TECHNISCHE UNIVERSITÄT DARMSTADT

AIML Lab

Winter Semester 2025/26 Lecture

Causality for AI & ML

“Causal Abstractions”

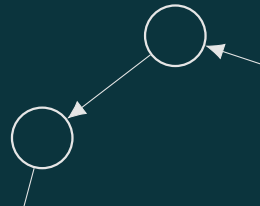
Prof. Dr. Kristian Kersting

Moritz Willig — Today’s speaker

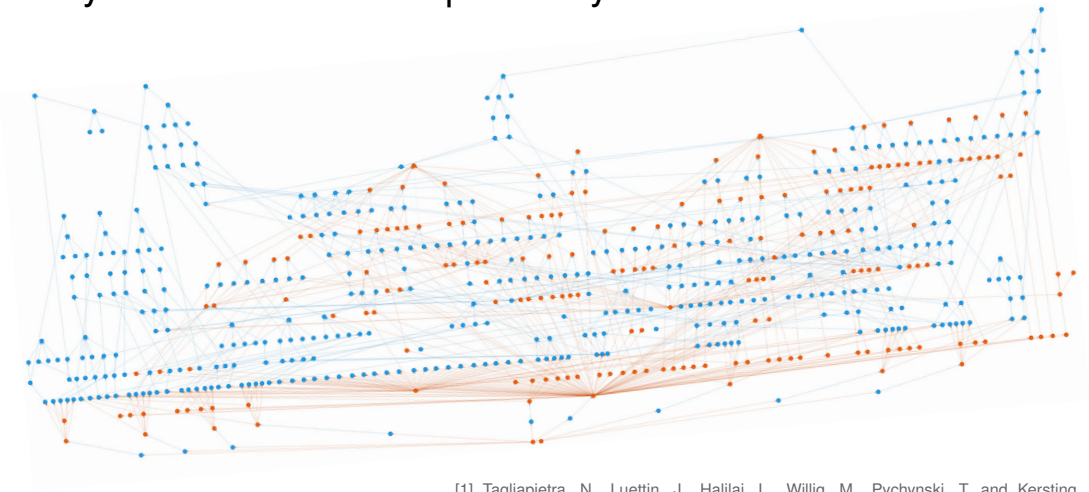
Tim Woydt

Florian Busch

Matej Zečević

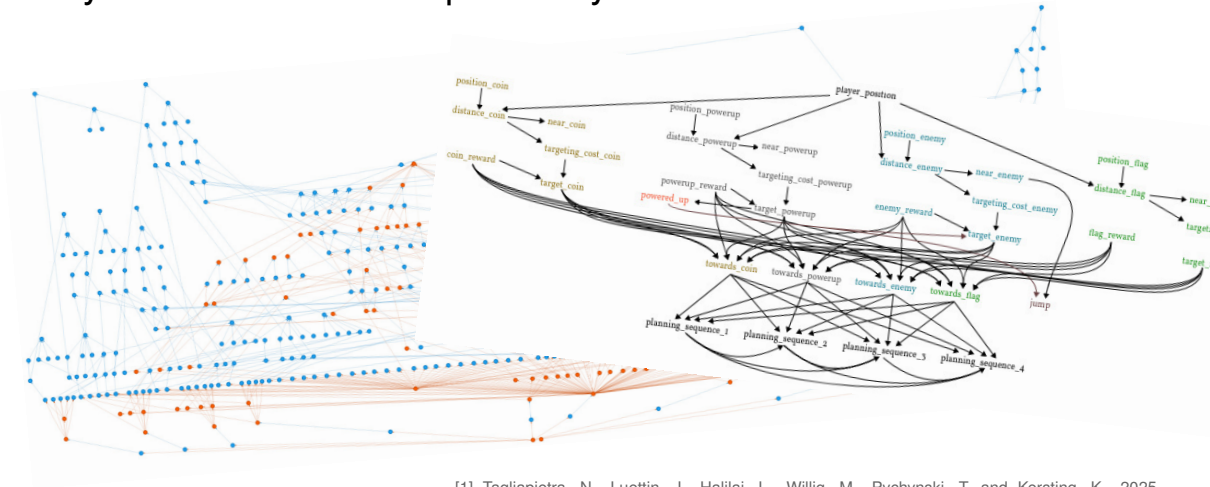


Why Abstractions?: Interpretability



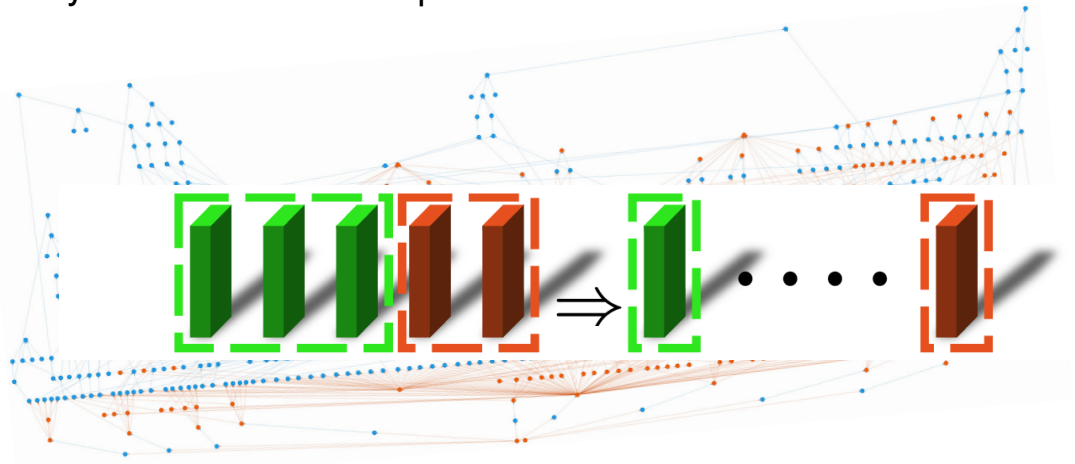
[1] Tagliapietra, N., Luettin, J., Halilaj, L., Willig, M., Pychynski, T. and Kersting, K., 2025. CausalMan: A physics-based simulator for large-scale causality. arXiv preprint arXiv:2502.12707.
[2] Willig, M., Zečević, M., Dhami, D. and Kersting, K., 2023. Do not marginalize mechanisms, rather consolidate!. Advances in Neural Information Processing Systems, 36, pp.60947-60965.

Why Abstractions?: Interpretability



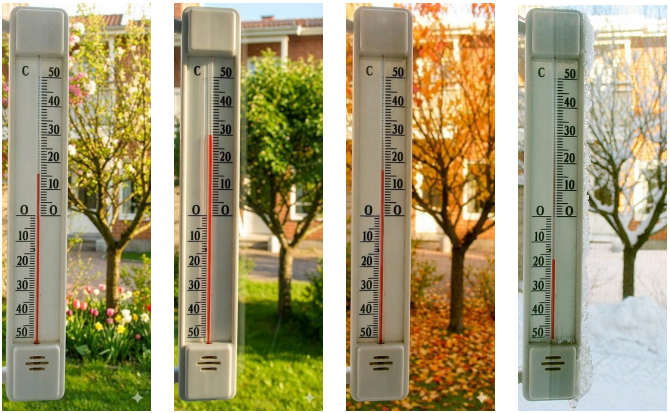
[1] Tagliapietra, N., Luettin, J., Halilaj, L., Willig, M., Pychynski, T. and Kersting, K., 2025. CausalMan: A physics-based simulator for large-scale causality. arXiv preprint arXiv:2502.12707.
[2] Willig, M., Zečević, M., Dhami, D. and Kersting, K., 2023. Do not marginalize mechanisms, rather consolidate!. Advances in Neural Information Processing Systems, 36, pp.60947-60965.

Why Abstractions?: Simplification



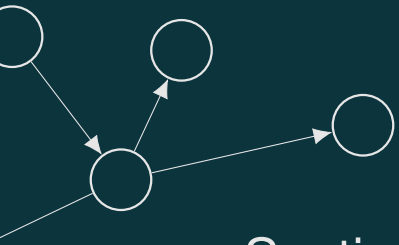
[1] Tagliapietra, N., Luettin, J., Halilaj, L., Willig, M., Pychynski, T. and Kersting, K., 2025. CausalMan: A physics-based simulator for large-scale causality. arXiv preprint arXiv:2502.12707.
[2] Willig, M., Zečević, M., Dhami, D. and Kersting, K., 2023. Do not marginalize mechanisms, rather consolidate!. Advances in Neural Information Processing Systems, 36, pp.60947-60965.

Why Abstractions?: Semantics



Variables are the 373×1024 pixels of the image.
$\rightarrow$ Is pixel $[23, 825]$ a useful variable?"

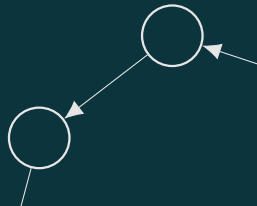
Image: https://en.wikipedia.org/wiki/File:Pakkanen.jpg



Section

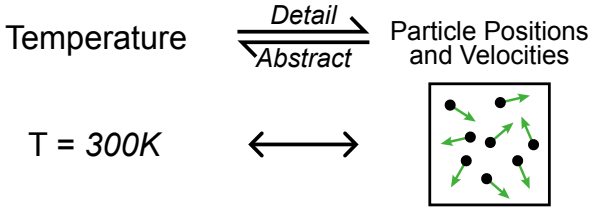
1

Granularity of (Causal) Models



Levels of Granularity

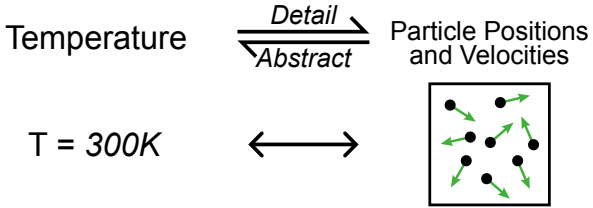
Systems can be expressed at different levels of detail:



Choose the right level of detail for the right purpose.

Levels of Granularity

Systems can be expressed at different levels of detail:

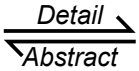


Choose the right level of detail for the right purpose.

→ We want to build causal models that are useful to us.

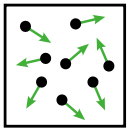
What is a Variable?

Temperature



Particle Positions and Velocities

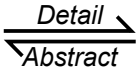
T = *300K*



What is a variable anyway?

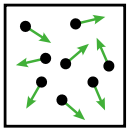
What is a Variable?

Temperature



Particle Positions and Velocities

T = *300K*



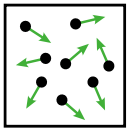
What is a variable anyway?

- Variables are abstract concepts in our models.

They might be observed or latent, real or purely hypothetical.

What is a Variable?

Temperature ⇌ (*Detail* / *Abstract*) Particle Positions and Velocities

T = *300K* ⟷ 

What is a variable anyway?

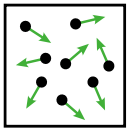
- Variables are abstract concepts in our models.

They might be observed or latent, real or purely hypothetical.

"An **observable** is a (physical) property or quantity that can be measured" - Wikipedia

What is a Variable?

Temperature ⇌ (Detail / Abstract) Particle Positions and Velocities

T = *300K* ⟷ 

What is a variable anyway?

- Variables are abstract concepts in our models.

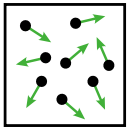
They might be observed or latent, real or purely hypothetical.

"An **observable** is a (physical) property or quantity that can be measured" - Wikipedia

- We often relate observations/measurements to our variables.

Observations

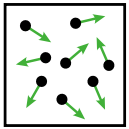
Temperature ⇌ (*Detail* / *Abstract*) Particle Positions and Velocities

T = *300K* ⟷ 

“But we can measure temperature. Does this mean this is the ‘true’ causal model?”

Observations

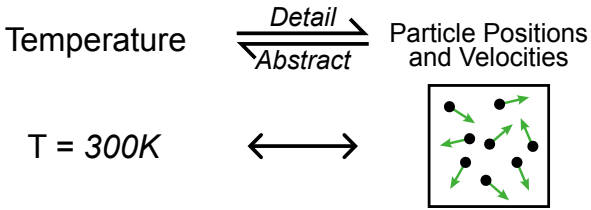
Temperature ⇌ (*Detail* / *Abstract*) Particle Positions and Velocities

T = *300K* ⟷ 

“But we can measure temperature. Does this mean this is the ‘true’ causal model?”

Every measurement quantifies some aspect of reality, but our measurements might not live on the same level of abstraction as our models.

Observations



"But we can measure temperature. Does this mean this is the 'true' causal model?"

Every measurement quantifies some aspect of reality, but our measurements might not live on the same level of abstraction as our models.

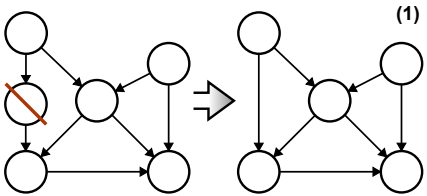
Objective of Designing Causal Abstractions: Define *transformations* that translate models of a given granularity into models of the *desired level of abstraction*.

Transforming Models

We might want to simplify models to

- … reduce model complexity
- … reduce computational effort
- … obtain high-level insights

Transforming Models



We might want to simplify models to

- … reduce model complexity
- … reduce computational effort
- … obtain high-level insights

Typical operations of abstractions:

(1) **Marginalization** of variables

Transforming Models

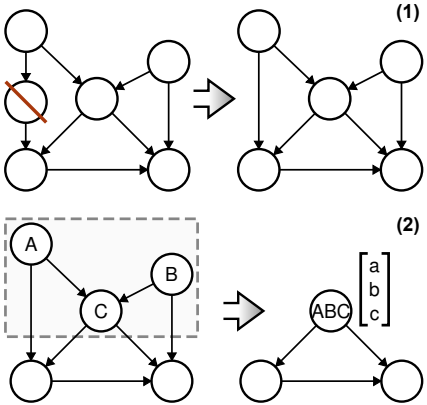
We might want to simplify models to

- … reduce model complexity
- … reduce computational effort
- … obtain high-level insights

Typical operations of abstractions:

(1) **Marginalization** of variables

(2) **Grouping** of variables



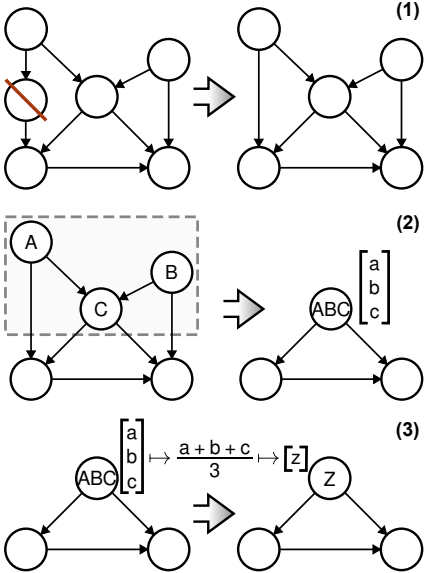
Transforming Models

We might want to simplify models to

- … reduce model complexity
- … reduce computational effort
- … obtain high-level insights

Typical operations of abstractions:

(1) **Marginalization** of variables
(2) **Grouping** of variables
(3) **Transformation** of variables



Transforming Models

We might want to simplify models to

- … reduce model complexity
- … reduce computational effort
- … obtain high-level insights

Typical operations of abstractions:

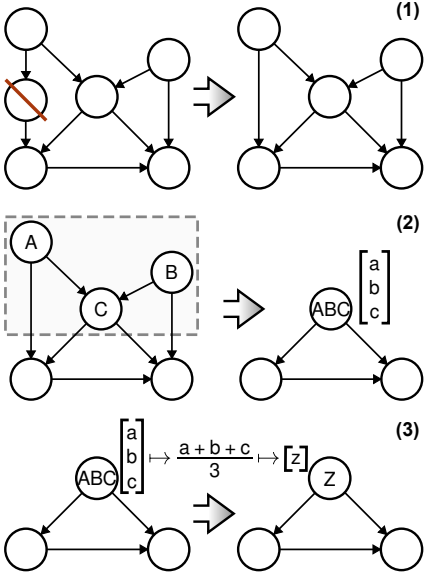
(1) **Marginalization** of variables

(2) **Grouping** of variables

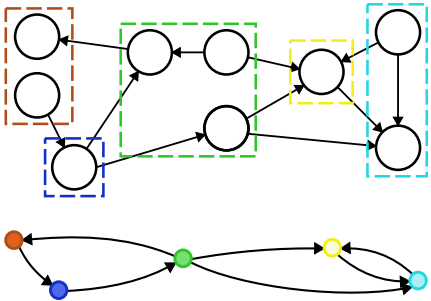
(3) **Transformation** of variables

Making models more *detailed* is also possible but usually requires further observables.

→ Simplifying models is often ‘easier’.



Grouping of Variables - Implications



- Grouping variables can induce cycles. ($\blacksquare \to \blacksquare \to \blacksquare; \blacksquare \to \blacksquare; \ldots$)
- Paths in the grouped graph do not need to be closed on the lower level. (Consider $\blacksquare \to \blacksquare \to \blacksquare$).

Forré, Patrick, and Joris M. Mooij. "Constraint-based Causal Discovery for Non-Linear Structural Causal Models with Cycles and Latent Confounders." Wahl, J., Ninad, U. and Runge, J., 2024. Foundations of causal discovery on groups of variables. Journal of Causal Inference, 12(1), p.20230041.

Transforming Models

$$\text{Diet} \longrightarrow \text{TC} \xrightarrow[+\,?]{-\,?} \text{HD}$$

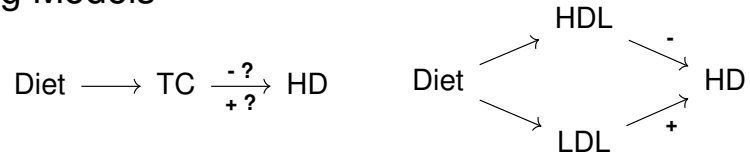
Total cholesterol in the blood (TC) was considered an indicator for a hearth disease (HD).

Conflicting evidence:

- In some studies TC lowered HD.
- In some studies TC increased HD.

Spirtes, Peter, and Richard Scheines. "Causal inference of ambiguous manipulations." Philosophy of Science 71.5 (2004): 833-845.
Rubenstein, Paul K., et al. "Causal Consistency of Structural Equation Models". In Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (2017).

Transforming Models



Total cholesterol in the blood (TC) was considered an indicator for a hearth disease (HD).

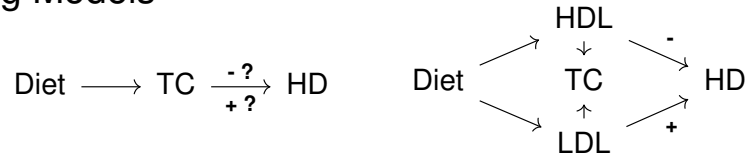
Conflicting evidence:

- In some studies TC lowered HD.
- In some studies TC increased HD.

Identified low-density lipoproteins (LDL) and high-density lipoproteins (HDL).

Spirtes, Peter, and Richard Scheines. "Causal inference of ambiguous manipulations." Philosophy of Science 71.5 (2004): 833-845.
Rubenstein, Paul K., et al. "Causal Consistency of Structural Equation Models". In Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (2017).

Transforming Models



Total cholesterol in the blood (TC) was considered an indicator for a hearth disease (HD).

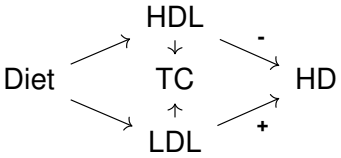
Conflicting evidence:

- In some studies TC lowered HD.
- In some studies TC increased HD.

Identified low-density lipoproteins (LDL) and high-density lipoproteins (HDL).
Diets that raise LDL or HDL both increase TC but have different effects on HD.

Spirtes, Peter, and Richard Scheines. "Causal inference of ambiguous manipulations." Philosophy of Science 71.5 (2004): 833-845.
Rubenstein, Paul K., et al. "Causal Consistency of Structural Equation Models". In Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (2017).

Transforming Models



Total cholesterol in the blood (TC) was considered an indicator for a hearth disease (HD).

Conflicting evidence:

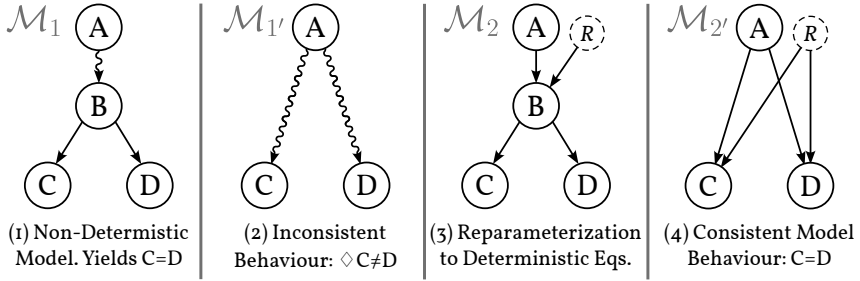
- In some studies TC lowered HD.
- In some studies TC increased HD.

The initial model was too coarse to model the causal effect.

Identified low-density lipoproteins (LDL) and high-density lipoproteins (HDL).
Diets that raise LDL or HDL both increase TC but have different effects on HD.

Spirtes, Peter, and Richard Scheines. "Causal inference of ambiguous manipulations." Philosophy of Science 71.5 (2004): 833-845.
Rubenstein, Paul K., et al. "Causal Consistency of Structural Equation Models". In Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence (2017).

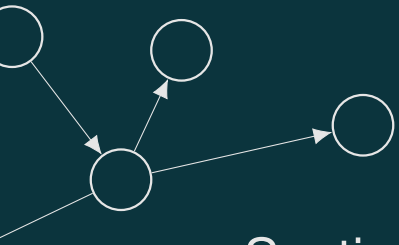
Note: Reparameterization of Non-deterministic Equations



Marginalizing variables in SCM with non-deterministic equations might lead to inconsistent behavior ($\mathcal{M}_1 \to \mathcal{M}_2$).

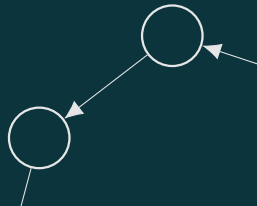
First reparameterize to deterministic equations and a random variable ($\mathcal{M}_1 \to \mathcal{M}_2$) and then refactor $\mathcal{M}_2 \to \mathcal{M}_2'$.

Willig, M., Zečević, M., Dhami, D. and Kersting, K., 2023. Do not marginalize mechanisms, rather consolidate!. Advances in Neural Information Processing Systems, 36, pp.60947-60965.



Section 2

Causal Abstractions



Causal Abstractions

Abstracting transforms usually abstract away information. *You don't say...*

$\rightarrow$ Irreversible process

$\rightarrow$ Loss of information

Causal Abstractions

Abstracting transforms usually abstract away information. *You don't say...*

$\rightarrow$ Irreversible process

$\rightarrow$ Loss of information

Still, abstractions are not arbitrary transformations.

We want to enforce **consistency** between the low-level model (M_L) and the high-level model (M_H):

1. **Consistency of variable mappings**:
   - There exists a **variable map** τ such that, $\tau(\mathbf{X}_L) = \mathbf{X}_H$.
2. **Consistency of interventions**:
   - There exists an **intervention map** ω such that, $\omega(\mathbf{I}_L) = \mathbf{I}_H$.

Rubenstein, P.K., Weichwald, S., Bongers, S., Mooij, J.M., Janzing, D., Grosse-Wentrup, M. and Schölkopf, B.. "Causal Consistency of Structural Equation Models". In Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, 2017

Consistency of Abstractions

Consistency of Variable Maps

Every variable configuration $\mathbf{x}_H$ must be explainable by some configuration in $\mathbf{x}_L$. Therefore, τ must be surjective such that:

Surjectiveness of τ: $\forall \mathbf{x}_H \in \mathcal{X}_H . \exists \mathbf{x}_L \in \mathcal{X}_L . \tau(\mathbf{x}_L) = \mathbf{x}_H$

Consistency of Intervention Maps

The abstraction of the effect must equal the effect of the abstraction:

Commutativity under ω: $\tau(M_L|I_L) = \tau(M_L)|\omega(I_L)$

Rubenstein, P.K., Weichwald, S., Bongers, S., Mooij, J.M., Janzing, D., Grosse-Wentrup, M. and Schölkopf, B.. "Causal Consistency of Structural Equation Models". In Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, 2017

(τ-ω)-Abstraction

An (exact) (τ-ω)-Abstraction is a pair of functions (τ, ω), where

1. $\tau : \mathbf{X}_L \to \mathbf{X}_H$ is a surjective map,
2. ω is a function $\omega : \mathcal{I}_L \to \mathcal{I}_H$ between sets of interventions,
3. and the following diagram commutes:

$$\begin{array}{ccc} \mathbb{P}_X & \xrightarrow{\mathrm{do}(i)} & \mathbb{P}_X^{\mathrm{do}(i)} \\ \downarrow{\scriptstyle \tau} & & \downarrow{\scriptstyle \tau} \\ \mathbb{P}_Y & \xrightarrow{\mathrm{do}(\omega(i))} & \mathbb{P}_Y^{\mathrm{do}(\omega(i))} \end{array}$$

Rubenstein, P.K., Weichwald, S., Bongers, S., Mooij, J.M., Janzing, D., Grosse-Wentrup, M. and Schölkopf, B.. "Causal Consistency of Structural Equation Models". In Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence, 2017

Types of abstractions

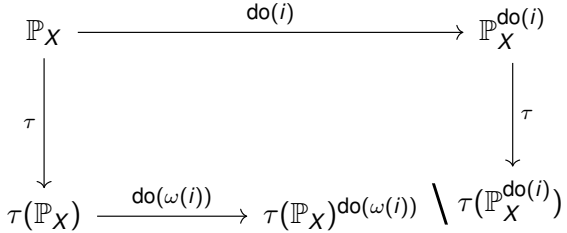
More refined notions of $\tau - \omega$-abstractions are given in Beckers et al., 2019:

- **Uniform**: Interventions should not only hold under a specific exogenous probability, but under all possible ones.
- **Strong**: All sets of interventions should be allowed. (Prevents cherry picking of interventions to hide inconsistencies of the mapping.)
- **Constructive**: High-level variables must be formed by grouping low-level variables into disjoint sets (clusters) with no overlap.
- **Natural Intervention Mapping**: τ tells us how to translate low- to high-level values. The ω mapping should reflect that in its translation of intervention values.

Beckers, Sander, and Joseph Y. Halpern. "Abstracting causal models." Proceedings of the aaai conference on artificial intelligence. Vol. 33. No. 01. 2019.

Approximate Abstractions

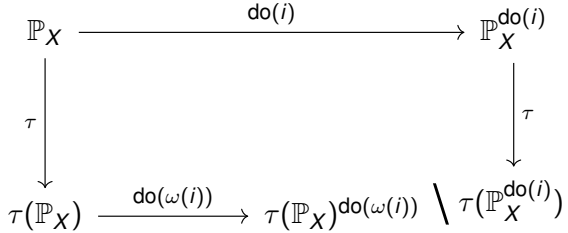
Moving from perfect models to the ‘messy’ reality:



Beckers, Sander, Frederick Eberhardt, and Joseph Y. Halpern. "Approximate causal abstractions." Uncertainty in artificial intelligence. PMLR, 2020.

Approximate Abstractions

Moving from perfect models to the ‘messy’ reality:



ϵ-Abstractions: Allow for slight deviations in the commutative property:

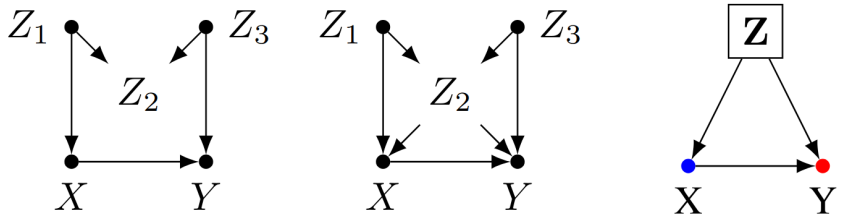
$$\left\| \tau(\mathbb{P}_X)^{\mathrm{do}(\omega(i))}, \tau(\mathbb{P}_X^{\mathrm{do}(i)}) \right\| < \epsilon$$

Beckers, Sander, Frederick Eberhardt, and Joseph Y. Halpern. "Approximate causal abstractions." Uncertainty in artificial intelligence. PMLR, 2020.

Cluster DAGS (C-DAGs)

"We often only know genes in cluster A regulate genes in cluster B, but do not know the exact 'per-variable' effects."

$\rightarrow$ Can we still compute causal effects in these graphs (e.g. between the groups)?



Anand, T.V., Ribeiro, A.H., Tian, J. and Bareinboim, E., 2023, June. Causal effect identification in cluster dags. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 10, pp. 12172-12179).

Definition of C-DAGs

A C-DAG is a graph where nodes are clusters rather than single variables.

Cluster DAG (C-DAG)

A **C-DAG** $\mathcal{G}_{\mathbf{C}}(\mathbf{C}, \mathbf{E}_{\mathbf{C}})$ is a graph, given some partition $\mathbf{C} = \{\mathbf{C}_1, .., \mathbf{C}_k\}$ of $\mathbf{X}$ and

1. A [bidirected] edge $\mathbf{C}_i \to \mathbf{C}_j$ [$\mathbf{C}_i \dashleftarrow\dashrightarrow \mathbf{C}_j$] is in $\mathbf{E}_{\mathbf{C}}$ if there exists some $V_i \in \mathbf{C}_i$ and $X_j \in \mathbf{C}_j$ such that $X_i \in pa(X_j)$ [the edge $\mathbf{X}_i \dashleftarrow\dashrightarrow \mathbf{X}_j$] is in $\mathcal{G}$.
2. $\mathcal{G}_{\mathbf{C}}(\mathbf{C}, \mathbf{E}_{\mathbf{C}})$ contains no cycles. (The partition is *admissible*).

Compatibility: A standard DAG is *compatible* with a C-DAG if:

1. Its edges respect the cluster edges $\mathbf{E}_{\mathbf{C}}$ (i.e., if there is no edge from Cluster A to Cluster B, there is no edge from any variable $a \in A$ to any variable $b \in B$).

Anand, T.V., Ribeiro, A.H., Tian, J. and Bareinboim, E., 2023, June. Causal effect identification in cluster dags. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 10, pp. 12172-12179).

Identification in C-DAGs

An effect is *identifiable in a C-DAG* only if:

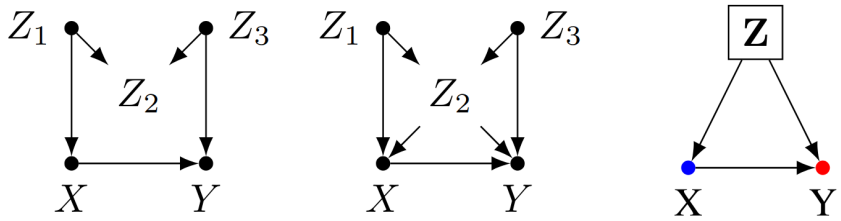
- The effect is identifiable in every possible DAG compatible with the cluster structure.
- The resulting estimand (the formula) is identical for all those DAGs. (Otherwise use bounding)

Note: Even if one compatible DAG allows for a backdoor path that cannot be blocked, the effect is non-identifiable…

Requires further assumptions/restrictions.

Anand, T.V., Ribeiro, A.H., Tian, J. and Bareinboim, E., 2023, June. Causal effect identification in cluster dags. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 10, pp. 12172-12179).

Compatability

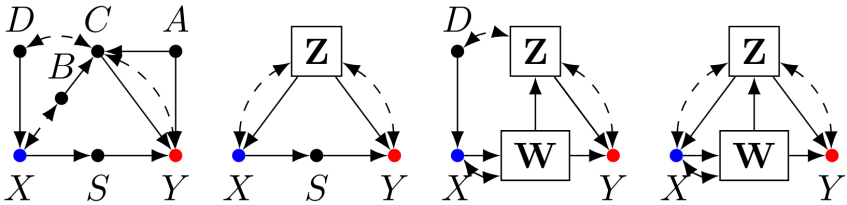


Given the two DAGs on the left, the corresponding C-DAG on the right is formed by the partition $\mathbf{C} = \{\{X\}, \{Y\}, \{Z_1, Z_2, Z_3\}\}$.

Note: Further DAGS with $Z_1 \leftarrow Z_2 \rightarrow Z_3$ might be compatible to the emerging C-DAG.

Anand, T.V., Ribeiro, A.H., Tian, J. and Bareinboim, E., 2023, June. Causal effect identification in cluster dags. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 10, pp. 12172-12179).

Identifiability



(a) (b) (c) (d)

The effect X on Y in (a) is identifiable via back-door (Adj=$\{B, D\}$) and front-door adjustment (via S).

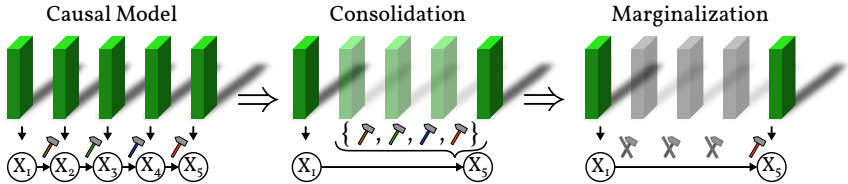
(b) still allows front-door adjustment via S.

(c) is no longer identifiable.

(d) is inadmissible. (Contains a cycle in (X, W, Z)).

Anand, T.V., Ribeiro, A.H., Tian, J. and Bareinboim, E., 2023, June. Causal effect identification in cluster dags. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 10, pp. 12172-12179).

Consolidation - An Intermediate Stage to Marginalization

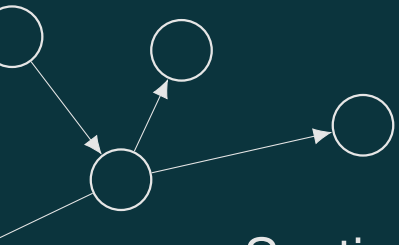


The state of the last domino simply depends on the state of the first, except if we intervene by holding on to any of the dominos along the way.

- **Low-Level SCM**: Detailed model. All variables intervenable.
- **Marginalized SCM**: Some interventions are no longer directly applicable.
- **Consolidated SCM**: Equations explicitly incorporate interventions.

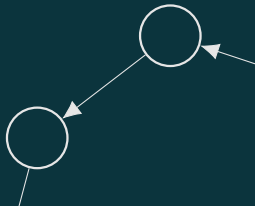
$$X_5 := \begin{cases} c & \textbf{if } do(X_i = c) \in \mathbf{I} \\ X_1 & \textbf{else} \end{cases}$$

Willig, M., Zečević, M., Dhami, D. and Kersting, K., 2023. Do not marginalize mechanisms, rather consolidate!. Advances in Neural Information Processing Systems, 36, pp.60947-60965.



Section 3

Learning (with) Causal Abstractions



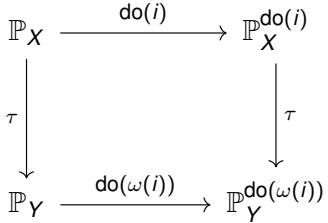
Learning Causal Abstractions

Goal: Discover τ and/or the high-level model from low-level data.

Learning Causal Abstractions

Goal: Discover τ and/or the high-level model from low-level data.

Again, consider the commutative diagram:

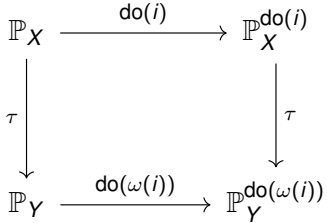


Setting τ to a constant function, e.g, $\tau : \mathbf{X}_L \to \mathbf{0}$ creates a **trivial abstraction**.
$\to$ Formally, a valid abstraction and easily learned by any ML method

Learning Causal Abstractions

Goal: Discover τ and/or the high-level model from low-level data.

Again, consider the commutative diagram:



Setting τ to a constant function, e.g, $\tau : \mathbf{X}_L \to \mathbf{0}$ creates a **trivial abstraction**.

$\to$ Formally, a valid abstraction and easily learned by any ML method

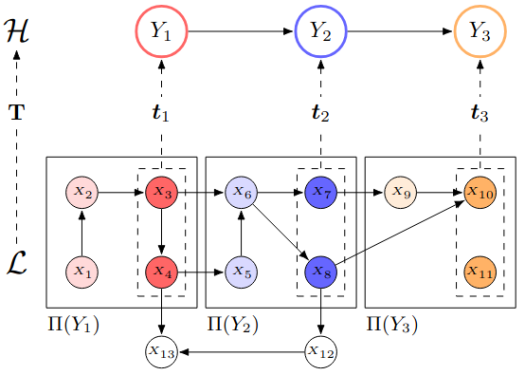
$\to$ Unfortunately uninformative…

We need some constraints/regularization on either τ or the high-level model.

Abs-LiNGAM

Utilize causal abstractions for causal discovery.

1. Learn the abstraction map τ using paired low- and high-level data.
2. Learn the abstract model by abstracting data to the high-level.
3. Infer constraints.
   (here:
   $\blacksquare \not\to \blacksquare$, $\blacksquare \not\to \blacksquare$, $\blacksquare \not\to \blacksquare$, $\blacksquare \not\to \blacksquare$)
4. Discover the low-level model from data and inferred constrains.



Massidda, Riccardo, Sara Magliacane, and Davide Bacciu, 2024. "Learning Causal Abstractions of Linear Structural Causal Models." The 40th Conference on Uncertainty in Artificial Intelligence.

Further Approaches

Further approaches with different assumptions:

- Leverage predictability between time steps and invertability of τ [1].
- Leverage the learning signal of a high-level target variable [2].
- Consider scenarios where different low-level interventions map to the same high-level intervention [3].
- ...

[1] Kumar, A., Gilra, A., Gonzalez-Soto, M., Meunier, A. and Grosse-Wentrup, M., 2023. BundDLe-Net: Neuronal Manifold Learning Meets Behaviour. bioRxiv, pp.2023-08.
[2] Kekić, Armin, Bernhard Schölkopf, and Michel Besserve. "Targeted Reduction of Causal Models." The 40th Conference on Uncertainty in Artificial Intelligence.
, and Elias Bareinboim. "Causal abstraction inference under lossy representations." arXiv preprint arXiv:2509.21607 (2025).

Causal Representation Learning (CRL)

Goal: Instead of matching high- and low-level models via abstractions, can we *learn* the high-level causal *variables* and structure from low-level data only?

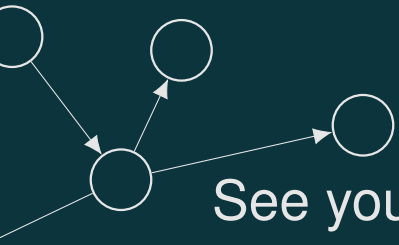
- **From Pixels to Attributes:** Separating factors of variation (e.g., color, shape, temperature readings, …) in latent space.
- **Causal Consistency:** Guarantee that the learned high-level factors follow the underlying system dynamics and the true causal relations.

Causal Representation Learning (CRL)

Goal: Instead of matching high- and low-level models via abstractions, can we *learn* the high-level causal *variables* and structure from low-level data only?

- **From Pixels to Attributes:** Separating factors of variation (e.g., color, shape, temperature readings, …) in latent space.
- **Causal Consistency:** Guarantee that the learned high-level factors follow the underlying system dynamics and the true causal relations.

More on that in two lectures…



See you next week!