Master Degree in Artificial Intelligence for Science and Technology

Introduction to Bayesian Networks

Fabio Stella
Department of Informatics, Systems and Communication
University of Milan-Bicocca
fabio.stella@unimib.it

OUTLOOK

- SIMPSON'S PARADOX: A CURIOUS CASE
- LADDER OF CAUSATION
- ASIA BAYESIAN NETWORK

- A group of sick patients are given the option to try a new drug (TREATMENT)
- Among those who took the drug (TREATMENT GROUP), a lower percentage recovered (OUTCOME) than among those who did not (CONTROL GROUP)
- However, when we partition by gender (COVARIATE), we see that:
  - *more* men taking the drug recover than do men are not taking the drug, and
  - *more* women taking the drug recover than do women are not taking the drug!

We record the number of recoveries of 700 patients who were given access to the drug.

A total of 350 patients chose to take the drug and 350 patients did not.

The results of the study are shown in the following Table.

	Drug		No Drug	
	patients	recovered	patients	recovered
Men	87	81	270	234
Women	263	192	80	55
Combined data	350	273	350	289

Should a doctor prescribe this drug or not?

Simpson's Paradox

- Named after Edward Simpson (born 1922), statistician
- A group of sick patients are given the option to try a new drug (TREATMENT)
- Among those who took the drug (TREATMENT GROUP), a lower percentage recovered (OUTCOME) than among those who did not (CONTROL GROUP)
- However, when we partition by gender (COVARIATE), we see that:
  - *more* men taking the drug recover than do men are not taking the drug, and
  - *more* women taking the drug recover than do women are not taking the drug!

Simpson's Paradox

- Named after Edward Simpson (born 1922), statistician
- A group of sick patients are given the option to try a new drug (TREATMENT)
- Among those who took the drug (TREATMENT GROUP), a lower percentage recovered (OUTCOME) than among those who did not (CONTROL GROUP)
- However, when we partition by gender (COVARIATE), we see that:
  - *more* men taking the drug recover than do men are not taking the drug, and
  - *more* women taking the drug recover than do women are not taking the drug!

We record the recovery rates of 700 patients who were given access to the drug. A total of 350 patients chose to take the drug and 350 patients did not. The results of the study are shown in **Table 1.1**.

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug			No Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Men	87	81	93%	270	234	87%
Women	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Simpson's Paradox

- Named after Edward Simpson (born 1922), statistician
- A group of sick patients are given the option to try a new drug (TREATMENT)
- Among those who took the drug (TREATMENT GROUP), a lower percentage recovered (OUTCOME) than among those who did not (CONTROL GROUP)
- However, when we partition by gender (COVARIATE), we see that:
  - *more* men taking the drug recover than do men are not taking the drug, and
  - *more* women taking the drug recover than do women are not taking the drug!

We record the recovery rates of 700 patients who were given access to the drug. A total of 350 patients chose to take the drug and 350 patients did not. The results of the study are shown in **Table 1.1**.

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug			No Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Men	87	81	93%	270	234	87%
Women	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Drug vs non-drug takers recovery rates:

93% vs 87% male

Simpson's Paradox

- Named after Edward Simpson (born 1922), statistician
- A group of sick patients are given the option to try a new drug (TREATMENT)
- Among those who took the drug (TREATMENT GROUP), a lower percentage recovered (OUTCOME) than among those who did not (CONTROL GROUP)
- However, when we partition by gender (COVARIATE), we see that:
  - *more* men taking the drug recover than do men are not taking the drug, and
  - *more* women taking the drug recover than do women are not taking the drug!

We record the recovery rates of 700 patients who were given access to the drug. A total of 350 patients chose to take the drug and 350 patients did not. The results of the study are shown in **Table 1.1**.

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug			No Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Men	87	81	93%	270	234	87%
Women	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Drug vs non-drug takers recovery rates:

73% vs 69% female

Simpson's Paradox

- Named after Edward Simpson (born 1922), statistician
- A group of sick patients are given the option to try a new drug (TREATMENT)
- Among those who took the drug (TREATMENT GROUP), a lower percentage recovered (OUTCOME) than among those who did not (CONTROL GROUP)
- However, when we partition by gender (COVARIATE), we see that:
  - *more* men taking the drug recover than do men are not taking the drug, and
  - *more* women taking the drug recover than do women are not taking the drug!

We record the recovery rates of 700 patients who were given access to the drug. A total of 350 patients chose to take the drug and 350 patients did not. The results of the study are shown in **Table 1.1**.

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug			No Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Men	87	81	93%	270	234	87%
Women	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Drug vs non-drug takers recovery rates:

78% vs 83% general population!

Simpson's Paradox

- Named after Edward Simpson (born 1922), statistician
- A group of sick patients are given the option to try a new drug (TREATMENT)
- Among those who took the drug (TREATMENT GROUP), a lower percentage recovered (OUTCOME) than among those who did not (CONTROL GROUP)
- However, when we partition by gender (COVARIATE), we see that:
  - *more* men taking the drug recover than do men are not taking the drug, and
  - *more* women taking the drug recover than do women are not taking the drug!

We record the recovery rates of 700 patients who were given access to the drug. A total of 350 patients chose to take the drug and 350 patients did not. The results of the study are shown in **Table 1.1**.

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug			**No Drug**		
	patients	recovered	% recovered	patients	recovered	% recovered
Men	87	81	93%	270	234	87%
Women	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Drug vs non-drug takers recovery rates:

93% vs 87% male

73% vs 69% female

78% vs 83% general population!

Should a doctor prescribe the drug; to whom?

Should a policy maker approve the drug for use?

Understand the causal story behind the data

- What mechanism generated the data?

Suppose estrogen has a negative effect on recovery

- women less likely to recover than men, regardless of the drug

From the data:

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug			No Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Men	87	81	93%	270	234	87%
Women	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Conclusion: the drug appears to be harmful but it is not

If we select a drug taker at random, that person is more likely to be a woman
Hence less likely to recover than a random person who doesn't take the drug

Causal Story

Being a woman is a common cause of both drug taking and failure to recover.
To assess the effectiveness we need to compare subjects of the same gender.
(Ensures that any difference in recovery rates is not ascribable to estrogen)

Data Segregation

- We have solved the problem using gender-segregated data
- Then let's just segregate the data whenever possible, right?

WRONG!!!

- Consider a drug affecting recovery by lowering blood pressure (BP)
- Unfortunately, it has also a toxic effect

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No Drug			Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Low BP	87	81	93%	270	234	87%
High BP	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Should a doctor prescribe this drug or not?

Data Segregation

- We have solved the problem using gender-segregated data
- Then let's just segregate the data whenever possible, right?

WRONG!!!

- Consider a drug affecting recovery by lowering blood pressure (BP)
- Unfortunately, it has also a toxic effect

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No Drug			Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Low BP	87	81	93%	270	234	87%
High BP	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Should a doctor prescribe this drug or not?

Once again, the answer follows from the way the data were generated.

Data Segregation

- We have solved the problem using gender-segregated data
- Then let's just segregate the data whenever possible, right?

WRONG!!!

- Consider a drug affecting recovery by lowering blood pressure (BP)
- Unfortunately, it has also a toxic effect

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No Drug			Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Low BP	87	81	93%	270	234	87%
High BP	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Should a doctor prescribe this drug or not?

Once again, the answer follows from the way the data were generated.

In the general population, the drug might improve recovery rates because of its effect on blood pressure.

Data Segregation

- We have solved the problem using gender-segregated data
- Then let's just segregate the data whenever possible, right?

WRONG!!!

- Consider a drug affecting recovery by lowering blood pressure (BP)
- Unfortunately, it has also a toxic effect

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No Drug			Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Low BP	87	81	93%	270	234	87%
High BP	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Should a doctor prescribe this drug or not?

Once again, the answer follows from the way the data were generated.

In the general population, the drug might improve recovery rates because of its effect on blood pressure.

in subpopulations—the group of people whose posttreatment BP is high and the group whose posttreatment BP is low—we, of course, would not see that effect; we would only see the drug's toxic effect.

Data Segregation

- We have solved the problem using gender-segregated data
- Then let's just segregate the data whenever possible, right?

WRONG!!!

- Consider a drug affecting recovery by lowering blood pressure (BP)
- Unfortunately, it has also a toxic effect

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No Drug			Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Low BP	87	81	93%	270	234	87%
High BP	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Should a doctor prescribe this drug or not?

Once again, the answer follows from the way the data were generated.

In the general population, the drug might improve recovery rates because of its effect on blood pressure.

in subpopulations—the group of people whose posttreatment BP is high and the group whose posttreatment BP is low—we, of course, would not see that effect; we would only see the drug's toxic effect.

Data Segregation

- We have solved the problem using gender-segregated data
- Then let's just segregate the data whenever possible, right?

WRONG!!!

- Consider a drug affecting recovery by lowering blood pressure (BP)
- Unfortunately, it has also a toxic effect

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No Drug			Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Low BP	87	81	93%	270	234	87%
High BP	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Should a doctor prescribe this drug or not? **YES**

Only by BP-segregating the data we can see the toxic effect

It makes no sense to segregate the data; we should use the combined data

Data Segregation

- We have solved the problem using gender-segregated data
- Then let's just segregate the data whenever possible, right?

WRONG!!!

- Consider a drug affecting recovery by lowering blood pressure (BP)
- Unfortunately, it has also a toxic effect

Table 1.1 Results of a study into a new drug, with gender being taken into account

	Drug			No Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Men	87	81	93%	270	234	87%
Women	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Table 1.2 Results of a study into a new drug, with posttreatment blood pressure taken into account

	No Drug			Drug		
	patients	recovered	% recovered	patients	recovered	% recovered
Low BP	87	81	93%	270	234	87%
High BP	263	192	73%	80	55	69%
Combined data	350	273	78%	350	289	83%

Note that the data are the same of Simpson's paradox.

Lessons Learned

- Information that allowed us to make a correct decision

the timing of the measurements

that the treatment affects blood pressure

that blood pressure affects recovery

- All this information was not in the data

as statisticians rightly say, **CORRELATION IS NOT CAUSATION**

hence there is no method that can determine the causal story from data alone

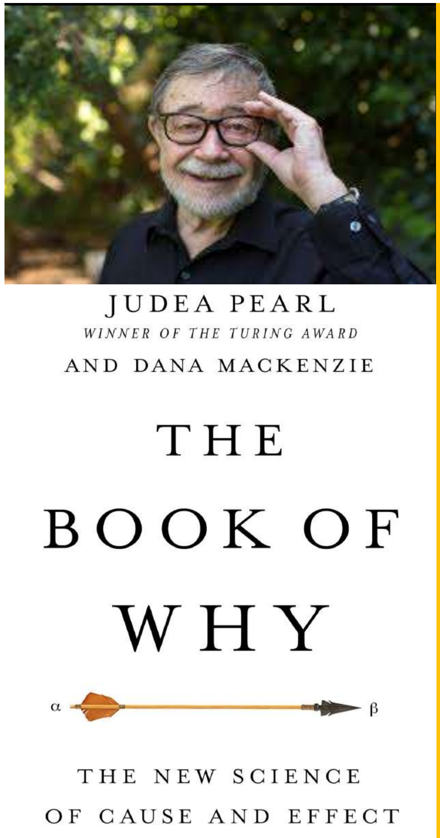
whence no ML method can aid in our decision

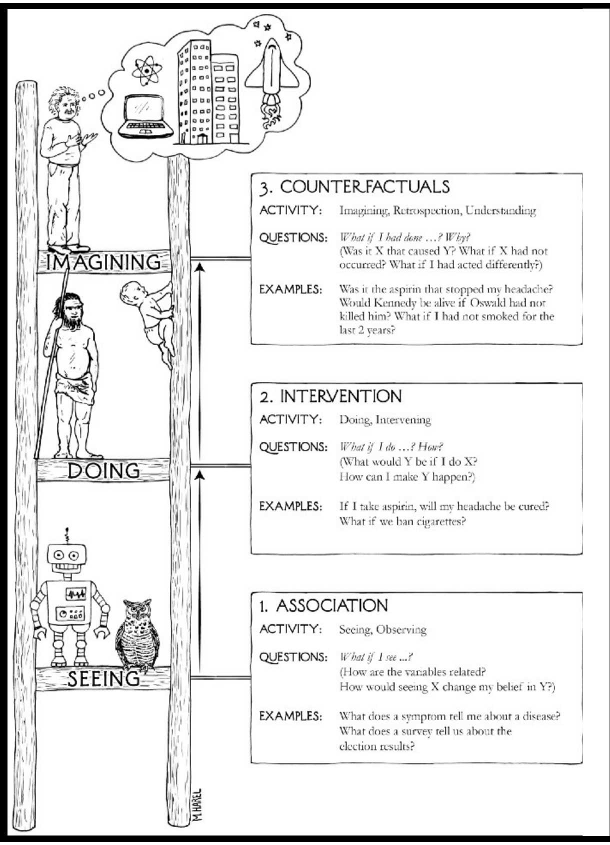
- The same holds for Simpson's paradox

the paradox arises out of our conviction that treatment cannot affect sex

if it could, we could explain it as in our blood pressure case

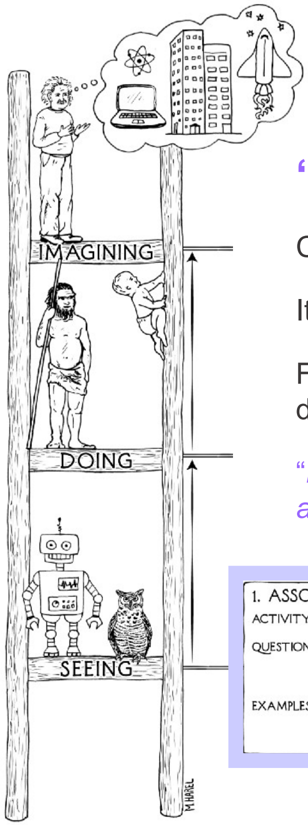
but we cannot test the assumption using the data





The Ladder of Causation

Seeing; we are looking for regularities in observations.



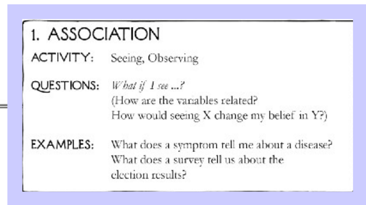
"What if I see …?"

Calls for predictions based on passive observations.

It is characterized by the question "*What if I see …?*"

For instance, imagine a marketing director at a department store who asks,

"How likely is a customer who bought toothpaste to also buy dental floss?"



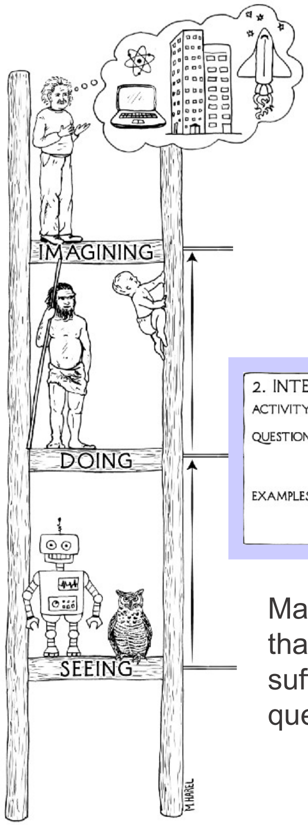
1. ASSOCIATION

ACTIVITY: Seeing, Observing

QUESTIONS: *What if I see …?*
(How are the variables related?
How would seeing X change my belief in Y?)

EXAMPLES: What does a symptom tell me about a disease?
What does a survey tell us about the election results?

Intervention; ranks higher than association because it involves not just seeing but changing what is.



"*What if do …?*" & "*How?*"

We step up to the next level of causal queries when we begin to change the world. A typical question for this level is

"What will happen to our floss sales if we double the price of toothpaste?"

2. INTERVENTION

ACTIVITY: Doing, Intervening

QUESTIONS: *What if I do …? How?*
(What would Y be if I do X?
How can I make Y happen?)

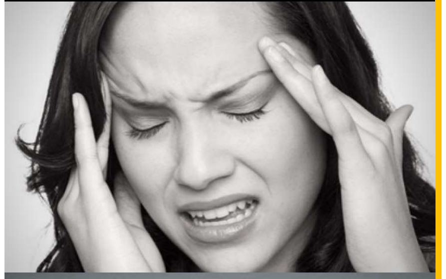
EXAMPLES: If I take aspirin, will my headache be cured?
What if we ban cigarettes?

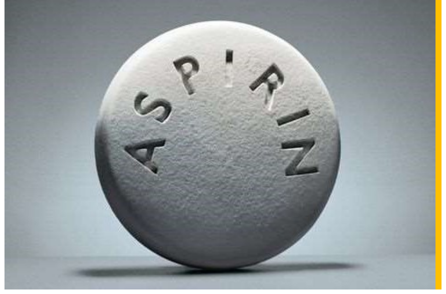
This already calls for a new kind of knowledge, absent from the data, which we find at rung two of the Ladder of Causation, **Intervention**.

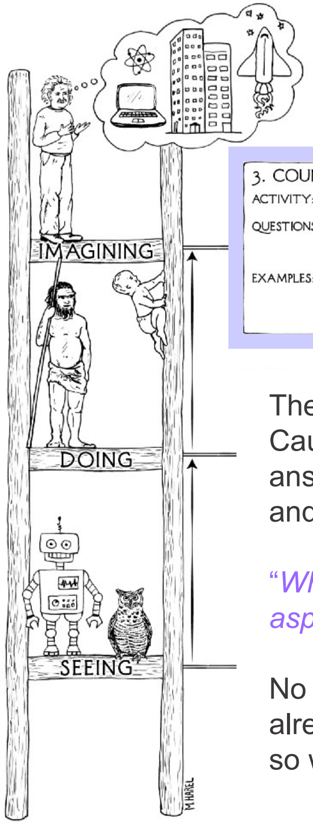
Many scientists have been quite traumatized to learn that none of the methods they learned in statistics is sufficient even to articulate, let alone answer, a simple question like

"What happens if we double the price?"

Counterfactuals; ranks higher than intervention because it involves **imagining**, **retrospection** and **understanding**.







"What if I had done …?" & "Why?"

3. COUNTERFACTUALS

ACTIVITY: Imagining, Retrospection, Understanding

QUESTIONS: *What if I had done …? Why?* (Was it X that caused Y? What if X had not occurred? What if I had acted differently?)

EXAMPLES: Was it the aspirin that stopped my headache? Would Kennedy be alive if Oswald had not killed him? What if I had not smoked for the last 2 years?

We might wonder, My headache is gone now, but

- *Why?*
- *Was it the aspirin I took?*
- *The food I ate?*
- *The good news I heard?*

These queries take us to the top rung of the Ladder of Causation, the level of **Counterfactuals**, because to answer them we must go back in time, change history, and ask,

"What would have happened if I had not taken the aspirin?"

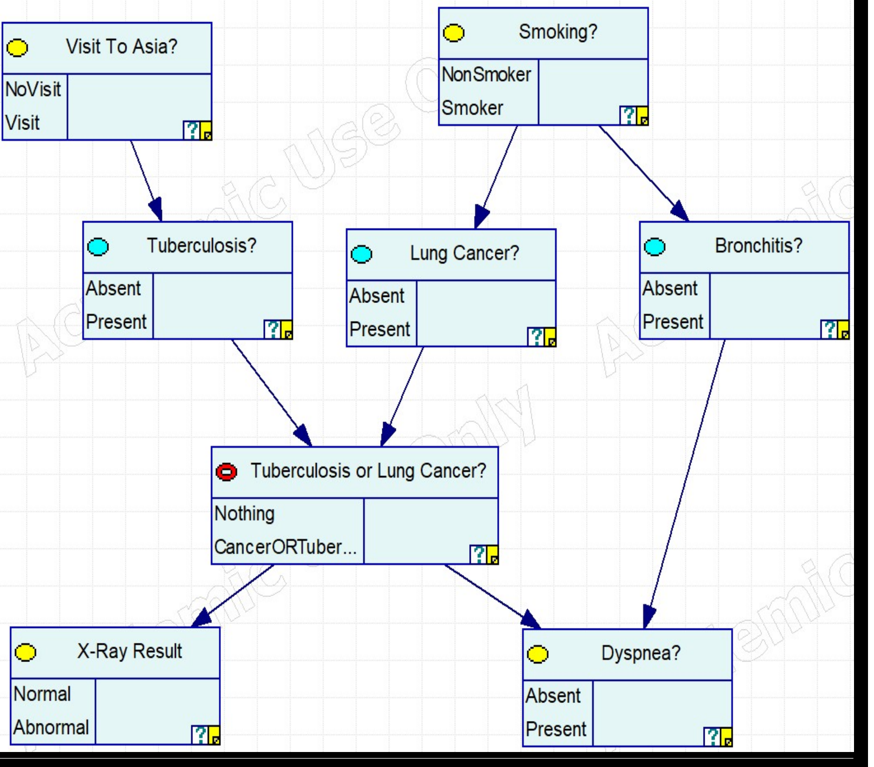
No experiment in the world can deny treatment to an already treated person and compare the two outcomes, so we must import a whole new kind of knowledge.

Node: is associated with a variable

- the question «**Visit to Asia?**» has two possible answers, i.e., «*NoVisit*» and «*Visit*»
- the question «**Dyspnea?**» has two possible answers, i.e., «*Absent*» and «*Present*»

Arc:

- in a Bayesian network the arc pointing from node «**Smoking**» to node «**Bronchitis**» means that «**Smoking**» influences «**Bronchitis**»
- in a causal Bayesian network the arc pointing from node «**Smoking**» to node «**Bronchitis**» means that «**Smoking**» directly causes «**Bronchitis**»

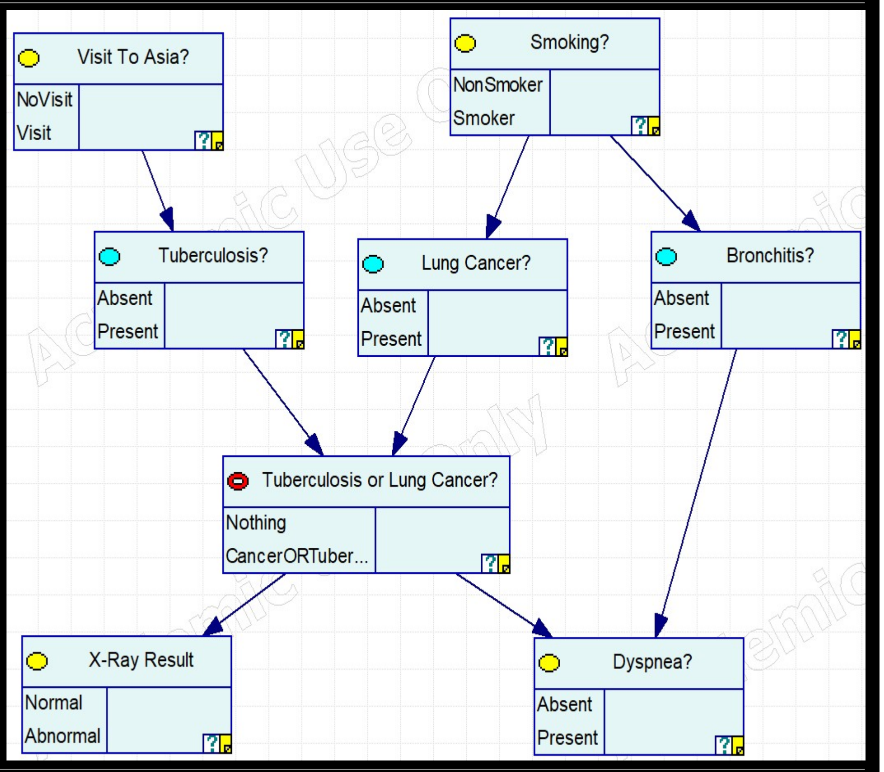


Node: is associated with a variable

- the question «**Visit to Asia?**» has two possible answers, i.e., «*NoVisit*» and «*Visit*»
- the question «**Dyspnea?**» has two possible answers, i.e., «*Absent*» and «*Present*»

Arc:

- in a Bayesian network the arc pointing from node «**Smoking**» to node «**Bronchitis**» means that «**Smoking**» influences «**Bronchitis**»
- in a causal Bayesian network the arc pointing from node «**Smoking**» to node «**Bronchitis**» means that «**Smoking**» directly causes «**Bronchitis**»

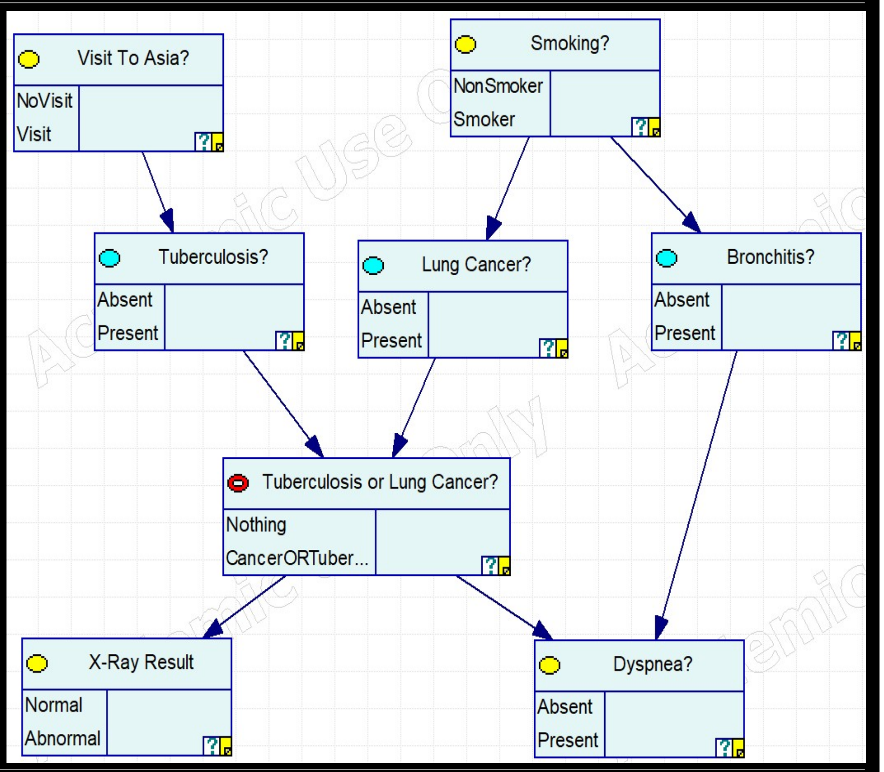


Parent of a node:

- «**Tubercolosis?**» and «**Lung Cancer?**» are parents of «**Tubercolosis or Cancer?**»
- «**Visit to Asia?**» is the parent node of «**Tubercolosis?**».
  In a causal Bayesian network we say that «**Visit to Asia?**» is a cause of «**Tubercolosis?**».

Children of a node:

- «**Tubercolosis?**» is a child node of «**Visit to Asia?**»
- «**Dyspnea?**» is a child node of «**Bronchitis?**».
  In a causal Bayesian network we say that «**Dyspnea?**» is an effect of «**Bronchitis?**».

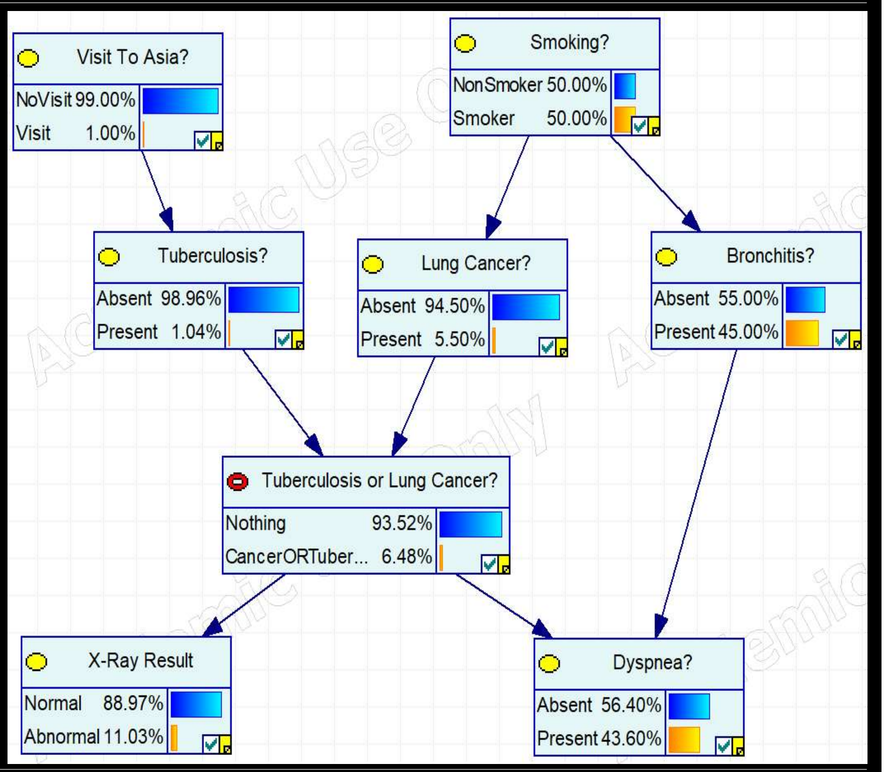


Probability: for each node (variable)

- «**Visit to Asia?**» has probability 0.99 to be in state «*NoVisit*» and probability 0.01 to be in state «*Visit*»
- «**Dyspnea?**» has probability 0.564 to be in state «*Absent*» and probability 0.436 to be in state «*Present*»

Conditional Probablity Table

- for each possibiile value of «**Visit to Asia?**», i.e., «*NoVisit*» or «*Visit*», we specify a probability value for each possible value of «**Tubercolosis?**», i.e., «*Absent*» and «*Present*»



Evidence: information available on a subset of variables

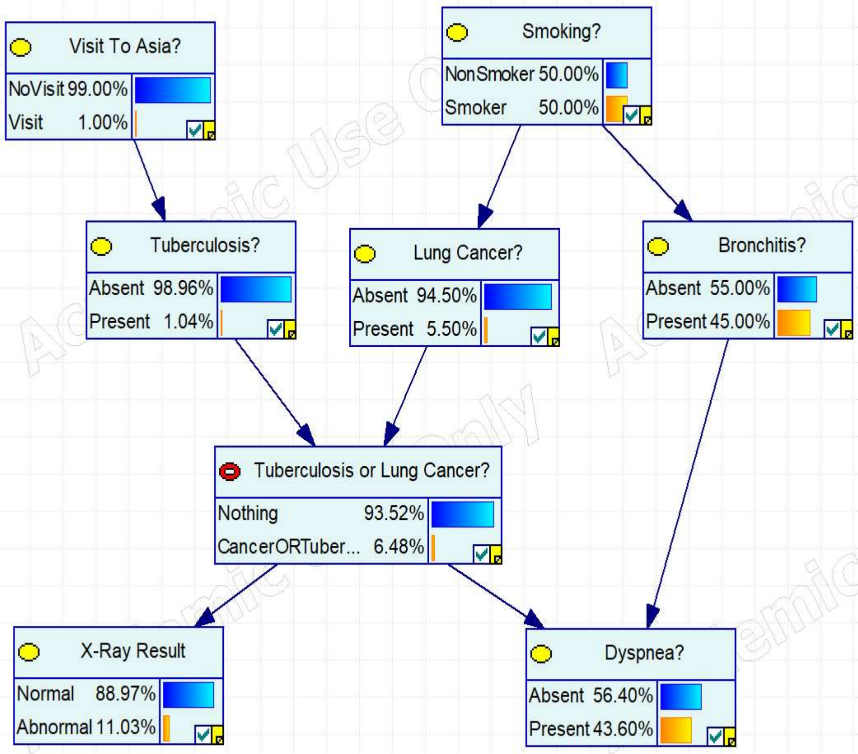
- the patient tells she suffers dyspnea

  «**Dyspnea**?=*Present*»

Query: we want to know how likely is the patient to have lung cancer

P(**Lung Cancer?** = *Present*) = ?

Evidence Propagation: evidence is propagated from the node **Dyspnea**? to the remaining nodes



Evidence: information available on a subset of variables

- the patient tells she suffers dyspnea

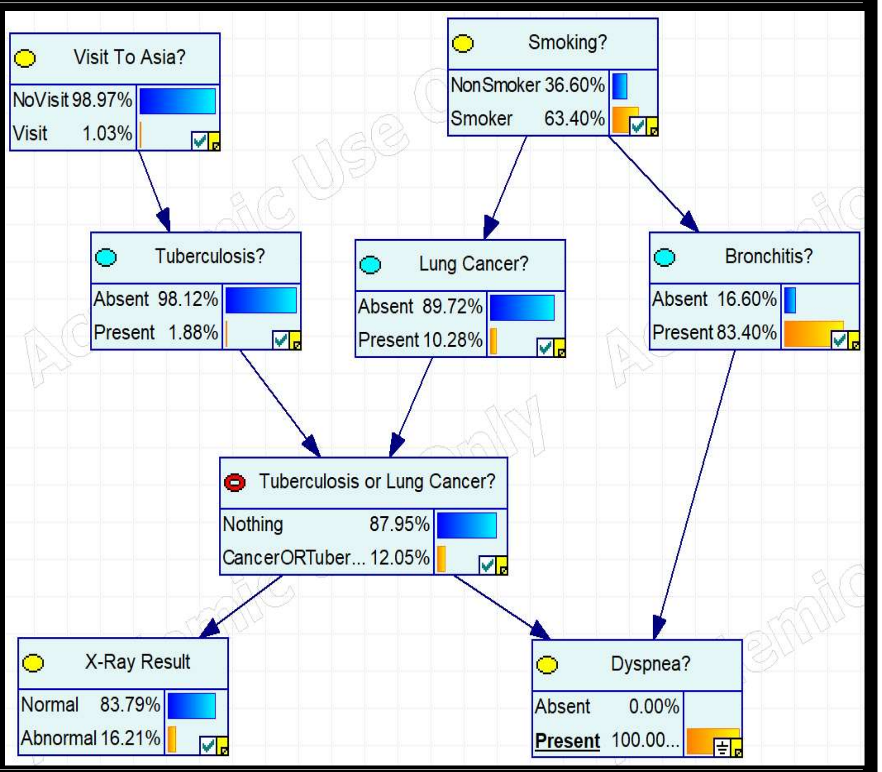
  «**Dyspnea**?=*Present*»

Query: we want to know how likely is the patient to have lung cancer

P(**Lung Cancer?** = *Present*) = ?

Evidence Propagation: evidence is propagated from the node **Dyspnea**? to the remaining nodes

P(**Lung Cancer?** = *Present*) = 0.1028

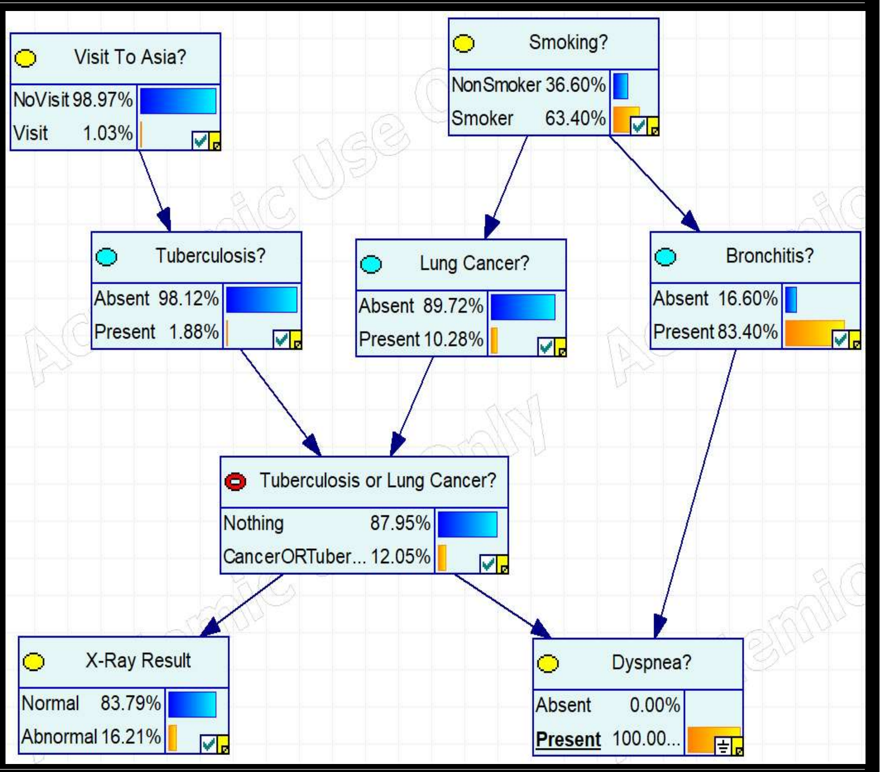


Evidence: we prescribe X-Ray to the patient and we got the following outcome

«**X-Ray Result**=*Abnormal*»

Query: we want to know how likely is the patient to have lung cancer

P(**Lung Cancer?** = *Present*) = ?



Evidence: we prescribe X-Ray to the patient and we got the following outcome

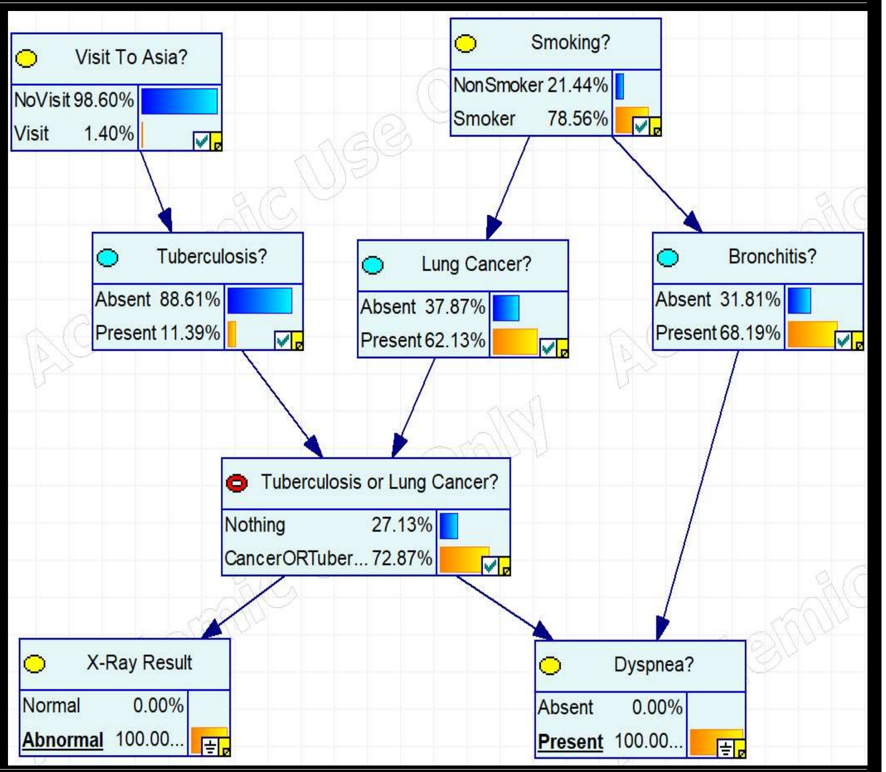
«**X-Ray Result**=*Abnormal*»

Query: we want to know how likely is the patient to have lung cancer

P(**Lung Cancer?** = *Present*) = ?

Evidence Propagation: evidence is propagated from the node **Dyspnea**? to the remaining nodes

P(**Lung Cancer?** = *Present*) = 0.6213

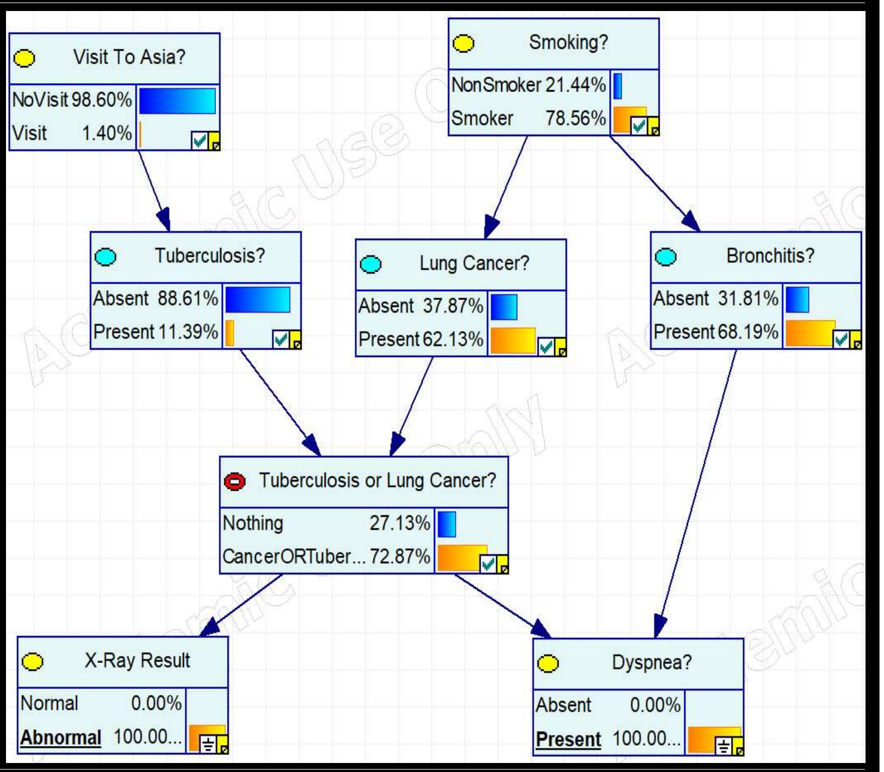


Evidence: you ask the patient whether she is a smoker or not and she replies that she never smoked

«**Smoking?**=*NonSmoker*»

Query: we want to know how likely is the patient to have lung cancer

P(**Lung Cancer?** = *Present*) = ?



Evidence: you ask the patient whether she is a smoker or not and she replies that she never smoked

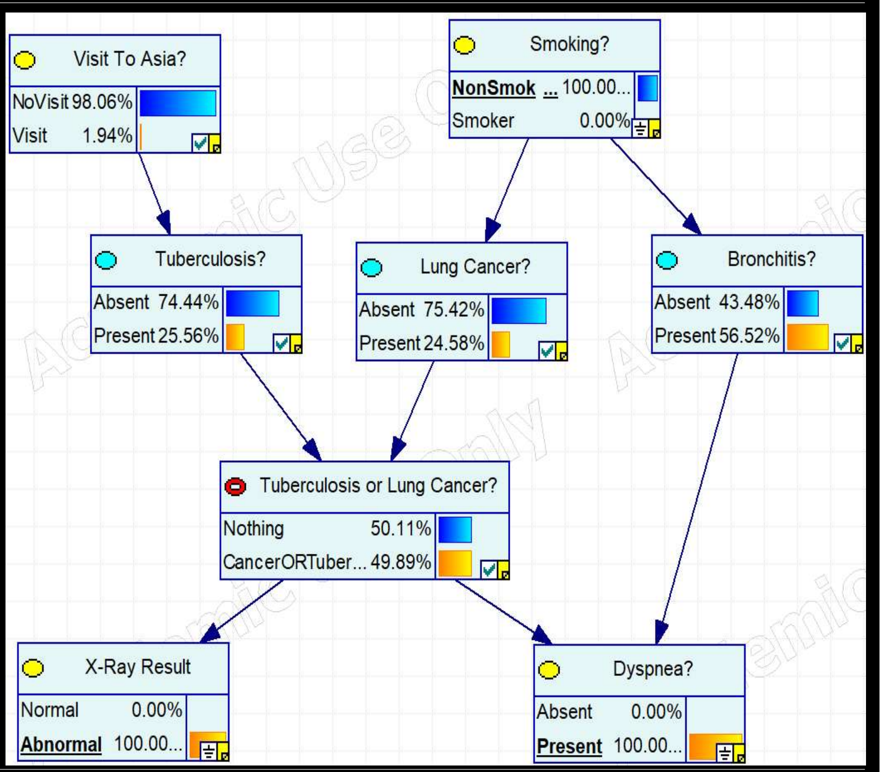
«**Smoking?**=*NonSmoker*»

Query: we want to know how likely is the patient to have lung cancer

P(**Lung Cancer?** = *Present*) = ?

Evidence Propagation: evidence is propagated from the node **Dyspnea**? to the remaining nodes

P(**Lung Cancer?** = *Present*) = 0.2458

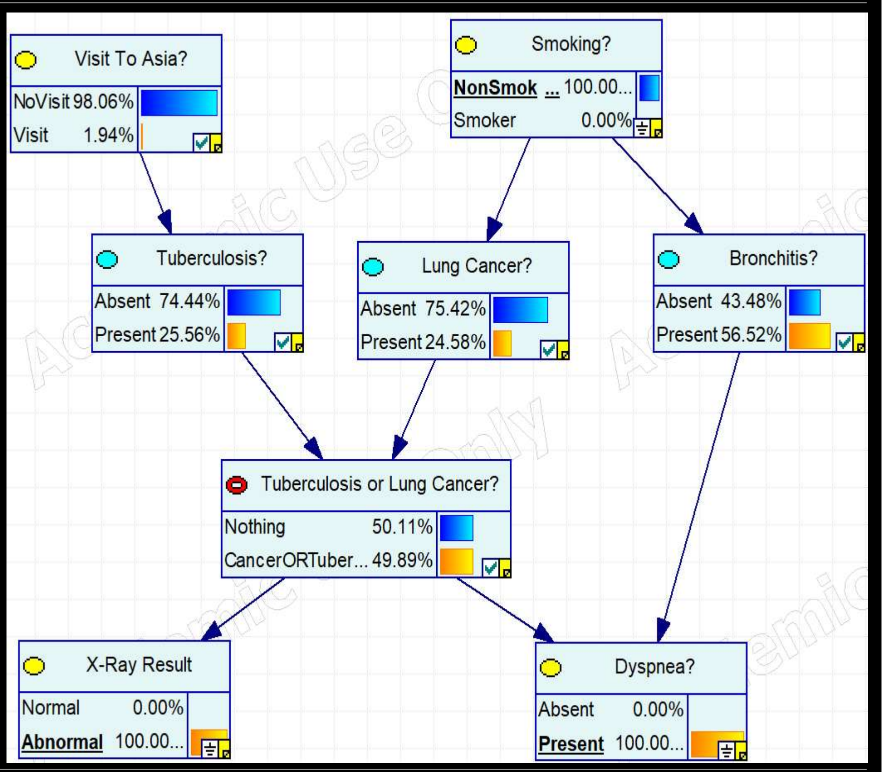


Evidence: you ask the patient whether she recently visited Asia and she replies indeed she visited Singapoor last week

«**Visit To Asia?**=*Visit*»

Query: we want to know how likely is the patient to have lung cancer

P(**Lung Cancer?** = *Present*) = ?



Evidence: you ask the patient whether she recently visited Asia and she replies indeed she visited Singapoor last week

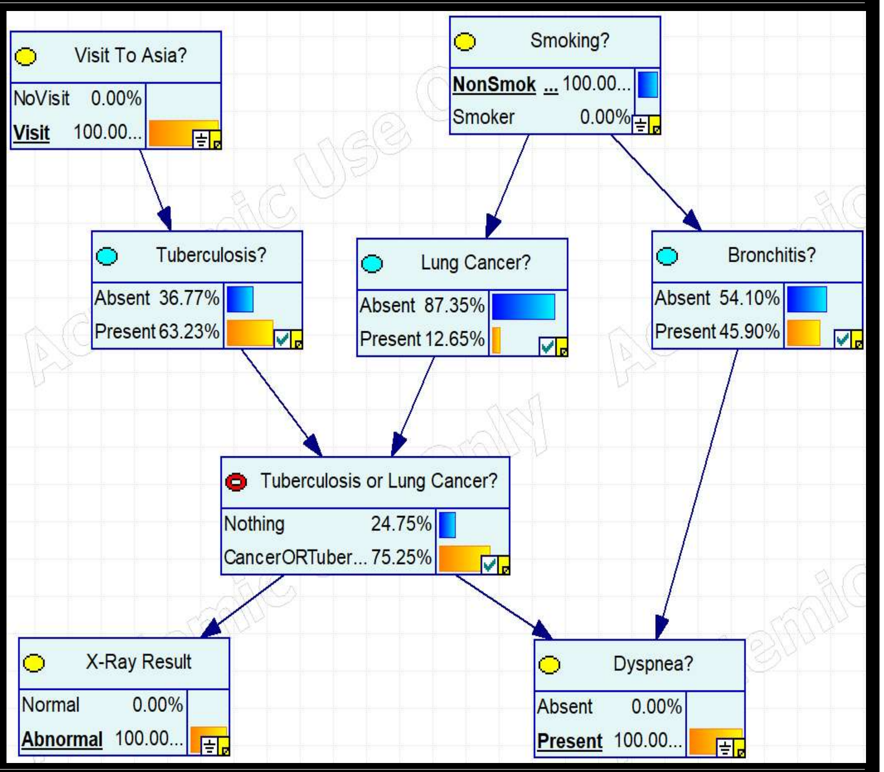
«**Visit To Asia?**=*Visit*»

Query: we want to know how likely is the patient to have lung cancer

P(**Lung Cancer?** = *Present*) = ?

Evidence Propagation: evidence is propagated from the node **Dyspnea**? to the remaining nodes

P(**Lung Cancer?** = *Present*) = 0.1265



RECAP

- SIMPSON'S PARADOX: A CURIOUS CASE
- LADDER OF CAUSATION
- ASIA BAYESIAN NETWORK