Data Structures and Algorithms

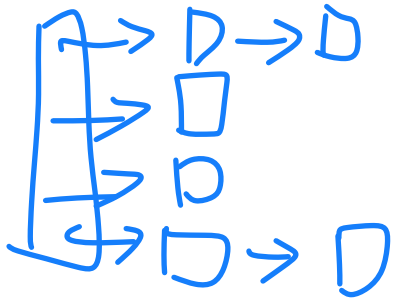
Hashing



CS 225

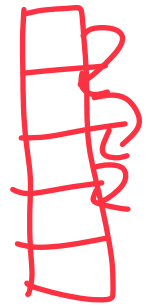
Brad Solomon

November 16, 2026



UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Department of Computer Science



Announcements

Exam 5 and Retake Exam after Fall break!

Retake Exam lets you average out one exam's grade

70% - 30% (Higher score, higher weight)

You do not need to specify the exam you want to retake

Only open one exam while in the CBTF

If you open multiple exams, run risk of losing retake

Learning Objectives

Motivate and formally define a hash table

Discuss what a 'good' hash function looks like

Identify the key weakness of a hash table

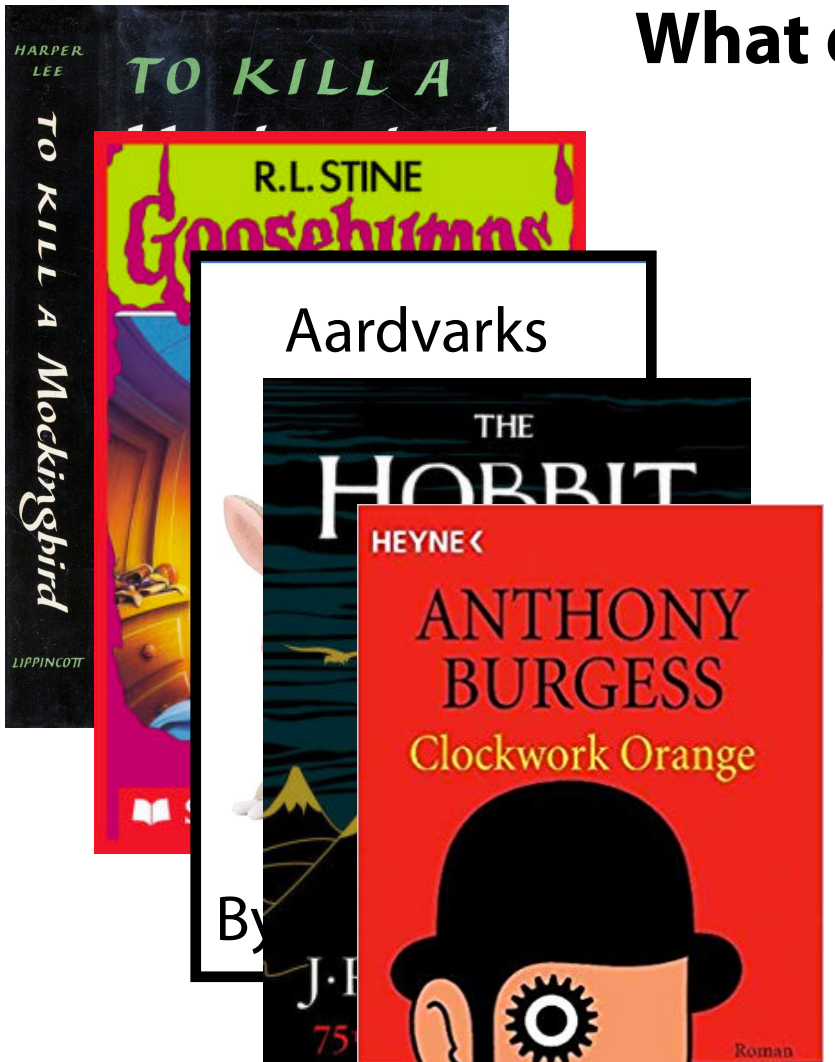
Introduce strategies to "correct" this weakness

We will see that this weakness is pretty bad

Data Structure Review

Key -> Value

I have a collection of books and I want to store them in a dictionary!



What data structures can I use here?

AVL Tree *

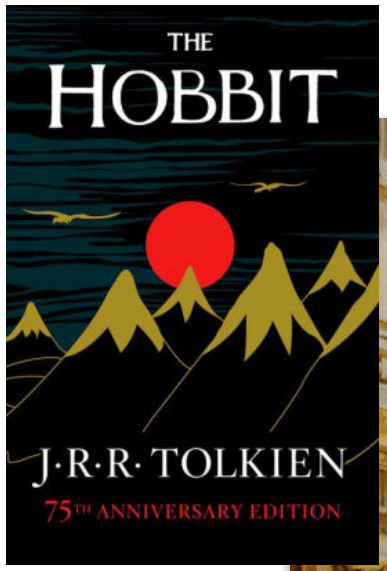
Insert: O(log n)
Find: O(log n)

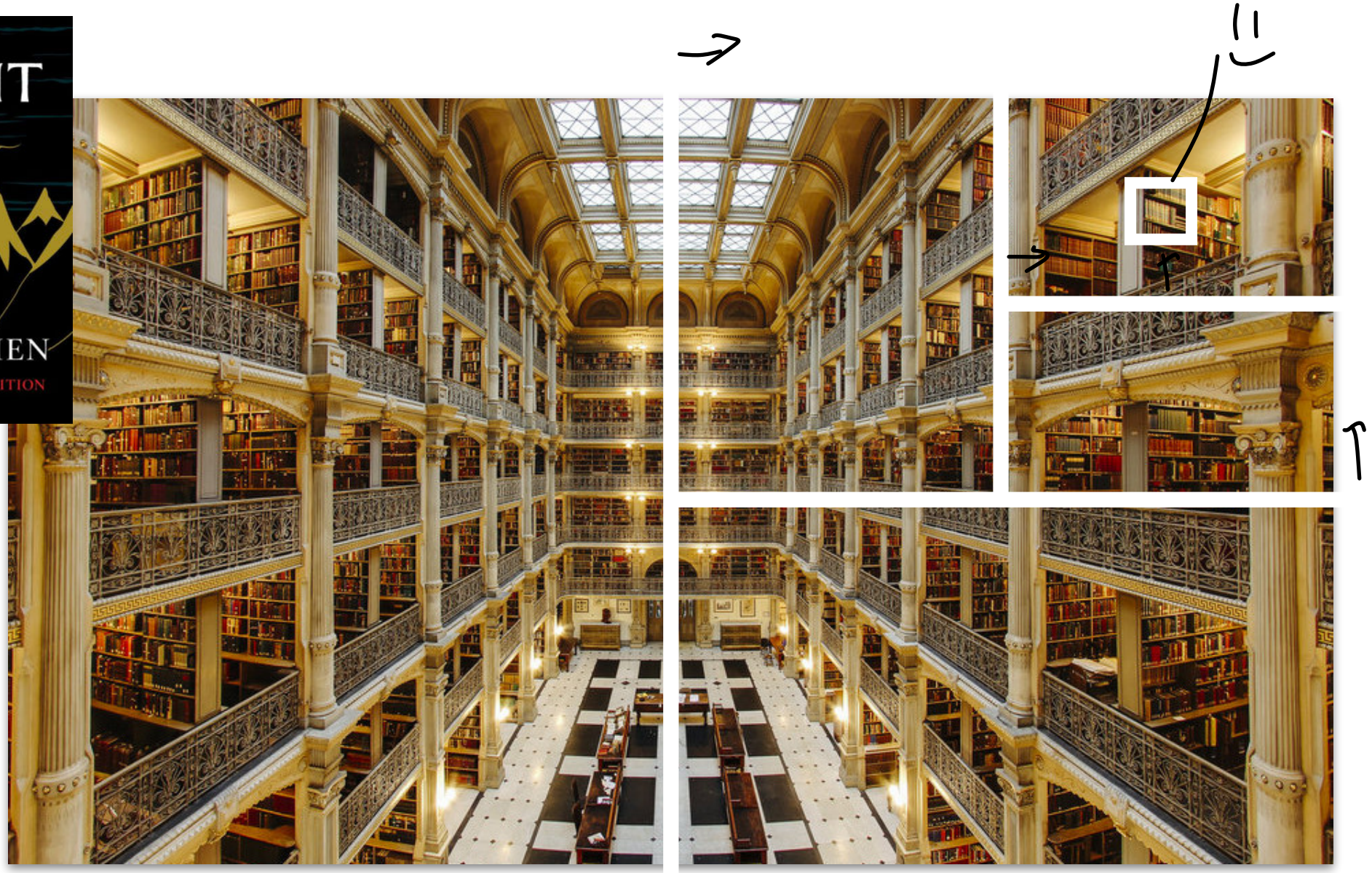
} n 's # books

Unsorted Array *

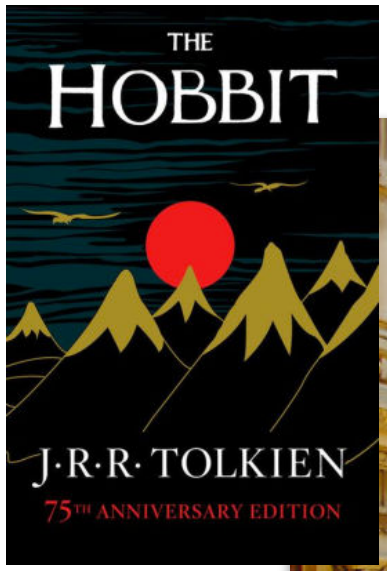
Insert: O(1)* ←
Find: O(n) ←

What if $O(log\ n)$ isn't good enough?





What if $O(log\ n)$ isn't good enough?



O(1)

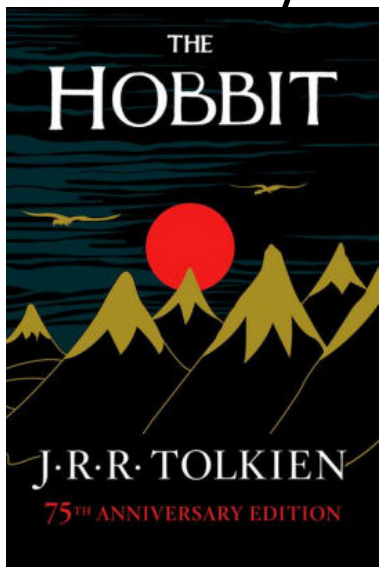
(compute)

O(1)



A Hash Table based Dictionary

key



O(1)

'Black Box'



O(1)

Integer address

ISBN: 9780062265722

Call #: PR
6068.O93
H35 1937

Address

ISBN: 9780062265722

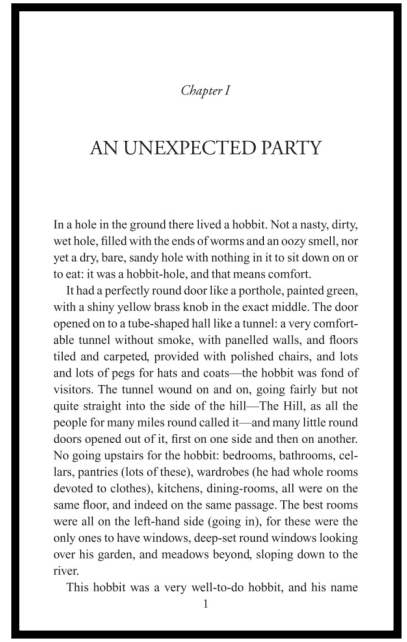
Call #: PR
6068.O93
H35 1937

O(1)



O(1)

value



Chapter I

AN UNEXPECTED PARTY

In a hole in the ground there lived a hobbit. Not a nasty, dirty, wet hole, filled with the ends of worms and an oozy smell, nor yet a dry, bare, sandy hole with nothing in it to sit down on or to eat: it was a hobbit-hole, and that means comfort.

It had a perfectly round door like a porthole, painted green, with a shiny yellow brass knob in the exact middle. The door opened on to a tube-shaped hall like a tunnel: a very comfortable tunnel without smoke, with panelled walls, and floors tiled and carpeted, provided with polished chairs, and lots and lots of pegs for hats and coats—the hobbit was fond of visitors. The tunnel wound on and on, going fairly but not quite straight into the side of the hill—The Hill, as all the people for many miles round called it—and many little round doors opened out of it, first on one side and then on another. No going upstairs for the hobbit: bedrooms, bathrooms, cellars, pantries (lots of these), wardrobes (he had whole rooms devoted to clothes), kitchens, dining-rooms, all were on the same floor, and indeed on the same passage. The best rooms were all on the left-hand side (going in), for these were the only ones to have windows, deep-set round windows looking over his garden, and meadows beyond, sloping down to the river.

This hobbit was a very well-to-do hobbit, and his name

1

Randomized Data Structures

Sometimes a data structure can be **too ordered / too structured**

↳ BBST (AVL) tree : $O(\log n)$ find & insert

No easy way to be better always

Randomized data structures rely on **expected** performance

↳ Goal: expected performance in $O(1)$***

Randomized data structures 'cheat' tradeoffs!

A Hash Table based Dictionary





User Code (is a map):

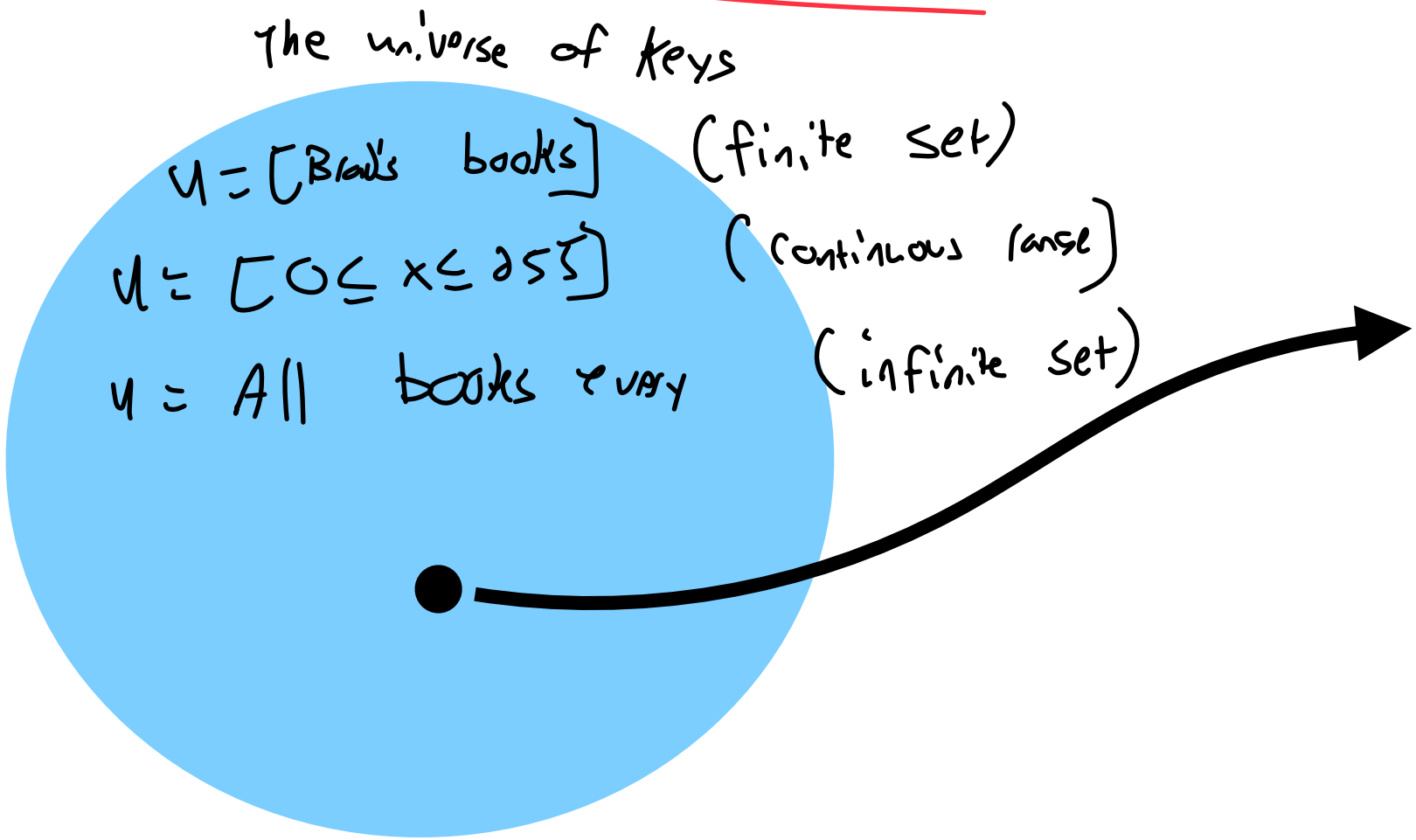
```
1  Dictionary<KeyType, ValueType> d;
2  d[k] = v;
```

A **Hash Table** consists of three things:

1. A hash function (key → int)
   address or index
2. A data storage structure
   80% 1'st ← unsorted array is best
   15% tree ← tree can't get O(1) performance
3. A method of handling "chaos"

Hash Function

Maps a **keyspace**, a (mathematical) description of the keys for a set of data, to a set of integers.



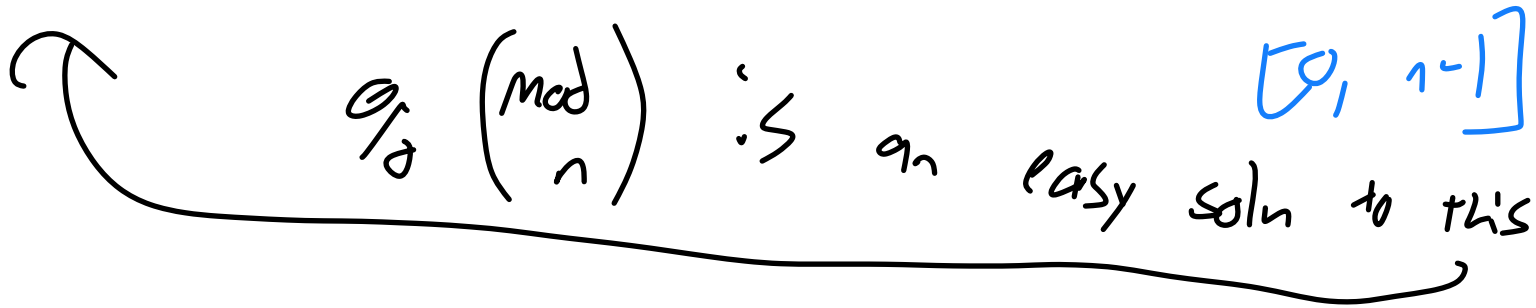
***m* elements**

Key	Value

Hash Function → General purpose is goal in CS 385

A hash function ***must*** be:

- **Deterministic**: $\forall K_1, K_2$ if $(K_1 == K_2)$, $H(K_1) == H(K_2)$
  - ↳ Same input always gives same output
- **Efficient**: $O(1)$
- **Defined for a certain size table**: Table of size n

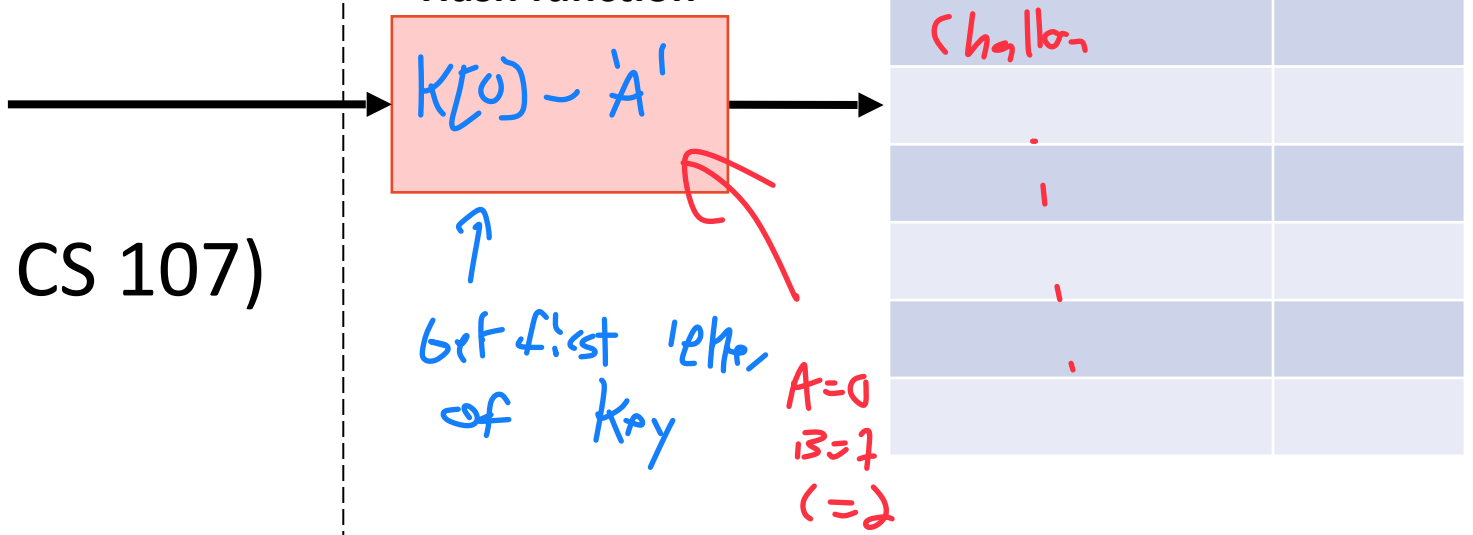


Hash Function

Professor (key) Last name

Value: a class they teach

(Angrave, CS 241)
(Beckman, CS 421)
(Challon, CS 125)
(Davis, CS 101)
(Evans, CS 225)
(Fagen-Ulmschneider, CS 107)
(Gunter, CS 422)
(Herman, CS 233)



Hash Function

Problem 1: Collisions when two keys hash to same address

(Angrave, CS 241)
(Beckman, CS 421)
(Challon, CS 125)
(Davis, CS 101)
(Evans, CS 225)
(Fagen-Ulmschneider, CS 107)
(Gunter, CS 422)
(Herman, CS 233)

Abhi Akwin

Hash function

(key[0] - 'A')

Key	Value
Angrave	241
Beckman	421
Challon	125
Davis	101
Evans	225
Fagen-U	107
Gunter	422
Herman	233

0 1 2 3 4 5 6 7

e.g. 8

Solomon, CS 225

Problem 2

Hash does not handle all inputs w/ current size table

General Hash Function



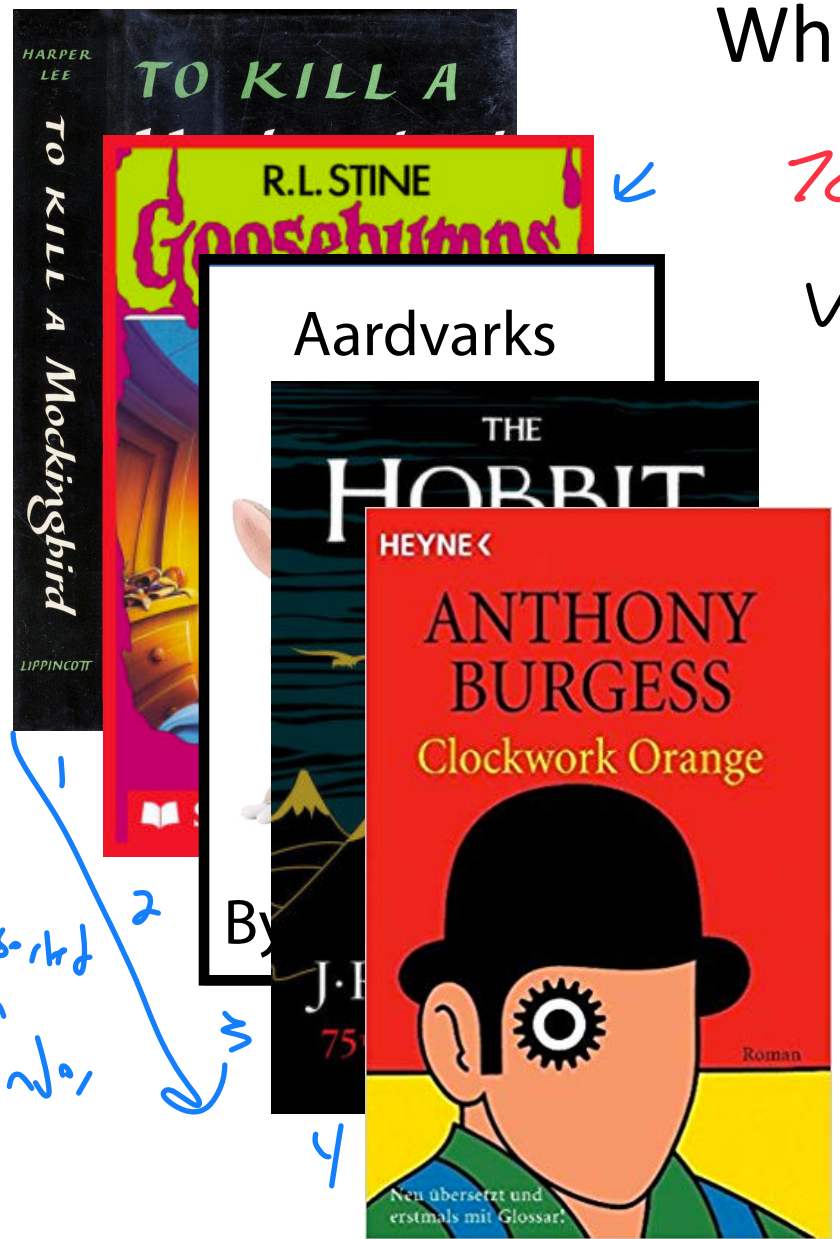
An $O(1)$ deterministic operation that maps all keys in a universe U to a defined range of integers $[0, \ldots, m-1]$

- A **hash**: h(k) → int
- A **compression**: h(k) % m

Choosing a good hash function is tricky…

- Don't create your own (yet*)

Hash Function



0 1 2 3 4 Inserted in order

Which of the following are valid hashes?

first letter of first + last name

70% $h_1(k) = (k.firstName[0] + k.lastName[0]) \% m$ ✓

↳ Deterministic? Yep!
↳ O(1) ↳ Yep!
↳ Fit in size m table ↳ Yes
↳ Any Nulls? → Debatable

# of pages mult by random #

30% $h_2(k) = (rand() * k.numPages) \% m$ ✗

↳ Not deterministic
↳ O(1) ↳ Yes
↳ All books ↳ fit in tbl ↳ Yep

The order I see a book is it's index

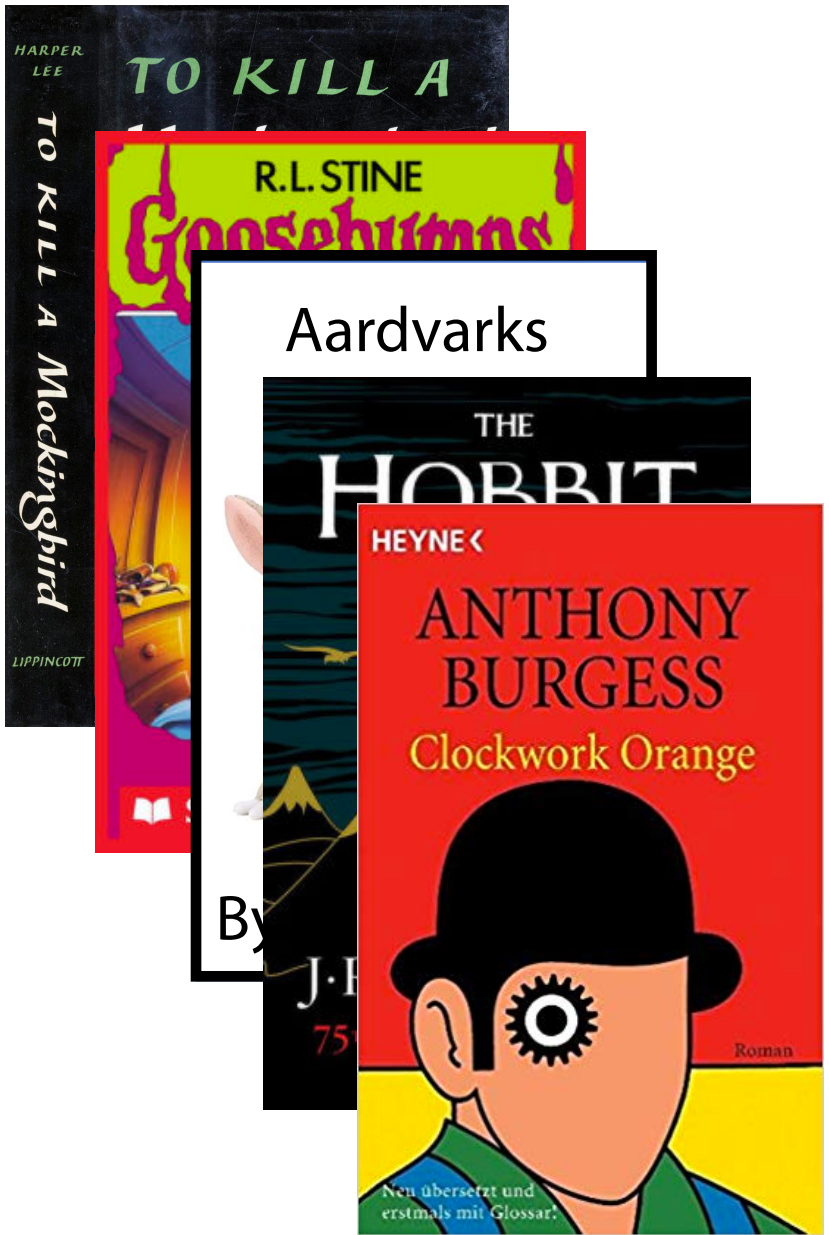
50% $h_3(k) = (\text{return Count++;}) \% m$ ✓

↳ O(1)
↳ All books ↳ fit in tbl

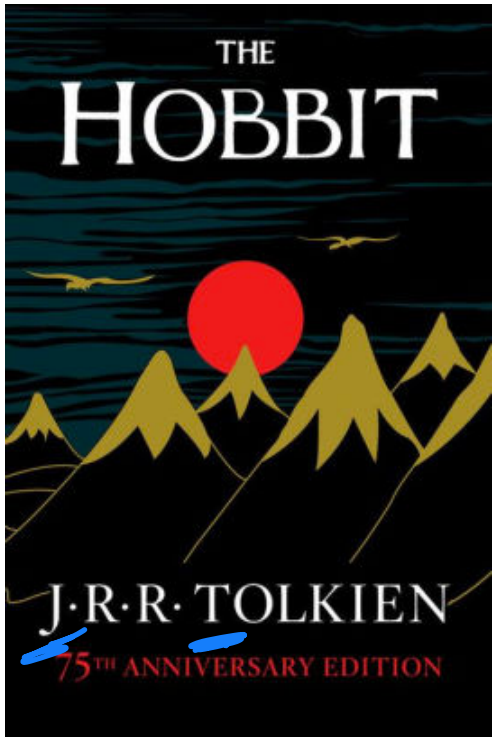
Lookup Goosebumps

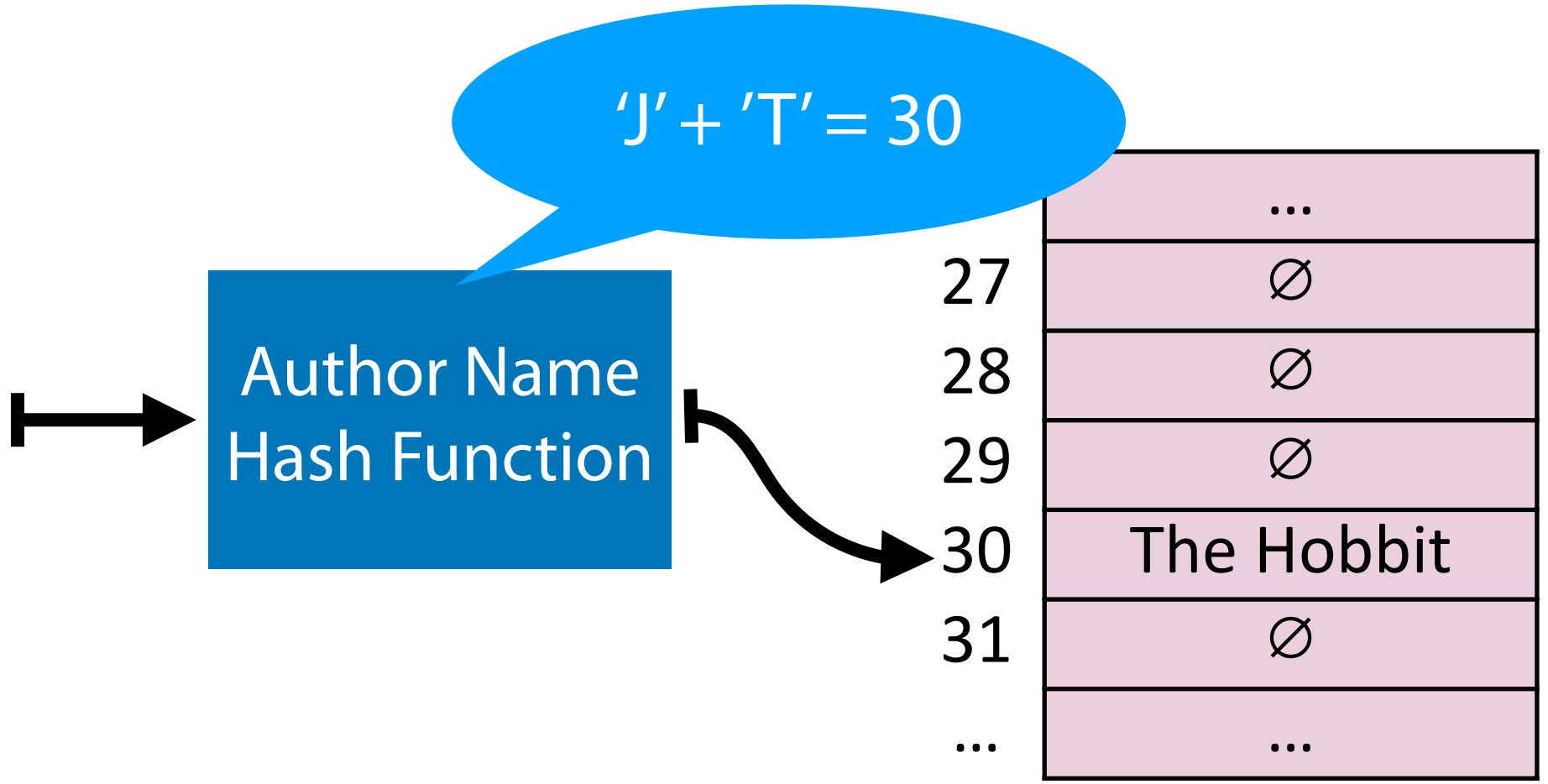
5

Hash Function

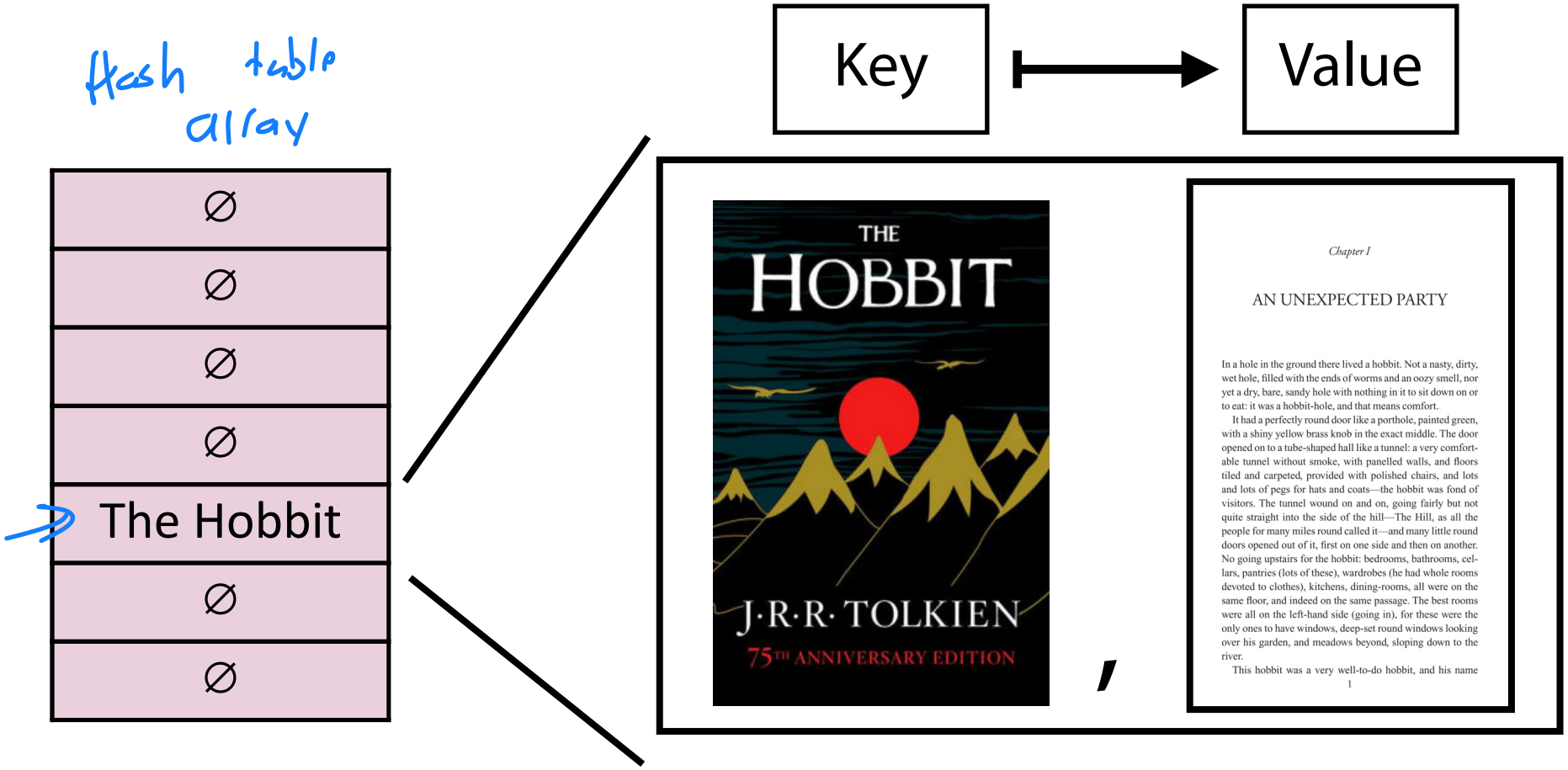


A Hash Table based Dictionary

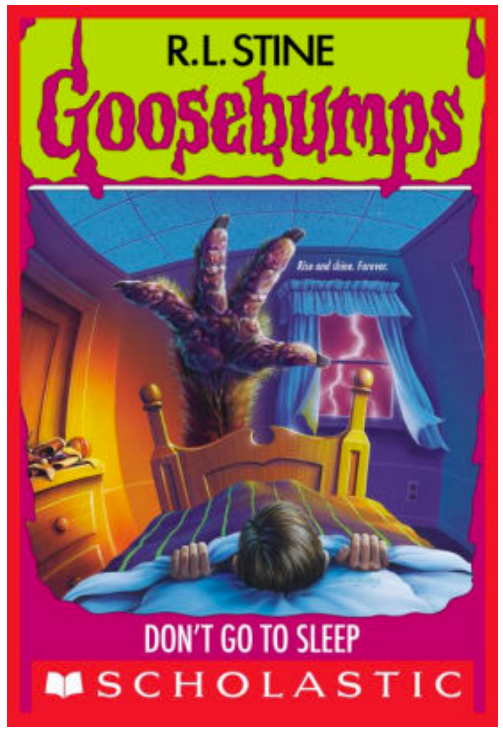




A Hash Table based Dictionary



A Hash Table based Dictionary

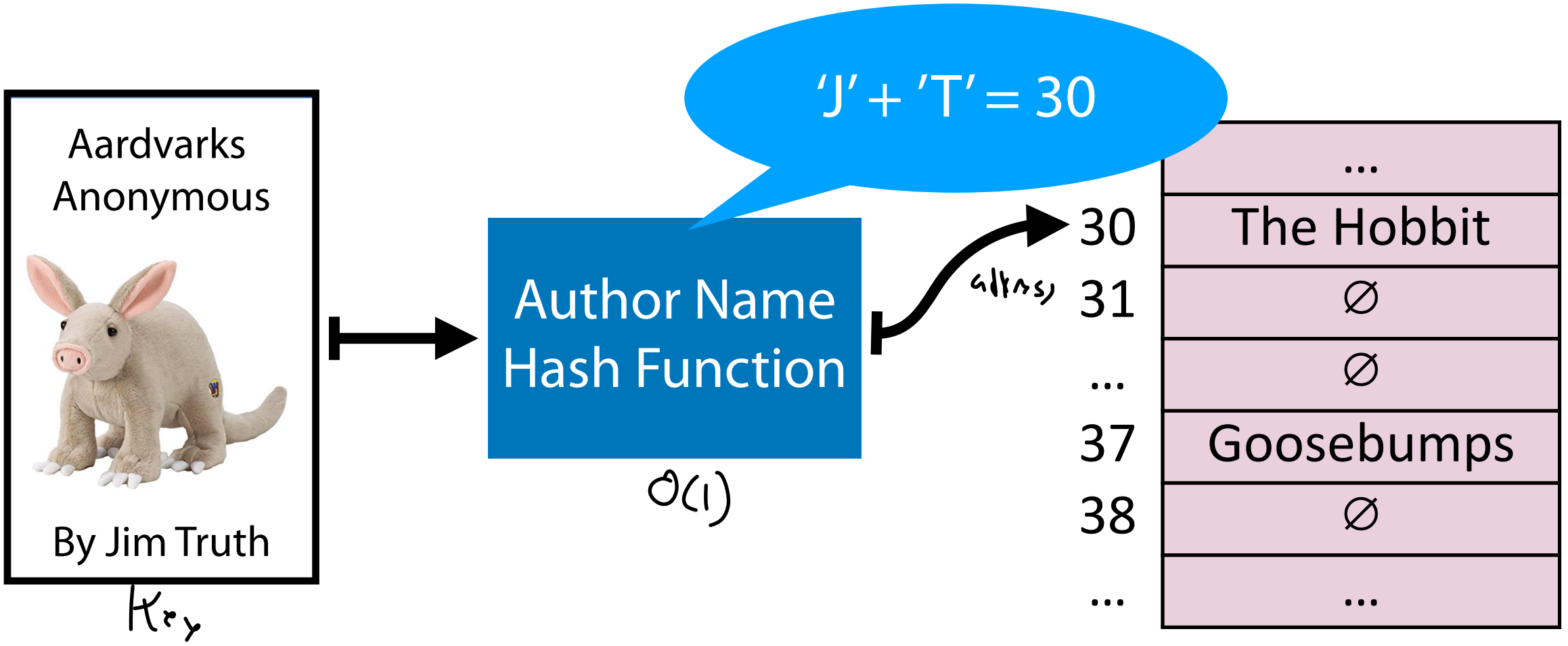


'R' + 'S' = 37

Author Name Hash Function

…	
30	The Hobbit
31	∅
…	∅
37	Goosebumps
38	∅
…	…

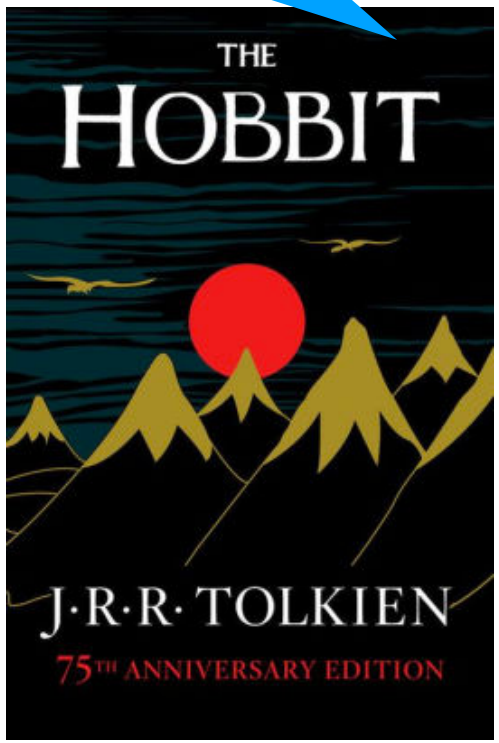
A Hash Table based Dictionary



Hash Collision

A ***hash collision*** occurs when multiple unique keys hash to the same value

J.R.R Tolkien = 30!



Jim Truth = 30!

Aardvarks Anonymous

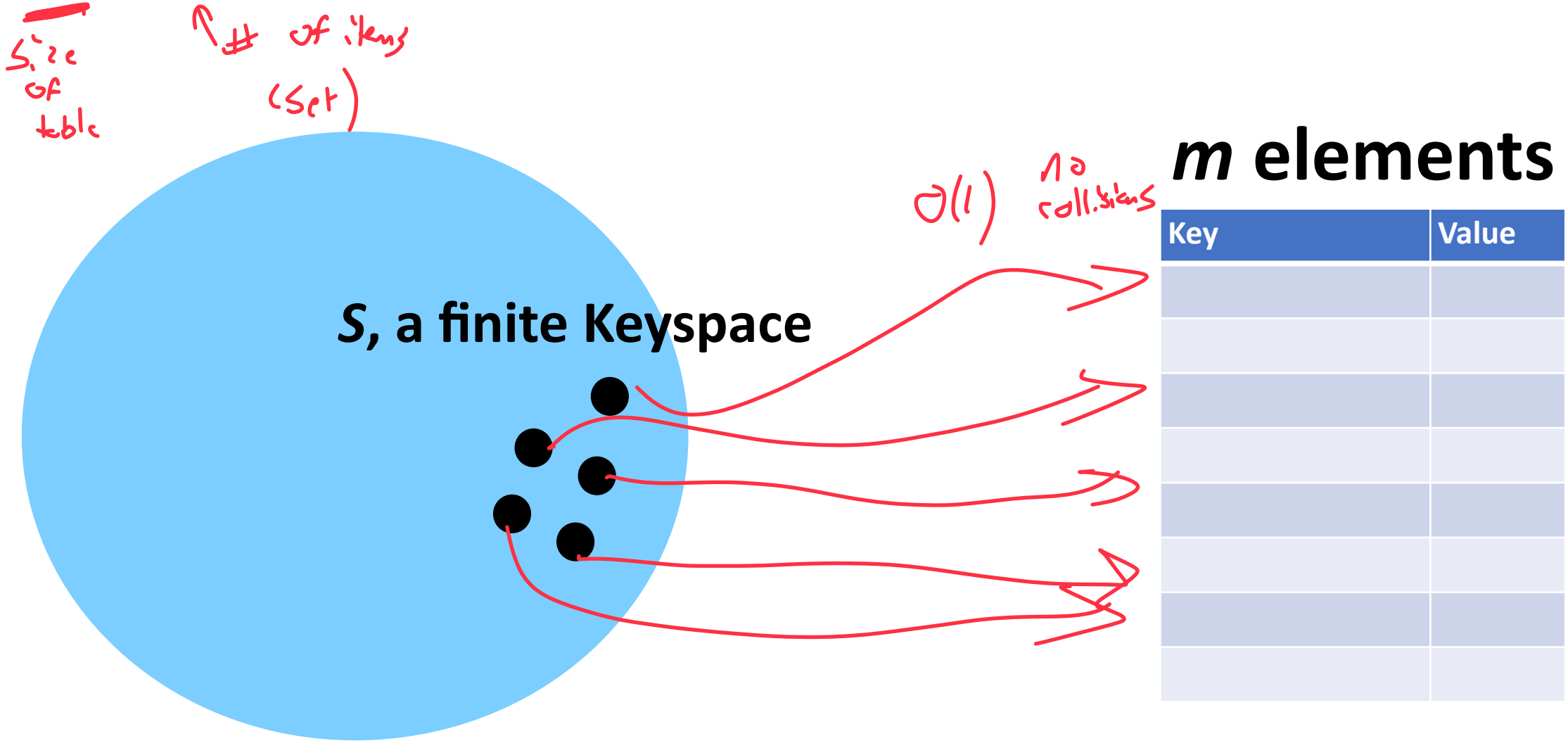


By Jim Truth

…	…
30	???
31	∅
…	∅
37	Goosebumps
38	∅
…	…

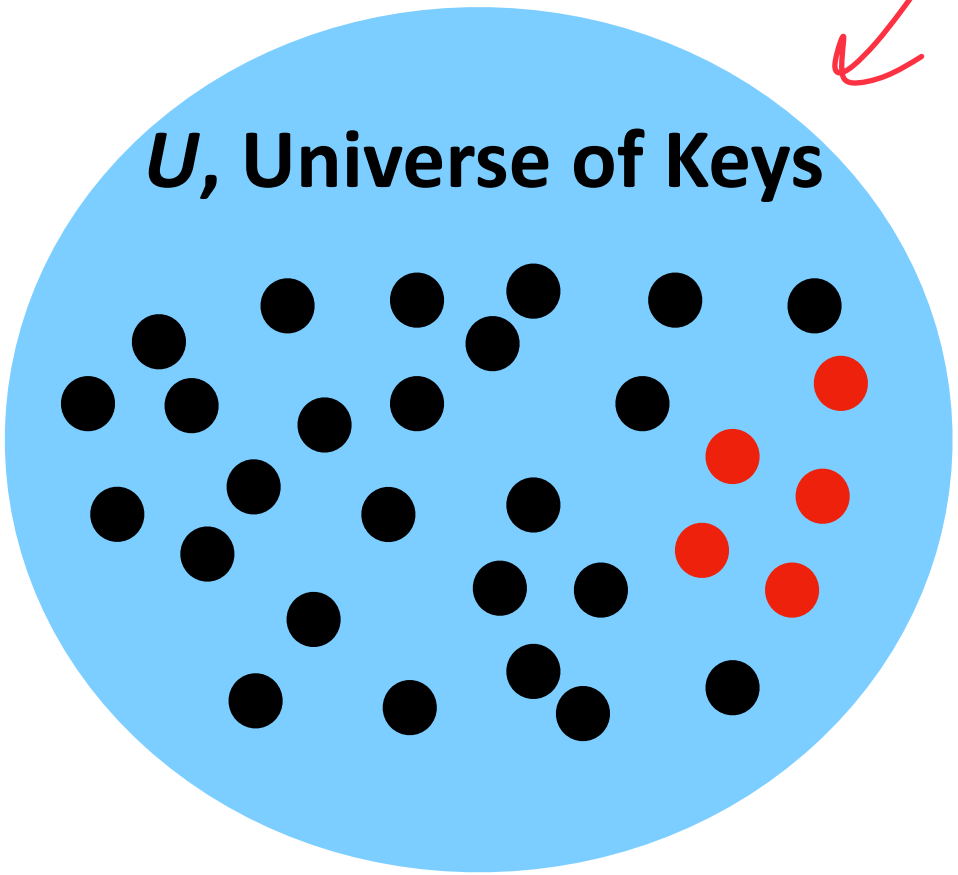
Perfect Hashing

If $m \geq S$, we can write a *perfect* hash with no collisions



General Purpose Hashing

In CS 225, we want our hash functions to work *in general*.

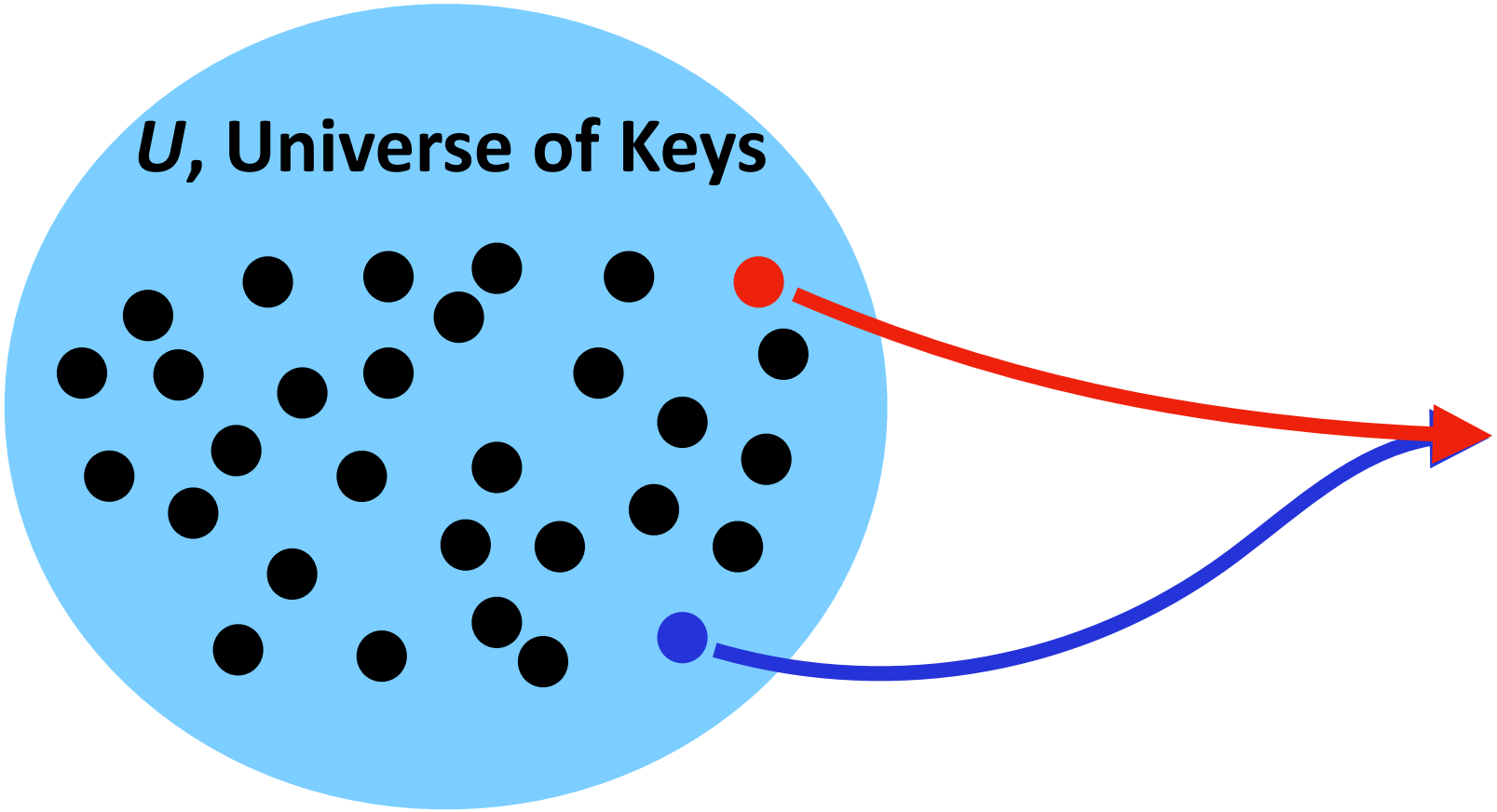


***m* elements**

Key	Value

General Purpose Hashing

If $m < U$, there must be at least one hash collision.

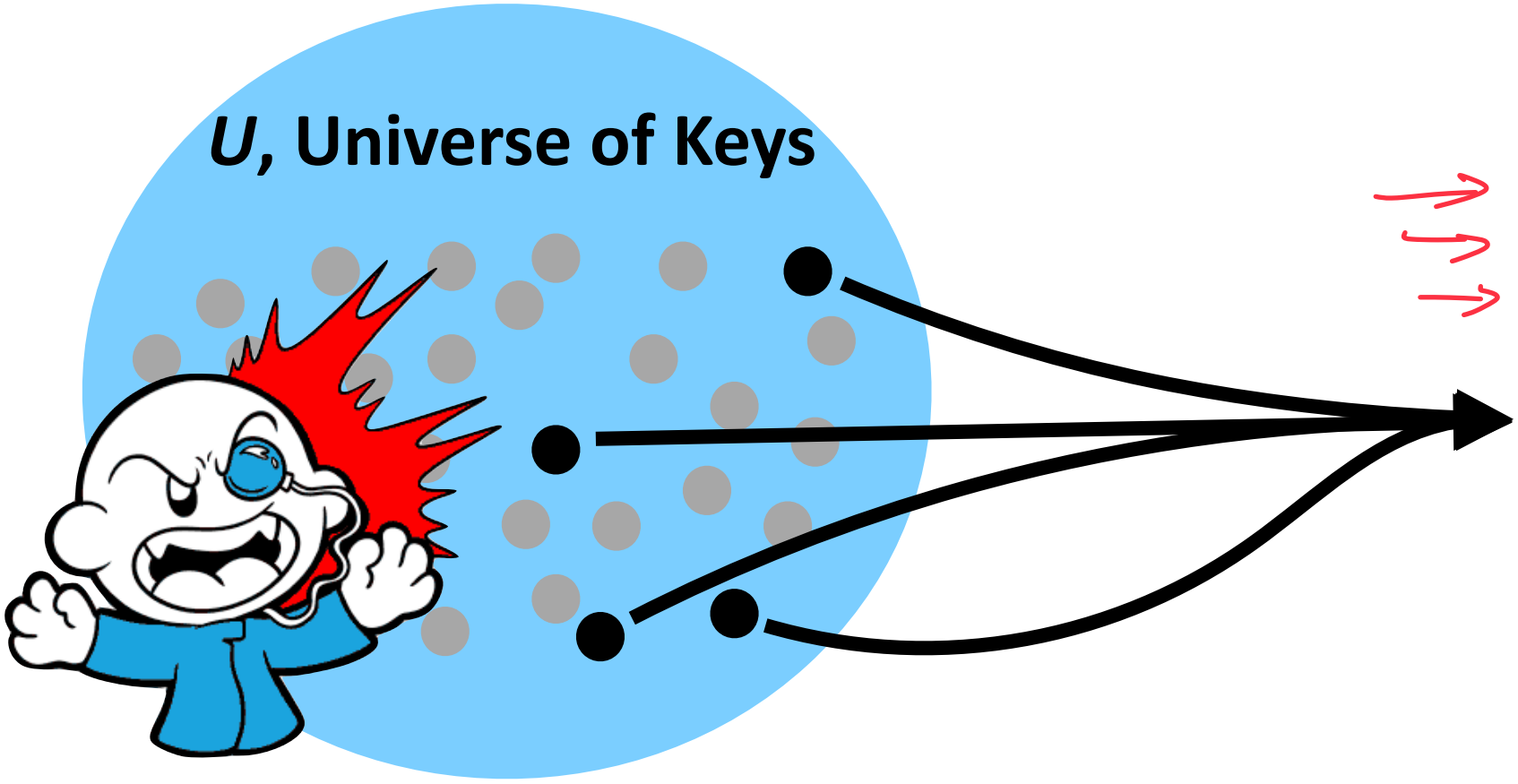


m elements

Key	Value

General Purpose Hashing

By fixing h, we open ourselves up to adversarial attacks.



***m* elements**

Key	Value
???	

Image by Matthew Loffhagen

A Hash Table based Dictionary



User Code (is a map):

```
1  Dictionary<KeyType, ValueType> d;
2  d[k] = v;
```

A **Hash Table** consists of three things:

1. A hash function

2. A data storage structure

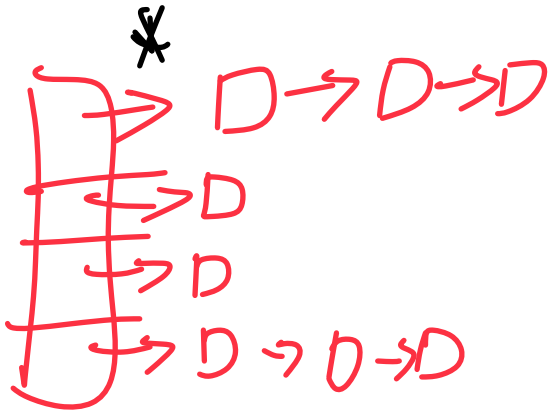
3. **A method of addressing *hash collisions***

Some prob of colliding where two keys hash to same address

Open vs Closed Hashing

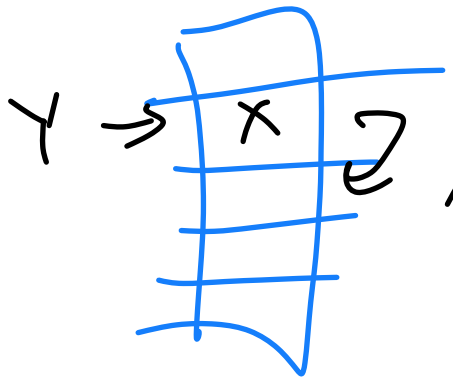
Addressing hash collisions depends on your storage structure.

- **Open Hashing:**



All K, V pairs stored externally

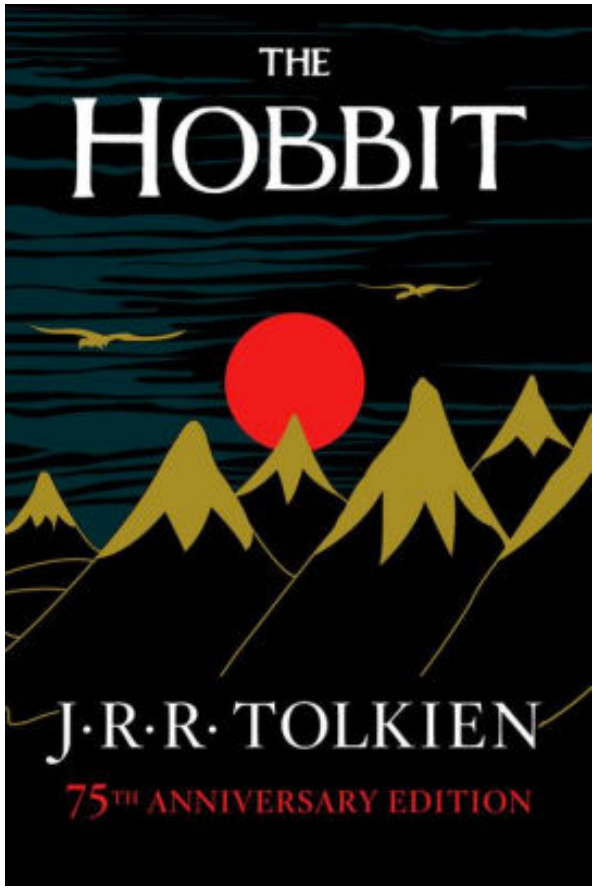
- **Closed Hashing:**

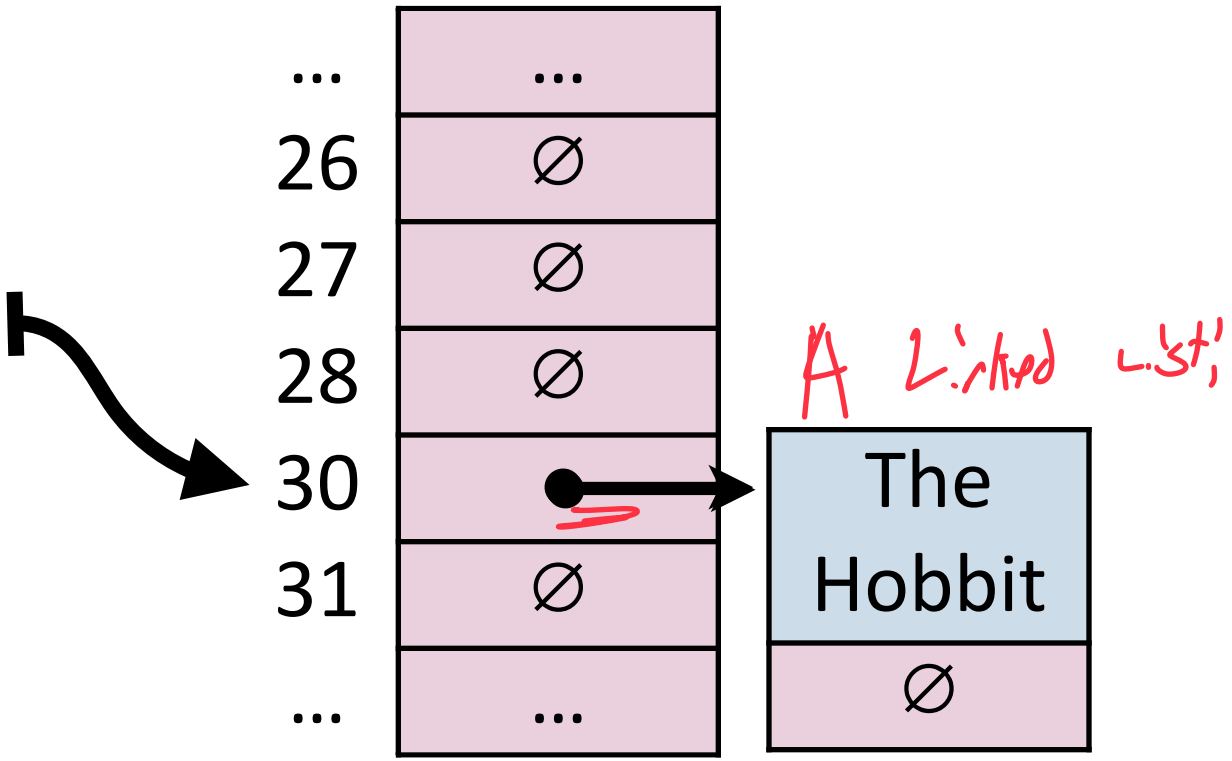


All K,V pairs stored in array

Open Hashing

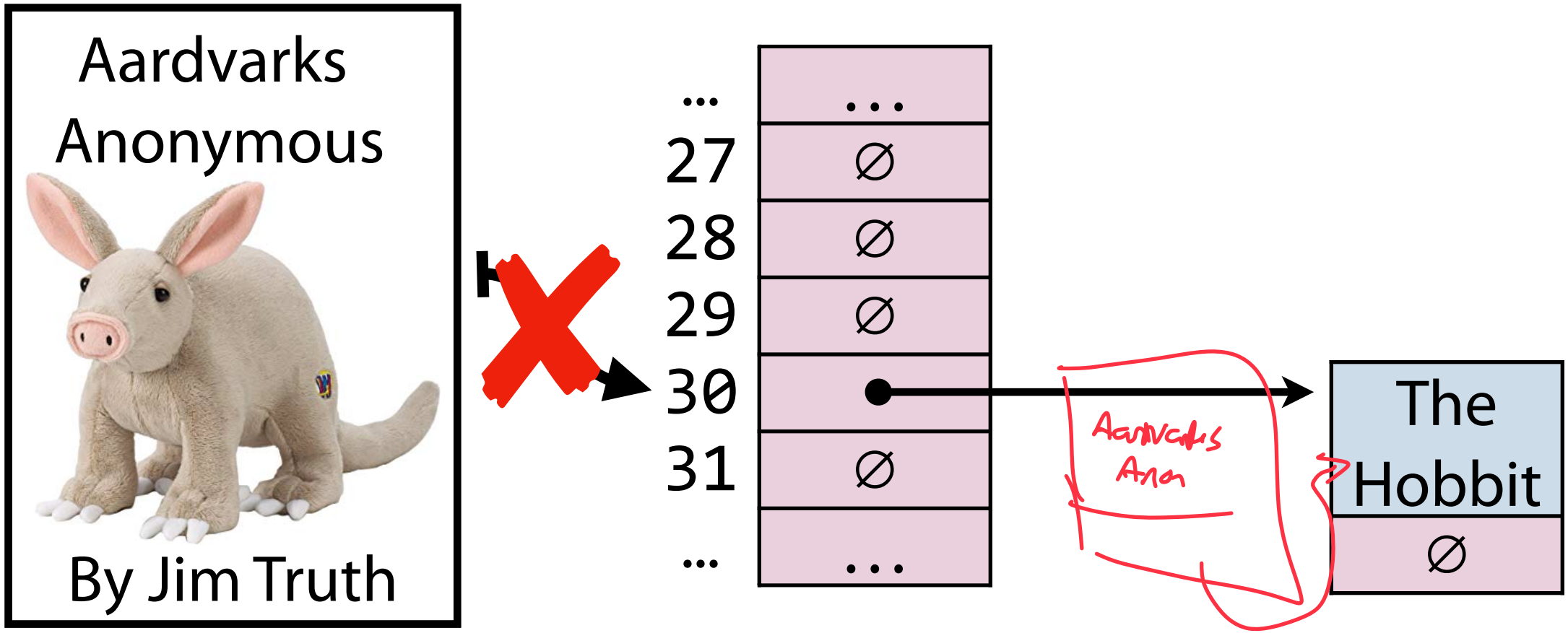
In an ***open hashing*** scheme, key-value pairs are stored externally (for example as a linked list).





Hash Collisions (Open Hashing)

A ***hash collision*** in an open hashing scheme can be resolved by adding it to the LL. This is called ***separate chaining***.



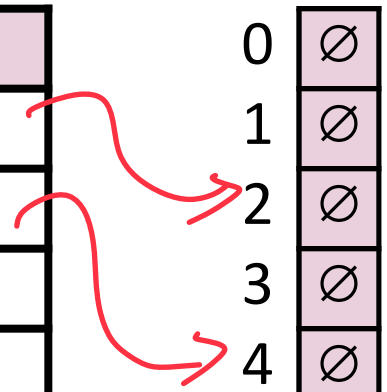
Insertion (Separate Chaining)

```
_insert("Bob")

_insert("Anna")
```

Address

Key	Value	Hash
Bob	**B+**	**2**
Anna	**A-**	**4**
Alice	A+	4
Betty	B	2
Brett	A-	2
Greg	A	0
Sue	B	7
Ali	B+	4
Laura	A	7
Lily	B+	7

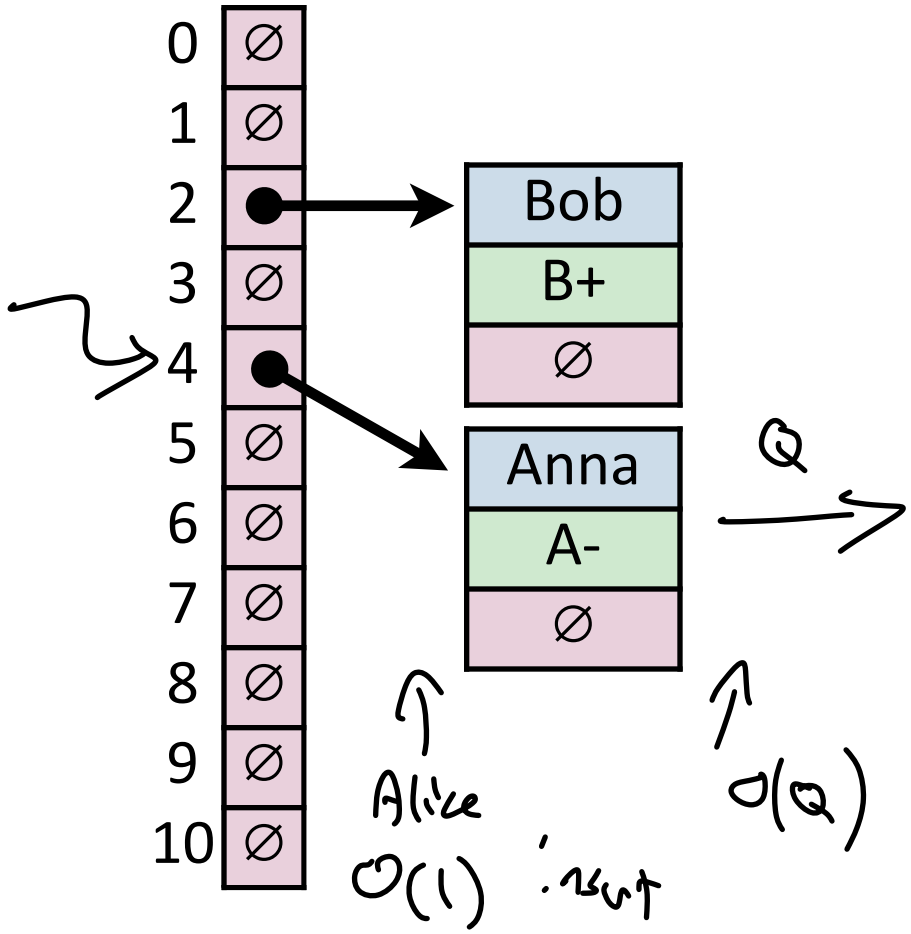


0	∅
1	∅
2	∅
3	∅
4	∅
5	∅
6	∅
7	∅
8	∅
9	∅
10	∅

Insertion (Separate Chaining) `_insert("Alice")`

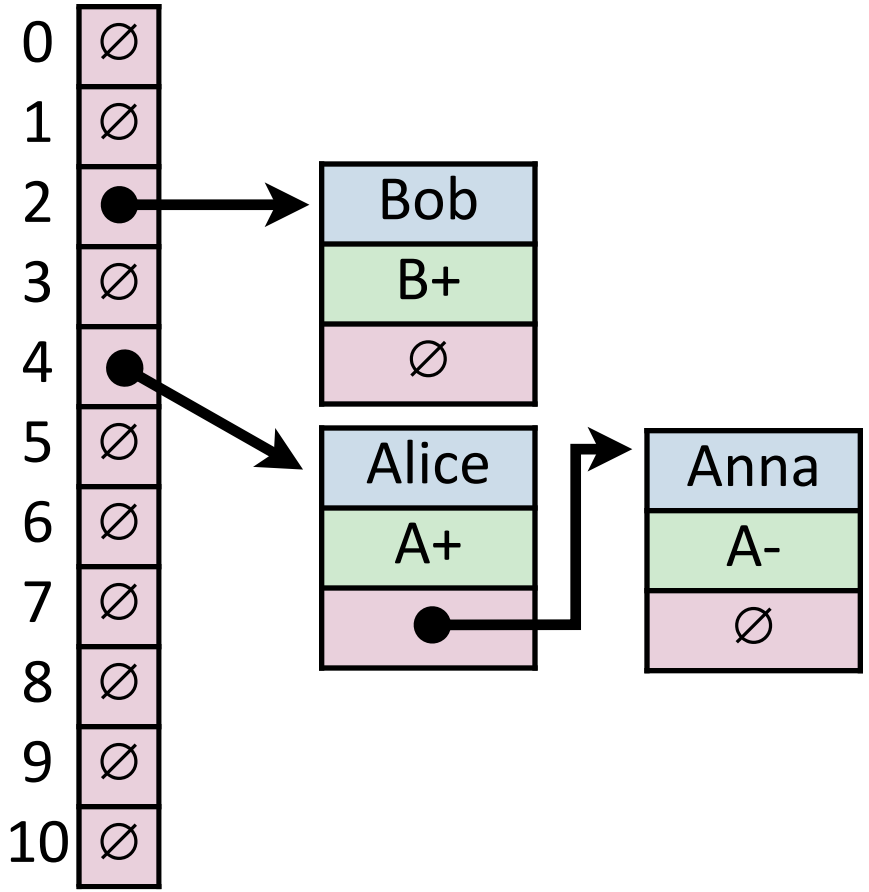
Where does Alice end up in the hash table?

Key	Value	Hash
Bob	B+	2
Anna	A-	4
Alice	**A+**	**4**
Betty	B	2
Brett	A-	2
Greg	A	0
Sue	B	7
Ali	B+	4
Laura	A	7
Lily	B+	7



Insertion (Separate Chaining)

Key	Value	Hash
Bob	B+	2
Anna	A-	4
Alice	A+	4
Betty	**B**	**2**
Brett	A-	2
Greg	A	0
Sue	B	7
Ali	B+	4
Laura	A	7
Lily	B+	7



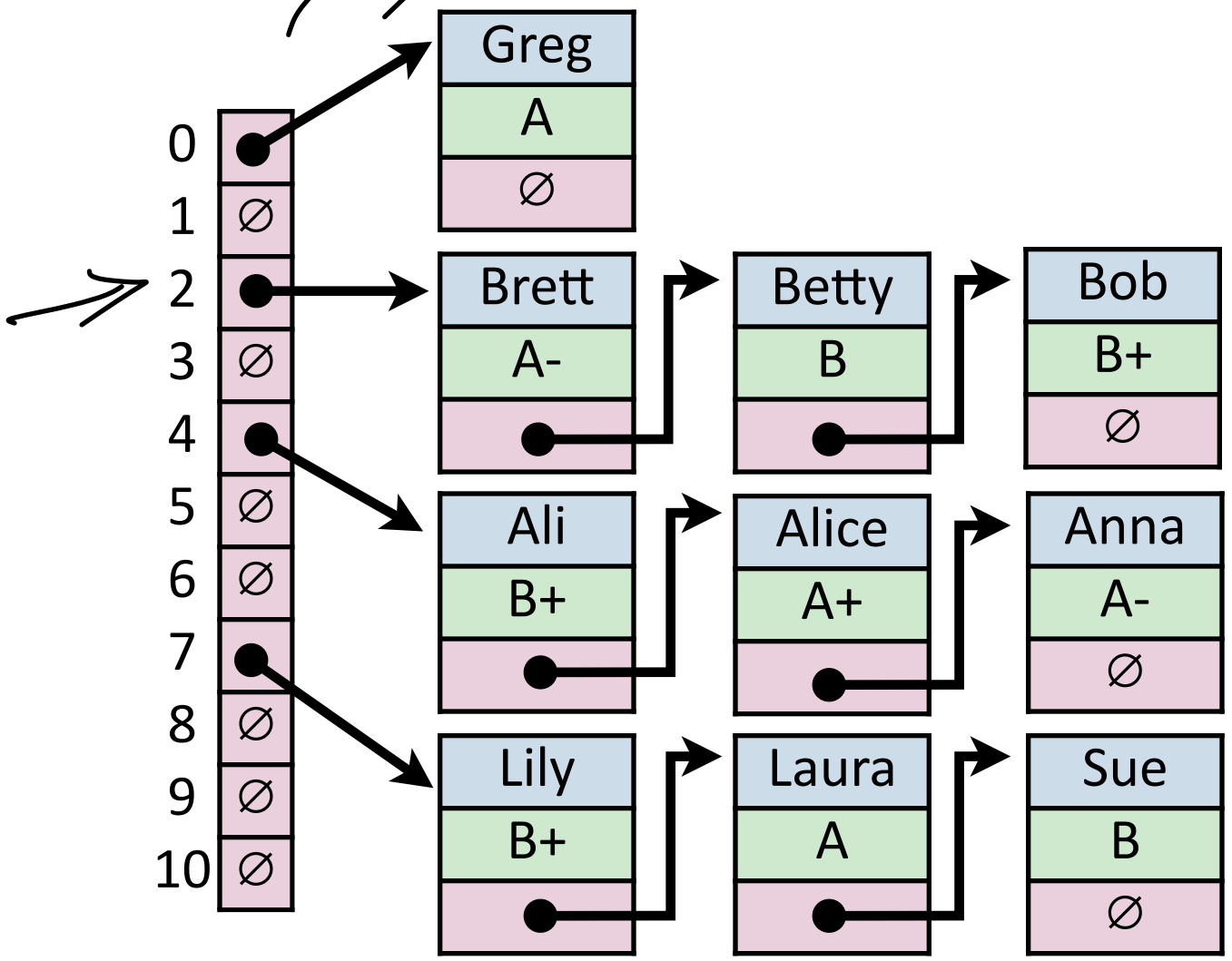
Insertion (Separate Chaining)

O(1)

O(1)

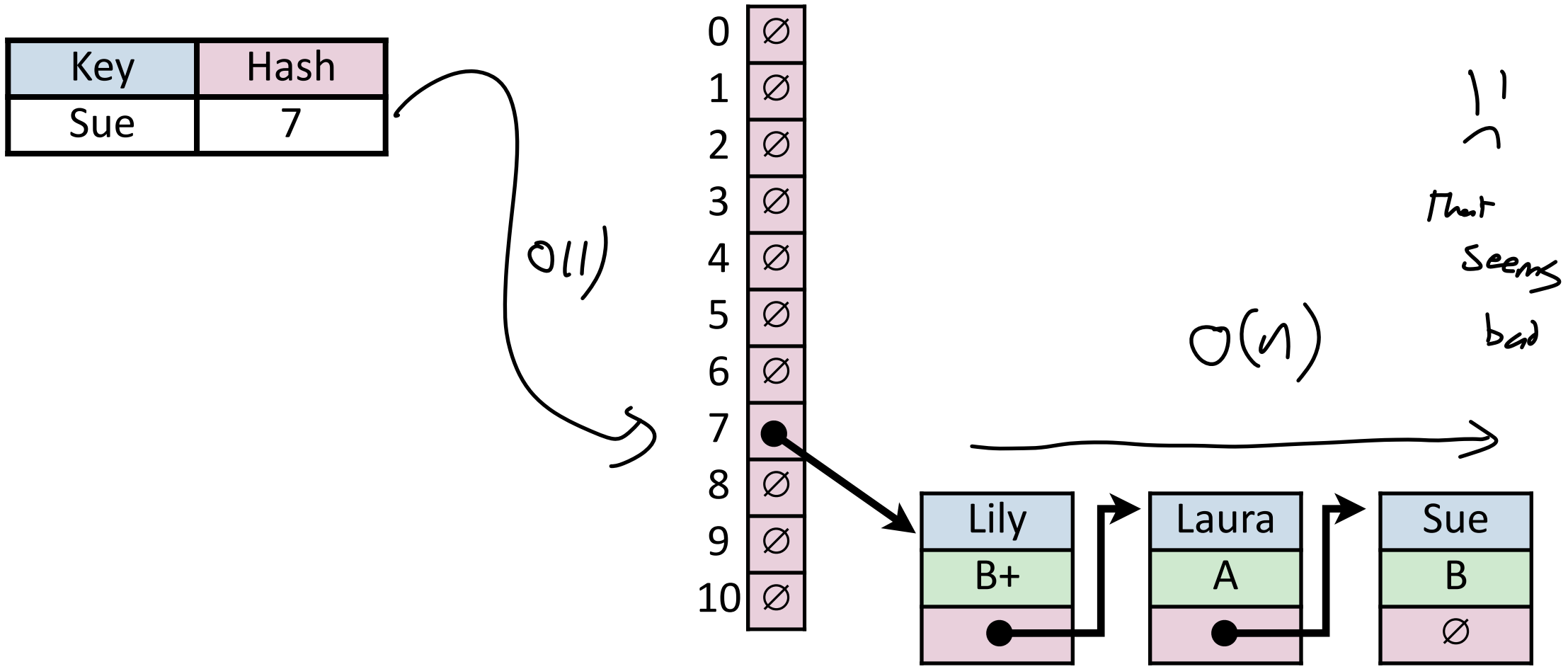
O(1)

Key	Value	Hash
Bob	B+	2
Anna	A-	4
Alice	A+	4
Betty	B	2
Brett	A-	2
Greg	A	0
Sue	B	7
Ali	B+	4
Laura	A	7
Lily	B+	7



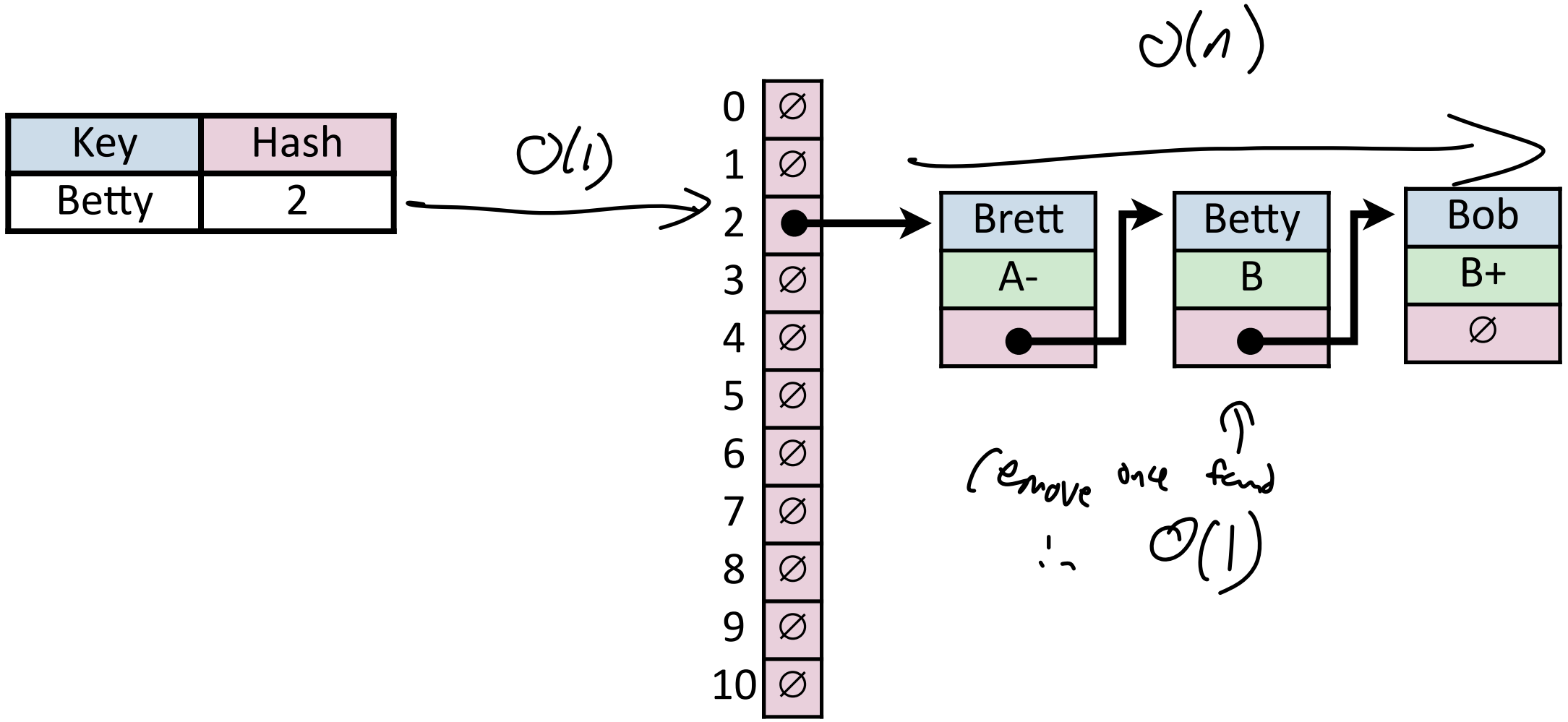
Find (Separate Chaining)

```
_find("Sue")
```



Remove (Separate Chaining)

`_remove("Betty")`



Hash Table (Separate Chaining)



For hash table of size *m* and *n* elements:

Find runs in: $O(n)$

Because collisions

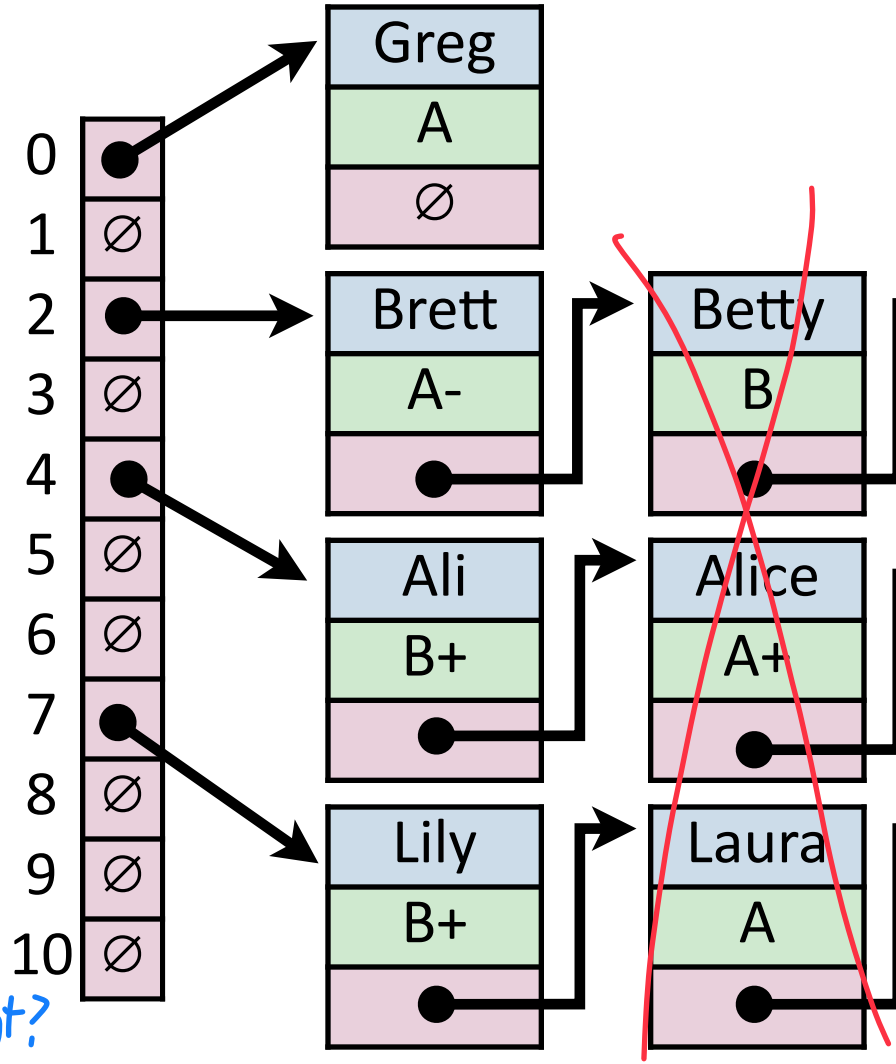
Insert runs in: $O(1)$

Maybe we can bound collision freq in some way

Remove runs in: $O(n)$

what can we do?

can we make this a constant?



Hash Table

Worst-Case behavior is bad — but what about randomness?

1) **Fix *h***, our hash, and assume it is good for ***all keys***:

2) Create a ***universal hash function family:***

Simple Uniform Hashing Assumption

Given table of size m, a simple uniform hash, h, implies

$$\forall k_1, k_2 \in U \text{ where } k_1 \neq k_2 \text{ , } Pr(h[k_1] = h[k_2]) = \frac{1}{m}$$

Uniform:

Independent:

Separate Chaining Under SUHA

Table Size: m

Num objects: n

Claim: Under SUHA, expected length of chain is $\frac{n}{m}$

$\alpha_j =$ expected # of items hashing to position j

$$\alpha_j = \sum_i H_{i,j}$$

$$H_{i,j} = \begin{cases} 1 \text{ if item i hashes to j} \\ 0 \text{ otherwise} \end{cases}$$

Separate Chaining Under SUHA

Table Size: m

Num objects: n

Claim: Under SUHA, expected length of chain is $\frac{n}{m}$

$\alpha_j =$ expected # of items hashing to position j

$$\alpha_j = \sum_i H_{i,j}$$

$$H_{i,j} = \begin{cases} 1 \text{ if item i hashes to j} \\ 0 \text{ otherwise} \end{cases}$$

$$E[\alpha_j] = E\left[\sum_i H_{i,j}\right]$$

Separate Chaining Under SUHA

Table Size: m

Num objects: n

Claim: Under SUHA, expected length of chain is $\frac{n}{m}$

$\alpha_j =$ expected # of items hashing to position j

$$\alpha_j = \sum_i H_{i,j}$$

$$H_{i,j} = \begin{cases} 1 \text{ if item i hashes to j} \\ 0 \text{ otherwise} \end{cases}$$

$$E[\alpha_j] = E[\sum_i H_{i,j}]$$

$$E[\alpha_j] = \sum_i Pr(H_{i,j} = 1) * 1 + Pr(H_{i,j} = 0) * 0$$

Separate Chaining Under SUHA

Table Size: m

Num objects: n

Claim: Under SUHA, expected length of chain is $\frac{n}{m}$

$\alpha_j =$ expected # of items hashing to position j

$$\alpha_j = \sum_i H_{i,j}$$

$$H_{i,j} = \begin{cases} 1 \text{ if item i hashes to j} \\ 0 \text{ otherwise} \end{cases}$$

$$E[\alpha_j] = E[\sum_i H_{i,j}]$$

$$E[\alpha_j] = \sum_i Pr(H_{i,j} = 1) * 1 + Pr(H_{i,j} = 0) * 0$$

$$E[\alpha_j] = n * Pr(H_{i,j} = 1)$$

Separate Chaining Under SUHA

Table Size: m

Num objects: n

Claim: Under SUHA, expected length of chain is $\frac{n}{m}$

$\alpha_j =$ expected # of items hashing to position j

$$\alpha_j = \sum_i H_{i,j}$$

$$H_{i,j} = \begin{cases} 1 \text{ if item i hashes to j} \\ 0 \text{ otherwise} \end{cases}$$

$$E[\alpha_j] = E[\sum_i H_{i,j}]$$

$$Pr[H_{i,j} = 1] = \frac{1}{m}$$

$$E[\alpha_j] = n * Pr(H_{i,j} = 1)$$

Separate Chaining Under SUHA



Claim: Under SUHA, expected length of chain is $\frac{n}{m}$

Table Size: m

Num objects: n

α_j = expected # of items hashing to position j

$$\alpha_j = \sum_i H_{i,j}$$

$$H_{i,j} = \begin{cases} 1 \text{ if item i hashes to j} \\ 0 \text{ otherwise} \end{cases}$$

$$E[\alpha_j] = E[\sum_i H_{i,j}]$$

$$Pr[H_{i,j} = 1] = \frac{1}{m}$$

$$E[\alpha_j] = n * Pr(H_{i,j} = 1)$$

$$\mathbf{E[\alpha_j] = \frac{n}{m}}$$

Separate Chaining Under SUHA

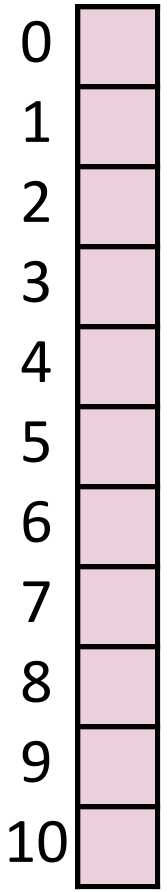


Under SUHA, a hash table of size *m* and *n* elements:

Find runs in: _____.

Insert runs in: _____.

Remove runs in: _____.



Separate Chaining Under SUHA



Pros:

Cons:

Next time: Closed Hashing

Closed Hashing: store *k,v* pairs in the hash table

S = { 1, 8 , 15}

h(k) = k % 7

