Bayesian Networks in Epistemology and Philosophy of Science

Lecture 1: Bayesian Networks

Stephan Hartmann

Center for Logic and Philosophy of Science
Tilburg University, The Netherlands

Formal Epistemology Course
Northern Institute of Philosophy
Aberdeen, June 2010

Motivation

- Bayesian Networks represent probability distributions over many variables X_i.
- They encode information about conditional probabilistic independencies between X_i.
- Bayesian Networks can be used to examine more complicated (=realistic) situations. This helps us to relax many of the idealizations that are usually made by philosophers.
- I introduce the theory of Bayesian Networks and present various applications to epistemology and philosophy of science.

Organizational Issues

Procedure: Mix of lecture and exercises units.

Literature:

1. Bovens, L. and S. Hartmann (2003): *Bayesian Epistemology*. Oxford: Oxford University Press (Ch. 3).
2. Dizadji-Bahmani, F., R. Frigg and S. Hartmann (2010): Confirmation and Reduction: A Bayesian Account. To appear in *Erkenntnis*.
3. Hartmann, S. and Meijs, W. (2010): Walter the Banker: The Conjunction Fallacy Reconsidered. To appear in *Synthese*.
4. Neapolitan, R. (2004): *Learning Bayesian Networks*. London: Prentice Hall (= the recommended textbook; Chs. 1 and 2).
5. Pearl, J. (1988): *Probabilistic Reasoning in Intelligent Systems*. San Francisco: Morgan Kaufmann (= the classic text).

Overview

Lecture 1: Bayesian Networks

1. Probability Theory
2. Bayesian Networks
3. Partially Reliable Sources

Lecture 2: Applications in Philosophy of Science

1. A Survey
2. Intertheoretic Reduction
3. Open Problems

Lecture 3: Applications in Epistemology

1. A Survey
2. Bayesianism Meets the Psychology of Reasoning
3. Open Problems

The Kolmogorov Axioms

Let $S = \{\mathrm{A}, \mathrm{B}, \ldots\}$ be a collection of sentences, and let P be a probability function. P satisfies the Kolmogorov Axioms:

Kolmogorov Axioms

1. $P(\mathrm{A}) \geq 0$
2. $P(\mathrm{A}) = 1$ if A true in all models
3. $P(\mathrm{A} \vee \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B})$ if A, B mutually exclusive

Some consequences:

1. $P(\neg\mathrm{A}) = 1 - P(\mathrm{A})$
2. $P(\mathrm{A} \vee \mathrm{B}) = P(\mathrm{A}) + P(\mathrm{B}) - P(\mathrm{A}, \mathrm{B})$; $(P(\mathrm{A}, \mathrm{B}) := P(\mathrm{A} \wedge \mathrm{B}))$
3. $P(\mathrm{A}) = \sum_{i=1}^{n} P(\mathrm{A} \wedge \mathrm{B}_i)$ if $\mathrm{B}_1, \ldots, \mathrm{B}_n$ are exhaustive and mutually exclusive ("Law of Total Probability")

Conditional Probabilities

Definition: Conditional Probability

$$P(\mathrm{A}|\mathrm{B}) := \frac{P(\mathrm{A}, \mathrm{B})}{P(\mathrm{B})} \quad \text{if} \quad P(\mathrm{B}) \neq 0$$

Bayes' Theorem:

$$\begin{aligned} P(B|A) &= \frac{P(\mathrm{A}|\mathrm{B})\, P(\mathrm{B})}{P(\mathrm{A})} = \frac{P(\mathrm{A}|\mathrm{B})\, P(\mathrm{B})}{P(\mathrm{A}|\mathrm{B})\, P(\mathrm{B}) + P(\mathrm{A}|\neg\mathrm{B})\, P(\neg\mathrm{B})} \\ &= \frac{P(\mathrm{B})}{P(\mathrm{B}) + P(\neg\mathrm{B})\, x} \end{aligned}$$

with the likelihood ratio

$$x := \frac{P(\mathrm{A}|\neg\mathrm{B})}{P(\mathrm{A}|\mathrm{B})}$$

Conditional Independence

Definition: (Unconditional) Independence

A and B are independent iff
$P(\mathrm{A}, \mathrm{B}) = P(\mathrm{A})\, P(\mathrm{B}) \Leftrightarrow P(\mathrm{A}|\mathrm{B}) = P(\mathrm{A}) \Leftrightarrow P(\mathrm{B}|\mathrm{A}) = P(\mathrm{B})$.

Definition: Conditional Independence

A is cond. independent of B given C iff $P(\mathrm{A}|\mathrm{B}, \mathrm{C}) = P(\mathrm{A}|\mathrm{C})$.

Example: A = yellow fingers, B = lung cancer, C = smoking

A and B are positively correlated, i.e. learning that a person has A raises the probability of B. Yet, if we know C, A leaves the probability of B unchanged.

C is called the common cause of A and B.

Propositional Variables

- We introduce two-valued propositional variables $A, B, \ldots$ (in italics). Their values are A and ¬A (in roman script) etc.
- Conditional independence, denoted by $A \perp\!\!\!\perp B | C$, is a relation between propositional variables (or sets of variables).
- $A \perp\!\!\!\perp B | C$ holds if $P(\mathrm{A}|\mathrm{B}, \mathrm{C}) = P(\mathrm{A}|\mathrm{C})$ for all values of A, B and C. (See exercise 4)
- The relation $A \perp\!\!\!\perp B | C$ is symmetrical: $A \perp\!\!\!\perp B | C \Leftrightarrow B \perp\!\!\!\perp A | C$
- Question: Which further conditions does the conditional independence relation satisfy?

Semi-Graphoid Axioms

The conditional independence relation satisfies the following conditions:

Semi-Graphoid Axioms

1. Symmetry: $X \perp\!\!\!\perp Y|Z \quad \Leftrightarrow \quad Y \perp\!\!\!\perp X|Z$
2. Decomposition: $X \perp\!\!\!\perp Y, W|Z \quad \Rightarrow \quad X \perp\!\!\!\perp Y|Z$
3. Weak Union: $X \perp\!\!\!\perp Y, W|Z \quad \Rightarrow \quad X \perp\!\!\!\perp Y|W, Z$
4. Contraction: $X \perp\!\!\!\perp Y|Z \;\&\; X \perp\!\!\!\perp W|Y, Z \quad \Rightarrow \quad X \perp\!\!\!\perp Y, W|Z$

With these axioms, new conditional independencies can be obtained from known independencies.

Joint and Marginal Probability

To specify the joint probability of two binary propositional variables A and B, three probability values have to be specified.

- Example: $P(\mathrm{A}, \mathrm{B}) = .2$, $P(\mathrm{A}, \neg\mathrm{B}) = .1$, and $P(\neg\mathrm{A}, \mathrm{B}) = .6$
- Note: $\sum_{A,B} P(A, B) = 1 \rightarrow P(\neg\mathrm{A}, \neg\mathrm{B}) = .1$

In general, $2^n - 1$ values have to be specified to specify the joint distribution over n variables.

With the joint probability, we can calculate marginal probabilities.

Definition: Marginal Probability

$P(\mathrm{A}) = \sum_B P(\mathrm{A}, B)$

Illustration: A: patient has lung cancer, B: X-ray test is reliable

Joint and Marginal Probability (cont'd)

The joint probability distribution contains everything we need to calculate all conditional and marginal probabilities involving the respective variables:

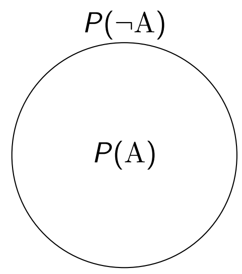
Conditional Probability

$$P(\mathrm{A}_1, \ldots, \mathrm{A}_m | \mathrm{A}_{m+1}, \ldots, \mathrm{A}_n) = \frac{P(\mathrm{A}_1, \ldots, \mathrm{A}_n)}{P(\mathrm{A}_{m+1}, \ldots, \mathrm{A}_n)}$$

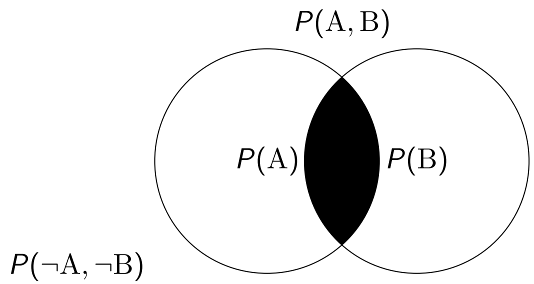
Marginal Probability

$$P(\mathrm{A}_{m+1}, \ldots, \mathrm{A}_n) = \sum_{A_1, \ldots, A_m} P(A_1, \ldots, A_m, \mathrm{A}_{m+1}, \ldots, \mathrm{A}_n)$$

A Venn Diagram Representation



Representing a Joint Probability Distribution



- Venn diagrams and the specification of all entries in $P(A_1, \ldots, A_n)$ are not the most efficient ways to represent a joint probability distribution.
- There is also a problem of computational complexity: Specifying the joint probability distribution over n variables requires the specification of $2^n - 1$ probability values.
- **The trick:** Use information about conditional independencies that hold between (sets of) variables. This will reduce the number of values that have to be specified.
- Bayesian Networks do just this . . .

An Example from Medicine

Two variables: T: Patient has tuberculosis; X: Positive X-ray
Given information:

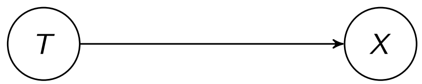
$$t := P(\mathrm{T}) = .01$$
$$p := P(\mathrm{X}|\mathrm{T}) = .95 = 1 - P(\neg\mathrm{X}|\mathrm{T}) = 1 - \text{rate of false negatives}$$
$$q := P(\mathrm{X}|\neg\mathrm{T}) = .02 = \text{rate of false positives}$$

Our **task** is to determine $P(\mathrm{T}|\mathrm{X})$. $\Rightarrow$ Apply Bayes' Theorem!

$$P(T|X) = \frac{P(\mathrm{X}|\mathrm{T})\,P(\mathrm{T})}{P(\mathrm{X}|\mathrm{T})\,P(\mathrm{T}) + P(\mathrm{X}|\neg\mathrm{T})\,P(\neg\mathrm{T})} = \frac{p\,t}{p\,t + q\,(1-t)} = \frac{t}{t + \bar{t}\,x} = .32$$

with the likelihood ratio $x := q/p$ and $\bar{t} := 1 - t$.

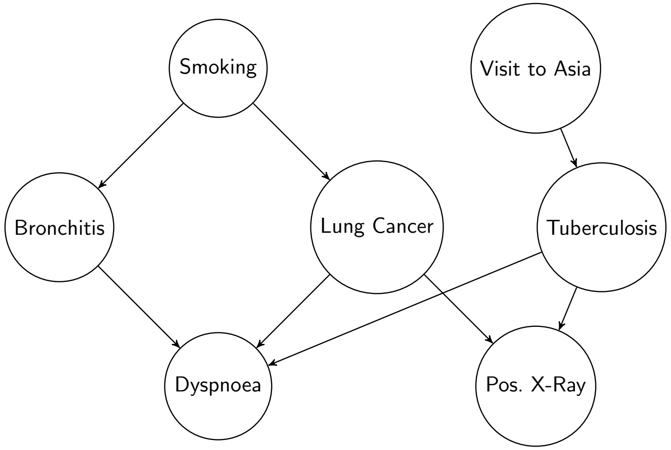
A Bayesian Network Representation



Parlance:

- "T causes X"
- "T directly influences X"

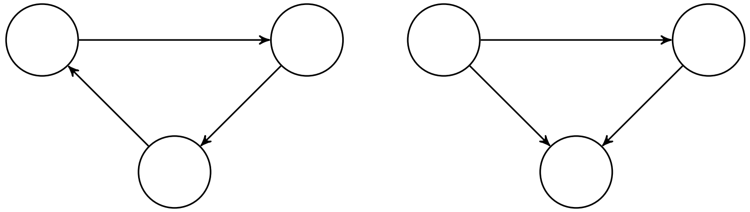
A More Complicated (= Realistic) Scenario



Directed Acyclic Graphs

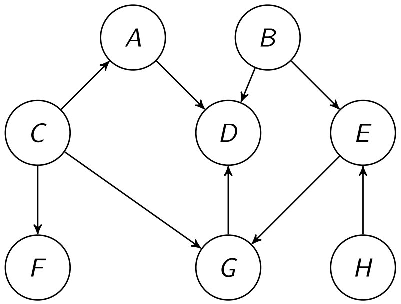
A directed graph $G(V, E)$ consists of a finite set of nodes V and an irreflexive binary relation E on V.

A directed acyclic graph (DAG) is a directed graph which does not contain cycles.



Some Vocabulary

- Parents of node A: $par(A)$
- Ancestor
- Child node
- Descendents
- Non-Descendents
- Root node



The Parental Markov Condition

Definition: The Parental Markov Condition (PMC)

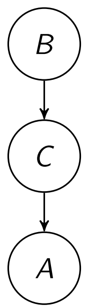
A variable is conditionally independent of its non-descendents given its parents.

Standard example: The common cause situation.

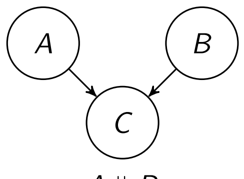
Definition: Bayesian Network

A Bayesian Network is a DAG with a probability distribution which respects the PMC.

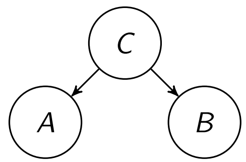
Three Examples



$A \perp\!\!\!\perp B | C$
"chain"



$A \perp\!\!\!\perp B$
"collider"



$A \perp\!\!\!\perp B | C$
"common cause"

Bayesian Networks at Work

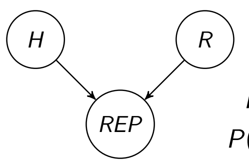
How can one calculate probabilities with a Bayesian Network?

The Product Rule

$$P(A_1, \ldots, A_n) = \prod_{i=1}^{n} P(A_i | par(A_i))$$

- Proof idea: Starts with a suitable anchestral ordering, then apply the Chain Rule and then the PMC (cf. exercises 3 & 6).
- I.e. the joint probability distribution is determined by the product of the prior probabilities of all root nodes ($par(A) = \emptyset$) and the conditional probabilities of all other nodes, given their parents.
- This requires the specification of no more than than $n \cdot 2^{m_{max}}$ values (m_{max} is the maximal number of parents).

An Example



$$P(\mathrm{H}) = h \quad , \quad P(\mathrm{R}) = r$$
$$P(\mathrm{REP}|\mathrm{H},\mathrm{R}) = 1 \quad , \quad P(\mathrm{REP}|\neg\mathrm{H},\mathrm{R}) = 0$$
$$P(\mathrm{REP}|\mathrm{H},\neg\mathrm{R}) = a \quad , \quad P(\mathrm{REP}|\neg\mathrm{H},\neg\mathrm{R}) = a$$

$$P(\mathrm{H}|\mathrm{REP}) = \frac{P(\mathrm{H},\mathrm{REP})}{P(\mathrm{REP})} = \frac{\sum_R P(\mathrm{H},R,\mathrm{REP})}{\sum_{H,R} P(H,R,\mathrm{REP})}$$
$$= \frac{P(\mathrm{H}) \sum_R P(R)\, P(\mathrm{REP}|\mathrm{H},R)}{\sum_{H,R} P(H)\, P(R)\, P(\mathrm{REP}|H,R)}$$
$$= \frac{h(r + a\,\bar{r})}{hr + a\,\bar{r}}$$

Some More Theory: *d*-Separation

We have already seen that there are more independencies in a Bayesian Network than the ones accounted for by the PMC.

Is there a systematic way to find *all* independences that hold in a given Bayesian Network?

Yes! *d*-separation

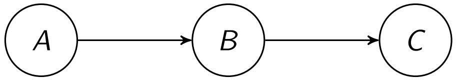
Let A, B, and C be sets of variables. Then the following theorem holds:

Theorem: *d*-Separation and Independence

$A \perp\!\!\!\perp B | C$ iff C *d*-separates A from B.

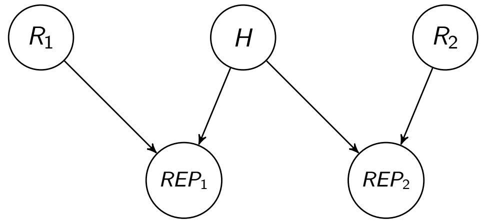
So what is *d*-separation?

Example 1



- PMC $\Rightarrow C \perp\!\!\!\perp A | B$
- But is it also the case that $A \perp\!\!\!\perp C | B$?
- This does not follow from PMC: PMC $\not\Rightarrow A \perp\!\!\!\perp C | B$
- $A \perp\!\!\!\perp C | B$ can, however, be derived from $C \perp\!\!\!\perp A | B$ and the **Symmetry Axiom** for Semi-Graphoids.

Example 2



- PMC $\Rightarrow REP_1 \perp\!\!\!\perp REP_2 | H, R_1$ (*)
- But: PMC $\not\Rightarrow REP_1 \perp\!\!\!\perp REP_2 | H$
- However: PMC $\Rightarrow R_1 \perp\!\!\!\perp H, REP_2$
- **Weak Union** $\Rightarrow R_1 \perp\!\!\!\perp REP_2 | H$ (**)
- (*), (**), **Symmetry & Contraction** $\Rightarrow R_1, REP_1 \perp\!\!\!\perp REP_2 | H$
- **Decomposition & Symmetry** $\Rightarrow REP_1 \perp\!\!\!\perp REP_2 | H$

d-Separation

Definition: $d-$Separation

A path p is d-separated (or blocked) by (a set) Z iff there is a node $w \in p$ satisfying either:

1. w has converging arrows ($u \rightarrow w \leftarrow v$) and none of w or its descendents are in Z.
2. w does not have converging arrows and $w \in Z$.

Theorem: $d-$Separation and Independence (again)

If Z blocks every path from X to Y, then Z d-separates X from Y and $X \perp\!\!\!\perp Y | Z$.

How to Construct a Bayesian Network

1. Specify all relevant variables.
2. Specify all conditional independences which hold between them.
3. Construct a Bayesian Network which exhibits these conditional independencies.
4. Check other (perhaps unwanted) independencies with the d-separation criterion. Modify the networks if necessary.
5. Specify the prior probabilities of all root nodes and the conditional probabilities of all other nodes, given their parents.
6. Calculate the (marginal or conditional) probabilities you are interested in using the Product Rule.

Plan

The proof of the pudding is in its eating!

Motivation

Guiding question: When we receive information from *independent* and *partially reliable sources*, what is our degree of confidence that this information is true?

Independence?

Partial reliability?

A. Independence

Assume that there are n facts (represented by propositional variables F_i) and there are n corresponding reports (represented by propositional variables REP_i) by partially reliable witnesses (testimonies, scientific instruments, etc.).

Assume that, given the corresponding fact, a report is independent of all other reports and of all other facts. They do not matter for the report. I.e., we assume that

Independent Reports

$REP_i \perp\!\!\!\perp F_1, REP_1, \ldots F_{i-1}, REP_{i-1}, F_{i+1}, REP_{i+1}, \ldots F_n, REP_n | F_i$ for all $i = 1, \ldots, n$.

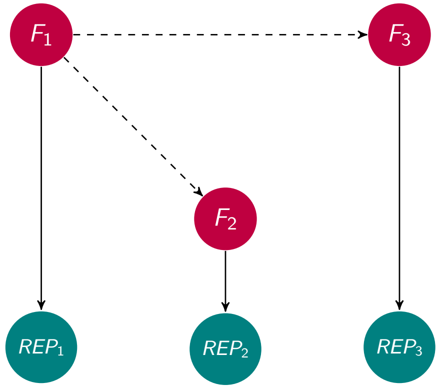
B. Partial Reliability

To model partially reliable information sources, additional model assumptions have to be made.

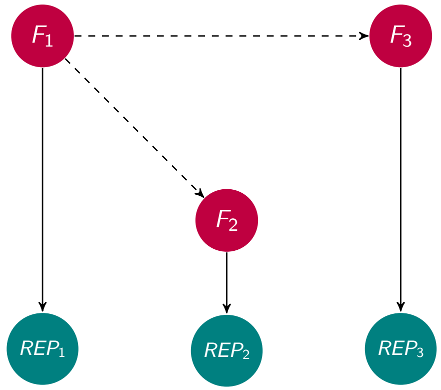
Examine two models!

Model I: Fixed Reliability

Paradigm: Medical Testing



Model I: Fixed Reliability (cont'd)



$$P(\mathrm{REP_i}|\mathrm{F_i}) = p$$
$$P(\mathrm{REP_i}|\neg\mathrm{F_i}) = q < p$$

Measuring Reliability

We assume positive reports. In the network, we specify two parameters that characterize the reliability of the sources, i.e. $p := P(\mathrm{REP_i}|\mathrm{F_i})$ and $q := P(\mathrm{REP_i}|\neg\mathrm{F_i})$.

Definition: Reliability

$r := 1 - q/p$ with $p > q$ (confirmatory reports)

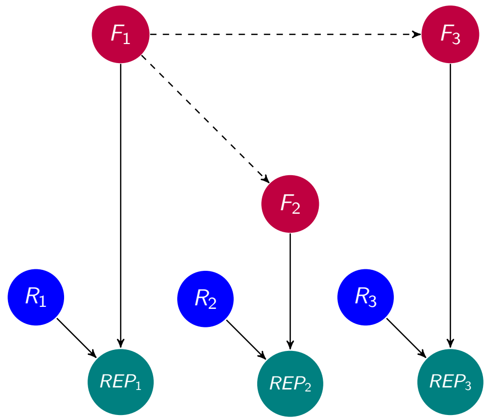
This definition makes sense:

1. If $q = 0$, then the source is maximally reliable.
2. If $p = q$, then the facts do not matter for the report and the source is maximally unreliable.

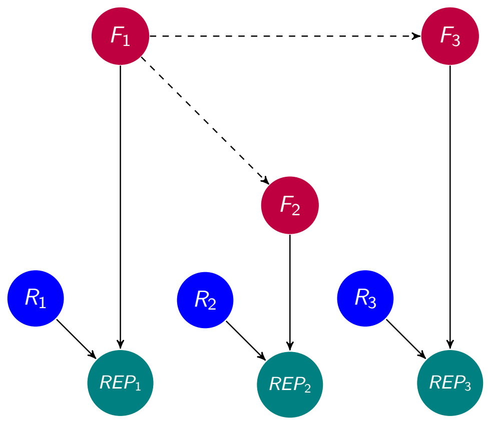
Note that any other normalized negative function of q/p also works and the results that obtain do not depend on this choice.

Model II: Variable Reliability, Fixed Random Parameter

Paradigm: Scientific Instruments

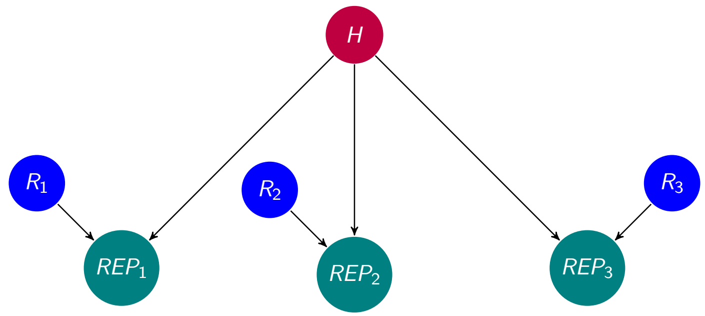


Model II: Variable Reliability, Fixed Random. Parameter



$$
\begin{aligned}
P(\mathrm{REP_i}|\mathrm{F_i}, \mathrm{R_i}) &= 1 \\
P(\mathrm{REP_i}|\neg\mathrm{F_i}, \mathrm{R_i}) &= 0 \\
P(\mathrm{REP_i}|\mathrm{F_i}, \neg\mathrm{R_i}) &= a \\
P(\mathrm{REP_i}|\neg\mathrm{F_i}, \neg\mathrm{R_i}) &= a
\end{aligned}
$$

Model IIa: Testing One Hypothesis



$$
P(\mathrm{REP_i}|\mathrm{H}, \mathrm{R_i}) = 1 \quad , \quad P(\mathrm{REP_i}|\neg\mathrm{H}, \mathrm{R_i}) = 0\,,
$$
$$
P(\mathrm{REP_i}|\mathrm{H}, \neg\mathrm{R_i}) = a \quad , \quad P(\mathrm{REP_i}|\neg\mathrm{H}, \neg\mathrm{R_i}) = a
$$

Outlook

1. The Parental Markov Condition is part of the definition of a Bayesian Network.
2. The *d*-separation criterion helps us to identify all conditional independences in a Bayesian Network.
3. We constructed two basic models of partially reliable information sources:
   (i) Endogenous reliability (paradigm: medical testing)
   (ii) Exogenous reliability (paradigm: scientific instruments)
4. In the following two lectures, we will examine applications of Bayesian Networks in philosophy of science (lecture 2) and epistemology (lecture 3).