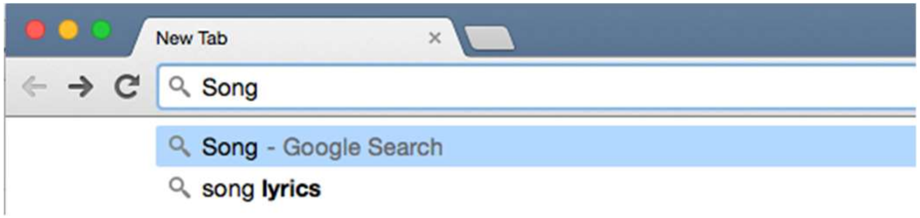
CSE 100: MARKOV, AVL

PA2 background

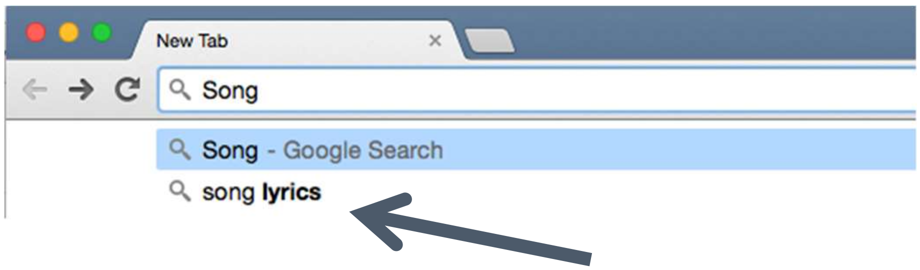
- Explain how Markov processes can be used to generate text
- Distinguish between training on text and generating text
- Create Markov models to generate text

Note: Punctuation is part of a word in my examples here. In the assignment, you may be asked to separate out "," "." "!" among others.

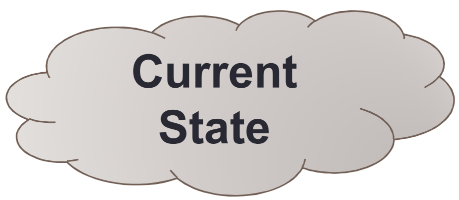
Predicting the future



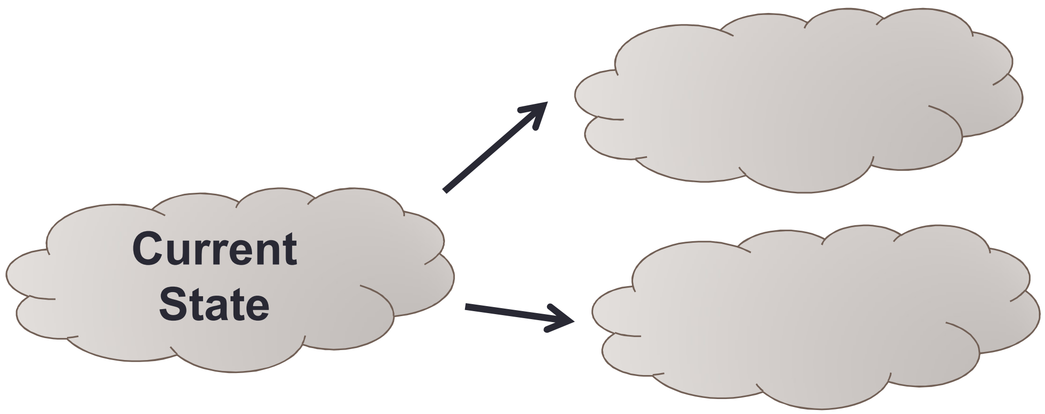
Predicting the future

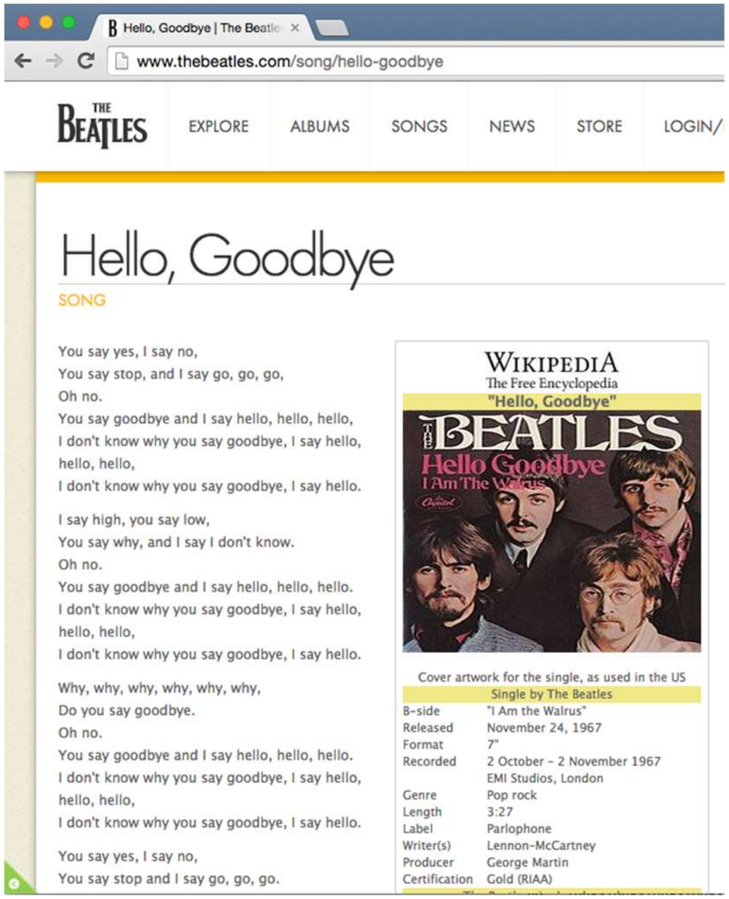


Predicting the future: Markov processes



Predicting the future: Markov processes




Hello, Goodbye | The Beatles

www.thebeatles.com/song/hello-goodbye

THE BEATLES | EXPLORE | ALBUMS | SONGS | NEWS | STORE | LOGIN/

Hello, Goodbye

SONG

You say yes, I say no,
You say stop, and I say go, go, go,
Oh no.
You say goodbye and I say hello, hello, hello,
I don't know why you say goodbye, I say hello,
hello, hello,
I don't know why you say goodbye, I say hello.

I say high, you say low,
You say why, and I say I don't know.
Oh no.
You say goodbye and I say hello, hello, hello.
I don't know why you say goodbye, I say hello,
hello, hello,
I don't know why you say goodbye, I say hello.

Why, why, why, why, why, why,
Do you say goodbye.
Oh no.
You say goodbye and I say hello, hello, hello.
I don't know why you say goodbye, I say hello,
hello, hello,
I don't know why you say goodbye, I say hello.

You say yes, I say no,
You say stop and I say go, go, go.

WIKIPEDIA
The Free Encyclopedia

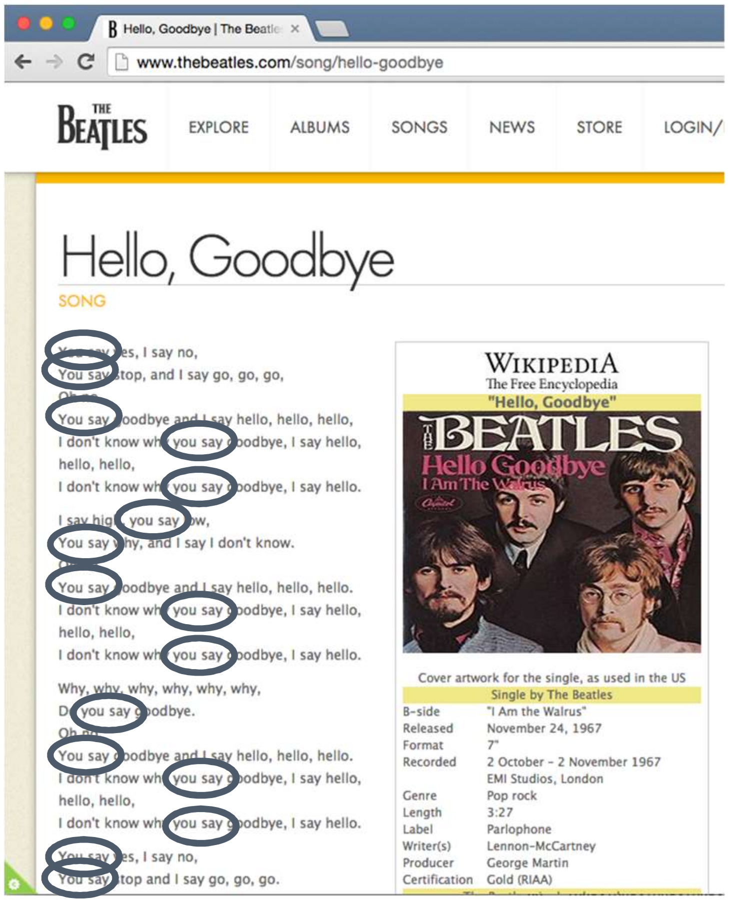
"Hello, Goodbye"

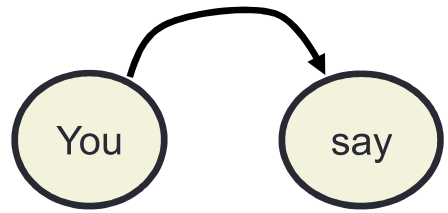
THE BEATLES
Hello Goodbye
I Am The Walrus

Cover artwork for the single, as used in the US

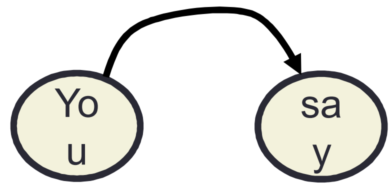
Single by The Beatles	
B-side	"I Am the Walrus"
Released	November 24, 1967
Format	7"
Recorded	2 October – 2 November 1967 EMI Studios, London
Genre	Pop rock
Length	3:27
Label	Parlophone
Writer(s)	Lennon-McCartney
Producer	George Martin
Certification	Gold (RIAA)

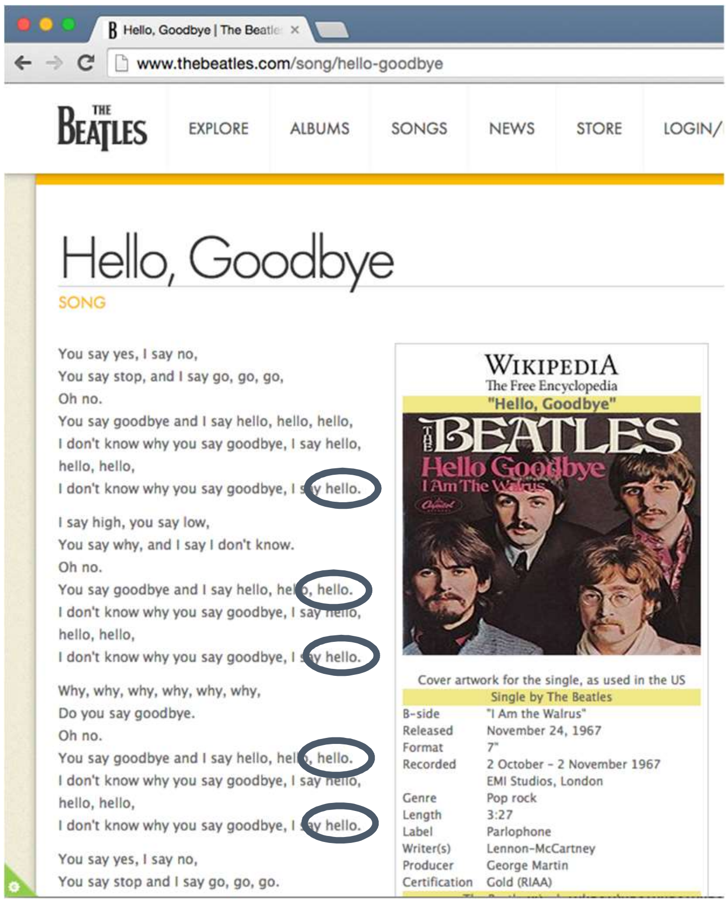
You say hello. I don't know why you say hello, hello. I say goodbye. Oh no. You say no, You


Hello, Goodbye | The Beatles
www.thebeatles.com/song/hello-goodbye
THE BEATLES
EXPLORE
ALBUMS
SONGS
NEWS
STORE
LOGIN/
Hello, Goodbye
SONG
WIKIPEDIA
The Free Encyclopedia
"Hello, Goodbye"
BEATLES
Hello Goodbye
I Am The Walrus
Cover artwork for the single, as used in the US
Single by The Beatles
B-side "I Am the Walrus"
Released November 24, 1967
Format 7"
Recorded 2 October – 2 November 1967
EMI Studios, London
Genre Pop rock
Length 3:27
Label Parlophone
Writer(s) Lennon–McCartney
Producer George Martin
Certification Gold (RIAA)


You
say

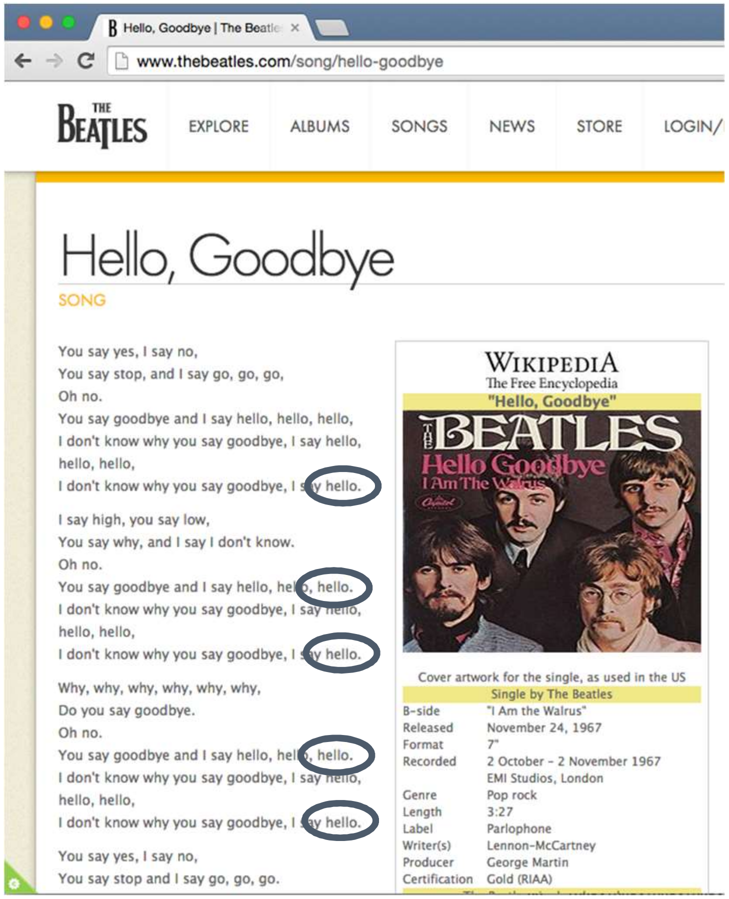
You say hello. I don't know why you say hello, hello. I say goodbye. Oh no. You say no, You

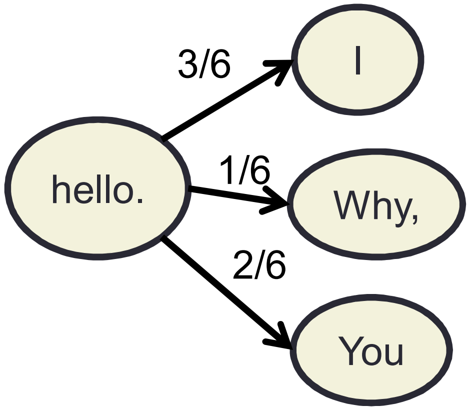




Suppose "hello." is followed by:
"I", "I", "Why", "I", "You", and "You"

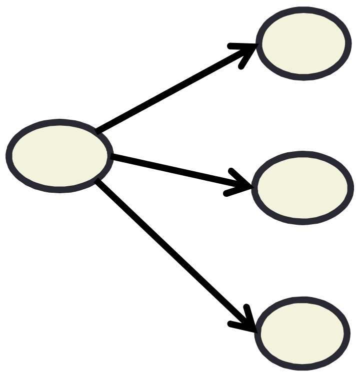
What should be the Markov Model for the word "hello."?


Hello, Goodbye | The Beatles
www.thebeatles.com/song/hello-goodbye
THE BEATLES
EXPLORE
ALBUMS
SONGS
NEWS
STORE
LOGIN/
Hello, Goodbye
SONG
You say yes, I say no,
You say stop, and I say go, go, go,
Oh no.
You say goodbye and I say hello, hello, hello,
I don't know why you say goodbye, I say hello,
hello, hello,
I don't know why you say goodbye, I say hello.
I say high, you say low,
You say why, and I say I don't know.
Oh no.
You say goodbye and I say hello, hello, hello.
I don't know why you say goodbye, I say hello,
hello, hello,
I don't know why you say goodbye, I say hello.
Why, why, why, why, why, why,
Do you say goodbye.
Oh no.
You say goodbye and I say hello, hello, hello.
I don't know why you say goodbye, I say hello,
hello, hello,
I don't know why you say goodbye, I say hello.
You say yes, I say no,
You say stop and I say go, go, go.
WIKIPEDIA
The Free Encyclopedia
"Hello, Goodbye"
THE BEATLES
Hello Goodbye
I Am The Walrus
Cover artwork for the single, as used in the US
Single by The Beatles
B-side "I Am the Walrus"
Released November 24, 1967
Format 7"
Recorded 2 October – 2 November 1967
EMI Studios, London
Genre Pop rock
Length 3:27
Label Parlophone
Writer(s) Lennon-McCartney
Producer George Martin
Certification Gold (RIAA)


hello.
I
Why,
You
3/6
1/6
2/6

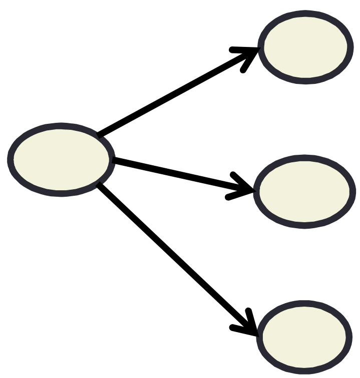
Stage 1: Train

Build model based on data



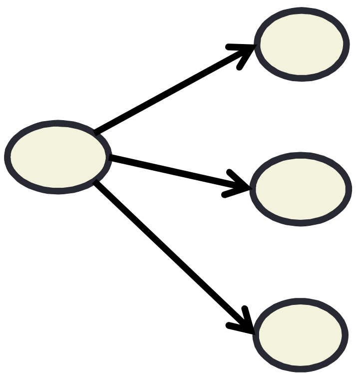
Stage 1: Train

Build model based on ~~data~~ input String



Stage 1: Train

Build model based on ~~data~~ input String



"You say yes, I say no, You say …"

Stage 1: Train

Build model based on ~~data~~ input String

"You say yes, I say no, You say …"

For each new word: need to keep track of word + next words

Stage 1: Train

"You"	"say"	"yes,"	"I"	"say"	"no,"	"You"	"say"

For each new word: need to keep track of word + next words

Stage 1: Train

For each new word: need to keep track of word + next words

"You"	"say"	"yes,"	"I"	"say"	"no,"	"You"	"say"

`word` → "You"

`nextWords` → "say"

Stage 1: Train

For each new word: need to keep track of word + next words

"You"	"say"	"yes,"	"I"	"say"	"no,"	"You"	"say"

`word` → "You"
`nextWords` → "say"

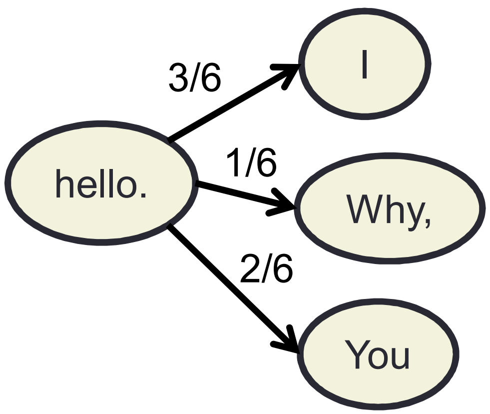
`word` → "say"
`nextWords` → "yes,"

Keep going through the next "say" to show duplicates

What about probabilities?

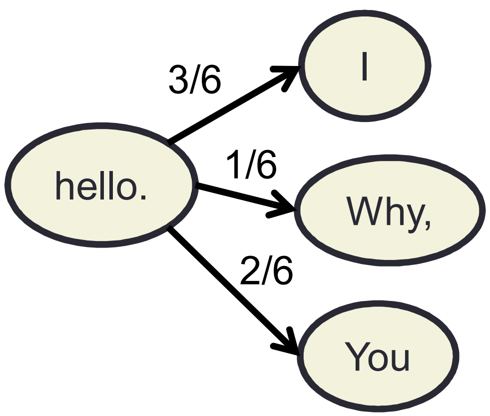
Stage 1: Train

"I don't know why you say goodbye, I say hello. I say high, you say low, …



Stage 1: Train

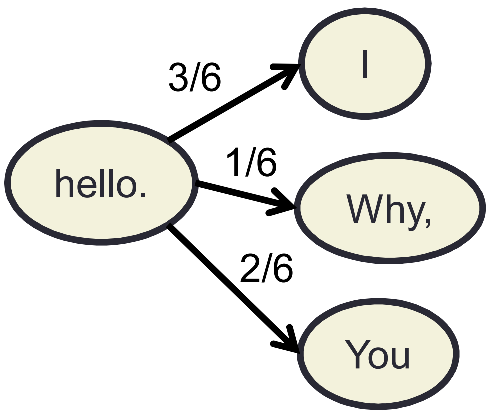
… I say hello, hello, hello. I don't know why you say goodbye …



What about probabilities?

Stage 1: Train

… I say hello. Why, why, why, why, why, why, …



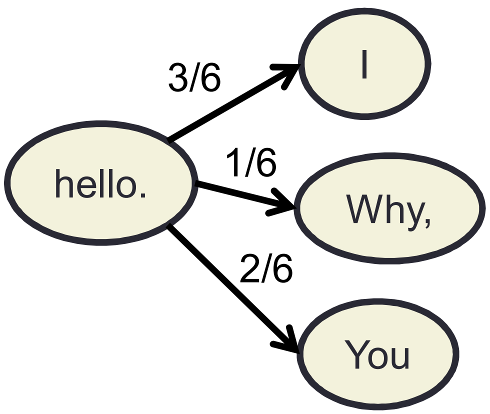
`word` → "hello."

`nextWords` → "I" , "I" , "Why,"

What about probabilities?

Stage 1: Train

… hello, hello, hello. I don't know …



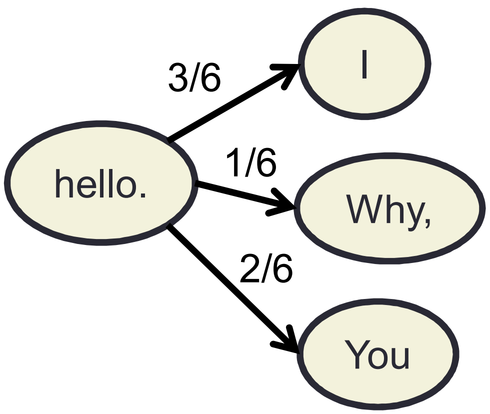
`word` → **"hello."**

`nextWords` → **"I" , "I" , "Why," , "I"**

What about probabilities?

Stage 1: Train

... I say hello. You say yes, I say no, …



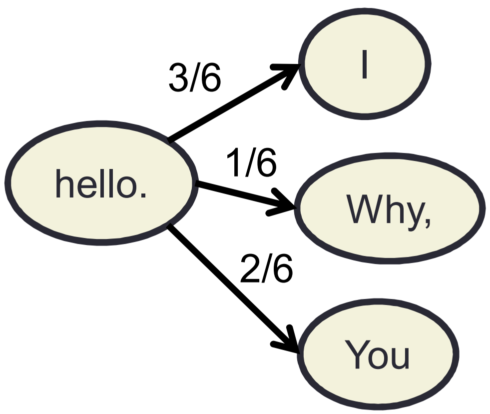
word → "hello."

nextWords → "I" , "I" , "Why," , "I" , "You"

What about probabilities?

Stage 1: Train

etc.

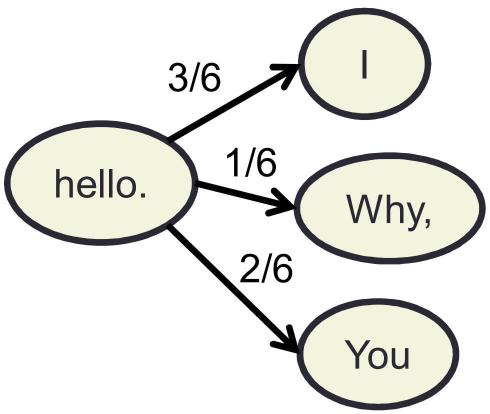


`word` → "hello."

`nextWords` → "I" , "I" , "Why," , "I" , "You" , "You"

Stage 2: Generate

What about probabilities?



`word` → "hello."

`nextWords` → "I" , "I" , "Why," , "I" , "You" , "You"

Pick Randomly
But we could do better than a list, right?

Stage 2: Generate

Until we have enough words:

- Find current word as the word of some node in present word list
- Generate a random number between 0 and the size of
  nextWords list of this node
- Print the word at that index,
- Repeat

Given the text below, build your structure

"I like my cat. I like to see my dog. My dog likes to see me."

Note – treat a "." as a separate word in this case.

Given the text below, build your structure

"I like my cat. I like to see my dog. My dog likes to see me."

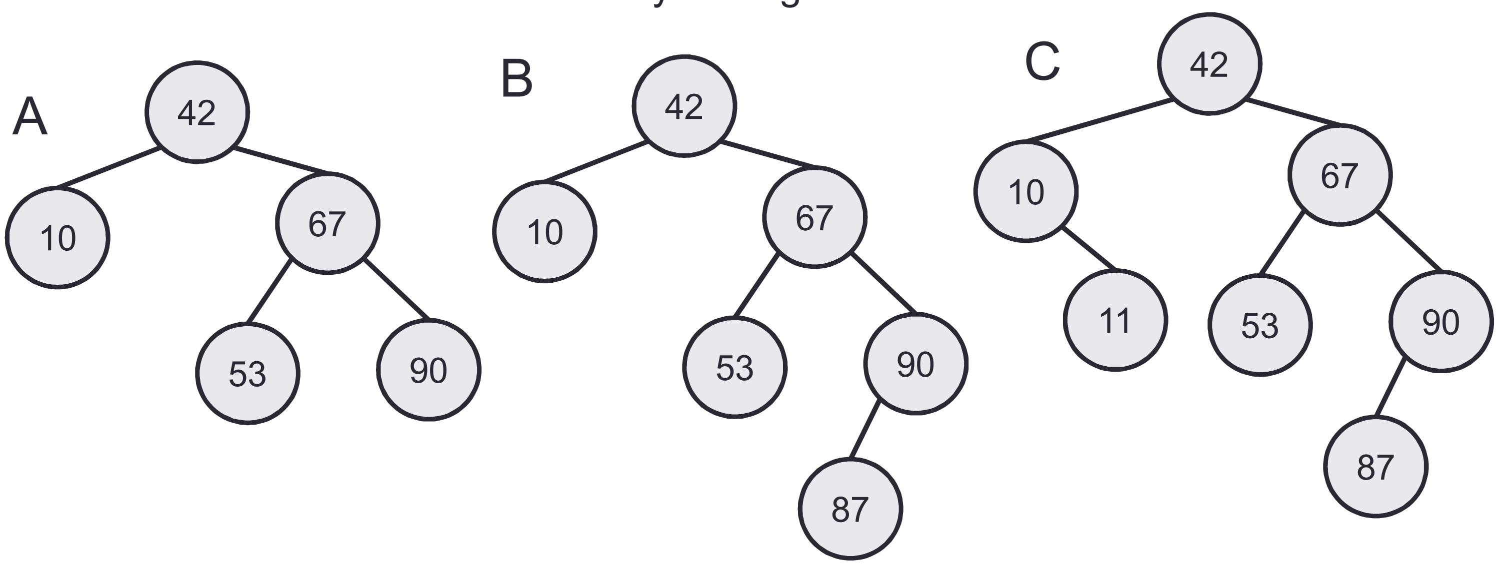
Note – treat a "." as a separate word in this case.

Which of the following Strings could not have been created with the text above?

A. I like to see my cat.
B. I like to see me.
C. I like my cat.
D. My dog likes to see my cat.
E. All of the above are possible.

Which of the following is/are balanced trees?

And thus can become AVL trees by adding the balance factors



D. A&C
E. A&B&C

Annotate the trees with balance factors

An AVL Tree is *worst case* O(logN) to find an element!

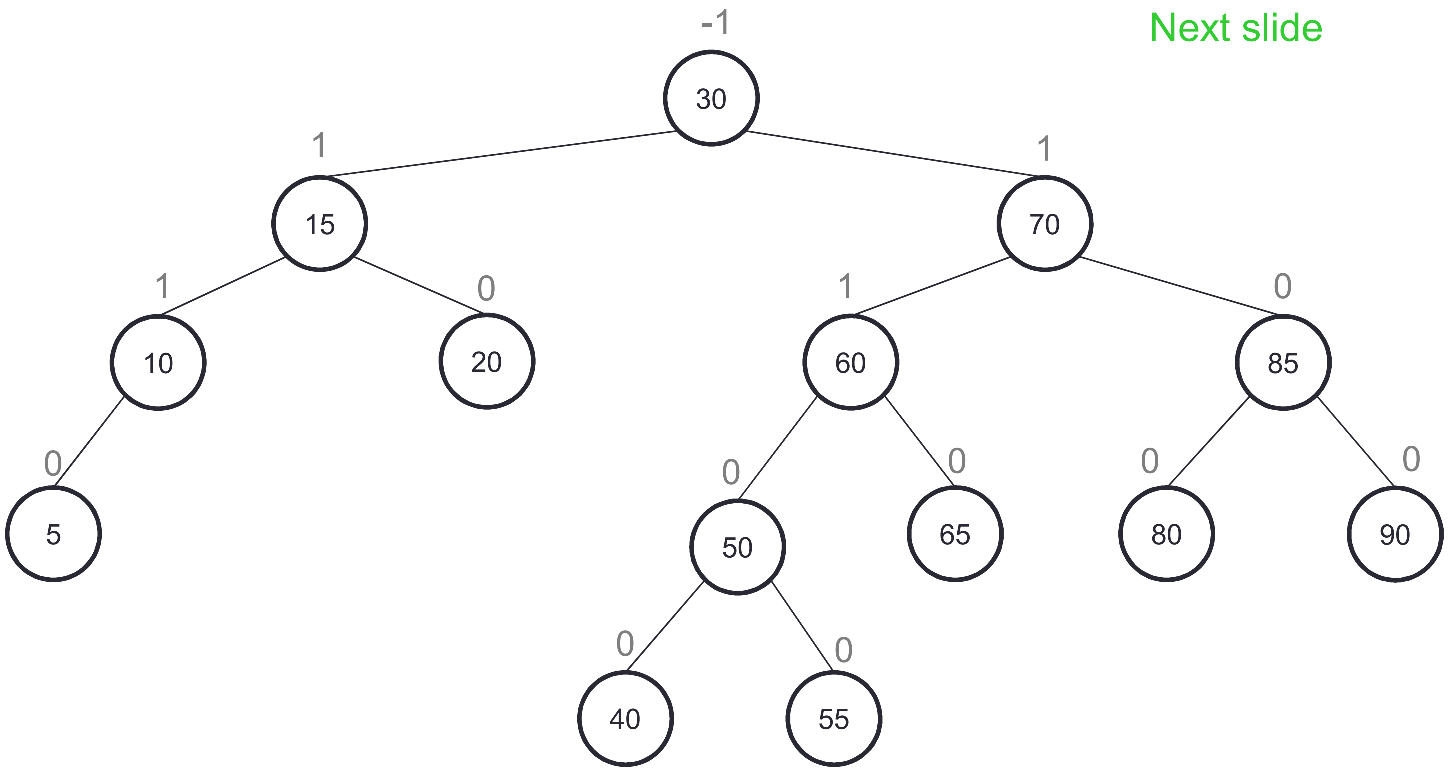
How would you prove this?
Come up with a formula that shows that the height of the tallest AVL tree with N nodes is never bigger than c*logN + k, for some c and k (assuming large N).

The key to this proof is showing that the height stays "small", no matter how legally "unbalanced" the tree is.

But how does the tree stay balanced??

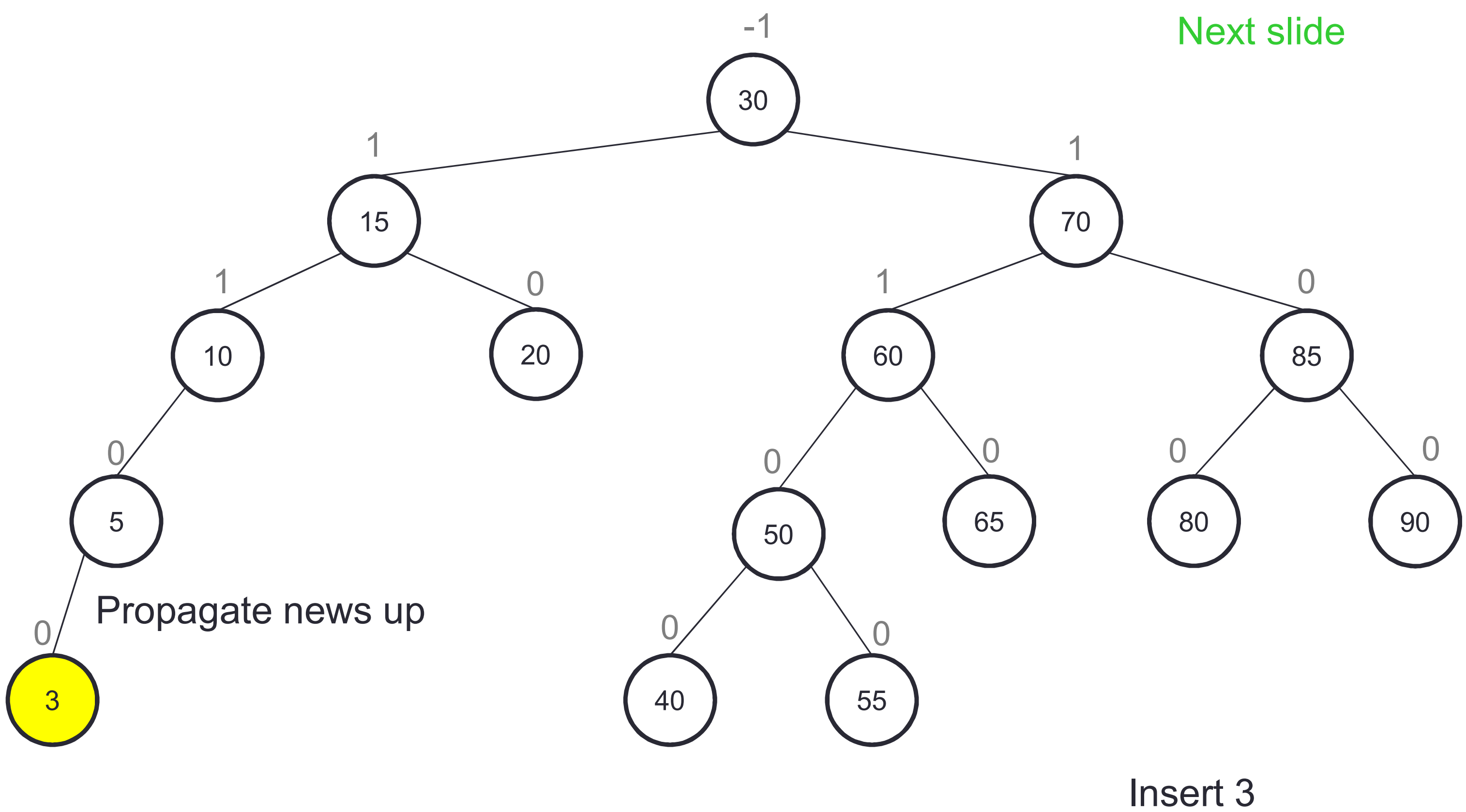
Inserting and rebalancing

Next slide

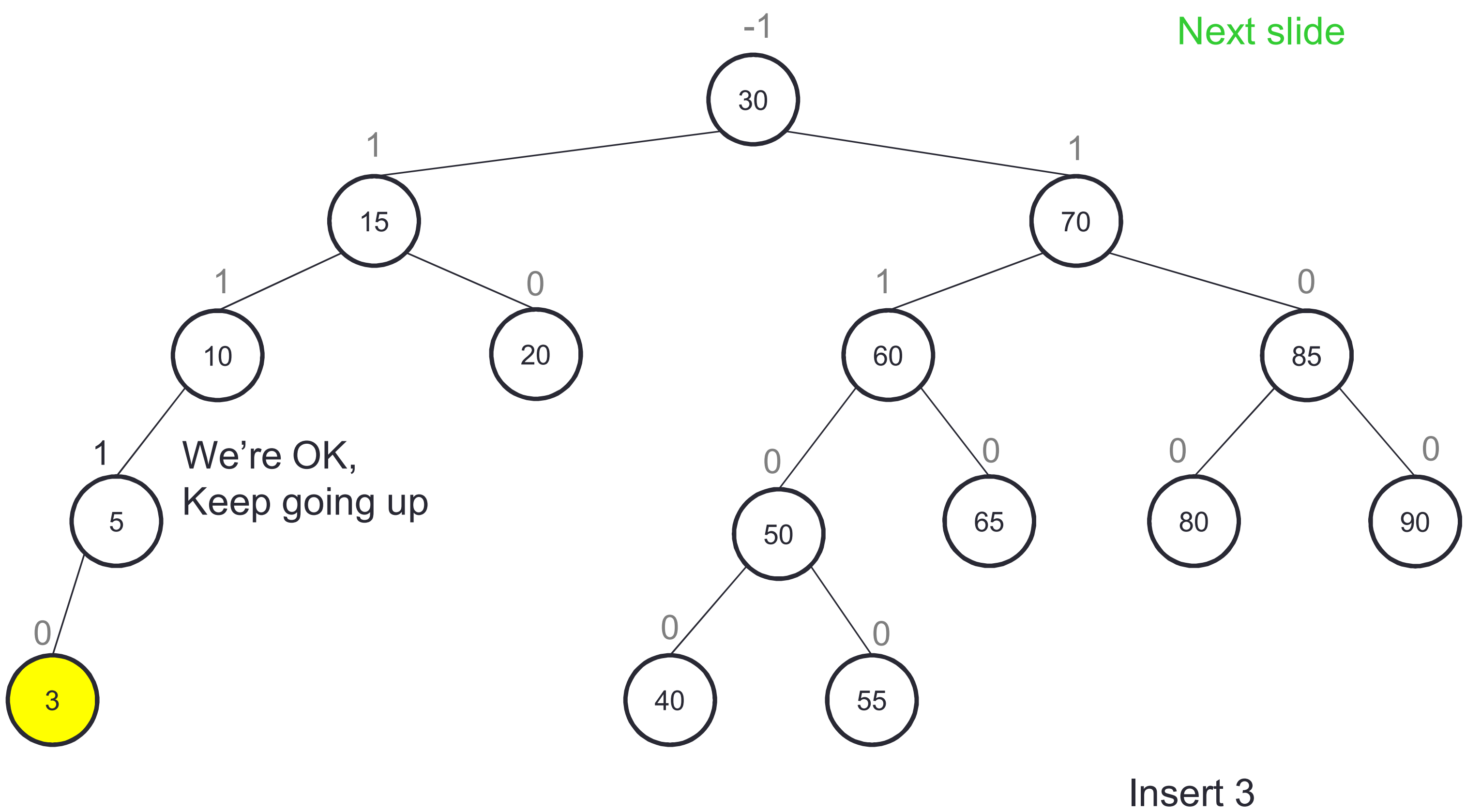


Insert 3

Inserting and rebalancing

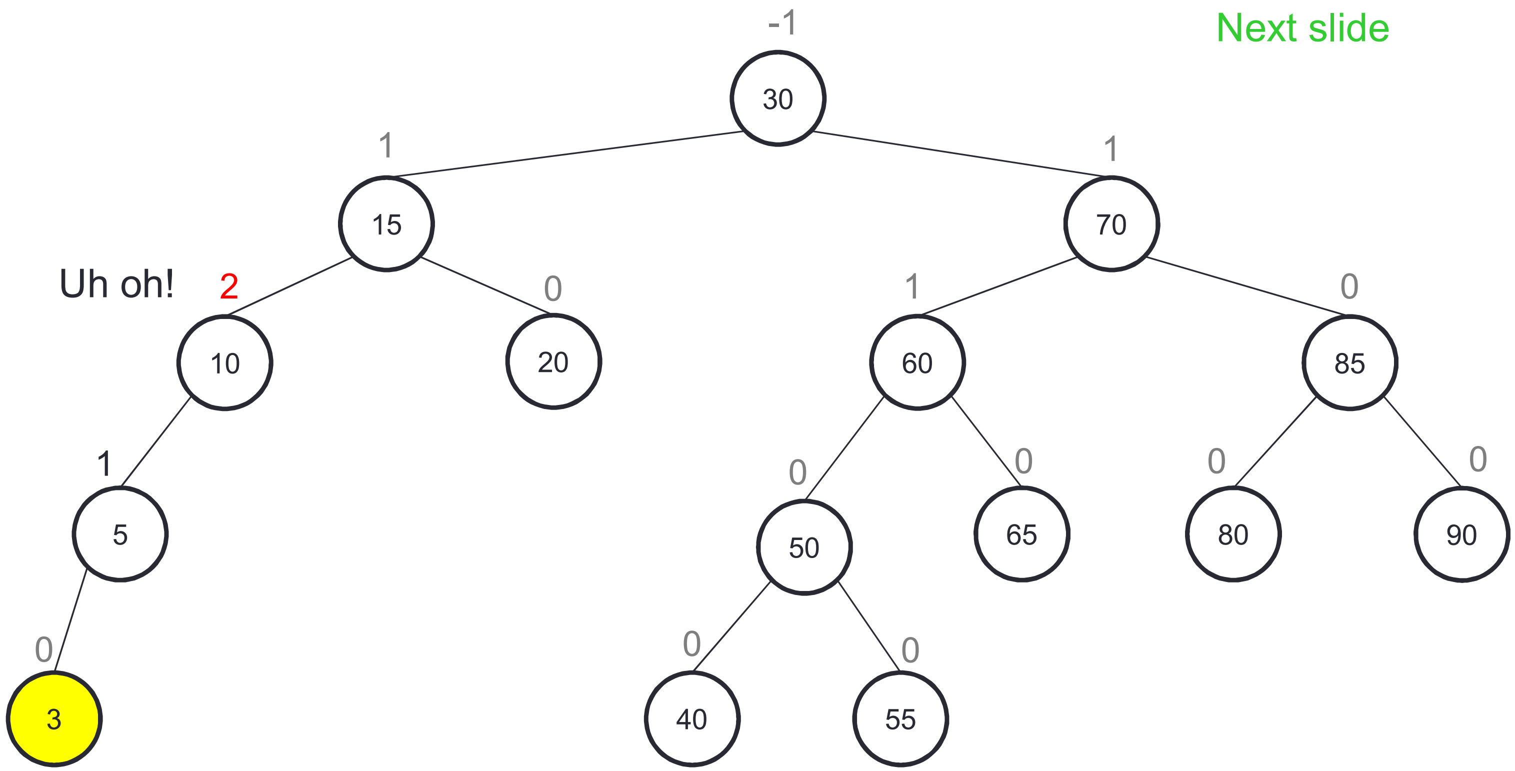


Inserting and rebalancing



Inserting and rebalancing

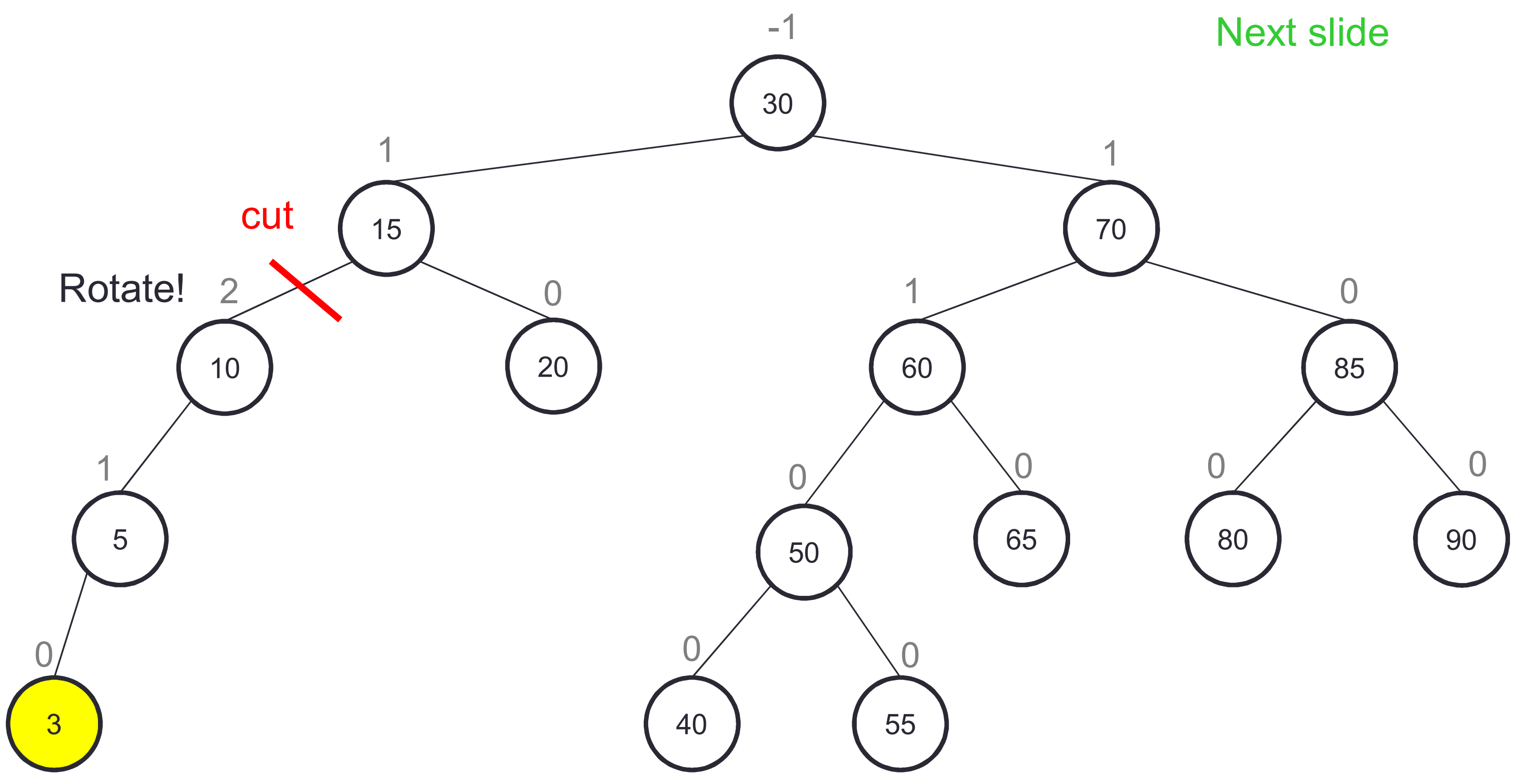
Next slide



Insert 3

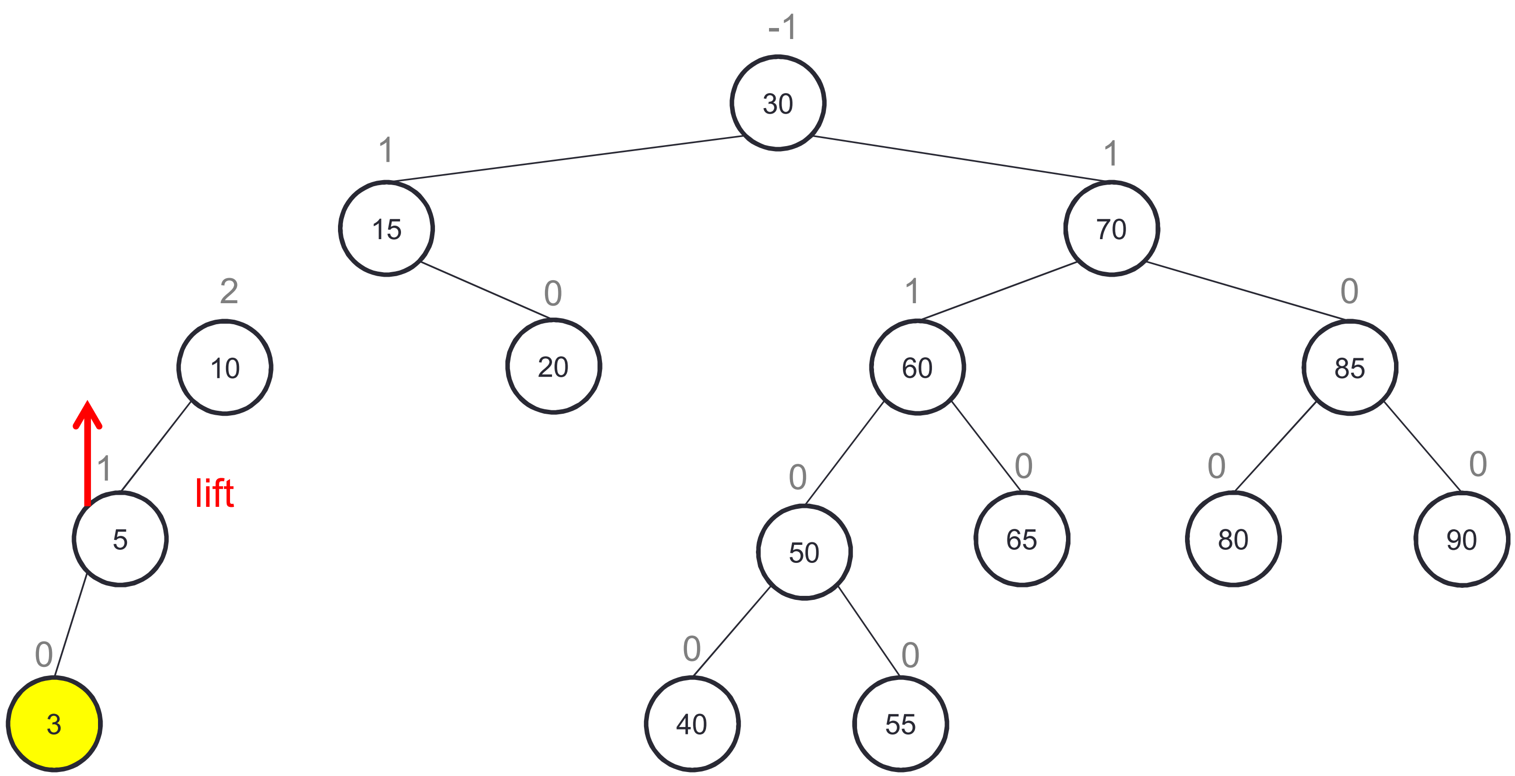
Inserting and rebalancing

Next slide



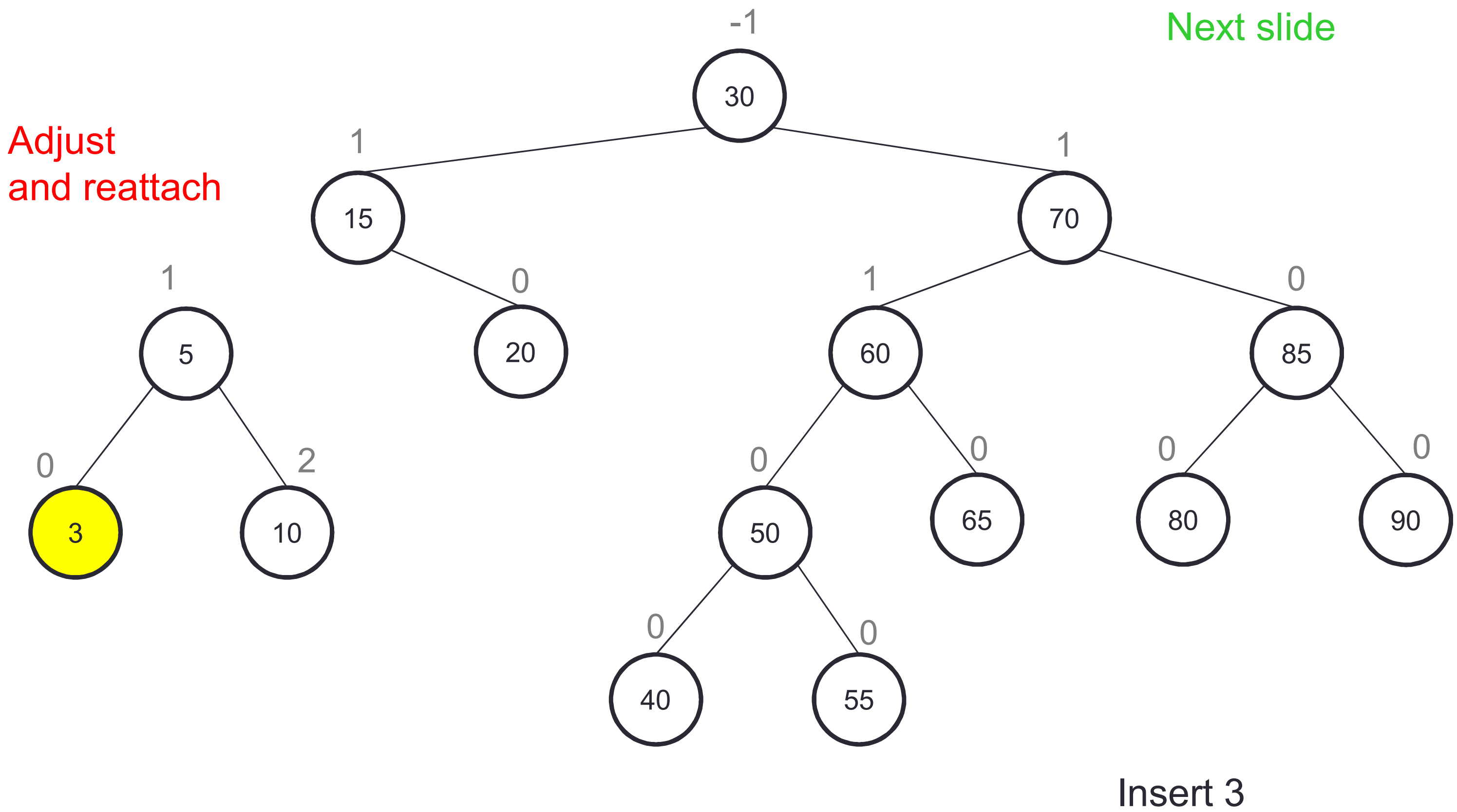
Insert 3

Inserting and rebalancing



Insert 3

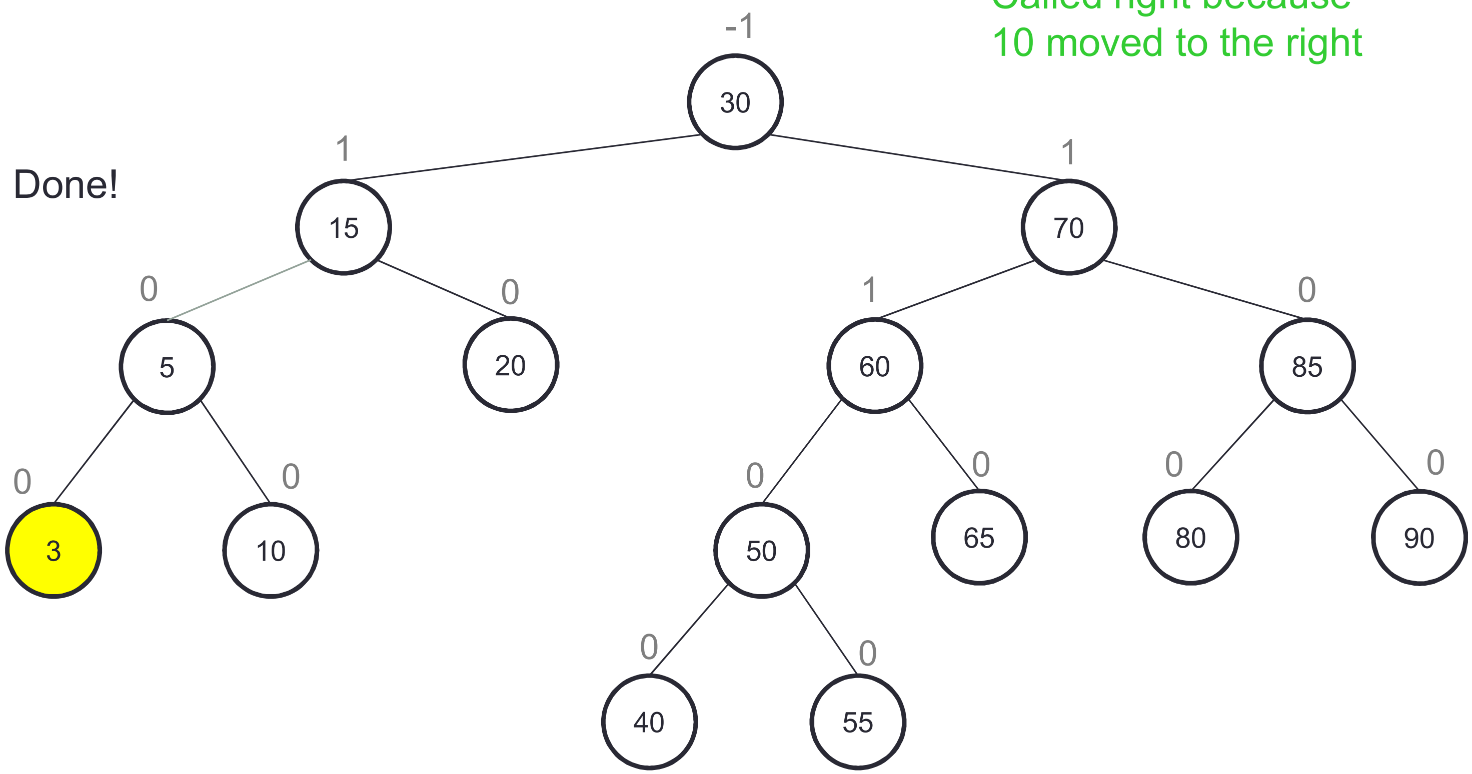
Inserting and rebalancing



We just did a single right rotation at 10

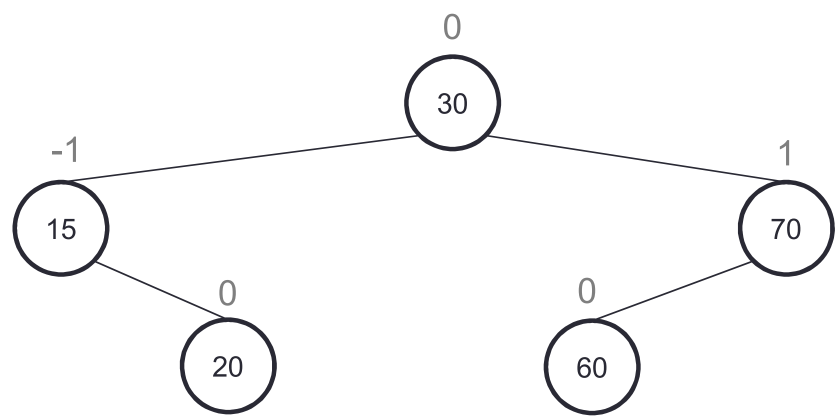
Called right because
10 moved to the right

Done!



Insert 3

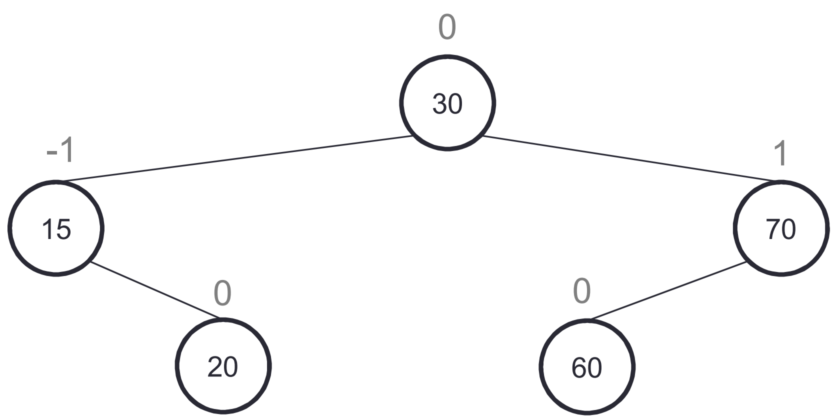
Single rotation practice



What could you insert into this AVL tree that would result in a single rotation?

A. 71
B. 10
C. 50
D. 66

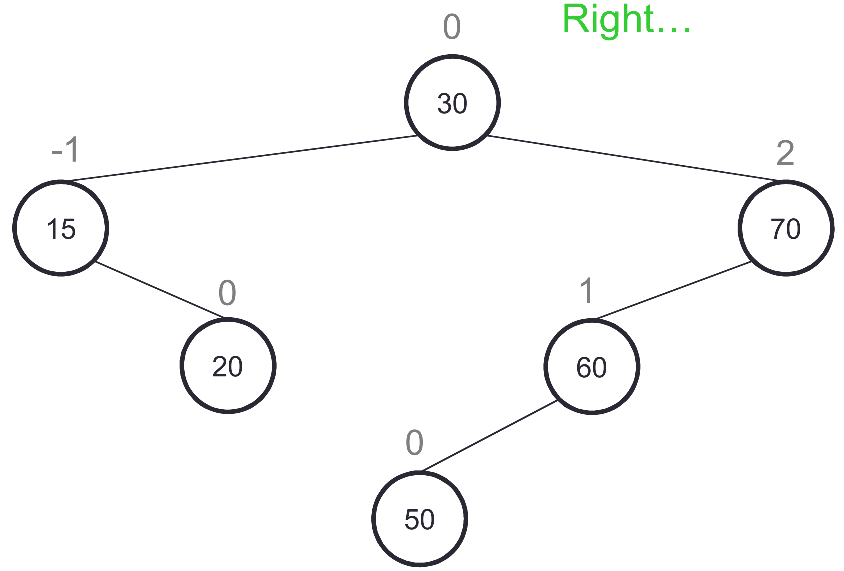
Single rotation practice



Insert 50. Draw the resulting AVL tree. (Don't peek)

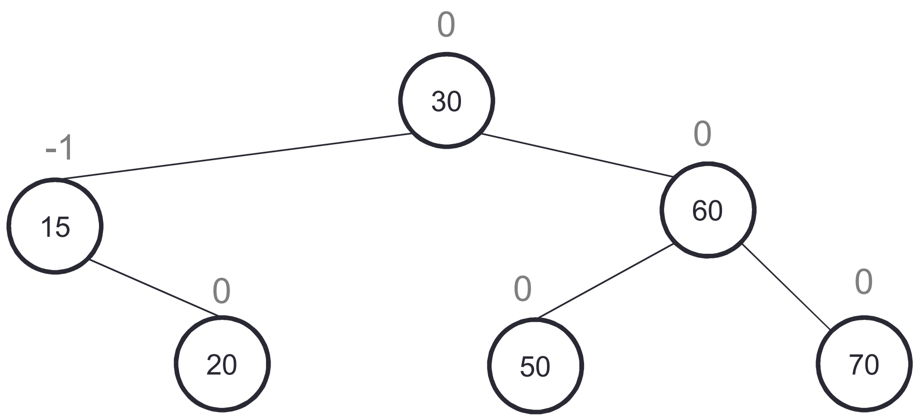
Single rotation practice

What rotation do we need to do?
Right…



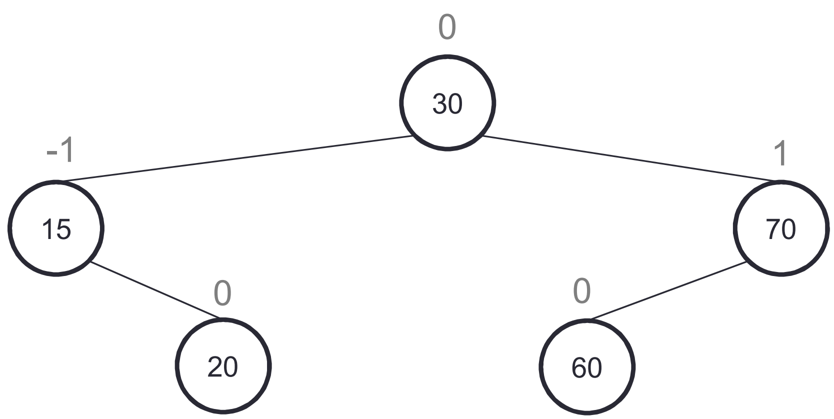
After insertion

Single rotation practice



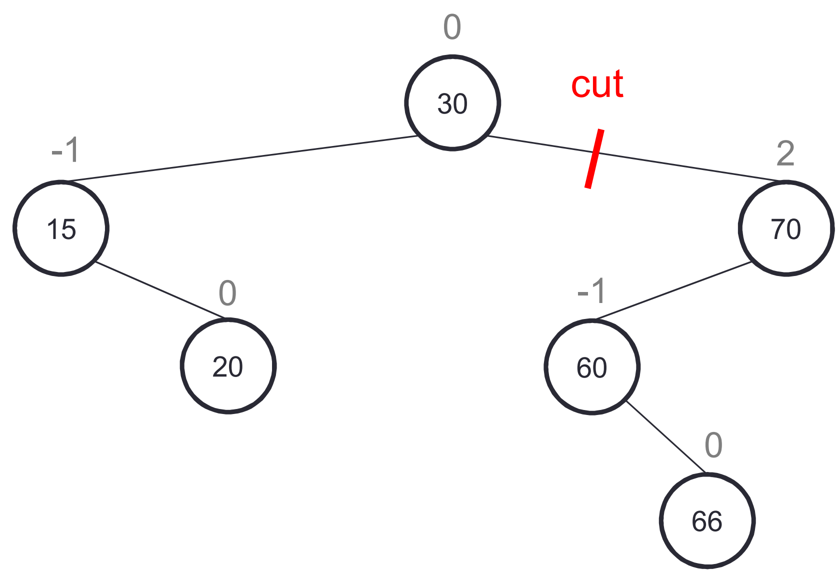
After rotation

Single rotation is not enough



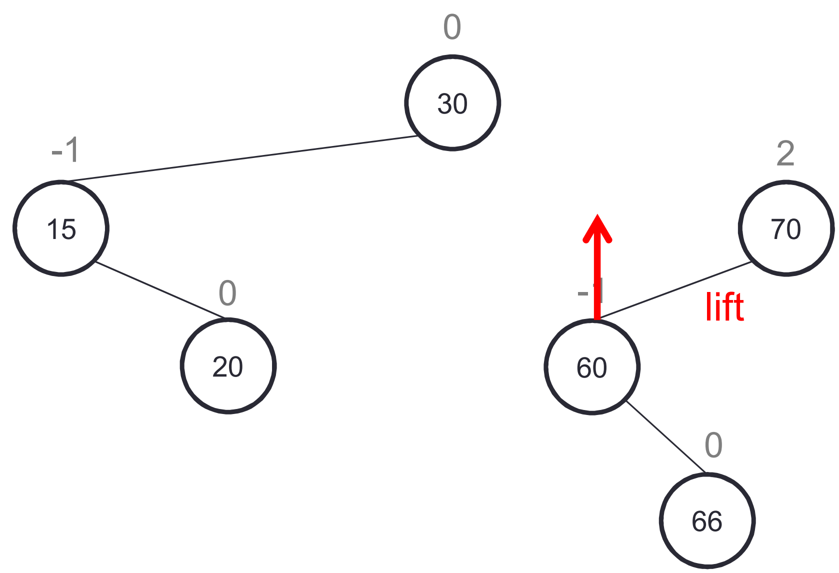
What happens if we insert 66?

Single rotation is not enough

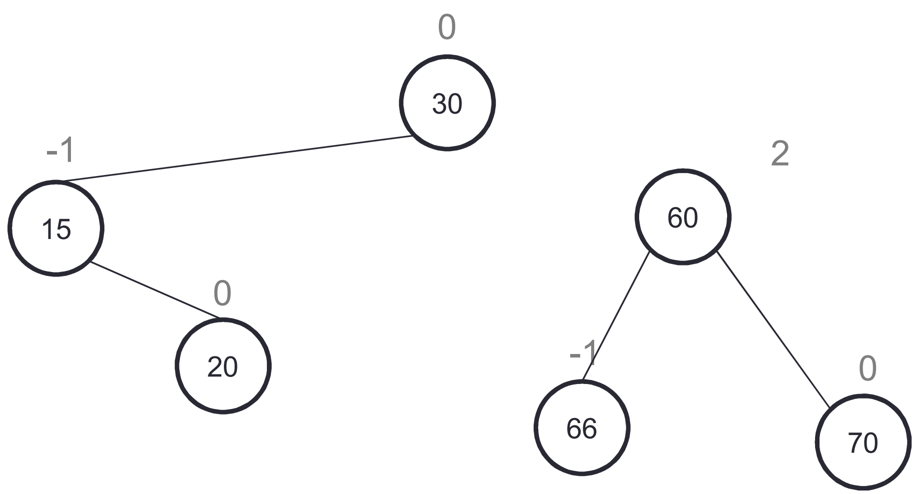


Why won't a single rotation work? Try it.

Single rotation is not enough

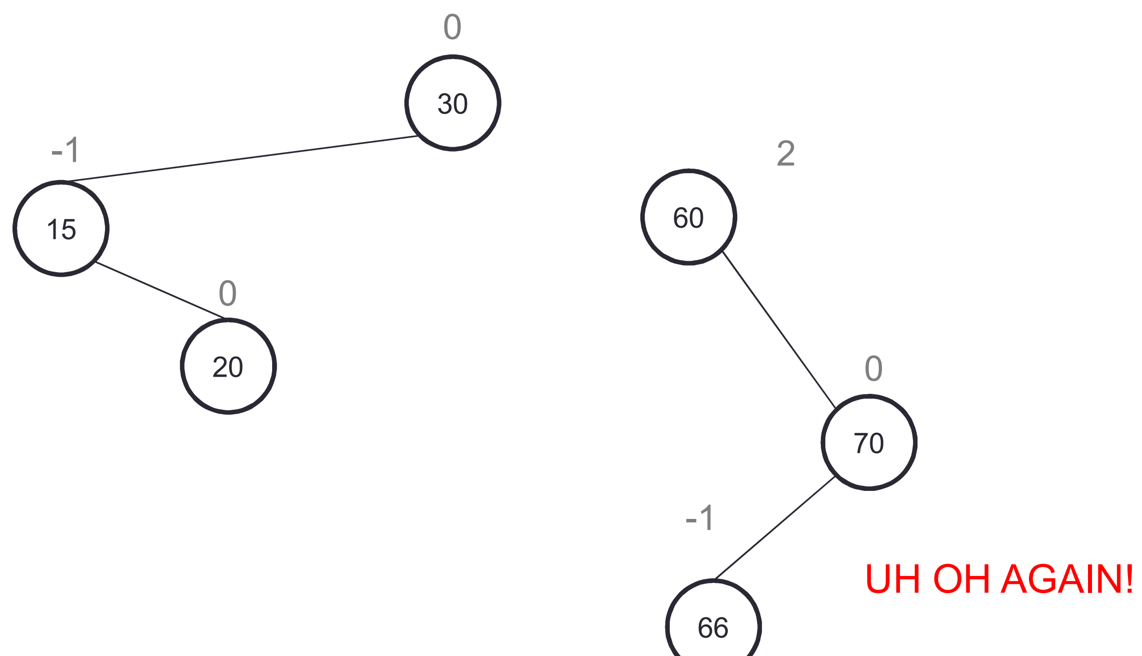


Single rotation is not enough



UH OH!

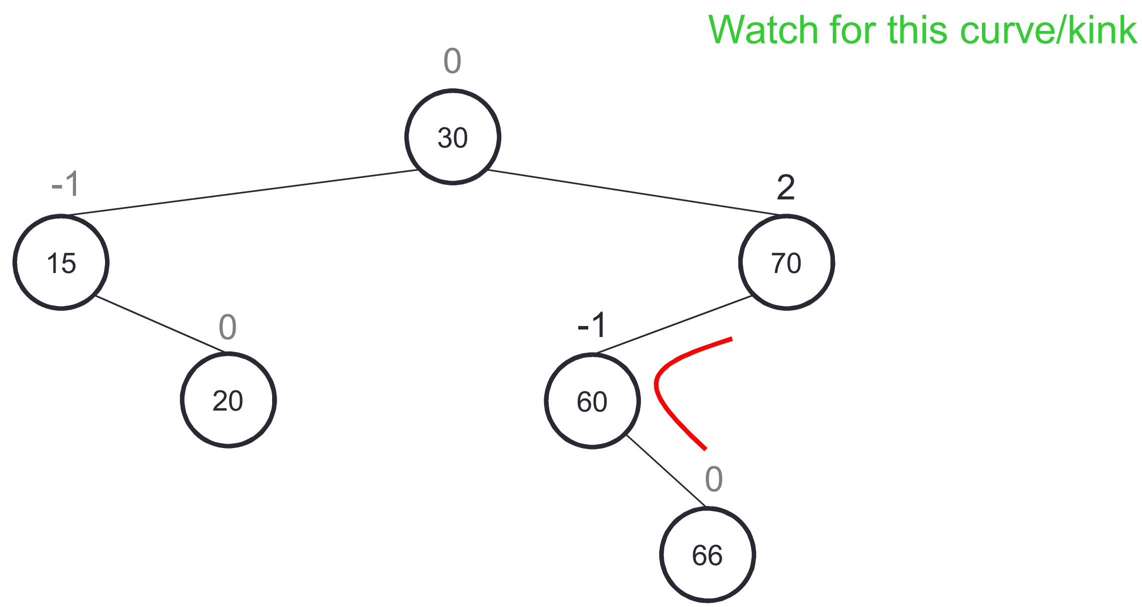
Single rotation is not enough



Reattaching 66 here will always work with respect to the BST properties, and we know that 66 will always fit here because 60 used to be 70s left child.

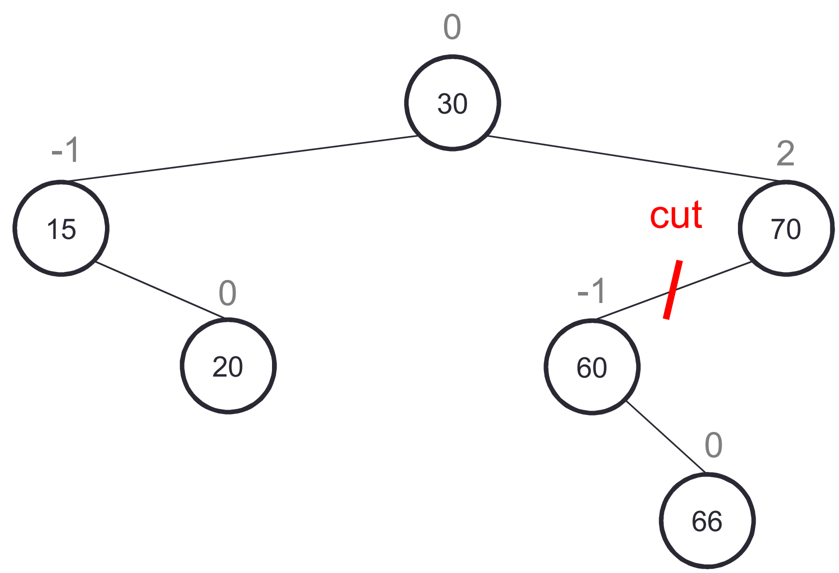
The problem is that this won't fix the balance issue!

Double rotation to the rescue



Single rotations only work to balance the tree when involved nodes are “in a line”
This is not the case here. We want 66 to be the top node, not 60.
So we will first rotate left at 60 to get 66 in the middle, then we can rotate right at 70.

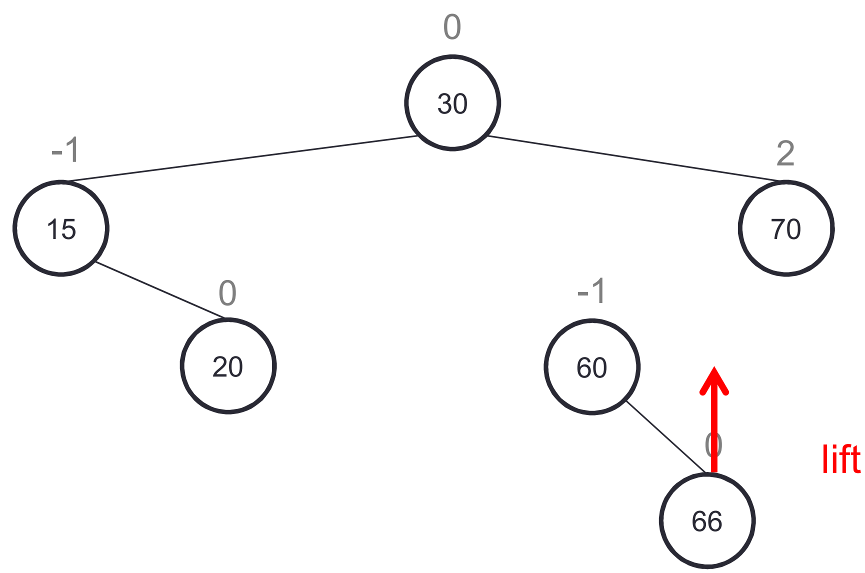
Double rotation to the rescue



Single rotations only work when involved nodes are “in a line”

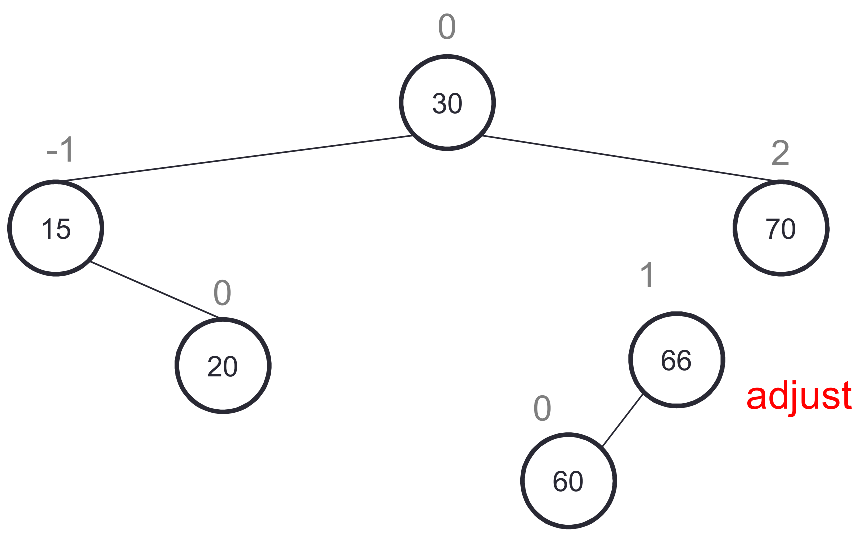
So we will first rotate left at 60, then we can rotate right at 70.

Double rotation to the rescue



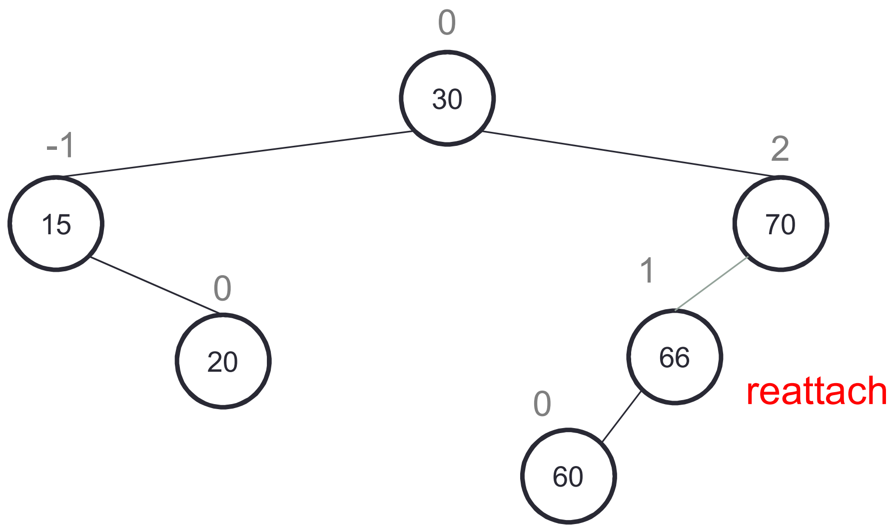
Single rotations only work when involved nodes are "in a line"
So we will first rotate left around 60, then we can rotate right around 70.

Double rotation to the rescue



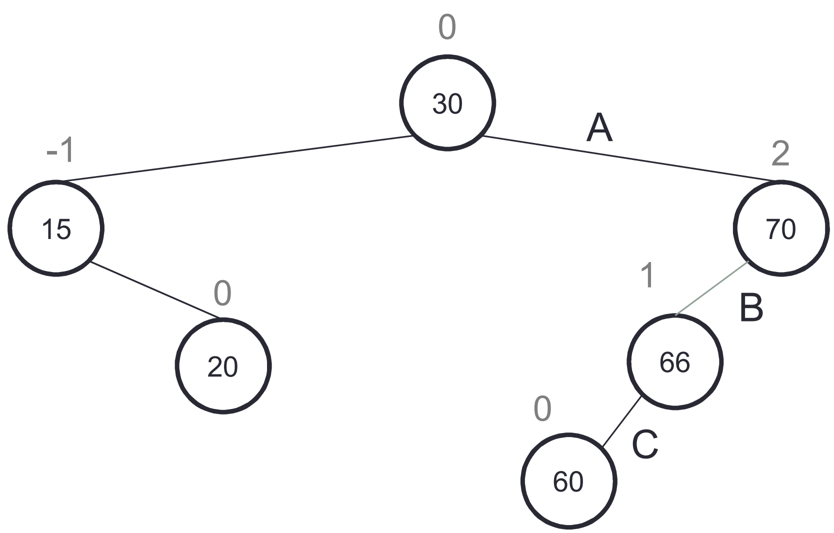
Single rotations only work when involved nodes are “in a line”
So we will first rotate left at 60, then we can rotate right at 70.

Double rotation to the rescue



Single rotations only work when involved nodes are “in a line”
So we will first rotate left at 60, then we can rotate right at 70.

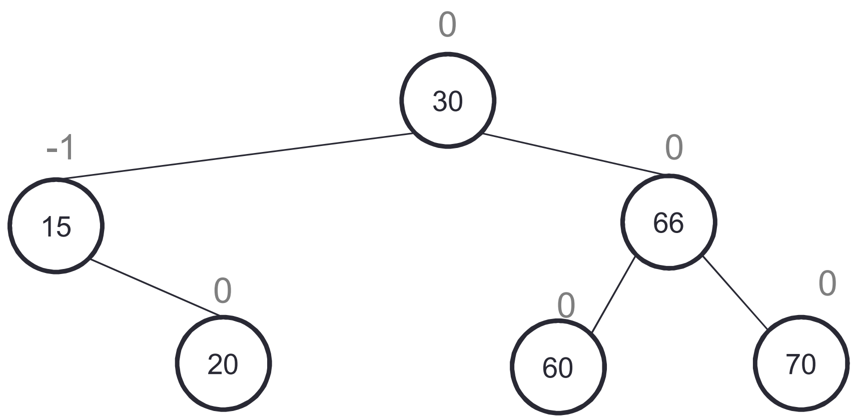
Double rotation to the rescue



Single rotations only work when involved nodes are “in a line”
So we will first rotate left around 60, then we can rotate right around 70.

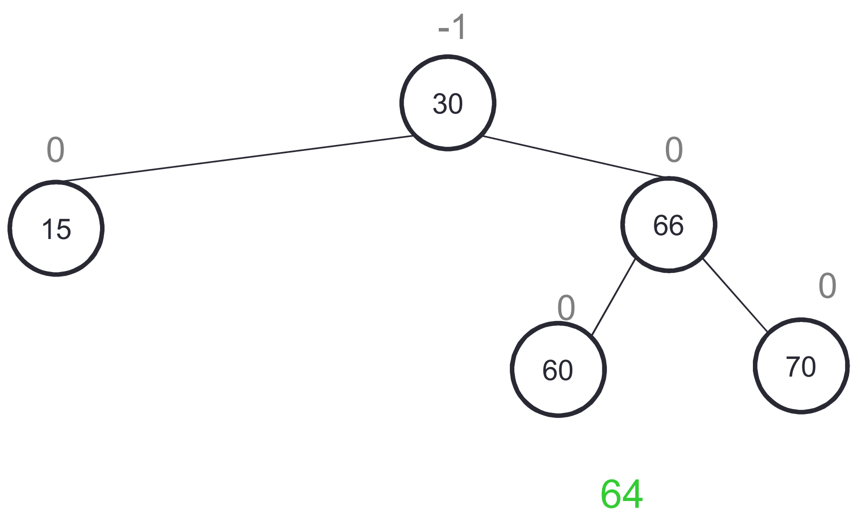
Where in the tree above should I cut to start the second rotation?

Double rotation to the rescue



Single rotations only work when involved nodes are “in a line”
So we will first rotate left at 60, then we can rotate right at 70.

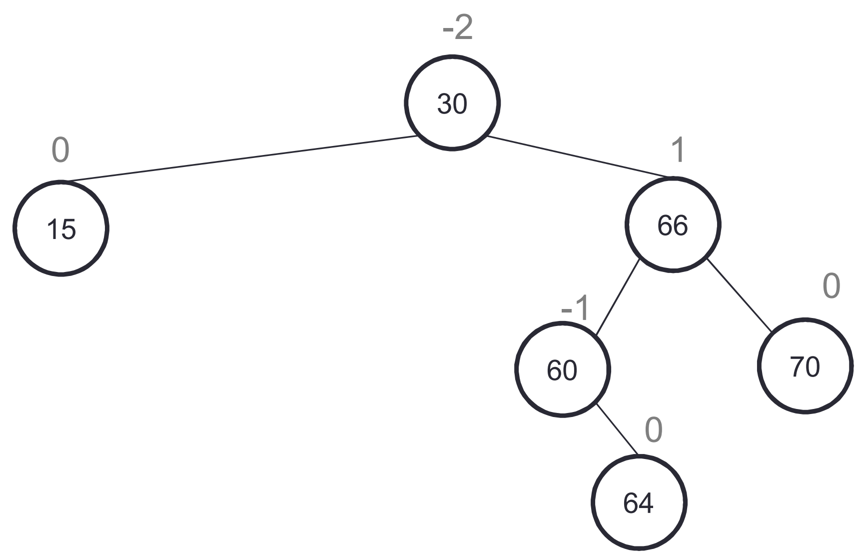
It's sometimes even more complicated



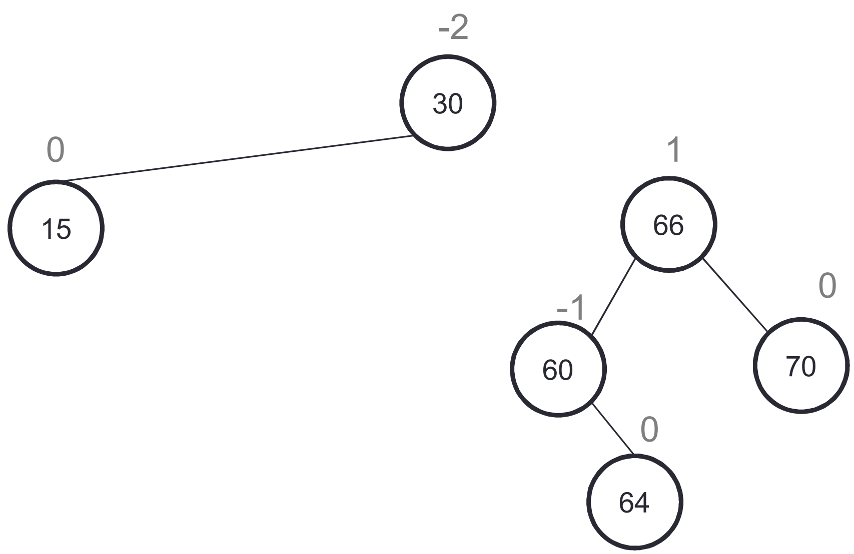
Insert 64… do we need a double or a single rotation?

A. Double
B. Single
C. No rotation needed

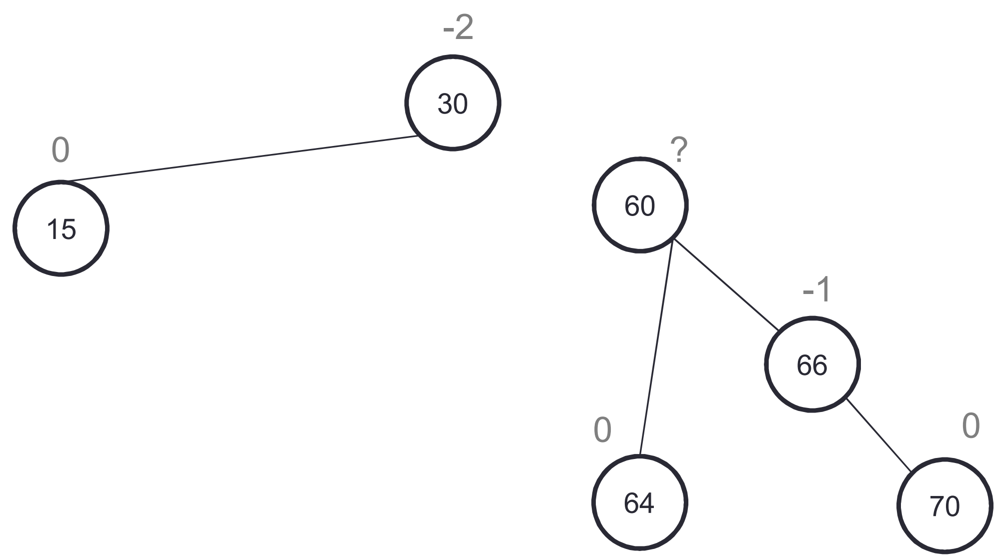
Rotate right at 66 to make a straight line



Rotate right at 66 to make a straight line

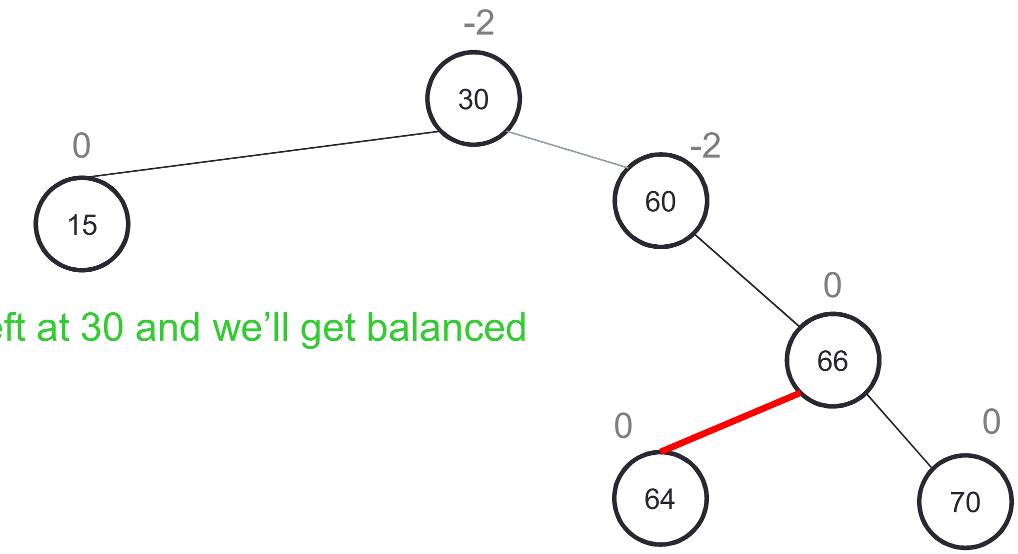


Rotate right at 66 to make a straight line



UH OH! Where do we put 64??
Are we stuck?

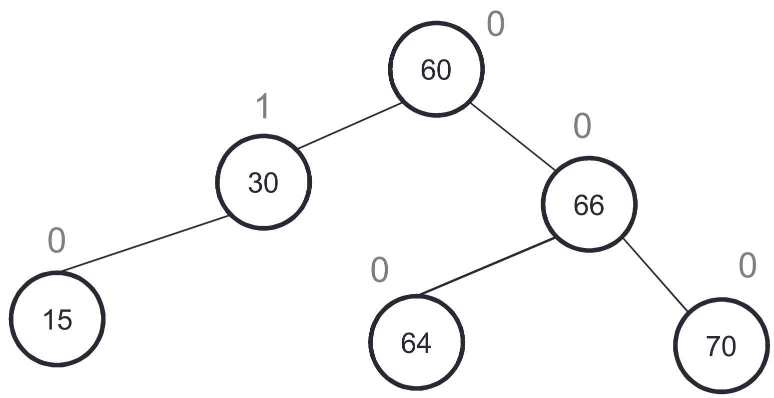
Rotate right at 66 to make a straight line



Now rotate left at 30 and we'll get balanced

Will 64 always reattach there?
Yes! Discuss why…

Finishing the rotation to balance the tree



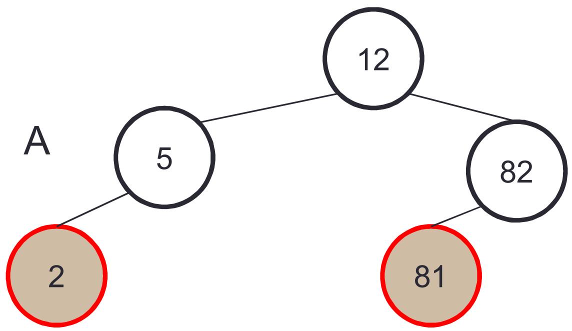
Summary

- AVLs are balanced because we enforce balance at insertion (and removal)
  - Rotations allow us to achieve this and are constant time operations
- By enforcing balance, we ensure O(log n) find operations

Red Black Trees, review

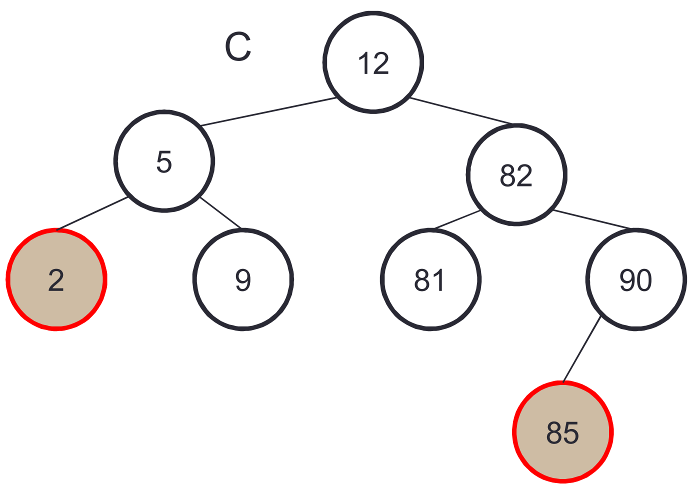
Which of the following are legal red-black trees?

A



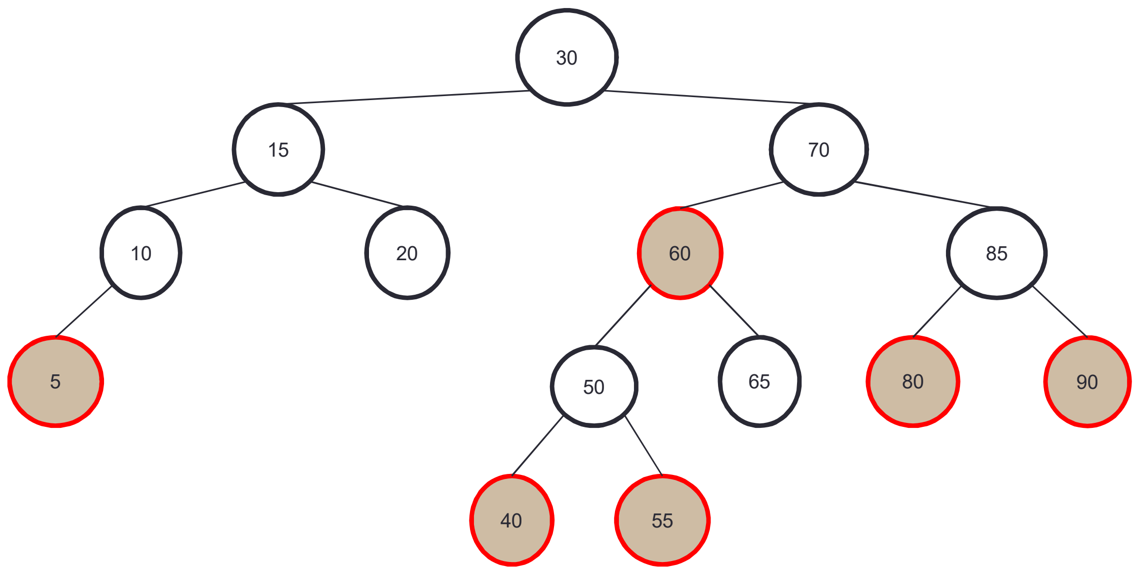
B

C



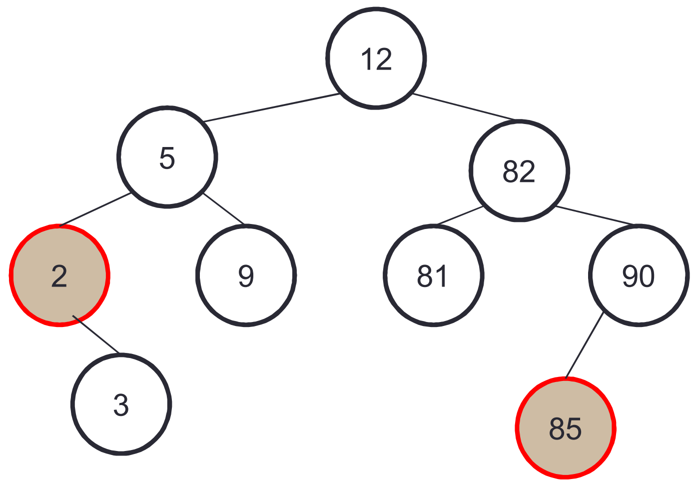
D. A&B
E. A&B&C

Red-Black Trees



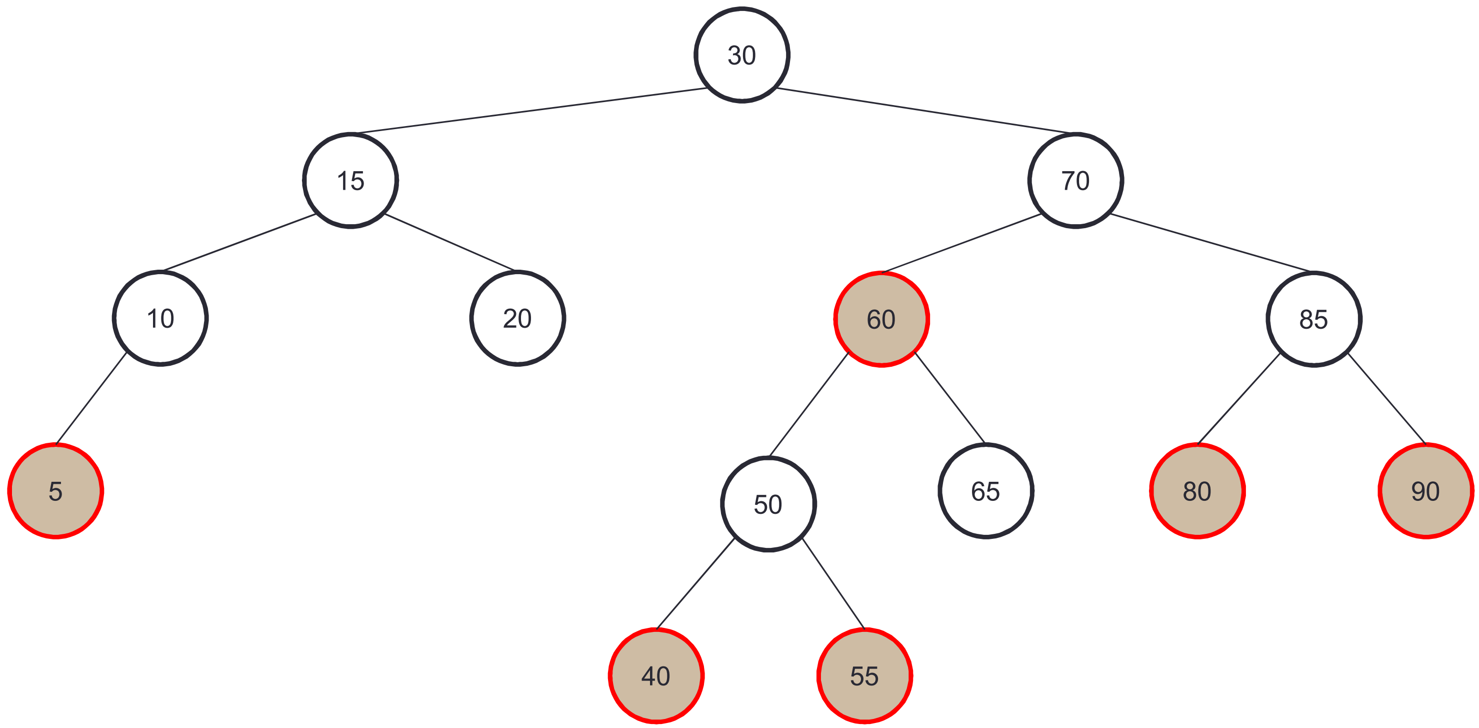
1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it's children must be black
4. For every node X, every path from X to a *null reference* must contain the same number of black nodes

Is this a legal Red-Black tree?



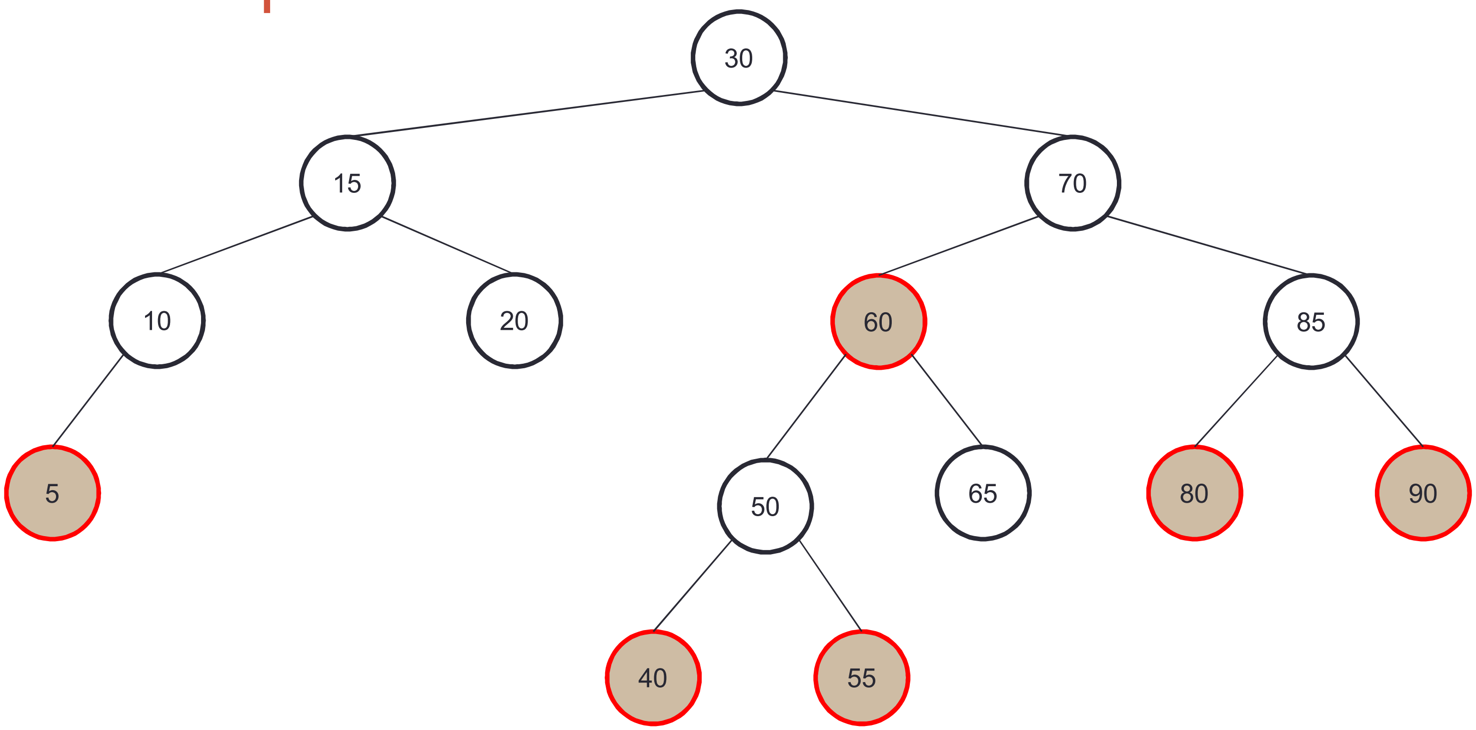
A. Yes
B. No

Red-Black trees have height O(logN)



Intuition for proof – remove all red nodes all black nodes are at the same height. Can then reason about relationship between n and the height. Height of just black is will be O(log n). Add back red and you get just 2*height black.

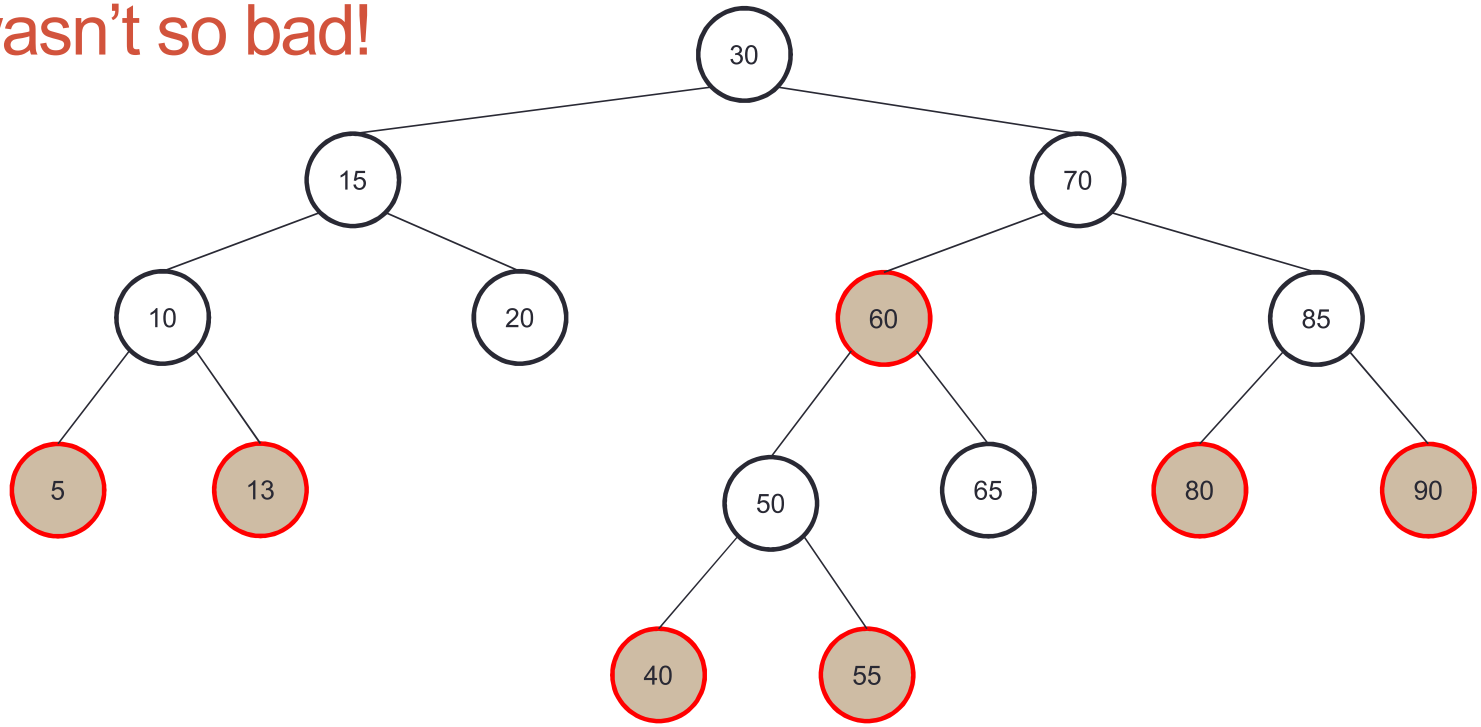
Now for the fun part… insertions



Non-root insertions will always be red

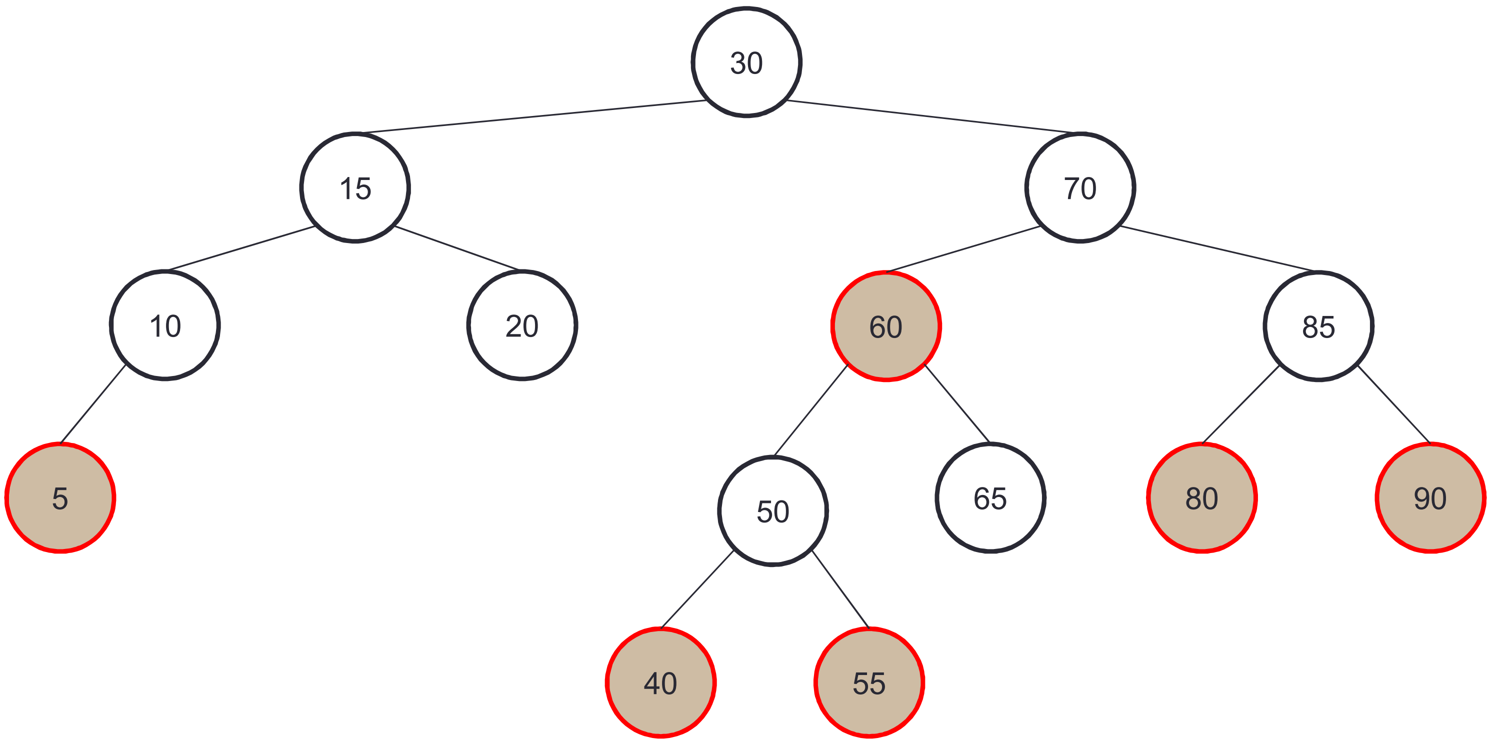
Try inserting 13

That wasn't so bad!



Case 0: Parent was black. Insert new leaf node (red) and you're done.

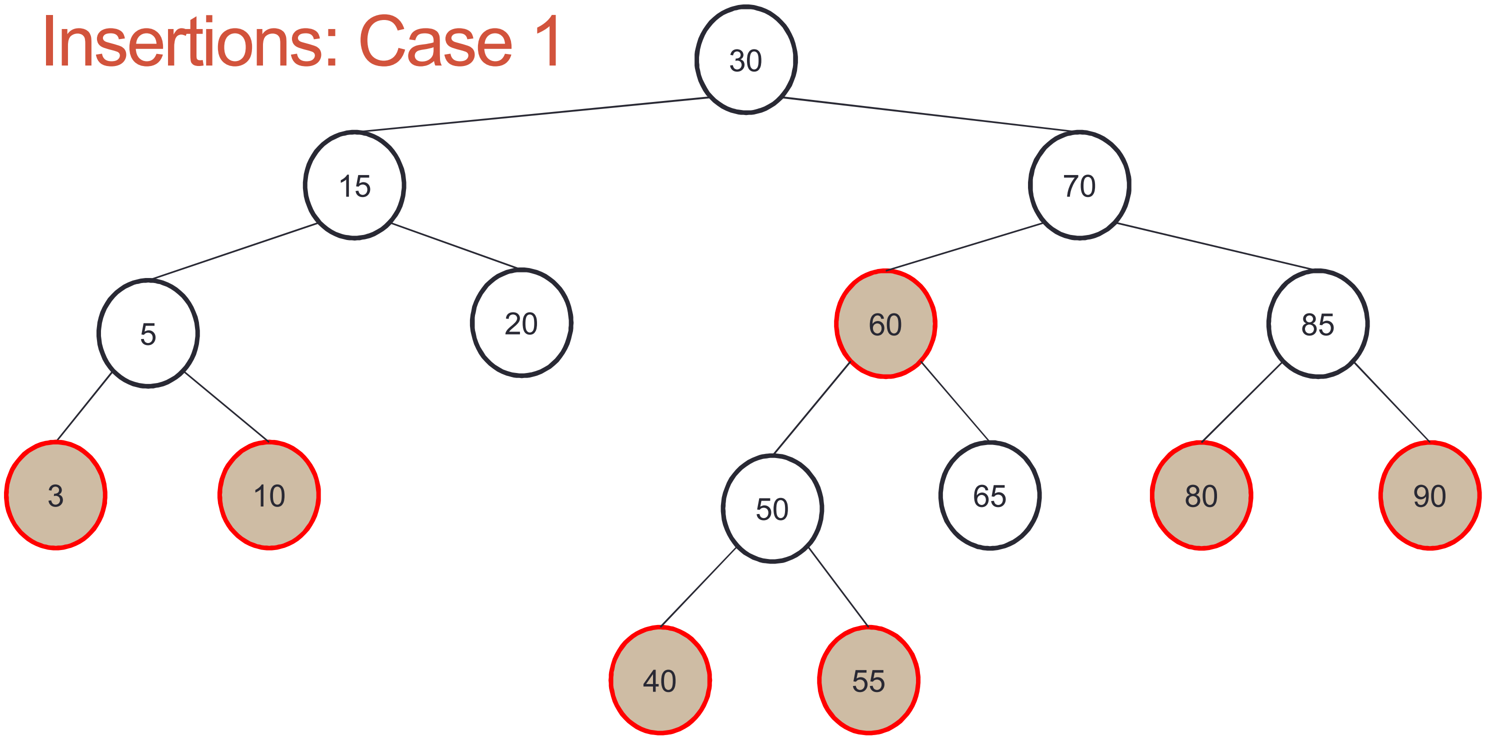
Insertions: More complicated case



Try inserting 3

Case 1: Parent of leaf is red, parent is left child of grandparent,
leaf is left child of parent, (& sibling of parent is black)

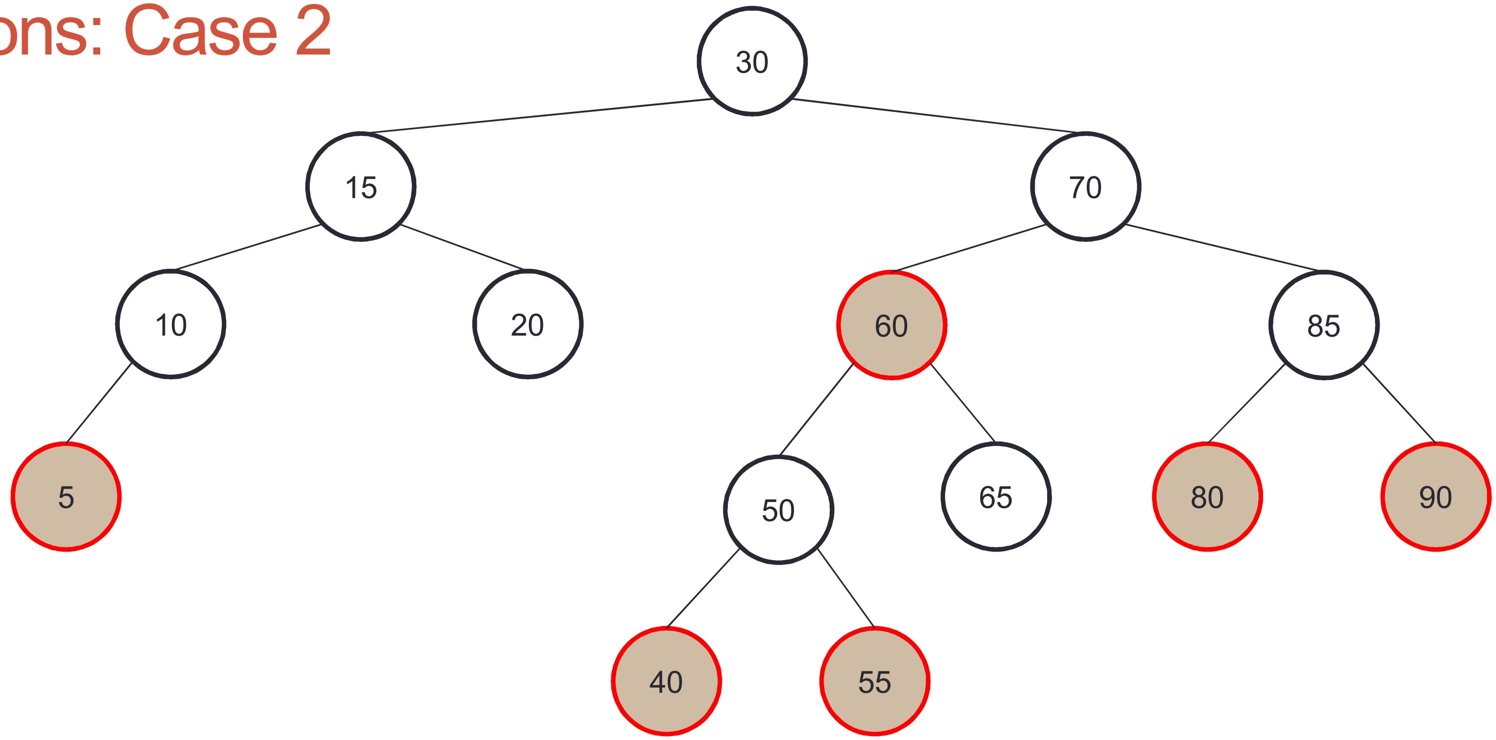
Insertions: Case 1



Right AVL rotation, and recolor

B

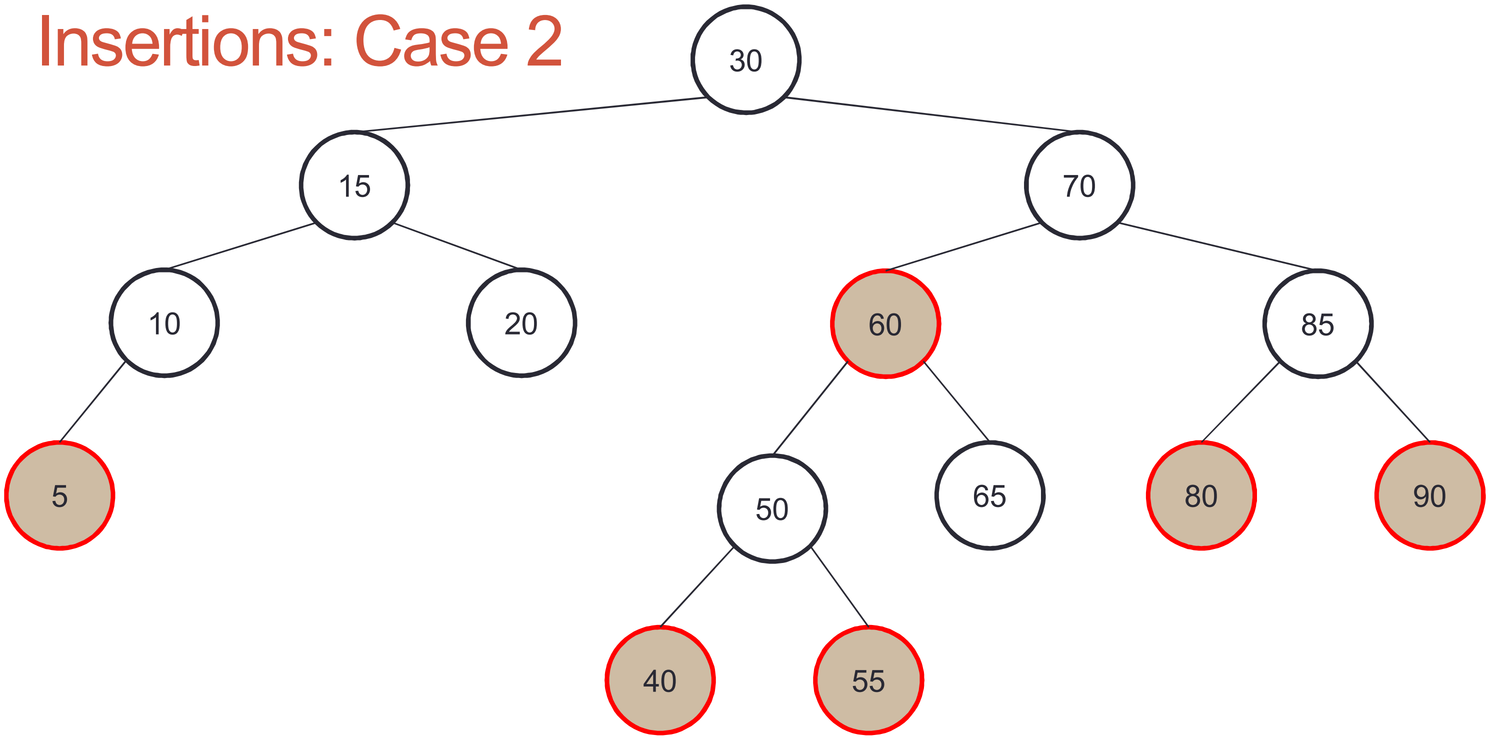
Insertions: Case 2



Which insertion can we not handle with the cases we've seen so far?

A. 1 B. 7 C. 12 D. 25

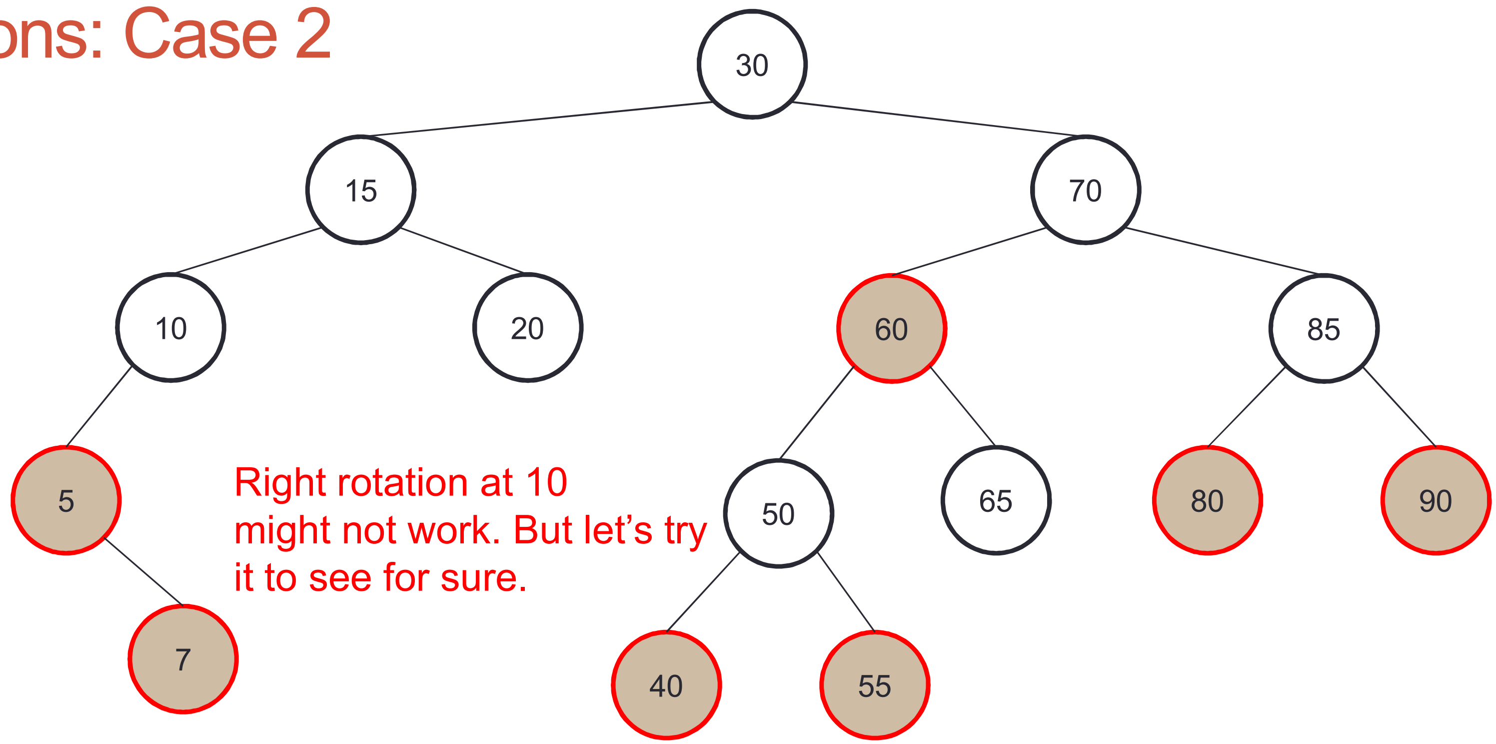
Insertions: Case 2



Insert 7

Case 2: Parent of leaf is red, parent is a left child of grandparent, leaf is right child of parent, (& sibling of parent is black)

Insertions: Case 2

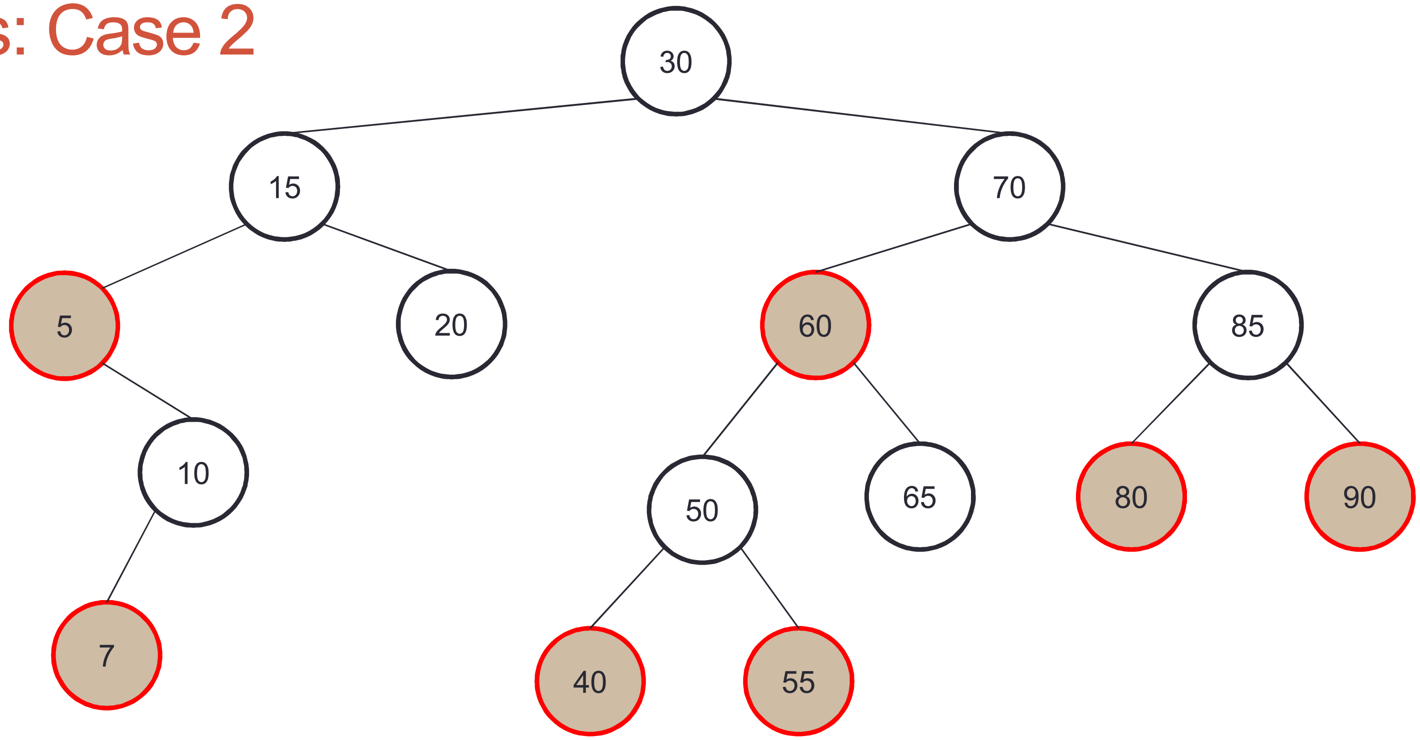


Right rotation at 10 might not work. But let's try it to see for sure.

Insert 7

Case 2: Parent of leaf is red, parent is a left child of grandparent, leaf is right child of parent, (& sibling of parent is black)

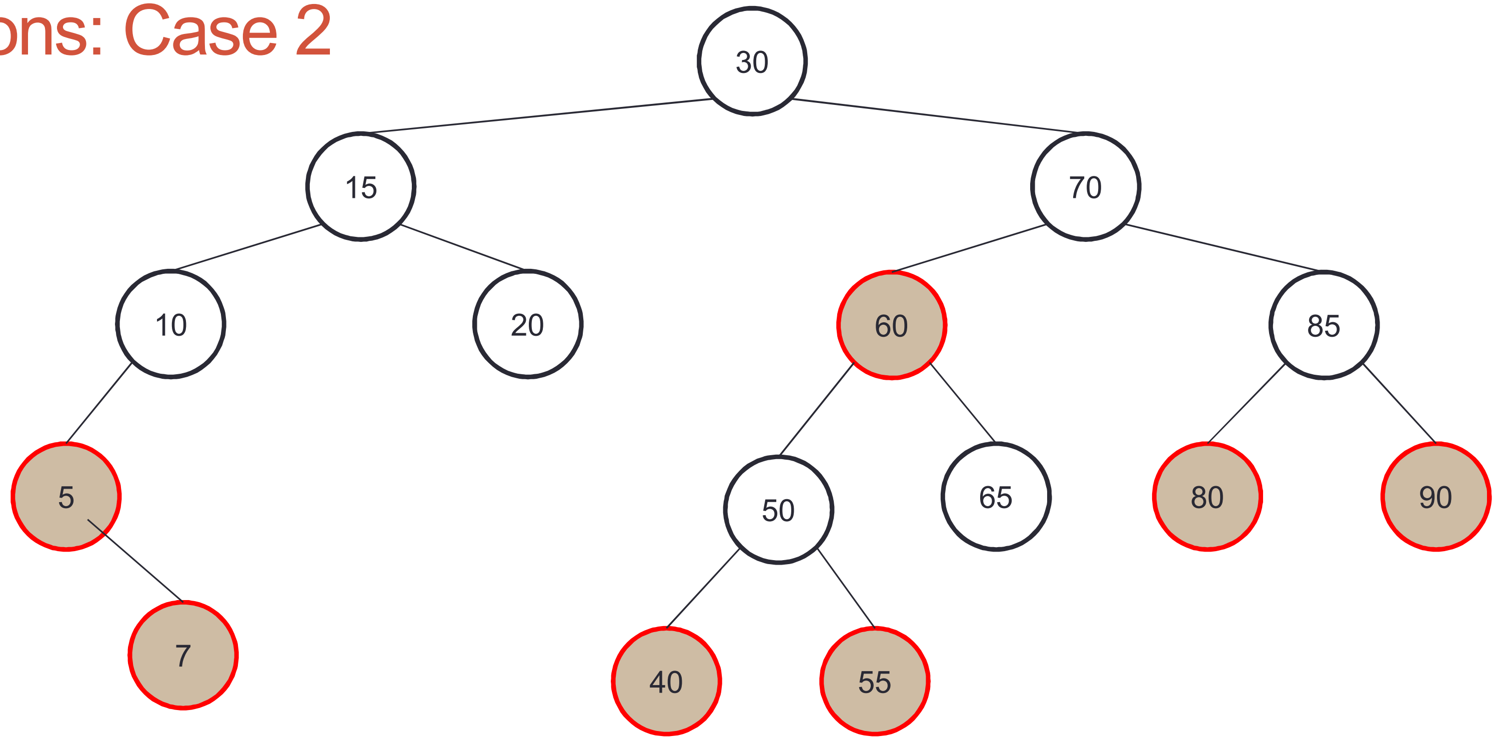
Insertions: Case 2



Why didn't this work?

A. It did! We're done!
B. The property about red nodes having only black children is violated.
C. The property about having the same number of black nodes on any path from the root through a null reference is violated.

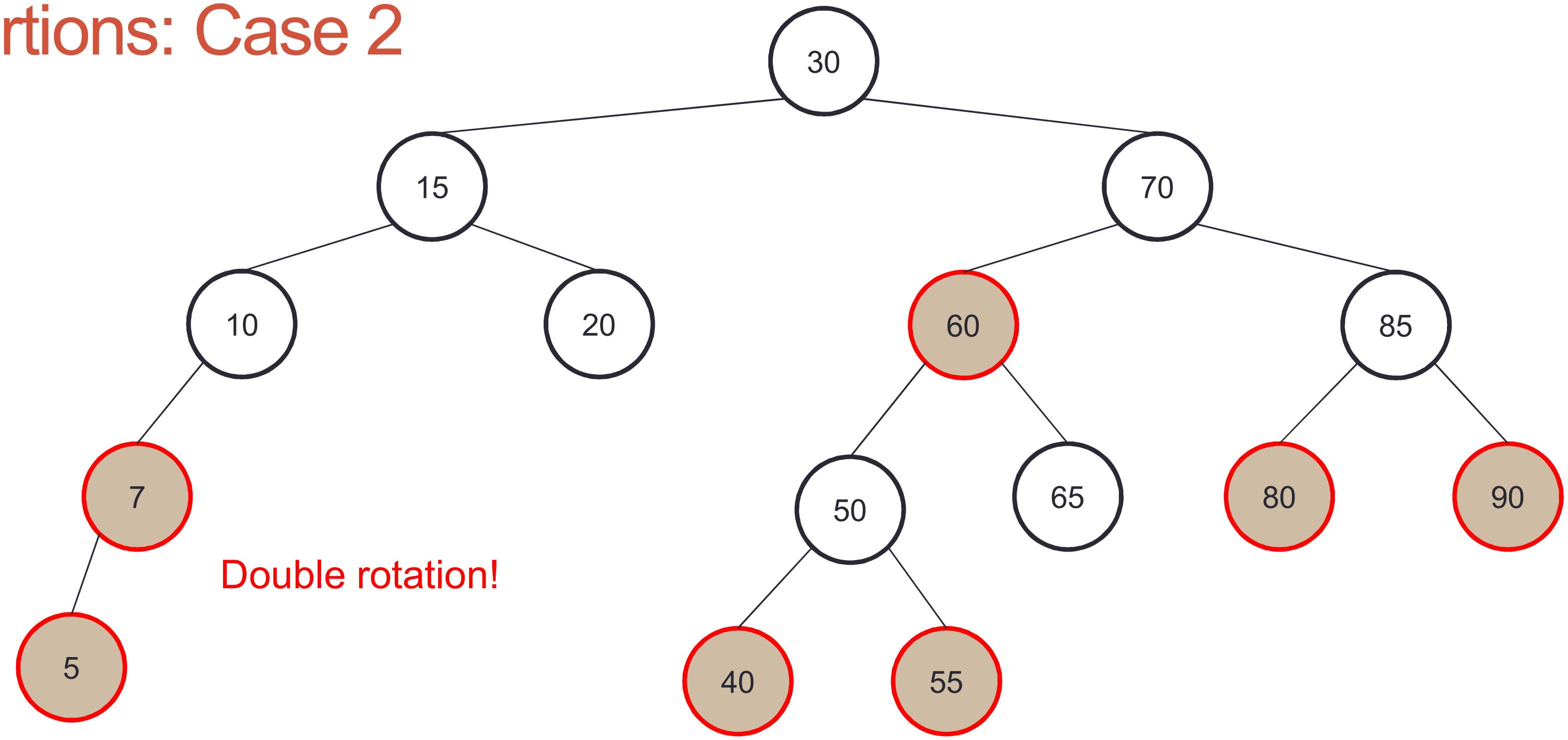
Insertions: Case 2



Insert 7

Case 2: Parent of leaf is red, parent is a left child of grandparent, leaf is right child of parent, (& sibling of parent is black)

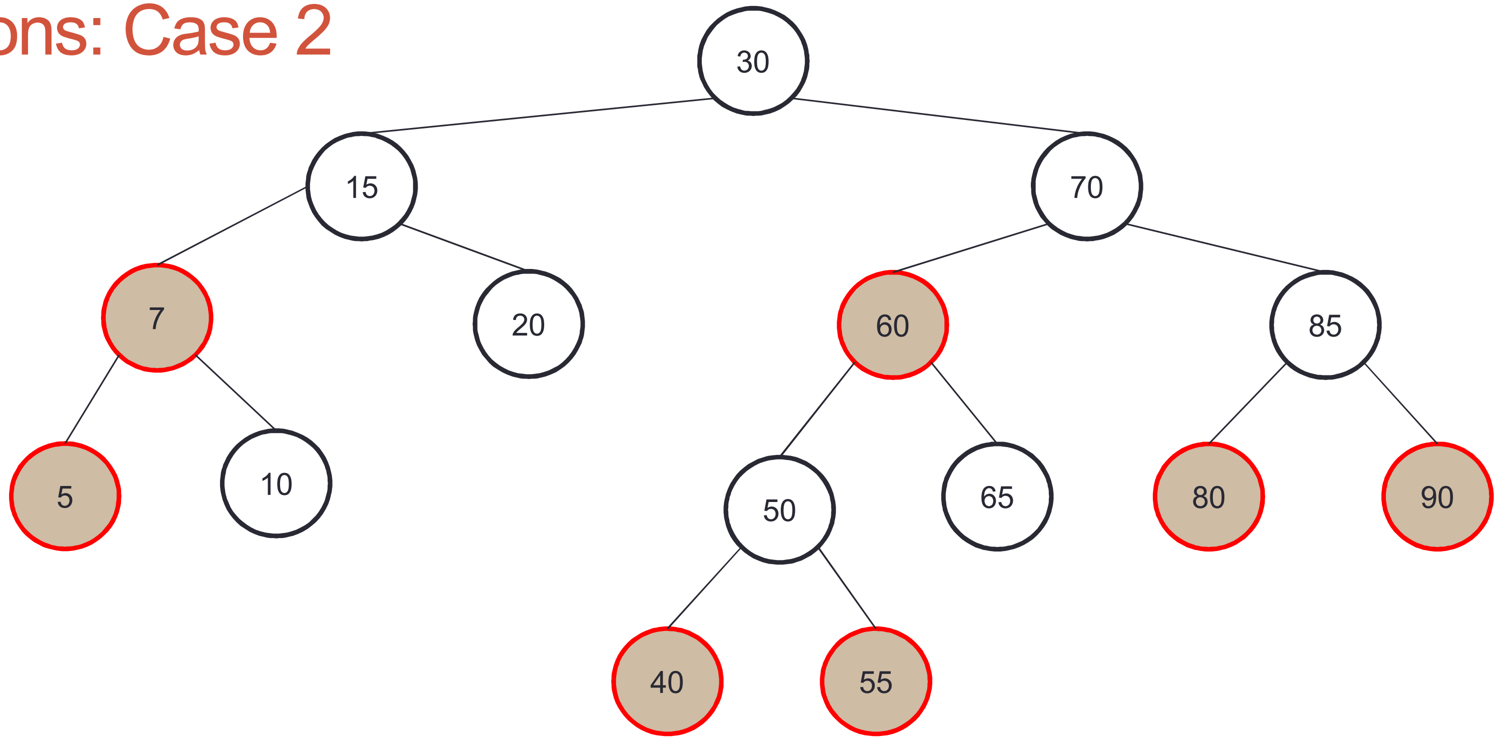
Insertions: Case 2



Insert 7

Case 2: Parent of leaf is red, parent is a left child of grandparent,
leaf is right child of parent, (& sibling of parent is black)

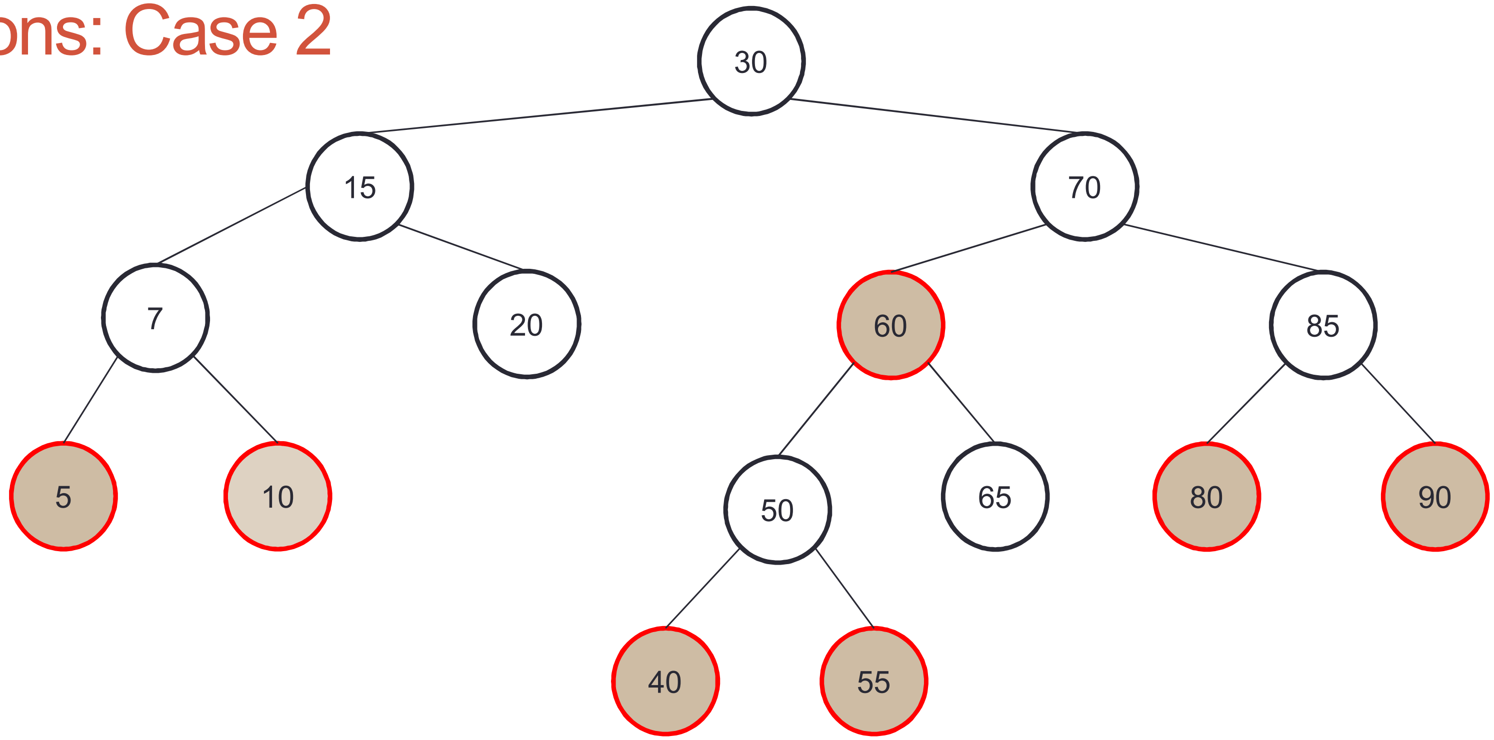
Insertions: Case 2



Insert 7 Recolor!

Case 2: Parent of leaf is red, parent is a left child of grandparent, leaf is right child of parent, (& sibling of parent is black)

Insertions: Case 2



Insert 7

Case 2: Parent of leaf is red, parent is a left child of grandparent,
leaf is right child of parent, (& sibling of parent is black)

Insertions: Summary, so far

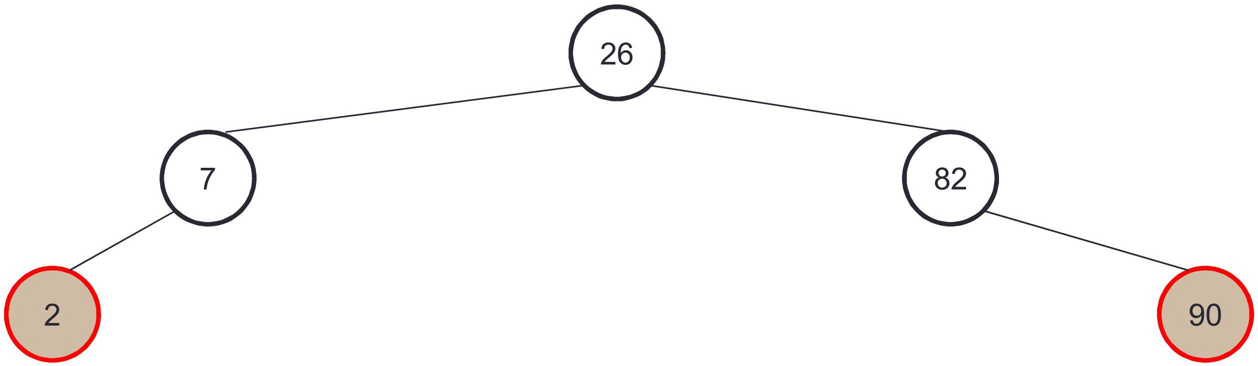
Case 0: The parent of the node you are inserting is black. Insert and you're done

For the remaining cases, the parent of the node is red, the sibling of the parent is black:

Case 1: P is left child of G, X is left child of P (single rotate then recolor)
Case 2: P is left child of G, X is right child of P (double rotate then recolor)

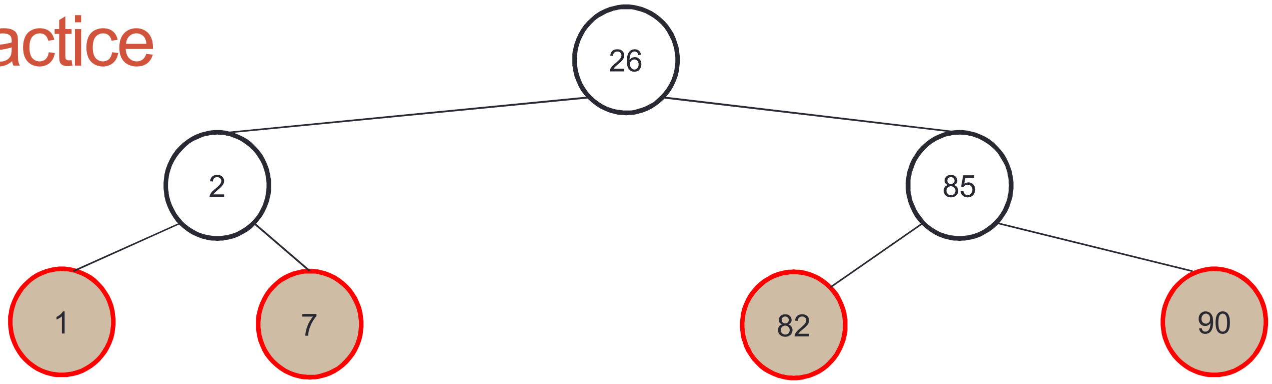
Case 3: P is right child of G, X is right child of P
Case 4: P is right child of G, X is left child of P

Practice



Insert 1 and then insert 85. Draw the resulting tree.

Practice



The final tree

Insertions: Summary, so far

Case 0: The parent of the node you are inserting is black. Insert and you're done

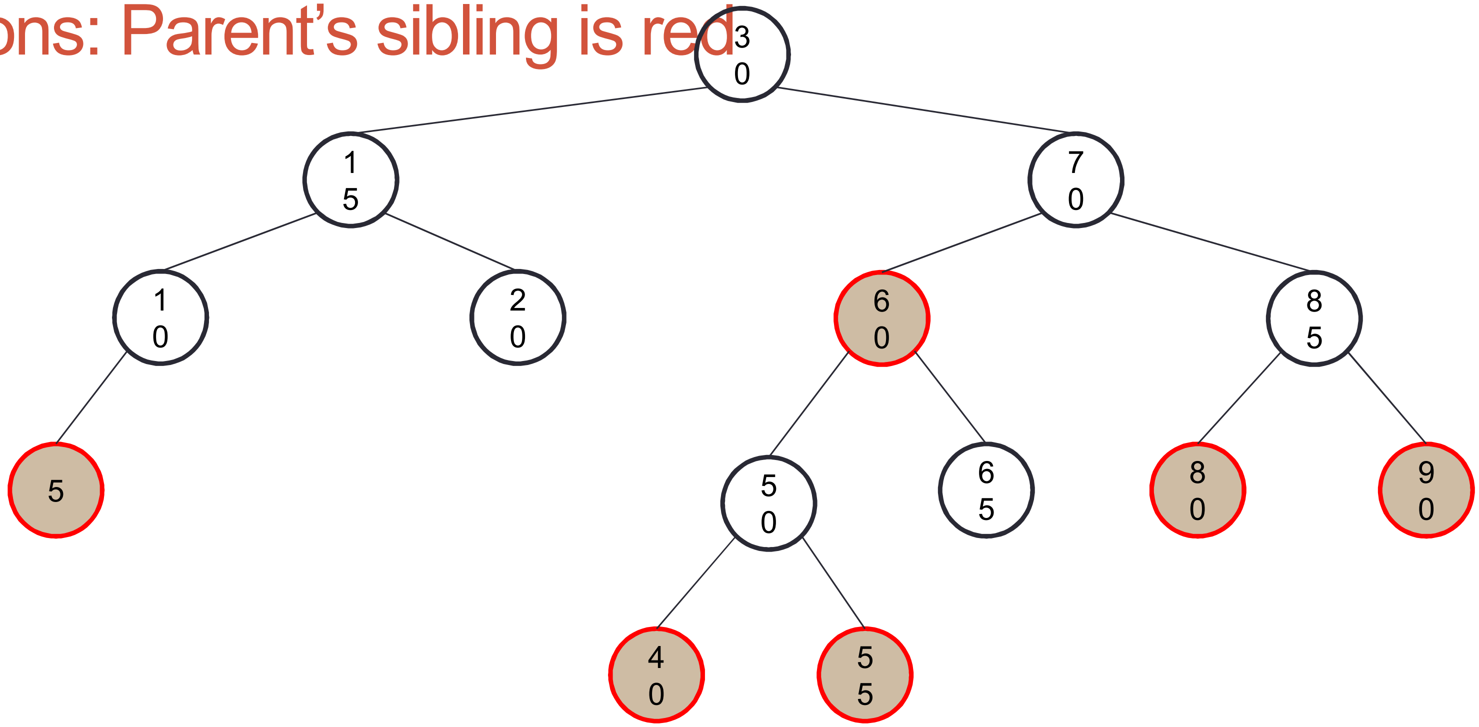
For the remaining cases, the parent of the node is red, **the sibling of the parent is black**:

Case 1: P is left child of G, X is left child of P (single rotate then recolor)
Case 2: P is left child of G, X is right child of P (double rotate then recolor)

Case 3: P is right child of G, X is right child of P
Case 4: P is right child of G, X is left child of P

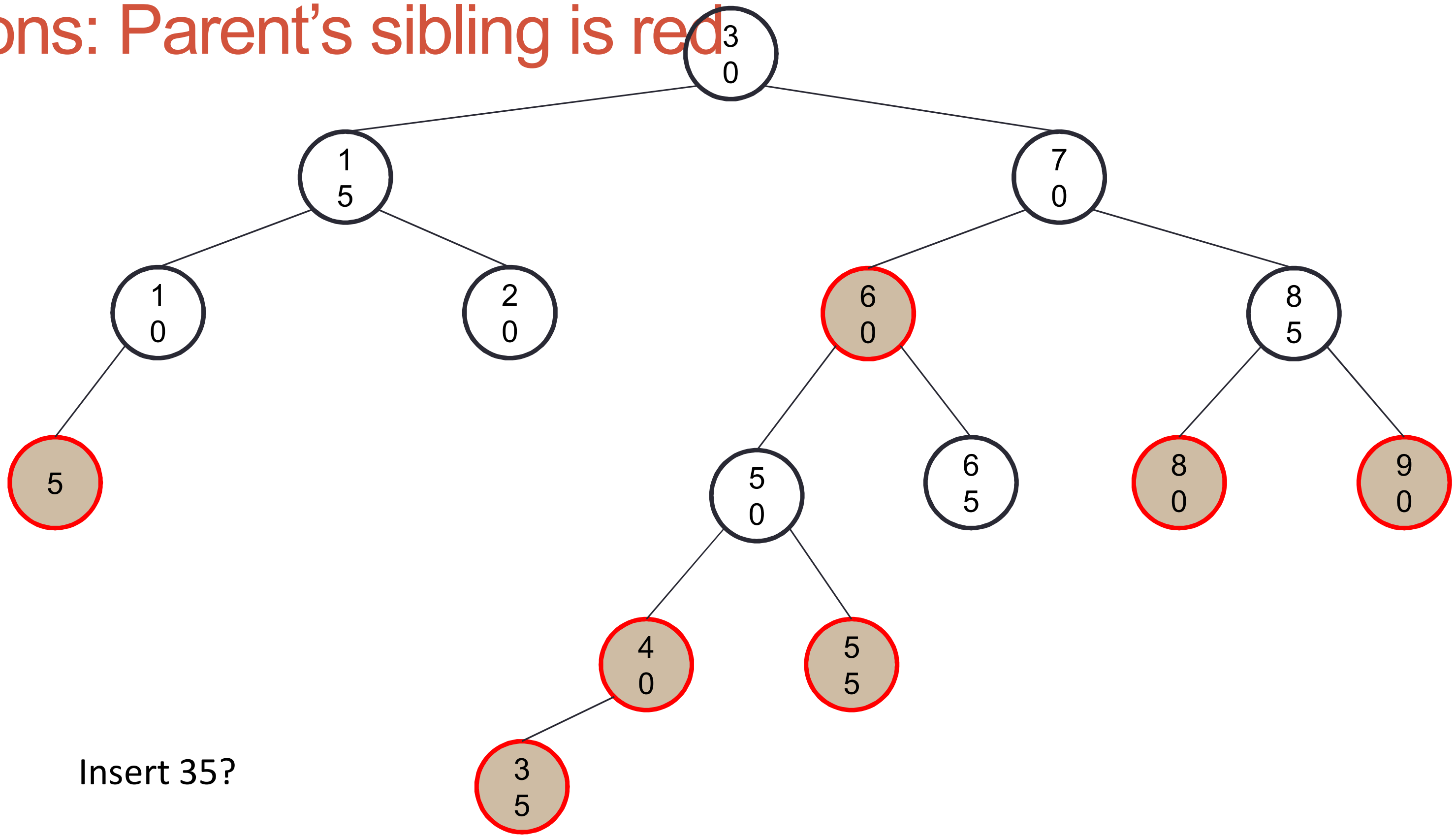
What if the sibling of the parent is red??

Insertions: Parent’s sibling is red



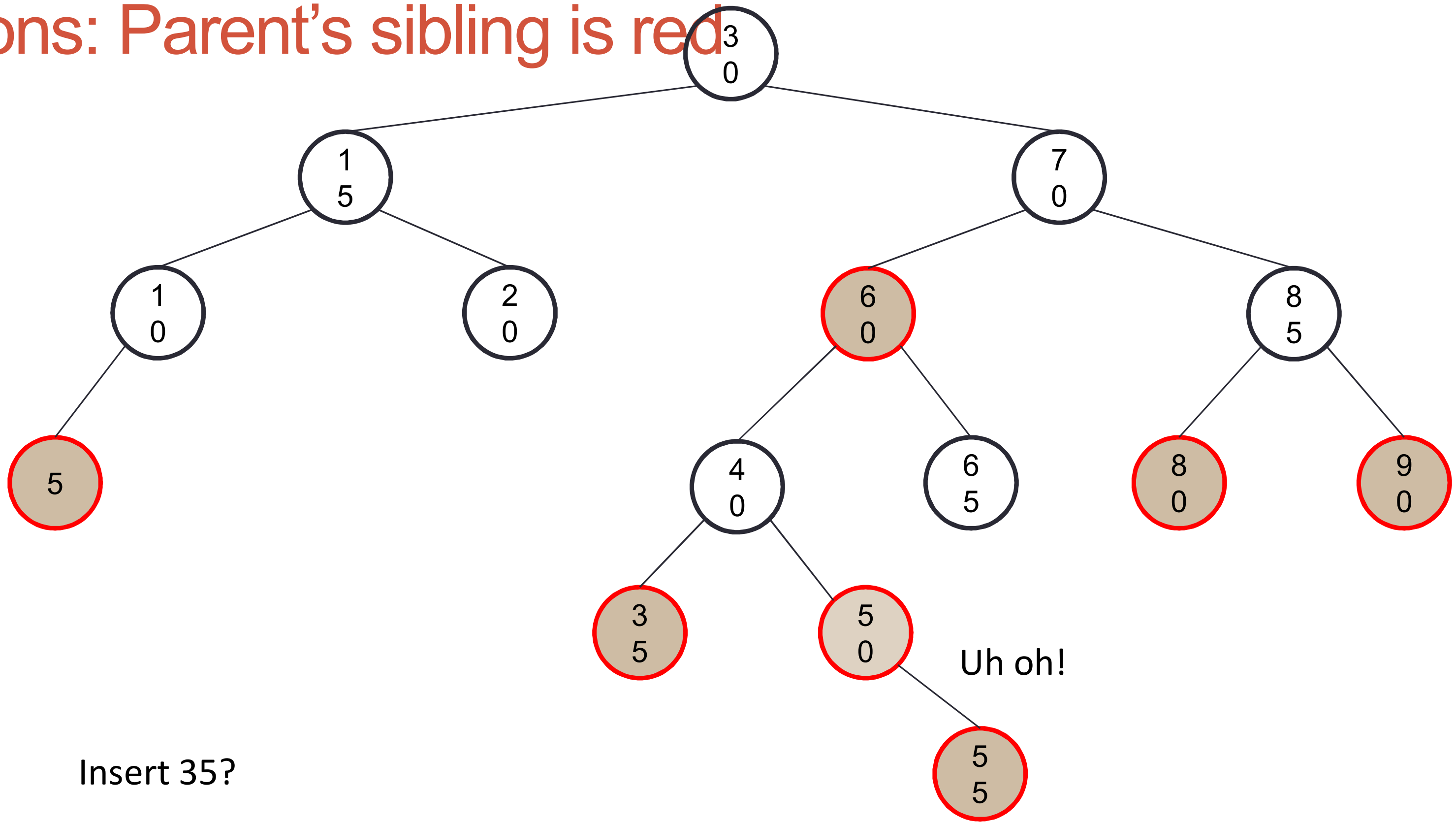
Insert 35?

Insertions: Parent's sibling is red

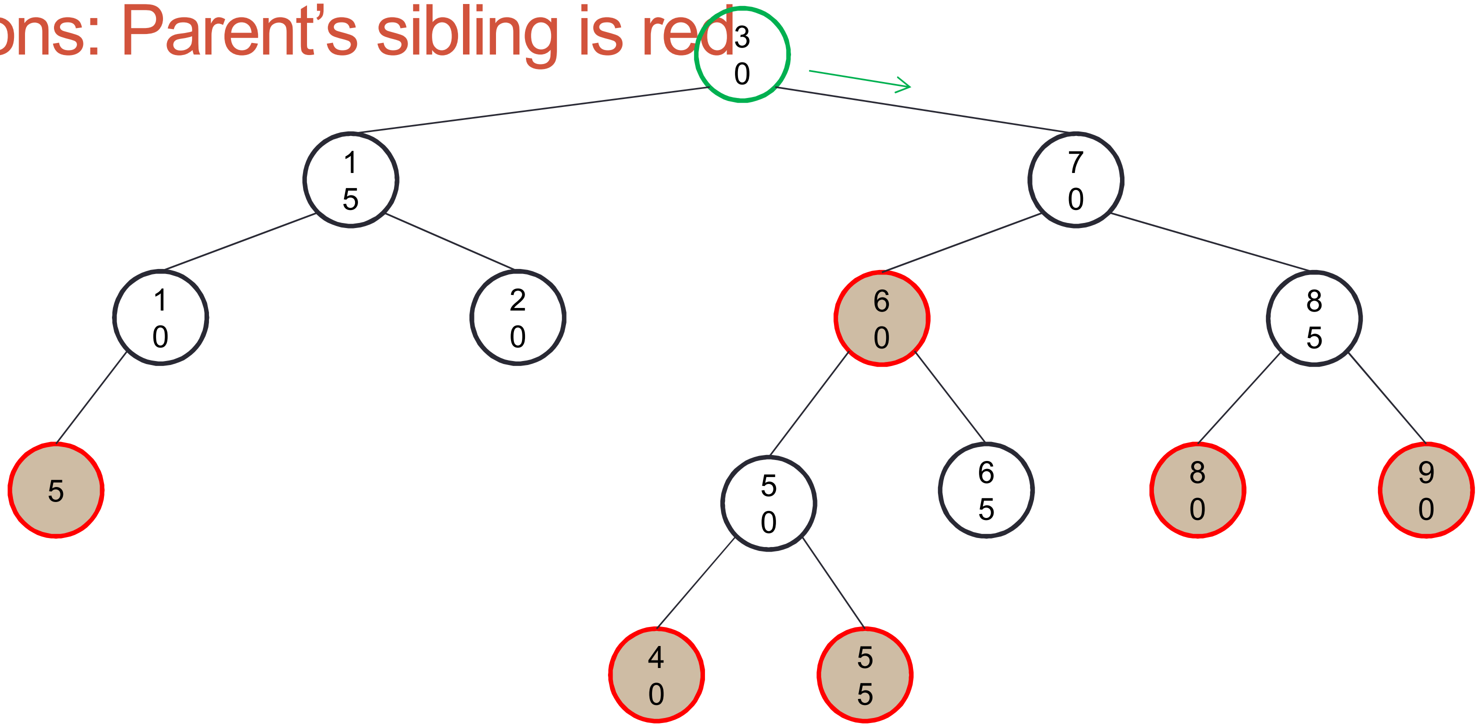


Insert 35?

Insertions: Parent's sibling is red



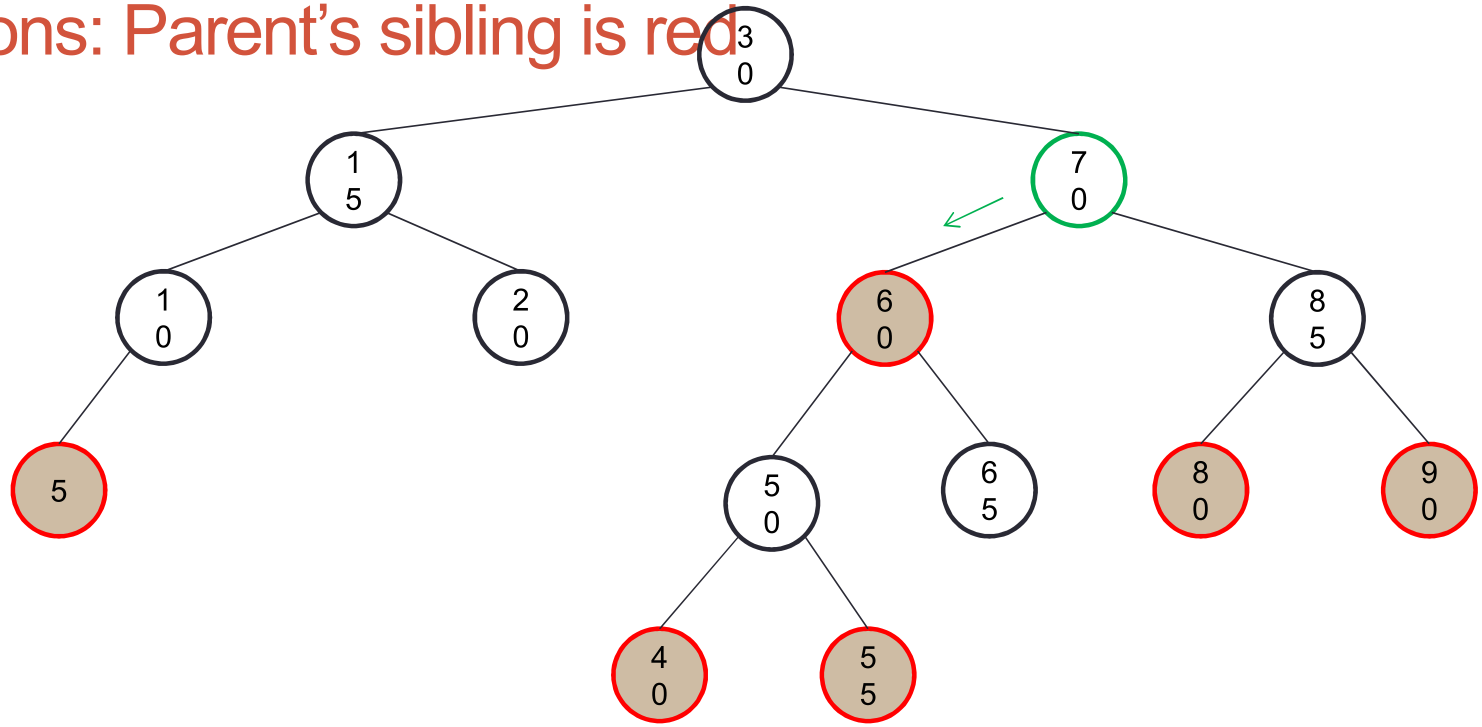
Insertions: Parent's sibling is red



Insert 35?

1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it's children must be black
4. For every node X, every path from X to a null reference must contain the same number of black nodes

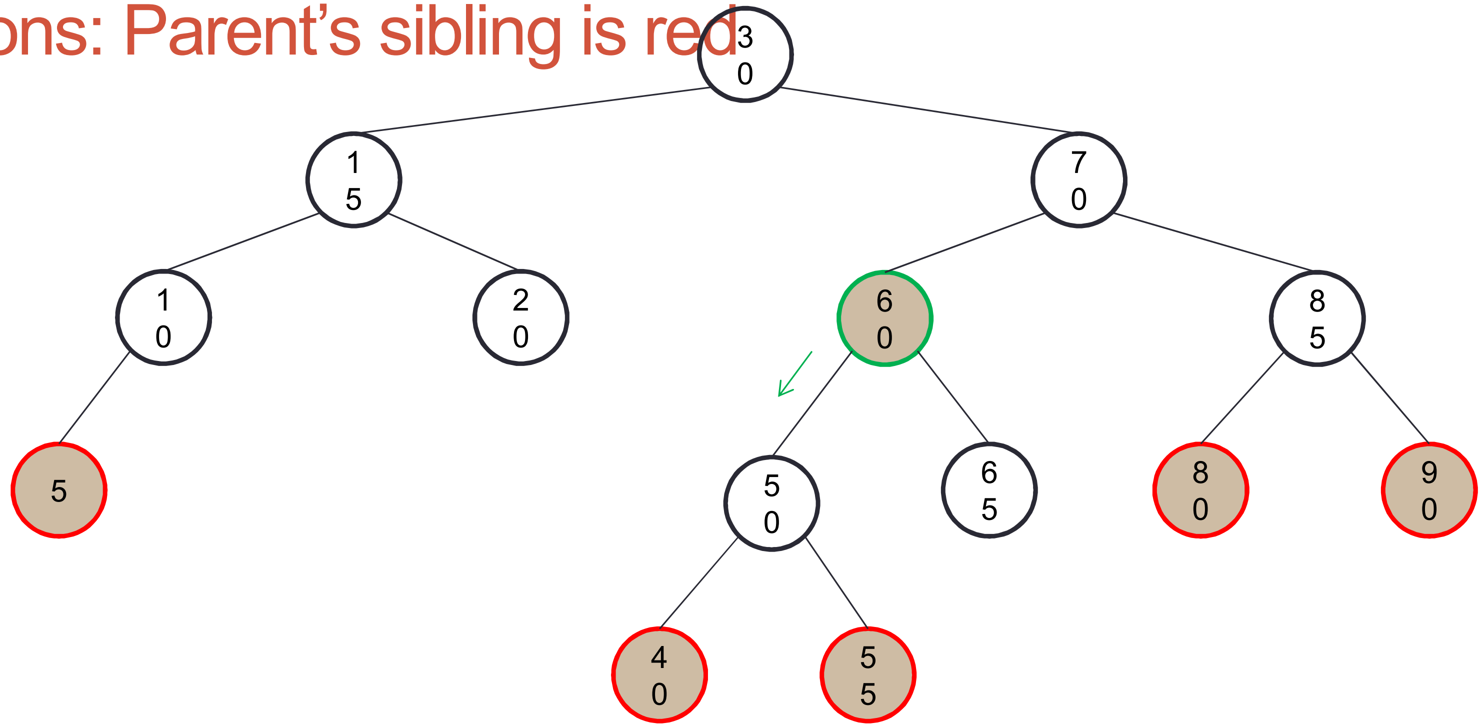
Insertions: Parent's sibling is red



Insert 35?

1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it's children must be black
4. For every node X, every path from X to a null reference must contain the same number of black nodes

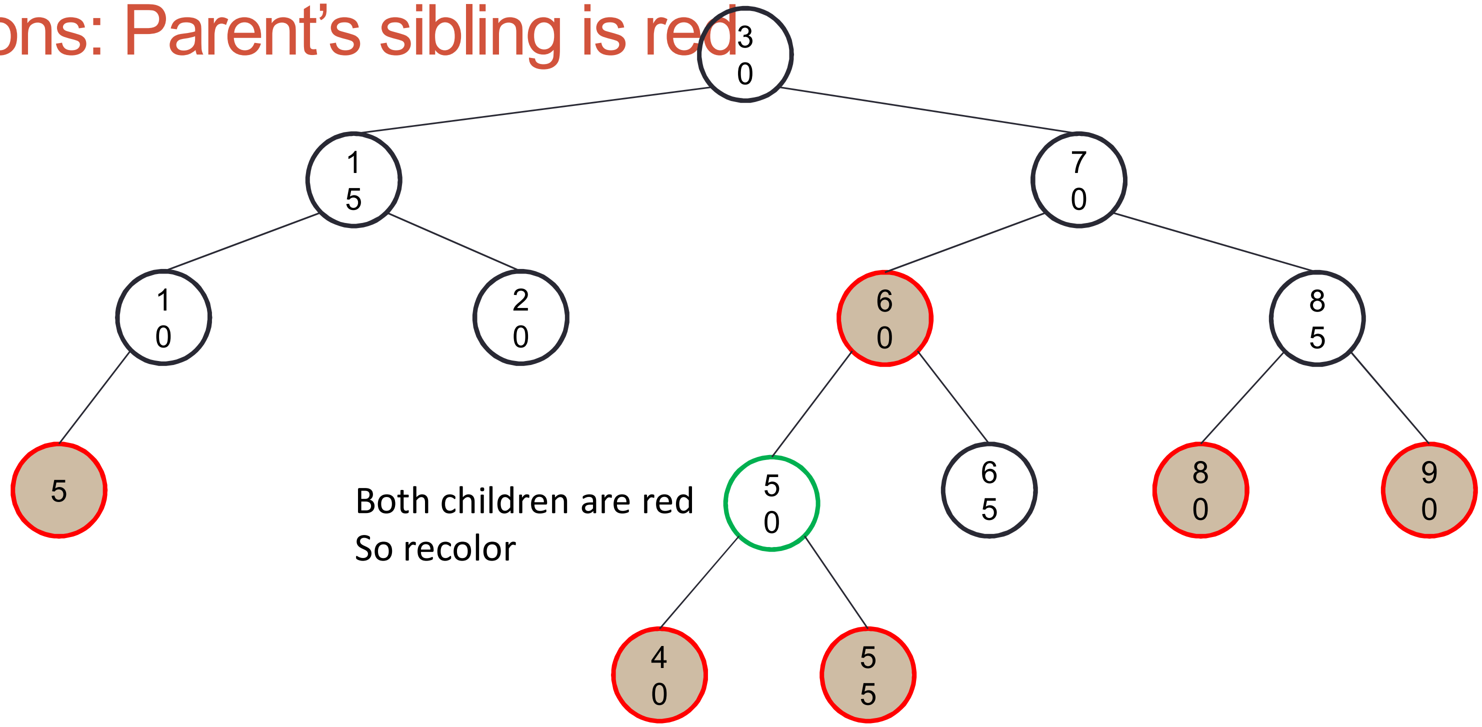
Insertions: Parent's sibling is red



Insert 35?

1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it's children must be black
4. For every node X, every path from X to a null reference must contain the same number of black nodes

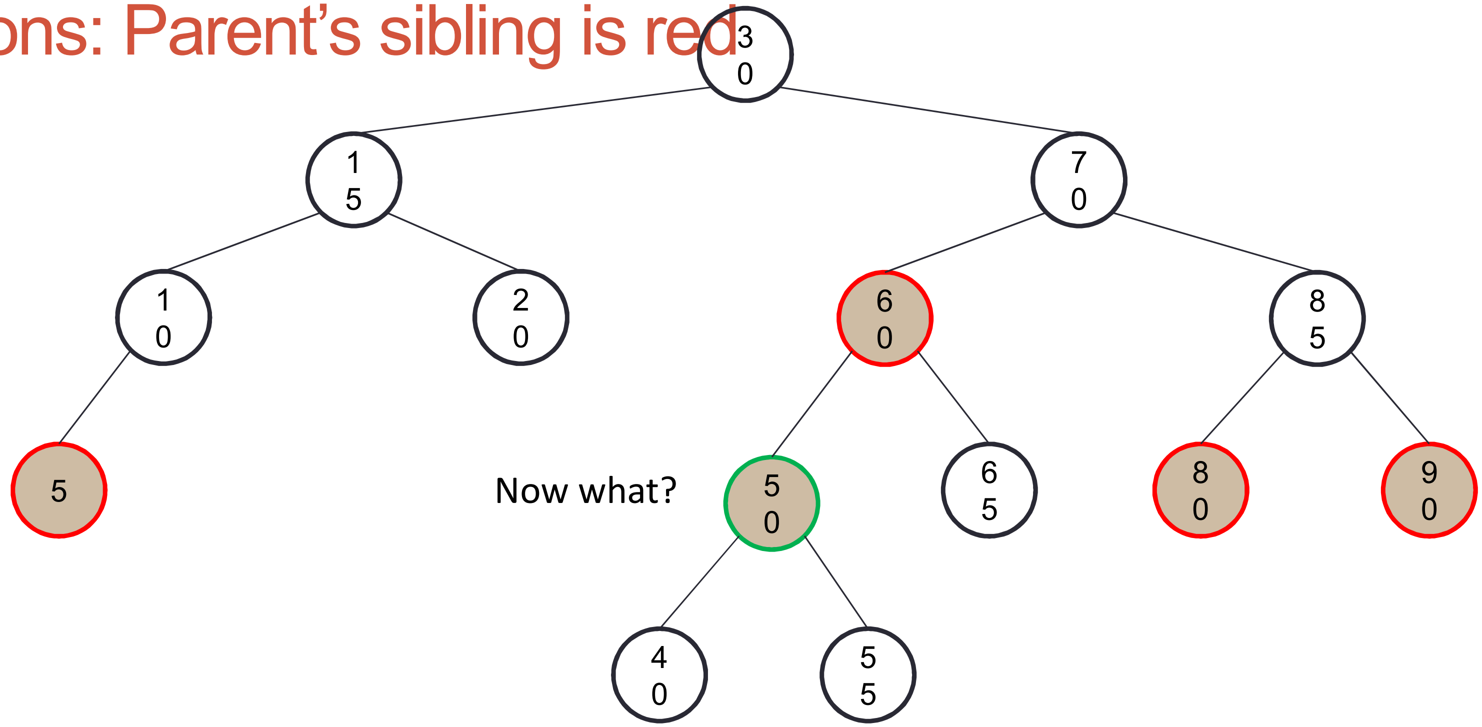
Insertions: Parent’s sibling is red



Insert 35?

1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it’s children must be black
4. For every node X, every path from X to a null reference must contain the same number of black nodes

Insertions: Parent's sibling is red

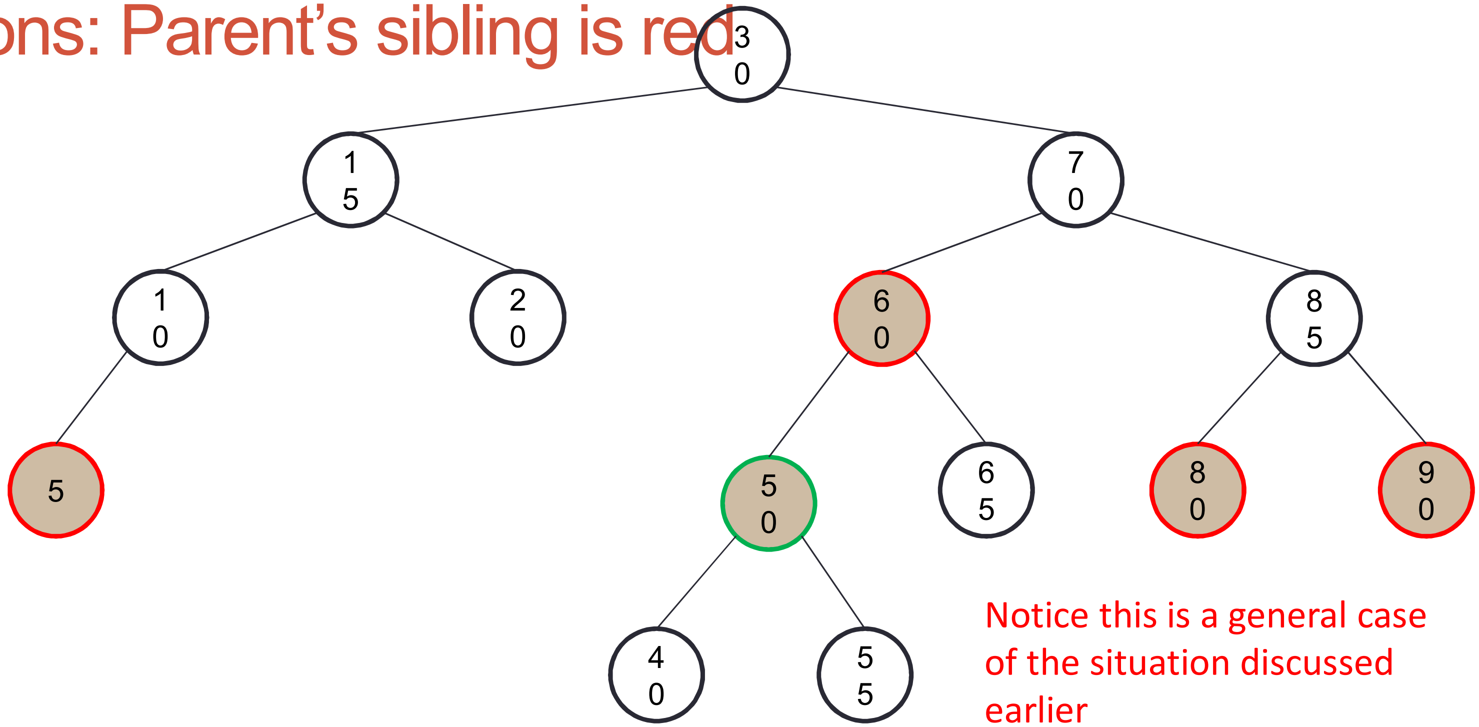


Insert 35?

Property 4 is okay, but 3 is not.
But wait – we've seen this before!

1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it's children must be black
4. For every node X, every path from X to a null reference must contain the same number of black nodes

Insertions: Parent’s sibling is red



Notice this is a general case of the situation discussed earlier

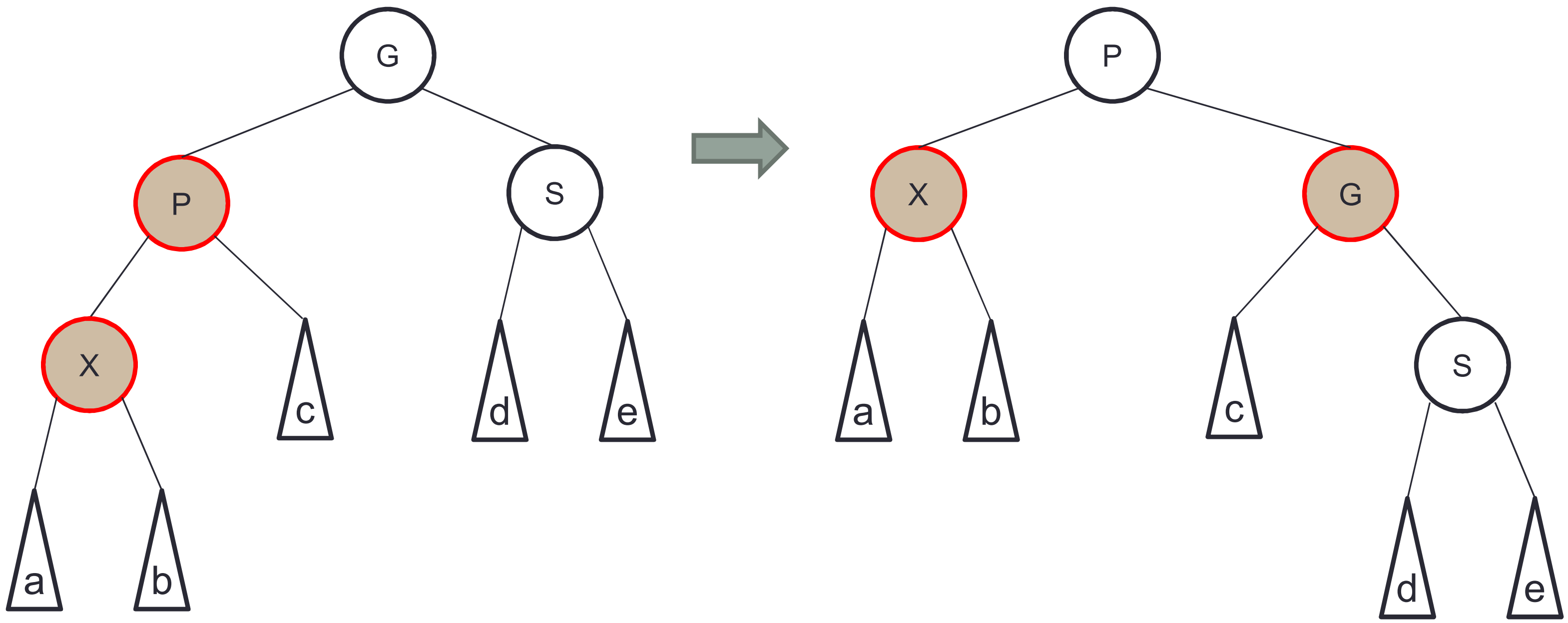
Insert 35?

We’ve seen this case, must Rotate at 70.

1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it’s children must be black
4. For every node X, every path from X to a null reference must contain the same number of black nodes

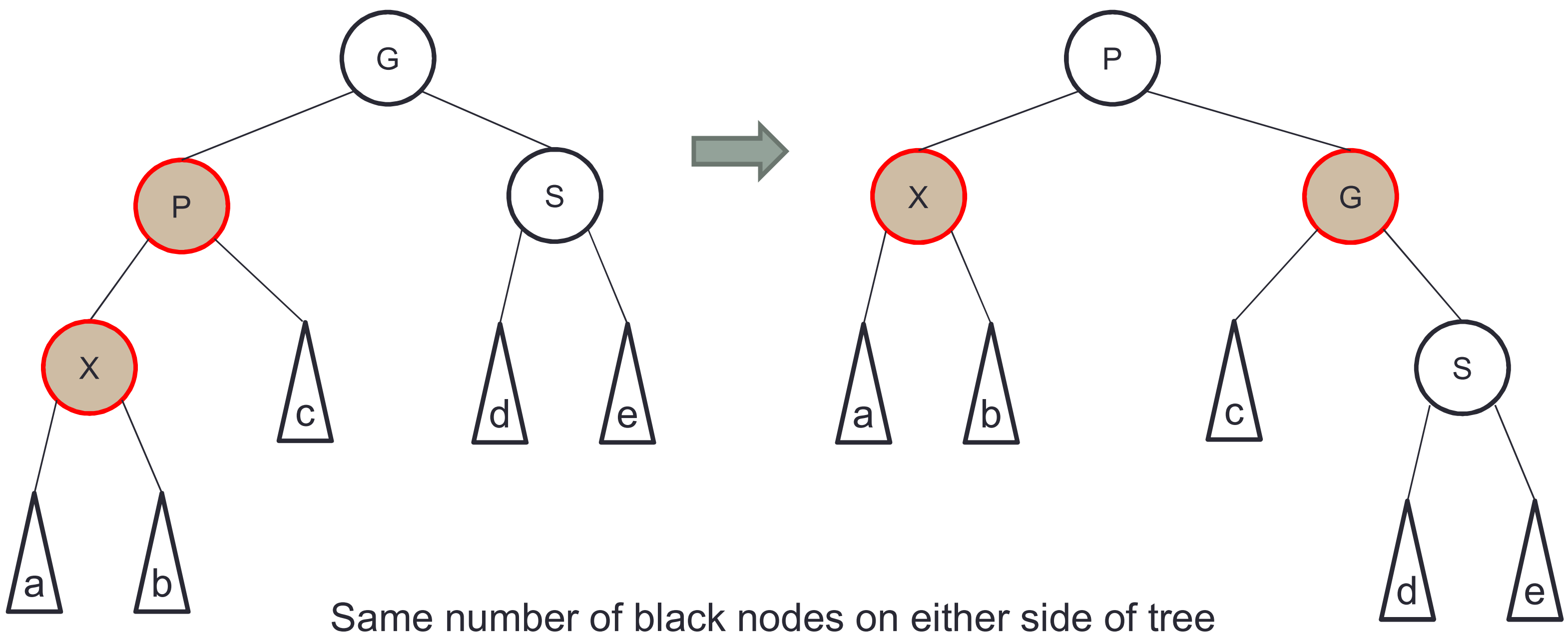
Case 1 in general

(assume this is a legal red-black tree, i.e. there are black nodes hidden in the subtrees)



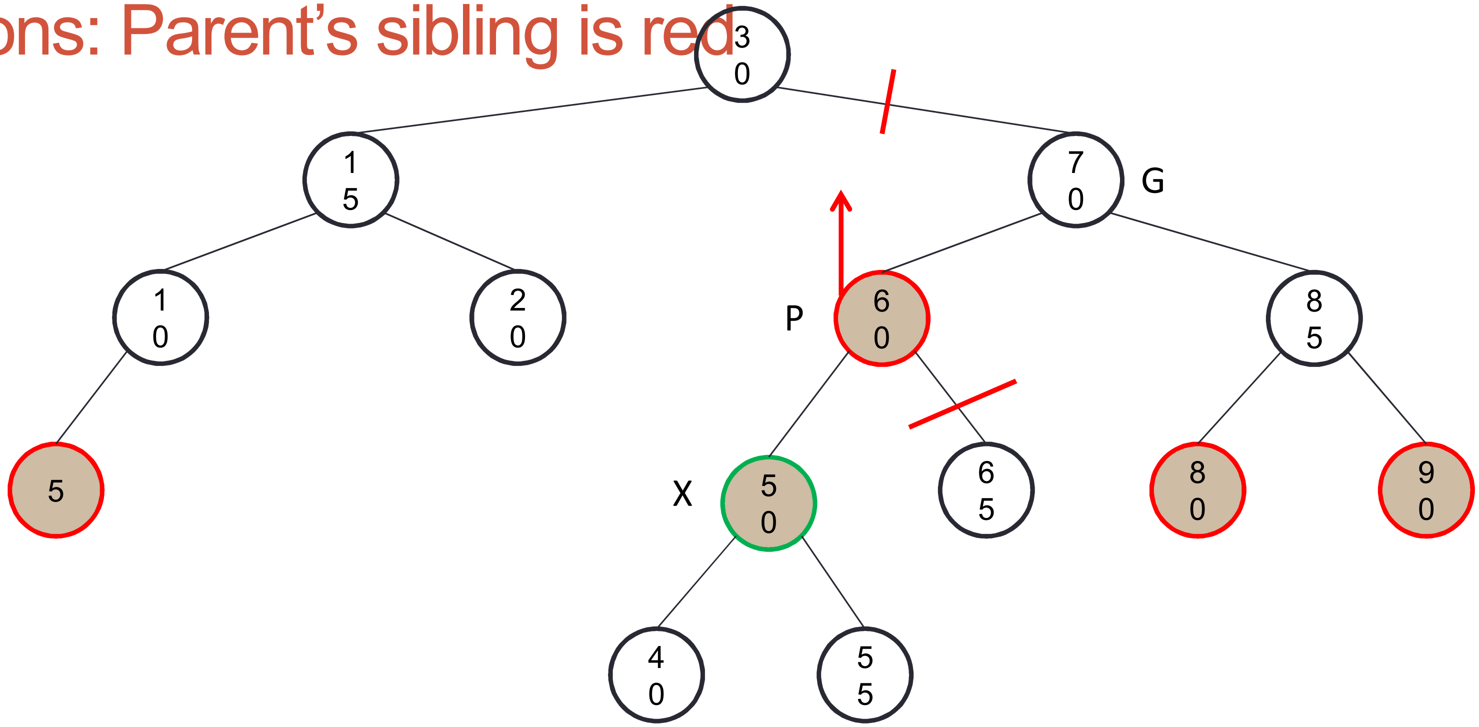
If X's Parent (P) is red, P is a left child of G, X is a left child of P, (and P's sibling (S) is black), then Rotate P right, flip colors of P and G

Case 1 in general



Same number of black nodes on either side of tree
Roots of subtrees a, b and c (and node S) must be black
X's and G's parent is now guaranteed to be black
BST property preserved through AVL rotations

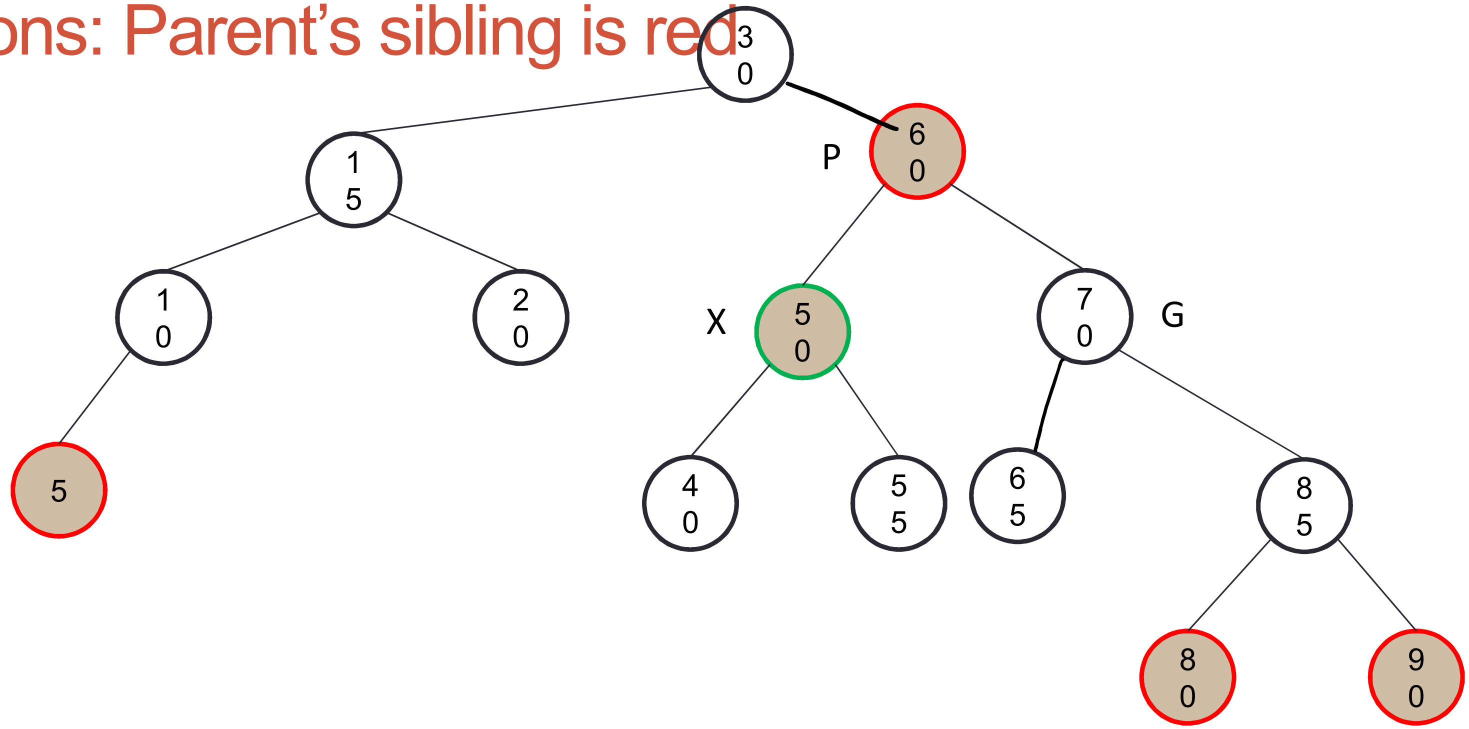
Insertions: Parent's sibling is red



Insert 35?

1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it's children must be black
4. For every node X, every path from X to a null reference must contain the same number of black nodes

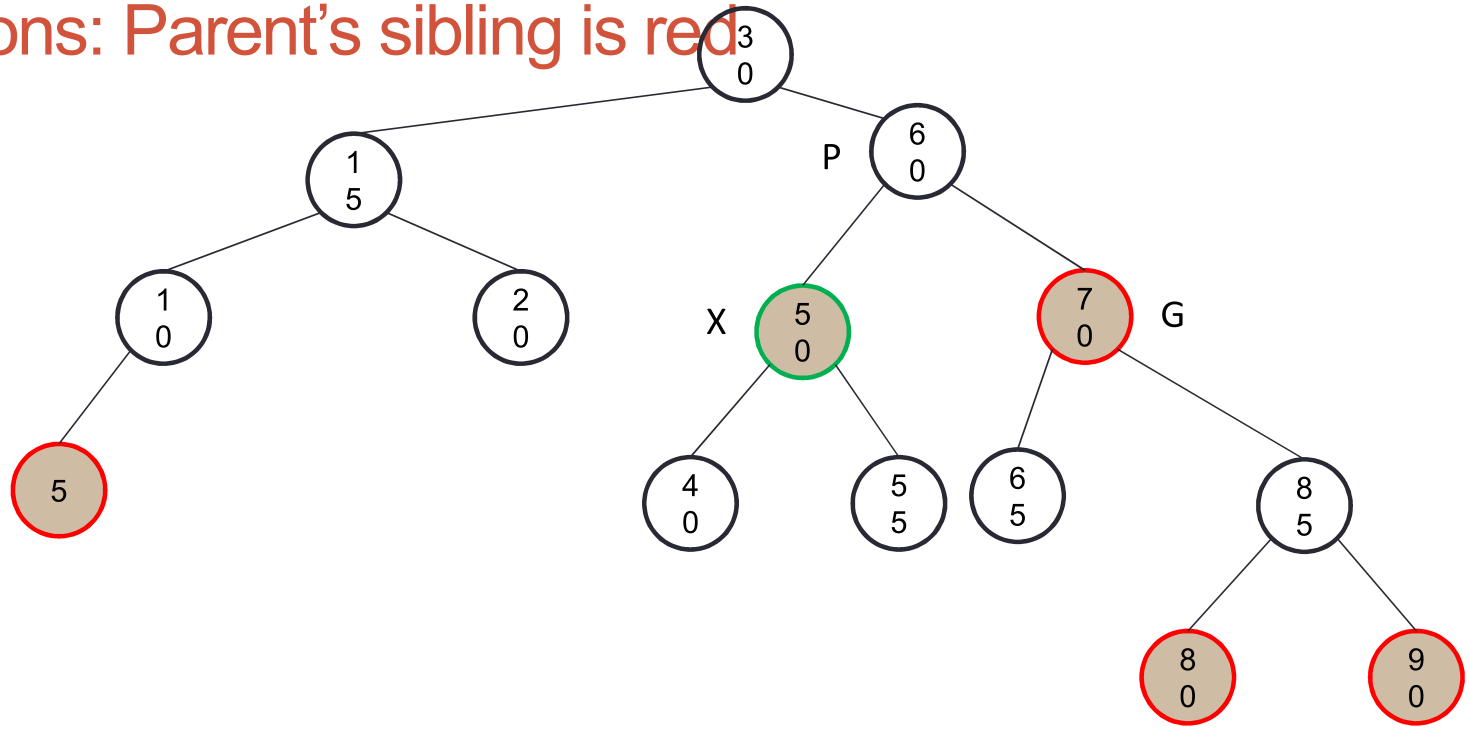
Insertions: Parent's sibling is red



Insert 35?

1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it's children must be black
4. For every node X, every path from X to a null reference must contain the same number of black nodes

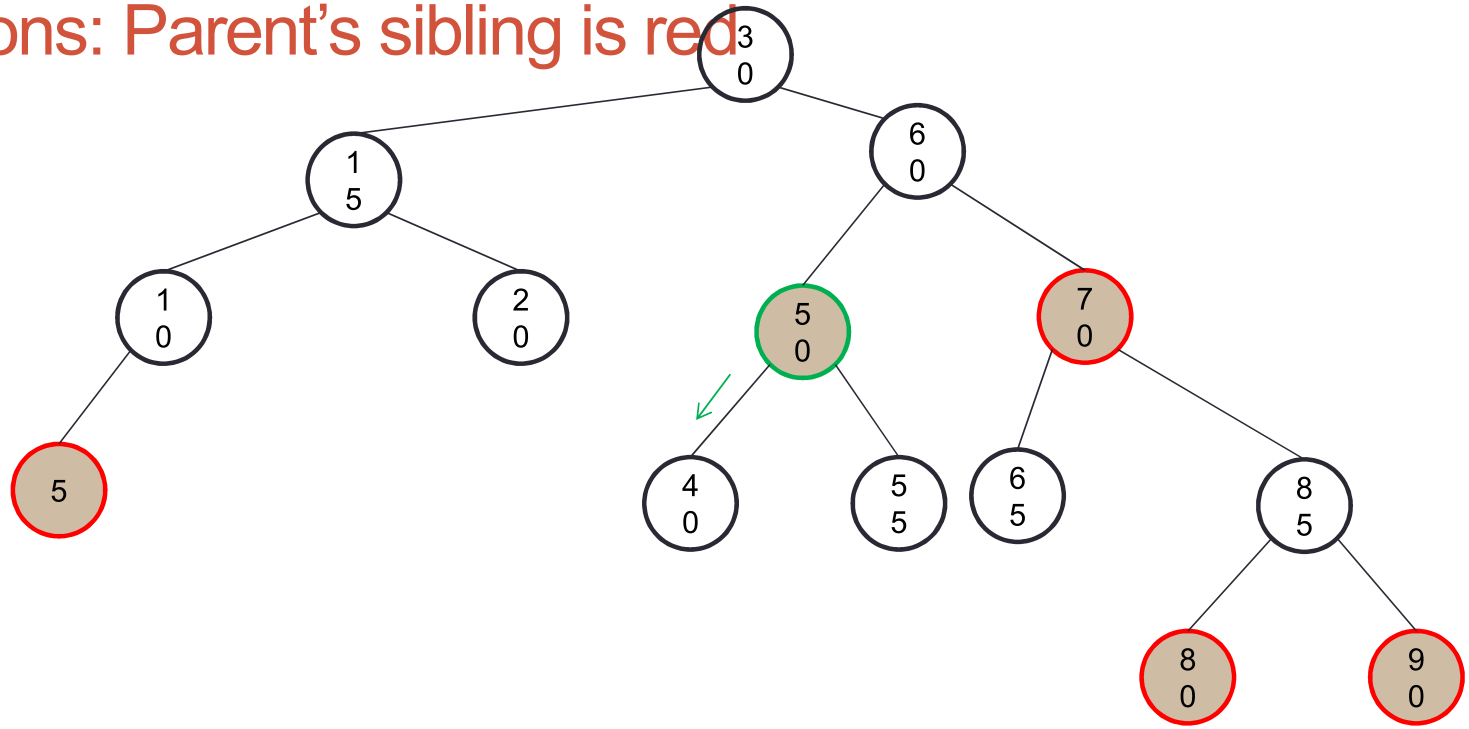
Insertions: Parent's sibling is red



Insert 35?

1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it's children must be black
4. For every node X, every path from X to a null reference must contain the same number of black nodes

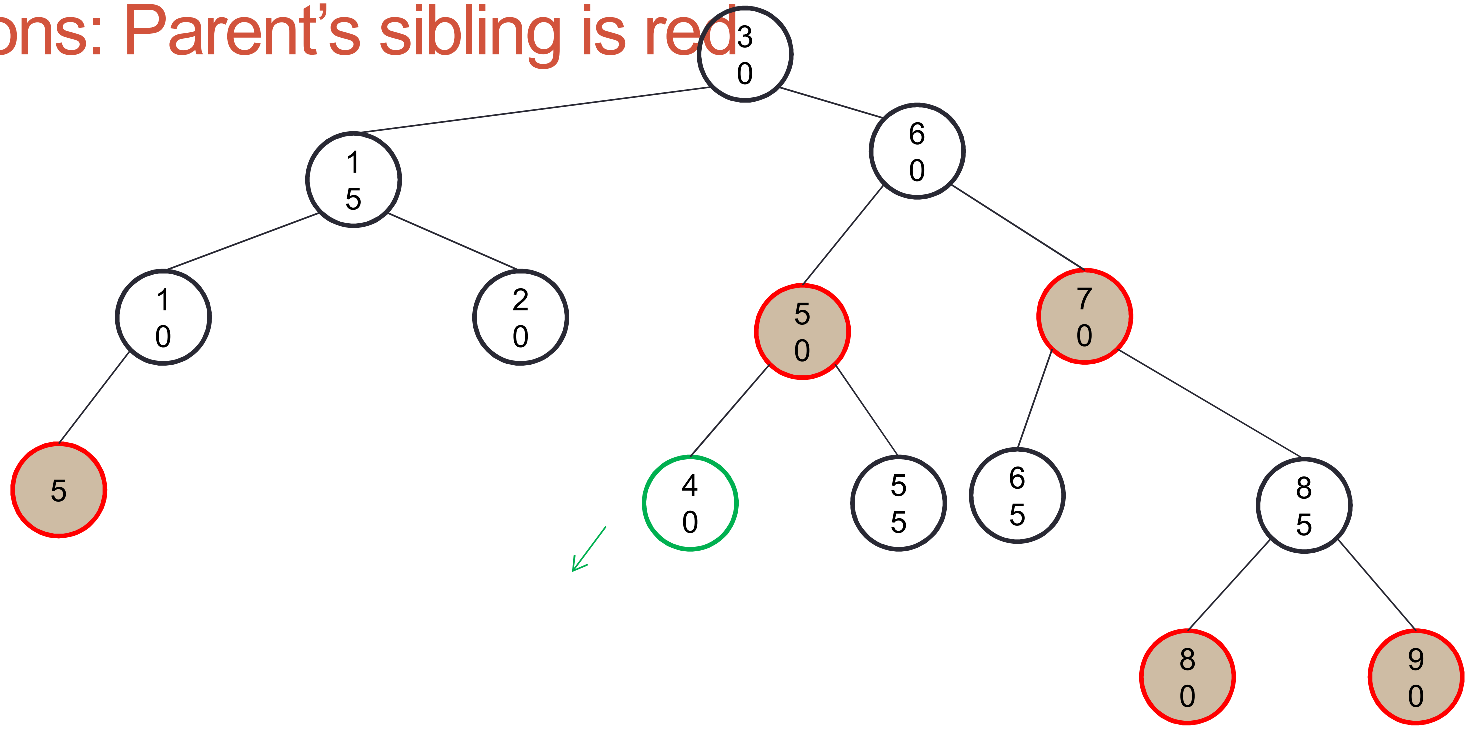
Insertions: Parent's sibling is red



Insert 35?

1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it's children must be black
4. For every node X, every path from X to a null reference must contain the same number of black nodes

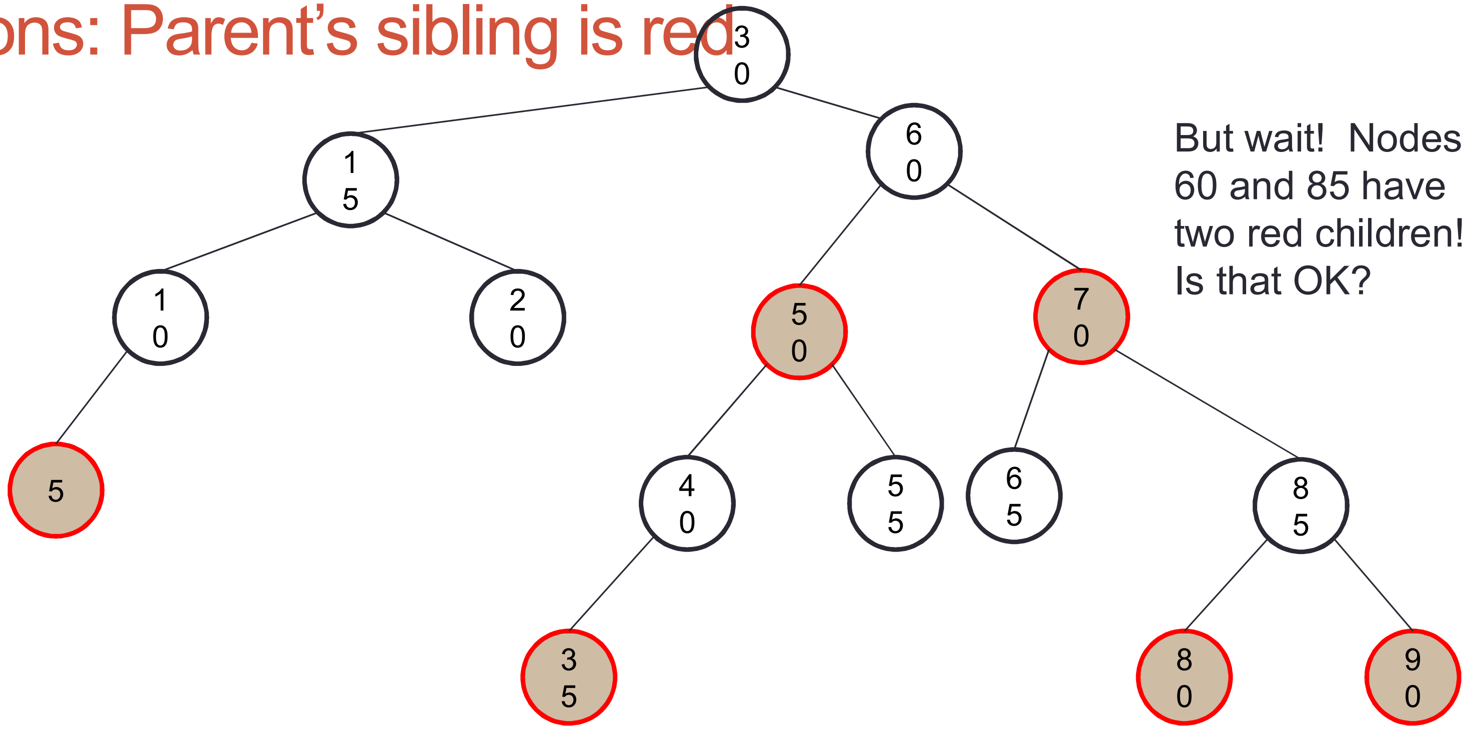
Insertions: Parent's sibling is red



Insert 35?

1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it's children must be black
4. For every node X, every path from X to a null reference must contain the same number of black nodes

Insertions: Parent's sibling is red



But wait! Nodes 60 and 85 have two red children! Is that OK?

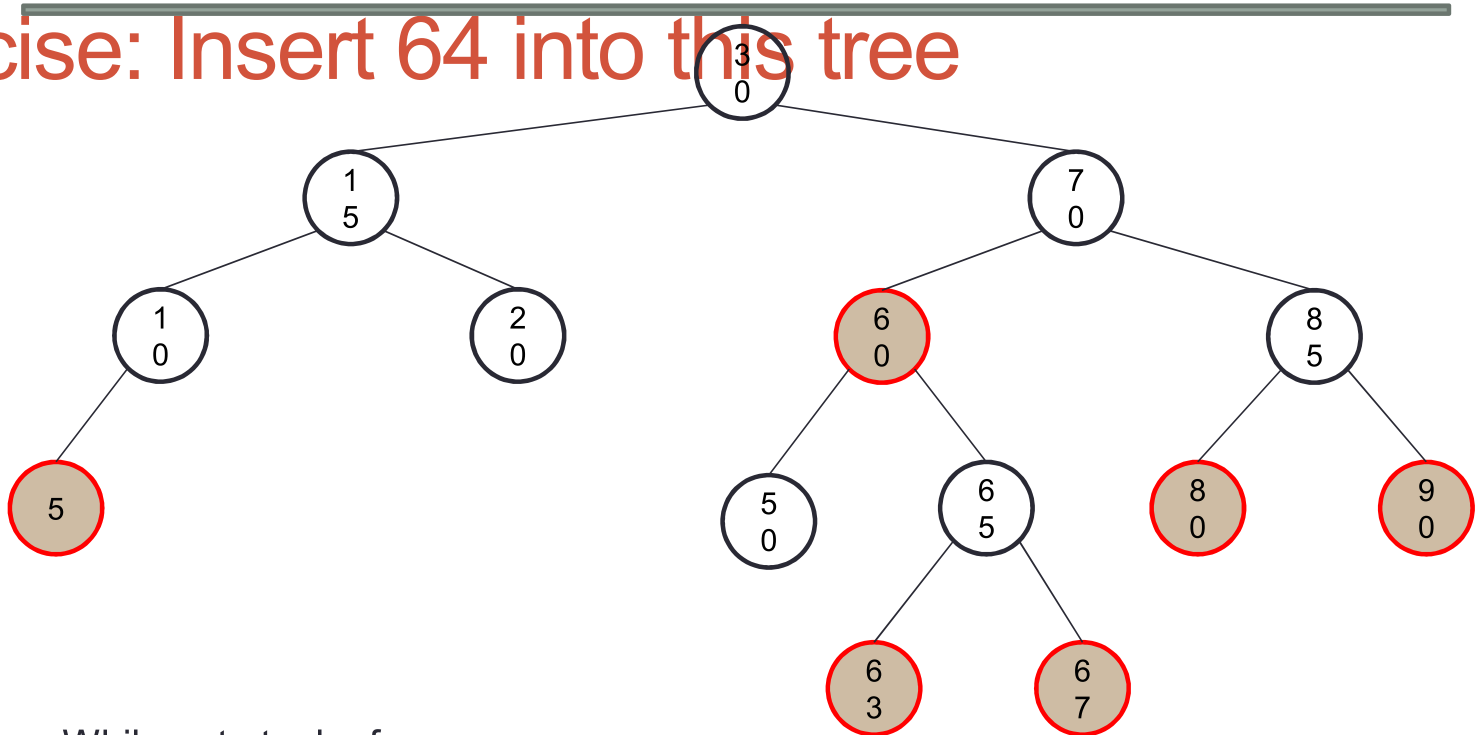
Insert 35? DONE!

1. Nodes are either red or black
2. Root is always black
3. If a node is red, all it's children must be black
4. For every node X, every path from X to a null reference must contain the same number of black nodes

Can any node have 2 red children?

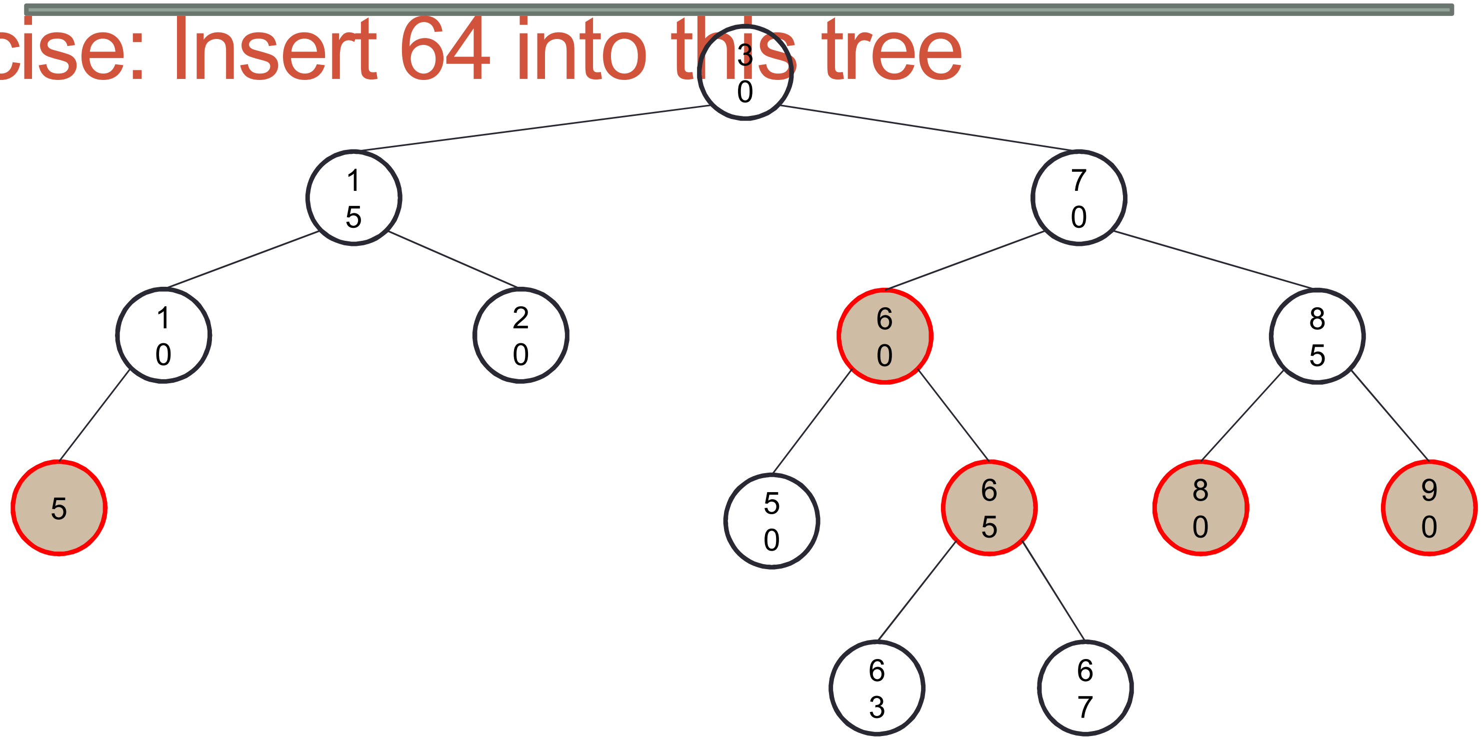
- As we descend the tree, we detect if a node X has 2 red children, and if so we do an operation to change the situation
- Note that in doing so:
  - we may change things so that a node above X now has 2 red children, where it didn't before! (example: node 60 after we insert 35)
  - if we have to do a double rotation, we will move X up and recolor it so that it becomes black, and has 2 red children itself! (example: work through inserting 64 in the tree on the following page)
- But neither of these is a problem, because
  - it never violates any of the properties of red-black trees (those 2 red nodes will always have a black parent, for example),
  - and the 2 red siblings will be too "high" in the tree for either of them to be the sibling of the parent of any red node that we find or create when we continue this descent of the tree

Exercise: Insert 64 into this tree



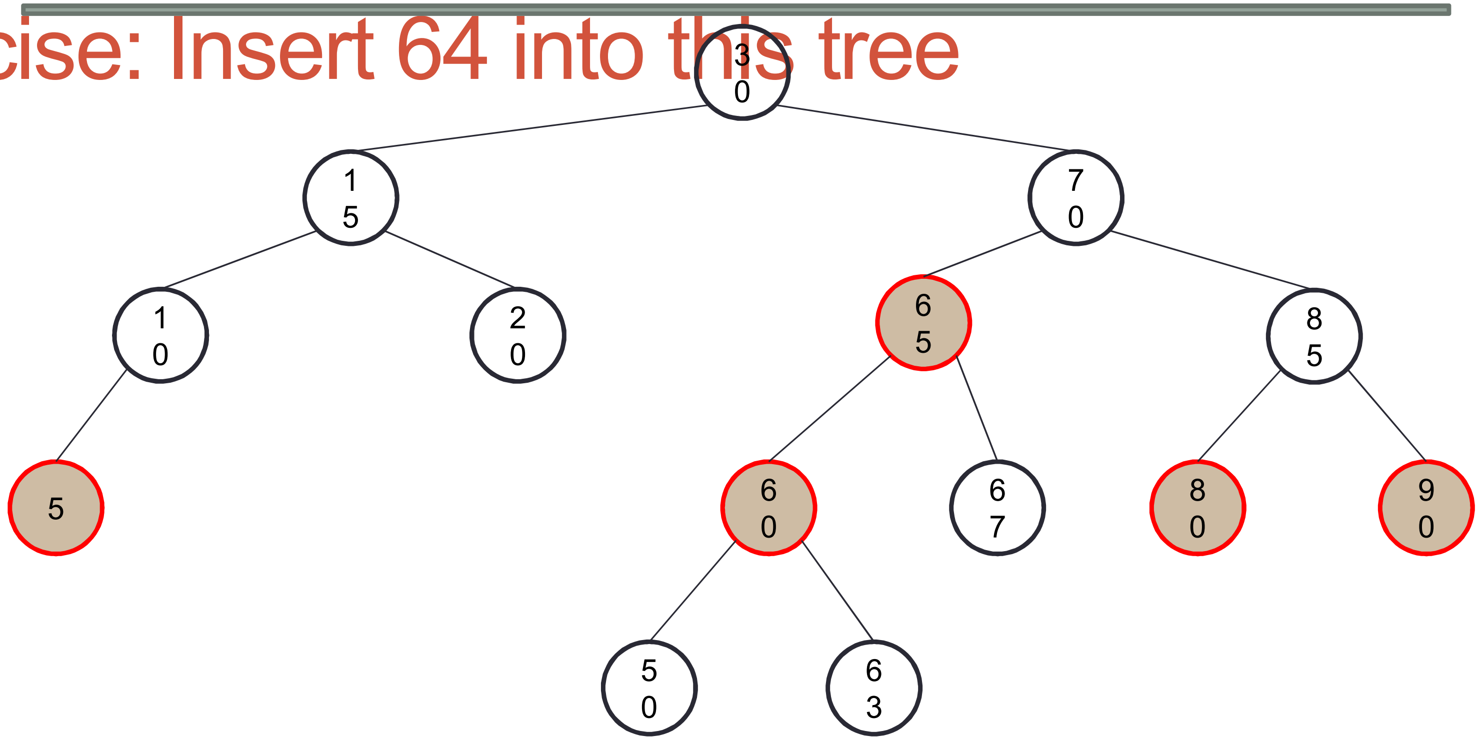
- While not at a leaf:
  - Move down the tree to where node should be placed
  - If you encounter a node with two red children, recolor, then perform any necessary rotations to fix the tree
- Insert the node
- Perform any necessary rotations to fix the tree

Exercise: Insert 64 into this tree



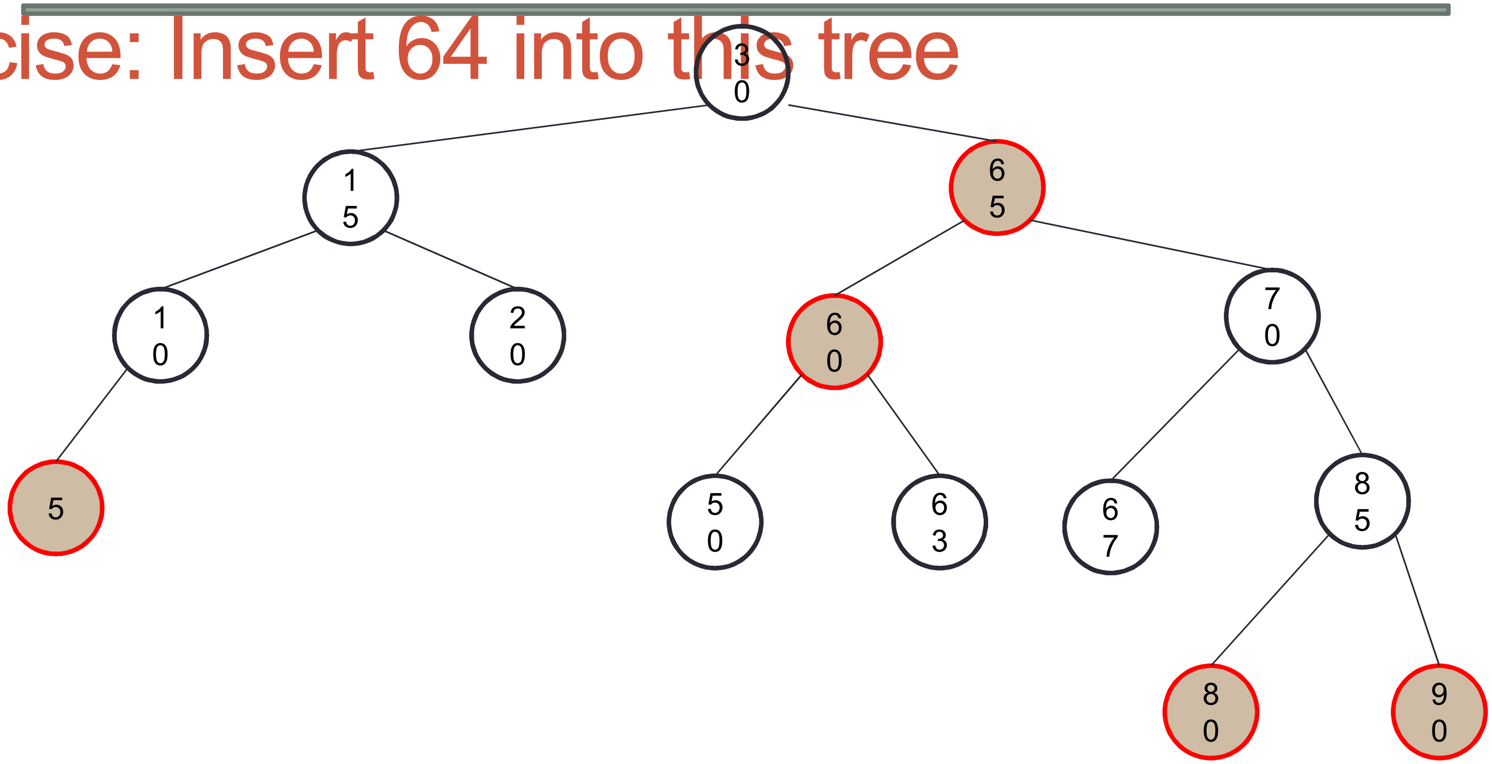
Recolor

Exercise: Insert 64 into this tree



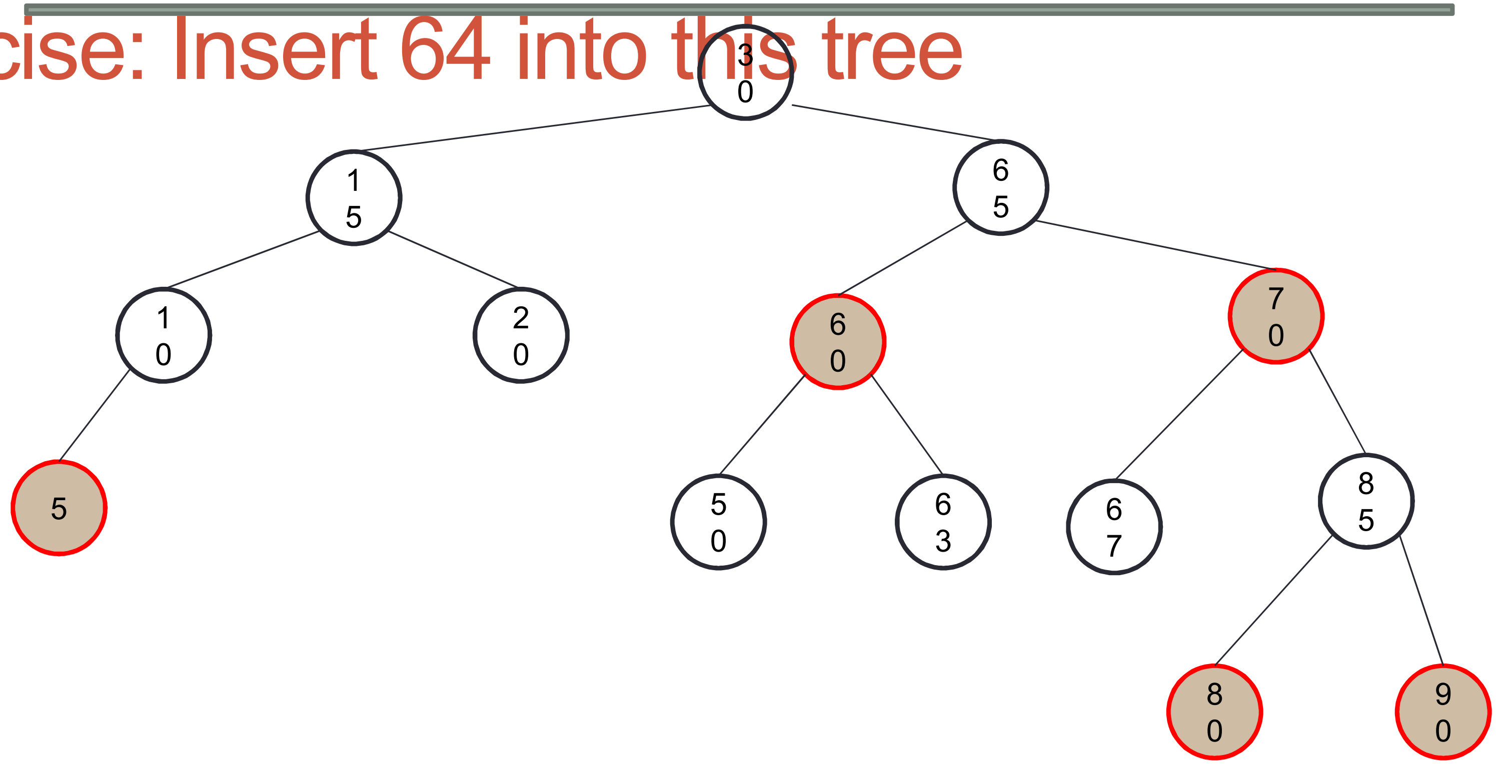
Double rotation (rotation 1)

Exercise: Insert 64 into this tree



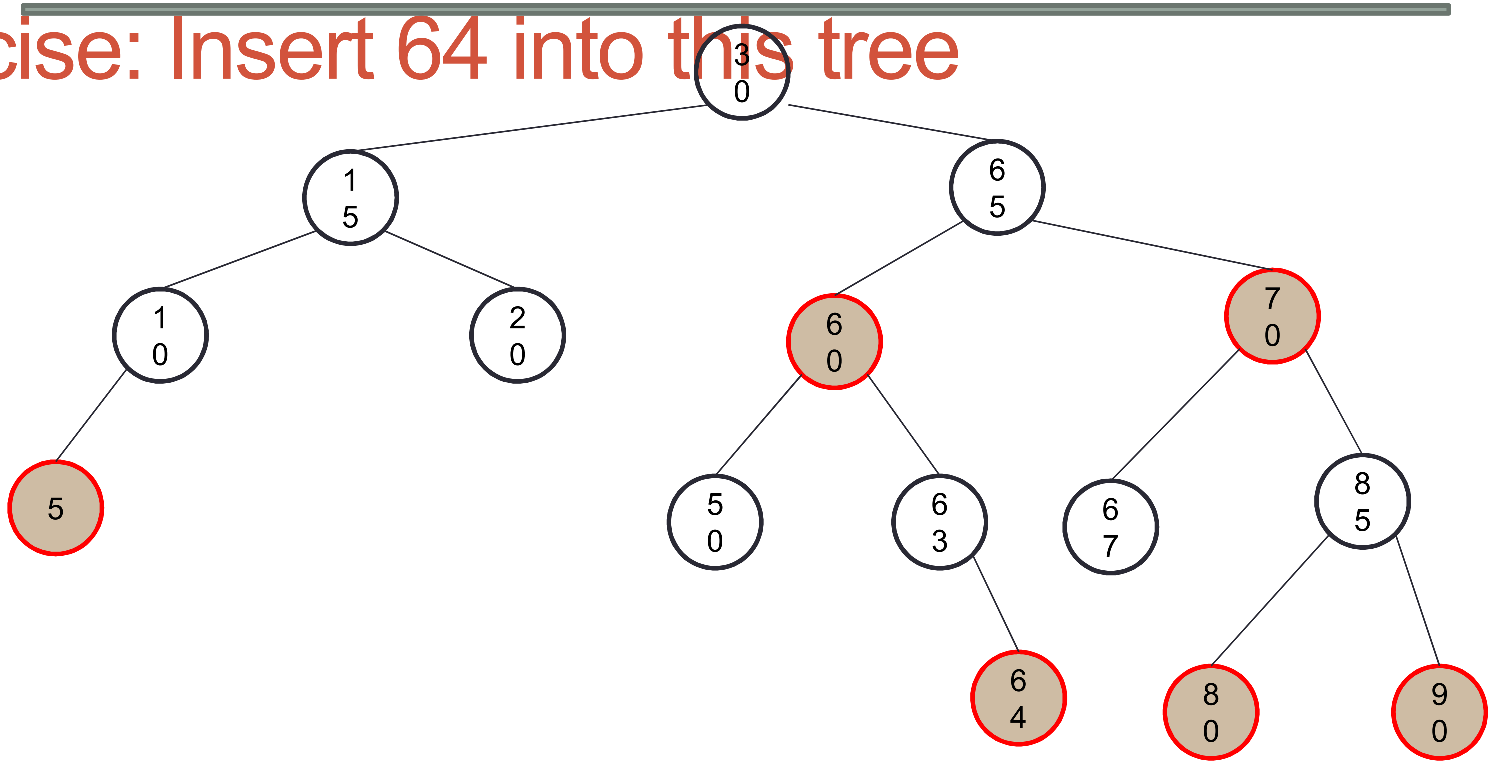
Double rotation (rotation 2)

Exercise: Insert 64 into this tree



Recolor

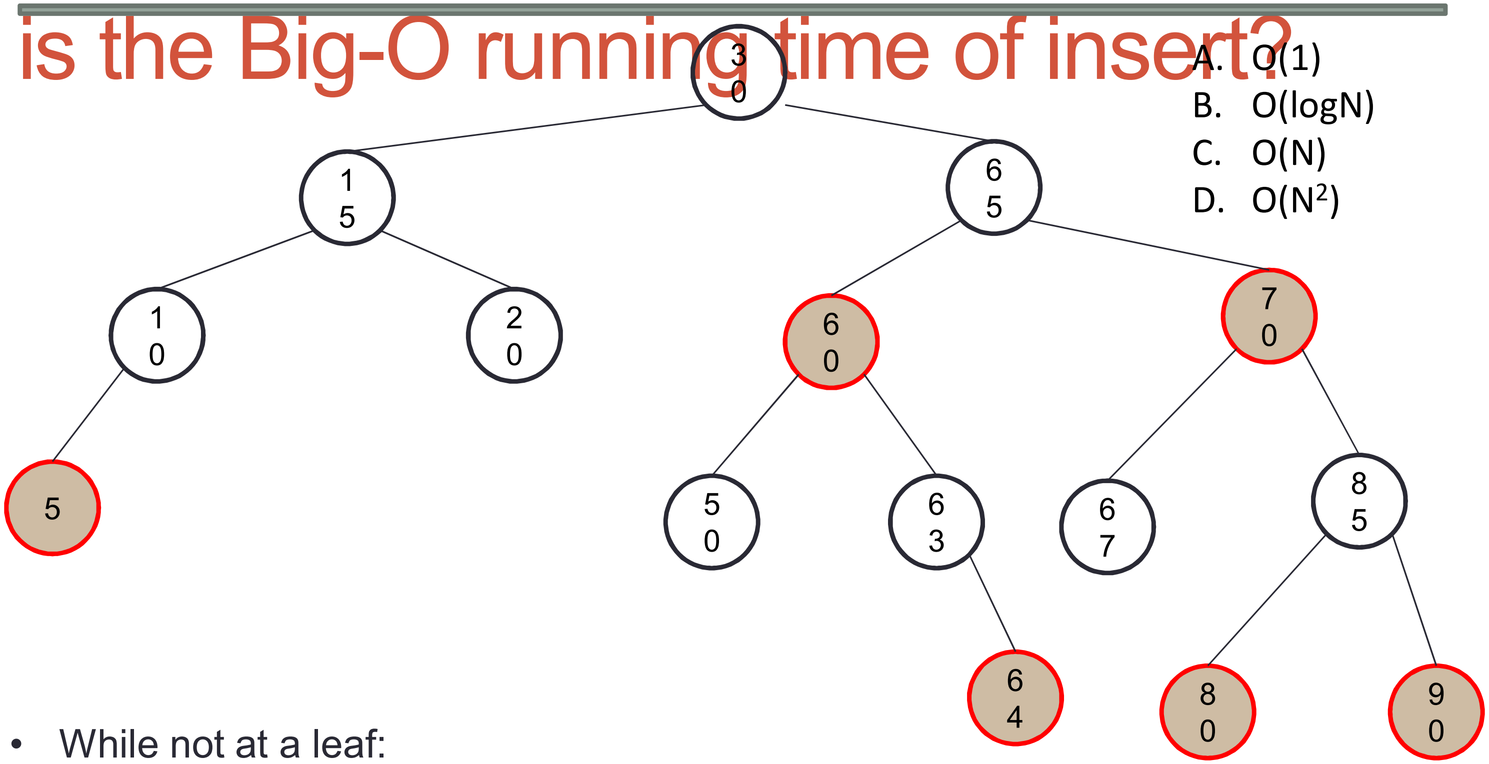
Exercise: Insert 64 into this tree



Insert

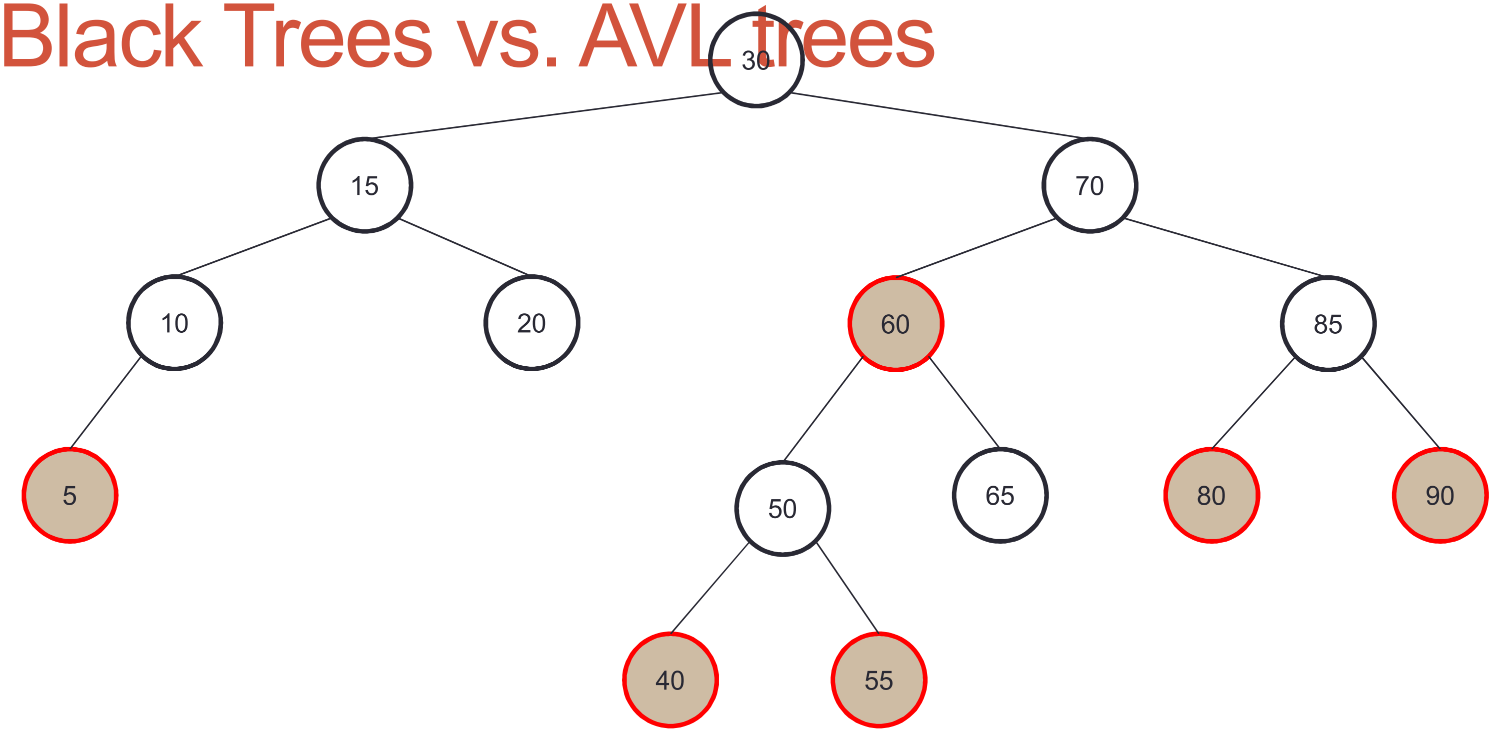
What is the Big-O running time of insert?

A. O(1)
B. O(logN)
C. O(N)
D. $O(N^2)$



- While not at a leaf:
  - Move down the tree to where node should be placed
  - If you encounter a node with two red children, recolor, then perform any necessary rotations to fix the tree
- Insert the node
- Perform any necessary rotations to fix the tree

Red-Black Trees vs. AVL trees

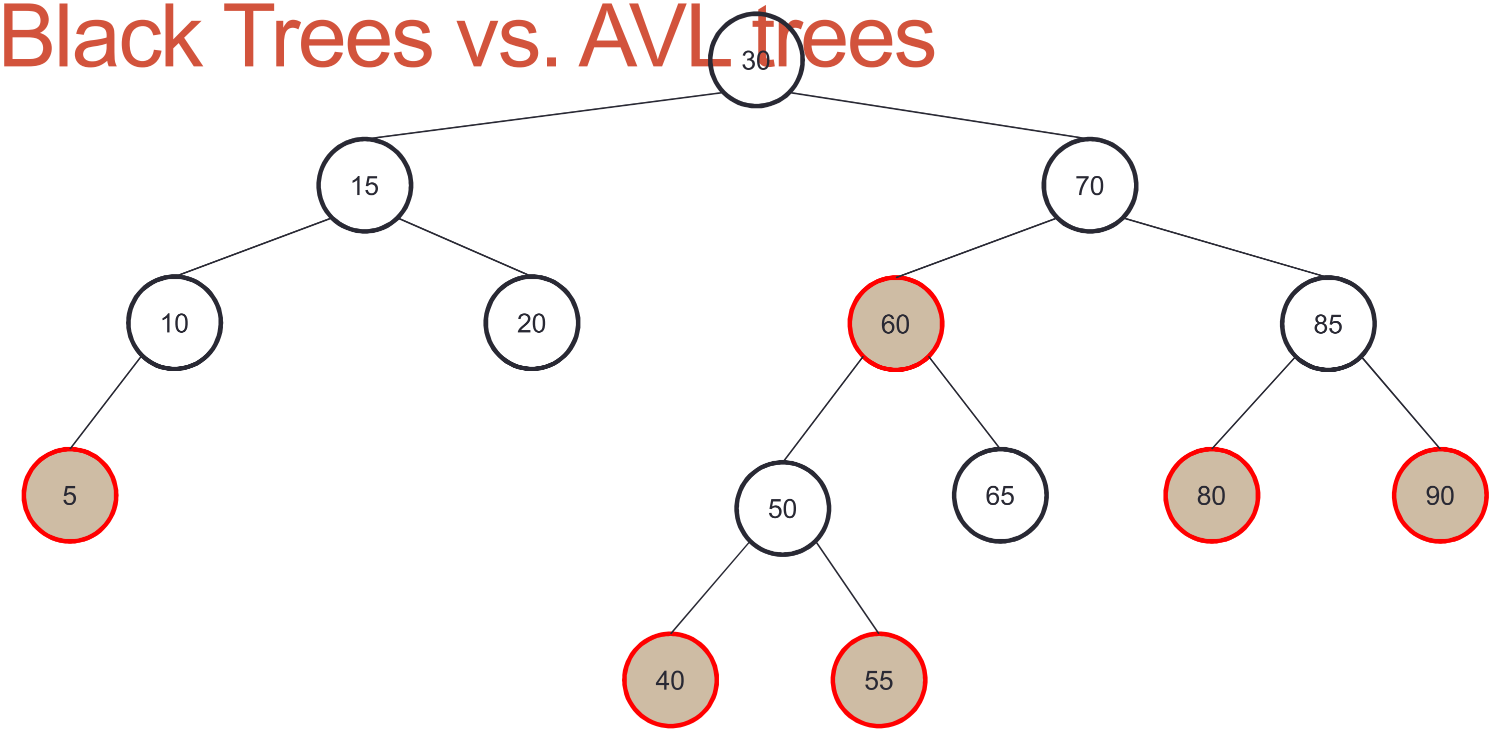


Is this an AVL tree?

A. Yes

B. No

Red-Black Trees vs. AVL trees

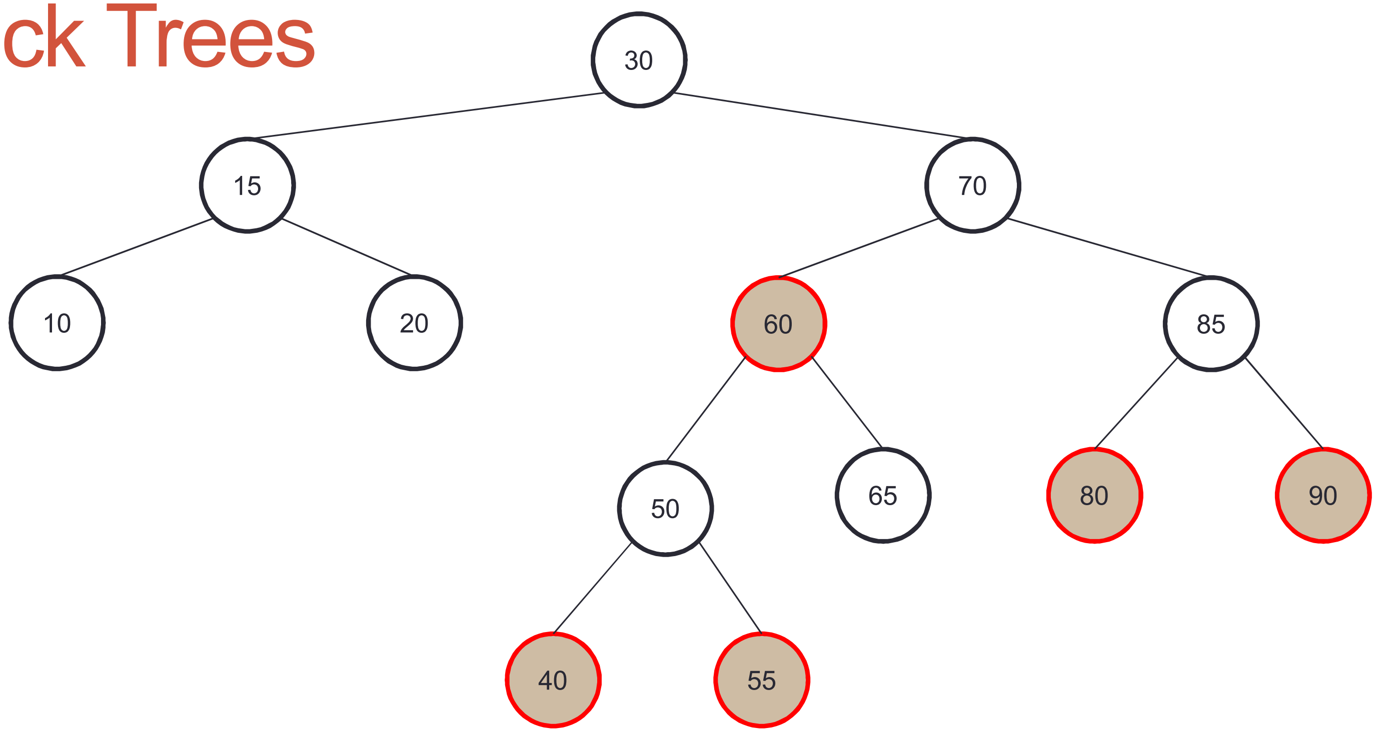


Is this an AVL tree?
Yes

Are all red black trees AVL trees?

A. Yes
B. No

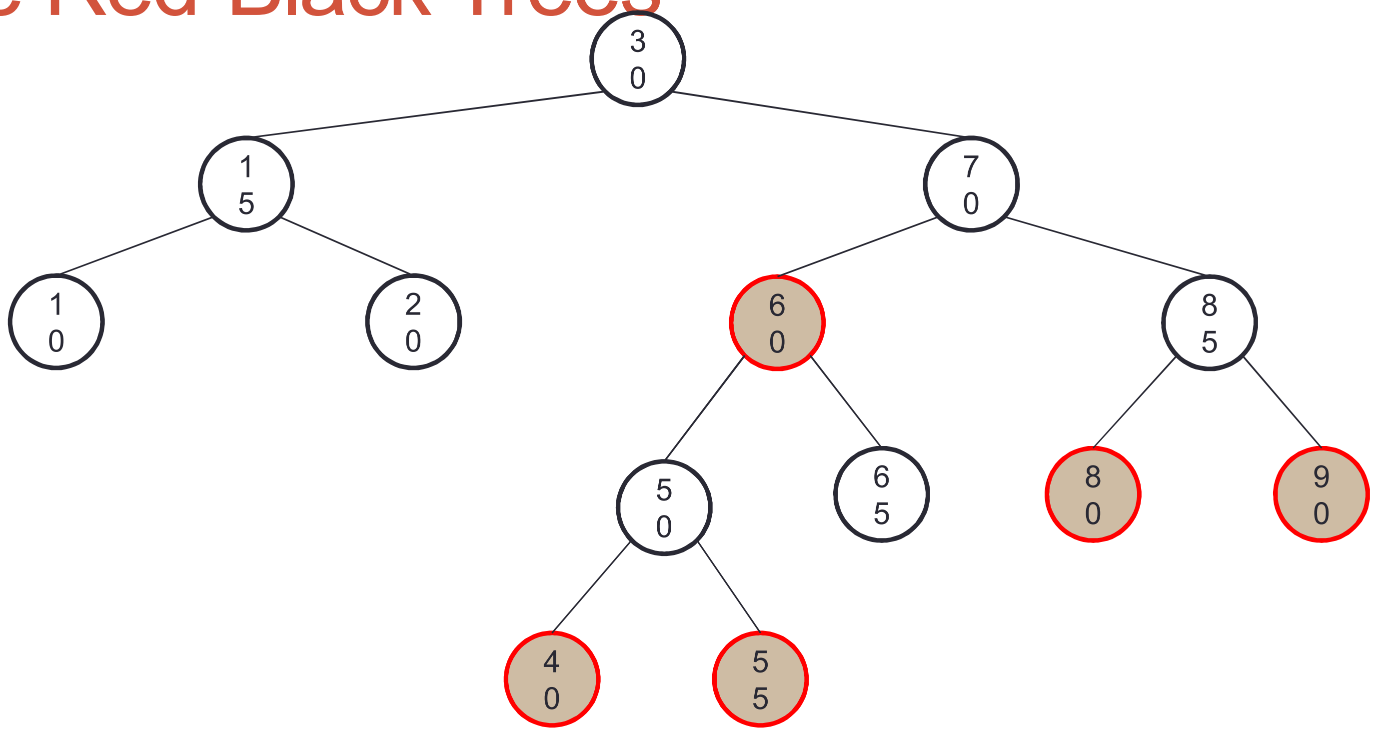
Red-Black Trees



Is this an AVL tree?
(Not anymore)

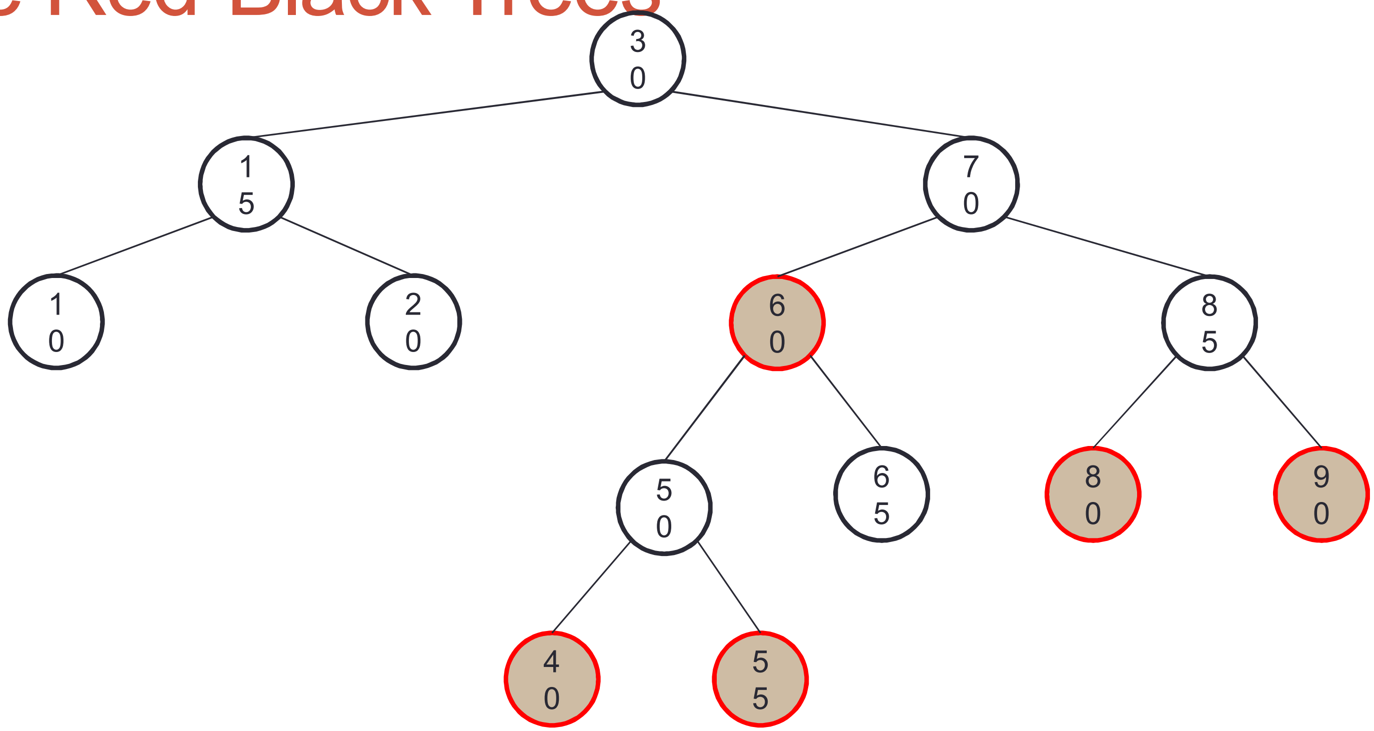
Are all red black trees AVL trees?
No

Why use Red-Black Trees



Fast to insert, slightly longer to find (but still guaranteed O(log(N)))

Why use Red-Black Trees



Faster to insert (than AVL): RBT insertion traverses the tree once instead of twice
Slower to find (that AVL): RBTs are generally slightly taller than AVL trees