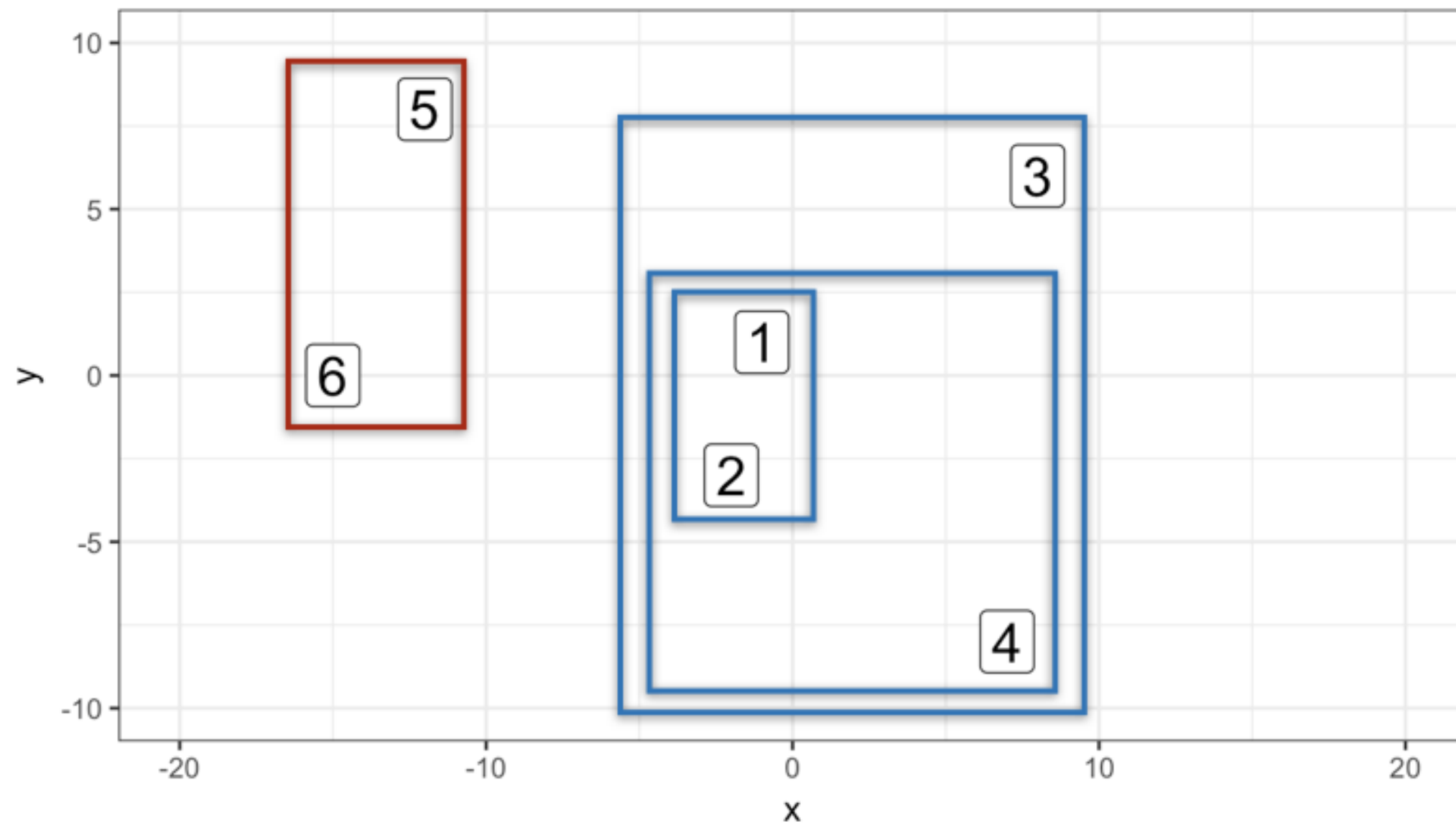


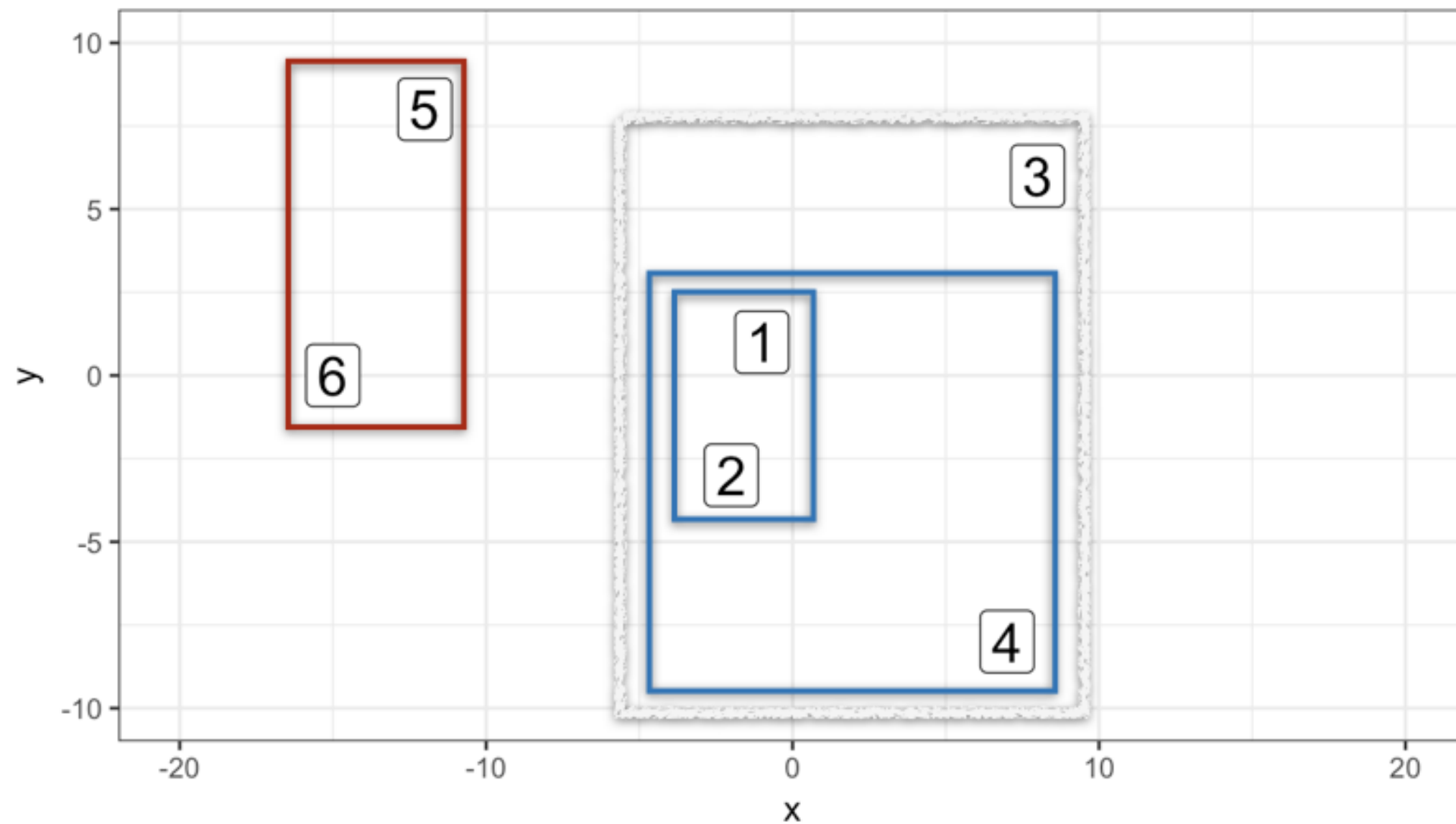


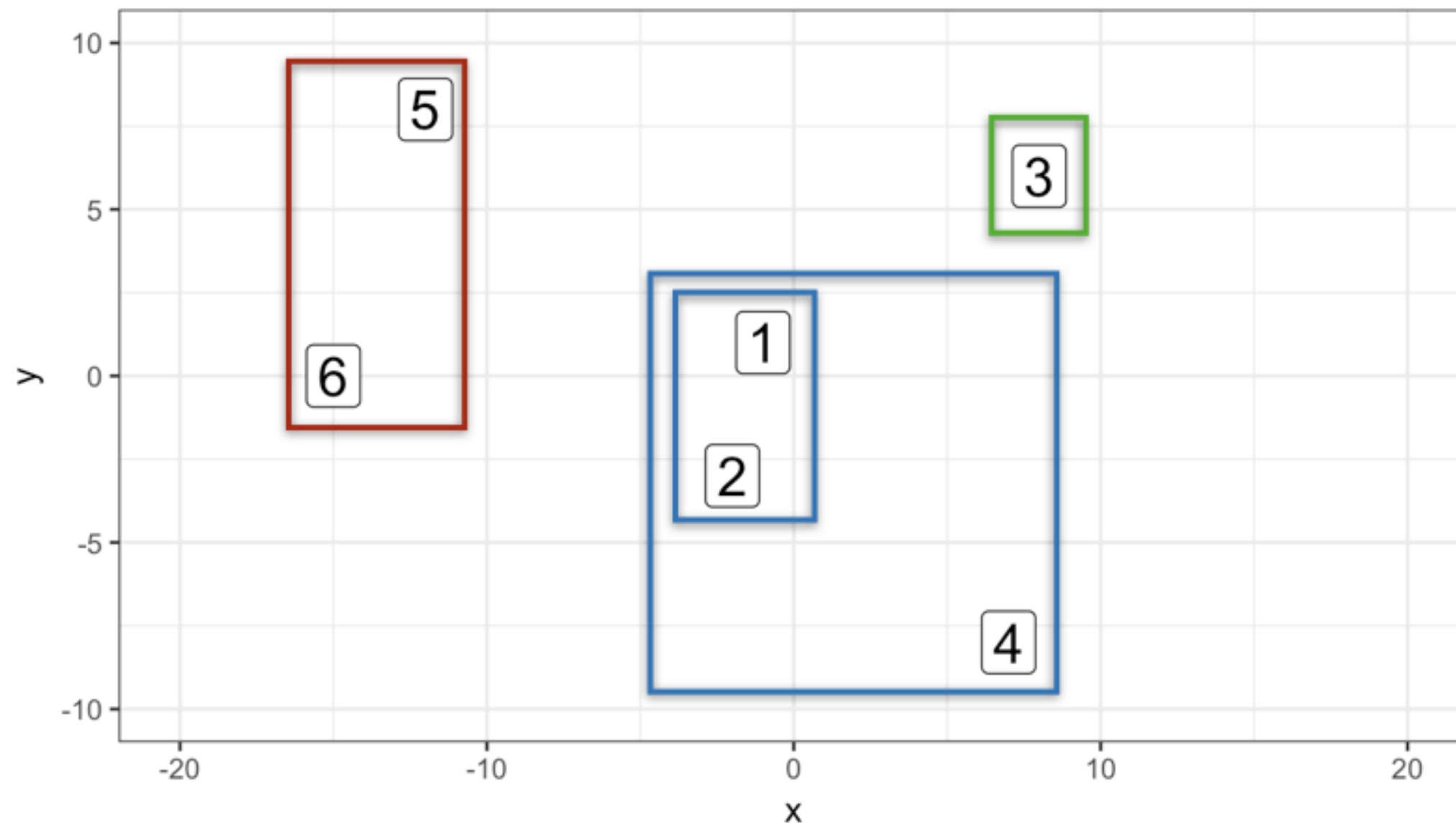












Hierarchical clustering in R

```
print(players)
      x      y
  <dbl> <dbl>
1    -1     1
2    -2    -3
3     8     6
4     7    -8
5   -12     8
6   -15     0

dist_players <- dist(players, method = 'euclidean')
hc_players  <- hclust(dist_players, method = 'complete')
```

Extracting K clusters

```
cluster_assignments <- cutree(hc_players, k = 2)
print(cluster_assignments)
[1] 1 1 1 1 2 2
library(dplyr)
players_clustered <- mutate(players, cluster = cluster_assignments)
```

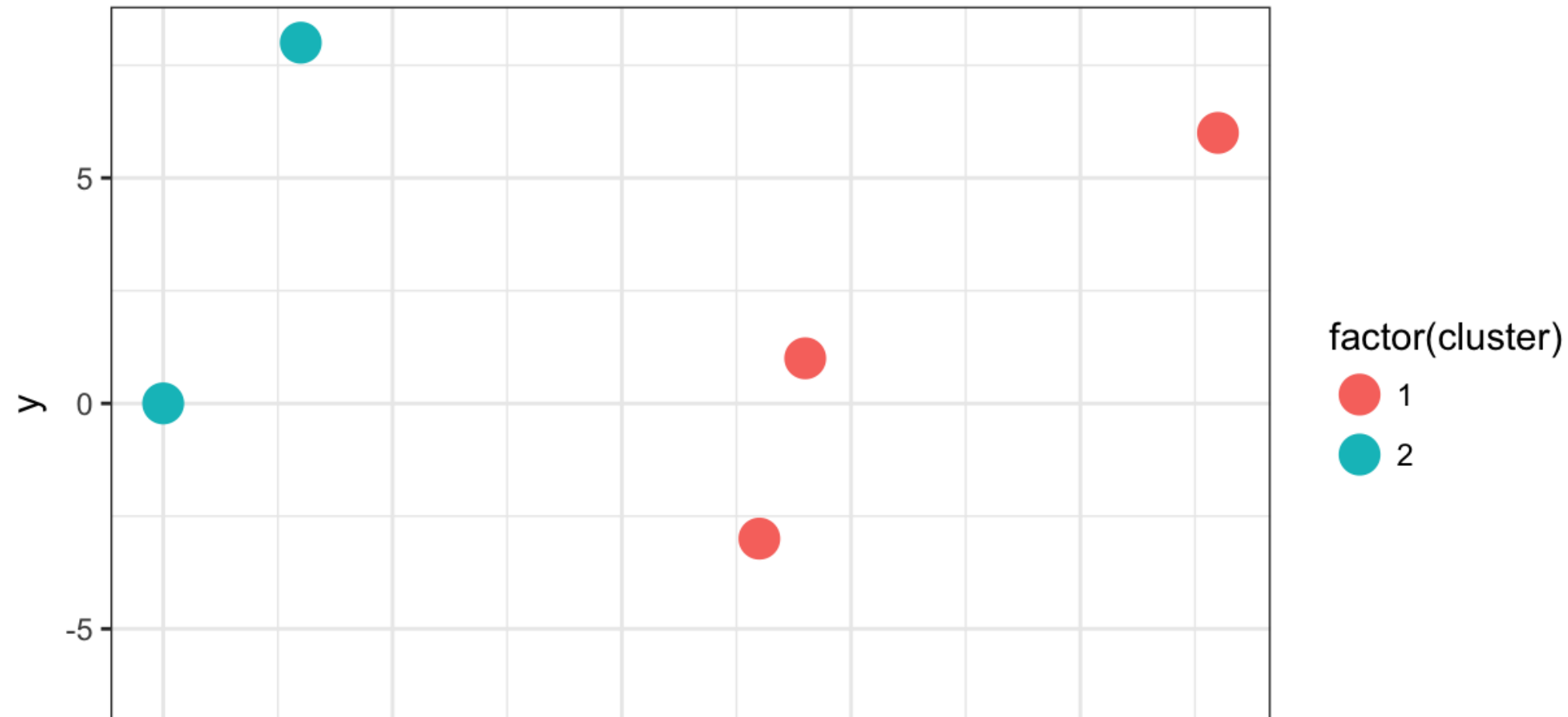
```
print(players_clustered)
```

	x	y	cluster
	<dbl>	<dbl>	<int>
1	-1	1	1
2	-2	-3	1
3	8	6	1
4	7	-8	1
5	-12	8	2
6	-15	0	2

Visualizing K Clusters

```
library(ggplot2)
```

```
ggplot(players_clustered, aes(x = x, y = y, color = factor(cluster))) +  
  geom_point()
```



Let's practice!
CLUSTER ANALYSIS IN R

Visualizing the dendrogram

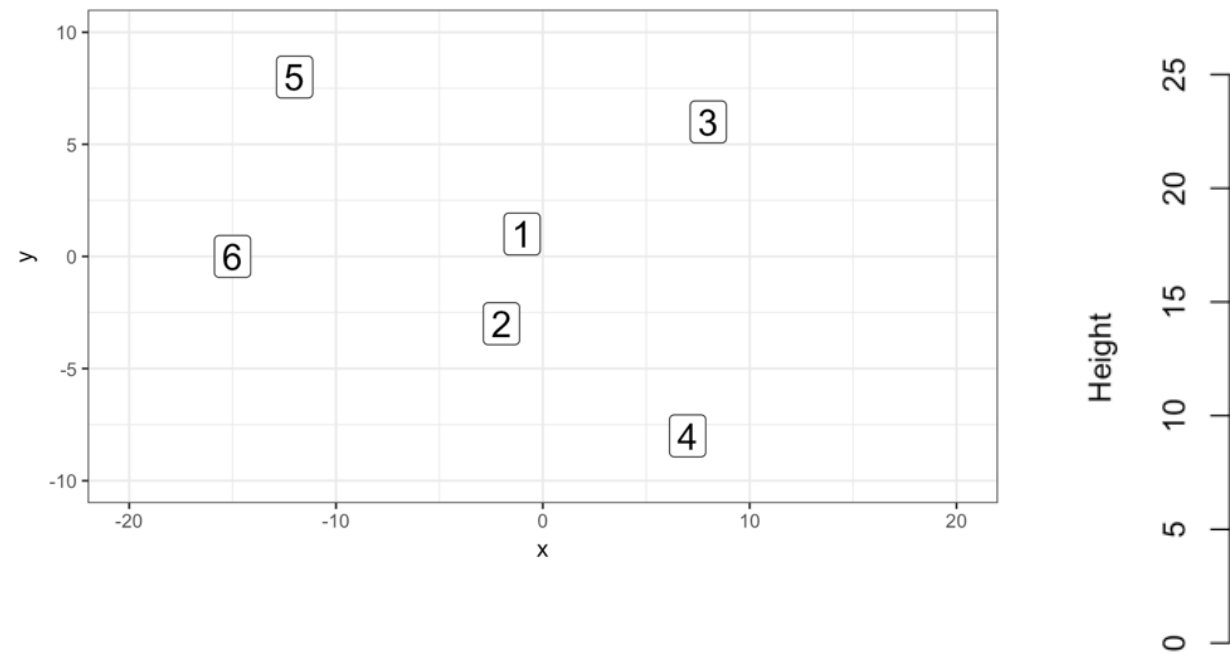
CLUSTER ANALYSIS IN R



Dmitriy Gorenshteyn

Lead Data Scientist, Memorial Sloan
Kettering Cancer Center

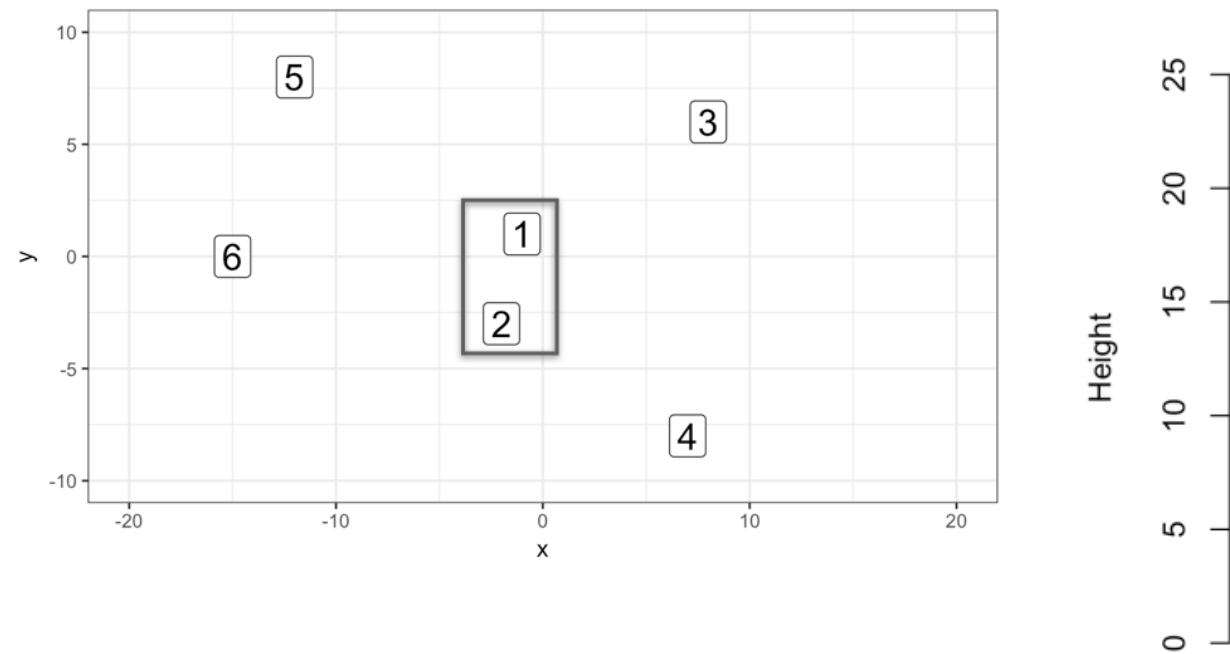
Building the dendrogram



Cluster Dendrogram

```
hclust (*, "complete")
```

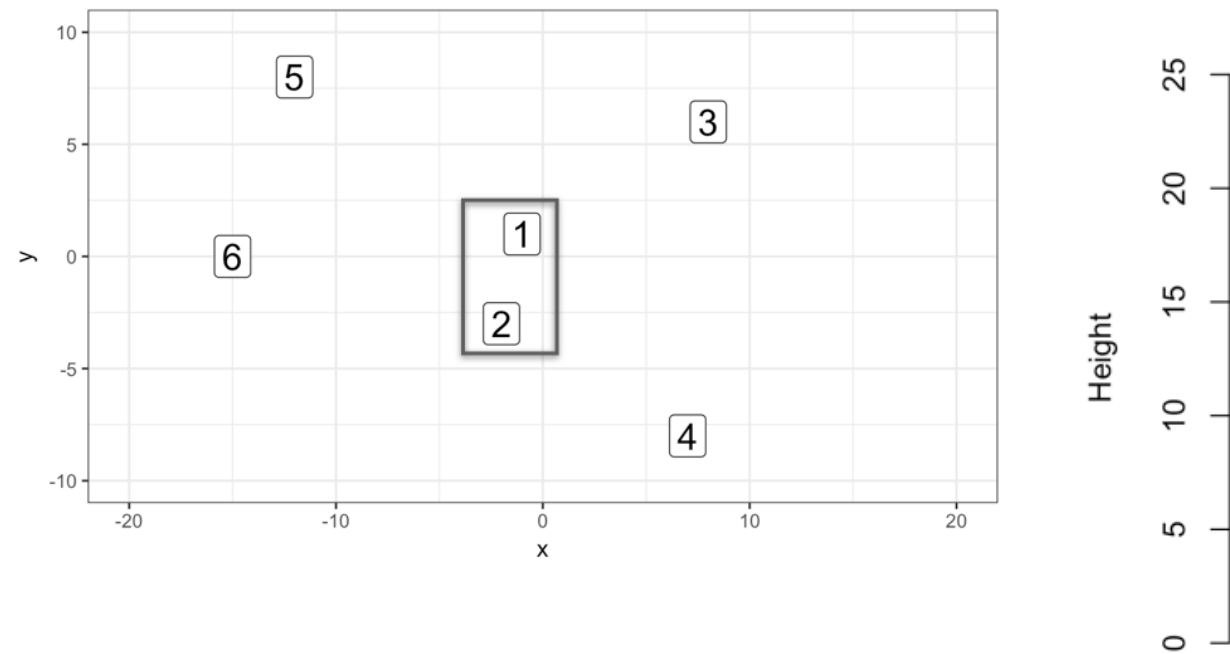
Building the dendrogram



Cluster Dendrogram

```
hclust (*, "complete")
```

Building the dendrogram

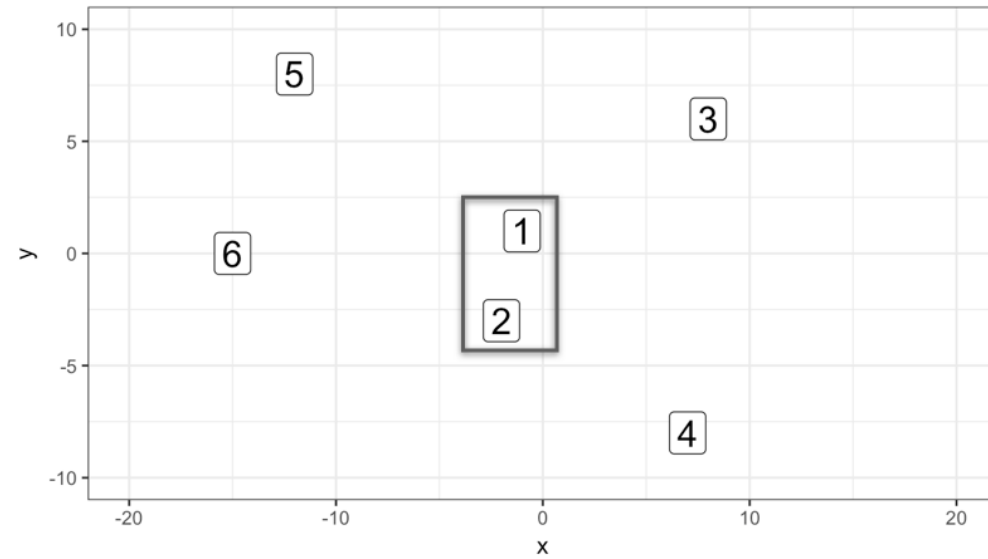


Cluster Dendrogram

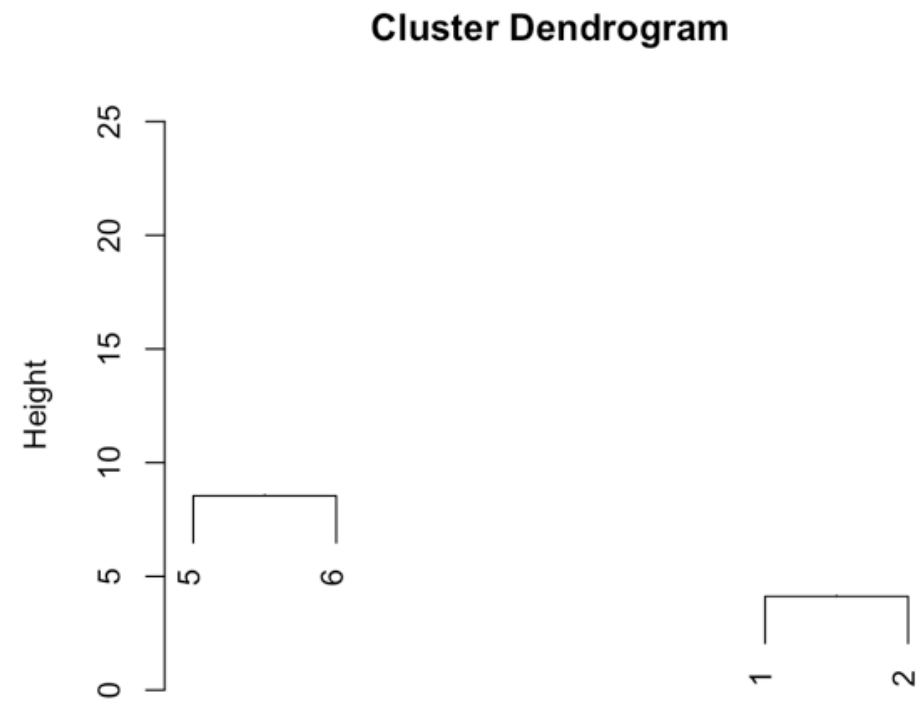
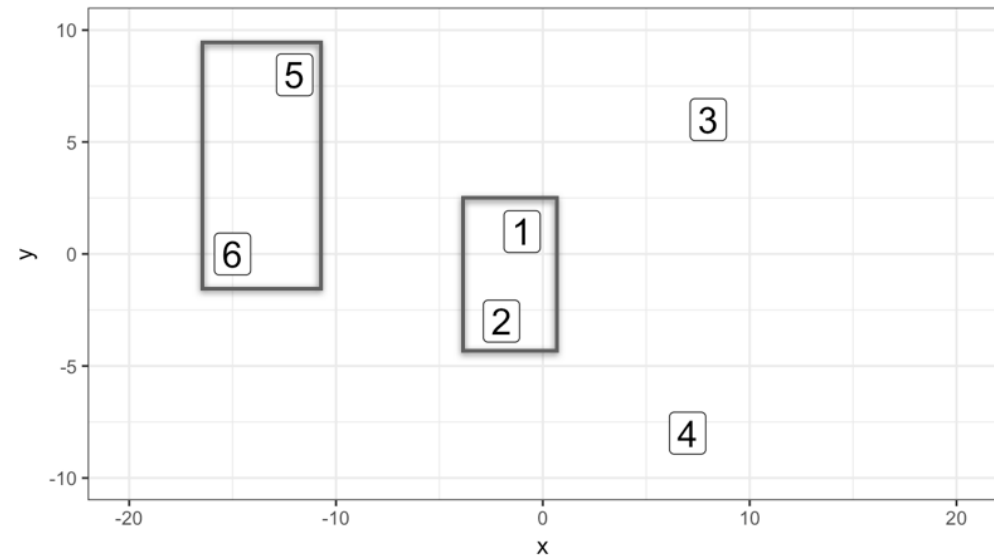


```
hclust (*, "complete")
```

Building the dendrogram

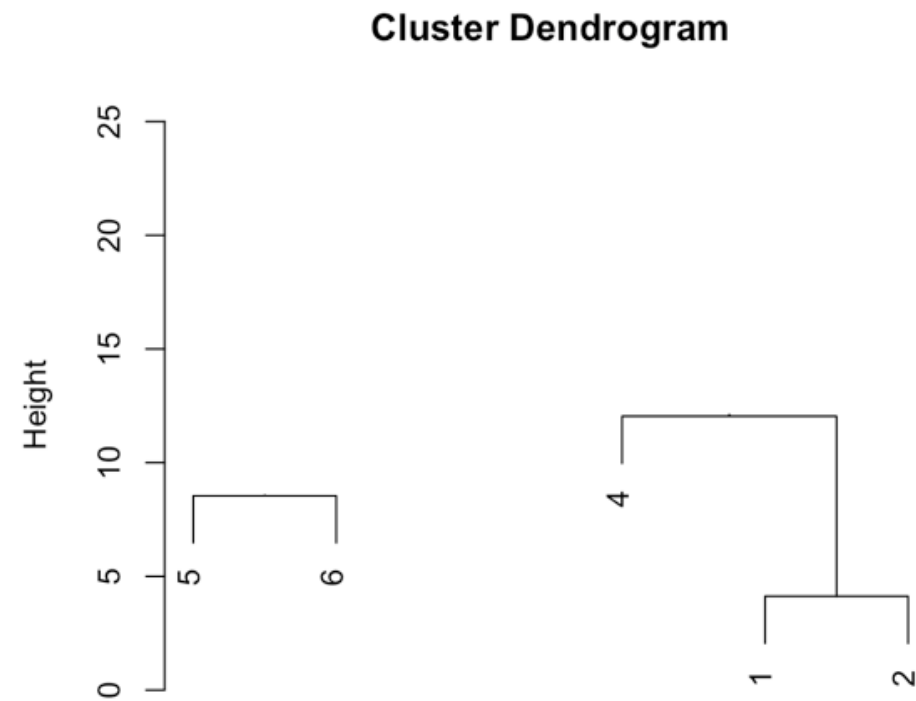
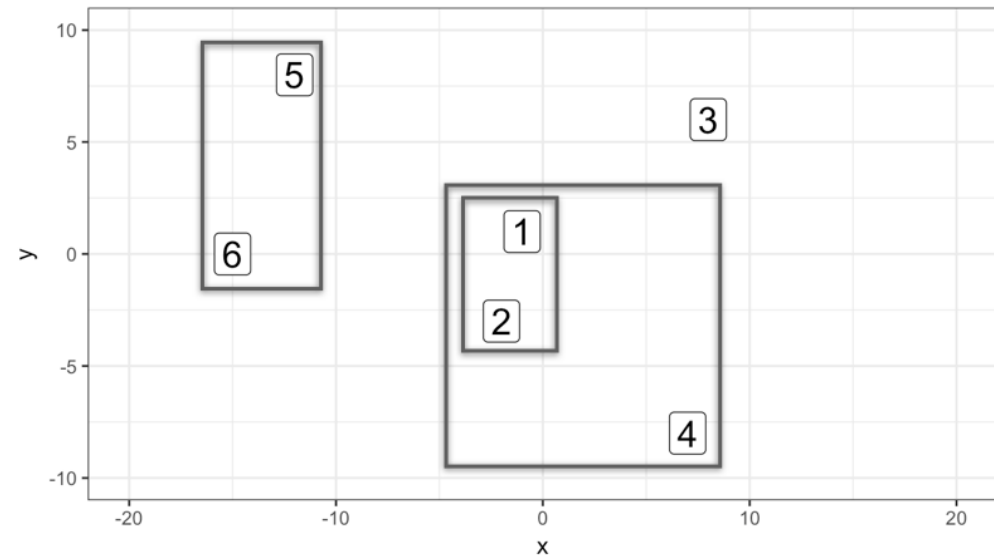


Building the dendrogram



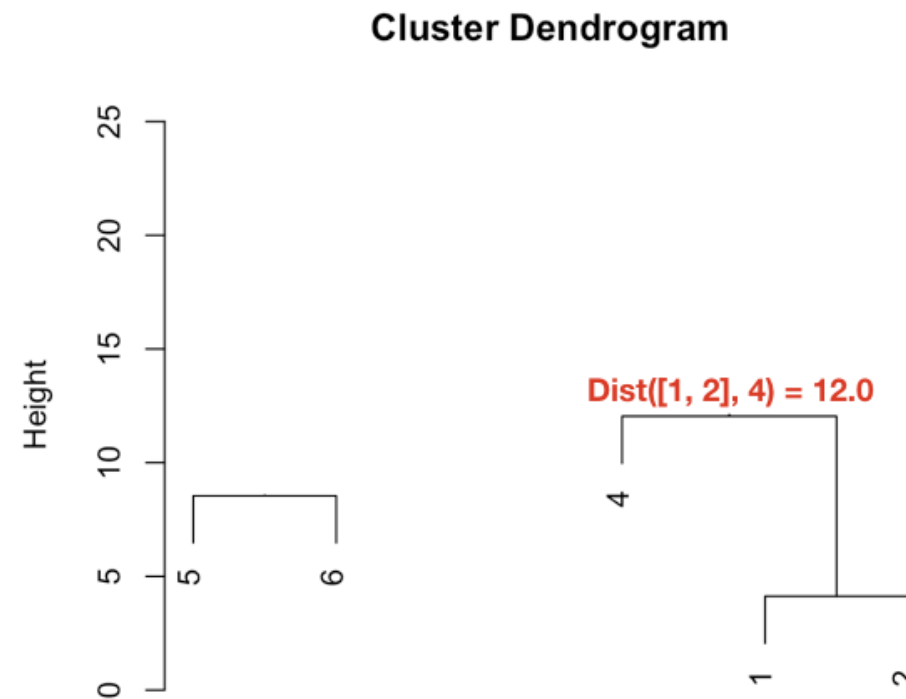
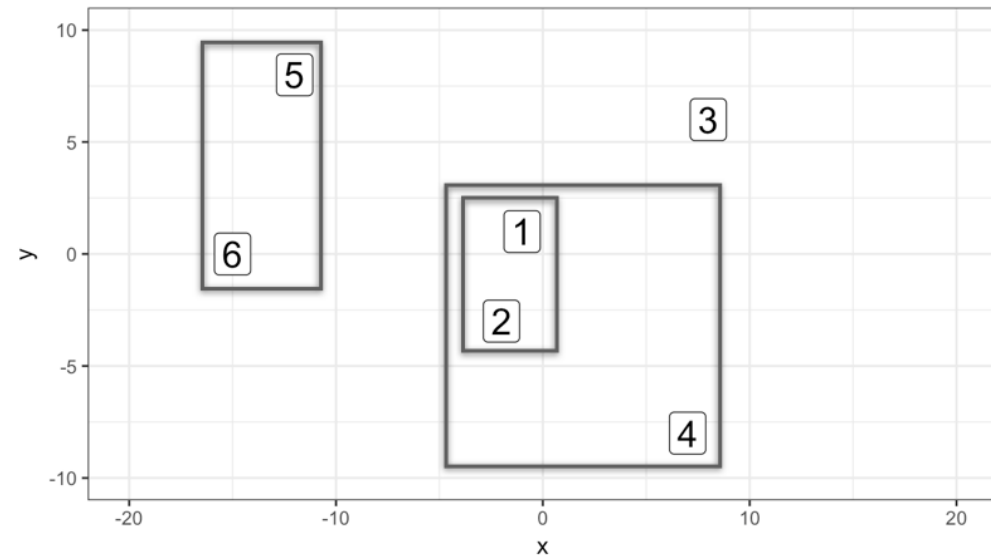
`hclust (*, "complete")`

Building the dendrogram



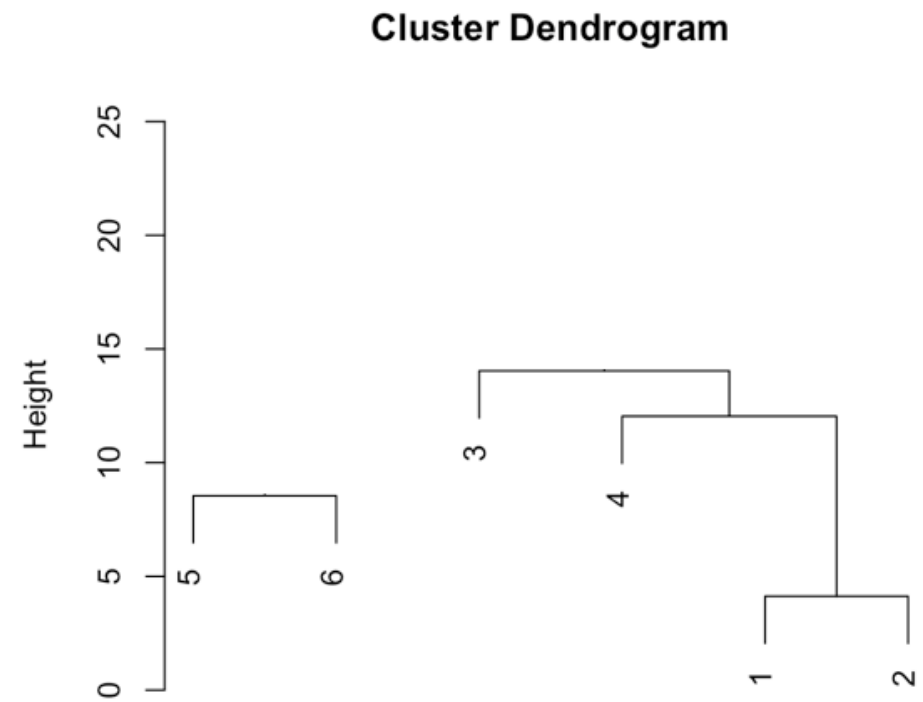
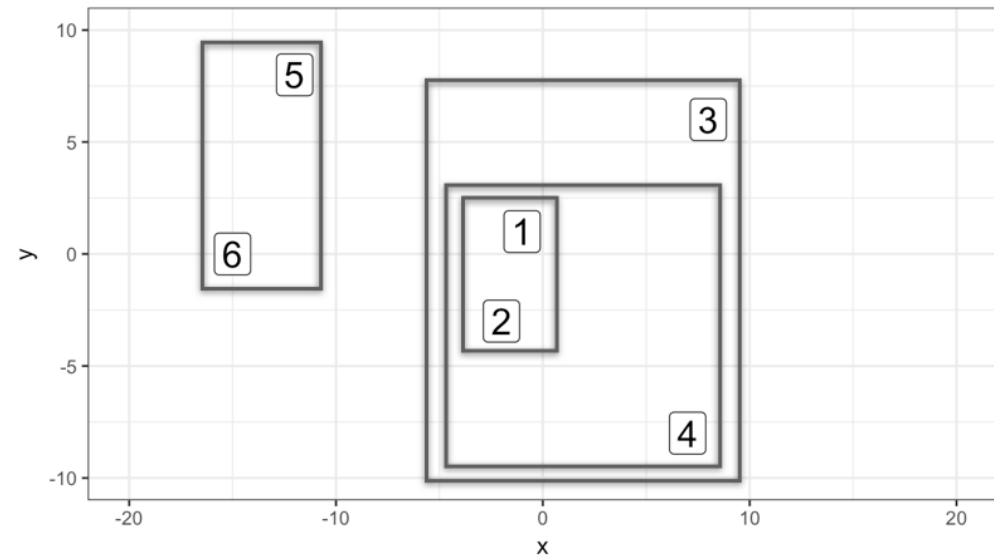
```
hclust (*, "complete")
```

Building the dendrogram



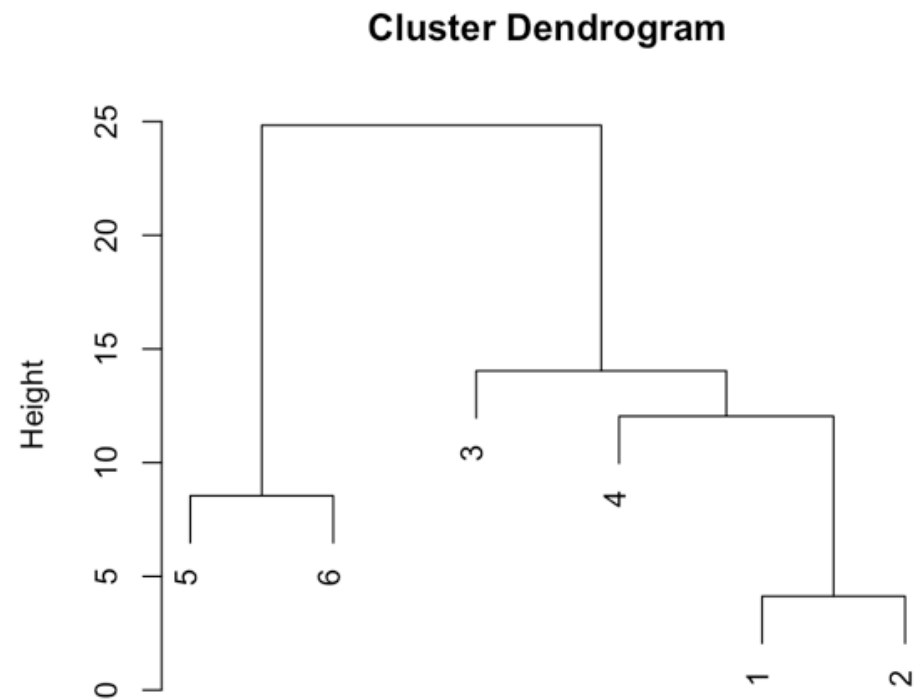
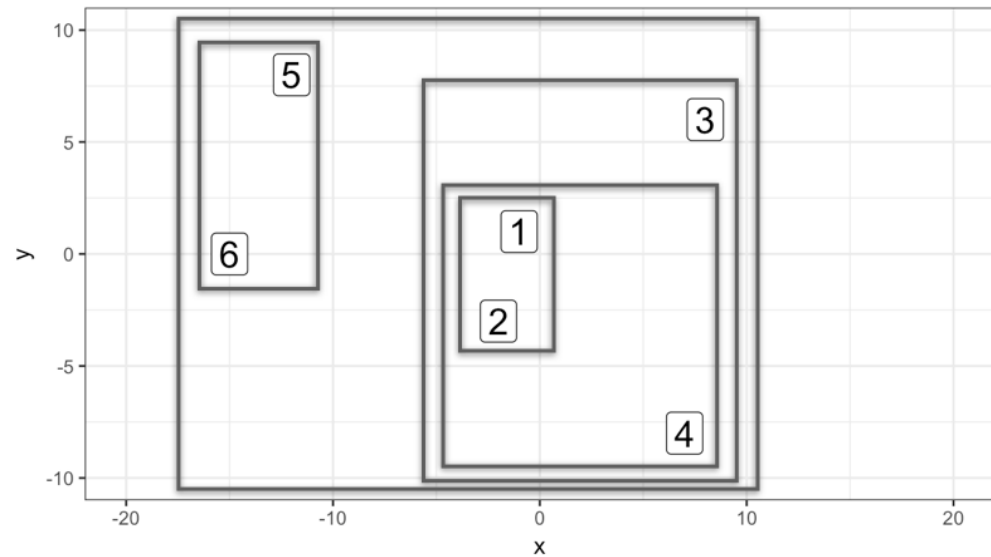
```
hclust (*, "complete")
```


Building the dendrogram



```
hclust (*, "complete")
```

Building the dendrogram

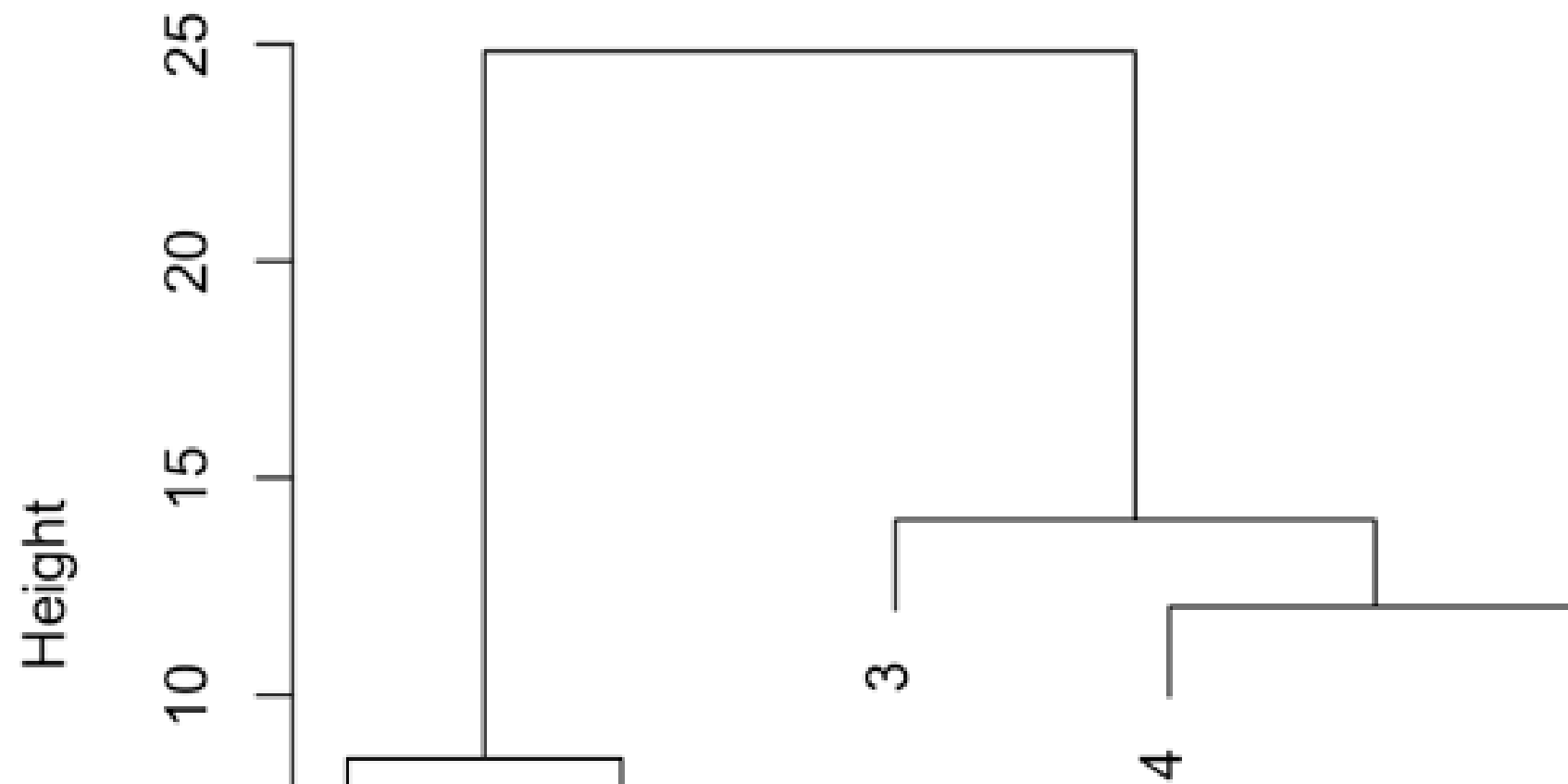


`hclust (*, "complete")`

Plotting the dendrogram

```
plot(hc_players)
```

Cluster Dendrogram



Let's practice!
CLUSTER ANALYSIS IN R

Cutting the tree

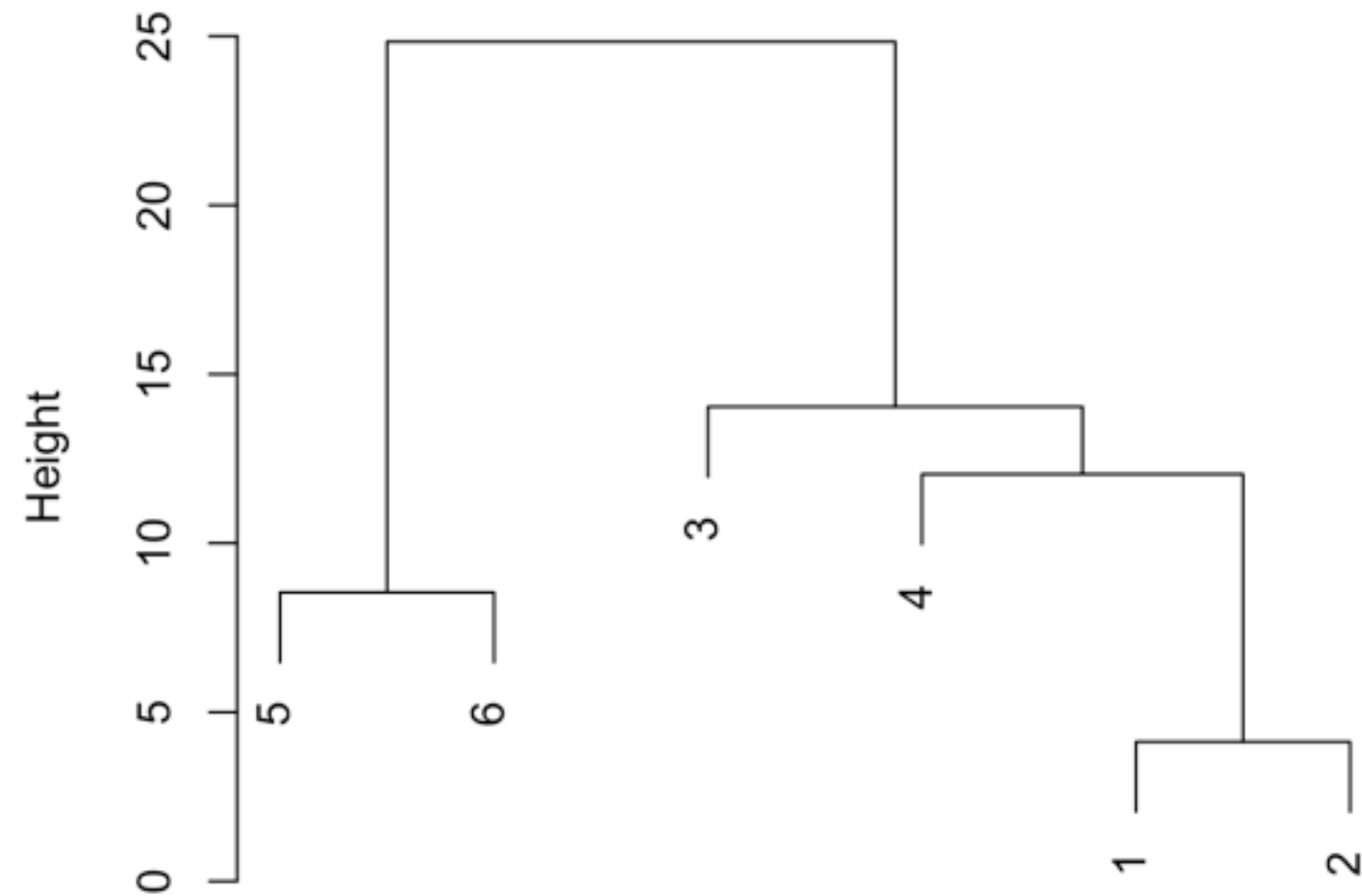
CLUSTER ANALYSIS IN R



Dmitriy Gorenshteyn

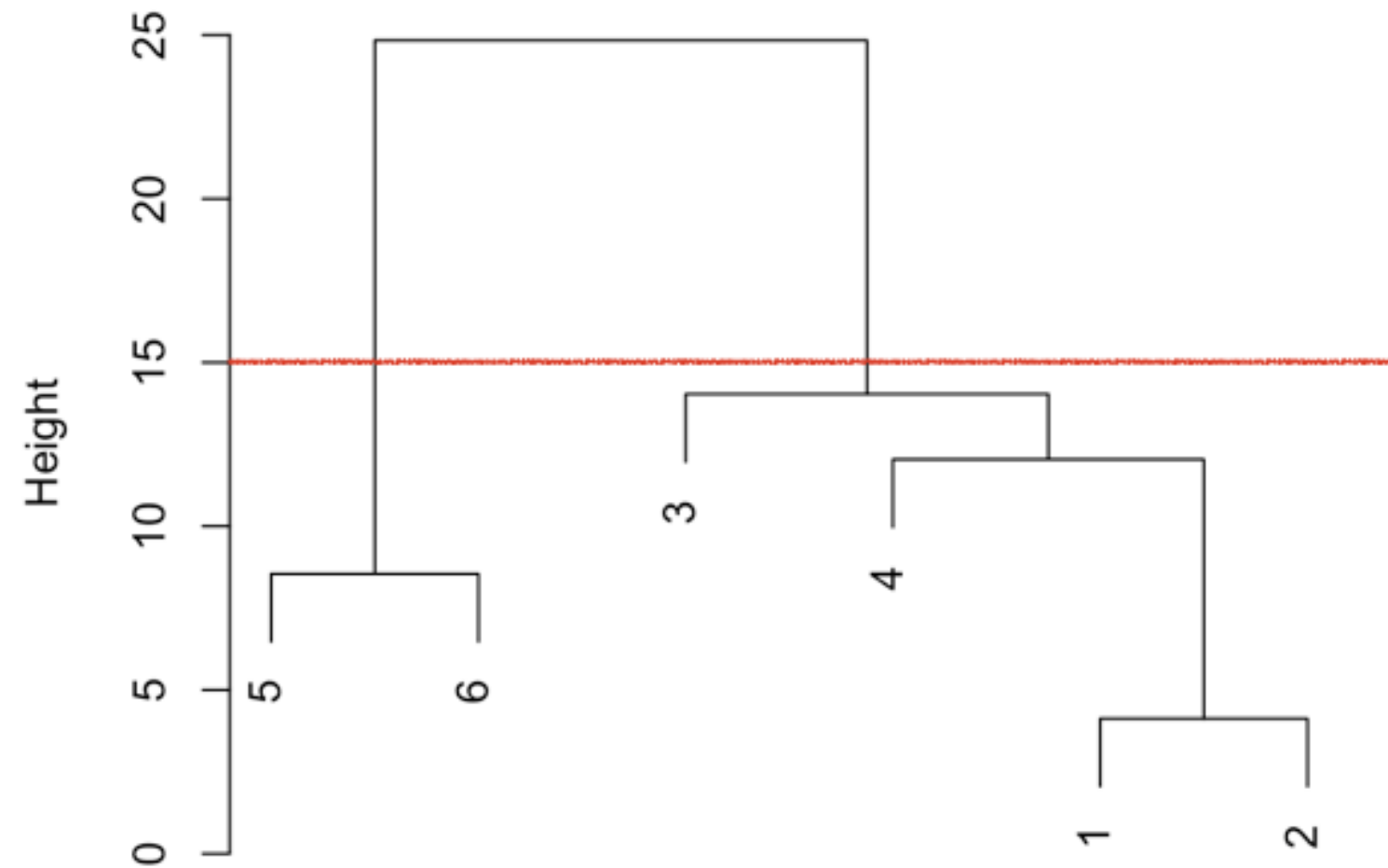
Lead Data Scientist, Memorial Sloan
Kettering Cancer Center

Cluster Dendrogram



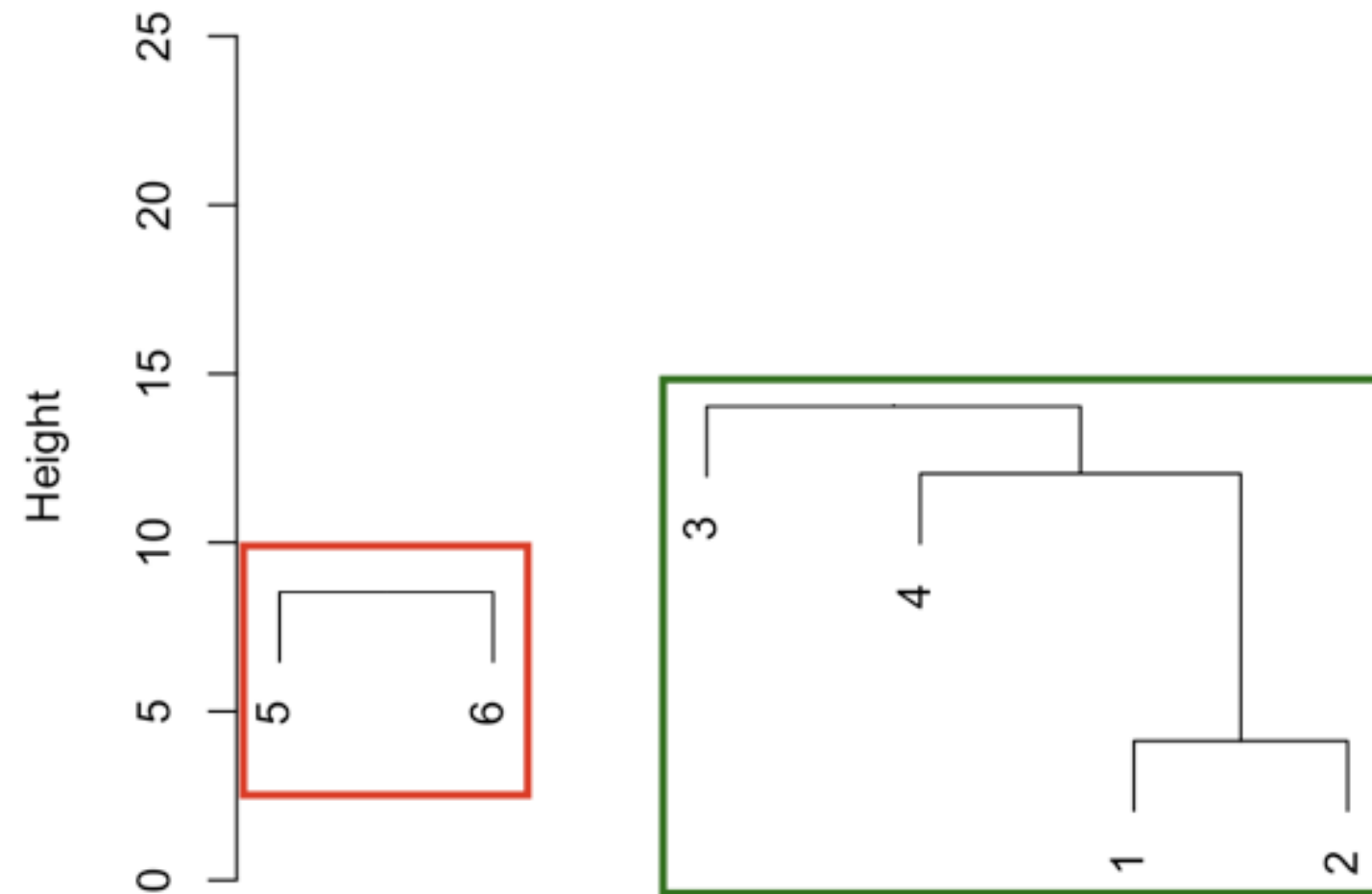
`hclust (*, "complete")`

Cluster Dendrogram



hclust (*, "complete")

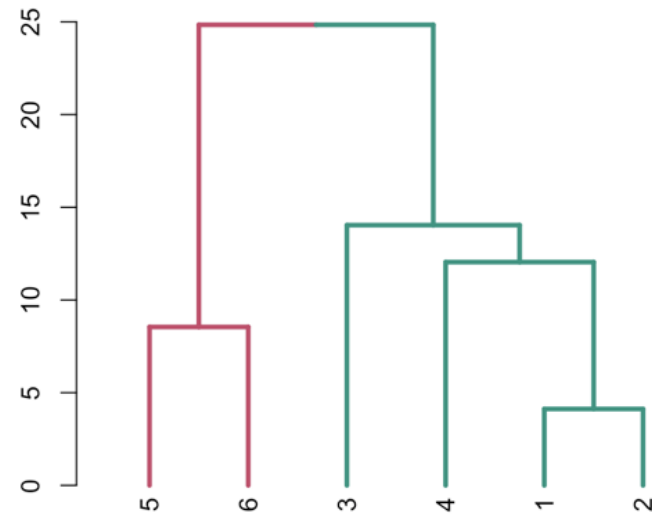
Cluster Dendrogram



`hclust (*, "complete")`

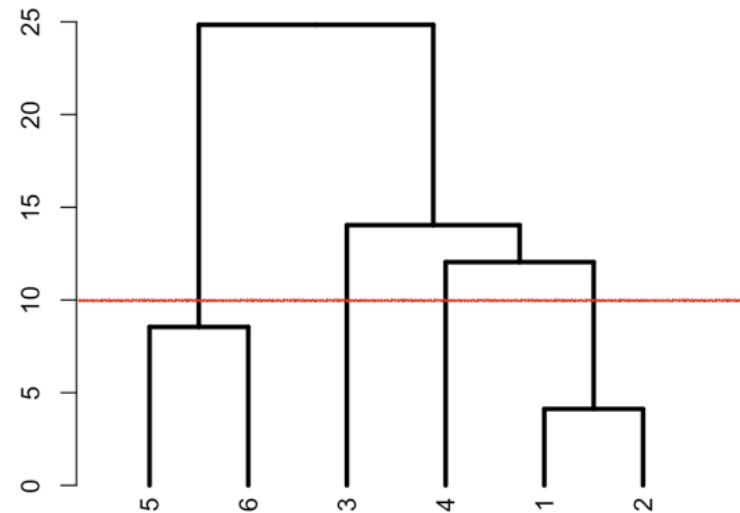
Coloring the dendrogram - height

```
library(dendextend)
dend_players <- as.dendrogram(hc_players)
dend_colored <- color_branches(dend_players, h = 15)
plot(dend_colored)
```



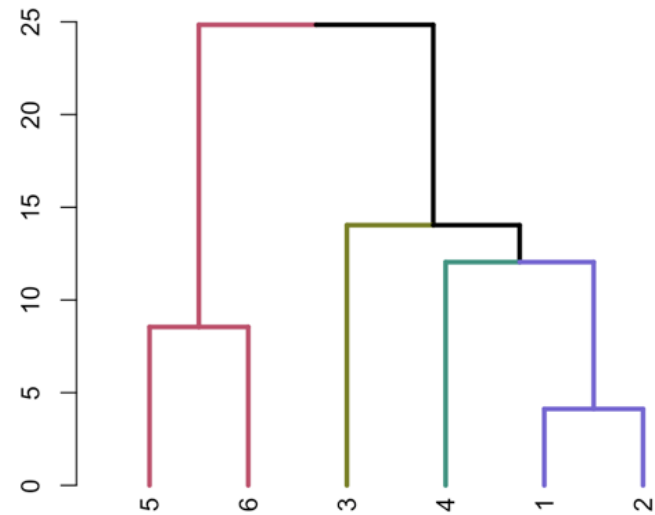
Coloring the dendrogram - height

```
library(dendextend)
dend_players <- as.dendrogram(hc_players)
dend_colored <- color_branches(dend_players, h = 15)
plot(dend_colored)
```



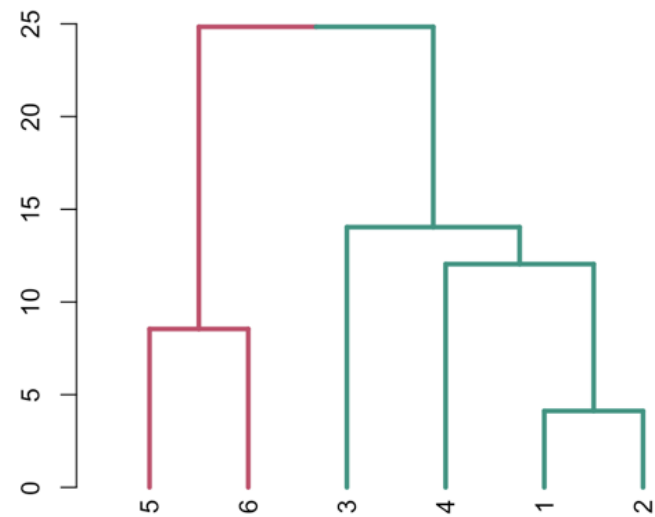
Coloring the dendrogram - height

```
library(dendextend)
dend_players <- as.dendrogram(hc_players)
dend_colored <- color_branches(dend_players, h = 10)
plot(dend_colored)
```



Coloring the dendrogram - K

```
library(dendextend)
dend_players <- as.dendrogram(hc_players)
dend_colored <- color_branches(dend_players, k = 2)
plot(dend_colored)
```



cutree() using height

```
cluster_assignments <- cutree(hc_players, h = 15)
print(cluster_assignments)
[1] 1 1 1 1 2 2
library(dplyr)
players_clustered <- mutate(players, cluster = cluster_assignments)
```

```
print(players_clustered)
```

	x	y	cluster
	<dbl>	<dbl>	<int>
1	-1	1	1
2	-2	-3	1
3	8	6	1
4	7	-8	1
5	-12	8	2
6	-15	0	2

Let's practice!
CLUSTER ANALYSIS IN R

Making sense of the clusters

CLUSTER ANALYSIS IN R



Dmitriy Gorenshteyn

Lead Data Scientist, Memorial Sloan
Kettering Cancer Center

Wholesale dataset

- 45 observations
- 3 features:
 - Milk Spending
 - Grocery Spending
 - Frozen Food Spending

Wholesale dataset

```
print(customers_spend)
      Milk Grocery Frozen
1  11103   12469    902
2   2013    6550    909
3   1897    5234    417
4   1304    3643   3045
5   3199    6986   1455
...     ...     ...     ...
```

Exploring more than 2 dimensions

- Plot 2 dimensions at a time
- Visualize using PCA
- Summary statistics by feature

Segment the customers

CLUSTER ANALYSIS IN R