Knowledge Representation and Reasoning

Introduction to Probabilistic Reasoning System

Nur Ulfa Maulidevi
(ulfa@informatika.org)

KK IF – Teknik Informatika – STEI ITB

Inteligensi Artifisial
(Artificial Intelligence)

Probabilistic Reasoning System (PRS)

Agent in real world need to handle uncertainty

Logical Agents handle uncertainty by disjunction → cannot tell us how likely the difference conditions are

Probability theory provides a quantitative way of encoding likelihood

Given state(s) e, what is the probabilty that x happens → P(x|e)

- Joint Probability Distribution
- Bayesian/ Belief Network

Joint Probability Distribution

- Random variables
  - Function: discrete domain → [0, 1]
  - Sums to 1 over the domain
    - Raining is a propositional random variable
    - Raining(true) = 0.2
      - P(Raining = true) = 0.2
    - Raining(false) = 0.8
      - P(Raining = false) = 0.8
- Joint distribution
  - Probability assignment to all combinations of values of random variables

Inference using Joint Probability Distribution

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- For any proposition ϕ, sum the atomic events where it is true: $P(\phi) = \Sigma_{\omega:\omega \models \phi} P(\omega)$
- P(*toothache*) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2
- P(*cavity*) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2
- P(cavity U toothache) = ?
- P(¬*cavity* | *toothache*) = P(¬*cavity* ∩ *toothache*) / P(*toothache*)

  = (0.016+0.064) / (0.108 + 0.012 + 0.016 + 0.064)

  = 0.4

Inference using Joint Probability Distribution

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

- If you have n binary propositional variables → requires 2^n numbers to build Joint Probability Distribution

→ Bayesian Network (We want to exploit independences in the domain)

Bayes' Rule:

$P(A \mid B) = P(A \cap B) / P(B)$

$= P(B \mid A)\, P(A) / P(B)$

Conditioning:

$P(A) = P(A \mid B)\, P(B) + P(A \mid \neg B)\, P(\neg B)$

$= P(A \cap B) + P(A \cap \neg B)$

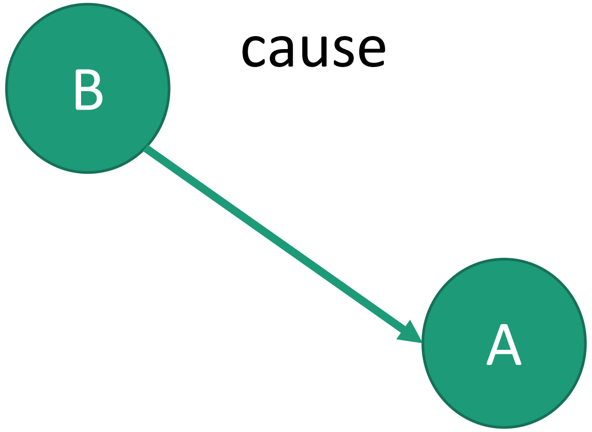
Structure of Bayesian Network

Nodes (variable)

Directed arc

P(B)

cause



effect

$P(A|B)$
$P(A|\neg B)$

Should be
Directed Acyclic Graph (DAG)

Numerical Parameters

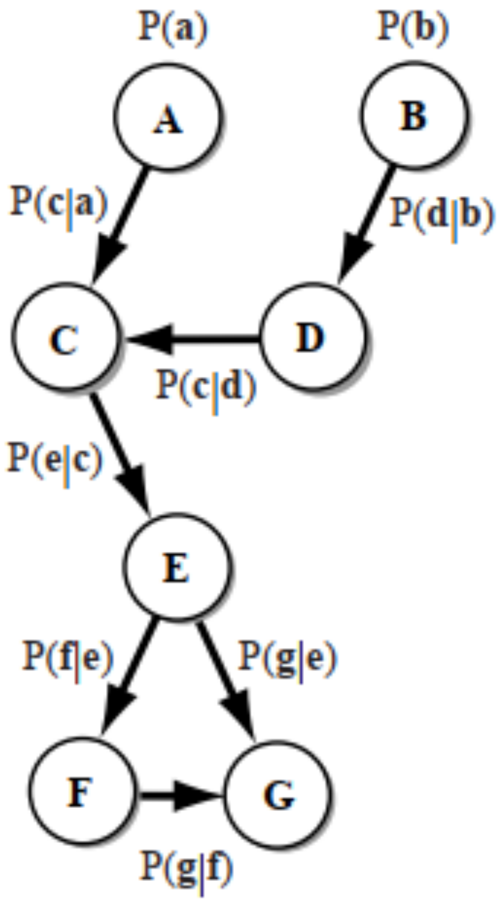
- Prior probability
- Probability conditional tables

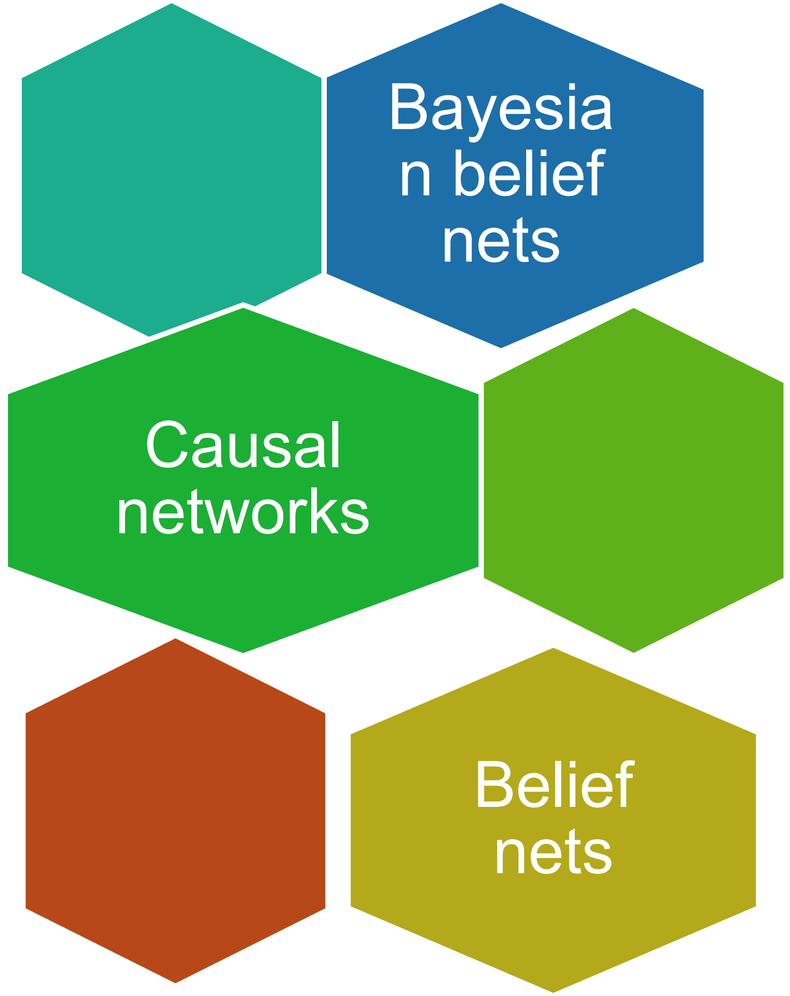
Structure of Bayesian Network

Representation of causal dependencies graphically (Hart et al.,2001)

BN have capability probabilistic reasoning like full joint probability distribution. It can answer any question about the domain.

-- How we exploit Independence?



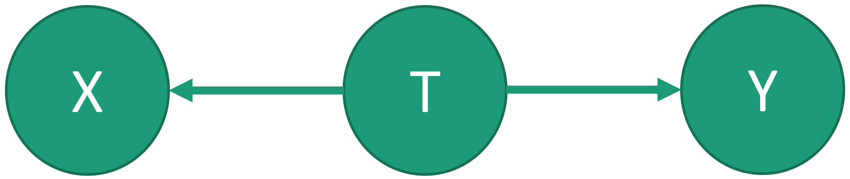


Independence

- A and B are independent iff
  - $P(A \cap B) = P(A) \cdot P(B)$
  - $P(A \mid B) = P(A)$
  - $P(B \mid A) = P(B)$
- Independence is essential for efficient probabilistic reasoning
- A and B are conditionally independent given C iff
  - $P(A \mid B, C) = P(A \mid C)$
  - $P(B \mid A, C) = P(B \mid C)$
  - $P(A \cap B \mid C) = P(A \mid C) \cdot P(B \mid C)$

Example of Independence

- X is late (X)
- Traffic Jam (T)
- Y is late (Y)
- None of these propositions are independent of one other
- X and Y are conditionally independent given T

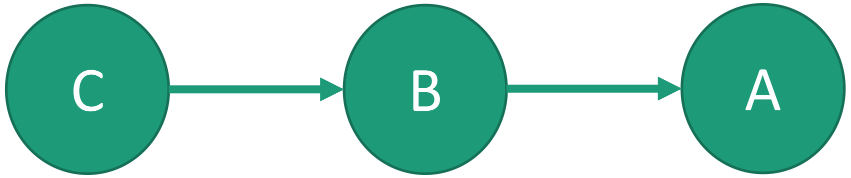


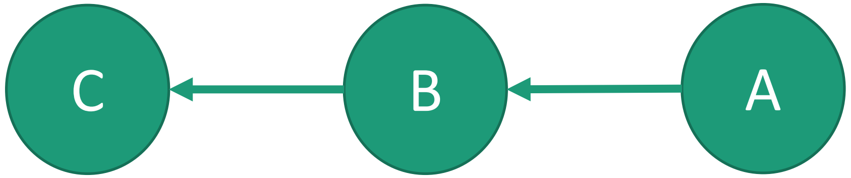
Types of Connections in Bayesian Network

Forward

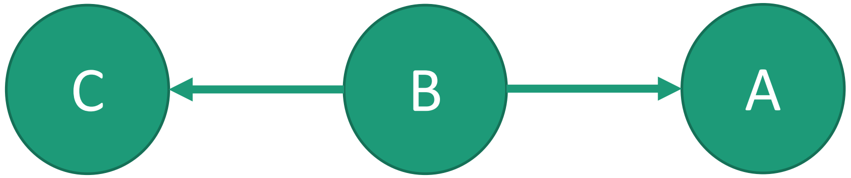
Backward

Serial



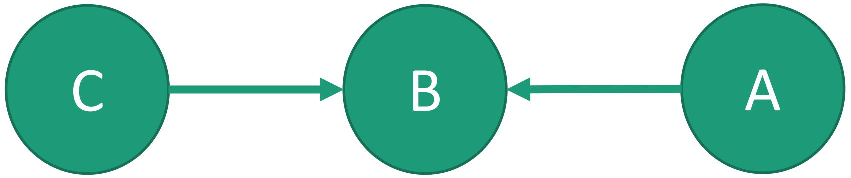


Diverging



Why? To exploit Independence

Converging



Independence in Connection

Serial

$C \rightarrow B \rightarrow A$

Diverging

$C \leftarrow B \rightarrow A$

Converging

$C \rightarrow B \leftarrow A$

Knowing C will tell us about A, but if we know B, knowing C will tell us nothing about A (C and A conditionally independent or d-separated)

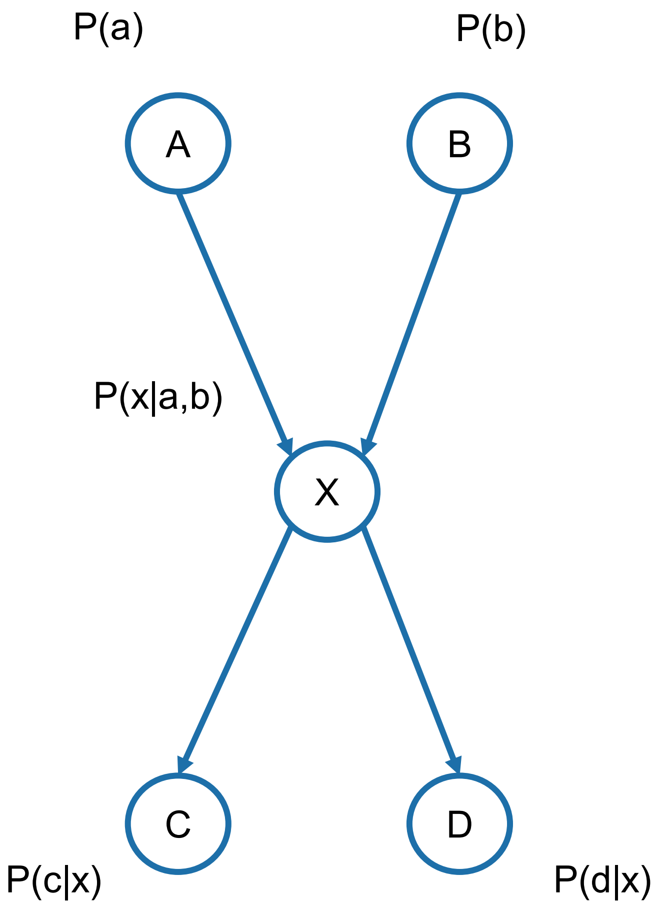
Knowing C will tell us nothing about A without knowing B, but if we see evidence about B, C and A becomes dependent

Reduction Number of Probabilities

In a domain with N binary propositional variables (2 possibilities value), one needs 2^N numbers to specify the joint probability distribution. N=5: need 32 probabilities

Independence and conditional independence relationships among variables can greatly reduce the number of probabilities that need to be specified in order to define the full joint distribution (Russel & Norvig, 2013)

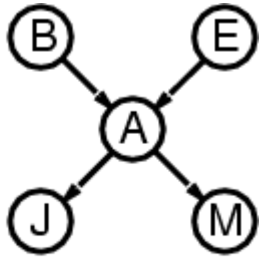
For 5 binary variables with causal networks: need 2+2+8+4+4=20 probabilities (or 10 with complements).



Bayesian Network as Joint Probability Distribution (chain rule)

The full joint distribution is defined as the product of the local conditional distributions:

$$\mathbf{P}(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i | Parents(X_i))$$



e.g., $\mathbf{P}(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

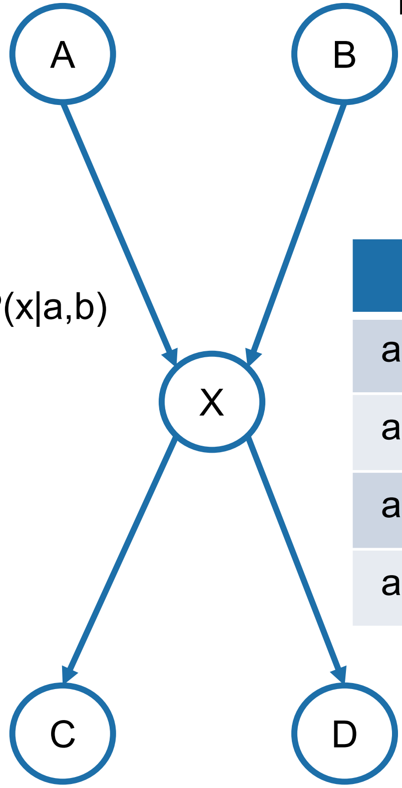
$= \mathbf{P}(j \mid a)\ \mathbf{P}(m \mid a)\ \mathbf{P}(a \mid \neg b, \neg e)\ \mathbf{P}(\neg b)\ \mathbf{P}(\neg e)$

Example

- We can determine the value of any entry in the joint probability.
- P(a2,b1,x2,c2,d1)
=P(a2)P(b1)P(x2|a2,b1)P(c2|x2)P(d1|x2)
=0.75*0.6*0.4*0.4*0.6
=0.0432

P(a1)=0.25

P(b1)=0.6



P(x|a,b)

	P(x1|a,b)
a1,b1	0.3
a1,b2	0.7
a2,b1	0.6
a2,b2	0.8

P(c|x)

	P(c1|x)
x1	0.3
x2	0.6

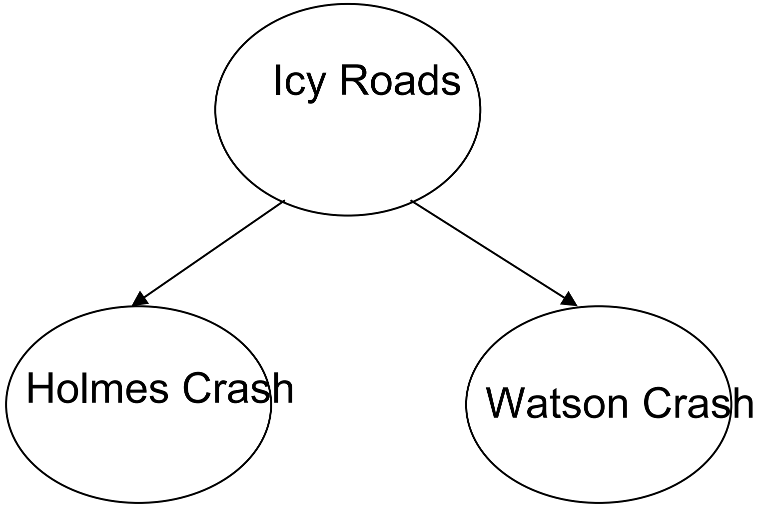
P(d|x)

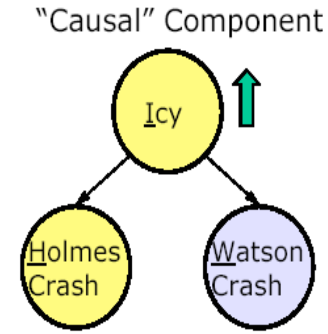
	P(d1|x)
x1	0.3
x2	0.6

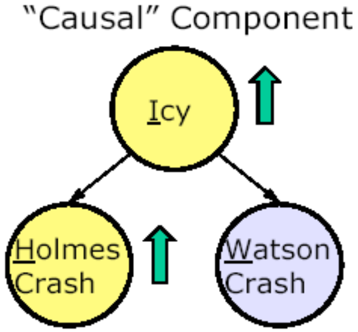
Example of Bayesian Network

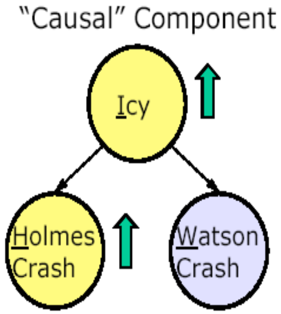
Icy Roads

Inspector Smith is waiting for Holmes and Watson, who are driving (separately) to meet him. It is winter. His secretary tells him that Watson has had an accident. He says, “It must be that the roads are icy. I bet that Holmes will have an accident too. I should go to lunch.” But, his secretary says, “No, the roads are not icy, look at the window.” So, he says, “I guess I better wait for Holmes.”

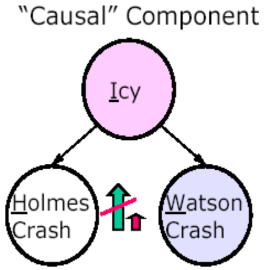




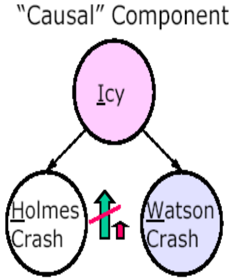




H and W are dependent,



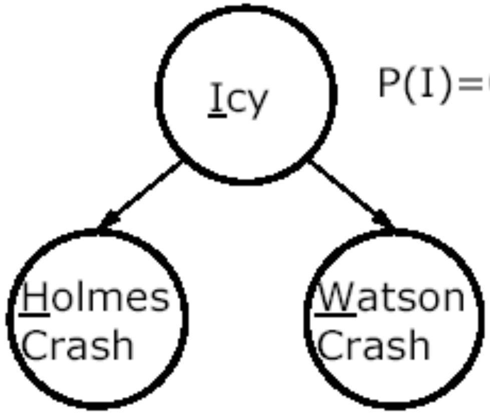
H and W are dependent,



H and W are dependent, but conditionally independent given I

Icy Roads with Numbers

P(W | I) = prob watson kecelakaan JIKA icy roads
P(W | ¬I) = prob watson kecelakaan JIKA TIDAK icy roads



$P(I)=0.7$

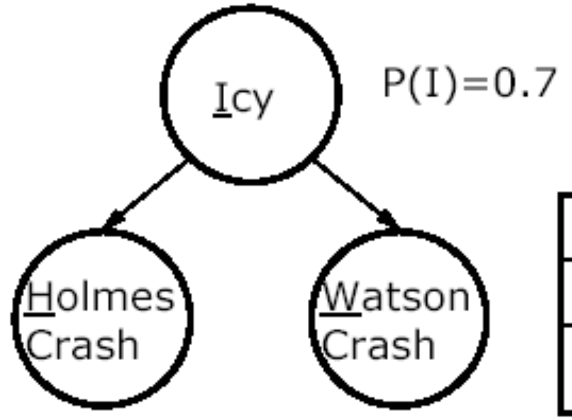
	P(H| I)
I	0.8
¬I	0.1

	P(W| I)
I	0.8
¬I	0.1

Probability that Watson Crashes:

$$
\begin{aligned}
P(W) &= P(W|\ I)\ P(I) + P(W|\ \neg I)\ P(\neg I) \\
&= 0.8 \cdot 0.7 + 0.1 \cdot 0.3 \\
&= 0.56 + 0.03 \\
&= 0.59
\end{aligned}
$$

Icy Roads with Numbers



	P(H| I)
I	0.8
¬I	0.1

	P(W| I)
I	0.8
¬I	0.1

Probability of Icy given Watson (Bayes' Rule):

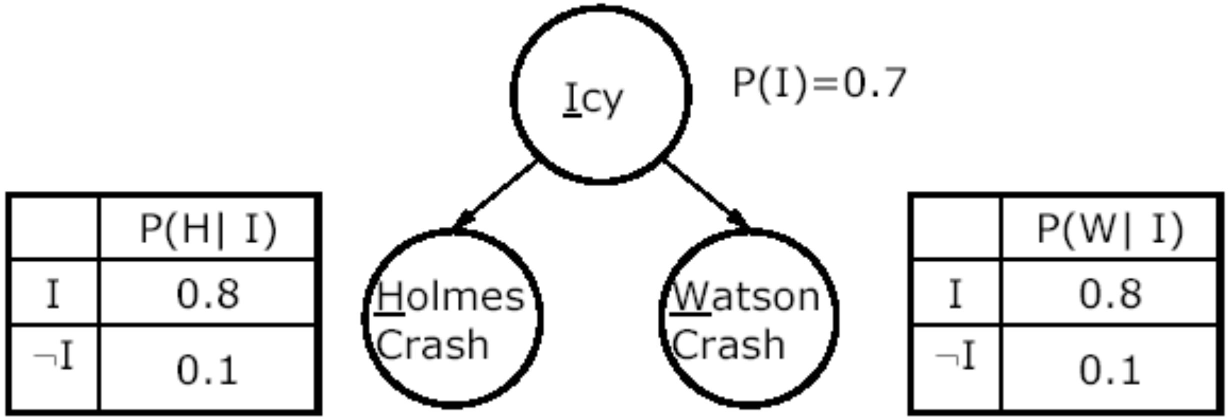
$P(I \mid W) = P(W \mid I)\, P(I) / P(W)$

$= 0.8 \cdot 0.7 / 0.59$

$= 0.95$

We started with P(I) = 0.7; knowing that Watson crashed raised the probability to 0.95

Icy Roads with Numbers



Probability of Holmes given Watson :

$P(H|W) = P(H|W,I)P(I|W) + P(H|W,\neg I)\, P(\neg I|\, W)$

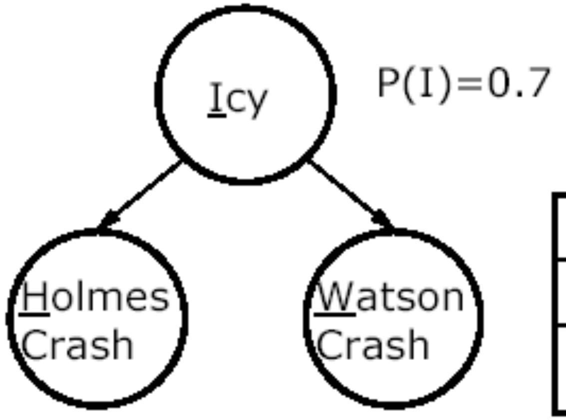
$= P(H|I)P(I|W) + P(H|\neg I)\, P(\neg I|\, W)$

$= 0.8 \cdot 0.95 + 0.1 \cdot 0.05$

$= 0.765$

We started with P(H) = 0.59; knowing that Watson crashed raised the probability to 0.765

Icy Roads with Numbers



	P(H| I)
I	0.8
¬I	0.1

	P(W| I)
I	0.8
¬I	0.1

Probability of Holmes given Icy and Watson :
$P(H|W, \neg I) = P(H|\neg I) = 0.1$

H and W are d-separated given I, so H and W are conditionally independent given I

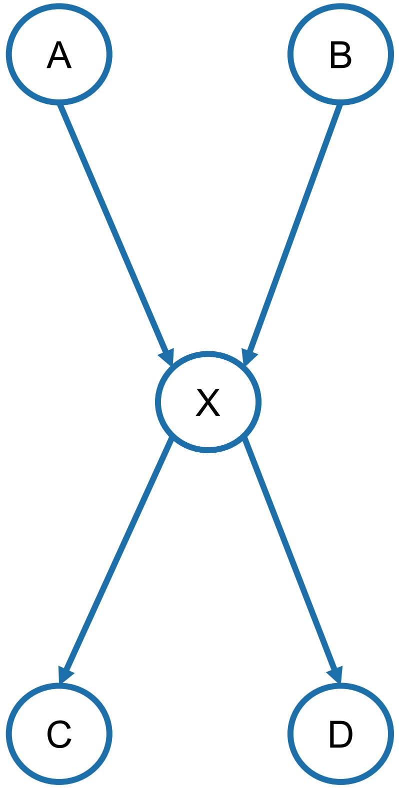
Classification using Bayesian Network

Masayu Leylia Khodra (masayu@informatika.org)

KK IF – Teknik Informatika – STEI ITB

Inteligensi Artifisial
(*Artificial Intelligence*)

- X represents the fish : x1=salmon and x2=sea bass.
- X is influenced by A and B.
- A represents time of year: a1 = winter, a2 = spring, a3 = summer and a4 = autumn. Probability distribution on A in uniform.
- B represents geographical area where the fish was caught: b1 = north Atlantic and b2 = south Atlantic. The probabilities that any fish came from those areas are 0.6 and 0.4.
- C represents lightness with c1 = light, c2 = medium and c3 = dark
- D represents thickness with d1 = wide and d2 = thin.



The probability that the fish was caught in the summer in the north Atlantic and is a sea bass that is dark and thin.

The probability that the fish was caught in the summer (a3) in the north Atlantic (b1) and is a sea bass (x2) that is dark (c3) and thin (d2).

P(a3,b1,x2,c3,d2)

Inference: Example

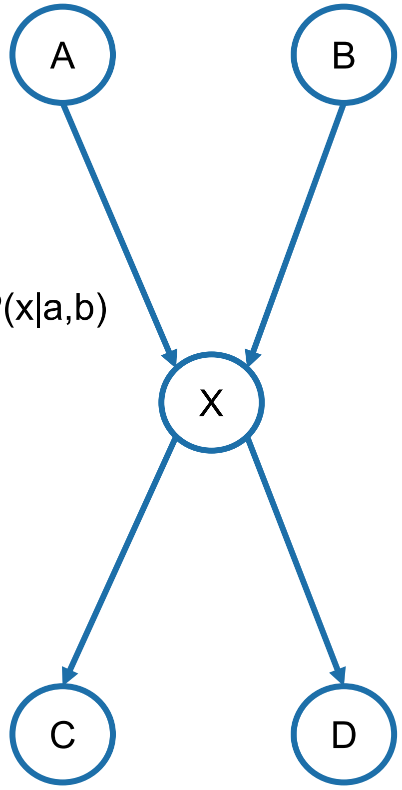
$$P(a3,b1,x2,c3,d2) = P(a3)P(b1)P(x2|a3,b1)P(c3|x2)P(d2|x2) = 0.012$$

P(a)

P(a1)	P(a2)	P(a3)	P(a4)
0.25	0.25	0.25	0.25

P(b)

P(b1)
0.6



P(x|a,b)

	P(x1|a,b)
a1,b1	0.3
a1,b2	0.7
a2,b1	0.6
a2,b2	0.8
a3,b1	0.4
a3,b2	0.1
a4,b1	0.2
a4,b2	0.3

P(c|x)

	P(c1|x)	P(c2|x)	P(c3|x)
x1	0.6	0.2	0.2
x2	0.2	0.3	0.5

P(d|x)

	P(d1|x)
x1	0.3
x2	0.6

Classification

Classify the fish that is light (c1) and caught in the south Atlantic (b2), but we do not know what time of year the fish was caught nor its thickness.

Maximum a posterior probability:
P(x1|c1,b2) vs P(x2|c1,b2)

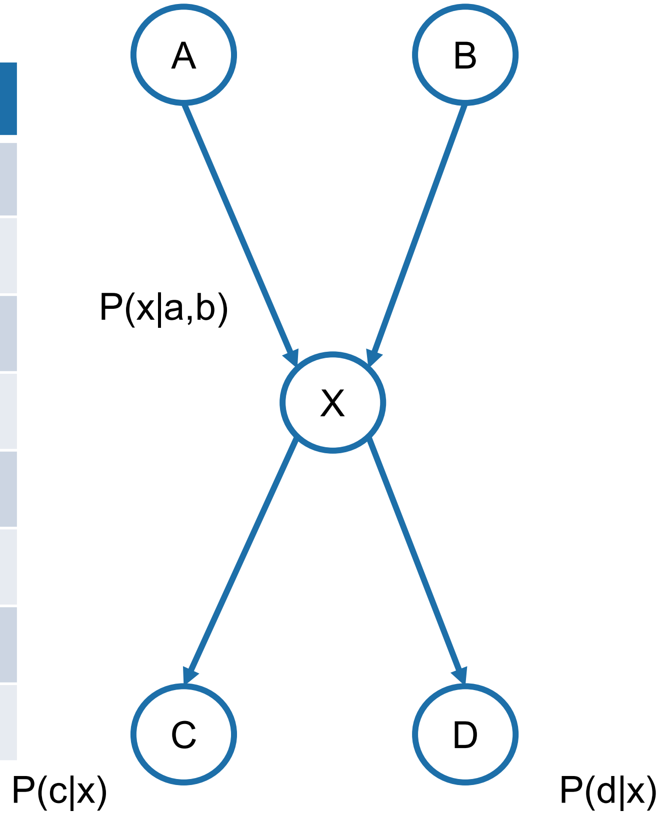
P(a)

P(a1)	P(a2)	P(a3)	P(a4)
0.25	0.25	0.25	0.25

P(b)

P(b1)
0.6

	P(x1|a,b)
a1,b1	0.3
a1,b2	0.7
a2,b1	0.6
a2,b2	0.8
a3,b1	0.4
a3,b2	0.1
a4,b1	0.2
a4,b2	0.3



	P(c1|x)	P(c2|x)	P(c3|x)
x1	0.6	0.2	0.2
x2	0.2	0.3	0.5

	P(d1|x)
x1	0.3
x2	0.6

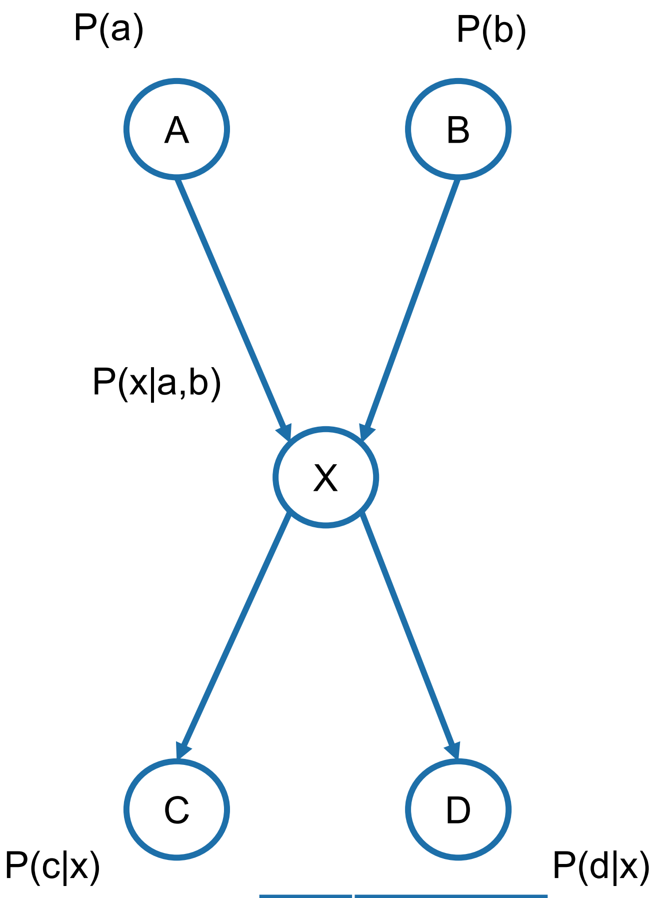
Classification (2)

Q: query
e: evidence of all variables
P(Q|**e**)=P(q,e)/P(e)=αP(Q,e)

P(x1|c1,b2)=P(x1,c1,b2)/P(c1,b2)
=α ∑ P(x1,**a**,b2,c1,**d**)
=α∑P(**a**).P(b2).P(x1|**a**,b2). P(c1|x1). P(**d**|x1)
=αP(b2).P(c1|x1) ∑P(**a**). P(x1|**a**,b2). P(**d**|x1)
=αP(b2).P(c1|x1) [∑P(**a**). P(x1|**a**,b2)][∑P(**d**|x1)]= α 0.114

P(a1). P(x1|a1,b2)+
P(a2). P(x1|a2,b2)+
P(a3). P(x1|a3,b2)+
P(a4). P(x1|a4,b2)

P(d1|x1)+P(d2|x1)=1.0



	P(d1|x)
x1	0.3
x2	0.6

Classification (3)

$P(x1|c1,b2)=P(x1,c1,b2)/P(c1,b2)$
$=\alpha P(b2).P(c1|x1).[\sum P(a).P(x1|a,b2)]. [\sum P(d|x1)]= \alpha\ 0.114$

$P(x2|c1,b2)=P(x2,c1,b2)/P(c1,b2)$
$=\alpha P(b2).P(c1|x2)\ [\sum P(a).P(x2|a,b2)][\sum P(d|x2)]= \alpha\ 0.042$

Normalize:
$P(x1|c1,b2)=0.73$
$P(x2|c1,b2)=0.27$
Decision: x1=salmon

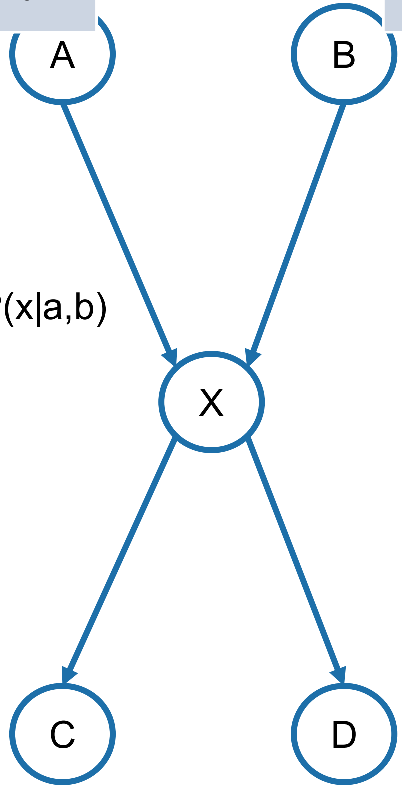
P(a)

P(a1)	P(a2)	P(a3)	P(a4)
0.25	0.25	0.25	0.25

P(b)

P(b1)
0.6

	P(x1|a,b)
a1,b1	0.3
a1,b2	0.7
a2,b1	0.6
a2,b2	0.8
a3,b1	0.4
a3,b2	0.1
a4,b1	0.2
a4,b2	0.3



P(c|x)

	P(c1|x)	P(c2|x)	P(c3|x)
x1	0.6	0.2	0.2
x2	0.2	0.3	0.5

P(d|x)

	P(d1|x)
x1	0.3
x2	0.6

Knowledge Representation and Reasoning

Learning from Data

Masayu Leylia Khodra (masayu@informatika.org)

KK IF – Teknik Informatika – STEI ITB

Inteligensi Buatan
(*Artificial Intelligence*)

Problems in BN constructed by human expert

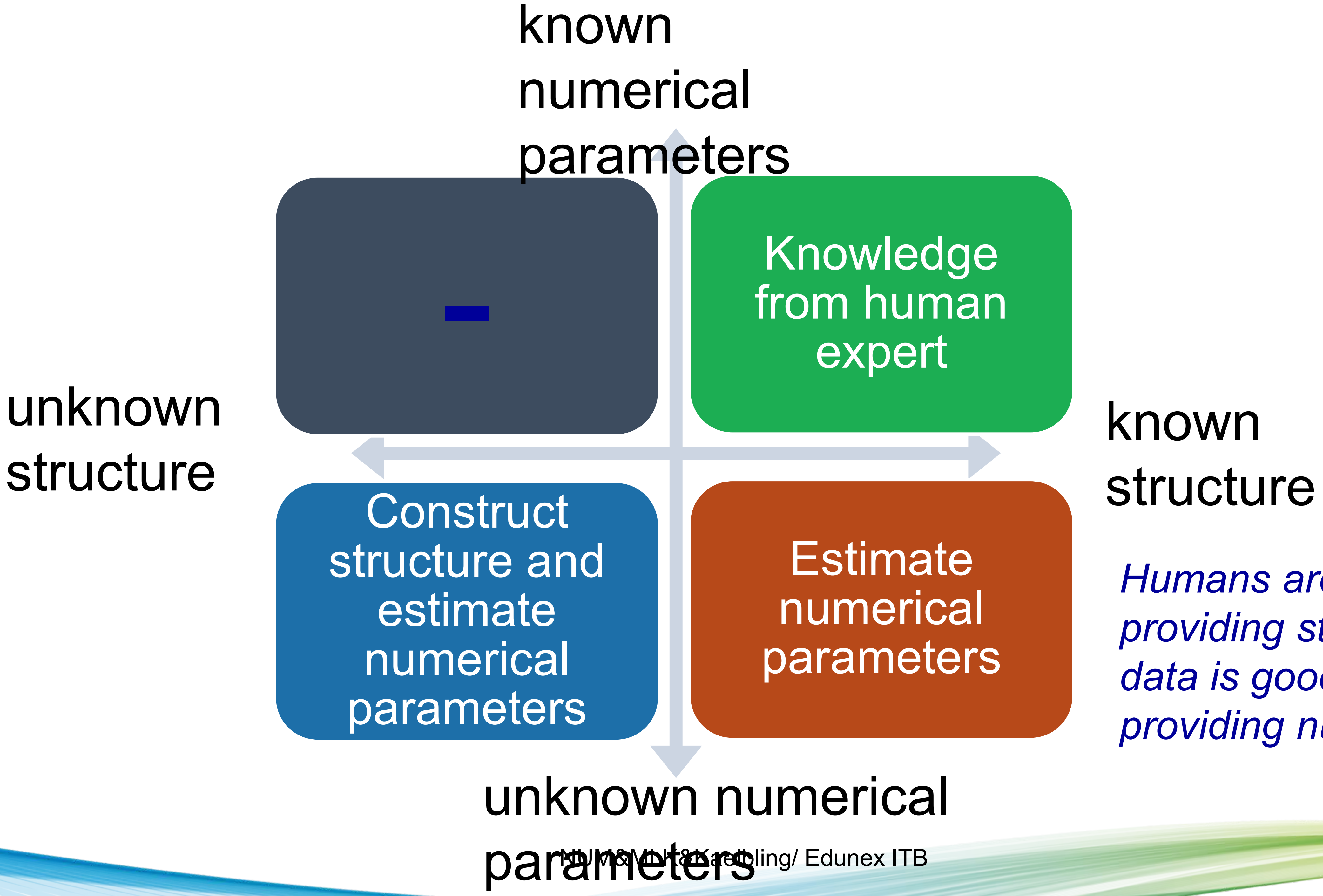
Knowledge acquisition bottleneck problem

Knowledge elicitation problem: slow speed, and inability of expert to express the knowledge they posses.

Very difficult in getting reliable probability estimates.

Ideally, determine probability by data distribution, experience, and assumption (subjective).

What is difference between P(a)=0.5 and P(a)=0.6 ?



Humans are good at providing structure, data is good at providing numbers

Parameter Estimation

Given structure with m nodes

Given a data set

D = $\{<v_1^1,..,v_m^1>,..,<v_1^k,..,v_m^k>\}$

Count $\#(V_i=T)$, $\#(V_i=F)$, $\#(V_i=T, V_j=T)$, $\#(V_i=T, V_j=F)$

Variable V_i with no parent

$$P(V_i) \approx \frac{\#(V_i = T)}{k}$$

Variable V_i with parent V_j

$$P(V_i|V_j) \approx \frac{\#(V_i = T, V_j = T)}{\#(V_j = T)}$$

$$P(V_i|\neg V_j) \approx \frac{\#(V_i = T, V_j = F)}{\#(V_j = F)}$$

Smoothing: Avoid Probability = 0

Variable V_i with no parent

$$P(V_i) \approx \frac{\#(V_i = T) + 1}{k + 2}$$

kemungkinan nilainya True/False
=> ada 2

Variable V_i with parent V_j

$$P(V_i|V_j) \approx \frac{\#(V_i = T, V_j = T) + 1}{\#(V_j = T) + 2}$$

$$P(V_i|\neg V_j) \approx \frac{\#(V_i = T, V_j = F) + 1}{\#(V_j = F) + 2}$$

Construct BN Structure

Determine set of variables

Determine ordering of variables $X_1,..X_m$

Loop: Add X_i, and select set of parents for X_i from $X_1,..,X_{i-1}$

Construct BN Structure: Example

Based on causal knowledge:

- Set of variables: A,B,C,D,X
- Ordering of variables: A,B,C,D,X
- Add A: no parent.
- Add B: is A parent of B ?
- Add C: parent of C ?
- Add D: parent of D ?
- Add X: parent of X ?

Based on data:

- Set of variables: A,B,C,D,X
- Ordering of variables: A,B,C,D,X
- Add A: no parent.
- Add B: $P(B|A)=P(B)$?
- Add C:
  - $P(C|A)=P(C)$?
  - $P(C|B)=P(C)$?
  - $P(C|A,B)=P(C|A)$?
  - $P(C|A,B)=P(C|B)$?
  - $P(C|A,B)=P(C)$?
- Etc.

is A the parent of B??

cara ceknya computer akan hitung $P(B|A) = P(B)$
kalau iya berarti mereka saling independen dan A bukan parent dari B

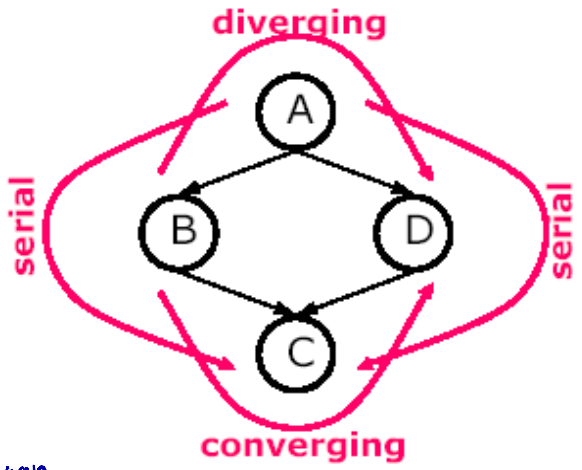
$P(C|A,B) = P(C,B)$
A tidak mempengaruhi C,
A bukan parent C.
$P(C|B) \neq P(C)$.
berarti parent C adalah B.

D Separation

Two variables A and B are d-separated iff for every path between them, there is an intermediate variable V such that either

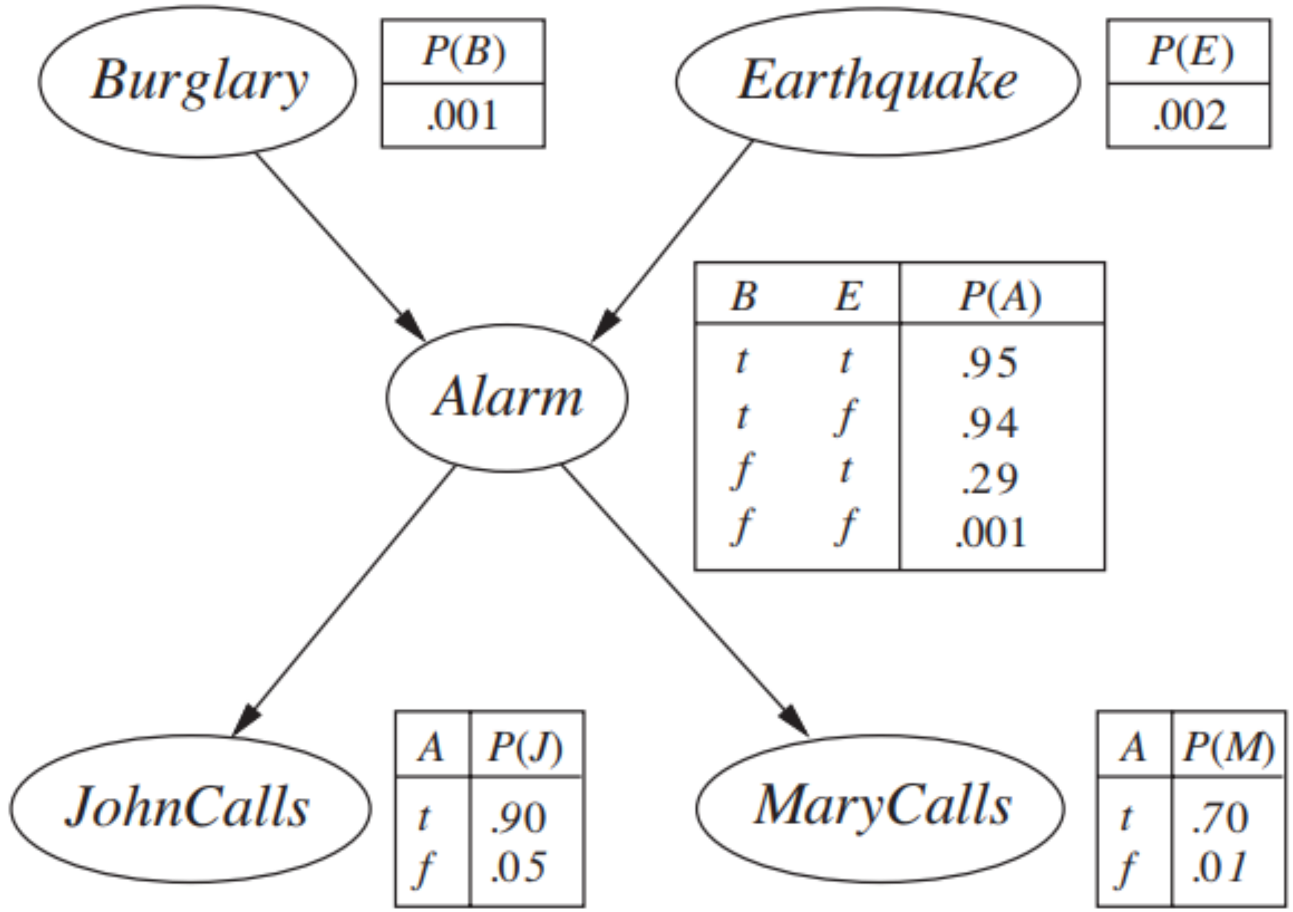
- The connection is serial or diverging and V is known
- The connection is converging and neither V nor any descendant is instantiated

Two variables are d-connected iff they are not d-separated



- A-B-C: serial, blocked when B is known, connected otherwise
- A-D-C: serial, blocked when D is known, connected otherwise
- B-A-D: diverging, blocked when A is known, connected otherwise
- B-C-D: converging, blocked when C has no evidence, connected otherwise

dalam BN,
boleh ubah² panah asal
ga bikin converging
baru + ga menghilangkan
converging lama



P(B)
.001

P(E)
.002

B	E	P(A)
t	t	.95
t	f	.94
f	t	.29
f	f	.001

A	P(J)
t	.90
f	.05

A	P(M)
t	.70
f	.01

Figure 14.2 A typical Bayesian network, showing both the topology and the conditional probability tables (CPTs). In the CPTs, the letters B, E, A, J, and M stand for *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, and *MaryCalls*, respectively.

Variable Ordering Influences Structure

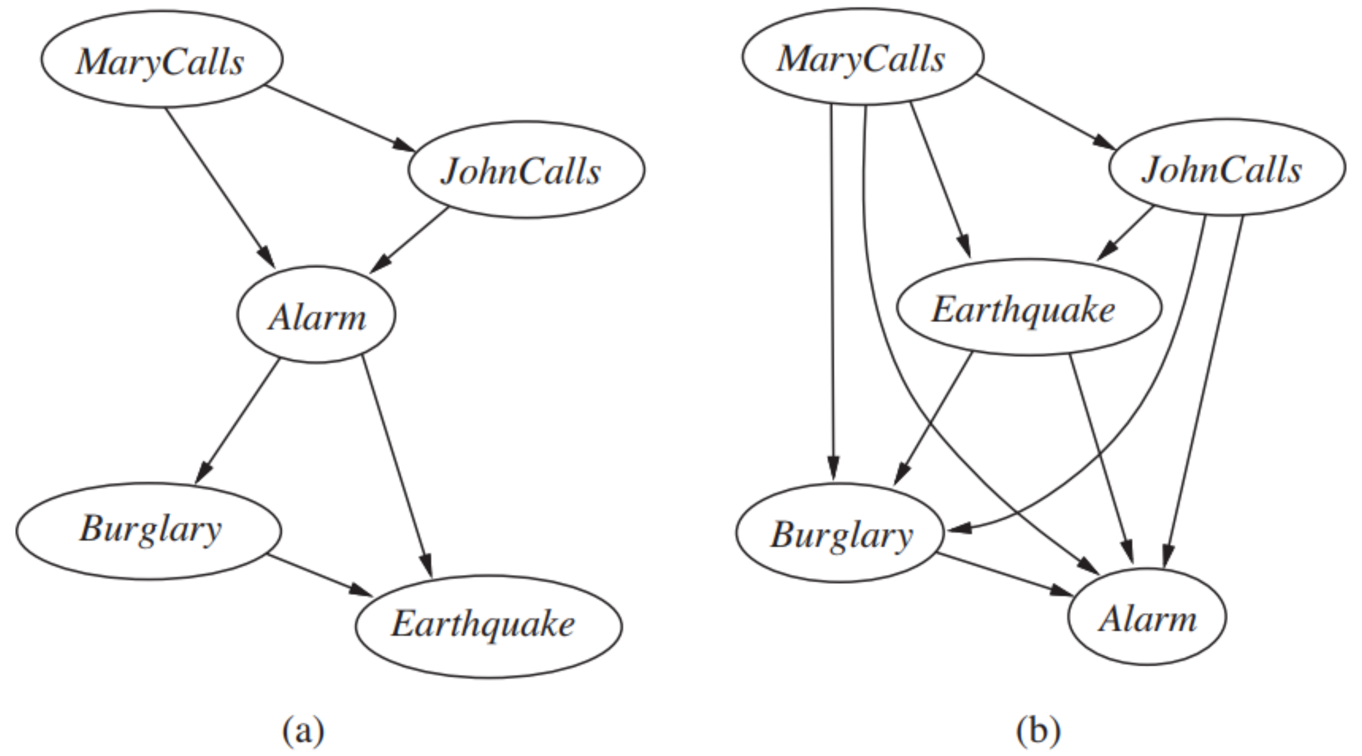


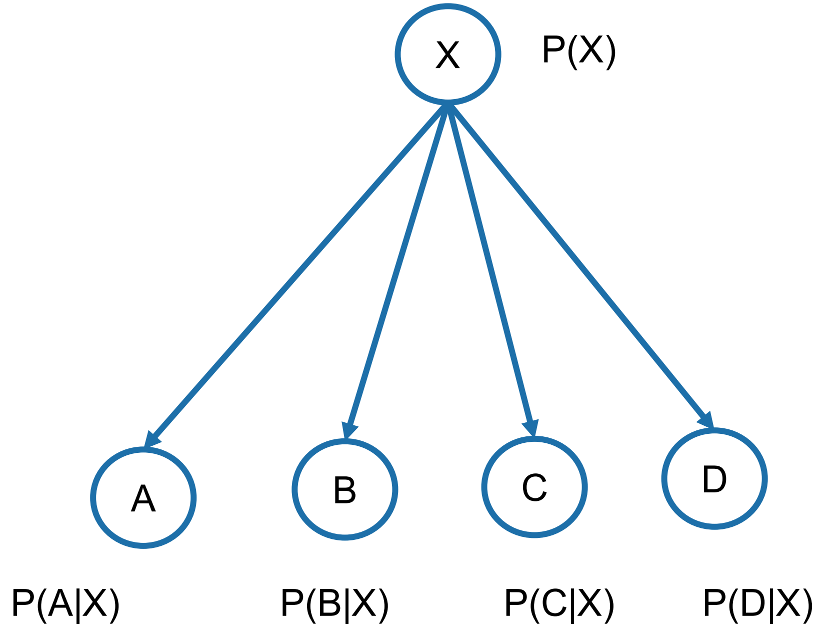
Figure 14.3 Network structure depends on order of introduction. In each network, we have introduced nodes in top-to-bottom order.

Russel & Norvig (2010). AI.

Construct BN Structure: Example 2

Based on causal knowledge:

- Set of variables: X,A,B,C,D
- Ordering of variables: X,A,B,C,D
- Add X: no parent.
- Add A: parent(A)=X.
- Add B: parent(B)=X. P(B|A,X)=P(B|X)
- Add C: parent(C)=X. P(C|A,B,X)=P(C|X)
- Add D: parent(D)=X. P(D|A,B,C,X)=P(D|X)

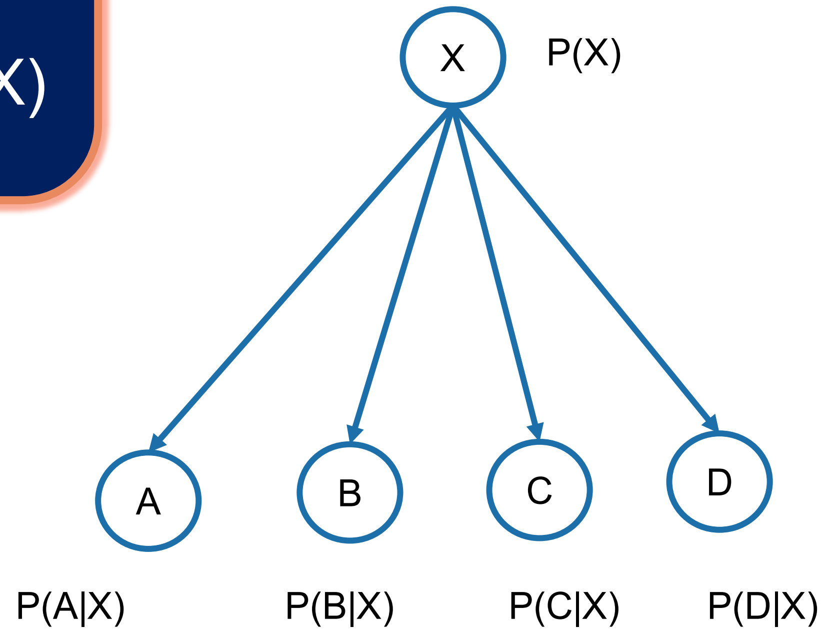


Naive Bayes as “Special” Bayesian Network

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

Probability Model of P(X) and $P(a_i|X)$
P(X, A, B, C, D) = P(X). P(A|X). P(B|X). P(C|X). P(D|X)

Classification:
Find the maximum P(X | A,B,C,D)



Example: Play Tennis Dataset

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes

outlook	temp.	humidity	windy	play
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Frequency of (sunny|yes) → 2

Frequency of (overcast|yes) → 4

Frequency of (hot|yes) → 2

…

Frequency of class 'yes'→ 9

Frequency of (sunny|no) → 3

Frequency of (overcast|no) → 0

Frequency of (hot|no) → 2

Frequency of class 'no' → 5

Example: Play Tennis Dataset

outlook	temp.	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes

outlook	temp.	humidity	windy	play
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Frequency o
(sunny|yes) →

Frequency o
(sunny|no) →

outlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								

Frequency of
class 'yes'→ 9

Frequency of
class 'no' → 5

Example: Play Tennis Dataset

Frequency

outlook			temperature			humidity			windy			play	
	yes	no		yes	no		yes	no		yes	no	yes	no
sunny	2	3	hot	2	2	high	3	4	false	6	2	9	5
overcast	4	0	mild	4	2	normal	6	1	true	3	3		
rainy	3	2	cool	3	1								

$P(a_i|v_j)$

outlook			temperature			humidity			windy		
	yes	no		yes	no		yes	no		yes	no
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5
rainy	3/9	2/5	cool	3/9	1/5						

$P(v_j)$

play	
yes	no
9/14	5/14

Probability Model

Classify New Instance: <Sunny, Cool, High, True>

$P(a_i|v_j)$

	outlook			temperature			humidity			windy	
	yes	no		yes	no		yes	no		yes	no
sunny	2/9	3/5	hot	2/9	2/5	high	3/9	4/5	false	6/9	2/5
overcast	4/9	0/5	mild	4/9	2/5	normal	6/9	1/5	true	3/9	3/5
rainy	3/9	2/5	cool	3/9	1/5						

$P(v_j)$

play	
yes	no
9/14	5/14

$$P(v_j \mid a_1, a_2, \ldots, a_n) = P(v_j) \, . \, \prod_i P(a_i \mid v_j)$$

P(yes|sunny, cool, high, true)

= P(yes). P(sunny|yes).P(cool|yes).P(high|yes).P(true|yes)

= 9/14 . 2/9 . 3/9 . 3/9 . 3/9 = 0.0053

P(no |sunny, cool, high, true)

= P(no). P(sunny|no).P(cool| no).P(high| no).P(true| no)

= 5/14 . 3/5 . 1/5 . 4/5 . 3/5 = $\boxed{0.0206}$

THANK YOU