Hidden Markov Model

Ramaseshan Ramachandran

Markov Assumption

- Let us consider a sequence of state variables $q_1, q_2, \ldots, q_i$.
- The future state is predicted based only on the present state - the other past states are not required
- In general, Markov assumption simplifies $P(q_i = a \mid q_1 \ldots q_{i-1})$ into $P(q_i = a \mid q_{i-1})$
- $P(q_i = a \mid q_{i-1})$ is the familiar bigram language model
- Markov chain uses the bigram LM assumption

Markov chain

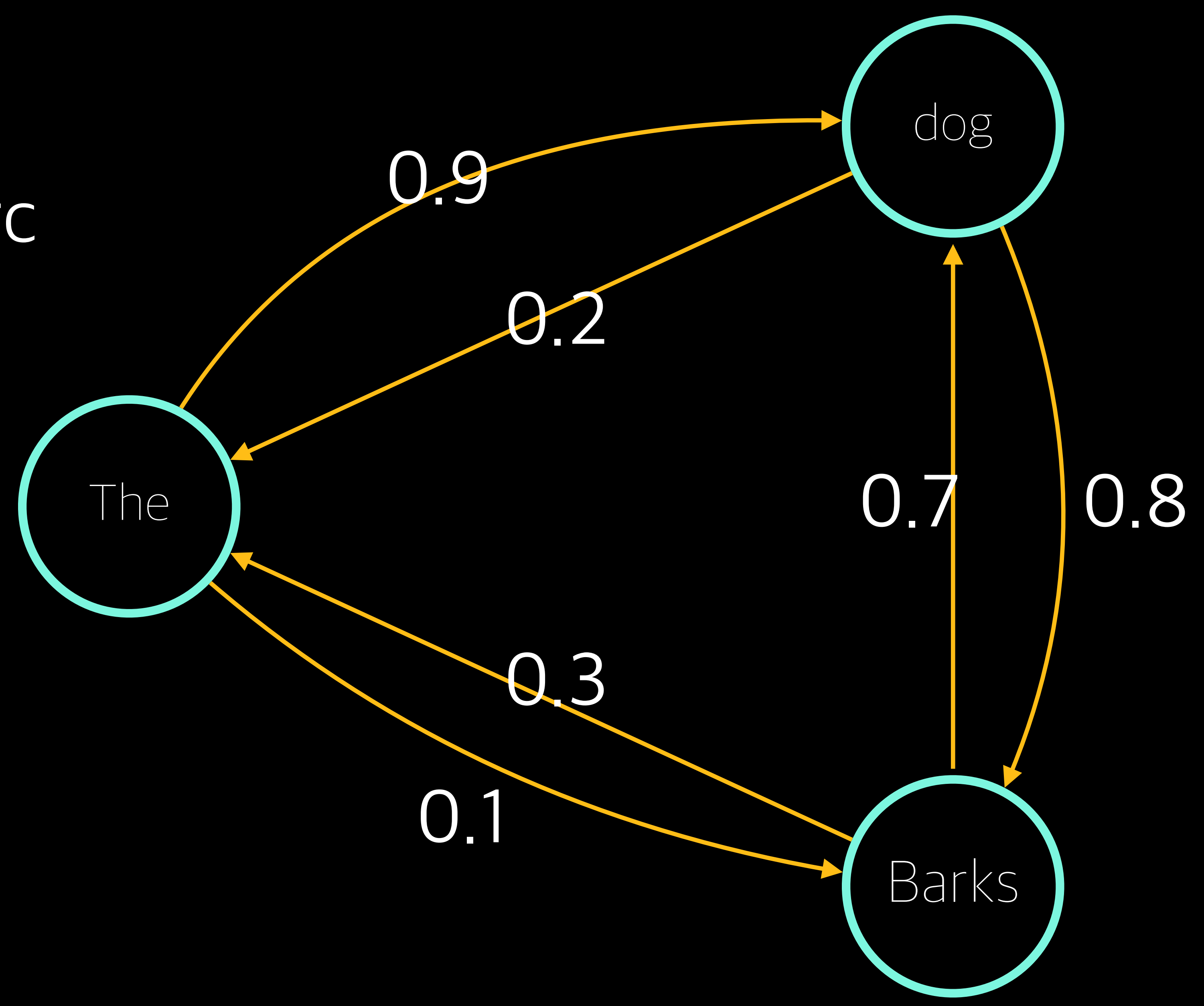
A Markov chain is a *stochastic* process characterized by the Markov property: the probability of transitioning to the next state depends only on the current state, not on the history of previous states.

$$p(q_i = a \,|\, q_1, q_2, \ldots, q_{i-1}) = p(q_i = a \,|\, q_{i-1})$$

This is a first order Markov chain - useful to compute a probability for a sequence of observable events using the just the current and the predecessor state

A Sample Markov Chain with Transitions

- Words are represented as states
- Transition probabilities are represented as edges Values leaving the arc must sum to 1 $\sum_i p_i('the') = 1$
- Initial probability distribution over states
  - $\Pi = \{\pi_1, \pi_2, \ldots, \pi_n\}$. π_i is the probability of any state that the Markov chain will start at i
    - $\Pi = \{0.7, 0.2, 0.1\}$ corresponds to the states {The, dog, barks}



Transition Examples

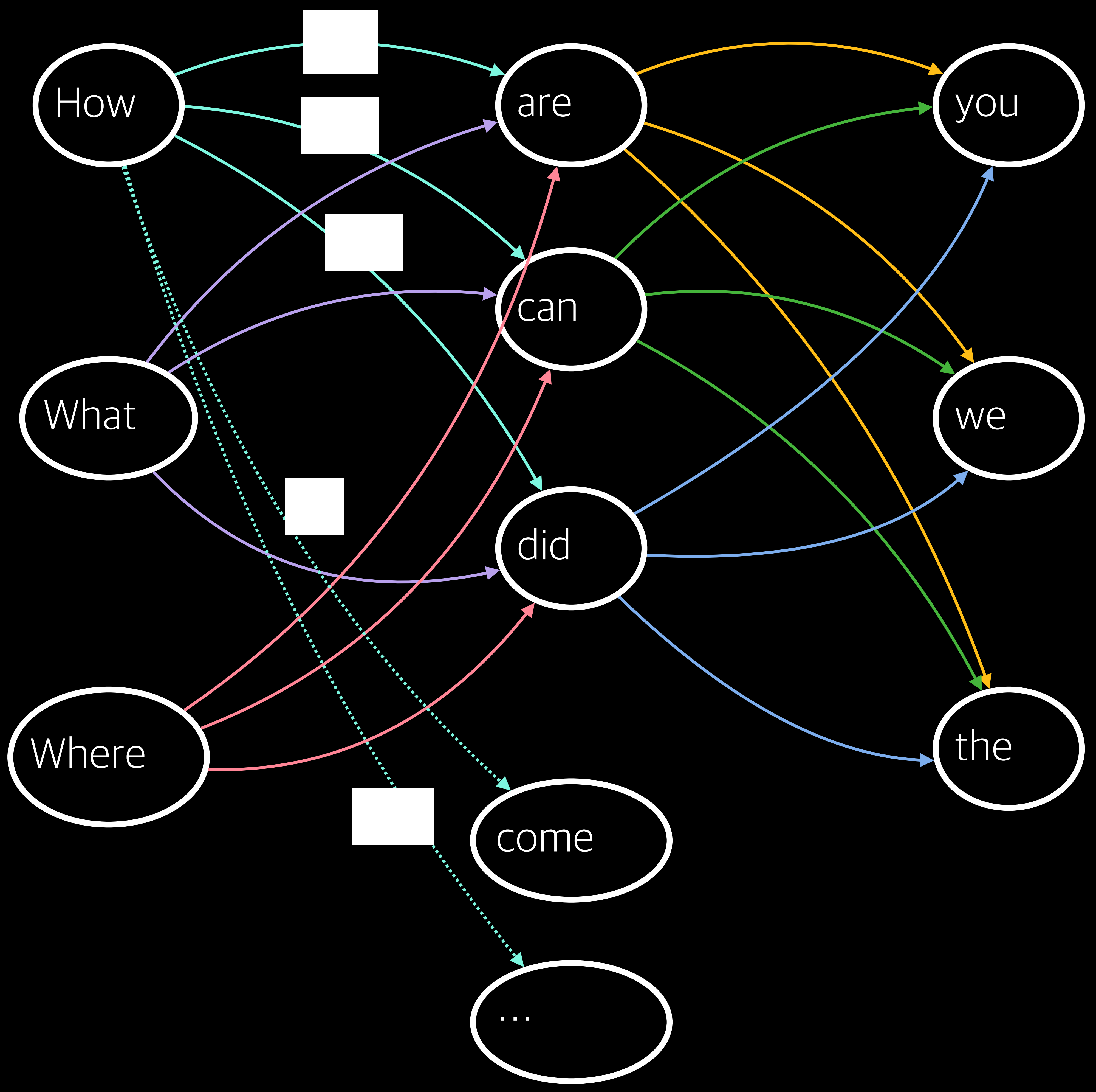
Transitioning between common POS tags

- $P(Noun \mid Determiner)$: This represents the probability of a noun following a determiner (e.g., "the", "a"). In English, this is a very high probability, as determiners typically introduce nouns.
- $P(Verb \mid Noun)$: This represents the probability of a verb following a noun. This is also quite common, as verbs often describe actions performed by the noun
- $P(Adjective \mid Comma)$: This represents the probability of an adjective following a comma. This is common for listing multiple adjectives describing the same noun

Emission Probability

- Common words and their likely POS tags
  - $P('the' \mid Determiner)$: Very high probability because "the" is almost always a determiner
  - $P('dog' \mid Noun)$: High probability
  - $P(run \mid Verb)$: High probability
- Ambiguous words
  - $P('book' \mid Noun)$: Mostly high probability
  - $P('book' \mid Verb)$Depending on the context, this could be high
  - $P('dust' \mid Noun)$: High probability (around 0.8-0.9) because "dust" is mostly a noun.
  - $P('dust' \mid Verb)$: Lower probability (around 0.1-0.2) because "dust" can also be a verb in specific cases.

Markov chain



$q = \{How, What, are, can, did, your, we, the\}$

$\pi = \{p_{How} = 0.4, p_{What} = 0.35, p_{Where} = 0.25\}$

The edges $\sum_i p_i('how') = 1$. Transition Probability

Matrix, $A = a_{ij}$, where a_{ij} represents the probability of moving from state i to j.

$a_{ij} = p(q_t = s_j | q_{t-1} s_i)$, for $1 \le i, j \le N$, with $a_{ij} \ge 0$ and $\sum_j a_{ij} = 1, \forall i$

Partial Transition Probability Matrix - A

	how	what	where	are	can	did	come	you	…
how	0.0	0	0	0.21	0.15	0.18	0.11		
what	0	0.0	0	0.2	0.12	0.16	0.001		
where	0	0	0.0	0.18	0.2	0.1	0.001		
are	0	0	0	0.0	0	0	0		
can	0	0	0	0	0.0				
did						0.0			
come							0.0		
…									

Hidden Markov Model

Markov chain is useful to compute a probability for a sequence of observable events

In all the sentences we

- Observe: Words
- Hidden/inferred: Parts of Speech tags

The POS-HMM can be used to compute the probability of a given sequence of words, as well as the most likely sequence of POS tags for a given sentence

Components

- States – The set of possible speech tags
- Observations – The words in the vocabulary
- Transition Probability – The probability of transitioning from one state to another
- Emission Probability – The probability of observing a particular word given a particular state

A Simple Example

- Want to determine the average annual temperature (Hot or Cold) of a location on earth over a period of time
- There is no record of temperature available
- The following information is available
- Assuming that there is a correlation between tree ring size and the temperature

$$A = \begin{bmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{bmatrix}$$

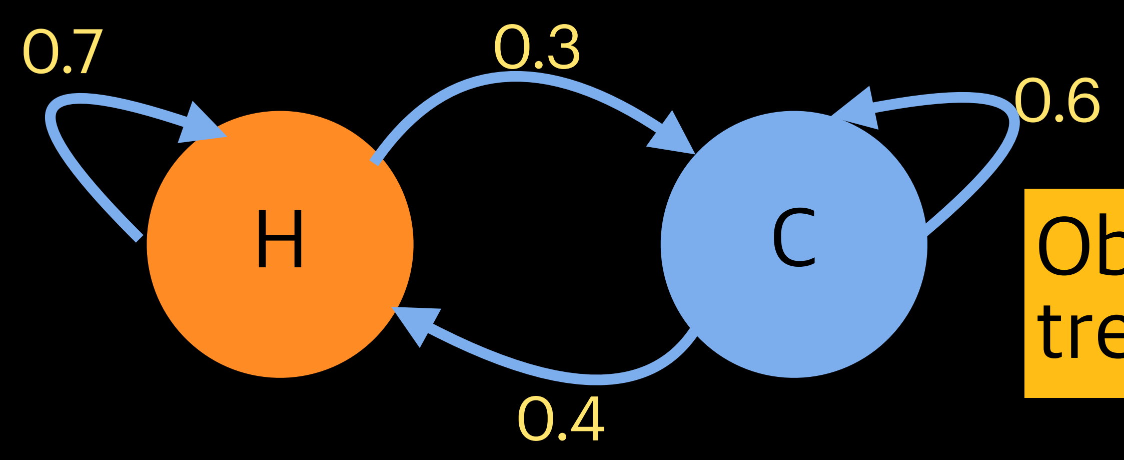
The sequence of HH is 0.7
Probability of hot year followed by another hot year

$$B = \begin{bmatrix} 0.1 & 0.4 & 0.5 \\ 0.7 & 0.2 & 0.1 \end{bmatrix}$$

Ring Sizes – Small(S), Medium(M) and Large(L)

$$\pi = \begin{bmatrix} 0.6 & 0.4 \end{bmatrix}$$

$$O = \begin{bmatrix} 0 & 1 & 0 & 2 \end{bmatrix}$$



Probability of the starting state

Observation sequence of tree rings for 4 years

Given the observation, O, we need to estimate the state sequence $\{H, C\}$

HMM (λ)

$T =$ length of the observed sequence

$N =$ Number of states in this model

$Q = \{q_q, q_2, \ldots, q_N\}$

$V =$ set of possible observations

$A =$ Probabilities of the state transitions - row stochastic

$B =$ Matrix of the probability of the Observed sequence

$\pi =$ Probability of the starting states

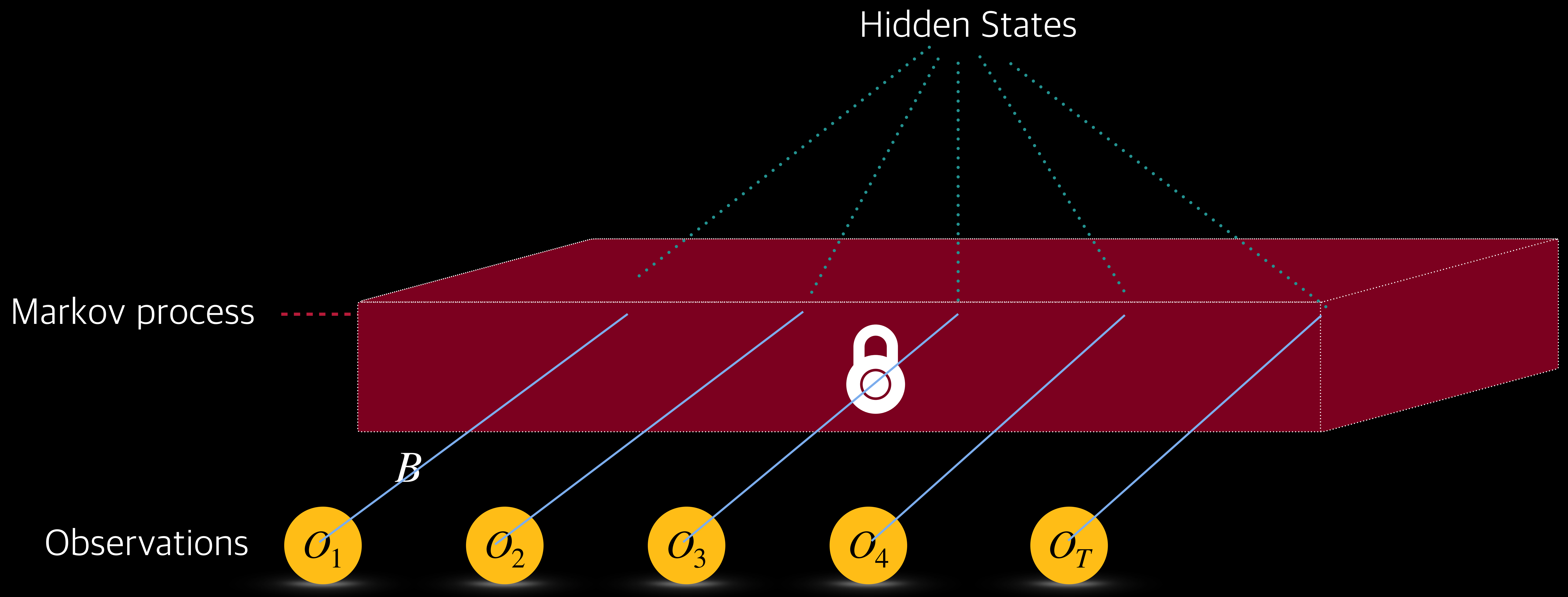
$O = \{O_1, O_2, \ldots, O_T\}$ – observation sequence

$a_{ij} = P($ state q_j at $t+1 \mid$ state q_i at $t)$

$b_j(k) = P($ observation k at $t \mid$ state q_j at $t)$ and independent of t

$$\lambda = (A, B, \pi)$$

Hidden Markov Model



$$P(X, O) = \pi_{x_1} \cdot b_{x_1}(O_1) \cdot a_{x_1 \to} a_{x_2} \cdot b_{x_2}(O_2) \cdot a_{x_2 \to} a_{x_3} \cdot b_{x_3}(O_3) \cdot a_{x_3 \to} a_{x_4} \cdot b_{x_4}$$

Probability of initially observing O_1

$$P(HHCC) = 0.6 \cdot 0.1 \cdot 0.7 \cdot 0.4 \cdot 0.3 \cdot 0.7 \cdot 0.6 \cdot 0.1 = 0.000212$$

HMM Probabilities

state	probability	normalized probability
$HHHH$	.000412	.042787
$HHHC$	.000035	.003635
$HHCH$	.000706	.073320
$HHCC$	.000212	.022017
$HCHH$	.000050	.005193
$HCHC$	.000004	.000415
$HCCH$	.000302	.031364
$HCCC$	.000091	.009451
$CHHH$	.001098	.114031
$CHHC$	.000094	.009762
$CHCH$	.001882	.195451
$CHCC$	.000564	.058573
$CCHH$	.000470	.048811
$CCHC$	.000040	.004154
$CCCH$	.002822	.293073
$CCCC$	.000847	.087963

State Sequence Probabilities

	1	2	3	4
$P(H)$	0.188182	0.519576	0.228788	0.804029
$P(C)$	0.811818	0.480424	0.771212	0.195971

HMM Probabilities

Normalised sum all the probabilities of H occurring at the first position

Normalised sum all the probabilities of C occurring at the first position