Hash tables & probability

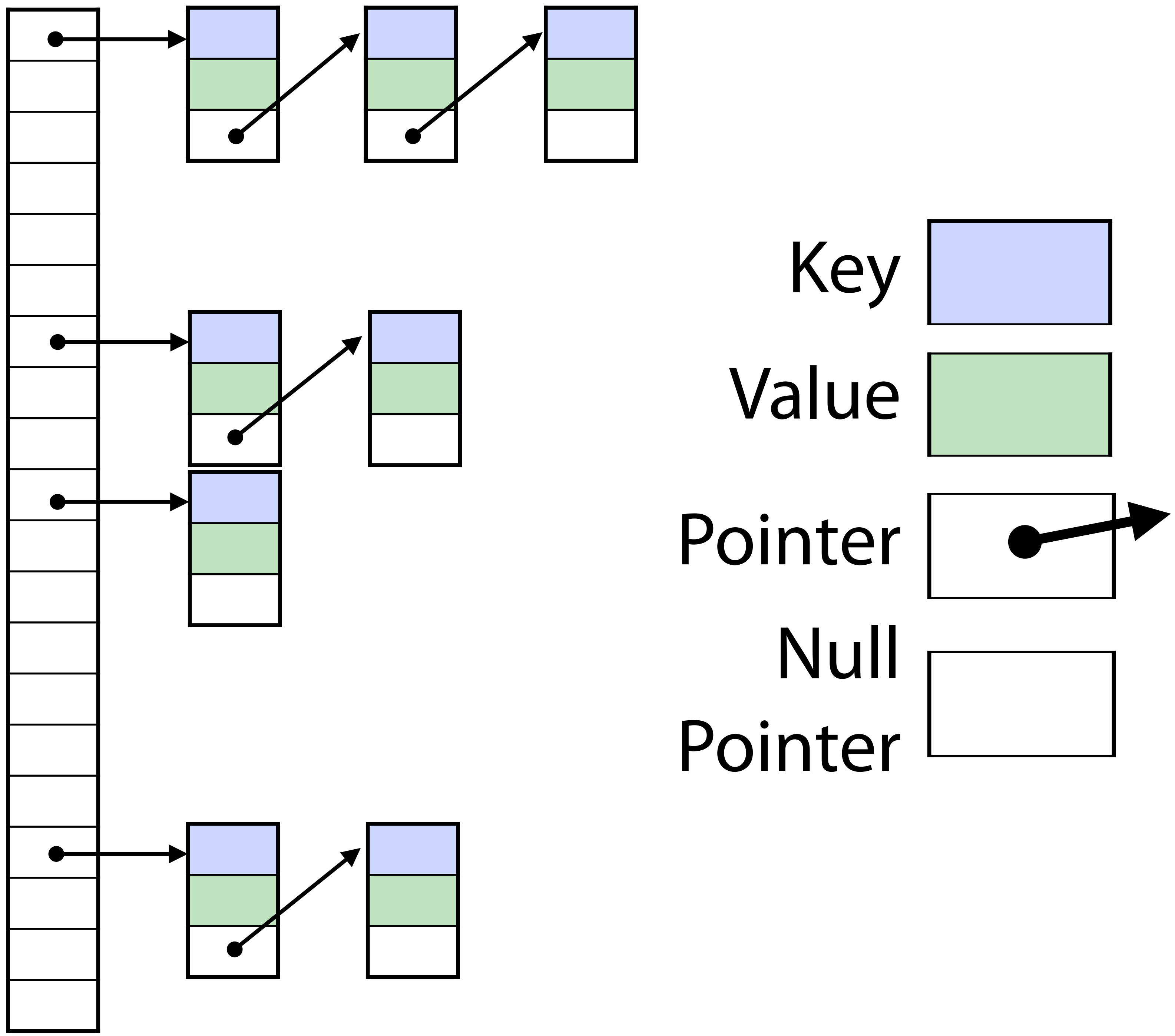
Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL of ENGINEERING

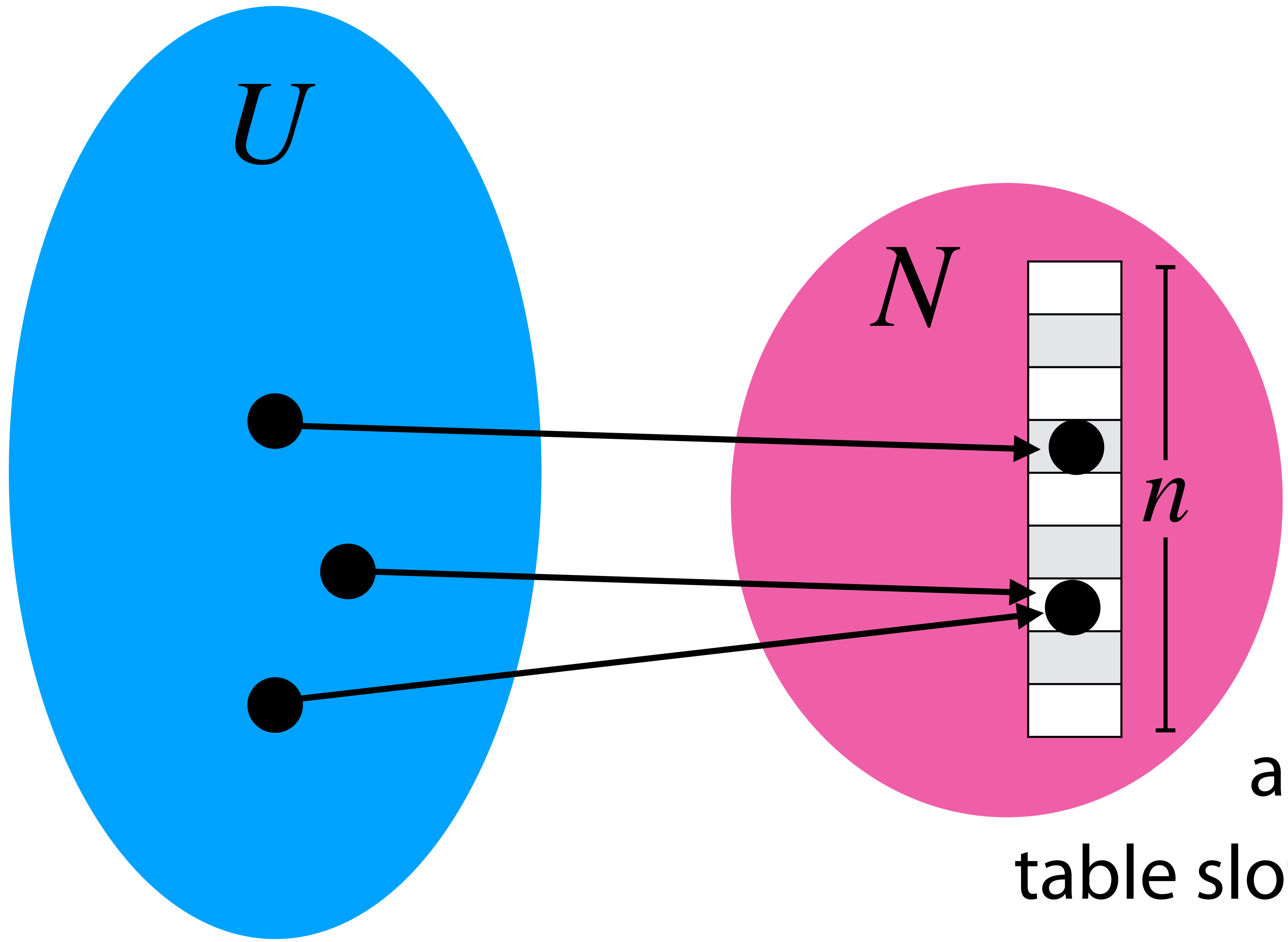
Department of Computer Science

Hash Table

"*Hashing with chaining*" or "*chain hashing*"



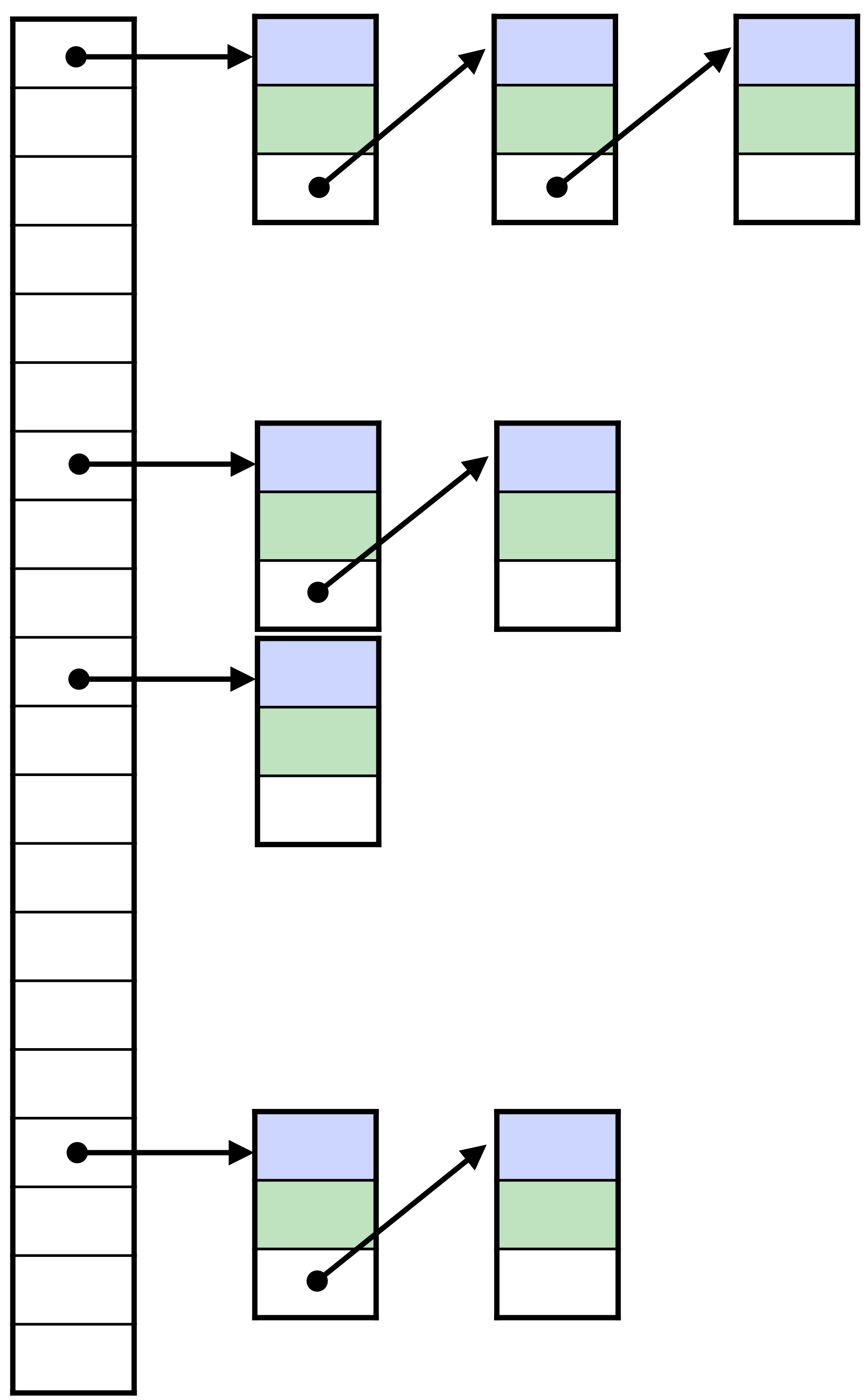
Hash Function



Assume accessing table slot is $O(1)$

Assume hash function operates on any item from U (integers, strings, etc) and is $O(1)$ time

Hash Table



What "abstract data types" can we implement with this?

map	set	counter
$<k_1, v_1>$	$<k_1>$	$<k_1, 7>$
$<k_2, v_2>$	$<k_2>$	$<k_2, 4>$
$<k_3, v_3>$	$<k_3>$	$<k_3, 8>$
$<k_4, v_4>$	$<k_4>$	$<k_4, 5>$

Hash Table

I add m items to an n-bucket hash table

Without probability, what can I say?

Question	Assumption	Statement	Comment
Does any bucket have more the one item?	$m > n$	Yes	Pigeonhole principle
Is any bucket empty?	$m < n$	Yes	"Empty pigeonhole" principle
What is the average bucket occupancy?	-	m/n	-

Nothing profound here

Hash Table

I have added m items to a n-bucket hash table. What "interesting questions" can I ask about the table's state?

How many buckets are empty?

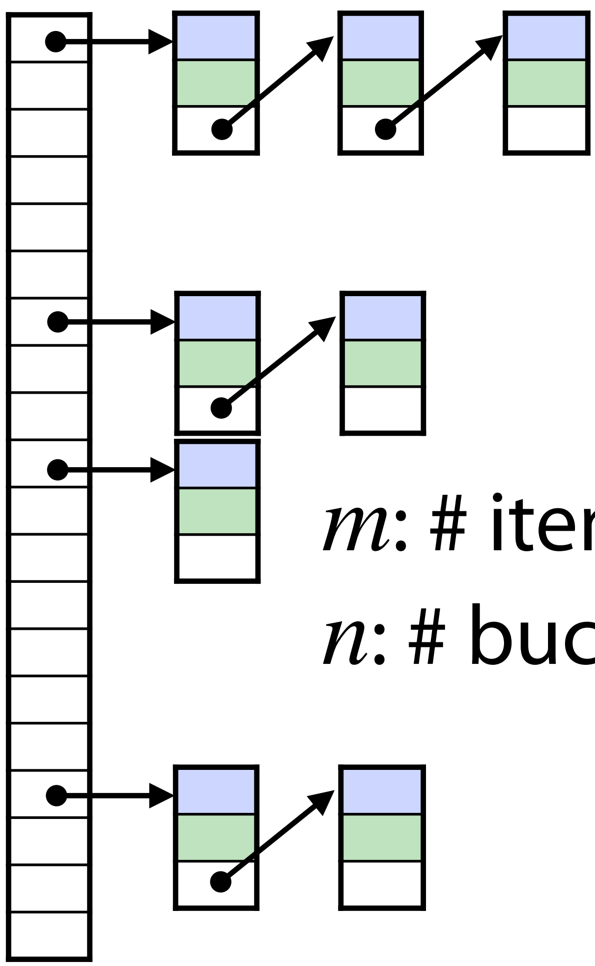
How many items are in the *average* bucket?

How many items are in the *fullest* bucket?

How many items are in the *median* bucket?

What's the chance all buckets are non-empty?

What's the chance no bucket has >1 item?



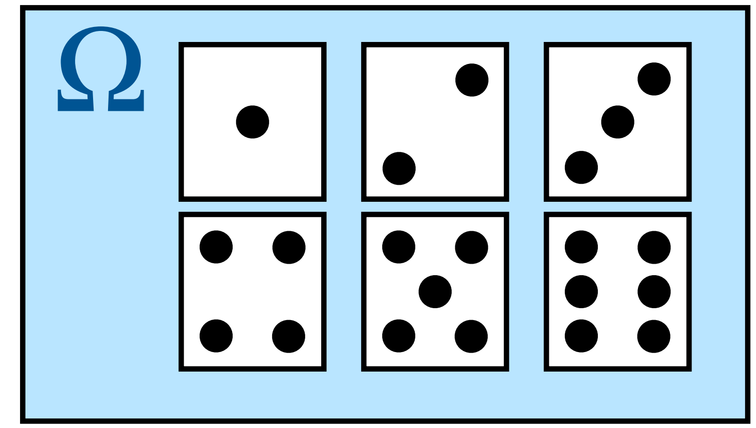
m: # items

n: # buckets

Probability

Sample space (Ω) is ***set*** of all possible outcomes

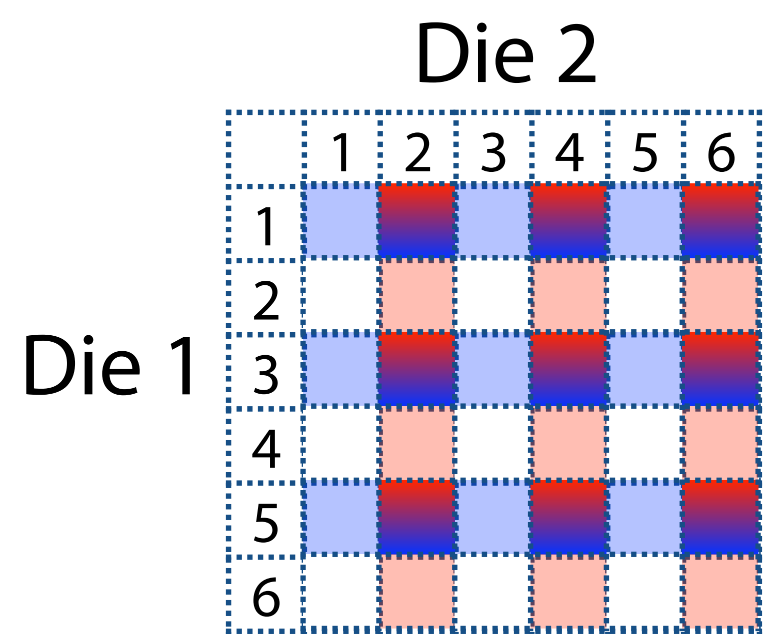
E.g. Ω = { all possible rolls of 2 dice }



An *event* is a ***subset*** of Ω

A = { rolls where 1st die is odd }

B = { rolls where 2nd die is even }



When outcomes are equally likely, can use "naive definition of probability"

$\Pr(A)$: fraction of outcomes that are in *A*

$$\Pr(A) = |A| / |\Omega| = 18/36 = 0.5$$

Probability

"Naive definition" of probability fails to apply when outcomes are not equally probable

Loaded coin

goals scored in soccer game

Probability function $\Pr$

$\Pr : \mathscr{P}(\Omega) \to \mathbb{R}$, where $\mathscr{P}(\Omega)$ is "power set" (set of all subsets) of Ω, satisfies conditions:

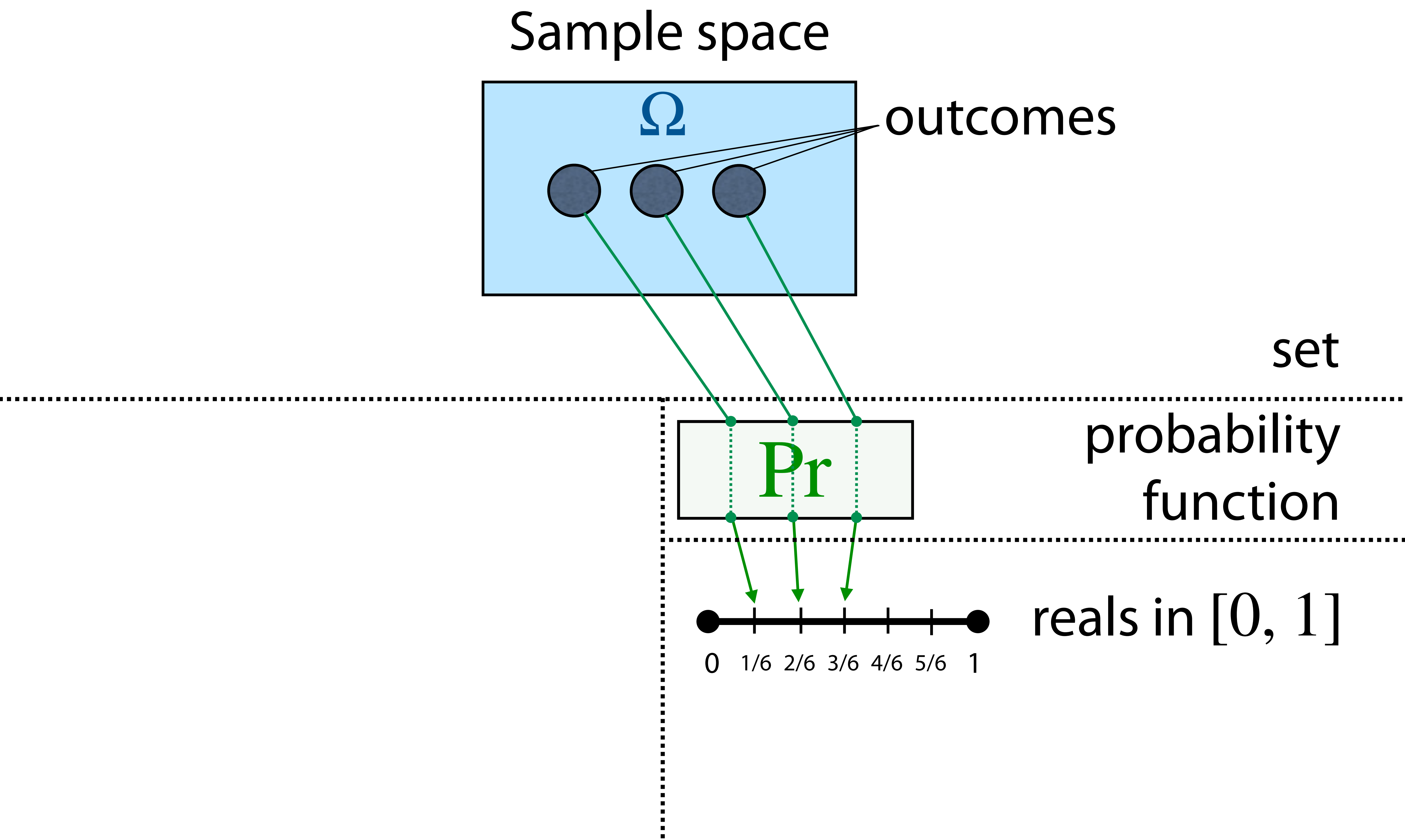
1. For any event E, $0 \leq \Pr(E) \leq 1$

2. $\Pr(\Omega) = 1$

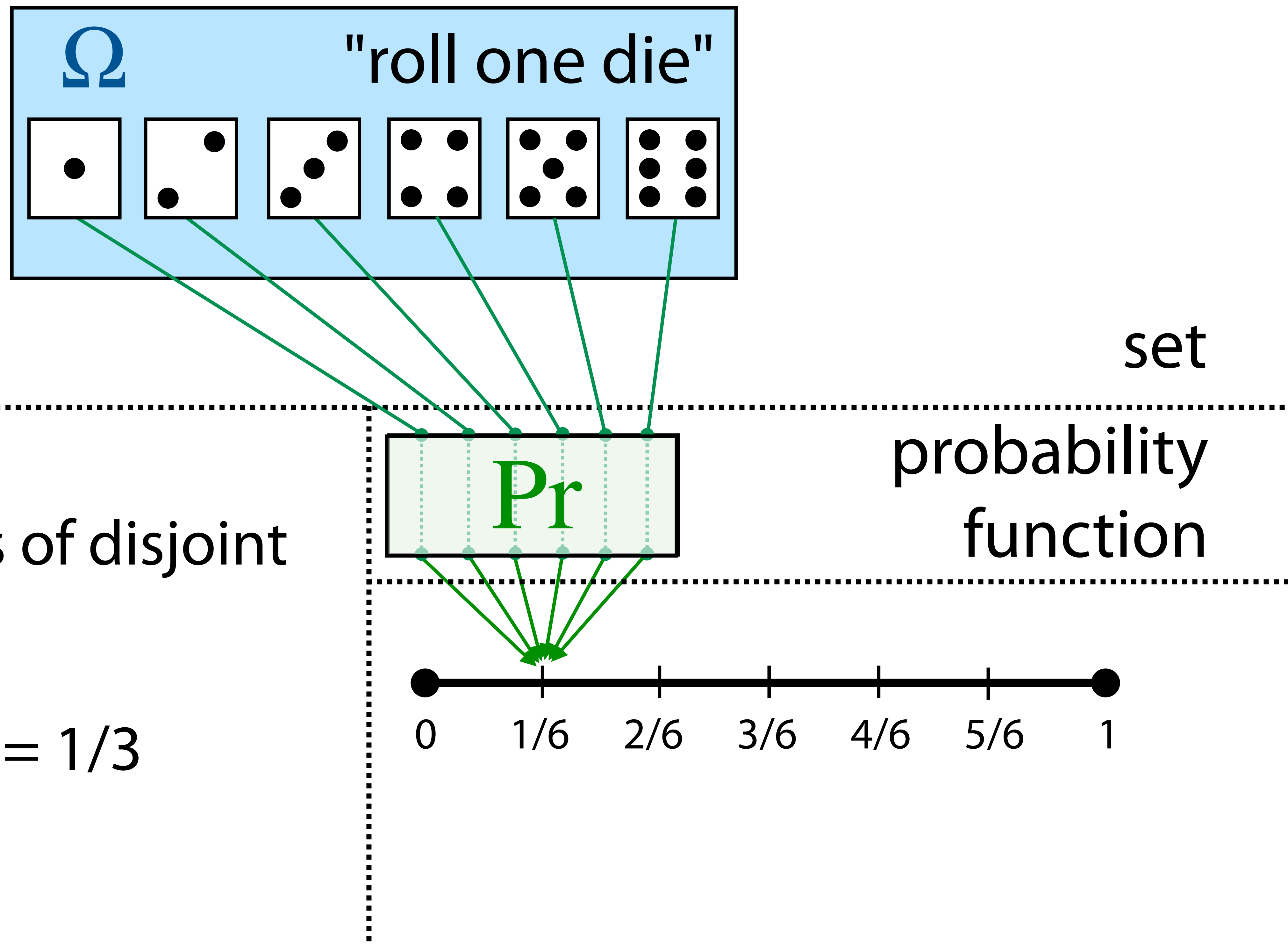
3. Probabilities of disjoint events $E_1, E_2, \ldots$ add:

$$\Pr\left(\bigcup_{i\geq 1} E_i\right) = \sum_{i\geq 1} \Pr(E_i)$$

Probability function Pr



Probability function Pr



Probabilities of disjoint events add;

Pr({⚀, ⚄}) = 1/3

Random variable

Random variables have two "natures"

$$X$$

Function, mapping *outcomes* from Ω to *numbers* (in $\mathbb{R}$)

$X($⚃$) = 4$

$$Y = 3.5 - X$$

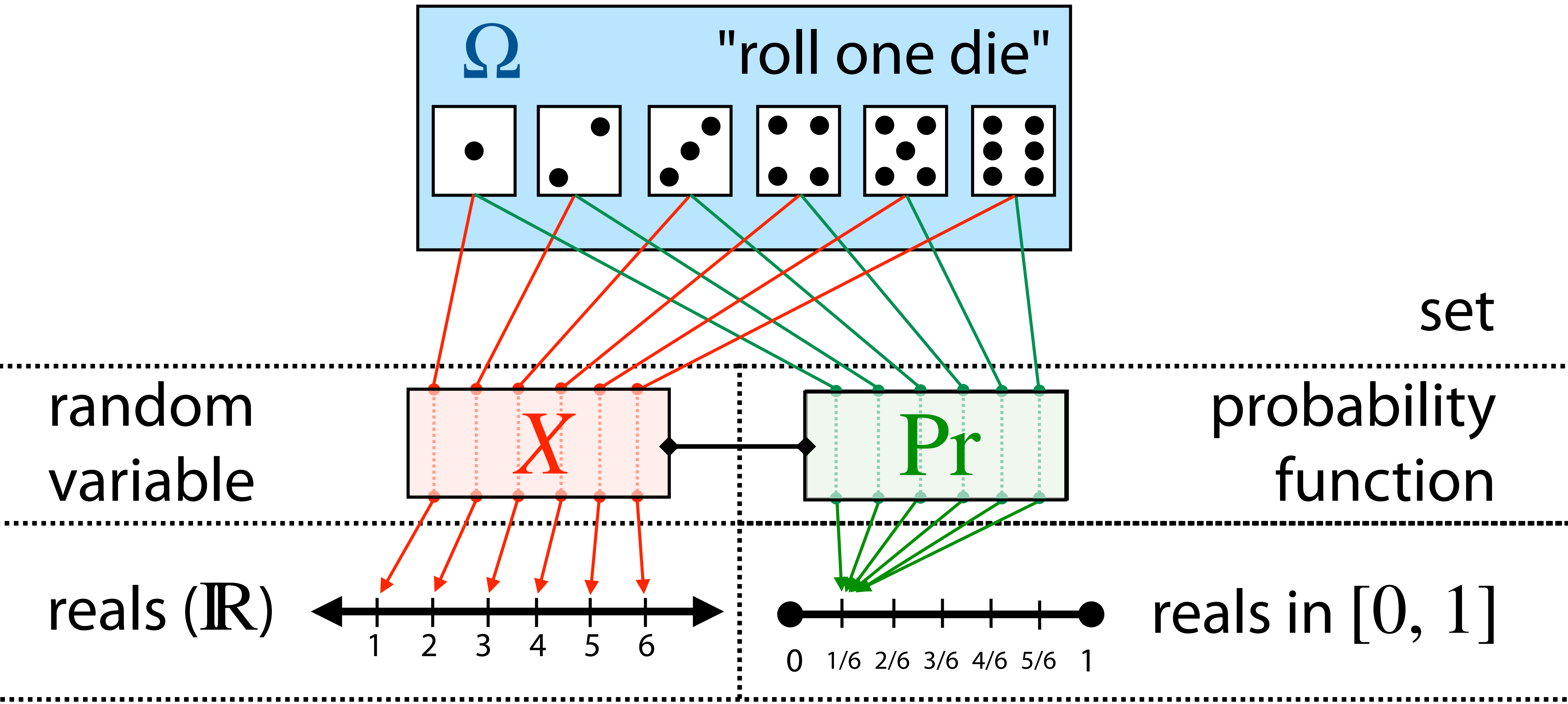
Potential experiment with a *distribution* (a $\Pr$ for its Ω) and numerical result

Random variable

We use capitals e.g. X, Y to denote a random variable

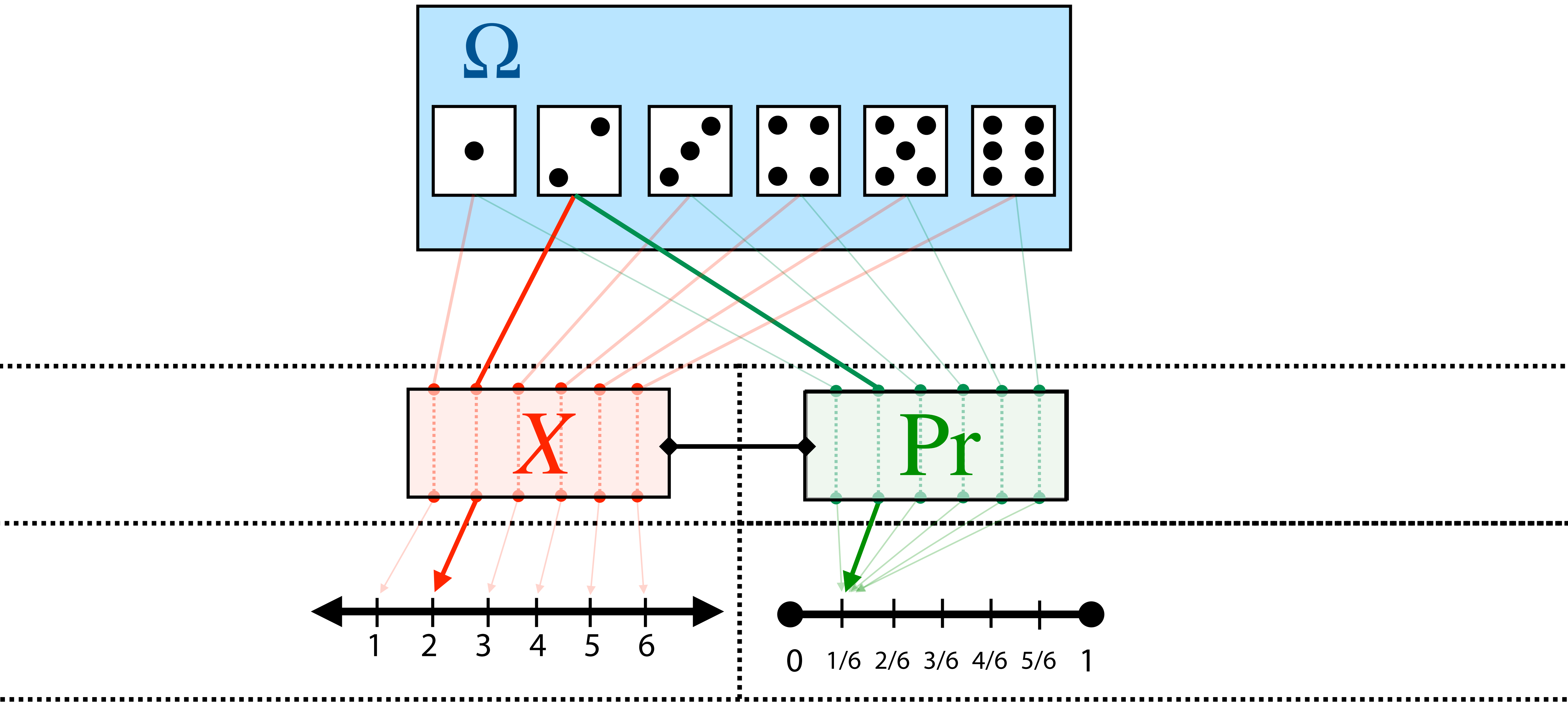
Abbreviate with "r.v."

Random variable & probability function



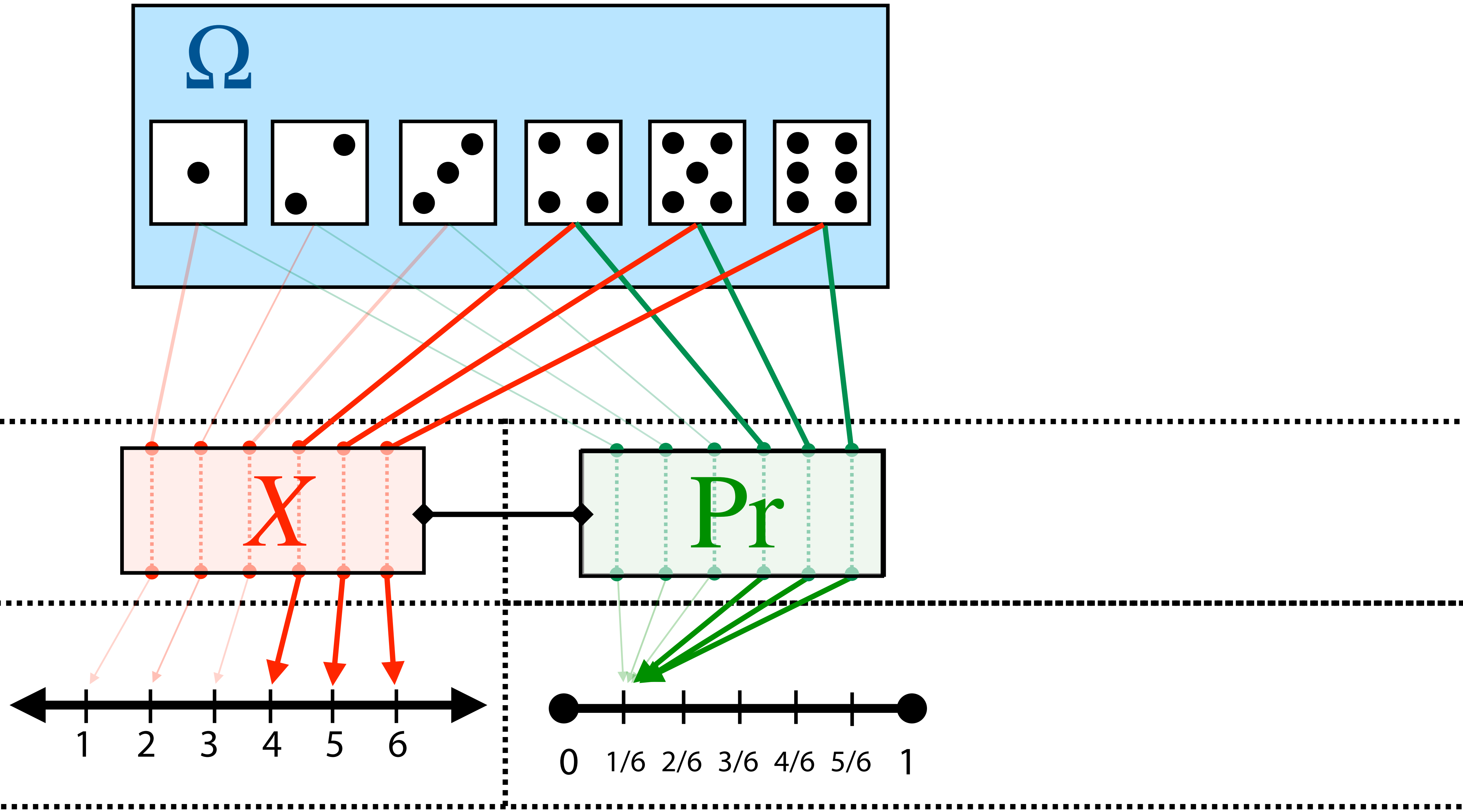
$$\Pr(X = 2) = \Pr(⚁) = 1/6$$

Random variable & probability function



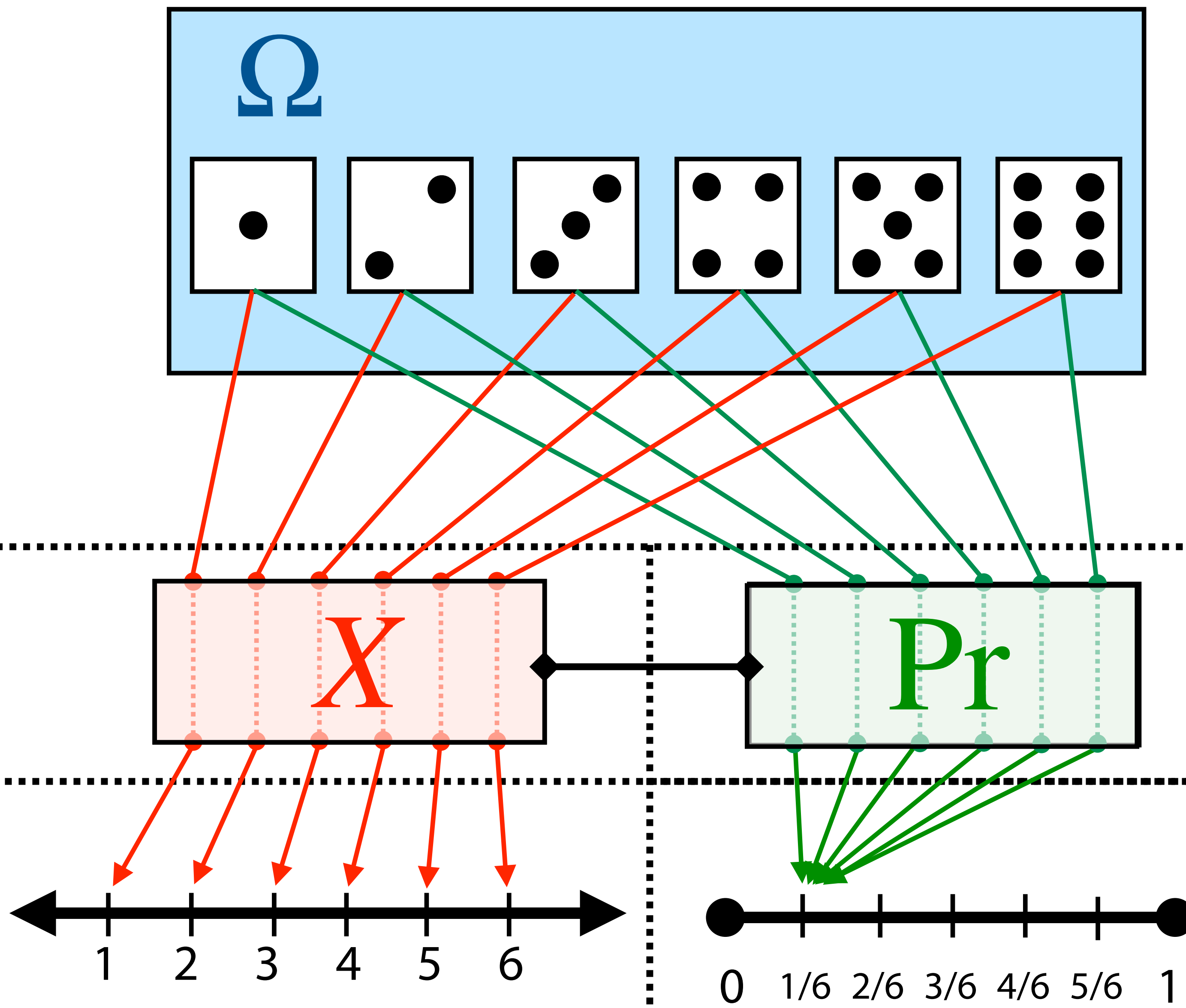
$$\Pr(X = 2) = \Pr(\text{[die face 2]}) = 1/6$$

Random variable & probability function



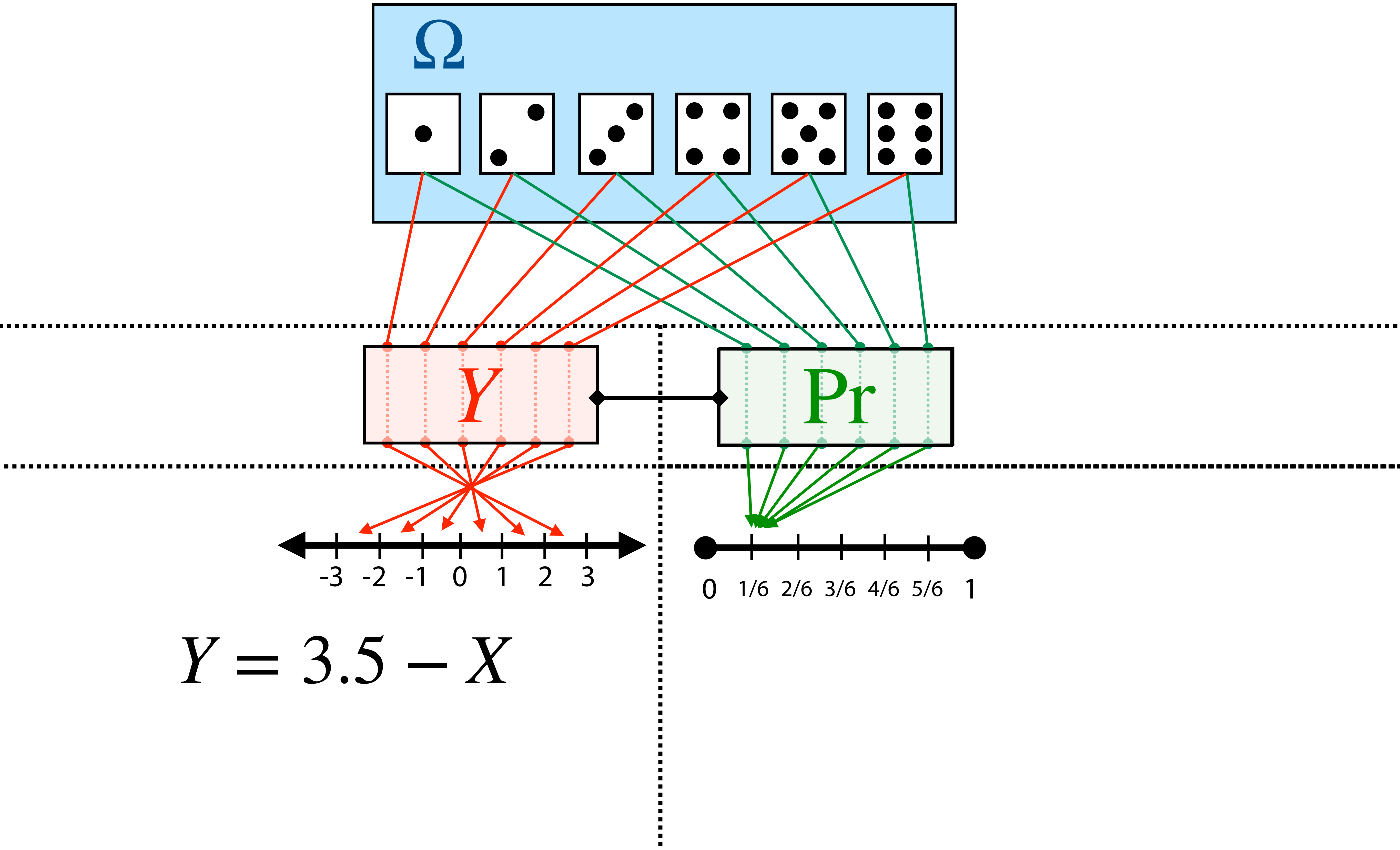
$$\Pr(X \geq 4) = \Pr(\text{⚃}) + \Pr(\text{⚄}) + \Pr(\text{⚅}) = 1/2$$

Random variable & probability function



$$Y = 3.5 - X$$

Random variable & probability function



$$Y = 3.5 - X$$

Expected value

Expectation ("expected value") of a discrete r.v. X, called $\mathbf{E}[X]$, is given by

$$\mathbf{E}[X] = \sum_{x} x \cdot \Pr(X = x)$$

where summation is over values in range of X.

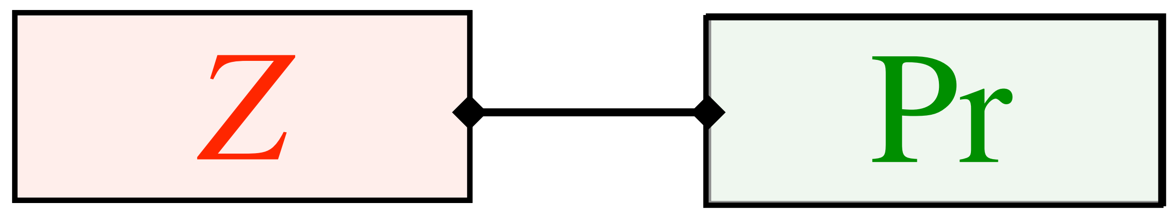
Linearity of expectation

For discrete r.v.s $X_1, X_2, \ldots, X_n$

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

True whether or not X_is are independent

Expected value



$Z = X + Y$ where X is fair die roll & Y is fair coin flip

When Z is a linear combination of other r.v.s, $\mathbf{E}[Z]$ can be easier to get than $\Pr$

$\mathbf{E}[Z] = \mathbf{E}[X] + \mathbf{E}[Y]$ is simple (3.5 + 0.5 = 4)

Hash Table

I have added m items to a n-bucket hash table

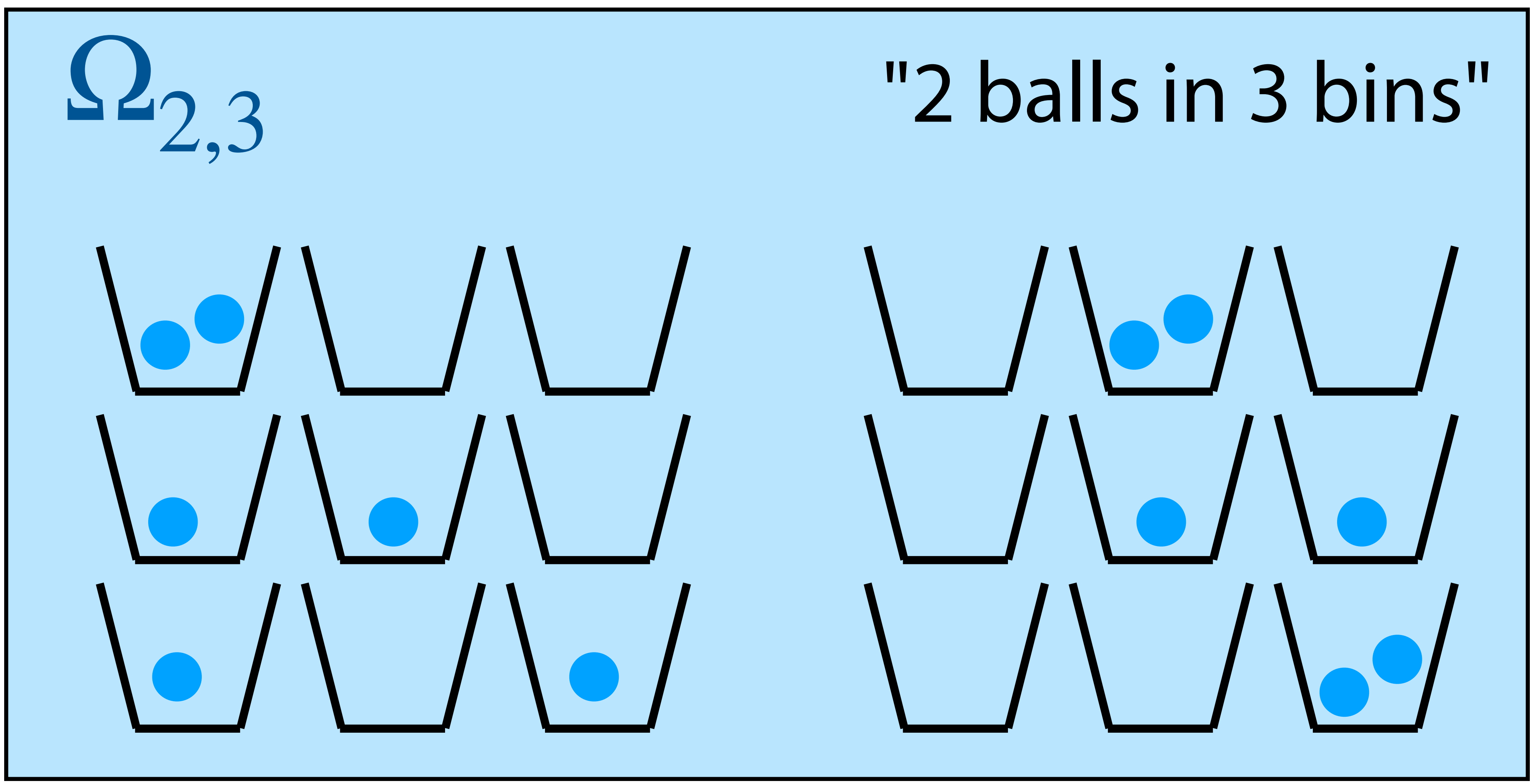
Besides this setup, what else do we need to define a random variable describing the table?

1. Sample space Ω — Possible allocation of items to buckets
2. Probability func. $\Pr$
3. Map X from outcomes to reals

(2 and 3:) Depend on question asked, assumptions made about hash function

Balls & bins

Throw m balls into n bins uniformly and independently



Hash Table

I have added m items to a n-bucket hash table. What "interesting questions" can I ask about the table's state?

How many buckets are empty?

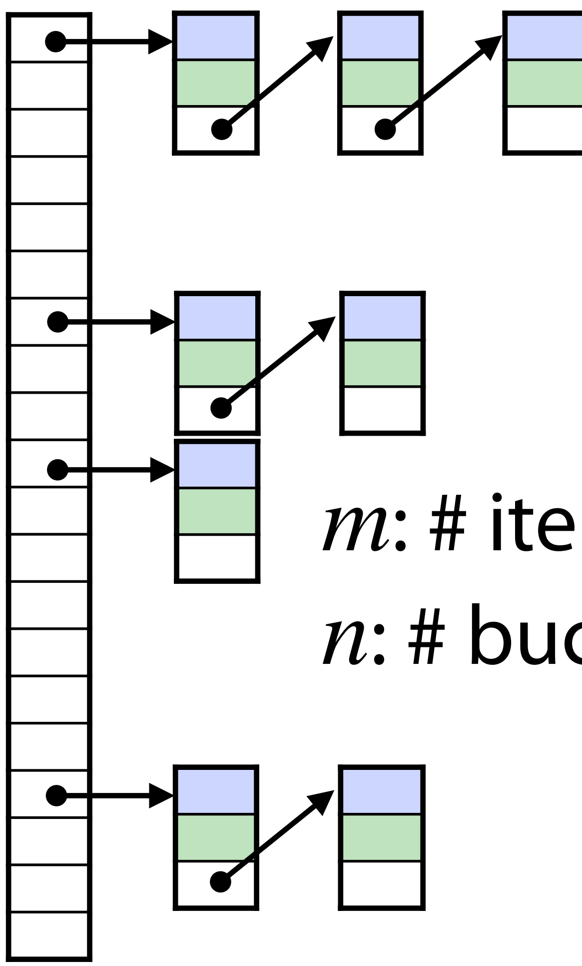
How many items are in the *average* bucket?

How many items are in the *fullest* bucket?

How many items are in the *median* bucket?

What's the chance all buckets are non-empty?

What's the chance no bucket has >1 item?



m: # items
n: # buckets

Balls & bins

I throw m balls into n bins uniformly and independently.
What can I ask about the bins and their contents?

How many bins are empty?

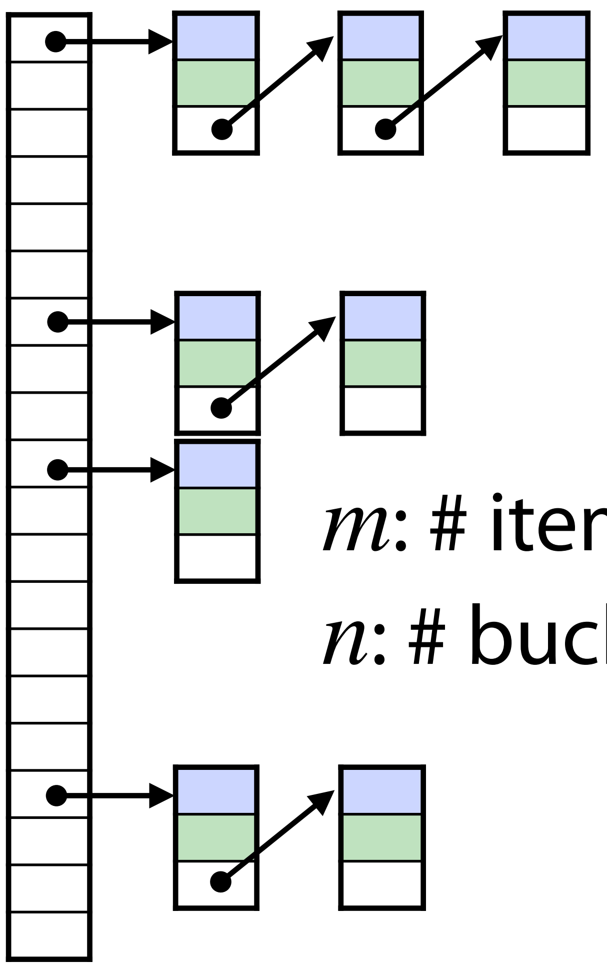
How many balls are in the *median* bin?

How many balls are in the *average* bin?

What's the chance all bins are non-empty?

How many balls are in the *fullest* bin?

What's the chance no bin has >1 item?



m: # items

n: # buckets

Balls & bins

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?	**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, Geometric, Poisson r.v.s**
Global occupancy	What is the *median* bucket occupancy?	What is the *maximum* bucket occupancy?	**Often hard**