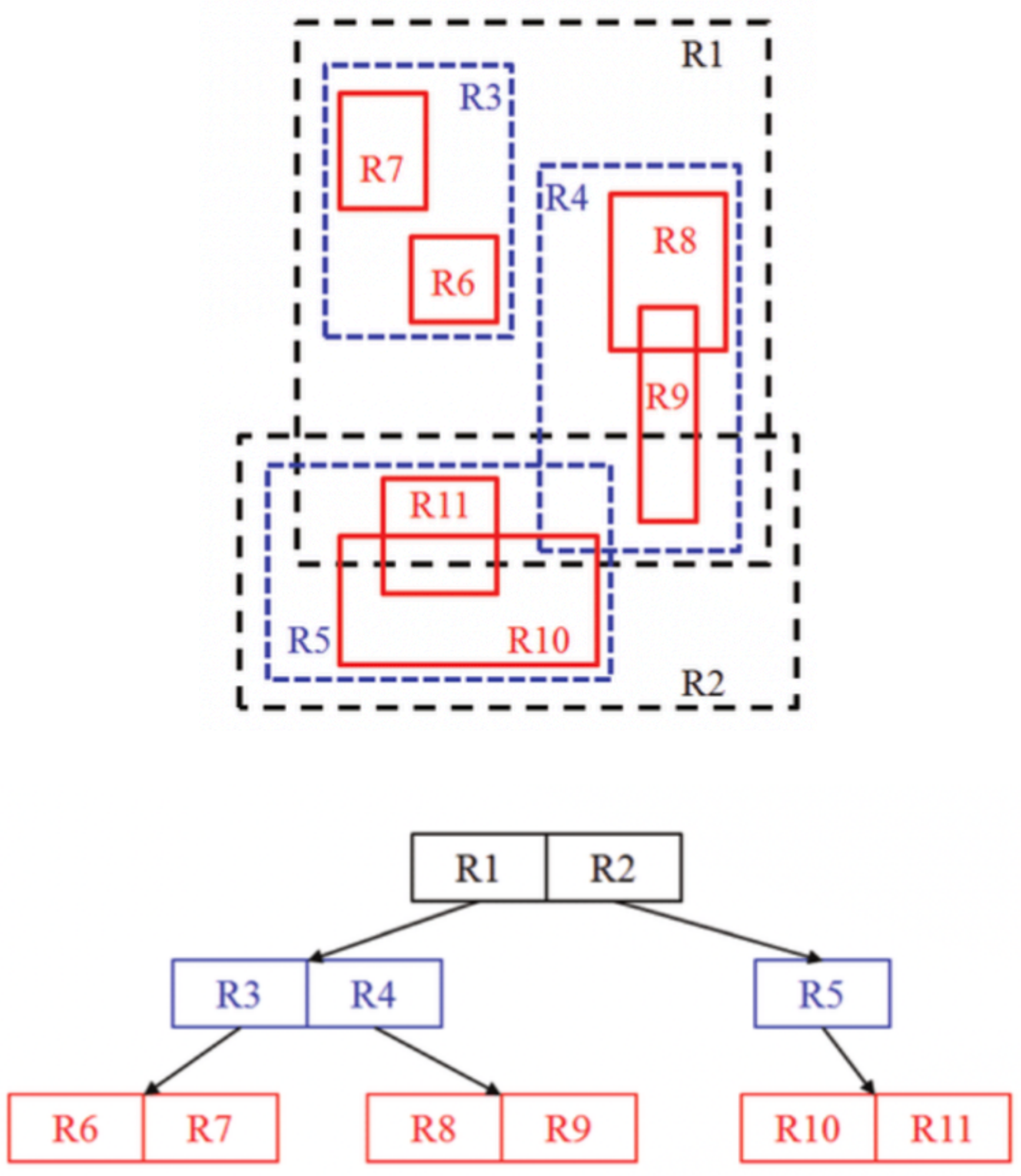
R-Trees

NDBI007: Practical class 6

R-Tree

- Height-*balanced* tree
  - Extension of B+-tree for *spatial data*
- Nodes correspond to disk pages
- Inner node contains n-dimensional bounding box I
  - MBRs (*minimum bounding rectangle*)
  - MBR of a node in MBR of all children
- Leaf level contains pointers to the spatial objects



Splitting In R-Tree: Guttman

- First, we *identify a pair of elements* which would result in the *largest dead space*
  - I.e., we apply method *PickSeeds*
- Next, remaining elements are added one by one
  - If remaining entries need to be assigned into node in order to have the *minimum number of entries*, then assign them
  - Otherwise, Pick the one that would make the *biggest difference in area enlargement* when put to one of two groups (method *PickNext*)
    - Add in to the one with the least difference

SplitNode(P, PP, E)
Input: node P, new node PP, m original entries, new entry E
Output: modified P, PP

PickSeeds(); {chooses first E_i and E_j for P and PP }
WHILE not assigned entry exists DO
  IF remaining entries need to be assigned to P or PP in order to have the minimum number of entries m
    THEN assign them;
  ELSE
    $E_i \longleftarrow$ **PickNext()** {choose where to assign next entry}
    Add E_i into group that will have to be enlarged least to accommodate it. Resolve ties by adding the entry to the group with smaller area, then to the one with fewer entries;

PickSeeds()
FOREACH E_i, E_j $(i \neq j)$ DO
  $d_{ij} \longleftarrow$ area(J) - area(E_i.) - area (E_J.I);
    { J is the MBR covering E_i and E_j }
  pick E_i and E_j with maximal d_{ij};

PickNext()
FOREACH remaining E_i DO
  $d_1 \longleftarrow$ area increase required for MBR of P and E_i.I;
  $d_2 \longleftarrow$ area increase required for MBR of PP and E_i.I;
pick E_i with maximal $d_1 - d_2$;

Example 6.1: Guttman's Split

- Split the following overflown node with Guttman's splitnode method
    - The maximum number of items in a node is $M = 8$
    - The minimum number of items in a node is $m = 3$

A	A		F	F			D
	A			B	B		
			B	B	B		
E	E	E					
E	E	E					
			C	C	C	G	
	H					G	
	H					I	I

Example 6.1: Guttman's Split (Continued)

- We apply Guttman's PickSeeds method to find two elements having the largest dead space if being placed together

Pair	Overall area	Area of the objects	Dead space
AB	18	8	10
AD	16	4	12
…			
AI	64	5	59
…			
DH	56	3	53
…			
HI	14	4	10

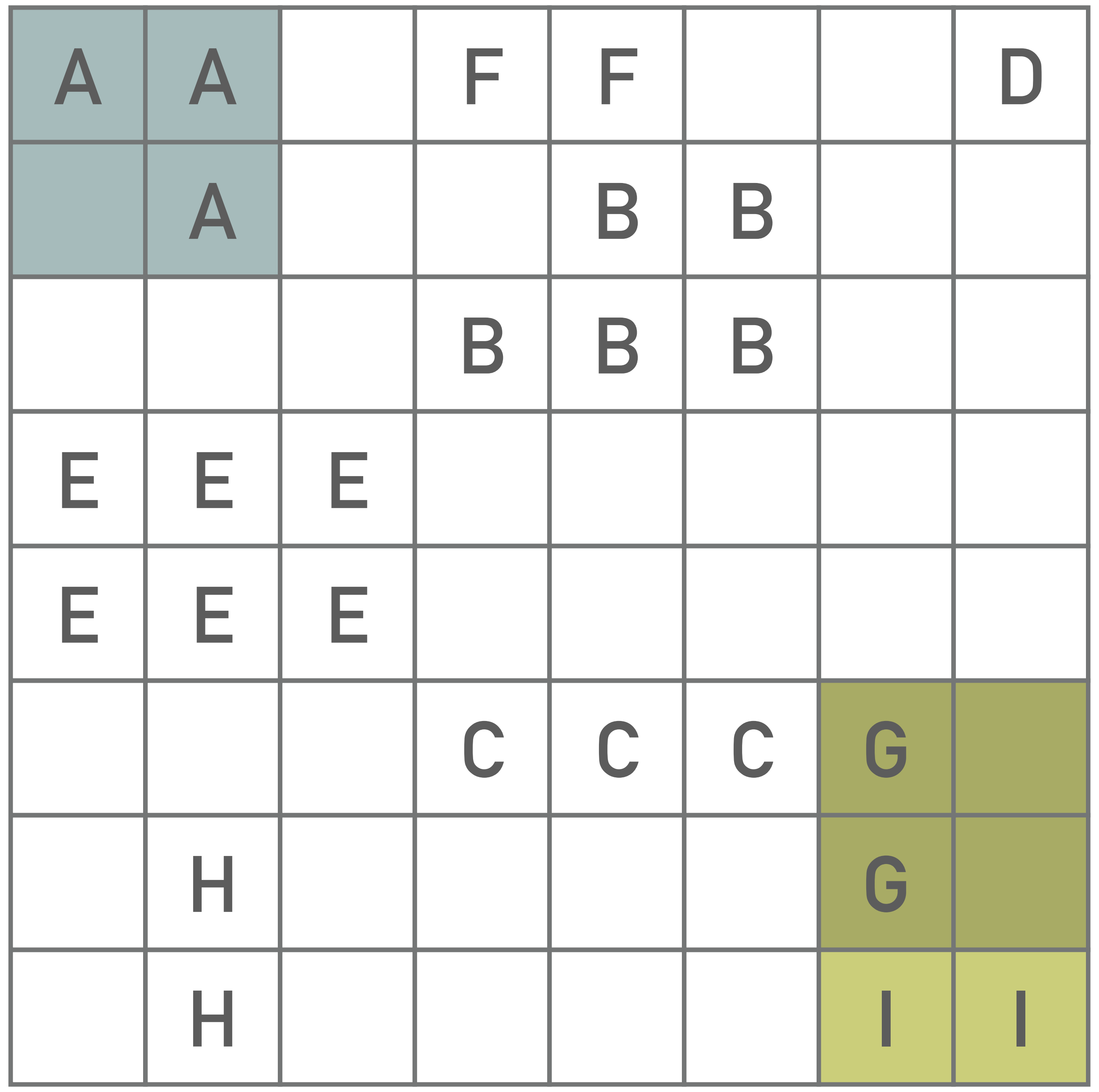
A	A		F	F			D
	A			B	B		
			B	B	B		
E	E	E					
E	E	E					
			C	C	C	G	
	H					G	
	H					I	I

- The largest dead space has AI thus those will be the seeds of the splitting method

Example 6.1: Guttman's Split (Continued)

- Next, iteratively add such an object into a node which will maximize the difference in the *node area enlargements* if the object was inserted into the first or second node

Object	A	I	Difference
B	6x3-4=14	5x7-2=33	|14-33|=19
C	6x6-4=32	5x3-2=13	|32-13|=19
D	8x2-4=14	2x8-2=16	|14-16|=2
E	3x5-4=11	8x5-2=38	|11-38|=27
F	5x2-4=6	5x8-2=38	|6-38|=32
G	7x7-4=45	2x3-2=4	|45-4|=41
H	2x8-4=14	7x2-2=12	|14-12|=2

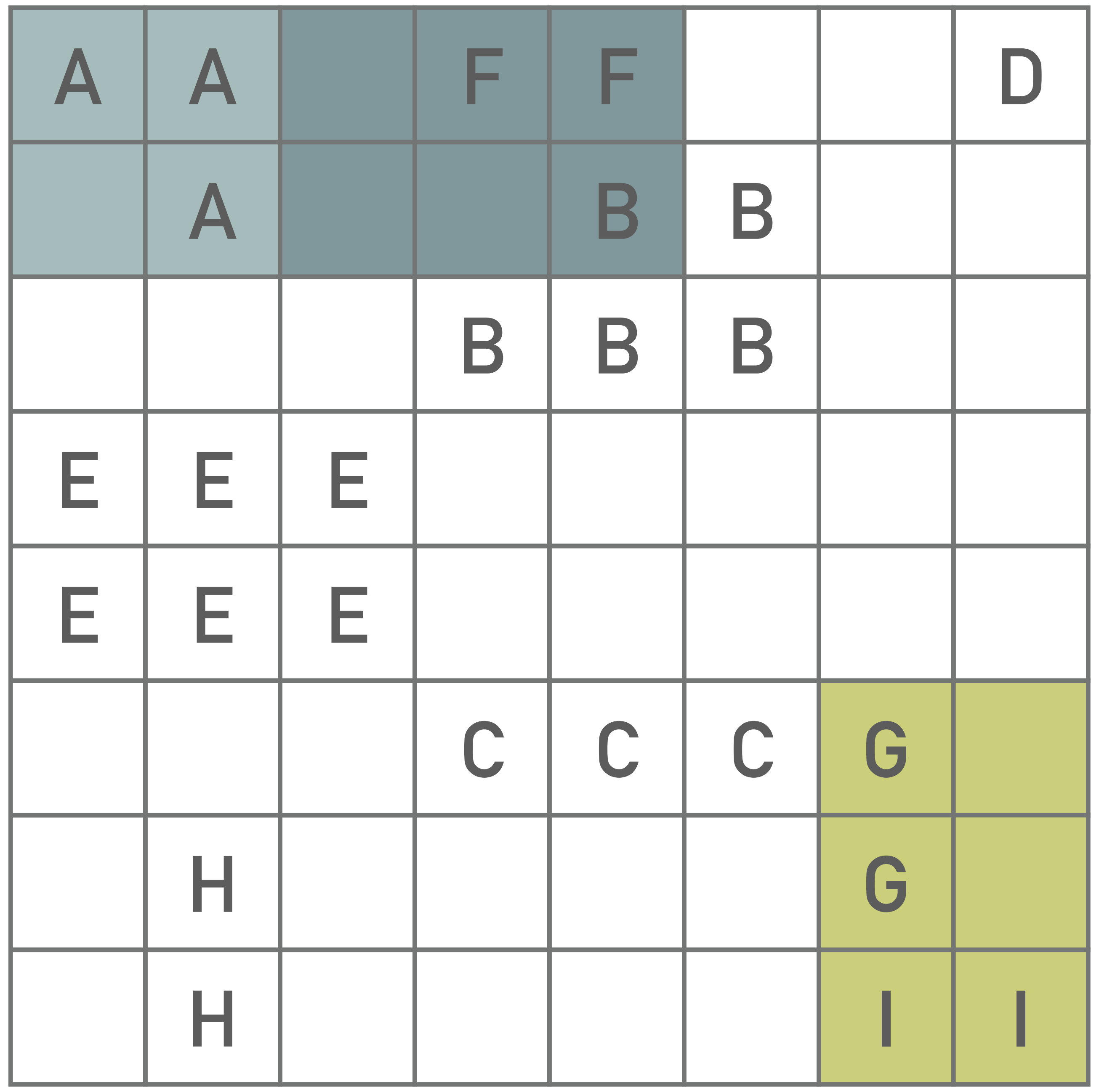


- The biggest difference shows the object G, hence it will be inserted into the node which is closer
- Thus, we have nodes A and GI

Example 6.1: Guttman's Split (Continued)

- Next, iteratively add such an object into a node which will maximize the difference in the node area enlargements if the object was inserted into the first or second node

Object	A	GI	Difference
B	6x3-4=14	5x7-6=29	|14-29|=15
C	6x6-4=32	5x3-6=9	|32-9|=23
D	8x2-4=14	2x8-6=10	|14-10|=4
E	3x5-4=11	8x5-6=34	|11-34|=23
F	5x2-4=6	5x8-6=34	|6-34|=28
H	2x8-4=14	7x3-6=15	|14-15|=1

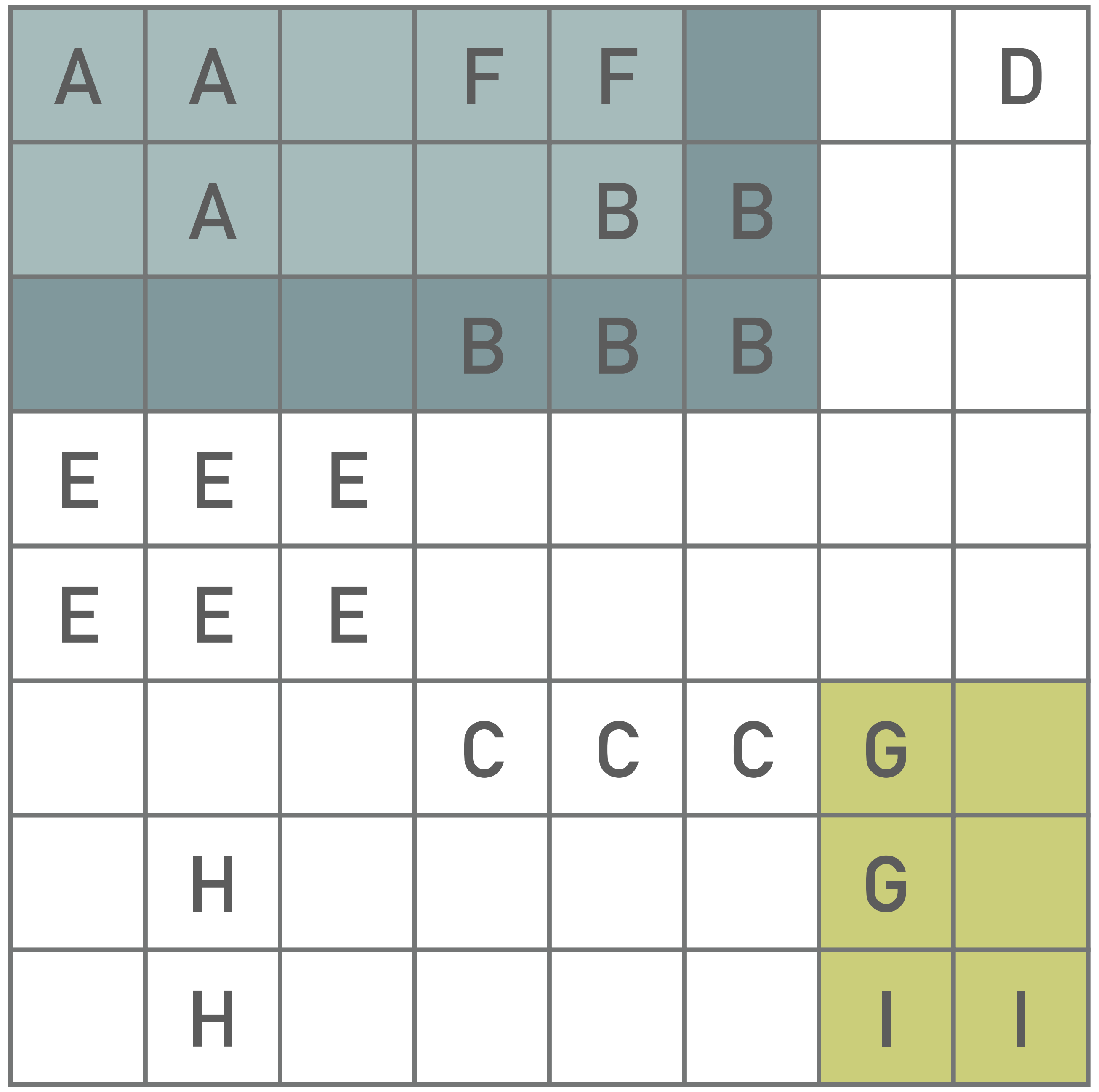


- The biggest difference shows the object F, hence it will be inserted into the node which is closer, i.e., A
- Thus, we have nodes AF and GI

Example 6.1: Guttman's Split (Continued)

- Next, iteratively add such an object into a node which will maximize the difference in the node area enlargements if the object was inserted into the first or second node

Object	AF	GI	Difference
B	6x3-10=8	5x7-6=29	|8-29|=21
C	6x6-10=26	5x3-6=9	|26-9|=17
D	8x2-10=6	2x8-6=10	|6-10|=4
E	5x5-10=15	8x5-6=34	|15-34|=19
H	5x8-10=30	7x3-6=15	|30-15|=15

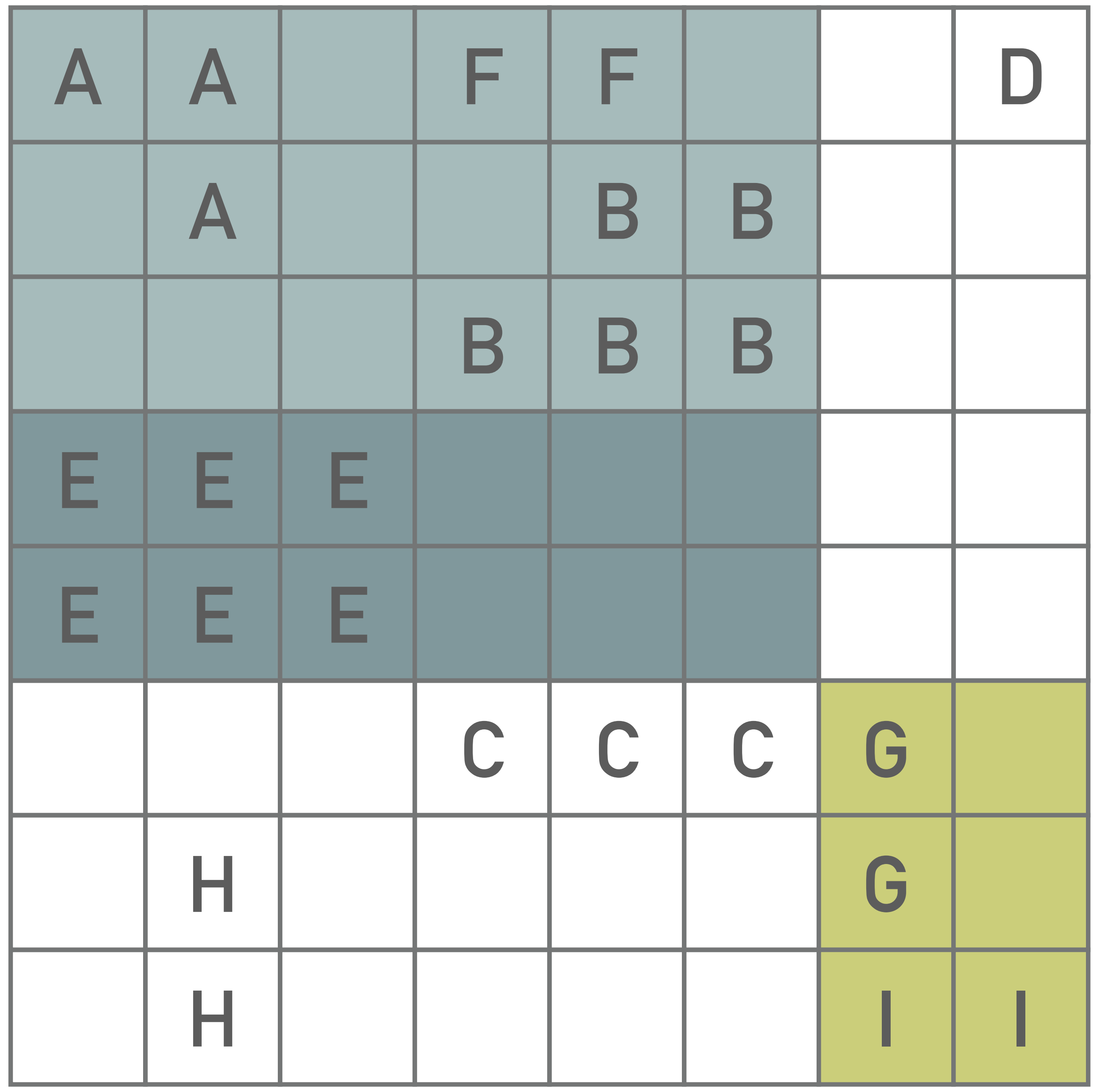


- The biggest difference shows the object B, hence it will be inserted into the node which is closer, i.e., AF
- Thus, we have nodes ABF and GI

Example 6.1: Guttman's Split (Continued)

- Next, iteratively add such an object into a node which will maximize the difference in the node area enlargements if the object was inserted into the first or second node

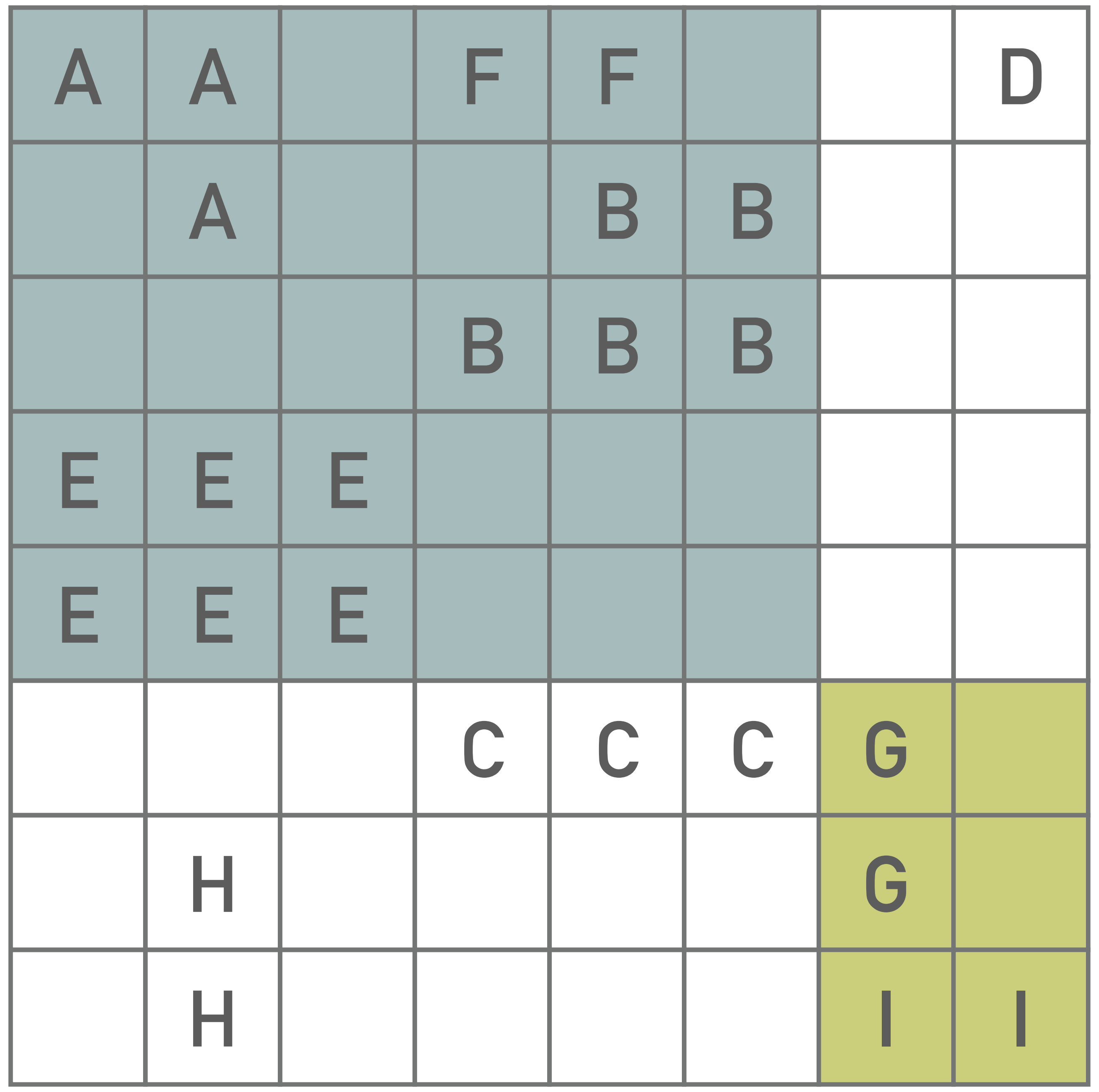
Object	ABF	GI	Difference
C	6x6-18=18	5x3-6=9	|18-9|=9
D	8x3-18=6	2x8-6=10	|6-10|=4
E	6x5-18=12	8x5-6=34	|12-34|=22
H	6x8-18=30	7x3-6=15	|30-15|=15



- The biggest difference shows the object E, hence it will be inserted into the node which is closer, i.e., ABF
- Thus, we have nodes ABEF and GI

Exercise 6.2

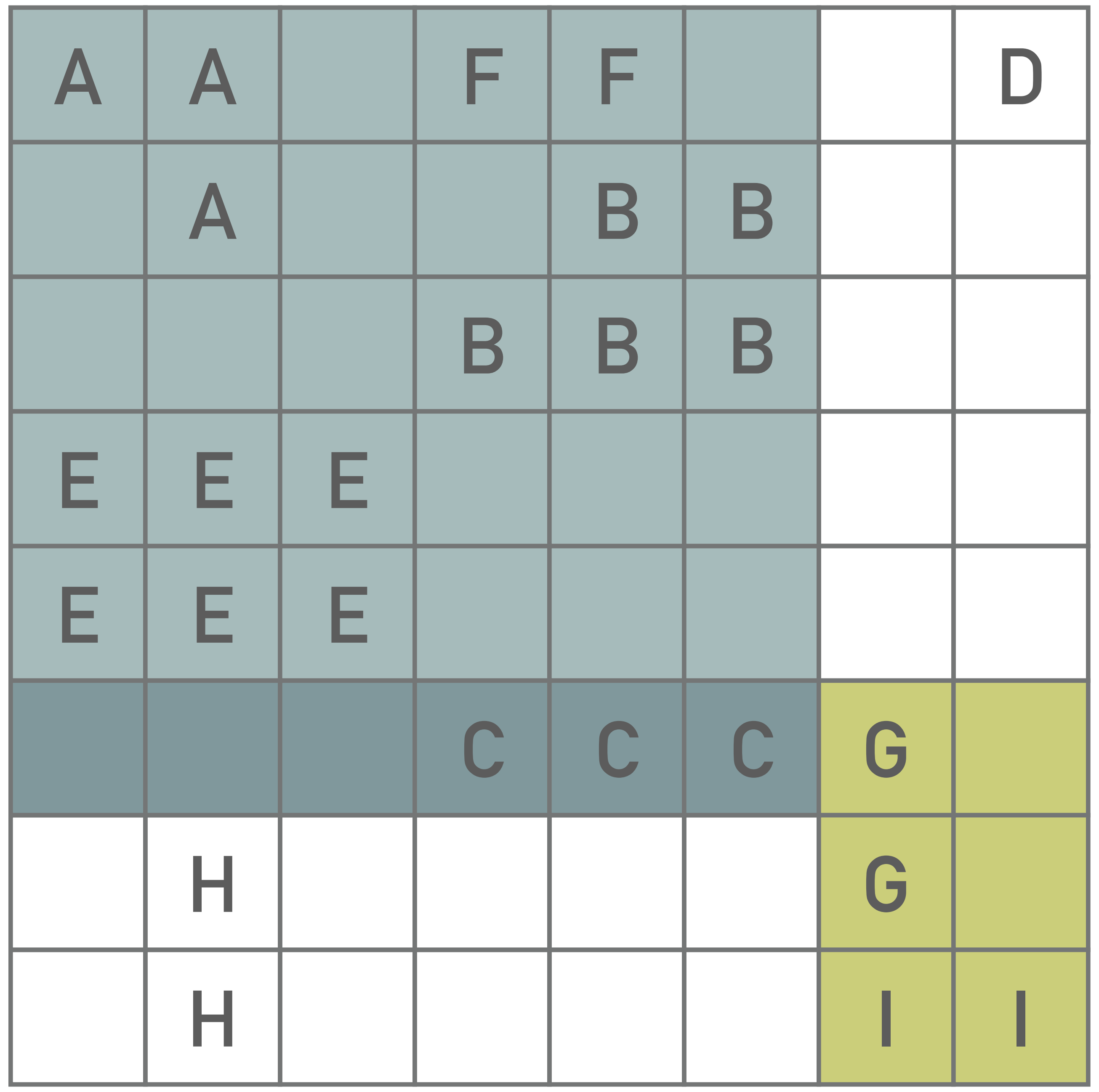
- Finish splitting of the overflown node
  - Continue with Guttman's method
  - The maximum number of items in a node is $M = 8$
  - The minimum number of items in a node is $m = 3$
- If there are more options to choose, explain the reason of yours choice



Exercise 6.2 (Solution)

- Next, iteratively add such an object into a node which will maximize the difference in the node area enlargements if the object was inserted into the first or second node

Object	ABEF	GI	Difference
C	6x6-30=6	5x3-6=9	|6-9|=3
D	8x5-30=10	2x8-6=10	|10-10|=0
H	6x8-30=18	7x3-6=15	|18-15|=3

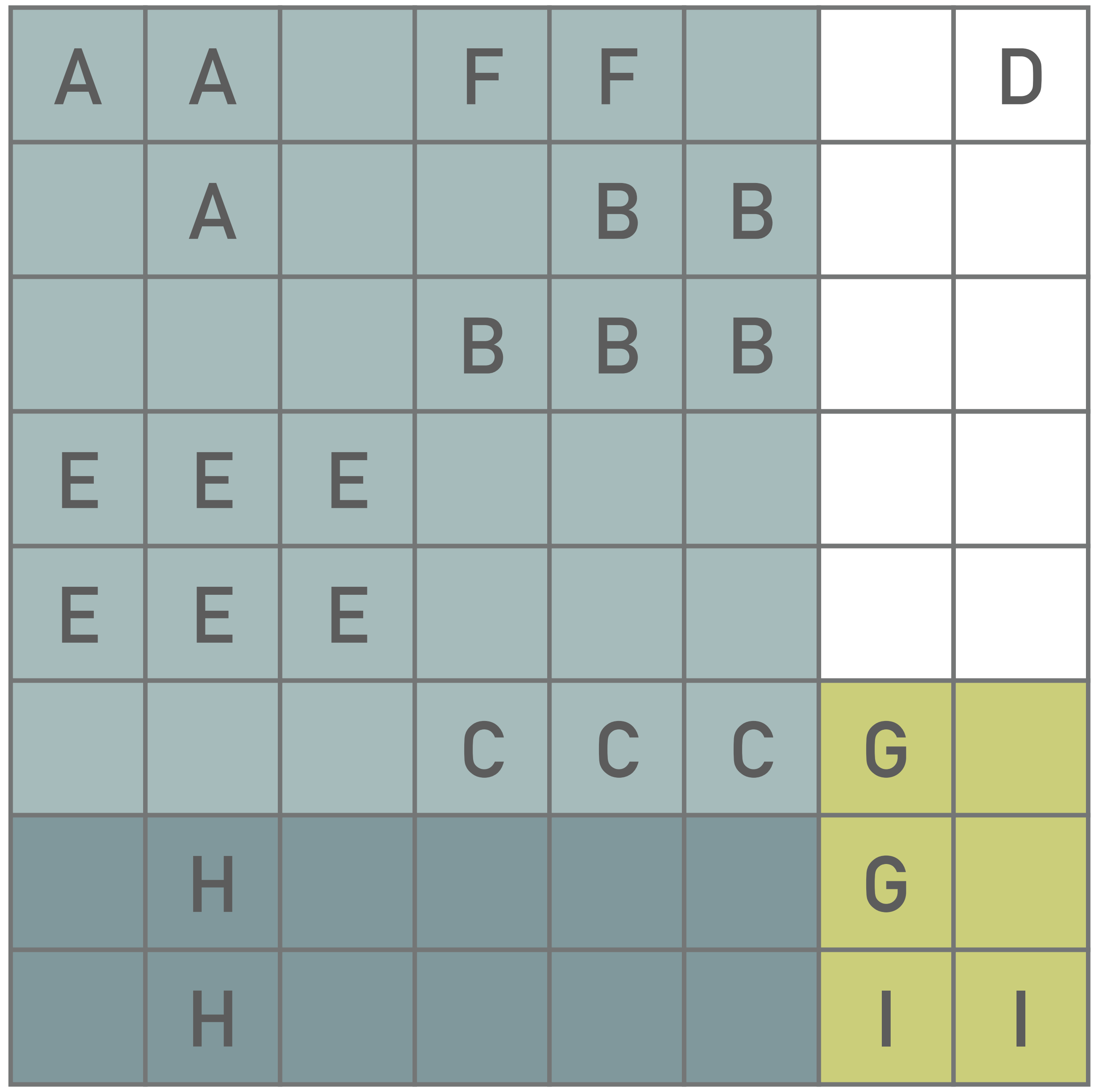


- The biggest difference shows the object C, hence it will be inserted into the node which is closer, i.e., ABEF
- Thus, we have nodes ABCEF and GI

Exercise 6.2 (Solution Continued)

- Next, iteratively add such an object into a node which will maximize the difference in the node area enlargements if the object was inserted into the first or second node

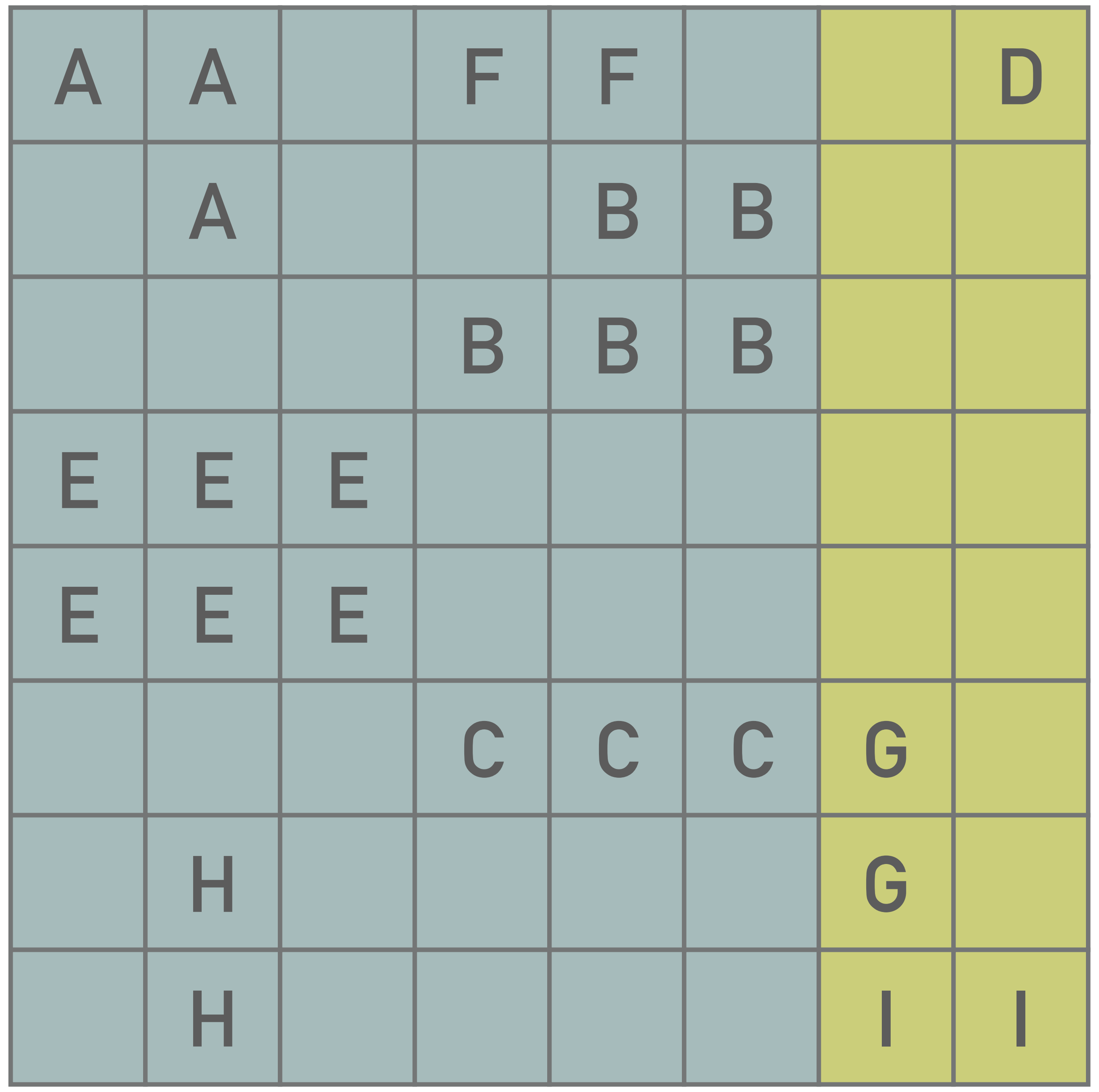
Object	ABCEF	GI	Difference
D	8x6-36=12	2x8-6=10	|12-10|=2
H	6x8-36=12	7x3-6=15	|12-15|=3



- The biggest difference shows the object H, hence it will be inserted into the node which is closer, i.e., ABCEF
- Thus, we have nodes ABCEFH and GI

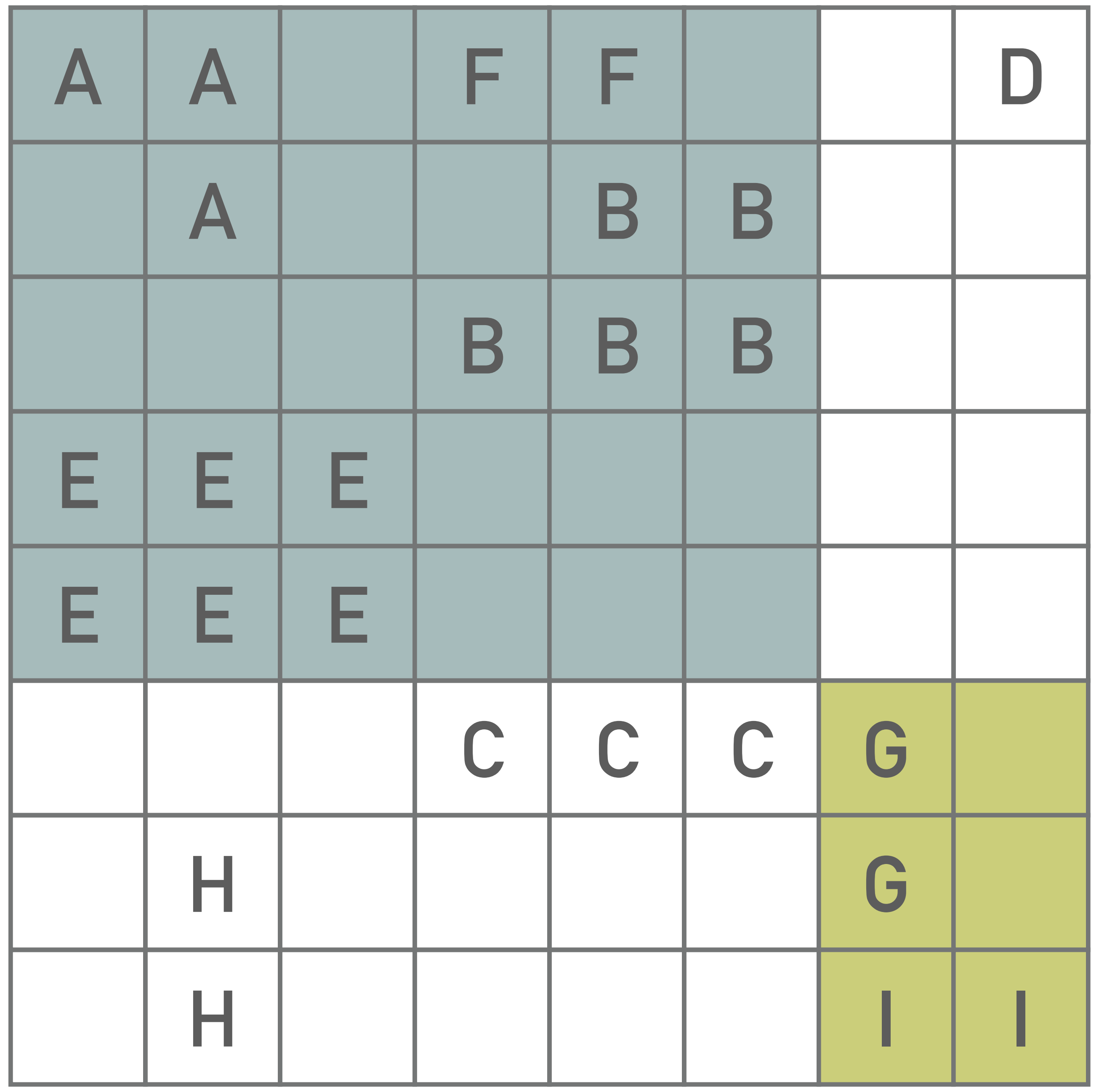
Exercise 6.2 (Solution Continued)

- Finally, object D must be placed in the node GI because the minimum number of items per node is $m = 3$ and $GI = 2$, that is $GI < m$
  - As a result, we have nodes ABCEFH and DGI



Exercise 6.3

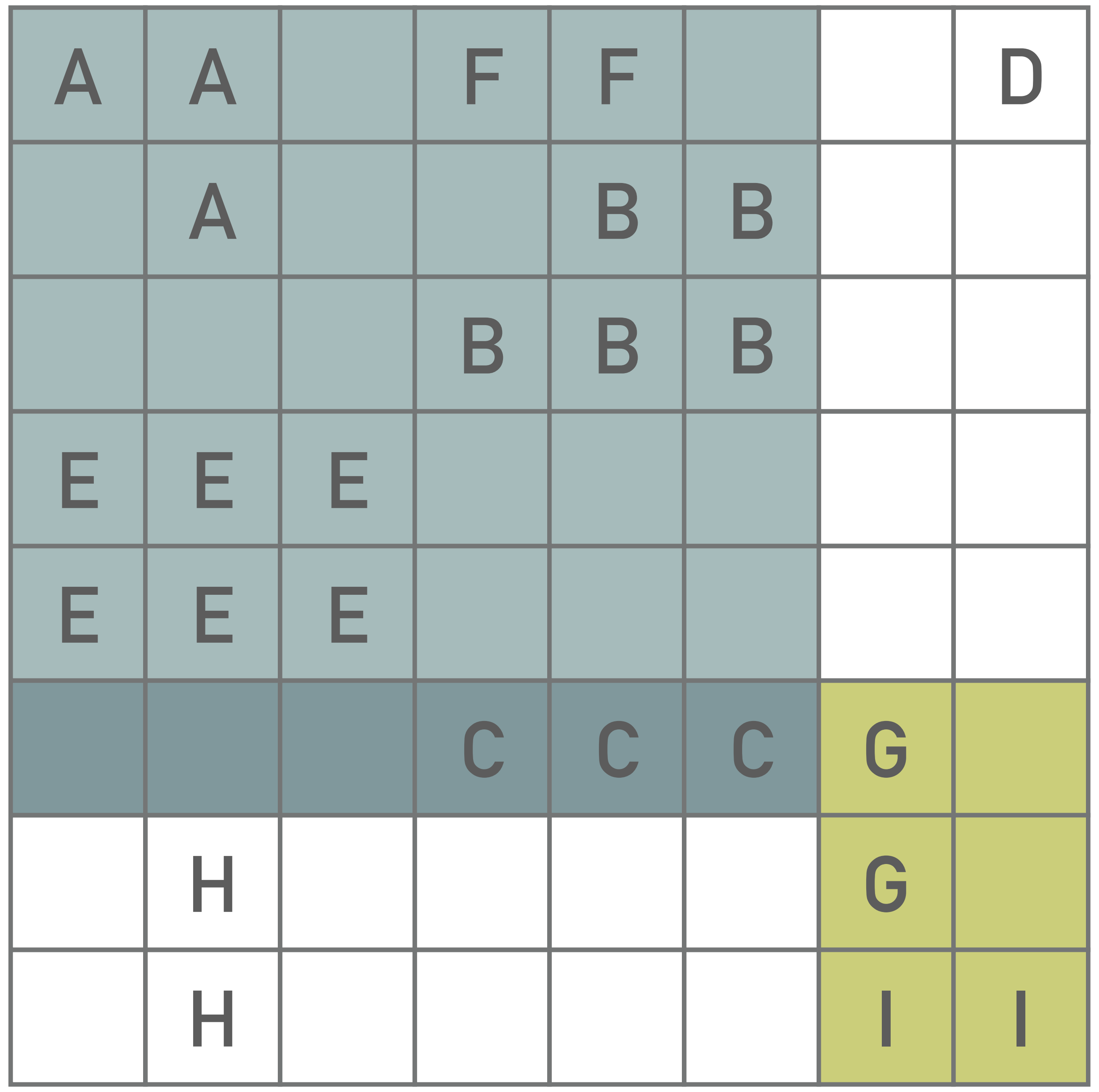
- Finish splitting of the overflown node
  - Continue with Guttman's method
  - The maximum number of items in a node is $M = 8$
  - This time, the minimum number of items in a node is $m = 4$, i.e., $m = M/2$
- If there are more options to choose, explain the reason of yours choice
- Compare and comment the results of exercises 6.2 and 6.3



Exercise 6.3 (Solution)

- Next, iteratively add such an object into a node which will maximize the difference in the node area enlargements if the object was inserted into the first or second node

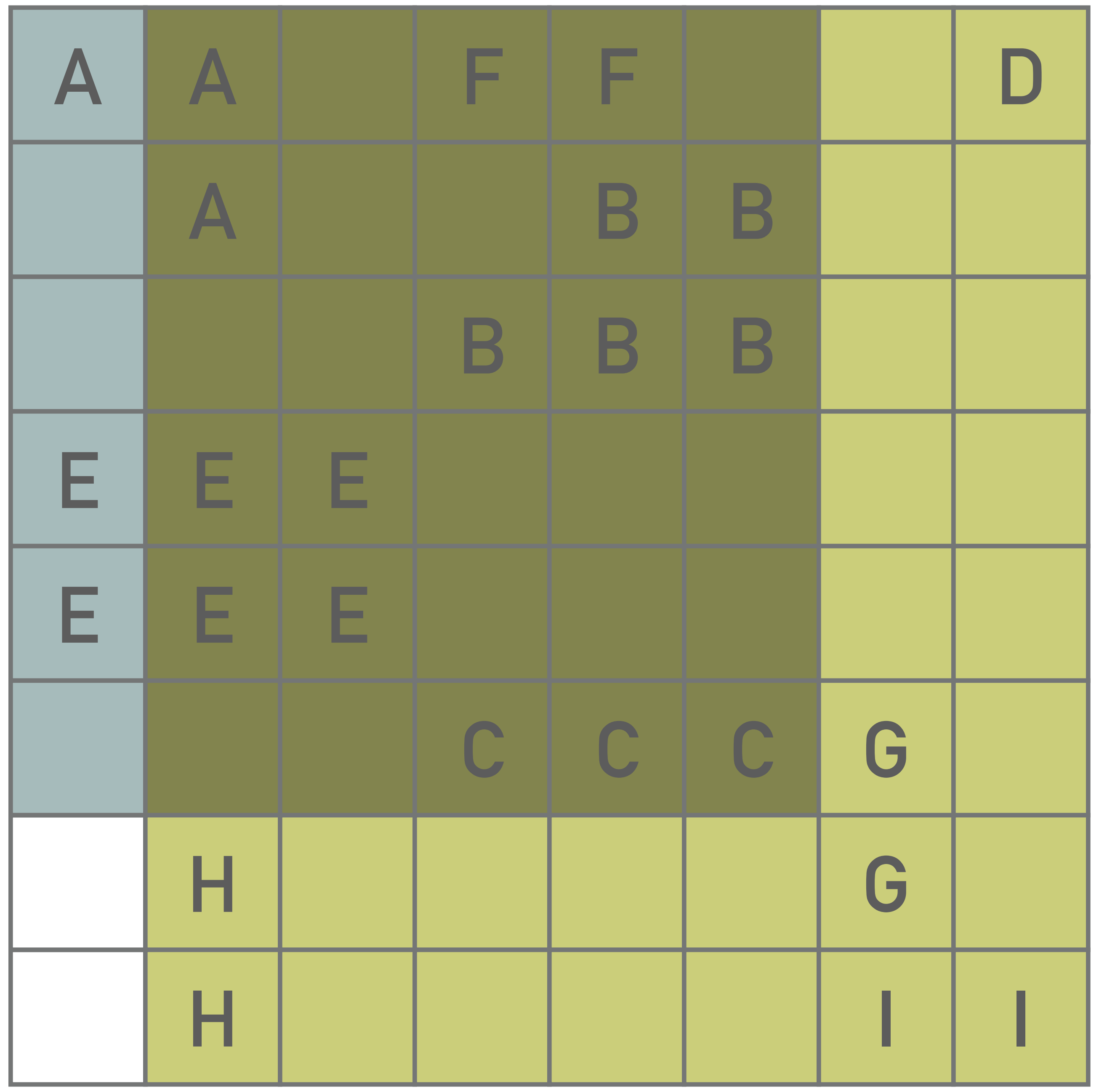
Object	ABEF	GI	Difference
C	6x6-30=6	5x3-6=9	|6-9|=3
D	8x5-30=10	2x8-6=10	|10-10|=0
H	6x8-30=18	7x3-6=15	|18-15|=3



- The biggest difference shows the object C and H, yet we choose C being inserted into the node which is closer, i.e., ABEF
- Thus, we have nodes ABCEF and GI

Exercise 6.3 (Solution Continued)

- Finally, objects H and D must be places in the node GI because the minimum number of items per node is $m = 4$ and $GI = 2$, i.e., $GI < m$
  - As a result, we have nodes ABCEF and DGHI
  - There is a smaller death space in node ABCEF but for a price of a huge overlapping area, therefore it is already better to use smaller value of m in this particular case



Splitting in R-Tree: Greene

- Modification of the split algorithm in original R-Tree (Guttman)
- Splitting is based on a hyperplane which defines in which node the objects will fall
  - I.e., it splits objects into two groups
- We choose an Axis
  - *PickSeeds* work identically to the Guttman's
  - We compute the *normalized distances of each axis* and select the axis having the highest value
- Next, we *order the objects* based on the selected axis
- Finally, we *redistribute* the objects

SplitNode(P, PP, E)
ChooseAxis();
Distribute();

ChooseAxis()
PickSeeds; { from Guttman's version - returns seeds E_i and E_j }
For every axis compute the distance between MBRs E_i, E_j;
Normalize the distance by the respective edge length of the bounding rectangle of the original node;
Pick the axis with greatest normalized separation;

Distribute()
Sort E_is in the choosen axis j based on the j-th coordinate;
Add first $\lceil (M+1)/2 \rceil$ records into P and rest of them into PP;

Example 6.4: Greene's Split

- Split the following overflown node with Greene's split method
  - The maximum number of items in a node is $M = 8$
  - The minimum number of items in a node is $m = 3$

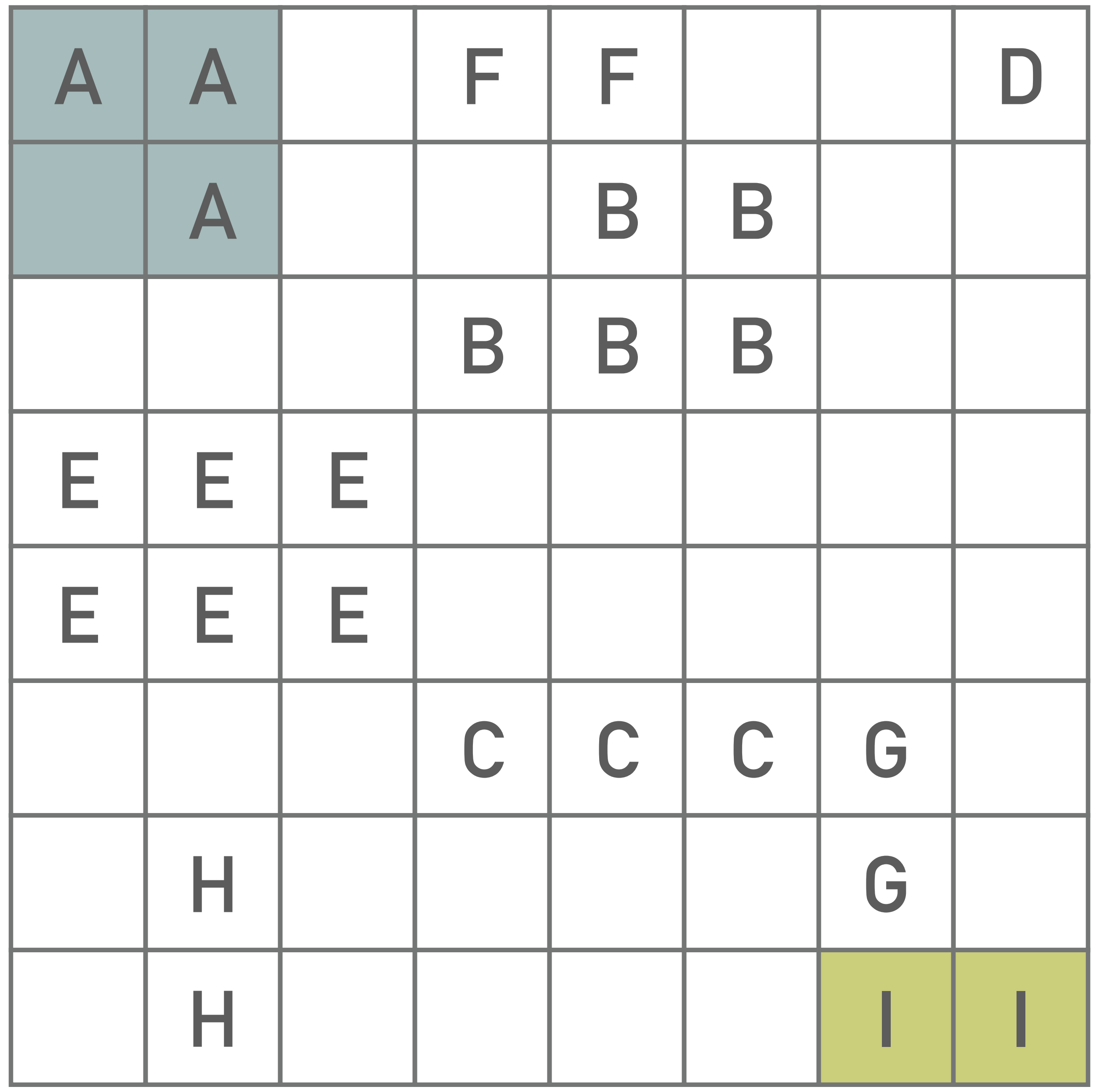
- I.e., execute the following methods:
  - PickSeeds (Guttman's)
  - ChooseAxis
  - Distribute (ordering and placement)

A	A		F	F			D
	A			B	B		
			B	B	B		
E	E	E					
E	E	E					
			C	C	C	G	
	H					G	
	H					I	I

Example 6.4: Greene's Split (Continued)

- We apply Guttman's PickSeeds method to find two elements having the largest dead space if being placed together

Pair	Overall area	Area of the objects	Dead space
AB	18	8	10
AD	16	4	12
…			
AI	64	5	59
…			
DH	56	3	53
…			
HI	14	4	10



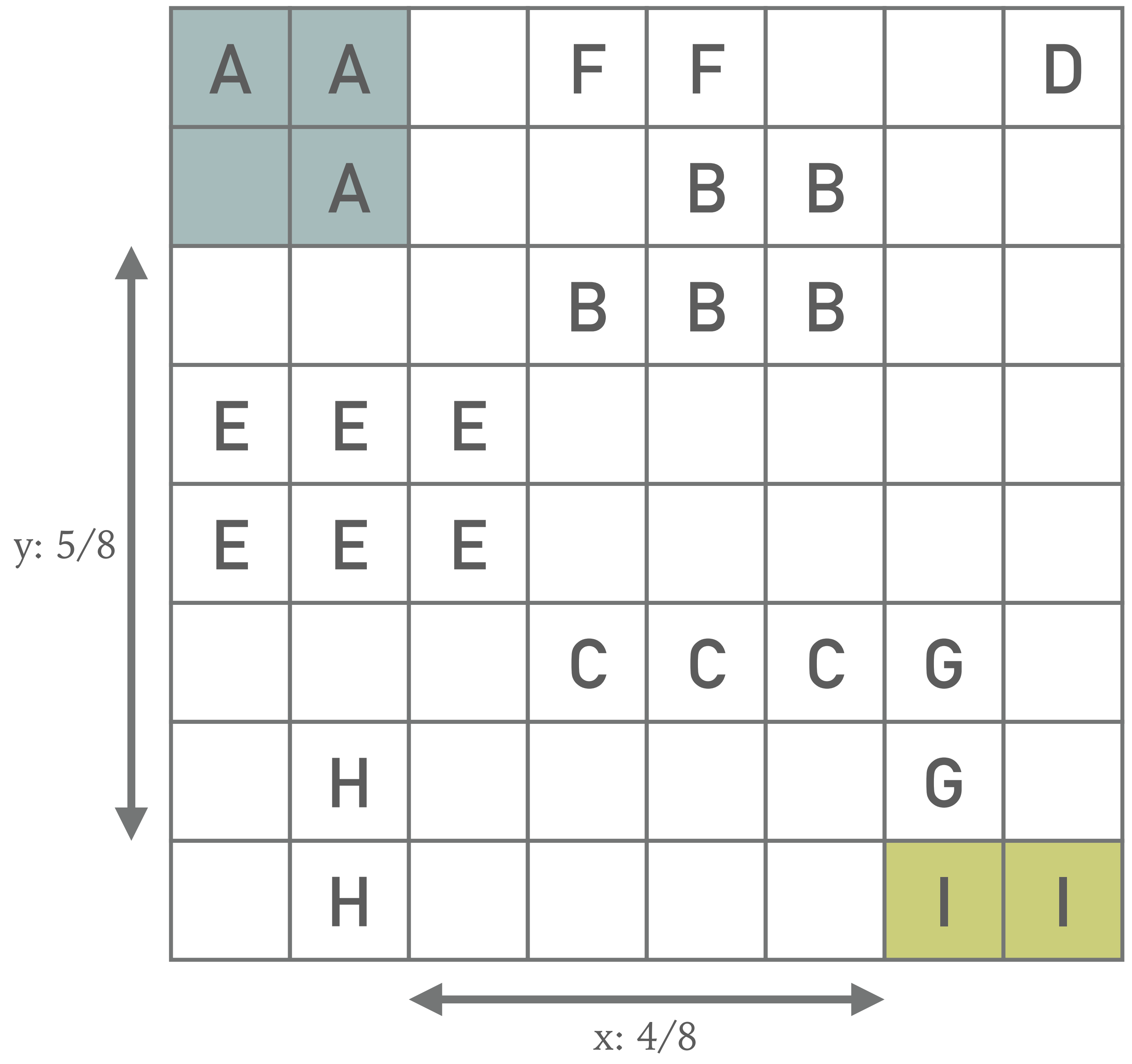
- The largest dead space has AI thus those will be the seeds of the splitting method

Example 6.4: Greene's Split (Continued)

- Having selected seeds, we compute the normalized distances of A and I along each of the axis and pick the axes with higher distance (better separation)

- x: 4/8 = 0.5
- y: 5/8 = 0.625

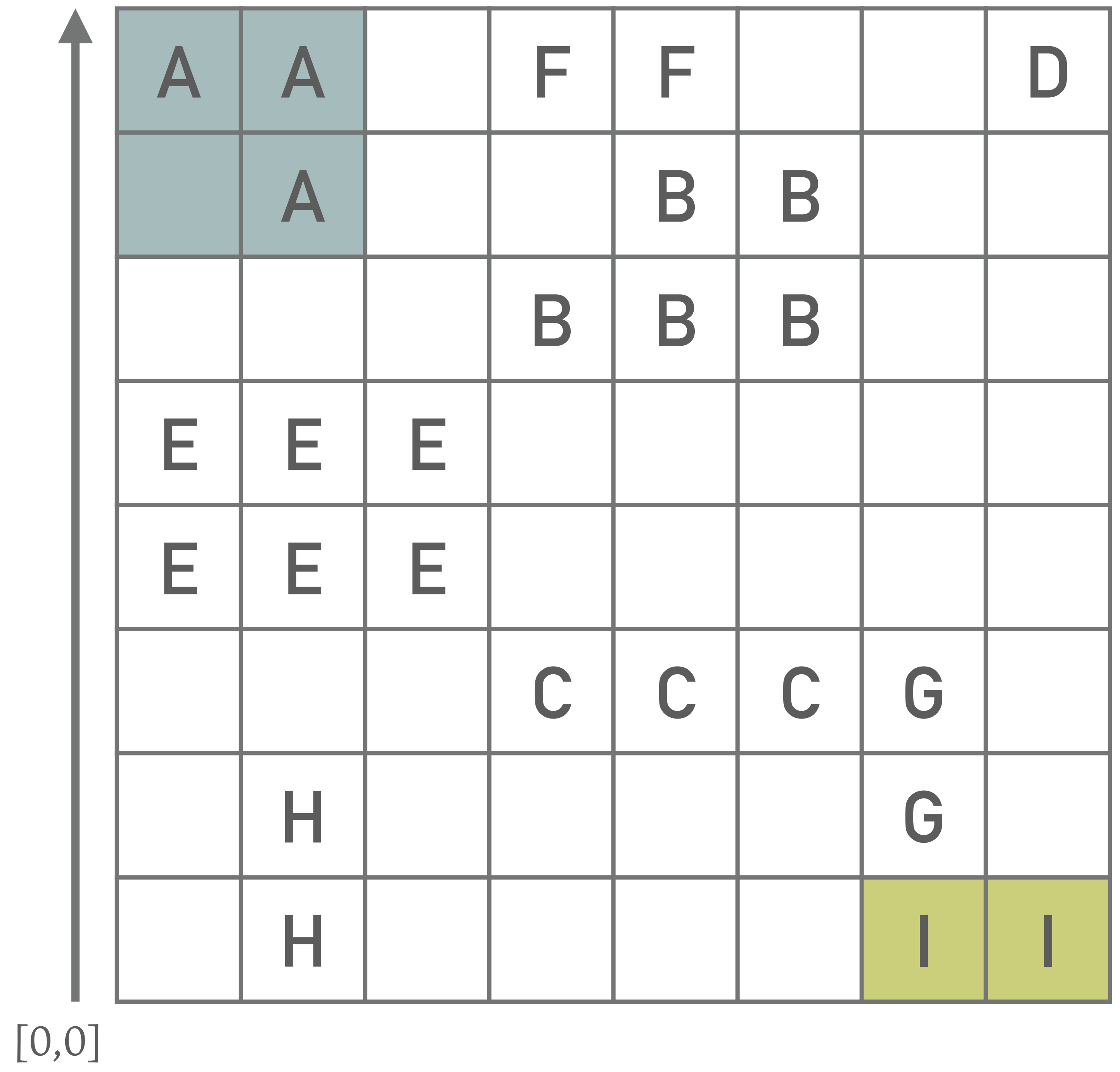
- In this particular case, the axis y is better separating A and I



Example 6.4: Greene's Split (Continued)

- Now we order the objects based on their y-axis
  - I.e., we start from the coordination [0,0]

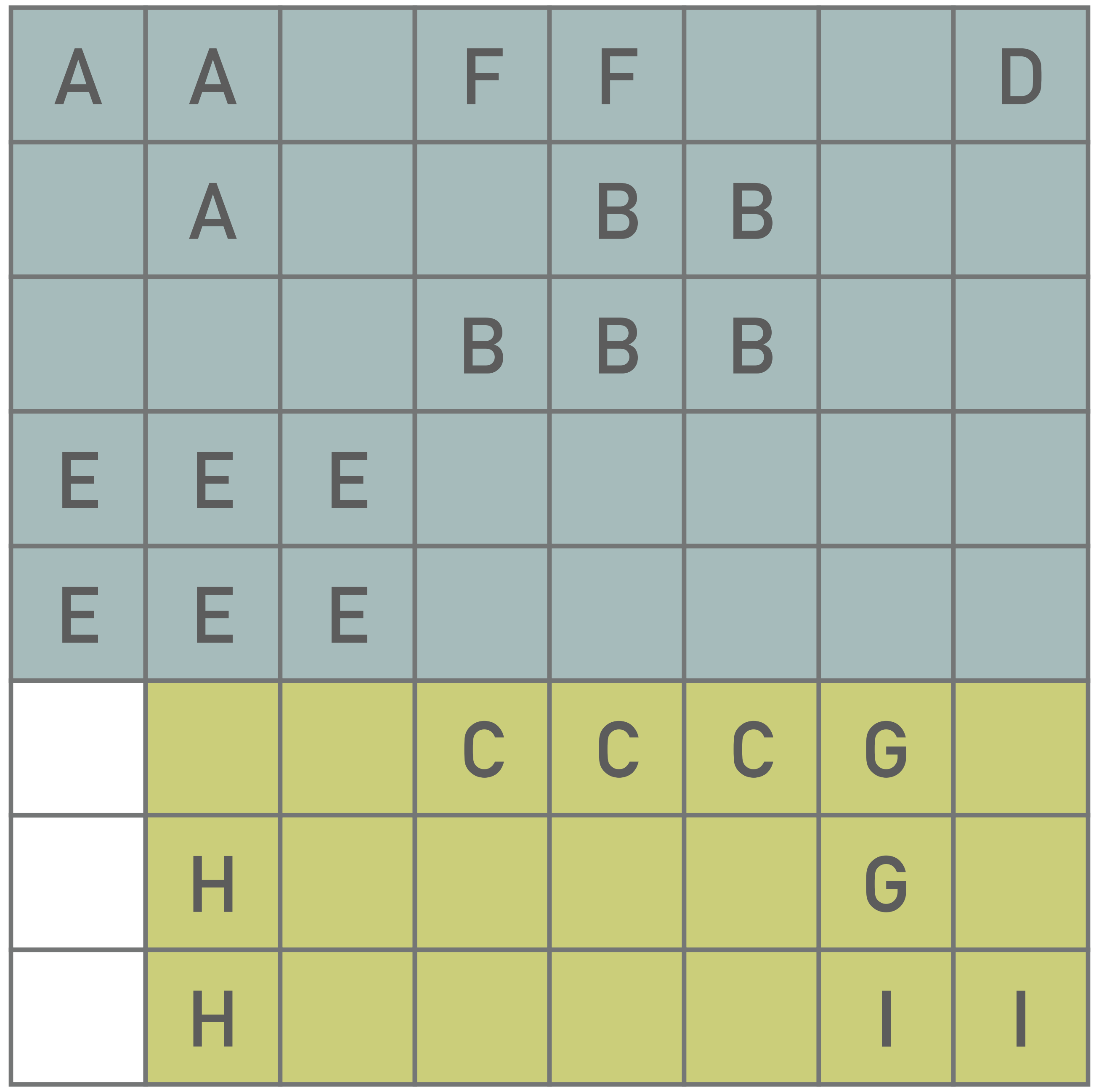
Object	I	H	G	C	E	B	A	F	D
Start	0	0	1	2	3	5	6	7	7
end	0	1	2	2	4	6	7	7	7

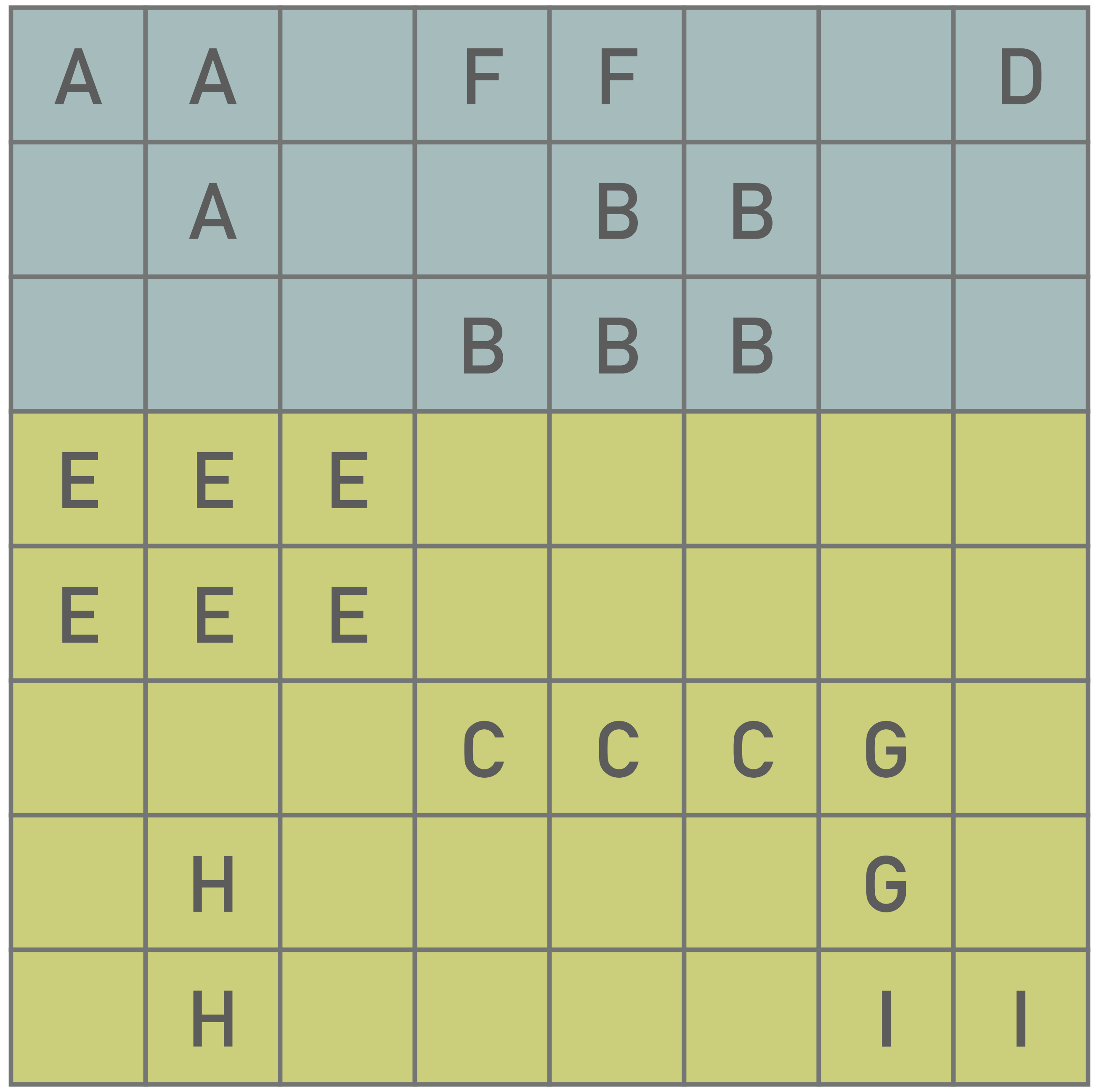


- If two objects start at the same level, we select first the one that ends at lower level
- If two or mode objects starts and ends at the same level, the order is arbitrary

Example 6.4: Greene's Split (Continued)

- We place half of the objects in one node and the other half into the second node
  - There are two possible solutions:





Exercise 6.5

- Split the following overflown node with Greene's split method
  - The maximum number of items in a node is $M = 9$
  - The minimum number of items in a node is $m = 3$

G			A			I	I				J
G		A	A	A							
			A					C			
	F							C	C		
	F				H	H		C			
	F										
				E	E	E		B	B	B	
D	D	D							B		

- That is, execute the following methods:
  - PickSeeds
  - ChooseAxis
  - Distribute (ordering and placement)

Exercise 6.5 (Solution)

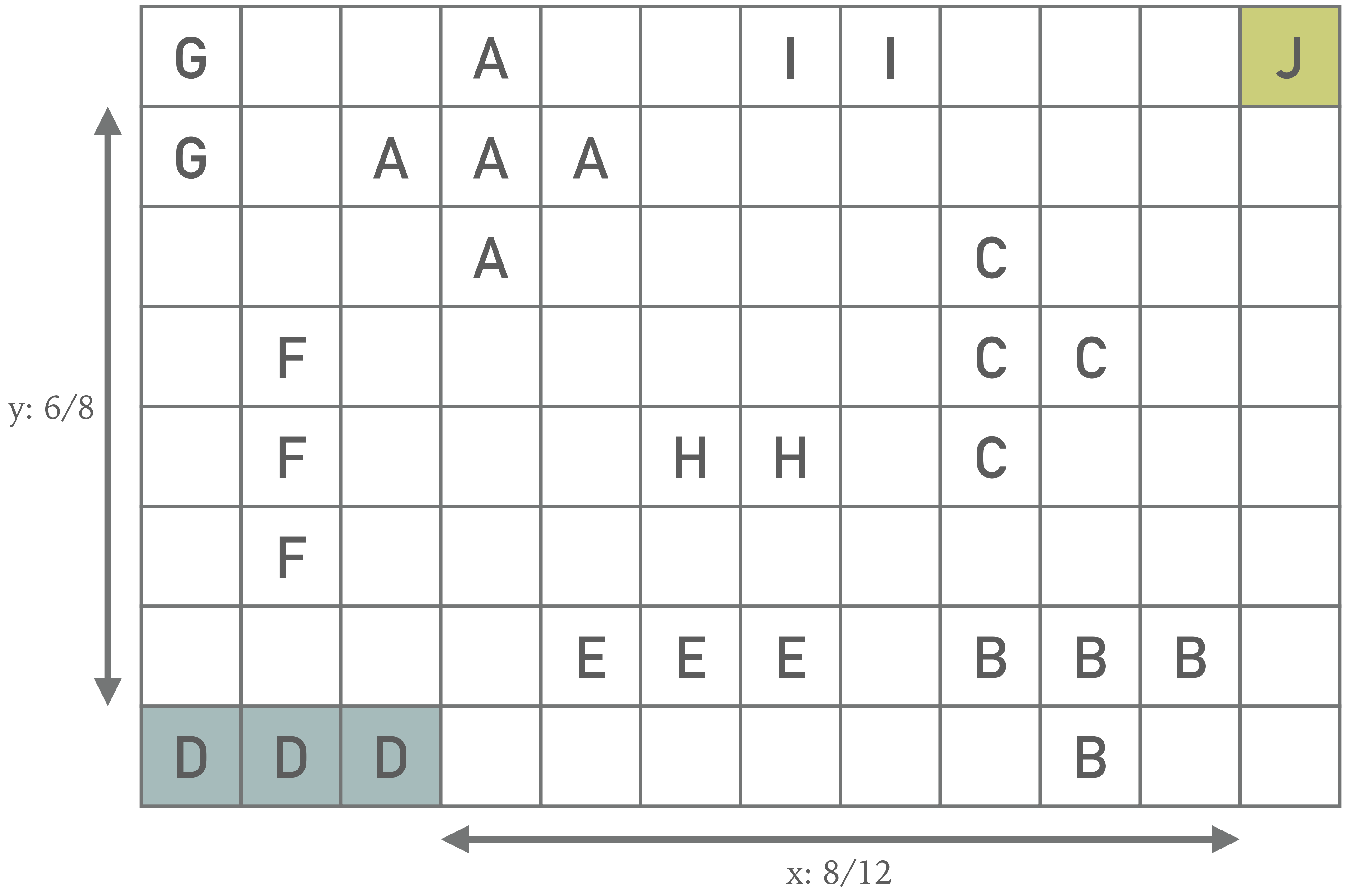
- PickSeeds
  - The largest dead space has DJ thus those will be the seeds of the splitting method

Pair	Overall area	Area of the objects	Dead space
AB	9x8=72	5+4=9	72-9=63
AC	8x5=40	5+4=9	40-9=31
…			
BG	11x8=88	4+2=6	88-6=82
…			
DJ	12x8=96	3+1=4	96-4=92
…			
IJ	6x1=6	2+1=3	6-3=3

G			A			I	I				J
G		A	A	A							
			A					C			
	F							C	C		
	F				H	H		C			
	F										
				E	E	E		B	B	B	
D	D	D							B		

Exercise 6.5 (Solution Continued)

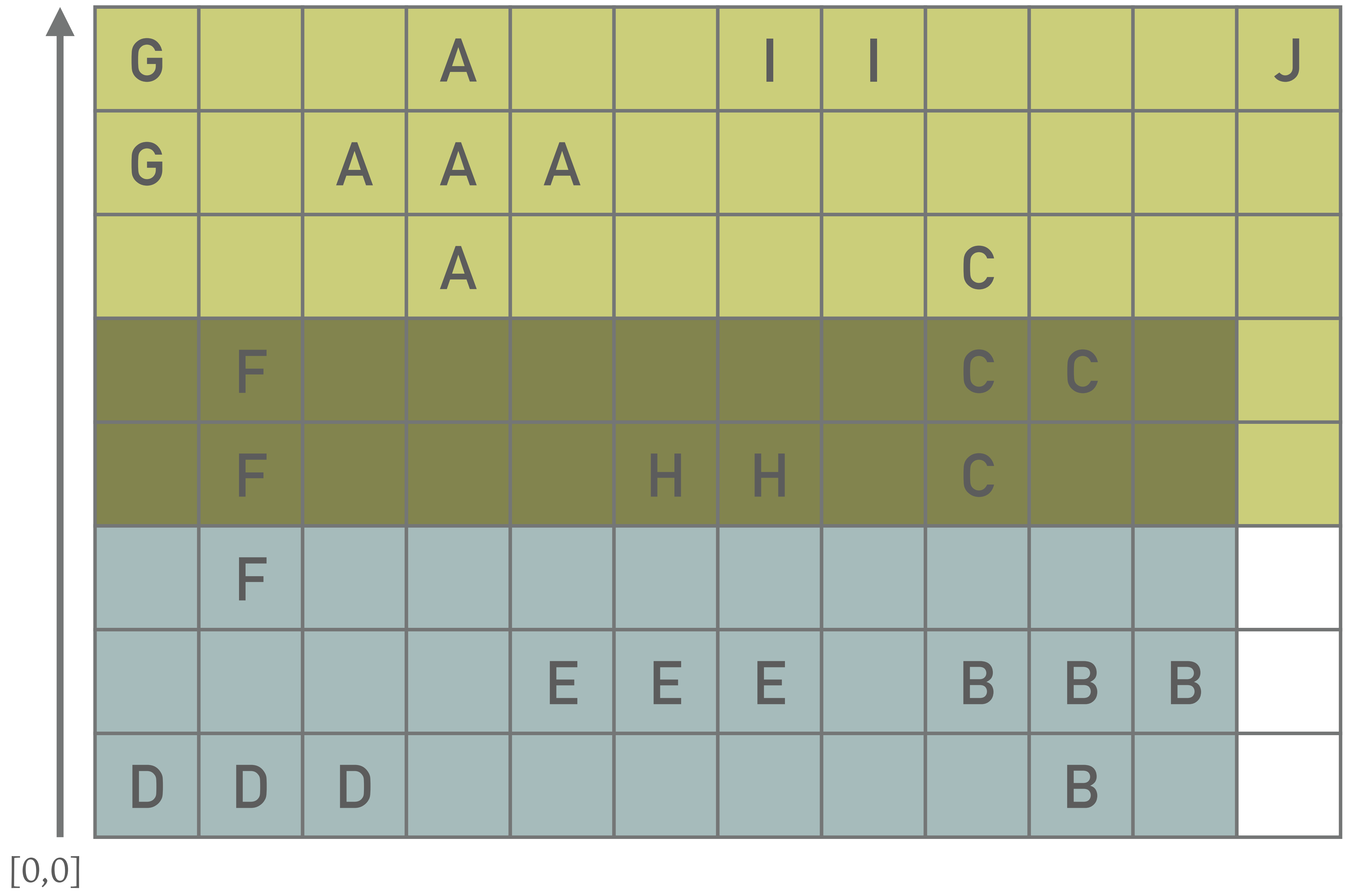
- ChooseAxis
  - x: 8/12 = 0.667
  - y: 6/8 = 0.750
  - In this particular case, the axis y is better separating D and J



Exercise 6.5 (Solution Continued)

- Distribute according to axis y

Object	D	B	E	F	H	C	A	G	I	J
Start	0	0	1	2	3	3	5	6	7	7
end	0	1	1	4	3	5	7	7	7	7



- The solution:
  - BDEFH || ACGIJ

Splitting in R* Tree

- R* tree tries to minimize coverage (area) and overlap by adding another criterion, i.e., margin
  - It is used only for the level above the leaf level
  - Other levels are split based on Guttman
- For every two group we can compute the following auxiliary values:
  - *margin-value*, i.e., sum of margins (surfaces) of the two groups
  - *overlap-value*, i.e., volume of the overlap of the two groups
  - *area-value*, i.e., sum of volumes of the two groups

Split_RS(P, PP, E)
ChooseSplitAxis();
Distribute();

ChooseSplitAxis()
FOREACH axis DO
Sort the entries along given axis;
$S \longleftarrow$ sum of all margin-values of all different distributions;
Choose the axis with the minimum S as split axis;

Distribute()
Along the split axis, choose the distribution with minimum overlap-value. Resolve ties by choosing the distribution with minimum area-value;

Example 6.6: Splitting in R* Tree

- Split the following overflown node with R* Tree split method
    - The maximum number of items in a node is $M = 8$
    - The minimum number of items in a node is $m = 3$
- That is, execute the following methods:
    - ChooseSplitAxis (i.e., compute margin-value)
    - Distribute (i.e., compute overlap-value and area-value)

A	A		F	F			D
	A			B	B		
			B	B	B		
E	E	E					
E	E	E					
			C	C	C	G	
	H					G	
	H					I	I

Example 6.6: Splitting in R* Tree (Continued)

- First, we compute the margin-values for every possible distributions of objects with regard to the x and y axis
  - margin-value: $margin(MBR(G_1)) * 2 + margin(MBR(G_2)) * 2$
- These are summed and such an axis is chosen which minimizes the sum
- Ordering* based on the x-axis: AEHFBCGID
  - margin-value (AEH || FBCGID) = (3+8)*2+(5+8)*2=22+26=48
  - margin-value (AEHF || BCGID) = (5+8)*2+(5+8)*2=26+26=52
  - margin-value (AEHFB || CGID) = (6+8)*2+(5+8)*2=28+26=54
  - margin-value (AEHFBC || GID) = (6+8)*2+(2+8)*2=28+20=48
  - Sum = 48 + 52 + 54 + 48 = 202

A	A		F	F			D
	A			B	B		
			B	B	B		
E	E	E					
E	E	E					
			C	C	C	G	
	H					G	
	H					I	I

AEHFBCGID

* If two objects start at the same level, we select first the one that ends at lower level. Or if two or mode objects starts and ends at the same level, the order is arbitrary.

Example 6.6: Splitting in R* Tree (Continued)

- Ordering* based on the y-axis: IHGCEBAFD
  - margin-value (IHG || CEBAFD) = (8+3)*2+(8+6)*2=22+28=50
  - margin-value (IHGC || EBAFD) = (8+3)*2+(8+5)*2=22+26=48
  - margin-value (IHGCE || BAFD) = (8+5)*2+(8+3)*2=26+22=48
  - margin-value (IHGCEB || AFD) = (8+7)*2+(8+2)*2=30+20=50
  - Sum = 50+48+48+50=196

- X-axis: 202
- Y-axis: 196
  - Therefore we chose splitting along the y-axis

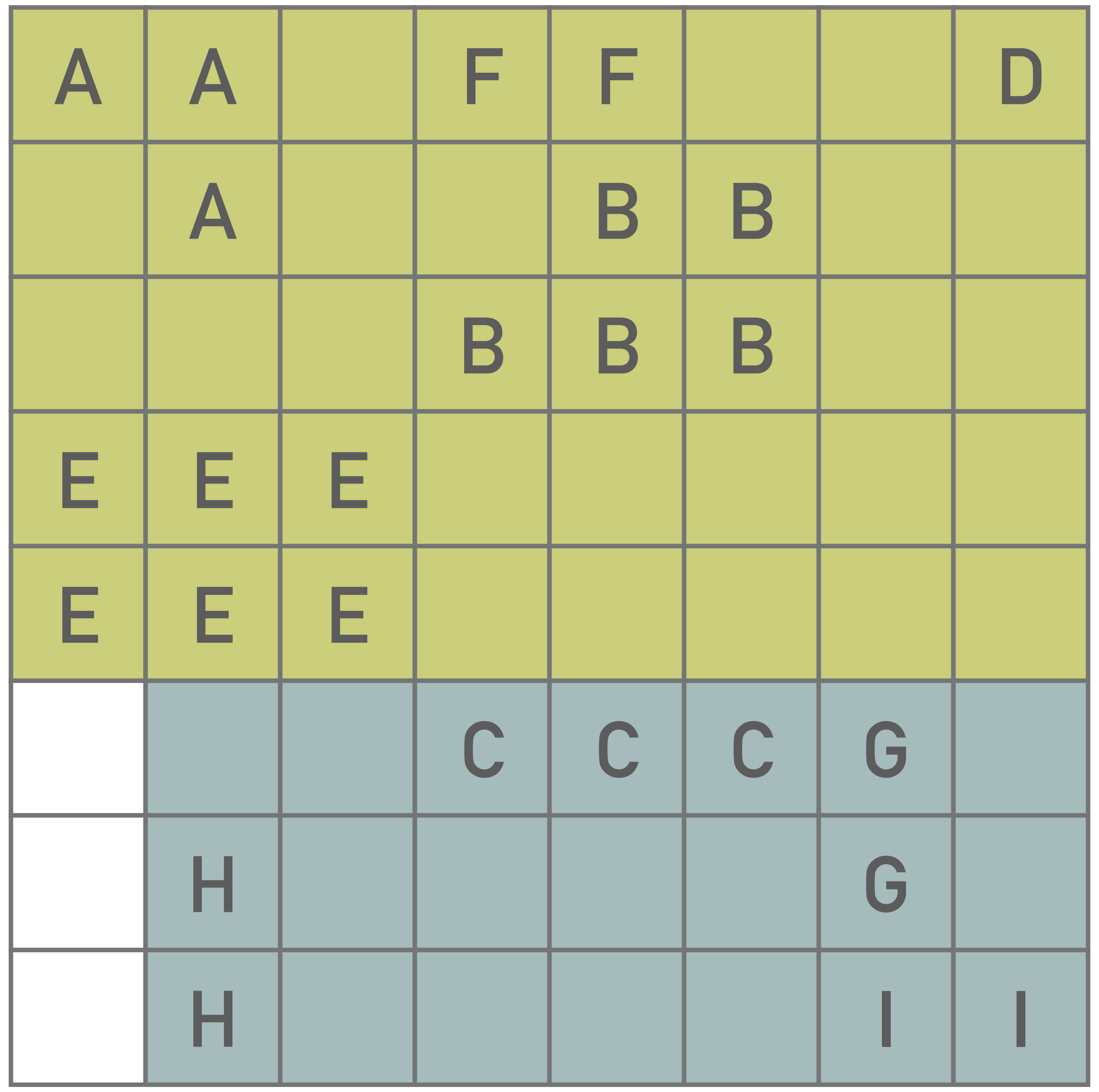
A	A		F	F			D
	A			B	B		
			B	B	B		
E	E	E					
E	E	E					
			C	C	C	G	
	H					G	
	H					I	I

IHGCEBAFD

* If two objects start at the same level, we select first the one that ends at lower level. Or if two or mode objects starts and ends at the same level, the order is arbitrary.

Example 6.6: Splitting in R* Tree (Continued)

- Now we compute the overlap-values among all the distributions (along y-axis) and pick the distribution that minimizes the overlap
  - overlap-value (IHG || CEBAFD) = 7 (row CCCG)
  - overlap-value (IHGC || EBAFD) = 0
  - overlap-value (IHGCE || BAFD) = 0
  - overlap-value (IHGCEB || AFD) = 8 (row ABB)



- If more distributions lead to the minimum overlap, the one is chosen which shows the smalles area-value
  - area-value (IHGC || EBAFD) = (7*3)+(8*5)=21+40=61
  - area-value (IHGCE || BAFD) = (8*5)+(8*3)=40+24=64

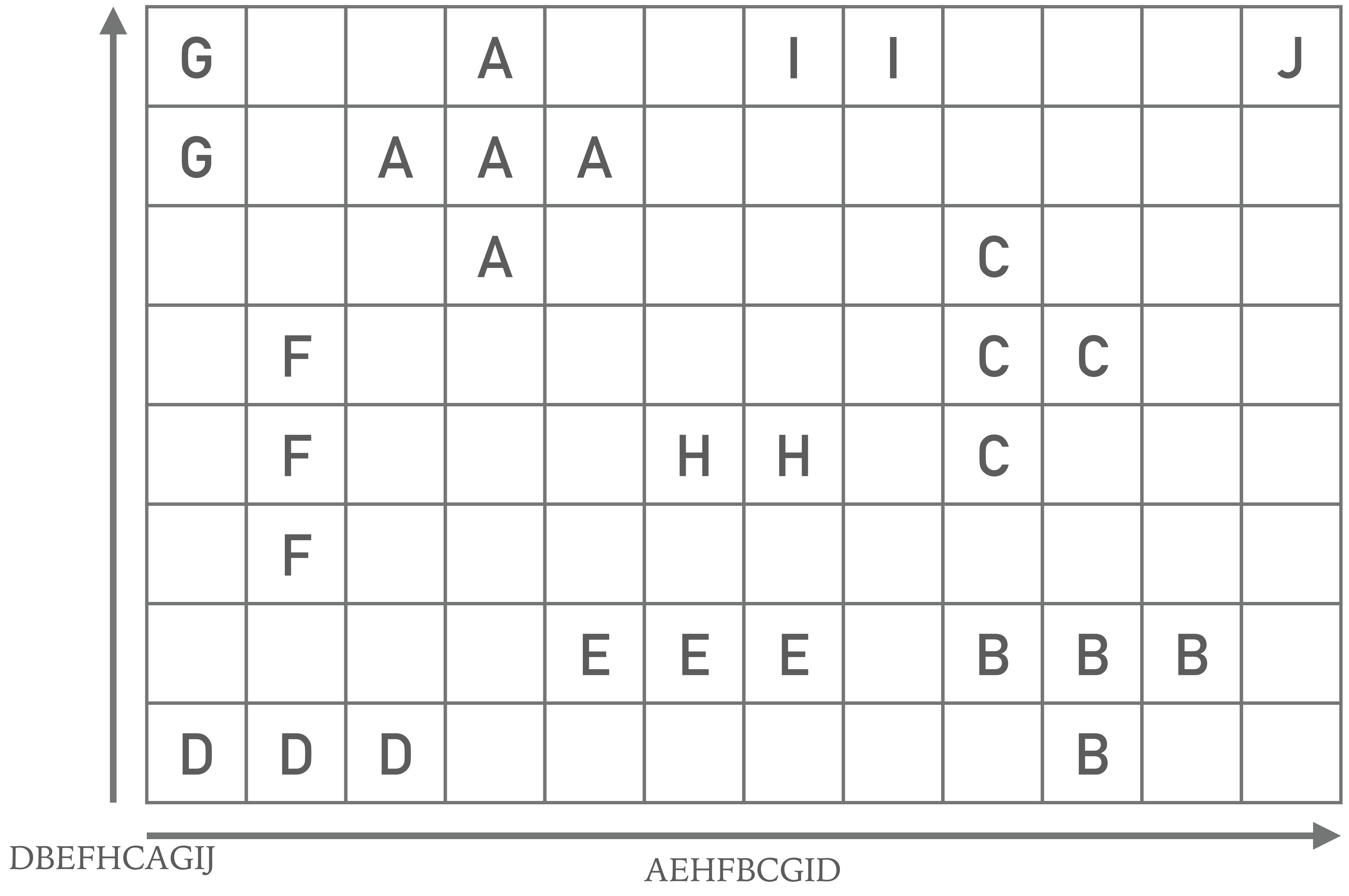
Exercise 6.7

- Split the following overflown node with R* Tree split method
  - The maximum number of items in a node is $M = 9$
  - The minimum number of items in a node is $m = 3$
- That is, execute the following methods:
  - ChooseSplitAxis
  - Distribute
- Illustrate the result

G			A			I	I				J
G		A	A	A							
			A					C			
	F							C	C		
	F				H	H		C			
	F										
				E	E	E		B	B	B	
D	D	D							B		

Exercise 6.7 (Solution)

- Ordering* based on the x-axis: GCFAEHICBJ
  - margin-value (GCF || AEHICBJ) = (3+8)*2+(10+8)*2=22+36=58
  - margin-value (GCFA || EHICBJ) = (5+8)*2+(8+8)*2=26+32=58
  - margin-value (GCFAE || HICBJ) = (7+8)*2+(7+8)*2=30+30=60
  - margin-value (GCFAEH || ICBJ) = (7+8)*2+(6+8)*2=30+28=58
  - margin-value (GCFAEHI || CBJ) = (8+8)*2+(4+8)*2=32+24=56
  - Sum = 58+58+60+58+56 = 290



- Ordering* based on the y-axis: DBEFHCAGIJ
  - margin-value (DBE || FHCAGIJ) = (11+2)*2+(12+6)*2=26+36=62
  - margin-value (DBEF || HCAGIJ) = (11+5)*2+(12+5)*2=32+34=66
  - margin-value (DBEFH || CAGIJ) = (11+5)*2+(12+5)*2=32+34=66
  - margin-value (DBEFHC || AGIJ) = (11+6)*2+(12+3)*2=34+30=64
  - margin-value (DBEFHCA || GIJ) = (11+8)*2+(12+2)*2=38+24=62
  - Sum = 62+66+66+64+62 = 320

* If two objects start at the same level, we select first the one that ends at lower level. Or if two or mode objects starts and ends at the same level, the order is arbitrary.

Exercise 6.7 (Solution Continued)

- We chose splitting along the x-axis (smaller sum)
  - overlap-value (GCF || AEHICBJ) = 8 (column AD)
  - overlap-value (GCFA || EHICBJ) = 8 (column AE)
  - overlap-value (GCFAE || HICBJ) = 16 (columns HE; IHE)
  - overlap-value (GCFAEH || ICBJ) = 8 (column IHE)
  - overlap-value (GCFAEHI || CBJ) = 0

G			A			I	I				J
G		A	A	A							
			A					C			
	F							C	C		
	F				H	H		C			
	F										
				E	E	E		B	B	B	
D	D	D							B		

- There is only one distribution having the smallest overlap, therefore the area-value does not have to be computed

- The result is: GCFAEHI || CBJ