CS 581

Advanced Artificial Intelligence

March 20, 2024

Announcements / Reminders

- **Please follow the Week 09 To Do List instructions (if you haven't already)**

- **Programming Assignment #02 due on Sunday (04/07) at 11:59 PM CST**
- **Written Assignment #03 due on Sunday (03/31) at 11:59 PM CST**

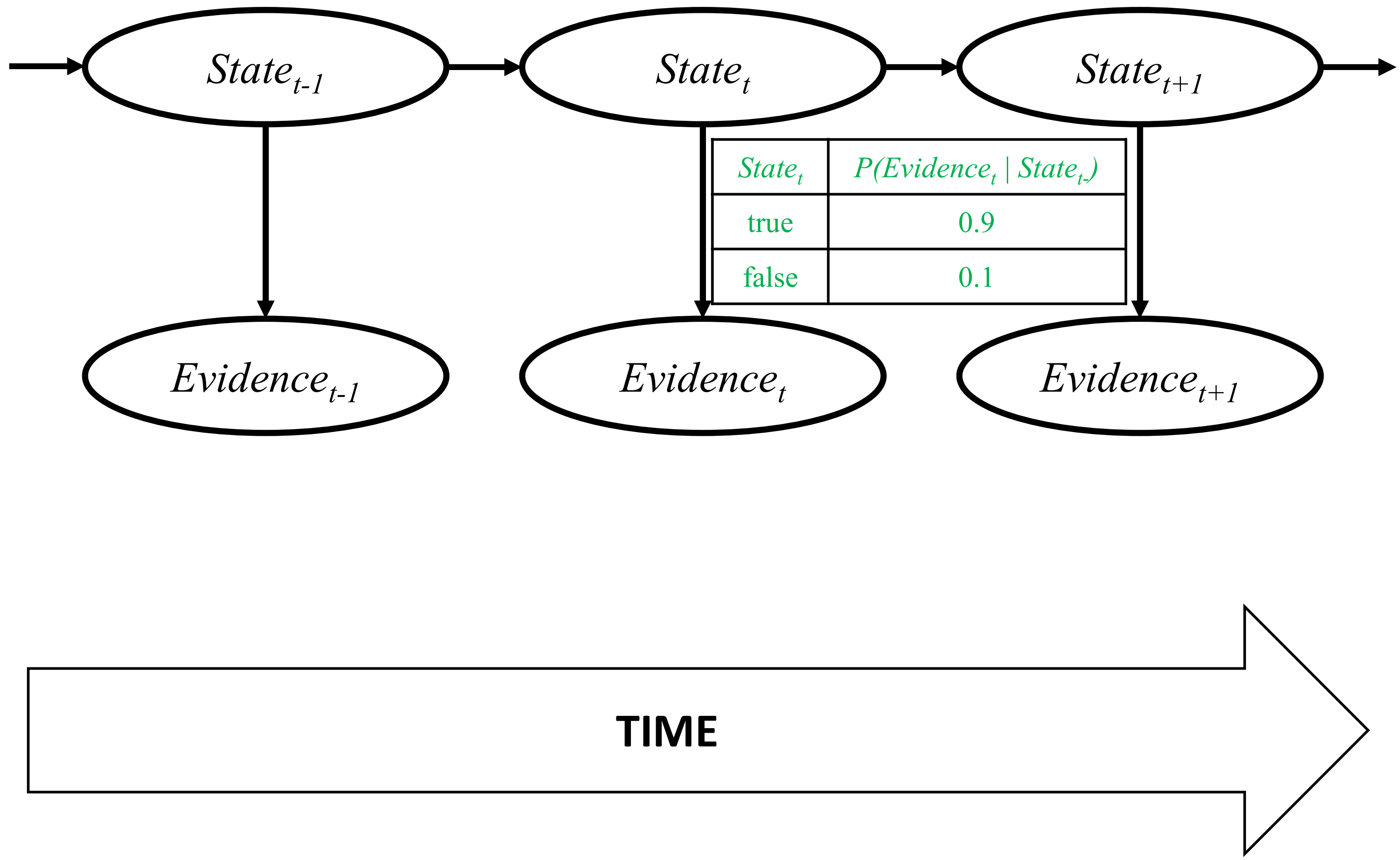
Plan for Today

- **Probabilistic Reasoning over Time**

Inference in Temporal Models

Transition and Sensor Models

$State_{t-1}$	$P(State_{t-1} \mid State_{t-1})$
true	0.7
false	0.3

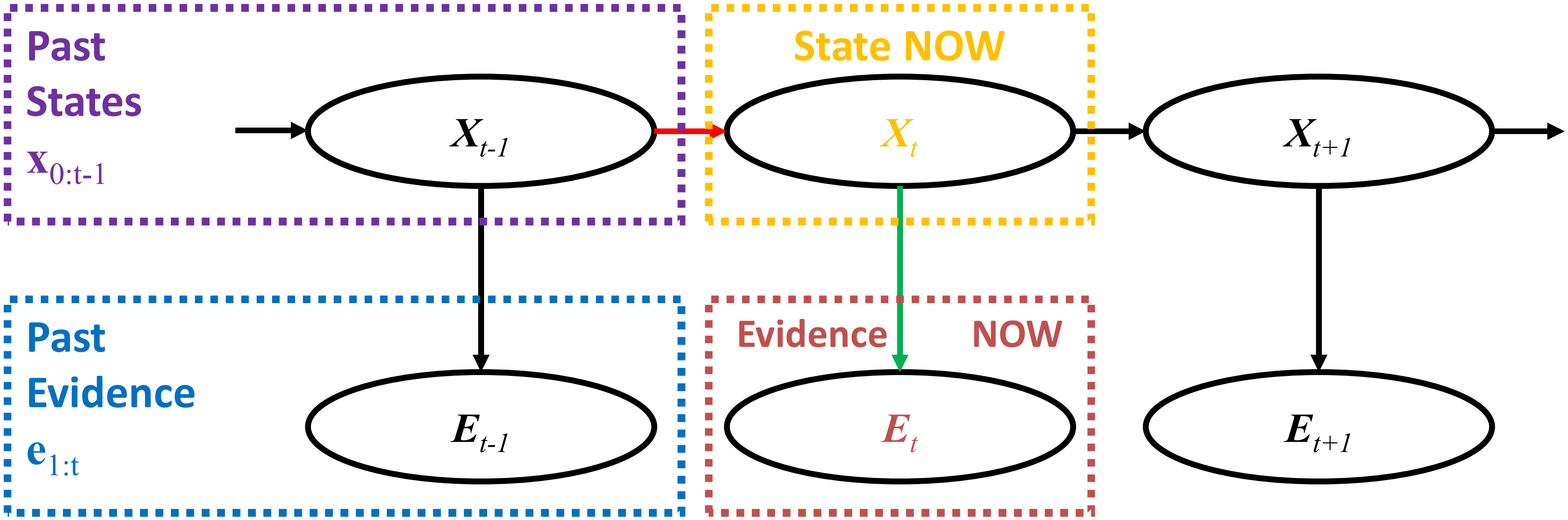


$State_t$	$P(Evidence_t \mid State_t)$
true	0.9
false	0.1

Transition and Sensor Model

The transition model specifies the probability distribution over the latest state variables, given the previous values:

$$P(X_t \mid X_{0:t-1})$$



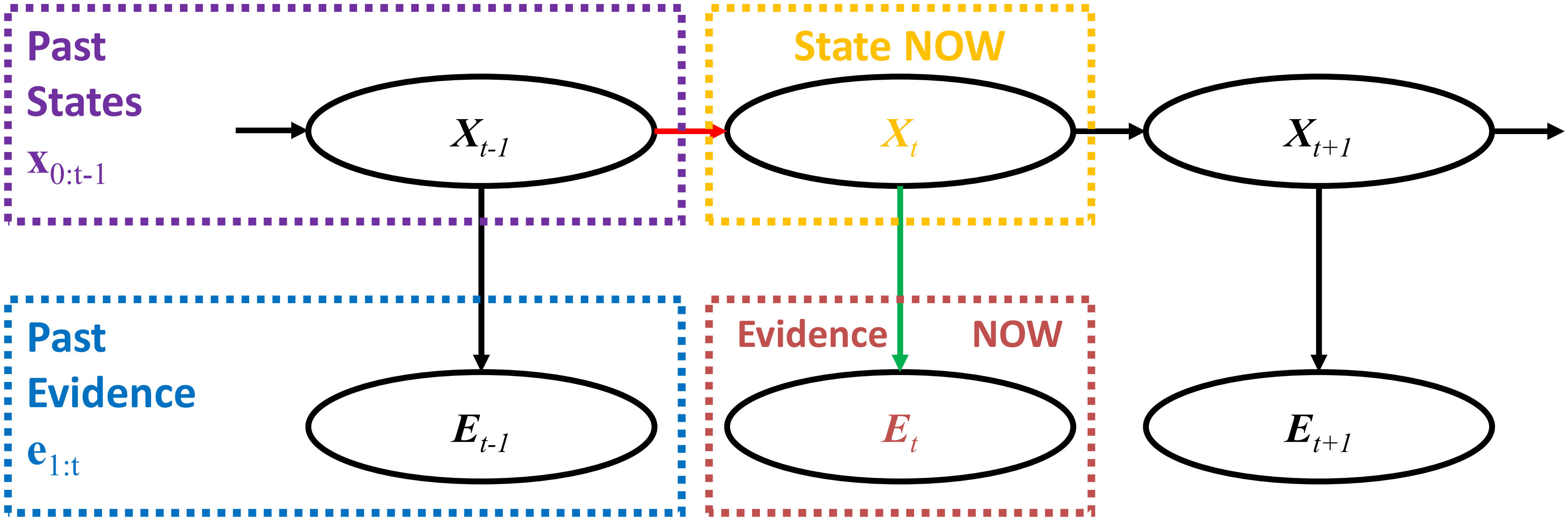
The sensor model: current evidence variables could depend on previous (past) evidence values as well as previous (past) and current state values

$$P(E_t \mid X_{0:t}, E_{1:t-1}) = P(E_t \mid X_{0:t-1}, X_t, E_{1:t-1})$$

Transition and Sensor Model

The **transition model** specifies the probability distribution over the **latest state variables**, given the **previous values**:

$$P(X_t \mid X_{0:t-1})$$



The **sensor model**: current evidence variables could depend on **previous (past) evidence values** as well as **previous (past)** and **current state** values

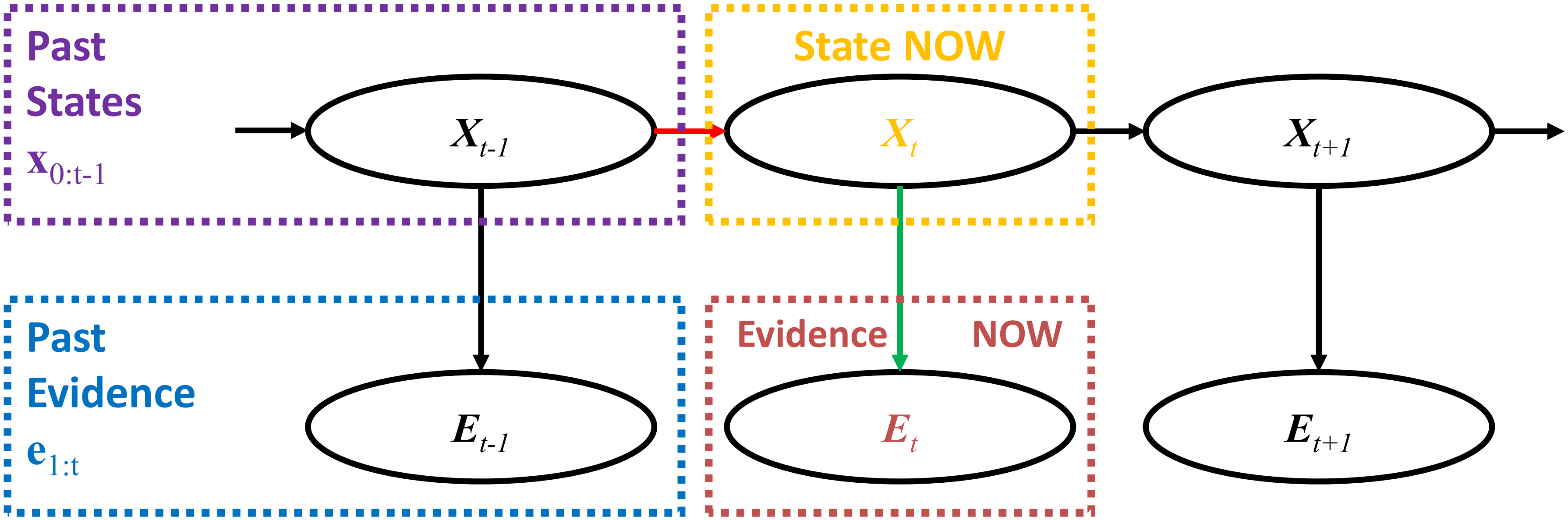
$$P(E_t \mid X_{0:t}, E_{1:t-1}) = P(E_t \mid X_{0:t-1}, X_t, E_{1:t-1})$$

What is the problem here?

Transition and Sensor Model

The transition model specifies the probability distribution over the latest state variables, given the previous values:

$$P(X_t \mid X_{0:t-1})$$



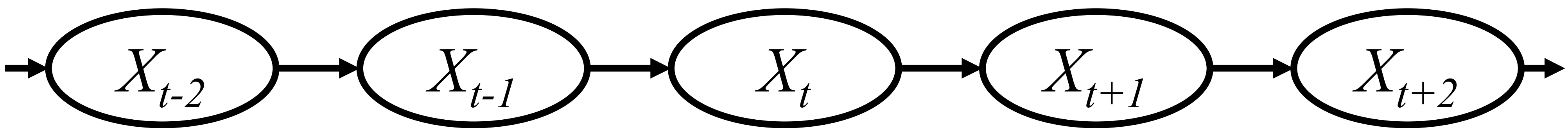
The sensor model: current evidence variables could depend on previous (past) evidence values as well as previous (past) and current state values

$$P(E_t \mid X_{0:t}, E_{1:t-1}) = P(E_t \mid X_{0:t-1}, X_t, E_{1:t-1})$$

Unbounded sets as t grows!

Markov Assumption

Markov Process (Chain) is a random process that generates a sequence of states:



Bayesian Network?? Anyone? Indeed!

$$P(X_{t+1} \mid X_t, X_{t-2}, X_{t-2}) = P(X_t \mid Parents(X_t)) = P(X_{t+1} \mid X_t)$$

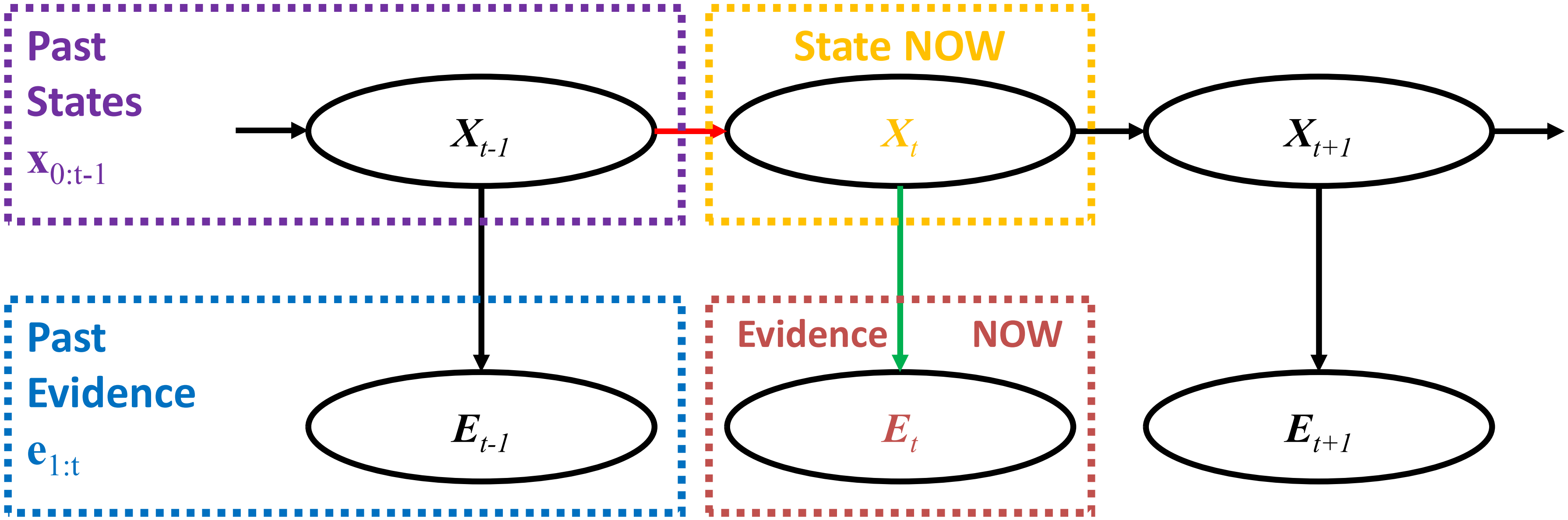
$$\color{red}{P(X_{t+1} \mid X_t, X_{t-2}, X_{t-2}) = P(X_{t+1} \mid X_t)}$$

(First-order) Markov ASSUMPTION

T / S Models /w Markov Assumption

The transition model specifies the probability distribution over the latest state variables, given the previous values:

$$P(X_t \mid X_{0:t-1}) = P(X_t \mid X_{t-1})$$



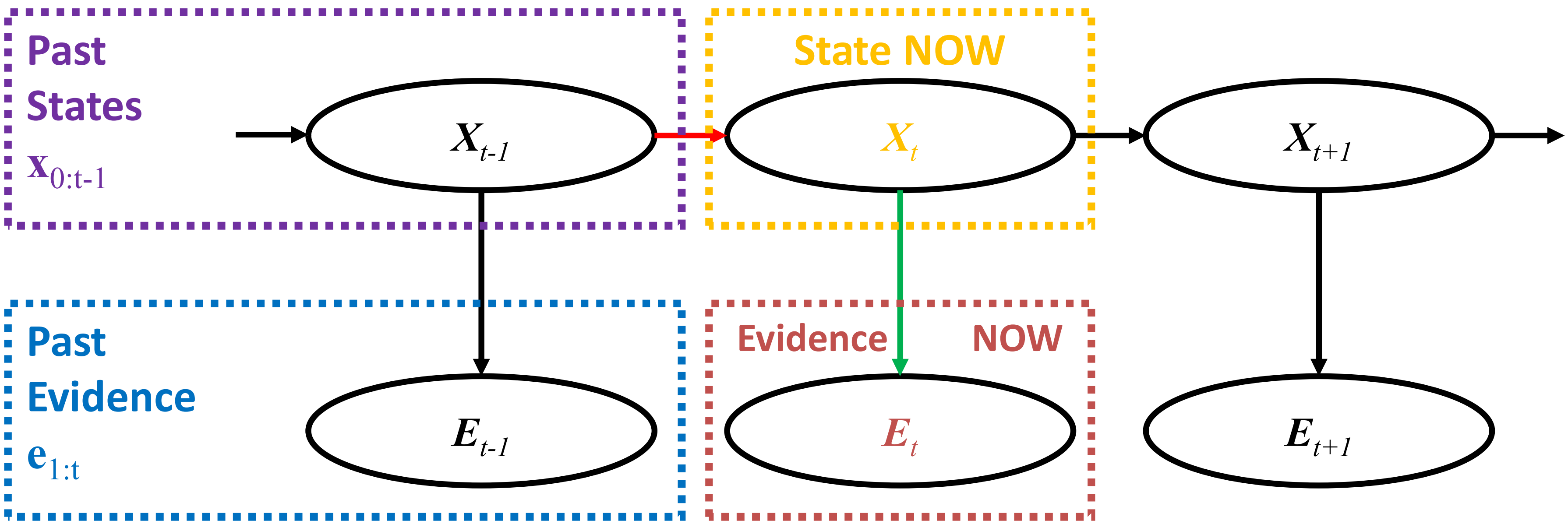
The sensor model: current evidence variables could depend on previous (past) evidence values as well as previous (past) and current state values

$$P(E_t \mid X_{0:t}, E_{1:t-1}) = P(E_t \mid X_t)$$

T / S Models /w Markov Assumption

The transition model specifies the probability distribution over the latest state variables, given the previous values:

$$P(X_t \mid X_{t-1})$$



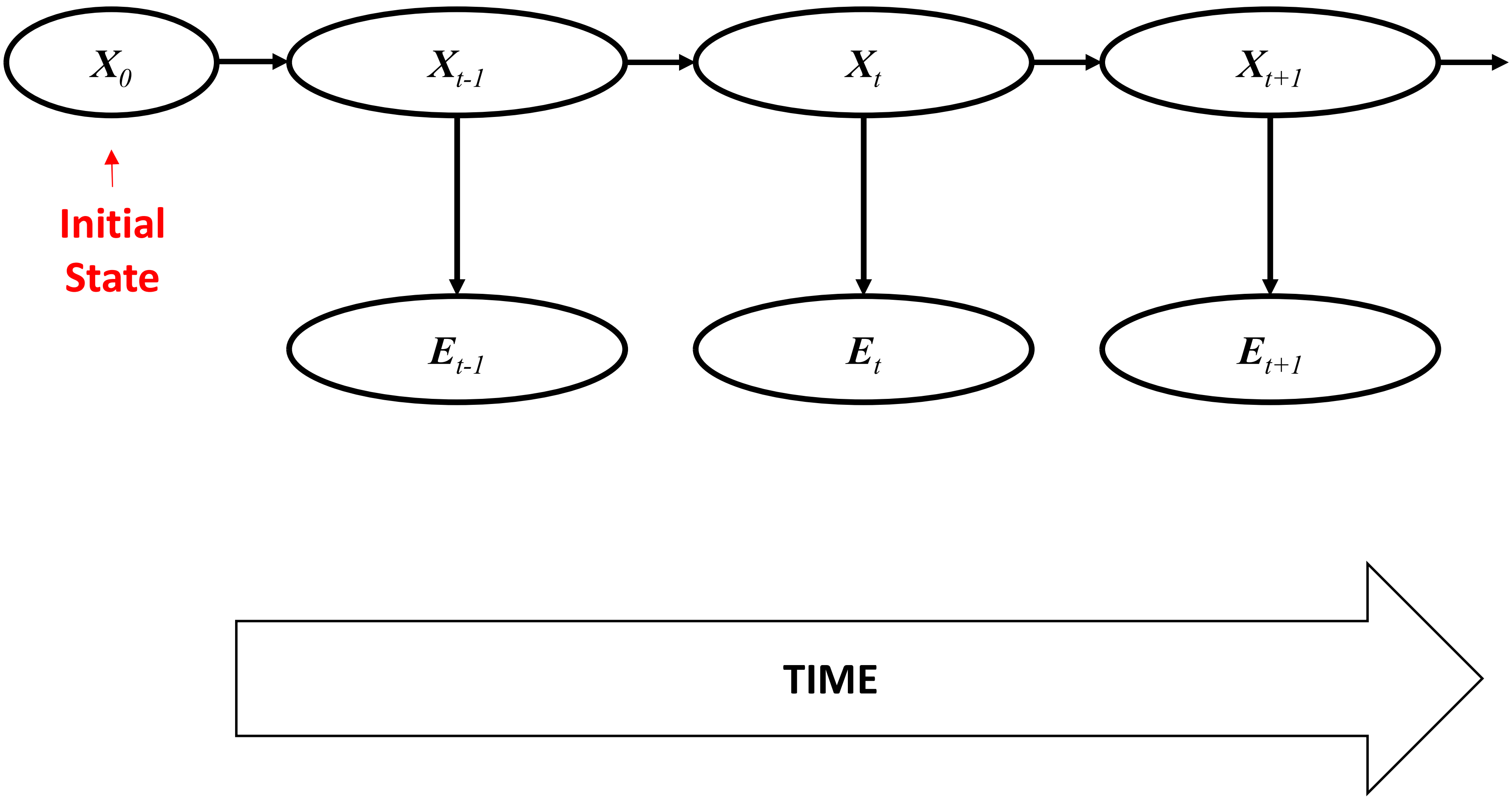
The sensor model: current evidence variables could depend on previous (past) evidence values as well as previous (past) and current state values

$$P(E_t \mid X_t)$$

Complete Joint Distribution

The complete (including initial state distribution | for any t) joint probability distribution for a sequence of transitions and emissions:

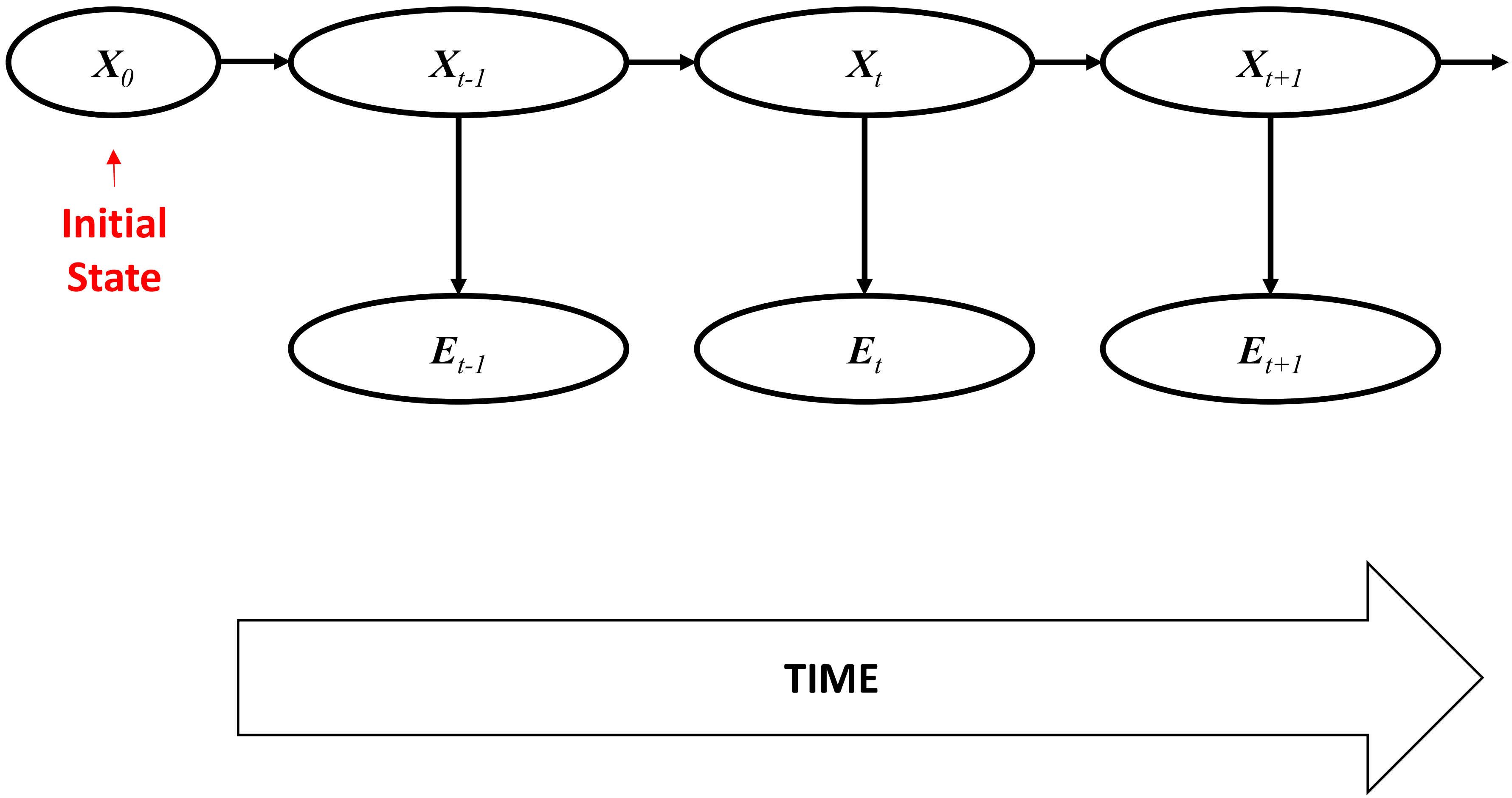
$$\mathbf{P}(\mathbf{X}_{0:t}, \mathbf{E}_{1:t}) = \mathbf{P}(\mathbf{X}_0) \prod_{i=1}^{t} \mathbf{P}(\mathbf{X}_i \mid \mathbf{X}_{i-1}) \, \mathbf{P}(\mathbf{E}_i \mid \mathbf{X}_i)$$



Complete Joint Distribution

The complete (including initial state distribution | for any t) joint probability distribution for a sequence of transitions and emissions:

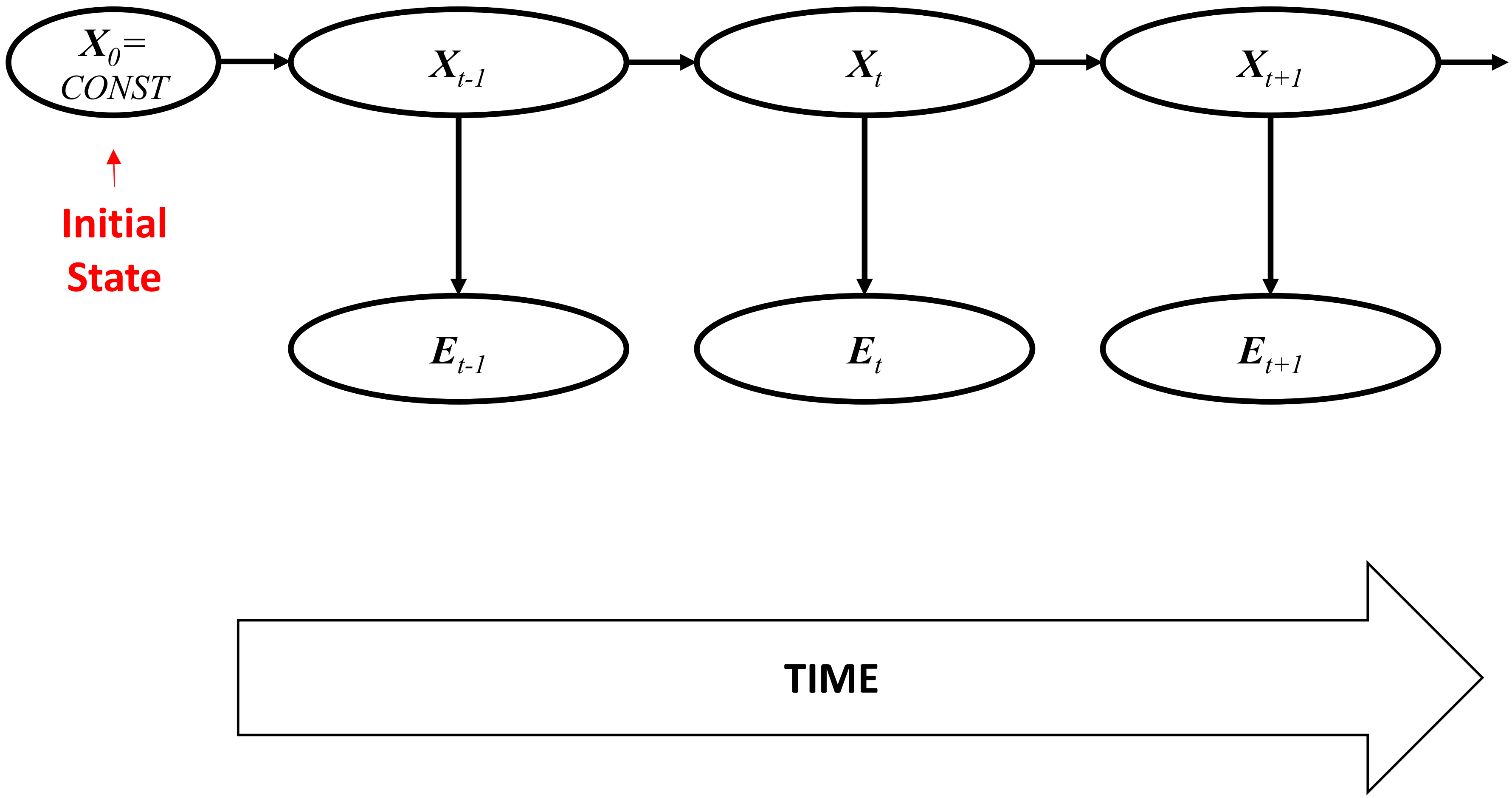
$$\mathbf{P}(\mathbf{X}_{0:t}, \mathbf{E}_{1:t}) = \underbrace{\mathbf{P}(\mathbf{X}_0)}_{\text{Initial}} \prod_{i=1}^{t} \underbrace{\mathbf{P}(\mathbf{X}_i \mid \mathbf{X}_{i-1})}_{\text{Transition}} \underbrace{\mathbf{P}(\mathbf{E}_i \mid \mathbf{X}_i)}_{\text{Emission}}$$



Complete Joint Distribution

The complete (including initial state distribution | for any t) joint probability distribution for a sequence of transitions and emissions:

$$\mathbf{P}(\mathbf{X}_{0:t}, \mathbf{E}_{1:t}) = \prod_{i=1}^{t} \underbrace{\mathbf{P}(\mathbf{X}_i \mid \mathbf{X}_{i-1})}_{\text{Transition}} \underbrace{\mathbf{P}(\mathbf{E}_i \mid \mathbf{X}_i)}_{\text{Emission}}$$



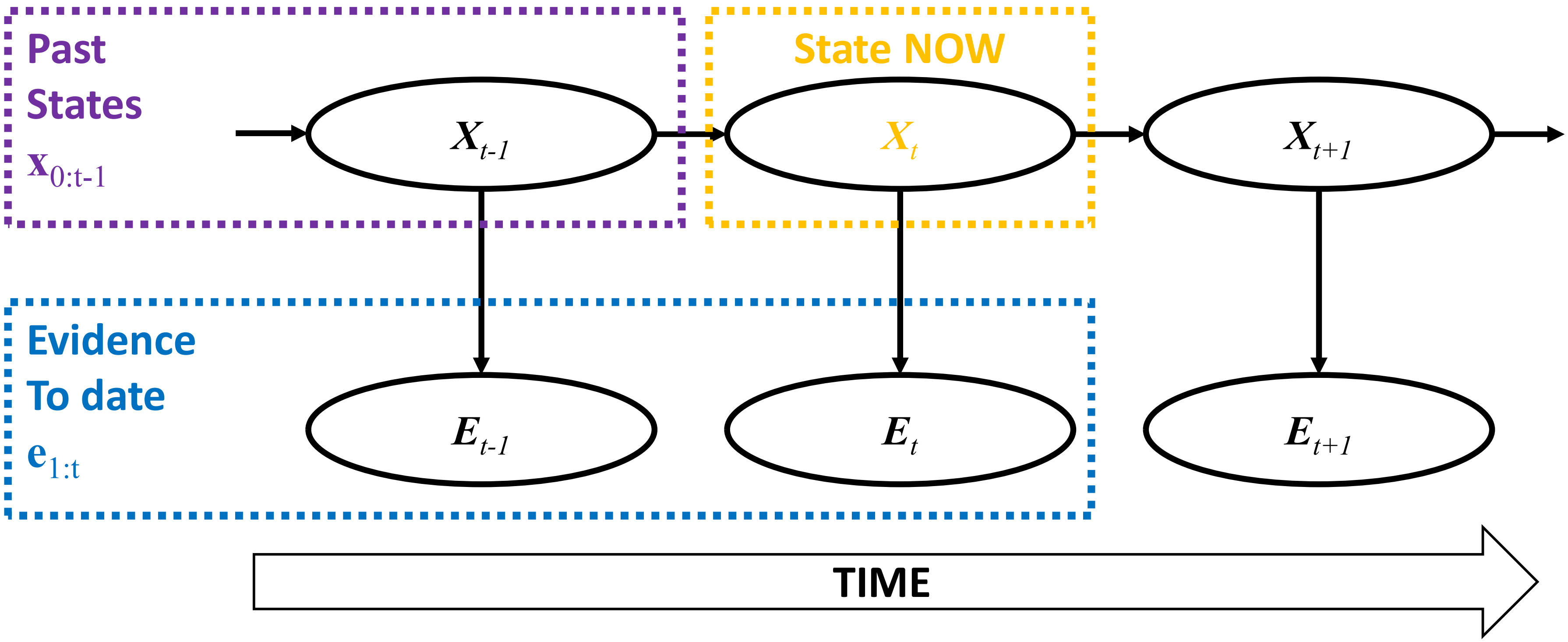
Inference in Temporal Models

- **Filtering / State Estimation**
  - **Current Belief State given Evidence/Percept/Observations so far**
- **Prediction**
  - **Future Belief State given Evidence/Percept /Observations so far**
- **Smoothing**
  - **Past Belief State given Evidence/Percept/Observations so far**
- **Most likely explanation:**
  - **Use sequence of observations to find sequence of states that generated them**
- **Learning:**
  - **Learn the transition and sensor models based on observations ("emissions")**

Inference: Filtering

This is the task of computing the belief state — the posterior distribution over the most recent state — given all evidence to date. Filtering is also called STATE ESTIMATION.

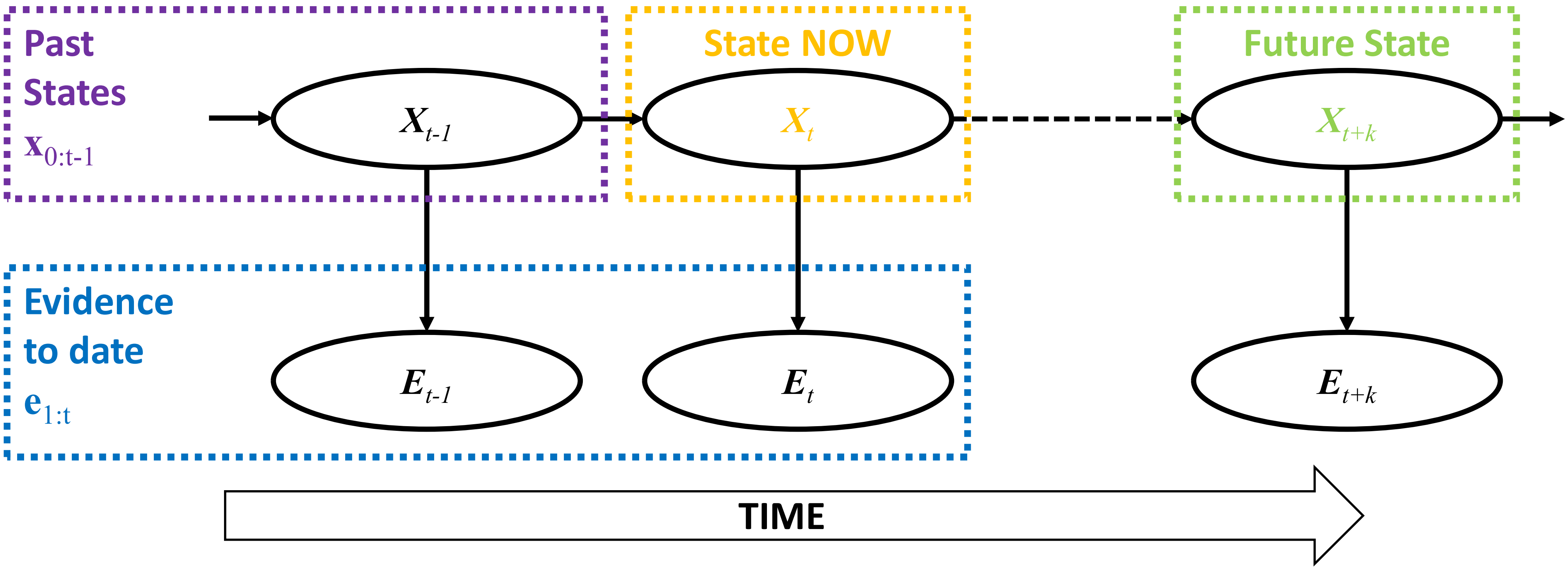
$$P(X_t \mid \mathbf{e}_{1:t})$$



Inference: Prediction

This is he task of computing the posterior distribution over the future state (time $t+k$, for some $k > 0$), given all evidence to date. Useful for evaluating possible courses of action.

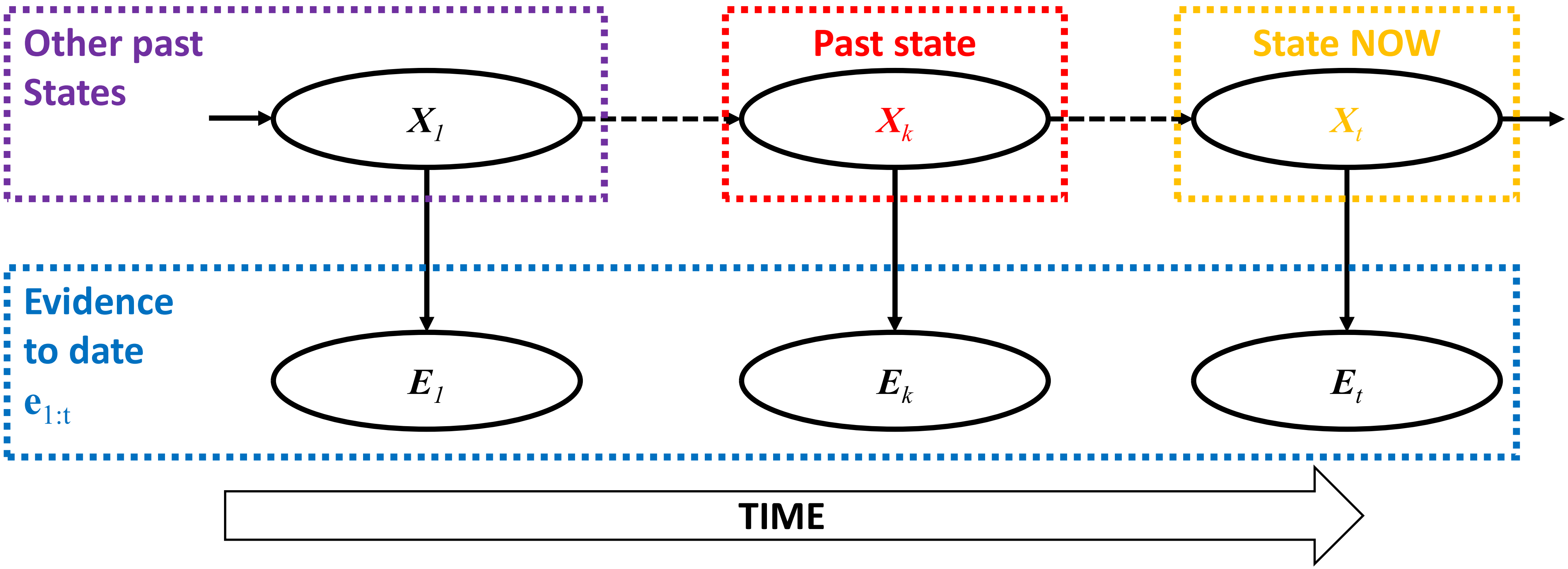
$$P(X_{t+k} \mid \mathbf{e}_{1:t})$$



Inference: Smoothing

This is the task of computing the posterior distribution over the past state (time k, for some $0 \leq k < t$), given all evidence to date. Provides a better state estimate of, because it incorporates more evidence.

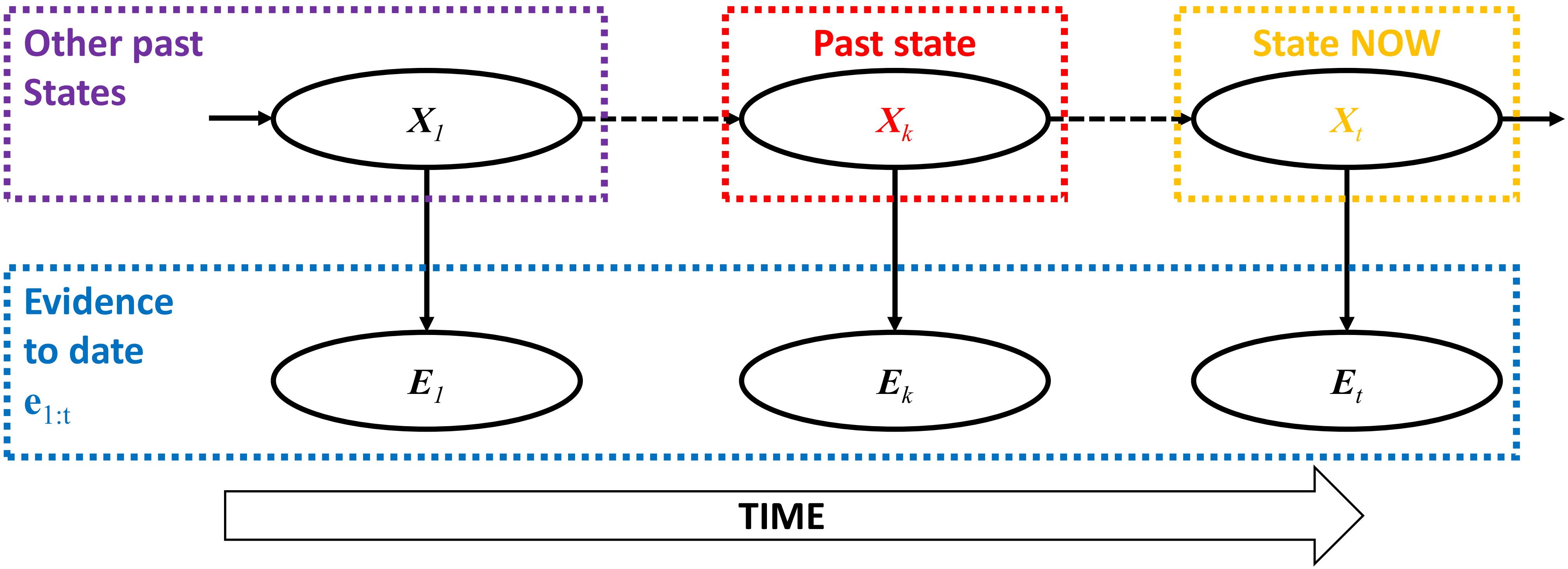
$$P(\mathbf{X}_k \mid \mathbf{e}_{1:t})$$



Inference: Smoothing

This is the task of computing the posterior distribution over the past state (time k, for some $0 \leq k < t$), given all evidence to date. Provides a better state estimate of, because it incorporates more evidence.

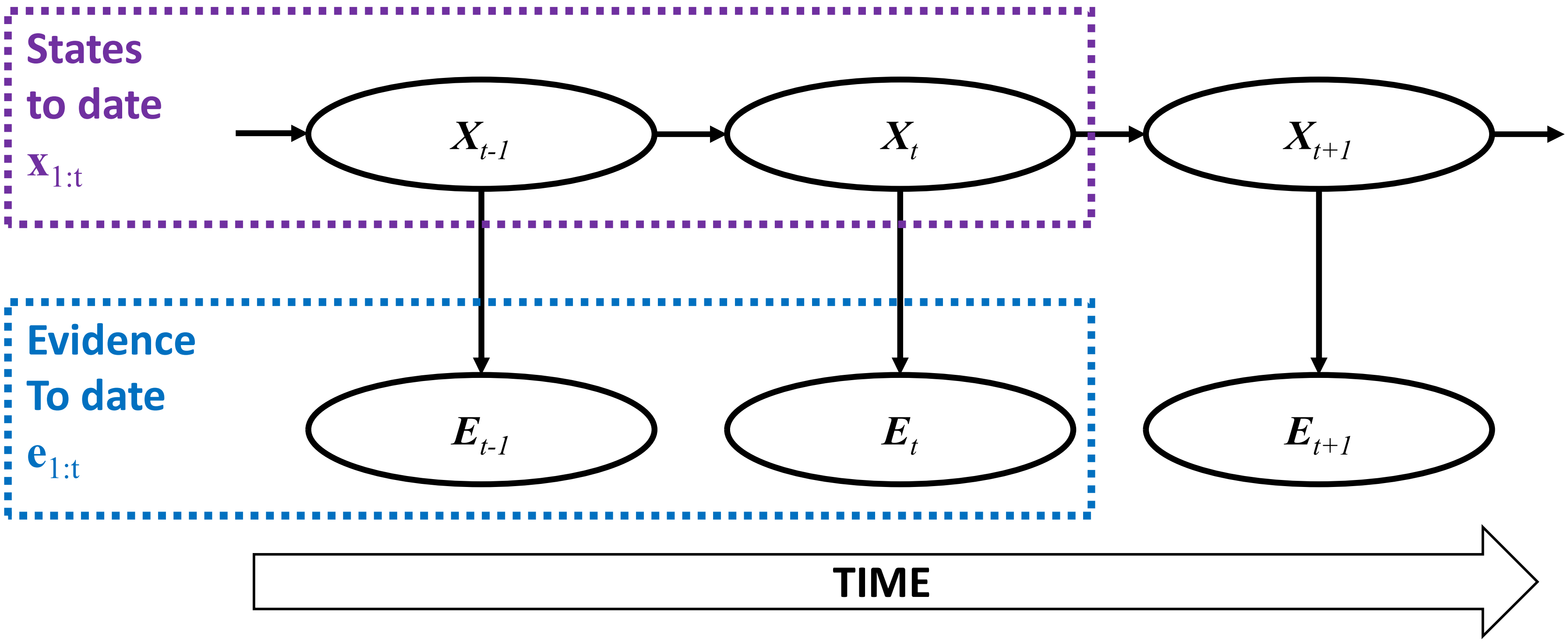
$$P(X_k \mid e_{1:t})$$



Inference: Most Likely Explanation

Given a sequence of observations, we might wish to find the sequence of states that is most likely to have generated those observations.

$$argmax_{\mathbf{x}_{1:t}} P(\mathbf{x}_{1:t} \mid \mathbf{e}_{1:t})$$

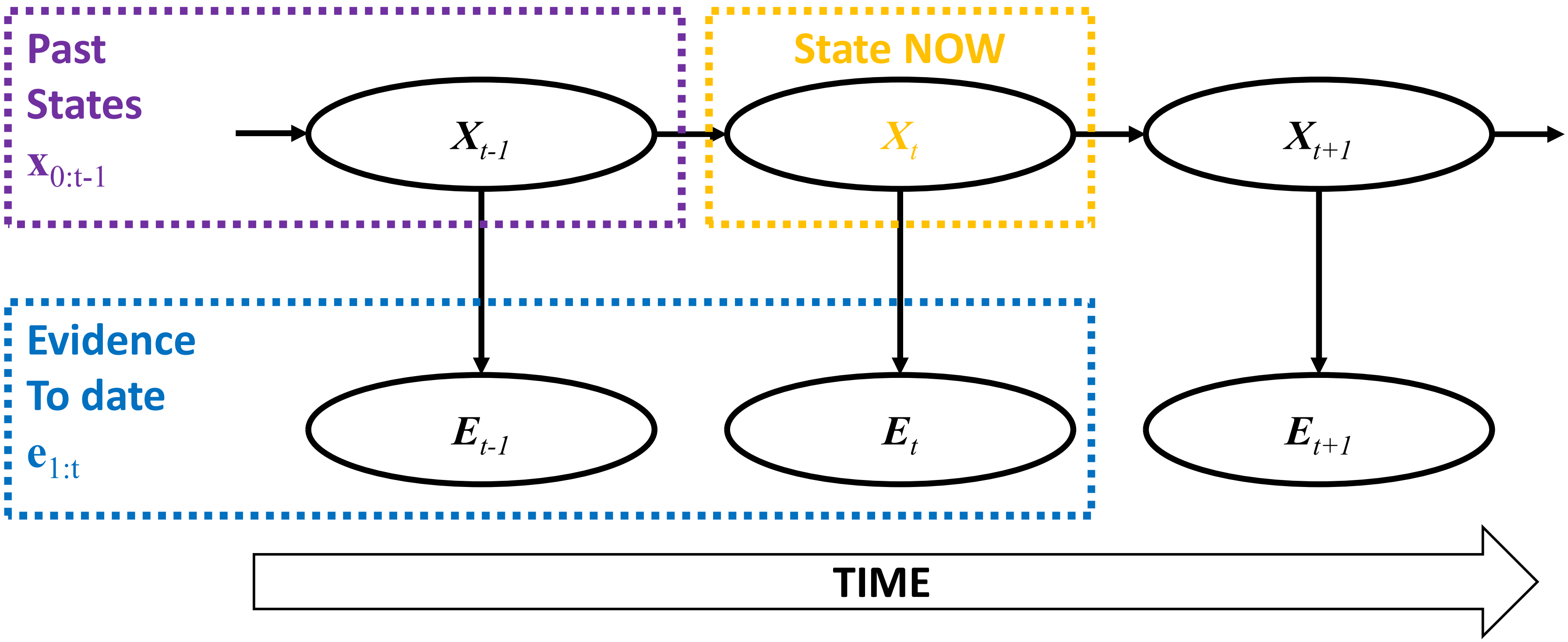


Filtering

Inference: Filtering

This is the task of computing the belief state — the posterior distribution over the most recent state — given all evidence to date. Filtering is also called STATE ESTIMATION.

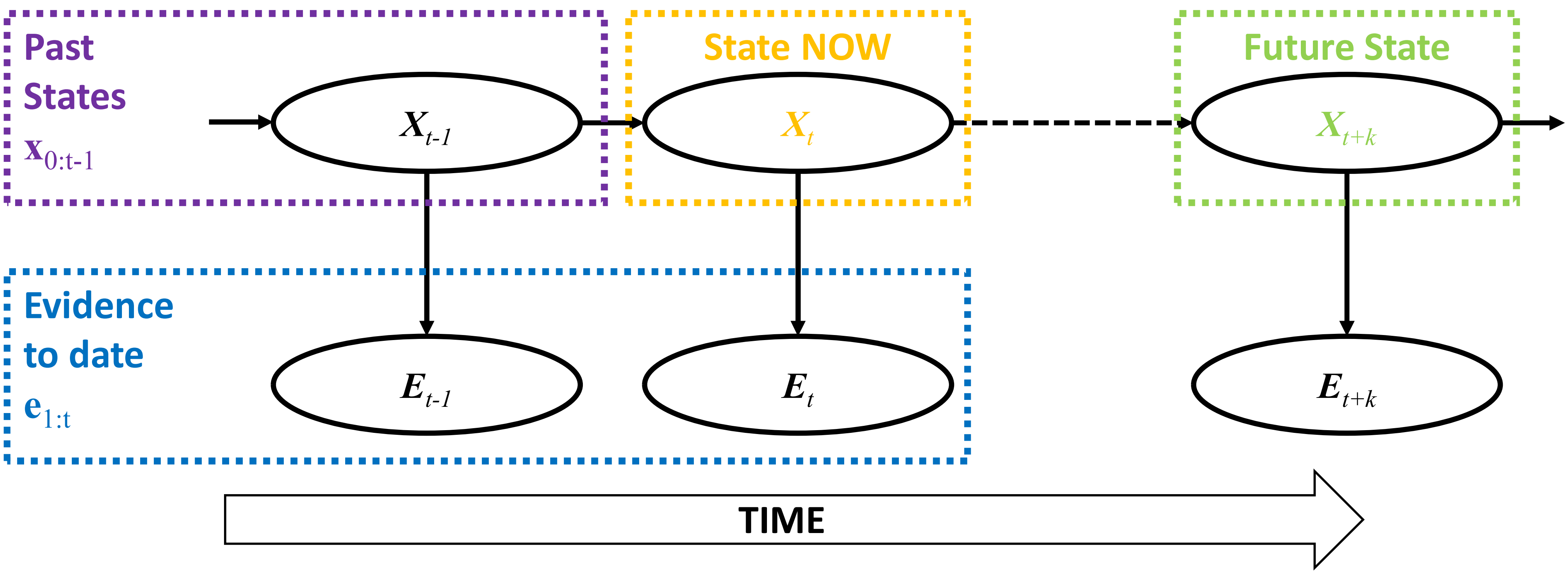
$$P(X_t \mid \mathbf{e}_{1:t})$$



Inference: Prediction

This is he task of computing the posterior distribution over the future state (time $t+k$, for some $k > 0$), given all evidence to date. Useful for evaluating possible courses of action.

$$P(\mathbf{X}_{t+k} \mid \mathbf{e}_{1:t})$$

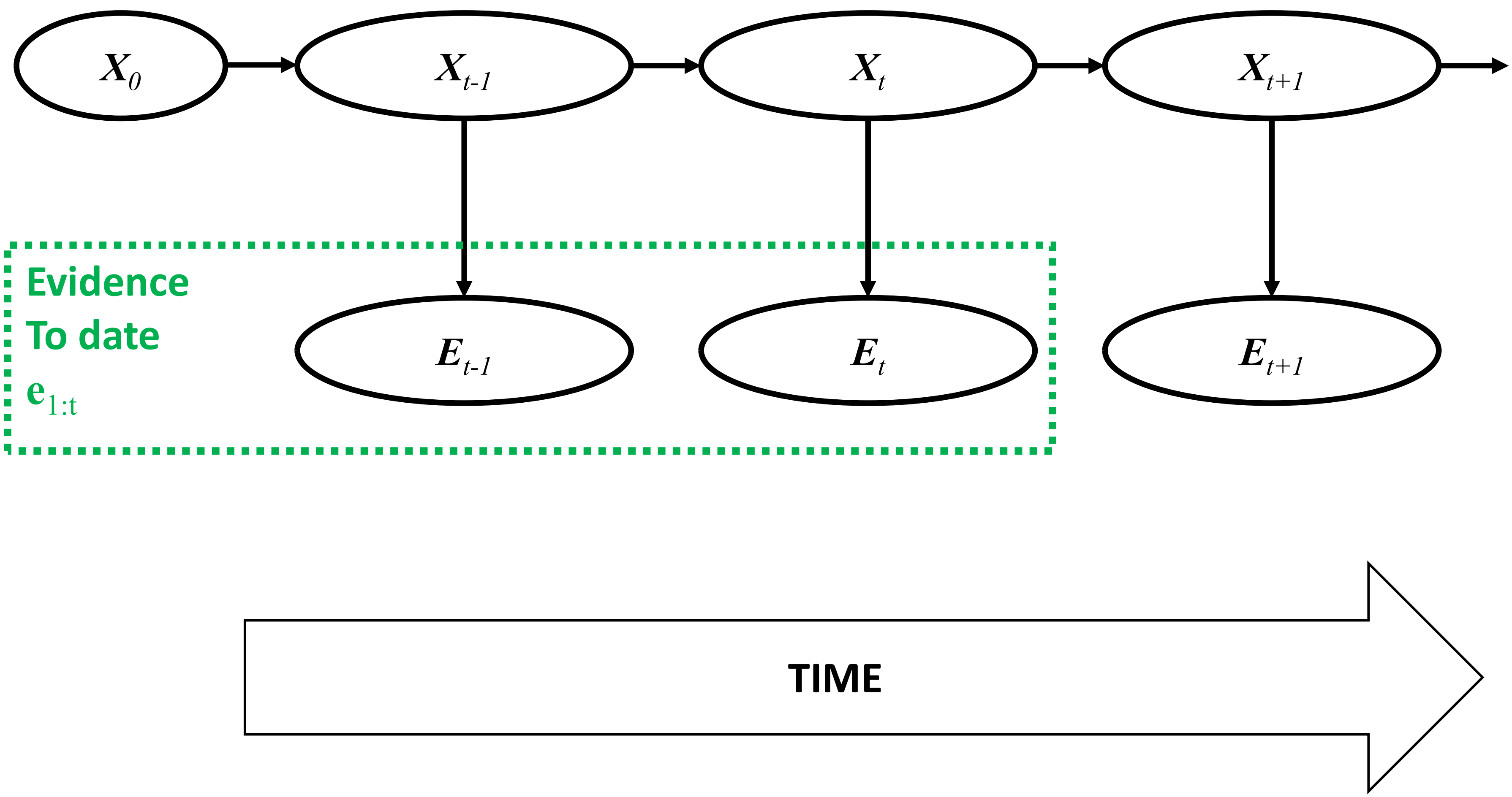


What Is Known?

The complete (including initial state distribution | for any t) joint probability distribution for a sequence of transitions and emissions:

Initial State Distribution

$$\mathbf{P}(\mathbf{X}_{0:t}, \mathbf{E}_{1:t}) = \underbrace{\mathbf{P}(\mathbf{X}_0)}_{\text{Initial}} \prod_{i=1}^{t} \underbrace{\mathbf{P}(\mathbf{X}_i \mid \mathbf{X}_{i-1})}_{\text{Transition}} \underbrace{\mathbf{P}(\mathbf{E}_i \mid \mathbf{X}_i)}_{\text{Emission}}$$



Filtering / Recursive Estimation

For filtering (state estimation) we are interested in calculating:

$$P(X_t \mid e_{1:t})$$

Given all the evidence/observation to date ($e_{1:t}$), what is the state X_t? What about estimation for the following time slice?

$$P(X_{t+1} \mid e_{1:t+1}) = ???$$

Ideally, we would like to take advantage of what we already have and UPDATE:

$$P(X_{t+1} \mid e_{1:t+1}) = f(e_{t+1}, P(X_t \mid e_{1:t}))$$

for some function f. This is called recursive estimation.

Filtering / Recursive Estimation

Let's take a look at this recursive relationship again:

$$\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t+1}) = f(\boldsymbol{e}_{t+1}, \boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{e}_{1:t}))$$

Recursive function arguments:

- $\boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{e}_{1:t}))$ **is the current (at time t) state probability distribution obtained once we have all the evidence/observation to date ($\boldsymbol{e}_{1:t}$)**
- $\boldsymbol{e}_{t+1}$ **is the observation for next (at time $t+1$) state X_{t+1}**

In other words: if we know current state (at time $t+1$) probability distribution AND next (at time $t+1$) observation -> next state probability distribution.

How do we calculate it, though?

Filtering / Recursive Estimation

Evidence can be separated: $e_{1:t+1} = e_{1:t}, e_{t+1}$

$$P(X_{t+1} \mid e_{1:t+1}) = P(X_{t+1} \mid e_{1:t}, e_{t+1})$$

Applying Bayes Rule yields:

Normalizing constant

$$P(X_{t+1} \mid e_{1:t}, e_{t+1}) = \alpha * P(e_{t+1} \mid X_{t+1}, e_{1:t}) * P(X_{t+1} \mid e_{1:t})$$

How did we get there?

Conditional probability: $P(A \mid B) = \frac{P(A,B)}{P(B)}$

$$P(X_{t+1} \mid e_{1:t}, e_{t+1}) = \frac{P(X_{t+1}, e_{1:t}, e_{t+1})}{P(e_{1:t}, e_{t+1})}$$

Ordering can be changed: $P(A,B) = P(B,A)$

$$P(X_{t+1} \mid e_{1:t}, e_{t+1}) = \frac{P(e_{1:t}, X_{t+1}, e_{t+1})}{P(e_{t+1}, e_{1:t})}$$

Filtering / Recursive Estimation

Evidence can be separated: $e_{1:t+1} = e_{1:t}, e_{t+1}$

$$P(X_{t+1} \mid e_{1:t+1}) = P(X_{t+1} \mid e_{1:t}, e_{t+1})$$

Applying Bayes Rule yields:

$$P(X_{t+1} \mid e_{1:t}, e_{t+1}) = \alpha * P(e_{t+1} \mid X_{t+1}, e_{1:t}) * P(X_{t+1} \mid e_{1:t})$$

How did we get there?

$$P(X_{t+1} \mid e_{1:t}, e_{t+1}) = \frac{P(e_{1:t}) * P(X_{t+1} \mid e_{1:t}) * P(e_{t+1} \mid X_{t+1}, e_{1:t})}{P(e_{t+1}, e_{1:t})}$$

by Chain Rule

$$P(X_{t+1} \mid e_{1:t}, e_{t+1}) = \frac{P(e_{1:t}) * P(X_{t+1} \mid e_{1:t}) * P(e_{t+1} \mid X_{t+1}, e_{1:t})}{P(e_{t+1}, e_{1:t})}$$

Filtering / Recursive Estimation

Evidence can be separated: $e_{1:t+1} = e_{1:t}, e_{t+1}$

$$P(X_{t+1} \mid e_{1:t+1}) = P(X_{t+1} \mid e_{1:t}, e_{t+1})$$

Applying Bayes Rule yields:

$$P(X_{t+1} \mid e_{1:t}, e_{t+1}) = \alpha * P(e_{t+1} \mid X_{t+1}, e_{1:t}) * P(X_{t+1} \mid e_{1:t})$$

How did we get there?

$$P(X_{t+1} \mid e_{1:t}, e_{t+1}) = \frac{P(e_{1:t},\ X_{t+1},\ e_{t+1})}{P(e_{t+1}, e_{1:t})}$$

by Product Rule

$$P(X_{t+1} \mid e_{1:t}, e_{t+1}) = \frac{P(e_{1:t}) * P(X_{t+1} \mid e_{1:t}) * P(e_{t+1} \mid X_{t+1}, e_{1:t})}{P(e_{t+1} \mid e_{1:t}) * P(e_{1:t})}$$

Filtering / Recursive Estimation

Evidence can be separated: $\boldsymbol{e}_{1:t+1} = \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1}$

$$\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t+1}) = \boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1})$$

Applying Bayes Rule yields:

$$\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1}) = \alpha * \boldsymbol{P}(\boldsymbol{e}_{t+1} \mid \boldsymbol{X}_{t+1}, \boldsymbol{e}_{1:t}) * \boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t})$$

How did we get there?

Cancel terms

$$\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1}) = \frac{\cancel{\boldsymbol{P}(\boldsymbol{e}_{1:t})} * \boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}) * \boldsymbol{P}(\boldsymbol{e}_{t+1} \mid \boldsymbol{X}_{t+1}, \boldsymbol{e}_{1:t})}{\boldsymbol{P}(\boldsymbol{e}_{t+1} \mid \boldsymbol{e}_{1:t}) * \cancel{\boldsymbol{P}(\boldsymbol{e}_{1:t})}}$$

$$\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1}) = \frac{\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}) * \boldsymbol{P}(\boldsymbol{e}_{t+1} \mid \boldsymbol{X}_{t+1}, \boldsymbol{e}_{1:t})}{\boldsymbol{P}(\boldsymbol{e}_{t+1} \mid \boldsymbol{e}_{1:t})}$$

Filtering / Recursive Estimation

Evidence can be separated: $e_{1:t+1} = e_{1:t}, e_{t+1}$

$$P(X_{t+1} \mid e_{1:t+1}) = P(X_{t+1} \mid e_{1:t}, e_{t+1})$$

Applying Bayes Rule yields:

$$P(X_{t+1} \mid e_{1:t}, e_{t+1}) = \alpha * P(e_{t+1} \mid X_{t+1}, e_{1:t}) * P(X_{t+1} \mid e_{1:t})$$

How did we get there?

We don't really need to know this

$$P(X_{t+1} \mid e_{1:t}, e_{t+1}) = \frac{P(X_{t+1} \mid e_{1:t}) * P(e_{t+1} \mid X_{t+1}, e_{1:t})}{P(e_{t+1} \mid e_{1:t})}$$

We can always normalize later

$$P(X_{t+1} \mid e_{1:t}, e_{t+1}) = \alpha * P(X_{t+1} \mid e_{1:t}) * P(e_{t+1} \mid X_{t+1}, e_{1:t})$$

Filtering / Recursive Estimation

Evidence can be separated: $\boldsymbol{e}_{1:t+1} = \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1}$

$$P(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t+1}) = P(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1})$$

Applying Bayes Rule yields:

$$P(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1}) = \alpha * P(\boldsymbol{e}_{t+1} \mid \boldsymbol{X}_{t+1}, \boldsymbol{e}_{1:t}) * P(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t})$$

Recall the Markov assumption for the sensor model (emission):

$$P(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1}) = \alpha * \underbrace{P(\boldsymbol{e}_{t+1} \mid \boldsymbol{X}_{t+1})}_{\text{update}} * \underbrace{P(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t})}_{\text{prediction}}$$

$$P(X_{t+k} \mid \mathbf{e}_{1:t})$$

Prediction

Filtering / Recursive Estimation

Evidence can be separated: $\boldsymbol{e}_{1:t+1} = \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1}$

$$\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t+1}) = \boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1})$$

Applying Bayes Rule yields:

$$\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1}) = \alpha * \boldsymbol{P}(\boldsymbol{e}_{t+1} \mid \boldsymbol{X}_{t+1}, \boldsymbol{e}_{1:t}) * \boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t})$$

Recall the Markov assumption for the <u>sensor</u> model (emission):

$$\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t+1}) = \alpha * \boldsymbol{P}(\boldsymbol{e}_{t+1} \mid \boldsymbol{X}_{t+1}) * \boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}) =$$

$$= \alpha * \boldsymbol{P}(\boldsymbol{e}_{t+1} \mid \boldsymbol{X}_{t+1}) * \sum_{\boldsymbol{x}_t} \boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{x}_t, \boldsymbol{e}_{1:t}) * \boldsymbol{P}(\boldsymbol{x}_t \mid \boldsymbol{e}_{1:t})$$

Recall the Markov assumption for the <u>transition</u> model:

$$\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t+1}) = \alpha * \boldsymbol{P}(\boldsymbol{e}_{t+1} \mid \boldsymbol{X}_{t+1}) * \sum_{\boldsymbol{x}_t} \boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{x}_t) * \boldsymbol{P}(\boldsymbol{x}_t \mid \boldsymbol{e}_{1:t})$$

Filtering: State Estimate

State estimate equation:

$$P(X_{t+1} \mid e_{1:t+1}) = \alpha * \underbrace{\underbrace{P(e_{t+1} \mid X_{t+1})}_{\text{Emission / sensor model}} * \sum_{x_t} \underbrace{P(X_{t+1} \mid x_t)}_{\text{Transition model}}}_{\text{KNOWN model of the World / Environment}} * \underbrace{\underbrace{P(x_t \mid e_{1:t})}_{\text{Recursion}}}_{\text{KNOWN Previous State Estimate}}$$

This is a recursive (estimation) relationship:

$$P(X_{t+1} \mid e_{1:t+1}) = f(e_{t+1}, P(X_t \mid e_{1:t}))$$

New state estimate depends only (can be updated based) on previous state estimate and new observation.

Filtering: State Estimate

State estimate equation:

$$\boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{e}_{1:t}) = \alpha * \underbrace{\underbrace{\boldsymbol{P}(\boldsymbol{e}_t \mid \boldsymbol{X}_t)}_{\text{Emission model}} * \sum_{\boldsymbol{x}_{t-1}} \underbrace{\boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{x}_{t-1})}_{\text{Transition model}}}_{\text{KNOWN model of the World / Environment}} * \underbrace{\underbrace{\boldsymbol{P}(\boldsymbol{x}_{t-1} \mid \boldsymbol{e}_{1:t-1})}_{\text{Recursion}}}_{\text{KNOWN Previous State Estimate}}$$

This is a recursive (estimation) relationship:

$$\boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{e}_{1:t}) = f(\boldsymbol{e}_t, \boldsymbol{P}(\boldsymbol{X}_{t-1} \mid \boldsymbol{e}_{1:t-1}))$$

New state estimate depends only (can be updated based) on previous state estimate and new observation.

Filtered Estimate as a "Message"

Filtered estimate

$$\boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{e}_{1:t})$$

can be thought of as a message $f_{1:t}$ propagated forward along the sequence, modified by each transition and updated by each new observation. The process is given by:

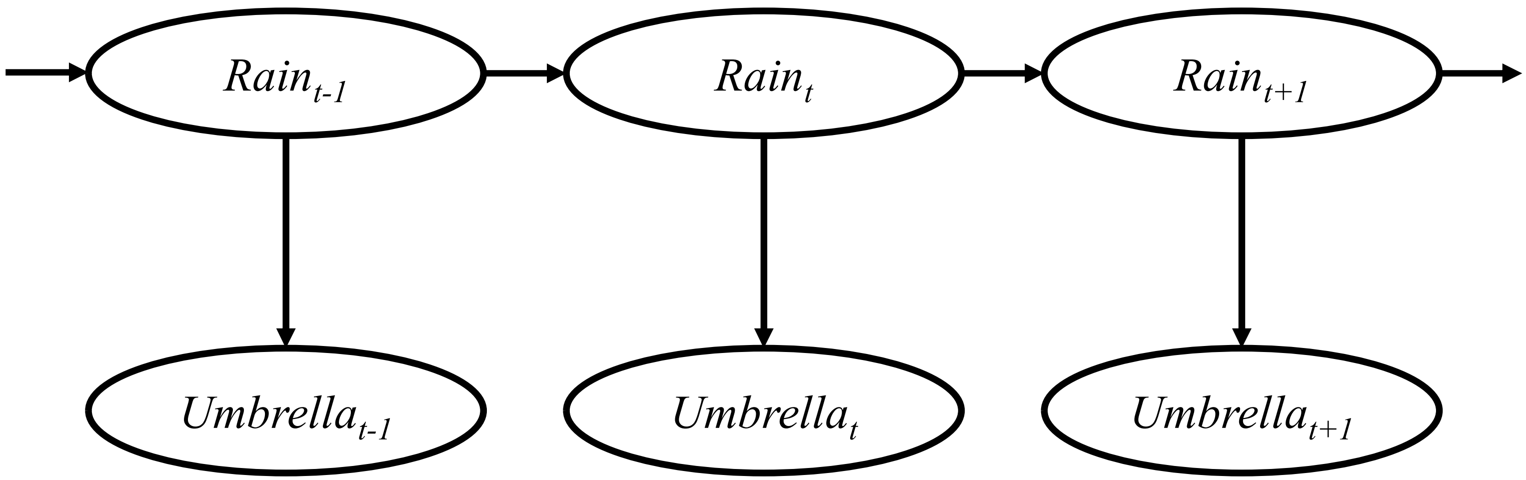
$$\boldsymbol{f}_{1:t} = FORWARD(\boldsymbol{f}_{1:t-1}, \boldsymbol{e}_t)$$

$$\textbf{where: } \boldsymbol{f}_{1:0} = \boldsymbol{P}(\boldsymbol{X}_0)$$

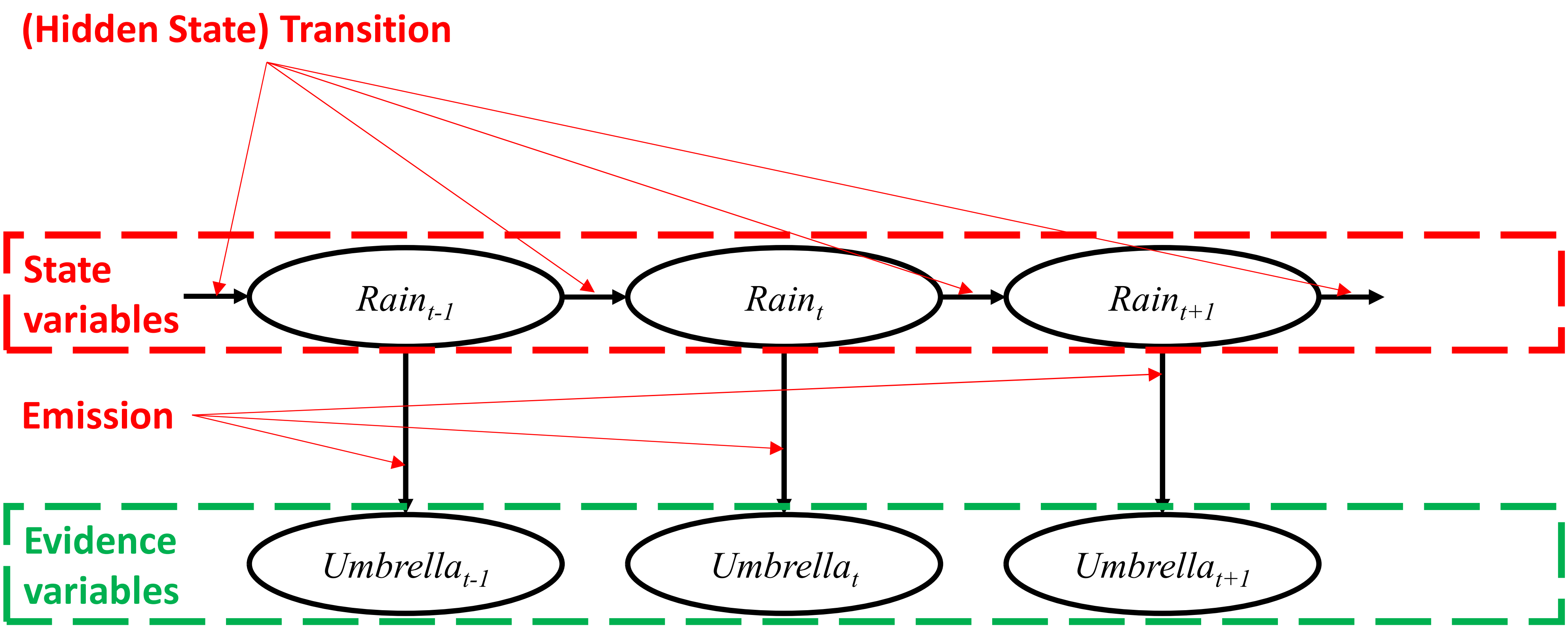
$$\boldsymbol{f}_{1:t} = \alpha * \boldsymbol{P}(\boldsymbol{e}_t \mid \boldsymbol{X}_t) * \sum_{\boldsymbol{x}_{t-1}} \boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{x}_{t-1}) * \boldsymbol{P}(\boldsymbol{x}_{t-1} \mid \boldsymbol{e}_{1:t-1})$$

With all state variables discrete: update time is constant (independent of t) and the space required is constant as well.

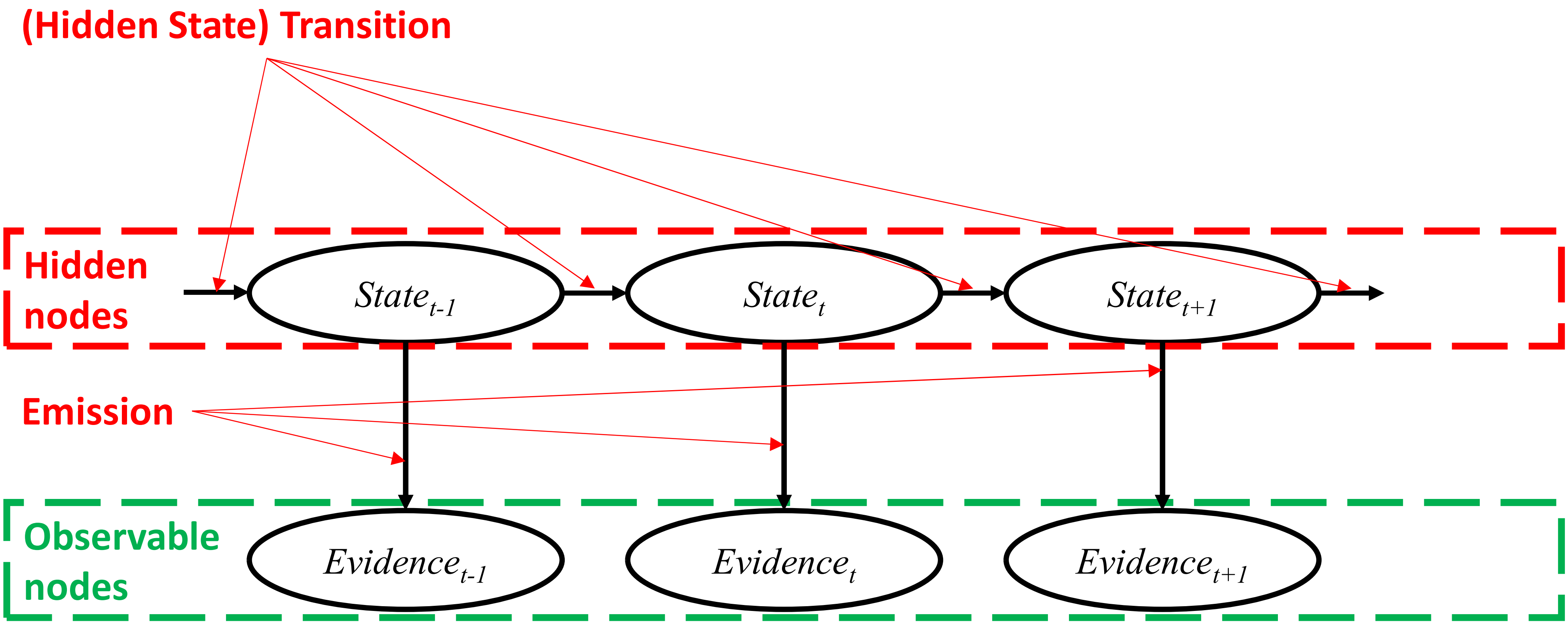
Filtering: Example



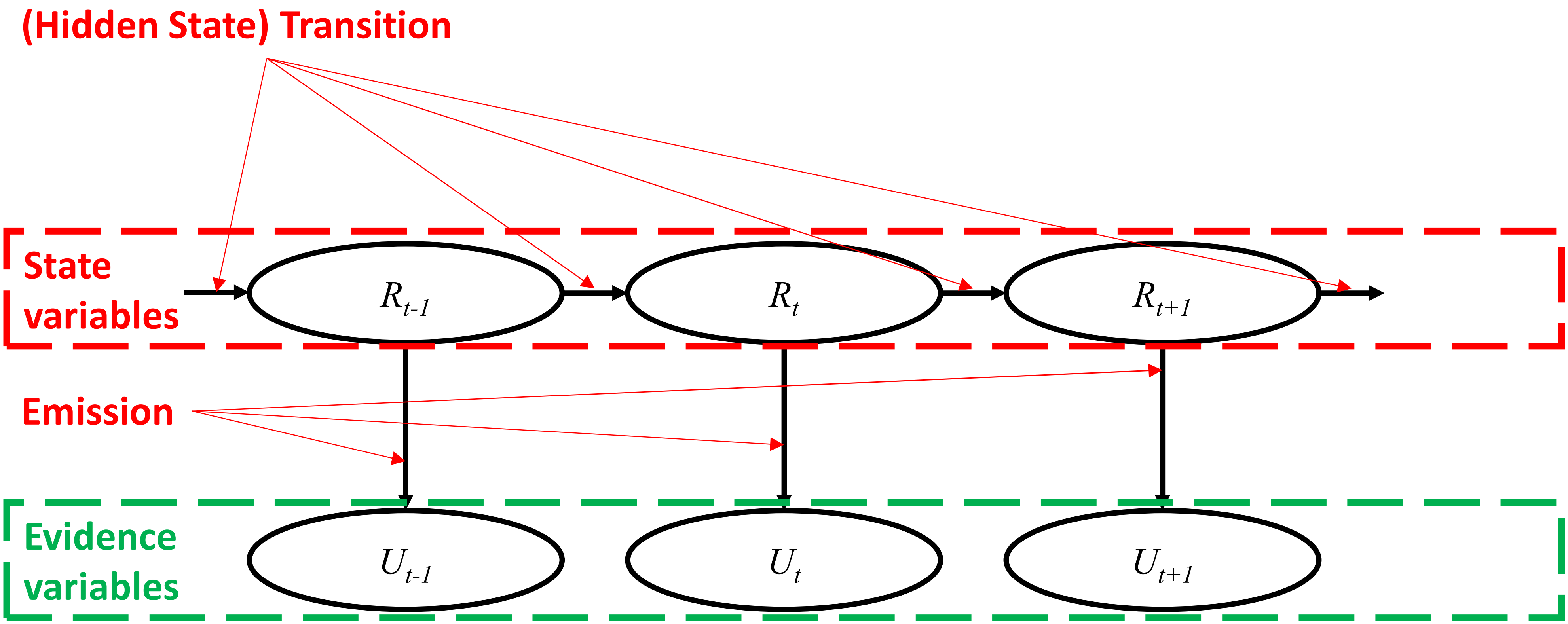
Inference in Temporal Model: Example



Inference in Temporal Model: Example



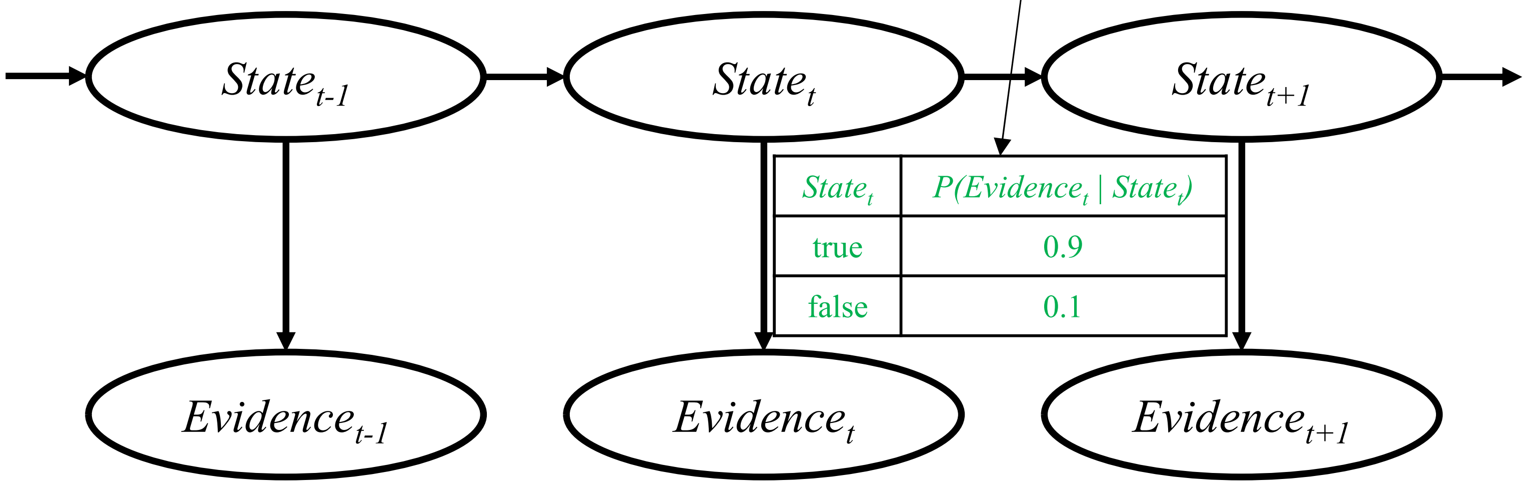
Inference in Temporal Model: Example



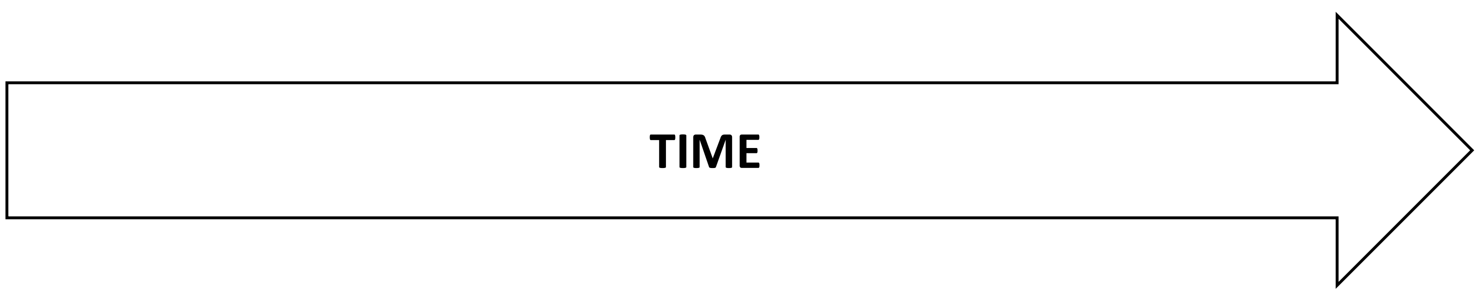
Transition and Sensor Models

Stationary system:
Transition and Emission probabilities stay FIXED

$State_{t-1}$	$P(State_t \mid State_{t-1})$
true	0.7
false	0.3



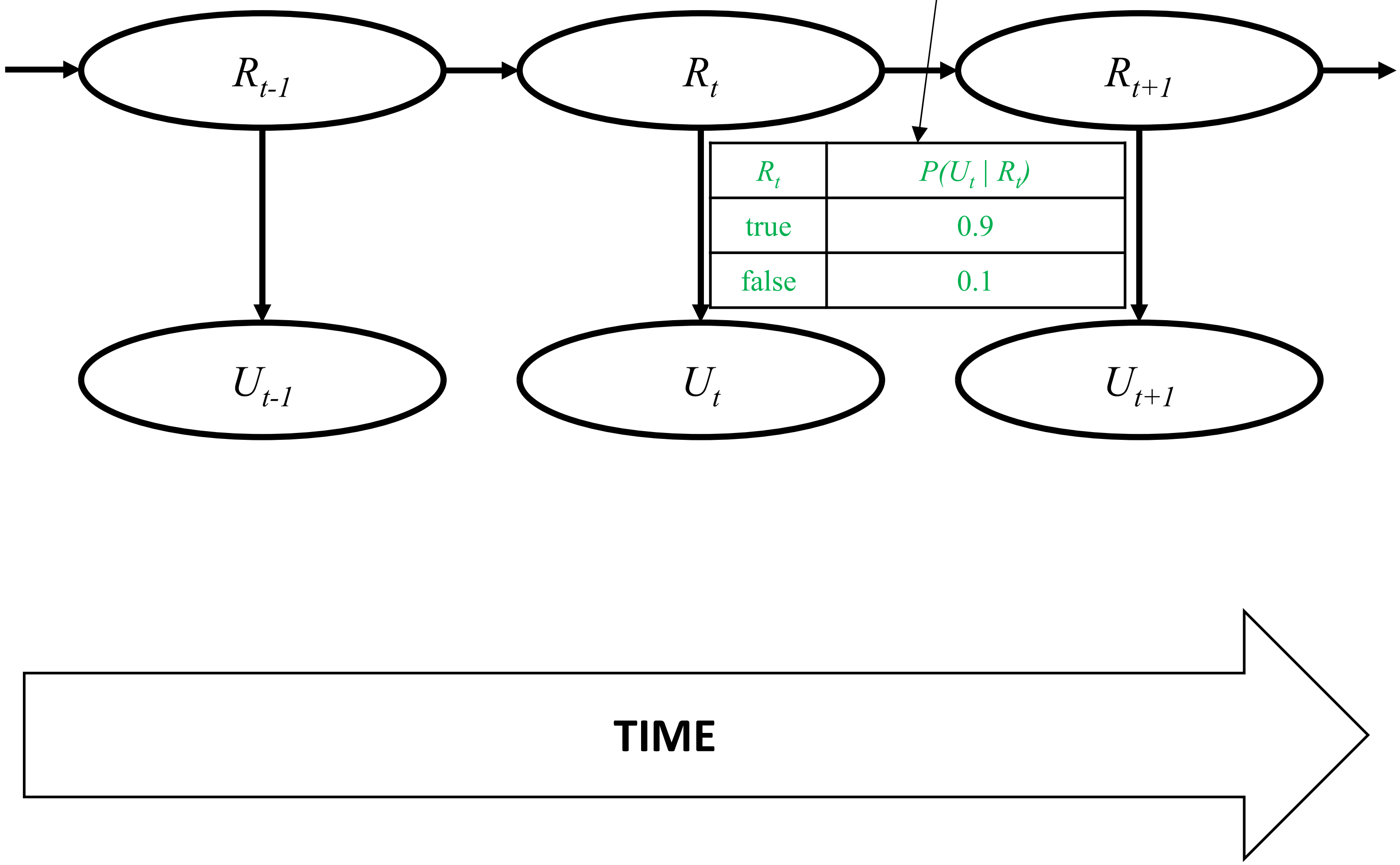
$State_t$	$P(Evidence_t \mid State_t)$
true	0.9
false	0.1



Transition and Sensor Models

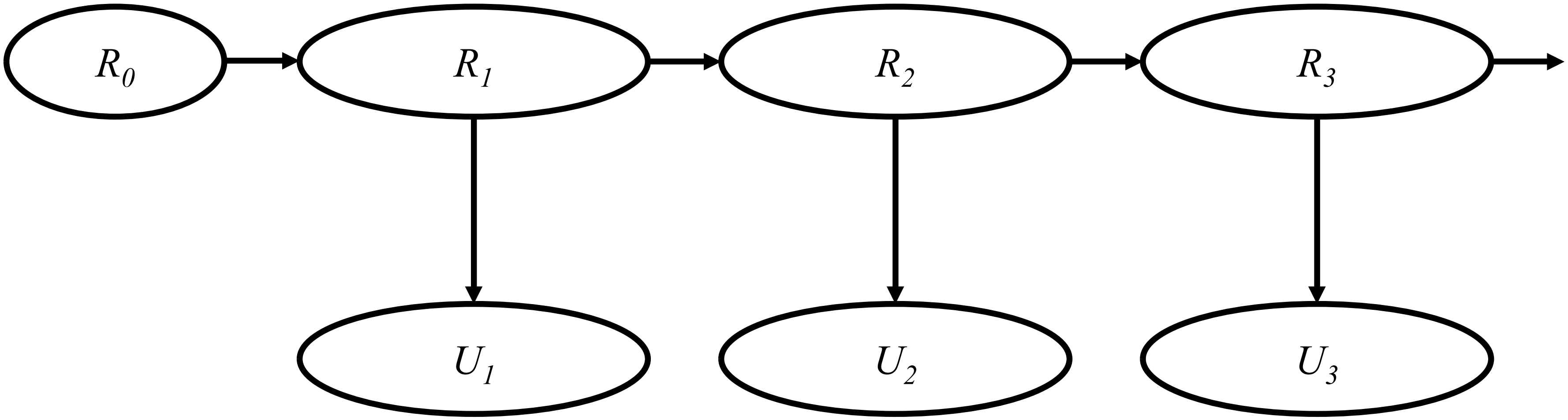
Stationary system:
Transition and Emission probabilities stay FIXED

R_{t-1}	$P(R_t \mid R_{t-1})$
true	0.7
false	0.3



R_t	$P(U_t \mid R_t)$
true	0.9
false	0.1

Inference in Temporal Model: Example



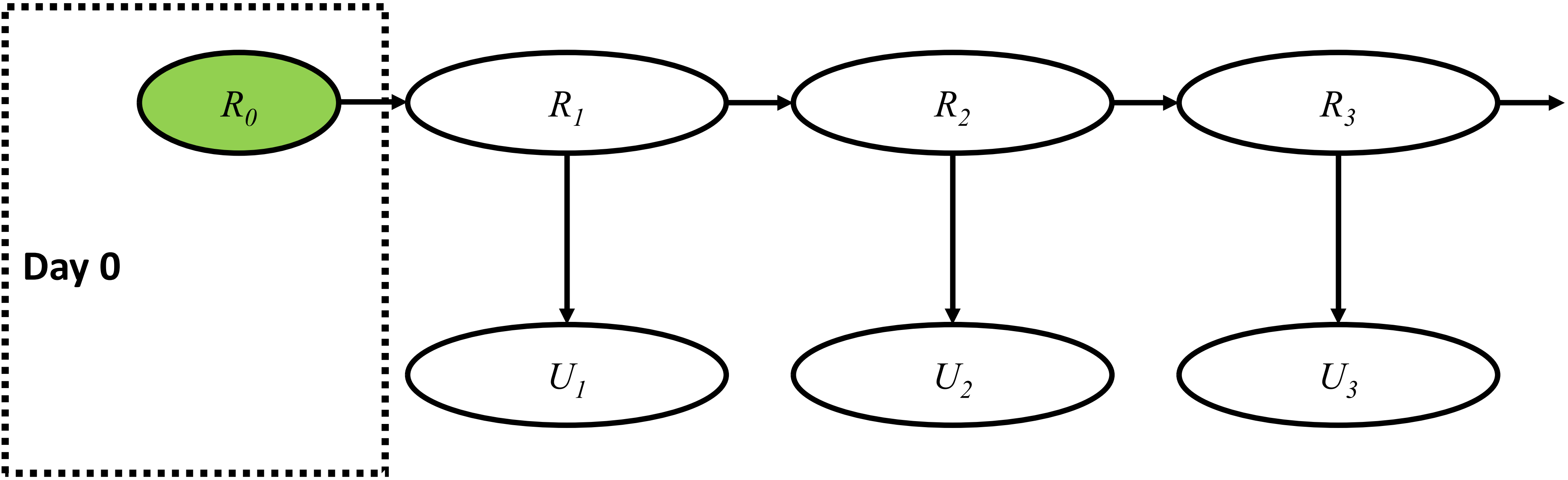
Possible values for R_i **:** $true, false$ **| Possible values for** U_i**:** $true, false$

Say we want to estimate $\mathbf{P}(R_2 \,|\, \mathrm{u}_{1:2})$**.**

Recall that:

$$\boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{e}_{1:t}) = \alpha * \boldsymbol{P}(\boldsymbol{e}_t \mid \boldsymbol{X}_t) * \sum_{\boldsymbol{x}_{t-1}} \boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{x}_{t-1}) * \boldsymbol{P}(\boldsymbol{x}_{t-1} \mid \boldsymbol{e}_{1:t-1})$$

Inference in Temporal Model: Example



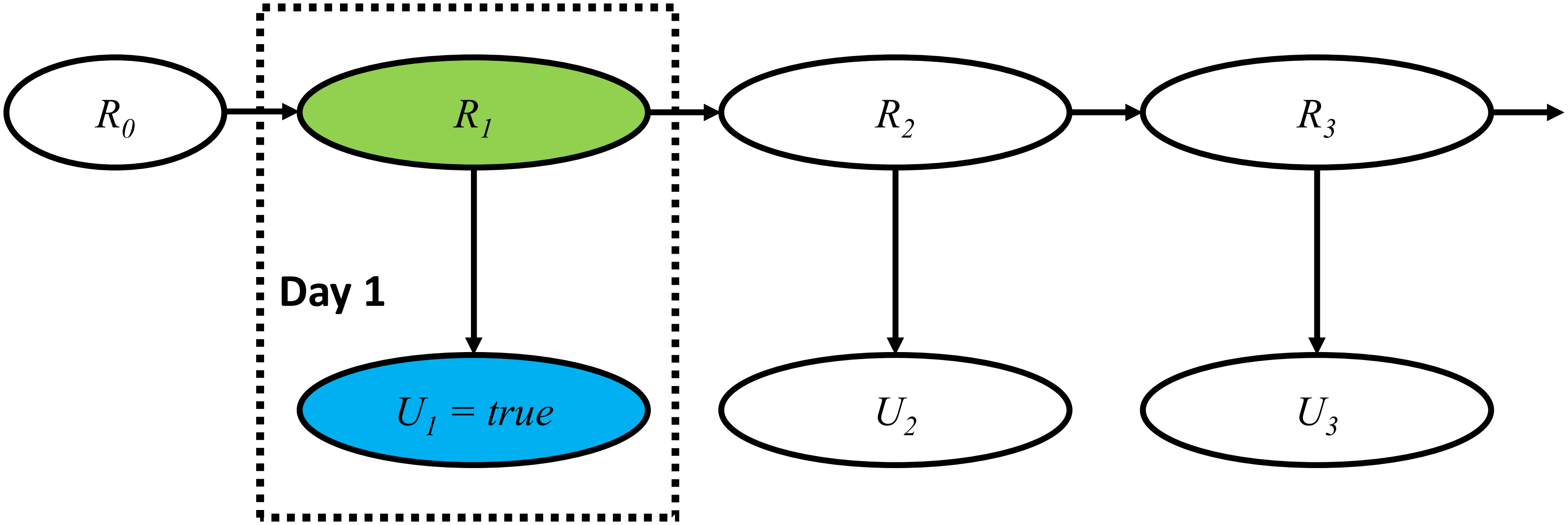
$$\boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{e}_{1:t}) = \alpha * \boldsymbol{P}(\boldsymbol{e}_t \mid \boldsymbol{X}_t) * \sum_{\boldsymbol{x}_{t-1}} \boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{x}_{t-1}) * \boldsymbol{P}(\boldsymbol{x}_{t-1} \mid \boldsymbol{e}_{1:t-1})$$

Let's assume that:

$$\boldsymbol{P}(R_0) = < 0.5, 0.5 >$$

And there are no observations on Day 0.

Inference in Temporal Model: Example



$$\boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{e}_{1:t}) = \alpha * \boldsymbol{P}(\boldsymbol{e}_t \mid \boldsymbol{X}_t) * \sum_{\boldsymbol{x}_{t-1}} \boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{x}_{t-1}) * \boldsymbol{P}(\boldsymbol{x}_{t-1} \mid \boldsymbol{e}_{1:t-1})$$

$$\boldsymbol{P}(R_1 \mid u_{1:1}) = \alpha * \boldsymbol{P}(u_1 \mid R_1) * \sum_{r_0} \boldsymbol{P}(R_1 \mid r_0) * \boldsymbol{P}(r_0 \mid u_{1:0})$$

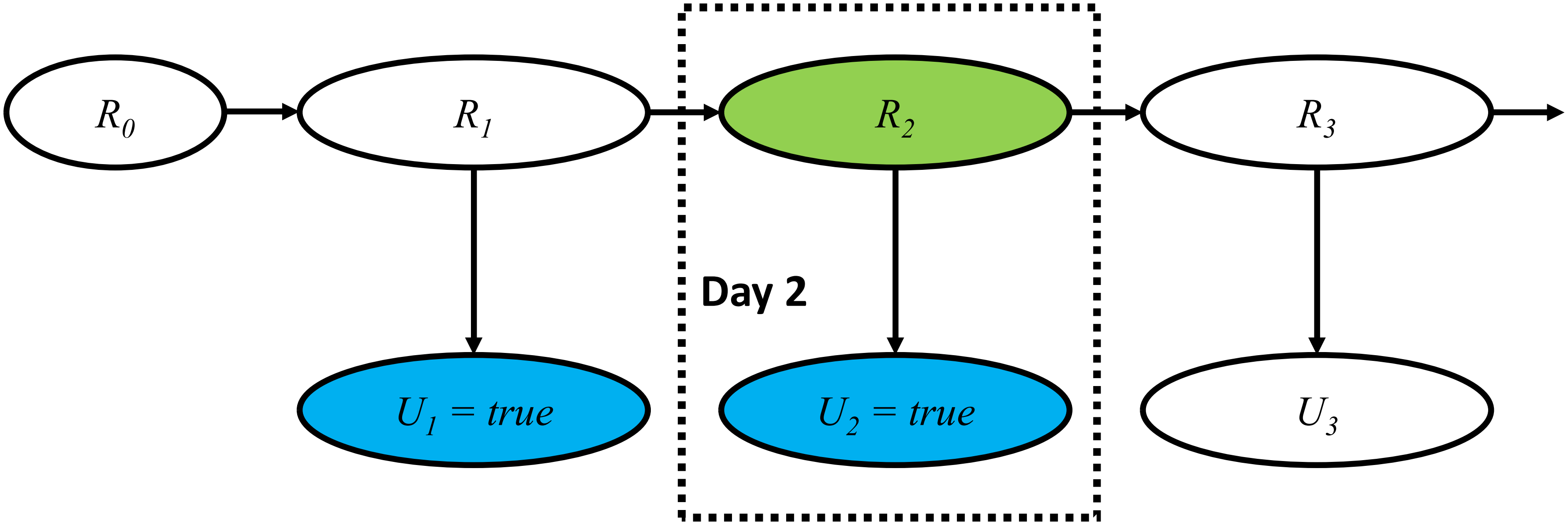
$$\boldsymbol{P}(R_1 \mid u_1) = \alpha * \boldsymbol{P}(u_1 \mid R_1) * \sum_{r_0} \boldsymbol{P}(R_1 \mid r_0) * \boldsymbol{P}(r_0)$$

$$\boldsymbol{P}(R_1 \mid u_1) = \alpha * < 0.9, 0.1 > * [< 0.7, 0.3 > * 0.5 + < 0.3, 0.7 > * 0.5]$$

$$\boldsymbol{P}(R_1 \mid u_1) = \alpha * < 0.9, 0.1 > * [< 0.35, 0.15 > + < 0.15, 0.35 >]$$

$$\boldsymbol{P}(R_1 \mid u_1) = \alpha * < 0.9, 0.1 > * < 0.5, 0.5 > \approx < 0.9, 0.1 >$$

Inference in Temporal Model: Example



$$\boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{e}_{1:t}) = \alpha * \boldsymbol{P}(\boldsymbol{e}_t \mid \boldsymbol{X}_t) * \sum_{\boldsymbol{x}_{t-1}} \boldsymbol{P}(\boldsymbol{X}_t \mid \boldsymbol{x}_{t-1}) * \boldsymbol{P}(\boldsymbol{x}_{t-1} \mid \boldsymbol{e}_{1:t-1})$$

$$\boldsymbol{P}(R_2 \mid u_{1:2}) = \alpha * \boldsymbol{P}(u_2 \mid R_2) * \sum_{r_1} \boldsymbol{P}(R_2 \mid r_1) * \boldsymbol{P}(r_1 \mid u_{1:1})$$

$$\boldsymbol{P}(R_2 \mid u_{1:2}) = \alpha * \boldsymbol{P}(u_2 \mid R_2) * \sum_{r_1} \boldsymbol{P}(R_2 \mid r_1) * \boldsymbol{P}(r_1 \mid u_1)$$

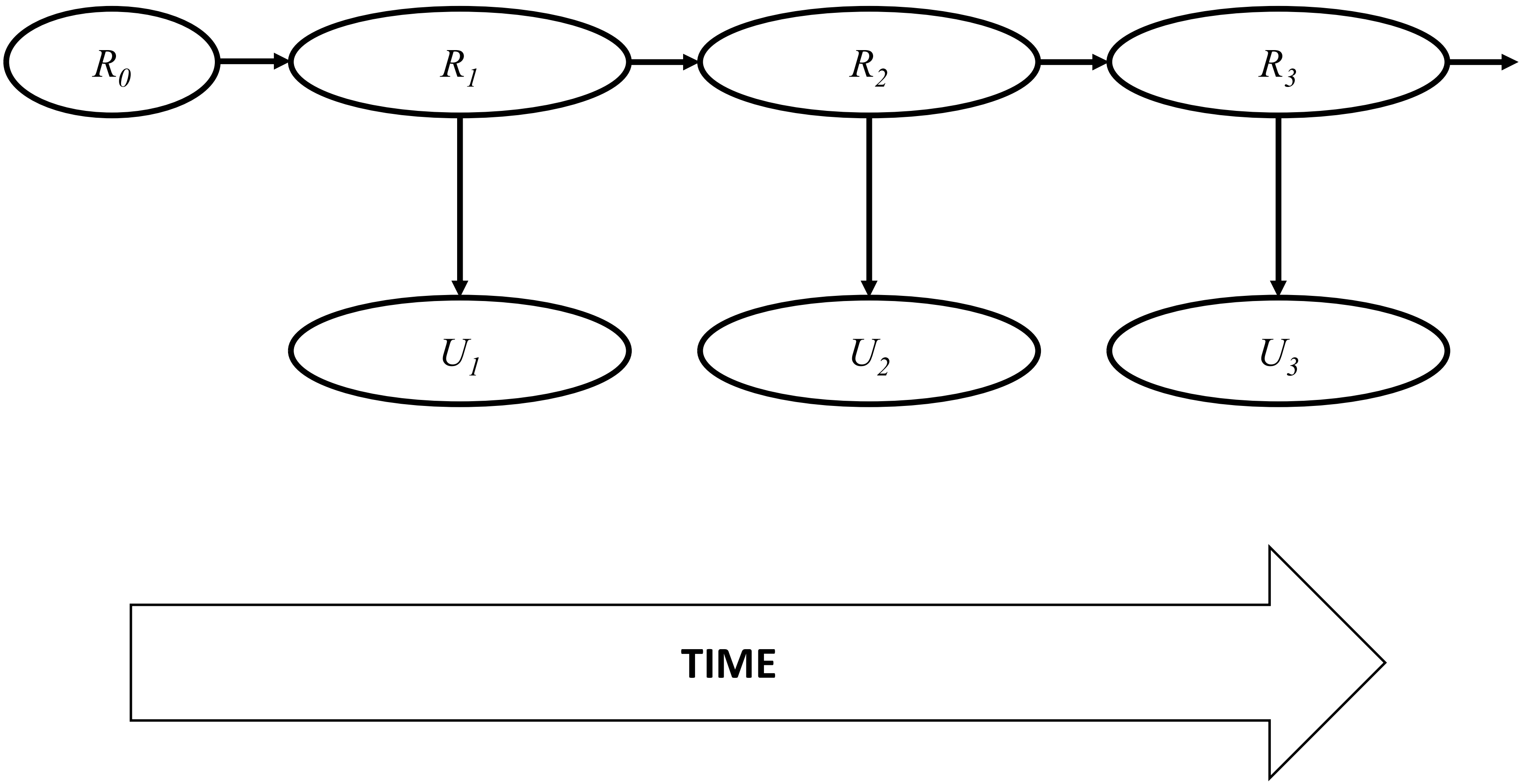
$$\boldsymbol{P}(R_2 \mid u_{1:2}) = \alpha * < 0.9, 0.1 > * [< 0.7, 0.3 > * 0.9 + < 0.3, 0.7 > * 0.1]$$

$$\boldsymbol{P}(R_2 \mid u_{1:2}) = \alpha * < 0.9, 0.1 > * [< 0.63, 0.27 > + < 0.03, 0.07 >]$$

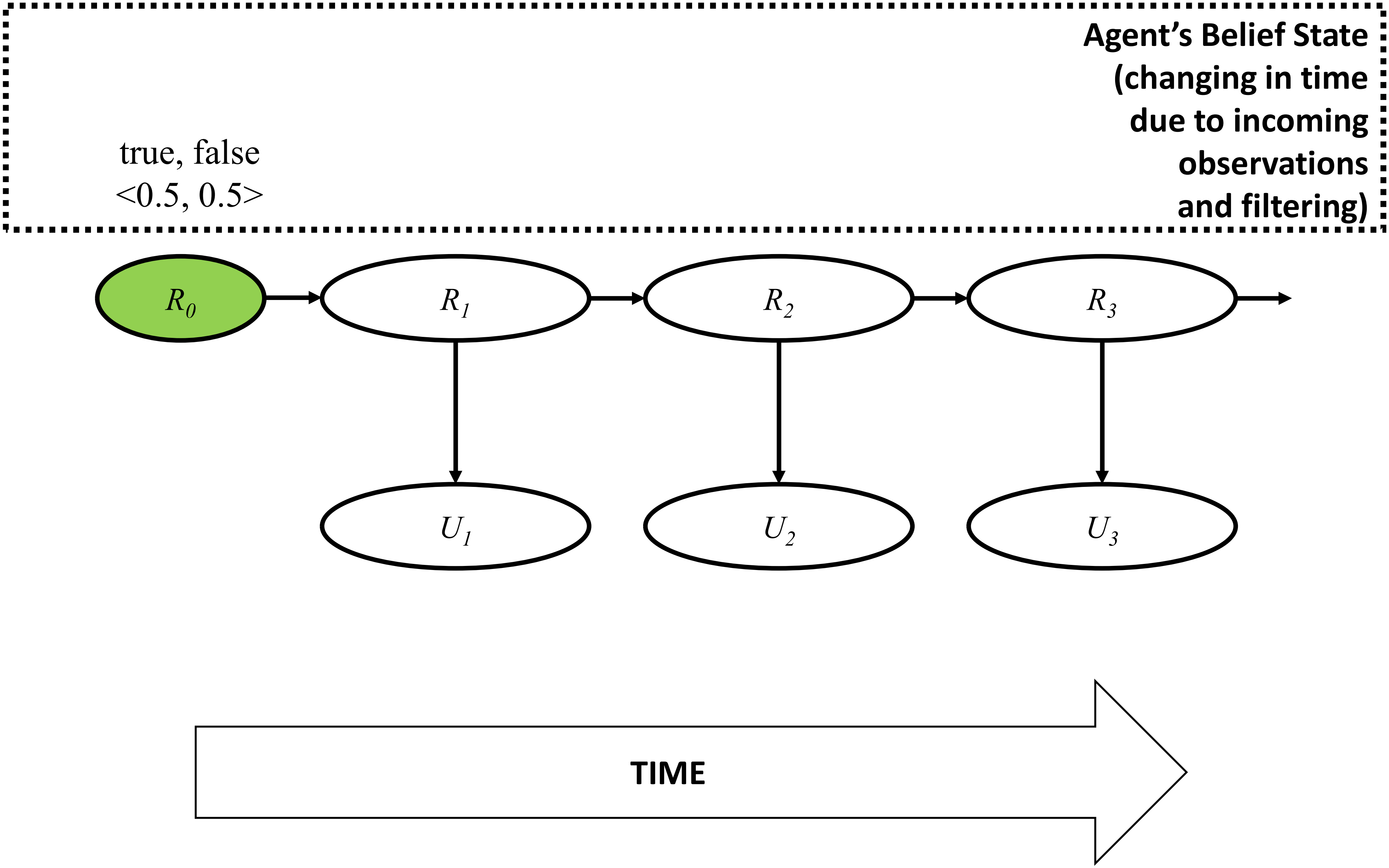
$$\boldsymbol{P}(R_2 \mid u_{1:2}) = \alpha * < 0.9, 0.1 > * < 0.66, 0.34 > \approx < 0.946, 0.054 >$$

Inference in Temporal Model: Example

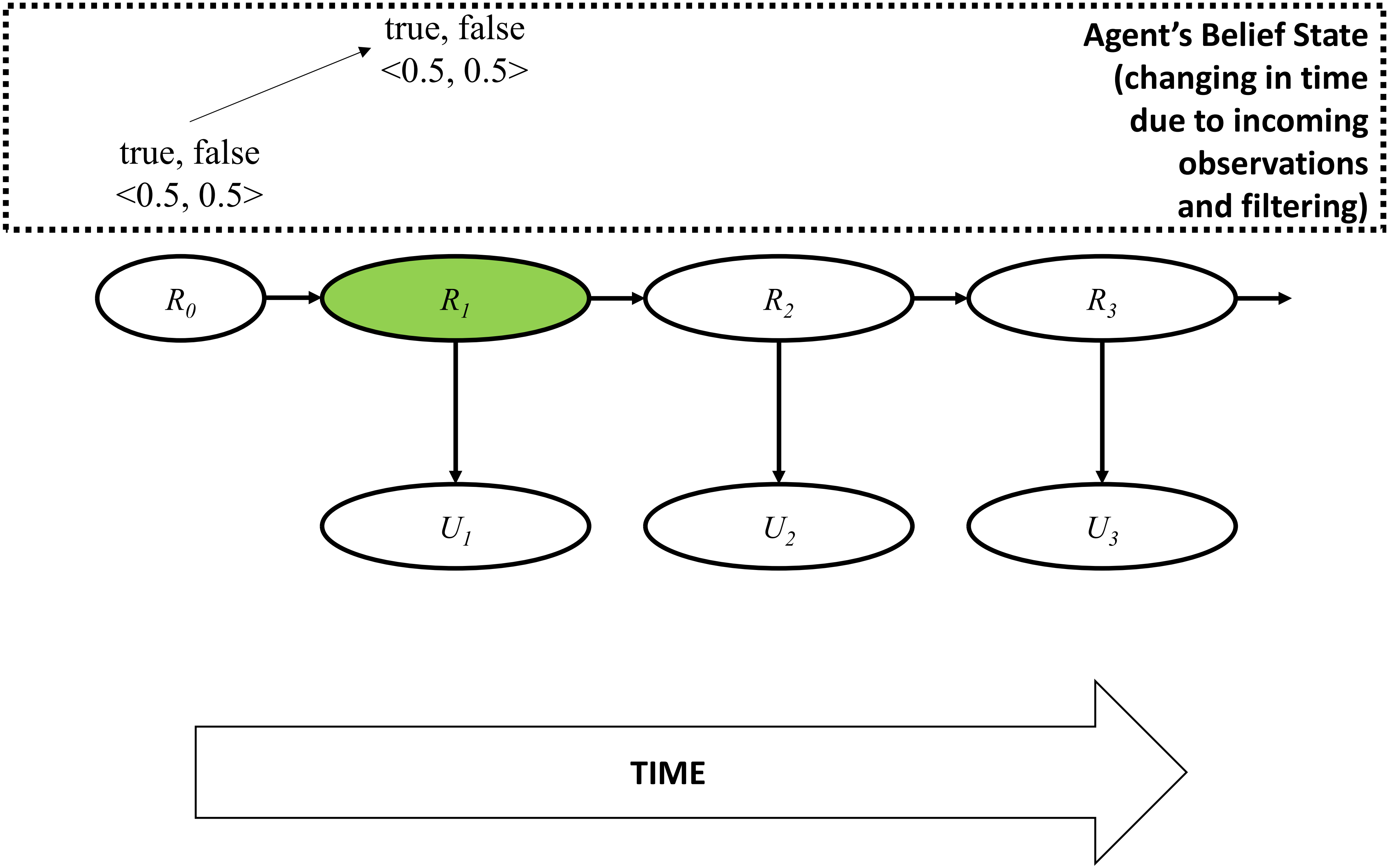
Agent's Belief State
(changing in time
due to incoming
observations
and filtering)



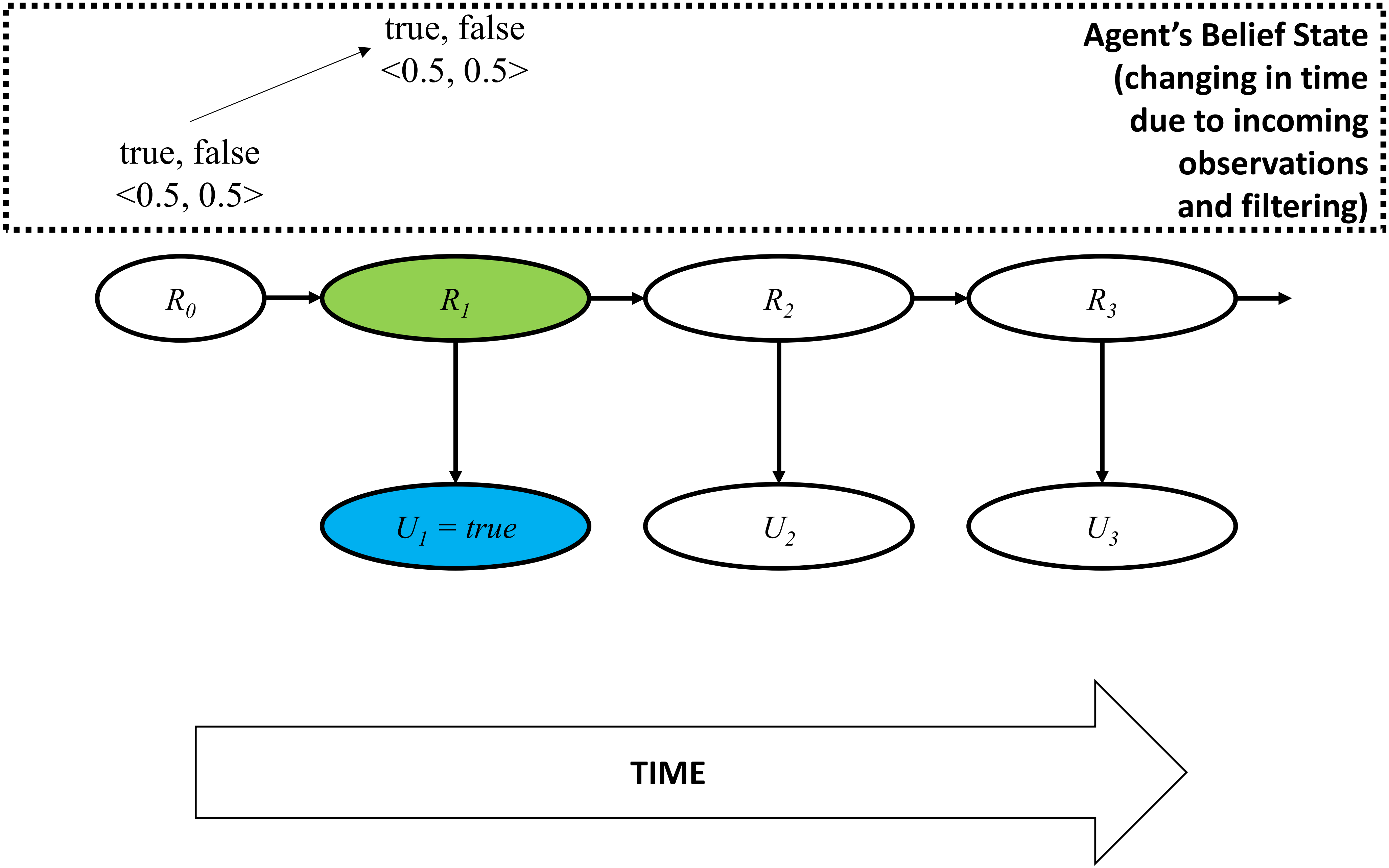
Inference in Temporal Model: Example



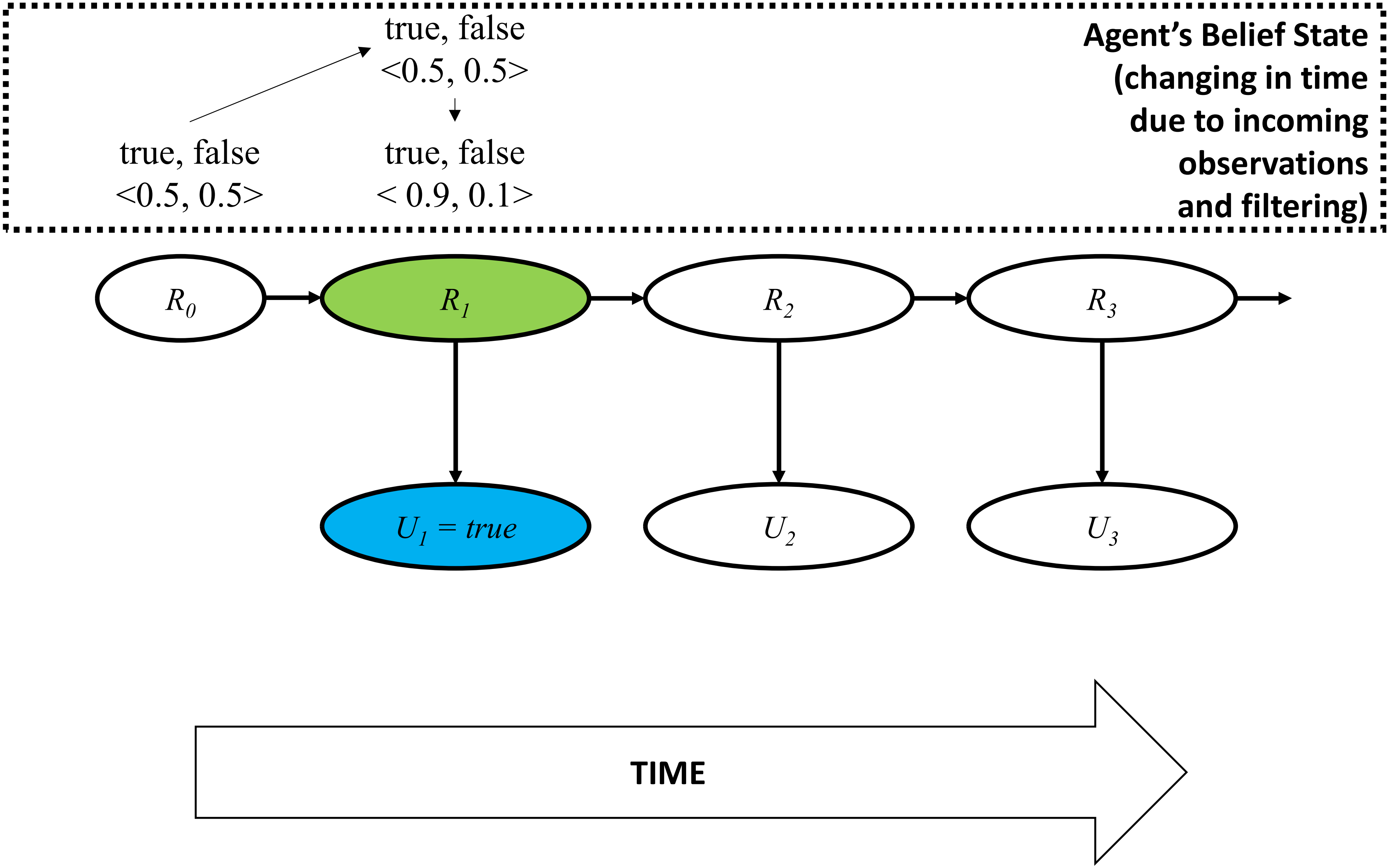
Inference in Temporal Model: Example



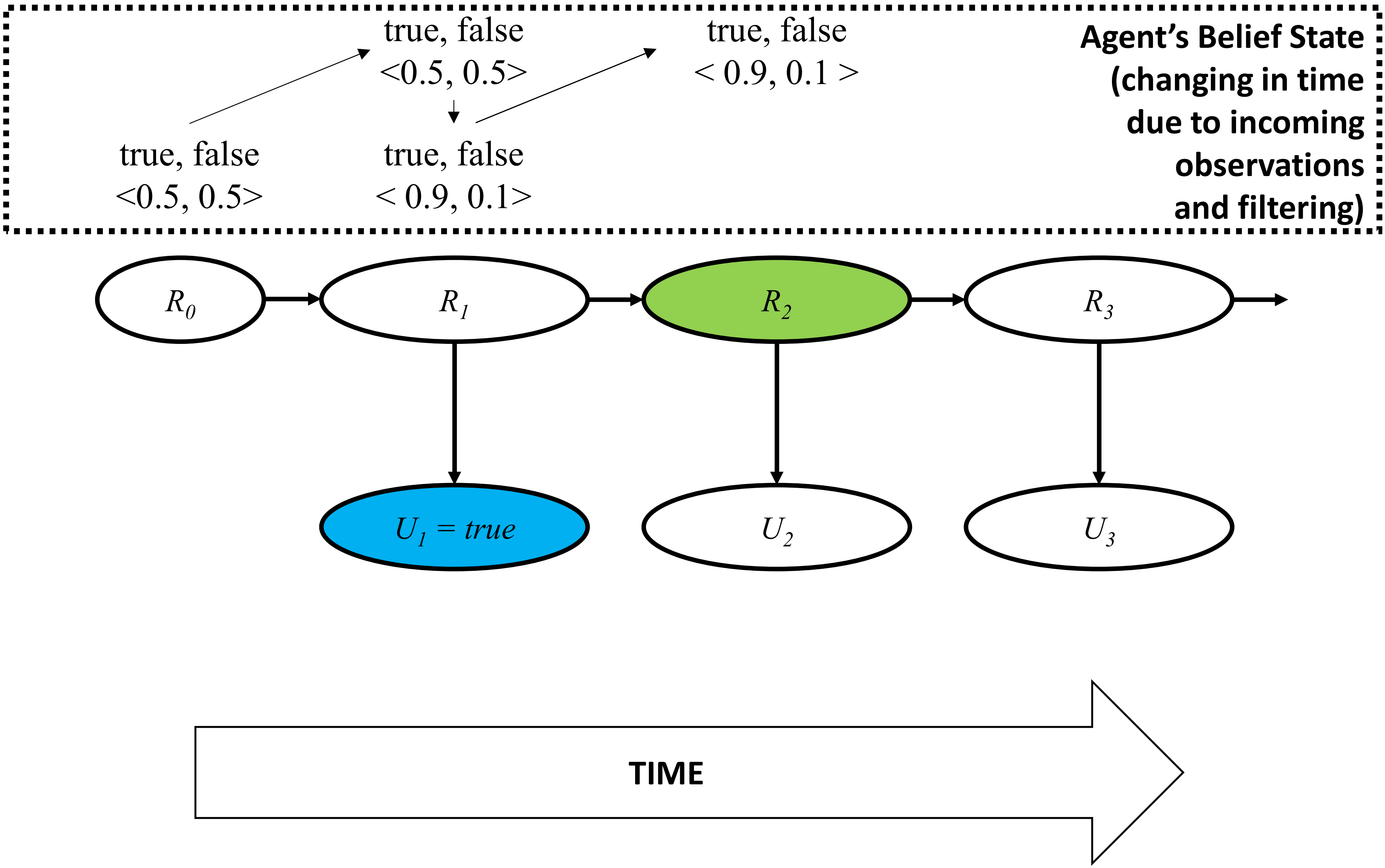
Inference in Temporal Model: Example



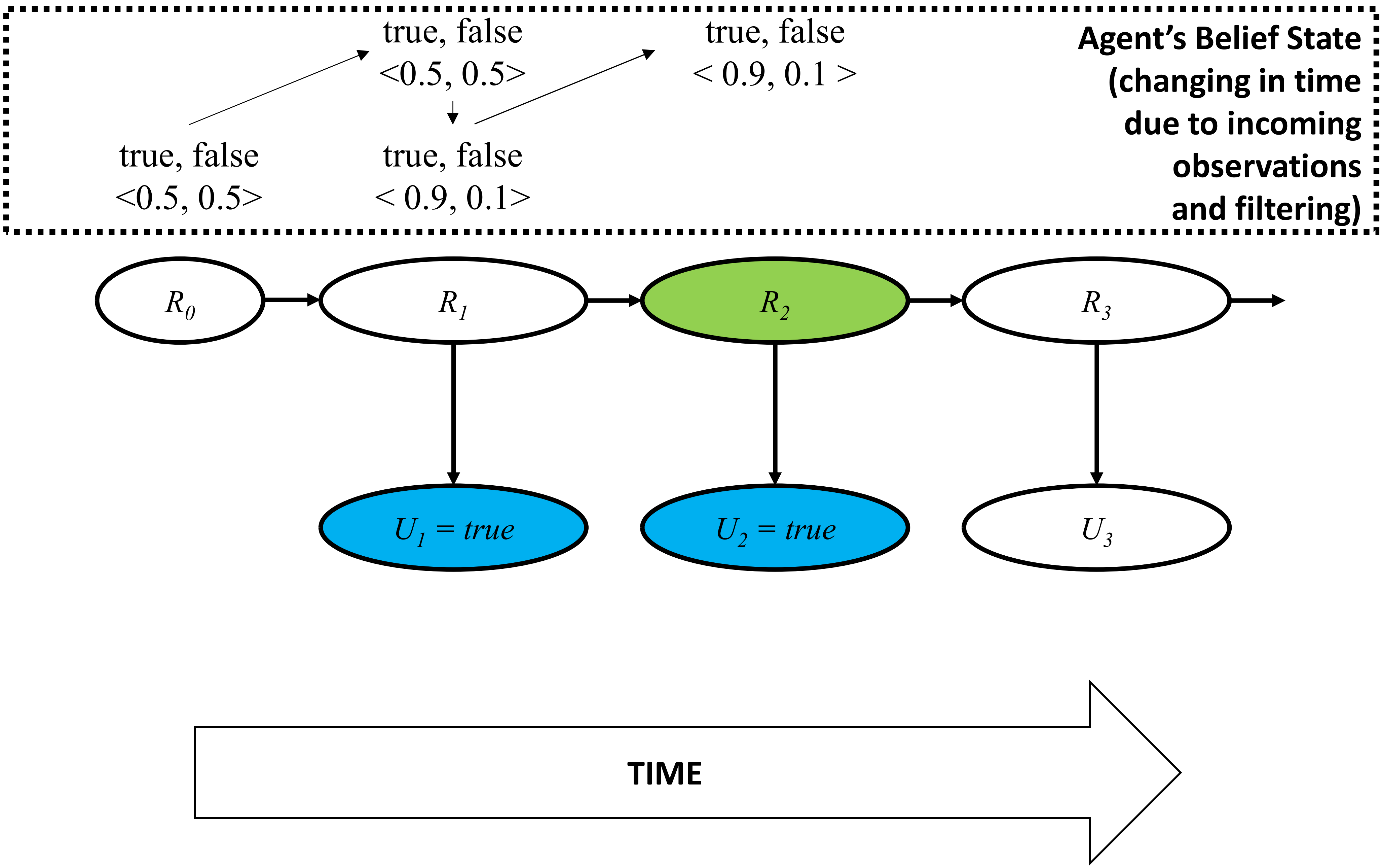
Inference in Temporal Model: Example



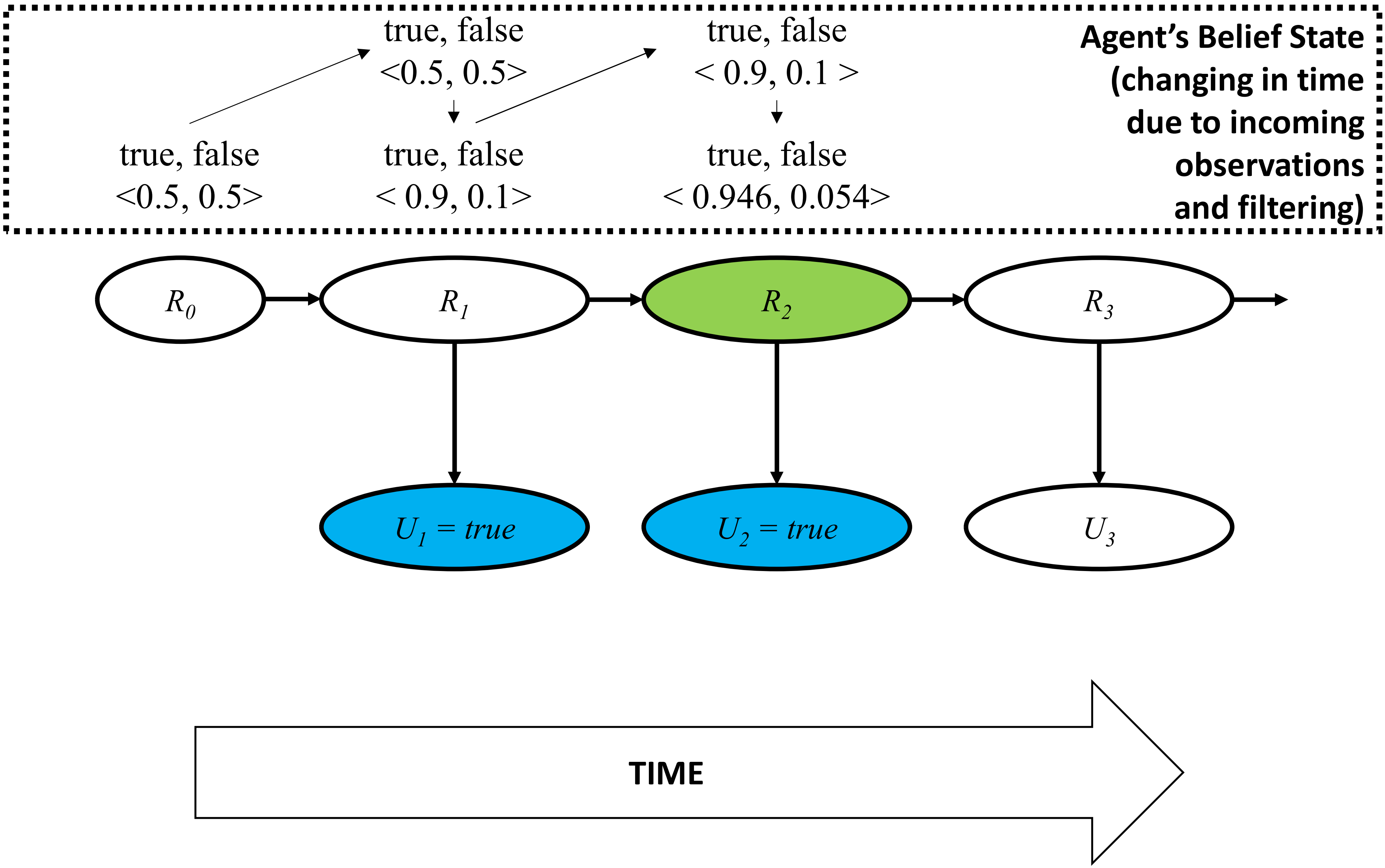
Inference in Temporal Model: Example



Inference in Temporal Model: Example



Inference in Temporal Model: Example

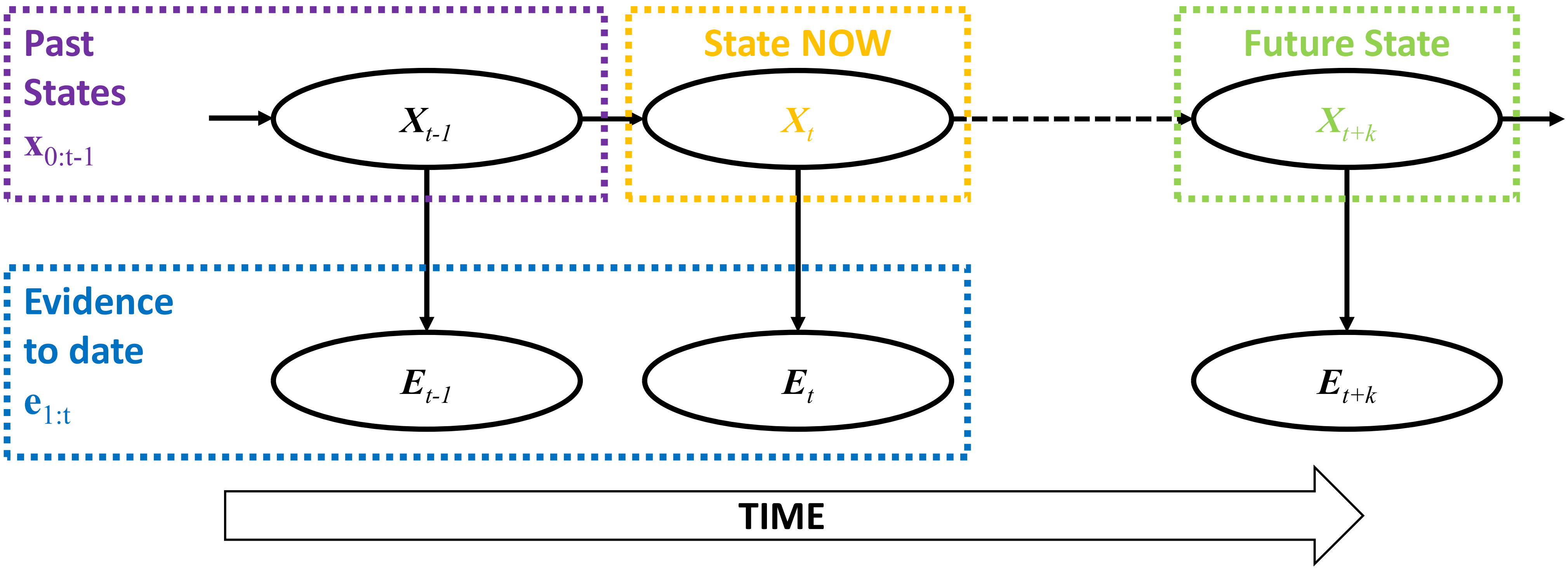


Prediction

Inference: Prediction

This is he task of computing the posterior distribution over the future state (time $t+k$, for some $k > 0$), given all evidence to date. Useful for evaluating possible courses of action.

$$P(\mathbf{X}_{t+k} \mid \mathbf{e}_{1:t})$$

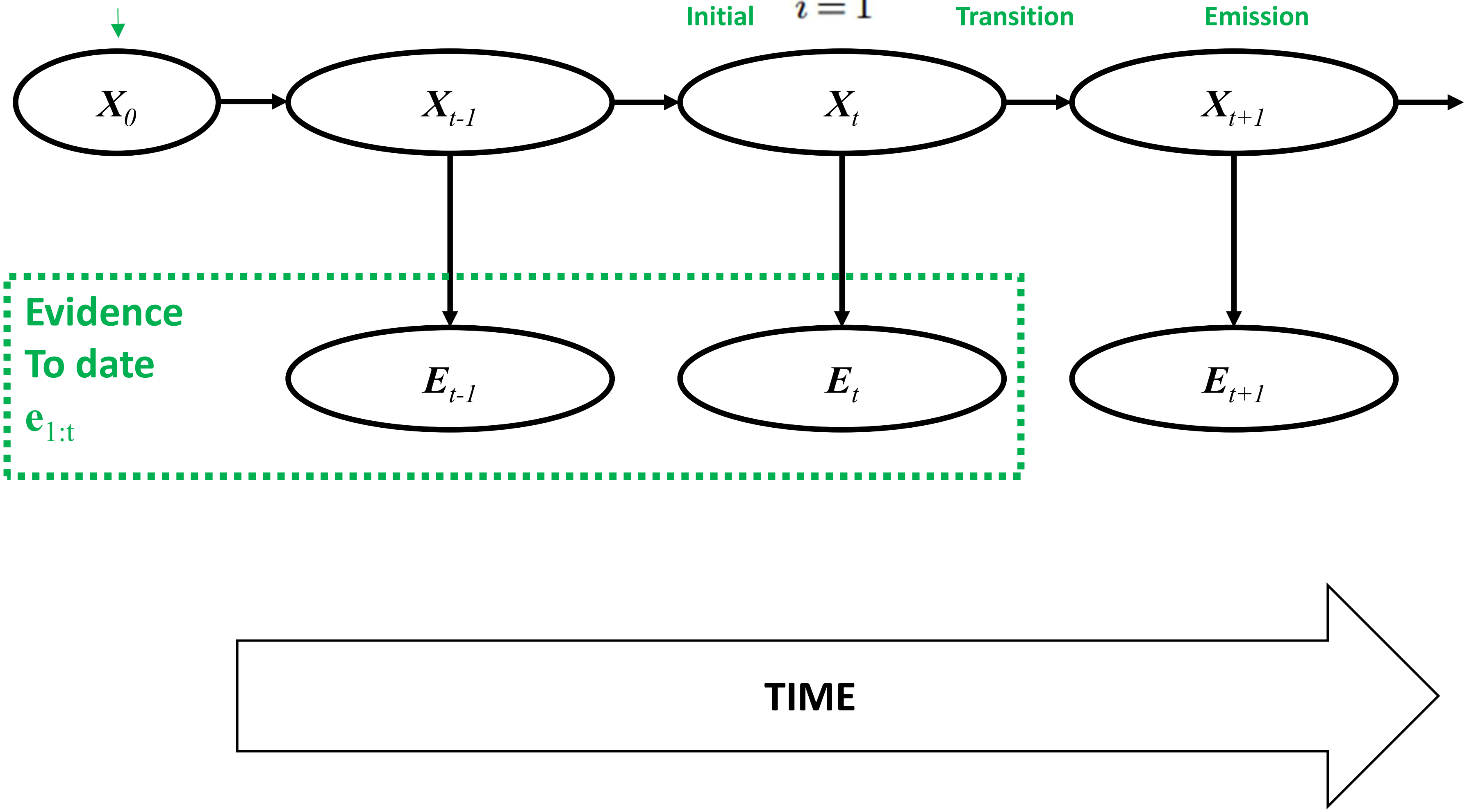


What Is Known?

The complete (including initial state distribution | for any t) joint probability distribution for a sequence of transitions and emissions:

$$\mathbf{P}(\mathbf{X}_{0:t}, \mathbf{E}_{1:t}) = \underbrace{\mathbf{P}(\mathbf{X}_0)}_{\text{Initial}} \prod_{i=1}^{t} \underbrace{\mathbf{P}(\mathbf{X}_i \mid \mathbf{X}_{i-1})}_{\text{Transition}} \underbrace{\mathbf{P}(\mathbf{E}_i \mid \mathbf{X}_i)}_{\text{Emission}}$$

Initial State Distribution



Filtering / Recursive Estimation

Evidence can be separated: $\boldsymbol{e}_{1:t+1} = \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1}$

$$\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t+1}) = \boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1})$$

Applying Bayes Rule yields:

$$\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1}) = \alpha * \boldsymbol{P}(\boldsymbol{e}_{t+1} \mid \boldsymbol{X}_{t+1}, \boldsymbol{e}_{1:t}) * \boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t})$$

Recall the Markov assumption for the sensor model (emission):

$$\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t}, \boldsymbol{e}_{t+1}) = \alpha * \underbrace{\boldsymbol{P}(\boldsymbol{e}_{t+1} \mid \boldsymbol{X}_{t+1})}_{\text{update}} * \underbrace{\boldsymbol{P}(\boldsymbol{X}_{t+1} \mid \boldsymbol{e}_{1:t})}_{\text{prediction}}$$

$\boldsymbol{P}(\boldsymbol{X}_{t+k} \mid \mathbf{e}_{1:t})$

Prediction

Prediction: Future State Estimate

Prediction = filtering WITHOUT adding new evidence (UPDATE):

$$P(X_{t+k+1} \mid e_{1:t}) = * \sum_{x_{t+k}} \underbrace{\underbrace{P(X_{t+k+1} \mid x_{t+k})}_{\text{Transition model}}}_{\text{KNOWN model of the World / Environment}} * \underbrace{\underbrace{P(x_{t+k} \mid e_{1:t})}_{\text{Recursion}}}_{\text{KNOWN Previous State Estimate}}$$

This is a recursive (estimation) relationship:

$$P(X_{t+k+1} \mid e_{1:t}) = f(P(X_{t+k} \mid e_{1:t}))$$

New state estimate depends only (can be updated based) on previous state estimate.

Likelihood of Evidence Sequence

Likelihood of evidence sequence (useful in comparing temporal models):

$$\boldsymbol{P}(\boldsymbol{e}_{1:t})$$

We can think of another "message" (likelihood message) here, $\boldsymbol{l}_{1:t}$. A message propagated forward along the sequence. The process is given by:

$$l_{1:t} = FORWARD(l_{1:t-1}, \boldsymbol{e}_t)$$

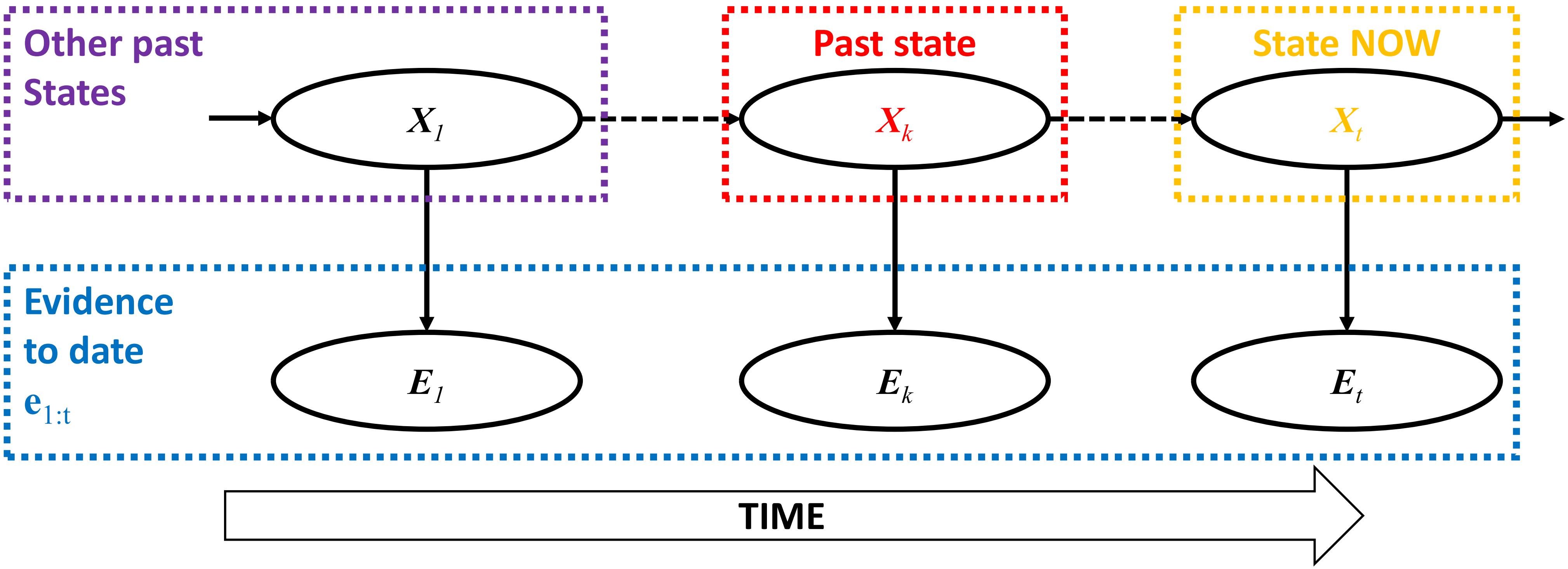
$$f_{1:t} = P(\boldsymbol{e}_{1:t}) = \sum_{\boldsymbol{x}_t} l_{1:t}(\boldsymbol{x}_t)$$

Smoothing

Inference: Smoothing

This is the task of computing the posterior distribution over the past state (time k, for some $0 \leq k < t$), given all evidence to date. Provides a better state estimate of, because it incorporates more evidence.

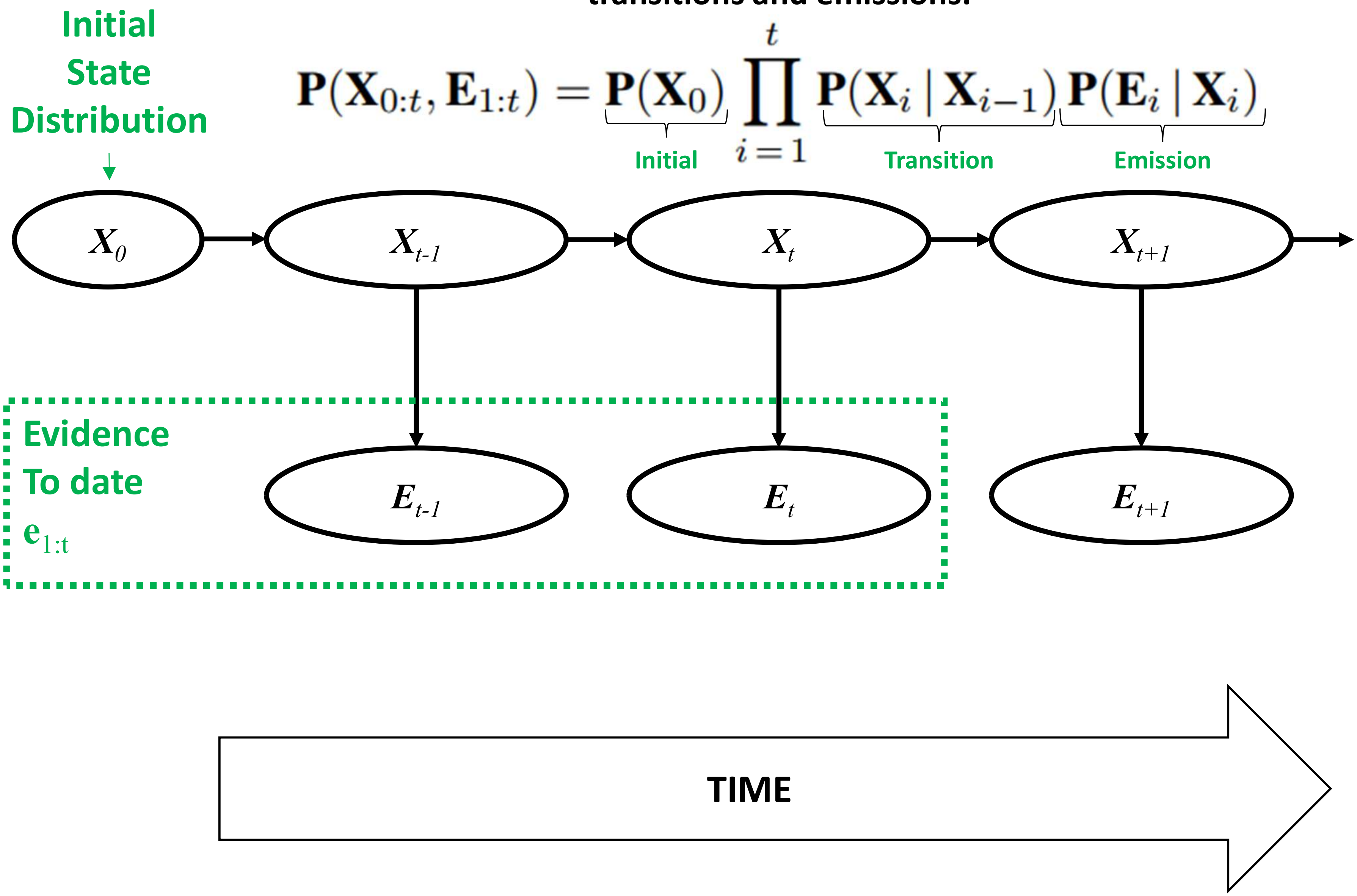
$$P(\mathbf{X}_k \mid \mathbf{e}_{1:t})$$



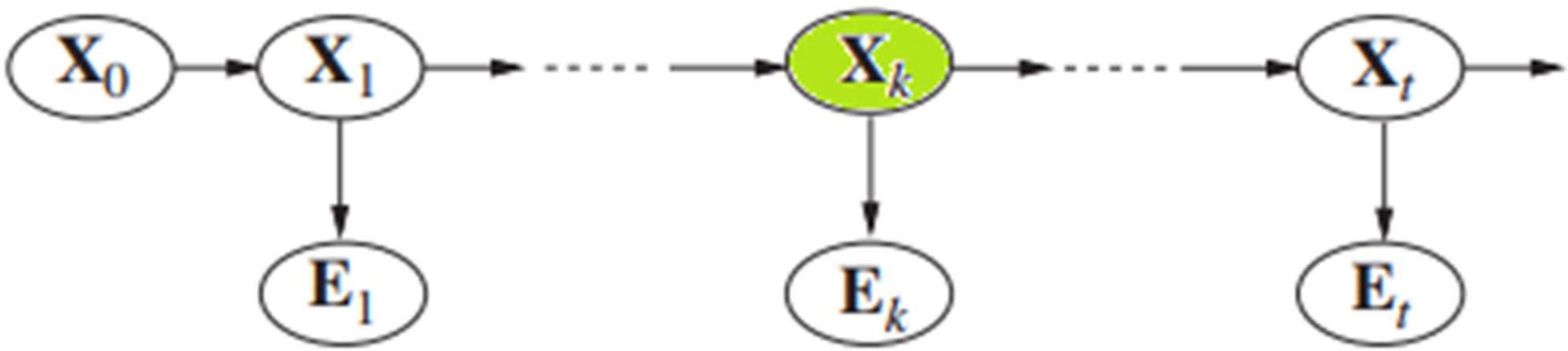
What Is Known?

The complete (including initial state distribution | for any t) joint probability distribution for a sequence of transitions and emissions:

$$\mathbf{P}(\mathbf{X}_{0:t}, \mathbf{E}_{1:t}) = \underbrace{\mathbf{P}(\mathbf{X}_0)}_{\text{Initial}} \prod_{i=1}^{t} \underbrace{\mathbf{P}(\mathbf{X}_i \mid \mathbf{X}_{i-1})}_{\text{Transition}} \underbrace{\mathbf{P}(\mathbf{E}_i \mid \mathbf{X}_i)}_{\text{Emission}}$$



Smoothing / Recursive Estimation



This time we are after:

$$P(X_k \mid e_{1:t})$$

Evidence can be separated: $e_{1:t} = e_{1:k}, e_{k+1:t}$

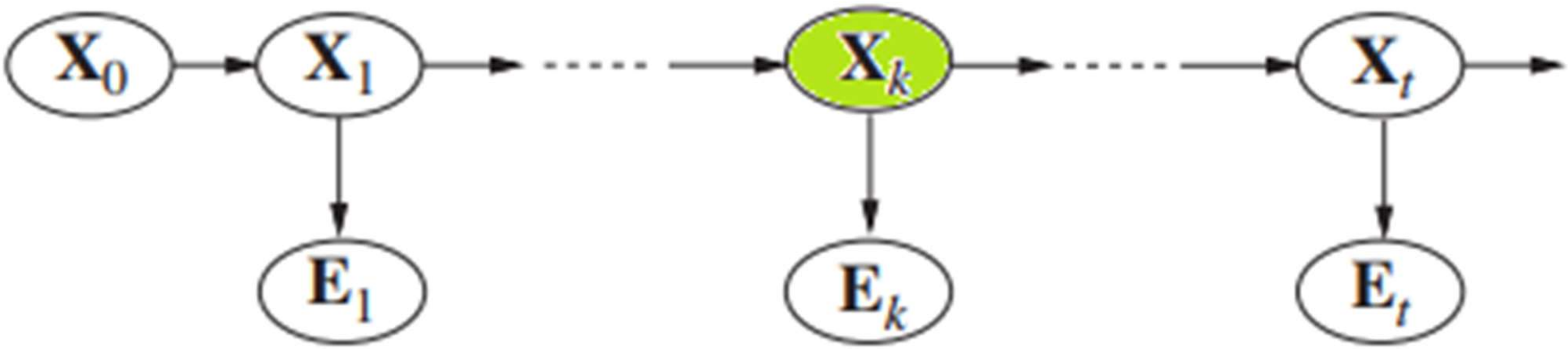
$$P(X_k \mid e_{1:t}) = P(X_k \mid e_{1:k}, e_{k+1:t})$$

Applying Bayes Rule yields:

Normalizing constant

$$P(X_k \mid e_{1:k}, e_{k+1:t}) = \alpha * P(X_k \mid e_{1:k}) * P(e_{k+1:t} \mid X_k, e_{1:k})$$

Smoothing / Recursive Estimation



From previous slide:

$$P(X_k \mid e_{1:k}, e_{k+1:t}) = \alpha * P(X_k \mid e_{1:k}) * P(e_{k+1:t} \mid X_k, e_{1:k})$$

After applying conditional independence:

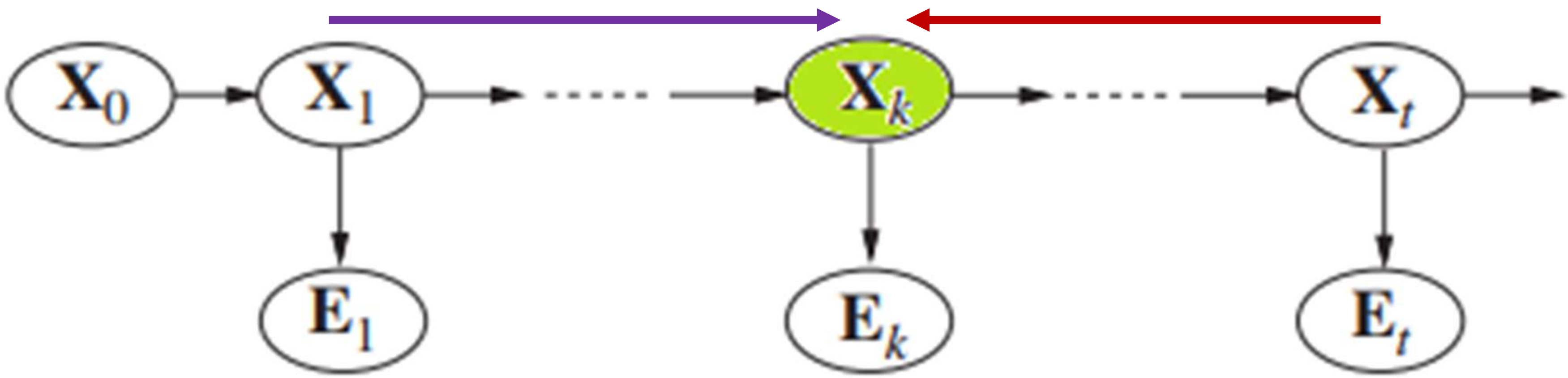
$$P(X_k \mid e_{1:k}, e_{k+1:t}) = \alpha * P(X_k \mid e_{1:k}) * P(e_{k+1:t} \mid X_k)$$

Let's express it in terms of messages ("forward" and "backward"):

$$P(X_k \mid e_{1:k}, e_{k+1:t}) = \alpha * f_{1:k} \times b_{k+1:t}$$

$\times$: Pointwise vector multiplication

Smoothing / Recursive Estimation



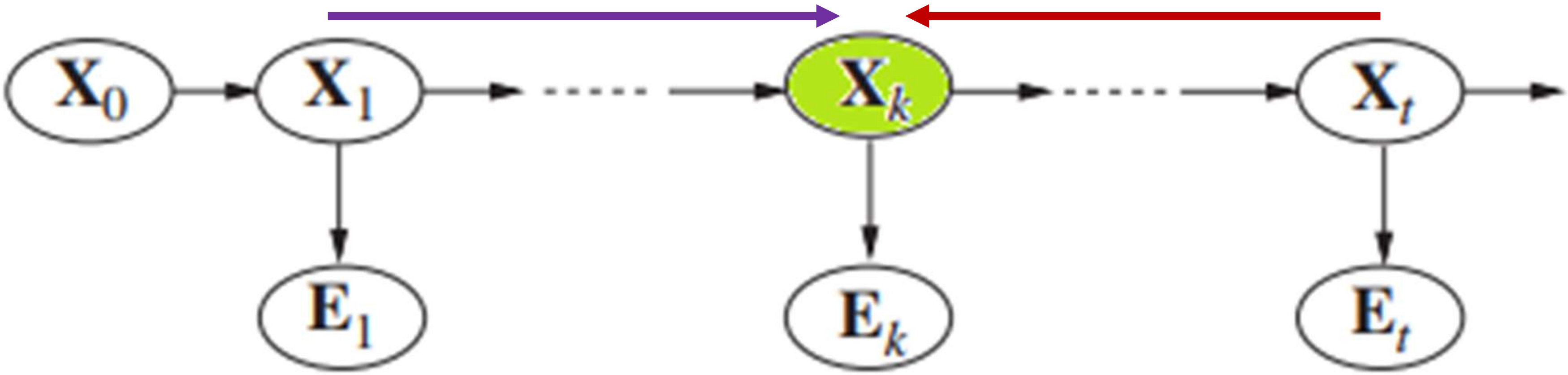
Expressed in terms of messages ("forward" and "backward"):

$$P(X_k \mid e_{1:k}, e_{k+1:t}) = \alpha * f_{1:k} \times b_{k+1:t}$$

where:

$$b_{k+1:t} = P(e_{k+1:t} \mid X_k)$$

Smoothing / Recursive Estimation



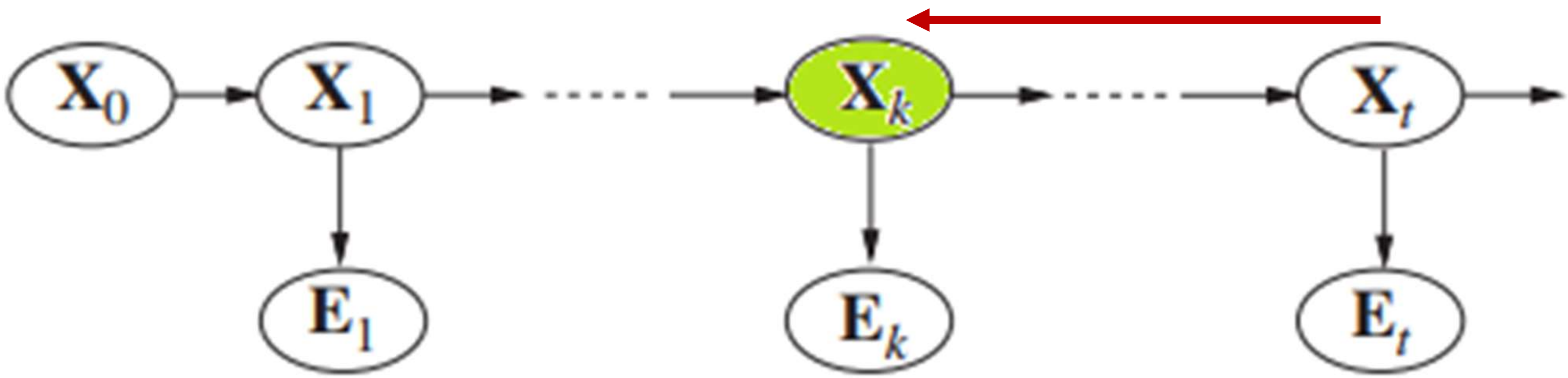
Expressed in terms of messages ("forward" and "backward"):

$$P(X_k \mid e_{1:k}, e_{k+1:t}) = \alpha * f_{1:k} \times b_{k+1:t}$$

where:

$$b_{k+1:t} = P(e_{k+1:t} \mid X_k)$$

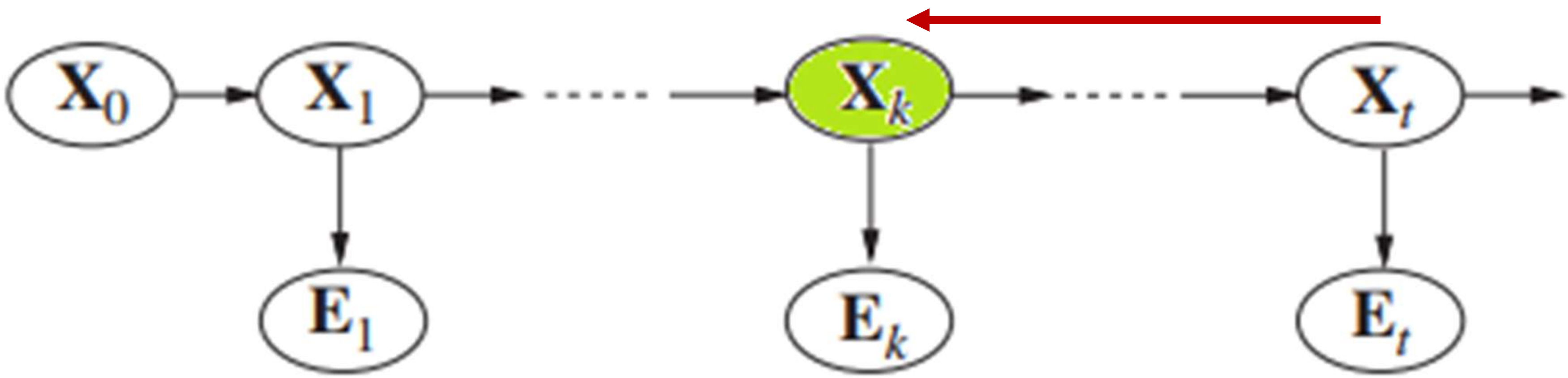
Smoothing / Recursive Estimation



"backward" message:

$$\begin{aligned}
\mathbf{P}(\mathbf{e}_{k+1:t} \mid \mathbf{X}_k) &= \sum_{\mathbf{x}_{k+1}} \mathbf{P}(\mathbf{e}_{k+1:t} \mid \mathbf{X}_k, \mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1} \mid \mathbf{X}_k) \quad \text{(conditioning on } \mathbf{X}_{k+1}\text{)} \\
&= \sum_{\mathbf{x}_{k+1}} P(\mathbf{e}_{k+1:t} \mid \mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1} \mid \mathbf{X}_k) \quad \text{(by conditional independence)} \\
&= \sum_{\mathbf{x}_{k+1}} P(\mathbf{e}_{k+1}, \mathbf{e}_{k+2:t} \mid \mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1} \mid \mathbf{X}_k) \\
&= \sum_{\mathbf{x}_{k+1}} P(\mathbf{e}_{k+1} \mid \mathbf{x}_{k+1})P(\mathbf{e}_{k+2:t} \mid \mathbf{x}_{k+1})\mathbf{P}(\mathbf{x}_{k+1} \mid \mathbf{X}_k),
\end{aligned}$$

Smoothing / Recursive Estimation



"**backward**" **message recursive relationship:**

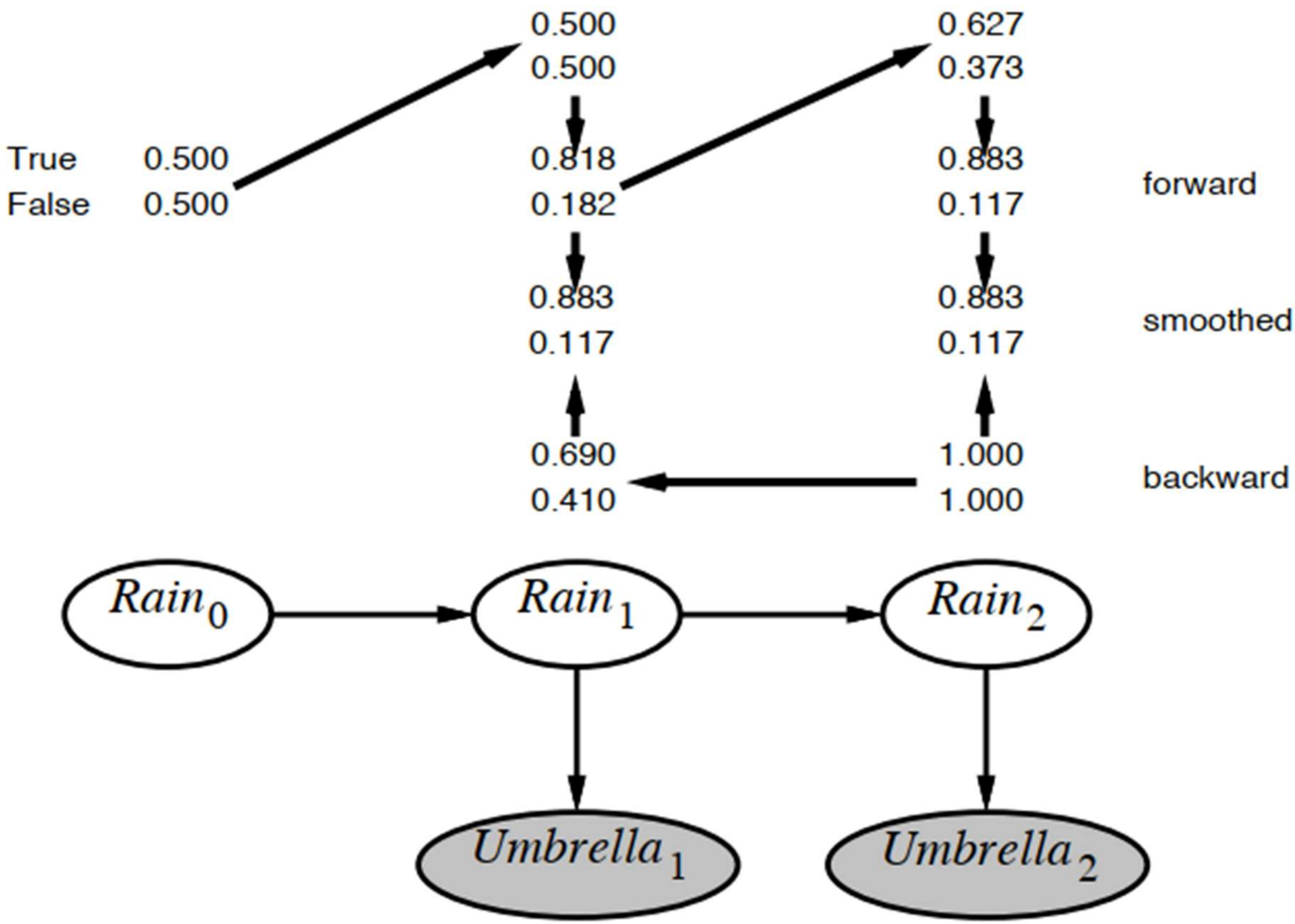
$$\boldsymbol{b}_{k+1:t} = BACKWARD(\boldsymbol{b}_{k+2:t}, \boldsymbol{e}_{k+1})$$

Forward-Backward Algorithm

function FORWARD-BACKWARD(**ev**, *prior*) **returns** a vector of probability distributions
  inputs: **ev**, a vector of evidence values for steps $1, \ldots, t$
    prior, the prior distribution on the initial state, $\mathbf{P}(\mathbf{X}_0)$
  local variables: **fv**, a vector of forward messages for steps $0, \ldots, t$
    b, a representation of the backward message, initially all 1s
    sv, a vector of smoothed estimates for steps $1, \ldots, t$

  $\mathbf{fv}[0] \leftarrow prior$
  for $i = 1$ **to** t **do**
    $\mathbf{fv}[i] \leftarrow$ FORWARD($\mathbf{fv}[i-1], \mathbf{ev}[i]$)
  for $i = t$ **downto** 1 **do**
    $\mathbf{sv}[i] \leftarrow$ NORMALIZE($\mathbf{fv}[i] \times \mathbf{b}$)
    $\mathbf{b} \leftarrow$ BACKWARD($\mathbf{b}, \mathbf{ev}[i]$)
  return sv

Smoothing: Example

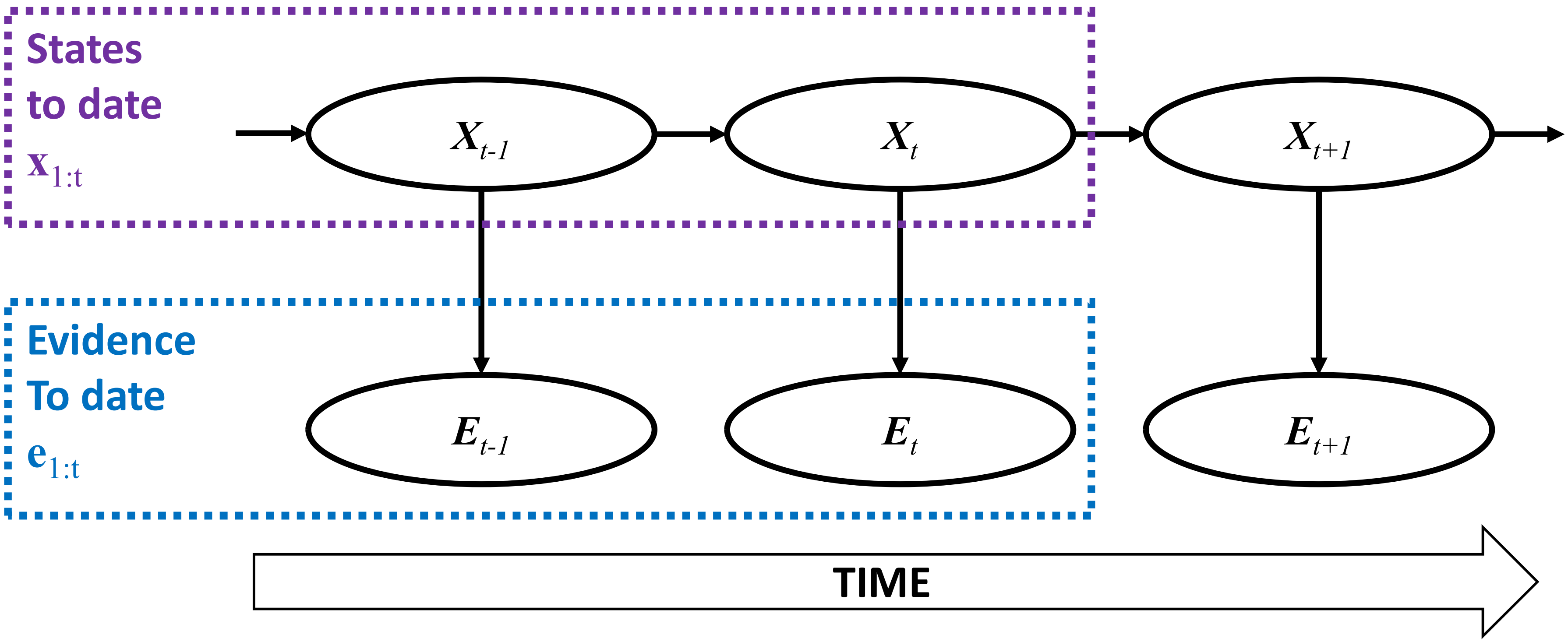


Most Likely Explanation
Hidden Markov Model
+
Viterbi Algorithm

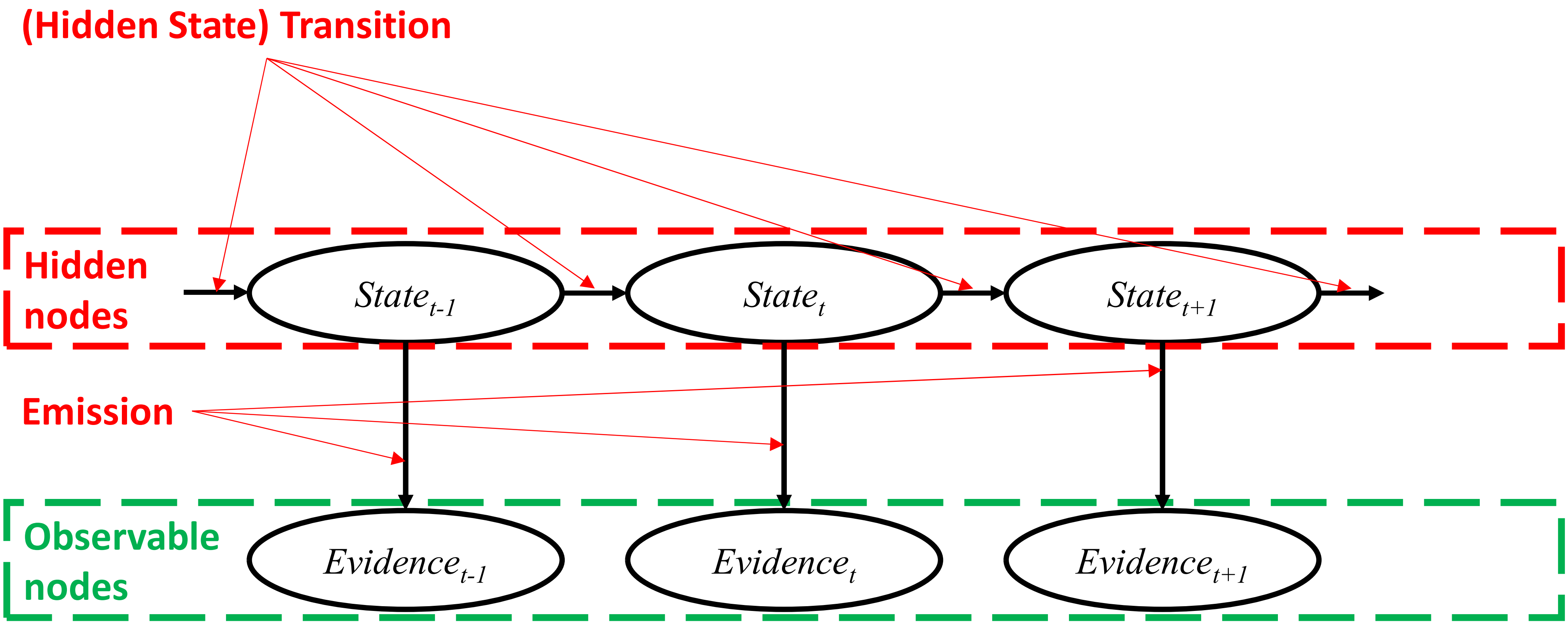
Inference: Most Likely Explanation

Given a sequence of observations, we might wish to find the sequence of states that is most likely to have generated those observations.

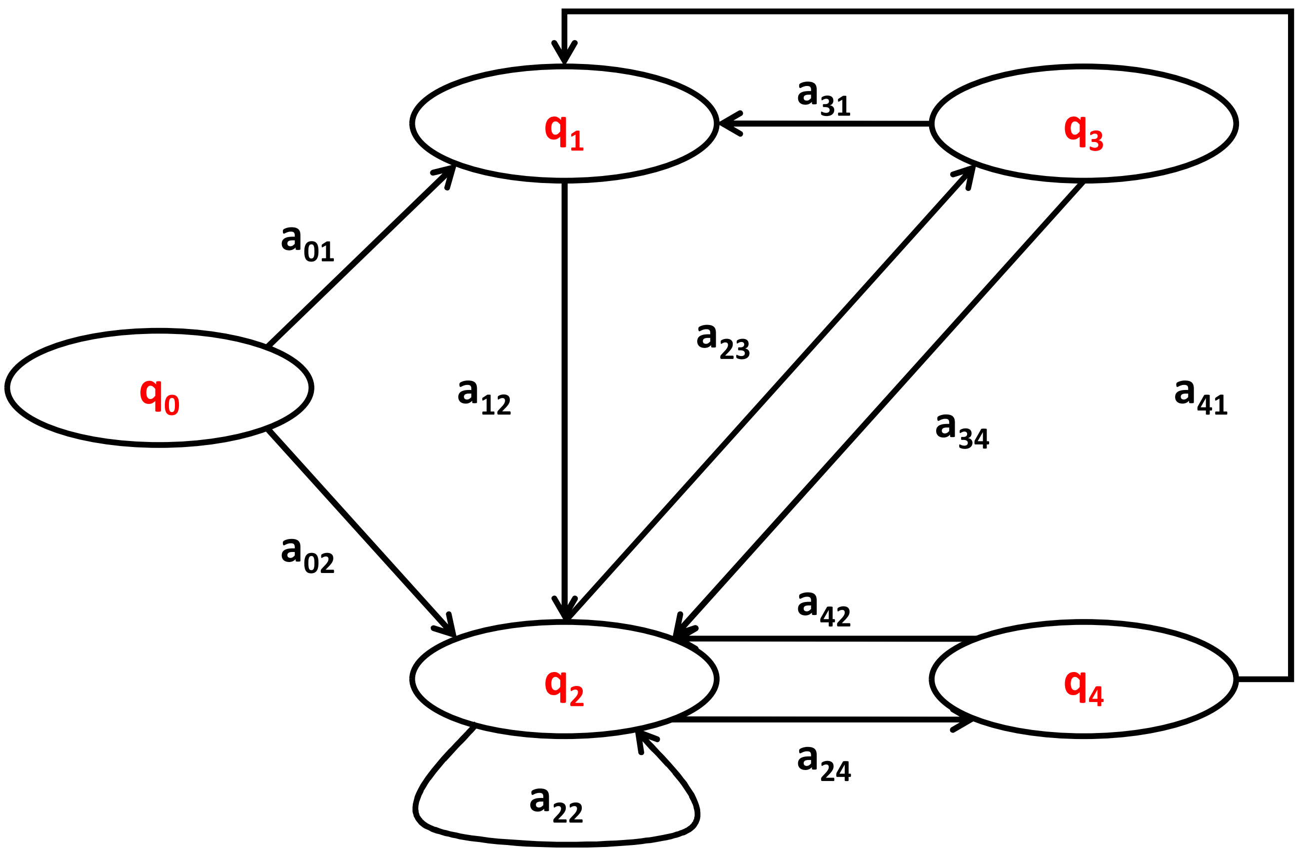
$$\operatorname{argmax}_{\mathbf{x}_{1:t}} P(\mathbf{x}_{1:t} \mid \mathbf{e}_{1:t})$$



Transition and Sensor Models



Hidden Markov Model



Transition probability matrix A						
	q_0	q_1	q_2	q_3	q_4	Notes
q_0	$a_{0,0}$	$a_{0,1}$	$a_{0,2}$	$a_{0,3}$	$a_{0,4}$	row sum = 1
q_1	$a_{1,0}$	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	row sum = 1
q_2	$a_{2,0}$	$a_{2,1}$	$a_{2,2}$	$a_{2,3}$	$a_{2,4}$	row sum = 1
q_3	$a_{3,0}$	$a_{3,1}$	$a_{3,2}$	$a_{3,3}$	$a_{3,4}$	row sum = 1
q_4	$a_{4.,0}$	$a_{4,1}$	$a_{4,2}$	$a_{4,3}$	$a_{4,4}$	row sum = 1

HMMs are specified with:

- A set of N states:

  $Q = \{q_1, q_2, \ldots, q_N\}$

- A transition probability matrix A, where each $a_{i,j}$ represents the probability of moving from state q_i to state q_j
- A sequence of T observations O:

  $O = o_1, o_2, \ldots, o_T$

- A sequence of observation likelihoods (emission probabilities): probability of observation o_t being generated by a state q_i:

  $B = b_i(o_t)$

- Special start (<s>) and end (final: not here) states

  q_0 and q_F

Hidden Markov Models: Decoding

The task of determining which sequence of variables is the underlying source of some sequence of observations is called the decoding:

Given as input an HMM $\alpha = (A, B)$ *and a sequence of observations* $o_1, o_2, \ldots, o_T$ *find the most probable sequence of states* $q_1, q_2, \ldots, q_T$.

or in our case:

Given as input an HMM $\alpha = (A, B)$ *and a sequence of* ***words*** $w_1, w_2, \ldots, w_T$ *find the most probable sequence of* ***tags/states*** $C_1, C_2, \ldots, C_T$.

A - transition probabilities matrix

B - emission probabilities matrix

Viterbi Algorithm: Pseudocode

function VITERBI(*observations* of len T,*state-graph* of len N) **returns** *best-path*, *path-prob*

create a path probability matrix *viterbi[N,T]*
for each state s **from** 1 **to** N **do** ; initialization step
    $viterbi[s,1] \leftarrow \pi_s * b_s(o_1)$
    $backpointer[\mathrm{s},1] \leftarrow 0$
for each time step t **from** 2 **to** T **do** ; recursion step
  for each state s **from** 1 **to** N **do**
    $$viterbi[\mathrm{s,t}] \leftarrow \max_{s'=1}^{N} \; viterbi[s',t-1] * a_{s',s} * b_s(o_t)$$
    $$backpointer[\mathrm{s,t}] \leftarrow \operatorname{argmax}_{s'=1}^{N} \; viterbi[s',t-1] * a_{s',s} * b_s(o_t)$$
$$bestpathprob \leftarrow \max_{s=1}^{N} \; viterbi[s,T]$$; termination step
$$bestpathpointer \leftarrow \operatorname{argmax}_{s=1}^{N} \; viterbi[s,T]$$; termination step
$bestpath \leftarrow$ the path starting at state *bestpathpointer*, that follows backpointer[] to states back in time
return *bestpath*, *bestpathprob*

Summary

Inference: Tasks and Applications

Filtering: $\mathbf{P}(\mathbf{X}_t|\mathbf{e}_{1:t})$
belief state—input to the decision process of a rational agent

Prediction: $\mathbf{P}(\mathbf{X}_{t+k}|\mathbf{e}_{1:t})$ for $k > 0$
evaluation of possible action sequences;
like filtering without the evidence

Smoothing: $\mathbf{P}(\mathbf{X}_k|\mathbf{e}_{1:t})$ for $0 \leq k < t$
better estimate of past states, essential for learning

Most likely explanation: $\arg\max_{\mathbf{x}_{1:t}} P(\mathbf{x}_{1:t}|\mathbf{e}_{1:t})$
speech recognition, decoding with a noisy channel