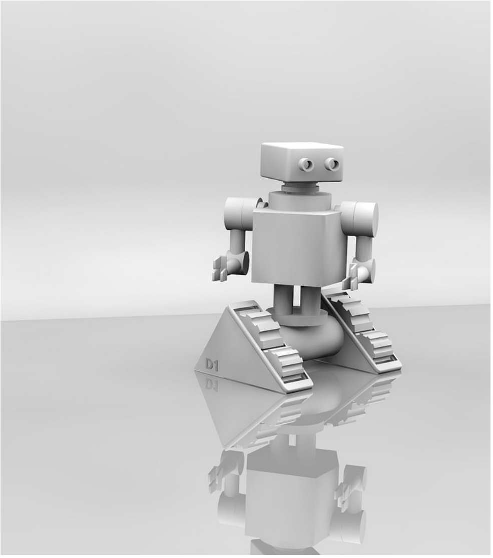


CPSC 481 Artificial Intelligence

Dr. Mira Kim

Mira.kim@fullerton.edu

What we will cover this week

- Bayesian Networks

Naïve Bayes Classifier

- What if we have multiple tests to decide is a person has cancer?

- Simplified assumption: attributes are conditionally independent

Naïve Bayes advantages

- Independence allows parameters to be estimated on different data sets, e.g.
  - Estimate content features from messages with headers omitted
  - Estimate header features from messages with content missing

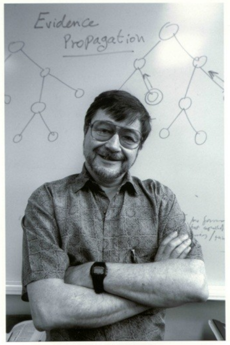
Naïve Bayes disadvantages

- What if we have many more variables?
  - Want to execute different types of queries
  - Cannot assume independence between all variables

Bayesian Network Motivation

- We want a representation and reasoning system that is based on conditional independence
  - Compact yet expressive representation
  - Efficient reasoning procedures
- Bayesian Networks are such a representation
  - Named after Thomas Bayes (ca. 1702 –1761)
  - Term coined in 1985 by Judea Pearl (1936 –)
    - Turing Award winner, 2011
  - Invention changed the focus on AI from logic to probability!

Thomas Bayes



Judea Pearl

What are Bayesian networks?

- Bayesian networks (BN) are a graph-based framework for representing and analyzing models involving uncertainty

- BN are different from other knowledge-based systems tools because uncertainty is handled in a mathematically rigorous yet efficient and simple way

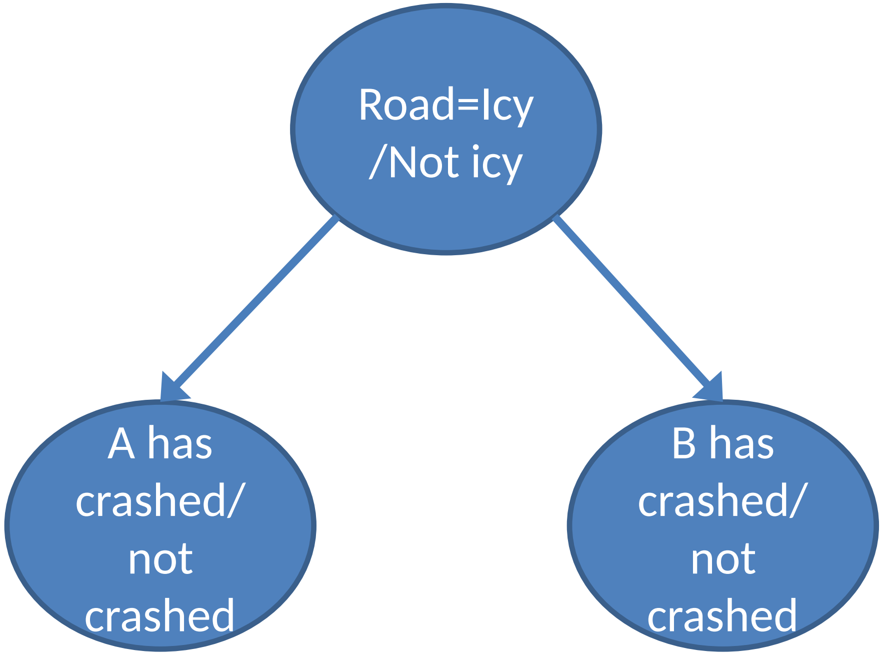
- BN are different from other probabilistic analysis tools because of graphical network representation of problems and use of Bayesian inference

Concept of a Bayesian Network

- Graph-based knowledge structure:
  - Nodes: variables
  - Edges: represent probabilistic dependence between variables
  - conditional probabilities encode the strength of the dependencies
- Computational architecture:
  - computes posterior probabilities given evidence about some nodes
  - exploits probabilistic independence for efficient computation

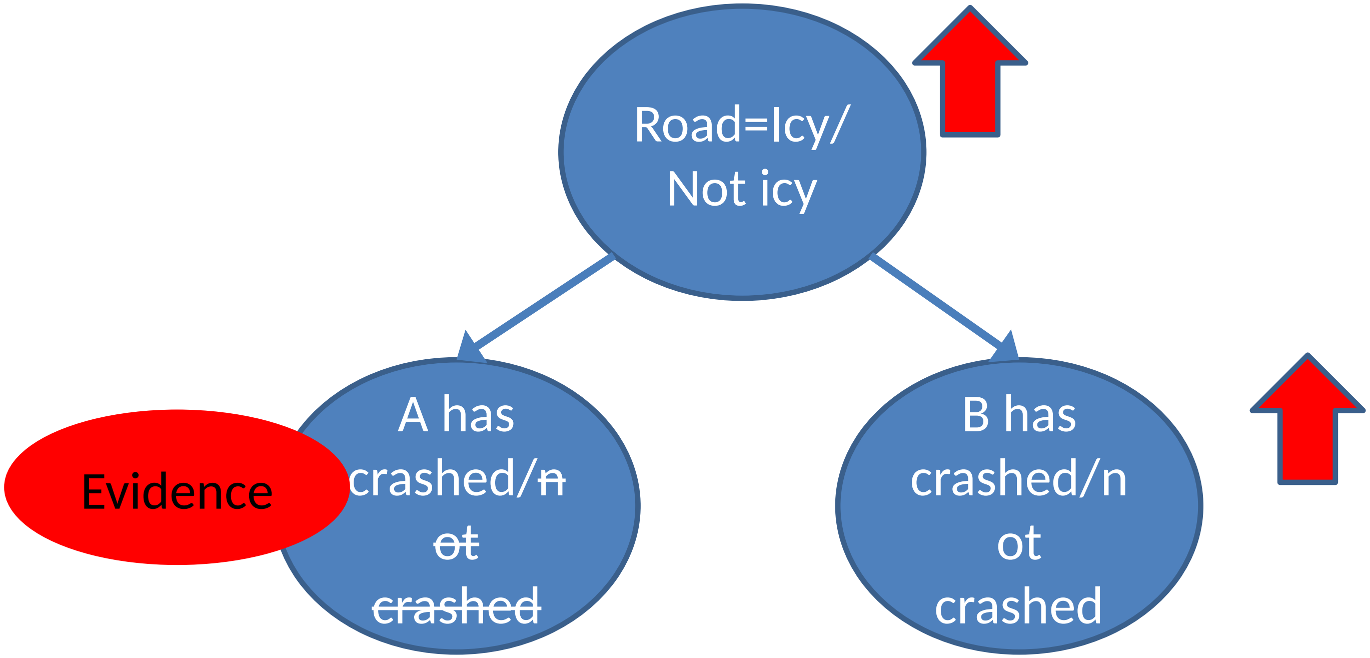
“Icy roads” example

- A road can be icy or not
- There is a chance that driving on that road will result in an accident
  - Accident is much more likely if road is icy
- Consider two people, A and B, driving on that road


Road=Icy
/Not icy
A has
crashed/
not
crashed
B has
crashed/
not
crashed

“Icy roads” example

- Initially, don’t know if road is icy, or if A/B had an accident
- You learn that A had an accident
  - What can we say about the condition of the road and B’s chances of an accident?
  - Our thinking:
    - “If A has crashed, then it probably means the roads are icy, and so it is more likely that B will have an accident”
  - The probability that B will have an accident **has increased given this information**



Road=Icy/ Not icy

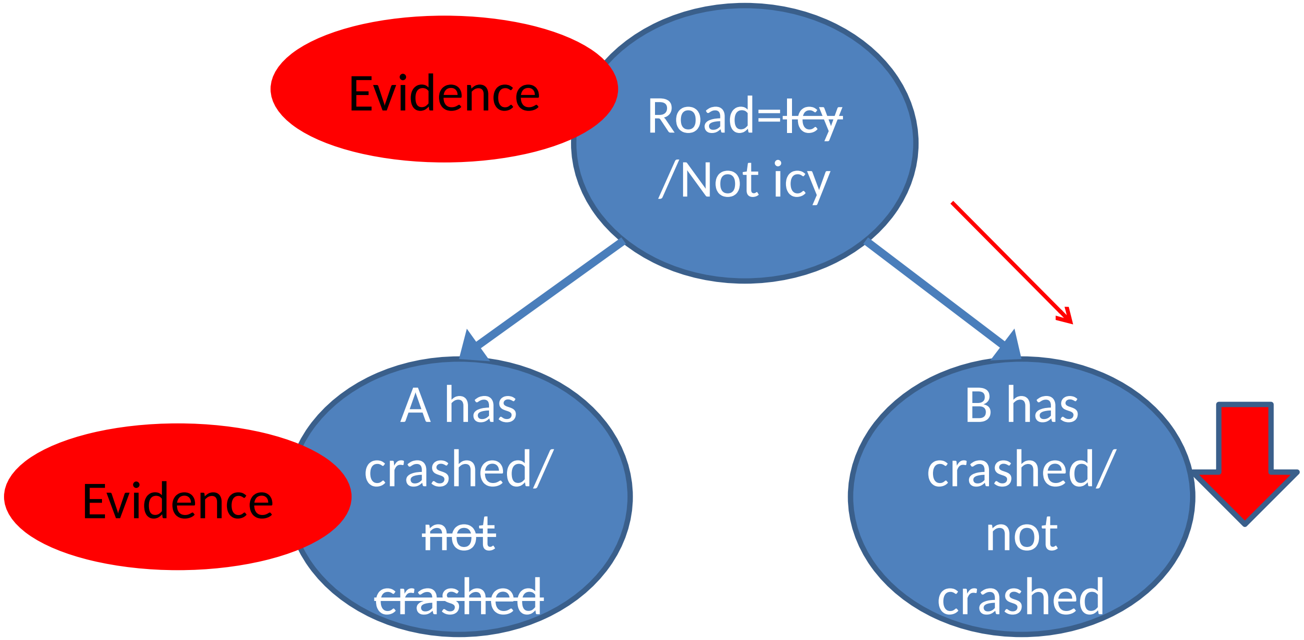
Evidence

A has crashed/~~not crashed~~

B has crashed/not crashed

“Icy roads” example

- You now find out that roads are not icy
  - Our thinking:
    - “A was just unlucky; B will probably not have an accident”
  - The probability that B will have an accident **has decreased given this information**



Evidence

Road=~~Icy~~/Not icy

Evidence

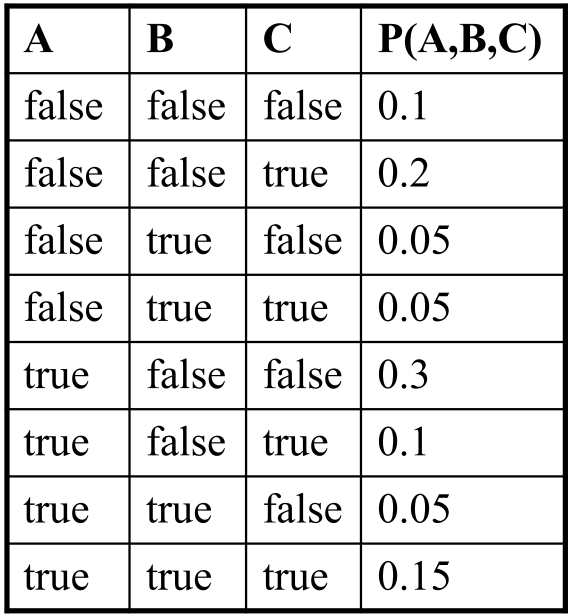
A has crashed/ ~~not crashed~~

B has crashed/ not crashed

The Joint Probability Distribution

- Joint probabilities can be between any number of variables

  eg. *P(A = true, B = true, C = true)*
- For each combination of variables, we need to say how probable that combination is
- The probabilities of these combinations need to sum to 1



A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Sums to 1

The Joint Probability Distribution

- Once you have the joint probability distribution, you can calculate **any** probability involving *A*, *B*, and *C*
- Note: use marginalization and definition of conditional probability

A	**B**	**C**	**P(A,B,C)**
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Examples of things you can compute:

- *P*(*A*=true) = sum of *P*(*A,B,C*) in rows with *A*=true
- *P*(*A*=true, *B* = true *C*=true)

Marginalization

- To get probability ***without*** a variable *A*:
  - Sum up all rows containing different values of *A*, without changing other variable values

Marginalization

Road is icy?	A has accident?	B has accident?	P(Road is icy?, A has accident?, B has accident?)
False	False	False	0.648
False	False	True	0.072
False	True	False	0.072
False	True	True	0.008
True	False	False	0.098
True	False	True	0.042
True	True	False	0.042
True	True	True	0.018

In-class Exercise:
Compute

- (Prior) probability that A has accident
- Probability that A has accident if it is known that the road is icy
- Probability that A has accident if it is known that B had an accident (and it is not known if the road is icy or not)
- Probability that A has accident if it is known that B had an accident and Road is not icy

Classwork: Compute

- (Prior) probability that A has accident
  - P(A=true)

  = 0.072+0.08+0.042+0.018 = 0.14
- Probability that A has accident if it is known that the road is icy
  - P(A=true|Icy=true) = P(A=true,Icy=true)/P(Icy=true)

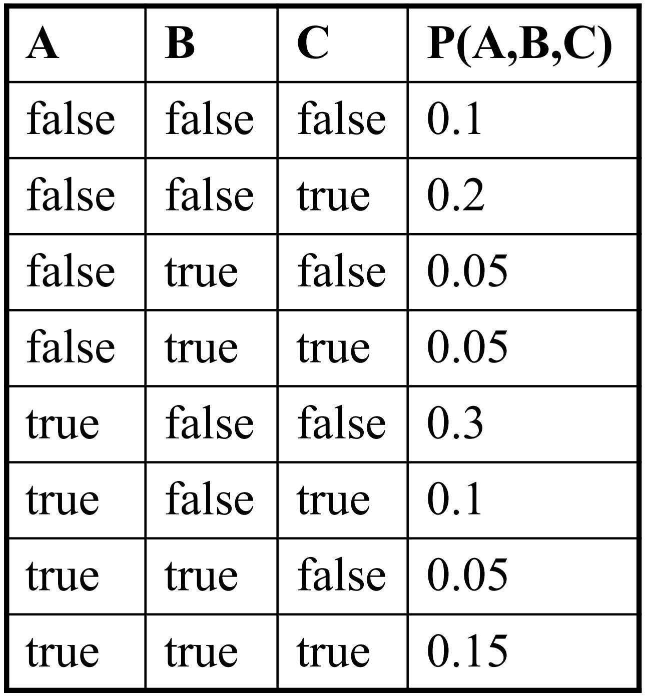
    = (0.042+0.018)/(.098 + .042 + .042 + .018) = 0.06/0.2 = 0.3
- Probability that A has accident if it is known that B had an accident (and it is not known if the road is icy or not)
  - P(A=true|B=true) = P(A=true,B=true)/P(B=true)

    = (0.008+0.018)/(0.072+0.008+0.042+0.018) = 0.1857
- Probability that A has accident if it is known that B had an accident and Road is not icy
  - P(A=true|B=true, Icy=false) = P(A=true,B=true, Icy=false) /P(B=true, Icy=false)

    =0.008/(0.072+0.008) = 0.1

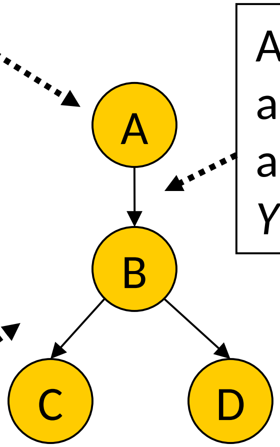
The Problem with the Joint Distribution

- Lots of entries in the table
- For *k* Boolean random variables, table of size 2^k
- How do we use fewer numbers?
- Exploit the concept of **conditional independence**



A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Bayesian network: A Directed Acyclic Graph



Each node in the graph is a random variable

A node *X* is a parent of another node *Y* if there is an arrow from node *X* to node *Y* eg. *A* is a parent of *B*

Informally, an arrow from node *X* to node *Y* means *X* has a direct influence on *Y*

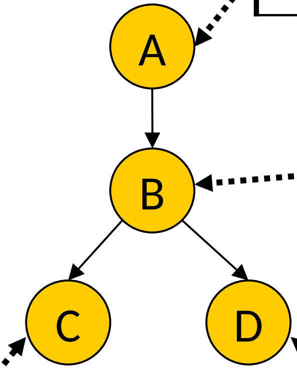
A Bayesian Network

1. A Directed Acyclic Graph
   - A node represents a variable
2. A conditional probability table for each node

Each node X_i has a conditional probability distribution P(X_i | Parents(X_i)) that quantifies the effect of the parents on the node

P(+a)	~~P(-a)~~
0.4	~~0.6~~

A	P(A)
false	0.6
true	0.4



A	P(+b|A)	~~P(-b|A)~~
false	0.99	~~0.01~~
true	0.3	~~0.7~~

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

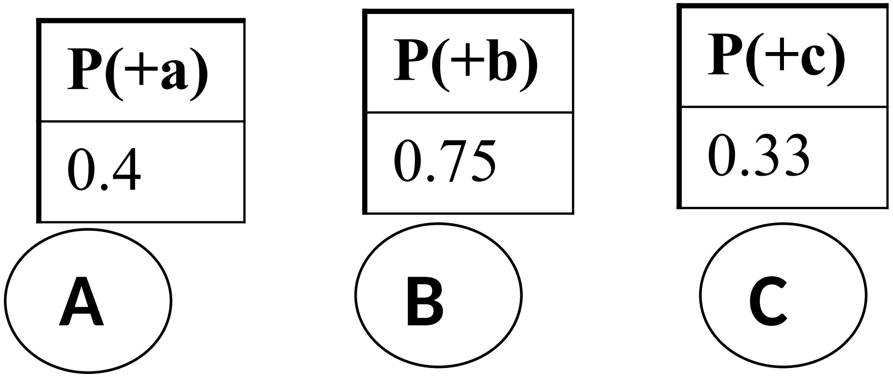
B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

B	P(+c|B)
false	0.6
true	0.1

B	P(+d|B)
false	0.98
true	0.95

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

Examples of 3-way Bayesian Networks



P(+a)
0.4

P(+b)
0.75

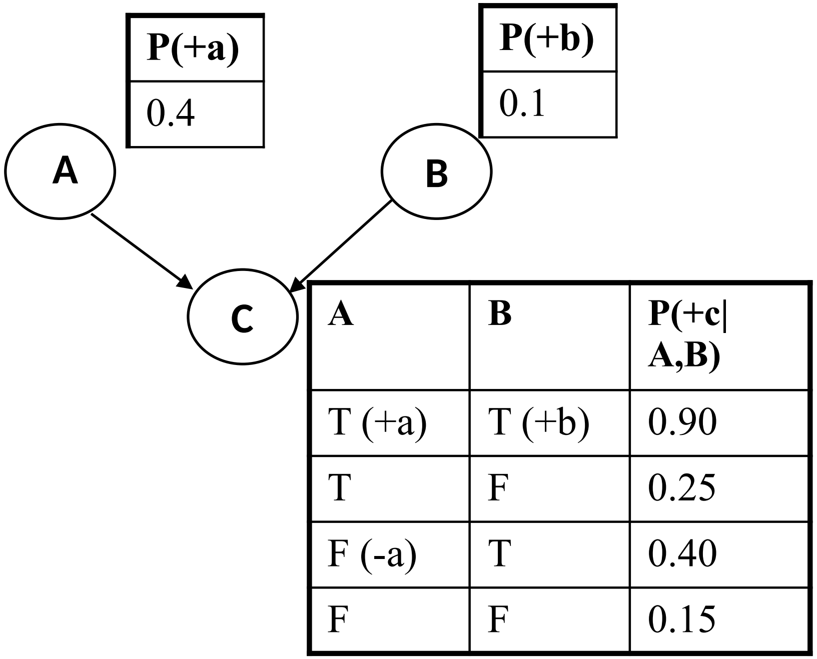
P(+c)
0.33

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

- What do the 3 tables look like?
  - Assume all variables are binary

Examples of 3-way Bayesian Networks

- Independent Causes
  - e.g., A is “exposure to toxins”, B is “exposure to radiation”
  - C is “cancer”



P(+a)
0.4

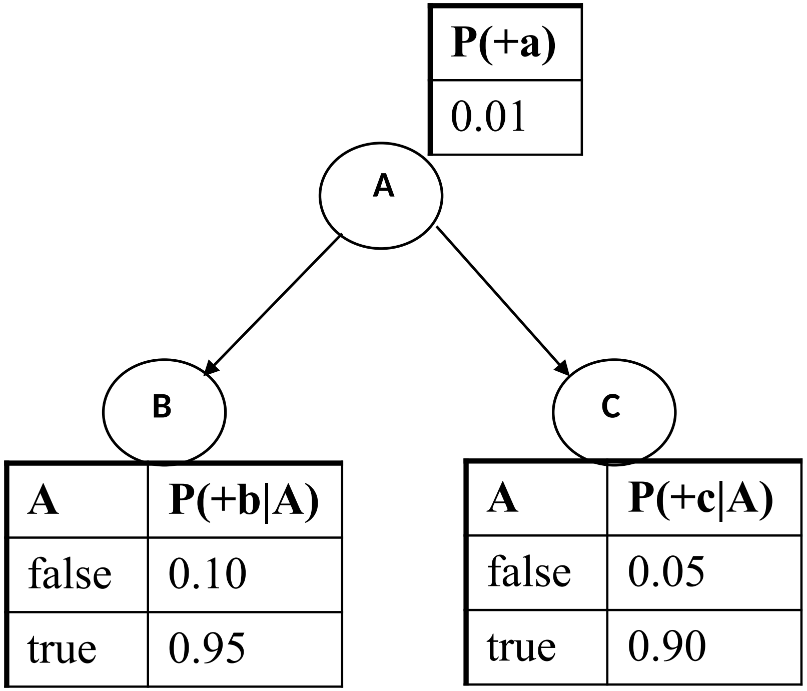
P(+b)
0.1

A	B	P(+c| A,B)
T (+a)	T (+b)	0.90
T	F	0.25
F (-a)	T	0.40
F	F	0.15

- What do the 3 tables look like?
  - Assume all variables are binary

Examples of 3-way Bayesian Networks

- Conditionally independent effects
- B and C are *not* independent
- But B and C **become conditionally independent** given A
  - e.g., A is a disease, and
  - B and C are conditionally independent symptoms given A



P(+a)
0.01

A	P(+b|A)
false	0.10
true	0.95

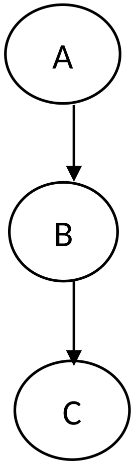
A	P(+c|A)
false	0.05
true	0.90

- What do the 3 tables look like?
  - Assume all variables are binary

Examples of 3-way Bayesian Networks

- Example: *causal chain*
- E.g.,
  - A: Low pressure
  - B: Rain
  - C: Traffic

P(+a)
0.01



A	**P(+b|A)**
false	0.10
true	0.95

B	**P(+c|B)**
false	0.05
true	0.90

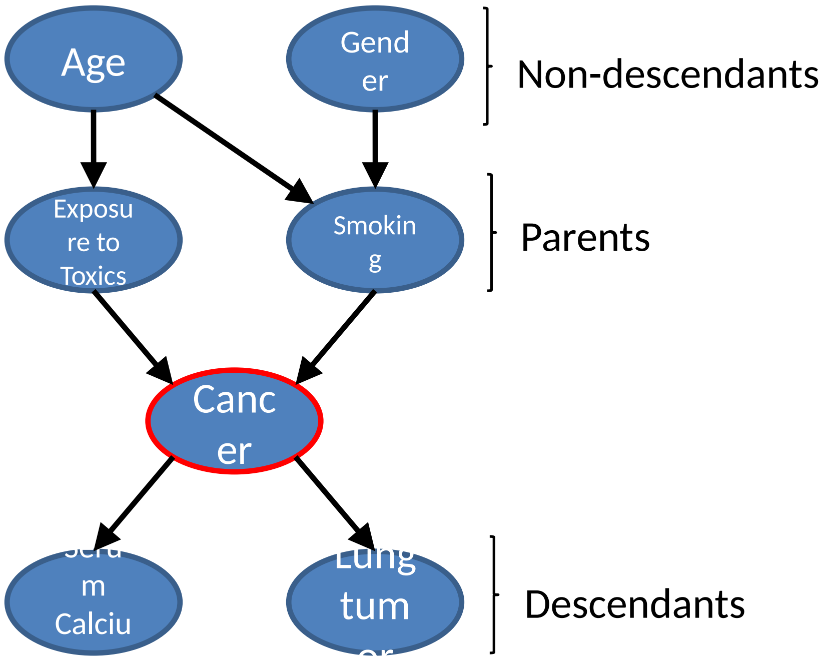
- What do the 3 tables look like?
  - Assume all variables are binary

The Chain Rule

- The product rule

- More generally, can always write **any** joint distribution as an incremental product of conditional distributions

Conditional independence



Non-descendants

Parents

Descendants

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

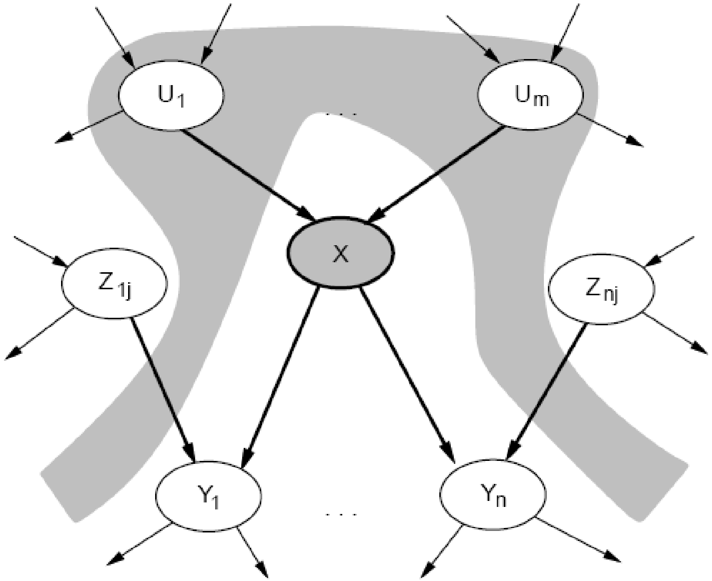
Cancer is independent of *Age* and *Gender* given *Exposure to Toxics* and *Smoking*

P is conditionally independent of Q if

(1) Q is a non-descendant of P
(2) given *all* parents(P)

Conditional Independence

A node (*X*) is conditionally independent of its non-descendants (Z_{1j}, Z_{nj}), given its parents (U_1, U_m).



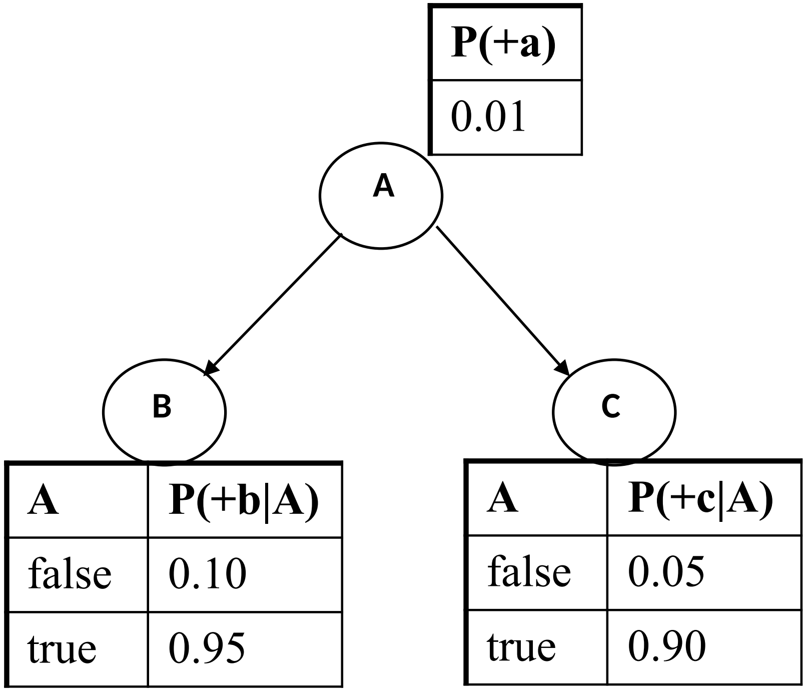
The Joint Probability Distribution

Assuming conditional independence due to parents,
the joint probability distribution over all the variables X_1, …, X_n in the Bayesian network is:

Compare to the general case:

Examples of 3-way Bayesian Networks

- Conditionally independent effects
- B and C are *not* independent
- But B and C **become conditionally independent** given A
  - – e.g., A is a disease, and
  - – B and C are conditionally independent symptoms given A



P(+a)
0.01

A	P(+b|A)
false	0.10
true	0.95

A	P(+c|A)
false	0.05
true	0.90

- What do the 3 tables look like?
  - Assume all variables are binary

Using a Bayesian Network Example

Using the network in the example (Slide 21), suppose you want to calculate:

P(A = true, B = true, C = true, D = true)

= P(A = true) * P(B = true | A = true) *

P(C = true | B = true) P(D = true | B = true)

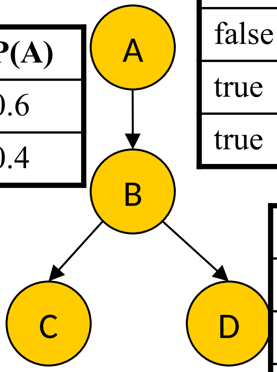
= (0.4)*(0.3)*(0.1)*(0.95)

This is from the graph structure

These numbers are from the conditional probability tables

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3



B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

Classwork: Using a Bayesian Network

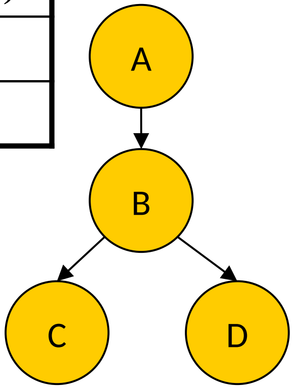
Using the network shown below (also see Slide 21), calculate the following:

P(A = true, B = false, C = false, D = true)

P(A = false, B = true, C = false, D = false)

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3



B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

Classwork: Using a Bayesian Network

Using the network shown below (also see Slide 21), calculate the following:

P(A = true, B = false, C = false, D = true)

P(A = True) * P (B = False|A = True) * P(C=False|B=False)*P(D=True|B=False)

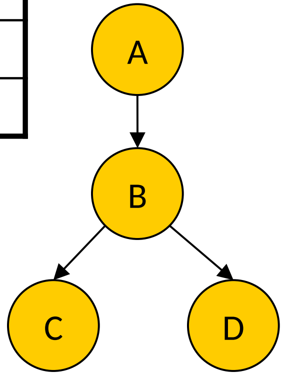
=.4 * .7 * .4 * .98 = .10976

P(A = false, B = true, C = false, D = false)

P(A = False) * P (B = True|A = False) * P(C=False|B=True)*P(D=False|B=True)

=.6 * .99 * .9 * .05 = .02673

A	P(A)
false	0.6
true	0.4



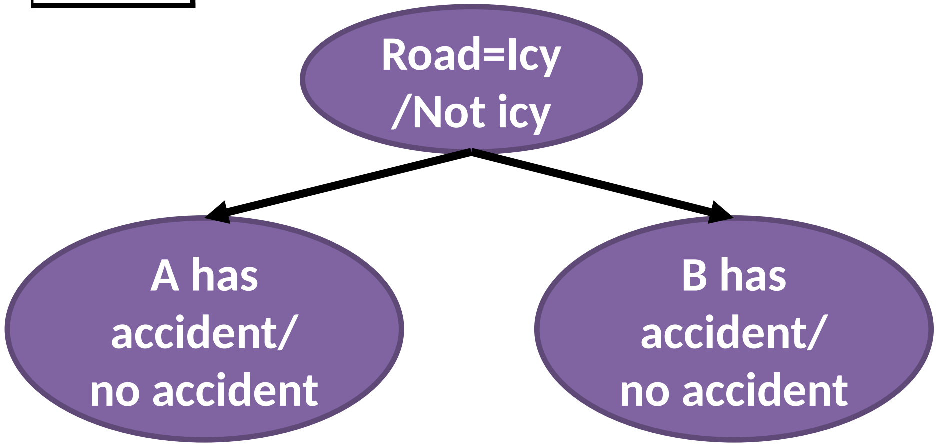
A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

What is the joint distribution represented by this BN?

P(+icy)
0.2



ICY	P(+a|ICY)
false	0.1
true	0.3

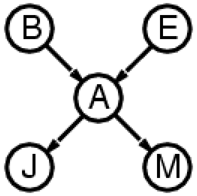
ICY	P(+b|ICY)
false	0.1
true	0.3

What is the joint distribution represented by this BN?

Road is icy?	A has accident?	B has accident?	P(Road is icy?, A has accident?, B has accident?)
False	False	False	0.648
False	False	True	0.072
False	True	False	0.072
False	True	True	0.008
True	False	False	0.098
True	False	True	0.042
True	True	False	0.042
True	True	True	0.018

Compactness

- Suppose we have a Boolean variable X_i with k Boolean parents. How many rows does its conditional probability table have?
  - – 2^k rows for all the combinations of parent values
  - – Each row requires one number p for X_i = true
- If each variable has no more than k parents, how many numbers does the complete network require?
  - – $O(n \cdot 2^k)$ numbers – vs. $O(2^n)$ for the full joint distribution



Inference

- Using a Bayesian network to compute probabilities is called inference
- In general, inference involves queries of the form:

  $P(X \mid E)$

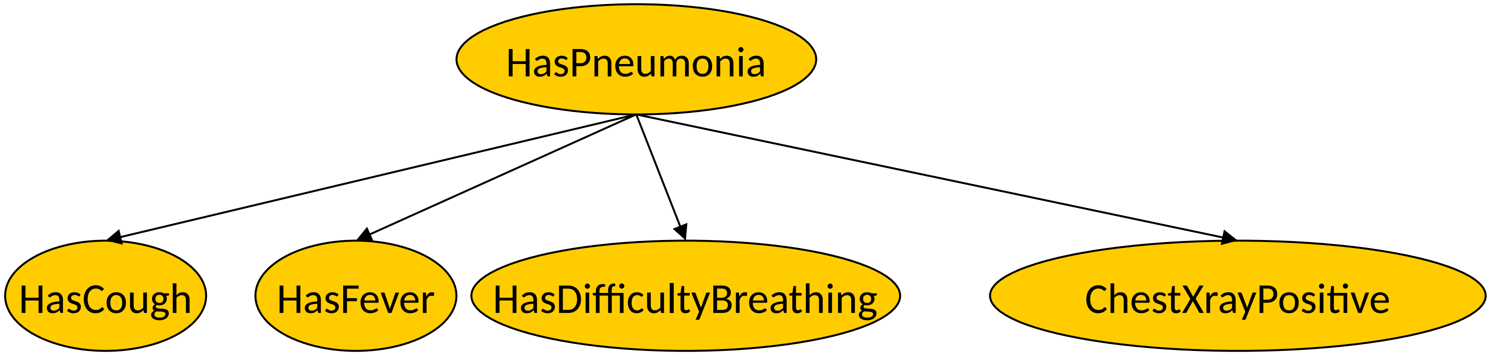
E = The evidence variable(s)

X = The query variable(s)

Inference

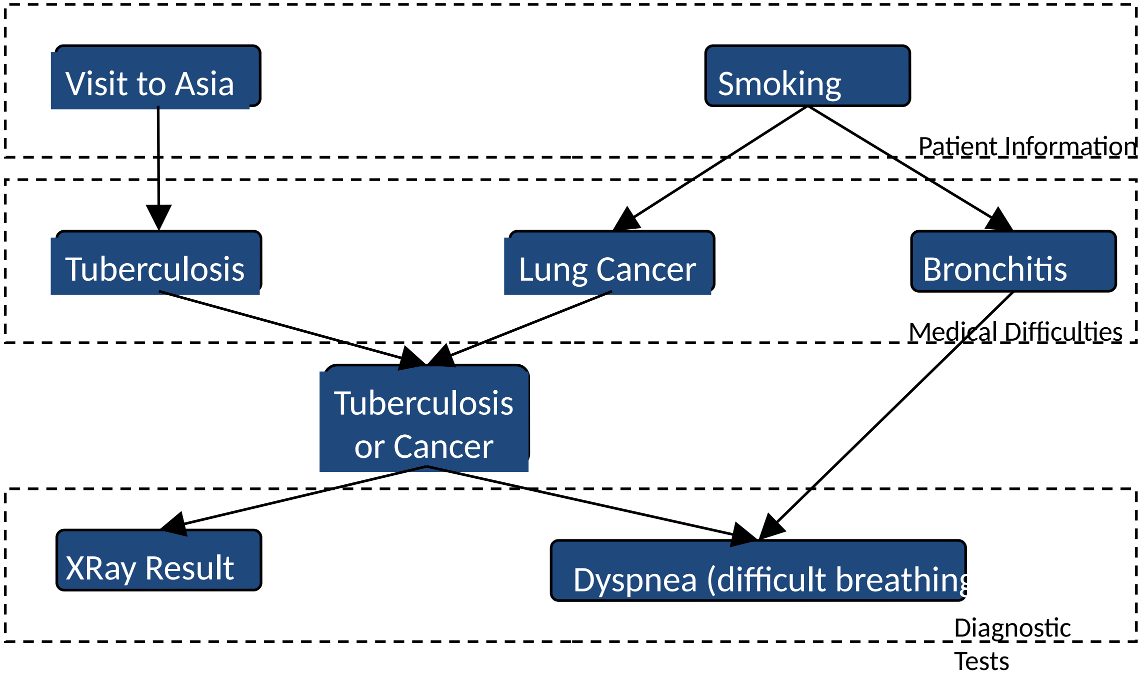
- An example query:

  P(*HasPneumonia* = *true* | *HasFever* = *true*, *HasCough* = *true*)?
- Note: Even though *HasDifficultyBreathing* and *ChestXrayPositive* are in the Bayesian network, they are not given values in the query
  - Neither query variables nor evidence variables
  - They are treated as unobserved or hidden variables

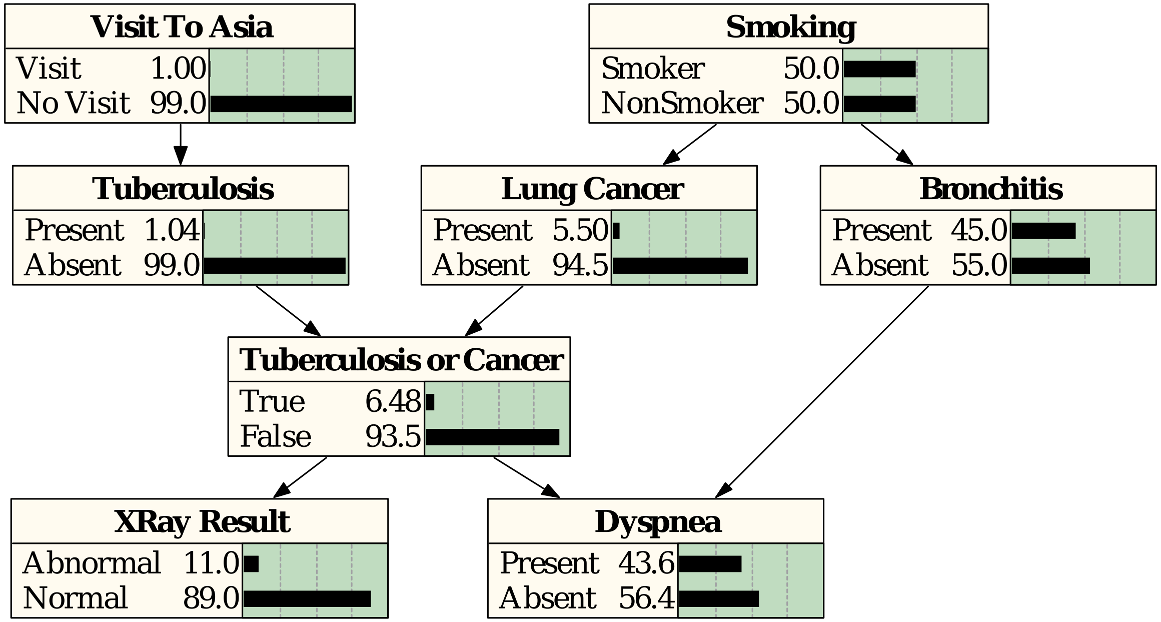


Example from Medical Diagnostics

- Network represents a knowledge structure that models the relationship between medical difficulties, their causes and effects, patient information and diagnostic tests



From D.M. Buede, J.A. Tatman, T.A. Bresnick "Introduction to Bayesian Networks," Tutorial for the 66th MORS Symposium



Visit To Asia
Visit 1.00
No Visit 99.0

Smoking
Smoker 50.0
NonSmoker 50.0

Tuberculosis
Present 1.04
Absent 99.0

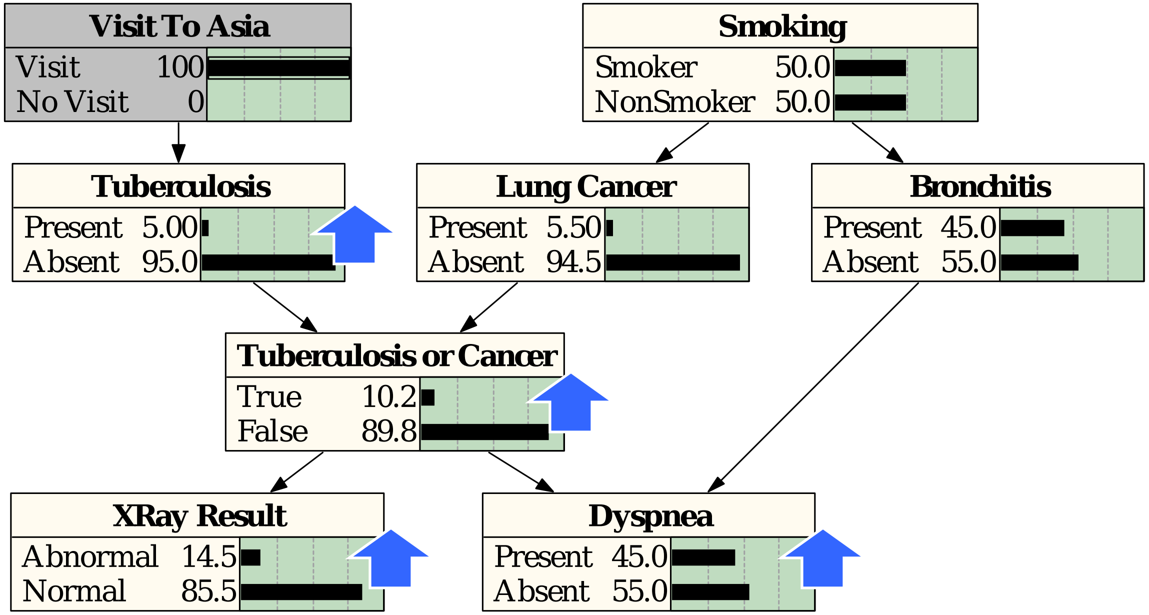
Lung Cancer
Present 5.50
Absent 94.5

Bronchitis
Present 45.0
Absent 55.0

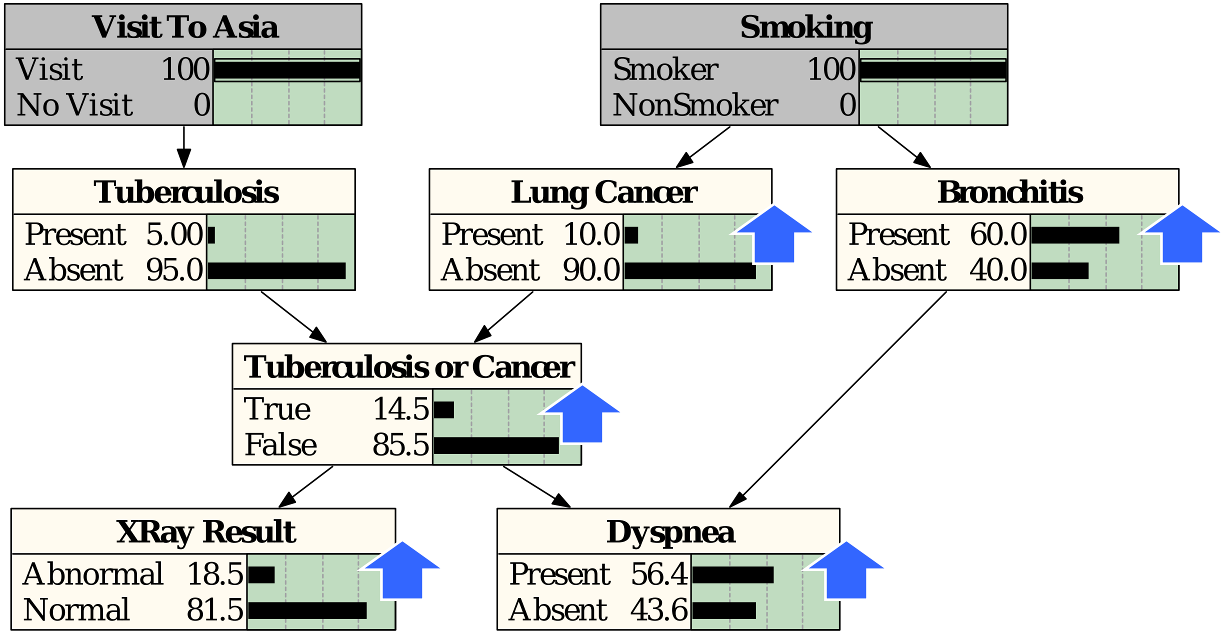
Tuberculosis or Cancer
True 6.48
False 93.5

XRay Result
Abnormal 11.0
Normal 89.0

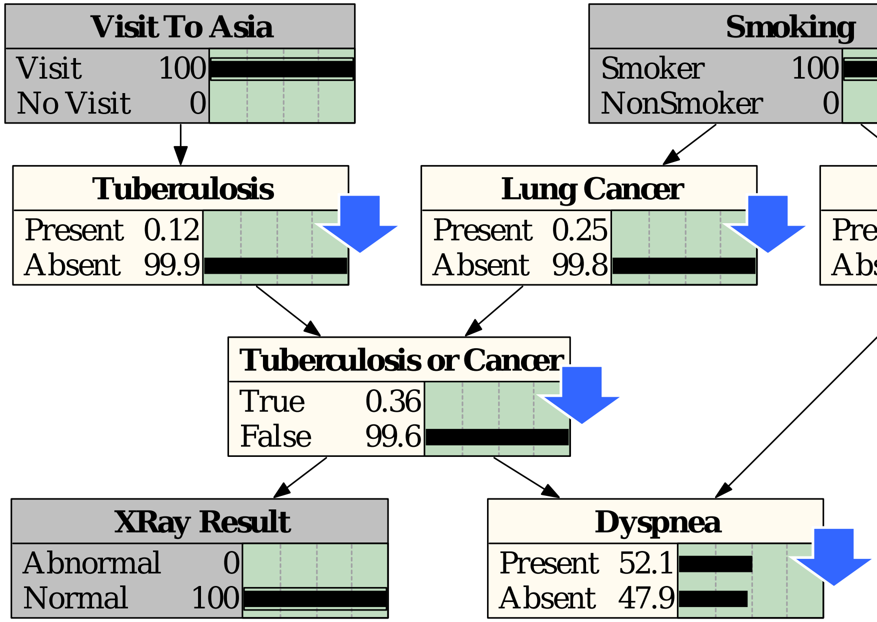
Dyspnea
Present 43.6
Absent 56.4



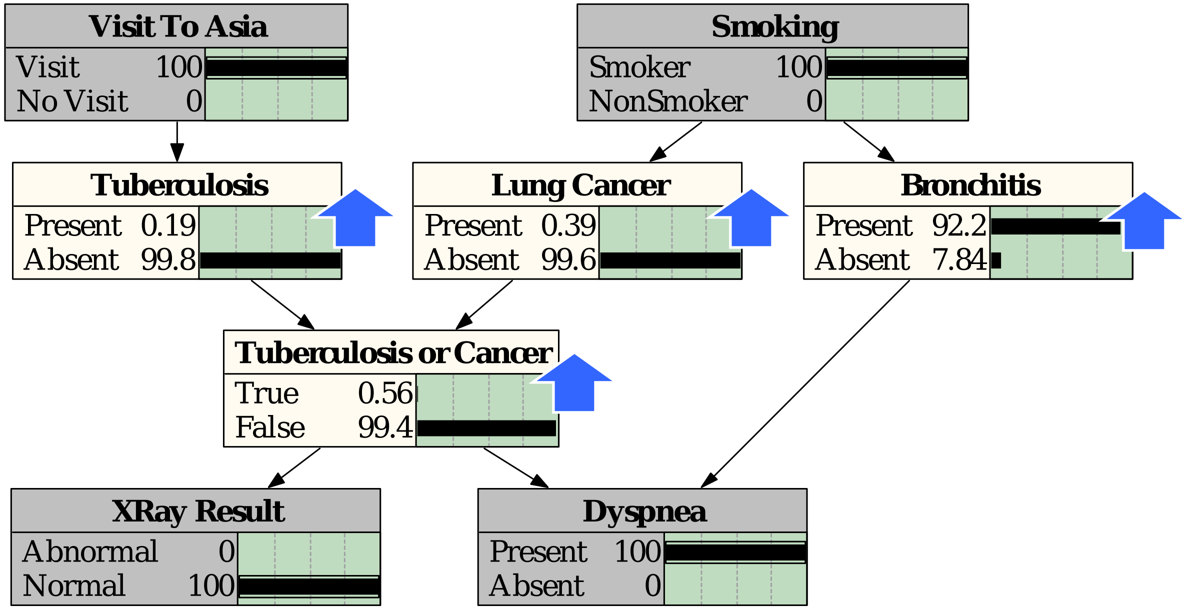
- As a finding is entered, the propagation algorithm updates the beliefs attached to each relevant node in the network
- Interviewing the patient produces the information that “Visit to Asia” is “Visit”
- This finding propagates through the network and the belief functions of several nodes are updated



- Further interviewing of the patient produces the finding “Smoking” is “Smoker”
- This information propagates through the network



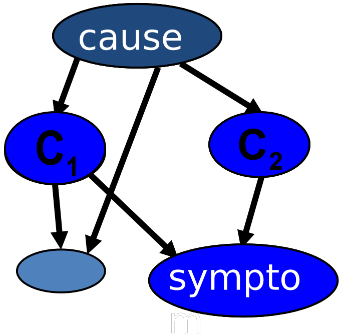
- Finished with interviewing the patient, the physician begins the examination
- The physician now moves to specific diagnostic tests such as an X-Ray, which results in a “Normal” finding which propagates through the network
- Note that the information from this finding propagates backward and forward through the arcs



- The physician also determines that the patient is having difficulty breathing, the finding “Present” is entered for “Dyspnea” and is propagated through the network
- The doctor might now conclude that the patient has bronchitis and does not have tuberculosis or lung cancer

Applications of BNs

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

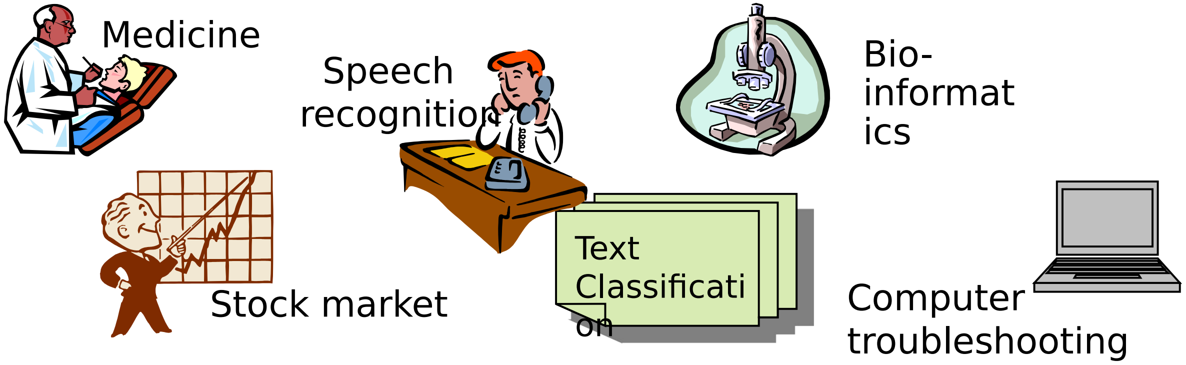


cause

C_1

C_2

sympto
m



Medicine

Speech recognition

Bio-
informat
ics

Stock market

Text
Classificati
on

Computer troubleshooting

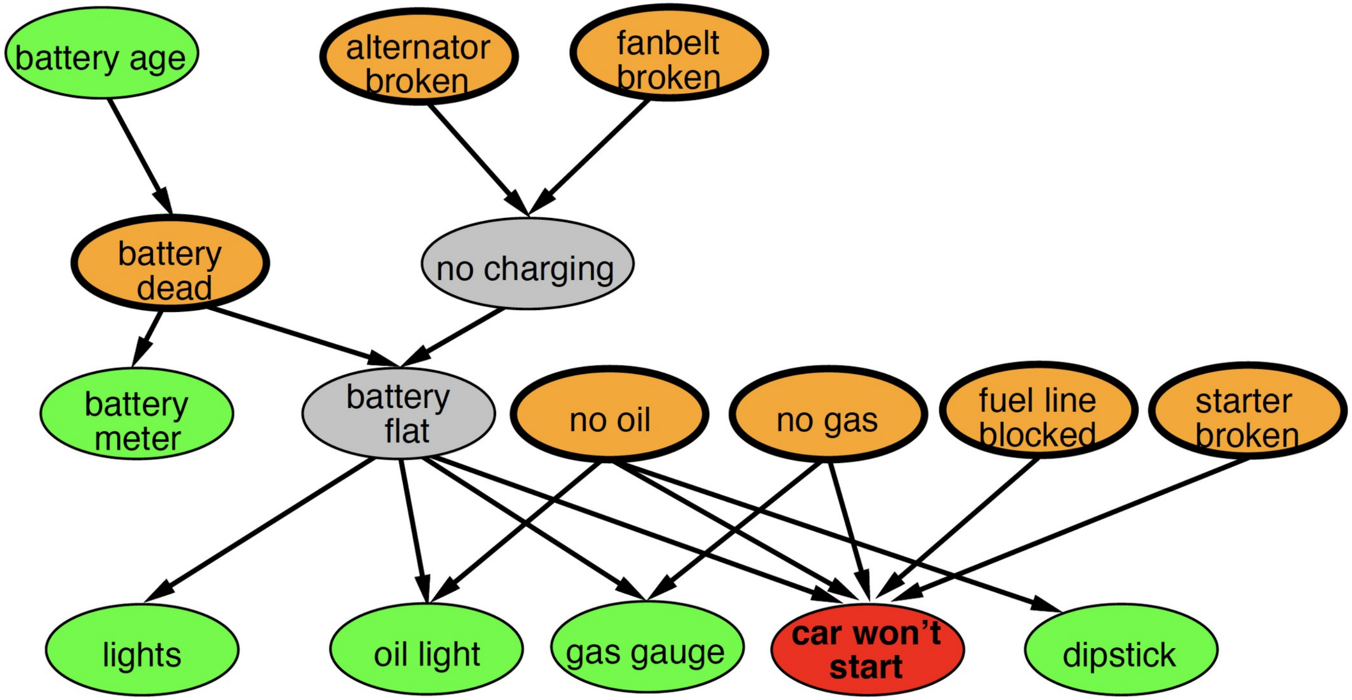
Application: car diagnosis

Example: Car diagnosis

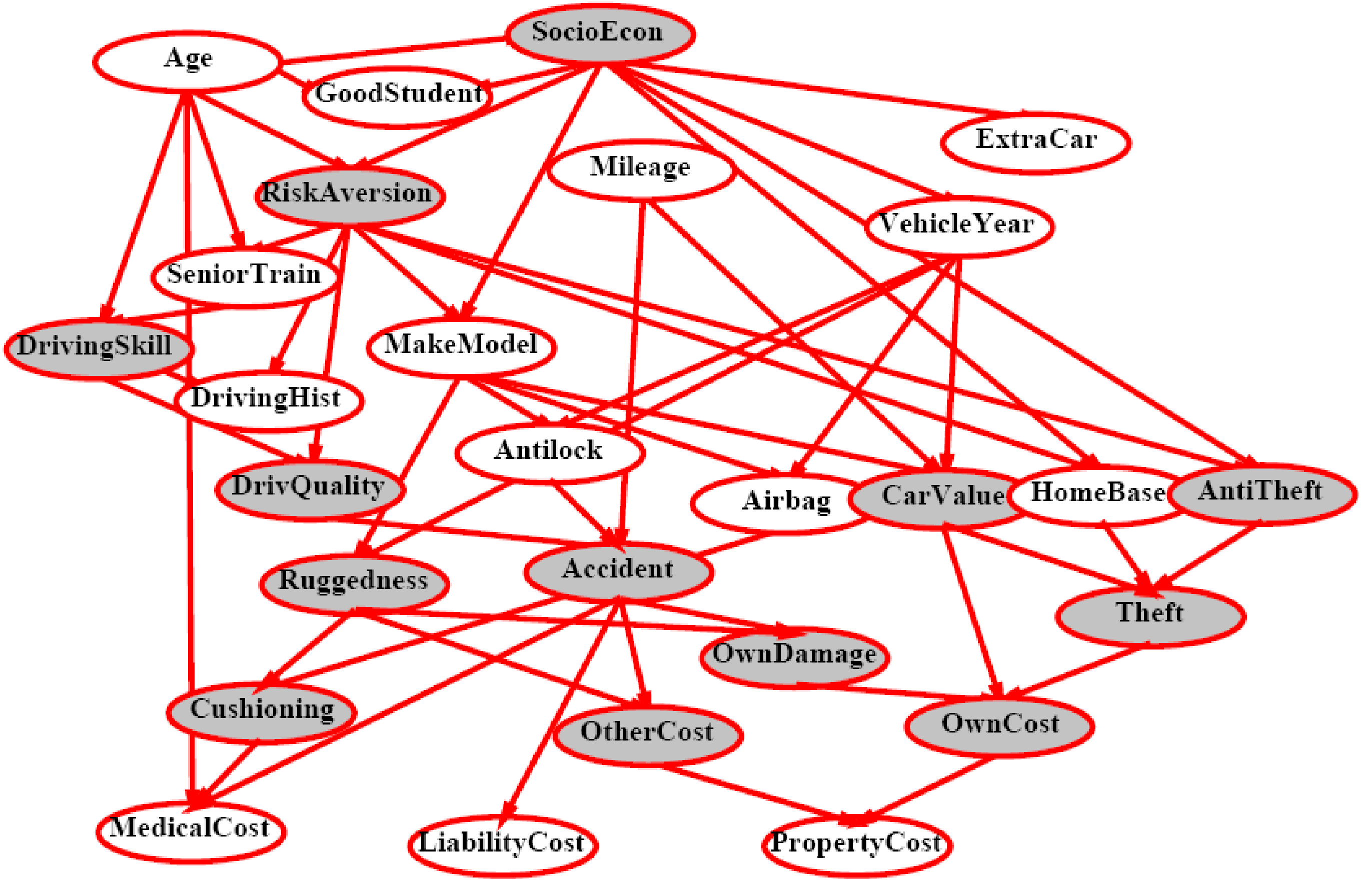
Initial evidence: car won't start

Testable variables (green), "broken, so fix it" variables (orange)

Hidden variables (gray) ensure sparse structure, reduce parameters

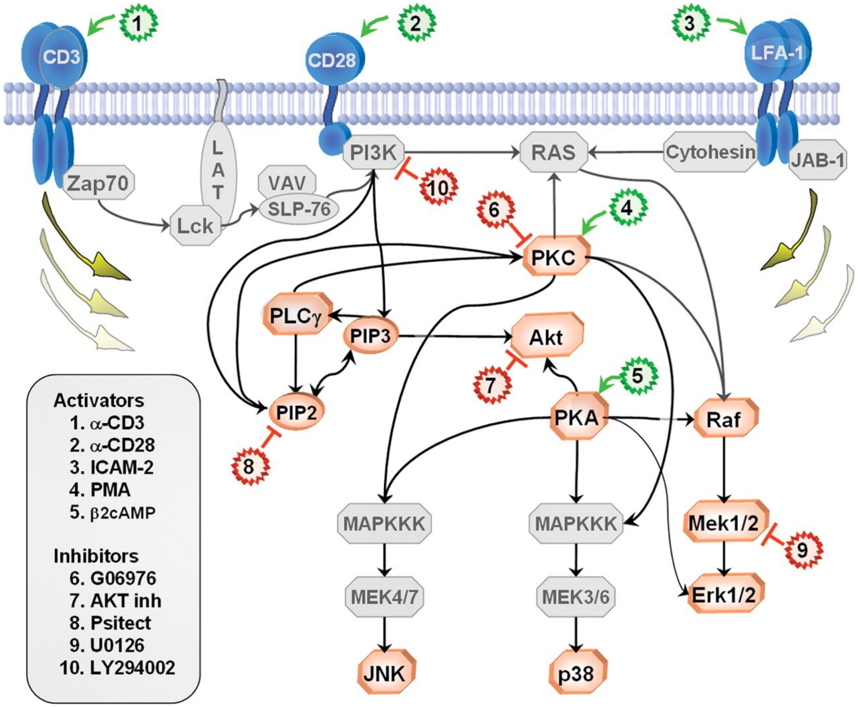


Car insurance



In research literature…

Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data
Karen Sachs, Omar Perez, Dana Pe'er, Douglas A. Lauffenburger, and Garry P. Nolan
(22 April 2005) *Science* **308** (5721), 523.



In research literature…

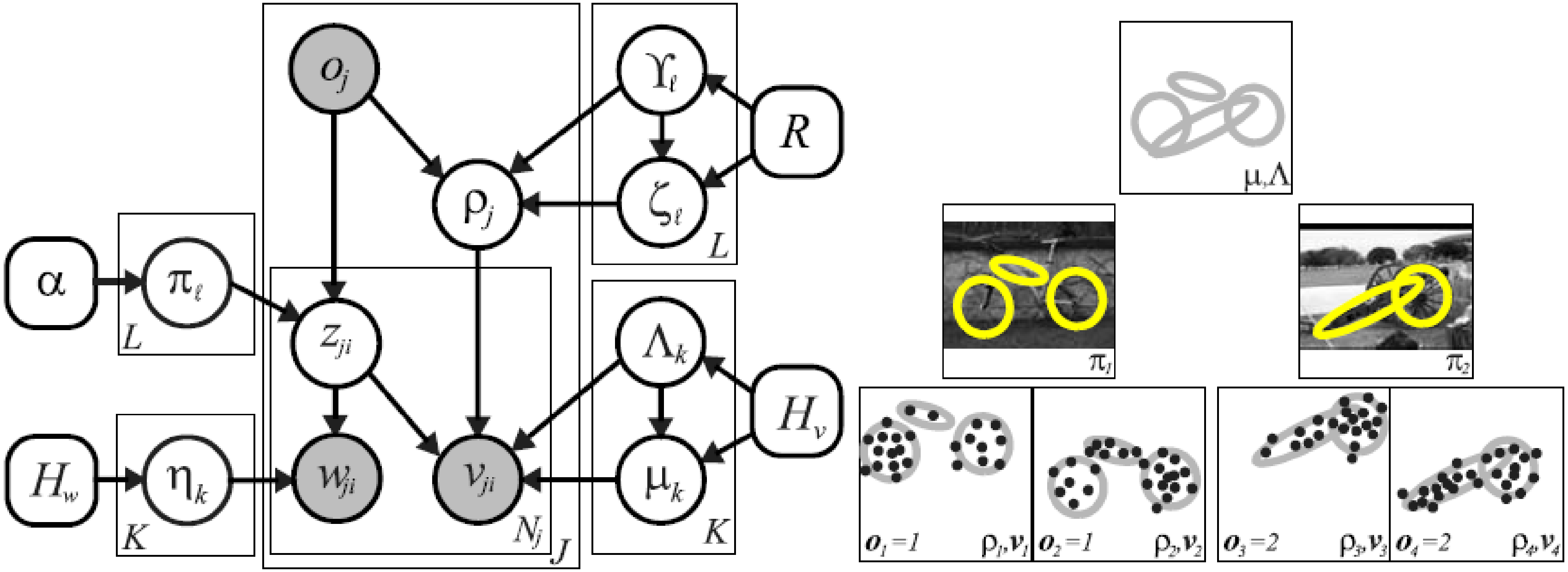
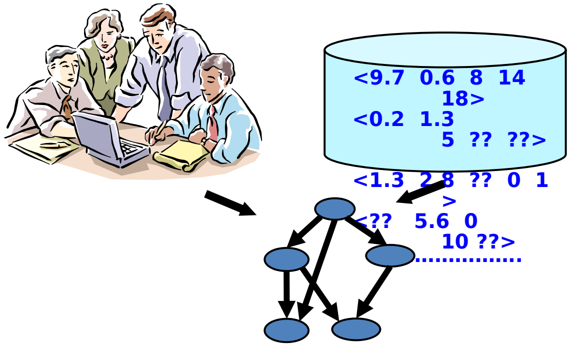


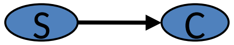
Fig. 3 A parametric, fixed-order model which describes the visual appearance of L object categories via a common set of K shared parts. The j^{th} image depicts an instance of object category o_j, whose position is determined by the reference transformation ρ_j. The appearance w_{ji} and position v_{ji}, relative to ρ_j, of visual features are determined by assignments $z_{ji} \sim \pi_{o_j}$ to latent parts. The cartoon example illustrates how a wheel part might be shared among two categories, *bicycle* and *cannon*. We show feature positions (but not appearance) for two hypothetical samples from each category

Advantages of BNs

- Combining domain expert knowledge with data
- Efficient representation and inference
- Handling missing data: Not all variable states need to be known in a query
- Learning causal relationships:



<9.7 0.6 8 14 18>
<0.2 1.3 5 ?? ??>
<1.3 2.8 ?? 0 1 >
<?? 5.6 0 10 ??>
....



How is the Bayesian network created?

1. Get an expert to design it
   - Expert must determine the structure of the Bayesian network
     - This is best done by modeling direct causes of a variable as its parents
   - Expert must determine the values of the CPT entries
     - These values could come from the expert's informed opinion
     - Or an external source e.g. census information
     - Or they are estimated from data
     - Or a combination of the above

2. **Learn** it from data
   - This is a much better option but it requires a large amount of data

Constructing Bayesian networks

1. Choose an ordering of variables $X_1, \dots , X_n$
2. For i = 1 to n
   - add X_i to the network
   - select parents from $X_1, \dots ,X_{i-1}$ such that
     $P(X_i \mid Parents(X_i)) = P(X_i \mid X_1, \dots X_{i-1})$

Example for BN construction: Fire Diagnosis

You want to diagnose whether there is a fire in a building

- If there is a fire, there *may* be smoke
- If there is a fire alarm, it *may* have been caused by a fire *or* by tampering
- If everyone is leaving, this *may* have been caused by a fire alarm
- You receive a *noisy* report about whether everyone is leaving the building

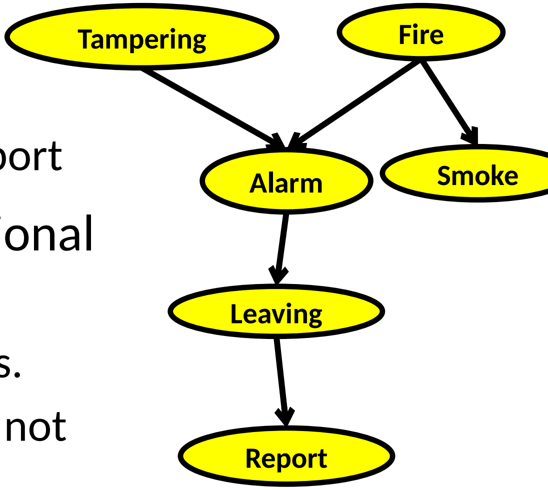
Example for BN construction: Fire Diagnosis

First you choose the variables. In this case, all are Boolean:

- Fire is true when there is a fire
- Smoke is true when there is smoke
- Alarm is true when there is an alarm
- Tampering is true when the alarm has been tampered with
- Leaving is true if there are lots of people leaving the building
- Report is true if the news reports that lots of people are leaving the building

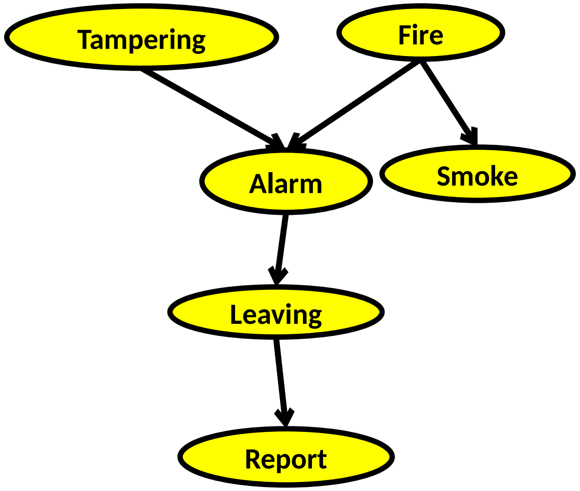
- Let’s construct the Bayesian network for this

Example for BN construction: Fire Diagnosis



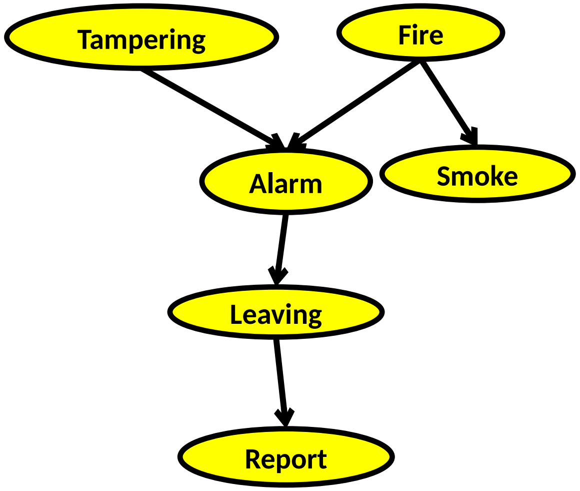
- Using the total ordering of variables:
  - (1)Fire, (2) Tampering, (3) Alarm, (4) Smoke, (5) Leaving, (6) Report
- Choose the parents for each variable by evaluating conditional independencies
  - Fire is the first variable in the ordering. It does not have parents.
  - Tampering independent of fire (learning that one is true would not change your beliefs about the probability of the other)
  - Alarm depends on both Fire and Tampering: it could be caused by either or both
  - Smoke is caused by Fire, and so is independent of Tampering and Alarm given whether there is a Fire
  - Leaving is caused by Alarm, and thus is independent of the other variables given Alarm
  - Report is caused by Leaving, and thus is independent of the other variables given Leaving

Example for BN construction: Fire Diagnosis



P(Tampering, Fire, Alarm, Smoke, Leaving, Report) =
P(Tampering) x P(Fire) x P(Alarm|Tampering, Fire) x P(Smoke|Fire) x P(Leaving|Alarm) x P(Report|Leaving)

Example for BN construction: Fire Diagnosis

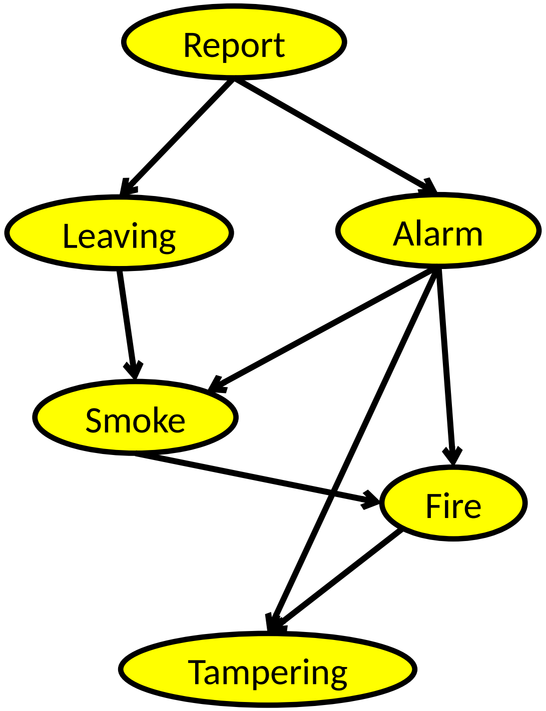


- How many probabilities do we need to specify for this Bayesian network?
  - 1+1+4+2+2+2 = 12

BN construction

- Order matters!
- The complexity of the BN depends on the order in which the variables are introduced
- A “bad” order -> many unnecessary links
    - Not exploiting conditional independence as much

Example for BN construction: Fire Diagnosis



- How many probabilities do we need to specify for this Bayesian network?
  - 1+2+2+4+4+4 = 17

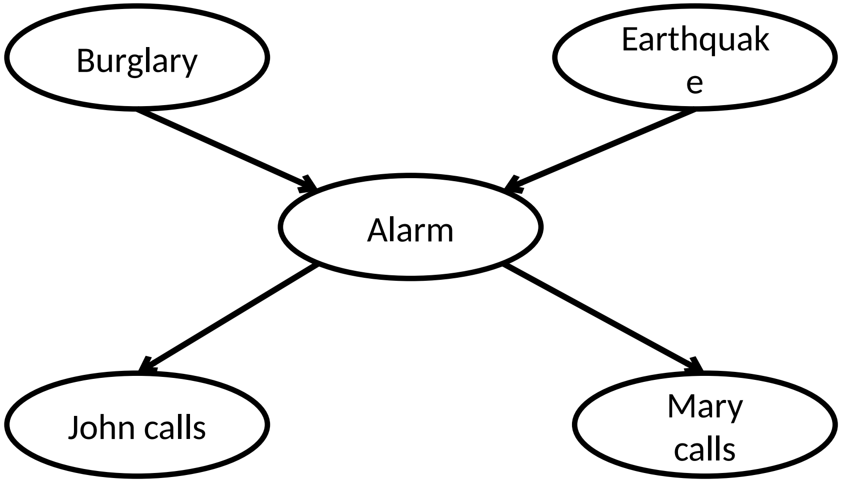
Classwork: Burglar Alarm

- I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm
  - Example inference task: suppose Mary calls and John doesn’t call. What is the probability of a burglary?
- What are the random variables?
  - *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- Create a Bayesian Network (only the graph) for this application

Classwork: Burglar Alarm

- I have a burglar alarm that is sometimes set off by minor earthquakes. My two neighbors, John and Mary, promised to call me at work if they hear the alarm
  - Example inference task: suppose Mary calls and John doesn't call. What is the probability of a burglary?
- What are the random variables?
  - *Burglary*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*
- Create a Bayesian Network (only the graph) for this application
- What are the direct influence relationships?
  - A burglar can set the alarm off
  - An earthquake can set the alarm off
  - The alarm can cause Mary to call
  - The alarm can cause John to call

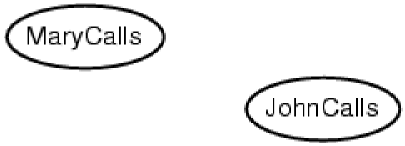
The ideal network based on *causality*



Example

- Suppose we choose the ordering M, J, A, B, E

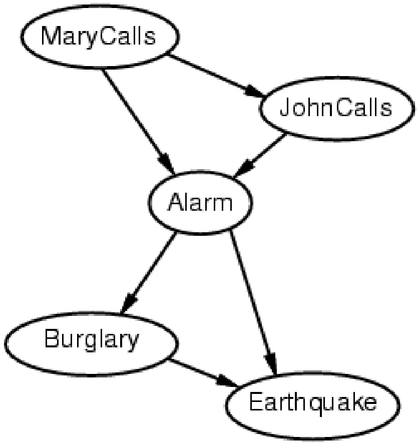
P(J | M) = P(J)?



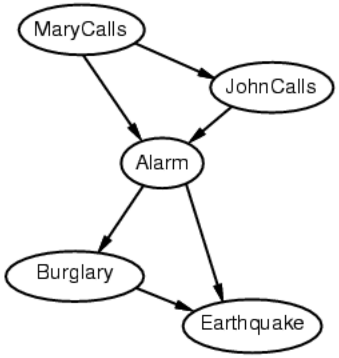
Example

- Suppose we choose the ordering M, J, A, B, E

P(J | M) = P(J)?	No
P(A | J, M) = P(A)?	No
P(A | J, M) = P(A | J)?	No
P(A | J, M) = P(A | M)?	No
P(B | A, J, M) = P(B)?	No
P(B | A, J, M) = P(B | A)?	Yes
P(E | B, A ,J, M) = P(E)?	No
P(E | B, A, J, M) = P(E | A, B)?	Yes



Example: The network we constructed

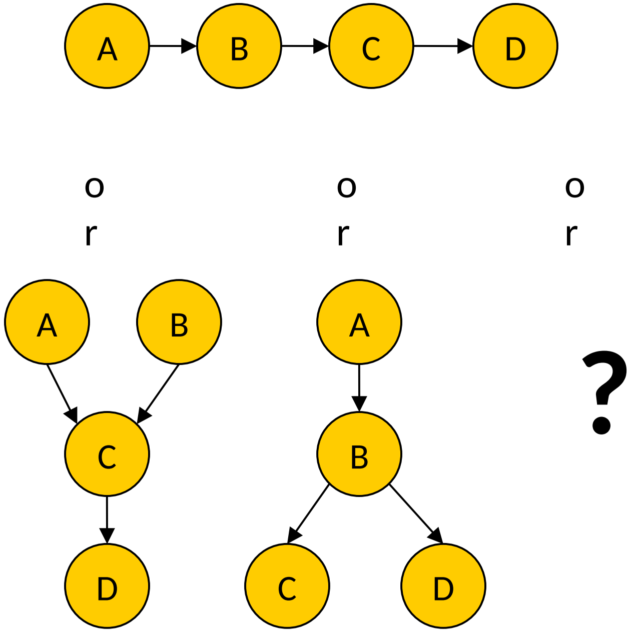


- Deciding conditional independence is hard in non-causal directions
  - The causal direction seems much more natural
- Network is less compact: 1 + 2 + 4 + 2 + 4 = 13 numbers needed

Learning Bayesian Networks from Data

Given a data set, which Bayesian network with variables A, B, C and D best represents the data?

A	B	C	D
true	false	false	true
true	false	true	false
true	false	false	true
false	true	false	false
false	true	false	true
false	true	false	false
false	true	false	false
:	:	:	:



Bayesian Networks summary

Two important properties:

1. Is a compact representation of the joint probability distribution over the variables
2. Encodes the conditional independence relationships between the variables in the graph structure

References

- George F. Luger, Artificial Intelligence: Structures and Strategies for Complex Problem Solving, 6th edition, Addison Wesley, 2009. **Chapters 9.3.1-9.3.3**.
- Russel and Norvig, Artificial Intelligence: A Modern Approach, 3rd edition, Prentice Hall, 2010. **Chapter 14.1-14.4**