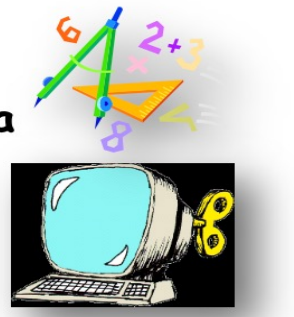
Bayesian networks

Mahdi Shafiee Kamalabad
Utrecht University

In a nutshell

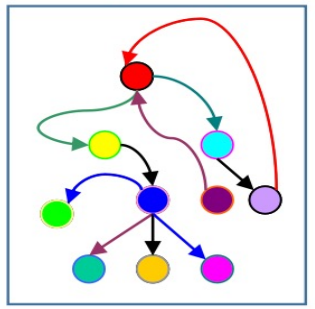
Raw data

Cleaned data



Machine Learning

Statistical Methods



Network inference

We aim to estimate the network structure, determining what depends on what and how, in the form of a *directed* network.

Data structure

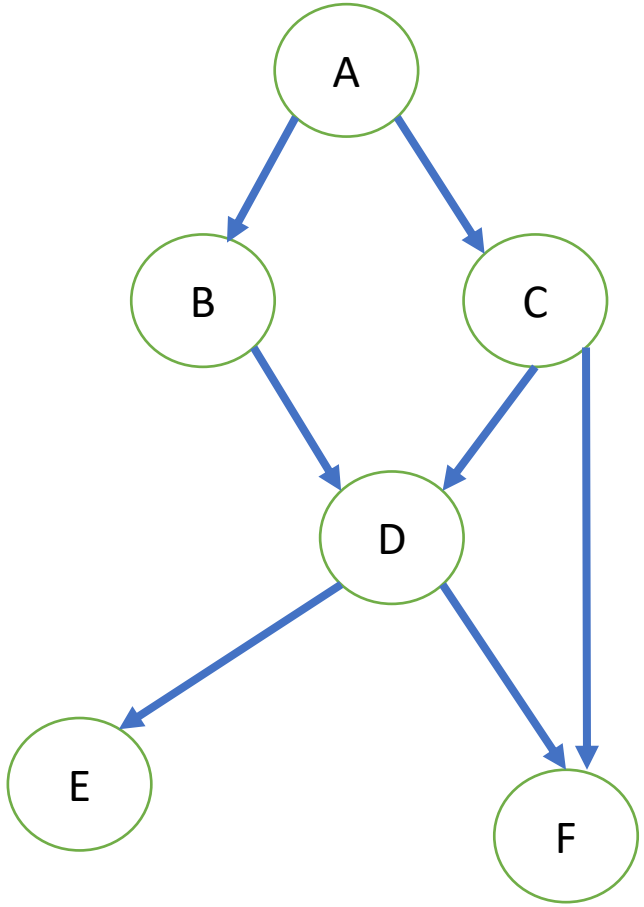
Data: Train Use Survey

The data structure remains the same as before, as we discussed that today morning:

"Age"	"Residence"	"Education"	"Occupation"	"Sex"	"Travel"
"adult"	"big"	"high"	"emp"	"F"	"car"
"adult"	"small"	"uni"	"emp"	"M"	"car"
"adult"	"big"	"uni"	"emp"	"F"	"train"
"adult"	"big"	"high"	"emp"	"M"	"car"
"adult"	"big"	"high"	"emp"	"M"	"car"
"adult"	"small"	"high"	"emp"	"F"	"train"
"adult"	"big"	"high"	"emp"	"F"	"car"
"young"	"big"	"uni"	"emp"	"F"	"train"

Bayesian networks (BNs)

- **Marriage between** graph theory and probability theory.
- **Nodes** represent variables and **edges** represent (conditional) dependence between variables.
- Bayesian Networks are **directed networks.**
- **If the directions of the dependencies are important in your project, go for a Bayesian network.**

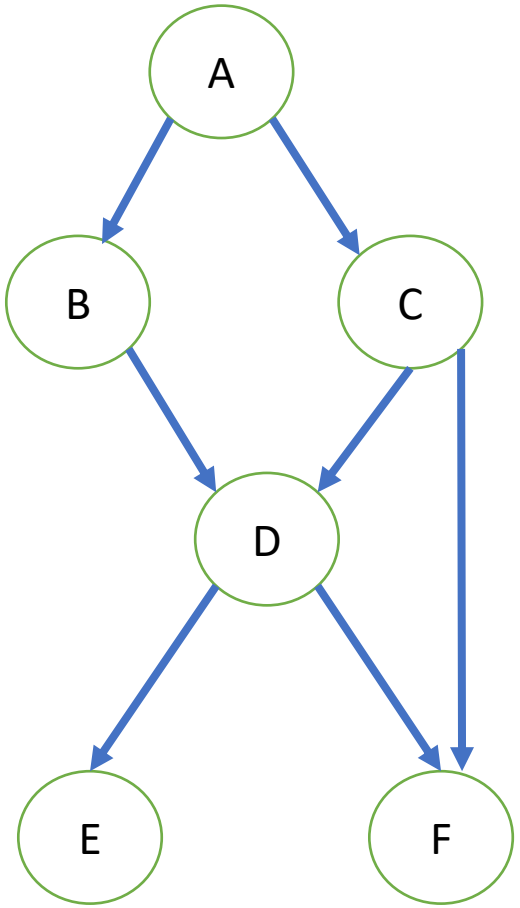


Bayesian networks

- **The first component** of a BN is a graph.

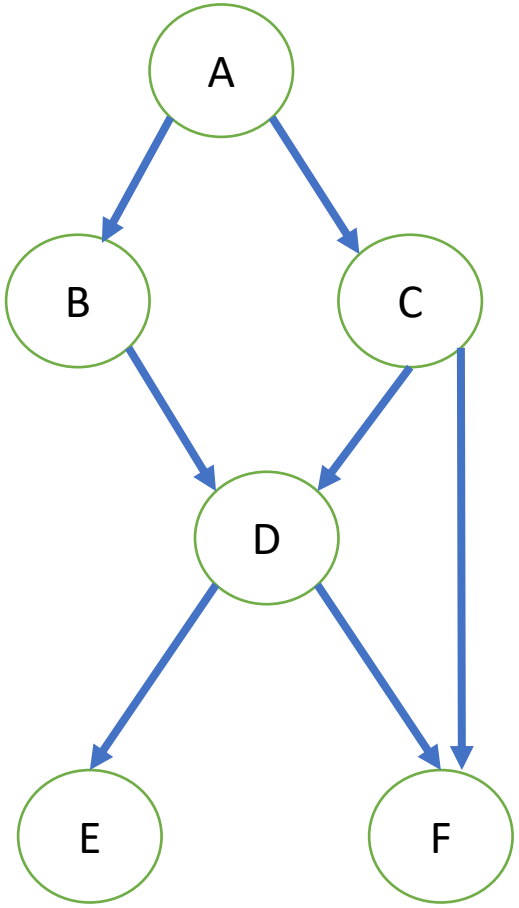
  A graph G is a mathematical object with:
- a set of nodes $V = \{v_1, \ldots, v_N\}$;
- a set of **arcs** $\boldsymbol{A}$ which are identified by pairs for nodes in V, e.g. $a_{ij} = (v_i, v_j)$.

- **The second component** of a BN is the probability distribution $P(X)$, should be such that the BN:
- can be learned efficiently from data;
- is flexible (distributional assumptions should not be too strict);
- is easy to query to perform inference.



Directed Acyclic Graph (DAG) in static Bayesian networks

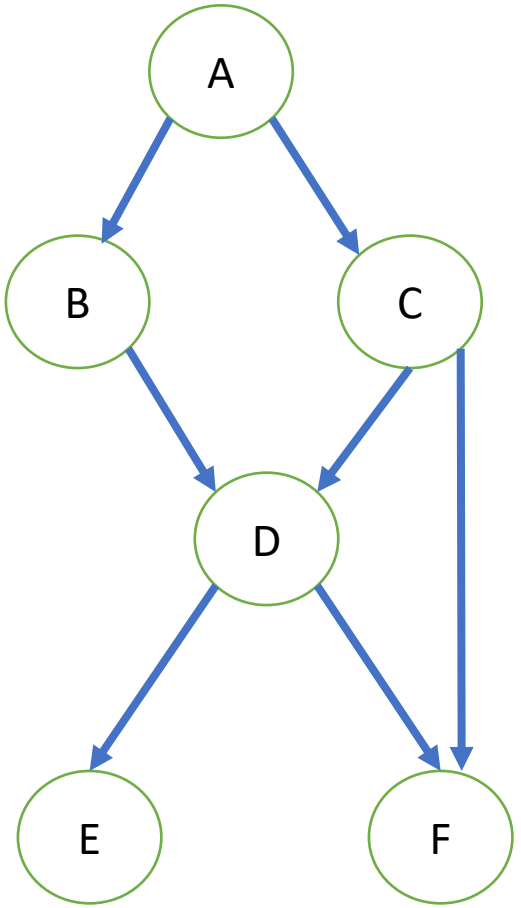
- Contains only **directed edges.**
- **No cycles**: you can’t return to the same node by following arrows.
- Can be used with **discrete or continuous variables.**
- Often used to represent **causal** or **statistical** dependencies.



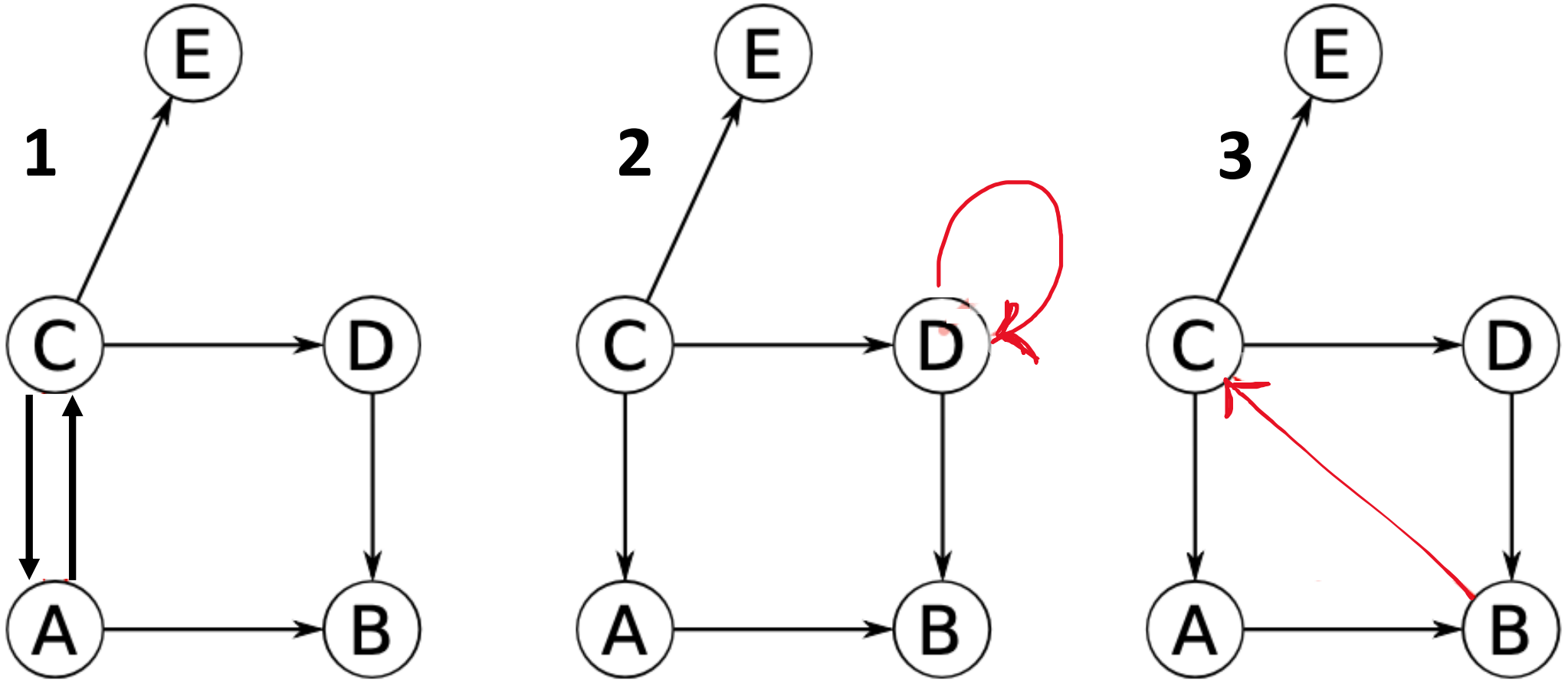
Interpretation of edges in BNs

- If there is an edge from one variable to another, the later depends on the former.

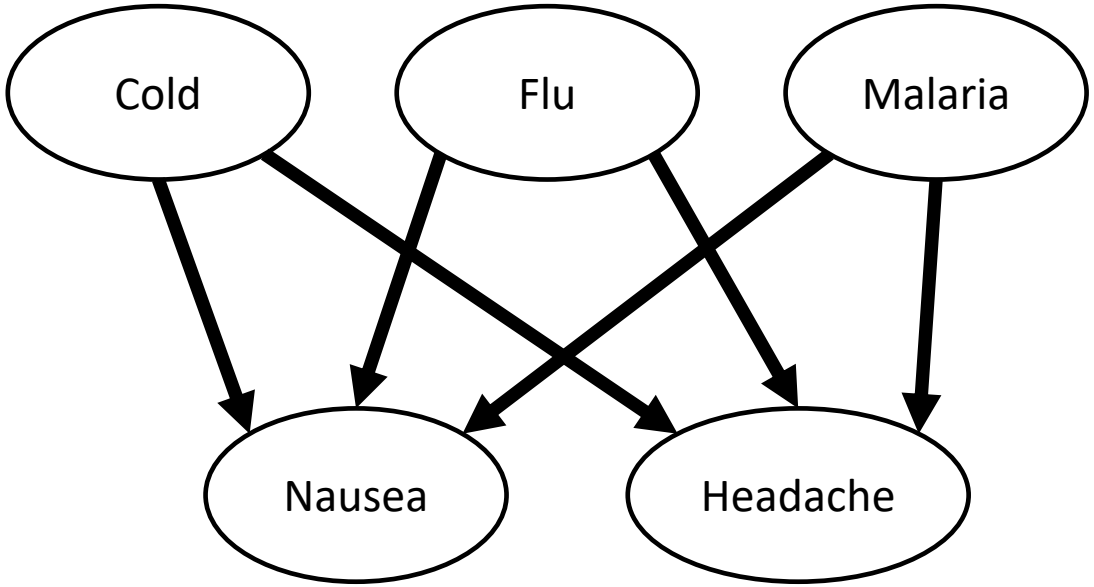
- The absence of an edge **does not always** mean independence. To determine conditional independence, we use the rules of **d-separation**. We will discuss about this later.



DAG: Which one is Directed acyclic graph(No loops/no cycle)?



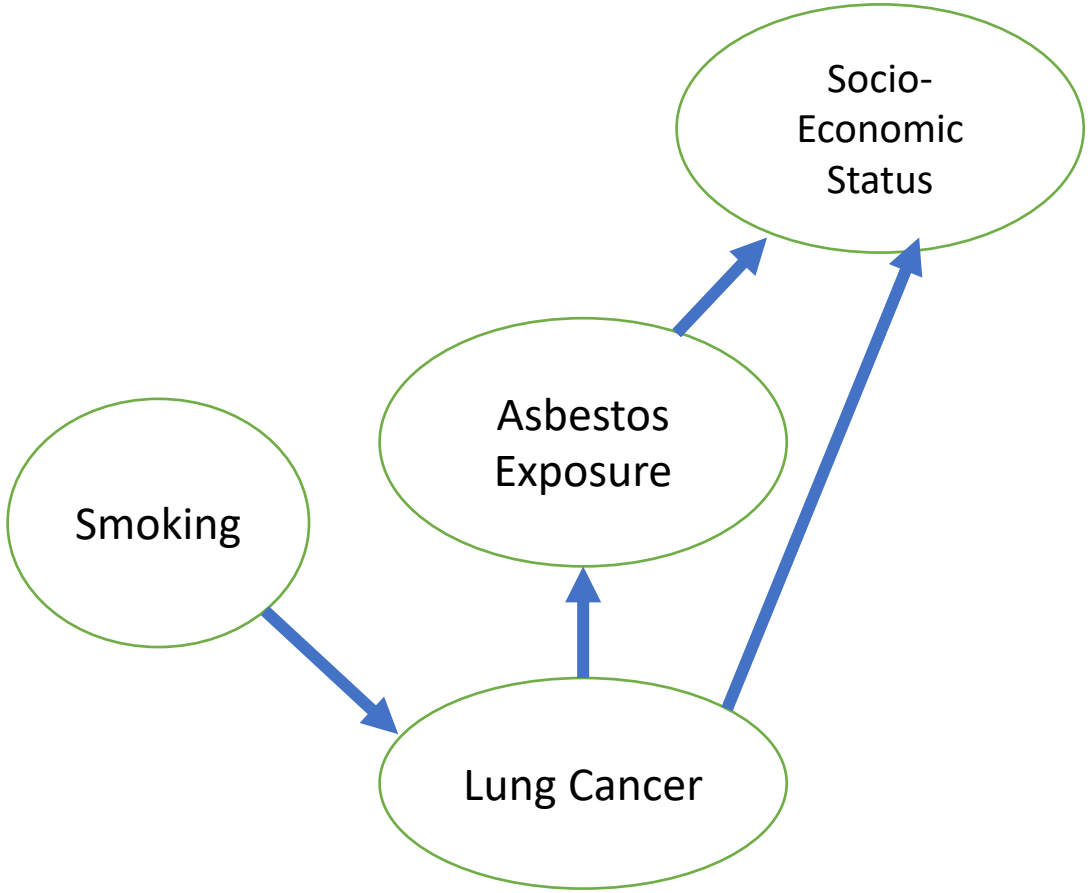
Example 1:



- A Bayesian network could represent the **probabilistic relationships** between **diseases** and **symptoms**.

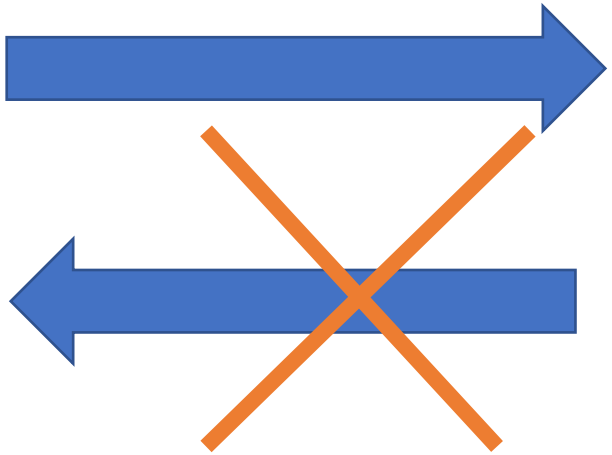
- Bayesian network with causes (diseases) Cold, Flu, and Malaria and effects (symptoms) Nausea and Headache.

Do you agree?


Socio-
Economic
Status
Asbestos
Exposure
Smoking
Lung Cancer

Note

Causal Network

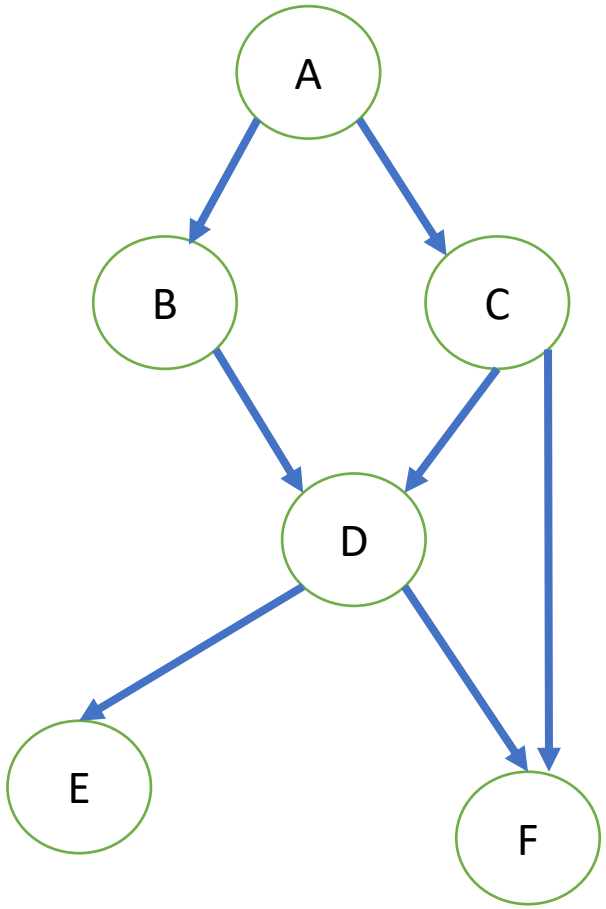


Not necessarily

Bayesian Network

Some applications of BNs

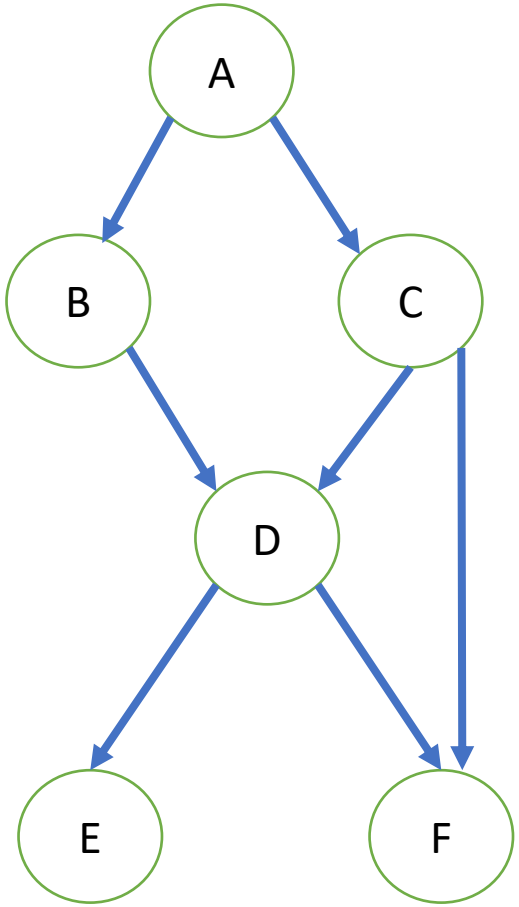
- Biology (Gene Regulatory Network, …)
- Medicine
- Document Classification. ...
- Image Processing. ...
- Spam Filter
- Psychology
- Economics
- …



See here
https://data-flair.training/blogs/bayesian-network-applications/

Bayesian networks

- A is a parent node of B
- B is a child node of A
- The parent node set of D is the set $\{B,C\}$
- D is a common child node of B and C.
- A has no parents. That is the parent set of A is the empty set.
- (B)C is (co-)parent node of D (another parent).



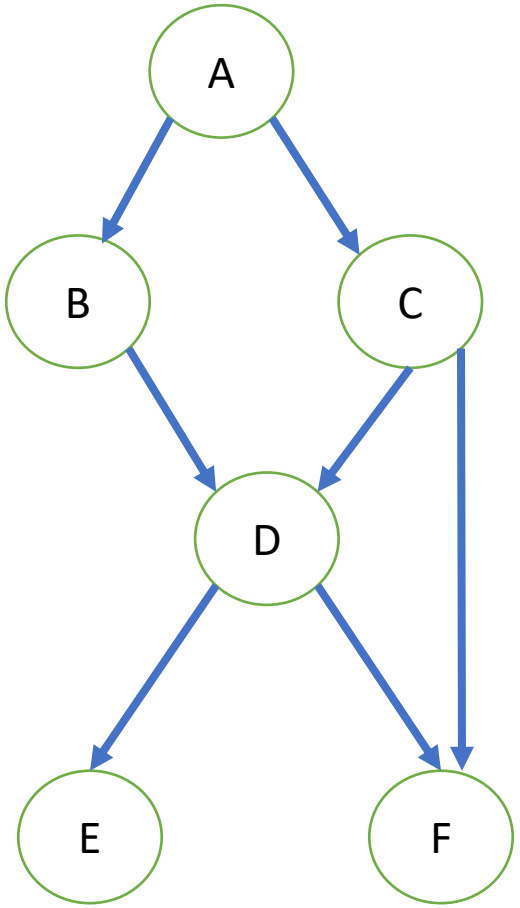
Bayesian networks

- Node X is an **ancestor** of node Y

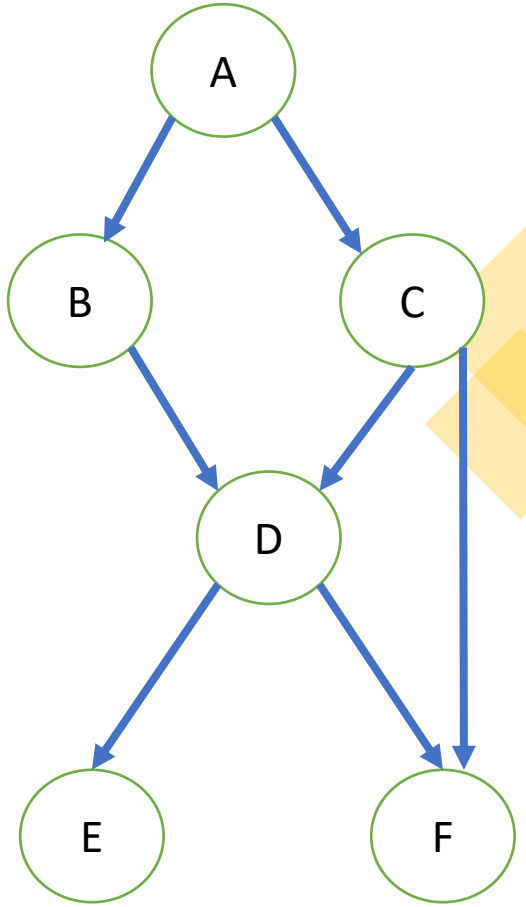
if there is a **path** of directed edges leading from X and Y:

$$\boldsymbol{X \rightarrow \cdots \rightarrow Y}$$

- Y is then called a **descendant** of X.



Bayesian networks and Markov assumption



- **Markov assumption** leads to a factorization of the joint probability distribution:

$P(A, B, C, D, E, F) =$

$$P(A)\ P(B|A)\ P(C|A)\ P(D|B, C)\ P(E|D)\ P(F|C, D)$$

$\mathbf{P(child|\ parent(s))}$

- **Markov assumption:**

Every **node** in a Bayesian network is **conditionally independent** of its non-descendants, given its parents (**only the parents**).

That is, a node only depends on its parents ,
once you know the parents, other non descendant variables don't matter.

Example: Train Use Survey

- Consider a survey whose aim is to **investigate the usage patterns** of different means of **transport**, with a **focus** on **cars** and **train**.
- Such surveys are used to assess **customer satisfaction** across different social groups, to evaluate public policies or for urban planning.
- Some other real-world examples can also be found, in **Kenett et al. (2012).**

Data: Train Use Survey

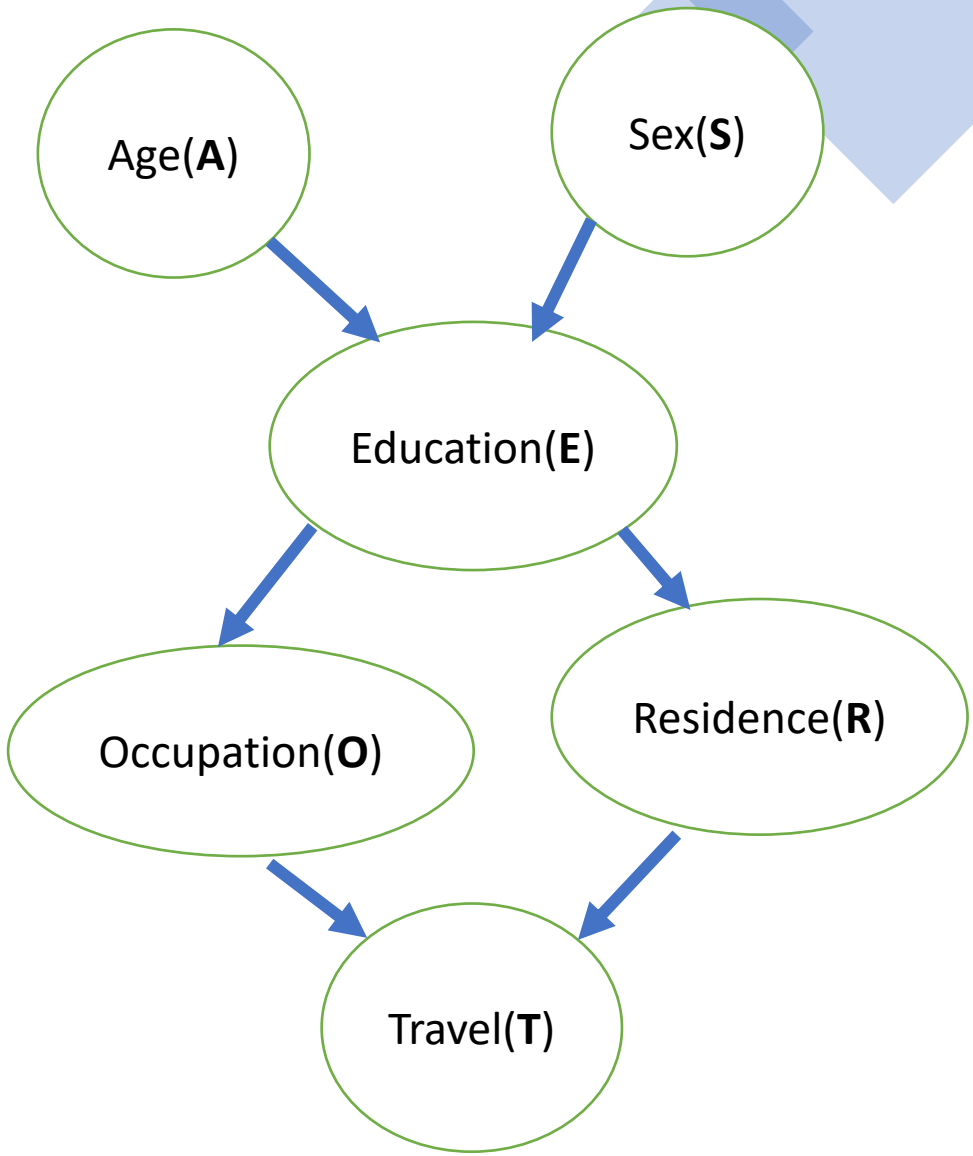
Age	Residence	Education	Occupation	Sex	Travel
"adult"	"big”	"high"	"emp”	"F”	"car"
"adult"	"small"	"uni"	"emp"	"M"	"car"
"adult"	"big"	"uni"	"emp"	"F"	"train"
"adult"	"big"	"high"	"emp"	"M"	"car"
"adult"	"big"	"high"	"emp"	"M"	"car"
"adult"	"small"	"high"	"emp"	"F"	"train"
"adult"	"big"	"high"	"emp"	"F"	"car"
"young"	"big"	"uni"	"emp"	"F"	"train"

Example: Train Use Survey

- **Age** and **sex** are not influenced by any other variable.
- **Age** and **sex** have a direct influence on **Education.**
- **Education** strongly influence both **occupation** and **residence**
- **Transports** are directly influenced by both **occupation** and **residence**.

- This **DAG** represents the **dependence relationships** between: **Age** , **Sex**, **Education**, **Occupation**, **Residence** and **Travel**.

- Later, you'll also learn how to **quantify the strength of the relationships**.



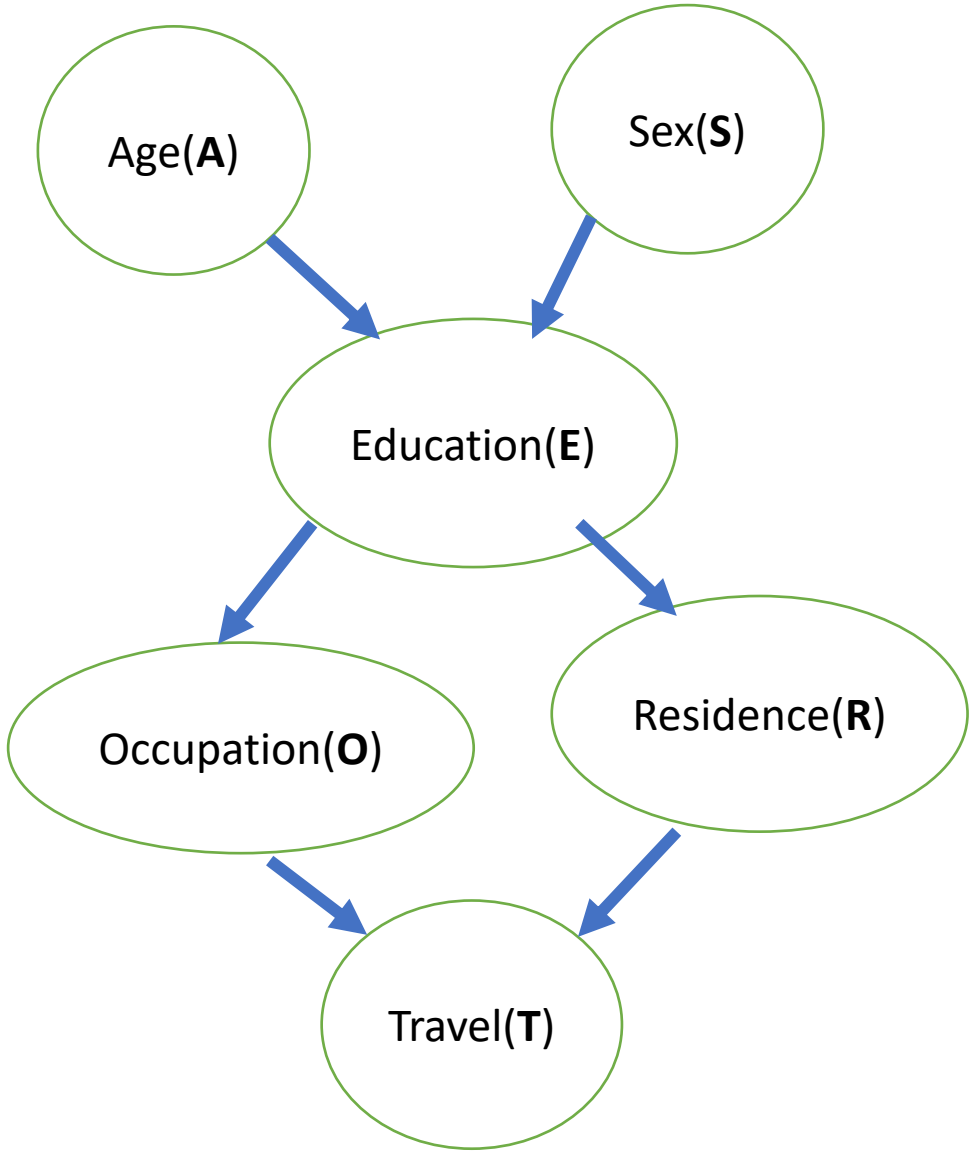
Example: Train Use Survey

The probabilistic relationship can be represented as below:

[A] [S] [E|A:S] [O|E] [R|E] [T|O:R]

[child|parents]

This is the type of representation you'll see when we use the ***"bnlearn"*** package in practice.



- Think about the application of this concept in your field for a few minutes and discuss that in pairs.

- Consider the following questions for reflection in your field:

✓What variables are present in your potential project?

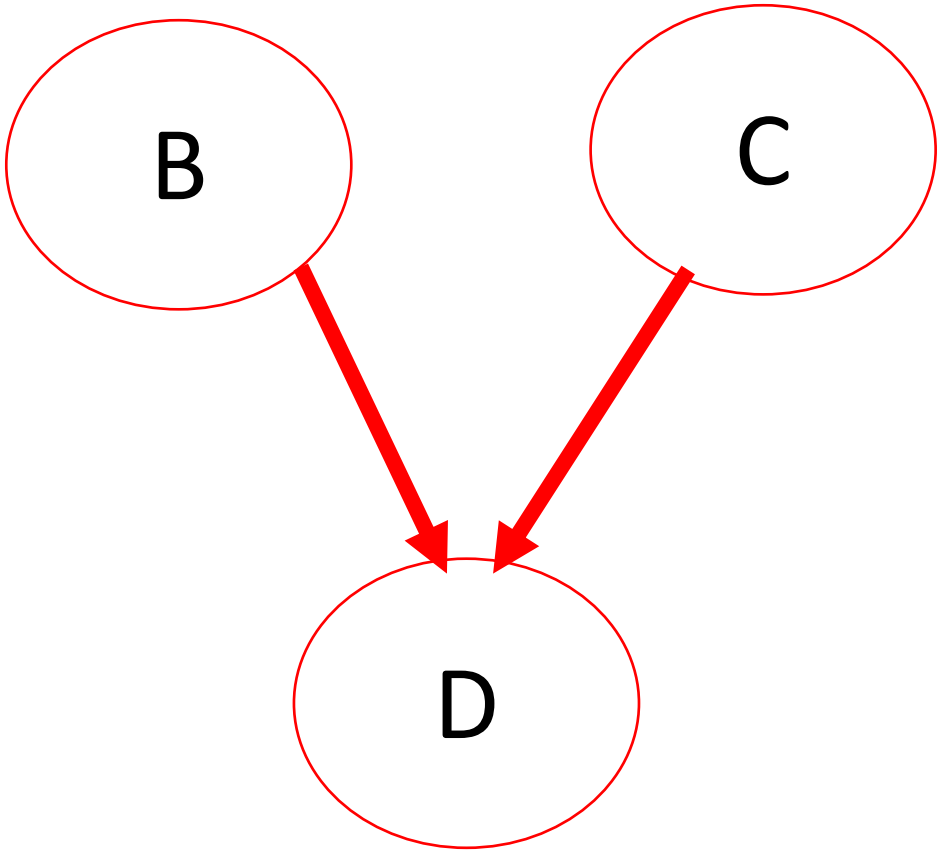
✓Why are these variables important?

✓What motivates your interest in understanding their interdependencies?

✓How does this understanding contribute to your work or goals?

Some terminologies

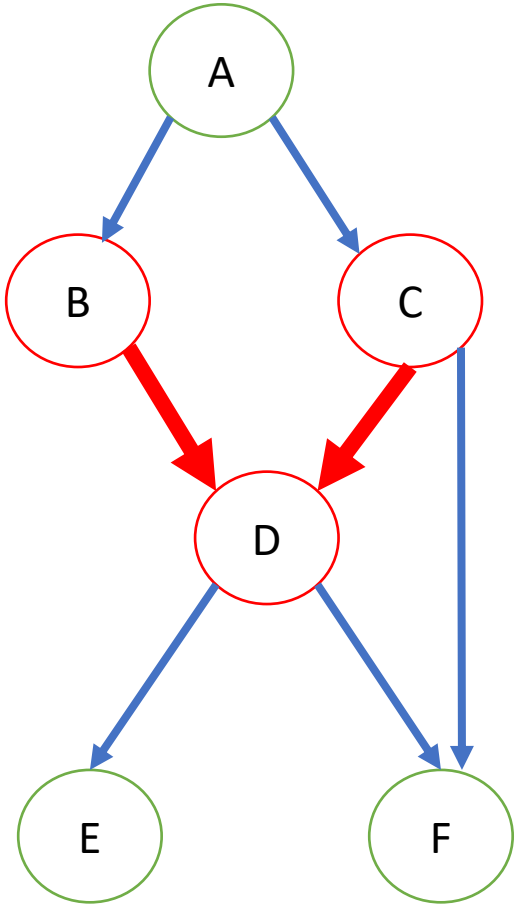
v-structure
Important for interpretation



v-structure

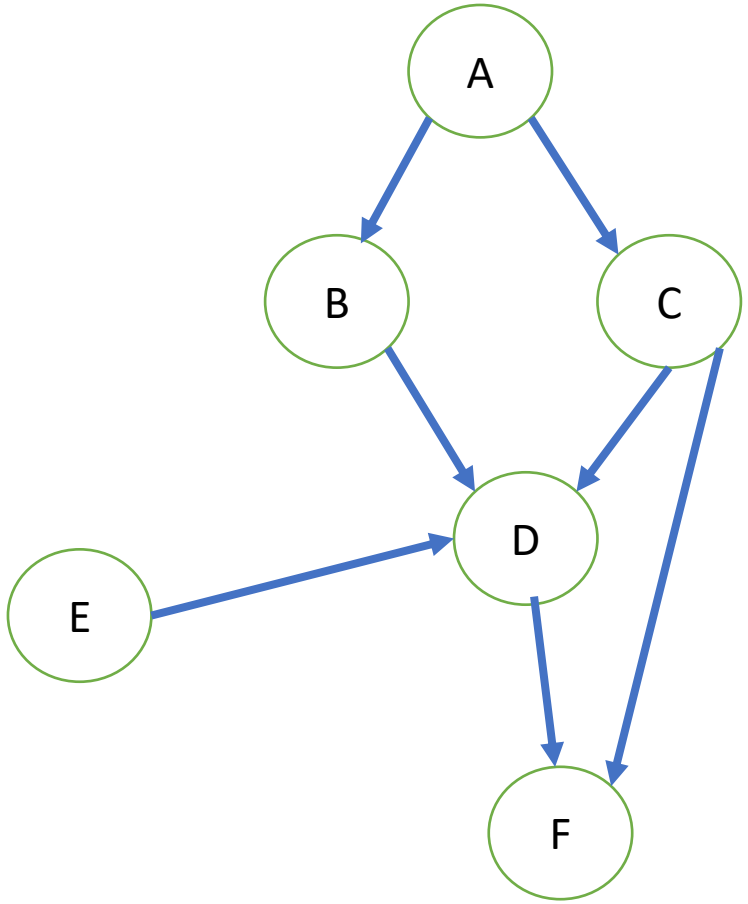
Important for interpretation

- D has the parent nodes B and C, and there is **no connection** between the nodes **B** and **C**.



v-structure

How many v-sructure do you see?



Markov Blanket

Graph **G** has **n** nodes, $\boldsymbol{X_1, \ldots, X_n}$.

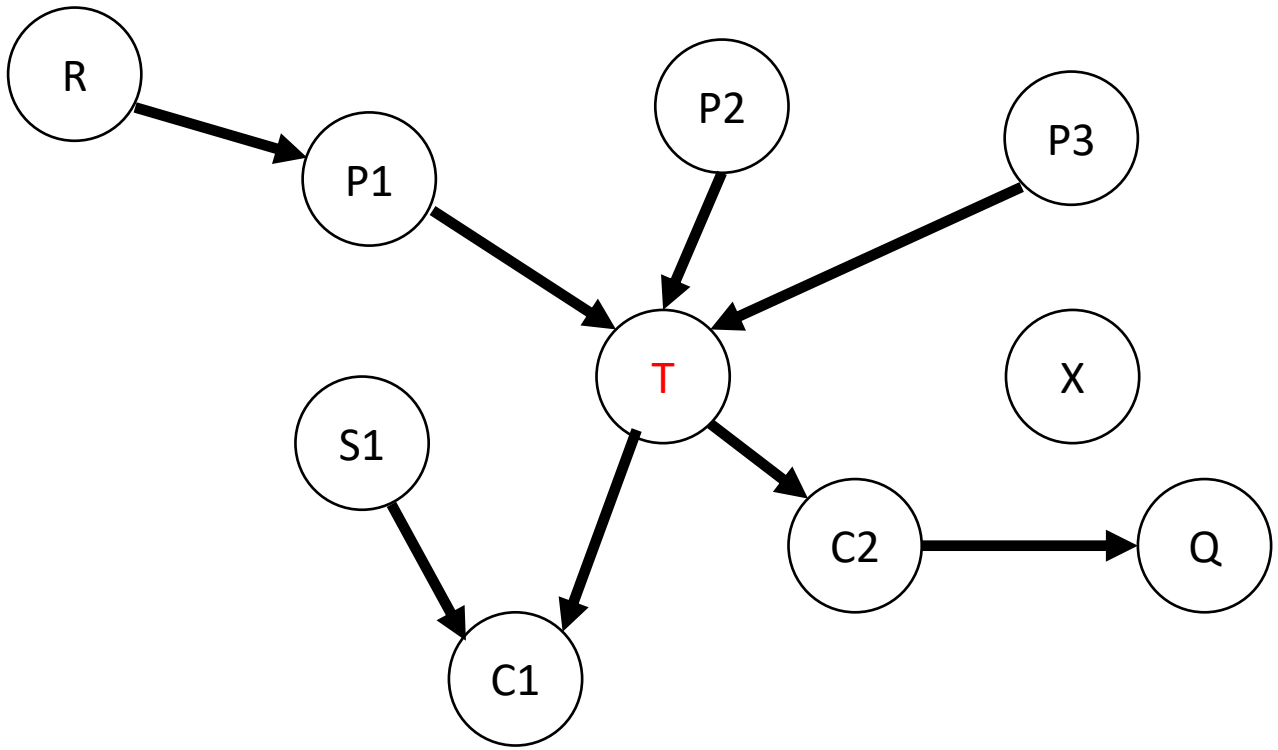
The Markov blanket of the node X_i $(i = 1, \ldots, n)$ includes:

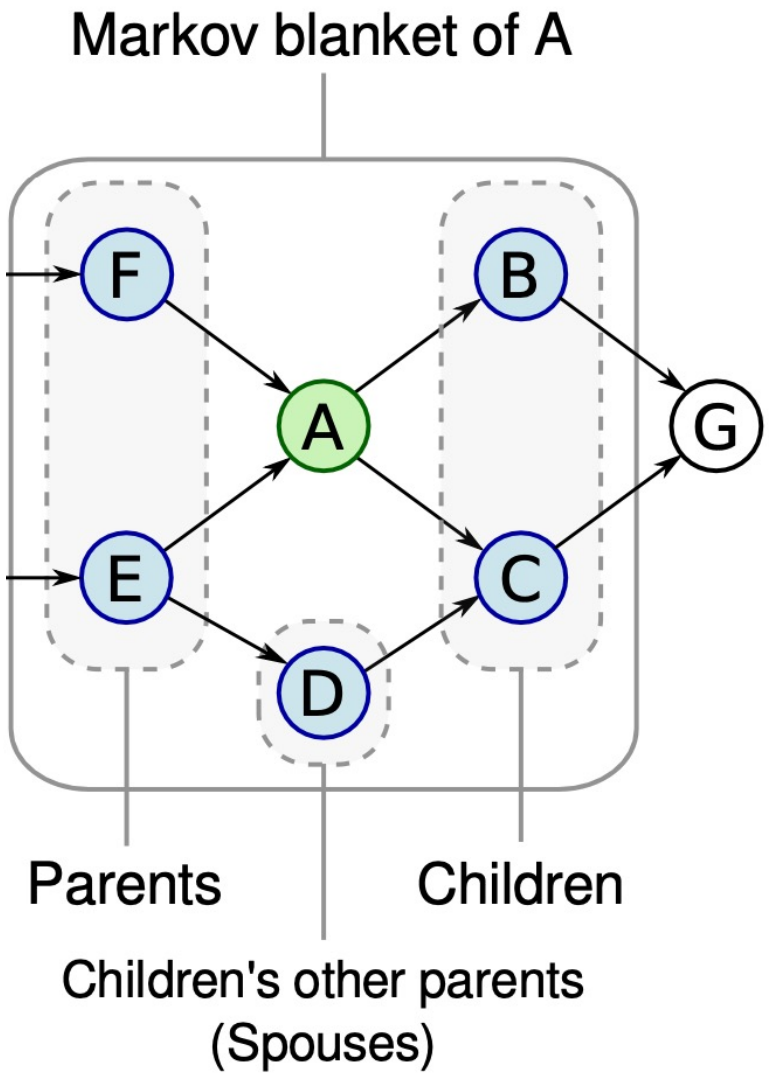
- **all parent** nodes of X_i
- **all child** nodes of X_i
- **all "co-parent** node" of X_i

We denote the **Markov Blanket** of X_i symbolically as $\mathbf{MB}(\boldsymbol{X_i})$.

Using **bnlearn** package, **mb()** functions can be used to show the Markov blankets.

Markov Blanket of T?





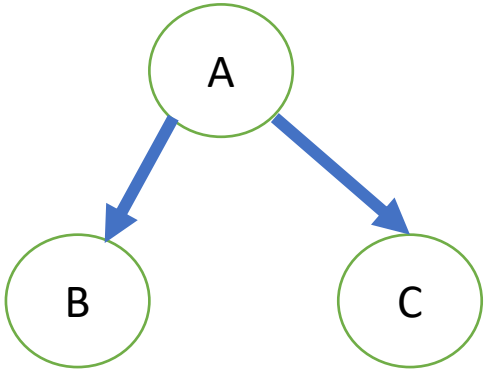
Markov blanket = minimal feature set

The **parents**, **children**, and **co-parents** of the target node contain all information you need.

Ignore everything else

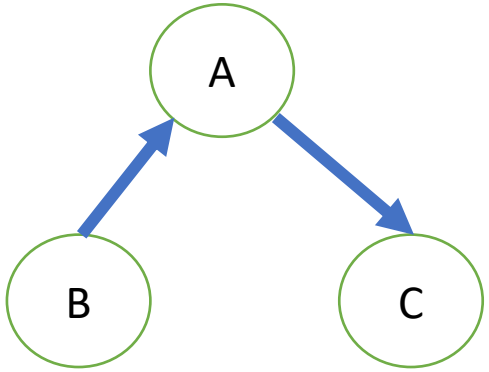
Once you condition on the Markov blanket, the target node is independent of every other variable, so the remaining features can be safely dropped.

Fundamental connections



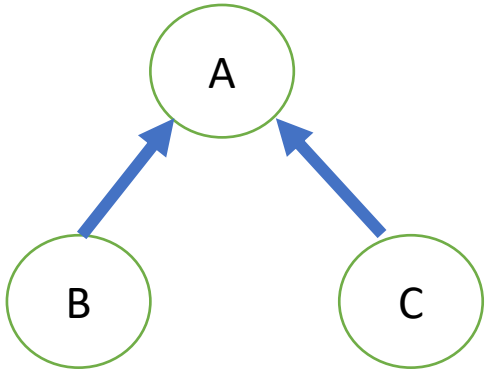
B ←A →C

Divergent connection



B →A →C

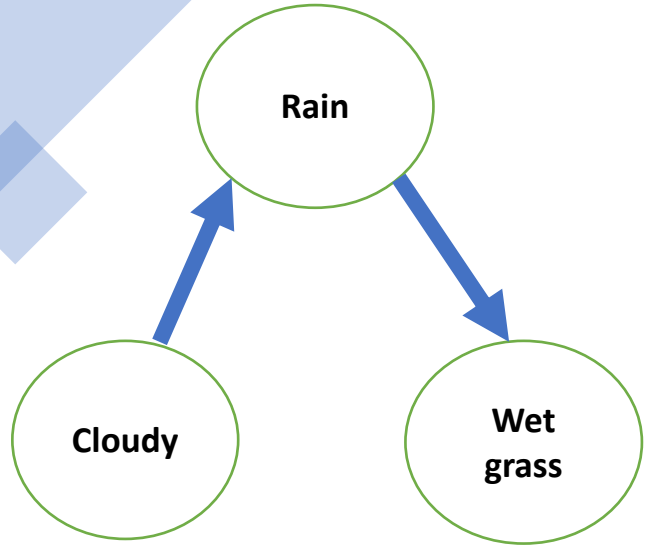
Serial connection

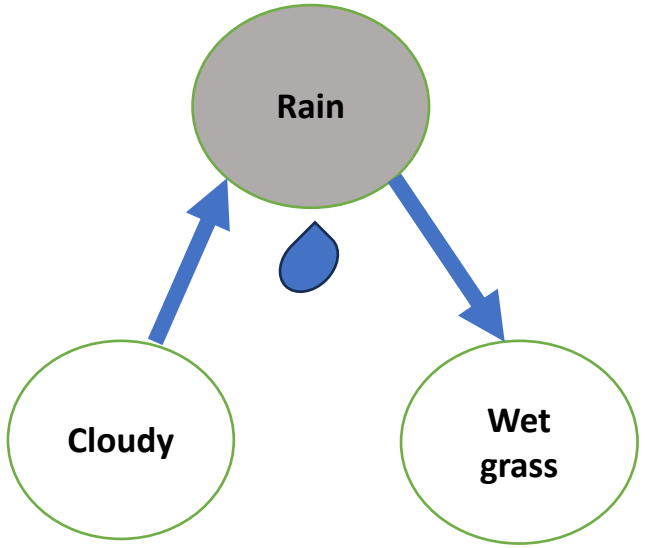


B →A ←C

Convergent connection

Some examples:



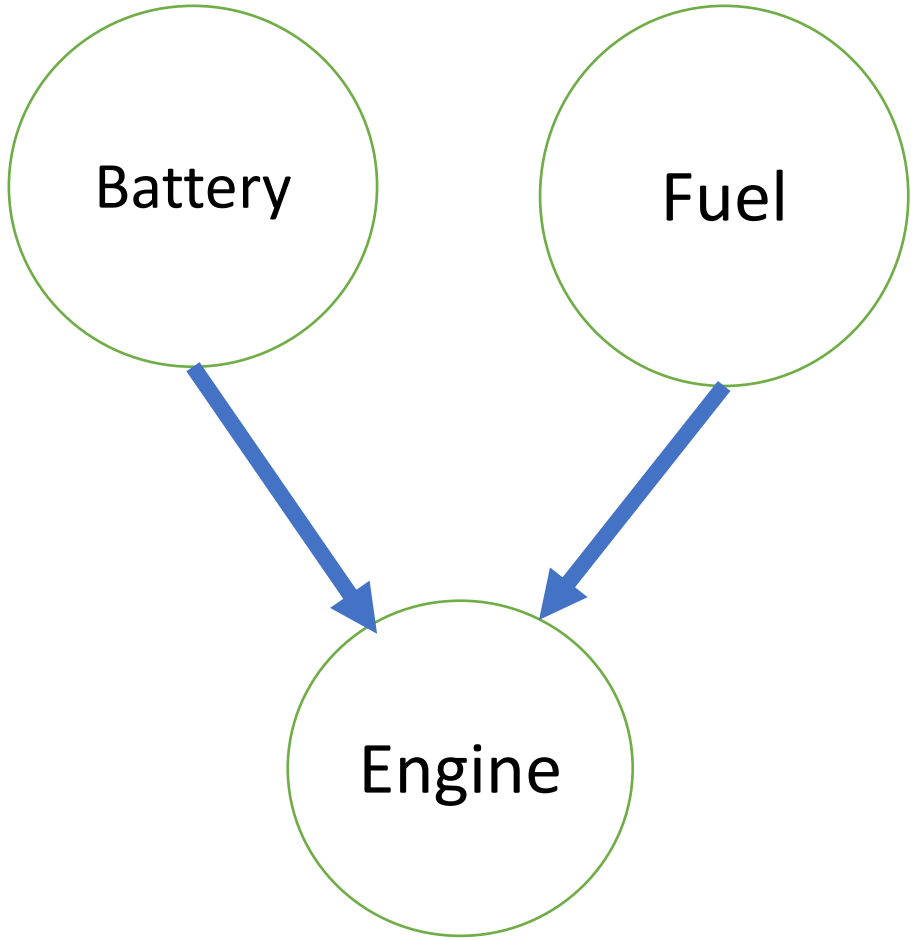


- Both variables **Cloudy** and **Grass** wet are **statistically dependent**.
- **Cloudiness** increases the probability of **rain** and thus indirectly the probability of a **wet ground**.
- **Conditional** on the variable **Rain**, the two variables **Cloudy and wet grass** are stochastically independent of each other.

1) If it is known whether it rains or not, the state of cloudiness has no influence on the probability that the ground is wet.

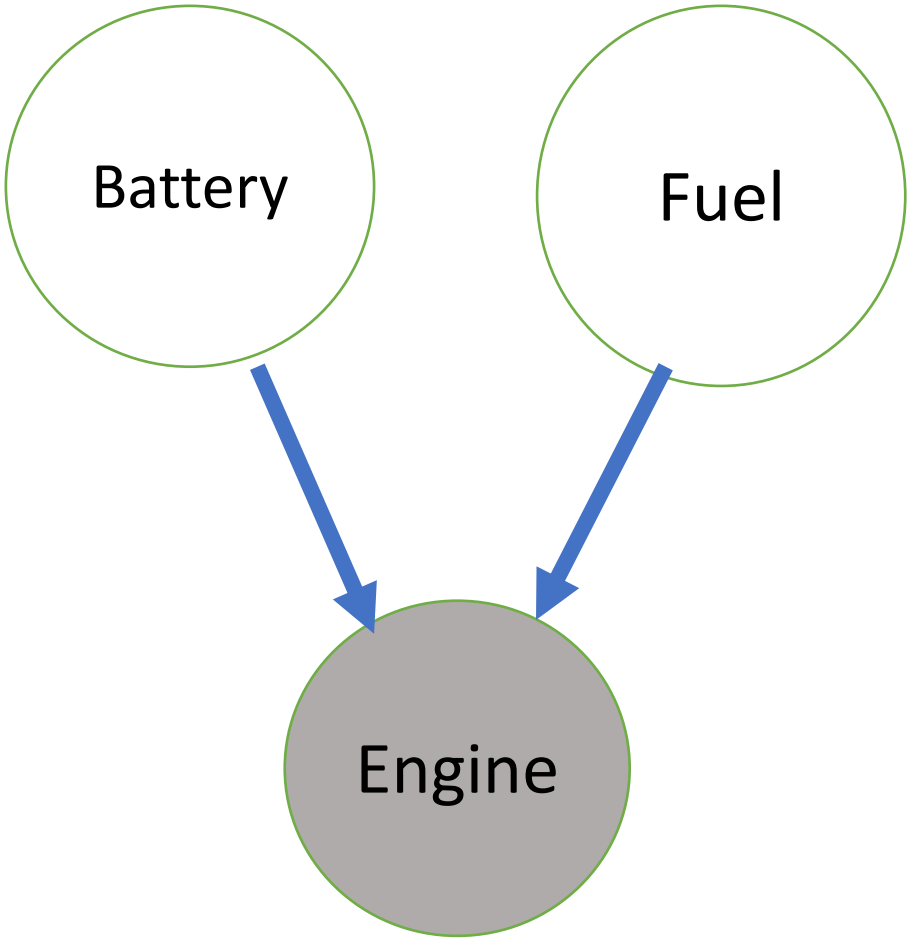
2) **If it is known whether it rains or not**, the condition of the grass has no influence on the probability of the state of the clouds.

Some examples:



- The binary variable **Battery** indicates if the car battery is working or not.
- The binary variable **Fuel** indicates whether the tank of the car is empty or not.
- The binary variable **Engine** indicates whether the car can be started or not.
- **Battery** and **Fuel** are **stochastically independent**.

Some examples:

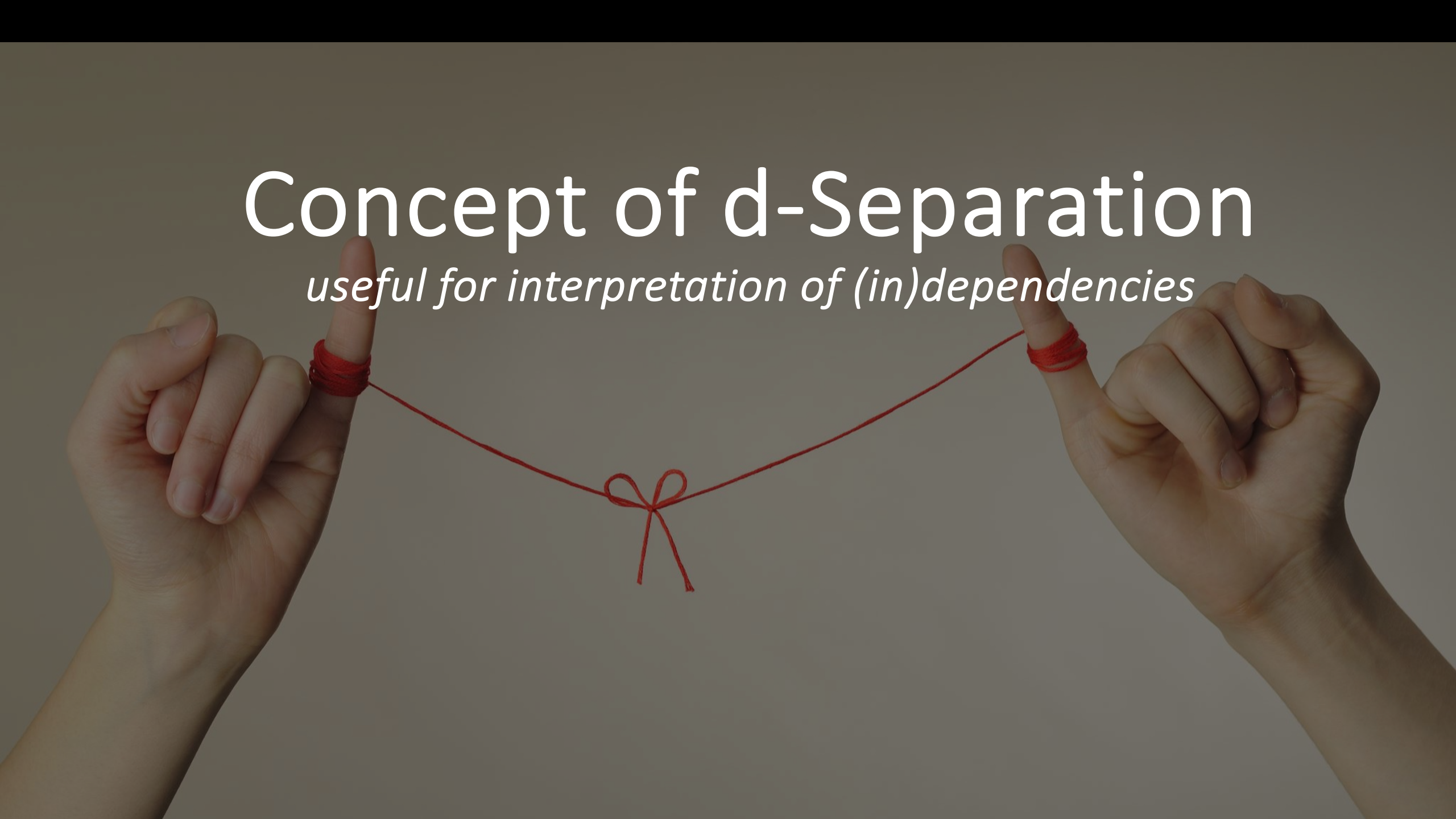


- However, **conditional** on the variable **Engine**, the two variables **Battery** and **Fuel** become **stochastically dependent**.

- When the car cannot be started, the probability that the fuel tank of the car is empty increases with the information that the battery is operating.

Concept of d-Separation

useful for interpretation of (in)dependencies



What is d-separation?

- **D-separation** is a way to check whether two variables in a Bayesian network **influence each other,** given that we know certain other variables (conditionally independency).

- **If variables are d-separated**, they **do not affect each other** once you know certain other variables (**no edge**).

- **If they are not d-separated**, then knowing one variable can still **tell you something useful** about the other (**edge**).

- **dsep()** in **bnlearn** package.

Path

- **Definition of directed path**

There is a **directed** path from node $\mathbf{X_i}$ to node $\mathbf{X_j}$ in a graph G if one can move from $\mathbf{X_i}$ to $\mathbf{X_j}$ by **following directed edges** (according to their edge directions) .

$$\mathbf{X_i \rightarrow \ldots \rightarrow X_j}$$

- **Definition of (any) path (path, trail)**

There is a path (trail) between the nodes X_i and X_j , if the two nodes are connected to each other through **a sequence of edges** (does not matter in which direction).

- In **a path or trail**, each node can appear **only once.**

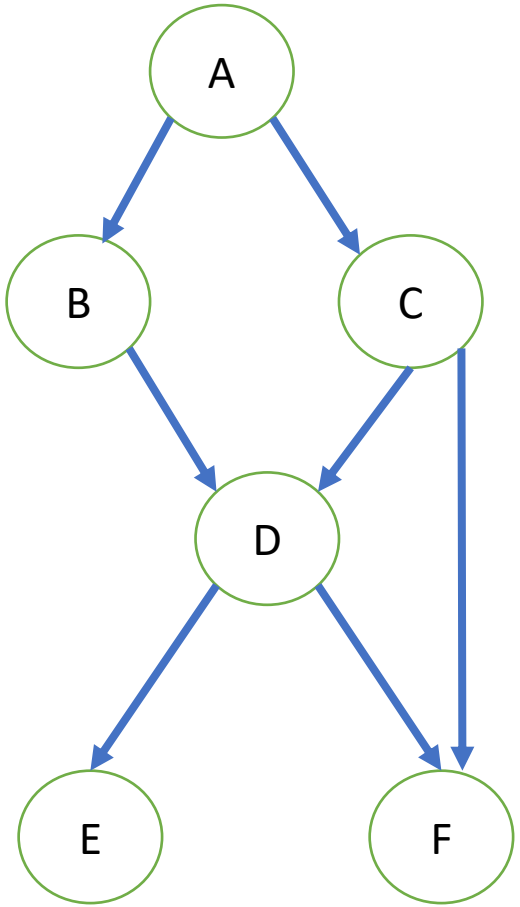
$$\mathbf{X_i \rightarrow X_{i+1} \leftarrow \ldots \rightarrow X_j}$$

- Is this a valid path?
- A→C→F←D←B←A←C

- What about this:

- A→D←B←C

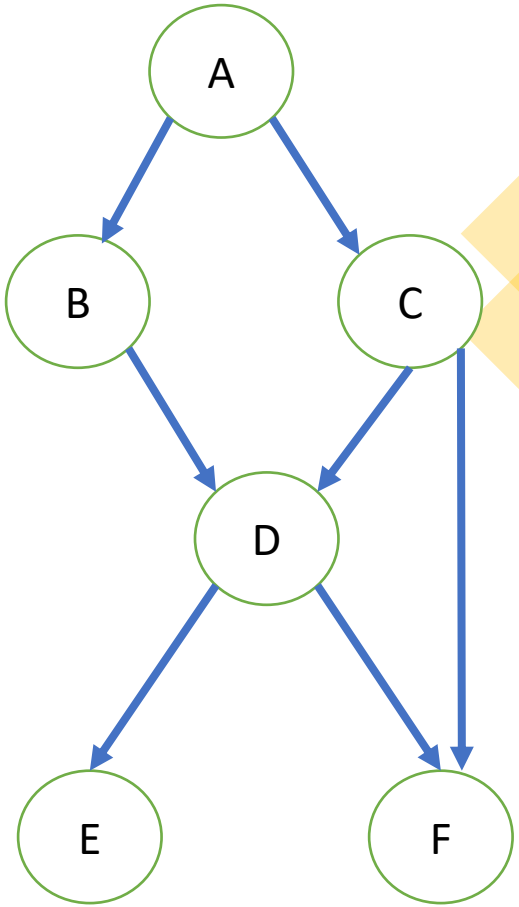
- Any more path?



Example of (directed) paths

- Examples of directed paths:
- A→B→D→F
- A→C→D

- We have the paths (trails):
- A→B→D←C
- B→D← C
- A→C→F←D←B



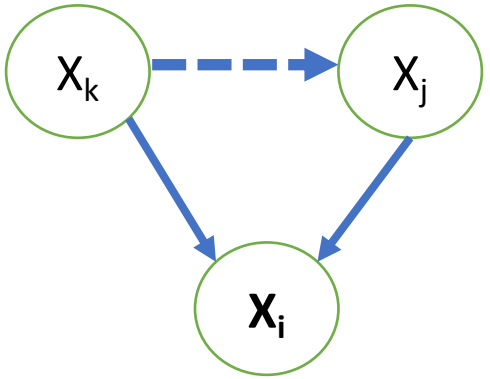
Collider

In a trail / path the node X_i $(i = 1, \dots, n)$ is a collider if two edges converge on X_i.

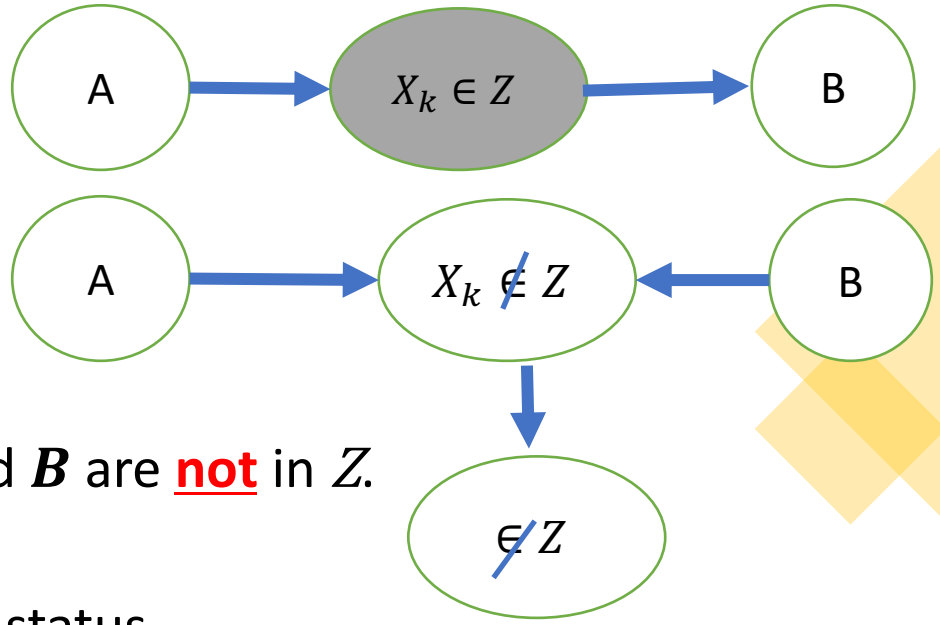
E.g.: $\mathbf{X_w \rightarrow X_k \rightarrow X_i \leftarrow X_j \rightarrow X_m}$

- **Note**: This definition does not require that $\mathbf{X_k \rightarrow X_i \leftarrow X_j}$ is a v-structure.

The nodes X_k and X_j can be in a parent-child relationship.



Blocked path

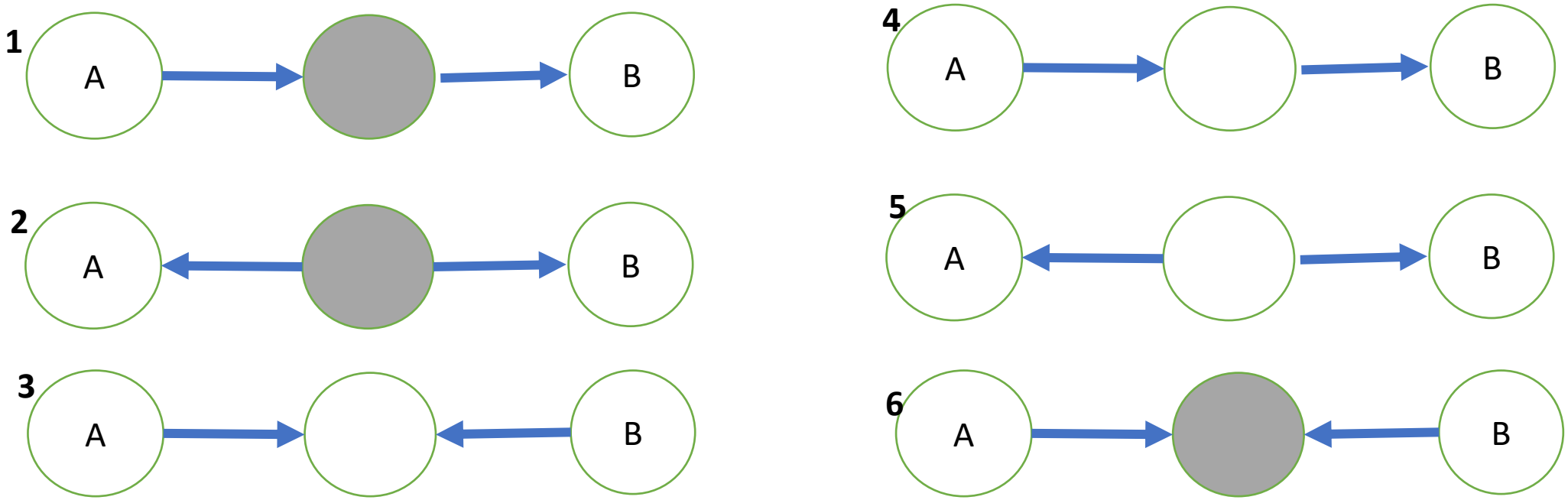


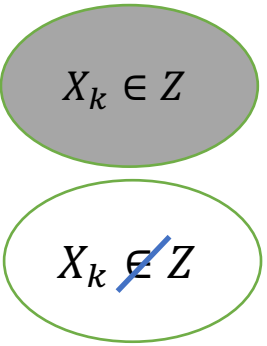
- We consider the nodes $\boldsymbol{A}$ and $\boldsymbol{B}$ and a subset Z , where $\boldsymbol{A}$ and $\boldsymbol{B}$ are **not** in Z.
- **Being in Z means the node is observed**. That is, we know its status.
- A path (trail) between $\boldsymbol{A}$ and $\boldsymbol{B}$ is **blocked** when the trail leads through any node X_k and:

(1) X_k **is not** a collider and X_k **is** an element of Z (X_k is observed).

(2) X_k **is** a collider and neither X_k nor a descendant of X_k is an element of Z (they are **not** observed).

Which one is blocked?





The filled (grey) nodes are elements of the set Z (known). The empty (white) nodes are not elements of Z.

Definition: d-Separation

d-Separation

- Two variables (X_i and X_j) are d-separated (no edge or disconnected) by a third set of variables (Z) if:
  - All paths between X_i and X_j are blocked when we know Z.
  - That is, if X_i and X_j are d-separated, they are conditionally independent.

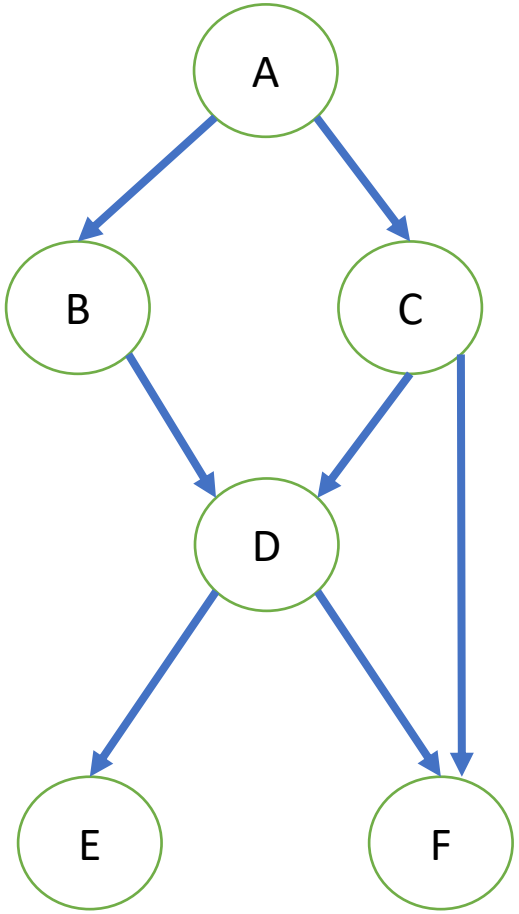
Example

A and D are d-separated conditional on Z={B,C}?

A and F are d-separated conditional on Z={D,C}?

B and C are d-separated conditional on Z={A}?

B and C are d-separated conditional on Z={A,D}?



Remark:

- **The software does the heavy lifting** - all the math and checks run automatically.

What you need to know:
D-separation is the rule that tells us whether two variables in a Bayesian network still **influence each other** after we fix (condition on) other variables.

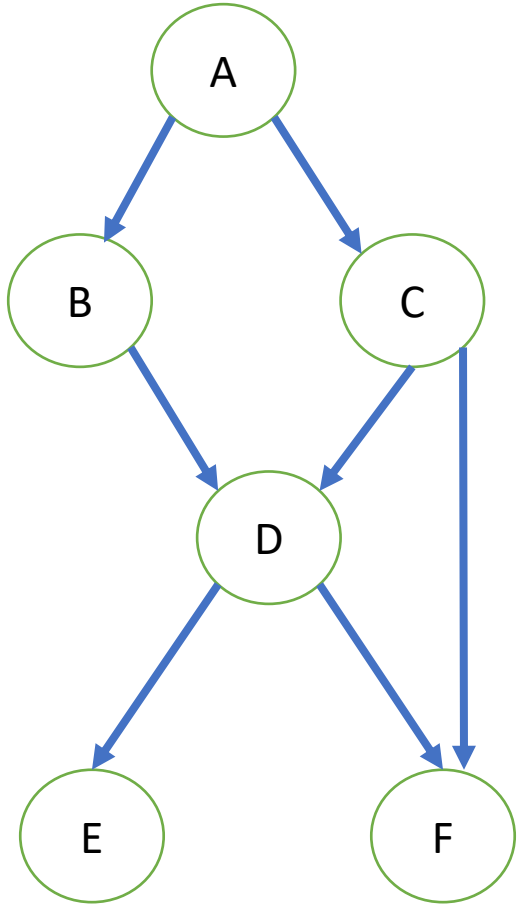
Why it matters:

- **Faster reasoning** - lets us ignore parts of the network that can't affect our result (Used to identify which variables are unrelated, and which ones matter for making predictions).
- **Smarter feature selection** - shows which variables you actually need for prediction.
- **Clearer interpretation** - spells out where information really flows in the model.

Static Bayesian networks

- **The first component** of a BN is a ***graph***.

- The second component of a BN is the ***probability distribution*** $\boldsymbol{P(X)}$.



- Second component:

The probability distribution $P(X)$

Common Types of Bayesian Networks (BNs)

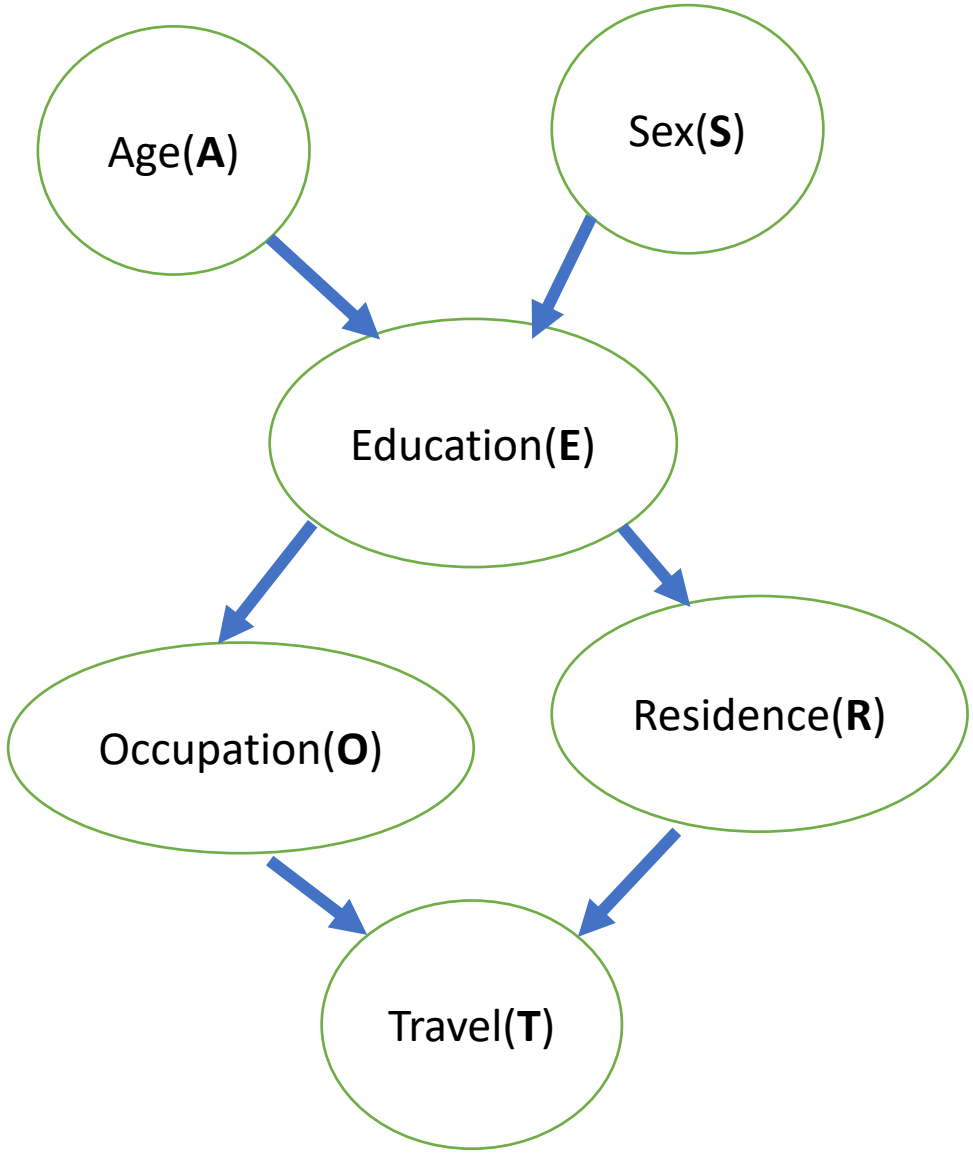
- **Discrete BNs**
  - All variables are **categorical** (e.g. yes/no, red/blue).
  - The relationships are modeled with **multinomial distributions**.
- **Gaussian BNs (GBNs)**
  - All variables are **continuous**.
  - The overall system follows a **multivariate normal distribution**, and each variable (given its parents) is **normally distributed**.
- **Conditional Linear Gaussian BNs (CLGBNs)**
  - A **mix of discrete and continuous** variables.
  - Combines the ideas of **Discrete BNs** and **Gaussian BNs.**

Discrete BNs

- The joint probability distribution is a **multinomial distribution,**
- **That is,** assigning a probability to each **combination of states of variables**.

Example of Discrete BNs

- **Age** and sex are not influenced by any other variable.
- **Age** and sex have a direct influence on **Education**
- **Education** strongly influence both **occupation** and **residence**
- **Transports** are directly influenced by both **occupation** and residence.
- If no BN: $\mathbf{2^6 - 1 = 63}$



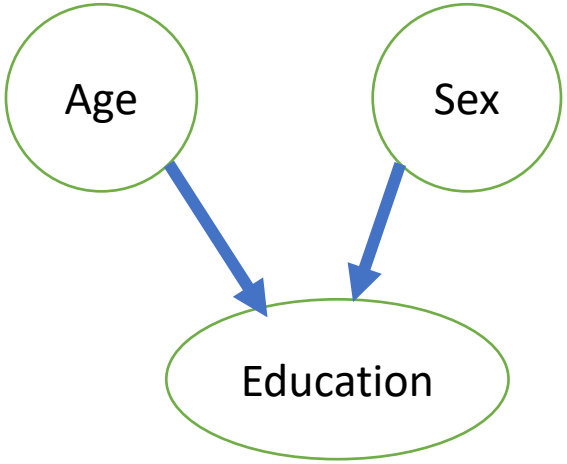
	Yong	Adult	Old
	0.3	0.5	0.2

Age

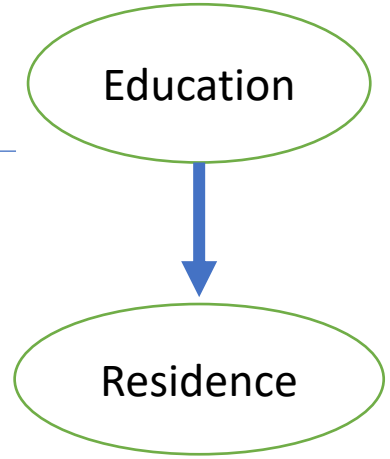
	M	F
	0.6	0.4

Sex

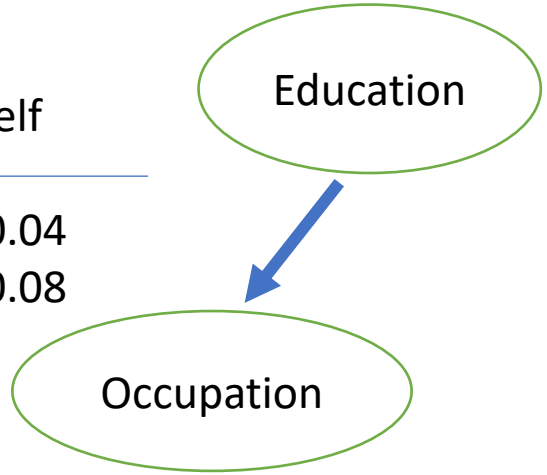
	High	Uni
Young& M	0.75	0.25
Adult& M	0.72	0.28
Old& M	0.88	0.12
Young& F	0.64	0. 36
Adult& F	0.70	0.30
Old& F	0.90	0.10



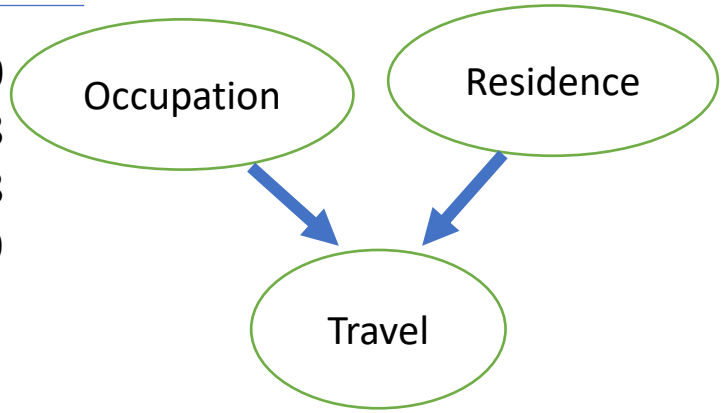
	Small	Big
High	0.25	0.75
Uni	0.20	0.80



	Emp	Self
High	0.96	0.04
Uni	0.92	0.08



	Car	Train	Other
Small&Emp	0.48	0.42	0.10
Small&Self	0.56	0.36	0.08
Big&Emp	0.58	0.24	0.18
Big&Self	0.70	0.21	0.09



If BN: Overall, fewer parameters: **21** **parameters only !**

Learning the dag structure:

- It is not always possible or desired to rely on prior knowledge on the phenomenon we are modeling to decide which arcs are present in the graph and which are not.

- **Therefore, the structure of the DAG itself maybe the object of our investigation.**

- E.g. in genetics and systems biology, BNs are used to **reconstruct** molecular pathways and networks underlying complex diseases and metabolic processes (Sachs et al., 2005).
- In **economics**, they help discover relationships between variables like interest rates, inflation, and employment.
- For casual inference (considering some assumptions)

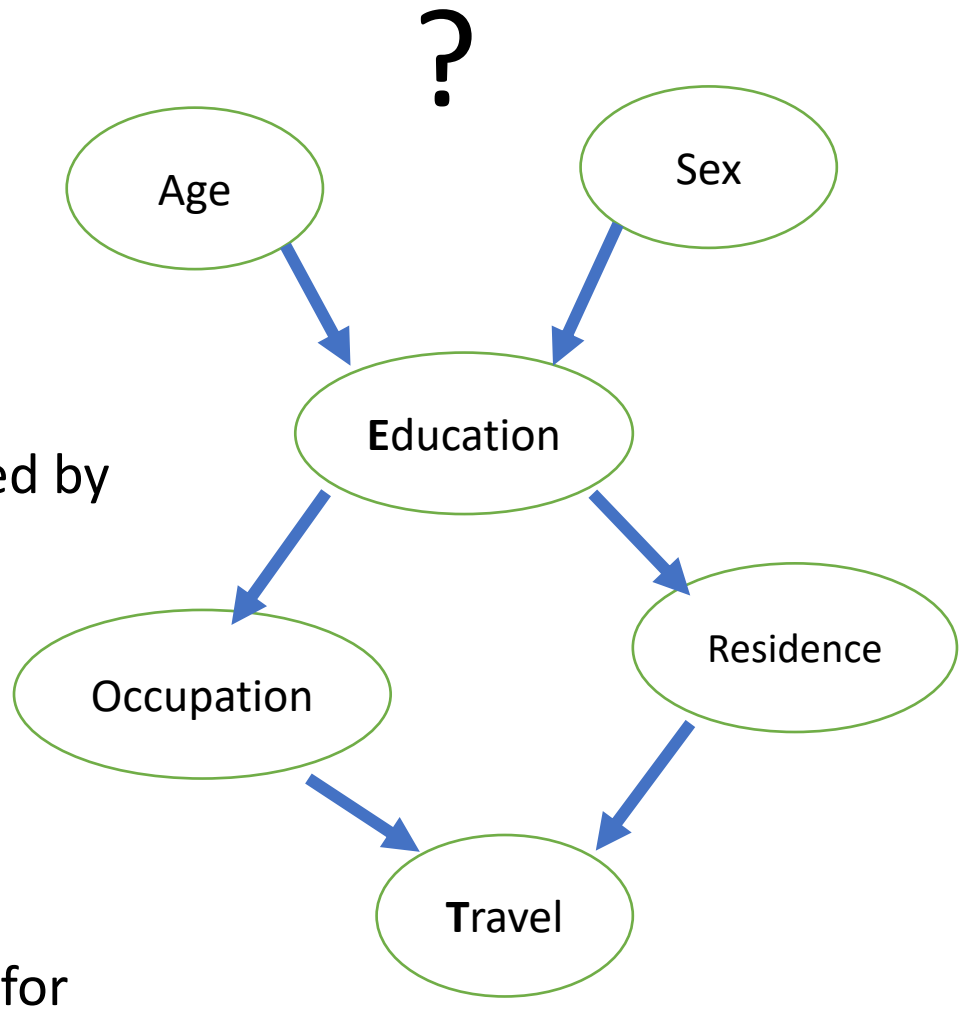
Learning the dag structure:

- Several algorithms have been presented in literature for this problem,
- They can all be traced to three main approaches:

  - **Constraint-based:** Uses statistical tests to find conditional independencies.
  - **Score-based:** Tries different graphs and scores them based on how well they fit the data.
  - **Hybrid:** Combines both

Constraint-based :

Conditional independence test
(Pearl 1990, Verma and Pearl, 1991)



- It focus on presence of individual arcs.
- It can be used to assess whether the dependency is supported by the data.

- H_0: T $\perp\!\!\!\perp_P$ E |{O,R}, they are independent, no edge
- H_1: T $\not\perp\!\!\!\perp_P$ E |{O,R}, they are dependent, edge

- If the **null hypothesis** is rejected, the arcs can be considered for **inclusion** otherwise for **exclusion**.

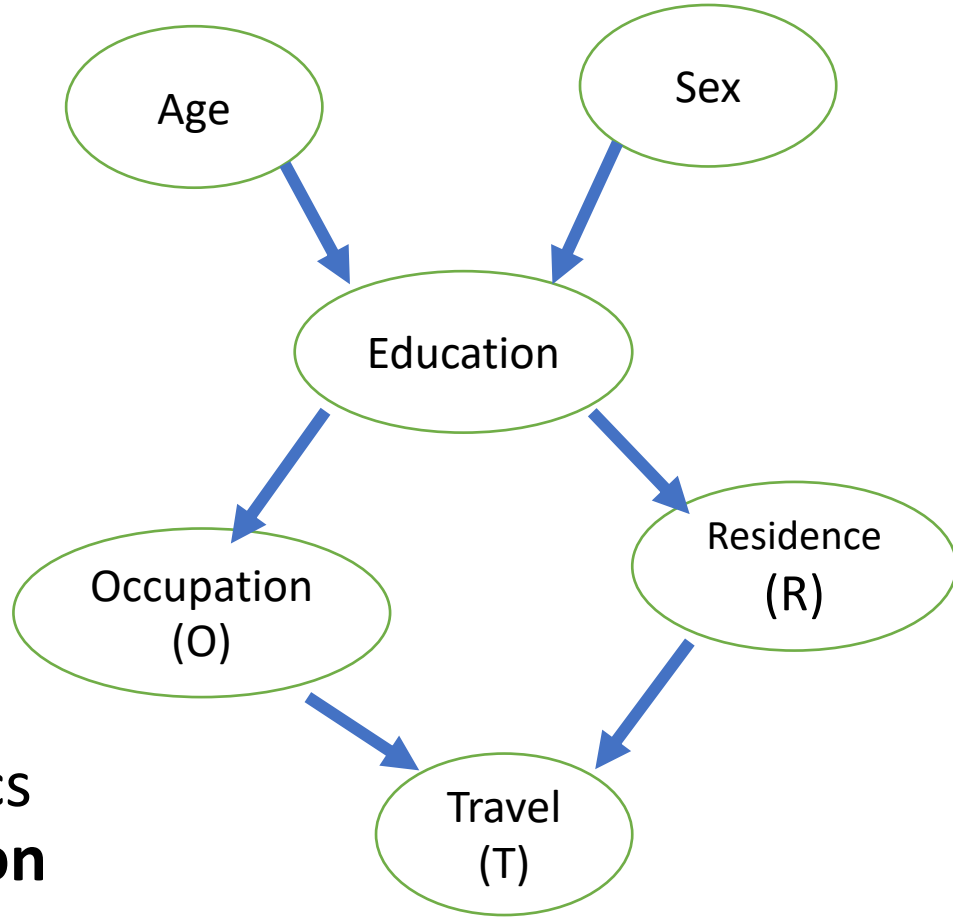
- This approach operates edgewise.

Conditional independence test

- H_0: T $\perp\!\!\!\perp_P$ E |{O,R},
- H_1: T $\not\perp\!\!\!\perp_P$ E |{O,R}

- Performing this test H_0 by adapting the **log-likelihood ratio**, G^2, or **Pearson's** X^2 test.

- If the **null hypothesis** is ***rejected*** (p-value > α), the arcs can be considered for **inclusion** otherwise for **exclusion** (very small p-value).



Conditional independence test

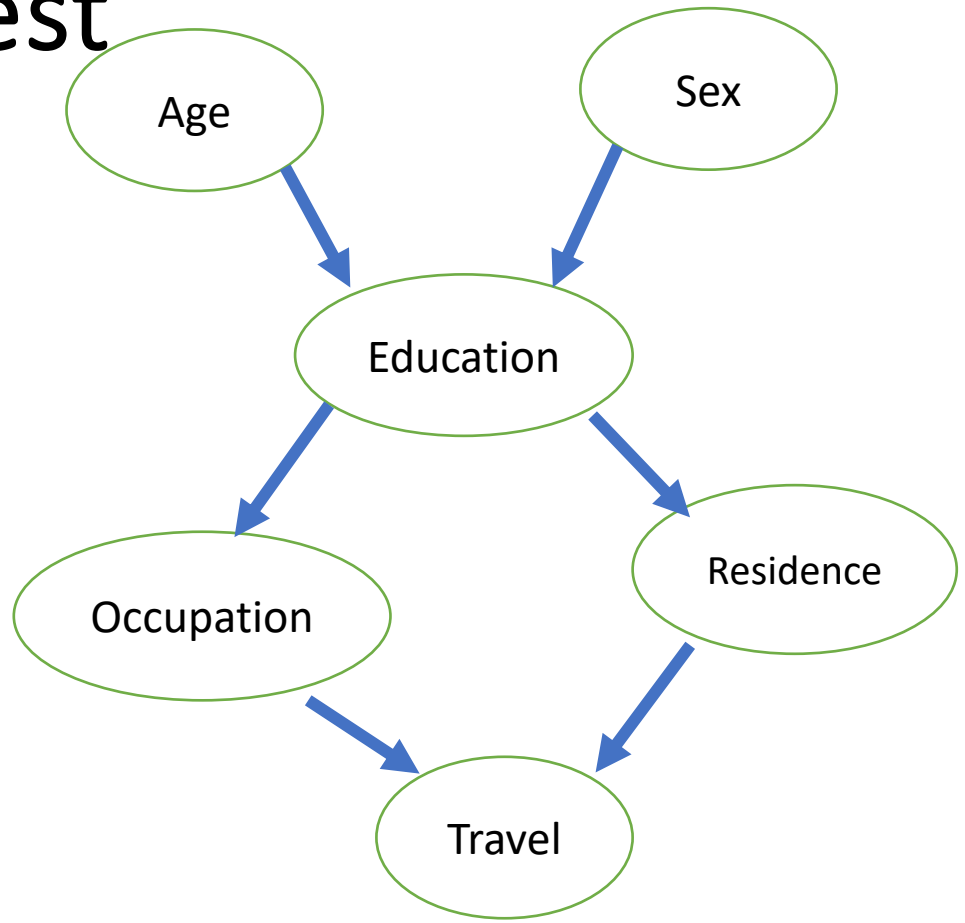
- **the null hypothesis *<u>is not</u>*** rejected,

  p-value > α ---- there is no edge

- **The null hypothesis** is rejected

  p-value < α----- there is an edge

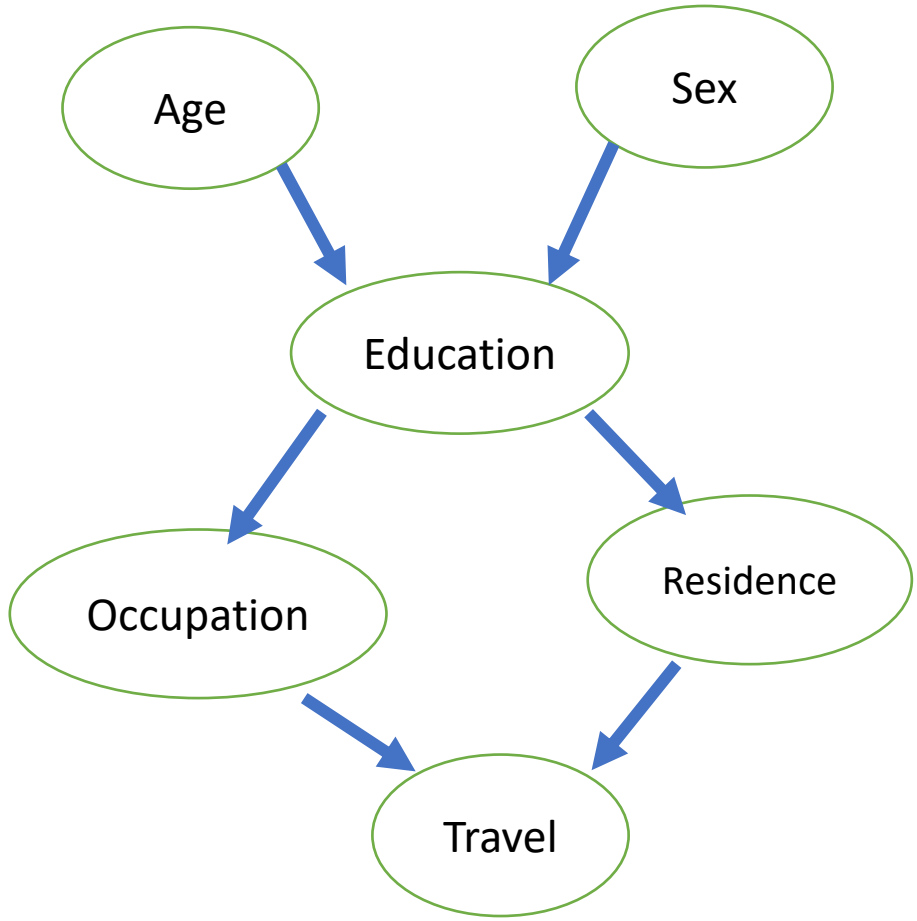
- α is usually 0.05 or 0.01
- We will discuss this in practical.



Network score

- Each candidate BN is assigned a network score reflecting its **goodness of fit.**

- Bayesian information criterion (**BIC**)
- Bayesian Dirichlet equivalent uniform(**BDeu**)



Algorithms that search for the DAG given the data (maximize a given network score)

- **Greedy search algorithm** (**e.g. Hill-climbing**). 
  - Start with an initial graph and **add, delete, or reverse one edge at a time**, always taking the move that most improves the score until no single change helps.
- **Genetic algorithm**.
- **Simulated annealing**

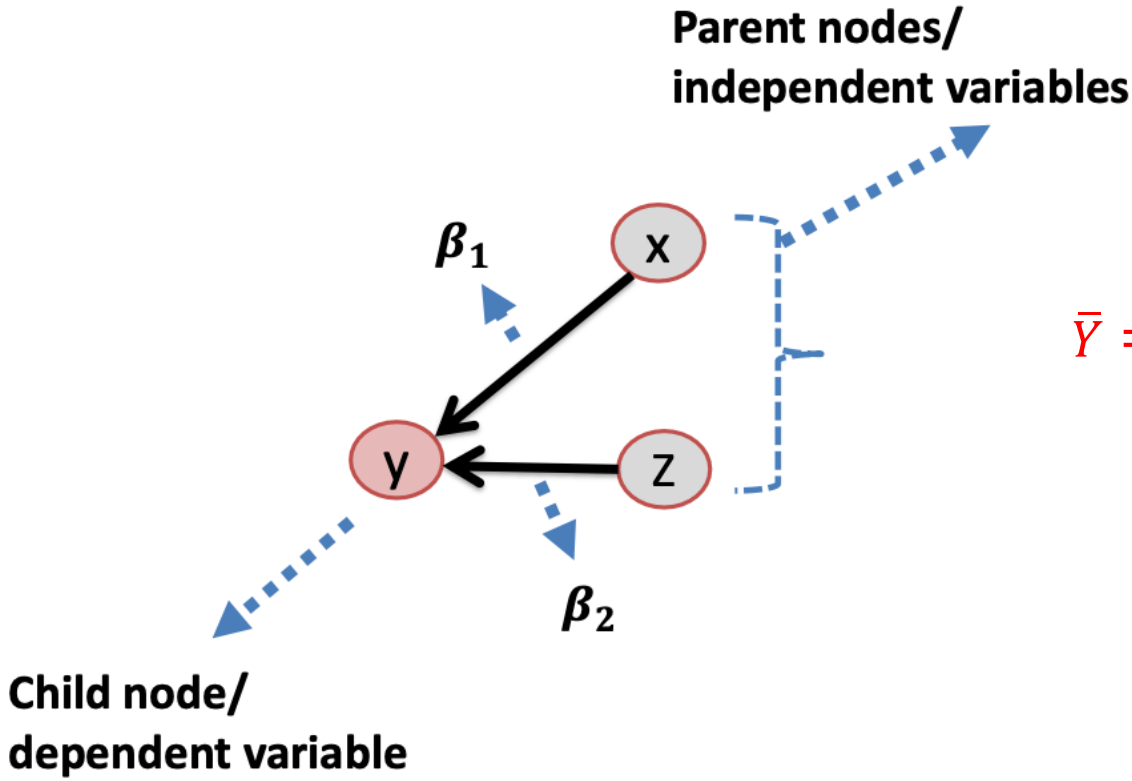
- We will return to this in the lab session.

Continuous (Gaussian) Bayesian Network

Continuous BNs

- Every node follows a **normal distribution**.
- Nodes without any parents, are discribed by the univariate normal distribution.
- The local distribution of each node can be equivalently expressed as **a Guassian linear model** which includes an intercept and the node's parents as explanatory variables (predictors), **without any intraction term**.
- **Child~ β_1*Parent1 + β_2 *parent2 +…**

Edges:presence or absense of coefficient



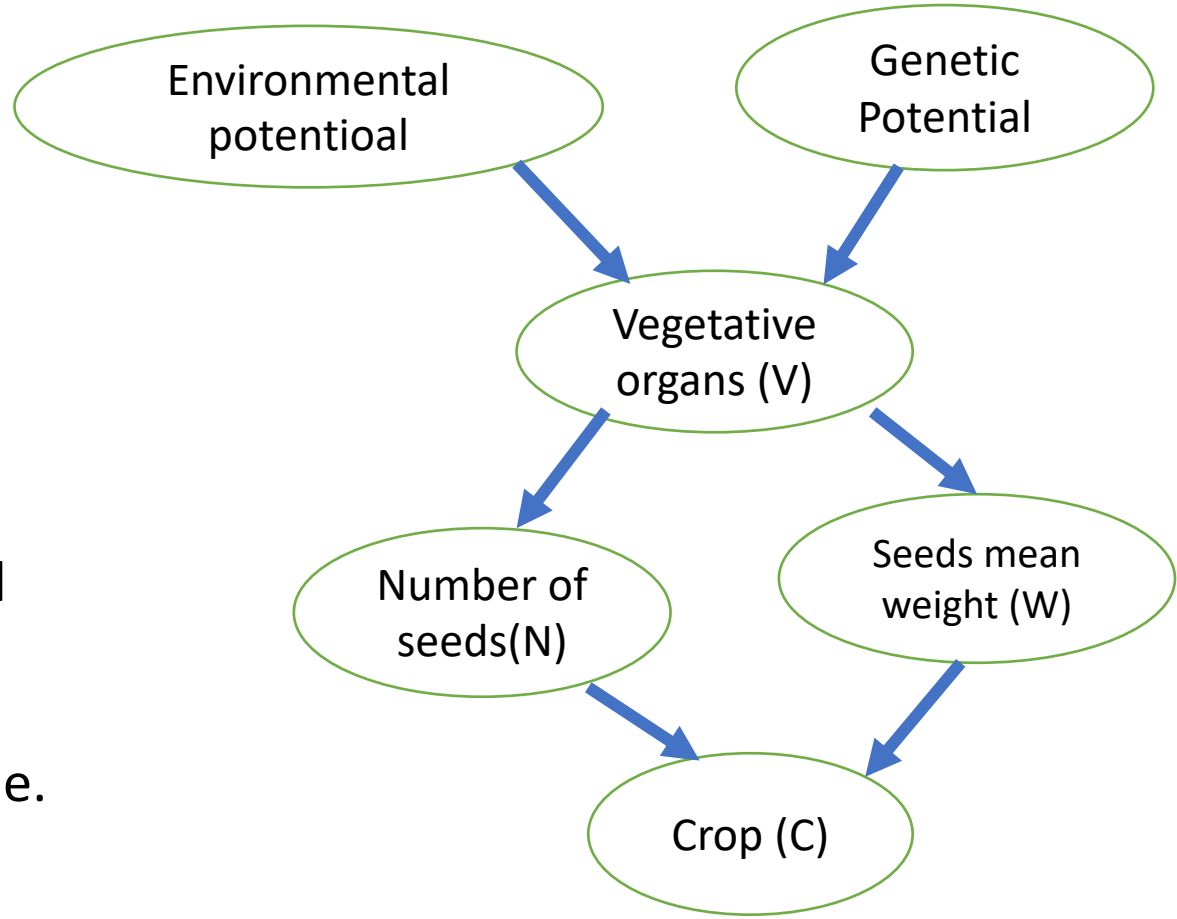
$$\bar{Y} = 0.1\boldsymbol{X} + 0.8\boldsymbol{Z}$$

$\overline{Child}$ ~ β_1*Parent1 + β_2 *parent2 +…

Example of continuous BNs

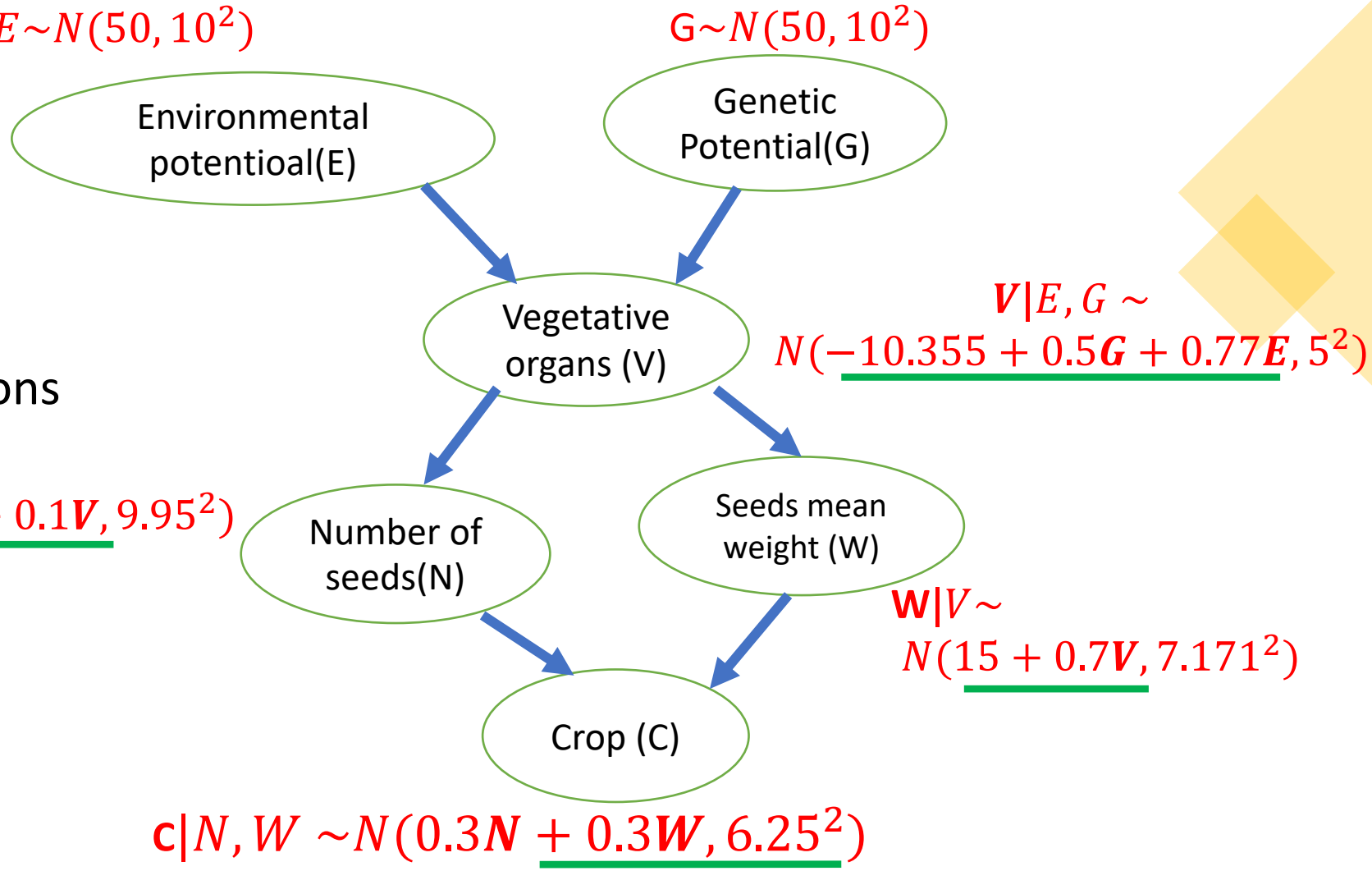
For the analysis of a particular plant:

- **Genetic Potential(G):** Genotype effect (a single score)
- **Environmental potentioal(E)**: Environmental (location and season) effect (a single score).
- **Vegetative organs (V):** Roots, stems, etc., grow and accumulate reserves exploited for reproduction and summarises all the information available on constituted reserves.
- **Number of seeds(N)** is determined at the flowering time.
- **Seeds mean weight (W)** is assessed in the plant's life.
- **Crop (C):** The harvasted grain mass.



Example

- Six variables and six arcs corresponding to the direct dependencies linking them.
- The local probability distributions are shown for each node.



N means normal/ Gaussian distribution.

$\overline{Child}$ ~ β_1*Parent1 + β_2 *parent2 +…

Some remarks

- WHY linear dependencies?
- Closed-form results for many inference procedures.

- Relatively **simple models** often perform better than very **sophisticated models**.

- $\overline{Child}$ ~ β_1*Parent1 + β_2 *parent2 +…

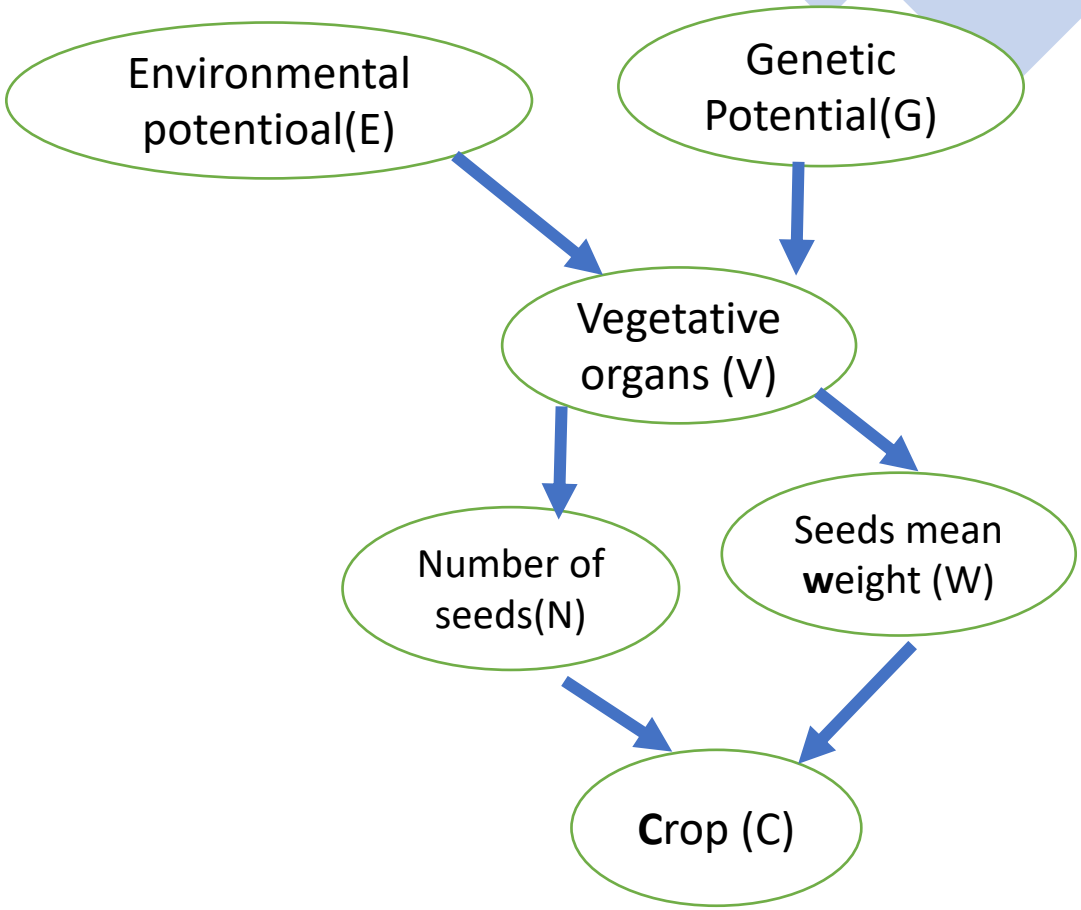
Learning the DAG Structure: Tests and Scores

- Often expert knowledge on the data is not detailed enough to completely specify the structure of the DAG. In such cases, if sufficient data are available, we can infer a sparse BN.

- The two classes of criteria used to learn the structure of the DAG are
- **conditional independence tests** and
- **network scores**.

Conditional Independence Tests

- **Most common**: exact test for *partial correlations*.
- H_0: C $\perp\!\!\!\perp_P$ W |N → **no edge**
- H_1: C $\not\perp\!\!\!\perp_P$ W |N → edge between **C** and **W**



Conditional Independence Tests using bnlearn package

- ✓ Using “bnlearn” package
- ✓ Test for partial correlations
- ✓ Computing the corresponding statistics.
- ✓ P-value

- ✓ **The null hypothesis *is not*** rejected :

  p-value > α ---- there is no edge

- ✓ **The null hypothesis** is rejected:

  p-value < α----- there is an edge

- ✓ We will discuss this in practical.

Network Scores

Same as Discrete BN:

- **BIC**
- **BGe (**Bayesian Guassian equvalent score)

- We search for the best network structure.
- Let’s discuss all in lab session.

Bayesian networks

- are a combination of a DAG and a global distribution, both defined on the same variables.

- provide a principled solution to the problem of feature selection using Markov blankets.

- ***can be very useful tool for Network reconstruction.***

Finally, practically: *MRFs vs BNs*

- **MRFs** have more power than **BNs**, but are more difficult to interpret and deal with computationally.

- A general rule of thumb is to use **Bayesian networks** whenever possible, and only switch to MRFs if there is no natural way to model the problem with a directed graph .

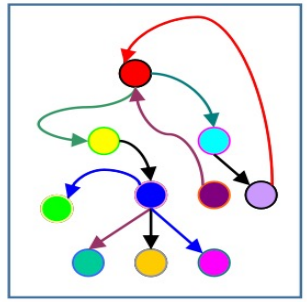
In a nutshell

Raw data

Cleaned data

Machine Learning

Statistical Methods



Network inference

- Think about the application of this concept in your field for a few minutes and discuss that in pairs.

- Consider the following questions for reflection in your field:

✓What variables are present in your project?

✓Why are these variables important?

✓What motivates your interest in understanding their interdependencies?

✓How does this understanding contribute to your work or goals?

References

- Dechter, R. (2019). ***Reasoning with Probabilistic and Deterministic Graphical Models****: Exact Algorithms*. Morgan & Claypool publishers. https://doi.org/10.2200/S00893ED2V01Y201901AIM041
- Højsgaard, S., Edwards, D., & Lauritzen, S. (2012). ***Graphical Models with R***. Springer New York, NY. https://doi.org/10.1007/978-1-4614-2299-0
- Koller, D.& Friedman, N. (2010). ***Probabilistic Graphical Models: Principles and Techniques***. The MIT Press Cambridge, Massachusetts.
- Nagarajan, R., Scutari, M. & Lébre, S. (2013). ***Bayesian Networks in R****: with Applications in Systems Biology*. Springer New York, NY. https://doi.org/10.1007/978-1-4614-6446-4