prob 7.

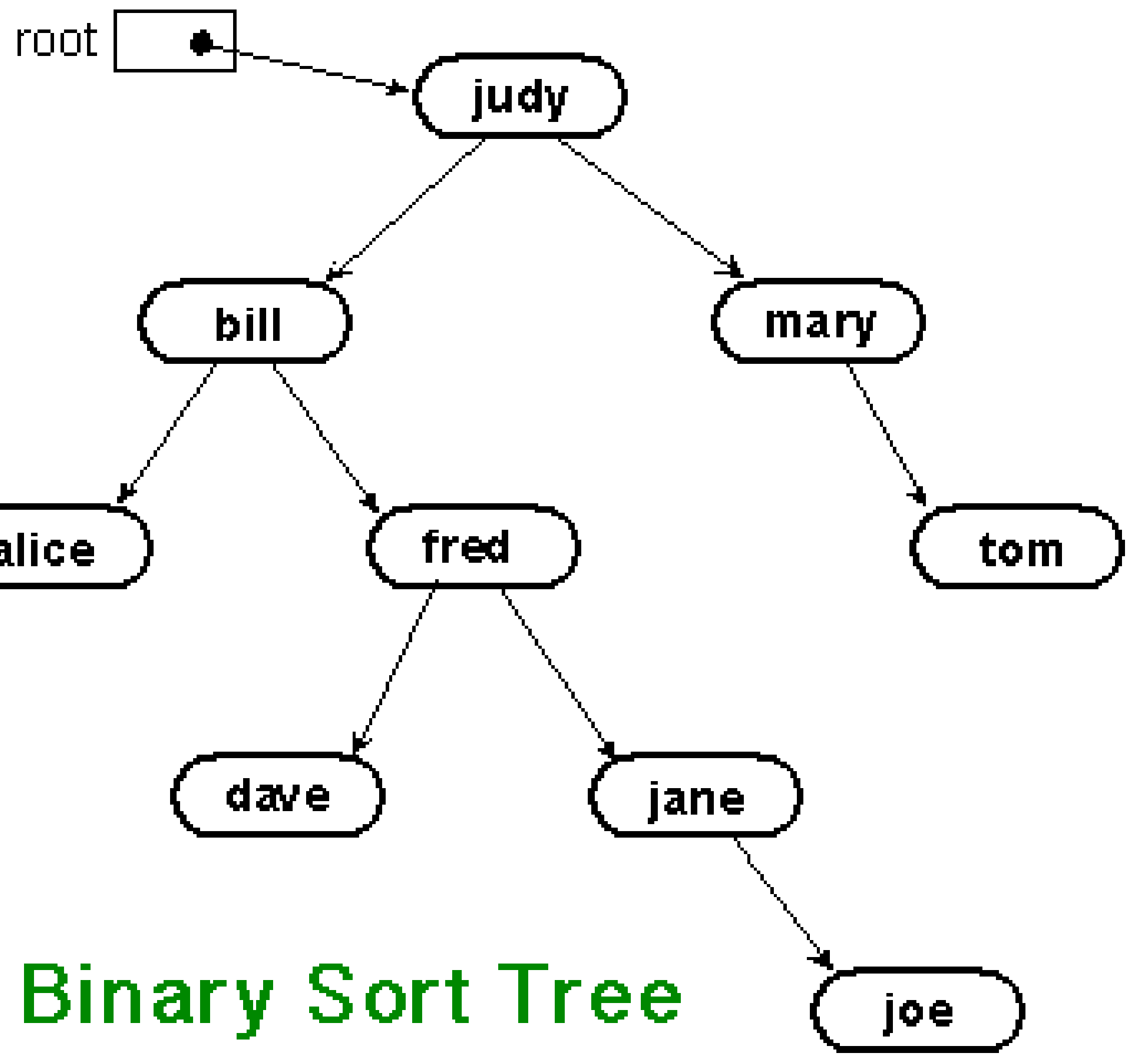
Optimal Binary Search Tree

- Suppose we are implementing a dictionary for words. The dictionary could be meaning of words, or reference to related information.
- Searching for a word in arbitrary placed words is O(n).
- A binary search tree reduces the time for individual search to O(log n).
- However, this will be true, if it is a balanced BST, otherwise it could be up to O(n).
- Note, the BST tree can be organized in number of ways.

- We need the BST tree that serves our purpose.
- What is the purpose of building BST?
- To search for words.
- If the frequently searched words are near the root, then it is okay.
- If those words are towards bottom of the tree, then overall search time is going to be very high.
- Is there a way to optimize it?

- Judy searched 100 times
- Joe searched 10 times

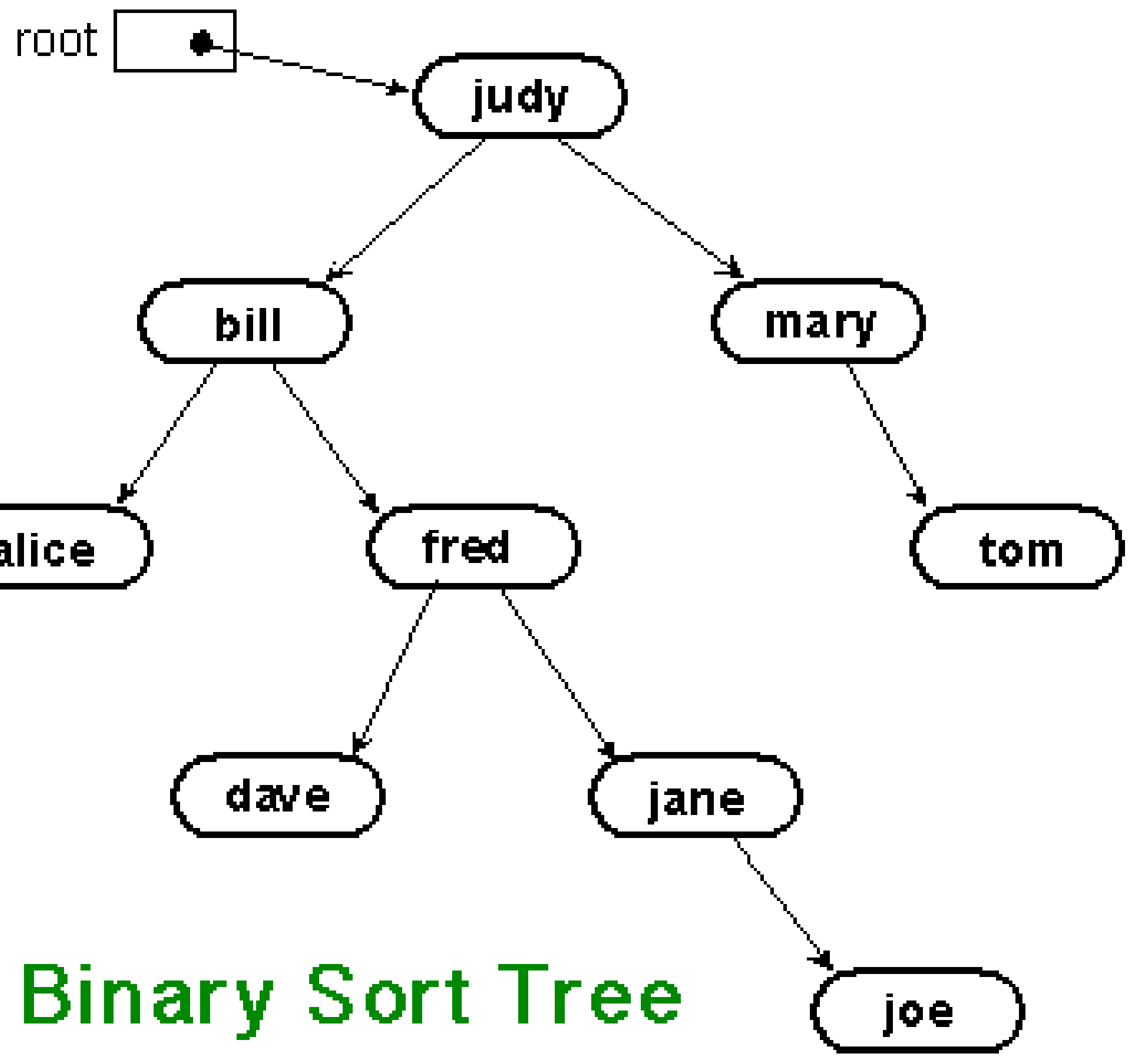
OKAY



A Binary Sort Tree

- Judy searched 10 times
- Joe searched 100 times

Not Okay



A Binary Sort Tree

- We want to create such a BST tree
- where more frequently searched words appear towards the top

- We call it an OPTIMAL BST

- The OPTIMAL BST tree is one, where
- not only *alphabetical order*, but
- *frequency of search* also is taken into account.

- Consider a case where we have only 3 elements
- keys = { 10, 12, 20 } with search freq. { 34, 8, 50 }
- There can be 5 following possible BSTs
- 10 12 20 10 20
- \ / \ / \ /
- 12 10 20 12 20 10
- \ / / \
- 20 10 12 12
- I II III IV V

- Cost for tree II: 8+ 68 + 100 = 176
- Cost for tree V: 50+ 68 + 24 = 142

- A collection of words leads to a very large number of possible BSTs.
- Suppose we try to find cost of all possible BST trees,
- The time complexity of searching for best BST, may turn out to have exponential complexity.

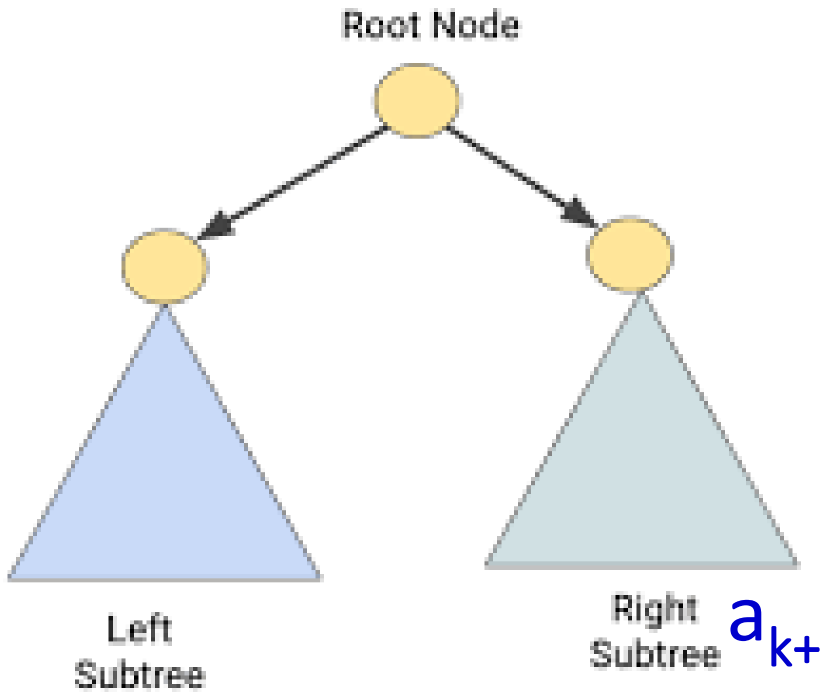
$$\frac{{}^{2n}C_n}{n+1}$$

- So we use Dynamic Programming to solve this problem.

- Let a_1 , a_2, ... , a_n be the nodes of a BST, arranged in ascending order, and let p_1 , p_2, ... , p_n be the probabilities of searching these items.

- Suppose a_k is the root node of the BST,
- nodes a_1, . .. , a_{k-1} are in the left subtree and
- a_{k+1}, . .. , a_n are in the right subtree.

a_k



$a_1, \ldots, a_{k-1}$

$a_{k+1}, \ldots, a_n$

- Let the average search time to process the left subtree is C[1. . . k-1] plus p_1 , p_2, ... , p_{k-1} to process the items in the root.

- By same logic, search time for right subtree is

  C[k+1…n] plus p_{k+1} , . . . , p_n .

- Given n elements, we have to construct a 2Dimensional cost matrix C[i , j],
- whose rows correspond to i = 1 , 2, 3, . . ., n+1
- and whose columns correspond to j = 0, 1, 2, . . . , n

- Cost of tree from node i to node j. k is the root node which splits the nodes in two parts. We need to figure out best value of k.
- $C[\, i, j\,] = \min \{C[i, k\text{-}1] + C[k\text{+}1\,, j]\} + \sum p_i$.
- $\quad\quad = \min \{C[i\,, k\text{-}1] + C[k\text{+}1, j]\} + p_i + \ldots + p_j$.
- $C[\, i\,, i\,] = p_i$.
- All diagonal entries are zero
- $C[\, i, i–1] = 0$.

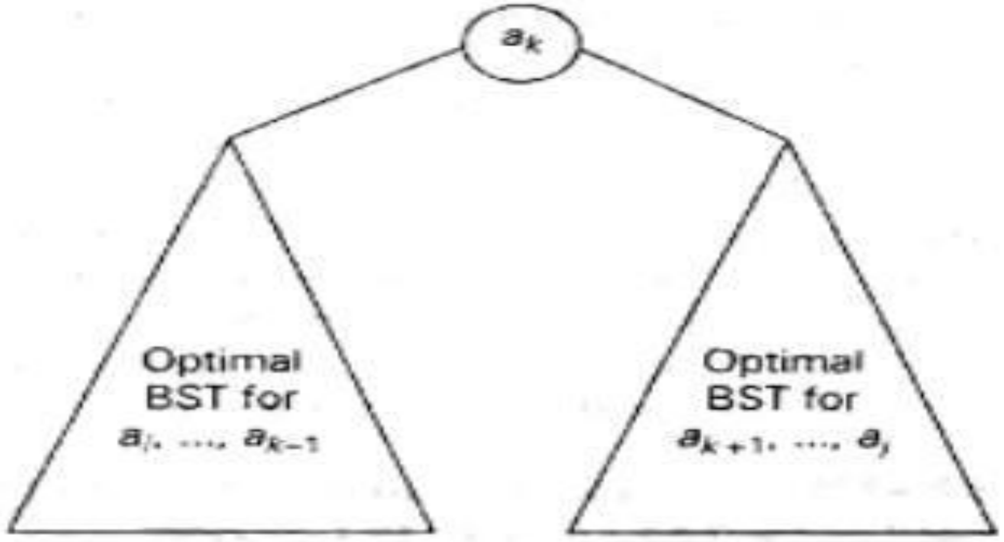


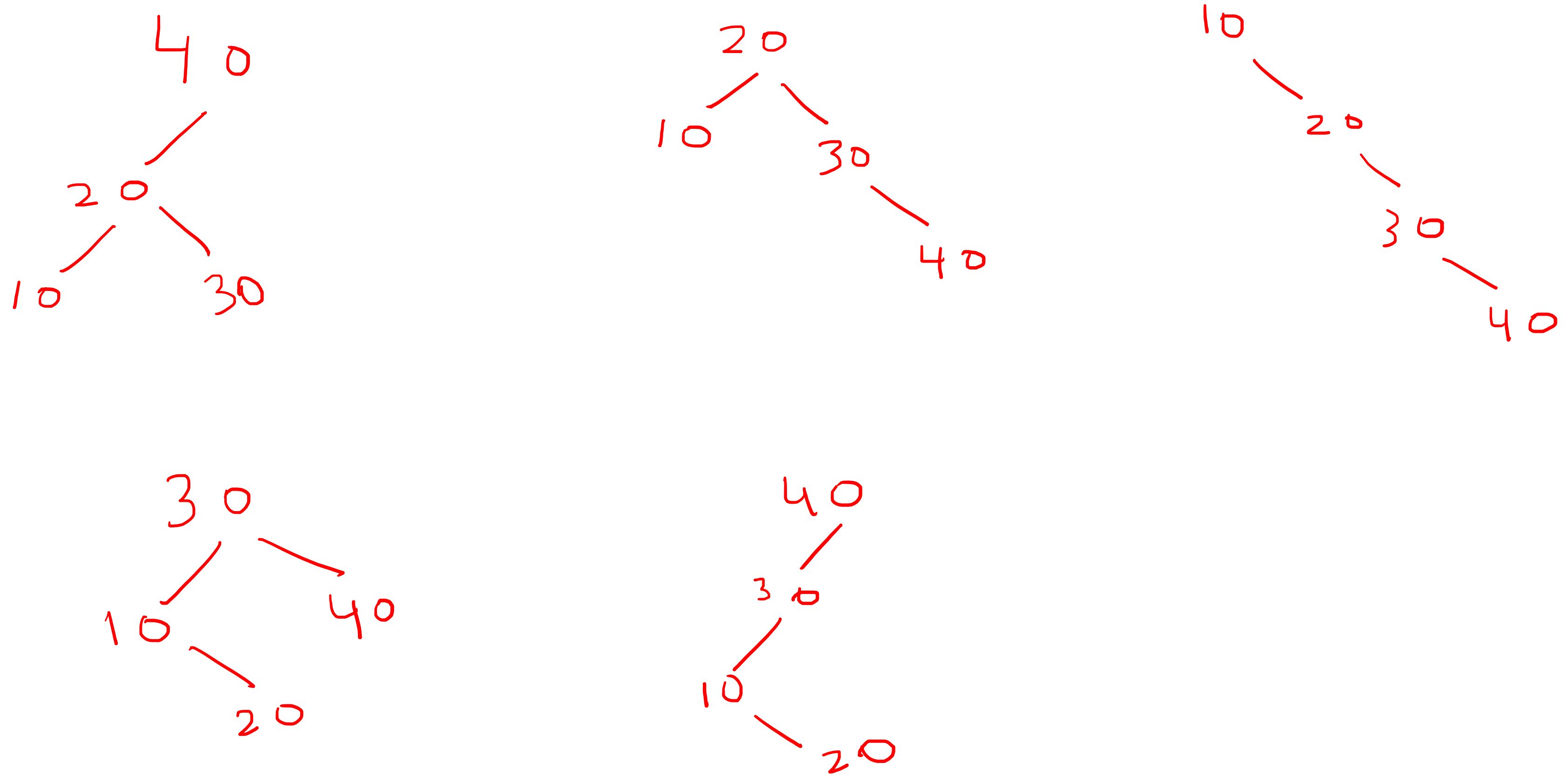
Fig: Binary search tree with root a_k and two optimal binary search subtrees and

- Example: Consider a 4 node tree with

	1	2	3	4
Keys →	10	20	30	40
Frequency →	4	2	6	3

- Note, key values are put in increasing order

- 14 trees are possible, we have to find min cost tree



- Brute force technique would involve examining each tree to find min cost tree.

- As n increases, possible number of BST will keep on growing.
- For n=6, the number is 132

- So Brute force technique will be cumbersome for bigger trees.

.

Dynamic Programming Approach for BST

.

- Let us now take up the DP approach.
- We need 2 matrices, the COST matrix (based on Frequency)
- and the BST order matrix (based on Key values)

	1	2	3	4
Keys →	10	20	30	40
Frequency →	4	2	6	3

- We create a COST matrix C[,] using dynamic programming
- $C[i\ , i\ ] = p_i$.
- $C[i, i - 1] = 0$.
- Each C[i, j] is computed for all possible values of k and taking the minimum
- $C[i,j] \ = \min \{C[i, k-1] + C[k+1\ ,j]\} + p_i + . . + p_j$.

- For our example fill in the cost of the elements (frequency values)
- $C[i \;, i\,] = p_i$.
- $C[i, i-1] = 0$.

	0	1	2	3	4	COST
1	0	4				
2		0	2			
3			0	6		
4				0	3	
5					0	

	1	2	3	4
Keys →	10	20	30	40
Frequency →	4	2	6	3

	0	1	2	3	4	COST
1	0	4				
2		0	2			
3			0	6		
4				0	3	
5					0	

	0	1	2	3	4	BST
1	0	1				
2		0	2			
3			0	3		
4				0	4	
5					0	

- The first value to be filled up is C[1,2]
- So here i = 1 and j = 2.
- k can take two values k=1 and k=2
- In the cost formula
- $C[i,j] = \min\{C[i, k-1] + C[k+1,j]\} + p_i + \ldots + p_j$.
- plug in the values for k=1 and k=2 and take the minimum.

C[1,2] with k =1

- $\{C[1,0] + C[2,2]\} + p_1 + p_2$.
- $= 2 + 4 + 2 = 8$
- C[1,2] with k =2
- $\{C[1,1] + C[3,2]\} + p_1 + p_2$.
- $= 4 + 4 + 2 = 10$

- Minimum for C[1,2] is 8.
- Instead of solving for two k values separately, we can do it in one go , as shown on next slide.

• 0	1	2	3	4	**COST**
1	0	4			
2		0	2		
3			0	6	
4				0	3
5					0

• 0	1	2	3	4	**BST**
1	0	1			
2		0	2		
3			0	3	
4				0	4
5					0

- C[i,j]= min {C[i, k-1] + C[k+1,j]} + p_i+ . . + p_j .
- calculate C[1,2] , k has 2 values
- k =1, k=2
- C[i,j]= min {C[1,0] + C[2,2] , C[1,1] +C[3,2] }

  + p_1+ p_2 .
- = min{ 2,4} + 4 + 2 = 8

- Min cost is found for k = 1,
- Update this information on the corresponding BST matrix

	0	1	2	3	4	**COST**
1	0	4	8			
2		0	2			
3			0	6		
4				0	3	
5					0	

	0	1	2	3	4	**BST**
1	0	1	1			
2		0	2			
3			0	3		
4				0	4	
5					0	

- $C[i,j]= \min \{C[i, k-1] + C[k+1,j]\} + p_i + \ldots + p_j$.
- calculate C[1,2] , k has 2 values
- k =1, k=2
- $C[i,j]= \min \{C[1,0] + C[2,2] ,\ C[1,1] + C[3,2] \}$
  $+ p_1 + p_2$.
- $= \min\{ 2,4\} + 4 + 2 = 8$

- Min cost is found for k = 1,
- Update this information on the corresponding BST matrix

	0	1	2	3	4	**COST**
1	0	4	8			
2		0	2	10		
3			0	6		
4				0	3	
5					0	

	0	1	2	3	4	**BST**
1	0	1	1			
2		0	2	3		
3			0	3		
4				0	4	
5					0	

- C[i,j]= min {C[i, k-1] + C[k+1,j]} + p_i+ . . + p_j .
- .
- calculate **C[2,3]** ,
- k has 2 values
- k =2, k=3
- C[i,j]= min { C[2,1] + C[3,3] ,C[2,2] +C[4,3] }
- + p_2+ p_3 .
- = min{ 0+ 6, 2+0} + 2 + 6
- = 2+8
- = 10

- The min. value is obtained for k =3
- so update BST matrix

COST

	0	1	2	3	4
1	0	4	8		
2		0	2	10	
3			0	6	12
4				0	3
5					0

BST

	0	1	2	3	4
1	0	1	1		
2		0	2	3	
3			0	3	3
4				0	4
5					0

- C[i,j]= min {C[i, k-1] + C[k+1,j]} + p_i+ . . + p_j .
- calculate C[3,4] , k has 2 values
- k =3, k=4
- C[i,j]=min { C[3,2] + C[4,4] ,
  C[3,3] +C[5,4] } + p_3+ p_4 .
- = min{ 0+3, 6+0} + 6 +3.
- = 3+9
- = 12

- The min. value is obtained for k = 3

	0	1	2	3	4	**COST**
1	0	4	8	20		
2		0	2	10		
3			0	6	12	
4				0	3	
5					0	

	0	1	2	3	4	**BST**
1	0	1	1	3		
2		0	2	3		
3			0	3	3	
4				0	4	
5					0	

- Calculate C[1,3] , k has 3 values
- k =1, k=2, k = 3
- =min { C[1,0] + C[2,3] ,
- C[1,1] + C[3,3],
- C[1,2] +C[4,3] } + p_1+ p_2 + p_3..
- = min{ 0+10,
- 4+6,
- 8+0 } + 4 + 2 + 6.
- = 8 + 12
- = 20

	0	1	2	3	4	**COST**
1	0	4	8	20		
2		0	2	10	16	
3			0	6	12	
4				0	3	
5					0	

	0	1	2	3	4	**BST**
1	0	1	1	3		
2		0	2	3	3	
3			0	3	3	
4				0	4	
5					0	

- calculate C[2,4] , k has 3 values
- k=2, k = 3, k =4
- =min { C[2,1] + C[3,4] ,
- C[2,2] + C[4,4],
- C[3,4] +C[5,4] } + p_2 + p_3 + p_4
- = min{ 0+12,
- 2+3,
- 12+0 } + 2 + 6 +3.
- = 5 + 11 = 16.
- Min value is obtained for k=3

COST

	0	1	2	3	4
1	0	4	8	20	26
2		0	2	10	**16**
3			0	6	12
4				0	3
5					0

BST

	0	1	2	3	4
1	0	1	1	3	3
2		0	2	3	3
3			0	3	3
4				0	4
5					0

- calculate C[1,4] , k has 4 values
- k = 1, k=2, k = 3, k =4
- C =min { C[1,0] + C[2,4] ,
- C[1,1] + C[3,4],
- C[1,2] +C[4,4],
- C[1,3]+C[5,4] }

  $+ p_1 + p_2 + p_3 + p_4$

- C= min { 0+16, 4+12, 8+3, 20+0} +4+2+6+3.
- = 11 + 15 = 26.
- Again best value obtained for k = 3

•	0	1	2	3	4	COST
1	0	4	8	20	26	
2		0	2	10	**16**	
3			0	6	12	
4				0	3	
5					0	

•	0	1	2	3	4	BST
1	0	1	1	3	3	
2		0	2	3	3	
3			0	3	3	
4				0	4	
5					0	

	1	2	3	4
Keys →	10	20	30	40
Frequency →	4	2	6	3

- We have now solved the problem using DP approach.
- To draw the final optimal BST, we make use of the BST matrix.

- Now to work out the structure of optimal BST.
- Note BST(1,4) is 3, so key 3 is the root (value 30)
- key 4 is greater than key 3, so forms right of 30
- R(1,2) is 1, so key 1 forms the root of left subtree of key 3.
- key 2 will be right child of key 1

- Let us verify the cost of optimal BST.
- 6*1 + 4*2 + 3*2 + 2*3
- = 6+8+6+6 = 26