Introduction to Causal Inference & Causal Discovery
Overview

Causal inference:
- Causal graphical models
- Interventions (the "do" operator)
- Example: Study of Kidney Stone Treatments
- Backdoor criterion
- The ladder of causation
- Counterfactuals

Causal discovery:
- Markov equivalence
- Faithfulness
- Structure identifiability
- Constraint-based methods
- Score-based methods
Causal Inference
Causal graphical models (CGM)

- A causal graphical model (CGM) is a pair \((p, G)\) s.t.

- \(G\) is a **directed acyclic graph** (DAG)

- \(p \in \mathcal{L}(G)\), i.e. \(p\) factorizes according to \(G\).

- \(G\) describes **causal relationships** between variables, i.e., how the system reacts to **interventions**.
A causal graphical model (CGM) is a pair \((p, G)\) s.t.

- \(G\) is a directed acyclic graph (DAG)
- \(p \in \mathcal{L}(G)\), i.e. \(p\) factorizes according to \(G\).
- \(G\) describes causal relationships between variables, i.e., how the system reacts to interventions.

Example: Kidney stone treatment

\[
T = \text{Treatment} \in \{A, B\} \\
S = \text{Stone size} \in \{\text{small}, \text{large}\} \\
R = \text{Patient recovered} \in \{0, 1\}
\]

\[
p(S, T, R) = p(S)p(T | S)p(R | S, T)
\]
Throughout, we will assume **perfect deterministic** interventions.

**Definition (The “do” operator)**

Given a causal graphical model \((p, G)\),

\[
p(x \mid do(x'_k)) := \delta(x_k, x'_k) \prod_{i \neq k} p(x_i \mid x_{\pi_i^G})
\]

Thus, \(p(x \mid do(x'_k))\) is a "new" distribution over \(X_V\).
Throughout, we will assume **perfect deterministic** interventions.

**Definition (The “do” operator)**

Given a causal graphical model \((p, G)\),

\[
p(x \mid do(x'_k)) := \delta(x_k, x'_k) \prod_{i \neq k} p(x_i \mid x_{\pi_i^G})
\]

- Thus, \(p(x \mid do(x'_k))\) is a "new" distribution over \(X_V\).

- Can compute marginals, e.g. \(p(x_i \mid do(x'_k)) = \sum_{X \setminus \{i\}} p(x \mid do(x'_k))\)
The "do" operator

Throughout, we will assume perfect deterministic interventions.

Definition (The “do” operator)

Given a causal graphical model \((p, G)\),

\[
p(x \mid do(x'_k)) := \delta(x_k, x'_k) \prod_{i \neq k} p(x_i \mid x_{\pi_i^G})
\]

- Thus, \(p(x \mid do(x'_k))\) is a "new" distribution over \(X_V\).
- Can compute marginals, e.g. \(p(x_i \mid do(x'_k)) = \sum_{x_{\setminus \{i\}}} p(x \mid do(x'_k))\)
- ... and conditionals, e.g. \(p(x_i \mid x_j, do(x'_k)) = \frac{p(x_i, x_j \mid do(x'_k))}{p(x_j \mid do(x'_k))}\)
The "do" operator

Throughout, we will assume **perfect deterministic** interventions.

**Definition (The “do” operator)**

Given a causal graphical model \((p, G)\),

\[
p(x \mid do(x'_k)) := \delta(x_k, x'_k) \prod_{i \neq k} p(x_i \mid x_{\pi_i^G})
\]

- **Thus**, \(p(x \mid do(x'_k))\) is a "new" distribution over \(X_V\).

- Can compute marginals, e.g. \(p(x_i \mid do(x'_k)) = \sum_{x_V \setminus \{i\}} p(x \mid do(x'_k))\)

- ... and conditionals, e.g. \(p(x_i \mid x_j, do(x'_k)) = \frac{p(x_i, x_j \mid do(x'_k))}{p(x_j \mid do(x'_k))}\)

- **Remark:** \(p(x_{V \setminus \{k\}} \mid do(x_k)) = \prod_{i \neq k} p(x_i \mid x_{\pi_i^G})\).
The "do" operator

Back to our example

\[
P(S, R \mid \text{do}(T)) = P(S) \underbrace{P(T \mid S)} \cdot P(R \mid S, T)
\]

The decision of taking treatment \( T \) does not depend on \( S \) anymore.
The "do" operator

Back to our example

\[ P(S, R \mid do(T)) = P(S) P(T \mid S) P(R \mid S, T) \]

The decision of taking treatment \( T \) does not depend on \( S \) anymore

Notice \( p(\cdot \mid do(x'_k)) \in \mathcal{L}(G') \), where \( G' \) is the mutilated graph, i.e.

\[ G' = (V, E') \quad E' = \{(i, j) \in E \mid j \neq k\} \]
The "do" operator

- Back to our example

\[ P(S, R \mid \text{do}(T)) = P(S) P(T \mid S) P(R \mid S, T) \]

The decision of taking treatment \( T \) does not depend on \( S \) anymore

- Notice \( p(\cdot \mid \text{do}(x'_k)) \in \mathcal{L}(G') \), where \( G' \) is the mutilated graph, i.e.

\[ G' = (V, E') \quad E' = \{(i, j) \in E \mid j \neq k\} \]
**Different types of interventions**

**Intervening on the treatment $T$**

$T = \text{Treatment} \in \{A, B\}$

$S = \text{Stone size} \in \{\text{small, large}\}$

$R = \text{Patient recovered} \in \{0, 1\}$

### Observations

$p(S)p(T \mid S)p(R \mid S, T)$

### Perfect intervention

$p(S)p(T)p(R \mid S, T)$

### Imperfect intervention

$p(S)p(T \mid S)p(R \mid S, T)$
Different types of interventions

### Intervening on the treatment $T$

- $T =$ Treatment $\in \{A, B\}$
- $S =$ Stone size $\in \{\text{small, large}\}$
- $R =$ Patient recovered $\in \{0, 1\}$

#### Observations

\[
p(S)p(T | S)p(R | S, T)
\]

#### Perfect intervention

\[
p(S)\tilde{p}(T)p(R | S, T)
\]

#### Imperfect intervention

\[
p(S)\tilde{p}(T | S)p(R | S, T)
\]

Definition presented previously is a perfect intervention with $\tilde{p}(T) := \delta(T, T')$. It is sometimes called a **perfect deterministic intervention**.
Why should I care!?! (Kidney Stone Treatment)

\[ T = \text{Treatment} \in \{A, B\} \]
\[ S = \text{Stone size} \in \{\text{small, large}\} \]
\[ R = \text{Patient recovered} \in \{0, 1\} \]

\[ p(S)p(T \mid S)p(R \mid S, T) \]

(Example taken from *Element of Causal Inference* by Peters et al. p111)
Pay attention to these two questions...
Why should I care!?! (Kidney Stone Treatment)

Pay attention to these two questions...

1- What is your chance of recovery knowing that the doctor gave you treatment A?

2- What is your chance of recovery if you decide to take treatment A?

(In both cases, assume you don’t know the size of your stone)
Why should I care!?! (Kidney Stone Treatment)

\[ T = \text{Treatment} \in \{A, B\} \]
\[ Z = \text{Stone size} \in \{\text{small}, \text{large}\} \]
\[ R = \text{Patient recovered} \in \{0, 1\} \]

What is your chance of recovery knowing that the doctor gave you treatment A?

- Compute \( P(R = 1 \mid T = A) \)! (we know how to do that :D)

- Knowing that your doctor gave you treatment A tells you that you probably have a large kidney stone ... \( P(S = \text{large} \mid T = A) = 0.75 \)

- ... which reduces your chance of recovery
  \[
  P(R = 1 \mid T = A, S = \text{large}) = 0.73 < 0.93 = P(R = 1 \mid T = A, S = \text{small})
  \]
Why should I care!?! (Kidney Stone Treatment)

\[ T = \text{Treatment} \in \{A, B\} \]
\[ Z = \text{Stone size} \in \{\text{small, large}\} \]
\[ R = \text{Patient recovered} \in \{0, 1\} \]

What is your chance of recovery knowing that the doctor gave you treatment A?

- Compute \( P(R = 1 \mid T = A) \) (we know how to do that :D)
- Knowing that your doctor gave you treatment A tells you that you probably have a large kidney stone ... \( P(S = \text{large} \mid T = A) = 0.75 \)
  - ... which reduces your chance of recovery
    \( P(R = 1 \mid T = A, S = \text{large}) = 0.73 < 0.93 = P(R = 1 \mid T = A, S = \text{small}) \)

What is your chance of recovery if you decide to take treatment A?

- \( P(R = 1 \mid do(T = A)) \)
- Your really don't know anything about your kidney stone
Why should I care!?! (Kidney Stone Treatment)

\[ T = \text{Treatment} \in \{A, B\} \]
\[ S = \text{Stone size} \in \{\text{small, large}\} \]
\[ R = \text{Patient recovered} \in \{0, 1\} \]

\[ P(S, R \mid do(T)) = P(S)P(T \mid S)P(R \mid S, T) \]

The decision of taking treatment \( T \)
does not depend on \( S \) anymore

Then simply marginalize as usual:

\[ P(R = 1 \mid do(T = A)) = \sum_S P(R = 1, S \mid do(T = A)) \]
\[ = \sum_S P(R = 1 \mid S, T = A)P(S) = 0.832 \]
Causal inference

Why should I care!?! (Kidney Stone Treatment)

\[ T = \text{Treatment} \in \{A, B\} \]
\[ S = \text{Stone size} \in \{\text{small, large}\} \]
\[ R = \text{Patient recovered} \in \{0, 1\} \]

What is your chance of recovery knowing that the doctor gave you treatment A?

\[ P(R = 1|T = A) = 0.78 \]
\[ P(R = 1|T = B) = 0.83 \]

What is your chance of recovery if you decide to take treatment A?

\[ P(R = 1|\text{do}(T = A)) = 0.832 \]
\[ P(R = 1|\text{do}(T = B)) = 0.782 \]
What just happened? We showed

$$P(R = 1\mid \text{do}(T = A)) = \sum_S P(R = 1\mid S, T = A)P(S)$$

Never observed data from $p(T, S, R \mid \text{do}(T = A))$ ...Yet I can estimate the query, since there is no "do" here :D
What just happened? We showed

\[
P(R = 1|do(T = A)) = \sum_S P(R = 1|S, T = A)P(S)
\]

Never observed data from \(p(T, S, R | do(T = A))\)

...Yet I can estimate the query, since there is no "do" here :D

Formally, this means \(p(R = 1 | do(T = A))\) is **identifiable from** \(p(R, T, S)\) and \(G\) (our computations *critically* relied on the causal graph).
Why should I care!?! (Kidney Stone Treatment)

- What just happened? We showed

\[
P(R = 1 | do(T = A)) = \sum_S P(R = 1 | S, T = A) P(S)
\]

Never observed data from \( p(T, S, R | do(T = A)) \)

...Yet I can estimate the query, since there is no "do" here :D

- Formally, this means \( p(R = 1 | do(T = A)) \) is identifiable from \( p(R, T, S) \) and \( G \) (our computations critically relied on the causal graph).

- Turns out what we just did is an instance of the backdoor criterion...
Theorem (Backdoor criterion)

\[ p(x_i \mid do(x_k)) = \sum_{x_S} p(x_i \mid x_k, x_S)p(x_S) \text{ if} \]

1. *S contains no descendants of* \( x_k \), *and*

2. *S blocks all paths from* \( x_i \) *to* \( x_k \) *entering* \( x_k \) *from "the backdoor", i.e. such that* \( x_k \leftarrow \ldots x_i \)
**Theorem (Backdoor criterion)**

\[ p(x_i \mid do(x_k)) = \sum_{x_S} p(x_i \mid x_k, x_S)p(x_S) \]

if

1. \( S \) contains no descendants of \( x_k \), and
2. \( S \) blocks all paths from \( x_i \) to \( x_k \) entering \( x_k \) from "the backdoor", i.e. such that \( x_k \leftarrow \ldots x_i \)

Say we want to compute \( p(y \mid do(x)) \):

**Left path:** Only backdoor path. Blocked by \( S = \{K\} \). **Right path:** Why we cannot include a descendant of \( X \) in \( S \).
Can all identifiable queries $p(x_i \mid do(x_k))$ be expressed with the backdoor criterion?
Can all identifiable queries $p(x_i \mid do(x_k))$ be expressed with the backdoor criterion?

Answer: No!

Since $U$ is unobserved, we cannot apply the backdoor criterion... Turns out we can nevertheless identify $p(y \mid do(x))$ from $p(X, Z, Y)$ using the front-door criterion. Look it up!
Can all identifiable queries $p(x_i \mid do(x_k))$ be expressed with the backdoor criterion?

Answer: No!

- Since $U$ is unobserved, we cannot apply the backdoor criterion...

- Turns out we can nevertheless identify $p(y \mid do(x))$ from $p(X, Z, Y)$ using the **front-door criterion**. Look it up!
Do-calculus

- Do-calculus is a set of **three rules** that can be applied to transform an interventional query (including a "do") into an observational expression (without any "do").

- Not enough time to present them...

- All identifiable queries can be found by a subsequent application of these rules, i.e. the rules are **complete**.
You now know about the first two steps of Pearl’s "ladder of causation".

<table>
<thead>
<tr>
<th>Level (Symbol)</th>
<th>Typical Activity</th>
<th>Typical Questions</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Association ( P(y</td>
<td>x) )</td>
<td>Seeing</td>
<td>What is? How would seeing ( X ) change my belief in ( Y )?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>What does a survey tell us about the election results?</td>
</tr>
<tr>
<td>2. Intervention ( P(y</td>
<td>do(x), z) )</td>
<td>Doing Intervening</td>
<td>What if? What if I do ( X )?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>What if we ban cigarettes?</td>
</tr>
<tr>
<td>3. Counterfactuals ( P(y_x</td>
<td>x’, y’) )</td>
<td>Imagining, Retrospection</td>
<td>Why? Was it ( X ) that caused ( Y )? What if I had acted differently?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Would Kennedy be alive had Oswald not shot him?</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>What if I had not been smoking the past 2 years?</td>
</tr>
</tbody>
</table>

Fig. 1. The Causal Hierarchy. Questions at level \( i \) can only be answered if information from level \( i \) or higher is available.

Taken from “The Seven Tools of Causal Inference with Reflections on Machine Learning” by Judea Pearl
You need **structural causal models (SCM)**. Let $\mathcal{G}$ be a DAG:

\[ X_1 := f_1(X_{\pi_1}) + N_1 \]  
\[ X_2 := f_2(X_{\pi_2}) + N_2 \]  
\[ \ldots \]  
\[ X_d := f_d(X_{\pi_d}) + N_d \]

- This induces an **observational** distribution
- Can define **interventions** as well
- Can define **counterfactual** statements (not possible with a causal graphical model). See Section 6.4 in ECI.
Causal Discovery
Recall: A Directed Graphical Model encodes the Conditional Independence of a distribution.

Multiple DAGs may encode the same Conditional Independence statements.

Two DAGs encoding the same Conditional Independence statements are called **Markov Equivalent**.
Markov Equivalence

Theorem (Verma & Pearl, 1991)

Two DAGs $G_1$ and $G_2$ are **Markov Equivalent** if and only if they have the same skeleton and the same v-structures.
Markov Equivalence

Theorem (Verma & Pearl, 1991)

Two DAGs $G_1$ and $G_2$ are **Markov Equivalent** if and only if they have the same skeleton and the same v-structures.

Markov Equivalence Classes can be represented as a **Completed Partially Directed Acyclic Graph** (CPDAG).
Faithfulness

Global Markov Property

$A \& B$ are d-separated given $C$ in $G$

$X_A \perp X_B \mid X_C$
Faithfulness

Global Markov Property

\[ A \ & \ B \] are d-separated
\[ \text{given} \ C \ \text{in} \ G \]

\[ X_A \perp X_B \mid X_C \]

Faithfulness
Faithfulness

Global Markov Property

\( A \) & \( B \) are d-separated given \( C \) in \( G \)

\[ X_A \perp X_B \mid X_C \]

Exercise: Violation of Faithfulness

\( X := N_X \)
\( Y := X + N_Y \)
\( Z := X - Y + N_Z \)

with \( N_X, N_Y, N_Z \sim iid \mathcal{N}(0, \sigma^2) \)

\( p(X, Y, Z) \) is a Multivariate Normal distribution, where the only conditional independence statements are: \( X \perp Z \) and \( X \not\perp Z \mid Y \).
**Faithfulness**

Global Markov Property

\[ A \& B \text{ are d-separated given } C \text{ in } G \]

\[ X_A \perp X_B \mid X_C \]

**Exercise: Violation of Faithfulness**

\[ p(X, Y, Z) \text{ is a Multivariate Normal distribution, where the only conditional independence statements are: } X \perp Z \text{ and } X \not\perp Z \mid Y. \]
Theorem

Assume that $p$ is faithful wrt. $\mathcal{G}^0$. The Markov Equivalence class of $\mathcal{G}^0$, represented by its CPDAG, is identifiable from $p$. Only the Markov Equivalence class is identifiable from observations, not an individual graph. Two Markov Equivalent graphs may lead to different causal conclusions! Under different assumptions, an individual DAG may be identifiable.
Assume that $p$ is faithful wrt. $\mathcal{G}^0$. The Markov Equivalence class of $\mathcal{G}^0$, represented by its CPDAG, is identifiable from $p$.

- Only the Markov Equivalence class is identifiable from observations, not an individual graph. Two Markov Equivalent graphs may lead to different causal conclusions!

- Under different assumptions, an individual DAG may be identifiable
Theorem

Assume that \( p \) is faithful wrt. \( G^0 \). The Markov Equivalence class of \( G^0 \), represented by its CPDAG, is identifiable from \( p \).

- Only the Markov Equivalence class is identifiable from observations, **not an individual graph**. Two Markov Equivalent graphs may lead to different causal conclusions!

\[
\begin{array}{c}
\require{amsmath}
\text{X} & \rightarrow & \text{Y} \\
\end{array}
\quad \text{or} \quad
\begin{array}{c}
\require{amsmath}
\text{X} & \leftarrow & \text{Y} \\
\end{array}
\]

- Under different assumptions, an individual DAG may be identifiable
  - Additive Noise Model (ANM): \( X_j \):= \( f_j(X_{Pa_j}) + N_j \), \( N_j \sim \mathcal{N}(0, \sigma^2) \), where \( f_j \) are nonlinear.
Structure Identifiability

Theorem

Assume that $p$ is faithful wrt. $\mathcal{G}^0$. The Markov Equivalence class of $\mathcal{G}^0$, represented by its CPDAG, is identifiable from $p$.

- Only the Markov Equivalence class is identifiable from observations, **not an individual graph**. Two Markov Equivalent graphs may lead to different causal conclusions!

- Under different assumptions, an individual DAG may be identifiable
  - Additive Noise Model (ANM): $X_j := f_j(X_{Pa_j}) + N_j$, $N_j \sim \mathcal{N}(0, \sigma^2)$, where $f_j$ are nonlinear.
  - Using **interventional data** (i.e. data resulting from controlled experiments).
Causal Structure Learning (Causal Discovery)

How to recover the (CP)DAG using a dataset $\mathcal{D}$?
Constraint-based methods

**Step 1: Identify the skeleton**

For each pair of nodes $X$ & $Y$, and $A \subseteq V\{X, Y\}$, test if $X \perp_D Y \mid A$.
If there is no set $A$ s.t. $X \perp_D Y \mid A$, then add an edge $X \rightarrow Y$. 

![Diagram of a causal graph with nodes A, S, T, L, E, B, X, D connected by edges.](attachment:image.png)
Constraint-based methods

**Step 1: Identify the skeleton**

For each pair of nodes $X$ & $Y$, and $A \subseteq V \setminus \{X, Y\}$, test if $X \perp_D Y \mid A$.

If there is no set $A$ s.t. $X \perp_D Y \mid A$, then add an edge $X \rightarrow Y$.

**Step 2: Identify the v-structures**

For each structure $X \leftarrow Z \rightarrow Y$ with no edge between $X$ & $Y$, orient $X \rightarrow Z \leftarrow Y$ iff $Z \not\in A$, where $A$ is such that $X \perp_D Y \mid A$. 

![Graph Diagram]
Constraint-based methods

**Step 1: Identify the skeleton**

For each pair of nodes $X$ & $Y$, and $A \subseteq V \setminus \{X, Y\}$, test if $X \perp_D Y \mid A$. If there is no set $A$ s.t. $X \perp_D Y \mid A$, then add an edge $X \rightarrow Y$.

**Step 2: Identify the v-structures**

For each structure $X \leftarrow Z \rightarrow Y$ with no edge between $X$ & $Y$, orient $X \rightarrow Z \leftarrow Y$ iff $Z \notin A$, where $A$ is such that $X \perp_D Y \mid A$.

**IC Algorithm**
Step 2’: Additional orientations

Use Meek’s orientation rules to orient some of the remaining edges.

Rule 1

Rule 2

Rule 3

Rule 4
Score-based methods

- Idea: treat the problem of learning the structure of the DAG as a **model selection problem**

\[
\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} | \mathcal{D})
\]

- Recall: choices of scores
Score-based methods

- Idea: treat the problem of learning the structure of the DAG as a **model selection problem**

\[
\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} \mid \mathcal{D})
\]

- Recall: choices of scores
  - **Likelihood score**:
    \[
    \text{score}_L(\mathcal{G} \mid \mathcal{D}) = \log p(\mathcal{D} \mid \hat{\theta}_M, \mathcal{G})
    \]
Score-based methods

- Idea: treat the problem of learning the structure of the DAG as a \textbf{model selection problem}.

\[
\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} \mid \mathcal{D})
\]

- Recall: choices of scores
  - \textbf{Likelihood score}:
    \[
    \text{score}_L(\mathcal{G} \mid \mathcal{D}) = \log p(\mathcal{D} \mid \hat{\theta}^{\text{MLE}}_\mathcal{G}, \mathcal{G})
    \]
  - \textbf{Bayesian score}:
    \[
    \text{score}_B(\mathcal{G} \mid \mathcal{D}) = \log p(\mathcal{D} \mid \mathcal{G}) + \log p(\mathcal{G})
    \]
Score-based methods

- Idea: treat the problem of learning the structure of the DAG as a **model selection** problem

\[
\max_{\mathcal{G} \in \text{DAG}} \text{score}(\mathcal{G} | \mathcal{D})
\]

- Recall: choices of scores
  - **Likelihood score:**
    \[
    \text{score}_L(\mathcal{G} | \mathcal{D}) = \log p(\mathcal{D} | \hat{\theta}^\text{MLE}_G, \mathcal{G})
    \]
  - **Bayesian score:**
    \[
    \text{score}_B(\mathcal{G} | \mathcal{D}) = \log p(\mathcal{D} | \mathcal{G}) + \log p(\mathcal{G})
    \]
  - **Bayesian Information Criterion (BIC):**
    \[
    \text{score}_{BIC}(\mathcal{G} | \mathcal{D}) = \log p(\mathcal{D} | \hat{\theta}^\text{MLE}_G, \mathcal{G}) - \frac{\log N}{2} \text{Dim}[\mathcal{G}]
    \]
Score-based methods

\[ \max_{G \in \text{DAG}} \text{score}(G \mid \mathcal{D}) \]

- How to search over the space of DAGs?

- The number of DAGs over \( n \) nodes is super-exponential in \( n \): \( 2^{\Theta(n^2)} \).
Score-based methods

\[
\max_{G \in \text{DAG}} \text{score}(G \mid D)
\]

- How to search over the space of DAGs?
- The number of DAGs over \( n \) nodes is super-exponential in \( n \): \( 2^{\Theta(n^2)} \).

**Theorem**

Let \( G_{\leq d} = \{G \text{ a DAG} \mid \text{every node has at most } d \text{ parents}\} \). Finding a DAG in \( G_{\leq d} \) that maximizes a score is NP-hard for \( d \geq 2 \).
Score-based methods

\[
\max_{G \in \text{DAG}} \text{score}(G \mid D)
\]

- How to search over the space of DAGs?
- The number of DAGs over \( n \) nodes is super-exponential in \( n \): \( 2^{\Theta(n^2)} \).

**Theorem**

Let \( G_{\leq d} = \{ G \text{ a DAG} \mid \text{every node has at most } d \text{ parents} \} \). Finding a DAG in \( G_{\leq d} \) that maximizes a score is **NP-hard** for \( d \geq 2 \).

- Heuristic solutions:
  - **Greedy algorithms**: Hill climbing, GES
  - **Genetic algorithms**
  - **Constrained continuous optimization**: NOTEARS, Gran-DAG, DCDI, etc...