IDSIA SUPSI

Imprecise Probabilistic Graphical Models in AI

Reasoning, Machine Learning and Causal Inference

Alessandro Antonucci (alessandro@idsia.ch)
Senior Lecturer-Researcher IDSIA USI-SUPSI

Sipta School 2024 - Ghent, August 14, 2024

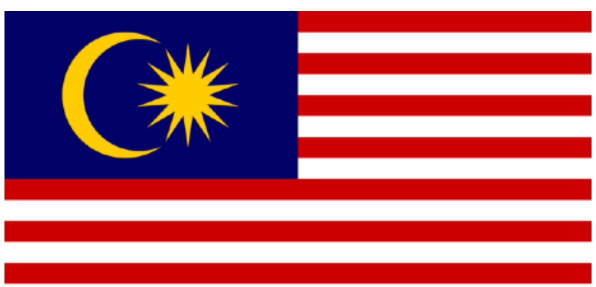
Sorry for arriving late,
yesterday I was in …

 V  ?

Sorry for arriving late,
yesterday I was in ...

 V  ?

No, Malaysia!



https://en.wikipedia.org/wiki/Stadthuys

Outline

I. $\underline{P} = \overline{P}$

II. $\text{AI} \neq \text{DL}$

III. (P)PGMs

IV. $\text{BN} + \text{CSs} = \text{CN}$

V. CN4DSS

VI. (C)ML

VII. $\text{SCM} \equiv \text{CN}$

Outline

I. $\underline{\mathrm{P}} = \overline{\mathrm{P}}$
II. AI $\neq$ DL
III. (P)PGMs
IV. BN + CSs = CN
V. CN4DSS
VI. (C)ML
VII. SCM $\equiv$ CN

I. probability theory
II. AI and deep learning
III. (precise) graphical models
IV. credal nets
V. decision-support by credal nets
VI. credal machine learing
VII. causality

Outline

I. $\underline{P} = \overline{P}$
II. AI $\neq$ DL
III. (P)PGMs
IV. BN + CSs = CN
V. CN4DSS
VI. (C)ML
VII. SCM $\equiv$ CN

I. probability theory II. AI and deep learning	~ slot #1
III. (precise) graphical models IV. credal nets	~ slot #2
V. decision-support by credal nets VI. credal machine learing	~ slot #3
VII. causality	~ slot #4

1. $\underline{\mathrm{P}} = \overline{\mathrm{P}}$

(~10 minutes of precise **P**robability)

(Precise) Probability Theory (in One Slide)

- X takes its values in Ω_X, generic value $x \in \Omega_X$
- We (initially) focus on categorical variables, i.e., $|\Omega_X| < +\inf$
- **Uncertainty** about X by a probability mass function (PMF) P
  - PMF $P : \Omega_X \to \mathbb{R}$, $P(x) \geq 0 \ \forall x \in \Omega_X$, $\sum_{x \in \Omega_X} P(x) = 1$
  - Expectation of $f : \Omega_X \to \mathbb{R}$? $\mathbb{E}[f] = \sum_{x \in \Omega_X} P(x) \cdot f(x)$
- Joint PMF $P(X, Y)$ (two or more variables)
  - **Marginalisation**, $P(X)$ s.t. $P(x) = \sum_{y \in \Omega_Y} P(x, y)$
  - **Conditioning**, $P(X|y)$ s.t. $P(x|y) = \frac{P(x,y)}{P(y)} = \frac{P(x,y)}{\sum_{x \in \Omega_X} P(x,y)}$ if $P(y) > 0$

IDSIA SUPSI

(First) Python Exercise

- Go to GDrive
- Install Colaboratory as an extension
- Create a notebook & install a package
- My notebooks in Github and GDrive
- Our **Notebook #1**
  - Learn $n(X,Y,Z) \rightarrow P(X,Y,Z)$
  - Marginalise gender Z to get $P(X,Y)$
  - Conditioning, e.g., recovery probability given treatment > or < than given no treatment?

```
>>> print("Hello World!")
Hello World!
>>>
```

Data from an observational study involving three Boolean variables [24, Section 4.1]. The state equal to one means *female* for Z, *treated* for X and *recovered* for Y.

Gender (Z)	Treatment (X)	Recovery (Y)	#
0	0	0	2
0	0	1	114
0	1	0	41
0	1	1	313
1	0	0	107
1	0	1	13
1	1	0	109
1	1	1	1

II. AI ≠ DL

Artificial **I**ntelligence is (not only) **D**eep **L**earning

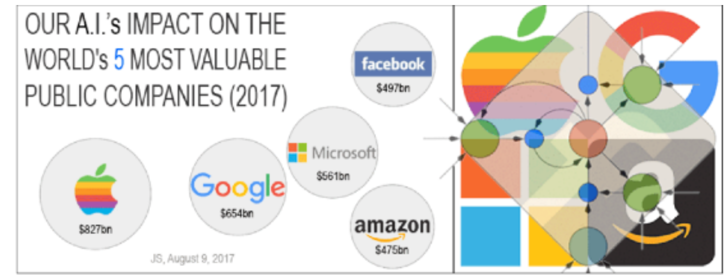

PIONEERS IN AI RESEARCH
Frameworks for Multi-GPU Pascal
Large-scale Deep Learning
Reinforcement Learning
Unsupervised and Transfer Learning
Natural Language Understanding
Autonomous Driving
Medical Applications
Berkeley
Carnegie Mellon University
MIT Massachusetts Institute of Technology
NYU
Stanford University
Université de Montréal
OXFORD
IDSIA
UNIVERSITY OF TORONTO

Istituto Dalle Molle di Studi sull'Intelligenza Artificiale (**IDSIA**)

USI/SUPSI

Istituto Dalle Molle di studi sull'intelligenza artificiale

IDSIA



Jürgen Schmidhuber (2017, reformatted in 2021)
Pronounce: You_again Shmidhoobuh

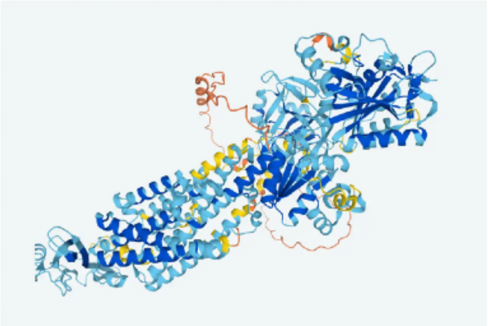
AI Blog
@SchmidhuberAI

- Research Institute founded in 1988 in Lugano to promote AI for **quality of life**
- Affiliated with both University of Lugano (USI) and University of Applied Sciences and Arts of Southern Switzerland (SUPSI)
- Staff ~100 people + 50 PhD
- Isipta '03&'17 + School '04 @Lugano

Angelo Dalle Molle (1908 - 2002)

AI is the new "electricity"?

- “About 100 years ago, electricity transformed every major industry. AI has advanced to the point where it has the power to transform every major sector in coming years.” Andrew Ng
- Recent (Deep Learning) Breakthroughs
  - Image Recognition (Super-Human) (~2015)
  - Translation (Near-Human) (~2016)
  - Go World-Champion Challenge (2017)
  - Protein Structure Prediction (2021)
  - (V)LLMs (~2022)



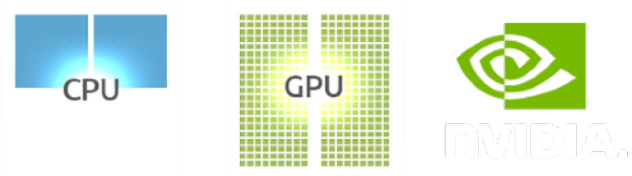
Or a new "bubble"?

(from Gary Marcus **substack**)

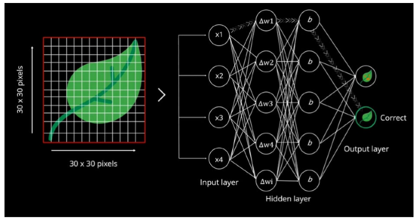
Deep Learning as a Series of (Fortunate) Events ...

DATA

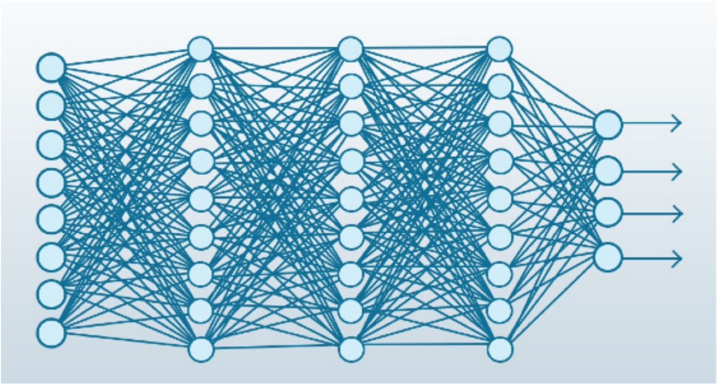
HARDWARE



THEORY

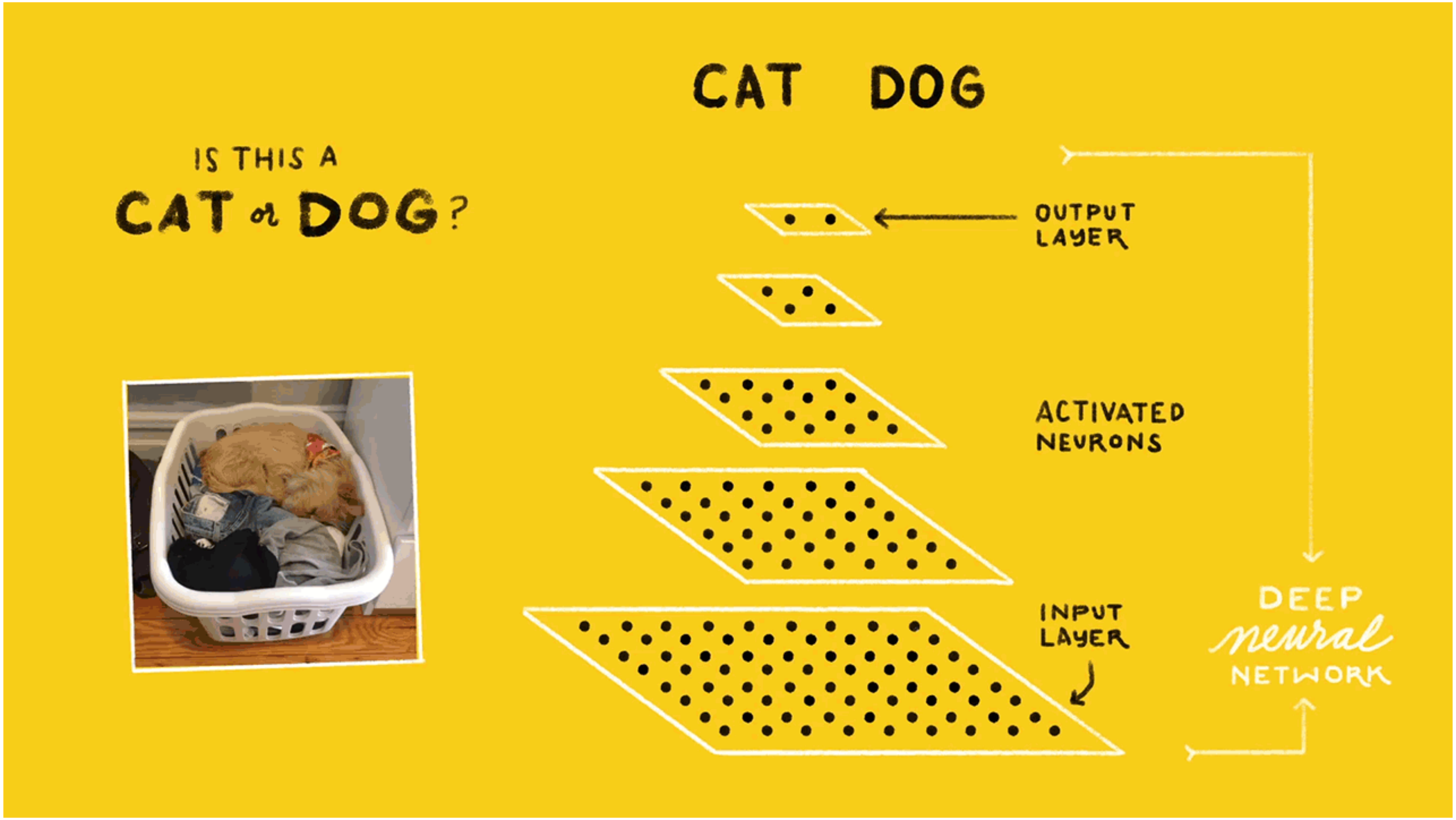


new technology
+ "earlier" theory



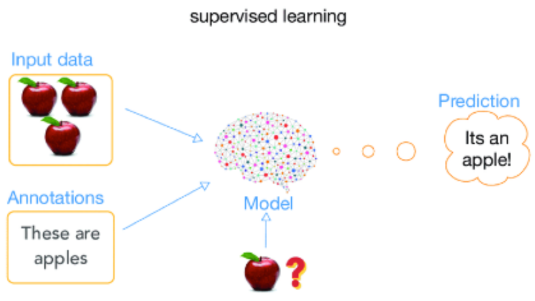
DEEP
NEURAL
NETS

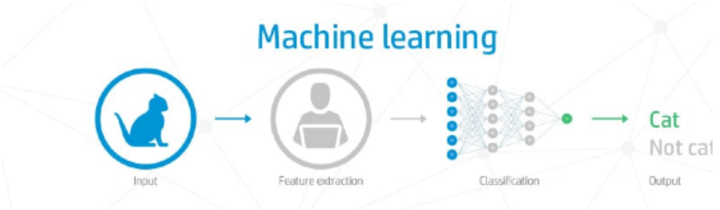
(Supervised) DL



Supervised Learning

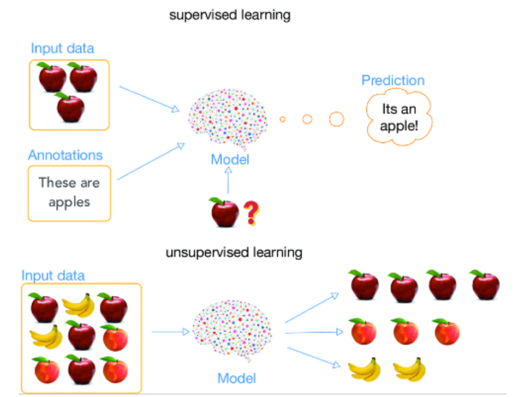
- Predictive task: find class Y from feature(s) X
- Based on annotated data $\{y_i, x_i\}_{i=1}^{d}$
- Algs to (optimally) learn $Y = f(X)$
- **Machine Learning** (ML), a two-step process
  - Feature Extraction (FE) $Z = g(X)$
  - Learn $Y = h(Z)$ from $\{y_i, g(x_i)\}_{i=1}^{d}$
  - $f := h \circ g$
- **Deep Learning** (DL) directly gets f
  - Automatic FE on initial layers
  - Unstructured features: training f requires more data than g

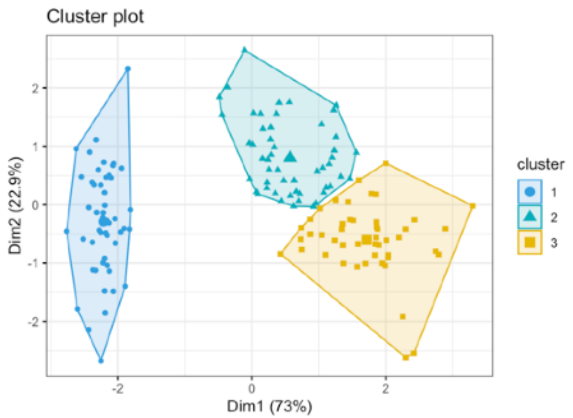


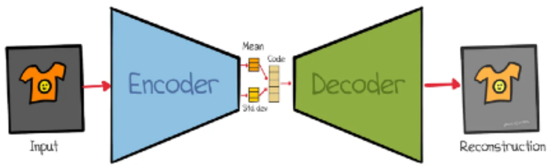


Unsupervised Learning

- Annotated data are costly (in many senses)
- Unannotated data $\{x_i\}_{i=1}^d$?
- Clustering: group together similar objects
- Again ML-vs-DL paradigm
- DL: good FE even in unsupervised settings
- Variational Autoencoders (VAEs)
- LLMs are a (super)sophistication of the (simple) VAE idea







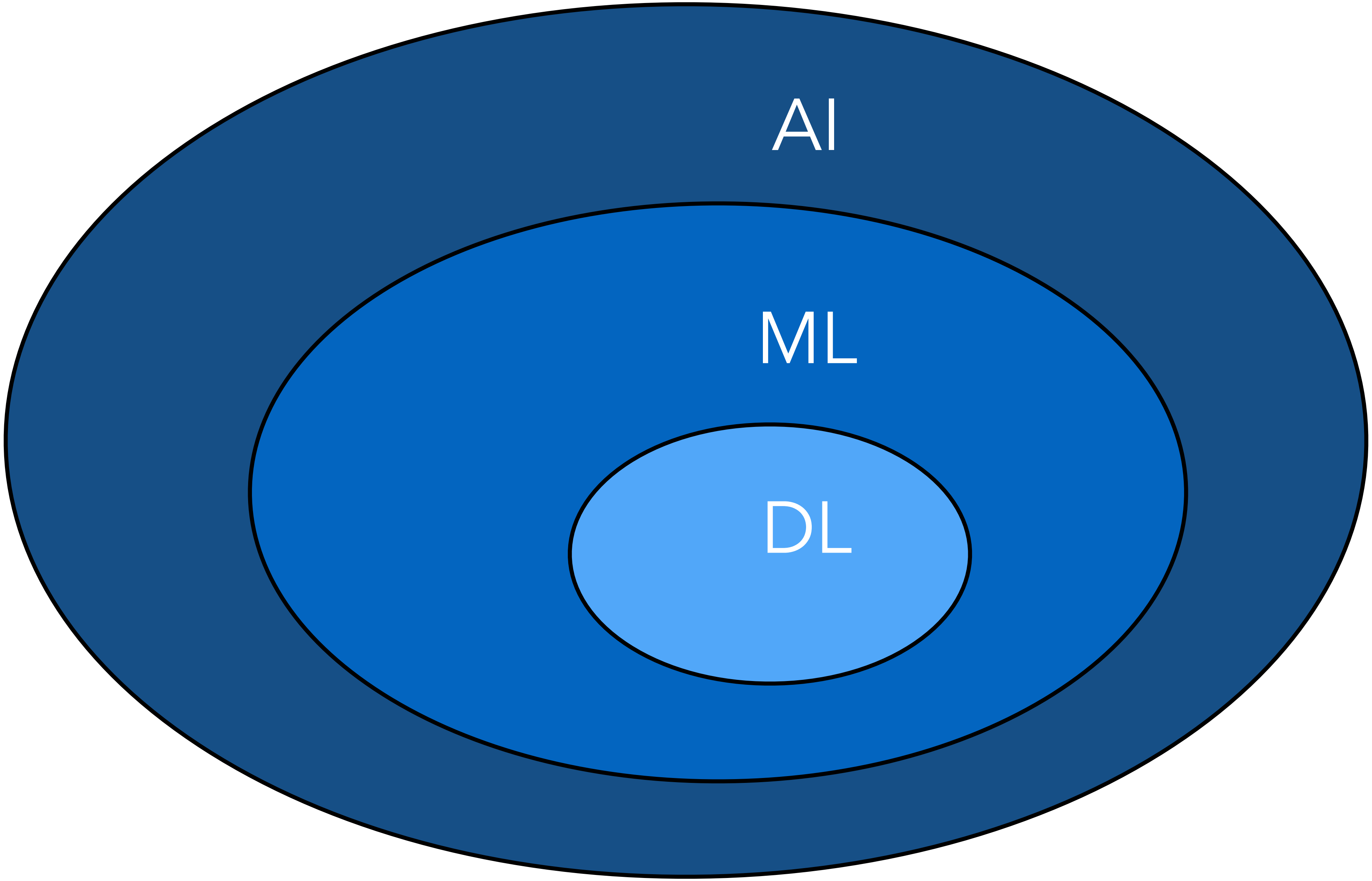
Let' (quickly) play with notebook #2

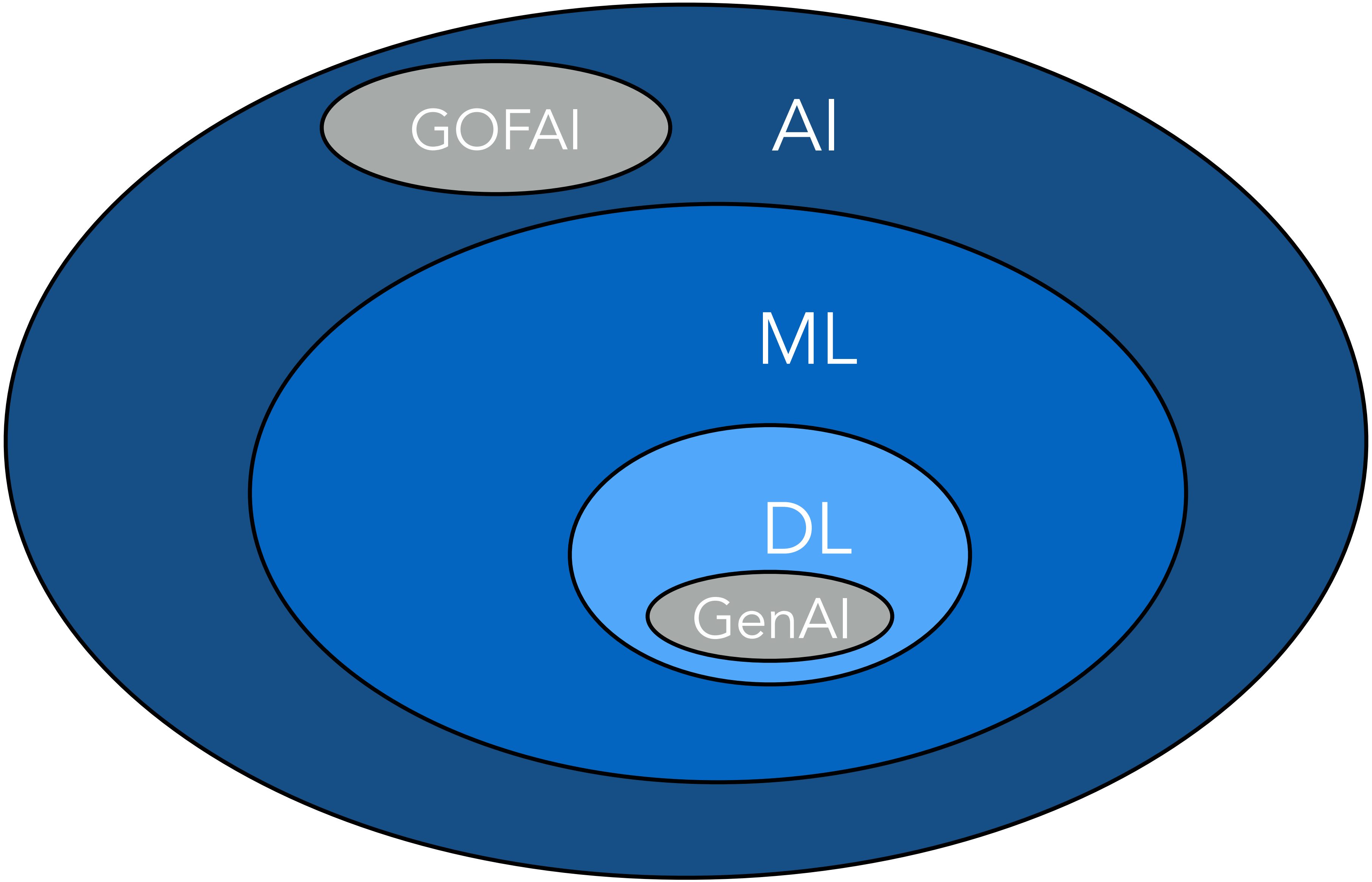
Discriminative vs. Generative

- Discriminative models designed to find Y given X
  - Deterministic models $\hat{y} = f(\hat{x})$
  - But also (conditional) probabilistic, i.e., $P(Y|\hat{x})$ and then

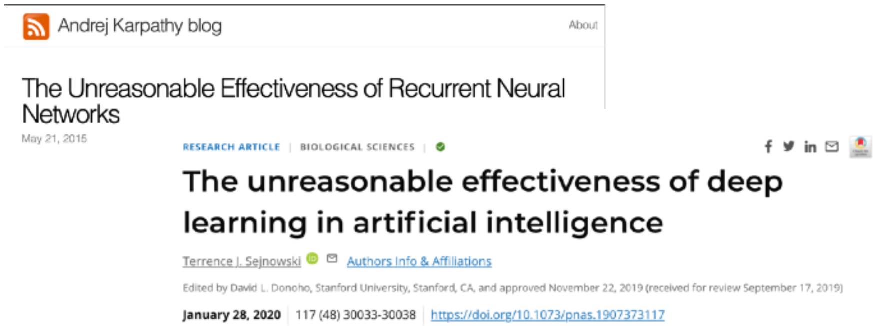
$$\hat{y} = \sum_{y \in \Omega_Y} y \cdot P(y|\hat{x}) \quad \text{or} \quad \hat{y} = \arg\max_{y \in \Omega_X} P(y|\hat{x})$$

- Generative models describe (joint) process behind (X, Y)
  - Joint PMF $P(X, Y)$
  - Predictions? $P(y|x) \propto P(y, x)$ (as $P(x)$ constant)
  - But also **reasoning**! E.g., MPE $\hat{x} = \arg\max_{x \in \Omega_X} P(x|\hat{y})$
  - More data (or knowledge) needed for training ...
    - Gen AI = neural generative models from unsupervised data
    - (Good Old-Fashioned) AI = symbolic generative models from experts


AI
ML
DL


GOFAI
AI
ML
DL
GenAI

AI > Deep Learning


Andrej Karpathy blog About

The Unreasonable Effectiveness of Recurrent Neural Networks

May 21, 2015

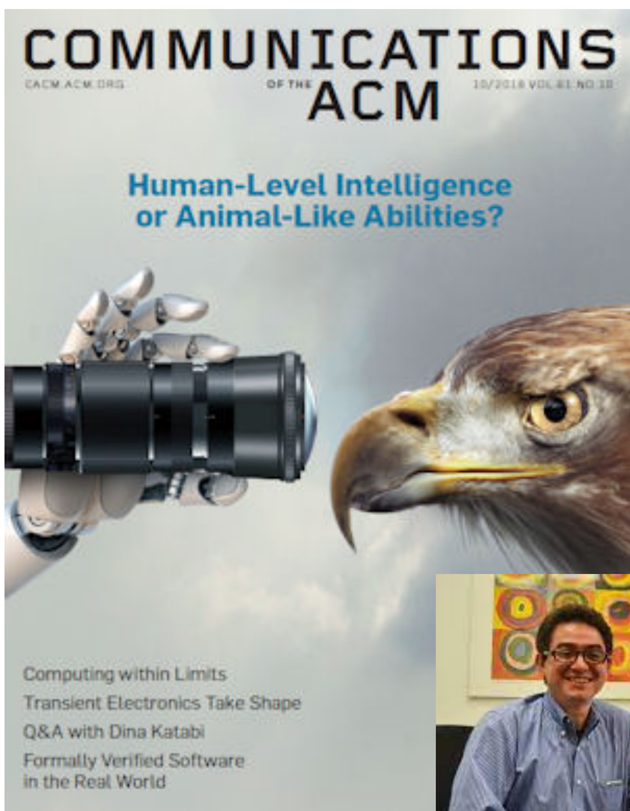
RESEARCH ARTICLE | BIOLOGICAL SCIENCES

The unreasonable effectiveness of deep learning in artificial intelligence

Terrence J. Sejnowski Authors Info & Affiliations

Edited by David L. Donoho, Stanford University, Stanford, CA, and approved November 22, 2019 (received for review September 17, 2019)

January 28, 2020 | 117 (48) 30033-30038 | https://doi.org/10.1073/pnas.1907373117

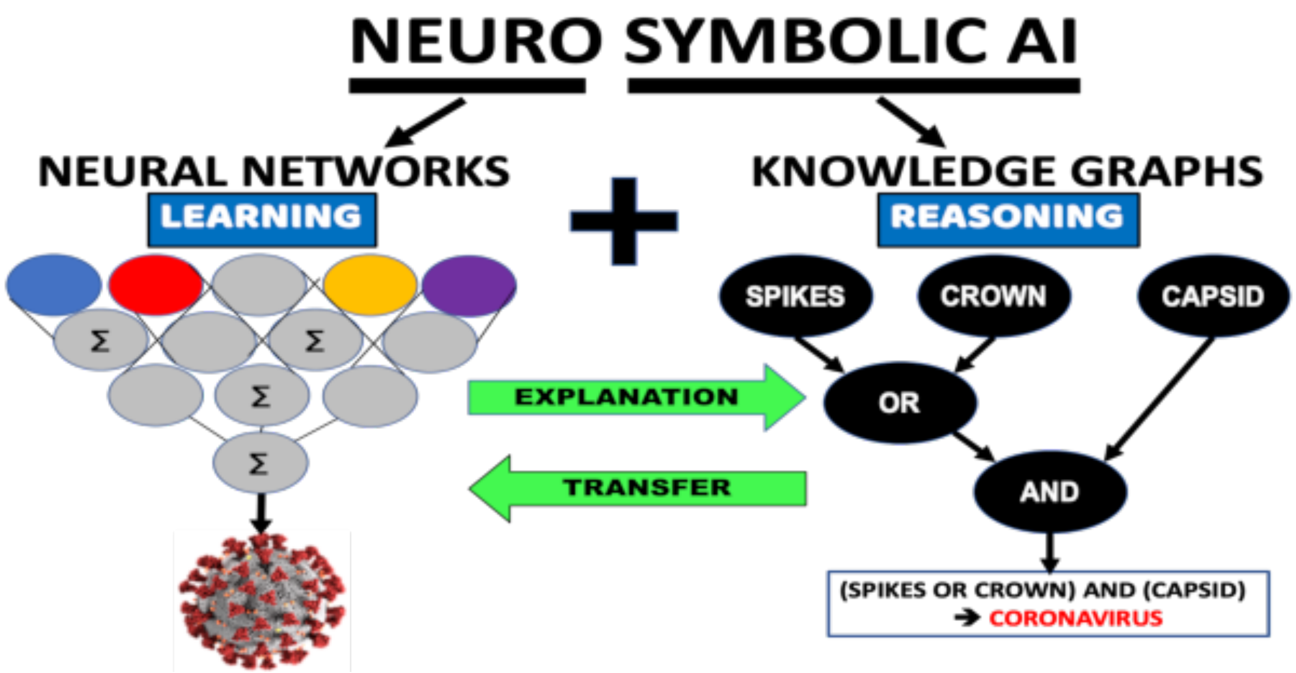


"Deep learning has instead given us machines with truly ***impressive abilities but no intelligence****.*

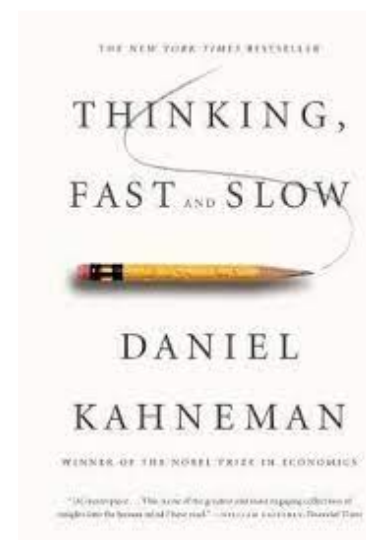
The difference is profound and lies in the ***absence of a model of reality****."*

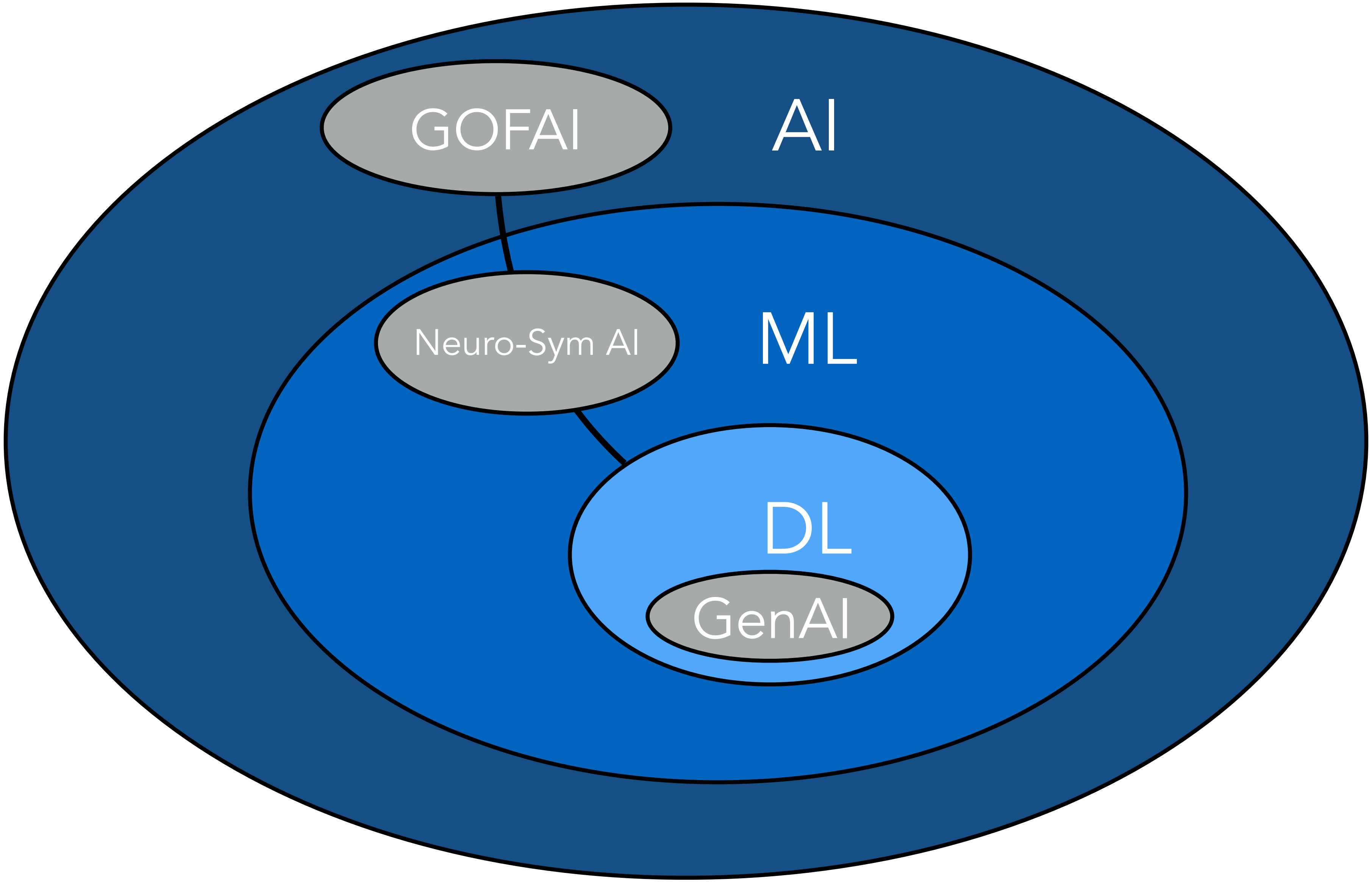
this seems to remain valid even in the LLMs age ...

A Unifying Framework: Neuro-Symbolic AI



System 1 "Fast"	System 2 "Slow"
DEFINING CHARACTERISTICS Unconscious Effortless Automatic	DEFINING CHARACTERISTICS Deliberate and conscious Effortful Controlled mental process
WITHOUT self-awareness or control	WITH self-awareness or control
"What you see is all there is."	Logical and skeptical
ROLE Assesses the situation Delivers updates	ROLE Seeks new/missing information Makes decisions




GOFAI
AI
Neuro-Sym AI
ML
DL
GenAI

AGI?

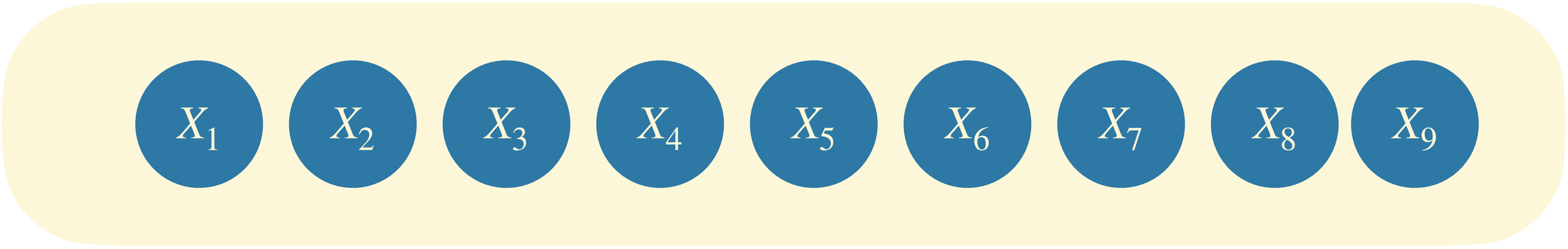
Strong AI enthusiasts say ~2028,
narrow AI people say not just around the corner ...

III. (P)PGMs

The sober elegance of (**P**recise) **P**robabilistic **G**raphical **M**odels

Assessing Generative Models (by Decomposition)

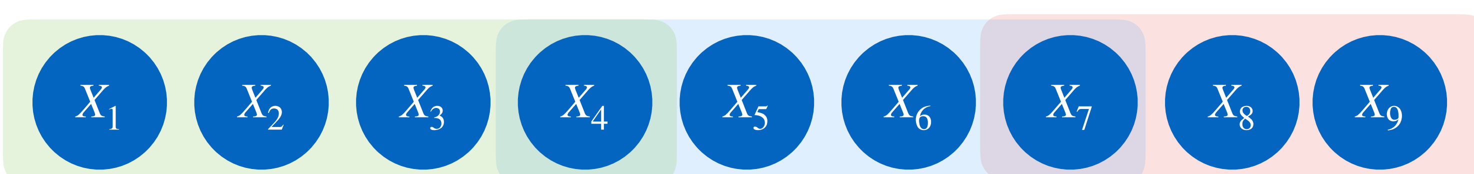
- Model variables $\mathbf{X} = (X_1, \ldots, X_n)$
- Joint PMF $P(\mathbf{X})$?
  - $O(2^n)$ humans are not good in eliciting small joint probabilities
  - Data? Sparse, risk of overfitting ...



$P(X_1, X_2, \ldots, X_9)$

Assessing Generative Models (by Decomposition)

- Model variables $\mathbf{X} = (X_1, \ldots, X_n)$
- Joint PMF $P(\mathbf{X})$?
  - $O(2^n)$ humans are not good in eliciting small joint probabilities
  - Data? Sparse, risk of overfitting ...
- Powerful idea: decomposition!
- Composition operator $\oplus$ (to be based on independence)



$$P(X_1, X_2, \ldots, X_9) = f(X_1, X_2, X_3, X_4) \oplus g(X_4, X_5, X_6, X_7) \oplus h(X_7, X_8, X_9)$$

Independence/Irrelevance as a Decomposition

- Independence $P(X_1, X_2) = P(X_1) \cdot P(X_2)$

Independence/Irrelevance as a Decomposition

- Independence $P(X_1, X_2) = P(X_1) \cdot P(X_2)$
- Equivalent to irrelevance $P(X_1 | X_2) = P(X_1)$

$P(X_1 | X_2) = \frac{P(X_1, X_2)}{P(X_2)} = P(X_1)$ implies $P(X_1, X_2) = P(X_1) \cdot P(X_2)$ and vice versa

Independence/Irrelevance as a Decomposition

SPOILER
The two concepts are not necessarily equivalent in imprecise settings

- Independence $P(X_1, X_2) = P(X_1) \cdot P(X_2)$
- Equivalent to irrelevance $P(X_1 | X_2) = P(X_1)$

$P(X_1 | X_2) = \dfrac{P(X_1, X_2)}{P(X_2)} = P(X_1)$ implies $P(X_1, X_2) = P(X_1) \cdot P(X_2)$ and vice versa

Independence/Irrelevance as a Decomposition

- Independence $P(X_1, X_2) = P(X_1) \cdot P(X_2)$
- Equivalent to irrelevance $P(X_1 | X_2) = P(X_1)$

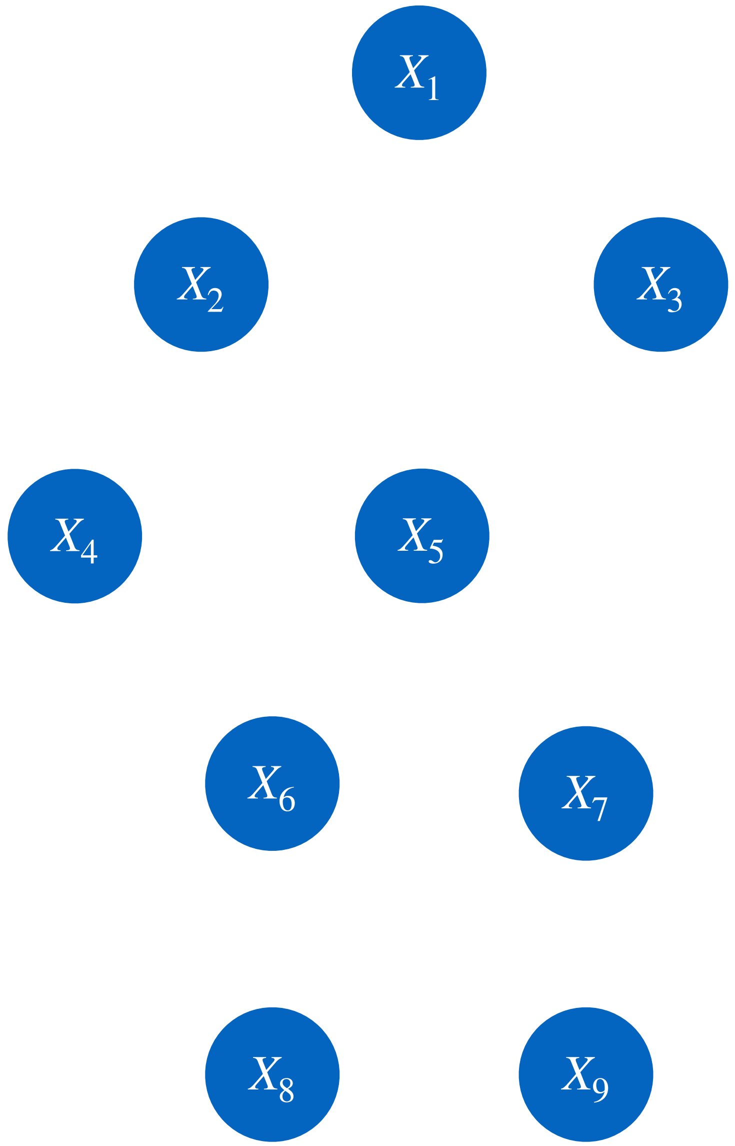
  $P(X_1 | X_2) = \dfrac{P(X_1, X_2)}{P(X_2)} = P(X_1)$ implies $P(X_1, X_2) = P(X_1) \cdot P(X_2)$ and vice versa
- But if X is irrelevant to $\mathbf{X} \setminus \{X\}$, we just don't need it ...
- More powerful concept:
  - conditional independence

    Es. knowing X_2 makes X_1 and X_3 indep $P(X_1, X_2 | X_3) = P(X_1 | X_3) \cdot P(X_2 | X_3)$
  - or, equivalently, conditional irrelevance

    Es. knowing X_2 makes X_3 irrelevant to X_1, i.e., $P(X_1 | X_2, X_3) = P(X_1 | X_3)$

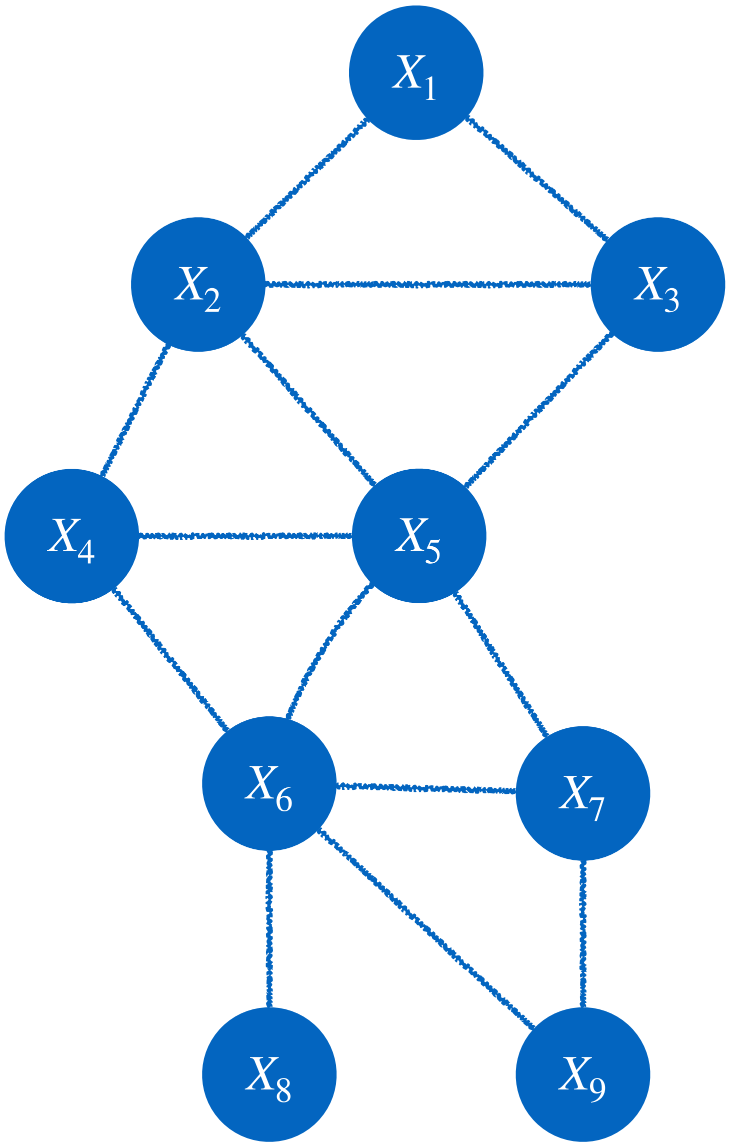
Graphical Models: Intuition

- Graphs (directed or undirected) as conditional independence maps
- Model variables $\mathbf{X} = (X_1, \ldots, X_n)$ as the nodes of a graph $\mathscr{G}$



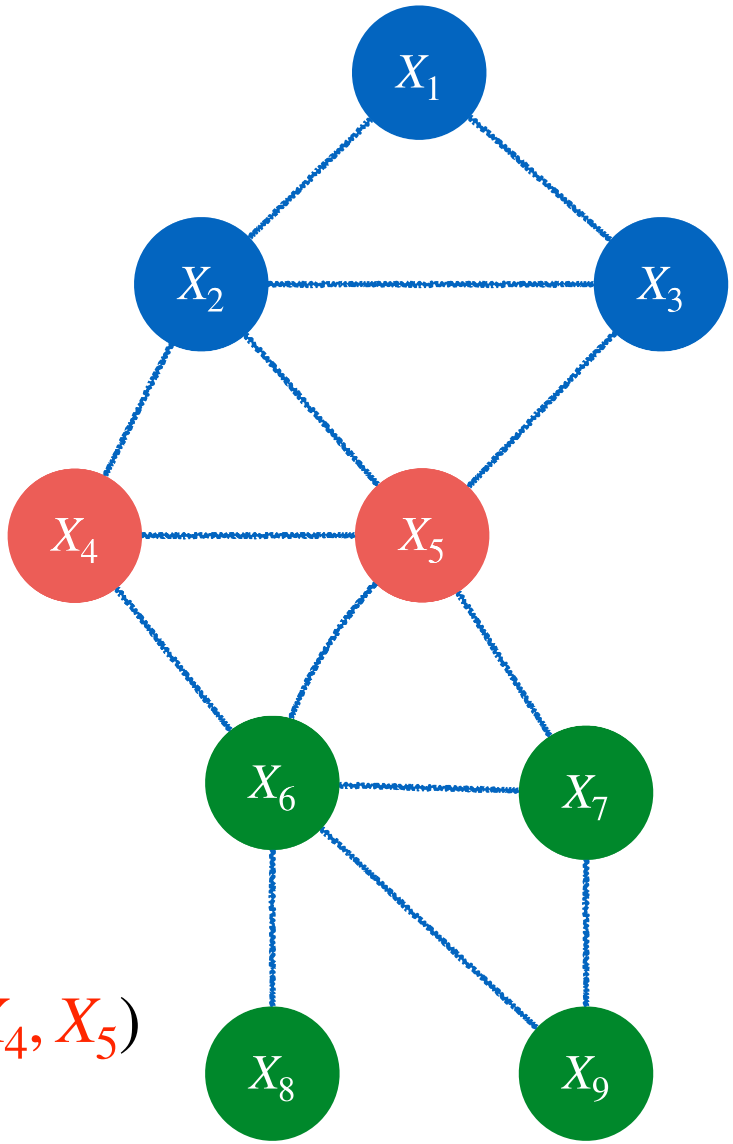
Graphical Models: Intuition

- Graphs (directed or undirected) as conditional independence maps
- Model variables $\mathbf{X} = (X_1, \ldots, X_n)$ as the nodes of a graph $\mathscr{G}$
- With undirected graphs



Graphical Models: Intuition

- Graphs (directed or undirected) as conditional independence maps
- Model variables $\mathbf{X} = (X_1, \ldots, X_n)$ as the nodes of a graph $\mathcal{G}$
- With undirected graphs, separation induced by a set of variables (roughly) corresponds to conditional independence

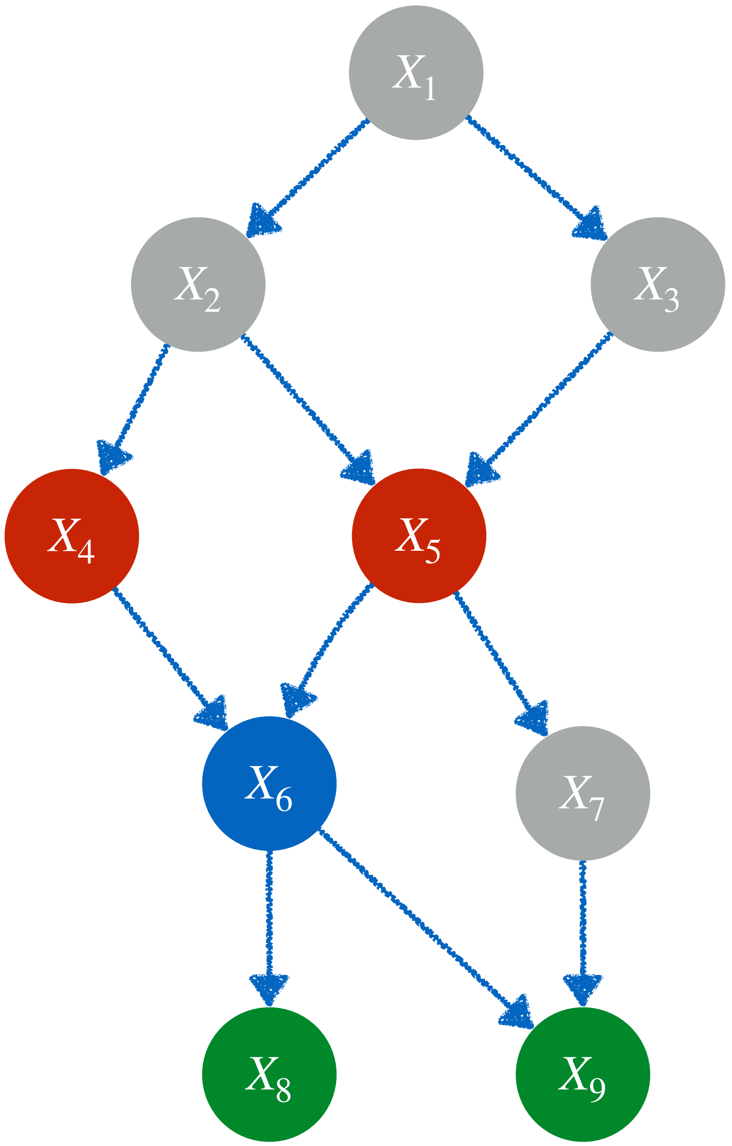


$$P(X_1, X_2, X_3, X_6, X_7, X_8, X_9 \,|\, X_4, X_5)$$
$$= P(X_1, X_2, X_3 \,|\, X_4, X_5) \cdot P(X_6, X_7, X_8, X_9 \,|\, X_4, X_5)$$

$$P(X_6 \mid X_1, X_2, X_3, X_4, X_5, X_7) = P(X_6 \mid X_4, X_5)$$

Graphical Models: Intuition

- Graphs (directed or undirected) as conditional independence maps
- Model variables $\mathbf{X} = (X_1, \ldots, X_n)$ as the nodes of a graph $\mathscr{G}$
- With undirected graphs, separation induced by a set of variables (roughly) corresponds to conditional independence
- **Markov condition** for di-graphs: *"every var independent of the non-desc non-parents given the parents"*



From Chain Rule to Bayesian Networks (Pearl, 1984)

- Chain rule based on (iterated) definition of conditional probability
  $$P(X_1, X_2, X_3) = P(X_1 \mid X_2, X_3) \cdot P(X_2, X_3) = P(X_1 \mid X_2, X_3) \cdot P(X_2 \mid X_3) \cdot P(X_3)$$

From Chain Rule to Bayesian Networks (Pearl, 1984)

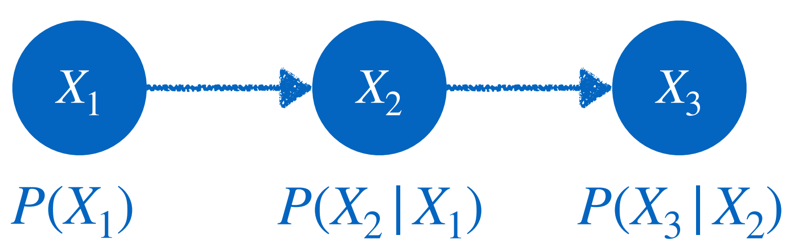
- Chain rule based on (iterated) definition of conditional probability
  $P(X_1, X_2, X_3) = P(X_1 | X_2, X_3) \cdot P(X_2, X_3) = P(X_1 | X_2, X_3) \cdot P(X_2 | X_3) \cdot P(X_3)$
- If X_3 irrelevant to X_1 given X_2: $P(X_1, X_2, X_3) = P(X_1 | X_2) \cdot P(X_2 | X_3) \cdot P(X_3)$

From Chain Rule to Bayesian Networks (Pearl, 1984)

- Chain rule based on (iterated) definition of conditional probability
  $P(X_1, X_2, X_3) = P(X_1 | X_2, X_3) \cdot P(X_2, X_3) = P(X_1 | X_2, X_3) \cdot P(X_2 | X_3) \cdot P(X_3)$
- If X_3 irrelevant to X_1 given X_2: $P(X_1, X_2, X_3) = P(X_1 | X_2) \cdot P(X_2 | X_3) \cdot P(X_3)$
- If $\mathcal{G}$ is acyclic and the variables are in **topological order**, the Markov condition implies $P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i | \mathrm{Pa}_{X_i})$ with Pa_{X_i} **parents** of X_i

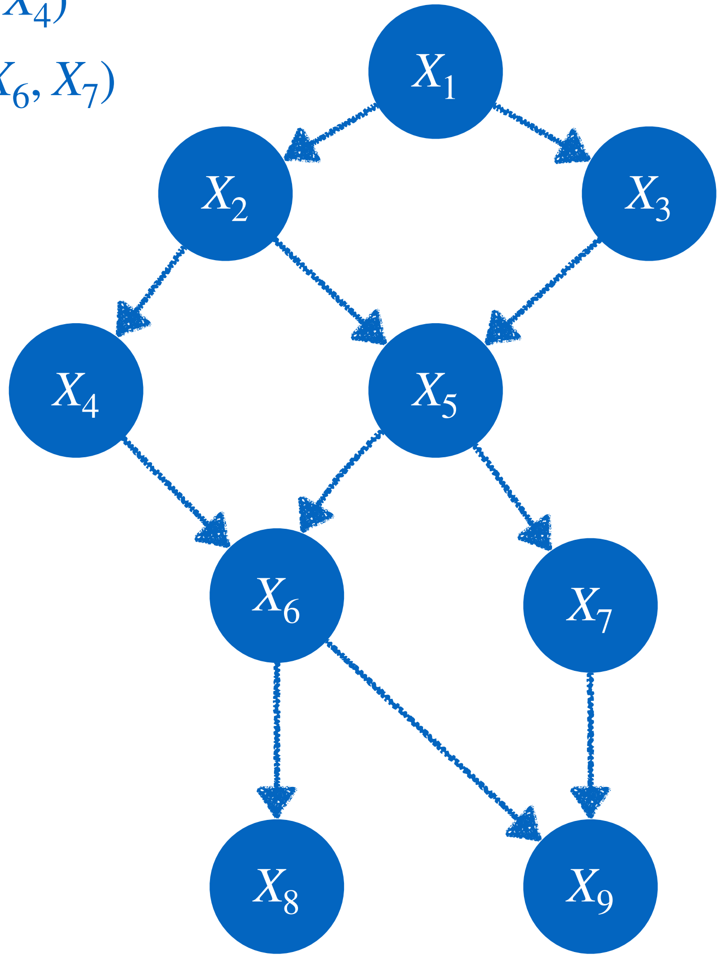
From Chain Rule to Bayesian Networks (Pearl, 1984)

- Chain rule based on (iterated) definition of conditional probability
  $P(X_1, X_2, X_3) = P(X_1 | X_2, X_3) \cdot P(X_2, X_3) = P(X_1 | X_2, X_3) \cdot P(X_2 | X_3) \cdot P(X_3)$
- If X_3 irrelevant to X_1 given X_2: $P(X_1, X_2, X_3) = P(X_1 | X_2) \cdot P(X_2 | X_3) \cdot P(X_3)$
- If $\mathscr{G}$ is acyclic and the variables are in **topological order**, the Markov condition implies $P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i | \mathrm{Pa}_{X_i})$ with Pa_{X_i} **parents** of X_i
- A joint model over $\mathbf{X}$ based "only" on the conditional probability tables (CPTs) for each variable given their parents
- Compact, $O(2^{\max_i |\mathrm{Pa}_{X_i}|})$, specification of generative models



Bayesian Net (BN) = Graph $\mathcal{G}$ + Conditional Probability Tables (CPTs)

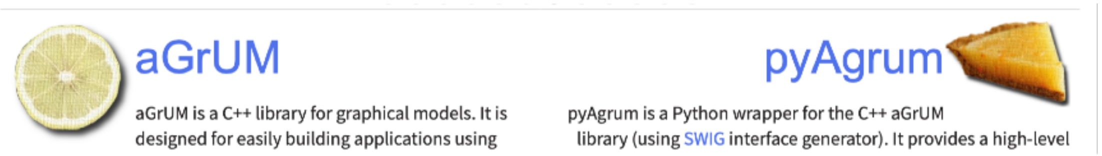
$$
\begin{aligned}
P(X_1, \ldots, X_9) = \; & P(X_1) \cdot P(X_2 \,|\, X_1) \cdot P(X_3 \,|\, X_1) \\
& P(X_4 \,|\, X_2) \cdot P(X_5 \,|\, X_2, X_3) \cdot P(X_6 \,|\, X_5, X_4) \\
& P(X_7 \,|\, X_5) \cdot P(X_8 \,|\, X_5) \cdot P(X_9 \,|\, X_6, X_7)
\end{aligned}
$$

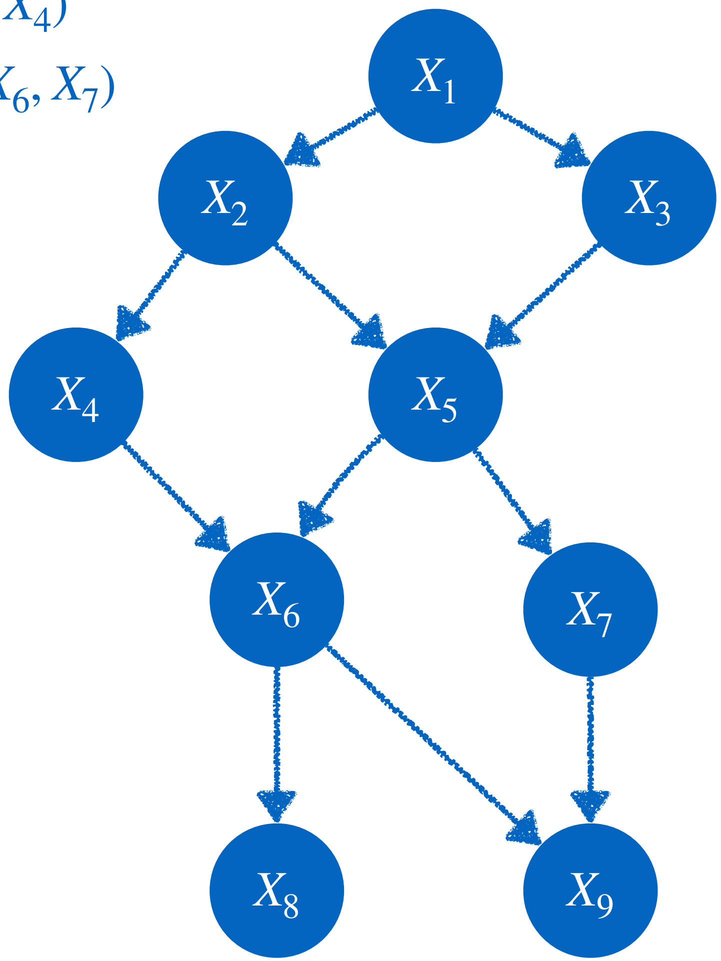


Bayesian Net (BN) = Graph $\mathscr{G}$ + Conditional Probability Tables (CPTs)

$$\begin{aligned} P(X_1, \ldots, X_9) = \; & P(X_1) \cdot P(X_2 \,|\, X_1) \cdot P(X_3 \,|\, X_1) \\ & P(X_4 \,|\, X_2) \cdot P(X_5 \,|\, X_2, X_3) \cdot P(X_6 \,|\, X_5, X_4) \\ & P(X_7 \,|\, X_5) \cdot P(X_8 \,|\, X_5) \cdot P(X_9 \,|\, X_6, X_7) \end{aligned}$$

Let' play with notebook #3

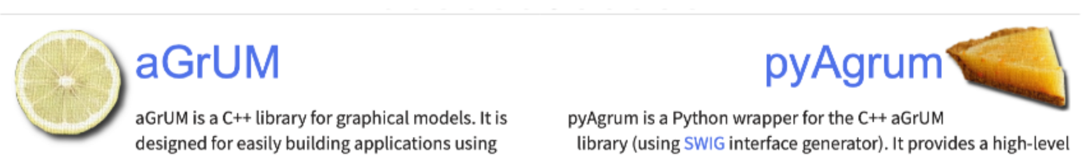




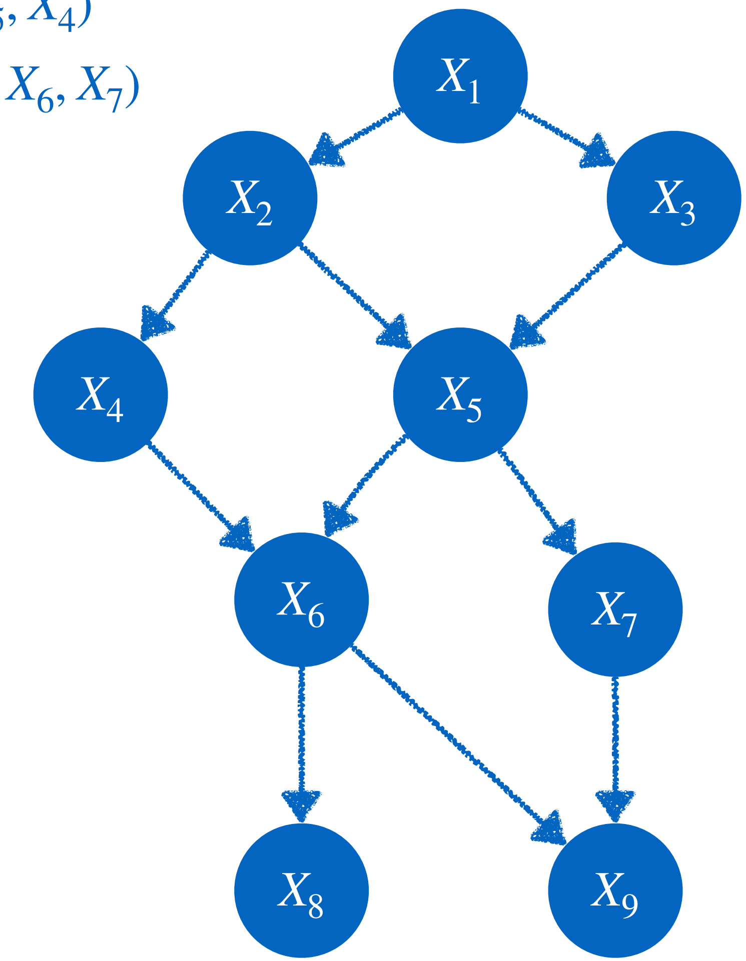
Bayesian Net (BN) = Graph $\mathscr{G}$ + Conditional Probability Tables (CPTs)

$$P(X_1, \ldots, X_9) = P(X_1) \cdot P(X_2 | X_1) \cdot P(X_3 | X_1)$$
$$P(X_4 | X_2) \cdot P(X_5 | X_2, X_3) \cdot P(X_6 | X_5, X_4)$$
$$P(X_7 | X_5) \cdot P(X_8 | X_5) \cdot P(X_9 | X_6, X_7)$$

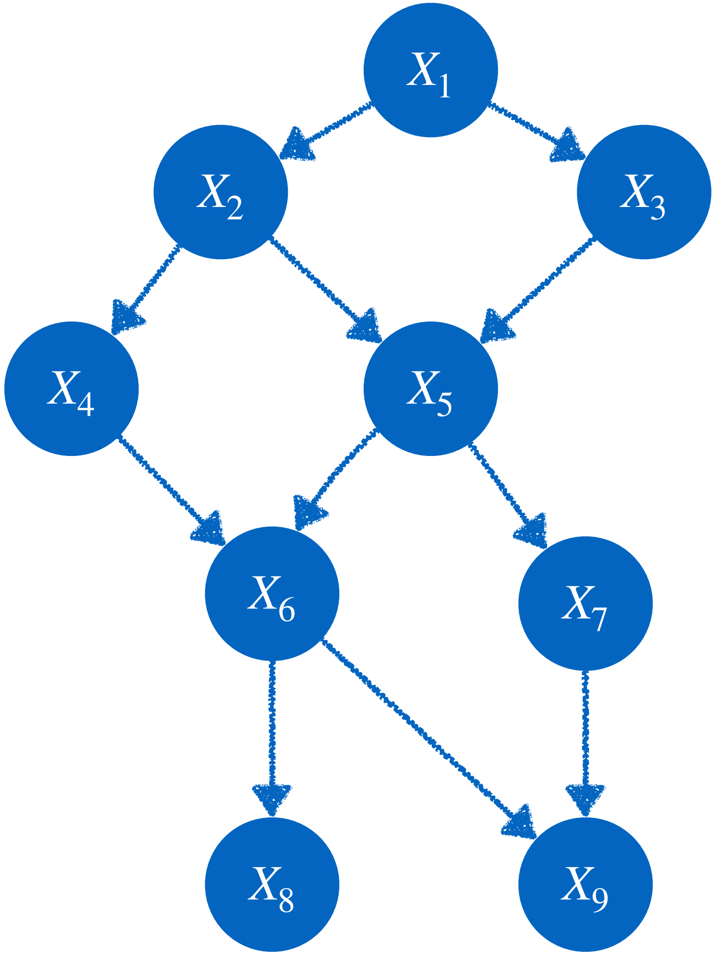
Let' play with notebook #3



SPOILER: We can generalise BNs by replacing the PMFs in the CPTs columns of the CPTs by sets of PMFs



Reasoning with Bayesian Networks (Inference)

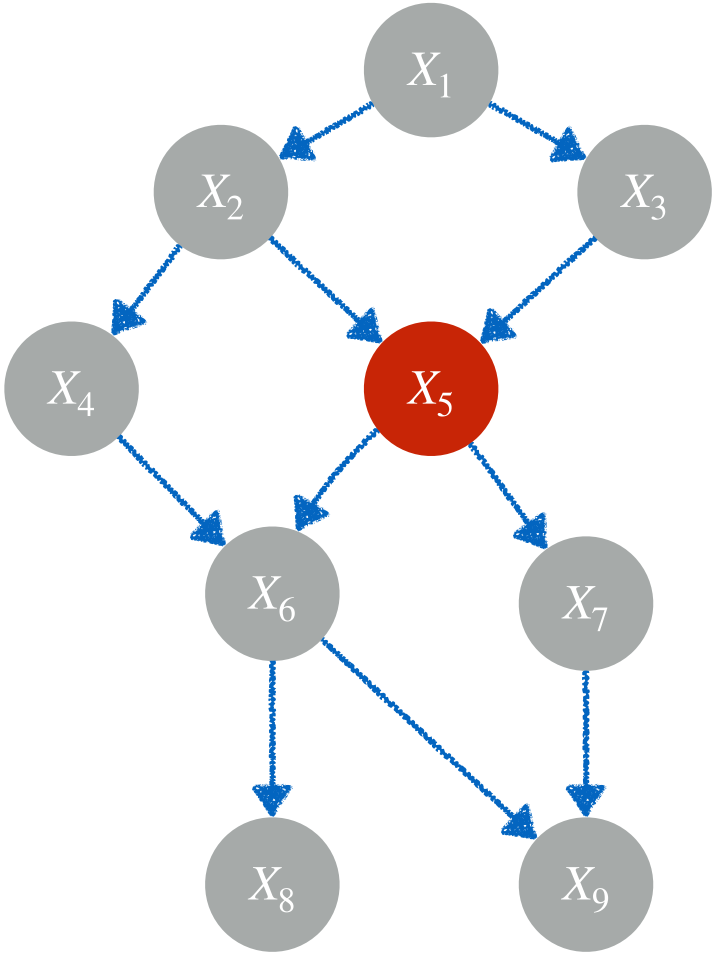


Reasoning with Bayesian Networks (Inference)

- **Marginal** on a queried var

$$P(x_q) = \sum_{X \in \mathbf{X} \setminus \{X_q\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})$$

$P(X_5) = ?$



Reasoning with Bayesian Networks (Inference)

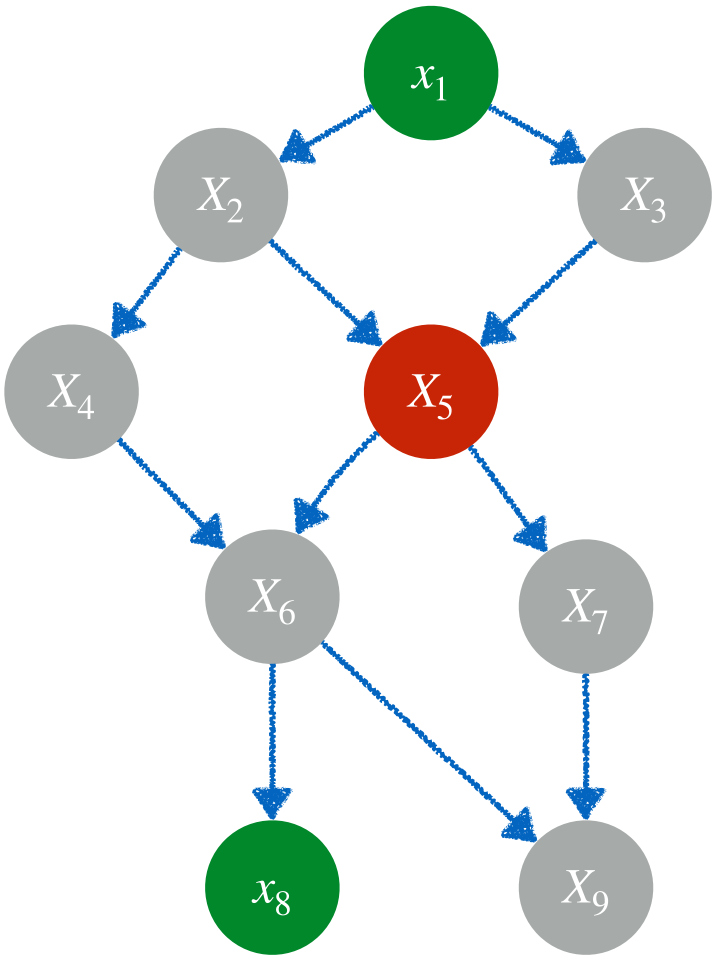
- **Marginal** on a queried var

$$P(x_q) = \sum_{X \in \mathbf{X} \setminus \{X_q\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})$$

- **Updating** query given evidence

$$P(x_q | \mathbf{x}_E) = \frac{\sum_{X \in \mathbf{X} \setminus \{X_q, X_E\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})}{\sum_{X \in \mathbf{X} \setminus \{X_q\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})}$$

$$P(X_5 | x_1, x_8) = ?$$



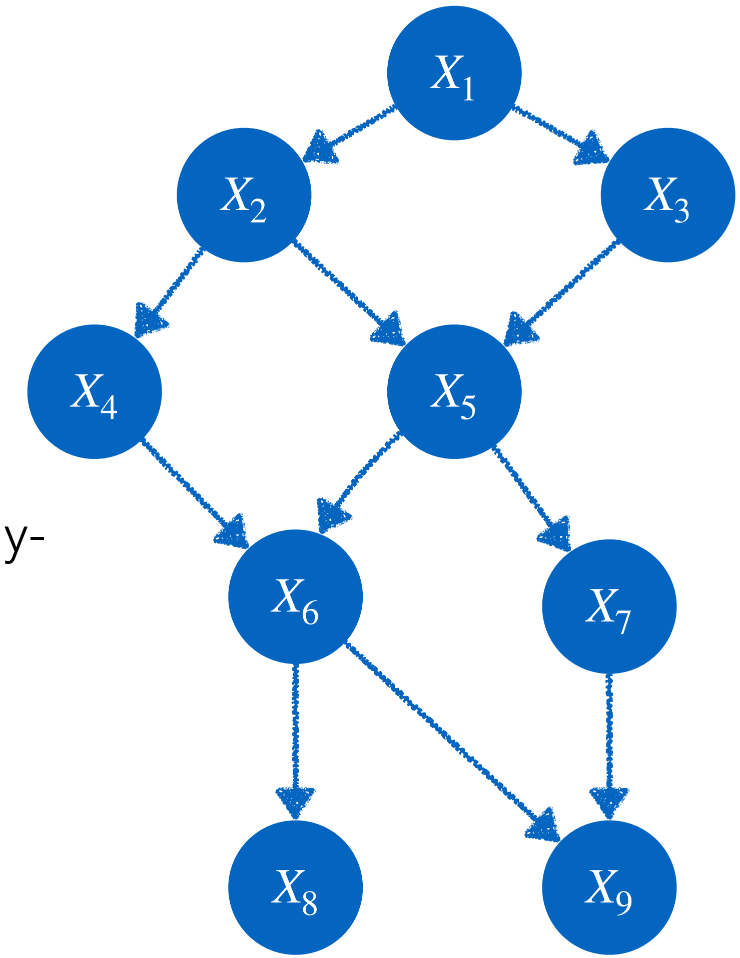
Reasoning with Bayesian Networks (Inference)

- **Marginal** on a queried var

$$P(x_q) = \sum_{X \in \mathbf{X} \setminus \{X_q\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})$$

- **Updating** query given evidence

$$P(x_q | \mathbf{x}_E) = \frac{\sum_{X \in \mathbf{X} \setminus \{X_q, X_E\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})}{\sum_{X \in \mathbf{X} \setminus \{X_q\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})}$$



Both (NP-)hard tasks, fast exact inference with singly-connected topologies, in general exponential wrt treewidth, many good approximate schemes

Reasoning with Bayesian Networks (Inference)

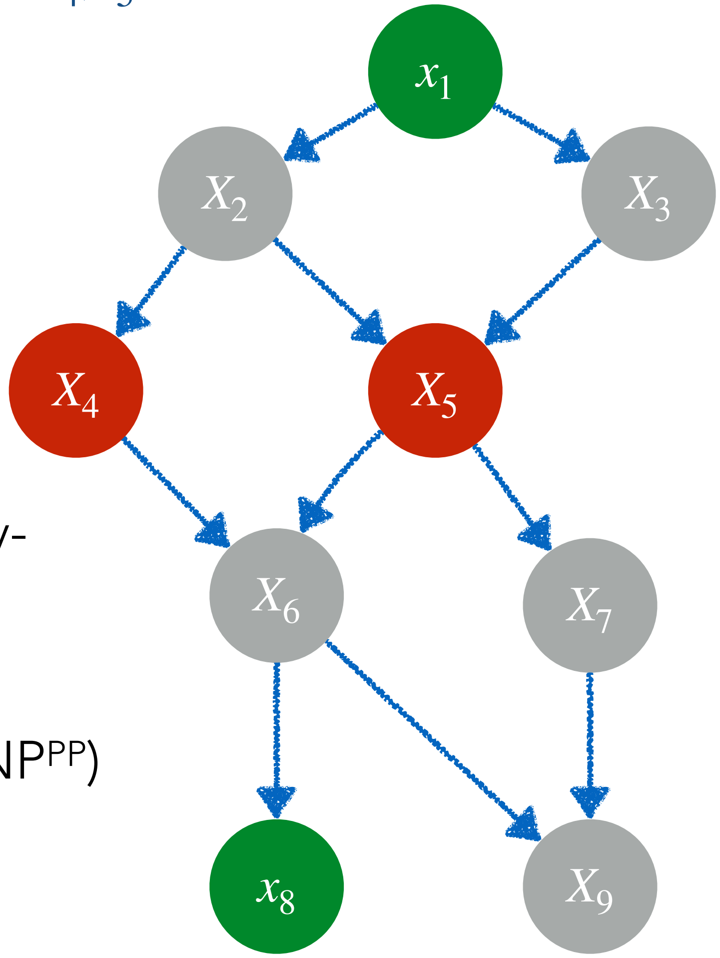
- **Marginal** on a queried var

$$P(x_q) = \sum_{X \in \mathbf{X} \setminus \{X_q\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})$$

- **Updating** query given evidence

$$P(x_q | \mathbf{x}_E) = \frac{\sum_{X \in \mathbf{X} \setminus \{X_q, X_E\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})}{\sum_{X \in \mathbf{X} \setminus \{X_q\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})}$$

$$\arg\max_{x_4, x_5} P(x_4, x_5 | x_1, x_8) = ?$$



Both (NP-)hard tasks, fast exact inference with singly-connected topologies, in general exponential wrt treewidth, many good approximate schemes

- Most probable **explanation** (MMAP) is harder (NP^{PP})

$$\mathbf{x}_q^* := \arg\max \sum_{X \in \mathbf{X} \setminus \{\mathbf{X}_q\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})$$

Reasoning with Bayesian Networks (Inference)

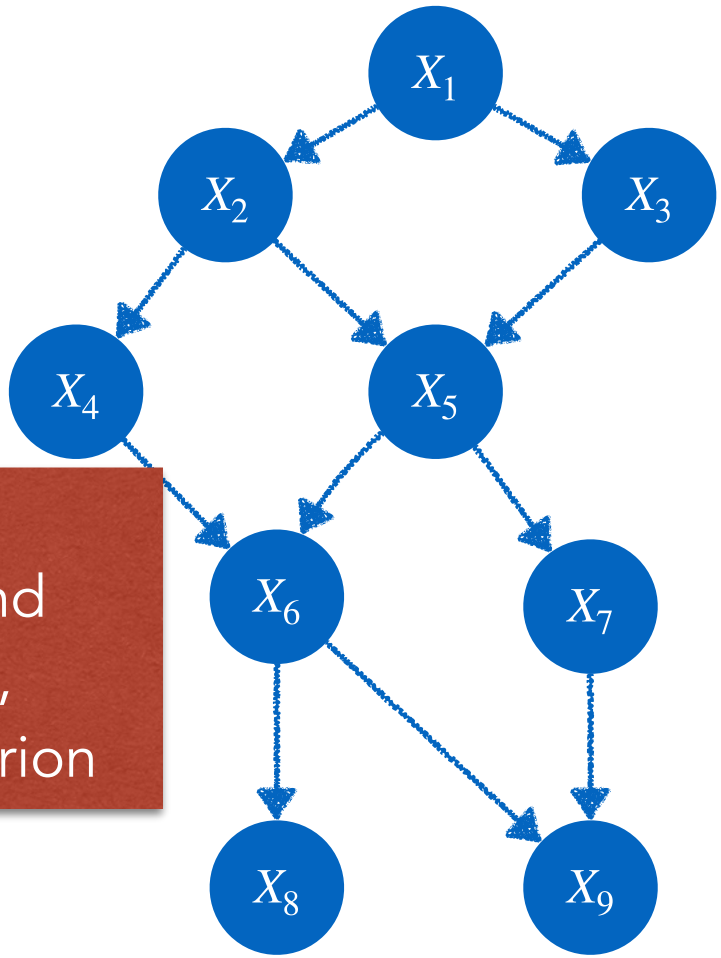
Let' keep playing with notebook #3

- **Marginal** on a queried var

$$P(x_q) = \sum_{X \in \mathbf{X} \setminus \{X_q\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})$$

- **Updating** query given evidence

$$P(x_q | \mathbf{x}_E) = \frac{\sum_{X \in \mathbf{X} \setminus \{X_q, X_E\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})}{\sum_{X \in \mathbf{X} \setminus \{X_q\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})}$$



Both (N
connec
treewid

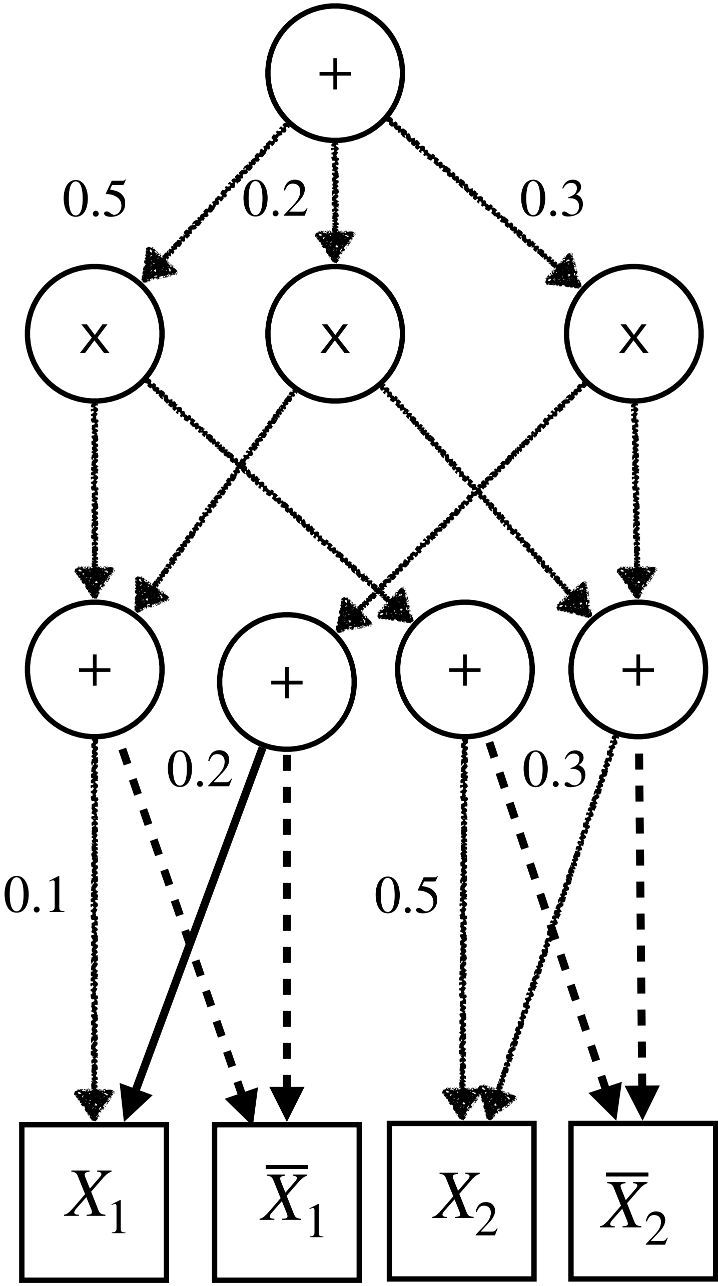
- Mos

SPOILER
With imprecise models, marginal and updating are non-equivalent tasks, MMAP depends on the decision criterion

$$\mathbf{x}_q^* := \arg\max \sum_{X \in \mathbf{X} \setminus \{X_q\}} \prod_{i=1}^{n} P(x_i | pa_{X_i})$$

Tractable Models: Sum-Product Networks

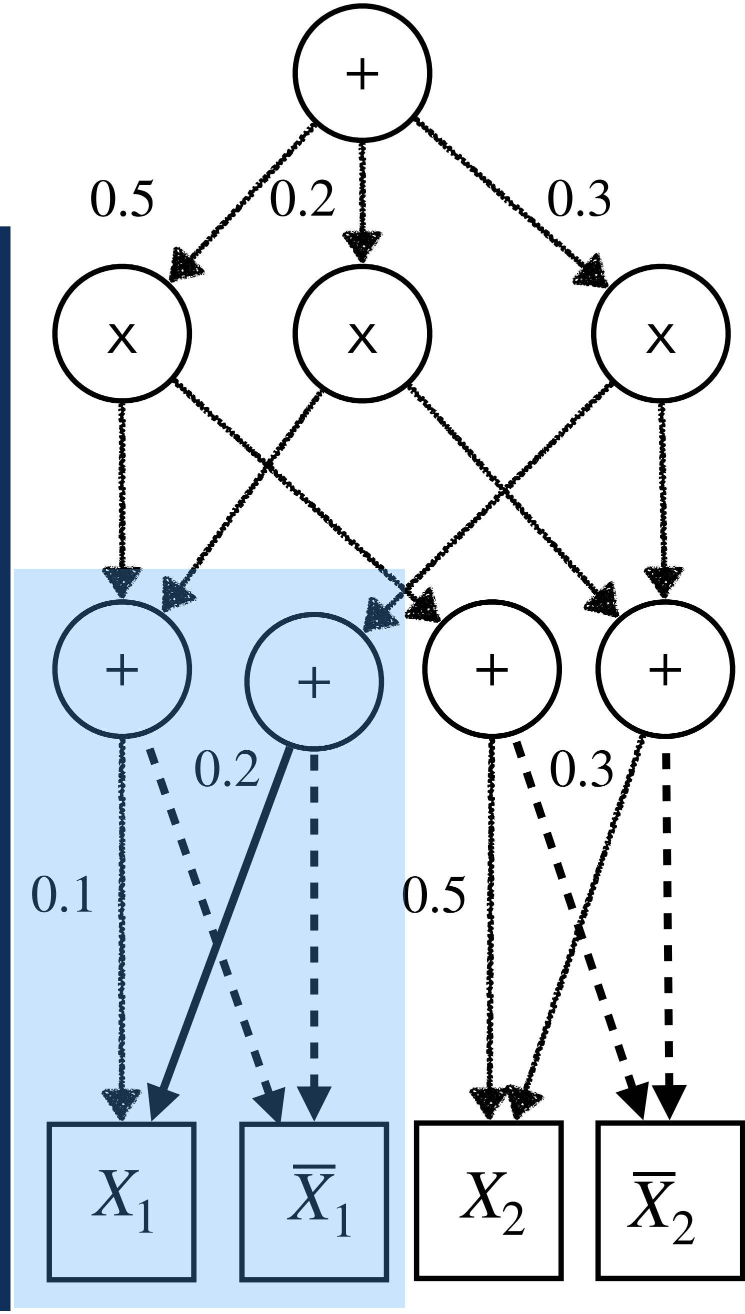
- Deep models (aka arithmetic/prob circuits)
- Generative models based on graph
- Tractable: inference $O(|\mathcal{G}|)$ for basic tasks
- $\mathcal{G}$ expresses an inferential computation schemes, not the (context-specific) independence relations
- Competitive performance wrt discriminative DL models, but less interpretable than BNs



Tractable Models: Sum-Product Networks

$$P_1(X_1) = 0.1 \cdot I_{X_1} + 0.9 \cdot I_{\overline{X_1}}$$

$$P_2(X_1) = 0.2 \cdot I_{X_1} + 0.8 \cdot I_{\overline{X_1}}$$



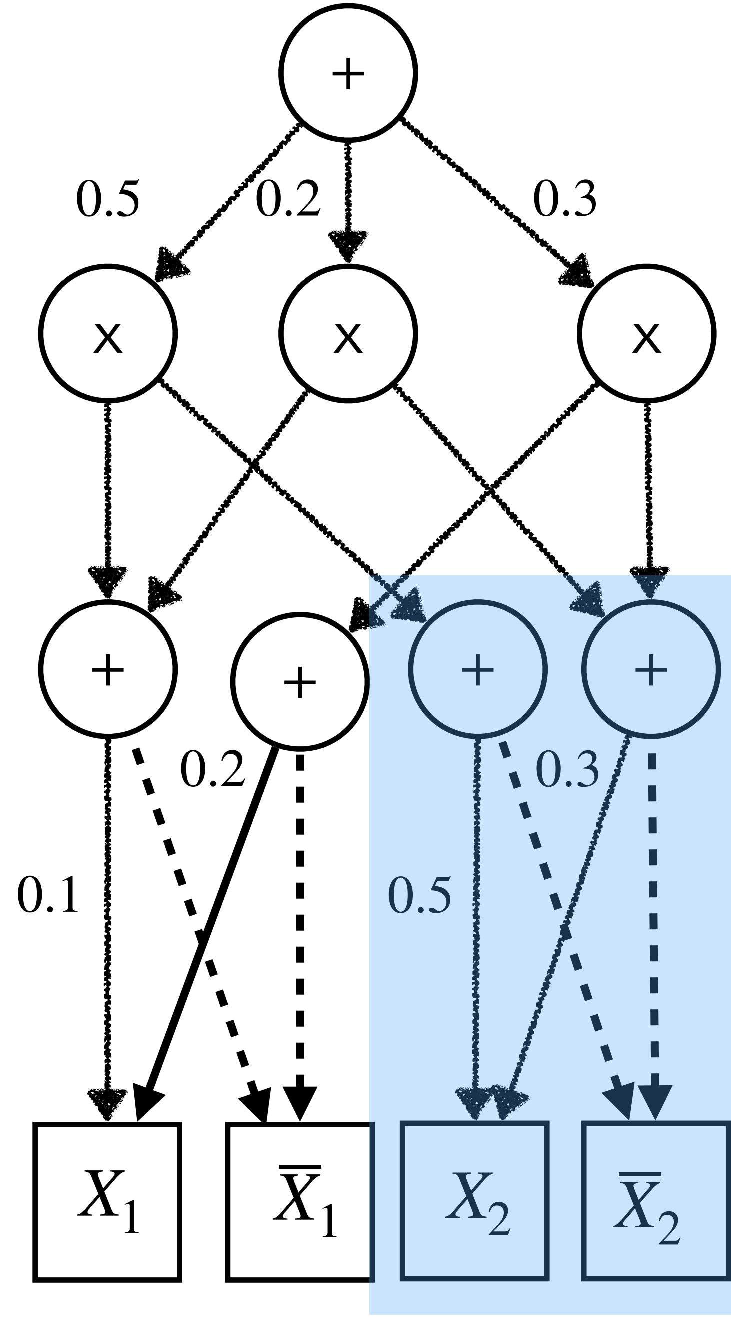
Tractable Models: Sum-Product Networks

$$P_1(X_1) = 0.1 \cdot I_{X_1} + 0.9 \cdot I_{\overline{X_1}}$$

$$P_2(X_1) = 0.2 \cdot I_{X_1} + 0.8 \cdot I_{\overline{X_1}}$$

$$P_1(X_2) = 0.5 \cdot I_{X_2} + 0.5 \cdot I_{\overline{X_2}}$$

$$P_2(X_2) = 0.3 \cdot I_{X_2} + 0.7 \cdot I_{\overline{X_2}}$$



Tractable Models: Sum-Product Networks

$$P_1(X_1) = 0.1 \cdot I_{X_1} + 0.9 \cdot I_{\overline{X_1}}$$

$$P_2(X_1) = 0.2 \cdot I_{X_1} + 0.8 \cdot I_{\overline{X_1}}$$

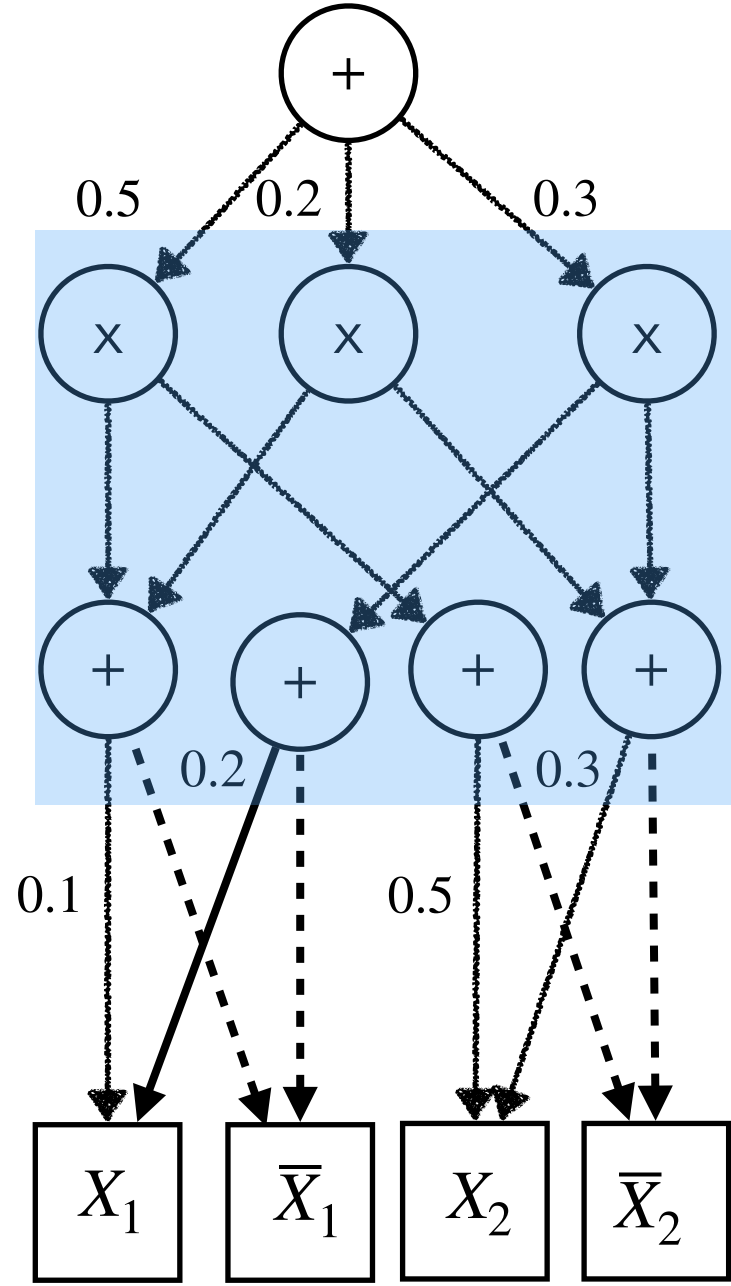
$$P_1(X_2) = 0.5 \cdot I_{X_2} + 0.5 \cdot I_{\overline{X_2}}$$

$$P_2(X_2) = 0.3 \cdot I_{X_2} + 0.7 \cdot I_{\overline{X_2}}$$

$$P_1(X_1, X_2) = P_1(X_1) \cdot P_2(X_2)$$

$$P_2(X_1, X_2) = P_1(X_1) \cdot P_2(X_2)$$

$$P_3(X_1, X_2) = P_2(X_1) \cdot P_2(X_2)$$



Tractable Models: Sum-Product Networks

$$P_1(X_1) = 0.1 \cdot I_{X_1} + 0.9 \cdot I_{\overline{X_1}}$$

$$P_2(X_1) = 0.2 \cdot I_{X_1} + 0.8 \cdot I_{\overline{X_1}}$$

$$P_1(X_2) = 0.5 \cdot I_{X_2} + 0.5 \cdot I_{\overline{X_2}}$$

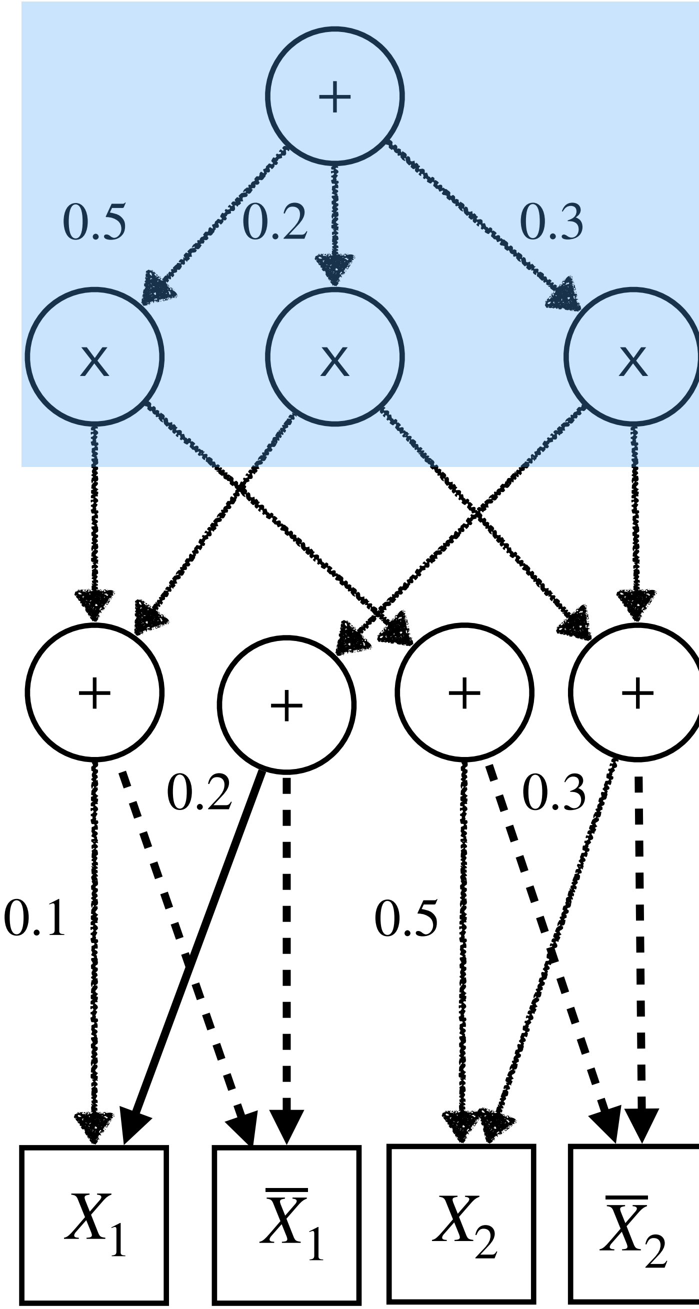
$$P_2(X_2) = 0.3 \cdot I_{X_2} + 0.7 \cdot I_{\overline{X_2}}$$

$$P_1(X_1, X_2) = P_1(X_1) \cdot P_2(X_2)$$

$$P_2(X_1, X_2) = P_1(X_1) \cdot P_2(X_2)$$

$$P_3(X_1, X_2) = P_2(X_1) \cdot P_2(X_2)$$

$$P(X_1, X_2) = 0.5 \cdot P_1(X_1, X_2) + 0.2 \cdot P_2(X_1, X_2) + 0.3 \cdot P_3(X_1, X_2)$$

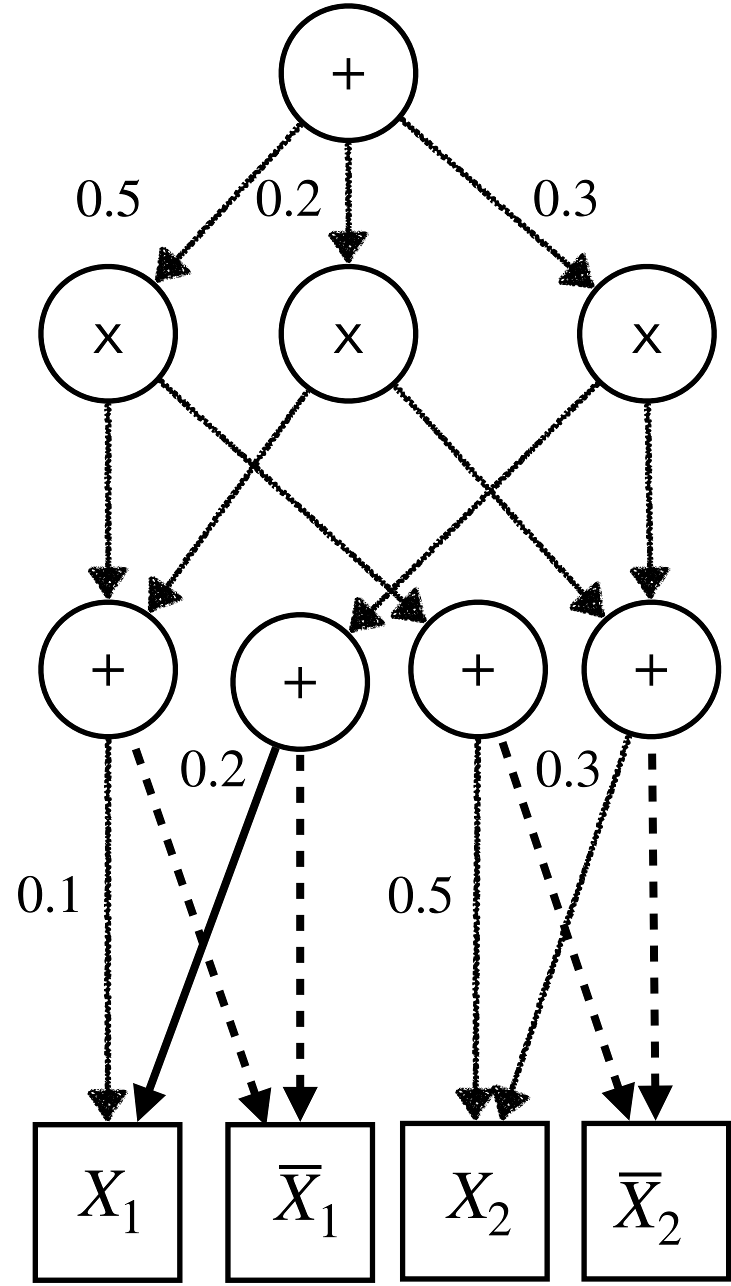


Tractable Models: Sum-Product Networks

$$P_1(X_1) = 0.1 \cdot I_{X_1} + 0.9 \cdot I_{\overline{X}_1}$$
$$P_2(X_1) = 0.2 \cdot I_{X_1} + 0.8 \cdot I_{\overline{X}_1}$$
$$P_1(X_2) = 0.5 \cdot I_{X_2} + 0.5 \cdot I_{\overline{X}_2}$$
$$P_2(X_2) = 0.3 \cdot I_{X_2} + 0.7 \cdot I_{\overline{X}_2}$$
$$P_1(X_1, X_2) = P_1(X_1) \cdot P_2(X_2)$$
$$P_2(X_1, X_2) = P_1(X_1) \cdot P_2(X_2)$$
$$P_3(X_1, X_2) = P_2(X_1) \cdot P_2(X_2)$$

$$P(X_1, X_2) = 0.5 \cdot P_1(X_1, X_2) + 0.2 \cdot P_2(X_1, X_2) + 0.3 \cdot P_3(X_1, X_2)$$

SPN2BN and BN2SPN transformations exists



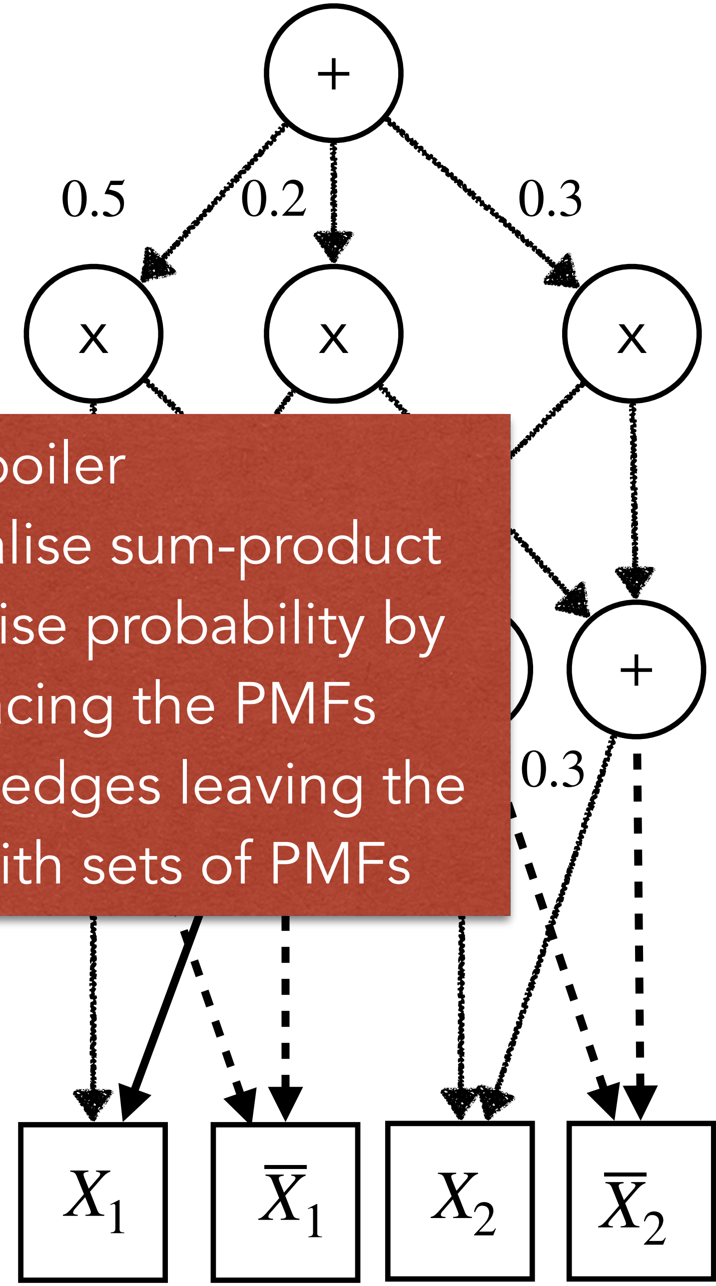
Tractable Models: Sum-Product Networks

$$P_1(X_1) = 0.1 \cdot I_{X_1} + 0.9 \cdot I_{\overline{X}_1}$$
$$P_2(X_1) = 0.2 \cdot I_{X_1} + 0.8 \cdot I_{\overline{X}_1}$$
$$P_1(X_2) = 0.5 \cdot I_{X_2} + 0.5 \cdot I_{\overline{X}_2}$$
$$P_2(X_2) = 0.3 \cdot I_{X_2} + 0.7 \cdot I_{\overline{X}_2}$$
$$P_1(X_1, X_2) = P_1(X_1) \cdot P_2(X_2)$$
$$P_2(X_1, X_2) = P_1(X_1) \cdot P_2(X_2)$$
$$P_3(X_1, X_2) = P_2(X_1) \cdot P_2(X_2)$$

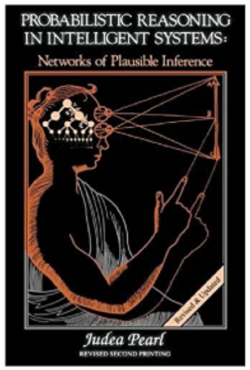
Spoiler
we can generalise sum-product nets to imprecise probability by simply replacing the PMFs annotating the edges leaving the sum nodes with sets of PMFs

$$P(X_1, X_2) = 0.5 \cdot P_1(X_1, X_2) + 0.2 \cdot P_2(X_1, X_2) + 0.3 \cdot P_3(X_1, X_2)$$

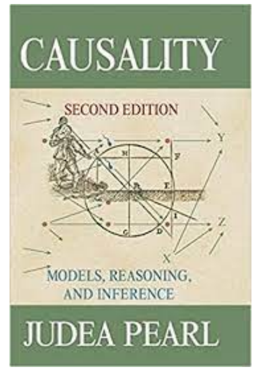
SPN2BN and BN2SPN transformations exists



(All PGMs Roads Lead to) Judea Pearl



Bayesian Nets
(~1988)

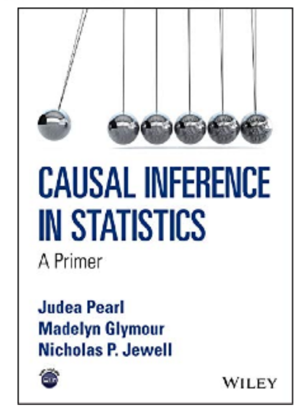


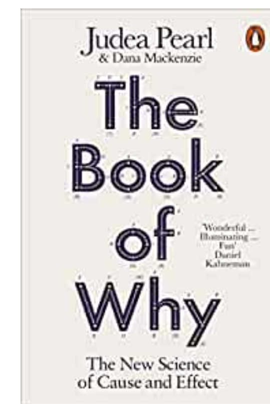
Do Calculus
(≤ 2000)

IV.

Credal
Nets
(~2000)

VII.





Structural Causal Models
(≤ 2016)

IV. CN = BN + CSs

Credal **N**ets as **B**ayesian **N**ets with (**C**redal) **S**et-Valued Parameters

Basic (Imprecise) Probability Theory

- Credal set (CS) over $K(X)$: (just) a set of PMFs $P(X)$ (over X)
- Expectation $\mathbb{E}[f] = \sum_{x\in\Omega_X} P(x)\cdot f(x)$ different for each $P(X) \in K(X)$, focus on lower (upper) bounds, e.g., $\underline{\mathbb{E}}[f] = \inf_{P(X)\in K(X)} \sum_x P(x)\cdot f(x)$

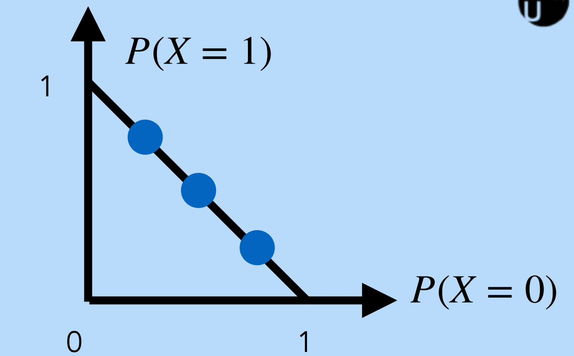
Basic (Imprecise) Probability Theory

- Credal set (CS) over $K(X)$: (just) a set of PMFs $P(X)$ (over X)
- Expectation $\mathbb{E}[f] = \sum_{x\in\Omega_X} P(x)\cdot f(x)$ different for each $P(X)\in K(X)$,

  focus on lower (upper) bounds, e.g., $\underline{\mathbb{E}}[f] = \inf_{P(X)\in K(X)} \sum_x P(x)\cdot f(x)$
- (for finite settings) bounds unaffected by convex hull (CH)
- This allows to focus on extreme points (ext), LP/combinatorial task:

$$\inf_{P(X)\in K(X)} \sum_x P(x)\cdot f(x) = \inf_{P\in \mathrm{CH}[K(X)]} \sum_x P(x)\cdot f(x) = \min_{P\in \mathrm{ext}[\mathrm{CH}[K(X)]]} \sum_x P(x)\cdot f(x)$$

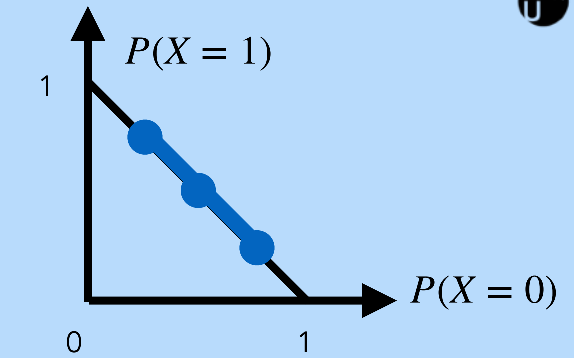
Basic (Imprecise) Probability Theory



- Credal set (CS) over $K(X)$: (just) a set of PMFs $P(X)$ (over X)
- Expectation $\mathbb{E}[f] = \sum_{x \in \Omega_X} P(x) \cdot f(x)$ different for each $P(X) \in K(X)$, focus on lower (upper) bounds, e.g., $\underline{\mathbb{E}}[f] = \inf_{P(X) \in K(X)} \sum_x P(x) \cdot f(x)$
- (for finite settings) bounds unaffected by convex hull (CH)
- This allows to focus on extreme points (ext), LP/combinatorial task:

$$\boxed{\inf_{P(X) \in K(X)} \sum_x P(x) \cdot f(x)} = \inf_{P \in \mathrm{CH}[K(X)]} \sum_x P(x) \cdot f(x) = \min_{P \in \mathrm{ext}[\mathrm{CH}[K(X)]]} \sum_x P(x) \cdot f(x)$$

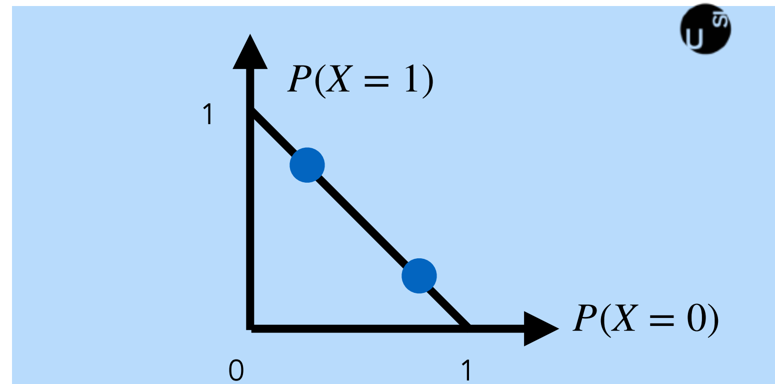
Basic (Imprecise) Probability Theory



- Credal set (CS) over $K(X)$: (just) a set of PMFs $P(X)$ (over X)
- Expectation $\mathbb{E}[f] = \sum_{x \in \Omega_X} P(x) \cdot f(x)$ different for each $P(X) \in K(X)$, focus on lower (upper) bounds, e.g., $\underline{\mathbb{E}}[f] = \inf_{P(X) \in K(X)} \sum_x P(x) \cdot f(x)$
- (for finite settings) bounds unaffected by convex hull (CH)
- This allows to focus on extreme points (ext), LP/combinatorial task:

$$\inf_{P(X) \in K(X)} \sum_x P(x) \cdot f(x) = \boxed{\inf_{P \in \mathrm{CH}[K(X)]} \sum_x P(x) \cdot f(x)} = \min_{P \in \mathrm{ext}[\mathrm{CH}[K(X)]]} \sum_x P(x) \cdot f(x)$$

Basic (Imprecise) Probability Theory



- Credal set (CS) over $K(X)$: (just) a set of PMFs $P(X)$ (over X)
- Expectation $\mathbb{E}[f] = \sum_{x\in\Omega_X} P(x)\cdot f(x)$ different for each $P(X)\in K(X)$, focus on lower (upper) bounds, e.g., $\underline{\mathbb{E}}[f] = \inf_{P(X)\in K(X)} \sum_x P(x)\cdot f(x)$
- (for finite settings) bounds unaffected by convex hull (CH)
- This allows to focus on extreme points (ext), LP/combinatorial task:

$$\inf_{P(X)\in K(X)} \sum_x P(x)\cdot f(x) = \inf_{P\in \mathrm{CH}[K(X)]} \sum_x P(x)\cdot f(x) = \min_{P\in \mathrm{ext}[\mathrm{CH}[K(X)]]} \sum_x P(x)\cdot f(x)$$

- Often focus on convex CSs with finite number of extreme points

Basic (Imprecise) Probability Theory

- Credal set (CS) over $K(X)$ = a set of PMFs $P(X)$ (over X)
- Expectation $\mathbb{E}[f] = \sum_{x \in \Omega_X} P(x) \cdot f(x)$ different for each $P(X) \in K(X)$,

  focus or

- (for finit
- This allo al task:

This is not enough to generalise BNs to CSs:
BNs are models based on independence,
we need an independence concept for CSs

$$\inf_{P(X) \in K(X)} \sum_x P(x) \cdot f(x) = \inf_{P \in \mathrm{CH}[K(X)]} \sum_x P(x) \cdot f(x) = \min_{P \in \mathrm{ext}[\mathrm{CH}[K(X)]]} \sum_x P(x) \cdot f(x)$$

- Often focus on convex CSs with finite number of extreme points

Independence Concepts with CSs

- Given CS $K(X, Y)$, what X independent of Y means? Irrelevant?
- X and Y (stochastically) independent $\forall P(X, Y) \in K(X, Y)$?
- So-called **strict** independence, does not preserve convexity ...

Independence Concepts with CSs

- Given CS $K(X,Y)$, what X independent of Y means? Irrelevant?
- X and Y (stochastically) independent $\forall P(X,Y) \in K(X,Y)$?
- So-called **strict** independence, does not preserve convexity …

Exercise: With X_1 and X_2 Boolean,

- $P(X_1, X_2)$ is a a 4-element normalised array
- The probability simplex is a tetrahedron
- Strictly independent PMFs are on a "maximally non-convex" surface inside such volume

Draw/imagine the surface (Solution: **https://www.geogebra.org/3d/e3rxtqjh**)

Independence Concepts with CSs

- Given CS $K(X, Y)$, what X independent of Y means? Irrelevant?
- X and Y (stochastically) independent $\forall P(X, Y) \in K(X, Y)$?
- So-called **strict** independence, does not preserve convexity ...
- X and Y (stochastically) independent $\forall P(X, Y) \in \mathrm{ext}[K(X, Y)]$!
- So-called **strong**, convenient choice for sensitivity analysis

Independence Concepts with CSs

- Given CS $K(X,Y)$, what X independent of Y means? Irrelevant?
- X and Y (stochastically) independent $\forall P(X,Y) \in K(X,Y)$?
- So-called **strict** independence, does not preserve convexity ...
- X and Y (stochastically) independent $\forall P(X,Y) \in \mathrm{ext}[K(X,Y)]$!
- So-called **strong**, convenient choice for sensitivity analysis
- Another popular concept is **epistemic irrelevance**: *lower and upper expectation of functions of X unaffected by Y (note that the concept is asymmetric and epistemic irrelevance $\neq$ independence)*
- *In general*, epistemic irrelevance gives more conservative inferences than strong independence, in some cases equal results

Credal Networks (CNs, Cozman, 2000)

- Simple idea: replace conditional PMFs in the CPTs with conditional CSs and obtain a joint CS (instead of a joint PMF)
- Each combination of valid CPT specifications defines a joint PMF satisfying the (stochastic) independence relations depicted by $\mathscr{G}$
- This would be a *strict* CN, inducing a non-convex joint CS
- In this sense, we are doing sensitivity analysis for BNs

Credal Networks (CNs, Cozman, 2000)

- Simple idea: replace conditional PMFs in the CPTs with conditional CSs and obtain a joint CS (instead of a joint PMF)
- Each combination of valid CPT specifications defines a joint PMF satisfying the (stochastic) independence relations depicted by $\mathscr{G}$
- This would be a *strict* CN, inducing a non-convex joint CS
- In this sense, we are doing sensitivity analysis for BNs
- Let us take the convex hull of the strict CS
- Good news: the vertices of convex hull of the strict joint CS are joint PMFs induced by BN whose parameters are the extreme points of the joint local CSs
- This corresponds to strong CNs and it allows to maintain a combinatorial nature in the model

(Strong) Credal Networks in Practice

- Directed acyclic graph $\mathscr{G}$ over variables $\mathbf{X} := (X_1, \ldots, X_n)$
- Assess CS $K(X_i \,|\, \mathrm{pa}_{X_i})$ for each $X_i \in \mathbf{X}$ and $\mathrm{pa}_{X_i} \in \Omega_{\mathrm{Pa}_{X_i}}$
- Build the joint CS ("strict extension") $K(\mathbf{X})$, i.e.,

$$K(\mathbf{X}) := \left\{ P(\mathbf{X}) : P(\mathbf{x}) = \prod_{i=1}^{n} P(x_i \,|\, \mathrm{pa}_{X_i}), \forall P(X_i \,|\, \mathrm{pa}_{X_i}) \in K(X_i \,|\, \mathrm{pa}_{X_i}) \right\}$$

(Strong) Credal Networks in Practice

- Directed acyclic graph $\mathscr{G}$ over variables $\mathbf{X} := (X_1, \ldots, X_n)$
- Assess CS $K(X_i | \mathrm{pa}_{X_i})$ for each $X_i \in \mathbf{X}$ and $\mathrm{pa}_{X_i} \in \Omega_{\mathrm{Pa}_{X_i}}$
- Build the joint CS ("strict extension") $K(\mathbf{X})$, i.e.,

$$K(\mathbf{X}) := \left\{ P(\mathbf{X}) : P(\mathbf{x}) = \prod_{i=1}^{n} P(x_i | \mathrm{pa}_{X_i}), \forall P(X_i | \mathrm{pa}_{X_i}) \in K(X_i | \mathrm{pa}_{X_i}) \right\}$$

- Compute inferences wrt such the strong extension, i.e., $\mathrm{CH}[K(\mathbf{X})]$
- Good news, the problem is combinatorial

$$\mathrm{ext}[\mathrm{CH}[K(\mathbf{X}]] \subseteq \{P(\mathbf{X} : P(\mathbf{x} = \prod_i P(x_i | \mathrm{pa}_{X_i}), P(X_i | \mathrm{pa}_{X_i}) \in \mathrm{ext}[K(X_i | \mathrm{pa}_{X_i})]\}$$

- CN as a (finite) collection of ("extreme") BNs over same $\mathscr{G}$

Let' play with notebook #4

Inference in Credal Networks

- Marginal query $\underline{P}(x_q) = \min_{P(X_i|\mathrm{pa}_{X_i}) \in \mathrm{ext}[K(X_i|\mathrm{pa}_{X_i})]} \sum_{\mathbf{X} \setminus \{X_q\}} \prod_{i=1}^{n} P(x_i \,|\, \mathrm{pa}_{X_i})$
- Closer to MMAP inference in BN, $\mathrm{NP}^{\mathrm{PP}}$-hard task

Inference in Credal Networks

- Marginal query $\underline{P}(x_q) = \min_{P(X_i|\mathrm{pa}_{X_i}) \in \mathrm{ext}[K(X_i|\mathrm{pa}_{X_i})]} \sum_{\mathbf{X}\setminus\{X_q\}} \prod_{i=1}^{n} P(x_i \,|\, \mathrm{pa}_{X_i})$
- Closer to MMAP inference in BN, NP^{PP}-hard task
- Posterior as a fractional task (num/den non indep optimisations)

$$\underline{P}(x_q \,|\, \mathbf{x}_E) = \min_{P(X_i|\mathrm{pa}_{X_i}) \in \mathrm{ext}[K(X_i|\mathrm{pa}_{X_i})]} \frac{\sum_{\mathbf{X}\setminus\{X_q\}} \prod_{i=1}^{n} P(x_i \,|\, \mathrm{pa}_{X_i})}{\sum_{\mathbf{X}\setminus\{X_q\}} \prod_{i=1}^{n} P(x_i \,|\, \mathrm{pa}_{X_i})} \neq \frac{\underline{P}(x_q, \mathbf{x}_E)}{\overline{P}(\mathbf{x}_E)}$$

Inference in Credal Networks

- Marginal query $\underline{P}(x_q) = \min_{P(X_i|\mathrm{pa}_{X_i}) \in \mathrm{ext}[K(X_i|\mathrm{pa}_{X_i})]} \sum_{\mathbf{X}\setminus\{X_q\}} \prod_{i=1}^{n} P(x_i|\mathrm{pa}_{X_i})$
- Closer to MMAP inference in BN, $\mathrm{NP}^{\mathrm{PP}}$-hard task
- Posterior as a fractional task (num/den non indep optimisations)

$$\underline{P}(x_q|\mathbf{x}_E) = \min_{P(X_i|\mathrm{pa}_{X_i}) \in \mathrm{ext}[K(X_i|\mathrm{pa}_{X_i})]} \frac{\sum_{\mathbf{X}\setminus\{X_q\}} \prod_{i=1}^{n} P(x_i|\mathrm{pa}_{X_i})}{\sum_{\mathbf{X}\setminus\{X_q\}} \prod_{i=1}^{n} P(x_i|\mathrm{pa}_{X_i})} \neq \frac{\underline{P}(x_q, \mathbf{x}_E)}{\overline{P}(\mathbf{x}_E)}$$

- Fast exact inference only in binary poly-trees (2U, Zaffalon, 1998) and epistemic trees (de Cooman et al., 2008)
- Fast approximated schemes and libraries
- E.g., LP inner approx (Antonucci et al., 2014)
- Credal MMAP? Harder than updating, depends on decision rule

Crema
Credal Models Algorithms

Inference in Credal Networks

- Marginal query $\underline{P}(x_q) = \min_{P(X_i|\mathrm{pa}_{X_i}) \in \mathrm{ext}[K(X_i|\mathrm{pa}_{X_i})]} \sum_{\mathbf{X}\setminus\{X_q\}} \prod_{i=1}^{n} P(x_i\,|\,\mathrm{pa}_{X_i})$
- Closer to MMAP inference in BN, $\mathrm{NP}^{\mathrm{PP}}$-hard task
- Posterior as a fractional task (num/den non indep optimisations)

Let's check this with notebook #4

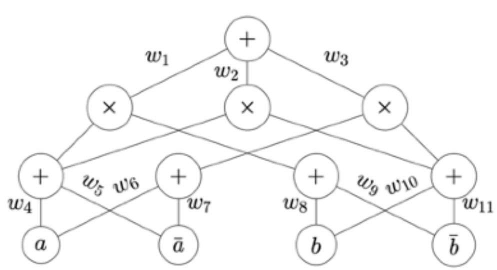
$$\underline{P}(x_q\,|\,\mathbf{x}_E) = \min_{P(X_i|\mathrm{pa}_{X_i}) \in \mathrm{ext}[K(X_i|\mathrm{pa}_{X_i})]} \frac{\sum_{\mathbf{X}\setminus\{X_q\}} \prod_{i=1}^{n} P(x_i\,|\,\mathrm{pa}_{X_i})}{\sum_{\mathbf{X}\setminus\{X_q\}} \prod_{i=1}^{n} P(x_i\,|\,\mathrm{pa}_{X_i})} \neq \frac{\underline{P}(x_q, \mathbf{x}_E)}{\overline{P}(\mathbf{x}_E)}$$

- Fast exact inference only in binary poly-trees (2U, Zaffalon, 1998) and epistemic trees (de Cooman et al., 2008)
- Fast approximated schemes and libraries
- E.g., LP inner approx (Antonucci et al., 2014)
- Credal MMAP? Harder than updating, depends on decision rule

Crema
Credal Models Algorithms

Credal Sum-Product Networks (Mauà et al., 2017)

- "Strict" semantics: SPNs do not rely on independence concepts
- Credality makes inference harder, but many tasks remain tractable
- Also credal SDDs (~ SPNs + logical constraints) (Mattei et al.,2019)
- No straightforward mappings CN2CSPN or CSPN2CN



$$
\begin{array}{ll}
(w_1, w_2, w_3) \in \mathrm{CH}(& [0.28, 0.45, 0.27], \\
\quad [0.18, 0.55, .27], & [0.18, 0.45, 0.37]), \\
0.54 \leq w_4 \leq 0.64, & 0.36 \leq w_5 \leq 0.46, \\
0.09 \leq w_6 \leq 0.19, & 0.81 \leq w_7 \leq 0.91, \\
0.27 \leq w_8 \leq 0.37, & 0.63 \leq w_9 \leq 0.73, \\
0.72 \leq w_{10} \leq 0.82, & 0.18 \leq w_{11} \leq 0.28, \\
w_4 + w_5 = 1, & w_6 + w_7 = 1, \\
w_8 + w_9 = 1, & w_{10} + w_{11} = 1
\end{array}
$$

Figure 2: A credal sum-product network over variables A and B.

Algorithm 3 Lower probability of evidence.

input: CSDD, evidence $\mathbf{e}$
for $n \leftarrow 1, \ldots, N$ **do**
  $\underline{\pi}(n) \leftarrow 0$
  if n is terminal, $n \neq \perp$ **then**
    $v \leftarrow$ leaf vtree node that n is normalized for
    $\underline{\pi}(n) \leftarrow \underline{\mathbb{P}}_n(\mathbf{e}_v)$
  else
    $((p_i, s_i)_{i=1}^k, \mathbb{K}_n(P)) \leftarrow n$ (decision node)
    $\underline{\pi}(n) \leftarrow \min_{[\theta_1,\ldots,\theta_k] \in \mathbb{K}_n(P)} \sum_{i=1}^k \underline{\pi}(p_i) \cdot \underline{\pi}(s_i) \cdot \theta_i$
  end if
end for
output: $\underline{\mathbb{P}}(\mathbf{e}) \leftarrow \underline{\pi}(N)$

V. CN4DSS

Knowledge-Based **D**ecision-**S**upport **S**ystems by **C**redal **N**etworks

Knowledge-Based Decision-Support Systems

- Aka **Expert Systems**, very popular GOFAI tools
- Less hype in the DL age, but annotated data are costly and in practice lot of people still use such models (e.g. BNs)
- Why CNs?

Knowledge-Based Decision-Support Systems

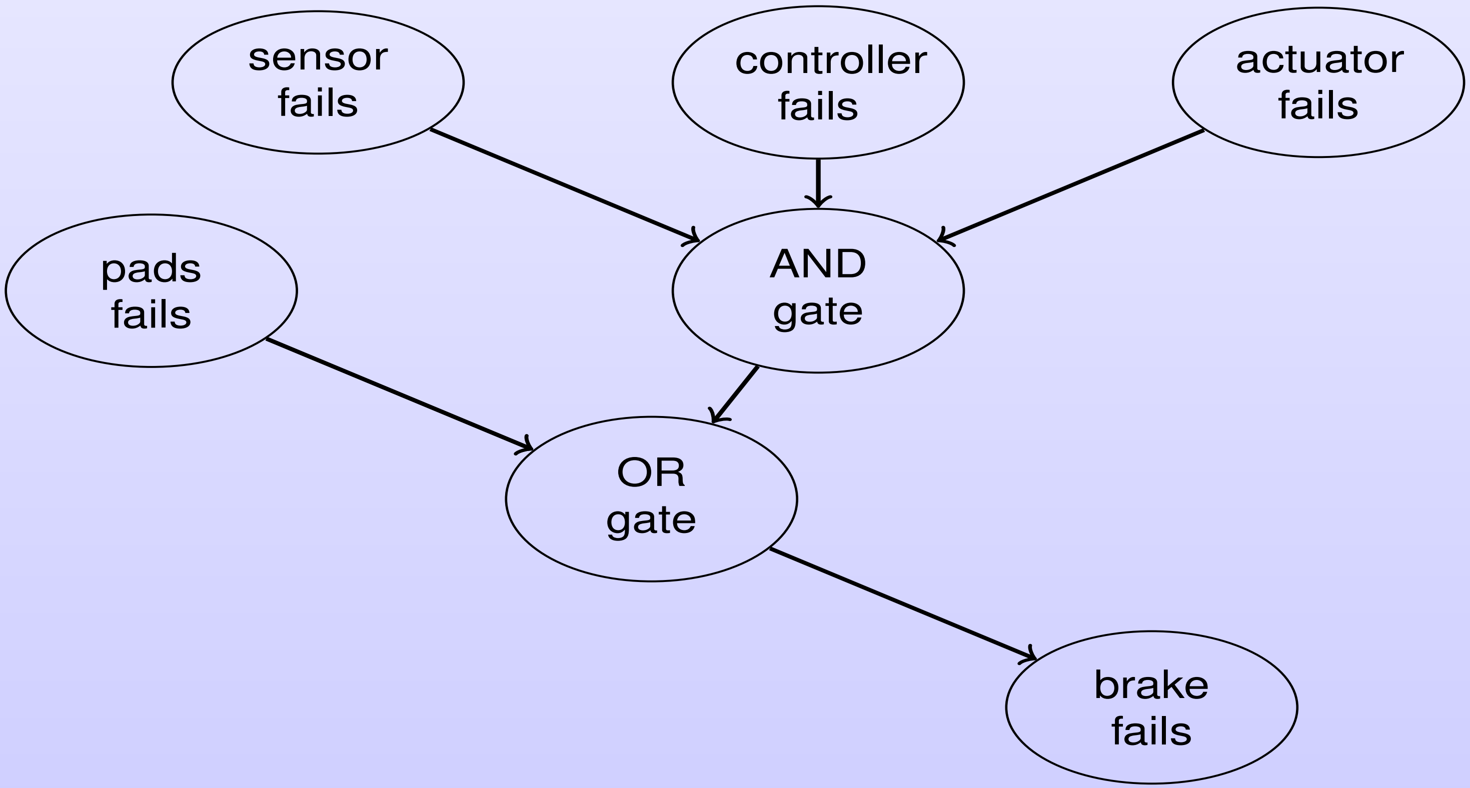
- Aka **Expert Systems**, very popular GOFAI tools
- Less hype in the DL age, but annotated data are costly and in practice lot of people still use such models (e.g. BNs)
- Why CNs? CS can be a better model of the expert (uncertain) knowledge. Knowledge engineering by CSs:
  - Models of complete **ignorance** (vacuous CS)
  - E.g., conservative updating for non-MAR missing data
  - Qualitative assessments by **probability intervals**
  - Preferences as inequality constraints
  - Positive/negative influence or synergy

Let's check gallery #1

Fault Trees (Vesely, 1981)

brake fails = [pads $\vee$ (sensor $\wedge$ controller $\wedge$ actuator)]

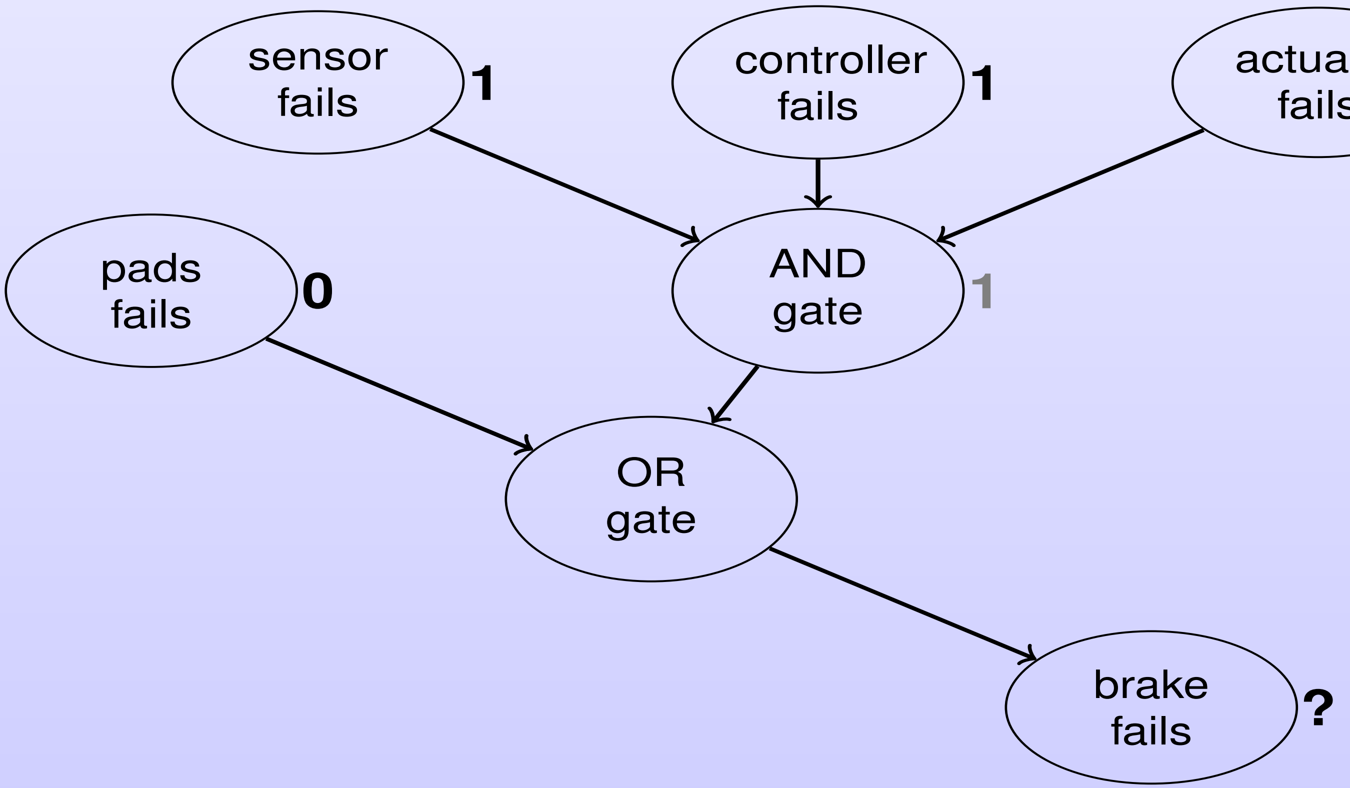
devices failures are independent



Fault Trees (Vesely, 1981)

brake fails = [pads ∨ (sensor ∧ controller ∧ actuator)]

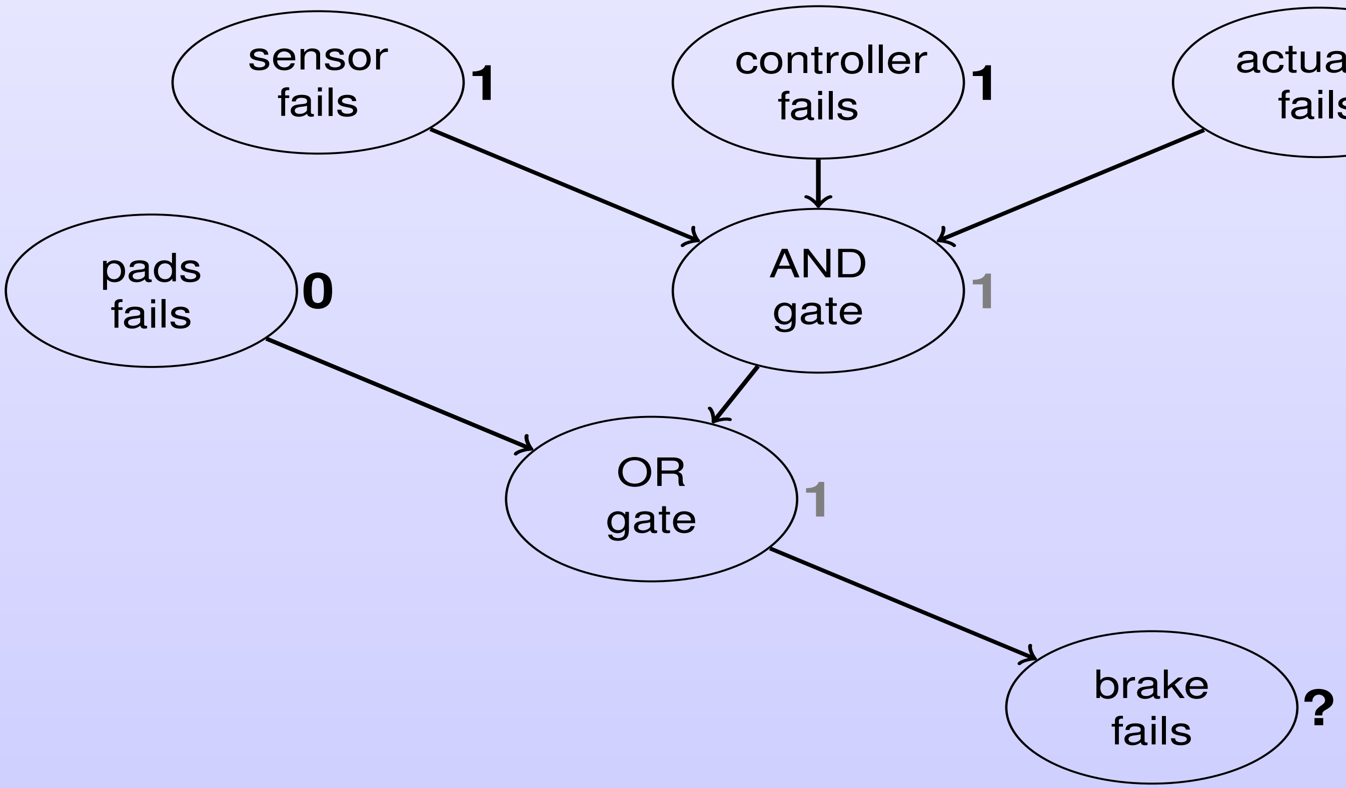
devices failures are independent



Fault Trees (Vesely, 1981)

brake fails = [pads $\vee$ (sensor $\wedge$ controller $\wedge$ actuator)]

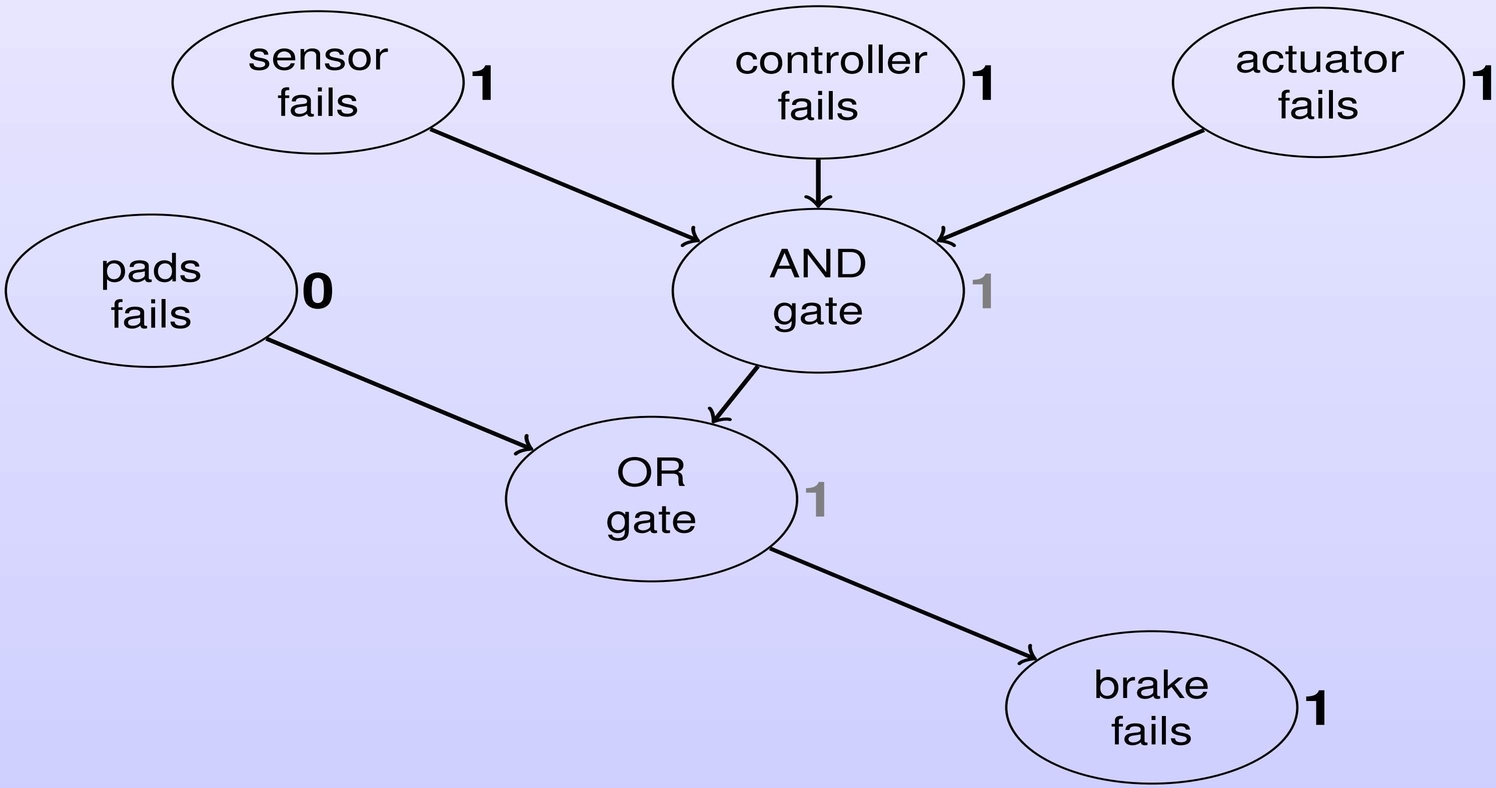
devices failures are independent



Fault Trees (Vesely, 1981)

brake fails = [pads $\vee$ (sensor $\wedge$ controller $\wedge$ actuator)]

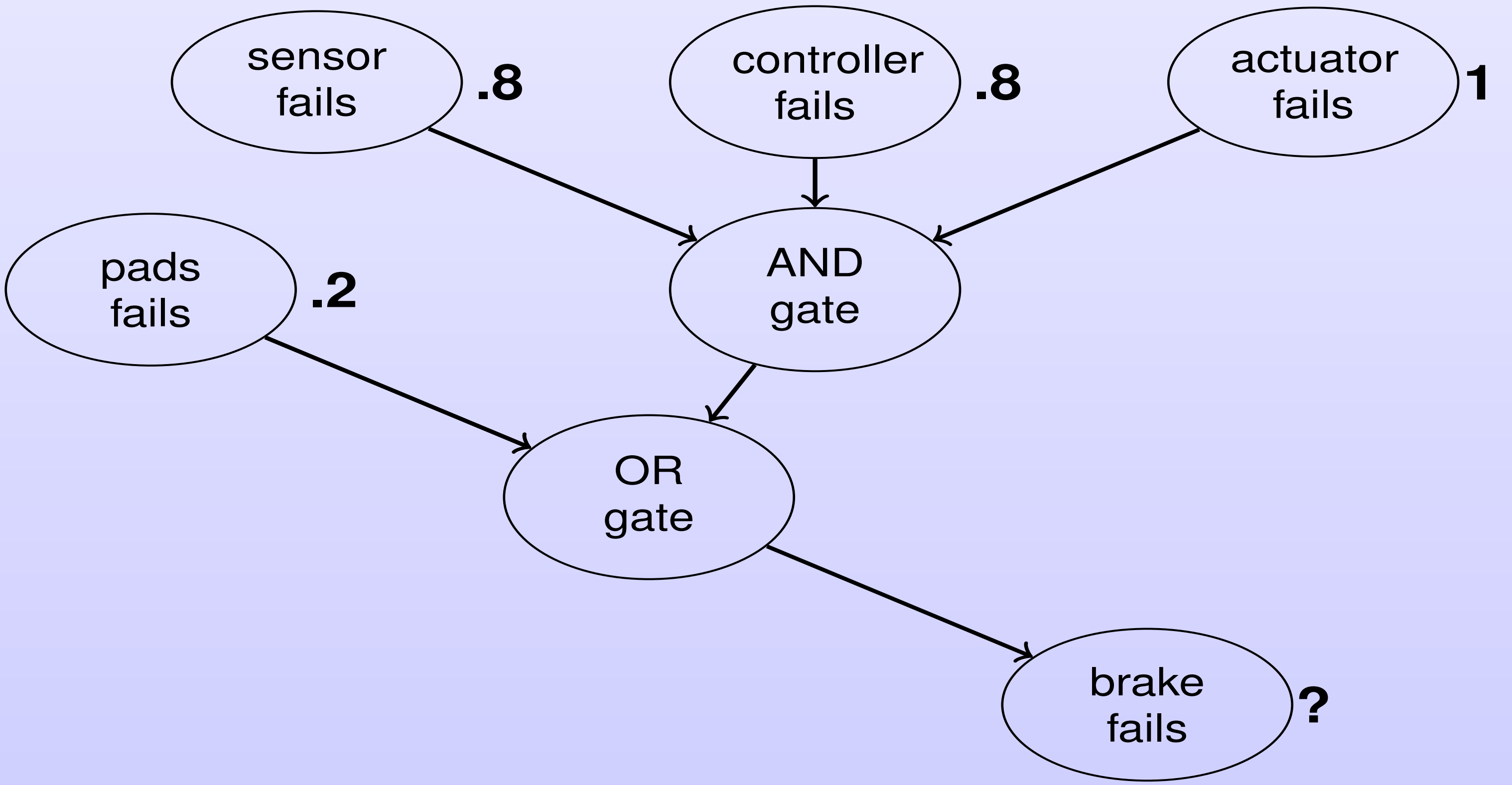
devices failures are independent



Fault Trees (Vesely, 1981)

brake fails = [pads $\vee$ (sensor $\wedge$ controller $\wedge$ actuator)]

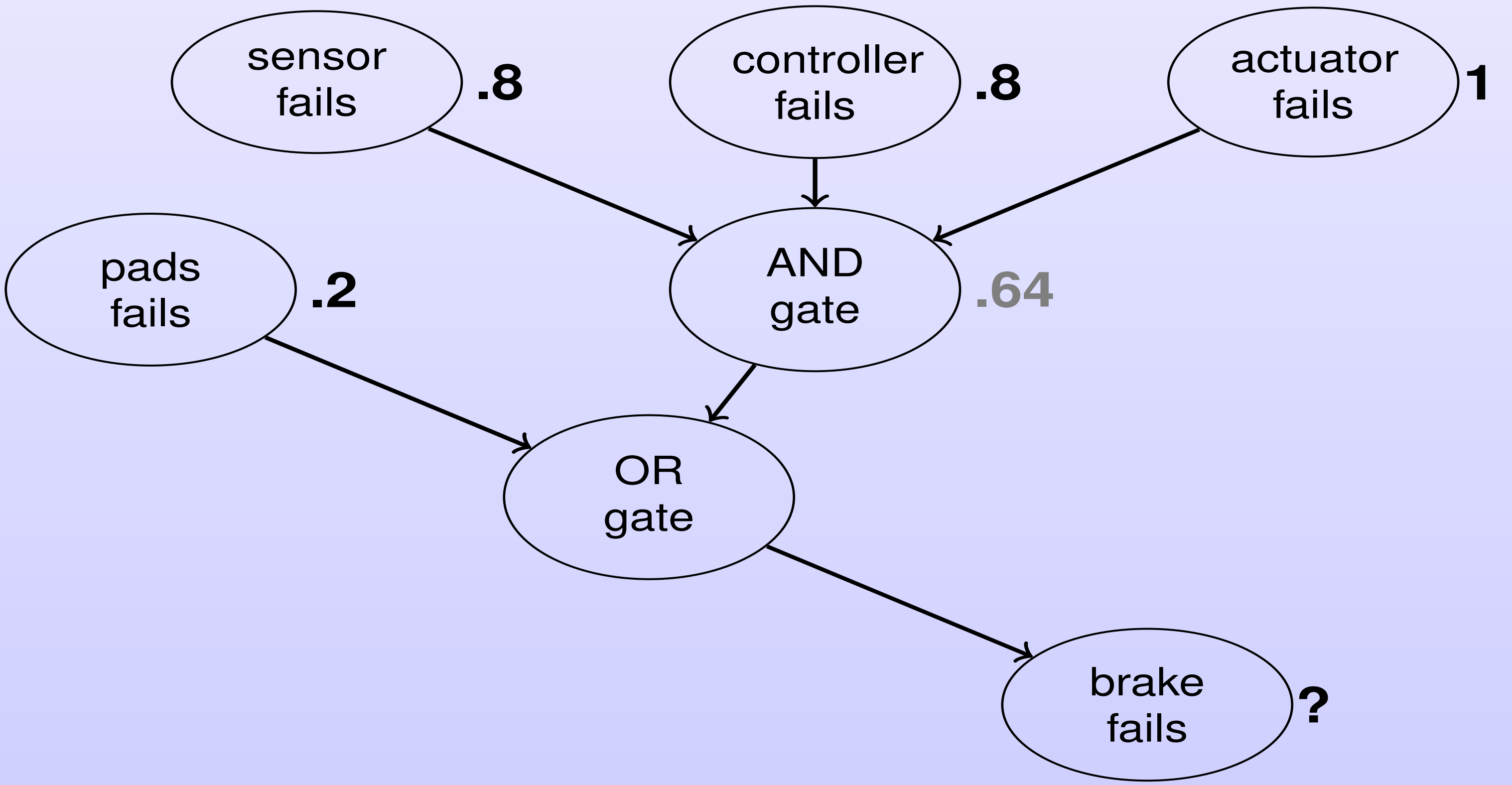
devices failures are independent



Fault Trees (Vesely, 1981)

brake fails = [pads $\vee$ (sensor $\wedge$ controller $\wedge$ actuator)]

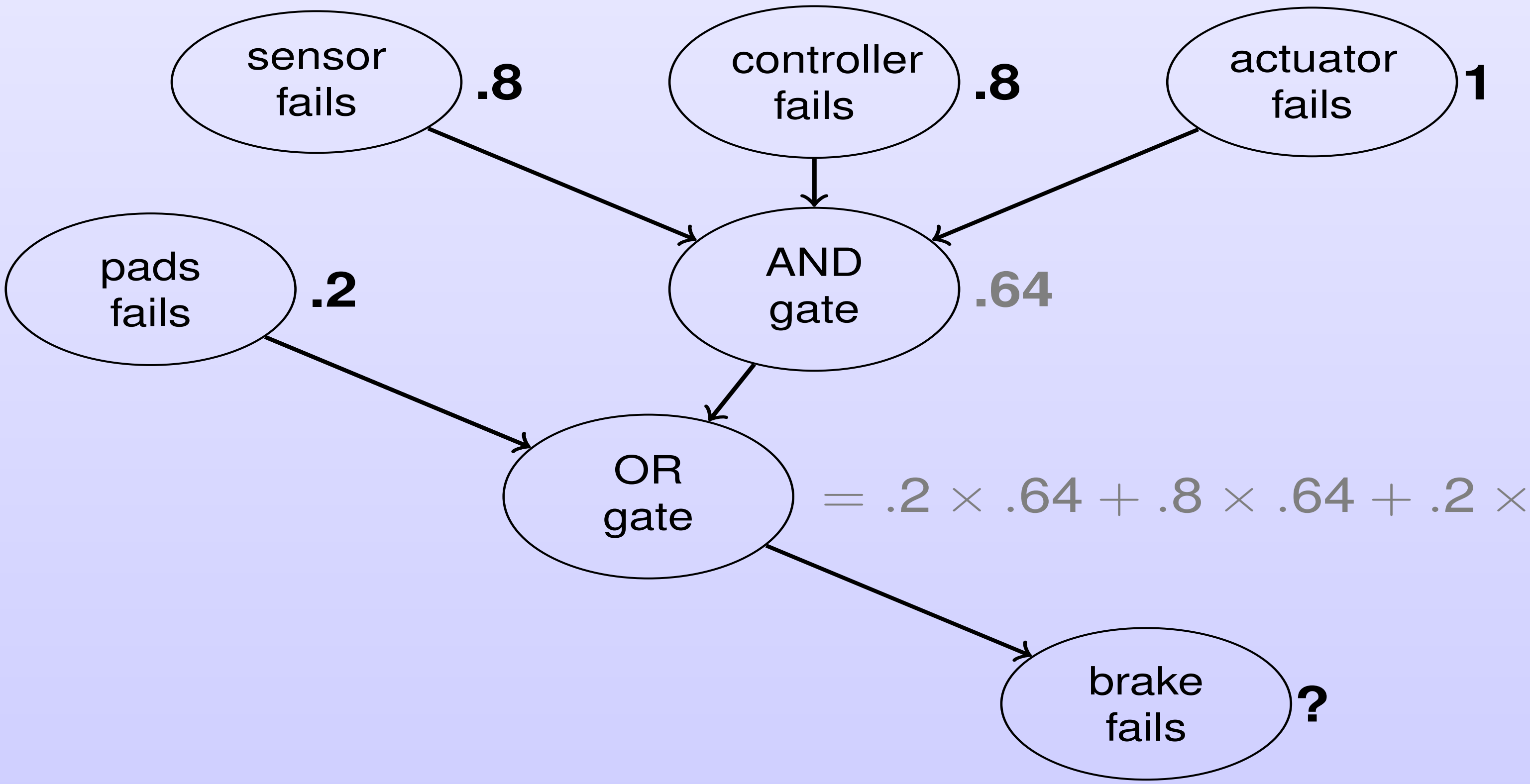
devices failures are independent



Fault Trees (Vesely, 1981)

brake fails = [pads $\vee$ (sensor $\wedge$ controller $\wedge$ actuator)]

devices failures are independent

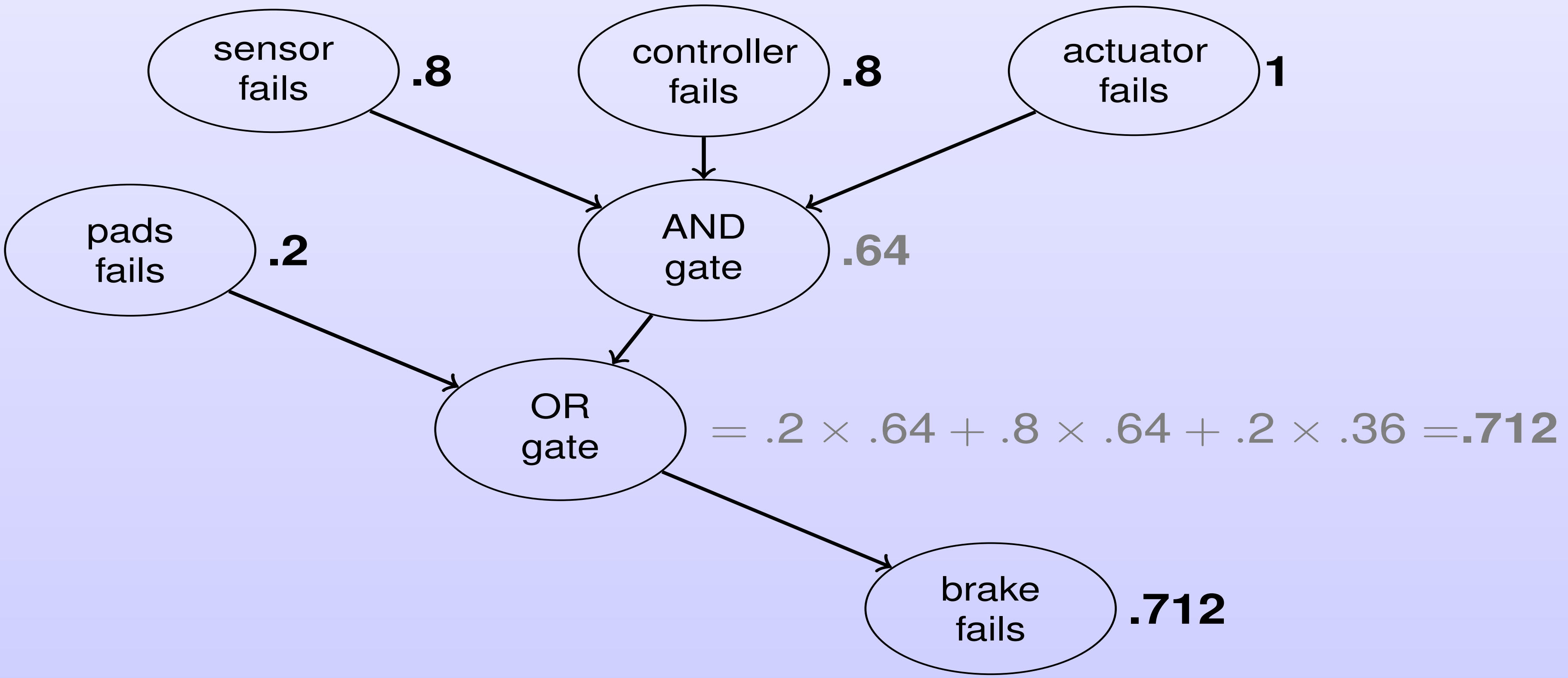


$= .2 \times .64 + .8 \times .64 + .2 \times .36 = \mathbf{.712}$

Fault Trees (Vesely, 1981)

brake fails = [pads $\vee$ (sensor $\wedge$ controller $\wedge$ actuator)]

devices failures are independent

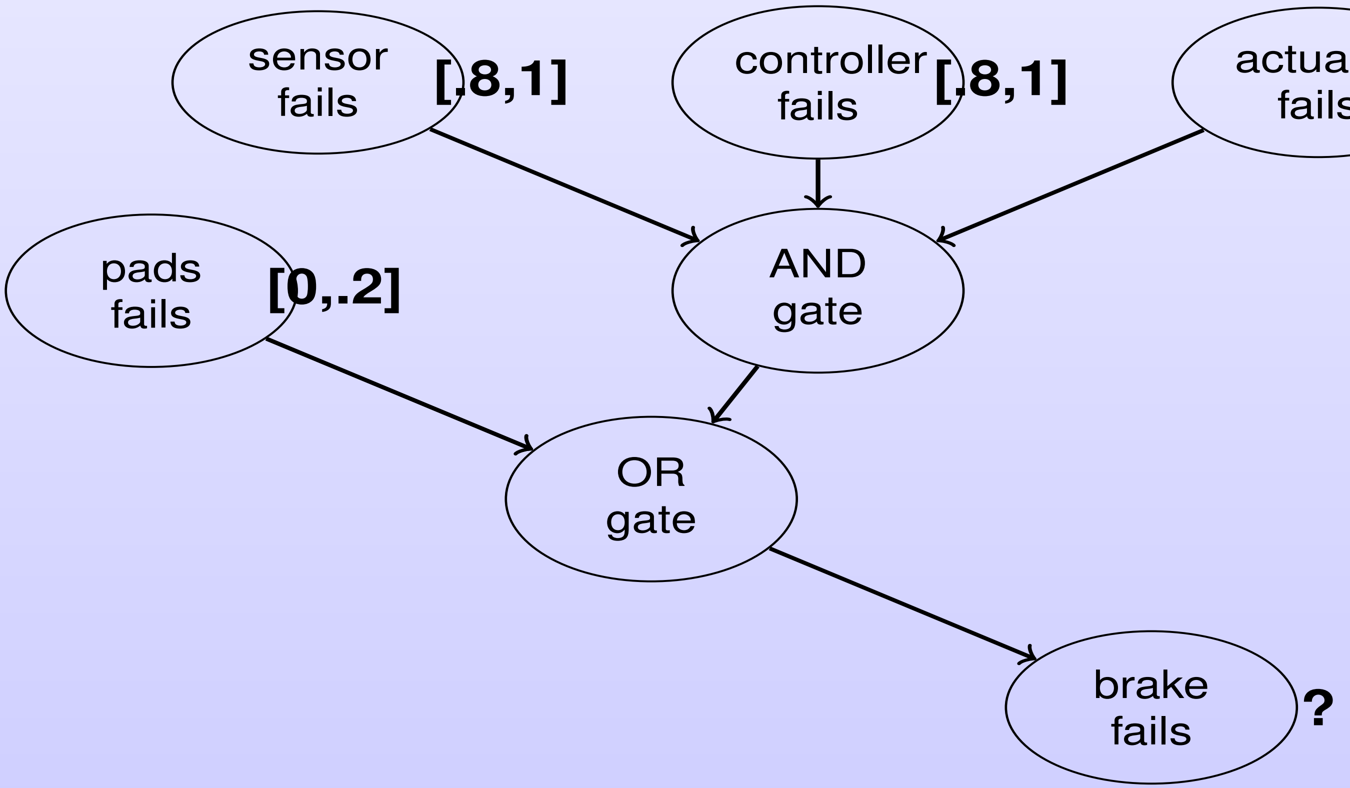


IDSIA SUPSI

Fault Trees (Vesely, 1981)

brake fails = [pads $\vee$ (sensor $\wedge$ controller $\wedge$ actuator)]

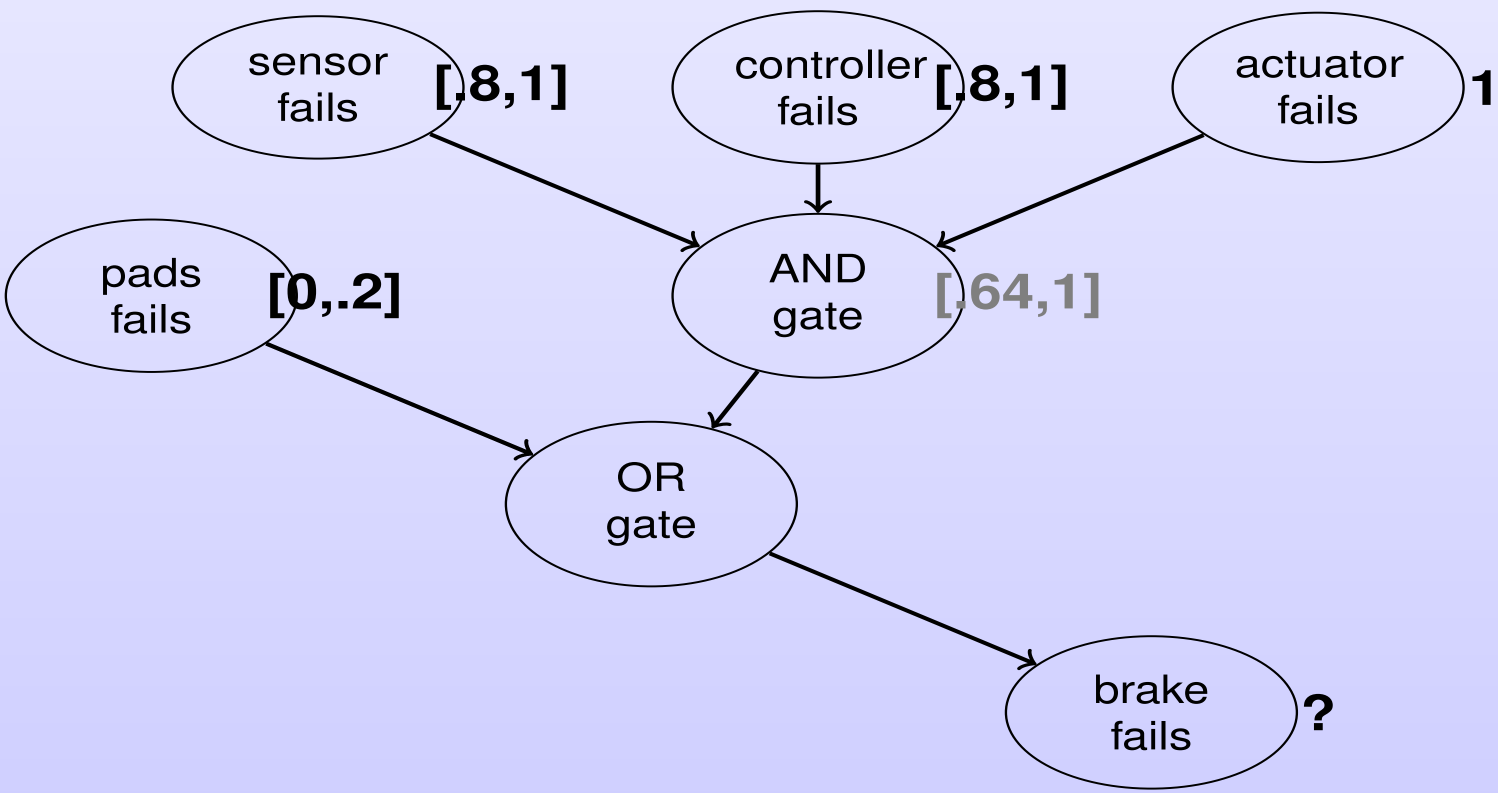
devices failures are independent



Fault Trees (Vesely, 1981)

brake fails = [pads $\vee$ (sensor $\wedge$ controller $\wedge$ actuator)]

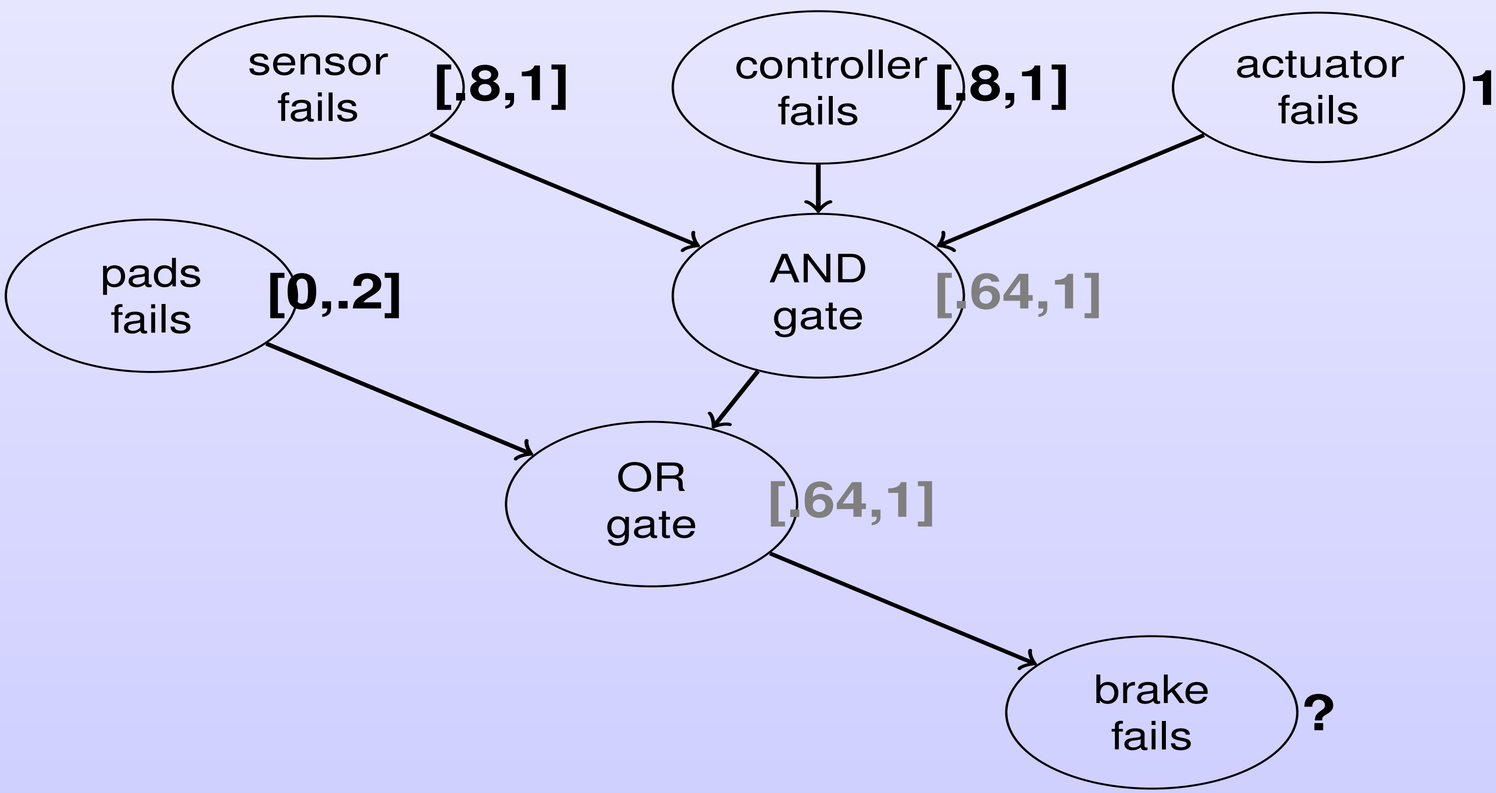
devices failures are independent



Fault Trees (Vesely, 1981)

brake fails = [pads $\vee$ (sensor $\wedge$ controller $\wedge$ actuator)]

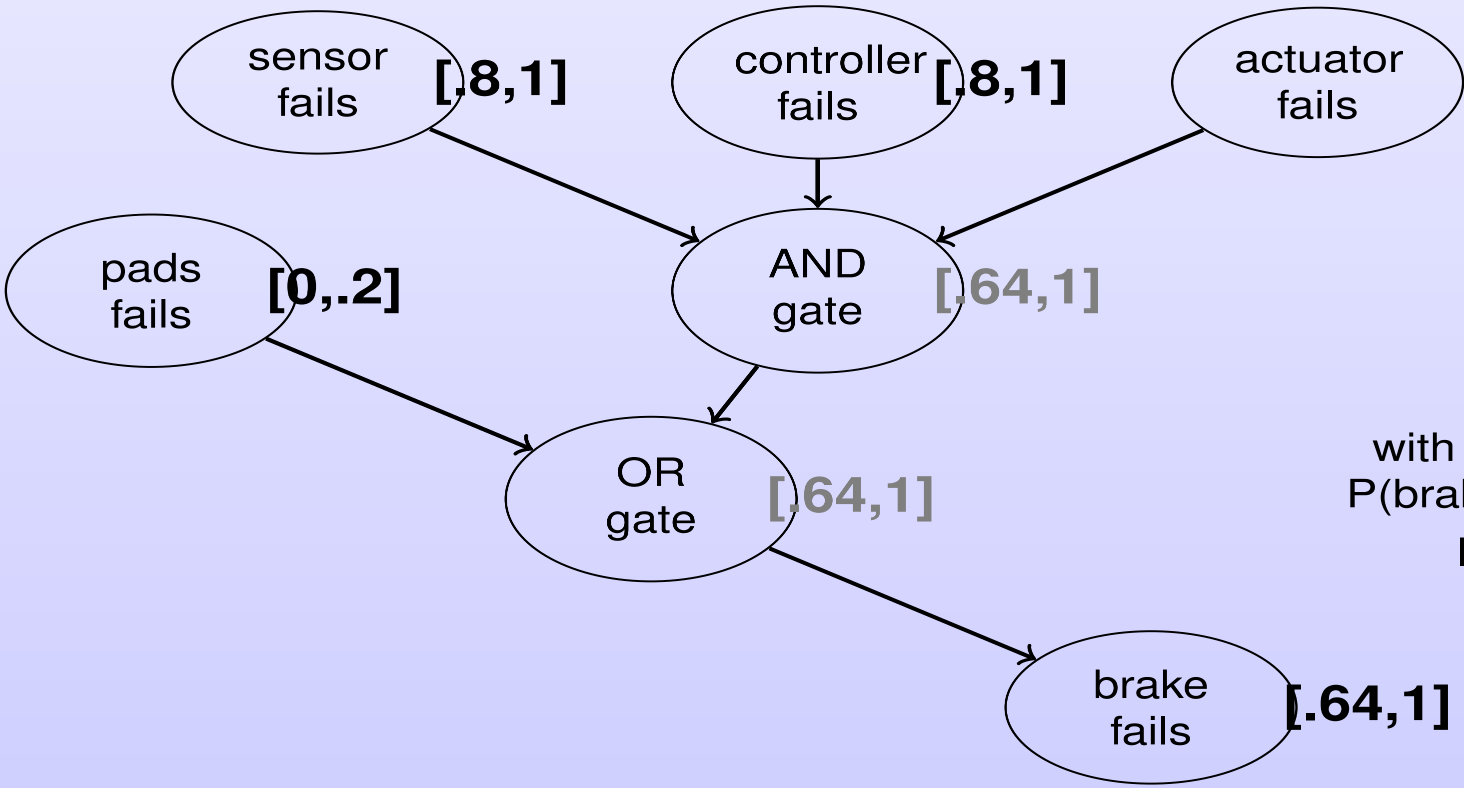
devices failures are independent



Fault Trees (Vesely, 1981)

brake fails = [pads $\vee$ (sensor $\wedge$ controller $\wedge$ actuator)]

devices failures are independent



with [.7, 1] instead
P(brake fails)$\in$[.49,1]
Indecision!

Exercise on Knowledge-Based Decision-Support Systems

- **Build a DSS** based on your knowledge based on a CN
- Decide the variables (few)
- Define a (correlational) graph over them
- Express your (uncertain) knowledge about the state of each variable given its parents
- Use this CN to extract decision-support information (for inference we do brute-force here)
- Lack of ideas? Let's check the examples in **gallery #2**
- Or let's define a simple 3-node CN/DSS together

VI. (C)ML

Credal Machine Learning

(Credal) Machine Learning with CNs

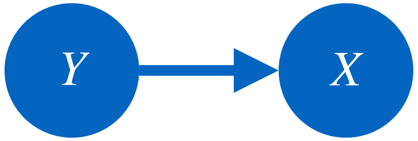
- Statistical learning with CNs?
  - $\mathcal{G}$? Structural BN learning or assumptions (ex. naive/TAN)
  - CSs? IDM (or likelihood-based) approaches
  - Decisions? E.g., maximality, undominated classes.
- These are **credal classifiers**, possibly returning multiple options
- With IDM we say that an instance is:
  - indeterminate if different priors lead to different classes
  - robust if all the priors gives the same class
- BN compatible classifier? Good accuracy on robust instances, inaccurate on the indeterminate ones

(Credal) Machine Learning with CNs

- Statistical learning with CNs?
  - $\mathscr{G}$? Structural BN learning or assumptions (ex. naive/TAN)
  - CSs? IDM (or likelihood-based) approaches
  - Decisions? E.g., maximality, undominated classes.
- These ar ble options
- With ID
  - indeterminate if different priors lead to different classes
  - robust if all the priors gives the same class
- BN compatible classifier? Good accuracy on robust instances, inaccurate on the indeterminate ones

Let's quickly browse **gallery 3** with its credal classifiers and their evaluation

IDM & (Personal) Considerations on CML



- Global IDM constraints between CSs specification
- Local? Easier optimisation, but more imprecise
- ML conferences can be very selective and credal approaches not always well-perceived. However, strong IP papers have been accepted (e.g., Hüllermeier/Caprio/Destercke/de Campos/...)
- Distinguishing problem formulation from the corresponding optimisation might be a good practice, if the experiments are good, not having an exact ad hoc solution might be still ok

X	Y	n
0	0	3+t
0	1	4+u
1	0	2+w
1	1	1+(1-t-u-w)

one virtual instance (ESS=1)

IDM constraints

$$0 \leq t, u, w \leq 1 \qquad t + u + w \leq 1$$

$$P(Y=0) = \frac{5+t+w}{11}$$

$$P(X=0 \mid Y=0) = \frac{3+t}{5+t+w}$$

Constraints between

$K(X)$ and $K(X=0 \mid Y=0)$

VII. SCMs ≡ CNs

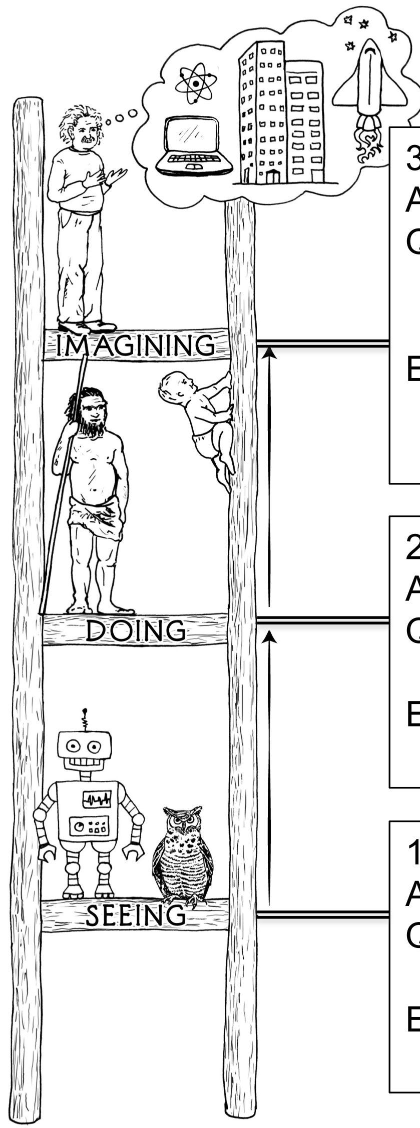
Structural Causal Models are (solvable by) Credal Networks

Pearl's Ladder of Causation and the Need for a Causal AI

(Causal) AI?

RL

ML/DL



3-LEVEL HIERARCHY

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 the last 2 years?

2. INTERVENTION
ACTIVITY: Doing, Intervening
QUESTIONS: *What if I do . . . ? How?*
(What would Y be if I do X?)
EXAMPLES: If I take aspirin, will my headache be cured? What if we ban cigarettes?

1. ASSOCIATION
ACTIVITY: Seeing, Observing
QUESTIONS: *What if I see . . . ?*
(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?

Source: The Book of Why, Pearl & Mc Kenzie

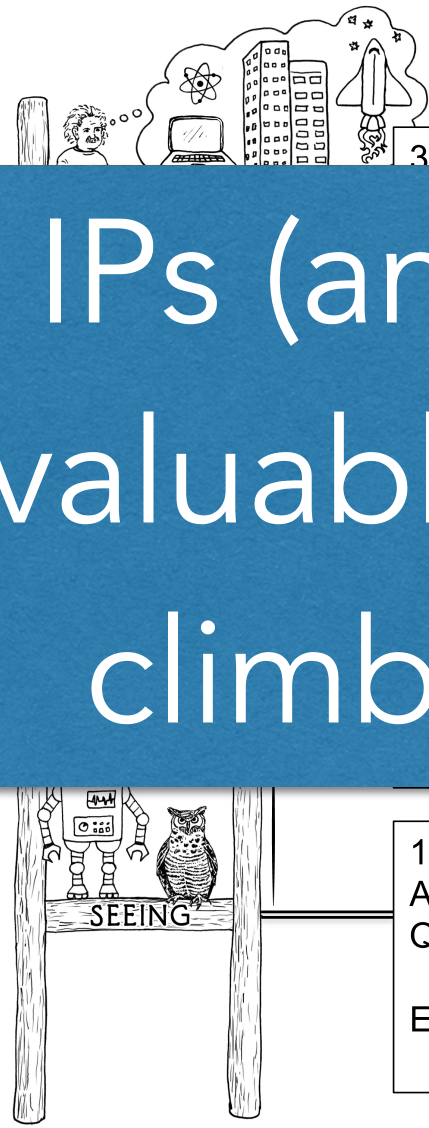
Pearl's Ladder of Causation and the Need for a Causal AI

(Cau l)

A

R

ML/DL



3-LEVEL HIERARCHY

3. COUNTERFACTUALS

s?

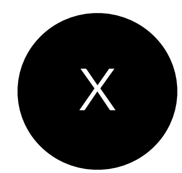
IPs (and CNs) as a valuable support to climb the ladder

1. ASSOCIATION
ACTIVITY: Seeing, Observing
QUESTIONS: *What if I see . . . ?*
(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?

Source: The Book of Why, Pearl & Mc Kenzie

Structural Causal Models

- Manifest **endogenous** variable X
- Observations $\mathscr{D}$ available
- From $\mathscr{D}$ statistical learning of $P(X)$



Boolean X

$P(X = 0) = p$

Structural Causal Models

- Manifest **endogenous** variable X
- Observations $\mathscr{D}$ available
- From $\mathscr{D}$ statistical learning of $P(X)$
- A latent **exogenous** variable U

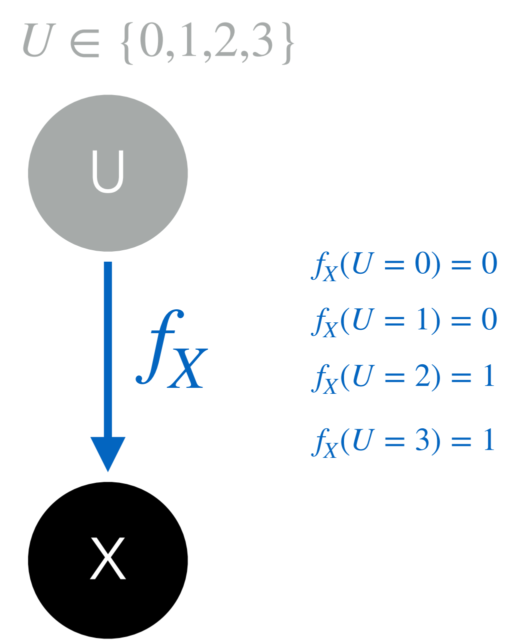
$U \in \{0,1,2,3\}$

Boolean X

$P(X = 0) = p$

Structural Causal Models

- Manifest **endogenous** variable X
- Observations $\mathscr{D}$ available
- From $\mathscr{D}$ statistical learning of $P(X)$
- A latent **exogenous** variable U
- States of U determines those of X through a **structural equation** f_X
  f_X surjective but not invertible

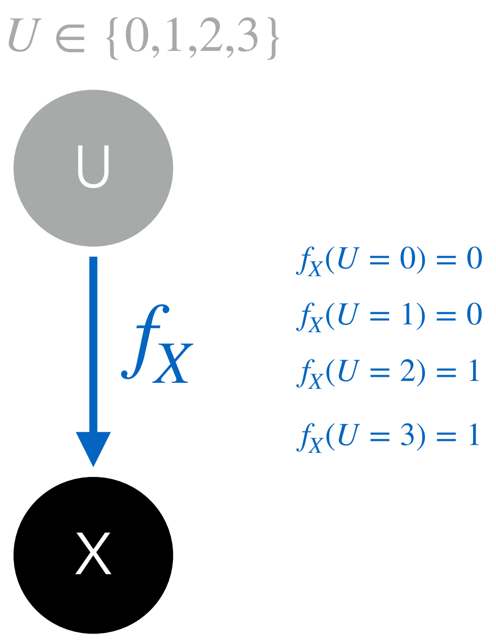


Boolean X

$P(X=0)=p$

Structural Causal Models

- Manifest **endogenous** variable X
- Observations $\mathscr{D}$ available
- From $\mathscr{D}$ statistical learning of $P(X)$
- A latent **exogenous** variable U
- States of U determines those of X through a **structural equation** f_X
  f_X surjective but not invertible
- $P(x) = \sum_x P(x|u)P(u) = \sum_u \delta_{f(u),x} P(u)$



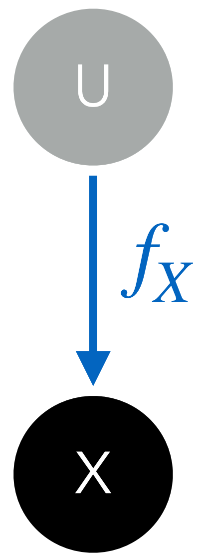
Boolean X
$P(X=0)=p$

Structural Causal Models

- Manifest **endogenous** variable X
- Observations $\mathscr{D}$ available
- From $\mathscr{D}$ statistical learning of $P(X)$
- A latent **exogenous** variable U
- States of U determines those of X through a **structural equation** f_X
  f_X surjective but not invertible
- $P(x) = \sum_x P(x\,|\,u)P(u) = \sum_u \delta_{f(u),x} P(u)$
- A $P(U)$ giving $P(X)$? More than one!

$$P(U) = \left[\frac{p}{2}, \frac{p}{2}, \frac{1-p}{2}, \frac{1-p}{2}\right]$$

$U \in \{0,1,2,3\}$



$f_X(U=0) = 0$
$f_X(U=1) = 0$
$f_X(U=2) = 1$
$f_X(U=3) = 1$

Boolean X
$P(X=0) = p$

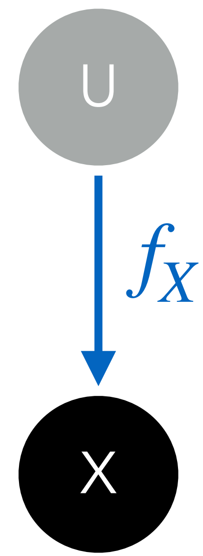
Structural Causal Models

- Manifest **endogenous** variable X
- Observations $\mathscr{D}$ available
- From $\mathscr{D}$ statistical learning of $P(X)$
- A latent **exogenous** variable U
- States of U determines those of X through a **structural equation** f_X
  f_X surjective but not invertible
- $P(x) = \sum_x P(x|u)P(u) = \sum_u \delta_{f(u),x} P(u)$
- $P(U)$ giving $P(X)$? More than one!
- Causal inference to be based on the credal set $K(U)$ compatible with $P(X)$

$K(U) = \{P(U) : P(U=0) + P(U=1) = p\}$

$P(U) = \left[\frac{p}{2}, \frac{p}{2}, \frac{1-p}{2}, \frac{1-p}{2}\right]$

$U \in \{0,1,2,3\}$



$f_X(U=0)=0$

$f_X(U=1)=0$

$f_X(U=2)=1$

$f_X(U=3)=1$

Boolean X

$P(X=0)=p$

Structural Causal Models

$$K(U) = \{P(U) : P(U=0) + P(U=1) = p\}$$

$$P(U) = \left[\frac{p}{2}, \frac{p}{2}, \frac{1-p}{2}, \frac{1-p}{2}\right]$$

$U \in \{0,1,2,3\}$

- Manifest **endogenous** variable X
- Observations $\mathscr{D}$ available
- From $\mathscr{D}$ statistical learning of $P(X)$
- A lat
- State
  throu
  f_X su

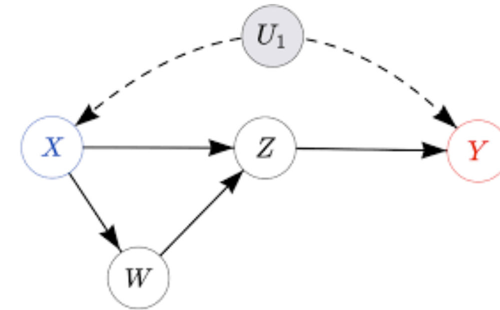
This is a (minimalistic) structural causal model

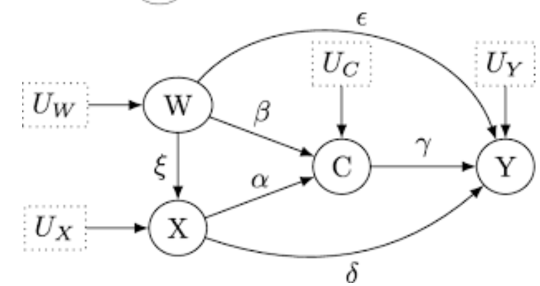
- $P(x) = \sum_x P(x|u)P(u) = \sum_u \delta_{f(u),x} P(u)$
- $P(U)$ giving $P(X)$? More than one!
- Causal inference to be based on the credal set $K(U)$ compatible with $P(X)$

Boolean X
$P(X=0) = p$

Structural Causal Models (General Definition)

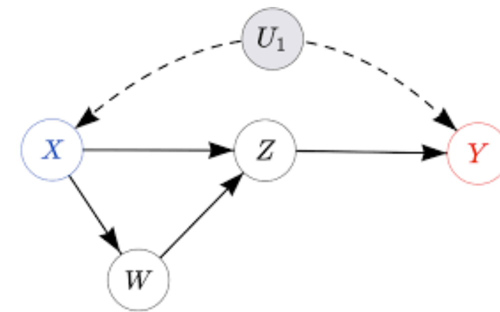
- $\mathbf{X} := (X_1, \ldots, X_n)$ (endogenous variables)
- $\mathbf{U} := (U_1, \ldots, U_m)$ (exogenous variables)
- Directed graph $\mathscr{G}$ assumed to be semi-Markovian = root in $\mathbf{U}$, non-root in $\mathbf{X}$
- Equation $X = f_X(\mathrm{Pa}_X)$ for each $X \in \mathbf{X}$
- Marginal $P(U)$ for $U \in \mathbf{U}$ (assessed if possible)

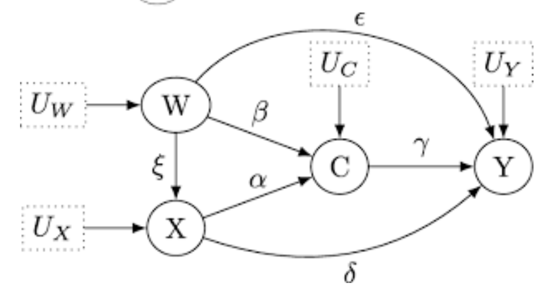




Structural Causal Models (General Definition)

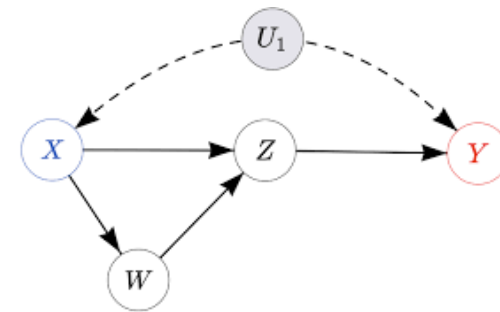
- $\mathbf{X} := (X_1, \ldots, X_n)$ (endogenous variables)
- $\mathbf{U} := (U_1, \ldots, U_m)$ (exogenous variables)
- Directed graph $\mathscr{G}$ assumed to be semi-Markovian = root in $\mathbf{U}$, non-root in $\mathbf{X}$
- Equation $X = f_X(\mathrm{Pa}_X)$ for each $X \in \mathbf{X}$
- Marginal $P(U)$ for $U \in \mathbf{U}$ (assessed if possible)
- SCM = BN with CPTs $P(X|\mathrm{Pa}_X) = \delta_{X, f_X(\mathrm{Pa}_X)}$
- Joint PMF $P(\mathbf{x}, \mathbf{u}) = \prod_{U \in \mathbf{U}} P(u) \prod_{X \in \mathbf{X}} \delta_{f_X(\mathrm{pa})_X, x}$

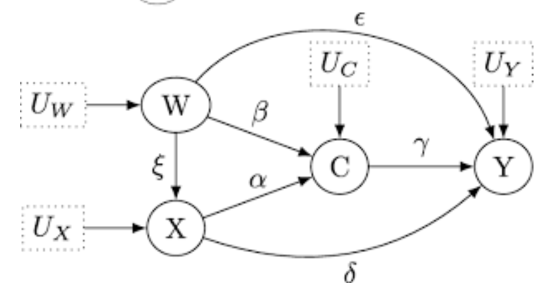


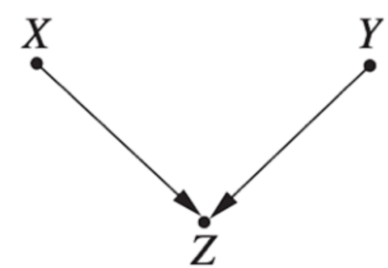


Structural Causal Models (General Definition)

- $\mathbf{X} := (X_1, \ldots, X_n)$ (endogenous variables)
- $\mathbf{U} := (U_1, \ldots, U_m)$ (exogenous variables)
- Directed graph $\mathcal{G}$ assumed to be semi-Markovian = root in $\mathbf{U}$, non-root in $\mathbf{X}$
- Equation $X = f_X(\mathrm{Pa}_X)$ for each $X \in \mathbf{X}$
- Marginal $P(U)$ for $U \in \mathbf{U}$ (assessed if possible)
- SCM = BN with CPTs $P(X|\mathrm{Pa}_X) = \delta_{X, f_X(\mathrm{Pa}_X)}$
- Joint PMF $P(\mathbf{x}, \mathbf{u}) = \prod_{U \in \mathbf{U}} P(u) \prod_{X \in \mathbf{X}} \delta_{f_X(\mathrm{pa})_X, x}$
- Here discrete vars, continuous case analogous







$U = \{X, Y\}, \quad V = \{Z\}, \quad F = \{f_Z\}$

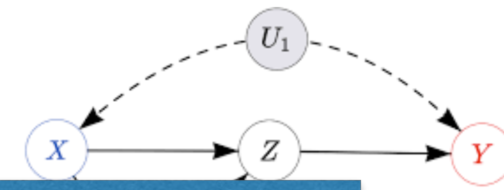
$f_Z : Z = 2X + 3Y$

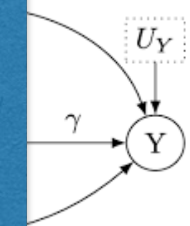
Structural Causal Models (General Definition)

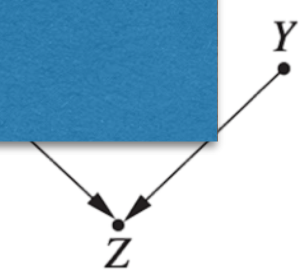
- $\mathbf{X} := (X_1, \ldots, X_n)$ (endogenous variables)
- $\mathbf{U} :=$
- Direc
  semi
- Equa
- Marg
- SCM

SCMs as (one of) the most powerful tools for causal analyses

- Joint PMF $P(\mathbf{x}, \mathbf{u}) = \prod_{U \in \mathbf{U}} P(u) \prod_{X \in \mathbf{X}} \delta_{f_X(\mathrm{pa})_X, x}$
- Here discrete vars, continuous case analogous







$U = \{X, Y\}, \quad V = \{Z\}, \quad F = \{f_Z\}$
$f_Z : Z = 2X + 3Y$

Headache Example (Staying on the First Rung)

- You take aspirin ($X=1$) and headache vanishes ($Y=1$)
- Probability that this has been due to aspirin?
- Observational data $\mathscr{D}$ about the two variables available
- From $\mathscr{D}$, $P(Y=0 \,|\, X=0) = 0.5 > P(Y=0 \,|\, X=1) = 0.1$

$$X \bullet\!\longrightarrow\!\bullet\, Y$$

X	Y	n
0	0	...
0	1	....
1	0	....
1	1	....

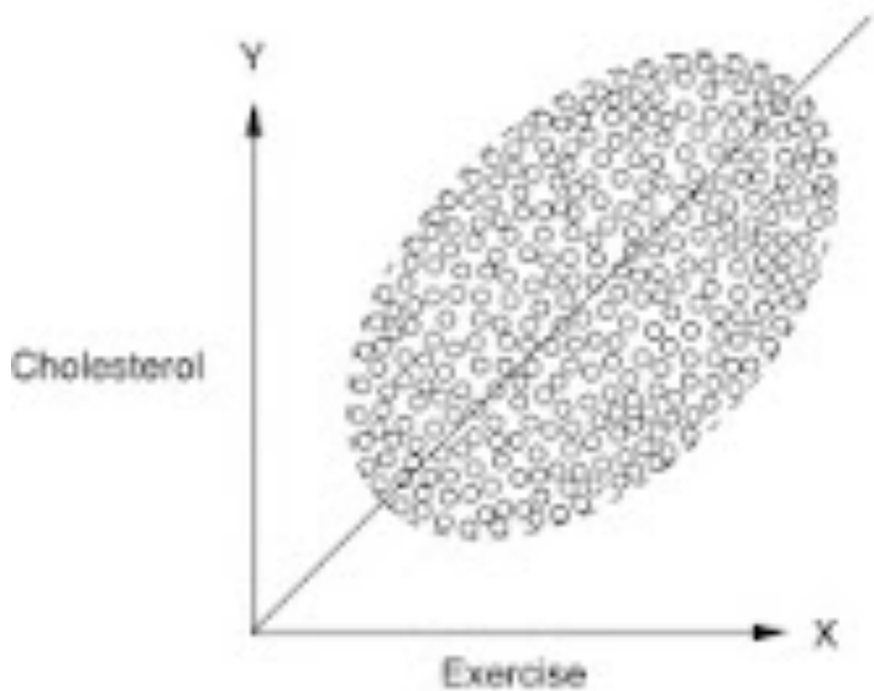
Headache Example (Staying on the First Rung)

$$X \longrightarrow Y$$

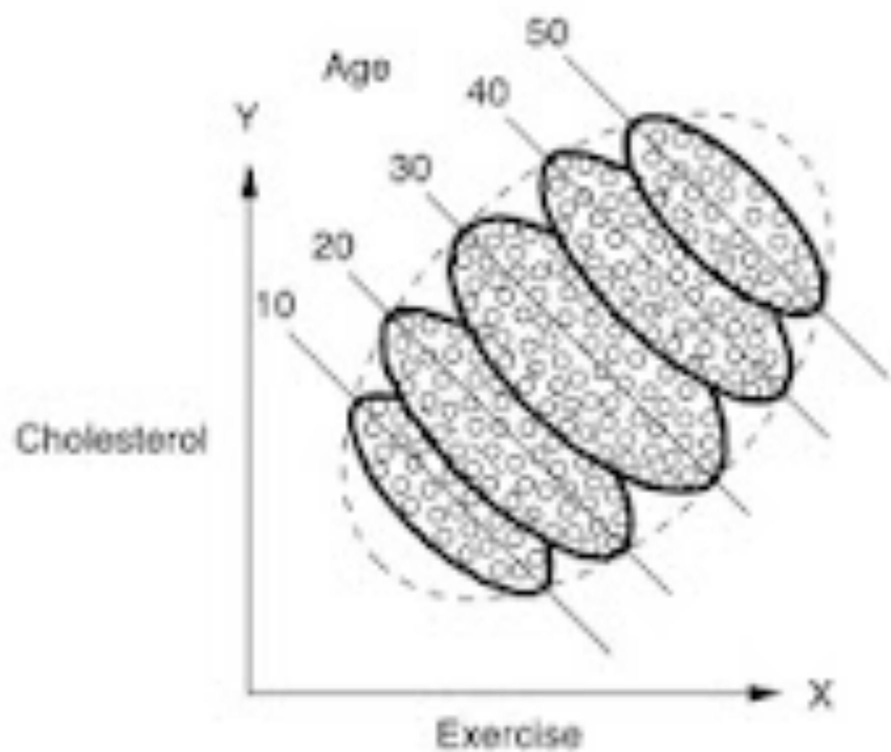
X	Y	n
0	0	...
0	1	....
1	0	....
1	1	....

- You take aspirin ($X = 1$) and headache vanishes ($Y = 1$)
- Probability that this has been due to aspirin?
- Observational data $\mathscr{D}$ about the two variables available
- From $\mathscr{D}$, $P(Y = 0 | X = 0) = 0.5 > P(Y = 0 | X = 1) = 0.1$
- Not genuine causal analysis: adding further covariates might give contradictory results (Simpson's paradox)
- $P(Y = 0 | X = 0, Z = z) < P(Y = 0 | X = 1, Z = z) \ \forall z$

Sport might seriously hurt your vascular health?



Sport might seriously hurt your vascular health? No!



Simpson's Paradox or Gender Bias?

UC Berkley in 1973

	TOTAL		MEN		WOMEN	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	4526	39%	2691	45%	1835	30%

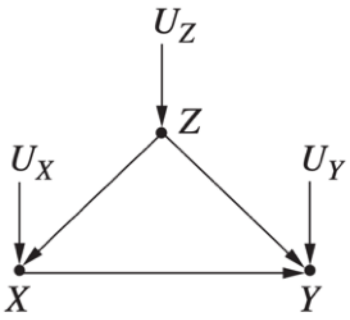
DEPT.	TOTAL		MEN		WOMEN	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	4526	39%	2691	45%	1835	30%
A	933	64%	825	62%	108	82%
B	585	63%	560	63%	25	68%
C	918	35%	325	37%	593	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	393	24%
F	714	6%	373	6%	341	7%

DEPT.	TOTAL		MEN		WOMEN	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	4526	39%	2691	45%	1835	30%
A	933	64%	**825**	62%	108	82%
B	585	63%	**560**	63%	25	68%
C	918	35%	325	37%	**593**	34%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%

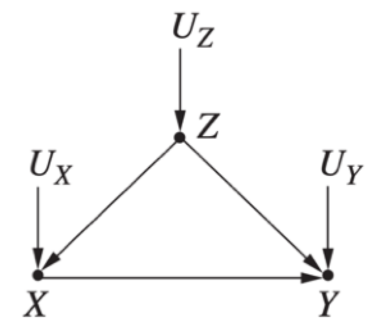
DEPT.	TOTAL		MEN		WOMEN	
	Applicants	Admitted	Applicants	Admitted	Applicants	Admitted
Total	4526	39%	2691	45%	1835	30%
D	792	34%	417	33%	375	35%
E	584	25%	191	28%	**393**	24%
F	714	6%	373	6%	341	7%

Time to climb up the ladder

Take the Aspirin! (Interventions = Second Rung)

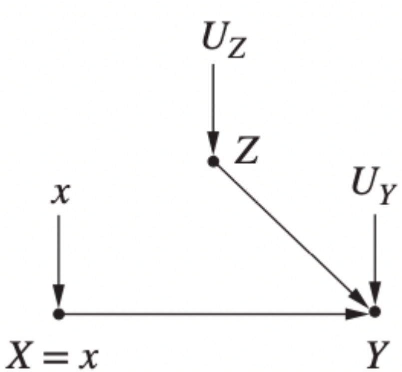


- Gender Z as an additional (endogenous) variable
- Markovian $\mathscr{G}$ (one exo parent for each endo)
- Force people to take aspirin = **intervention** $\mathrm{do}(X = 1)$

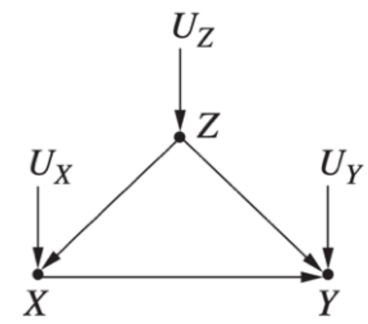


Take the Aspirin! (Interventions = Second Rung)

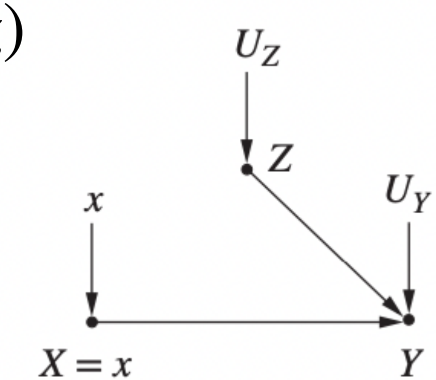
- Gender Z as an additional (endogenous) variable
- Markovian $\mathscr{G}$ (one exo parent for each endo)
- Force people to take aspirin = **intervention** $\mathrm{do}(X=1)$
- f_X should be modified (constant output), after a **surgery** on $\mathscr{G}$ (incoming arcs removed) intervention = observation



Take the Aspirin! (Interventions = Second Rung)

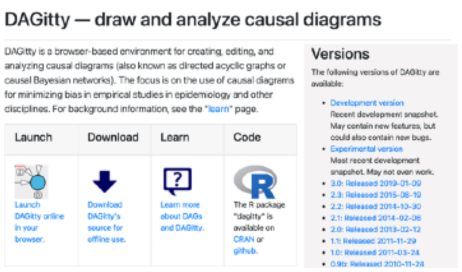


- Gender Z as an additional (endogenous) variable
- Markovian $\mathscr{G}$ (one exo parent for each endo)
- Force people to take aspirin = **intervention** $\mathrm{do}(X=1)$
- f_X should be modified (constant output), after a **surgery** on $\mathscr{G}$ (incoming arcs removed) intervention = observation
- Pearl's **do calculus** allows to reduce interventional queries to observational ones (solved by BN inference)
- E.g., backdoor $P(y\,|\,\mathrm{do}(X=x)) = \sum_z P(y\,|\,x,z)\cdot P(z)$
- Do calculus only needs $\mathscr{G}$ (and not the SCM)!



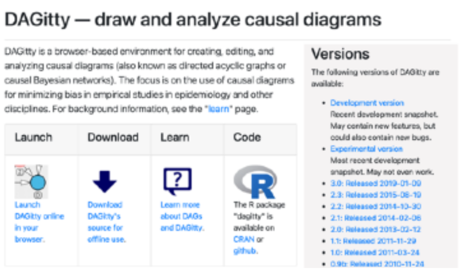
Identifiability of Causal Queries

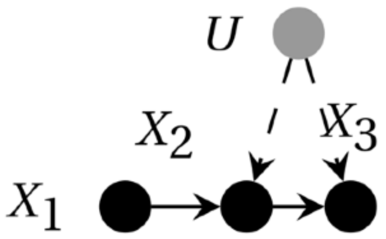
- Do calculus reduces interventional to observational queries by exploiting d-separation in SCMs
- Sound and complete (graph-theoretic) algorithm + inference in the empirical joint PMF
- Alternatively: surgery and inference in the SCM ...



Identifiability of Causal Queries

- Do calculus reduces interventional to observational queries by exploiting d-separation in SCMs
- Sound and complete (graph-theoretic) algorithm + inference in the empirical joint PMF
- Alternatively: surgery and inference in the SCM …
- Not all queries can be computed by do calculus. If not we call the query **unidentifiable**
- Emerging idea: unidentifiable queries are only partially identifiable (bounds can be estimated!)
- Recent works in this field by various groups



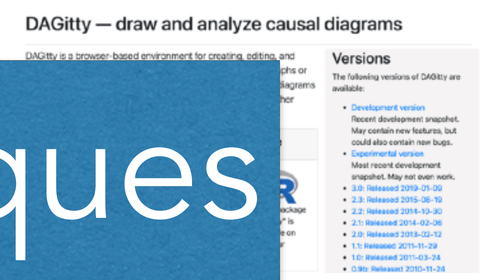


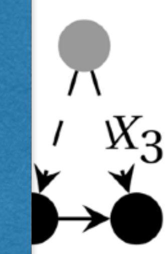
$P(x_3 \,|\, \mathrm{do}(x_2)) \in [l, u]$

Identifiability of Causal Queries

- Do calculus reduces interventional to observational quer
- Sour + inf
- Alter
- Not If no
- Emerging idea: unidentifiable queries are only partially identifiable (bounds can be estimated!)
- Recent works in this field by various groups

Optimisation techniques for IPs to be used for partial identifiability

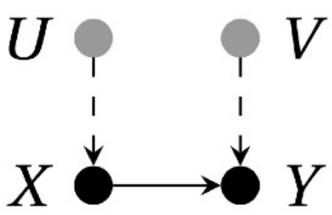




$P(x_3 \,|\, \mathrm{do}(x_2) \in [l, u]$

Back to Headache (Moving to the Third Rung)

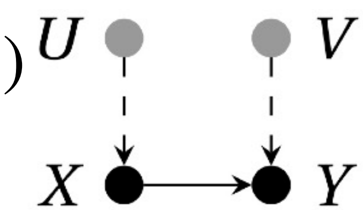
- **What if** I had not taken the aspirin, would have headache stayed?
- An intervention contrasting the current observation ...
- This is a **counterfactual** query $P(Y_{X=0} = 0 | X = 1, Y = 1)$ (called probability of necessity, PN, sub denote do)



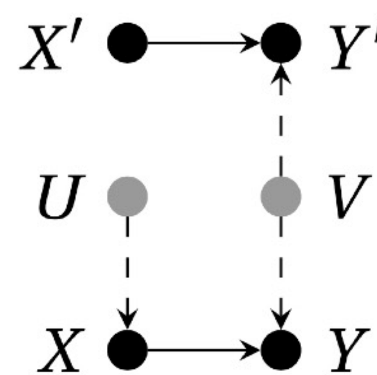
Back to Headache (Moving to the Third Rung)

- **What if** I had not taken the aspirin, would have headache stayed?

- An intervention contrasting the current observation ...
- This is a **counterfactual** query $P(Y_{X=0} = 0 \,|\, X = 1, Y = 1)$ (called probability of necessity, PN, sub denote do)



- We need the complete SCM: $\mathcal{G} + \{f_X\}_{X \in \mathbf{X}} + \{P(U)\}_{U \in \mathbf{U}}$
- With complete SCM, an augmented model called **twin network** with duplicated endogenous variables is used for counterfactual analysis after surgery



- (Non-trivial) **counterfactuals are unidentifiable**!

To Compute Counterfactuals ...

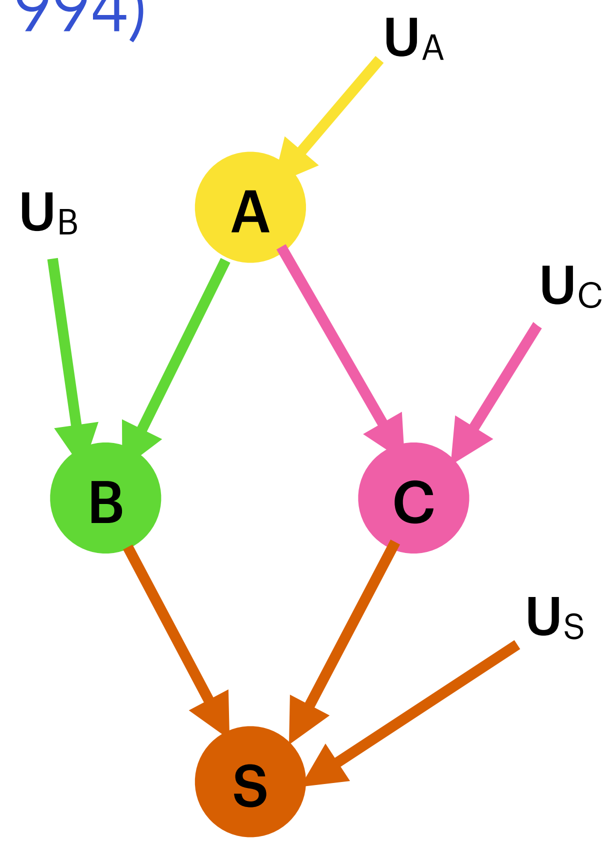
- We need a fully specified SCM, i.e.,
    1. Graph $\mathscr{G}$ over $(\mathbf{X}, \mathbf{U})$
       (often available by domain expert or Markovian assumption)
    2. Endogenous equations $\{f_X\}_{X \in \mathbf{X}}$
       (available or obtained by complete enumeration)
    3. Exogenous marginals $\{P(U)\}_{U \in \mathbf{U}}$ (rarely available)

To Compute Counterfactuals ...

- We need a fully specified SCM, i.e.,
  1. Graph $\mathcal{G}$ over $(\mathbf{X}, \mathbf{U})$
     (often available by domain expert or Markovian assumption)
  2. Endogenous equations $\{f_X\}_{X \in \mathbf{X}}$
     (available or obtained by complete enumeration)
  3. Exogenous marginals $\{P(U)\}_{U \in \mathbf{U}}$ (rarely available)
- Latent $P(\mathbf{U}) = \prod P(U)$ unavailable? We have data $\mathcal{D}$ about $\mathbf{X}$
- Compute counterfactual = Compute $\{P(U)\}_{U \in \mathbf{U}}$ from $\mathcal{D}$
- Not a new problem: LP approach for special cases already in Balke and Pearl (1994), but do-calculus reduced attention to CFs

Causal Analysis at the Party (Balke & Pearl 1994)

Ann sometimes goes to parties
Bob is not a party guy,
but he likes Ann
and he might be there
Carl broke up with Ann,
he tries to avoid Ann,
but he likes parties
Carl and Bob hate each other,
they might have a Scuffle
if both at the party



besides such knowledge assume
we have observations $\mathscr{D}$ corresponding
to a joint mass function $P(A, B, C, S)$
(e.g., in the form of a BN)

Causal Analysis at the Party (Balke & Pearl 1994)

CAUSAL GOSSIP

INTERVENTIONAL

"**A**nn must not be at the party, or **B**ob would be there instead of home"

$$P(B \mid \mathrm{do}(\overline{a})) = ?$$

COUNTERFACTUAL

"If **B**ob were at the party, then Bob and Carl would surely **S**cuffle"

$$P(S_b \mid \overline{b}) = ?$$

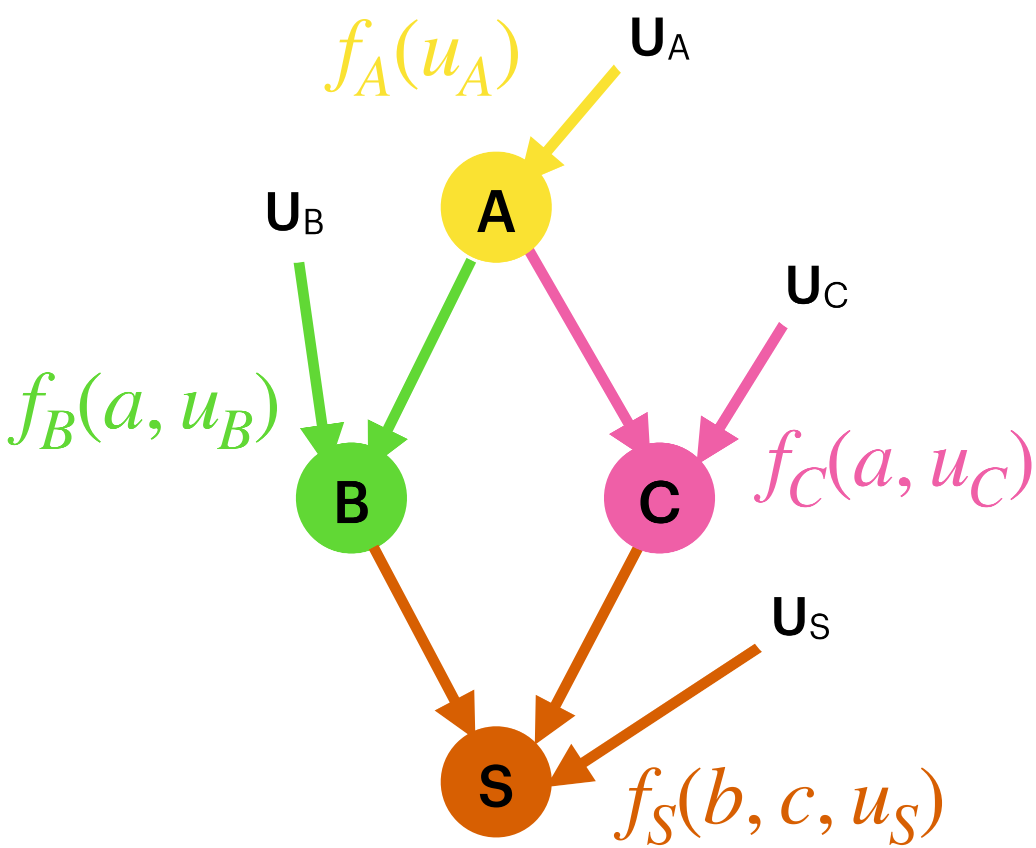
a (fully specified) SCM can answer these questions

(e.g., in the form of a BN)

Let's (Eventually) Use IPs!

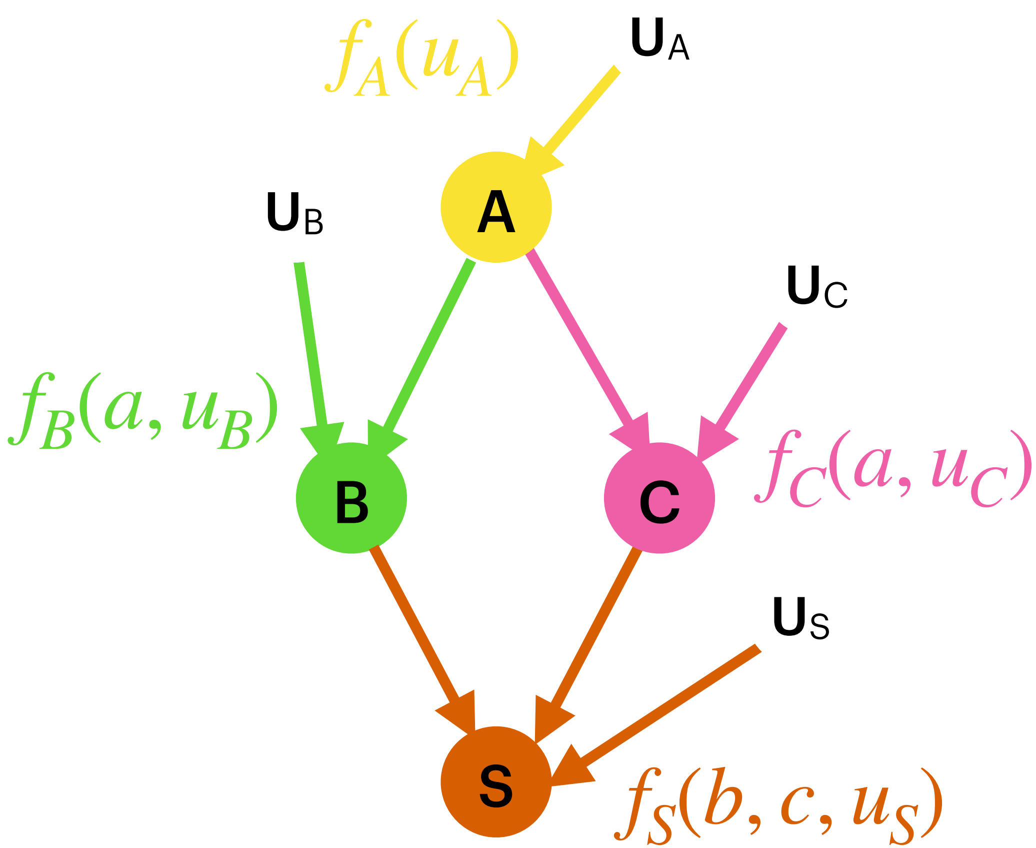
- Find the exogenous marginals?

  $P(U_A)P(U_B)P(U_C)P(U_S)$



Let's (Eventually) Use IPs!

- Find the exogenous marginals?
  $P(U_A)P(U_B)P(U_C)P(U_S)$
- **Endogenous** (= with $\mathscr{D}$) **consistency**



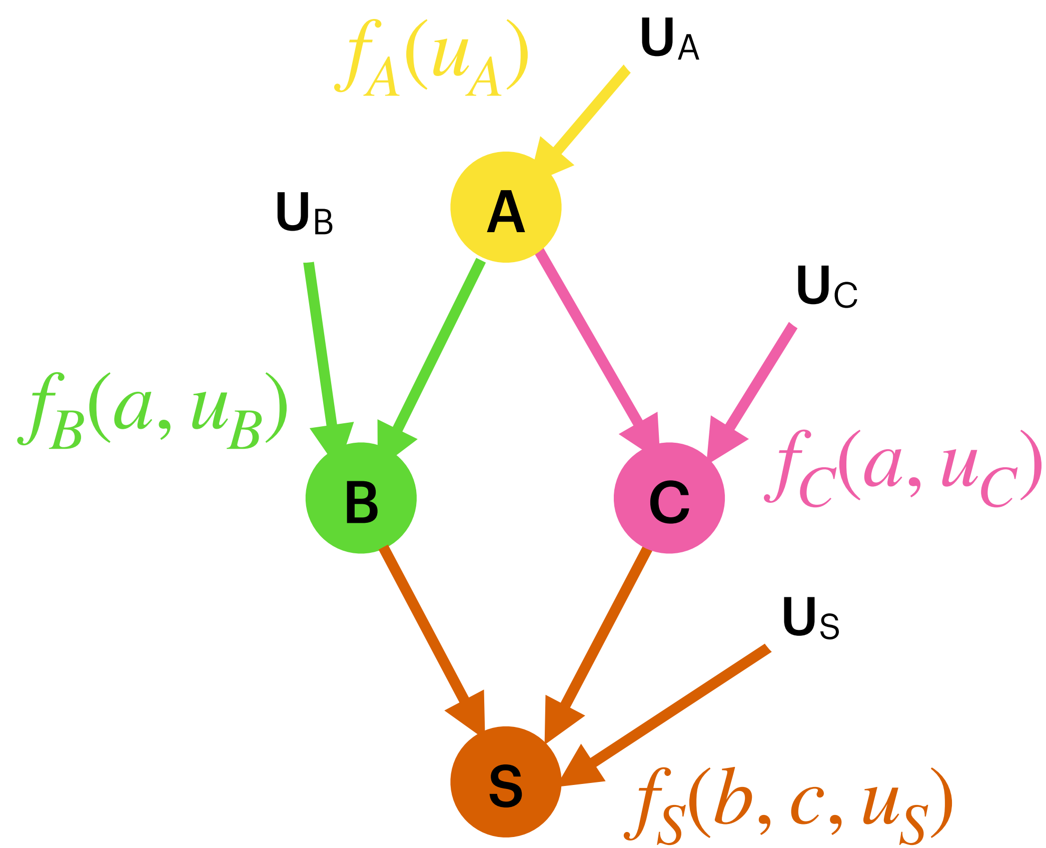
$$\sum_{u_A,u_B,u_C,u_D} \left[p(u_A) \cdot \delta_{a,f_A(u_A)} \cdot p(u_B) \cdot \delta_{b,f_B(a,u_B)} \cdot p(u_C) \cdot \delta_{c,f_C(a,u_c)} \cdot p(u_S) \cdot \delta_{s,f_S(b,c,u_S)} \right] = \tilde{p}(a,b,c,s)$$

Let's (Eventually) Use IPs!

- Find the exogenous marginals?

  $P(U_A)P(U_B)P(U_C)P(U_S)$

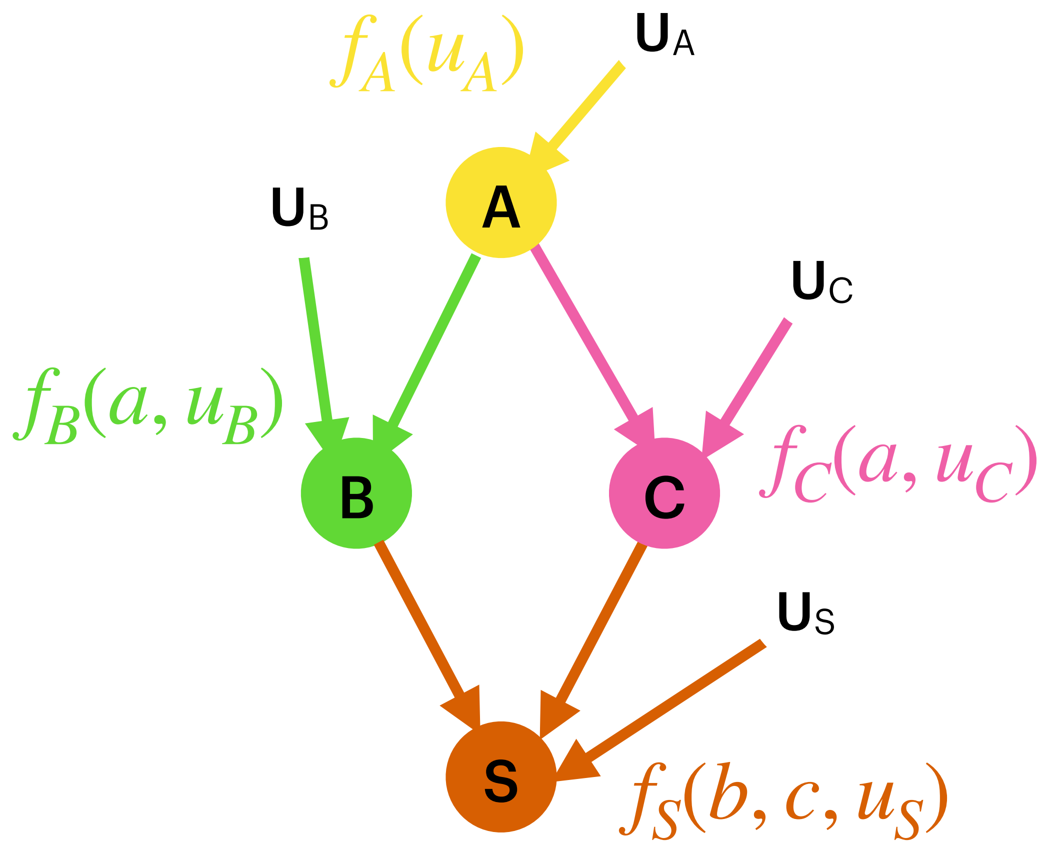
- **Endogenous** (= with $\mathscr{D}$) **consistency**
- This induces global non-linear (so-called Verma) constraints



$$\sum_{u_A,u_B,u_C,u_D} \left[\overset{\text{Unknown}}{p(u_A)} \cdot \delta_{a,f_A(u_A)} \cdot \overset{\text{Unknown}}{p(u_B)} \cdot \delta_{b,f_B(a,u_B)} \cdot \overset{\text{Unknown}}{p(u_C)} \cdot \delta_{c,f_C(a,u_c)} \cdot \overset{\text{Unknown}}{p(u_S)} \cdot \delta_{s,f_S(b,c,u_S)} \right] = \overset{\text{Empirical, known}}{\tilde{p}(a,b,c,s)}$$

Let's (Eventually) Use IPs!

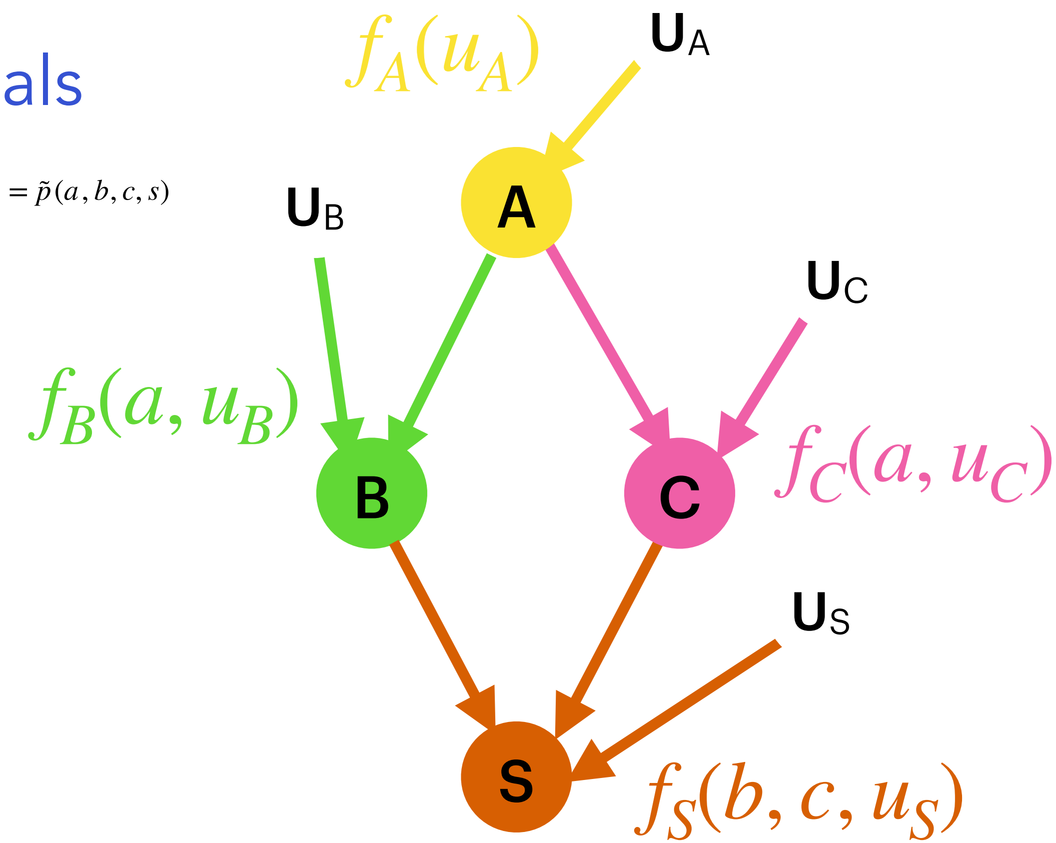
- Find the exogenous marginals?
  $P(U_A)P(U_B)P(U_C)P(U_S)$
- **Endogenous** (= with $\mathscr{D}$) **consistency**
- This induces global non-linear (so-called Verma) constraints
- Constraints became local and linear ones by marginalisation and conditioning (Zaffalon et al., 2020)



$$\sum_{u_A,u_B,u_C,u_D} \Big[\overset{\text{Unknown}}{p(u_A)} \cdot \delta_{a,f_A(u_A)} \cdot \overset{\text{Unknown}}{p(u_B)} \cdot \delta_{b,f_B(a,u_B)} \cdot \overset{\text{Unknown}}{p(u_C)} \cdot \delta_{c,f_C(a,u_c)} \cdot \overset{\text{Unknown}}{p(u_S)} \cdot \delta_{s,f_S(b,c,u_S)} \Big] = \overset{\text{Empirical, known}}{\tilde{p}(a,b,c,s)}$$

Constraining Exogenous Marginals

$$\sum_{u_A,u_B,u_C,u_D}\left[p(u_A)\cdot\delta_{a,f_A(u_A)}\cdot p(u_B)\cdot\delta_{b,f_B(a,u_B)}\cdot p(u_C)\cdot\delta_{c,f_C(a,u_c)}\cdot p(u_S)\cdot\delta_{s,f_S(b,c,u_S)}\right]=\tilde{p}(a,b,c,s)$$



Constraining Exogenous Marginals

$$\sum_{u_A,u_B,u_C,u_D}\left[p(u_A)\cdot\delta_{a,f_A(u_A)}\cdot p(u_B)\cdot\delta_{b,f_B(a,u_B)}\cdot p(u_C)\cdot\delta_{c,f_C(a,u_C)}\cdot p(u_S)\cdot\delta_{s,f_S(b,c,u_S)}\right]=\tilde{p}(a,b,c,s)$$

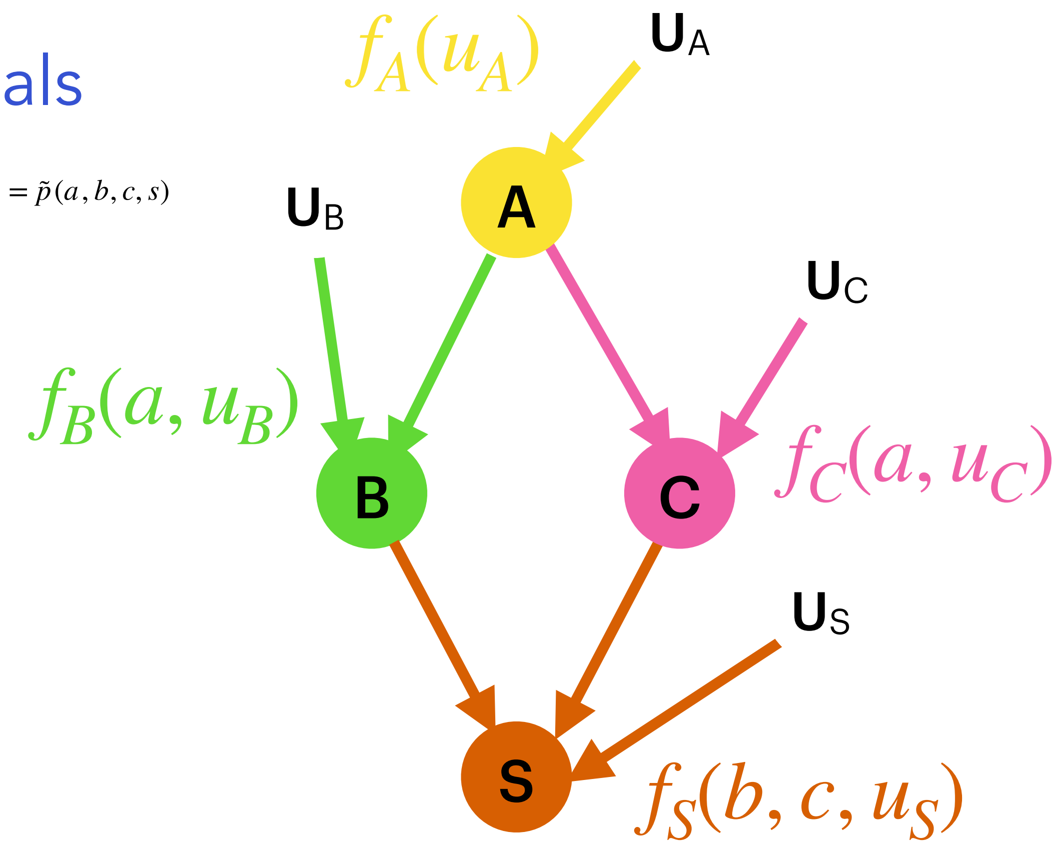
$$\Downarrow$$

$$P(a)=\sum_{u_A}P(a\,|\,u_A)\cdot P(u_A)$$

$$P(b\,|\,a)=\sum_{u_B}P(b\,|\,a,u_B)\cdot P(u_B)$$

$$P(c\,|\,a)=\sum_{u_C}P(c\,|\,a,u_C)\cdot P(u_C)$$

$$P(s\,|\,b,c)=\sum_{u_S}P(s\,|\,b,c,u_S)\cdot P(u_S)$$



Constraining Exogenous Marginals

$$\sum_{u_A,u_B,u_C,u_D} \left[p(u_A)\cdot\delta_{a,f_A(u_A)}\cdot p(u_B)\cdot\delta_{b,f_B(a,u_B)}\cdot p(u_C)\cdot\delta_{c,f_C(a,u_C)}\cdot p(u_S)\cdot\delta_{s,f_S(b,c,u_S)}\right] = \tilde{p}(a,b,c,s)$$

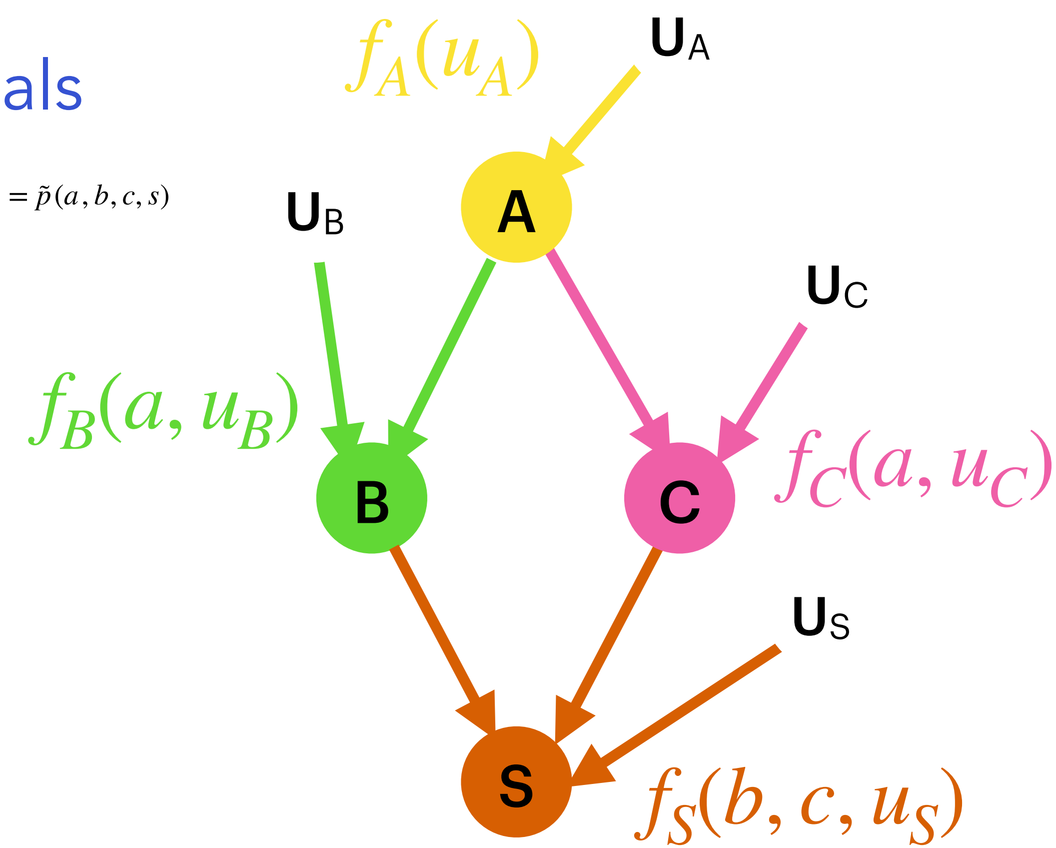
$$\Downarrow$$

$$P(a) = \sum_{u_A} P(a\,|\,u_A)\cdot P(u_A)$$

$$P(b\,|\,a) = \sum_{u_B} P(b\,|\,a,u_B)\cdot P(u_B)$$

$$P(c\,|\,a) = \sum_{u_C} P(c\,|\,a,u_C)\cdot P(u_C)$$

$$P(s\,|\,b,c) = \sum_{u_S} P(s\,|\,b,c,u_S)\cdot P(u_S)$$



- Linear constraints on marginal exogenous probabilities leading to the credal sets specification $K(U_A)$, $K(U_B)$, $K(U_C)$, $K(U_S)$
- Structural equations (= endogenous CPTS) remain unaffected

Constraining Exogenous Marginals

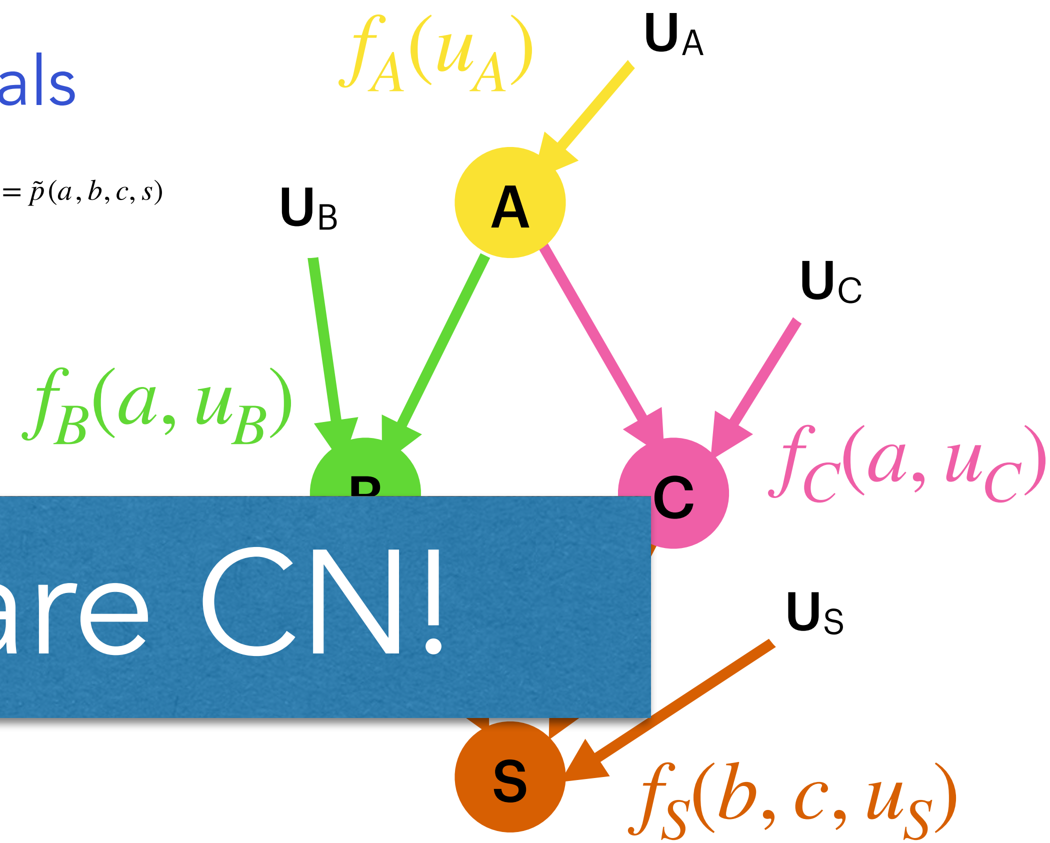
$$\sum_{u_A,u_B,u_C,u_D}\left[p(u_A)\cdot\delta_{a,f_A(u_A)}\cdot p(u_B)\cdot\delta_{b,f_B(a,u_B)}\cdot p(u_C)\cdot\delta_{c,f_C(a,u_C)}\cdot p(u_S)\cdot\delta_{s,f_S(b,c,u_S)}\right]=\tilde{p}(a,b,c,s)$$

$$P(a)=\sum_{u_A}P(a\,|\,u_A)\cdot P(u_A)$$

$$P(b\,|\,a)=\sum_{u_B}\ldots$$

$$P(c\,|\,a)=\sum_{u_C}\ldots$$

$$P(s\,|\,b,c)=\sum_{u_S}P(s\,|\,b,c,u_S)\cdot P(u_S)$$

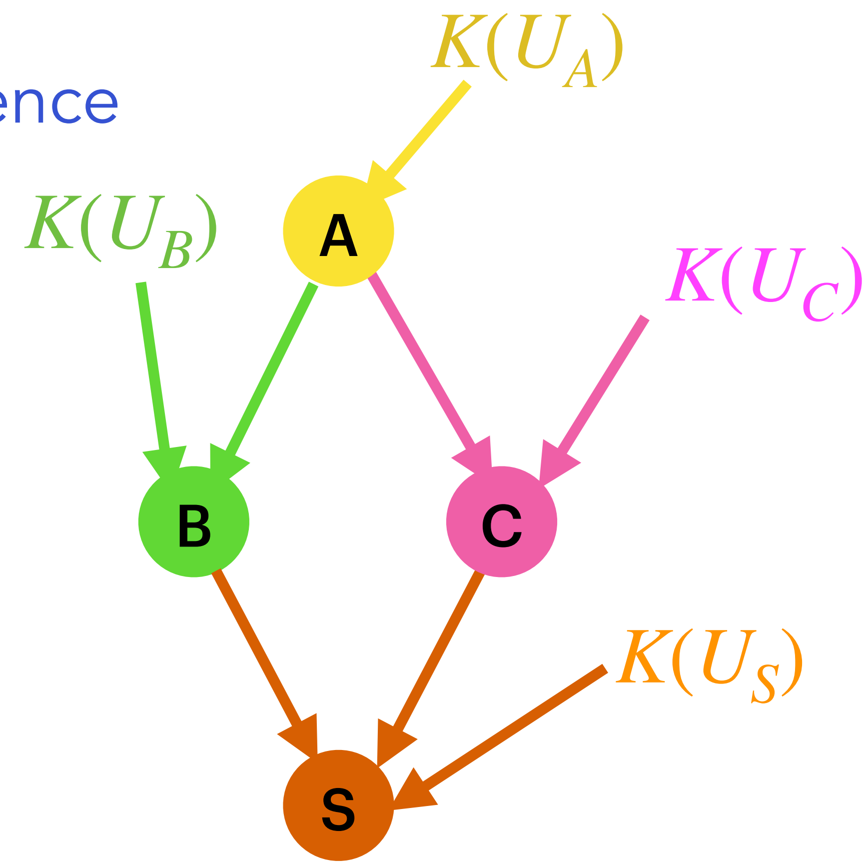


SCMs are CN!

- Linear constraints on marginal exogenous probabilities leading to the credal sets specification $K(U_A)$, $K(U_B)$, $K(U_C)$, $K(U_S)$
- Structural equations (= endogenous CPTS) remain unaffected

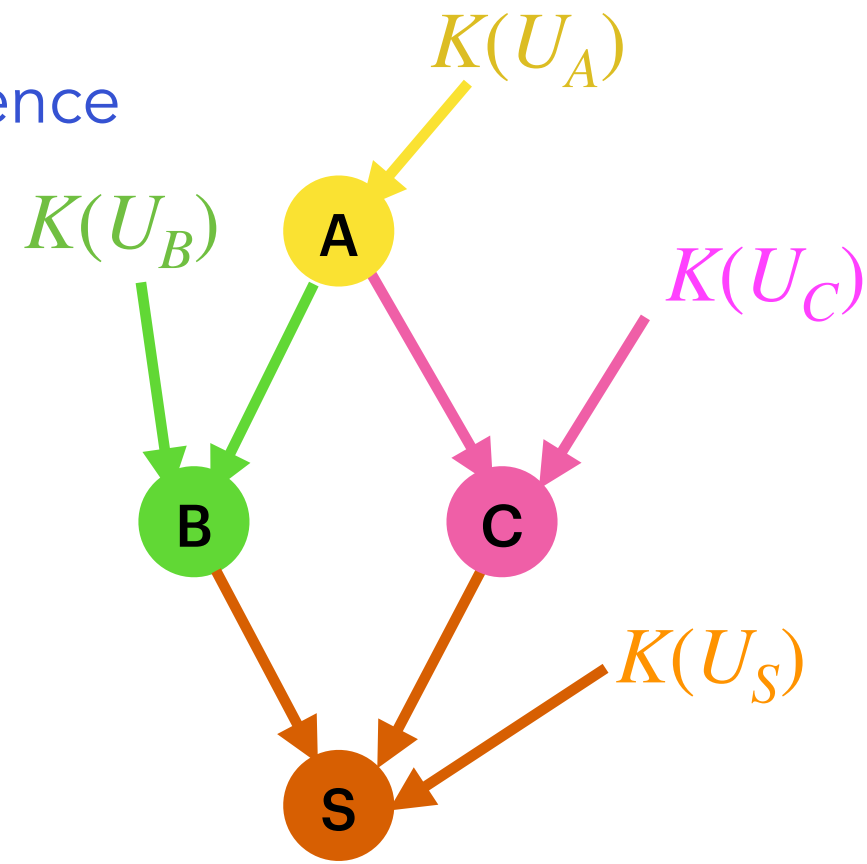
Reducing Causal Queries to CN Inference

- Consistent SCMs as a single CN



Reducing Causal Queries to CN Inference

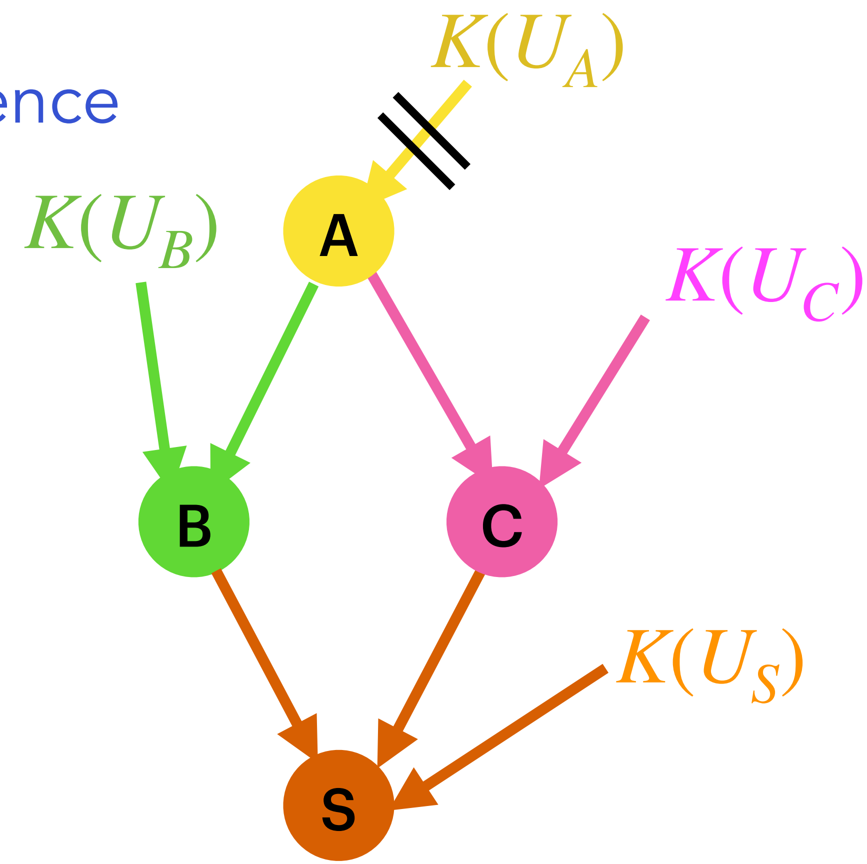
- Consistent SCMs as a single CN
- d-separation holds for CNs, we can do surgery à la Pearl
- CN algs to compute bounds!



Reducing Causal Queries to CN Inference

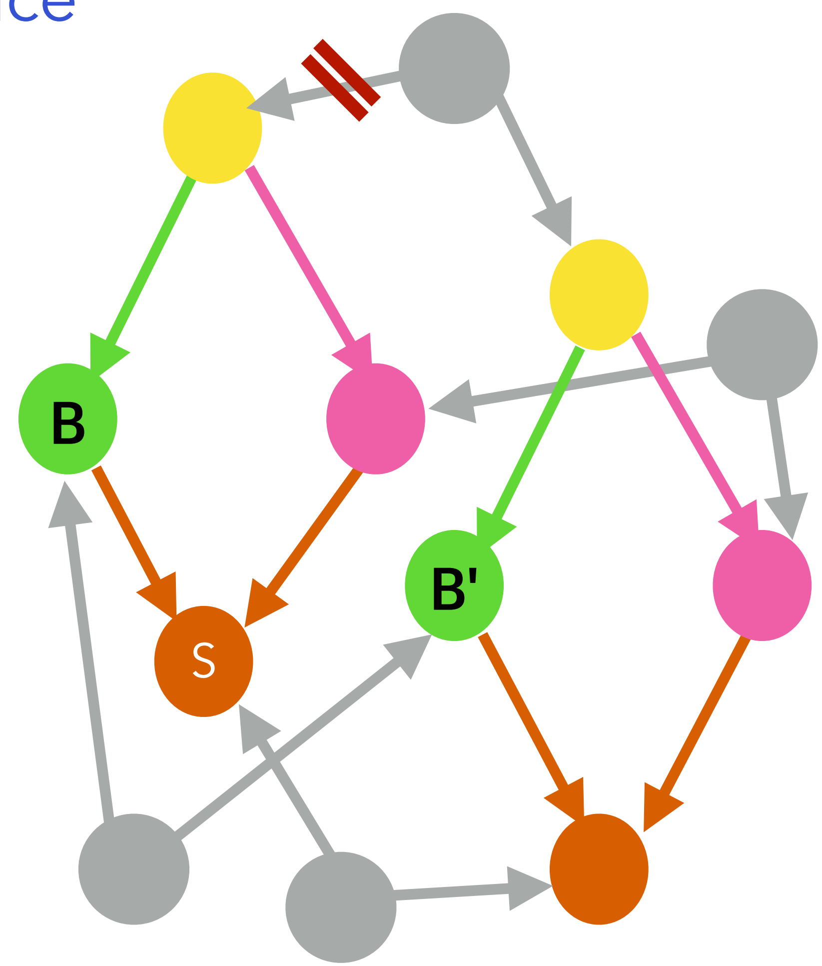
- Consistent SCMs as a single CN
- d-separation holds for CNs, we can do surgery à la Pearl
- CN algs to compute bounds!
- Interventions are straightforward

$$P(B \,|\, \mathrm{do}(\overline{a})) \in [\underline{P}'(B \,|\, \overline{a}), \overline{P}'(B \,|\, \overline{a})]$$



Reducing Causal Queries to CN Inference

- Consistent SCMs as a single CN
- d-separation holds for CNs, we can do surgery à la Pearl
- CN algs to compute bounds!
- Interventions are straightforward $P(B\,|\,\mathrm{do}(\overline{a})) \in [\underline{P}'(B\,|\,\overline{a}), \overline{P}'(B\,|\,\overline{a})]$
- Counterfactuals require twin nets $P(S_b\,|\,\overline{b}) \in [\underline{P}(S\,|\,b,\overline{b}'), \overline{P}(S\,|\,b,\overline{b}')]$
- Identifiable? $\underline{P} = \overline{P}$

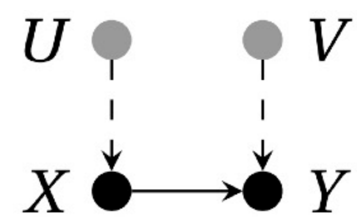


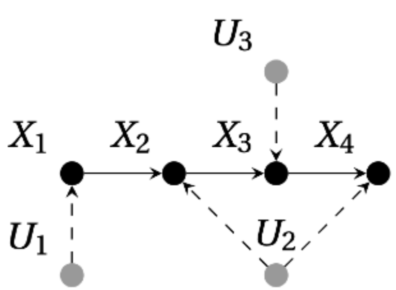
Markovian and Quasi-Markovian SCMs as CNs

Algorithm 1 Given an SCM M and a PMF $\tilde{P}(\boldsymbol{X})$, return CSs $\{K(U)\}_{U \in \boldsymbol{U}}$

for $X \in \boldsymbol{X}$ **do**
    $U \leftarrow \mathrm{Pa}(X) \cap \boldsymbol{U}$ // U as the unique exogenous parent of X
    $\underline{\mathrm{Pa}}(X) \leftarrow \mathrm{Pa}(X) \setminus \{U\}$ // Endogenous parents of X
    if $\underline{\mathrm{Pa}}(X) = \emptyset$ **then**
        $K(U) \leftarrow \{P'(U) : \sum_{u \in f_X^{-1}} P'(u) = \tilde{P}(x), \forall x \in \Omega_X\}$ // Eq. (4)
    else
        $K(U) \leftarrow \{P'(U) : \sum_{u \in f^{-1}_{X|\underline{\mathrm{pa}}(X)}(x)} P'(u) = \tilde{P}(x|\underline{\mathrm{pa}}(X)), \forall x \in \Omega_X, \forall \underline{\mathrm{pa}}(X) \in \Omega_{\underline{\mathrm{Pa}}_X}\}$ // Eq. (6)
    end if
end for

Markovian Models





Quasi-Markovian Models

Algorithm 2 Given an SCM M and a PMF $\tilde{P}(\boldsymbol{X})$, return CSs $\{K(U)\}_{U \in \boldsymbol{U}}$

for $U \in \boldsymbol{U}$ **do**
    $\{X_U^k\}_{k=1}^{n_U} \leftarrow \mathrm{Sort}[X \in \boldsymbol{X} : U \in \mathrm{Pa}(X)]$ // Children of U in topological order
    $\gamma \leftarrow \emptyset$
    for $(x_U^1, \ldots, x_U^{n_U}) \in \times_{k=1}^{n_U} \Omega_{X_U^k}$ **do**
        for $(\underline{\mathrm{pa}}(X_U^1), \ldots, \underline{\mathrm{pa}}(X_U^{n_U})) \in \times_{k=1}^{n_U} \Omega_{\underline{\mathrm{Pa}}(X_U^k)}$ **do**
            $\Omega'_U \leftarrow \bigcap_{k=1}^{n_U} f^{-1}_{X_U^k|\underline{\mathrm{pa}}(X_U^k)}(x_U^k)$
            $\gamma \leftarrow \gamma \cup \left\{\sum_{u \in \Omega'_U} P(u) = \prod_{k=1}^{n_U} \tilde{P}(x_U^k | x_U^1, \ldots, x_U^{k-1}, \underline{\mathrm{pa}}(X_U^1)), \ldots, \underline{\mathrm{pa}}(X_U^k))\right\}$
        end for
    end for
    $K(U) \leftarrow \{P(U) : \gamma\}$ // CS by linear constraints on $P(U)$
end for

GitHub

Software and Experiments

Crema
Credal Models Algorithms

Java library for CNs

Credici
Credal Inference for Causal Inference

Java library for Causal Inference built on the top of CREMA

GitHub

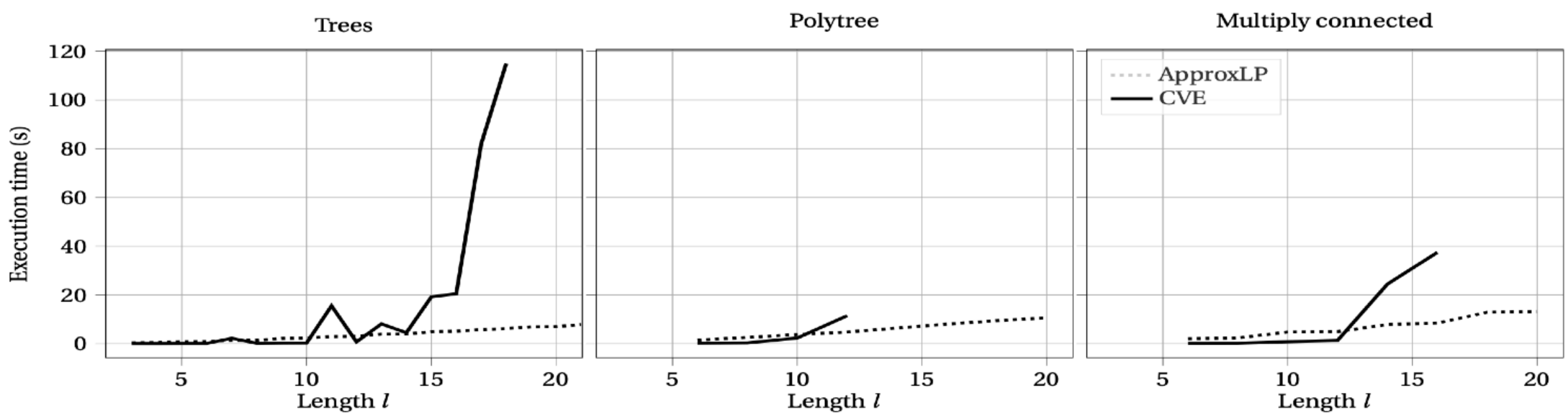
Software and Experiments

Crema
Credal Models Algorithms

Java library for CNs

Credici
Credal Inference for Causal Inference

Java library for Causal Inference built on the top of CREMA



Exact inference by credal variable elimination only for small models

ApproxLP (Antonucci et al., 2014) allows to process larger models

RMSE always <0.7%

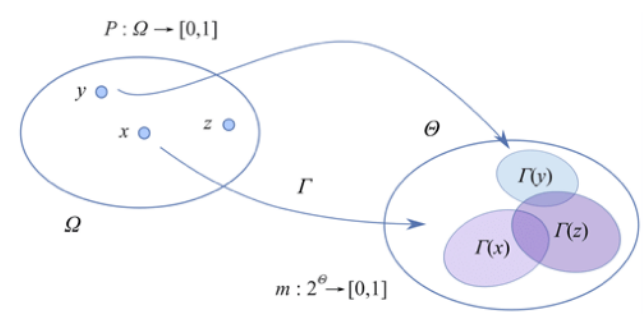
Intermezzo: Belief Functions (as Credal Sets)

- Linear constraints for CN induced by SCM have a peculiar form
- These are CS corresponding to **belief functions** (Dempster '68, Shafer '76)
- Class of generalised probabilistic models
- PMF distributes mass over the singletons, BF over (poss. overlapping) sets
- Dempster's **multi-valued mapping**, in SCMs $\mathbf{U} = f^{-1}(\mathbf{X})$, $\mathrm{BF}(\mathbf{U}) := f^{-1}[P(\mathbf{X})]$
- Dedicated conditioning/combination rules

$$\sum_{u:\text{condition}} P(u) = \text{const}$$

$$\sum_{u\in\Omega'_U} P(u) = \prod_{k=1}^{n_U} \tilde{P}(x_U^k|x_U^1,\dots,x_U^{k-1},\underline{\mathrm{pa}}(X_U^1)),\dots,\underline{\mathrm{pa}}(X_U^k))$$

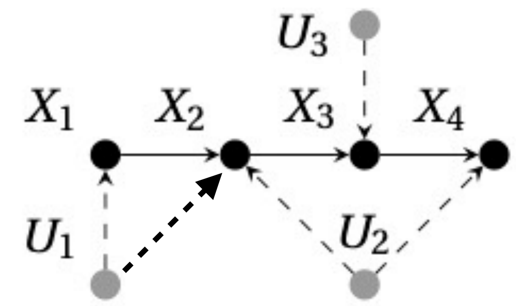
$$\{P'(U) : \sum_{u\in f^{-1}_{X|\underline{\mathrm{pa}}(X)}(x)} P'(u) = \tilde{P}(x|\underline{\mathrm{pa}}(X))\,,\forall x\in\Omega_X,\forall\underline{\mathrm{pa}}(X)\in\Omega_{\underline{\mathrm{Pa}}_X}\}$$



Credits: Fabio Cuzzolin

Back to SCM2CN: Non Quasi-Markovian Case

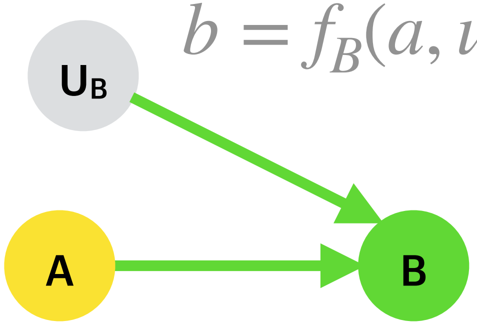
- Non Quasi-Markovian? Non-Linear constraint
- E.g., $\sum P(u_1) \cdot P(u_2) = \dots$
- Merge exogenous variables $U := (U_1, U_2)$
- Independence constraints can be disregarded (but higher exogenous dimensionality)
- Again CN approximate inference to solve causal queries
- State space dimensionality affects complexity
- We might have very large latent spaces ...



Conservative Specification of Structural Equations

- Finding the equations given $\mathcal{G}$ only
- $P(B \mid A)$ should be a deterministic CPT

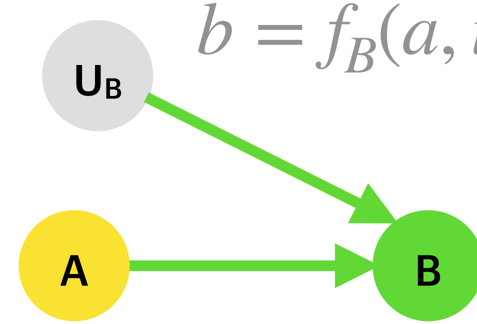
$$b = f_B(a, u_B)$$



Conservative Specification of Structural Equations

- Finding the equations given $\mathscr{G}$ only
- $P(B\,|\,A)$ should be a deterministic CPT

$$b = f_B(a, u_B)$$



$$P(B\,|\,A)$$

	A=0	A=1
B=0	1	1
B=1	0	0

$B = 0$

	A=0	A=1
B=0	1	0
B=1	0	1

$B = A$

	A=0	A=1
B=0	0	1
B=1	1	0

$B = \neg A$

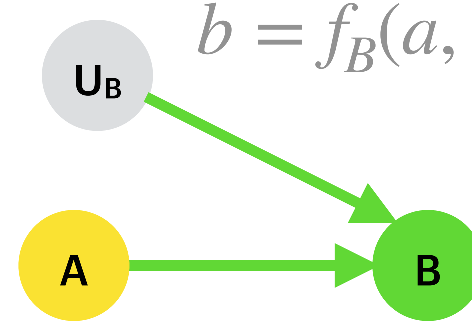
	A=0	A=1
B=0	0	0
B=1	1	1

$B = 1$

Conservative Specification of Structural Equations

- Finding the equations given $\mathscr{G}$ only
- $P(B\,|\,A)$ should be a deterministic CPT
- U_B indexing all these deterministic CPTs

$$b = f_B(a, u_B)$$



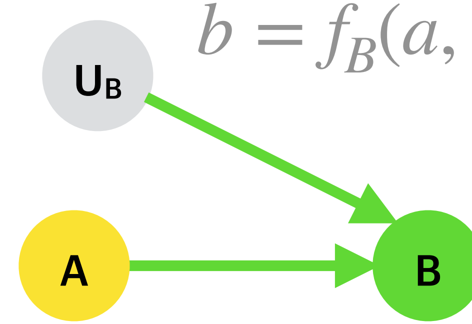
$$P(B\,|\,A, U)$$

	A=0	A=1	A=0	A=1	A=0	A=1	A=0	A=1
B=0	1	1	1	0	0	1	0	0
B=1	0	0	0	1	1	0	1	1
	U=0		U=1		U=2		U=3	
	$B=0$		$B=A$		$B=\neg A$		$B=1$	

Conservative Specification of Structural Equations

- Finding the equations given $\mathscr{G}$ only
- $P(B \mid A)$ should be a deterministic CPT
- U_B indexing all these deterministic CPTs
- Knowledge might discard some states (ex., Bob goes to the party if Ann does)

$$b = f_B(a, u_B)$$

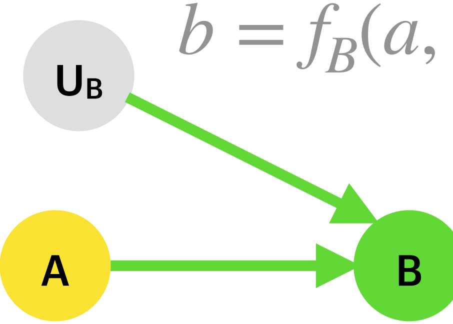


$$P(B \mid A, U)$$

	A=0	A=1	A=0	A=1	A=0	A=1	A=0	A=1
B=0			1	0			0	0
B=1			0	1			1	1
	U=0		U=1		U=2		U=3	
	$B=0$		$B=A$		$B=\neg A$		$B=1$	

Conservative Specification of Structural Equations

$b = f_B(a, u_B)$



- Finding the equations given $\mathcal{G}$ only
- $P(B\,|\,A)$ should be a deterministic CPT
- U_B indexing all these deterministic CPTs
- Knowledge might discard some states (ex., Bob goes to the party if Ann does)
- With Boolean parent & child) $|\,U\,| = 4$ in general (exp size) :

$$|\,U\,| = |\,X\,|^{\prod_{Y \in \mathrm{Pa}_Y} |\mathrm{Y}|}$$

$$P(B\,|\,A, U)$$

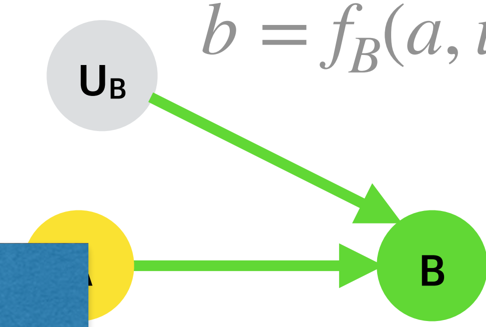
	A=0	A=1	A=0	A=1	A=0	A=1	A=0	A=1
B=0			1	0			0	0
B=1			0	1			1	1
	U=0		U=1		U=2		U=3	
	$B = 0$		$B = A$		$B = \neg A$		$B = 1$	

Conservative Specification of Structural Equations

- Finding the equations given $\mathscr{G}$ only
- $P(B|A)$ should be a deterministic CPT
- U_B indexing
- Knowledge r
  (ex., Bob goe
- With Boolea
  in general (exp size) :

$$|U| = |X|^{\prod_{Y \in \mathrm{Pa}_Y} |Y|}$$

$$b = f_B(a, u_B)$$



$B|A,U)$

	A=0	A=1	A=0	A=1	A=0	A=1	A=0	A=1
B=0			1	0			0	0
B=1			0	1			1	1
	U=0		U=1		U=2		U=3	
	$B=0$		$B=A$		$B=\neg A$		$B=1$	

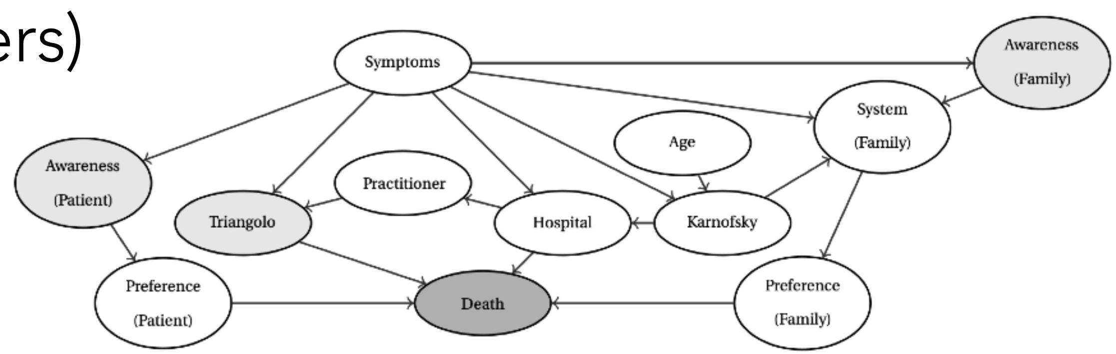
CFs based on $\mathscr{G}$ and $\mathscr{D}$ only

An Application: Counterfactual Analysis in Palliative Cares

- Study of terminally ill cancer patients' preferences wrt their place of death (home or hospital)
- $\mathscr{G}$ obtained by expert knowledge and data
- Exogenous variables?
- Markovian assumption (= no confounders)

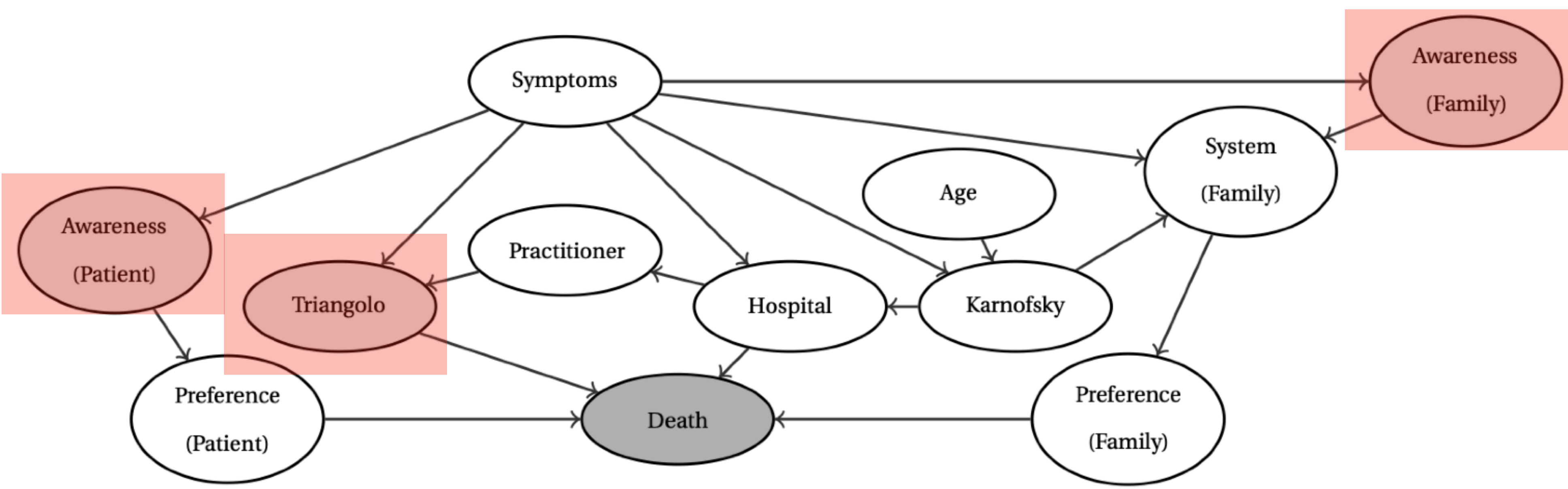
Impact on place of death in cancer patients: a causal exploration in southern Switzerland

Heidi Kern [1], Giorgio Corani [2], David Huber [2], Nicola Vermes [2], Marco Zaffalon [2], Marco Varini [3], Claudia Wenzel [4], André Fringer [5]



An Application: Counterfactual Analysis in Palliative Cares

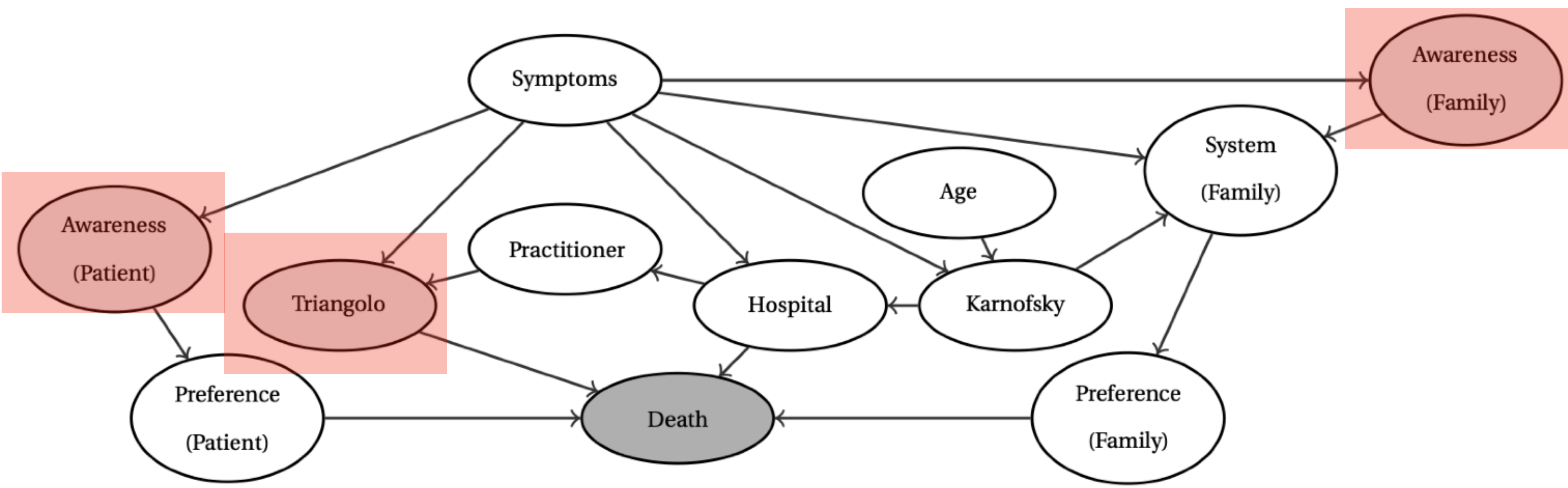
- Most patients prefer to die at home
- But a majority actually die in institutional settings
- Interventions by health care professionals can facilitate dying at home?



An Application: Counterfactual Analysis in Palliative Cares

- Importance of a variable?
- Probability of necessity and sufficiency

$$PNS := P(Y_{X=1} = 1, Y_{X=0} = 0)$$



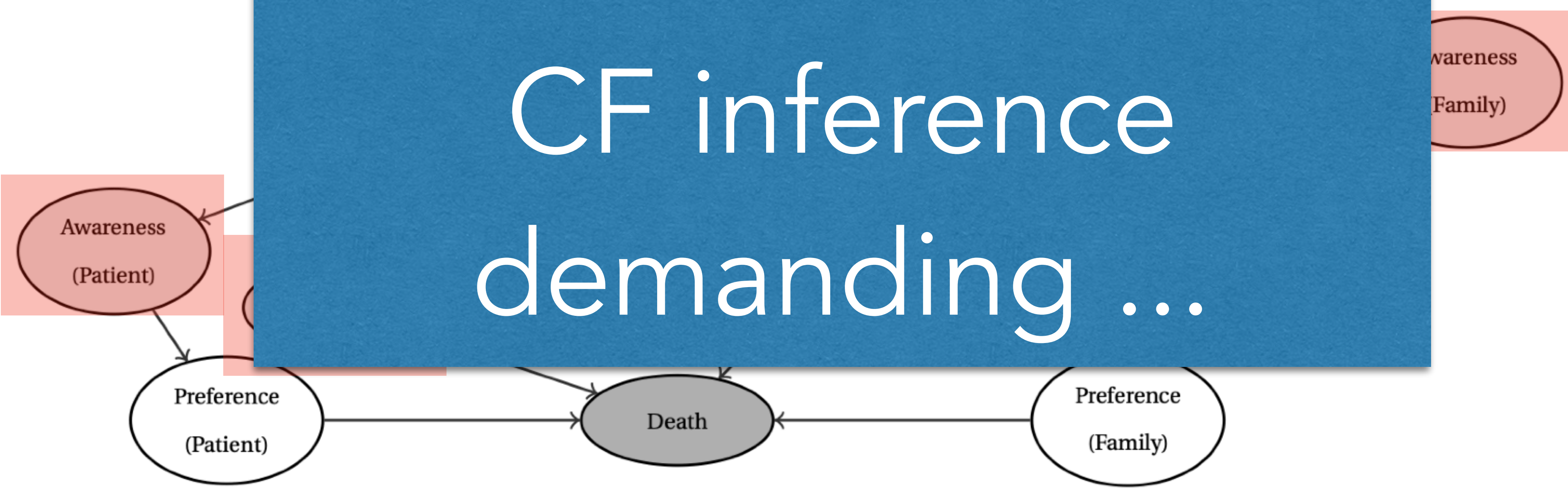
An Application: Counterfactual Analysis in Palliative Cares

- Importance of a variable?
- Probab

P

Small CN but large cardinalities

CF inference demanding …



Causal Expectation Maximisation (Zaffalon et al., 2021)

- Exogenous variables are always missing (MAR, asystematic, way)
- Expectation Maximisation (Dempster 1977)
  - Random initialisation of P(U)
  - E-step: Missing data completion by expected (fractional) counts
  - M-step: "completed" data to retrain P(U)
  - Iterate until convergence
- EM goes to a (local/global) max of $\log P(\mathscr{D})$

U1	U2	X1	X2	n
*	*	0	0	...
*	*	0	1	...
*	*	1	0	...
*	*	1	1	...

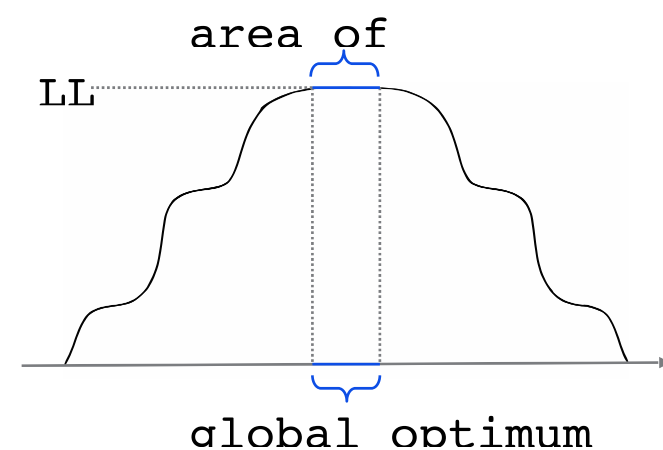
Casual EM: Likelihood Unimodality

- Causal EM reduce should converge to global maxima only the corresponding $P(U)$ belongs to credal set $K(U)$
- Sampling initialisations = sampling of $K(U)$
- For each sample we obtain an inner point

Theorem 1. *Let $\mathcal{K}$ denote the set of quantifications for $\{P(U)\}_{U\in\boldsymbol{U}}$ consistent with the following constraint to be satisfied for each $c \in \mathscr{C}$ and each $\boldsymbol{y}^{(c)}$:*

$$\sum_{\substack{\boldsymbol{u}^{(c)}: f_X(\mathrm{pa}_X)=x \\ \forall X\in\boldsymbol{X}^{(c)}}} \prod_{U\in\boldsymbol{U}^c} P(u) = \prod_{X\in\boldsymbol{X}^{(c)}} \hat{P}(x|\boldsymbol{y}_X^{(c)}), \tag{8}$$

where the values of u, x and $y_X^{(c)}$ are those consistent with $\boldsymbol{u}^{(c)}$ and $\boldsymbol{y}^{(c)}$. If $\mathcal{K} \neq \emptyset$, the log-likelihood in Eq. (7) achieves its global maximum if and only if $\{P(U)\}_{U\in\boldsymbol{U}} \in \mathcal{K}$. If $\mathcal{K} = \emptyset$, the marginal log-likelihood in Eq. (7) can only take values strictly lower than the global maximum.



Casual EM: Guarantees?

- We first reduced causal queries to CN inference
- Causal EM reduces CN inference to (iterated) BN inference
- Identifiable queries? Each sample gives the same values (a numerical alternative to do-calculus)
- Unidentifiable? Each sample as an inner point
- Credible intervals can be derived

Theorem 5. *Let $[a^*, b^*]$ denote the exact probability bounds of a causal query. Say that $\rho := \{r_i\}_{i=1}^n$ are the outputs of n EMCC iterations, while $[a, b]$ is the interval induced by ρ, i.e., $a := \min_{i=1}^n r_i$ and $b := \max_{i=1}^n r_i$. By construction $a^* \leq a \leq b \leq b^*$. The following inequality holds:*

$$P\left(a - \varepsilon L \leq a^* \leq b^* \leq b + \varepsilon L \,\middle|\, \rho\right) = \frac{1 + (1+2\varepsilon)^{2-n} - 2(1+\varepsilon)^{2-n}}{(1 - L^{n-2}) - (n-2)(1-L)L^{n-2}}, \quad (13)$$

where $L := (b - a)$ and $\varepsilon := \delta/(2L)$ is the relative error at each extreme of the interval obtained as a function of the absolute allowed error $\delta \in (0, L)$.

Casual EM: Guarantees?

- We first reduced causal queries to CN inference
- Cau ... e
- Ider
  (a n
- Uni
- Cre

In practice?

20 EM runs to get close to the actual bounds with 95% credibility

For identifiable queries 9 runs to be sure with 99% credibility

Theore

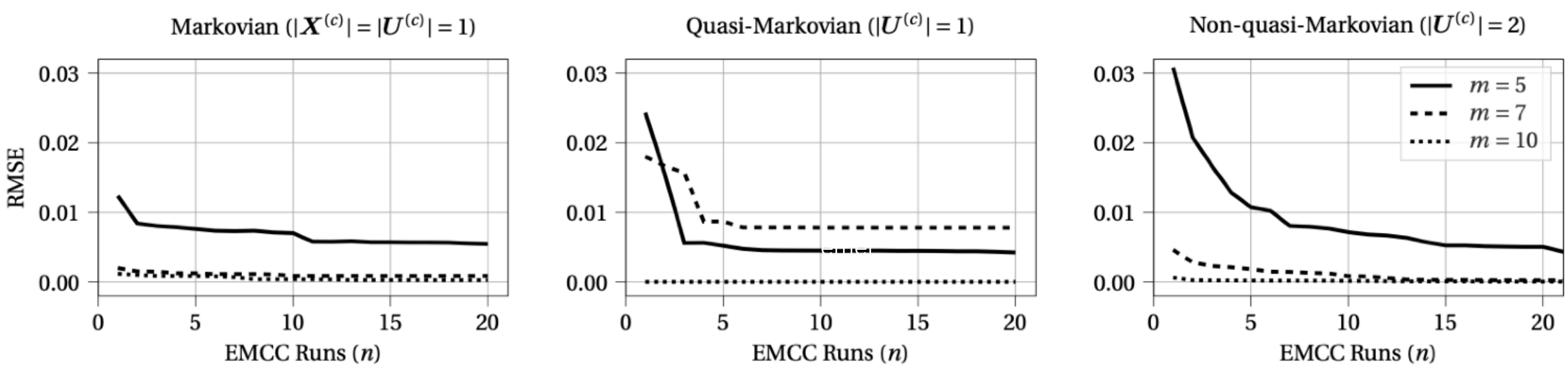
are the

and $b := \max_{i=1} r_i$. *By construction* ... *The following inequality holds:*

$$P\left(a-\varepsilon L \leq a^* \leq b^* \leq b+\varepsilon L \,\middle|\, \rho\right) = \frac{1+(1+2\varepsilon)^{2-n}-2(1+\varepsilon)^{2-n}}{(1-L^{n-2})-(n-2)(1-L)L^{n-2}}, \quad (13)$$

where $L := (b-a)$ *and* $\varepsilon := \delta/(2L)$ *is the relative error at each extreme of the interval obtained as a function of the absolute allowed error* $\delta \in (0, L)$.

Causal EM: Experiments



PNS for artificial SMCs: quick convergence
(= much faster than direct CN approach)

Credici
Credal Inference for Causal Inference

Causal Analysis from **Biased** Data

- Selective data acquisition
  (untreated M and treated F missing)

Treatment X	Recovery Y	Gender Z	counts
0	0	0	2
1	0	0	41
0	1	0	114
1	1	0	313
0	0	1	107
1	0	1	109
0	1	1	13
1	1	1	1

[Müeller et al., 2022]

Causal Analysis from **Biased** Data

- Selective data acquisition
  (untreated M and treated F missing)
- A (Boolean) **selector** variable $S \equiv (X \neq Z)$

Treat, X	Recover y	Gender Z	Selector S	counts
*	*	*	0	2
1	0	0	1	41
*	*	*	0	114
1	1	0	1	313
0	0	1	1	107
*	*	*	0	109
0	1	1	1	13
*	*	*	0	1

[Müeller et al., 2022]

Causal Analysis from **Biased** Data

- Selective data acquisition
  (untreated M and treated F missing)
- A (Boolean) **selector** variable $S \equiv (X \neq Z)$
- Assume we know $n(S=0) \propto P(S=0)$

Treat, X	Recovery Y	Gender Z	Selector S	counts
1	0	0	1	41
1	1	0	1	313
0	0	1	1	107
0	1	1	1	13
*	*	*	0	226

[Müeller et al., 2022]

Causal Analysis from **Biased** Data

- Selective data acquisition
  (untreated M and treated F missing)
- A (Boolean) **selector** variable $S \equiv (X \neq Z)$
- Assume we know $n(S=0) \propto P(S=0)$
- Interventional queries with bias?
- Do calculus for selection bias
  Barenboim & Tian (AAAI, 2015)

Treat, X	Recovery Y	Gender Z	Selector S	counts
1	0	0	1	41
1	1	0	1	313
0	0	1	1	107
0	1	1	1	13
*	*	*	0	226

[Müeller et al., 2022]

Recovering Causal Effects from Selection Bias

Elias Bareinboim*
Computer Science Department
University of California, Los Angeles
Los Angeles, CA. 90095
eb@cs.ucla.edu

Jin Tian*
Department of Computer Science
Iowa State University
Ames, IA. 50011
jtian@iastate.edu

Causal Analysis from **Biased** Data

- Selective data acquisition
  (untreated M and treated F missing)
- A (Boolean) **selector** variable $S \equiv (X \neq Z)$
- Assume we know $n(S=0) \propto P(S=0)$
- Interventional queries with bias?
- Do calculus for selection bias
  Barenboim & Tian (AAAI, 2015)
- Unidentifiable queries?
- Our EM(CC) can be used for that!

Treat, X	Recovery Y	Gender Z	Selector S	counts
1	0	0	1	41
1	1	0	1	313
0	0	1	1	107
0	1	1	1	13
*	*	*	0	226

[Müeller et al., 2022]

Recovering Causal Effects from Selection Bias

ience

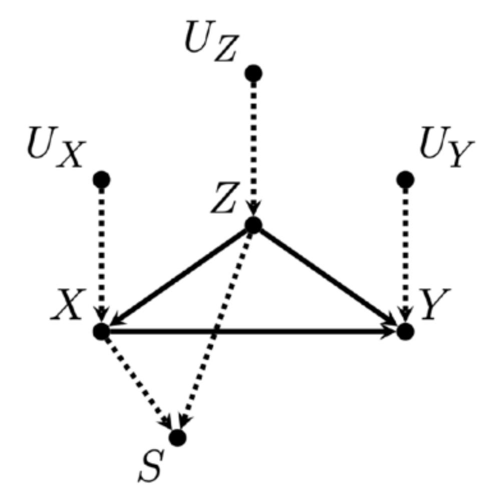
Bounding Counterfactuals under Selection Bias

Marco Zaffalon ZAFFALON@IDSIA.CH
IDSIA, Lugano (Switzerland)
Alessandro Antonucci ALESSANDRO@IDSIA.CH
IDSIA, Lugano (Switzerland)
Rafael Cabañas RCABANAS@UAL.ES
Department of Mathematics, University of Almería, Almería (Spain)
David Huber DAVID.HUBER@IDSIA.CH
IDSIA, Lugano (Switzerland)
Dario Azzimonti DARIO.AZZIMONTI@IDSIA.CH
IDSIA, Lugano (Switzerland)

Back to the Biased Data …

- S determined by an equation, a SCM!

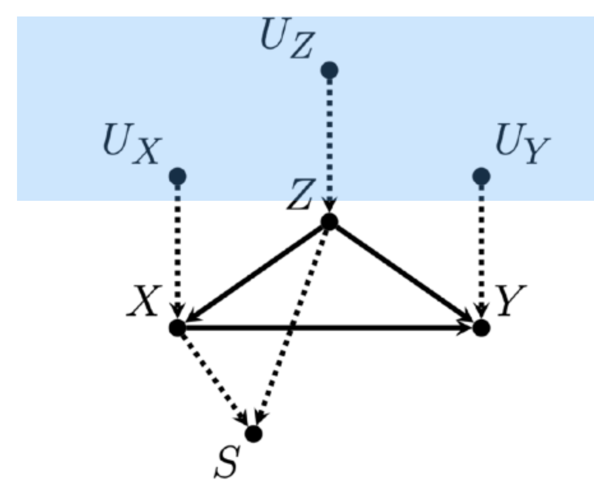
UX	UY	UZ	X	Y	Z	S	n
*	*	*	1	0	0	1	41
*	*	*	1	1	0	1	313
*	*	*	0	0	1	1	107
*	*	*	0	1	1	1	13
*	*	*	*	*	*	0	226



UX	UY	UZ	X	Y	Z	S	n
*	*	*	1	0	0	1	41
*	*	*	1	1	0	1	313
*	*	*	0	0	1	1	107
*	*	*	0	1	1	1	13
*	*	*	*	*	*	0	226

Back to the Biased Data …

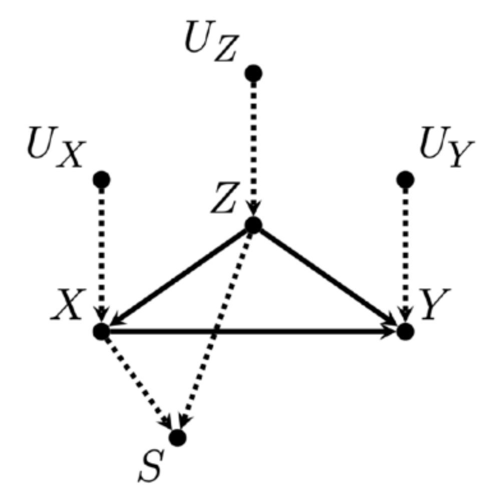
- S determined by an equation, a SCM!
- CN approach? No, $S = 1$ induces relations between $P(U)$'s in the CN



Back to the Biased Data ...

- S determined by an equation, a SCM!
- CN approach? No, $S = 1$ induces relations between $P(U)$'s in the CN
- EM? Maybe, but "non-rectangular" missingness, might kill unimodality ...
- Convergence to max preserved? (hence inner points of $[\underline{P}, \overline{P}]$)

UX	UY	UZ	X	Y	Z	S	n
*	*	*	1	0	0	1	41
*	*	*	1	1	0	1	313
*	*	*	0	0	1	1	107
*	*	*	0	1	1	1	13
*	*	*	*	*	*	0	226

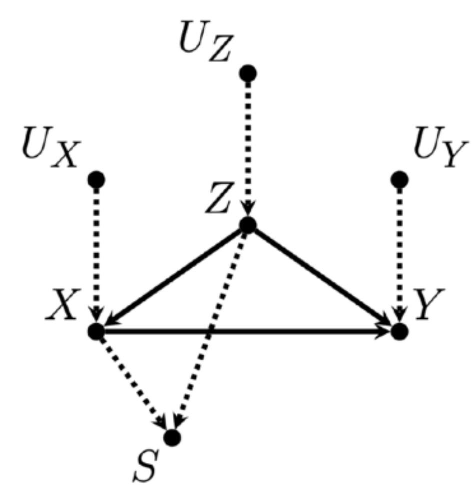


Back to the Biased Data ...

- S determined by an equation, a SCM!
- CN approach? No, $S=1$ induces relations between $P(U)$'s in the CN
- EM? Maybe, but "non-rectangular" missingness, might kill unimodality ...
- Convergence to max preserved? (hence inner points of $[\underline{P},\overline{P}]$) Yes!

UX	UY	UZ	X	Y	Z	S	n
*	*	*	1	0	0	1	41
*	*	*	1	1	0	1	313
*	*	*	0	0	1	1	107
*	*	*	0	1	1	1	13
*	*	*	*	*	*	0	226

Theorem 4 *As a function of $\{P(U)\}_{U\in\boldsymbol{U}}$, the log-likelihood in Eq. (7) has no local maxima and a global maximum equal to the value LL^* in Eq. (6). Such a maximum is achieved if and only if the M-compatibility constraints in Eqs. (8) and (9) are satisfied.*



Extensions: Hybrid Data

Learning to Bound Counterfactual Inference from Observational, Biased and Randomised Data

Marco Zaffalon[a], Alessandro Antonucci[a,*], Rafael Cabañas[b], David Huber[a]

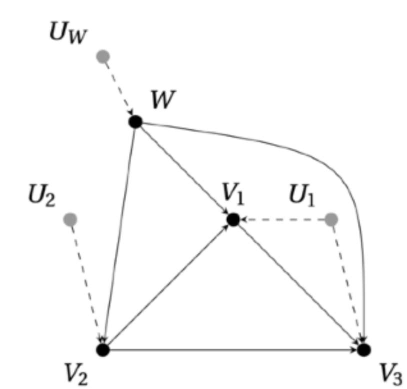
[a] IDSIA, Lugano (Switzerland)
[b] Department of Mathematics, University of Almería, Almería (Spain)

Study	Treatment	Gender	Survival	Counts
interventional	do(drug)	female	survived	489
	do(drug)	female	dead	511
	do(drug)	male	survived	490
	do(drug)	male	dead	510
	do(no drug)	female	survived	210
	do(no drug)	female	dead	790
	do(no drug)	male	survived	210
	do(no drug)	male	dead	790
observational	drug	female	survived	378
	drug	female	dead	1022
	drug	male	survived	980
	drug	male	dead	420
	no drug	female	survived	420
	no drug	female	dead	180
	no drug	male	survived	420
	no drug	male	dead	180

Table 1: Data from interventional and observational studies on the potential effects of a drug on patients affected by a deadly disease.

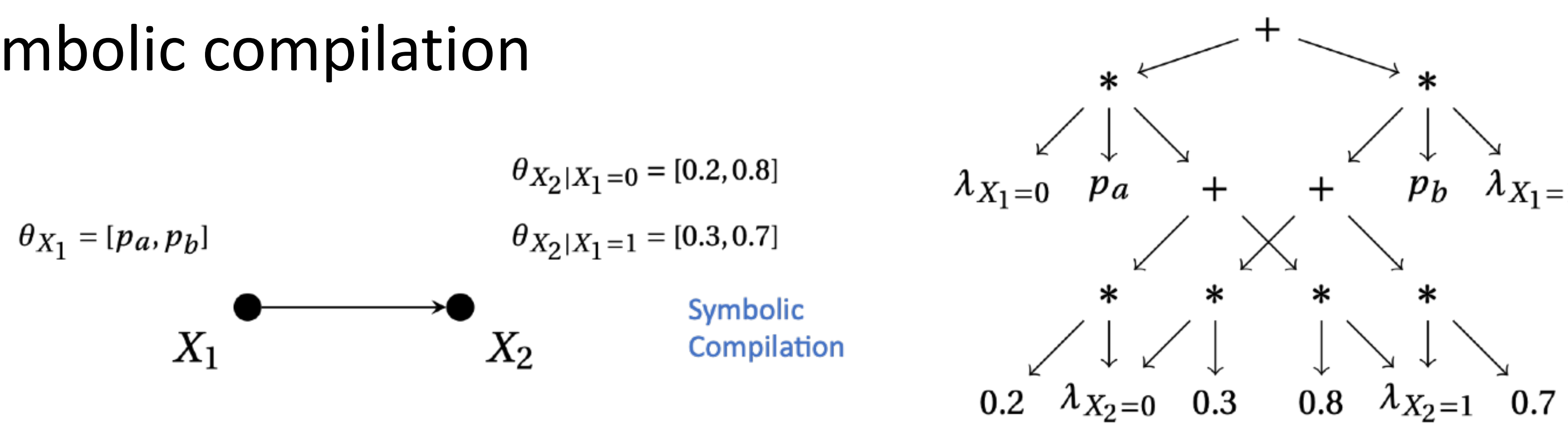
Treatment	Gender	Survival	W	Counts
drug	female	survived	drug	489
drug	female	dead	drug	511
drug	male	survived	drug	490
drug	male	dead	drug	510
no drug	female	survived	no drug	210
no drug	female	dead	no drug	790
no drug	male	survived	no drug	210
no drug	male	dead	no drug	790
drug	female	survived	$w_\emptyset$	378
drug	female	dead	$w_\emptyset$	1022
drug	male	survived	$w_\emptyset$	980
drug	male	dead	$w_\emptyset$	420
no drug	female	survived	$w_\emptyset$	420
no drug	female	dead	$w_\emptyset$	180
no drug	male	survived	$w_\emptyset$	420
no drug	male	dead	$w_\emptyset$	180

Table 2: A merged version of the two datasets in Table 1 with the index variable W.



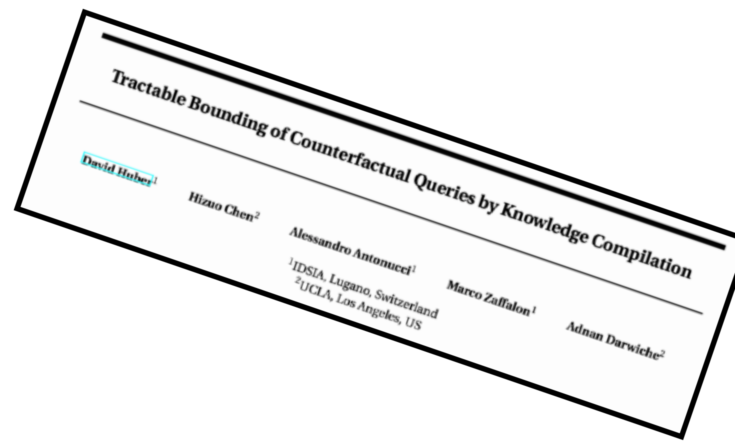
Symbolic Knowledge Compilation (TPM 2023)

- Joint work with Adnan Darwiche and Hizuo Chen
- Our EM requires many (BN) queries
- Equations remain constant
- Compile BN once, use many times
- Symbolic compilation



Current Work: Symbolic Knowledge Compilation (TPM 2023)

- Joint work with Adnan Darwiche and Hizuo Chen
- Our EM requires many (BN) queries
- Equations remain constant
- Compile BN once, use many times
- Symbolic compilation



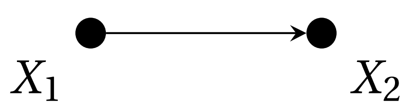
Tractable Bounding of Counterfactual Queries by Knowledge Compilation

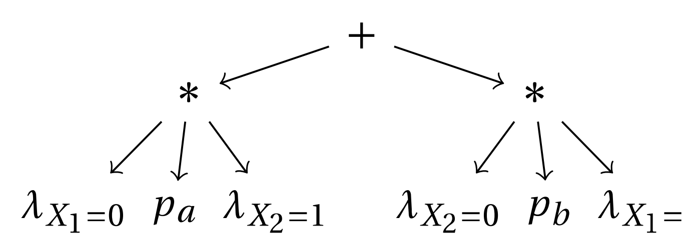
David Huber[1] Hizuo Chen[2] Alessandro Antonucci[1] Marco Zaffalon[1] Adnan Darwiche[2]

[1]IDSIA, Lugano, Switzerland
[2]UCLA, Los Angeles, US

$\theta_{X_2|X_1=0} = [0.0, 1.0]$

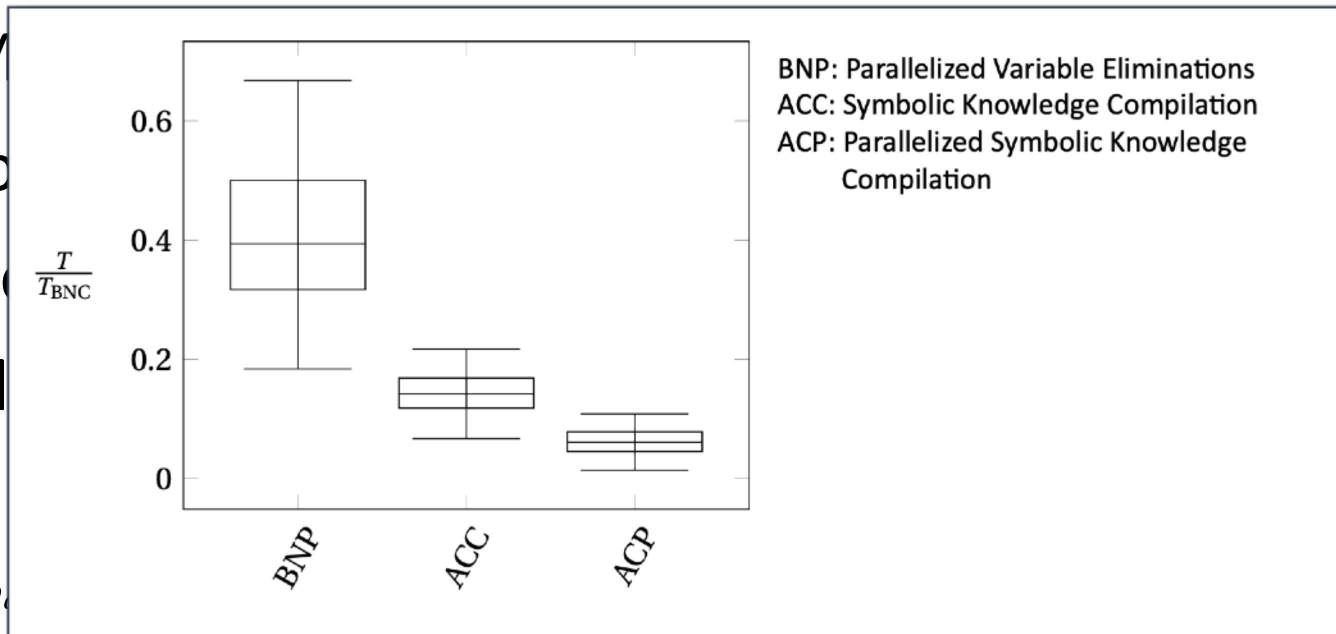
$\theta_{X_1} = [p_a, p_b]$ $\quad$ $\theta_{X_2|X_1=1} = [1.0, 0.0]$

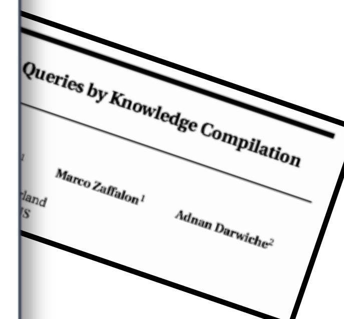




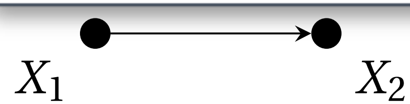
Symbolic Knowledge Compilation (TPM 2023)

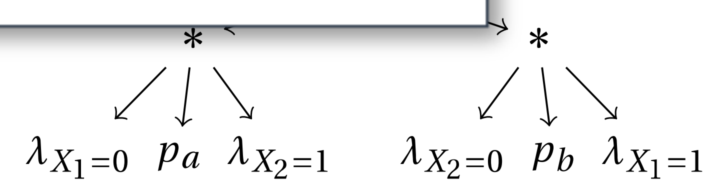
- Joint work with Adnan Darwiche and Hizuo Chen
- Our EM
- Equatio
- Compil
- Symbol





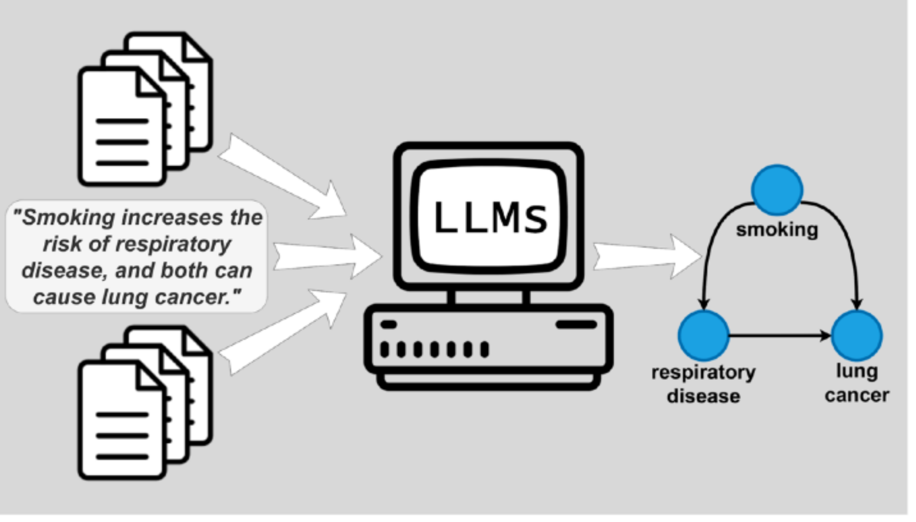
$\theta_{X_1} = [p$





Current Work: Causal Graphs by LLMs (AI4SCIENCE 2024)

- GPT parsing causal statements in natural language



- Link with IPs? Multiple causal graphs might be returned!
- Many recent papers on bounding counterfactual wrt **ignorance** about the causal structure (credal structures?)

Our Earlier Work:

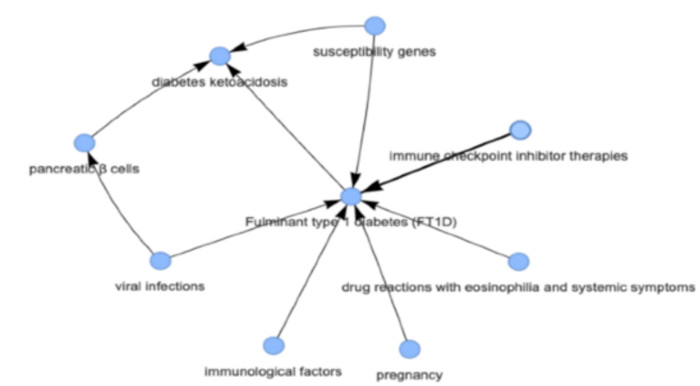
[Submitted on 22 Dec 2023]

Zero-shot Causal Graph Extrapolation from Text via LLMs

Alessandro Antonucci, Gregorio Piqué, Marco Zaffalon

Fulminant type 1 diabetes (FT[illegible]) is [illegible]novel type of type 1 diabetes that is caused b[illegible]remely [illegible]pid [illegible]he pancrea[illegible]c β cells. Early diagnosis [illegible] treatment of diabetes ketoacidos[illegible], wh[illegible]an be life-threatening. Understanding its triggers or promoting factors plays an important role in the prevention and treatment of FT1D. In this review, we summarised the various triggering factors of FT1D, including susceptibility genes, immunological factors (cellular and humoural immunity), immune checkpoint inhibitor therapies, drug reactions with eosinophilia and systemic symptoms or drug-induced hypersensitivity syndrome, pregnancy, viral infections, and vaccine inoculation. This review provides the basis for future research into the pathogenetic mechanisms that regulate FT1D development and progression to further improve the prognosis and clinical management of patients with FT1D.

LLMs



Prompt Engineering

```
You will be provided with a text delimited by the <Text></Text>
xml tags, and a pair of entities delimited by the <Entity></Entity>
xml tags representing entities extracted from the given text.

  Text:
   <Text>Cobalt metal fume and dust cause upper respiratory
   tract irritation, chronic interstitial pneumonitis,
   and skin sensitization.</Text>

  Entities:
   <Entity>fume[illegible]</Entity>
   <Entity>sens[illegible]

Read the provid[illegible] text care[illegible]y [illegible] the [illegible]ontext
and content. Examine the roles, interactio[illegible], and details
surrounding the entities within the text.

Bas[illegible]y on the infor[illegible]ation in the text, determ[illegible]e the
mos[illegible] likel[illegible] relat[illegible] the entities
fro[illegible]

  Options:
   A: "fume" caus[illegible]"sensitization";
   B: "sensitization" causes "fume";
   C: "fume" and "sensitization" are not directly causally related;

Your response should analyze the situation in a step-by-step manner
ensuring the correctness of the ultimate conclusion,
which should accurately reflect the likely causal connection
between the two entities, based on the information
presented in the text. If no clear causal relationship is apparent,
select the appropriate option accordingly.

Then provide your final answer within the tags
<Answer>[answer]</Answer>, (e.g. <Answer>C</Answer>).
```

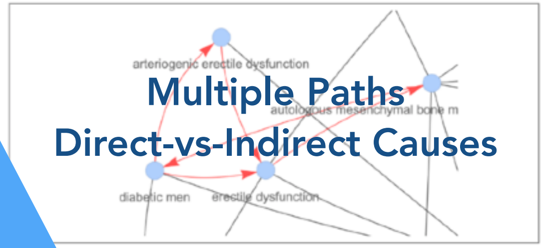
🙂

Sentence	Orientation
Zinc is essential for growth and cell division.	A → B
The infection came from a wound.	A ← B
As we saw earlier, helicobacter is responsible for causing stomach ulcer.	A → B
The pseudolesion was caused by drainage of the paraumbilical vein.	A ← B

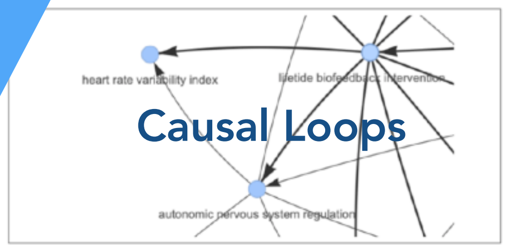
Good Causal Relation Identification/Orientation

		Ground Truth	
		A → B	A ← B
GPT	A → B	335	7
	A ← B	6	650

☹️



Multiple Paths
Direct-vs-Indirect Causes



Causal Loops

Results (LLM vs. Fine-Tuning, F1 score)

- Bert $\gg$ (FS) LLM

Model	Approach	Binary class	Multi-class
GPT 3.5 turbo	Zero shot (ZS)	0.59	0.37
GPT 3.5 turbo	Zero shot with Cues (ZS-Cues)	0.66	0.51
GPT 3.5 turbo	Few shot with Cues (FS-Cues)	0.62	0.50
BERT-base-cased	Full Fine Tuning (FFT)	0.92	0.87

With 10-shot small improvement of ZS, but no wrt Cues ...

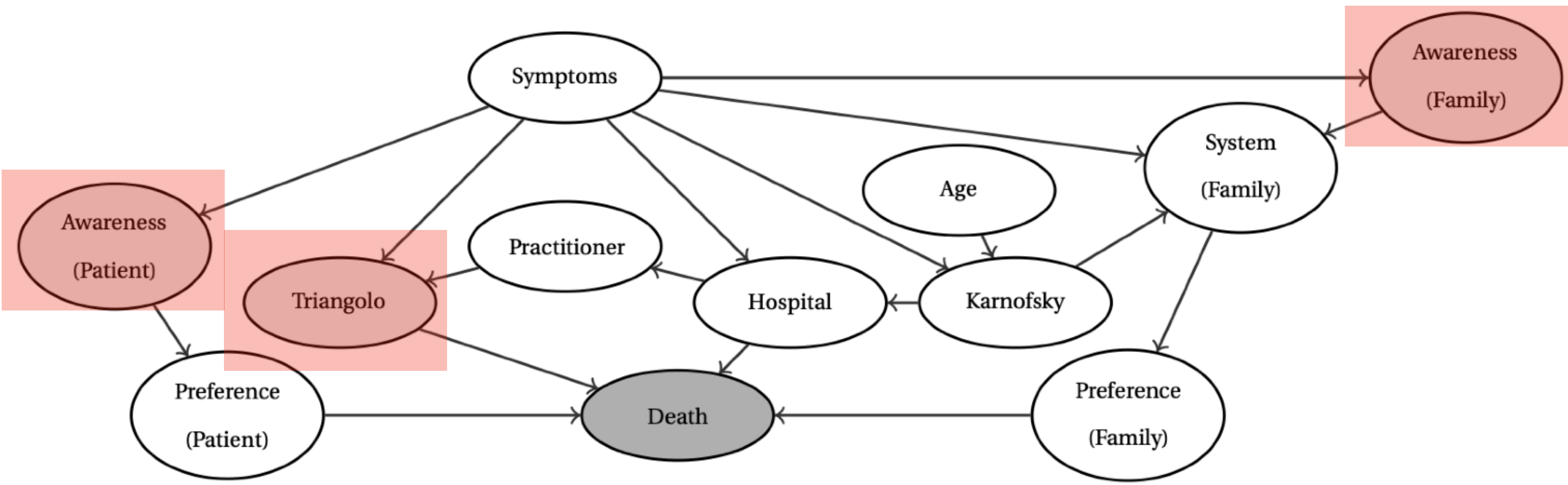
Counterfactual Analysis in Palliative Cares by Causal EM

- Importance of a variable?
- Probability of necessity and sufficiency

$$PNS := P(Y_{X=1} = 1, Y_{X=0} = 0)$$

- 15 EM runs before convergence

PNS(Family_Awareness) ∈ [0.06,0.10]



PNS(Patient_Awareness) ∈ [0.03,0.10]

PNS(Triangolo) ∈ [0.30,0.31]

Coun

- Im
- Pr
- 15

One should act on Triangolo first: for instance, by making Triangolo available to all patients, we should expect a reduction of people at the hospital by 30%

This would save money too, and would allow politicians to do economic considerations as to which amount it is even economically profitable to fund Triangolo, and have patients die at home, rather than spending more to have patients die at the hospital

.06,0.10]

reness
mily)

Aware
(Patie

PNS(Patient_Awareness) ∈ [0.03,0.10]

PNS(Triangolo) ∈ [0.30,0.31]

Conclusions

- Causality theories have an intimate connection with IPs
- Past research about CNs might offer new tools for causal analysis
- IPs offer formalism for a deeper SCMs understanding
- (Our) current challenge: learn non-canonical structural equations
- This also involves neuro-symbolic approaches with neural nets playing the role of (approximating) structural equations
- Plugging causal symbolic knowledge into (large) neural models can be a promising direction to solve current limitations (halucinations)
- Lot of works has to be done, causal machine (and reinforcement) learning is just at the beginning!

Conclusions

- Causality theories have an intimate connection with IPs
- Past research about CNs might offer new tools for causal analysis

I'll be here Tue&Fri
but also **alessandro@idsia.ch**

- Plugging causal symbolic knowledge into (large) neural models can be a promising direction to solve current limitations (halucinations)
- Lot of works has to be done, causal machine (and reinforcement) learning is just at the beginning!

Friday's Project

Practical Bounding of Counterfactual Inferences by Credal Networks

Consider an observational (or interventional or hybrid) dataset.

Say that you are interested in causal inference and in particolar in a counterfactual analysis.

You can use the dataset based on recovery/treatment/gender data, but if you have your own data is even better. I can support you during the project.

The main steps are:

- Identification of the causal, counterfactual, query we want to answer.
- Identification of the underlying causal graph and possible latent confounders.
- Specification (expert-based or canonical) of the structural equations.
- Implementation of the equivalent credal network.
- Computation of the bounds and analysis of the results.

Even if we have dedicated software tools for that, for small models like the one proposed to the participants, the analysis can also be sketched on paper (or in a Python notebook).