COMSM0152, Foundations of ProAI: Causality

James Cussens

School of Computer Science
University of Bristol

15th January 2026

Agenda

- Identifying causal effects in the potential outcomes framework
- Causal graphs
- Causal discovery
- Causal representation learning

The potential outcomes framework

Quoting the presentation in [WRR25], let the observed data be (L, A, Y) where:

- L is a vector of covariates,
- $A \in \{0, 1\}$ is a binary treatment indicator, and
- Y is the outcome of interest

Let:

- $Y(0)$ be value of Y that would be observed for a given unit if assigned $A = 0$ (placebo), and
- $Y(1)$ be value of Y that would be observed for a given unit if assigned $A = 1$ (treatment)

The *Fundamental Problem of Causal Inference* is that we can never simultaneously observe both potential outcomes $Y(0)$ and $Y(1)$ for the same individual.

Causal effects

Note: a statistical model is *identifiable* if its parameters could be found were we to have infinite data, i.e. different parameters determine different probability distributions (over observable data).

Individual causal effect Defined as $Y_i(1) - Y_i(0)$ for unit i. This is not observable or, typically, identifiable.

Average causal effect (ACE) Defined as $\mathrm{ACE} = \mathbb{E}\{Y(1) - Y(0)\}$. This is typically what we want to estimate and is identifiable if e.g. we have a *randomised controlled trial (RCT)*.

Conditional average treatment effect (CATE) Defined as $\mathrm{CATE} = \mathbb{E}\{Y(1) - Y(0)|X = x\}$, where $X \subset L$. Needed for personalised medicine.

Some notation

- As we shall see later, there are important connections between causality and *conditional independence*.
- If A, B and S are three disjoint sets of random variables, we write $A \perp B|S$ for some probability distribution P if:

$$P(A, B|S) = P(A|S)P(B|S)$$

Equivalently,

$$P(A|B, S) = P(A|S)$$

- Intuitively, "once we know the value of S then observing A provides no information about B (and vice-versa)."

Estimating average causal effect

To estimate ACE, we need some assumptions:

Consistency $Y = Y(a)$ if $A = a$.

No interference Each unit's potential outcome is not affected by the treatment of others.

No unmeasured confounding $A \perp Y(a)|L,\ a = 0, 1$

If these assumptions are met ACE is identified using this *g-formula* [Rob86]:

$$\mathrm{ACE} = \mathbb{E}_L\{\mathbb{E}(Y|A = 1, L) - \mathbb{E}(Y|A = 0, L)\}$$

Estimating CATE with metalearners: T-learner

- Basic idea: Decompose the estimation of CATE into several subregression problems, each of which can be solved by any regression model.
- In the relevant paper [KSBY19], the covariates are denoted X (not L) and the treatment indicator is W (not A).

T-learner:

- Get estimate $\hat{\mu}_0(x)$ of the control response function $\mu_0(x) = \mathbb{E}(Y(0)|X = x)$ using data $\{(X_i, Y_i\}_{W_i=0}$.
- Get estimate $\hat{\mu}_1(x)$ of the treatment response function $\mu_1(x) = \mathbb{E}(Y(1)|X = x)$ using data $\{(X_i, Y_i\}_{W_i=1}$.
- Get T-learner estimate of CATE by simply subtracting: $\hat{\tau}_T(x) = \mu_1(x) - \mu_0(x)$

Estimating CATE with metalearners: S-learner

- Treatment indicator treated like just another covariate and we estimate the combined response function $\mu(x, w) = \mathbb{E}(Y|X = x, W = w)$, using some regression method.
- Get S-learner estimate of CATE as: $\hat{\tau}_S(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$

Estimating CATE with metalearners: X-learner

- Estimate the response functions $\mu_0(x) = \mathbb{E}(Y(0)|X = x)$ and $\mu_1(x) = \mathbb{E}(Y(1)|X = x)$ somehow. Denote the estimated functions $\hat{\mu}_0$ and $\hat{\mu}_1$.
- Impute the individual treatment effects for units in both the treatment and control groups:

$$\begin{aligned} \tilde{D}_i^1 &:= Y_i^1 - \hat{\mu}_0(X_i^1) \\ \tilde{D}_i^0 &:= \hat{\mu}_1(X_i^0) - Y_i^0 \end{aligned}$$

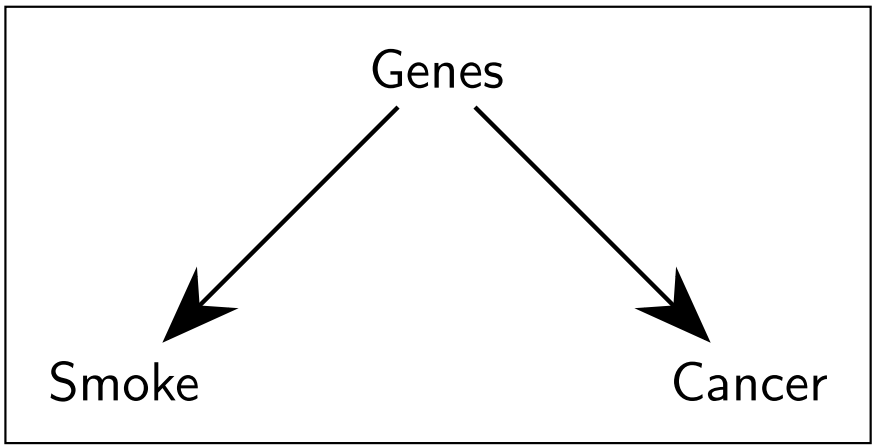
- Note that if we had $\hat{\mu}_0 = \mu_0$ and $\hat{\mu}_1 = \mu_1$ then we would have $\tau(x) = \mathbb{E}[\tilde{D}^1|X = x] = \mathbb{E}[\tilde{D}^0|X = x]$.
- So use the $\tilde{D}^1$ data to get an estimate (somehow) $\tau_1(x)$ of $\tau(x)$ and also the $\tilde{D}^0$ data to get an estimate $\tau_0(x)$ of $\tau(x)$.
- Combine the $\tau_1(x)$ and $\tau_0(x)$ estimates to get our final X-learner estimate of the CATE!

$$\hat{\tau}(x) = g(x)\tau_0(x) + (1 - g(x))\tau_1(x)$$

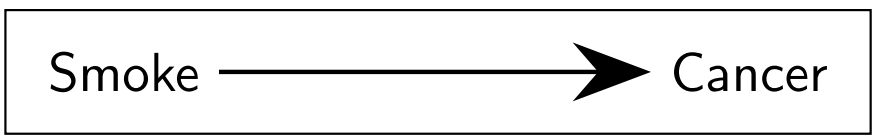
where $g(x) \in [0, 1]$.

Visualising (alleged!) confounding with a causal graph

- If we observe a correlation between smoking and cancer, what is the causal story?



or



(Causal) Bayesian networks

- The structure of a Bayesian network is a directed acyclic graph (DAG) where each node represents a random variable.
- Together with some parameters, the Bayesian network represents a joint probability distribution over the random variables in the DAG. This distribution factorises as follows:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | \mathrm{Pa}_i)$$

where Pa_i denotes the *parents* of x_i.

- This factorisation implies certain *conditional independence relations* which you can read off the DAG (with a bit of training!).
- In a *causal* Bayesian network the arrows represent causal relations.

Causal models and interventions

Schölkopf *et al.*: Toward Causal Representation Learning

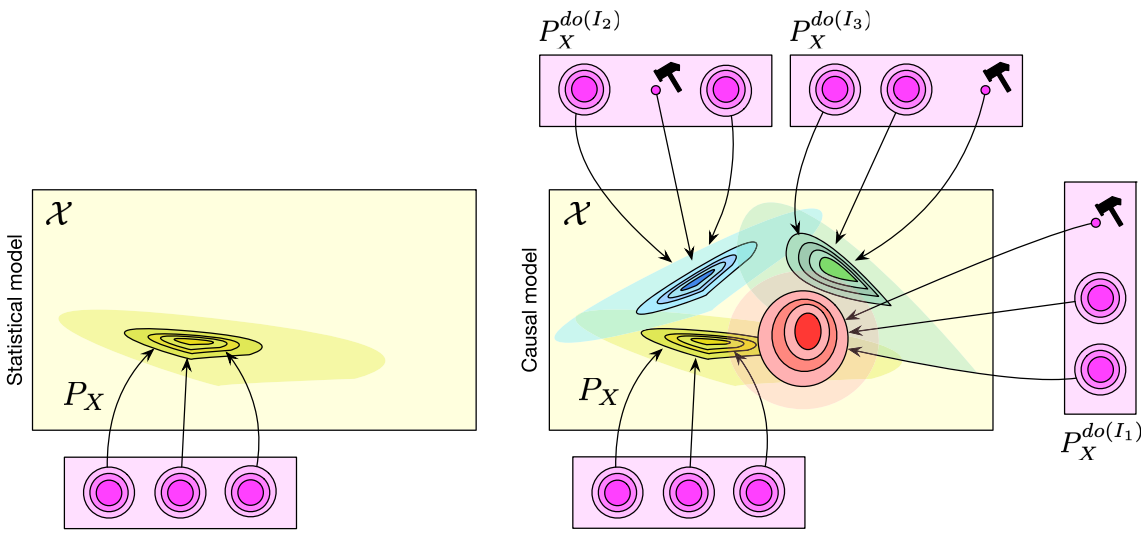


Fig. 1. ***Difference between statistical (left) and causal models (right) on a given set of three variables. While a statistical model specifies a single probability distribution, a causal model represents a set of distributions, one for each possible intervention (indicated with a 🔨).***

Causal DAGs with DAGitty

- Let's look at what causal DAGs represent, using DAGitty [TvdZGML16]

The adjustment formula

- If Z is a common cause of X and Y then the *adjustment formula* is:

$$P(Y = y|\mathrm{do}(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z)$$

- $\mathrm{do}(X = x)$ indicates *intervening* to set X to x
- Here we are "adjusting for" or "controlling for" Z.

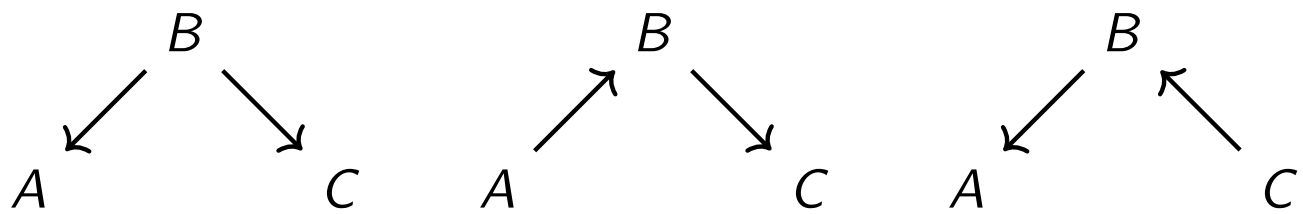
Causal discovery

- One can learn non-causal Bayesian networks in much the same way as any other statistical model, i.e. trade-off fit to data and number of parameters.
- I use an *integer programming* approach since I like being able to easily incorporate constraints on DAG structure.
- But is this a reliable way to construct *causal* graphs?
- I agree with Didelez [Did24]: causal discovery algorithms can suggest causal models for later (human) checking.

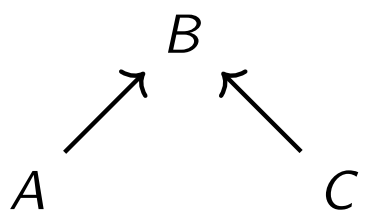
Difficulties in causal discovery

Computational Learning Bayesian networks, even if all variables observed, is known to be an NP-complete problem. It doesn't get any easier once you allow for latent (i.e. unobserved) variables!

Observational equivalence of distinct causal structures The following 3 causally distinct DAGs are *Markov equivalent* (encode the same conditional Independence relations):



However



is not Markov equivalent to any other DAG.

Causal representation learning

Schölkopf *et al.*: Toward Causal Representation Learning

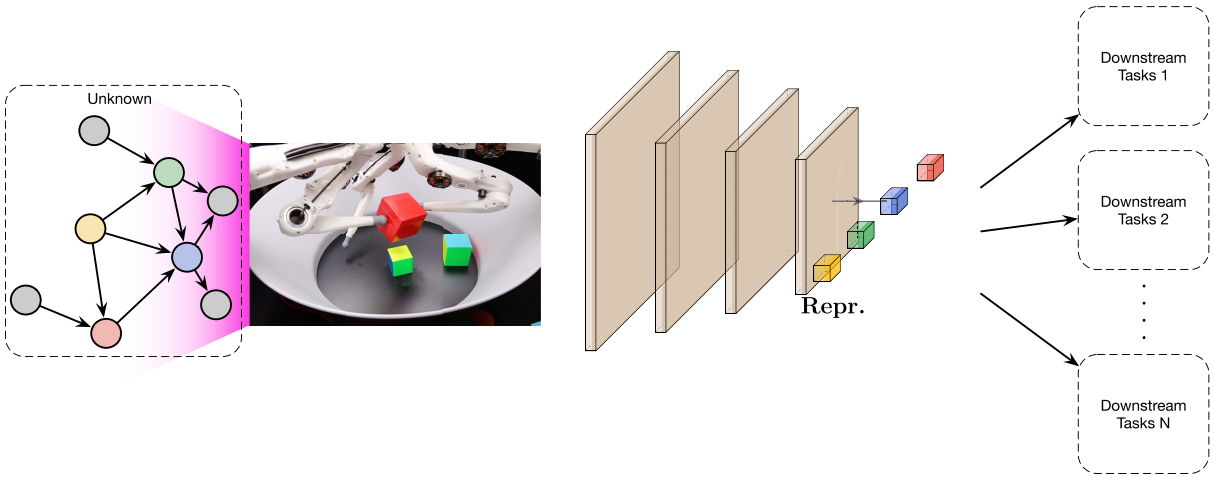


Fig. 2. ***Illustration of the causal representation learning problem setting. Perceptual data, such as images or other high-dimensional sensor measurements, can be thought of as entangled views of the state of an unknown causal system, as described in (10). With the exception of possible task labels, none of the variables describing the causal variables generating the system may be known. The goal of causal representation learning is to learn a representation (partially) exposing this unknown causal structure (e.g., which variables descri the system, and their relations). As full recovery may often be unreasonable, neural networks may map the low-level features to some high-level variables supporting causal statements relevant to a set of downstream tasks of interest. For example, if the task is to detect th manipulable objects in a scene, the representation may separate intrinsic object properties from their pose and appearance to achieve robustness to distribution shifts on the latter variables. Usually, we do not get labels for the high-level variables, but the properties of causal models can serve as useful inductive biases for learning (e.g., the SMS hypothesis).***

Causal representation learning

- Some assumptions used in causal representation learning are the *Independent Causal Mechanisms (ICM)* principle and its consequence the *Sparse Mechanism Shift (SMS)* hypothesis.

"ICM principle: The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. In the probabilistic case, this means that the conditional distribution of each variable given its causes (i.e., its mechanism) does not inform or influence the other mechanisms." [SLB+21]

"Sparse Mechanism Shift (SMS): Small distribution changes tend to manifest themselves in a sparse or local way in the causal/disentangled factorization, that is, they should usually not affect all factors simultaneously." [SLB+21]

The SMS hypothesis

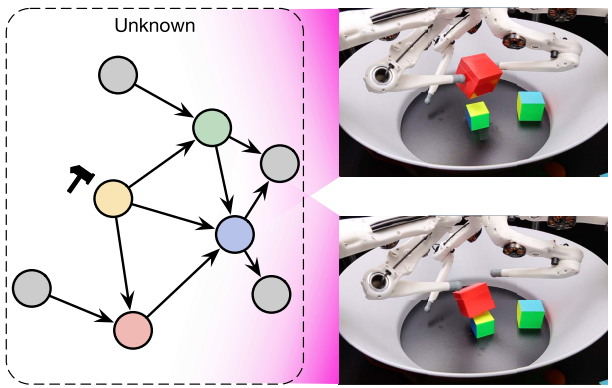


Fig. 3. ***Example of the SMS hypothesis where an intervention (which may or may not be intentional/observed) changes the position of one finger (🔨), and as a consequence, the object falls. The change in pixel space is entangled (or distributed), in contrast to the change in the causal model.***


Vanessa Didelez.
Invited commentary: where do the causal dags come from?
American Journal of Epidemiology, 193(8):1075–1078, 04 2024.


Sören R Künzel, Jasjeet S Sekhon, Peter J Bickel, and Bin Yu.
Metalearners for estimating heterogeneous treatment effects using machine learning.
Proceedings of the National Academy of Sciences, 116(10):4156–4165, 2019.


J. M. Robins.
A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect.
Mathematical Modelling, 7(9–12):1393–1512, 1986.


Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio.
Toward causal representation learning.
Proceedings of the IEEE, 109(5):612–634, 2021.


Johannes Textor, Benito van der Zander, Mark K. Gilthorpe, and George T.H. Maciej Liskiewicz, Ellison.
Robust causal inference using directed acyclic graphs: the R package 'dagitty'.
International Journal of Epidemiology, 45(6):1887–1894, 2016.


Linbo Wang, Thomas Richardson, and James Robins.
Causal inference: A tale of three frameworks, 2025.