Markov Models
Motivation: Casino security

- A gambler is winning too often in a dice game
- We detain him and seize the die being used
- Did he manage to sneak in a rigged die?

A legit die

	Next roll					
	1	2	3	4	5	6
	1/6	1/6	1/6	1/6	1/6	1/6

A legit die

	Next roll					
	1	2	3	4	5	6
	51	48	50	49	52	50

A rigged die

	Next roll					
	1	2	3	4	5	6
	27	35	134	34	31	39

The suspect die

	Next roll					
	1	2	3	4	5	6
	47	51	48	50	49	53

Not the only way to cheat

- Over the years, we've confiscated:
  - Dice that favor a specific number
  - Dice that like to roll the same number in a row
  - Dice that alternate between low and high
  - Dice that like to sum up to even numbers

The same number in a row

- Not a specific number
- 75% chance that the next number will be the same
- 25% chance it rolls another number
    - Each equally likely

- 6, 6, 6, 1, 1, 4, 5, 5, 6, 3, 3, 3, 3, …

Problem: 1/6 still holds

- Die can get stuck on one number for a while
- But if it takes turns getting stuck on each:
  - Over the long run, each still 1/6
- Simple averages don’t detect this

Alternating low and high

- Designed for a game with turns
- If the number is 1, 2, or 3:
  - Next number very likely to be 4, 5, or 6
- And vice versa
- 3, 6, 2, 6, 3, 4, 1, 5, …

Which type of die is it?

- We can use a **Markov Model**
  - A stochastic model that attempts to determine by probability the behavior of a randomized independent process.
- Throw different rigged dice many times, record what they do
- Throw suspect die many times and record its behavior

Training models

- Roll the rigged dice many times
- Measure the probabilities that apply to each die
- Fill them in to some kind of… table?

Testing

- Roll the suspect die many times
- Calculate the probability that each rigged die would have rolled the same sequence
- Choose the most likely match

A specific number

	Next roll					
	1	2	3	4	5	6
	5%	5%	75%	5%	5%	5%

What type of table?

- For this die, probabilities weren't affected by previous rolls
    - One-dimensional table
- For our other rigged dies, they depend upon the (one) immediately previous roll
    - More generally, a Markov Model attempts to explain a random process that depends on the current event but not on previous events.

Two-dimensional table

- Columns are still the next roll
- Rows are the *previous* roll
- This matrix is known as the Transition Probability Matrix

The same number in a row

		Next roll					
		1	2	3	4	5	6
Prev roll	1						
	2						
	3						
	4						
	5						
	6						

The same number in a row

		Next roll					
		1	2	3	4	5	6
Prev roll	1	77	4	6	3	5	5
	2	6	72	5	6	5	6
	3	4	6	74	5	5	6
	4	5	5	6	75	4	5
	5	6	5	4	5	73	7
	6	6	5	4	3	6	76

The same number in a row

		Next roll					
		1	2	3	4	5	6
Prev t roll	1	75%	5%	5%	5%	5%	5%
	2	5%	75%	5%	5%	5%	5%
	3	5%	5%	75%	5%	5%	5%
	4	5%	5%	5%	75%	5%	5%
	5	5%	5%	5%	5%	75%	5%
	6	5%	5%	5%	5%	5%	75%

If we had only used 1-D

	Next roll					
	1	2	3	4	5	6
	1/6	1/6	1/6	1/6	1/6	1/6

Alternating low and high

		Next roll					
		1	2	3	4	5	6
Prev roll	1	8%	8%	8%	25%	25%	25%
	2	8%	8%	8%	25%	25%	25%
	3	8%	8%	8%	25%	25%	25%
	4	25%	25%	25%	8%	8%	8%
	5	25%	25%	25%	8%	8%	8%
	6	25%	25%	25%	8%	8%	8%

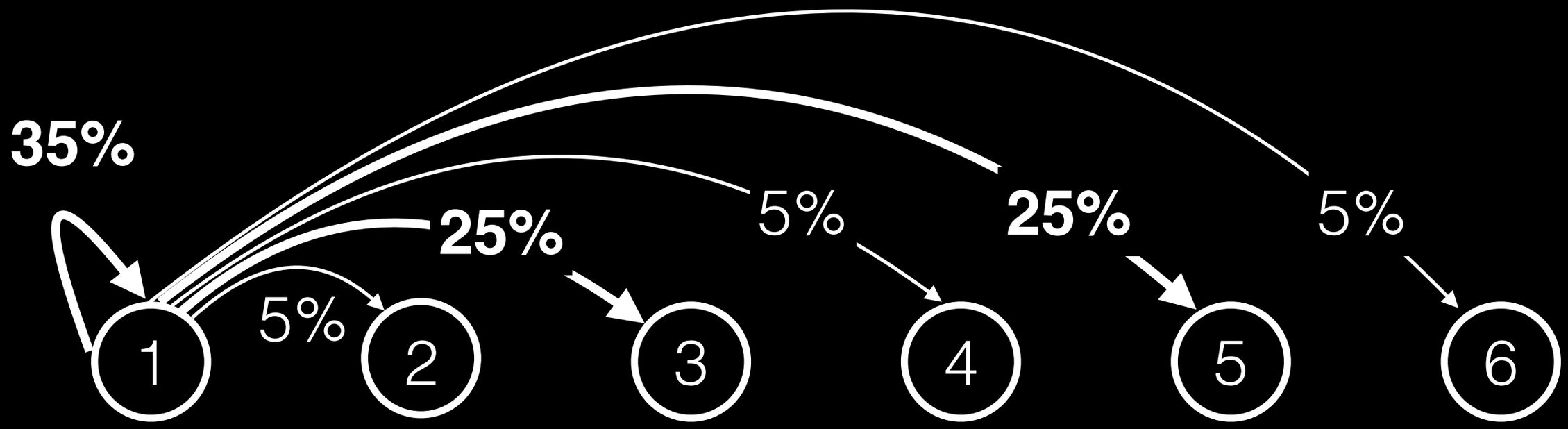
A weird one

		Next roll					
		1	2	3	4	5	6
Prev roll	1	12%	7%	19%	34%	8%	20%
	2	…	…	…	…	…	…
	3	…	…	…	…	…	…
	4	…	…	…	…	…	…
	5	…	…	…	…	…	…
	6	…	…	…	…	…	…

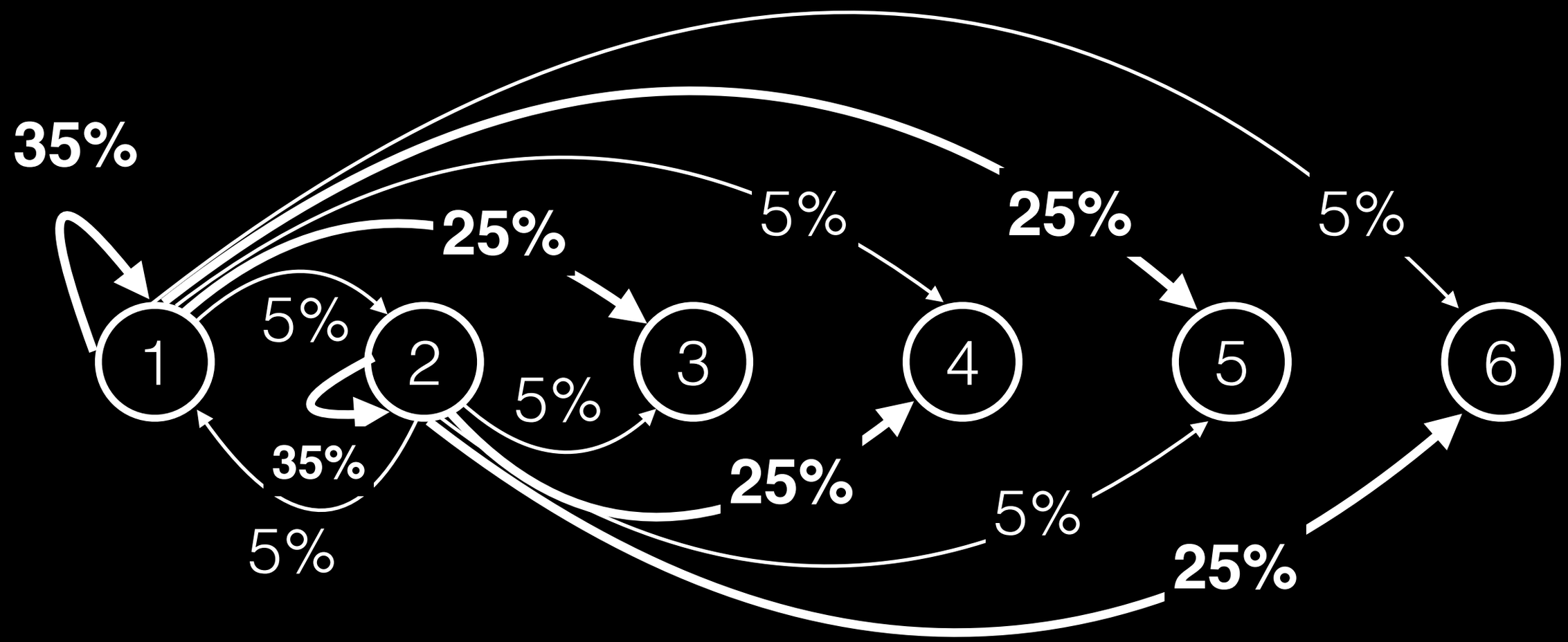
A specific number

		Next roll					
		1	2	3	4	5	6
Prev roll	1	5%	5%	75%	5%	5%	5%
	2	5%	5%	75%	5%	5%	5%
	3	5%	5%	75%	5%	5%	5%
	4	5%	5%	75%	5%	5%	5%
	5	5%	5%	75%	5%	5%	5%
	6	5%	5%	75%	5%	5%	5%

Another way to draw it…



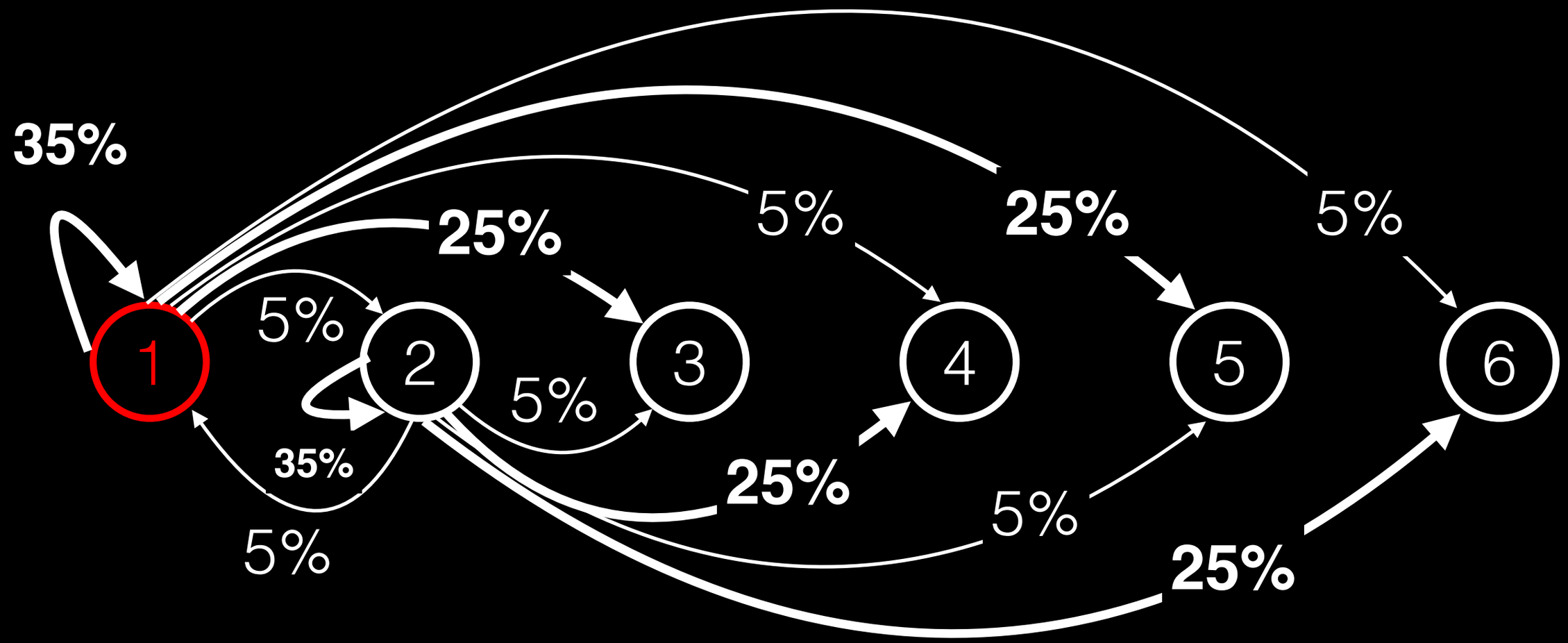
Another way to draw it…



Using these models

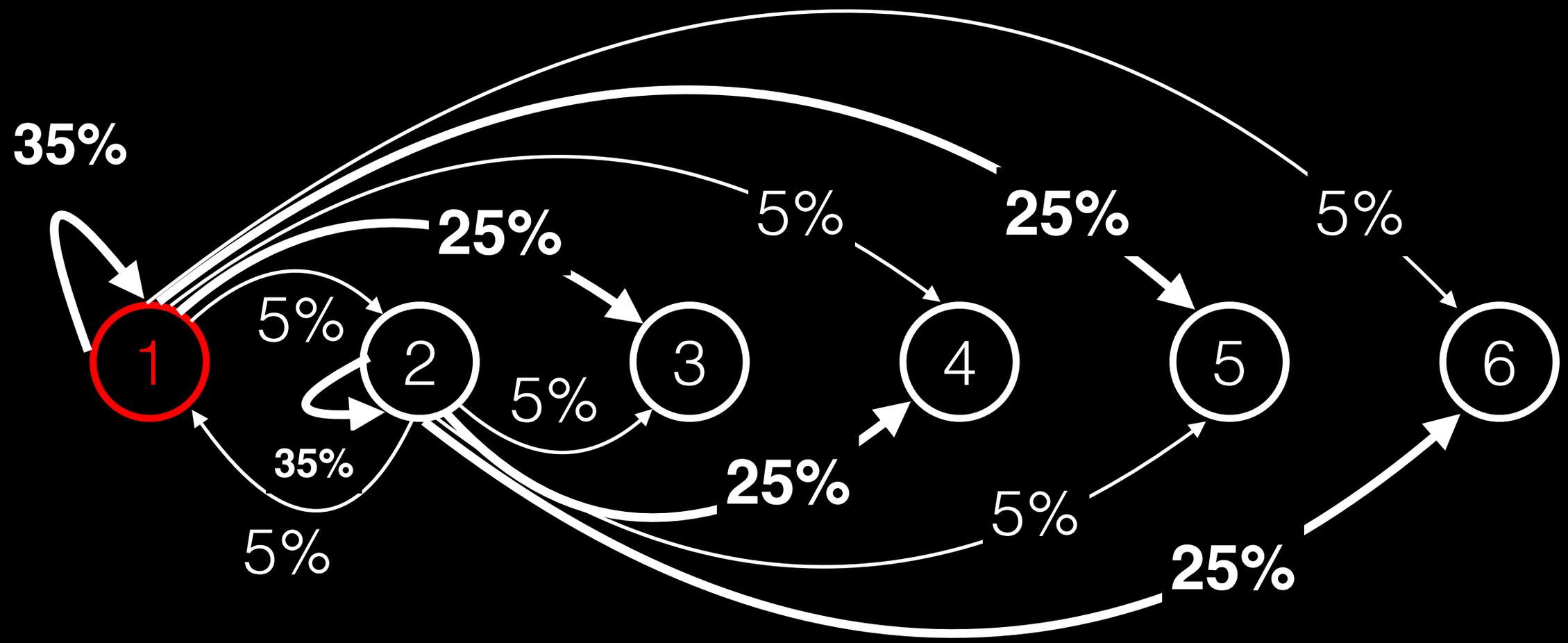
- Simulate the behavior of a die
- Compare a real die to the modeled die
- Mathematically analyze the behavior

Simulating this die



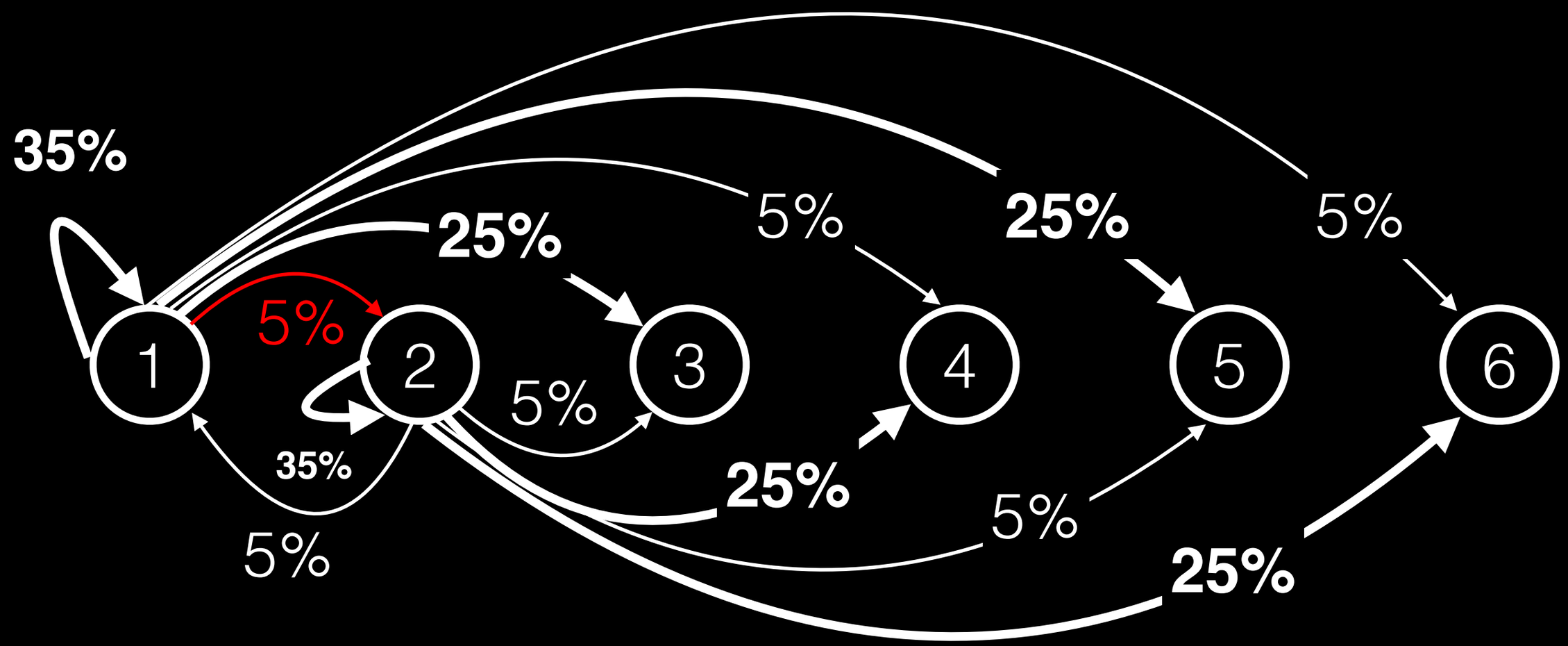
Randomly choose first roll. First = 1.

Simulating this die



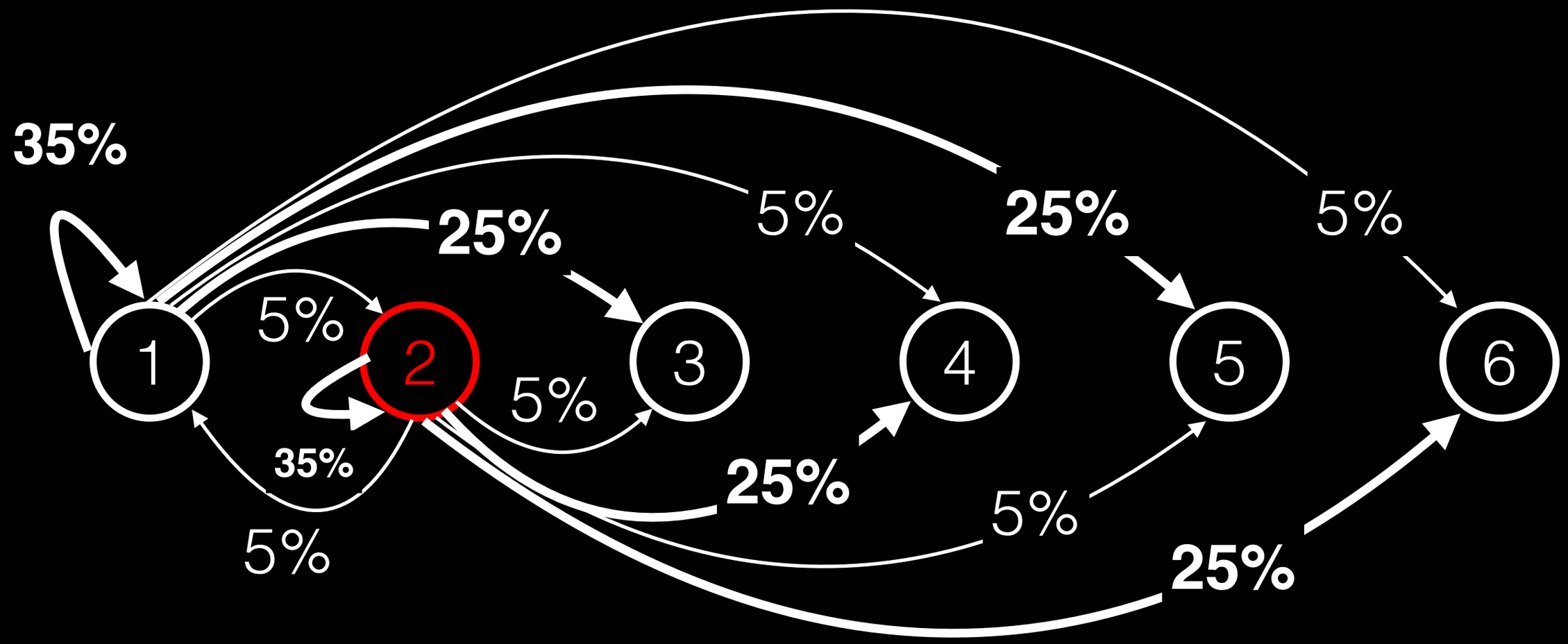
Choose next state based on random

Simulating this die



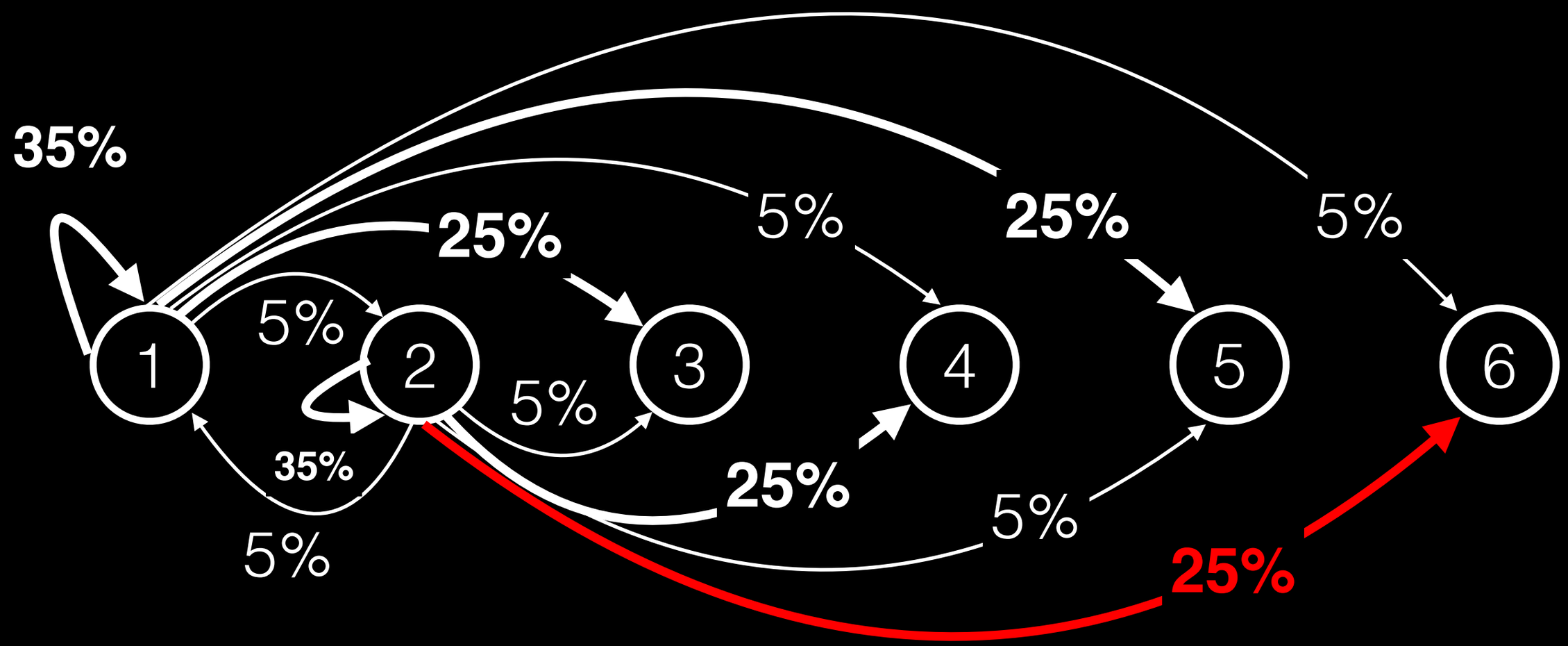
Randomly choose 2. Next roll = 2.

Simulating this die



Choose next state based on random

Simulating this die



Randomly choose 6. Next roll = 6…

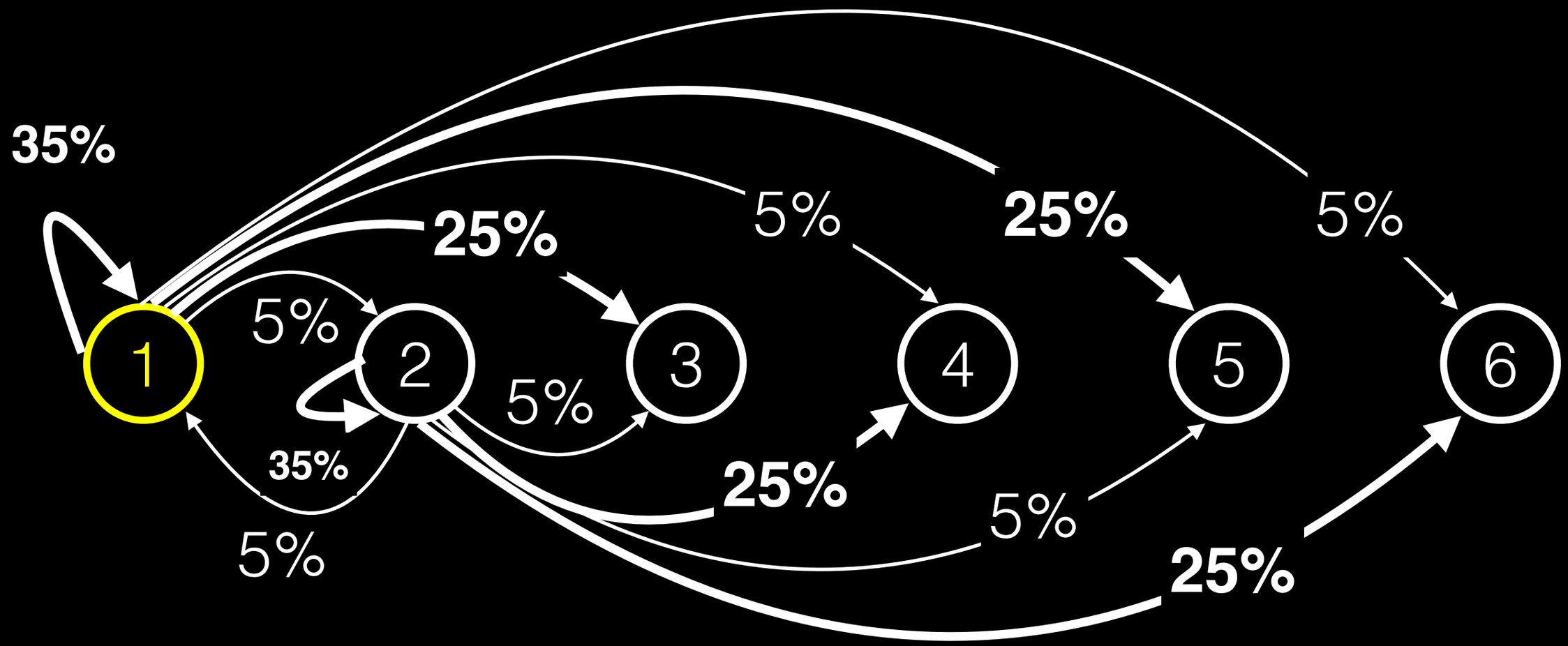
Simulated sequence

- 1, 2, 6, ...

Likelihood of a sequence

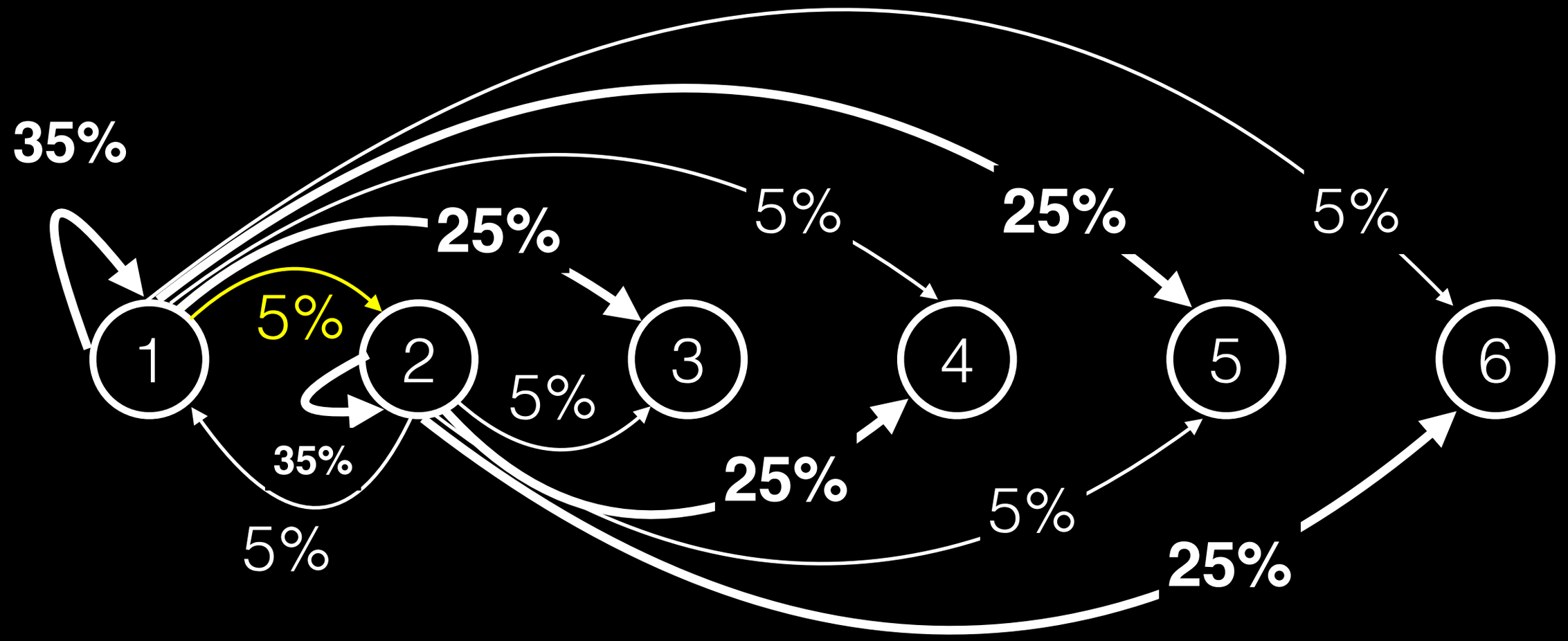
- 1, 2, 1, 3

Walk through a series of rolls



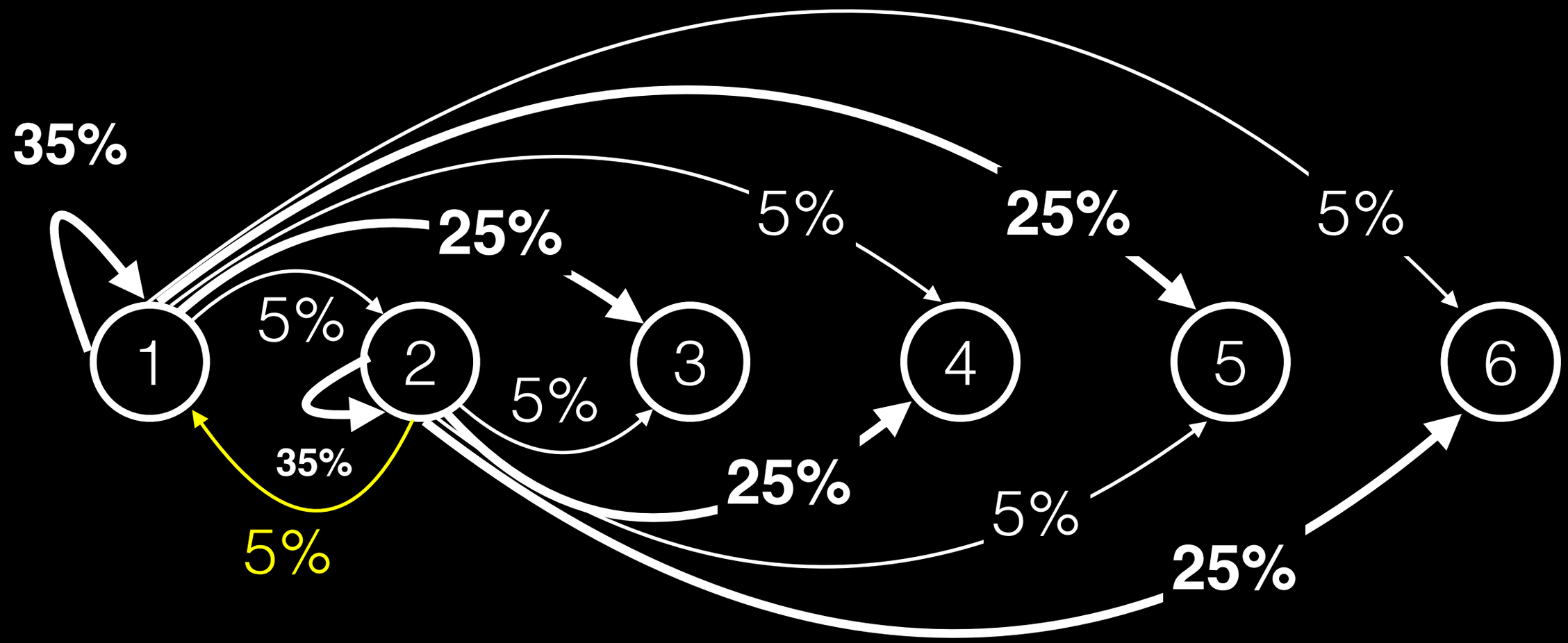
First roll is a 1

Walk through a series of rolls



Next roll is a 2
Probability of transition = 5%
Probability of sequence= 5%

Walk through a series of rolls

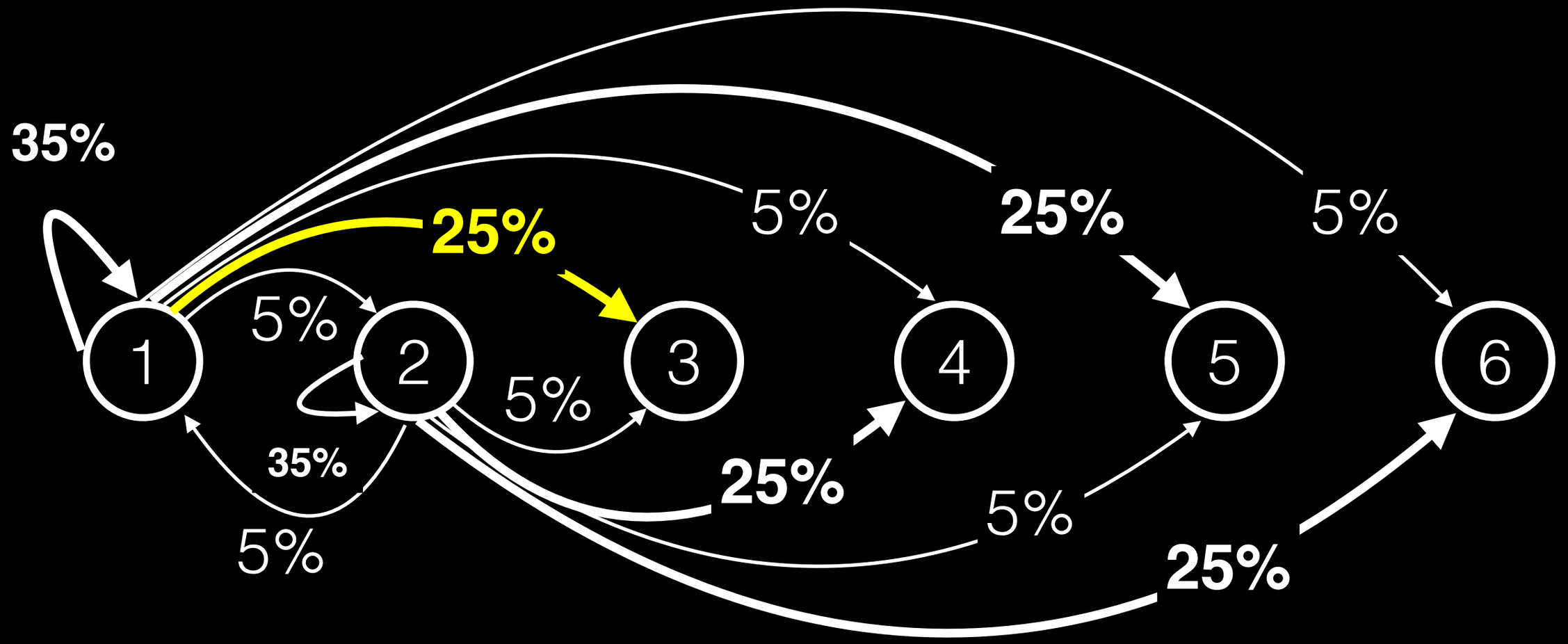


Next roll is a 1

Probability of transition = 5%

Probability of sequence= 5% x 5% = 0.25%

Walk through a series of rolls



Next roll is a 3

Probability of transition = 25%

Prob. of seq.= 0.25% x 25% = 0.0625%

Which model fits best?

Model	Prob
Specific number	0.0625%
Same in a row	**0.1600%**
Alternating low high	0.0512%
Sum to even	0.0625%

Patterns in English text

- Markov models over sequences of words or letters
- This video: Generate sequences of words
- PA #1: Recognize sequences of letters
  - Attribute to specific speaker

Generating word sequences

		Next word					
		the	is	that	to	be	…
Prev word	the	0	0	0	0	0	…
	is	17	1	14	12	2	…
	that	45	23	12	7	2	…
	to	37	0	9	0	8	…
	be	44	4	25	13	0	…
	…	…					

Data representation?

- Dictionary of dictionaries?
  - counts[“to”][“be”] = 8

Dictionary of dictionaries

the	question = 1, answer = 3, whether = 9
is	to = 7, that = 6, not = 14
that	…
to	…
be	…

Required operations

- Adding a new word to the table
- Incrementing the frequency of a word pair
- Randomly choosing the next word based on the previous word, using probabilities

Required operations

- Adding a new word to the table: **CHEAP**
- Incrementing frequency of a word pair: **CHEAP**
- Randomly choosing the next word based on the previous word, using probabilities: **EXPENSIVE**

Improving random selection

- Need a data structure that lets us “throw dart” and pick the word we hit

- Lists: cheap lookup for specific index

  - But how to bias the probabilities?

Duplicate entries

- Words can appear multiple times in list
  - Appear more times → more likely to be randomly chosen

- In list that corresponds to specific previous word:
- Words appear once per occurrence in training text
  - Probability of selection is:
    num occurrences / num times preceding word appeared

Hash table of lists

the	question, answer, question
is	not, to, to, that, not, to, not, that
that	…
to	…
be	…

Required operations

- Adding a new word to the table: **?**
- Incrementing frequency of a word pair: **?**

- Randomly choosing the next word based on the previous word, using probabilities: **CHEAP**

Killing three birds

- Adding new word: just append to end
- Adding another occurrence: just append to end

Required operations

- Adding a new word to the table: **CHEAP**
- Incrementing frequency of a word pair: **CHEAP**

- Randomly choosing the next word based on the previous word, using probabilities: **CHEAP**

```
class TextGen:
    def __init__(self,words):
        self._nextwords = {}
        self.learn(words)
```

```
    def learn(self,words):
        prev = words[-1]  # wrap around
        for word in words:
            if prev in self._nextwords:
                self._nextwords[prev].append(word)
            else:
                self._nextwords[prev] = [word]
            prev = word
```

```
def generate(self,nwords):
    # choose random starting word
    start_words = self._nextwords.keys()
    prev = start_words[rand.randint(0, len(start_words) - 1)]
    rv = prev + " "
    for i in range(nwords):
        next_words = self._nextwords.get(prev, None)
        if next_words is None:
            # last word in training data
            # and no information on possible successors
            return rv
        next_word = next_words[rand.randint(0, len(next_words) - 1)]
        rv = rv + next_word + " "
        prev = next_word
    return rv[:-1]
```

Presidential debates

- “… certainty. It's in Kosovo, supporting me. Why? Because that's the Internet to use American troops down, and where the authority wisely….”

- “… ports. And we have been doing a while. And I meet with insurance that money out there. The middle-class need more…”

- “… campaign contributions. Senator Obama has not in the fundamental difference. Tragically, I know our rhetoric. That's what's the people, and I…”

What type of table?

- For simple die, probabilities weren't affected by previous rolls
  - One-dimensional table
- For our other rigged dies, they depend upon the (one) immediately previous roll
  - Two-dimensional table
- Suppose it depends upon the last *two* rolls?

3-D without glasses

		Next roll					
		1	2	3	4	5	6
Prev rolls	1,1	25%	8%	25%	8%	25%	8%
	1,2	8%	25%	8%	25%	8%	25%
	1,3	25%	8%	25%	8%	25%	8%
	…	…					
	2,1	25%	8%	25%	8%	25%	8%
	…	…					

Text generation

is the	question answer weather problem
to be	is or what that someone something
be that	…
in that	…
in the	…

```
    def learn(self,words):
        prev = words[-self._k:]  # wrap around
        for word in words:
            prev_str = " ".join(prev)
            if prev_str in self._nextwords:
                self._nextwords[prev_str].append(word)
            else:
                self._nextwords[prev_str] = [word]
            prev = prev[1:]
            prev.append(word)
```

```
def generate(self, nwords):
    # choose random starting words
    start_words = list(self._nextwords.keys())
    prev = start_words[rand.randint(0, len(start_words) - 1)].split()
    rv = " ".join(prev) + " "
    for i in range(nwords):
        prev_str = " ".join(prev)
        next_words = self._nextwords.get(prev_str, None)
        if next_words is None:
            # no information on possible successors
            return rv
        next_word = next_words[rand.randint(0, len(next_words) - 1)]
        rv = rv + next_word + " "
        prev = prev[1:]
        prev.append(next_word)
    return rv[:-1]
```

3rd order example

- “… is that for 30 years, politicians in Washington haven't done anything. What Senator McCain refers to is a measure in the Senate that would try to broaden the mandate inside of Iraq. To deal with Iran. And you know what insurance companies will do? They will find a state -- maybe Arizona, maybe…”

3rd order example

- “… That's why I'm running for president, and I'm hopeful that all of you are prepared to continue this extraordinary journey that we call America. But we're going to the emergency room for treatable illnesses like asthma. And when Senator McCain proposes a $300 billion tax cut, $200 billion of it to the larges…”

3rd order example

- “… been doing that. Let me talk about North Korea. It is naive and dangerous to take a policy that he suggested the other day, which is to have litigation reform. As I told you, we've just got a report that said America is safer but not yet safe. There is more work to…”