**HASHMAPS AND BLOOM FILTERS:**

**Bloom Filters:**

Suppose you want to store set S of size N in filter T of size M, where T is an array of bits

To do this, use multiple has functions, h1 … hk

To add x to T, compute h1(x), h2(x), etc, and set all those bits to 1

For a false positive z, z is not part of S, and T(hi(z)) = 1 for all 1 to k has functions

K = M/N \* ln2

For a set of strings with average length l, we would need 8lN bits to store S

With a bloom filter, we only need 8N bits

**Applications of Bloom Filters:**

One early application of Bloom filters was as a spell checker

The dictionary of valid words was stored in a BF

If the word check said it was not in the Dictionary, then it definitely was not

If it said it was, then it most likely was

Suppose you want to store a set S, and want to know if query q belongs to S

Suppose also that the queries usually don’t belong to S

It is expensive to store S in main memory

Store S as bloom filter instead, with actual S in secondary memory

One application of the above is google chrome, which wanted to track a few million malicious websites, and store the list on your local machine, along with chrome

Chrome takes a few MB, whereas the list is over 1GB

Instead, create bloom filter of malicious websites

**Bloom Filter Drawbacks:**

The false positives are the most obvious drawback

You are also unable to remove an element easily

Counting is also difficult (can implement counter for each cell)

Cannot retrieve elements

**Hash functions for Bloom Filters:**

We always assume we have a hash function from U to T that is uniformly at random

We want to know how to store the hash function

There are two common practices for storing a hash function

One is to use a deterministic hash function, the other is to use a limited random hash function

**Limited Random Hash Function:**

To store S in table T of size M, start by picking a prime number P

Uniformly randomly pick a, b elements of {0, 1, … p-1}

Hab(x) = (ax + b)%P

**Deterministic Hash Functions:**

Most programming languages use this