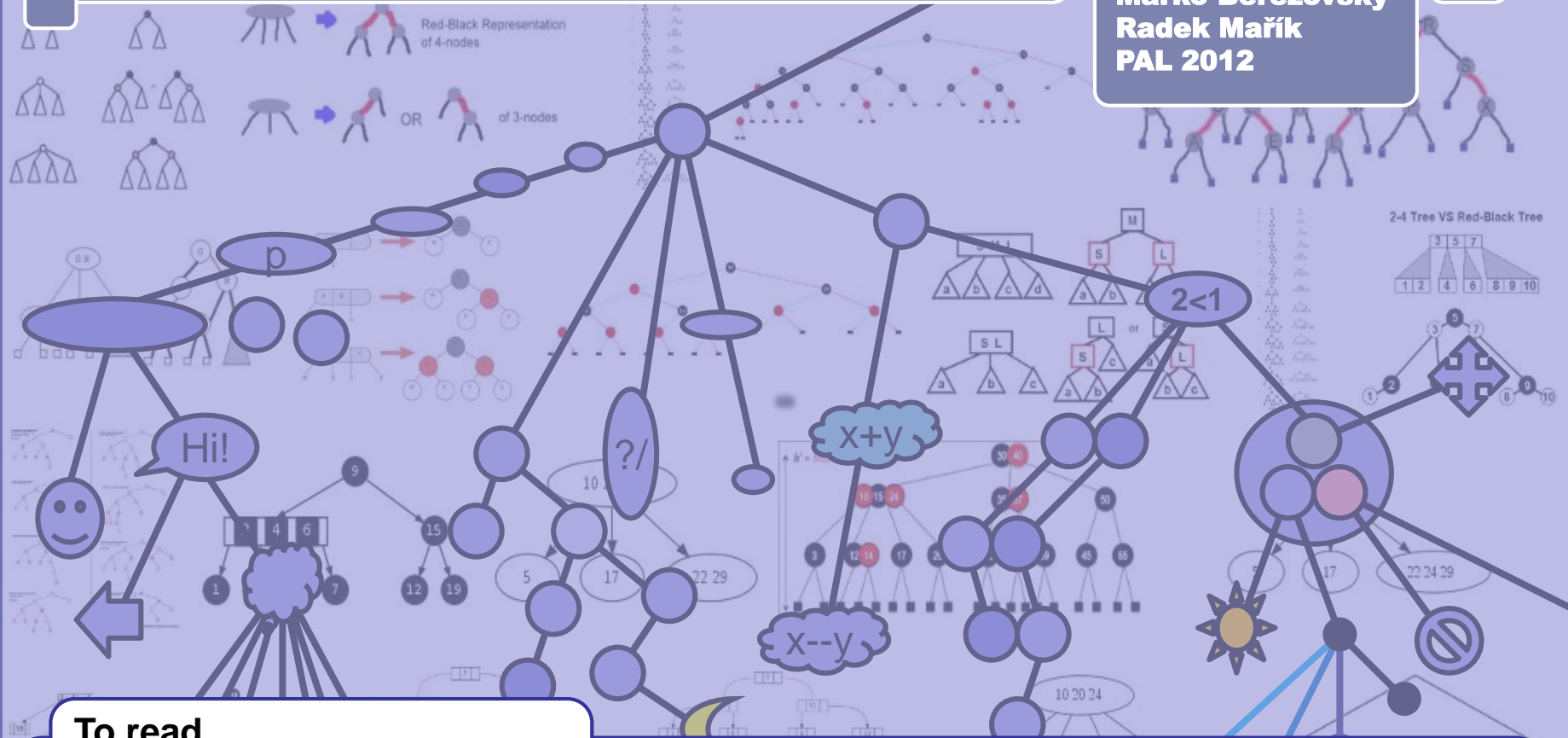


# Search trees, k-d tree

Marko Berezovský  
Radek Mařík  
PAL 2012



## To read

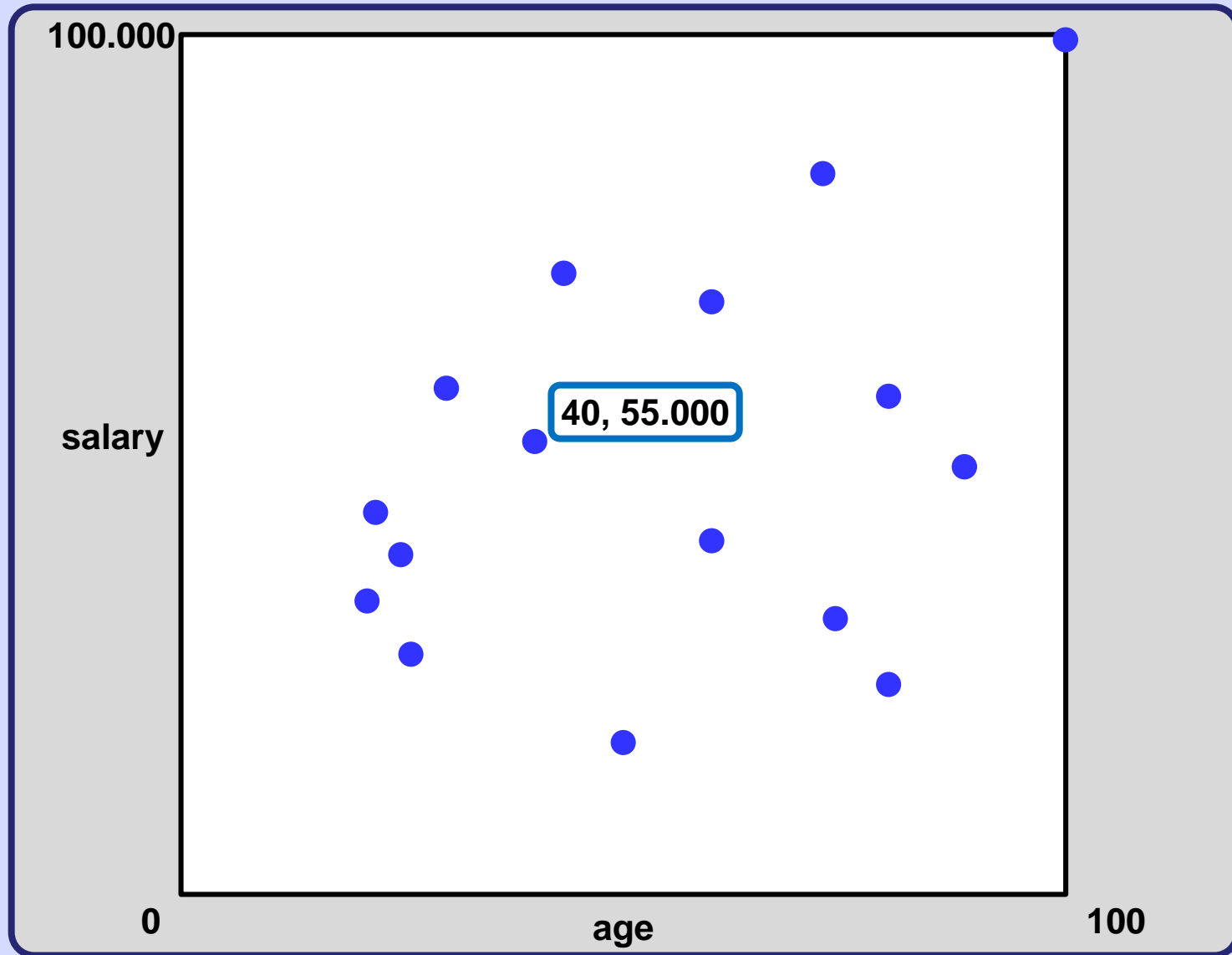
Dave Mount: *CMSC 420: Data Structures1 Spring 2001*, Lessons 17&18.

<http://www.cs.umd.edu/~mount/420/Lects/420lects.pdf>

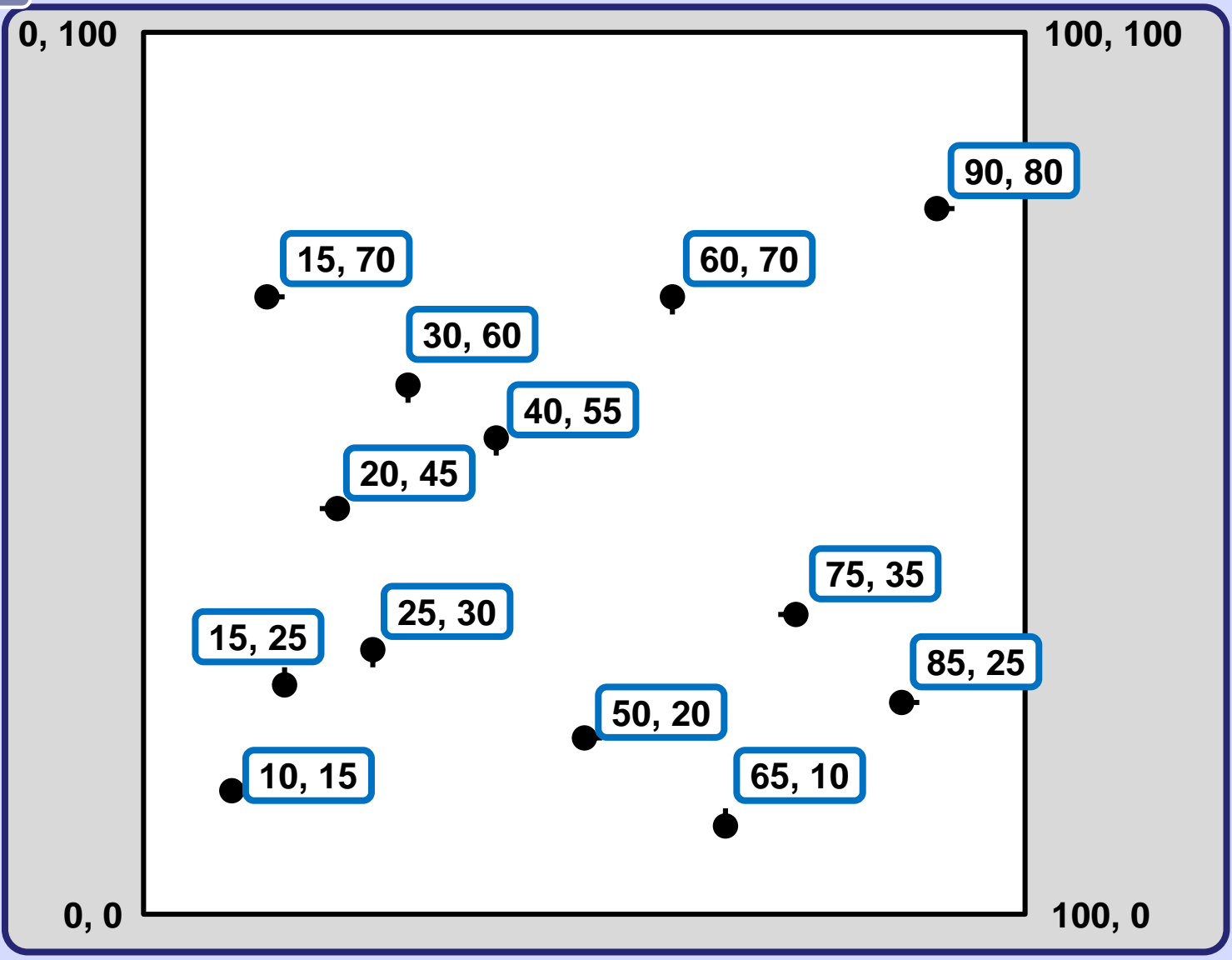
Hanan Samet: *Foundations of multidimensional and metric data structures*, Elsevier, 2006, chapter 1.5.

<http://www.amazon.com/Foundations-Multidimensional-Structures-Kaufmann-Computer/dp/0123694469>

See PAL webpage for references

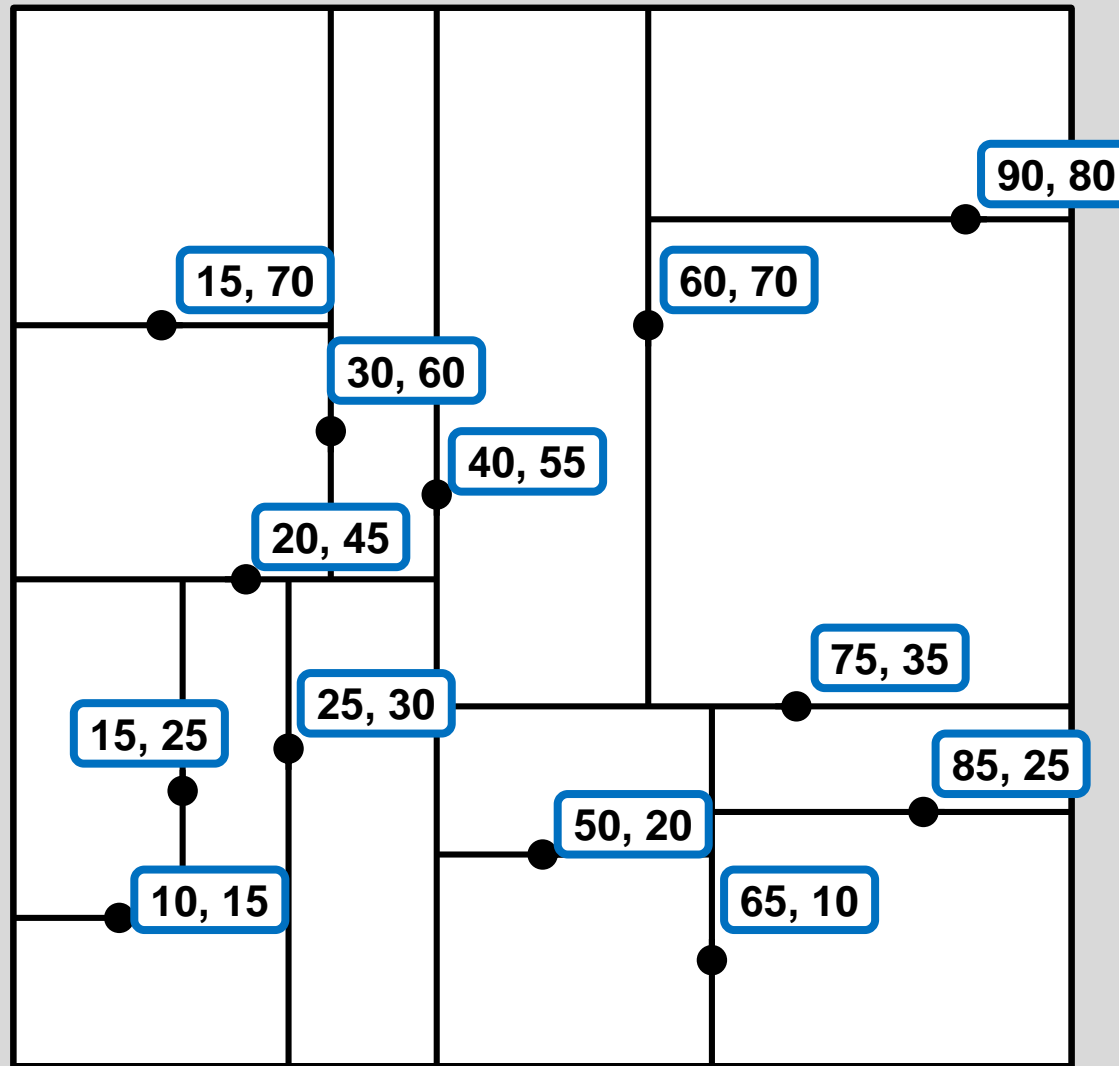


Points represent employees, coordinates are derived based on age and salary.



- 40, 55
- 20, 45
- 75, 35
- 60, 70
- 30, 60
- 65, 10
- 50, 20
- 15, 70
- 85, 25
- 90, 80
- 15, 25
- 10, 15
- 25, 30

Points in plane in general position are given, suppose no two are identical.



Scheme of area division exploited in k-d tree.

K-d tree is a binary search tree representing a rectangular area in D-dimensional space. The area is divided (and recursively subdivided) into **rectangular cells**. Denote dimensions naturally by their index 0, 1, 2, ... D-1. Denote by R the root of a tree or a subtree.

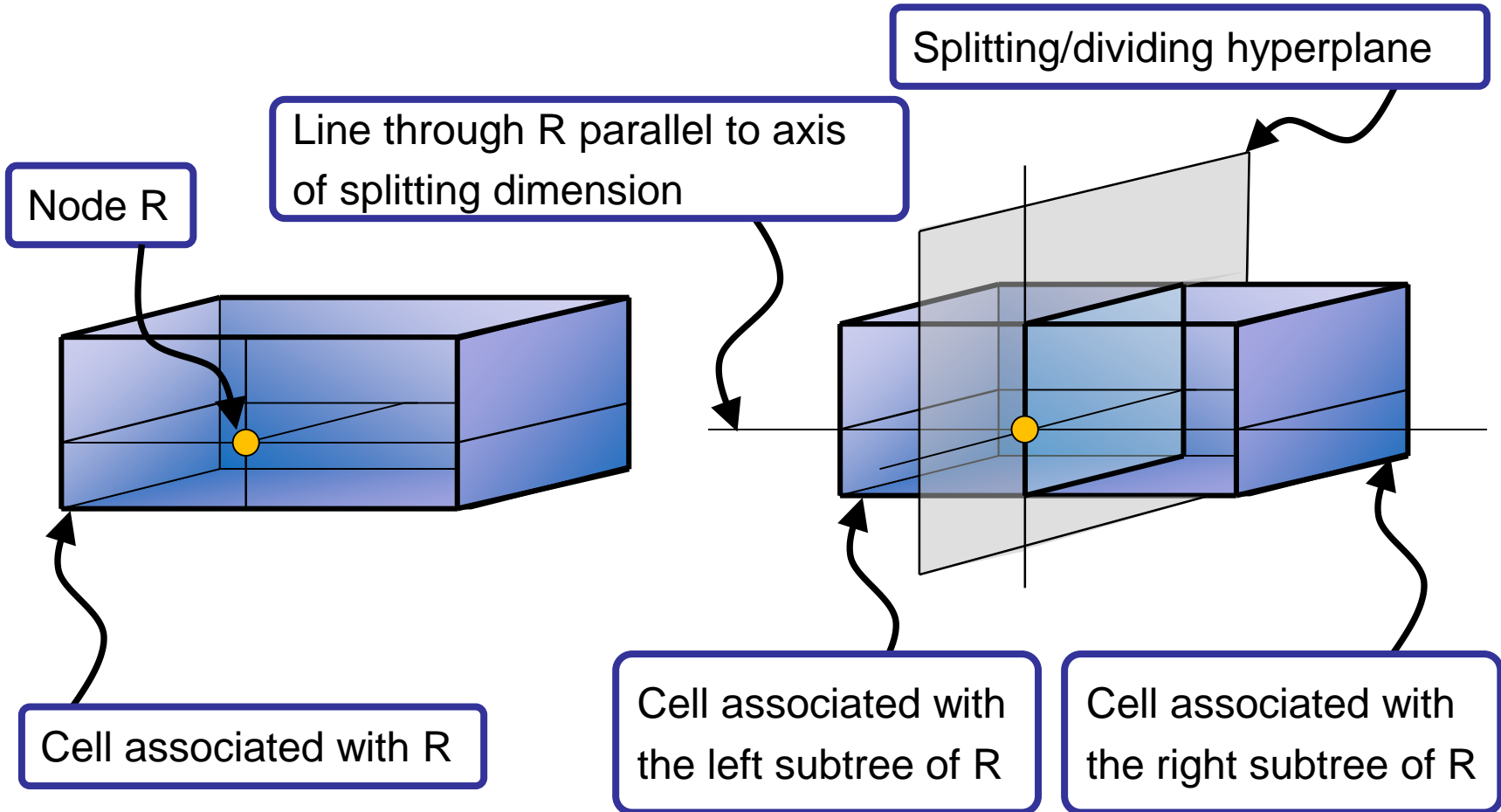
A rectangular D-dimensional cell  $C(R)$  (hyperrectangle) is associated with R. Let R coordinates be  $R[0], R[2], \dots, R[D-1]$  and let h be its depth in the tree. The cell  $C(R)$  is splitted into two subcells by such **hyperplane** of dim D-1, for whose all points y it holds:  $y[h\%D] = R[h\%D]$ .

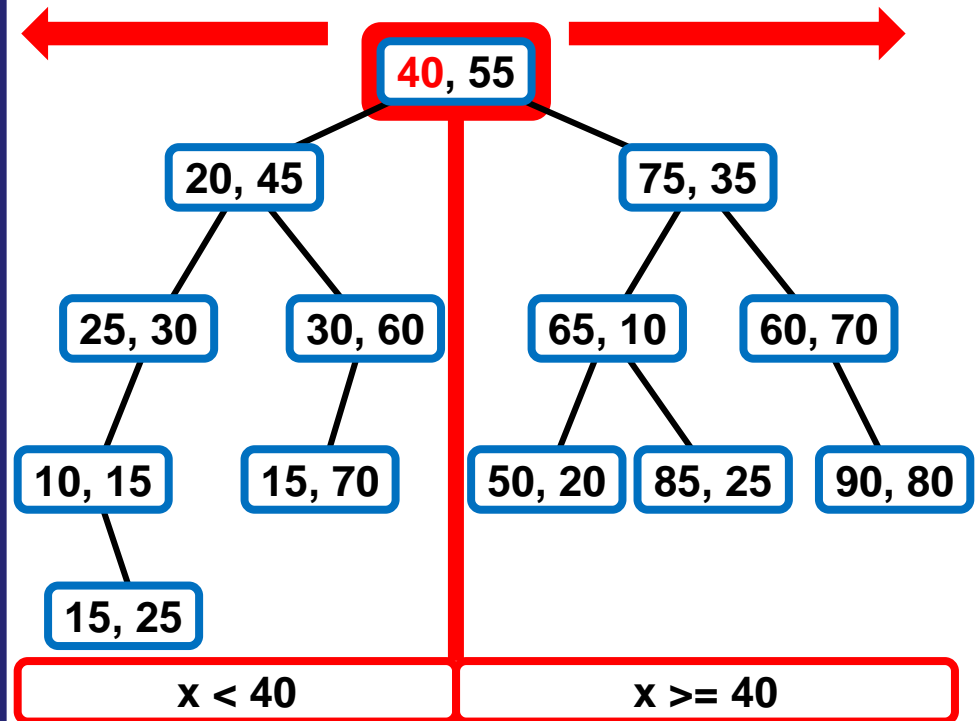
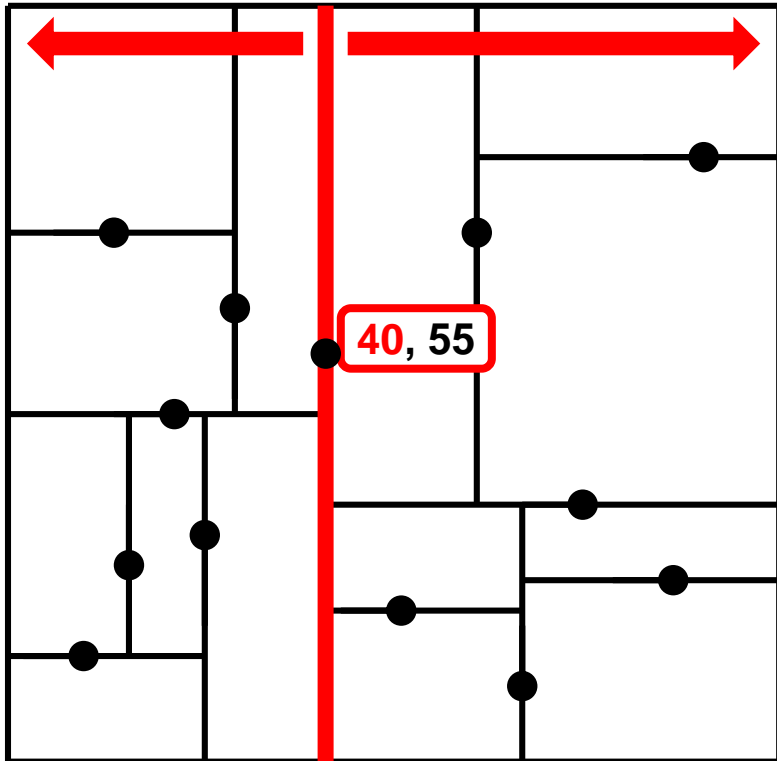
All nodes in the left subtree of R are characterised by their  $(h\%D)$ -th coordinate being **less than**  $R[h\%D]$ .

All nodes in the right subtree of R are characterised by their  $(h\%D)$ -th coordinate being **greater than or equal to**  $R[h\%D]$ .

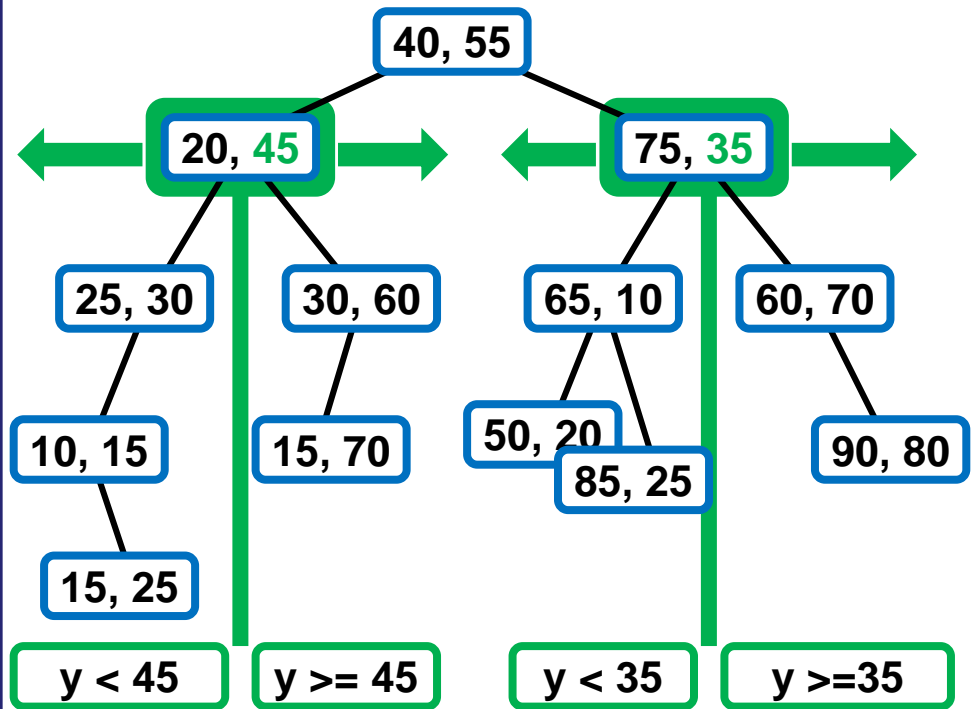
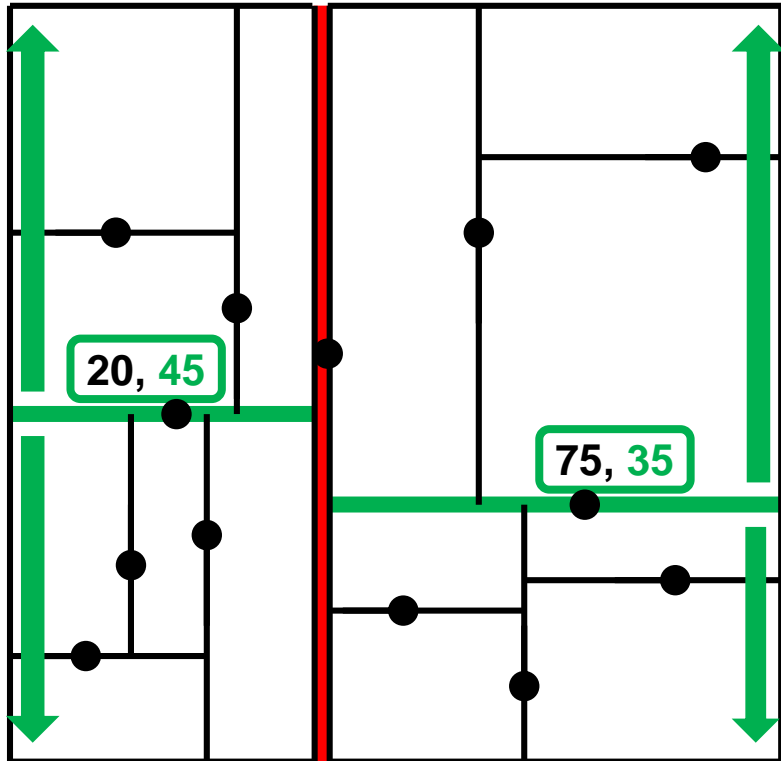
Let us call the value  $h\%D$  **splitting /cutting dimension** of a node in depth h.

Note that k-d tree presented here is a basic simple variant, many other, more sophisticated variants do exist.



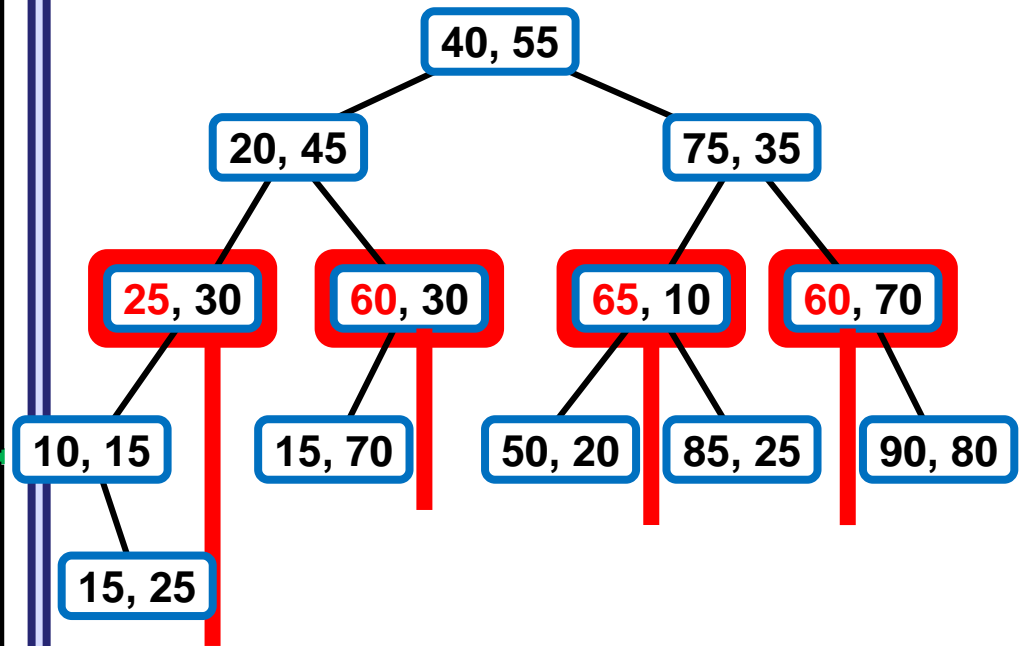
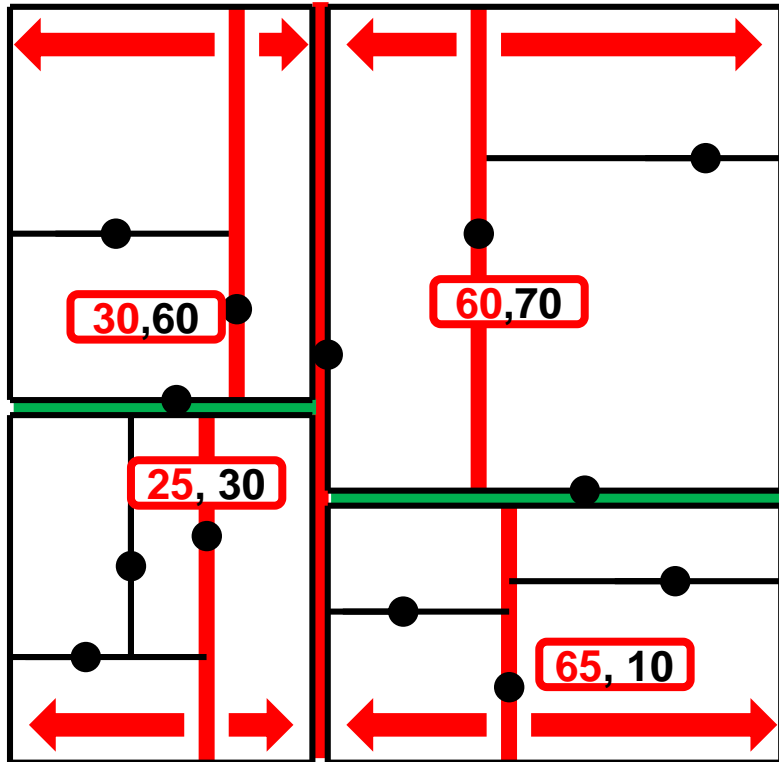


Scheme of area division exploited in k-d tree.

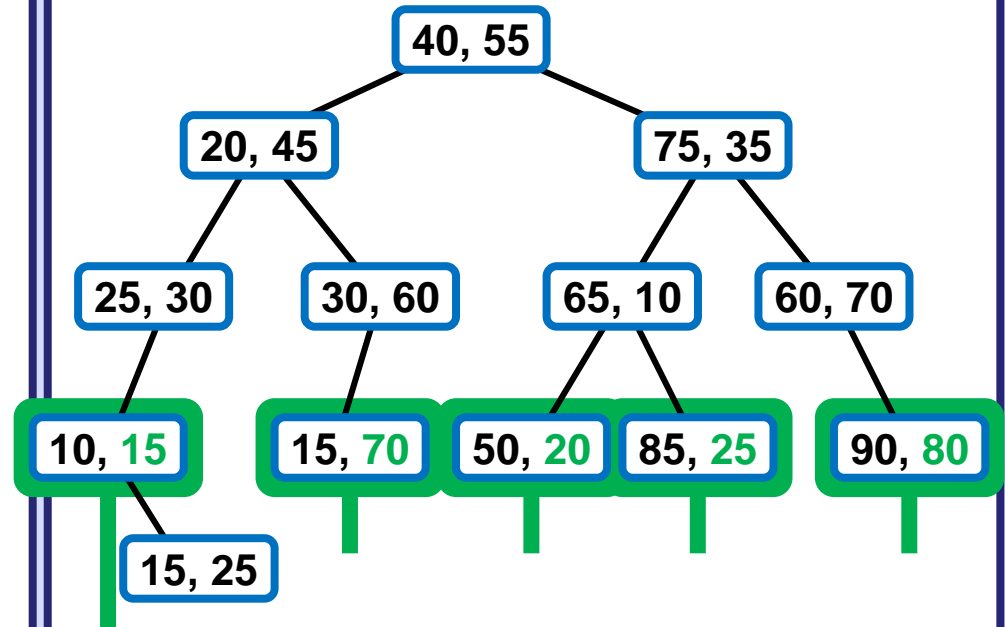
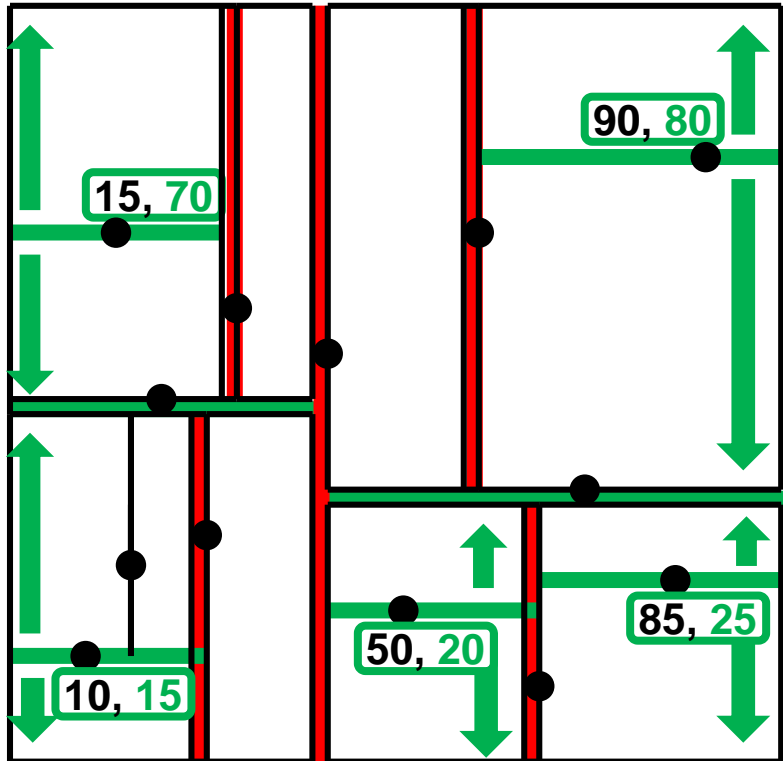


Scheme of area division exploited in k-d tree.

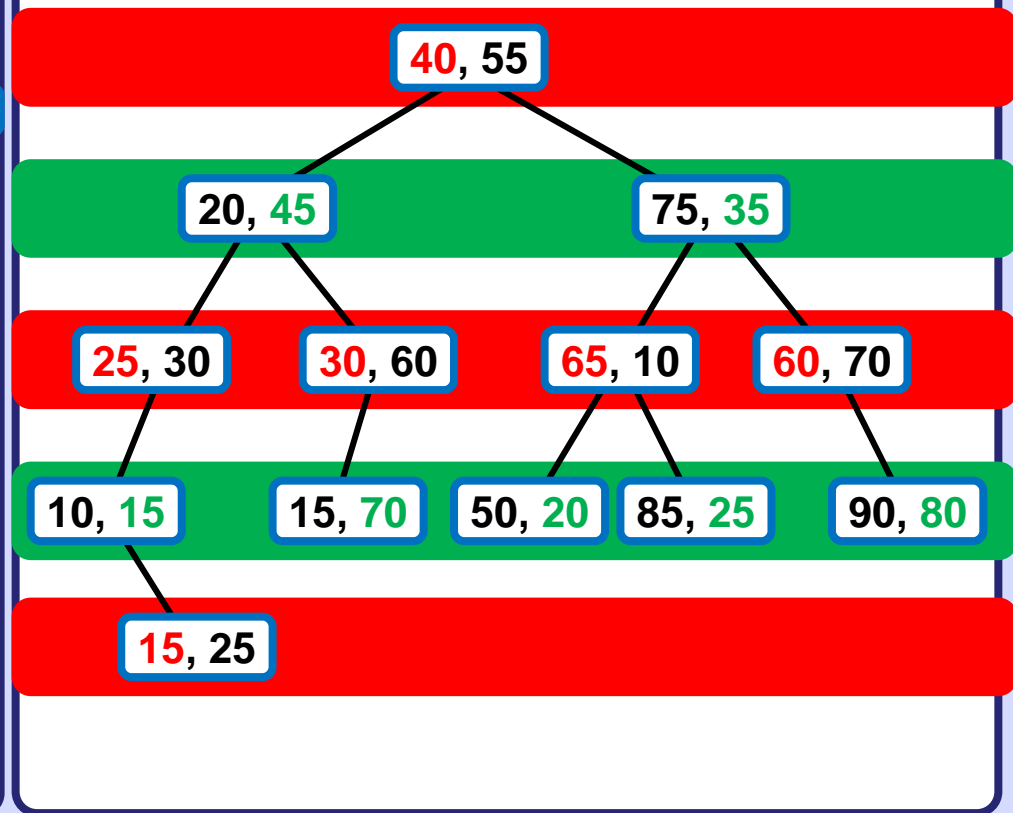
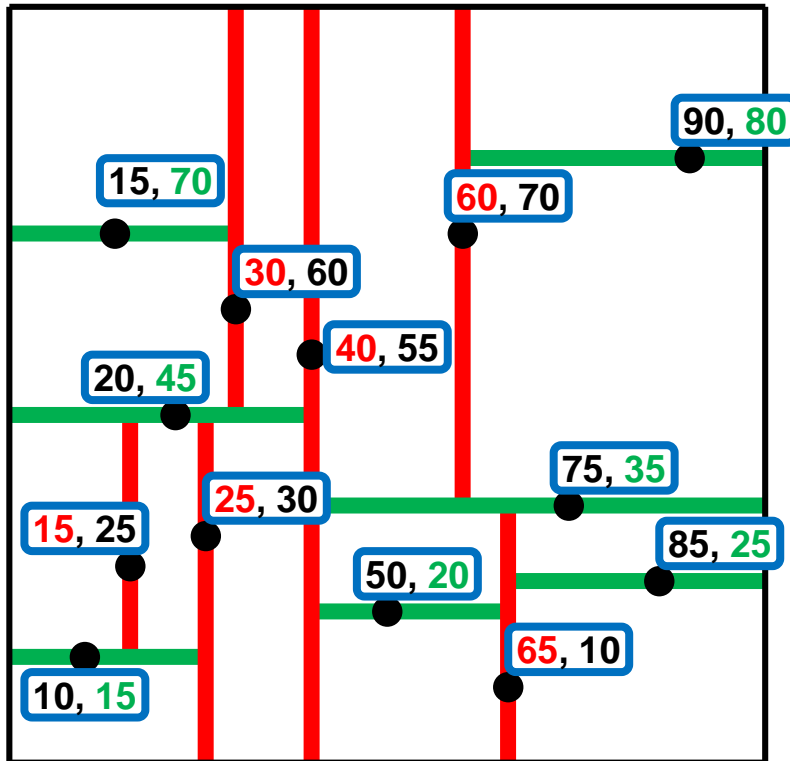




Scheme of area division exploited in k-d tree.



Scheme of area division exploited in k-d tree.



Complete k-d tree with with marked area division.

Operation **Find**(Q) is analogous to 1D trees.

Let

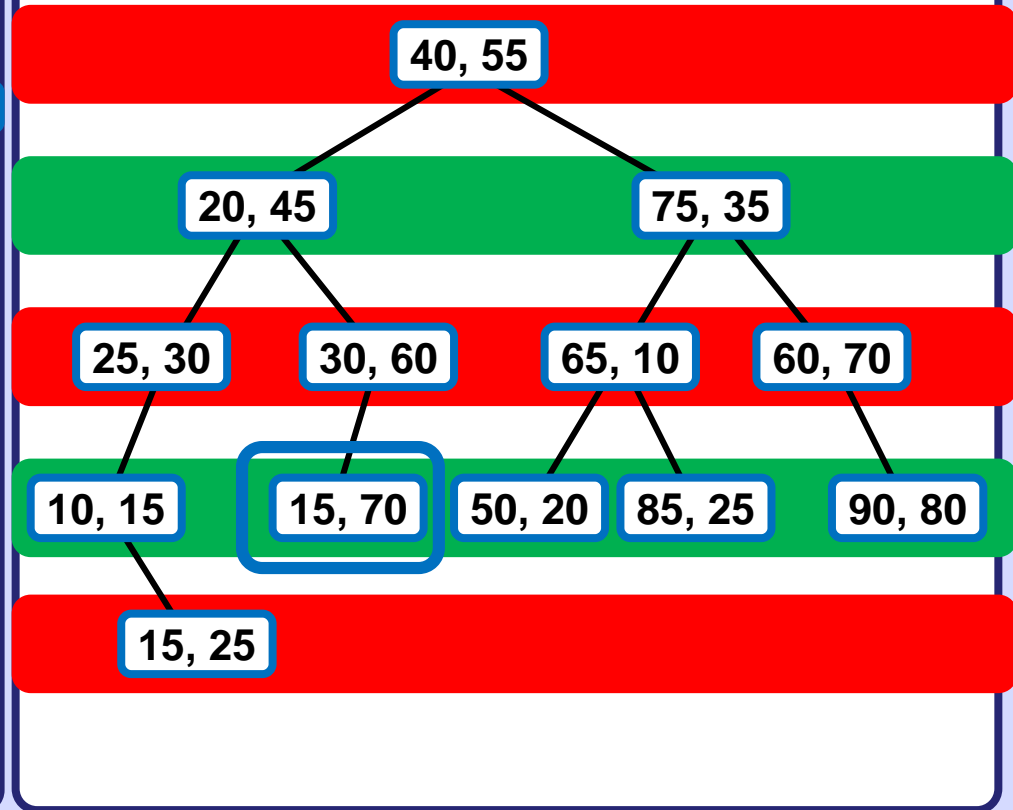
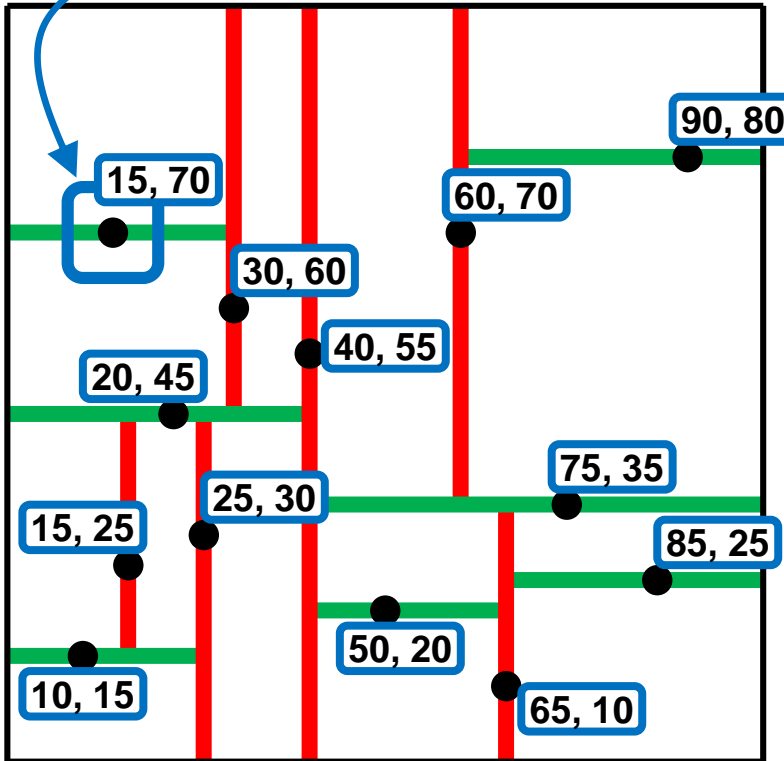
$Q[ ] = ( Q[0], Q[1], \dots, Q[D-1] )$  be the coordinates of the query point Q,  
 $N[ ] = ( N[0], N[1], \dots, N[D-1] )$  be the coordinates of the current node N,  
 $h = h(N)$  be the depth of current node N.

If  $Q[ ] == N[ ]$  stop, Q was found

if  $Q[h\%D] < N[h\%D]$  continue search recursively in left subtree of N.

if  $Q[h\%D] \geq N[h\%D]$  continue search recursively in right subtree of N.

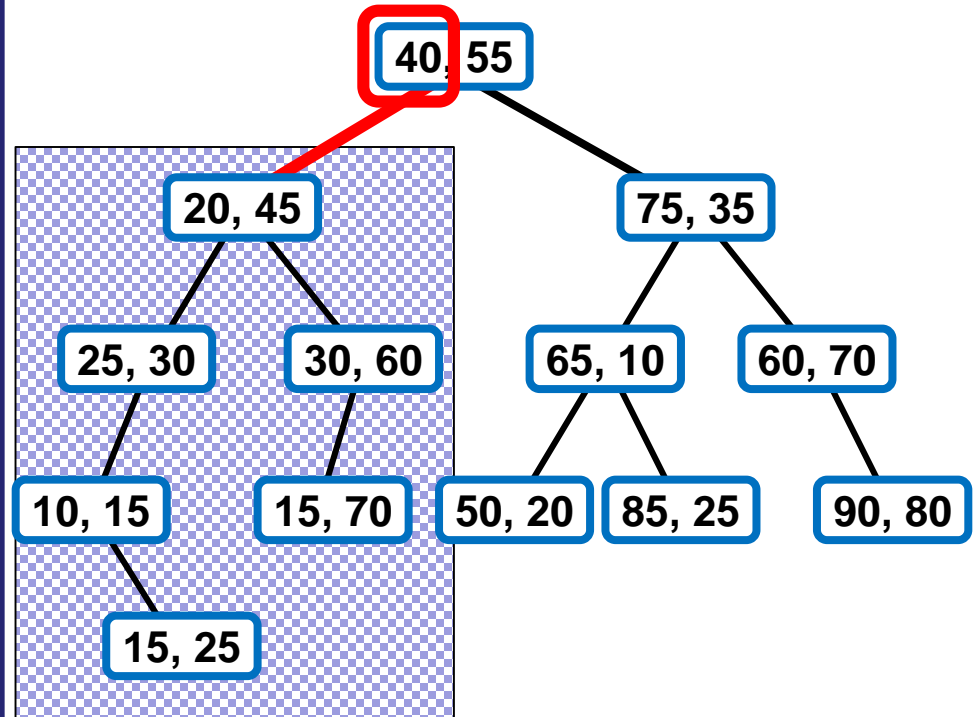
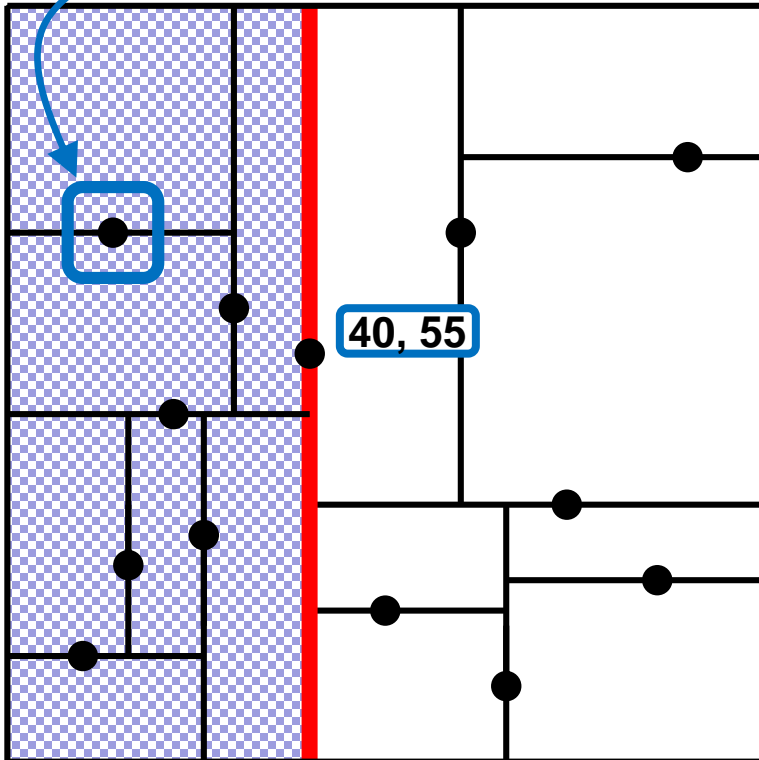
Find [15, 70]



Operation Find works analogously as in other (1D) trees.  
 Note how cutting dimension along which the tree is searched alternates regularly with the depth of the currently visited node .

Find **[15, 70]**

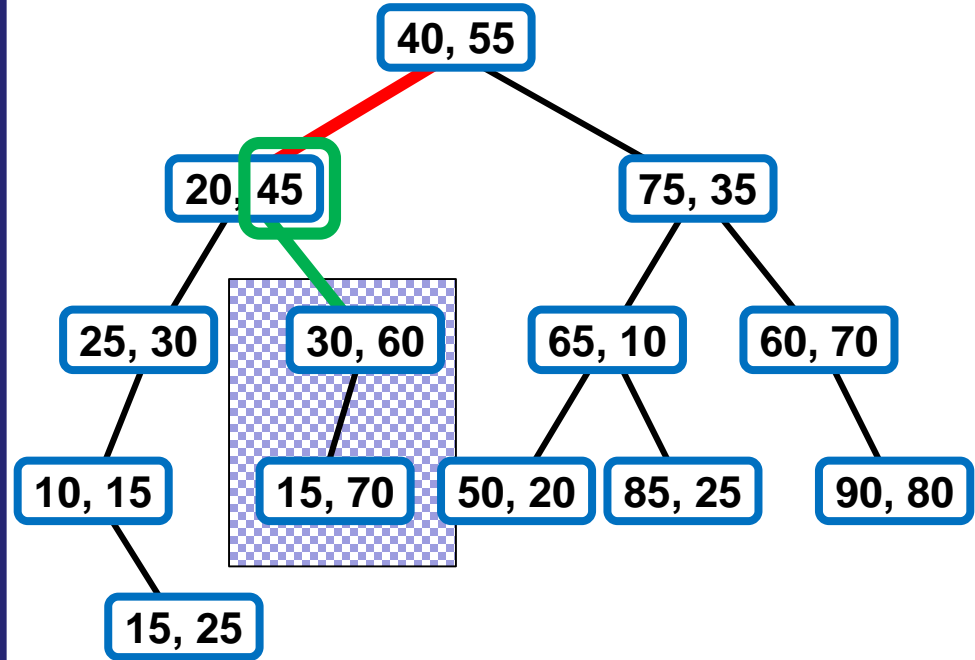
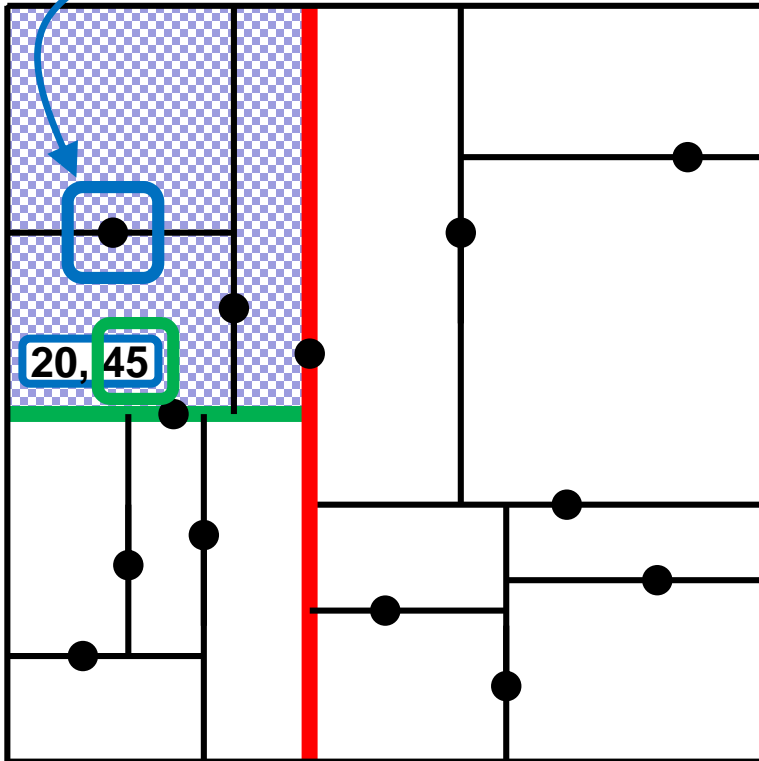
**15 < 40**



$Q = [15, 70]$ ,  $N = [40, 55]$ ,  $Q \neq N$ ,  $h(N) = 0$ . Compare **x-coordinate** of searched key  $Q$  to **x-coordinate** of the current node  $N$  and continue search accordingly in the left or in the right subtree of  $N$ .

Find [15, 70]

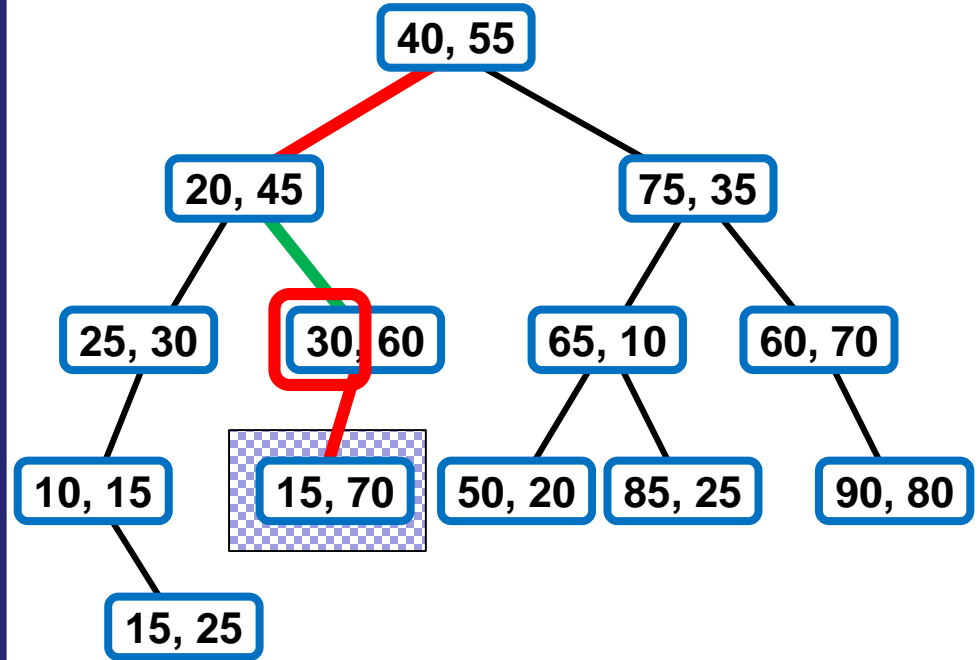
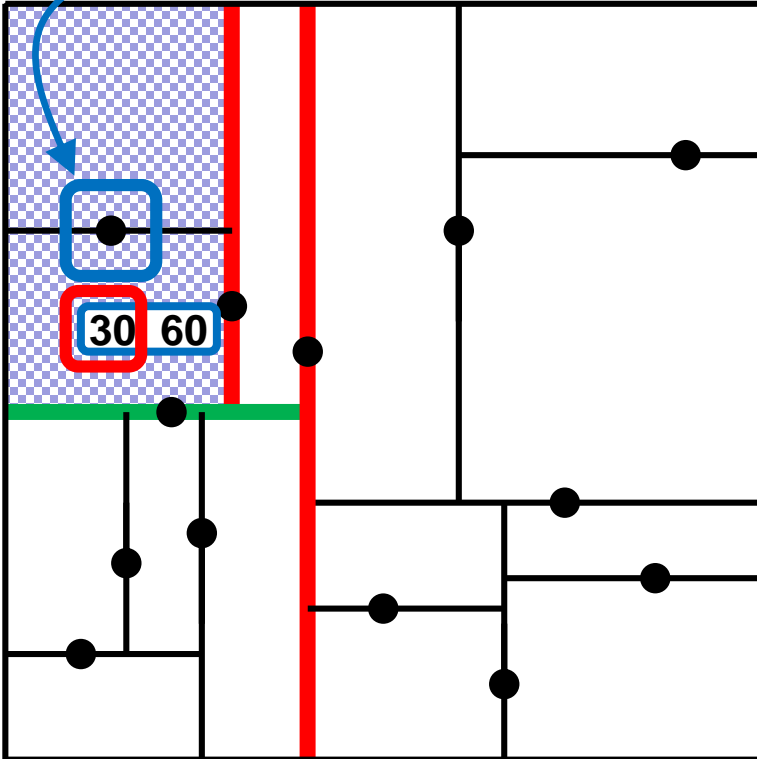
70 >= 45



$Q = [15, 70]$ ,  $N = [20, 45]$ ,  $Q \neq N$ ,  $h(N) = 1$ . Compare **y-coordinate** of searched key  $Q$  to **y-coordinate** of the current node  $N$  and continue search accordingly in the left or in the right subtree of  $N$ .

Find [15, 70]

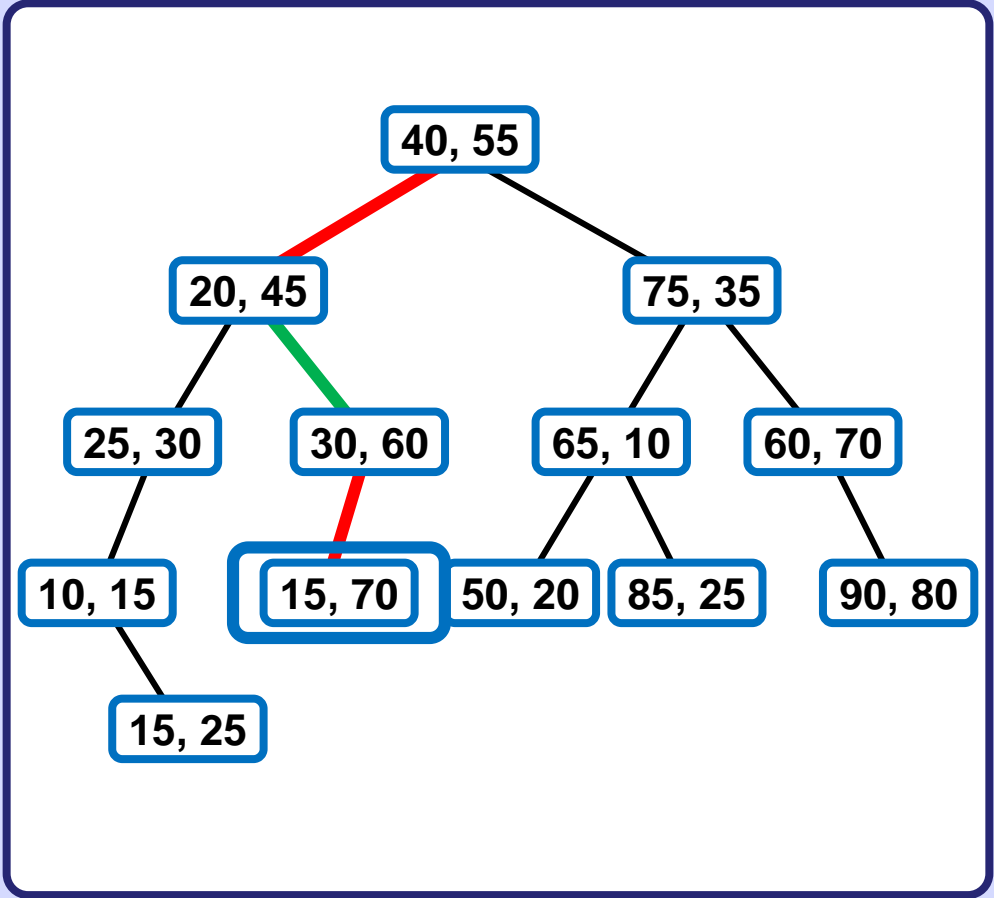
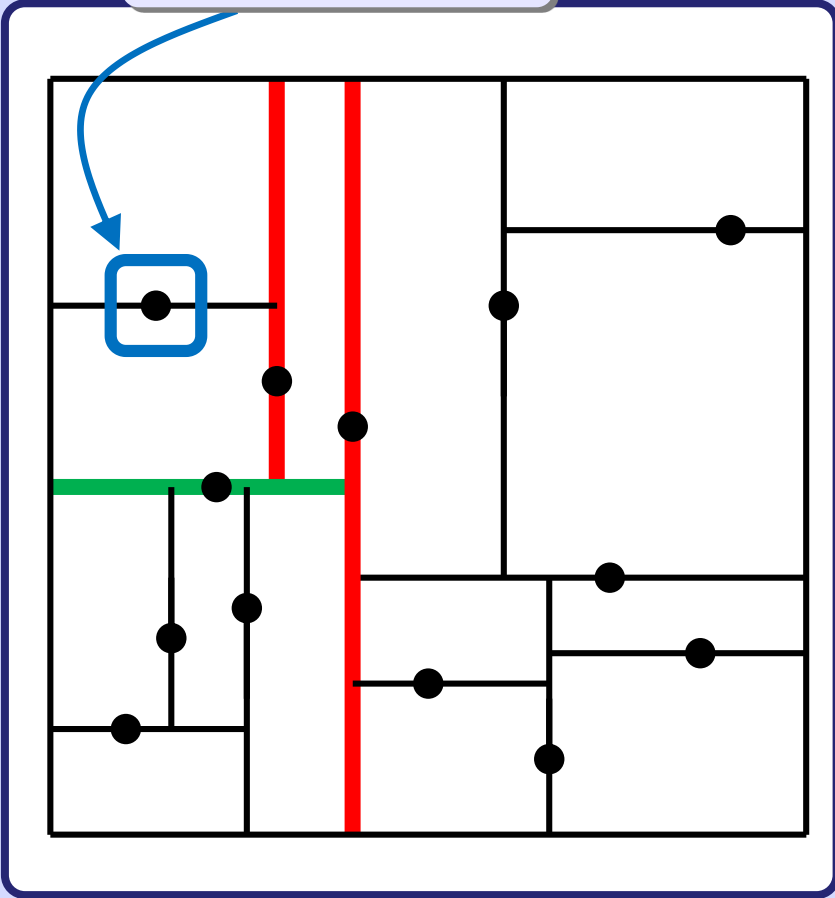
15 < 30



$Q = [15, 70]$ ,  $N = [30, 60]$ ,  $Q \neq N$ ,  $h(N) = 0$ . Compare **x-coordinate** of searched key  $Q$  to **x-coordinate** of the current node  $N$  and continue search accordingly in the left or in the right subtree of  $N$ .



Found [15, 70]



$Q = [15, 70]$ ,  $N = [15, 70]$ , found.

Operation **Insert**(P) is analogous to 1D trees.

Let  $P[ ] = ( P[0], P[1], \dots, P[D-1] )$  be the coordinates of the inserted point P.  
Perform search for P in the tree.

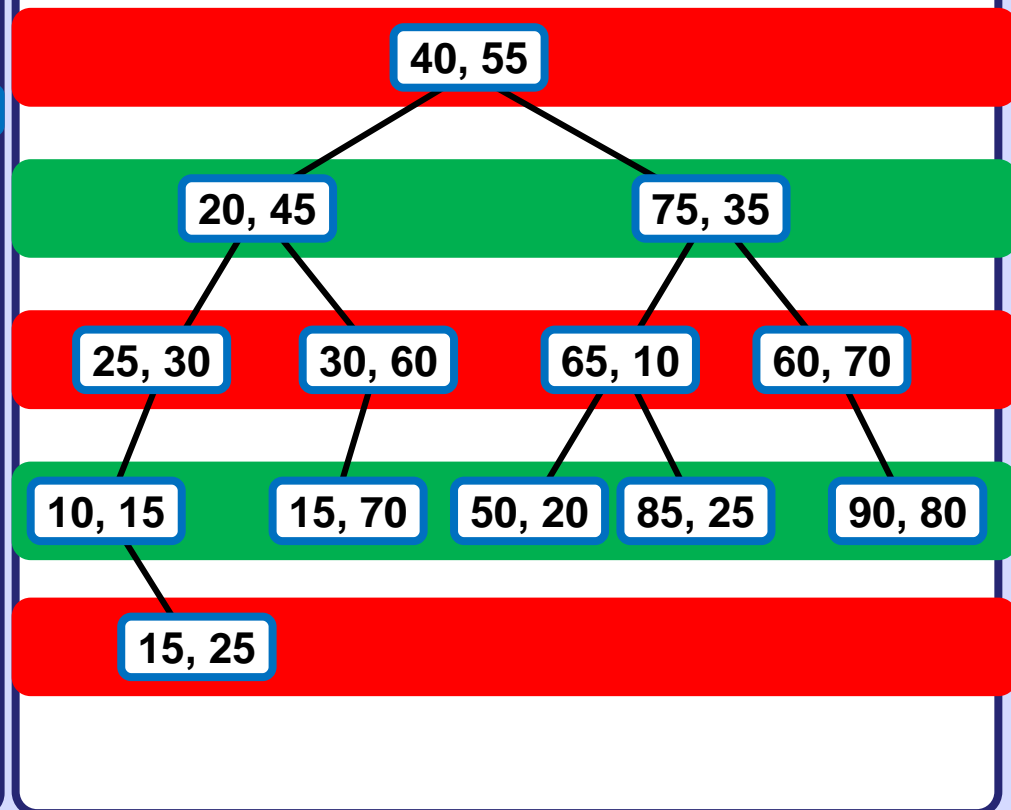
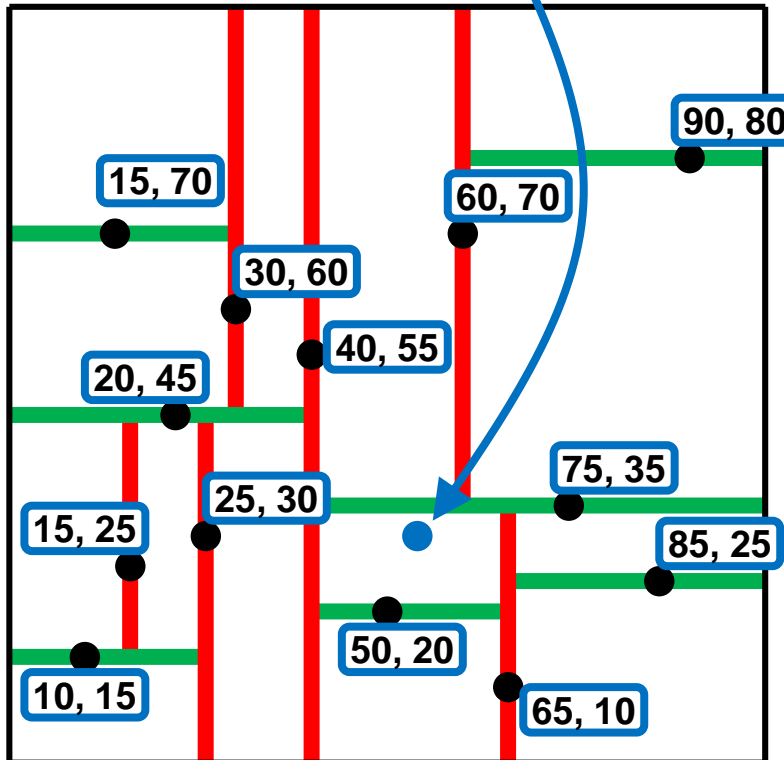
Let  $L[ ] = ( L[0], L[1], \dots, L[D-1] )$  be the coordinates of the leaf L which was the last node visited during the search. Let  $h = h(L)$  be the depth of L.

Create node N containing P as a key.

If  $P[h\%D] < L[h\%D]$  set N as the left child of L.

If  $P[h\%D] \geq L[h\%D]$  set N as the right child of L.

Insert [55, 30]

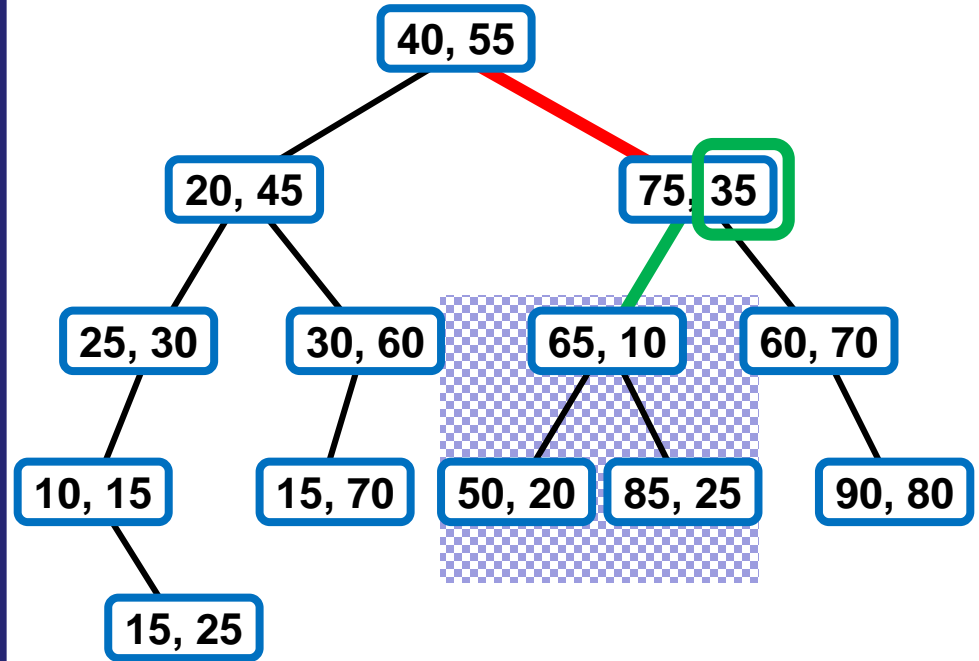
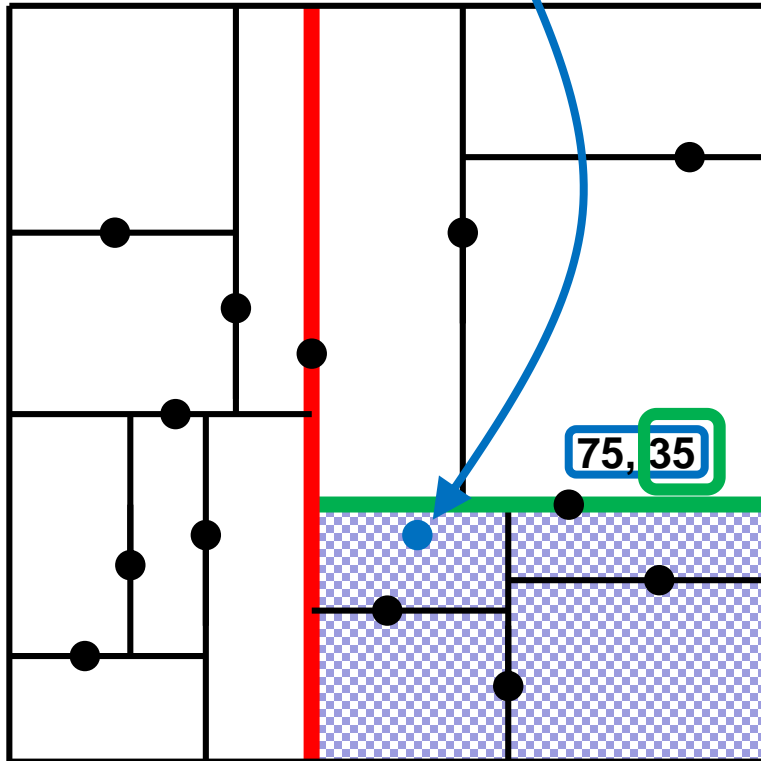


Operation Insert works analogously as in other (1D) trees.  
 Find the place for the new node under some of the leaves and insert node there.  
 Do not accept key which is identical to some other key already stored in the tree.



Insert [55, 30]

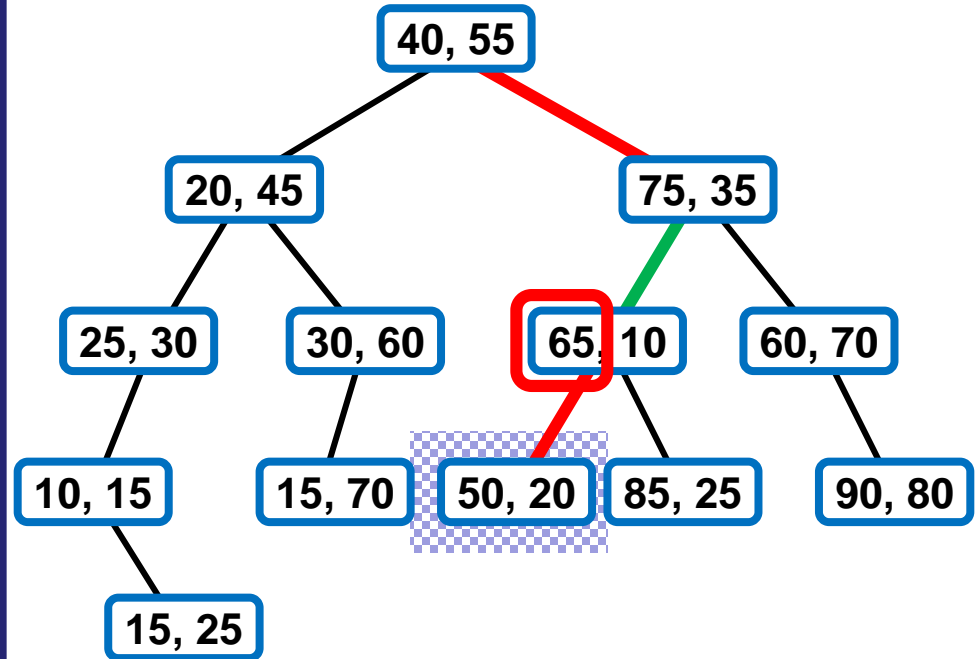
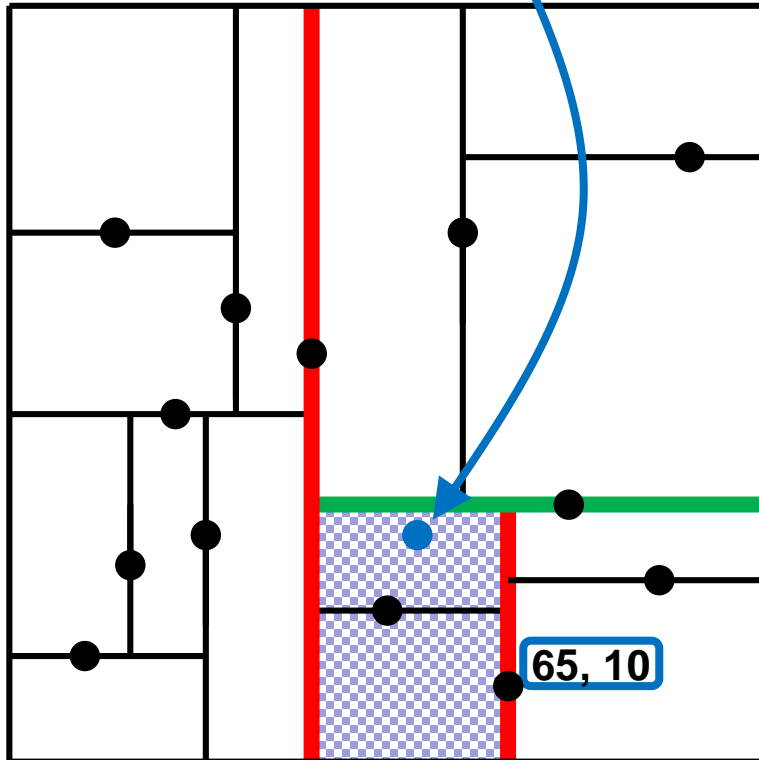
30 < 35



Operation Insert works analogously as in other (1D) trees.  
Searching for the place for the inserted key/node.

Insert [55 30]

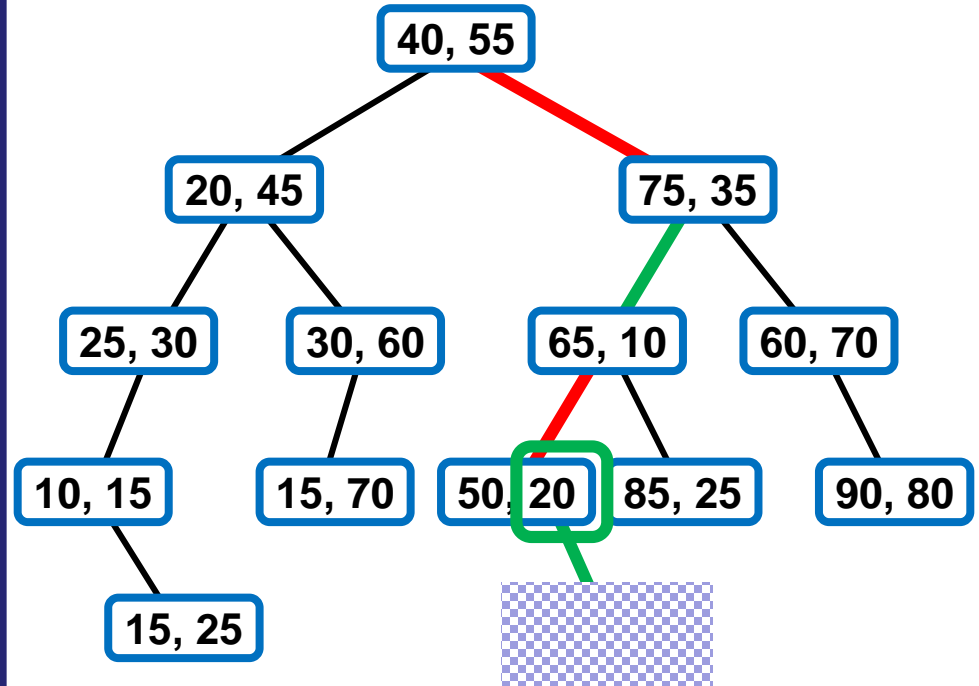
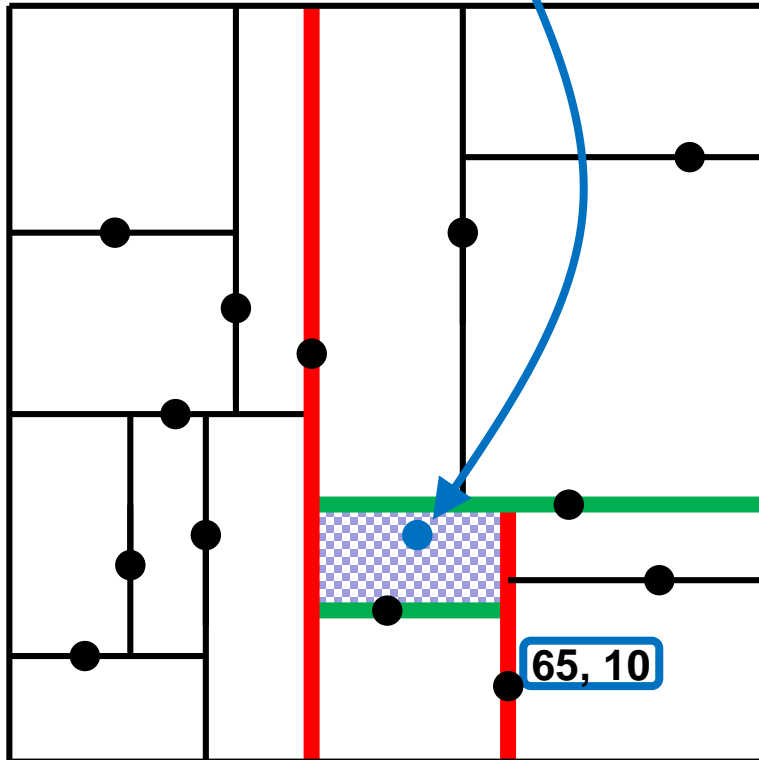
55 < 65



Operation Insert works analogously as in other (1D) trees.  
Searching for the place for the inserted key/node.

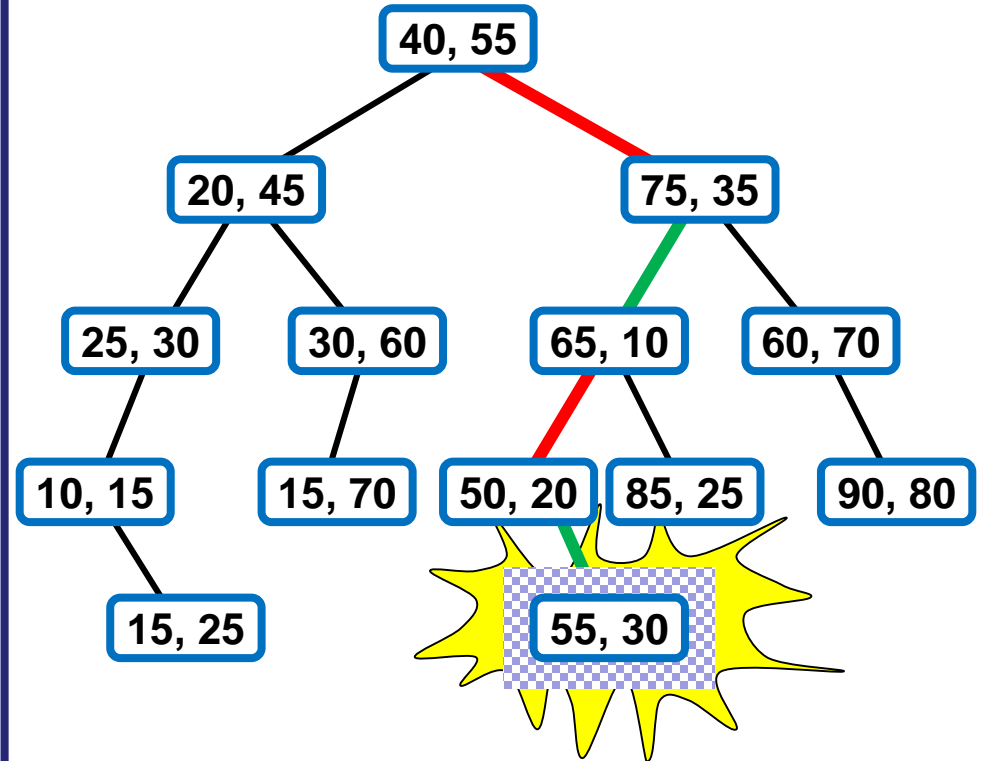
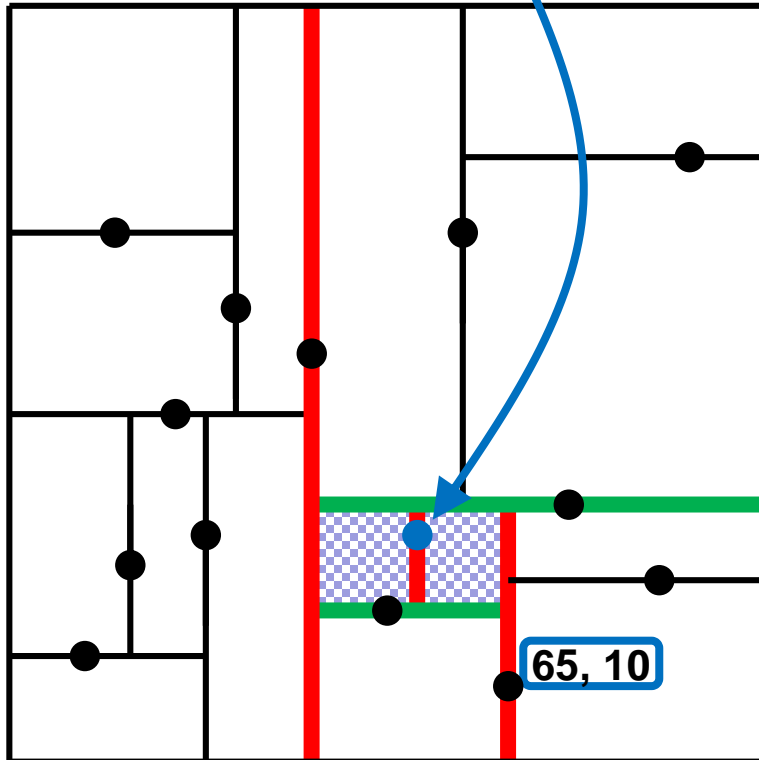
Insert [55, 30]

$30 \geq 20$



Operation Insert works analogously as in other (1D) trees.  
Searching for the place for the inserted key/node.

Inserted [55, 30]



Operation Insert works analogously as in other (1D) trees.  
 The place for the inserted key/node was found, the node/key was inserted.



```

// cd .. current dimension
Node Insert(Point P, Node N, Node parent, int cd) {
    if (N == null) // under a leaf
        N = new Node( P, parent );
    else if( P.coords.equals(N.coords) )
        throw new ExceptionDuplicatePoint();
    else if( P.coords[cd] < N.coords[cd] )
        N.left = insert( P, N.left, N, (cd+1)%D );
    else
        N.right = insert( P, N.right, N, (cd+1)%D );
    return N;
}
```

### Operation **FindMin(dim = k)**

Searching for a key whose k-th coordinate is minimal of all keys in the tree.

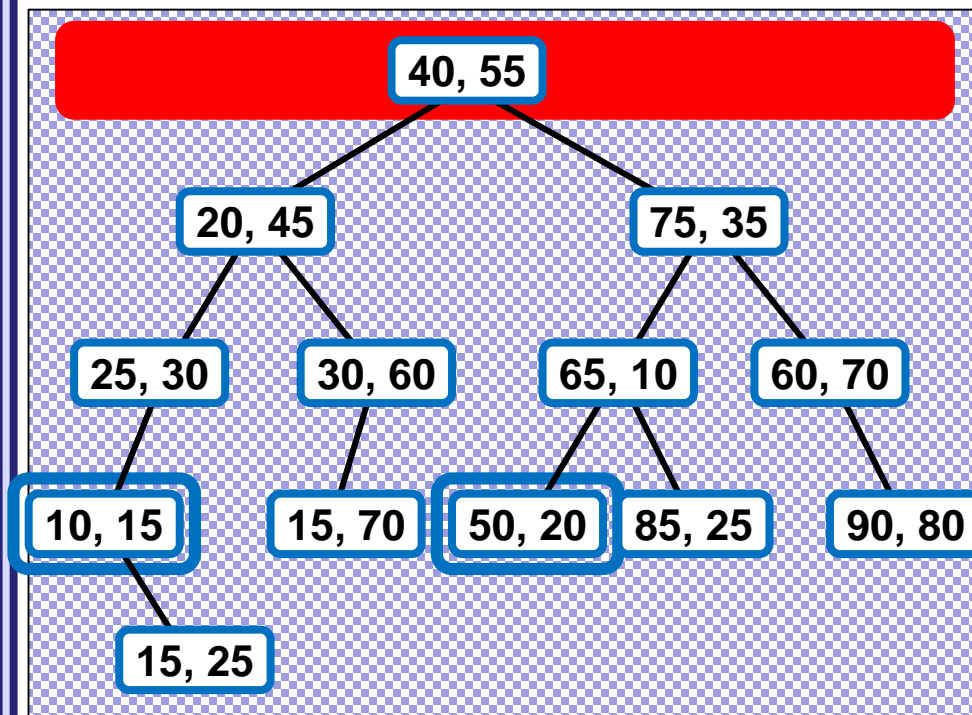
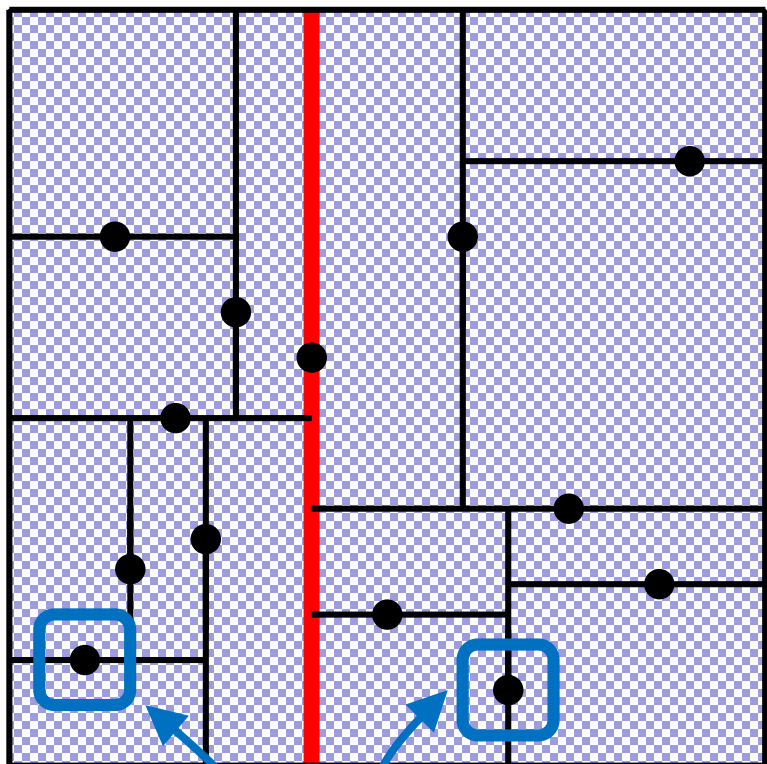
FindMin(dim = k) is performed as part of Delete operation.

The k-d tree offers no simple method of keeping track of the keys with minimum coordinates in any dimension because Delete operation may often significantly change the structure of the tree.

FindMin(dim = k) is the most costly operation, with complexity  $O(n^{1-1/d})$ , in a k-d tree with n nodes and dimension d.

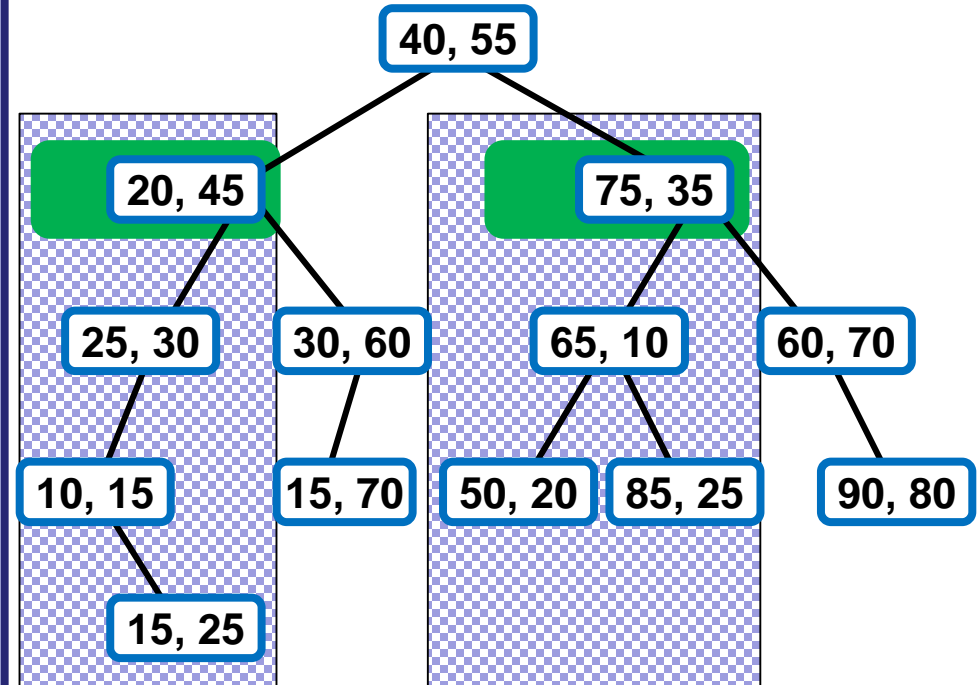
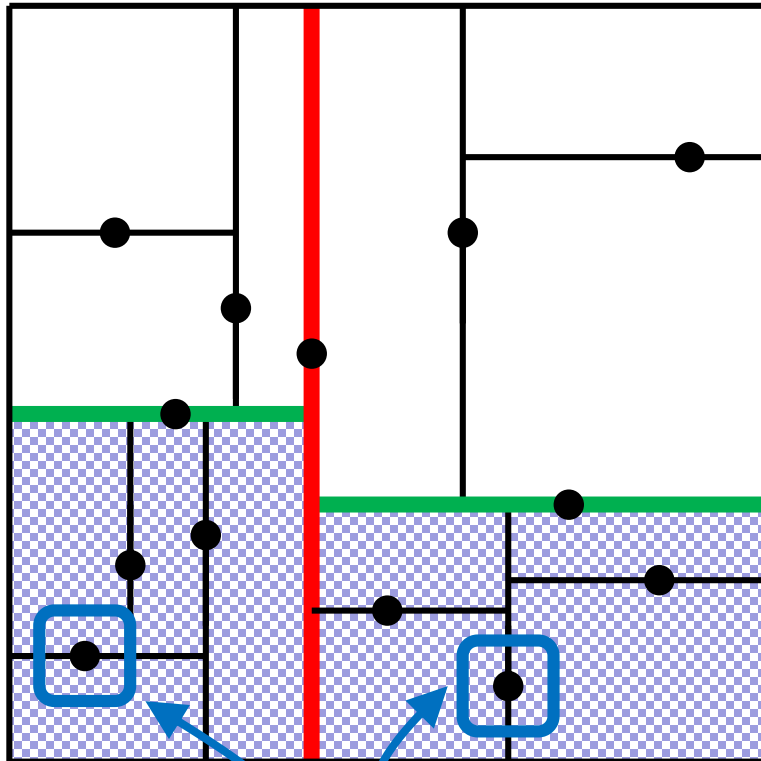
When  $d = 2$  the complexity is  $O(n^{1/2})$ .

FindMin(dim = y)



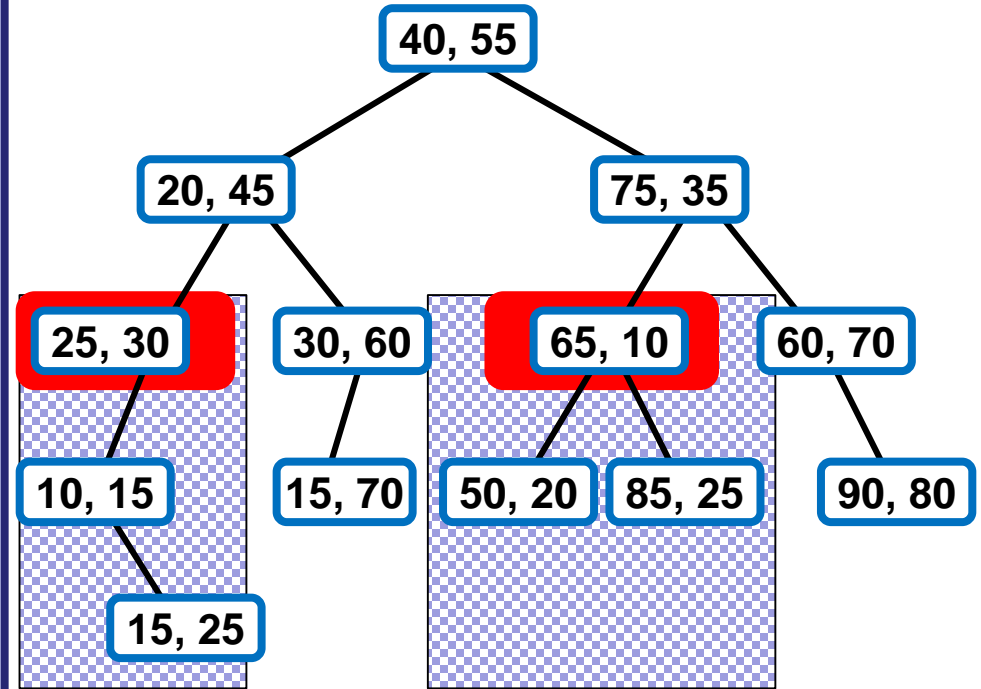
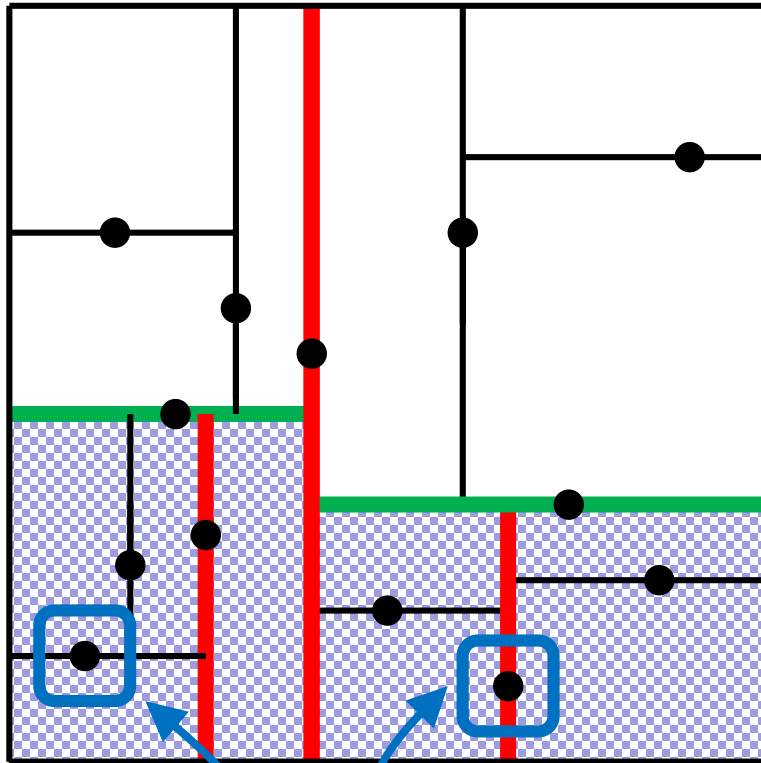
Node with minimal y-coordinate can be in L or R subtree of a node N corresponding to cutting dimension other than y, thus both subtrees of N (including N) must be searched.

FindMin(dim = y)



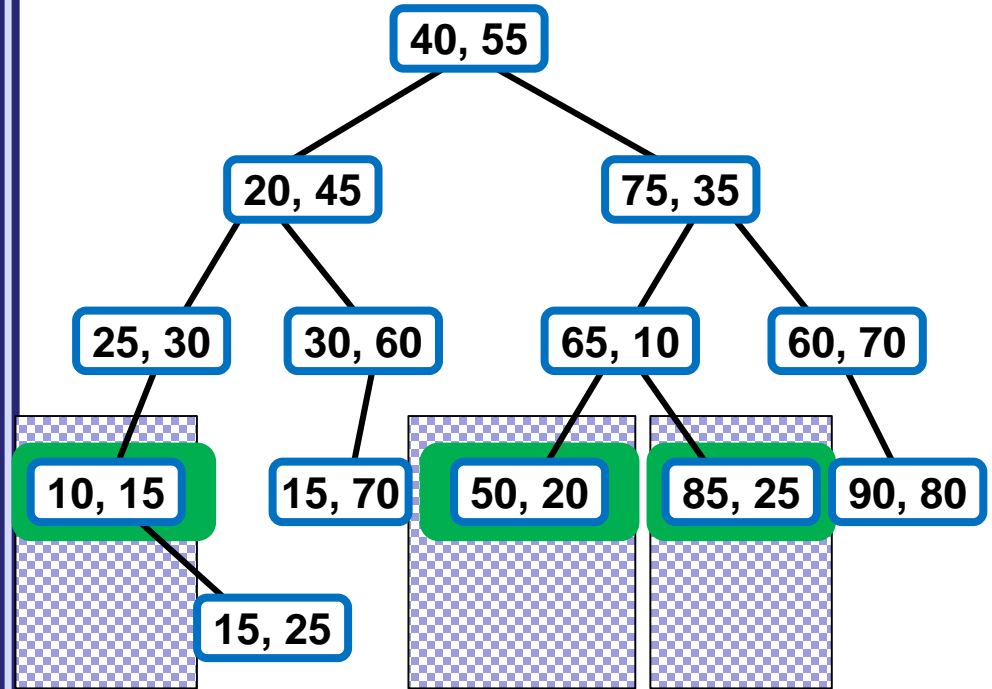
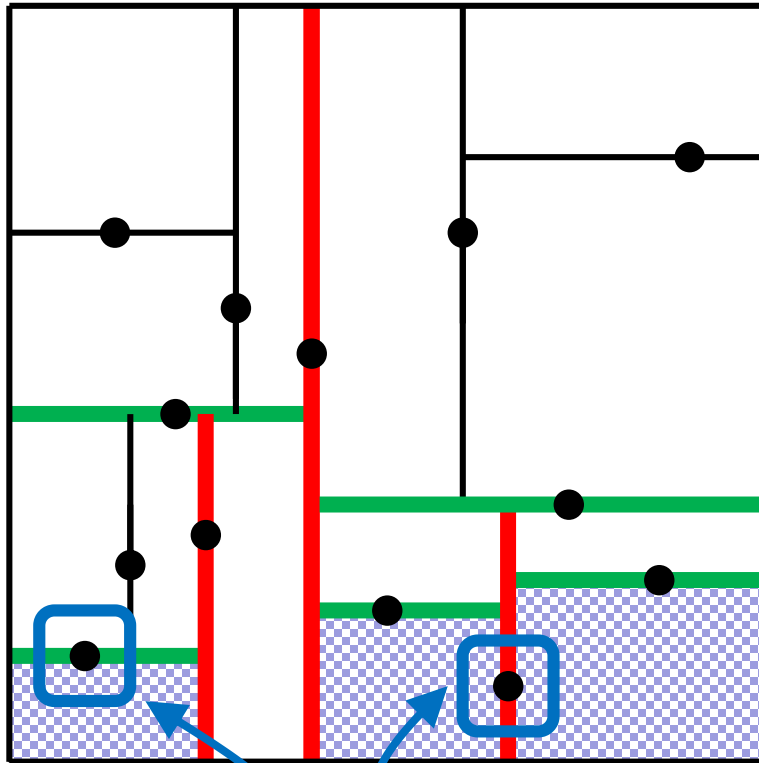
Node with minimal  $y$ -coordinate can be only in L subtree of a node  $N$  corresponding to cutting dimension  $y$ , thus only L subtree of  $N$  (including  $N$ ) must be searched.

FindMin(dim = y)



Node with minimal y-coordinate can be in L or R subtree of a node N corresponding to cutting dimension other than y, thus both subtrees of N (including N) must be searched.

FindMin(dim = y)



Node with minimal y-coordinate can be only in L subtree of a node N corresponding to cutting dimension y, thus only L subtree of N (including N) must be searched.

```
Node findMin( Node N, int dim, int cd ) {  
    if( N == null ) return null;  
    if( cd == dim )  
        if( N.left == null ) return N;  
        else return findMin( N.left, dim, (cd+1)%D );  
    else  
        return min( dim, // see the description bellow  
                    N,  
                    findMin(N.left, dim, (cd+1)%D),  
                    findMin(N.right, dim, (cd+1)%D) );  
}
```

Function min(int dim; Node N1, N2, N3) returns that node out of N1, N2, N3 whose coordinate in dimension dim is the smallest:

```
if( N1.coords[dim] <= N2.coords[dim] && N1.coords[dim] <= N3.coords[dim] ) return N1;  
if( N2.coords[dim] <= N1.coords[dim] && N2.coords[dim] <= N3.coords[dim] ) return N2;  
if( N3.coords[dim] <= N1.coords[dim] && N3.coords[dim] <= N2.coords[dim] ) return N3;
```

Only leaves are physically deleted.

Deleting an inner node  $X$  is done by substituting its key values by key values of another suitable node  $Y$  deeper in the tree.

If  $Y$  is a leaf physically delete  $Y$  otherwise set  $X := Y$  and continue recursively.

Denote cutting dimension of  $X$  by  $cd$ .

If right subtree  $X.R$  of  $X$  is unempty

use operation FindMin to find node  $Y$  in  $X.R$  whose coordinate in  $cd$  is minimal. (It may be sometimes even equal to  $X$  coordinate in  $cd$ .)

If right subtree  $X.R$  of  $X$  is empty

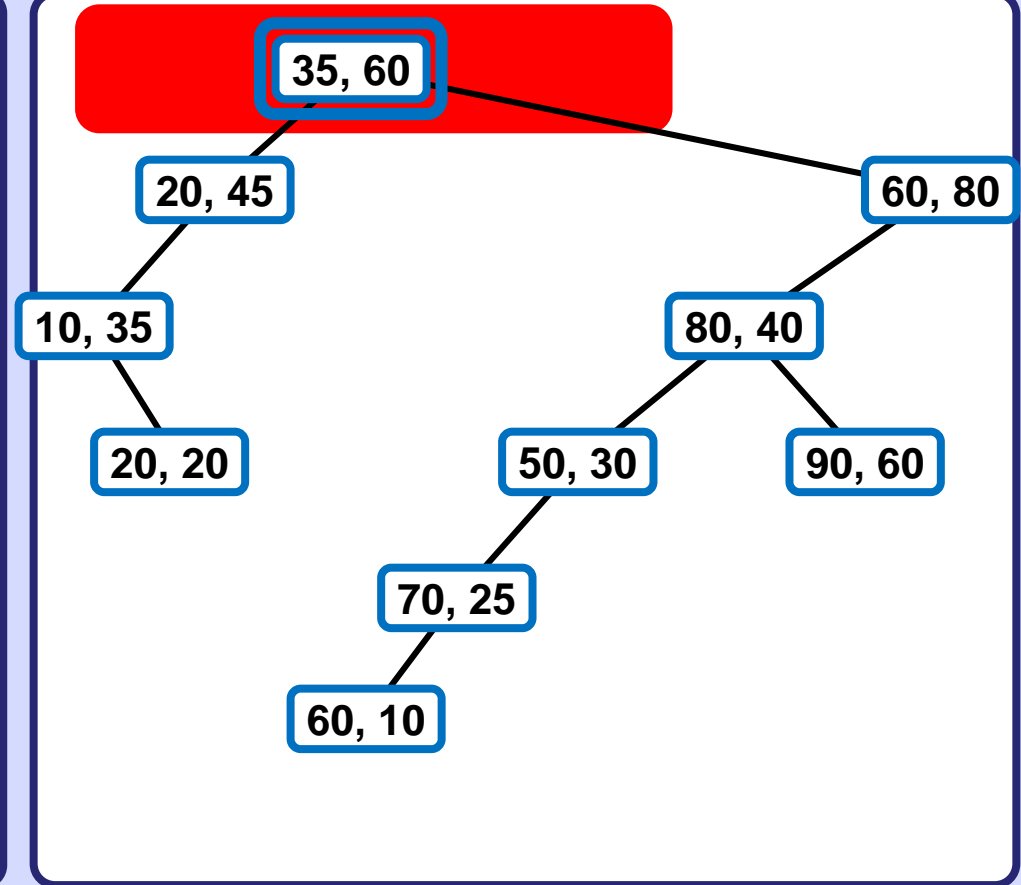
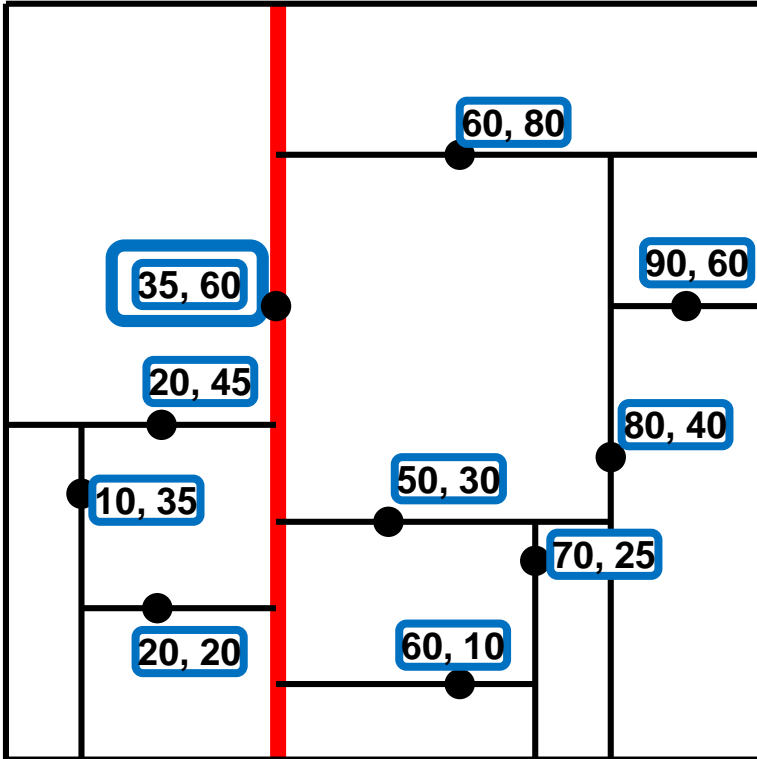
use operation FindMin to find in the left subtree  $X.L$  such node  $Y$  whose coordinate in  $cd$  is minimal. Substitute key values of  $X$  by those of  $Y$ .

Move  $X.L$  to the (empty) right subtree of updated  $X$  (swap  $X.R$  and  $X.L$ ).

Now  $X$  has unempty right subtree, continue the process with previous case.

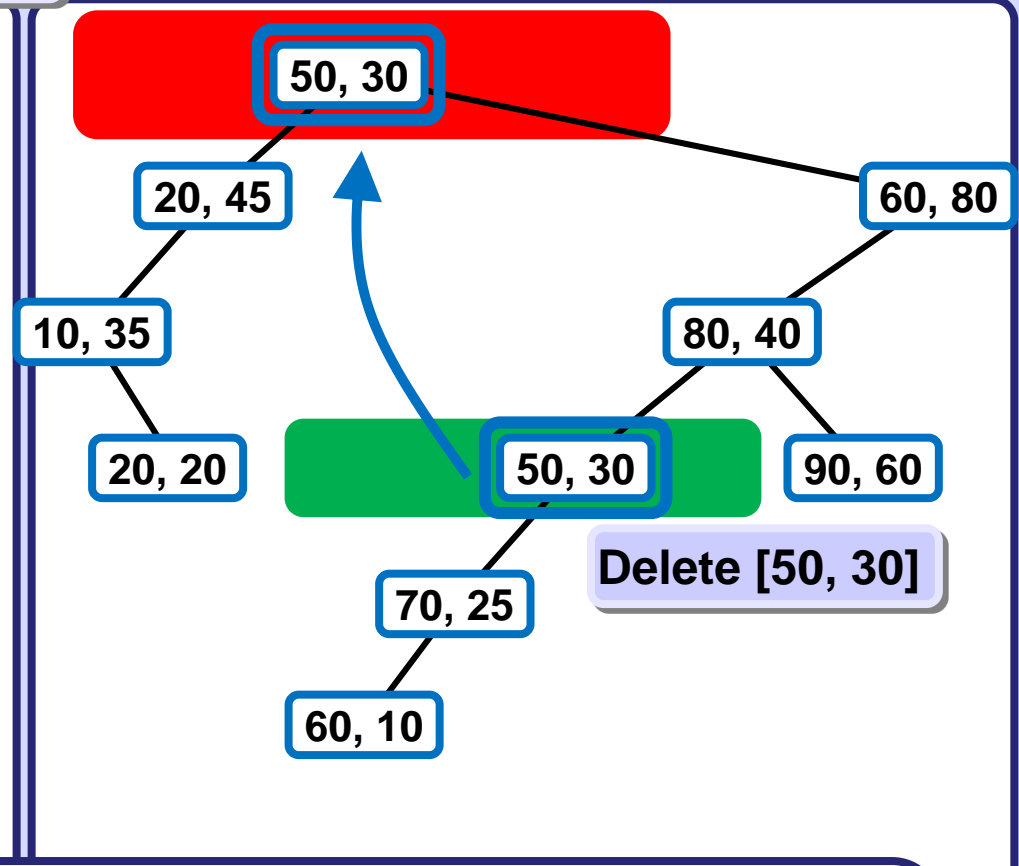
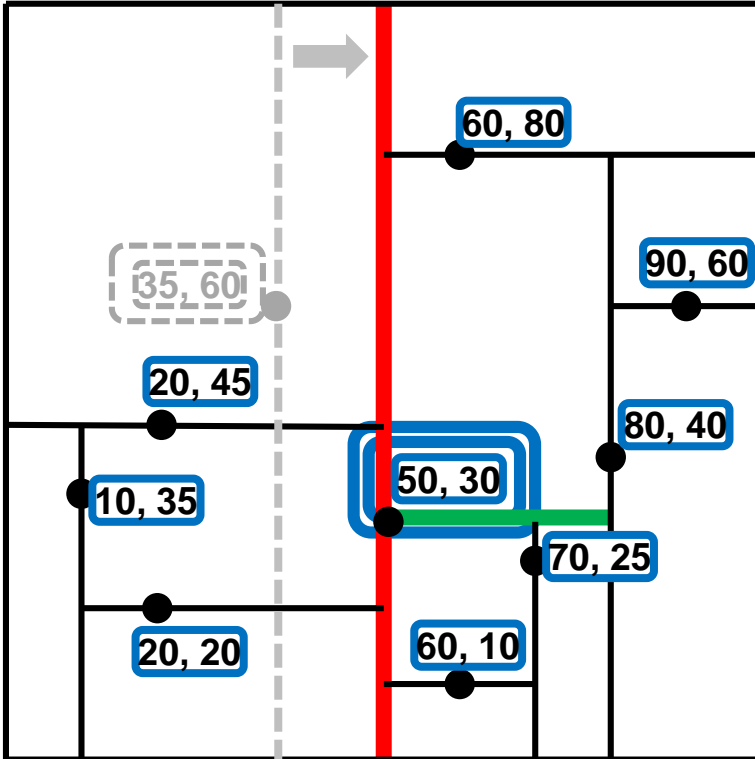


Delete [35, 60]



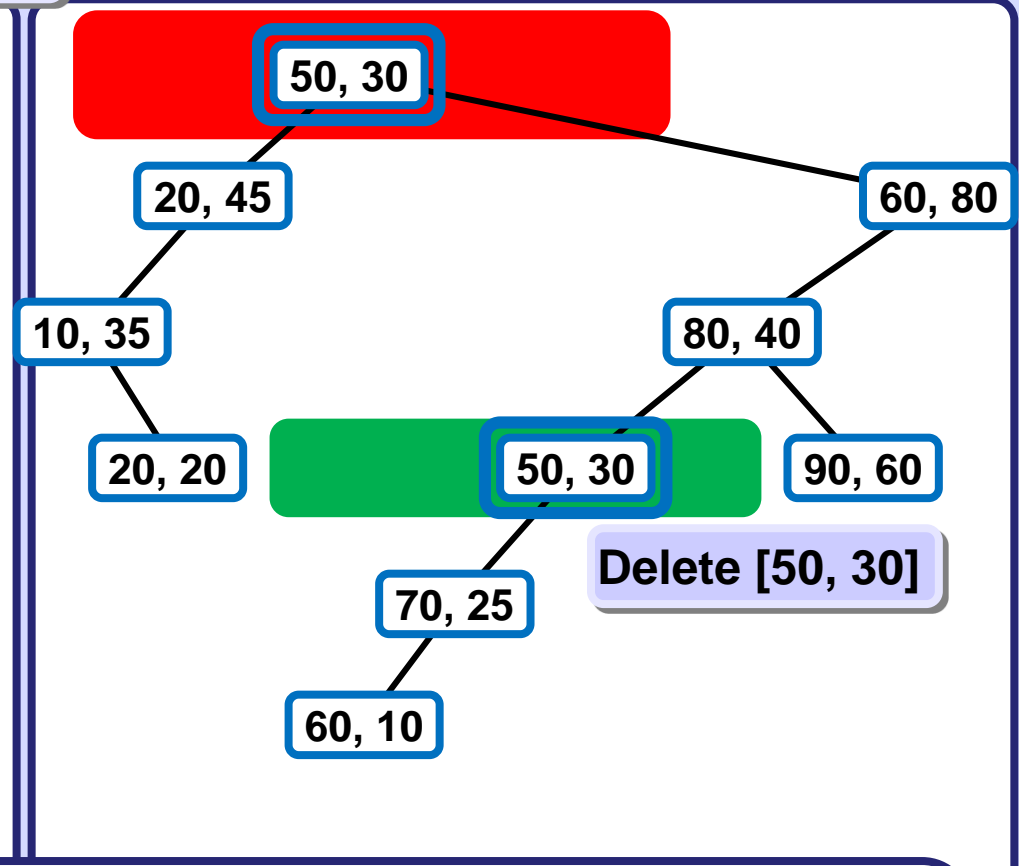
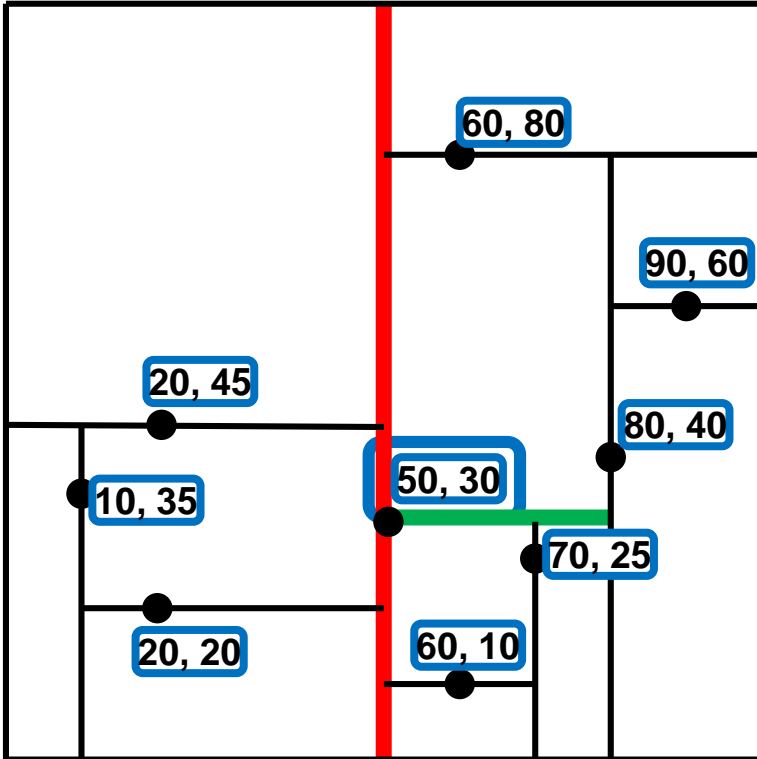
Deleting node [35, 60], its cutting dimension is x.  
 Find node Y with minimum x-coordinate in right subtree of [35, 60].  
 Note that Y might have different cutting dimension.

Delete [35, 60] ... In progress....



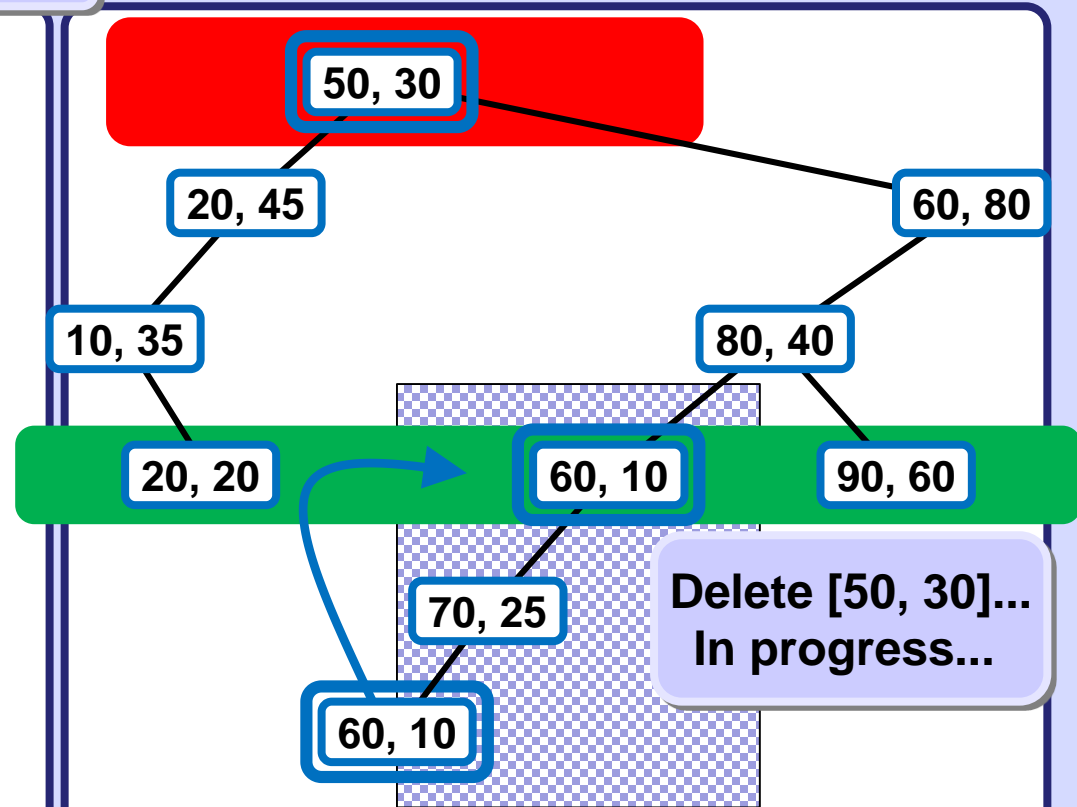
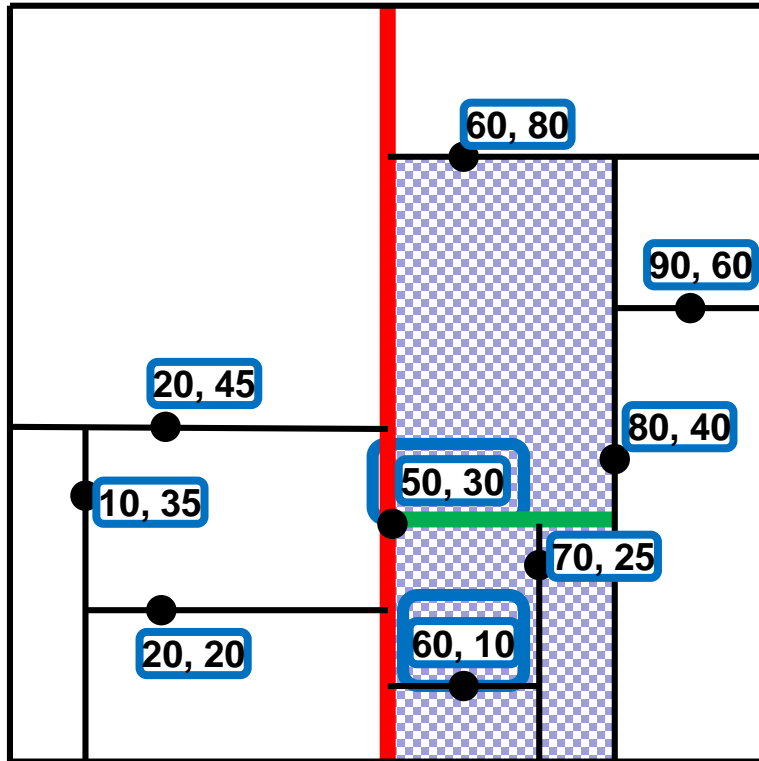
Deleting node [35, 60], its cutting dimension is x.  
 Find node Y with minimum x-coordinate in right subtree of [35, 60].  
 Fill node [35, 60] with keys of Y  
 and if Y is not a leaf continue by recursively deleting Y.

Delete [35, 60] ... In progress....



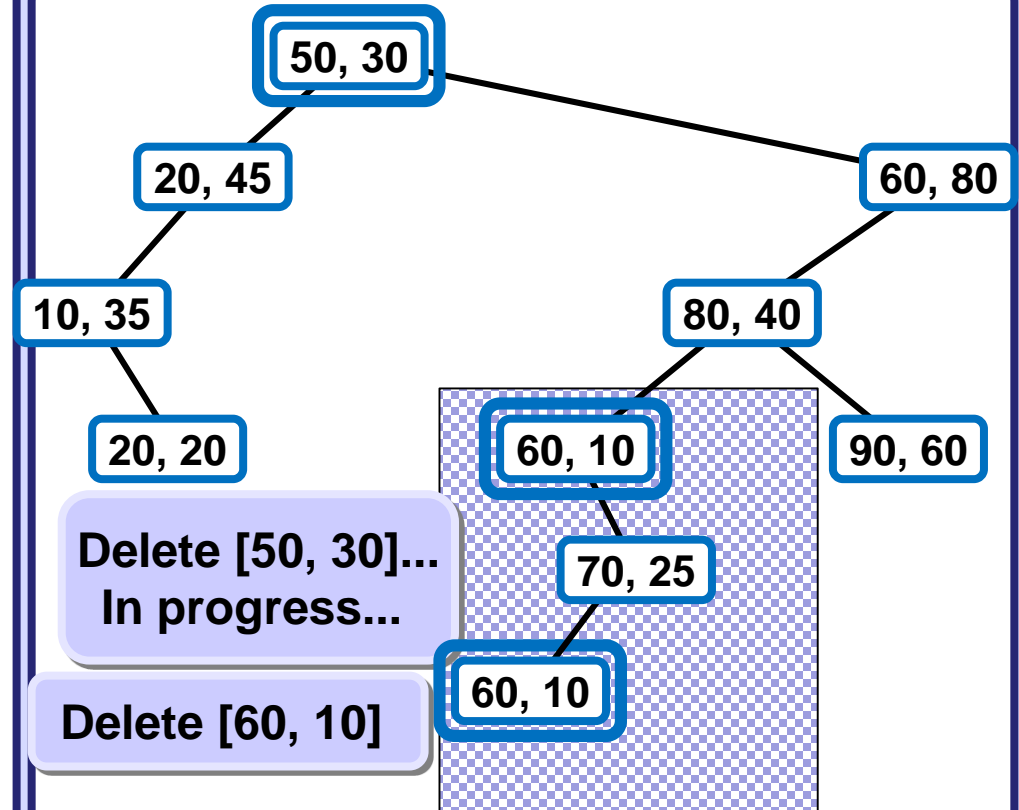
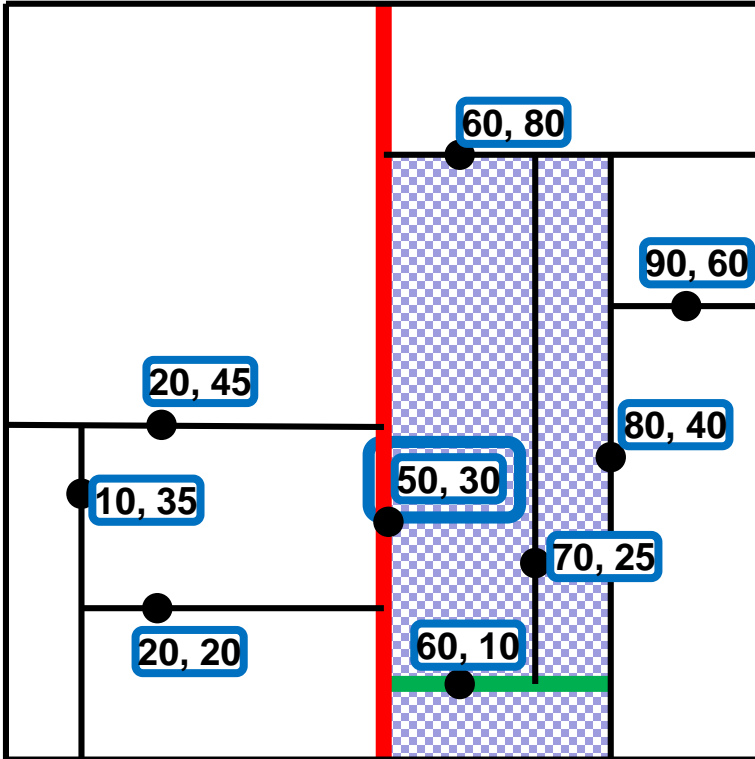
Deleting node [50, 30], its cutting dimension is y, it has no R subtree.  
 Find node Z with minimum y-coordinate in LEFT subtree of [50, 30],  
 Fill [50, 30] with keys of Z and move L subtree of [50, 30] to its R subtree.


Delete [35, 60] ... In progress....

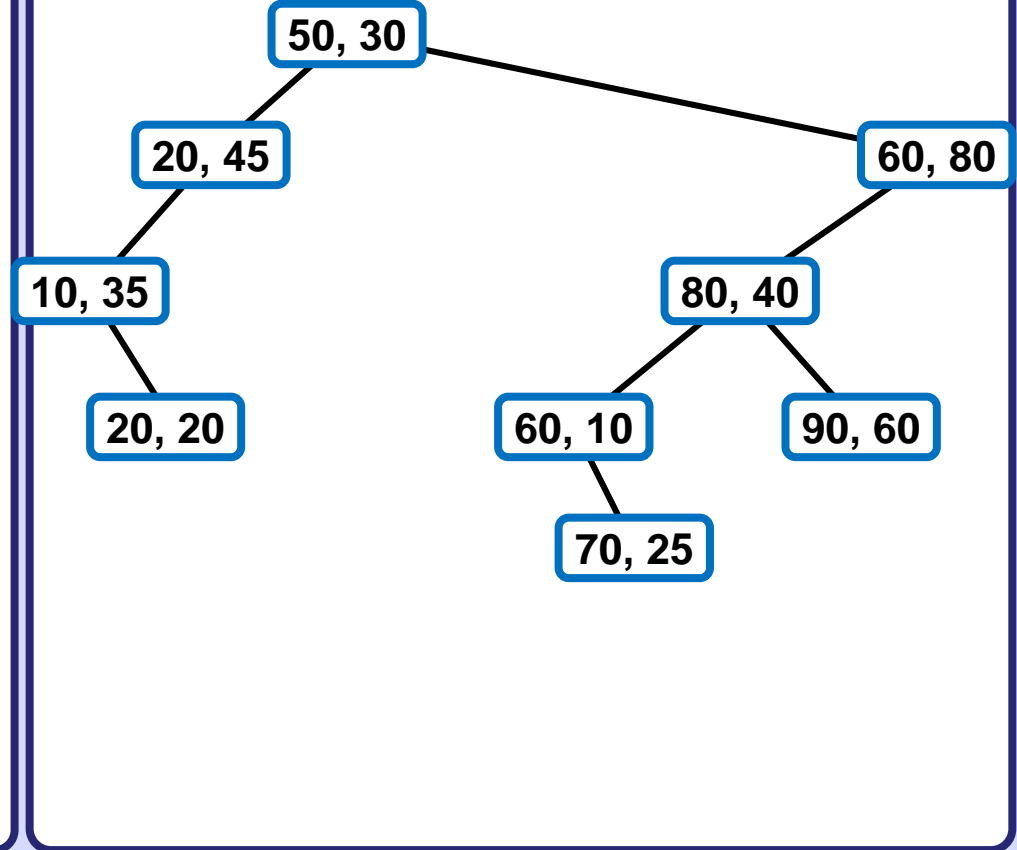
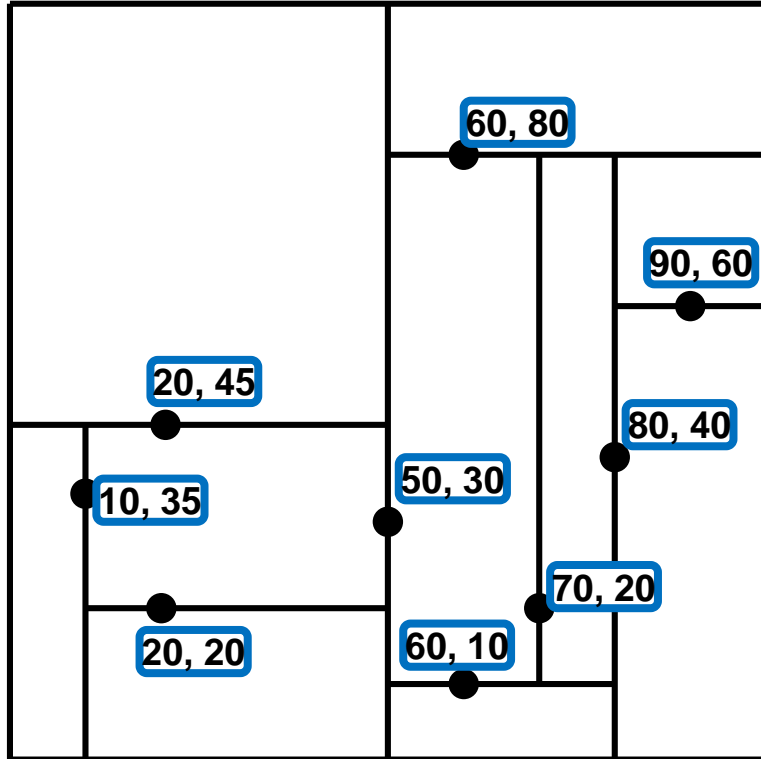


Deleting node [50, 30], its cutting dimension is y, it has no R subtree. Find node Z with minimum y-coordinate in LEFT subtree of [50, 30], Fill [50, 30] with keys of Z and move L subtree of [50, 30] to its R subtree. If Z is not a leaf continue by recursively deleting Z.

Delete [35, 60] ... In progress....

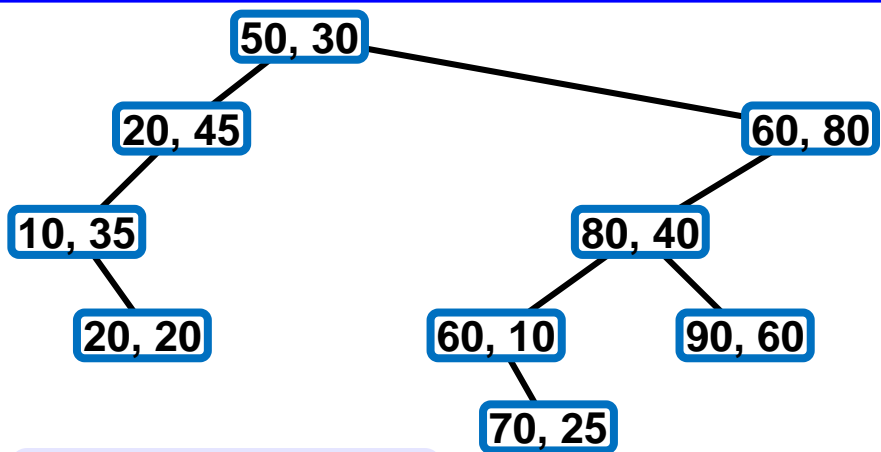
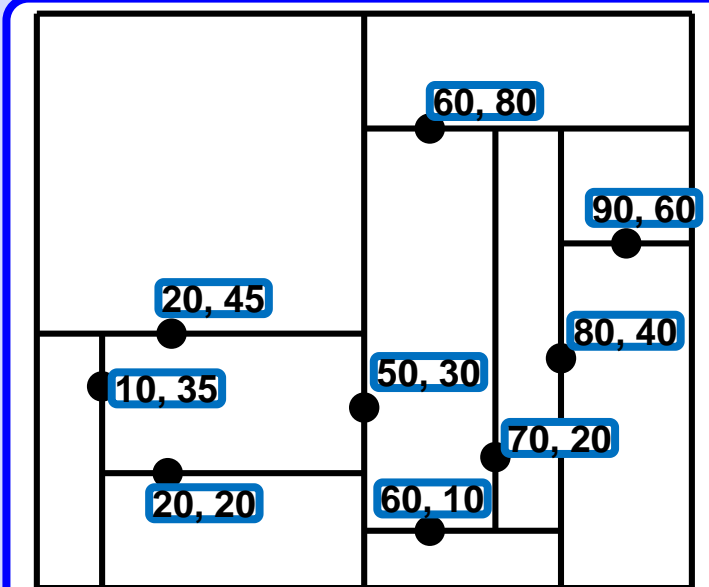
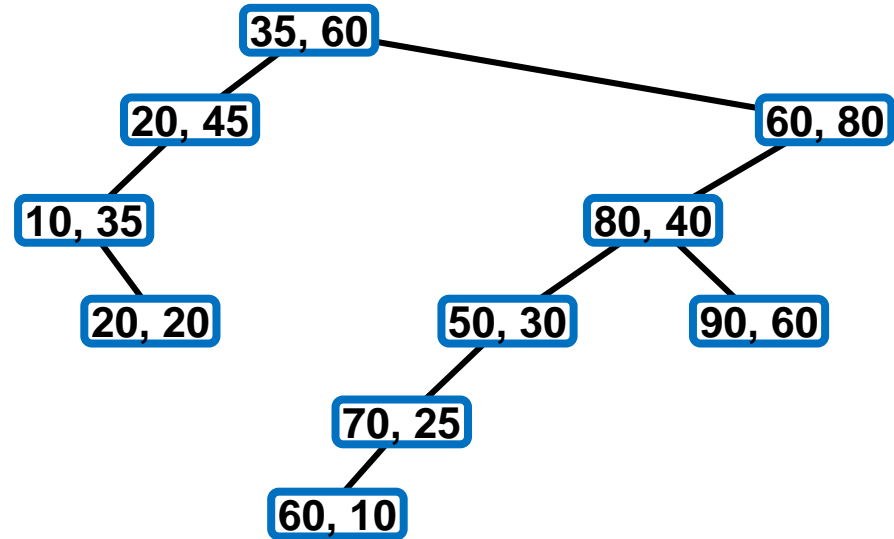
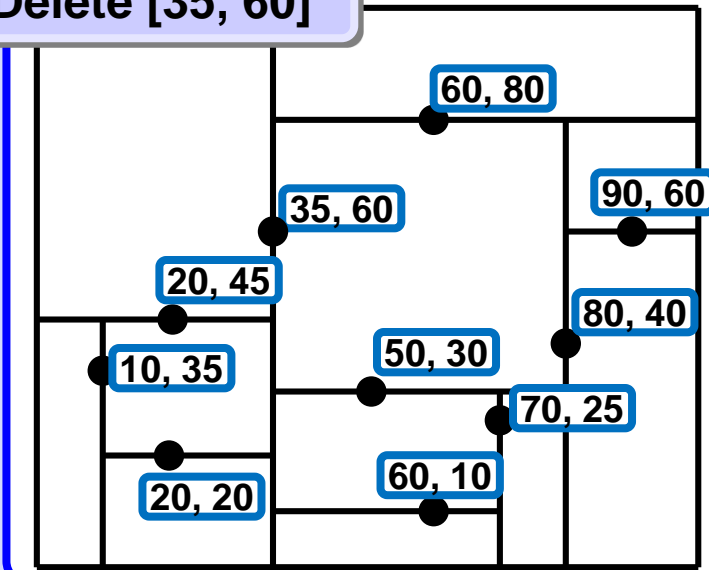


Deleting original node [60, 10], it it is a leaf, delete it and and stop.  
 Note the change in the cell division left to [80, 40], the node with minimal y-coordinate becomes the splitting node for the corresponding area .



Deleted [35, 60]

Delete [35, 60]

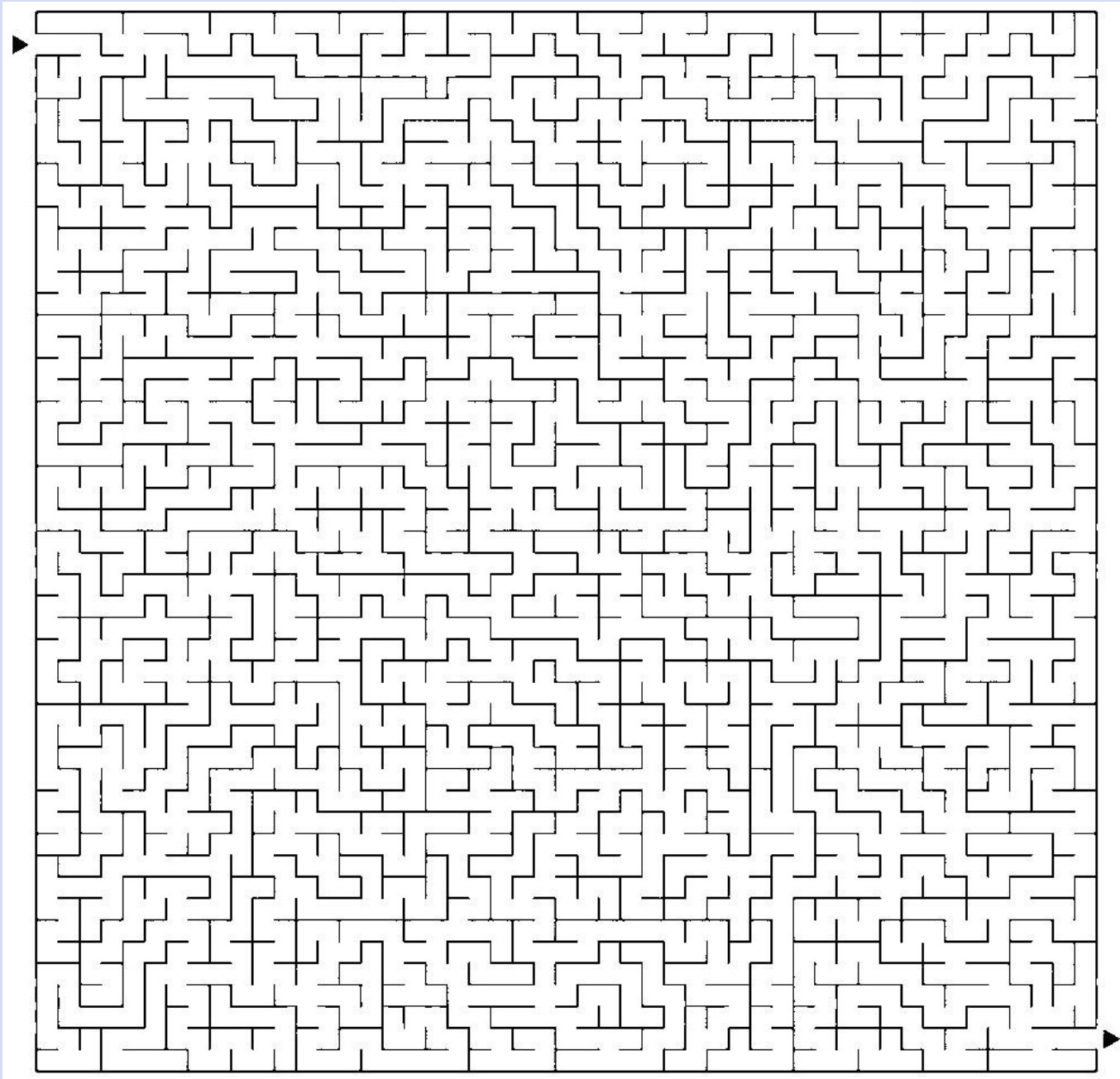


Deleted [35, 60]

```
Node delete(Point P, Node N, int cd) {
    if( N == null ) throw new ExceptionDeleteNonexistentPoint();
    else if( P.equals(N.coords) ){ // point P found in N
        if( N.right != null ){ // replace deleted from right
            N.coords = findMin( N.right, cd, (cd+1)%D ).coords();
            N.right = delete( N.coords, N.right, (cd+1)%D );
        }
        else if( N.left != null ){ // replace deleted from left
            N.coords = findMin( N.left, cd, (cd+1)%D ).coords();
            N.right = delete( N.coords, N.left, (cd+1)%D );
            N.left = null;
        }
        else N = null; // destroy leaf N
    }
    else // point P not found yet
        if( P.coords[cd] < N.coords[cd] ) // search left subtree
            N.left = delete( P, N.left, (cd+1)%D );
        else // search right subtree
            N.right = delete( P, N.right, (cd+1)%D );
    return N;
}
```



	1							9
			5					3
	7	8	2			1		
8		5	3					
				5	6			
7	6							
3			6				5	7
9						4	1	6
				8				



**Nearest Neighbour search**

using

**k-d tree**

Search starts in the root  
and runs recursively in both L and R subtrees of the current node.

Register and update **partial results**:

Object *close* = {*close.point*, *close.dist*}.

Field *.point* refers to the node (point) which is closest to the query so far,  
field *.dist* contains euclidean distance from *.point* to the query.

Perform **pruning**:

During the search dismiss the cells (and associated subtrees)  
which are too far from query. Object *close* helps to accomplish this task.

**Traversal order** (left or right subtree is searched first) depends on simple  
(in other variants of k-d tree on more advanced) heuristic:

First search the subtree whose cell associated with it is closer to the query.

This does not guarantee better results but in practice it helps.

To implement Nearest Neighbour Search suppose existence of the following:

1. Class HyperRectangle (or Box, in 2D just Rectangle) representing cells of particular nodes in k-d tree. This class offers two methods:

HyperRectangle trimLeft(int cd, coords c)

HyperRectangle trimRight(int cd, coords c)

When hyperrectangle *this* represents the current cell, cd represents cutting dimension, c represents coordinates of a point (or node)

then trimLeft returns the hyperrectangle associated with the left subtree of the point/node with coordinates c. Analogously trimRight returns hyperrectangle associated with the right subtree.

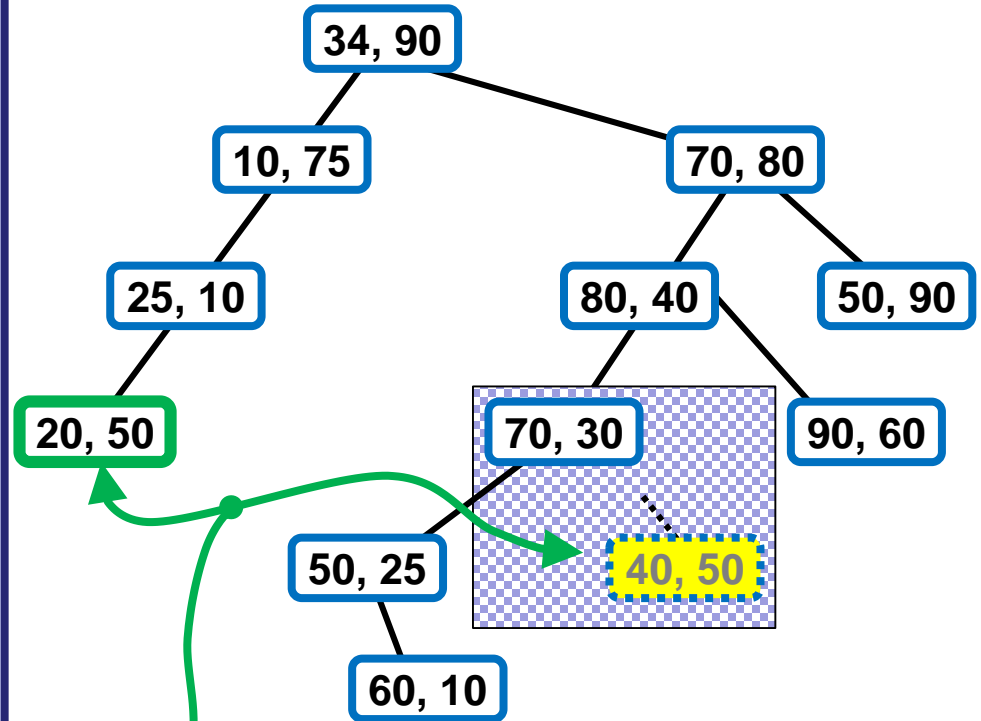
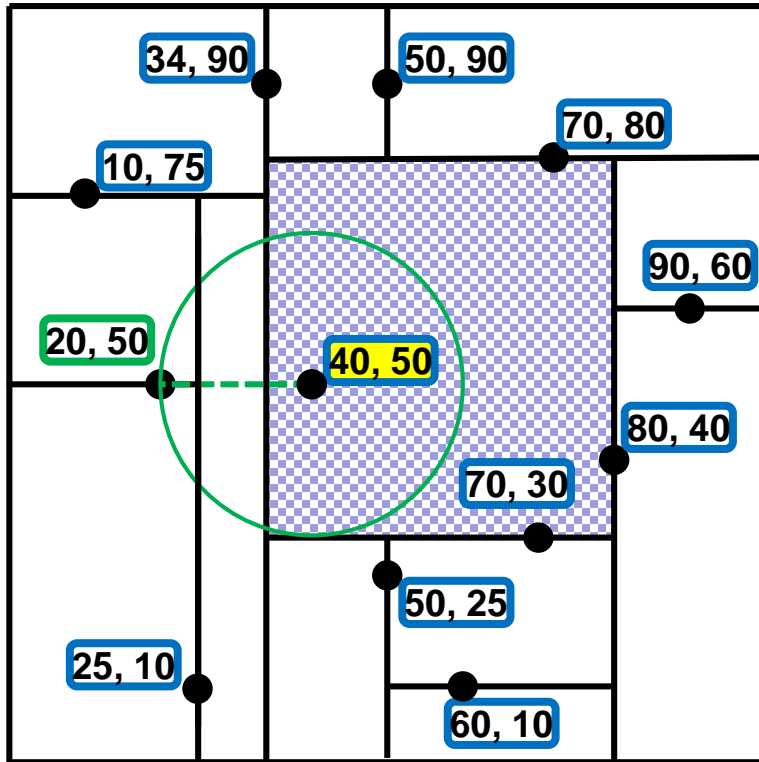
2. Class or utility G (like Geometry) equipped with methods

G.distance(Point p, Point q) with obvious functionality

G.distance(point p, Hyperrectangle r) which computes distance from q to the point x of r which is nearest to q.

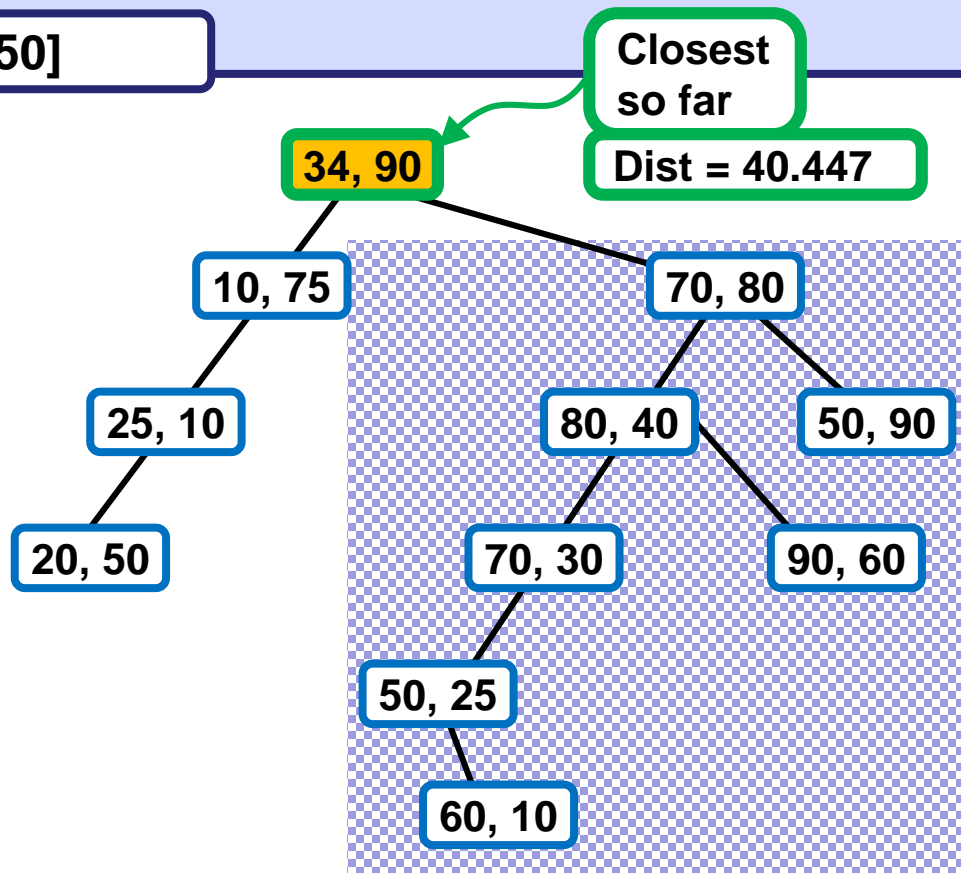
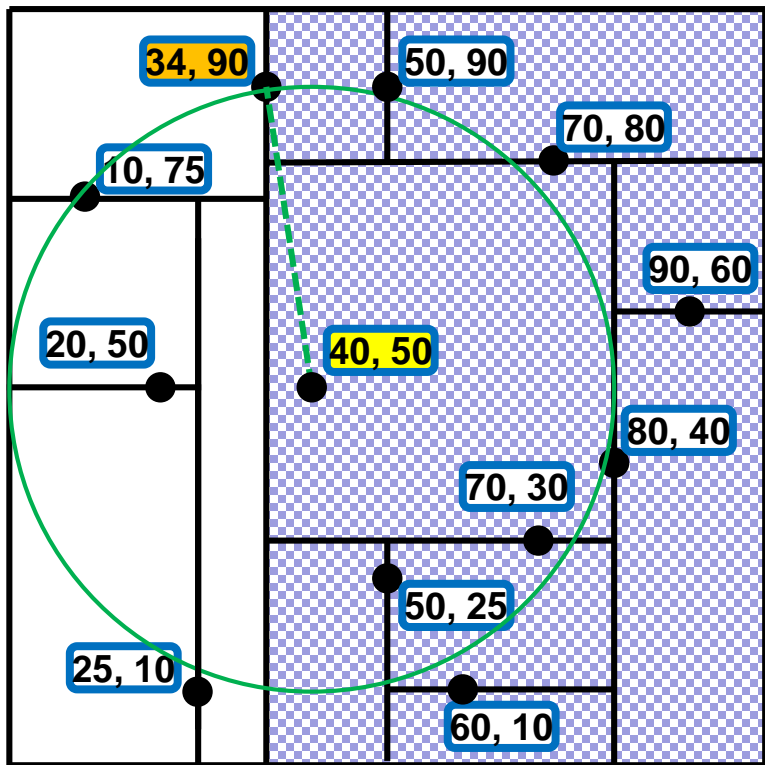
3. Object *close* with fields *dist* and *point*, storing the best distance found so far and reference to the point at which it was attained. Initialize by *dist* = inf, *point* = null.

Find Nearest Neighbour to [40, 50]



The query point  $Q = [40, 50]$  lies inside (empty) leaf cell right to the node  $[70, 30]$ . The closest point to query  $Q = [40, 50]$  is the point  $[20, 50]$  which, in fact, lies in a distant part of the tree.

Find Nearest Neighbour to [40, 50]

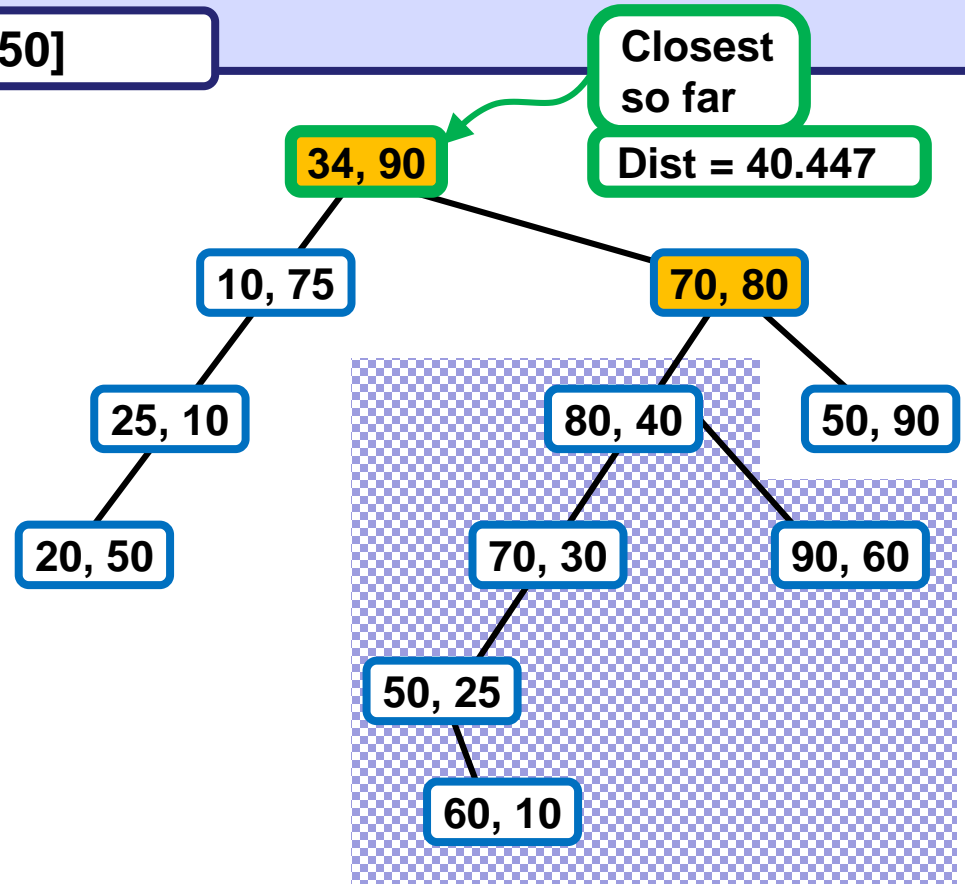
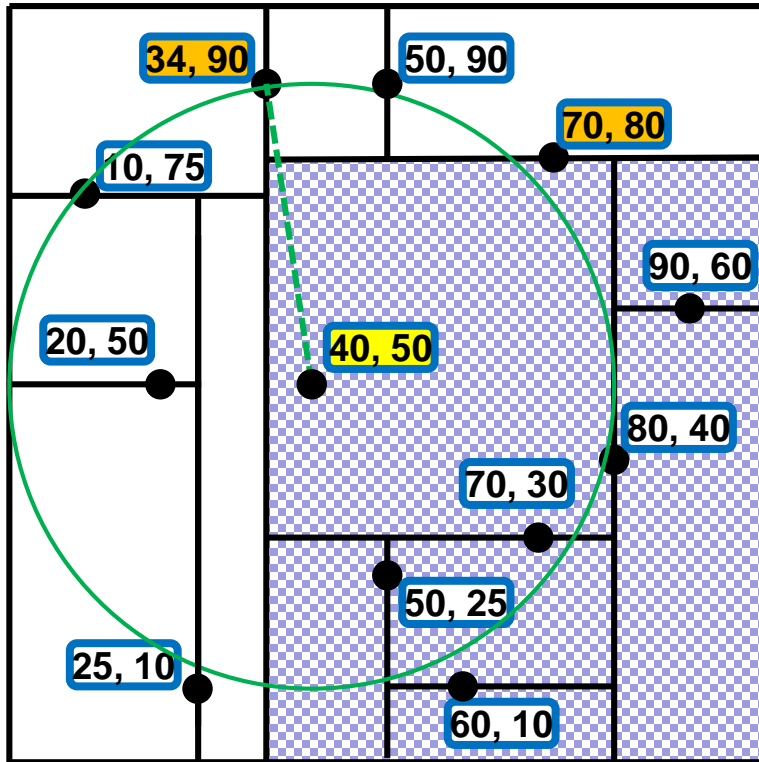


Distance  $(Q, [34, 90]) = 40.447$ .

Heuristic: The query point  $Q = [40, 50]$  lies inside the (hyper) rectangle  $r_1$  associated with the right subtree of the root  $[34, 90]$ , so the distance  $Q$  to  $r_1$  is 0. The search starts in the right subtree of the root.

Searched nodes

Find Nearest Neighbour to [40, 50]



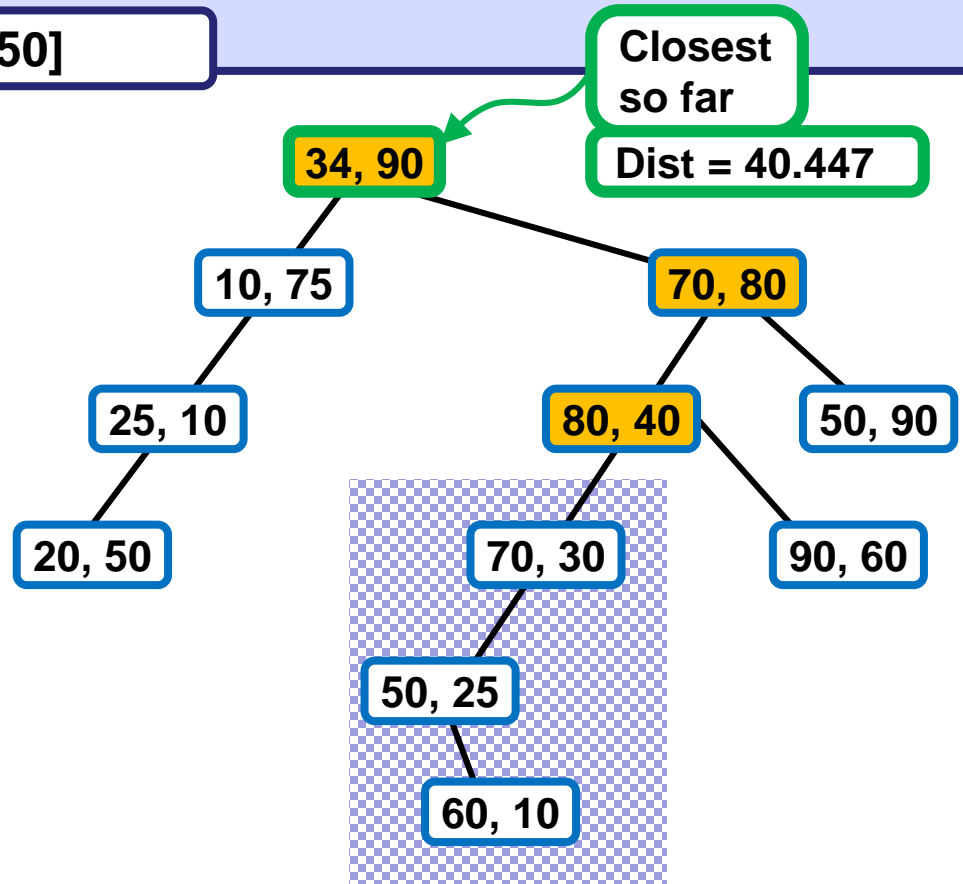
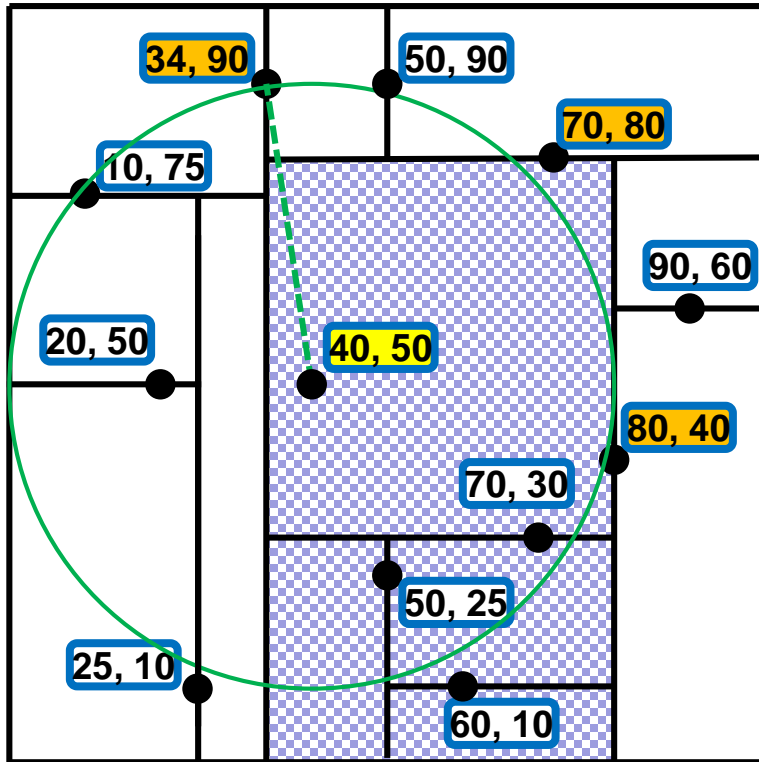
Distance  $(Q, [70, 80]) = 42.426 > 40.447$ .

Heuristic: The query point  $Q = [40, 50]$  lies inside the (hyper) rectangle  $r_2$  associated with the left subtree of the node  $[70, 80]$ , so the distance  $Q$  to  $r_2$  is 0. The search continues in the left subtree of  $[70, 80]$ .

Searched nodes



Find Nearest Neighbour to [40, 50]

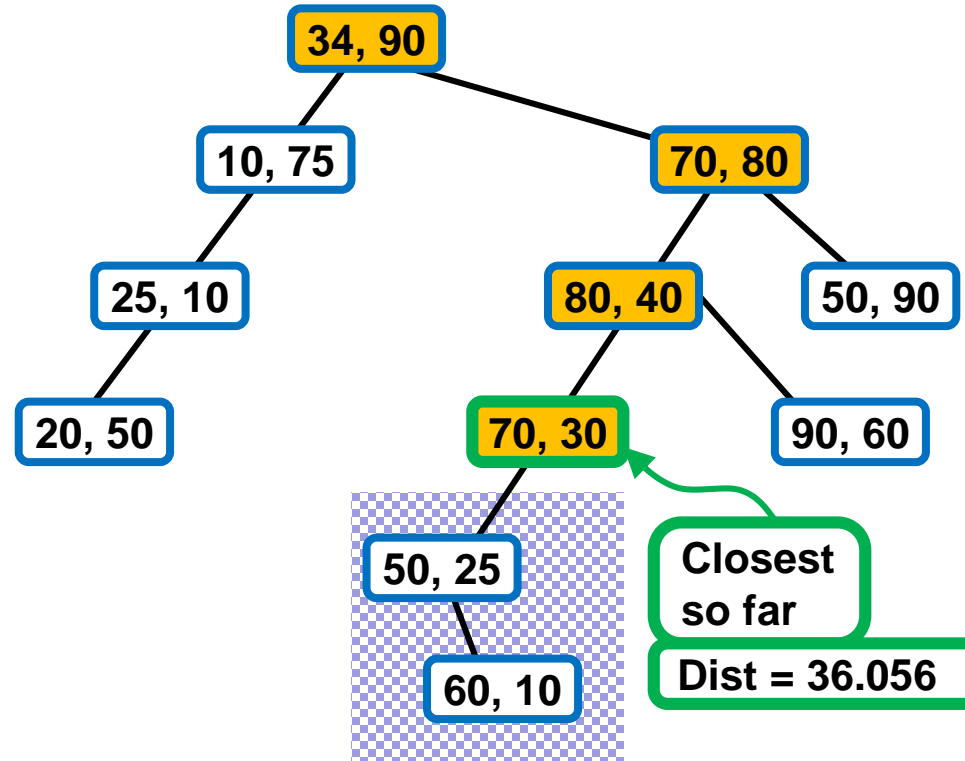
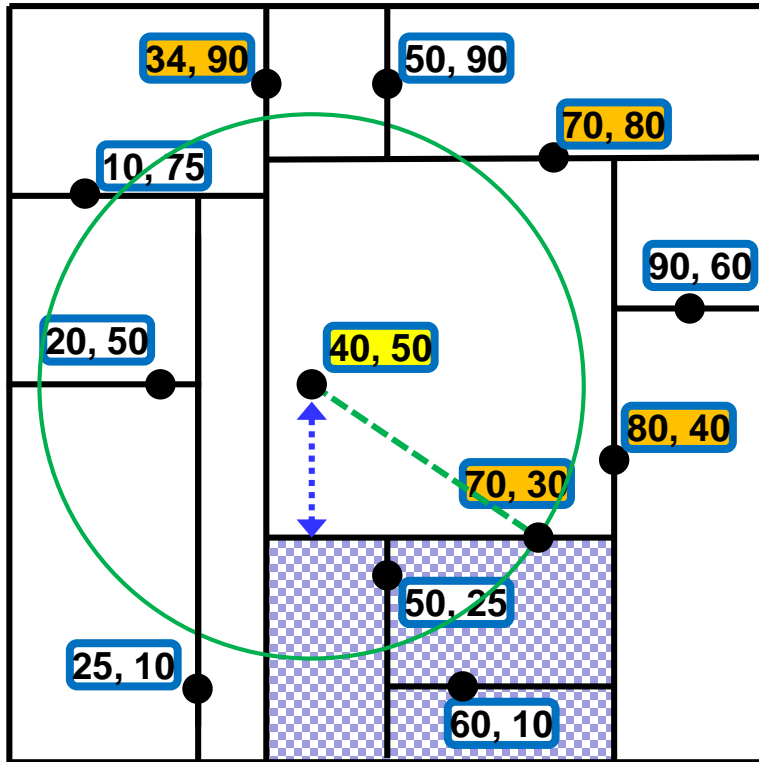


Distance  $(Q, [80, 40]) = 41.231 > 40.447$ .

Heuristic: The query point  $Q = [40, 50]$  lies inside the (hyper) rectangle  $r_3$  associated with the left subtree of the node  $[80, 40]$ , so the distance  $Q$  to  $r_3$  is 0. The search continues in the left subtree of  $[80, 40]$ .

Searched nodes

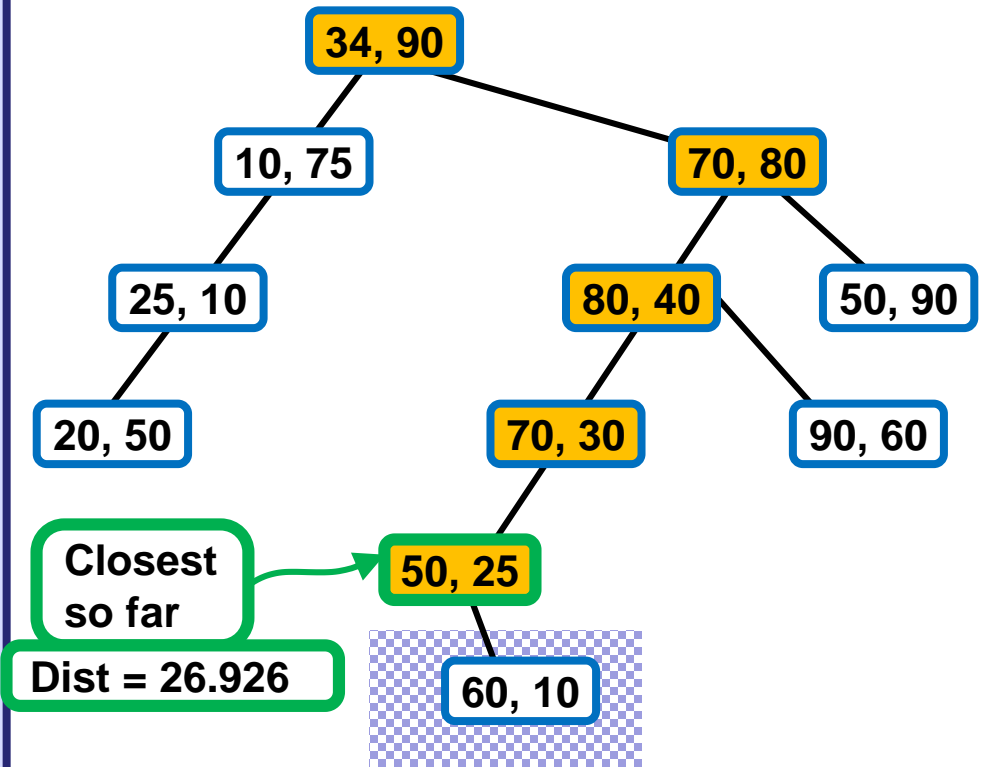
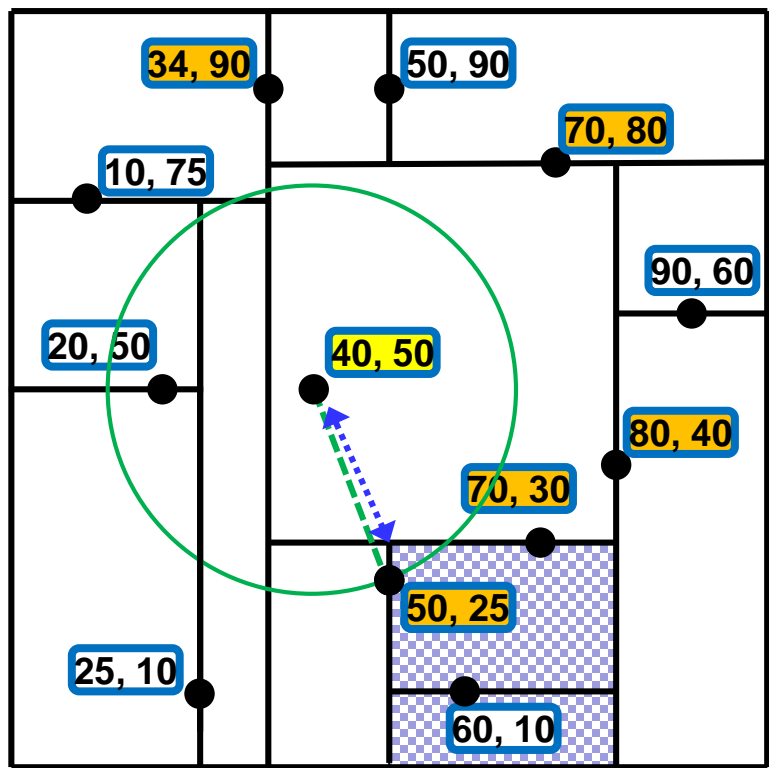
Find Nearest Neighbour to [40, 50]



Distance  $(Q, [70, 30]) = 36.056 < 40.447$ ,  $[70, 30]$  becomes new *close* node.  
 Pruning?: The the distance from  $Q = [40, 50]$  to the (hyper) rectangle  $r_4$  associated with the left subtree of  $[70, 30]$  is  $20.0 < 36.056$ . No pruning occurs.  
 The search continues in the left subtree of  $[70, 30]$ .

Searched nodes

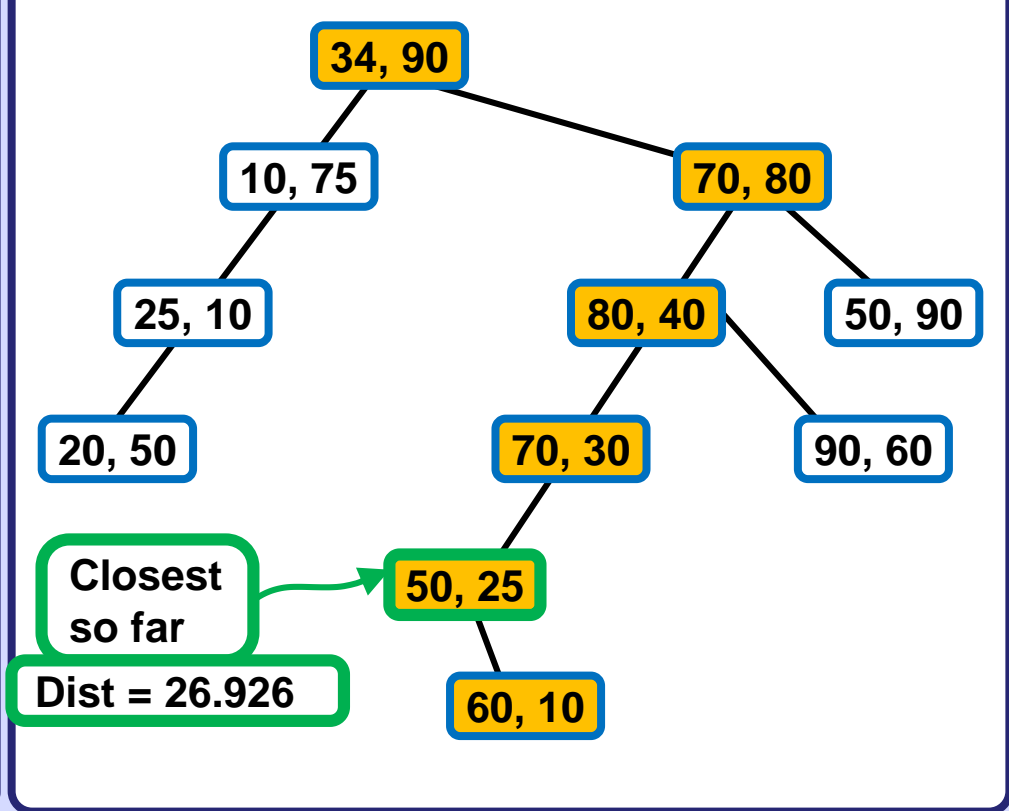
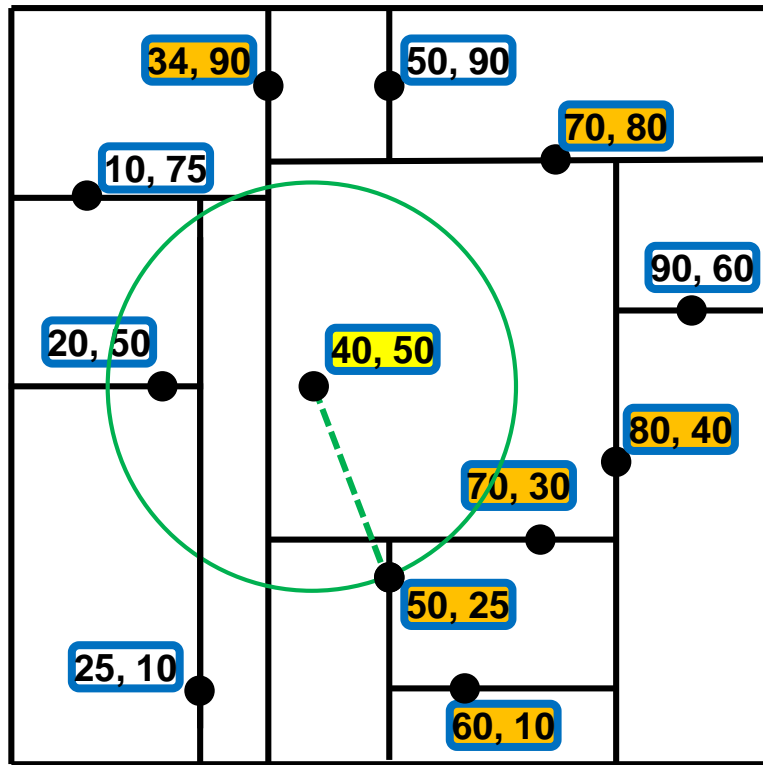
Find Nearest Neighbour to [40, 50]



Distance  $(Q, [50, 25]) = 26.926 < 36.056$ ,  $[50, 25]$  becomes new *close* node.  
 Pruning?: The the distance from  $Q = [40, 50]$  to the (hyper) rectangle  $r_5$  associated with the right subtree of  $[50, 25]$  is  $22.361 < 26.926$ . No pruning occurs.  
 The search continues in the right subtree of  $[50, 25]$ .

Searched nodes

Find Nearest Neighbour to [40, 50]

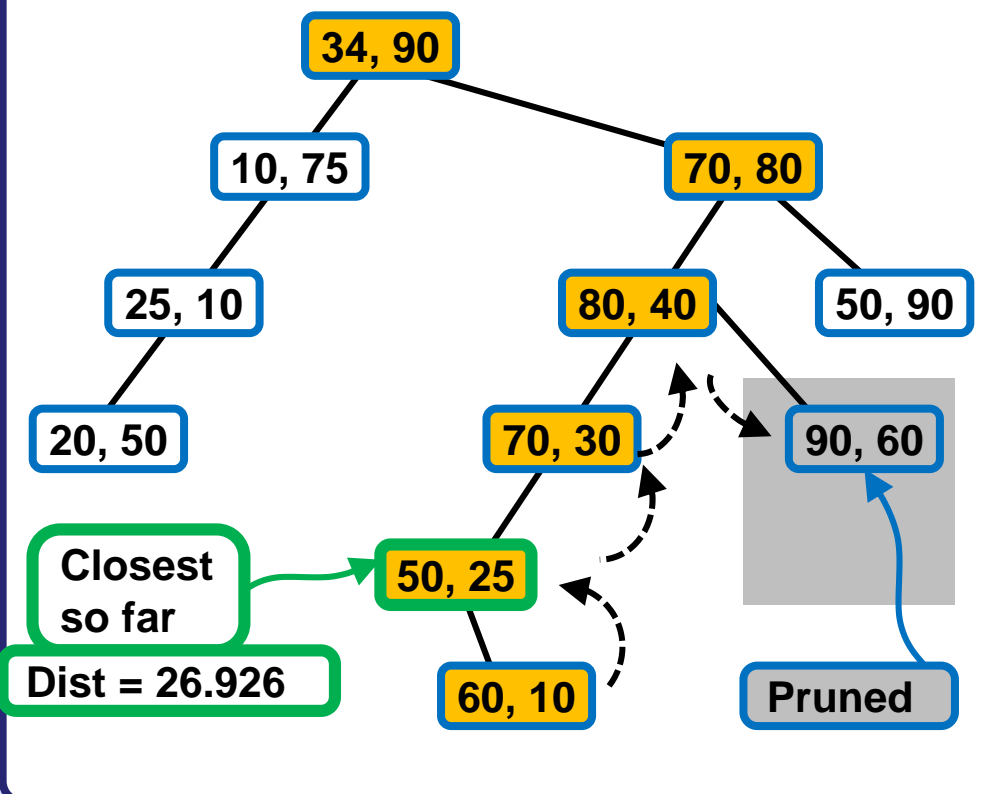
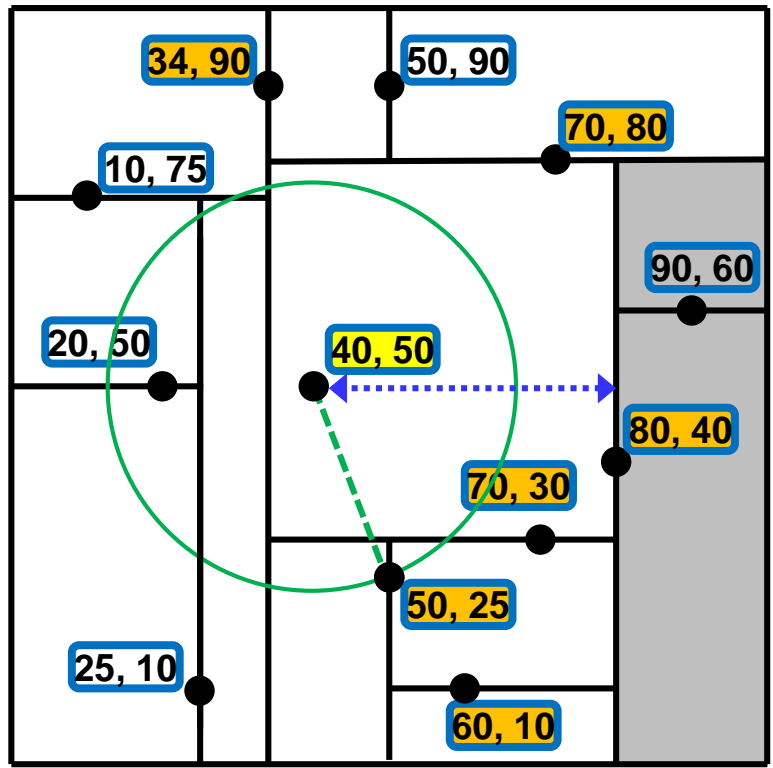


Distance (Q, [60, 10] ) = 44.721 > 26.926.

The search has reached a leaf and returns (due to recursion) to the last unexplored branch .

Searched nodes

Find Nearest Neighbour to [40, 50]



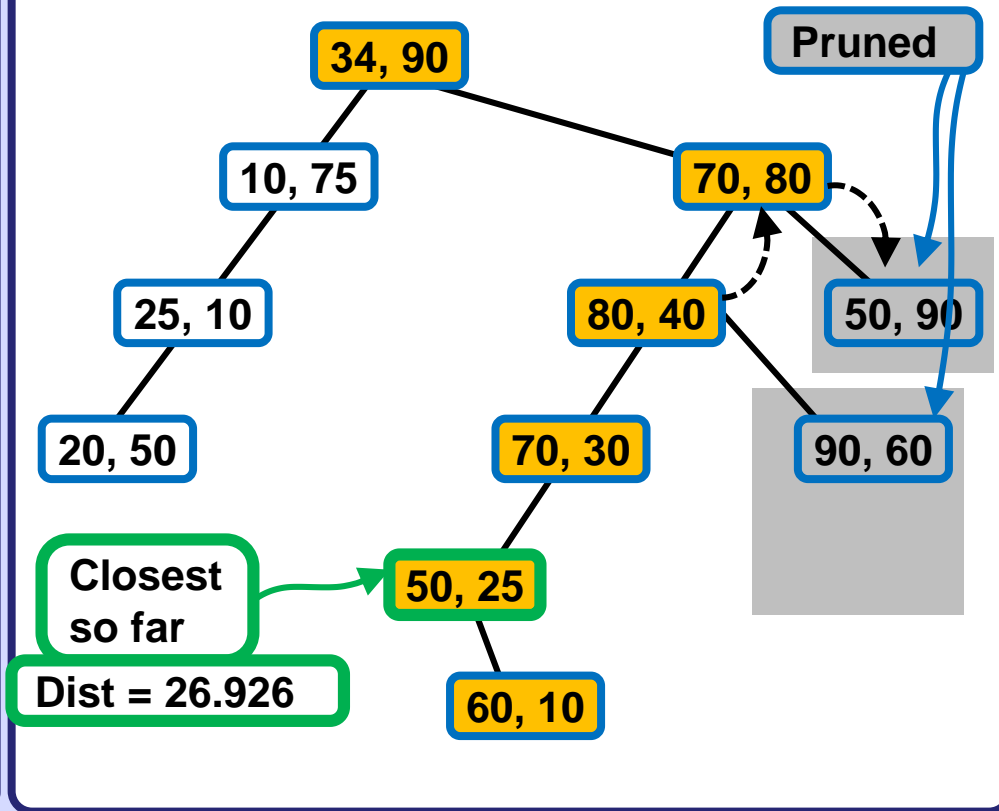
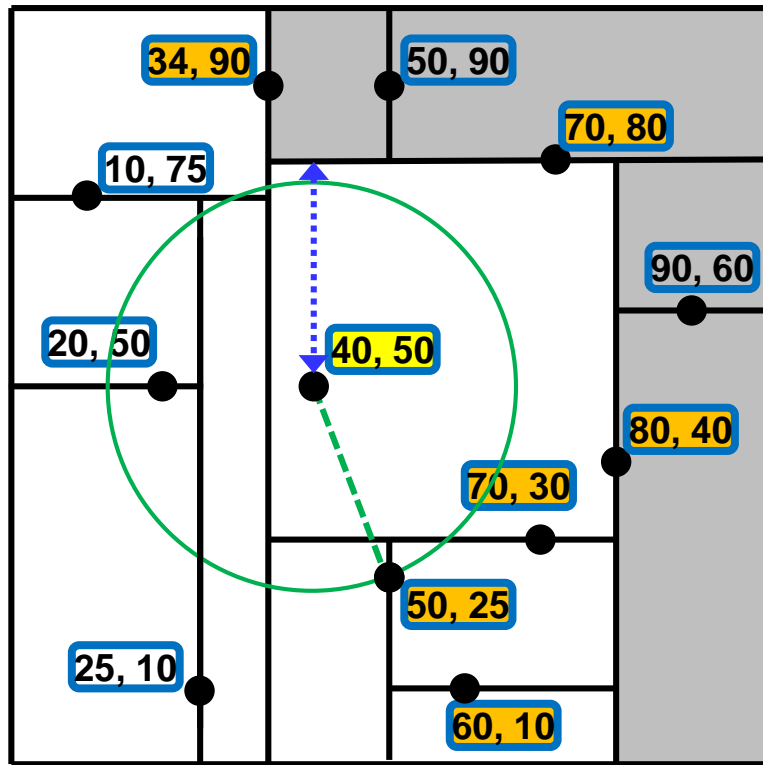
The search has returned to the last unexplored branch.

Pruning?: The the distance from  $Q = [40, 50]$  to the (hyper) rectangle  $r_6$  associated with the right subtree of  $[80, 40]$  is  $40.0 > 26.926$ . The whole branch is pruned.

The search returns back to the previous unexplored branch.

Searched nodes

Find Nearest Neighbour to [40, 50]



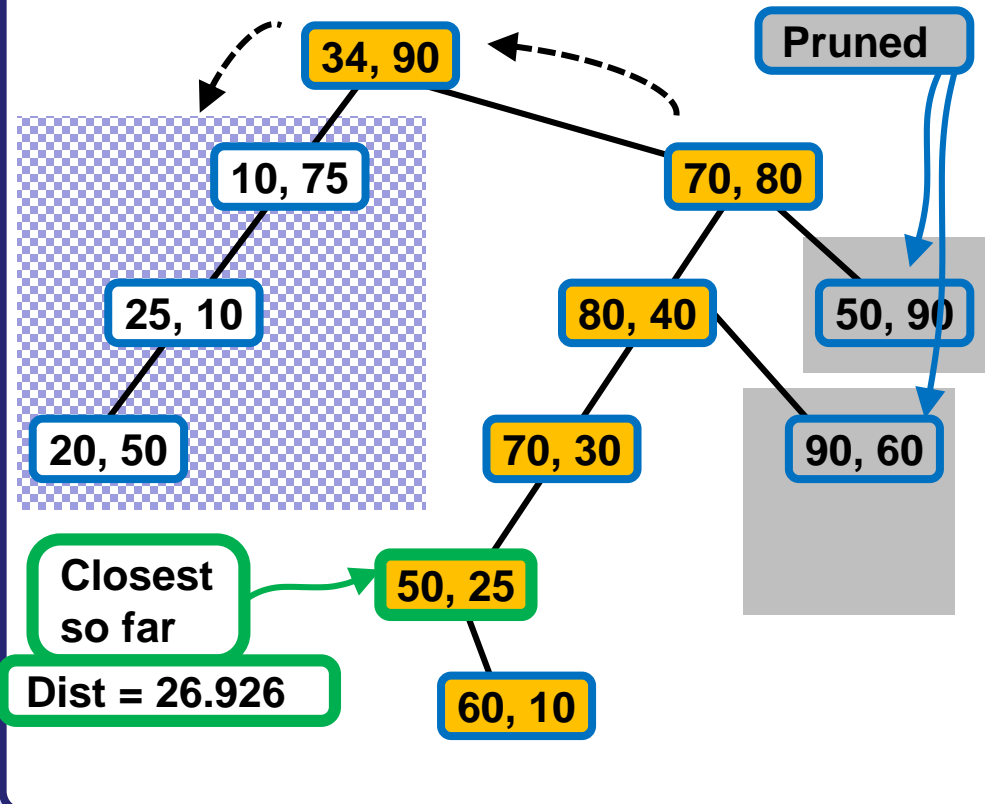
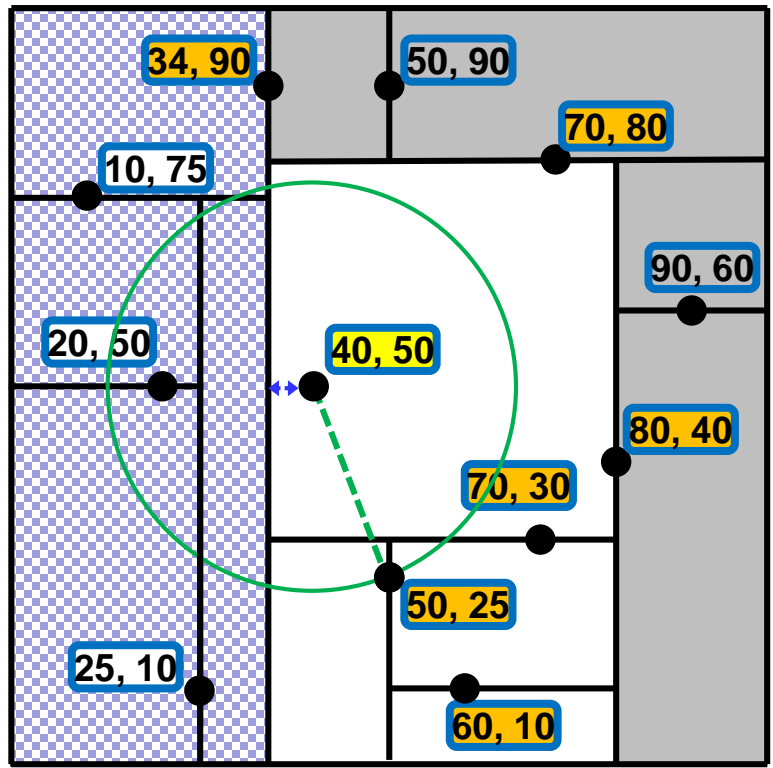
The search has returned to the last unexplored branch.

Pruning?: The the distance from  $Q = [40, 50]$  to the (hyper) rectangle  $r_7$  associated with the right subtree of  $[70, 80]$  is  $30.0 > 26.926$ . The whole branch is pruned.

The search returns back to the previous unexplored branch.

Searched nodes

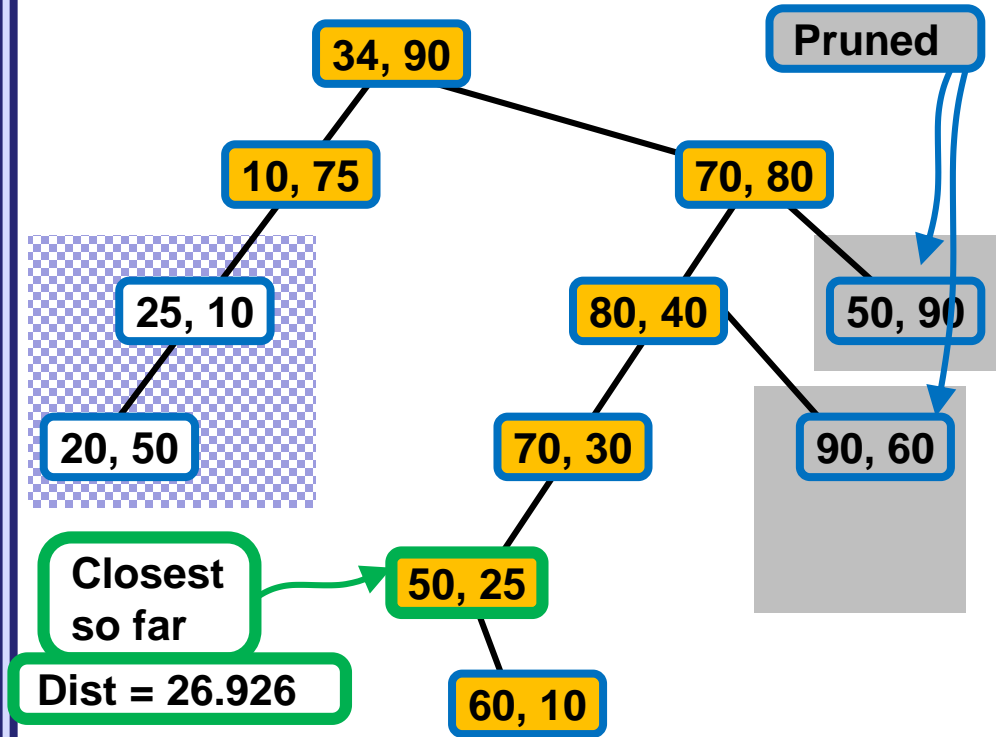
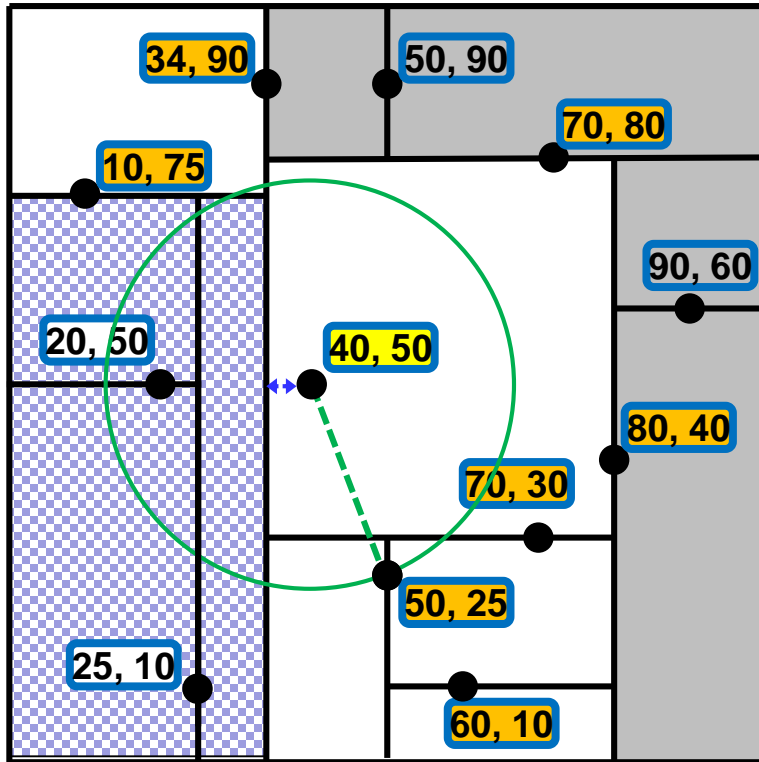
Find Nearest Neighbour to [40, 50]



The search has returned to the last unexplored branch.  
 Pruning?: The distance from  $Q = [40, 50]$  to the (hyper) rectangle  $r_8$  associated with the left subtree of  $[34, 90]$  is  $6.0 < 26.926$ . No pruning occurs.  
 The search continues in the left subtree of  $[34, 90]$ .

Searched nodes

Find Nearest Neighbour to [40, 50]



Distance  $(Q, [10, 75]) = 39.051 > 26.926$ .

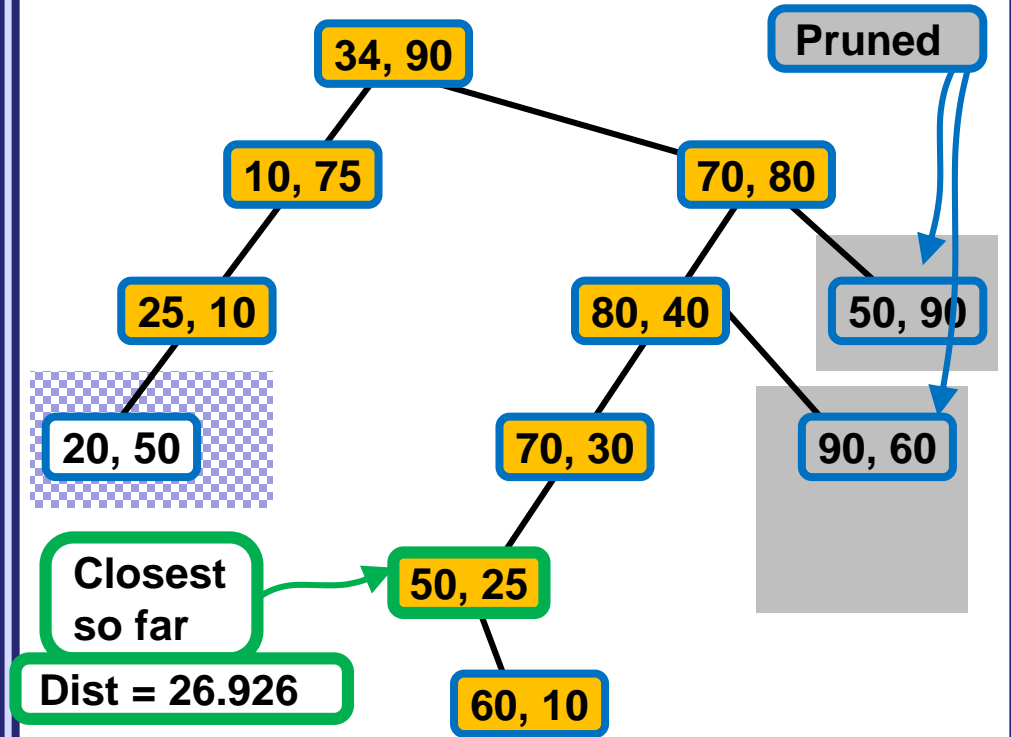
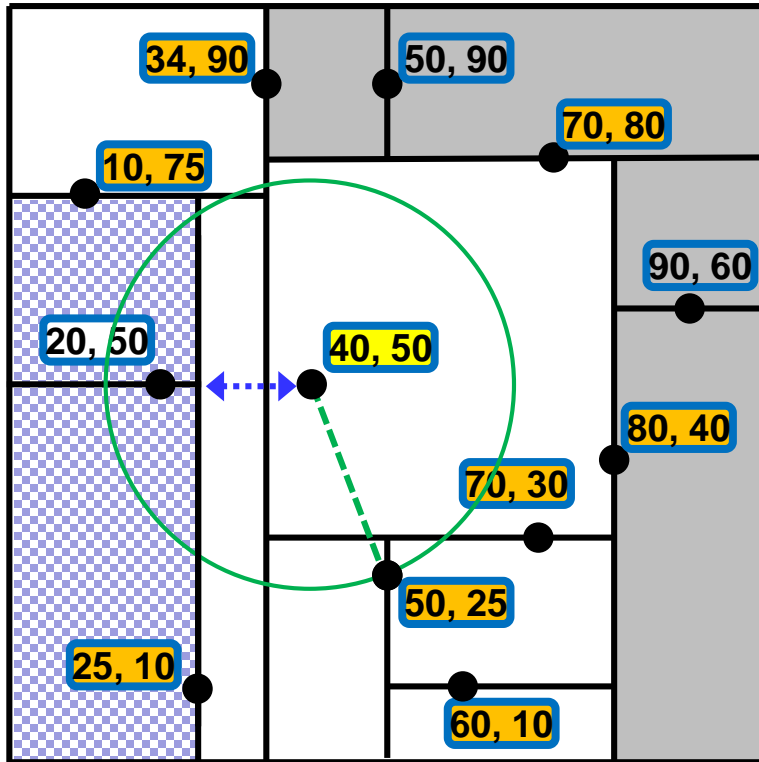
Pruning?: The the distance from  $Q = [40, 50]$  to the (hyper) rectangle  $r_9$  associated with the left subtree of  $[10, 75]$  is  $6.0 < 26.926$ . No pruning occurs.

The search continues in the left subtree of  $[10, 75]$ .

Searched nodes



Find Nearest Neighbour to [40, 50]

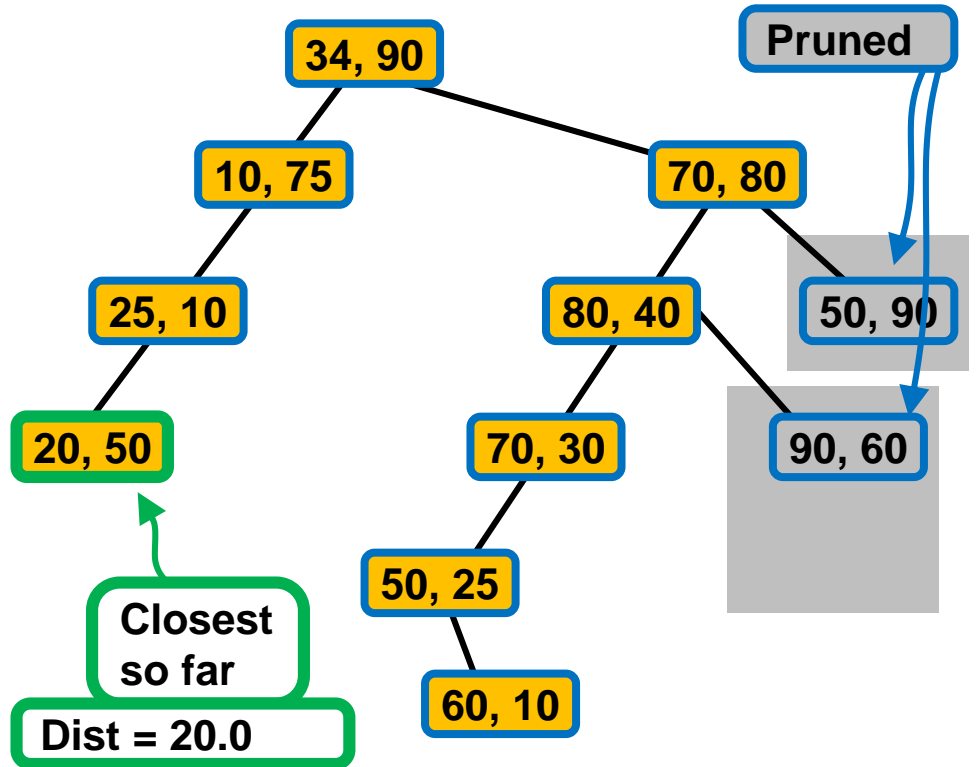
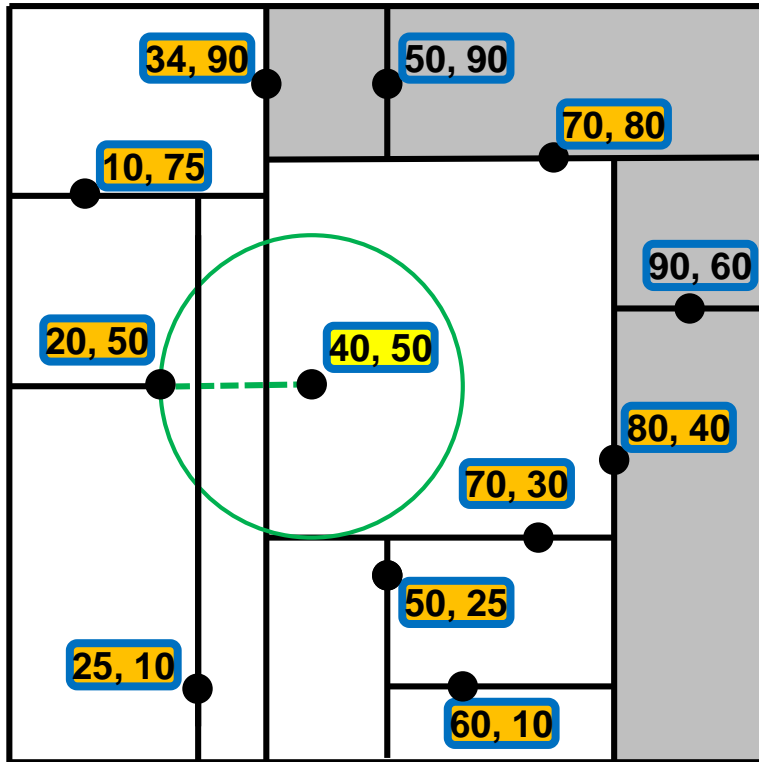


Distance (Q, [25, 10] ) = 42.72 > 26.926.

Pruning?: The the distance from Q = [40, 50] to the (hyper) rectangle r10 associated with the left subtree of [25, 10] is 15.0 < 26.926. No pruning occurs. The search continues in the left subtree of [25, 10].

Searched nodes

Find Nearest Neighbour to [40, 50]



Distance  $(Q, [20, 50]) = 20.0 < 26.926$ .  $[20, 50]$  becomes new *close* node. The search returns to the root and terminates.

Searched nodes

```
NNres nn(point q, Node t, int cd, HypRec r, NNres close) {  
    if (t == null) return close;           // out of tree  
    if (G.distance(q, r) >= close.dist)  
        return close;                       // cell of t is too far from q  
    Number dist = G.distance(q, t.coords);  
    if (dist < close.dist)                 // upd close if necessary  
        { close.coords = t.coords; close.dist = dist; }  
    if (q[cd] < t.coords[cd] {              // q closer to L child  
        close = nn(q, t.left, (cd+1)%D,  
                    r.trimLeft(cd, t.coords), close);  
        close = nn(q, t.right, (cd+1)%D,  
                    r.trimRight(cd, t.coords), close);  
    } else {                                // q closer to R child  
        close = nn(q, t.right, (cd+1)%D,  
                    r.trimRight(cd, t.coords), close);  
        close = nn(q, t.left, (cd+1)%D,  
                    r.trimLeft(cd, t.coords), close);  
    }  
    return close;  
}
```

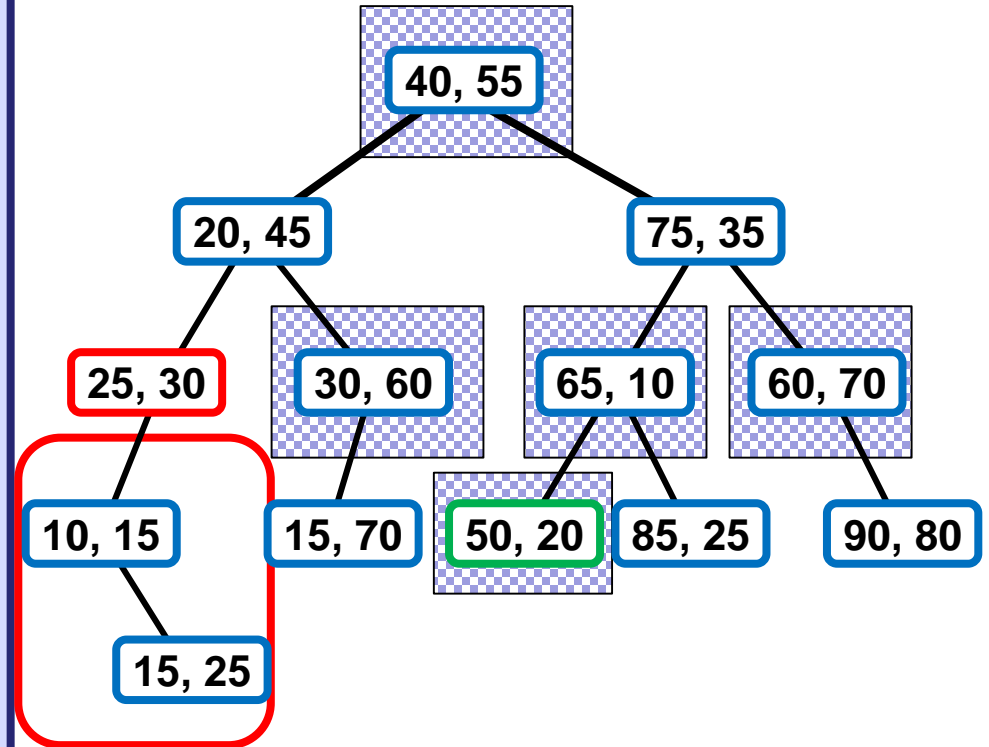
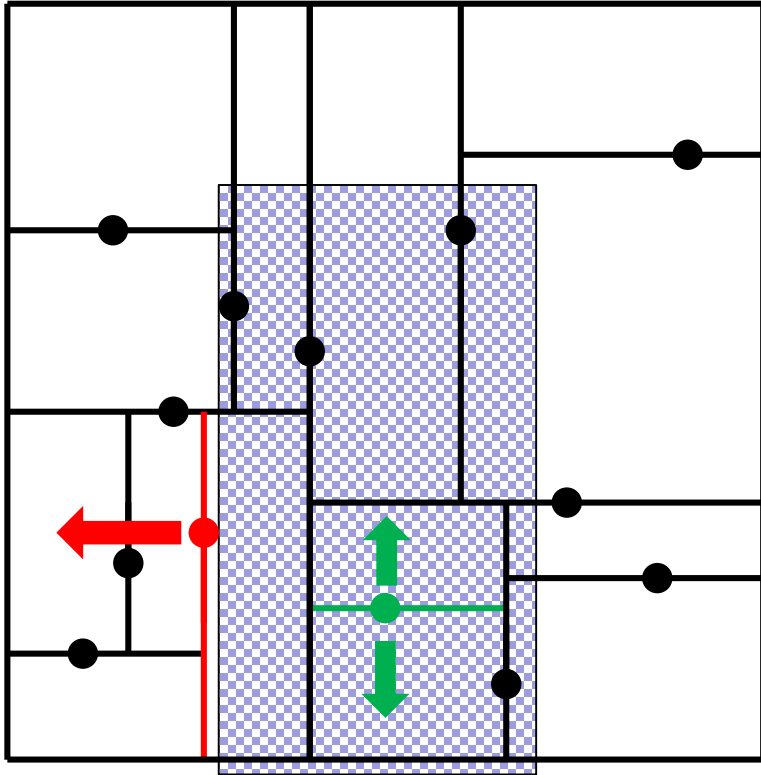
## Complexity of Nearest Neighbour search

Complexity of Nearest Neighbour search might be close to  $O(n)$  when data points and query point are unfavorably arranged. However, this happens only when:

- A. The dimension  $D$  is relatively high, 7,8... and more, 10 000 etc... , or
- B. The arrangement of points in low dimension  $D$  is very special (artificially constructed etc.).

Expected time of NN search is close to  
 **$O(2^D + \log n)$**   
with uniformly distributed data.

Thus it is effective only when  $2^D$  is significantly smaller than  $n$ .



Complexity of Range search

$$O(D \cdot n^{1-\frac{1}{D}})$$