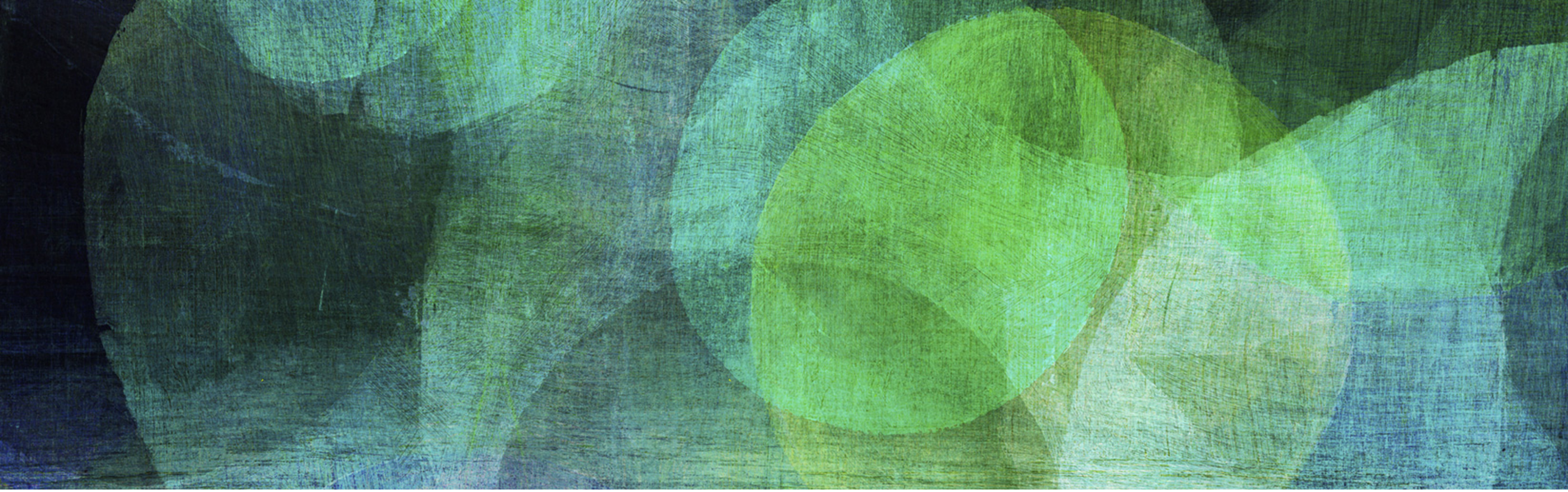
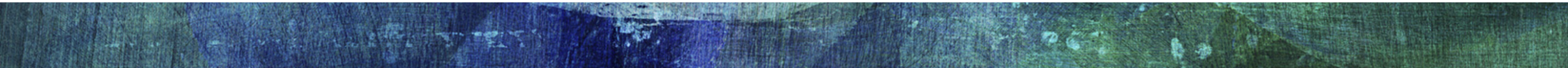


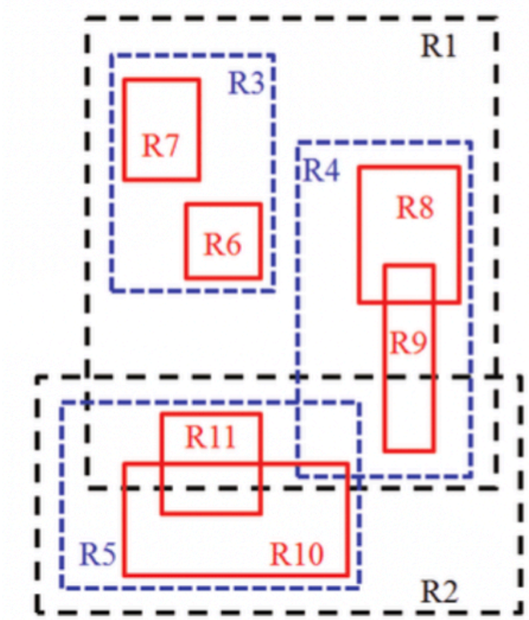
R-TREES

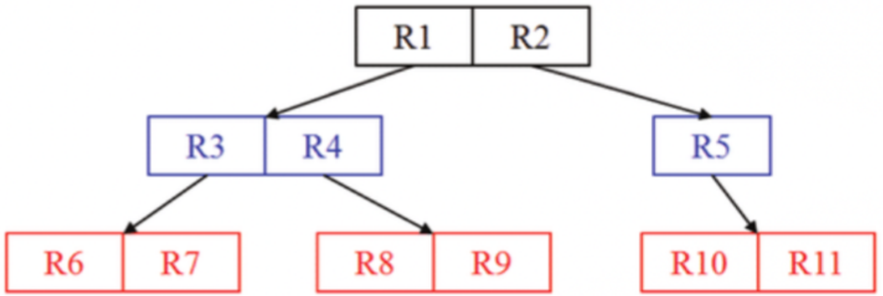
NDBI007: Practical Class 6



R-TREE

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





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 the 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
    $E_i \leftarrow$ **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
  $d_{ij} \leftarrow$ area(J) − area(E_i.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
  $d_1 \leftarrow$ area increase required for MBR of P and E_i.I;
  $d_2 \leftarrow$ area increase required for MBR of PP and E_i.I;
pick E_i with maximal $|d_1 - d_2|$;

EXAMPLE 1: GUTTMAN'S SPLIT

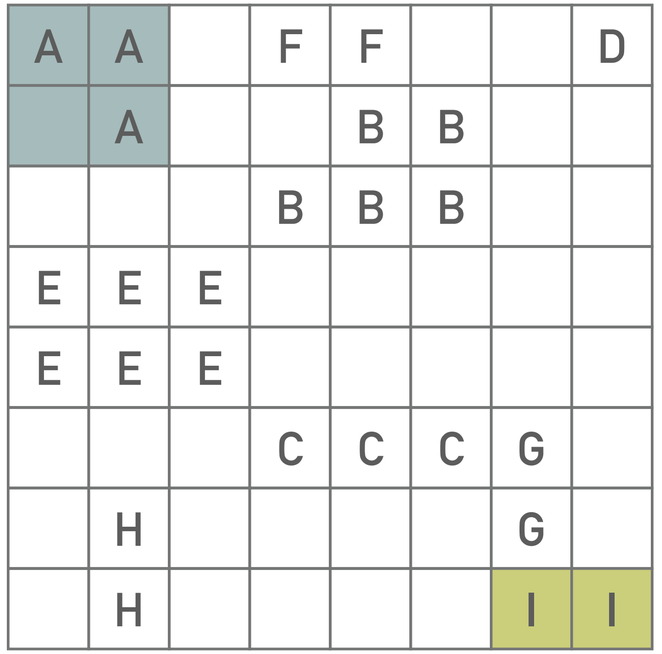
- Split the following overflown node with Guttman's splitnode method
  - The maximum number of items in a node is $M = 8$
  - The minimum number of items in a node is $m = 3$

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 1: GUTTMAN'S SPLIT

- 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

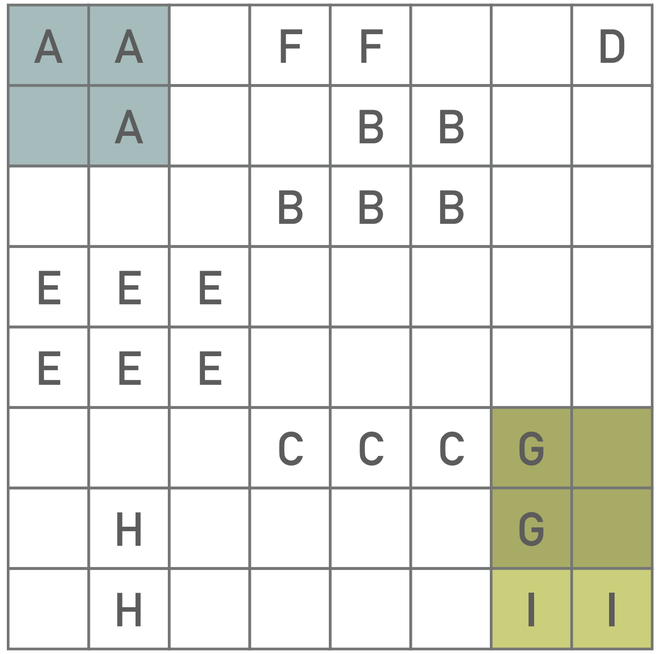


- The largest dead space has AI thus those will be the seeds of the splitting method

EXAMPLE 1: GUTTMAN'S SPLIT

- Next, iteratively add such an object into a node which will maximise 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

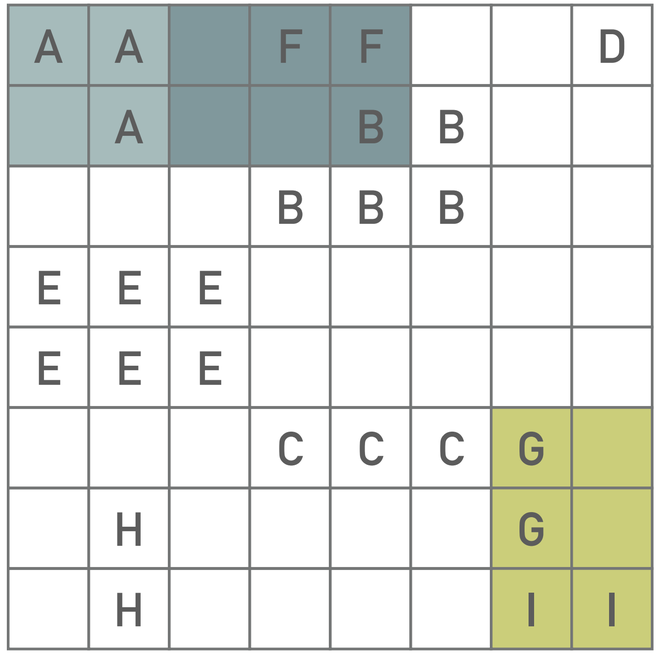


- The biggest difference shows the object G so it will be inserted into the node which is closer
- So now we have nodes A and GI

EXAMPLE 1: GUTTMAN'S SPLIT

- Next, iteratively add such an object into a node which will maximise the difference in the node areas 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

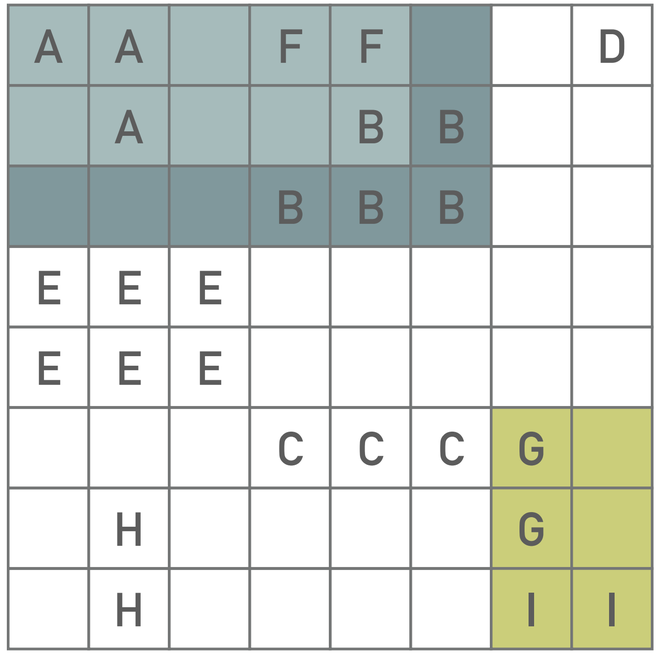


- The biggest difference shows the object F so it will be inserted into the node which is closer, i.e., A
- So now we have nodes AF and GI

EXAMPLE 1: GUTTMAN'S SPLIT

- Next, iteratively add such an object into a node which will maximise the difference in the node areas 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

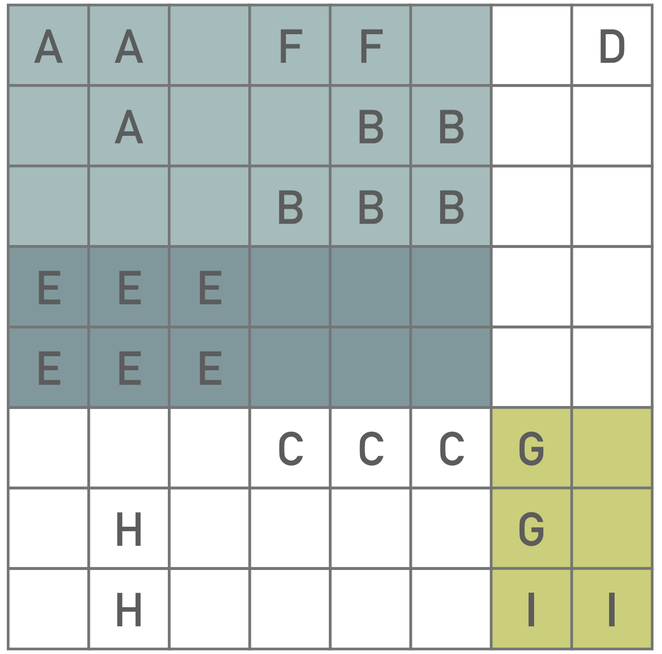


- The biggest difference shows the object B so it will be inserted into the node which is closer, i.e., AF
- So now we have nodes ABF and GI

EXAMPLE 1: GUTTMAN'S SPLIT

- Next, iteratively add such an object into a node which will maximise the difference in the node areas 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 so it will be inserted into the node which is closer, i.e., ABF
- So now we have nodes ABEF and GI

EXERCISE 1

- 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

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

EXERCISE 2

- 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 1 and 2

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

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 *normalised distances of each axis* and select the axis having the highest value
- Next, we *order the objects* based on the selected axis
- Finally, we *distribute* the objects

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

ChooseAxis()
PickSeads; { 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 distances 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 chosen axis j based on the j-th coordinate;
Add first ⌈ (M+1)/2 ⌉ records into P and rest of them into PP;

EXAMPLE 2: 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 $m = 3$

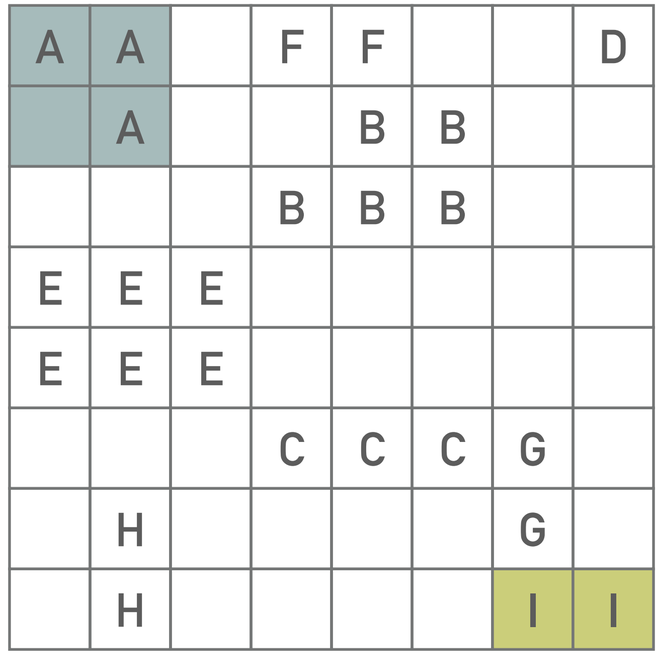
A	A		F	F			D
	A			B	B		
			B	B	B		
E	E	E					
E	E	E					
			C	C	C	G	
	H					G	
	H					I	I

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

EXAMPLE 2: GREENE'S SPLIT

- 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



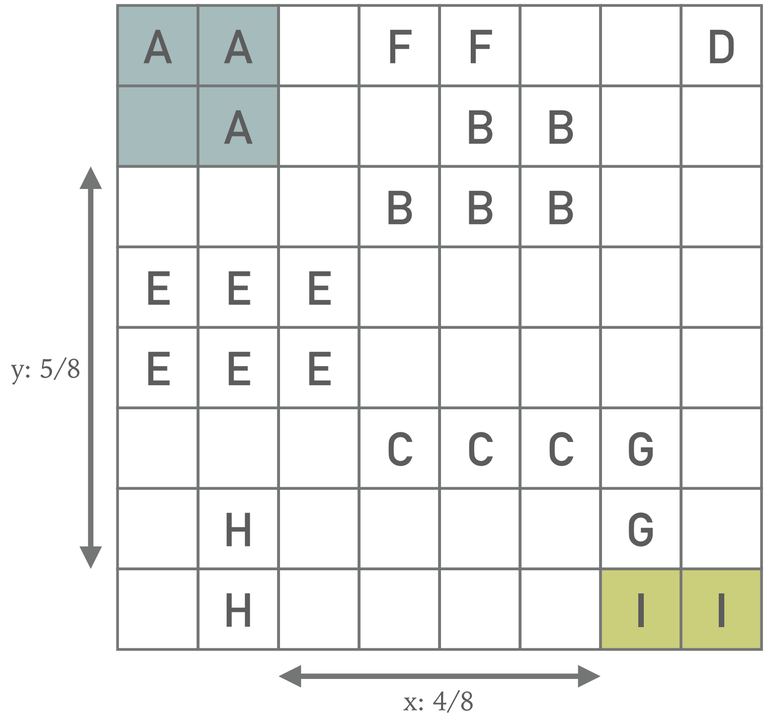
- The largest dead space has AI thus those will be the seeds of the splitting method

EXAMPLE 2: GREENE'S SPLIT

- Having selected seeds, we compute the normalised 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: $5/8 = 0.625$

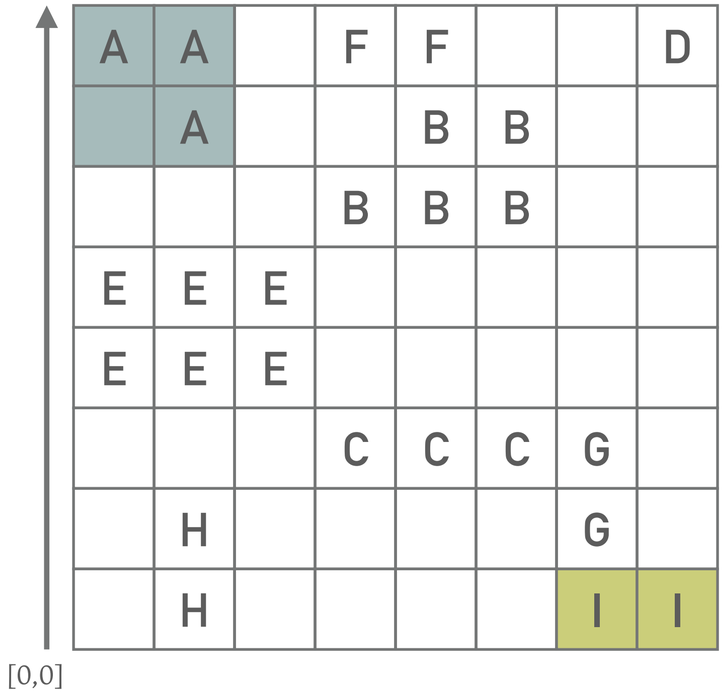
- In our case, the axis better separating A and I is y



EXAMPLE 2: GREENE'S SPLIT

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

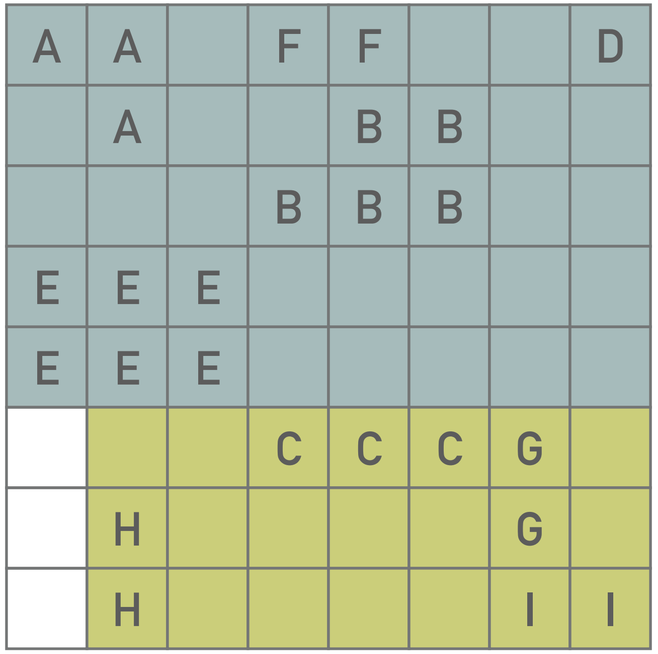
Object	I	H	G	C	E	B	A	F	D
Start	0	0	1	2	3	5	6	7	7
end	0	1	2	2	4	6	7	7	7

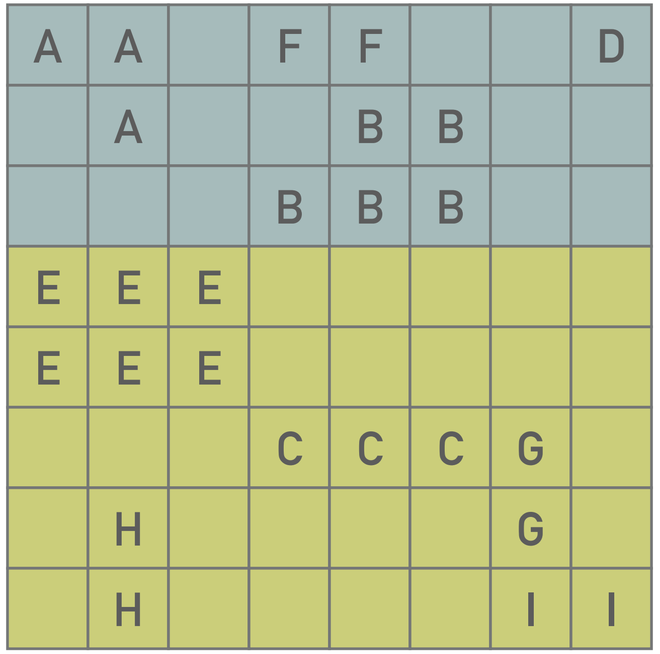


- If two objects start at the same level, we select first the one that ends at lower level
- If two or more objects starts and ends at the same level, the order is arbitrary

EXAMPLE 2: GREENE'S SPLIT

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





EXERCISE 3

- 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		

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

SPLITTING IN R* TREE

- R* tree tries to minimise 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 following auxiliary values:
  - *margin-value* - sum of margins (surfaces) of the two groups
  - *overlap-value* - volume of the overlap of the two groups
  - *area-value* - sum of volumes of the two groups

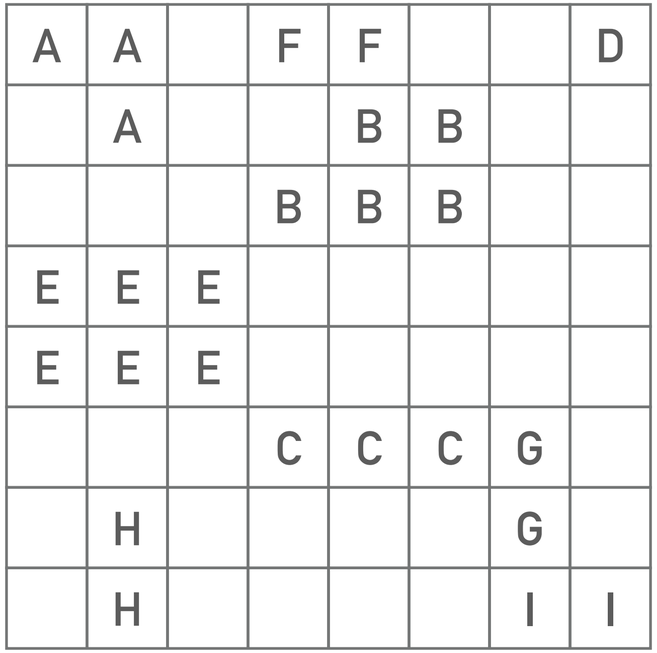
Split_RS(P,PP,E)
ChooseSplitAxis();
Distribute();

ChooseSplitAxis
FOREACH axis DO
  Sort the entries along given axis;
  S ← 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 3: SPLITTING IN R* TREE

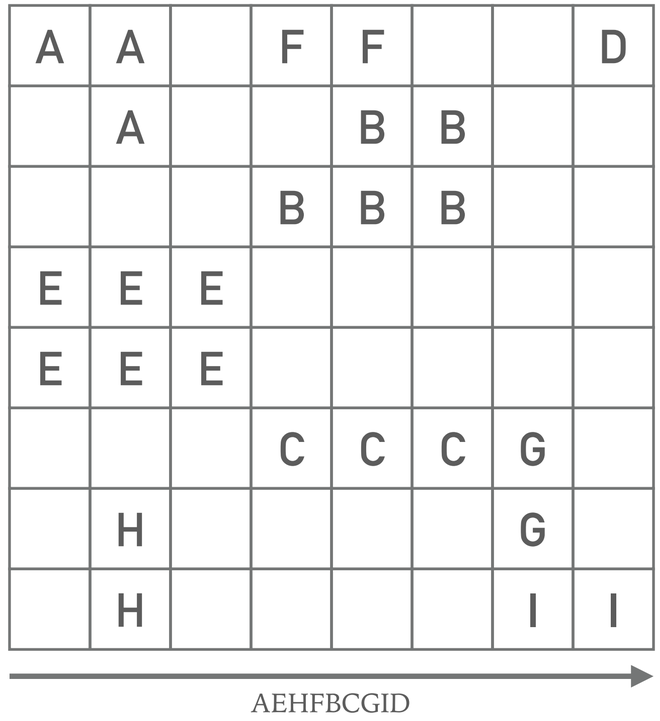
- Split the following overflown node with R*Tree split method
  - The maximum number of items in a node is $M = 8$
  - The minimum number of items in a node is $m = 3$



- I.e., execute the following methods:
  - ChooseSplitAxis
  - Distribute

EXAMPLE 3: SPLITTING IN R* TREE

- 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 minimises the sum



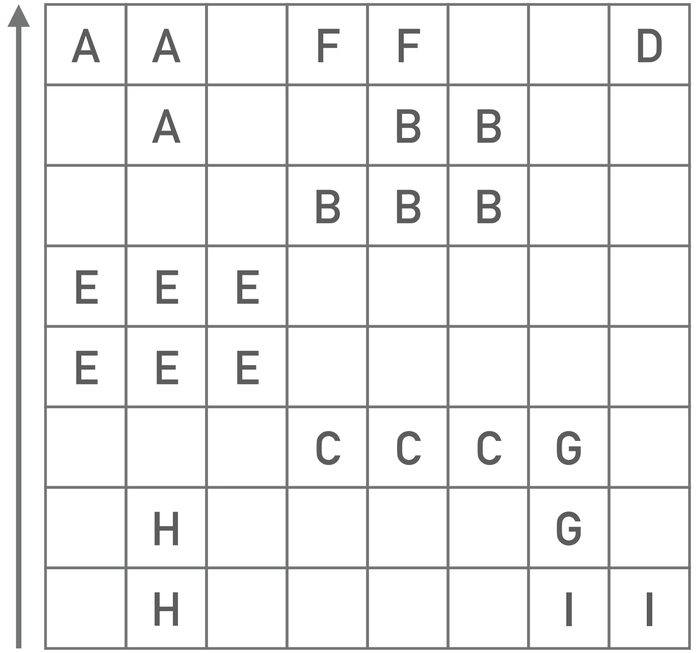
- Ordering* based on the x-axis: AEHFBCGID
  - margin-value (AEH||FBCGID) = (3+8)*2+(5+8)*2=22+26=48
  - margin-value (AEHF||BCGID) = (5+8)*2+(5+8)*2=26+26=52
  - margin-value (AEHFB||CGID) = (6+8)*2+(5+8)*2=28+26=54
  - margin-value (AEHFBC||GID) = (6+8)*2+(2+8)*2=28+20=48
  - Sum = 48+52+54+48=202

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

EXAMPLE 3: SPLITTING IN R* TREE

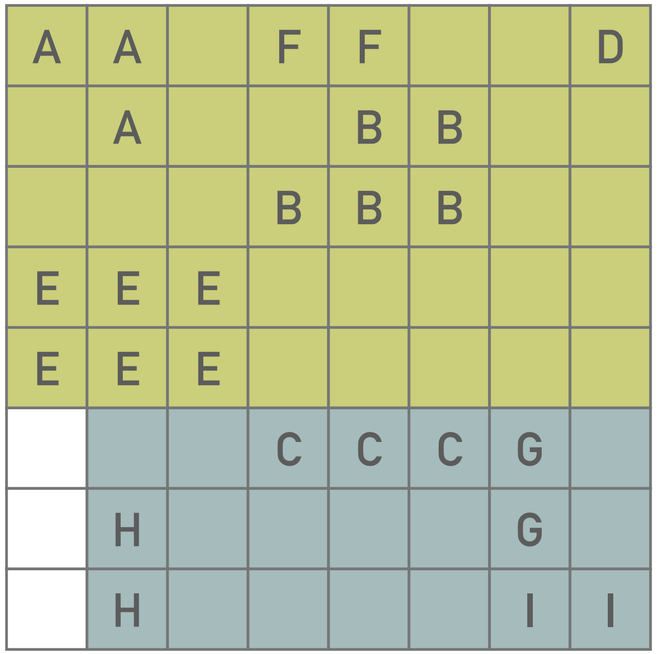
- Ordering* based on the y-axis: IHGCEBAFD
  - margin-value (IHG||CEBAFD) = (8+3)*2+(8+6)*2=22+28=50
  - margin-value (IHGC||EBAFD) = (8+3)*2+(8+5)*2=22+26=48
  - margin-value (IHGCE||BAFD) = (8+5)*2+(8+3)*2=26+22=48
  - margin-value (IHGCEB||AFD) = (8+7)*2+(8+2)*2=30+20=50
  - Sum = 50+48+48+50 = 196

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



EXAMPLE 3: SPLITTING IN R* TREE

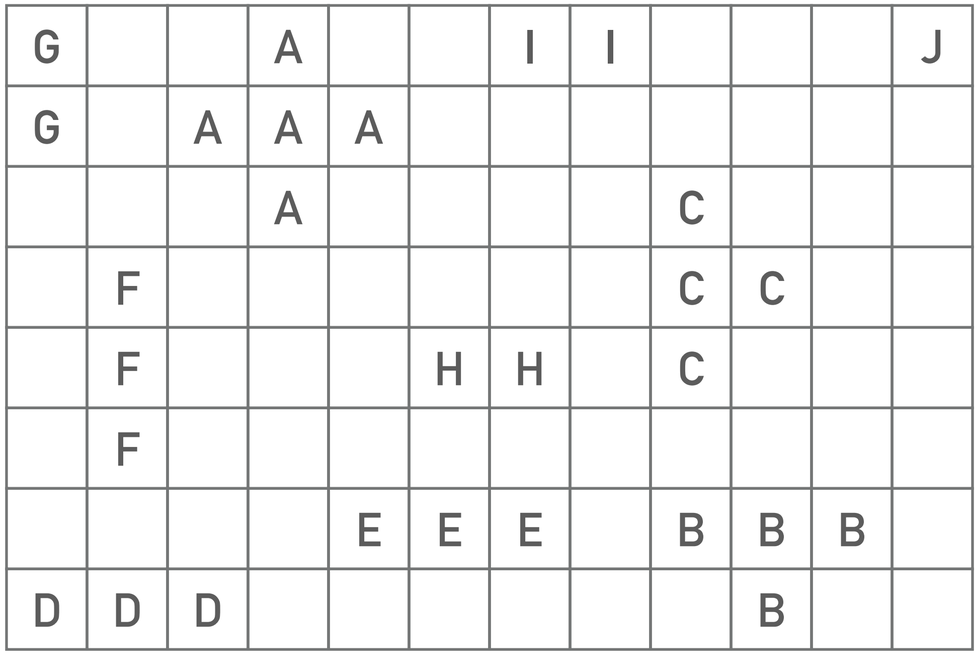
- Now we compute the overlap-values among all the distributions (along y-axis) and pick the distributions minimising the overlap
  - overlap-value (IHG||CEBAFD) = 8 (row CCCG)
  - overlap-value (IHGC||EBAFD) = 0
  - overlap-value (IHGCE||BAFD) = 0
  - overlap-value (IHGCEB||AFD) = 8 (row ABB)



- If more distributions lead to the minimum overlap, the one is chosen which shows the smallest 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 4

- 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 $m = 3$



- I.e., execute the following methods:
  - ChooseSplitAxis
  - Distribute

  - Illustrate the result