



# Machine Learning 10-601

Tom M. Mitchell  
Machine Learning Department  
Carnegie Mellon University

February 18, 2015

## Today:

- Graphical models
- Bayes Nets:
  - Representing distributions
  - Conditional independencies
  - Simple inference
  - Simple learning

## Readings:

- Bishop chapter 8, through 8.2

# Graphical Models

- Key Idea:
  - Conditional independence assumptions useful
  - but Naïve Bayes is extreme!
  - Graphical models express sets of conditional independence assumptions via graph structure
  - Graph structure plus associated parameters define joint probability distribution over set of variables
- Two types of graphical models:
  - Directed graphs (aka Bayesian Networks)
  - Undirected graphs (aka Markov Random Fields)

10-601



# Graphical Models – Why Care?

- Among most important ML developments of the decade
- Graphical models allow combining:
  - Prior knowledge in form of dependencies/independencies
  - Prior knowledge in form of priors over parameters
  - Observed training data
- Principled and ~general methods for
  - Probabilistic inference
  - Learning
- Useful in practice
  - Diagnosis, help systems, text analysis, time series models, ...

# Conditional Independence

*Definition:*  $X$  is conditionally independent of  $Y$  given  $Z$ , if the probability distribution governing  $X$  is independent of the value of  $Y$ , given the value of  $Z$

$$(\forall i, j, k) P(X = x_i | Y = y_j, Z = z_k) = P(X = x_i | Z = z_k)$$

Which we often write  $P(X|Y, Z) = P(X|Z)$

E.g.,  $P(\text{Thunder} | \text{Rain}, \text{Lightning}) = P(\text{Thunder} | \text{Lightning})$

# Marginal Independence

*Definition:*  $X$  is marginally independent of  $Y$  if

$$(\forall i, j) P(X = x_i, Y = y_j) = P(X = x_i)P(Y = y_j)$$

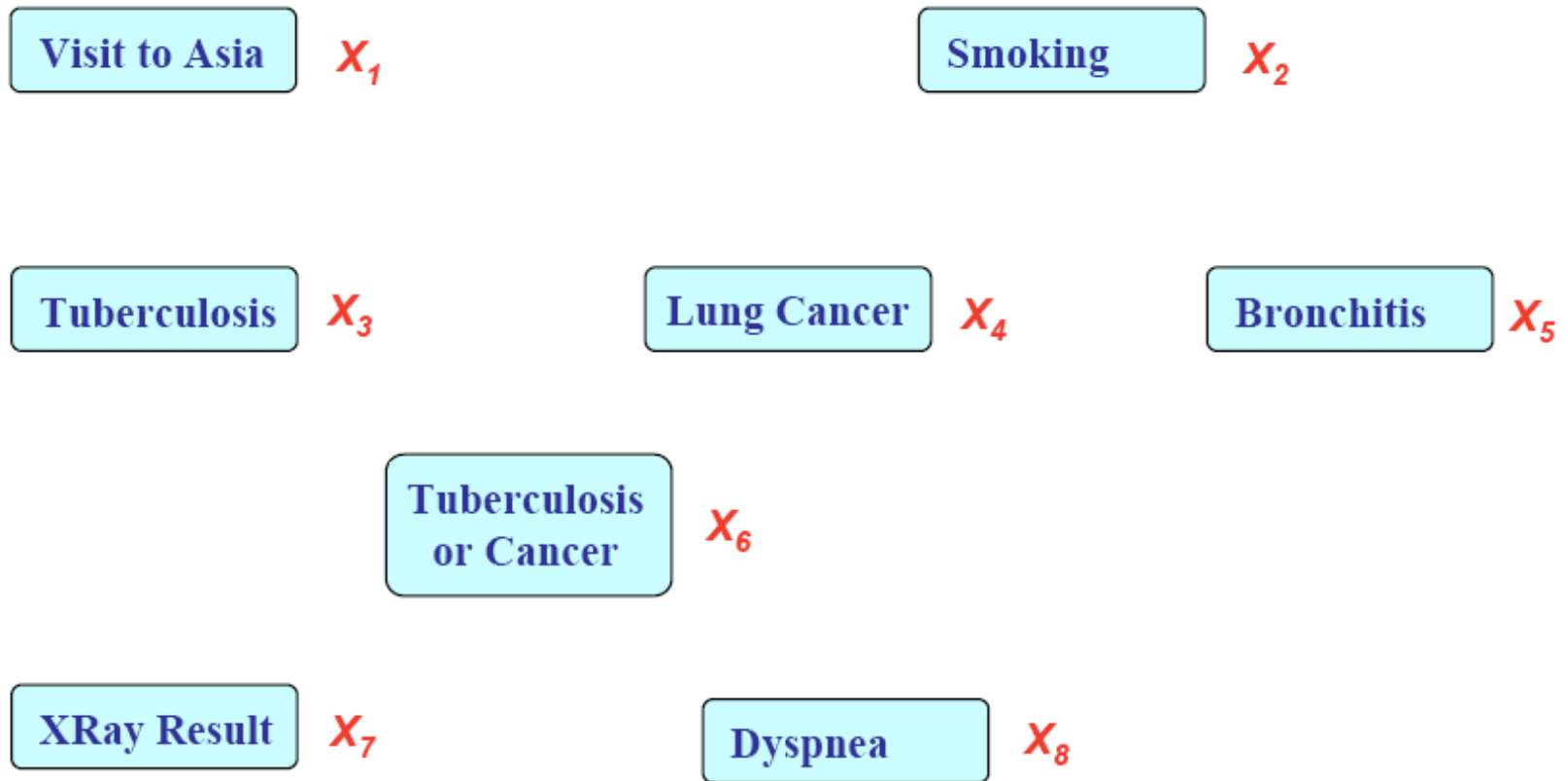
Equivalently, if

$$(\forall i, j) P(X = x_i | Y = y_j) = P(X = x_i)$$

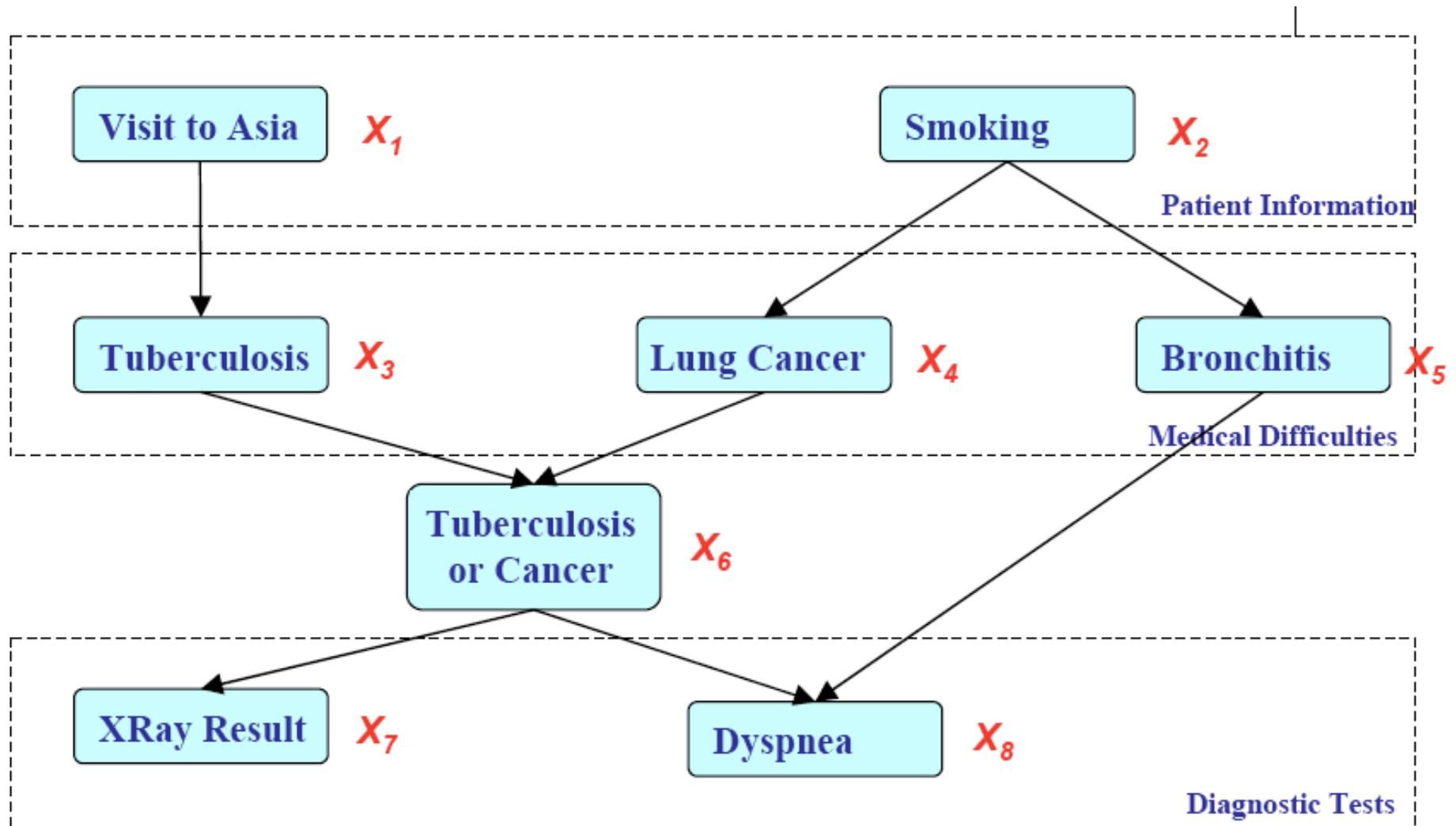
Equivalently, if

$$(\forall i, j) P(Y = y_i | X = x_j) = P(Y = y_i)$$

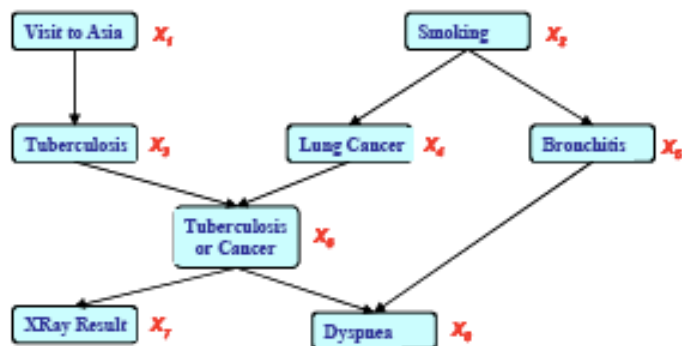
## Represent Joint Probability Distribution over Variables



# Describe network of dependencies



Bayes Nets define Joint Probability Distribution in terms of this graph, plus parameters



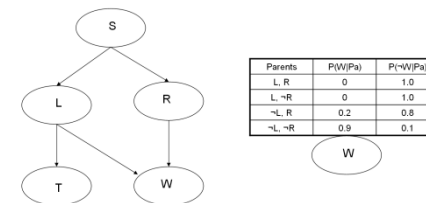
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\ = P(X_1) P(X_2) P(X_3 | X_1) P(X_4 | X_2) P(X_5 | X_2) \\ P(X_6 | X_3, X_4) P(X_7 | X_6) P(X_8 | X_5, X_6)$$

Benefits of Bayes Nets:

- Represent the full joint distribution in fewer parameters, using prior knowledge about dependencies
- Algorithms for inference and learning



# Bayesian Networks Definition



A Bayes network represents the joint probability distribution over a collection of random variables

A Bayes network is a directed acyclic graph and a set of conditional probability distributions (CPD's)

- Each node denotes a random variable
- Edges denote dependencies
- For each node  $X_i$  its CPD defines  $P(X_i | Pa(X_i))$
- The joint distribution over all variables is defined to be

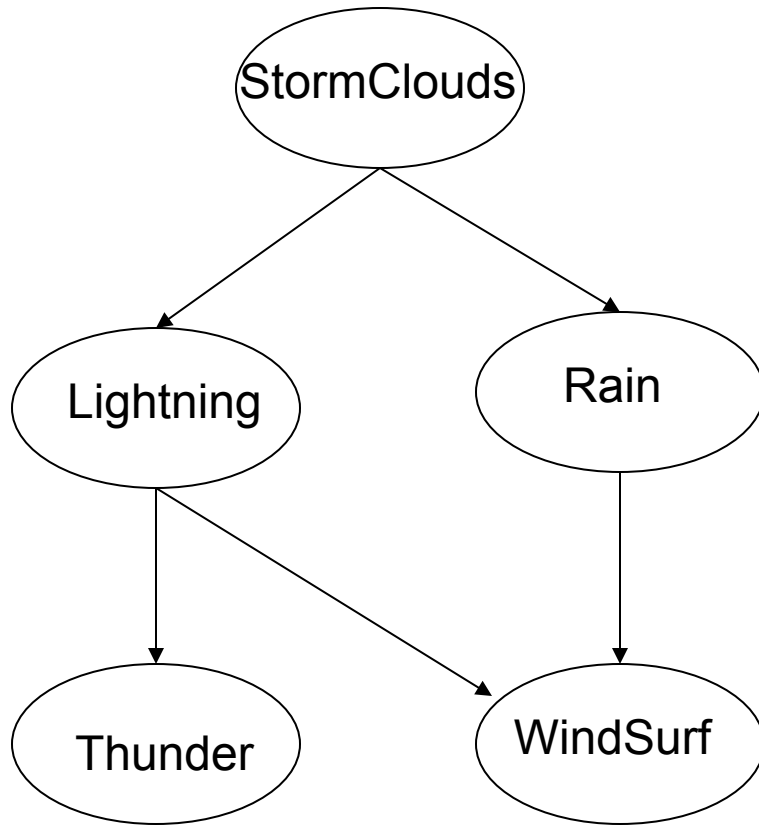
$$P(X_1 \dots X_n) = \prod_i P(X_i | Pa(X_i))$$

$Pa(X)$  = immediate parents of  $X$  in the graph

# Bayesian Network

Nodes = random variables

A conditional probability distribution (CPD) is associated with each node  $N$ , defining  $P(N \mid \text{Parents}(N))$



Parents	$P(W Pa)$	$P(\neg W Pa)$
L, R	0	1.0
L, $\neg R$	0	1.0
$\neg L$ , R	0.2	0.8
$\neg L$ , $\neg R$	0.9	0.1



The joint distribution over all variables:

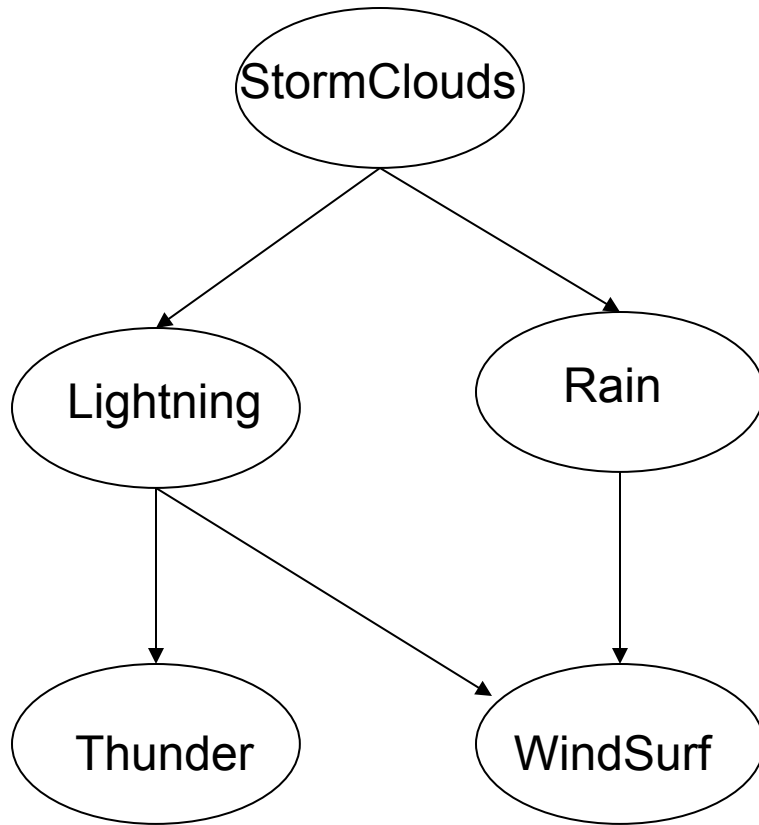
$$P(X_1 \dots X_n) = \prod_i P(X_i | Pa(X_i))$$

# Bayesian Network

What can we say about conditional independencies in a Bayes Net?

One thing is this:

Each node is conditionally independent of its non-descendents, given only its immediate parents.



Parents	$P(W Pa)$	$P(\neg W Pa)$
L, R	0	1.0
L, $\neg R$	0	1.0
$\neg L$ , R	0.2	0.8
$\neg L$ , $\neg R$	0.9	0.1



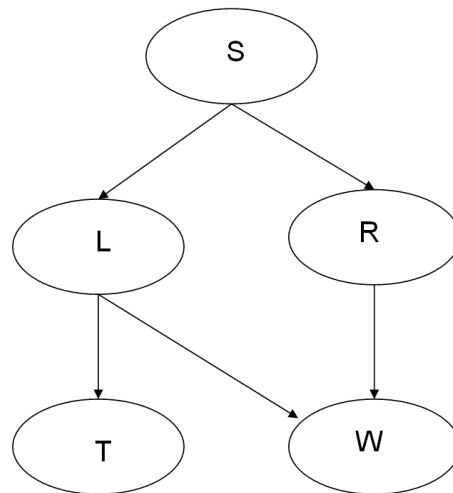
# Some helpful terminology

Parents =  $\text{Pa}(X)$  = immediate parents

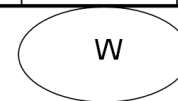
Antecedents = parents, parents of parents, ...

Children = immediate children

Descendants = children, children of children, ...

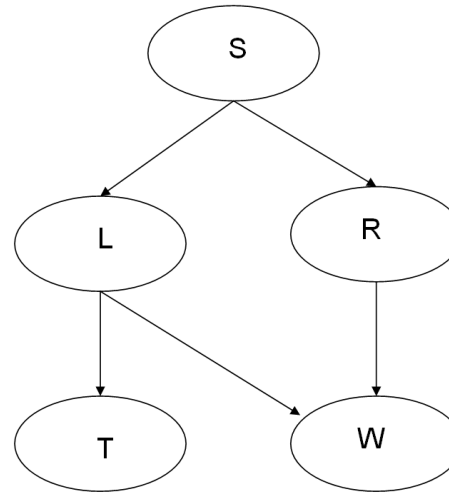


Parents	$P(W \text{Pa})$	$P(\neg W \text{Pa})$
L, R	0	1.0
L, $\neg R$	0	1.0
$\neg L$ , R	0.2	0.8
$\neg L$ , $\neg R$	0.9	0.1



# Bayesian Networks

- CPD for each node  $X_i$  describes  $P(X_i \mid Pa(X_i))$



Parents	$P(W Pa)$	$P(\neg W Pa)$
L, R	0	1.0
L, $\neg R$	0	1.0
$\neg L$ , R	0.2	0.8
$\neg L$ , $\neg R$	0.9	0.1

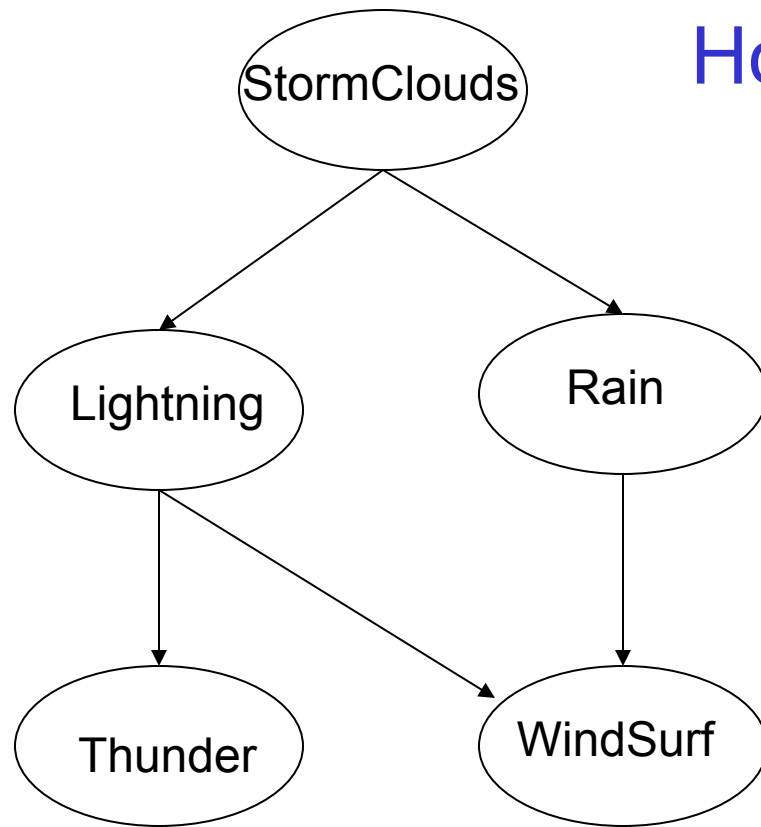


Chain rule of probability says that in general:

$$P(S, L, R, T, W) = P(S)P(L|S)P(R|S, L)P(T|S, L, R)P(W|S, L, R, T)$$

But in a Bayes net:  $P(X_1 \dots X_n) = \prod_i P(X_i | Pa(X_i))$

## How Many Parameters?



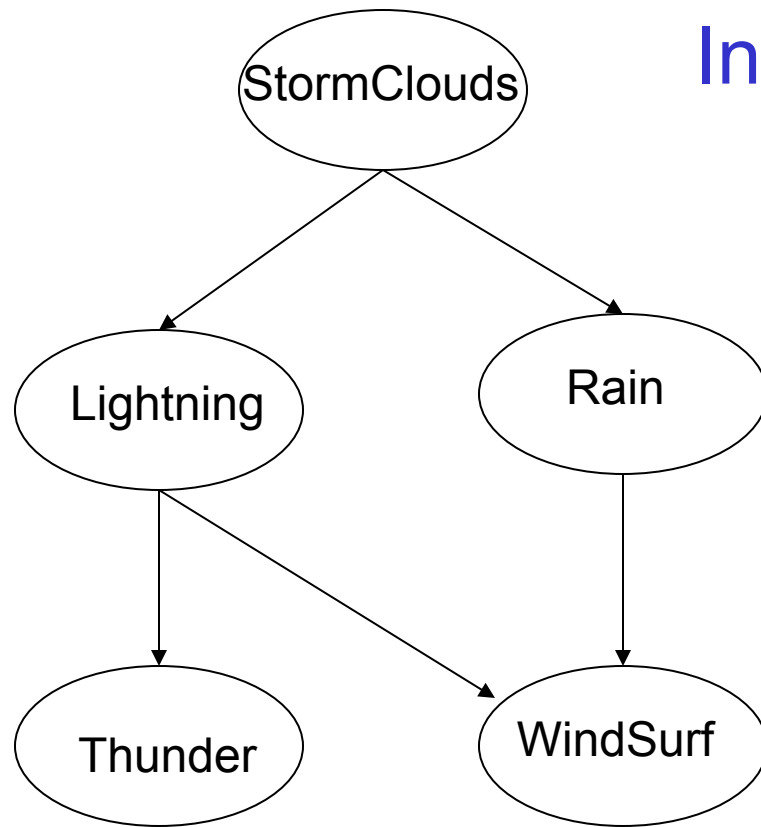
Parents	$P(W Pa)$	$P(\neg W Pa)$
L, R	0	1.0
L, $\neg R$	0	1.0
$\neg L$ , R	0.2	0.8
$\neg L$ , $\neg R$	0.9	0.1



To define joint distribution in general?

To define joint distribution for this Bayes Net?

## Inference in Bayes Nets

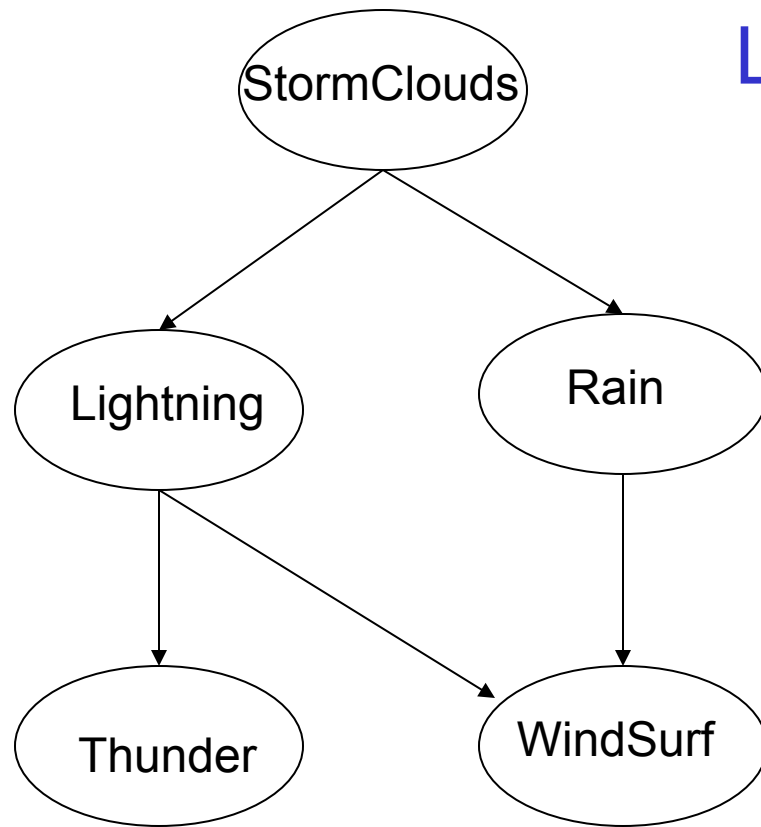


Parents	$P(W Pa)$	$P(\neg W Pa)$
L, R	0	1.0
L, $\neg R$	0	1.0
$\neg L$ , R	0.2	0.8
$\neg L$ , $\neg R$	0.9	0.1



$$P(S=1, L=0, R=1, T=0, W=1) =$$

## Learning a Bayes Net



Parents	$P(W Pa)$	$P(\neg W Pa)$
L, R	0	1.0
L, $\neg R$	0	1.0
$\neg L$ , R	0.2	0.8
$\neg L$ , $\neg R$	0.9	0.1



Consider learning when graph structure is given, and data = { <s,l,r,t,w> }

What is the MLE solution? MAP? 



# Algorithm for Constructing Bayes Network

- Choose an ordering over variables, e.g.,  $X_1, X_2, \dots, X_n$
- For  $i=1$  to  $n$ 
  - Add  $X_i$  to the network
  - Select parents  $Pa(X_i)$  as minimal subset of  $X_1 \dots X_{i-1}$  such that

$$P(X_i | Pa(X_i)) = P(X_i | X_1, \dots, X_{i-1})$$

Notice this choice of parents assures

$$\begin{aligned} P(X_1 \dots X_n) &= \prod_i P(X_i | X_1 \dots X_{i-1}) && \text{(by chain rule)} \\ &= \prod_i P(X_i | Pa(X_i)) && \text{(by construction)} \end{aligned}$$

# Example

- Bird flu and Allergies both cause Nasal problems
- Nasal problems cause Sneezes and Headaches

What is the Bayes Network for  $X_1, \dots, X_4$  with NO assumed conditional independencies?

What is the Bayes Network for Naïve Bayes?

What do we do if variables are mix of discrete and real valued?

