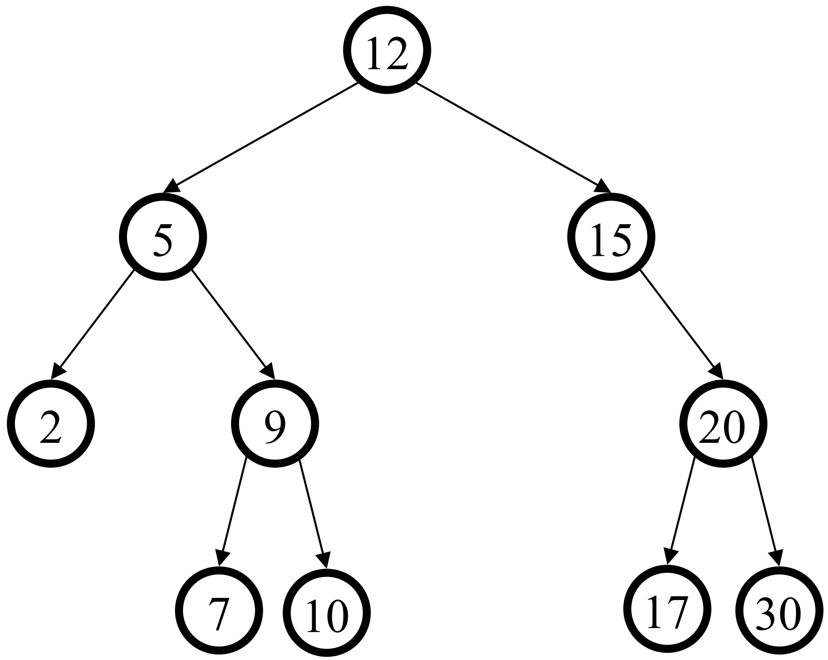
Announcements

- Lilian’s office hours rescheduled: Fri 2-4pm
- HW2 out tomorrow, due Thursday, 7/7

Deletion in BST



Why might deletion be harder than insertion?

Deletion

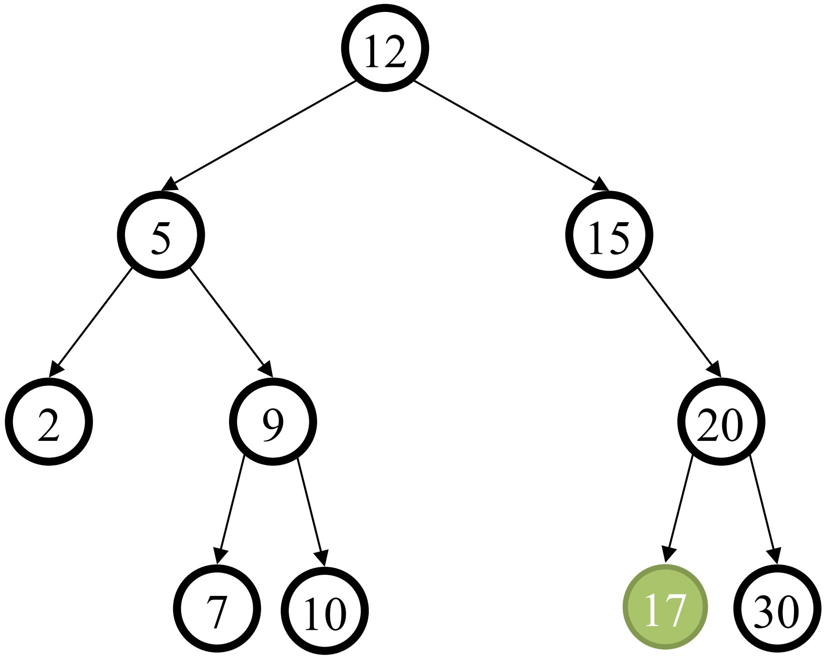
- Removing an item disrupts the tree structure

- Basic idea: **`find`** the node to be removed, then "fix" the tree so that it is still a binary search tree

- Three cases:
  - Node has no children (leaf)
  - Node has one child
  - Node has two children

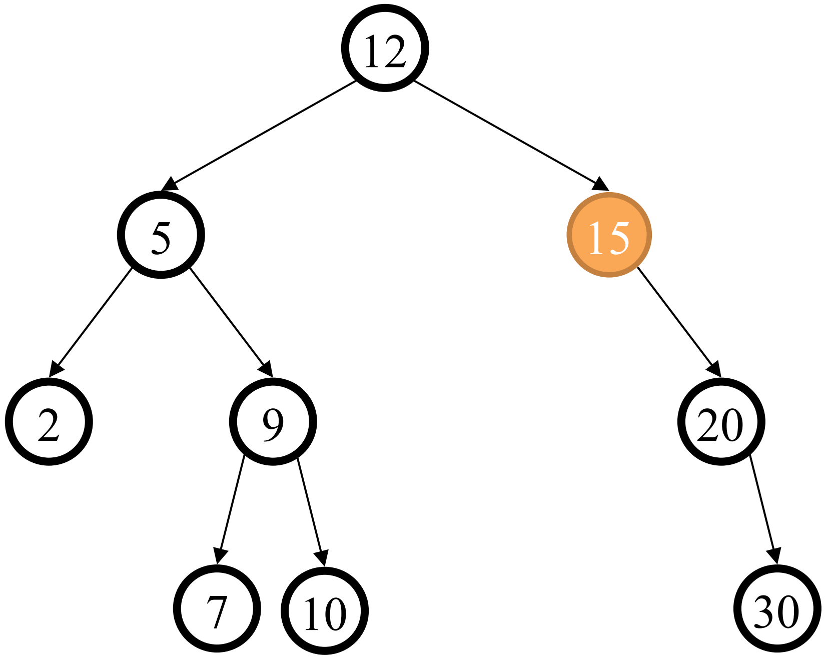
Deletion – The Leaf Case

`delete(17)`



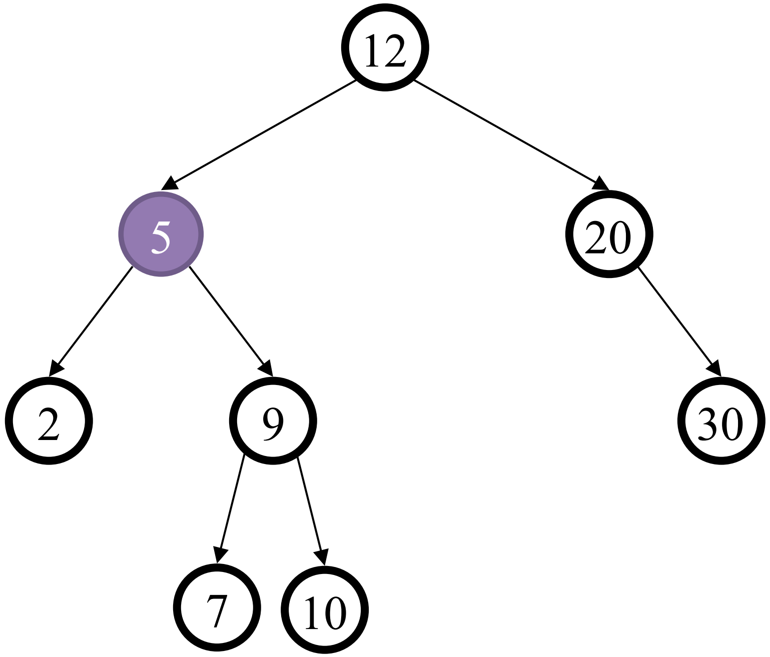
Deletion – The One Child Case

```
delete(15)
```



Deletion – The Two Child Case

`delete(5)`



What can we replace 5 with?

Deletion – The Two Child Case

Idea: Replace the deleted node with a value guaranteed to be between the two child subtrees

Options:

- *successor* from right subtree: **`findMin(node.right)`**
- *predecessor* from left subtree: **`findMax(node.left)`**
  - These are the easy cases of predecessor/successor

Now delete the original node containing *successor* or *predecessor*

- Leaf or one child case – easy cases of delete!

Lazy Deletion

- Lazy deletion can work well for a BST
  - Simpler
  - Can do “real deletions” later as a batch
  - Some inserts can just “undelete” a tree node

- But
  - Can waste space and slow down find operations
  - Make some operations more complicated:
    - How would you change **`findMin`** and **`findMax`**?

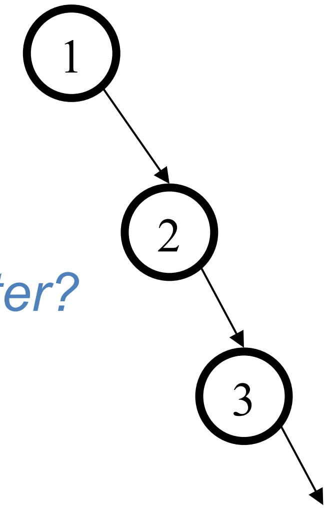
BuildTree for BST

- Let's consider **`buildTree`**
  - Insert all, starting from an empty tree

- Insert keys 1, 2, 3, 4, 5, 6, 7, 8, 9 into an empty BST
  - If inserted in given order, what is the tree?
  - What big-O runtime for this kind of sorted input?
  - Is inserting in the reverse order any better?

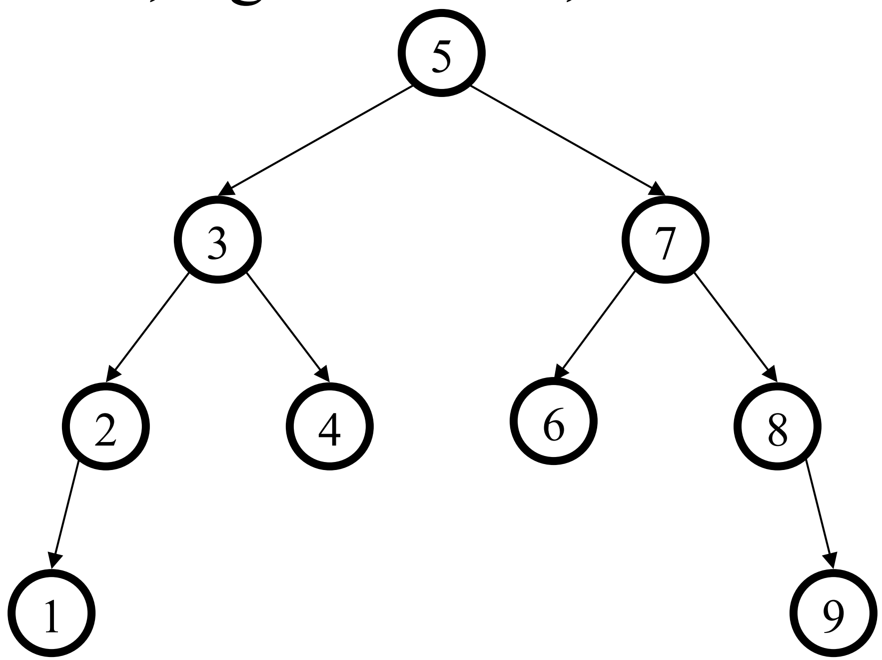
$O(n^2)$

Can we do better?



BuildTree for BST

- Insert keys 1, 2, 3, 4, 5, 6, 7, 8, 9 into an empty BST
- What we if could somehow re-arrange them
  - median first, then left median, right median, etc.
  - 5, 3, 7, 2, 1, 4, 8, 6, 9
  - What tree does that give us?
  - What big-O runtime?

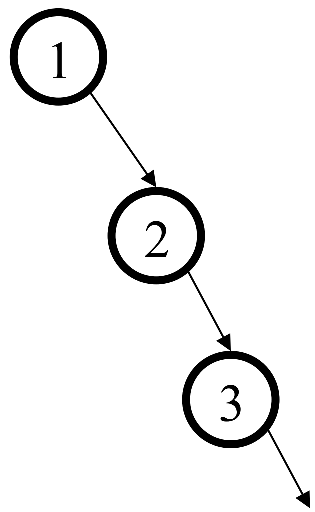


O(n log *n), awesome!*

Unbalanced BST

- Balancing a tree at build time is insufficient, as sequences of operations can eventually transform that carefully balanced tree into the dreaded list

- At that point, everything is *O(n)* and nobody is happy
  - **`find`**
  - **`insert`**
  - **`delete`**



Balanced BST

Observation

- BST: the shallower the better!
- For a BST with n nodes inserted in arbitrary order
  - Average height is $O(\texttt{log}\ n)$ – see text for proof
  - Worst case height is $O(n)$
- Simple cases, such as inserting in key order, lead to the worst-case scenario

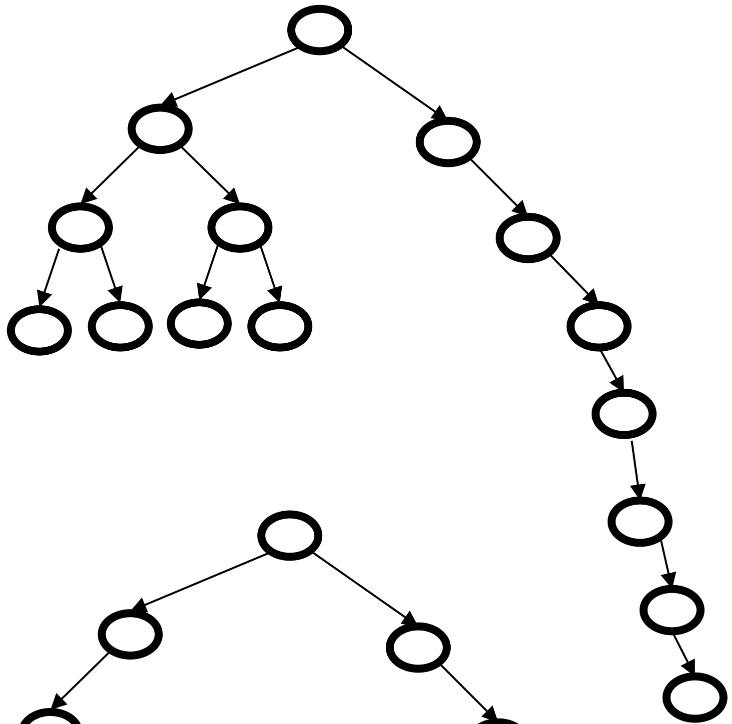
Solution: Require a **Balance Condition** that

1. Ensures depth is always $O(\texttt{log}\ n)$ – strong enough!
2. Is efficient to maintain – not too strong!

Potential Balance Conditions

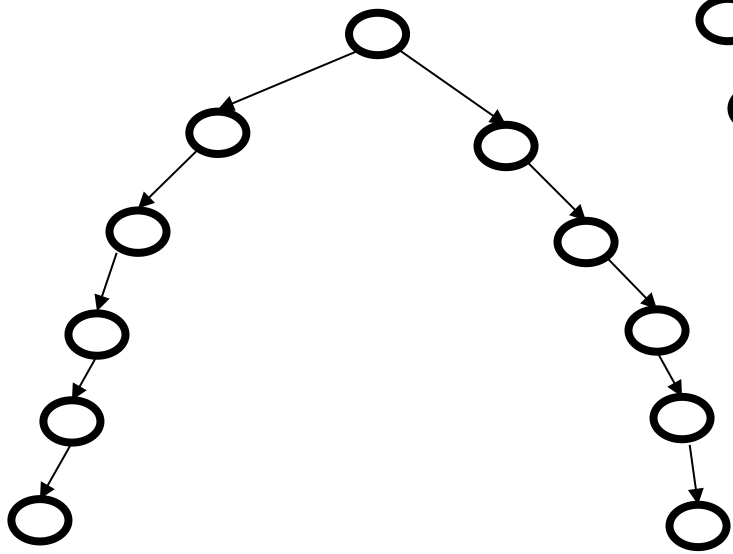
1. Left and right subtrees of the *root* have equal number of nodes

Too weak!
Height mismatch example:



2. Left and right subtrees of the *root* have equal *height*

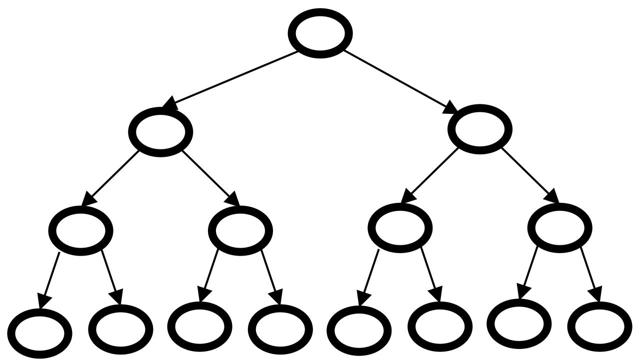
Too weak!
Double chain example:



Potential Balance Conditions

3. Left and right subtrees of every node have equal number of nodes

Too strong!
Only perfect trees (2^n – 1 nodes)



4. Left and right subtrees of every node have equal *height*

Too strong!
Only perfect trees (2^n – 1 nodes)

The AVL Balance Condition

Left and right subtrees of *every node* have *heights* **differing by at most 1**

Definition: **balance**(*node*) = height(*node*.left) – height(*node*.right)

AVL *property*: **for every node *x*, –1 ≤ balance(*x*) ≤ 1**

- Ensures small depth
  - Will prove this by showing that an AVL tree of height *h* must have a number of nodes *exponential* in *h*
- Efficient to maintain
  - Using single and double rotations

CSE373: Data Structures & Algorithms

Lecture 5: AVL Trees

Hunter Zahn

Summer 2016

Thanks to Kevin Quinn and Dan Grossman for slide materials

The AVL Tree Data Structure

Structural properties

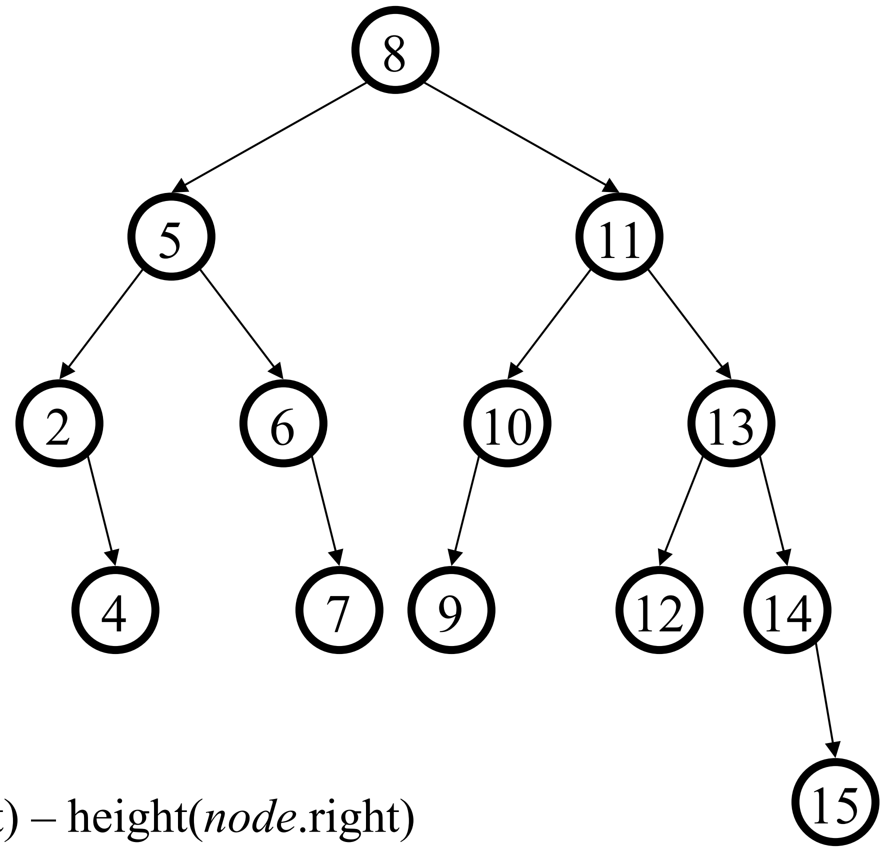
1. Binary tree property
2. Balance property: balance of every node is between -1 and 1

Result:

Worst-case depth is $O(\log n)$

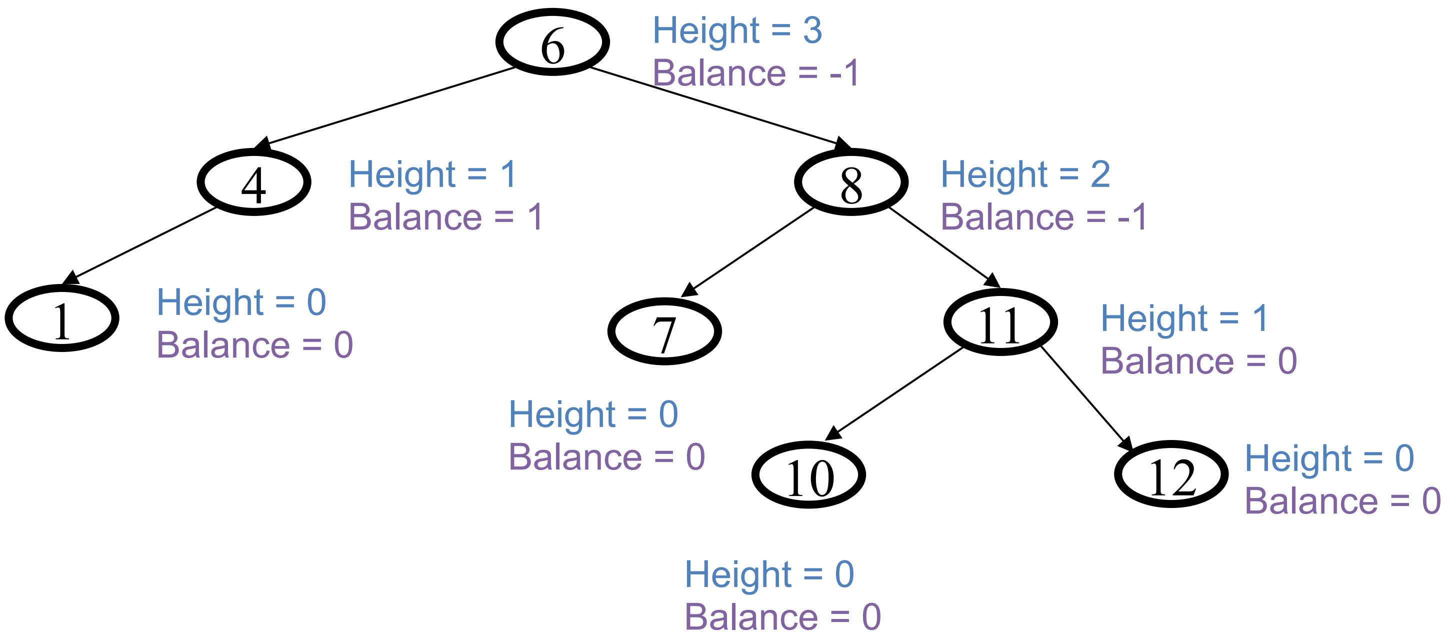
Ordering property

– Same as for BST

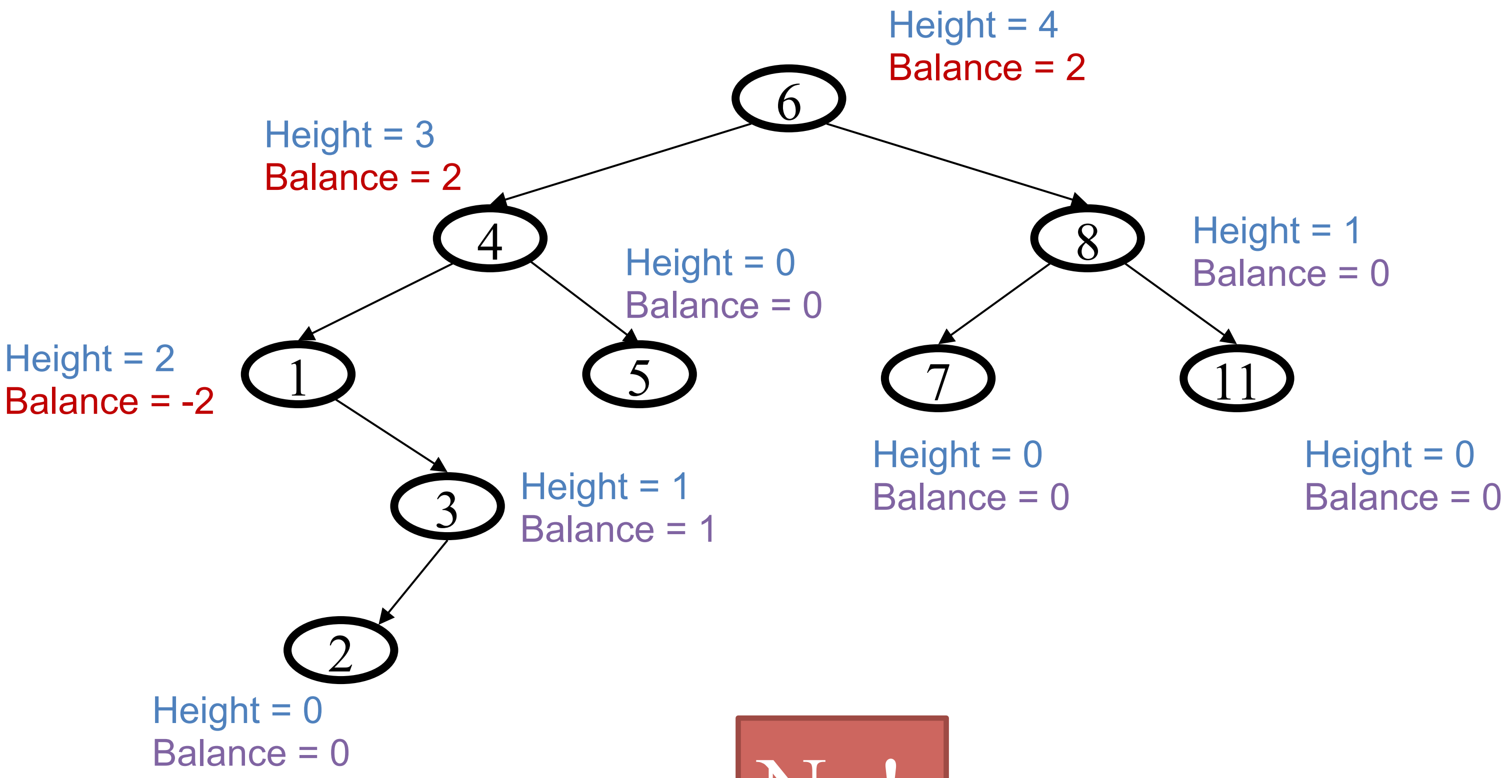


Definition: **balance**(*node*) = height(*node*.left) – height(*node*.right)

An AVL tree?


6
Height = 3
Balance = -1
4
Height = 1
Balance = 1
8
Height = 2
Balance = -1
1
Height = 0
Balance = 0
7
Height = 0
Balance = 0
11
Height = 1
Balance = 0
10
Height = 0
Balance = 0
12
Height = 0
Balance = 0

An AVL tree?



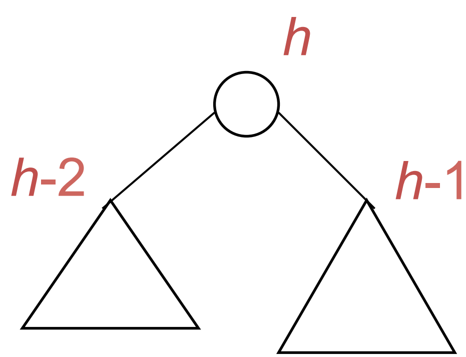
No!

The shallowness bound

Let $S(h)$ = the minimum number of nodes in an AVL tree of height h

– If we can prove that $S(h)$ grows exponentially in h, then a tree with n nodes has a logarithmic height

- Step 1: Define $S(h)$ inductively using AVL property
  – $S(-1)=0$, $S(0)=1$, $S(1)=2$
  – *For* $h \geq 1$, $S(h) = 1+S(h-1)+S(h-2)$



- Step 2: Bound $S(h)$
  – Using everybody's favorite: **Induction!**

Fibonacci Numbers

- Sequence of numbers where each number is the sum of the preceding two numbers:
  - 0, 1, 1, 2, 3, 5, 8, …

  $F(0) = 0$

  $F(1) = 1$

  $F(n+1) = F(n-1) + F(n)$

  Grows exponentially

S(h) and F(h)

h	0	1	2	3	4	5	6	7	8	9
S(h)	1	2	4	7	12	20	33	54	88	143
F(h)	0	1	1	2	3	5	8	13	21	34

h	0	1	2	3	4	5	6	7	8	9
S(h)	1	2	4	7	12	20	33	54	88	143
F(h+3)	2	3	5	8	13	21	34	55	89	144

$S(h)$ appears to be equal to $F(h+3) - 1$

The proof

Remember:
1) $S(-1)=0$, $S(0)=1$, $S(1)=2$
2) For $h \geq 1$, $S(h) = 1+S(h-1)+S(h-2)$

$F(h + 1) = F(h - 1) + F(h)$

Let $P(h)$ be $S(h) == F(h+3)-1$. *We will prove this for all* $h >= 0$

Base cases:

$S(0) = 1 \qquad F(0 + 3) – 1 = 2 – 1 = 1$
$S(1) = 2 \qquad F(1 + 3) – 1 = 3 – 1 = 2$

Inductive Hypothesis:

Assume $P(k)$ for an arbitrary $k > 1$.

$P(k)$: $S(k) == F(k+3)-1$

Inductive Step:

$$
\begin{aligned}
S(k+1) &= S(k-1) + S(k) + 1 \\
&= [F((k-1) + 3) – 1] + [F(k+3) – 1] + 1 && \textit{I.H.} \\
&= F((k-1) + 3) + F(k + 3) – 1 && \textit{simplify} \\
&= F((k+1) + 3) – 1 && \textit{def of F}
\end{aligned}
$$

$P(k) \rightarrow P(k+1)$

Conclusion:

We have proven by induction that $P(h)$ *holds for all* $h >= 0$.

O(log n)!!

The minimum number of nodes in an AVL tree grows exponentially with respect to the height!
Therefore, the height grows logarithmically w.r.t. the number of nodes in an AVL tree!

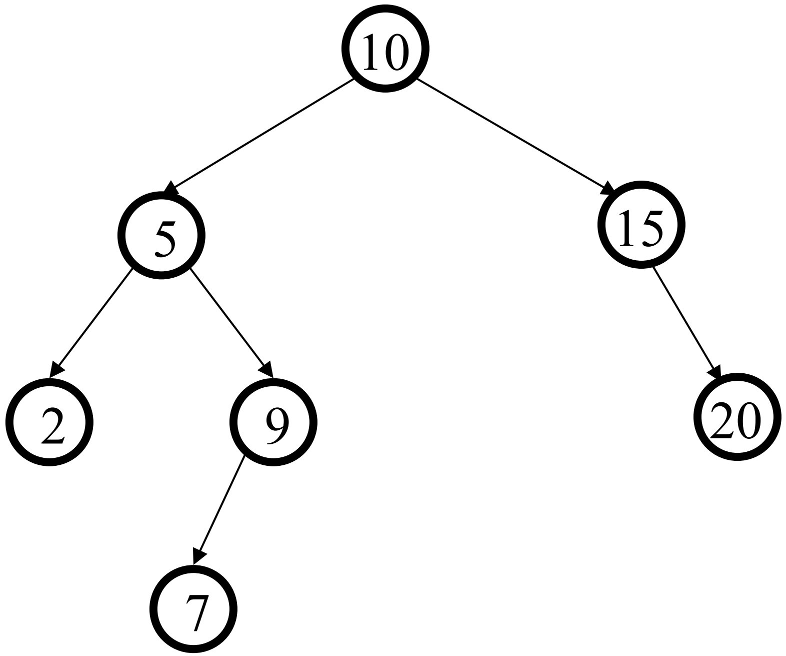
Good news

Proof means that if we have an AVL tree, then **`find`** is $O($**`log`** $n)$

– Recall logarithms of different bases > 1 differ by only a constant factor

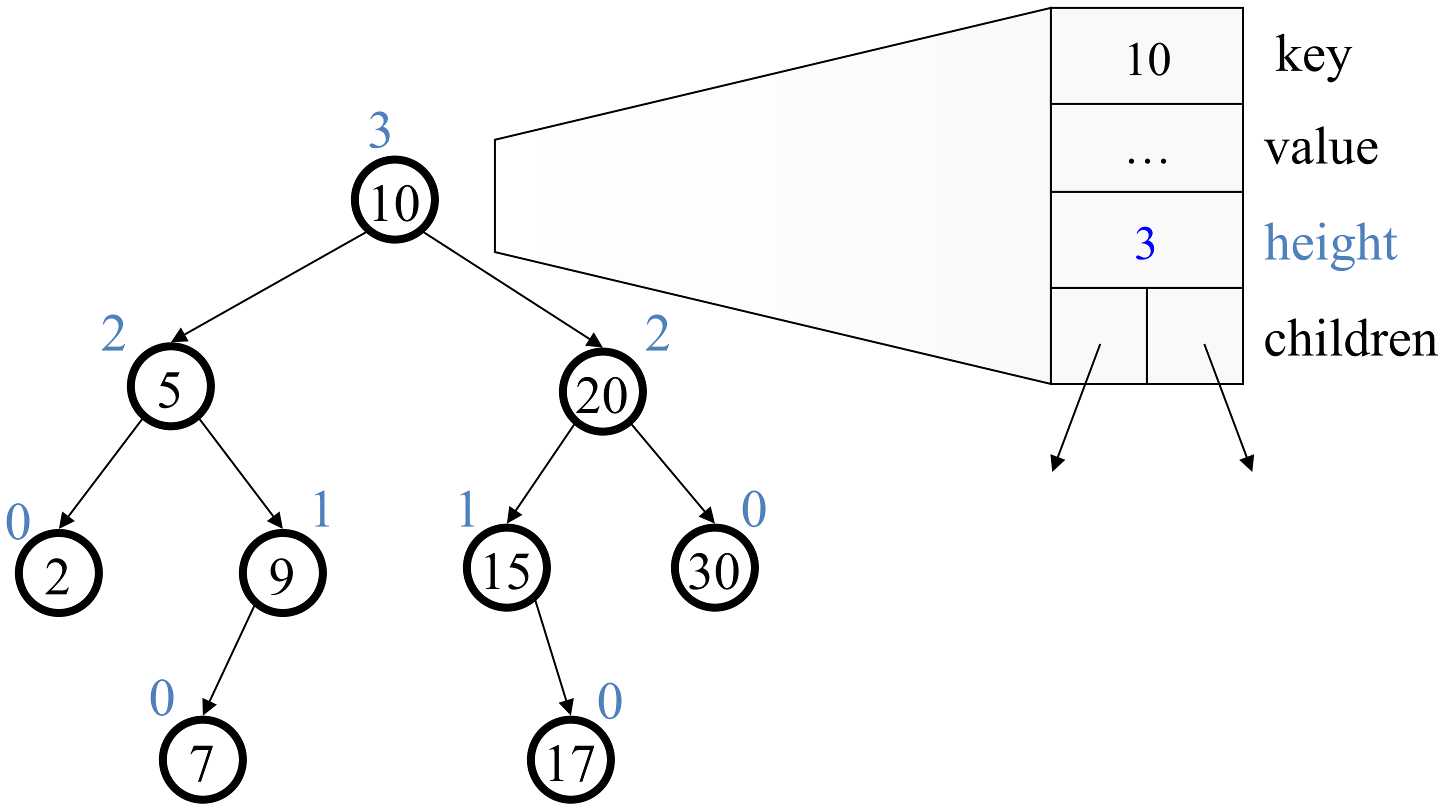
But as we insert and delete elements, we need to:

1. **Track balance**
2. **Detect imbalance**
3. **Restore balance**



Is this AVL tree balanced?
How about after `insert(30)`?

An AVL Tree



Track height at all times!

AVL tree operations

- **AVL `find`**:
  - Same as BST **`find`**
- **AVL `insert`**:
  - First BST **`insert`**, *then* check balance and potentially "fix" the AVL tree
  - Four different imbalance cases
- **AVL `delete`**:
  - The "easy way" is lazy deletion
  - Otherwise, do the deletion and then have several imbalance cases (we will likely skip this but post slides for those interested)

Insert: detect potential imbalance

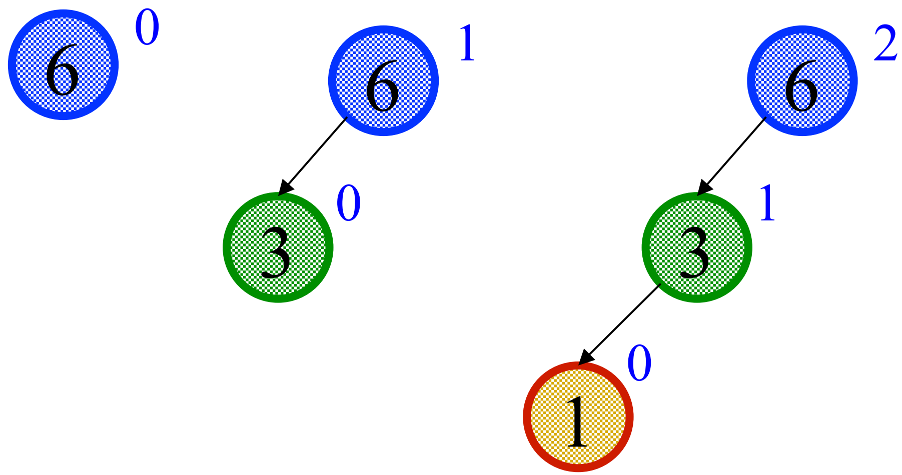
1. Insert the new node as in a BST (a new leaf)
2. For each node on the path from the root to the new leaf, the insertion may (or may not) have changed the node's height
3. So after recursive insertion in a subtree, detect height imbalance and perform a ***rotation*** to restore balance at that node

Type of rotation will depend on the location of the imbalance (if any)

Facts that an implementation can ignore:

- There must be a deepest element that is imbalanced after the insert (all descendants still balanced)
- After rebalancing this deepest node, every node is balanced
- So at most one node needs to be rebalanced

Case #1: Example



Insert(6)

Insert(3)

Insert(1)

Third insertion violates balance property

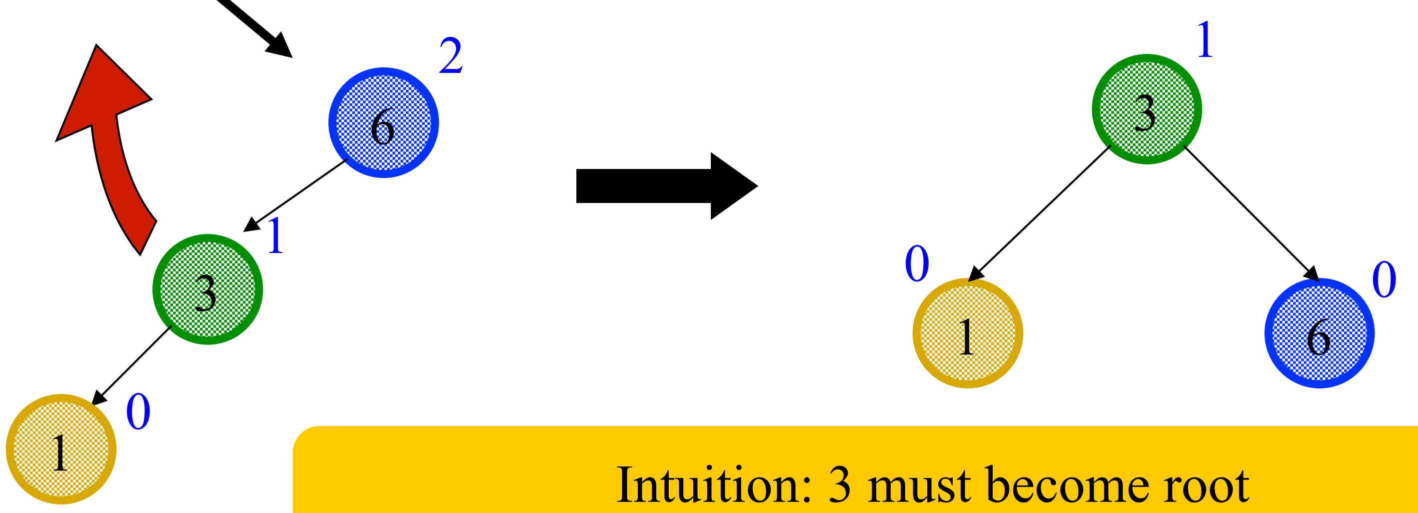
- happens to be at the root

What is the only way to fix this?

Fix: Apply "Single Rotation"

- *Single rotation:* The basic operation we'll use to rebalance
  - Move child of unbalanced node into parent position
  - Parent becomes the "other" child (always okay in a BST!)
  - Other subtrees move in only way BST allows (next slide)

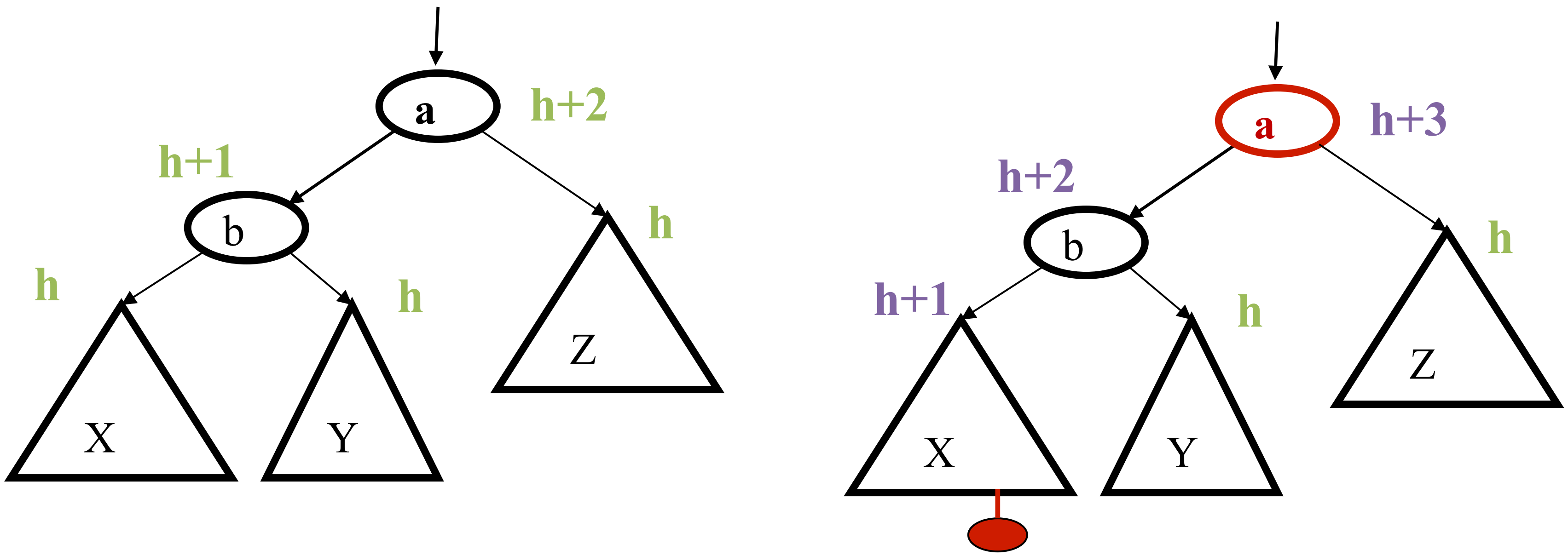
AVL Property violated here



Intuition: 3 must become root
New parent height is now the old parent's height before insert

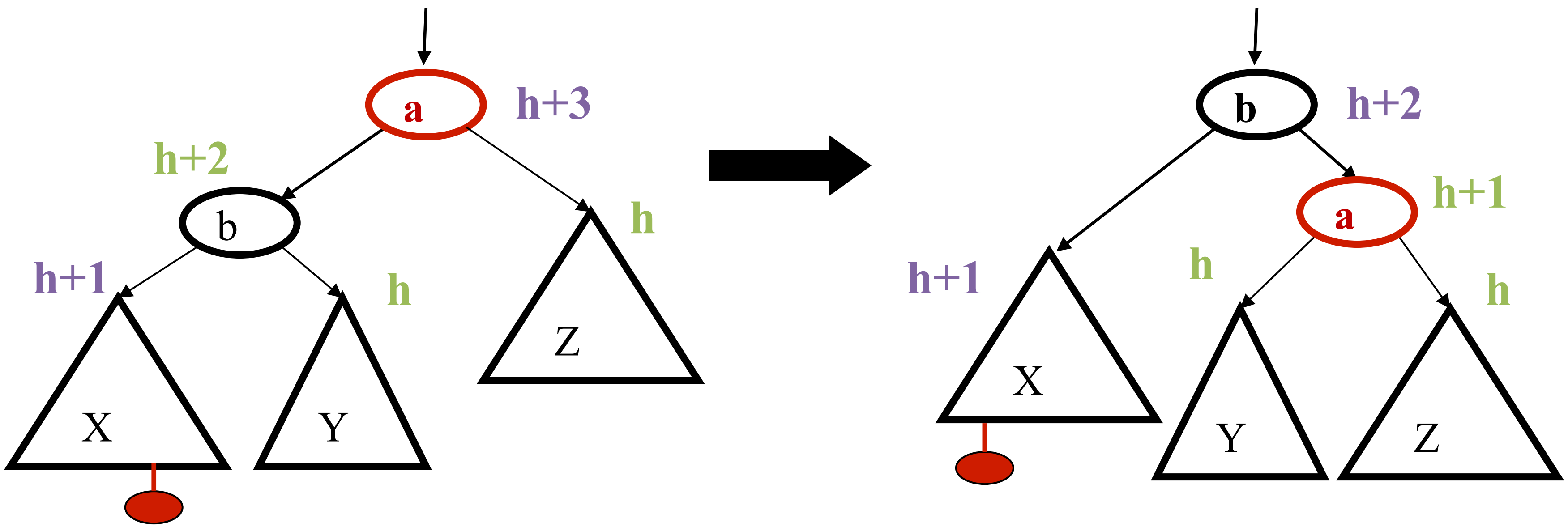
The example generalized

- Node imbalanced due to insertion *somewhere* in **left-left grandchild** that causes an increasing height
  - 1 of 4 possible imbalance causes (other three coming)
- First we did the insertion, which would make ***a*** imbalanced



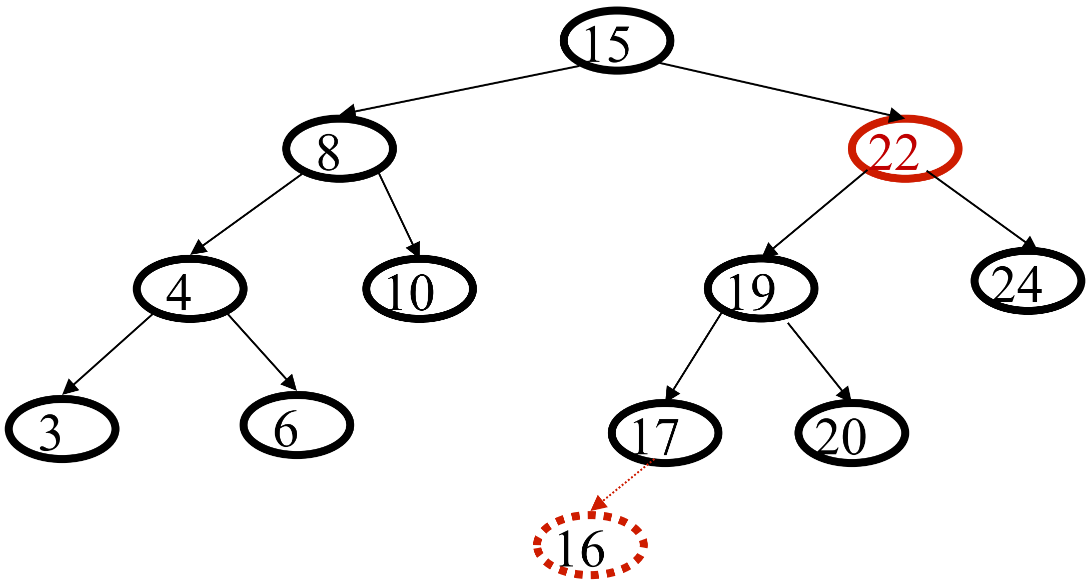
The general left-left case

- Node imbalanced due to insertion *somewhere* in **left-left grandchild**
  - 1 of 4 possible imbalance causes (other three coming)
- So we rotate at `a`, using BST facts: X < b < Y < a < Z

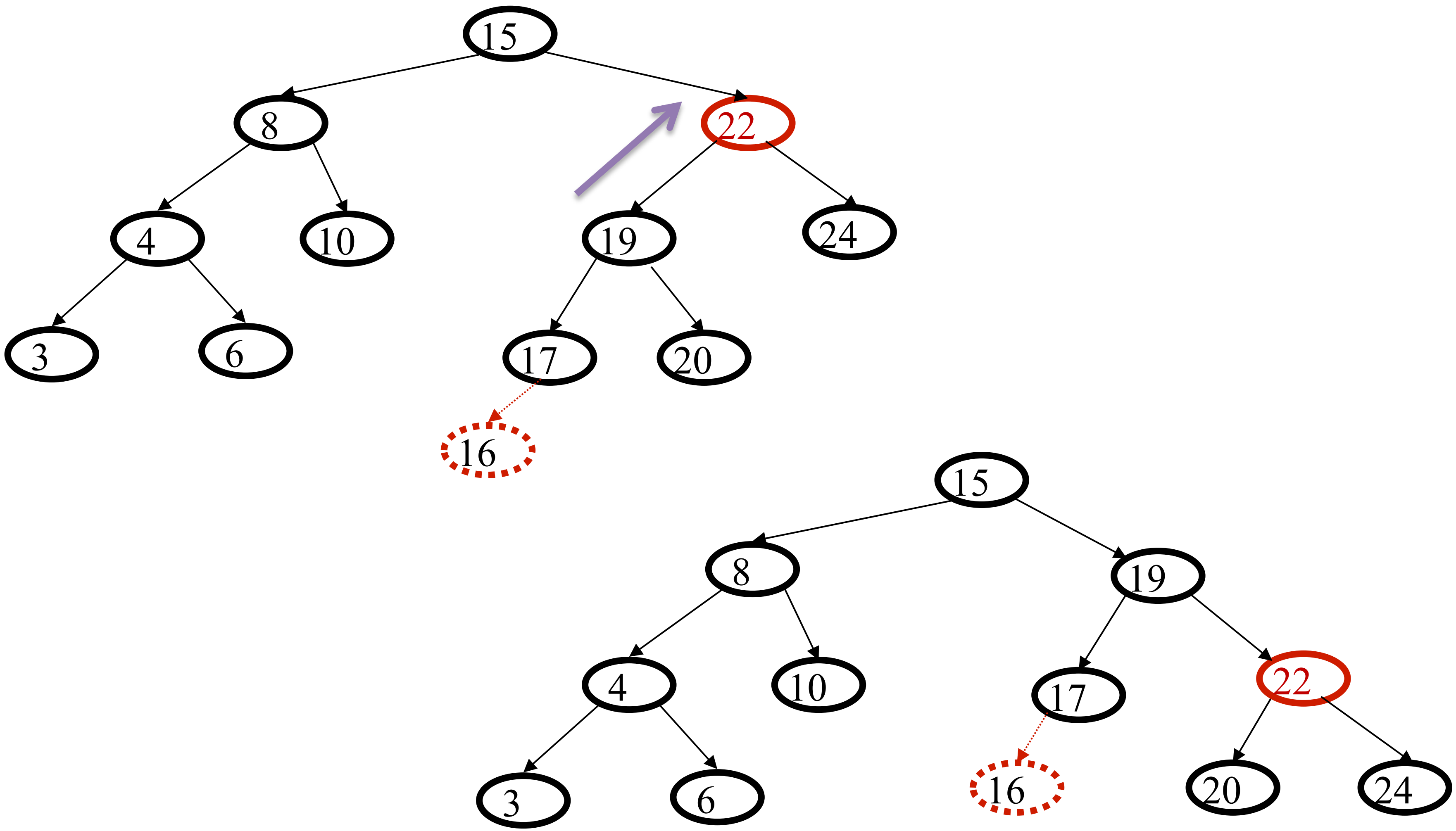


- A single rotation restores balance at the node
  - To same height as before insertion, so ancestors now balanced

Another example: `insert(16)`

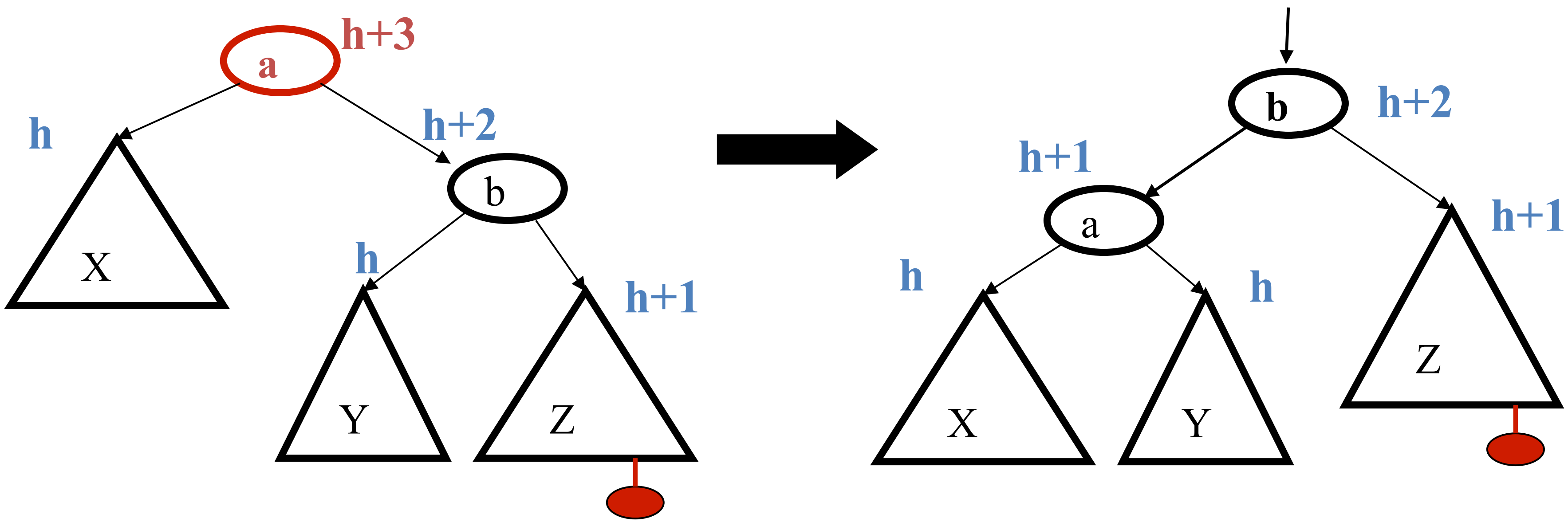


Another example: `insert(16)`



The general right-right case

- Mirror image to left-left case, so you rotate the other way
  - Exact same concept, but need different code

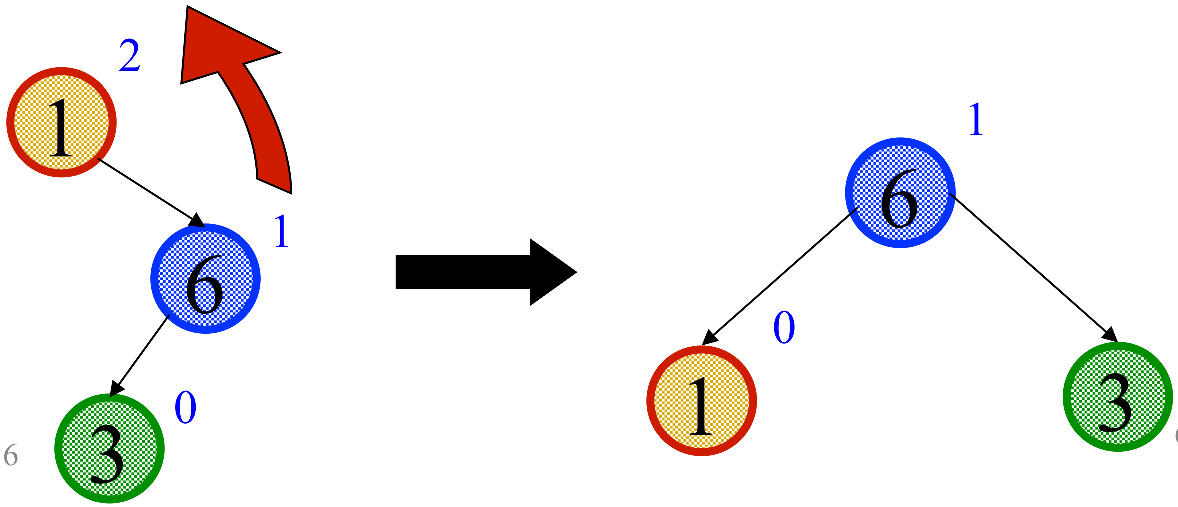


Two cases to go

Unfortunately, single rotations are not enough for insertions in the **left-right** subtree or the **right-left** subtree

Simple example: **`insert`**(1), **`insert`**(6), **`insert`**(3)

- First wrong idea: single rotation like we did for left-left

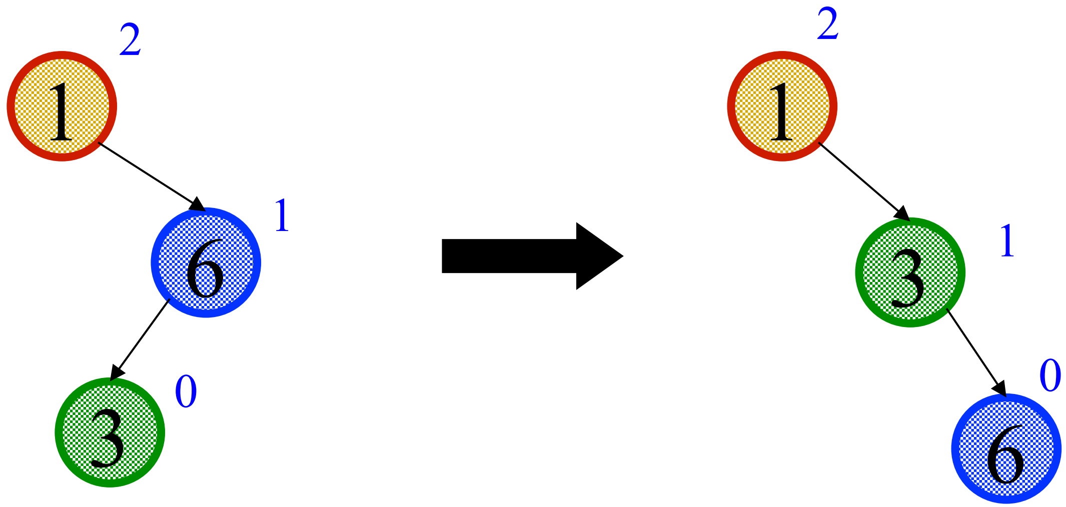


Two cases to go

Unfortunately, single rotations are not enough for insertions in the left-right subtree or the right-left subtree

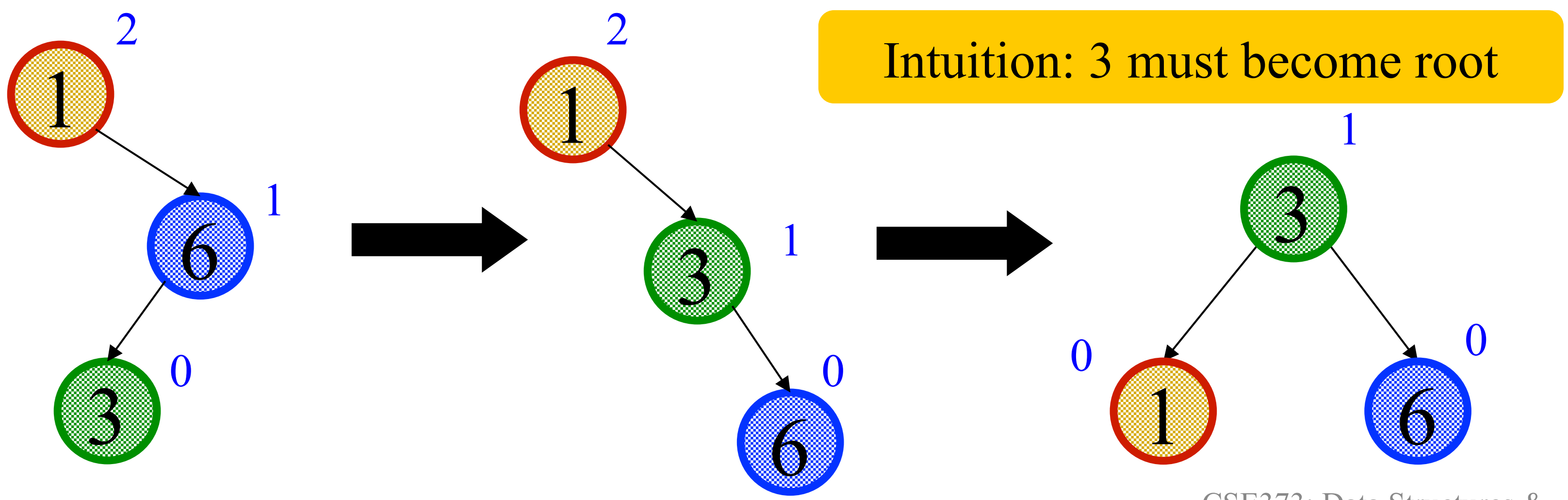
Simple example: **`insert`**(1), **`insert`**(6), **`insert`**(3)

– Second wrong idea: single rotation on the child of the unbalanced node

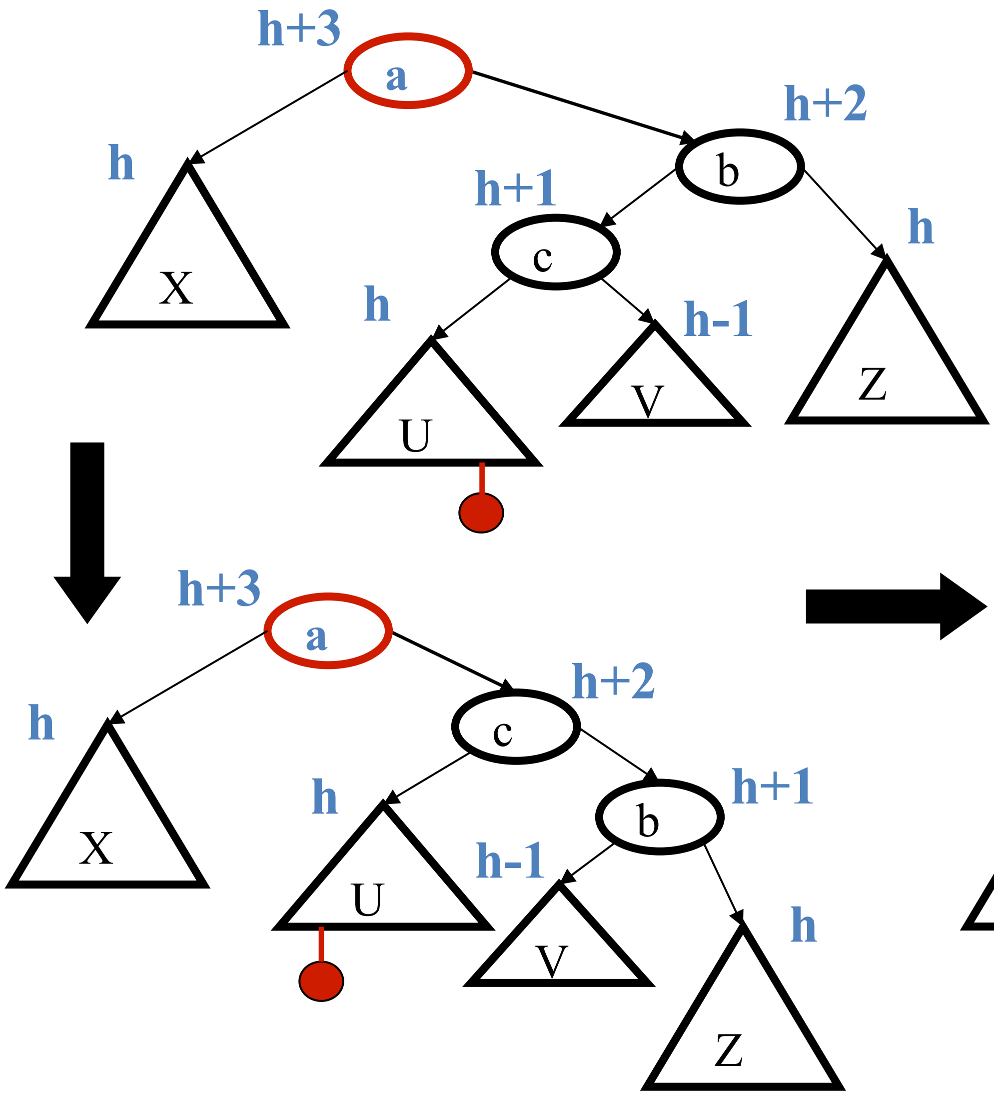


Sometimes two wrongs make a right

- First idea violated the BST property
- Second idea didn't fix balance
- But if we do both single rotations, starting with the second, it works! (And not just for this example.)
- Double rotation:
  1. Rotate problematic child and grandchild
  2. Then rotate between self and new child



The general right-left case

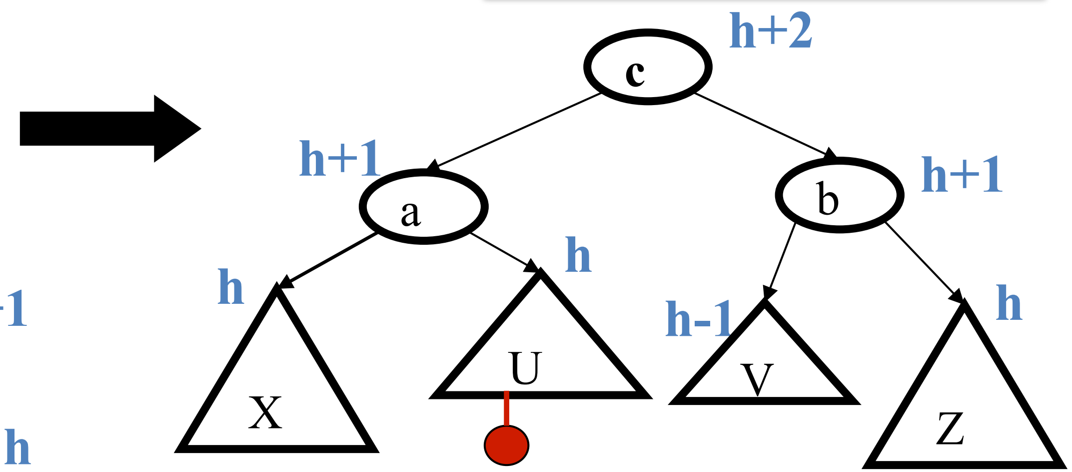


Rotation 1:

```
b.left = c.right
c.right = b
a.right = c
```

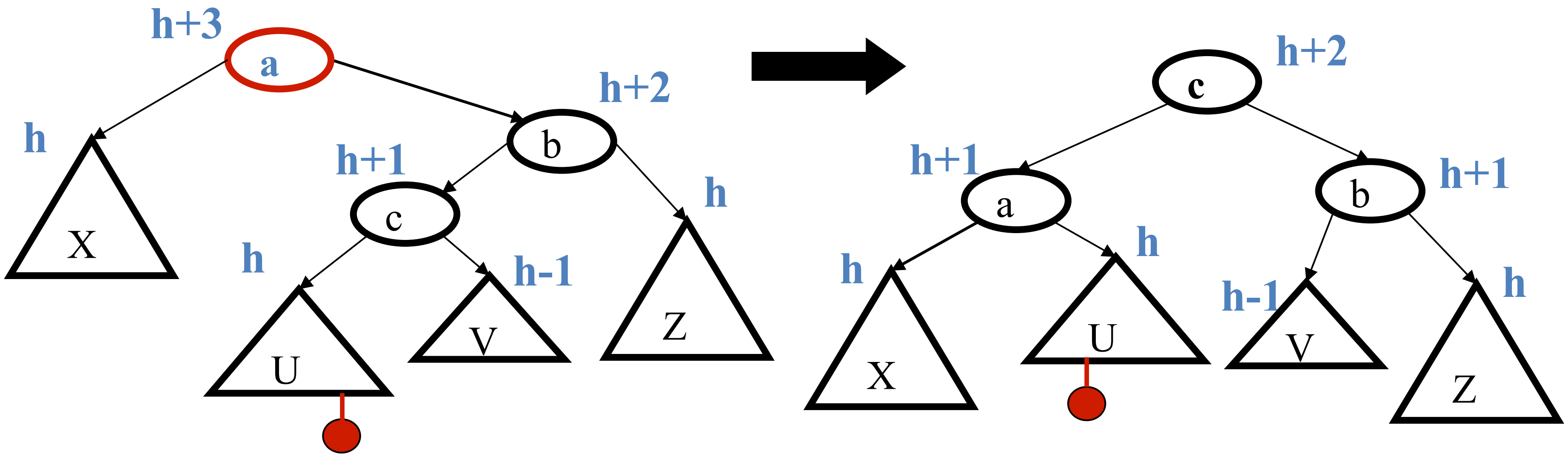
Rotation 2:

```
a.right = c.left
c.left = a
root = c
```



Comments

- Like in the left-left and right-right cases, the height of the subtree after rebalancing is the same as before the insert
  - So no ancestor in the tree will need rebalancing
- Does not have to be implemented as two rotations; can just do:



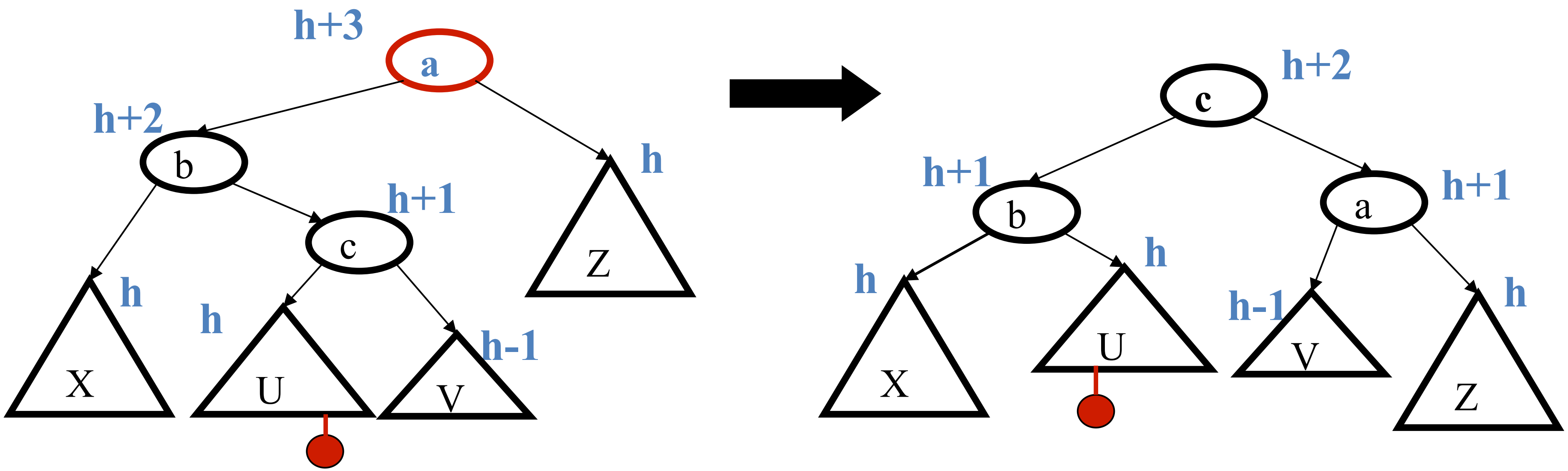
Easier to remember than you may think:

1) Move c to grandparent's position

2) Put a, b, X, U, V, and Z in the only legal positions for a BST

The last case: left-right

- Mirror image of right-left
  - Again, no new concepts, only new code to write



Insert, summarized

- Insert as in a BST
- Check back up path for imbalance, which will be 1 of 4 cases:
  – Node's left-left grandchild is too tall (**left-left single rotation**)
  – Node's left-right grandchild is too tall (**left-right double rotation**)
  – Node's right-left grandchild is too tall (**right-left double rotation)**
  – Node's right-right grandchild is too tall (**right-right double rotation**)
- Only one case occurs because tree was balanced before insert
- After the appropriate single or double rotation, the smallest-unbalanced subtree has the same height as before the insertion
  – So all ancestors are now balanced

Now efficiency

- Worst-case complexity of **`find`**: $O(\texttt{log}\ n)$
  - Tree is balanced
- Worst-case complexity of **`insert`**: $O(\texttt{log}\ n)$
  - Tree starts balanced
  - A rotation is $O(1)$ and there's an $O(\texttt{log}\ n)$ path to root
  - (Same complexity even without one-rotation-is-enough fact)
  - Tree ends balanced
- Worst-case complexity of **`buildTree`**: $O(n\ \texttt{log}\ n)$

Takes some more rotation action to handle **`delete`**…

Pros and Cons of AVL Trees

Arguments for AVL trees:

1. All operations logarithmic worst-case because trees are *always* balanced
2. Height balancing adds no more than a constant factor to the speed of `insert` and `delete`

Arguments against AVL trees:

1. Difficult to program & debug [but done once in a library!]
2. More space for height field
3. Asymptotically faster but rebalancing takes a little time
4. Most large searches are done in database-like systems on disk and use other structures (e.g., *B*-trees, a data structure in the text)
5. If *amortized* (later, I promise) logarithmic time is enough, use splay trees (also in text)