# Conditional Independence: Representation and Learning

Generative and Deep Learning (GDL)

Riccardo Massidda (riccardo.massidda@unipi.it)
Davide Bacciu (davide.bacciu@unipi.it)


UNIVERSITÀ DI PISA

# Graphical Models: Probability and Causality

- Bayesian Networks (Tuesday 24th, **today**!)
  - Compact representation of joint probabilities
  - Plate Notation
  - Local Markov Property
  - Ancestral Sampling
- d-separation, Markov blankets (Wednesday 25th)
- Graphical Causal Models (Thursday 26th)
- Probabilistic and Causal Structure Learning (Tuesday 3rd)

# Representing Joint Distributions

- The main goal of **probabilistic modeling** is to define models able to represent the **joint distribution** of a set of variables.
- Probabilistic models enable
  - **Sampling** new instances
  - Inferencing values of **hidden** variables
  - Estimating the **likelihood** of a configuration
  - ...

# Representing Joint Distributions

- Assume N discrete random variables with k distinct values.
- How many parameters in the **joint probability distribution**?

| $Y_1$ | $Y_2$ | $Y_3$ | $P(Y_1, Y_2, Y_3)$ |
|---|---|---|---|
| 0 | 0 | 0 | 0.03 |
| 0 | 0 | 1 | 0.12 |
| 0 | 1 | 0 | 0.31 |
| ⋮ | ⋮ | ⋮ | ⋮ |
| 1 | 1 | 1 | 0.04 |

$$k^N - 1$$

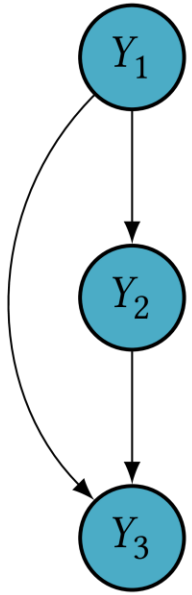
# Representing Joint Distributions

- What if we compute the probability **one variable** at the time?
- We can exploit the **chain rule** to decompose the joint.

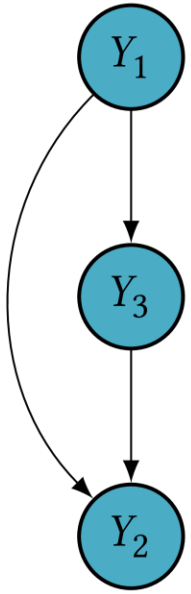
$$
\begin{aligned}
P(Y_1, Y_2, Y_3) &= P(Y_1)P(Y_2 \mid Y_1)P(Y_3 \mid Y_1, Y_2) \\
&= P(Y_2)P(Y_1 \mid Y_2)P(Y_3 \mid Y_1, Y_2) \\
&= \dots \\
&= P(Y_3)P(Y_2 \mid Y_3)P(Y_1 \mid Y_2, Y_3).
\end{aligned}
$$

# Representing Joint Distributions

- The **order** of the variables can be represented by **directed graphs**.

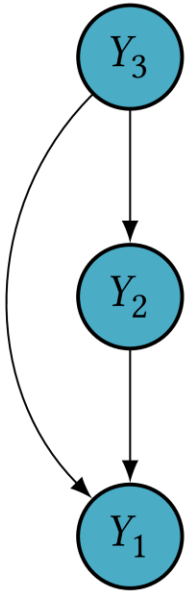



$P(Y_1)P(Y_2 \mid Y_1)P(Y_3 \mid Y_1, Y_2)$




$P(Y_1)P(Y_3 \mid Y_1)P(Y_2 \mid Y_1, Y_3)$

...




$P(Y_3)P(Y_2 \mid Y_3)P(Y_1 \mid Y_2, Y_3)$

# Representing Joint Distributions

- Decomposing the joint with the **chain rule** reduces the **number of parameters**?
- No! 🥶

$$P(Y_1, Y_2, Y_3) = \underbrace{P(Y_1)}_{1}\underbrace{P(Y_2 \mid Y_1)}_{2}\underbrace{P(Y_3 \mid Y_1, Y_2)}_{4}$$

$$\sum_{i=0}^{N-1}(k-1)k^i = k^N - 1$$

# Marginal and Conditional Independence

- Two random variables X and Y are **independent** if knowledge about X does not change the uncertainty about Y and vice versa

$$I(X,Y) \iff X \perp Y \iff P(X,Y) = P(X \mid Y)P(Y)$$
$$= P(Y \mid X)P(X) = P(X)P(Y).$$

# Representing Joint Distributions

- When variables are **independent**, we only need Nk parameters.

$$P(Y_1, Y_2, Y_3) = P(Y_1)P(Y_2 \mid Y_1)P(Y_3 \mid Y_1, Y_2)$$
$$= P(Y_1)P(Y_2)P(Y_3)$$

$\underbrace{P(Y_1)}_{1}\ \underbrace{P(Y_2)}_{1}\ \underbrace{P(Y_3)}_{1}$

# Marginal and Conditional Independence

- Two random variables X and Y are **conditionally independent** given Z if knowledge about X does not change the uncertainty about Y and vice versa on the conditional distribution

$$I(X, Y \mid Z) \iff X \perp Y \mid Z \iff P(X, Y \mid Z) = P(X \mid Y, Z)P(Y \mid Z)$$
$$= P(Y \mid X, Z)P(X \mid Z)$$
$$= P(X \mid Z)P(Y \mid Z).$$

# Representing Joint Distributions

- Conditional independences reduce the **number of parameters**
- Yes! 🥳

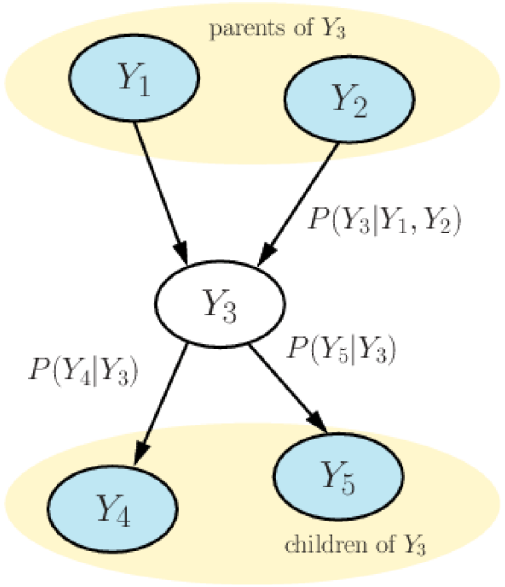
$$Y_1 \perp Y_3 \mid Y_2$$

$$\implies P(Y_1, Y_2, Y_3) = P(Y_1)P(Y_2 \mid Y_1)P(Y_3 \mid Y_1, Y_2)$$

$$= P(Y_1)P(Y_2 \mid Y_1)P(Y_3 \mid Y_2)$$

$$\underbrace{P(Y_1)}_{1}\ \underbrace{P(Y_2 \mid Y_1)}_{2}\ \underbrace{P(Y_3 \mid Y_2)}_{2}$$

# Bayesian Network

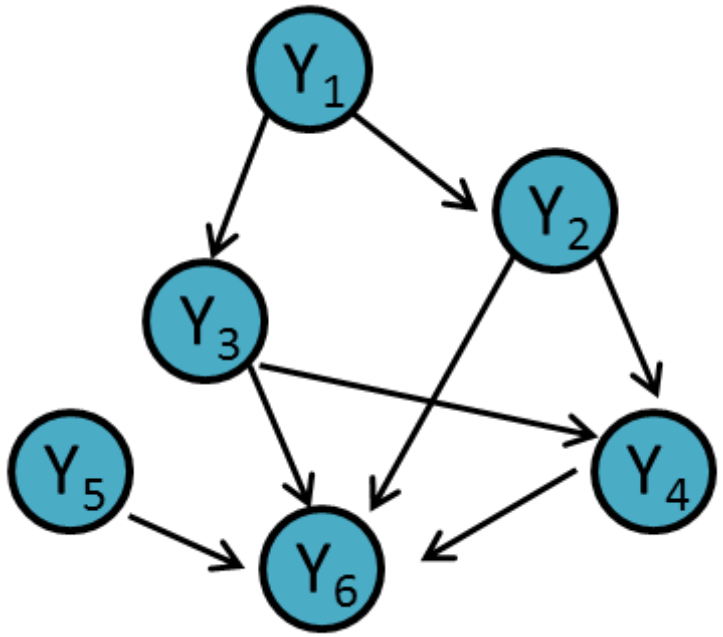



- Directed Acyclic Graph (DAG) $\mathcal{G} = (\mathcal{V}, \mathcal{E})$
- Nodes $v \in \mathcal{V}$ represent random variables
  - Shaded ⇒ observed
  - Empty ⇒ un-observed
- Edges $e \in \mathcal{E}$ describe the conditional independence relationships

In a Bayesian Network, the **joint probability** is decomposed as

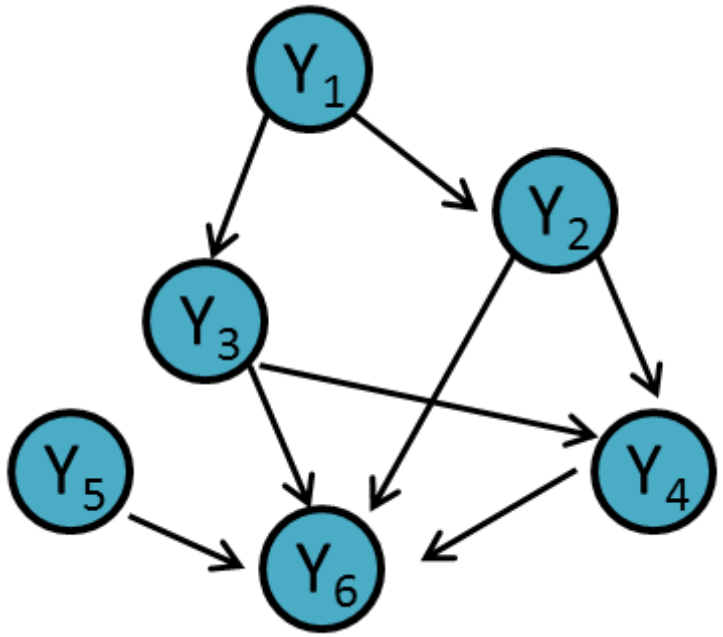
$$P(Y_1, \dots, Y_N) = \prod_{i=1}^{N} P(Y_i \mid \mathrm{pa}(Y_i))$$

# Discrete Bayesian Networks




- In a **discrete** BN parameters are represented by Conditional Probability Tables, or **CPT**.
- Let $\mathrm{L}$ be the **maximum number of ingoing edges** in a Bayes Net.
- Then, the number of parameters is **at most** $\mathrm{N}\cdot(\mathrm{k}\text{-}1)^{\mathrm{L}}$
- ⇒ The **sparser** the network, the less "complex" the parameters.

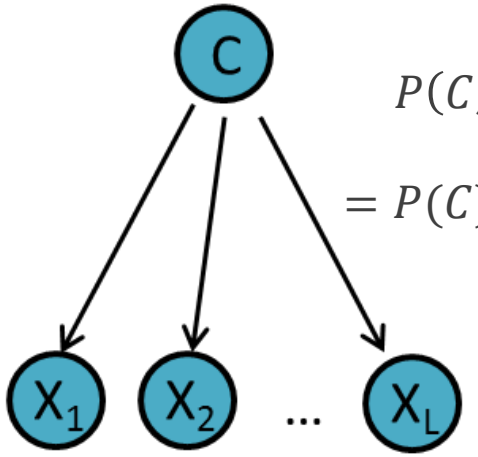
# (Causal?) Bayesian Networks




- Are edge relations **causal**?
- In general **no**, a Bayesian Network represent **statistical dependence** relations.
- However, they **might** coincide with causal dependence under further **assumptions**.
  - Focus of the upcoming "**Causality**" lecture!

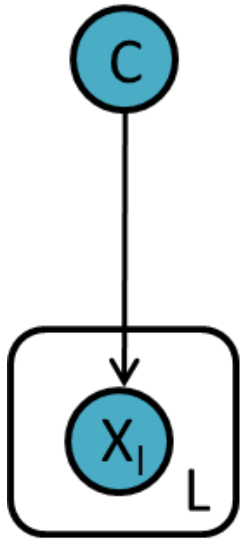
# Compact Graphical Representation

If the same **dependencies are replicated** over different variables, we can compactly represent it by **plate notation**.

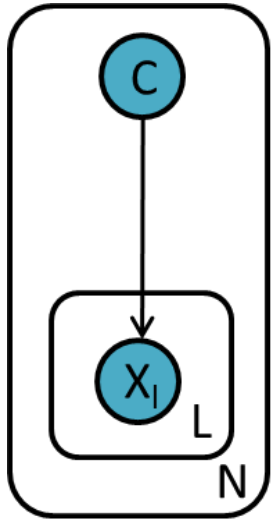



$$P(C, Y_1, \dots, Y_N) = P(C) \prod_{i=1}^{N} P(Y_i|C)$$

The Naive Bayes Classifier

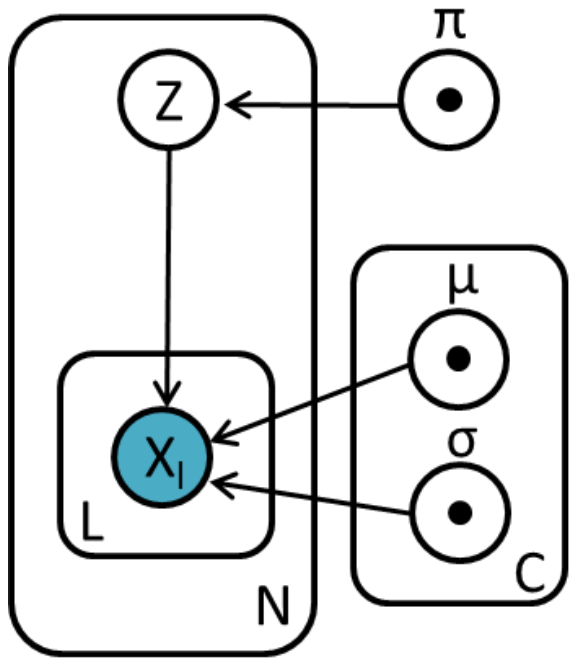



Replication for $L$ attributes




Replication for $N$ data samples

# Full Plate Notation




Gaussian Mixture Model

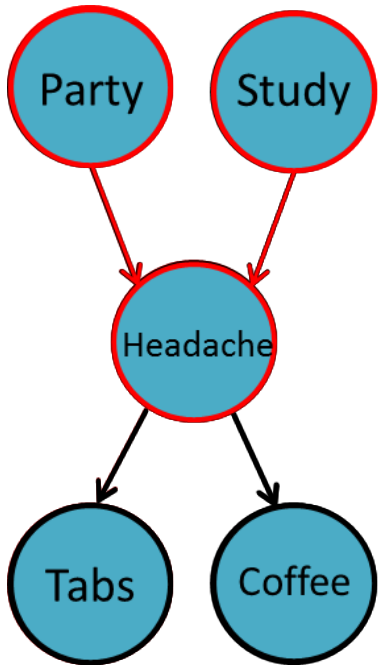
- Boxes denote replication for a number of times denoted by the letter in the corner
- Shaded nodes are observed variables
- Empty nodes denote un-observed latent variables
- Black seeds (optional) identify model parameters
  - $\pi \rightarrow$ multinomial prior distribution
  - $\mu \rightarrow$ means of the $C$ Gaussians
  - $\sigma \rightarrow$ std of the $C$ Gaussians

# Local Markov Property

**Definition (Local Markov property)**

Each node / random variable is conditionally independent of all its non-descendants given a joint state of its parents

$$Y_v \perp Y_{V \setminus \mathrm{ch}(v)} \mid Y_{pa(v)} \text{ for all } v \in V$$




$Party$ and $Study$ are marginally independent

- $Party \perp Study$

However, local Markov property does not support

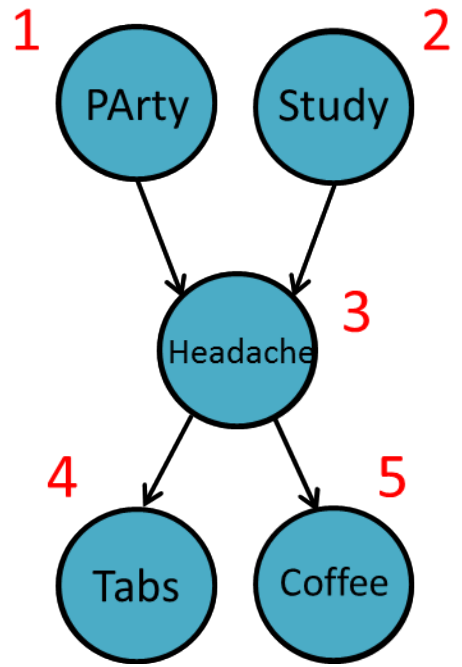
- $Party \perp Study \mid Headache$
- $Tabs \perp Party$

But $Party$ and $Tabs$ are independent given $Headache$

# Joint Probability Factorization

An application of **Chain rule** and **Local Markov Property**

1. Pick a topological ordering of nodes
2. Apply chain rule following the order
3. Use the conditional independence assumptions




$$P(PA, S, H, T, C) =$$
$$P(PA) \cdot P(S|PA) \cdot P(H|S, PA) \cdot P(T|H, S, PA) \cdot P(C|T, H, S, PA)$$
$$= P(PA) \cdot P(S) \cdot P(H|S, PA) \cdot P(T|H) \cdot P(C|H)$$

# (Ancestral) Sampling in Bayesian Networks

A BN describes a generative process for observations

1. Pick a topological ordering of nodes
2. Generate data by sampling from the local conditional probabilities following this order

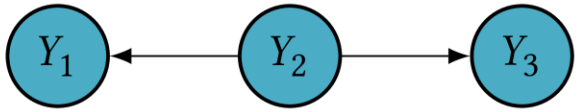
Generate $i$-th sample for each variable $PA, S, H, T, C$

1. $pa_i \sim P(PA)$
2. $s_i \sim P(S)$
3. $h_i \sim P(H|S = s_i, PA = pa_i)$
4. $t_i \sim P(T|H = h_i)$
5. $c_i \sim P(C|H = h_i)$

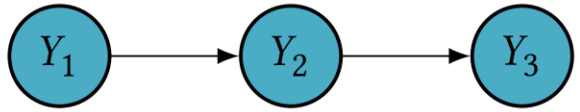
# Fundamental BN structures

There exist **three fundamental substructures** that determine the conditional independence relationships in a Bayesian Network.

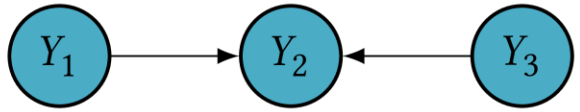
- **Tail-to-Tail** (Fork, "Common Cause")



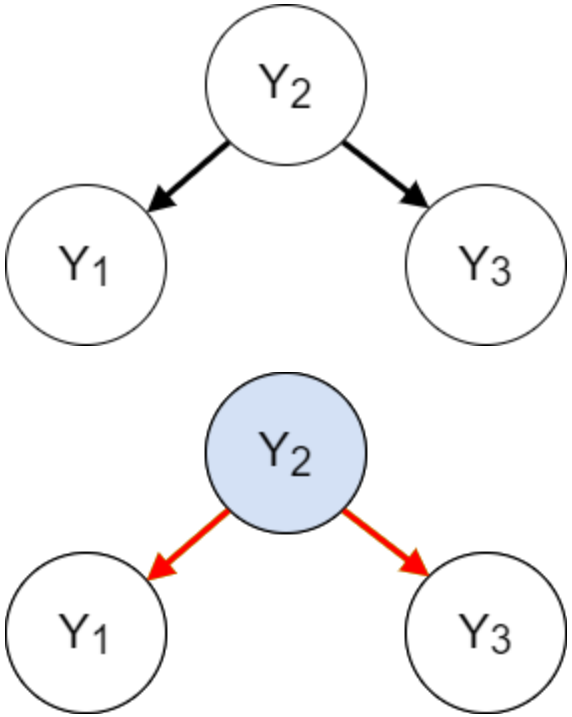

- **Head-to-Tail** (Chain, "Causal Chain")




- **Head-to-Head** (Collider, "Common Effect")



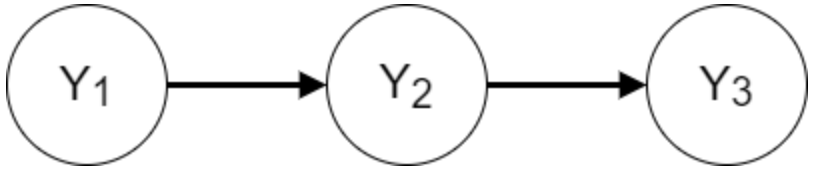
# Tail-to-Tail Connections

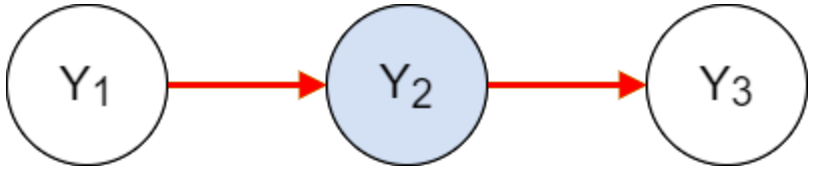



- Corresponds to
$$P(Y_1, Y_3|Y_2)P(Y_2) = P(Y_1|Y_2)P(Y_3|Y_2)P(Y_2)$$
- If $Y_2$ is unobserved then $Y_1$ and $Y_3$ are marginally dependent
$$Y_1 \not\perp Y_3$$
- If $Y_2$ is observed then $Y_1$ and $Y_3$ are conditionally independent
$$Y_1 \perp Y_3|Y_2$$

When $Y_2$ in observed is said to **block the path** from $Y_1$ to $Y_3$

# Head-to-Tail Connections




Observed $Y_2$ **blocks the path** from $Y_1$ to $Y_3$

- Corresponds to

$$\mathrm{P}(\mathrm{Y}_1, \mathrm{Y}_2, \mathrm{Y}_3) = \mathrm{P}(\mathrm{Y}_1)P(\mathrm{Y}_2|\mathrm{Y}_1)\mathrm{P}(\mathrm{Y}_3|\mathrm{Y}_2)$$
$$= \mathrm{P}(\mathrm{Y}_1|\mathrm{Y}_2)\mathrm{P}(\mathrm{Y}_3|\mathrm{Y}_2)\mathrm{P}(\mathrm{Y}_2)$$

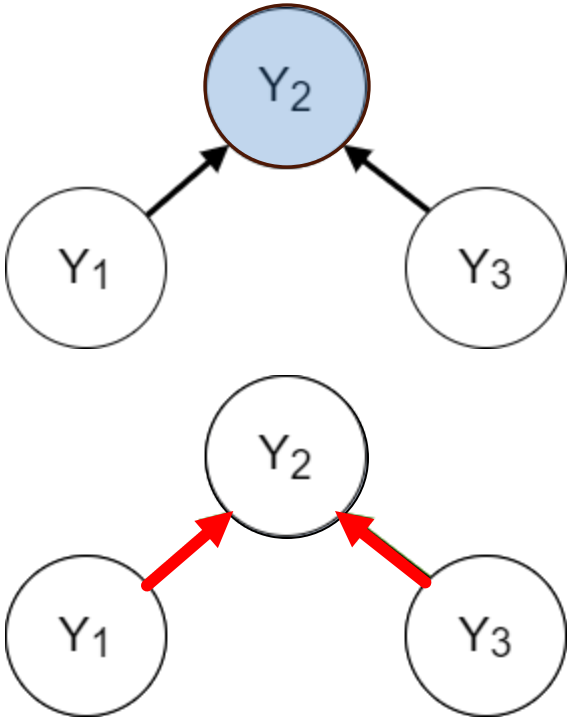
- If $Y_2$ is unobserved then $Y_1$ and $Y_3$ are marginally dependent Type equation here.

$$Y_1 \not\perp Y_3$$

- If $Y_2$ is observed then $Y_1$ and $Y_3$ are conditionally independent

$$Y_1 \perp Y_3|Y_2$$

# Head-to-Head Connections




- Corresponds to

$$P(Y_1, Y_2, Y_3) = P(Y_1)P(Y_3)P(Y_2|Y_1, Y_3)$$

- If $Y_2$ is observed then $Y_1$ and $Y_3$ are conditionally dependent

$$Y_1 \perp\!\!\!\not\perp Y_3 | Y_2$$

- If $Y_2$ is unobserved then $Y_1$ and $Y_3$ are marginally independent

$$Y_1 \perp Y_3$$

If any $Y_2$ **descendants** is observed it **unlocks the path**

# Graphical Models: Probability and Causality

- Bayesian Networks (Tuesday 24th)
- d-separation, Markov blankets (Wednesday 25th, **next**!)
  - d-separation
  - Markov Property and Faithfulness
  - Markov Blanket
- Graphical Causal Models (Thursday 26th)
- Structure Learning and Causal Discovery (Tuesday 3rd)