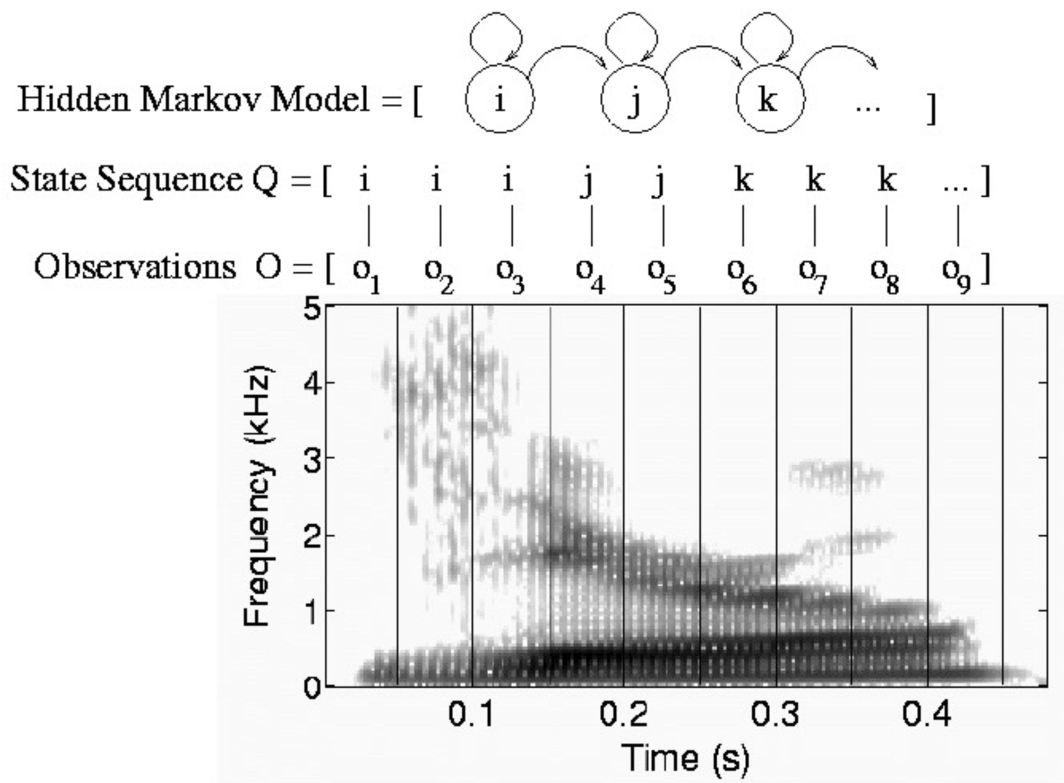
CS440/ECE448 Lecture 22: Hidden Markov Models

Mark Hasegawa-Johnson, 3/2022



Outline

- HMM: Probabilistic reasoning over time
- Two views of an HMM: as a Bayes Net, as an FSM
- Inference: Belief propagation in an HMM
- Parameter learning: Maximum likelihood
- Parameter learning: EM

Probabilistic reasoning over time

- So far, we've mostly dealt with *episodic* environments
  - Exceptions: games with multiple moves, planning
- In particular, the Bayesian networks we've seen so far describe static situations
  - Each random variable gets a single fixed value in a single problem instance
- Now we consider the problem of describing probabilistic environments that evolve over time
  - Examples: robot localization, human activity detection, tracking, speech recognition, machine translation,

Probabilistic reasoning over time

- At each time slice *t*, the state of the world is described by an unobservable **state variable** Y_t and an observable **observation variable** X_t
- **State Transitions:** in general, the value of Y_t depends on the whole past history:

$$P(Y_t \mid Y_0, \ldots, Y_{t-1}) = P(Y_t \mid \mathbf{Y}_{0:t-1})$$

- **Observation model:** in general, the value of X_t depends on all current and past states and observations:

$$P(X_t \mid Y_0, \ldots, Y_t, X_1, \ldots, X_{t-1}) = P(X_t \mid \mathbf{Y}_{0:t}, \mathbf{X}_{1:t-1})$$

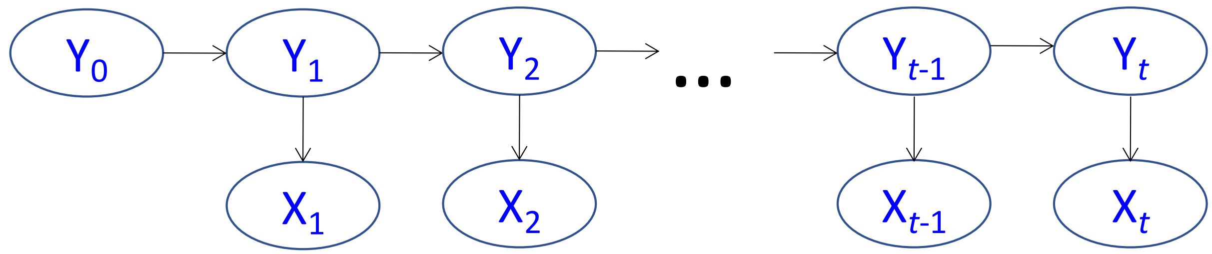
Hidden Markov Model

- A hidden Markov model assumes that both the state and the observation are Markov.
- **<u>State Transitions</u>:** the Markov assumption means that each state variable depends only on the preceding time step:

$$P(Y_t \mid Y_0, \ldots, Y_{t-1}) = P(Y_t \mid Y_{t-1})$$

- **<u>Observation model</u>:** the Markov assumption means that each state variable depends only on the current state:

$$P(X_t \mid Y_0, \ldots, Y_t, X_1, \ldots, X_{t-1}) = P(X_t \mid Y_t)$$



Example Scenario: UmbrellaWorld

Characters from the novel *Hammered* by Elizabeth Bear,
Scenario from chapter 15 of Russell & Norvig

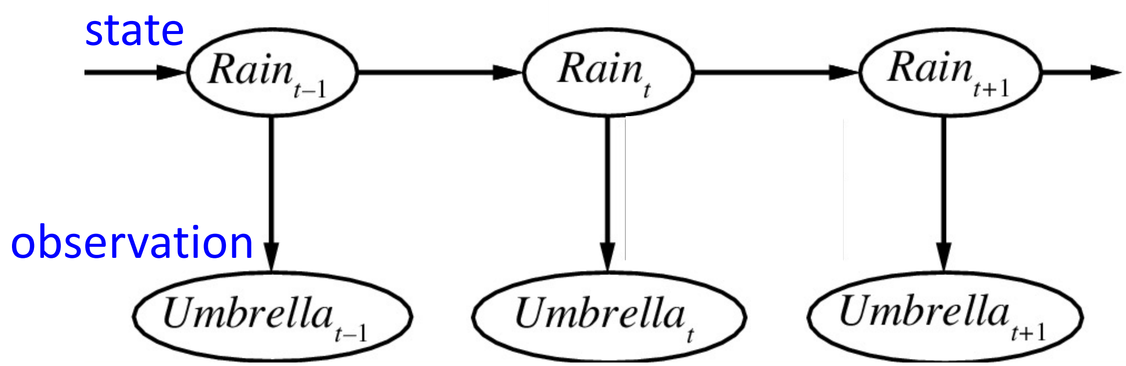
- Elspeth Dunsany is an AI researcher at the Canadian company Unitek.
- Richard Feynman is an AI, named after the famous physicist, whose personality he resembles.
- To keep him from escaping, Richard's workstation is not connected to the internet. He knows about rain but has never seen it.
- He has noticed, however, that Elspeth sometimes brings an umbrella to work. He correctly infers that she is more likely to carry an umbrella on days when it rains.

Example Scenario: UmbrellaWorld

Characters from the novel *Hammered* by Elizabeth Bear,
Scenario from chapter 15 of Russell & Norvig

Since he has read a lot about rain, Richard proposes a hidden Markov model:

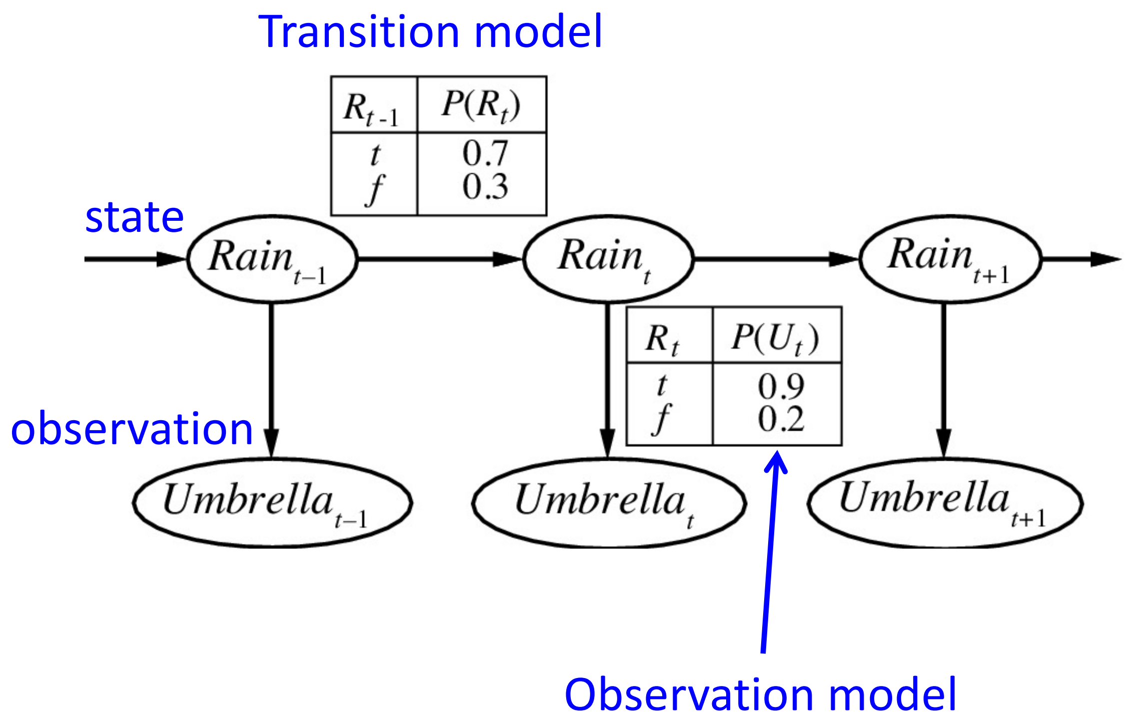
- Rain on day t-1 (R_{t-1}=T) makes rain on day t ($R_t = T$) more likely.
- Elspeth usually brings her umbrella ($U_t = T$) on days when it rains ($R_t = T$), but not always.



Example Scenario: UmbrellaWorld

Characters from the novel *Hammered* by Elizabeth Bear,
Scenario from chapter 15 of Russell & Norvig

- Richard learns that the weather changes on 3 out of 10 days, thus
  $$P(R_t = T|R_{t-1} = T) = 0.7$$
  $$P(R_t = T|R_{t-1} = F) = 0.3$$
- He also learns that Elspeth sometimes forgets her umbrella when it's raining, and that she sometimes brings an umbrella when it's not raining. Specifically,
  $$P(U_t = T|R_t = T) = 0.9$$
  $$P(U_t = T|R_t = F) = 0.2$$



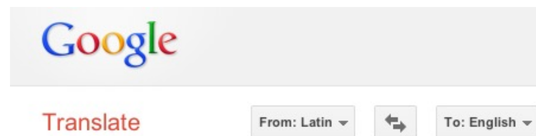
R_{t-1}	$P(R_t)$
t	0.7
f	0.3

R_t	$P(U_t)$
t	0.9
f	0.2

Applications of HMMs

- Speech recognition HMMs:
  - Observations are acoustic signals (continuous valued)
  - States are specific positions in specific words (so, tens of thousands)

- Machine translation HMMs:
  - Observations are words (tens of thousands)
  - States are cross-lingual alignments



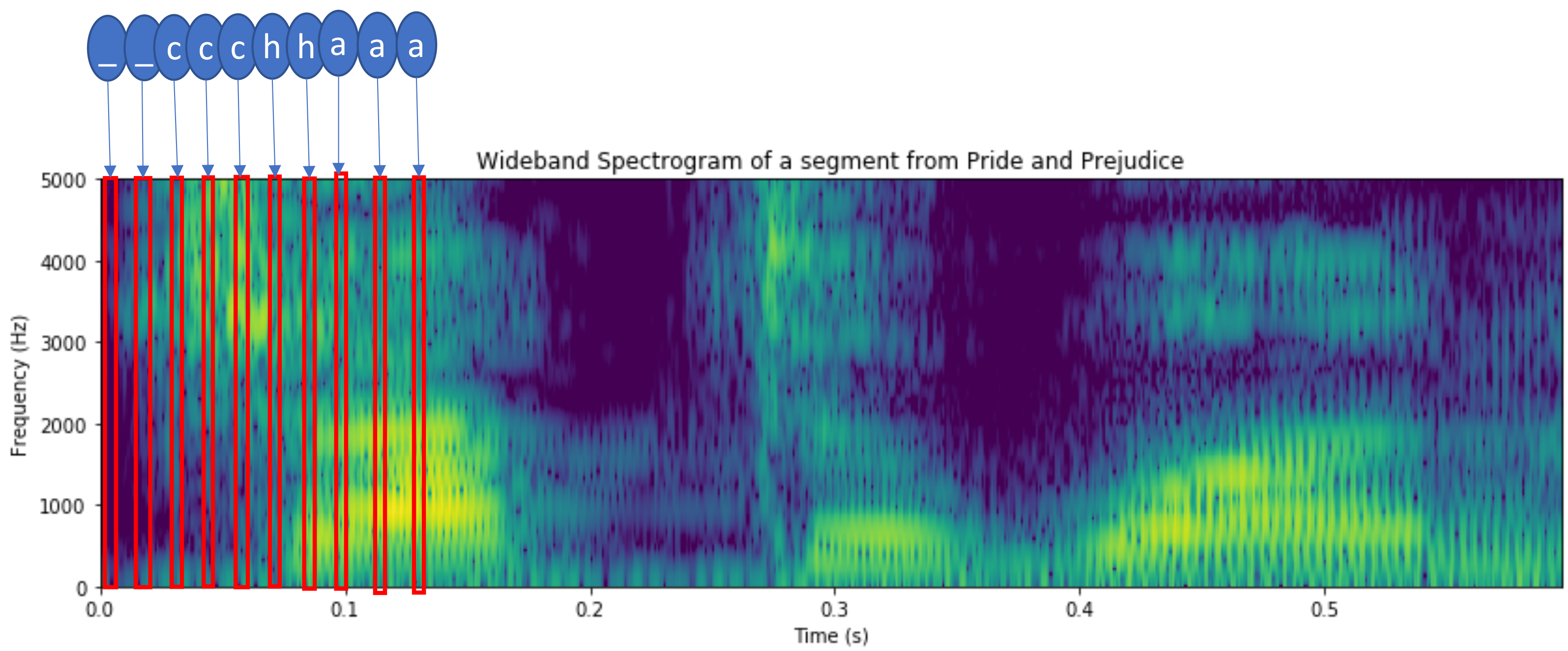
- Robot tracking:
  - Observations are range readings (continuous)
  - States are positions on a map

Source: Tamara Berg

Example: Speech Recognition

- Observations: X_t = spectrum of 25ms frame of the speech signal.
- State: Y_t = phoneme or letter being currently produced

Example utterance: “chapter one,” from a Librivox recording of *Pride and Prejudice.*



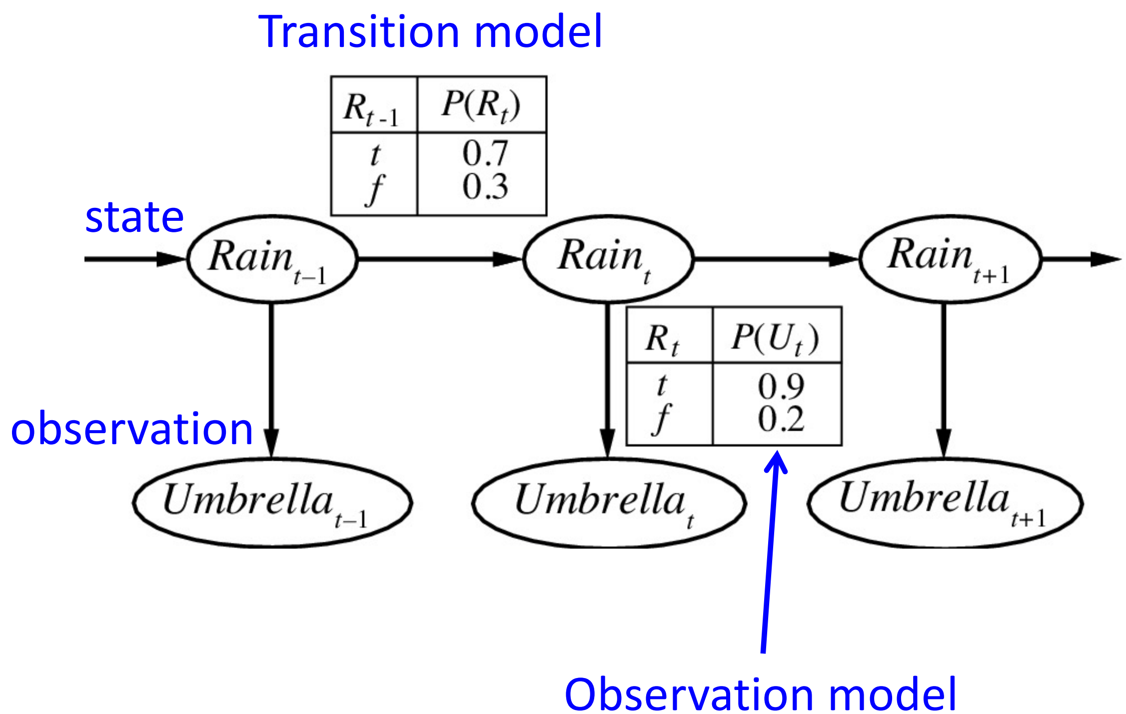
Outline

- HMM: Probabilistic reasoning over time
- Two views of an HMM: as a Bayes Net, as an FSM
- Inference: Belief propagation in an HMM
- Parameter learning: Maximum likelihood
- Parameter learning: EM and Hard EM

HMM as a Bayes Net

This slide shows an HMM as a Bayes Net. You should remember the graph semantics of a Bayes net:

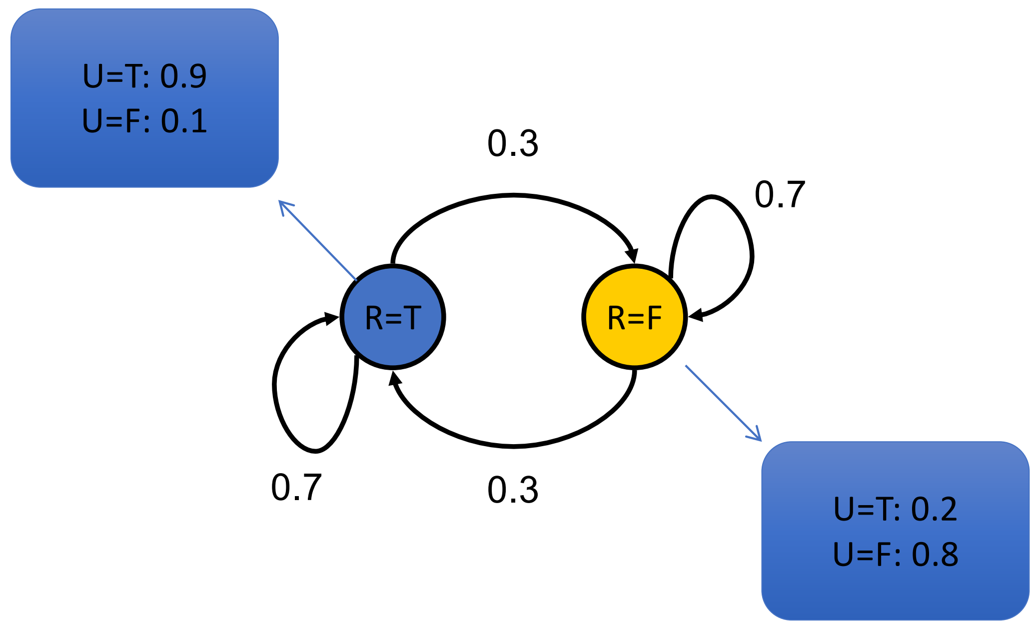
- Nodes are random variables.
- Edges denote stochastic dependence.



HMM as a Finite State Machine

This slide shows ***exactly the same HMM,*** viewed in a totally different way. Here, we show it as a finite state machine:

- Nodes denote states.
- Edges denote possible transitions between the states.
- Observation probabilities must be written using little table thingies, hanging from each state.



Transition probabilities

	R_t = T	R_t = F
R_{t-1} = T	0.7	0.3
R_{t-1} = F	0.3	0.7

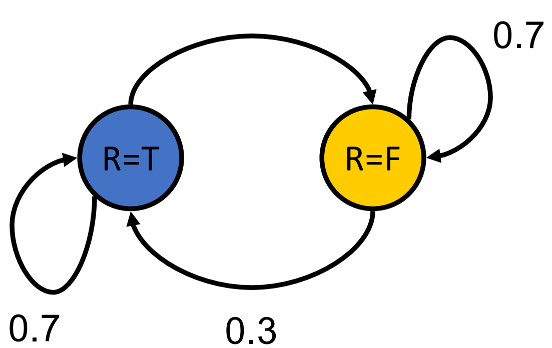
Observation probabilities

	U_t = T	U_t = F
R_t = T	0.9	0.1
R_t = F	0.2	0.8

Bayes Net vs. Finite State Machine

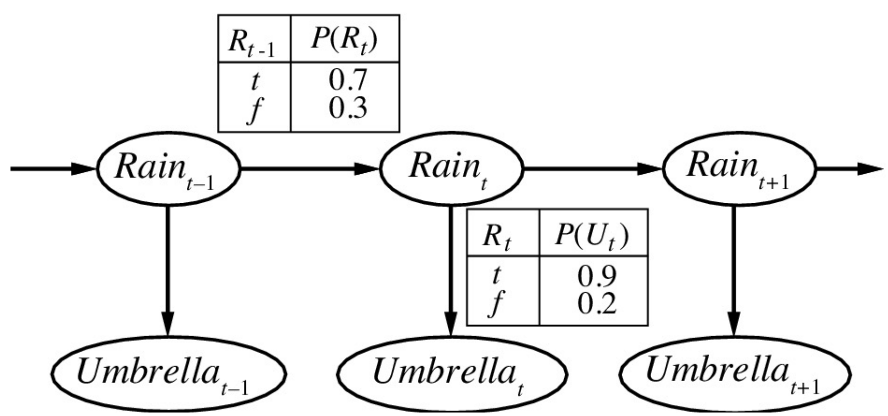
Finite State Machine:

- Lists the different possible states that the world can be in, at one particular time.
- Evolution over time is not shown.



Bayes Net:

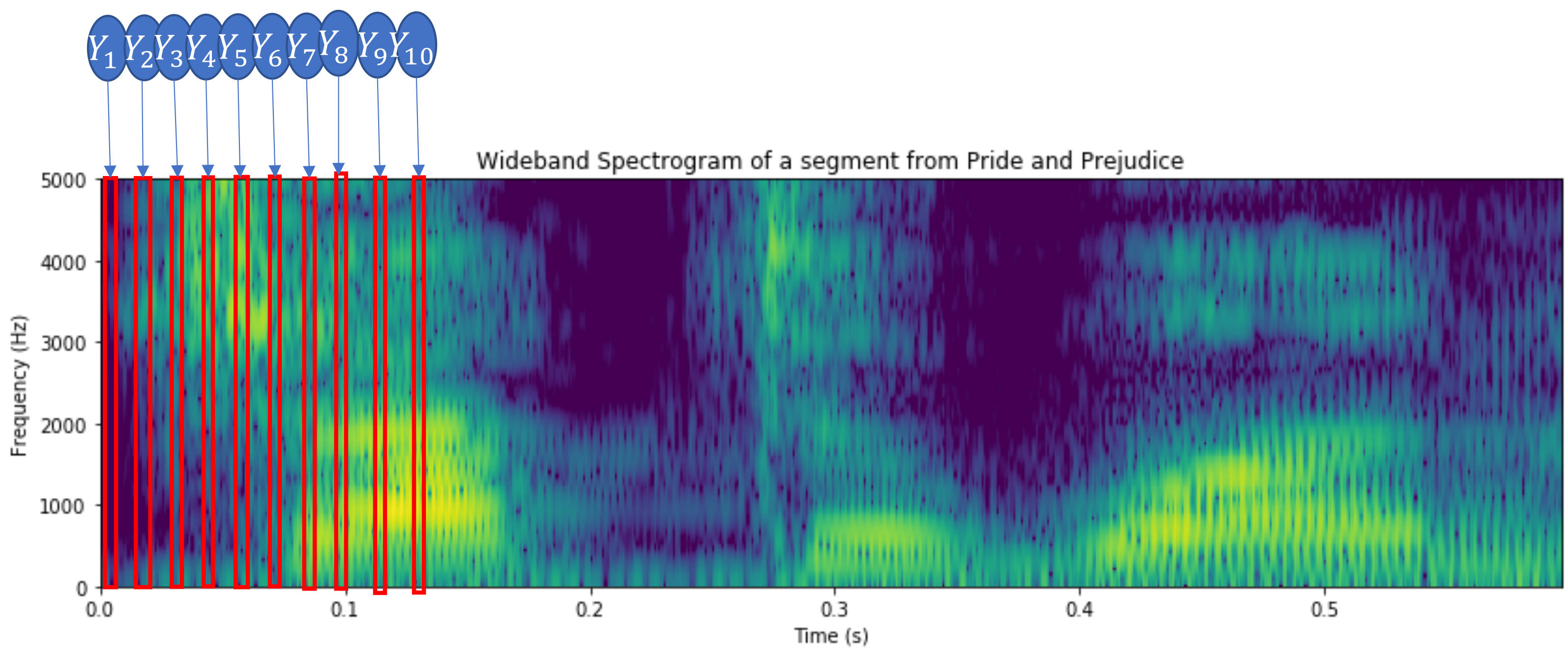
- Lists the different time slices.
- The various possible settings of the state variable are not shown.



Speech Recognition as a Bayes Net

- Observations: X_t = spectrum of 25ms frame of the speech signal.
- State: Y_t = phoneme or letter being currently produced

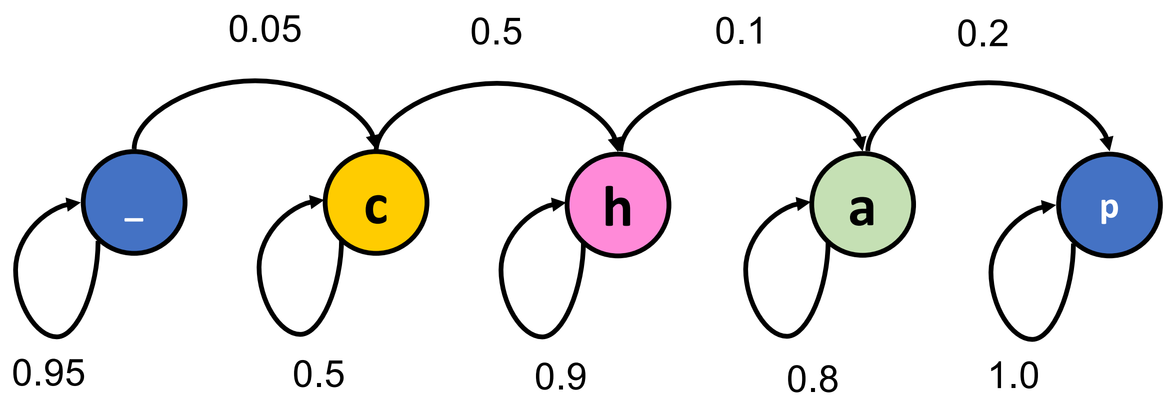
Example utterance: “chapter one,” from a Librivox recording of *Pride and Prejudice.*



Speech Recognition as a Finite State Machine

- Observations: X_t = spectrum of 10ms "frame" of the speech signal.
- States: Y_t = letter or phoneme.

Finite State Machine model of the first part of the word "chapter"



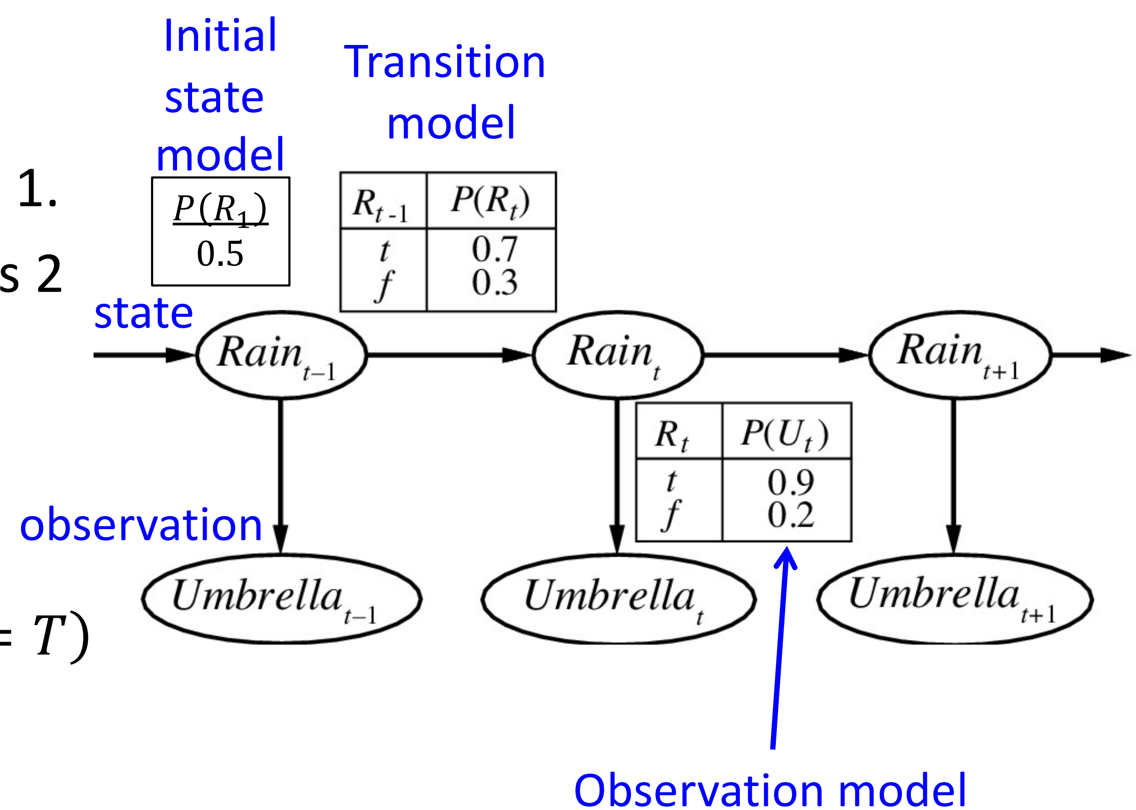
Outline

- HMM: Probabilistic reasoning over time
- Two views of an HMM: as a Bayes Net, as an FSM
- Inference: Belief propagation in an HMM
- Parameter learning: Maximum likelihood
- Parameter learning: EM and Hard EM

Belief propagation in an HMM: Example

- Elspeth has no umbrella on day 1.
- Elspeth has an umbrella on days 2 and 3.
- What is the probability that it's raining on day 3?

$$P(R_3 = T | U_1 = F, U_2 = T, U_3 = T)$$



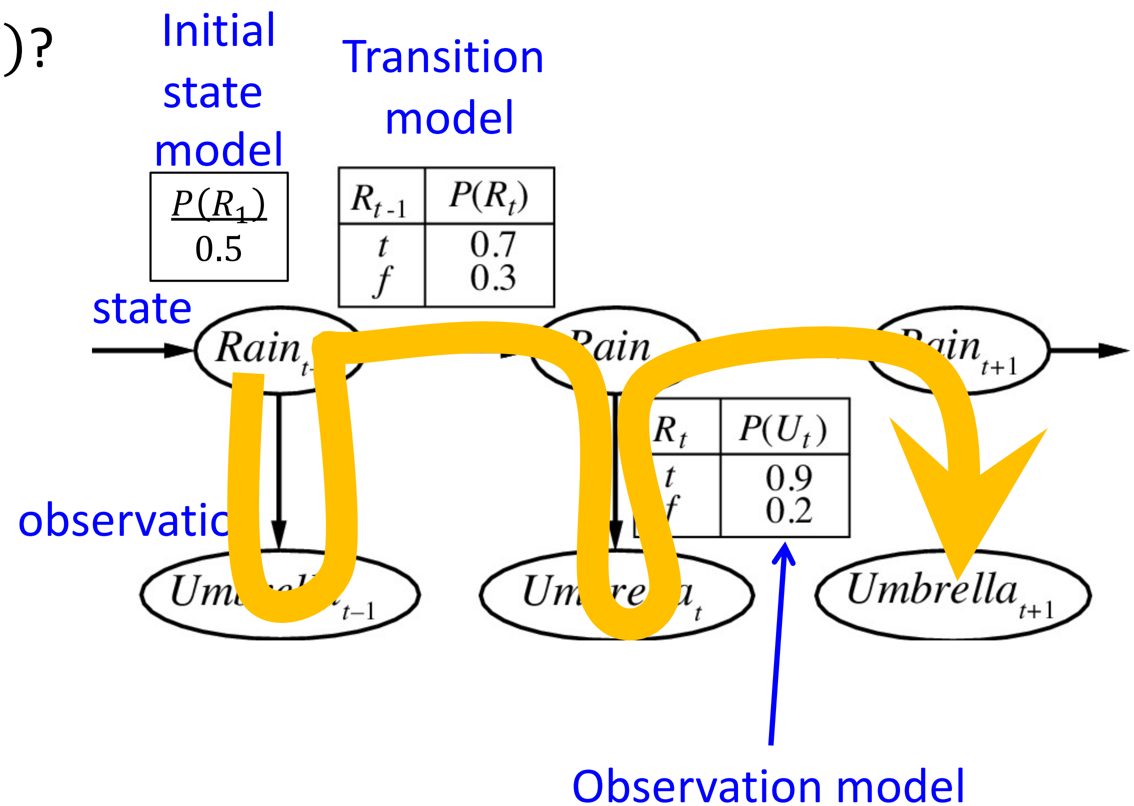
Belief propagation, step by step

1. Identify a path through the Bayesian network that includes all variables, including the query variable and all observed variables, starting at their common ancestor
2. Calculate the joint probability of the query variable and all observed variables, iteratively marginalizing out all intermediate variables step-by-step along the path.
3. Apply Bayes’ rule to get the desired conditional probability

Step 1: Identify a path starting at their common ancestor

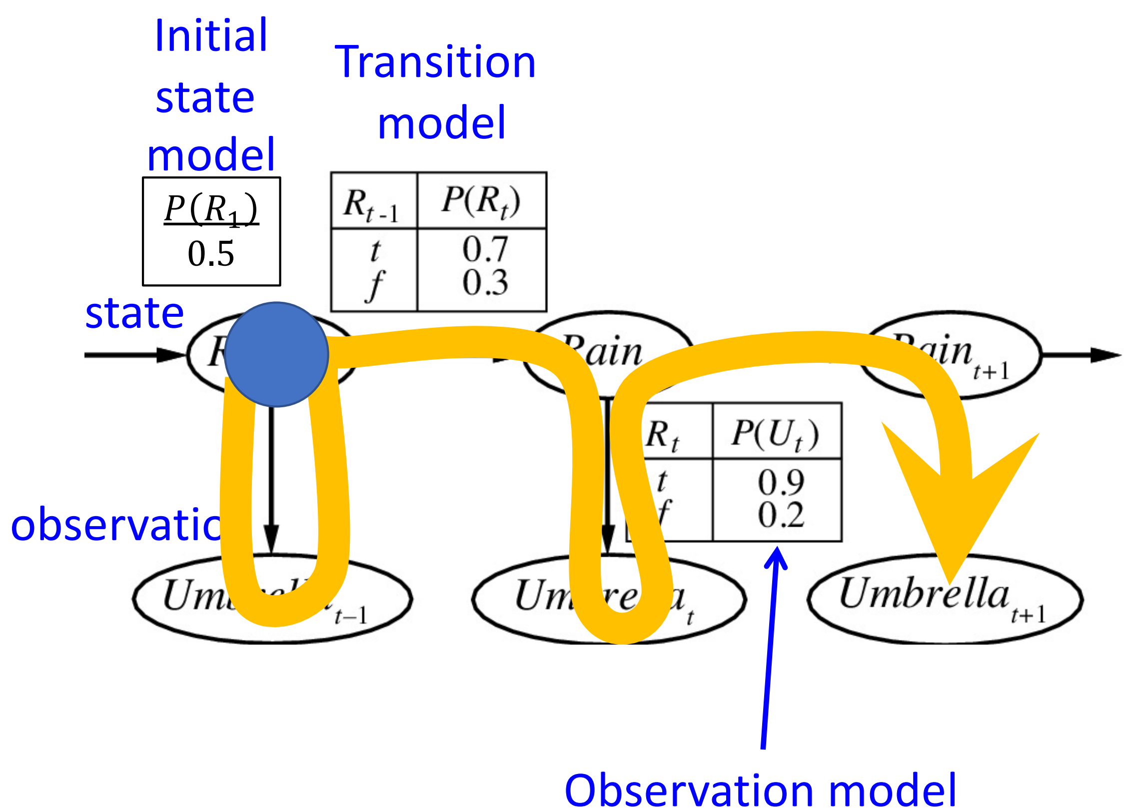
$P(R_3 = T | U_1 = F, U_2 = T, U_3 = T)$?

- Query variable: R_3
- Observed variables:
  - $U_1 = F$
  - $U_2 = T$
  - $U_3 = T$



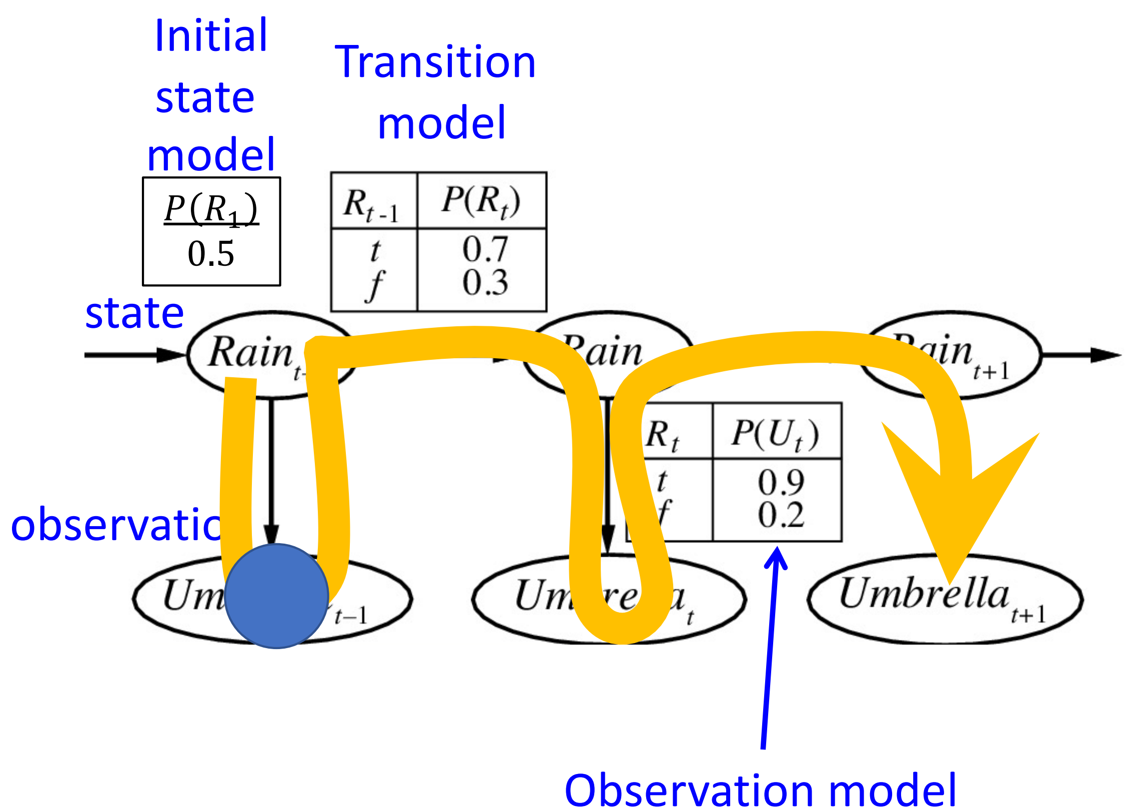
Step 2: Calculate the joint probability, step-by-step…

- $P(R_1 = T) = 0.5$
- $P(R_1 = F) = 0.5$



Step 2: Calculate the joint probability, step-by-step…

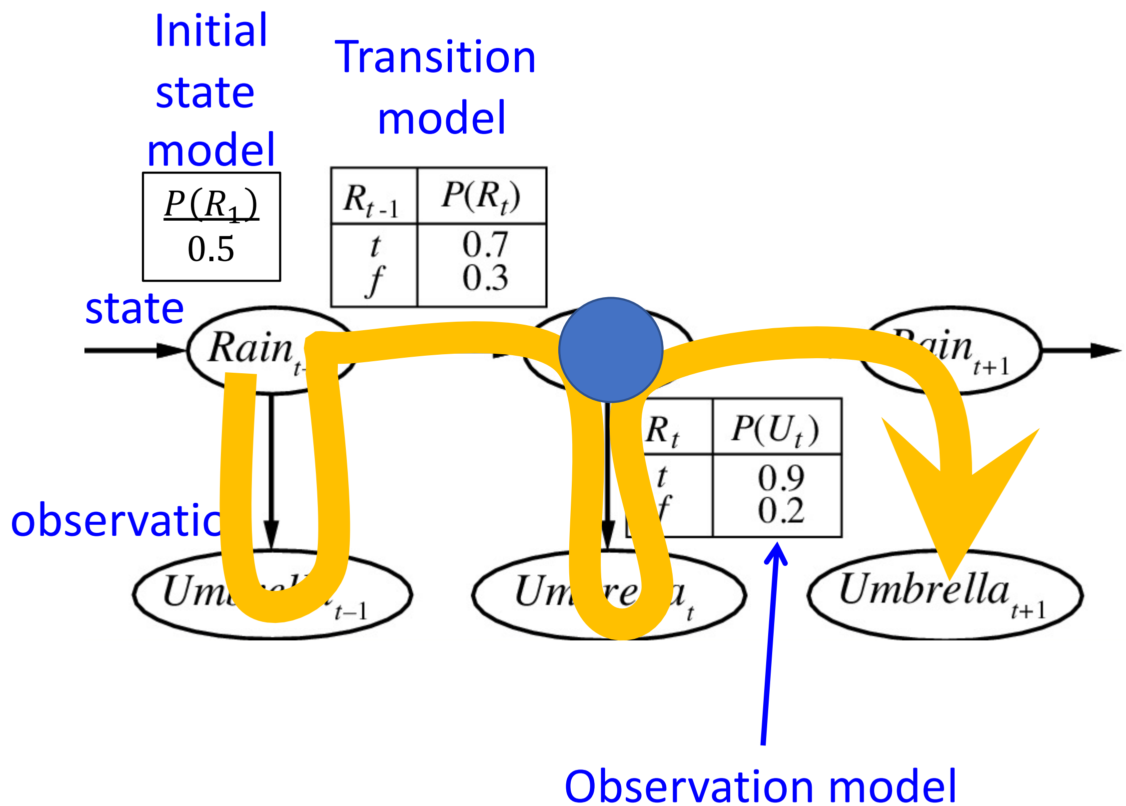
- $P(R_1 = T, U_1 = F) = (0.5)(0.1)$
- $P(R_1 = F, U_1 = F) = (0.5)(0.8)$



Step 2: Calculate the joint probability, step-by-step…

- $P(R_1 = T, U_1 = F, R_2 = T) = (0.5)(0.1)(0.7)$
- $P(R_1 = T, U_1 = F, R_2 = F) = (0.5)(0.1)(0.3)$

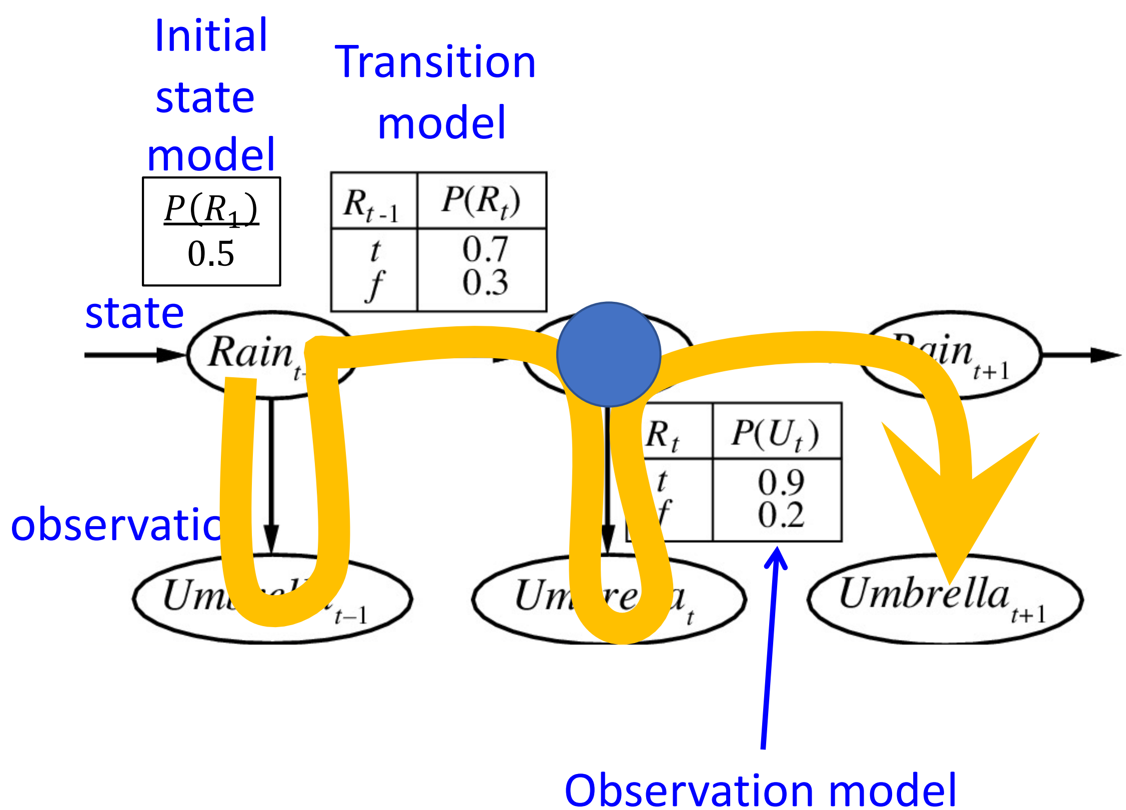
- $P(R_1 = F, U_1 = F, R_2 = T) = (0.5)(0.8)(0.3)$
- $P(R_1 = F, U_1 = F, R_2 = F) = (0.5)(0.8)(0.7)$



…iteratively marginalizing out intermediate variables as you go…

$$P(U_1 = F, R_2 = T) =$$
$$P(R_1 = T, U_1 = F, R_2 = T)$$
$$+ P(R_1 = F, U_1 = F, R_2 = T) =$$
$$(0.5)(0.1)(0.7) + (0.5)(0.8)(0.3) = 0.155$$

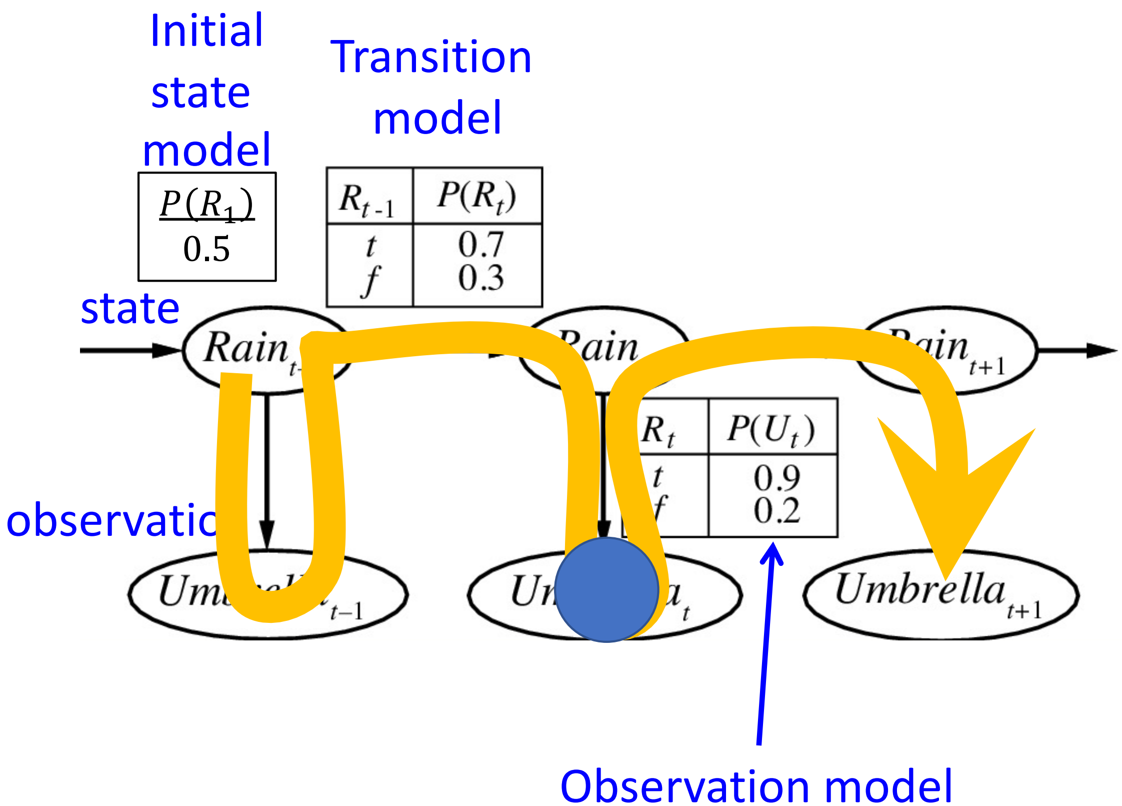
$$P(U_1 = F, R_2 = F) =$$
$$P(R_1 = T, U_1 = F, R_2 = F)$$
$$+ P(R_1 = F, U_1 = F, R_2 = F) =$$
$$(0.5)(0.1)(0.3) + (0.5)(0.8)(0.7) = 0.295$$



Step 2: Calculate the joint probability, step-by-step…

- $P(U_1 = F, R_2 = T, U_2 = T) =$
  $P(U_1 = F, R_2 = T) \times$
  $P(U_2 = T | R_2 = T) =$
  $(0.155)(0.9)$

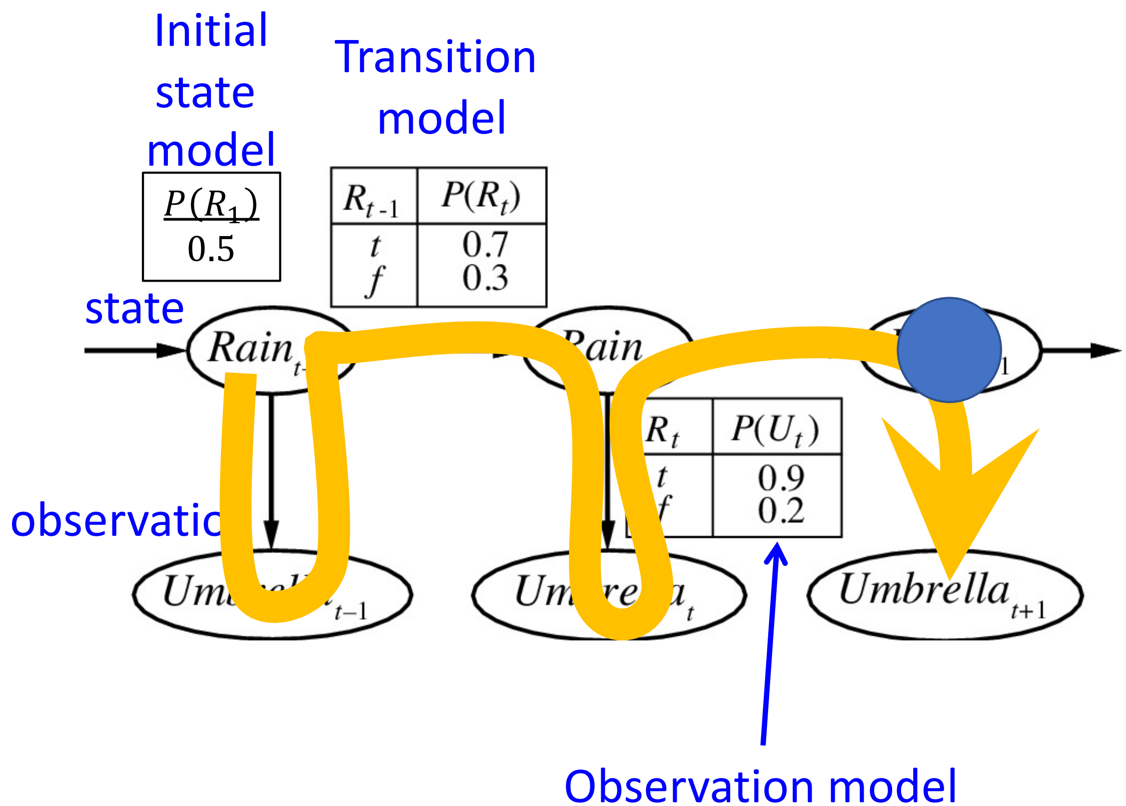
- $P(U_1 = F, R_2 = F, U_2 = T) =$
  $P(U_1 = F, R_2 = F) \times$
  $P(U_2 = T | R_2 = F) =$
  $(0.295)(0.2)$



Step 2: Calculate the joint probability, step-by-step…

- $P(U_1 = F, R_2 = T, U_2 = T, R_3) =$
  $P(U_1 = F, R_2 = T, U_2 = T) \times$
  $P(R_3 | R_2 = T) =$
  $(0.155)(0.9)(0.7 \text{ or } 0.3)$

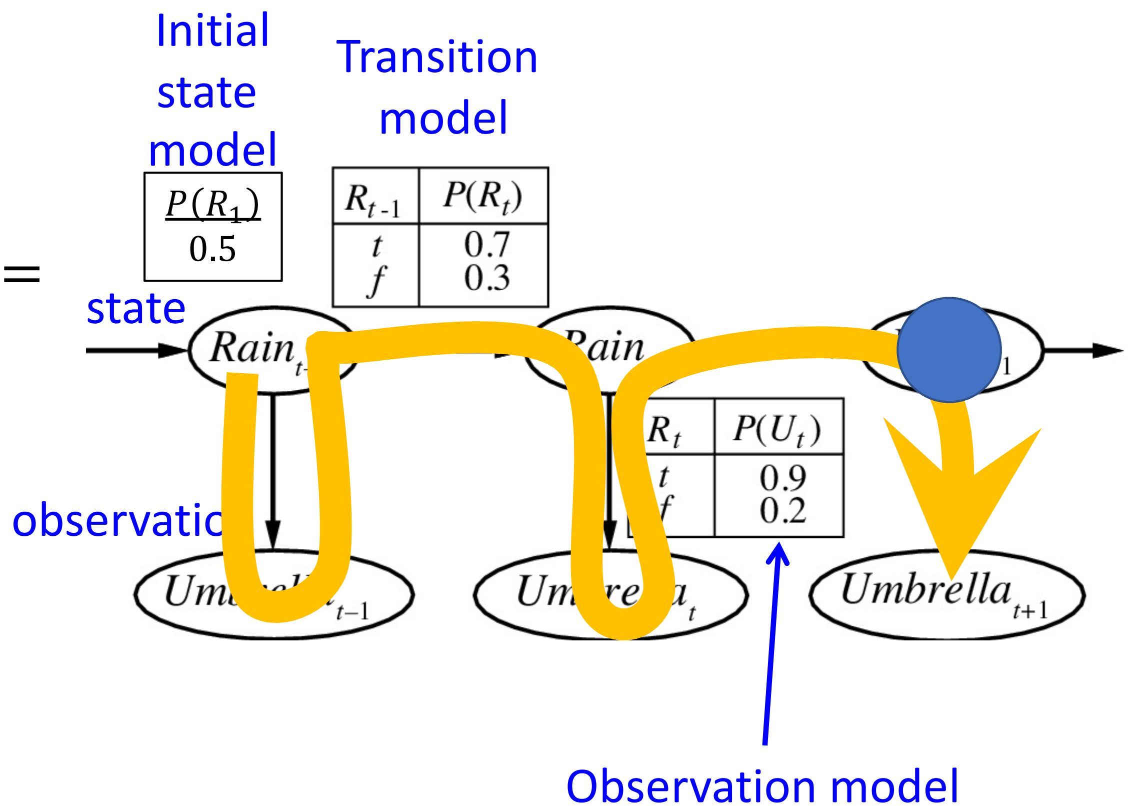
- $P(U_1 = F, R_2 = F, U_2 = T, R_3) =$
  $P(U_1 = F, R_2 = F, U_2 = T) \times$
  $P(R_3 | R_2 = F) =$
  $(0.295)(0.2)(0.3 \text{ or } 0.7)$



...iteratively marginalizing out intermediate variables as you go...

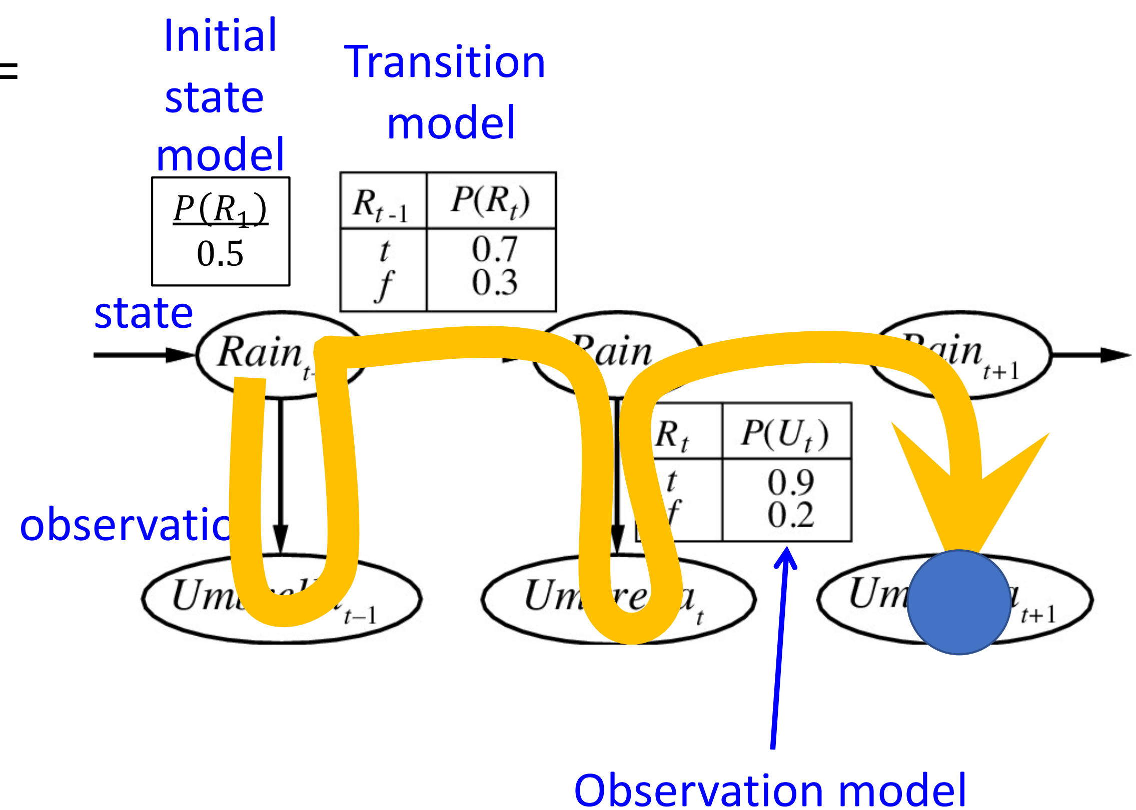
$$P(U_1 = F, U_2 = T, R_3) =$$
$$P(U_1 = F, R_2 = T, U_2 = T, R_3)$$
$$+ P(U_1 = F, R_2 = F, U_2 = T, R_3) =$$

$$\begin{cases} 0.11535 & R_3 = T \\ 0.08315 & R_3 = F \end{cases}$$



Step 2: Calculate the joint probability, step-by-step…

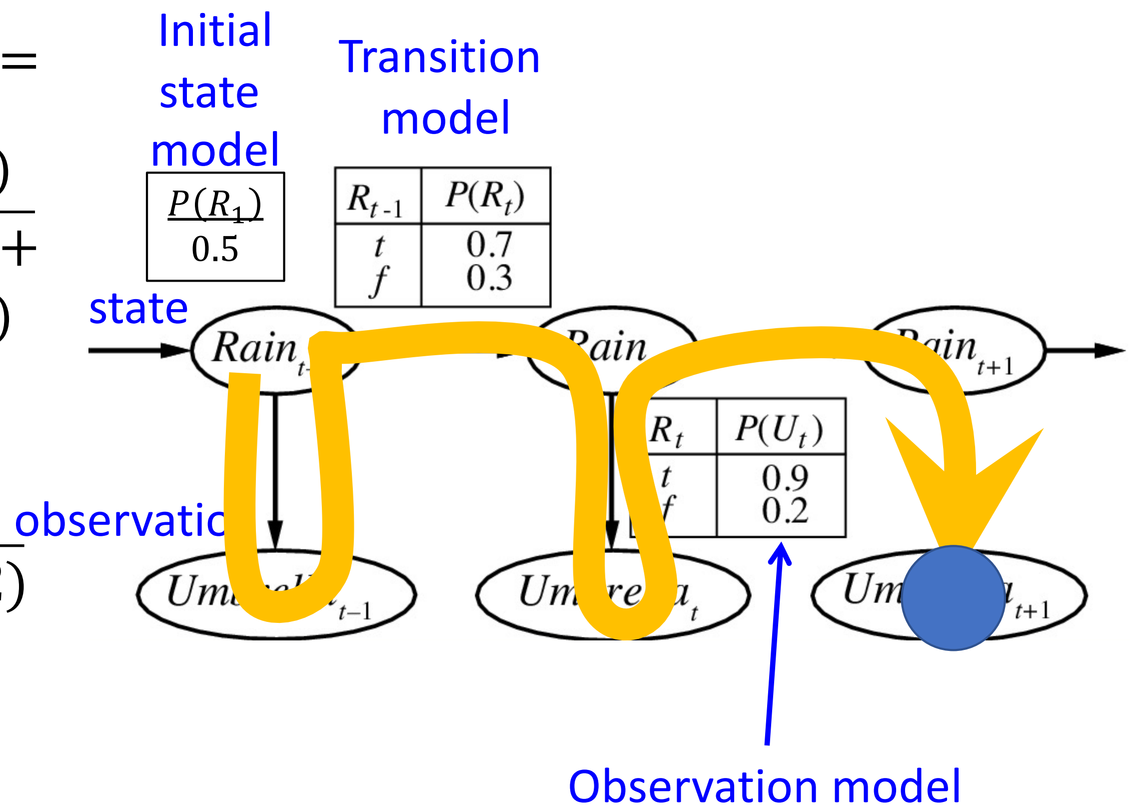
$$P(U_1 = F, U_2 = T, U_3 = T, R_3) =$$
$$P(U_1 = F, U_2 = T, R_3) \times$$
$$P(U_3 = T | R_3) =$$
$$\begin{cases} (0.11535)(0.9) & R_3 = T \\ (0.08315)(0.2) & R_3 = F \end{cases}$$



Step 3: Apply Bayes' rule to get conditional probability

$$P(R_3 = T | U_1 = F, U_2 = T, U_3 = T) = \frac{P(R_3 = T, U_1 = F, U_2 = T, U_3 = T)}{P(R_3 = T, U_1 = F, U_2 = T, U_3 = T) + P(R_3 = F, U_1 = F, U_2 = T, U_3 = T)}$$

$$= \frac{(0.11535)(0.9)}{(0.11535)(0.9) + (0.08315)(0.2)}$$



Belief propagation, step by step

1. Identify a path through the Bayesian network that includes all variables, including the query variable and all observed variables, starting at their common ancestor
2. Calculate the joint probability of the query variable and all observed variables, iteratively marginalizing out all intermediate variables step-by-step along the path.
    1. Product Step: $P(A, B, C) = P(A, B)P(C|A, B)$
    2. Sum Step: $P(A, C) = \sum_b P(A, B = b, C)$
3. Apply Bayes' rule to get the desired conditional probability

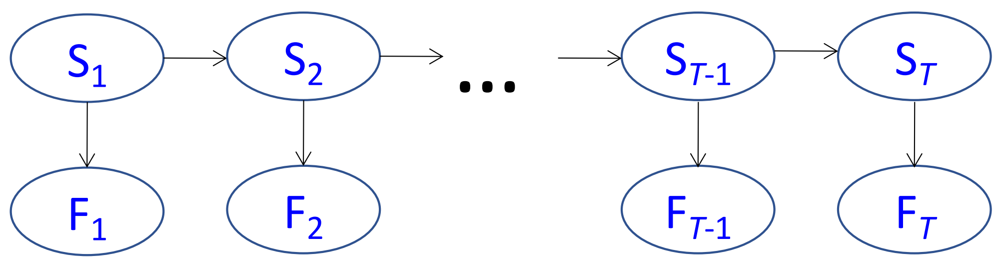
Outline

- HMM: Probabilistic reasoning over time
- Two views of an HMM: as a Bayes Net, as an FSM
- Inference: Belief propagation in an HMM
- Parameter learning: Maximum likelihood
- Parameter learning: EM and Hard EM

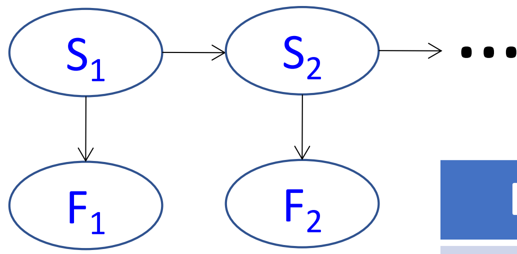
Flying Cows

The University of Illinois Vaccavolatology Department has a new model of the way in which cows learn to fly.

- If a smart cow arrives in the pasture, it tends to remain for more than one day. There is a transition probability, $P(S_t|S_{t-1})$.
- If there is smart cow present, then on that day, it is likely that one or more cows will fly away: $P(F_t|S_t)$.



Flying cows

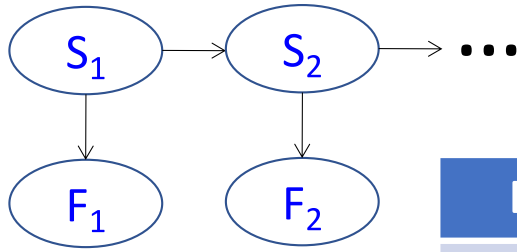


The Vaccavolatologists went out to watch a nearby pasture for ten days.

- Their results are shown in the table at left (True is marked as "T"; False is shown with a blank).

Day	S	F
1		
2		
3	T	
4	T	T
5	T	
6	T	T
7	T	T
8		
9		T
10		

Maximum Likelihood



The transition probabilities can be estimated as:

$$P(S_t = T | S_{t-1} = T) =$$

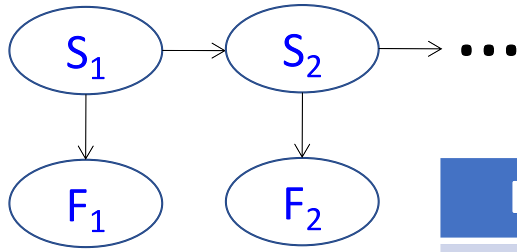
$$\frac{\#\text{ days }(S_t = T, S_{t-1}=\text{T})}{\#\text{ days }(S_{t-1} = T)} = \frac{4}{5}$$

$$P(S_t = T | S_{t-1} = F)$$

$$= \frac{\#\text{ days }(S_t = T, S_{t-1} = F)}{\#\text{ days }(S_{t-1} = F)} = \frac{1}{4}$$

Day	S	F
1		
2		
3	T	
4	T	T
5	T	
6	T	T
7	T	T
8		
9		T
10		

Maximum Likelihood



The observation probabilities can be estimated as:

$$P(F_t = T | S_t = T) =$$

$$\frac{\#\text{ days }(F_t = T, S_t = T)}{\#\text{ days }(S_t = T)} = \frac{3}{5}$$

$$P(F_t = T | S_t = F) =$$

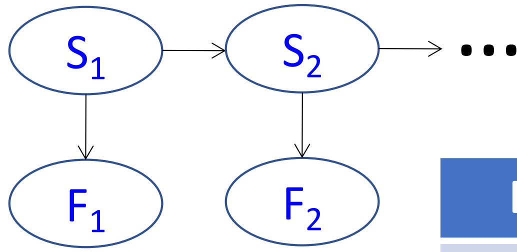
$$\frac{\#\text{ days }(F_t = T, S_t = F)}{\#\text{ days }(S_t = F)} = \frac{1}{5}$$

Day	S	F
1		
2		
3	T	
4	T	T
5	T	
6	T	T
7	T	T
8		
9		T
10		

Outline

- HMM: Probabilistic reasoning over time
- Two views of an HMM: as a Bayes Net, as an FSM
- Inference: Belief propagation in an HMM
- Parameter learning: Maximum likelihood
- Parameter learning: EM

Missing data



What can we do if some of the observations are missing?

Day	S	F
1		
2		
3	T	
4	T	T
5	T	
6	T	T
7	?	T
8	?	
9	?	T
10	?	

Missing data

What can we do if some of the observations are missing?

- Answer: we can use EM, just like any other Bayes Net.

$$P(S_t = T | S_{t-1} = T) =$$

$$\frac{E[\#\text{ days } (S_t = T, S_{t-1}=\mathrm{T})]}{\mathrm{E}[\#\text{ days } (S_{t-1} = T)]} =$$

$$\frac{\sum_{t=1}^{T} P(S_t = T, S_{t-1} = T | \text{ observations})}{\sum_{t=1}^{T} P(S_{t-1} = T | \text{ observations})}$$

Outline

- HMM: Probabilistic reasoning over time

$$P(Y_{0:T}, X_{1:T}) = P(Y_0) \prod_{t=1}^{T} P(Y_t|Y_{t-1})\, P(X_t|Y_t)$$

- Two views of an HMM: as a Bayes Net, as an FSM
- Inference: Belief propagation in an HMM
- Parameter learning: Maximum likelihood

$$P(S_t|S_{t-1}) = \frac{\#\text{ days }(S_t, S_{t-1})}{\#\text{ days }(S_{t-1})}$$

- Parameter learning: EM

$$P(S_t|S_{t-1}) = \frac{E[\#\text{ days }(S_t, S_{t-1})]}{\mathrm{E}[\#\text{ days }(S_{t-1})]}$$