https://algs4.cs.princeton.edu

5.2 Tries

- *string symbol tables*
- *R-way tries*
- *ternary search tries*
- *character-based operations*

https://algs4.cs.princeton.edu

5.2 TRIES

- *string symbol tables*
- *R-way tries*
- *ternary search tries*
- *character-based operations*

Summary of the performance of symbol-table implementations

Order of growth of the frequency of operations.

implementation	typical case			ordered operations	operations on keys
	search	insert	delete		
red–black BST	$\log n$	$\log n$	$\log n$	✔	`compareTo()`
hash table	1 †	1 †	1 †		`equals()` `hashCode()`

† under uniform hashing assumption

Q. Can we do better?

A. Yes, if we can avoid examining the entire key, as with string sorting.

String symbol table implementations cost summary

Challenge. Efficient performance for string keys.

exchange rate: around $\Theta(\log n)$ character compares per string compare

	character accesses (typical case)				dedup	
implementation	**search hit**	**search miss**	**insert**	**space (references)**	**moby.txt**	**actors.txt**
red-black BST	$L + \log^2 n$	$\log^2 n$	$\log^2 n$	$4n$	1.4	97.4
hashing (linear probing)	L	L	L	$4n$ to $16n$	0.76	40.6

n = number of string

L = length of string

R = radix

file	size	words	distinct
moby.txt	1.2 MB	210 K	32 K
actors.txt	82 MB	11.4 M	900 K

String symbol table basic API

String symbol table. Symbol table specialized to string keys.

	`public class StringST<Value>`	
	`StringST()`	*create an empty symbol table*
`void`	`put(String key, Value val)`	*put key–value pair into the symbol table*
`Value`	`get(String key)`	*return value paired with given key*
`void`	`delete(String key)`	*delete key and corresponding value*
	⋮	⋮

Goal. Faster than hashing, more flexible than BSTs.

Algorithms

Robert Sedgewick | Kevin Wayne

https://algs4.cs.princeton.edu

5.2 Tries

- *string symbol tables*
- ***R-way tries***
- *ternary search tries*
- *character-based operations*

Tries

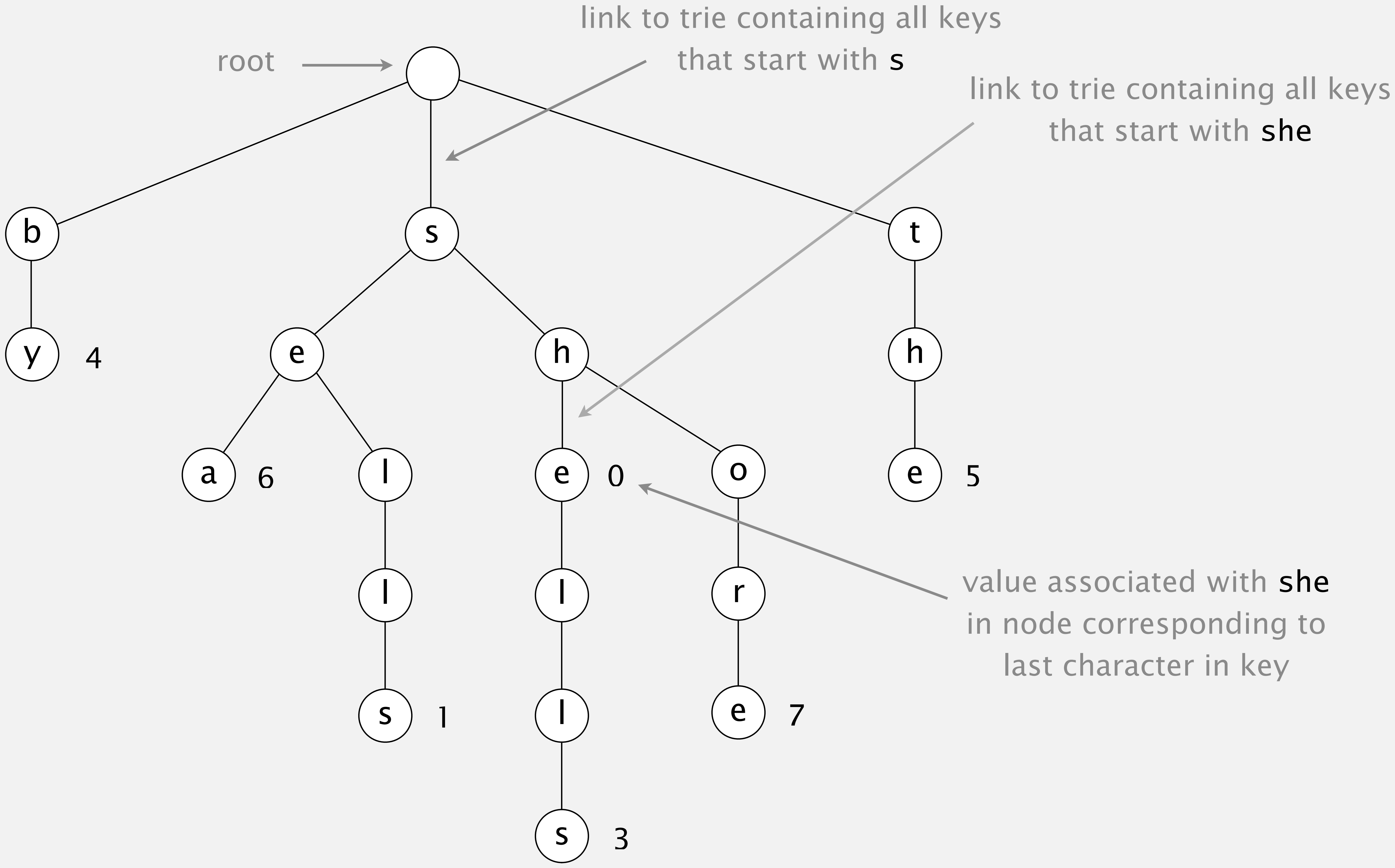
Tries

Tries. [from re**trie**val, but pronounced "try"]

- Store characters in nodes (not keys).
- Each node has R children, one for each possible character.

  (for now, we do not draw null links)

key	value
by	4
sea	6
sells	1
she	0
shells	3
shore	7
the	5

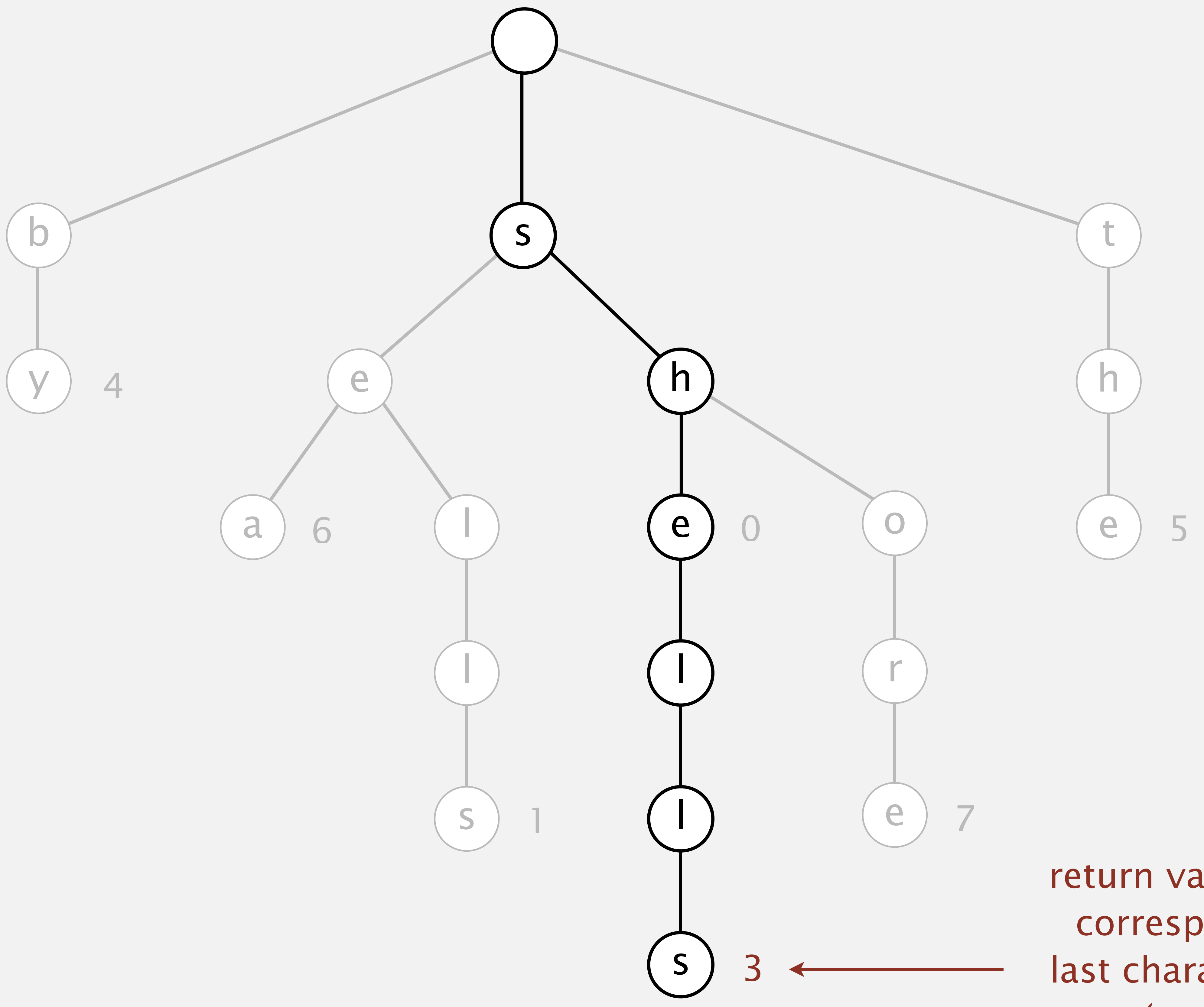


Search in a trie

Follow links corresponding to each character in the key.

- Search hit: node where search ends has a non-null value.
- Search miss: reach null link or node where search ends has null value.

get("shells")



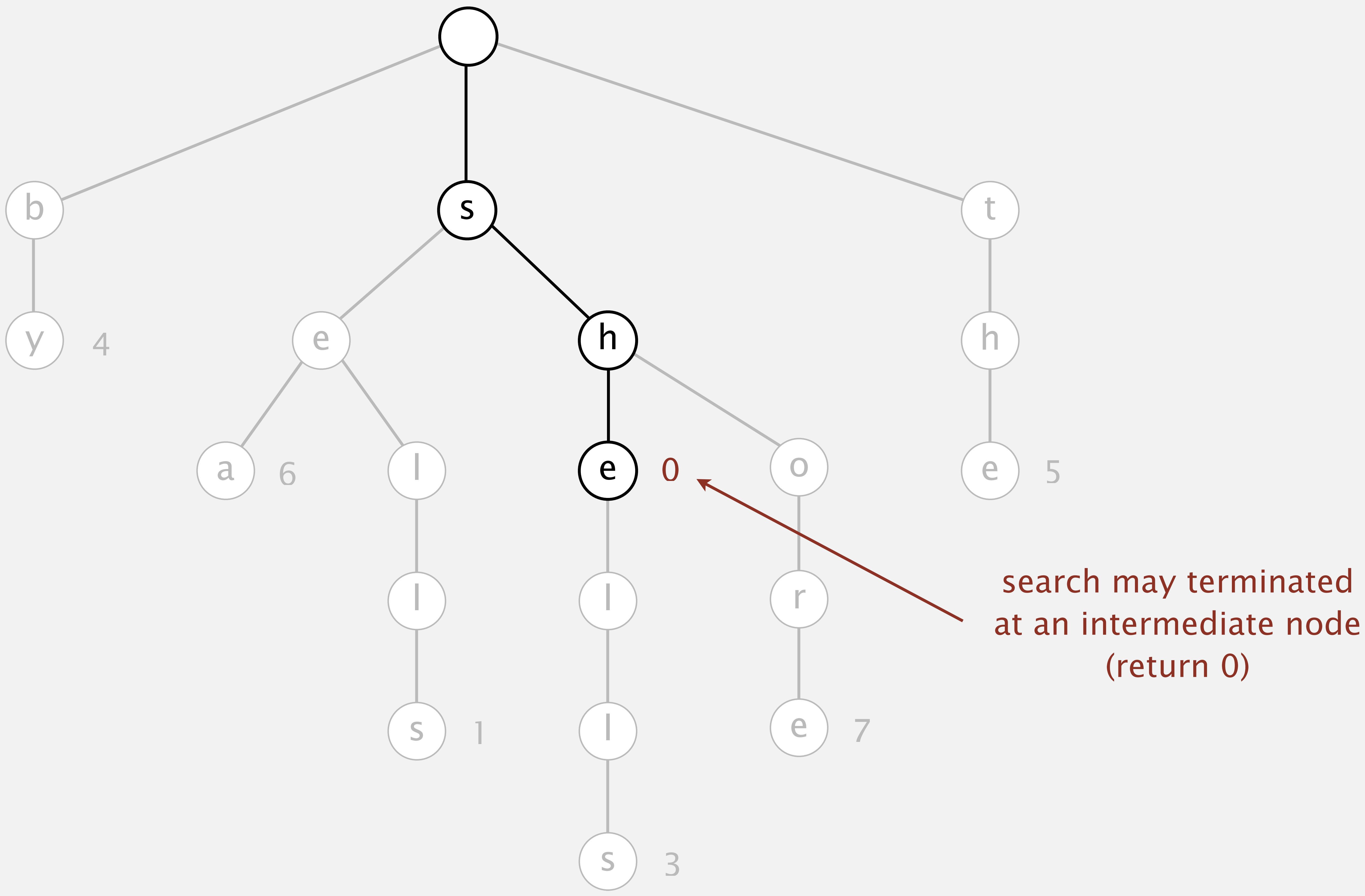
return value in node corresponding to last character in key (return 3)

Search in a trie

Follow links corresponding to each character in the key.

- Search hit: node where search ends has a non-null value.
- Search miss: reach null link or node where search ends has null value.

get("she")

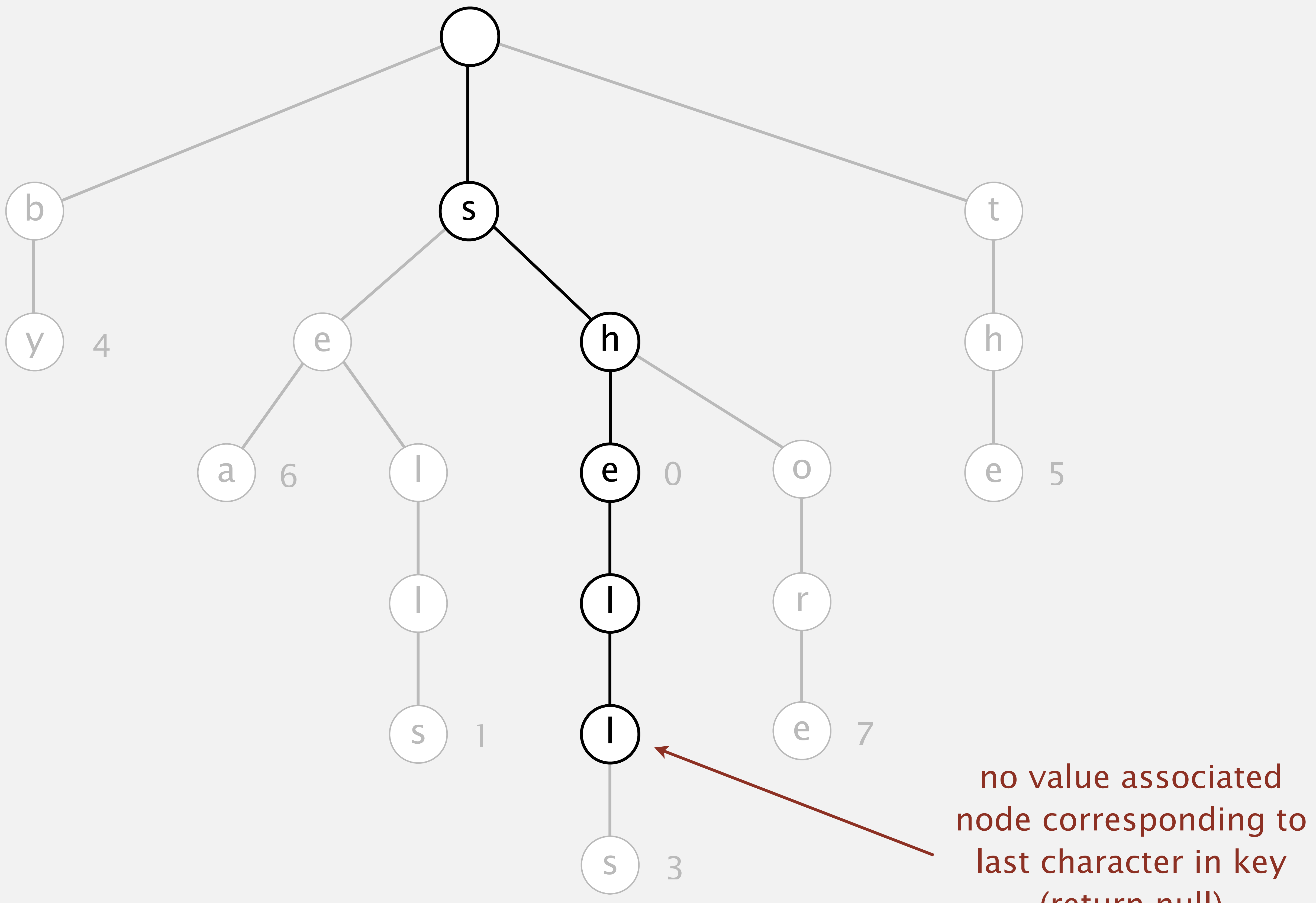


Search in a trie

Follow links corresponding to each character in the key.

- Search hit: node where search ends has a non-null value.
- Search miss: reach null link or node where search ends has null value.

get("shell")

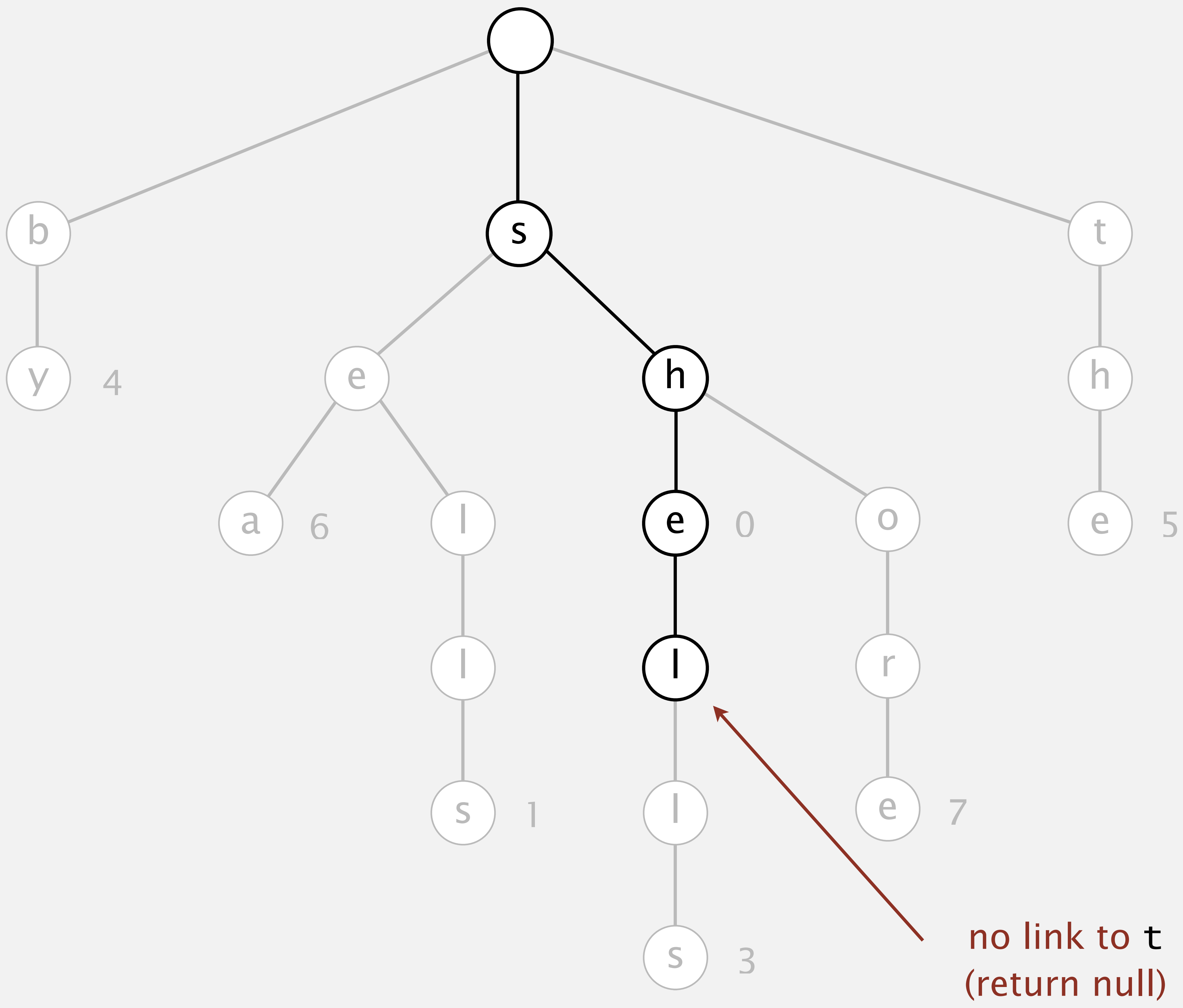


Search in a trie

Follow links corresponding to each character in the key.

- Search hit: node where search ends has a non-null value.
- Search miss: reach null link or node where search ends has null value.

get("shelter")

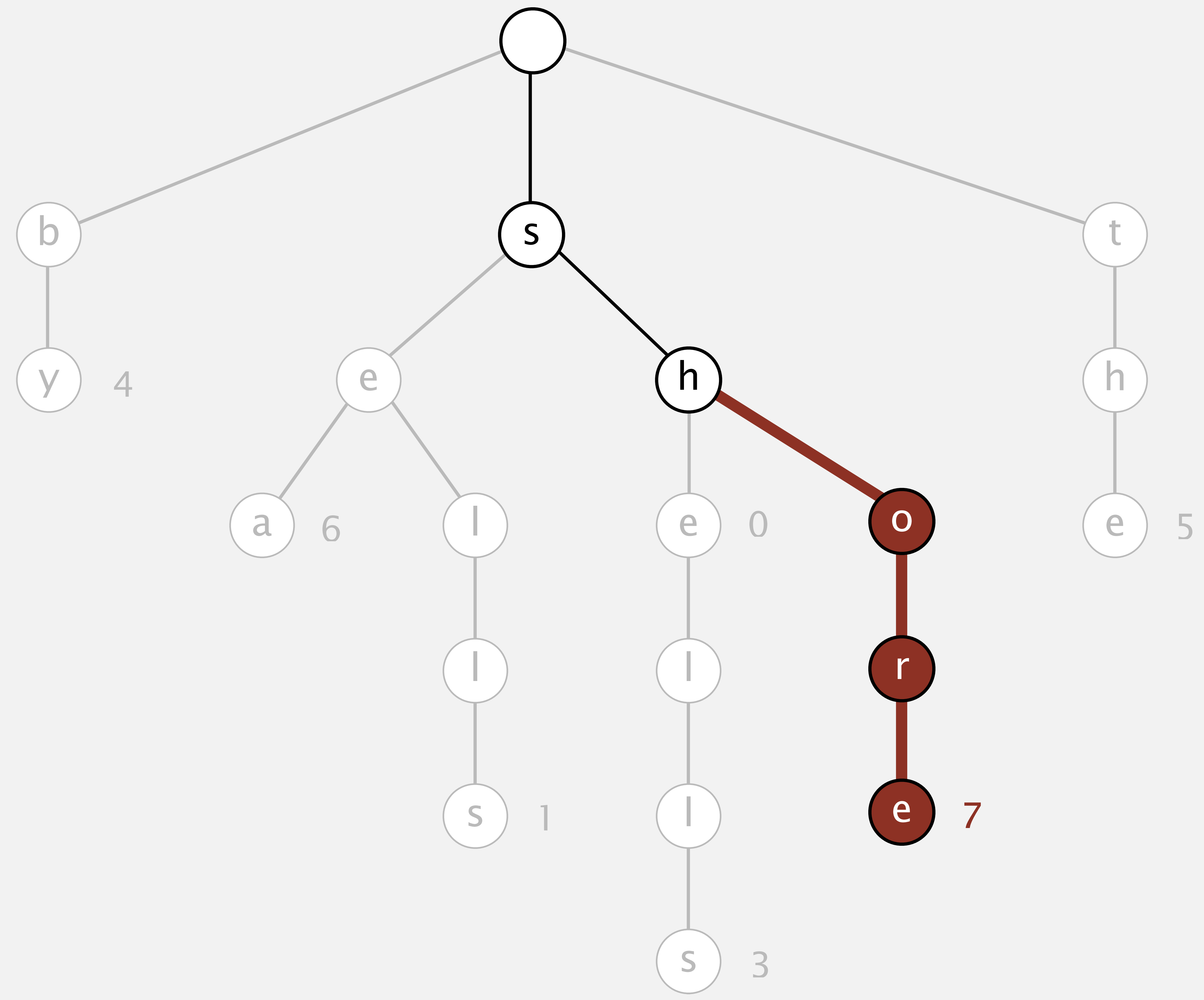


Insertion into a trie

Follow links corresponding to each character in the key.

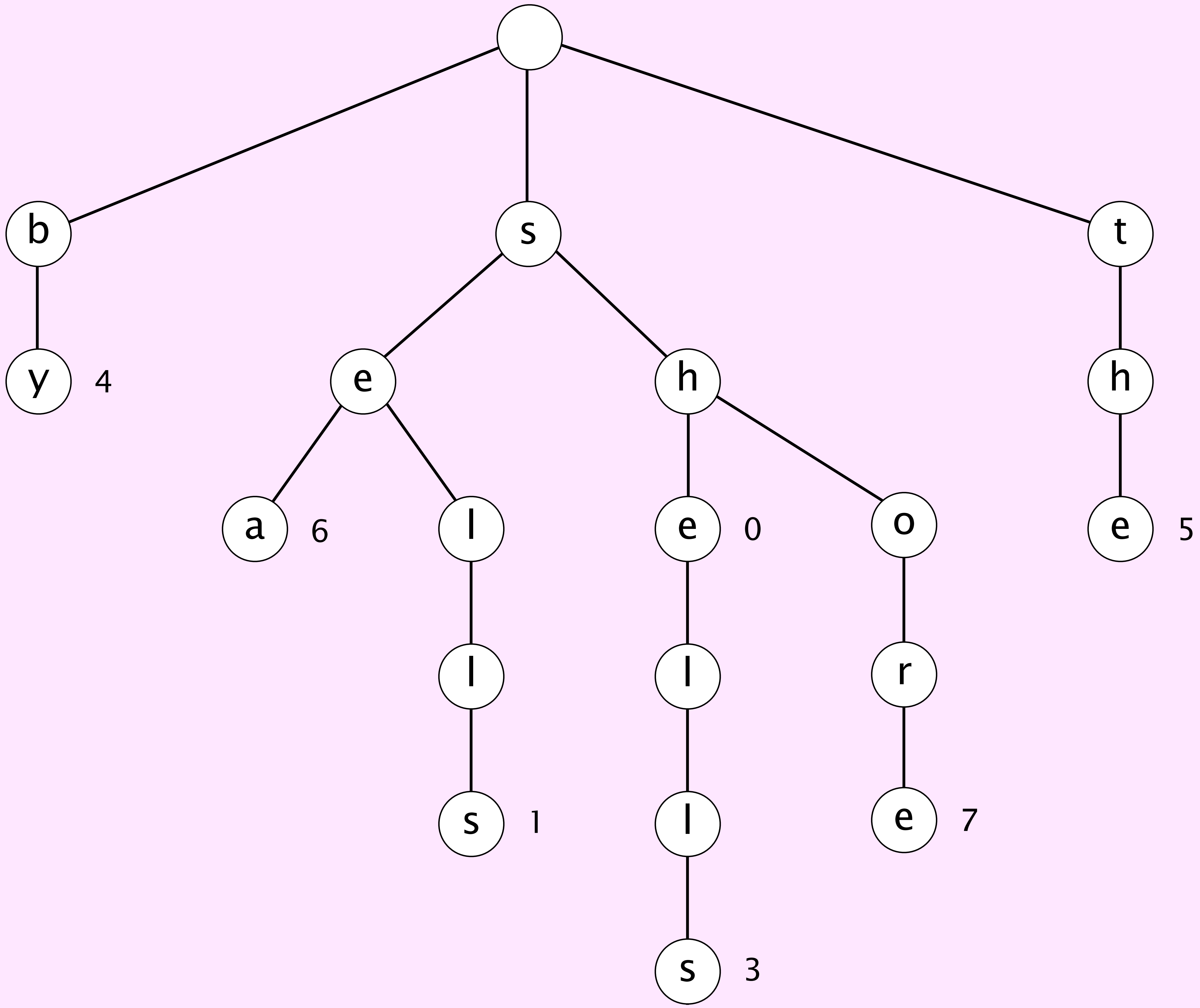
- Encounter a null link: create new node.
- Encounter the last character of the key: set value in that node.

put("shore", 7)



Trie construction demo

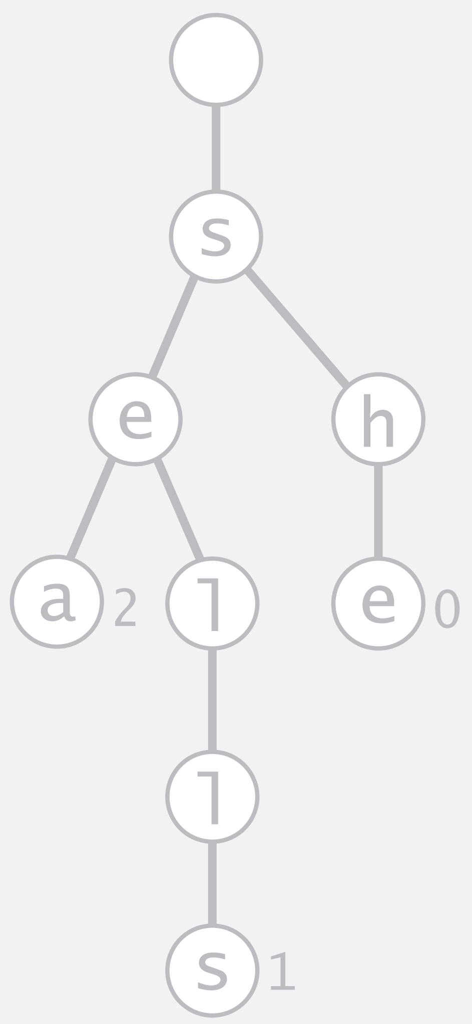
trie

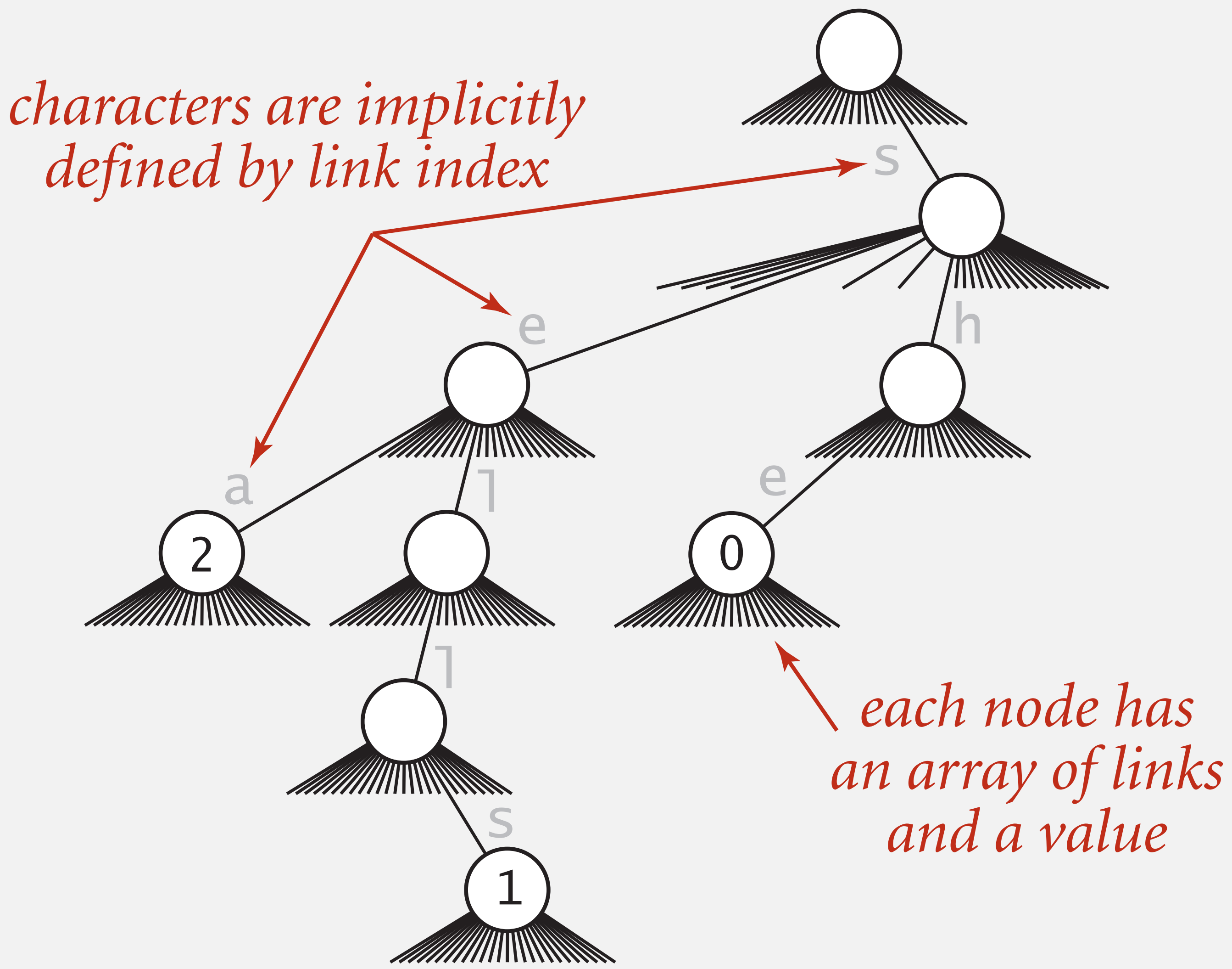


R-way trie representation: Java implementation

Node. A value, plus references to R nodes.

```
private static class Node
{
   private Object val;   ←——  no generic array creation
   private Node[] next = new Node[R];
}
```





Remark. An R-way trie stores neither keys nor characters explicitly.

R-way trie: Java implementation

```
public class TrieST<Value>
{
   private static final int R = 256;  ←—— extended ASCII
   private Node root = new Node();

   private static class Node
   {  /* see previous slide */  }

   public void put(String key, Value val)
   {  root = put(root, key, val, 0);  }

   private Node put(Node x, String key, Value val, int d)
   {
      if (x == null) x = new Node();
      if (d == key.length()) { x.val = val; return x; }
      char c = key.charAt(d);
      x.next[c] = put(x.next[c], key, val, d+1);
      return x;
   }

   private Value get(String key)
   {  /* similar, see book or booksite */  }
}
```

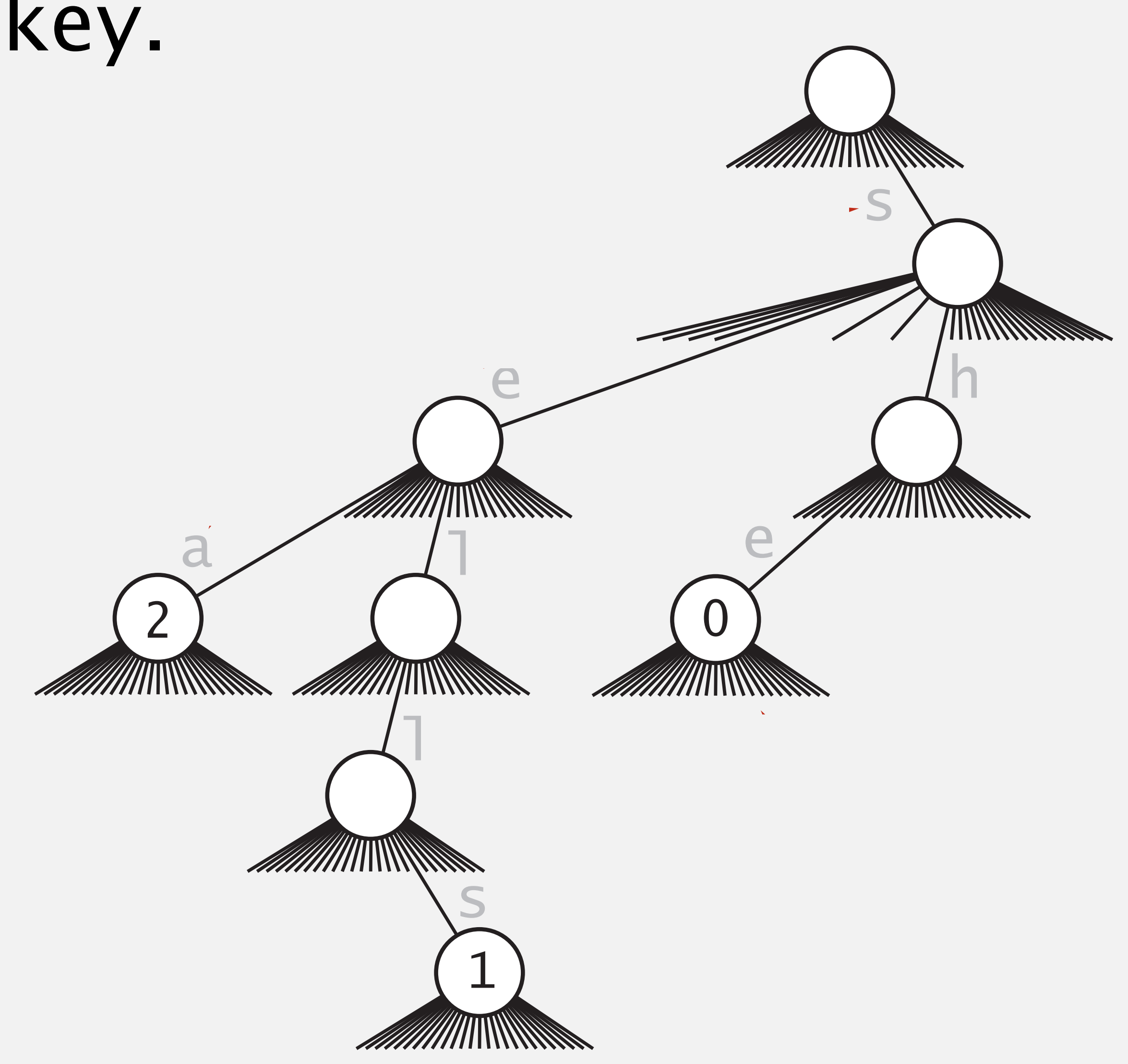
Trie performance

Search hit. Need to examine all L characters for equality.

Search miss.

- Worst case: examine L characters.
- Typical case: examine only a few characters before mismatch (sublinear).

Space. At least R links per key.



Bottom line. Fast search hit and even faster search miss, but wastes space.

Trie quiz 1

What is worst-case running time to insert a key of length L into an R-way trie that contains n key-value pairs?

R = alphabet size
L = length of key
n = number of keys

A. $\Theta(L)$

B. $\Theta(R + L)$

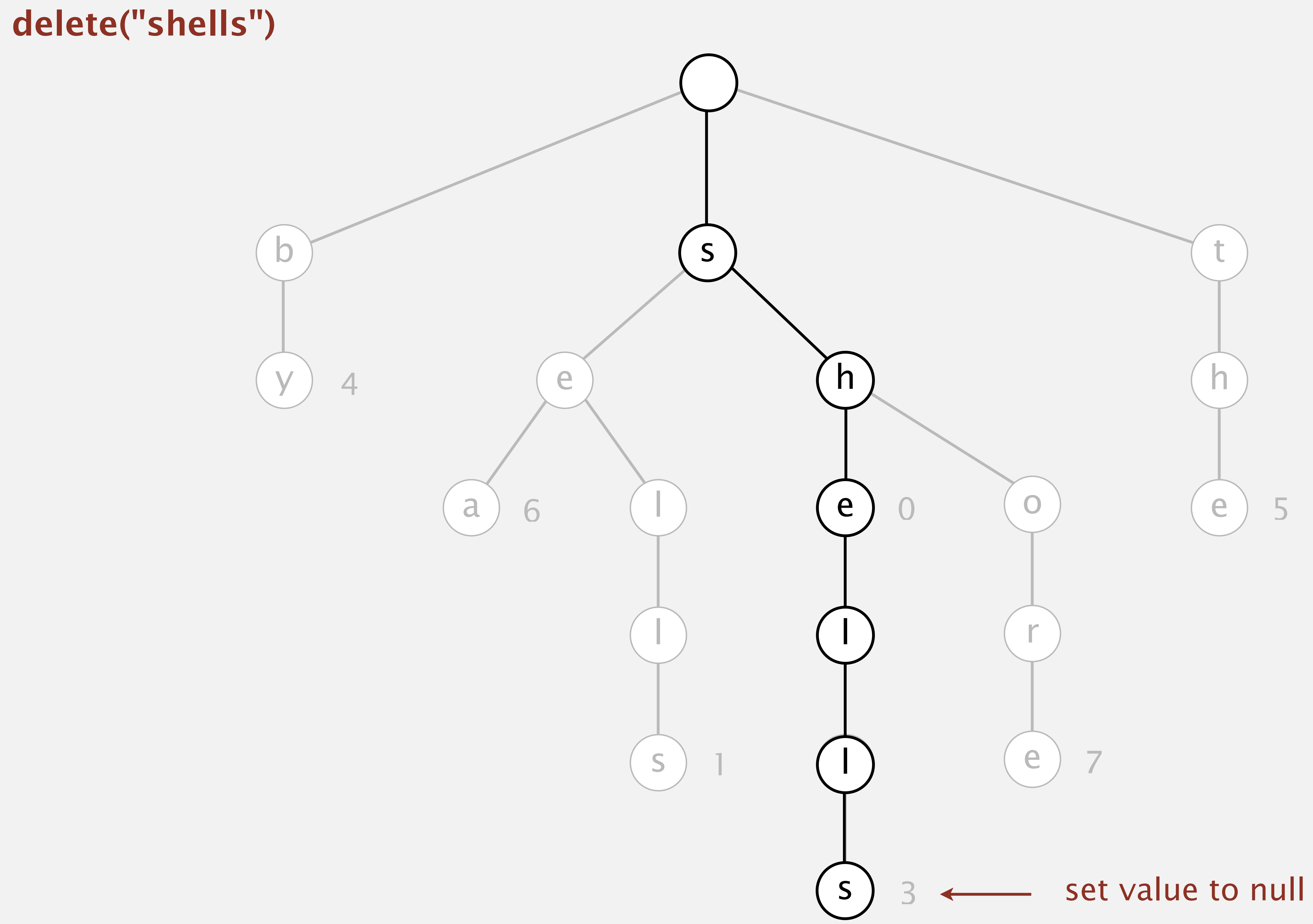
C. $\Theta(n + L)$

D. $\Theta(R\ L)$

Deletion in an R-way trie

To delete a key–value pair:

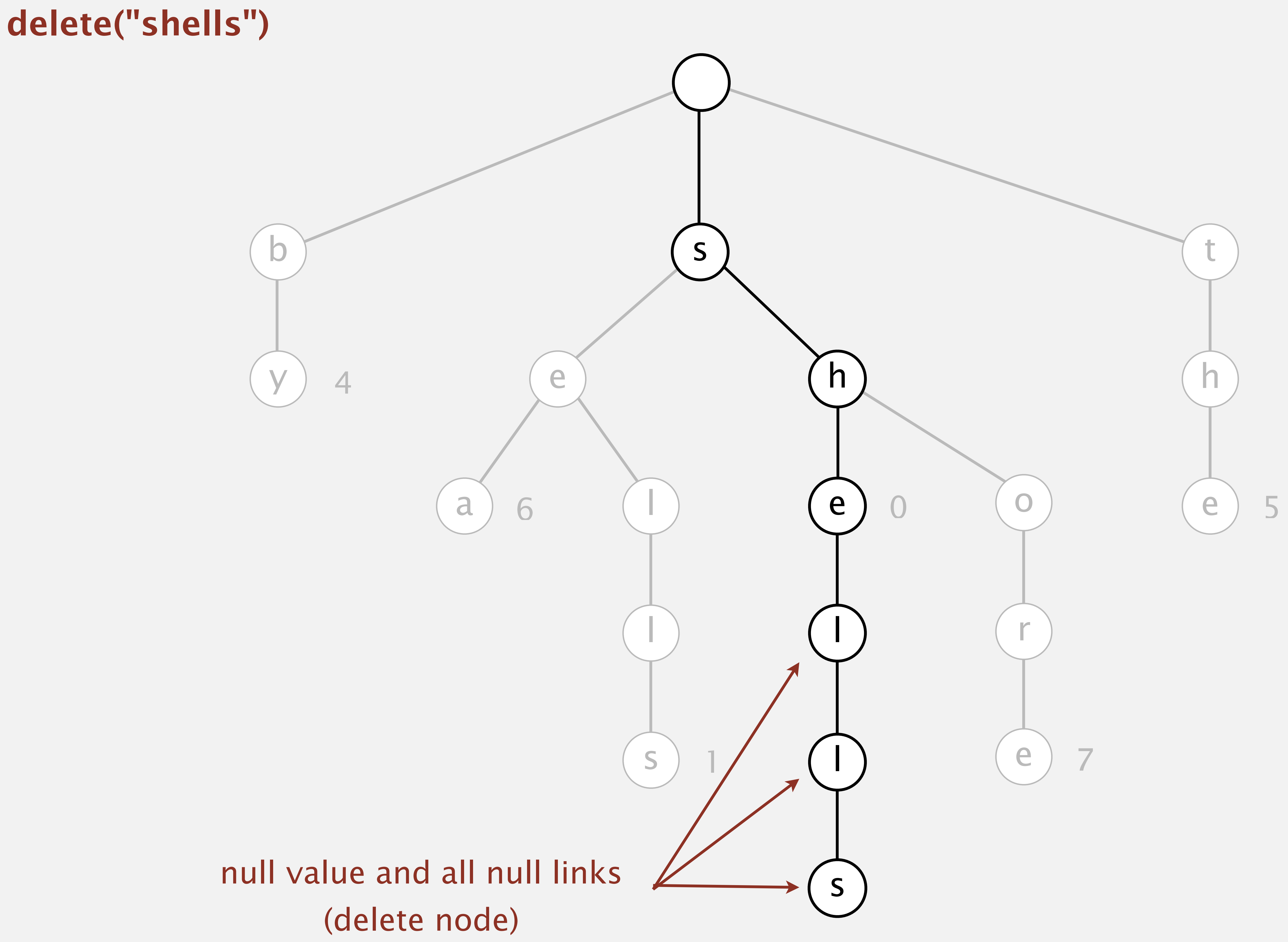
- Find the node corresponding to key and set value to null.
- If node has null value and all null links, remove that node (and recur).



Deletion in an R-way trie

To delete a key–value pair:

- Find the node corresponding to key and set value to null.
- If node has null value and all null links, remove that node (and recur).



String symbol table implementations cost summary

	character accesses (typical case)				dedup	
implementation	**search hit**	**search miss**	**insert**	**space (references)**	**moby.txt**	**actors.txt**
red-black BST	$L + \log^2 n$	$\log^2 n$	$\log^2 n$	$4\,n$	1.4	97.4
hashing (linear probing)	L	L	L	$4\,n$ to $16\,n$	0.76	40.6
R-way trie	L	$\log_R n$	$R + L$	$(R+1)\,n$	1.12	*out of memory*

R-way trie.

- Method of choice for small R.
- Works well for medium R.
- Too much memory for large R.

Challenge. Use less memory, e.g., a 65,536-way trie for Unicode!

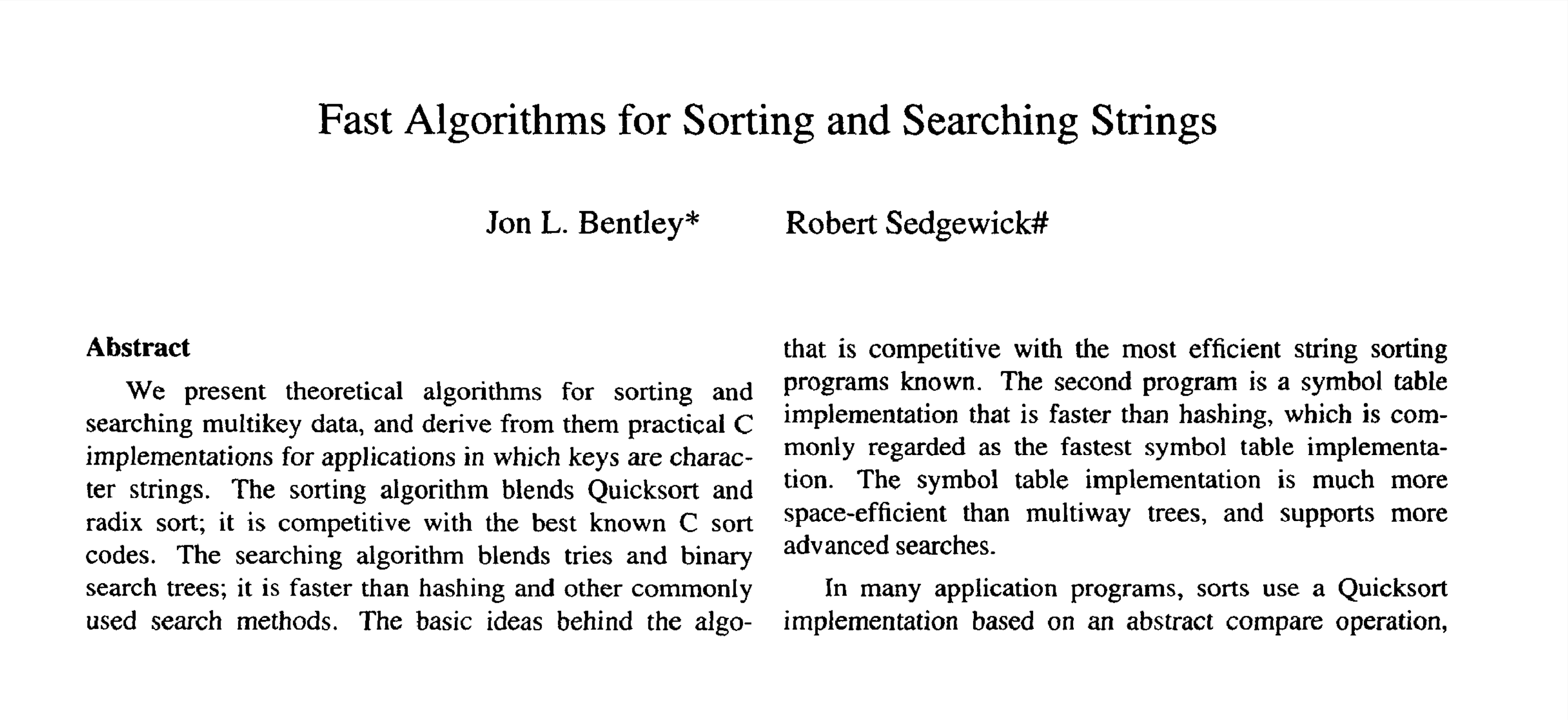
https://algs4.cs.princeton.edu

5.2 Tries

- *string symbol tables*
- *R-way tries*
- *ternary search tries*
- *character-based operations*

Ternary search tries

- Store characters and values in nodes (not keys).
- Each node has three children: smaller (left), equal (middle), larger (right).



Fast Algorithms for Sorting and Searching Strings

Jon L. Bentley*   Robert Sedgewick#

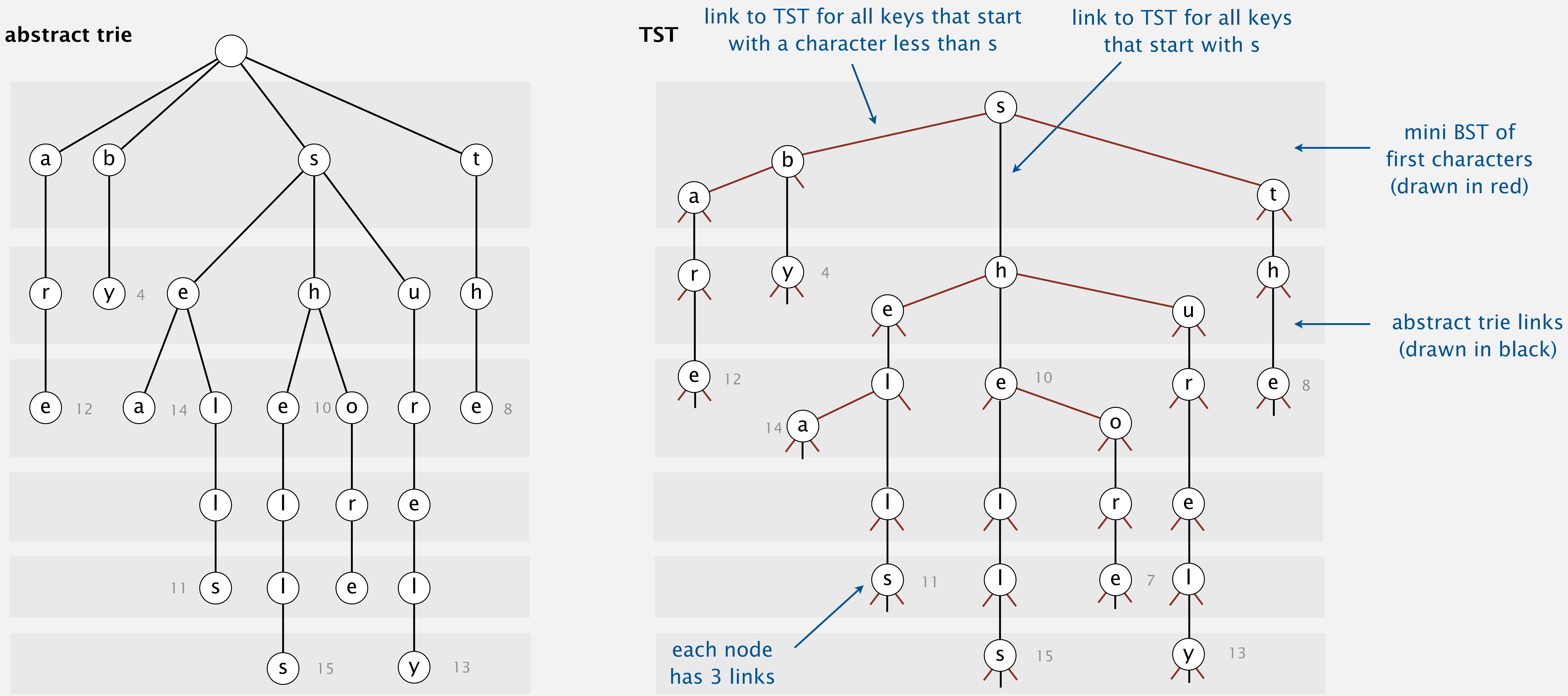
Abstract

We present theoretical algorithms for sorting and searching multikey data, and derive from them practical C implementations for applications in which keys are character strings. The sorting algorithm blends Quicksort and radix sort; it is competitive with the best known C sort codes. The searching algorithm blends tries and binary search trees; it is faster than hashing and other commonly used search methods. The basic ideas behind the algo- that is competitive with the most efficient string sorting programs known. The second program is a symbol table implementation that is faster than hashing, which is commonly regarded as the fastest symbol table implementation. The symbol table implementation is much more space-efficient than multiway trees, and supports more advanced searches.

In many application programs, sorts use a Quicksort implementation based on an abstract compare operation,

Ternary search tries

- Store characters and values in nodes (not keys).
- Each node has three children: smaller (left), equal (middle), larger (right).



Trie quiz 2

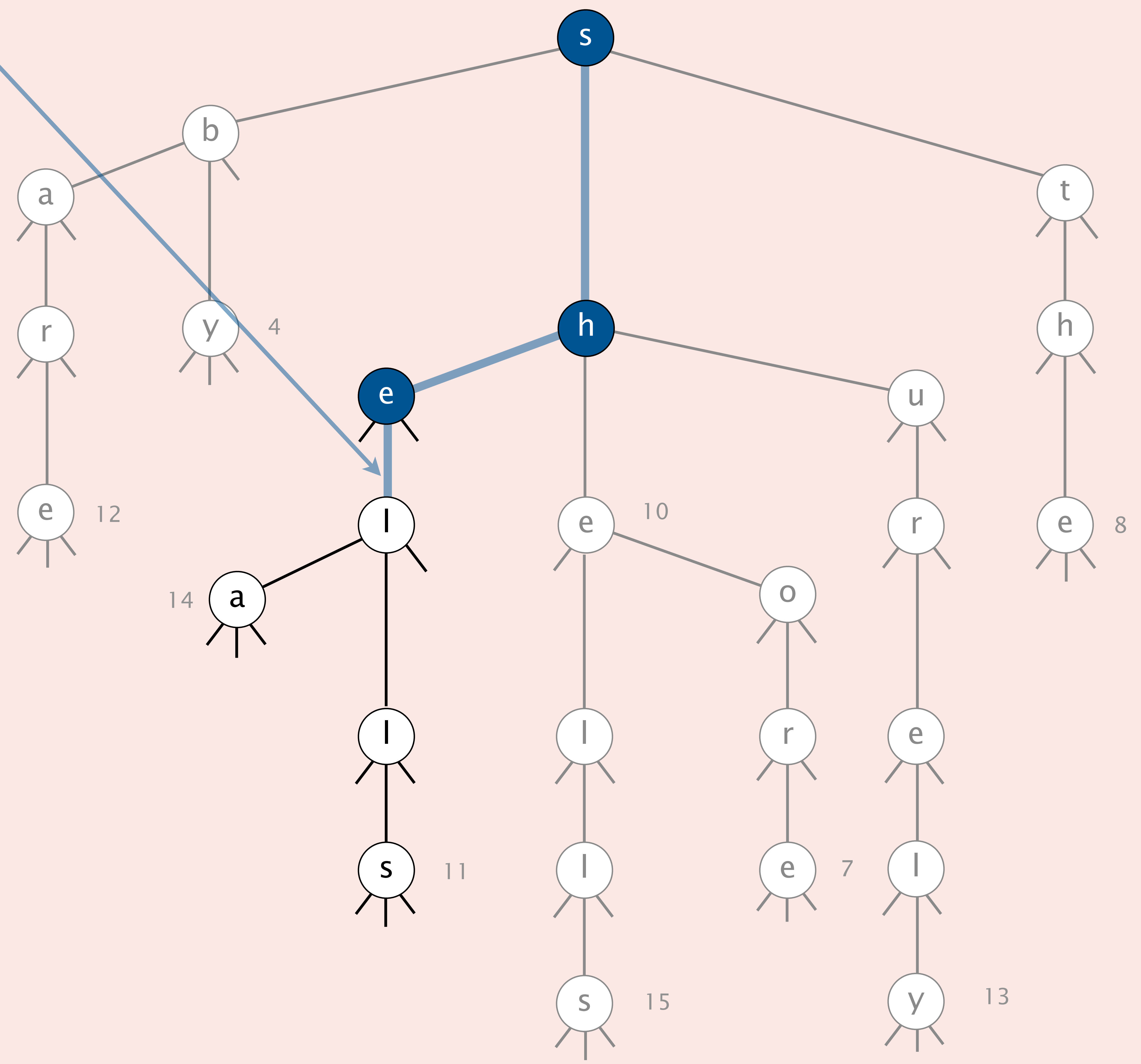
Which keys are stored in this subtrie of the given TST?

A. Strings that start with `s`.

B. Strings that start with `se`.

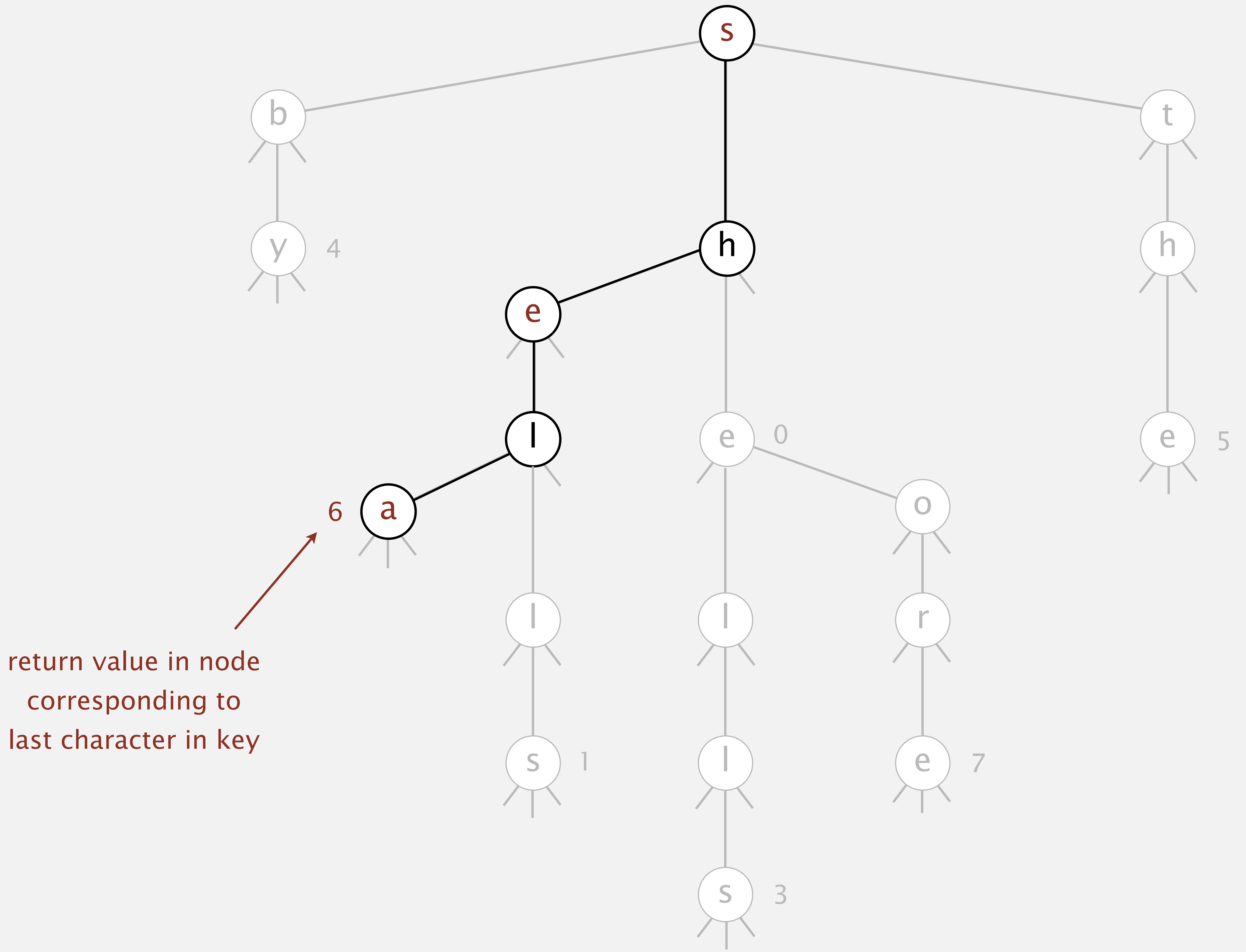
C. Strings that start with `sh`.

D. Strings that start with `she`.



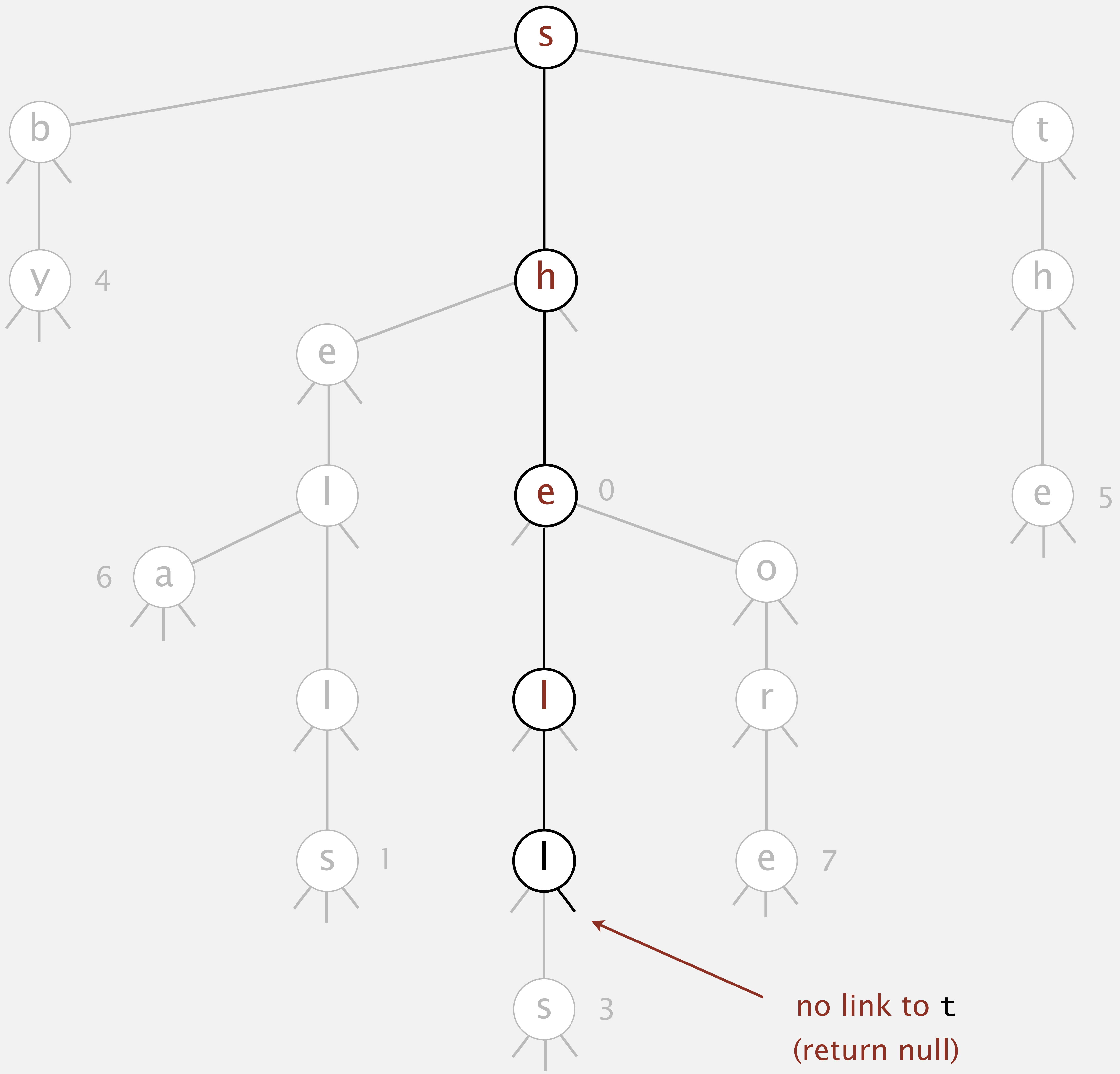
Search hit in a TST

get("sea")



Search miss in a TST

get("shelter")



Search in a TST

Compare key character to key in node and follow links accordingly:

- If less, go left.
- If greater, go right.
- If equal, go middle and advance to the next key character.

Search hit. Node where search ends has a non-null value.

Search miss. Either (1) reach a null link or (2) node where search ends has null value.

Trie quiz 2

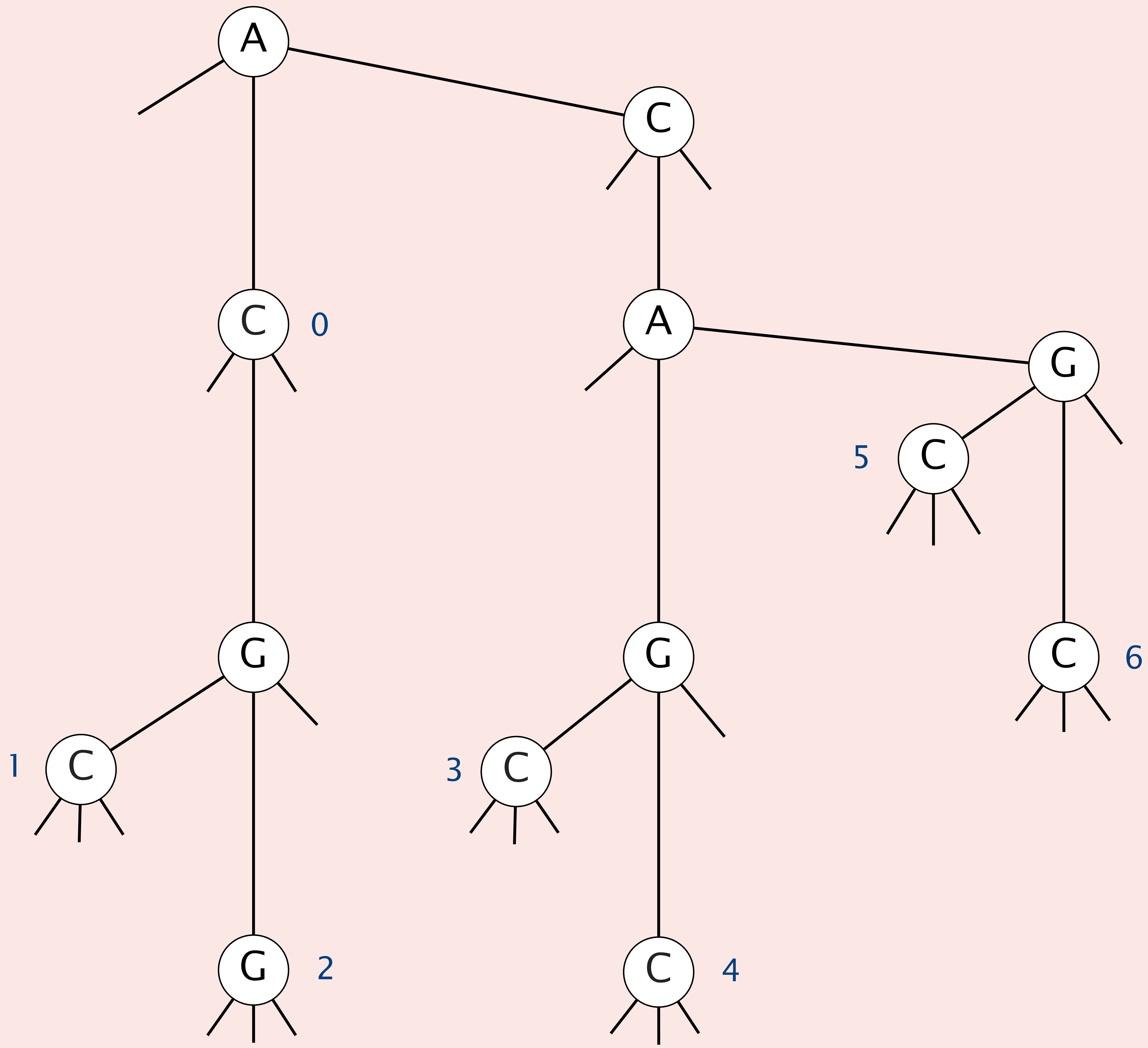
Which value is associated with the key CAC ?

A. 3

B. 4

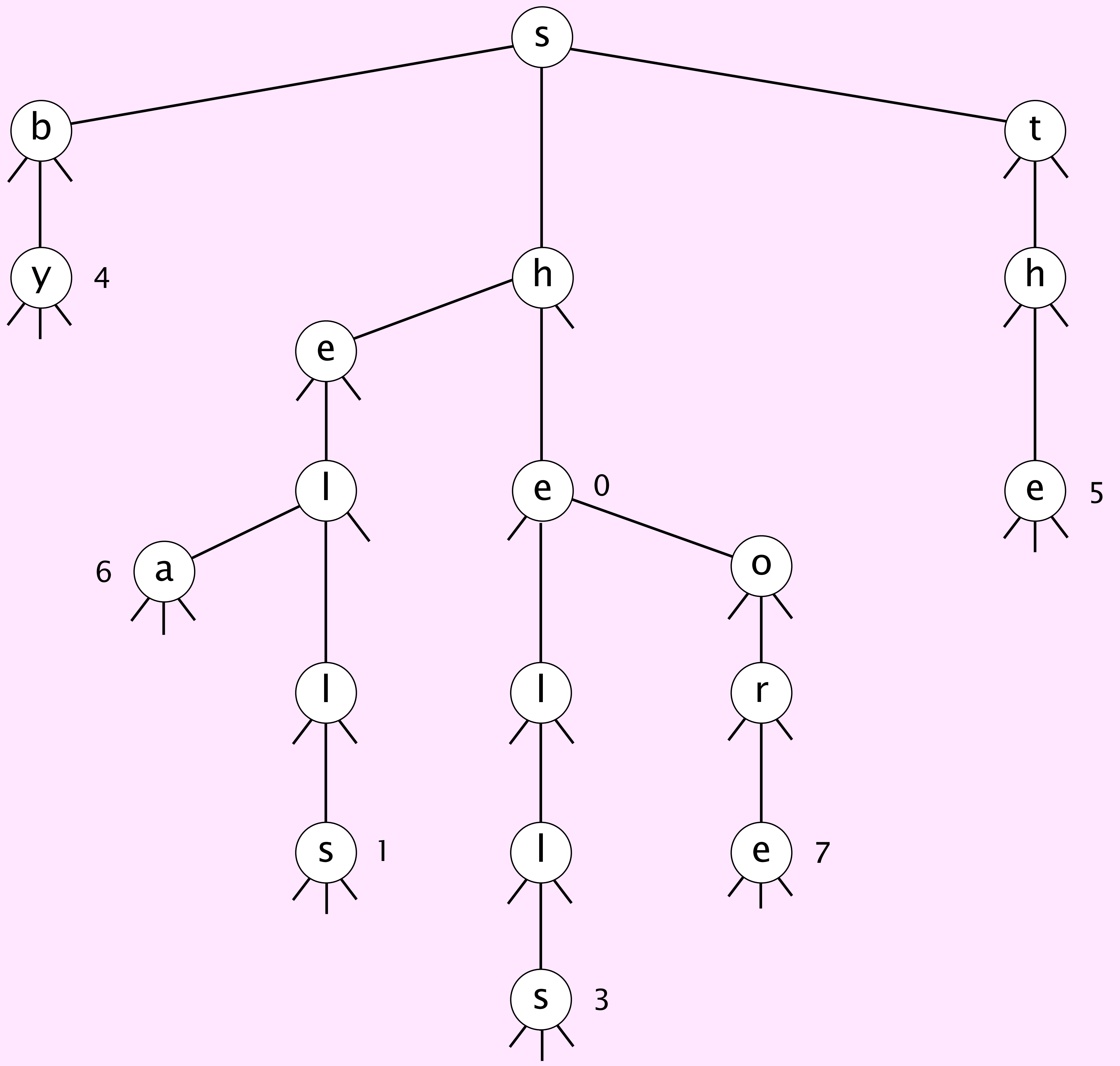
C. 5

D. *null*



Ternary search trie construction demo

ternary search trie



Trie quiz 3

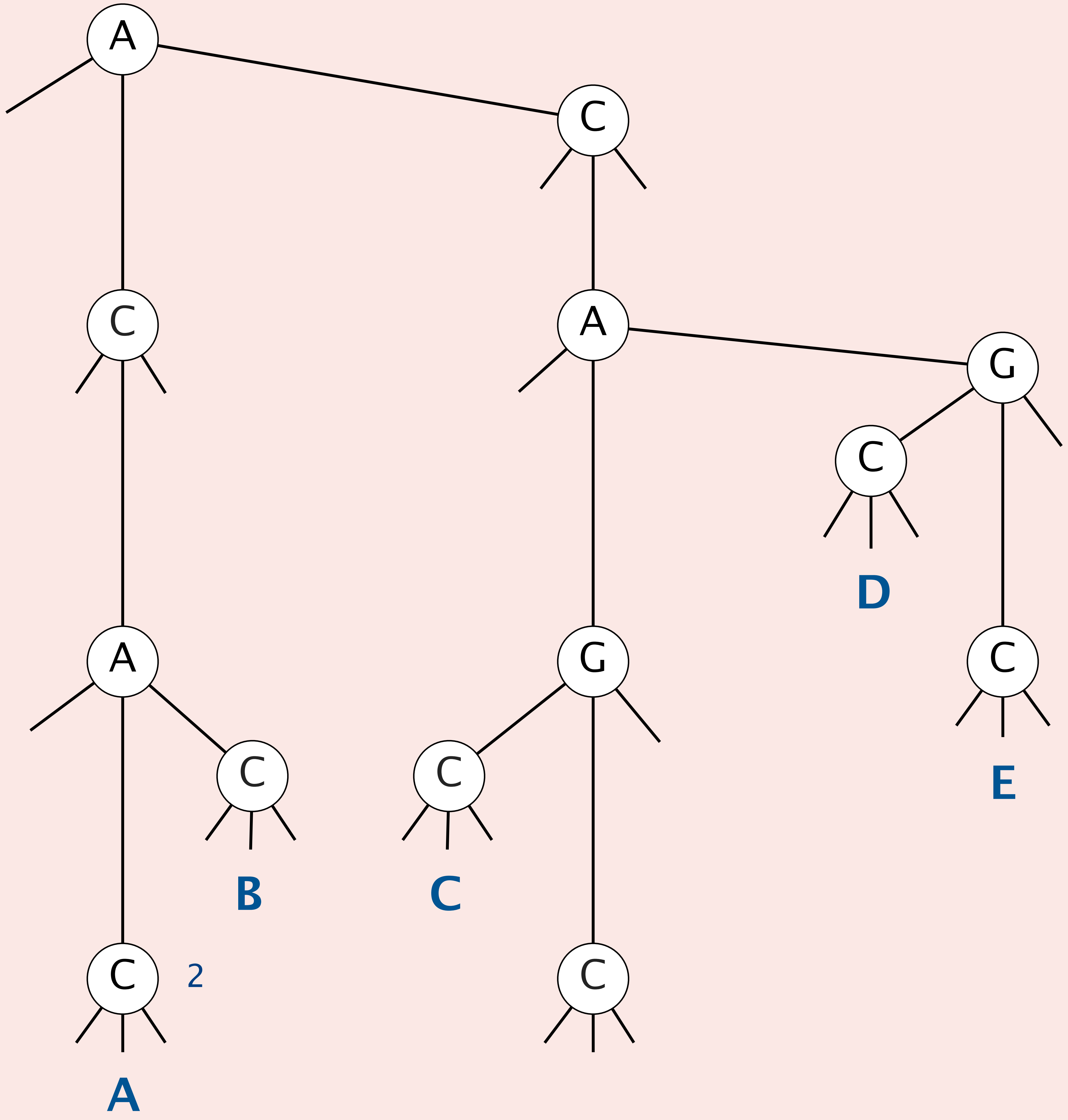
In which subtrie would the key CCC be inserted?

A.

B.

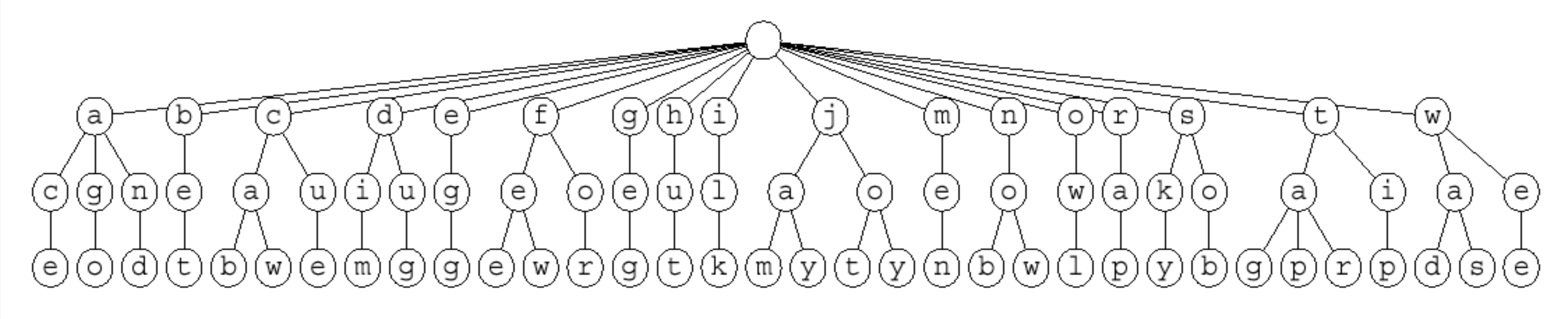
C.

D.



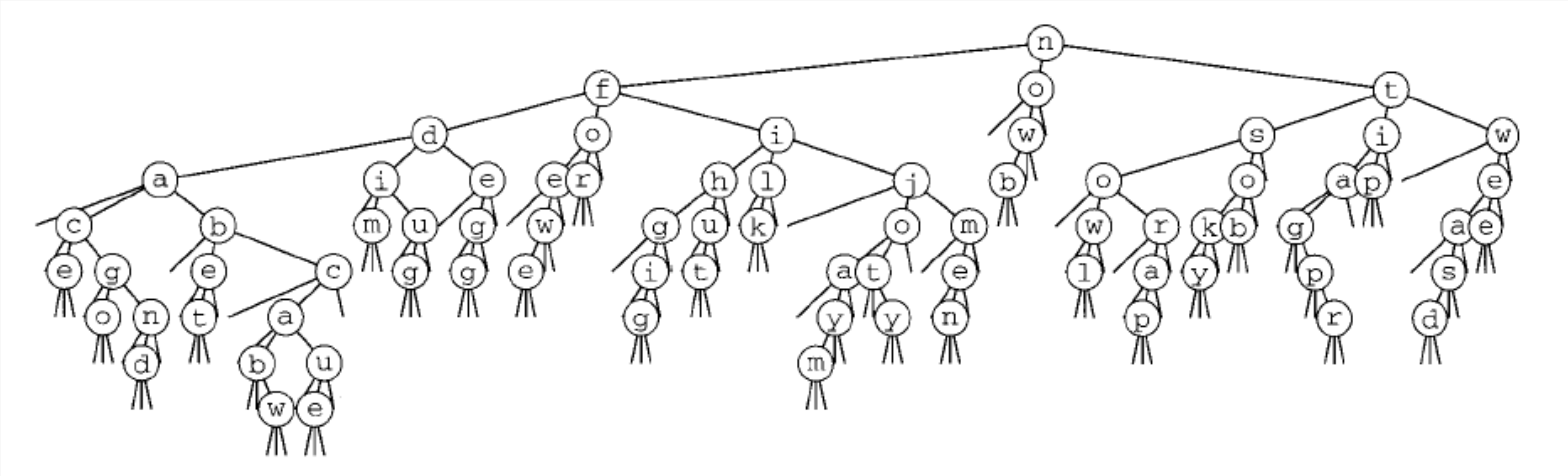
26-way trie vs. TST

26-way trie. 26 null links in each leaf.



26-way trie (1035 null links, not shown)

TST. 3 null links in each leaf.



TST (155 null links)

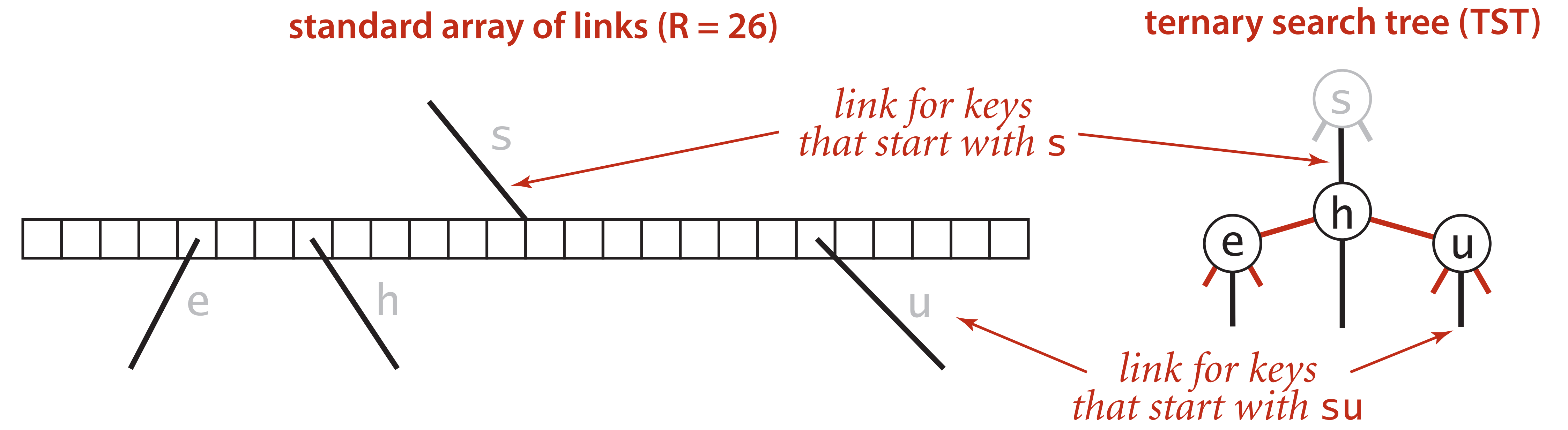
```
now
for
tip
ilk
dim
tag
jot
sob
nob
sky
hut
ace
bet
men
egg
few
jay
owl
joy
rap
gig
wee
was
cab
wad
caw
cue
fee
tap
ago
tar
jam
dug
and
```

TST representation in Java

A TST node is five fields:

- A value.
- A character.
- A reference to a left TST.
- A reference to a middle TST.
- A reference to a right TST.

```
private class Node
{
   private Value val;
   private char c;
   private Node left, mid, right;
}
```



Trie node representations

TST: Java implementation

```
public class TST<Value>
{
   private Node root;
   private class Node
   {  /* see previous slide */  }

   public Value get(String key)
   {  return get(root, key, 0);  }

   private Value get(Node x, String key, int d)
   {
      if (x == null) return null;
      char c = key.charAt(d);
      if      (c < x.c)              return get(x.left,  key, d);
      else if (c > x.c)              return get(x.right, key, d);
      else if (d < key.length() - 1) return get(x.mid,   key, d+1);
      else                           return x.val;
   }

   public void put(String Key, Value val)
   {  /* similar, see book or booksite */  }
}
```

String symbol table implementation cost summary

	character accesses (typical case)				dedup	
implementation	search hit	search miss	insert	space (references)	moby.txt	actors.txt
red-black BST	$L + \log^2 n$	$\log^2 n$	$\log^2 n$	$4n$	1.4	97.4
hashing (linear probing)	L	L	L	$4n$ to $16n$	0.76	40.6
R-way trie	L	$\log_R n$	$R + L$	$(R+1)\, n$	1.12	*out of memory*
TST	$L + \log n$	$\log n$	$L + \log n$	$4n$	0.72	38.7

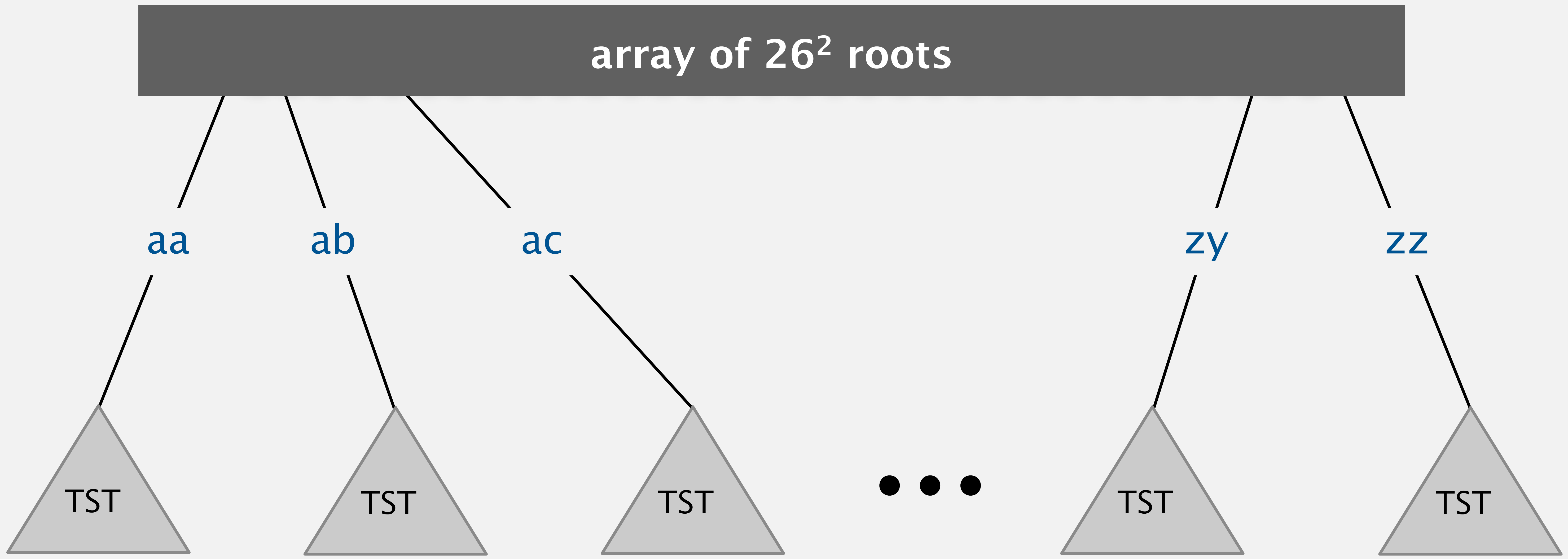
Remark. Can build balanced TSTs via rotations to achieve $\Theta(L + \log n)$ in worst case.

Bottom line. TST is as fast as hashing (for string keys) and space efficient.

TST with R^2 branching at root

Hybrid of R-way trie and TST.

- Do R^2-way branching at root.
- Each of R^2 root nodes points to a TST.



Q. What about one- and two-letter words?

String symbol table implementation cost summary

	character accesses (typical case)				dedup	
implementation	**search hit**	**search miss**	**insert**	**space (references)**	**moby.txt**	**actors.txt**
red-black BST	$L + \log^2 n$	$\log^2 n$	$\log^2 n$	$4n$	1.4	97.4
hashing (linear probing)	L	L	L	$4n$ to $16n$	0.76	40.6
R-way trie	L	$\log_R n$	$R + L$	$(R+1)\, n$	1.12	*out of memory*
TST	$L + \log n$	$\log n$	$L + \log n$	$4n$	0.72	38.7
TST with R^2	$L + \log n$	$\log n$	$L + \log n$	$4n + R^2$	0.51	32.7

Bottom line. Faster than hashing for our benchmark client.

TST vs. hashing

Hashing.

- Need to examine entire key.
- Search hits and misses cost about the same.
- Performance relies on hash function.
- Does not support ordered symbol table operations.

TSTs.

- Works only for string (or digital) keys.
- Search miss may involve only a few characters.
- Supports ordered symbol table operations (plus extras!).

Bottom line. TSTs are:

- Faster than hashing (especially for search misses).
- More flexible than red–black BSTs. [ahead]

https://algs4.cs.princeton.edu

5.2 Tries

- *string symbol tables*
- *R-way tries*
- *ternary search tries*
- *character-based operations*

String symbol table API

Character-based operations. The string symbol table API supports several useful character-based operations.

key	value
by	4
sea	6
sells	1
she	0
shells	3
shore	7
the	5

Prefix match. Keys with prefix `sh`: `she`, `shells`, and `shore`.

Longest prefix. Key that is the longest prefix of `shellsort`: `shells`.

String symbol table API

public class StringST<Value>		
	StringST()	*create a symbol table with string keys*
void	put(String key, Value val)	*put key-value pair into the symbol table*
Value	get(String key)	*value paired with key*
void	delete(String key)	*delete key and corresponding value*
Iterable<String>	keys()	*all keys (in sorted order)*
	⋮	
Iterable<String>	keysWithPrefix(String s)	*keys having s as a prefix*
String	longestPrefixOf(String s)	*longest key that is a prefix of s*

Remark. Can also add other ordered ST methods, e.g., `floor()` and `rank()`.

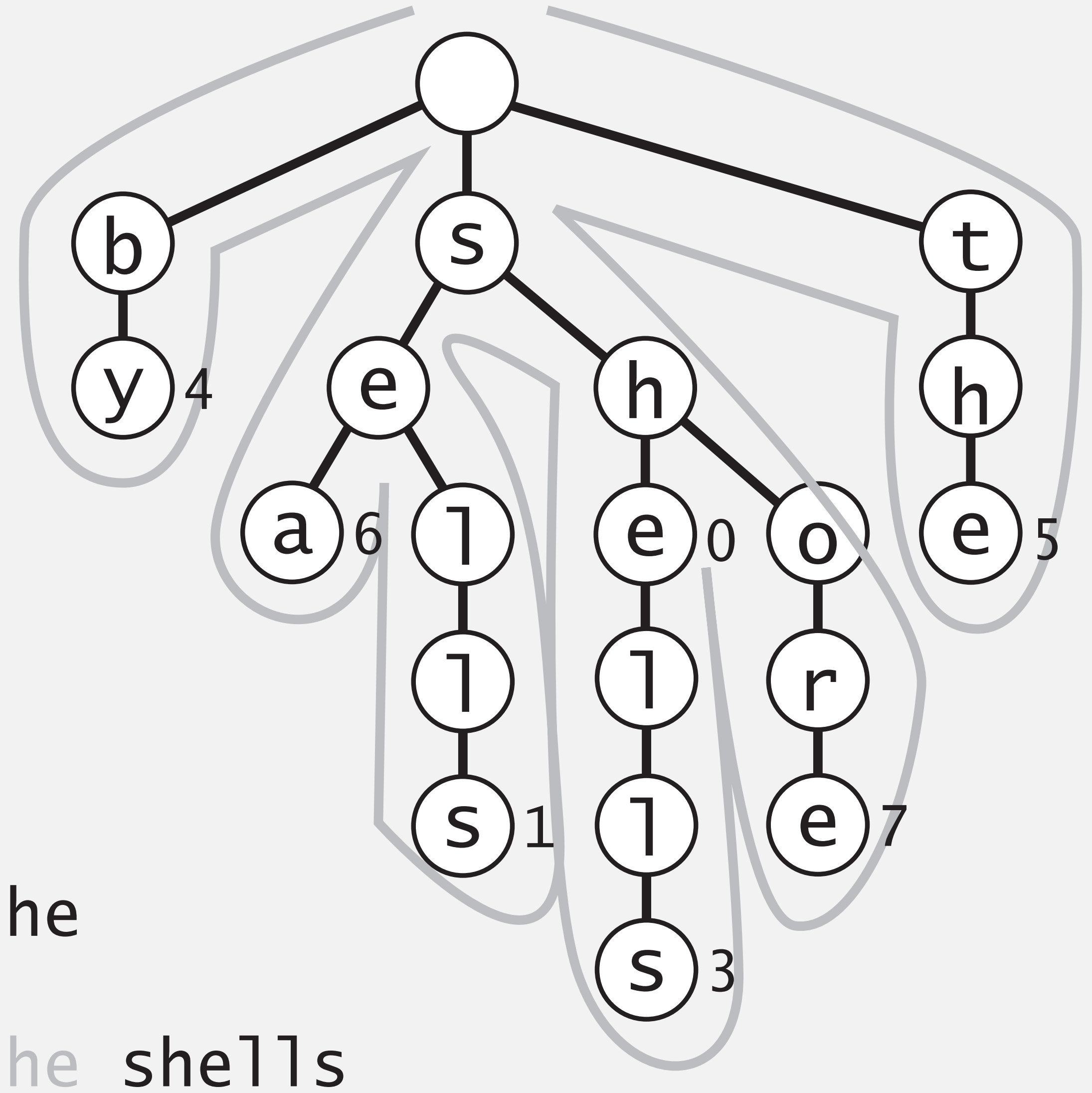
Warmup: ordered iteration

To iterate through all keys in sorted order:

- Do inorder traversal of trie; add keys encountered to a queue.
- Maintain sequence of characters on path from root to node.

`keys()`

key	q
b	
by	by
s	
se	
sea	by sea
sel	
sell	
sells	by sea sells
sh	
she	by sea sells she
shell	
shells	by sea sells she shells
sho	
shor	
shore	by sea sells she shells shore
t	
th	
the	by sea sells she shells shore the



Ordered iteration: Java implementation

To iterate through all keys in sorted order:

- Do inorder traversal of trie; add keys encountered to a queue.
- Maintain sequence of characters on path from root to node.

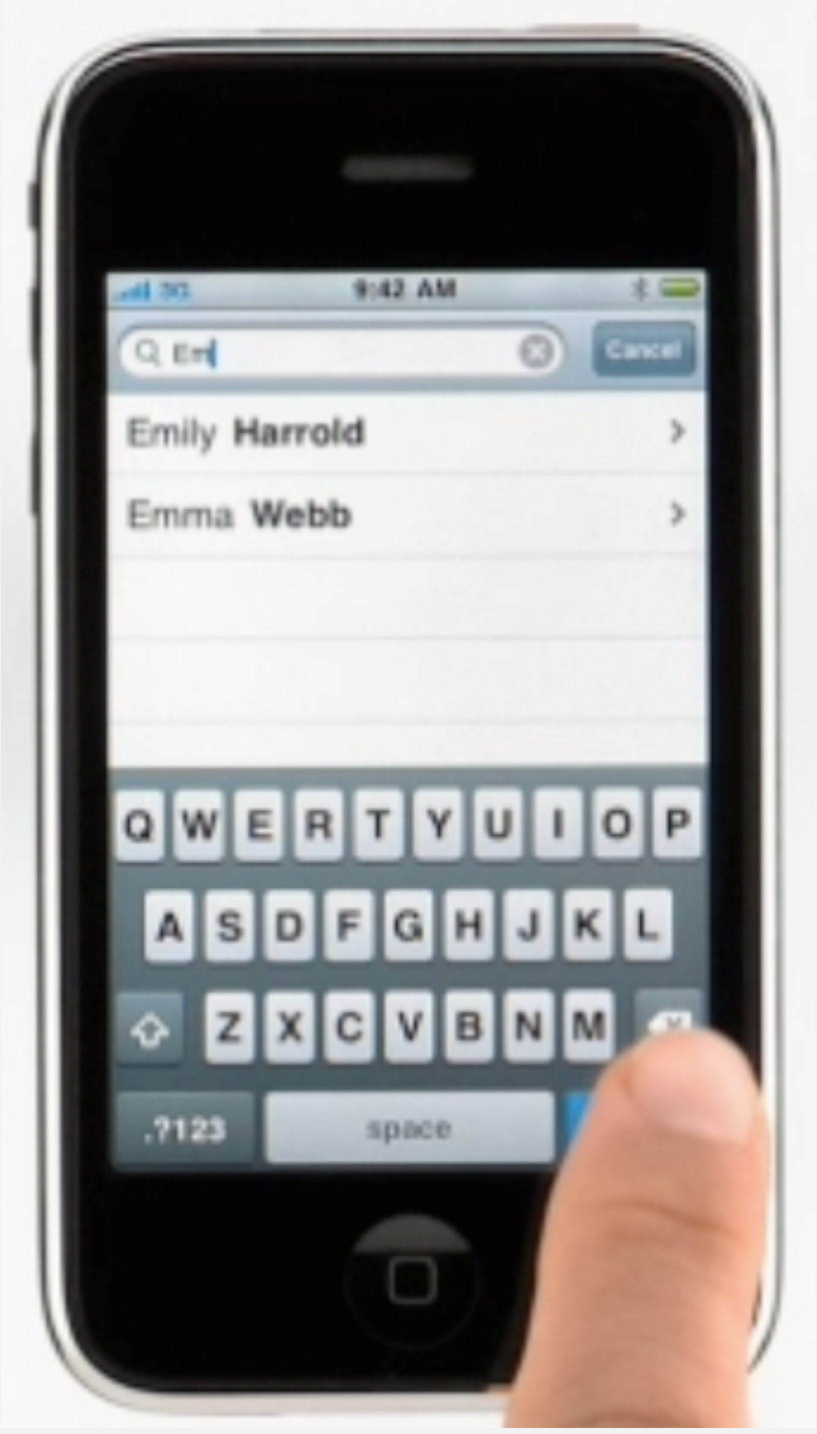
```
public Iterable<String> keys()
{
   Queue<String> queue = new Queue<String>();
   collect(root, "", queue);
   return queue;
}
                                          sequence of characters
                                          on path from root to x

private void collect(Node x, String prefix, Queue<String> queue)
{
   if (x == null) return;
   if (x.val != null) queue.enqueue(prefix);
   for (char c = 0; c < R; c++)
      collect(x.next[c], prefix + c, queue);
}
                                    or use StringBuilder
```

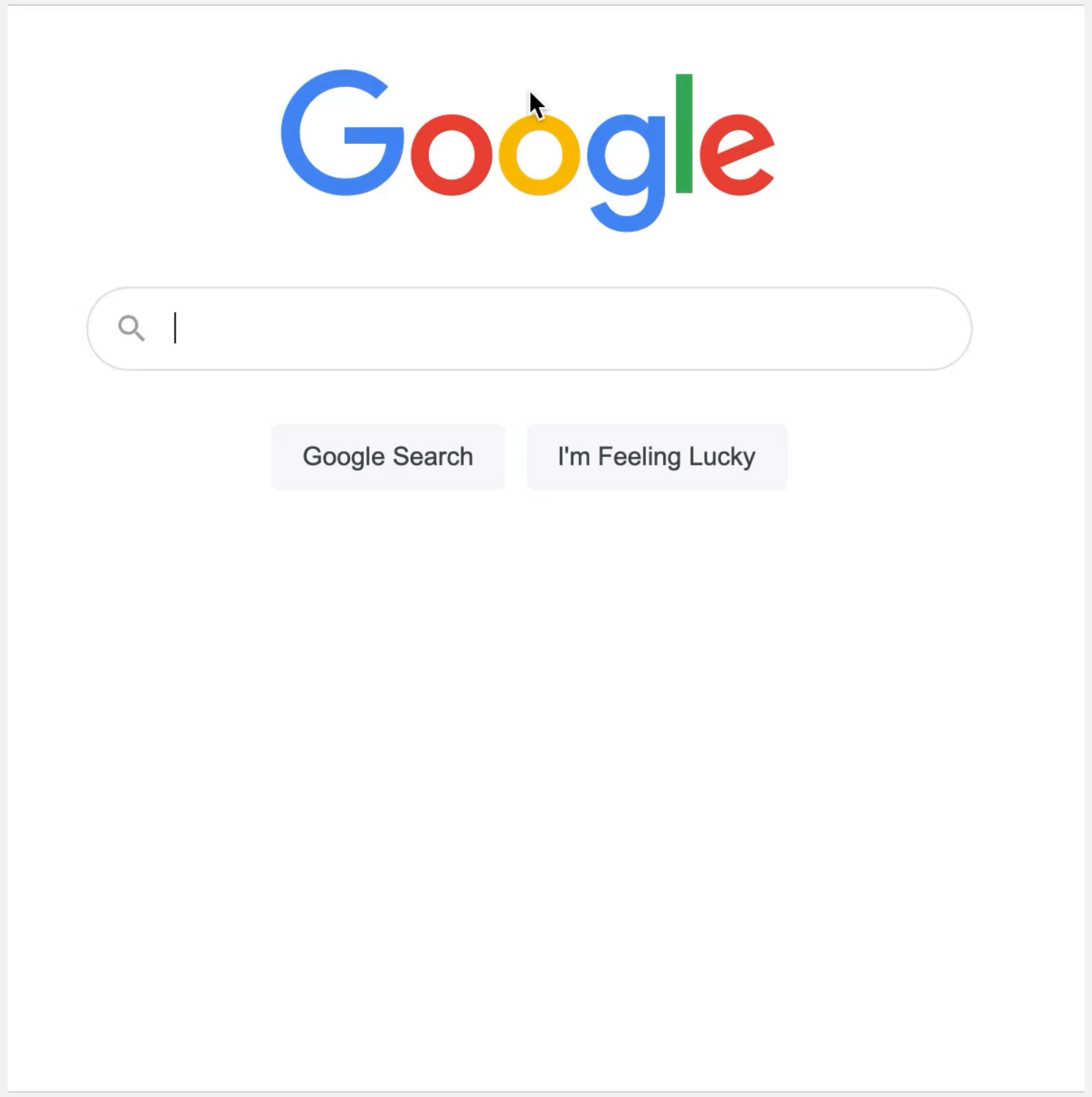
Prefix matches

Find all keys in a symbol table starting with a given prefix.

Ex. Autocomplete in a cell phone, search bar, text editor, or shell.

- User types characters one at a time.
- System reports all matching strings.

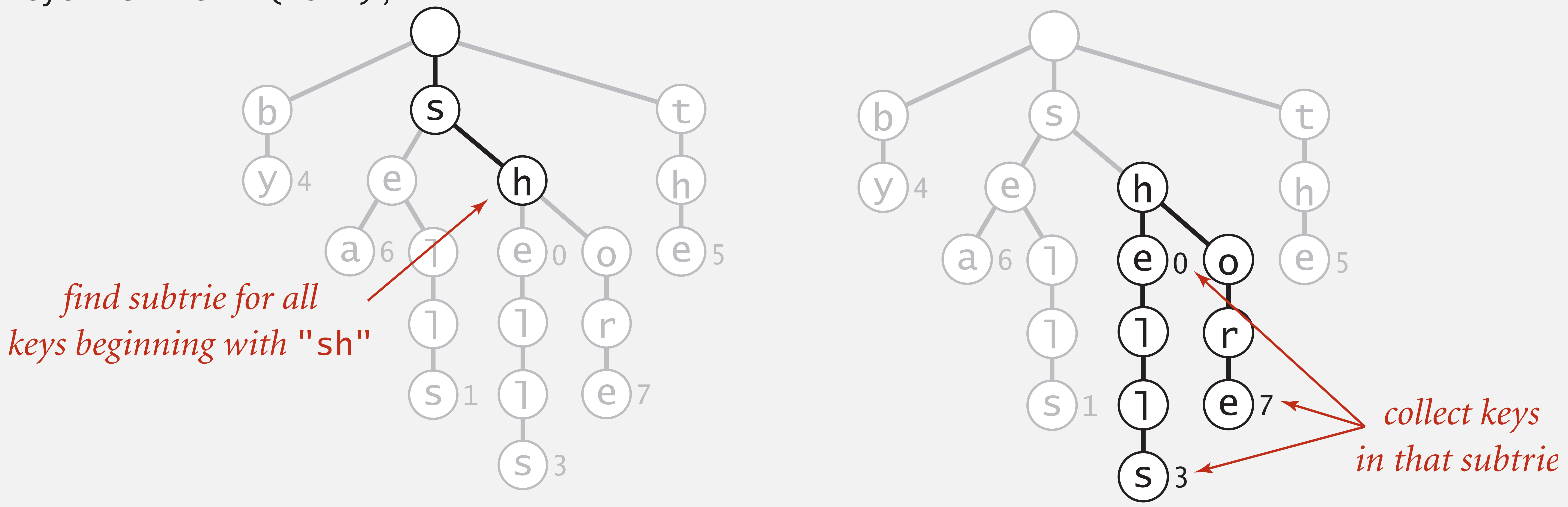




Prefix matches in an R-way trie

Find all keys in a symbol table starting with a given prefix.

```
keysWithPrefix("sh");
```



Longest prefix

Find longest key in symbol table that is a prefix of query string.

Ex 1. To send packet toward destination IP address, router chooses IP address in routing table that is longest prefix match.

```
"128"
"128.112"
"128.112.055"
"128.112.055.15"
"128.112.136"
"128.112.155.11"
"128.112.155.13"
"128.222"
"128.222.136"
```

represented as 32-bit binary number for IPv4 (instead of string)

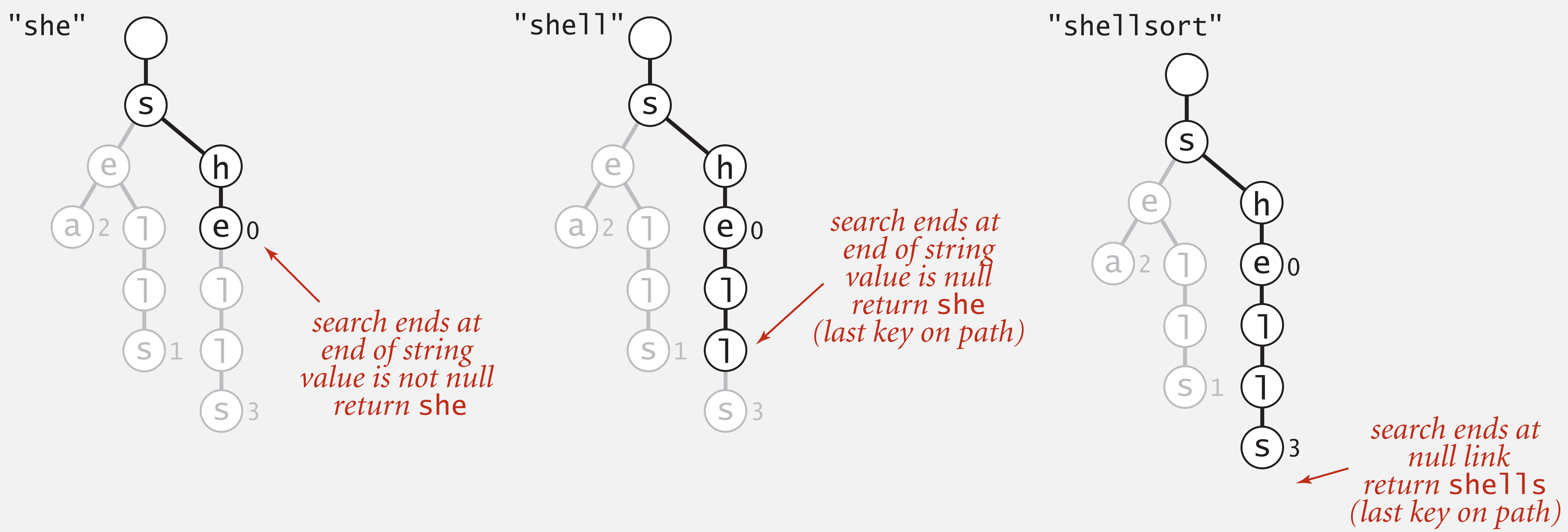
```
longestPrefixOf("128.112.136.11") = "128.112.136"
longestPrefixOf("128.112.100.16") = "128.112"
longestPrefixOf("128.166.123.45") = "128"
```

Note. Not the same as floor: `floor("128.112.100.16") = "128.112.055.15"`

Longest prefix in an R-way trie

Find longest key in symbol table that is a prefix of query string.

- Search for query string.
- Keep track of longest key encountered.



Possibilities for `longestPrefixOf()`

T9 texting (predictive texting)

Goal. Type text messages on a phone keypad.

Multi-tap input. Enter a letter by repeatedly pressing a key.

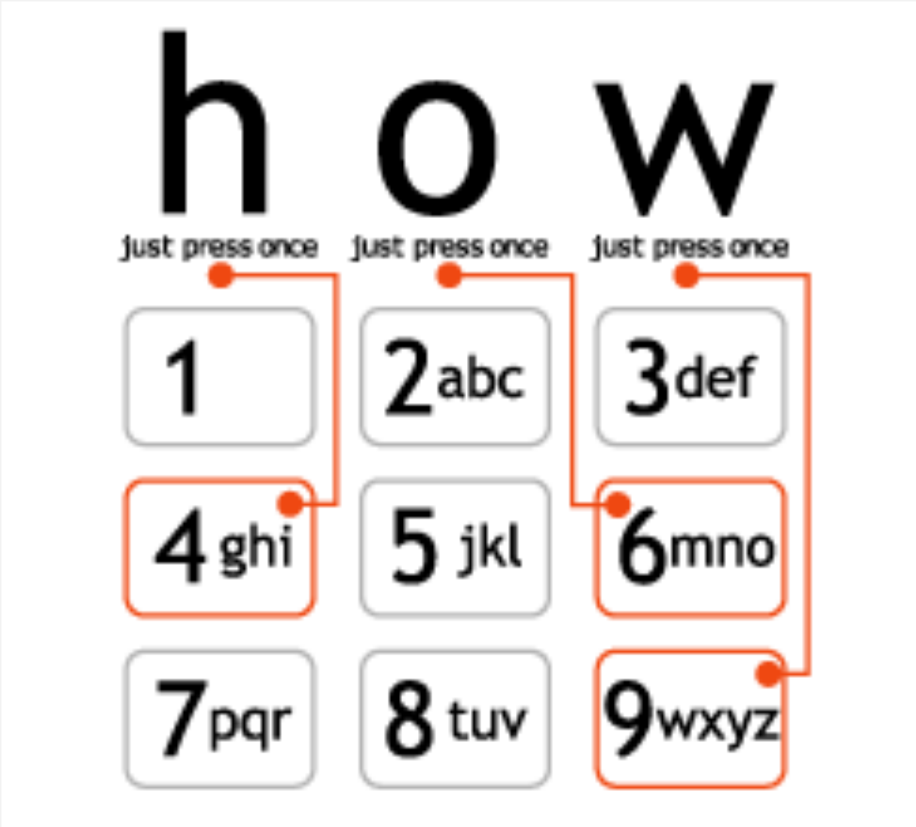
Ex. good: 4 6 6 6 6 6 6 3

"a much faster and more fun way to enter text"

T9 text input (on 4 billion handsets).

- Find all words that correspond to given sequence of numbers.

  4663: good, home, gone, hoof. ← textonyms
- Press * to select next option.
- Press 0 to see all completion options.
- System adapts to user's tendencies.



http://www.t9.com

T9 TEXTING

Q. How to implement T9 texting on a mobile phone?

SONY verizon SIEMENS

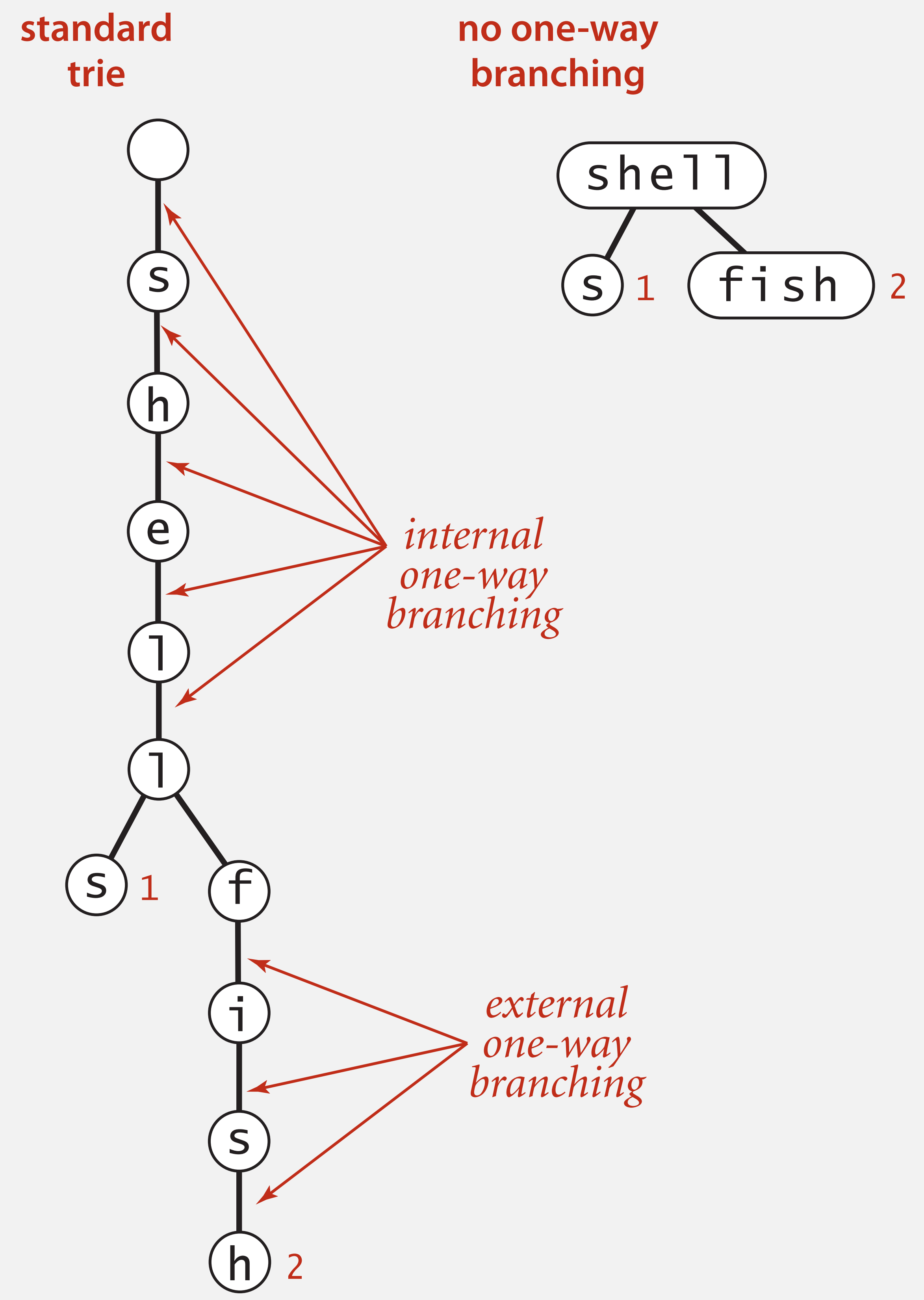
NEC SANYO SAMSUNG ELECTRONICS

Patricia trie

Patricia trie. [Practical Algorithm to Retrieve Information Coded in Alphanumeric]

- Remove one-way branching.
- Each node represents a sequence of characters.
- Implementation: one step beyond this course.

```
put("shells", 1);
put("shellfish", 2);
```



Applications.

- Database search.
- P2P network search.
- IP routing tables: find longest prefix match.
- Compressed quad-tree for n-body simulation.
- Efficiently storing and querying XML documents.

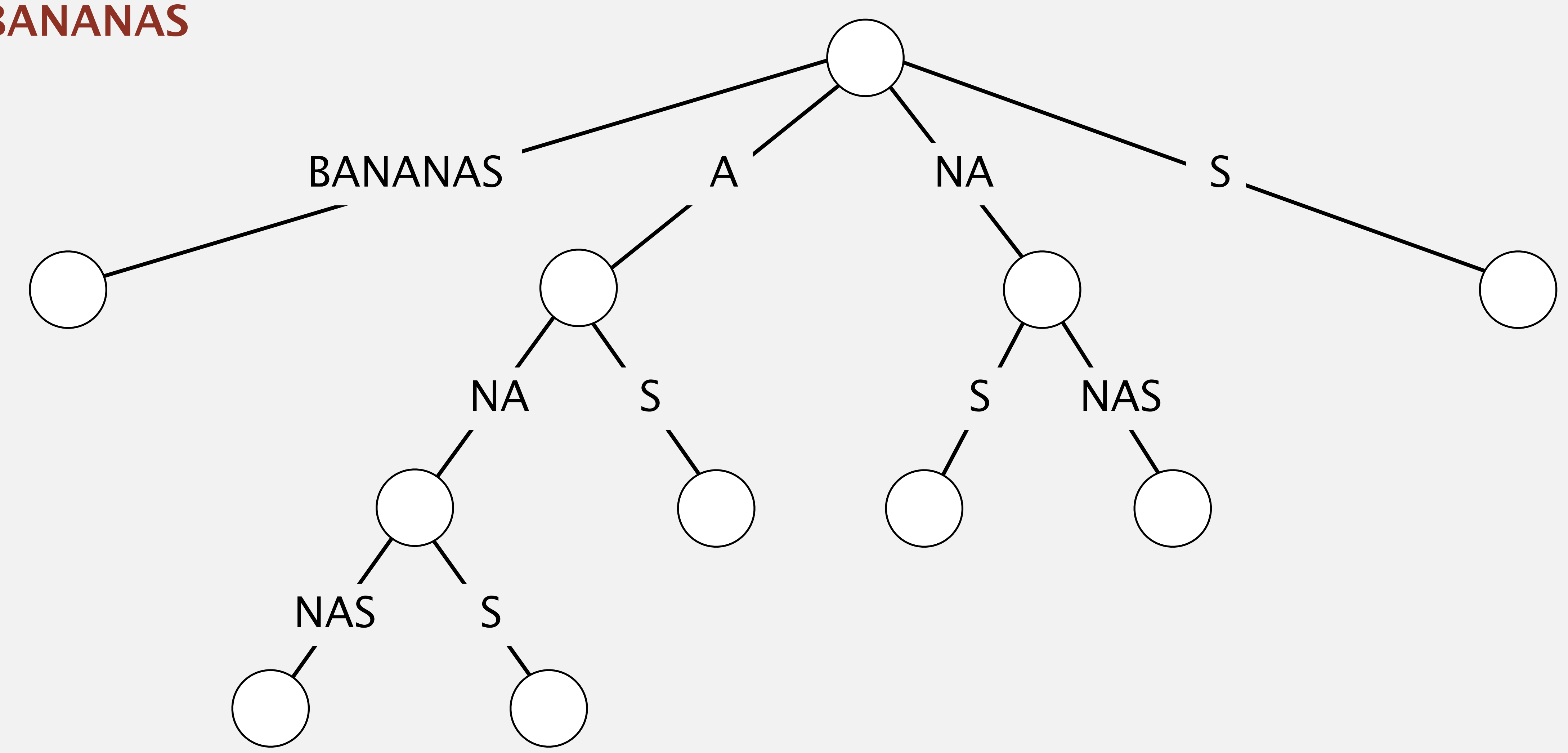
Also known as: crit-bit tree, radix tree.

Suffix tree

Suffix tree.

- Patricia trie of suffixes of a string.
- Linear-time construction: well beyond scope of this course.

suffix tree for BANANAS



Applications.

- Linear-time: longest repeated substring, longest common substring, longest palindromic substring, substring search, tandem repeats, ...
- Computational biology databases (BLAST, FASTA).

String symbol tables summary

A success story in algorithm design and analysis.

Red–black BSTs.

- $\Theta(\log n)$ key compares per search/insert. ← worst case
- Supports ordered symbol table API.

Hash tables.

- $\Theta(1)$ probes per search/insert. ← uniform hashing assumption
- Requires good hash function for key type.

Tries. R-way, TST.

- $O(\log n)$ characters accessed per search/insert. ← typical applications
- Supports character-based operations.