

Late class start: Enjoy the solar eclipse!

Class begins at 2:15 PM.



Algorithms and Data Structures for Data Science

Hashing

CS 277

Brad Solomon

April 8, 2024



UNIVERSITY OF
ILLINOIS
URBANA - CHAMPAIGN

Department of Computer Science

Learning Objectives

Motivate and define a hash table

Discuss what a 'good' hash function looks like

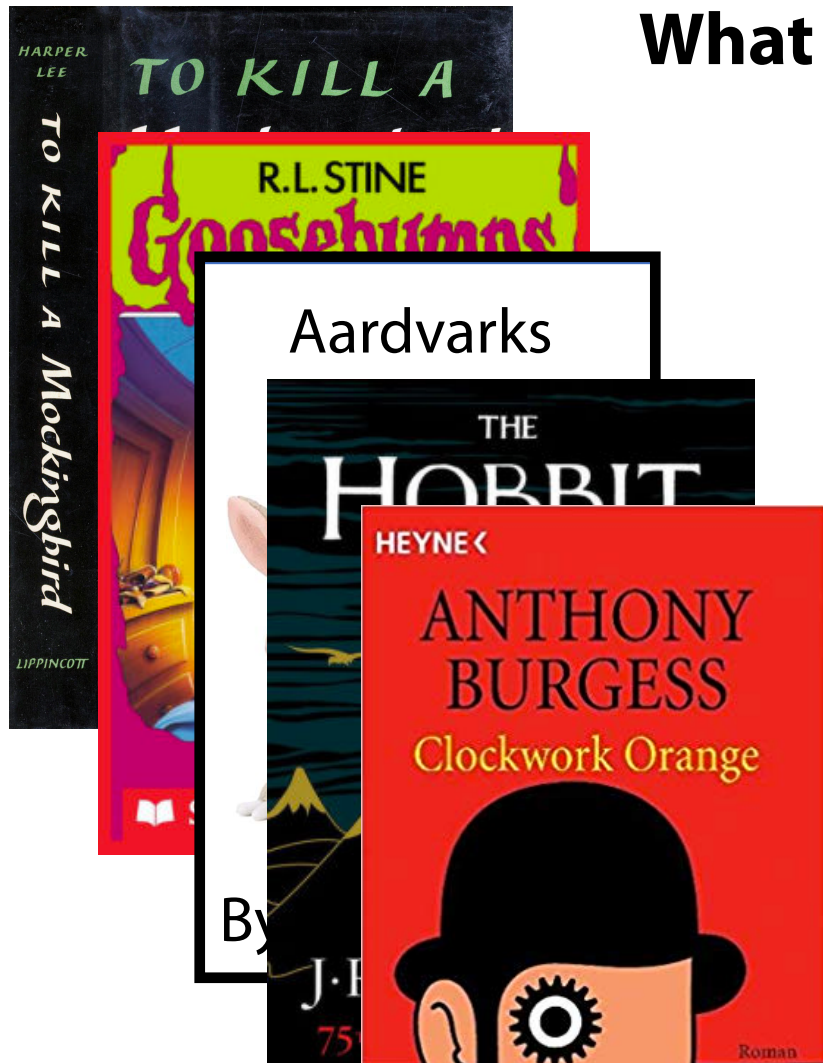
Identify a key weakness of the hash table

Introduce strategies to 'correct' this weakness

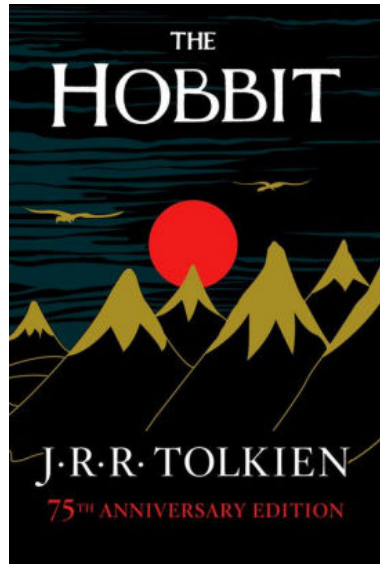
Data Structure Review

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

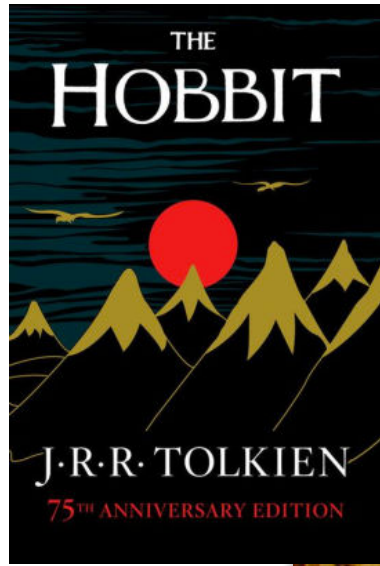
What data structures can I use here?



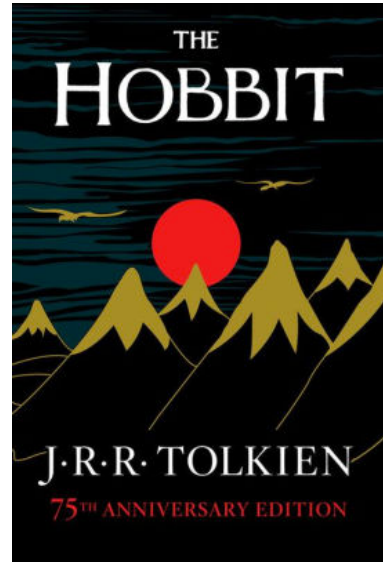
If we recognize that libraries are ordered: $O(\log n)$



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

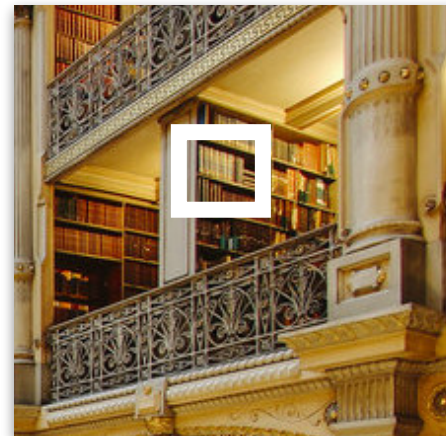


A Hash Table based Dictionary



ISBN: 9781526602381
Call #: PR
6068.093
H35 1998

ISBN: 9781526602381
Call #: PR
6068.093
H35 1998



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

A Hash Table based Dictionary



```
1 d = {}  
2 d[k] = v  
3 print(d[k])
```

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

1.

2.

3.

Hash Function

A hash function *must* be:

- **Deterministic:**
- **Efficient:**
- **Defined for a certain size table:**

Hash Function

(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

(key[0] - 'A')

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



General Hash Function

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

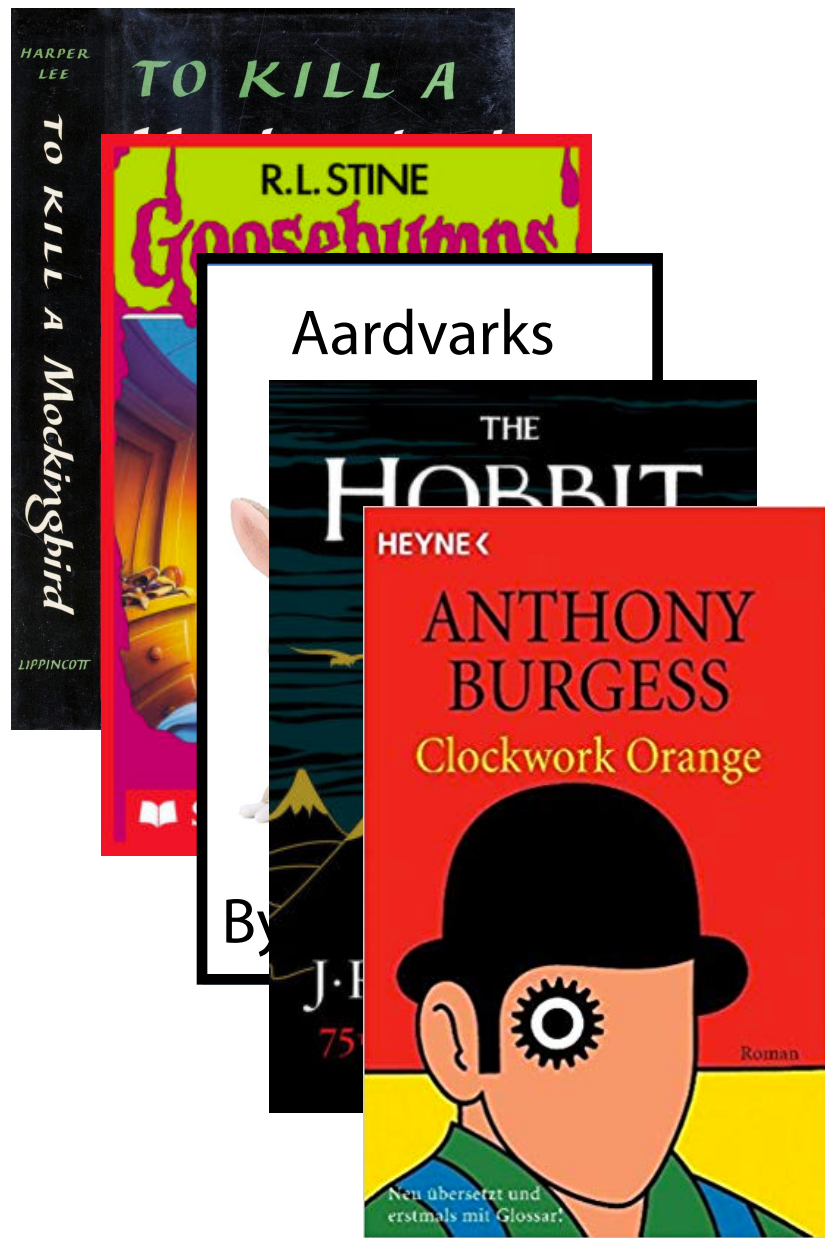
- A **hash**:

- A **compression**:

Choosing a good hash function is tricky...

- Don't create your own!

Hash Function

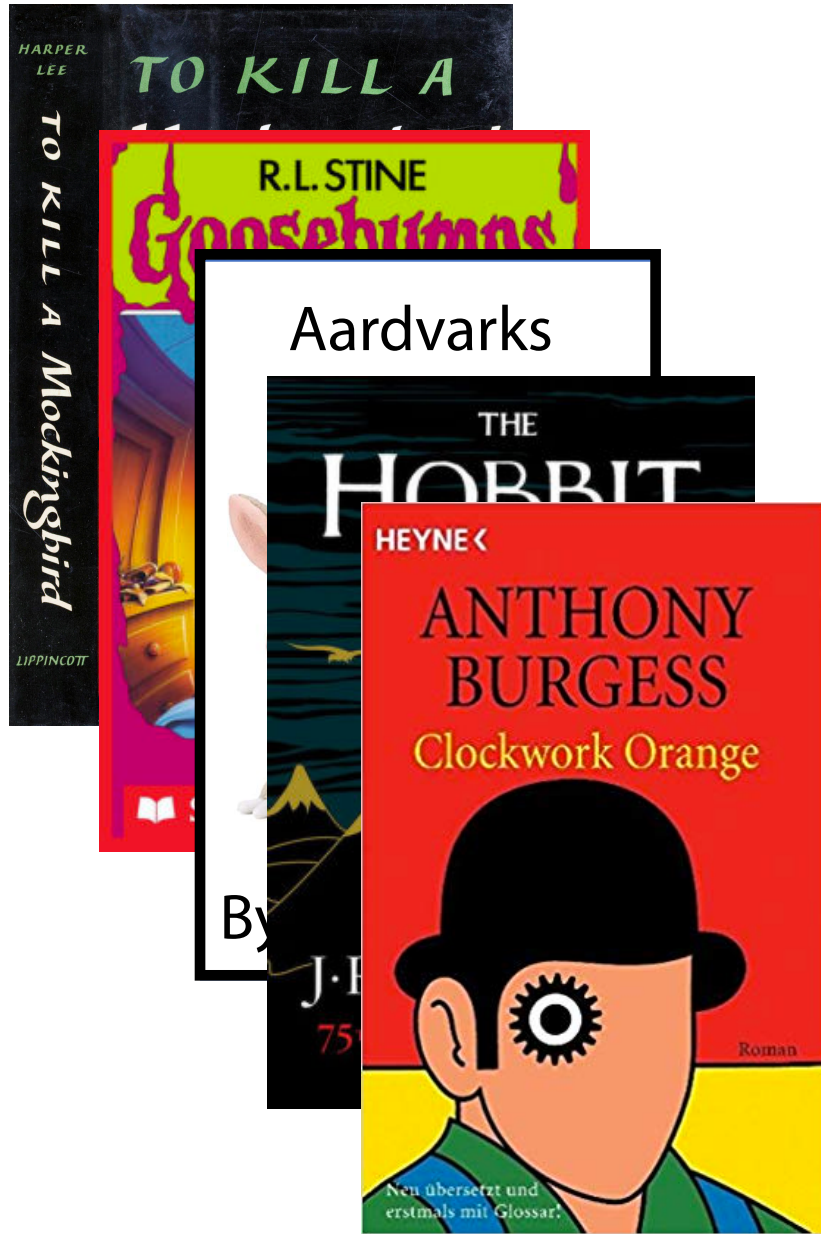


$$h(k) = (k.firstName[0] + k.lastName[0]) \% m$$

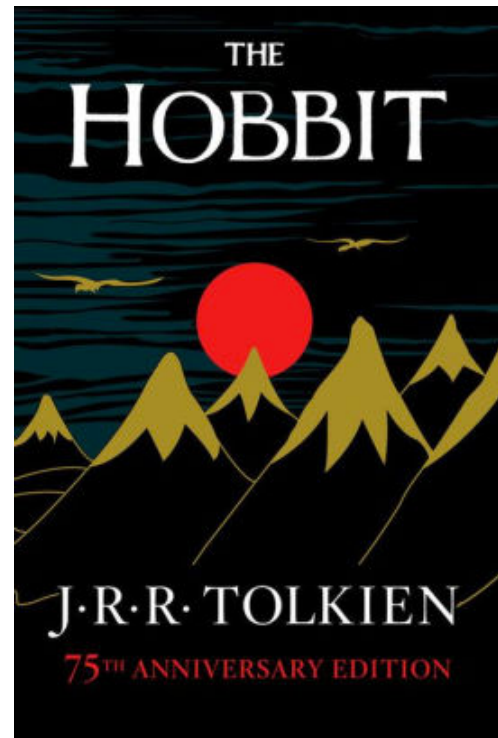
$$h(k) = (rand() * k.numPages) \% m$$

$$h(k) = (\text{Order I insert} [\text{Order seen}]) \% m$$

Hash Function

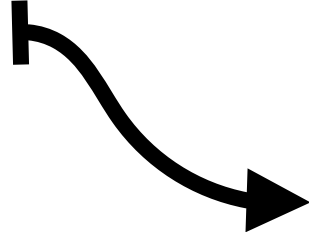


A Hash Table based Dictionary



Author Name
Hash Function

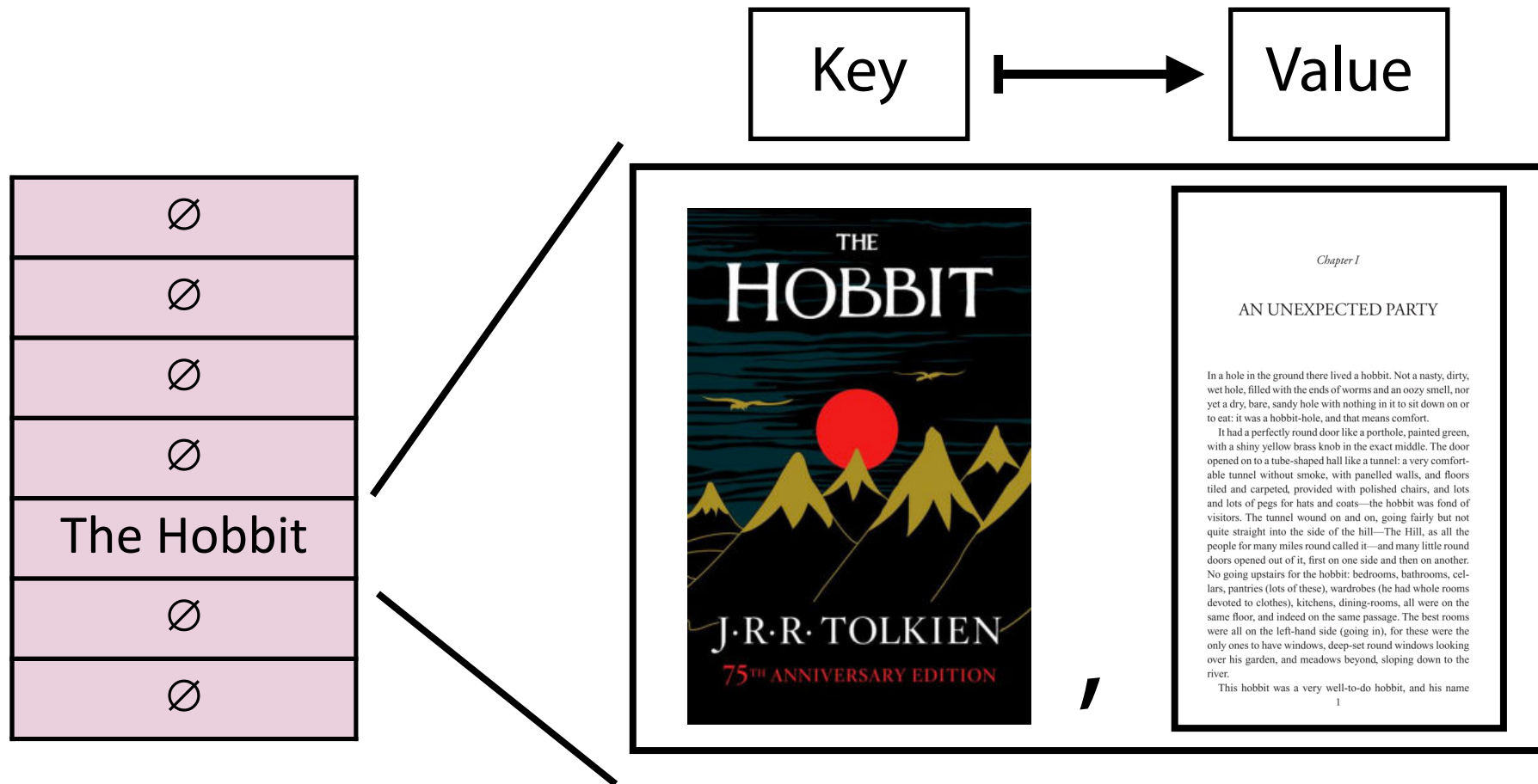
'J' + 'T' = 30



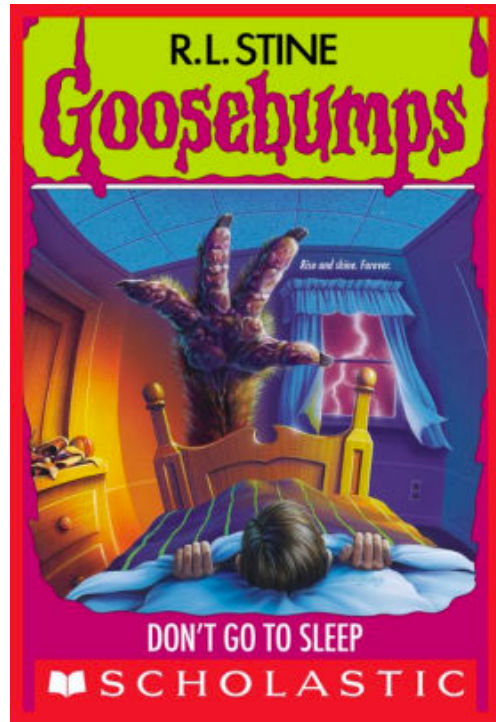
27
28
29
30
31
...

...
∅
∅
∅
The Hobbit
∅
...

A Hash Table based Dictionary

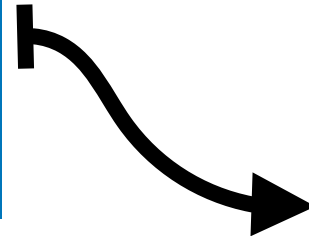


A Hash Table based Dictionary



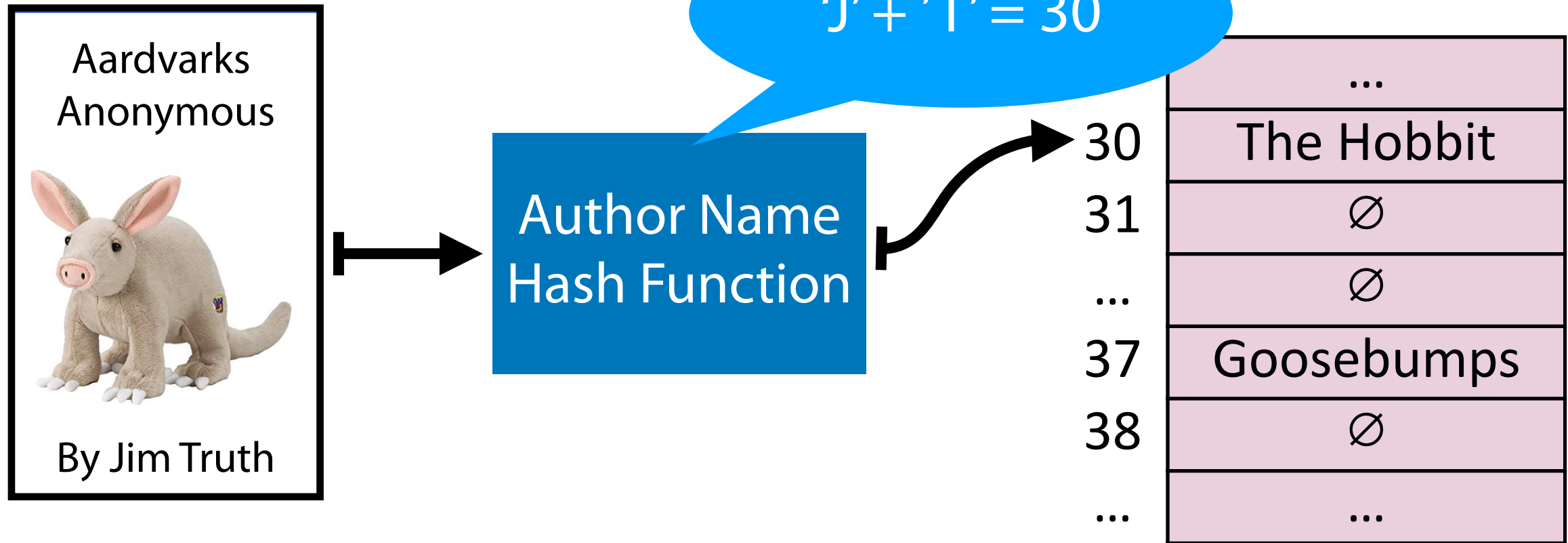
Author Name
Hash Function

'R' + 'S' = 37



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

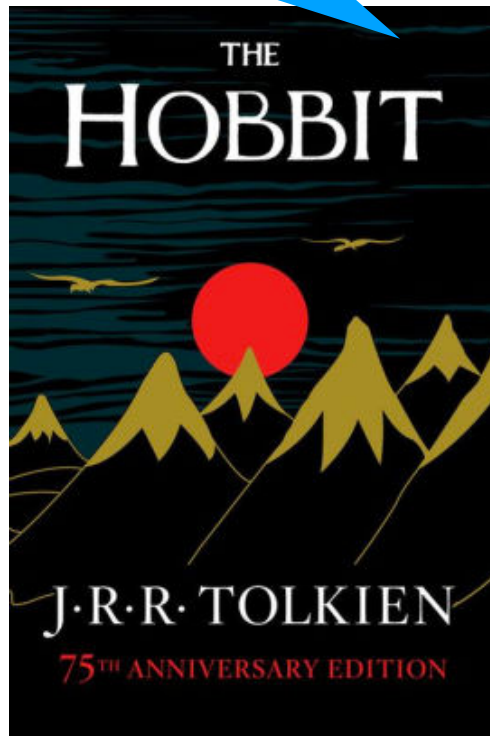
A Hash Table based Dictionary



Hash Collision

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

J.R.R Tolkien = 30!



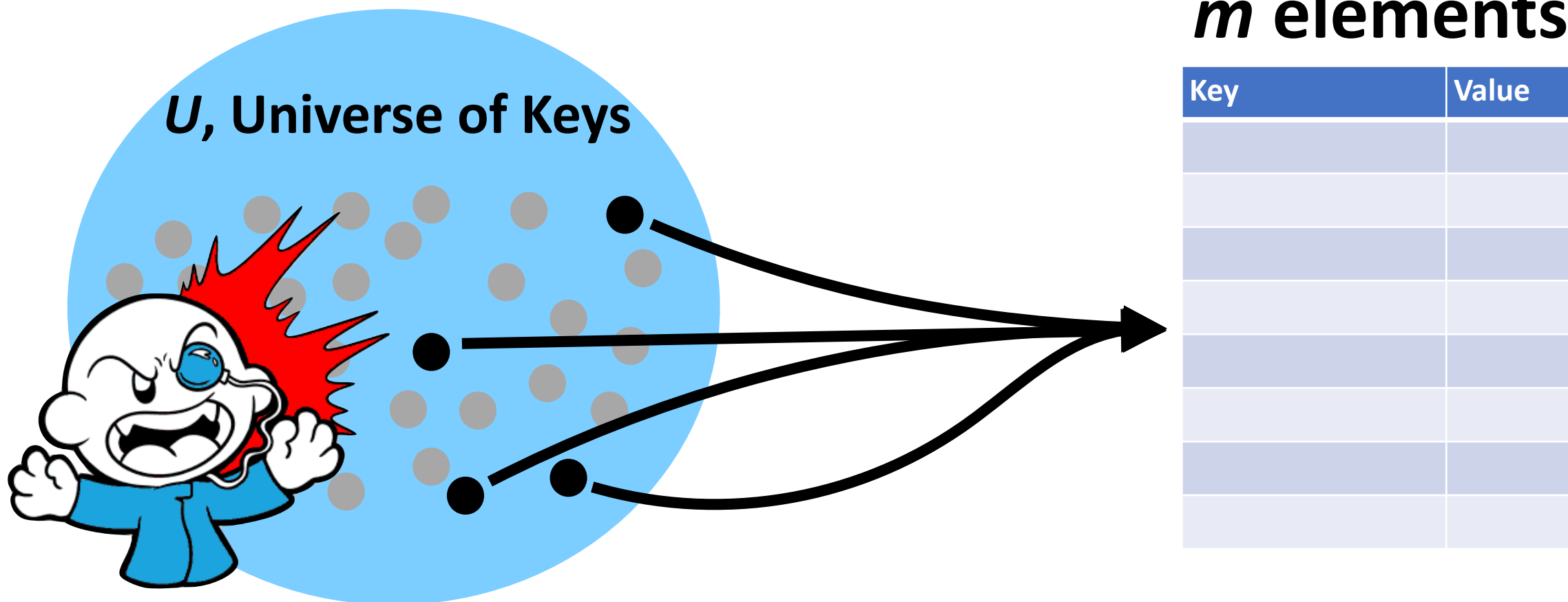
Jim Truth = 30!



...	...
30	???
31	∅
...	∅
37	Goosebumps
38	∅
...	...

General Purpose Hashing

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





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*

Open vs Closed Hashing

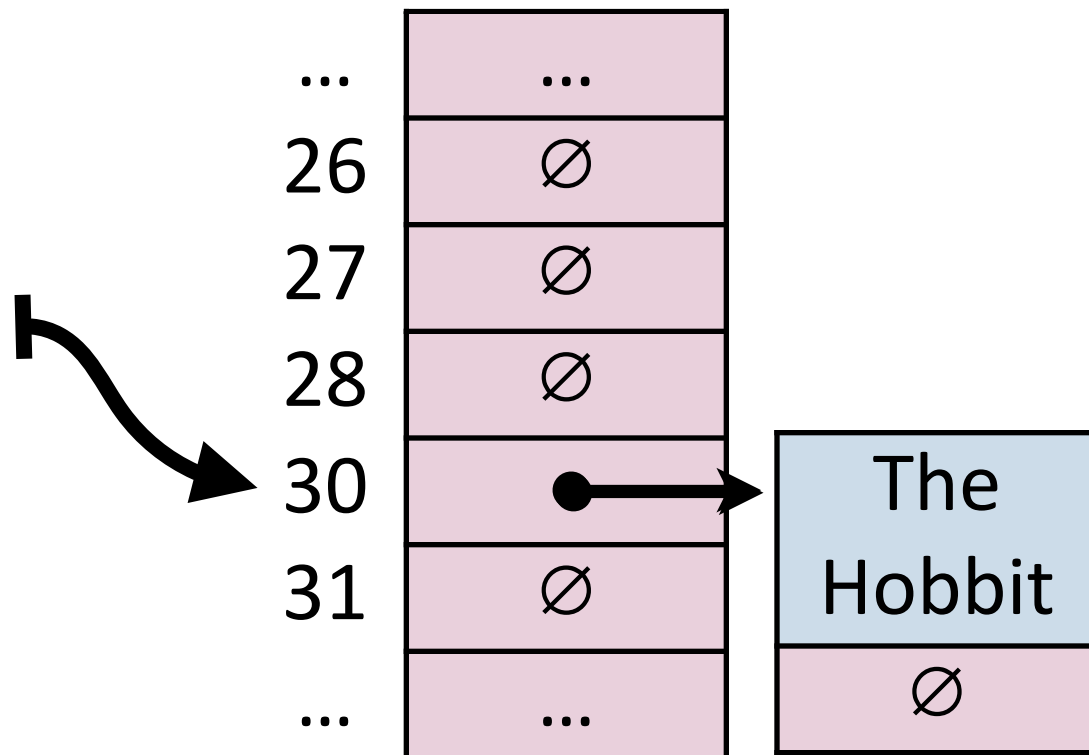
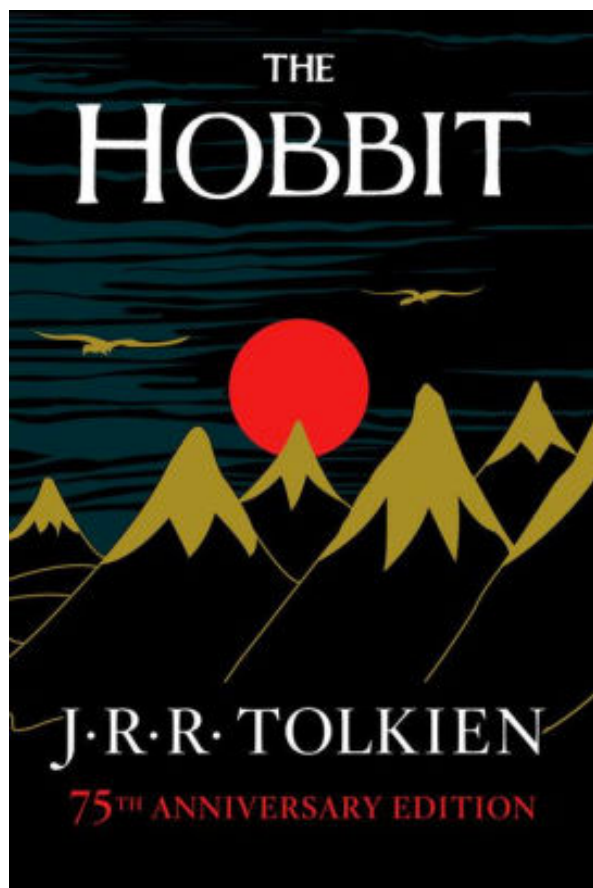
Addressing hash collisions depends on your storage structure.

- **Open Hashing:**

- **Closed Hashing:**

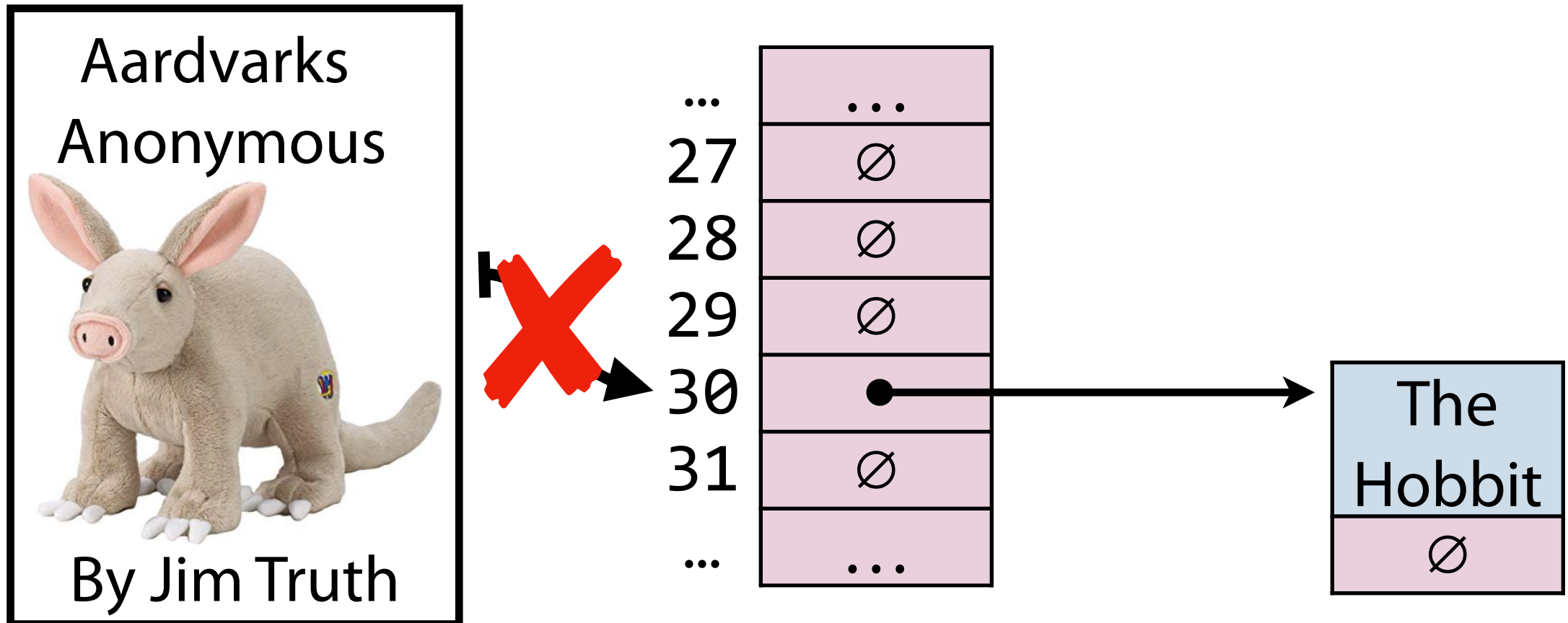
Open Hashing

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 _____ . This is called **separate chaining**.



Insertion (Separate Chaining)

`_insert("Bob")`

`_insert("Anna")`

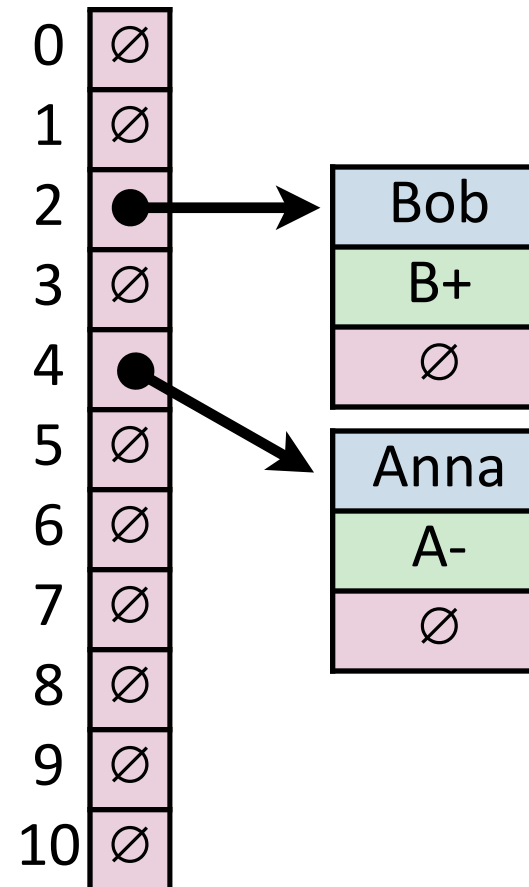
Key	Value	Hash
Bob	B+	2
Anna	A-	4
Alice	A+	4
Betty	B	2
Brett	A-	2
Greg	A	0
Sue	B	7
Ali	B+	4
Laura	A	7
Lily	B+	7

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

Insertion (Separate Chaining)

`_insert("Alice")`

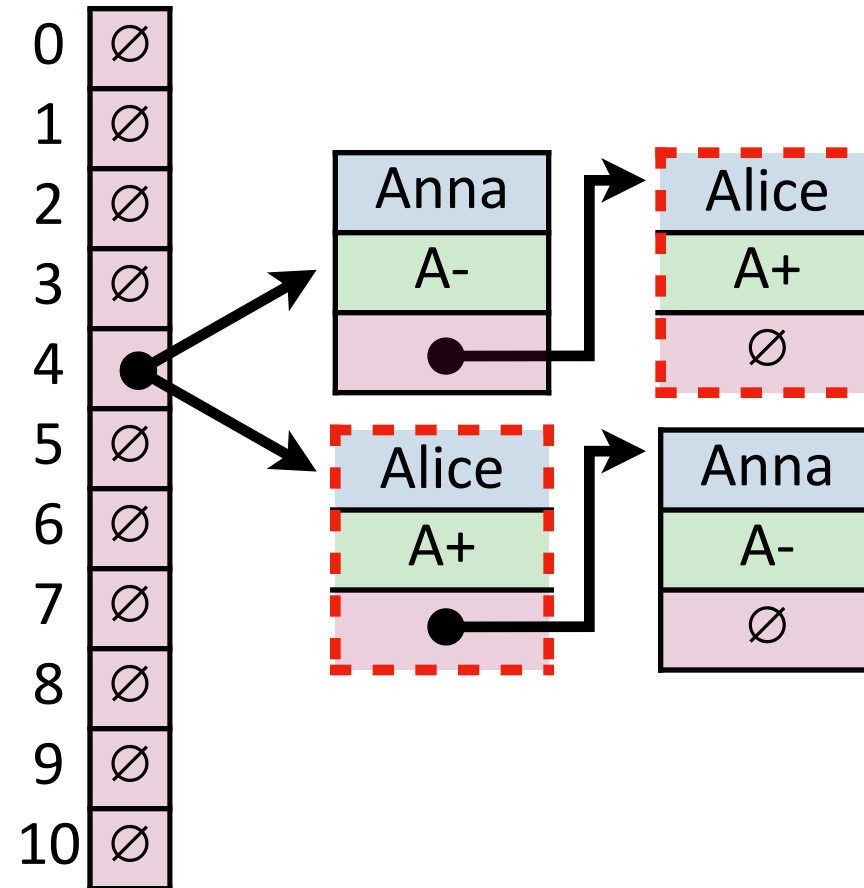
Key	Value	Hash
Bob	B+	2
Anna	A-	4
Alice	A+	4
Betty	B	2
Brett	A-	2
Greg	A	0
Sue	B	7
Ali	B+	4
Laura	A	7
Lily	B+	7



Insertion (Separate Chaining)

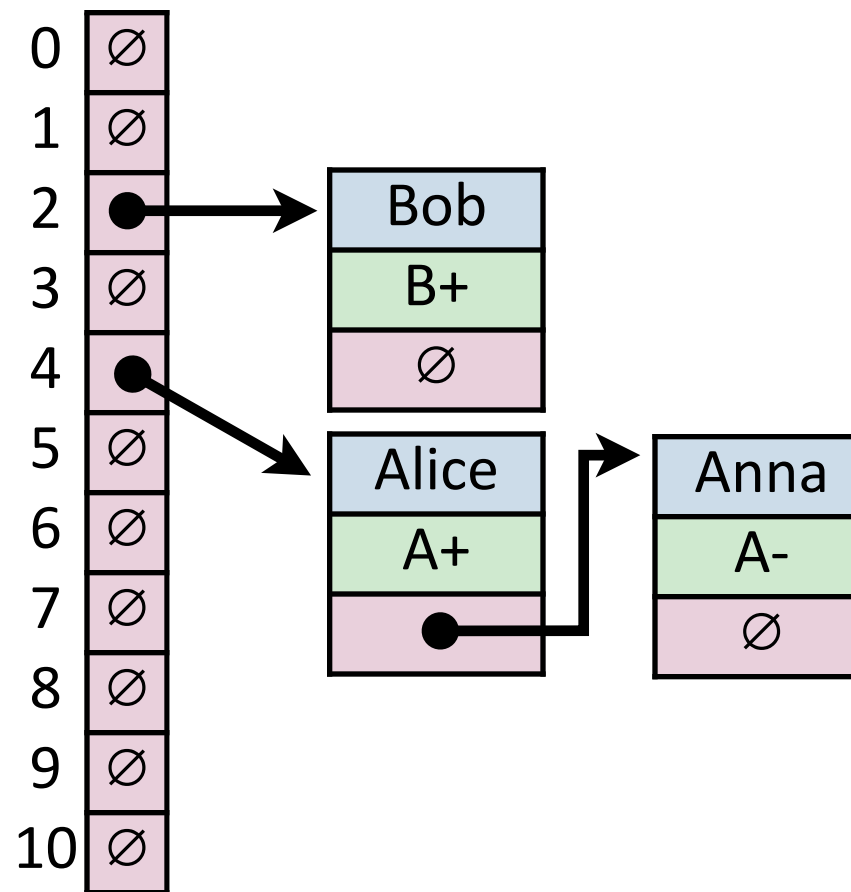
Where does Alice end up relative to Anna in the chain?

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



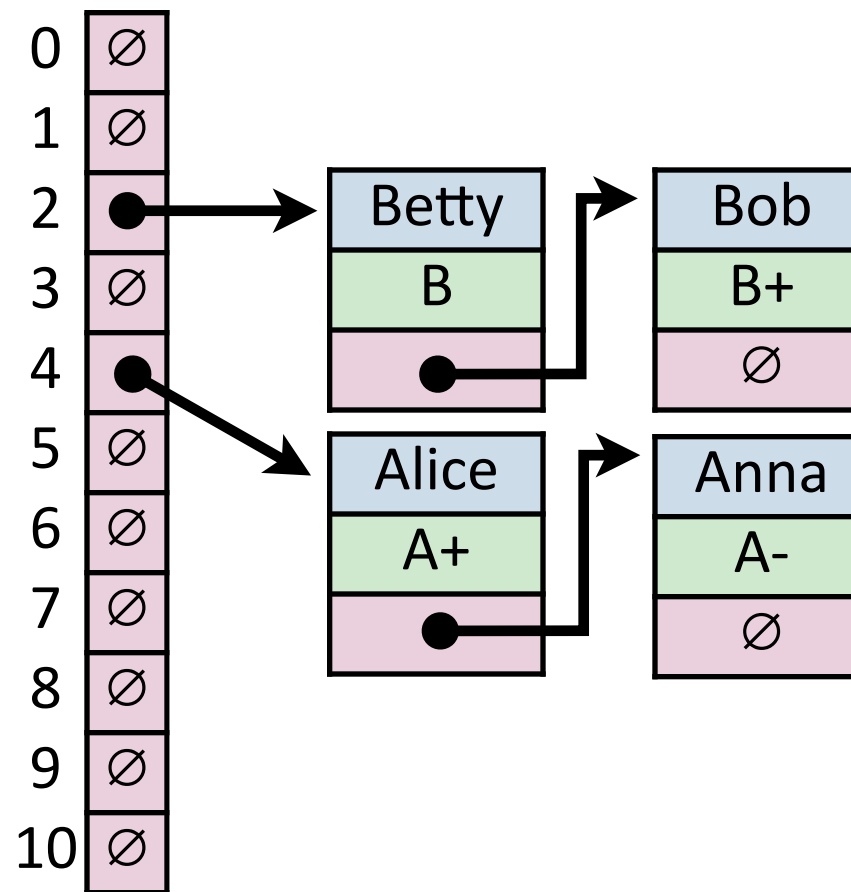
Insertion (Separate Chaining)

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



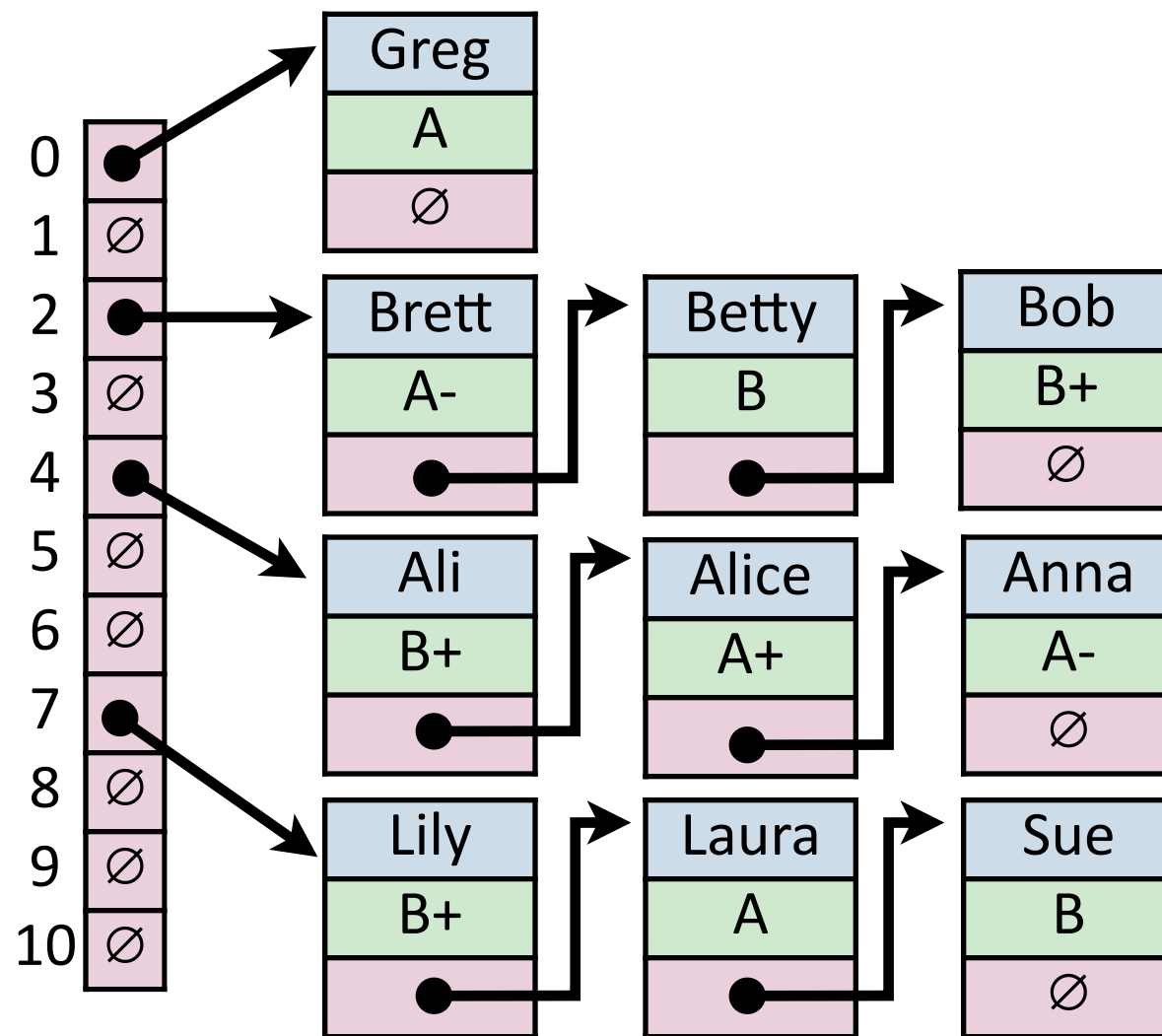
Insertion (Separate Chaining)

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)

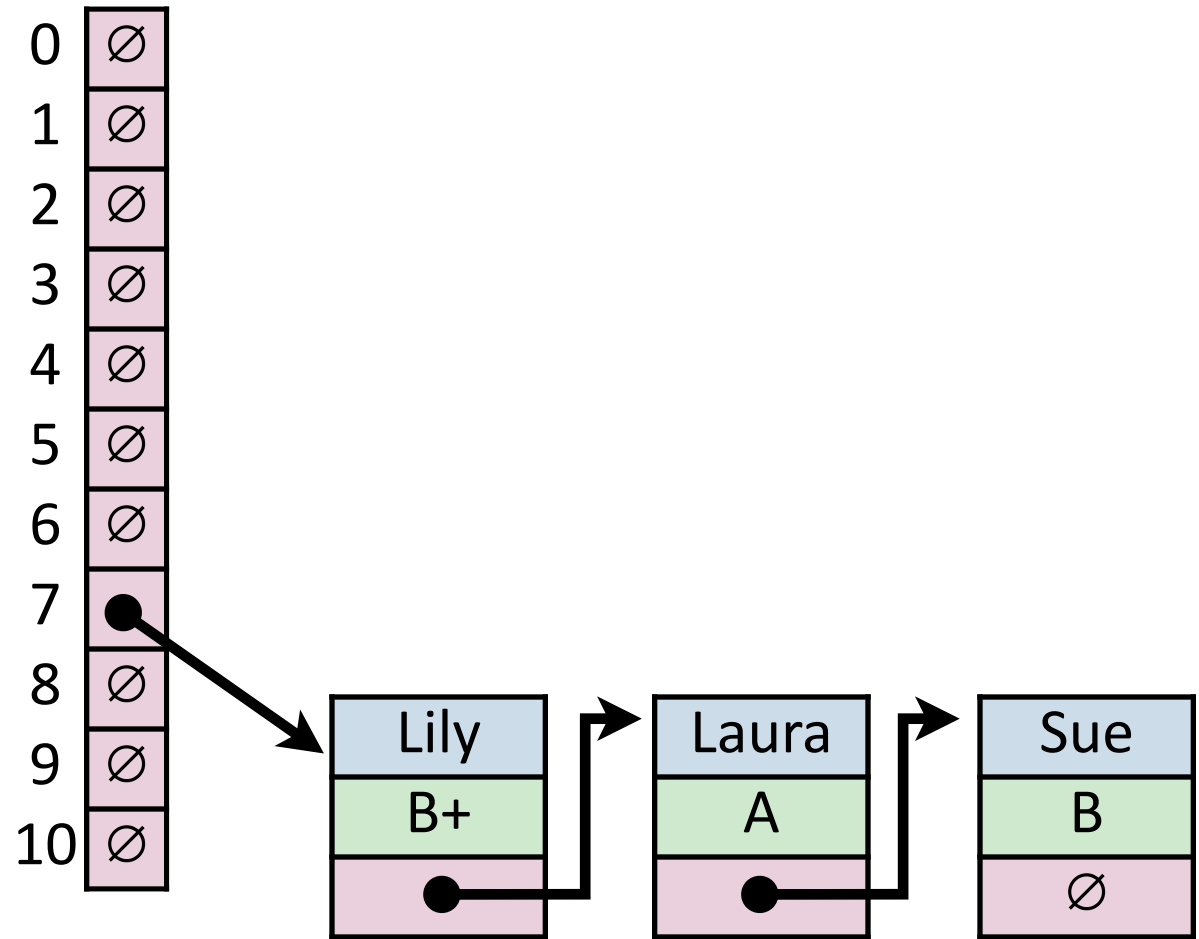
Key	Value	Hash
Bob	B+	2
Anna	A-	4
Alice	A+	4
Betty	B	2
Brett	A-	2
Greg	A	0
Sue	B	7
Ali	B+	4
Laura	A	7
Lily	B+	7



Find (Separate Chaining)

`_find("Sue")`

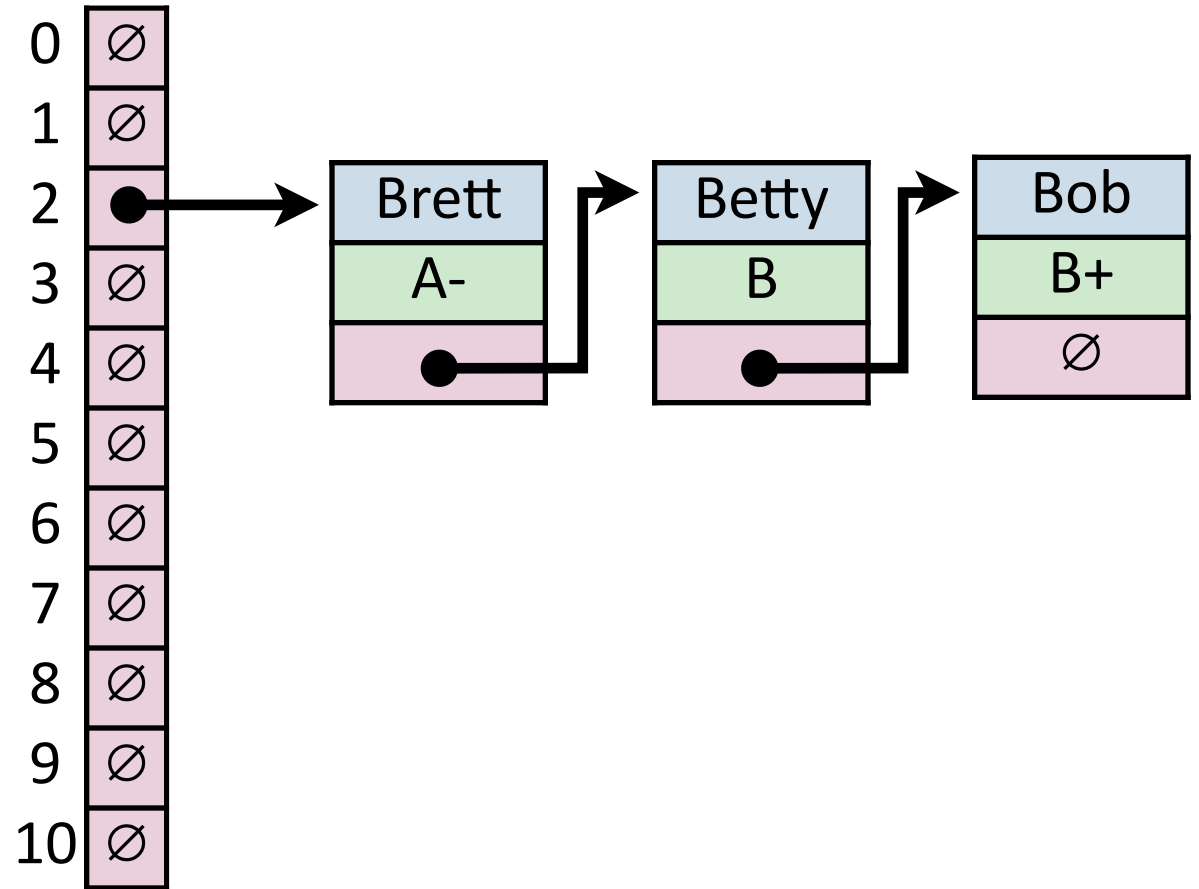
Key	Hash
Sue	7



Remove (Separate Chaining)

`_remove("Betty")`

Key	Hash
Betty	2



Hash Table (Separate Chaining)

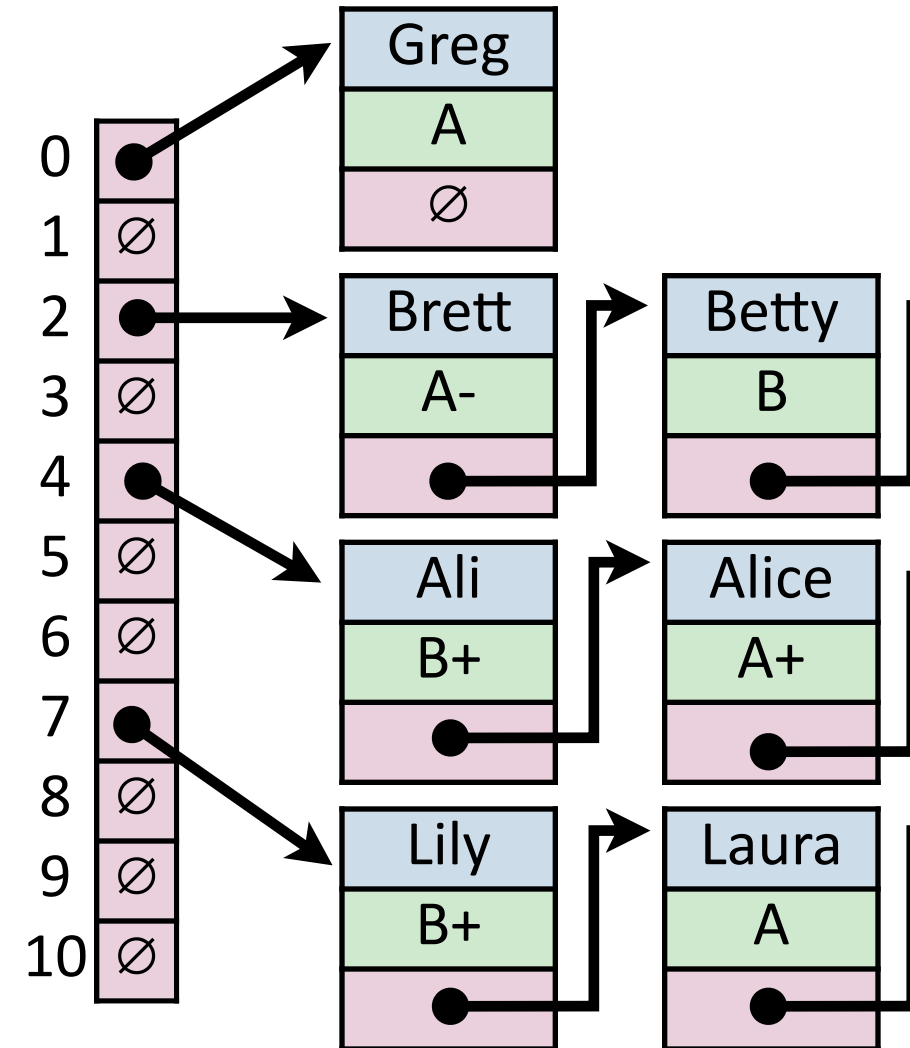


For hash table of size m and n elements:

Find runs in: _____

Insert runs in: _____

Remove runs in: _____



Fundamentals of Probability

Imagine you roll a pair of six-sided dice.

The **sample space** Ω is the set of all possible outcomes.

An **event** $E \subseteq \Omega$ is any subset.

Fundamentals of Probability

Imagine you roll a pair of six-sided dice. What is the expected value?

The **expectation** of a (discrete) random variable is:

$$E[X] = \sum_{x \in \Omega} Pr\{X = x\} \cdot x$$

Fundamentals of Probability

Imagine you roll a pair of six-sided dice. What is the expected value?

Linearity of Expectation: For any two random variables X and Y ,

$$E[X + Y] = E[X] + E[Y]$$

Fundamentals of Probability

Imagine you roll a pair of six-sided dice. What is the expected value?

Linearity of Expectation: For any two random variables X and Y ,

$$\begin{aligned} E[X + Y] &= E[X] + E[Y] \\ &= \sum_x \sum_y Pr\{X = x, Y = y\}(x + y) \end{aligned}$$

Fundamentals of Probability

Imagine you roll a pair of six-sided dice. What is the expected value?

Linearity of Expectation: For any two random variables X and Y ,

$$E[X + Y] = E[X] + E[Y]$$

$$= \sum_x \sum_y \Pr\{X = x, Y = y\}(x + y)$$

$$= \sum_x x \sum_y \Pr\{X = x, Y = y\} + \sum_y y \sum_x \Pr\{X = x, Y = y\}$$

Fundamentals of Probability

Imagine you roll a pair of six-sided dice. What is the expected value?

Linearity of Expectation: For any two random variables X and Y ,

$$E[X + Y] = E[X] + E[Y]$$

$$= \sum_x \sum_y \Pr\{X = x, Y = y\}(x + y)$$

$$= \sum_x x \sum_y \Pr\{X = x, Y = y\} + \sum_y y \sum_x \Pr\{X = x, Y = y\}$$

$$= \sum_x x \cdot \Pr\{X = x\} + \sum_y y \cdot \Pr\{Y = y\}$$

Fundamentals of Probability



Imagine you roll a pair of six-sided dice. What is the expected value?

Linearity of Expectation: For any two random variables X and Y ,

$$E[X + Y] = E[X] + E[Y]$$

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

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

Uniform:

Independent:

Separate Chaining Under SUHA

Table Size: m

Num objects: n

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

α_j = expected # of items hashing to position j

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

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

Separate Chaining Under SUHA

Table Size: m

Num objects: n

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

α_j = expected # of items hashing to position j

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

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

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

Separate Chaining Under SUHA

Table Size: m

Num objects: n

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

α_j = expected # of items hashing to position j

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

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

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

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

Separate Chaining Under SUHA

Table Size: m

Num objects: n

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

α_j = expected # of items hashing to position j

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

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

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

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

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

Separate Chaining Under SUHA

Table Size: m

Num objects: n

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

α_j = expected # of items hashing to position j

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

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

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

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

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

Separate Chaining Under SUHA



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

α_j = expected # of items hashing to position j

Num objects: n

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

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

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

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

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

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

Separate Chaining Under SUHA

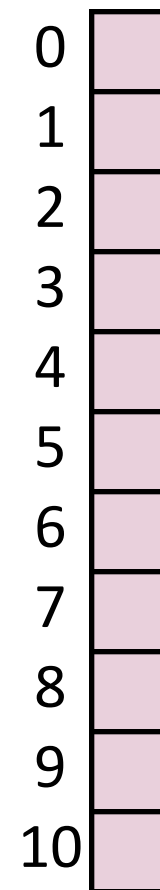


Under SUHA, a hash table of size m and n elements:

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

$$S = \{ 1, 8, 15 \}$$

$$h(k) = k \% 7$$

