THE BENEFITS AND CHALLENGES OF CAUSAL MACHINE LEARNING

DR ANTHONY CONSTANTINOU

SCHOOL OF ELECTRONIC ENGINEERING AND COMPUTER SCIENCE

Bayesian Artificial Intelligence Research Lab

Queen Mary University of London

ABOUT

About me:

- Senior Lecturer (Associate Prof) in Data Analytics at Queen Mary University of London.
- Lead the Bayesian Artificial Intelligence research lab:
  - http://bayesian-ai.eecs.qmul.ac.uk/
  - At QMUL since Oct 2009: joined as a PhD student!

About the presentation:

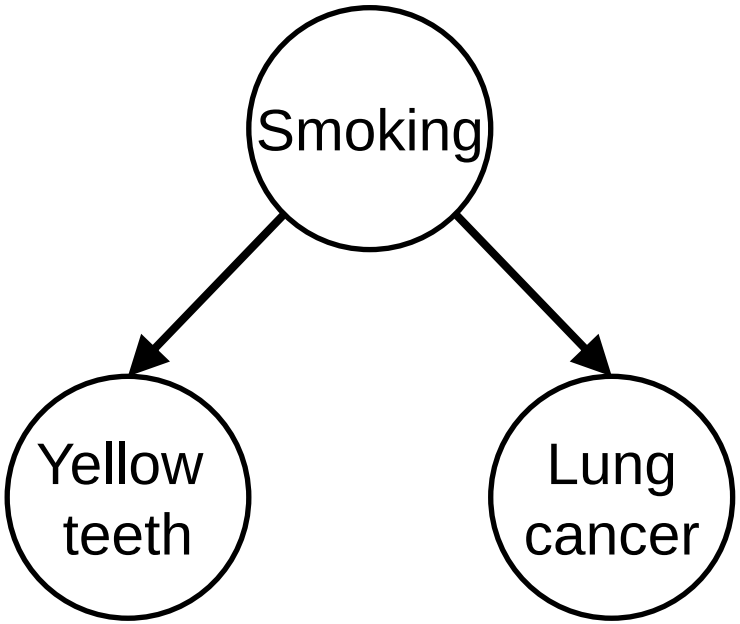
- Association and causation.
- What is causal machine learning.
- Why causal structure learning is difficult (limitations).
- Applied work.
- Why causal structure learning is important (benefits).
- Q&A.

ASSOCIATION AND CAUSATION

ASSOCIATION AND CAUSATION

Causation claims:

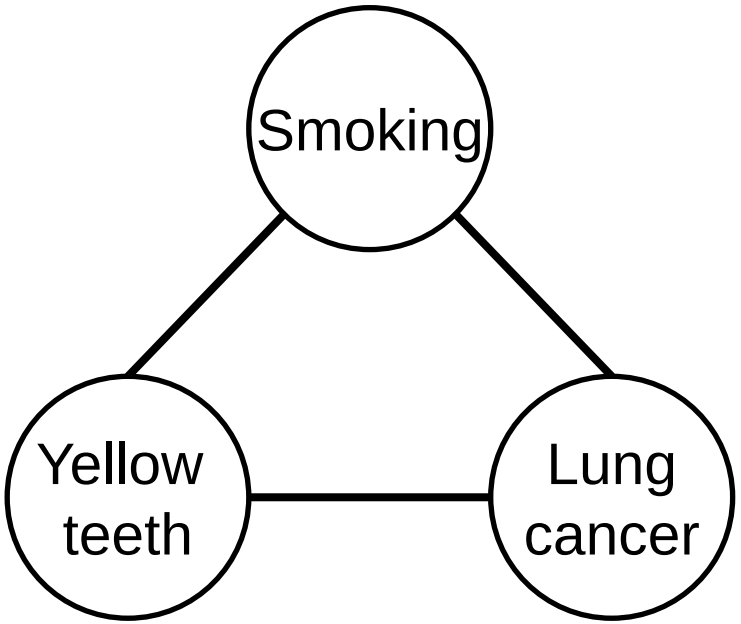
- **Smoking** is a cause for **Lung cancer** and **Yellow teeth**.
- **Lung cancer** and **Yellow teeth** are effects of **Smoking**.
- Intervening on **Smoking** will affect **Lung cancer** and **Yellow teeth**.
- Intervening on **Yellow teeth** has no effect on **Smoking** nor on **Lung cancer**.



Association and Causation

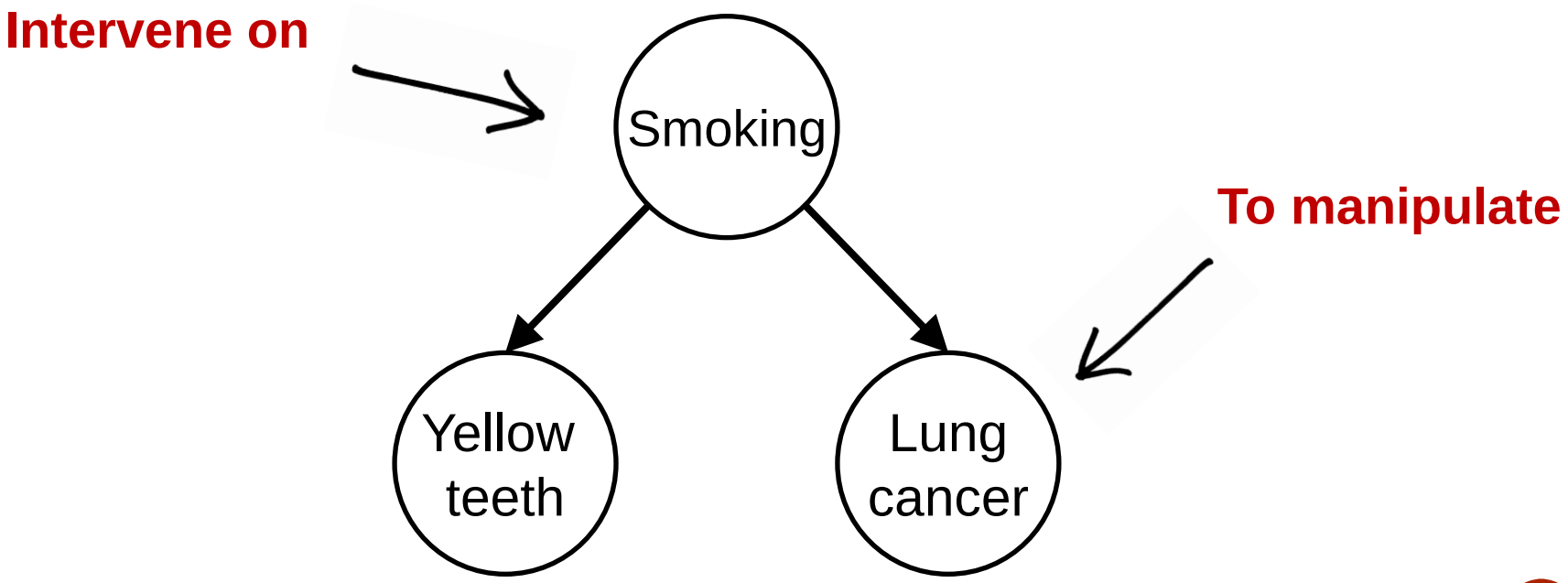
Association claims:

- Observing **Smoking** is predictive of both **Lung cancer** and **Yellow teeth**.
- Observing **Lung cancer** is predictive of both **Yellow teeth** and **Smoking**.
- Observing **Yellow teeth** is predictive of both **Smoking** and **Lung cancer**.
- Association makes no claims about interventions.



ASSOCIATION AND CAUSATION

- Causal models enable us to simulate the effect of **hypothetical interventions**, which is important for decision making.
- If we want to **reduce the risk** of **Lung cancer**, a causal model enables us to determine that we should:



Association and Causation

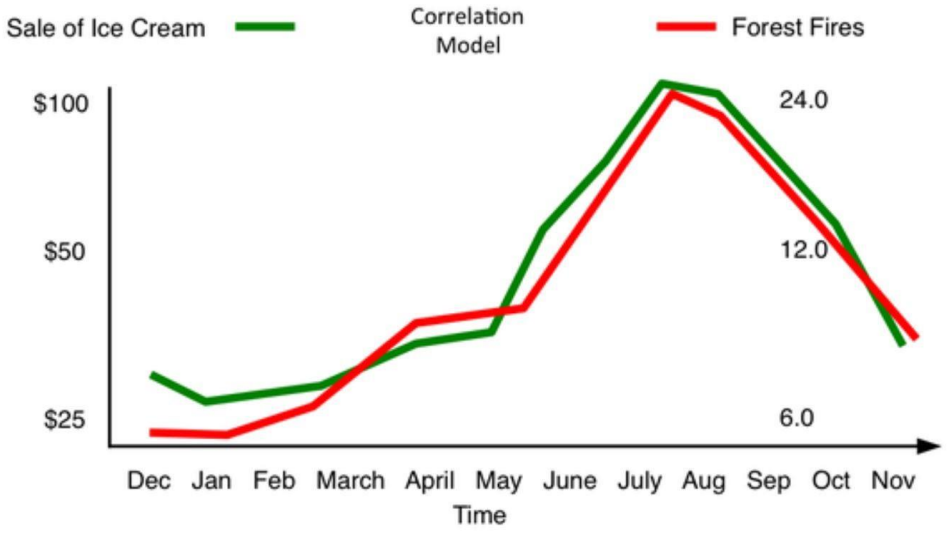
- Correlation does not imply causation!

Why is this statement so important for machine learning?

- Because the **best predictors** of X **are often NOT the causes** of X!

ASSOCIATION AND CAUSATION

Spurious correlations represent events that are associated but which are not causally related.



Images taken from www.kdnuggets.com/2019/09/risk-ai-big-data.html

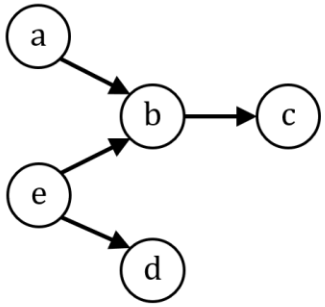
- Can ice cream sales predict forest fires or shark attacks?
- **Yes!**
- Can we intervene on ice cream sales to manipulate the outcomes of forest fire or shark attacks?
- **No!**

CAUSAL MACHINE LEARNING

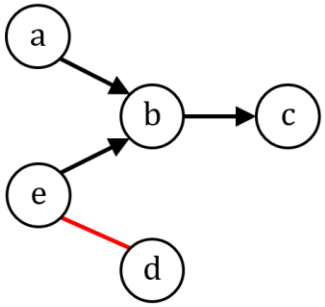
CAUSAL MACHINE LEARNING

- The ML field that focuses on learning **causal relationships** or some form of **causal representation** from data.

- Different works fall within this area of research.
  - Produce different kind of graphs: DAG, DCG, PAG, MAG, PDAG, CPDAG…



DAG



CPDAG

  - Each graph can be converted into a model (with some uncertainty), such as a BN, a CBN, an ID, SCM, SEMs, Markov network, etc.

- Our work primarily focuses on two well-established classes of **unsupervised learning**, known as **constraint-based** and **score-based** learning, and on (causal?) Bayesian networks.

SUPERVISED AND UNSUPERVISED LEARNING

Supervised learning:

Gender	Age	Work	Motivation for treatment	Positive life goals	Uncooperativeness	Anger	Delusions	Anxiety	Self control	Procriminal attitudes	Prior serious offences	Length of stay as inpatient	Violence at 6 months after release
Female	52	No	Yes	N/A	Partly	Partly	Partly	No	No	No	4	55	No
Female	46	Yes	Yes	Yes	No	No	Partly	Partly	Yes	No	0	78	Yes
Female	31	No	Yes	Yes	No	No	No	Partly	Yes	No	1	7	No
Male	27	Yes	No	Yes	Partly	Yes	No	No	Yes	Yes	3	5	No
Male	23	No	No	N/A	Yes	Partly	Yes	Partly	No	Yes	2	7	Yes
Male	51	No	Yes	N/A	Partly	No	Partly	Yes	No	No	5	34	No

Features x

Target variable y

Unsupervised learning:

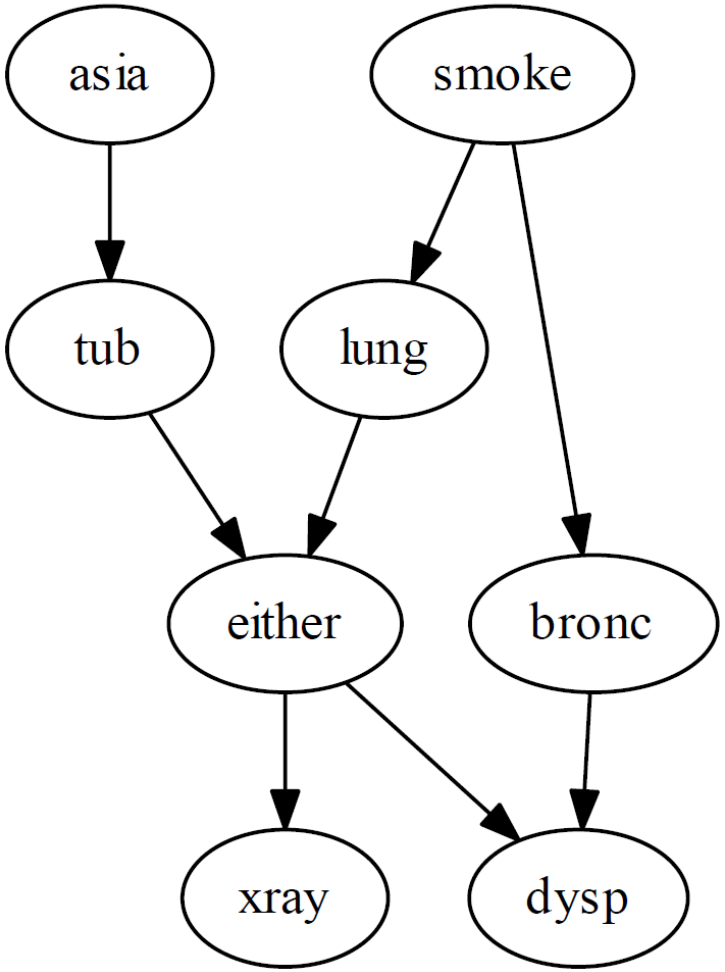
Usually do no data labels, but **we expect to have these labels** in the case of causal structure learning.

Gender	Age	Work	Motivation for treatment	Positive life goals	Uncooperativeness	Anger	Delusions	Anxiety	Self control	Procriminal attitudes	Prior serious offences	Length of stay as inpatient	Violence at 6 months after release
Female	52	No	Yes	N/A	Partly	Partly	Partly	No	No	No	4	55	No
Female	46	Yes	Yes	Yes	No	No	Partly	Partly	Yes	No	0	78	Yes
Female	31	No	Yes	Yes	No	No	No	Partly	Yes	No	1	7	No
Male	27	Yes	No	Yes	Partly	Yes	No	No	Yes	Yes	3	5	No
Male	23	No	No	N/A	Yes	Partly	Yes	Partly	No	Yes	2	7	Yes
Male	51	No	Yes	N/A	Partly	No	Partly	Yes	No	No	5	34	No

Entire data set is give as input without specifying x and y

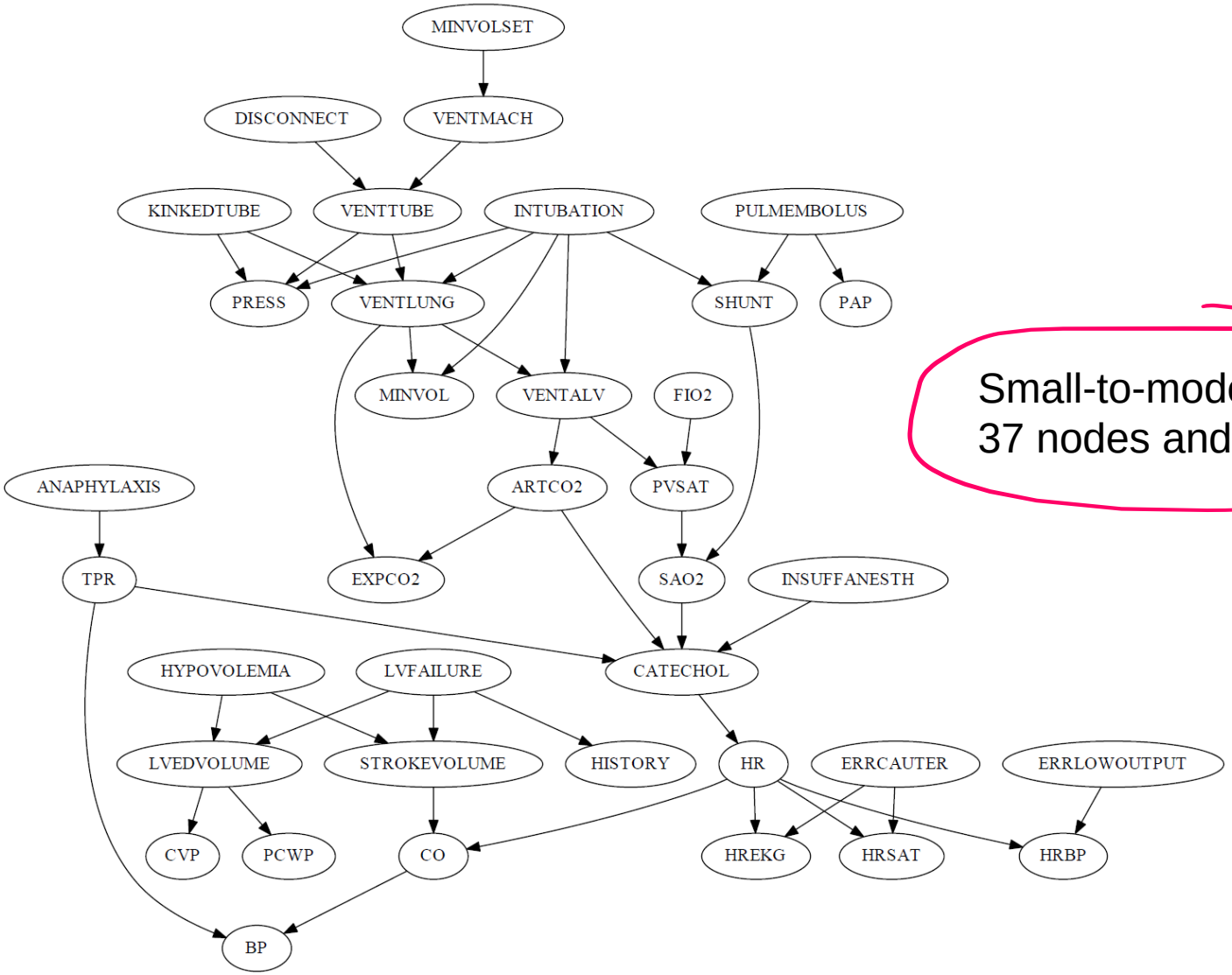
There is no target variable y to predict

POSSIBLE OUTPUT OF A STRUCTURE LEARNING ALGORITHM: THE ASIA NETWORK



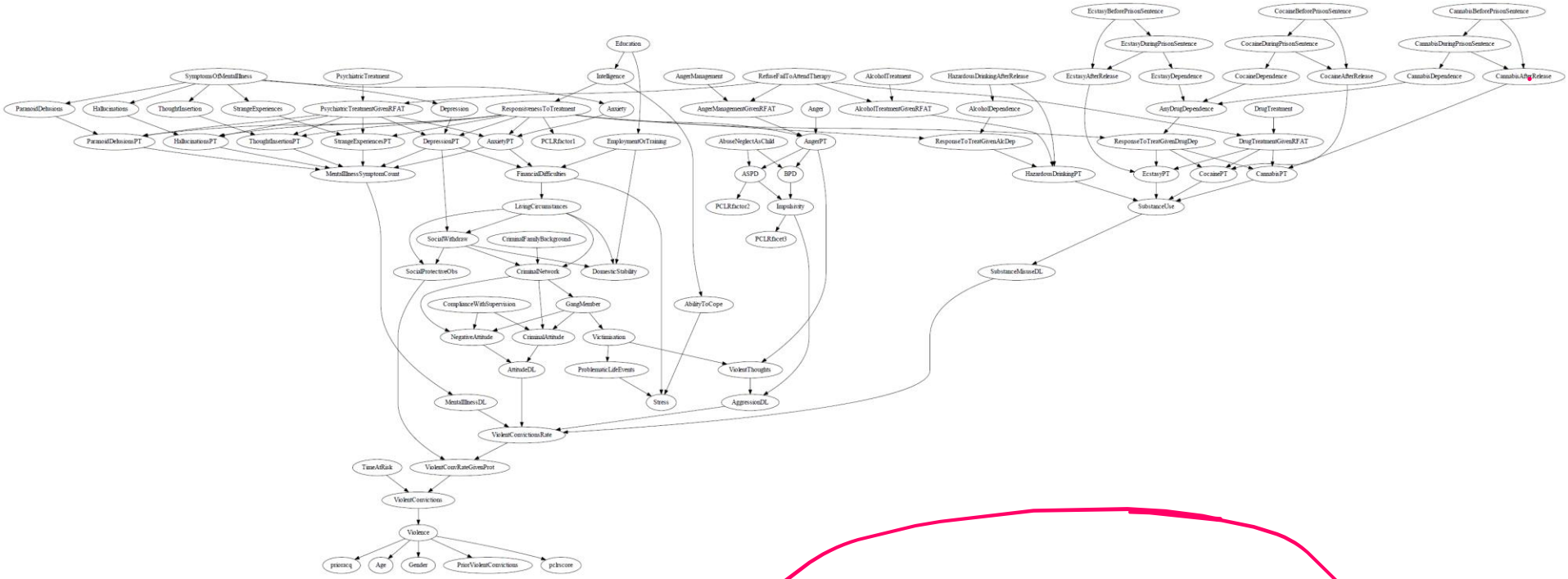
Small size network: 8 nodes and 8 edges.

POSSIBLE OUTPUT OF A STRUCTURE LEARNING ALGORITHM: THE ALARM NETWORK



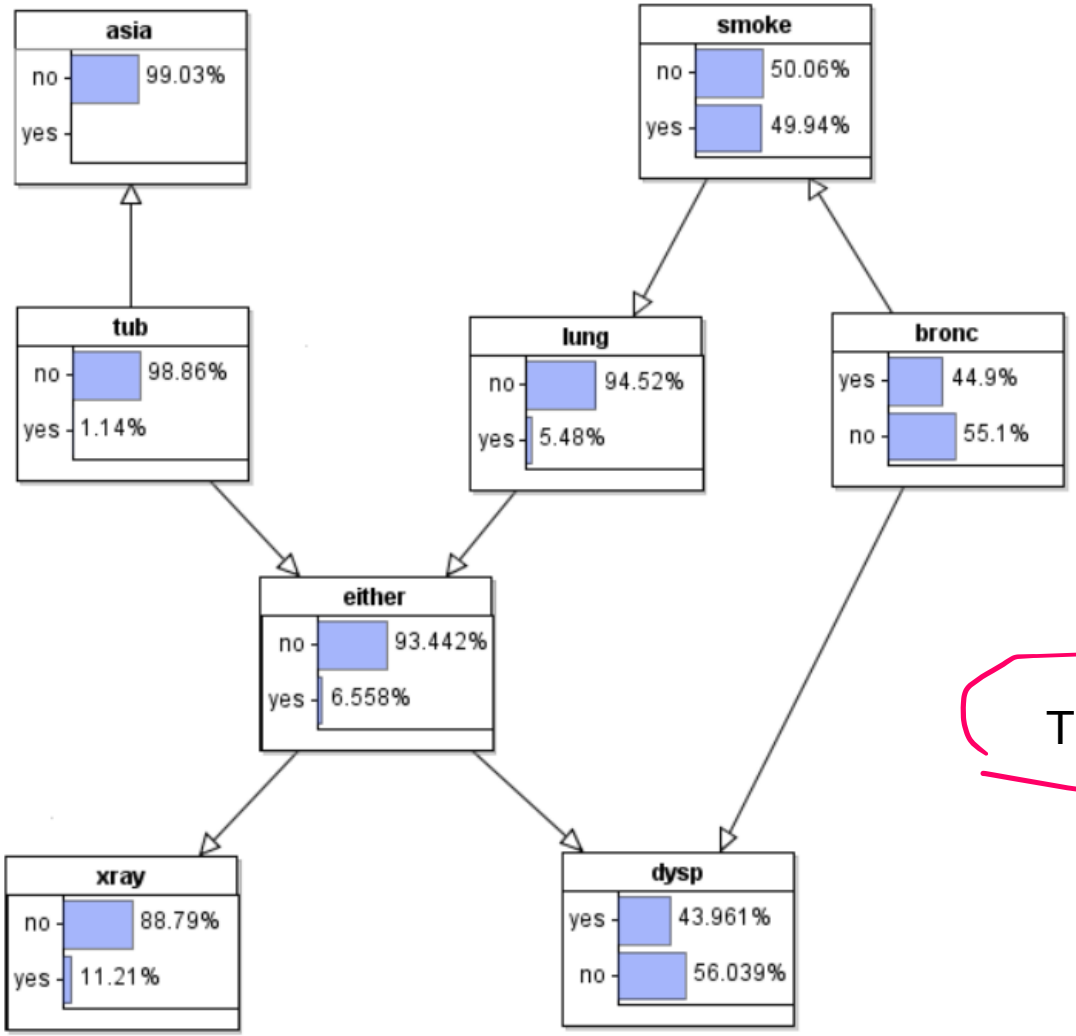
Small-to-moderate size network:
37 nodes and 46 edges.

POSSIBLE OUTPUT OF A STRUCTURE LEARNING ALGORITHM: THE FORMED NETWORK



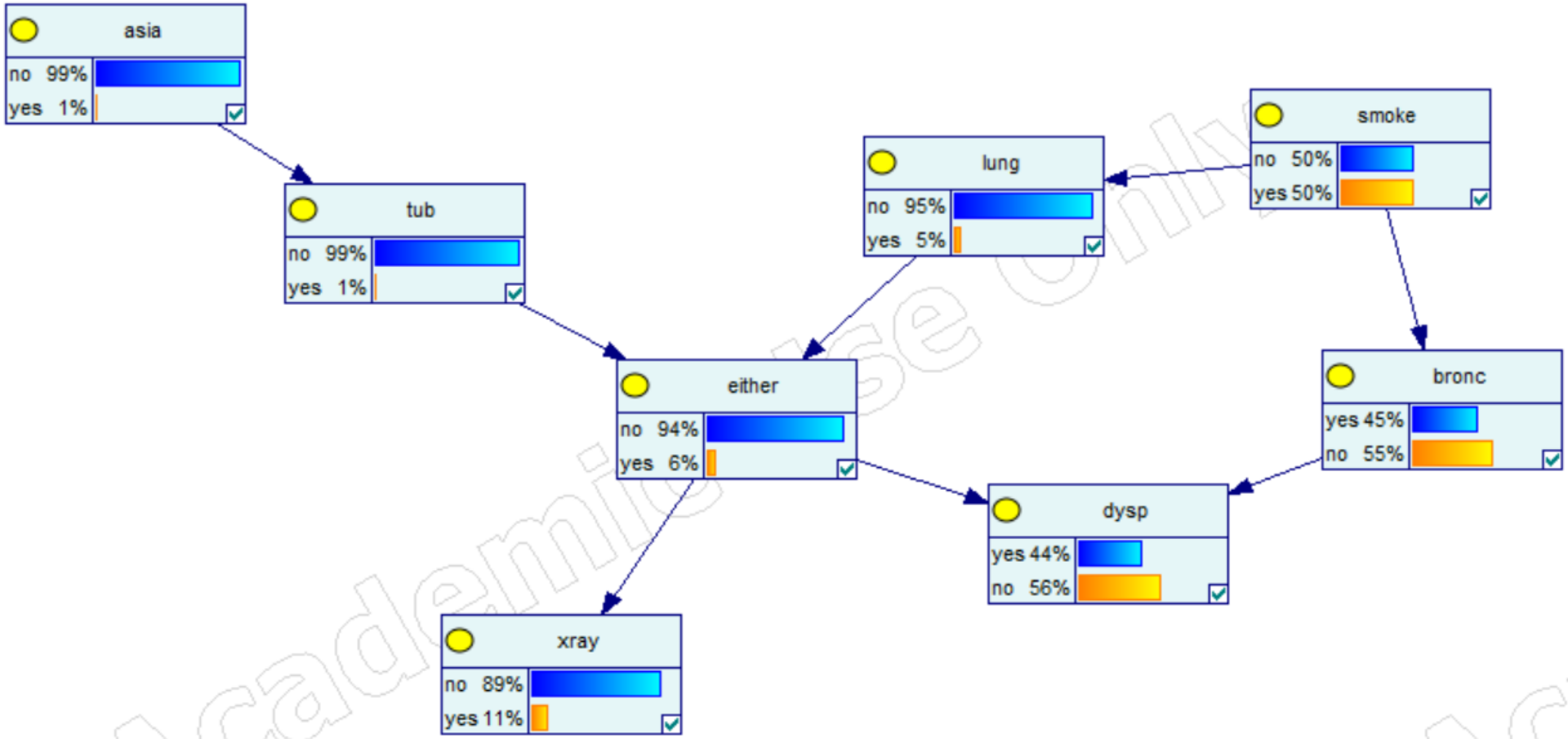
Moderate size network: 88 nodes and 138 edges.

GRAPHS CAN BE CONVERTED INTO MODELS: A BAYESIAN NETWORK



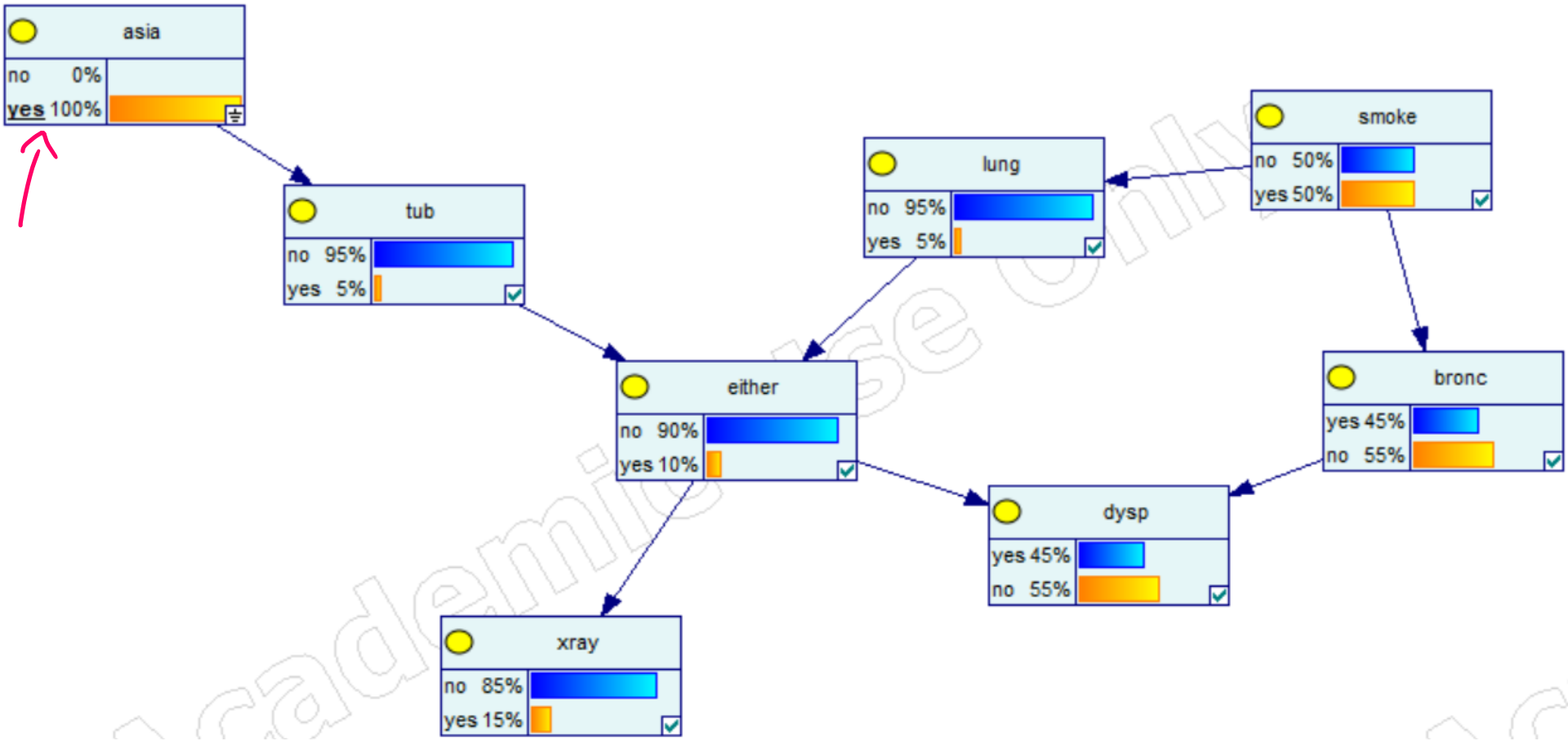
The Asia BN in AgenaRisk

GRAPHS CAN BE CONVERTED INTO MODELS: A BAYESIAN NETWORK



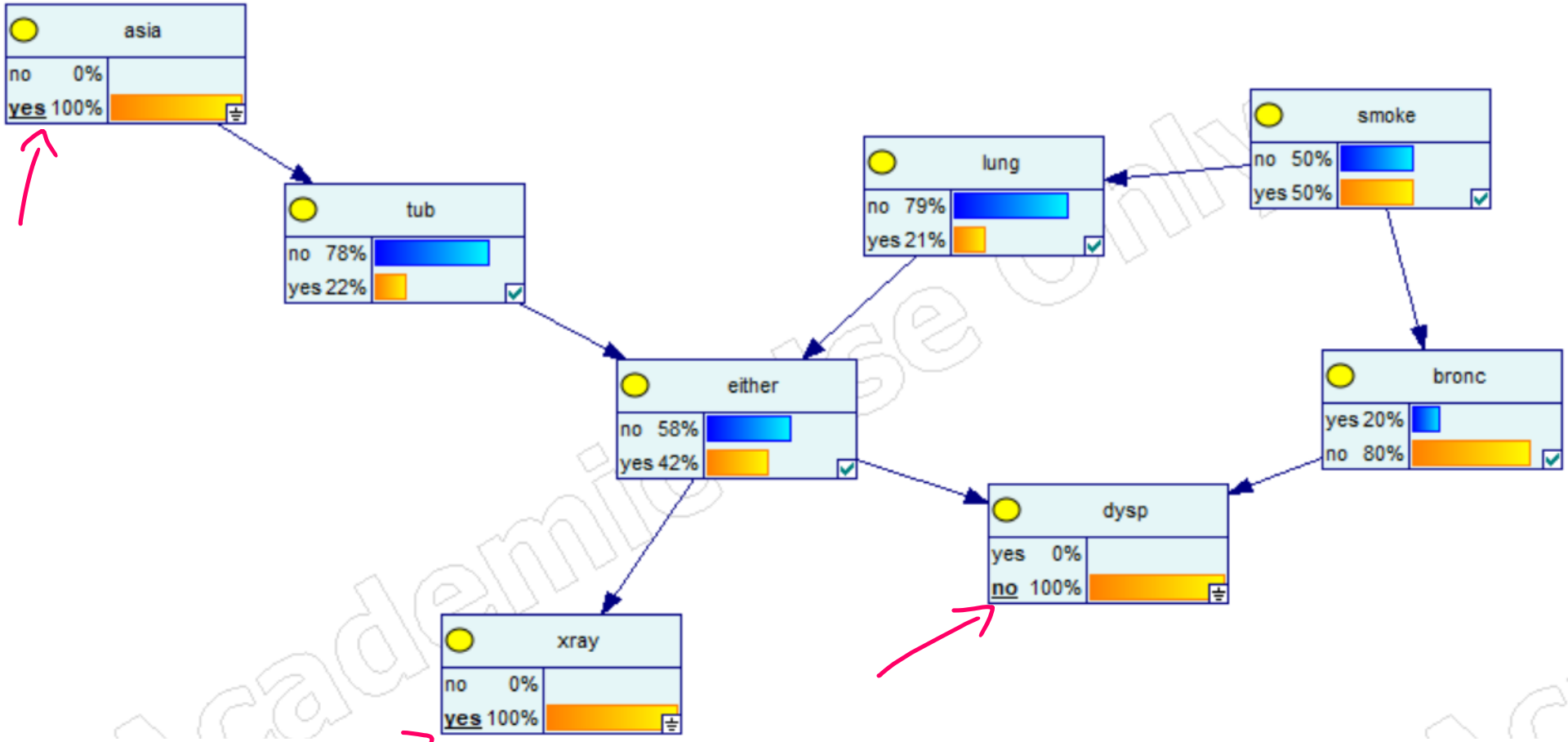
The Asia BN in GeNIe

GRAPHS CAN BE CONVERTED INTO MODELS: A BAYESIAN NETWORK



The Asia BN in GeNIe

GRAPHS CAN BE CONVERTED INTO MODELS: A BAYESIAN NETWORK



The Asia BN in GeNIe

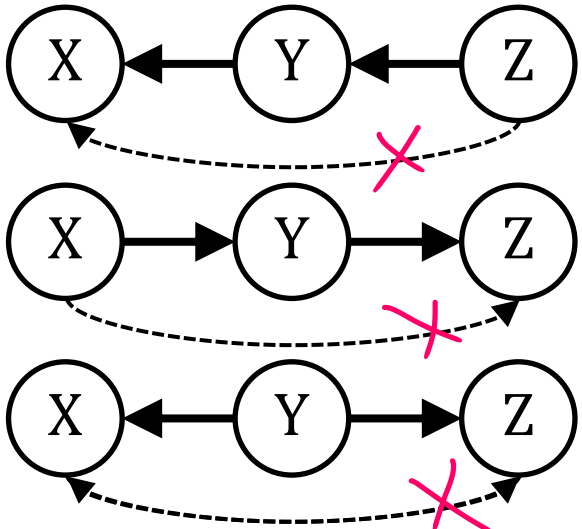
CONSTRAINT-BASED LEARNING

- **Constraint-based:** they return a graph that is consistent with the conditional independencies found in the data.
  - They perform a series of conditional independence and conditional dependence tests, usually in sets of triples.

Conditional independence:

Removes spurious edges.

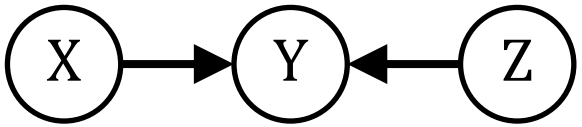
The local graphs that could have produced $X \perp Z \mid Y$



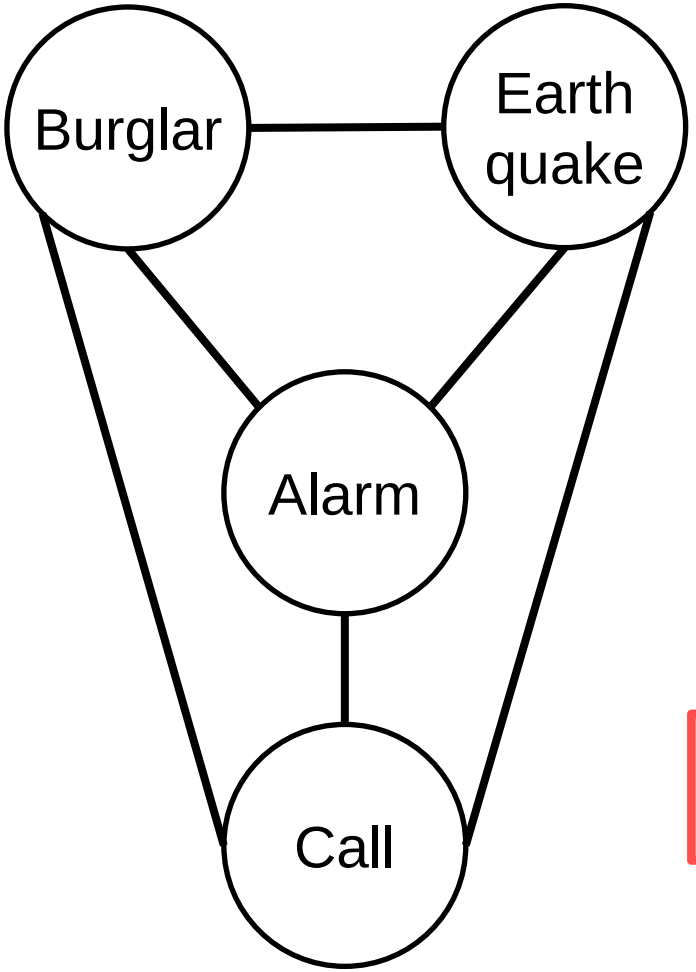
Conditional dependence:

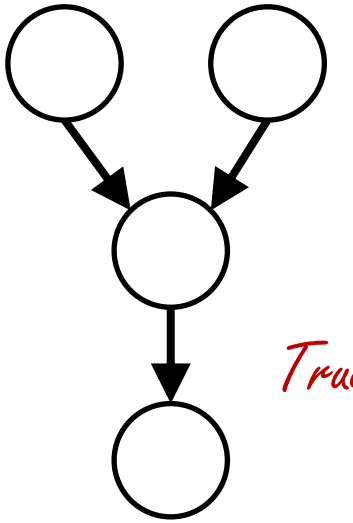
Orientates some of the edges.

The local graph that could have produced $X \top Z \mid Y$.



TYPICAL PROCESS OF CONSTRAINT-BASED LEARNING: THE ALARM NETWORK

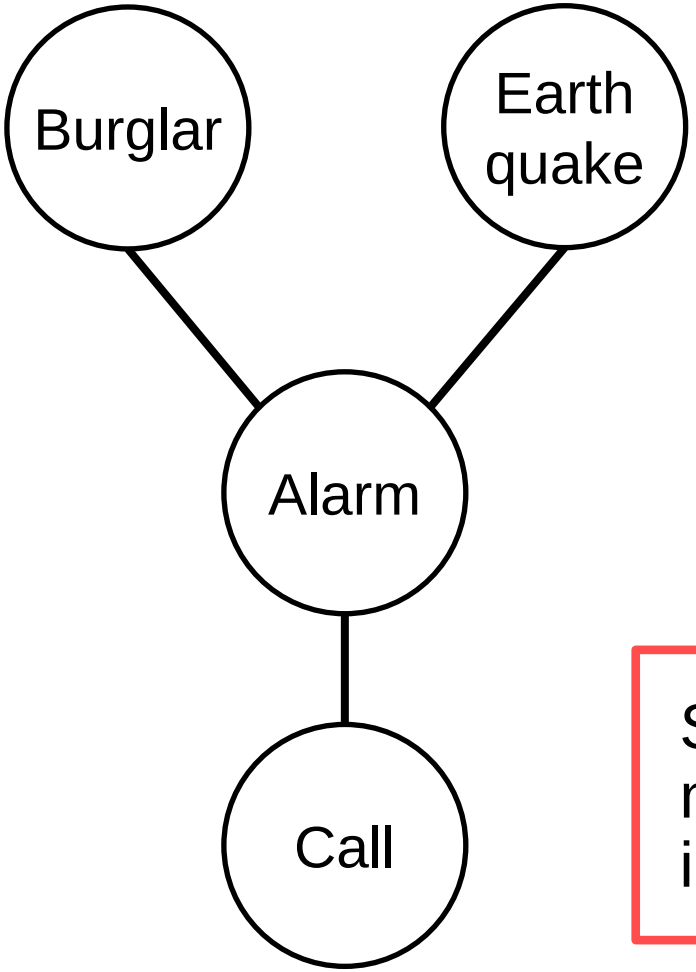


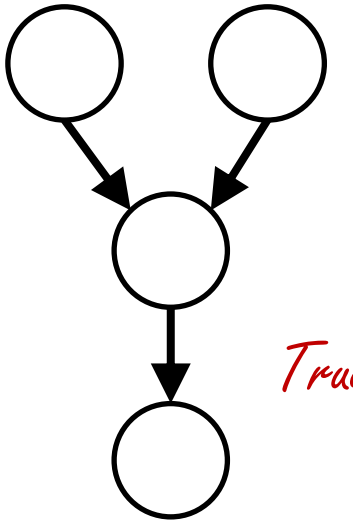


True graph

Step 1: A fully connected graph.

Typical process of constraint-based learning: The alarm network

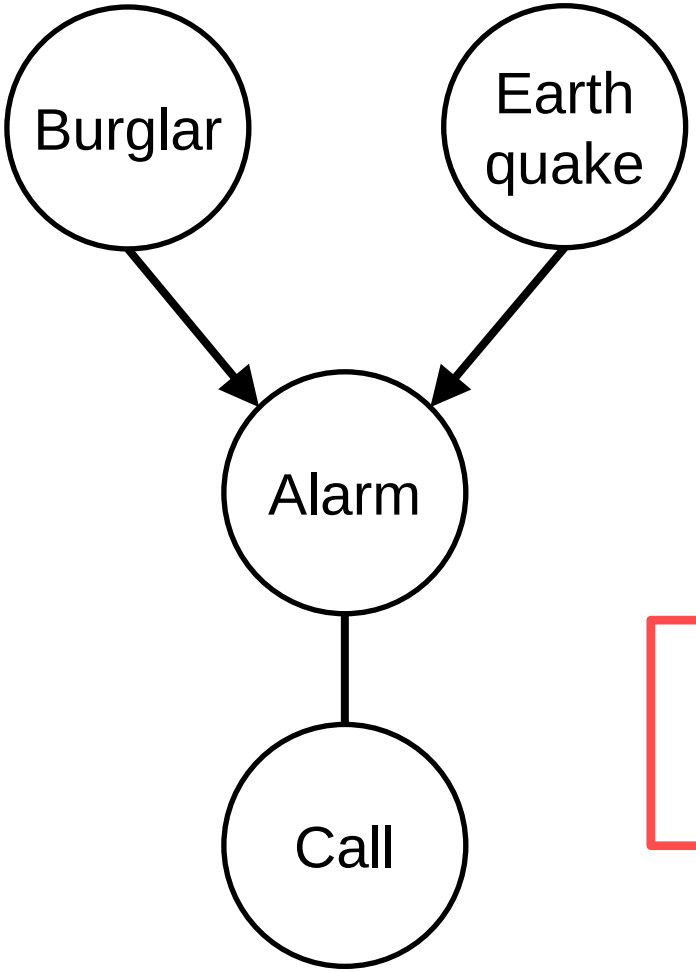


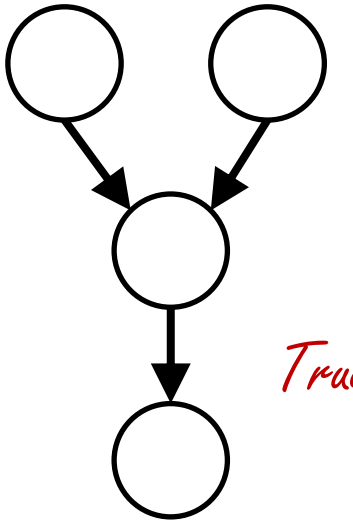


True graph

Step 2: Remove edges based on marginal and conditional independencies.

Typical process of constraint-based learning: The alarm network

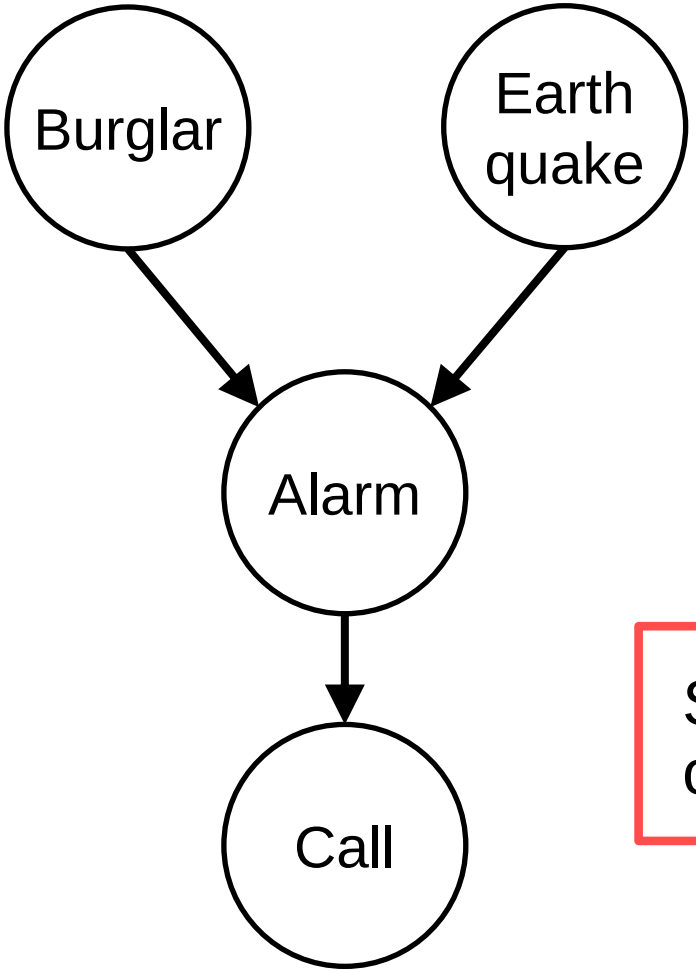


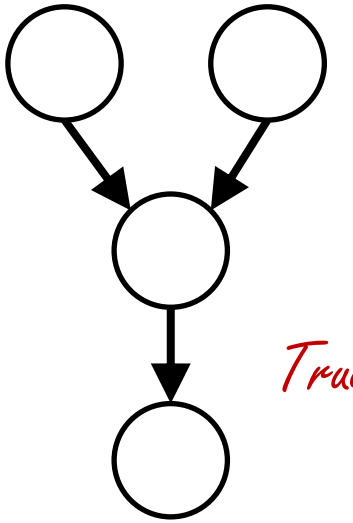


True graph

Step 3: **Orientate edges** based on conditional dependency tests.

Typical process of constraint-based learning: the alarm network





True graph

Step 4: **Orientate edges** based on additional directionality rules.

SCORE-BASED LEARNING

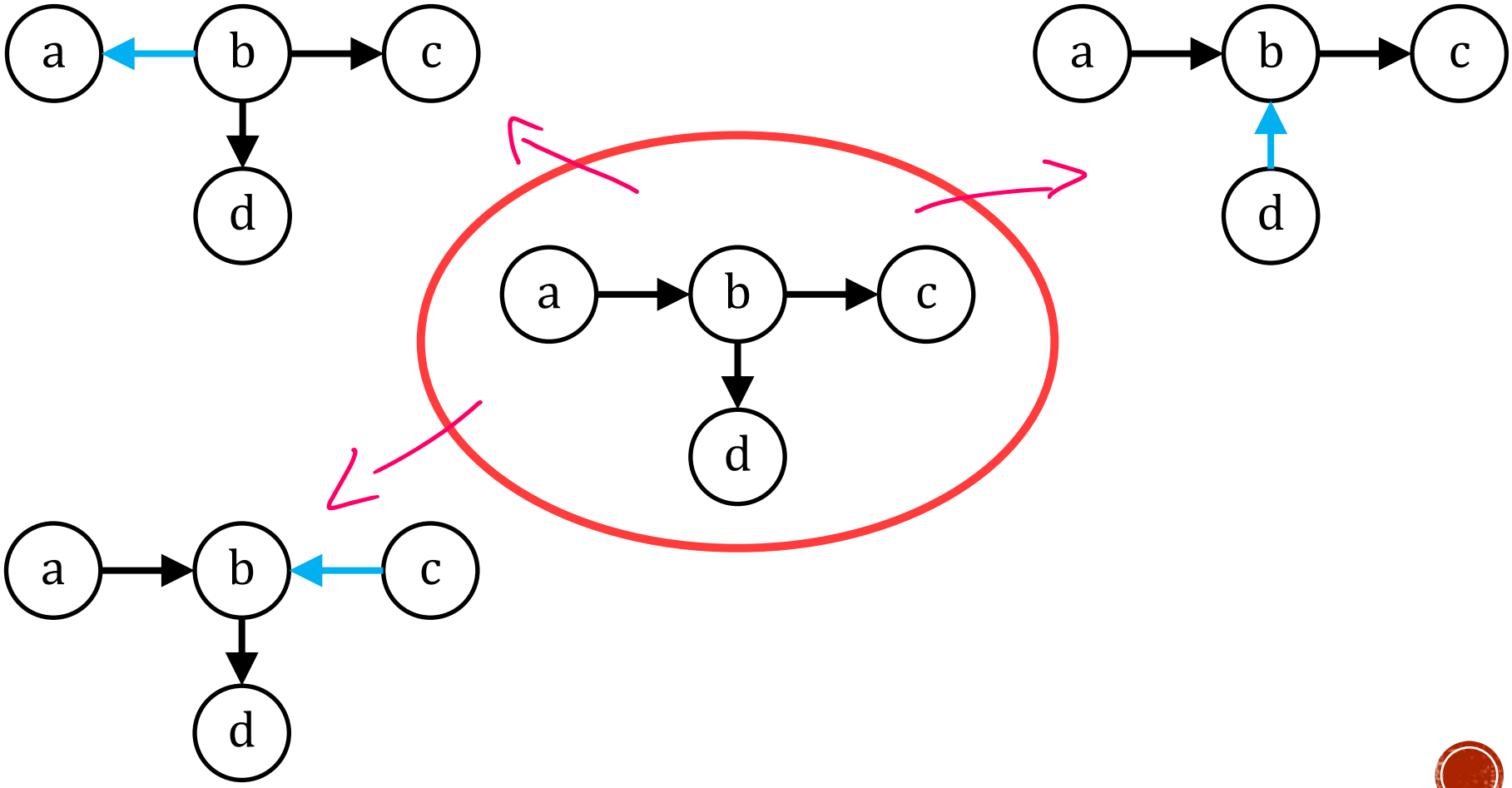
- **Score-based:** Traditional machine learning process that involves:
  - **Search:** to explore the search-space of graphs; e.g., heuristics or pruned combinatorial optimisation.
  - **Score:** to evaluate each graph visited; e.g., BIC, BD/BDe/BDeu

- The solution space of graphs grows super-exponentially with the number of variables.
  - **Exhaustive search** not a practical solution.
  - Algorithms will often **explore well below 1%** of possible graphs; especially in large networks.

Variables	DCGs	DAGs	$\frac{\text{DAGs}}{\text{DCGs}}$
2	3	3	100%
3	27	25	92.59%
4	729	543	74.49%
5	59,049	29,281	49.59%
6	14,349,000	3,781,500	26.35%
7	1.0460×10^{10}	1.1388×10^{9}	10.89%
8	2.2877×10^{13}	7.8730×10^{11}	3.42%
9	1.5009×10^{17}	1.2314×10^{15}	0.81%
10	2.9543×10^{21}	4.1751×10^{18}	0.14%

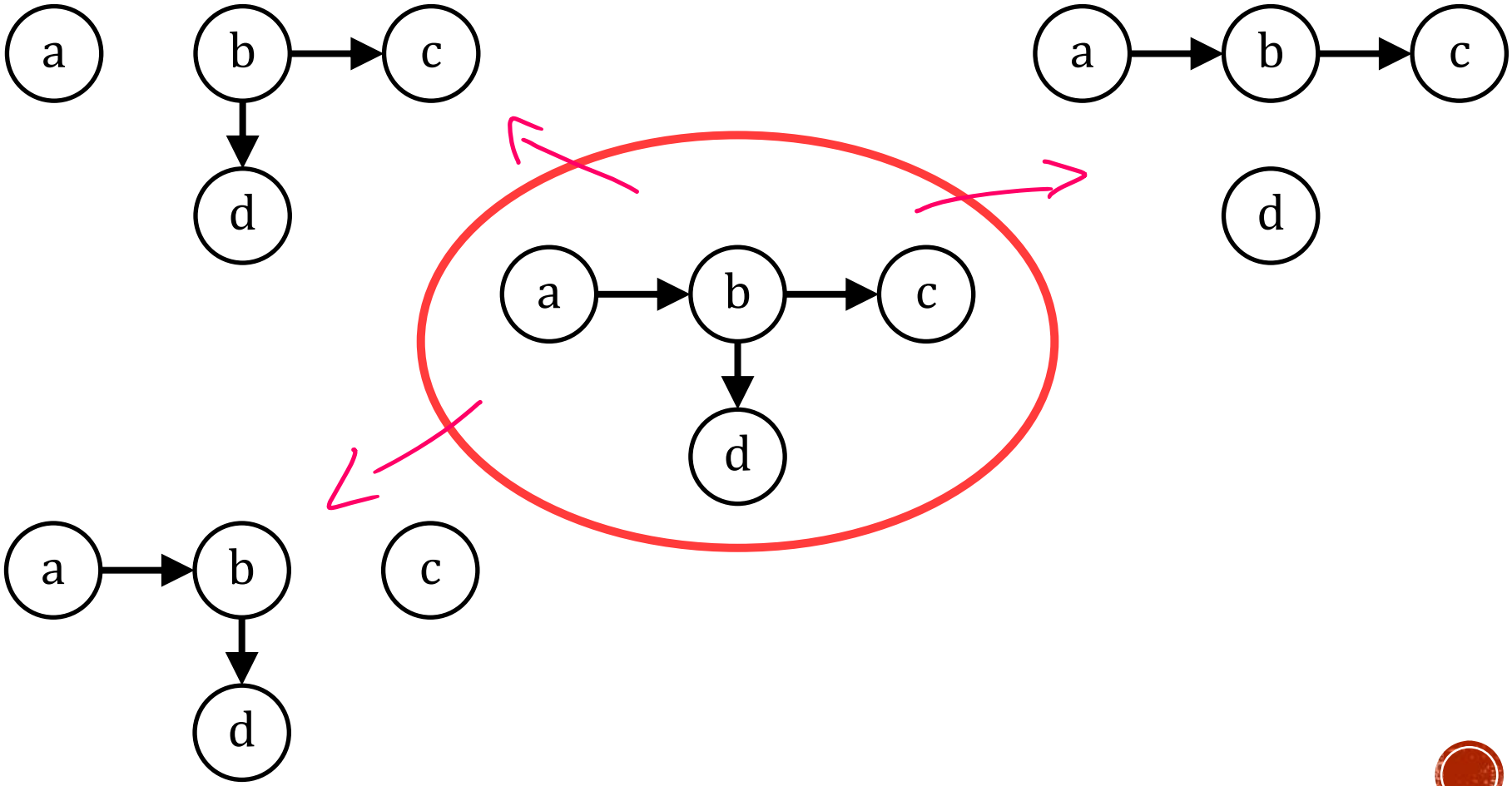
NEIGHBOURING DAGs

The neighbouring DAGs in which an existing edge is **reversed**.



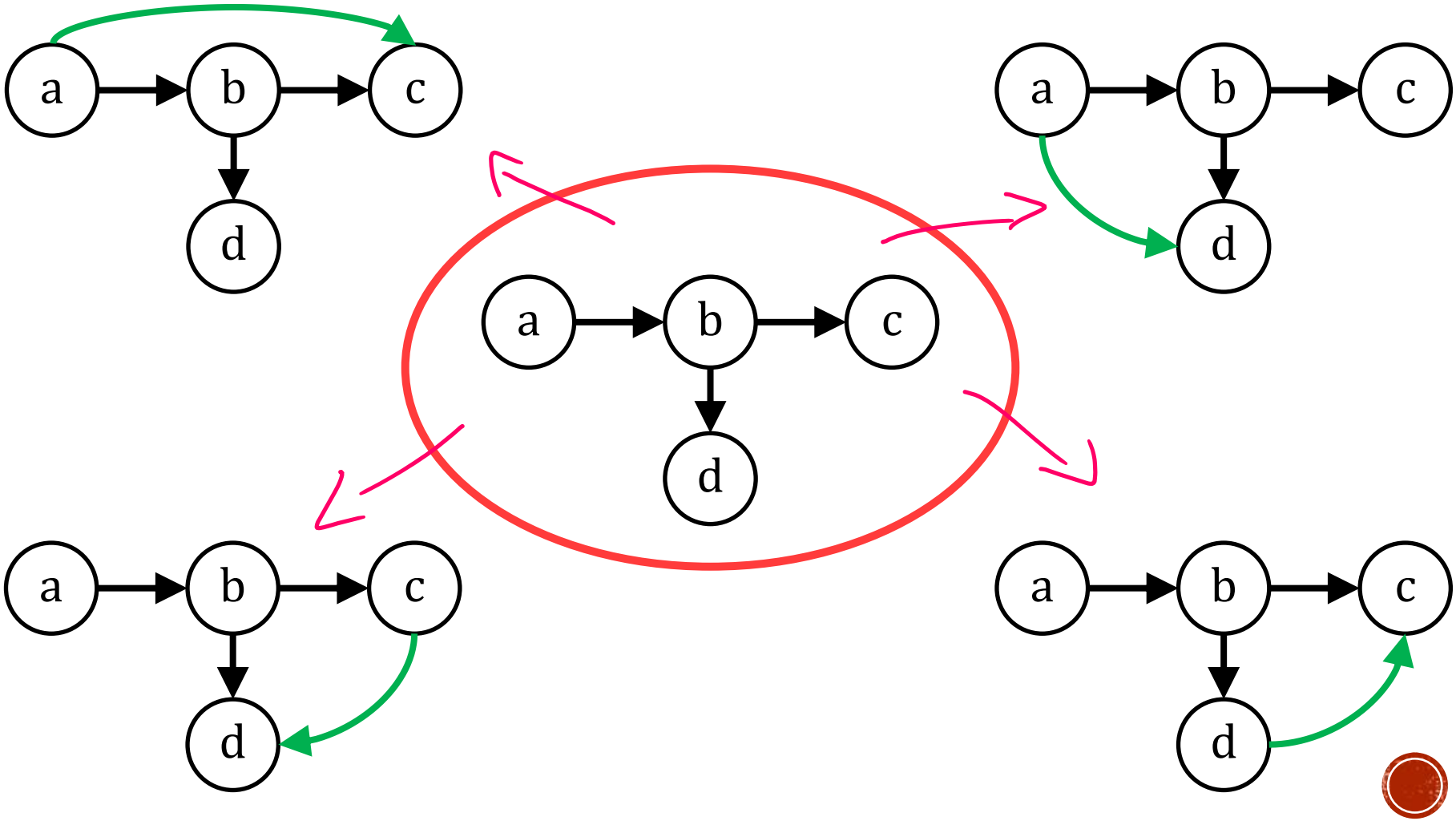
NEIGHBOURING DAGs

The neighbouring DAGs in which an existing edge is **removed**.



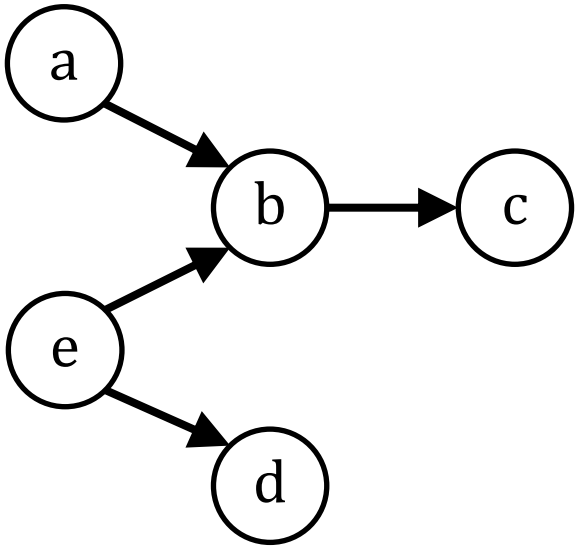
NEIGHBOURING DAGs

Some of the neighbouring DAGs which include an **additional** edge.

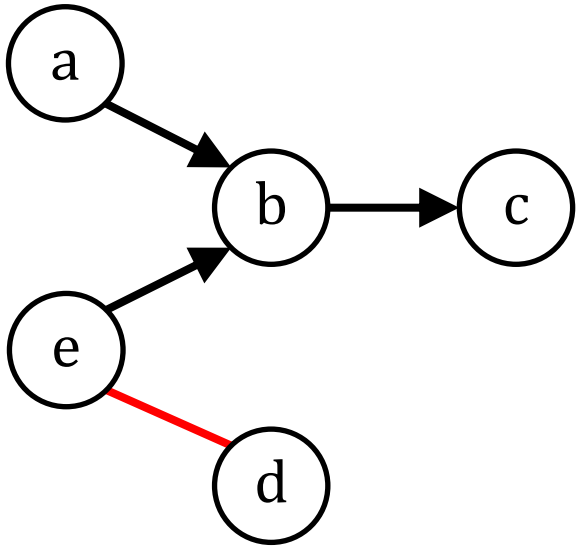


WHY CAUSAL STRUCTURE LEARNING IS DIFFICULT (LIMITATIONS)

EQUIVALENCE CLASSES



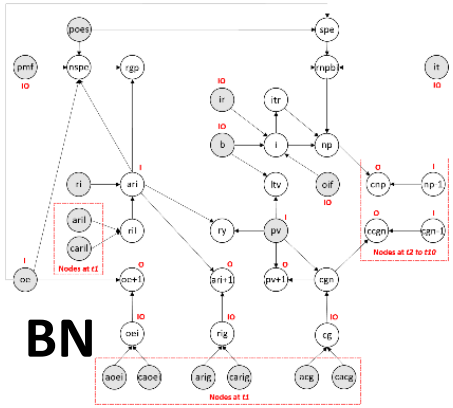
DAG



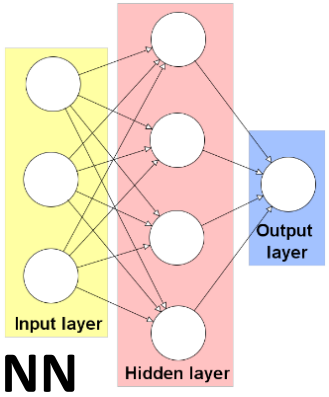
CPDAG

- Algorithms typically rely on score-equivalent functions that produce **Markov equivalence DAGs**.
  - A set of DAGs that encode the **same set of conditional independencies** is represented by a CPDAG.
  - Orientation of such undirected edges **cannot be determined by observational data alone** (unless the temporal order of the variables is given).

CHALLENGES OF STRUCTURE LEARNING:
COMPUTATIONAL COMPLEXITY



Optimising probability distributions across causal structures

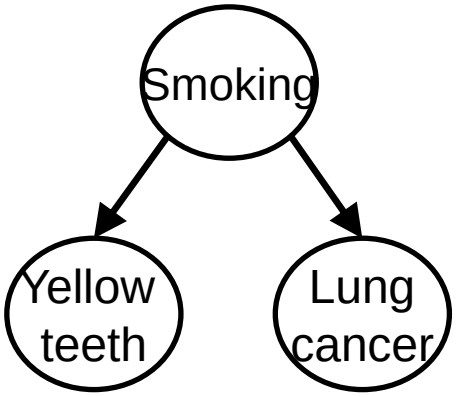


Optimising real numbers across hidden layers

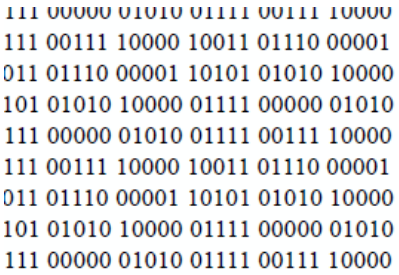
- **Number of graphs** grows super-exponentially with the number of variables.
- Optimising for conditional **probability distributions** further increases complexity.

CHALLENGES OF STRUCTURE LEARNING: SYNTHETIC DATA

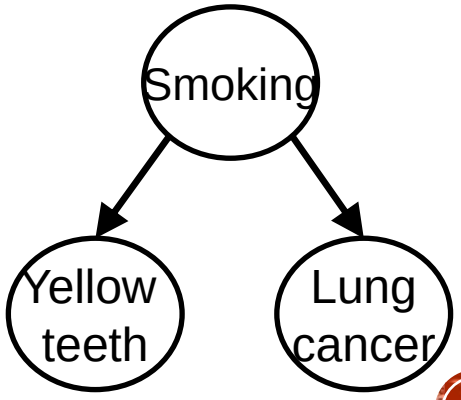
- Generated based on hypothetical models assumed to represent the ground truth.
- Synthetic data are **clean** and adhere to causal representation.
- Algorithms typically evaluated by **reverse-engineering** the synthetic data generating process.
- Unlike synthetic data, real data suffer from numerous known and unknown problems and hence, are **noisy** and do not adhere to causal representation in the same way synthetic data do.



Hypothetical model



Synthetic data



Learnt graph

THE IMPACT OF DATA NOISE ON STRUCTURE LEARNING

Based on (Constantinou et al., 2021):

- Tested **15 algorithms** .
- Considered multiple **case studies**, **sample sizes**, and **evaluation criteria**.
- Considered multiple **types of data noise** to investigate their impact, independently and jointly:
  - Missing values, incorrect values/measurement error, merged states/dimensionality reduction, latent variables and confounders.
- Work involved learning approximately 10,000 graphs with a total structure learning **runtime of seven months**.

Constantinou, A. C., Liu, Y., Chobtham, K., Guo, Z., and Kitson, N. K. (2021). Large-scale empirical validation of Bayesian Network structure learning algorithms with noisy data. *International Journal of Approximate Reasoning*, Vol. 131, pp. 151–188. [Open-access DOI]

THE IMPACT OF DATA NOISE ON STRUCTURE LEARNING

Table
The strengths and weaknesses of the algorithms for each of the categories, based on the empirical results presented in this study, where 0% and 100% represent the weakest and strongest performance for each category.

	Performance									
Algorithm	Ranking	Smaller networks	Larger networks	Limited data	Big data	Variance	Resilience to noise	Under/Over-fitting	Computational speed	Reliability
FCI	46%	35%	61%	55%	33%	0%	50%	31%	52%	99%
FGES	60%	52%	71%	58%	67%	96%	9%	64%	84%	87%
GFCI	60%	52%	66%	58%	63%	83%	0%	62%	86%	93%
GS	16%	1%	0%	0%	16%	75%	47%	0%	100%	100%
H2PC	79%	100%	81%	74%	99%	97%	100%	36%	74%	81%
HC	100%	100%	99%	100%	95%	55%	51%	49%	100%	100%
ILP	82%	98%	81%	93%	83%	43%	39%	62%	63%	77%
Inter-IAMB	37%	26%	41%	31%	45%	29%	60%	29%	100%	100%
MMHC	71%	84%	60%	68%	75%	64%	92%	25%	99%	100%
NOTEARS	0%	0%	0%	29%	0%	100%	28%	100%	100%	100%
PC-Stable	56%	42%	73%	63%	39%	18%	73%	29%	55%	95%
RFCI-BSC	11%	18%	52%	35%	0%	99%	74%	11%	0%	0%
SaiyanH	74%	81%	76%	58%	100%	36%	48%	94%	77%	100%
TABU	99%	93%	100%	95%	95%	78%	43%	51%	100%	100%
WINASOBS	72%	76%	73%	60%	88%	64%	28%	48%	97%	97%

Constantinou, A. C., Liu, Y., Chobtham, K., Guo, Z., and Kitson, N. K. (2021). Large-scale empirical validation of Bayesian Network structure learning algorithms with noisy data. *International Journal of Approximate Reasoning*, Vol. 131, pp. 151–188. [Open-access DOI]

THE IMPACT OF DATA NOISE ON STRUCTURE LEARNING

The results suggest that traditional synthetic performance may overestimate real-world performance by anywhere between **10% and more than 50%**.

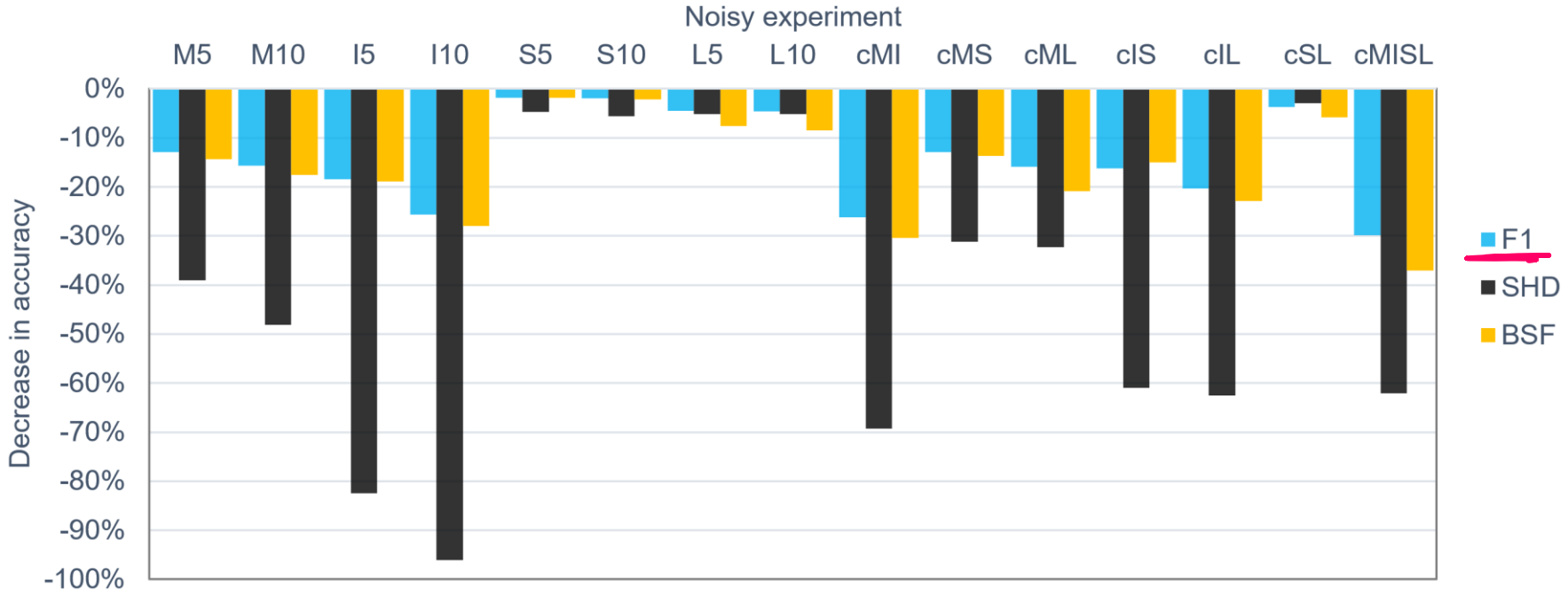


Fig. The overall decrease in accuracy (F1 and BSF), and increase in error (SHD), over all 15 structure learning algorithms and for each type of noise added to the data.

Constantinou, A. C., Liu, Y., Chobtham, K., Guo, Z., and Kitson, N. K. (2021). Large-scale empirical validation of Bayesian Network structure learning algorithms with noisy data. *International Journal of Approximate Reasoning*, Vol. 131, pp. 151–188. [Open-access DOI]

Reliance on Approximate Solutions

Typically the Bayesian Information Criterion (BIC)

- Model selection function that balances model **fitting** with model **dimensionality**.
- The highest BIC scoring graph is *not* the ground truth graph **in the presence of data noise**; e.g., real data.
- Still, it is good at recovering a graph that is close to the ground truth.

What do we conclude from this?

- Finding the global maximum graph **DOES NOT IMPLY** finding **a superior** causal graph.
  - Limited incentive to pursue **exact learning** solutions.
    - Also restricted to smaller graphs.
  - Less computationally expensive **approximate learning** solutions might be a better option?

LEARNING FROM IMPERFECT DATA

- Each algorithm is based on **a set of assumptions** about the input data.
- Almost all real-world data sets are '**imperfect**' for the algorithms.
- **Imperfect data:** when the data set violates algorithmic assumptions.
- Data can be **imperfect** for one algorithm and **perfect** or another algorithm.
- No algorithm takes into consideration all possible data 'imperfections'.

  - Systematic missing data; i.e., when data missingness is not random.
  - Limited data; e.g., BIC assumes sufficient data instances.
  - Causal insufficiency; i.e., latent confounders.
  - Dimensionality reduction leading to information loss.
  - Measurement error; e.g., biased or inaccurate data records.
  - Heterogeneous data; i.e., high variability of data types and formats.
  - Ordinal or nominal data; e.g., low/medium/high vs red/blue/green.
  - Distributional assumptions; e.g., normally distributed data.
  - Observational/experimental data.
  - Non-stationary distributions; e.g., distribution shift over time.
  - Time-varying causality; i.e., when causal relationships change over time.

CAUSATION WITHOUT CORRELATION

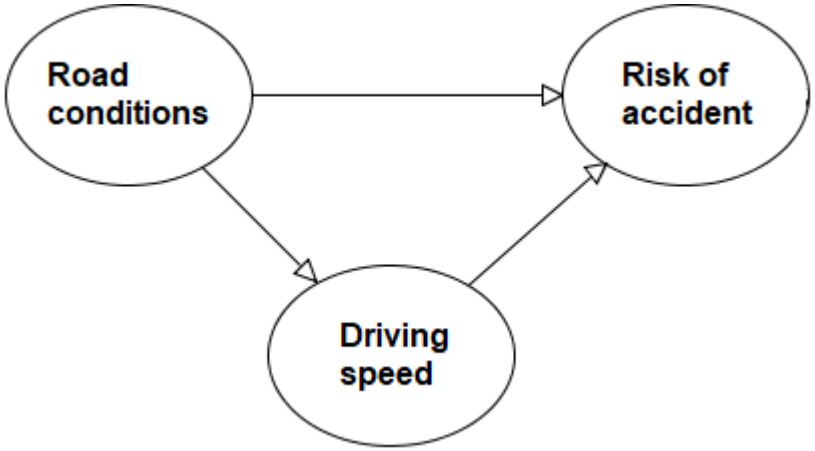
- Assume the variables are causally related:

  a) road conditions influence risk of accident,

  b) road conditions influence driving speed,

  c) driving speed influences risk of accident.

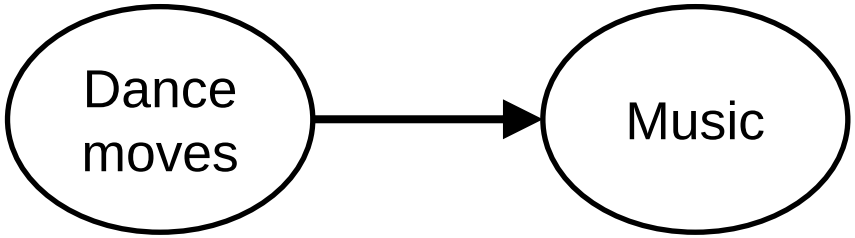
- It is possible to observe no correlation between **Road conditions** and **Risk of accident**.
  - E.g., poor road conditions increase the risk of accident, but they also decrease driving speed which in turn decreases the risk of accident.



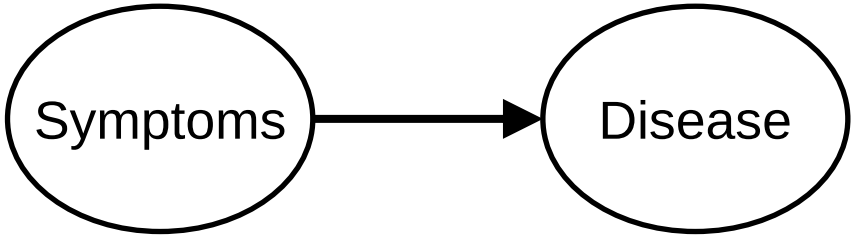
"*…the **best predictors** of Risk of accident **are often NOT the causes** of Risk of accident!*"

COMMON SENSE FAILURES

- An algorithm may discover that the music is influenced **by the moves of the dancer**, rather than concluding that the **dancer is dancing to the music**.



  - **Can we intervene on 'Dance moves' to manipulate 'Music' ?**



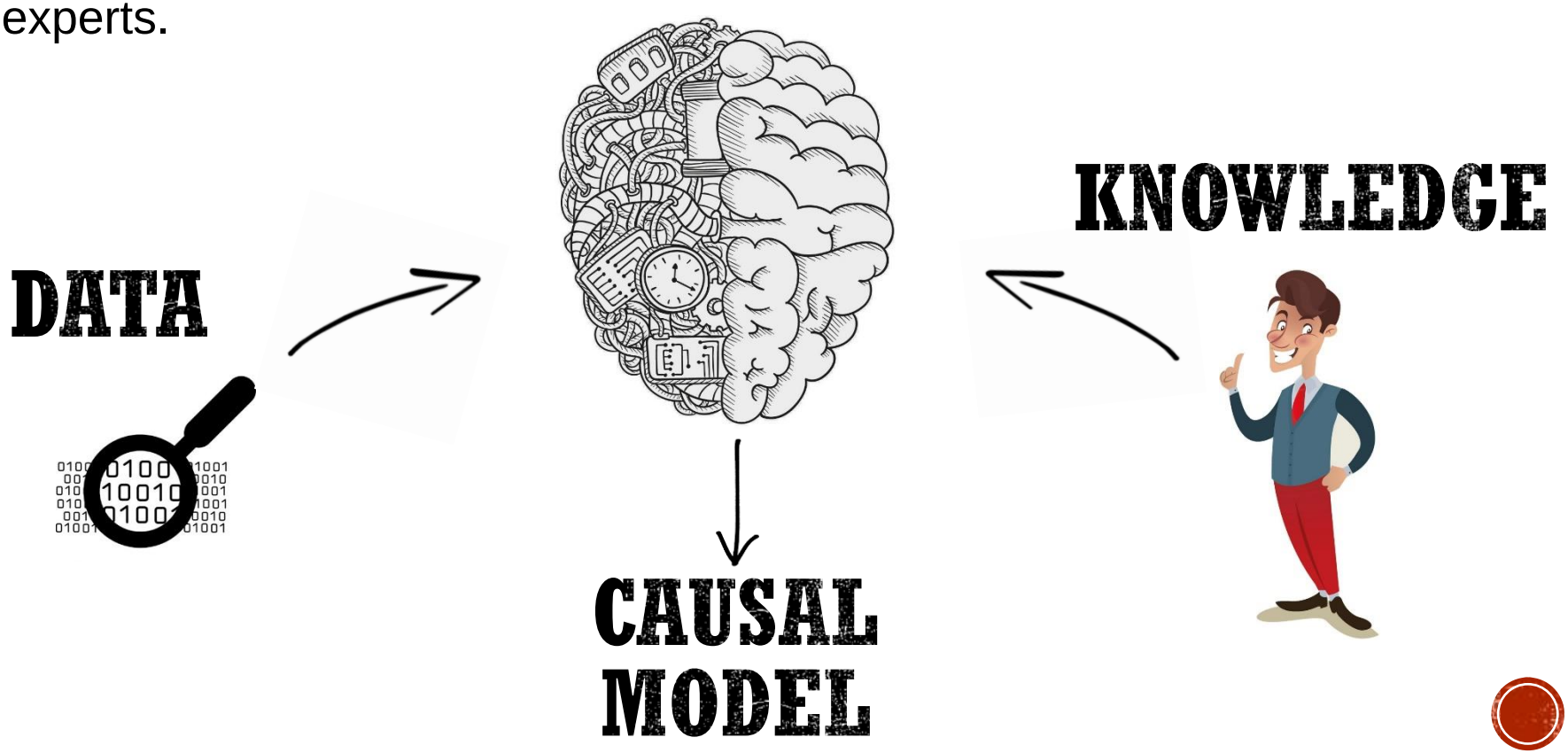
  - **Can we intervene on 'Symptoms' to manipulate 'Disease' ?**

- Such counterintuitive relationships are **viewed as failures of causal common-sense** and raise questions as to whether machine learning is capable of achieving human level causal understanding.
  - **Is this true?** ML has historically been poor at some common-sense tasks and good at some tasks that are difficult for humans.

CAUSAL KNOWLEDGE

One solution is to guide algorithms using causal knowledge.

- **Benefits:** Effective use of causal knowledge helps algorithms **avoid common sense errors**.
- **Limitations:** bias, implementation effort, elicitation effort, cost, access to experts.

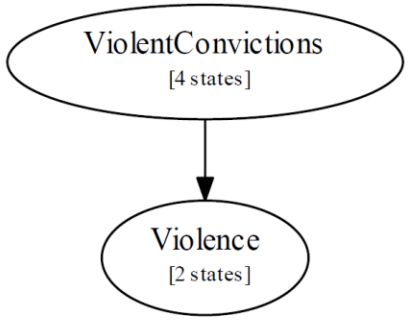


KNOWLEDGE CONSTRAINTS

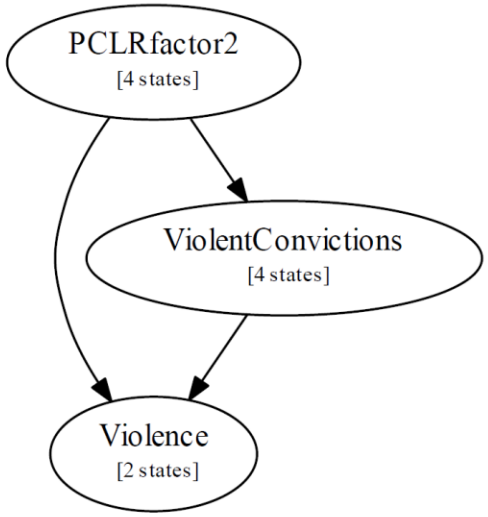
Approach	Input example	Knowledge	Constrain or guide
Directed edge	$A \rightarrow B$	Directed dependency (or causal relationship).	Constrains search space to those containing $A \rightarrow B$.
Undirected edge	$A - B$	Dependency without knowledge of causal direction.	Constrains search space to those containing $A \rightarrow B$ or $A \leftarrow B$.
Forbidden edge	$A \perp B$	No dependency.	Constrains search space to those *not* containing $A \rightarrow B$ and $A \leftarrow B$.
Temporal order	Tier 1: $\{A\}$ Tier 2: $\{B, C\}$	B and C occur after observing A and hence, B and C cannot be parents nor ancestors of A.	Constrains search space to those not containing $A \leftarrow B$, $A \leftarrow C$, or B and/or C as ancestors of A.
Initial best-guess graph	DAG	An initial best guess graph	Sets the starting point in the search space of graphs to the initial best-guess graph.
Variables are relevant	n/a	All variables in the input data are relevant.	The learnt graph must not contain disjoint subgraphs or unconnected nodes.
Target nodes	A node A or a set of nodes $\{A, B\}$	Variable/s targeted for identification of more causes.	Relaxes the dimensionality penalty in BIC for targeted variables.

Constantinou, A. C., Guo, Z., and Kitson, N. K. (2021). Information fusion between knowledge and data in Bayesian network structure learning. [arXiv:2102.00473] [cs.AI]

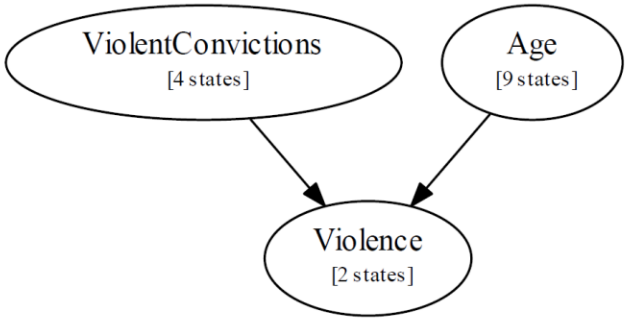
EXAMPLE OF KNOWLEDGE APPROACH 'TARGET NODE' (NODE VIOLENCE)



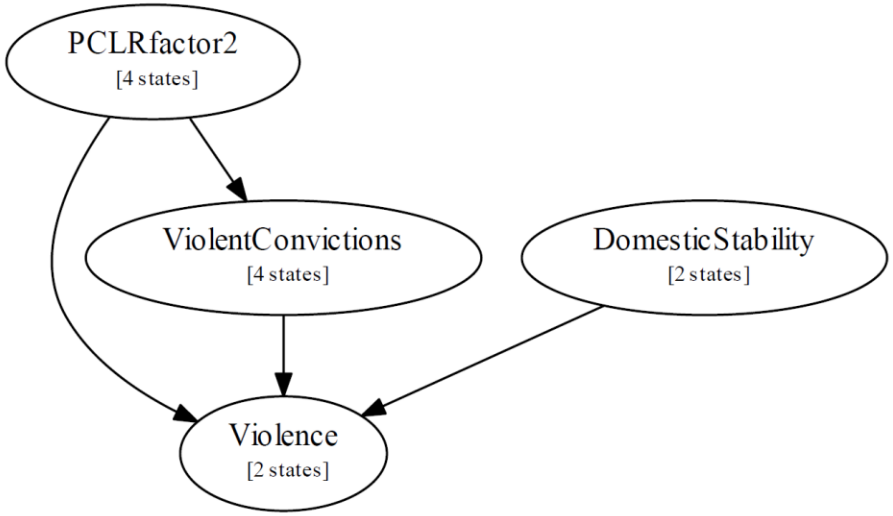
standard



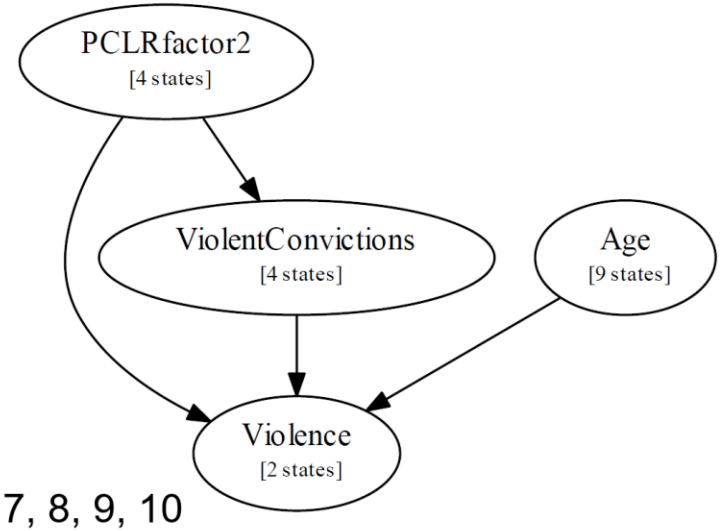
r = 2



r = 3, 5, 6



r = 4



r = 7, 8, 9, 10

CHALLENGES OF STRUCTURE LEARNING: EVALUATION

- There is **no agreed evaluation process**.
- Korb and Nicholson (2011) state that "*every publication in the field attempts to make some kind of empirical case for the particular algorithm being described in that publication*".
  - **Graph-based metrics:** Precision, Recall, F1, SHD, BSF, SID, etc.
  - **Inference-based metrics:** LL, BIC, BD/BDe/BDeu, any other objective function.
- Different evaluation methods lead to **inconsistencies** whereby one evaluator determines algorithm A to be superior to algorithm B, whereas another evaluator concludes the opposite.
- **Solution** (requires effort)**:** Evaluate algorithms across different metrics, case studies, sample sizes, data settings, hyperparameters.

Korb, K., & Nicholson, A. (2011). *Bayesian Artificial Intelligence* (Second Edition). CRC Press, London, UK

CAUSAL MODEL CONSTRUCTION

- Causal modelling has evolved in **different directions**.
  - E.g., causal ML algorithms, expert systems, eliciting causal knowledge, combining knowledge with data, causal data pre-processing, dealing with latent confounders, randomised control trials for causal interventions, etc.

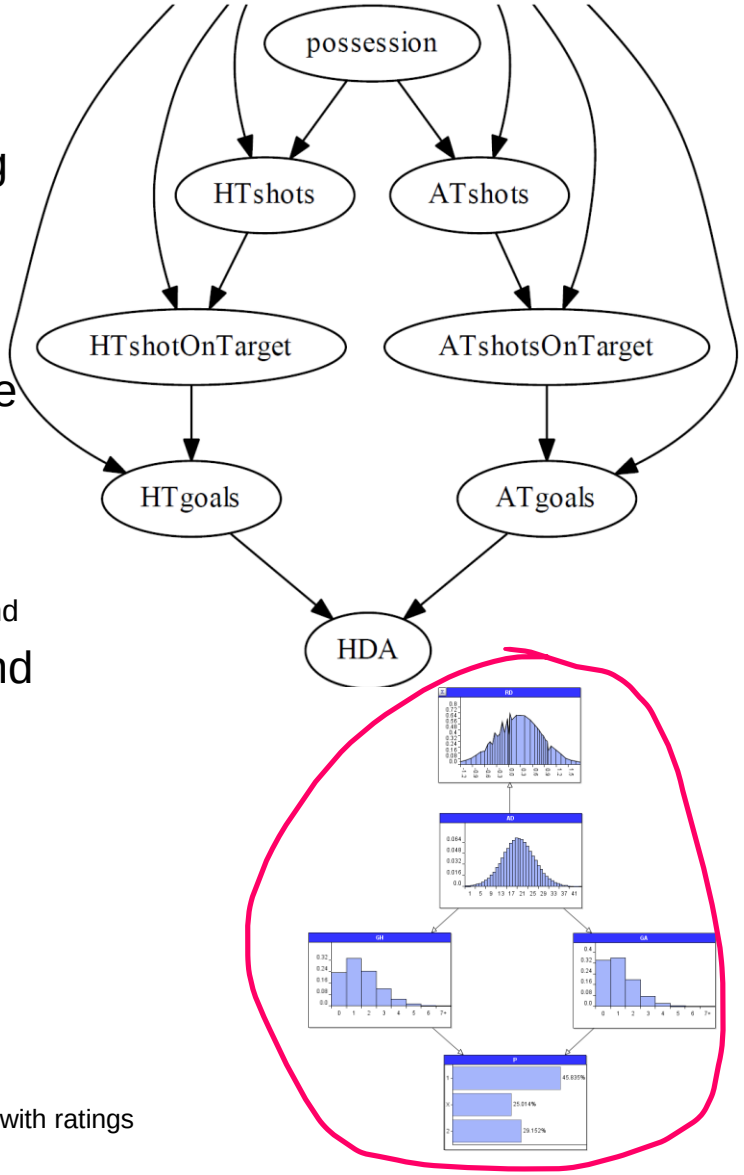
- Most of these research directions have **evolved independently** with little interaction between them.

- Building causal models requires solutions available in different **directions**, often coming from different **disciplines** that rely on different **terminology**, implemented in different **programming languages** and **statistical packages**, some of which will be **open-source** and others based on industry software and **proprietary technology**.

APPLIED WORK

CAUSAL MODELS FOR BETTING MARKET EFFICIENCY AND FOOTBALL PREDICTION

- Assessment of the **efficiency** of the Asian handicap betting market (Constantinou, 2021)
  - Is it possible for automated betting decision making models to beat the market?
  - Asian handicap introduces a **hypothetical score advantage** in favour of one team.
  - Graphical structure determined by knowledge of the natural temporal chain of events: **Possession** → **Shots** → **Shots on Target** → **Goals scored**.
- Football prediction competition (Constantinou, 2018):
  - Hosted by the *Machine Learning* journal (ranked 2nd with a predictive error 0.94% higher than the top and 116.78% lower than the bottom participants).
  - Involved 52 football leagues worldwide.



Time-series data an issue: Had to combine BNs with rating systems in both studies.

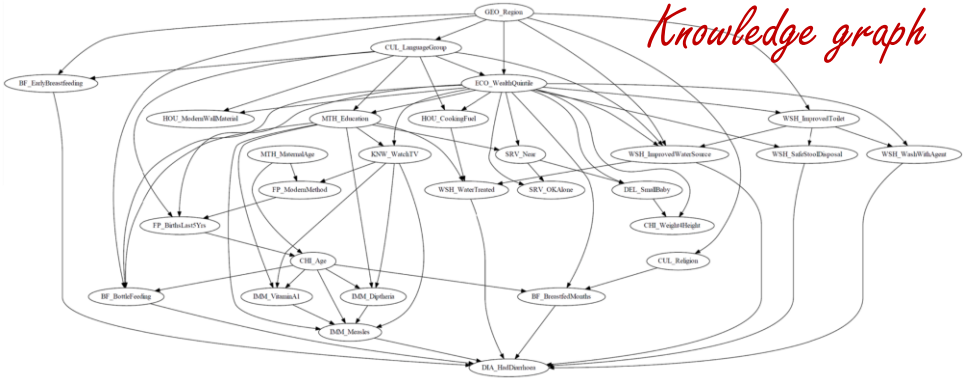
Constantinou, A. (2018). Dolores: A model that predicts football match outcomes from all over the world. *Machine Learning*, pp. 1–27. [Free view, DOI]

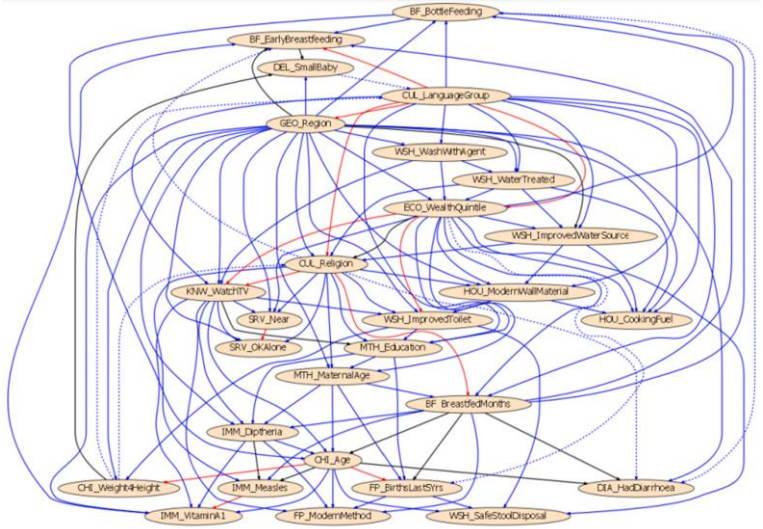
Constantinou, A. (2021). Investigating the efficiency of the Asian handicap football betting market with ratings and Bayesian networks. *Journal of Sports Analytics*, TBA [Open-access DOI]

LEARNING BAYESIAN NETWORKS FROM DEMOGRAPHIC AND HEALTH SURVEY DATA

- Investigated the factors associated with **childhood diarrhoea** in India.
- **Demographic and health survey data** is modelled using BNs.
- Uses structure learning and knowledge to construct BN graphs.
- Found **large variations** in the graphs learnt by the different algorithms explored.
- **Applied knowledge** to the learning process to reduce variation between algorithms.

Knowledge graph



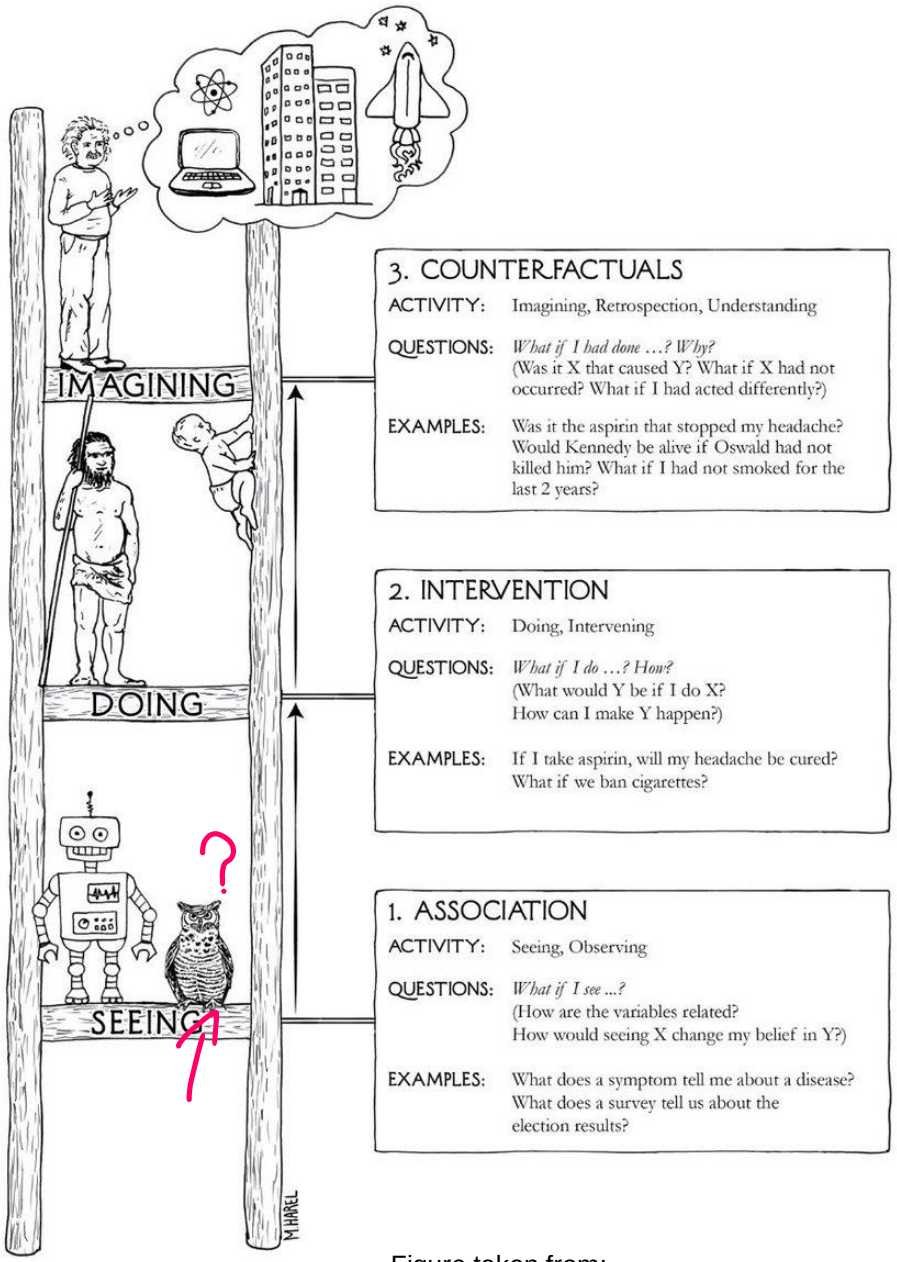


TABU graph (most edges) compared to the GS graph (least edges). Solid blue edges appear exclusively in the TABU graph, dashed blue edges exclusively in the GS graph, black edges appear in both graphs with the same orientation, and red edges appear in both graphs with different orientation.

Kitson, N. K., & Constantinou, A. (2021). Learning Bayesian networks from demographic and health survey data. *Journal of Biomedical Informatics*, Vol. 113, Article 103588 [Open-access DOI]

WHY CAUSAL STRUCTURE LEARNING IS IMPORTANT (BENEFITS)

WHY CAUSAL MODELS?



- Pearl's ladder of causation suggests that there are three steps to achieving true AI (Pearl and Mackenzie, 2018).

Figure taken from:
Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The new science of cause and effect.* Basic books.

FUTURE OF CAUSAL MACHINE LEARNING

Learning by association is not always a problem:

- Achievements in deep learning might be **blind to causality** but are **clearly impressive** in some areas.
  - Yet, these learning achievements made clearer than ever that **black-box solutions cannot satisfactorily inform** human decision-making.
- Causal ML to emerge as a **crucial approach in complementing predictive ML** and to support **verified** human decision-making.
  - We already observe a shift, both in academia and industry, towards white-box ML solutions that offer **transparency and explainability**.
- Distinguished deep learning researchers acknowledge the need to move towards causal representation learning: "*there is, now, cross-pollination and increasing interest in both fields* [deep learning and causal representation] *to benefit from the advances of the other*" (Schölkopf et al., 2021).

Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., and Bengio, Y. (2021). Toward causal representation learning. In *Proceeding of the IEEE*, 109(5), pp. 612–634

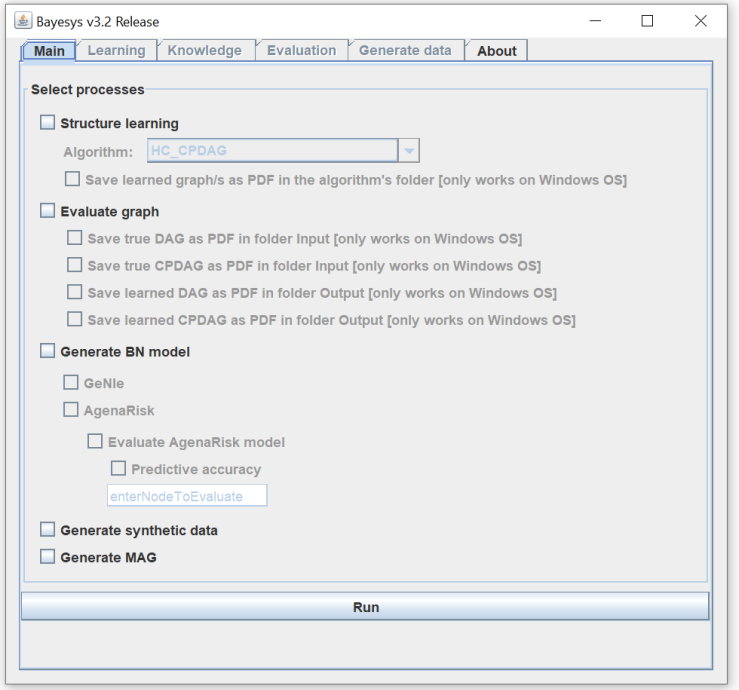
BAYESYS OPEN-SOURCE SYSTEM

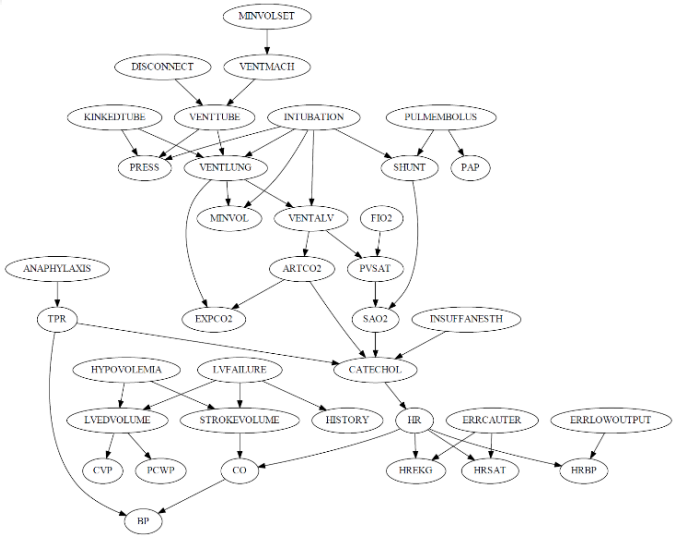
http://bayesian-ai.eecs.qmul.ac.uk/bayesys/

- Bayesys is a Java NetBeans project.
- Comes with a **user manual** and a **repository** of data sets, networks and case studies.

Provides access to:

- 6 structure learning algorithms.
- 10 knowledge-based approaches.
- Enables learning using multiple structure learning algorithms and data sets with a single click.
- Metrics to evaluate structure learning.
- Methods to generate clean or noisy synthetic data.
- Methods to draw learnt graphs in PDF.
- Converts learnt graphs that can be imported into AgenaRisk and GeNIe BNs and IDs.





THANK YOU

QUESTIONS?

Bayesian Artificial Intelligence
Research Lab

Queen Mary
University of London