Statistical Methods in Artificial Intelligence
CSE471 - Monsoon 2015 : Lecture 23

Avinash Sharma

CVIT, IIIT Hyderabad

Lecture Plan

- Basic concepts in Probability Theory
- Introduction to Probabilistic Graphical Models
- Bayesian Network
  - Representation
  - Conditional Independence
  - Inference
- Combining Classifiers (Next Class)

Probabilities

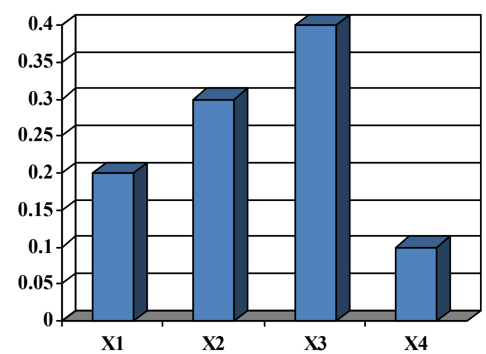
- Probability distribution $P(X|\xi)$
  - X is a random variable
    - Discrete
    - Continuous
  - ξ is background state of information

Discrete Random Variables

- Finite set of possible outcomes

$$X \in \{x_1, x_2, x_3, \ldots, x_n\}$$

$$P(x_i) \geq 0$$



$$\sum_{i=1}^{n} P(x_i) = 1$$

X binary: $P(x) + P(\overline{x}) = 1$

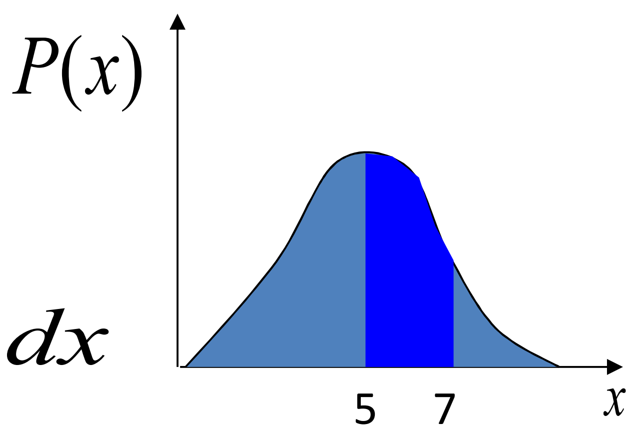
Continuous Random Variable

- Probability distribution (density function) over continuous values

$$X \in [0,10] \qquad P(x) \geq 0$$

$$\int_0^{10} P(x)dx = 1$$

$$P(5 \leq x \leq 7) = \int_5^7 P(x)dx$$



More Probabilities

- Joint

$$P(x, y) \equiv P(X = x \wedge Y = y)$$

  – Probability that both *X*=*x* and Y=y

- Conditional

$$P(x \mid y) \equiv P(X = x \mid Y = y)$$

  – Probability that *X*=*x* given we know that *Y*=*y*

Rules of Probability

- Product Rule

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

- Marginalization

$$P(Y) = \sum_{i=1}^{n} P(Y, x_i)$$

X binary: $P(Y) = P(Y,x) + P(Y,\bar{x})$

Bayes Rule

$$P(H,E) = P(H \mid E)P(E) = P(E \mid H)P(H)$$

$$\boxed{P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}}$$

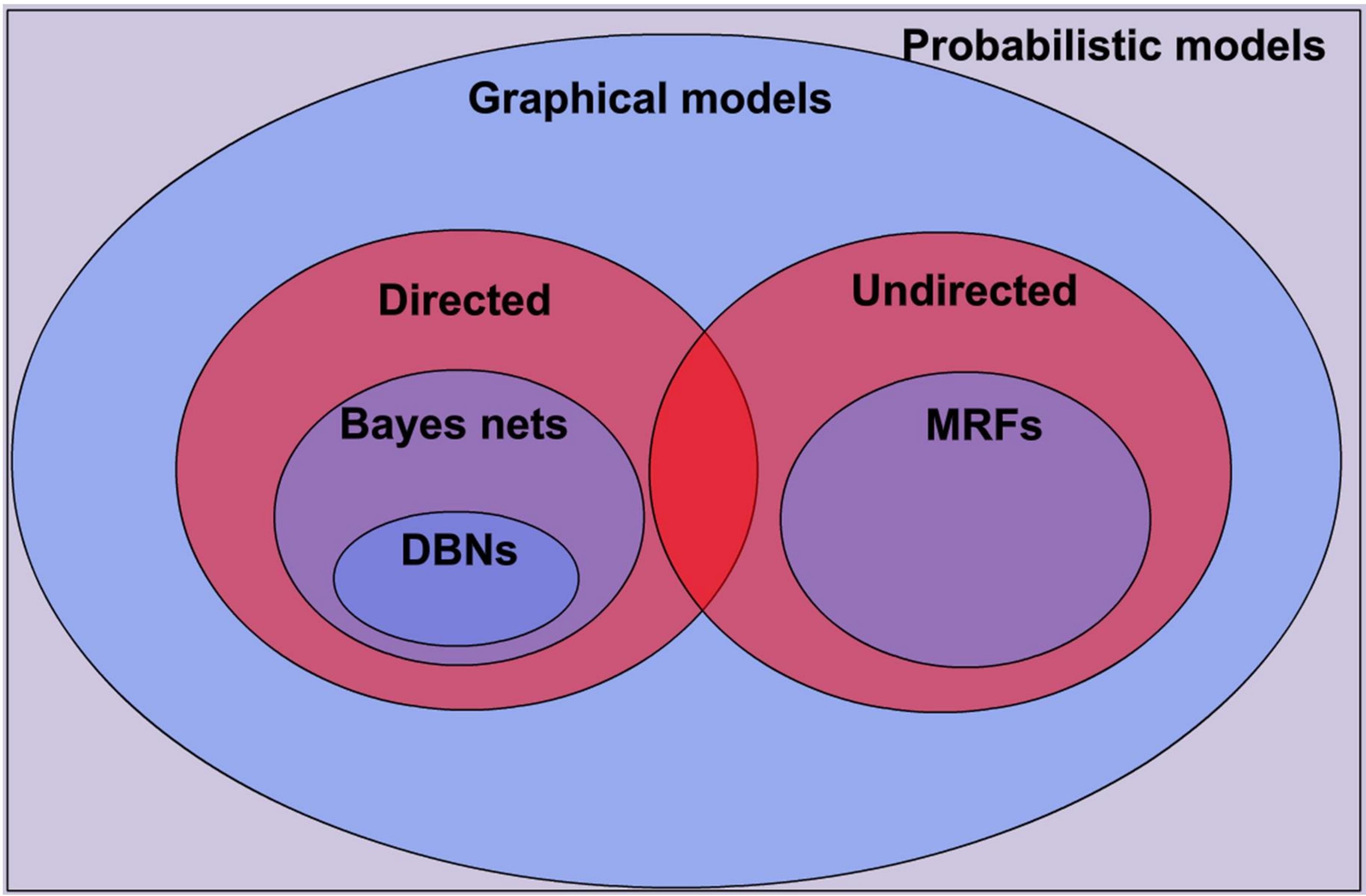
Introduction to GM's

- A Graphical Model (GM) is a visual, abstract, and mathematically formal description of properties of families of probability distributions (densities, mass functions)
- GM's are a marriage between probability theory and graph theory where the graph theoretic side provides
  - an intuitively appealing interface for modeling highly-interacting sets of variables
  - a data structure that lends itself naturally to the design of efficient general-purpose algorithms.
- GM's provide a natural tool for dealing with uncertainty and complexity in engineering & applied mathematics domain.
- The notion of modularity is fundamental to the idea of a GM
  - a complex system is built by combining simpler parts.
  - Probability theory provides the glue whereby the parts are combined.

Introduction to GM's

- GM's provides:
  - **Structure**: A method to explore the structure of "natural" phenomena (causal vs. correlated relations, properties of natural signals)
  - **Algorithms**: A set of algorithms that provide "efficient" probabilistic inference and statistical decision making
  - **Language**: A mathematically formal, abstract, visual language with which to efficiently discuss families of probabilistic models.
  - **Approximation**: Methods to explore systems of approximation and their implications. E.g., what are the consequences of a (perhaps known to be) wrong assumption?
  - **Data-base**: Provide a probabilistic "data-base" and corresponding "search algorithms" for making queries about properties in such model families.

Classes of GM's



Representation in Bayesian Net

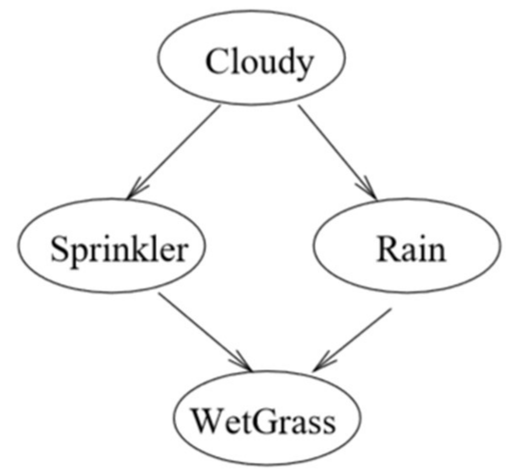
- Probabilistic GM's are graphs in which nodes represent random variables, and the (absence of) arcs represent conditional (in)dependence. **Hence they provide a compact representation of joint probability distributions.**

- *Undirected* graphical models, also called Markov Random Fields (MRFs) or Markov networks, have a simple definition of independence that two (sets of) nodes A and B are conditionally independent given a third set, C, if all paths between the nodes in A and B are separated by a node in C.

- *Directed* graphical models also called Bayesian Networks or Belief Networks (BNs), have a more complicated notion of independence, which takes into account the directionality of the arcs.

Representation in Bayesian Net

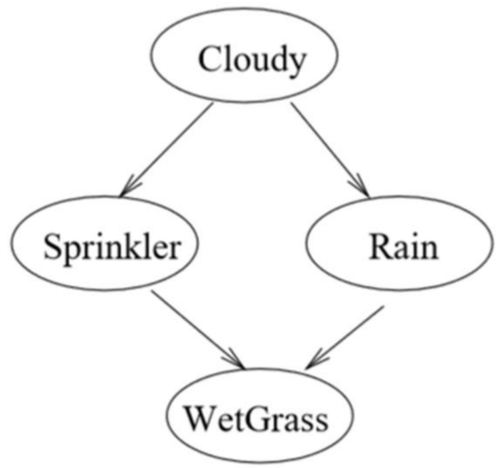
- The directed graphical models have a more complicated notion of independence than undirected models, they do have following advantages.
  - One can regard an arc from A to B as indicating that A ``causes'' B. This can be used as a guide to construct the graph structure.
  - Directed models can encode deterministic relationships, and are easier to learn (fit to data).



- Bayesian networks are so called as they use Bayes' rule for inference.

Representation in Bayesian Net

P(C=F)	P(C=T)
0.5	0.5



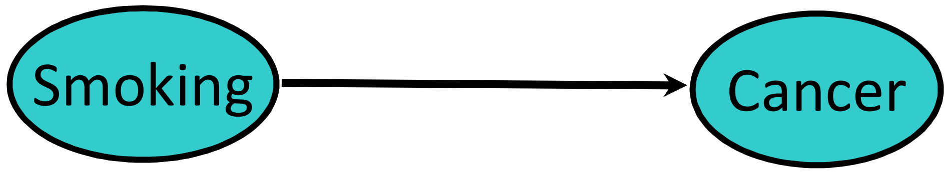
C	P(S=F)	P(S=T)
F	0.5	0.5
T	0.9	0.1

C	P(R=F)	P(R=T)
F	0.8	0.2
T	0.2	0.8

S	R	P(W=F)	P(W=T)
F	F	1.0	0.0
T	F	0.1	0.9
F	T	0.1	0.9
T	T	0.01	0.99

Bayesian Networks

$S \in \{no, light, heavy\}$



P(S=no)	0.80
P(S=light)	0.15
P(S=heavy)	0.05

$C \in \{none, benign, malignant\}$

Smoking=	no	light	heavy
P(C=none)	0.96	0.88	0.60
P(C=benign)	0.03	0.08	0.25
P(C=malig)	0.01	0.04	0.15

Product Rule

- $P(C,S) = P(C|S)\,P(S)$

$S\Downarrow \quad C\Rightarrow$	*none*	*benign*	*malignant*
no	0.768	0.024	0.008
light	0.132	0.012	0.006
heavy	0.035	0.010	0.005

Marginalization

$S\Downarrow \quad C\Rightarrow$	*none*	*benign*	*malig*	total
no	0.768	0.024	0.008	.80
light	0.132	0.012	0.006	.15
heavy	0.035	0.010	0.005	.05
total	0.935	0.046	0.019	

P(Smoke)

P(Cancer)

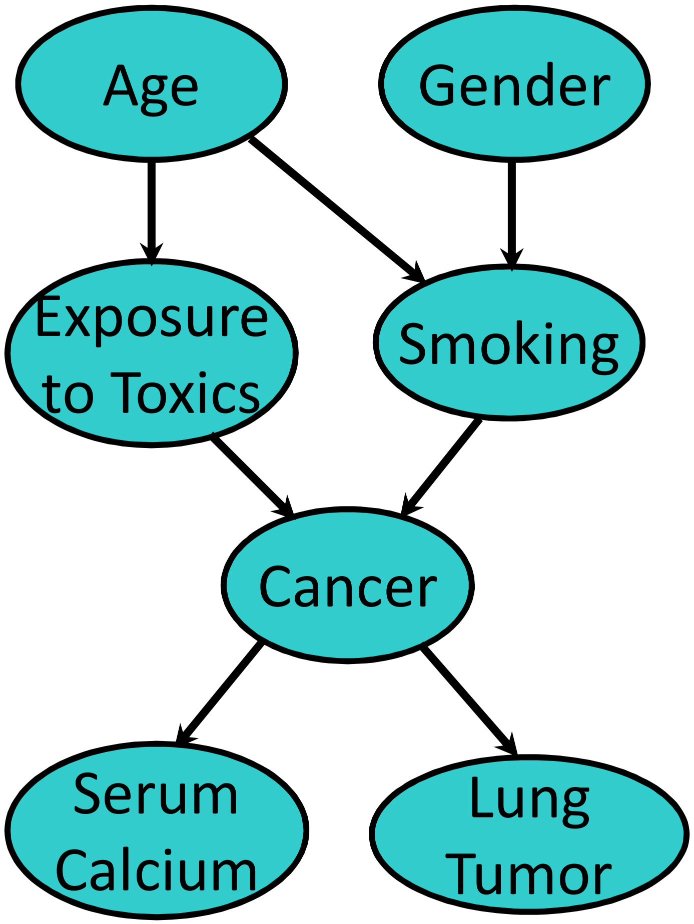
Bayes Rule Revisited

$$P(S \mid C) = \frac{P(C \mid S)P(S)}{P(C)} = \frac{P(C,S)}{P(C)}$$

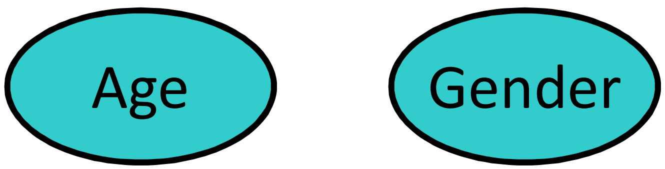
$S\Downarrow \quad C\Rightarrow$	*none*	*benign*	*malig*
no	0.768/.935	0.024/.046	0.008/.019
light	0.132/.935	0.012/.046	0.006/.019
heavy	0.030/.935	0.015/.046	0.005/.019

Cancer=	none	benign	malignant
P(S=no)	0.821	0.522	0.421
P(S=light)	0.141	0.261	0.316
P(S=heavy)	0.037	0.217	0.263

A Bayesian Network



Independence



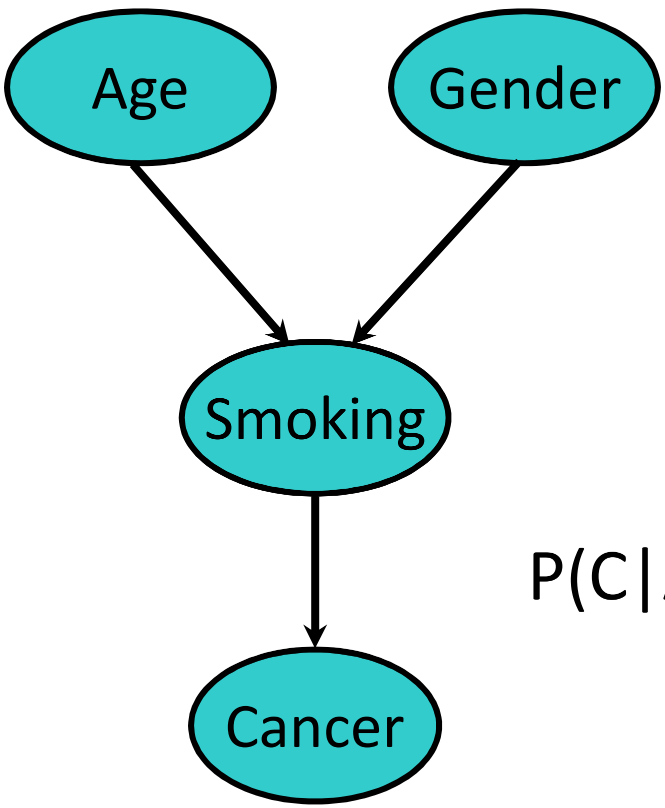
Age and Gender are independent.

$$P(A,G) = P(G)P(A)$$

$$P(A|G) = P(A) \quad A \perp G$$
$$P(G|A) = P(G) \quad G \perp A$$

$$P(A,G) = P(G|A)\,P(A) = P(G)P(A)$$
$$P(A,G) = P(A|G)\,P(G) = P(A)P(G)$$

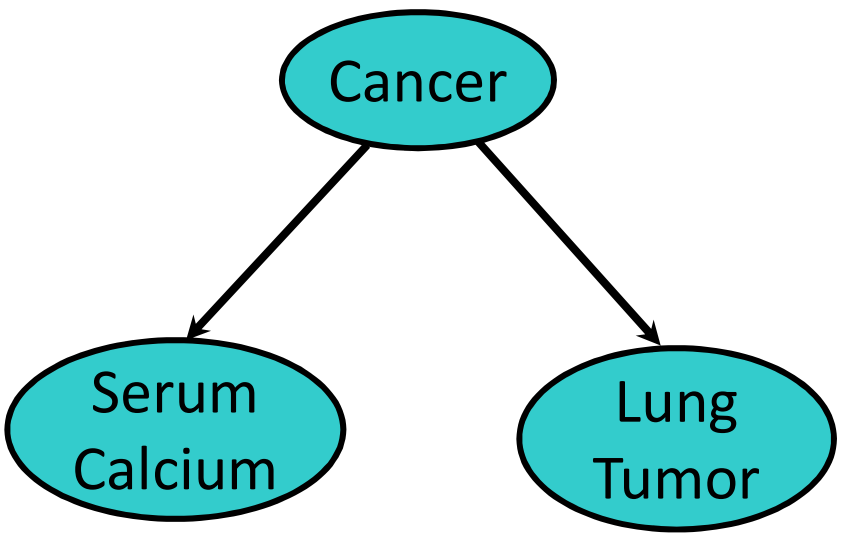
Conditional Independence



Cancer is independent of Age and Gender given Smoking.

$P(C|A,G,S) = P(C|S) \quad C \perp A,G \mid S$

More Conditional Independence: Naïve Bayes

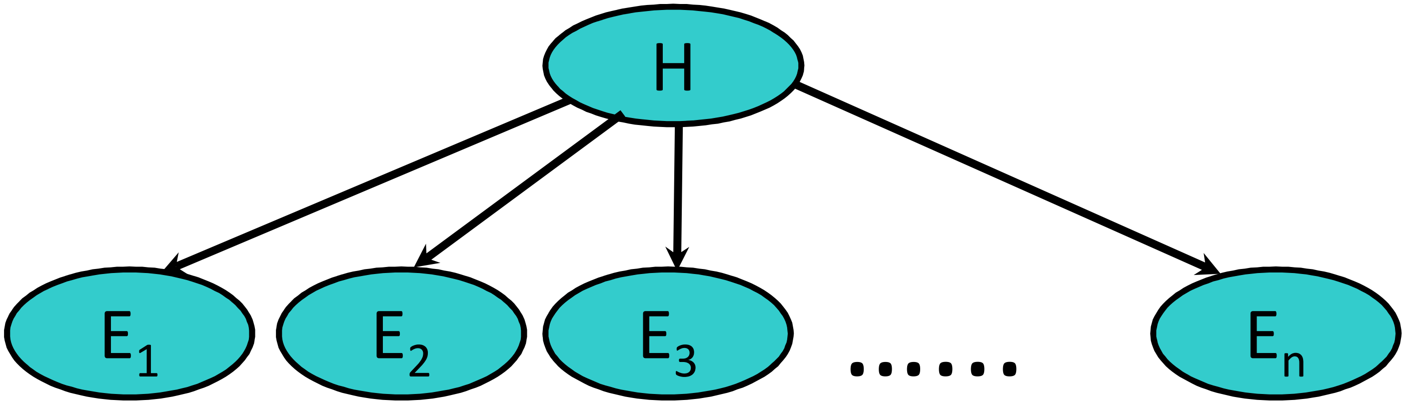


Serum Calcium and Lung Tumor are dependent

Serum Calcium is independent of Lung Tumor, given Cancer

$$P(L|SC,C) = P(L|C)$$

Naïve Bayes in general

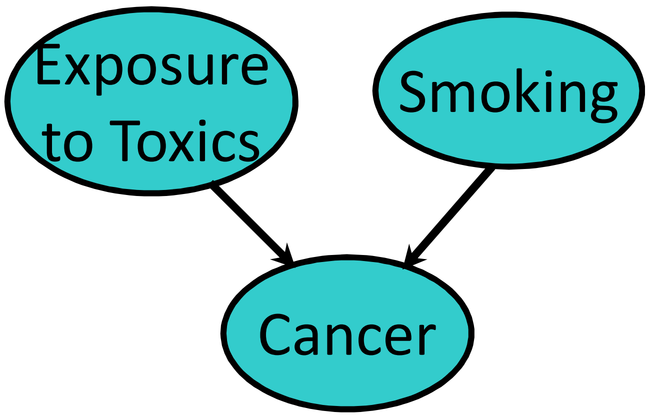


2n + 1 parameters:

$$P(h)$$

$$P(e_i \mid h), P(e_i \mid \overline{h}),\ \ i = 1, \ldots, n$$

More Conditional Independence: Explaining Away



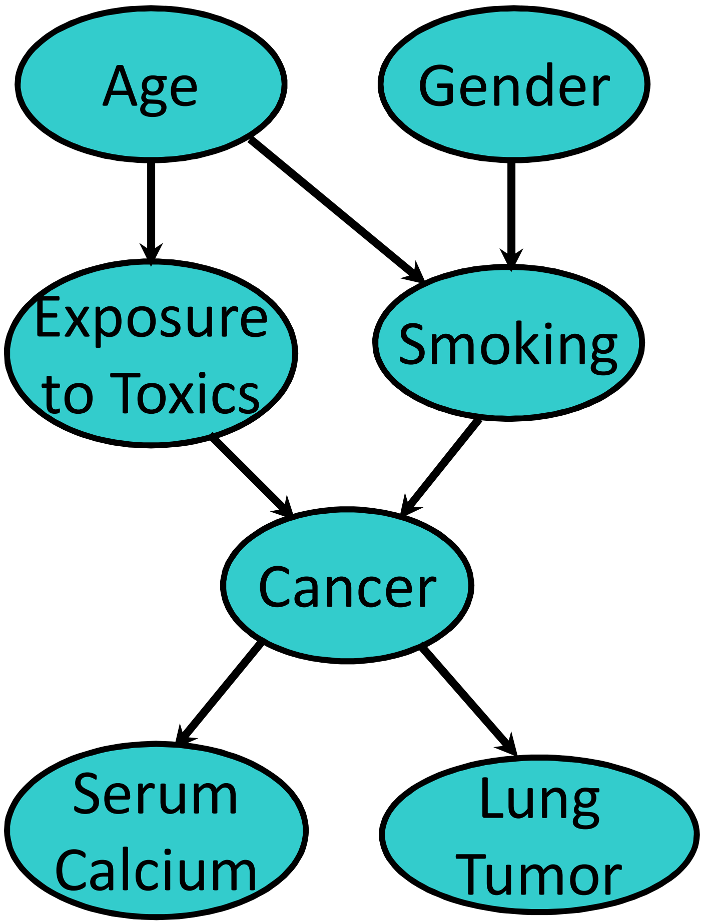
Exposure to Toxics and Smoking are independent

$$E \perp S$$

Exposure to Toxics is **dependent** on Smoking, given Cancer

$$P(E = \text{heavy} \mid C = \text{malignant}) >$$
$$P(E = \text{heavy} \mid C = \text{malignant}, S=\text{heavy})$$

Put it all together



$$P(A, G, E, S, C, L, SC) =$$

$$P(A) \cdot P(G) \cdot$$

$$P(E \mid A) \cdot P(S \mid A, G) \cdot$$

$$P(C \mid E, S) \cdot$$

$$P(SC \mid C) \cdot P(L \mid C)$$

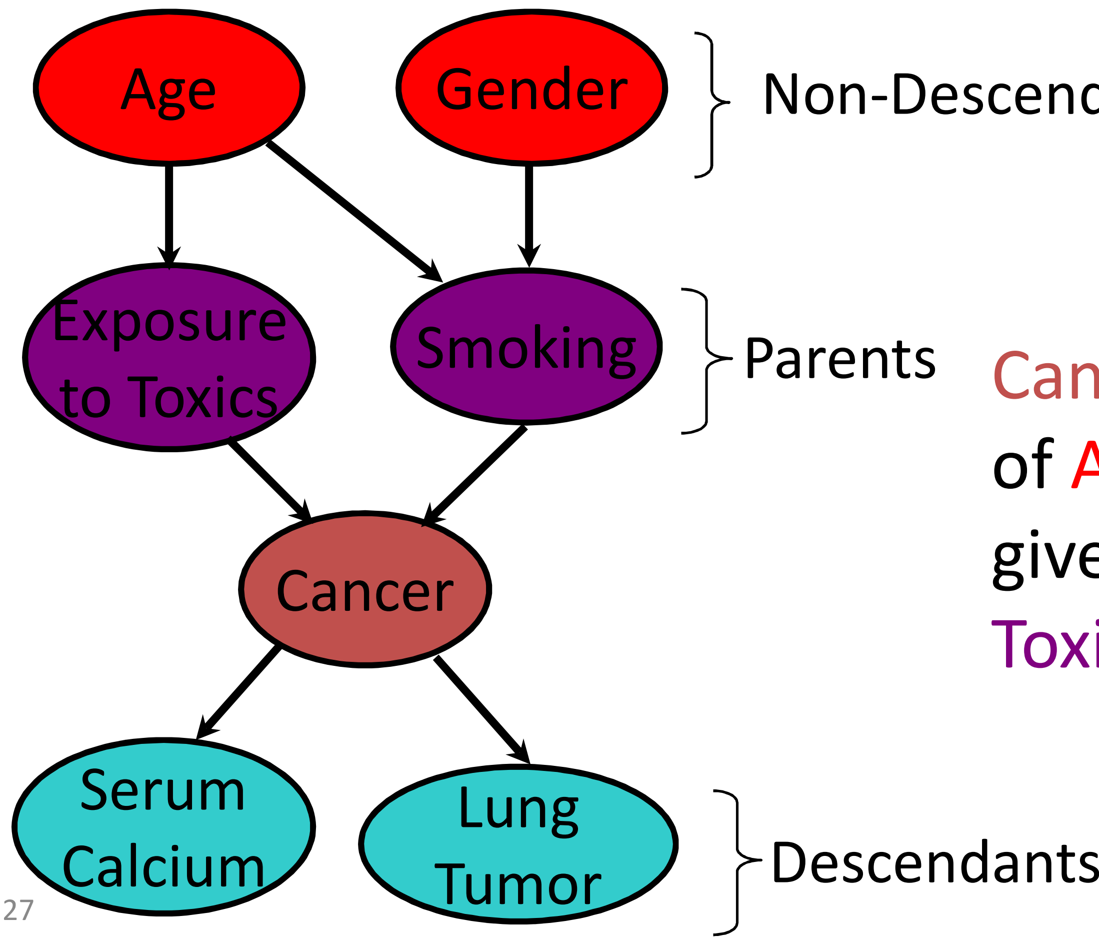
General Product (Chain) Rule for Bayesian Networks

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \boldsymbol{Pa_i})$$

$\mathbf{Pa_i}$=parents(X_i)

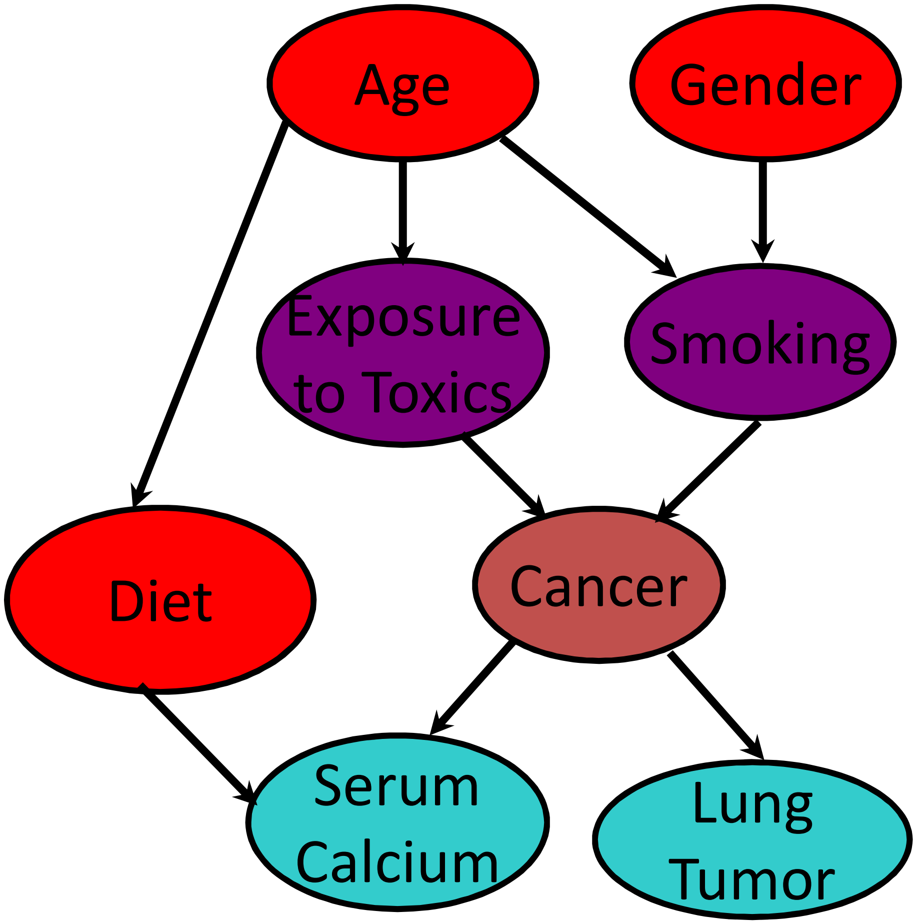
Conditional Independence

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

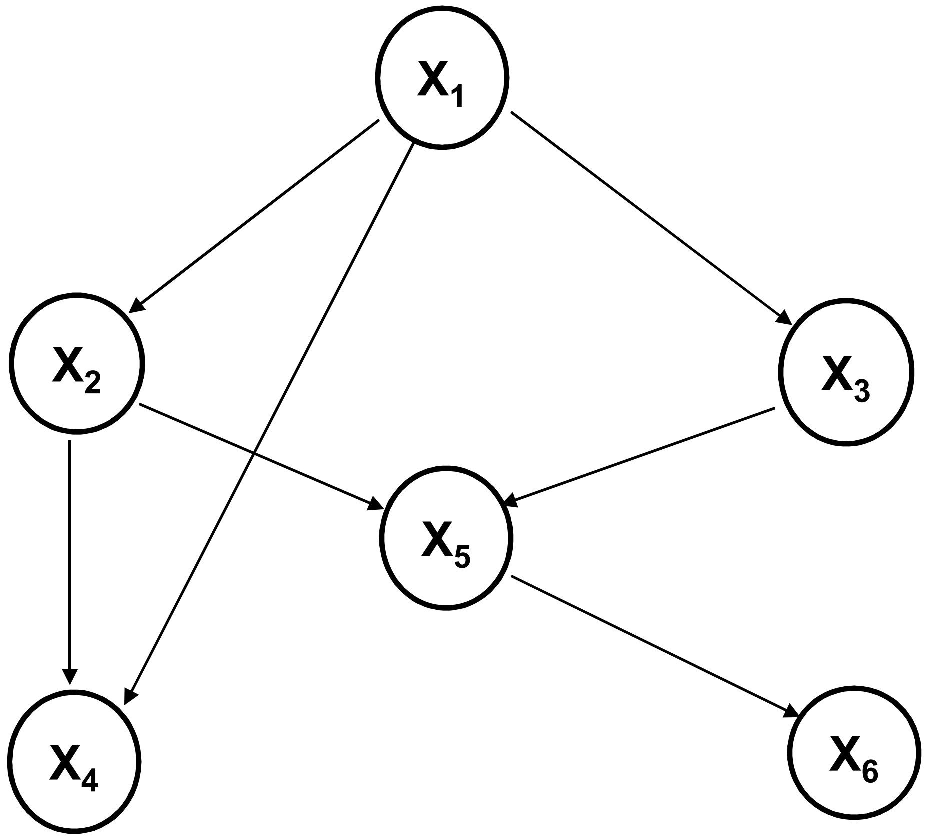


Cancer is independent of Diet given Exposure to Toxics and Smoking.

Independence and Graph Separation

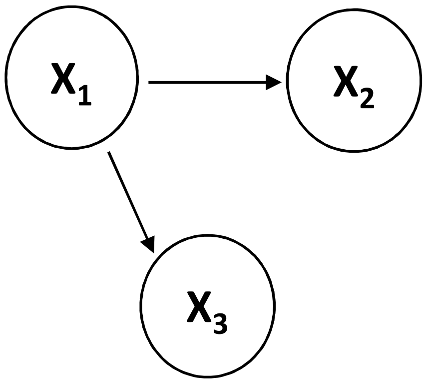
- Given a set of observations, is one set of variables dependent on another set?
- Observing effects can induce dependencies.
- d-separation (Pearl 1988) allows us to check conditional independence graphically.

Sample of General Product Rule



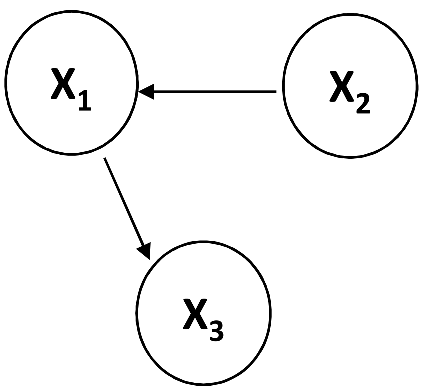
$$p(x_1, x_2, x_3, x_4, x_5, x_6) = p(x_6 \mid x_5)\, p(x_5 \mid x_3, x_2)\, p(x_4 \mid x_2, x_1)\, p(x_3 \mid x_1)\, p(x_2 \mid x_1)\, p(x_1)$$

Arc Reversal - Bayes Rule



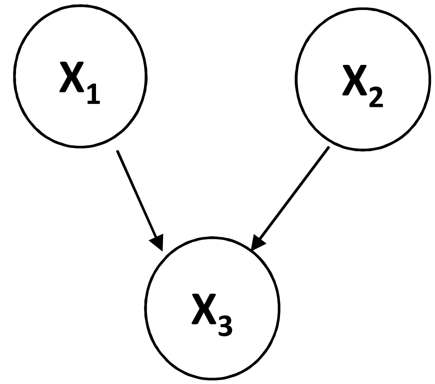
$$p(x_1, x_2, x_3) = p(x_3 \mid x_1)\, p(x_2 \mid x_1)\, p(x_1)$$

is equivalent to



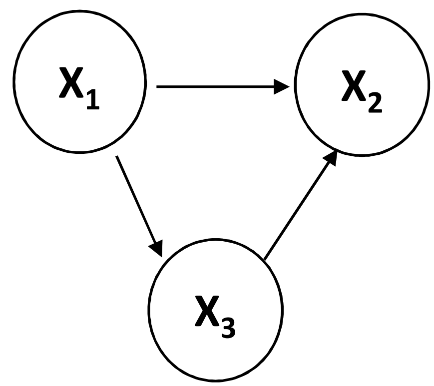
$$p(x_1, x_2, x_3) = p(x_3 \mid x_1)\, p(x_2 , x_1)$$

$$= p(x_3 \mid x_1)\, p(x_1 \mid x_2)\, p(x_2)$$



$$p(x_1, x_2, x_3) = p(x_3 \mid x_2, x_1)\, p(x_2)\, p(x_1)$$

is equivalent to



$$p(x_1, x_2, x_3) = p(x_3, x_2 \mid x_1)\, p(x_1)$$

$$= p(x_2 \mid x_3, x_1)\, p(x_3 \mid x_1)\, p(x_1)$$

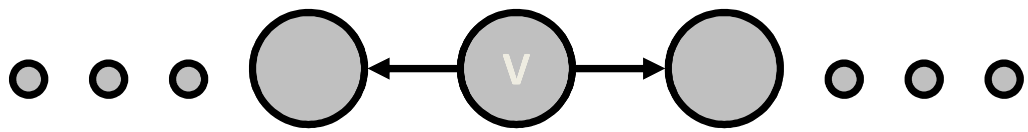
D-Separation of variables

- Fortunately, there is a relatively simple algorithm for determining whether two variables in a Bayesian network are conditionally independent: ***d-separation***.

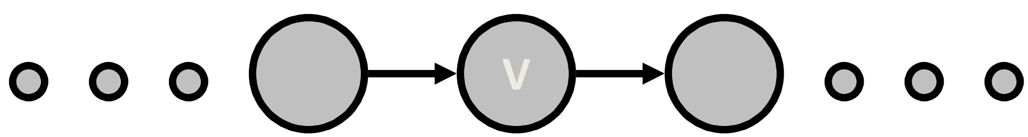
- Definition: X and Z are *d-separated* by a set of evidence (observed) variables E iff every undirected path from X to Z is "blocked".
- A path is "blocked" iff one or more of the following conditions is true: ...

A path is blocked when:

- There exists a variable *V* on the path such that
  - it is in the evidence set *E* (Observed Variables)
  - the arcs putting *V* in the path are "tail-to-tail"



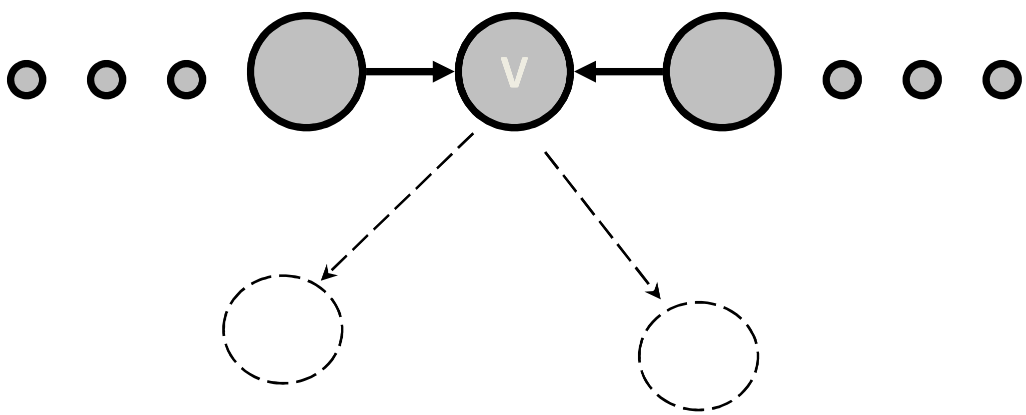
- Or, there exists a variable *V* on the path such that
  - it is in the evidence set *E* (Observed Variables)
  - the arcs putting *V* in the path are "tail-to-head"



- Or, ...

… a path is blocked when:

- … Or, there exists a variable V on the path such that
  - it is NOT in the evidence set E (Observed Variables)
  - neither are any of its descendants
  - the arcs putting V on the path are "head-to-head"



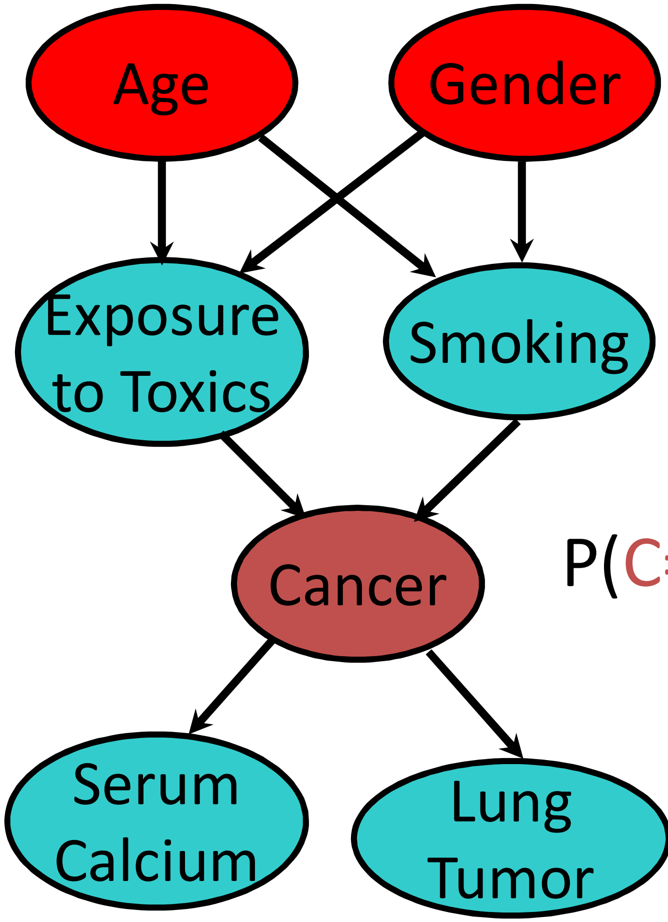
D-Separation and independence

- Theorem [Verma & Pearl, 1998]:
  - If a set of evidence variables E **d-separates** X and Z in a Bayesian network's graph, then X and Z will be independent.
- *d*-separation can be computed in linear time.
- Thus we now have a fast algorithm for automatically inferring whether learning the value of one variable might give us any additional hints about some other variable, given what we already know.

Inference in Bayesian Network

- The general probabilistic inference problem is to find the probability of an event given a set of evidence;
- This can be done in Bayesian nets with sequential applications of Bayes Theorem;
- In 1986 Judea Pearl published an innovative algorithm for performing inference in Bayesian nets.

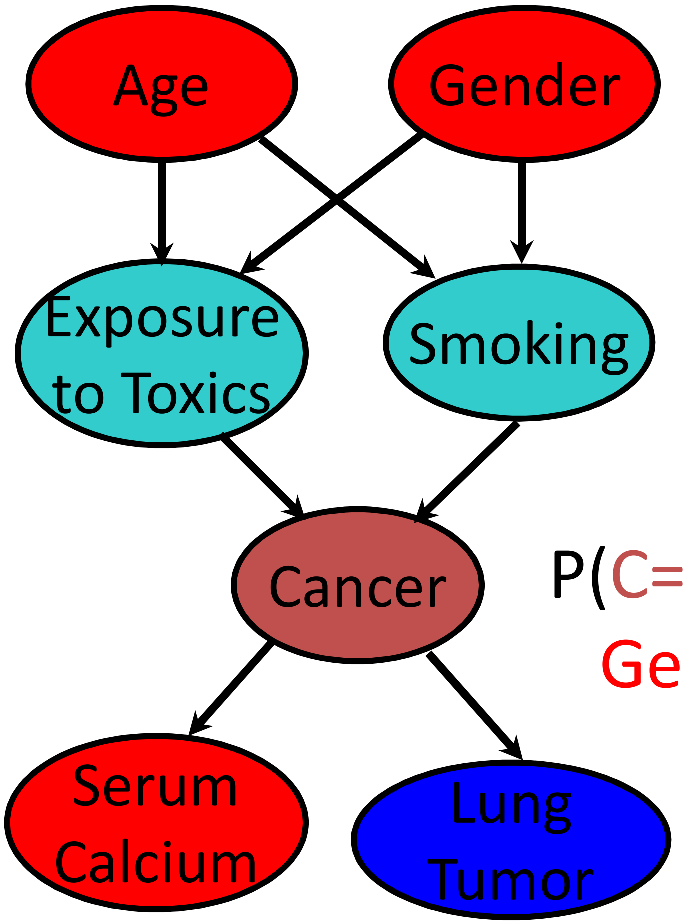
Predictive Inference



How likely are elderly males to get malignant cancer?

P(C=malignant | Age>60, Gender= male)

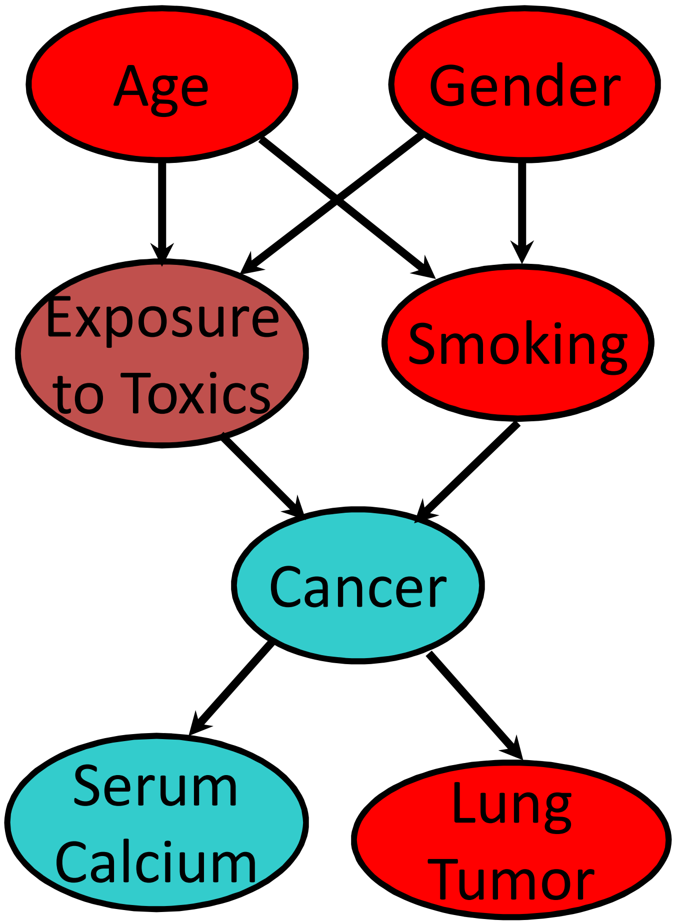
Combined



How likely is an elderly male patient with high Serum Calcium to have malignant cancer?

P(C=malignant | Age>60, Gender= male, Serum Calcium = high)

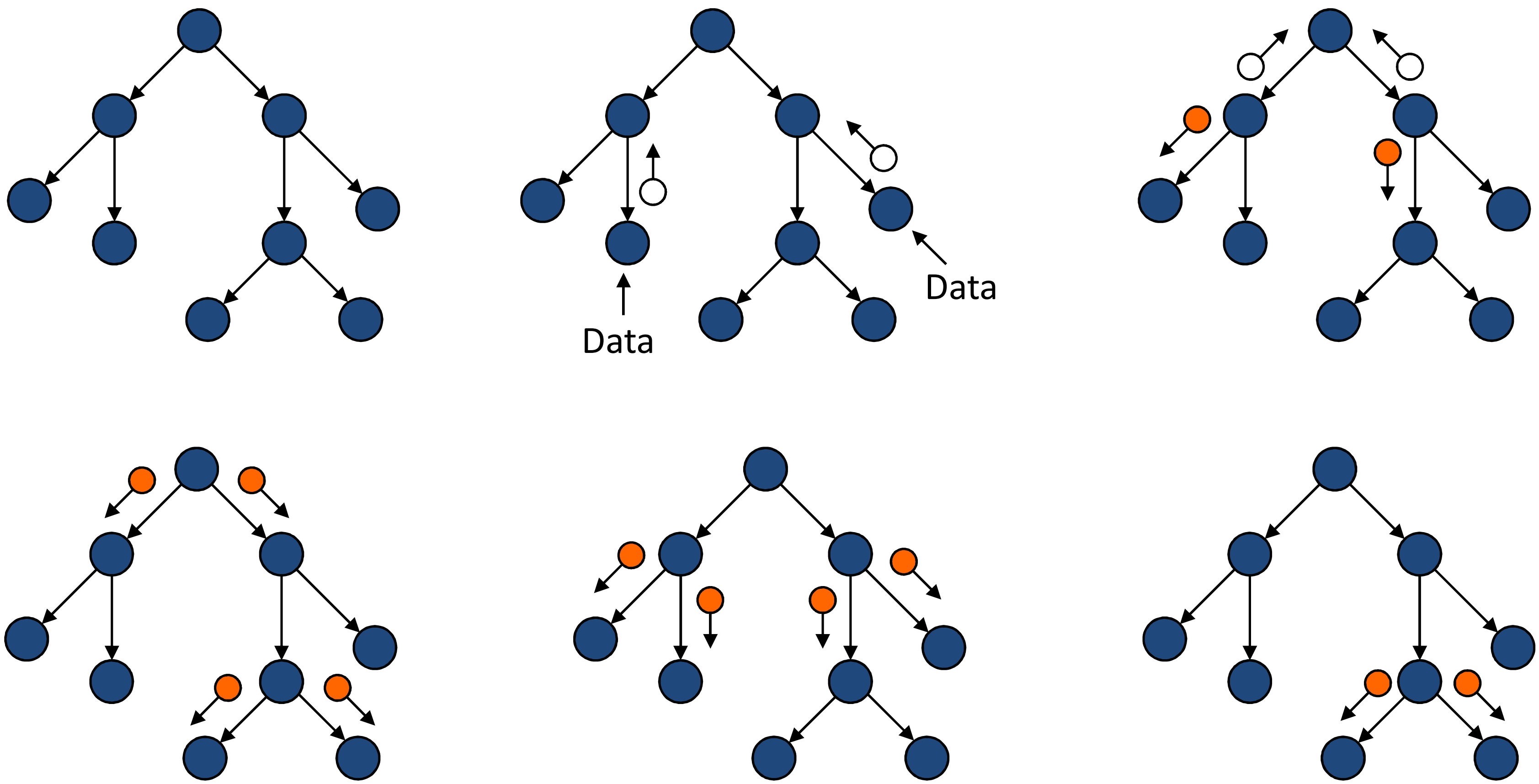
Explaining away



- If we see a lung tumor, the probability of heavy smoking and of exposure to toxics both go up.
- If we then observe heavy smoking, the probability of exposure to toxics goes back down.

Propagation Example

"The impact of each new piece of evidence is viewed as a perturbation that propagates through the network via message-passing between neighboring variables . . ." (Pearl, 1988, p 143`



- The example above requires five time periods to reach equilibrium after the introduction of data (Pearl, 1988, p 174)

References

- http://www.cs.ubc.ca/~murphyk/Bayes/bnintro.html
- Introduction to Bayesian Networks
- Tutorial on Bayesian Networks