Probabilistic models, such as **Bayesian Networks**, enable the decomposition of joint probabilities

\[ P(\text{Rain}, \text{Music}) \]
Probabilistic models, such as **Bayesian Networks**, enable the decomposition of joint probabilities

\[
P(\text{\ding{109}}, \text{\ding{102}})
\]

\[
P(\text{\ding{102}} \mid \text{\ding{109}}) \cdot P(\text{\ding{109}})
\]
Motivation

Bayesian Networks

Probabilistic models, such as **Bayesian Networks**, enable the decomposition of joint probabilities

\[ P(\text{💧, 🧂}) \]

**Causal ordering** is not necessary.
In a **Causal Bayesian Network**, edges represent **causal** relations.

Given causal ordering, we can represent external **interventions** on the model.
In a **Causal Bayesian Network**, edges represent causal relations.

Given causal ordering, we can represent external interventions on the model.
Motivation

What’s a causal variable?

---

**Independent Mechanism Principle (Peters et al. 2017)**

The causal generative process of a system’s variables is composed of autonomous modules that do not inform or influence each other.
Motivation

Causal Models

Figure 1: A probabilistic model represents a distribution $\mathbb{P}_X$ on a set of random variables $X$. For each intervention $i$, a causal model represents a distinct distribution $\mathbb{P}_X^i$ on the same variables, where the observational distribution corresponds to the empty intervention. Illustration from Schölkopf et al. (2021).
Motivation

Causal Models

**Figure 1:** A probabilistic model represents a distribution $\mathbb{P}_X$ on a set of random variables $X$. For each intervention $i$, a causal model represents a distinct distribution $\mathbb{P}^i_X$ on the same variables, where the observational distribution corresponds to the empty intervention. Illustration from Schölkopf et al. (2021).
1. Structural Causal Models

2. Causal Reasoning

3. Causal Discovery

4. Causal Abstraction

5. Causal Representation Learning
Structural Causal Models
A Structural Causal Model

\[ M = (X, E, f, P_E), \]

specifies the deterministic mechanisms \( f \) between a set of endogenous variables \( X \) and a set of exogenous variables \( E \) with distribution \( P_E \).
To each *endogenous* variable $X \in \mathbf{X}$, we assign an *exogenous* variable $E_X \in \mathbf{E}$. The endogenous mechanism $f_X$ of $X$ is then defined as a function

$$f_X : \mathcal{D}(\text{Pa}(X) \cup E_X) \to \mathcal{D}(X).$$

Due to acyclicity, we can define the model reduction

$$\mathcal{M} : \mathcal{D}(\mathbf{E}) \to \mathcal{D}(\mathbf{X}).$$
Linear Gaussian SCMs

Given the exogenous distribution $P_E$, the deterministic mechanisms $f$ induce a distribution on the endogenous variables $P_X$.

Structural Causal Model

\[
\begin{align*}
X_1 &= E_1 \\
X_2 &= a \cdot X_1 + E_2 \\
X_3 &= b \cdot X_1 + c \cdot X_2 + E_3 \\
E_1, E_2, E_3 &\sim \mathcal{N}(0, I)
\end{align*}
\]
Given the exogenous distribution $P_E$, the deterministic mechanisms $f$ induce a distribution on the endogenous variables $P_X$.

**Structural Causal Model**

\[
X_1 = E_1 \\
X_2 = a \cdot X_1 + E_2 \\
X_3 = b \cdot X_1 + c \cdot X_2 + E_3 \\
E_1, E_2, E_3 \sim \mathcal{N}(0, I)
\]

**Causal Bayesian Network**

\[
X_1 \sim \mathcal{N}(0, 1) \\
X_2 \mid X_1 \sim \mathcal{N}(a \cdot X_1, 1) \\
X_3 \mid X_1, X_2 \sim \mathcal{N}(b \cdot X_1 + c \cdot X_2, 1)
\]
Linear Gaussian SCMs

Given the exogenous distribution \( P_E \), the deterministic mechanisms \( f \) induce a distribution on the endogenous variables \( P_X \).

**Structural Causal Model**

\[
\begin{align*}
X_1 &= 2 \cdot E_1 \\
X_2 &= a \cdot X_1 + 2 \cdot E_2 \\
X_3 &= b \cdot X_1 + c \cdot X_2 + 2 \cdot E_3 \\
E_1, E_2, E_3 &\sim \mathcal{N}(0, I \cdot 1/2)
\end{align*}
\]

**Causal Bayesian Network**

\[
\begin{align*}
X_1 &\sim \mathcal{N}(0, 1) \\
X_2 \mid X_1 &\sim \mathcal{N}(a \cdot X_1, 1) \\
X_3 \mid X_1, X_2 &\sim \mathcal{N}(b \cdot X_1 + c \cdot X_2, 1)
\end{align*}
\]
Intervened Models

How to intervene on an SCM?

**Hard Intervention**

Given an SCM

\[ \mathcal{M} = (X, E, f, P_E), \]

a subset of variables \( V \subset X \) and a setting \( v \in \mathcal{D}(V) \), an hard intervention \( i = (V \leftarrow v) \) results in a SCM \( \mathcal{M}^i = (X, E, f^i, P_E) \), where

\[
  f^i_X = \begin{cases} 
    v_X & X \in V \\
    f_X & X \notin V, 
  \end{cases}
\]

for each endogenous variable \( X \in X \).
Intervened Models

Hard Intervention

Given an SCM

\[ M = (X, E, f, P_E), \]

a subset of variables \( V \subset X \) and a setting \( v \in \mathcal{D}(V) \), an hard intervention \( i = (V \leftarrow v) \) results in a SCM \( M^i = (X, E, f^i, P_E) \), where

\[
f^i_X = \begin{cases} 
  v_X & X \in V \\
  f_X & X \notin V,
\end{cases}
\]

for each endogenous variable \( X \in X \).

Interventions can be: soft, stochastic, perfect, imperfect, ...
Causal Reasoning
### Causal Queries

<table>
<thead>
<tr>
<th>Layer</th>
<th>Query</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Association</td>
<td>$P(Y \mid X = x)$</td>
<td>BN</td>
</tr>
<tr>
<td>Intervention</td>
<td>$P(Y \mid \text{do}(X \leftarrow x))$</td>
<td>CBN</td>
</tr>
<tr>
<td>Counterfactual</td>
<td>$P(Y = y' \mid Y = y, X = x, \text{do}(X \leftarrow x'), x \neq x')$</td>
<td>SCM</td>
</tr>
</tbody>
</table>
Causal Queries

# Pearl’s Ladder of Causation

<table>
<thead>
<tr>
<th>Layer</th>
<th>Query</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Association</strong></td>
<td>( P(Y \mid X = x) )</td>
<td>BN</td>
</tr>
<tr>
<td>Intervention</td>
<td>( P(Y \mid \text{do}(X \leftarrow x)) )</td>
<td>CBN</td>
</tr>
<tr>
<td>Counterfactual</td>
<td>( P(Y = y' \mid Y = y, X = x, \text{do}(X \leftarrow x'), x \neq x' )</td>
<td>SCM</td>
</tr>
</tbody>
</table>
### Causal Queries

Pearl’s Ladder of Causation

<table>
<thead>
<tr>
<th>Layer</th>
<th>Query</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Association</td>
<td>$P(Y \mid X = x)$</td>
<td>BN</td>
</tr>
<tr>
<td><strong>Intervention</strong></td>
<td>$P(Y \mid \text{do}(X \leftarrow x))$</td>
<td>CBN</td>
</tr>
<tr>
<td>Counterfactual</td>
<td>$P(Y = y' \mid Y = y, X = x, \text{do}(X \leftarrow x'))$, $x \neq x'$</td>
<td>SCM</td>
</tr>
</tbody>
</table>
Average Treatment Effect

Suppose that we have a binary *treatment* variable $X$ and an *outcome* variable $Y$. Then, we can compute the **Average Treatment Effect** as

$$\text{ATE}(X, Y) = \mathbb{E}_{Y \sim Y|\text{do}(X\leftarrow 1)} [y] - \mathbb{E}_{Y \sim Y|\text{do}(X\leftarrow 0)} [y]$$
Suppose that we have a binary \textit{treatment} variable \( X \) and an \textit{outcome} variable \( Y \). Then, we can compute the \textbf{Average Treatment Effect} as

\[
\text{ATE}(X, Y) = \mathbb{E}_{y \sim Y | \text{do}(X \leftarrow 1)} [y] - \mathbb{E}_{y \sim Y | \text{do}(X \leftarrow 0)} [y]
\]

\[
= \sum_{y \in \mathcal{D}(Y)} y \cdot p_{y}^{\text{do}(X \leftarrow 1)}(y) - \sum_{y \in \mathcal{D}(Y)} y \cdot p_{y}^{\text{do}(X \leftarrow 0)}(y).
\]
To compute $\text{ATE}(X, Y)$ or any other causal estimate, we need to compute the interventional distribution $p_{Y}^{\text{do}(X \leftarrow x)}$. 
To compute $\text{ATE}(X, Y)$ or any other causal estimate, we need to compute the interventional distribution $p_{Y}^{\text{do}(X \leftarrow x)}$.

The distribution is **causally identifiable** if it can be computed from the observational distribution.
To compute $\text{ATE}(X, Y)$ or any other causal estimate, we need to compute the interventional distribution $p_{Y}^{\text{do}(X \leftarrow x)}$.

The distribution is **causally identifiable** if it can be computed from the observational distribution.

In general, this is not possible, but we can use the **do-calculus** to identify the conditions under which it is possible.
To compute $\text{ATE}(X, Y)$ or any other causal estimate, we need to compute the interventional distribution $p_{Y}^{\text{do}(X \leftarrow x)}$.

The distribution is **causally identifiable** if it can be computed from the observational distribution.

In general, this is not possible, but we can use the **do-calculus** to identify the conditions under which it is possible.

The do-calculus is a **complete** set of rules that can be easily applied to any causal graph.
An hospital offers two distinct surgeries ($T$) for kidney stones which they assign depending on the whether the stones are large ($Z = 1$) or small ($Z = 0$). Depending on the size of the stones and the treatment, the success rate ($R$) varies.
Backdoor Criterion

\[
\begin{align*}
Z &= 0 & Z &= 1 \\
T = a & 0.93 (81/87) & 0.73 (192/263) \\
T = b & 0.87 (234/270) & 0.69 (55/80)
\end{align*}
\]
We can define the best treatment by computing the interventional success rate.

\[
\begin{align*}
P(R = 1 \mid \text{do}(T \leftarrow a)) &= P(R = 1 \mid T = a, Z = 0)P(Z = 0) \\
&\quad + P(R = 1 \mid T = a, Z = 1)P(Z = 1) \\
&= 0.93 \cdot 0.51 + 0.73 \cdot 0.49 = 0.832, \\
P(R = 1 \mid \text{do}(T \leftarrow b)) &= P(R = 1 \mid T = b, Z = 0)P(Z = 0) \\
&\quad + P(R = 1 \mid T = b, Z = 1)P(Z = 1) \\
&= 0.87 \cdot 0.51 + 0.69 \cdot 0.49 = 0.782.
\end{align*}
\]
By computing the conditional probabilities, we can easily see how conditioning differs from intervening.

\[
P(R = 1 \mid T = a) = 0.780, \quad P(R = 1 \mid \text{do}(T \leftarrow a)) = 0.832, \]
\[
P(R = 1 \mid T = b) = 0.830, \quad P(R = 1 \mid \text{do}(T \leftarrow b)) = 0.782. \]

Given the probability of success when observing the treatments, we would have, arguably incorrectly, chosen treatment \(b\).
## Causal Queries

<table>
<thead>
<tr>
<th>Layer</th>
<th>Query</th>
<th>Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Association</td>
<td>( P(Y \mid X = x) )</td>
<td>BN</td>
</tr>
<tr>
<td>Intervention</td>
<td>( P(Y \mid \text{do}(X \leftarrow x)) )</td>
<td>CBN</td>
</tr>
<tr>
<td><strong>Counterfactual</strong></td>
<td>( P(Y = y' \mid Y = y, X = x, \text{do}(X \leftarrow x')) ), ( x \neq x' )</td>
<td>SCM</td>
</tr>
</tbody>
</table>
Example 3.4 (Eye disease)  There exists a rather effective treatment for an eye disease. For 99% of all patients, the treatment works and the patient gets cured ($B = 0$); if untreated, these patients turn blind within a day ($B = 1$). For the remaining 1%, the treatment has the opposite effect and they turn blind ($B = 1$) within a day. If untreated, they regain normal vision ($B = 0$).

Which category a patient belongs to is controlled by a rare condition ($N_B = 1$) that is unknown to the doctor, whose decision whether to administer the treatment ($T = 1$) is thus independent of $N_B$. We write it as a noise variable $N_T$.

Assume the underlying SCM

$$
\begin{align*}
\mathcal{C} : \quad T & := N_T \\
B & := T \cdot N_B + (1 - T) \cdot (1 - N_B)
\end{align*}
$$

with Bernoulli distributed $N_B \sim \text{Ber}(0.01)$; note that the corresponding causal graph is $T \rightarrow B$.

Now imagine a specific patient with poor eyesight comes to the hospital and goes blind ($B = 1$) after the doctor administers the treatment ($T = 1$). We can now ask the counterfactual question “What would have happened had the doctor administered treatment $T = 0$?”
Causal Discovery
Causal Discovery, or causal learning, consists of determining causal relations between variables $X$ from their observational distribution $\mathbb{P}_X$. 
Causal Discovery, or causal learning, consists of determining causal relations between variables $X$ from their observational distribution $P_X$. 
Graph Identifiability

Given a set of assumptions $A$, we say that the graph $G$ of a causal model $M$ is identifiable from the distribution $P_X$ whenever there does not exist another causal model $M'$ satisfying $A$ with a different graph $G'$ but the same observational distribution $P_X$.

Lachapelle et al. 2019
Structure Identifiability

Can we even learn a causal model?

**Graph Identifiability**

Given a set of assumptions $\mathcal{A}$, we say that the graph $\mathcal{G}$ of a causal model $\mathcal{M}$ is identifiable from the distribution $\mathbb{P}_X$ whenever there does not exist another causal model $\mathcal{M}'$ satisfying $\mathcal{A}$ with a different graph $\mathcal{G}'$ but the same observational distribution $\mathbb{P}_X$.

Lachapelle et al. 2019

$\mathcal{A} = \emptyset \implies$ No identifiability.
A model is causally sufficient whenever there are no unobserved confounders. This equates to assuming that there is no selection bias and the exogenous terms $E$ are marginally independent, i.e.,

$$\forall i \neq j. \quad E_i \perp \perp E_j.$$
A model is causally sufficient whenever there are no unobserved confounders. This equates to assuming that there is no selection bias and the exogenous terms $E$ are marginally independent, i.e.,

$$\forall i \neq j. \quad E_i \perp \perp E_j.$$
A model is causally faithful whenever all conditional independences in the distribution $\mathbb{P}_X$ imply $d$-separations in the graph $G$, i.e.,

$$A \perp \perp B \mid C \implies A \perp \perp_G B \mid C.$$
**Common Assumptions**

\[ A = \{\text{Faithfulness}\} \]

A model is causally faithfull whenever all conditional independences in the distribution \( P_x \) imply \( d \)-separations in the graph \( G \), i.e.,

\[ A \perp\!\!\!\perp B \mid C \implies A \perp\!\!\!\perp_G B \mid C. \]

Faithfulness Violation,

\[ X_1 \perp\!\!\!\perp X_4 \text{ but } X_1 \not\perp\!\!\!\perp_G X_4. \]
The PC algorithm is a constraint-based discovery method that iteratively removes and orients edges.
The PC algorithm is a constraint-based discovery method that iteratively removes and orients edges.
The PC algorithm is a constraint-based discovery method that iteratively removes and orients edges.

The algorithm returns the Markov equivalence class of the true causal graph, which can contain multiple graphs.

$A = \{\text{Sufficiency, Faithfulness}\}$
<table>
<thead>
<tr>
<th>Description</th>
<th>Mechanism</th>
<th>$f$</th>
<th>Id.</th>
</tr>
</thead>
<tbody>
<tr>
<td>General SCM</td>
<td>$X := f_X(\text{Pa}(X), E_X)$</td>
<td></td>
<td>$\times$</td>
</tr>
<tr>
<td>Description</td>
<td>Mechanism</td>
<td>$f$</td>
<td>Id.</td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------------------------------------</td>
<td>-----------</td>
<td>-----</td>
</tr>
<tr>
<td>General SCM</td>
<td>$X := f_X(Pa(X), E_X)$</td>
<td>—</td>
<td>✗</td>
</tr>
<tr>
<td>ANM</td>
<td>$X := f_X(Pa(X)) + E_X$</td>
<td>Nonlinear</td>
<td>✓</td>
</tr>
<tr>
<td>CAM</td>
<td>$X := \sum_{X' \in Pa(X)} f(X') + E_X$</td>
<td>Nonlinear</td>
<td>✓</td>
</tr>
<tr>
<td>Gaussian ANM</td>
<td>$X := \langle w, Pa(X) \rangle + E_X$</td>
<td>Linear</td>
<td>✗</td>
</tr>
<tr>
<td>Non-Gaussian ANM</td>
<td>$X := f_X(Pa(X)) + E_X$</td>
<td>Linear</td>
<td>✓</td>
</tr>
<tr>
<td>Gaussian Eq. Var</td>
<td>$X := \langle w, Pa(X) \rangle + E_X$</td>
<td>Linear</td>
<td>✓</td>
</tr>
</tbody>
</table>
Continuous Causal Discovery (CCD) approaches try to recast combinatorial discovery algorithms as optimization problems.

\[
\min_{\mathcal{G}} S(\mathcal{G}, D_x) \\
\text{s.t. } \mathcal{G} \text{ is acyclic.}
\]

There are two main open problems:
Continuous Causal Discovery (CCD) approaches try to recast combinatorial discovery algorithms as optimization problems.

\[
\min_{\mathcal{G}} S(\mathcal{G}, \mathcal{D}_x)
\]

s.t. \( \mathcal{G} \) is acyclic.

There are two main open problems:

- How to \textit{efficiently} enforce acyclicity on the solution \( \mathcal{G} \)?
Continuous Causal Discovery (CCD) approaches try to recast combinatorial discovery algorithms as optimization problems.

\[
\min_{\mathcal{G}} S(\mathcal{G}, \mathcal{D}_X)
\]

s.t. \(\mathcal{G}\) is acyclic.

There are two main open problems:

- How to \textit{efficiently} enforce acyclicity on the solution \(\mathcal{G}\)?
- How to encode assumptions in the \textit{score} function \(S\)?
Causal Abstraction
Traditional causal discovery algorithms require a large number of samples and are computationally expensive.
Traditional causal discovery algorithms require a large number of samples and are computationally expensive.

Can we represent a system with a *simpler* model at an *higher-level* of abstraction?
Causal Abstraction

Main Intuition

Given an abstraction function $\tau$, an SCM $\mathcal{H}$ is an abstraction of $\mathcal{L}$ if the diagram commutes, i.e.,

\[ u \xrightarrow{\mathcal{H}} y \]

\[ e \xrightarrow{\mathcal{L}} x \]
Causal Abstraction

Main Intuition

Given an abstraction function $\tau$, an SCM $\mathcal{H}$ is an abstraction of $\mathcal{L}$ if the diagram commutes, i.e.,

$$\tau \circ \mathcal{L} = \mathcal{H} \circ \tau.$$
Given an abstraction function $\tau$, an SCM $\mathcal{H}$ is an abstraction of $\mathcal{L}$ if the diagram commutes, i.e.,

$$\tau \circ \mathcal{L}^i = \mathcal{H}^j \circ \tau,$$

for any intervention $i$. 
In the *linear* case, we can use abstraction to cluster larger causal graphs.
In the *linear* case, we can use abstraction to cluster larger causal graphs.

This improves the complexity of causal reasoning and allows for more interpretable models.
Introducing abstract information in the LiNGAM pipeline, we gain significant speedup (2x) in execution time (b, right) without performance loss (a, left).
Causal Representation Learning
In many contexts, we can assume that high-dimensional observations $x$ are generated through a decoder function $f: \mathcal{D}(Z) \to \mathcal{D}(X)$ from a set of latent causal variables $Z$. 
Disentanglement

Factors are \textit{statistically} independent. Altering a factor should only affect a single dimension of the data.

Causal Representation

Factors are \textit{causally} independent. Altering a factor might affect other factors, but we can independently manipulate them.
We need interventional samples to learn causal representations such as TempoRal Intervened Sequences (TRIS),

\[ \mathcal{D} = \{(x_t, i, x_{t+1})\}, \]

where we can observe the state of the model before \textit{and} after an intervention \( i \).
We need interventional samples to learn causal representations such as TempoRal Intervened Sequences (TRIS),

$$D = \{(x_t, i, x_{t+1})\},$$

where we can observe the state of the model before and after an intervention $i$. 
We need interventional samples to learn causal representations such as TempoRal Intervened Sequences (TRIS),

\[ D = \{(x_t, i, x_{t+1})\} , \]

where we can observe the state of the model before and after an intervention \( i \).
Conclusion
Main References

Elements of Causal Inference

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf

Elements of Causal Inference: Foundations and Learning Algorithms

Jonas Peters, Dominik Janzing, Bernhard Schölkopf

The MIT Press, 2017
Main References

Causality

Causality:
Models, Reasoning, and Inference

Judea Pearl

Cambridge University Press, 2009
Conclusion

• Causal Graphical Models are a powerful tool to represent causal relationships between variables.
• Efficient causal discovery from observational data is a challenging problem.
• Causal Abstraction and Causal Representation Learning are promising research directions for future applications.

Riccardo Massidda
Università di Pisa
Dept. of Computer Science
Office 298B, 2nd Floor
pages.di.unipi.it/massidda/
riccardo.massidda@phd.unipi.it
Conclusion

• Causal *Graphical Models* are a powerful tool to represent causal relationships between variables.
• Causal *Graphical Models* are a powerful tool to represent causal relationships between variables.

• Efficient causal *discovery* from observational data is a challenging problem.
• Causal *Graphical Models* are a powerful tool to represent causal relationships between variables.

• Efficient causal *discovery* from observational data is a challenging problem.

• Causal *Abstraction* and Causal *Representation Learning* are promising research directions for future applications.
Conclusion

• Causal *Graphical Models* are a powerful tool to represent causal relationships between variables.
• Efficient causal *discovery* from observational data is a challenging problem.
• Causal *Abstraction* and Causal *Representation Learning* are promising research directions for future applications.

Riccardo Massidda
Università di Pisa
Dept. of Computer Science
Office 298B, 2nd Floor

pages.di.unipi.it/massidda/
riccardo.massidda@phd.unipi.it

