CS 156:Introduction to Artificial Intelligence

Instructor: Dr. Sayma Akther

San José State University

Issues

- If a state is described by n propositions, then a belief state contains 2^n states
  - $\rightarrow$ Modeling difficulty: many numbers must be entered in the first place
  - $\rightarrow$ Computational issue: memory size and time

	Toothache		¬Toothache	
	PCatch	¬PCatch	PCatch	¬PCatch
Cavity	0.108	0.012	0.072	0.008
¬Cavity	0.016	0.064	0.144	0.576

- Bayesian networks explicitly represent independence among propositions to reduce the number of probabilities defining a belief state

Bayesian Networks

- Two problems with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time

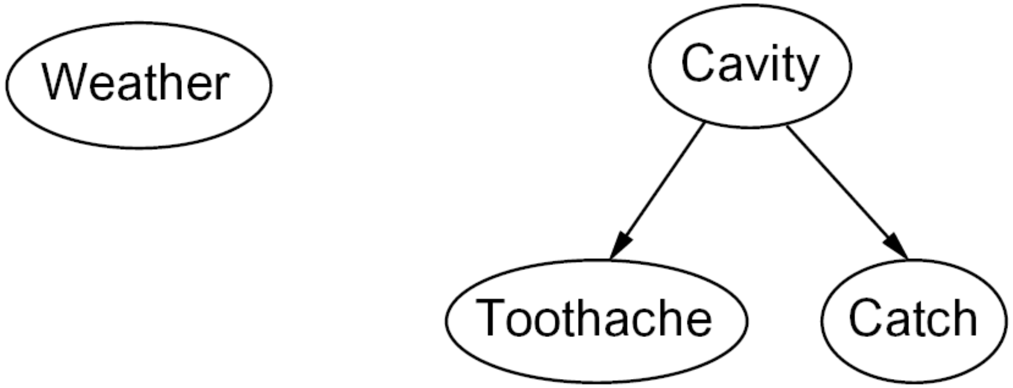
- Bayes' nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)

- Bayesian Networks, also known as Belief Networks, are graphical models that represent the probabilistic relationships among a set of variables.

- Representing and reasoning about uncertainty in AI.

Structure of Bayesian Networks

- **Nodes**: Represent random variables (can be observed or hidden).
- **Edges**: Direct arrows indicate a direct probabilistic influence or causality.
- **Directed Acyclic Graph (DAG)**: The structure ensures there are no closed loops or cycles.



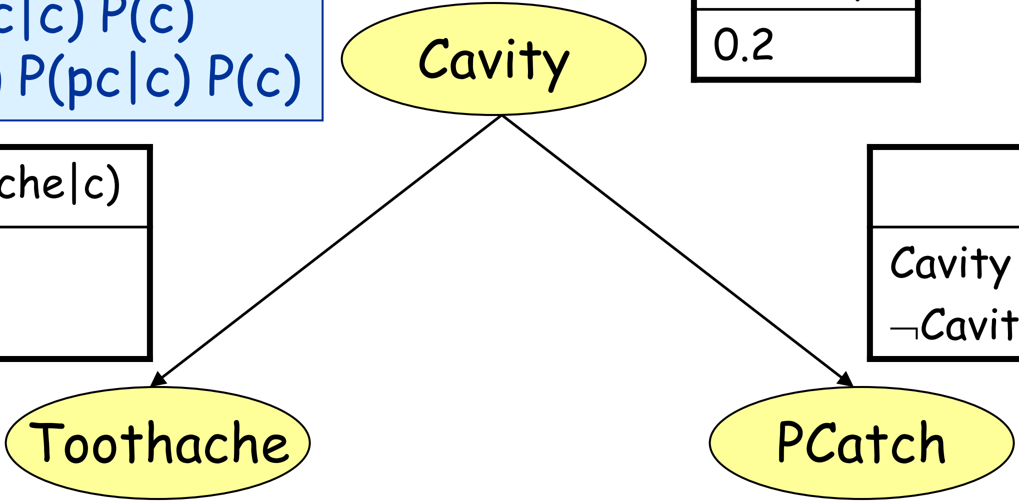
Local Conditional Probability Tables (CPTs)

- Each node has a **conditional probability table** (**CPT**) that gives the probability of each of its values given every possible combination of values for its parents (conditioning case).
  - – Roots (sources) of the DAG that have no parents are given prior probabilities.

- Notice that Cavity is the "cause" of both Toothache and PCatch, and represent the causality links explicitly
- Give the prior probability distribution of Cavity
- Give the conditional probability tables of Toothache and PCatch

$$
\begin{aligned}
P(c \wedge t \wedge pc) &= P(t \wedge pc \mid c)\, P(c) \\
&= P(t \mid c)\, P(pc \mid c)\, P(c)
\end{aligned}
$$

P(Cavity)
0.2



	P(Toothache|c)
Cavity	0.6
¬Cavity	0.1

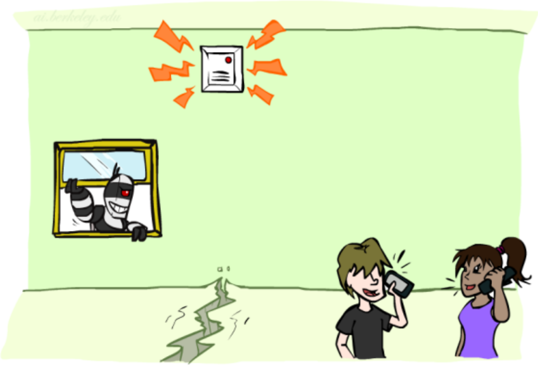
	P(PCatch|c)
Cavity	0.9
¬Cavity	0.02

A More Complex BN

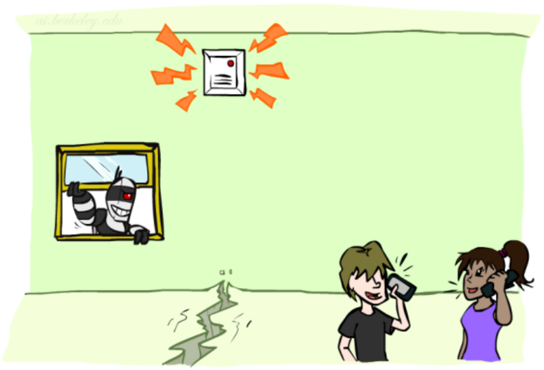
Imagine you have a burglar alarm installed at your home. There are a few potential causes for the alarm to ring: there might be a burglar, or maybe there was just an earthquake that set it off. You also have two neighbors, John and Mary, who promised to call you at work if they hear the alarm. John always listens to loud music and might miss the alarm, while Mary might misinterpret the phone ringing as the alarm. This situation sets up a network of conditional dependencies which can be represented as a Bayesian Network.

- What are the random variables?

  Burglary, Earthquake, Alarm, JohnCalls, MaryCalls



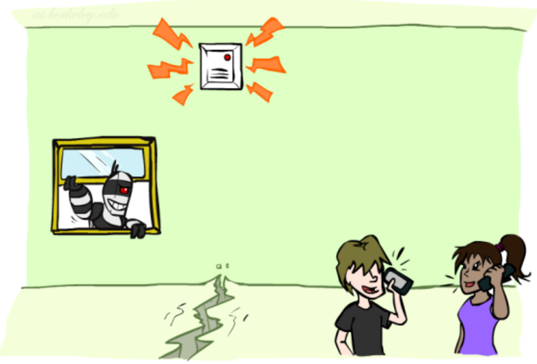
A More Complex BN

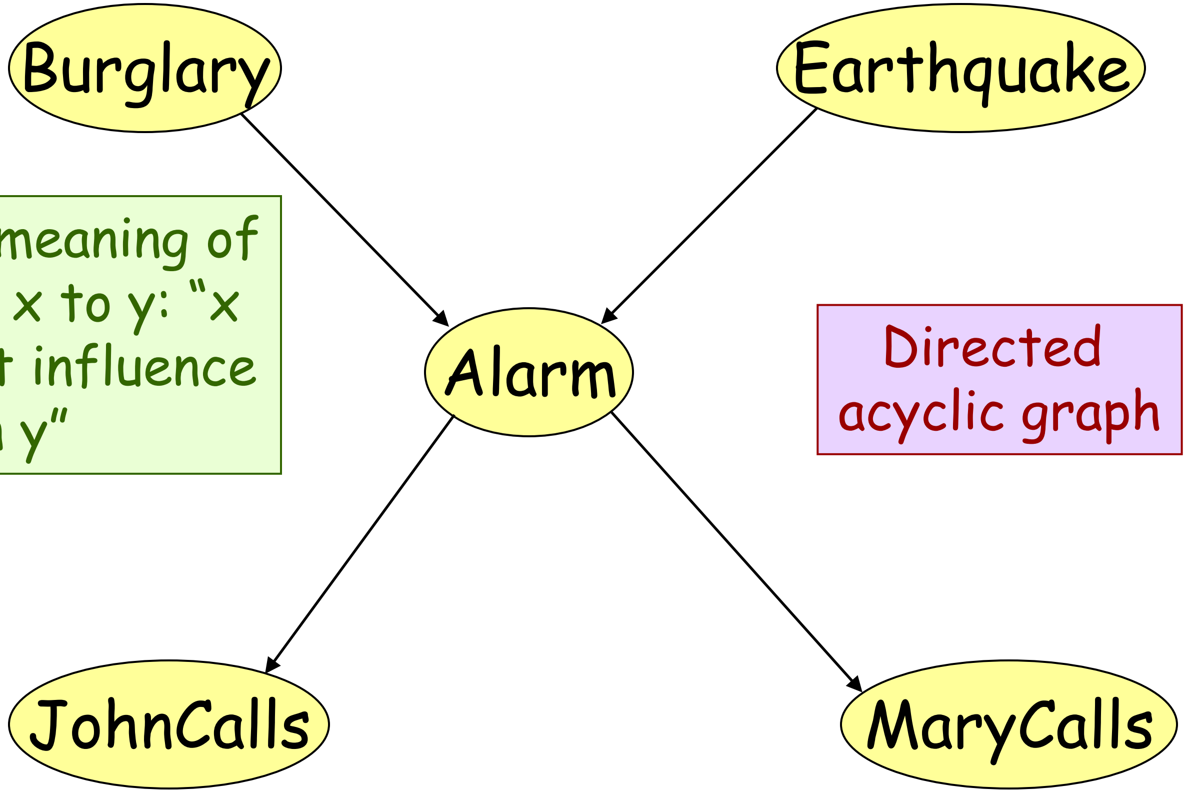


- What are the random variables?
  - Burglary, Earthquake, Alarm, JohnCalls, MaryCalls

- Nodes:
  - Burglar: Represents whether a burglar is present (True or False).
  - Earthquake: Represents whether there was an earthquake (True or False).
  - Alarm: Represents whether the alarm goes off (True or False).
  - JohnCalls: Represents whether John calls (True or False).
  - MaryCalls: Represents whether Mary calls (True or False).

A More Complex BN





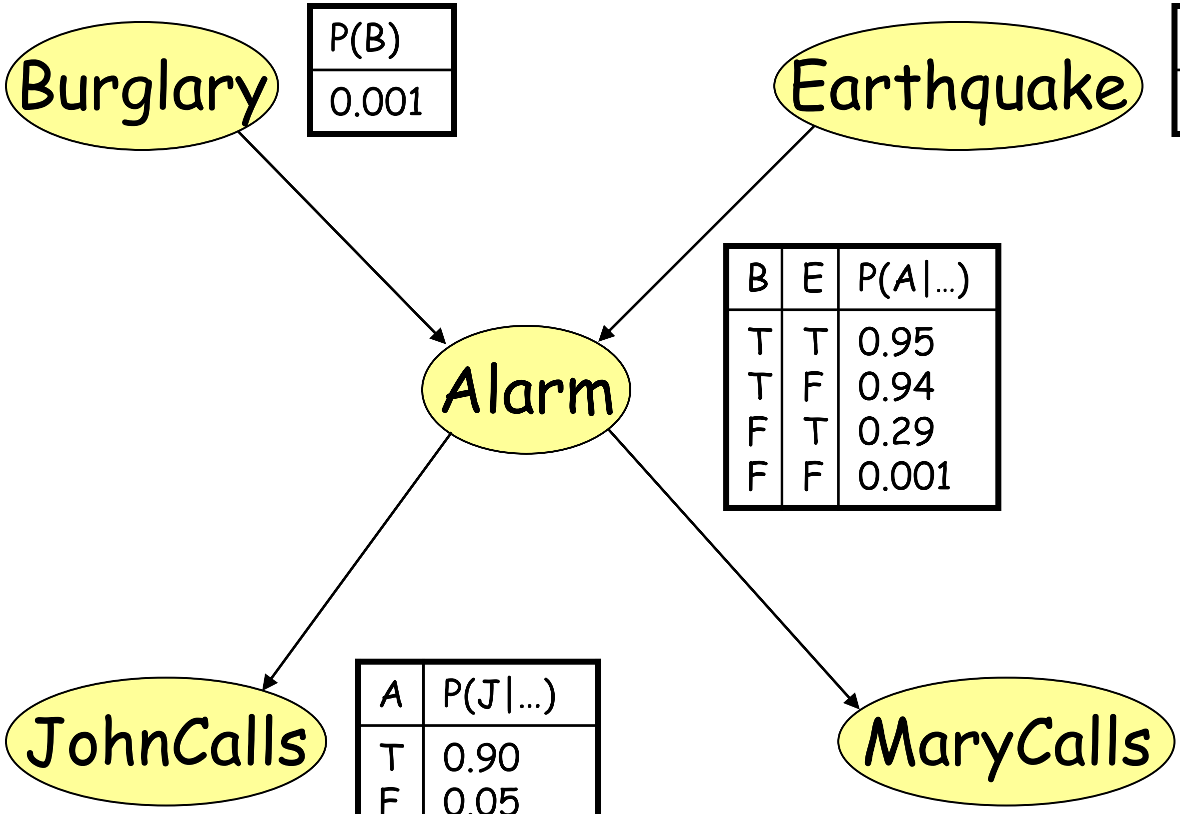
Intuitive meaning of arc from x to y: "x has direct influence on y"

Directed acyclic graph

causes

effects

A More Complex BN



P(B)
0.001

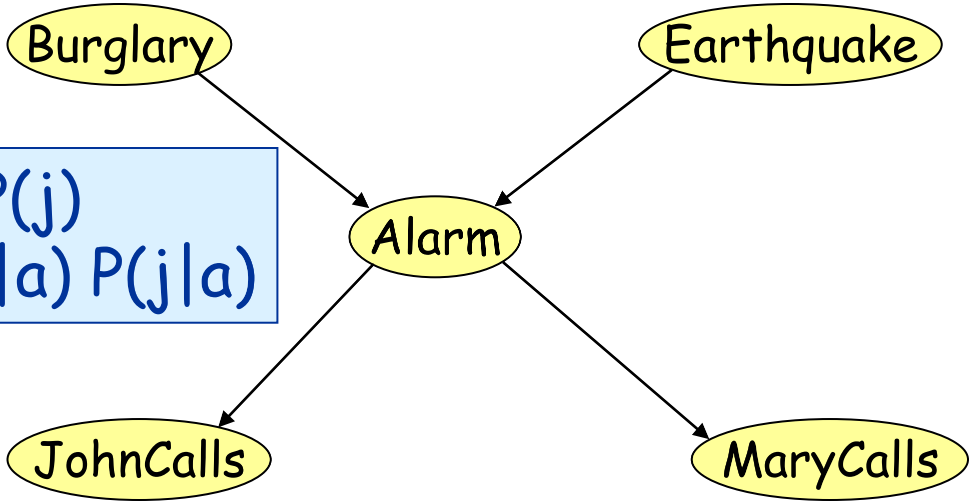
P(E)
0.002

B	E	P(A|…)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|…)
T	0.90
F	0.05

A	P(M|…)
T	0.70
F	0.01

A More Complex BN



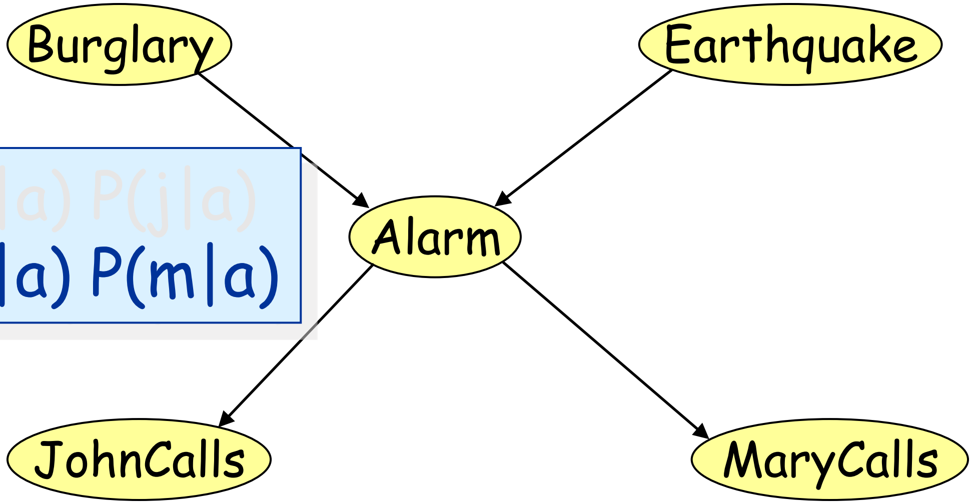
$P(b \wedge j) = P(b)\, P(j)$

$P(b \wedge j \mid a) = P(b \mid a)\, P(j \mid a)$

Each of the beliefs JohnCalls and MaryCalls is independent of Burglary and Earthquake given Alarm or $\neg$Alarm

For example, John does not observe any burglaries directly

A More Complex BN



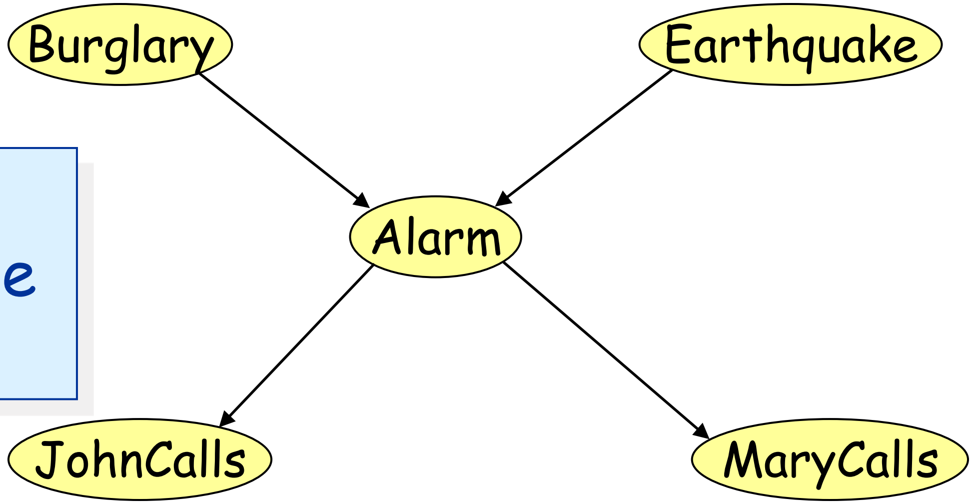
$P(b \wedge j|a) = P(b|a)\, P(j|a)$

$P(j \wedge m|a) = P(j|a)\, P(m|a)$

The beliefs JohnCalls and MaryCalls are independent given Alarm or ¬Alarm

For instance, the reasons why John and Mary may not call if there is an alarm are unrelated

A More Complex BN

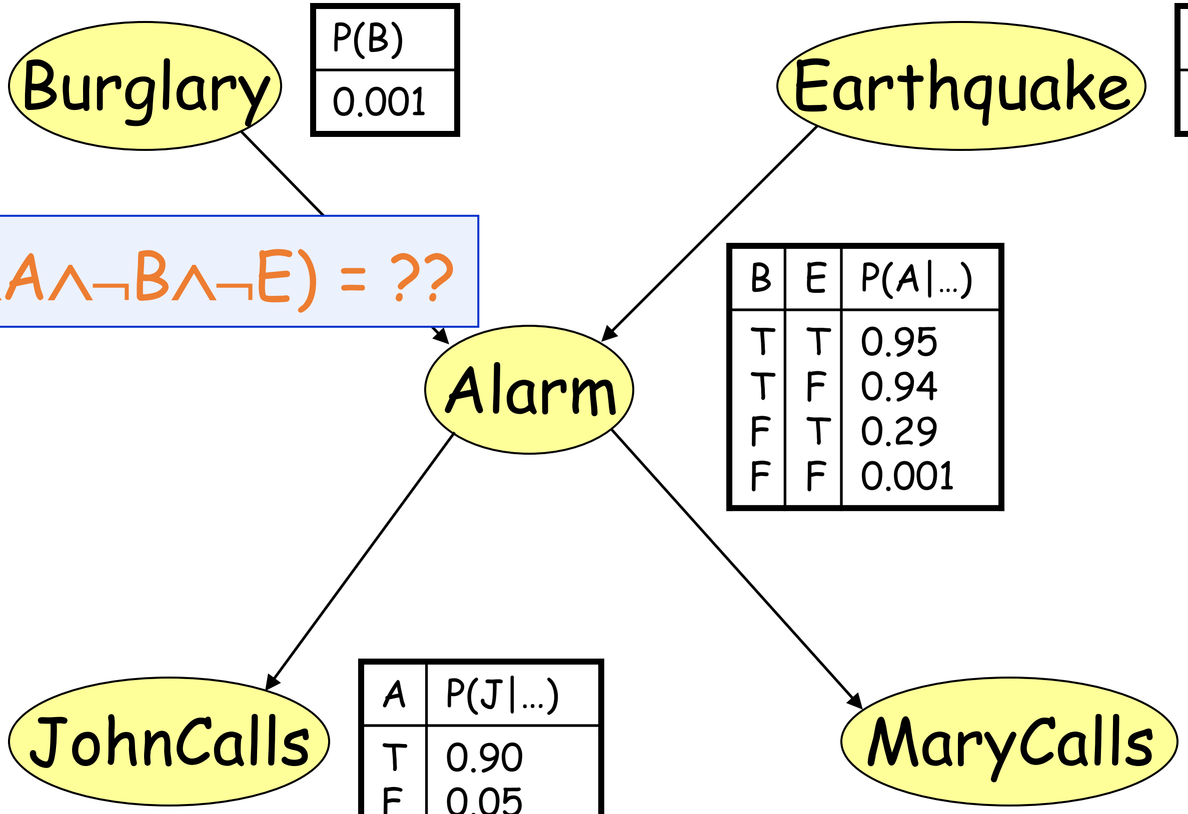


Burglary and Earthquake are independent

The beliefs JohnCalls and MaryCalls are independent given Alarm or ¬Alarm

For instance, the reasons why John and Mary may not call if there is an alarm are unrelated

Calculation of Joint Probability



$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = ??$

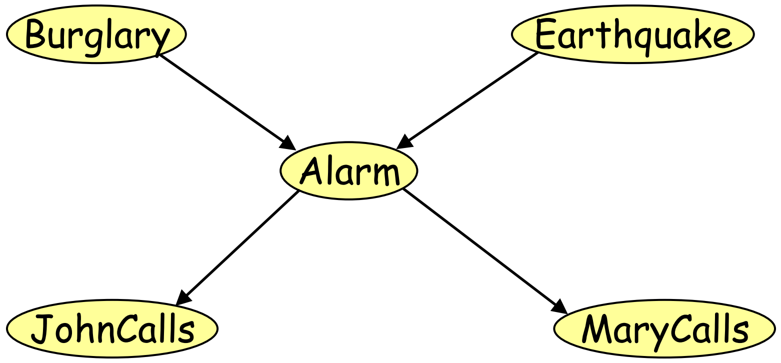
P(B)
0.001

P(E)
0.002

B	E	P(A|…)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

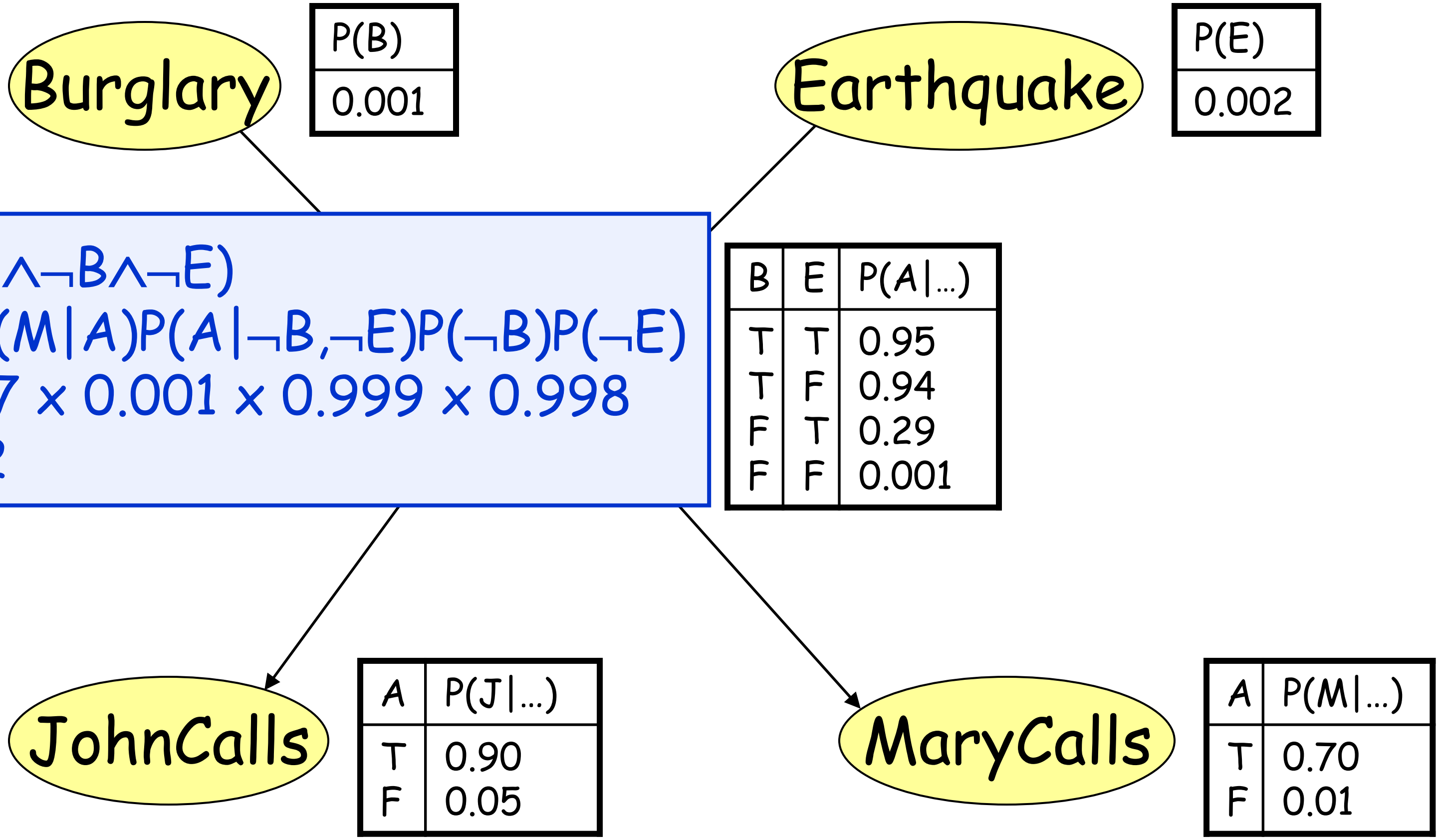
A	P(J|…)
T	0.90
F	0.05

A	P(M|…)
T	0.70
F	0.01



- $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$
  $= P(J \wedge M | A, \neg B, \neg E) \times P(A \wedge \neg B \wedge \neg E)$
  $= P(J | A, \neg B, \neg E) \times P(M | A, \neg B, \neg E) \times P(A \wedge \neg B \wedge \neg E)$
  (J and M are independent given A)
- $P(J | A, \neg B, \neg E) = P(J | A)$
  (J and $\neg B \wedge \neg E$ are independent given A)
- $P(M | A, \neg B, \neg E) = P(M | A)$
- $P(A \wedge \neg B \wedge \neg E) = P(A | \neg B, \neg E) \times P(\neg B | \neg E) \times P(\neg E)$
  $= P(A | \neg B, \neg E) \times P(\neg B) \times P(\neg E)$
  ($\neg B$ and $\neg E$ are independent)
- $P(J \wedge M \wedge A \wedge \neg B \wedge \neg E) = P(J | A) P(M | A) P(A | \neg B, \neg E) P(\neg B) P(\neg E)$

Calculation of Joint Probability



P(B)
0.001

P(E)
0.002

$P(J \wedge M \wedge A \wedge \neg B \wedge \neg E)$
$= P(J|A)P(M|A)P(A|\neg B,\neg E)P(\neg B)P(\neg E)$
$= 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998$
$= 0.00062$

B	E	P(A|…)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|…)
T	0.90
F	0.05

A	P(M|…)
T	0.70
F	0.01

Probability Recap

- Conditional probability

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

- Product rule

$$P(x,y) = P(x|y)P(y)$$

- Chain rule

$$\begin{aligned} P(X_1, X_2, \ldots X_n) &= P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots \\ &= \prod_{i=1}^{n} P(X_i|X_1, \ldots, X_{i-1}) \end{aligned}$$

- X, Y independent if and only if:

$$\forall x, y : P(x,y) = P(x)P(y)$$

- X and Y are conditionally independent given Z if and only if:

$$X \perp\!\!\!\perp Y | Z$$

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

Markov Blanket

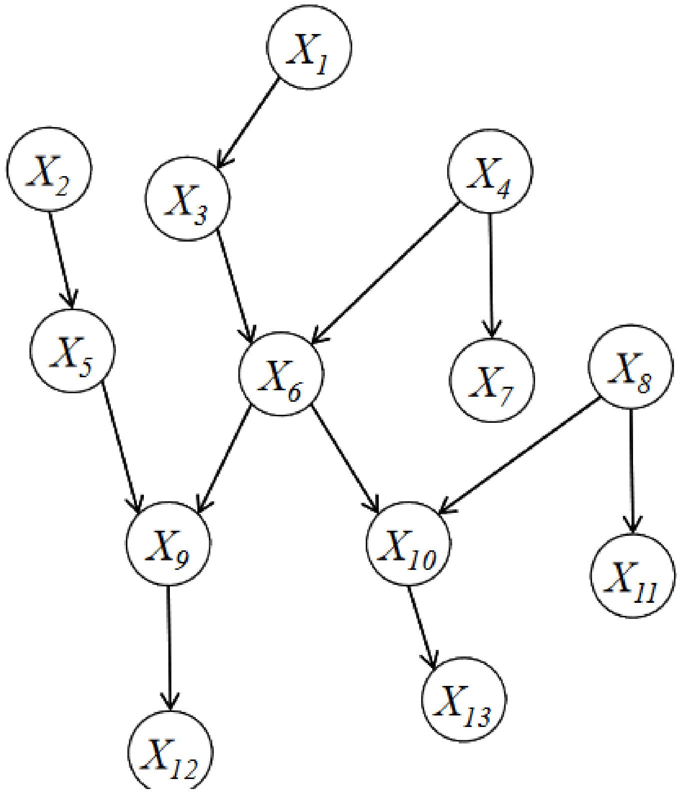
- The **Markov blanket** for a node in a graphical model contains all the variables that shield the node from the rest of the network. This means if you know the values of the variables in the Markov blanket, then the node is independent of all other variables in the network.
- In a Bayesian network (a type of directed graphical model), the Markov blanket for a node X includes:
  1. **Parents of X**
  2. **Children of X**
  3. **Co parents of X's children**

Markov Blanket

Def: the **co-parents** of a node are the parents of its children

Def: the **Markov Blanket** of a node is the set containing the node's parents, children, and co-parents.

Thm: a node is **conditionally independent** of every other node in the graph given its **Markov blanket**



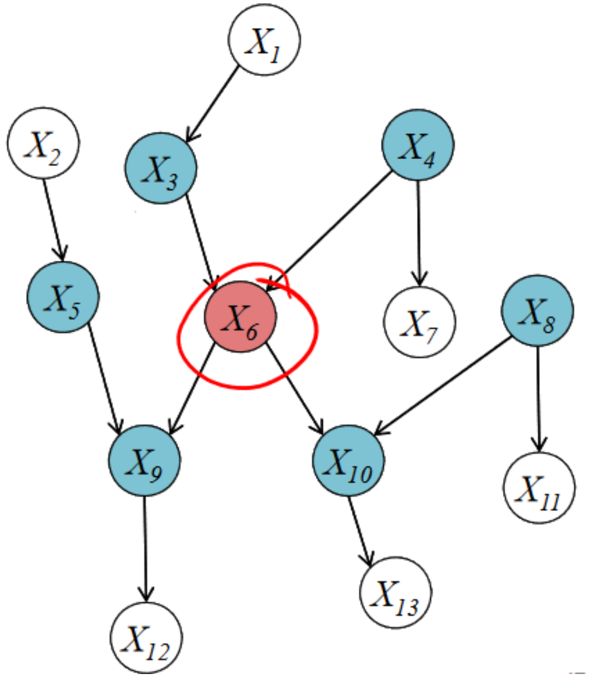
Markov Blanket

Def: the **co-parents** of a node are the parents of its children

Def: the **Markov Blanket** of a node is the set containing the node's parents, children, and co-parents.

Theorem: a node is **conditionally independent** of every other node in the graph given its **Markov blanket**

Example: The Markov Blanket of X_6 is $\{X_3, X_4, X_5, X_8, X_9, X_{10}\}$



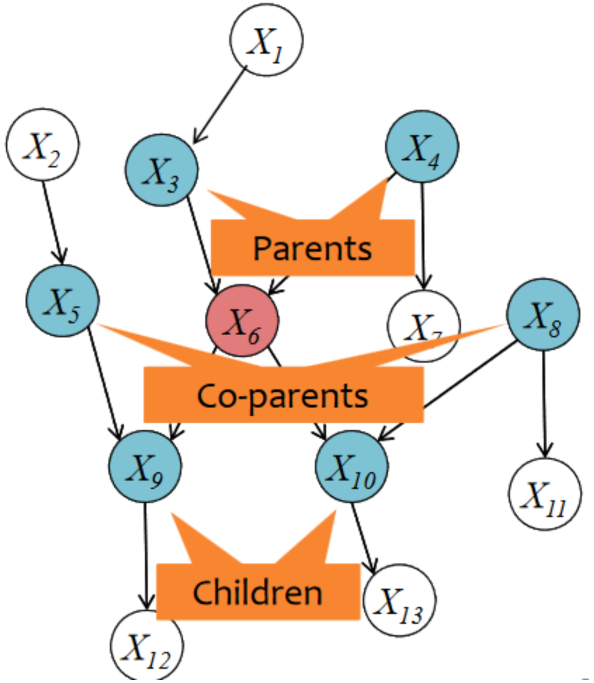
Markov Blanket

Def: the **co-parents** of a node are the parents of its children

Def: the **Markov Blanket** of a node is the set containing the node's parents, children, and co-parents.

Theorem: a node is **conditionally independent** of every other node in the graph given its **Markov blanket**

Example: The Markov Blanket of X_6 is $\{X_3, X_4, X_5, X_8, X_9, X_{10}\}$



Bayesian Networks and Markov Models

- Both Bayesian Networks (BNs) and Markov Models (specifically, Markov Chains or Hidden Markov Models) are graphical representations that help in understanding and calculating probabilities for a set of variables

Markov Models

- **Definition**: A stochastic model describing a sequence of events in which the probability of each event depends only on the state attained in the previous event.
- **Memoryless Property**: The future state of a process only depends on the current state and is independent of the past states.
- **Components of a Markov Model**
  - **States**: The distinct scenarios or configurations the model can exist in.
  - **Transition Probabilities**: Probabilities of moving from one state to another.

Markov Models

Types of Markov Models

- **Markov Chains**: The simplest form, where we can observe the state directly.
- **Hidden Markov Models (HMMs)**: The true state is hidden, but there's observable data that depends on the state.
- **Markov Decision Processes (MDPs)**: Like Markov Chains but with decisions and rewards.

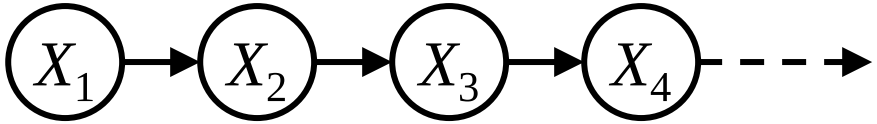
Applications of Markov Models

- **Weather Forecasting**: Predicting weather transitions (e.g., sunny to rainy).
- **Economics**: Modeling market transitions.
- **Biology**: Representing DNA sequences.
- **Speech Recognition (using HMMs)**.

Markov Chains

Definition: A sequence of random variables where the future variable is independent of the past variables given the present.

- Value of X at a given time is called the state



$$P(X_1) \qquad P(X_t|X_{t-1})$$

- Parameters: called transition probabilities or dynamics, specify how the state evolves over time (also, initial state probabilities)

~UC Berkeley

Joint Distribution of a Markov Model

$$X_1 \to X_2 \to X_3 \to X_4$$

$P(X_1)$ $\quad$ $P(X_t|X_{t-1})$

- Joint distribution:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

- More generally:

$$\begin{aligned} P(X_1, X_2, \ldots, X_T) &= P(X_1)P(X_2|X_1)P(X_3|X_2)\ldots P(X_T|X_{T-1}) \\ &= P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1}) \end{aligned}$$

Chain Rule and Markov Models

$$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4$$

- From the chain rule, every joint distribution over X_1, X_2, X_3, X_4 can be written as:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)$$

- Assuming that

$$X_3 \perp\!\!\!\perp X_1 \mid X_2 \quad \text{and} \quad X_4 \perp\!\!\!\perp X_1, X_2 \mid X_3$$

results in the expression posited on the previous slide:

$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_2)P(X_4|X_3)$$

Chain Rule and Markov Models

$$X_1 \rightarrow X_2 \rightarrow X_3 \rightarrow X_4 \dashrightarrow$$

- From the chain rule, every joint distribution over $X_1, X_2, \ldots, X_T$ can be written as:

$$P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=2}^{T} P(X_t | X_1, X_2, \ldots, X_{t-1})$$

- Assuming that for all *t*:

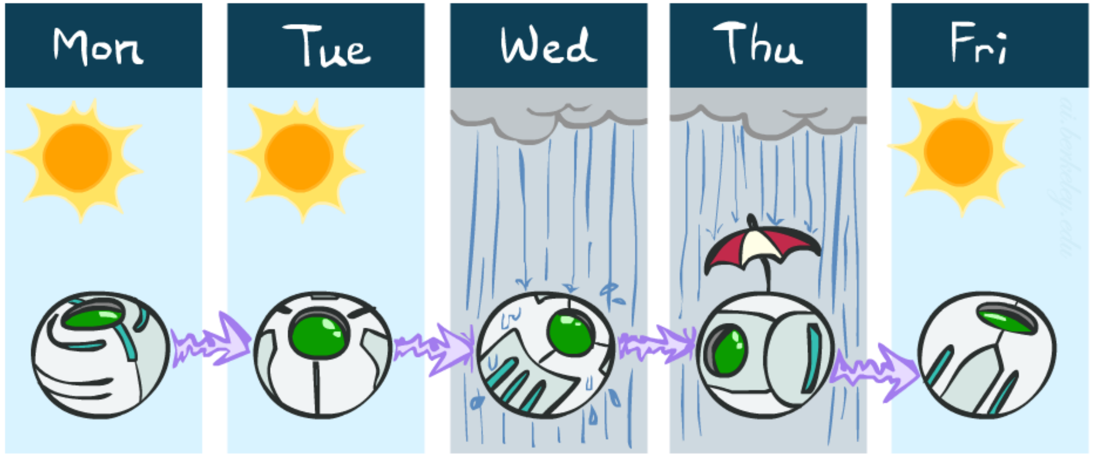
$$X_t \perp\!\!\!\perp X_1, \ldots, X_{t-2} \mid X_{t-1}$$

gives us the expression posited on the earlier slide:

$$P(X_1, X_2, \ldots, X_T) = P(X_1) \prod_{t=2}^{T} P(X_t | X_{t-1})$$

Example Markov Chain: Weather

- States: $X = \{\text{rain}, \text{sun}\}$

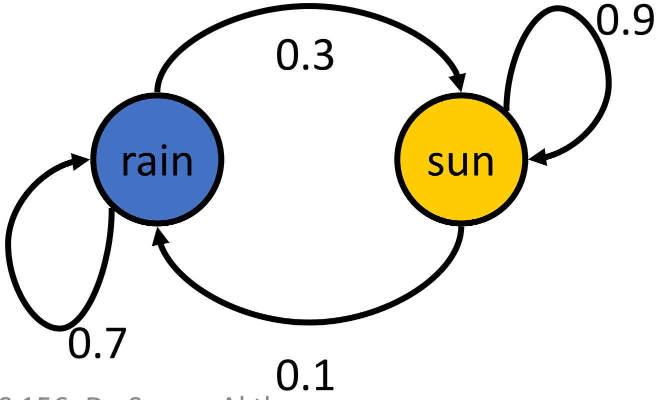


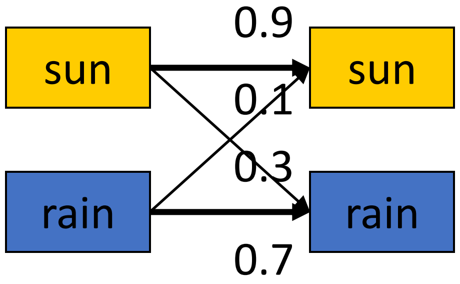
- Initial distribution: 1.0 sun

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

X_{t-1}	X_t	$P(X_t \mid X_{t-1})$
sun	sun	0.9
sun	rain	0.1
rain	sun	0.3
rain	rain	0.7

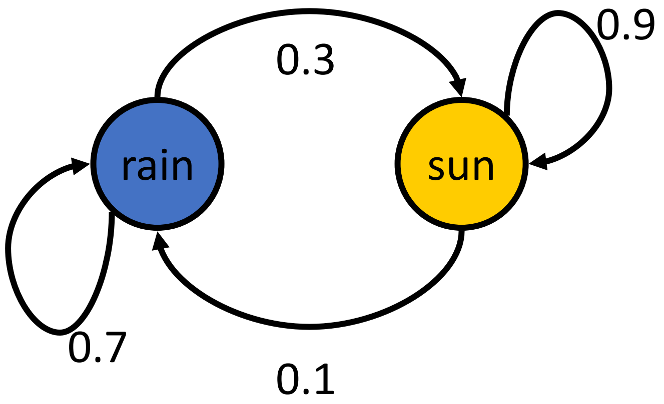
Two new ways of representing the same CPT





Example Markov Chain: Weather

- Initial distribution: 1.0 sun



- What is the probability distribution after one step?

$$P(X_2 = \text{sun}) = \quad P(X_2 = \text{sun}|X_1 = \text{sun})P(X_1 = \text{sun}) + P(X_2 = \text{sun}|X_1 = \text{rain})P(X_1 = \text{rain})$$

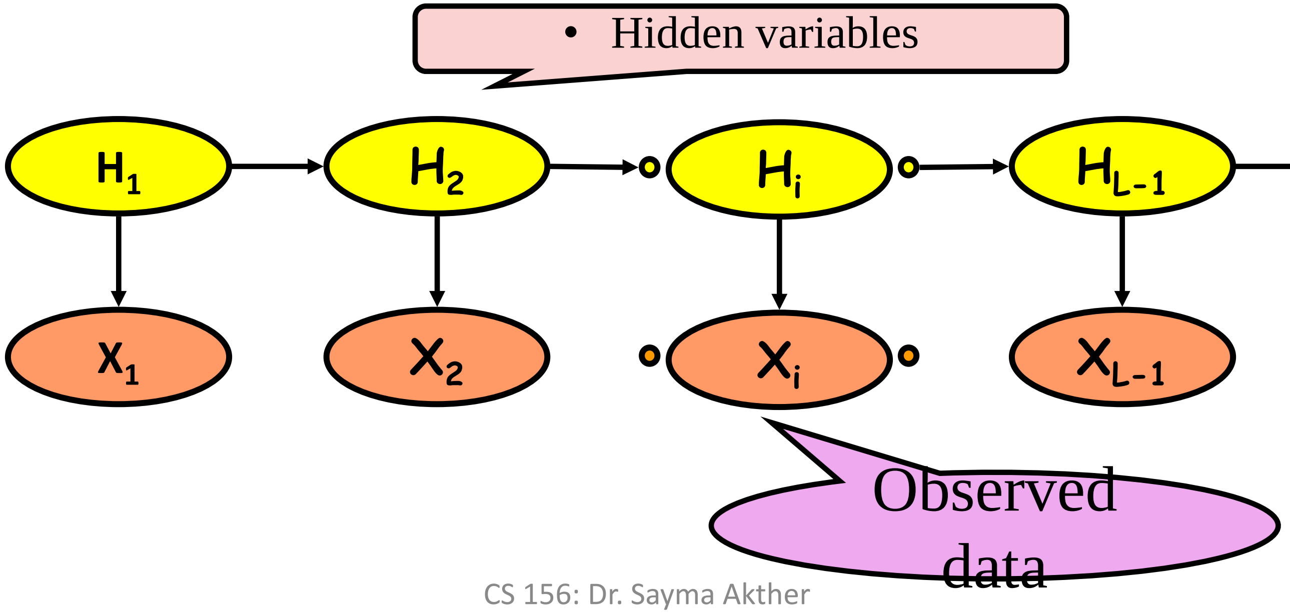
$$0.9 \cdot 1.0 + 0.3 \cdot 0.0 = 0.9$$

Hidden Markov Models

- A **Hidden Markov Model (HMM)** is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved states.
- HMMs can be thought of as a generalization of mixture models where the **hidden variables (or latent variables)** have temporal properties.
- Widely used in applications such as speech recognition, bioinformatics, and finance.

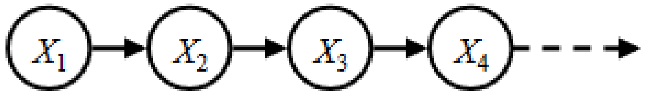
Structure of an HMM:

- **States**: A finite set of N states. These states are "hidden" meaning they are not directly observable.
- **Observations**: A set of M possible observations, which can be seen.
- **State Transition Probabilities**: Probabilities of transitioning from one state to another.
- **Emission Probabilities**: The probability of an observation being generated from a state.
- **Initial State Probabilities**: Probabilities regarding which state the model is in when it starts.



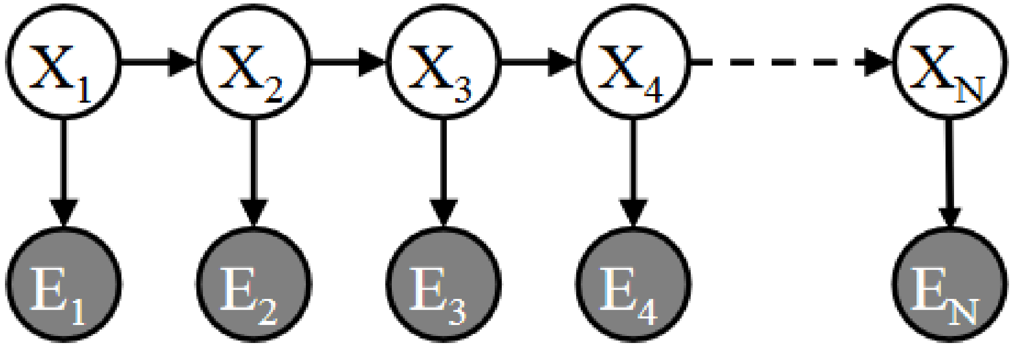
Markov Chain and Hidden Markov Models

- Markov chains not so useful for most agents
  - Eventually you don't know anything anymore
  - Need observations to update your beliefs



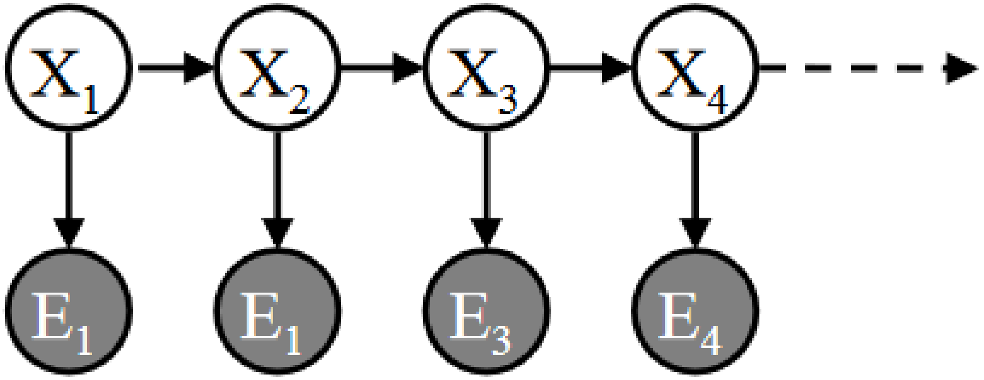
$P(X_1)$ $\qquad P(X|X_{-1})$

- Hidden Markov models (HMMs)
  - Underlying Markov chain over states S
  - You observe outputs (effects) at each time step



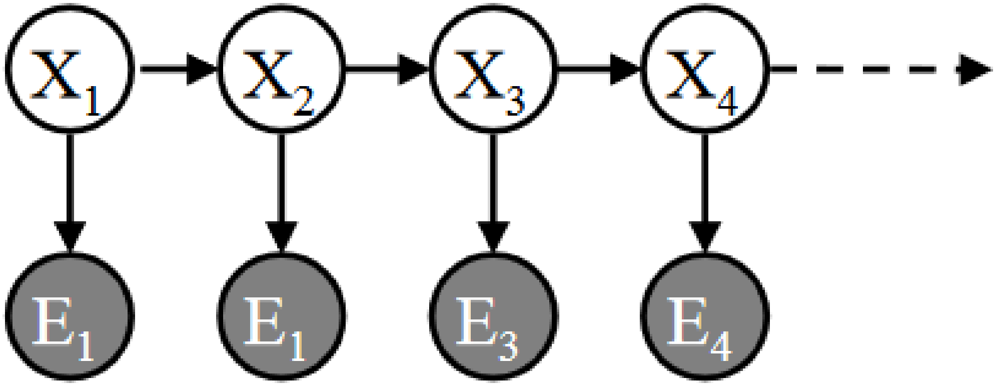
Real HMM Examples

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

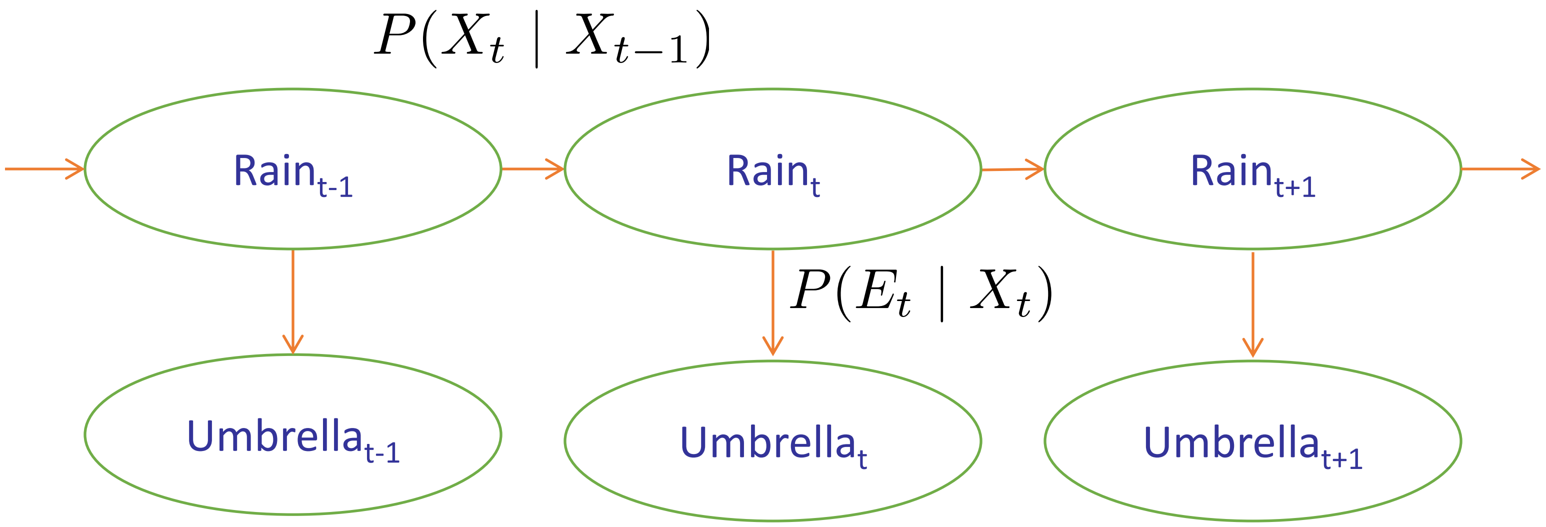


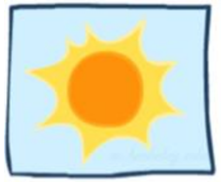
Real HMM Examples

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



Example: Weather HMM



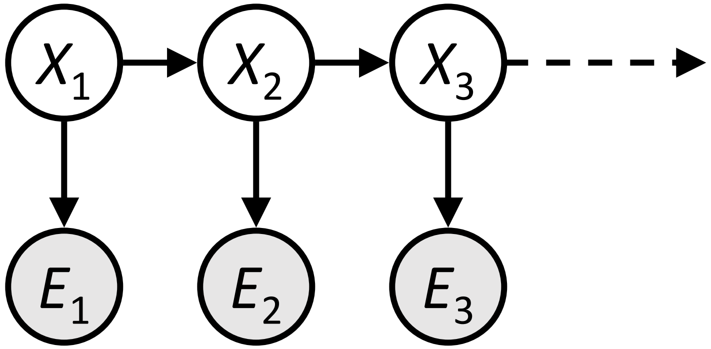


- An HMM is defined by:
  - Initial distribution: $P(X_1)$
  - Transitions: $P(X_t \mid X_{t-1})$
  - Emissions: $P(E_t \mid X_t)$

R_t	R_{t+1}	$P(R_{t+1} \mid R_t)$
+r	+r	0.7
+r	-r	0.3
-r	+r	0.3
-r	-r	0.7

R_t	U_t	$P(U_t \mid R_t)$
+r	+u	0.9
+r	-u	0.1
-r	+u	0.2
-r	-u	0.8

Joint Distribution of an HMM



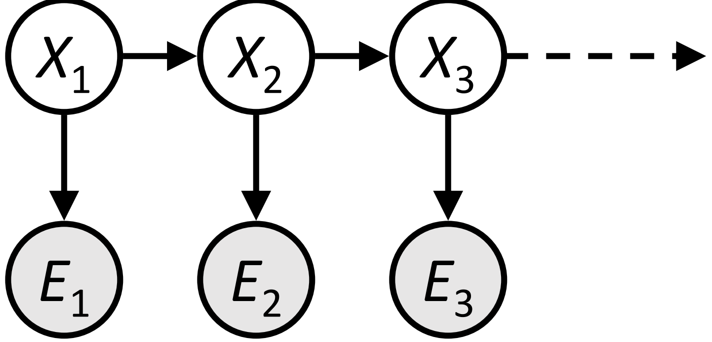
- Joint distribution:

$$P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1|X_1)P(X_2|X_1)P(E_2|X_2)P(X_3|X_2)P(E_3|X_3)$$

- More generally:

$$P(X_1, E_1, \ldots, X_T, E_T) = P(X_1)P(E_1|X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})P(E_t|X_t)$$

Chain Rule and HMMs



- From the chain rule, *every* joint distribution over $X_1, E_1, X_2, E_2, X_3, E_3$ can be written as:

$$P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1|X_1)P(X_2|X_1, E_1)P(E_2|X_1, E_1, X_2)$$
$$P(X_3|X_1, E_1, X_2, E_2)P(E_3|X_1, E_1, X_2, E_2, X_3)$$

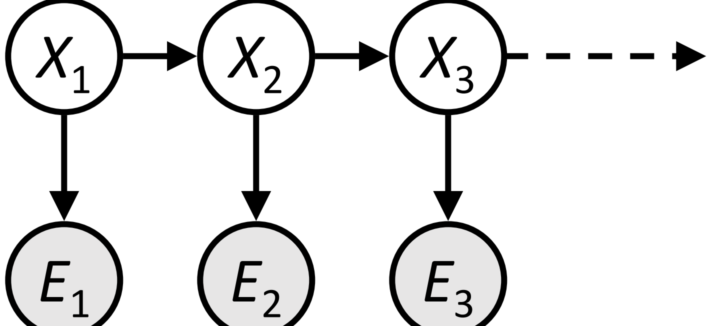
- *Assuming* that

$$X_2 \perp\!\!\!\perp E_1 \mid X_1, \quad E_2 \perp\!\!\!\perp X_1, E_1 \mid X_2, \quad X_3 \perp\!\!\!\perp X_1, E_1, E_2 \mid X_2, \quad E_3 \perp\!\!\!\perp X_1, E_1, X_2, E_2 \mid X_3$$

gives us the expression posited on the previous slide:

$$P(X_1, E_1, X_2, E_2, X_3, E_3) = P(X_1)P(E_1|X_1)P(X_2|X_1)P(E_2|X_2)P(X_3|X_2)P(E_3|X_3)$$

Chain Rule and HMMs



- From the chain rule, *every* joint distribution over $X_1, E_1, \ldots, X_T, E_T$ can be written as:

$$P(X_1, E_1, \ldots, X_T, E_T) = P(X_1)P(E_1|X_1) \prod_{t=2}^{T} P(X_t|X_1, E_1, \ldots, X_{t-1}, E_{t-1})P(E_t|X_1, E_1, \ldots, X_{t-1}, E_{t-1}, X_t)$$

- *Assuming* that for all *t*:
  - State independent of all past states and all past evidence given the previous state, i.e.:

$$X_t \perp\!\!\!\perp X_1, E_1, \ldots, X_{t-2}, E_{t-2}, E_{t-1} \mid X_{t-1}$$

  - Evidence is independent of all past states and all past evidence given the current state, i.e.:

$$E_t \perp\!\!\!\perp X_1, E_1, \ldots, X_{t-2}, E_{t-2}, X_{t-1}, E_{t-1} \mid X_t$$

gives us the expression posited on the earlier slide:

$$P(X_1, E_1, \ldots, X_T, E_T) = P(X_1)P(E_1|X_1) \prod_{t=2}^{T} P(X_t|X_{t-1})P(E_t|X_t)$$