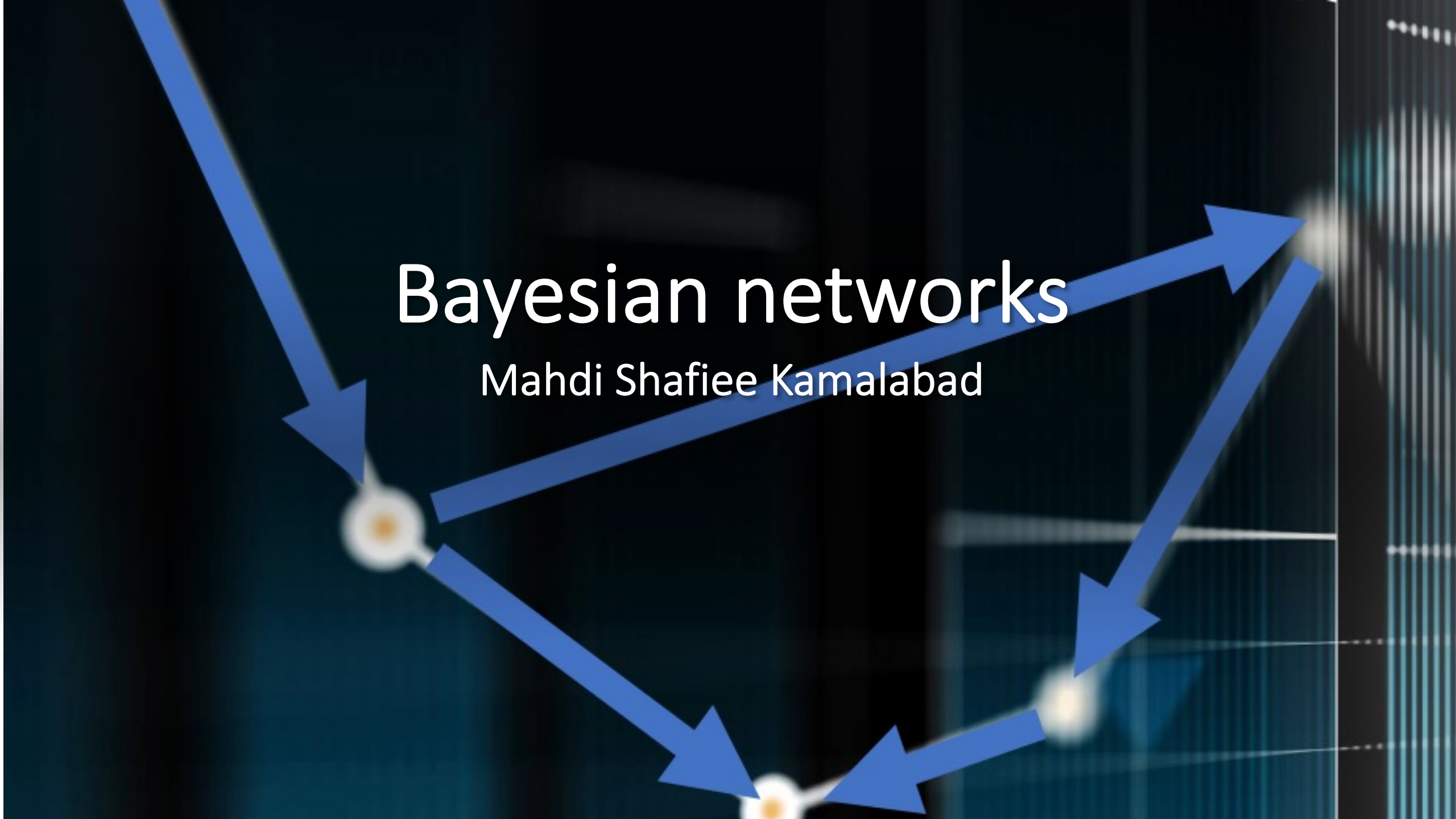
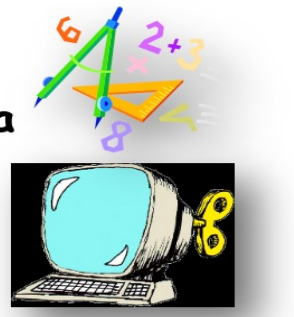

Bayesian networks
Mahdi Shafiee Kamalabad

In a nutshell

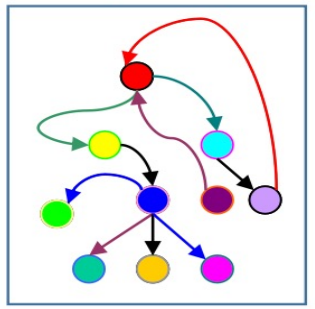
Raw data

Cleaned data



Machine Learning

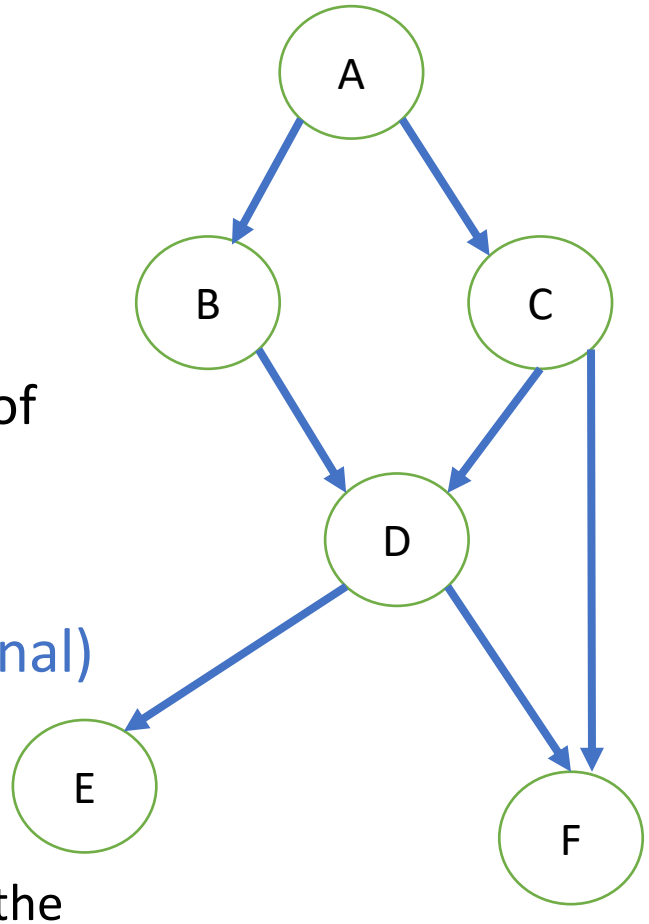
Statistical Methods



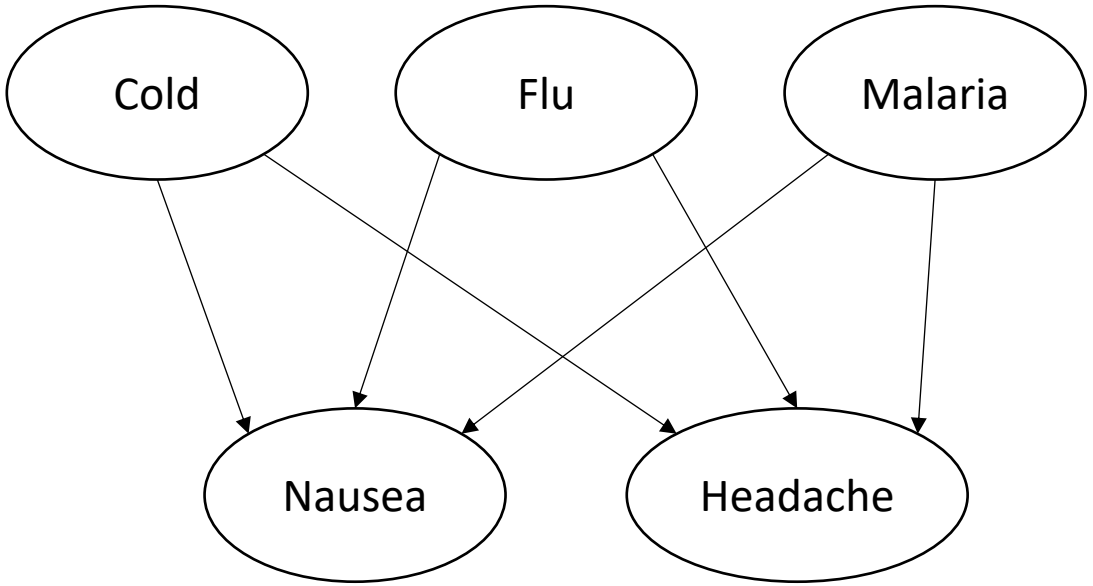
Network inference

Bayesian networks (BNs)

- **Marriage between** graph theory and probability theory.
- A graphical model for **probabilistic relationships** among a set of variables
- **Nodes** represent variables and **edges** represent (conditional) dependence between variables.
- Real world applications are probabilistic in nature, and to represent the relationship between multiple events, we may need a Bayesian network.



Example 1:

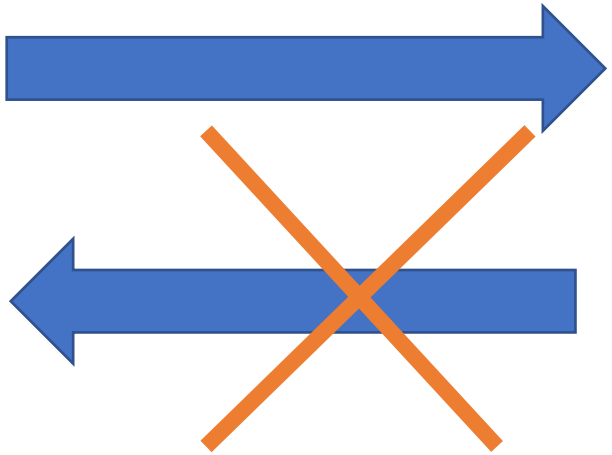


- A Bayesian network could represent the **probabilistic relationships** between **diseases** and **symptoms**.

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

Note

Causal Network



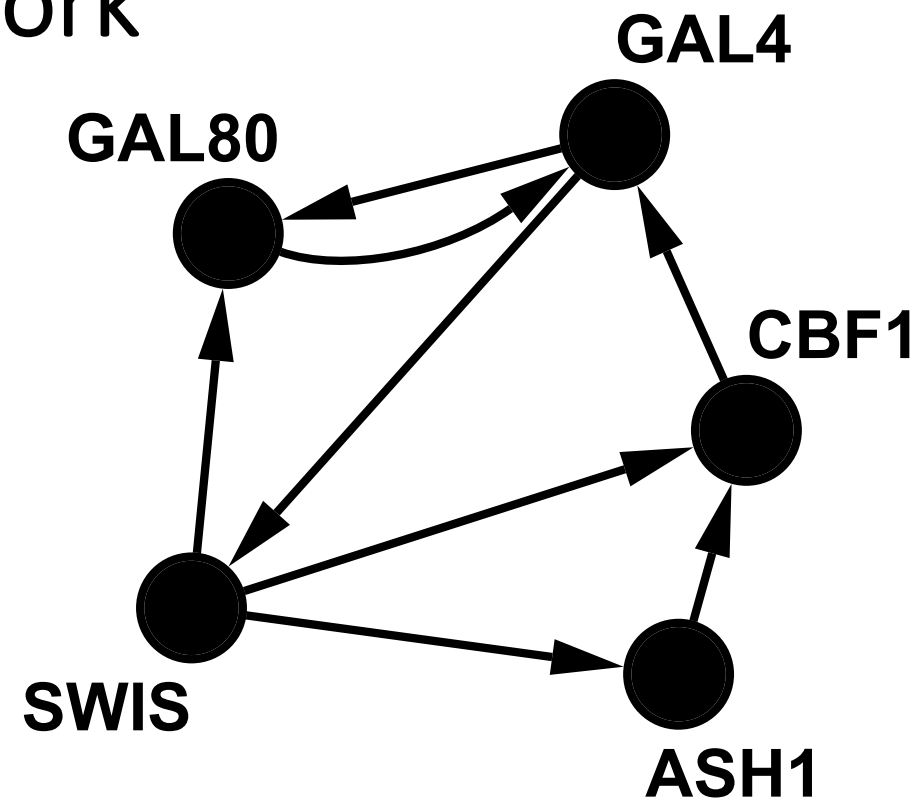
Not necessarily

Bayesian Network

Example 2: Gene regularity network

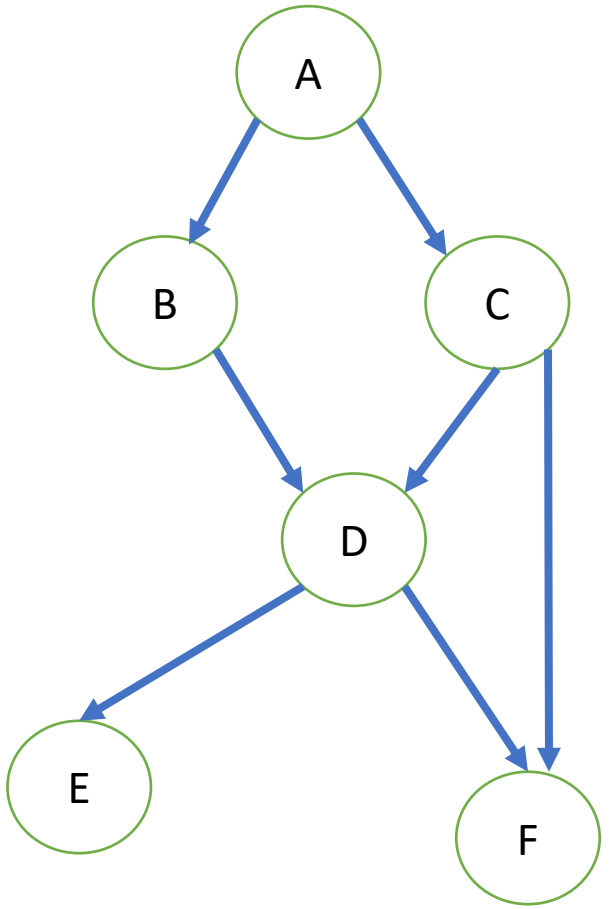
Network of n = 5 genes in Saccharomyces cerevisiae (yeast). The data obtained from synthetically designed yeast cells grown with different carbon sources: galactose (“switch on”) or glucose (“switch off ”), Cantone et al. (2009) .

- **Dynamic Bayesian Network Models**



Some applications of BNs

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



See here

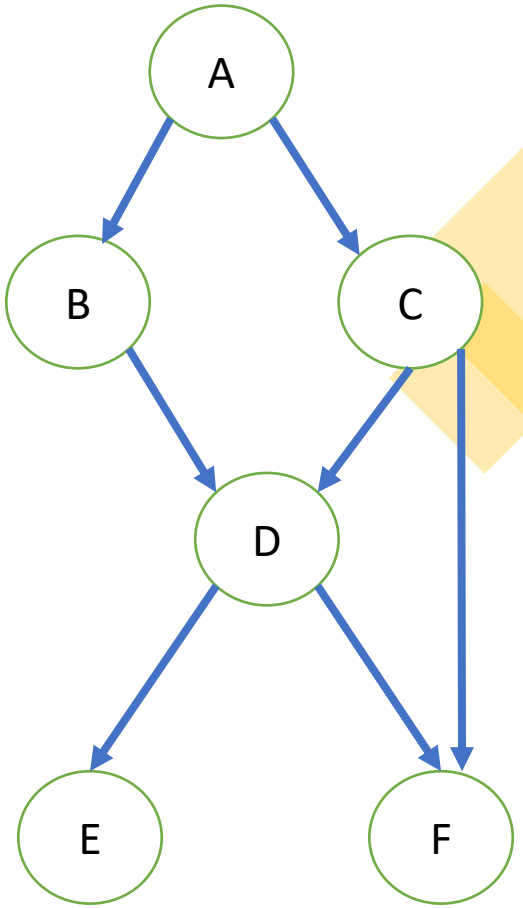
https://data-flair.training/blogs/bayesian-network-applications/

Bayesian networks

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

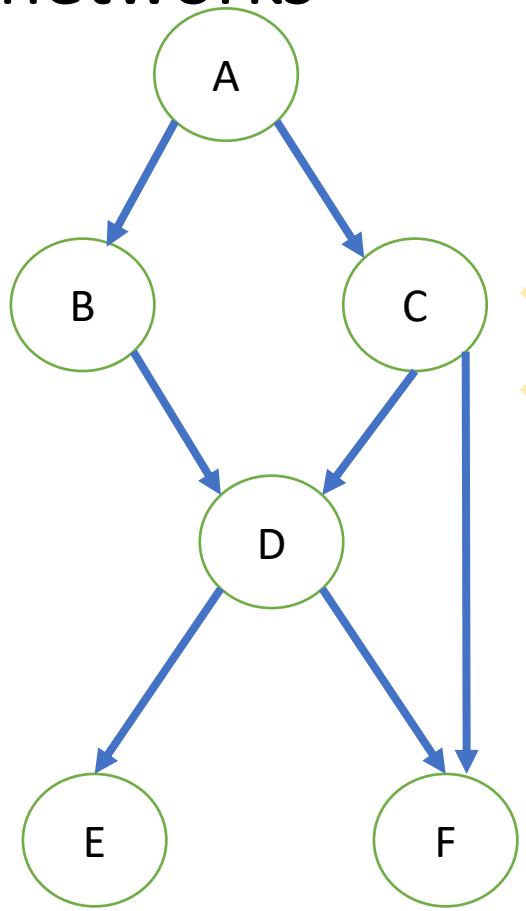
  A graph G is a mathematical object with:
- a set of nodes $V = \{v_1, \ldots, v_N\}$;
- a set of **arcs** A which are identified by pairs for nodes in V, e.g. $a_{ij} = (v_i, v_j)$.

- **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.

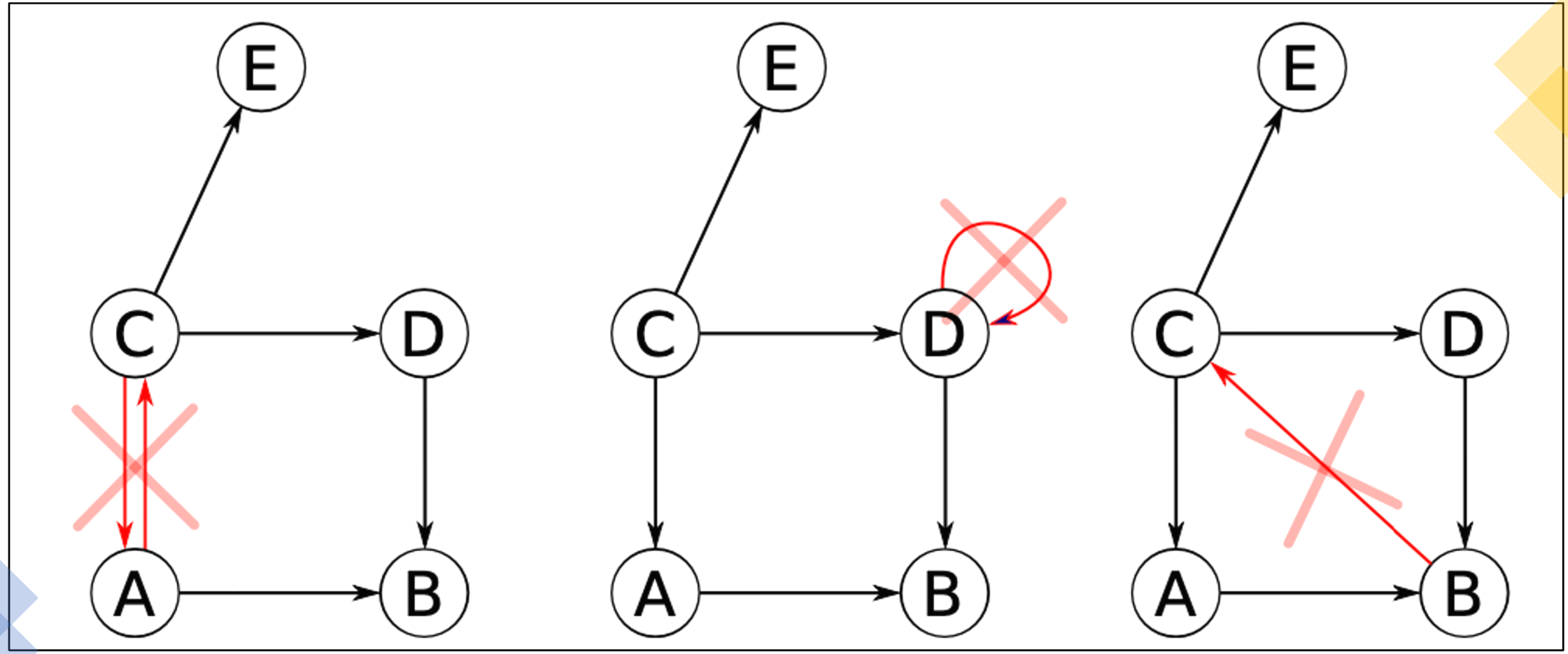


Directed Acyclic Graph (DAG) in static Bayesian networks

- contains only **directed** arcs and <u>does not contain any loop/cycle</u>.
- **does not depend on the nature** of the variables under considerations.



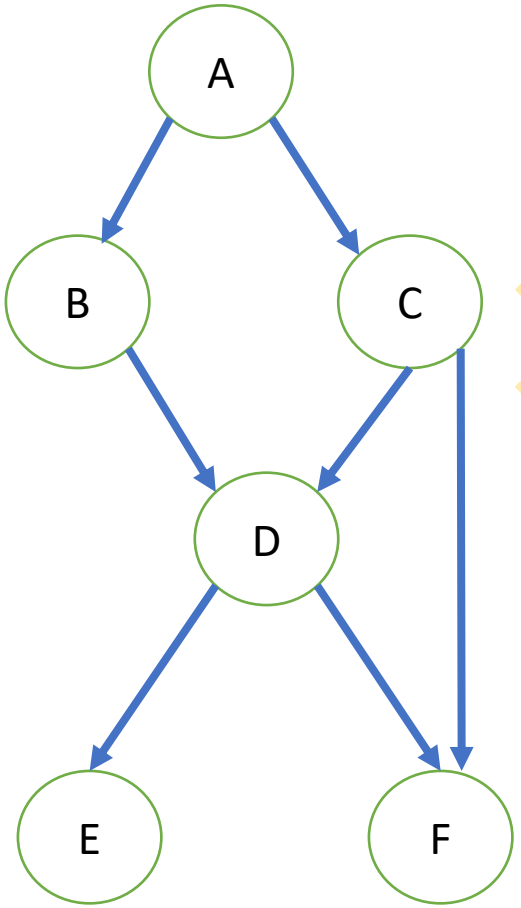
DAG: Directed acyclic graph: No loops/no cycle



Interpretation of edges in BNs

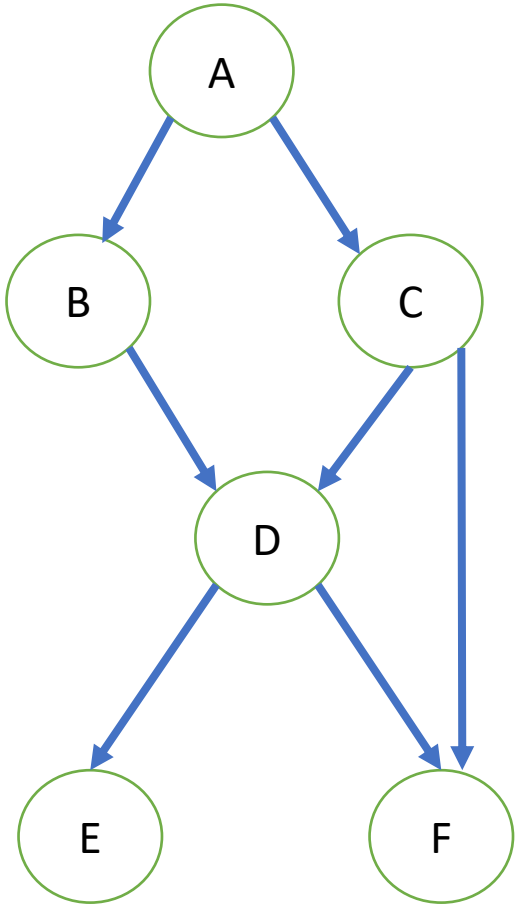
- If there is an **edge** from one variable to another, **the later depends on the former**.

- Variables that **are not linked (lack of edges)** are **conditionally independent**.



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).



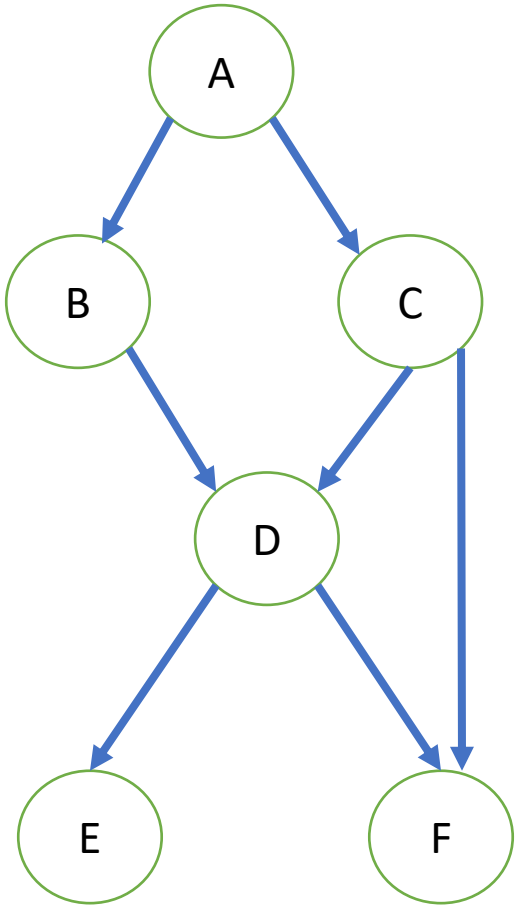
Bayesian networks

- 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:

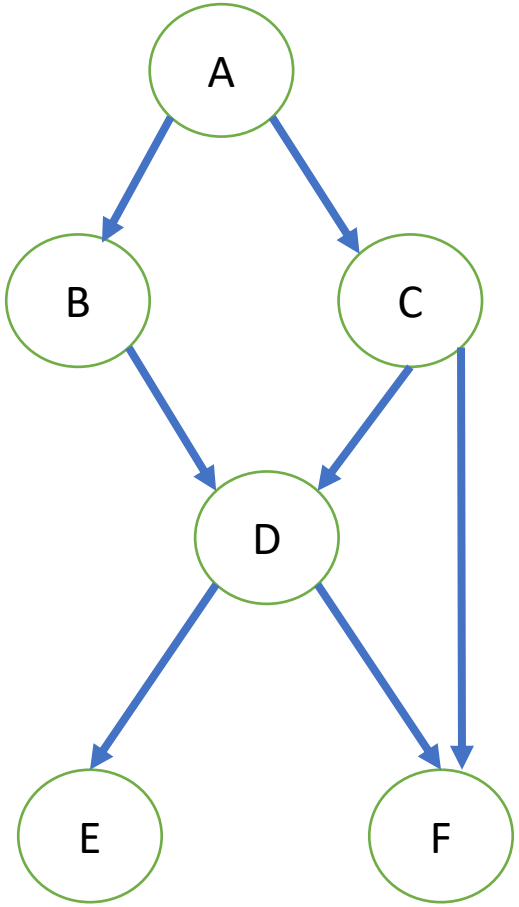
- $P(A, B, C, D, E, F) =$

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

- **Markov assumption:**

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

We will see in the next slides, what it means.



Markov assumption and chain rule

- The local Markov assumption :

$$P(X_1, \ldots, X_n) = \prod_1^n P(X_i | pa(X_i))$$

where $\boldsymbol{pa(X_i)}$ **is the set of parent** nodes for X_i.

- This is the decomposition of the global distribution $P(X_1, \ldots, X_n)$ into the local distributions for the X_i given their parents $pa(X_i)$.

- This decomposition is preferable to that obtained from the **chain rule**,

$$P(X) = \prod_{i=1}^{i=n} P(X_i | X_{i+1}, \ldots, X_n)$$

(the conditioning sets are typically smaller)

BNs & MRFs

- BNs are a special case of MRFs with a very specific type of clique factor (one that corresponds to a conditional probability distribution and implies a directed acyclic structure in the graph), and a normalizing constant of one.

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 .
- 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).**

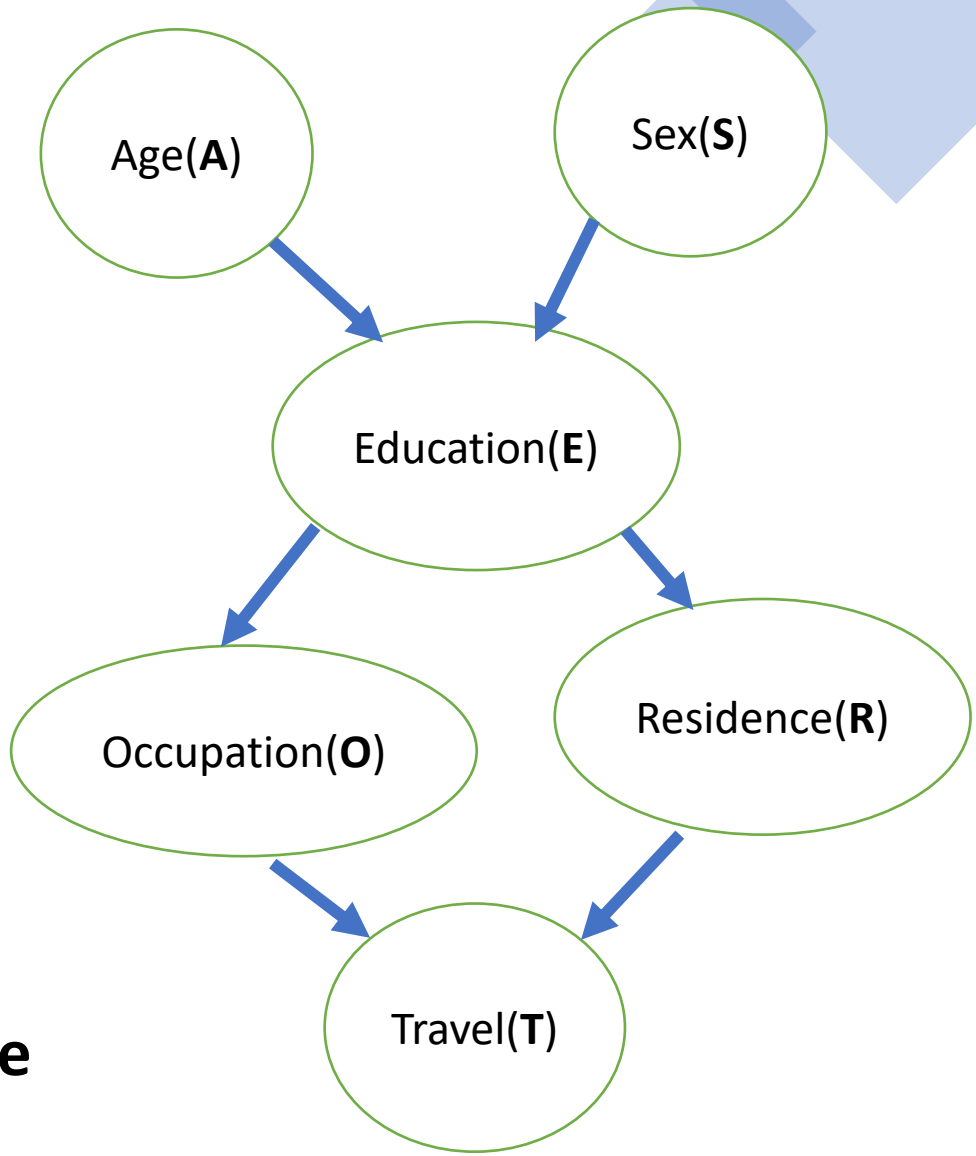
Data: Train Use Survey

"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**.



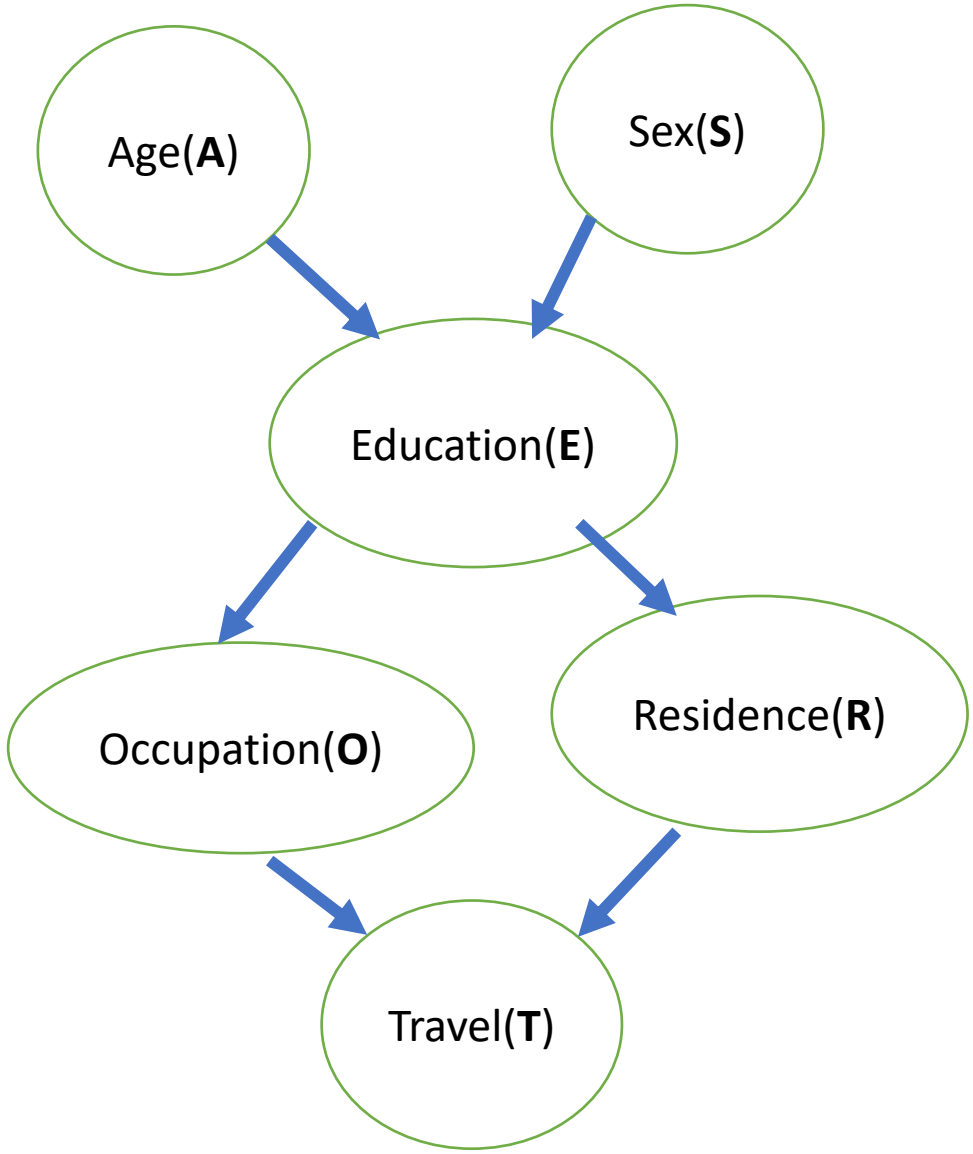
Example: Train Use Survey

The probabilistic relationship:

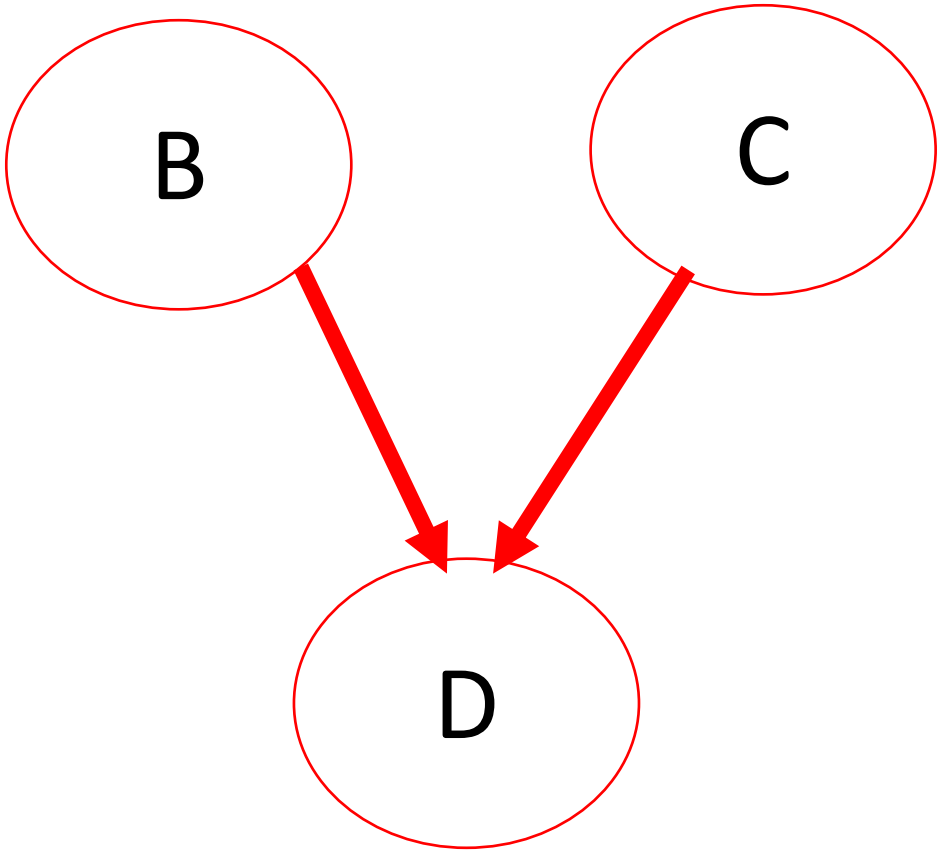
[A][S][E|A:S][O|E][R|E][T|O:R]

[child|parents]

This type of representation
is what we will see when we use
bnlearn package in practical.



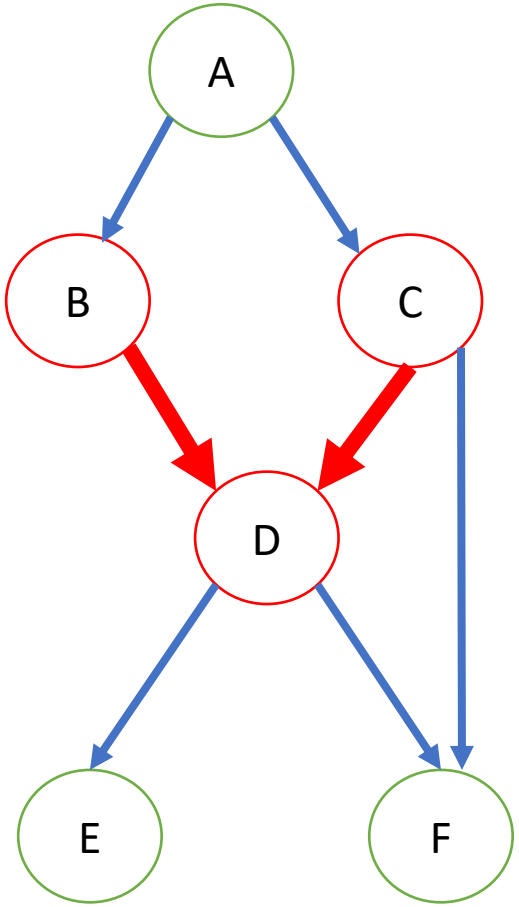
v-structure
Important for interpretation



v-structure

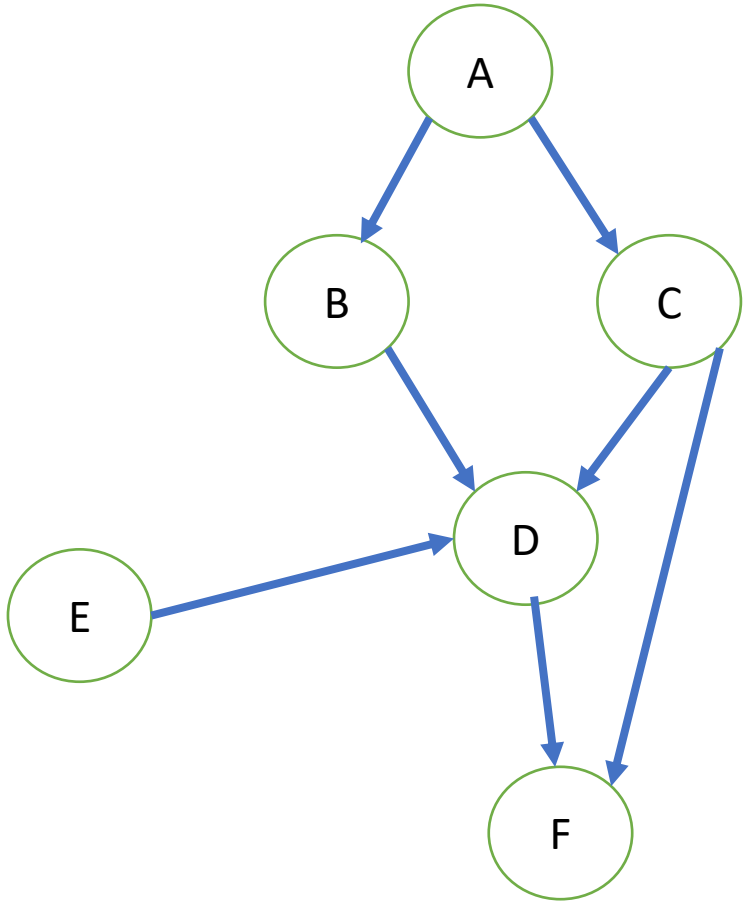
Important for interpretation

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



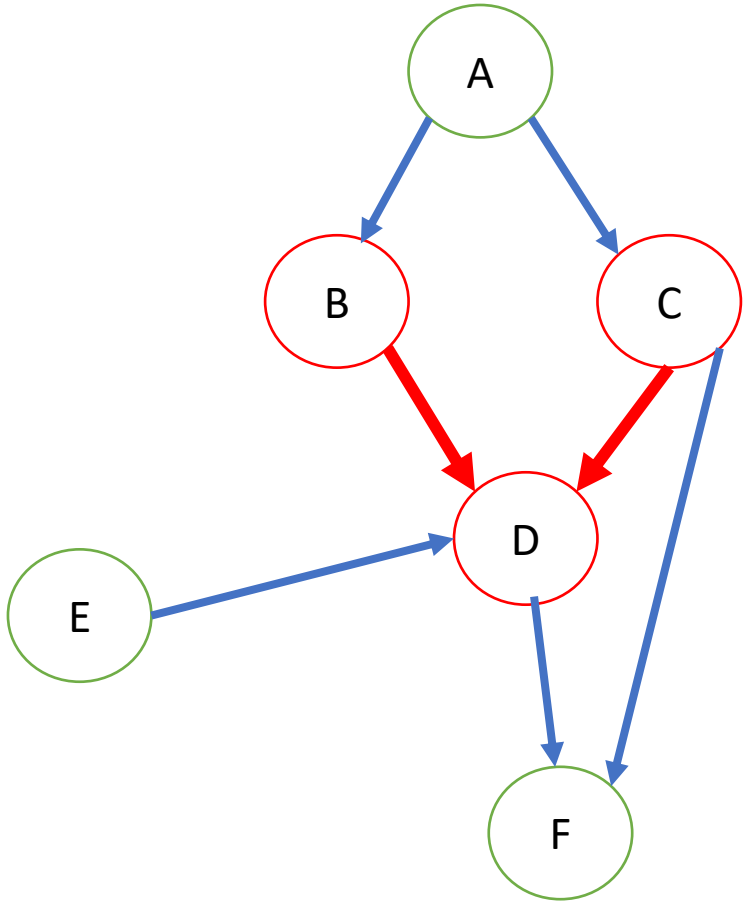
v-structure

How many v-sructure do you see?



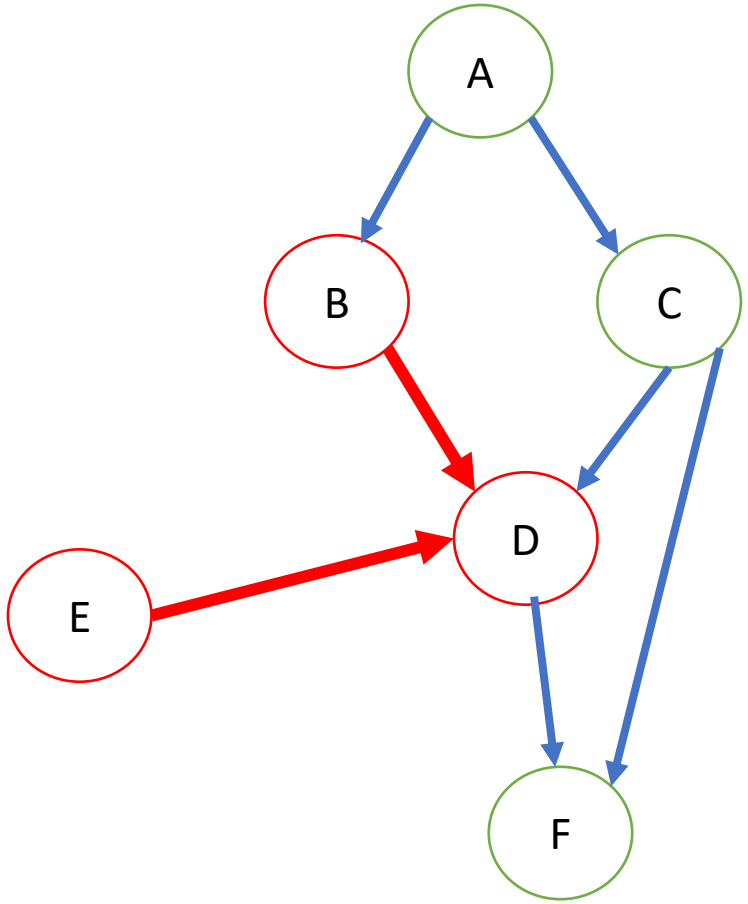
v-structure

1



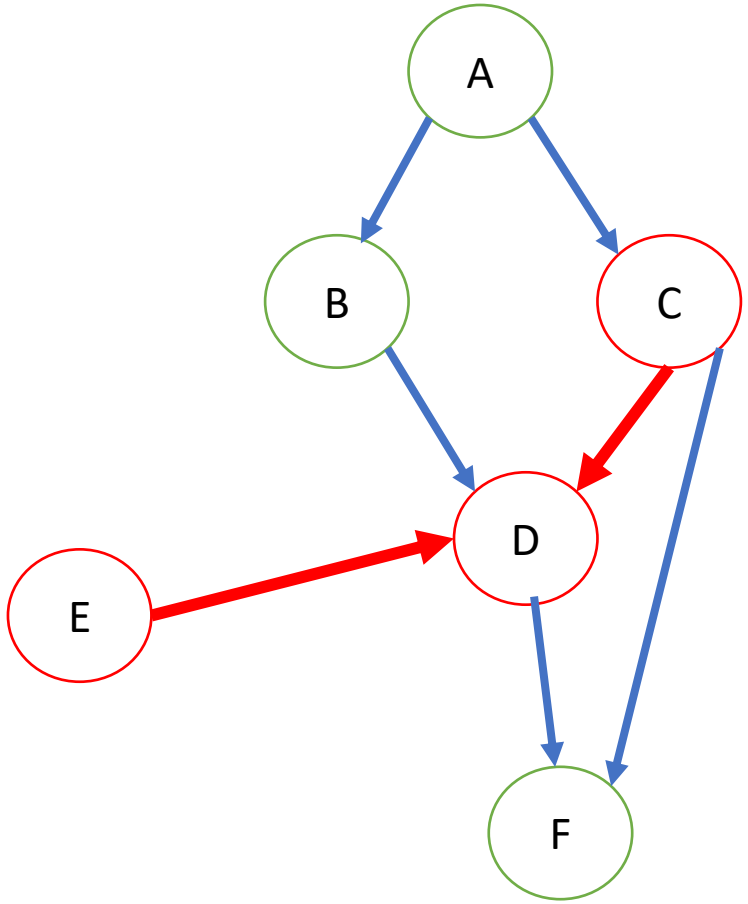
v-structure

2



v-structure

3



Markov Blanket

Graph: **G** and n nodes: $\boldsymbol{X_1, \ldots, X_n}$

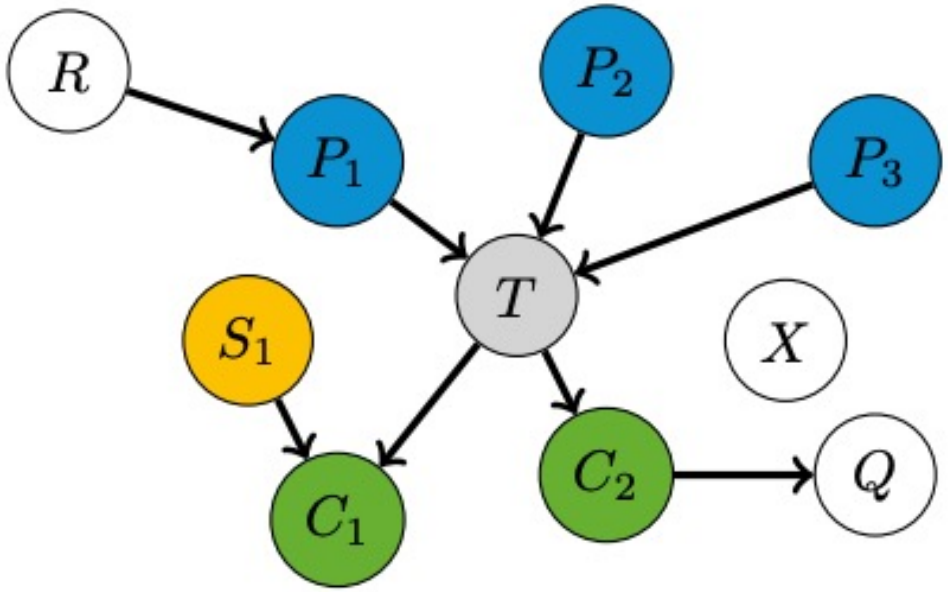
The Markov blanket of the node X_i $(i = 1, \ldots, n)$ includes:

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

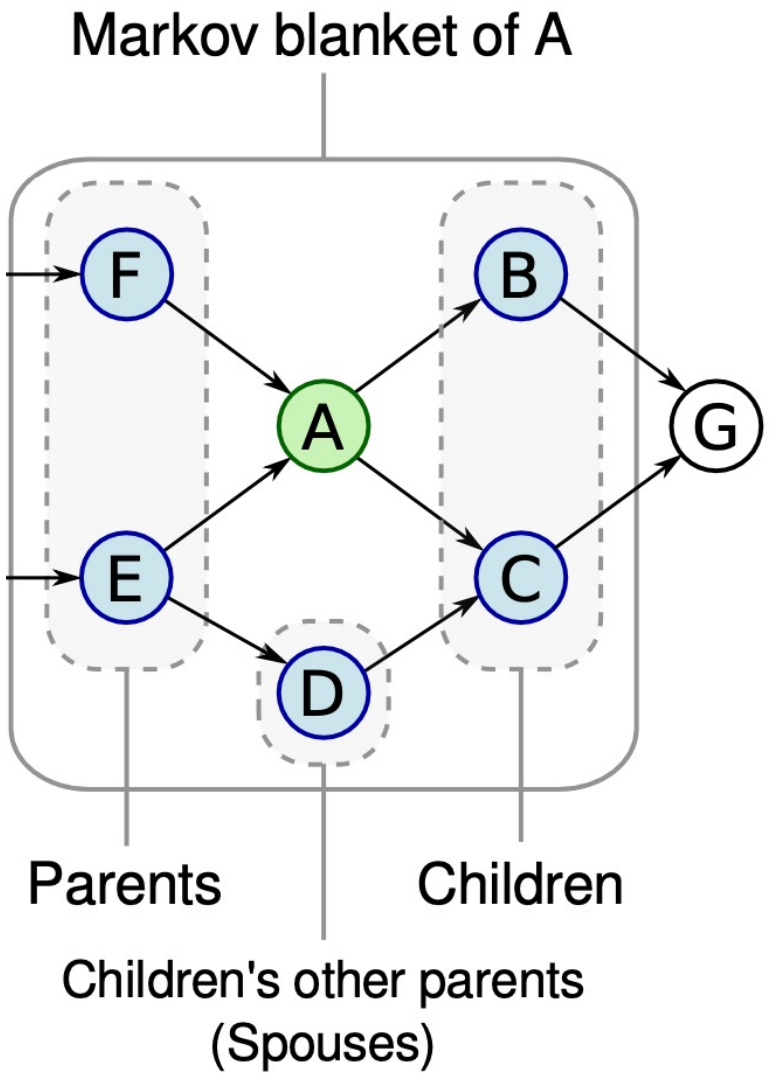
We denote the **Markov Blanket** of X_i symbolically as $\mathbf{MB}(\boldsymbol{X_i})$.

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

Markov Blanket



Example Markov blanket of a target T, consisting of **three parents (blue nodes), two children (green)** and **one spouse (orange)**. All other nodes are conditionally independent of T given MBT .

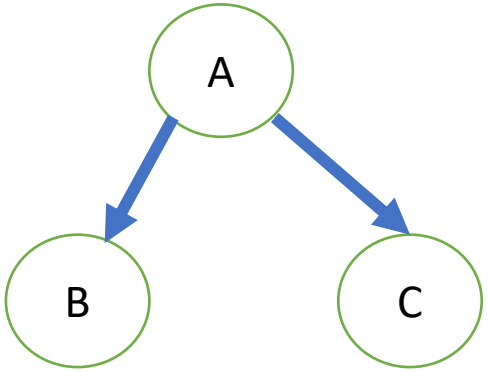


We can easily use the DAG to solve the **feature selection** problem.

We can restrict ourselves to the Markov blanket to perform any kind of inference on the target node, and disregard the rest.

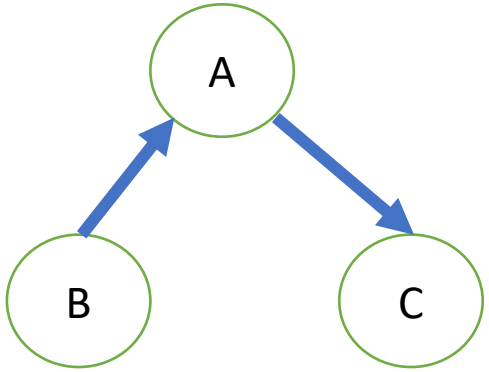
Since graphical separation implies probabilistic independence.

Fundamental connections



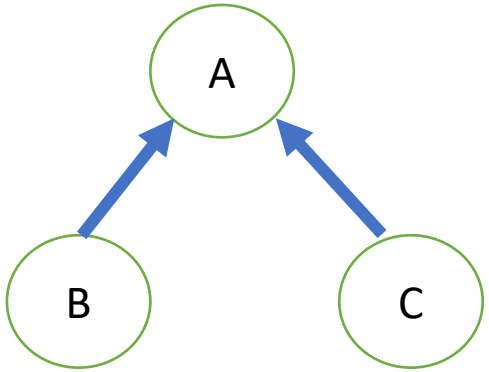
B ←A →C

Divergent connection



B →A →C

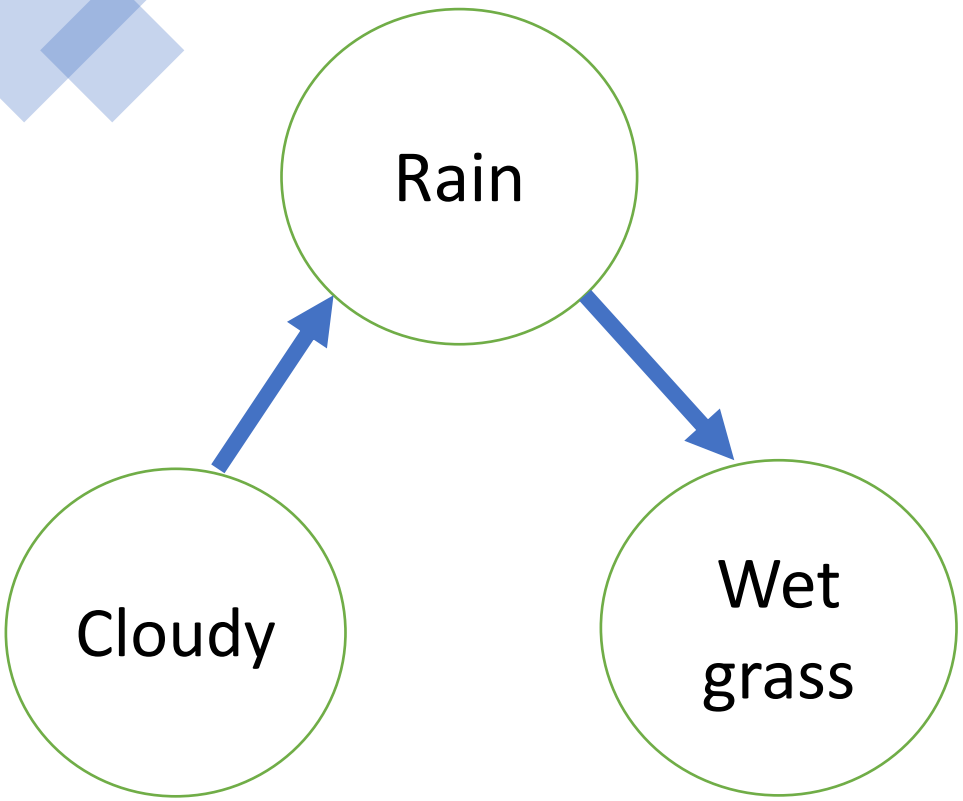
Serial connection



B →A ←C

Convergent connection

Some examples:

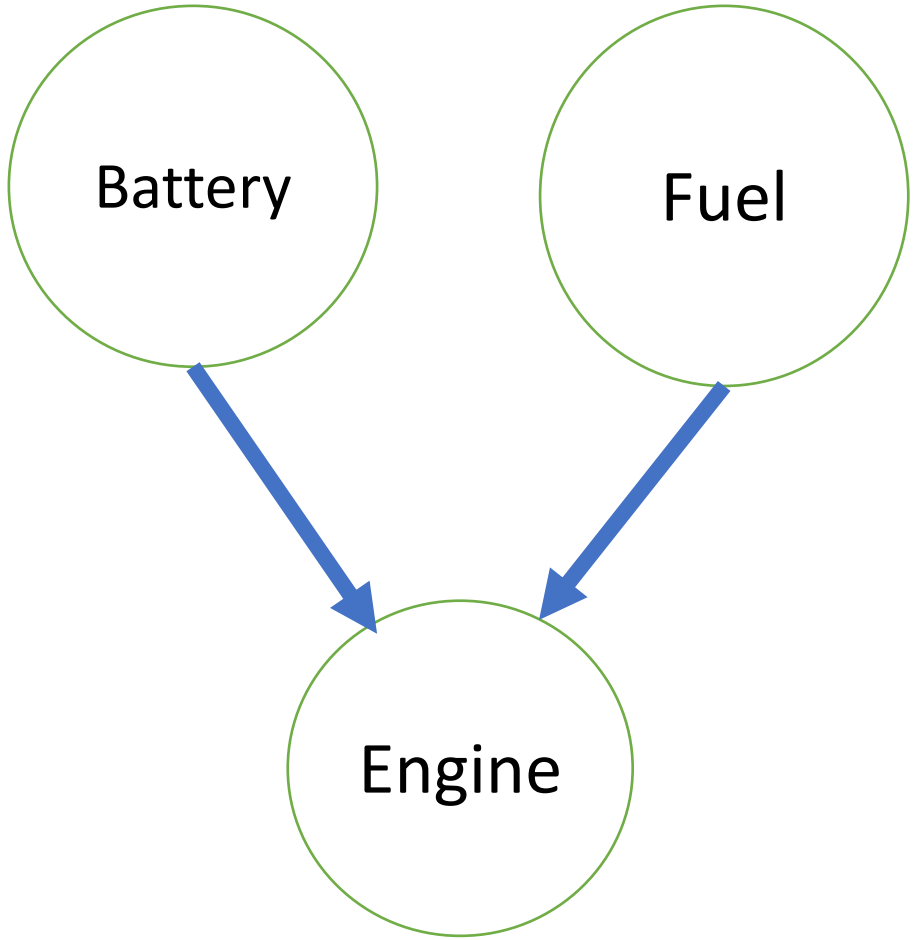


- Both variables **Cloudy** and **Grass** wet are **statistically dependent**.
- **Cloudiness** increases the probability of **rain** and thus indirectly the probability of a **wet ground**.
- **Conditional** on the variable **Rain**, the two variables **Cloudy and wet grass** are stochastically independent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

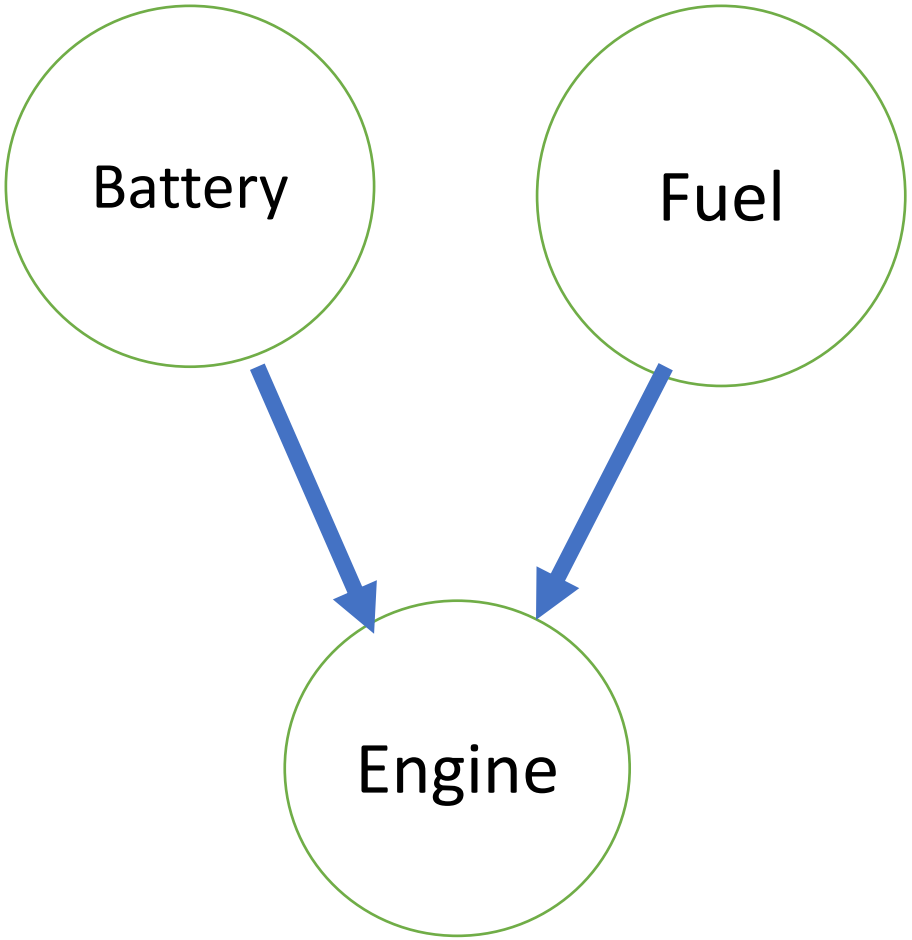
2) If it is known whether it rains or not, the condition of the grass has no influence on the probability of the state of the clouds.

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.

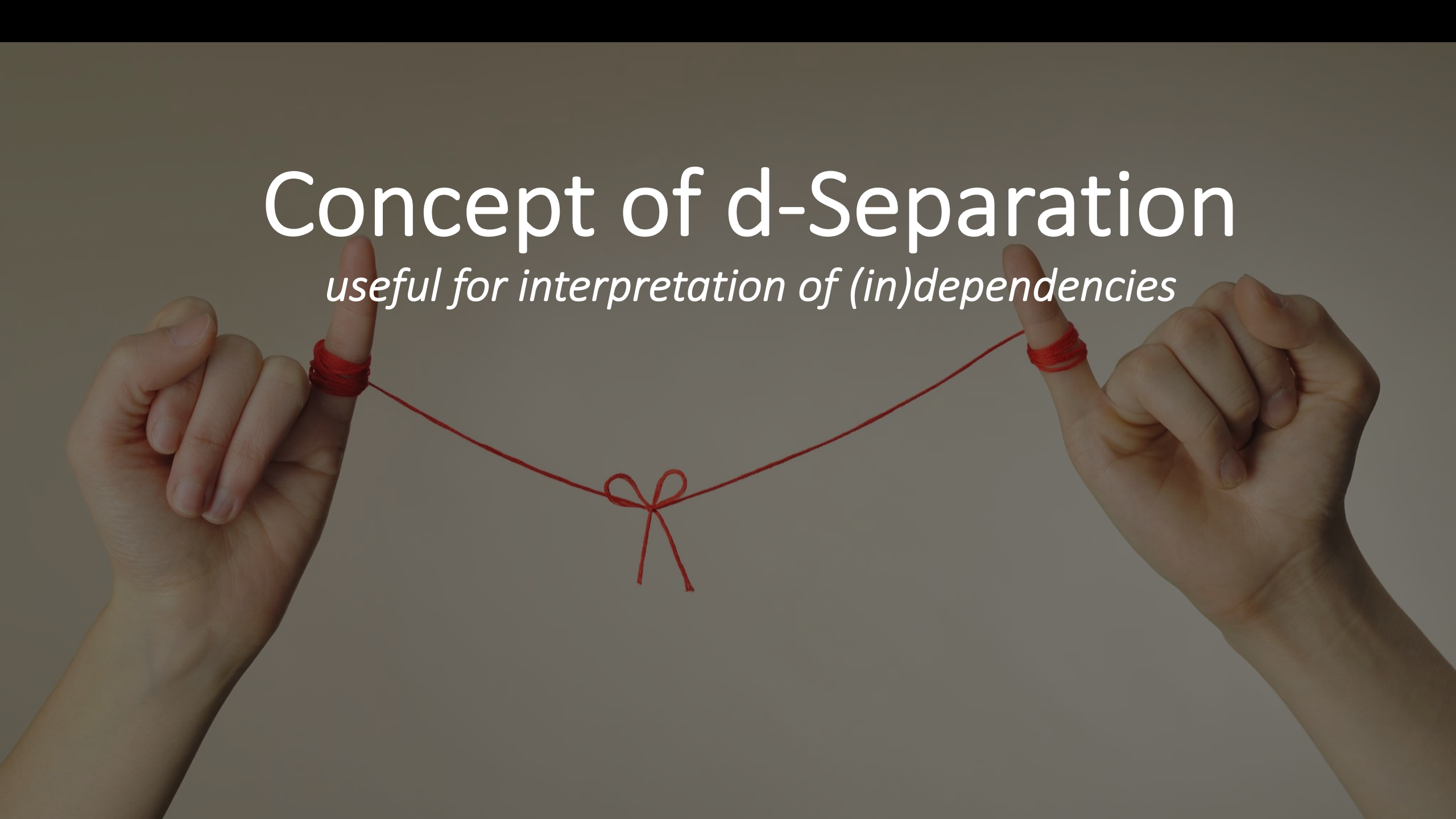
Some examples:



- **Battery** and **Fuel** are **stochastically independent**.
- 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



Concept of d-Separation

- In Bayesian networks, **the (in-)dependence relations** between the nodes (or variables) are easily obtained with the help of the **d-separation** .

- It is very useful for interpretation.

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

Path

- **Definition of directed path**

There is a directed path from node X_i to node X_j in a graph G if one can move from X_i to X_j by **following directed edges** (according to their edge directions) . A directed path from X_i to X_j implies that X_i is an ancestor of X_j .

$$X_i \rightarrow \ldots \rightarrow X_j$$

- **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.**

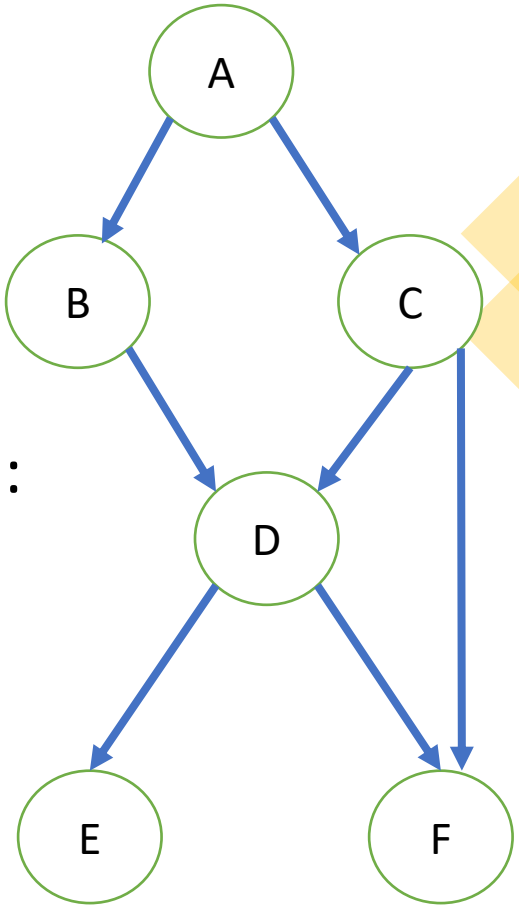
$$X_i \rightarrow X_{i+1} \leftarrow \ldots \rightarrow X_j$$

Example of (directed) paths

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

- In addition to the directed paths we have the paths (trails):
- A→B→D←C
- B→D← C
- A→C→F←D←B

- Not a valid path is e.g..:
- **A**→C→F←D←B←**A**←C why?



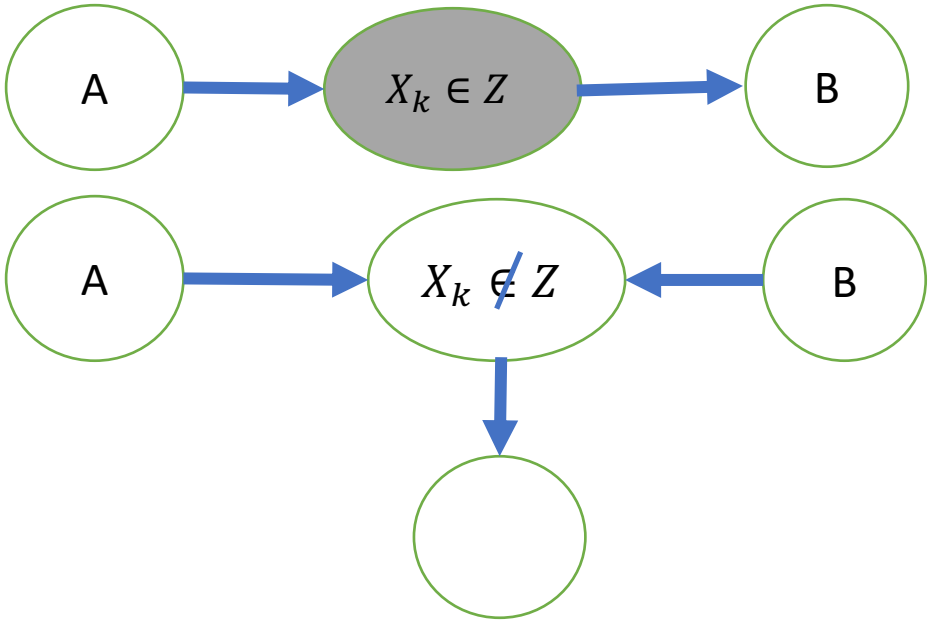
Collider

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

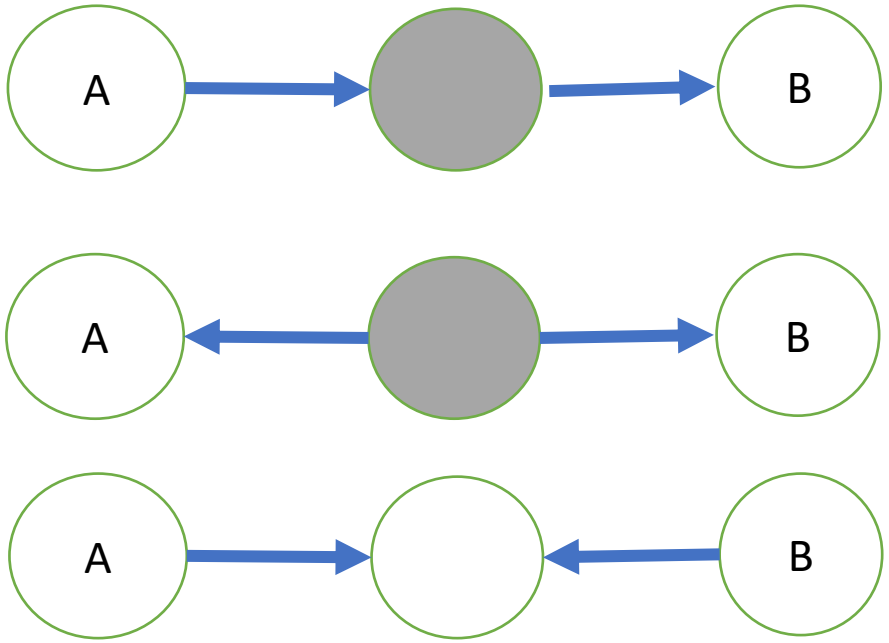


- Given nodes $X_1, \ldots, X_n$, we consider the nodes X_i and X_j $(i \neq j)$ and a subset Z of $\{X_1, \ldots, X_n\}$, where X_i and X_j are **not** in Z.

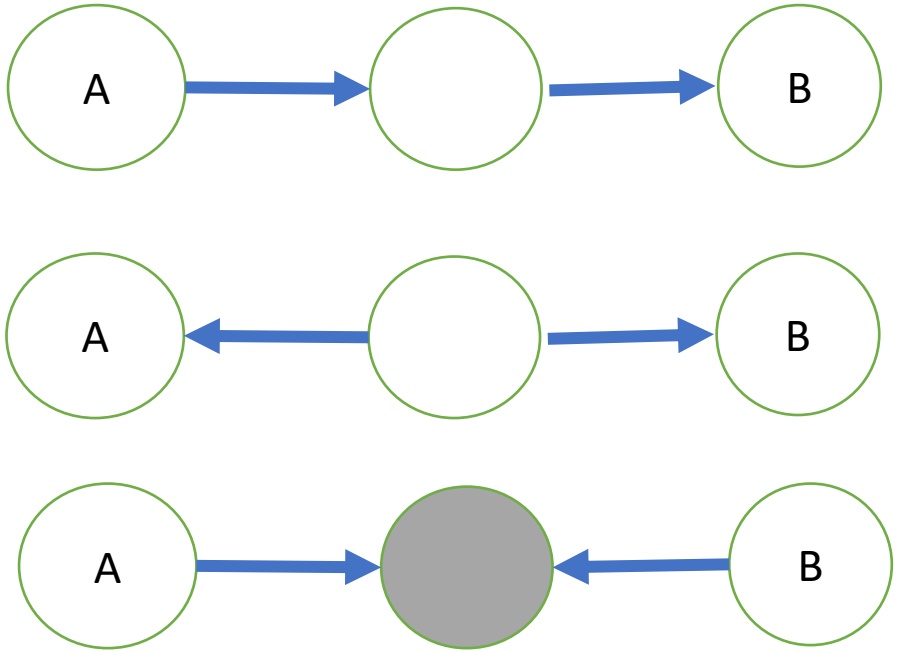
- A path (trail) between X_i and X_j is **blocked** conditional on Z when the trail leads through any node X_k and:

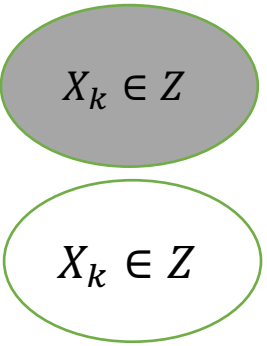
- (1) X_k **is not** a collider and X_k **is** an element of Z.
- (2) X_k **is** a collider and neither X_k nor a descendant of X_k is an element of Z.

Blocked paths between A and B



Open paths between A and B





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

Definition: d-Separation

d-Separation

- We consider the nodes X_i and X_j (i ≠j) and a subset Z of $\{X_1, \ldots, X_n\}$, with X_i and X_j being not in Z.

- **The nodes X_i and X_j are d-separated with respect to Z, when every path between X_i and X_j is blocked conditional on Z.**

- When X_i and X_j are d-separated with regard to Z, then X_i and X_j are **stochastically independent conditional on Z.**

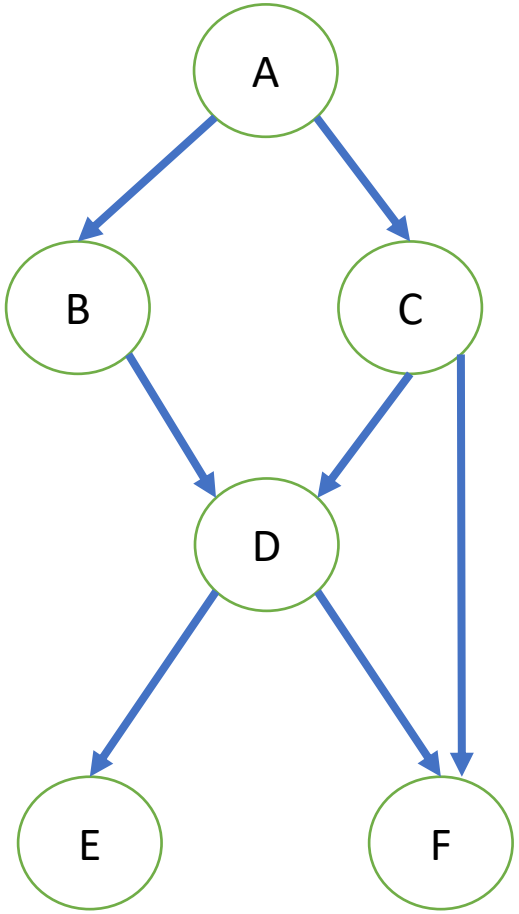
Example

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}

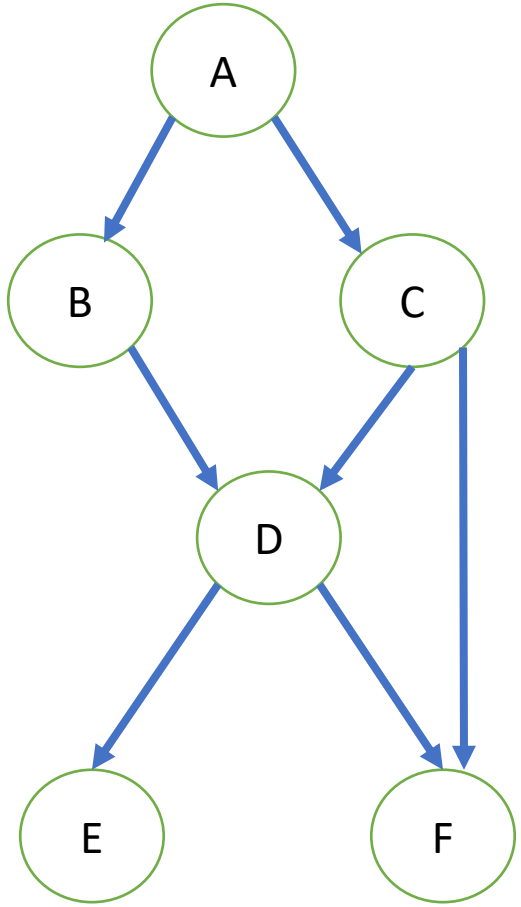
B and C are **not** d-separated conditional on Z={A,D} **why?**



Static 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 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.



- Second component:

The probability distribution $P(X)$

Types of BN:

- The **three most common choices** in the literature (by far), are:

- **Discrete BNs:** X and the $X_i \mid \Pi X_i$ are **multinomial**;

- **Gaussian BNs (GBNs):** X is **multivariate normal** and the $X_i \mid \Pi X_i$ are univariate normal;

Conditional linear Gaussian BNs (CLGBNs): CLGBNs contain both discrete and continuous nodes, and combine **discrete BNs** and **GBNs** to obtain a mixture-of-Gaussians network.

Discrete BNs

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

Multinomial distribution

- A multinomial distribution involves a process that has a set of k possible results (X_1, X_2, X_3, …, X_k) with associated probabilities (p_1, p_2, p_3, …, p_k) such that $\sum p_i = 1$.
- Then for n repeated trials of the process, let x_i indicate the number of times that the result X_i occurs, subject to the restraints that $0 \leq x_i \leq n \; and \; \sum x_i = n$.

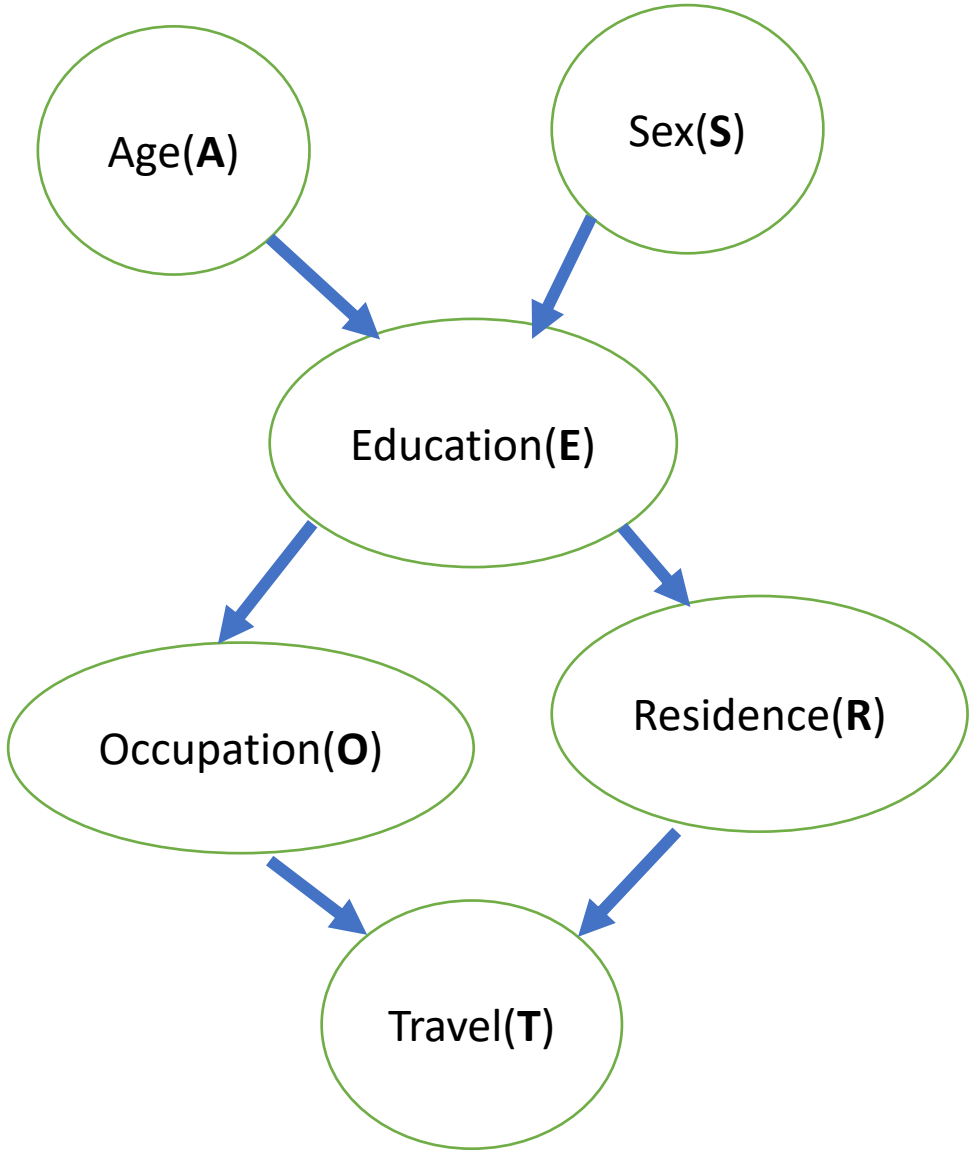
$$P(X_1 = x_1, X_2 = x_2, \dots, X_k = x_k) = \frac{n!}{x_1!\, x_2!\, \dots\, x_k!} p_1^{x_1} p_2^{x_2} \cdots p_k^{x_k}$$

Using BNs it can be decomposed

Example of Discrete BNs

Prior knowledge (experts):

- The network is known.
- **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.



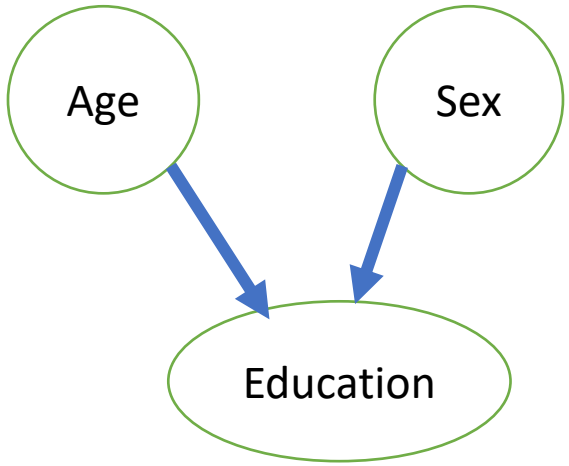
	Yong	Adult	Old
	0.3	0.5	0.2

Age

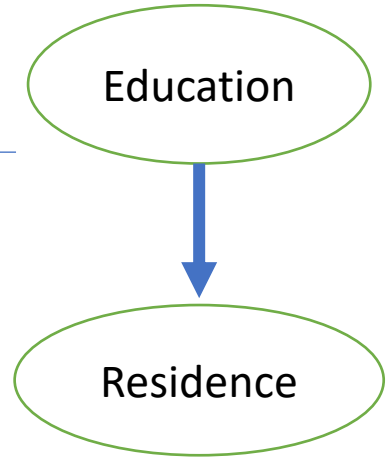
	M	F
	0.6.	0.4

Sex

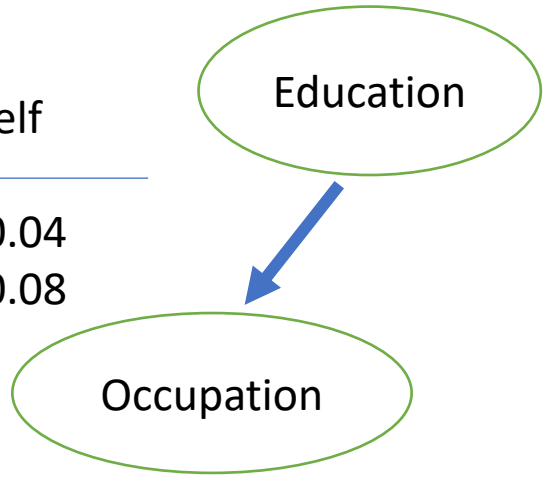
	High	Uni
Young& M	0.75	0.25
Adult& M	0.72	0.28
Old& M	0.88	0.12
Young& F	0.64	0. 36
Adult& F	0.70	0.30
Old& F	0.90	0.10



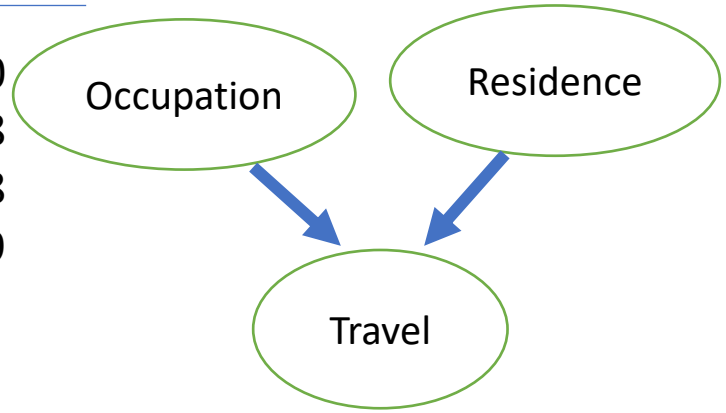
	Small	Big
High	0.25	0.75
Uni	0.20	0.80



	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

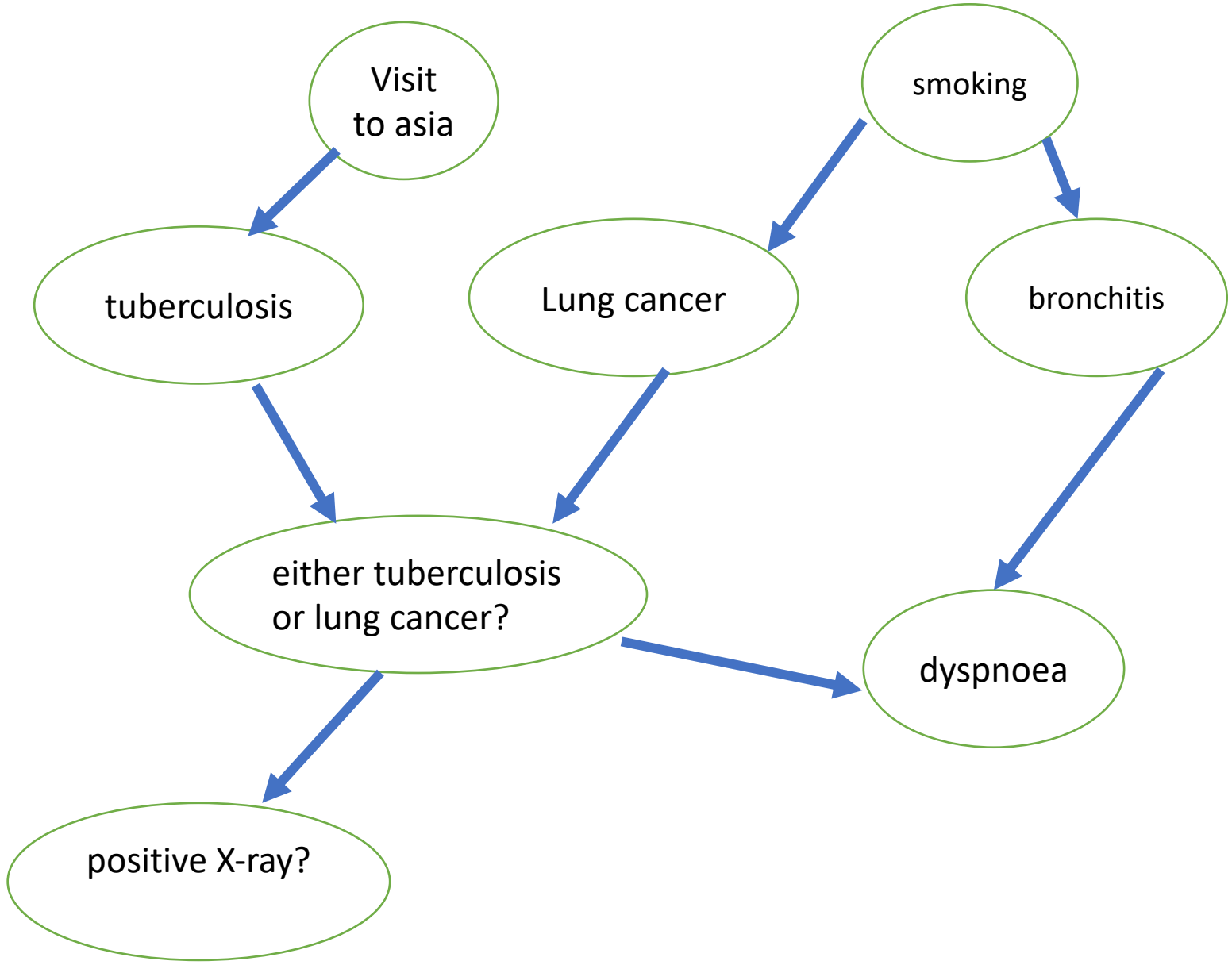


The set of all **local distributions** has, overall, fewer parameters than global distribtion.

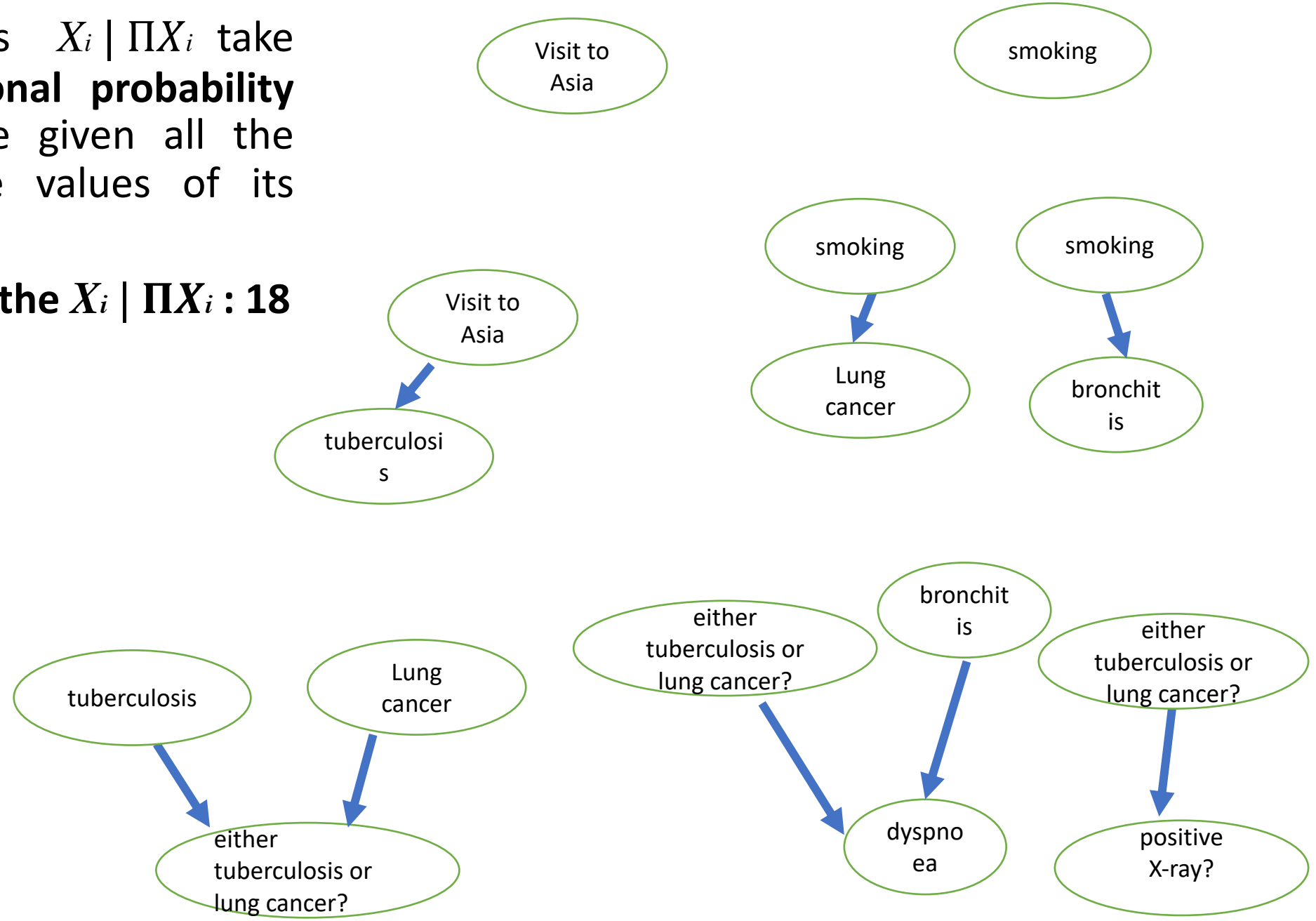
A **classic example of BN** is **the ASIA network** from Lauritzen & Spiegelhalter (1988), which includes a collection of binary variables. It describes a simple diagnostic problem for tuberculosis and lung cancer.

Total parameters of X :

$2^8 - 1 = 255$



- The local distributions $X_i \mid \Pi X_i$ take the form of **conditional probability tables** for each node given all the configurations of the values of its parents.
- **Overall parameters of the $X_i \mid \Pi X_i$: 18**



The set of all **local distributions** has, overall, fewer parameters than global distribtion.

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 invistigation.**

- E.g. genetics and systems biology to recunstruct the molecular pathways and networks underlying complex disease and metaboloc processes (sachs et al. (2005))

Learning the dag structure:

- Several algorithms have been presented in literature for this problem,
- Despite the variety of theoretical backgrounds and terminology they can all be traced to only three approaches:

- *constraint-based*,
- *score-based* and
- *hybrid*.

Constraint-based based:
Conditiona independence test

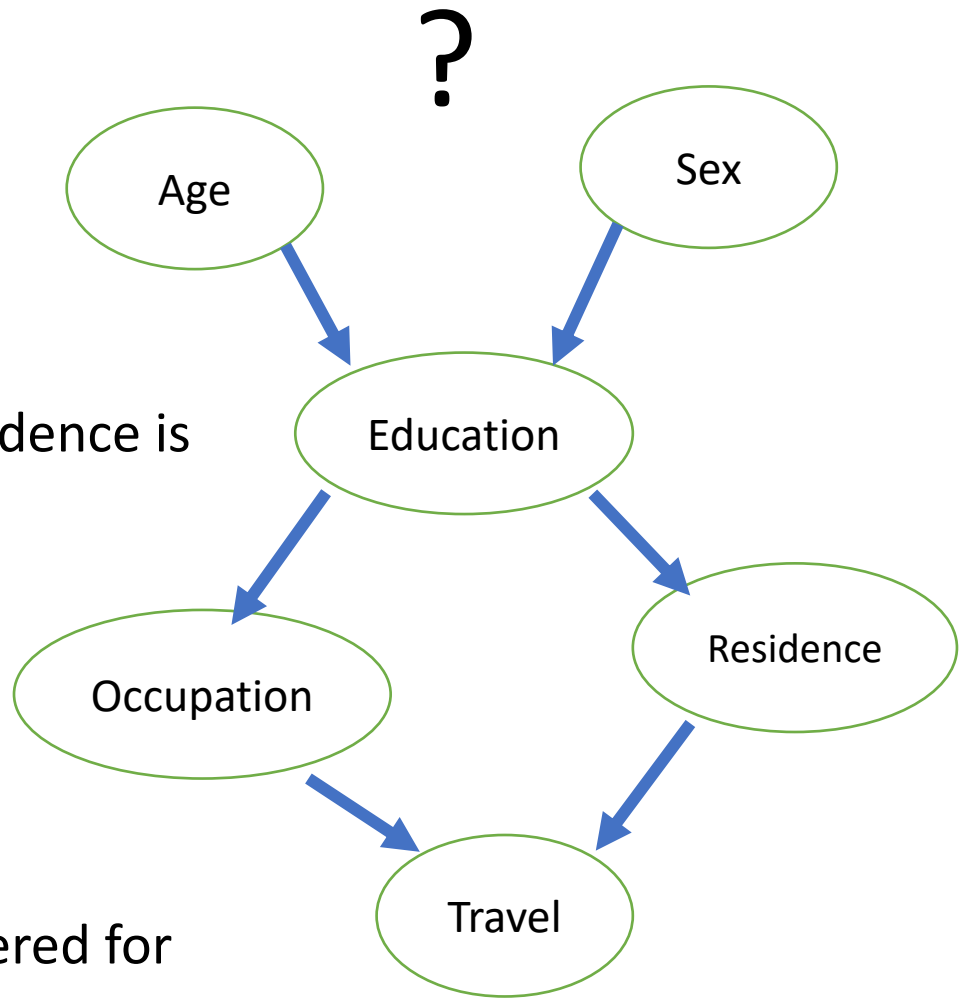
(Pearl 1990, Verma and Pearl, 1991)

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

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

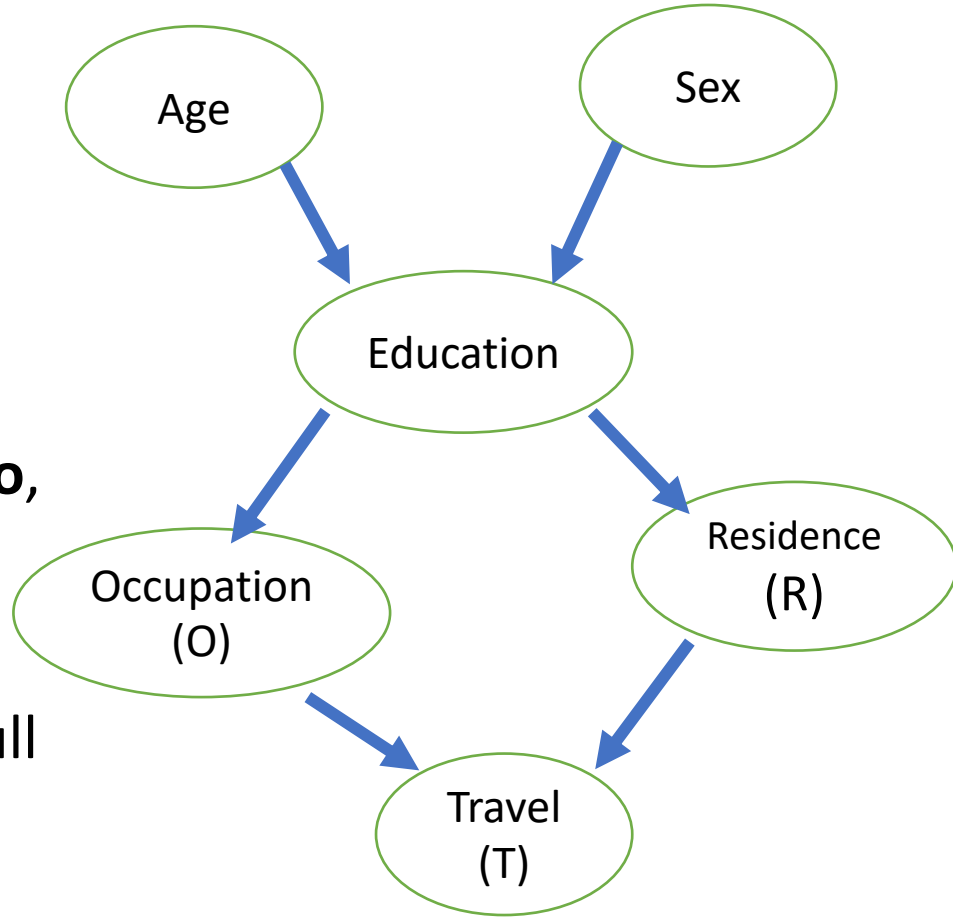
- If the **null hypothesis** is rejected, the arcs can be considered for **inclusion** otherwise for **exclusion**.

- 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.

- Both test have an asymptotic χ^2 **distribution** under the null hypothesis.

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

Conditional independence test

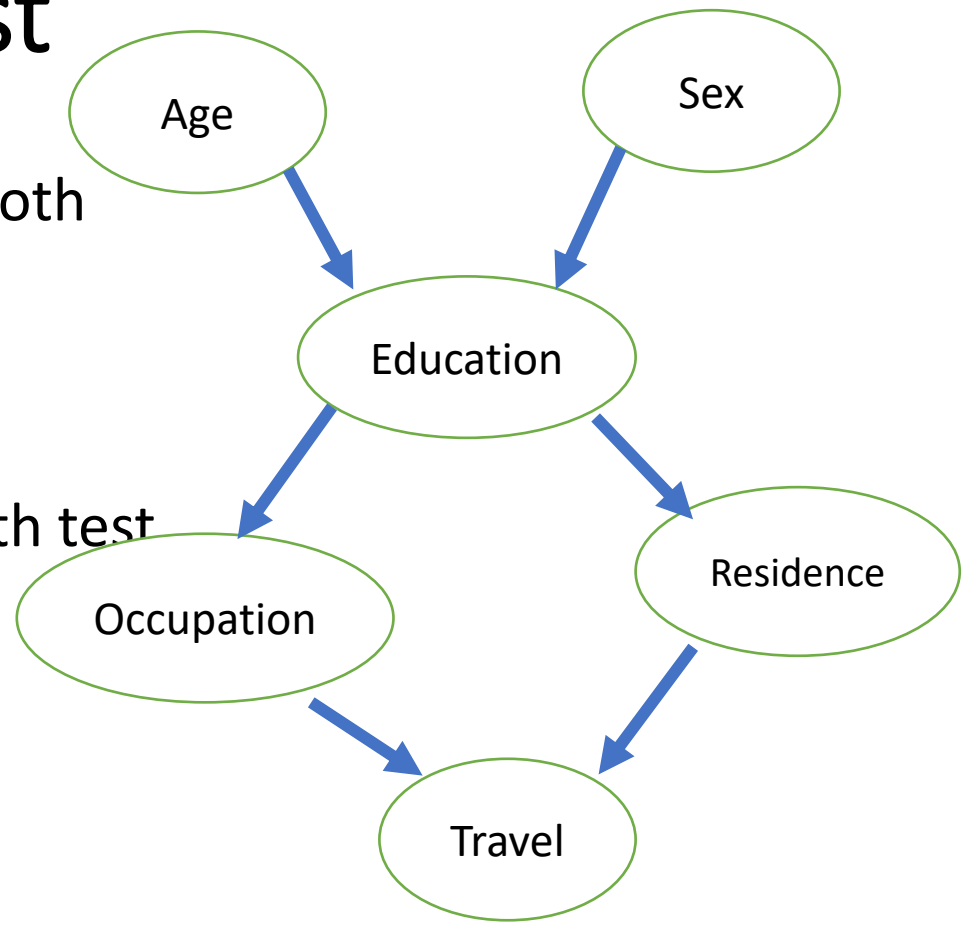
- **Conditional independence** results in **small values** of both tests statistics, **the null hypothesis** ***<u>is not</u>*** rejected),

  p-value > α ---- there is no edge

- **The null hypothesis** is rejected for **large values** of both test statistics.

  p-value < α----- there is an edge

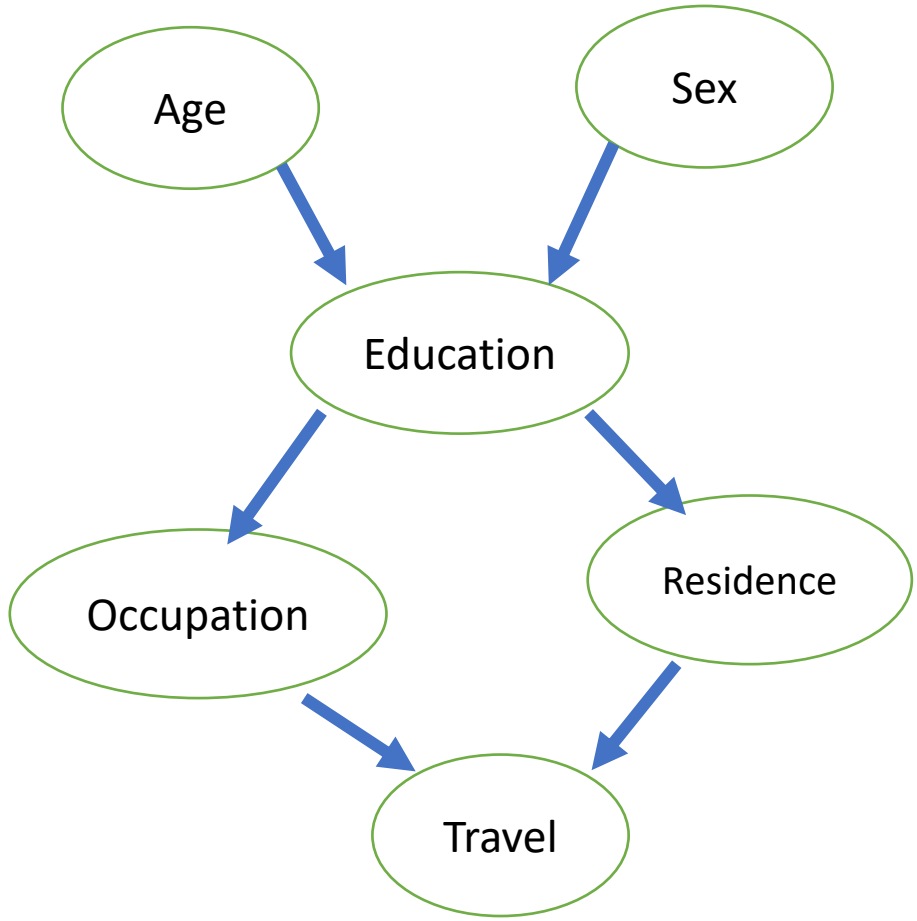
- We will discuss this in practical.



Network score

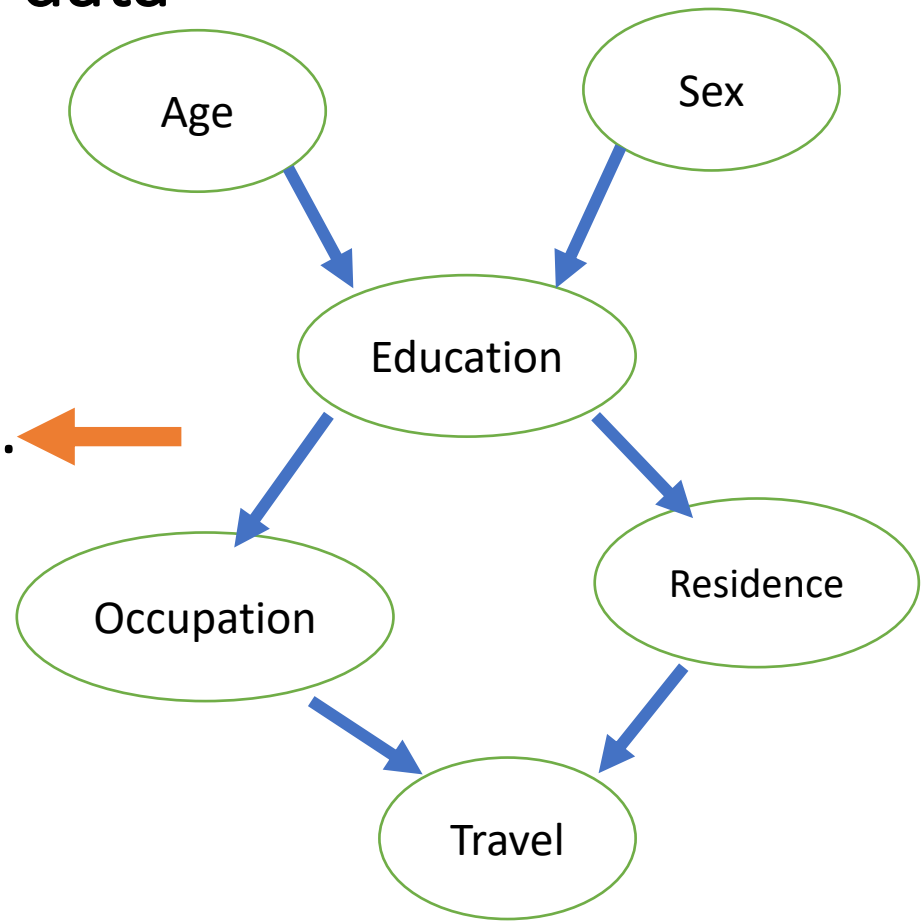
- Each candidate BN is assigned a network score reflecting its **goodness of fit.**

- Bayesian information criterion (**BIC**)
- Bayesian Dirichlet equivalent uniform(**BDeu**) (posterior probability of the DAG based on the uniform prior over the space of the DAG and the parameters.)



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

- **Greedy search algorithm (such as Hill-climbing).**
- **Genetic algorithm.**
- **Simulated annealing** (Bouckaert, 1995)

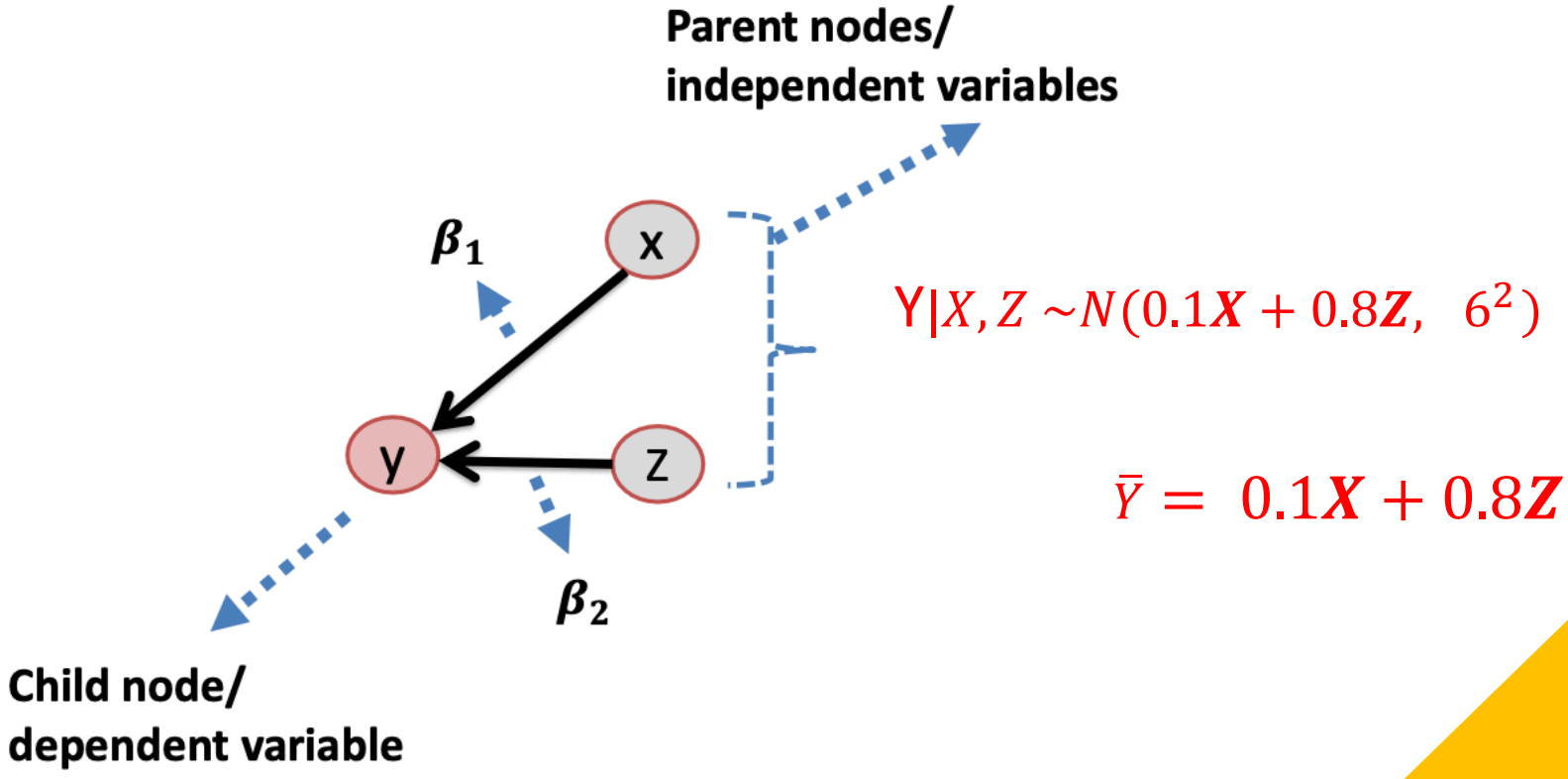


Continuous (Gaussian) Bayesian Network

Continuous BNs

- Every node follows a **normal distribution**.
- Nodes without any parents (root nodes), are discribed by the univariate normal distribution.
- The conditional effect of the parent nodes is given by an **aditive liner term in the mean (predictores)**, and does not effect the variance.
- 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, **without any intraction term**.

Edges:presence or absense of coefficient



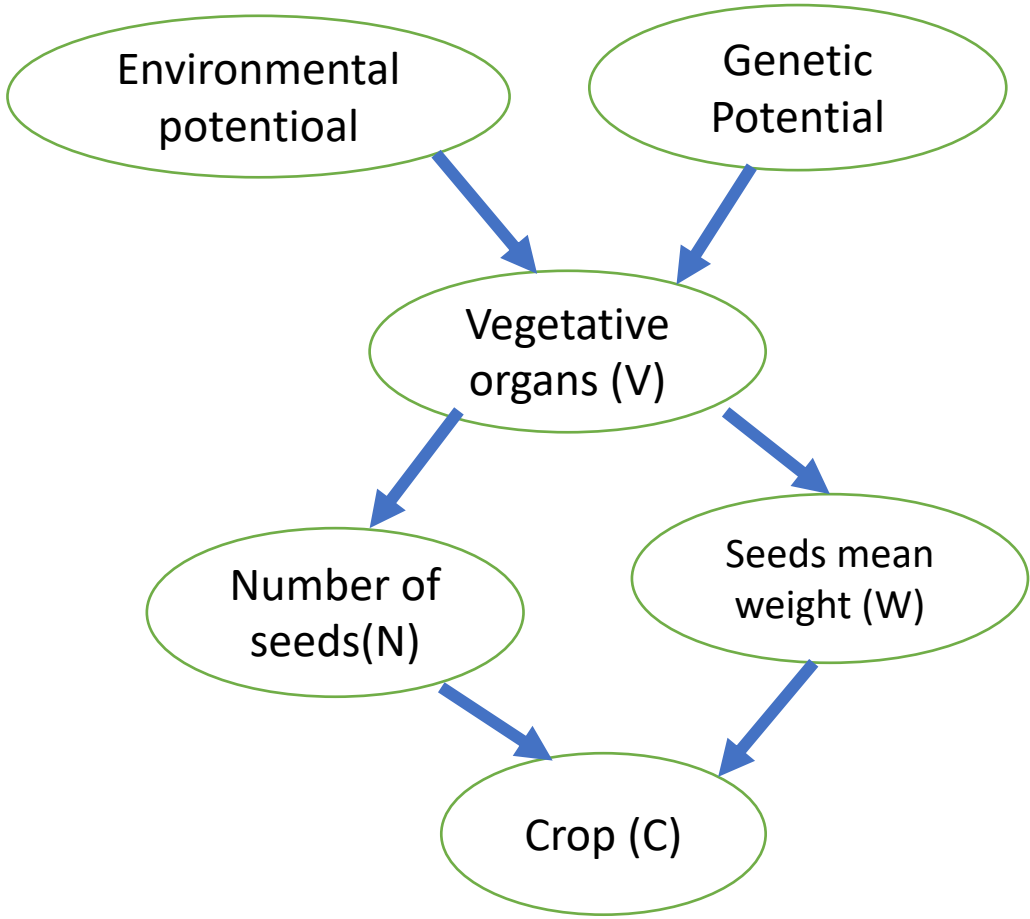
Example of continuous BNs

- For the analisis of a particular plant:

- The potential of the plant and of environment;

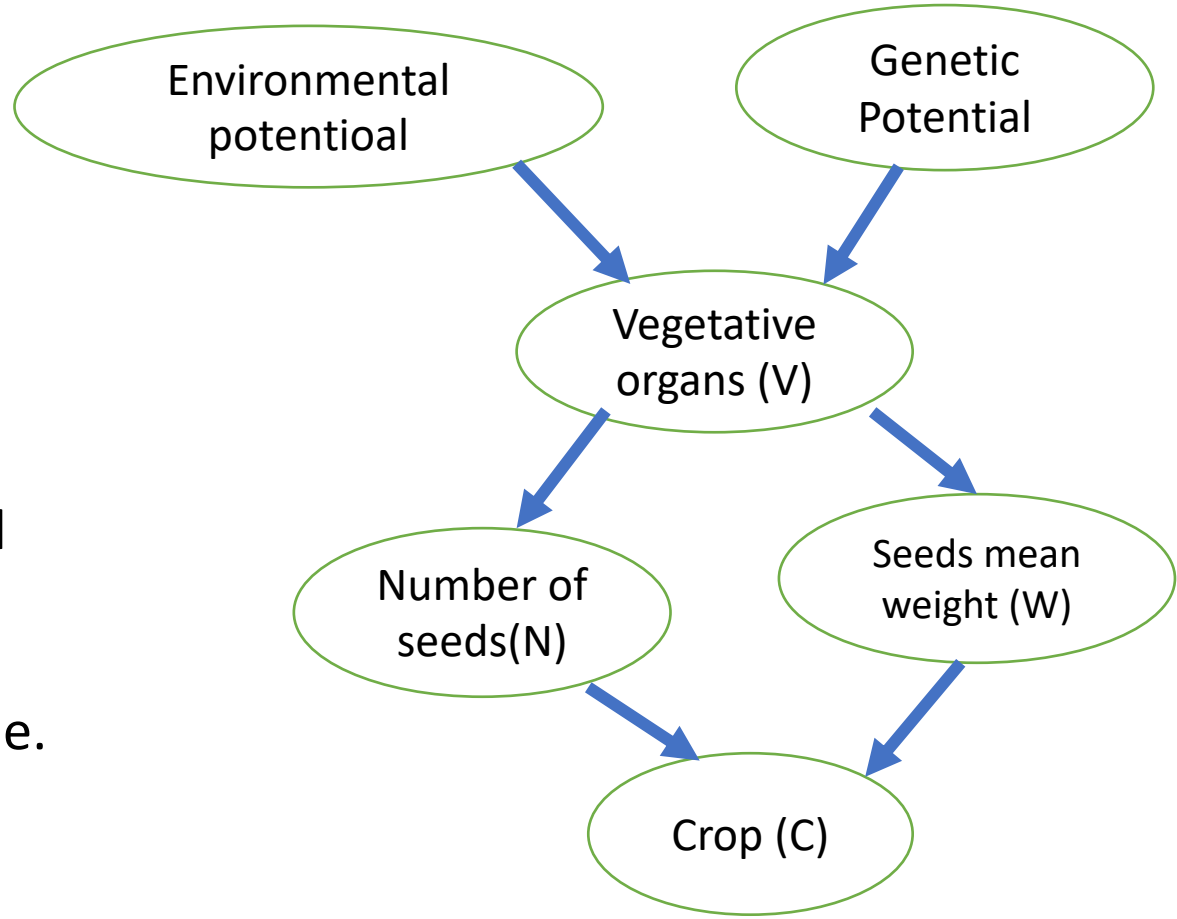
- The production of vegatitive mass;

- The harvasted grain mass, which is called the crop.



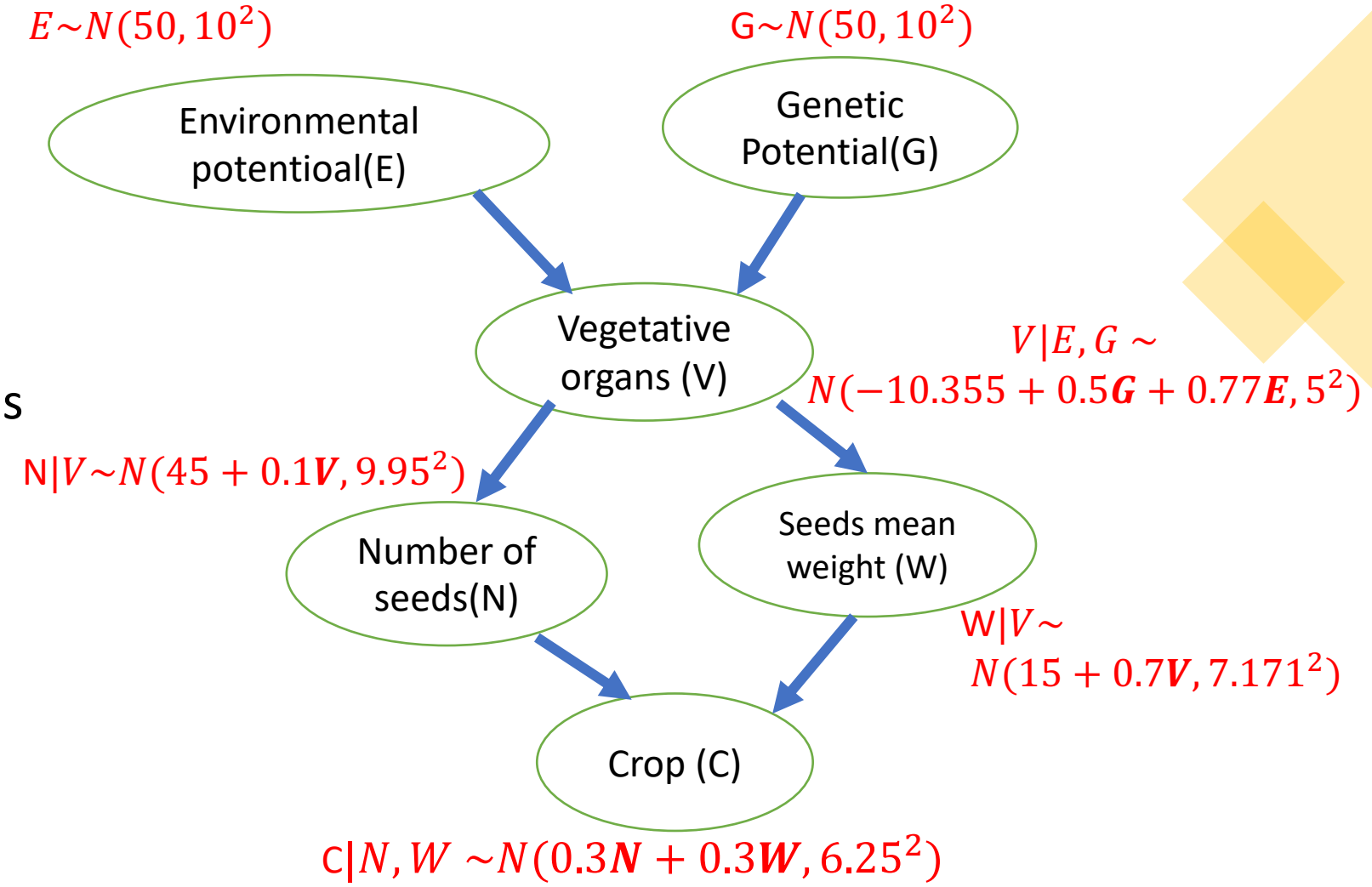
Example of continuous BNs

- **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.



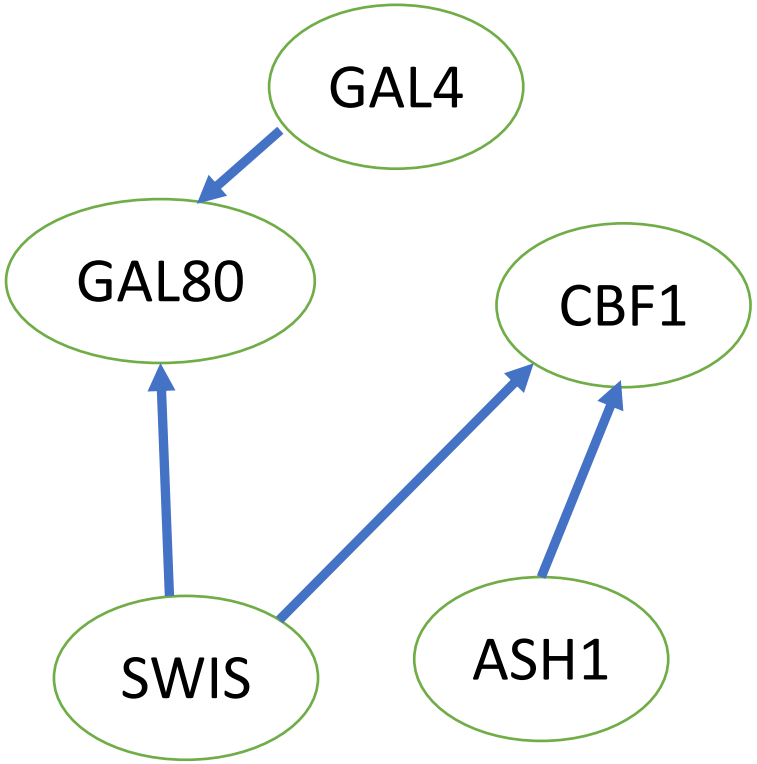
Example

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



Example 2: Infered Yeast network using BNs.

Network of n = 5 genes in Saccharomyces cerevisiae (yeast). The data obtained from synthetically designed yeast cells grown with different carbon sources: galactose (“switch on”) or glucose (“switch off ”), Cantone et al. (2009) .



$$GAL80 \mid \text{GAL4}, \text{SWIS} \sim N(0.6 \cdot \text{GAL4} + 0.8 \cdot \text{SWIS}, 2^2)$$

there is no loop here

Some remarks

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

- For small variations any continuous function can be approximated by an **additive function**, *e.g.*, a first-order Taylor expansion.

- Relatively **simple models** often perform better than very **sophisticated** ones when few observations are available.

$$G \sim N(50, 10^2)$$

$$E \sim N(50, 10^2)$$

$$V|E, G \sim N(-10.355 + 0.5G + 0.77E, 5^2)$$

$$N|V \sim N(45 + 0.1V, 9.95^2)$$

$$W|V \sim N(15 + 0.7V, 7.171^2)$$

$$C|N, W \sim N(0.3N + 0.3W, 6.25^2)$$

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**.

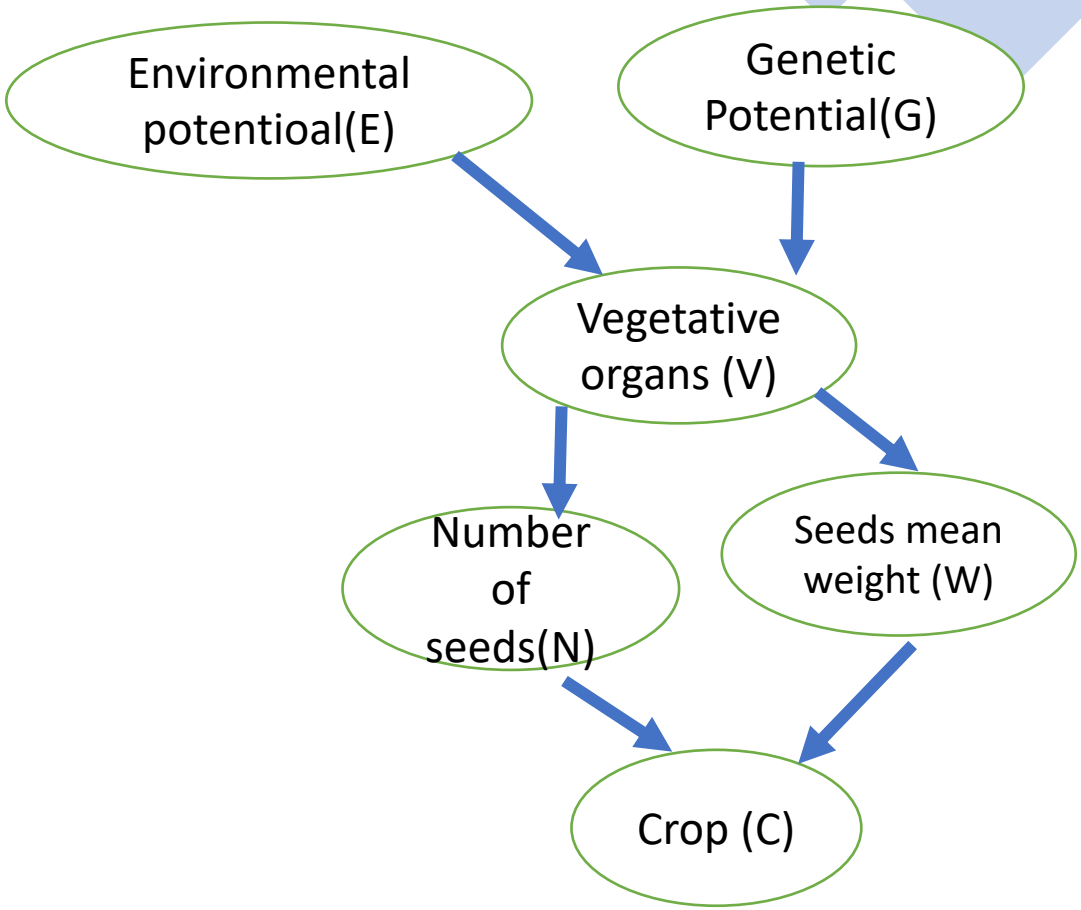
Conditional Independence Tests

- **Most common**: exact test for *partial correlations*. (equivalent to setting $\beta_W = 0$ in the regression model).

- H_0: C $\perp\!\!\!\perp_{\boldsymbol{P}}$ W |N
- H_1: C $\not\perp\!\!\!\perp_{\boldsymbol{P}}$ W |N

- The correlation we need to test is the **partial correlation** between C and W given N, say $\rho_{C,W|N}$,
- $C \perp\!\!\!\perp_{\boldsymbol{P}} W \,|N$ if and only if $\boldsymbol{\rho_{C,W|N}}$ is not significantly **different from zero**; it can be shown that

$\boldsymbol{\beta_W = 0}$ **if and only if** $\boldsymbol{\rho_{C,W|N} = 0}$



Conditional Independence Tests using bnlearn package

- Using “bnlearn” package
- Test for partial correlations
- Computing the corresponding statistics.
- The distribution for the test under the null hypothesis is a **Student’s t distribution**.

Conditional Independence Tests

- If the tested variables are conditionally independent, statistics is close to zero;
- large values, either positive or negative, are indicative of the presence and of the direction of the conditional dependence.

- So, in the case of $\rho_{C,W|N}$, we can say that C has a significant positive correlation with W given N and reject the null hypothesis of independence with an extremely **small p-value**.

- More in practical.

Network Scores

Same as Discrete BN:

- **BIC**,
- **Posterior probability score** (uniform prior over the space of DAGs and of the parametes, Bayesian Guassian equvalent score (**BGe**))

Summary

- Bayesian networks

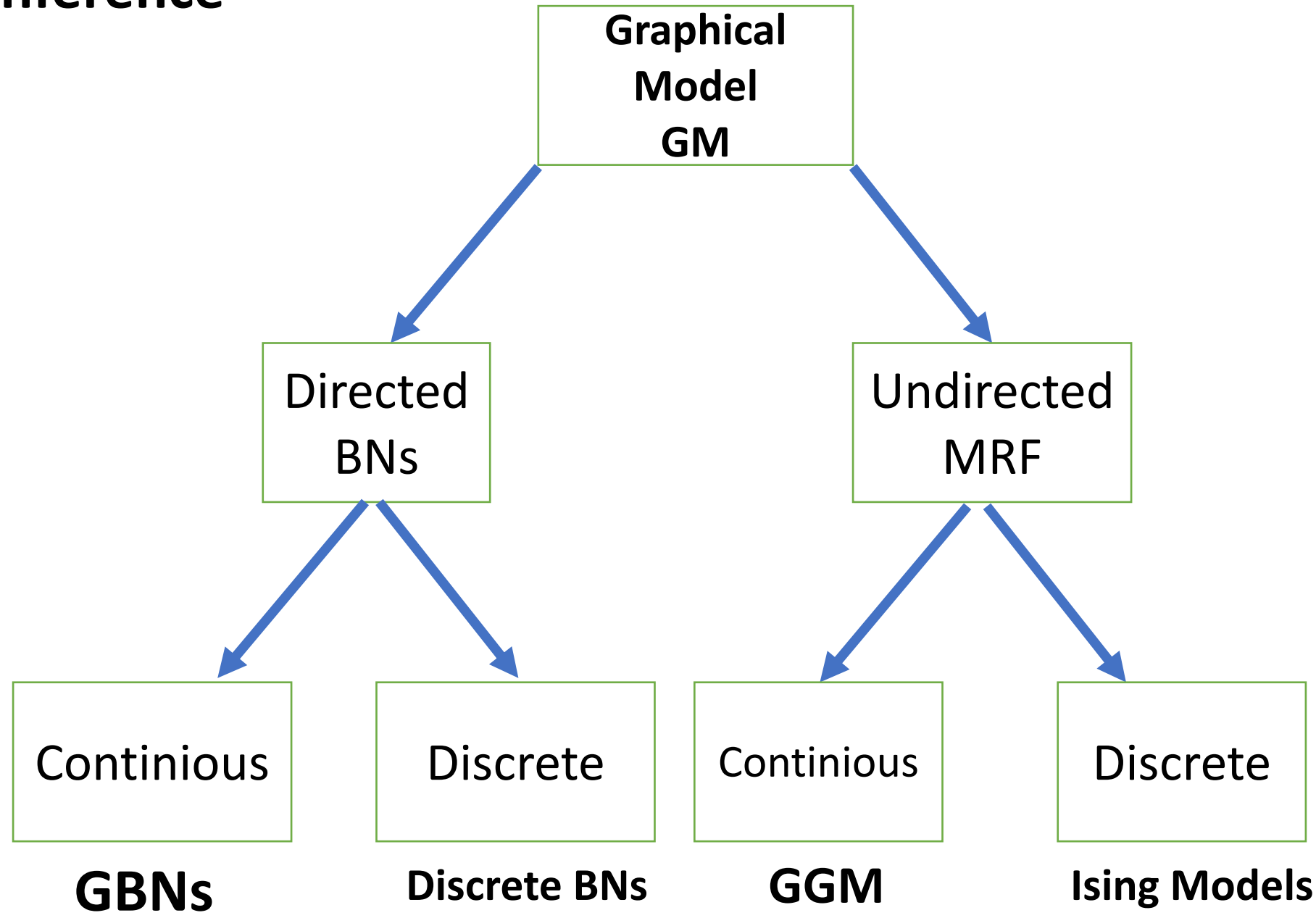
- are a combination of a DAG and a global distribution, both defined on the same variables.

- provide a systematic decomposition of the global distribution into lower-dimensional local distributions, in a divide-and-conquer way.

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

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

Network Inference



Finally, practically: *MRFs vs BNs*

- **MRFs** have more power than **BNs**, but are more difficult to 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 .

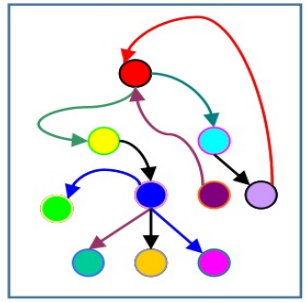
In a nutshell

Raw data

Cleaned data

Machine Learning

Statistical Methods



Network inference

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