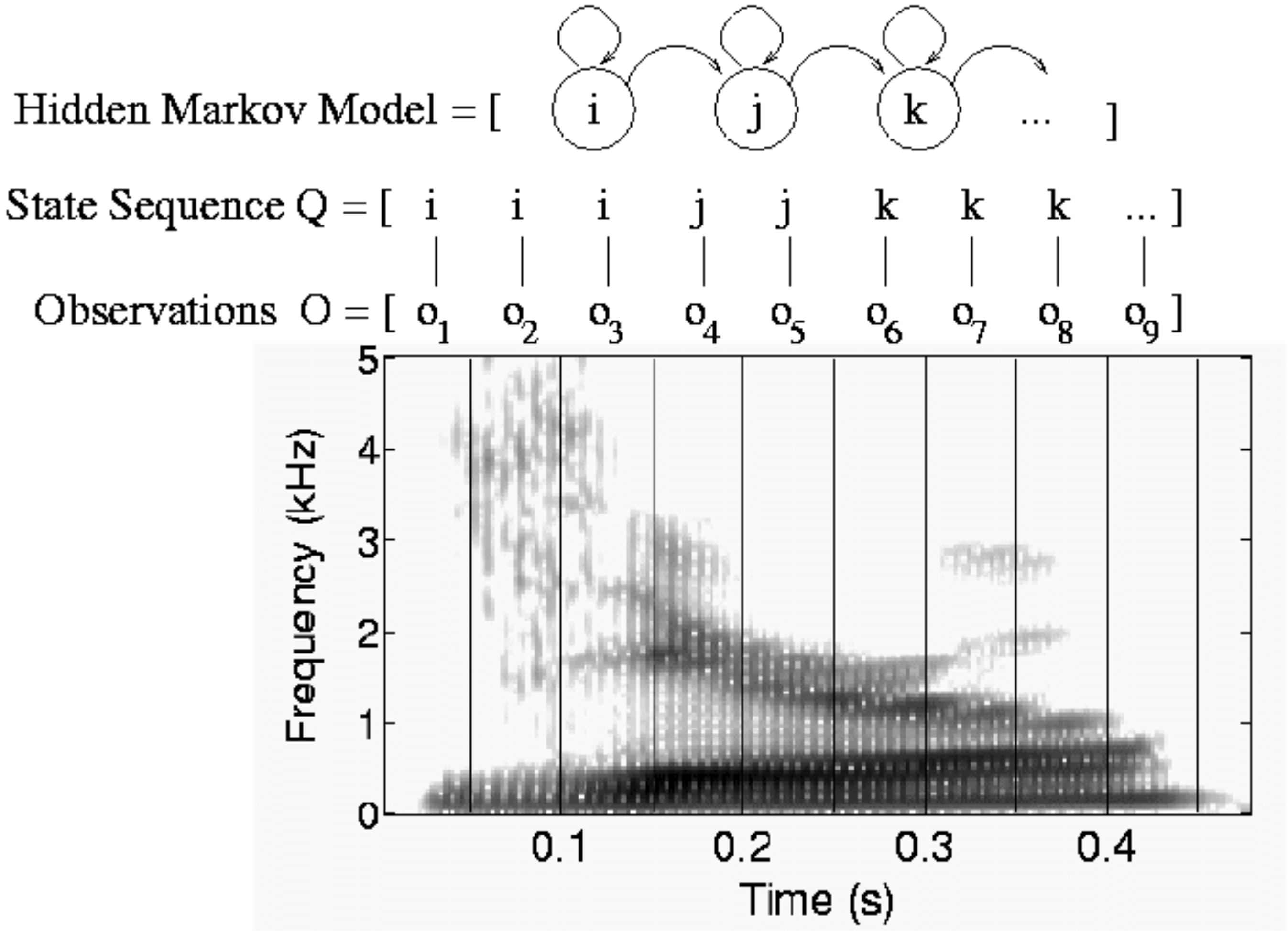
CS440/ECE448 Lecture 18: Hidden Markov Models

Mark Hasegawa-Johnson, 3/2020

Including slides by Svetlana Lazebnik

CC-BY 3.0

You may remix or redistribute if you cite the source.

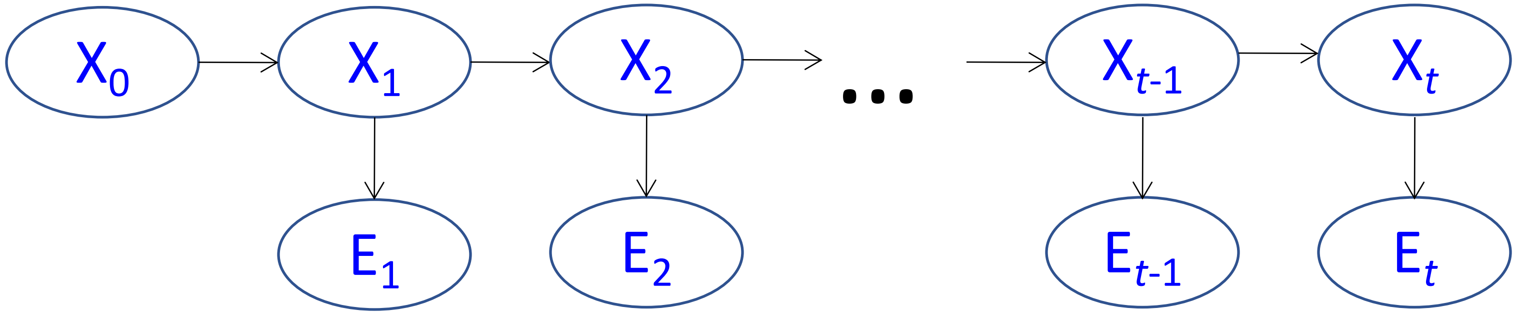


Probabilistic reasoning over time

- 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,

Hidden Markov Models

- At each time slice t, the state of the world is described by an unobservable variable X_t and an observable *evidence* variable E_t
- **Transition model:** distribution over the current state given the whole past history:
  $P(X_t \mid X_0, \ldots, X_{t-1}) = P(X_t \mid \mathbf{X}_{0:t-1})$
- **Observation model:** $P(E_t \mid \mathbf{X}_{0:t}, \mathbf{E}_{1:t-1})$

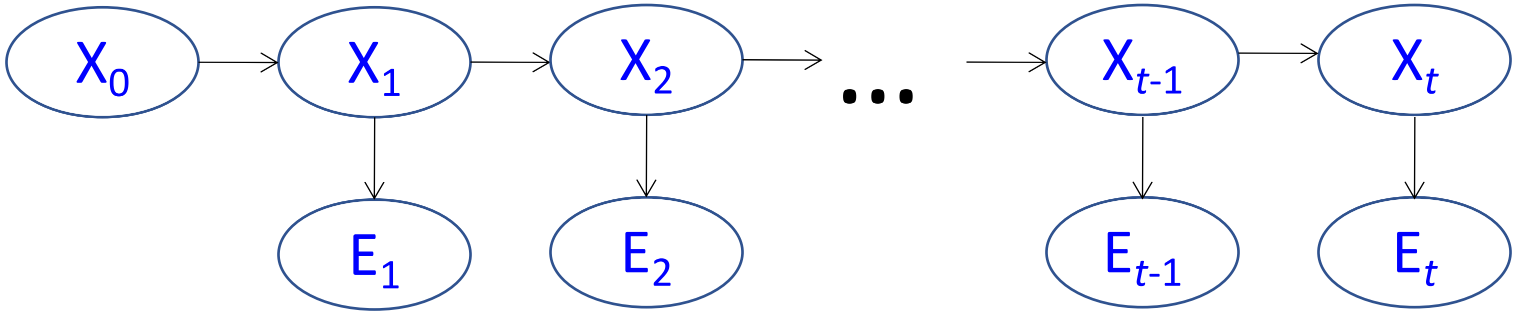


Hidden Markov Models

- **Markov assumption** (first order)
  - The current state is conditionally independent of all the other states given the state in the previous time step
  - What does $P(X_t \mid \mathbf{X}_{0:t-1})$ simplify to?

    $P(X_t \mid \mathbf{X}_{0:t-1}) = P(X_t \mid X_{t-1})$
- Markov assumption for observations
  - The evidence at time *t* depends only on the state at time *t*
  - What does $P(E_t \mid \mathbf{X}_{0:t}, \mathbf{E}_{1:t-1})$ simplify to?

    $P(E_t \mid \mathbf{X}_{0:t}, \mathbf{E}_{1:t-1}) = P(E_t \mid X_t)$



Example Scenario: UmbrellaWorld

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

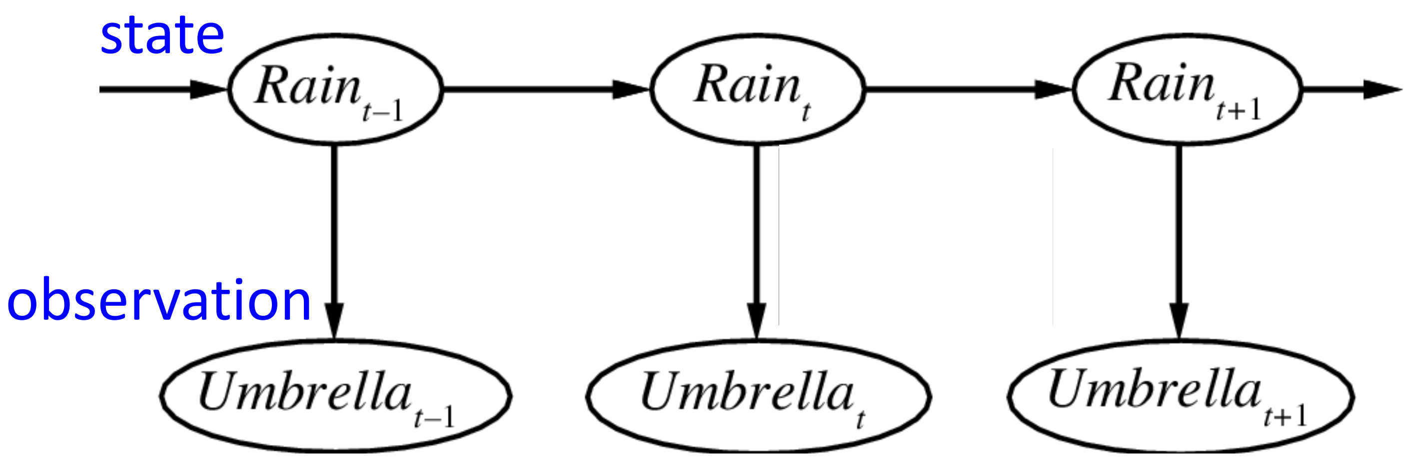
- Elspeth Dunsany is an AI researcher at the Canadian company Unitek.
- Richard Feynman is an AI, named after the famous physicist, whose personality he resembles.
- To keep him from escaping, Richard's workstation is not connected to the internet. He knows about rain but has never seen it.
- He has 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}) makes rain on day t (R_t) more likely.
- Elspeth usually brings her umbrella (U_t) on days when it rains (R_t), but not always.



Example Scenario: UmbrellaWorld

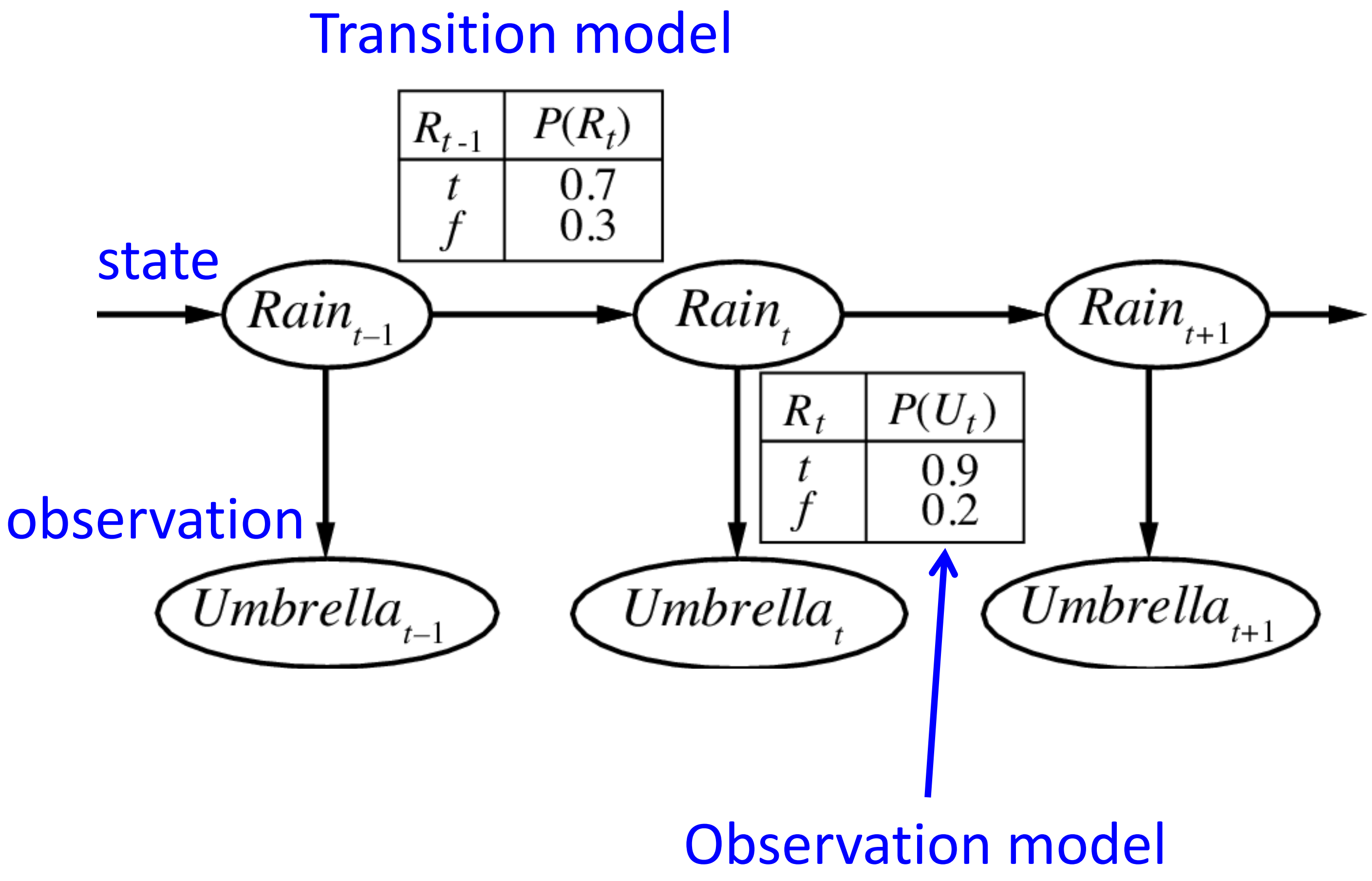
Characters from the novel *Hammered* by Elizabeth Bear,
Scenario from chapter 15 of Russell & Norvig

- Richard learns that the weather changes on 3 out of 10 days, thus

$$P(R_t|R_{t-1}) = 0.7$$
$$P(R_t|\neg R_{t-1}) = 0.3$$

- 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,

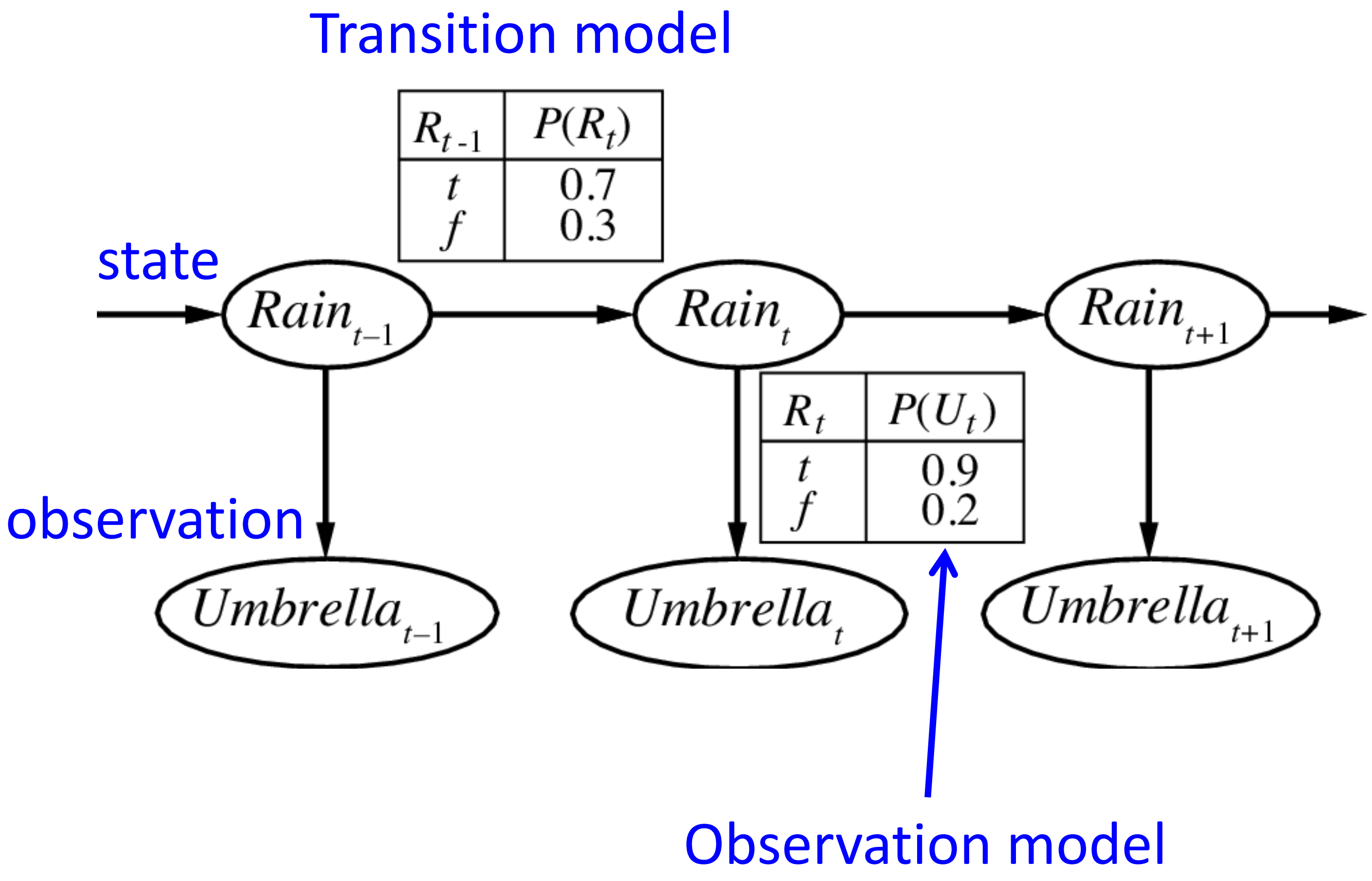
$$P(U_t|R_t) = 0.9$$
$$P(U_t|\neg R_t) = 0.2$$



HMM as a Bayes Net

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

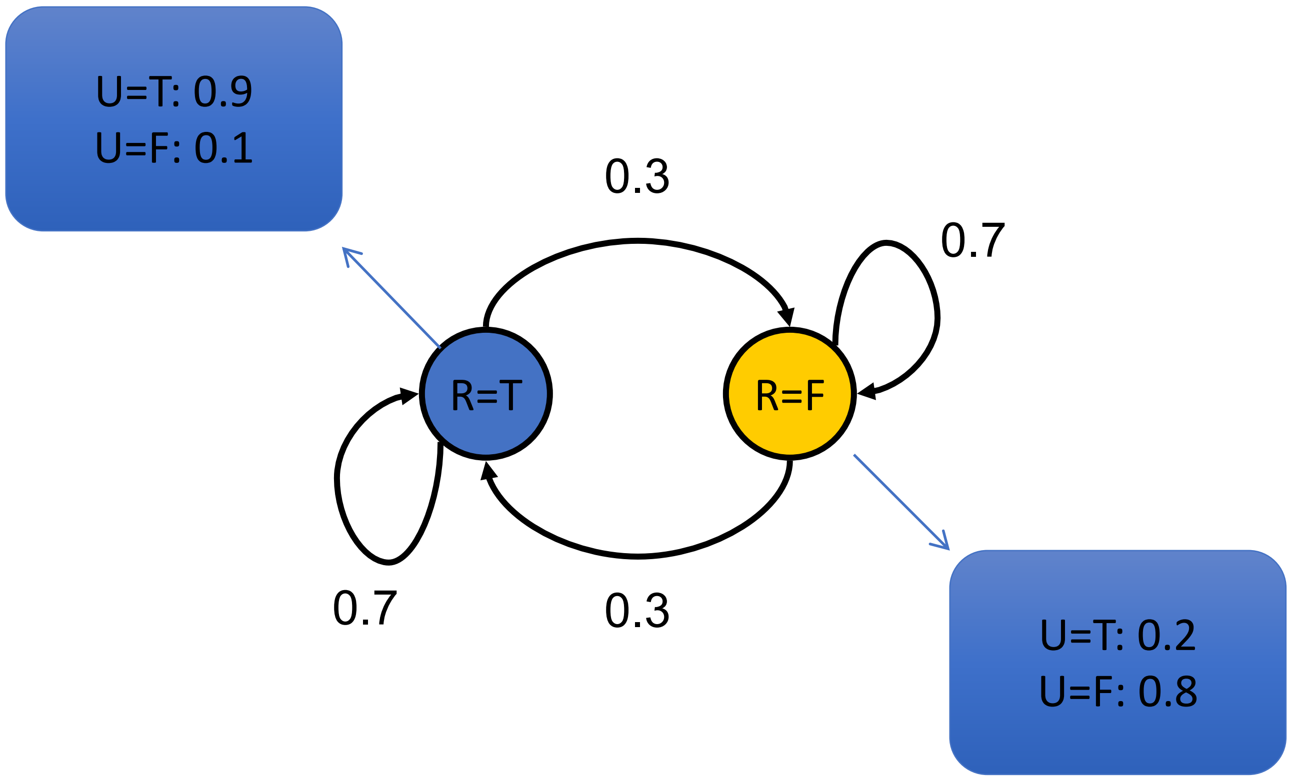
- Nodes are random variables.
- Edges denote stochastic dependence.



HMM as a Finite State Machine

This slide shows ***<u>exactly the same HMM,</u>*** 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

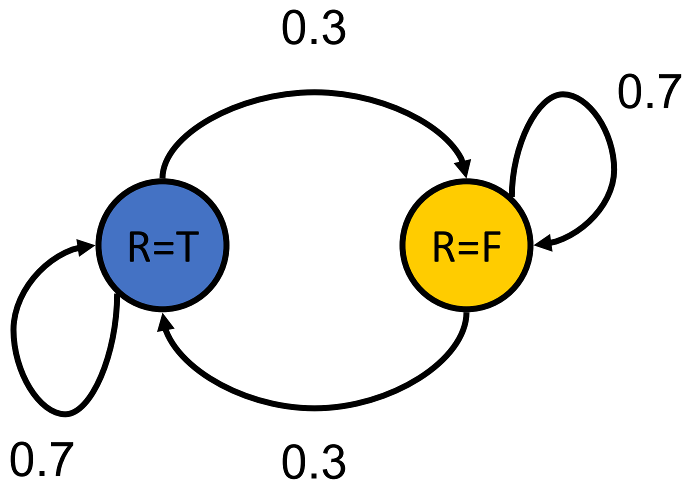
Observation probabilities

	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

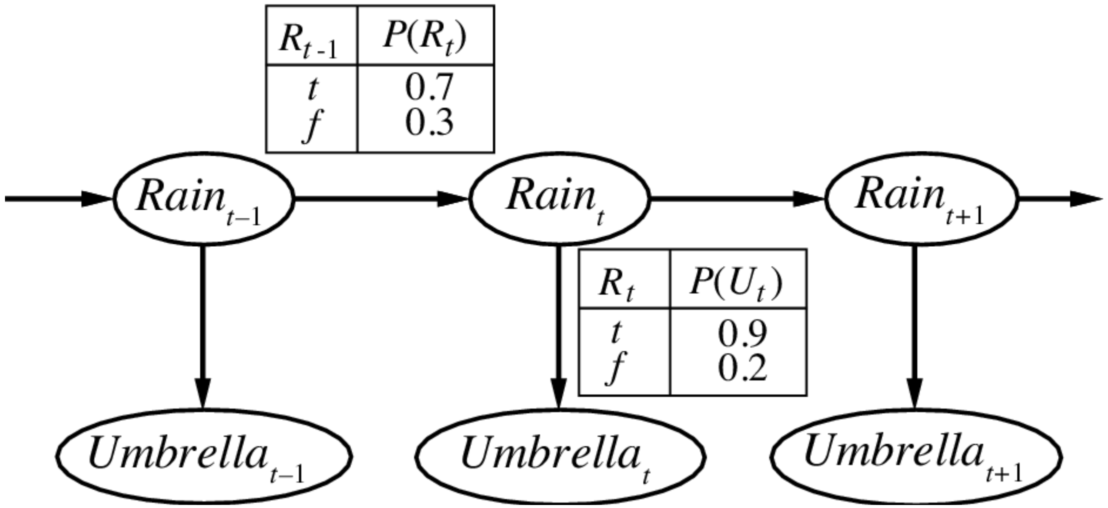
Finite State Machine:

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



Bayes Net:

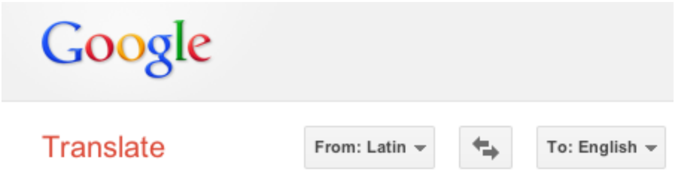
- Lists the different time slices.
- The various possible settings of the state variable are not shown.



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 translation options

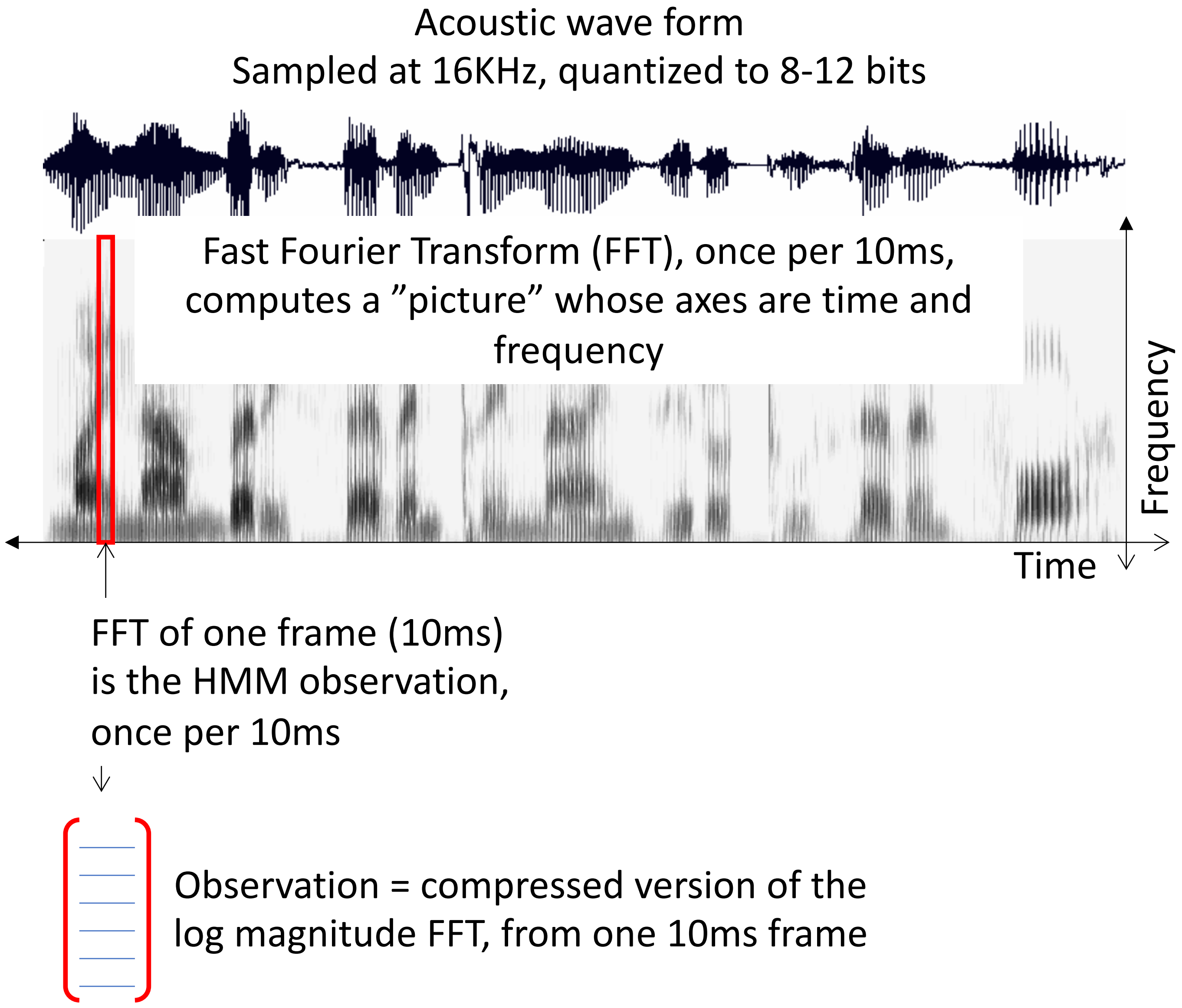


- Robot tracking:
  - Observations are range readings (continuous)
  - States are positions on a map (continuous)

Source: Tamara Berg

Example: Speech Recognition

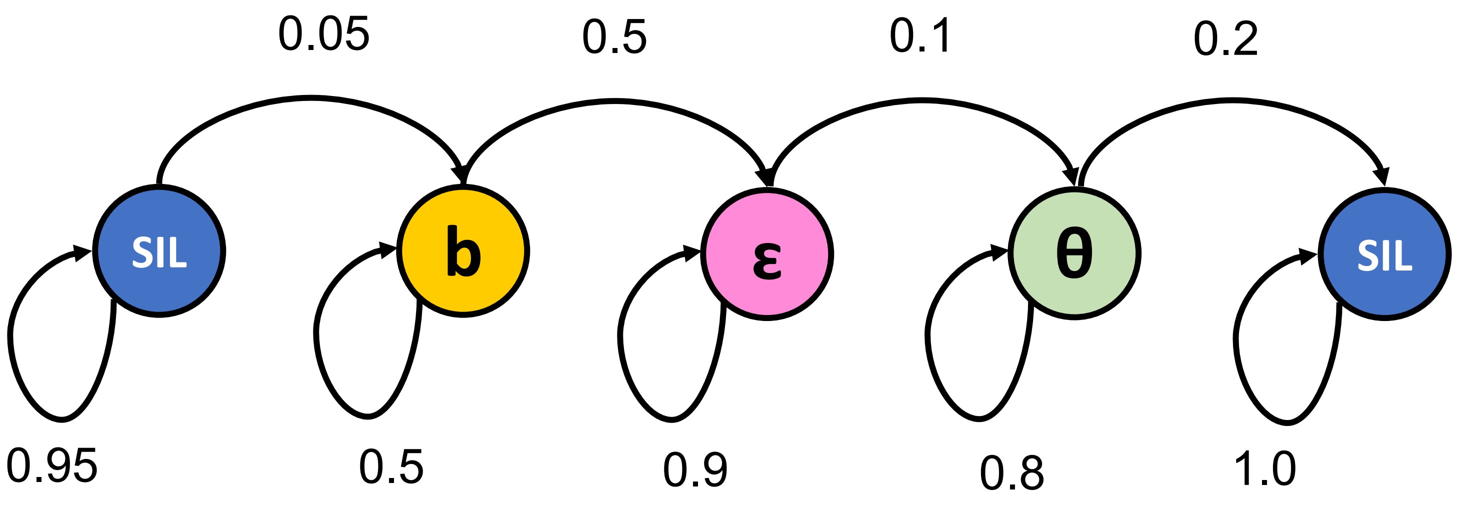
- Observations: E_t = FFT of 10ms “frame” of the speech signal.



Example: Speech Recognition

- Observations: E_t = FFT of 10ms "frame" of the speech signal.
- States: X_t = a specific position in a specific word, coded using the international phonetic alphabet:
  - b = first sound of the word "Beth"
  - ε = second sound of the word "Beth"
  - θ = third sound in the word "Beth"

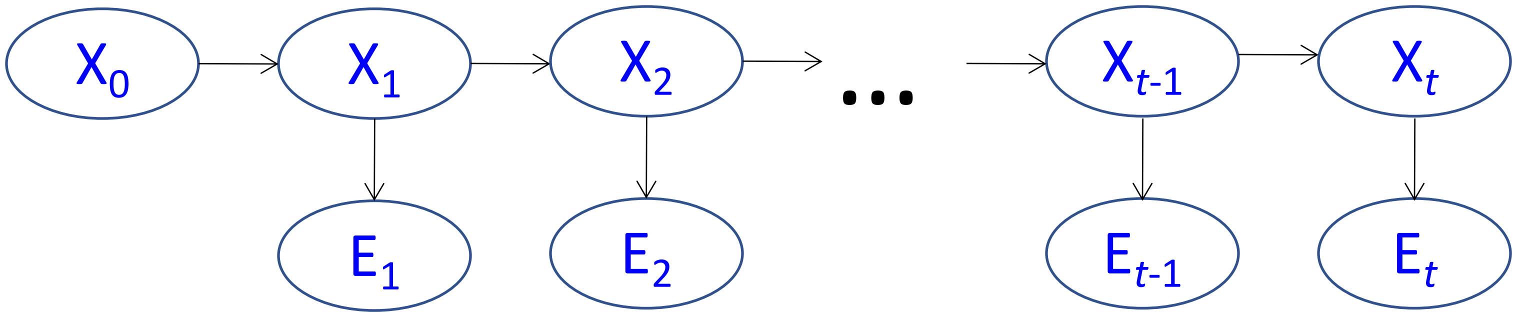
Finite State Machine model of the word "Beth"



The Joint Distribution

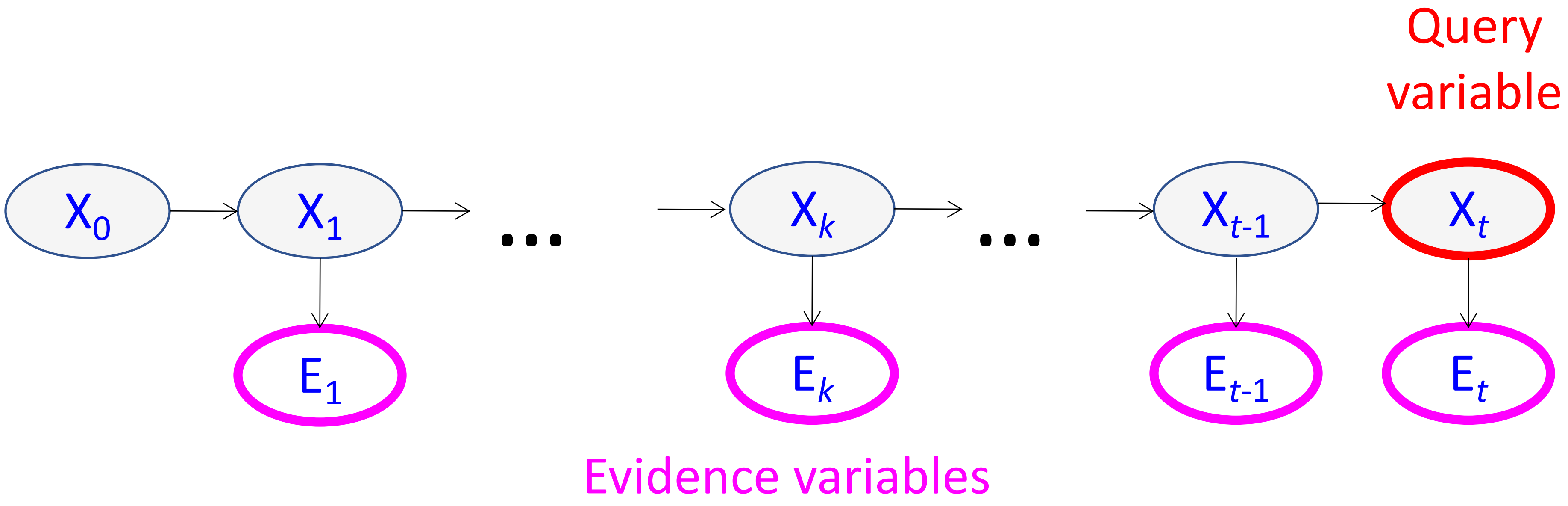
- Transition model: $P(X_t \mid \mathbf{X}_{0:t-1}) = P(X_t \mid X_{t-1})$
- Observation model: $P(E_t \mid \mathbf{X}_{0:t}, \mathbf{E}_{1:t-1}) = P(E_t \mid X_t)$
- How do we compute the full joint probability table
  $P(\mathbf{X}_{0:t}, \mathbf{E}_{1:t})$?

$$P(\boldsymbol{X}_{0:t}, \boldsymbol{E}_{1:t}) = P(X_0) \prod_{i=1}^{t} P(X_i | X_{i-1}) P(E_i | X_i)$$



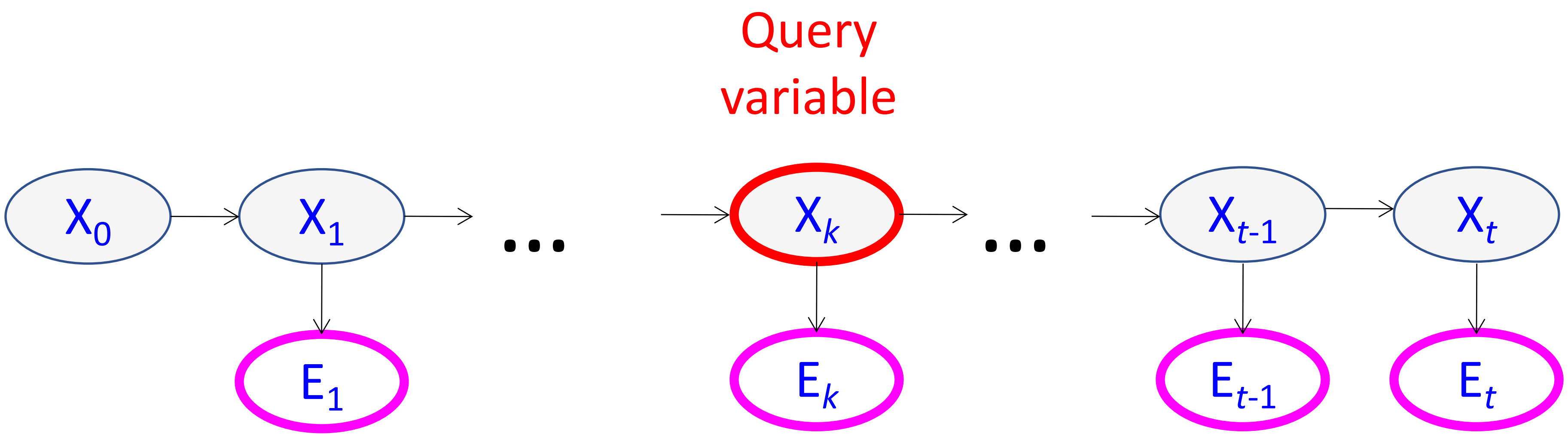
HMM inference tasks

- **Filtering:** what is the distribution over the current state X_t given all the evidence so far, $\mathbf{E}_{1:t}$?   (example: is it currently raining?)



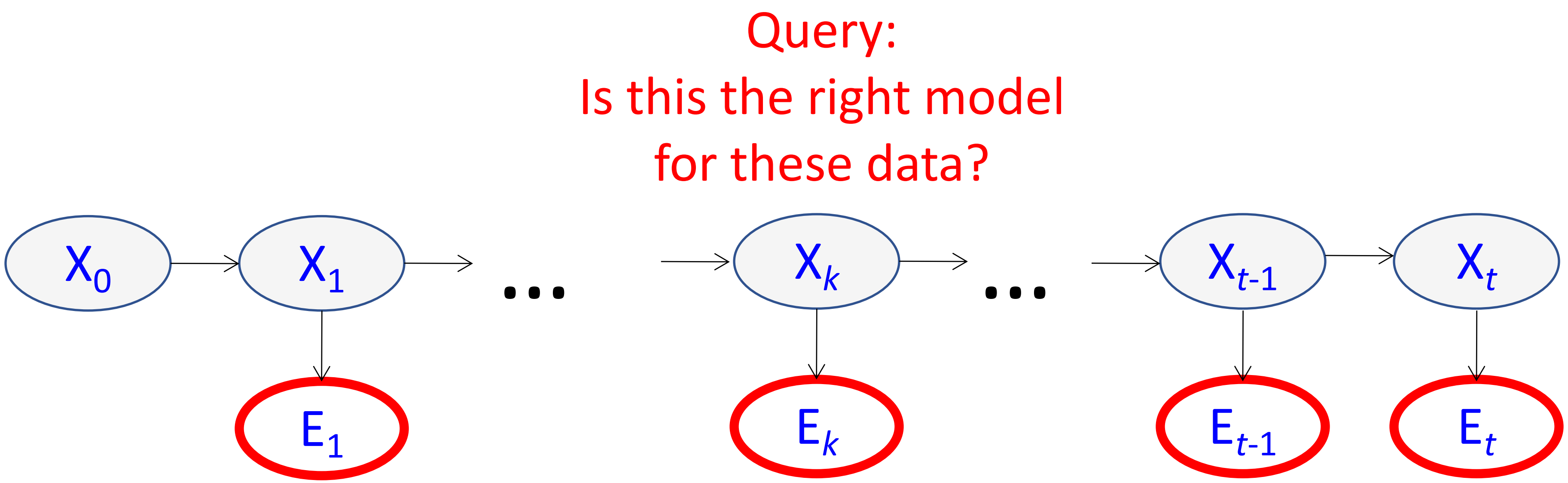
HMM inference tasks

- **Filtering:** what is the distribution over the current state X_t given all the evidence so far, $\mathbf{E}_{1:t}$?
- **Smoothing:** what is the distribution of some state X_k (k<t) given the entire observation sequence $\mathbf{E}_{1:t}$? (example: did it rain on Sunday?)



HMM inference tasks

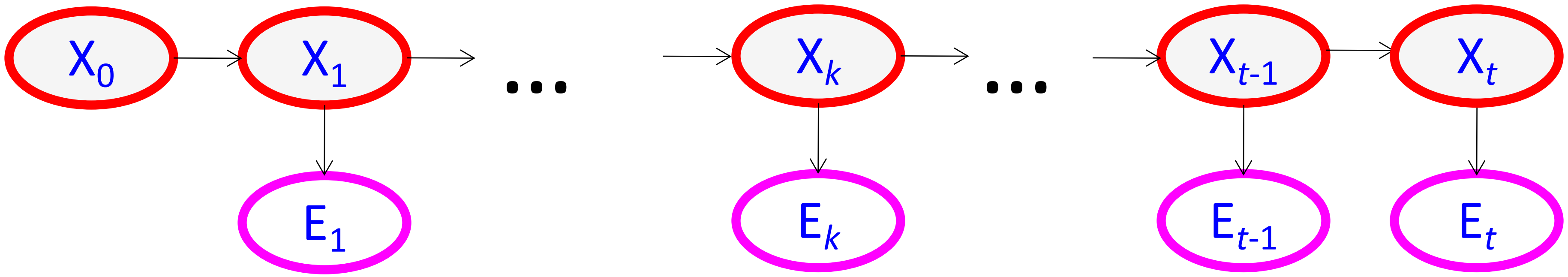
- **Filtering:** what is the distribution over the current state X_t given all the evidence so far, $\mathbf{E}_{1:t}$?
- **Smoothing:** what is the distribution of some state X_k (k<t) given the entire observation sequence $\mathbf{E}_{1:t}$?
- **Evaluation:** compute the probability of a given observation sequence $\mathbf{E}_{1:t}$ (example: is Richard using the right model?)



HMM inference tasks

- **Filtering:** what is the distribution over the current state X_t given all the evidence so far, $\mathbf{E}_{1:t}$
- **Smoothing:** what is the distribution of some state X_k ($k<t$) given the entire observation sequence $\mathbf{E}_{1:t}$?
- **Evaluation:** compute the probability of a given observation sequence $\mathbf{E}_{1:t}$
- **Decoding:** what is the most likely state sequence $\mathbf{X}_{0:t}$ given the observation sequence $\mathbf{E}_{1:t}$? (example: what's the weather every day?)

Query variables: all of them



HMM Learning and Inference

- Inference tasks
  - **Filtering:** what is the distribution over the current state X_t given all the evidence so far, $\mathbf{E}_{1:t}$
  - **Smoothing:** what is the distribution of some state X_k ($k<t$) given the entire observation sequence $\mathbf{E}_{1:t}$?
  - **Evaluation:** compute the probability of a given observation sequence $\mathbf{E}_{1:t}$
  - **Decoding:** what is the most likely state sequence $\mathbf{X}_{0:t}$ given the observation sequence $\mathbf{E}_{1:t}$?
- Learning
  - Given a training sample of sequences, learn the model parameters (transition and emission probabilities)

Filtering and Decoding in UmbrellaWorld

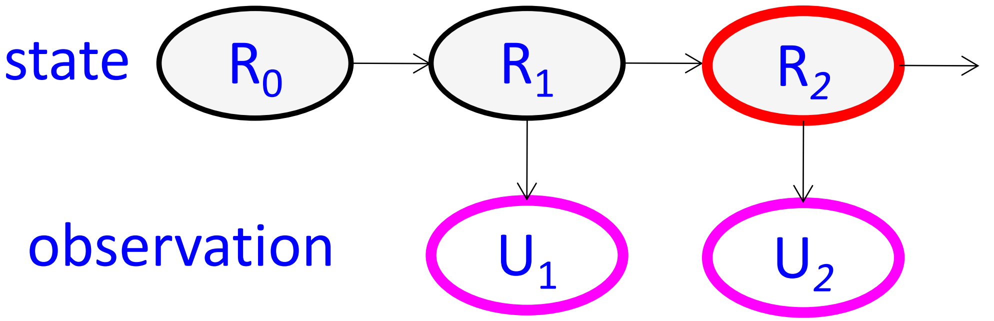
Filtering: Richard observes Elspeth's umbrella on day 2, but not on day 1. What is the probability that it's raining on day 2?

$$P(R_2|\neg U_1, U_2)?$$

Decoding: Same observation. What is the most likely sequence of hidden variables?

$$\underset{R_1,R_2}{\operatorname{argmax}} P(R_1, R_2|\neg U_1, U_2)\ ?$$

Transition model



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 Inference for HMMs

To calculate a probability $P(R_2|U_1,U_2)$:

1. **Select:** which variables do we need, in order to model the relationship among U_1, U_2, and R_2?
   - We need also R_0 and R_1.
2. **Multiply** to compute joint probability:

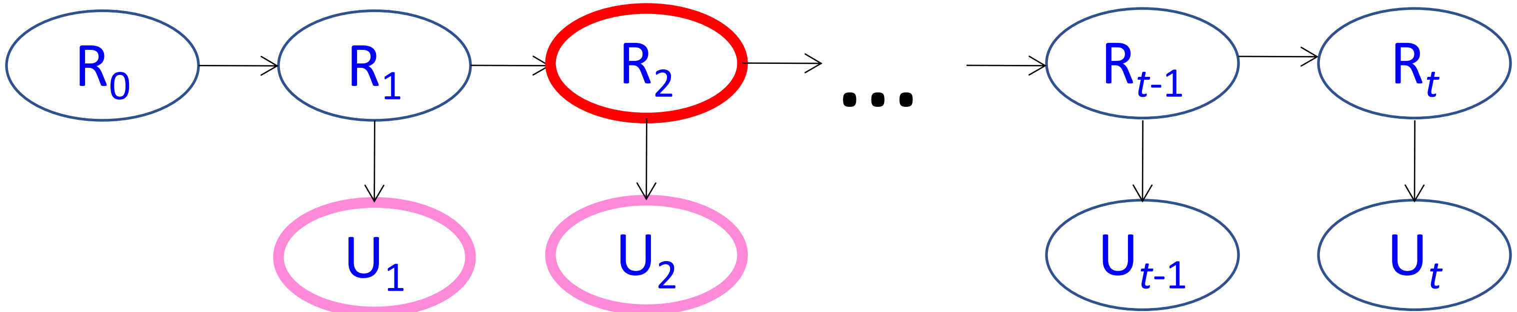
$$P(R_0,R_1,R_2,U_1,U_2) = P(R_0)P(R_1|R_0)P(U_1|R_1) \ldots P(U_2|R_2)$$

3. **Add** to eliminate those we don't care about

$$P(R_2,U_1,U_2) = \sum_{R_0,R_1} P(R_0,R_1,R_2,U_1,U_2)$$

4. **Divide:** use Bayes' rule to get the desired conditional

$$P(R_2|U_1,U_2) = P(R_2,U_1,U_2)/P(U_1,U_2)$$



Filtering and Decoding in UmbrellaWorld

1. **Select**:

To represent the relationship among

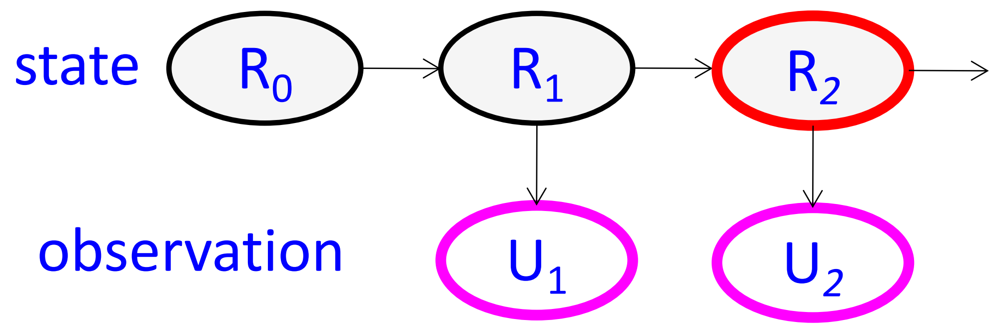
$$P(R_2|\neg U_1, U_2)?$$

…we also need knowledge of R_0 and R_1.

- In particular, we need the initial state probability, $P(R_0)$.
- It wasn't specified in the problem statement! Therefore we are justified in making any reasonable assumption, and clearly stating our assumption. Let's assume

$$P(R_0) = 0.5$$

Transition model



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

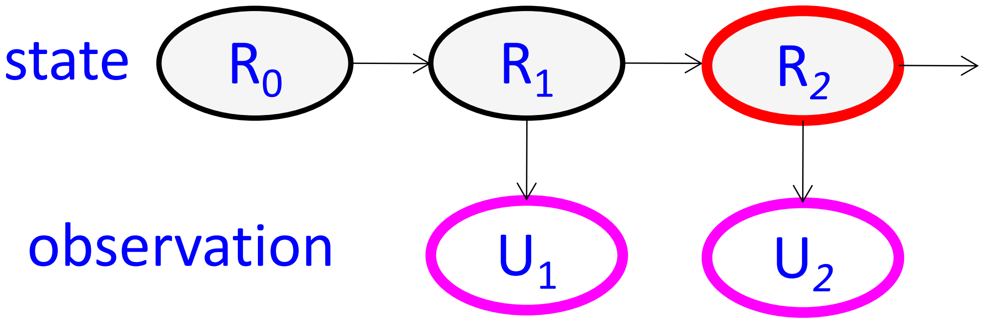
Filtering and Decoding in UmbrellaWorld

2. **Multiply:**

$$P(R_0, R_1, R_2, U_1, U_2) =$$
$$P(R_0)P(R_1|R_0)P(U_1|R_1) \ldots P(U_2|R_2)$$

	$\neg R_2 \neg U_2$	$\neg R_2 U_2$	$R_2 \neg U_2$	$R_2 U_2$
$\neg R_0 \neg R_1 \neg U_1$	0.1568	0.0392	0.0084	0.0756
$\neg R_0 \neg R_1 U_1$	0.0392	0.0098	0.0021	0.0189
$\neg R_0 R_1 \neg U_1$	0.0036	0.0009	0.0011	0.0095
$\neg R_0 R_1 U_1$	0.0324	0.0081	0.0095	0.0851
$R_0 \neg R_1 \neg U_1$	0.0672	0.0168	0.0036	0.0324
$R_0 \neg R_1 U_1$	0.0168	0.0042	0.009	0.0081
$R_0 R_1 \neg U_1$	0.0084	0.0021	0.0025	0.0221
$R_0 R_1 U_1$	0.0756	0.0189	0.0221	0.1985

Transition model



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

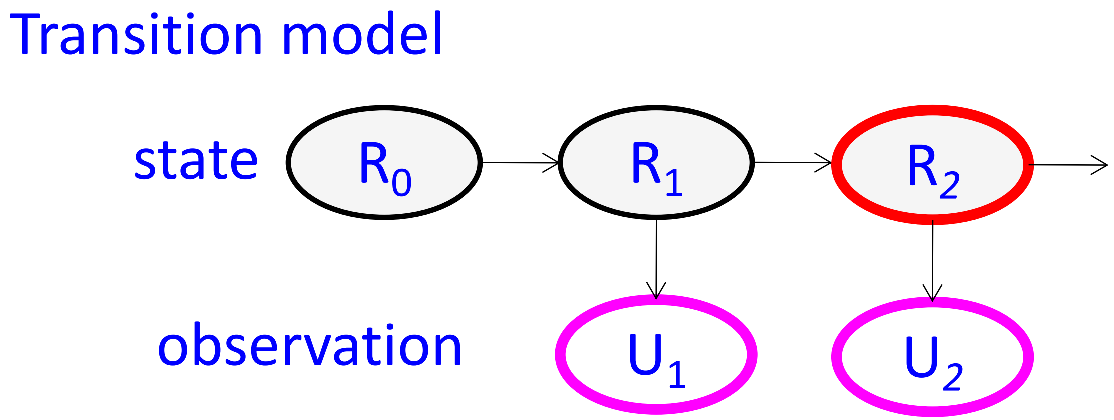
	U_t = T	U_t = F
R_t = T	0.9	0.1
R_t = F	0.2	0.8

Filtering and Decoding in UmbrellaWorld

3. <u>Add:</u>

$$P(R_2, U_1, U_2) = \sum_{R_0,R_1} P(R_0, R_1, R_2, U_1, U_2)$$

	$\neg U_1 \neg U_2$	$\neg U_1 U_2$	$U_1 \neg U_2$	$U_1 U_2$
$\neg R_2$	0.236	0.059	0.164	0.041
R_2	0.0155	0.1395	0.0345	0.3105



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

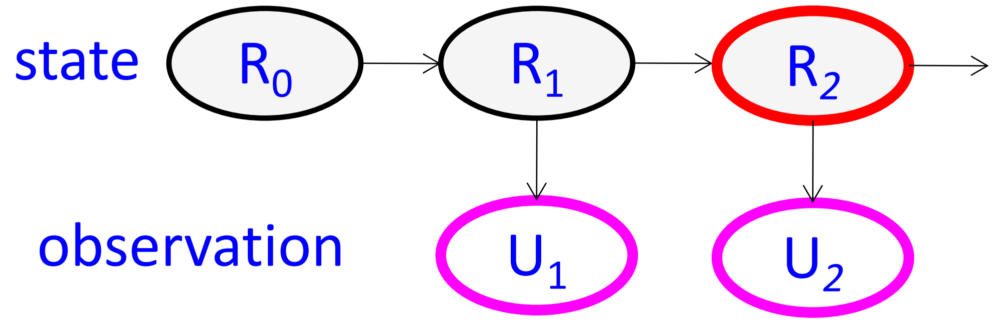
Filtering and Decoding in UmbrellaWorld

4. Divide:

$$P(R_2|U_1,U_2) = P(R_2,U_1,U_2)/P(U_1,U_2)$$

	$\neg U_1 \neg U_2$	$\neg U_1 U_2$	$U_1 \neg U_2$	$U_1 U_2$
$\neg R_2$	0.94	0.30	0.83	0.12
R_2	0.06	0.70	0.17	0.88

Transition model



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

Filtering and Decoding in UmbrellaWorld

- Wow! That was insanely difficult! Why was it so difficult?
- Answer: The select step chose 5 variables that were necessary, so the multiply step needed to construct a table with 32 numbers in it.
- In general:
  - If the select step chooses N variables, each of which has k values, then
  - The multiply step needs to create a table with k^N entries!
  - Complexity is O{k^N}!
- For example: to find $P(R_9|U_1, \ldots, U_9)$
  - Select: there are 19 relevant variables ($R_0, \ldots, R_9, U_1, \ldots, U_9$)
  - …so complexity is 2^19 = 524288

Better Algorithms for HMM Inference

- This can be made much, much more computationally efficient by taking advantage of the structure of the HMM.
- Since each node has only 2 children, the complexity can be reduced from $O\{k^N\}$ to only $O\{k^2\}$.
- The algorithm has two variants: the forward algorithm, and the Viterbi algorithm.
- I’ll tell you the secret on Monday.