Surfing the waves of explanation

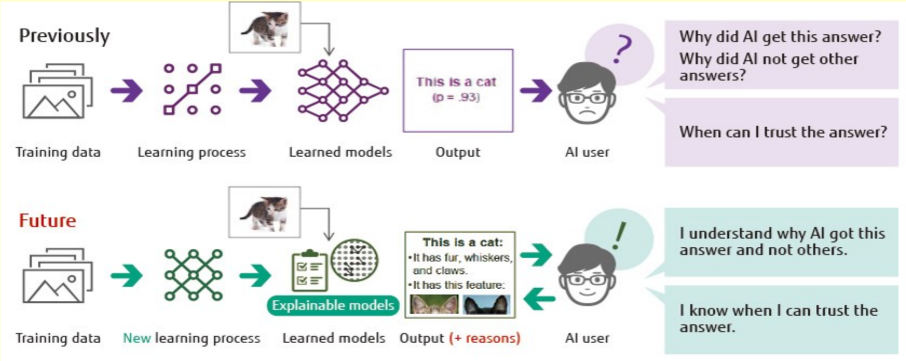
Silja Renooij

Universiteit Utrecht

SUM 2022

Img: `https://www.surfertoday.com/surfing/the-basic-rules-of-surf-etiquette`

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/

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.

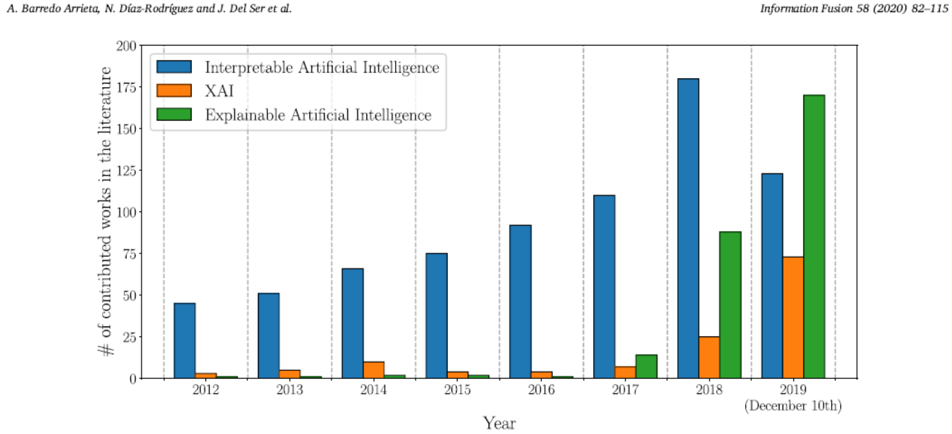
Miller, T. (2019) *Explanation in Artificial Intelligence: Insights from the social sciences*

Miller [2019]:

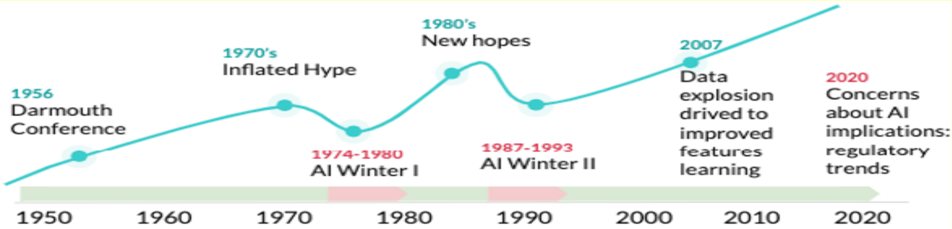
> *For over two decades, cognitive psychologists and scientists have investigated how people generate explanations and how they evaluate their quality.*

When did AI start generating and evaluating explanations?

XAI output past decade

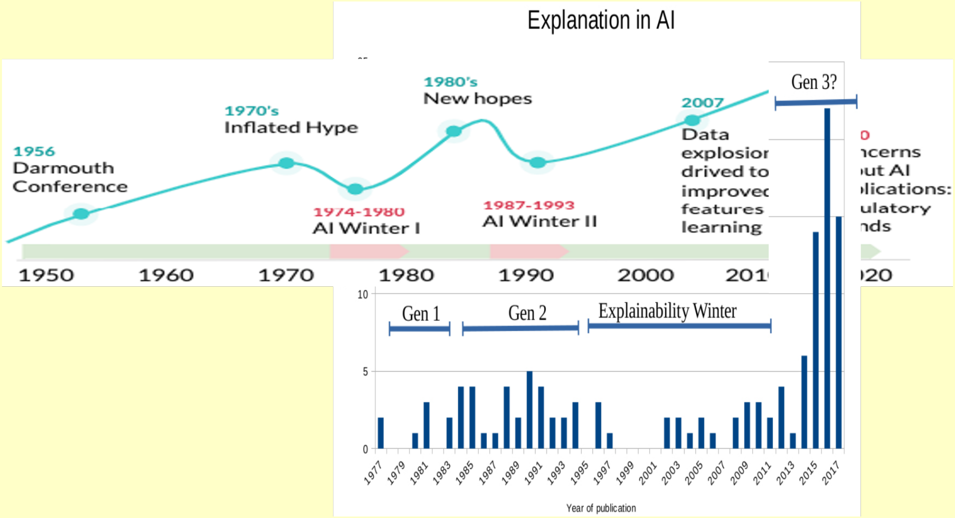


Waves of AI output



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

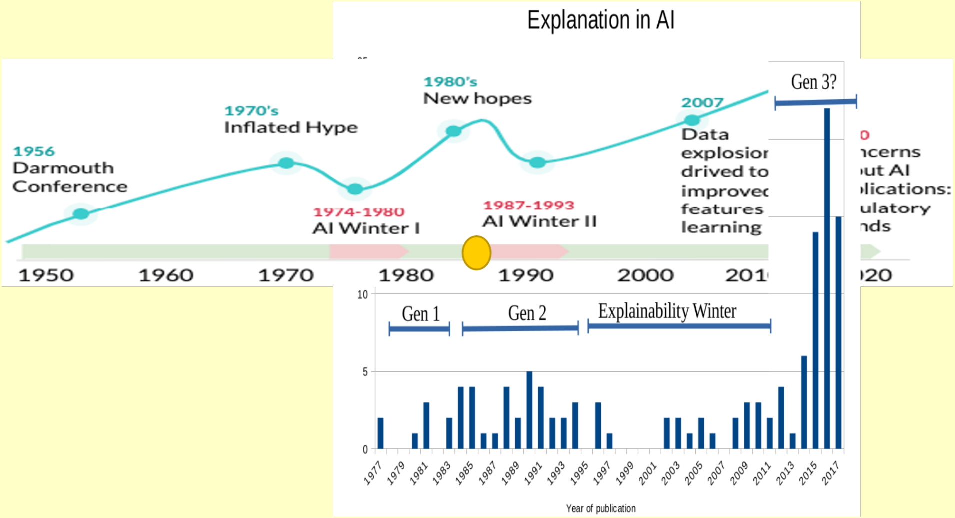
Waves of AI and XAI output



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*

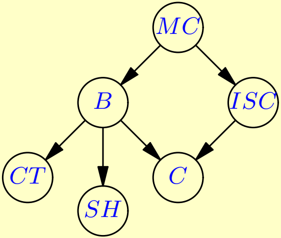
Waves of AI and XAI output



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*

Bayesian network (BN)

- late 1980s: introduced by J. Pearl;
- model $\mathcal{B}$ of discrete joint probability distribution $P(\boldsymbol{V})$;
- qualitative part: intuitive (?) DAG G of independence relation;
- quantitative part: distributions $P(V_i \mid pa_G(V_i))$;



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

- can be handcrafted or learned from data;
- $P(\boldsymbol{V}) = \prod_{i=1}^{n} P(V_i \mid pa_G(V_i))$

Reasoning in Bayesian networks: queries

Let $\boldsymbol{V} = \boldsymbol{H} \cup \boldsymbol{I} \cup \boldsymbol{E}$ be composed of three disjoint subsets.

Typical queries posed to a BN are:

MAP/MPE: $\arg\max_{\boldsymbol{h}} P(\boldsymbol{H} = \boldsymbol{h} \mid \boldsymbol{E} = \boldsymbol{e})$ (classification)

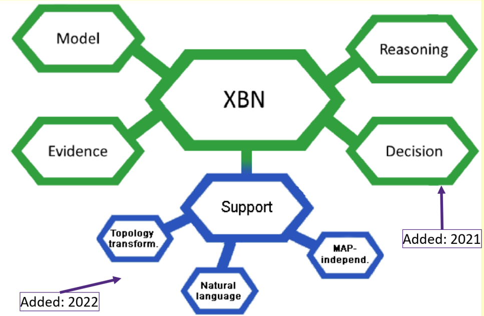
Inference: $P(\boldsymbol{H} = \boldsymbol{h} \mid \boldsymbol{E} = \boldsymbol{e})$ (What if?)

(typically $\boldsymbol{H}$ is a single V_i)

where $\boldsymbol{e}$ and $\boldsymbol{h}$ denote value assignments to $\boldsymbol{E}, \boldsymbol{H}$.

Explaining Bayesian networks

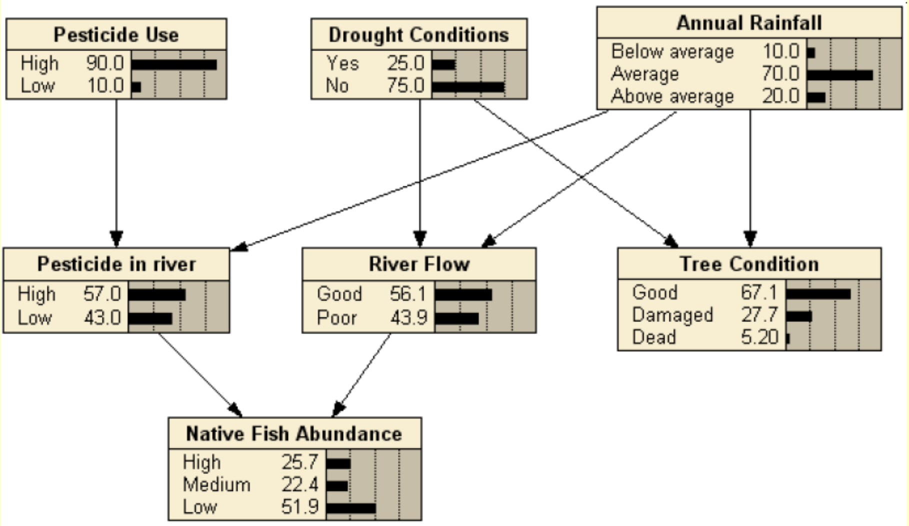
- 1992: *Explanation in Bayesian belief networks* (Stanford PhD thesis by H.J. Suermondt)
- 2001: *A Review of Explanation Methods for Bayesian Networks* (KER paper by C. Lacave and F.J. Díez)



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)

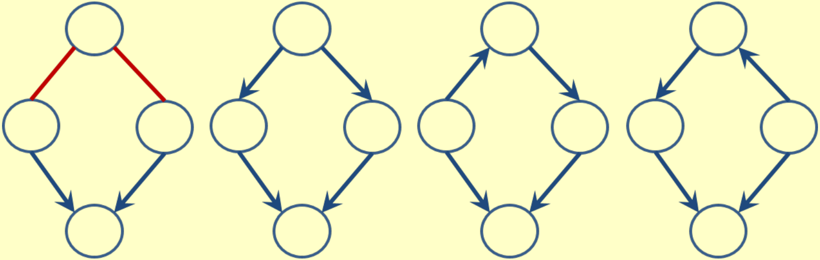
Explanation of the model: graph and visual priors



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

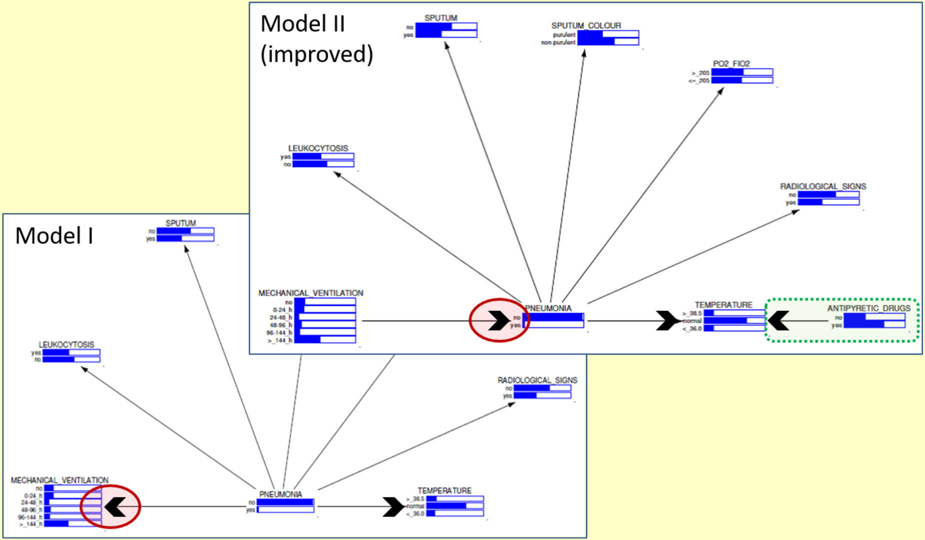
Beware of the DAG!

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



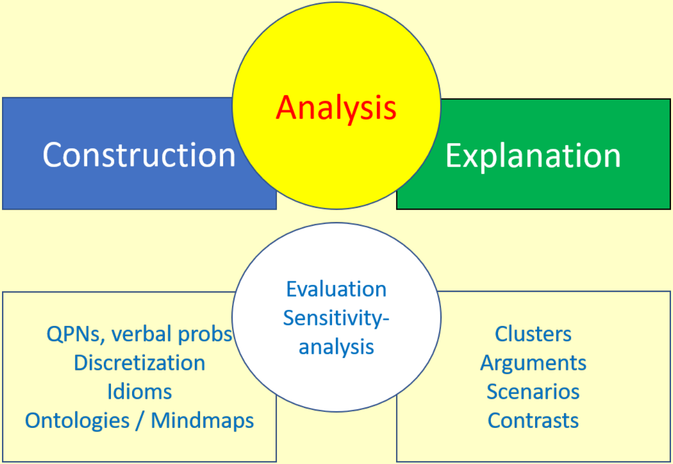
$\Longrightarrow$ BNs with different graphs and different 'causal' interpretation can represent same $P(\boldsymbol{V})$!

Causal anecdote



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

Intermezzo: general overview of my research



Analysis for explaining decisions

Derks & De Waal (2021):

Explanation of decisions supports the following questions:

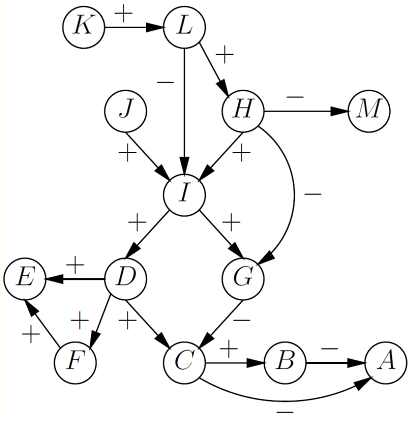
- “Given the available information, are we ready to make a decision?”, and if not
- “ What additional information do we require to make an informed decision?”

using threshold-based solutions:

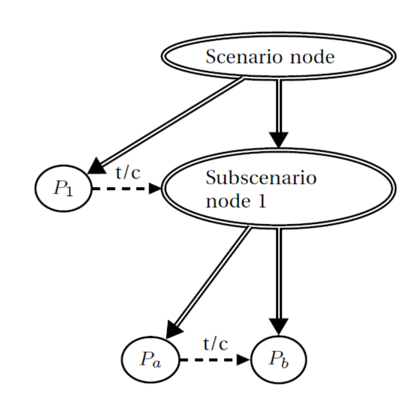
- SDP: probability that same decision is made upon obtaining additional evidence (2012 –)
- sensitivity analysis: to what extent does the outcome depend on the specified conditional probabilities? (1995 –)

Construction: using monotonicity & idioms

QPNs, ∼1990 –



idioms, ∼2000 –



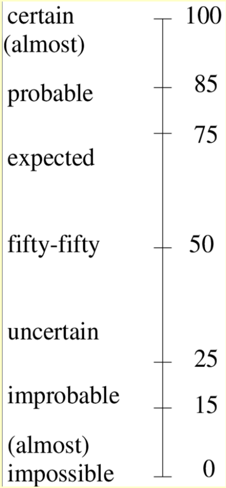
QPN: *Qualitative approaches to quantifying probabilistic networks* (S. Renooij, PhD Thesis, UU, 2001)
Narrative idiom: *When stories and numbers meet in court* (C.S. Vlek, PhD Thesis, RUG, 2016)

Construction: probability elicitation

Eliciting $P(Conjunctivitis = yes \mid Mucositis = no)$:

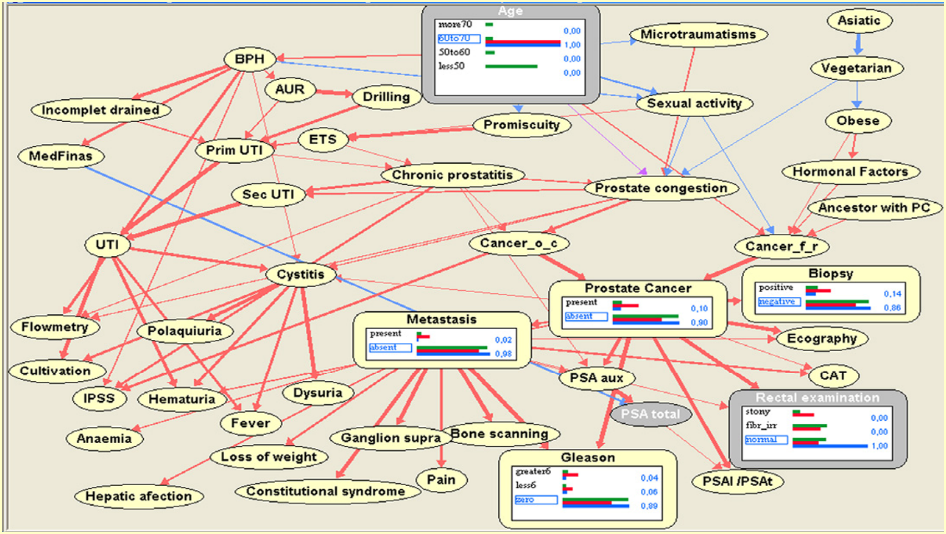
Conjunctivitis | Mucositis (1)

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



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

Explanation of reasoning: monotonicity (visual)



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

Explanation of reasoning: scenarios (textual)

1991:

```
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).
```

2016:

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)

Explanation of reasoning: relevance of evidence

1997:

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.

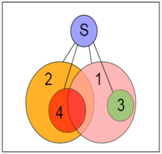
Presentation of the evidence results in a posterior probability of 0.227 for the presence of GALLSTONES.

2015:

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)**

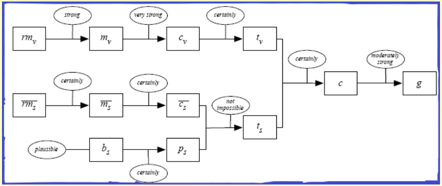
- Argument 1: Node **Endosono-mediast** has value **no**
  Node **Bronchoscopy** has value **no**
  Node **Lapa-diagragm** has value **no**
  Node **CT-organs** has value **none**
  Node **X-fistula** has value **no**
  Node **CT-liver** has value **no**
  Node **X-lungs** has value **no**
  Node **CT-lungs** 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**



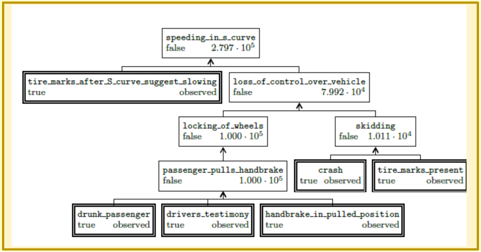
1997: *BANTER: a Bayesian network tutoring shell* (P. Haddawy, J. Jacobson, Ch.E. Kahn Jr., AI in Med.)
2015: *Explaining the reasoning of Bayesian networks with intermediate nodes and clusters* (J. van Leersum, MSc Thesis, UU)

Explanation of reasoning: argument graphs

2011:



2017:



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)

Persuasive contrastive explanation
(explanation of reasoning: classification)

Consider evidence $\mathbf{e} \in \Omega(\boldsymbol{E})$, resulting in output t instead of t'.

A persuasive contrastive explanation combines

- sufficient explanation $\mathbf{s}$
  ► *minimal* sub-configuration of evidence $\mathbf{e}$ that suffices for concluding t, regardless of the values for $\boldsymbol{E} \backslash \boldsymbol{S}$

  " evidence $\mathbf{s}$ would already be enough to conclude t "

- counterfactual explanation $\mathbf{c}$
  ► *minimal* sub-configuration of unobserved values $\overline{\mathbf{e}} \in \Omega(\boldsymbol{E})$ that in combination with the remaining evidence for $\boldsymbol{E} \backslash \boldsymbol{C}$ suffices to conclude t'

  " t' would result if the evidence contains $\mathbf{c}$ instead "

Persuasive contrastive Explanations for Bayesian networks (T. Koopman, S. Renooij, ECSQARU 2021)

Computing Explanations

- # of potential sufficient explanations: $2^{|\boldsymbol{E}|}$
- # of potential counterfactual explanations: $\prod_{k=1}^{|\boldsymbol{E}|} |\Omega(E_k)| - 1$
- we need to compute the outcome for the associated value-assignments from the network
- in Bayesian networks, probabilistic inference is NP-hard....

Various properties of these explanations allow for their computation

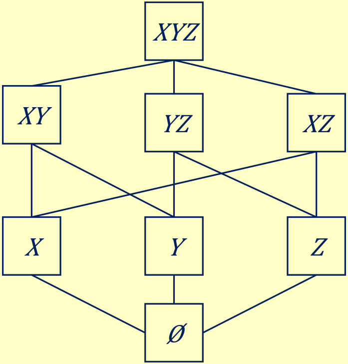
- using a breadth first search: `BFS-SFX-CFX`
- on a dynamically annotated subset lattice

Explanation lattice I

Lattice $\mathcal{L} = (\mathcal{P}(\boldsymbol{E}), \subseteq)$ and each element $\boldsymbol{S} \subseteq \boldsymbol{E}$ annotated with:

1. $\mathbf{s} \subseteq \mathbf{e}$

e.g. $\mathbf{x_1 y_1 z_1}$ for $\boldsymbol{S} = \{X, Y, Z\}$
$\mathbf{x_1 z_1}$ for $\boldsymbol{S} = \{X, Z\}$
$\mathbf{y_1}$ for $\boldsymbol{S} = \{Y\}$



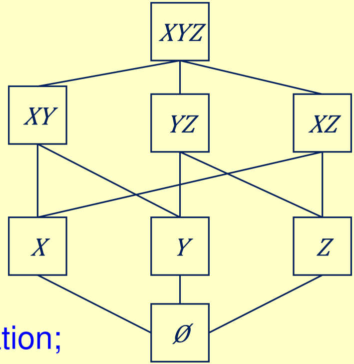
s is potentially a sufficient explanation;
(s should be as small as possible)

Explanation lattice II

Lattice $\mathcal{L} = (\mathcal{P}(\boldsymbol{E}), \subseteq)$ and each element $\boldsymbol{S} \subseteq \boldsymbol{E}$ annotated with:

2. all pairs $(\mathbf{c}, t^*)$ with $\mathbf{c} \in \Omega(\boldsymbol{E} \backslash \boldsymbol{S})$, $\mathbf{c} \subseteq \overline{\mathbf{e}}$, and t^* is output for input $\mathbf{sc}$

e.g. $(\mathbf{z_2}, t'),(\mathbf{z_3}, t)$ for $\boldsymbol{S} = \{X, Y\}$
$(\mathbf{x_2}, t'')$ for $\boldsymbol{S} = \{Y, Z\}$
$(\mathbf{x_2 y_2}, \mathrm{unkn})$ for $\boldsymbol{S} = \{Z\}$

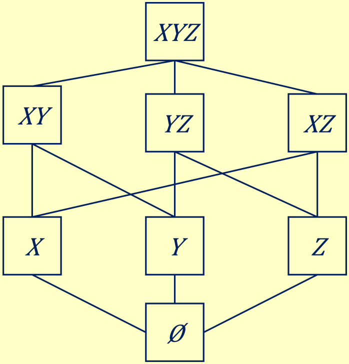


$\mathbf{c}$ is potentially a counterfactual explanation;
($\mathbf{c}$ should be as small as possible)

Explanation lattice III

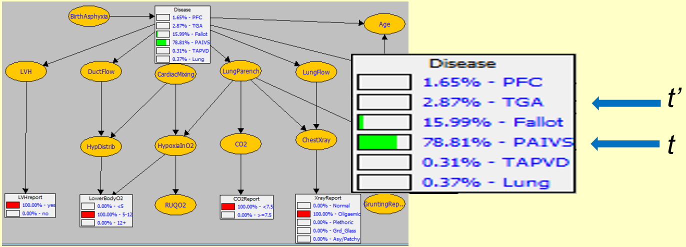
Lattice $\mathcal{L} = (\mathcal{P}(\boldsymbol{E}), \subseteq)$ and each element $\boldsymbol{S} \subseteq \boldsymbol{E}$ annotated with:

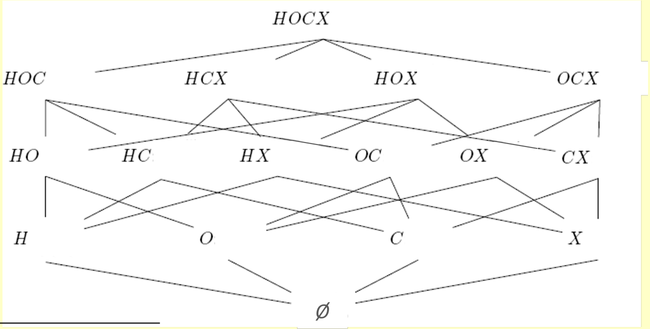
3. $l_{\boldsymbol{S}} \in \{\text{true, exp, oth}\}$
   - true: all t^* in $(\mathbf{c}, t^*)$ are t
     $\Rightarrow$ cue for continuing SFX
   - exp: all t^* are t'
     $\Rightarrow$ cue for stopping CFX
   - oth: t^* mix of $t, t', t'', \ldots$
     $\Rightarrow$ cue for SFX and CFX



Initially all labels $l_{\boldsymbol{S}}$ are empty

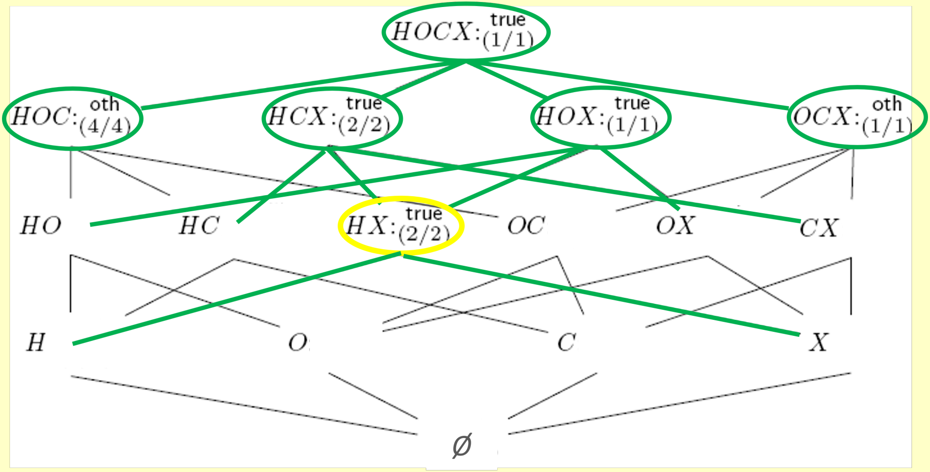
Example





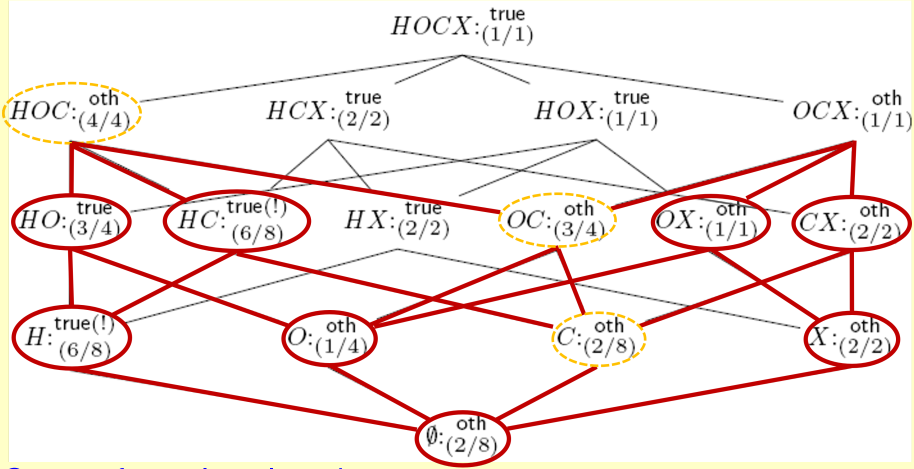
CHILD network (Spiegelhalter et al., 1993) implemented in **Samlam** (UCLA, AR Group)

Example: finding sufficient explanations



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

Example: finding counterfactual explanations



Counterfactual explanations:

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

Explanation support: MAP-independence

Recall: MAP $h^* = \arg\max_{\boldsymbol{h}} P(\boldsymbol{H} = \boldsymbol{h} \mid \boldsymbol{E} = \boldsymbol{e})$.

h^* is MAP-*independent* of subset $\boldsymbol{R}$ of intermediate variables, if for all $\boldsymbol{r} \in \Omega(\boldsymbol{R})$: (Kwisthout, 2021)

$$\arg\max_{h' \in \Omega(\boldsymbol{H})} \Pr(h' \wedge \boldsymbol{r} \mid \mathbf{e}) = h^*$$

If $\operatorname{argmax} h' \neq h^*$ for some $\boldsymbol{r}$ then

- $\boldsymbol{r}$ provides for a *counterfactual*;
- that *contrasts* outputs h^* and h'.

Note that the explanation concerns the effects of possible future observations rather than current!

Explainable AI using MAP-independence (J. Kwisthout, ECSQARU 2021)

Relevance for Robust Bayesian Network MAP-Explanations (S. Renooij, PGM 2022)

Interactive explanation

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)

Take home message

- explanations are more than ever necessary
- not everything needs explanation

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

- effective explanations are not always accurate

Utrecht University

DISCLAIMER