Coupon collector & more Bloom filters

Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

Department of Computer Science

Balls and Bins

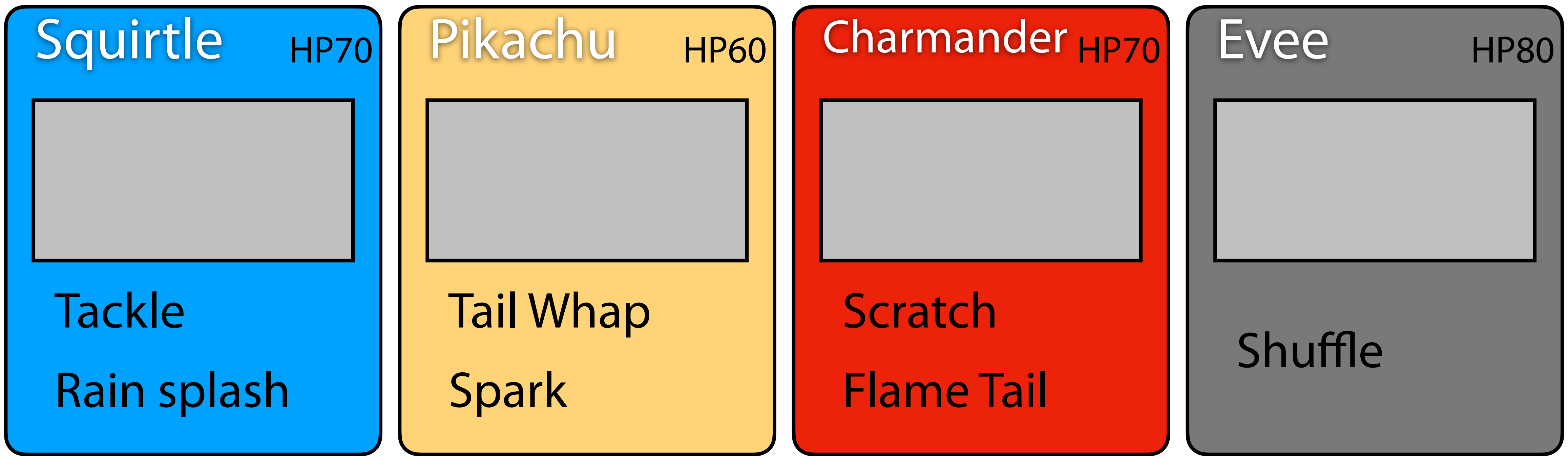
Say we've chosen n & k, and our set-bit fraction target is 50%

Can we use Balls-and-Bins thinking to estimate what m would get us there?

I throw m balls into n bins uniformly and independently. What can I ask?

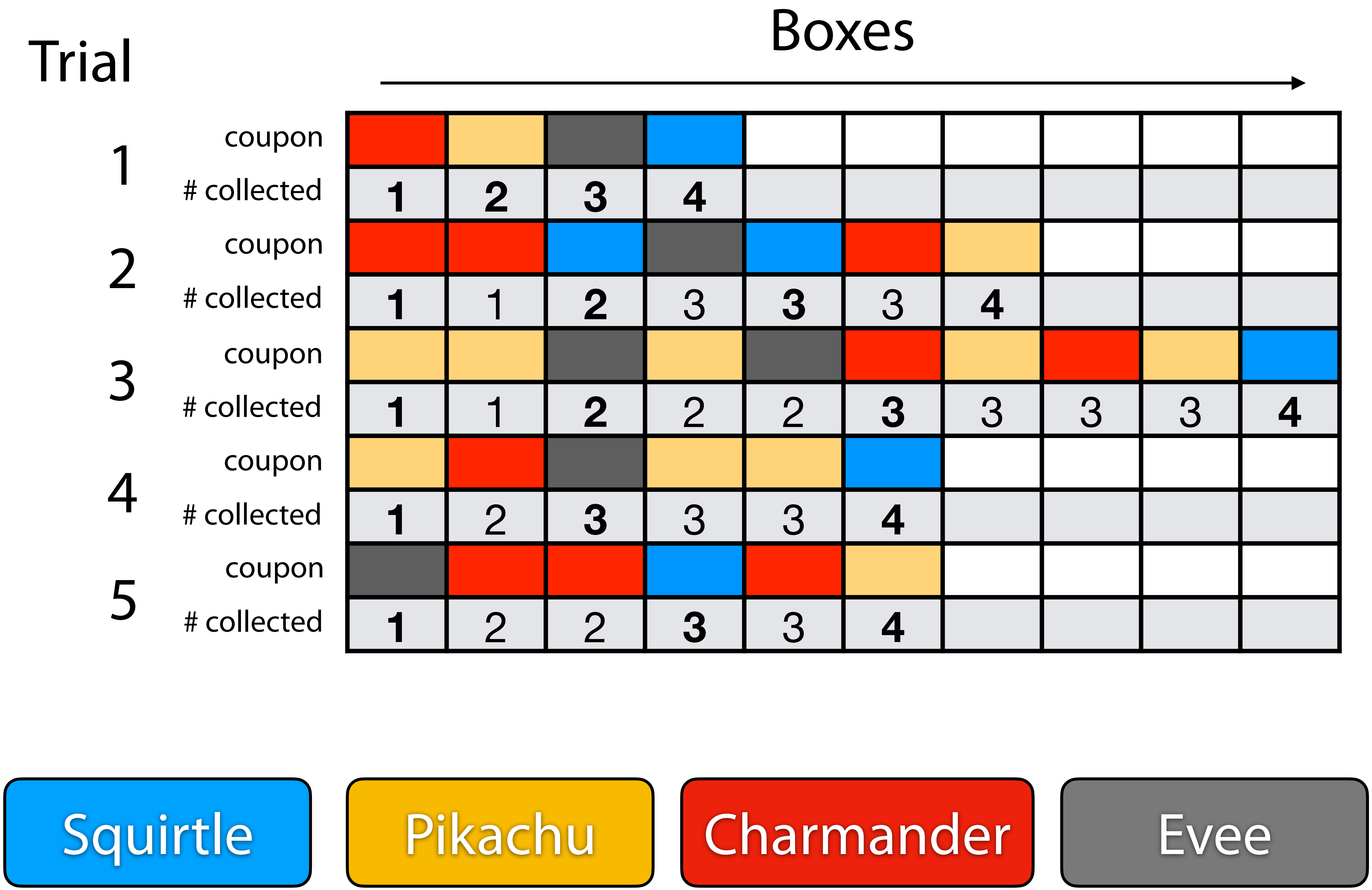
Category	Questions			Approach
Empty/ non empty	How many buckets are empty?	What's the chance all buckets are non-empty?	**How many balls until 1/2 of bins are non-empty?**	**Coupon collector**
Collisions / no collisions	How many throws until there is a >0.5 chance of a collision?	What is the chance no bin has >1 item?		**Birthday problem**
Local (single bin) occupancy	What's the occupancy of a given bucket?	What is the chance a given bucket has >2 items?		**Binomial & Poisson r.v.s**
Global occupancy	What is the *median* bucket occupancy?	What is the *maximum* bucket occupancy?		**Often hard** E.g. M&U Lemma 5.1 on p100

Coupon collector



We have n "coupons" to collect. We collect them by opening boxes of cereal. Each box has 1 random coupon; probability is uniform ($1/n$ chance of each) and independent (no box affects another).

Coupon collector



Coupon collector

How many boxes until we collect em all?

Formally: if X is an r.v. for # boxes up to and including box with final coupon, what is $\mathbf{E}[X]$?

Idea: partition sequence into *stages* by # coupons collected so far

1	**2**	**3**	**4**						
1	1	**2**	3	**3**	3	**4**			
1	1	**2**	2	2	**3**	3	3	3	**4**

Bernoulli random variable

An r.v. X is a Bern(p) (Bernoulli) random variable if it takes value 1 with probability p, 0 otherwise

A fair coin is Bern(0.5), letting heads=1 and tails=0. A *loaded* coin that lands heads with probability 0.75 is Bern(0.75).

$$\mathbf{E}[X] = p$$

Geometric random variable

Geom(p) random variable X equals # trials of a Bern(p) r.v. up to the first success

$$\Pr(X = n) = (1 - p)^{n-1}\, p$$

failures 1st success

$$\mathbf{E}[X] = \frac{1}{p}$$

Coupon collector

Let X_i for $i = 1, 2, \ldots, n$ be r.v.s for # boxes bought while holding $i - 1$ coupons

For X as just defined, $X = \sum_{i=1}^{n} X_i$

Sum is stratified by "stage"

Coupon collector

If we hold $i - 1$ coupons, probability p_i that next box has a new coupon is: $\dfrac{n-(i-1)}{n}$

A "loaded coin" we flip repeatedly until success; sounds like ... a geometric

Each X_i is Geom(p_i)

Coupon collector

Each X_i is a Geom(p_i) r.v. and

$$\mathbf{E}[X_i] = \frac{1}{p_i} = \frac{n}{n-i+1}$$

With linearity of expectation:

$$\mathbf{E}[X] = \mathbf{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbf{E}[X_i]$$

$$= \sum_{i=1}^{n} \frac{n}{n-i+1} = n\sum_{i=1}^{n} \frac{1}{n-i+1} = n\sum_{i=1}^{n} \frac{1}{i}$$

Coupon collector

Say want to keep a Bloom filters's set-bit ratio near 50%. What # items can we add until the expected set-bit ratio exceeds 0.5?

$$n \sum_{i=1}^{n} \frac{1}{n-i+1}$$

Instead of summing to n (100%), stop after 50% of coupons

$$n \sum_{i=1}^{\lceil 0.5 \cdot n \rceil} \frac{1}{n-i+1}$$

Coupon collector

$$n \sum_{i=1}^{\lceil \alpha \cdot n \rceil} \frac{1}{n-i+1}$$

n	n/2	Coupon collector until 50%	Coupon collector until 100%
100	50	68.82	518.74
1,000	500	692.65	7,485.47
10,000	5,000	6,930.97	97,876.06
100,000	50,000	69,314.22	1,209,014.61

Approaching $n \ln 2$

Tracking with $n \ln n$

Can you see why?