Explainable AI and the user:

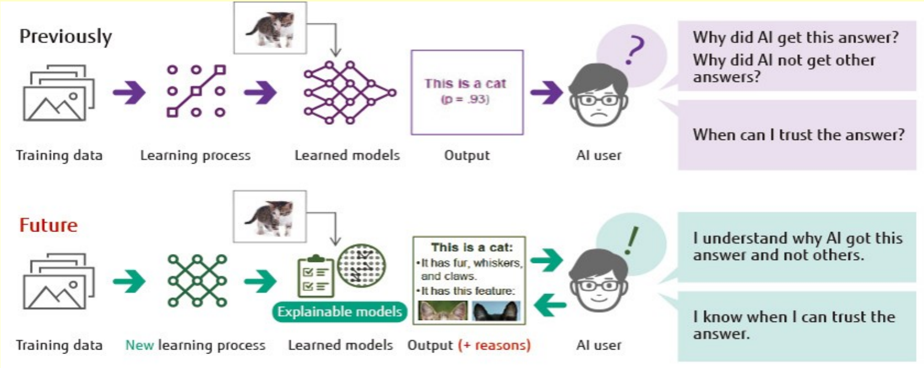
the perspective of a typical(?) computer scientist

Silja Renooij

Universiteit Utrecht

RAIL LAB

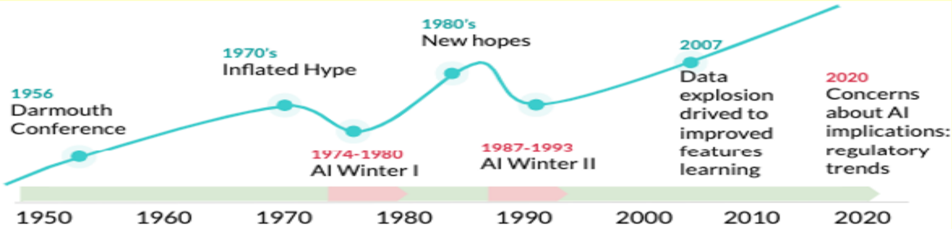
Human-AI interaction: the goal of explainable AI



Wikipedia:
Explainable AI (XAI) refers to methods and techniques in the application of artificial intelligence technology (AI) such that the results of the solution can be understood by human experts.

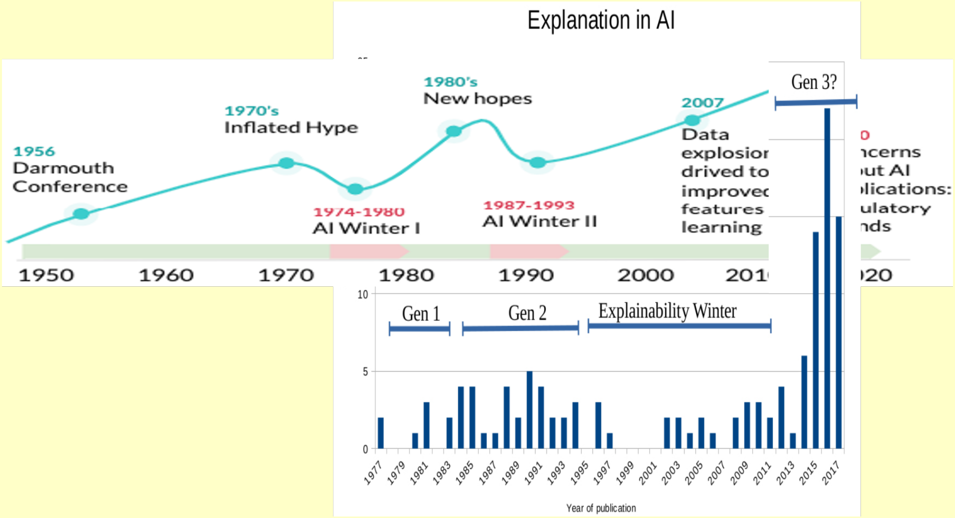
Img: https://blog.global.fujitsu.com/fgb/2019-08-01/why-ai-got-the-answer-explainable-ai-showing-bases/

History of AI output



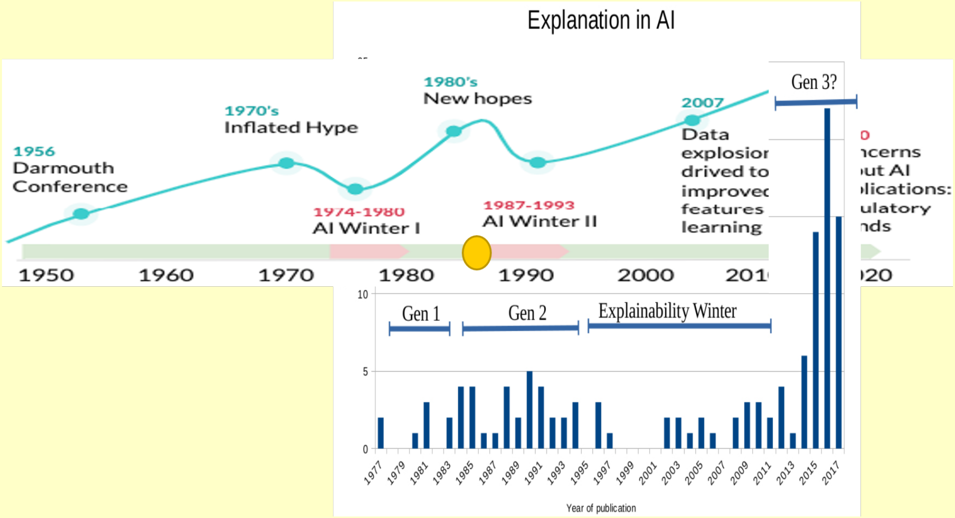
AI: https://www.finextra.com/the-long-read/62/what-should-be-taken-into-account-if-artificial-intelligence-is-to-be-regulated

History of AI and XAI output



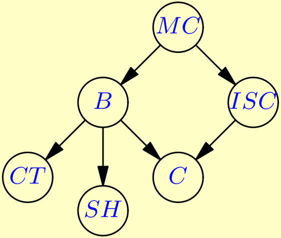
AI: `https://www.finextra.com/the-long-read/62/what-should-be-taken-into-account-if-artificial-intelligence-is-to-be-regulated`
XAI: 2019 DARPA report *Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI*

History of AI and XAI output



AI: `https://www.finextra.com/the-long-read/62/what-should-be-taken-into-account-if-artificial-intelligence-is-to-be-regulated`
XAI: 2019 DARPA report *Explanation in Human-AI Systems: A Literature Meta-Review Synopsis of Key Ideas and Publications and Bibliography for Explainable AI*

An AI model: the Bayesian network (BN)



$$
\begin{aligned}
P(b \mid mc) &= 0.20 \\
P(b \mid \neg mc) &= 0.05 \\[1ex]
P(sh \mid b) &= 0.80 \\
P(sh \mid \neg b) &= 0.60 \\[1ex]
P(ct \mid b) &= 0.95 \\
P(ct \mid \neg b) &= 0.10
\end{aligned}
\qquad
\begin{aligned}
P(mc) &= 0.20 \\[1ex]
P(c \mid b \wedge isc) &= 0.80 \\
P(c \mid \neg b \wedge isc) &= 0.80 \\
P(c \mid b \wedge \neg isc) &= 0.80 \\
P(c \mid \neg b \wedge \neg isc) &= 0.02 \\[1ex]
P(isc \mid mc) &= 0.80 \\
P(isc \mid \neg mc) &= 0.20
\end{aligned}
$$

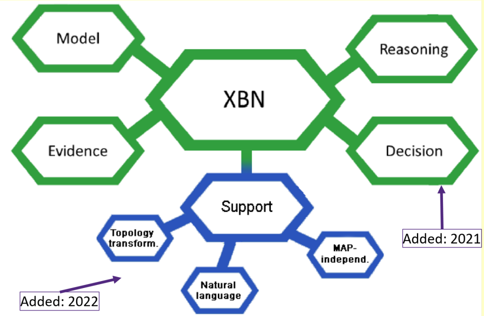
$$P(V_1, \ldots, V_n) = \prod_{i=1}^{n} P(V_i \mid pa_G(V_i))$$

Typical outputs:

- the probability of some hypothesis given evidence ($P(c \mid sh)$)
- the most likely hypothesis given evidence

Explaining Bayesian networks

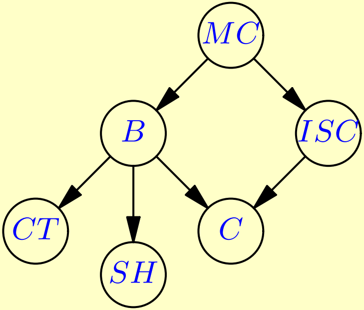
- 1992: *Explanation in Bayesian belief networks* (Stanford PhD thesis by H.J. Suermondt)
- 2001: *A Review of Explanation Methods for Bayesian Networks* (KER paper by C. Lacave and F.J. Díez)



2021: *A taxonomy of explainable Bayesian networks* (I.P. Derks, A. de Waal)

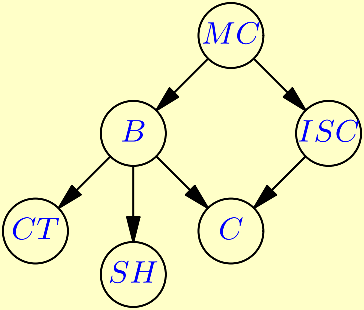
2022: *Extending MAP-independence for Bayesian network explainability* (E. Valero-Leal, P. Larrañaga, C. Bielza)

Explanation of the model



$$
\begin{array}{llll}
P(b \mid mc) = & 0.20 & P(mc) = & 0.20 \\
P(b \mid \neg mc) = & 0.05 & & \\
 & & P(c \mid b \land isc) = & 0.80 \\
P(sh \mid b) = & 0.80 & P(c \mid \neg b \land isc) = & 0.80 \\
P(sh \mid \neg b) = & 0.60 & P(c \mid b \land \neg isc) = & 0.80 \\
 & & P(c \mid \neg b \land \neg isc) = & 0.02 \\
P(ct \mid b) = & 0.95 & & \\
P(ct \mid \neg b) = & 0.10 & P(isc \mid mc) = & 0.80 \\
 & & P(isc \mid \neg mc) = & 0.20
\end{array}
$$

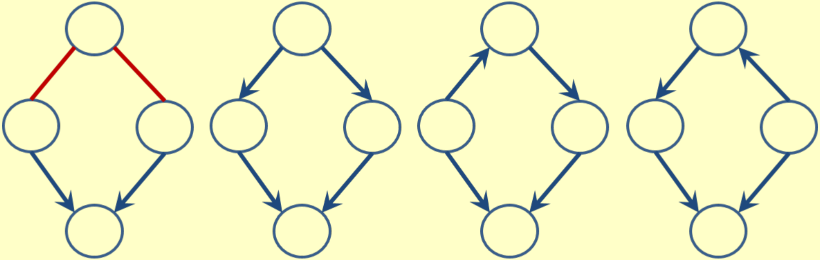
Explanation of the model



$$
\begin{array}{llll}
P(b \mid mc) = & 0.20 & P(mc) = & 0.20 \\
P(b \mid \neg mc) = & 0.05 & & \\
 & & P(c \mid b \wedge isc) = & 0.80 \\
P(sh \mid b) = & 0.80 & P(c \mid \neg b \wedge isc) = & 0.80 \\
P(sh \mid \neg b) = & 0.60 & P(c \mid b \wedge \neg isc) = & 0.80 \\
 & & P(c \mid \neg b \wedge \neg isc) = & 0.02 \\
P(ct \mid b) = & 0.95 & & \\
P(ct \mid \neg b) = & 0.10 & P(isc \mid mc) = & 0.80 \\
 & & P(isc \mid \neg mc) = & 0.20
\end{array}
$$

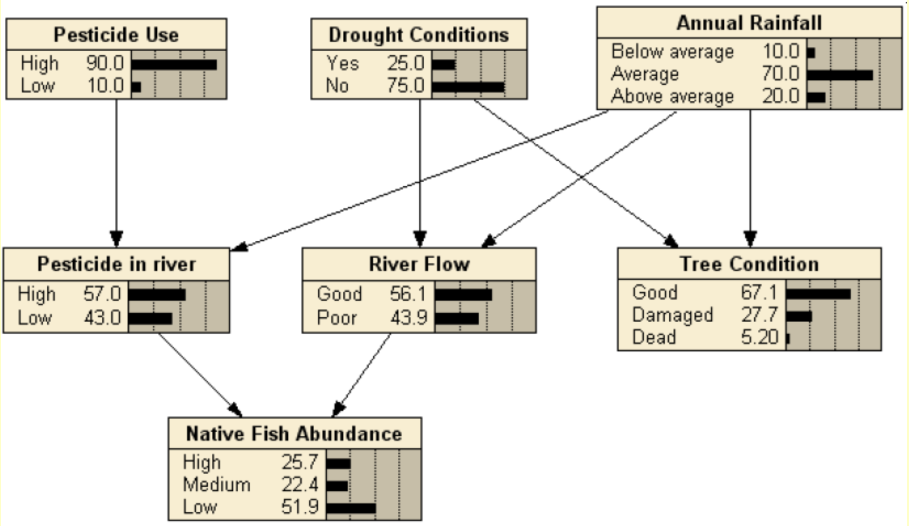
Beware of the DAG! (Directed Acyclic Graph)

- DAG suggests causal interpretation;
- DAGs in the same equivalence class represent the same probabilistic independences



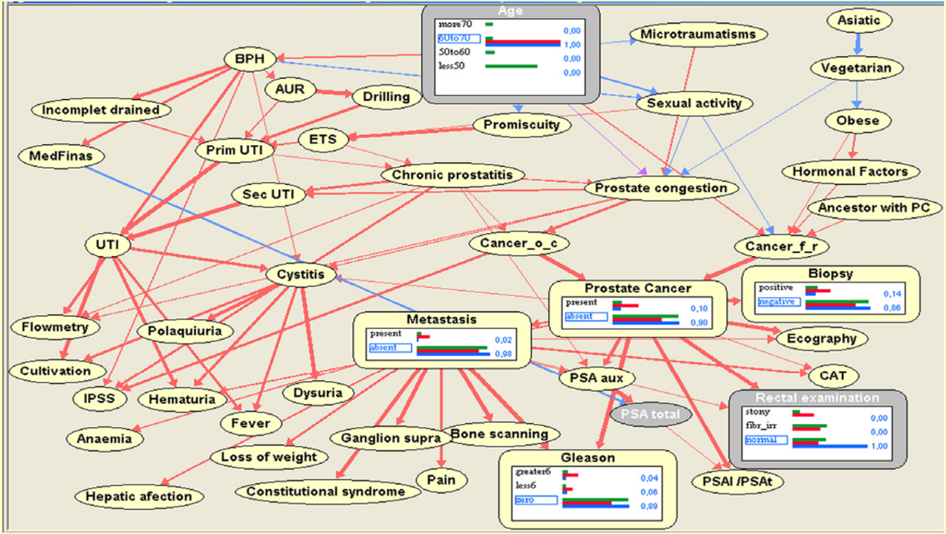
$\Longrightarrow$ BNs with different graphs and different 'causal' interpretation can represent the exact same distribution!

Explanation of the model: priors



BN: *The Native Fish Bayesian networks* (A. Nicholson, O. Woodberry, Ch. Twardy, Bayesian Intelligence Tech.Rep. 2010)

Explanation of reasoning



Img: *Explanation of Bayesian Networks and Influence Diagrams in Elvira* (C. Lacave, M. Luque, F.J. Díez, IEEE Trans., 2007)

Explanation with scenarios (in natural language)

Scenarios H, E (in)compatible with most likely h^*:

```
The following scenario(s) are
compatible with cold:
A. Cold and no cat hence no
   allergy                        0.47
   Other less probable
   scenario(s)                    0.06

The following scenario(s) are
incompatible with cold:
B. No Cold and cat causing
   allergy                        0.48

Scenario A is about as likely as
scenario B (0.47/0.48)
because cold in A is a great deal
less likely than no cold in B
(0.08/0.92),
although no cat in A is a great deal
more likely than cat in B (0.9/0.1).

Therefore cold is slightly more
likely than not (p=0.52).
```

Scenario h^* most likely, with evidence for and against it:

Scenario 2: Sylvia and Tom committed the burglary. (prior probability: 0.0001, posterior probability: 0.2326)

Scenario: Sylvia and Tom committed the burglary: Sylvia and Tom had debts and a window was already broken. Then, Sylvia and Tom climbed through the window. Then, Tom stole a laptop.

Scenario 2 is complete and consistent. It contains the evidential gap 'Sylvia and Tom had debts' and the supported implausible element 'A window was already broken'.

Evidence for and against scenario 2:

* Broken window: moderate evidence to support scenario 2.
* Statement: Tom sold laptop: moderate evidence to support scenario 2.
* Testimony: window was already broken: weak evidence to support scenario 2.
* All evidence combined: very strong evidence to support scenario 2.

1991: *Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning* (M. Henrion, M.J. Druzdzel, UAI)

2016: *When stories and numbers meet in court* (C.S. Vlek, PhD Thesis, RUG)

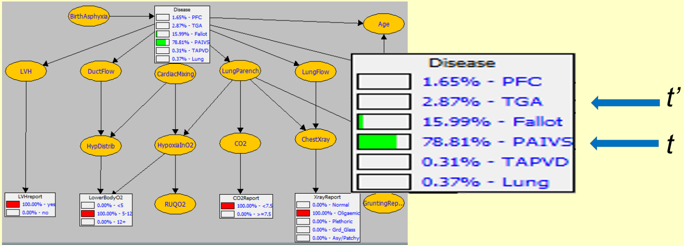
Explanations: a social science perspective

It is important to realise that [Miller, 2019]:

1. explanations are contrastive: “why P instead of Q?”
2. explanations are selected (in a biased manner): people include just one or two relevant causes as explanation; this selection is influenced by cognitive biases.
3. explanations do not refer to probabilities or statistical relationships; the most likely explanation is not always the best explanation.
4. explanations are social: presented as part of a conversation or interaction.

Miller, T. (2019) *Explanation in Artificial Intelligence: Insights from the social sciences*

Explanation: robustness of classification



Sufficient explanation(s): $\{`H = \texttt{yes} \wedge X = \texttt{Oligaemic}'\}$

Counterfactual explanations:
$\{`X = \texttt{Plethoric}', `X = \texttt{Normal} \wedge H = \texttt{no}', `X = \texttt{Grd\_Glass} \wedge H = \texttt{no}', `X = \texttt{Asy/Patchy} \wedge H = \texttt{no} \wedge O = <5'\}$

CHILD network (Spiegelhalter et al., 1993) implemented in **Samlam** (UCLA, AR Group)

Interactive explanation: human-in-the-loop

S	Based on the given evidence, what is your expected value for *Acccident*?
U	I expect a moderate accident.
S	Based on the given evidence, it is found that *severe* is the most likely value for *Accident*. Do you prefer understandable or convincing explanations for this finding?
U	I prefer understandable explanations.
S	We found the following sufficient and counterfactual explanations:

sufficient → 1. If only $Age = adolescent, Experience < 3, Model = old$ was observed *severe* would always be the most probable value for *Accident* regardless of the values for the other evidence.

counterfactual → If $Model = normal$ was observed instead of the actual values and all other values would stay the same, *moderate* would be the most probable value for Accident

2. If only $Mileage = 2000, Experience < 3, Model = old$ was observed *severe* would always be the most probable value for *Accident* regardless of the values for the other evidence.

[+ same counterfactual]

S	Do you want to see more explanations?
U	No.
S	On what observation is your expectation for a moderate accident based?
U	On the observation that I have a cautious driving style.
S	You probable have the right expectations of how E_i influences the target. However, this influence was suppressed by the other observations.

Computing contrastive, counterfactual explanations for Bayesian networks (T. Koopman, MSc. Thesis, UU, 2020)

What do all these explanations have in common?

- mostly model-specific (for BNs)
- domain-independent
- focus on what is ‘technically’ possible
- hardly a real user involved
- …

What do all these explanations have in common?

- mostly model-specific (for BNs)
- domain-independent
- focus on what is ‘technically’ possible
- hardly a real user involved
- …

Mostly a computer scientist perspective. Why?

My two cents

AI was generating explanations before we even knew what (good) explanations are.

My two cents

AI was generating explanations before we even knew what (good) explanations are.

Miller [2019]:

For over two decades, cognitive psychologists and scientists have investigated how people generate explanations and how they evaluate their quality.

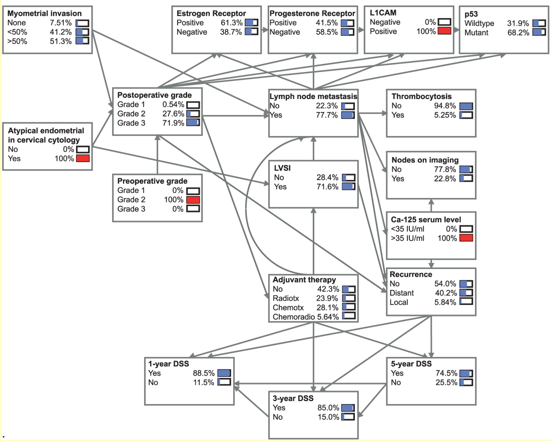
Human-centered XAI

Current research ‘involving’ users:

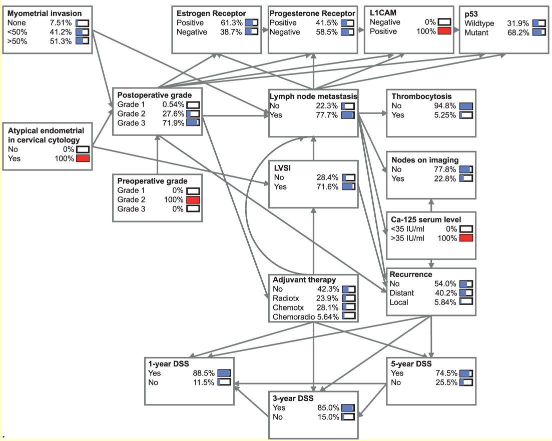
- papers that identify stakeholders
- papers that define quality, goals and types of explanation
- papers that introduce frameworks/questionnaires for user requirements concerning explanations
- many literature studies
- …

All general, model-agnostic, domain independent.

Asking a real user of a real application



Asking a real user of a real application



“We’d like a 95% confidence interval with each prediction.”

Take home message

Multi-disciplinary teams:

- need to know what is technically possible
- need to involve and interact with user more

In addition to what, whom and how, consider . . .

when:

- explanations are necessary,
  yet not everything needs explanation
- machine-in-the-loop?*

why:

- effective explanations are not always accurate

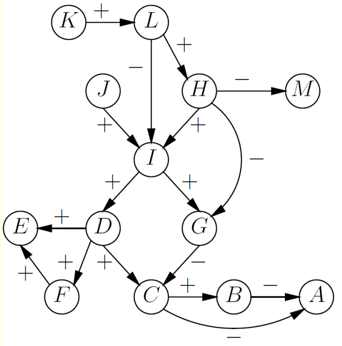
* Tim Miller (2023 arXiv preprint) *Explainable AI is Dead, Long Live Explainable AI! Hypothesis-driven decision support*

Utrecht University

DISCLAIMER

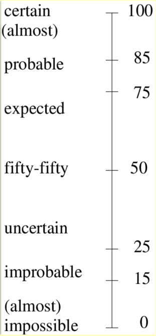
The information in this presentation has been compiled with the utmost care, but no rights can be derived from its contents.

Explanation of the model: probabilistic relations



Conjunctivitis | Mucositis (1)

Consider a pig *without an infection of the mucous*. How likely is it that this pig shows a *conjunctivitis* ?



Qualitative approaches to quantifying probabilistic networks (S. Renooij, PhD Thesis, UU, 2001)

Explanation of reasoning

Flow of influence from most relevant evidence

Before presenting any evidence, the probability of GALLSTONES being present is 0.128.

The following pieces of evidence are considered important (in order of importance):

- Presence of GUARDING results in a posterior probability of 0.175 for GALLSTONES.
- AGE of 41 results in a posterior probability of 0.172 for GALLSTONES.

Their influence flows along the following paths:

- GUARDING is caused by CHOLECYSTITIS, which is caused by GALLSTONES.
- AGE influences GALLSTONES.

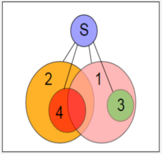
Presentation of the evidence results in a posterior probability of 0.227 for the presence of GALLSTONES.

Arguments built from most likely intermediate values

The value **scirrheus** of node **Shape** is certain (P = 1.00).

We were able to construct four arguments based on the evidence associated with the value **scirrheus** for node **Shape (S)**.
The arguments are ordered by how influential they are for the value of the node **Shape (S)**

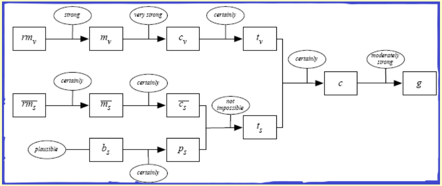
- Argument 1: Node **Endosono-mediast** has value **no**
  Node **Bronchoscopy** has value **no**
  Node **Lapa-diagragm** has value **no**
  Node **CT-organs** has value **none**
  Node **X-fistula** has value **no**
  Node **CT-liver** has value **no**
  Node **X-lungs** has value **no**
  Node **CT-lungs** has value **no**
  Node **Endosono-wall** has value **T3**
- Argument 2: Node **Gastro-shape** has value **scirrheus**
  Node **Gastro-circumf** has value **circulair**
  Node **Gastro-length** has value **5 <= x < 10**
  Node **Weightloss** has value **x<10%**
  Node **Endosono-wall** has value **T3**
  Node **Endosono-truncus** has value **non-determ**
  Node **Endosono-loco** has value **yes**
  Node **Gastro-necrosis** has value **no**
  Node **X-fistula** has value **no**
  Node **Endosono-mediast** has value **no**
  Node **Gastro-location** has value **distal**
- Argument 3: Node **Gastro-shape** has value **scirrheus**
- Argument 4: Node **X-fistula** has value **no**
  Node **Gastro-necrosis** has value **no**



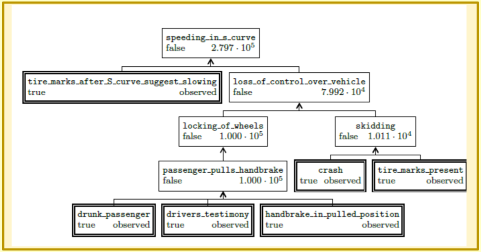
1997: *BANTER: a Bayesian network tutoring shell* (P. Haddawy, J. Jacobson, Ch.E. Kahn Jr., AI in Med.)
2015: *Explaining the reasoning of Bayesian networks with intermediate nodes and clusters* (J. van Leersum, MSc Thesis, UU)

Explanation of reasoning

Argument diagram:

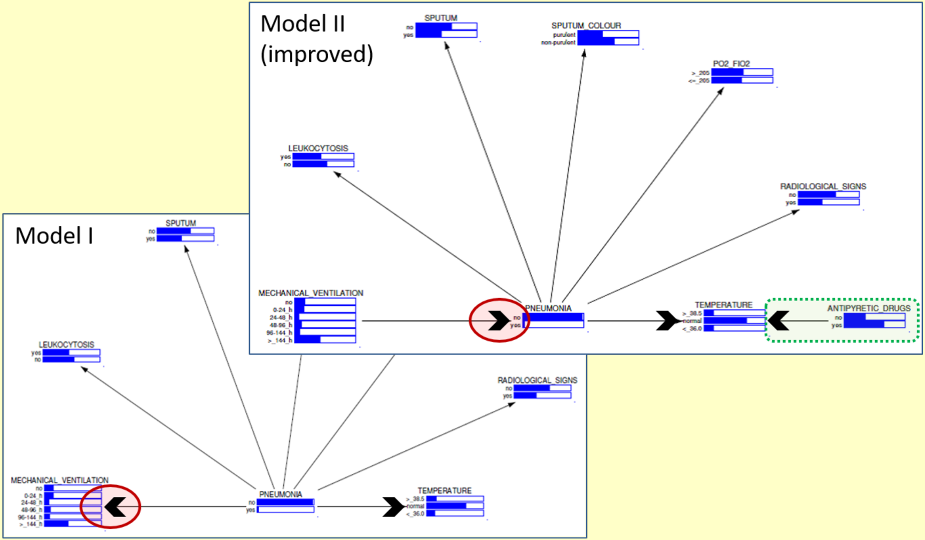


Argument tree:



2011: *On extracting arguments from Bayesian network representations of evidential reasoning* (J. Keppens, ICAIL)
2017: *Designing and understanding forensic Bayesian networks using argumentation* (S.T. Timmer, PhD Thesis, UU)

Causal anecdote



BNs: *Bayesian network models for the management of ventilator-associated pneumonia* (S. Visscher, PhD Thesis, UU, 2008)