Advanced Machine Learning Code: 18AI72

Module II - HMM

Dr. Varalatchoumy M
Prof.&Head – Dept. of AIML
Head –CHOSS,
Cambridge Institute of Technology, Bangalore

Markov Model is a stochastic process used to model sequential or temporal data that satisfies Markov property.

In probability theory and statistics, the term Markov property refers to the memoryless property of a stochastic process, i.e. it is assumed that future states depend only on the current state, not on the events that occurred before it.

There are four common Markov models used in different situations, depending on whether every sequential state is observable or not, and whether the system is to be adjusted based on observations made.

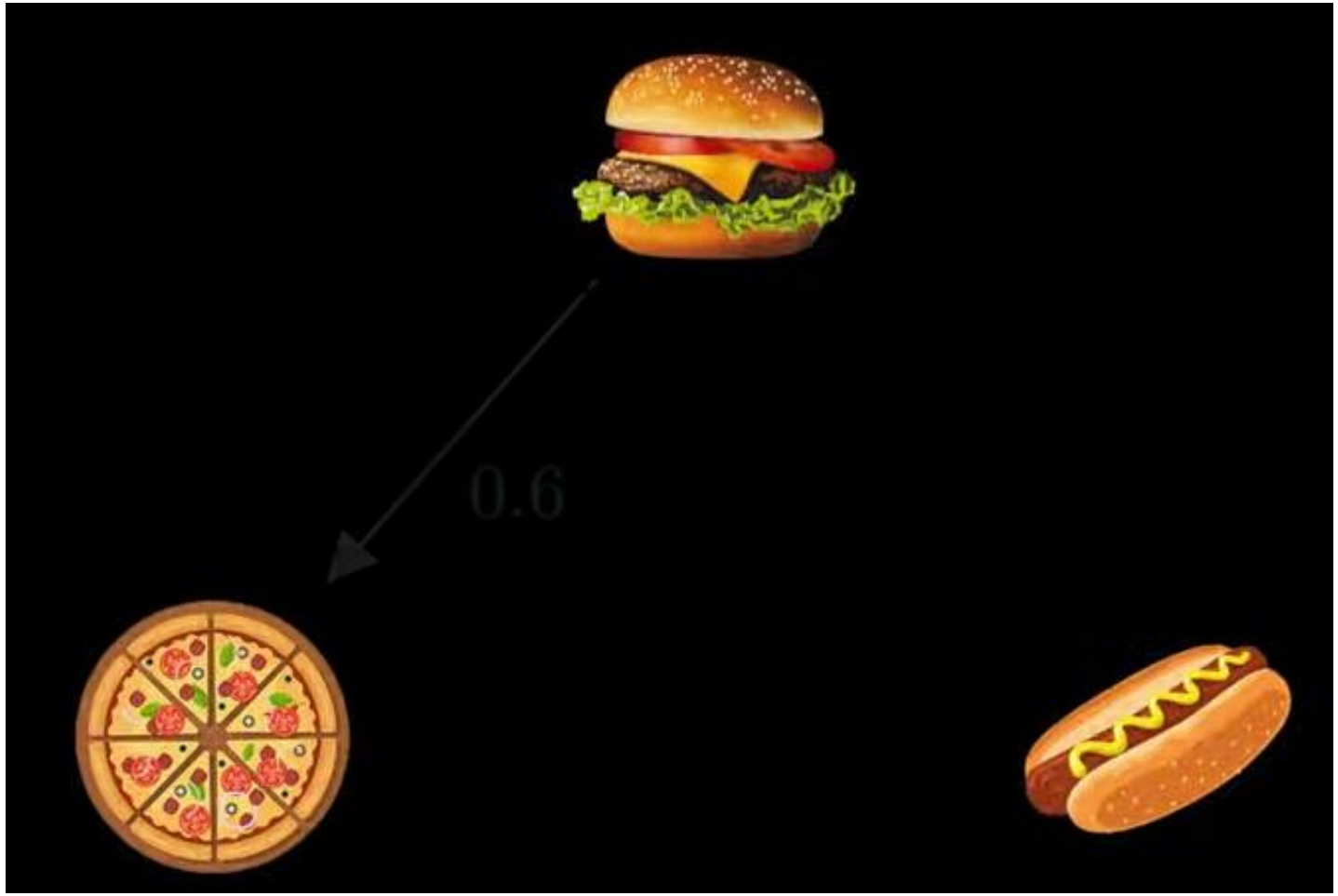
Andrey Markov, 1856-1922. Russian mathematician renowned for his work on stochastic models.

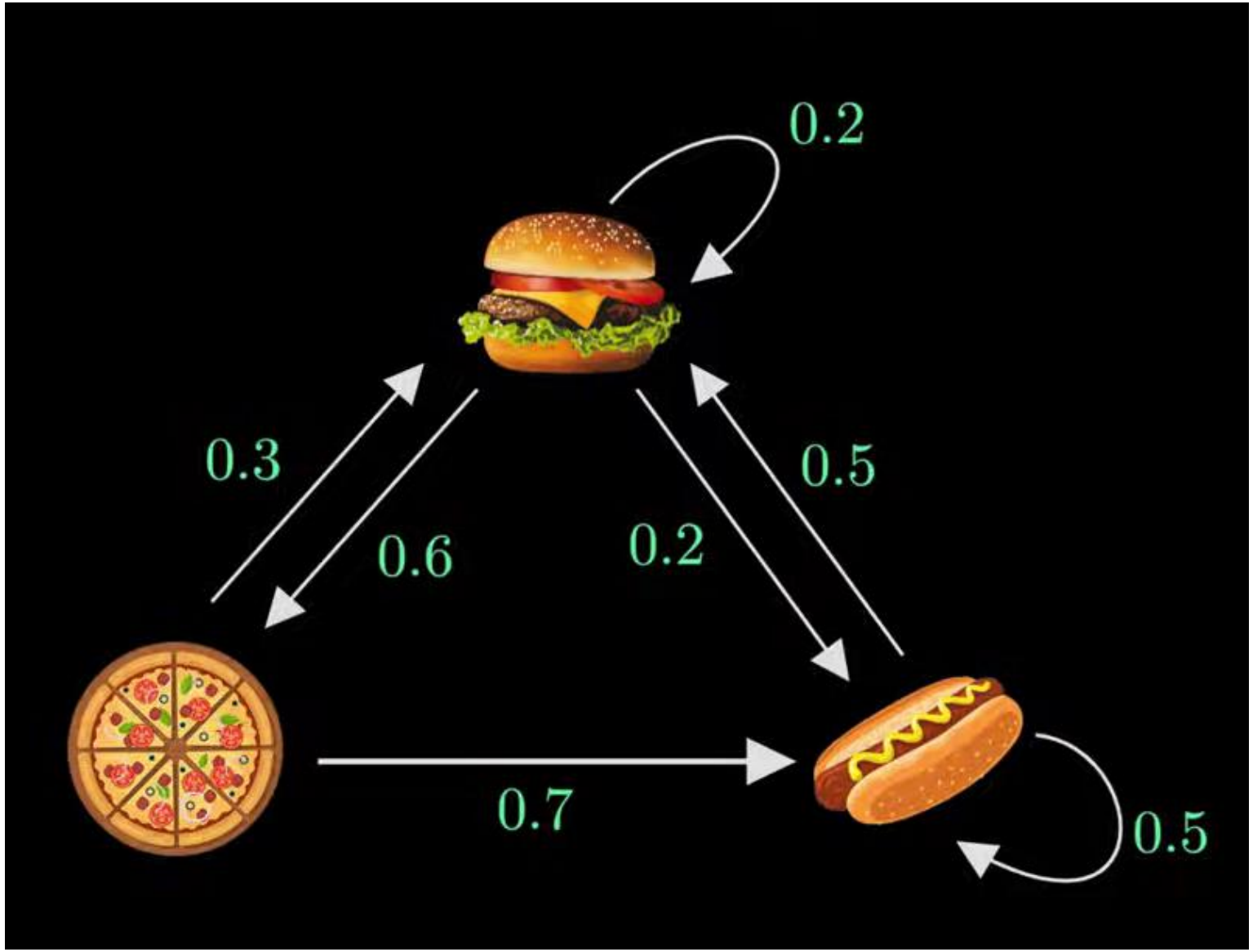
Type	States are fully observable	States are partially observable
Autonomous	Markov chain	Hidden Markov model
Controlled	Markov decision process	Partially observable Markov decision process

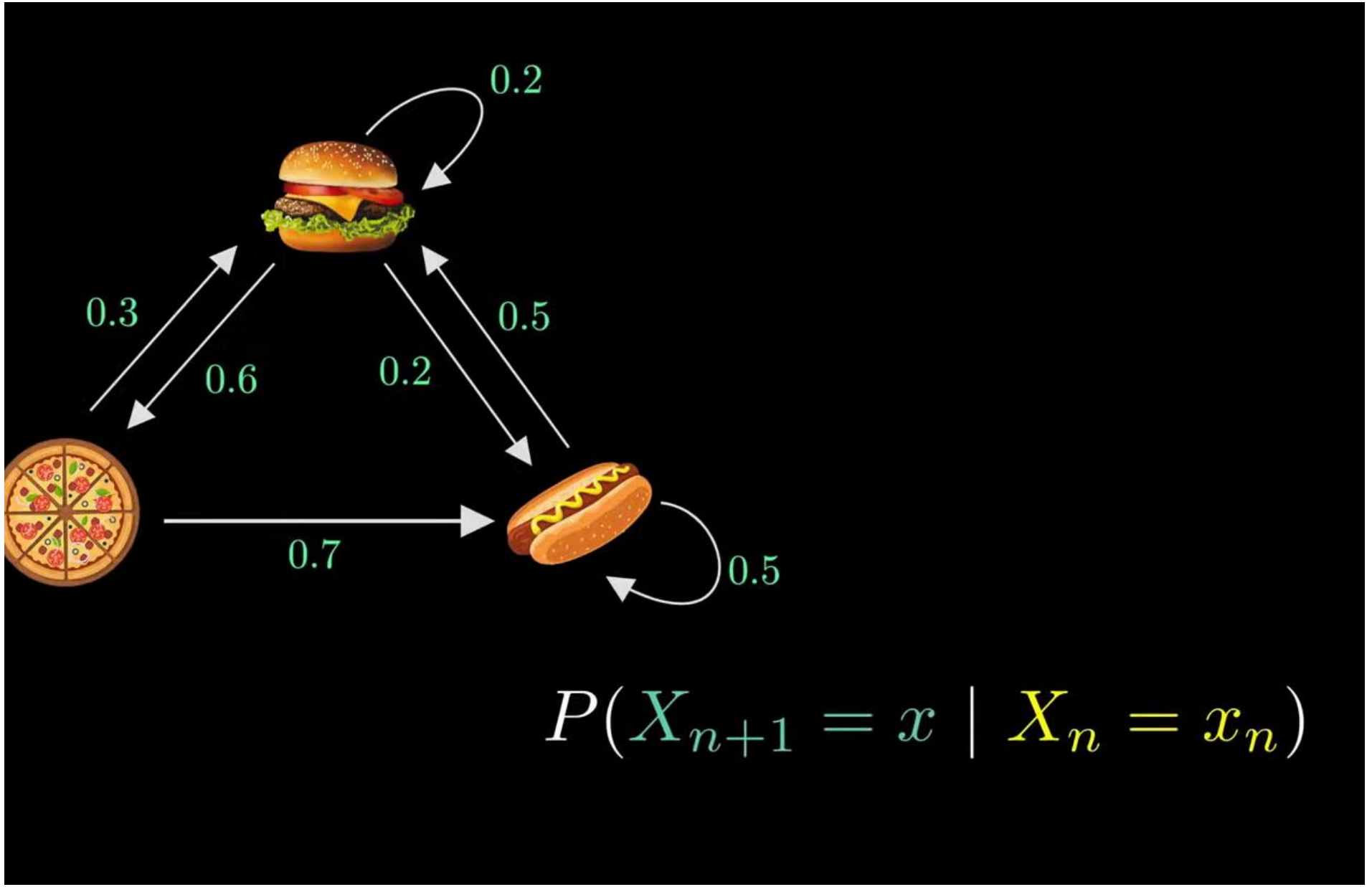
Markov Models have a very wide range applications and use different varieties of temporal or sequential data:

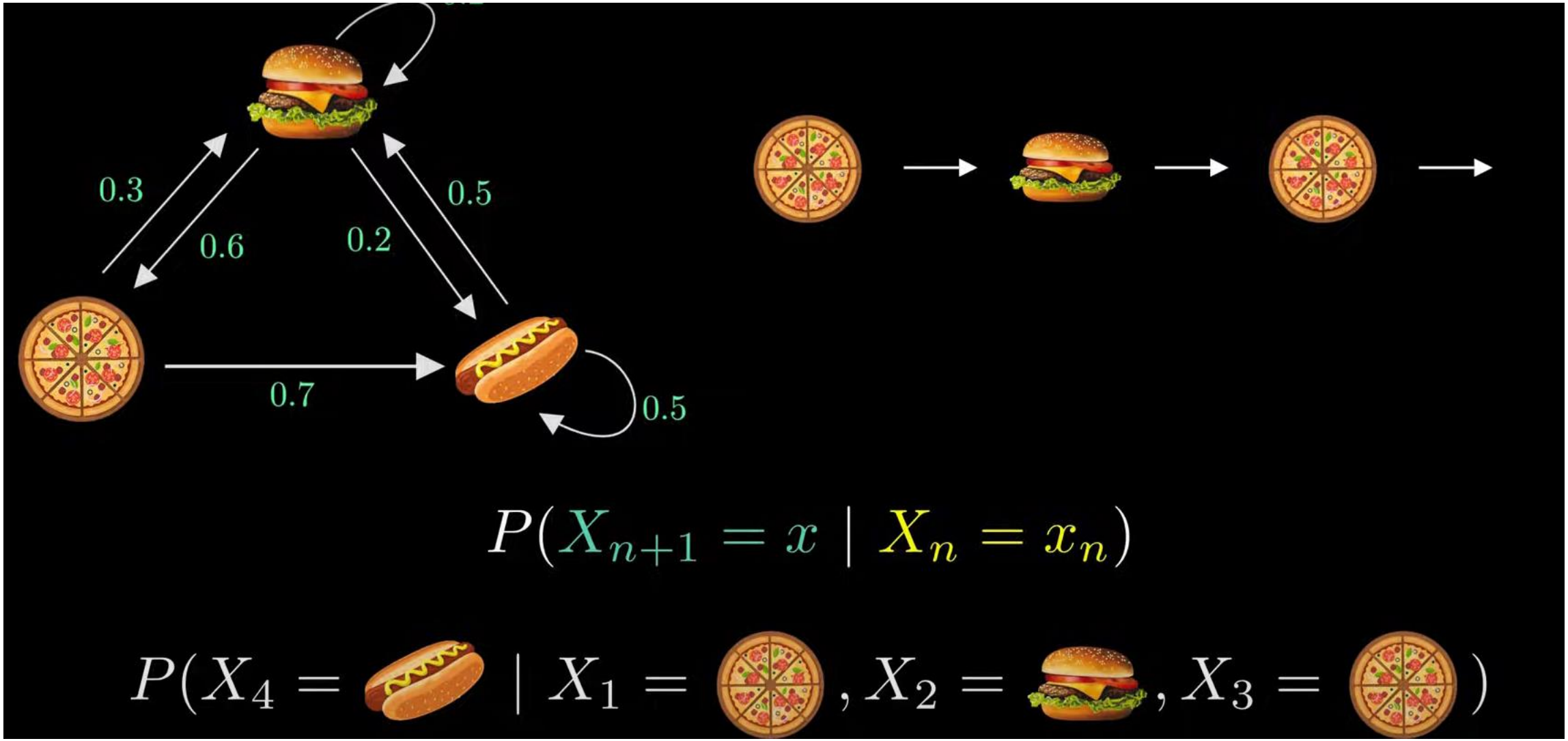
- Climate data (e.g. weather forecast)
- Financial data (e.g. investment analysis)
- Text (e.g. language modeling)
- Speech to text (e.g. transcribing speech to text data)
- Music (e.g. generate music based on certain patterns)
- Motion data (e.g. human activity analysis)

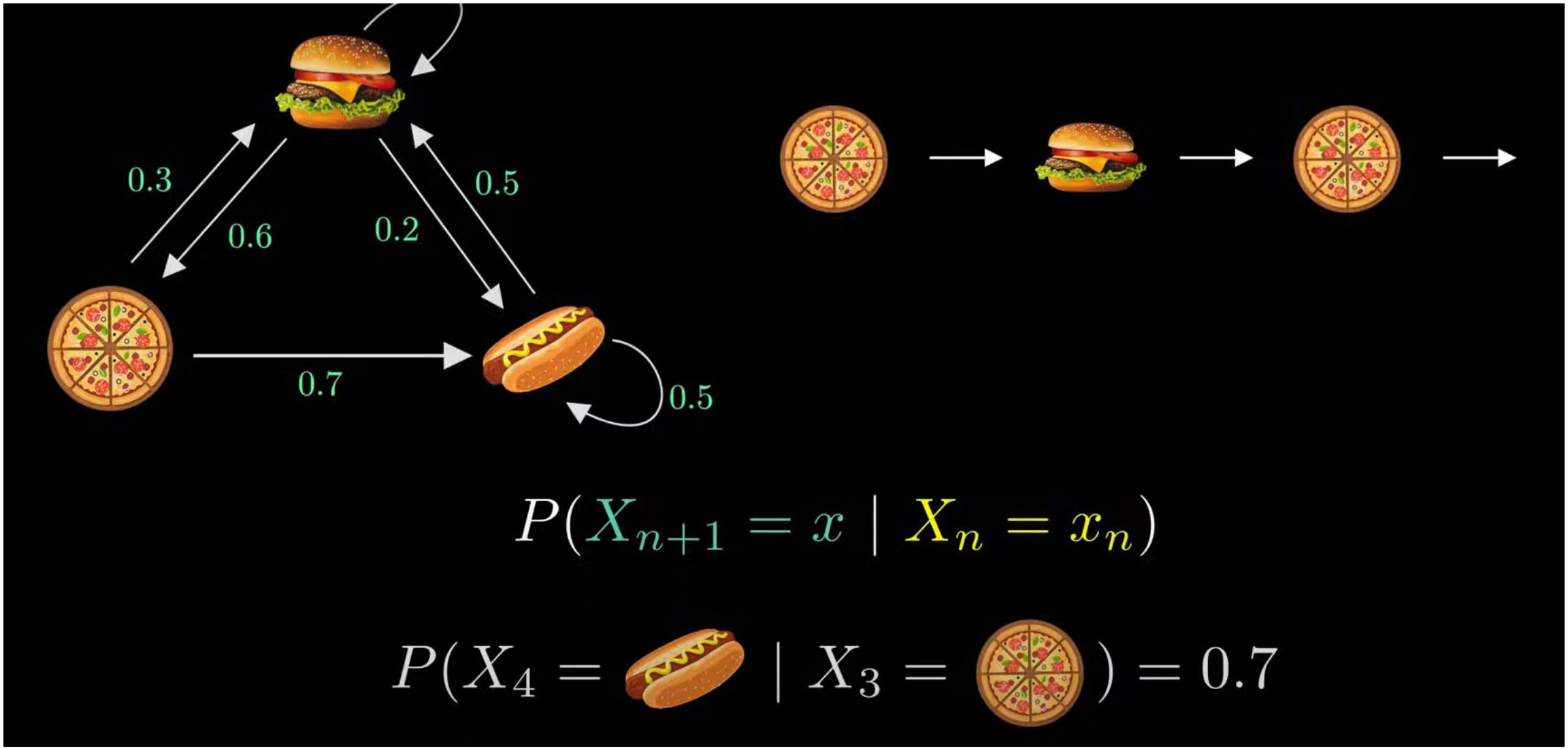
Markov chain and Hidden Markov Model

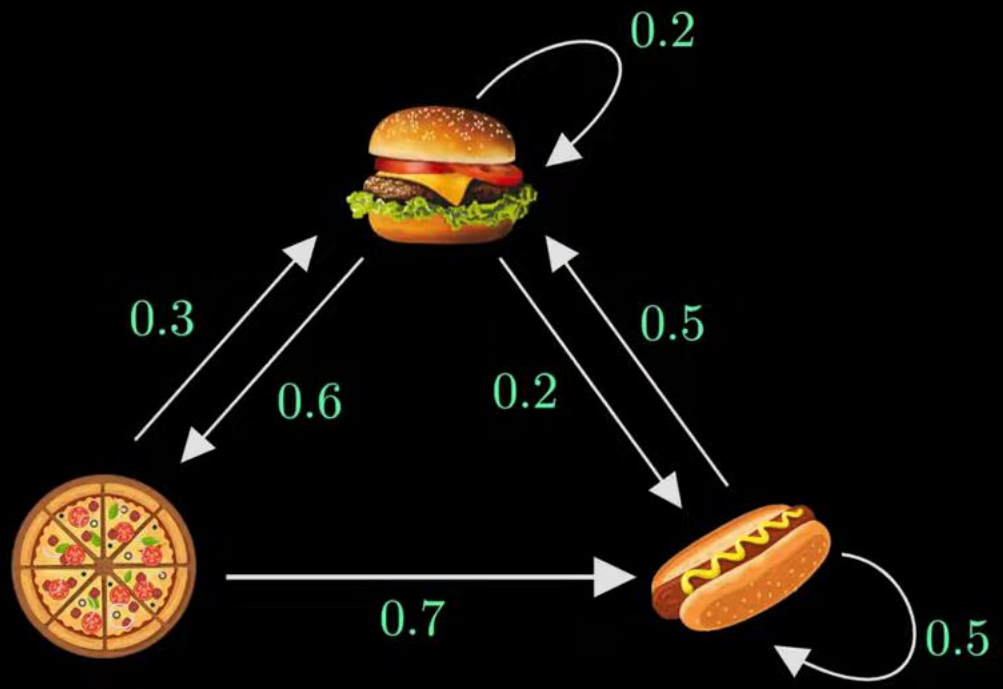

0.6


0.2
0.3
0.6
0.2
0.5
0.7
0.5


0.2
0.3
0.6
0.2
0.5
0.7
0.5
$P(X_{n+1} = x \mid X_n = x_n)$


0.3
0.6
0.2
0.5
0.7
0.5
$P(X_{n+1} = x \mid X_n = x_n)$
$P(X_4 =$ $\mid X_1 =$ $, X_2 =$ $, X_3 =$ $)$


0.3
0.6
0.2
0.5
0.7
0.5
$P(X_{n+1} = x \mid X_n = x_n)$
$P(X_4 =$ $\mid X_3 =$ $) = 0.7$



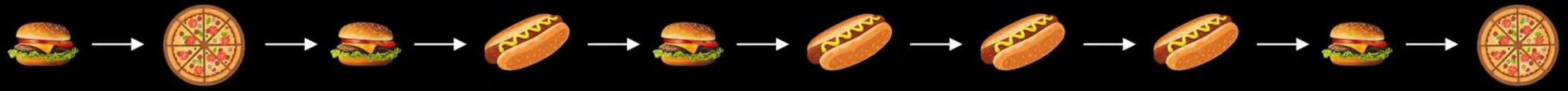
After 10 steps...

$P($  $)$ $\frac{4}{10}$

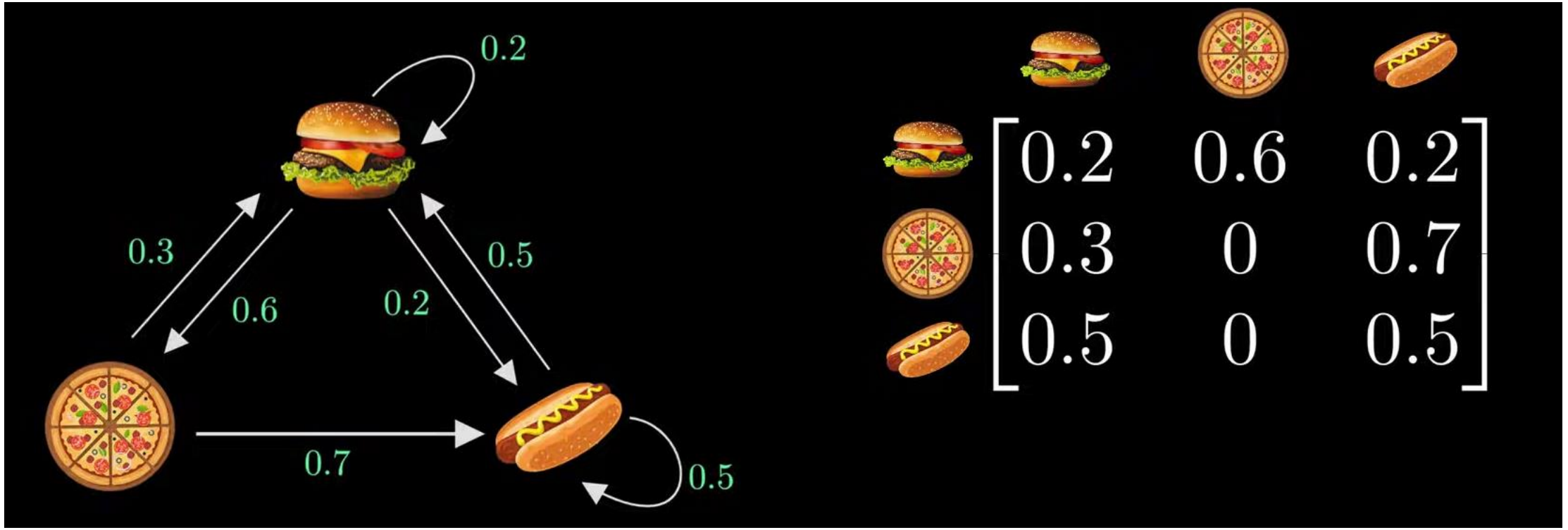
$P($  $)$ $\frac{2}{10}$

$P($  $)$ $\frac{4}{10}$

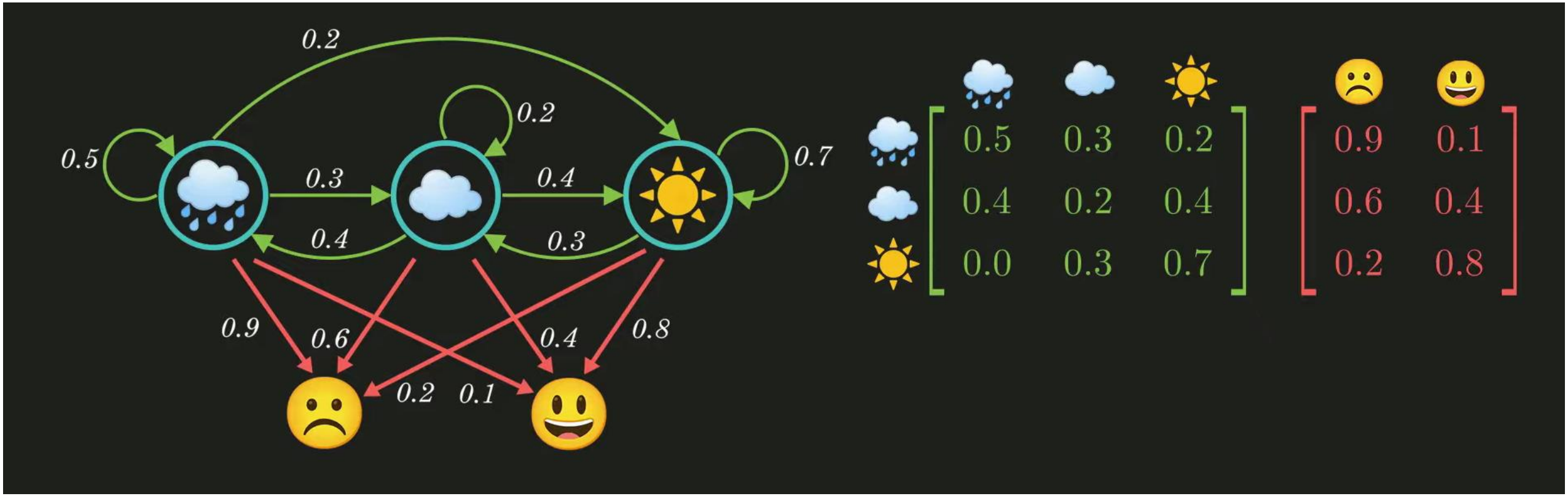
Random Walk

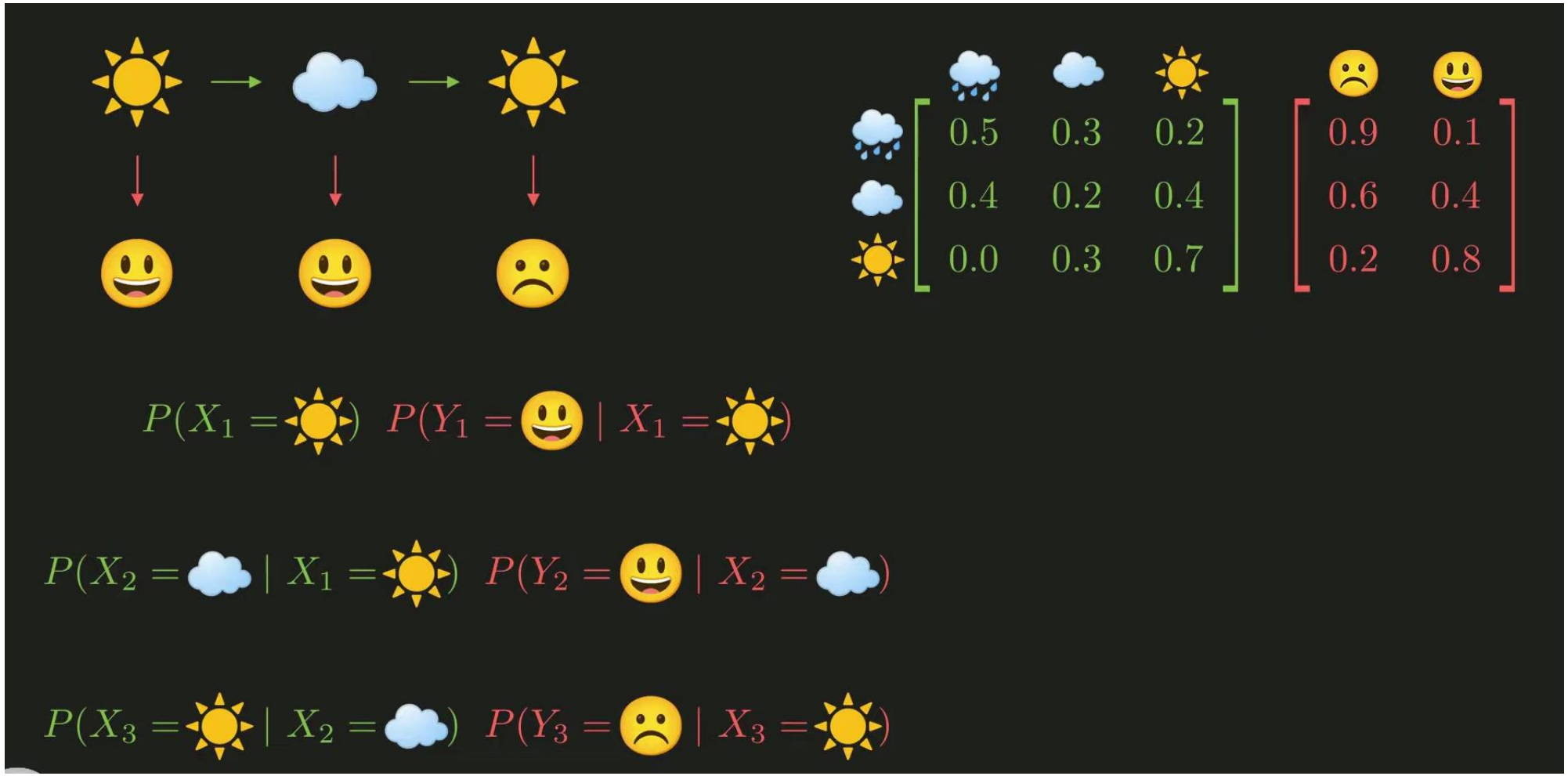


Transition matrices or adjacency matrices



Π denotes probability distribution of the states


0.2
0.2
0.5
0.7
0.3
0.4
0.4
0.3
0.9
0.6
0.4
0.8
0.2
0.1
0.5 0.3 0.2
0.4 0.2 0.4
0.0 0.3 0.7
0.9 0.1
0.6 0.4
0.2 0.8



BAYES THEOREM

12 Hidden Markov Model

LEARNING OBJECTIVES

- To introduce the category of recognition problems using temporal patterns.
- To introduce the basics of Hidden Markov Model (HMM).
- To understand the issues in HMM.

LEARNING OUTCOMES

- Students will understand and appreciate the use of HMM as a classifier.
- Students will be get an insight on speech recognition problems.
- Students will understand the issues in HMM.

12.1 Introduction to Hidden Markov Model

Speech recognition problems come under the category of recognition problems with temporal patterns. We can usually see some persons with some mannerisms. As an example, some may use lots of hand gestures while talking. Some frequently repeat some words like "Got it", "Catch my point". So in a temporal sequence, that in a speech signal if some patterns are frequently repeated its called as temporal pattern. Speech signals are time baring signals (i.e., at different instances of time we have different signal strength). For speech recognition purposes, the speech is divided into a number of phonemes. The word spoken can be identified based on the sequence in which the phonemes occur.

Activity recognition is another application of temporal pattern recognition. These activities can be recognized by sign language recognition using hand movements. Based on the sequence of hand gestures, we can interpret the message conveyed. An application area of activity recognition is security surveillance where abnormal movement or activity can be monitored. Movement in security surveillance is nothing but body postures at different instances of time. So if the movement is found to be abnormal, it implies that it would be from a suspicious person.

To recognize or identify such temporal sequence we need a machine which is similar to a sequential machine or finite state machine. In other words, a machine takes us through a finite set of output symbols from a finite set of input symbols. Thus, we have a finite set of states through which the machine makes a transition, i.e., at any time of time the machine moves from one state to another, it takes an input and emits an output.

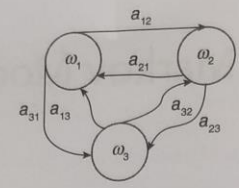


Figure 12.1 State transitions of HMM.

The machine makes a transition from one input state to an output state. If the set of outputs are limited to only two symbols, 0 and 1, the sequence machine becomes a finite state automaton and has huge application in sequence generation and sequence detection. Moreover, it is a useful concept in communication. So if a finite state automaton detects the beat sequence, the next portion of the sequence can also be detected.

Hidden Markov Model (HMM) is a statistical model that takes statistical property of signal into account. There are two types of states in HMM:

1. Visible state (V).
2. Hidden state (ω).

HMM can have a number of hidden states and visible states. As an example, let us consider an HMM model (θ) having three hidden states (ω) and three visible states (V). This is given by $\omega = \{\omega_1, \omega_2, \omega_3\}$ and $V = \{V_1, V_2, V_3\}$.

Figure 12.1 shows the state transitions of the HMM from state ω_{t-1} at time $t-1$ to state t with a probability a_{ij}.

In each state, the machine can emit one of the visible states. The probability of emission from a visible state is given by $P(v_k|\omega_j) = b_{jk}$, that is the probability of v_k given ω_j. So from ω_1 the machine emits v_2 with a probability v_{12} and v_3 with a probability v_{13}. These are the probabilities of emission of visible states from different hidden states of HMM. The above process is called *observable Markov model*, since the output of the process is the set of states at each instant of time, where each state corresponds to an observable event. Thus, we can see that given any state there will always be transition to some of the states including the original state itself.

Redrawing Fig. 12.1 with the visible states, hidden states, emission probabilities, and transition probabilities, we get Fig. 12.2.

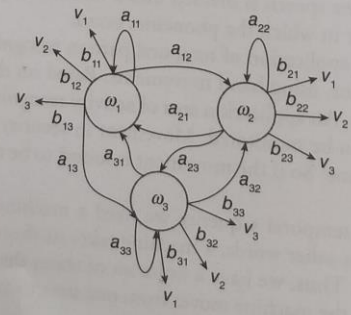


Figure 12.2 State transition diagrams showing emission and transition probabilities.

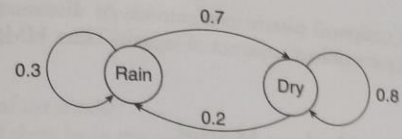


Figure 12.3 Markov chain for Rain and Dry.

It can also be seen that at any point of time

$$\sum a_{ij} = 1, \quad \forall i \quad \text{and} \quad \sum b_{jk} = 1, \quad \forall j$$

Let us consider the example of two hidden states Rain and Dry with the initial probabilities as shown in Fig. 12.3.

Figure 12.3 shows a simple example of Markov chain with two hidden states Rain and Dry. The different state transition probabilities when the state changes from state Rain to Rain, Rain to Dry, Dry to Dry, Dry to Rain are given below, along with the initial probability of Rain and Dry.

$$P(\text{Rain}|\text{Rain}) = 0.3$$
$$P(\text{Dry}|\text{Dry}) = 0.8$$
$$P(\text{Dry}|\text{Rain}) = 0.7$$
$$P(\text{Rain}|\text{Dry}) = 0.2$$
$$P(\text{Rain}) = 0.4$$
$$P(\text{Dry}) = 0.6$$

It is seen that the sum of all transition probabilities is equal to 1 and the sum of emission probabilities is also equal to 1, as given in the calculations below.

$$P(\text{Rain}|\text{Rain}) + P(\text{Dry}|\text{Rain}) = 0.3 + 0.7 = 1$$
$$P(\text{Dry}|\text{Dry}) + P(\text{Rain}|\text{Dry}) = 0.2 + 0.8 = 1$$

Finally, HMM has a specific hidden state called *receiving state* or *accepting state*. Once the machine reaches that state it cannot come out. Only one visible state can be emitted from this hidden state. Incorporating this concept in Fig. 12.2 and redrawing it we get Fig. 12.4. It is seen that v_0 is the final visible state that is emitted from the hidden state. There are no transitions to other hidden states from this state.

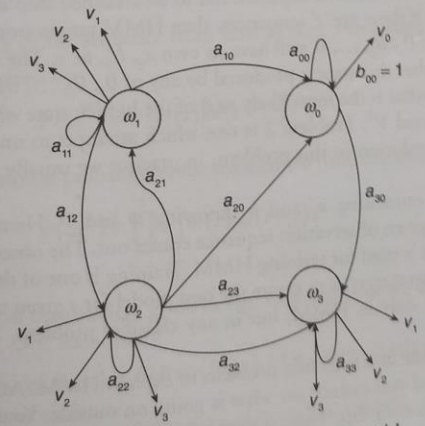


Figure 12.4 Markov model with visible and hidden states.

HIDDEN MARKOV MODEL

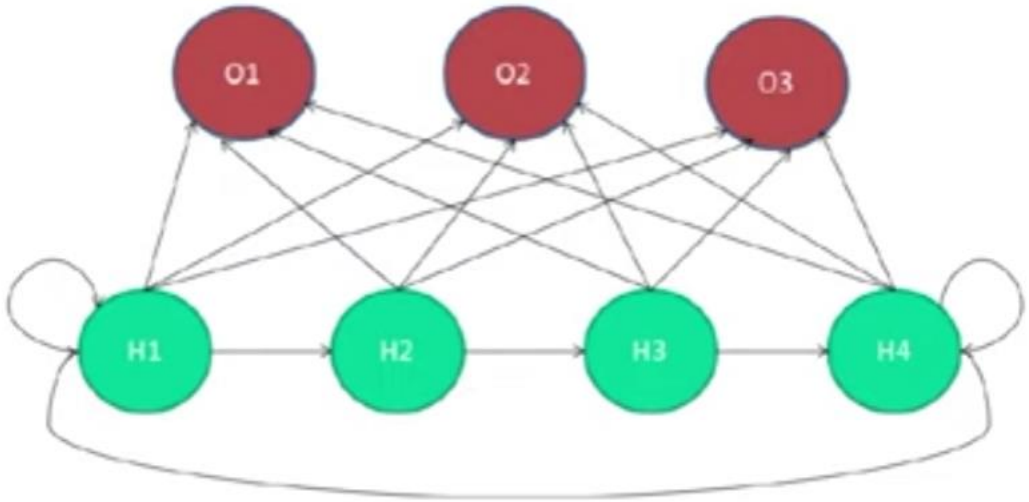
Hidden Markov Model is a statistical model, in which the system being modeled is assumed to be a Markov Process - call it X - with unobserved ("hidden") states

HMM assumes that there is another process Y (observable), whose behaviour depends on X

The goal is to learn about X by observing Y

Hidden Markov Model is very useful in,

- Bio informatics
- Speech Recognition
- Natural Language Processing etc

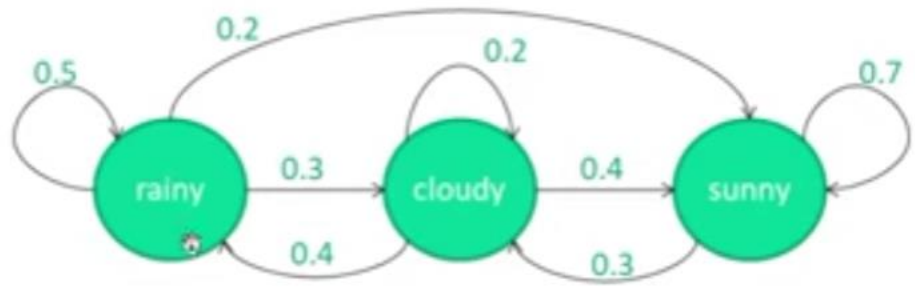


- Assume the kinds weather at my place are,
  1. Rainy
  2. Cloudy
  3. Sunny
- On any given day only one of them will occur
- The weather tomorrow only depends upon the weather today
- We can model this with a simple Markov chain

- Assume the kinds weather at my place are,
  1. Rainy
  2. Cloudy
  3. Sunny

- On any given day only one of them will occur
- The weather tomorrow only depends upon the weather today

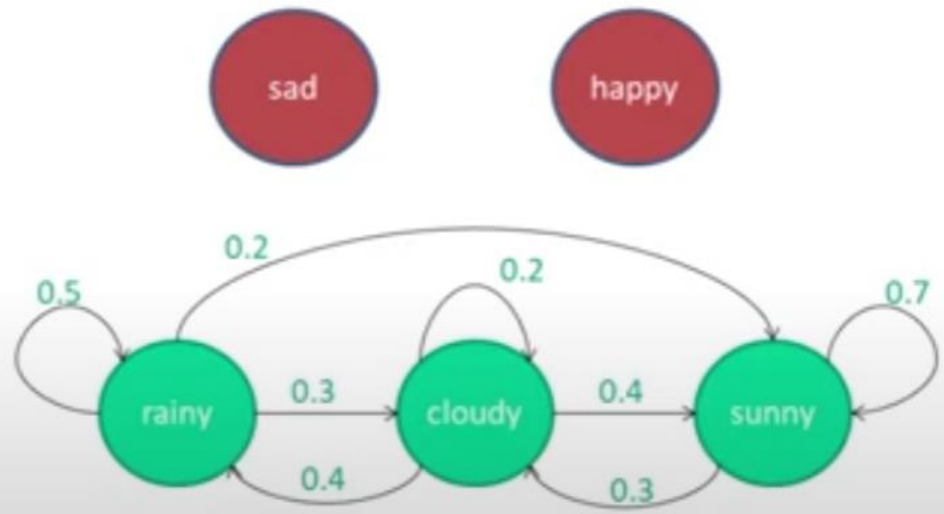
- We can model this with a simple Markov chain
- Adding the state transitions and transition probabilities



- Assume the kinds weather at my place are,
  1. Rainy
  2. Cloudy
  3. Sunny
- On any given day only one of them will occur
- The weather tomorrow only depends upon the weather today
- We can model this with a simple Markov chain
- Adding the state transitions and transition probabilities

At any given day, my mood can be

- happy or sad

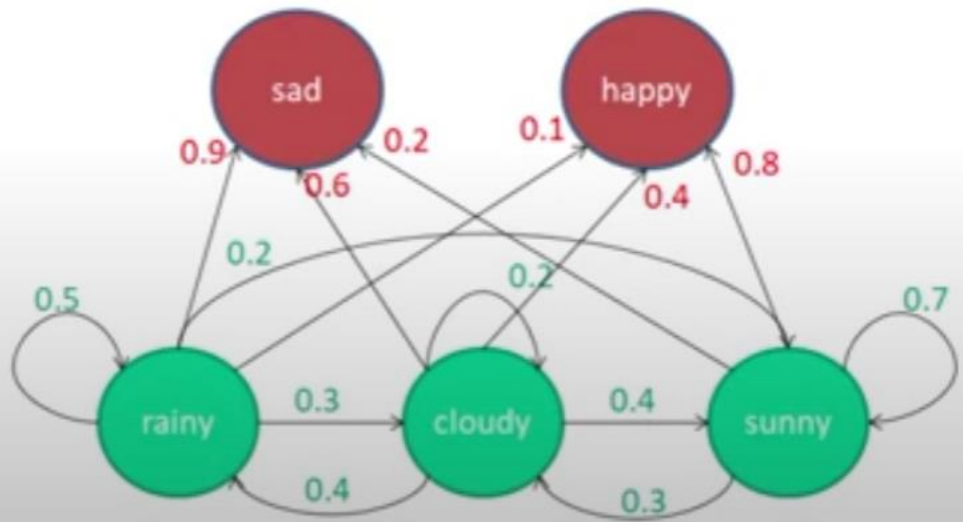


- Assume the kinds weather at my place are,
  1. Rainy
  2. Cloudy
  3. Sunny
- On any given day only one of them will occur
- The weather tomorrow only depends upon the weather today
- We can model this with a simple Markov chain
- Adding the state transitions and transition probabilities

At any given day, my mood can be

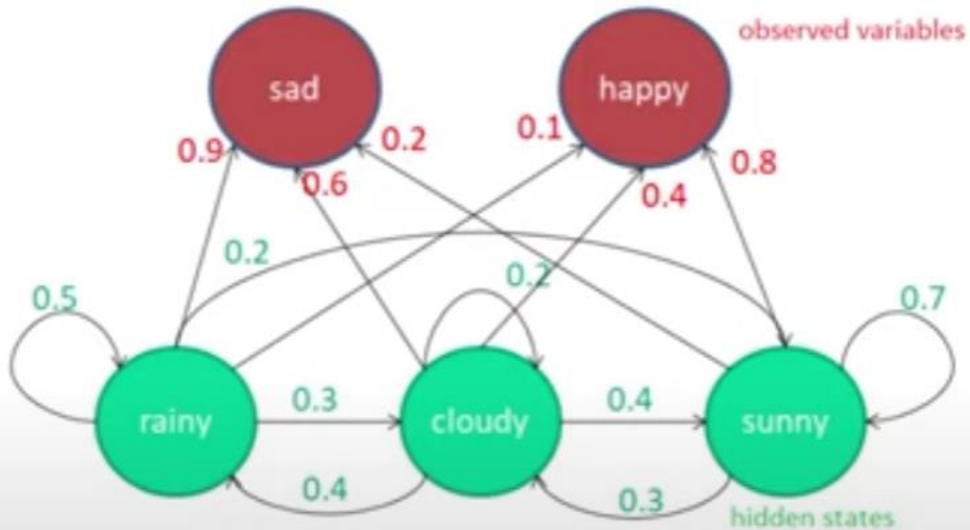
- happy or sad

This mood depends on the weather of that particular day
The probabilities are,

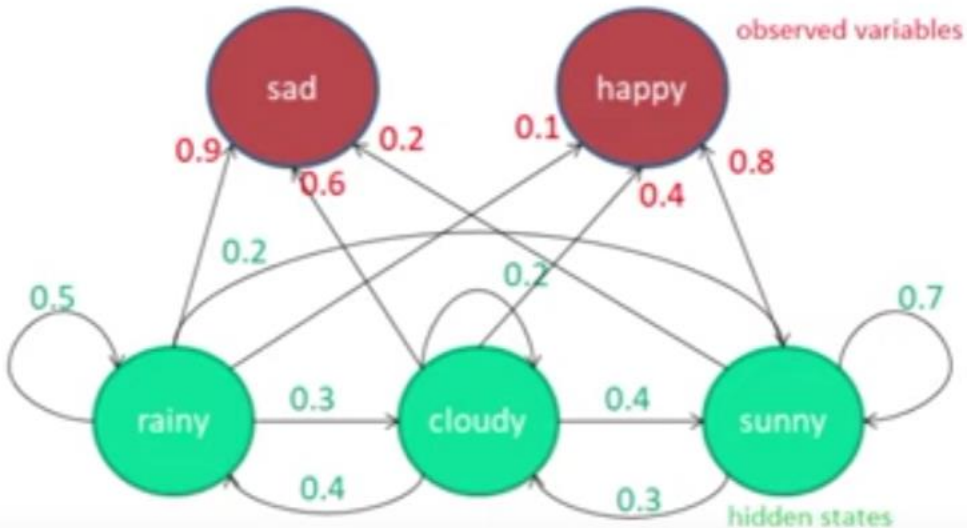


- The states of the Markov chain is hidden or unknown
- We can observe some variables that are dependent on these hidden states
- Such a model is called **Hidden Markov Model**

- HMM = Hidden MC + Observed variables



- Look at three consecutive days
  1. Sunny + Happy
  2. Cloudy + Happy
  3. Sunny + Sad
- We cant observe the hidden states (Let us assume this scenario)
- What is the probability of this scenario?
- It is same as the joint probability,
- P(happy-happy-sad, sunny-cloudy-sunny)



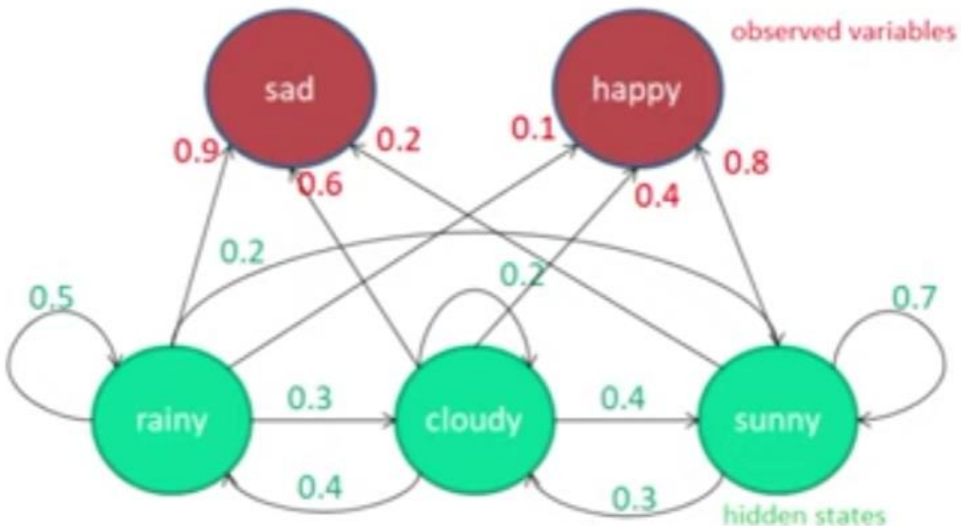
$$\begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.0 & 0.3 & 0.7 \end{bmatrix} \qquad \begin{bmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix}$$

By using the Markov property,

P(Y=happy-happy-sad, X=sunny-cloudy-sunny))

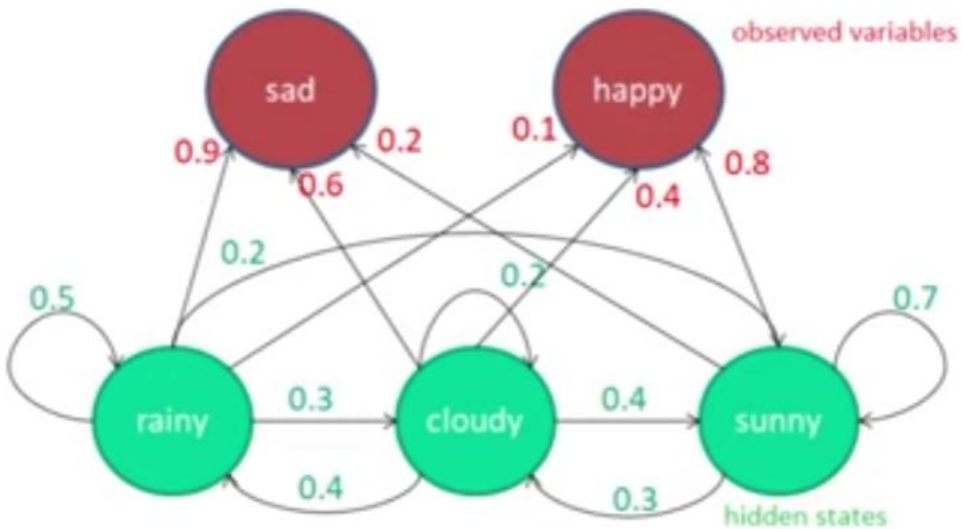
$$= P(X1=\text{sunny}) * P(Y1 = \text{happy} \mid X1=\text{sunny}) * P(X2=\text{cloudy} \mid X1=\text{sunny})$$
$$* P(Y2 = \text{happy} \mid X2=\text{cloudy}) * P(X3=\text{sunny} \mid X2=\text{cloudy}) * P(Y3 = \text{sad} \mid X3=\text{sunny})$$

$$= 0.509 * 0.8 * 0.3 * 0.4 * 0.4 * 0.2 = 0.00391$$



$$\begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.0 & 0.3 & 0.7 \end{bmatrix} \qquad \begin{bmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix}$$

- Hiding the states from this model
- We have only a sequence of observed variables
- What is the most likely weather sequence for this observed sequence?
- There are many possible permutations(eg, rainy-cloudy-sunny, rainy-rainy-sunny, sunny-sunny-rainy etc.
- Find the weather sequence with maximizes the joint probability
- Maximum Joint Probability : **P(happy-happy-sad, sunny-sunny-cloudy) = 0.04105**

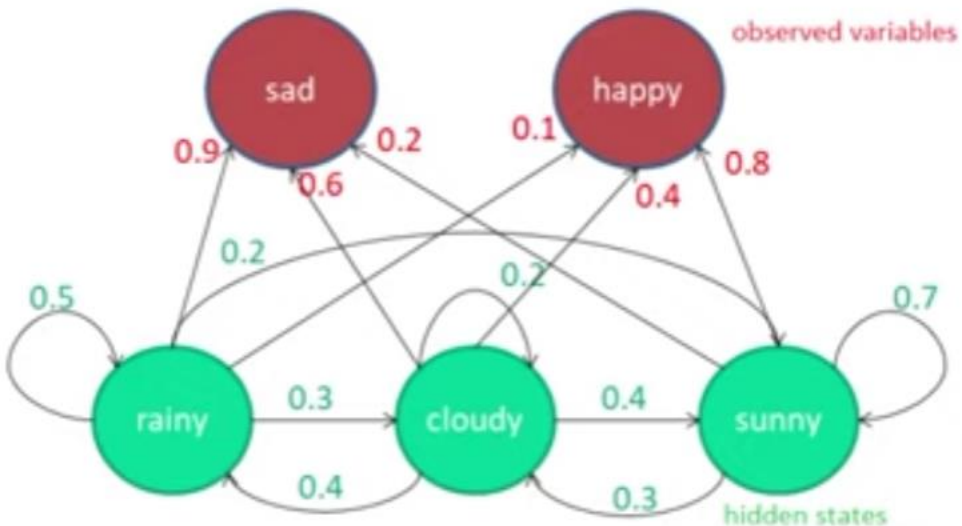


$$\begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.0 & 0.3 & 0.7 \end{bmatrix} \qquad \begin{bmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix}$$

- X: hidden variable, Y: observed variable

$$\underset{X=x_1,x_2,\ldots,x_n}{\arg\max} \; P(X = x_1, x_2, ..., x_n | Y = y_1, y2, ..., y_n)$$

- Bayes theorem can be used to calculate this result



$$\begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.4 & 0.2 & 0.4 \\ 0.0 & 0.3 & 0.7 \end{bmatrix} \qquad \begin{bmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \\ 0.2 & 0.8 \end{bmatrix}$$

$$\underset{X=x_1,x_2,\ldots,x_n}{\arg\max} \; P(X|Y)$$

Using Bayes theorem, we may rewrite this as ,

$$\underset{X=x_1,x_2,\ldots,x_n}{\arg\max} \; \frac{P(Y|X)P(X)}{P(Y)}$$

For all the practical problems, we can neglect the denominator,
The numerator is just the joint probability of X and Y)

$$P(Y|X) = P(Y_1|X_1) * P(Y_2|X_2) * \ldots * P(Y_n|X_n)$$

$$P(Y|X) = \prod P(Y_i|X_i)$$

$$P(Y|X) = P(Y_1|X_1) * P(Y_2|X_2) * \ldots * P(Y_n|X_n)$$

$$P(Y|X) = \prod P(Y_i|X_i)$$

Can be obtained from the red matrix

$$P(X) = \prod P(X_i|X_{i-1})$$

$$\underset{X=x_1,x_2,\ldots,x_n}{\arg\max} \prod P(Y_i|X_i) \quad P(x_i|X_{i-1})$$

An HMM is characterized by

- N, Number of hidden states
- M, Number of distinct observation symbols per state
- a_{ij}, The state transition probability distribution
- b_{jk}, Observation symbol probability distribution
- π_i, The initial probability ($\pi_i = P(S_1) = S_i$)
- $\lambda = (a_{ij}, b_{jk}, \pi_i)$: Complete parameter set of HMM

An HMM can be used to solve

1. Evaluation Problem
2. Decoding Problem
3. Learning Problem

EVALUATION MODEL

- Given the model, compute the probability that a particular output sequence is produced by this model
  - λ is given
  - A sequence 'O' is given
  - Find the probability that this observation sequence 'O' is generated by this model
- This problem is solved using a forward-backward algorithm

- An observation sequence is given

$$O = \{O_1, O_2, ..., O_t\}$$

- HMM model is given

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

- Find all the possible state sequences

$$Q = \{q_1, q_2, ..., q_t e\}$$

- Let us assume that the state sequence Q is given, then
- Probability of observing 'O' from the model λ and state sequence Q is

$$P(O|Q,\lambda) = \prod_{t=1}^{T} O_t|Q_t,\lambda$$
$$= b_{q_1}(o_1) * b_{q_2}(o_2) * ... * b_{q_t}(o_t) * ... * b_{q_T}(o_T)$$

- Let us assume that the state sequence Q is given, then
- Probability of observing 'O' from the model λ and state sequence Q is

$$P(O|Q,\lambda) = \prod_{t=1}^{T} O_t|Q_t,\lambda$$
$$= b_{q_1}(o_1) * b_{q_2}(o_2) * ... * b_{q_t}(o_t) * ... * b_{q_T}(o_T)$$

- Its not possible to calculate this, because
- The state sequence is not given
- So, the probability of the state sequence is determined by

- The probability of finding the state sequence Q given the model λ

$$P(Q|\lambda) = P_{q_1} \prod_{t=2}^{T} P(q_t|q_{t-1})$$
$$= \pi_{q_1} * a_{q_1 q_2} * a_{q_2 q_3} * \cdots * a_{q_T q_{T-1}}$$

- The probability of the observation sequence and the state sequence
- The joint probability can be obtained as,

$$P(O, Q|\lambda) = P(O|Q, \lambda) * P(Q|\lambda)$$
$$= \prod_{t=1}^{T} O_t|Q_t, \lambda * P_{q_1} \prod_{t=2}^{T} P(q_t|q_{t-1})$$

$$P(O, Q|\lambda) = P(O|Q, \lambda) * P(Q|\lambda)$$
$$= \prod_{t=1}^{T} O_t|Q_t, \lambda * P_{q_1} \prod_{t=2}^{T} P(q_t|q_{t-1})$$
$$= \pi_{q_1} * b_{q_1}(o_1)a_{q_1 q_2} * b_{q_2}(o_2)a_{q_1 q_2} * \ldots * b_{q_T}(o_T)a_{q_T q_{T-1}}$$

- The probability of observation sequence 'O' given λ is

$$P(O|\lambda) = \sum_{all-possible-Q} P(O, Q|\lambda)$$

- The difficulty with this method is that The number of calculations is very huge (N^T)
- Where N is number of states and T is sequence length

The observation sequence is divided into two parts

1. First part starting at time 1 to time t
2. Second part starting at t+1 to time T

$$O = \{O_1, O_2, ..., O_{t-1}, O_t \quad O_{t+1}, O_{t+2}, .., O_T\}$$

Observation Sequence	First part	Second Part
	Forward Algorithm	Backward Algorithm

Forward variable $\alpha_t(i) =$

- Probability of observing the partial sequence $\{O1, .., O_t\}$ until time t, and being in state S_i at time t, given the model λ

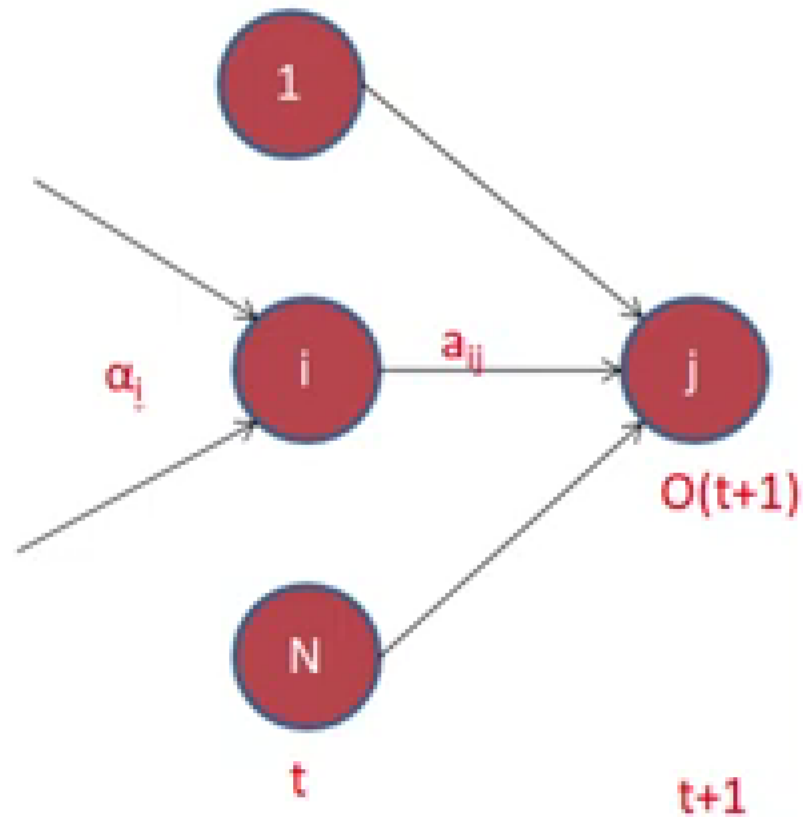
$$\alpha_t(i) = P(O1, ..., O_t, q_t = Si|\lambda$$

- There are two steps here,
  1. Initialization
     $\alpha_1(i) = P(O_1, q_1 = Si|\lambda$
     $\alpha_1(i) = P(O_1|q_1 = Si, \lambda) * P(q_1 = Si|\lambda)$
     $\alpha_1(i) = b_i(O_1) * \pi_i$
  2. By Recursion, find the partial sequence

By Recursion,

- By using the probability a_{ij}, transition from state t to state t+1 happens
- $\alpha_t(i)$ = Explains the first 't' observations and ends in state S_i
- Multiplying this probability with a_{ij} to move to state S_j

$$\alpha_{t+1}(j) = [\sum_{i=1}^{N} \alpha_t(i) a_{ij}] * b_j O(t+1)$$



→ In evaluation problem we calculated forward and backward probability.

→ From state 2, to have observation sequence O_T we had calculated the forward probability (i.e) α_i

→ From the last O_T we had been calculated the backward probability using β.

→ α_{Ej} → forward variable β_{tj} → Backward variable

↓ Sum of all the probability upto.

$\alpha_t(j)$ was representing the sum of all the probabilities upto state j by giving the transition

, Suppose initially the model was in state i, then the transition can be from state i to state j.

Now, the transition probability from state i to state j is given as a_{ij}

(i) a_{ij} → (j), and this happen from time 't' to 't+1'

this can be represented as $\alpha_t(i) a_{ij}$

So, now at a particular time 't' the model can be any state, eg: state 1, state 2, state 3 etc.

So, all probabilities are summed up

$$\sum_{t=1}^{N} \alpha_t(i) a_{ij}$$

Now joint probability of this and the emission probability of each state producing and an observation at the particular time (t+1) (is

$b_j O_{(t+1)}$ (Feeling (State) → happy, sad) j^{th} observations

is calculated as,

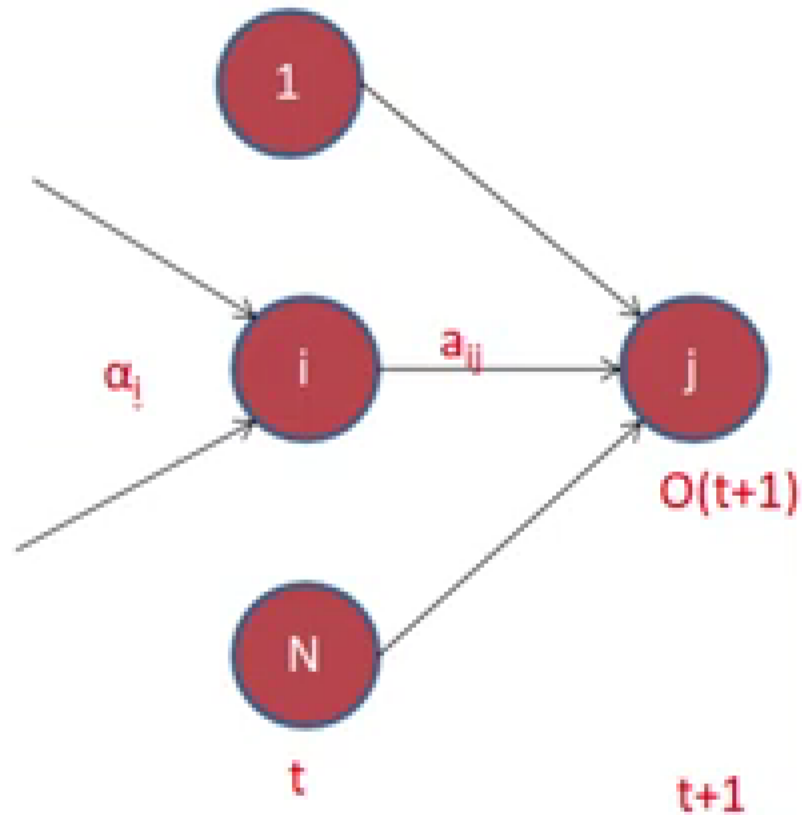
$$\alpha_{t+1}(j) = \left[\sum_{t=1}^{N} \alpha_t(i) a_{ij} \right] * b_j O_{(t+1)}$$

By Recursion,

- Probability of Observation Sequence

$$P(O|\lambda = \sum_{i=1}^{N} P(O, q_1 = S_i|\lambda)$$
$$= \sum_{i=1}^{N} \alpha_t(i)$$

- Total Time taken is $O(N^2 T)$



Backward variable $\beta_t(i) =$

- Probability of being in state S_i at time 't' and observing the partial sequence $O_{t+1}, ..., O_T$

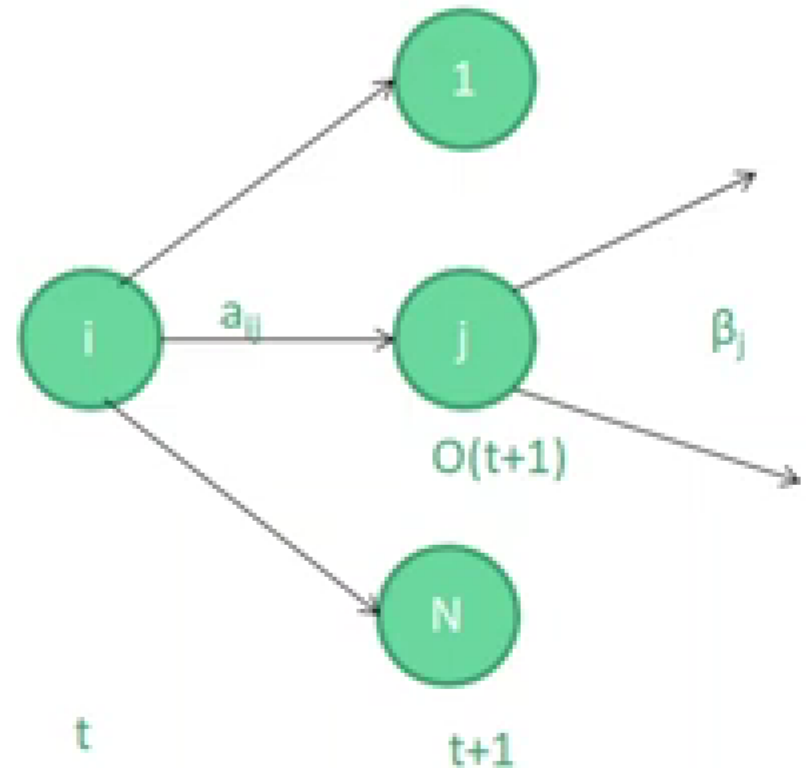
$$\beta_t(i) = P(O_{t+1}, ..., O_T, q_t = Si|\lambda$$

- There are two steps here,
  1. Initialization
     $\beta_t(i) = 1$
  2. Recursively,

By Recursion,

- From S_i it can move to N possible next states S_j, each with probability S_{ij}
- There the $t + 1^{th}$ observation is generated
- $\beta_{t+1}(j)$ represents all operations after time t+1, continuing from there

$$\beta_t(i) = [\sum_{j=1}^{N} \beta_{t+1}(j) a_{ij}] * b_j O(t+1)$$



- Multiplying α and β

$$P(O|\lambda) = \sum_{t=1}^{N} P(O, q_i = S_i|\lambda)$$

$$= \sum_{i=1}^{N} \alpha_t(i) * \beta_t(i)$$