Fall 2022 | Lecture 14

Hidden Markov Models

Ariel Procaccia | Harvard University

MARKOV CHAINS

Suppose the weather evolves according to the following Bayes net (with $t = sun$):

$\boldsymbol{P}[X_0]$
0.5

X_0	$\boldsymbol{P}[X_1]$
t	.9
f	.3

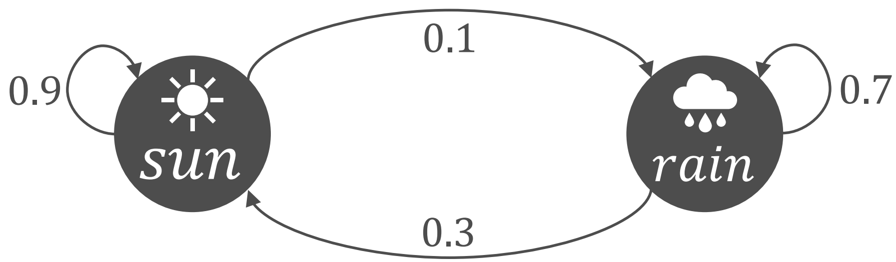
X_1	$\boldsymbol{P}[X_2]$
t	.9
f	.3

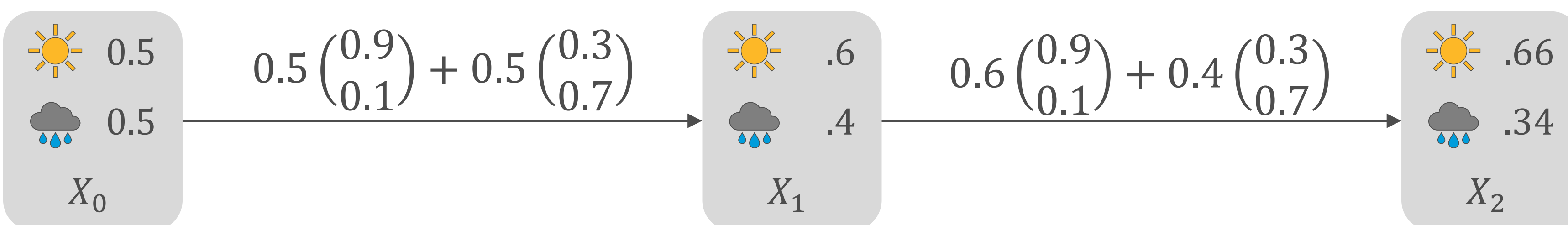
X_2	$\boldsymbol{P}[X_3]$
t	.9
f	.3

ASSUMPTIONS

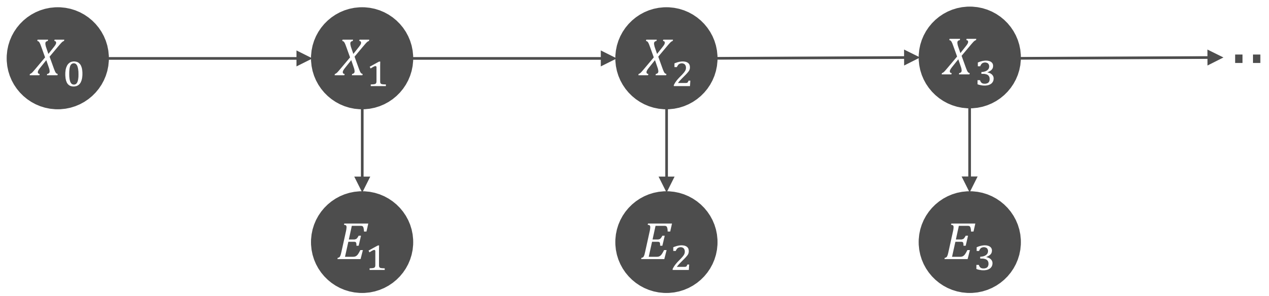
- We will think of infinite processes described by random variables $X_0, X_1, \ldots$
- We use $\boldsymbol{X}_{0:t}$ to denote $X_0, \ldots, X_t$
- Our simple Bayes net is assumed to satisfy:
  - Markov assumption:
    $$\boldsymbol{P}[X_t \mid \boldsymbol{X}_{0:t-1}] = \boldsymbol{P}[X_t \mid X_{t-1}]$$
  - Stationarity assumption: for all t, t',
    $$\boldsymbol{P}[X_t \mid X_{t-1}] = \boldsymbol{P}[X_{t'} \mid X_{t'-1}]$$

PREDICTING THE WEATHER





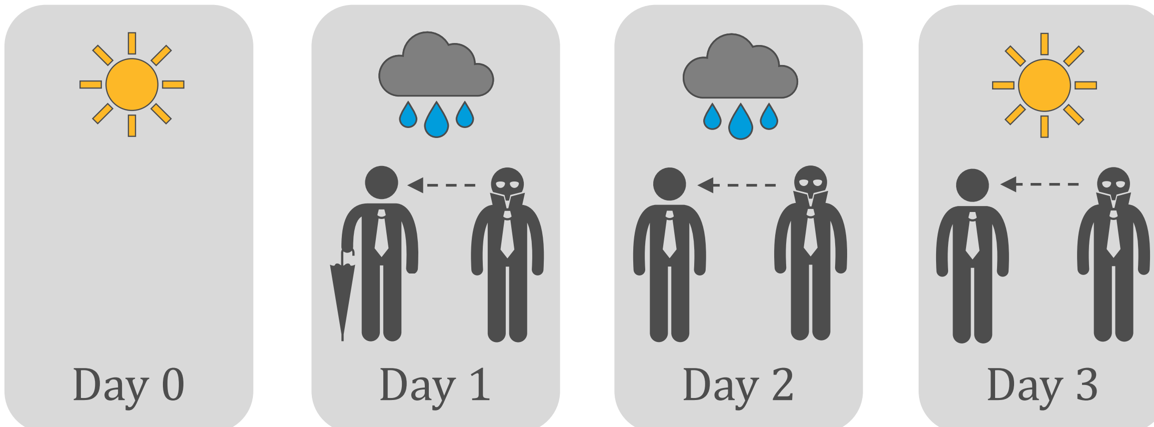
HIDDEN MARKOV MODELS



- Sometimes we can't directly observe the state of the world, but rather only observe evidence
- We are given $\boldsymbol{P}[E_t \mid \boldsymbol{X}_{0:t}, \boldsymbol{E}_{0:t-1}]$
- A hidden Markov model satisfies the same assumptions as before, plus the Markov sensor assumption:

$$\boldsymbol{P}[E_t \mid \boldsymbol{X}_{0:t}, \boldsymbol{E}_{0:t-1}] = \boldsymbol{P}[E_t \mid X_t]$$

HIDDEN MARKOV MODELS: EXAMPLE

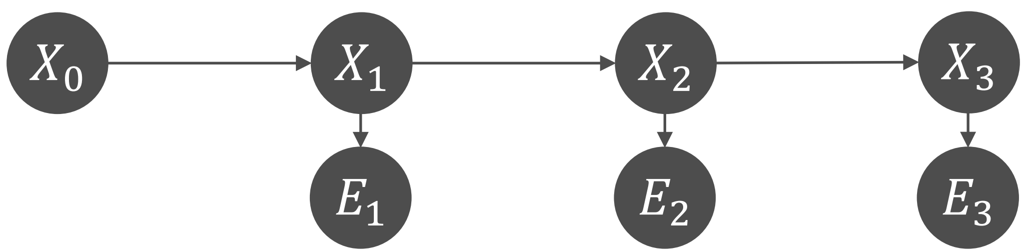


$\boldsymbol{P}[X_0]$
0.5

X_0	$\boldsymbol{P}[X_1]$
t	.9
f	.3

X_1	$\boldsymbol{P}[X_2]$
t	.9
f	.3

X_2	$\boldsymbol{P}[X_3]$
t	.9
f	.3



X_1	$\boldsymbol{P}[E_1]$
t	.2
f	.9

X_2	$\boldsymbol{P}[E_2]$
t	.2
f	.9

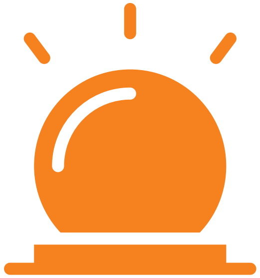
X_3	$\boldsymbol{P}[E_3]$
t	.2
f	.9

INFERENCE

Discuss

Filtering

$\boldsymbol{P}[X_t \mid \boldsymbol{e}_{1:t}]$



Follows from filtering

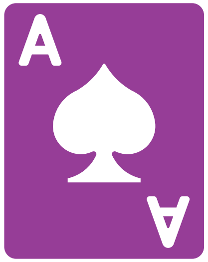
Prediction

$\boldsymbol{P}[X_{t+k} \mid \boldsymbol{e}_{1:t}]$

Skip

Smoothing

$\boldsymbol{P}[X_k \mid \boldsymbol{e}_{1:t}], k < t$



Discuss

Max likelihood

$\text{argmax}_{\boldsymbol{x}_{0:t}} \Pr[\boldsymbol{x}_{0:t} \mid \boldsymbol{e}_{1:t}]$

FILTERING

- We want an iterative algorithm that would compute $\boldsymbol{P}[X_{t+1} \mid \boldsymbol{e}_{1:t+1}]$ given e_{t+1} and our previous calculation of $\boldsymbol{P}[X_t \mid \boldsymbol{e}_{1:t}]$
- Given $\boldsymbol{e}_{1:t}$ we will first compute $\boldsymbol{P}[X_1 \mid e_1]$, use that estimate to compute $\boldsymbol{P}[X_2 \mid \boldsymbol{e}_{1:2}]$ based on e_2, and so on

FILTERING

$$\boldsymbol{P}[X_{t+1} \mid \boldsymbol{e}_{1:t+1}] = \boldsymbol{P}[X_{t+1} \mid \boldsymbol{e}_{1:t}, e_{t+1}]$$

Bayes' Rule

$$\propto \boldsymbol{P}[e_{t+1} \mid X_{t+1}, \boldsymbol{e}_{1:t}] \cdot \boldsymbol{P}[X_{t+1} \mid \boldsymbol{e}_{1:t}]$$

Conditional independence

$$= \boldsymbol{P}[e_{t+1} \mid X_{t+1}] \cdot \boldsymbol{P}[X_{t+1} \mid \boldsymbol{e}_{1:t}]$$

Condition on x_t

$$= \boldsymbol{P}[e_{t+1} \mid X_{t+1}] \cdot \sum_{x_t} \Pr[x_t \mid \boldsymbol{e}_{1:t}] \cdot \boldsymbol{P}[X_{t+1} \mid x_t, \boldsymbol{e}_{1:t}]$$

Conditional independence

$$= \underbrace{\boldsymbol{P}[e_{t+1} \mid X_{t+1}]}_{\text{Given}} \cdot \sum_{x_t} \underbrace{\Pr[x_t \mid \boldsymbol{e}_{1:t}]}_{\text{Previously computed}} \cdot \underbrace{\boldsymbol{P}[X_{t+1} \mid x_t]}_{\text{Given}}$$

FILTERING: EXAMPLE

$$\boldsymbol{P}[X_{t+1} \mid \boldsymbol{e}_{1:t+1}] \propto \boldsymbol{P}[e_{t+1} \mid X_{t+1}] \cdot \sum_{x_t} \Pr[x_t \mid \boldsymbol{e}_{1:t}] \cdot \boldsymbol{P}[X_{t+1} \mid x_t]$$

In our weather example:

$\boldsymbol{P}[X_0]$
0.5

X_0	$\boldsymbol{P}[X_1]$
t	.9
f	.3

X_1	$\boldsymbol{P}[E_1]$
t	.2
f	.9

Suppose the piece of evidence on Day 1 is that the director is carrying an umbrella ($E_1 = t$) then

$$\boldsymbol{P}[X_1 = t \mid E_1 = t] \propto 0.2 \cdot (0.5 \cdot 0.9 + 0.5 \cdot 0.3) = 0.12$$
$$\boldsymbol{P}[X_1 = f \mid E_1 = t] \propto 0.9 \cdot (0.5 \cdot 0.1 + 0.5 \cdot 0.7) = 0.36$$

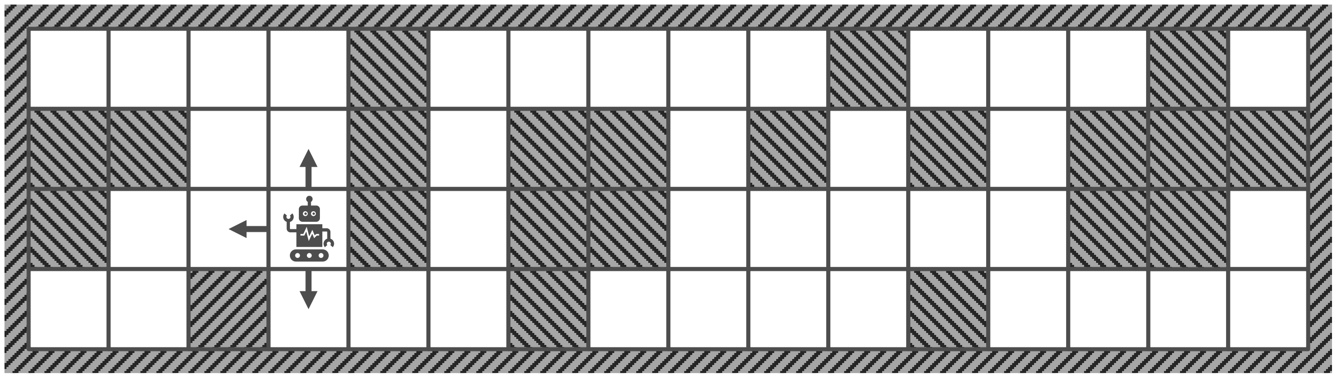
By normalizing we get a prediction of (0.25,0.75)

FILTERING: RUNNING TIME

$$\boldsymbol{P}[X_{t+1} \mid \boldsymbol{e}_{1:t+1}] \propto \boldsymbol{P}[e_{t+1} \mid X_{t+1}] \cdot \sum_{x_t} \Pr[x_t \mid \boldsymbol{e}_{1:t}] \cdot \boldsymbol{P}[X_{t+1} \mid x_t]$$

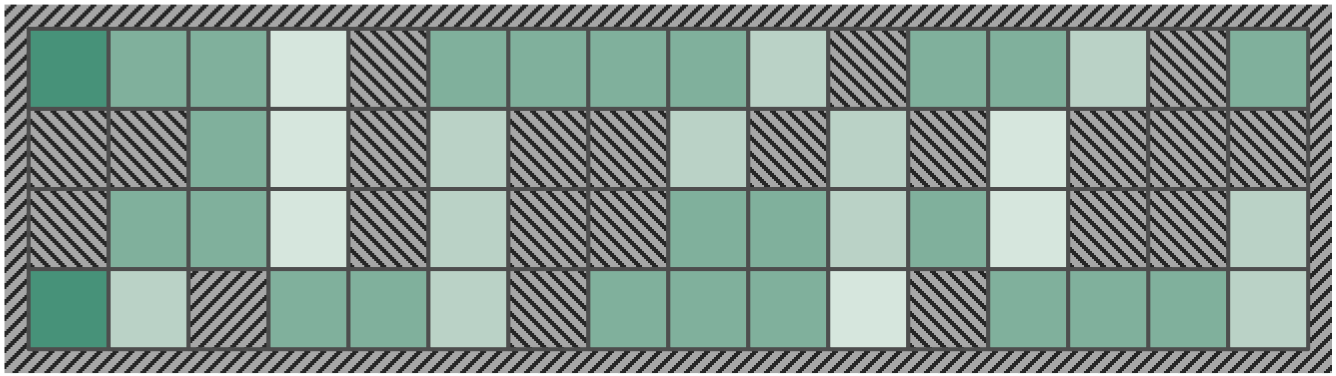
- Poll 1: What is the running time of computing $\boldsymbol{P}[X_t \mid e_{1:t}]$ as a function of t?
  - $\Theta(1)$
  - $\Theta(t)$
  - $\Theta(t^2)$
  - $\Theta(t^3)$

ROBOT LOCALIZATION AS HMM

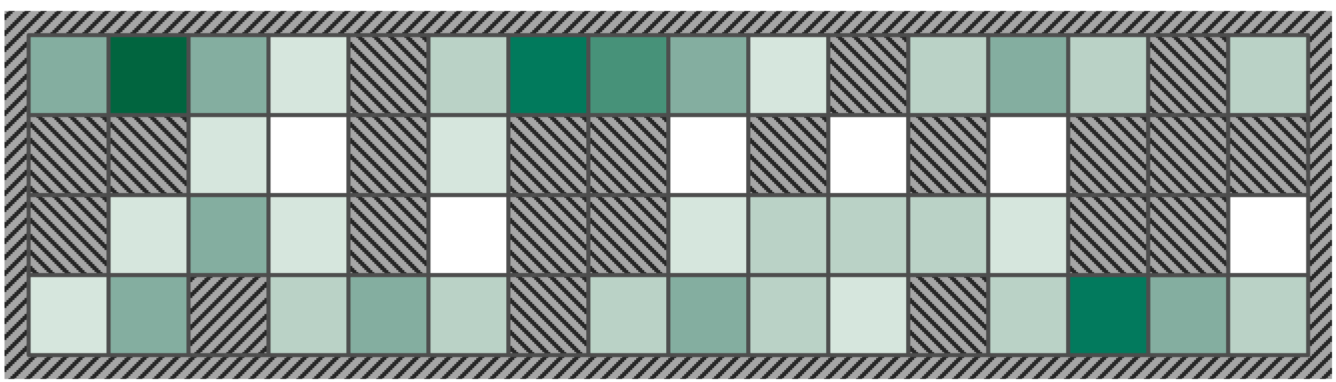


- State X_t is position of robot on grid
- In each step the robot moves in a random unblocked direction
- The robot starts in a random cell
- E_t consists of a string of four bits that indicate obstacles in N, E, S, W
- The sensor error rate for each bit is 0.2

ROBOT LOCALIZATION AS HMM

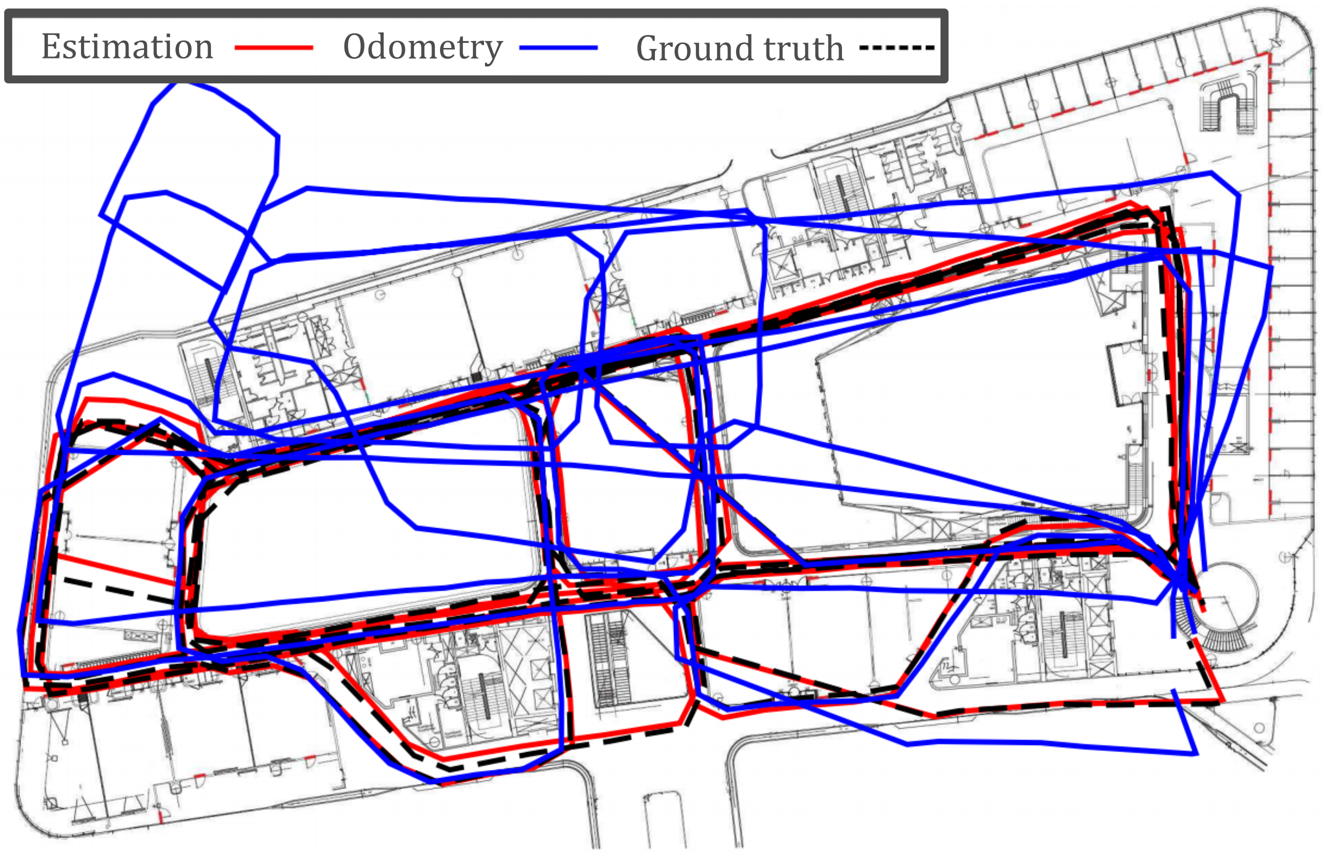


Posterior distribution over robot location after $E_1 = 1011$



Posterior distribution over robot location after $E_1 = 1011, E_2 = 1010$

LOCALIZATION IN THE REAL WORLD



[Liu et al., 2019]

MAXIMUM LIKELIHOOD

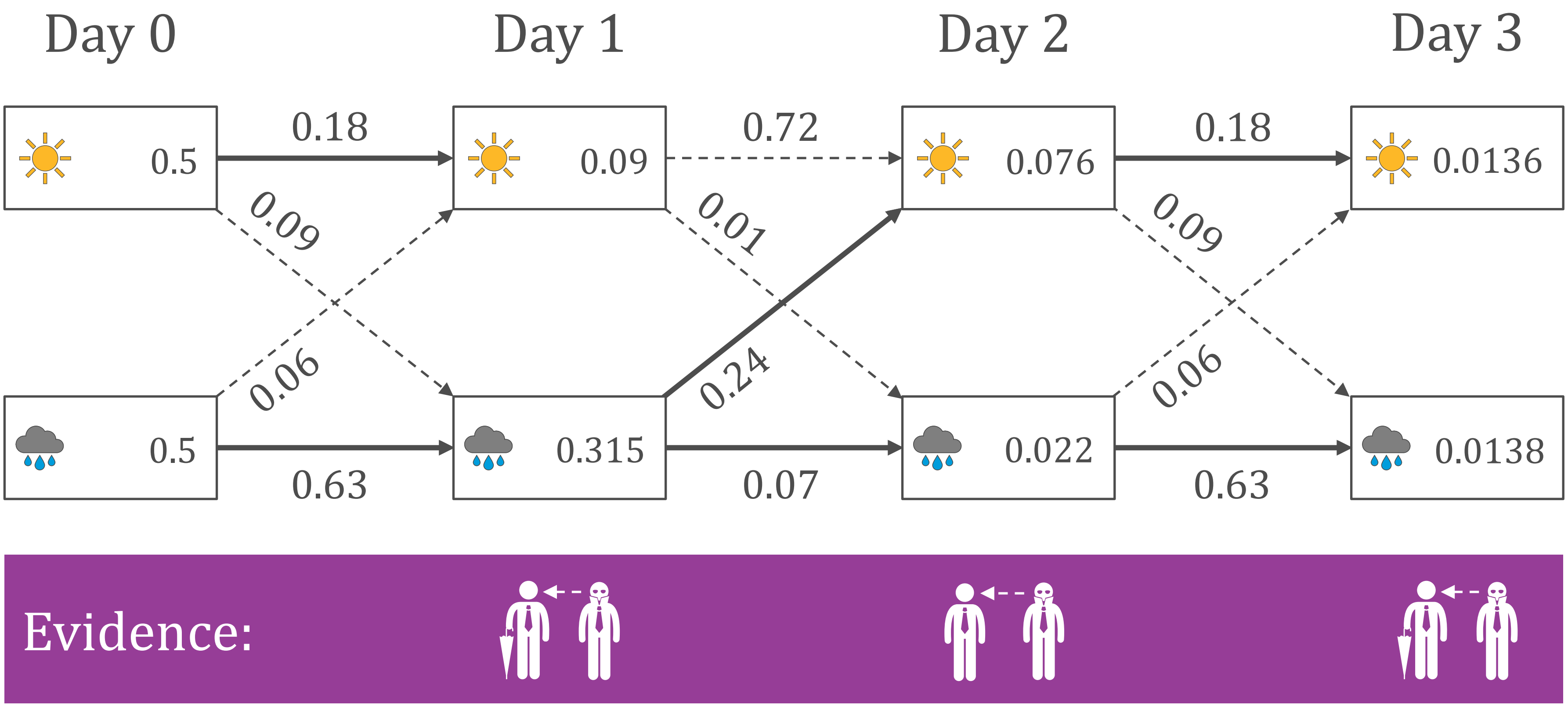
- We want an iterative algorithm to compute
$$\text{argmax}_{\boldsymbol{x}_{0:t}} \boldsymbol{P}[\boldsymbol{x}_{0:t}, X_{t+1} \mid \boldsymbol{e}_{1:t+1}]$$
given e_{t+1} and our previous computation of
$$\text{argmax}_{\boldsymbol{x}_{0:t-1}} \boldsymbol{P}[\boldsymbol{x}_{0:t-1}, X_t \mid \boldsymbol{e}_{1:t}]$$
- By a calculation similar to the one we did for filtering, we can show that
$$\max_{\boldsymbol{x}_{0:t}} \boldsymbol{P}[\boldsymbol{x}_{0:t}, X_{t+1} \mid \boldsymbol{e}_{1:t+1}]$$
$$\propto \boldsymbol{P}[e_{t+1} \mid X_{t+1}] \max_{x_t} \boldsymbol{P}[X_{t+1} \mid x_t] \max_{\boldsymbol{x}_{0:t-1}} \Pr[\boldsymbol{x}_{0:t} \mid \boldsymbol{e}_{1:t}]$$
- Andrew Viterbi showed in 1967 how to compute this efficiently

VITERBI: EXAMPLE

Previously computed

$$\max_{\boldsymbol{x}_{0:t}} \boldsymbol{P}[\boldsymbol{x}_{0:t}, X_{t+1} \mid \boldsymbol{e}_{1:t+1}] \propto \boldsymbol{P}[e_{t+1} \mid X_{t+1}] \max_{x_t} \boldsymbol{P}[X_{t+1} \mid x_t] \max_{\boldsymbol{x}_{0:t-1}} \Pr[\boldsymbol{x}_{0:t} \mid \boldsymbol{e}_{1:t}]$$

The weight on each edge is $\Pr[e_{t+1} \mid x_{t+1}] \cdot \Pr[x_{t+1} \mid x_t]$



VITERBI: RUNNING TIME

- Poll 2: What is the running time of the Viterbi Algorithm as a function of t?
  - $\Theta(1)$
  - $\Theta(t)$
  - $\Theta(t^2)$
  - $\Theta(t^3)$

APPLICATION: POS TAGGING