



Lecture Slides for

INTRODUCTION TO

# Machine Learning

## 2nd Edition

ETHEM ALPAYDIN

© The MIT Press, 2010

*alpaydin@boun.edu.tr*

*<http://www.cmpe.boun.edu.tr/~ethem/i2ml2e>*

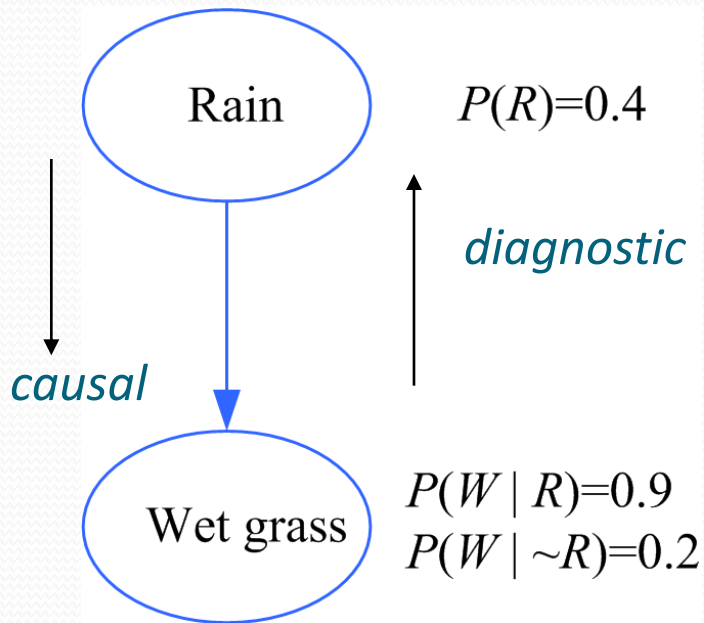
CHAPTER 16:

# Graphical Models

# Graphical Models

- Aka Bayesian networks, probabilistic networks
- Nodes are hypotheses (random vars) and the probabilities corresponds to our belief in the truth of the hypothesis
- Arcs are direct influences between hypotheses
- The structure is represented as a directed acyclic graph (DAG)
- The parameters are the conditional probabilities in the arcs (Pearl, 1988, 2000; Jensen, 1996; Lauritzen, 1996)

# Causes and Bayes' Rule



*Diagnostic inference:  
Knowing that the grass is wet,  
what is the probability that rain is  
the cause?*

$$\begin{aligned} P(R | W) &= \frac{P(W | R)P(R)}{P(W)} \\ &= \frac{P(W | R)P(R)}{P(W | R)P(R) + P(W | \sim R)P(\sim R)} \\ &= \frac{0.9 \times 0.4}{0.9 \times 0.4 + 0.2 \times 0.6} = 0.75 \end{aligned}$$

# Conditional Independence

- $X$  and  $Y$  are independent if

$$P(X, Y) = P(X)P(Y)$$

- $X$  and  $Y$  are conditionally independent given  $Z$  if

$$P(X, Y | Z) = P(X | Z)P(Y | Z)$$

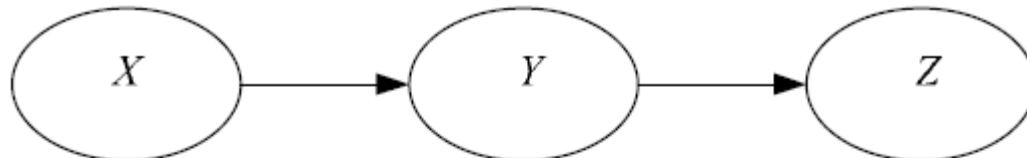
or

$$P(X | Y, Z) = P(X | Z)$$

- Three canonical cases: Head-to-tail, Tail-to-tail, head-to-head

# Case 1: Head-to-Head

- $P(X, Y, Z) = P(X)P(Y|X)P(Z|Y)$



(a) Model

$$P(C) = 0.4$$

$$P(R|C) = 0.8$$

$$P(R|\sim C) = 0.1$$

$$P(W|R) = 0.9$$

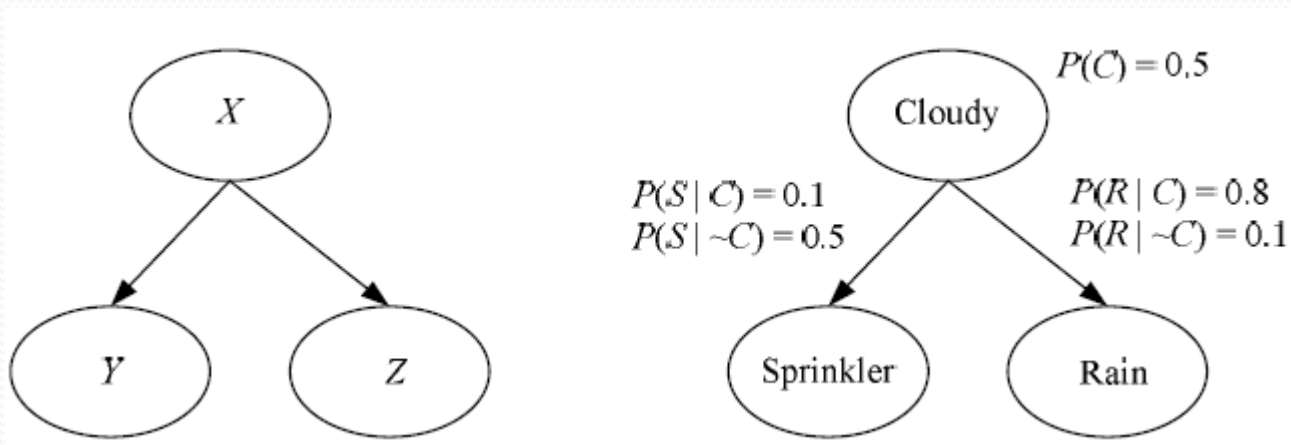
$$P(W|\sim R) = 0.2$$



- $P(W|C) = P(W|R)P(R|C) + P(W|\sim R)P(\sim R|C)$

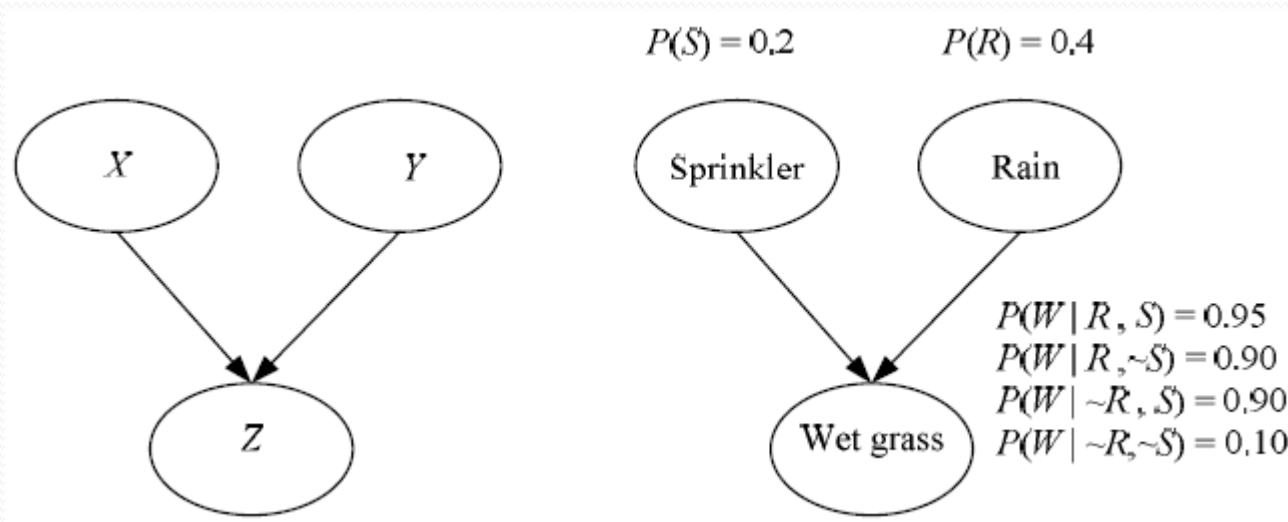
# Case 2: Tail-to-Tail

- $P(X,Y,Z)=P(X)P(Y|X)P(Z|X)$



# Case 3: Head-to-Head

- $P(X,Y,Z)=P(X)P(Y)P(Z|X,Y)$

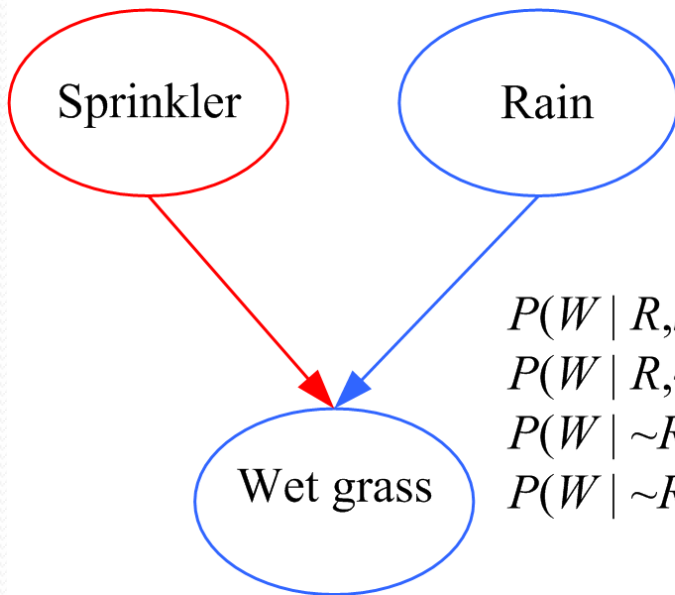




# Causal vs Diagnostic Inference

$$P(S)=0.2$$

$$P(R)=0.4$$



$$P(W | R,S)=0.95$$

$$P(W | R,\sim S)=0.90$$

$$P(W | \sim R,S)=0.90$$

$$P(W | \sim R,\sim S)=0.10$$

*Causal inference: If the sprinkler is on, what is the probability that the grass is wet?*

$$P(W|S) = P(W|R,S) P(R|S) + P(W|\sim R,S) P(\sim R|S)$$

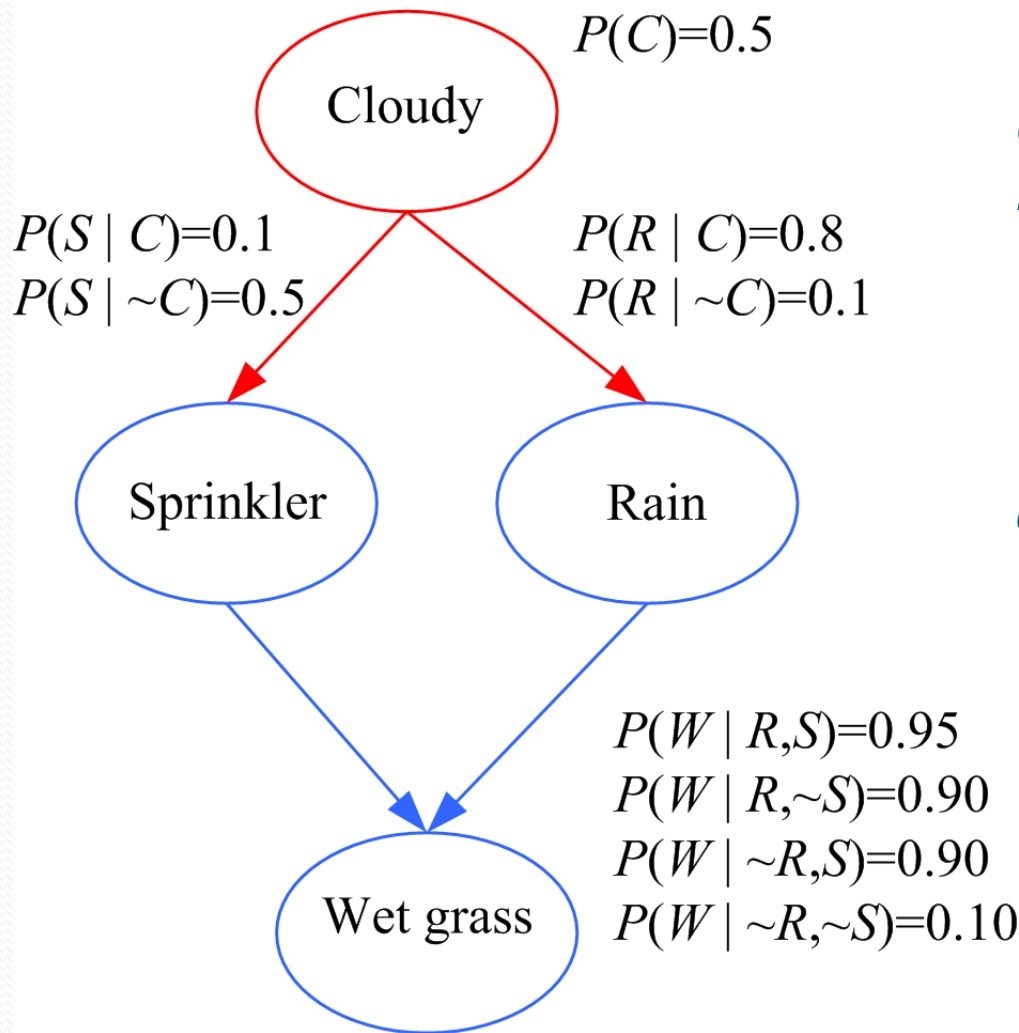
$$= P(W|R,S) P(R) + P(W|\sim R,S) P(\sim R)$$

$$= 0.95 \cdot 0.4 + 0.9 \cdot 0.6 = 0.92$$

*Diagnostic inference: If the grass is wet, what is the probability that the sprinkler is on?  $P(S|W) = 0.35 > 0.2 P(S)$*

*$P(S|R,W) = 0.21$  Explaining away: Knowing that it has rained decreases the probability that the sprinkler is on.*

# Causes



*Causal inference:*

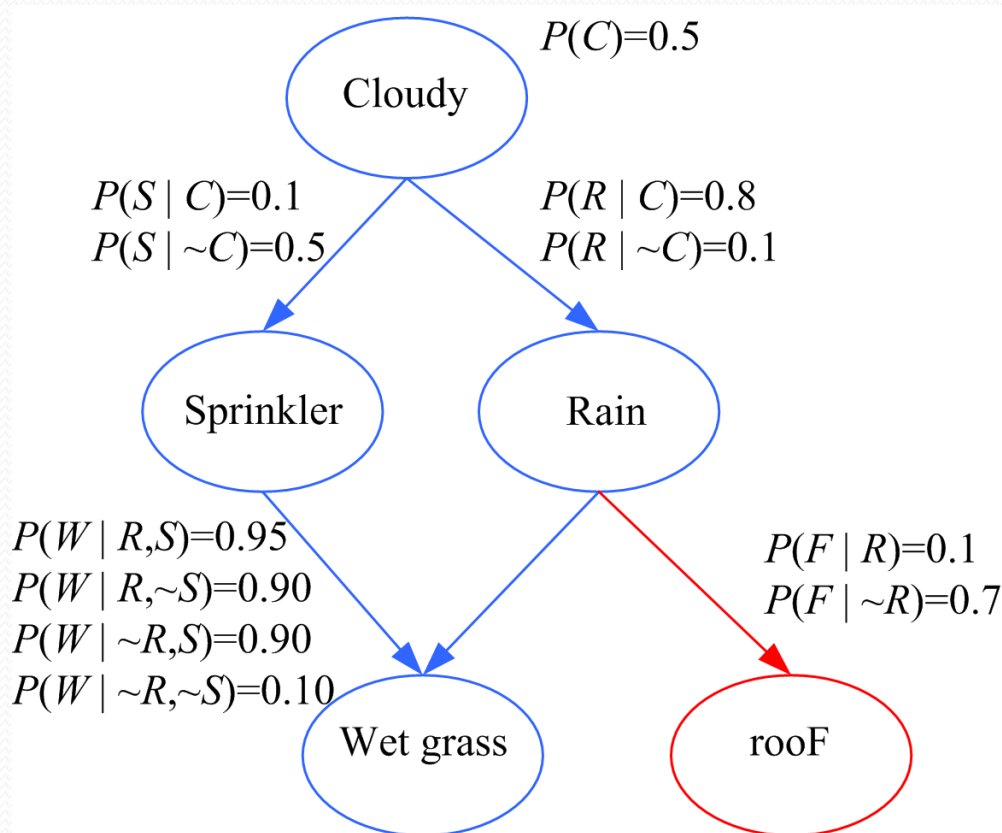
$$P(W|C) = P(W|R,S) P(R,S|C) + P(W|\sim R,S) P(\sim R,S|C) + P(W|R,\sim S) P(R,\sim S|C) + P(W|\sim R,\sim S) P(\sim R,\sim S|C)$$

*and use the fact that*

$$P(R,S|C) = P(R|C) P(S|C)$$

*Diagnostic:  $P(C|W) = ?$*

# Exploiting the Local Structure

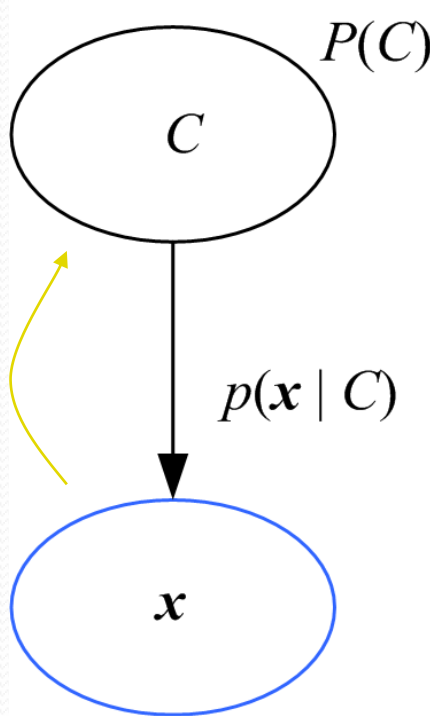


$$P(F | C) = ?$$

$$P(C, S, R, W, F) = P(C)P(S | C)P(R | C)P(W | S, R)P(F | R)$$

$$P(X_1, \dots, X_d) = \prod_{i=1}^d P(X_i | \text{parents}(X_i))$$

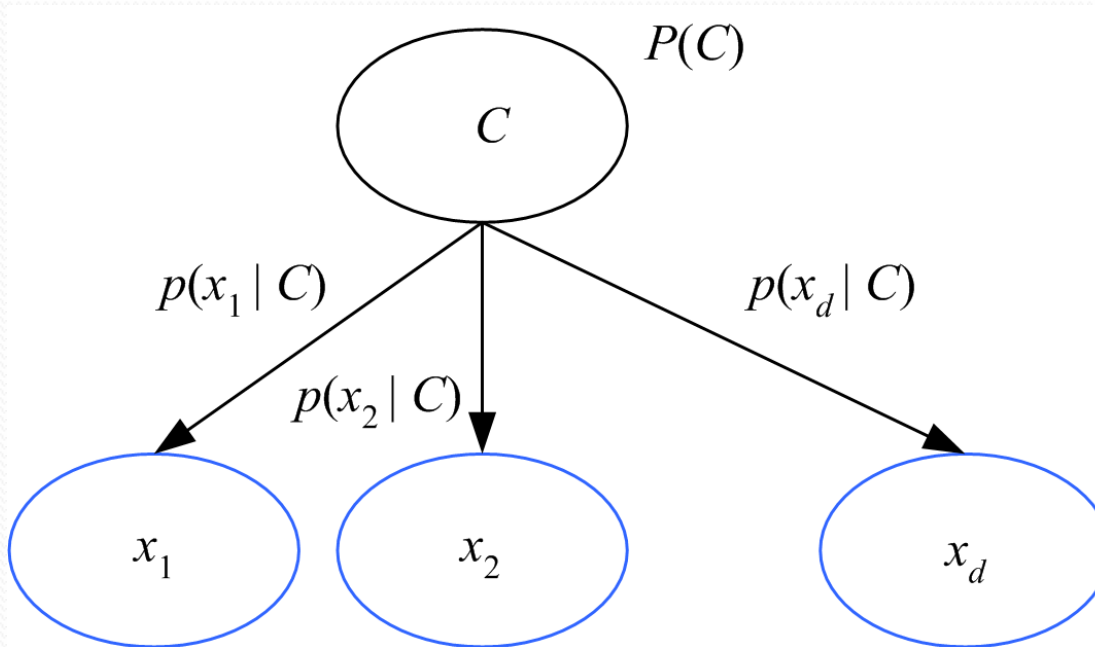
# Classification



Bayes' rule inverts the arc:

$$P(C | \mathbf{x}) = \frac{p(\mathbf{x} | C)P(C)}{p(\mathbf{x})}$$

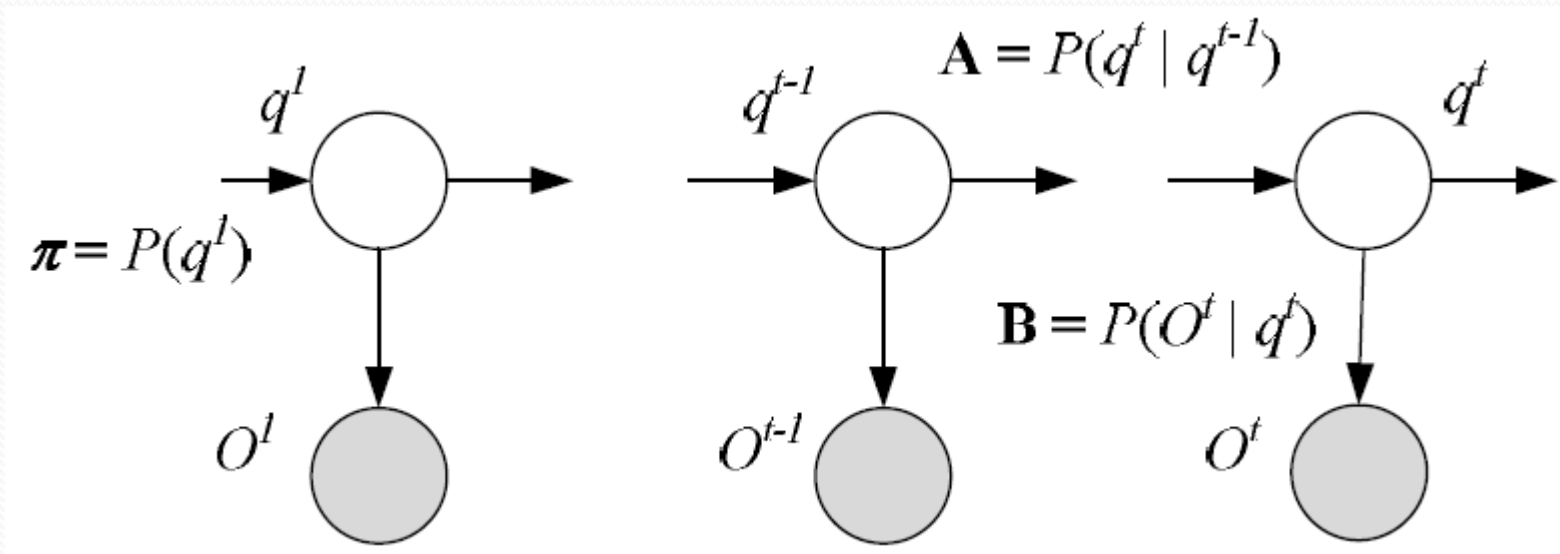
# Naive Bayes' Classifier

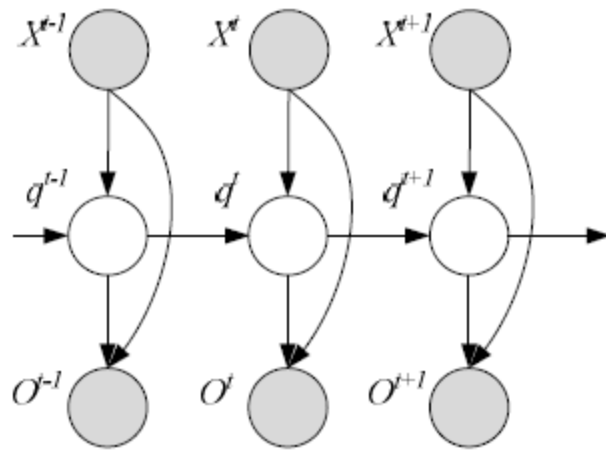


Given  $C$ ,  $x_j$  are independent:

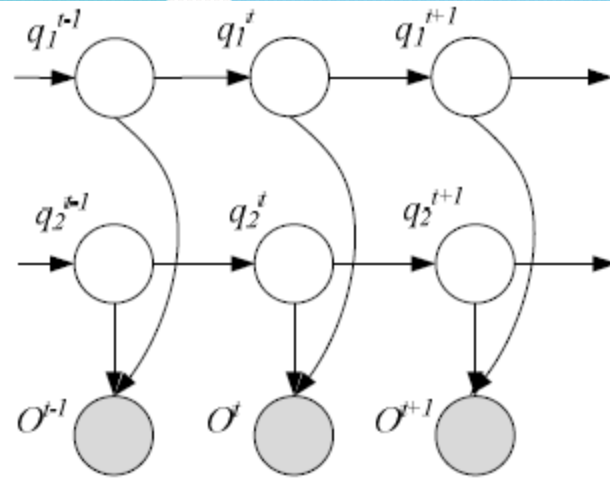
$$p(\mathbf{x}|C) = p(x_1|C) p(x_2|C) \dots p(x_d|C)$$

# Hidden Markov Model as a Graphical Model

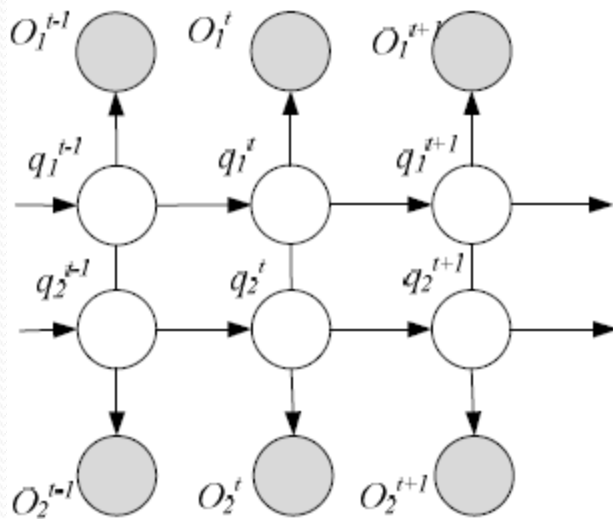




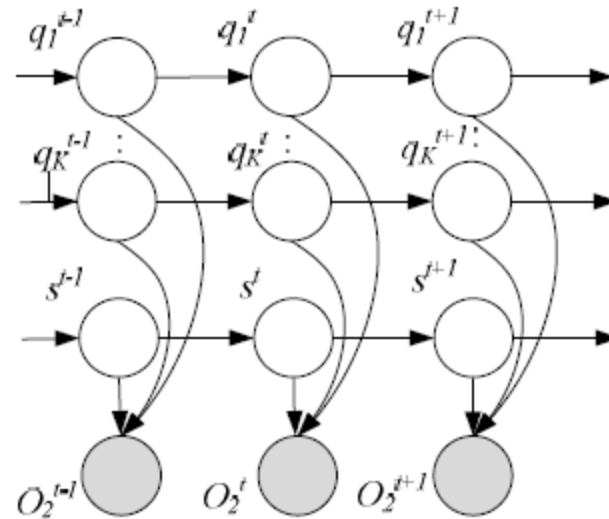
(a) Input-output HMM



(b) Factorial HMM

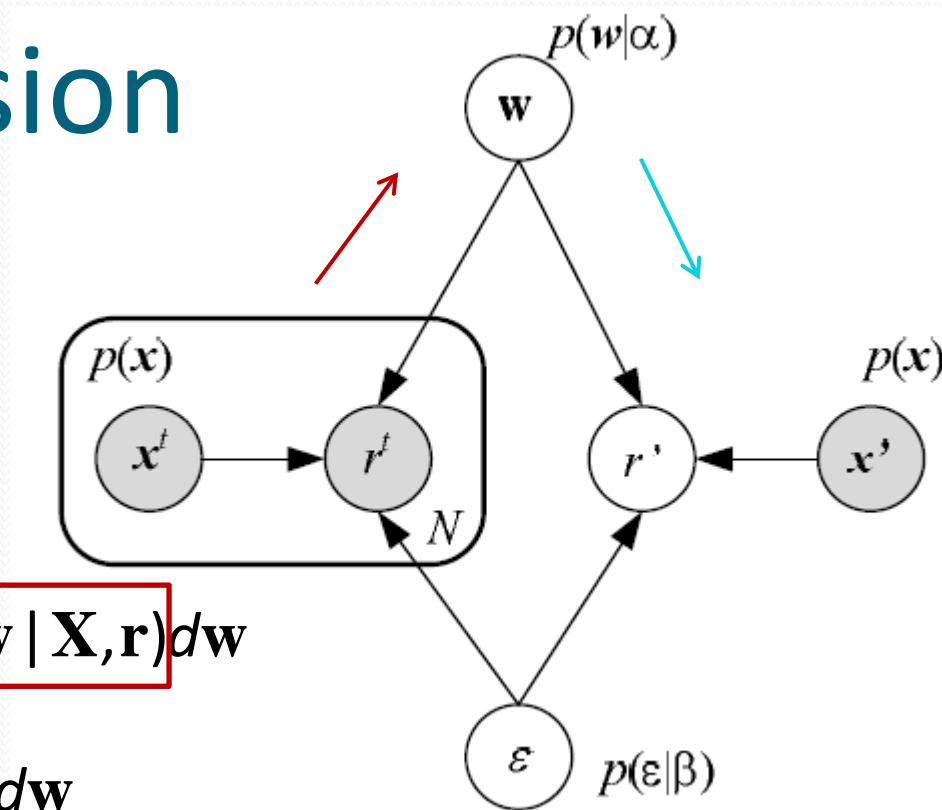


(c) Coupled HMM



(d) Switching HMM

# Linear Regression



$$p(r' | \mathbf{x}', \mathbf{r}, \mathbf{X}) = \int p(r' | \mathbf{x}', \mathbf{w}) p(\mathbf{w} | \mathbf{X}, \mathbf{r}) d\mathbf{w}$$

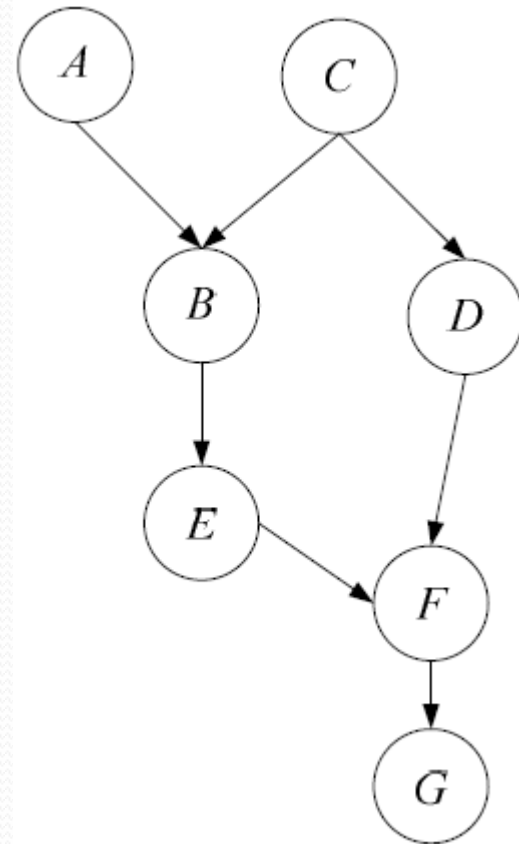
$$= \int p(r' | \mathbf{x}', \mathbf{w}) \frac{p(\mathbf{r} | \mathbf{X}, \mathbf{w}) p(\mathbf{w})}{p(\mathbf{r})} d\mathbf{w}$$

$$\propto \int p(r' | \mathbf{x}', \mathbf{w}) \prod_t p(r^t | \mathbf{x}^t, \mathbf{w}) p(\mathbf{w}) d\mathbf{w}$$



# d-Separation

- A path from node  $A$  to node  $B$  is blocked if
  - a) The directions of edges on the path meet head-to-tail (case 1) or tail-to-tail (case 2) and the node is in  $C$ , or
  - b) The directions of edges meet head-to-head (case 3) and neither that node nor any of its descendants is in  $C$ .
- If all paths are blocked,  $A$  and  $B$  are d-separated (conditionally independent) given  $C$ .



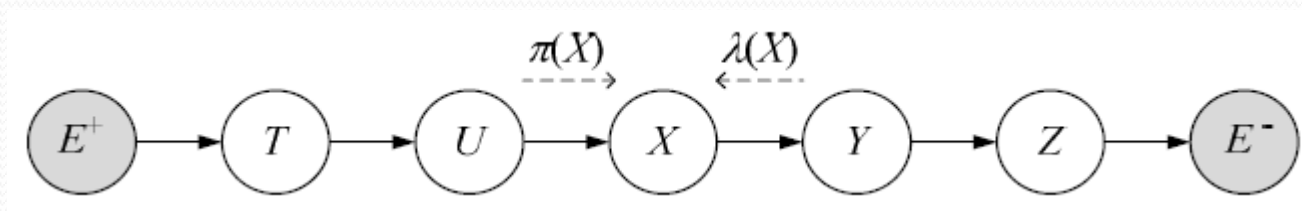
$BCDF$  is blocked given  $C$ .

$BEFG$  is blocked by  $F$ .

$BEFD$  is blocked unless  $F$  (or  $G$ ) is given.

# Belief Propagation (Pearl, 1988)

- Chain:



$$\begin{aligned} P(X|E) &= \frac{P(E|X)P(X)}{P(E)} = \frac{P(E^+, E^- | X)P(X)}{P(E)} \\ &= \frac{P(E^+ | X)P(E^- | X)P(X)}{P(E)} = \alpha \pi(X) \lambda(X) \end{aligned}$$

$$\pi(X) = \sum_U P(X|U)\pi(U)$$

$$\lambda(X) = \sum_Y P(Y|X)\lambda(Y)$$

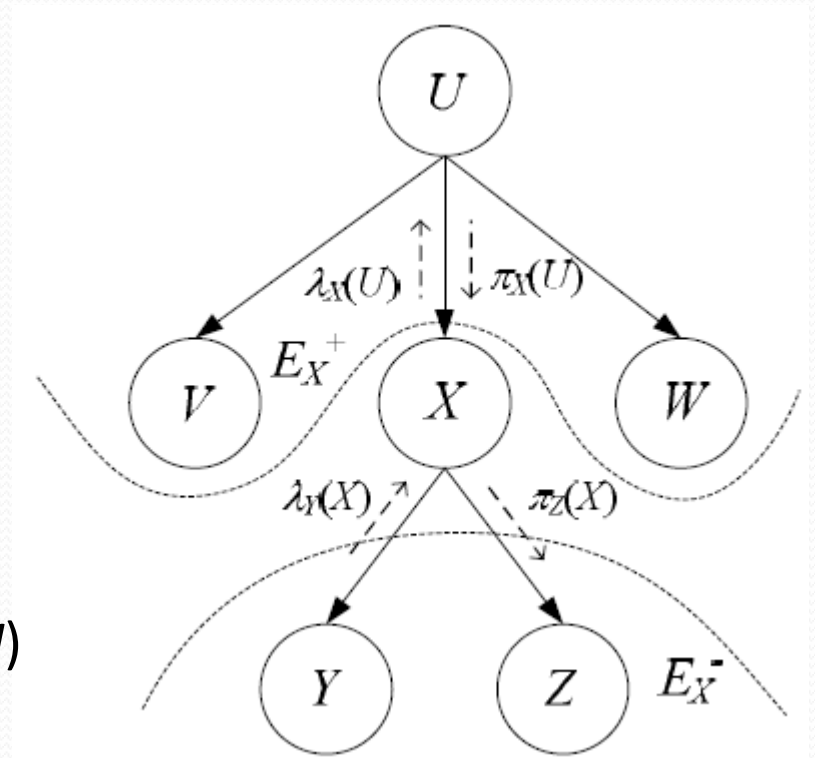
# Trees

$$\lambda(X) = P(E_x^- | X) = \lambda_y(X)\lambda_z(X)$$

$$\lambda_x(U) = \sum_x \lambda(X)P(X | U)$$

$$\pi(X) = P(X | E_x^+) = \sum_U P(X | U)\pi_x(U)$$

$$\pi_y(X) = \alpha\lambda_z(X)\pi(X)$$



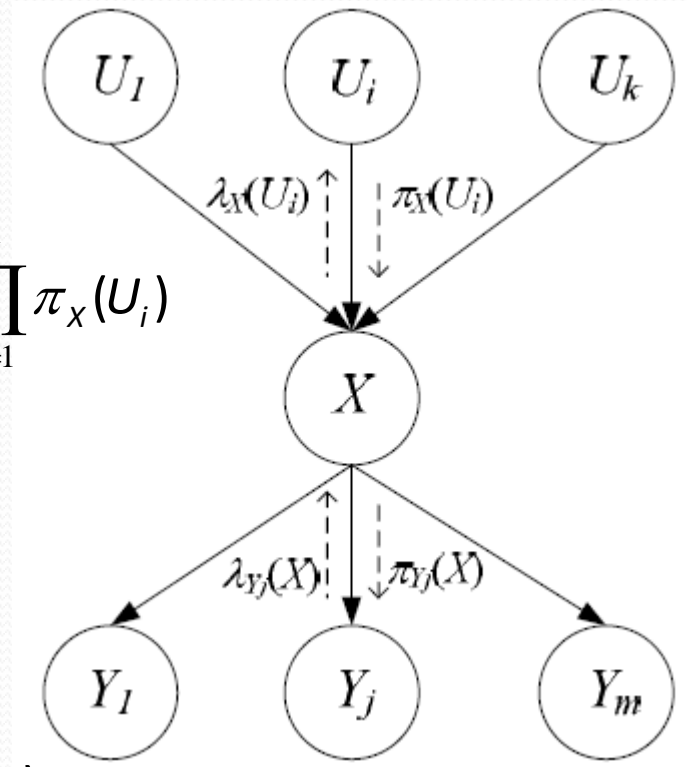
# Polytrees

$$\pi(X) = P(X | E_X^+) = \sum_{U_1} \sum_{U_2} \cdots \sum_{U_k} P(X | U_1, U_2, \dots, U_k) \prod_{i=1}^k \pi_X(U_i)$$

$$\pi_{Y_j}(X) = \alpha \prod_{s \neq j} \lambda_{Y_s}(X) \pi(X)$$

$$\lambda_X(U_i) = \beta \sum_X \lambda(X) \sum_{U_{r \neq i}} P(X | U_1, U_2, \dots, U_k) \prod_{r \neq i} \pi_X(U_r)$$

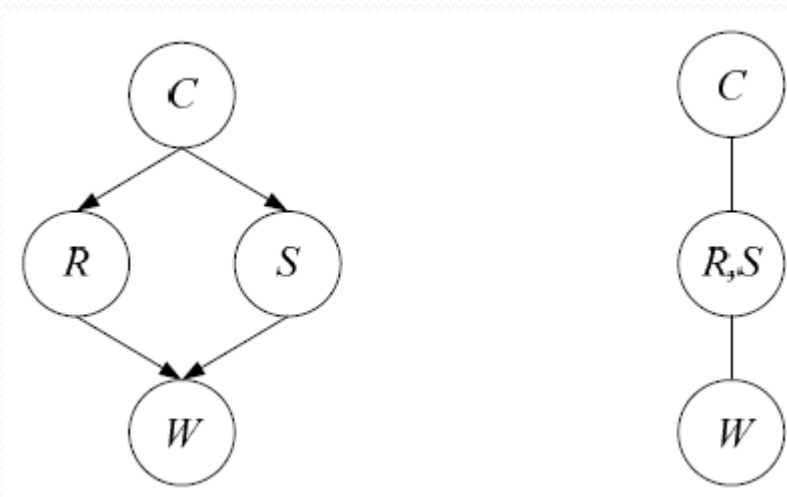
$$\lambda(X) = \prod_{j=1}^m \lambda_{Y_j}(X)$$



How can we model  $P(X | U_1, U_2, \dots, U_k)$  cheaply?

# Junction Trees

- If  $X$  does not separate  $E^+$  and  $E^-$ , we convert it into a junction tree and then apply the polytree algorithm



Tree of moralized,  
clique nodes

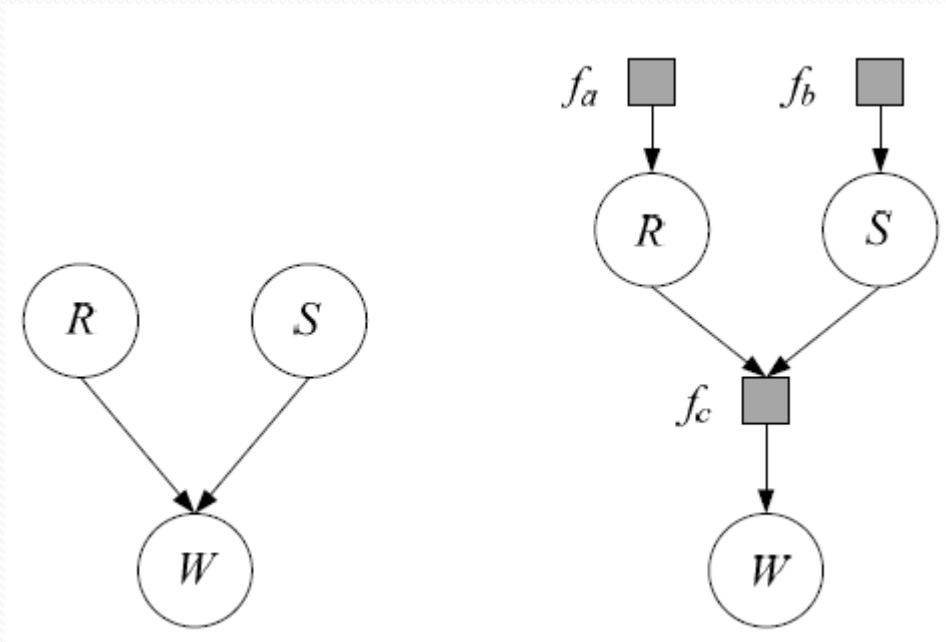
# Undirected Graphs: Markov Random Fields

- In a Markov random field, dependencies are symmetric, for example, pixels in an image
- In an undirected graph,  $A$  and  $B$  are independent if removing  $C$  makes them unconnected.
- Potential function  $\psi_c(X_c)$  shows how favorable is the particular configuration  $X$  over the clique  $C$
- The joint is defined in terms of the clique potentials

$$p(X) = \frac{1}{Z} \prod_c \psi_c(X_c) \text{ where normalizer } Z = \sum_x \prod_c \psi_c(X_c)$$

# Factor Graphs

- Define new factor nodes and write the joint in terms of them



$$p(X) = \frac{1}{Z} \prod_s f_s(X_s)$$

# Learning a Graphical Model

- Learning the conditional probabilities, either as tables (for discrete case with small number of parents), or as parametric functions
- Learning the structure of the graph: Doing a state-space search over a score function that uses both goodness of fit to data and some measure of complexity



# Influence Diagrams

