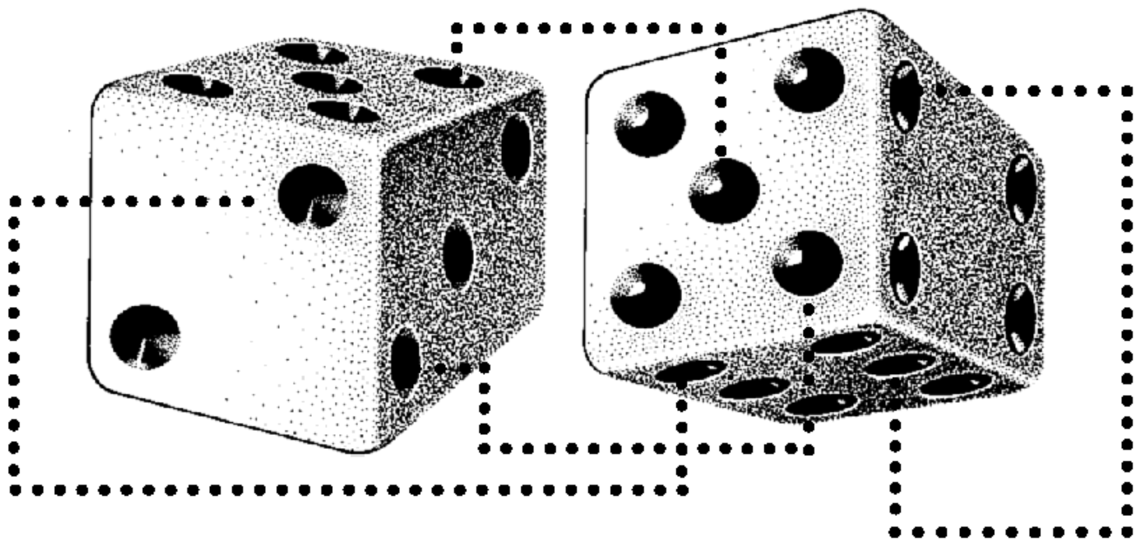
Introduction to Bayesian Networks



Based on the Tutorials and Presentations:

(1) Dennis M. Buede Joseph A. Tatman, Terry A. Bresnick;

(2) Jack Breese and Daphne Koller;

(3) Scott Davies and Andrew Moore;

(4) Thomas Richardson

(5) Roldano Cattoni

(6) Irina Rich

Introduction

Suppose you are trying to determine if a patient has inhalational anthrax. You observe the following symptoms:

- The patient has a cough
- The patient has a fever
- The patient has difficulty breathing

Introduction

You would like to determine how likely the patient is infected with inhalational anthrax given that the patient has a cough, a fever, and difficulty breathing

We are not 100% certain that the patient has anthrax because of these symptoms. We are dealing with uncertainty!

Introduction

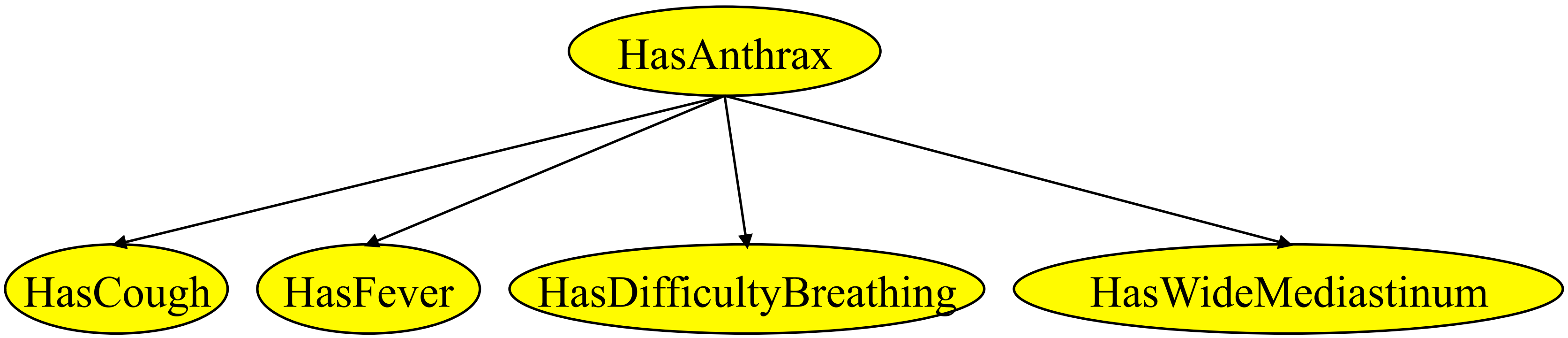
Now suppose you order an x-ray and observe that the patient has a wide mediastinum.

Your belief that that the patient is infected with inhalational anthrax is now much higher.

Introduction

- **In the previous slides, what you observed affected your belief that the patient is infected with anthrax**
- **This is called reasoning with uncertainty**
- **Wouldn't it be nice if we had some methodology for reasoning with uncertainty? Why in fact, we do…**

Bayesian Networks



- **In the opinion of many AI researchers, Bayesian networks are the most significant contribution in AI in the last 10 years**

Conditional probability

- A *conditional probability statement* has the form:
- "Given that B happened, the probability that A happened is x."
  - » This is written: $P(A \mid B) = x$.
- This does *not* imply that whenever B is true the probability of A is x
  - » P(Snow today | Snow yesterday)
  - » $\neq$ P(Snow today | Snow yesterday, the month is July)

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: $P(Y) = P(Y,x) + P(Y,\overline{x})$

Bayes rule

Likelihood of Evidence given Hypothesis

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

Posterior Probability of Hypothesis given evidence

Normalizing Constant

Prior Probability of Hypothesis

What are Bayesian nets?

- **Bayesian nets (BN) are a network-based framework for representing and analyzing models involving uncertainty;**

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

- **BN are different from other probabilistic analysis tools because of network representation of problems, use of Bayesian statistics, and the synergy between these**

Definition of a Bayesian Network

Knowledge structure:

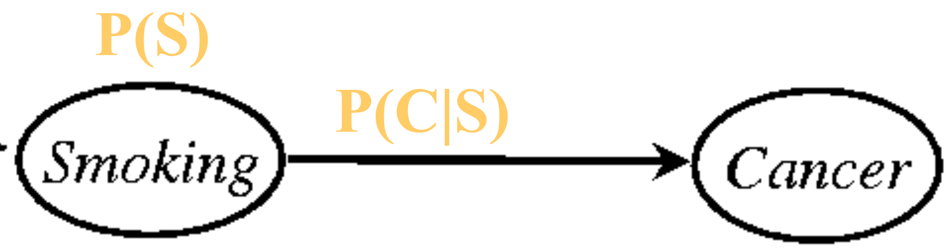
- variables are nodes
- arcs represent probabilistic dependence between variables
- conditional probabilities encode the strength of the dependencies

Computational architecture:

- computes posterior probabilities given evidence about some nodes
- exploits probabilistic independence for efficient computation

Bayesian Networks

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



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

P(S)

$P(S{=}no)$	0.80
$P(S{=}light)$	0.15
$P(S{=}heavy)$	0.05

P(C|S)

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

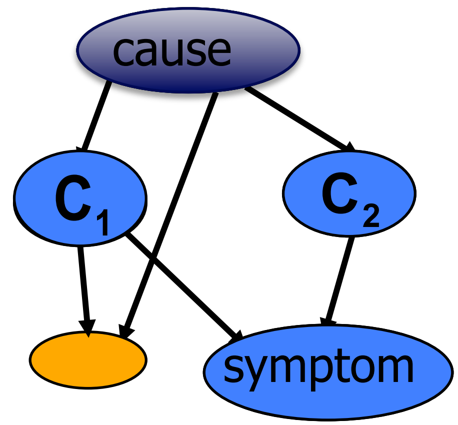
$$P(S \mid C) = \frac{P(C \mid S)P(S)}{P(C)} = \frac{P(C,S)}{P(C)}$$

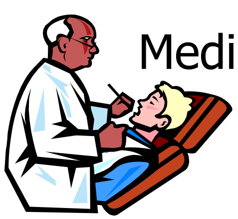
$S\Downarrow \quad C\Rightarrow$	*none*	*benign*	*malig*
no	0.768/.935	0.024/.046	0.008/.019
light	0.132/.935	0.012/.046	0.006/.019
heavy	0.030/.935	0.015/.046	0.005/.019

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

What Bayesian Networks are good for?

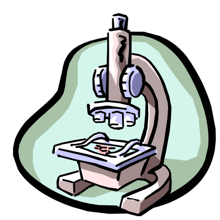
- Diagnosis: P(cause|symptom)=?
- Prediction: P(symptom|cause)=?
- Classification: $\max_{class}$ P(class|data)
- Decision-making (given a cost function)





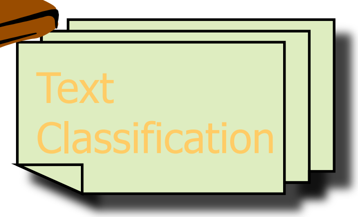
Medicine

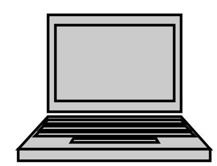
Speech recognition



Bio-
informatics

Stock market

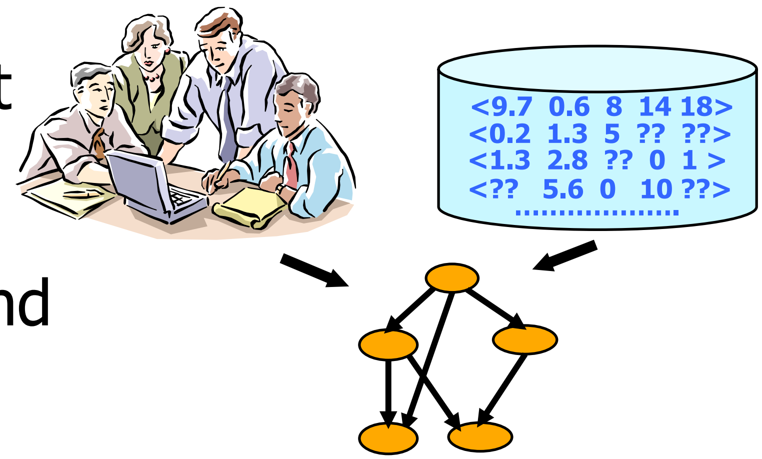




Computer troubleshooting

Why learn Bayesian networks?

- Combining domain expert knowledge with data

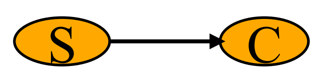


- Efficient representation and inference

- Incremental learning

- Handling missing data: **<1.3 2.8 ?? 0 1 >**

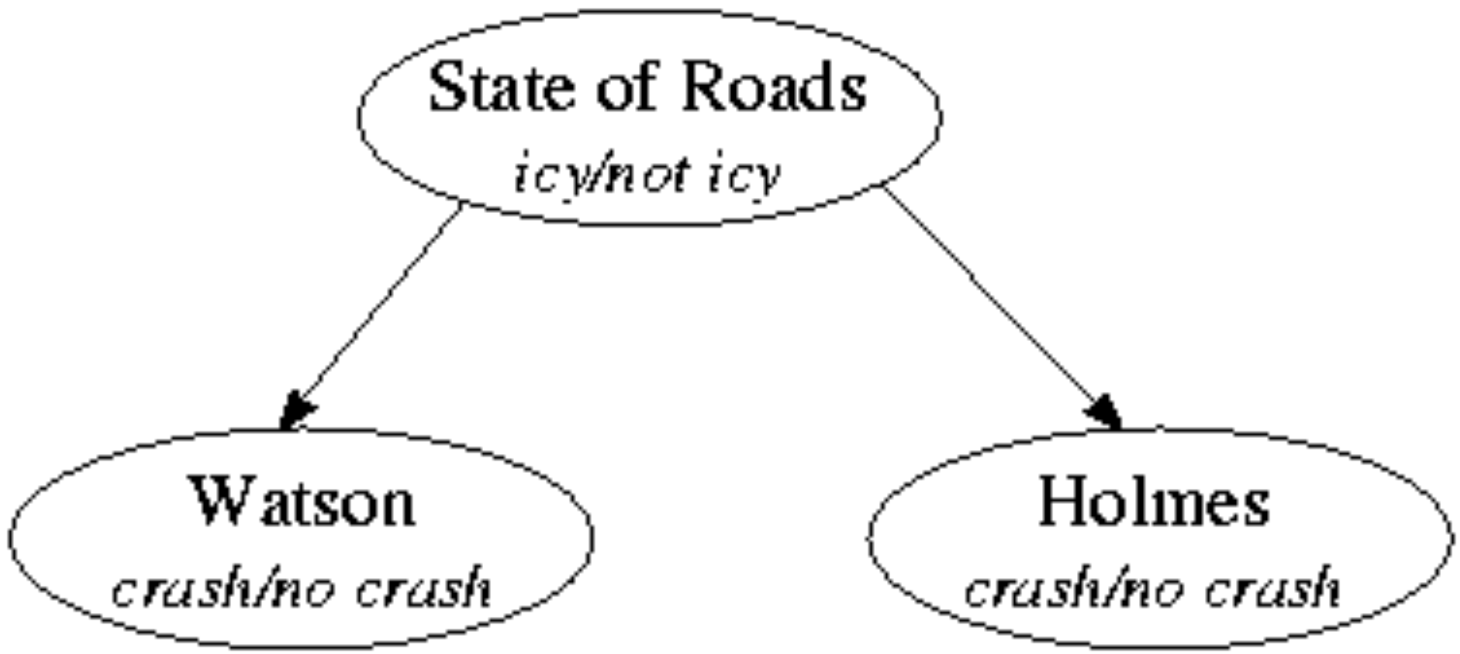
- Learning causal relationships:



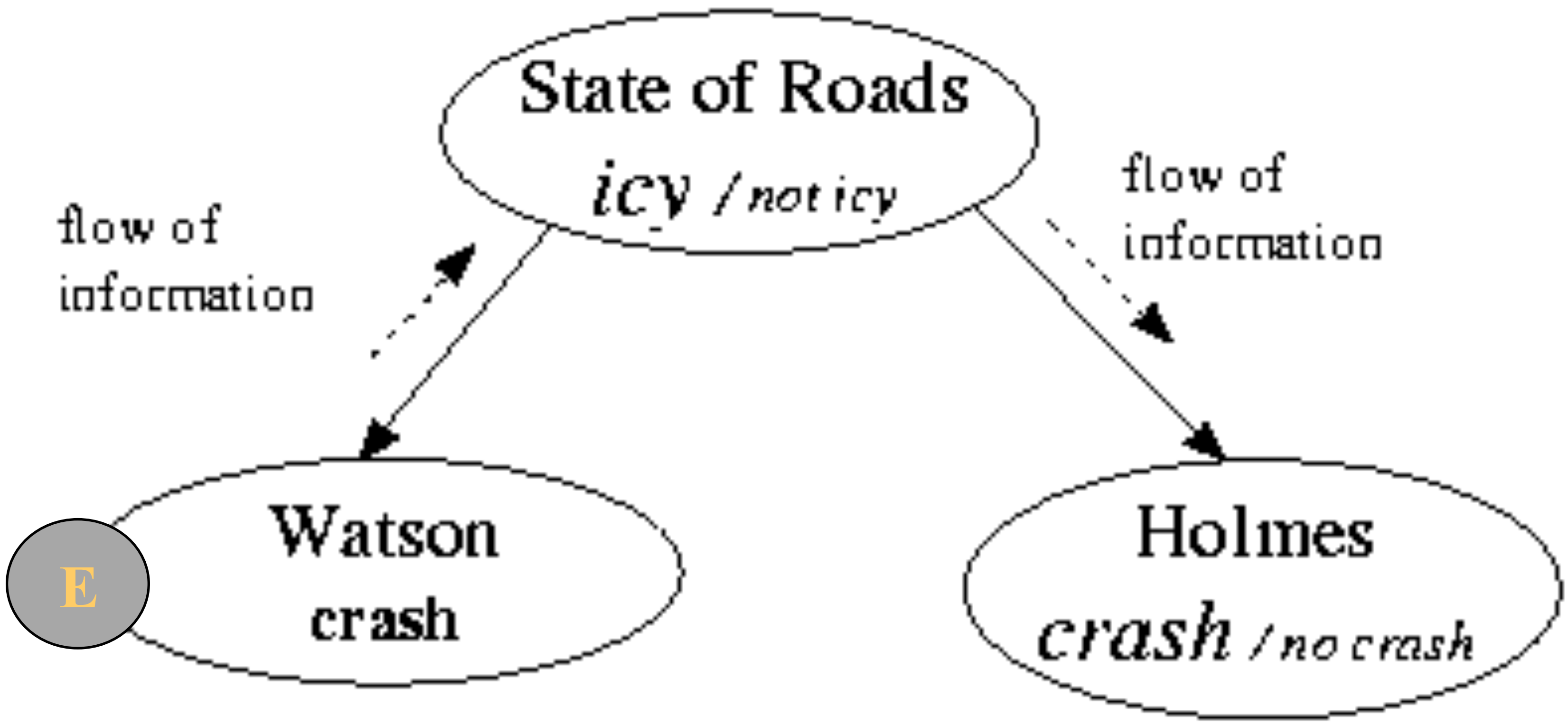
"Icy roads" example

- Inspector Smith is waiting for Holmes and Watson who are both late for an appointment.
- Smith is worried that if the roads are icy one or both of them may have crashed his car.
- Suddenly Smith learns that Watson has crashed.
- Smith thinks: *If Watson has crashed, probably the roads are icy, then Holmes has probably crashed too!*
- Smith then learns it is warm outside and roads are salted
- Smith thinks: *Watson was unlucky; Holmes should still make it.*

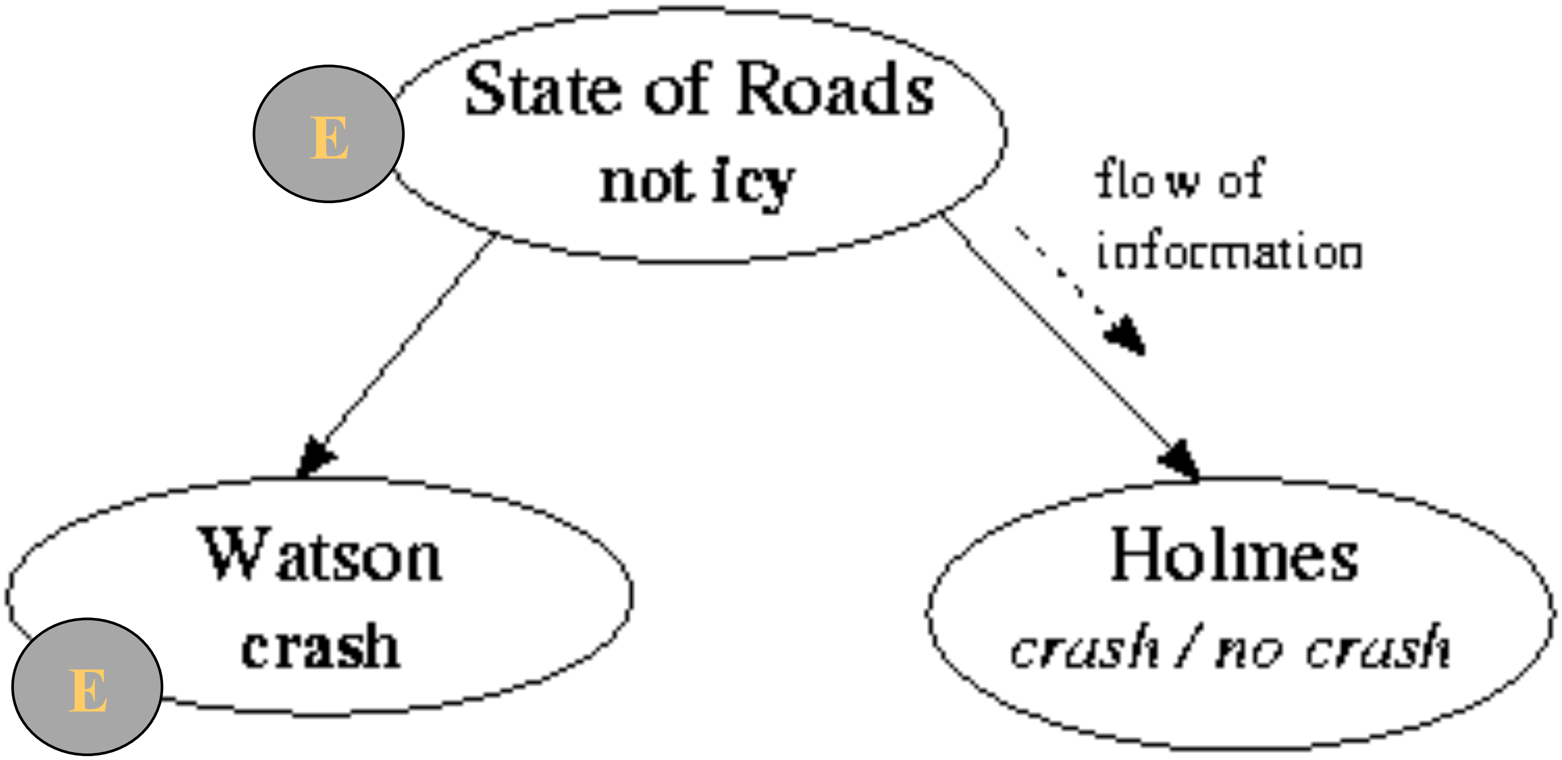
Causal relationships



Watson has crashed !



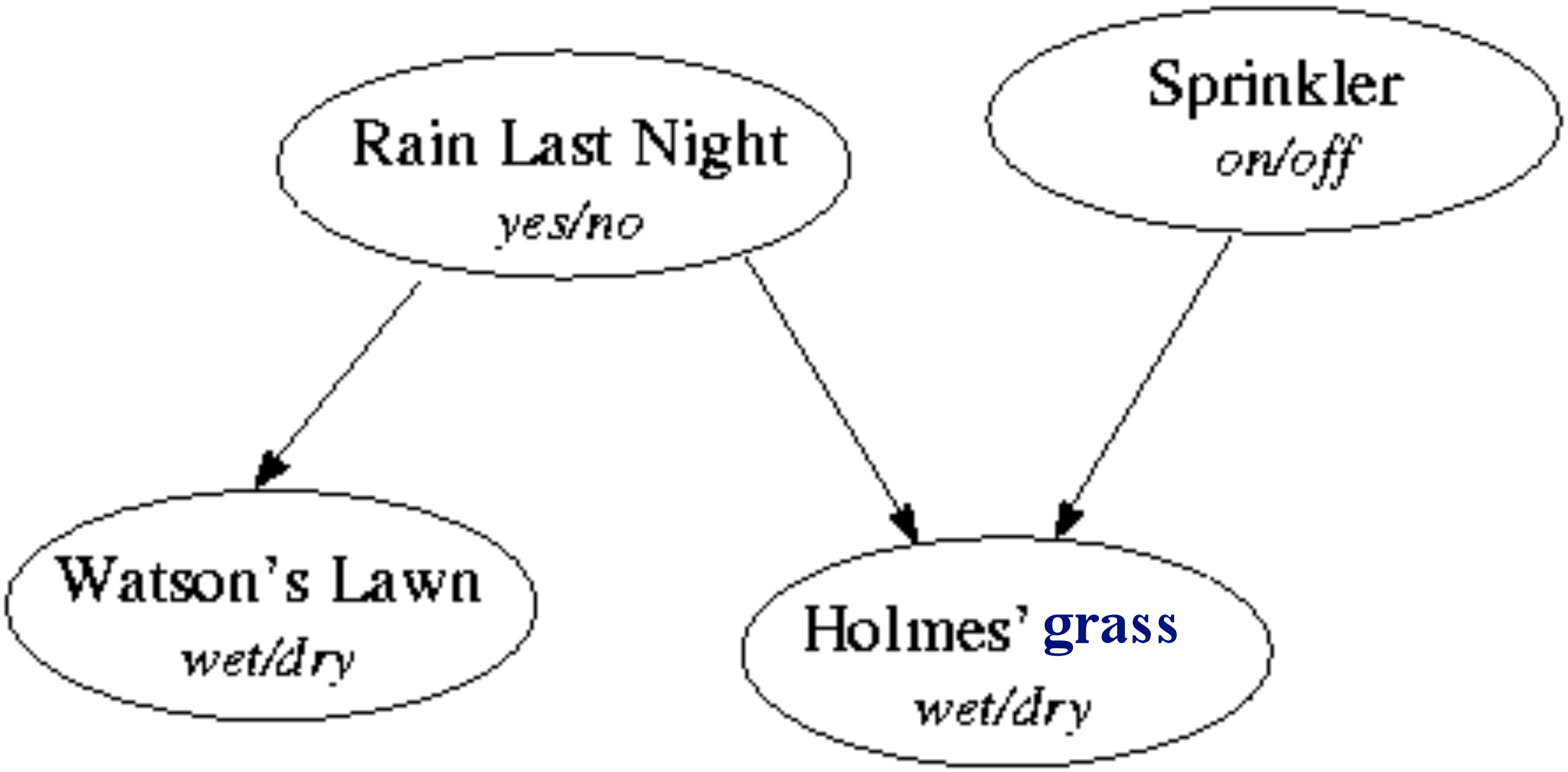
… But the roads are salted !



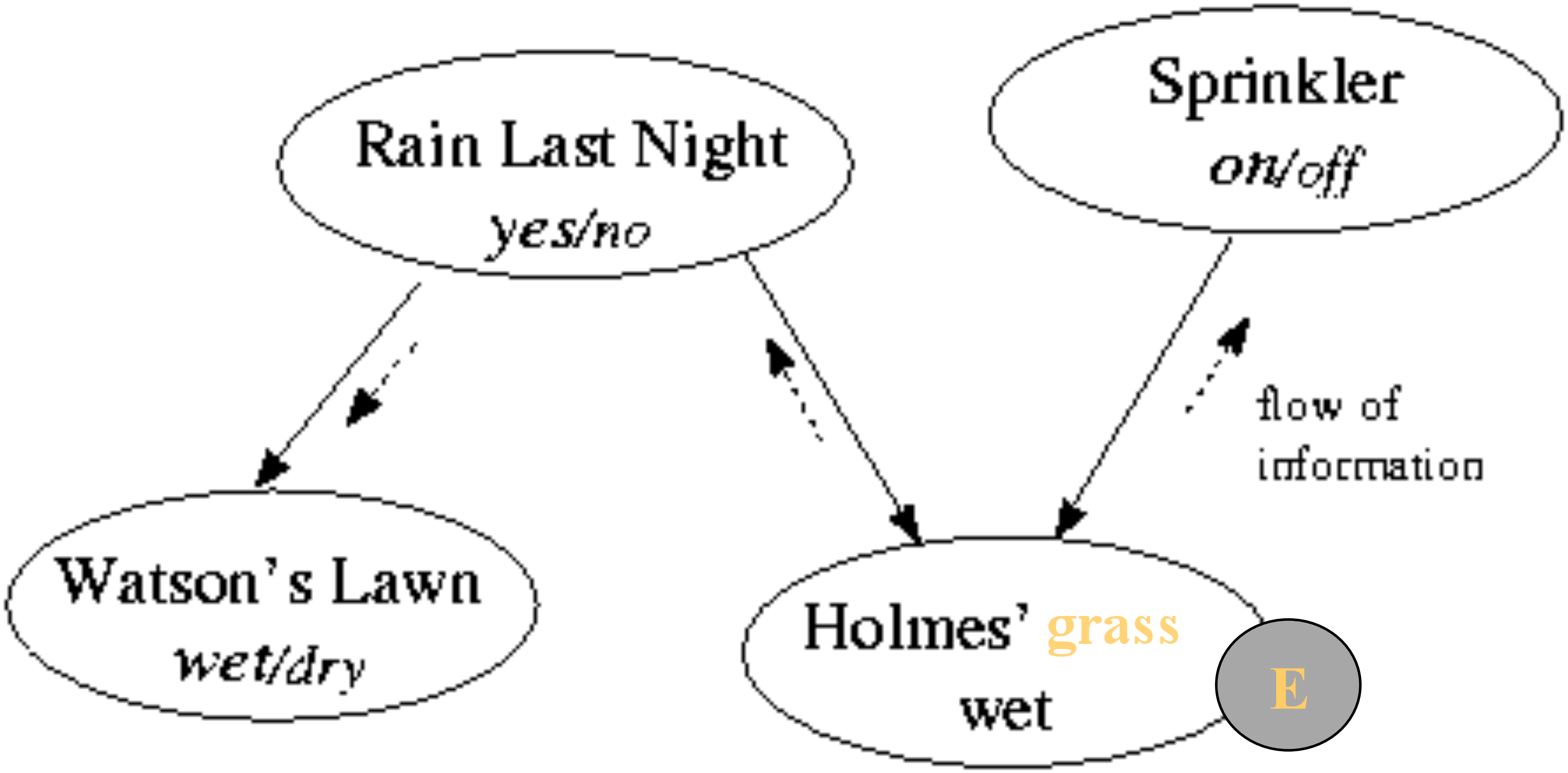
"Wet grass" example

- One morning as Holmes leaves for work, he notices that his grass is wet. He wonders whether he has left his sprinkler on, or it has rained.
- Glancing over to Watson's lawn he notices that it is also wet.
- Holmes thinks: *Since Watson's lawn is wet, it probably rained last night.*
- He then thinks: *If it rained then that explains why my grass is wet, so probably the sprinkler is off.*

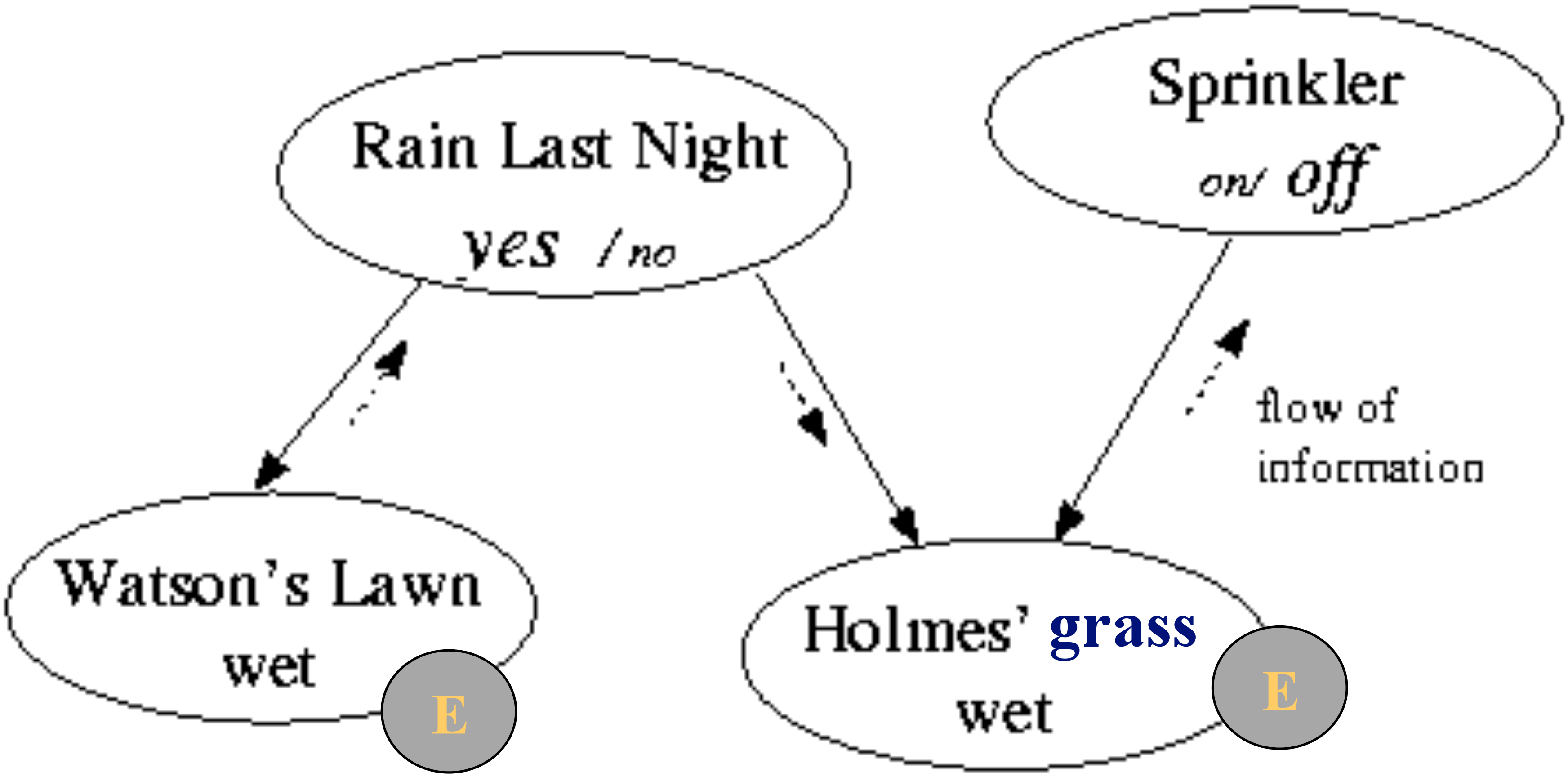
Causal relationships



Holmes' grass is wet !



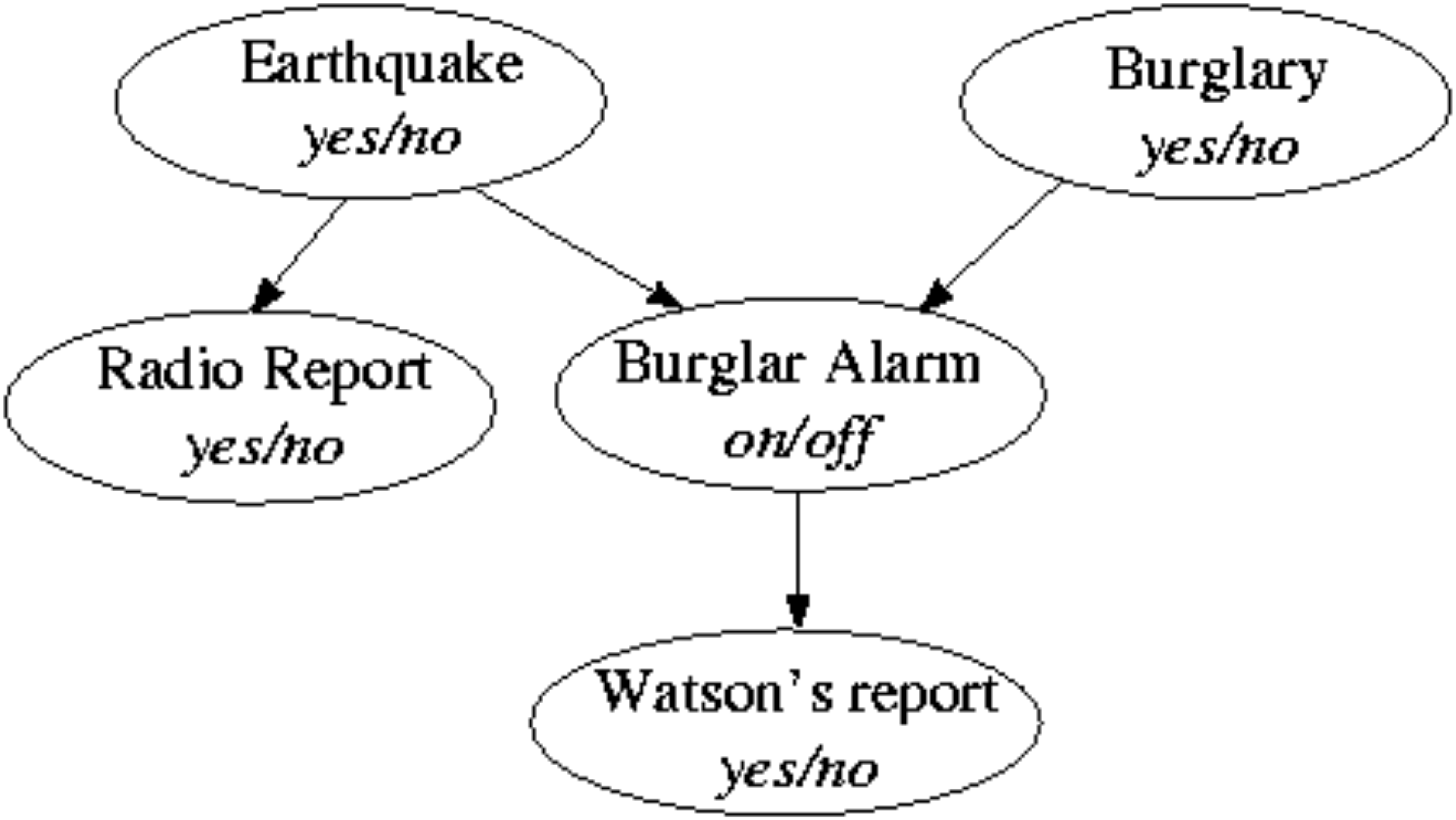
Watson's lawn is also wet !



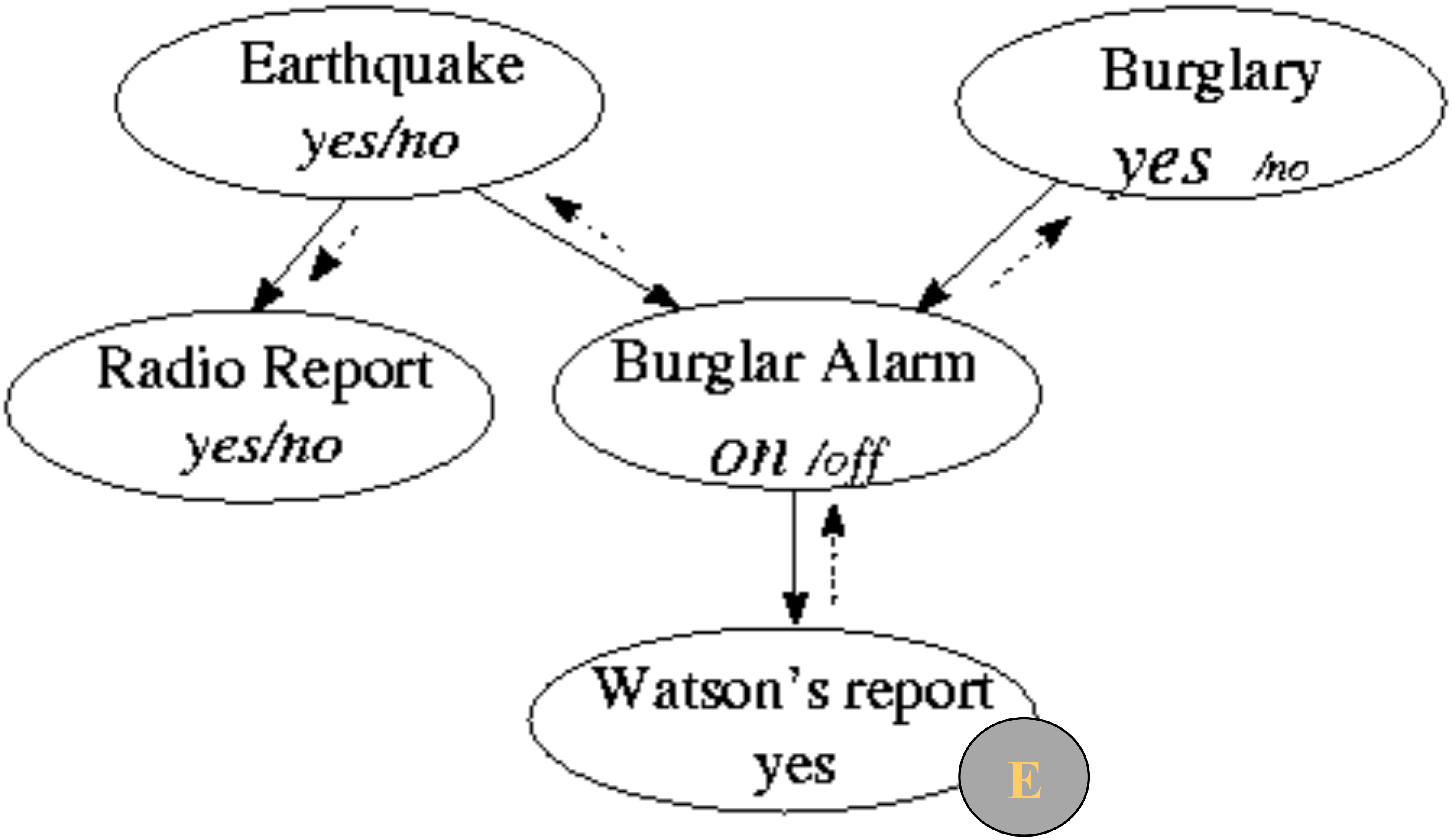
"Burglar alarm" example

- Holmes is at work when he receives a call from Watson, informing him that his alarm has gone off.
- Holmes thinks it is likely that the alarm really went off, although Watson sometimes play practical jokes.
- Holmes is on his way home when he hears a report on the radio, that there was an earthquake in the vicinity.
- Since the burglar alarm has been known to go off when there is an earthquake, Holmes reckons that a burglary is unlikely.
- Holmes goes back to work. (Leaving the noise for Watson)

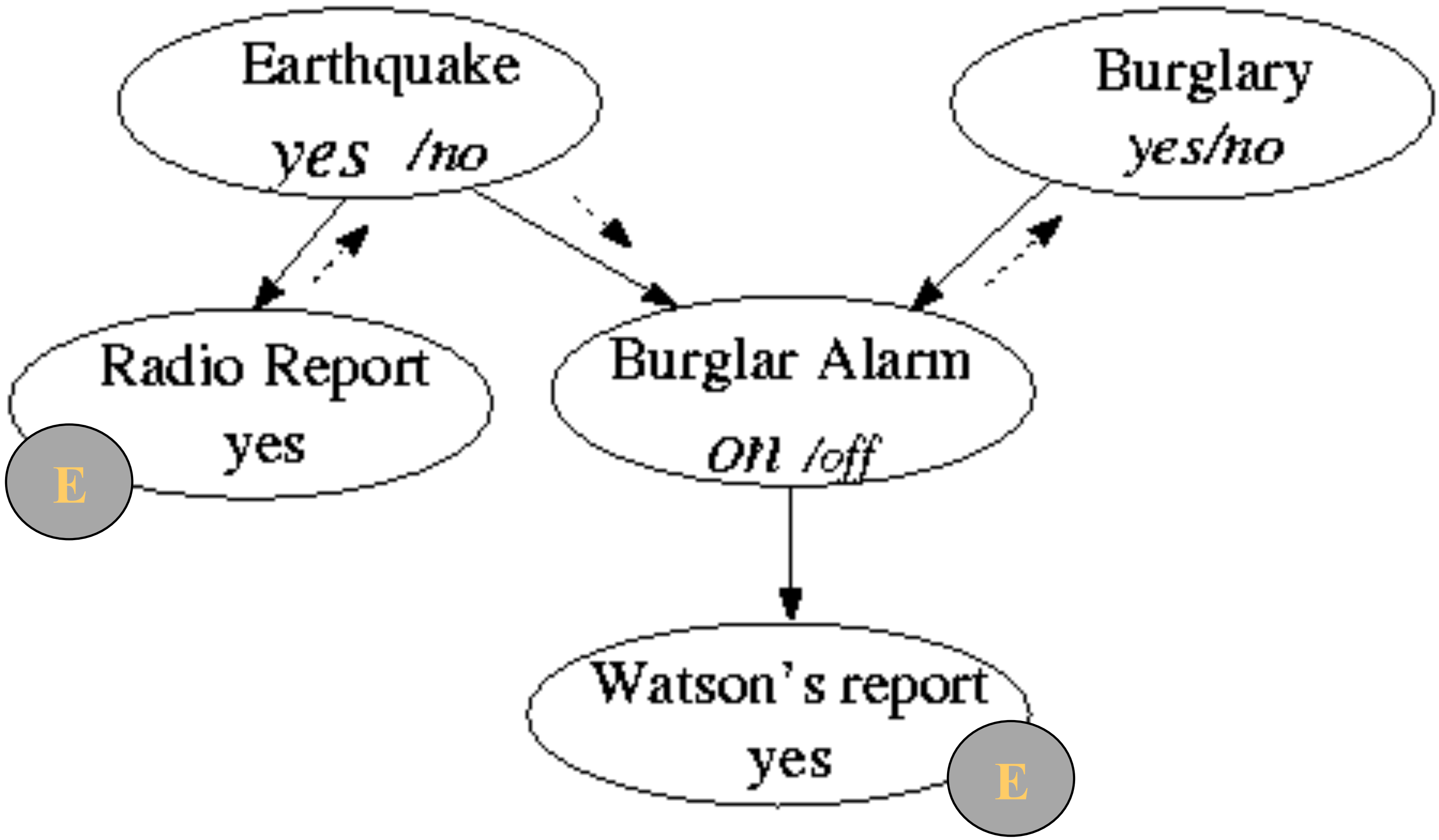
Causal relationships



Watson reports about alarm

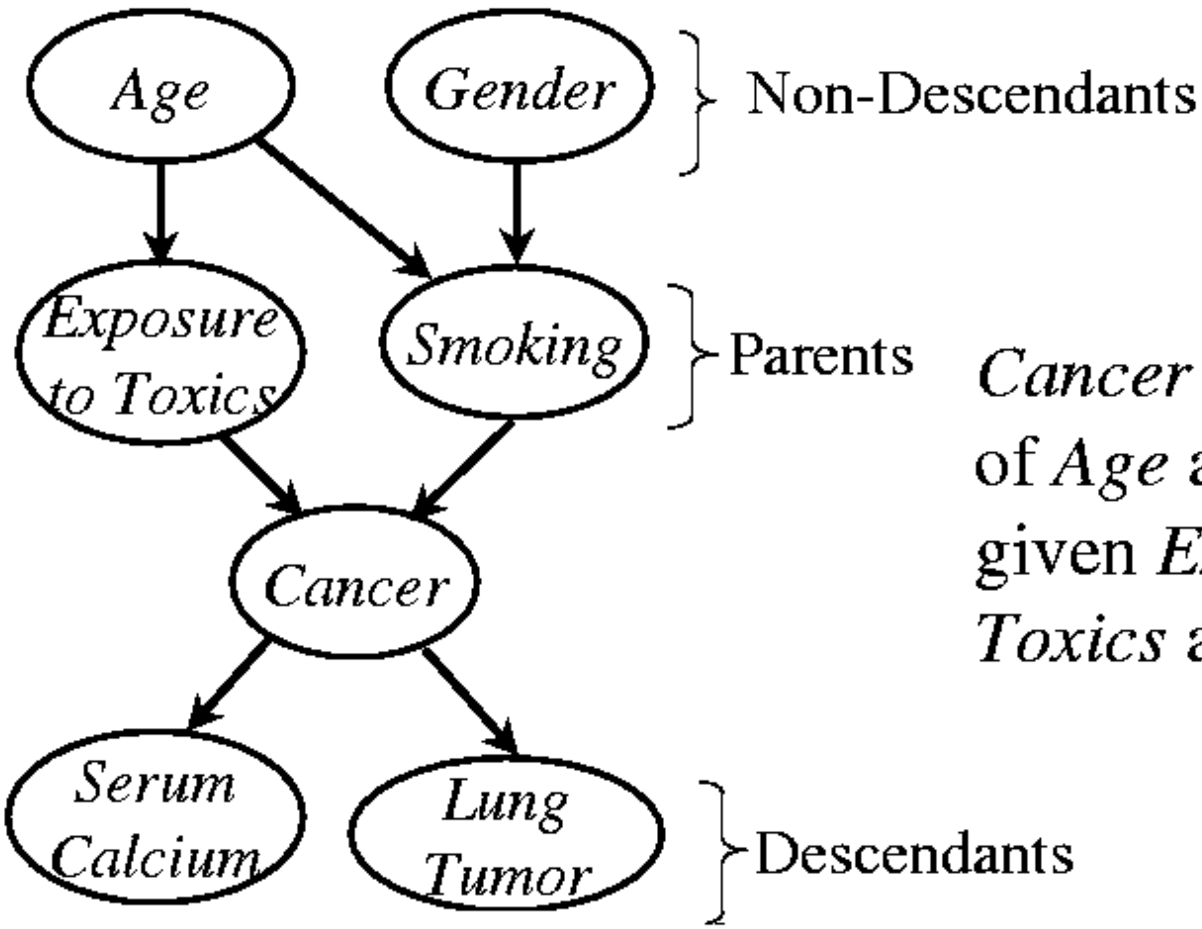


Radio reports about earthquake



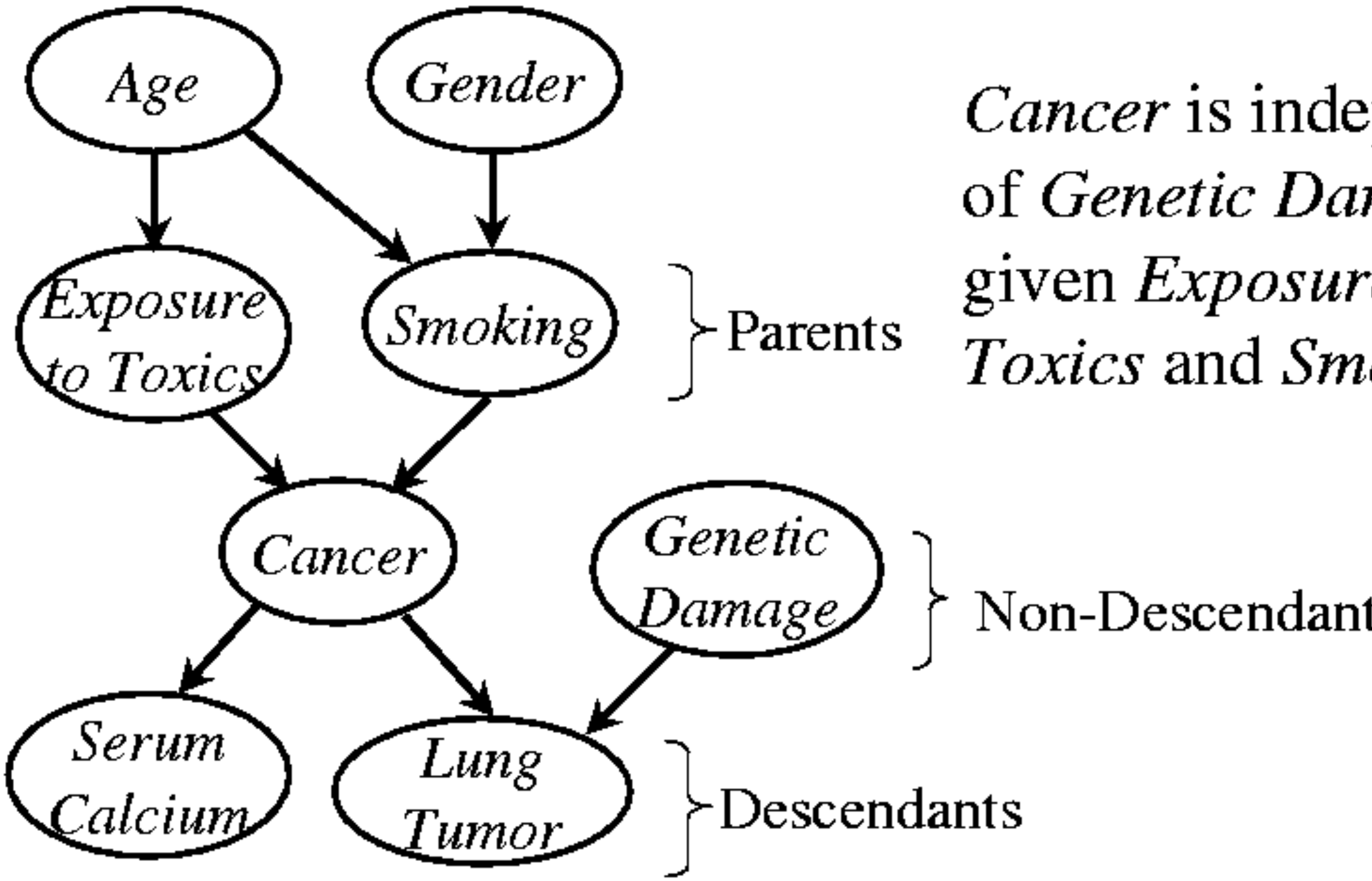
Conditional Independence

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



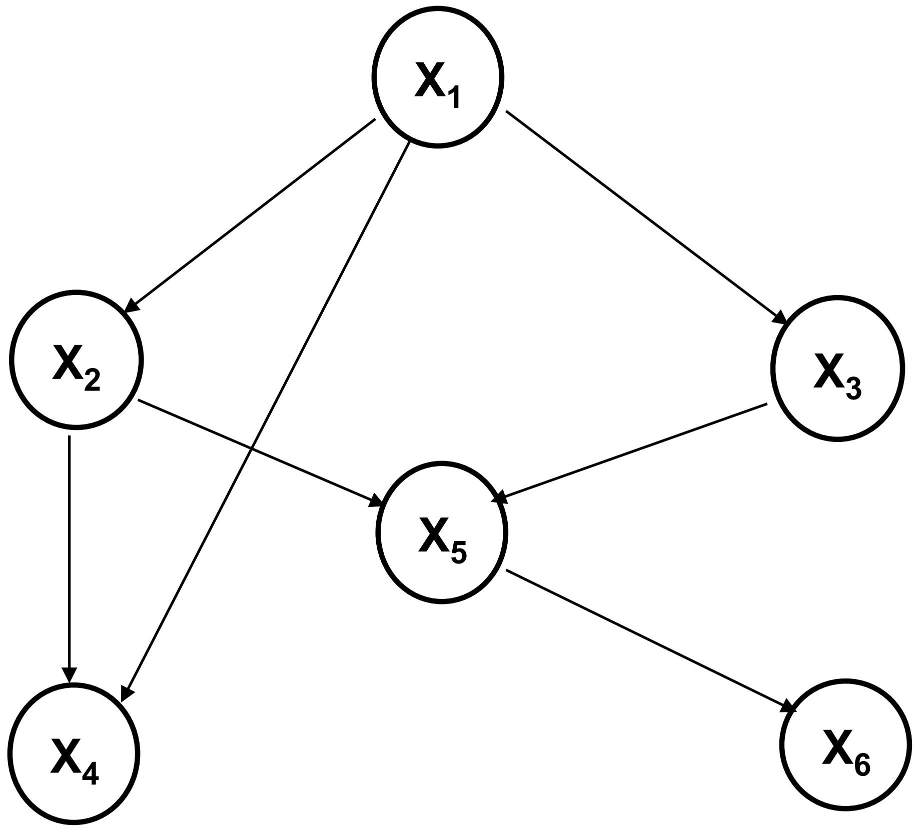
Cancer is independent of *Genetic Damage* given *Exposure to Toxics* and *Smoking*.

General Product (Chain) Rule for Bayesian Networks

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

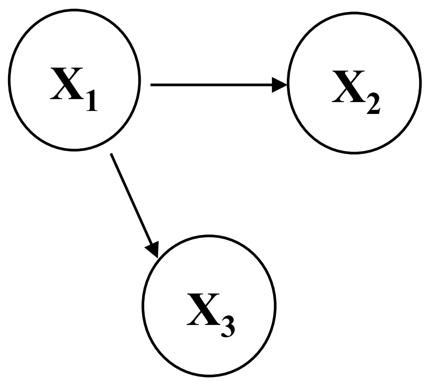
$$\boldsymbol{Pa_i} = parents(X_i)$$

Sample of General Product Rule



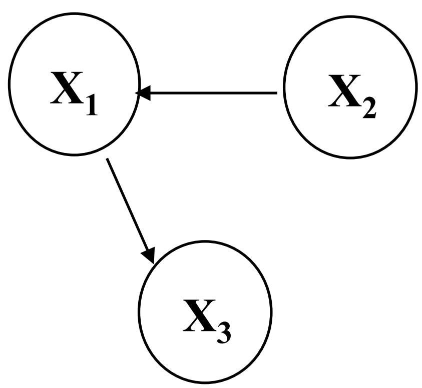
$$p(x_1, x_2, x_3, x_4, x_5, x_6) = p(x_6 \mid x_5)\, p(x_5 \mid x_3, x_2)\, p(x_4 \mid x_2, x_1)\, p(x_3 \mid x_1)\, p(x_2 \mid x_1)\, p(x_1)$$

Arc Reversal - Bayes Rule



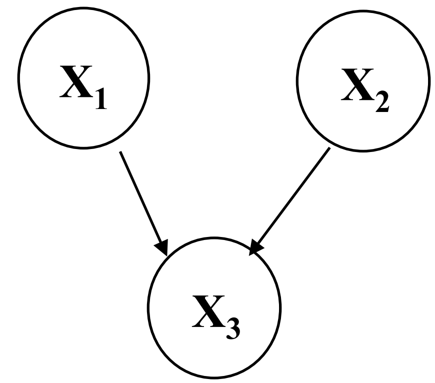
$p(x_1, x_2, x_3) = p(x_3 \mid x_1)\, p(x_2 \mid x_1)\, p(x_1)$

is equivalent to



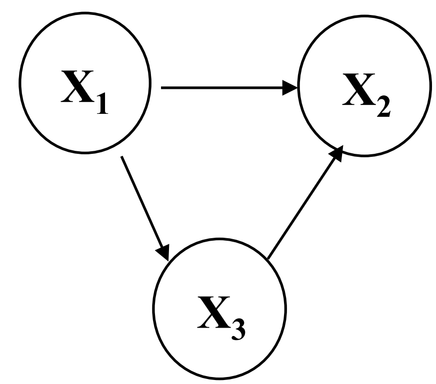
$p(x_1, x_2, x_3) = p(x_3 \mid x_1)\, p(x_2 , x_1)$

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



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

is equivalent to



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

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

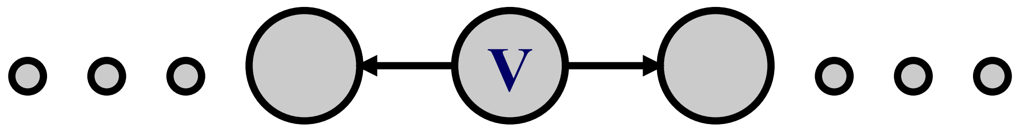
D-Separation of variables

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

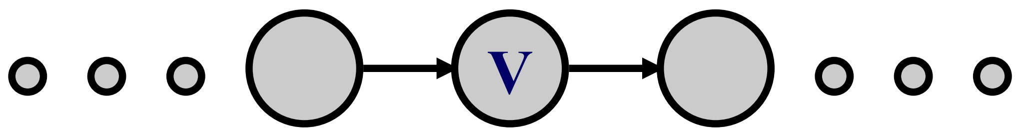
- **Definition: *X* and *Z* are *d-separated* by a set of evidence variables *E* iff every undirected path from *X* to *Z* is "blocked".**
- **A path is "blocked" iff one or more of the following conditions is true: ...**

A path is blocked when:

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



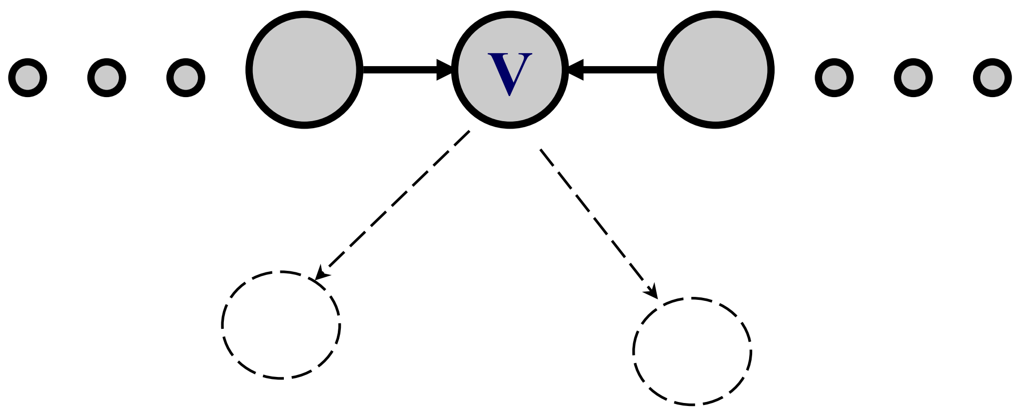
- **Or, there exists a variable *V* on the path such that**
  - **it** is **in the evidence set *E***
  - **the arcs putting *V* in the path are "tail-to-head"**



- **Or, ...**

… a path is blocked when:

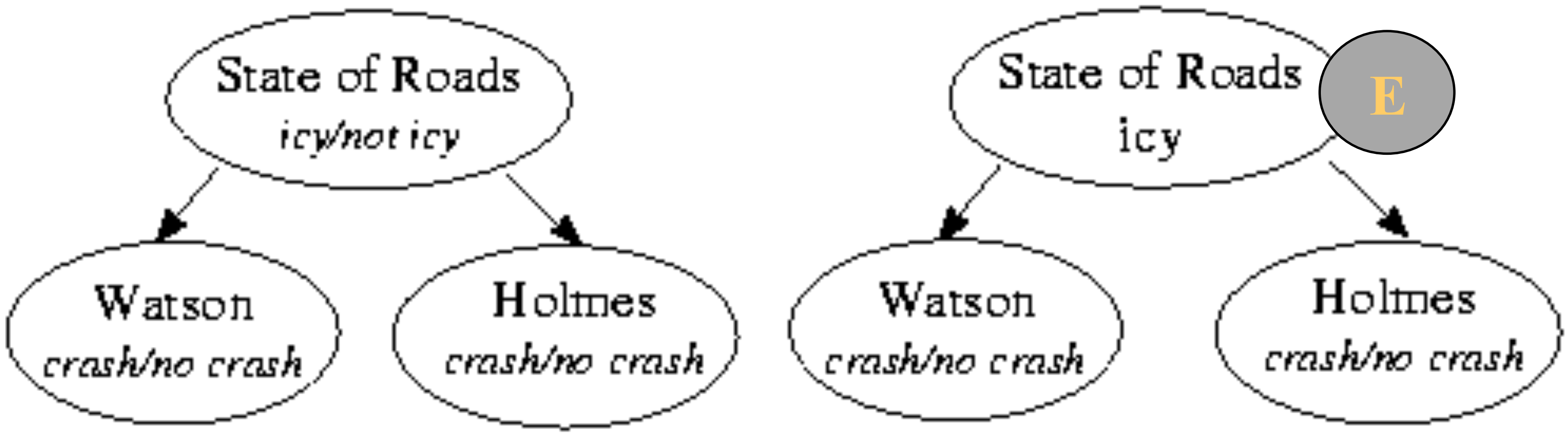
- **… Or, there exists a variable *V* on the path such that**
  - **it** is NOT **in the evidence set *E***
  - neither are any of its descendants
  - **the arcs putting *V* on the path are “head-to-head”**



D-Separation and independence

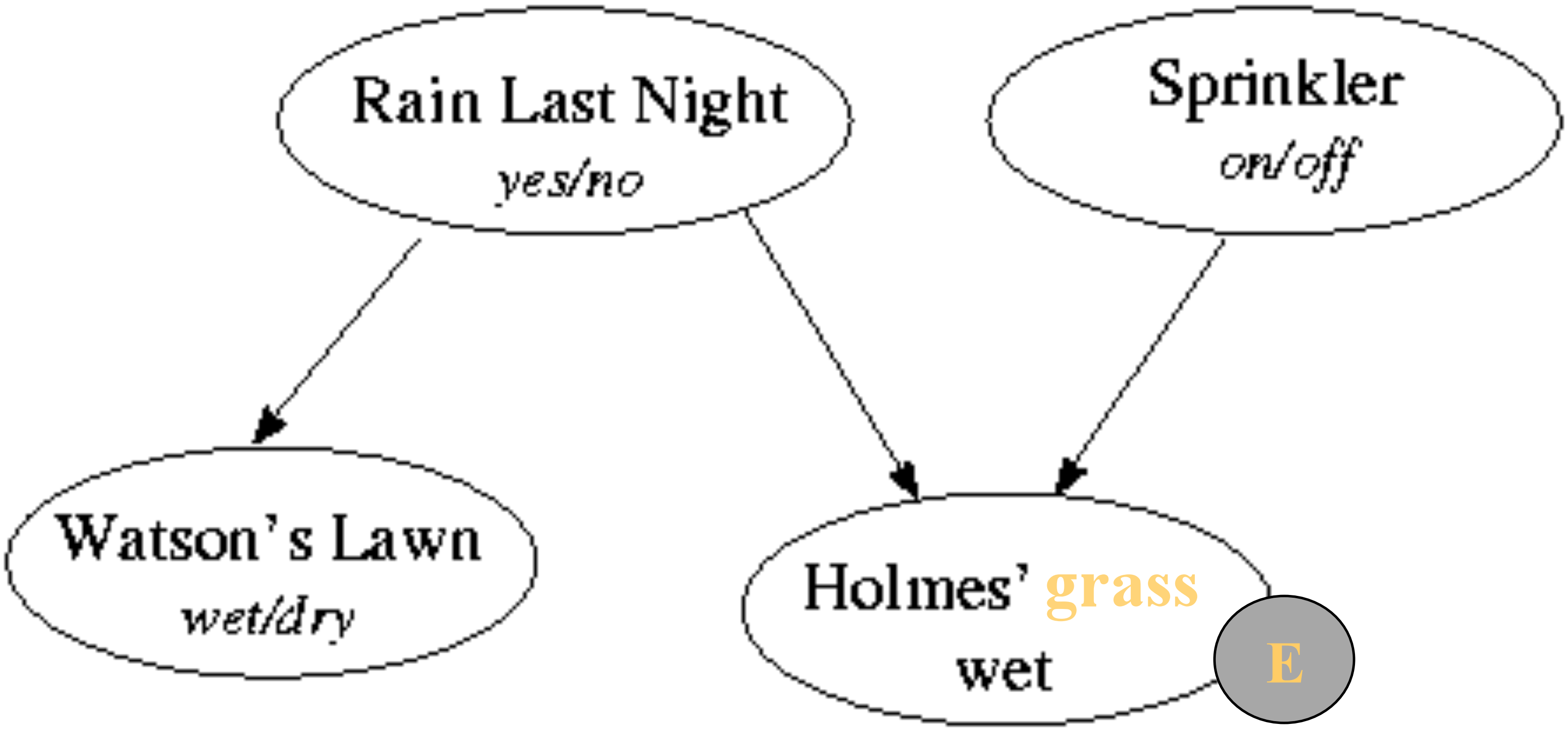
- **Theorem [Verma & Pearl, 1998]:**
  - **If a set of evidence variables *E* d-separates *X* and *Z* in a Bayesian network's graph, then *X* and *Z* will be independent.**
- ***d*-separation can be computed in linear time.**
- **Thus we now have a fast algorithm for automatically inferring whether learning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

Holmes and Watson: "Icy roads" example



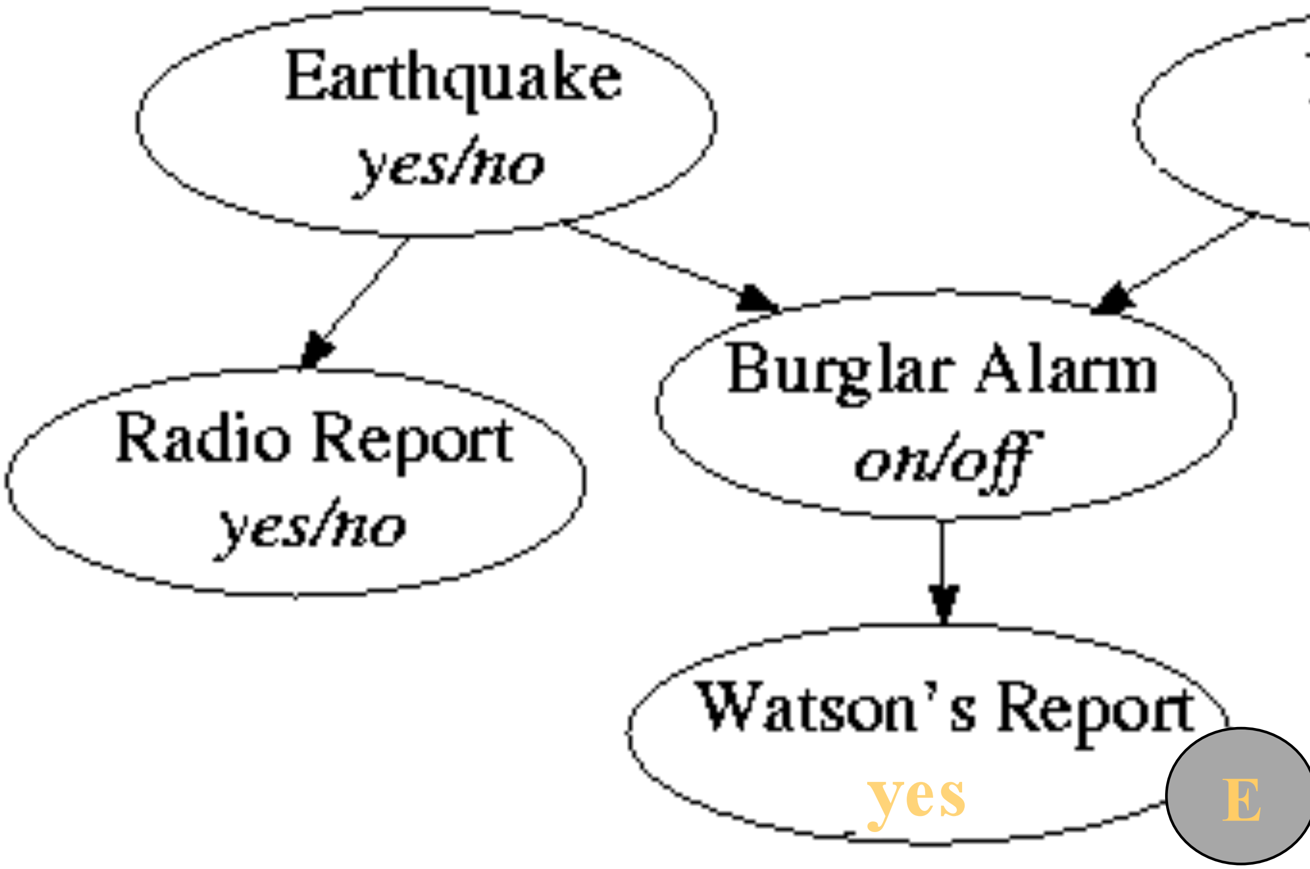
Watson and Holmes are d-connected given no evidence, but
Watson and Holmes are d-separated given State of Roads

Holmes and Watson: "Wet grass" example



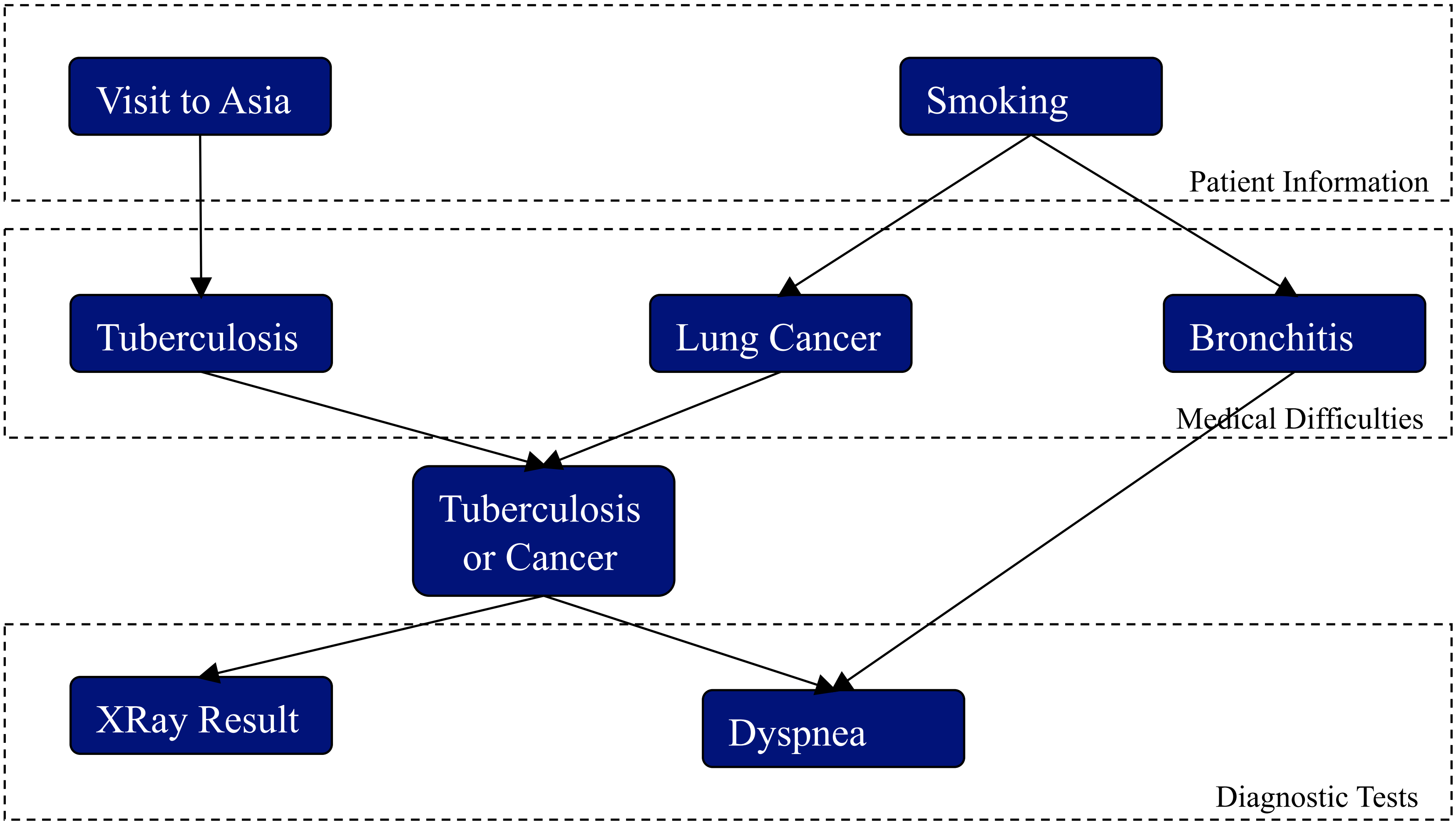
Watson's Lawn and Sprinkler are d-connected given Holmes' grass

Holmes and Watson: “Burglar alarm” example



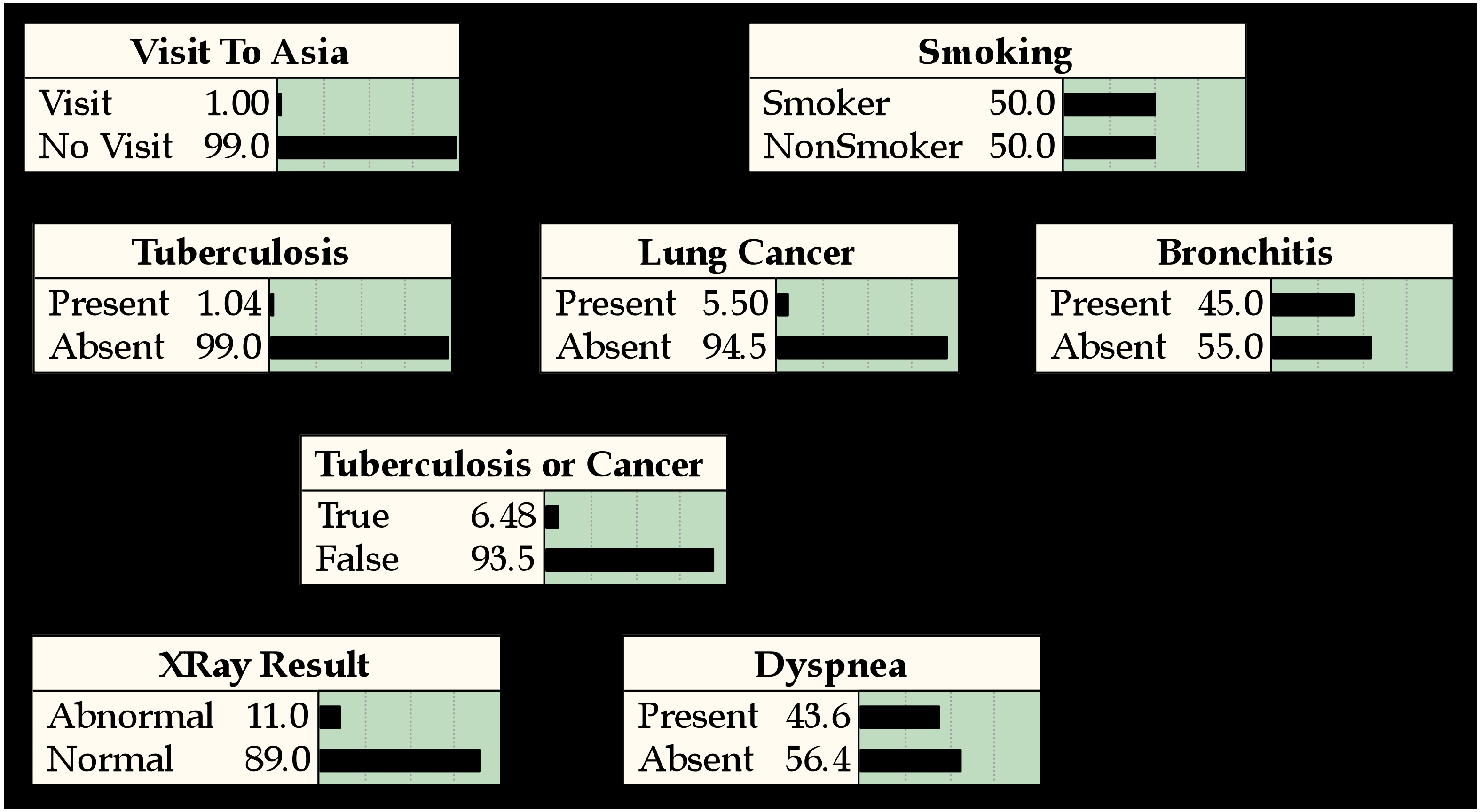
Radio Report and Burglary are d-connected given Watson's Report

Example from Medical Diagnostics



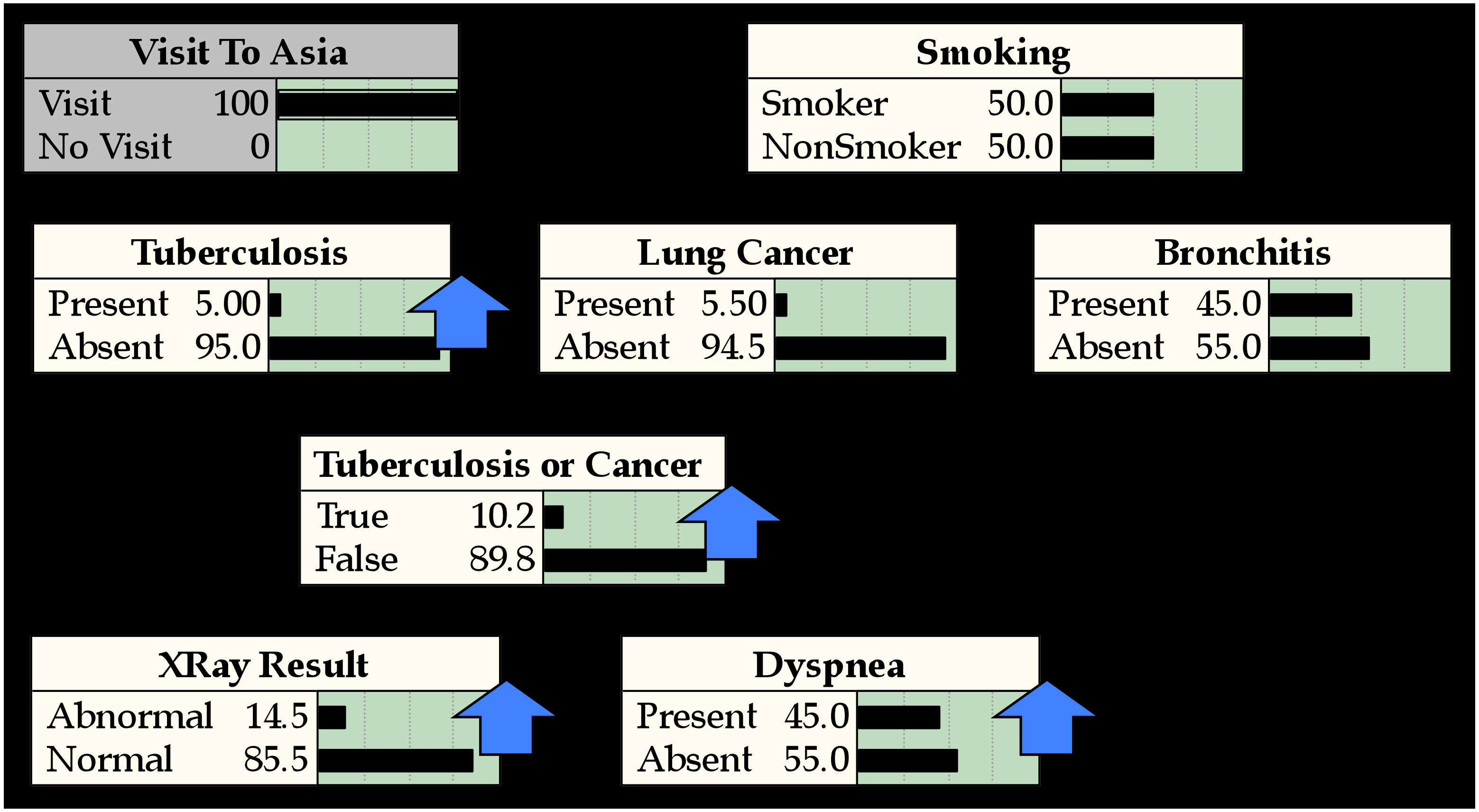
- **Network represents a knowledge structure that models the relationship between medical difficulties, their causes and effects, patient information and diagnostic tests**

Example from Medical Diagnostics



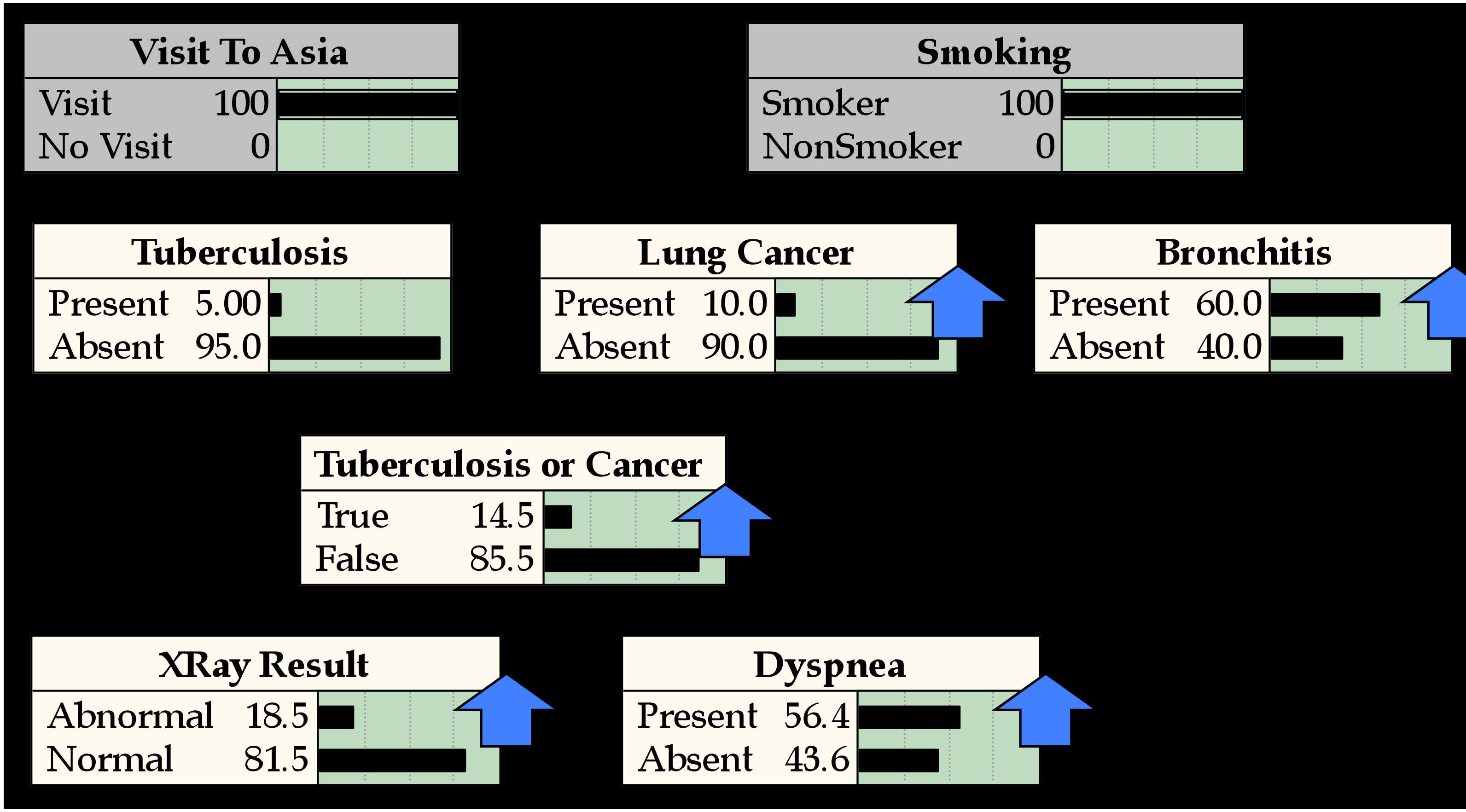
- **Propagation algorithm processes relationship information to provide an unconditional or marginal probability distribution for each node**
- **The unconditional or marginal probability distribution is frequently called the belief function of that node**

Example from Medical Diagnostics



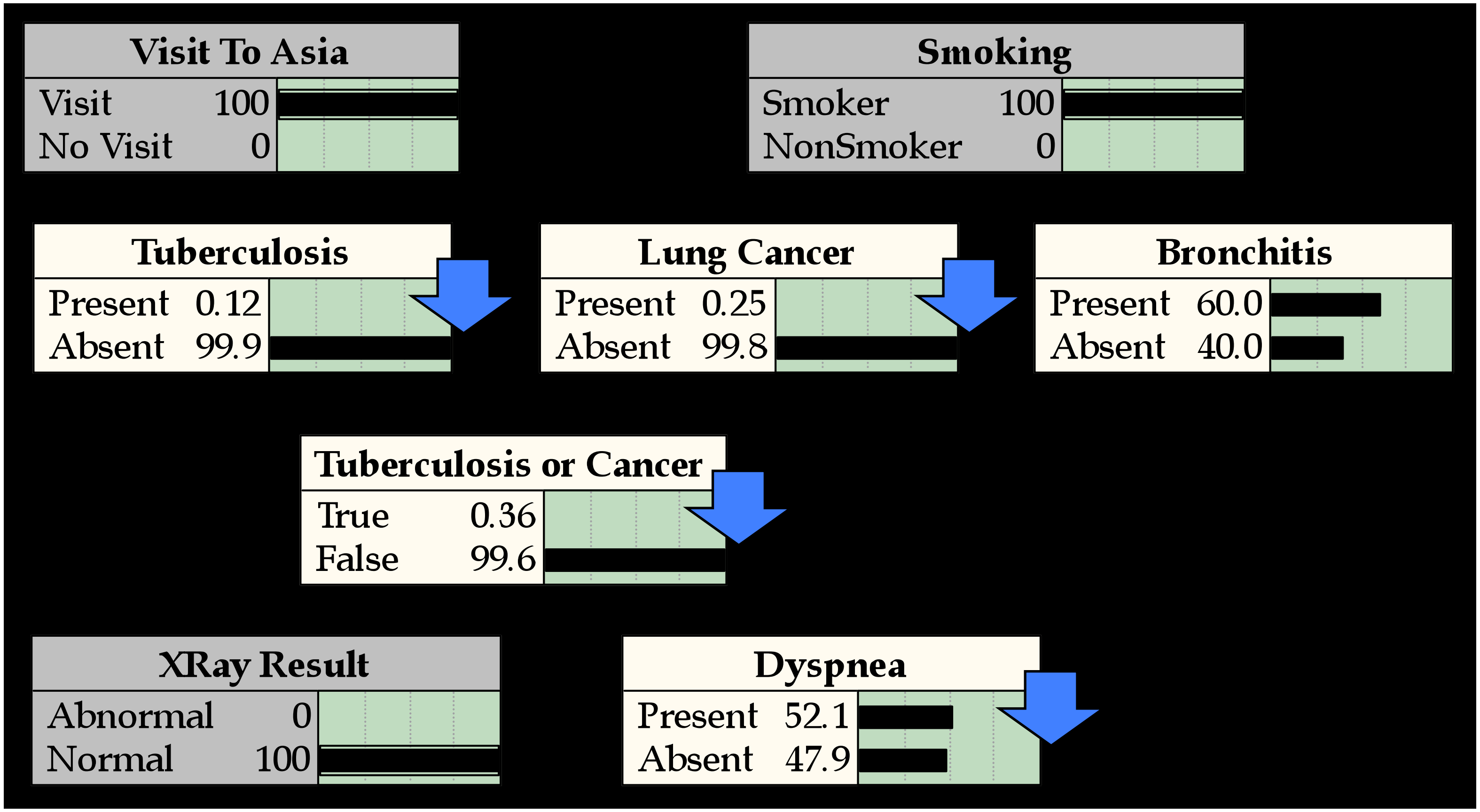
- **As a finding is entered, the propagation algorithm updates the beliefs attached to each relevant node in the network**
- **Interviewing the patient produces the information that "Visit to Asia" is "Visit"**
- **This finding propagates through the network and the belief functions of several nodes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

Example from Medical Diagnostics



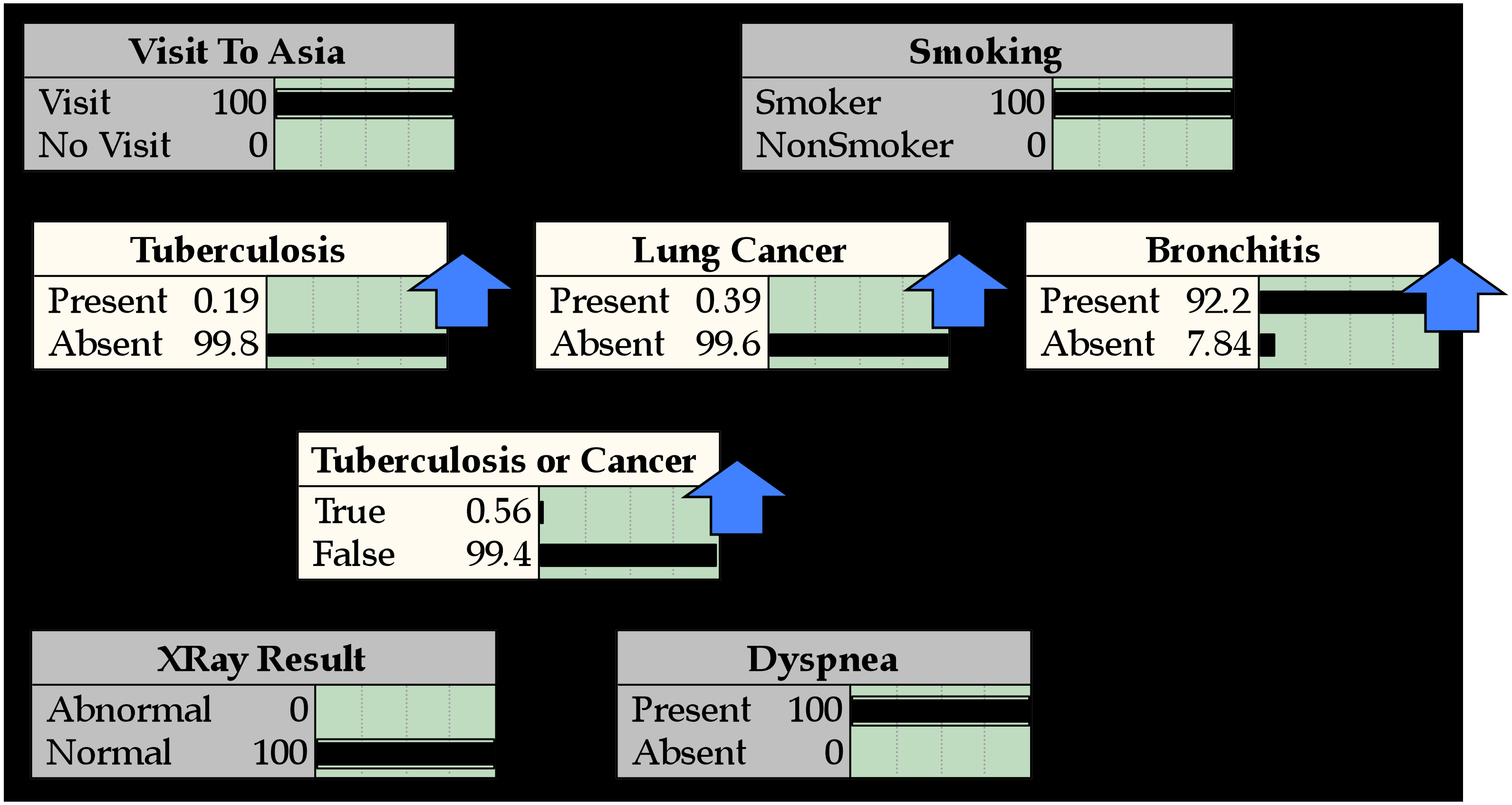
- **Further interviewing of the patient produces the finding “Smoking” is “Smoker”**
- **This information propagates through the network**

Example from Medical Diagnostics



- 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

Example from Medical Diagnostics



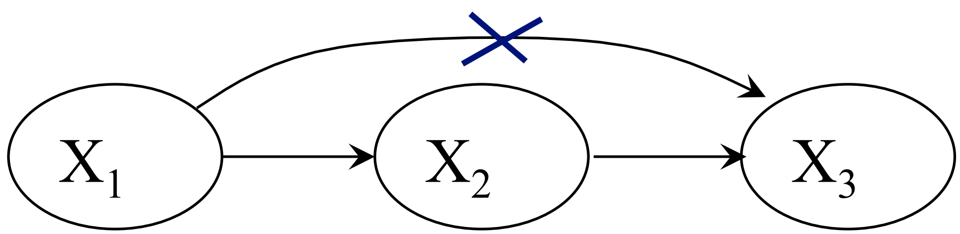
- **The physician also determines that the patient is having difficulty breathing, the finding “Present” is entered for “Dyspnea” and is propagated through the network**
- **The doctor might now conclude that the patient has bronchitis and does not have tuberculosis or lung cancer**

What is a Bayesian Network?

also called belief networks, and (directed acyclic) graphical models

- **Directed acyclic graph**
  - **Nodes are variables (discrete or continuous)**
  - **Arcs indicate dependence between variables.**
- **Conditional Probabilities (local distributions)**

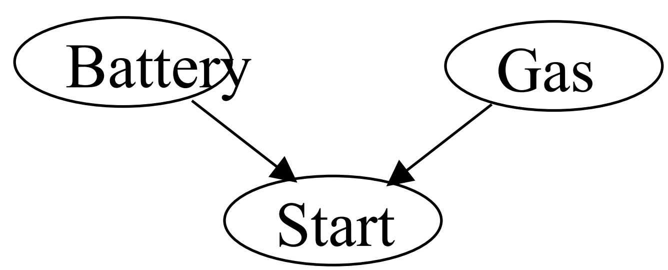
- **Missing arcs implies conditional independence**
- **Independencies + local distributions => modular specification of a joint distribution**



$$p(x_1)\ p(x_2 \mid x_1)\ p(x_3 \mid \cancel{x_1}, x_2) = p(x_1, x_2, x_3)$$

Why Bayesian Networks?

- Expressive language
  - Finite mixture models, Factor analysis, HMM, Kalman filter,…
- Intuitive language
  - Can utilize causal knowledge in constructing models
  - Domain experts comfortable building a network
- General purpose "inference" algorithms
  - P(Bad Battery | Has Gas, Won't Start)



  - Exact: Modular specification leads to large computational efficiencies
  - Approximate: "Loopy" belief propagation