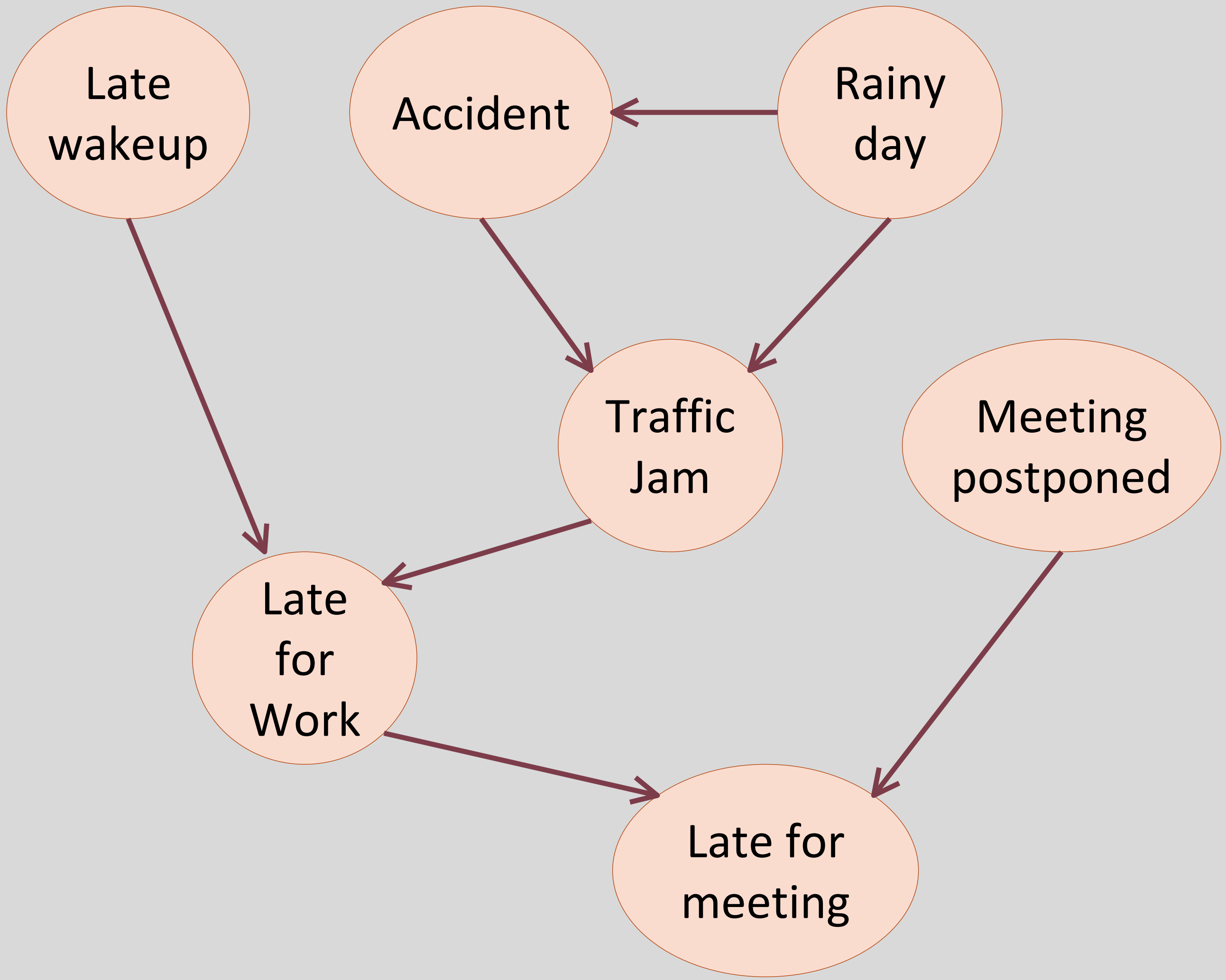
Foundations of Machine Learning

Module 4:
Part D: Bayesian Networks

Sudeshna Sarkar
IIT Kharagpur

Why Bayes Network

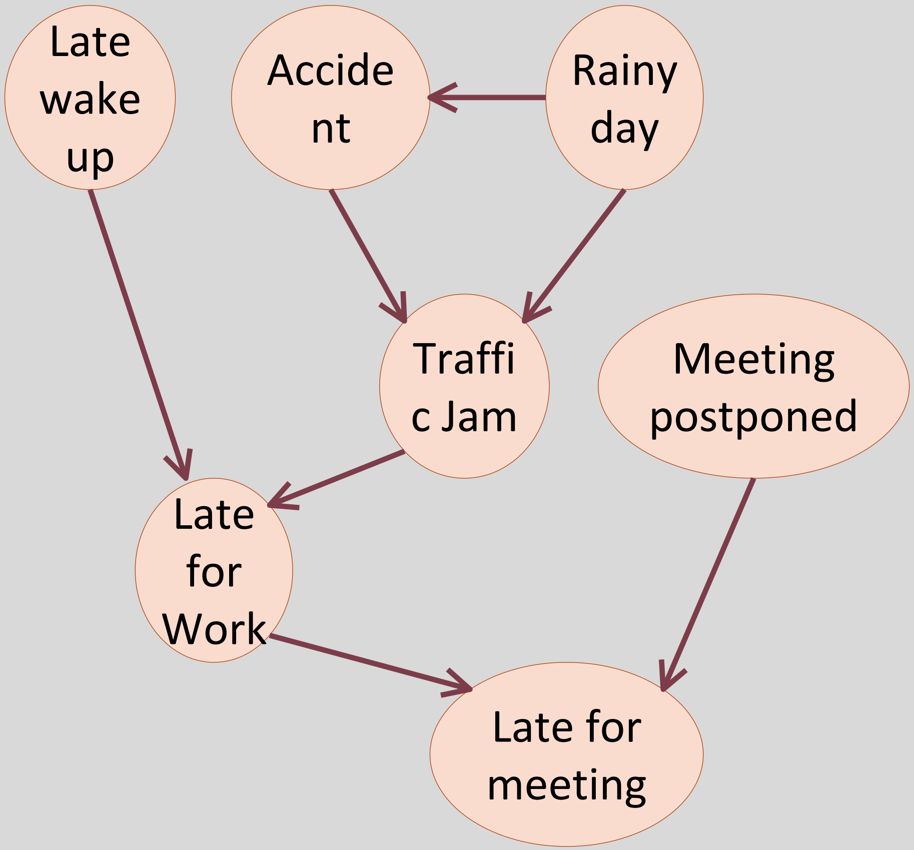
- Bayes optimal classifier is too costly to apply
- Naïve Bayes makes overly restrictive assumptions.
  - But all variables are rarely completely independent.
- Bayes network represents conditional independence relations among the features.
- Representation of causal relations makes the representation and inference efficient.


Late wakeup
Accident
Rainy day
Traffic Jam
Meeting postponed
Late for Work
Late for meeting

Bayesian Network

- A graphical model that efficiently encodes the joint probability distribution for a large set of variables
- A Bayesian Network for a set of variables (nodes) X = { X1,…….Xn}
- Arcs represent probabilistic dependence among variables
- Lack of an arc denotes a conditional independence
- The network structure S is a directed acyclic graph
- A set P of local probability distributions at each node (Conditional Probability Table)

Representation in Bayesian Belief Networks



Conditional probability table associated with each node specifies the conditional distribution for the variable given its immediate parents in the graph

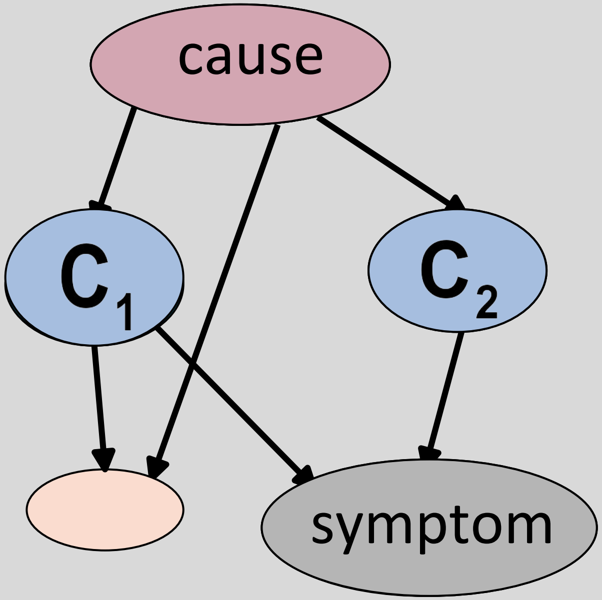
Each node is asserted to be conditionally independent of its non-descendants, given its immediate parents

Inference in Bayesian Networks

- Computes posterior probabilities given evidence about some nodes
- Exploits probabilistic independence for efficient computation.
- Unfortunately, exact inference of probabilities in general for an arbitrary Bayesian Network is known to be NP-hard.
- In theory, approximate techniques (such as Monte Carlo Methods) can also be NP-hard, though in practice, many such methods were shown to be useful.
- Efficient algorithms leverage the structure of the graph

Applications of Bayesian Networks

- Diagnosis: P(cause|symptom)=?
- Prediction: P(symptom|cause)=?
- Classification: P(class|data)
- Decision-making
  (given a cost function)



Bayesian Networks

- Structure of the graph $\Leftrightarrow$ Conditional independence relations

In general,

$$p(X_1, X_2, \ldots . X_N) = \Pi\, p(X_i \mid \text{parents}(X_i))$$

The full joint distribution

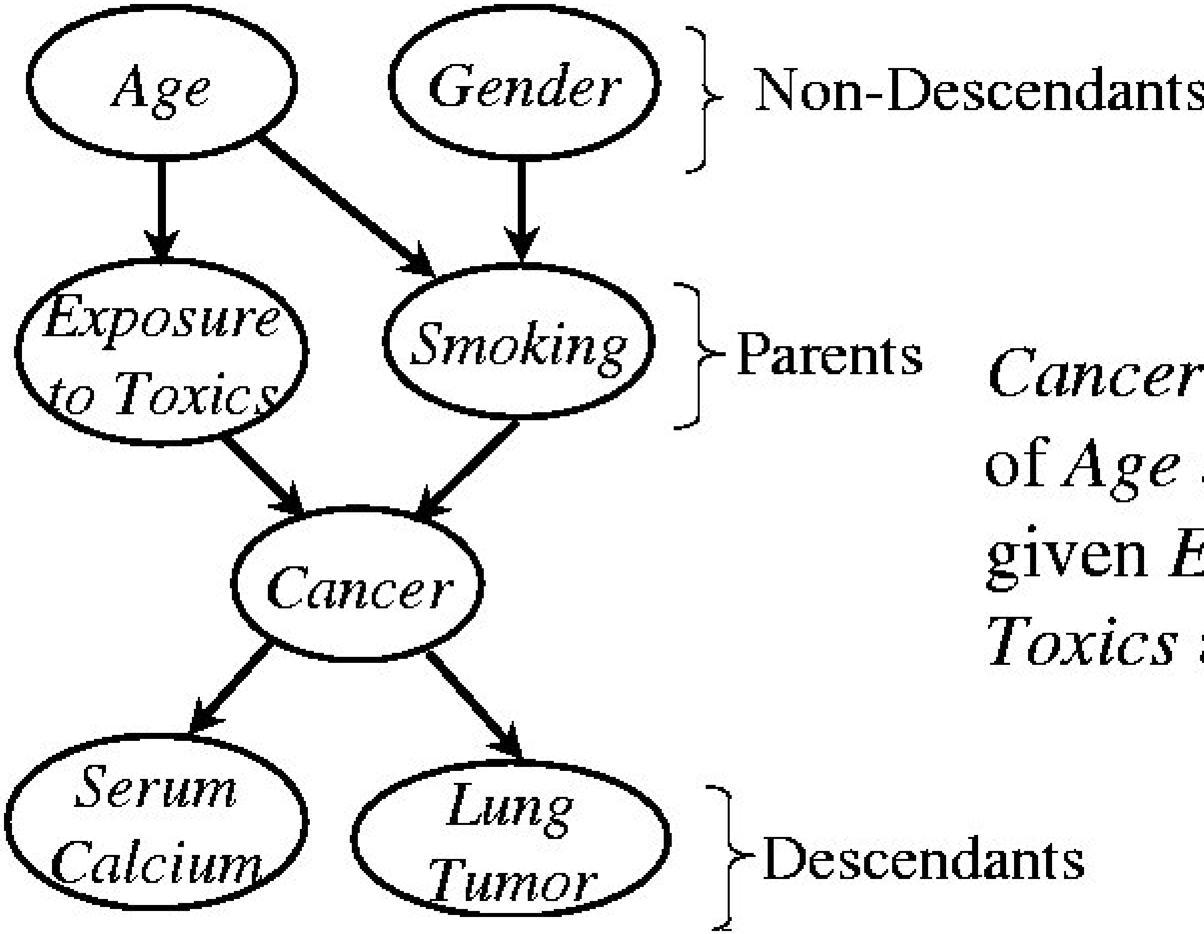
The graph-structured approximation

- Requires that graph is acyclic (no directed cycles)

- 2 components to a Bayesian network
  - The graph structure (conditional independence assumptions)
  - The numerical probabilities (for each variable given its parents)

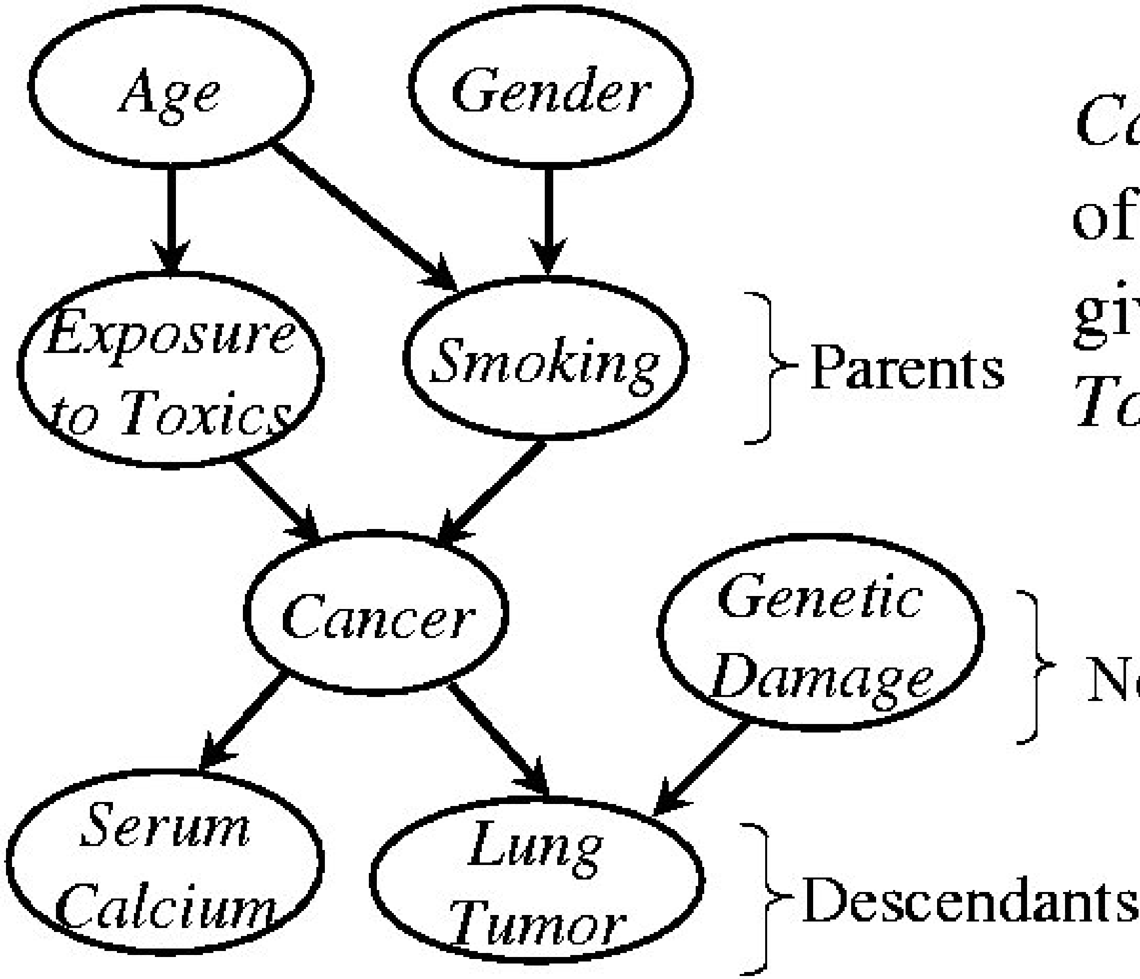
Conditional Independence

A variable (node) is conditionally independent of its non-descendants given its parents.



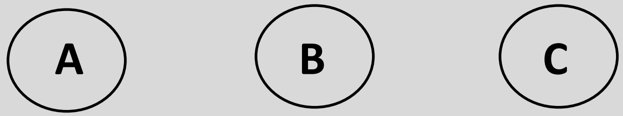
Cancer is independent of *Age* and *Gender* given *Exposure to Toxics* and *Smoking*.

Another non-descendant

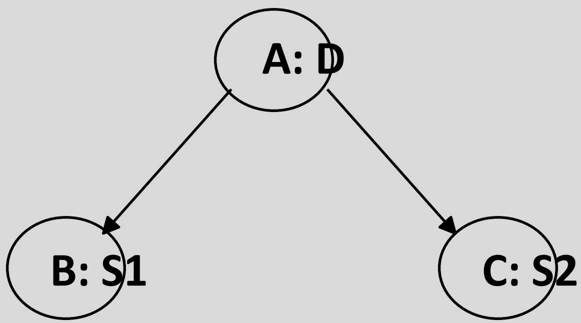


Cancer is independent of *Genetic Damage* given *Exposure to Toxics* and *Smoking*.

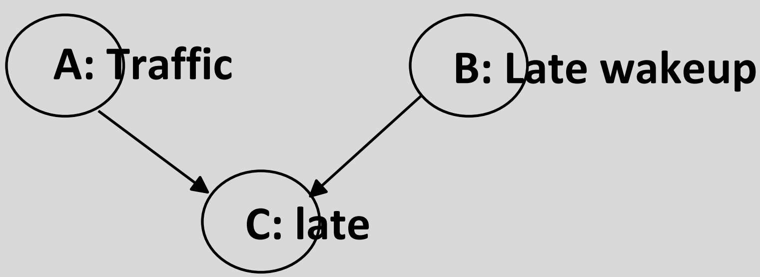
Examples



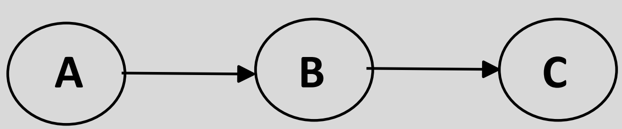
Marginal Independence:
p(A,B,C) = p(A) p(B) p(C)



Conditionally independent effects:
p(A,B,C) = p(B|A)p(C|A)p(A)
B and C are conditionally independent
Given A

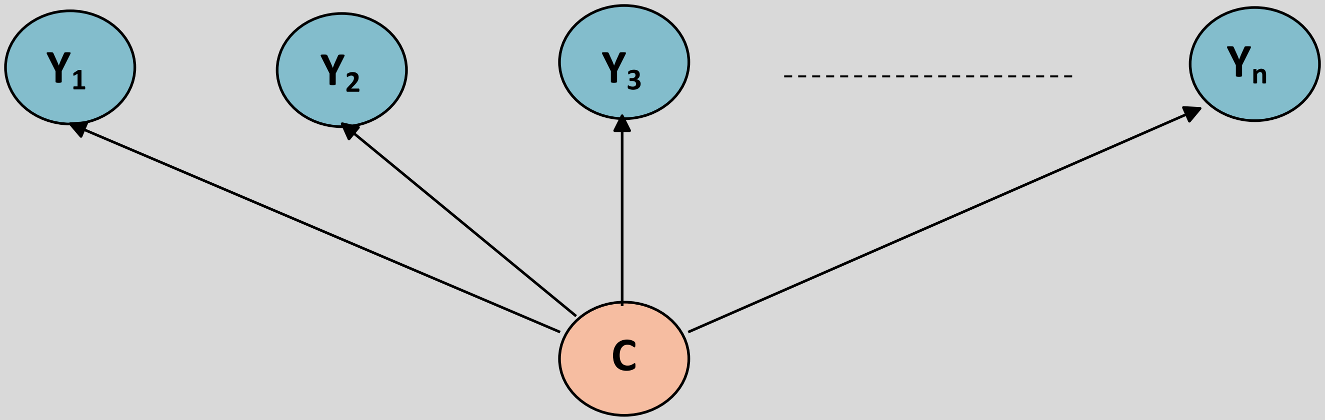


Independent Causes:
p(A,B,C) = p(C|A,B)p(A)p(B)
“Explaining away”

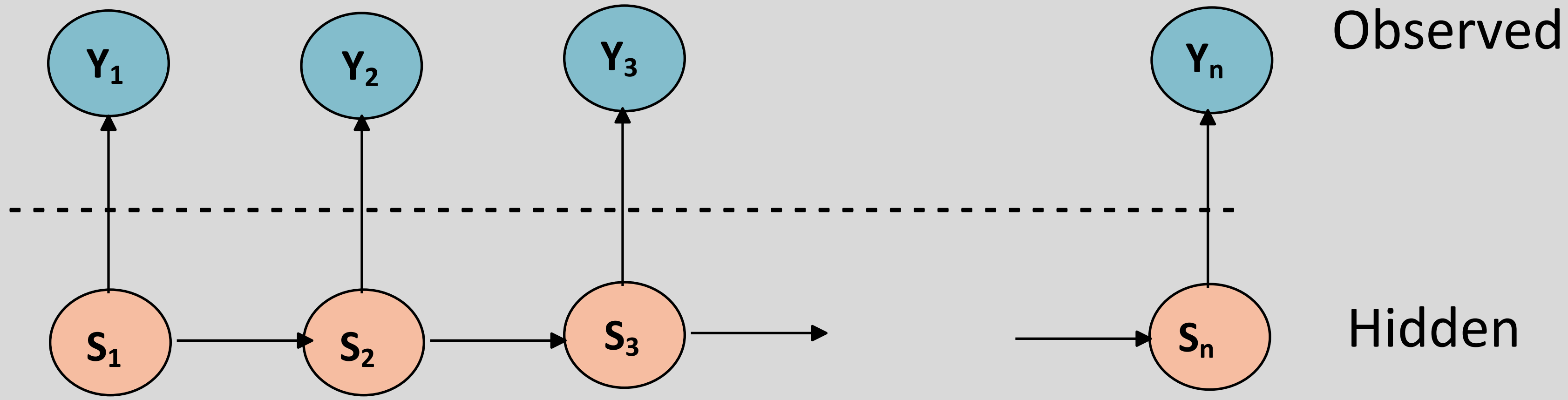


Markov dependence:
p(A,B,C) = p(C|B) p(B|A)p(A)

Naïve Bayes Model



Hidden Markov Model (HMM)



Assumptions:

1. hidden state sequence is Markov
2. observation Y_t is conditionally independent of all other variables given S_t

Widely used in sequence learning eg, speech recognition, POS tagging

Inference is linear in n

Learning Bayesian Belief Networks

1. The network structure is given in advance and all the variables are fully observable in the training examples.
   – estimate the conditional probabilities.
2. The network structure is given in advance but only some of the variables are observable in the training data.
   – Similar to learning the weights for the hidden units of a Neural Net: Gradient Ascent Procedure
3. The network structure is not known in advance.
   – Use a heuristic search or constraint-based technique to search through potential structures.

Thank You