Introduction to Causality

Ben Cardoen*, Ghassan Hamarneh*

* School of Computing Science, SFU

Presented in research meeting, 18th December, 2020

Abstract

In scientific discovery, engineering, imaging, and machine learning, it is often critical to understand what causes an event or observation, rather than focusing on correlation/association alone. In order to make this complex topic more accessible I would like to share what I learned on how causality can be applied and what concepts are essential in doing so. I will introduce the graphical causal model (Bayesian network), and show how you can translate human intuition on causality into formal axioms that fuse the causal graph with the probability space from observed events. We will discuss counterfactual causality, and end with an overview of recommendations on how to use causality in practice as well as current open issues and relevant papers that tackle those questions. [1]

[1] {bcardoen, hamarneh}@sfu.ca

Introduction to Causality

Ben Cardoen

Dec 18th, 2020

Medical Image Analysis Research Group, School of Computing Science, Simon Fraser University

Causality is key to thriving in an uncertain world

From intuition to causality

- Somehow humans learn to control a very diverse environment
- One part of this process is understanding how a certain need (effect) can be met
- Understanding $\sim$ causality

Cognitive psychology still cannot explain how a small child learns to understand and manipulate a complex world to not only survive, but thrive.

What is causality ?

Knowledge is power. –Francis Bacon

An event that 'causes' another event ?

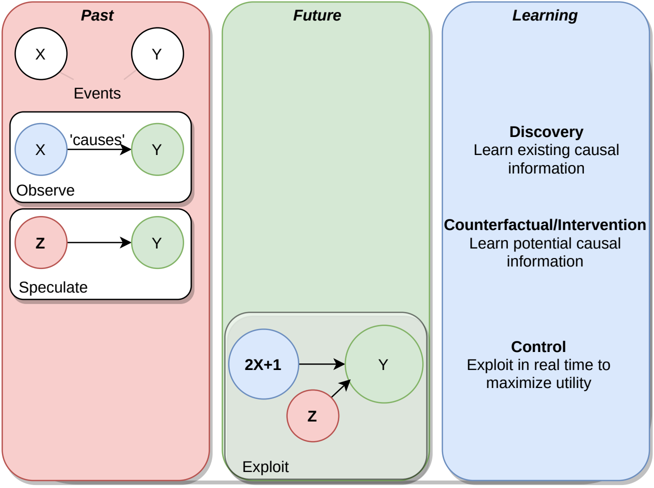
- Logical : $X \Rightarrow Y$
- Probabilistic : P(effect|cause) > P(effect) (Mellor, Suppes)
- Interventional : P(effect | do(cause)) > P(effect | do(!cause)) (Pearl [6])
- Boolean: $\exists (c, !c), (a, !a) \ \ \forall c, a, \in C, A$
- Continuous : Var(Cause) $\sim$ Var(Effect)
- How big of an effect ?

Not everyone believe{d,s} modelling causality can be formally done, or is feasible

- Bertrand Russell : *Causality is **pre-scientific**.*
- Karl Pearson : *Let's focus on correlation instead.*
- Fisher : *Let's develop experiment design* $\sim$ *test 1 while fixing all other variables.*

Causality is an intuitive, utilitarian relation of events

Causality is key to interacting with a complex world



Discovery
Learn existing causal information

Counterfactual/Intervention
Learn potential causal information

Control
Exploit in real time to maximize utility

An overview of the intuition behind causality, and the different tasks required in understanding the world.

A Universal Axiomatic Model

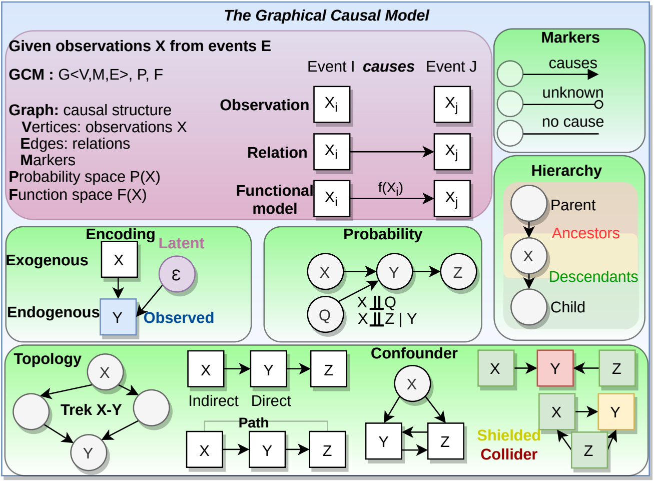
Translate intuitive understanding into an unambiguous model

Causality is

- A relation $X_i \rightarrow X_j,\ X_i, X_j \in X$ that is:
  - Irreflexive: $X_i \nrightarrow X_i$
  - Transitive: $X_i \rightarrow X_j, X_j \rightarrow X_k \Rightarrow X_i \rightarrow X_k$
  - Antisymmetric: $X_i \rightarrow X_j \Rightarrow (X_j \nrightarrow X_i)$
- A Graphical Causal Model on an event space **X**, described by a graph **G<V(X),E,M>** and probability space **P(X)**
- Link probability space and graph with axioms encoding intuition:
  - Markov Causal Condition
  - Minimality
  - Faithfulness

A Universal Axiomatic Model

Translate intuitive understanding into an unambiguous model



The Graphical Causal Model

Given observations X from events E

GCM : G<V,M,E>, P, F

Graph: causal structure
Vertices: observations X
Edges: relations
Markers
Probability space P(X)
Function space F(X)

Event I ***causes*** Event J

Observation X_i X_j

Relation $X_i \rightarrow X_j$

Functional model $X_i \xrightarrow{f(X_i)} X_j$

Markers
causes
unknown
no cause

Hierarchy
Parent
Ancestors
X
Descendants
Child

Encoding
Exogenous X
Latent ε
Endogenous Y
Observed

Probability
X → Y → Z
Q → Y
$X \perp\!\!\!\perp Q$
$X \perp\!\!\!\perp Z \mid Y$

Topology
X
Trek X-Y
Y

X → Y → Z
Indirect Direct

Path
X → Y → Z

Confounder
X
Y Z

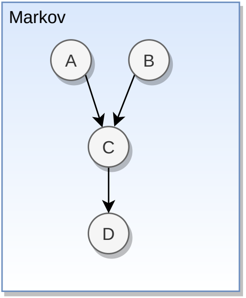
X → Y ← Z
X → Y
Z
Shielded
Collider

The complete cheat sheet for GCMs in all your future causal research.

The Markov condition

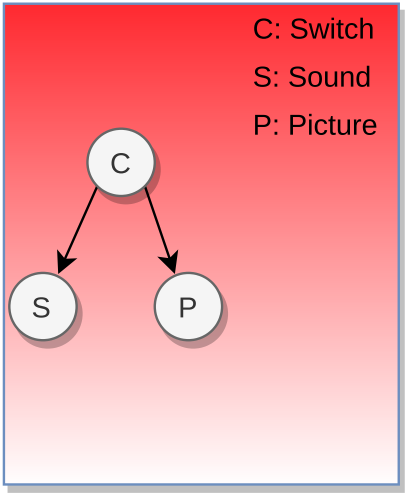
GCM G(V,E) and probability space P

- $W \in V$
- Q = $V - \{\text{Parents}(W) \cup \text{Descendants}(W)\}$
- $P(W, Q|\text{Parents}(W)) = P(W)P(Q)$



The Markov Condition illustrated: P(A,B,C,D) = P(A)P(B)P(D|C)P(C|A,B)

The Markov condition can be easily broken



A counter-example. A television set has a faulty switch that not always turns it on. When the TV is on, both sound and picture are produced. $P(S|C) < P(S|P,C)$. A more current example are my bluetooth headphones that connect only 90% of the time, where S, P are the left and right ears respectively.

Statistical Paradox 1 : mixing data (Yule, Pearson)

Suppose gender determines if a person has a certain trait.
Assume $P[\text{male} \wedge +] = 1/2$, $P[\text{female} \wedge +] = 1/10$

Joint probability table

F	S	Both	M	D	Both
+	+	0.25	+	+	0.01
+	-	0.25	+	-	0.09
-	+	0.25	-	+	0.09
-	-	0.25	-	-	0.81

Table: F(ather), S(on), M(other), D(aughter)

Joint probability table ignoring gender

Parent	Offspring	Both
+	+	0.13
+	-	0.13
-	+	0.17
-	-	0.53

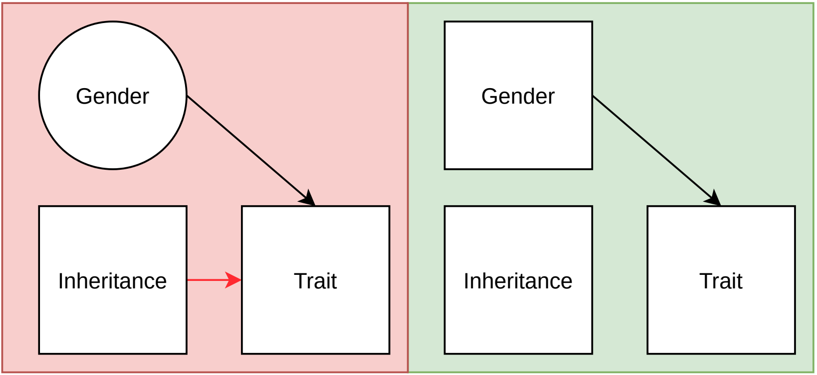
Table: P[Offspring, +] = 0.3, P[Offspring | Parent,+]=0.43. [10]

Conclusion on mixed data : Trait is caused by inheritance.

Conclusion on unmixed data : Trait is caused by gender.

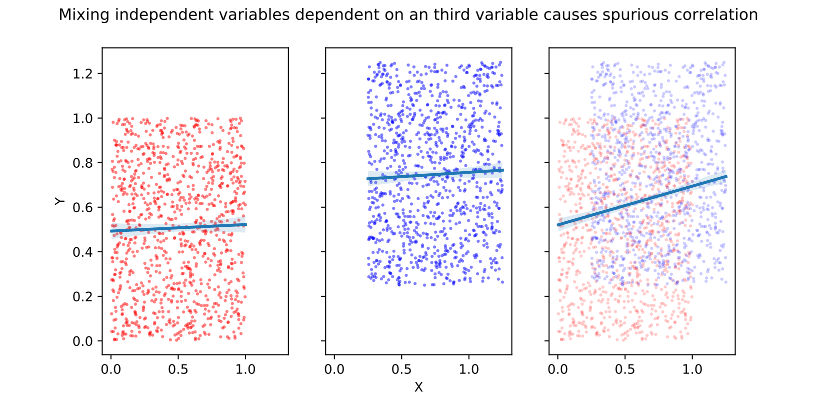
[1] Example adopted from [10]

Statistical Paradox 1 : mixing data (Yule, Pearson)



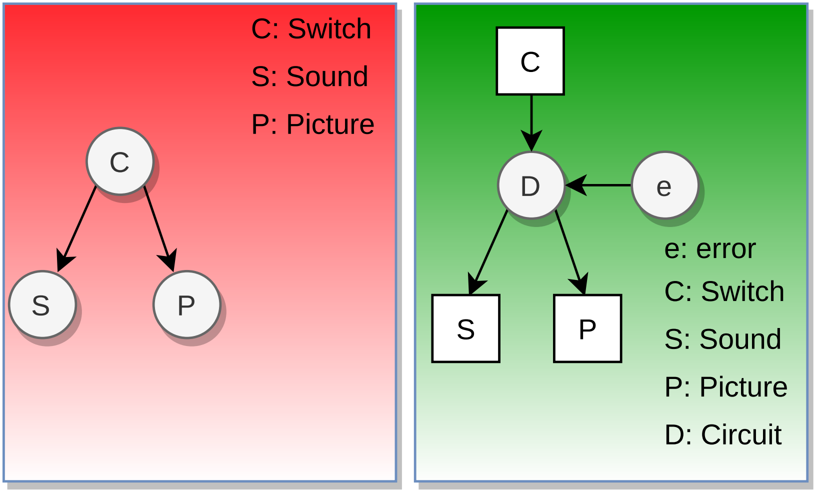
GCM of the previous example. Mixing data leads to confounders, and can break Markov condition, as well as lead to false causal observations.

Statistical Paradox 1 : mixing data (Yule, Pearson)



Karl Pearson showed that mixing 2 independent variables (red, blue) is certain to lead to **non-zero correlation** except in the case where the distributions of the 2 variables, given a third, overlap exactly (e.g. identical mean)

The Markov condition

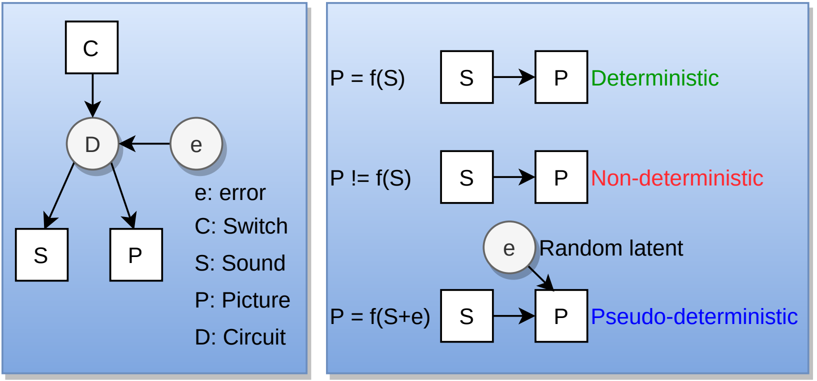


The correct model: introduce latent error and explicit circuit. This example was used to challenge the applicability of the Markov axiom, also leads to introducing the concept of determinism in causality.

The limits of Markovian conditions

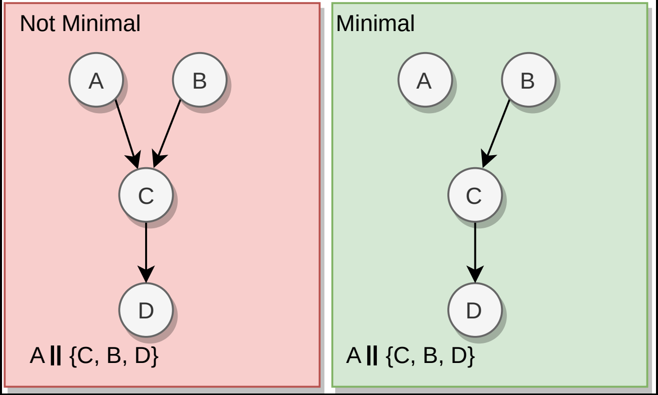
A quantum event E with a state S produces 2 particles (e.g. +, -).
Laws of conservation: particles must have correlated variables
Quantum physics (Bell's inequality):
$\exists$ 'hidden' state S $\Rightarrow$ it must be non-local (breaking Markov)
Markovian causal relations exist at **macro** scale only.

Determinism



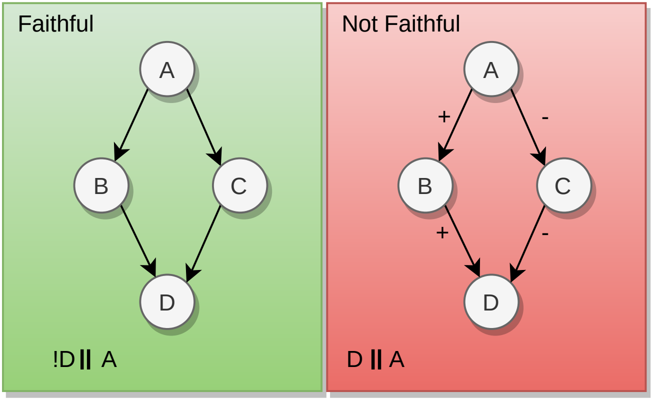
A causal relation can be (non)-deterministic. A indeterministic GCM that can be made deterministic with the addition of latent random variables is pseudo-indeterministic. More formally: P = f(g(S,e), d) where e is the random effect on S, d is the observation error.

Minimality



Minimality: If a GCM G with probability distribution P has a subgraph that is Markovian, then G is not minimal.

Faithfulness



Faithfulness: If the probability space P has A, D independent, yet there is a path in GCM G, then G and P are not faithful. Example: 2 linear effects that cancel each other out. (Kendall, Simpson)

Paradoxically causal

Simpson's paradox/reversal: Given a positive effect in both subpopulations, observe a negative effect in the joint population.

Male	E	!E	Recovery	**Female**	E	!E	Recovery	**Combined**	E	E!	Recovery
Drug	18	12	60%		2	8	20%		20	20	50%
!Drug	7	3	70%		9	21	30%		16	24	40%

Table: Illustration of Simpson's paradox, where E indicates 'Effect', e.g. survival. , [7] The paradox can lead to breaking the faithfulness condition.

Why does this reversal happen?

- Simpson's reversal is the consequence of gender inducing frequency shift in subpopulations (in this example men take the drug more often)
- The numerical effect is sound (algebraic)
- The paradox follows from our mismatched perception and interpretation.
- $P(E|C) > P(E|!C)$; C causes E ?
- Exercise for the reader: Can you have 2 positives that combine into a neutral effect?

[2] https://ftp.cs.ucla.edu/pub/sta_ser/r414.pdf

Paradoxically Causal

Simpson's paradox/reversal: Given a negative effect in both subpopulations, observe a **positive** effect in the joint population.

Male	E	!E	Recovery	**Female**	E	!E	Recovery	**Combined**	E	E!	Recovery
Drug	18	12	60%		2	8	20%		20	20	50%
!Drug	7	3	70%		9	21	30%		16	24	40%

Table: Illustration of Simpson's paradox, where E indicates 'Effect', e.g. survival. , [7] The paradox can lead to breaking the faithfulness condition.

Resolving the paradox by differentiating between evidential and interventional causality

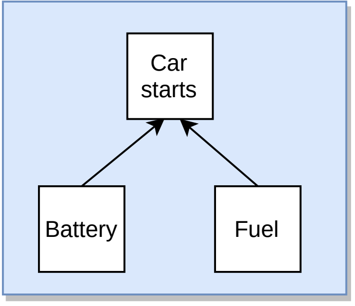
- $P(E|C) > P(E|\neg C)$: Does **not** mean : C increases/causes E
- $P(E|C) > P(E|\neg C)$: Means: C is **evidence** for E, but can be subject to common causes.
- $P(E|do(C)) > P(E|\neg do(C))$: C causes E (to improve).

[3] https://ftp.cs.ucla.edu/pub/sta_ser/r414.pdf

Common Effect Conditioning

Independent variables can become dependent on a common effect

- P(Battery=empty, Fuel=empty) = P(Battery=empty)P(Fuel=empty)
- P(Battery, Fuel | Car !Start) ?
- Empty fuel tank in combination with starting car => battery cannot be empty, independence is replaced by conditioning on common effect.



Causal Discovery Algorithms

Should be general, scalable, robust, efficient, converging

- Constraint based
  - Statistical independence tests, general
  - Cannot solve 2 variable case, large sample size needed
- Scoring Functions
  - Require a model of causality, can solve 2-var problem
  - Specific assumptions needed
  - Generalized Scoring Functions[4]

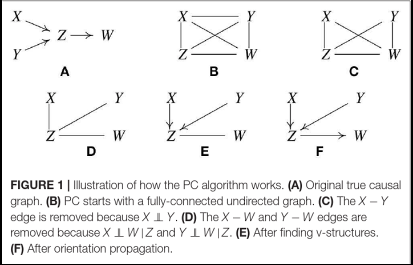
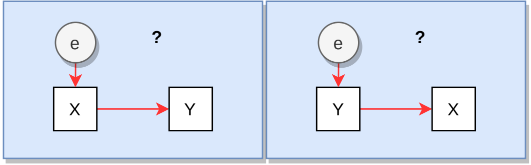


FIGURE 1 | Illustration of how the PC algorithm works. **(A)** Original true causal graph. **(B)** PC starts with a fully-connected undirected graph. **(C)** The $X - Y$ edge is removed because $X \perp\!\!\!\perp Y$. **(D)** The $X - W$ and $Y - W$ edges are removed because $X \perp\!\!\!\perp W \mid Z$ and $Y \perp\!\!\!\perp W \mid Z$. **(E)** After finding v-structures. **(F)** After orientation propagation.

Illustration of PC (constraint based), using (conditional) independence tests to retrieve the causal structure. (source [2])

Finding Causal Direction

- Given only 2 variables, can you find the causal direction ?
- Let's assume $Y = bX + \epsilon, X \perp\!\!\!\perp \epsilon$



Which is the cause and which the effect, and how do you find out?

Finding Causal Direction

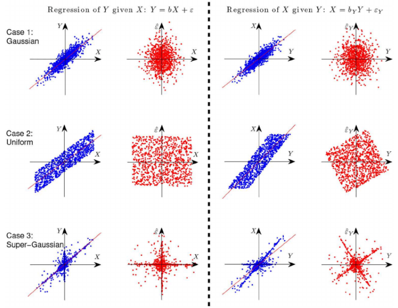


FIGURE 3 | Illustration of causal asymmetry between two variables with linear relations. The causal relation is $X \rightarrow Y$. From top to bottom: X and E both follow the Gaussian distribution (case 1), uniform distribution (case 2), and Laplace distribution (case 3). The two columns on the left show the scatter plot of X and Y and that of X and the regression residual for regressing Y on X, and the two columns on the right correspond to regressing X on Y.

Illustration why, when at least one of X or ϵ is non-Gaussian, you can always recover direction of causality in continuous variables [2]. The regression error (column 2,4) is not independent for the incorrect direction (ICA theory). Figure reproduced from [2].

[4] If the data has no noise, can you still find direction between cause/event?

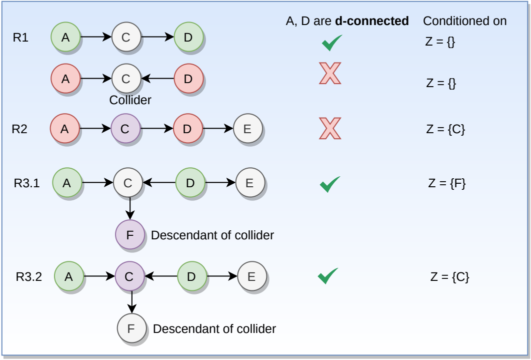
D-connectedness

D-connected : conditional independence on a Markov GCM

X, Y are d-connected, given W, iff

- undirected path U : X..Y has no collider (not in W)
- U has no vertex in W
- U has a collider in W

X, Y | W d-separated : P(X,Y|W) = P(X|W)P(Y|W)



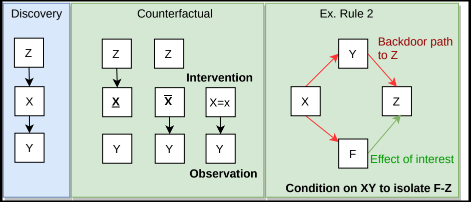
Directional connectedness in a GCM allows you to read the model to query which events can be related given a set of known events.

Do-operator (Pearl), or Causal Calculus

Intervention: P(cancer | Smoker = yes)?
Causal calculus can avoid expensive or unethical double blind experiments.

$do(x) \Rightarrow X = x$

- **Ignore observation:**
  - $p(y|do(x), z, w) = p(y|do(x), w)$ if $(Y \perp\!\!\!\perp Z|X, W)G_{\overline{X}}$
- **Action to observation:**
  - $p(y|do(x), do(z), w) = p(y|do(x), z, w)$ if $(Y \perp\!\!\!\perp Z|X, W)G_{\overline{X},\underline{Z}}$
- **Ignore action:**
  - $p(y|do(x), do(z), w) = p(y|do(x), w)$ if $(Y \perp\!\!\!\perp Z|X, W, G_{\overline{X},\overline{\neg(z\rightarrow w)}})$
- **d-connectedness** is essential to resolve $X \perp\!\!\!\perp Y|W$ queries in GCM



Causal calculus enables expressing **observations**, **actions|interventions** and their probabilities in the context of a GCM. Intervention graphs: $G_{\underline{X}}$(V,E'): $E' = E \setminus \{(\mathbf{x}, k)\forall k \in V\}$, $G_{\overline{X}}$(V,E'): $E' = E \setminus \{(k, \mathbf{x})\forall k \in V\}$. Intuition: if I control X, no other cause of X is relevant.

[5] Adapted from https://www.cs.ubc.ca/labs/lci/mlrg/slides/doCalc.pdf

Research frontier of open problems

- **Time series**[5]
- Heterogeneity/Non-stationary: The causal process changes over time or datasets [11]
- **Mixed** data [8]
- **Measurement error**: Cause X, effect Y, $Y = f(g(X, \epsilon), \delta)$ What is the impact of not knowing ϵ, δ? [12]
- **Selection bias** (Simpson, Berkson paradox)
- **Missing data**
- **Causal learning**: e.g., in semi-supervised learning, when features are causes of the outcome, unlabelled data will not necessarily improve performance [9]. Similarly, in transfer learning it is of interest to transfer the causal relation rather than the association/correlation function.
- **Scaling** (number of variables): number of acyclic paths in G(V,E) of length k = k! Causal algorithms are bound (for now) by non-linear relations to the number of variables (dimensions).

[6] Intuitive difference: Machine learning : learn transformation of data into output, Causality: learn nature of the process generating the data

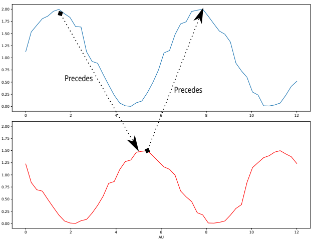
Temporal Causality $X \rightarrow Y \Rightarrow$ X ‘before‘ Y?

Granger ‘Causality’

- if cause (X) precedes effect (Y)
- if cause has unique information determining future values of effect
- Information of system at time t : I_t
- $P[Y_{t+1} \sim Z | I_t] \neq P[Y_{t+1} \sim Z | (I \setminus X)_t]$

Heidegger: Interpretable Causal Discovery[5]

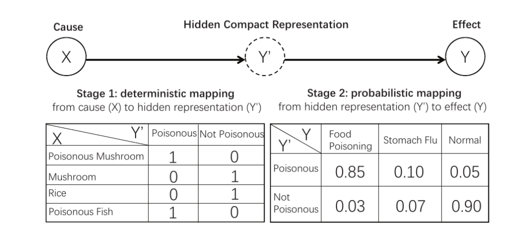
- Causal Profile Discovery
- Assume cause X, effect Y
- Which pattern (/time) should X follow to maximize effect Y?
- Applications: dementia (age), cancer (drugs), ...



Unravelling temporal statistics is a hard problem confounded (pun intended) by our ill-defined concepts of time(span), causes, effects. $sin((x - \mathbf{10}) + \epsilon$ versus $sin((x) + \epsilon$

Causality on non-continuous data

- Causal direction finding in continuous data is easier compared to discrete, categorical or mixed data. (defining additive noise for e.g. categories is non-trivial)
- Idea : Learn a hidden compact representation (HCR)
- HCR aims to encode low cardinality 'true' causal variable
- Learn $M : X \to Y' \to Y, M' : Y \to X' \to X$
- BIC score (M, M') decides which causal direction is likely.


Cause
Hidden Compact Representation
Effect
X
Y'
Y
Stage 1: deterministic mapping
from cause (X) to hidden representation (Y')
Stage 2: probabilistic mapping
from hidden representation (Y') to effect (Y)

X \ Y'	Poisonous	Not Poisonous
Poisonous Mushroom	1	0
Mushroom	0	1
Rice	0	1
Poisonous Fish	1	0

Y' \ Y	Food Poisoning	Stomach Flu	Normal
Poisonous	0.85	0.10	0.05
Not Poisonous	0.03	0.07	0.90

Figure 1: Food Poisoning: A Hidden Compact Representation Example in Real World.

Learning an encoding from categorical space enables more powerful continuous space causality inference algorithms.[1]

Explaining image classification with causal learning

- Cars are often seen in images with bicycles (correlation is high)
- Presence of car can be misleading 'explainer' for classification of bicycle
- Does the concept 'car' explain classification? : p(class=bicycle | do(car))

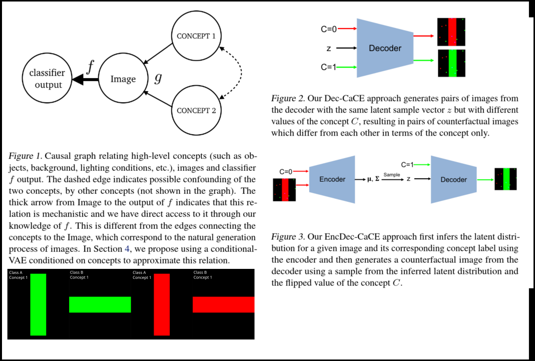


Figure 1. Causal graph relating high-level concepts (such as objects, background, lighting conditions, etc.), images and classifier f output. The dashed edge indicates possible confounding of the two concepts, by other concepts (not shown in the graph). The thick arrow from Image to the output of f indicates that this relation is mechanistic and we have direct access to it through our knowledge of f. This is different from the edges connecting the concepts to the Image, which correspond to the natural generation process of images. In Section 4, we propose using a conditional-VAE conditioned on concepts to approximate this relation.

Figure 2. Our Dec-CaCE approach generates pairs of images from the decoder with the same latent sample vector z but with different values of the concept C, resulting in pairs of counterfactual images which differ from each other in terms of the concept only.

Figure 3. Our EncDec-CaCE approach first infers the latent distribution for a given image and its corresponding concept label using the encoder and then generates a counterfactual image from the decoder using a sample from the inferred latent distribution and the flipped value of the concept C.

An example of recent work [3], images courtesy of original work. A network can learn to generate counterfactuals and learn which concepts **do** explain the classification. Orientation of the white rectangle is the class, the concept is the color (green,red)

Fin

Causality is a human intuition essential to understanding a complex world that can be formalized (it took >100 years to do so)

Takeaway message(s)

- Understand your data
- Understand the axioms under which a given causal model is valid
- Understand and formulate your query
- Have fun :)

Try it yourself

https://www.ccd.pitt.edu/

Acknowledgements

I would like to thank Prof. Ghassan Hamarneh, Dr. Weina Jin, Dr. Sieun Lee, Prof. Martin Ester, Darren Sutton, and all the members of my lab for the many discussions that made it possible for me to begin to understand this complex topic, and for their support.

References I

Ruichu Cai et al. “Causal discovery from discrete data using hidden compact representation”. In: **Advances in neural information processing systems** 31 (2018), pp. 2666–2674.

Clark Glymour, Kun Zhang, and Peter Spirtes. “Review of causal discovery methods based on graphical models”. In: **Frontiers in genetics** 10 (2019), p. 524.

Yash Goyal et al. “Explaining classifiers with causal concept effect (cace)”. In: **arXiv preprint arXiv:1907.07165** (2019).

Biwei Huang et al. “Generalized score functions for causal discovery”. In: **Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**. 2018, pp. 1551–1560.

Mehrdad Mansouri et al. “Heidegger: Interpretable Temporal Causal Discovery”. In: **Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining**. 2020, pp. 1688–1696.

Judea Pearl. **Causality**. Cambridge university press, 2009.

References II

- Judea Pearl. "Comment: understanding Simpsons paradox". In: **The American Statistician** 68.1 (2014), pp. 8–13.
- Vineet K Raghu et al. "Comparison of strategies for scalable causal discovery of latent variable models from mixed data". In: **International journal of data science and analytics** 6.1 (2018), pp. 33–45.
- Bernhard Schölkopf et al. "On causal and anticausal learning". In: **arXiv preprint arXiv:1206.6471** (2012).
- Peter Spirtes et al. **Causation, prediction, and search**. MIT press, 2000.
- Kun Zhang et al. "Causal discovery from nonstationary/heterogeneous data: Skeleton estimation and orientation determination". In: **IJCAI: Proceedings of the Conference**. Vol. 2017. NIH Public Access. 2017, p. 1347.
- Kun Zhang et al. "Causal discovery in the presence of measurement error: Identifiability conditions". In: **arXiv preprint arXiv:1706.03768** (2017).