HMM Review

Lecture Outline

1. Markov models
2. Hidden Markov models
3. Viterbi algorithm

MARKOV MODELS

One View of Text

- Sequence of symbols (bytes, letters, characters, morphemes, words, …)
  - Let Σ denote the set of symbols.
- Lots of possible sequences. (Σ* is infinitely large.)
- Probability distributions over Σ*?

Trivial Distributions over Σ*

- Give probability 0 to sequences with length greater than B; uniform over the rest.
- Use data: with N examples, give probability N^{-1} to each observed sequence, 0 to the rest.

- What if we want *every* sequence to get some probability?
  - Need a probabilistic *model family* and algorithms for constructing the model from *data*.

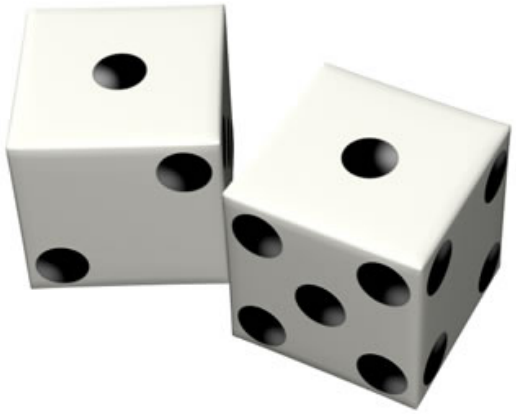
A History-Based Model

$$p(\text{start}, w_1, w_2, \ldots, w_n, \text{stop}) = \prod_{i=1}^{n+1} \gamma(w_i \mid w_1, w_2, \ldots, w_{i-1})$$

- Generate each word from left to right, conditioned on what came before it.

Die / Dice

one die

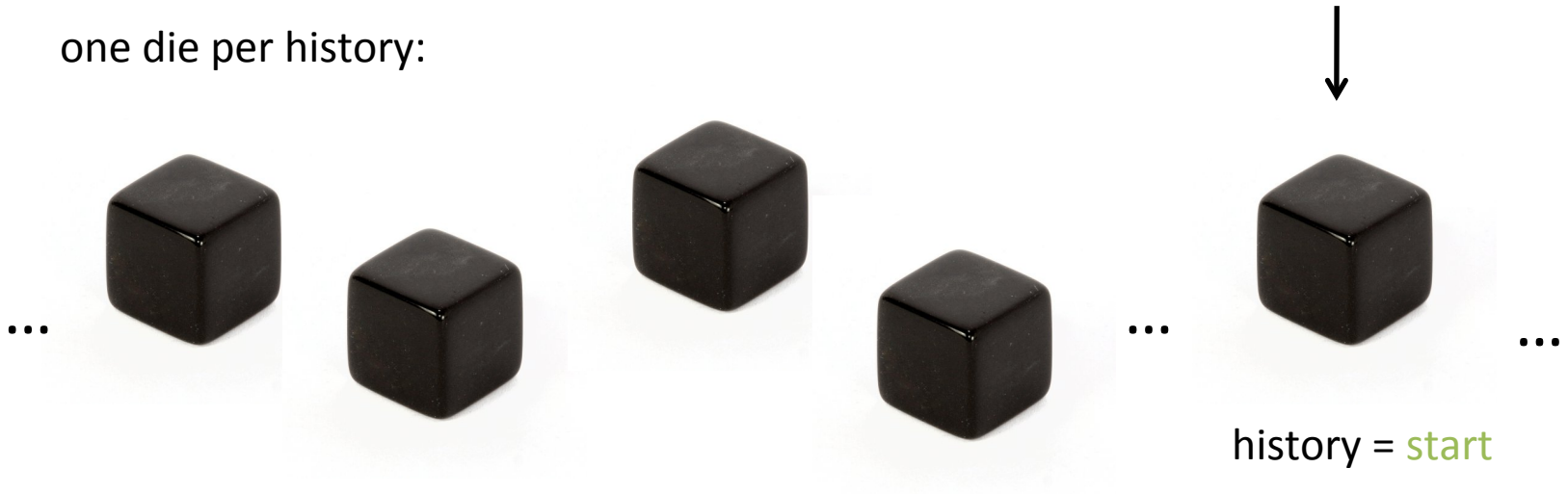


two dice

start

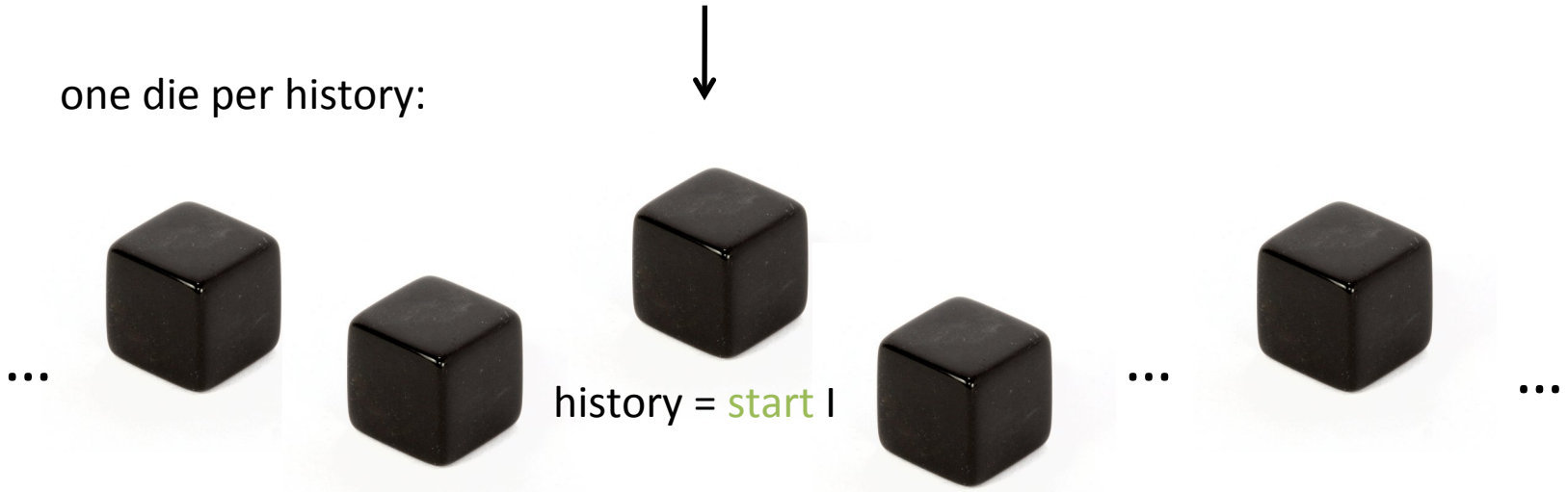
one die per history:

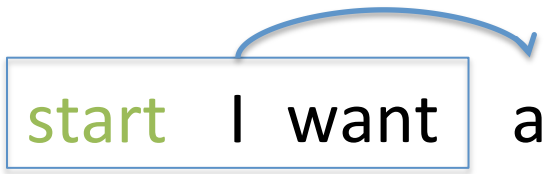
one die per history:



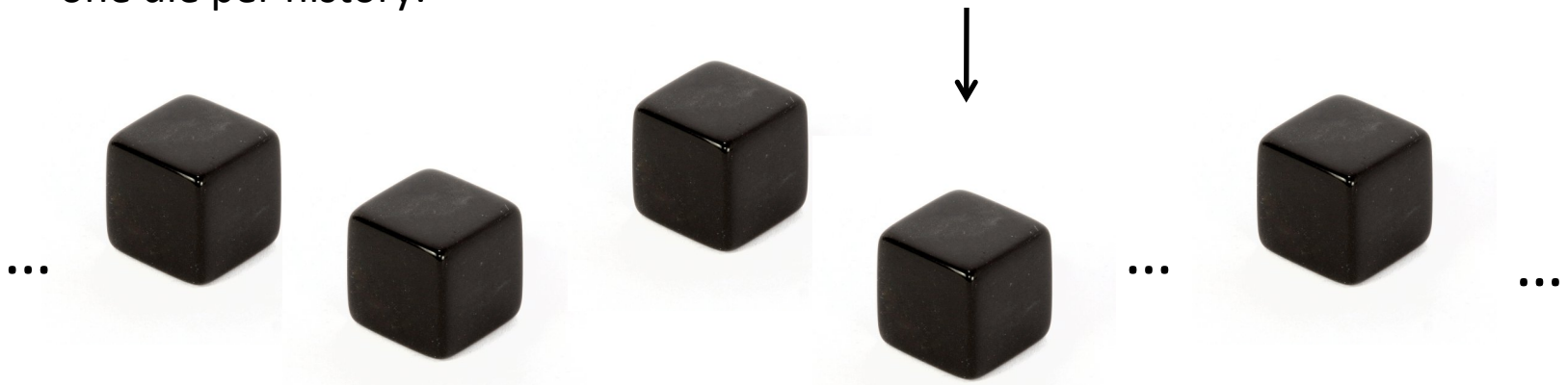
history = start

one die per history:





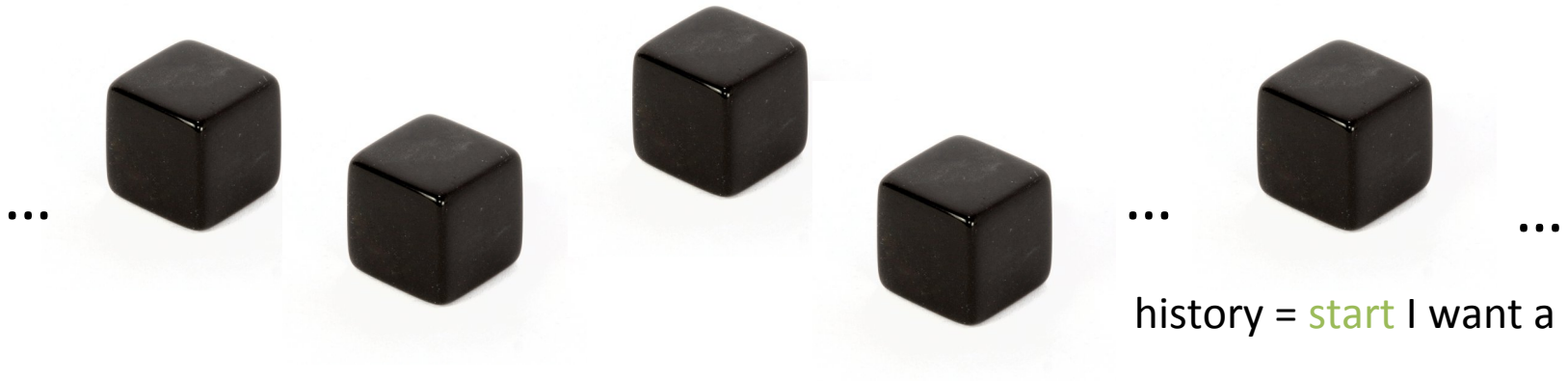
one die per history:



history = start I want

start I want a flight

one die per history:



history = start I want a

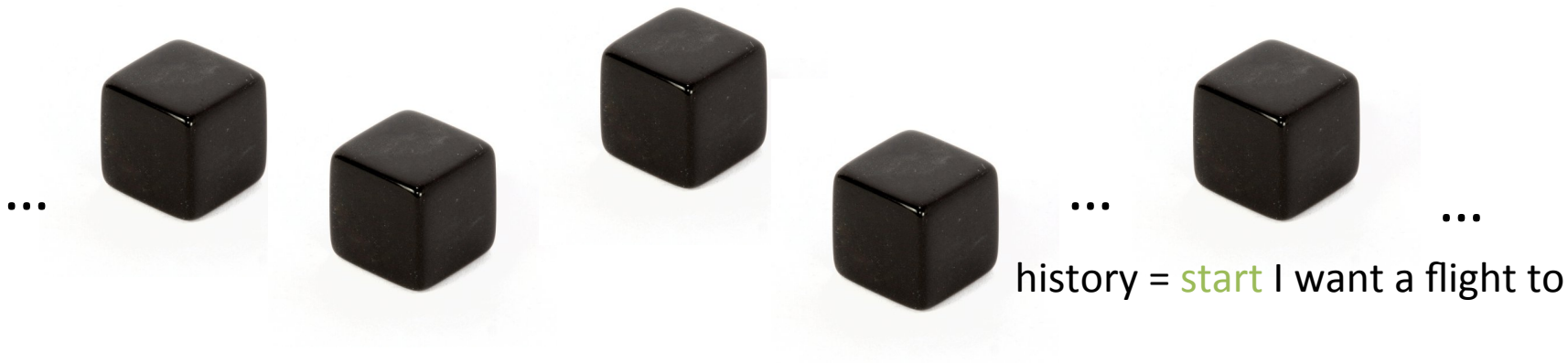
start I want a flight to

one die per history:

history = start I want a flight

start I want a flight to Lisbon

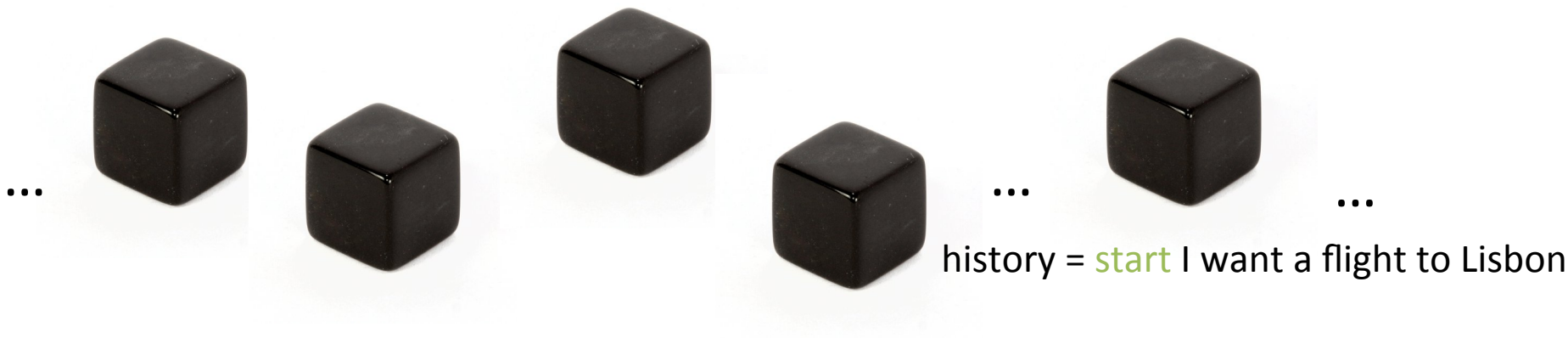
one die per history:



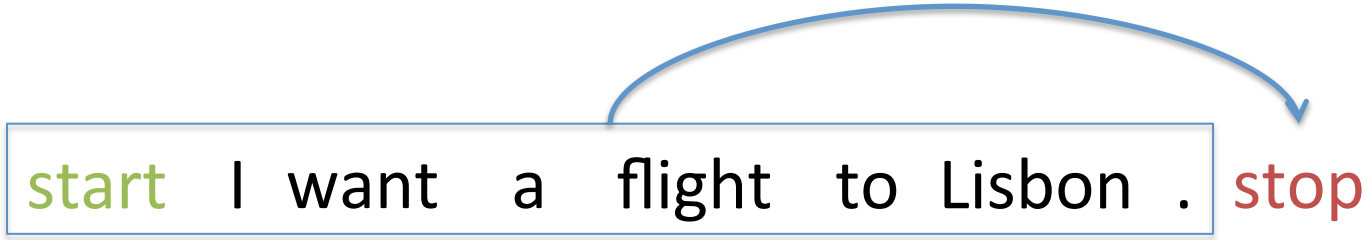
history = start I want a flight to

start I want a flight to Lisbon .

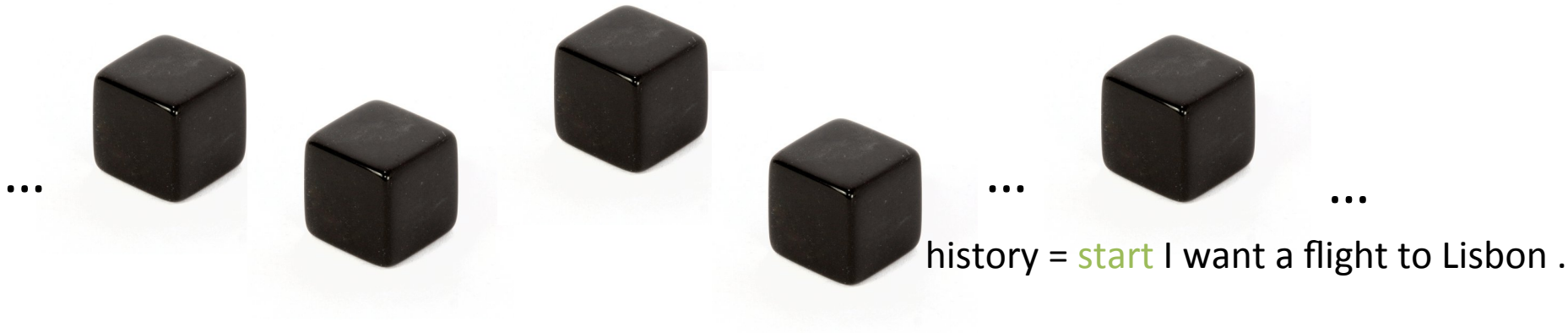
one die per history:



history = start I want a flight to Lisbon



one die per history:



history = start I want a flight to Lisbon .

A History-Based Model

$$p(\text{start}, w_1, w_2, \ldots, w_n, \text{stop}) = \prod_{i=1}^{n+1} \gamma(w_i \mid w_1, w_2, \ldots, w_{i-1})$$

- Generate each word from left to right, conditioned on what came before it.
- Very rich representational power!
- How many parameters?
- What is the probability of a sentence not seen in training data?

A Bag of Words Model

$$p(\text{start}, w_1, w_2, \ldots, w_n, \text{stop}) = \prod_{i=1}^{n+1} \gamma(w_i)$$

- Every word is independent of every other word.

start

one die:

start

one die:

start I want

one die:

start I want a

one die:

start I want a flight

one die:

start I want a flight to

one die:

start I want a flight to Lisbon

one die:

start I want a flight to Lisbon .

one die:

start I want a flight to Lisbon . stop

one die:

A Bag of Words Model

$$p(\text{start}, w_1, w_2, \ldots, w_n, \text{stop}) = \prod_{i=1}^{n+1} \gamma(w_i)$$

- Every word is independent of every other word.
- Strong assumptions mean this model cannot fit the data very closely.
- How many parameters?
- What is the probability of a sentence not seen in training data?

First Order Markov Model

- Happy medium?

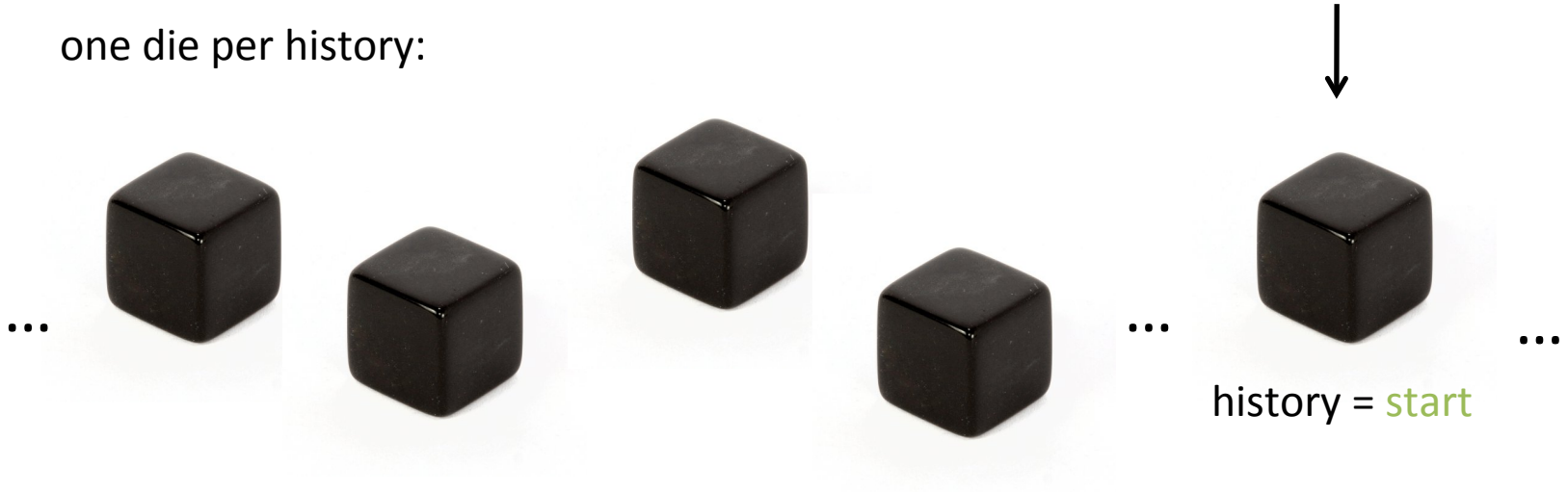
$$p(\text{start}, w_1, w_2, \ldots, w_n, \text{stop}) = \prod_{i=1}^{n+1} \gamma(w_i \mid w_{i-1})$$

- Condition on the most recent symbol in history.

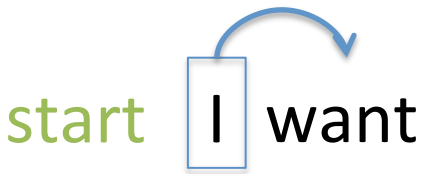
start

one die per history:

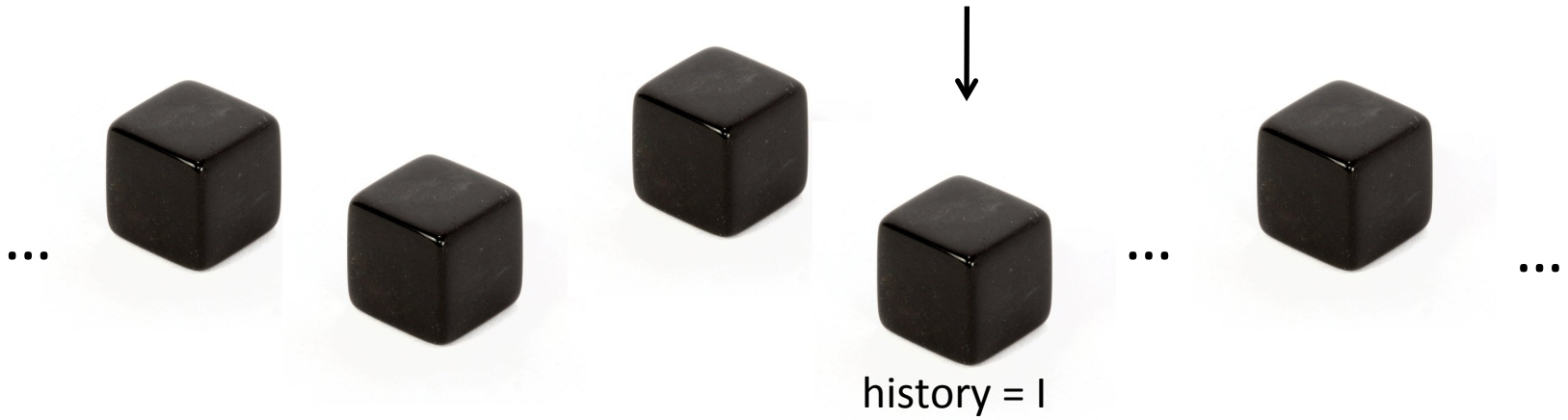
one die per history:

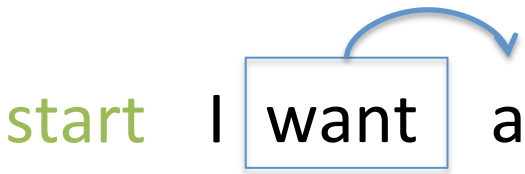


history = start

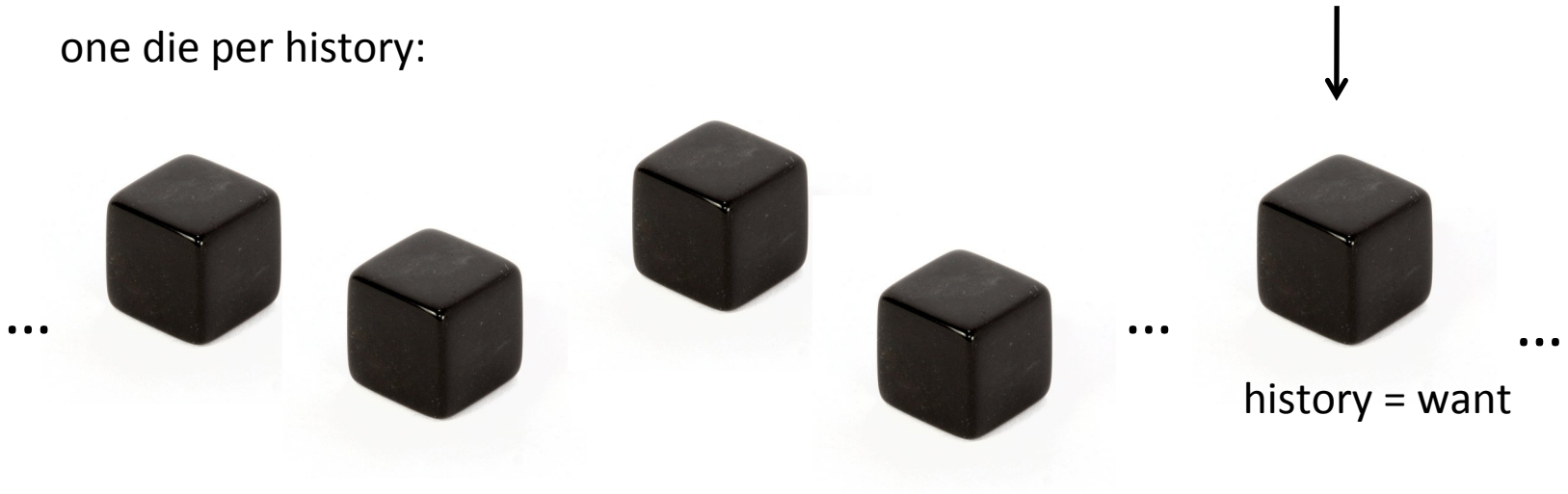


one die per history:



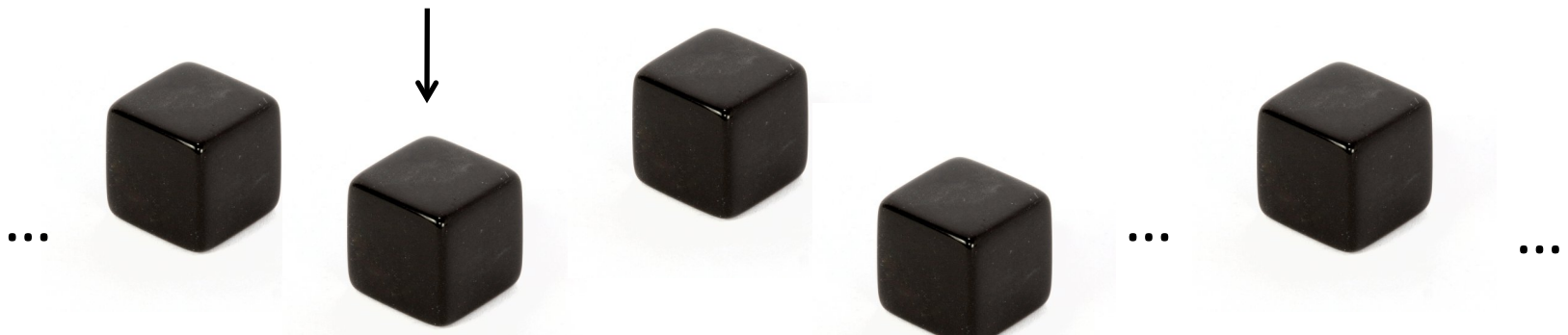


one die per history:



start I want a flight

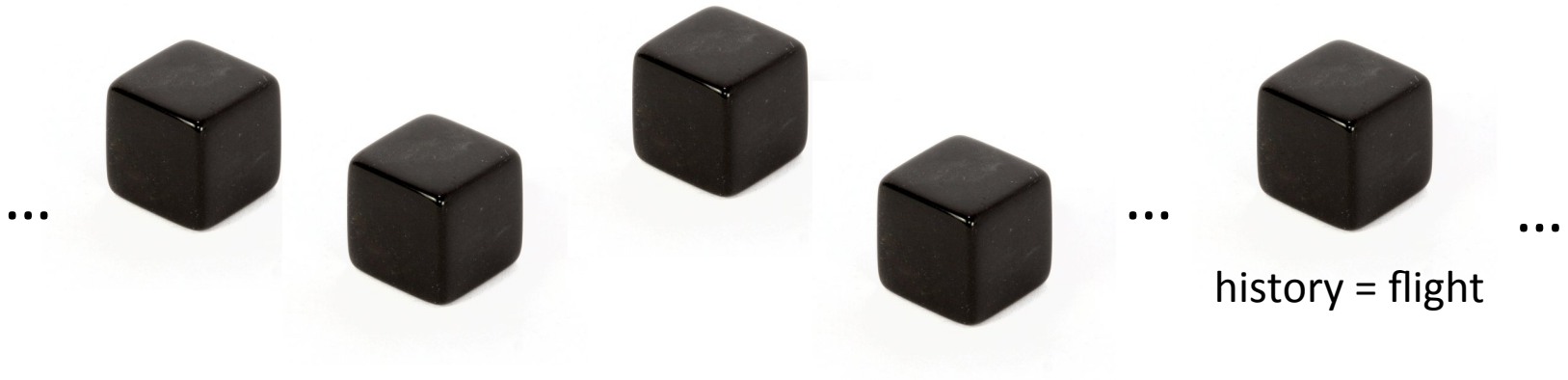
one die per history:



history = a

start I want a flight to

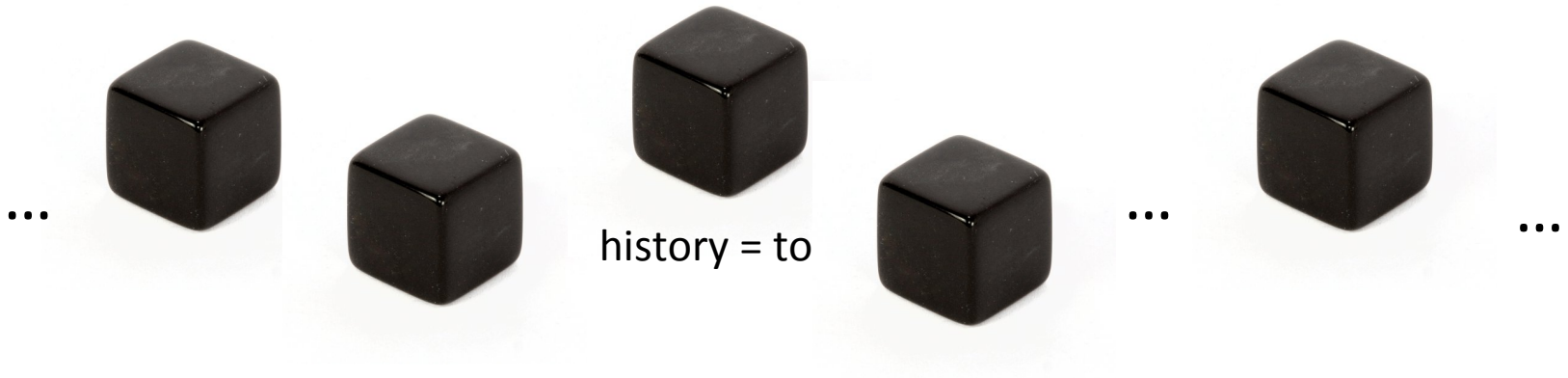
one die per history:



history = flight

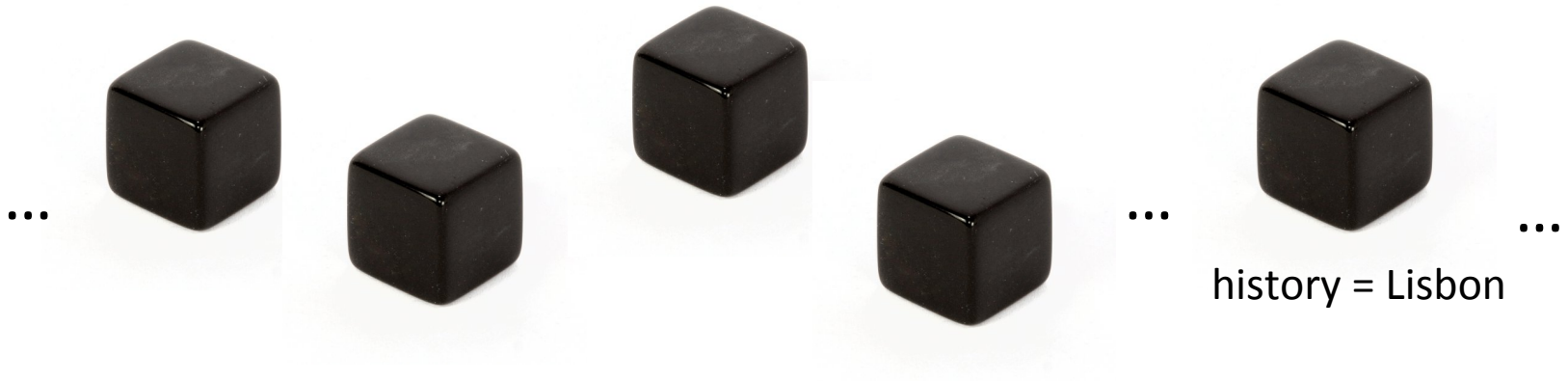
start I want a flight to Lisbon

one die per history:



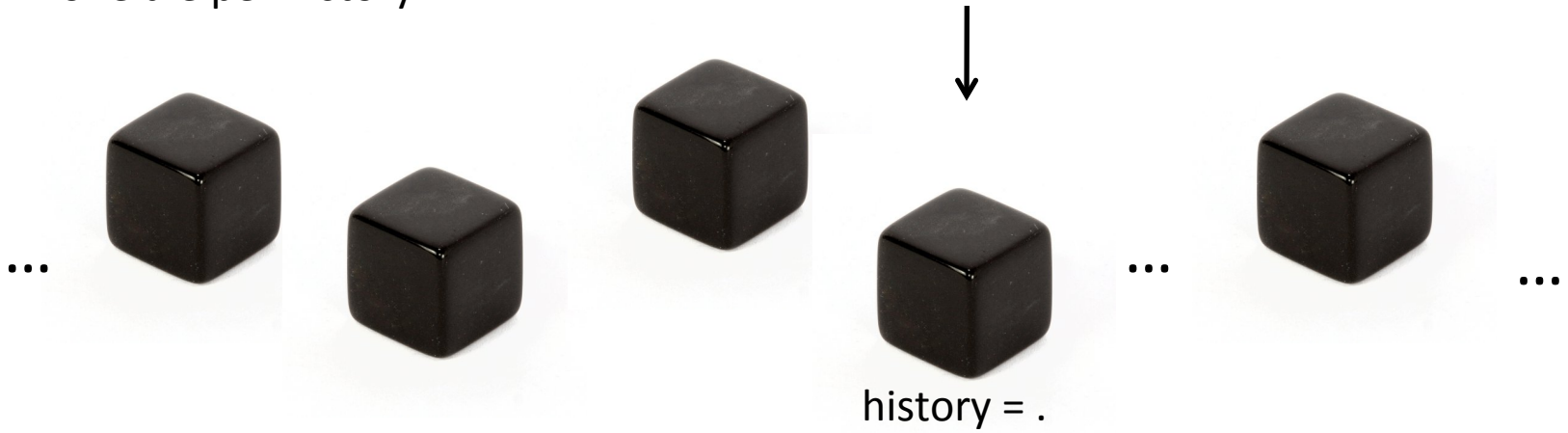
start I want a flight to Lisbon .

one die per history:



start I want a flight to Lisbon . stop

one die per history:



First Order Markov Model

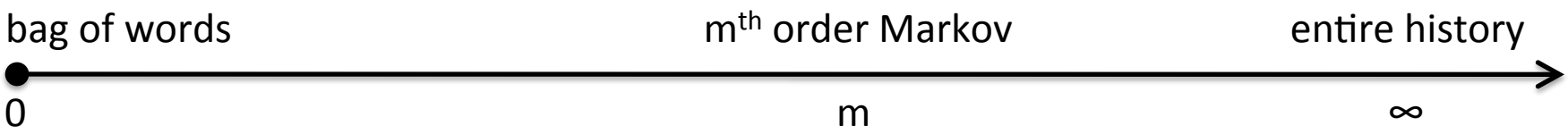
- Happy medium?

$$p(\text{start}, w_1, w_2, \ldots, w_n, \text{stop}) = \prod_{i=1}^{n+1} \gamma(w_i \mid w_{i-1})$$

- Condition on the most recent symbol in history.
- Independence assumptions?
- Number of parameters?
- Sentences not seen in training?

mth Order Markov Models

$$p(\text{start}, w_1, w_2, \ldots, w_n, \text{stop}) = \prod_{i=1}^{n+1} \gamma(w_i \mid w_{i-m}, \ldots, w_{i-1})$$



fewer parameters

stronger independence assumptions

richer expressive power

Example

- Unigram model estimated on 2.8M words of American political blog text.

```
this trying our putting and funny
 and among it herring it obama
 but certainly foreign my
 c on byron again but from i
 i so and i chuck yeah the as but but republicans if
this stay oh so or it mccain bush npr this with what
and they right i while because obama
```

Example

- Bigram model estimated on 2.8M words of American political blog text.

```
the lack of the senator mccain hadn t keep this story
backwards
while showering praise of the kind of gop weakness
it was mistaken for american economist anywhere in the
white house press hounded the absence of those he s as
a wide variety of this election day after the candidate
b richardson was polled ri in hempstead moderated by
the convention that he had zero wall street journal
argues sounds like you may be the primary
but even close the bill told c e to take the obama on
the public schools and romney
fred flinstone s see how a lick skillet road it s
little sexist remarks
```

Example

- Trigram model estimated on 2.8M words of American political blog text.

```
 as i can pin them all none of them want to bet that
any of the might be
 conservatism unleashed into the privacy rule book and
when told about what paul
 fans organized another massive fundraising initiative
yesterday and i don t know what the rams supposedly
want ooh
 but she did but still victory dinner
 alone among republicans there are probably best not
all of the fundamentalist community
 asked for an independent maverick now for
crystallizing in one especially embarrassing
```

Example

- 5-gram model estimated on 2.8M words of American political blog text.

```
 he realizes fully how shallow and insincere
conservative behavior has been he realizes that there
is little way to change the situation
 this recent arianna huffington item about mccain
issuing heartfelt denials of things that were actually
true or for that matter about the shia sunni split and
which side iran was on would get confused about this
any more than someone with any knowledge of us politics
would get confused about whether neo confederates were
likely to be supporting the socialist workers party
 at the end of the world and i m not especially
discouraged now that newsweek shows obama leading by
three now
```

Example

- 100-gram model estimated on 2.8M words of American political blog text.

```
 and it would be the work of many hands to catalogue all
the ridiculous pronouncements made by this man since his
long train of predictions about the middle east has been
gaudily disastrously stupefyingly misinformed just the
buffoon it seems for the new york times to award with a
guest column for if you object to the nyt rewarding
failure in quite this way then you re intolerant according
to the times editorial page editor andrew rosenthal
rosenthal doesn t seem to recognize that his choice of
adjectives to describe kristol serious respected are in
fact precisely what is at issue for those whom he
dismisses as having a fear of opposing views
```

N-Gram Models

Pros

- Easily understood **linguistic formalism**.
- Fully generative.
- Algorithms:
  - calculate probability of a sequence
  - choose a sequence from a set
  - training

Cons

- Obviously inaccurate linguistic formalism.
- As N grows, data sparseness becomes a problem.
  - Smoothing is a black art.
- How to deal with unknown words?