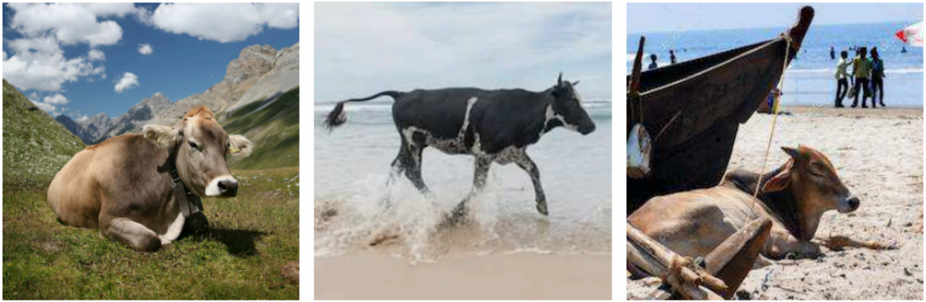
Highlights from the First CNRS AISSAI Thematic Quarter on Causality

Alessandro Leite

TAU, INRIA Saclay, LISN

December 1, 2023



(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

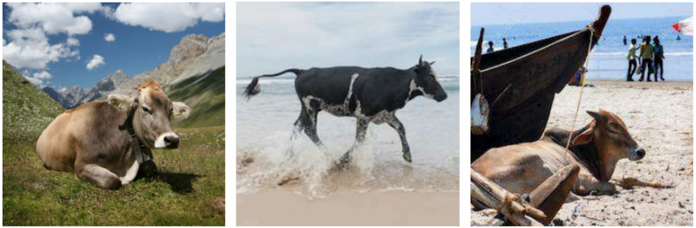
Figure 1: Cow and grass are spuriously correlated[1]

- Is the label "cow" really due to the presence of the cow in the image?

[1] Beery, Van Horn, and Perona, "Recognition in terra incognita".

Machine learning needs causal reasoning

- **React** to **events different from the training set**
- **Explain** what happened
- Capture **how** the **world works**
- Answer **what if**, **intervention**, and **counterfactuals** questions[2]
  - Was it the new tax policy that caused prices to increase?
  - How effective is a treatment in preventing a disease?
  - Can hiring records prove an employer's guilty of gender discrimination?



(A) **Cow: 0.99**, Pasture: 0.99, Grass: 0.99, No Person: 0.98, Mammal: 0.98

(B) No Person: 0.99, Water: 0.98, Beach: 0.97, Outdoors: 0.97, Seashore: 0.97

(C) No Person: 0.97, **Mammal: 0.96**, Water: 0.94, Beach: 0.94, Two: 0.94

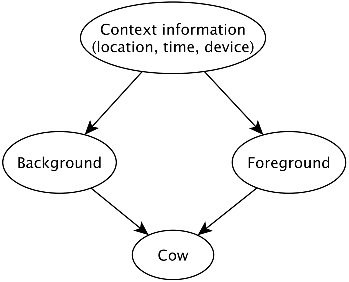


Figure 2: Possible modeling interpretation

[2] Pearl, "The seven tools of causal inference, with reflections on machine learning".

▶ If one can predict ...

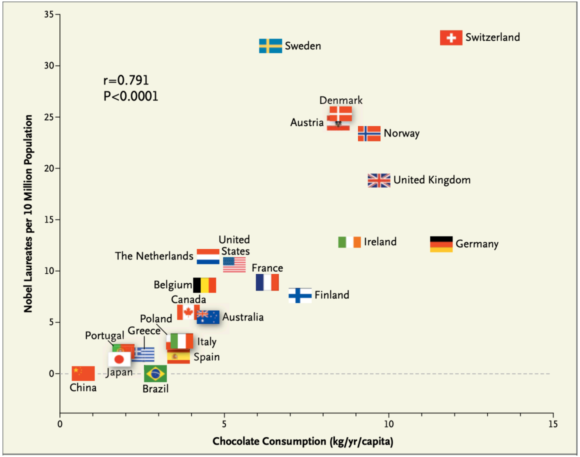


Figure 3: Messerli, "Chocolate consumption, cognitive function, and Nobel laureates"

▶ They can make things happen?

- Ask people to eat more chocolate to get more Nobel Prizes ...

Slide credit: *Sébag, M.(2020)*

A thematic quarter on causality

AI SSAI — AI for science, science for AI

- Opening session
- Three main symposiums
  - When Causal Inference Meets Statistical Analysis
  - Fundamental Challenges in Causality
  - Causality in Practice
- Two research schools
  - Spring School on Causality
  - Tools for Causality
- One Study Week on Causal Inference for Industry
  - Two industrial use cases:
    - Causal Discovery from Sequential Data (SAINT-GOBAIN)
    - Estimating Marketing Uplifts as Heterogeneous Treatment Effects with Meta-learners (Ekimetrics.)
- Materials available at quarter-on-causality.github.io


Alessandro Leite
TAU, Inria, Paris-Saclay University


Georges Oppenheim
Paris-Saclay University


Emilie Devijver
CNRS & University of Grenoble Alpes


Marianne Clausel
University of Lorraine


Eric Gaussier
University of Grenoble Alpes


Michèle Sébag

What is causality? – partial propositions

- **Causality** is what connects one process (the "**cause**") to another (the "**effect**"). The former is partially responsible for the latter, and the latter is partially dependent on the former [Pearl,2009].
  - If **X** is a **necessary cause** of **Y**, then, the presence of **Y** implies a prior occurrence of **X**; however, the presence of **X** does not imply that **Y** occur.
  - If **X** is a **sufficient cause** of **Y**, then the presence of **X** necessarily implies the subsequent occurrence of **Y**; however, the presence of **Y** does not imply the prior occurrence of **X**, as another cause may be responsible for it.
  - **X** is an **INUS condition** of Y if it is an **i**nsufficient but **n**on-redundant part of a condition which is itself **u**nncessary but sufficient for the occurrence of **Y** [Warr and Warr,2016].

Slide credit: *Chambaz, A. (2019)*

What is causality? – interventionists' interpretation

- "A necessary and sufficient condition for **X** to be a **direct cause** of **Y** with respect to some variable set $\boldsymbol{V}$ is that there is a ***possible intervention*** on **X** that will change **Y** (or the probability of **Y**) when all other variables are held fixed at some value by interventions" [Woodward,2005]
- The existence of a **possible intervention** is a necessary and sufficient condition for **direct type-level cause**.
- **Direct cause** $X \to Y$

$$P_{X_j|\boldsymbol{do}(X_i=x,\boldsymbol{X}_{\setminus ij}=c)} \neq P_{X_j|\boldsymbol{do}(X_i=x',\boldsymbol{X}_{\setminus ij}=\boldsymbol{c})}$$

- **Example**:
  C: Cancer, **S**: Smoking, **G**: Genetic factors

$$P(C|\boldsymbol{do}(S=0,G=0)) \neq P(C|\boldsymbol{do}(S=1,G=0))$$

- **X** is a cause of **Y** *iff*
  changing **X** leads to a change in **Y**,
  keeping everything else constant.

Gold standard: randomized controlled trials (RCTs)

- Draw **i.i.d.** samples, from two subsets:
  - T = 1: treatment group
  - T = 0: control group
- **Estimate the average treatment effect (ATE)**[3]

$$\text{ATE} = \mathbb{E}[Y(1) - Y(0)]$$

- where:
  - Y: outcome (survival)
  - X: covariates
  - T: treatment $(0 or 1)$
  - $Y_i(0)$: outcome of the i-th sample if it does not get the treatment
  - $Y_i(1)$: outcome of the i-th sample if it does get the treatment

One knows only one output of $Y_i(0)$ and $Y_i(1)$

[3] It is also known as average causal effect (ACE)

Potential outcomes – estimating average treatment effect (ATE) [Rubin,2005]

- It works under certain assumptions

$$
\begin{aligned}
\text{ATE} &= \mathbb{E}[Y(1) - Y(0)] \\
&= \mathbb{E}[Y(1)] - \mathbb{E}[Y(0)] \\
&\qquad\qquad\qquad \text{linearity of expectation} \\
&= \mathbb{E}_x[\mathbb{E}[Y(1)|X]] - \mathbb{E}_x[\mathbb{E}[Y(0)|X]] \\
&\qquad\qquad\qquad \text{expectation over covariates} \\
&= \mathbb{E}_x[\mathbb{E}(Y(1)T = 1, X)] - \mathbb{E}_x[\mathbb{E}(Y(1)T = 0, X)] \\
&\qquad\qquad\qquad \text{no hidden confounder; no unobserved common causes overlap assumption} \\
&\qquad\qquad\qquad \text{T=1 and T=0 are observed in the data} \\
&= \mathbb{E}_x[\mathbb{E}[Y|T = 1, X]] - E_x[\mathbb{E}[Y|T = 0, X]] \\
&\qquad\qquad\qquad \text{consistency} \, Y_i(1) \sim Y|T = 1, X = X_i
\end{aligned}
$$

Questions-Assumptions-Data (QAD) and the Pearl's Causal Hierarchy (PCH)

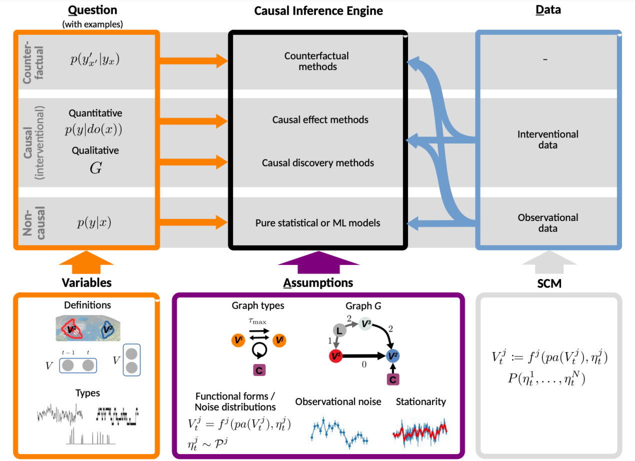


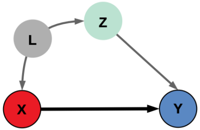
Figure 4: Question-Assumptions-Data template[4]

[4] Runge et al., "Causal inference for time series".

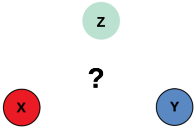
Causal inference framework

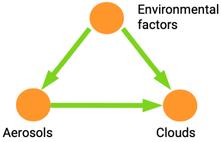
Two types of questions:

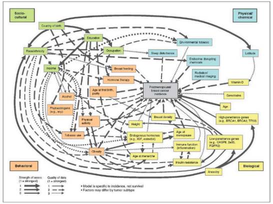
1. **Assume qualitative causal graph** to **quantify causal effects:**



2. Make **general assumptions** to **learn causal graph:**







Slide credit: *Jakob Runge*

Observational causal discovery

- **Similar to machine learning**
  - Given the data, infer the causal models
  - Data quality, quantity, and learning criterion may be challenging
- **Difference**: functional causal models
  - **Assumptions**
    - **Causal sufficiency**: no unobserved confounders
    - **Causal Markov**: all d-separations in the causal graph G imply conditional independence in the observational distribution P
    - **Causal faithfulness**: all conditional independence in P imply d-separations in G

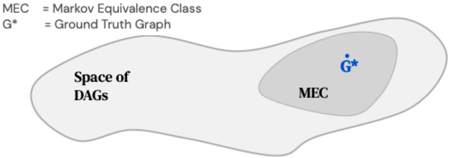


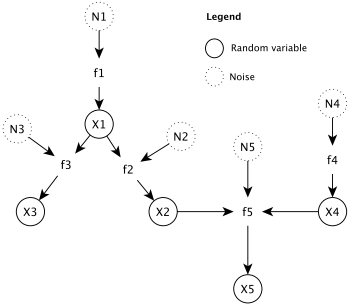
Image credit Rosemary and Bauer, 2021

Functional models a.k.a. structural causal models (SCM) [Pearl,2009]

- If a DAG G formalizes the **causal relation** between the random variables $X_1, \ldots, X_n$, then every X_j can be written as a deterministic function of $\boldsymbol{PA_j}$ and a noise variable N_j

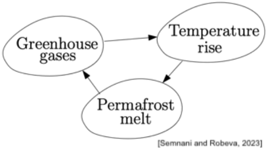
$$X_j = f_j(\boldsymbol{PA}_j, N_j)$$

where N_j are the noises (i.e., all unobserved influences), and they are all jointly independent

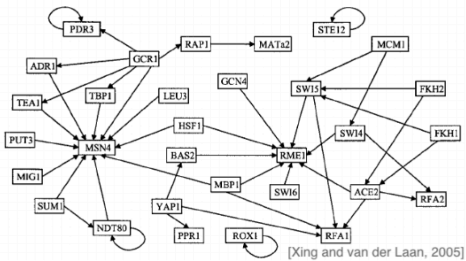


- **Markov condition in functional models**: every joint distribution $P(X_1, \ldots, X_n)$ generated according to the causal faithfulness condition satisfies the Markov conditions relative to G.
- **Functional models** formalize the conception that the outcome of an experiment is completely determined by the values of all relevant parameters where the only uncertainty stems from the fact that some of them are hidden.

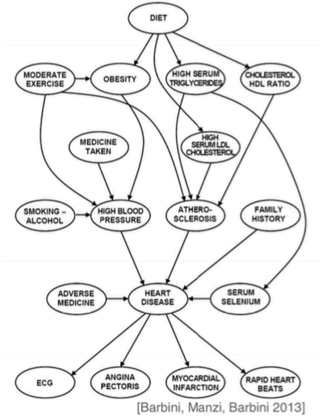
Acyclicity assumption does not hold in different domains



Climate science



GENE REGULATORY NETWORKS



DISEASE DIAGNOSIS GRAPHS

Image credit Robeva and Semnani, 2023

- How can we learn the structure of these graphs from observations?

Looking at higher order moment to introduce hidden variables

$$X = (I - \Lambda)^{-T}\varepsilon.$$

Definition

The *linear structural equation model* $\mathcal{M}^{(2,3)}(G)$ of second and third order moments corresponding to a DAG $G = (V, E)$ with $|V| = n$ is defined as

$$\begin{aligned}\mathcal{M}^{(2,3)}(G) = \{(S = (I - \Lambda)^{-T}\Omega^{(2)}(I - \Lambda)^{-1},\\ T = \Omega^{(3)} \bullet (I - \Lambda)^{-1} \bullet (I - \Lambda)^{-1} \bullet (I - \Lambda)^{-1}) :\\ \Omega^{(2)} \text{ is } n \times n \text{ positive definite diagonal matrix},\\ \Omega^{(3)} \text{ is } n \times n \times n \text{ diagonal 3-way tensor, and } \Lambda \in \mathbb{R}^{E}\}.\end{aligned}$$

Here, $\bullet$ denotes the *Tucker product*.

Theorem (Améndola, Drton, Grosdos, Homs-Pons, and R., 2021+)

The set of second and third order moments (T, S) of a linear non-Gaussian causal model corresponding to a tree DAG are precisely the ones that satisfy certain quadratic binomials which arise as the 2×2 minors of certain matrices constructed from the DAG.

Slide credit: *Robeva, 2023*

Vanishing of cumulants

- For a zero-mean random vector $X = (X_1, \ldots, X_d)$, its *k-th order cumulant* is an $d \times \cdots \times d$ (k times) tensor $C^{(k)}$ whose entries can be obtained from the moments of X, e.g. for $k = 4$:

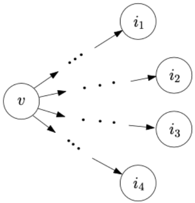
$$C^{(4)}_{i_1,i_2,i_3,i_4} = \mathbb{E}[X_{i_1}X_{i_2}X_{i_3}X_{i_4}] - \mathbb{E}[X_{i_1}X_{i_2}]\mathbb{E}[X_{i_3}X_{i_4}] - \mathbb{E}[X_{i_1}X_{i_3}]\mathbb{E}[X_{i_2}X_{i_4}] - \mathbb{E}[X_{i_1}X_{i_4}]\mathbb{E}[X_{i_2}X_{i_3}].$$

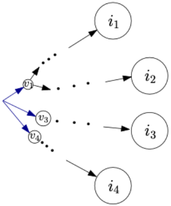
Theorem (Robeva and Seby, 2020)

If X comes from a linear non-Gaussian acyclic model with graph $G = (V, E, H)$ and X has cumulants $C^{(k)}$, then

$$C^{(k)}_{i_1,\ldots,i_k} = 0$$

if and only if there is no k-trek between the vertices *$i_1, \ldots, i_k$ in G.*





Slide credit: *[Robeva and Seby,2021]*

Causal Inference with Information Algebras

Information Dependency Models and Information Fields

- **Information dependency models**: causality with **information fields**
- **Information fields**: Witsenhausen's 1971 paper [1]
- Witsenhausen's motivation: control of multi-agent systems
- but in fact, it is a very generic tool
  - Used to revisit the foundations of game theory[2]
  - Theoretical toolbox for causality: **the Information Dependency Model** (IDM)

[1] *On information structures, feedback and causality.*
[2] *Kuhn's equivalence theorem for games in product form*

Making the case for Information Dependency Model (IDM)

- Unlock **mathematical toolboxes**
- **Unifying and generalizing** framework for causality[3]
- Elegant style of expression and proof : **equational reasoning**
- Potential to **bridge** causality, game theory, control and Reinforcement Learning

[3] can deal with spurious edges, cycles

Slide credit: *Heymann, Benjamin and De Lara, Michel and Chancelier, Jean-Philippe*

Scaling causal discovery through diffusion models [Sanchez et al.,2023]

Overview

1. Causal discovery can be efficiently done via **topological ordering**
2. Assuming additive noise models (ANM) the log-likelihood's Hessian can be used to find **leaf nodes.**
3. Diffusion models can approximate a **Hessian**

Algorithm - Greedy

$$\text{leaf} = \arg\min_{\mathrm{x}_i \in \mathbf{x}} \mathrm{Var}_{\boldsymbol{X}} \left[\nabla_{\mathbf{x}} \boldsymbol{\epsilon}_\theta(\boldsymbol{X}, t)\right]$$

For d-1 variables

1. Train diffusion model
2. Find leaf
   1. Pass data through diffusion model
   2. Backpropagate w.r.t. inputs to obtain Hessian
   3. Compute the variance
   4. Leaf is diagonal element with smallest variance
   5. $\pi = [\pi, \text{leaf}]$
   6. Remove leaf from data

Score with Diffusion Models $\quad \boldsymbol{\epsilon}_\theta$

Training

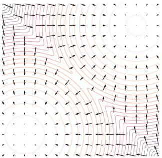
Denoise $\mathbf{x}_t$, a corrupted version of a data point.

$$\theta^* = \arg\min_\theta \mathbb{E}_{\mathbf{x}_0, t, \epsilon}\left[\lambda(t)\, \|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \epsilon\|_2^2\right]$$

t controls the amount of corruption.

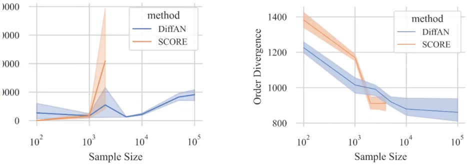
Intuition

Denoising approximate the score function $\nabla_x \log p(\boldsymbol{x})$



Contours: Density of a mixture of two Gaussians.
Vector field: Score $\nabla_x \log p(\boldsymbol{x})$

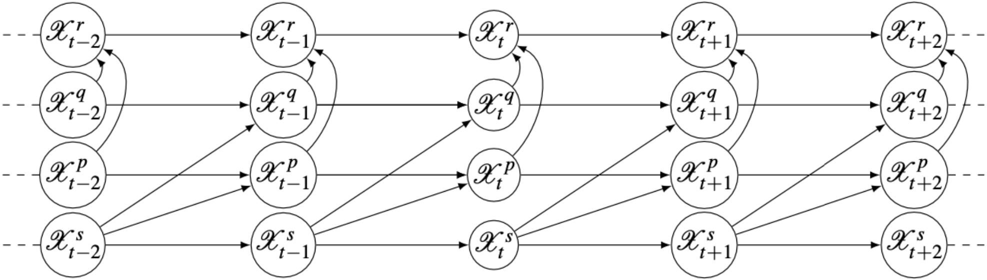
https://yang-song.net/blog/2021/score/ 8



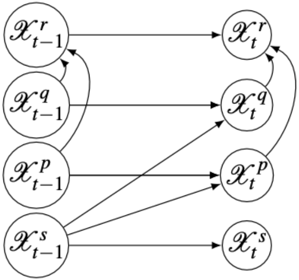
Synthetic data with d = 500 for different dataset sizes.
Error bars across different samples of ER and SF graphs.

Image credit [Sanchez et al.,2023]

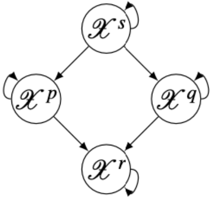
Time series and causal representation learning [Assaad, Devijver, and Gaussier,2022]



Full Time Causal Graph (a)



Window Causal Graph (b)

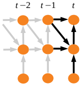


Summary Causal Graph (c)

Learning causal graphs

Given data and *general assumptions*, estimate causal graph from observational distribution

Ground truth



Time series case:

- ***PCMCI causal discovery framework – main idea:***

1. PC_1 lagged skeleton phase

→ Find supersets $\widehat{\mathcal{B}}^-_t(X^j_t)$ of lagged parents

for $p=0,1,\dots:$ $\mathbf{S}_p = \{\mathrm{adj}^p_{t-1:t-\tau_{\max}}(X^j_t)\}^p_{l=1}$

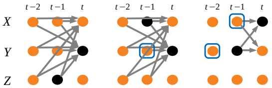
$X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \mathbf{S}_p$ for $\tau > 0$

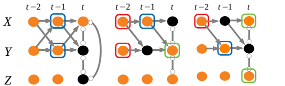
2. Contemp. + lagged MCI skeleton phase

→ Use these parent-supersets as added conditions

for $p=0,1,\dots:$ $\mathbf{S}_p \subseteq \mathrm{adj}^p_t(X^j_t)$ of cardinality p

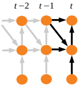
$X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \mathbf{S}_p, \widehat{\mathcal{B}}^-_t(X^j_t)\setminus\{X^i_{t-\tau}\}, \widehat{\mathcal{B}}^-_{t-\tau}(X^i_{t-\tau})$ for $\tau \geq 0$





3. Orientation phase

→ using collider and Meek rules



Better skeleton
→ better orientations

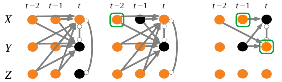
- ***standard PC algorithm:***

1. Skeleton phase

→ Iterate through subsets of adjacencies

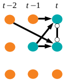
for $p=0,1,\dots:$ $\mathbf{S}_p \subseteq \mathrm{adj}^p_{t:t-\tau_{\max}}(X^j_t)$ of cardinality p

$X^i_{t-\tau} \perp\!\!\!\perp X^j_t \mid \mathbf{S}_p$ for $\tau \geq 0$



2. Orientation phase

→ using collider and Meek rules



143

Slide credit: *Runge, Jakob and Ninad, Urmi and Wahl, Jonas*

Causal representation learning and LLMs

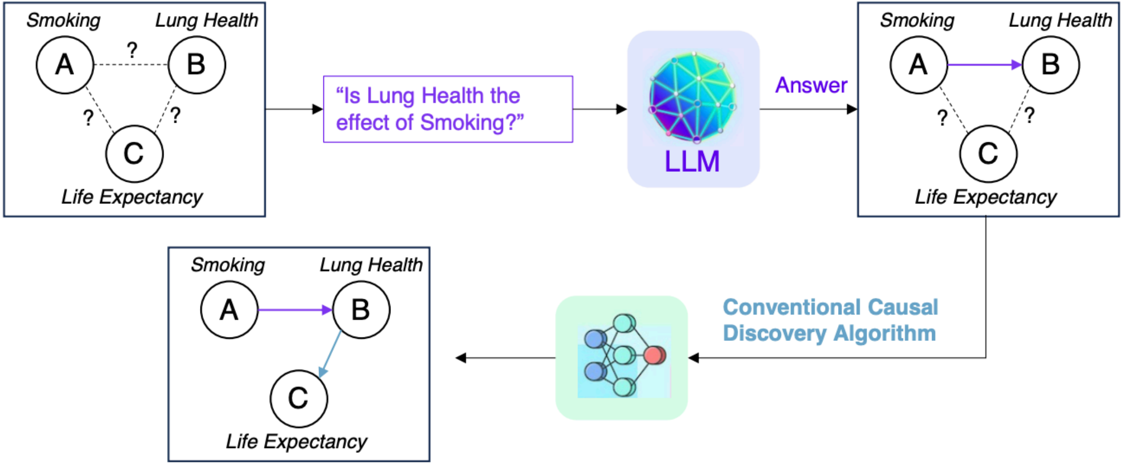


Figure 5: LLMs-based causal analysis pipeline [Kıcıman et al.,2023]

Combining LLMs and PCMCI algorithm for causal discovery

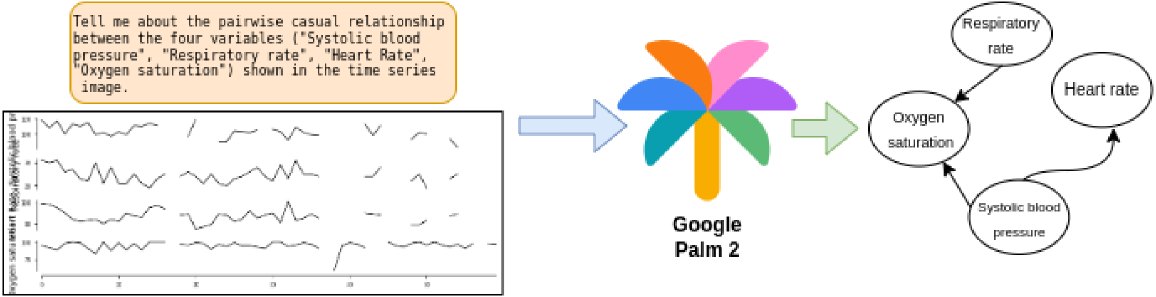


Figure 6: Google PaLM and the MIMIC III dataset

Counterfactual inference as a mass transportation problem

The effect of $\mathrm{do}(S = s'|S = s)$ is fully characterized by the coupling

$$\pi^*_{\langle s'|s\rangle} := \mathcal{L}\left((X, X_{S=s'})|S = s\right).$$

It assigns a probability to all the pairs (x, x') between an observable value x and a counterfactual counterpart x'.

This coupling admits $\mu_s := \mathcal{L}(X|S = s)$ as first marginal and $\mu_{\langle s'|s\rangle} := \mathcal{L}(X_{S=s'}|S = s)$ as second marginal.

Remark: Therefore, $\pi^*_{\langle s'|s\rangle} \in \Pi(\mu_s, \mu_{\langle s'|s\rangle}) \neq \Pi(\mu_s, \mu_{s'})$.

Slide credit: *De Lara, Lucas. See [De Lara et al.,2021]*

Take-home message

- Causal inference provides a framework that integrates statistical and machine learning methods to answer causal questions from observational data
- Two settings:
  1. Assume known causal graphs and learn causal effects
  2. Learning causal graphs
- Different approaches have been proposed to enable research questions to be framed as causal questions and analysis the underlying assumptions to answer them
- The First CNRS AISSAI Thematic Quarter on Causality enabled us to explore them under different viewpoints

AI for science, science for AI

References I

- Assaad, Charles K, Emilie Devijver, and Eric Gaussier. "Survey and evaluation of causal discovery methods for time series". In: Journal of Artificial Intelligence Research 73 (2022), pp. 767–819.
- Beery, Sara, Grant Van Horn, and Pietro Perona. "Recognition in terra incognita". In: European Conference on Computer Vision. 2018, pp. 456–473.
- De Lara, Lucas et al. "Transport-based Counterfactual Models". In: arXiv:2108.13025 (2021).
- Kıcıman, Emre et al. "Causal reasoning and large language models: Opening a new frontier for causality". In: arXiv:2305.00050 (2023).
- Messerli, Franz H. "Chocolate consumption, cognitive function, and Nobel laureates". In: The New England Journal of Medicine 367.16 (2012), pp. 1562–1564.
- Pearl, Judea. Causality. Cambridge University Press, 2009.
- — ."The seven tools of causal inference, with reflections on machine learning". In: Communications of the ACM 62.3 (2019), pp. 54–60.

References II

- Robeva, Elina and Jean-Baptiste Seby. "Multi-Trek Separation in Linear Structural Equation Models". In: SIAM Journal on Applied Algebra and Geometry 5.2 (2021), pp. 278–303.
- Rubin, Donald B. "Causal inference using potential outcomes: Design, modeling, decisions". In: Journal of the American Statistical Association 100.469 (2005), pp. 322–331.
- Runge, Jakob et al. "Causal inference for time series". In: Nature Reviews Earth & Environment 4.7 (2023), pp. 487–505.
- Sanchez, Pedro et al. "Diffusion Models for Causal Discovery via Topological Ordering". In: 11th International Conference on Learning Representations. 2023.
- Warr, Jason and Jason Warr. "INUS Conditions and Criminological Theory". In: An Introduction to Criminological Theory and the Problem of Causation (2016), pp. 55–70.
- Woodward, James. Making things happen: A theory of causal explanation. Oxford University Press, 2005.


That's all, Folks!
d/l