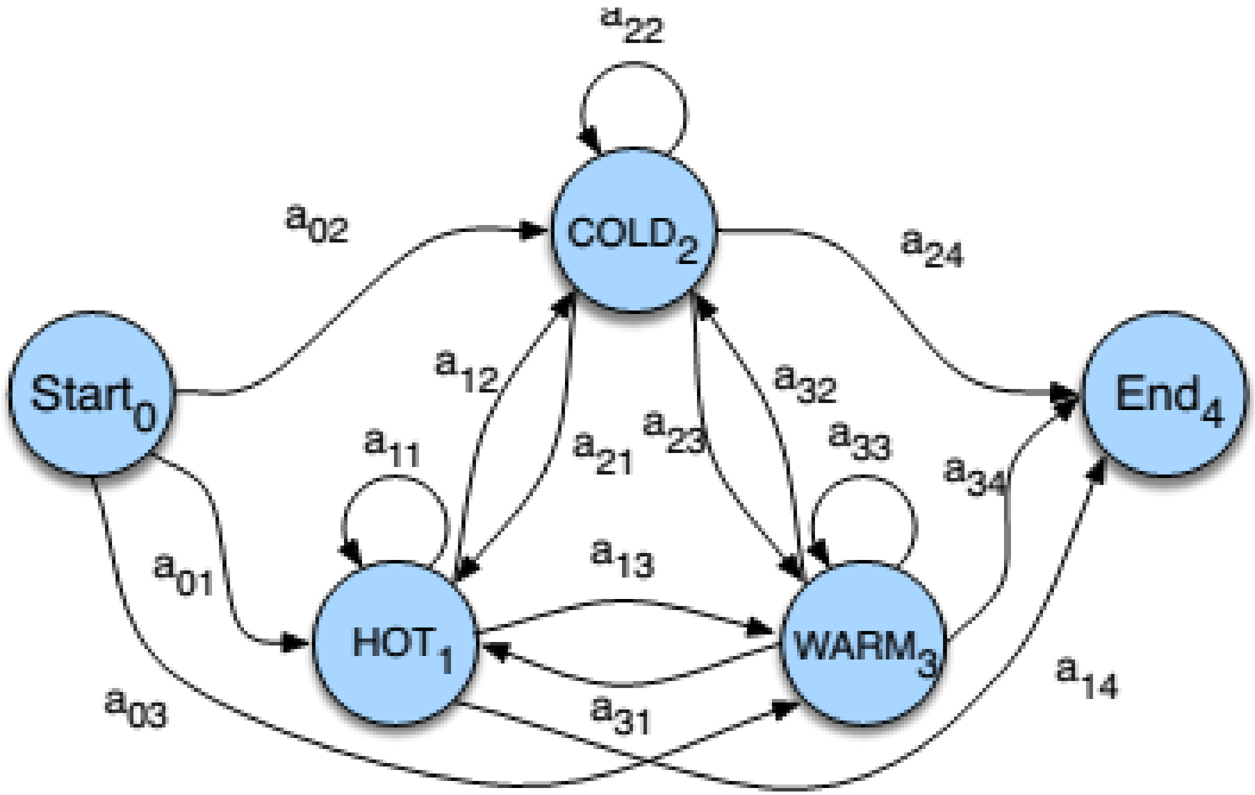
HMM

HMM Model

- The HMM is a **sequence model**.
- A sequence model or sequence classifier is a model whose job is to assign a label or class to each unit in a sequence, thus mapping a sequence of **observations** to a sequence of **labels**.
- An HMM is a **probabilistic sequence model**: given a sequence of units (words, letters, morphemes, sentences, whatever), they compute a probability distribution over possible sequences of labels and choose the best label sequence.
- **Sequence labeling task : PoS Tagging, NER, Speech Recognition**

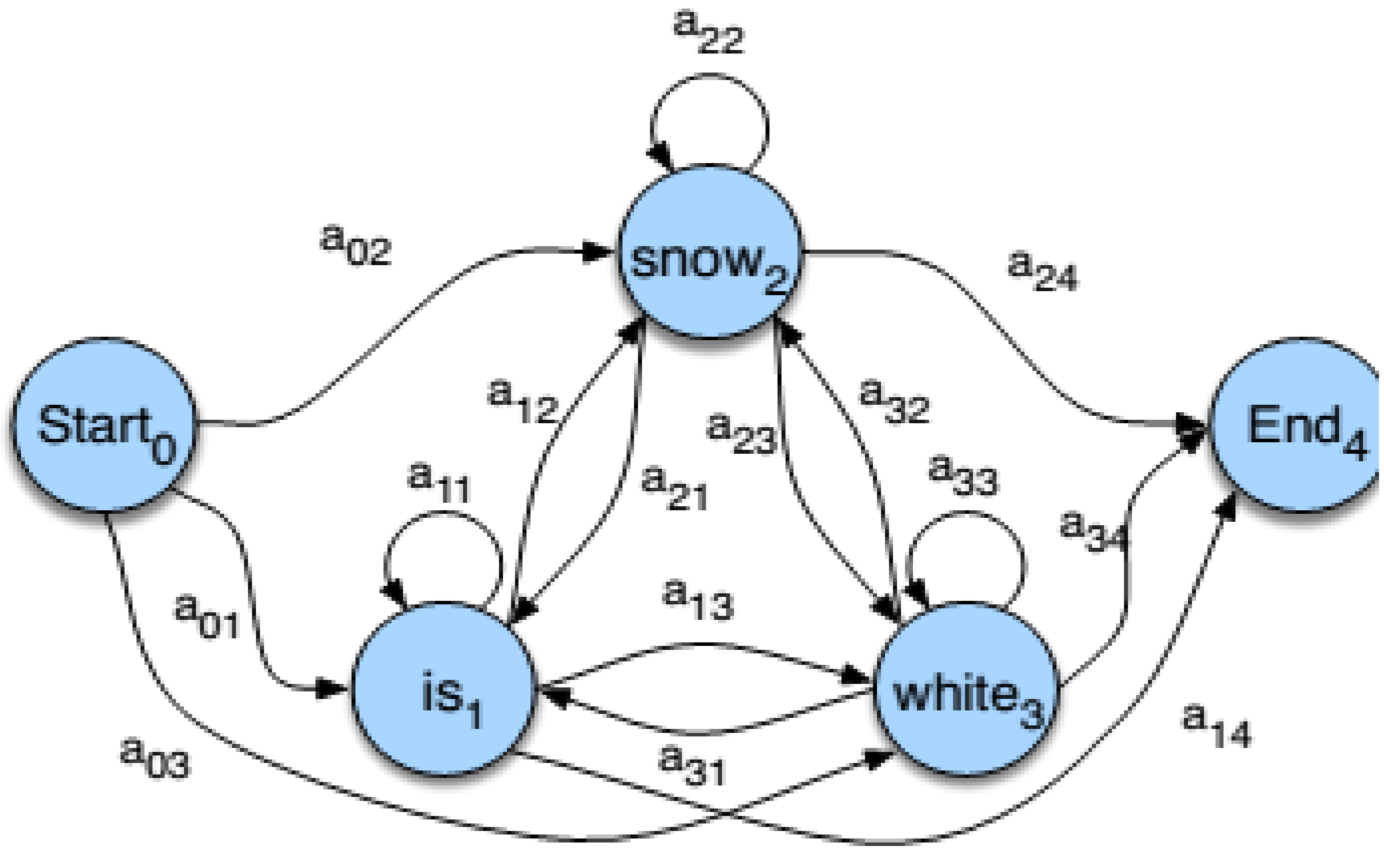
Markov Chains



Source : jurafsky

- Markov chain is an extension of Finite Automata, especially weighted finite automaton.
- Markov chain is a special case which the weights are probability so that it sums to 1, and not ambiguous

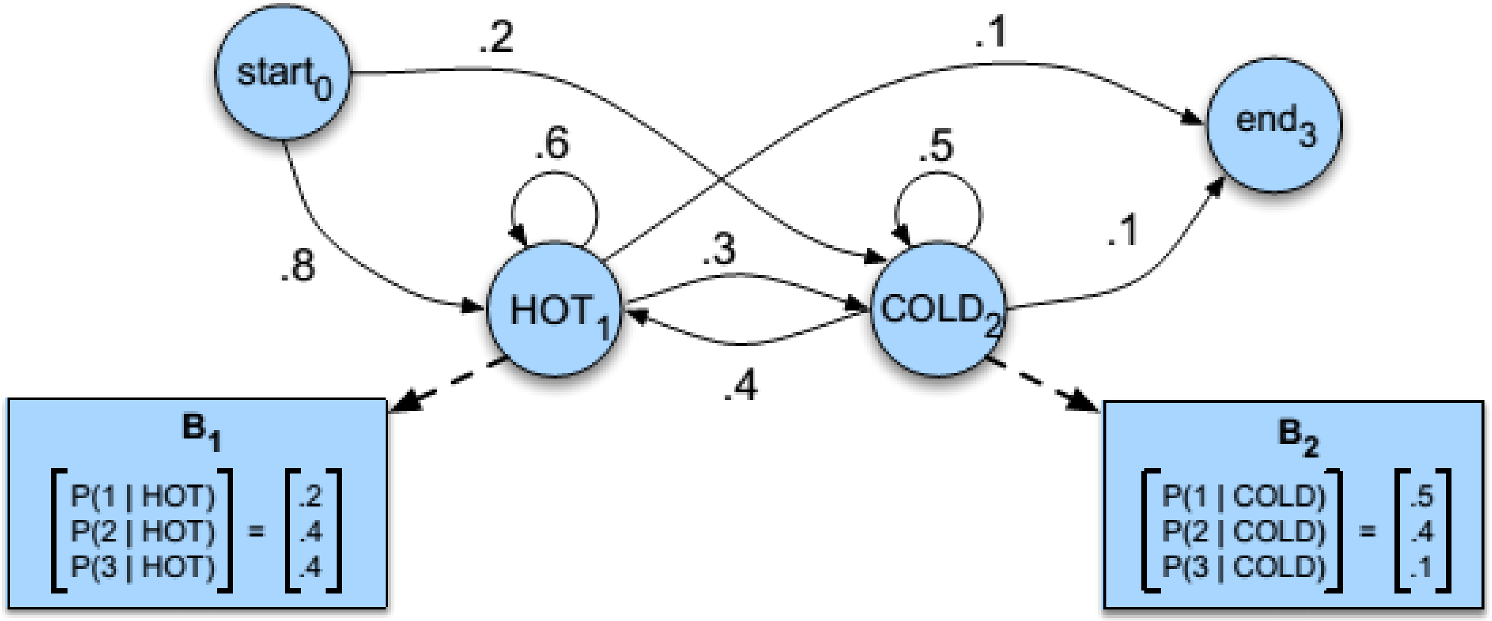
Markov Chain



Source : jurafsky

- This markov chain represent bigram language model. Can you see that?

The Hidden Markov Model



Source : jurafsky

- Given how many **Ice Cream[observation]** Jason Eisner eats everyday in summer, figure out the **weather status[hidden]** each day

HMM Components

$Q = q_1 q_2 \ldots q_N$	a set of N **states**
$A = a_{11} a_{12} \ldots a_{n1} \ldots a_{nn}$	a **transition probability matrix** A, each a_{ij} representing the probability of moving from state i to state j, s.t. $\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i$
$O = o_1 o_2 \ldots o_T$	a sequence of T **observations**, each one drawn from a vocabulary $V = v_1, v_2, ..., v_V$
$B = b_i(o_t)$	a sequence of **observation likelihoods**, also called **emission probabilities**, each expressing the probability of an observation o_t being generated from a state i
q_0, q_F	a special **start state** and **end (final) state** that are not associated with observations, together with transition probabilities $a_{01} a_{02} \ldots a_{0n}$ out of the start state and $a_{1F} a_{2F} \ldots a_{nF}$ into the end state

$\pi = \pi_1, \pi_2, ..., \pi_N$	an **initial probability distribution** over states. π_i is the probability that the Markov chain will start in state i. Some states j may have $\pi_j = 0$, meaning that they cannot be initial states. Also, $\sum_{i=1}^{n} \pi_i = 1$
$QA = \{q_x, q_y \ldots\}$	a set $QA \subset Q$ of legal **accepting states**

Some Probabilities

- We want to find : $q_1^n = \underset{q_1^n}{\operatorname{argmax}} P(q_1^n | o_1^n)$
- Using Bayes' rule : $q_1^n = \underset{q_1^n}{\operatorname{argmax}} \frac{P(o_1^n | q_1^n) P(q_1^n)}{P(o_1^n)}$
- Drop denominator (why?) : $q_1^n = \underset{q_1^n}{\operatorname{argmax}} P(o_1^n | q_1^n) P(q_1^n)$

Assumptions

- $q_1^n = \underset{q_1^n}{\operatorname{argmax}} P(o_1^n|q_1^n)P(q_1^n)$

There are 2 assumptions in HMM :

1. 1st order Markov Assumption : probability of a particular state depends only on the previous state
$$\boldsymbol{P(q_i|q_1, q_2, \dots, q_{i-1}) = P(q_i|q_{i-1})}$$
2. The probability of an output observation o_i depends only on the state that produce the observation which is q_i
$$\boldsymbol{P(o_i|q_1, \dots, q_i, \dots, q_N, o_1, \dots, o_i, \dots, o_N) = P(o_i|q_i)}$$

Problems related to HMM

1. Likelihood : Given an HMM $\lambda = (A,B)$ and an observation sequence O, determine the likelihood $P(O|\lambda)$
2. Decoding : Given an observation sequence O and an HMM $\lambda = (A,B)$, discover the best hidden state sequence Q.
3. Learning : Given an observation sequence O and the set of states in the HMM, learn the HMM parameters A and B

HMM for PoS Tagging

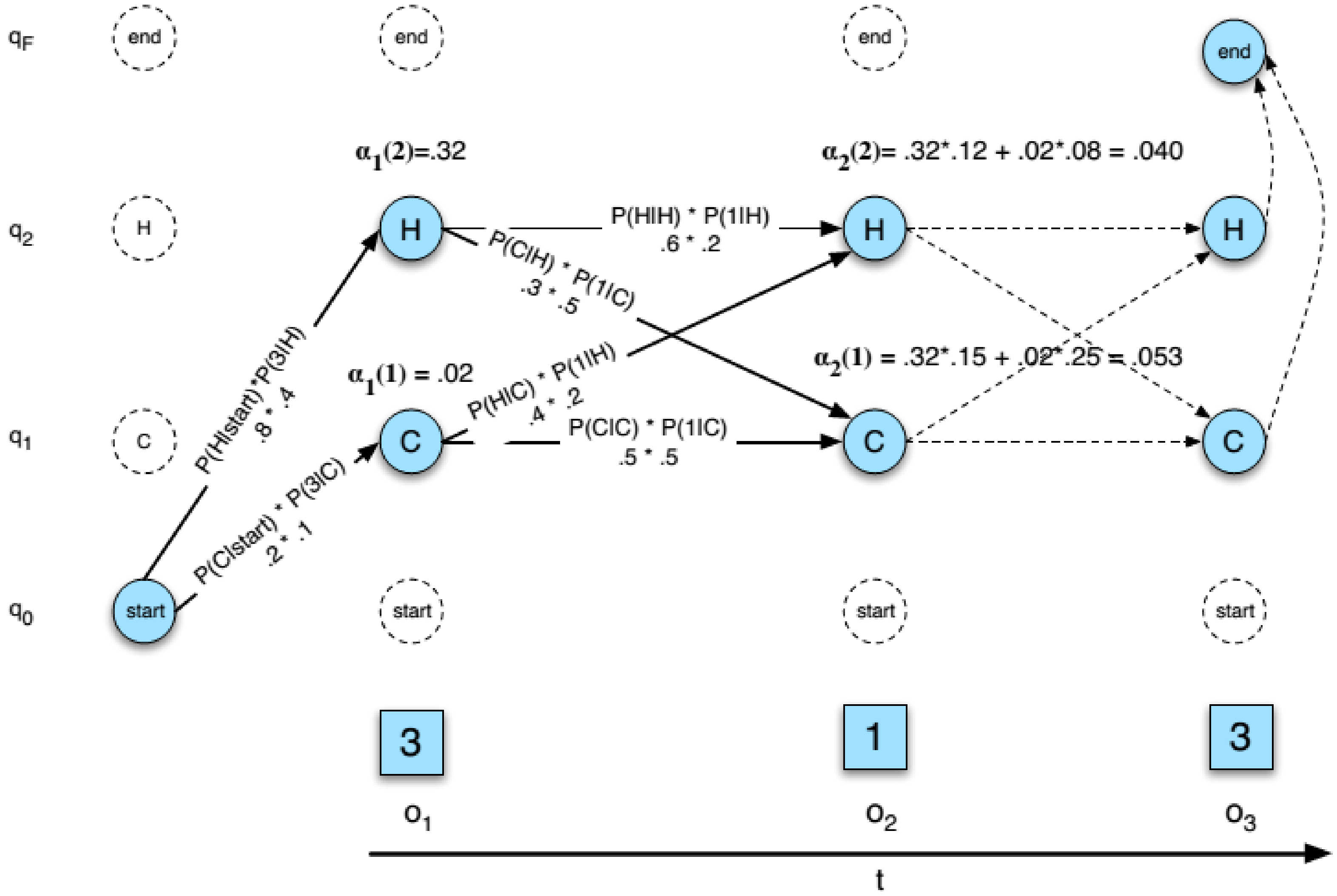
- From : Janet will back the bill → **OBSERVED**
- To : NNP MD VB DT NN → **HIDDEN**

- **Which problem is this ?**

Likelihood

- Ex : what is the likelihood of eating ice cream with a sequence of 3 1 3 ?
- P(3 1 3)= P(3 1 3, cold cold cold)+P(3 1 3, cold cold hot)+…. P(3 1 3, hot hot hot)

Likelihood

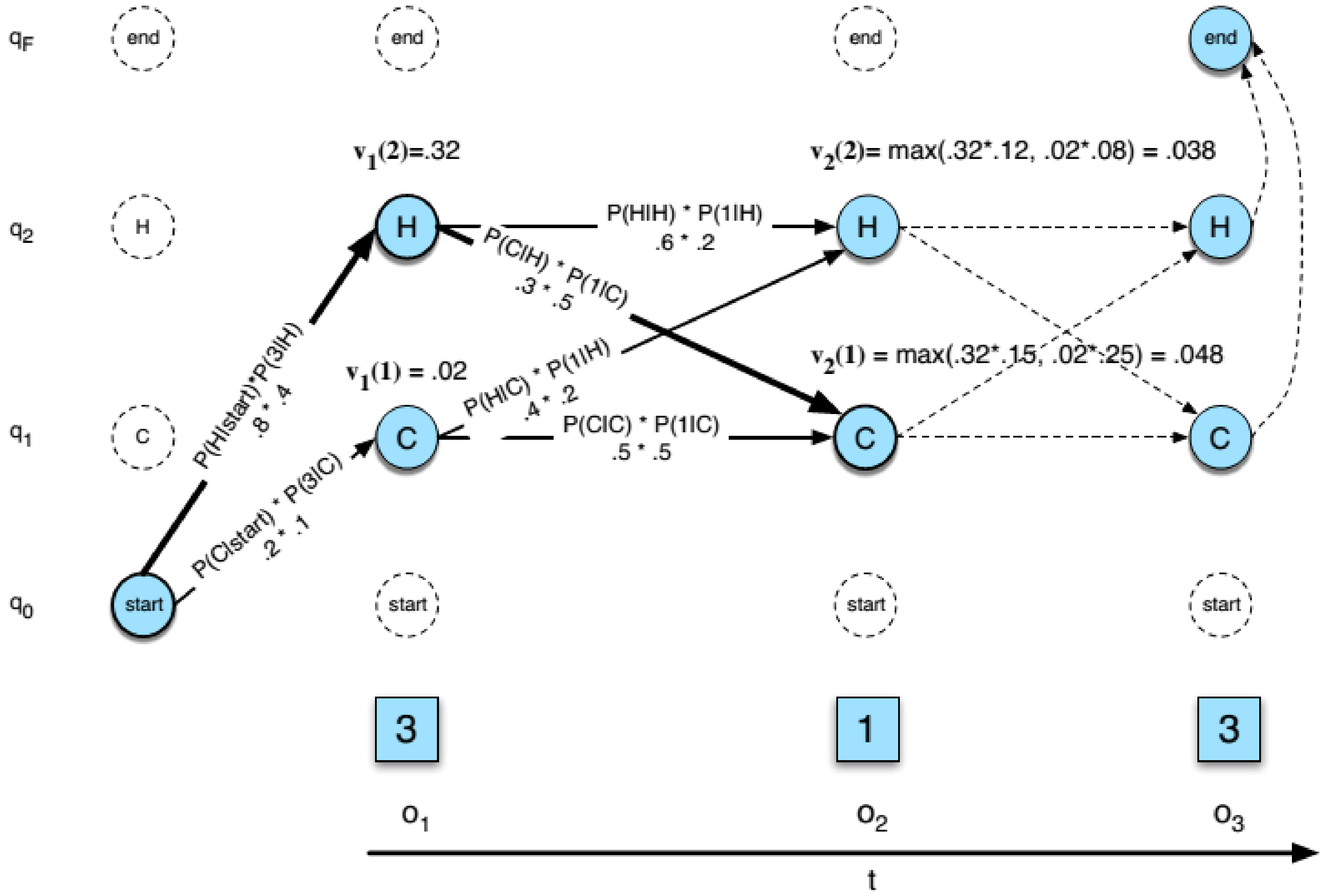


qF
q2
q1
q0
end
end
end
end
H
C
start
α1(2)=.32
α1(1) = .02
α2(2)= .32*.12 + .02*.08 = .040
α2(1) = .32*.15 + .02*.25 = .053
P(H|start)*P(3|H)
.8 * .4
P(C|start) * P(3|C)
.2 * .1
P(H|H) * P(1|H)
.6 * .2
P(C|H) * P(1|C)
.3 * .5
P(H|C) * P(1|H)
.4 * .2
P(C|C) * P(1|C)
.5 * .5
H
C
H
C
H
C
start
start
start
3
1
3
o1
o2
o3
t

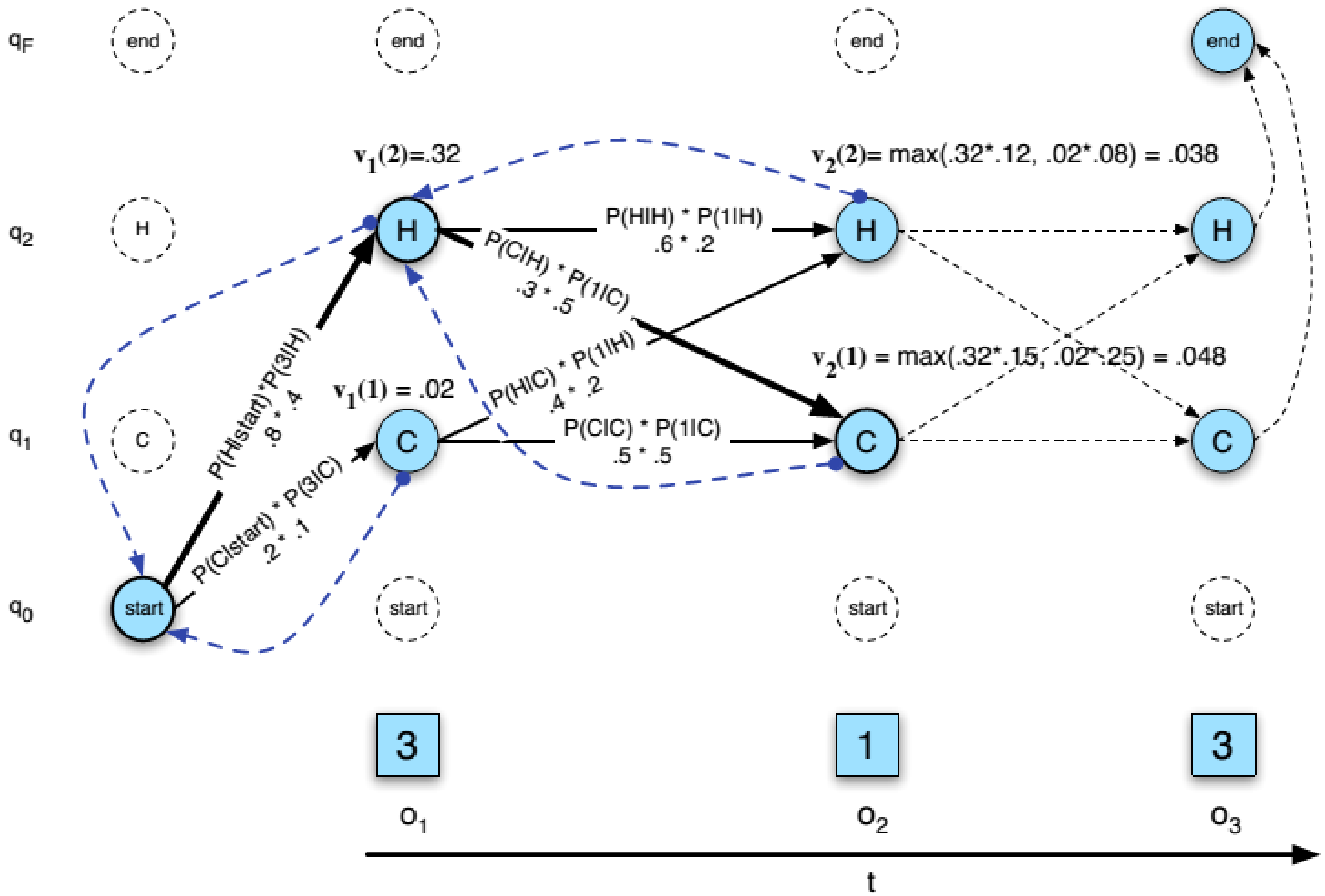
Decoding

- Finding **the best** hidden states given observations
- Ex : What is the best sequence of weather given ice cream observation of 3 1 3 ?
- Approach :
  - Brute force : 3 1 3, Find likelihood (problem 1) of all possible states combination with length of 3, ex : C C C, C C H, …, H H H, then choose sequence that give the maximum likelihood
  - **Viterbi Algorithm**
    - A kind of dynamic programming

Decoding : Viterbi



Viterbi Backtrace



PoS Tagging

- earnings growth took a **back/JJ** seat
- a small building in the **back/NN**
- a clear majority of senators **back/VBP** the bill
- Dave began to **back/VB** toward the door
- enable the country to buy **back/RP** about debt
- I was twenty-one **back/RB** then

- How to tag a word correctly ?
    1. Look at the word
    2. Look at the previous tag ?

- Janet will back the bill
- Janet/NNP will/MD back/VB the/DT bill/NN