Hidden Markov Models

Instructor: Dr. Muhammad Fahim

INNOPOLIS UNIVERSITY

Contents

- Introduction
- Markov Model
- Hidden Markov model (HMM)
- Three central issues of HMM
    - Model evaluation
    - Most probable path decoding
    - Model training
- Application Areas of HMM
- Summary

Recap!!

- Belief nets are a powerful method for representing the dependencies and independencies among variables
- In problems that have an inherent temporality — i.e., consist of a process that unfolds in time.

Hidden Markov Model

Introduction

A hidden Markov model (HMM) is a *statistical model* in which the system being modeled is assumed to be a *Markov process* with *hidden states*.

Introduction

Hidden Markov Model

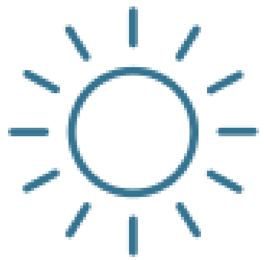
Hidden + **Markov Model**

Markov Models

- Lets talk about the weather in ***Innopolis***
  - **Sunny**
  - **Rainy**
  - **Foggy**

Markov Models

- *Weather prediction is all about trying to guess what the weather will be like tomorrow based on a history of observations of weather*

Markov Models

- Consider weather at day *n* is "w_n":

$$w_n \in \{Sunny, Rainy, Foggy\}$$

  - w_n depends on the known weathers of the past days $(w_{n-1}, w_{n-2}, w_{n-3}, \ldots)$

- **We want to find:**

$$P(w_n | (w_{n-1}, w_{n-2}, \ldots, w_1))$$

  - It means the probability of the unknown weather at day n, depending on the (known) weather $(w_{n-1}, w_{n-2}, w_{n-3}, \ldots)$ of the past days.

Markov Models

- **For example:** If we knew that the weather for the past three days was:

$$\{Sunny, Sunny, Foggy\}$$

- The probability that tomorrow would be rainy is given by:

$$P(w_4 = Rainy \mid w_3 = Foggy, w_2 = Sunny, w_1 = Rainy)$$

- This probability could be inferred from the relative frequency (the statistics) of past observations of weather sequences

- **Problem:** It's complex
- **Solution:** We will make a simplifying assumption, called the Markov assumption

Markov Models – Markov Assumption

- In a sequence $\{w_1, w_2, w_3, \dots, w_n\}$:

$$P(w_n | (w_{n-1}, w_{n-2}, \dots, w_1) \approx P(w_n | (w_{n-1})$$

- This is called a *first-order* Markov assumption, since we say that the probability of an observation at time *n* only depends on the observation at time *n-1*

- A *second-order* Markov assumption would have the observation at time *n* depend on *n - 1* and *n - 2*.

Markov Models – Markov Assumption

- So let's arbitrarily pick some numbers for $P(w_{tommorow} \,|w_{today})$:

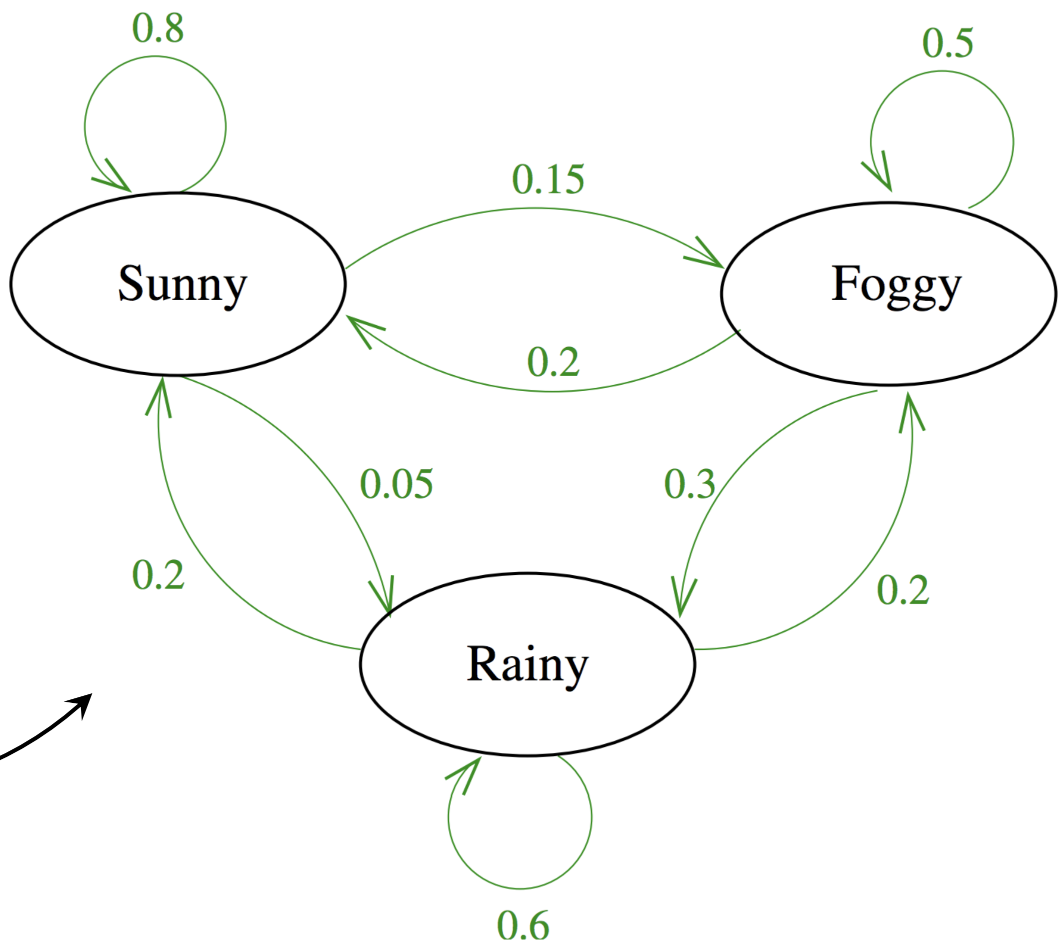
		Tomorrow's Weather		
Today's Weather		Sunny	Rainy	Foggy
	Sunny	0.8	0.05	0.15
	Rainy	0.2	0.6	0.2
	Foggy	0.2	0.3	0.5

Probabilities of Tomorrow's weather based on Today's Weather

Markov Models – Markov Assumption

		Tomorrow's Weather		
		Sunny	Rainy	Foggy
Today's Weather	Sunny	0.8	0.05	0.15
	Rainy	0.2	0.6	0.2
	Foggy	0.2	0.3	0.5

Probabilities of Tomorrow's weather based on Today's Weather



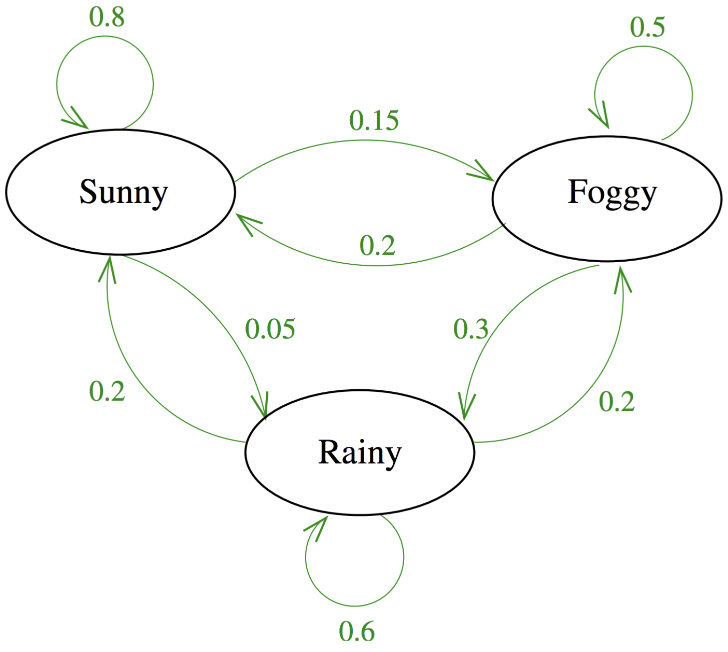
Markov Models

- **Question:** Given that *today is sunny* what's the probability that *tomorrow is sunny* and the *day after is rainy*?

$$P(w_2 = \text{Sunny}, w_3 = \text{Rainy} | w_1 = \text{Sunny})$$

$$\begin{aligned}
&= P(w_3 = \text{Rainy} \mid w_2 = \text{Sunny}, w_1 = \text{Sunny}) * \\
&\quad P(w_2 = \text{Sunny} \mid w_1 = \text{Sunny}) \\
&= P(w_3 = \text{Rainy} \mid w_2 = \text{Sunny}) * \\
&\quad P(w_2 = \text{Sunny} \mid w_1 = \text{Sunny}) \\
&= (0.05)(0.8) \\
&= 0.04
\end{aligned}$$

(Markov assumption)



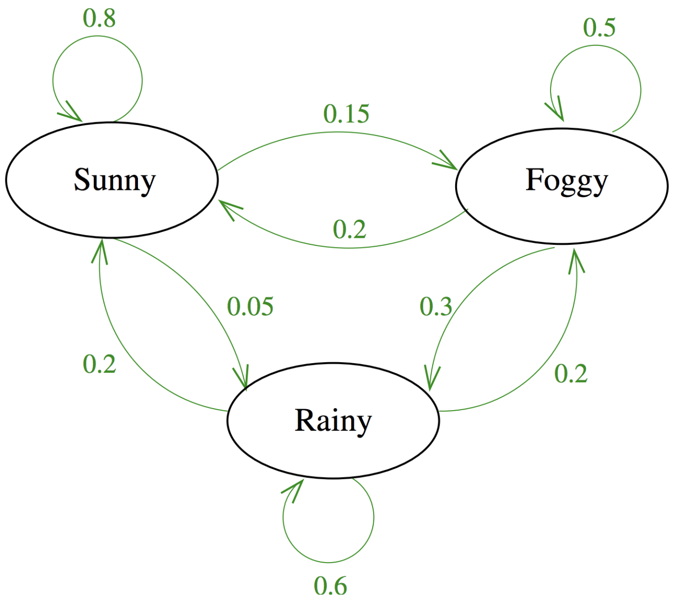
NOTE: You can also think about this as moving through the automaton, multiplying the probabilities along the path you go.

Markov Models

- **Question:** Given that today is foggy what's the probability that it will be rainy two days from now?
- **Answer:** There are three ways to get from foggy today to rainy two days from now:

{foggy, foggy, rainy}, {foggy, rainy, rainy} and {foggy, sunny, rainy}

- Therefore we have to sum over these paths:



$$
\begin{aligned}
& P(w_3 = \text{Rainy} \mid w_1 = \text{Foggy}) \\
= \quad & P(w_2 = \text{Foggy}, w_3 = \text{Rainy} \mid w_1 = \text{Foggy}) + \\
& P(w_2 = \text{Rainy}, w_3 = \text{Rainy} \mid w_1 = \text{Foggy}) + \\
& P(w_2 = \text{Sunny}, w_3 = \text{Rainy} \mid w_1 = \text{Foggy}) + \\
= \quad & P(w_3 = \text{Rainy} \mid w_2 = \text{Foggy}) P(w_2 = \text{Foggy} \mid w_1 = \text{Foggy}) + \\
& P(w_3 = \text{Rainy} \mid w_2 = \text{Rainy}) P(w_2 = \text{Rainy} \mid w_1 = \text{Foggy}) + \\
& P(w_3 = \text{Rainy} \mid w_2 = \text{Sunny}) P(w_2 = \text{Sunny} \mid w_1 = \text{Foggy}) \\
= \quad & (0.3)(0.5) + (0.6)(0.3) + (0.05)(0.2) \\
= \quad & 0.34
\end{aligned}
$$

Hidden Markov Models

Hidden Markov Models – Example

Alice

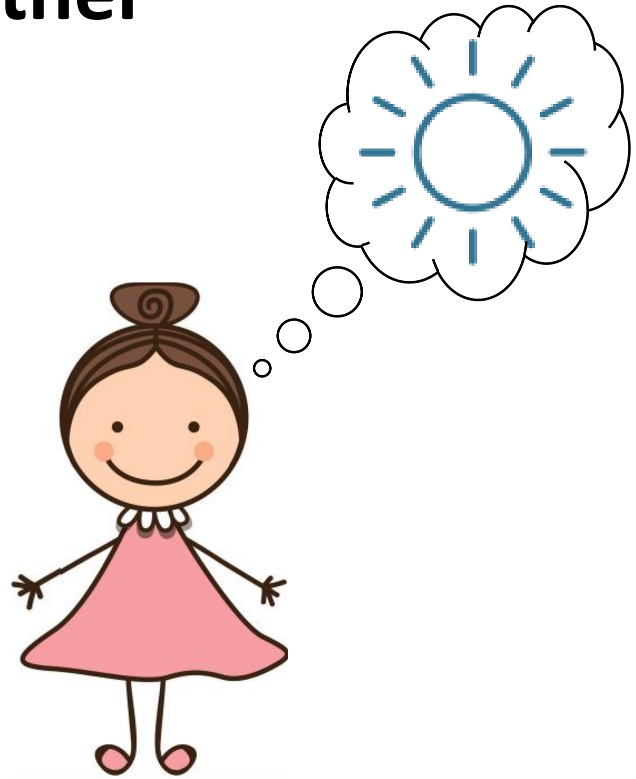
Bob

Images Source: vectorStock

This example is based on video lecture of Luis: https://www.youtube.com/watch?v=kqSzLo9fenk

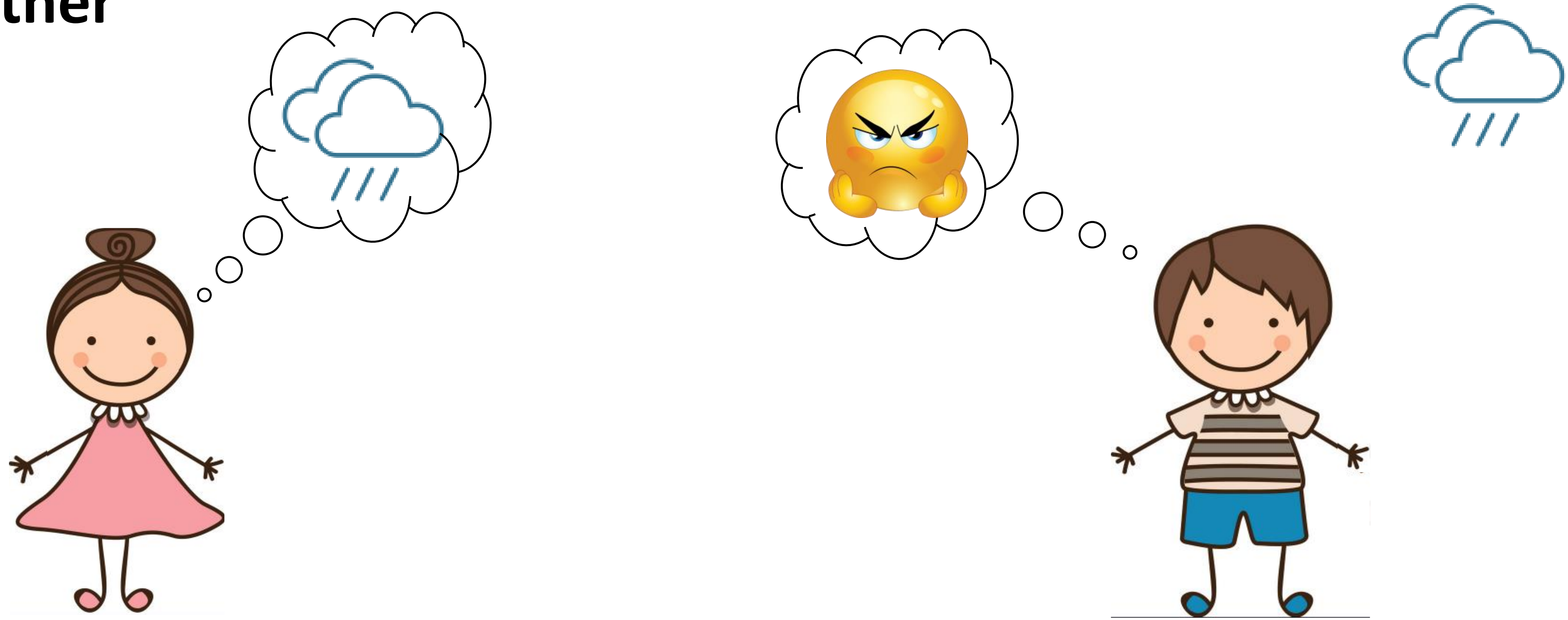
Hidden Markov Models

Weather



Hidden Markov Models

Weather

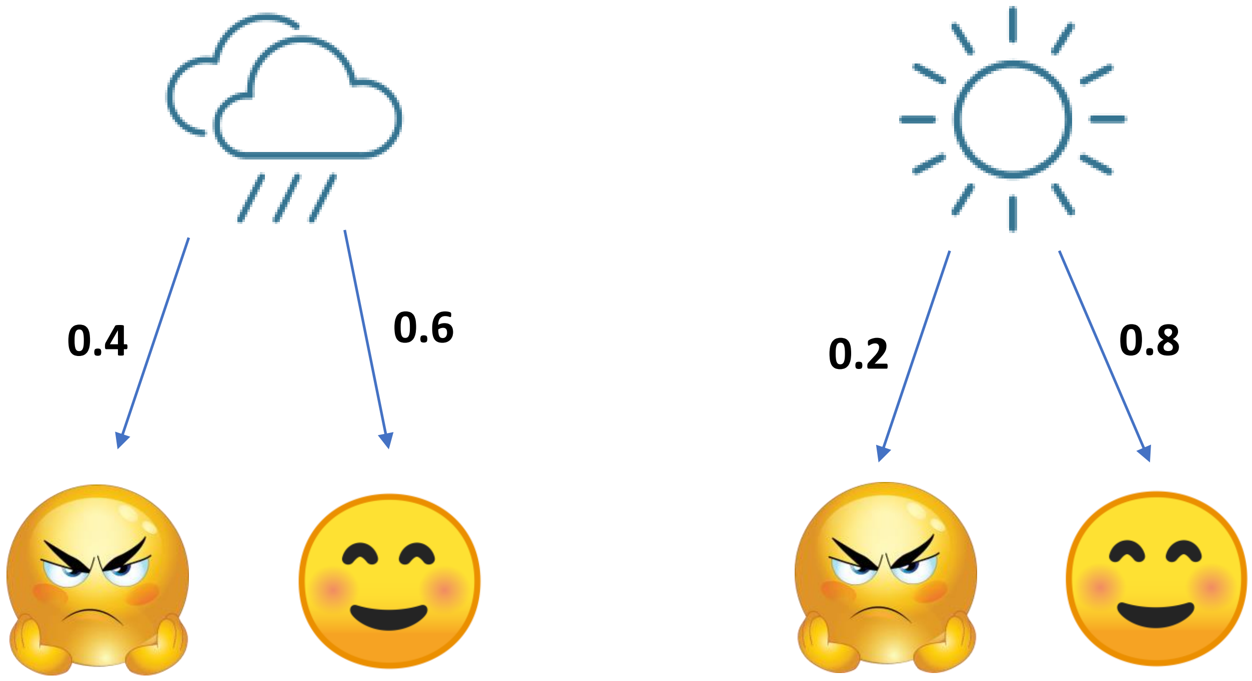


Hidden Markov Models

Weather

Hidden Markov Models

Weather

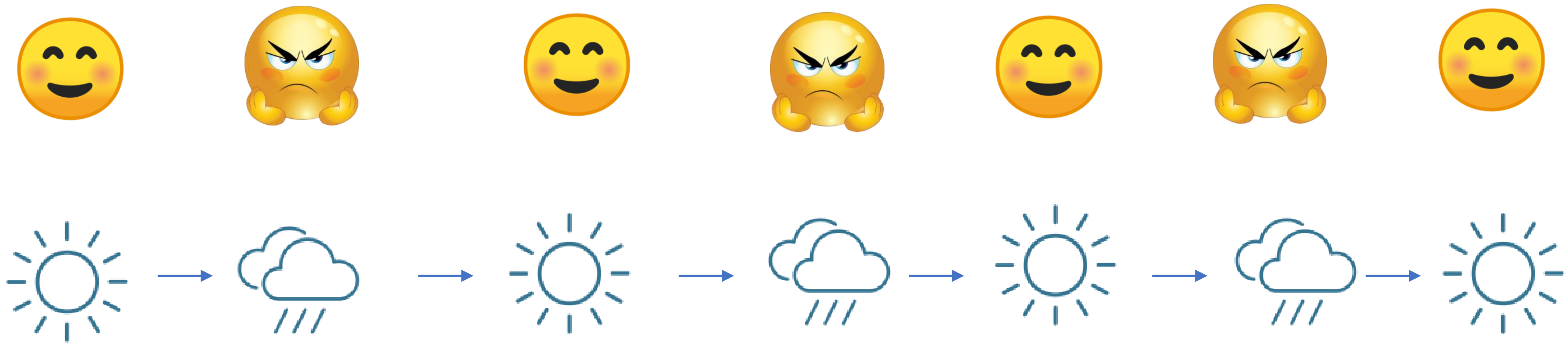


Hidden Markov Models

Weather

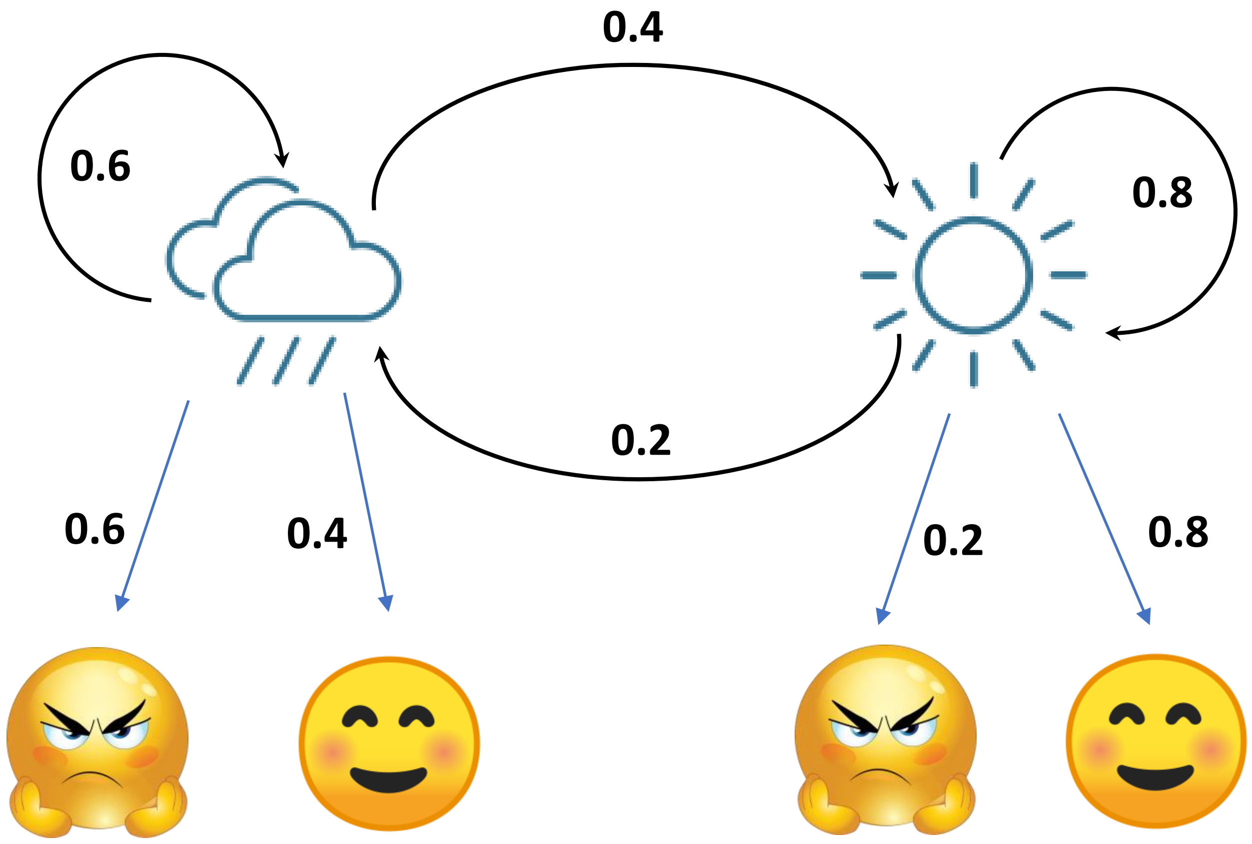
Hidden Markov Models

Monday Tuesday Wednesday Thursday Friday Saturday Sunday

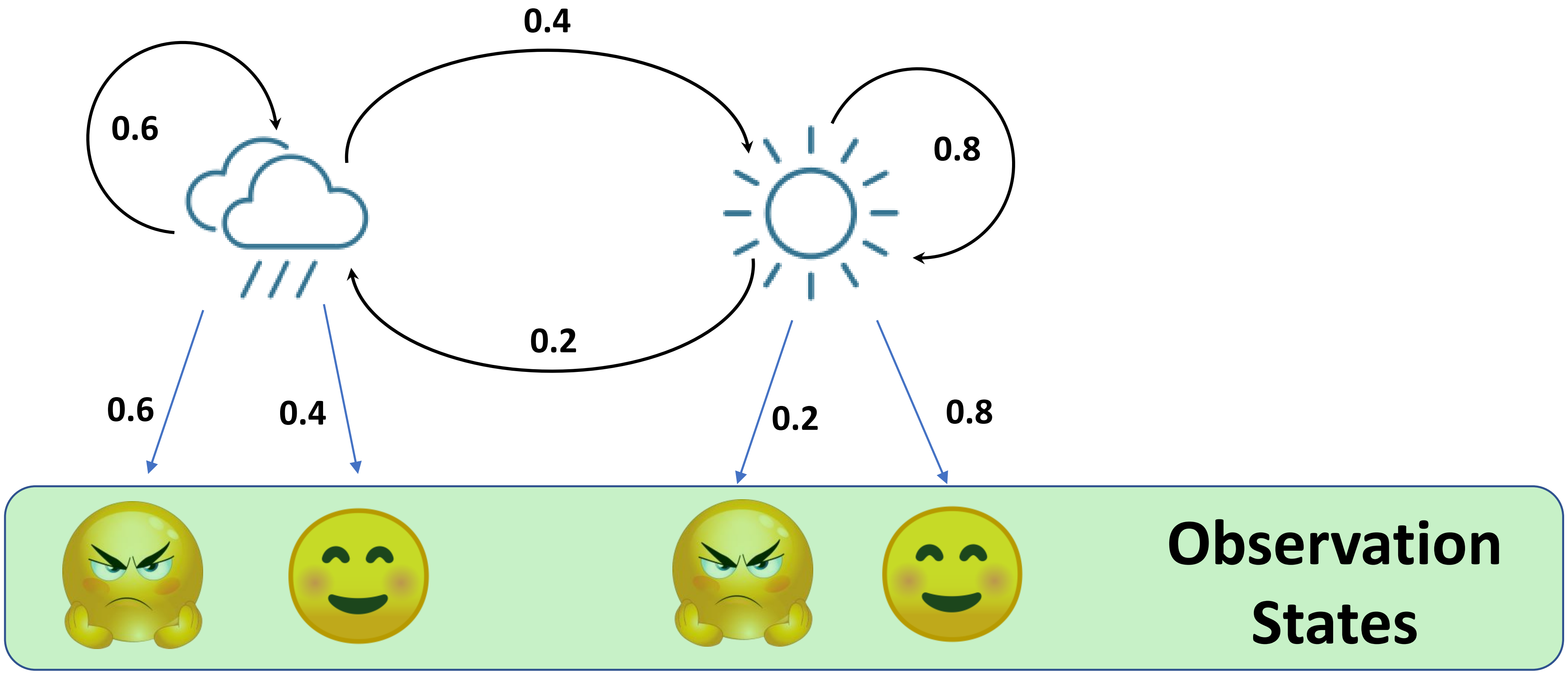


- If we look at the weather change!! It seems not practical ☺.
  So lets go back and see his mood.

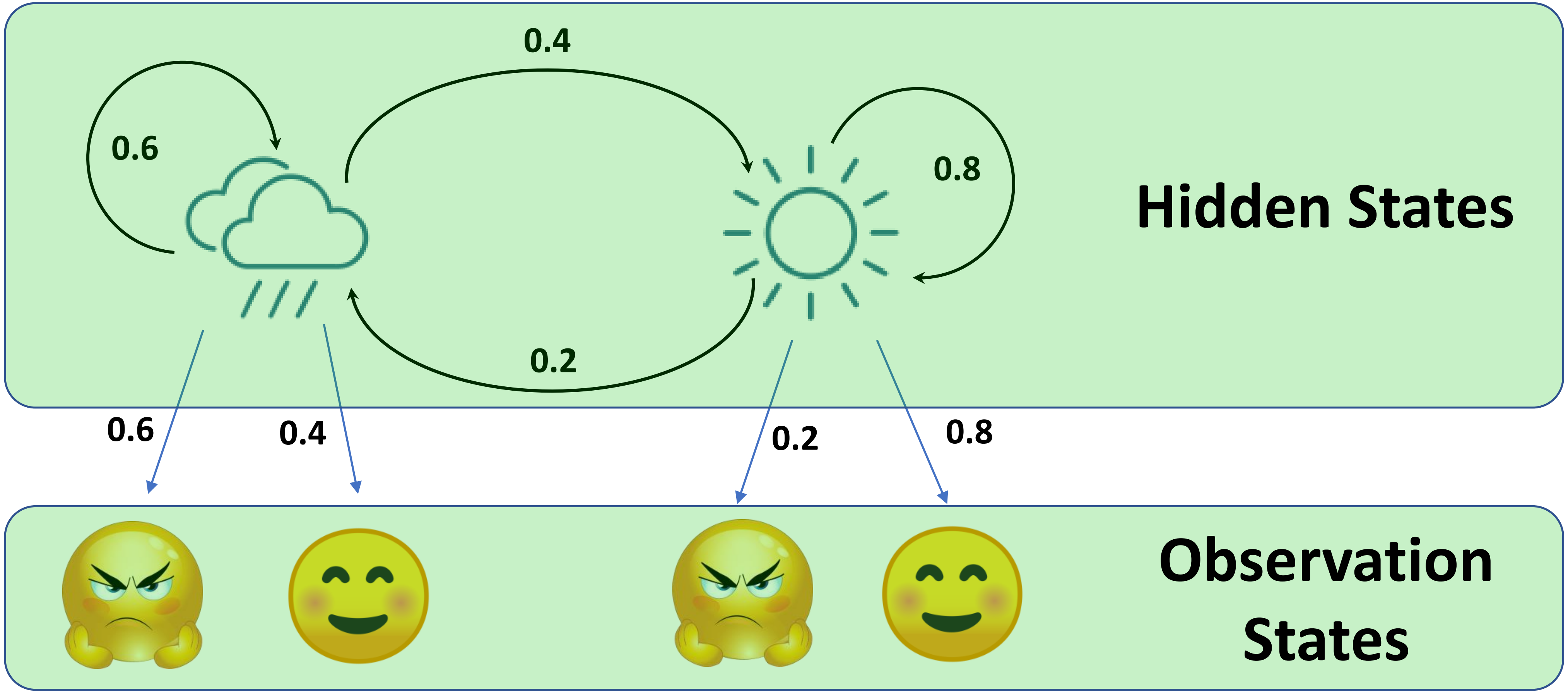
Hidden Markov Models



Hidden Markov Models

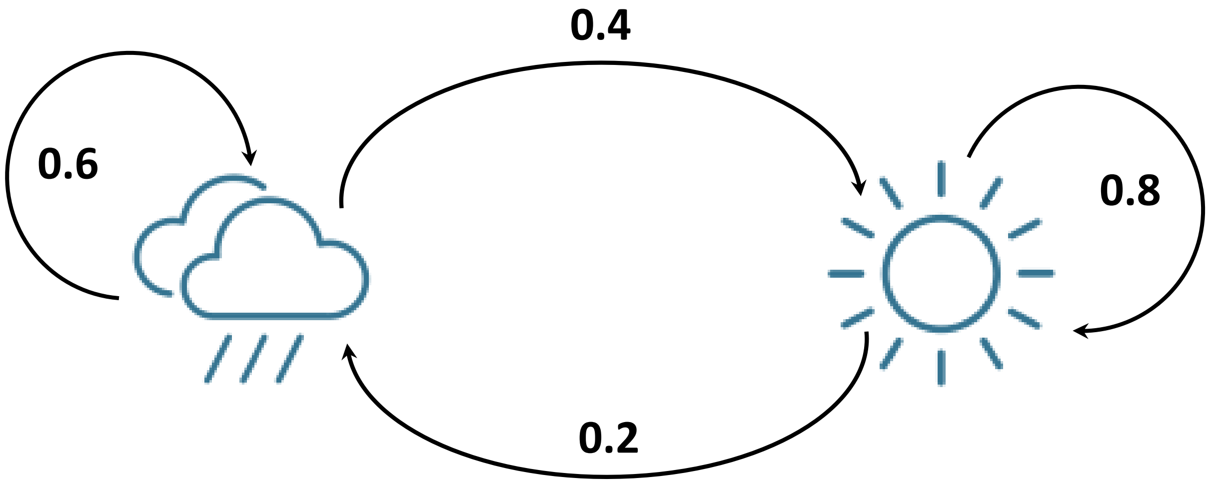


Hidden Markov Models



Hidden Markov Models

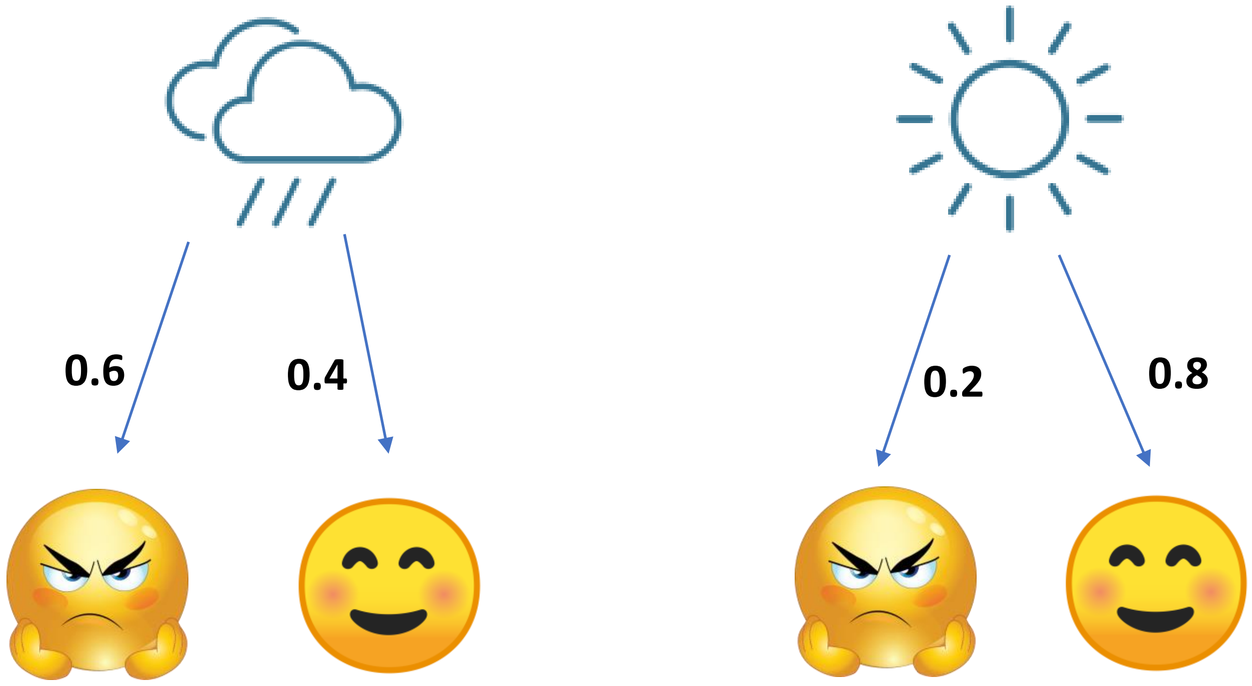
Hidden States



Transition Probabilities

Hidden Markov Models

Observation States



Emission Probabilities

Hidden Markov Models

- **Questions**

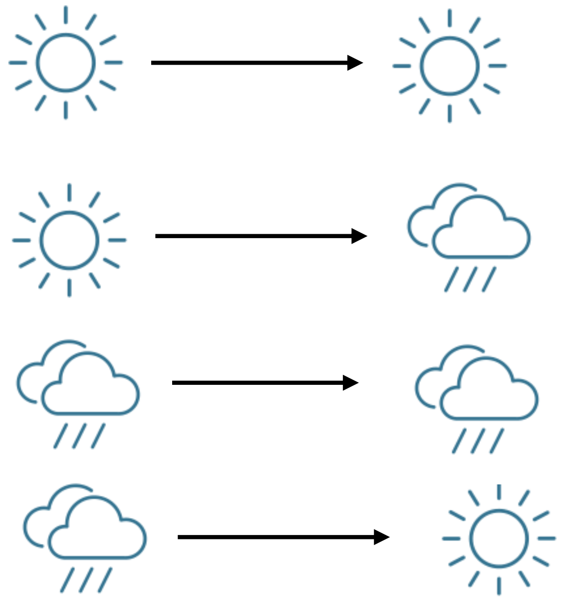
1. How did we find these probabilities?
2. What’s the probability that a random day is Sunny or Rainy?
3. If Bob is Happy today, what’s the probability that it’s Sunny or Rainy?
4. If for three days Bob is Happy, Grumpy, Happy, what was the weather?

Hidden Markov Models

1. How did we find these probabilities?

We have history of weather as follows:

- How many sunny days followed by sunny day?
- Similarly:



= 8 0.8 (Normalized Prob.)

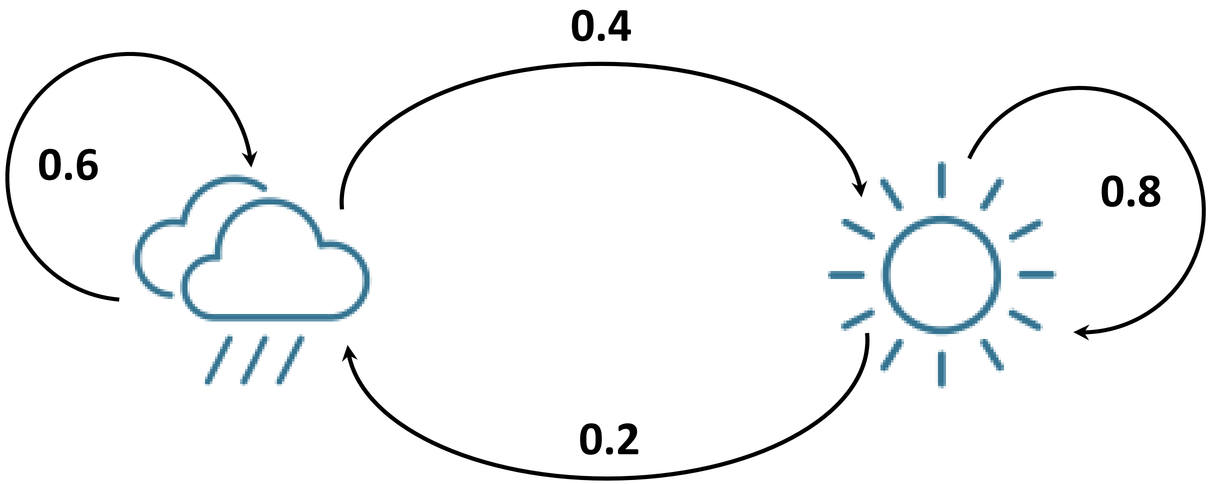
= 2 0.2 (Normalized Prob.)

= 3 0.6 (Normalized Prob.)

= 2 0.4 (Normalized Prob.)

Hidden Markov Models

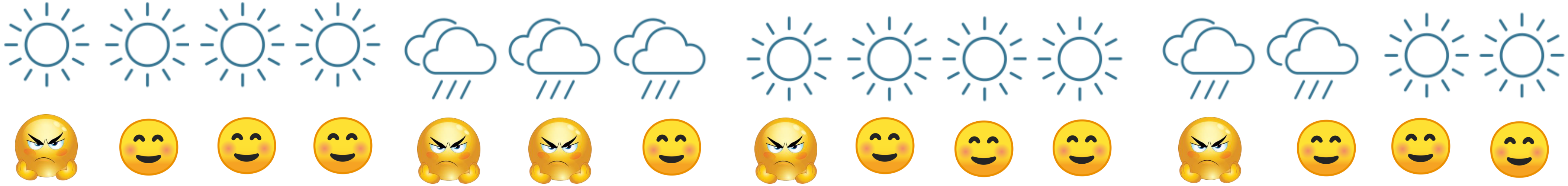
- In this way, we calculate the transition probabilities.

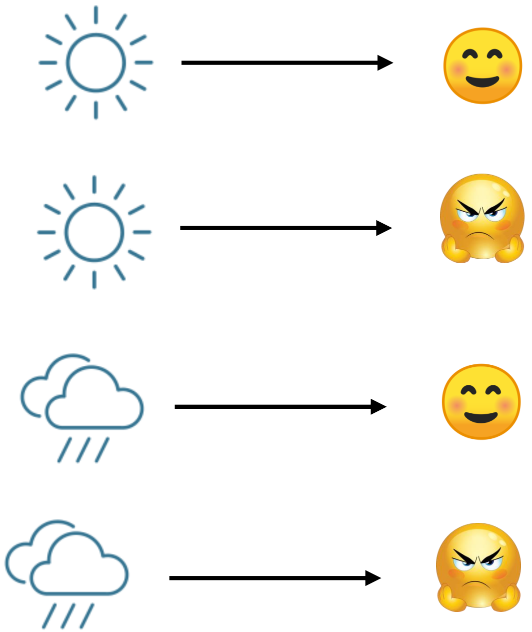


Transition Probabilities

Hidden Markov Models

1. How did we find these probabilities?





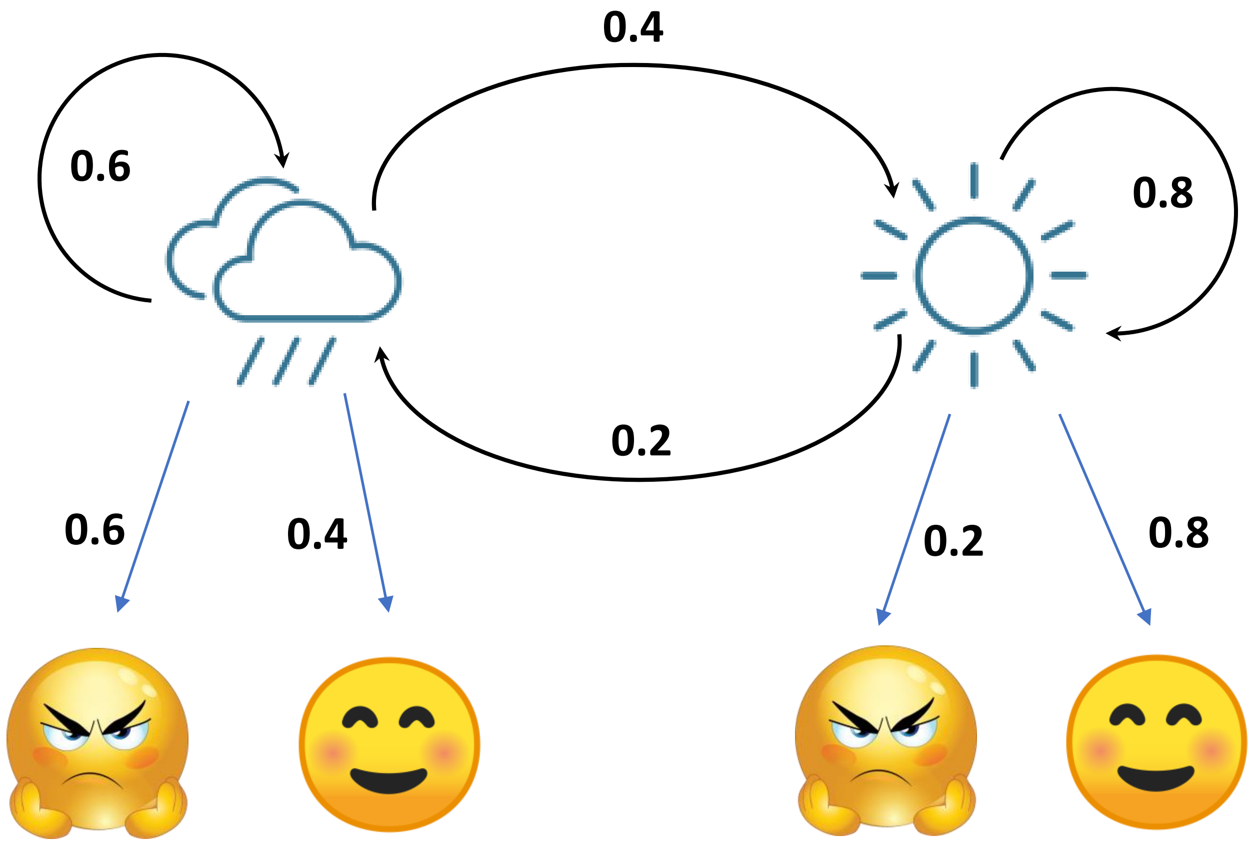
= 8 0.8 (Normalized Prob.)

= 2 0.2 (Normalized Prob.)

= 2 0.4 (Normalized Prob.)

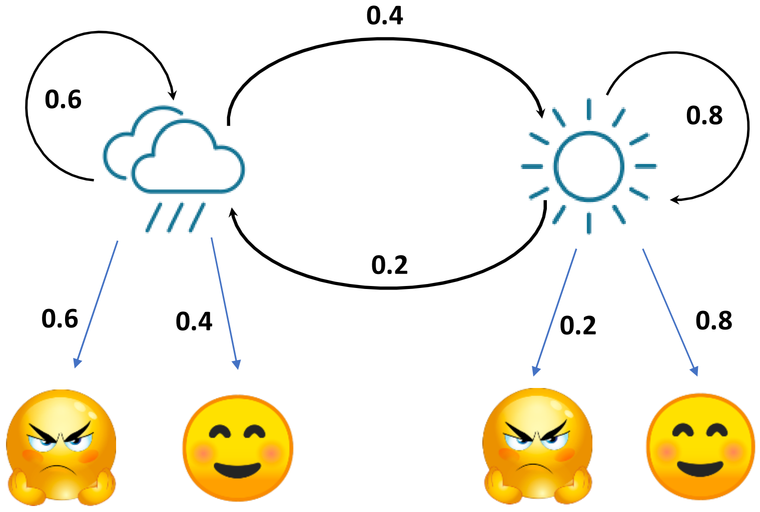
= 3 0.6 (Normalized Prob.)

Hidden Markov Models



Hidden Markov Models

2. What's the probability that a random day is Sunny or Rainy?



$$R = 0.6R + 0.2S \qquad S = 0.8S + 0.4R$$

$$S + R = 1$$

$$S = 2/3 \text{ and } R = 1/3$$

Hidden Markov Models

3. **If Bob is Happy today, what's the probability that it's Sunny or Rainy?**

Hidden Markov Models

3. **If Bob is Happy today, what's the probability that it's Sunny or Rainy?**

Hidden Markov Models

Hidden Markov Models

- **Bayes Theorem**

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

Thomas Bayes
1701 – 1761

- $P(h)$ = prior probability of hypothesis h
- $P(D)$ = prior probability of training data D
- $P(h|D)$ = probability of h given D
- $P(D|h)$ = probability of D given h

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h|D) \\ &= \arg\max_{h \in H} \frac{P(D|h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D|h)P(h) \end{aligned}$$

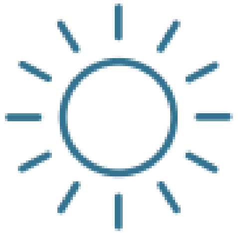
If assume $P(h_i) = P(h_j)$ then can further simplify, and choose the *Maximum likelihood* (ML) hypothesis

$$h_{ML} = \arg\max_{h_i \in H} P(D|h_i)$$

Hidden Markov Models

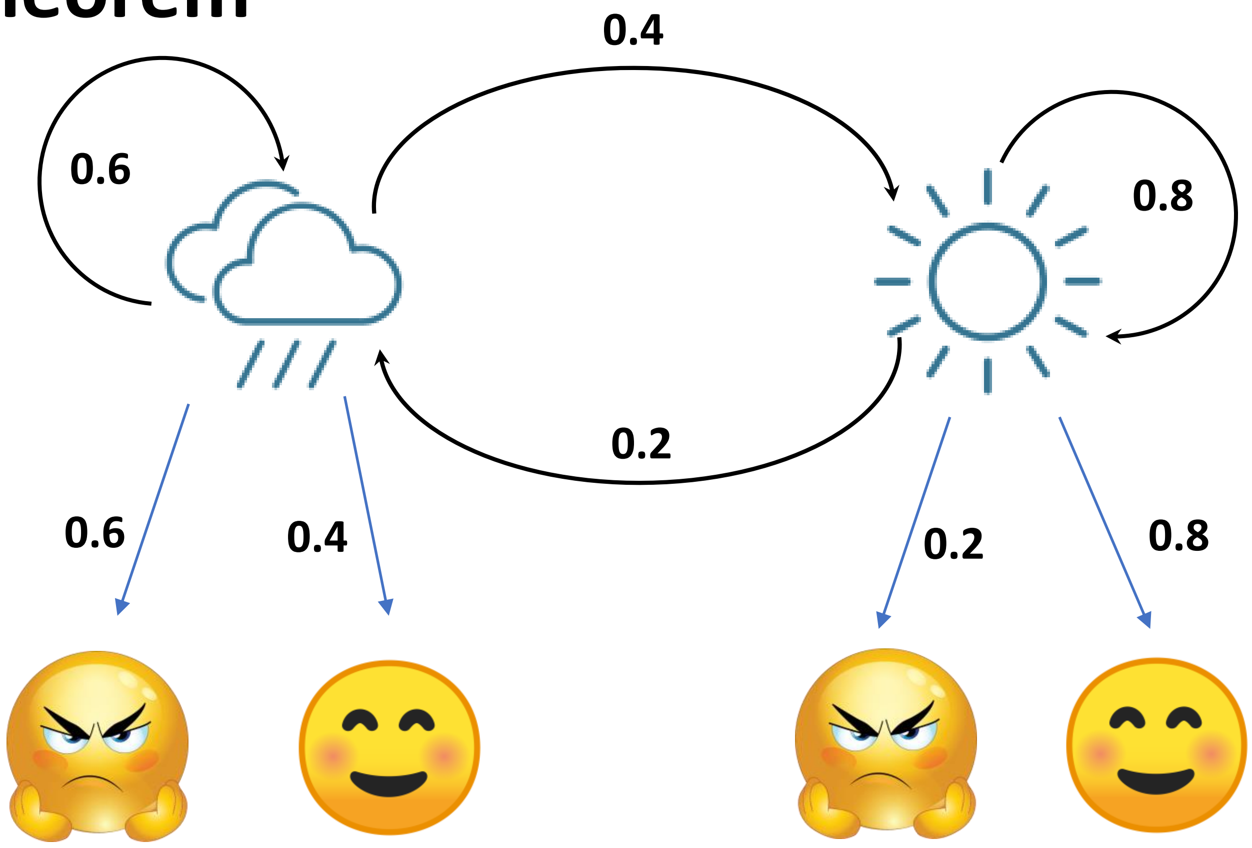
- **Today, Bob is feeling happy**

Monday



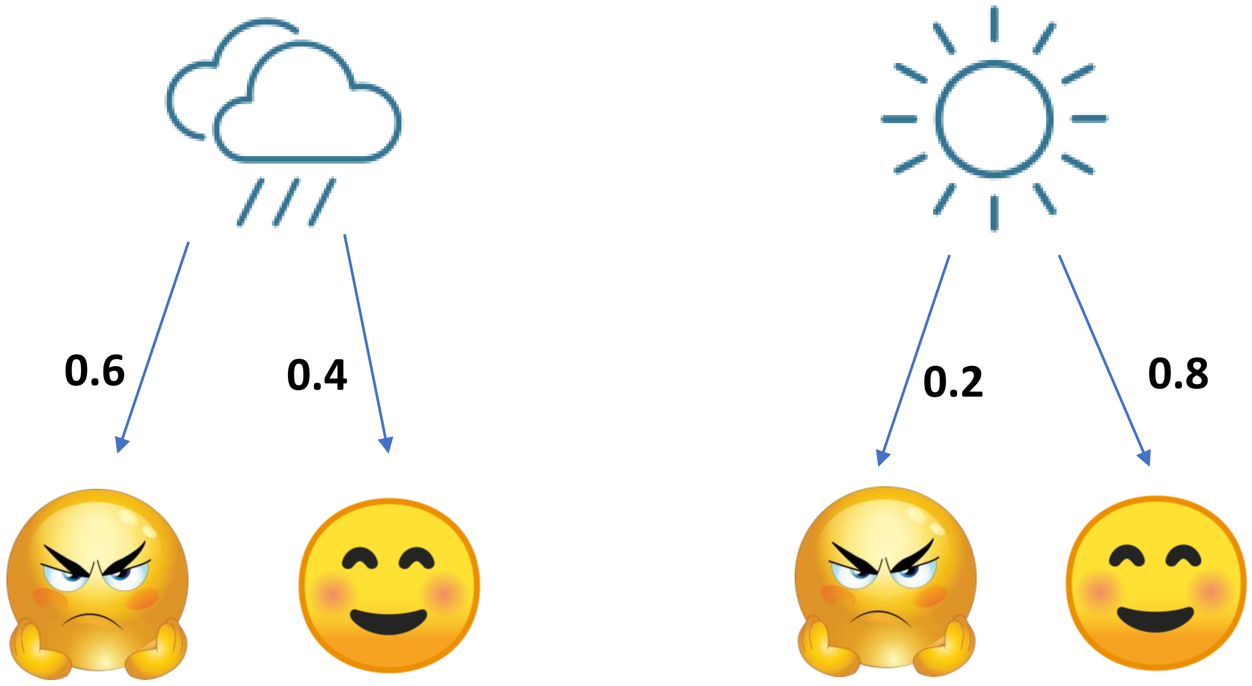
Hidden Markov Models

- **Bayes Theorem**



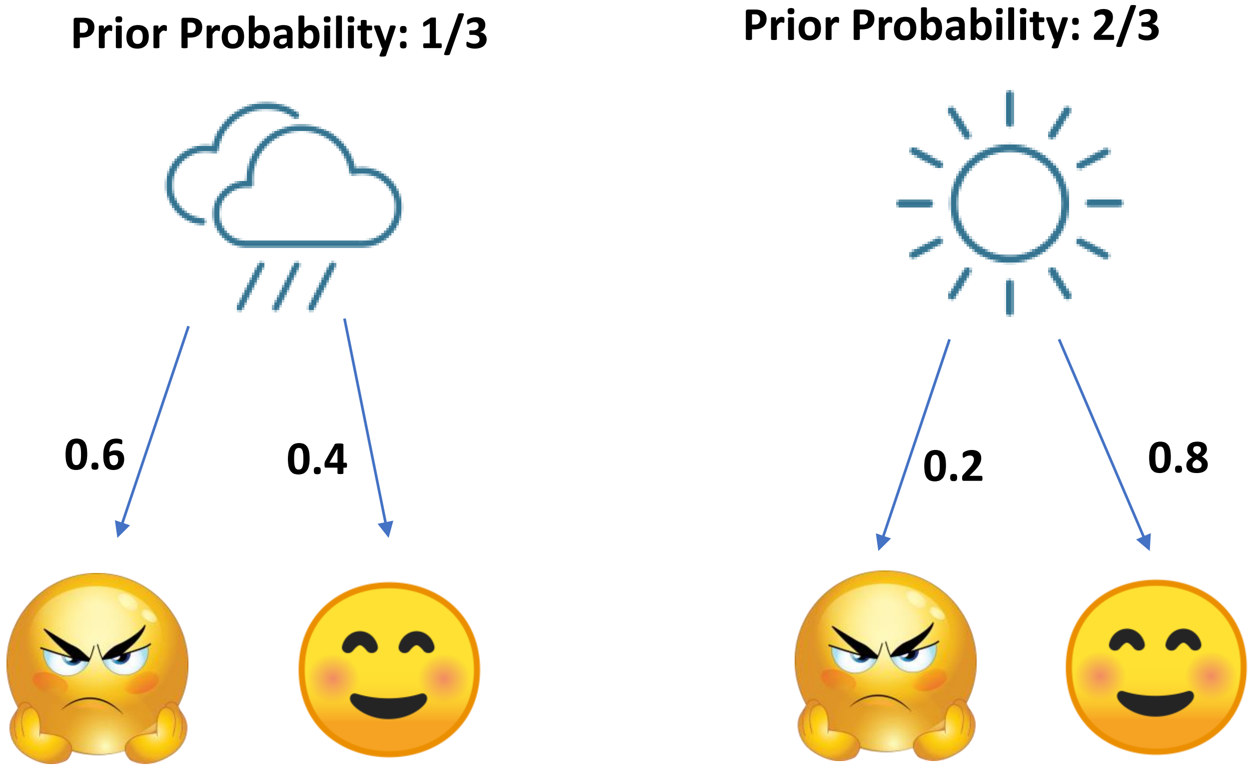
Hidden Markov Models

- **Bayes Theorem**



Hidden Markov Models

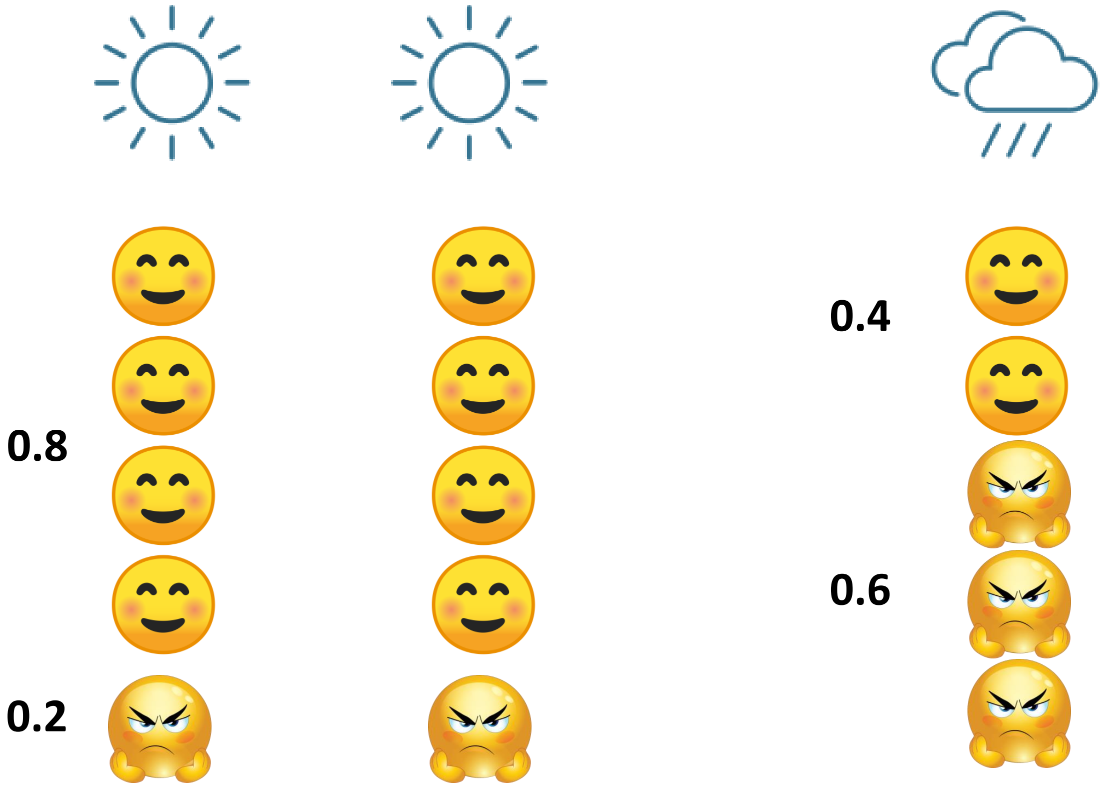
- **Bayes Theorem**

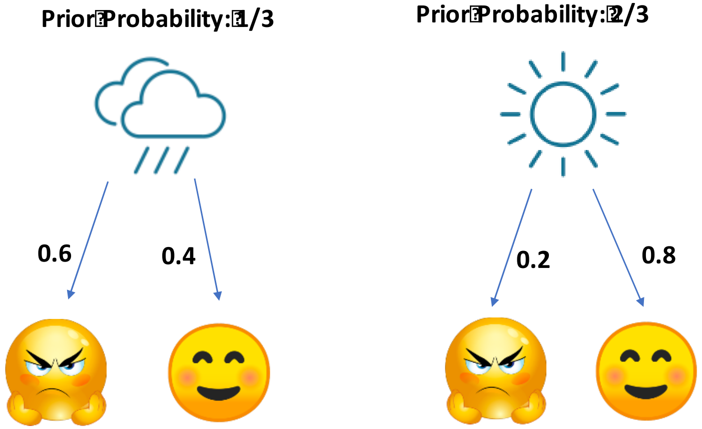


Hidden Markov Models

- **Bayes Theorem**

Prior Probability: 2/3 **Prior Probability: 1/3**

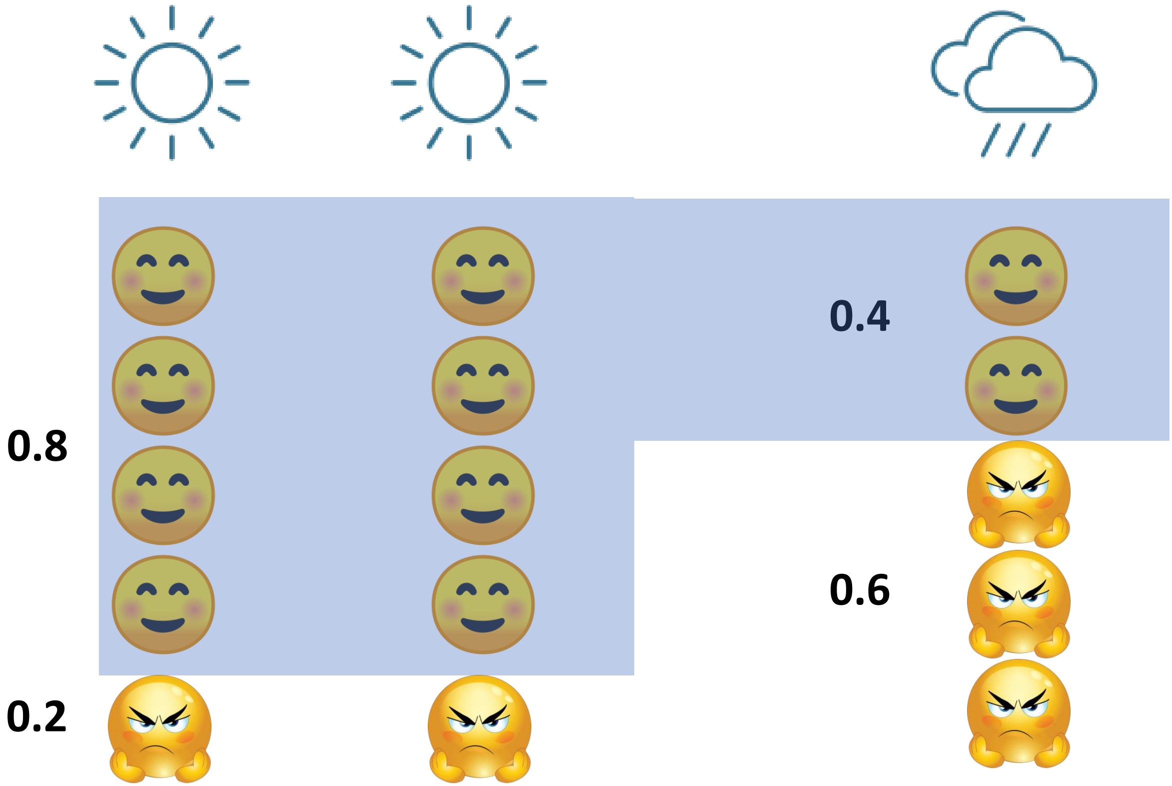




Hidden Markov Models

- **Bayes Theorem**

Prior Probability: 2/3 **Prior Probability: 1/3**



If (Bob = Happy)
P(Sunny | Happy) = 8/10
P(Rainy | Happy) = 2/10

If (Bob = Grumpy)
P(Sunny | Grumpy) = 2/5
P(Rainy | Grumpy) = 3/5

Hidden Markov Models

4. If for three days Bob is Happy, Grumpy, Happy, what was the weather?

We need to calculate most likelihood in this scenario.

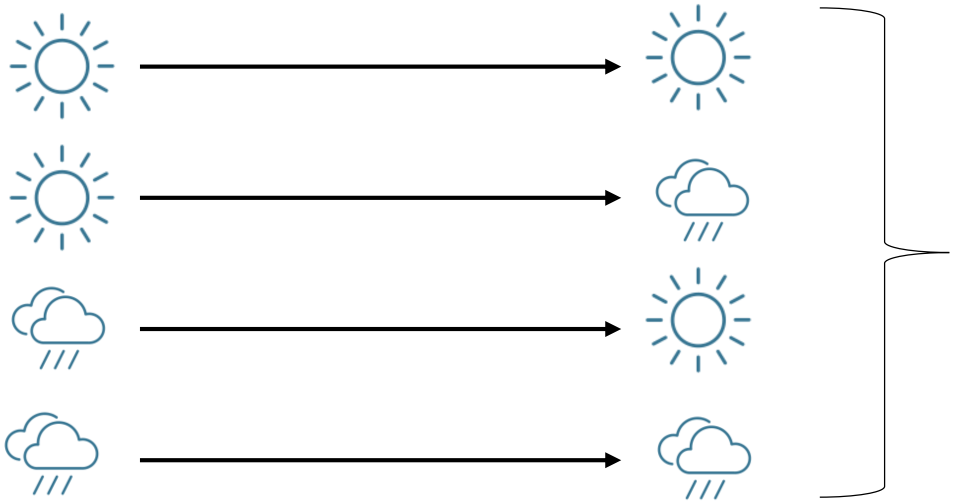
Hidden Markov Models

- **First Simplest Case:** if happy-grumpy, what's weather?

Wednesday

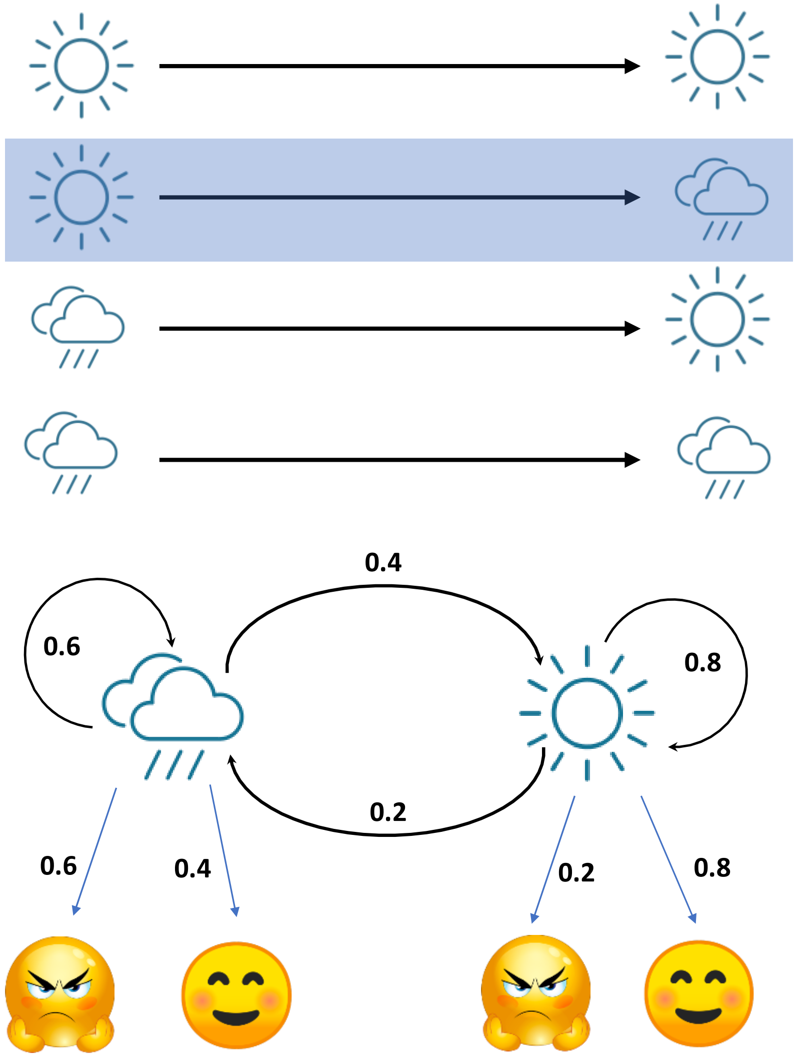
Thursday

How many possible scenarios for the weather?

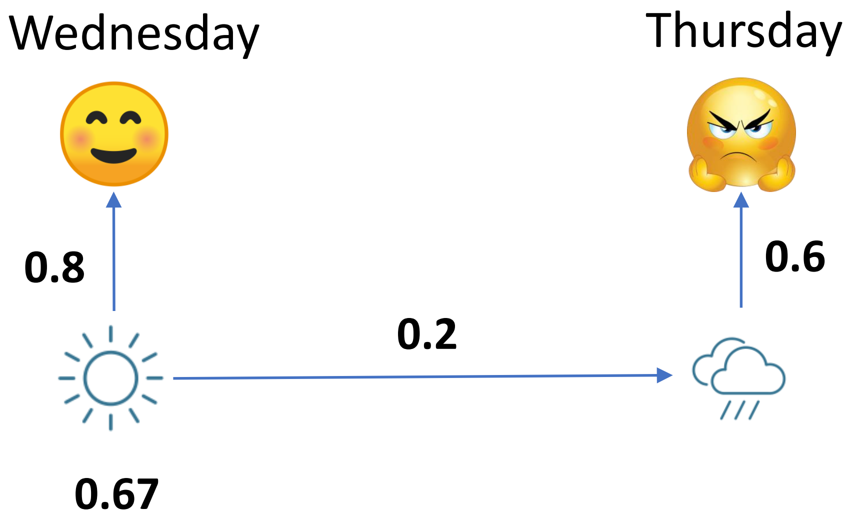


Calculate maximum likelihood

Calculate Maximum Likelihood

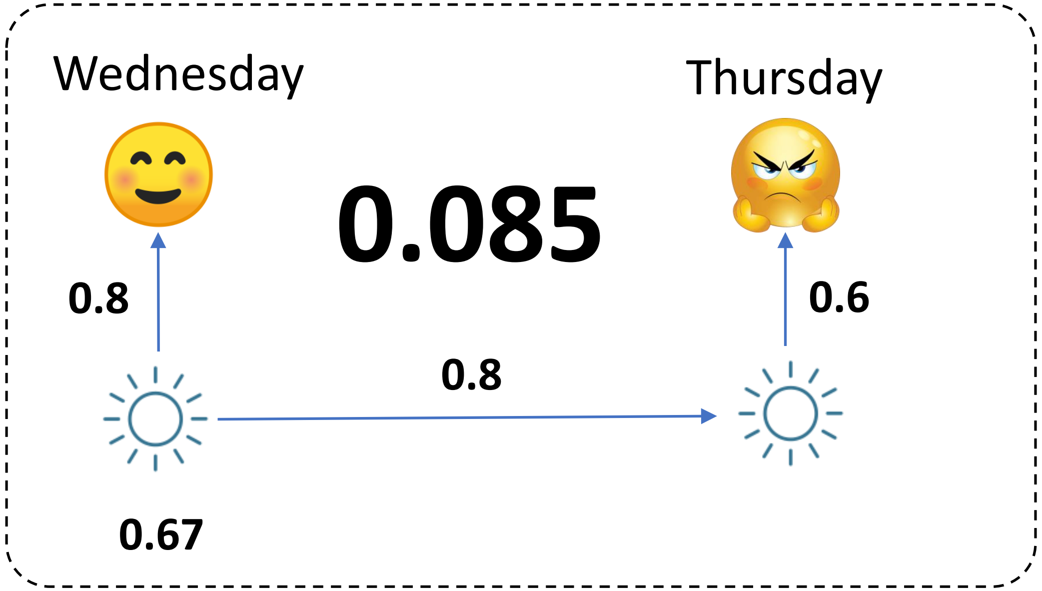


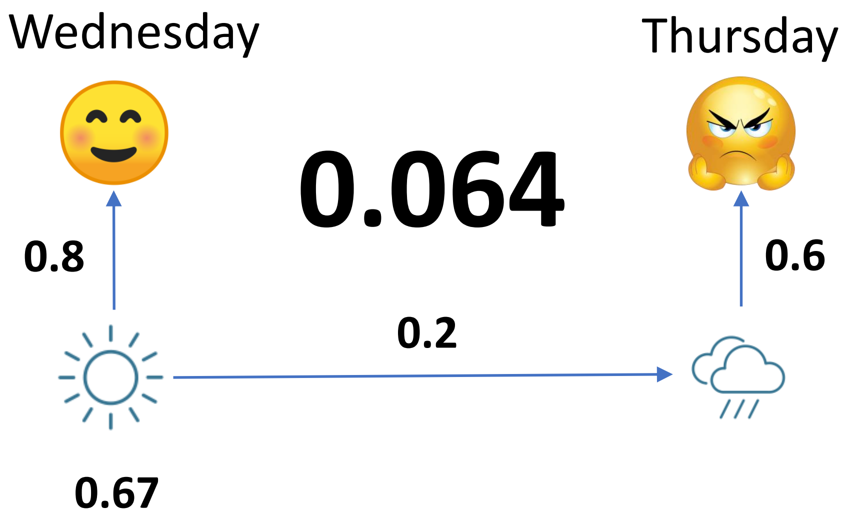
We know that:

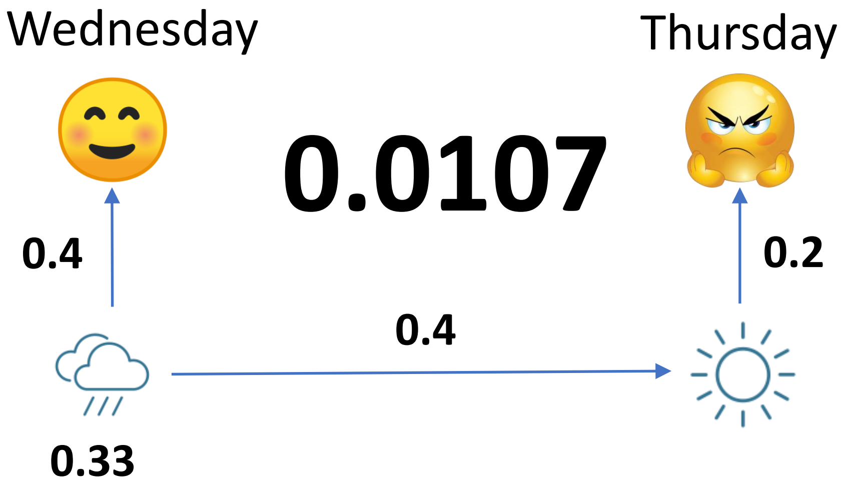


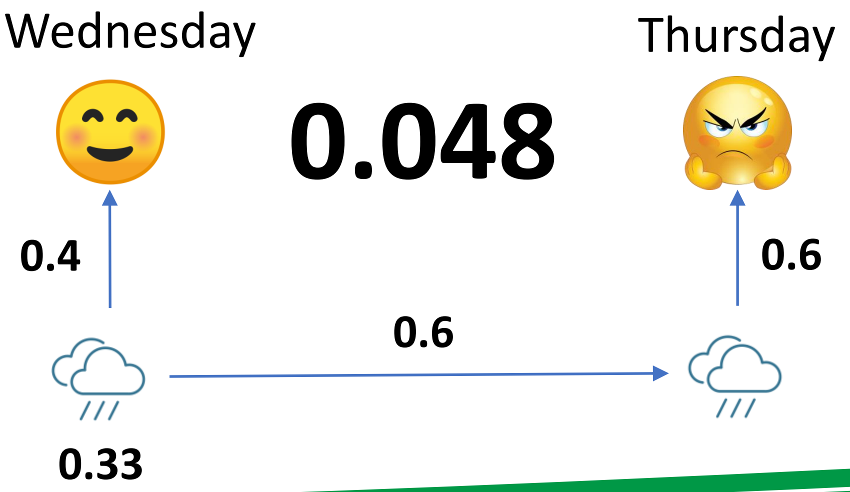
0.064

Calculate Maximum Likelihood







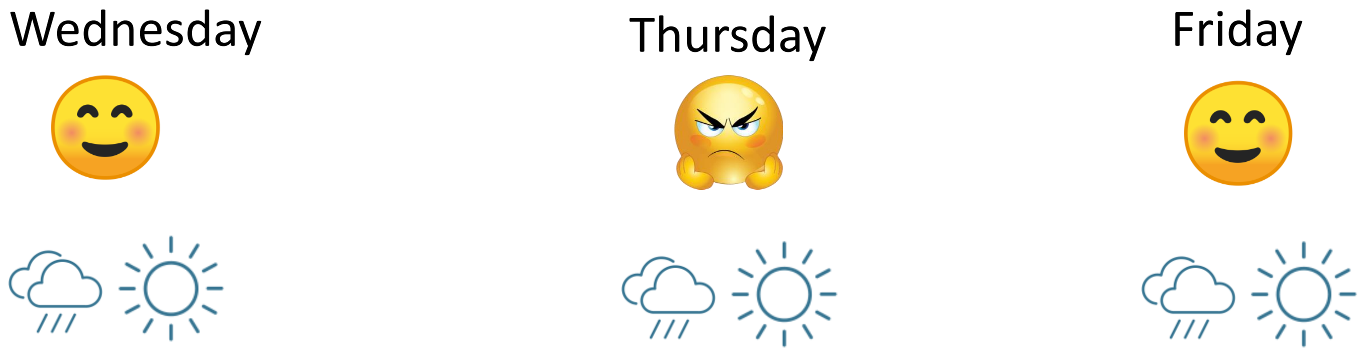


Calculate Maximum Likelihood

- **Question:** if happy-grumpy, what’s weather?
- **Answer:** Sunny, Sunny

Another Scenario

- If happy-grumpy-happy, what’s weather?



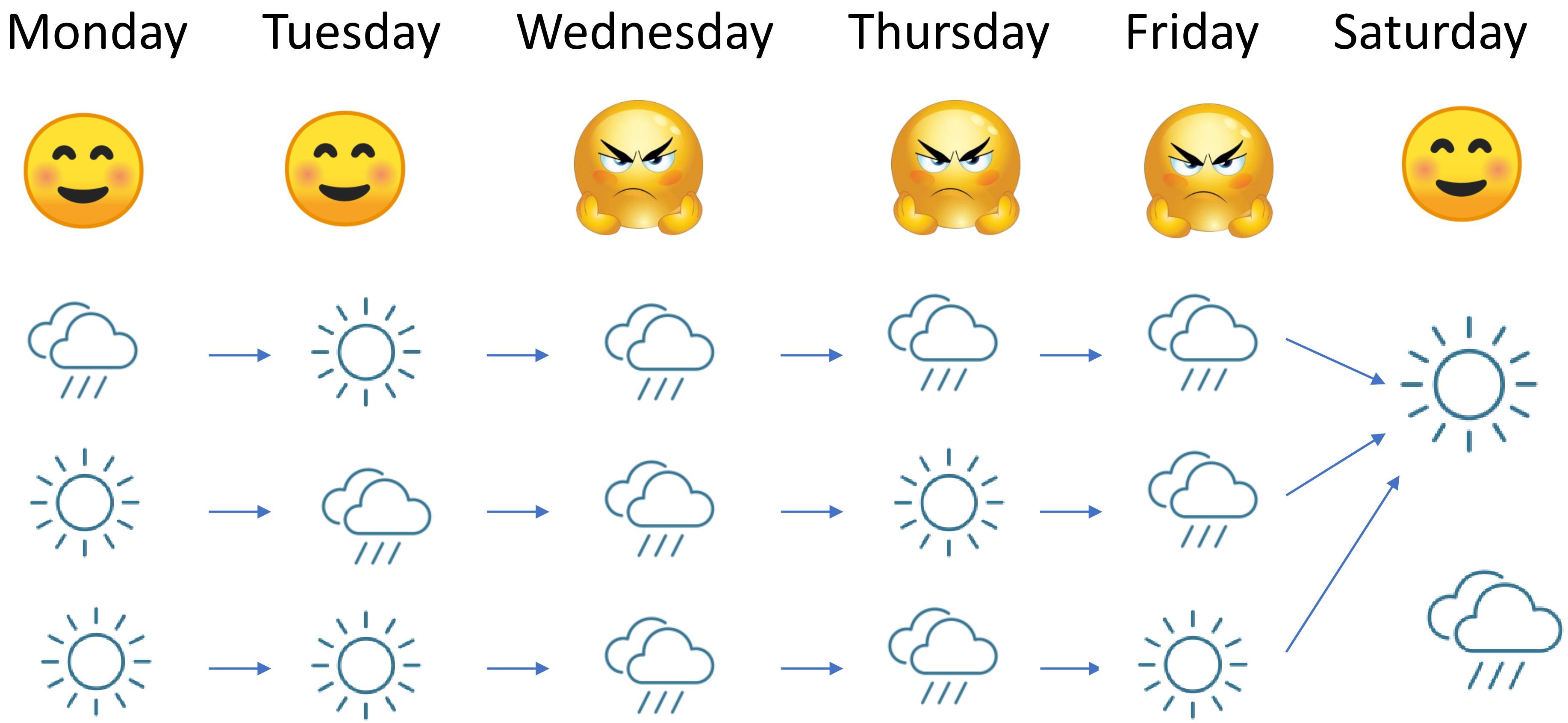
- So we have to look 8 possible scenarios to calculate the maximum likelihood.

Lets say we have 5 days. It’s getting more and more Complex!!

Viterbi Algorithm

It is a dynamic programming algorithm for finding the most likely sequence of hidden states—called the Viterbi path—that results in a sequence of observed events, especially in the context of Markov information sources and hidden Markov models.

Viterbi Algorithm

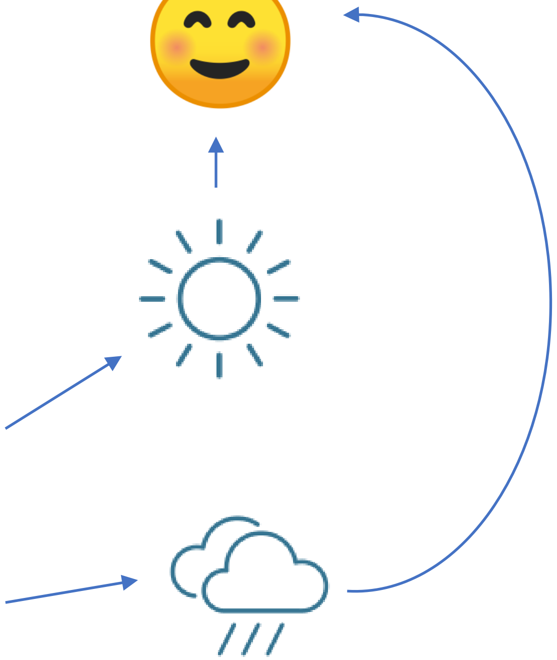


Viterbi Algorithm

Monday Tuesday Wednesday Thursday Friday Saturday

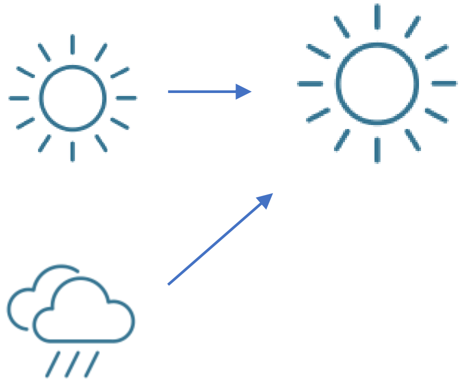
Best Path up to here!!

Best Path up to here!!

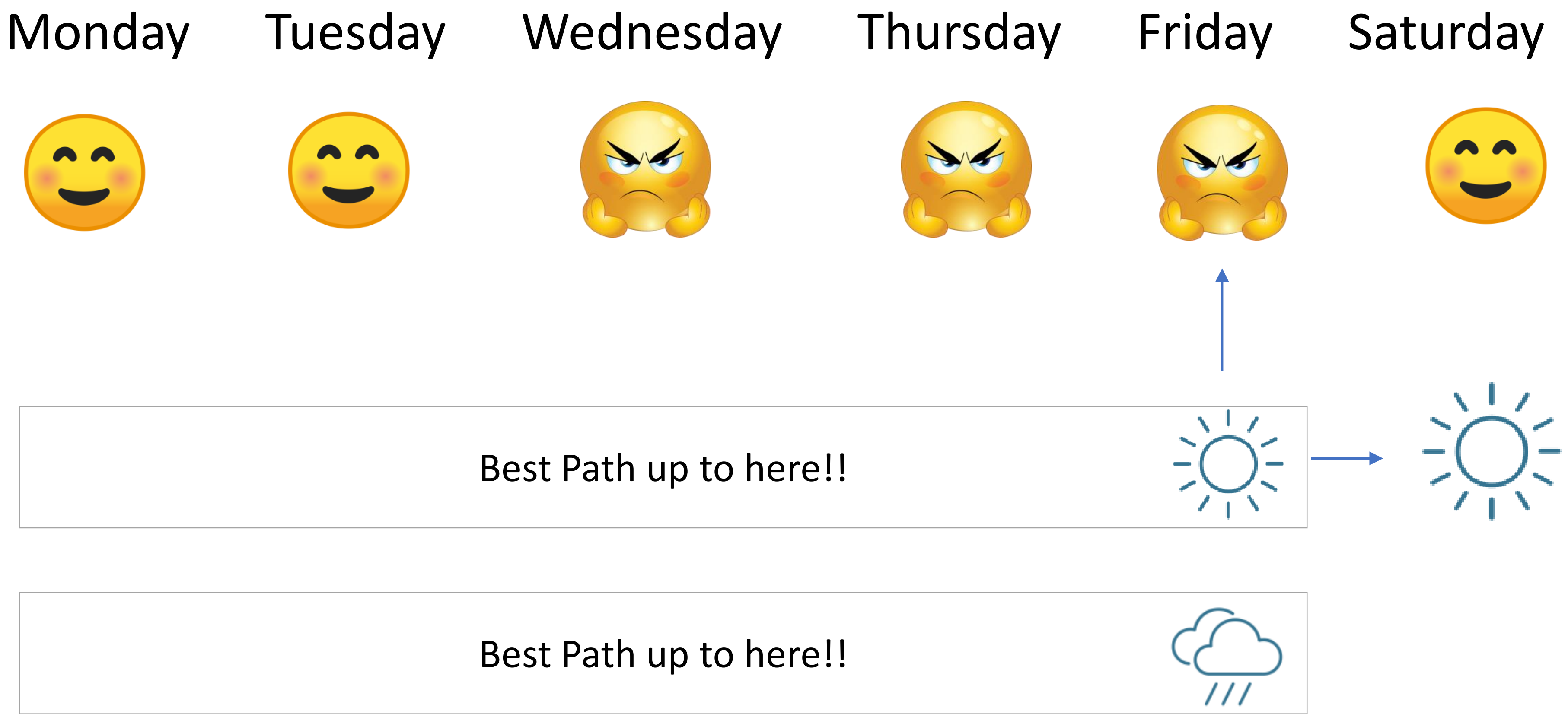


Viterbi Algorithm

Monday Tuesday Wednesday Thursday Friday Saturday

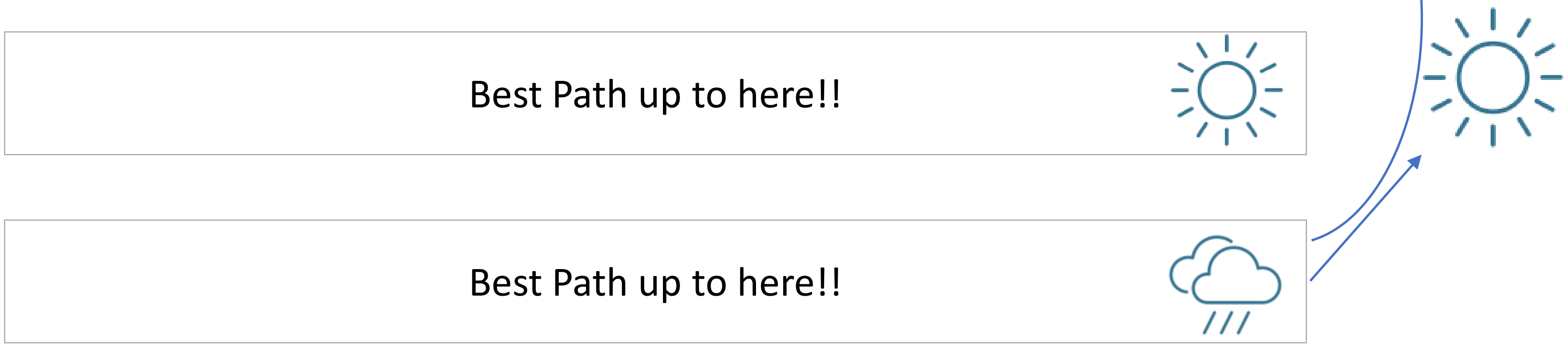


Viterbi Algorithm

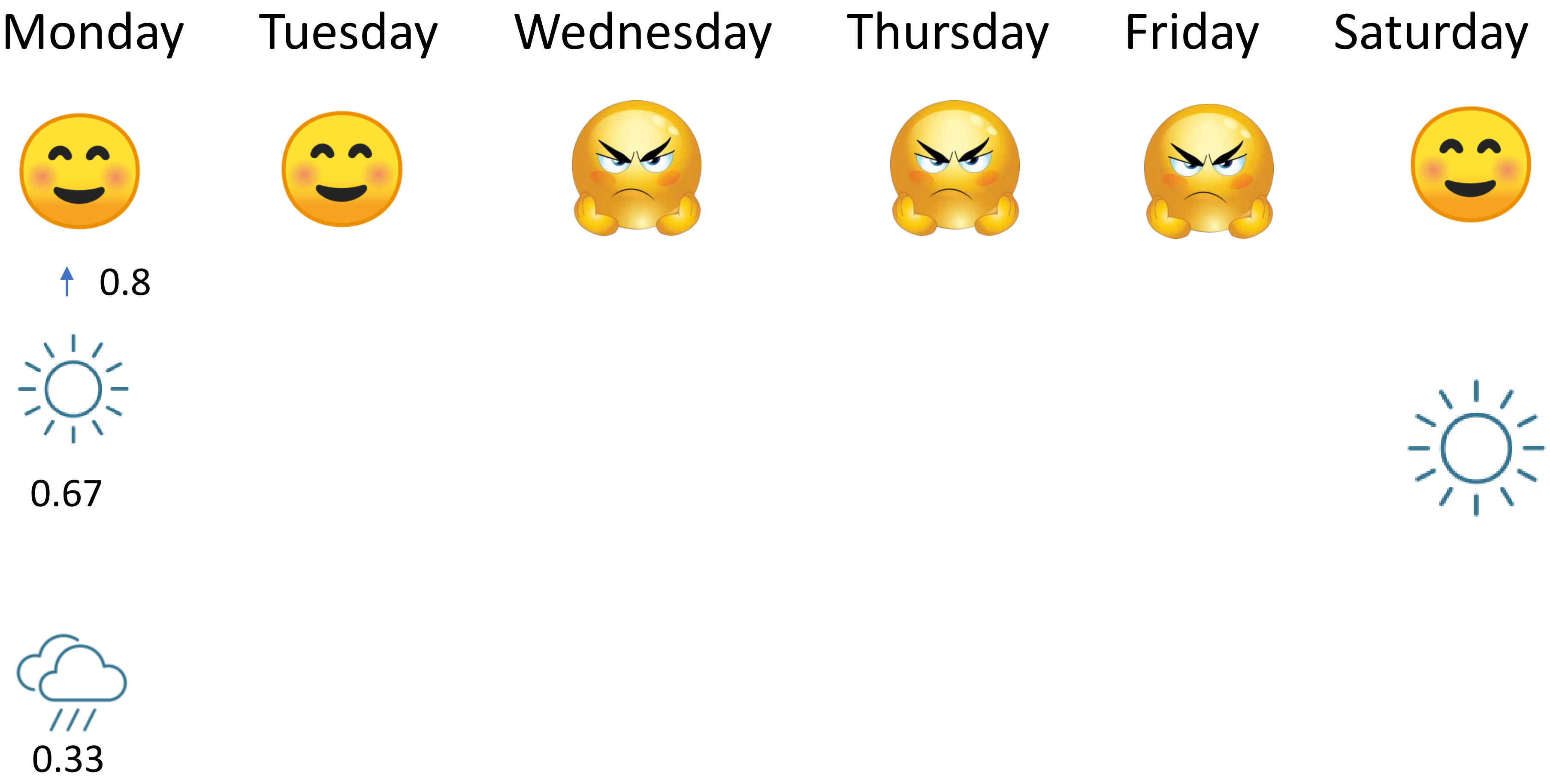


Viterbi Algorithm

Monday Tuesday Wednesday Thursday Friday Saturday



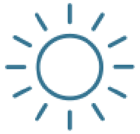
Viterbi Algorithm



Viterbi Algorithm

Monday Tuesday Wednesday Thursday Friday Saturday

↑ 0.8

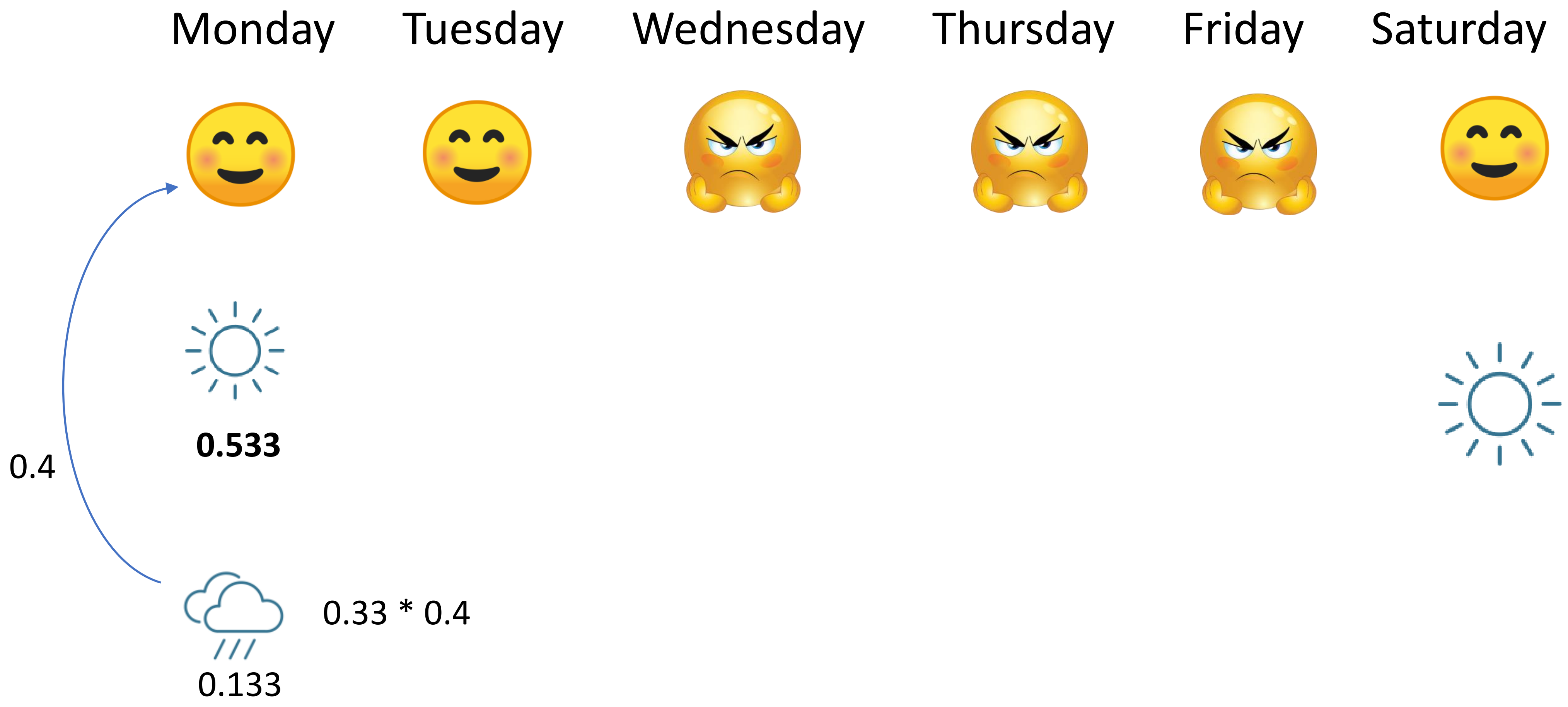


0.67 * 0.8

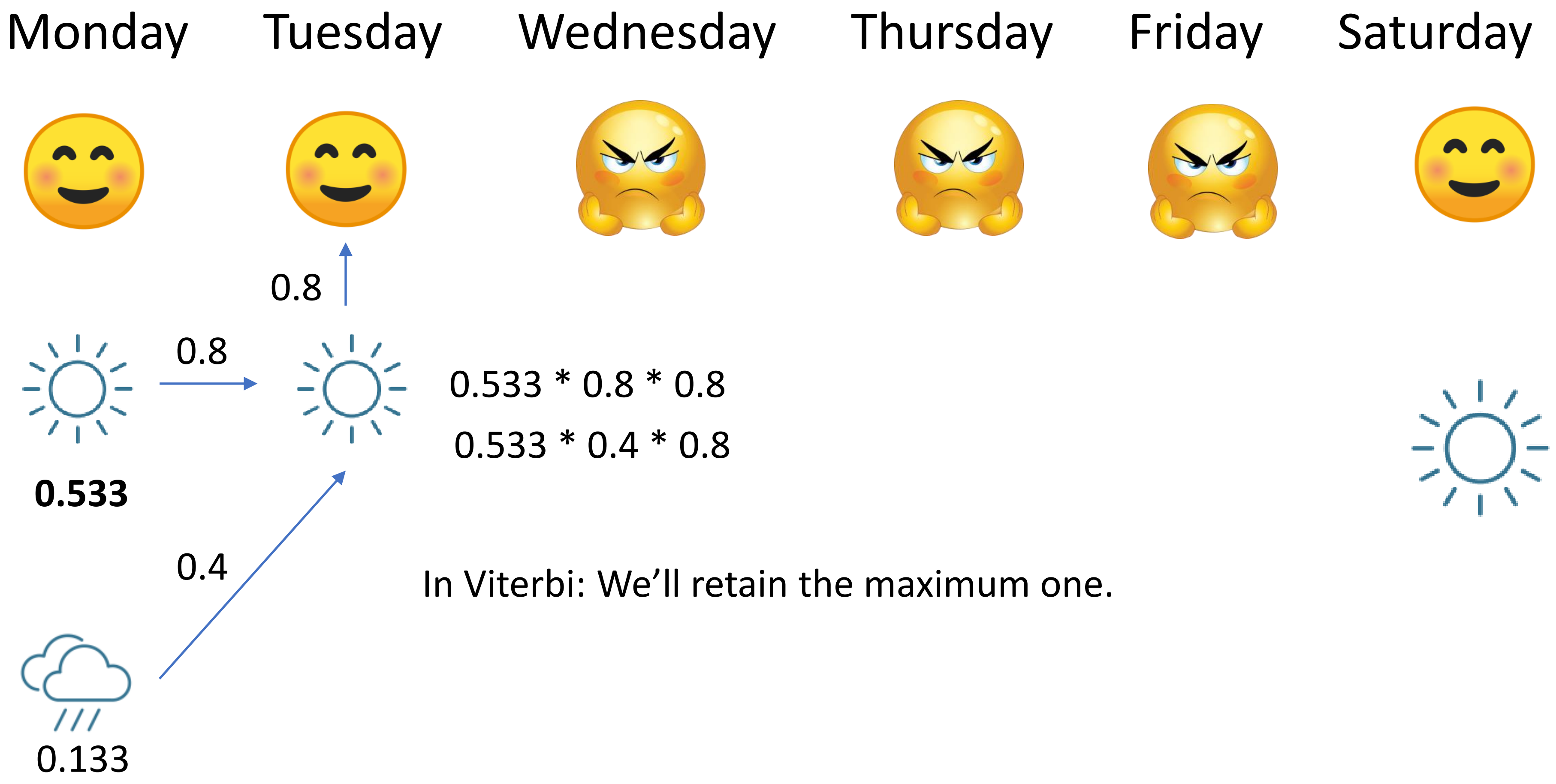
0.533

0.33

Viterbi Algorithm



Viterbi Algorithm

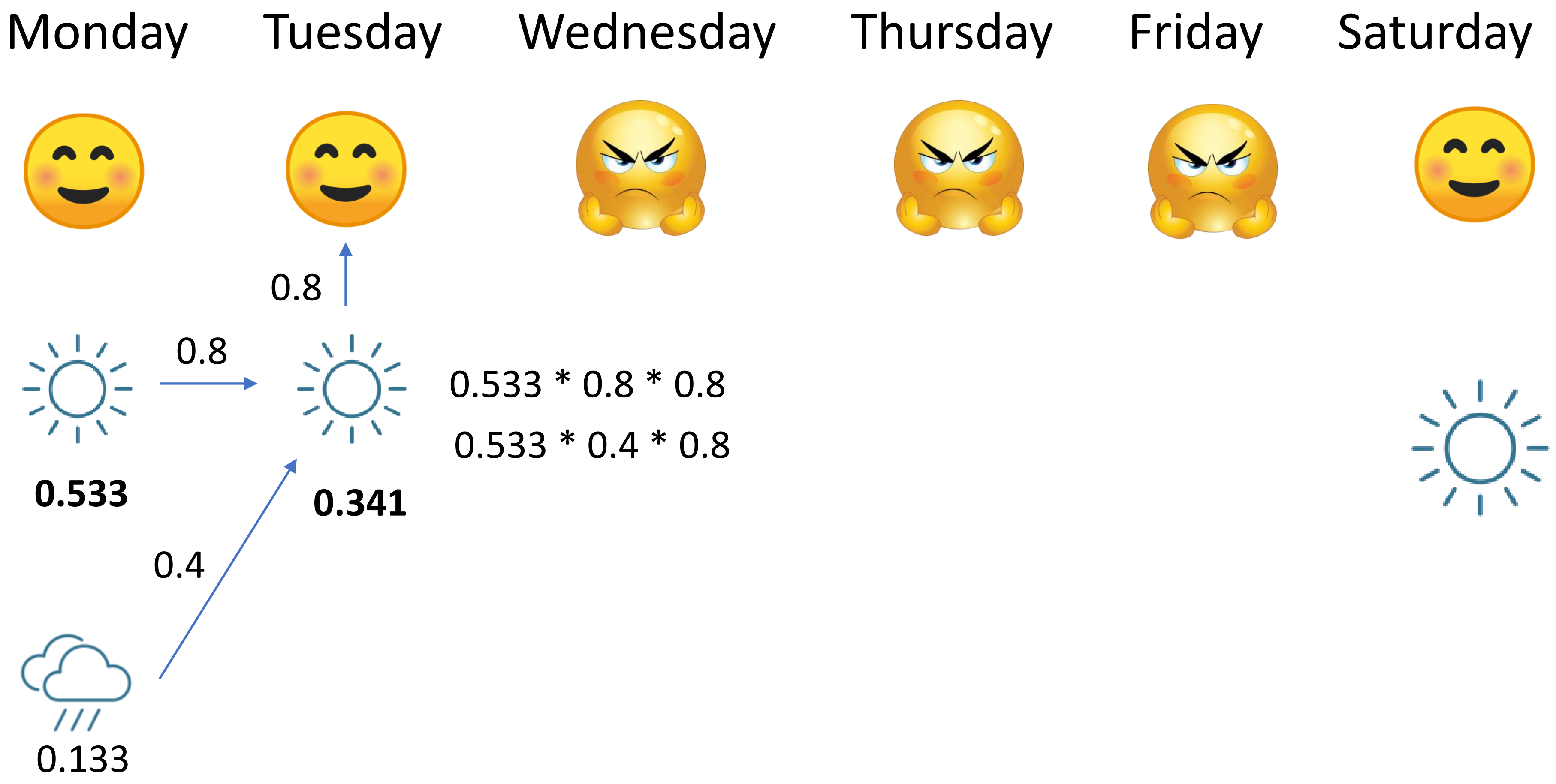


0.533 * 0.8 * 0.8

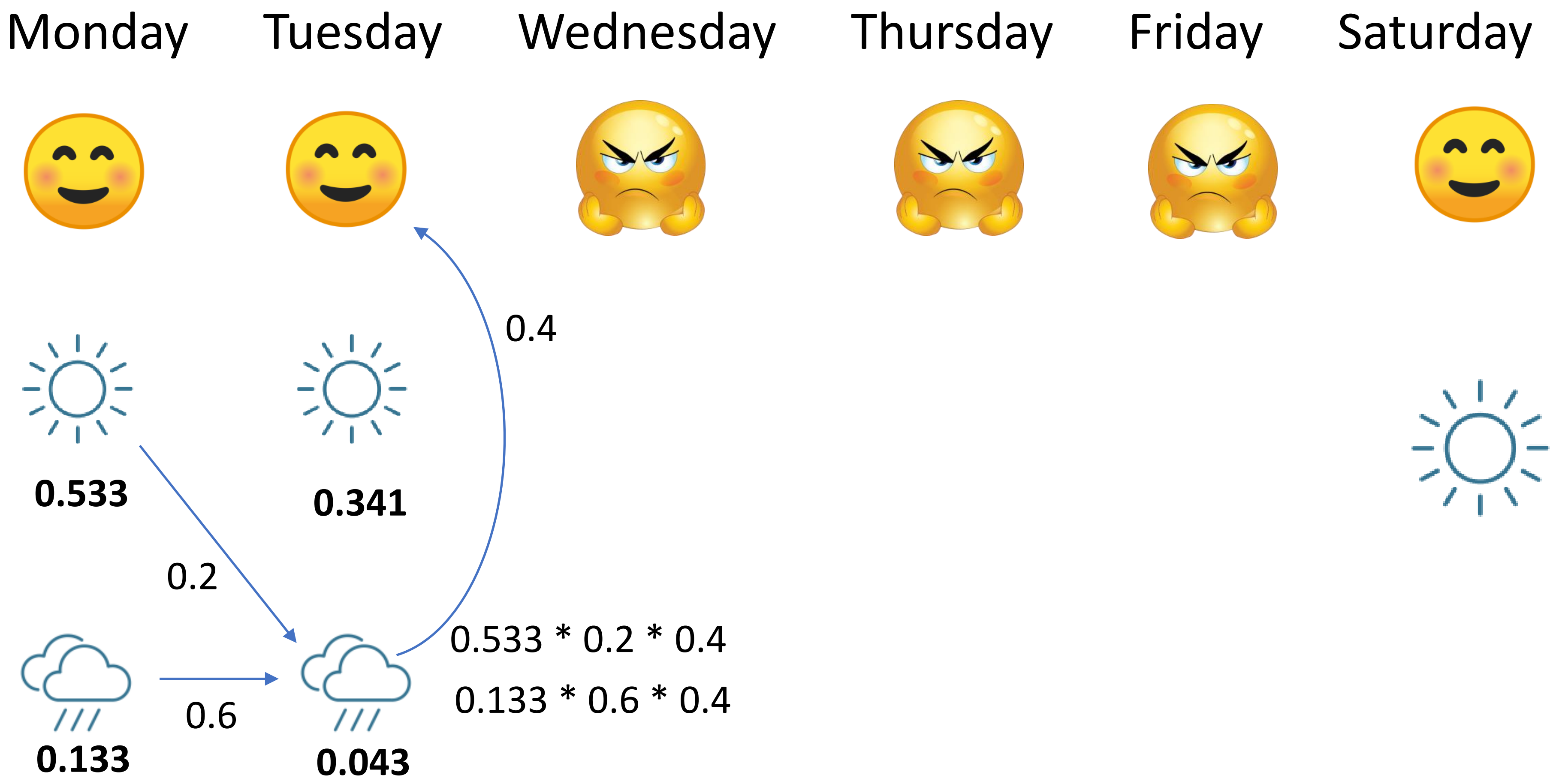
0.533 * 0.4 * 0.8

In Viterbi: We'll retain the maximum one.

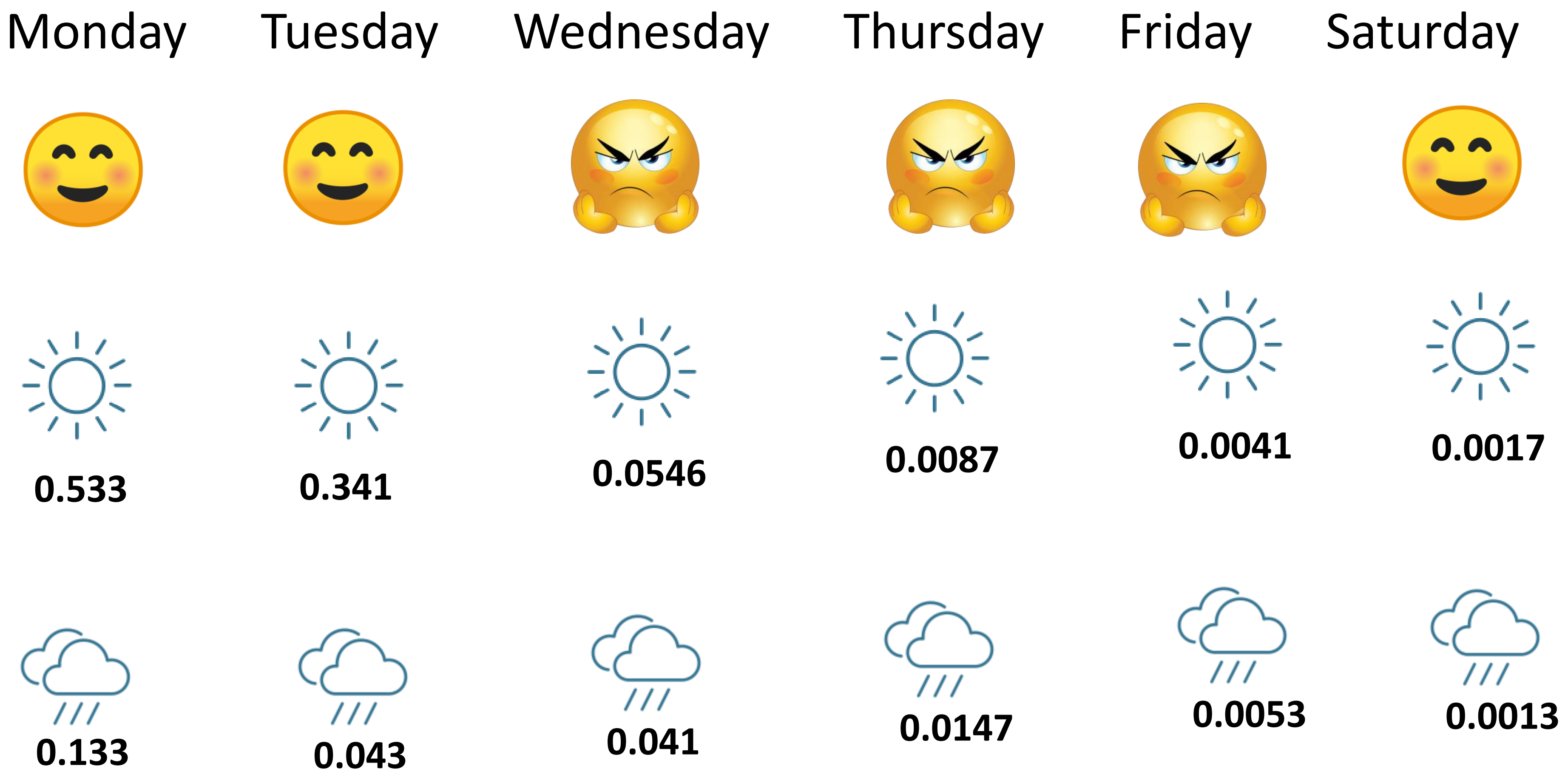
Viterbi Algorithm



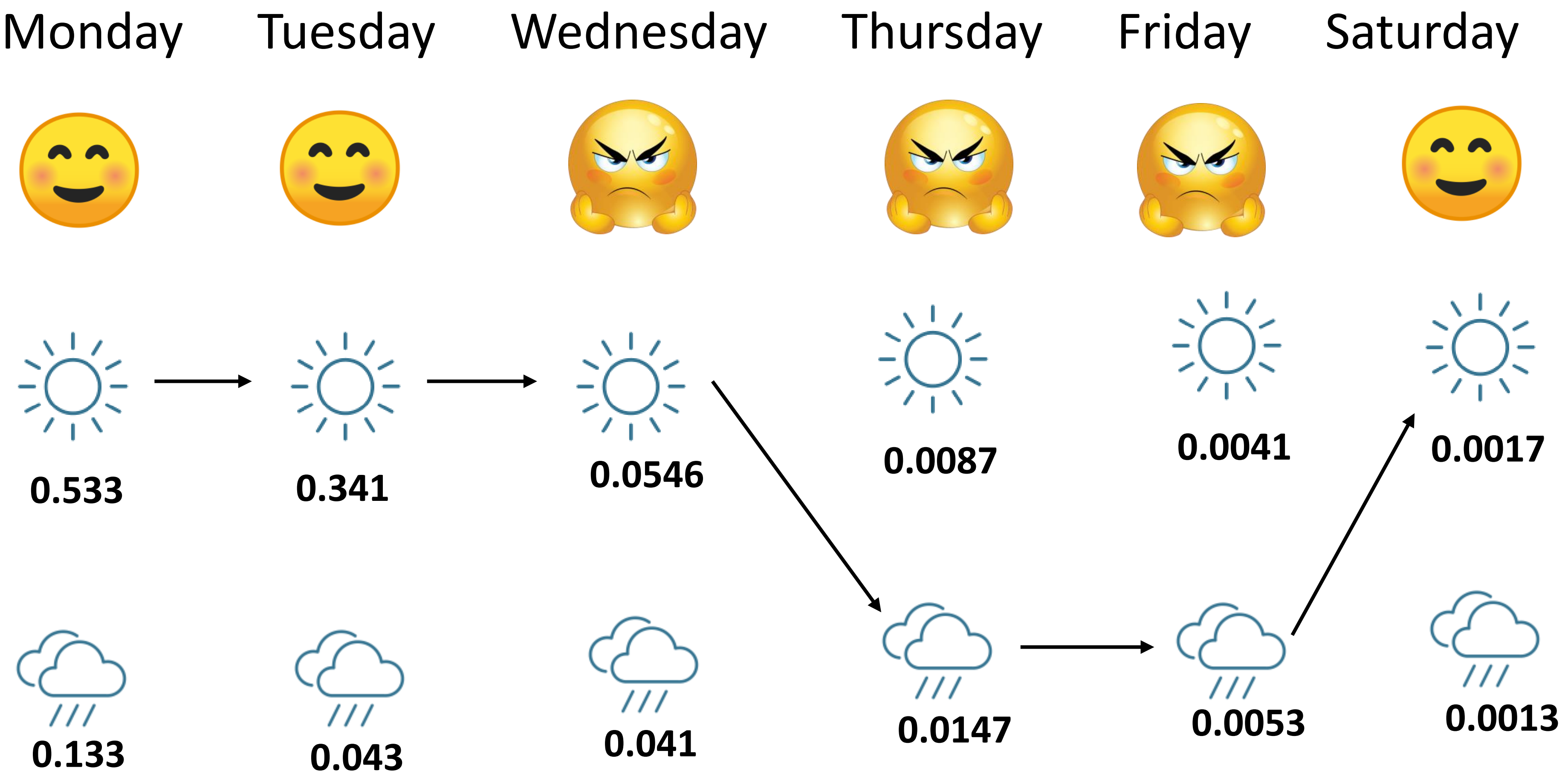
Viterbi Algorithm



Viterbi Algorithm



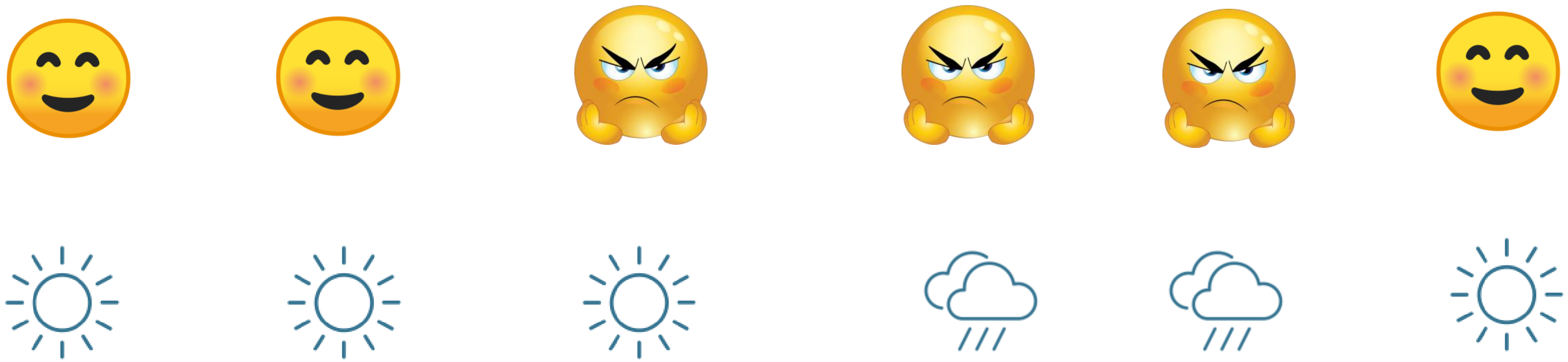
Viterbi Algorithm



Viterbi Algorithm

- **Our Answer is:**

Monday Tuesday Wednesday Thursday Friday Saturday



Hidden Markov Models (Revisited!!)

- **Questions**

1. How did we find these probabilities?
2. What’s the probability that a random day is Sunny or Rainy?
3. If Bob is Happy today, what’s the probability that it’s Sunny or Rainy?
4. If for three days Bob is Happy, Grumpy, Happy, what was the weather?

HMM Terminology

- Consider a system that can be described at any time as being in one of a set of ***N*** distinct states

$$\omega_1, \omega_2, \omega_3, \dots, \omega_N$$

- The state at any time *t* is denoted *ω(t),* where t = 1,2,3…

- The probability of the system being in state *ω(t)* is

$$P(\omega(t) \,/\, \omega(t\text{-}1), \ldots, \omega(1))$$

HMM Terminology

- We assume that the state *ω(t)* is conditional independent of the previous states given the predecessor state *ω(t-1)*

$$P(\omega(t) \mid \omega(t\text{-}1), \ldots, \omega(1)) = P(\omega(t) \mid \omega(t\text{-}1))$$

- A particular sequence of length *T* is denoted by

$$\omega^T = \{\omega(1), \omega(2), \ldots, \omega(T)\}$$

- For Example: $\omega^6 = \{\omega_1, \omega_4, \omega_2, \omega_2, \omega_1, \omega_4\}$

- **Note**: System can revisit a state at different steps and not every state need to be visited.

Transition probabilities

- Our model for the production of any sequence is described by *transition probabilities*

$$a_{ij} = P(w(t) = w_j | w(t-1) = w_i)$$

- *Where* $i, j \in \{1,2,\ldots,N\}$, $a_{ij} \geq 0$, and $\sum_{j-1}^{N} a_{ij} = 1, \forall i$

Transition probabilities

- State-transition probability matrix:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdot & \cdot & \cdot & a_{1N} \\ a_{21} & a_{23} & \cdot & \cdot & \cdot & a_{2N} \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ \cdot & \cdot & & & \cdot & \cdot \\ a_{N1} & a_{N2} & \cdot & \cdot & \cdot & a_{NN} \end{bmatrix}$$

$$a_{ij} = P(w(t) = w_j | w(t-1) = w_i)$$

$$a_{ij} \geq 0, \text{ and } \sum_{j-1}^{N} a_{ij} = 1, \forall i$$

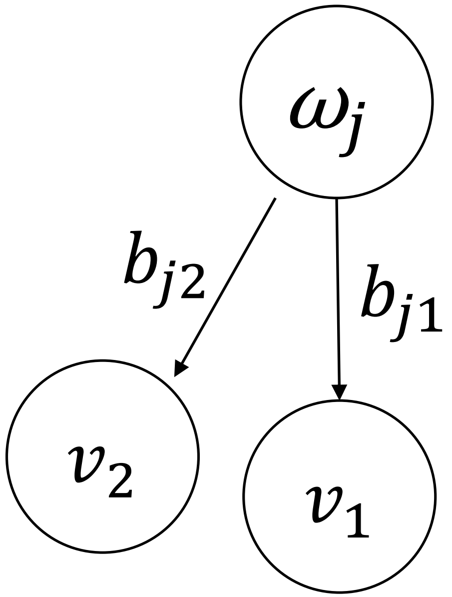
$\mathit{Transisiton\ probability\ from\ state}\ \omega_i$ to ω_j

$$\omega_i \xrightarrow{a_{ij}} \omega_j$$

Emission Probabilities

- We denote the observation at time *t* as *v(t)* and the probability of producing that observation in state $\omega(t)$ as

$$P(v(t)|\omega(t))$$



- When the observations are discrete, the distributions

$$b_{jk} = P(v(t) = v_k|\omega(t) = \omega_j)$$

are probability mass function where $j \in \{1,2,\dots,N\}, k \in \{1,2,\dots,M\}, b_{jk} \geq 0, and \ \sum_{k=1}^{M} b_{jk} = 1, \forall j.$

Emission Probabilities

$$B = \begin{bmatrix} b_1(v_1) & b_1(v_2) & \cdot & \cdot & \cdot & b_1(v_M) \\ b_2(v_1) & b_2(v_2) & \cdot & \cdot & \cdot & b_2(v_M) \\ \cdot & \cdot & & & & \cdot \\ \cdot & \cdot & \cdot & & & \cdot \\ \cdot & \cdot & & \cdot & & \cdot \\ b_N(v_1) & b_N(v_2) & \cdot & \cdot & \cdot & b_N(v_M) \end{bmatrix}$$

$$b_{jk} \geq 0, and$$

$$\sum_{k=1}^{M} b_{jk} = 1, \forall j.$$

Emission Probabilities

- When the observations are continuous, the distributions are typically specified using a parametric model family where the most common family is the **Gaussian mixture**

- We will restrict ourselves to discrete observations where a particular sequence of visible states of length T is denoted by

$$\mathbf{V}^T = \{v(1), v(2), ..., v(T)\}$$

Prior Probabilities

- **Initial (prior) probabilities:** these are the probabilities of starting the observation sequence in state ω_i

$$\pi = \begin{bmatrix} \pi_1 \\ \pi_2 \\ . \\ . \\ . \\ \pi_N \end{bmatrix}$$

$$\pi_i = P(\omega_i = s_i), \qquad 1 \leq i \leq N$$

$$\sum_{i=1}^{N} \pi_i = 1$$

Hidden Markov Models Parameters

- A HMM is governed by the following parameters:

$$\boldsymbol{\Theta} = \{A, B, \pi\}$$

  - State-transition probability matrix A
  - Emission/Observation/State Conditional Output probabilities B
  - Initial (prior) state probabilities π

- Determine the fixed number of states (N)

Three Central Issues in HMMs

- **The Evaluation problem.**
  - Give the model, compute the probability that a particular output sequence was produced by that model (solved by forward algorithm)

- **The Decoding problem.**
  - Given the model, find the most likely sequence of hidden states which could have generated a given output sequence (solved by Viterbi algorithm)

- **The Learning problem.**
  - Suppose we are given the coarse structure of a model (the number of states and the number of visible states) but not the probabilities a_{ij} and b_{jk}. Given a set of training observations of visible symbols, determine these parameters. (solved by Baum-Weltch algorithm)

The Evaluation Problem

- The probability that the model produces a sequence V^T of visible states is:

$$P(\mathbf{V}^T) = \sum_{r=1}^{r_{max}} P(\mathbf{V}^T|\boldsymbol{\omega}_r^T)P(\boldsymbol{\omega}_r^T)$$

- Where each *r* indexes a particular sequence of *T* hidden states.

$$\boldsymbol{\omega}_r^T = \{\omega(1), \omega(2), ..., \omega(T)\}$$

- In the general case of c hidden states, there will be $r_{max} = c^T$ possible terms in the sum of Eq. corresponding to all possible sequences of length T .

The Evaluation Problem

- The probability of a particular visible sequence is merely the product of the corresponding (hidden) transition probabilities a_{ij} and the (visible) output probabilities b_{jk} of each step.
- Because we are dealing here with a first-order Markov process

$$P(\boldsymbol{\omega}_r^T) = \prod_{t=1}^{T} P(\omega(t)|\omega(t-1))$$

$$P(\mathbf{V}^T|\boldsymbol{\omega}_r^T) = \prod_{t=1}^{T} P(v(t)|\omega(t)),$$

The Evaluation Problem

$$P(\mathbf{V}^T) = \sum_{r=1}^{r_{max}} P(\mathbf{V}^T|\boldsymbol{\omega}_r^T)P(\boldsymbol{\omega}_r^T)$$

$$P(\boldsymbol{\omega}_r^T) = \prod_{t=1}^{T} P(\omega(t)|\omega(t-1))$$

$$P(\mathbf{V}^T|\boldsymbol{\omega}_r^T) = \prod_{t=1}^{T} P(v(t)|\omega(t)),$$

$$P(\mathbf{V}^T) = \sum_{r=1}^{r_{max}} \prod_{t=1}^{T} P(v(t)|\omega(t))P(\omega(t)|\omega(t-1)).$$

- **NOTE:** All these are captured in our parameters a_{ij} and b_{kj} and can be evaluated directly
- Alas, this is an **$O(c^T\, T)$** calculation, which is quite prohibitive in practice.
- However, computationally simpler algorithm called the forward algorithm computes $P(V^T)$ recursively

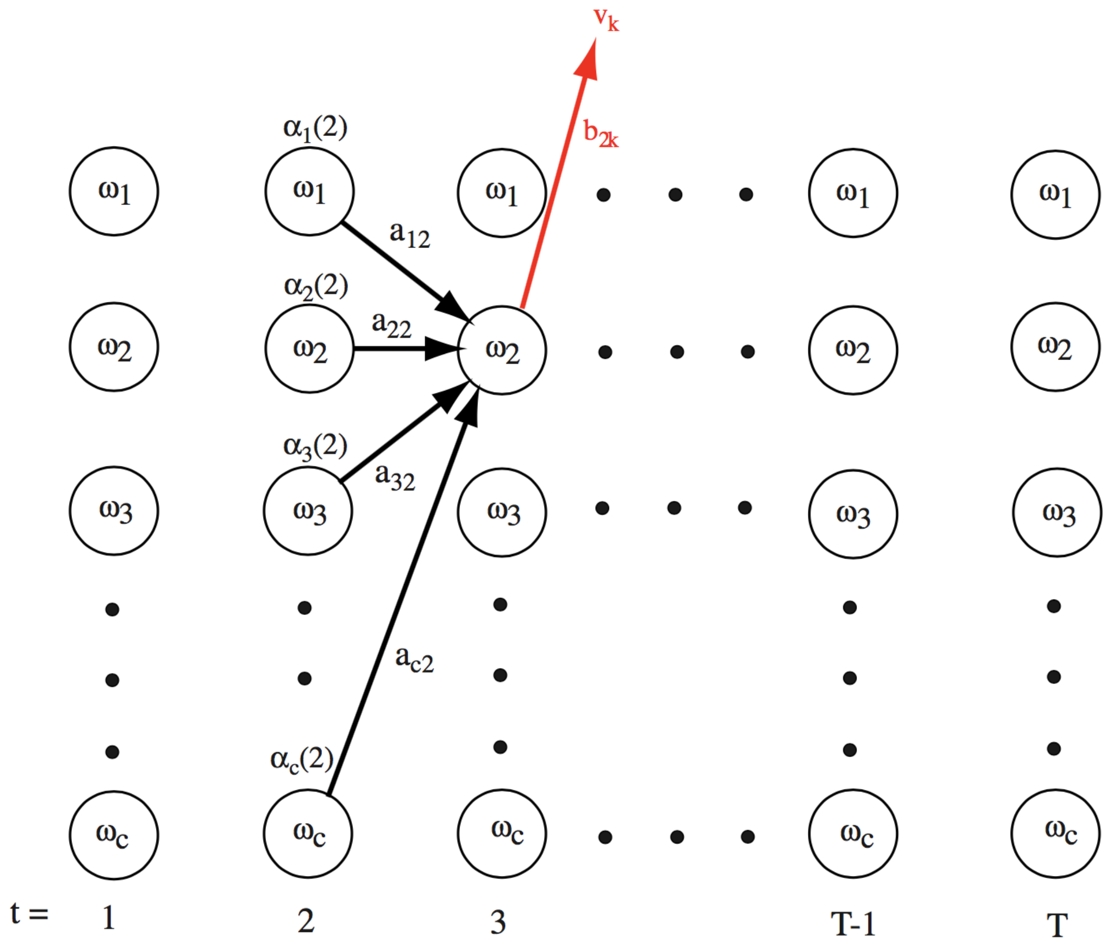
The Evaluation Problem

- Since each term *$P(v(t)|\omega(t))P(\omega(t)|\omega(t-1))$* involves only v(t), ω(t) and ω(t − 1).
- We do this by defining

$$\alpha_i(t) = \begin{cases} 0 & t = 0 \text{ and } i \neq \text{initial state} \\ 1 & t = 0 \text{ and } i = \text{initial state} \\ \sum_j \alpha(t-1) a_{ij} b_{jk} v(t) & \text{otherwise,} \end{cases}$$

- Where the notation *$b_{jk}v(t)$* means the transition probability b_{jk} selected by the visible state emitted at time t.
- Thus the only non-zero contribution to the sum is for the index k which matches the visible state v(t).

The Evaluation Problem – Forward Algorithm



The Evaluation Problem – Forward Algorithm

The computation of probabilities by the Forward algorithm can be visualized by means of a trellis — a sort of "unfolding" of the HMM through time. Suppose we seek the probability that the HMM was in state ω_2 at $t = 3$ and generated the observed visible up through that step (including the observed visible symbol v_k). The probability the HMM was in state $\omega_j(t = 2)$ and generated the observed sequence through $t = 2$ is $\alpha_j(2)$ for $j = 1, 2, ..., c$. To find $\alpha_2(3)$ we must sum these and multiply the probability that state ω_2 emitted the observed symbol v_k. Formally, for this particular illustration we have $\alpha_2(3) = b_{2k} \sum_{j=1}^{c} \alpha_j(2) a_{j2}$.

The Evaluation Problem – Forward Algorithm

initialize $\omega(1), t = 0, a_{ij}, b_{jk}$, visible sequence $\mathbf{V}^T, \alpha(0) = 1$

  for $t \leftarrow t + 1$

    $\alpha_j(t) \leftarrow \sum_{i=1}^{c} \alpha_i(t-1) a_{ij} b_{jk}$

  until $t = T$

return $P(\mathbf{V}^T) \leftarrow \alpha_0(T)$

end

- The Forward algorithm has, thus, a computational complexity of $O(c^2T)$ — far more efficient than the complexity associated with exhaustive enumeration of paths

The Backward Algorithm

- We shall have cause to use the Backward algorithm, which is the time-reversed version of the Forward algorithm.

initialize $\omega(T), t = T, a_{ij}, b_{jk}$, visible sequence V^T

    for $t \leftarrow t - 1;$

$$\beta_j(t) \leftarrow \sum_{i=1}^{c} \beta_i(t+1) a_{ij} b_{jk} v(t+1)$$

    until $t = 1$

return $P(V^T) \leftarrow \beta_i(0)$ for the known initial state

end

The Evaluation Problem

- If we denote our model — the a's and b's — by θ, we have by Bayes' formula that the probability of the model given the observed sequence is:

$$P(\boldsymbol{\theta}|\mathbf{V}^T) = \frac{P(\mathbf{V}^T|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\mathbf{V}^T)}$$

where $P(\Theta)$ is the prior for a particular class, and $P(\mathcal{V}^T|\Theta)$ is computed using the forward algorithm with the HMM for that class.

Then, we can select the class with the highest posterior.

NOTE!!

- In HMM pattern recognition we would have a number of HMMs, one for each category and classify a test sequence according to the model with the highest probability.

- Thus in HMM speech recognition we could have a model for “cat” and another one for “dog” and for a test utterance determine which model has the highest probability.

Application Areas

- On-line handwriting recognition
- Speech recognition
- Gesture recognition
- Language modeling
- Motion video analysis and tracking
- Stock price prediction and many more….

Summary

- Markov Model
- Hidden Markov model (HMM)
- Examples
- Three central issues of HMM
- Application areas of HMM

Reference

- Video lecture of Luis: https://www.youtube.com/watch?v=kqSzLo9fenk
- Chapter 3 of R.O. Duda, P.E. Hart, and D.G. Stork, Pattern Classification, New York: John Wiley, 2001
- Other materials found over the internet for HMM.

Thank You ☺