Machine Learning 2

Graphical Models
- Bayesian Networks

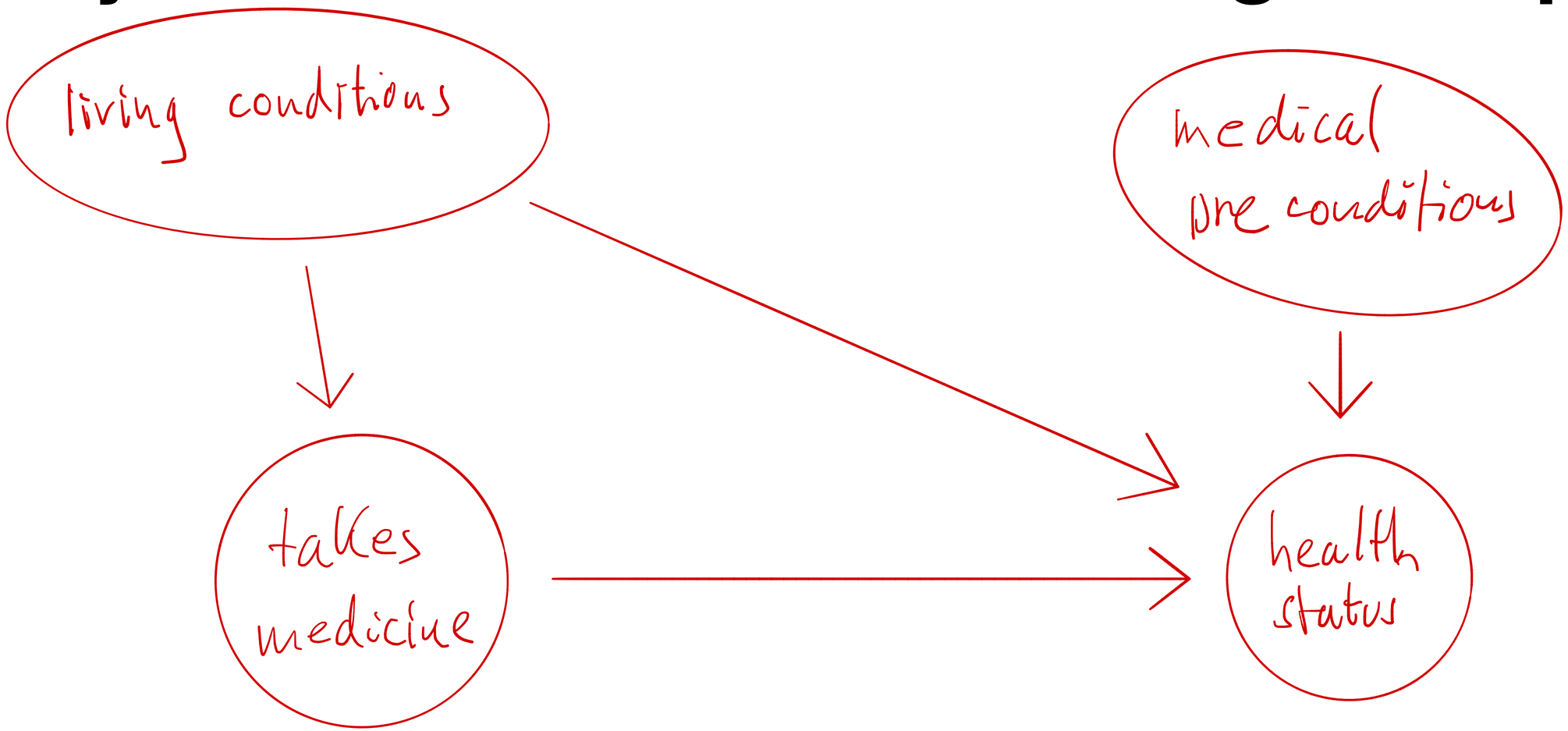
Patrick Forré

Machine Learning 2

Graphical Models
- Bayesian Networks
- Definitions

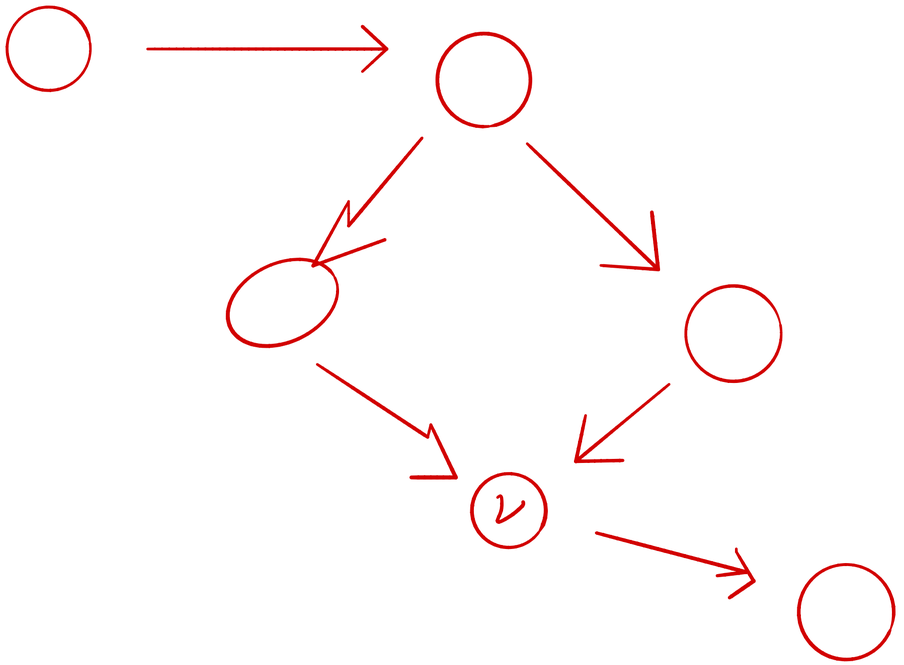
Patrick Forré

Bayesian Networks - Motivating Example

- 

Directed Acyclic Graphs (DAGs) - Definitions

- A Directed Acyclic Graph (DAG) is a directed graph $G = (V, E)$ with:
  - set of nodes/vertices V,
  - set of directed edges E
  - such that there is no directed cycle

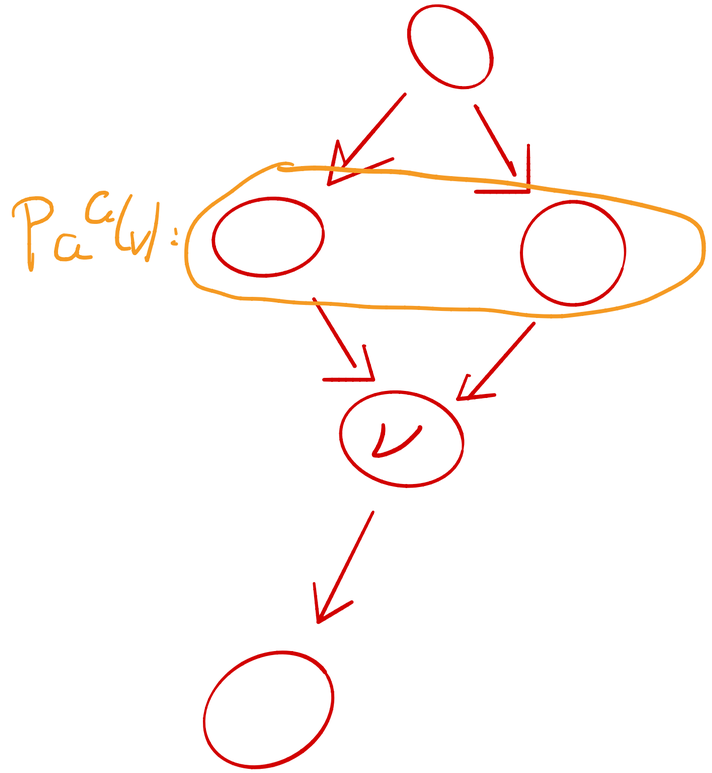


- Parents of v: nodes pointing to v.
- Children of v: nodes that v points to.
- Ancestors of v: nodes with directed path ending in v. (including v itself)
- Descendents of v: nodes that have a directed path from v pointing to it.

Factorization Property

- Let V be an (index) set of random variables $X_1, \ldots, X_M$ with a joint distribution $p(x_1, \ldots, x_M)$.
- Let $G = (V, E)$ be a DAG.
- We say that $p(x_1, \ldots, x_M)$ factorizes over G if:

$$p(x_1, \ldots, x_M) = \prod_{v \in V} p(x_v \mid x_{\mathrm{Pa}_G(v)})$$



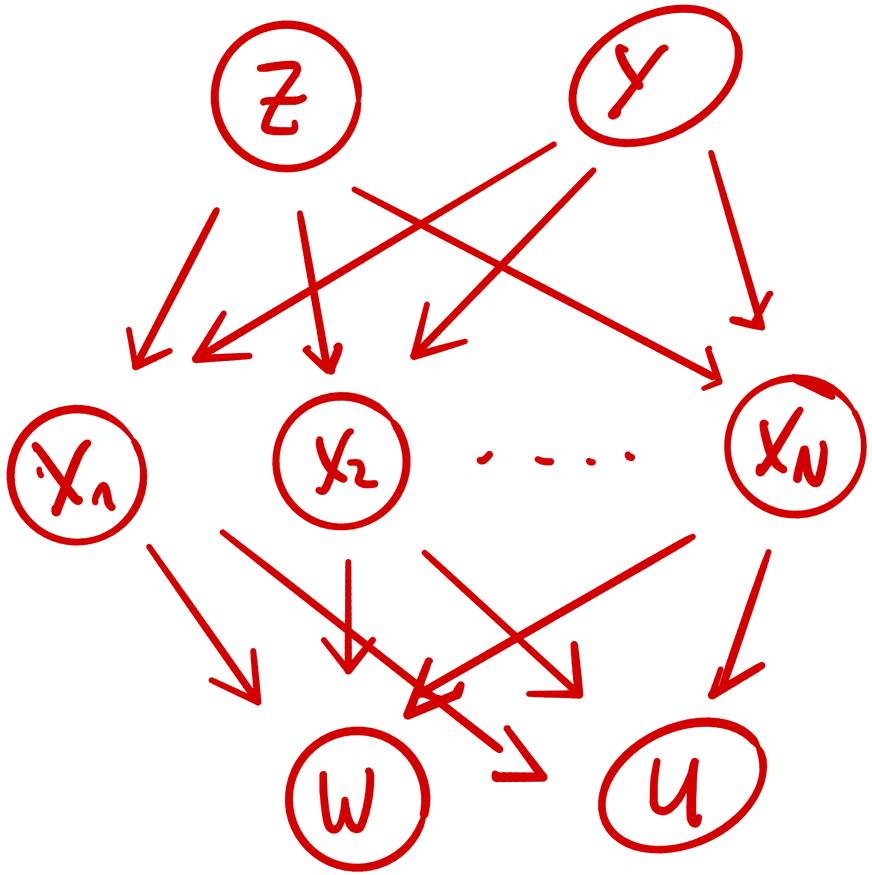
$x_A := (x_v)_{v \in A}$

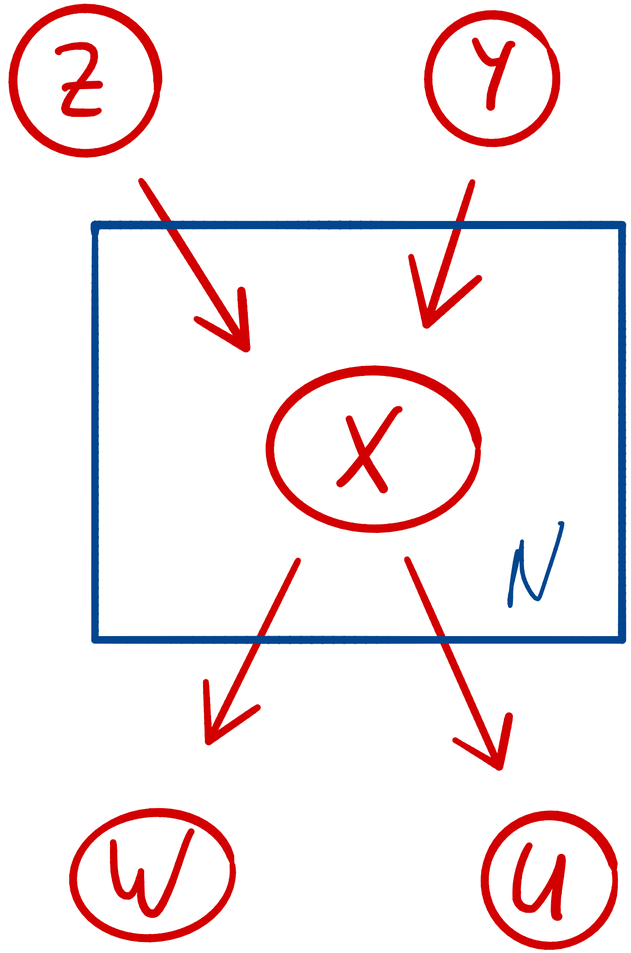
$A \subseteq V$

Bayesian Network (BN) - Definition

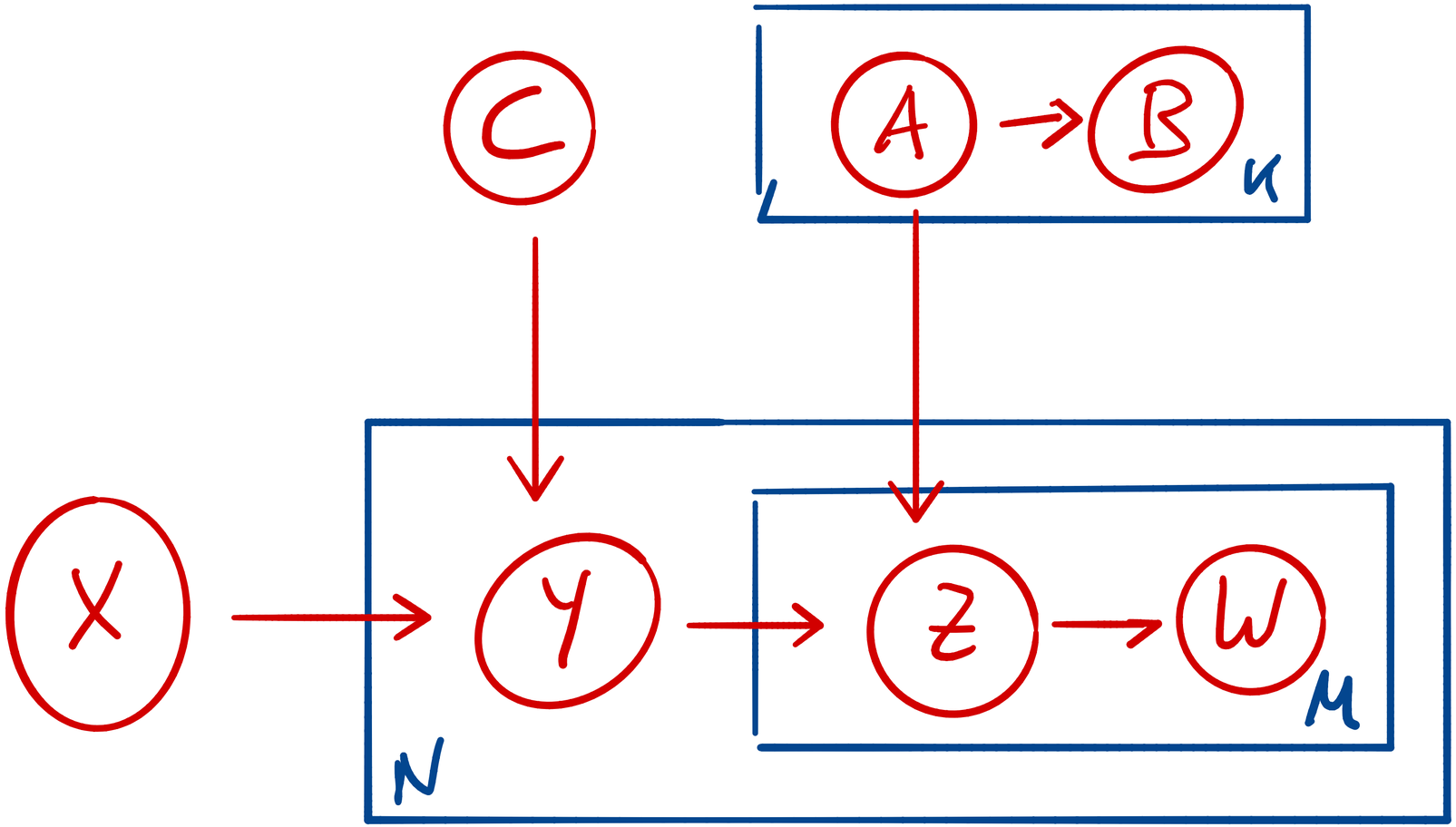
- A Bayesian Network (BN) (G, p) by definition consists of:
  - an (index) set V of random variables $X_1, \ldots, X_M$ with a joint distribution $p(x_1, \ldots, x_M)$,
  - a DAG structure $G = (V, E)$, such that:
  - $p(x_1, \ldots, x_M)$ factorizes over G.

Bayesian Networks - Plate Notation

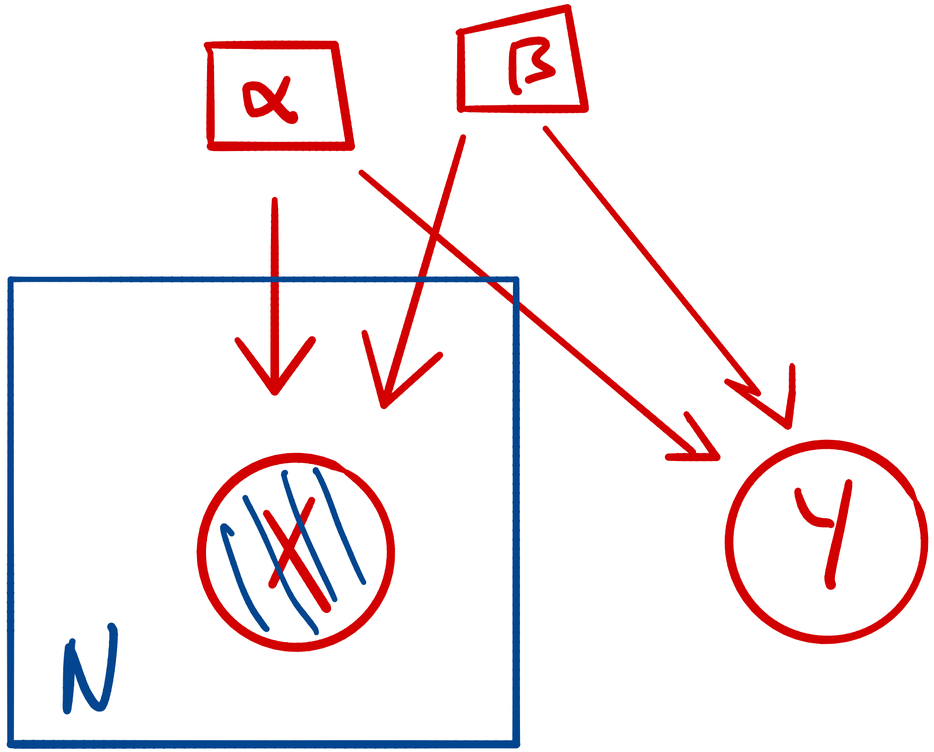
- 



Bayesian Networks - Nested Plate Notation

- 

Bayesian Networks - Parameters & Observations

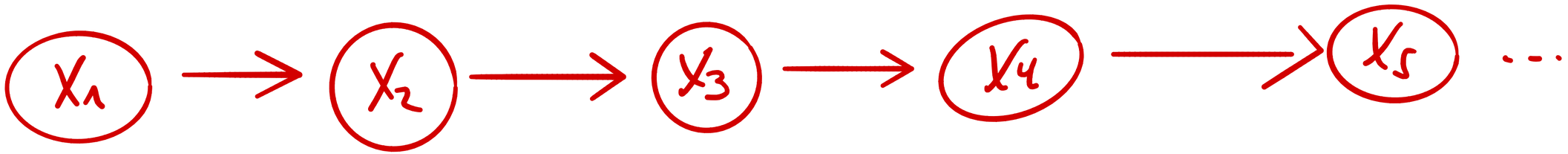
- 

Machine Learning 2

Graphical Models
- Bayesian Networks
- Examples

Patrick Forré

Example: Markov Chain

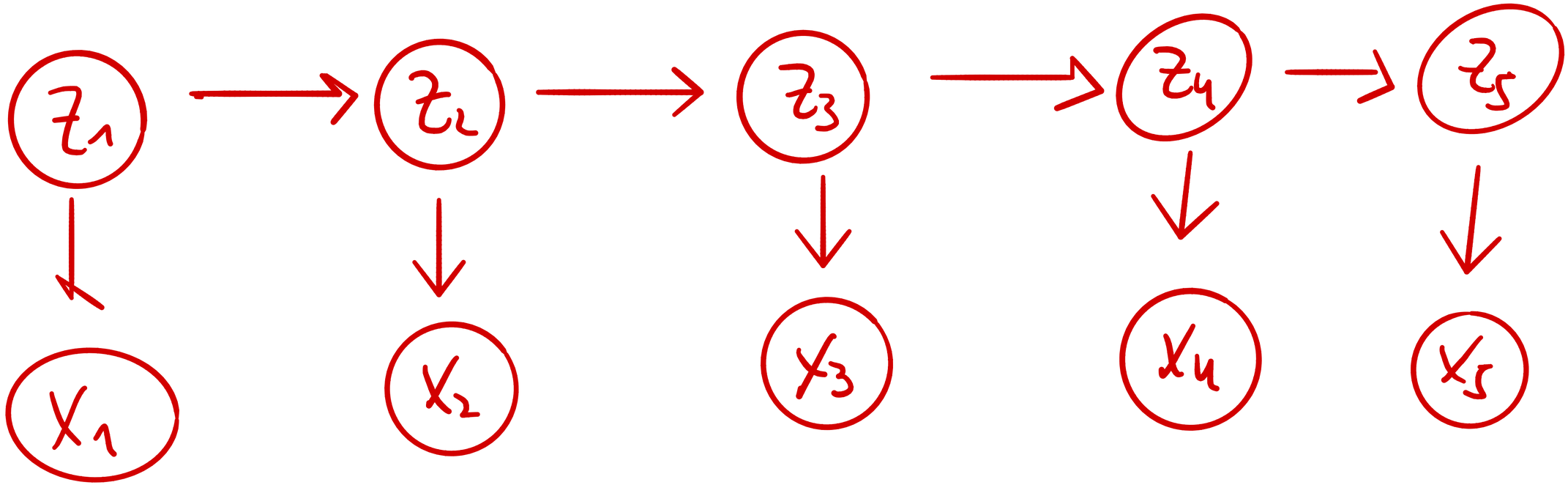
- 

$$p(x_V) = \prod_{v \in V} p(x_v \mid x_{Pa(v)})$$

capital V

$$= p(x_1) \cdot p(x_2 \mid x_1) \cdot p(x_3 \mid x_2) \cdot p(x_4 \mid x_3) \cdot p(x_5 \mid x_4) \cdots$$

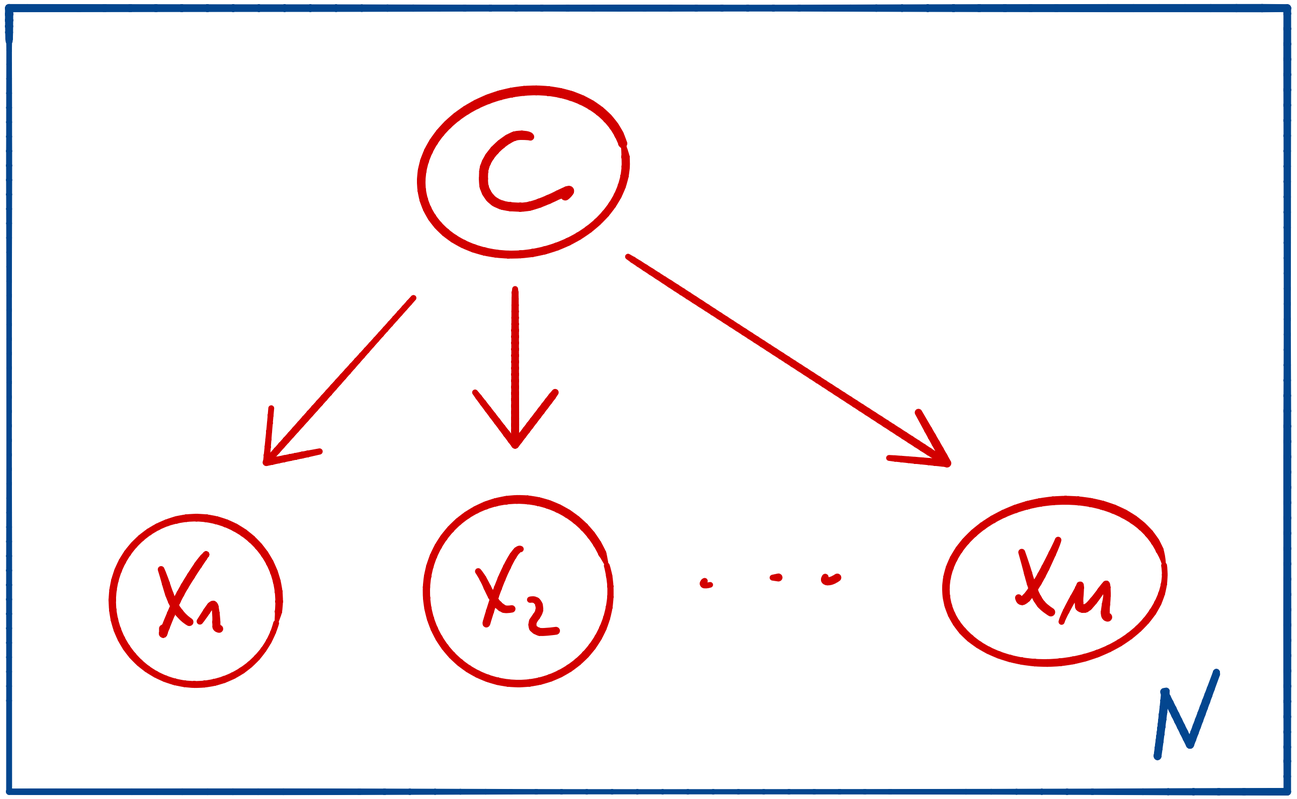
Example: State-Space Model

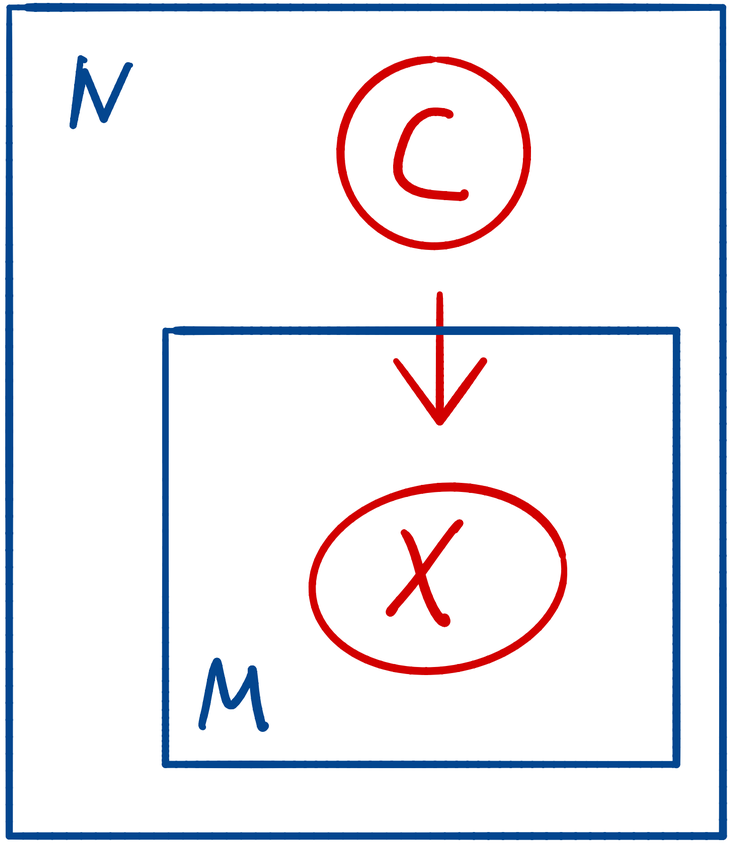
- 

$$p(x_V) = p(z_1)\, p(x_1 \mid z_1) \cdot p(z_2 \mid z_1)\, p(x_2 \mid z_2)\, p(z_3 \mid z_2) \cdot p(x_3 \mid z_3) \cdots$$

capital V

Example: Naive Bayes Model

- 

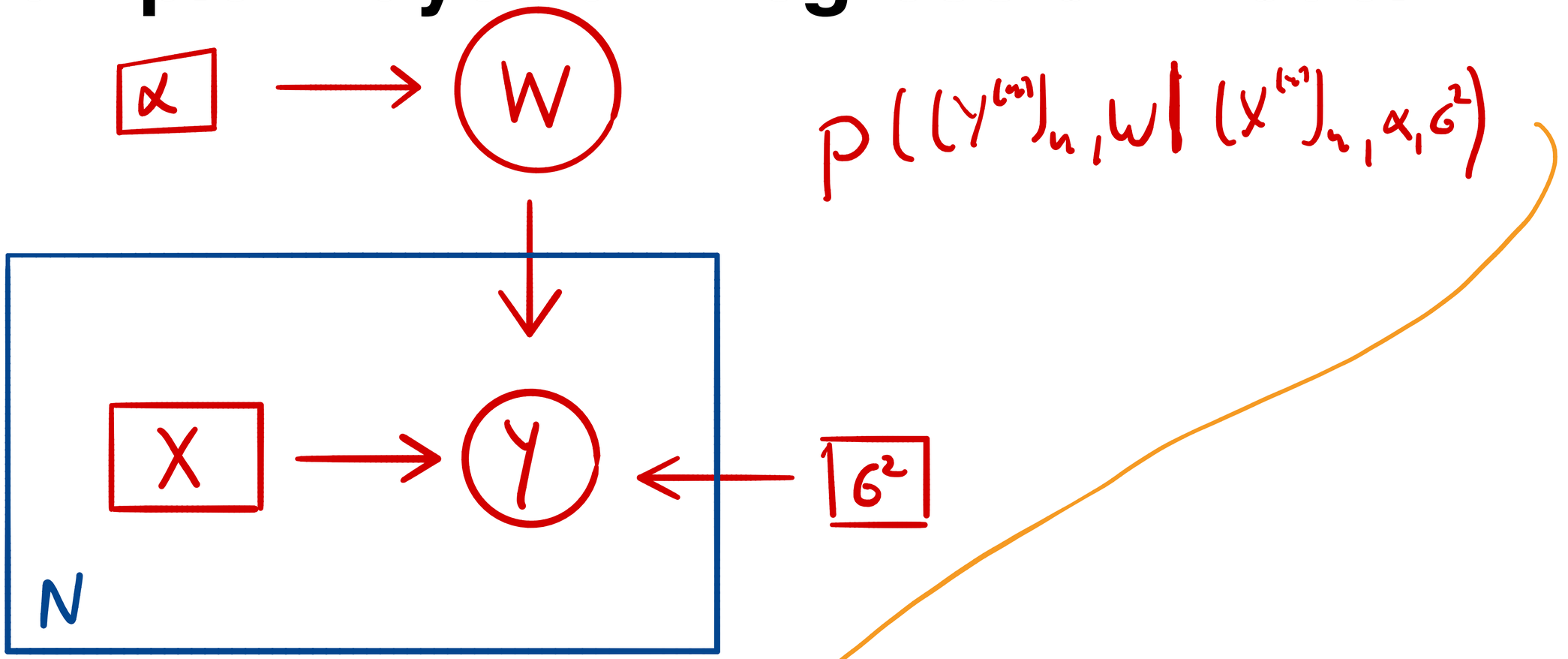


$$p(X_V) = \prod_{n=1}^{N} \left[p(C^{(n)}) \cdot p(x_1^{(n)} | C^{(n)}) \cdots p(x_M^{(n)} | C^{(n)}) \right]$$

C

capital V

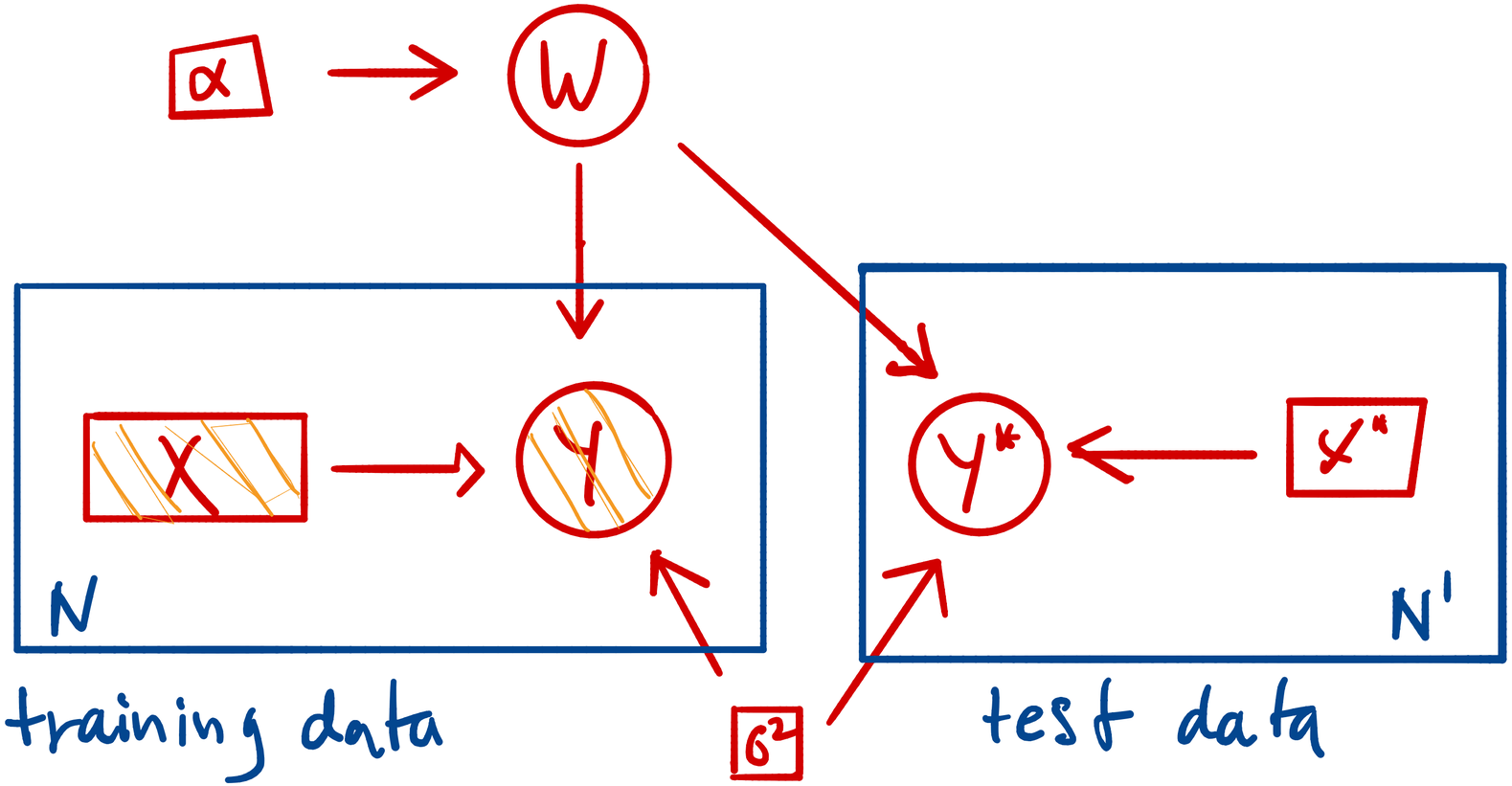
Example: Bayesian Regression Model

- 

$$p((y^{(n)})_n, w \mid (x^{(n)})_n, \alpha, \sigma^2)$$

$$= p(w \mid \alpha) \cdot \prod_{n=1}^{N} p(y^{(n)} \mid x^{(n)}, w, \sigma^2)$$

Example: Prediction in Baysian Regression Model

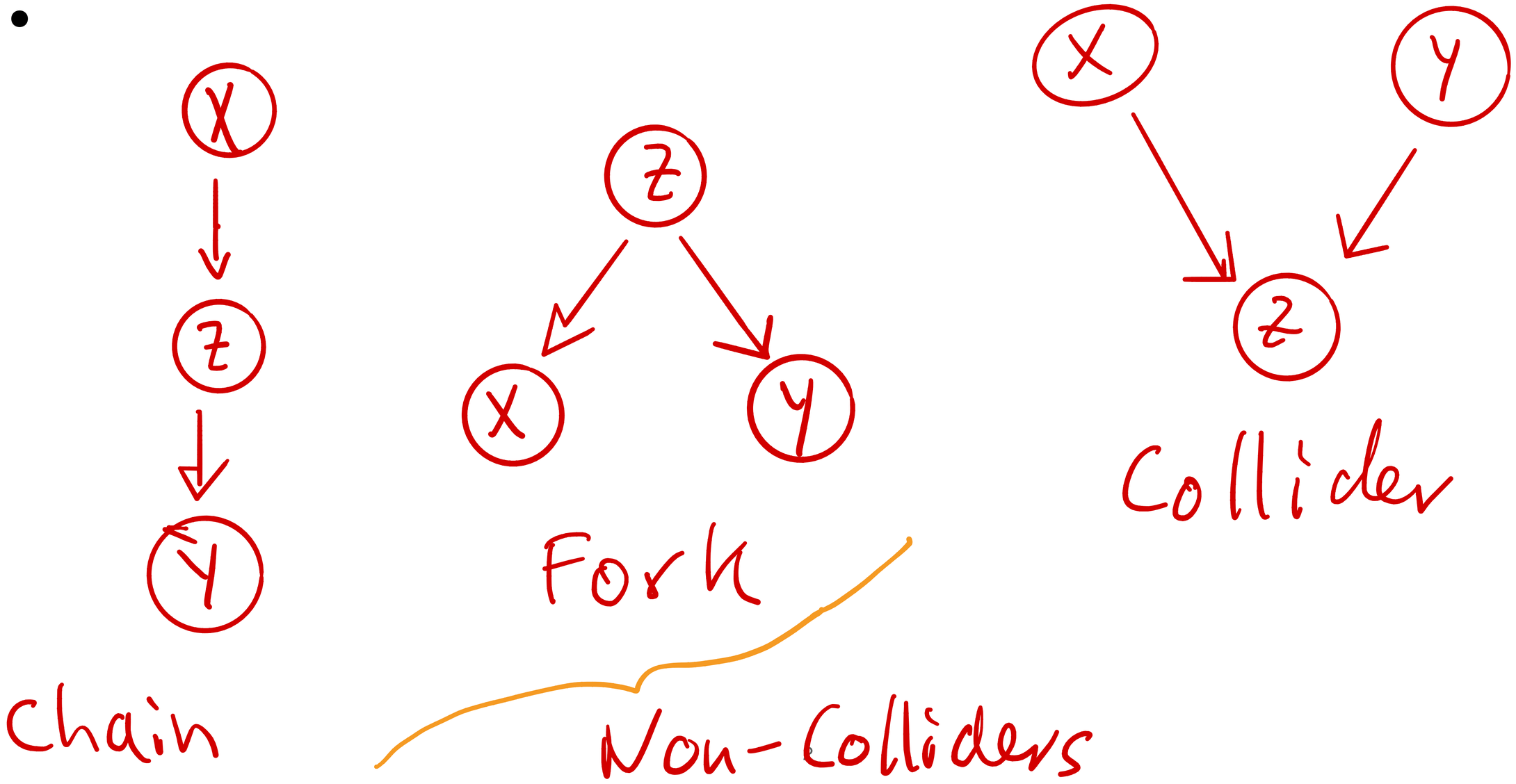
- 

Machine Learning 2

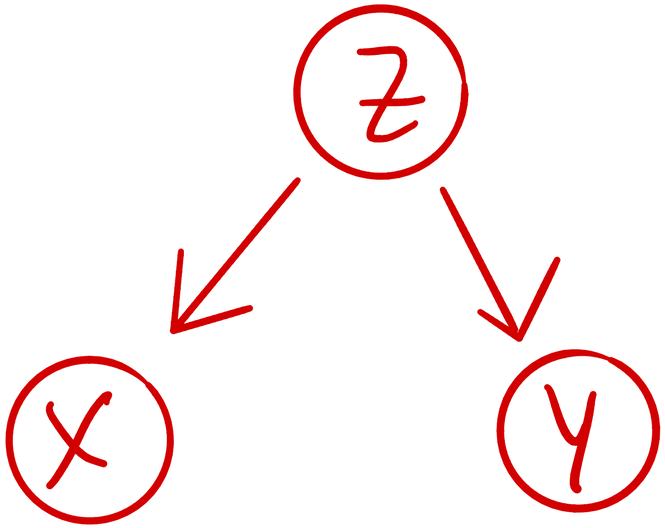
Graphical Models
- Bayesian Networks
- the 3 building blocks

Patrick Forré

The Good, the Bad and the Ugly

- 

The Fork

- 

$(*)\ p(x,y,z) = p(x|z)\ p(y|z) \cdot p(z)$

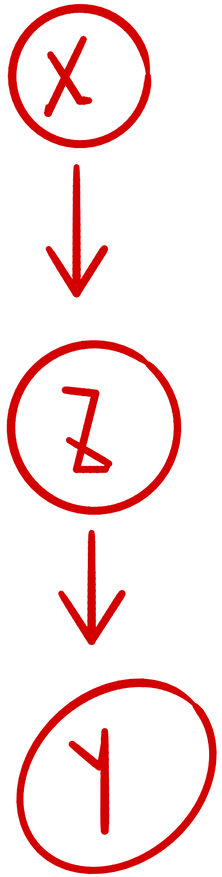
generally: $X \not\perp\!\!\!\perp Y$

always: $X \perp\!\!\!\perp Y \mid Z$

$(*) \Rightarrow p(x,y|z) = p(x|z) \cdot p(y|z)$

$\Rightarrow X \perp\!\!\!\perp Y \mid Z$

The Chain

- 

$$p(x,y,z) = p(y|z) \cdot p(z|x) \cdot p(x) \qquad (\alpha)$$

generally: $X \not\perp\!\!\!\perp Y$

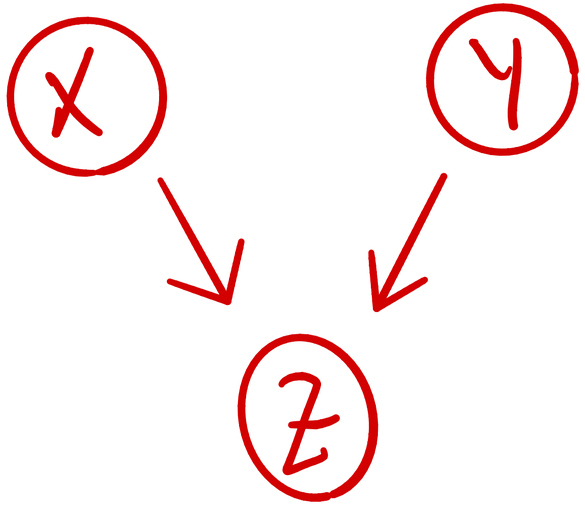
always: $X \perp\!\!\!\perp Y \mid Z$

$$(*) \Rightarrow p(x,y,z) = p(y|z) \cdot p(z,x)$$
$$= p(y|z) \cdot p(x|z) \cdot p(z)$$

$$\Rightarrow p(x,y|z) = p(x|z) \cdot p(y|z)$$

$$\Rightarrow X \perp\!\!\!\perp Y \mid Z$$

The Collider

- 

$$p(x,y,z) = p(z|x,y) \cdot p(x) \cdot p(y)$$

$$\overset{1-dz}{\Rightarrow} p(x,y) = p(x) \cdot p(y)$$

$\Rightarrow$ always: $X \perp\!\!\!\perp Y$

generally: $X \not\perp\!\!\!\perp Y \mid Z$

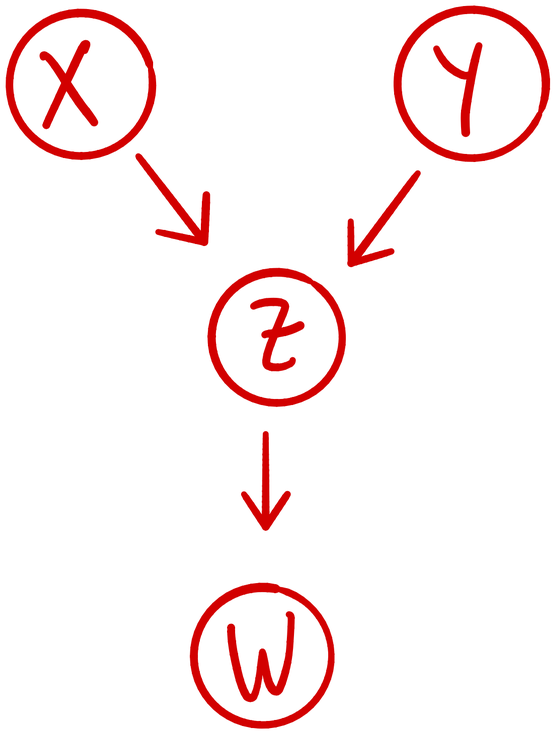
Example: Coin flips $X, Y \in \{0,1\}$

$Z = X + Y$

$Z=0 \rightsquigarrow X=0 \wedge Y=0$

$Z=1 : \quad X=0 \rightsquigarrow Y=1$

The Collider

- 

$p(x,y,z,w) = p(w|z) \cdot p(z|x,y) \cdot p(x) \cdot p(y)$

generally: $X \not\perp\!\!\!\perp Y \mid W$

always: $X \perp\!\!\!\perp Y$

Example: as before

$W := Z$

Summary

	always	generally
Fork: 	$X \perp Y \mid Z$	$X \not\perp Y$
Chain: 	$X \perp Y \mid Z$	$X \not\perp Y$
Collider: 	$X \perp Y$	$X \not\perp Y \mid Z$

Machine Learning 2

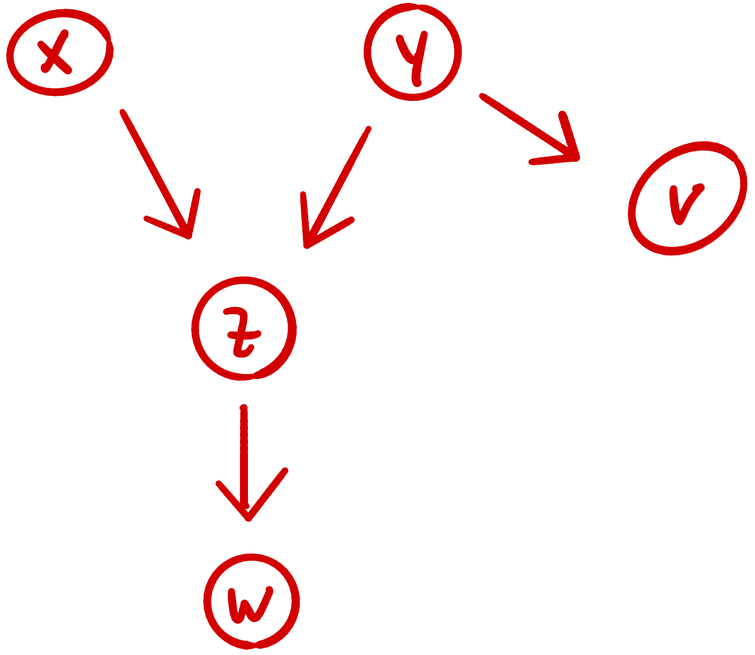
Graphical Models
- Bayesian Networks
- Global Markov Property

Patrick Forré

Blocked and open paths in DAGs - Definition

- Let $G = (V, E)$ be a DAG and $C \subseteq V$ a subset, $v, w \in V$ nodes.
- A path $v = v_0 \to \ldots \leftarrow \ldots \to \ldots v_n = w$ in G (any direction of arrow heads allowed) is called blocked by C if at least one of the following cases holds:
  - at least one of the end-nodes v or $w \in C$,
  - there is a non-collider ($\leftarrow v_i \to$ / $\to v_i \to$ / $\leftarrow v_i \leftarrow$) on the path with $v_i \in C$,
  - there is a collider ($\to v_i \leftarrow$) on the path with $v_i \notin \mathrm{Anc}^G(C)$.
- Otherwise the path is called C-open or not blocked by C.
- If $C = \emptyset$ then a path is blocked iff it contains at least one collider.

Examples: blocked and open paths in DAGs

- 

  - $X \to Z \leftarrow Y$: ϕ-blocked
    $\{Z\}$-open, $\{W\}$-open
  - $X \to Z \leftarrow Y \to V$: ϕ-blocked
    $\{Z\}$-open, $\{W\}$-open
    $\{Z, W\}$-open
    $\{Z, W, Y\}$-blocked
    $\{Y\}$-blocked

d-Separation in DAGs - Definition

- Let $G = (V, E)$ be a DAG and $A, B, C \subseteq V$ subset of nodes.
- We say that A and B are d-separated by C in G if every path from a node $v \in A$ to a node $w \in B$ in G is blocked by C.
- In symbols: $A \perp B \mid C$ or $A \overset{d}{\underset{G}{\perp}} B \mid C$

$A \perp B \iff A \perp B \mid \emptyset$

Exercise: Separoid Axioms for d-Separation

- Let $G = (V, E)$ be a DAG and $A, B, C, F \subseteq V$ be subset of nodes.
- Redundancy: $A \perp B \,|\, A$ always holds.
- Symmetry: $A \perp B \,|\, C \iff B \perp A \,|\, C$
- Decomposition: $A \perp B \cup F \,|\, C \implies A \perp B \,|\, C$
- Weak Union: $A \perp B \cup F \,|\, C \implies A \perp F \,|\, B \cup C$
- Contraction: $A \perp F \,|\, B \cup C \;\wedge\; A \perp B \,|\, C \implies A \perp B \cup F \,|\, C$
- So we have the equivalence:

$$A \perp B \cup F \,|\, C \iff A \perp F \,|\, B \cup C \;\wedge\; A \perp B \,|\, C$$

Theorem - Global Markov Property for BN

- Let (G, p) be a Bayesian Network (BN) with DAG $G = (V, E)$ and random variables $X_1, \ldots, X_M$ with joint distribution $p(x_V)$. For a set of nodes $F \subseteq V$ we write $X_F = (X_v)_{v \in F}$.

- For any three subsets $A, B, C \subseteq V$ of nodes we have the implication:

$$A \perp B \,|\, C \implies X_A \perp\!\!\!\perp X_B \,|\, X_C.$$

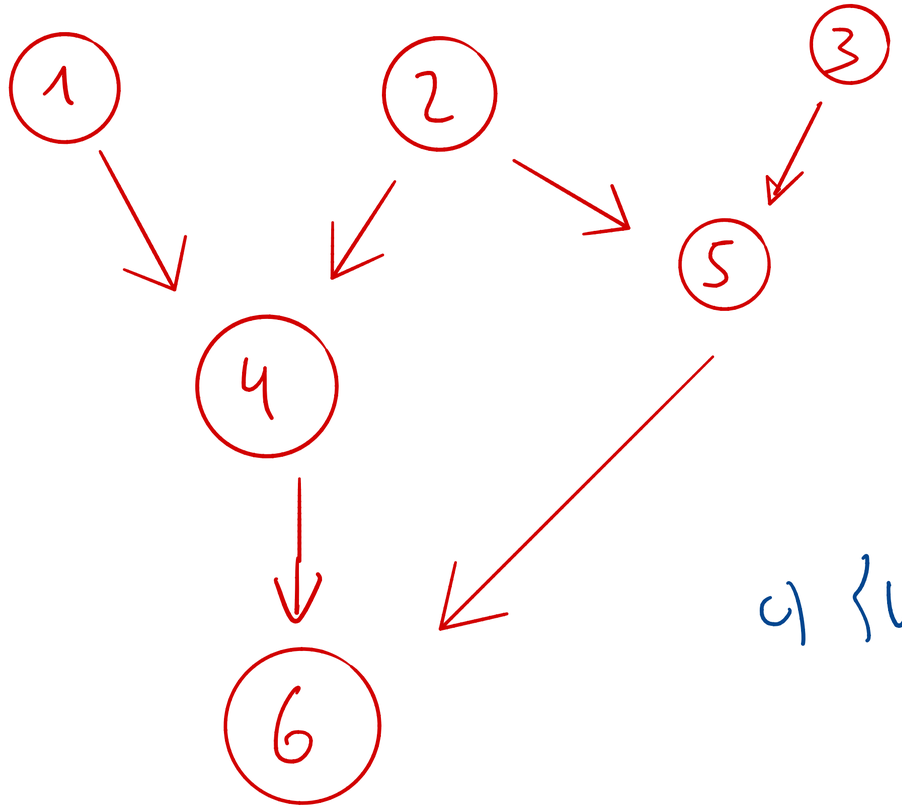
  i.e. d-separation implies corresponding conditional independence.

- The reverse implication is NOT true in general.

Exercise - Prove Global Markov Property - Sketch

- Induction by the number of nodes.
- Check that removing a child-less node $w \in V$ (exists because of acyclicity) doesn't hurt factorization property for remaining variables.
- Check (*): $X_w \perp\!\!\!\perp X_{-w} \,|\, X_{\mathrm{Pa}^G(w)}$.
- Check that w.l.o.g. A, B, C can be assumed to be pairwise disjoint.
- Check 4 cases: 0) $w \notin A \cup B \cup C$, 1) $w \in A$, 2) $w \in B$, 3) $w \in C$.
- For those use induction step, relation (*) and the separoid axioms.
- For case 3) show that $A \perp B \,|\, C' \cup \{w\}$ implies $A \cup \{w\} \perp B \,|\, C'$ or $A \perp B \cup \{w\} \,|\, C'$ and then use case 1) and 2).

Examples: d-Separation and Conditional Independence

- 

a) $\{V_4\} \perp \{V_5\} \mid \{V_2\} \Rightarrow X_4 \perp\!\!\!\perp X_5 \mid X_2$

b) $\{V_1, V_2, V_3\} \perp \{V_6\} \mid \{V_4, V_5\}$

$\Rightarrow (X_1, X_2, X_3) \perp\!\!\!\perp X_6 \mid (X_4, X_5)$

c) $\{V_3\} \perp \{V_6\} \mid \{V_5, V_4\} \Rightarrow X_3 \perp\!\!\!\perp X_6 \mid X_5, X_4$

Machine Learning 2

Graphical Models
- Bayesian Networks
- Construction of BNs

Patrick Forré

Bayesian Networks - What are they good for?

- Let $p(x_1, \ldots, x_M)$ be the joint distribution of discrete random variables $X_1, \ldots, X_M$, where each X_m can take on K values.
- A naive table representation of $p(x_V)$ needs $\sim K^M$ entries.

- If (G, p) is a 'sparse' BN, where every node has at most $L \ll M$ parents.
- Representing $p(x_v \mid x_{\mathrm{Pa}_G(v)})$ in table form each takes $\sim K^{L+1}$ entries.
- So $p(x_V) = \prod_{v \in V} p(x_v \mid x_{\mathrm{Pa}_G(v)})$ takes $\sim M \cdot K^{L+1} \ll K^M$ entries.
- $\Longrightarrow$ BNs model independence relations and allow for sparse representations.

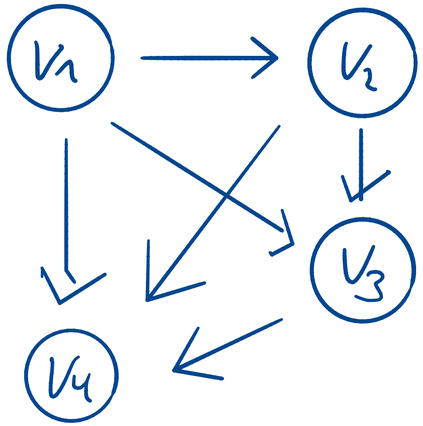
Construction of bad Bayesian Networks

- Let $p(x_1, \ldots, x_M)$ be the joint distribution and $V = \{1, \ldots, M\}$.
- We want to construct a DAG $G = (V, E)$ such that (G, p) is a BN.
- Product rule:

$$p(x_1, \ldots, x_M) = \prod_{m=1}^{M} p(x_m \mid x_1, \ldots, x_{m-1})$$

$\mathrm{Pa}_G(v_m)$

$p(x_m \mid x_1, \ldots x_{m-1}) \sim K^M$



- So fully connected DAG always turns (G, p) into BN.
- Note, different ordering of $X_1, \ldots, X_M$ results in different DAG.

Construction of better Bayesian Networks

- Inductively, take a minimal subset $P_m \subseteq \{1, \ldots, m-1\}$ such that:
  - $X_m \perp\!\!\!\perp (X_1, \ldots, X_{m-1}) \,|\, X_{P_m}$.
- Then: $p(x_m \,|\, x_1, \ldots, x_{m-1}) = p(x_m \,|\, x_{P_m})$
- Include in G edges from nodes $w \in P_m$ to v_m, then:
  - G is a DAG,
  - $\mathrm{Pa}^G(v_m) = P_m$,
  - (G, p) is a BN with minimal number of parents (given the total ordering).

$$p(x_V) = \prod_{m=1}^{M} p(x_m \,|\, x_{1:m-1}) = \prod_{m=1}^{M} p(x_m \,|\, x_{P_m})$$

- Note, different ordering of $X_1, \ldots, X_M$ results in different DAG and BN.

Machine Learning 2

Graphical Models
- Bayesian Networks
- Summary

Patrick Forré

Bayesian Networks - Summary

- BN consists of a DAG $G = (V, E)$ and distribution $p(x_V) = \prod_{v \in V} p(x_v \mid x_{\mathrm{Pa}^G(v)})$.
- Introduced Plate notation with parameters and observations.
- d-Separation: $A \perp B \mid C$ iff every path from A to B in G either contains a non-collider in C or a collider NOT in $\mathrm{Anc}^G(C)$.
- Global Markov Property of BN: $A \perp B \mid C \implies X_A \perp\!\!\!\perp X_B \mid X_C$.
- Construction of BN for $p(x_V)$: Inductively, take a minimal subset $\mathrm{Pa}_m \subseteq \{1, \ldots, m-1\}$ s.t. $X_m \perp\!\!\!\perp X_{[1:m]} \mid X_{\mathrm{Pa}_m}$ as the parents of v_m.
- BNs model independence relations and allow for sparse representations.