香港科技大學
THE HONG KONG
UNIVERSITY OF SCIENCE
AND TECHNOLOGY

COMP 5212
Machine Learning
Lecture 14

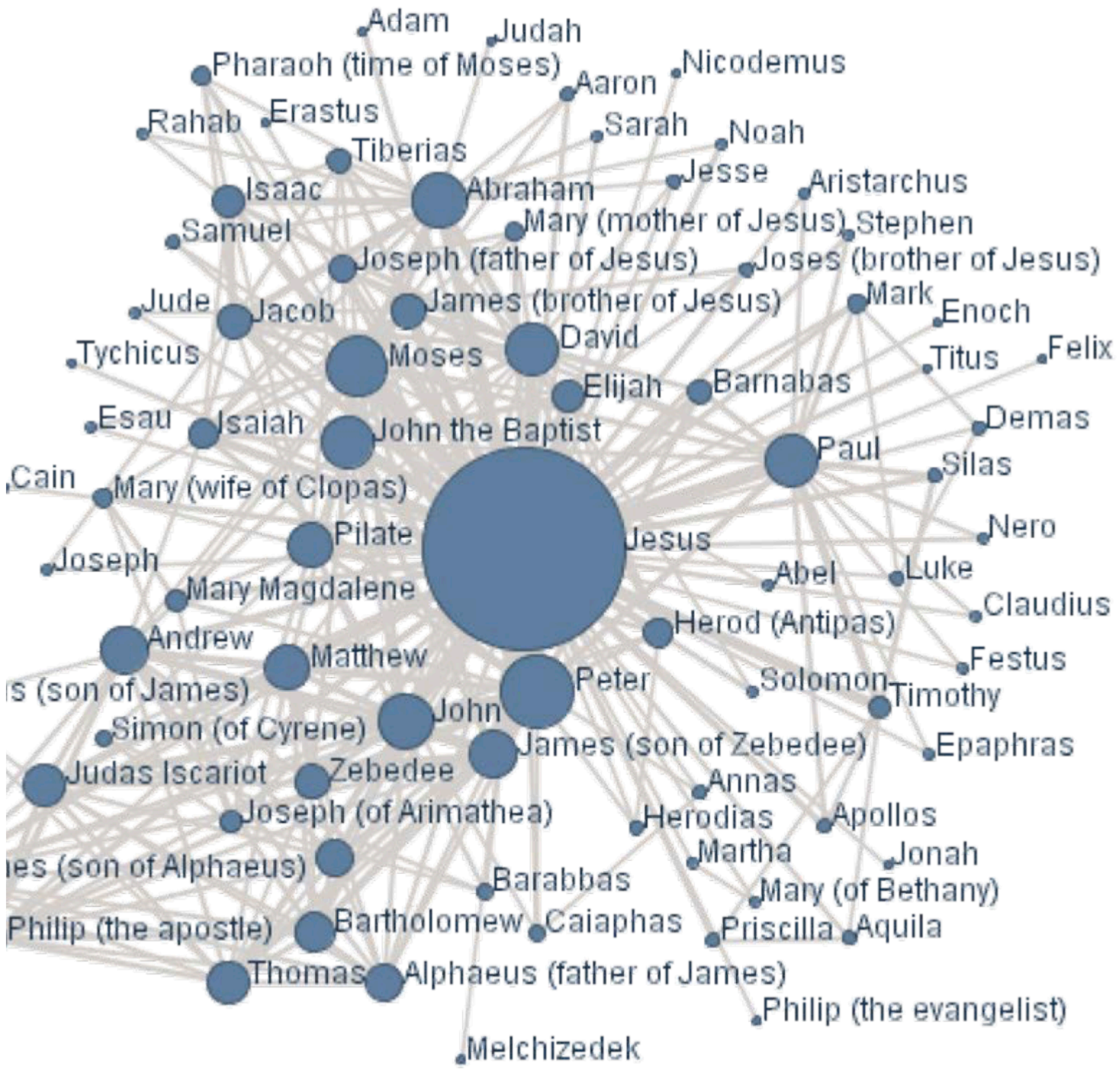
Probabilistic Graphical Models

Junxian He
Mar 22, 2024

What Are Graphical Models?

- Informally, a GM is just a graph representing relationship among random variables
  - Nodes: random variables (features, not examples)
  - Edges (or absence of edges): relationship

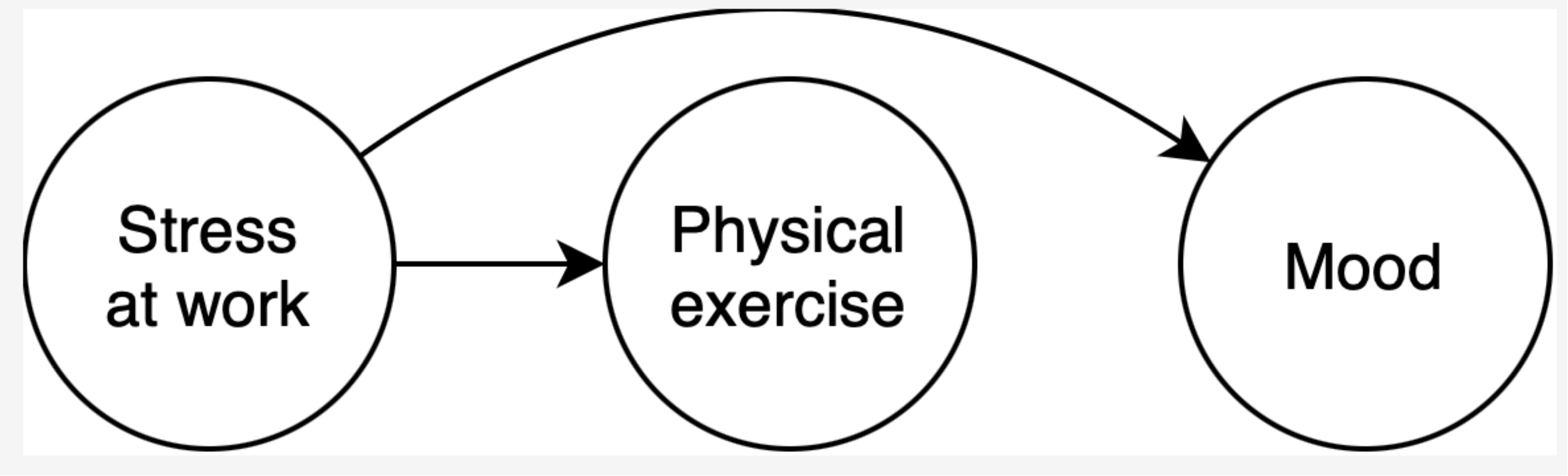
- Looks simple!
  - But detail matters, as always.
  - What exactly do we mean by relationship?



Relationship between two random variables

- Many types of relationships exist:
  - X and Y are correlated
  - X and Y are dependent
  - X and Y are independent
  - X and Y are partially correlated given Z
  - X and Y are conditionally dependent given Z
  - X and Y are conditionally independent given Z
  - X causes Y
  - Y causes X
  - ...

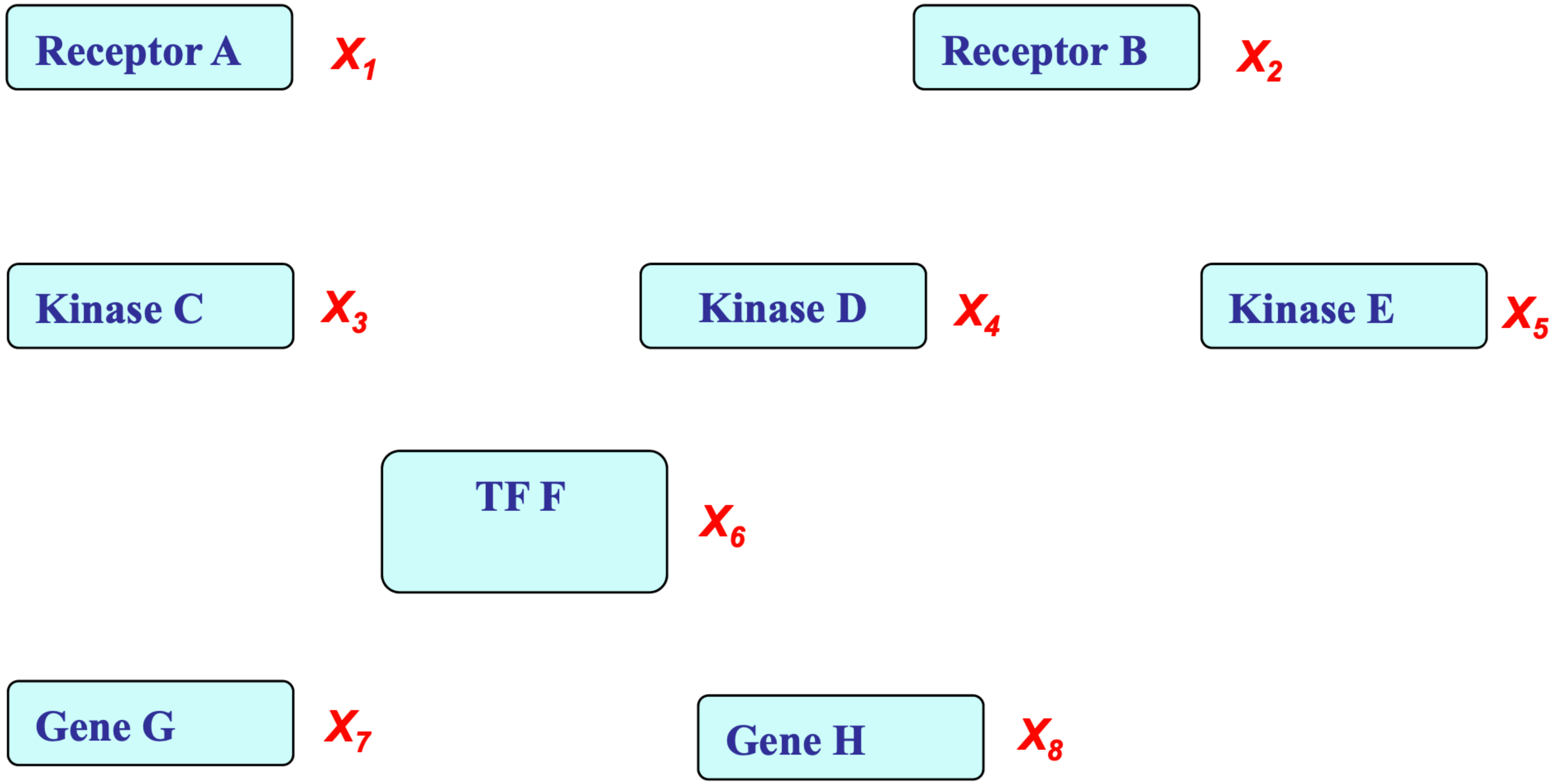
Correlation does not imply causation



What is a Graphical Model?

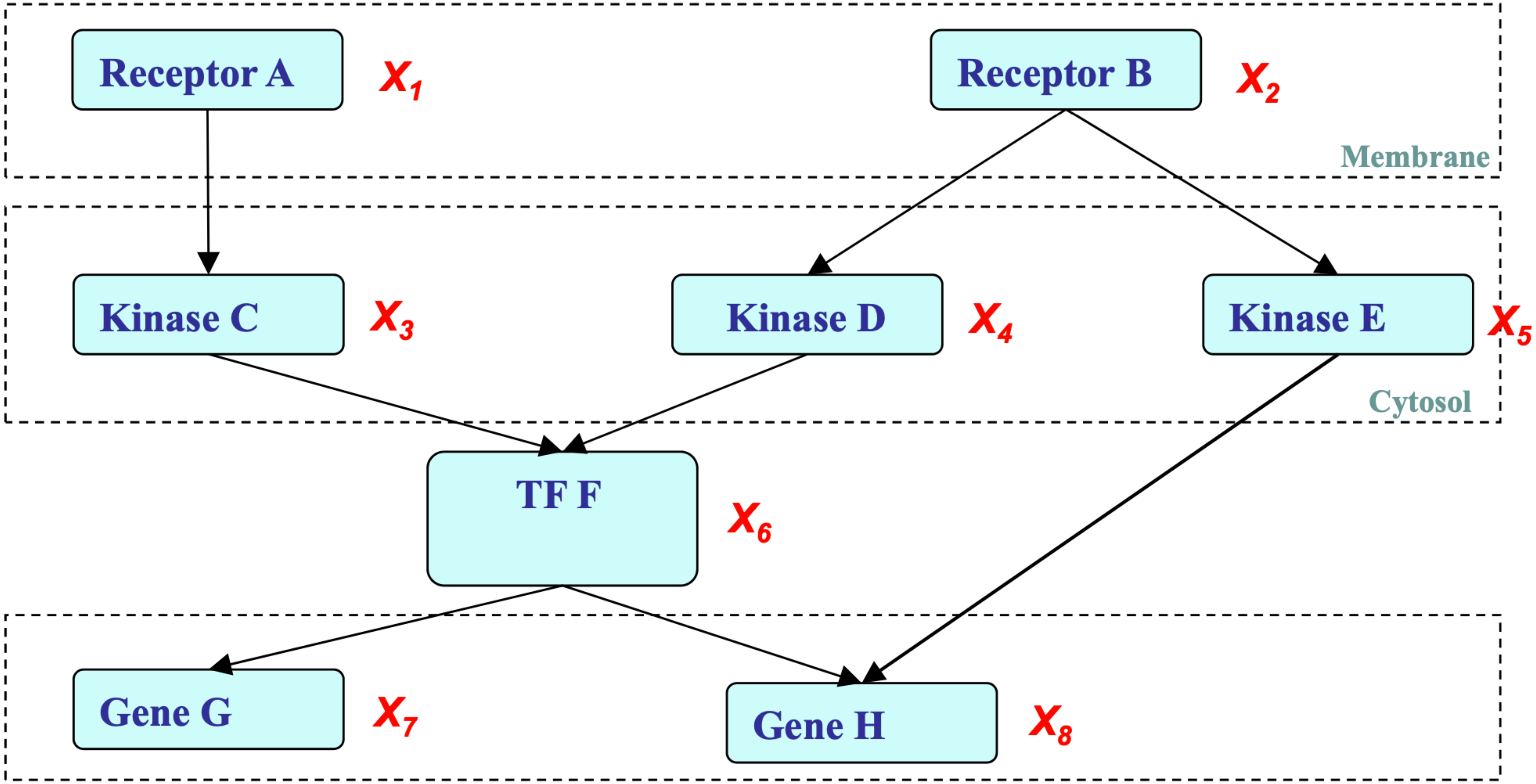
Graphical model represents a multivariate distribution in High-D space

A possible world for celluar signal transduction:



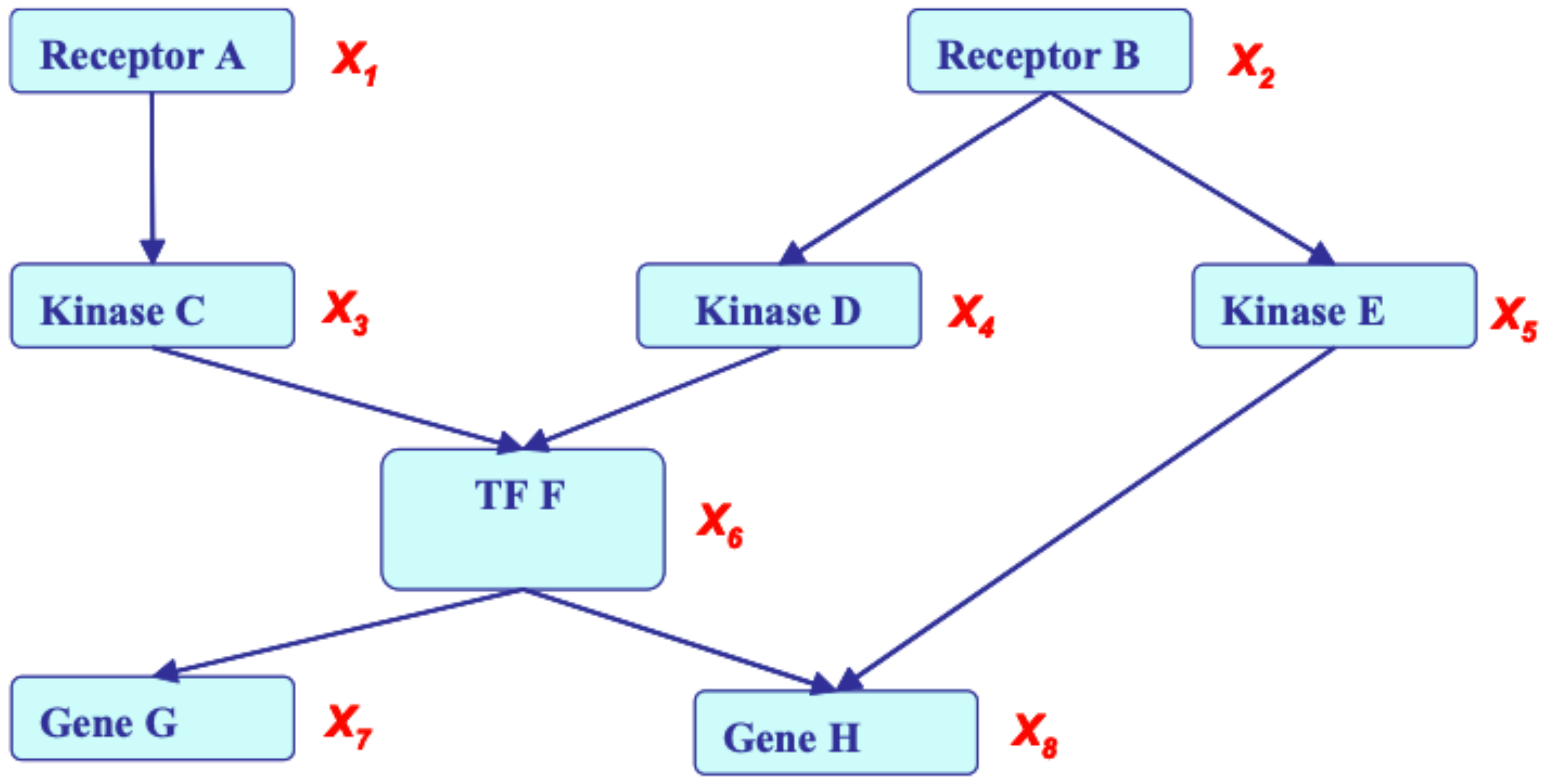
Structure Simplifies Representation

Dependencies among variables



Probabilistic Graphical Models

- If X_i's are **conditionally independent** (as described by a **PGM**), the joint can be factored to a product of simpler terms, e.g.,

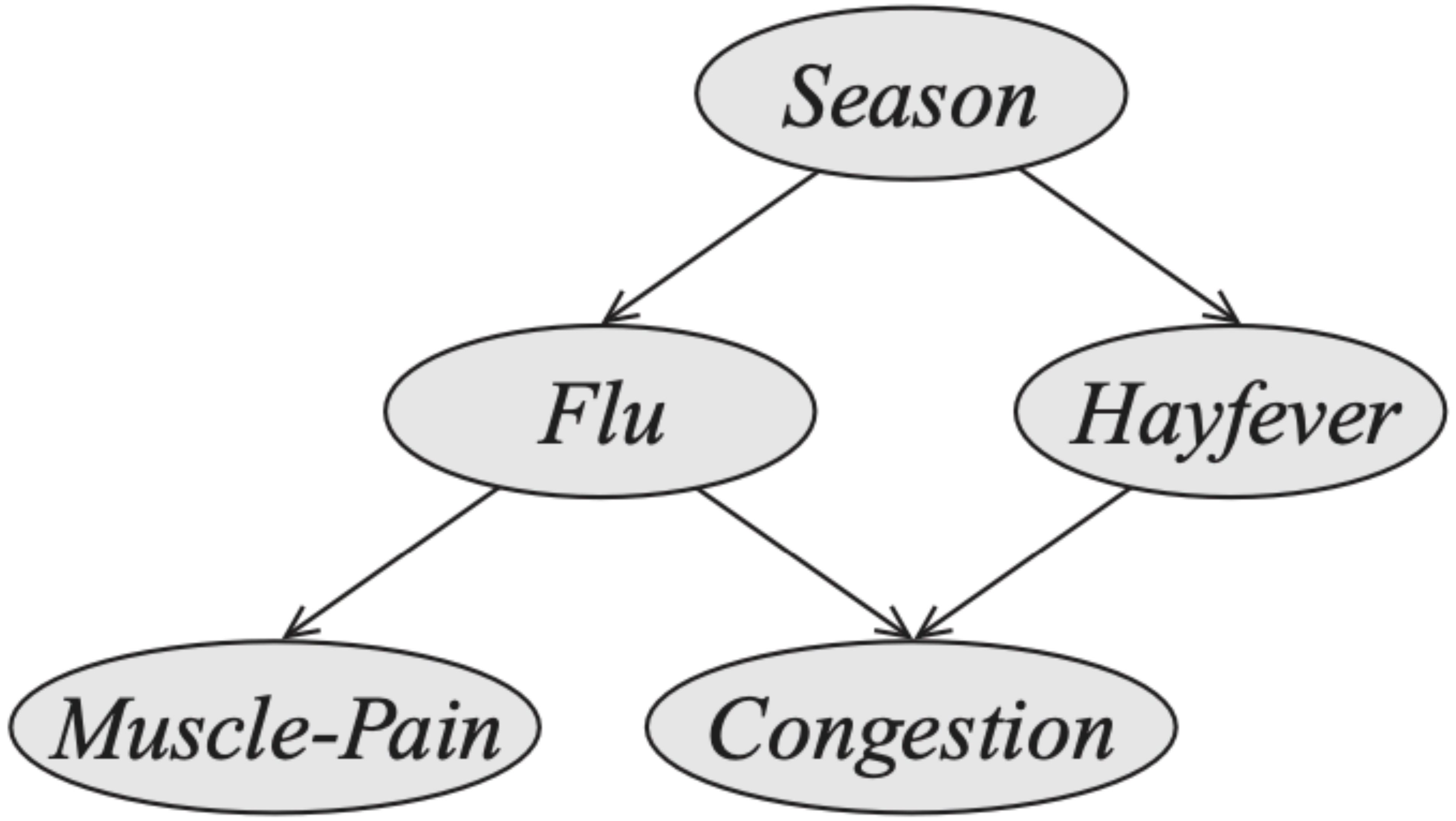


$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

Stay tune for what are these independencies!

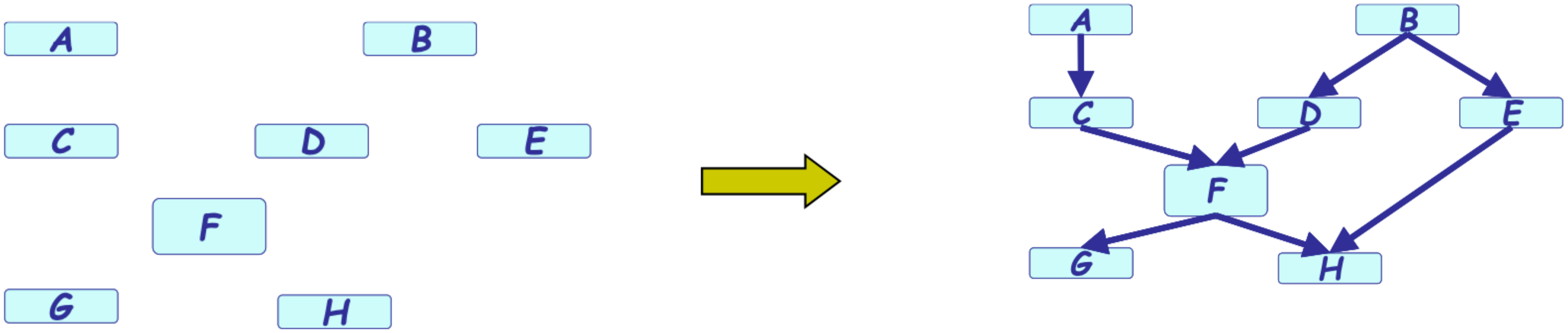
Another Example



$$P(\mathit{Congestion} \mid \mathit{Flu}, \mathit{Hayfever}, \mathit{Season}) = P(\mathit{Congestion} \mid \mathit{Flu}, \mathit{Hayfever});$$

What is a PGM After All

It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***



$$P(X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8)$$

$$P(X_{1:8}) = P(X_1)P(X_2)P(X_3 \mid X_1X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$
$$P(X_6 \mid X_3, X_4)P(X_7 \mid X_6)P(X_8 \mid X_5, X_6)$$

More formal definition: It refers to a family of distributions on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a graph that connects these variables

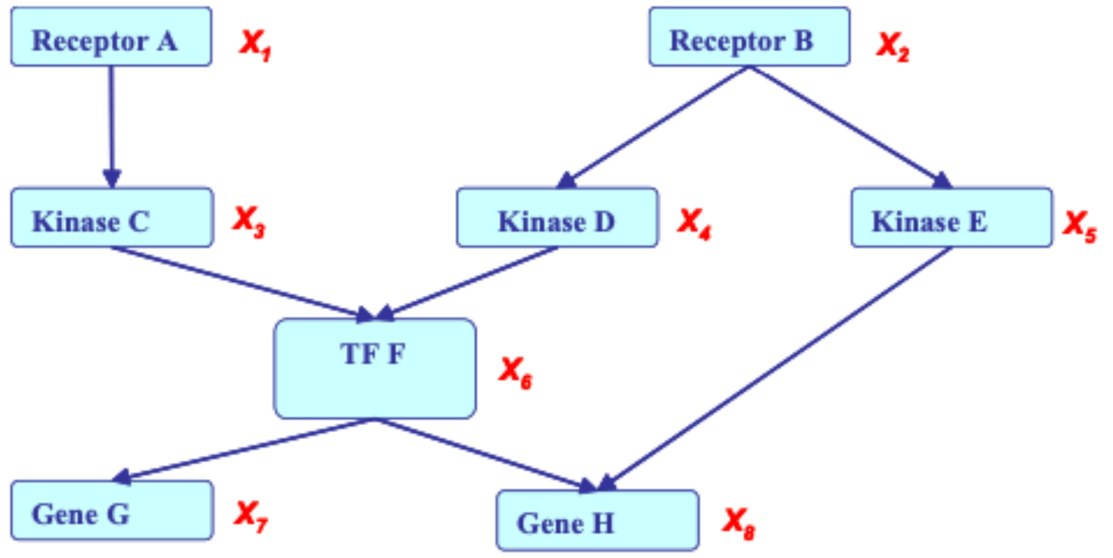
Probabilistic Graphical Model is a graphical language to express conditional indenpdence

Two types of Graphical Models

- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

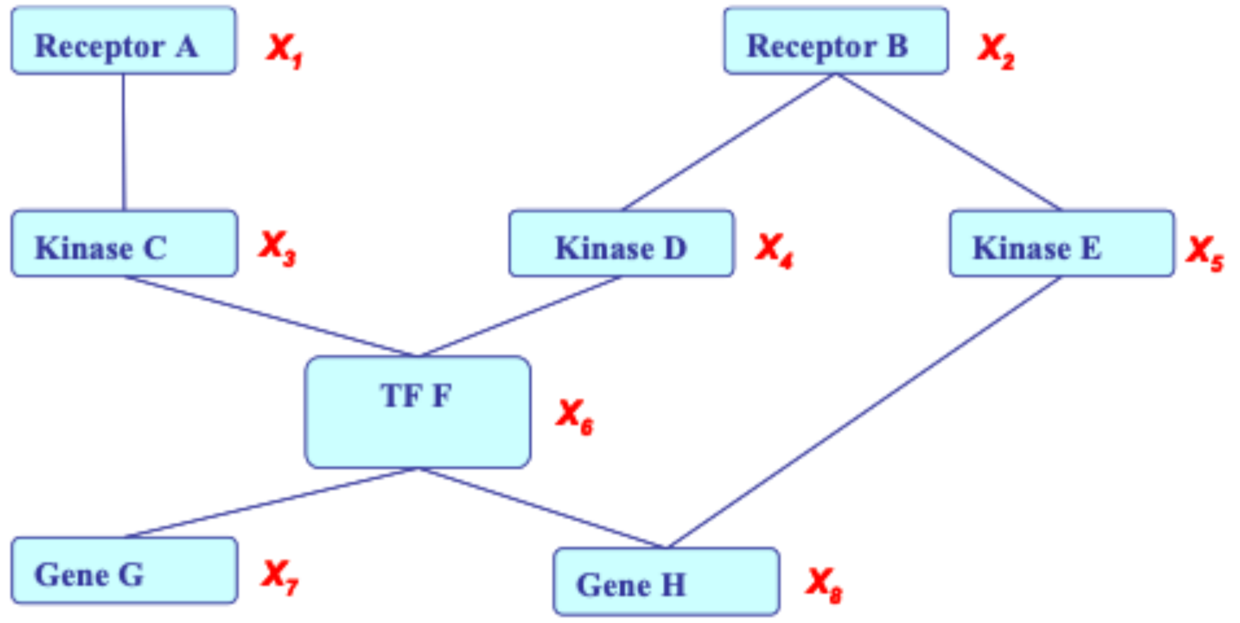
$$= P(X_1)\, P(X_2)\, P(X_3| X_1)\, P(X_4| X_2)\, P(X_5| X_2)\, P(X_6| X_3, X_4)\, P(X_7| X_6)\, P(X_8| X_5, X_6)$$



- Undirected edges simply give correlations between variables (**Markov Random Field** or **Undirected Graphical model**):

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$

$$= 1/Z \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2) + E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}$$



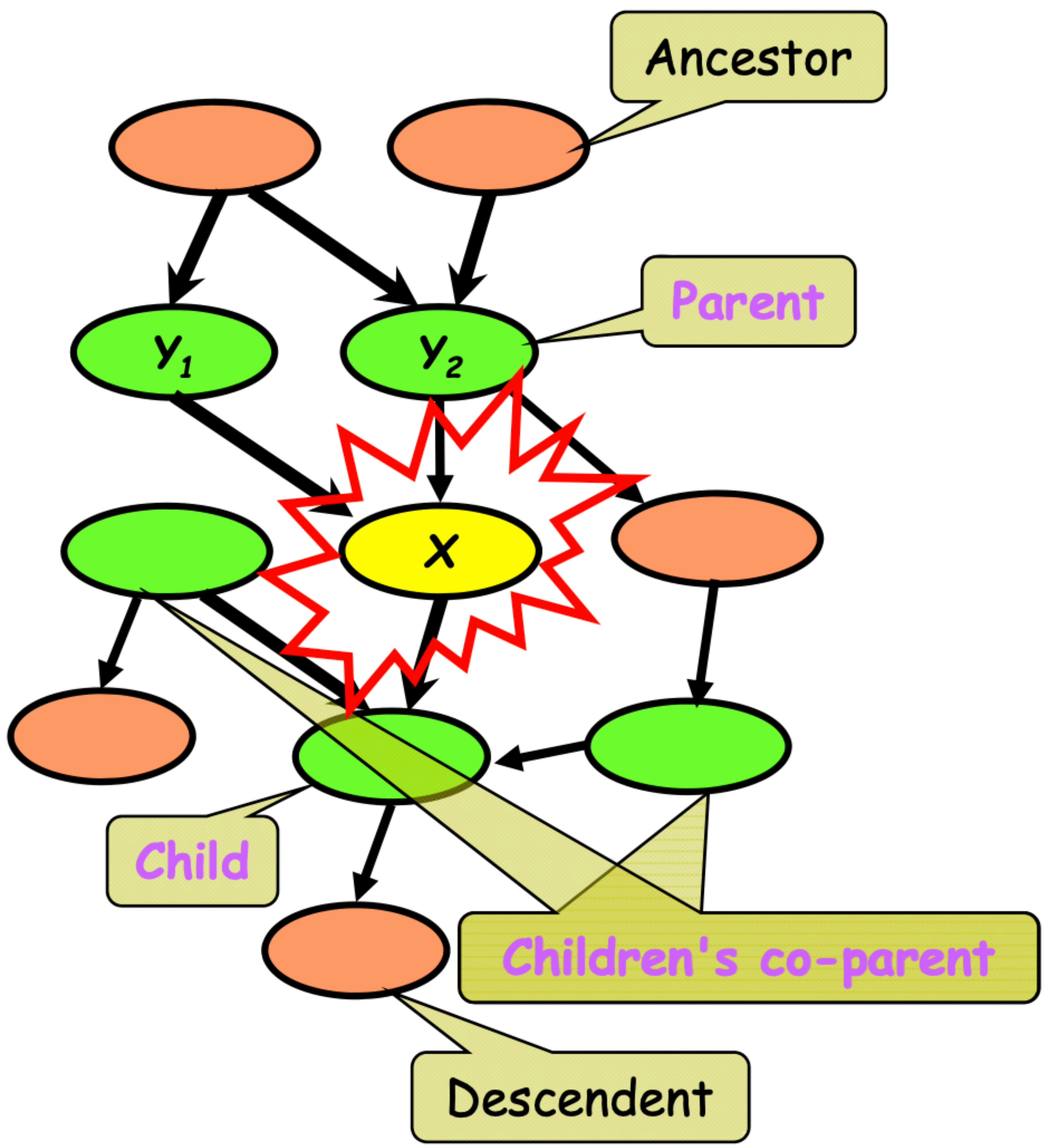
PGMs are Structural Specification of Probability Distribution

- Separation properties in the graph imply independence properties about the associated variables

- For the graph to be useful, any conditional independence properties we can derive from the graph should hold for the probability distribution that the graph represents

Markov Blanket for Directed Acyclic Graph (DAG)

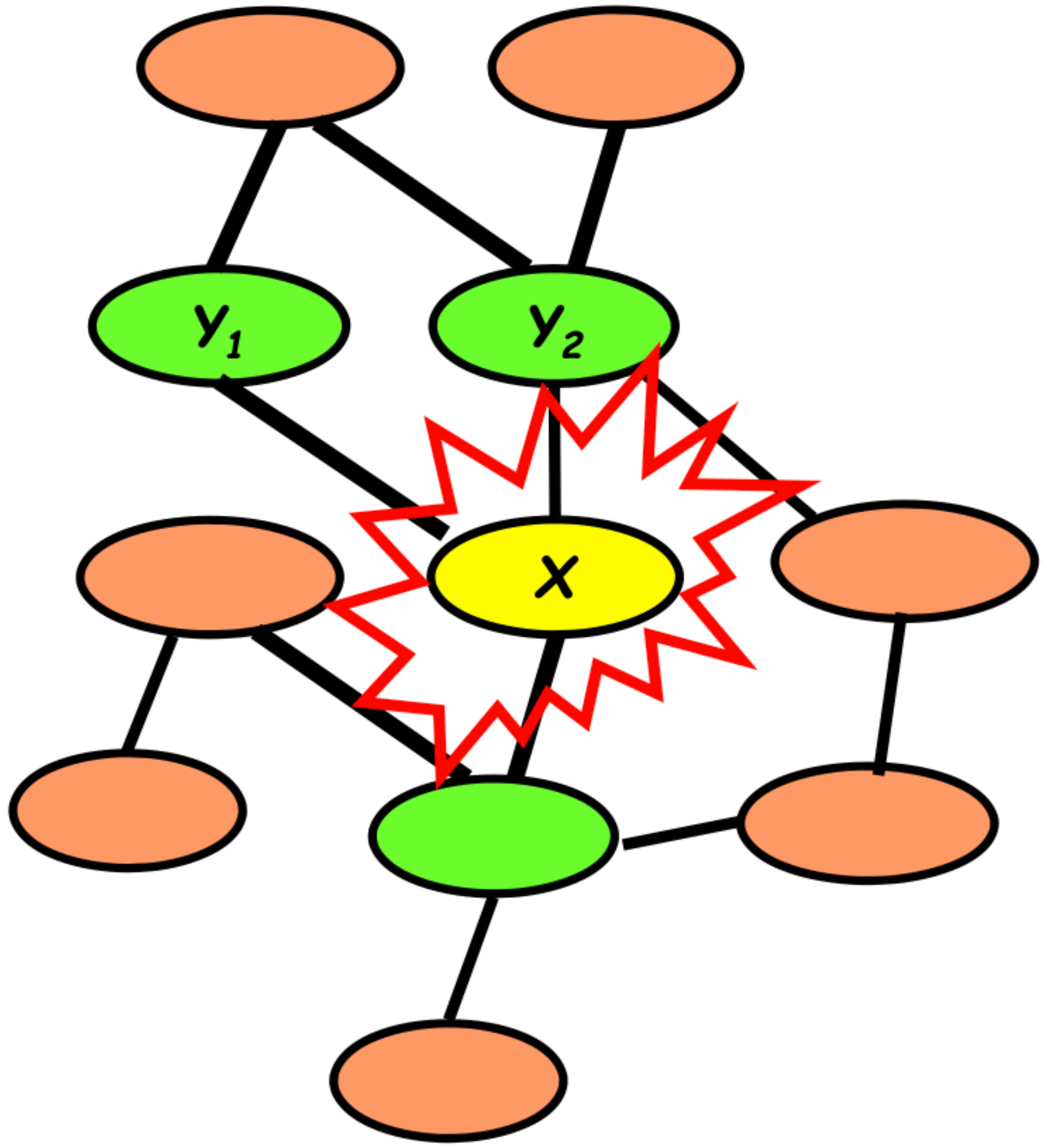
- Meaning: a node is **conditionally independent** of every other node in the network outside its **Markov blanket**



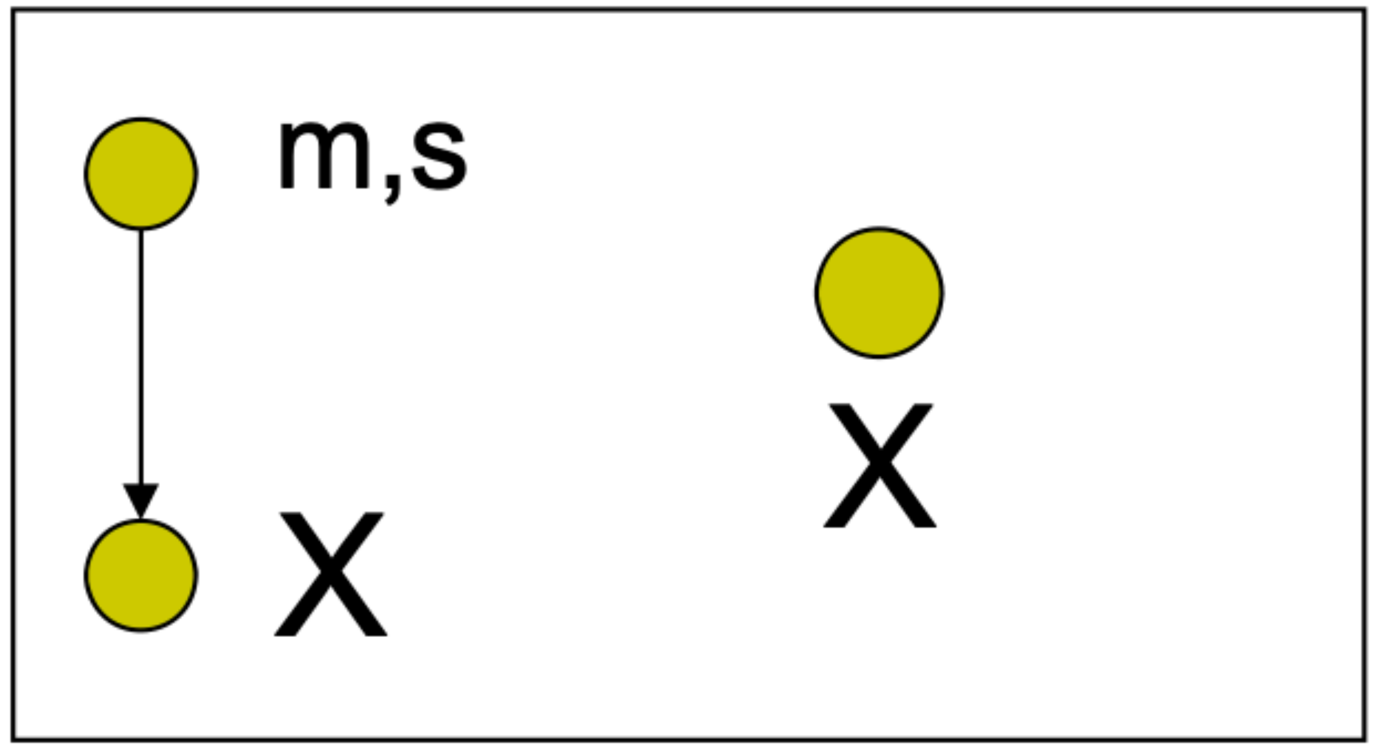
Markov blanket of a node is its parents + child + children’s co-parent

Conditional Independence of Undirected Graph

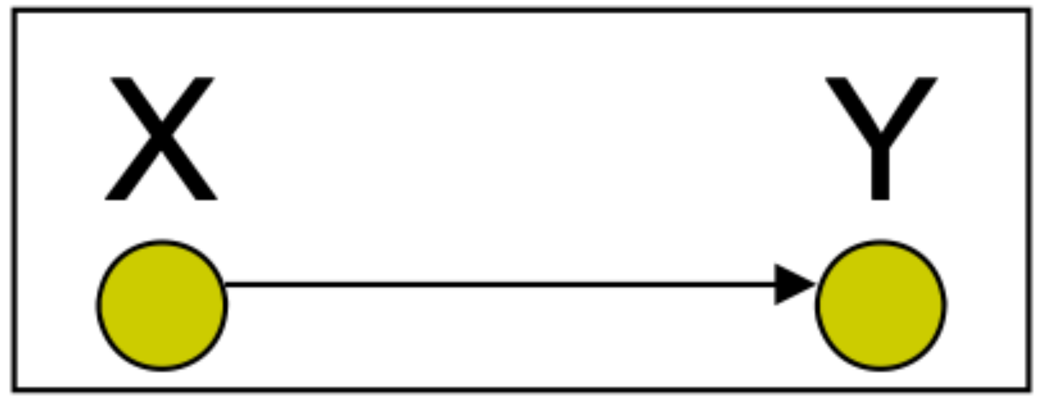
- Meaning: a node is **conditionally independent** of every other node in the network given its **Directed neighbors**



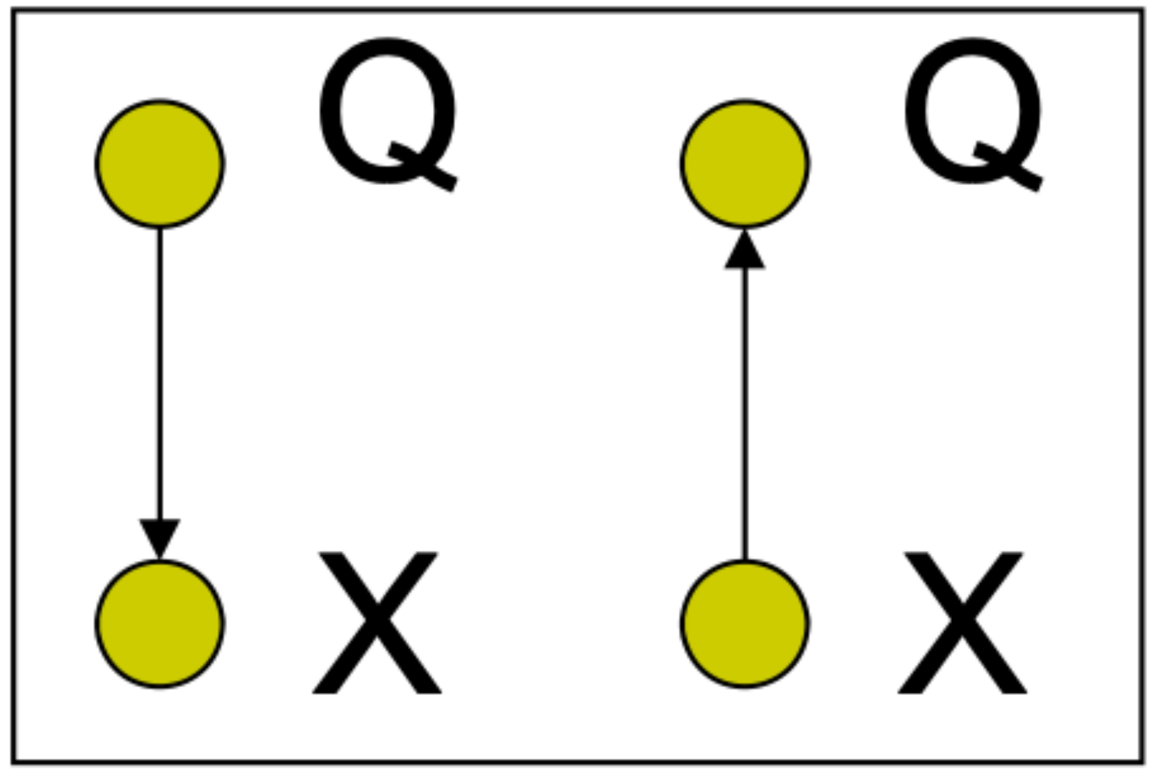
GMs are your old friends



P(x)



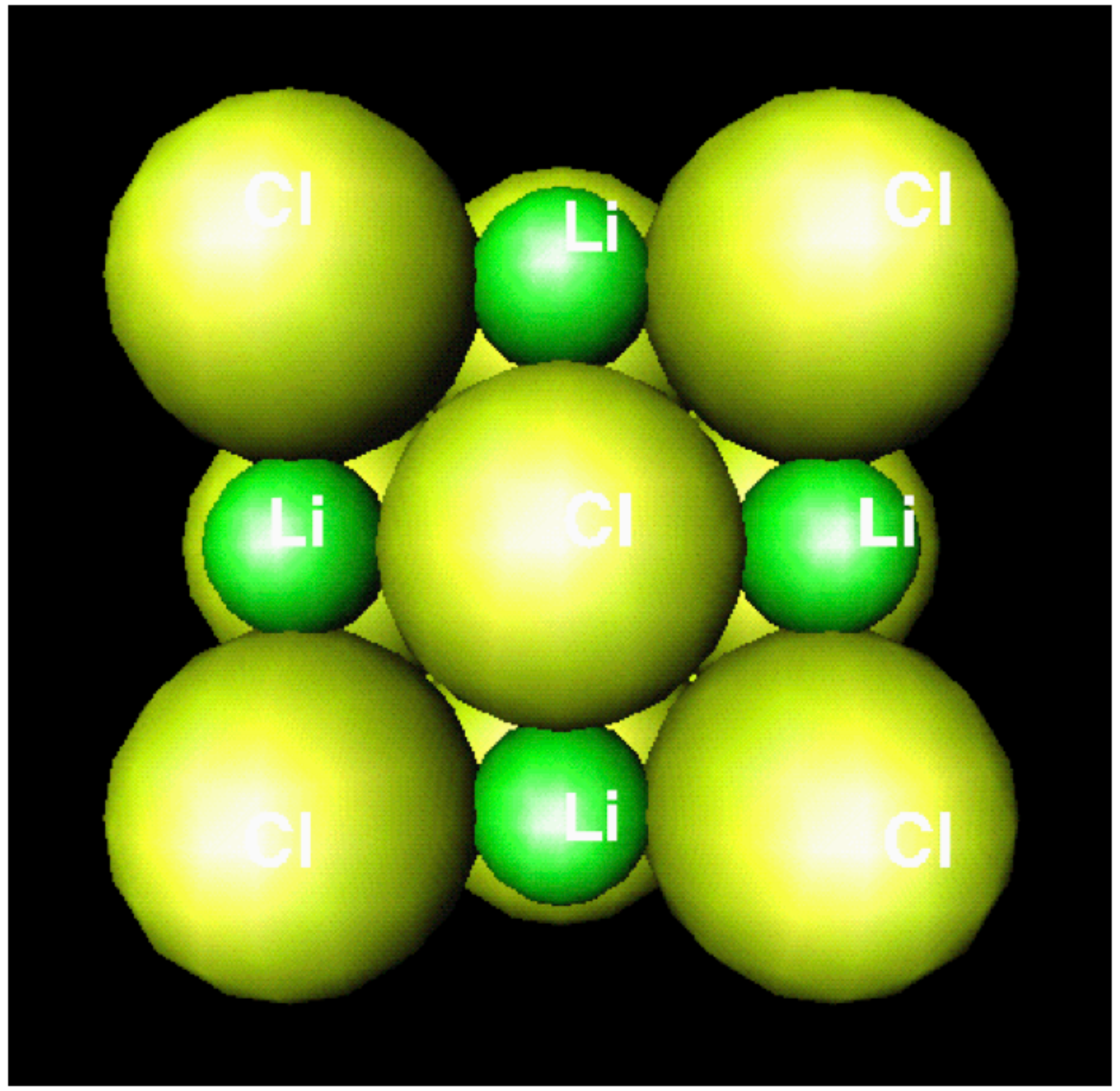
Regression, classification

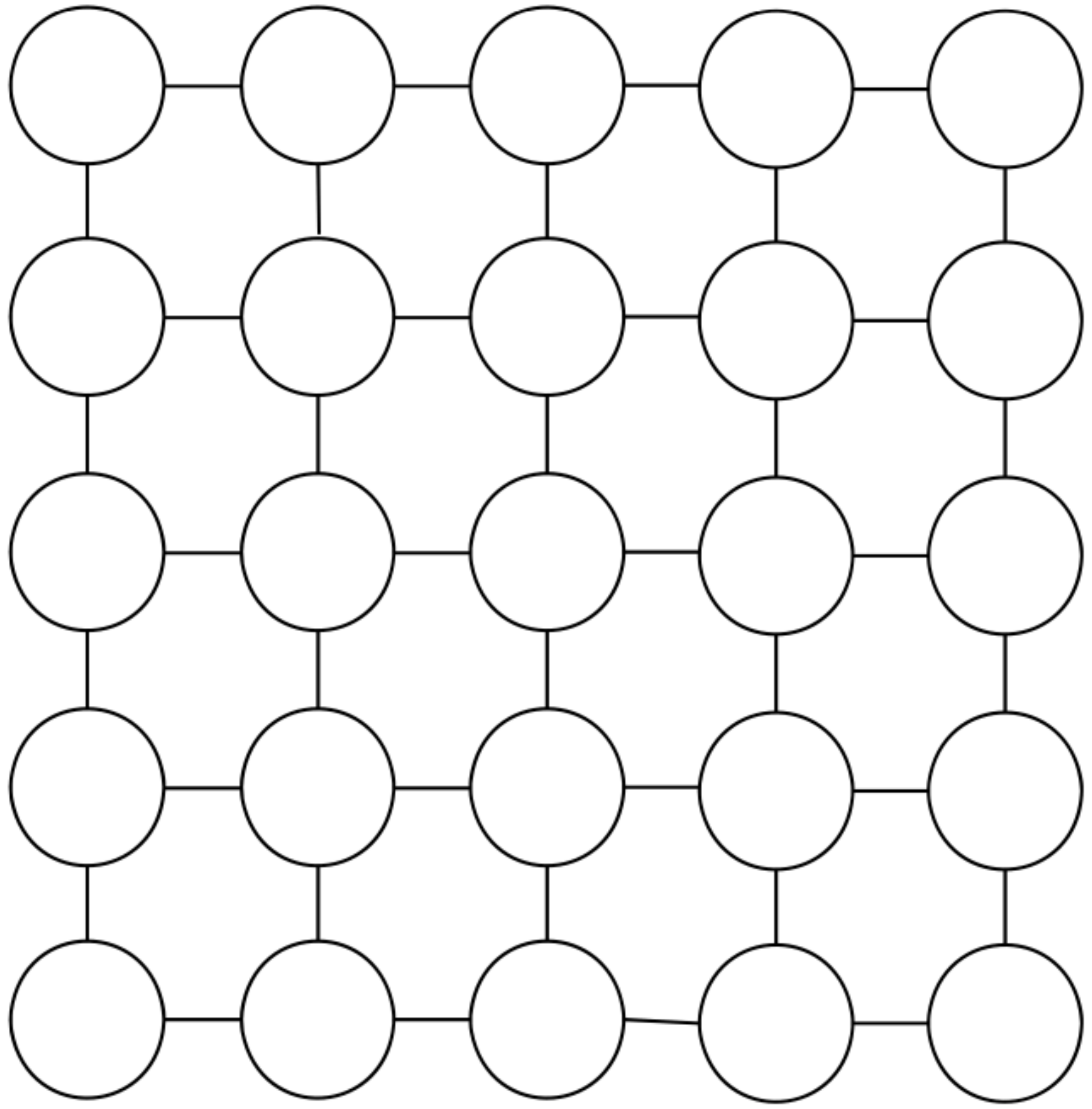


Generative vs
Discriminative Classification

Probabilistic Graphical Model is a language to express distributions

Fancier GMs: Solid State Physics





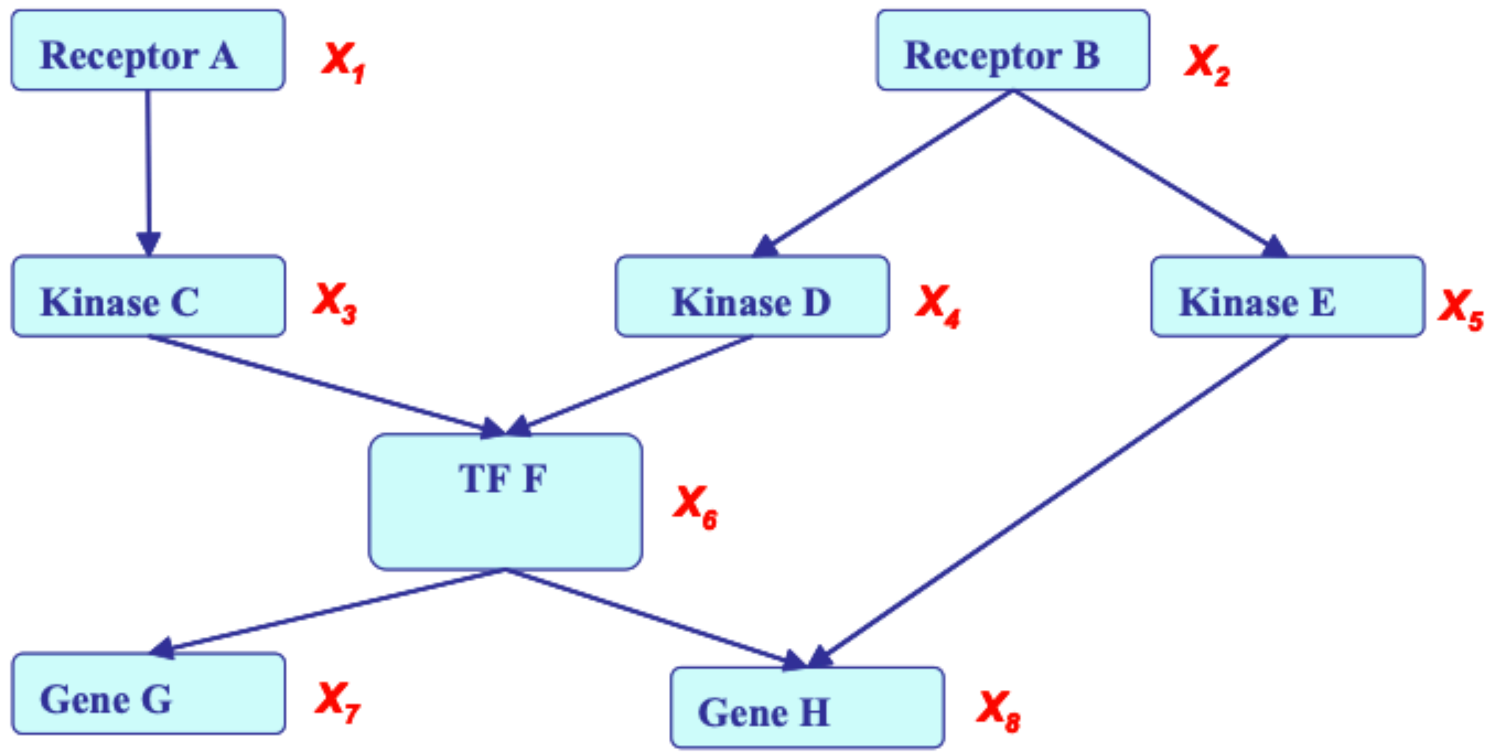
Ising/Potts model

Define the strengths/correlation between different atoms

Why Graphical Models

- A language for communication
- A language for computation
- A language for development

How to Factor a Distribution Given a DAG



$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

- **Theorem:**

  Given a DAG, The most general form of the probability distribution that is consistent with the (probabilistic independence properties encoded in the) graph factors according to "node given its parents":

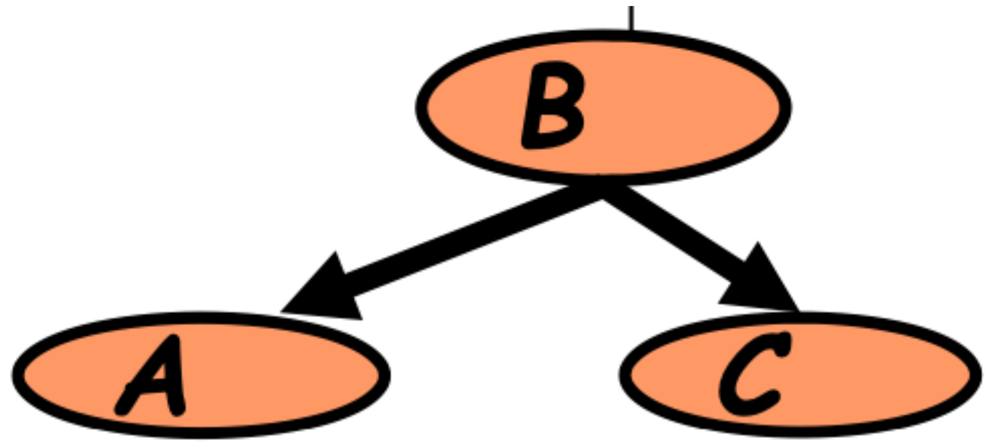
$$P(\mathbf{X}) = \prod_i P(X_i \,|\, \mathbf{X}_{\pi_i})$$

  where $\mathbf{X}_{\pi_i}$ is the set of parents of xi. d is the number of nodes (variables) in the graph.

Local Structures & Independence

- Common parent
  - Fixing B decouples A and C

    "given the level of gene B, the levels of A and C are independent"

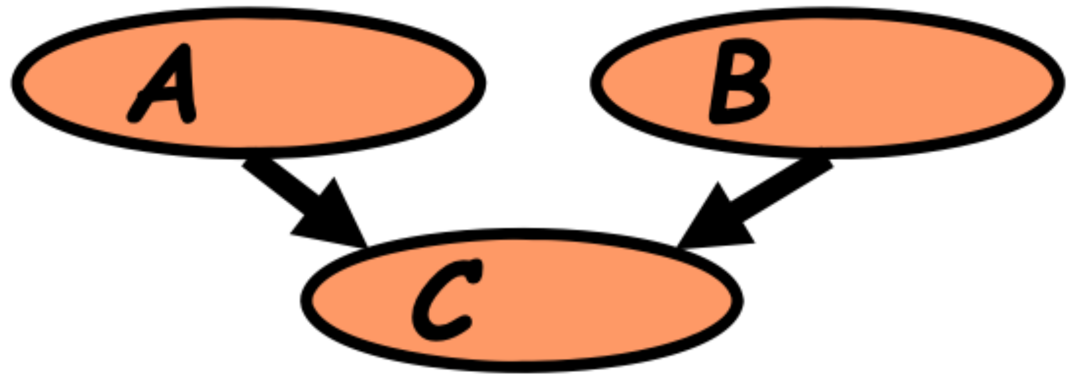


- Cascade
  - Knowing B decouples A and C

    "given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"

- V-structure
  - Knowing C couples A and B
    because A can "explain away" B w.r.t. C

    "If A correlates to C, then chance for B to also correlate to B will decrease"

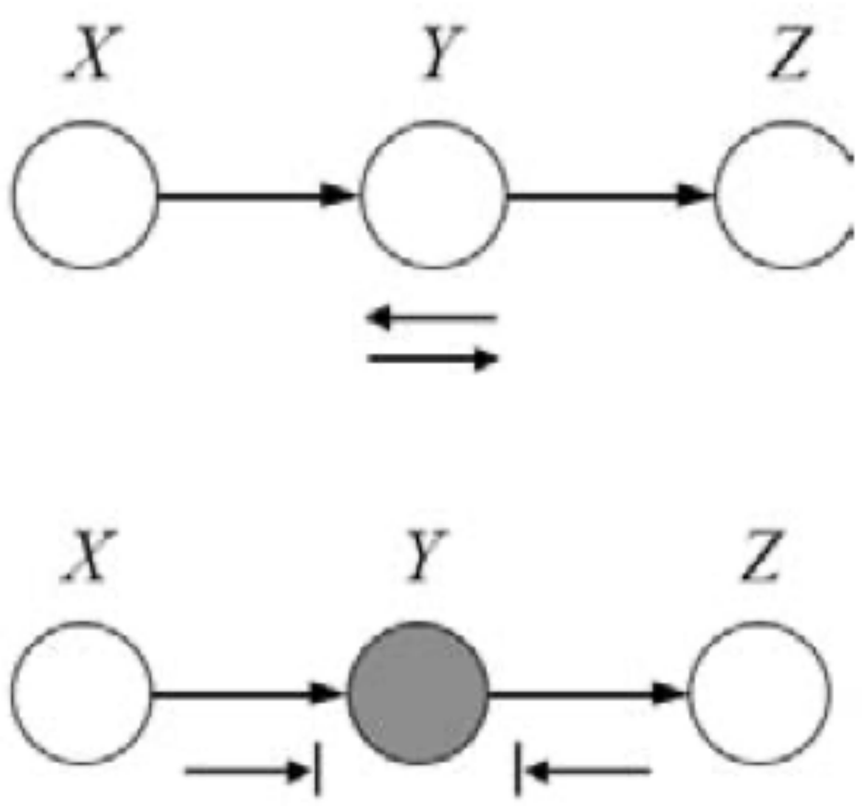


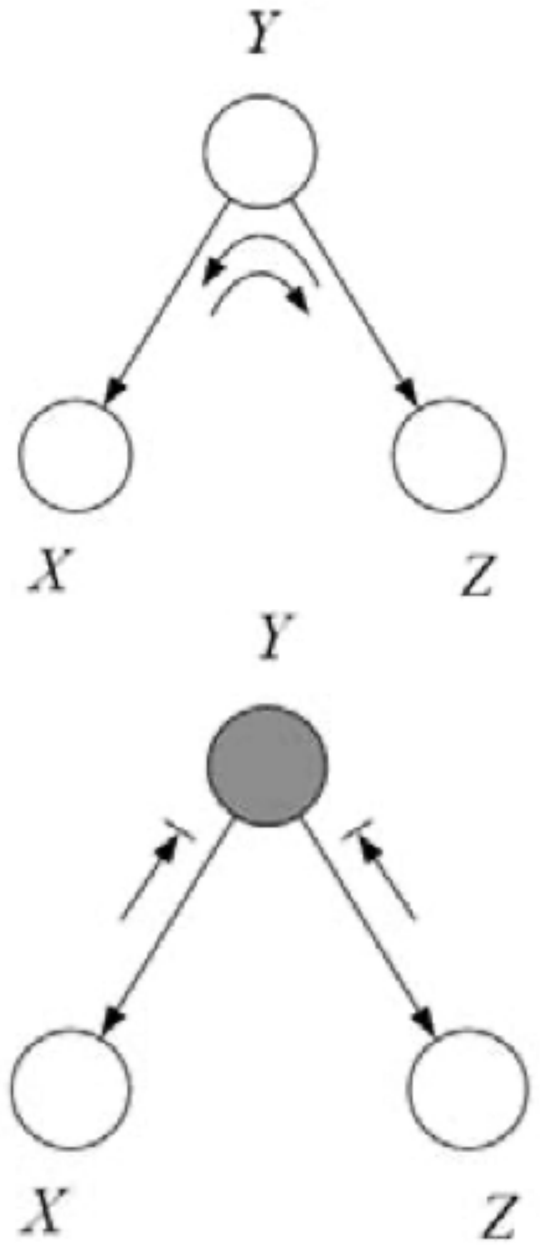
The language is compact, the concepts are rich!

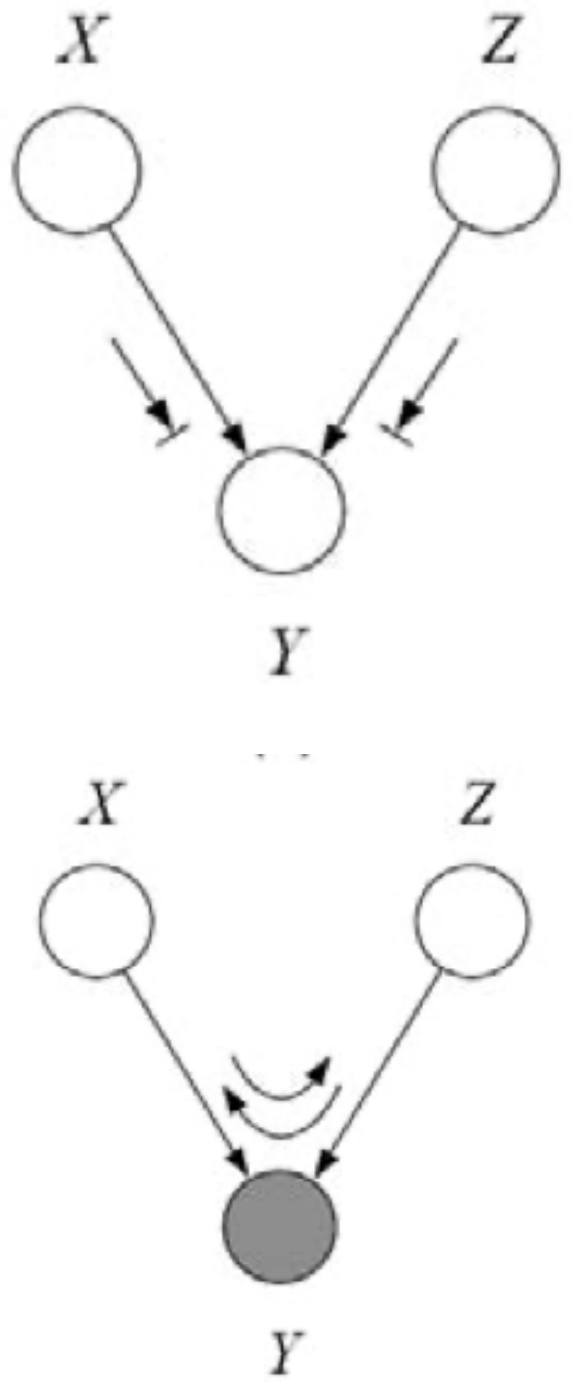
Global Markov Properties of DAGs

How to determine two variables are conditionally independent given another variable?

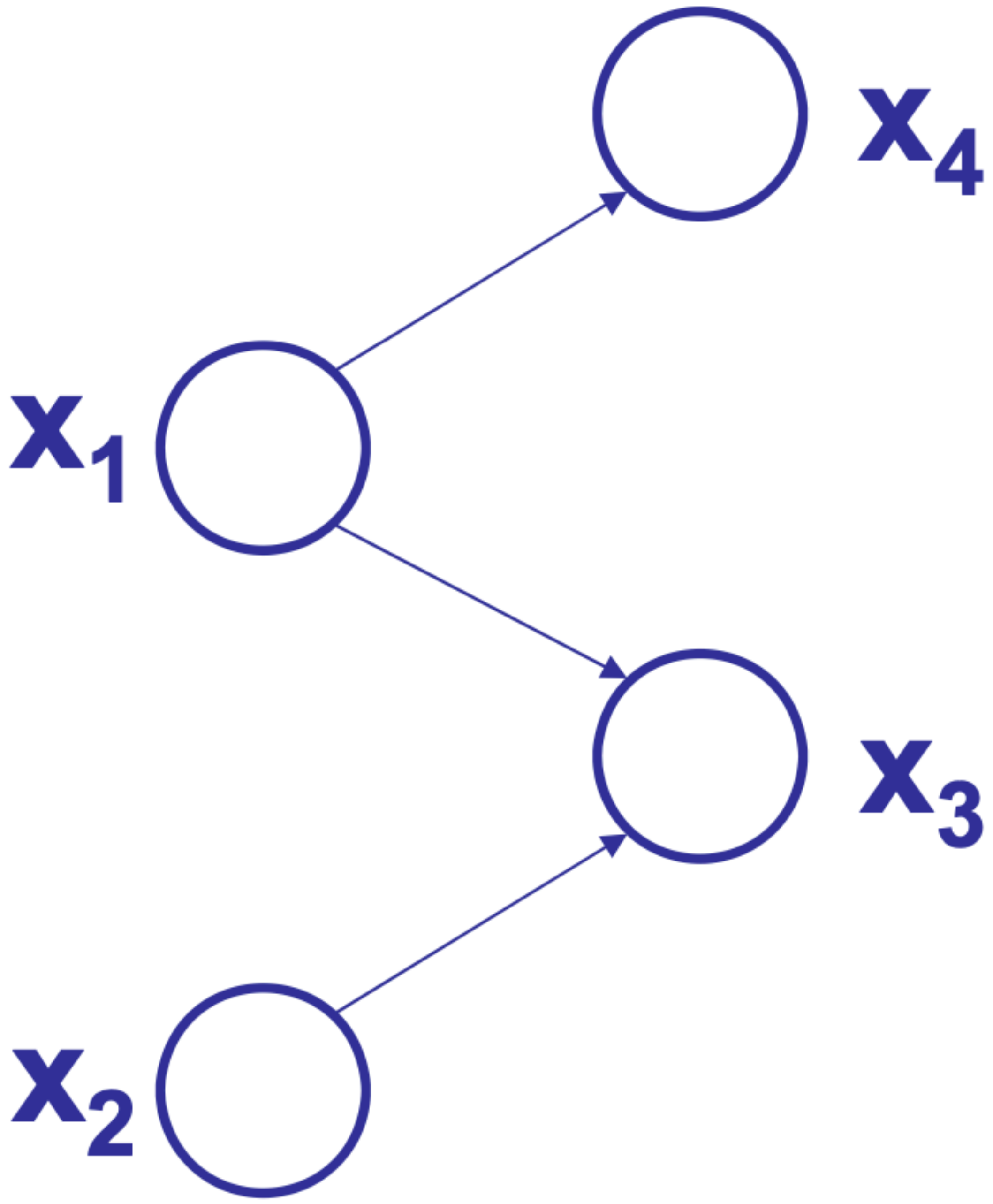
X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "*Bayes-ball*" algorithm illustrated bellow (and plus some boundary conditions):







Example



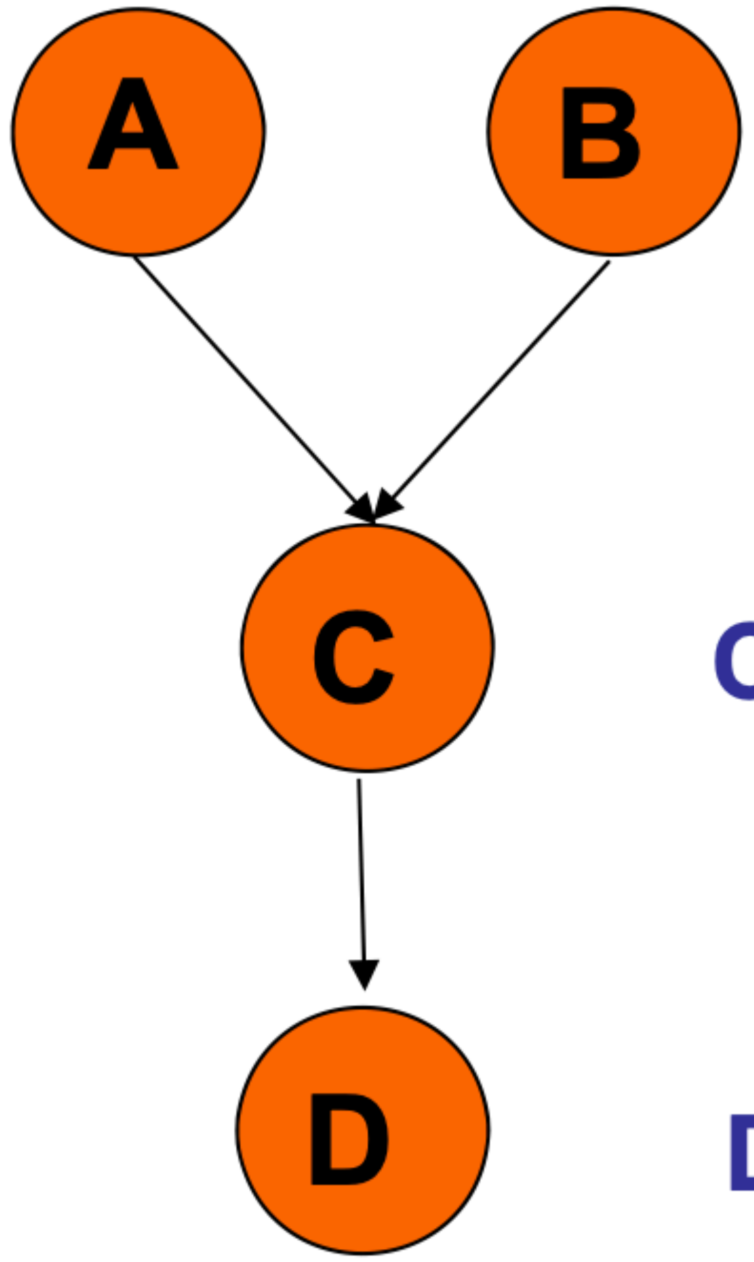
1. Are X2 and X4 independent?

2. Are X2 and X4 conditionally independent given X1?

3. Are X2 and X4 conditionally independent given X3?

Conditional Probability Density Func

$A \sim N(\mu_a, \Sigma_a)$ $B \sim N(\mu_b, \Sigma_b)$

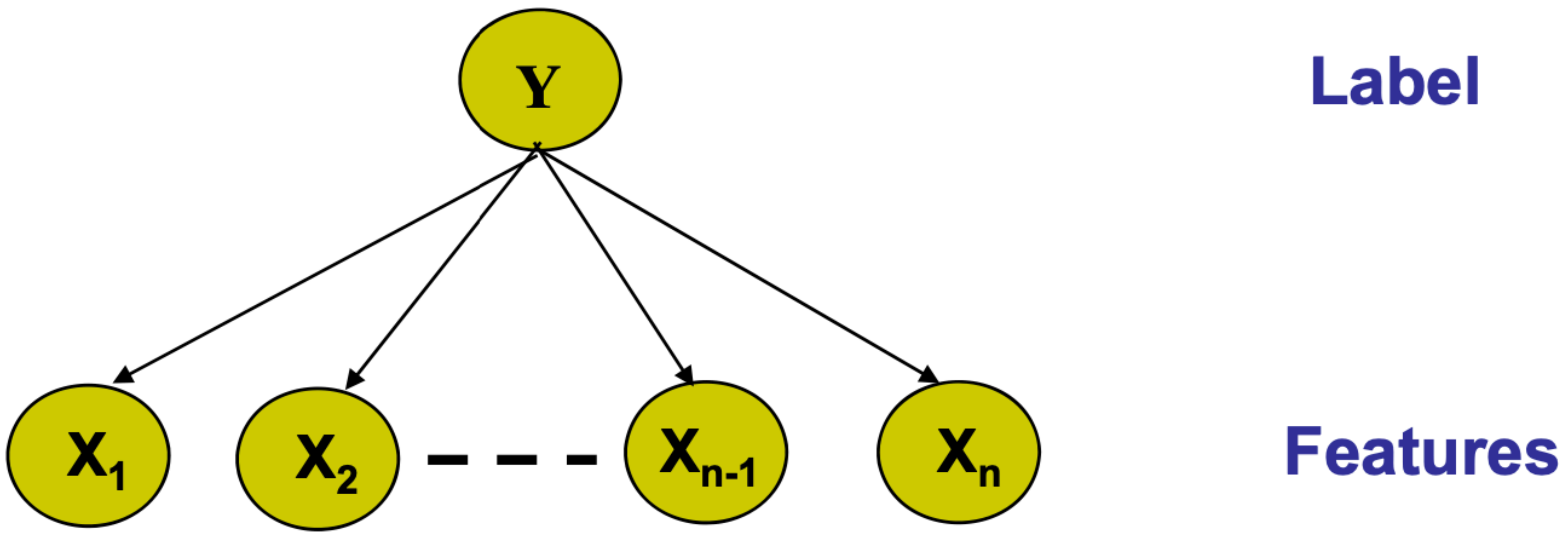


$C \sim N(A+B, \Sigma_c)$

$D \sim N(\mu_a + C, \Sigma_a)$

$$\boxed{\begin{gathered} P(a,b,c.d) = \\ P(a)P(b)P(c|a,b)P(d|c) \end{gathered}}$$

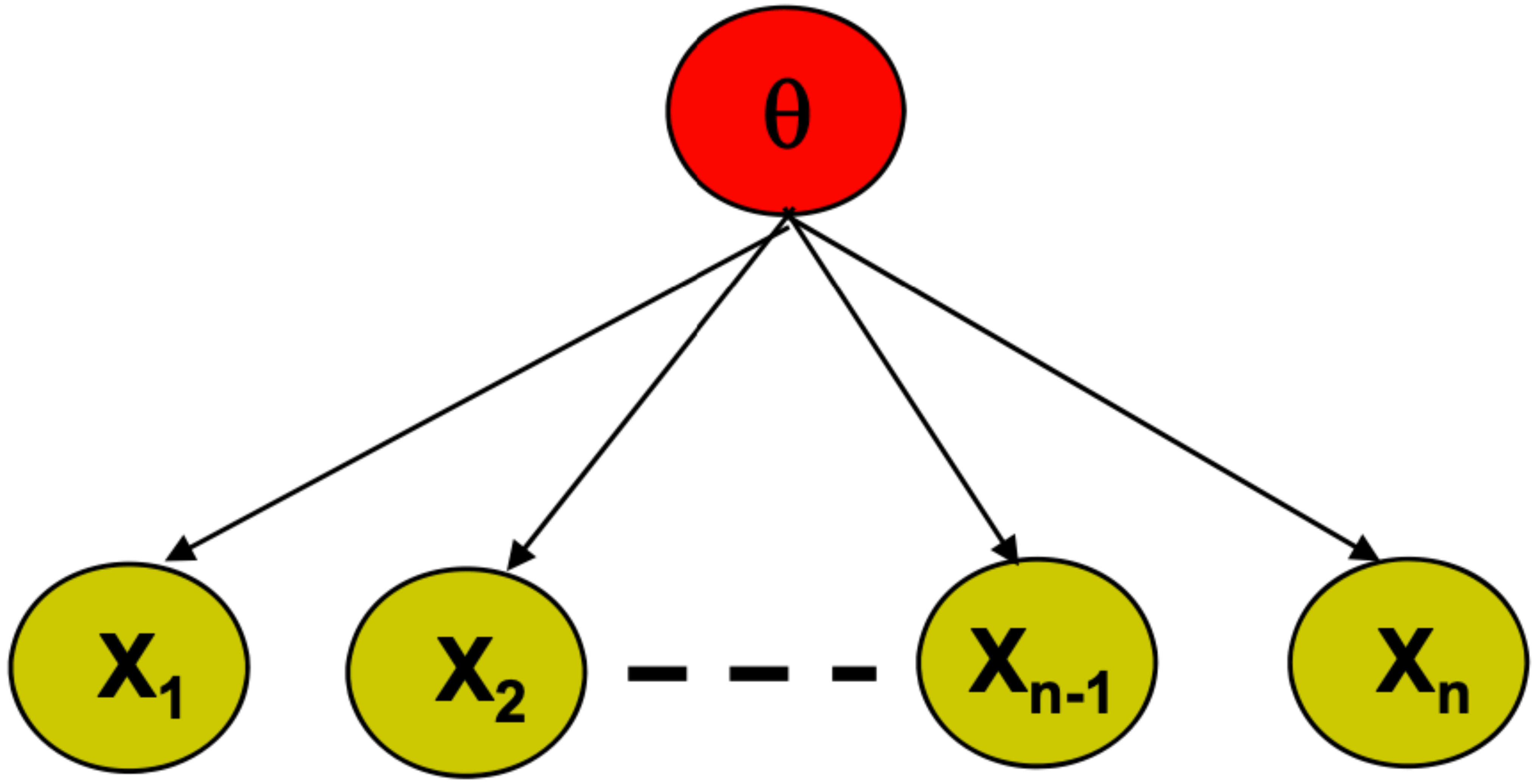
Conditional Independencies



Are Xi D-separated from Xj given Y?

What is this model when Y is observed?

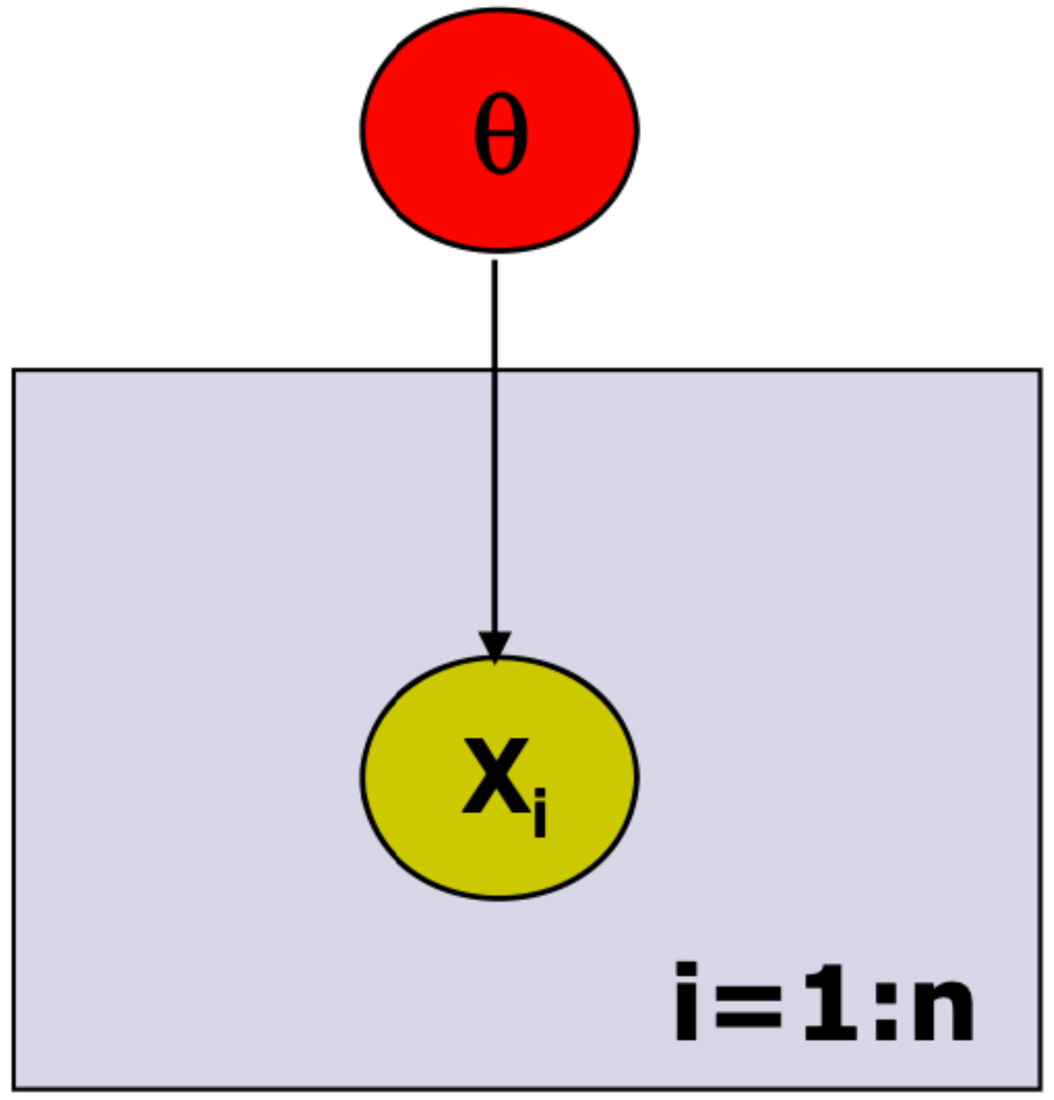
Conditionally Independent Observations



Model parameters

Data {X1, X2 …. Xn}

"Plate" Notation



Model parameters

Data = $\{x_1, \ldots x_n\}$

variables within a plate are replicated in a conditionally independent manner

Example: Gaussian Model

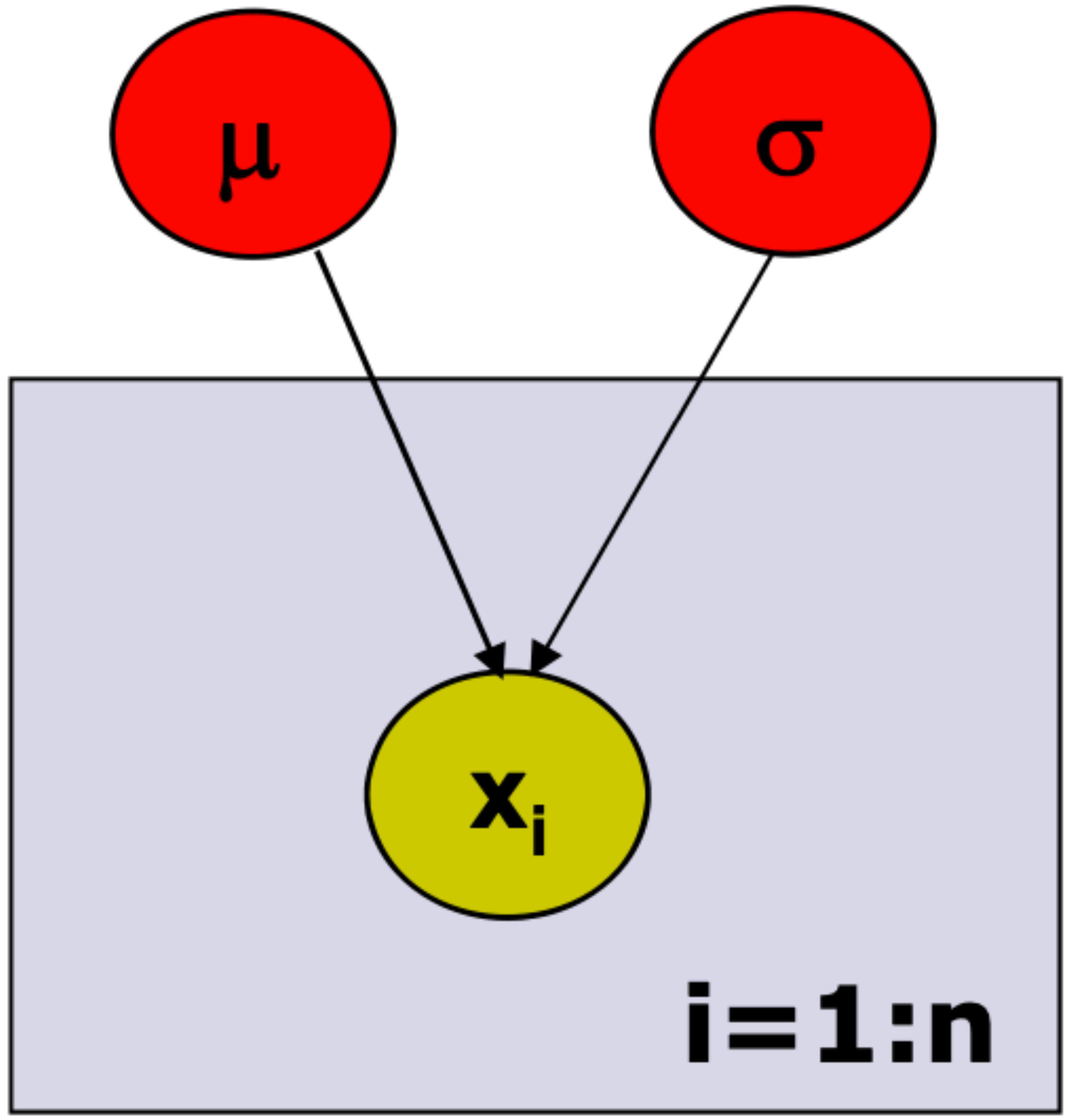
Generative model:

$p(x_1,\ldots x_n \mid \mu, \sigma) \quad = P \;\; p(x_i \mid \mu, \sigma)$

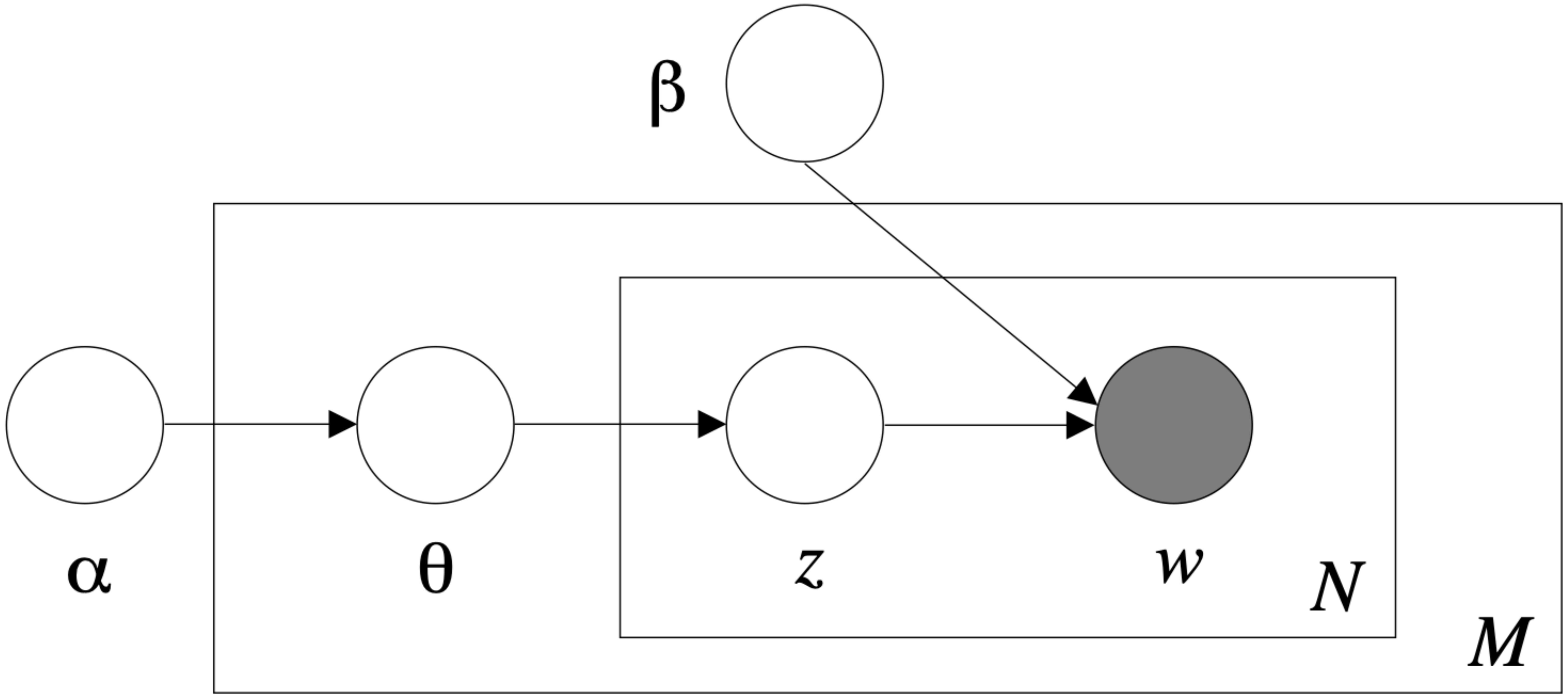
$= \; p(\text{data} \mid \text{parameters})$

$= \; p(D \mid \theta)$

where $\theta = \{\mu, \sigma\}$



Observed Variable and Latent Variable Notations



We typically use gray variables to denote observed variables

Gaussian Mixture Model / Gaussian Discriminative Analysis in PGMs

Inference and Learning

Query a node (random variable) in the graph

- Task 1: How do we answer **queries** about P?
  - We use **inference** as a name for the process of computing answers to such queries

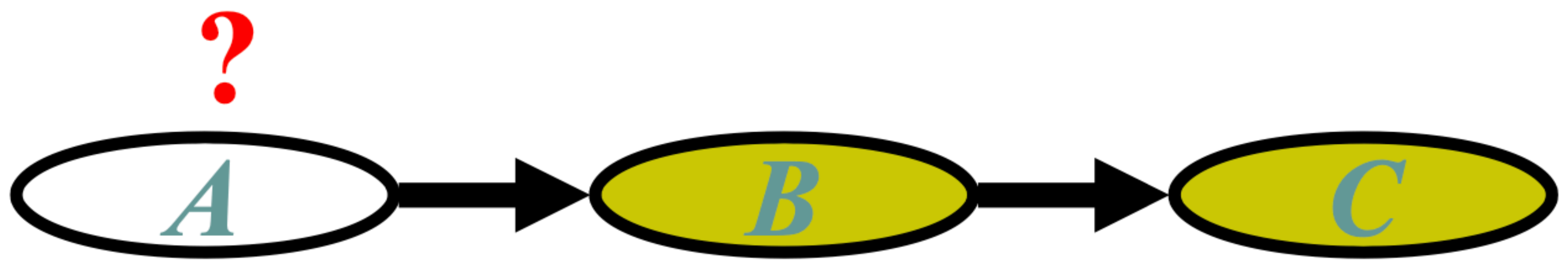
- Task 2: How do we estimate a **plausible model** M from data D?
  i. We use **learning** as a name for the process of obtaining point estimate of M.

Examples

- **Prediction**: what is the probability of an outcome given the starting condition

  - the query node is a descendent of the evidence

- **Diagnosis**: what is the probability of disease/fault given symptoms



  - the query node an ancestor of the evidence

In practice, the observed variable is often the data that is on the leaf nodes

How to Learn the Parameters

1. When θ is the parameter and does not have prior —> MLE

$$p(x, z; \theta)$$

2. When we add the prior over θ —> MAP (Bayesian)

$$p(x, z, \theta)$$

How to do MLE on Latent Variable Models?

Expectation Maximization!

The E-step computes the posterior distribution $p(z|x)$

This process is referred to as inference

Approaches to Inference

- Exact inference algorithms
  - The elimination algorithm
  - Belief propagation
  - The junction tree algorithms (but will not cover in detail here)

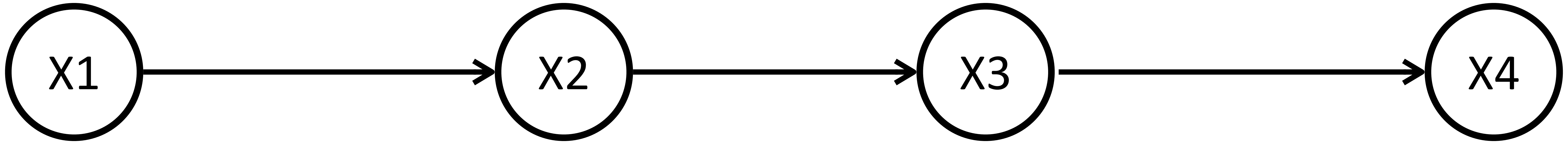
- Approximate inference techniques
  - Variational algorithms
  - Stochastic simulation / sampling methods
  - Markov chain Monte Carlo methods

Variational Autoencoders

Elimination Algorithm/ Marginalization

$$P(h) = \sum_g \sum_f \sum_e \sum_d \sum_c \sum_b \sum_a P(a,b,c,d,e,f,g,h)$$

a naïve summation needs to enumerate over an exponential number of terms



What if the random variables follow this chain structure?

Thank You!
Q & A