Bayes Net

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 plis associated parameters define *joint probability distribution over set of variables/nodes*

- Two types of graphical models :
  - Directed graphs (aka Bayesian Networks)
  - Undirected graphs (aka Markov Random Fields)

today

Graphical Models – Why Care ?

- **Among most important ML developments of the decade**

- **Graphical models allow combining**
  - Prior knowledge in form of dependencies/independencies
  - Observed data to estimate parameters

- **Principles and ~general methods for**
  - Probabilistic inference
  - Learning

- **Useful in practice**
  - Diagnosis, help system, text analysis, time series models, ...

Conditional Independence

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

$$(\forall_i, j, k) P(X = x_i \mid Y = y_j, Z = z_k) = P(X = x_i \mid Z = z_k)$$

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

E.g., $P(Thunder \mid Rain, Lightning) = P(Thunder \mid Lightning)$

Marginal Independence

Definition : X is marginally independent of Y if

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

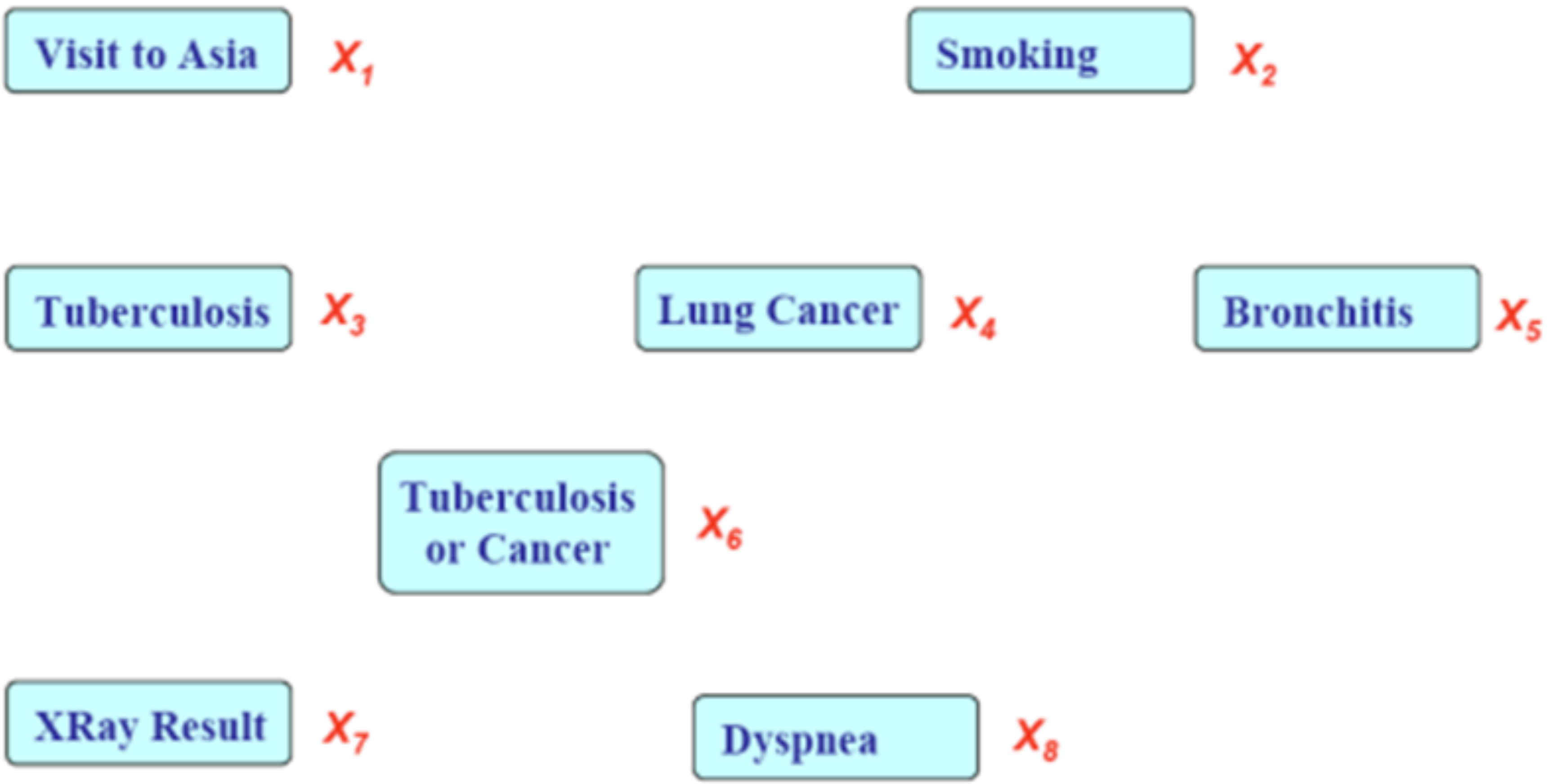
Equivalently, if

$$(\forall_i, j) P(X = x_i \mid Y = y_j) = P(X = x_i)$$

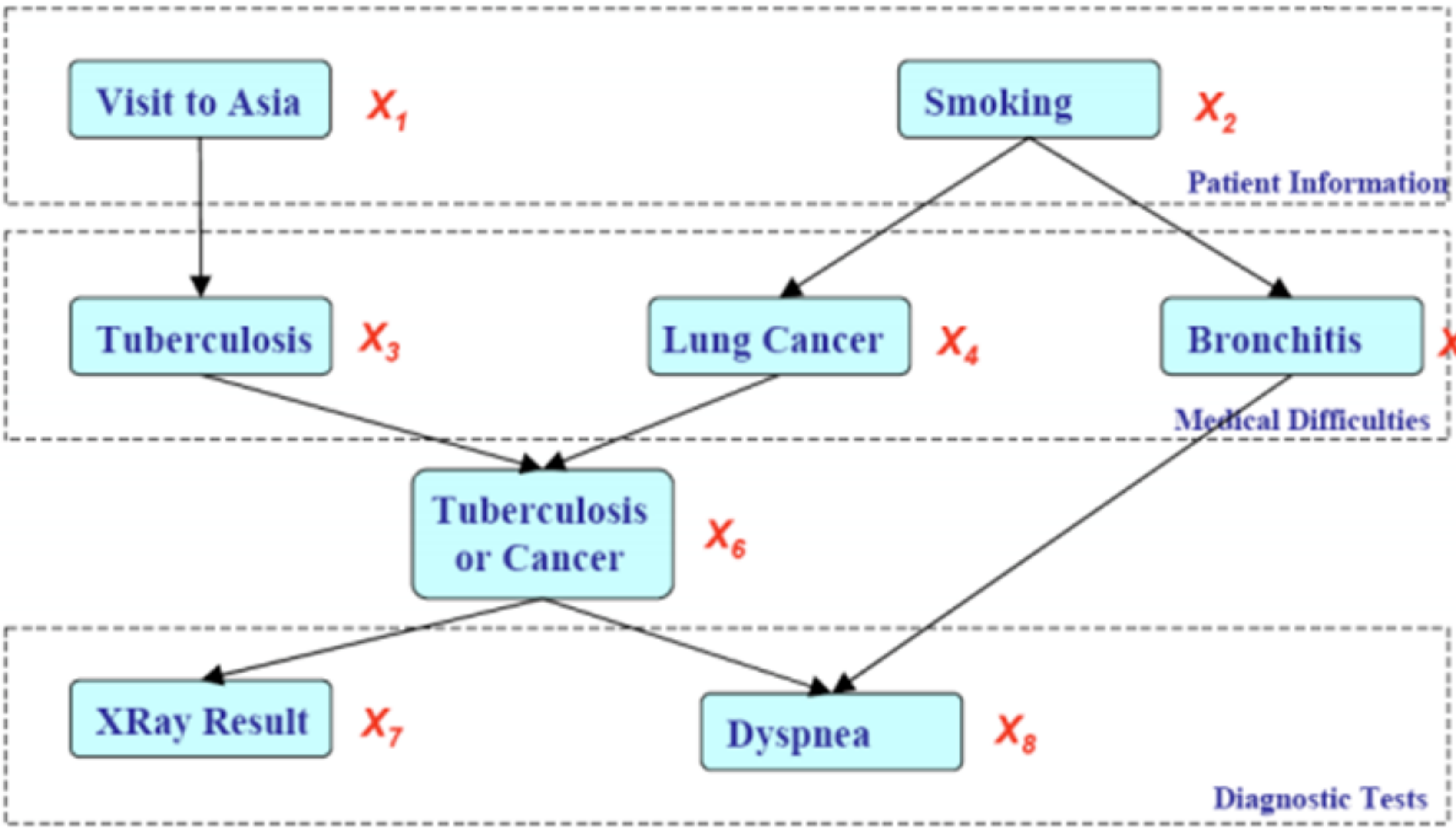
Equivalently, if

$$(\forall_i, j) P(Y = y_j \mid X = x_i) = P(Y = y_j)$$

Represent Joint Probability Distribution over Variables

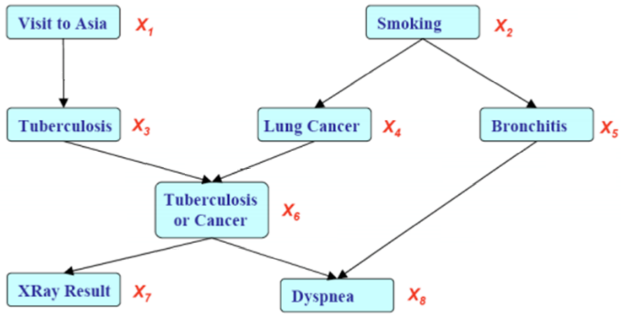


Discribe network of dependencies



Bayesian Networks define Joint Distribution in term of this graph, plus parameters

- If X_i's are conditionally independent (as described by a PGM) , the joint can be factored to a product of simples terms, e.g.,

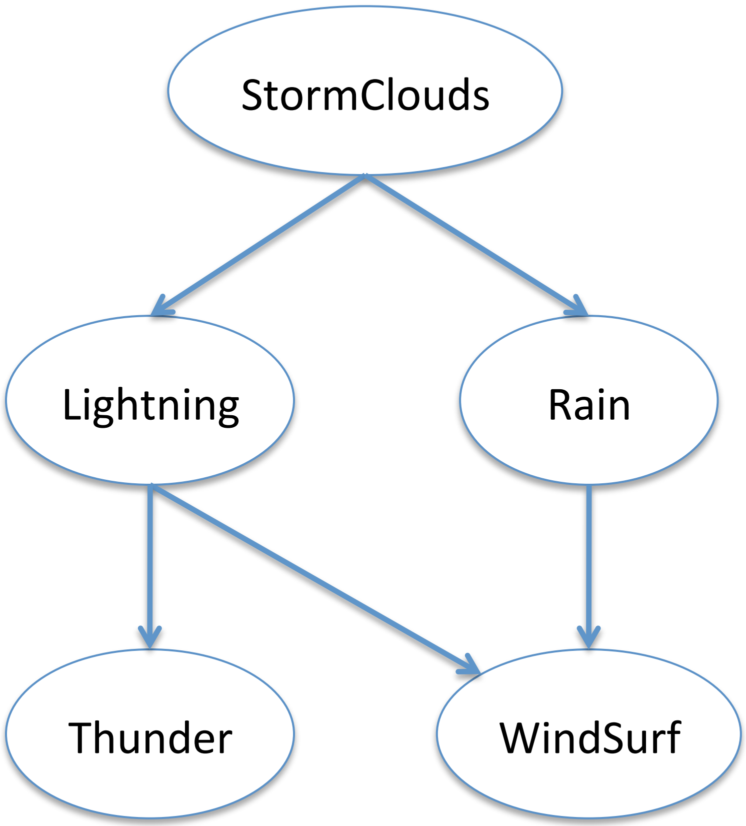


$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

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

- Why we may favor a PGM ?
  - Representation cost : how many probability statements are need ? 2+2+4+4+4+8+4+8=36, an 8-fold reduction from 2^8 !
  - Algoritthms for systemic and efficient inference/learning computation
    Exploring the graph structure and probabilistic (e.g., Bayesian, Markovian) semantics
  - Incorporation of domain knowledge and causal (logical) structures

Bayesian Network

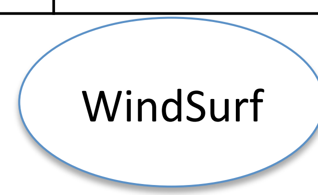


<u>Bayes network</u> : a directed acyclic graph defining a joint probability distribution over a set of variables

Each node denotes a random variable

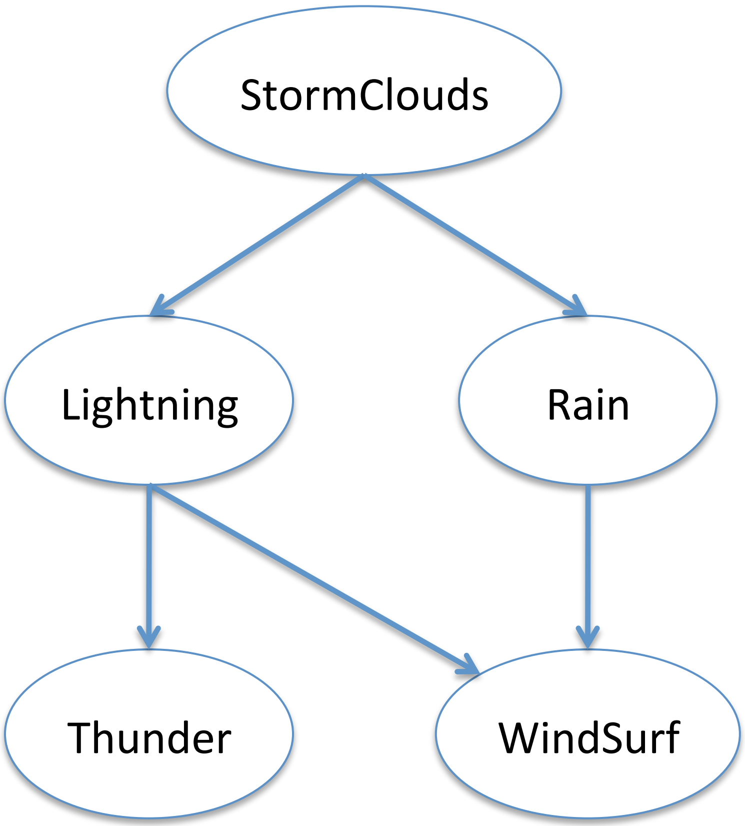
A conditional probability distribution (CPD) is associated with each node N, defining P(N | Parent(N))

Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



The joint distribution over all variables in the network is defined in terms of these CPD’s, plus the graph

Bayesian Network



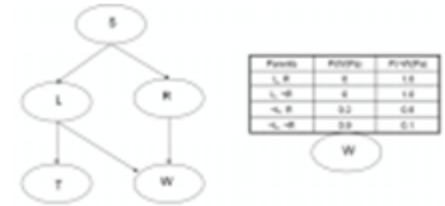
What can we say about conditional independencies in a Bayes Net ?

One thing is this :

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

Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1

Bayesian Network Definition



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

A Bayes network is a derected acyclic graph and a set of CPD's

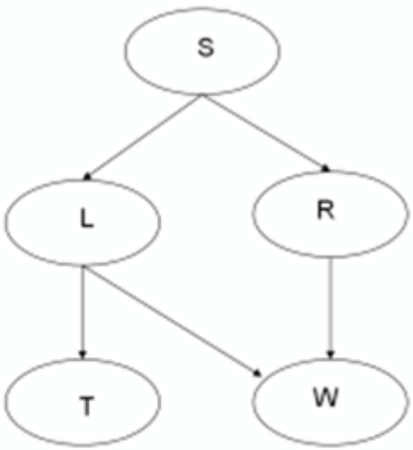
- Each node denotes a random variable
- Edges denote dependencies
- CPD for each node X_i define $P(X_i \mid Pa(X_i))$
- The joint distribution over all variables id defined as

$$P(X_1 \ldots X_n) = \prod_i P(X_i \mid Pa(X_i))$$

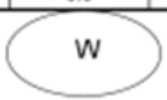
***Pa*(*X*)=immediate parent of X in the graph**

Some helpful terminology

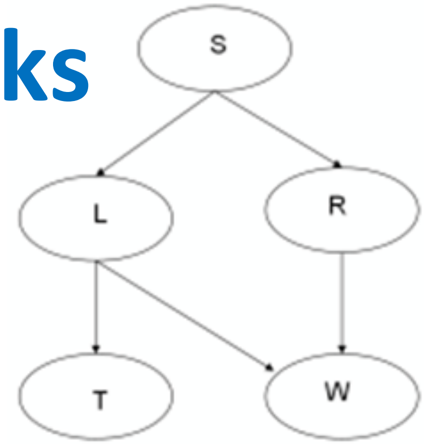
***Parents = Pa*(*X*) = immediate parent**
Antecedents = parents, parents of parents, …
Children = immediate children
Descendents = children, children of children, …



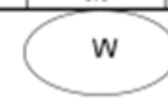
Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



Bayesian Networks



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



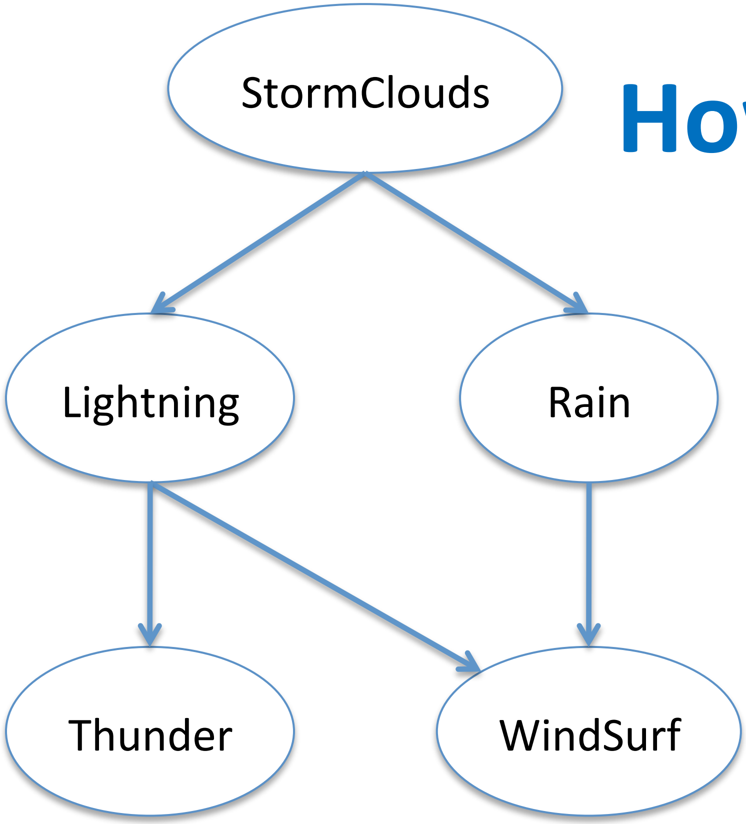
- CPD for each node X_i describes $P(X_i \mid Pa(X_i))$

Chain rule of probability :

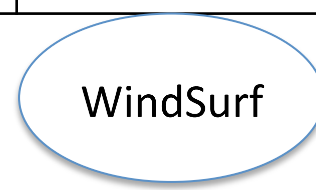
$$P(S,L,R,T,W) = P(S)P(L|S)P(R|S,L)P(T|S,L,R)P(W|S,L,R,T)$$

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

How Many Parameters ?



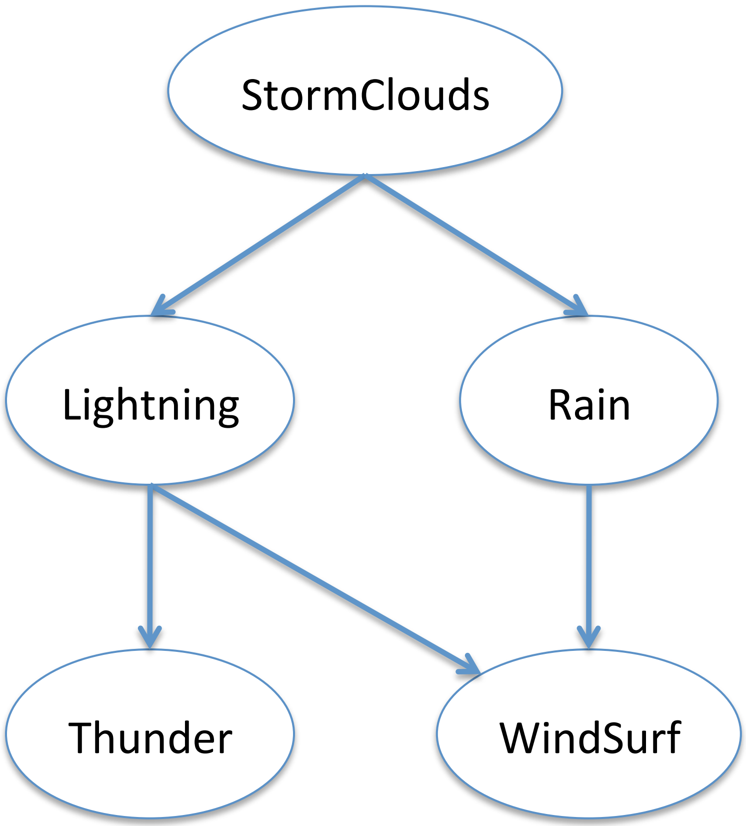
Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



In full joint distribution ?

Given this Bayes Net ?

Bayesian Network

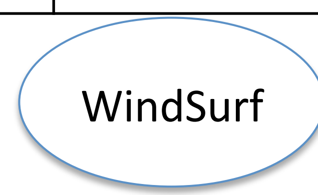


What can we say about conditional independencies in a Bayes Net ?

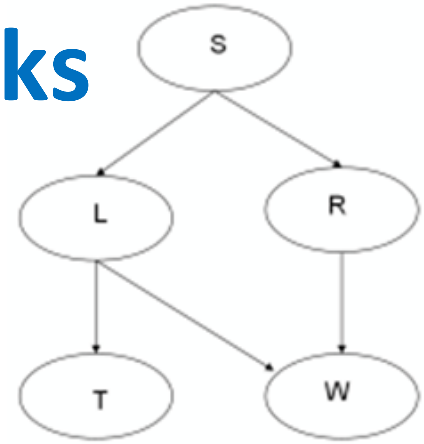
One thing is this :

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

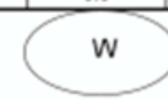
Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



Bayesian Networks



Parents	P(W|Pa)	P(¬W|Pa)
L, R	0	1.0
L, ¬R	0	1.0
¬L, R	0.2	0.8
¬L, ¬R	0.9	0.1



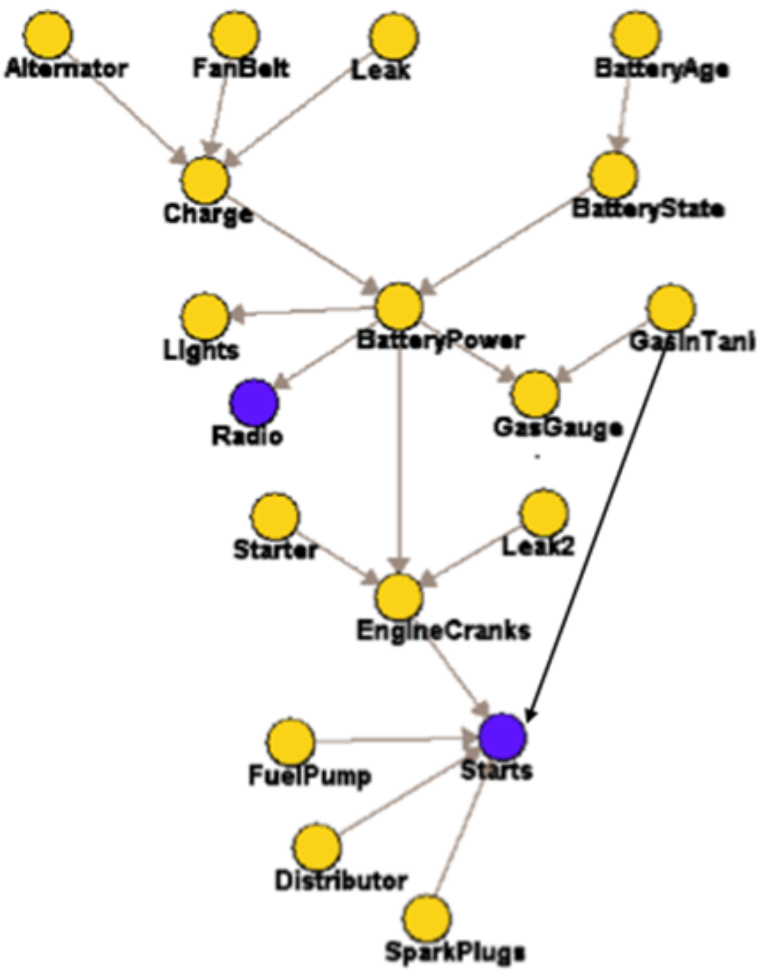
- CPD for each node X_i describes $P(X_i \mid Pa(X_i))$

Chain rule of probability :

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

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

Bayes Net



Inference:

P(BattPower=t | Radio=t, Start=f)

Most probable explanation:

What is most likely value of leak, BatteryPower given Start=f ?

Active data collection :

What is most useful variable to observe next, to improve our knowledge of node X ?

What is the Bayes Network for X1 , … Xn with NO assumed conditional independencies

Algorithm for Contructing Bayes Network

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

$$P(X_i \mid Pa(X_i)) = P(X_i \mid X_1, X_2, \ldots X_{i-1})$$

Notice this choice of parent assures

$$P(X_1 \ldots X_n) = \prod_i P(X_i \mid X_1 \ldots X_{i-1}) \quad \text{(by chain rule)}$$

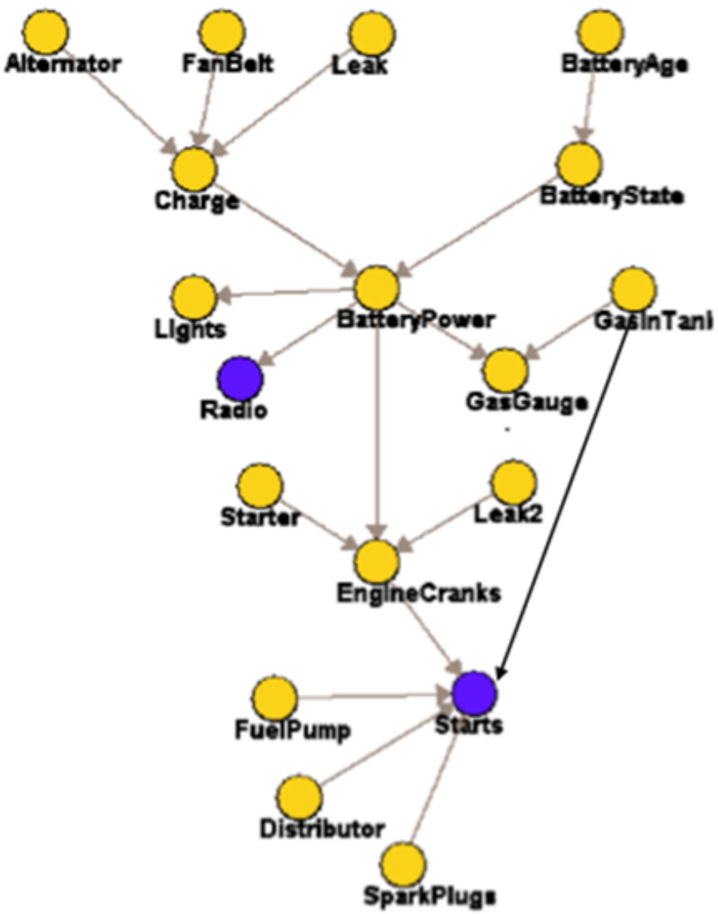
$$= \prod_i P(X_i \mid Pa(X_i)) \quad \text{(by construction)}$$

Example

- Bird flu and Allegies both cause Nasal problems
- Nasal problem cause Sneezes and Headaches

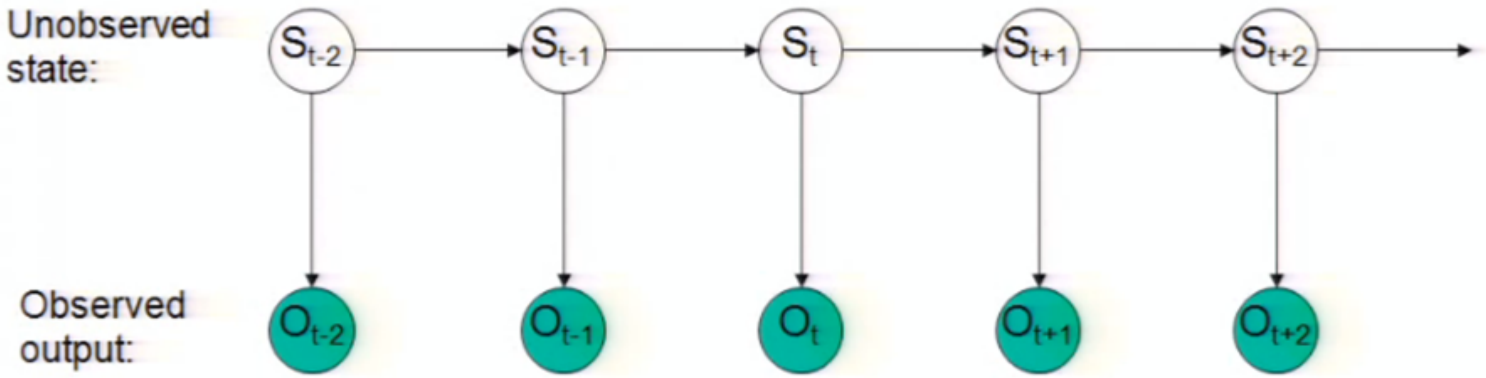
What is the Bayes Network for Naïve Bayes ?

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



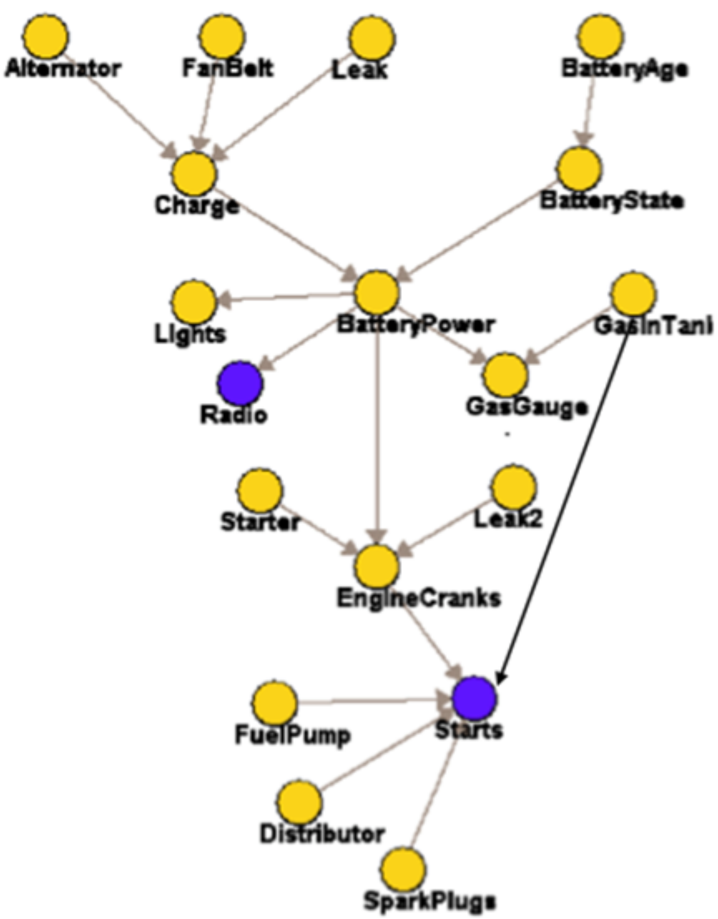
Bayes Network for a Hidden Markov Model

Assume the future is conditionally independent of the past, given the present



$$P(S_{t-2}, Q_{t-2}, S_{t-1}, \ldots, Q_{t+2}) =$$

How Can We Train a Bayes Net



1. When graph is given, and each training example gives values of every RV ?

   Easy : use data to obtain MLE or MAP estimates of θ for each CPD

   e.g. like training the CPD's of a naïve Bayes classiffier

   P(Xi | Pa(Xi); θ)

2. When graph unknown or some RV's unobserved ?

This is more difficult … later ….