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

COMP 5212
Machine Learning
Lecture 12

$p(z)$

z $\quad 1, 2, 3 - K$

x $\quad \mathcal{N}(\mu_{z_i}, \Sigma_{z_i})$

GMM

Probabilistic Graphical Models

VAE GANs

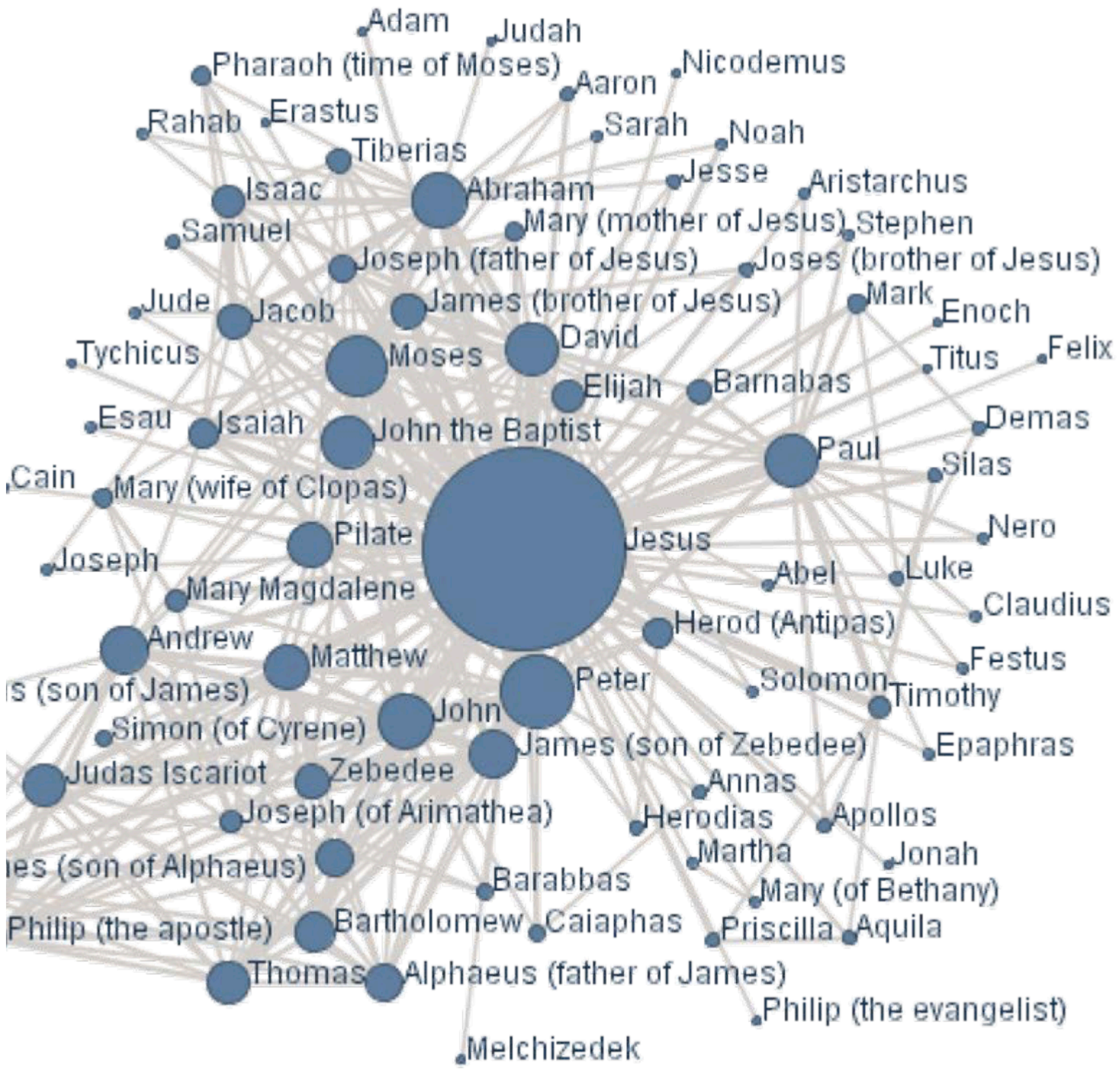
diffusion

Junxian He
Mar 24, 2026

Some Announcements

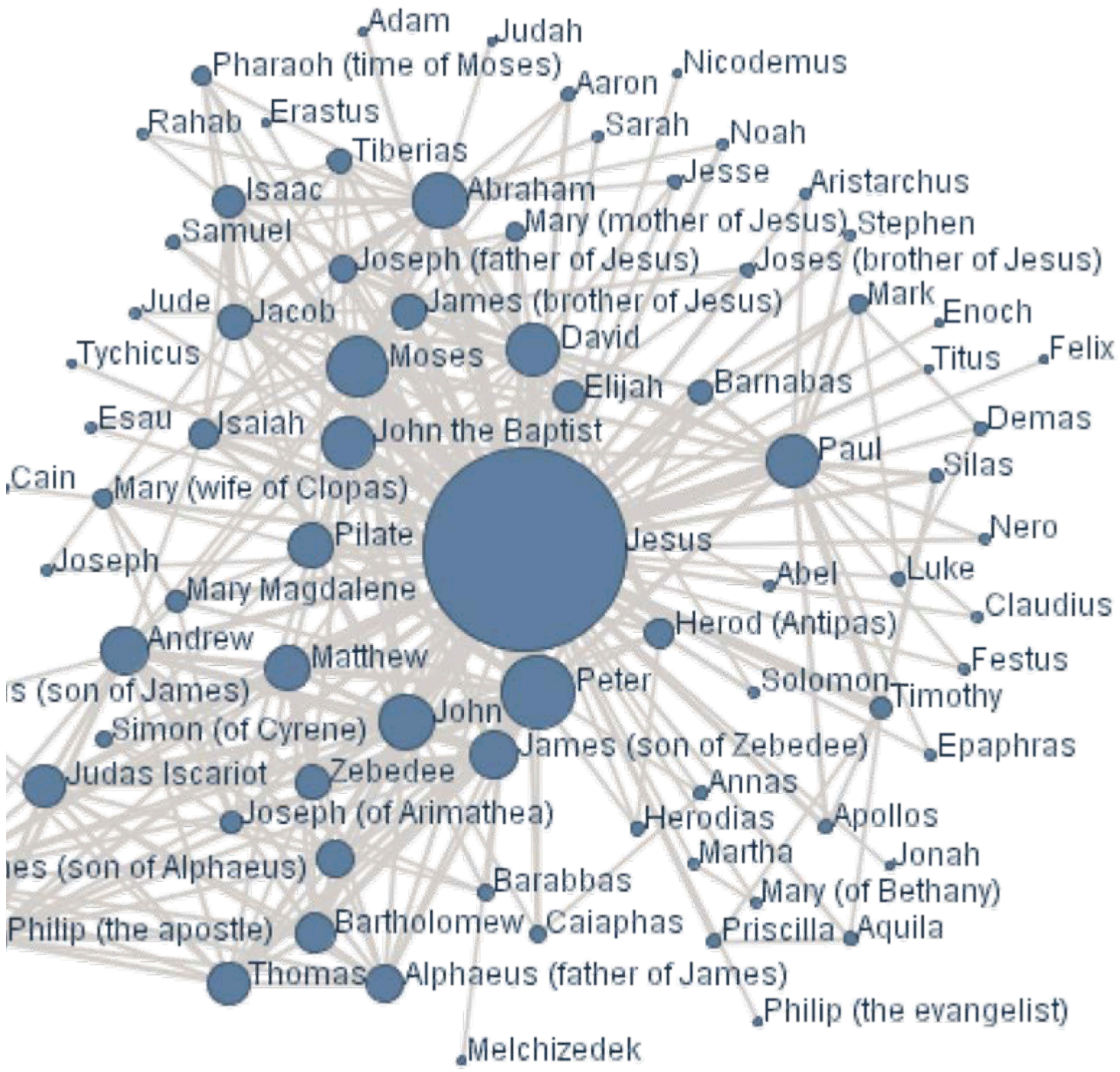
- Midterm Exam next Tuesday, in-class exam → , Em]
- Lecture recordings have been released on Canvas

What Are Graphical Models?



What Are Graphical Models?

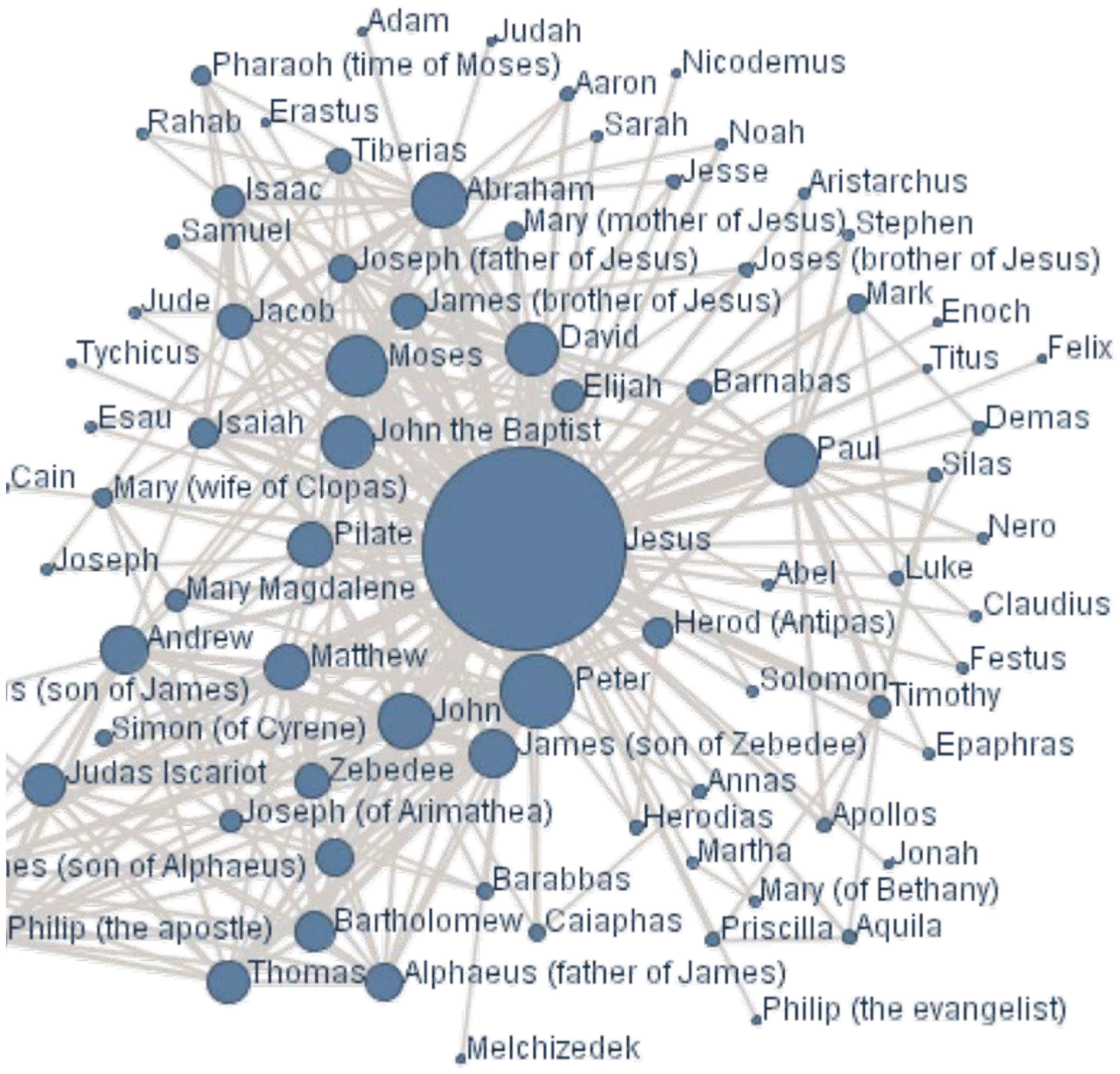
- Informally, a GM is just a graph representing relationship among random variables
  - Nodes: random variables (features, not examples)
  - Edges (or absence of edges): relationship



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
  - ...

causation

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

Relationship between two random variables

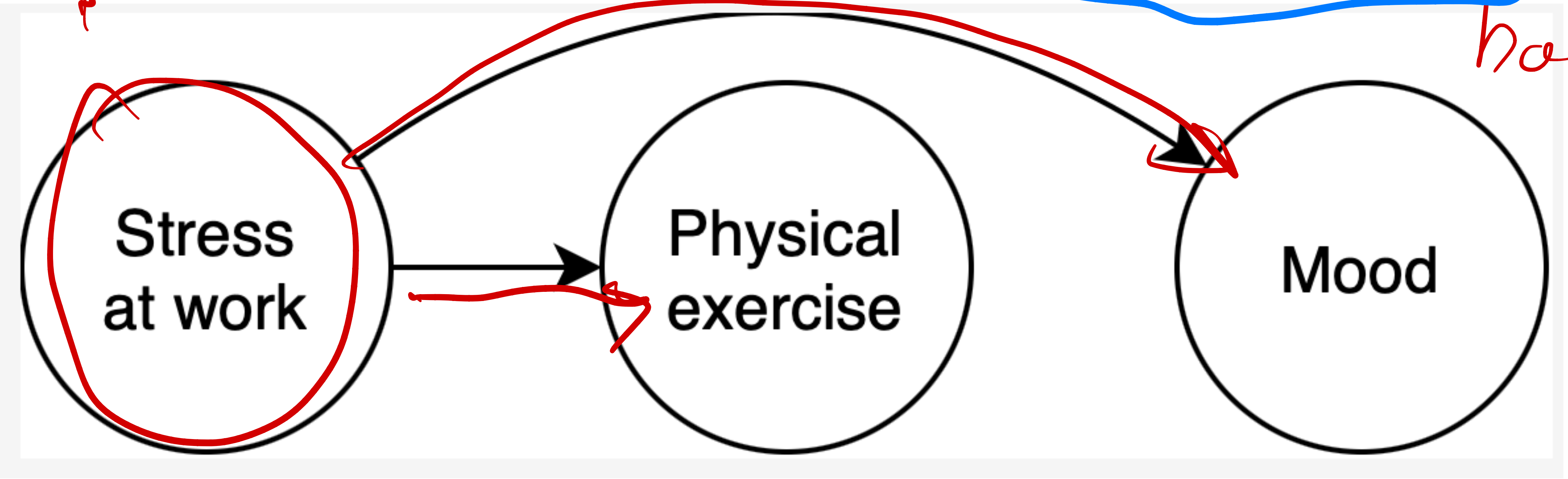
correlated?

physical exercise → bad mood

- 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

bad mood



What is a Graphical Model?

What is a Graphical Model?

Graphical model represents a multivariate distribution in High-D space

What is a Graphical Model?

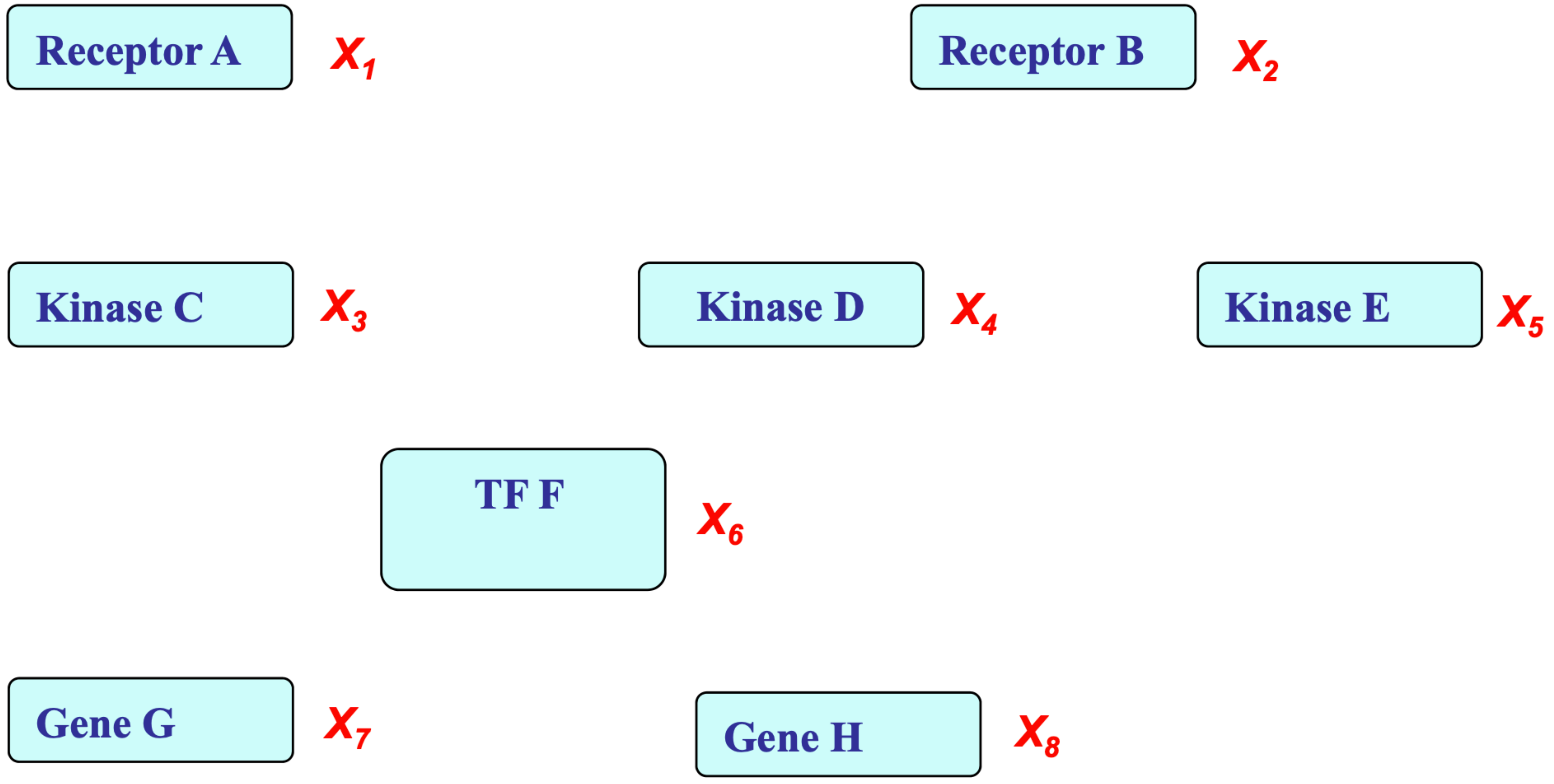
Graphical model represents a multivariate distribution in High-D space

A possible world for celluar signal transduction:

What is a Graphical Model?

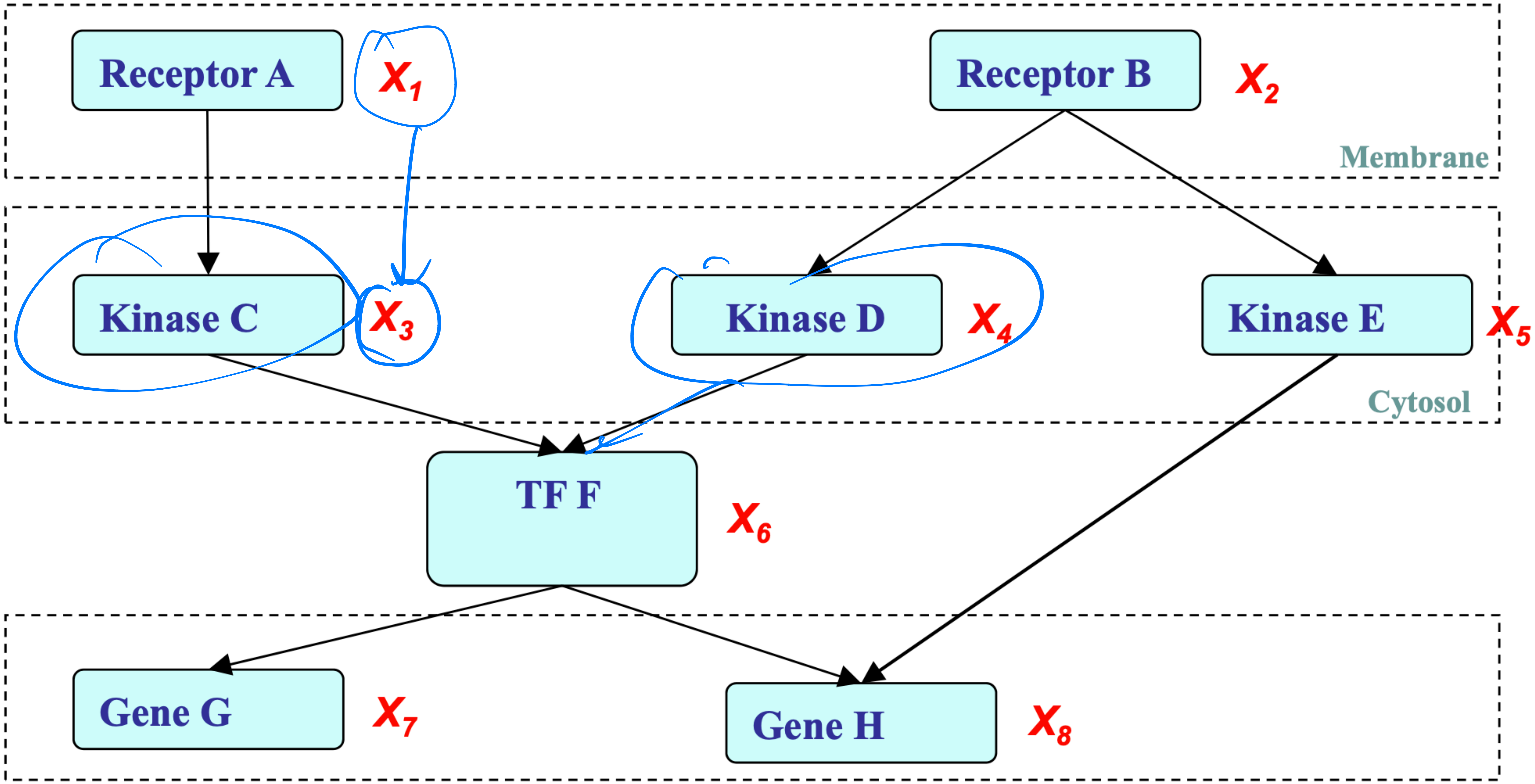
Graphical model represents a multivariate distribution in High-D space

A possible world for celluar signal transduction:



Structure Simplifies Representation

Dependencies among variables



Probabilistic Graphical Models

Probabilistic Graphical Models

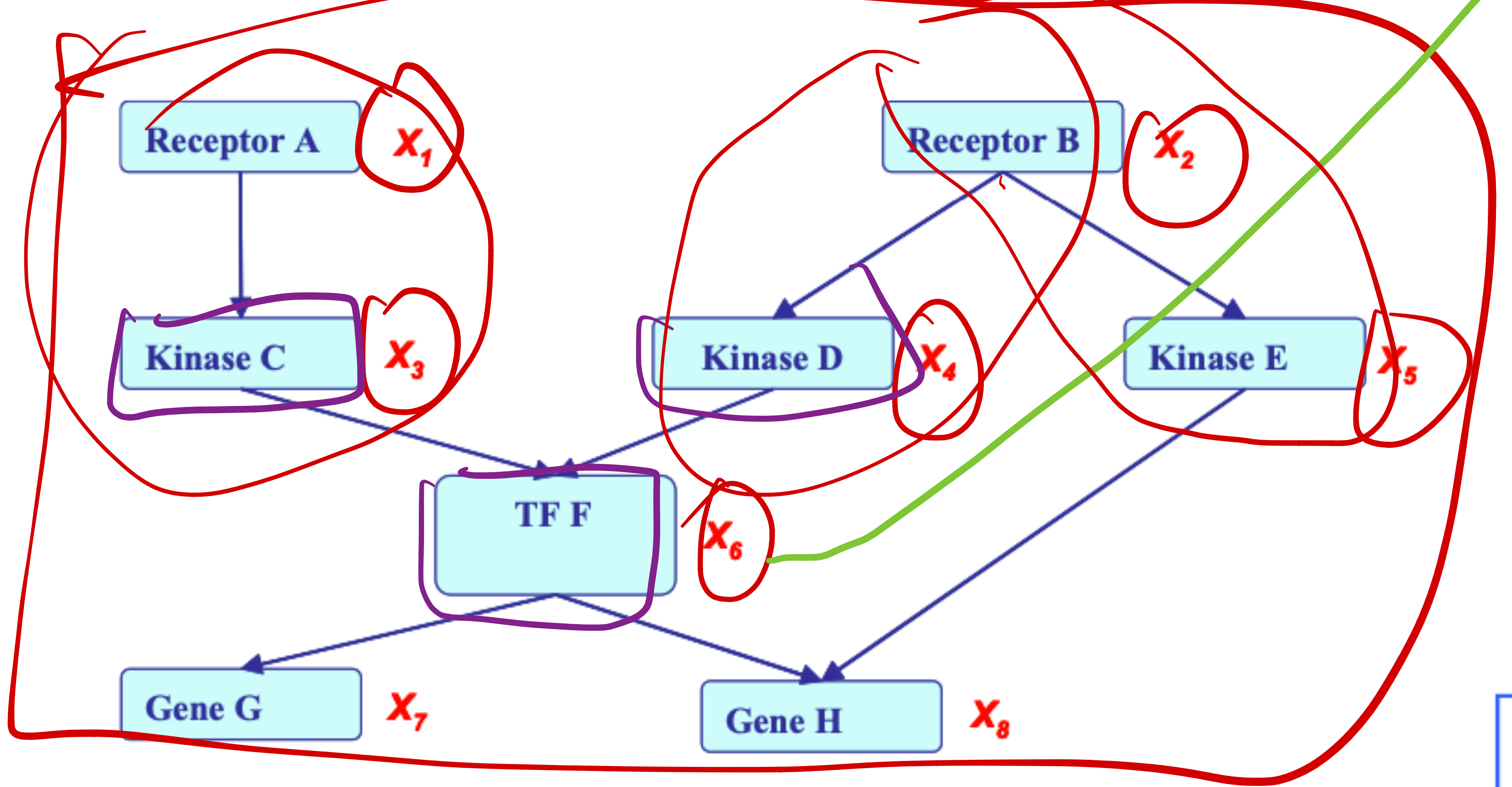
- If X_i's are **conditionally independent** (as described by a **PGM**), the joint can be factored to a product of simpler terms, e.g.,

Probabilistic Graphical Models

$X_6 \perp X_1, X_2 \dots$ given X_3, X_4

- 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.,

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



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

Stay tune for what are these independencies!

$P(X_1, X_2 \dots X_8) = P(X_1) P(X_2 | X_1) P(X_3 | X_1, X_2) \dots P(X_4 \dots | X_1, X_2, X_3)$

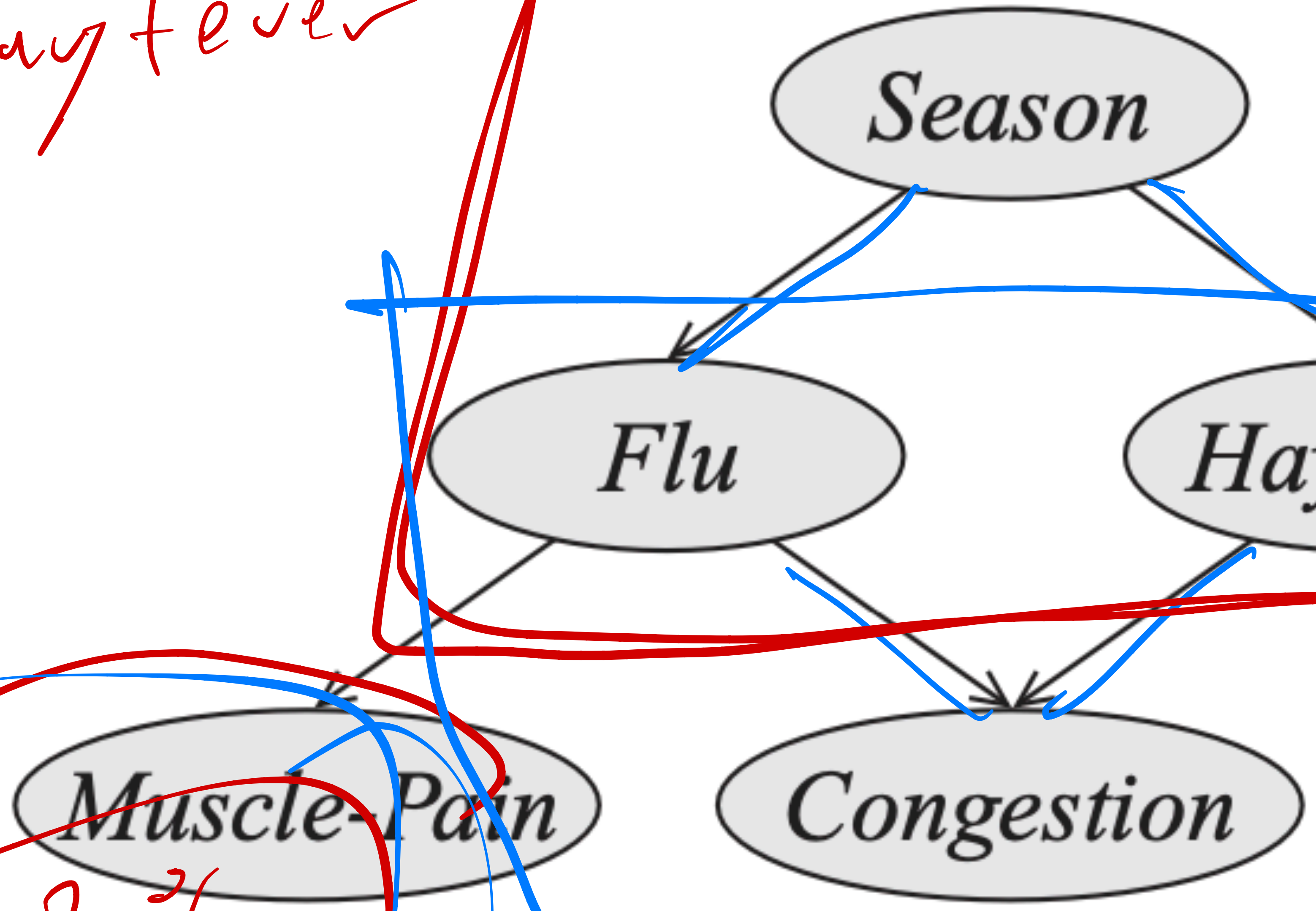
$\dots P(X_8 | X_1, X_2 \dots X_7)$

Another Example

Season

flu

Hayfever



Flu

Hayfever

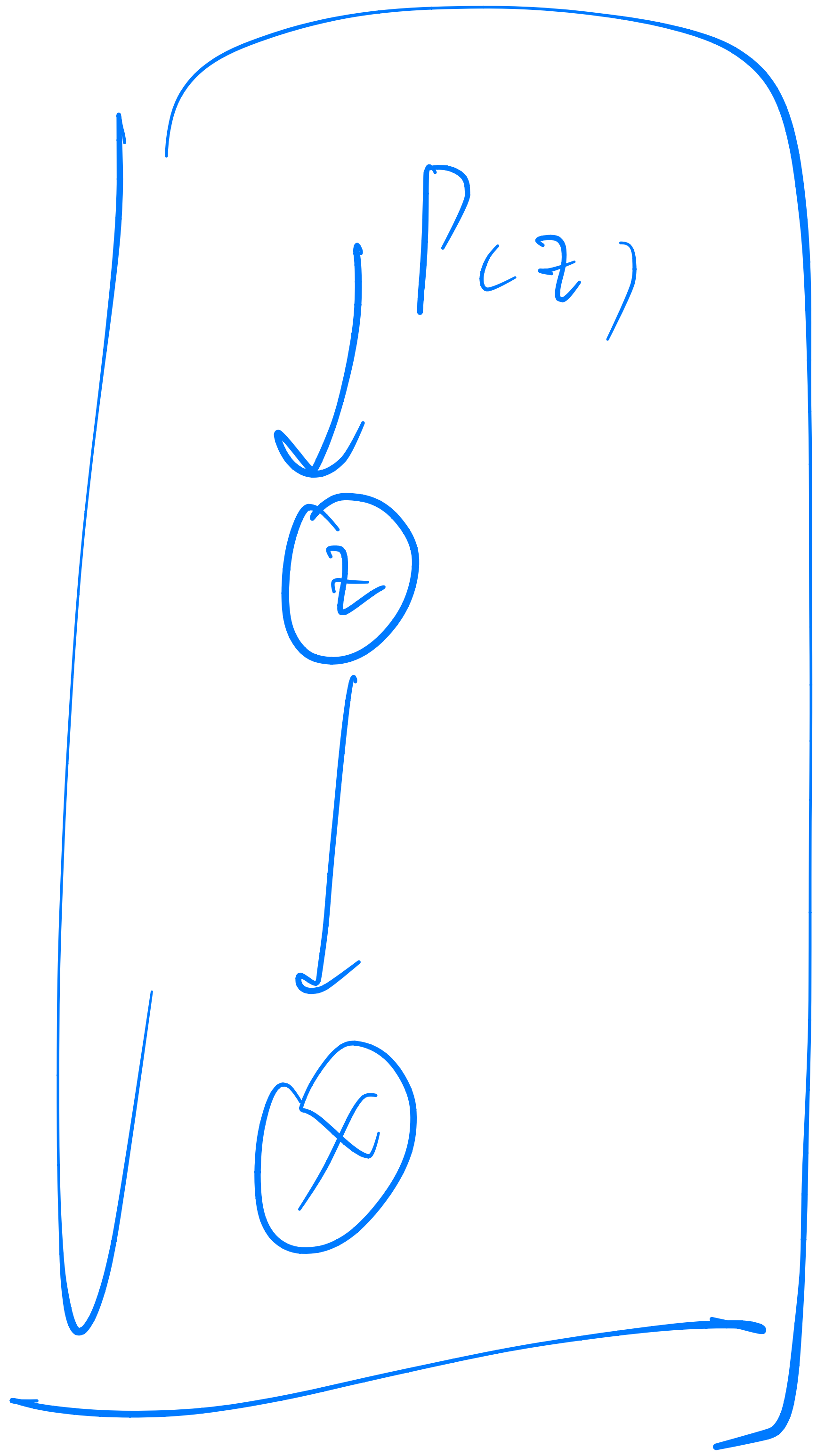
Congestion

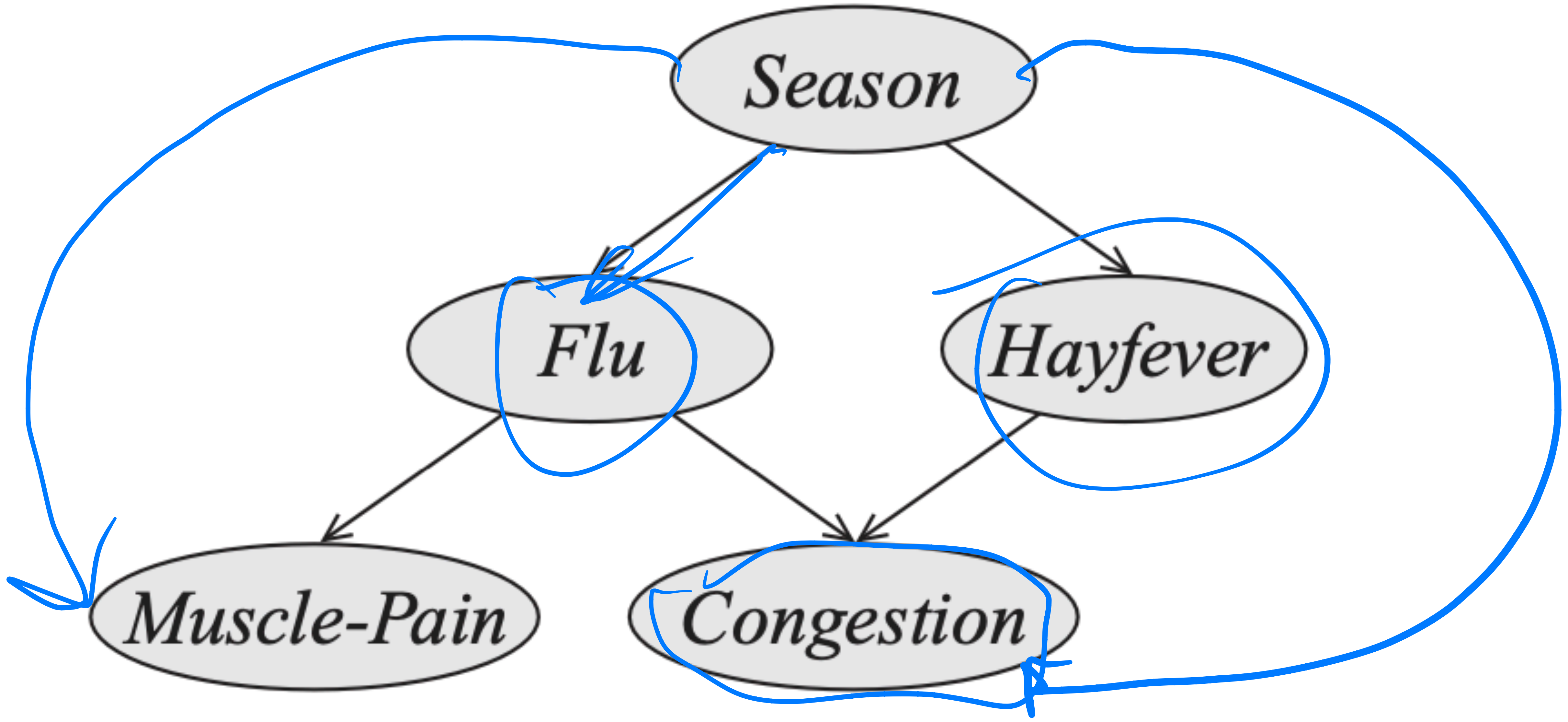
flu → 80% winter → Hayfever

Congestion Yes

~~Flu Yes~~

Another Example





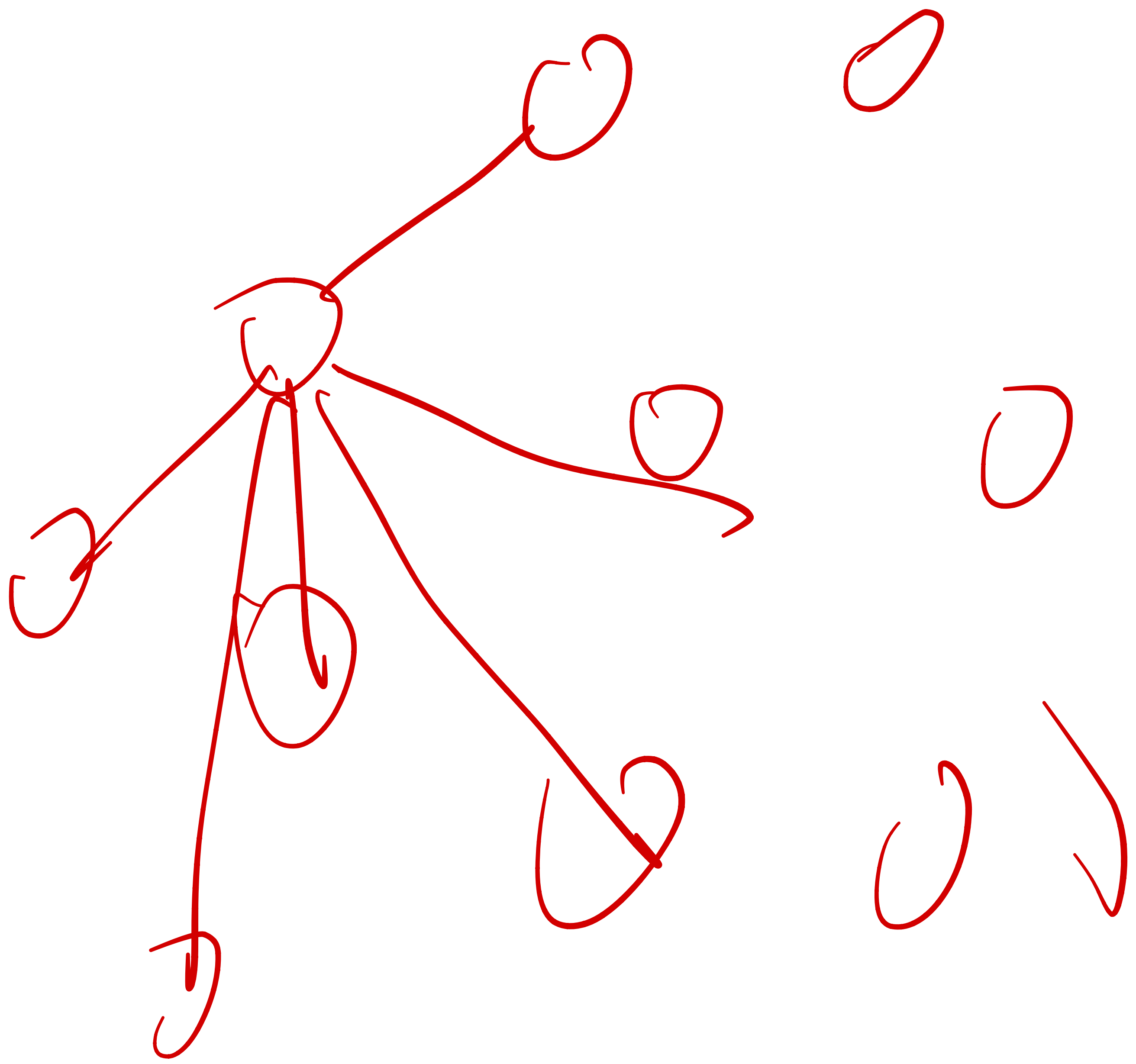
$P(\text{muscle-pain} \mid \text{flu}, \text{season}) = P(\text{muscle-pain} \mid \text{flu})$

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

What is a PGM After All

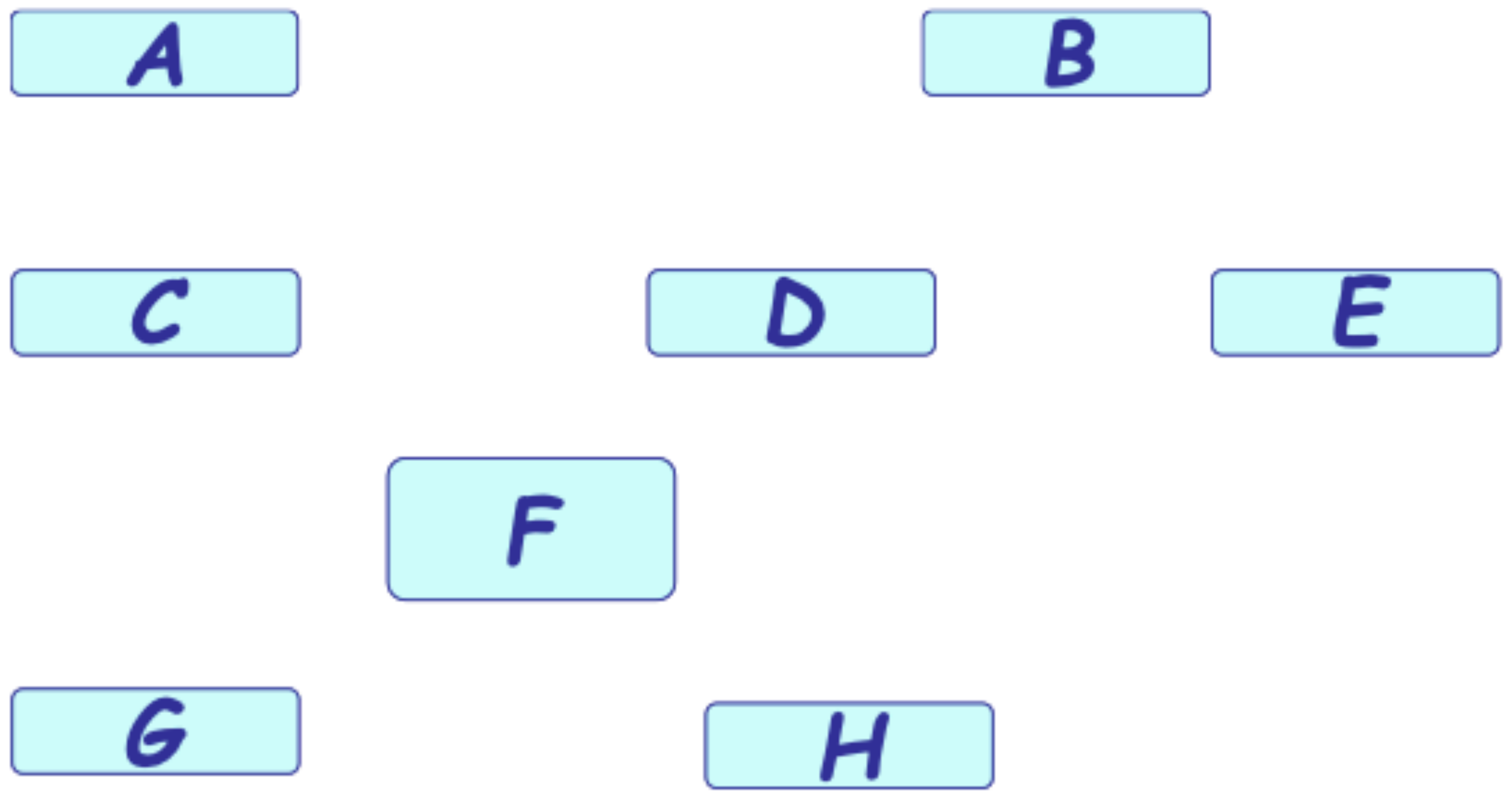
What is a PGM After All

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***



What is a PGM After All

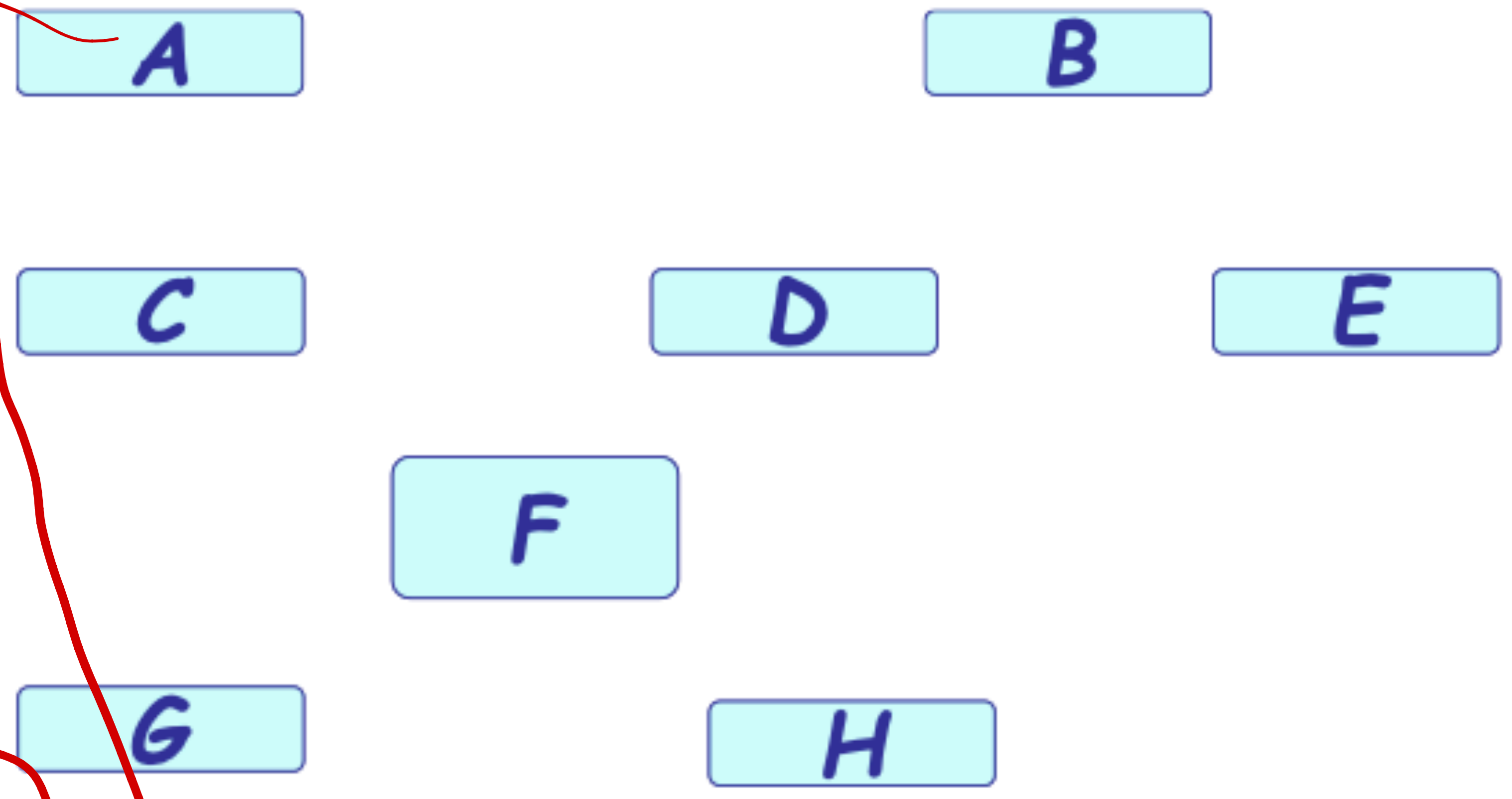
It is a smart way to write/specify/compose/design exponentially-large probability distributions without paying an exponential cost, and at the same time endow the distributions with ***structured semantics***



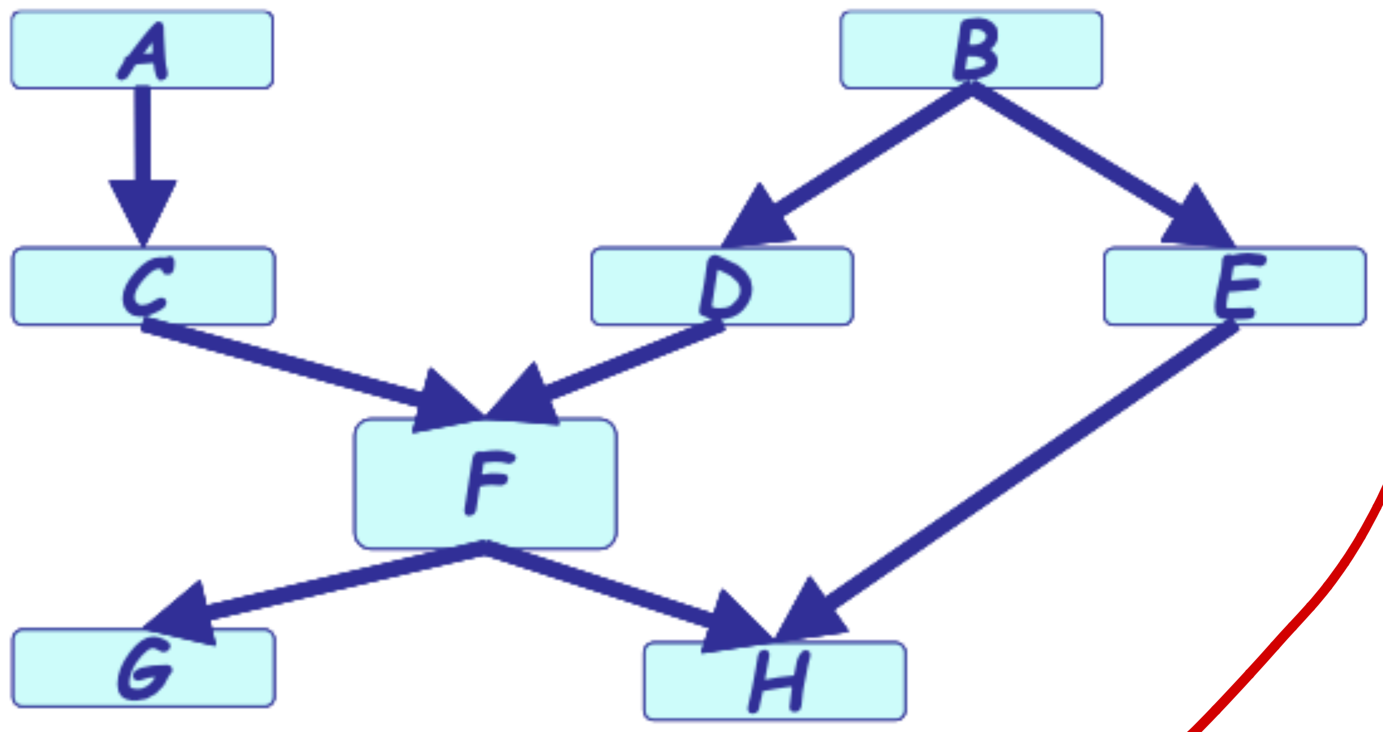
$$P(X_1,X_2,X_3,X_4,X_5,X_6,X_7,X_8)$$

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***



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

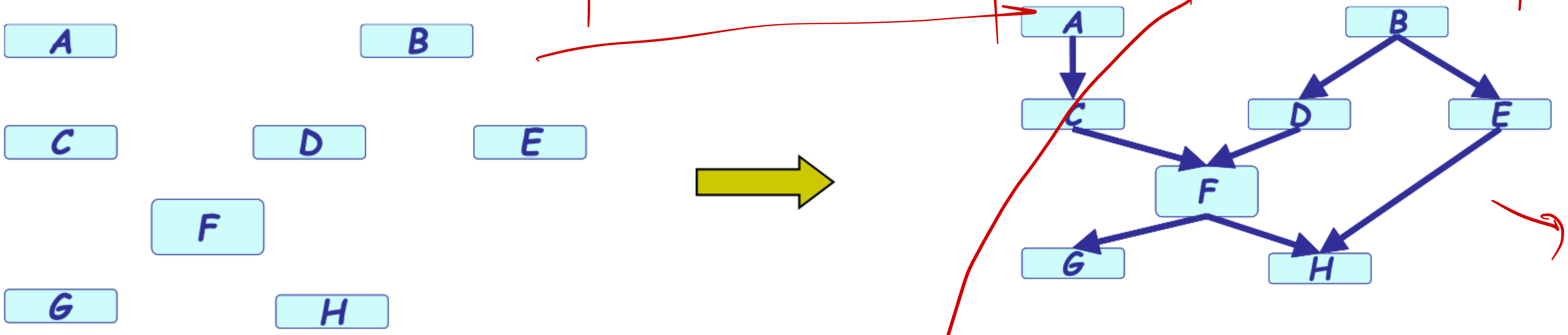


$$P(X_{1:8}) = P(X_1)P(X_2)P(X_3 \mid X_1 X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$
$$P(X_6 \mid X_3, X_4)P(X_7 \mid X_6)P(X_8 \mid X_5, X_6)$$

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***

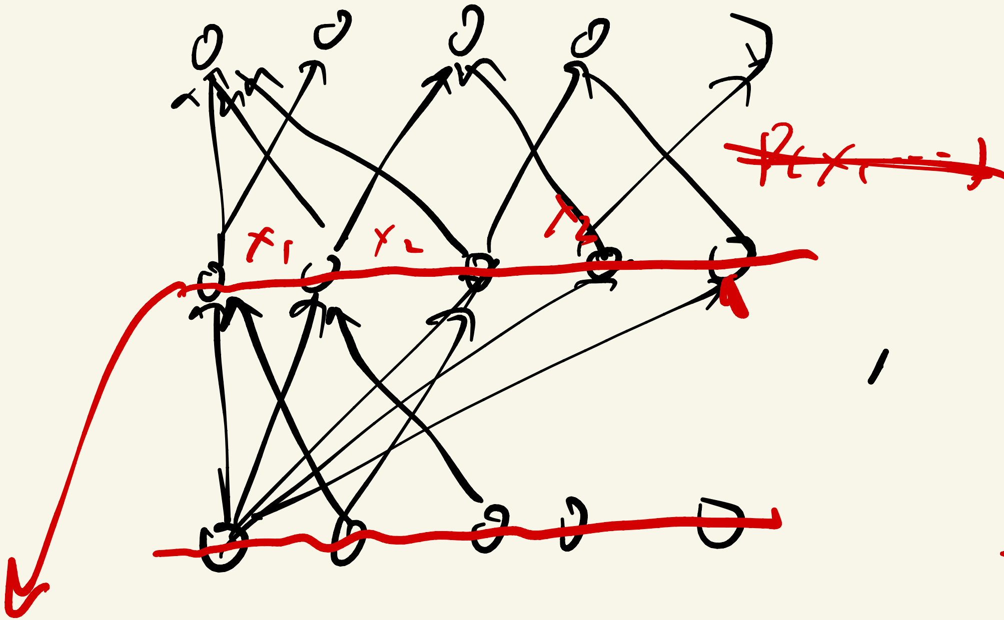
$X_1 - - - X_8$ $P(X_1, X_2 \ldots X_8)$



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

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

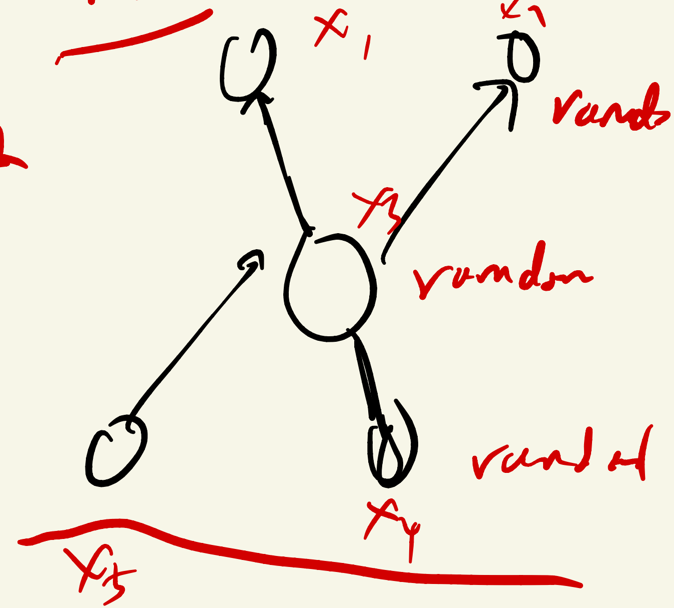
More formal definition: It refers to a family of distributions on a set of random variables that are compatible with all the probabilistic independence propositions encoded by a graph that connects these variables



deterministic

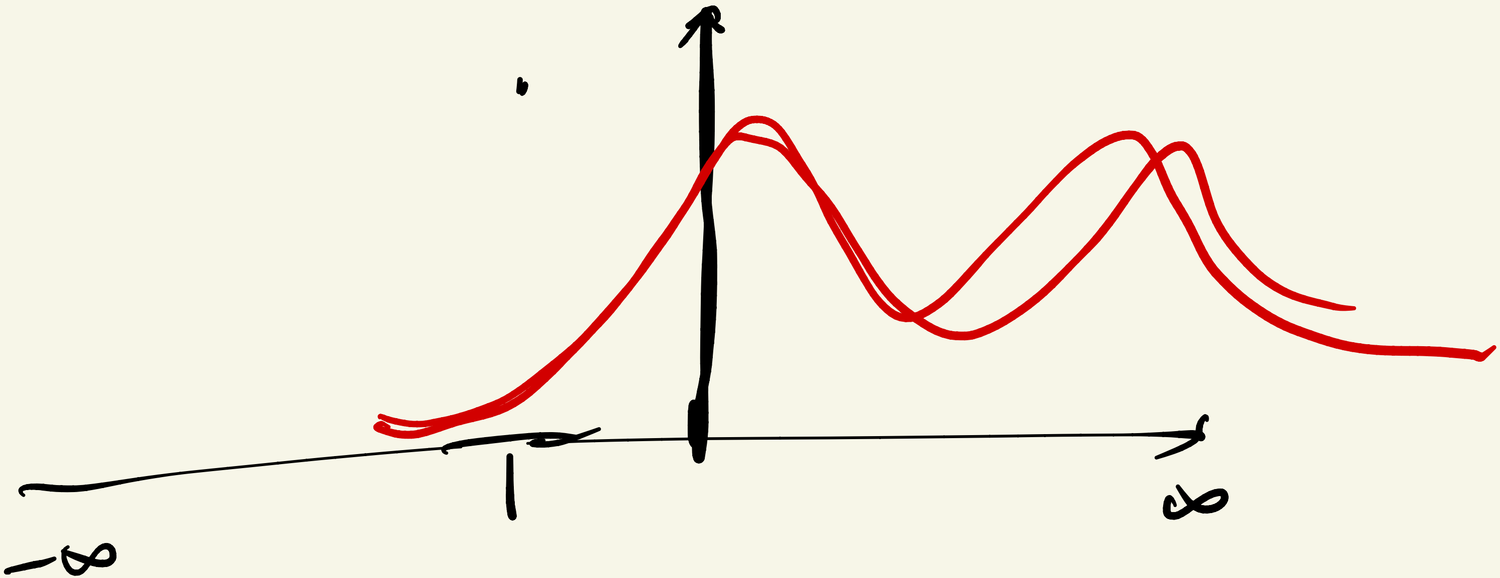
not random variable

PGM



$P(x_1, x_2 \ldots x_5)$

random variable vs. deterministic



Probabilistic Graphical Model is a graphical language to express conditional independence

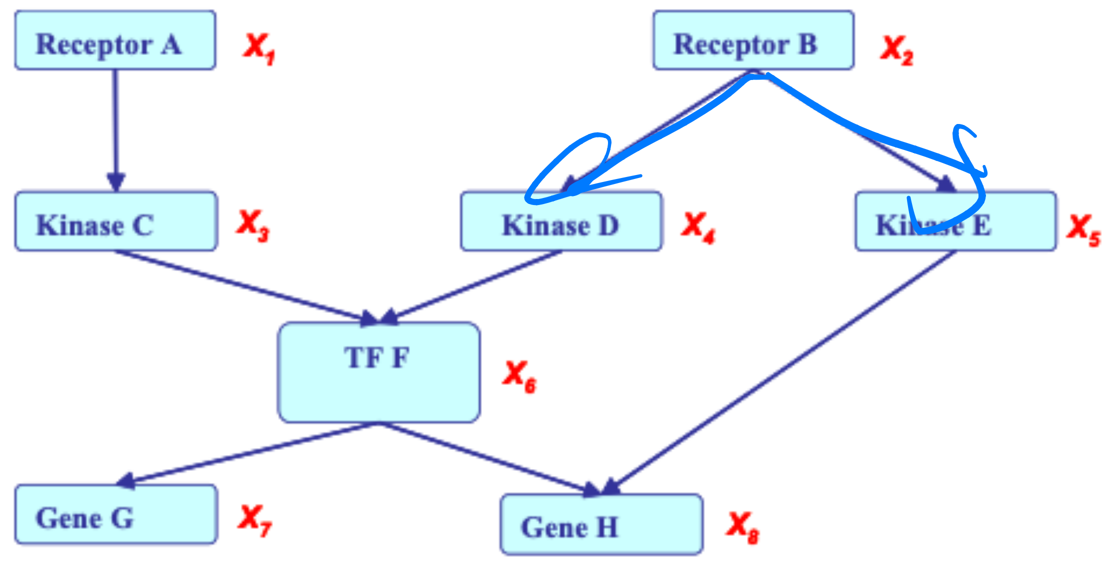
Two types of Graphical Models

Two types of Graphical Models

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

Two types of Graphical Models

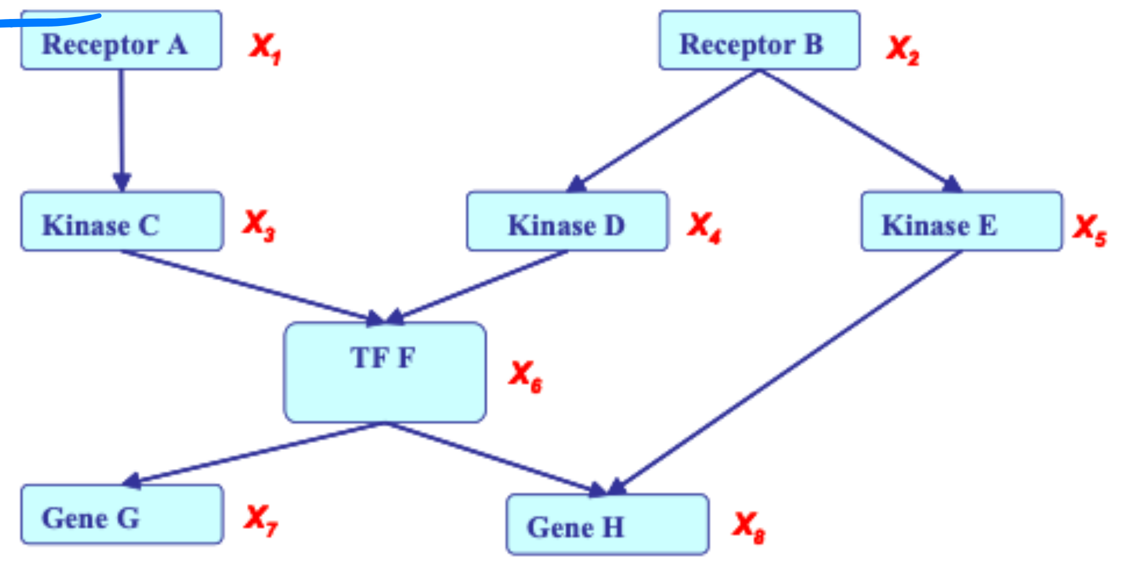
- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):



Two types of Graphical Models

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

DGM



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

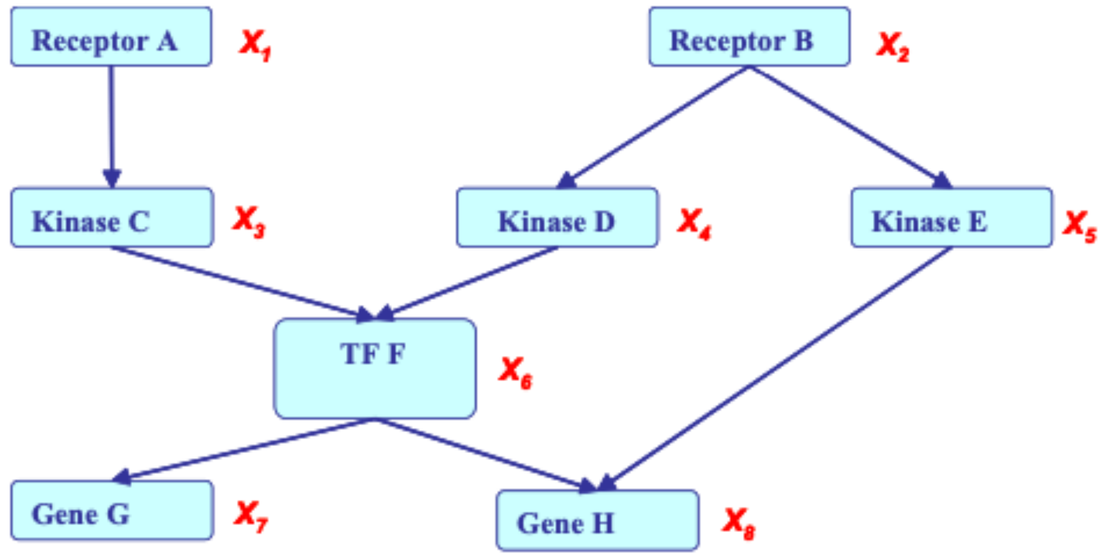
$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

Two types of Graphical Models

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

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

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



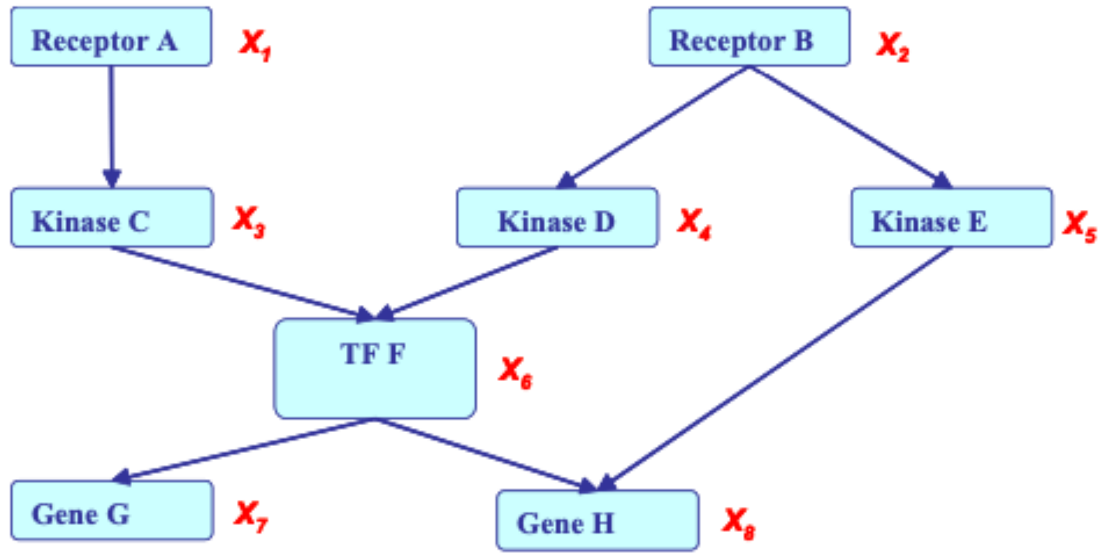
- Undirected edges simply give correlations between variables (**Markov Random Field** or **Undirected Graphical model**):

Two types of Graphical Models

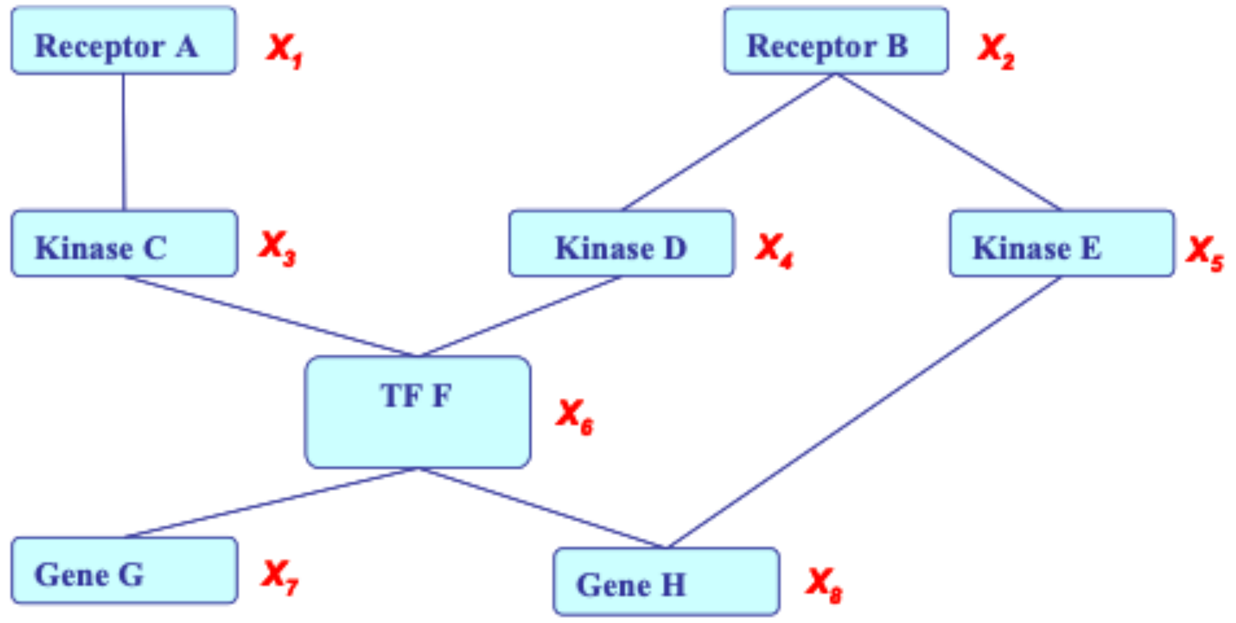
- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):

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

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



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



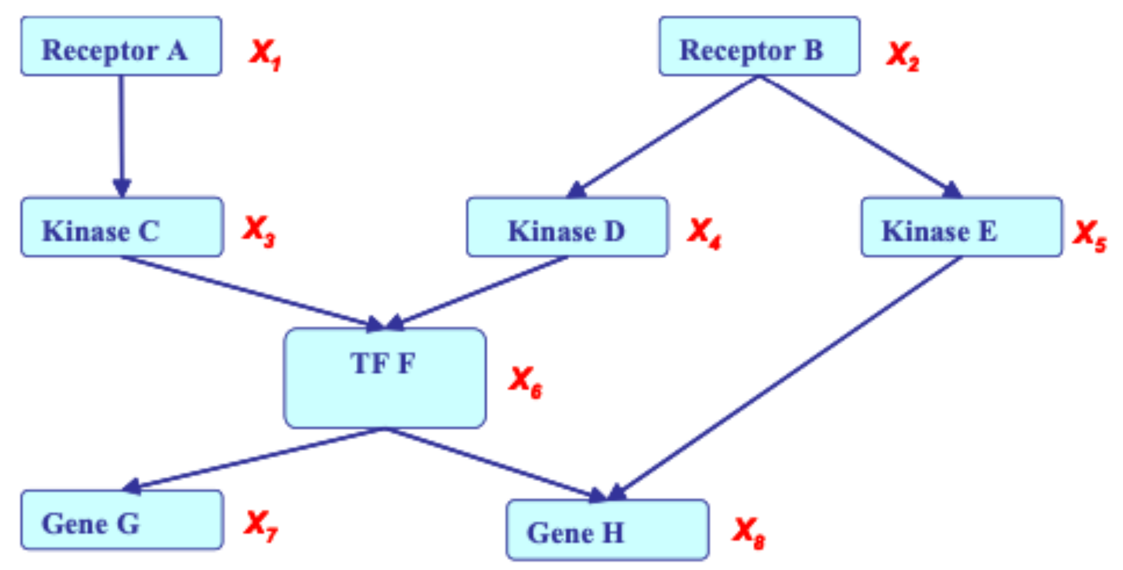
Two types of Graphical Models

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

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

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

DGM



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

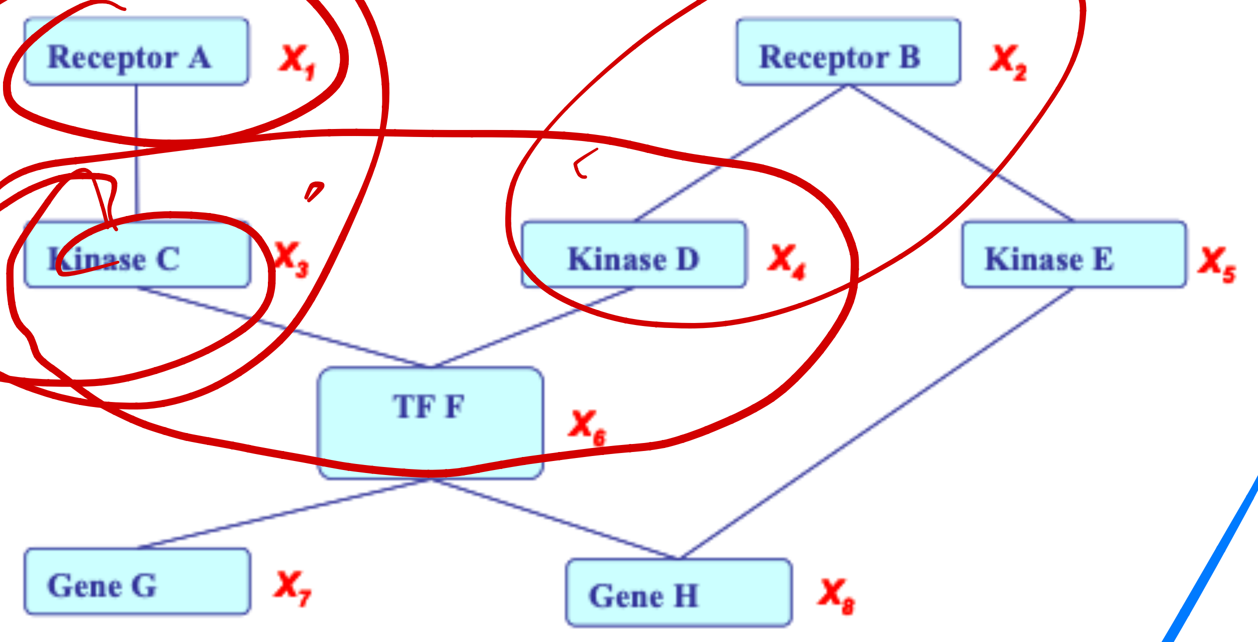
energy model

$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \propto \exp(E(x_i)) + E(x_i)$$

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

Z

energy

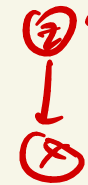


Ancestral Sampling

$P(x, z)$

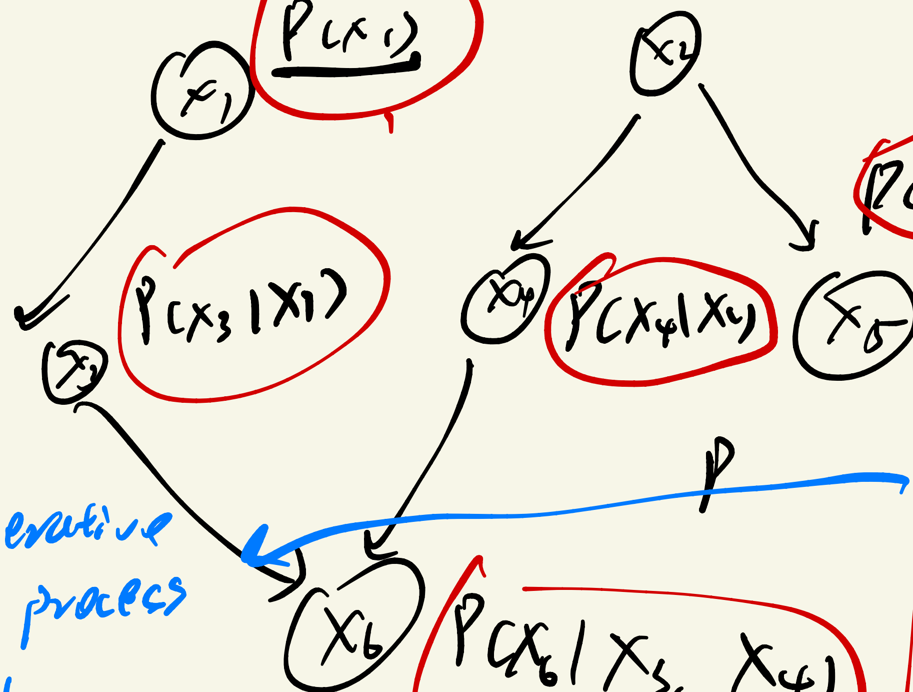
$P(x) = \sum P(x, z)$

$P(x_2)$



$P(z)$

$P(x_1, x_2 \ldots x_6)$



$x_1, x_2 \ldots x_6$

$x_1 \sim P(x_1)$

$x_2 \sim P(x_2)$

$x_3 \sim P(x_3 | x_1)$

$x_4 \sim P(x_4 | x_2, \ldots$

$x_6 \sim P(x_6 | x_3, x_4)$

generative process of x_6

$$P(x_1, \dots x_5 \mid x_6)$$

$$P(x_6) = \sum_{x_1 \dots x_5} P(x_1, \dots x_5, x_6)$$

$$P(x) = \sum_{z} P(x, z)$$

$$z = x_1, \dots x_5$$

$$P(x_6, x_5) = \sum_{x_1, \dots x_4} P(x_1, \dots x_6)$$

PGMs are Structural Specification of Probability Distribution

PGMs are Structural Specification of Probability Distribution

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

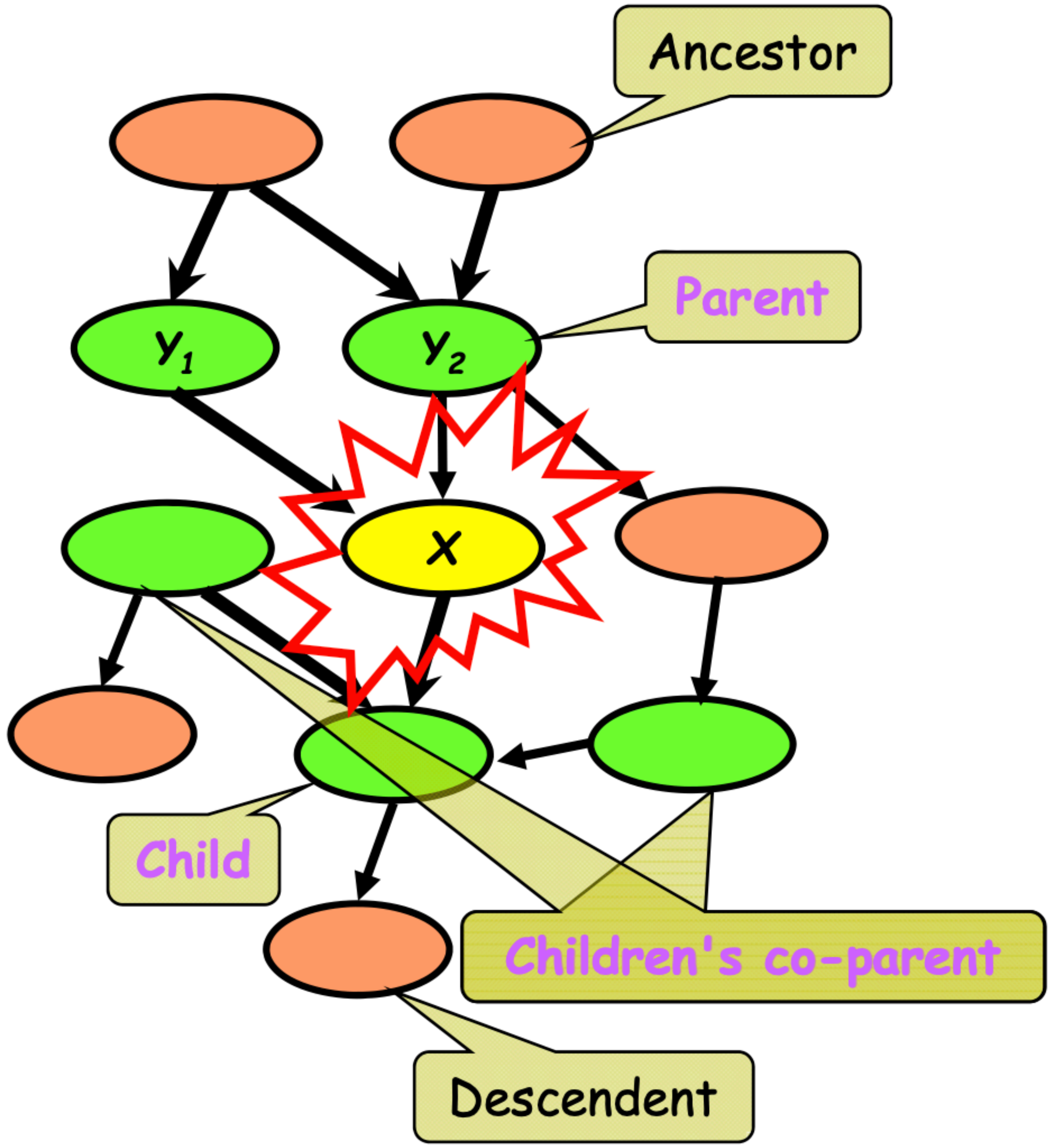
PGMs are Structural Specification of Probability Distribution

- 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

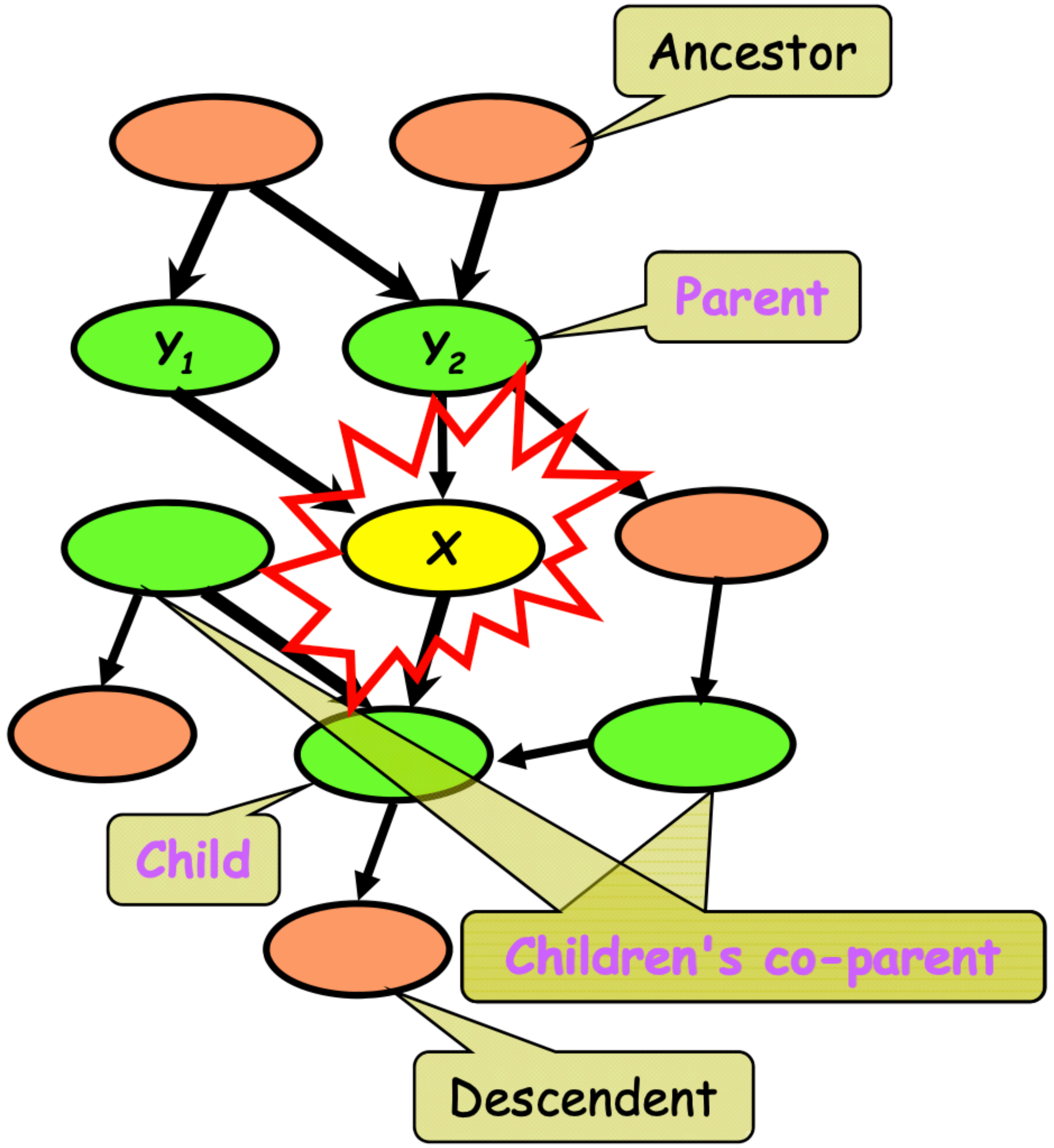
Markov Blanket for Directed Acyclic Graph (DAG)

Markov Blanket for Directed Acyclic Graph (DAG)



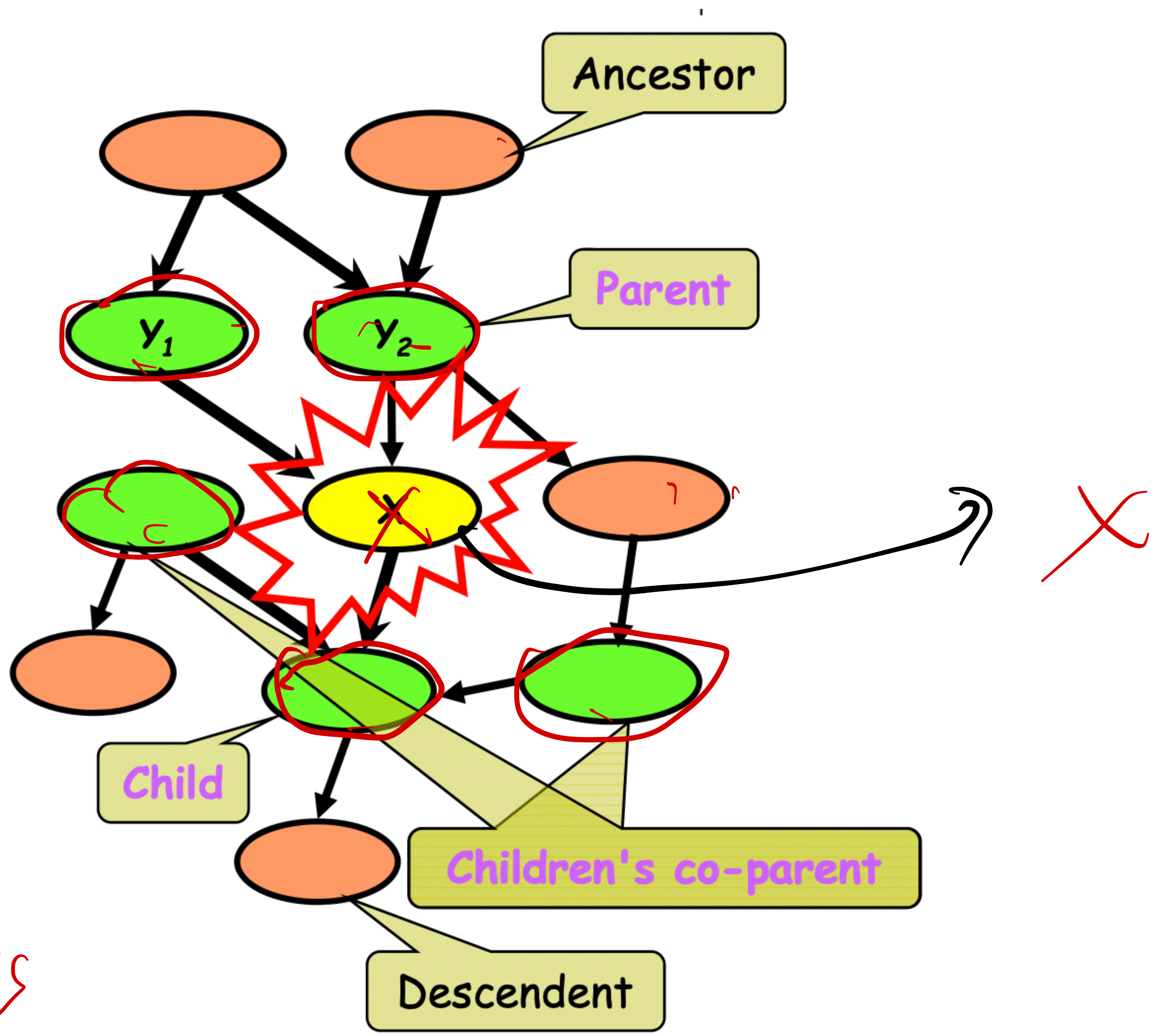
Markov Blanket for Directed Acyclic Graph (DAG)

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



Markov Blanket for Directed Acyclic Graph (DAG)

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



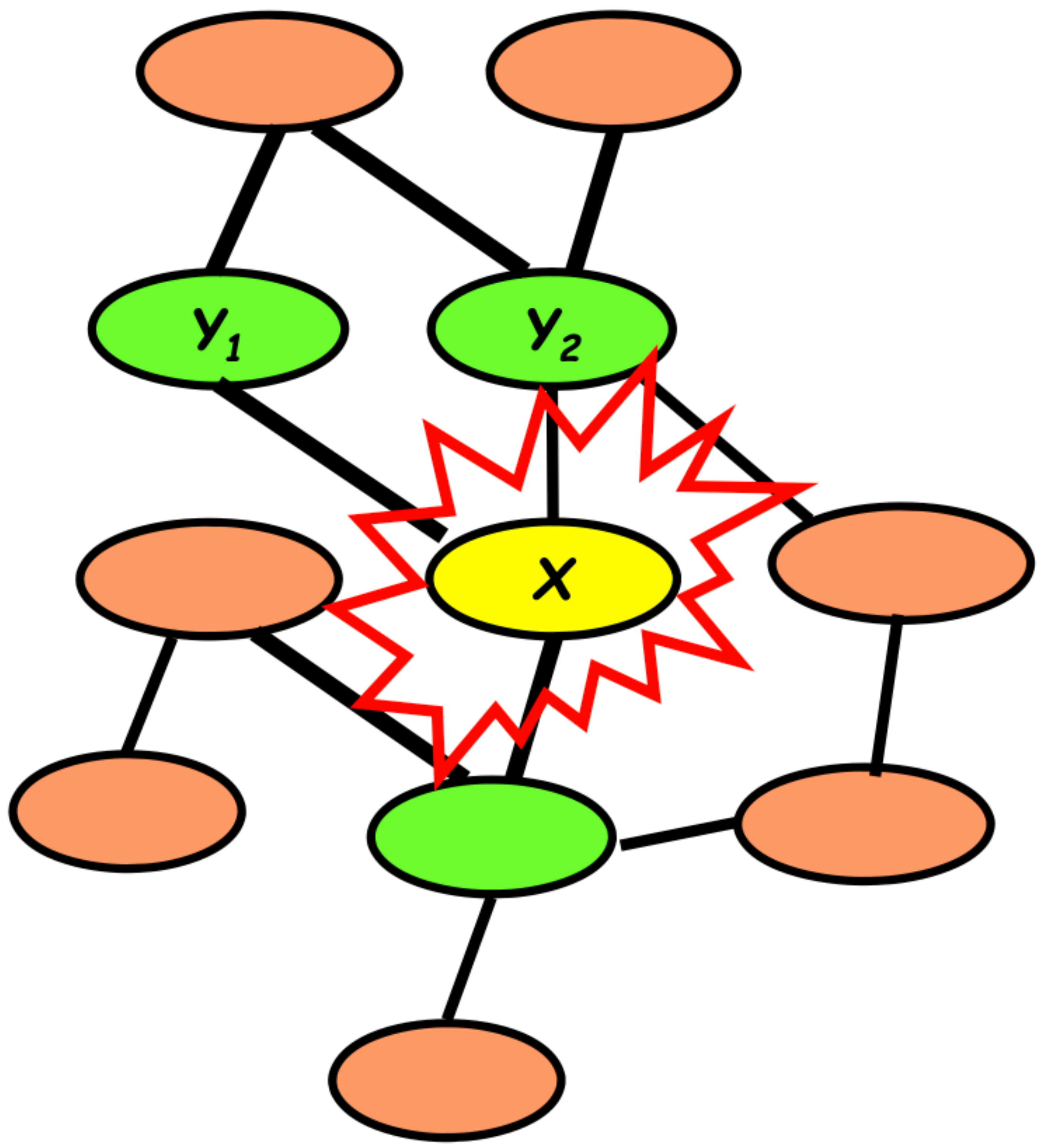
$X \perp$ all other nodes

condition: green ones

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

Conditional Independence of Undirected Graph

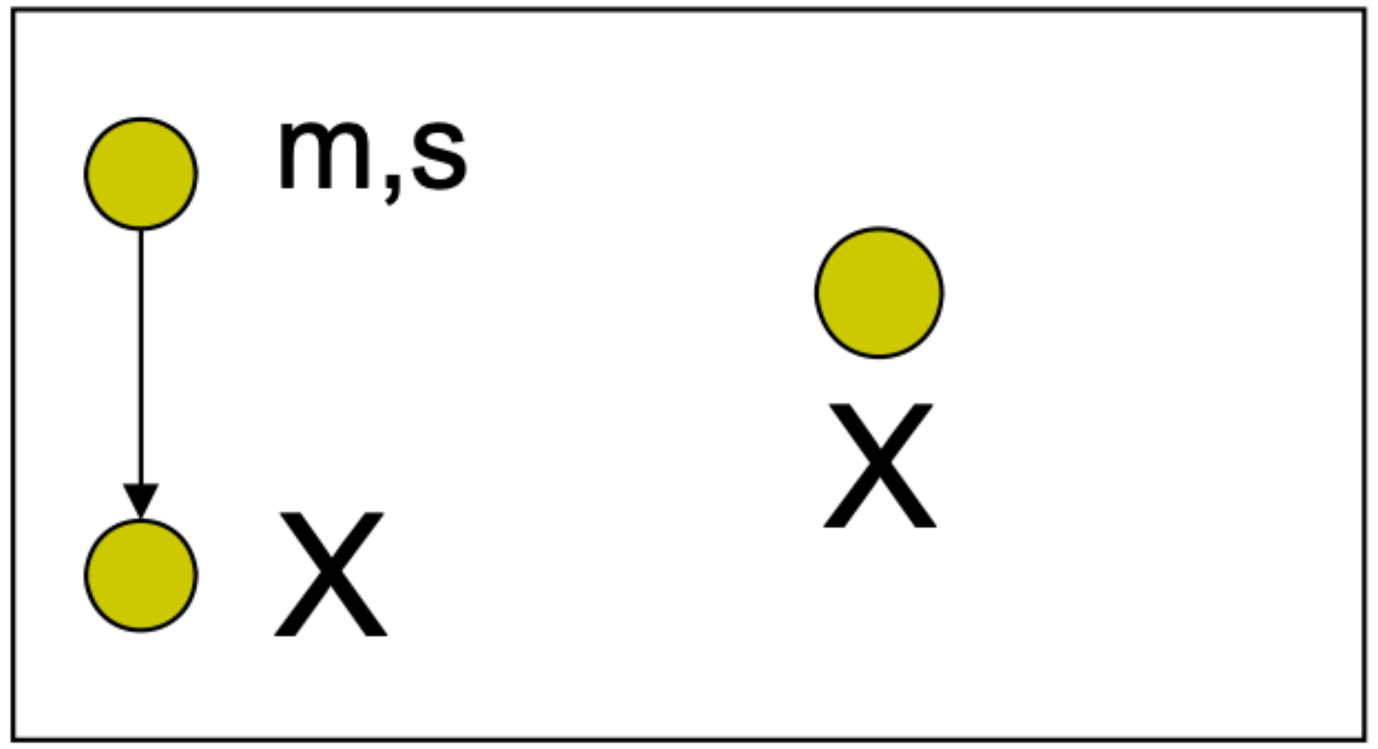
- Meaning: a node is **conditionally independent** of every other node in the network given its **Directed neighbors**



GMs are your old friends

Probabilistic Graphical Model is a language to express distributions

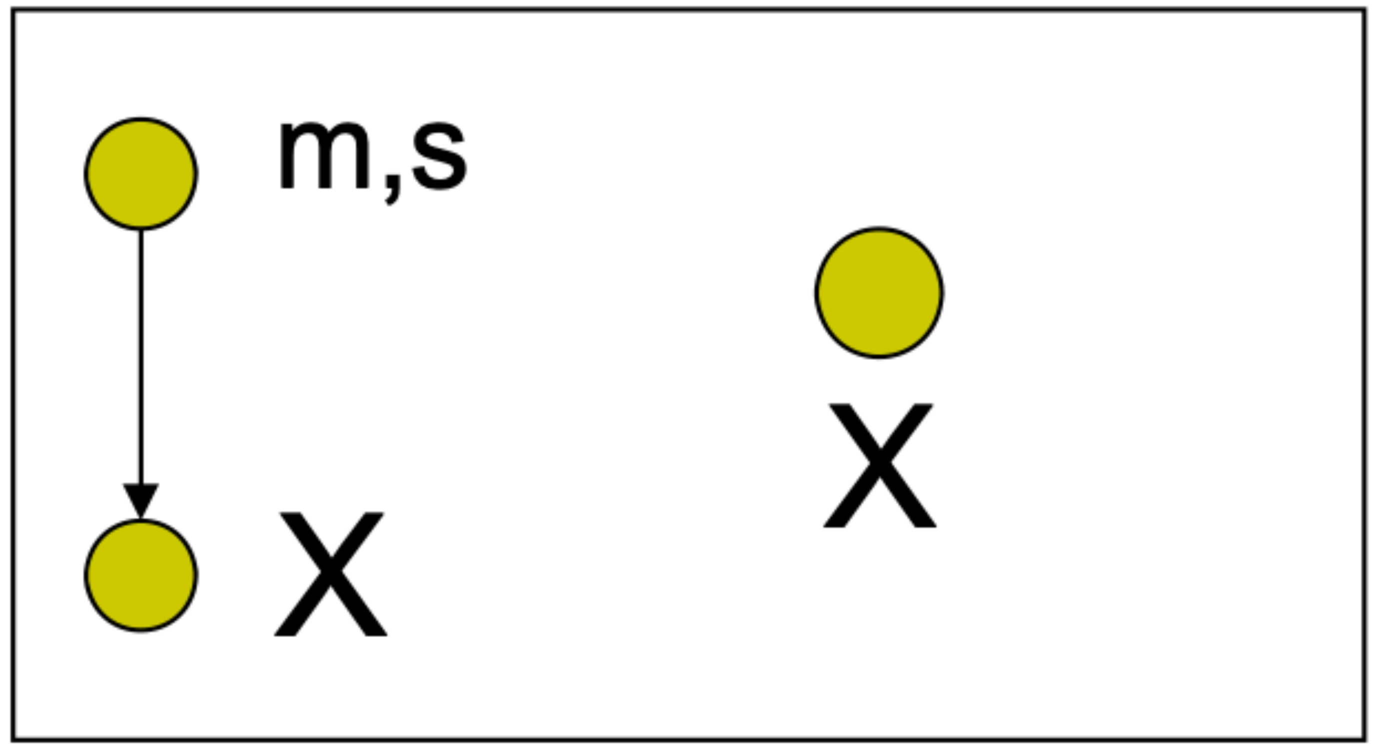
GMs are your old friends



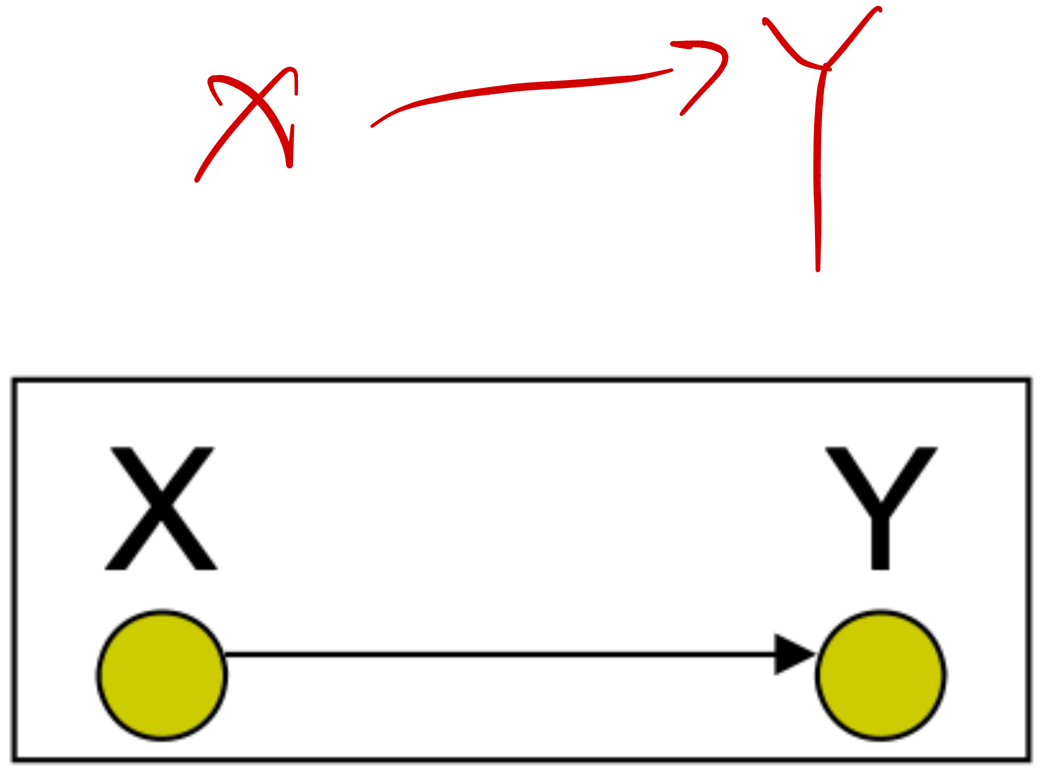
P(x)

Probabilistic Graphical Model is a language to express distributions

GMs are your old friends



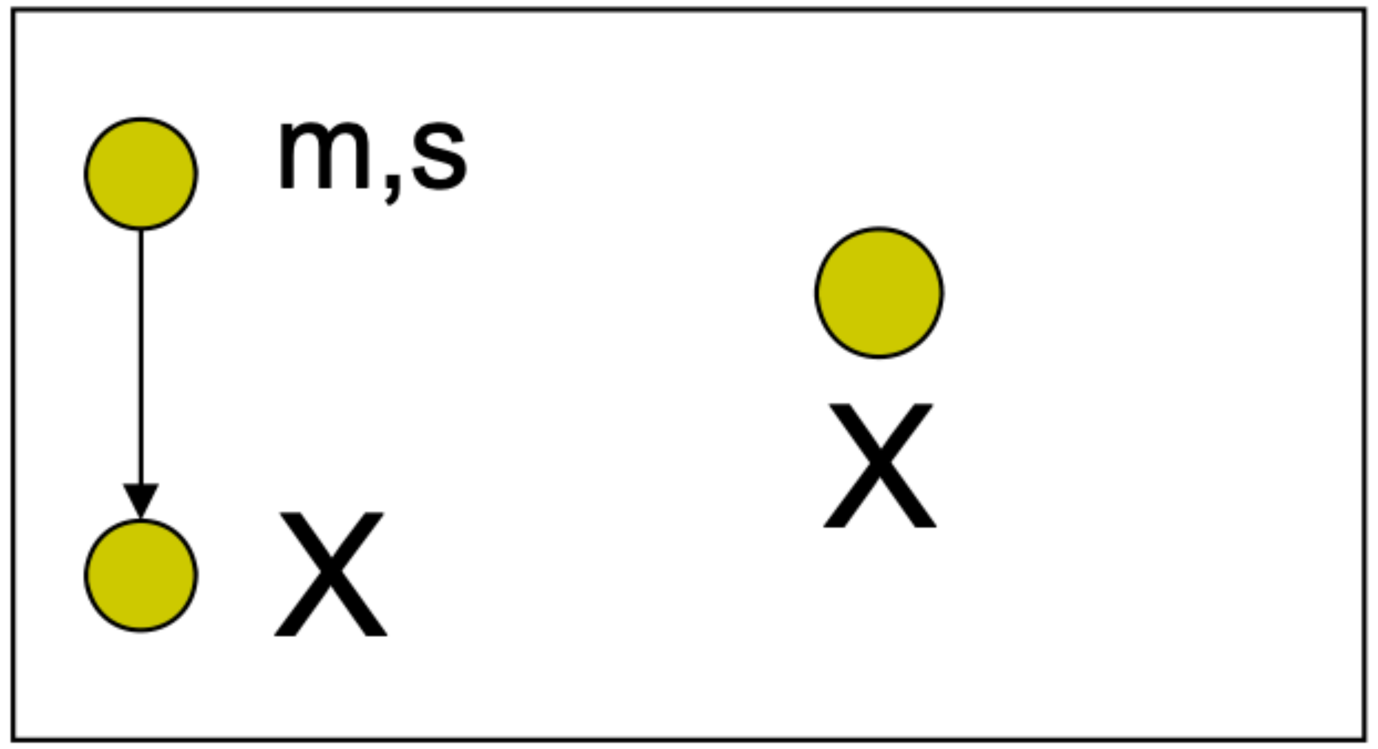
P(x)



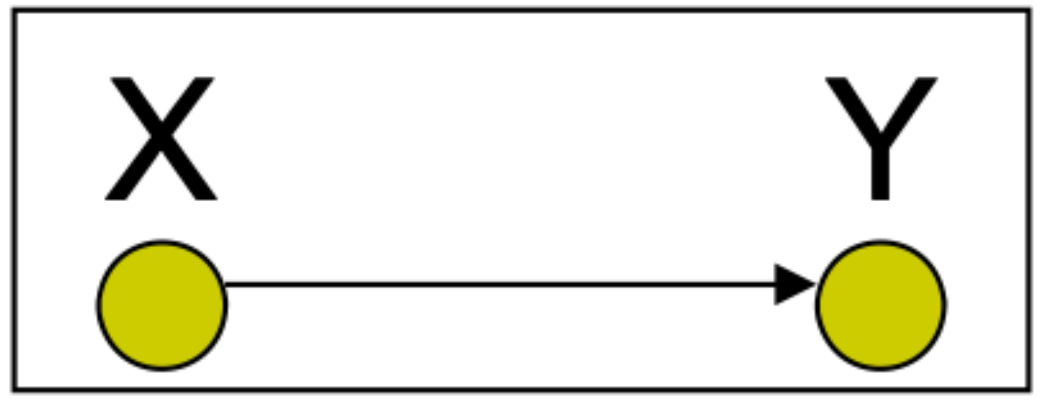
Regression, classification

Probabilistic Graphical Model is a language to express distributions

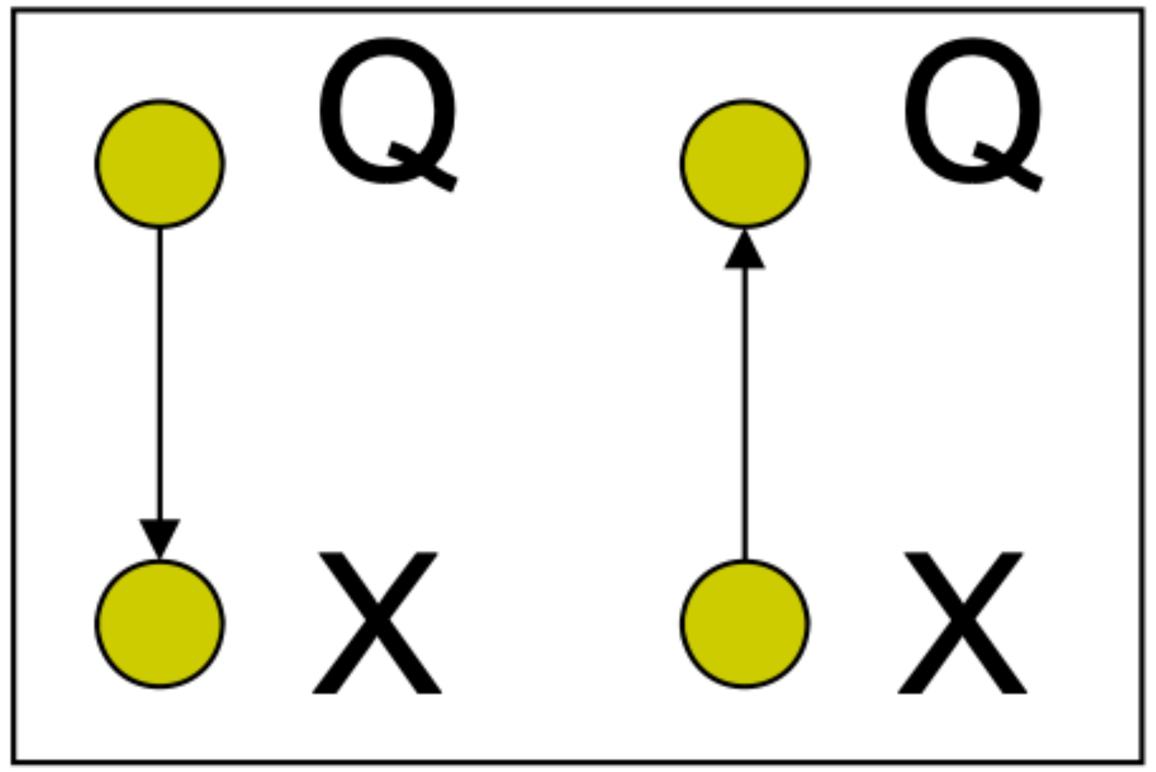
GMs are your old friends



P(x)



Regression, classification

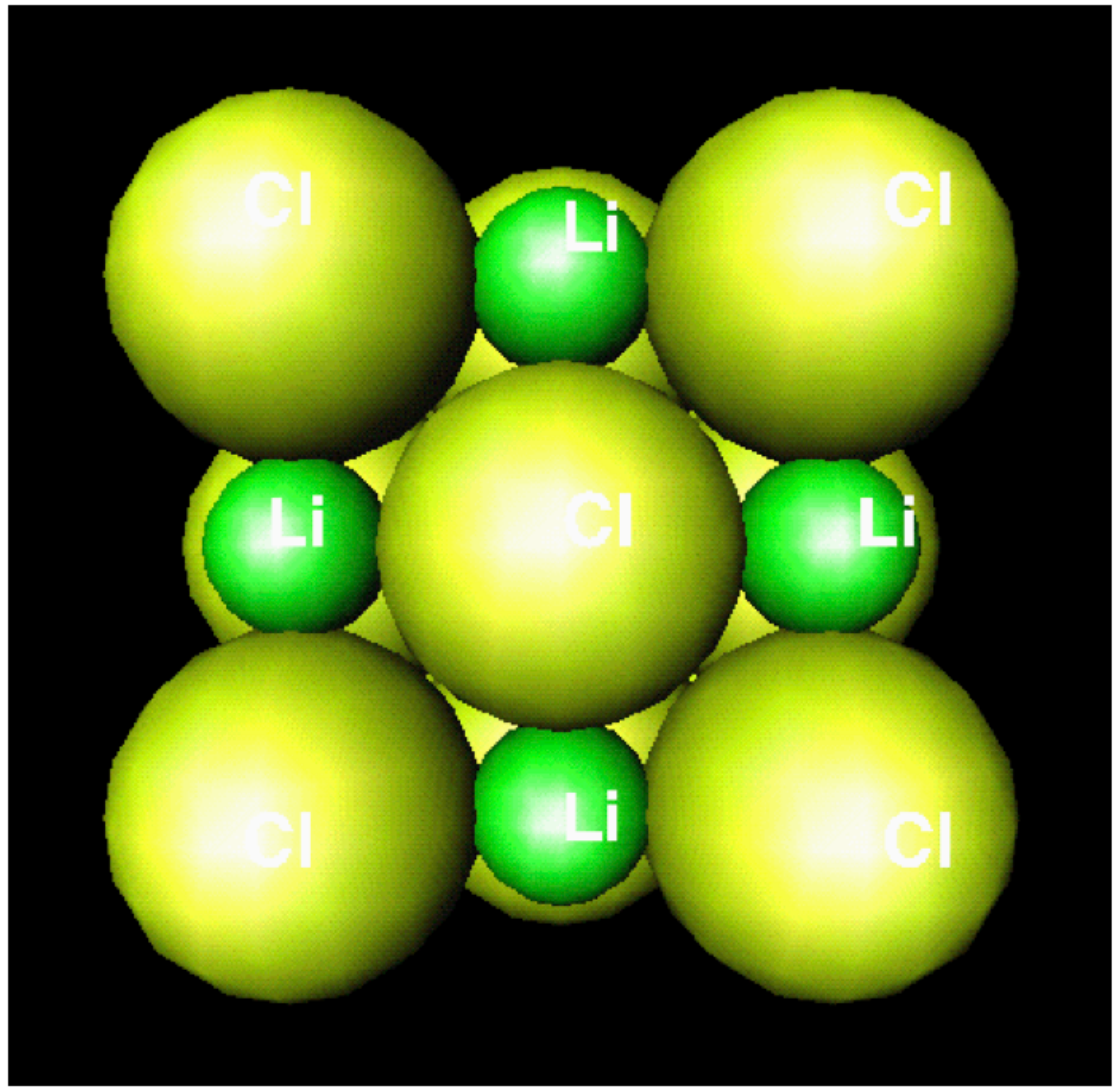


Generative vs
Discriminative Classification

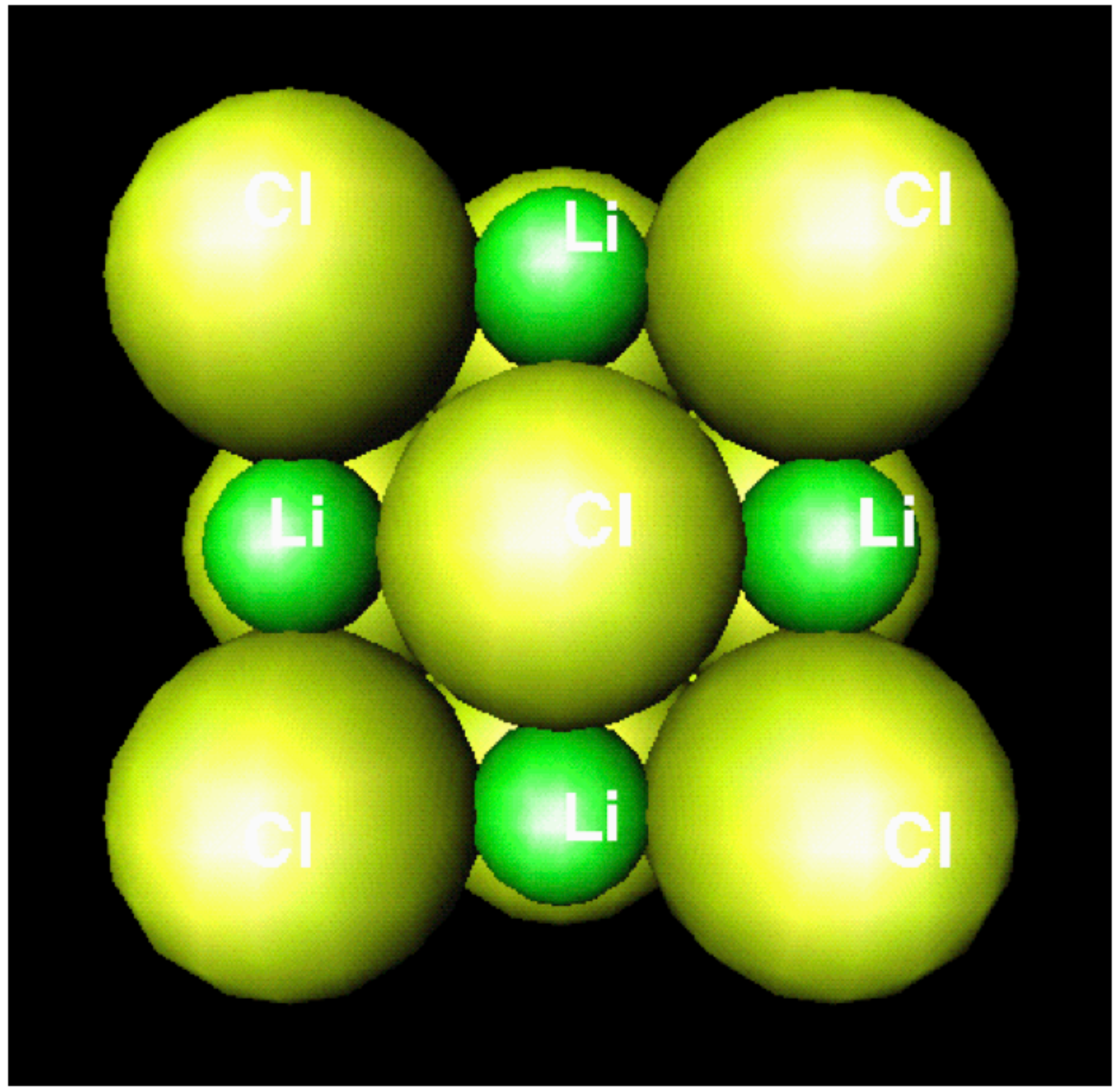
Probabilistic Graphical Model is a language to express distributions

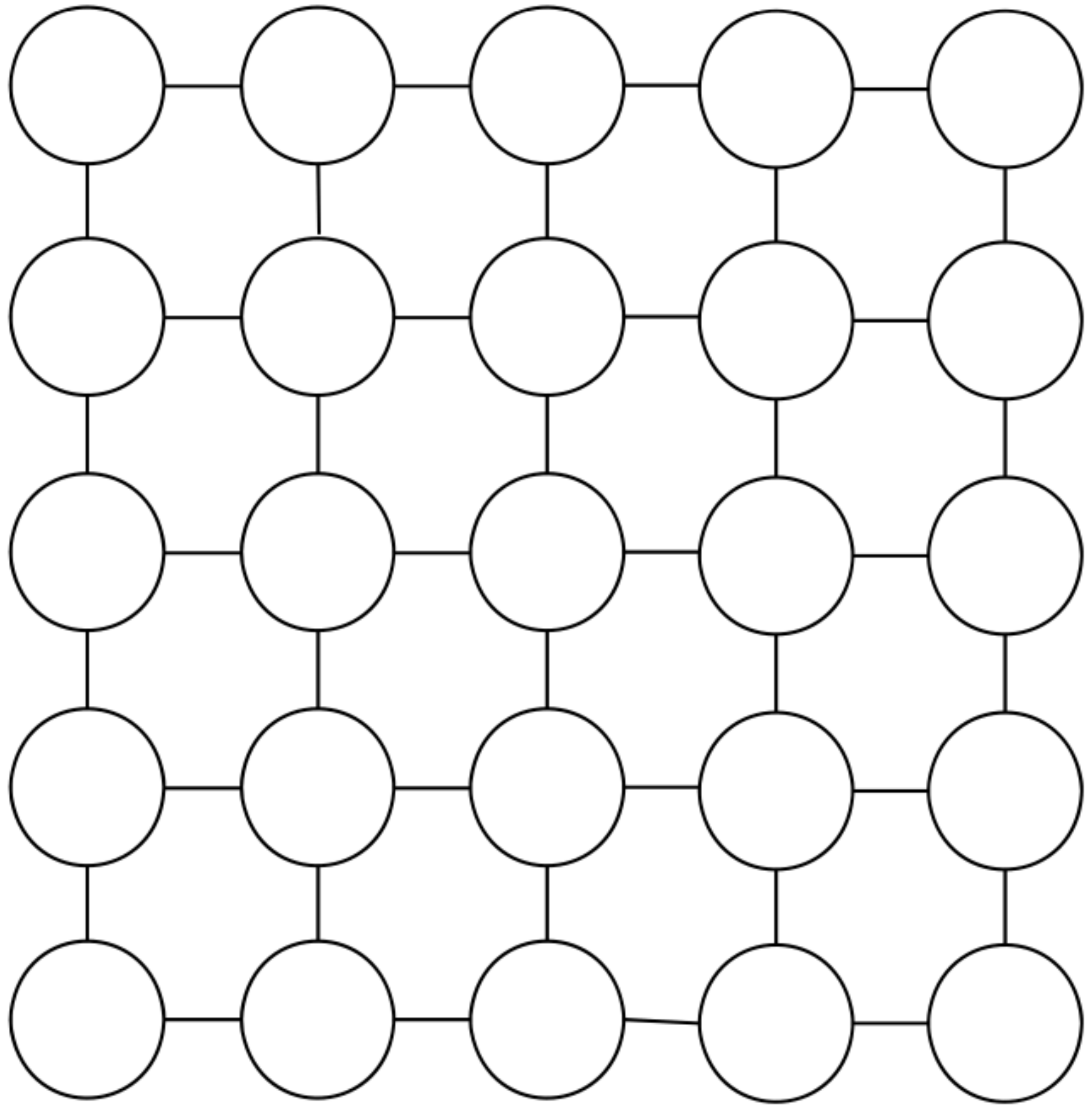
Fancier GMs: Solid State Physics

Fancier GMs: Solid State Physics



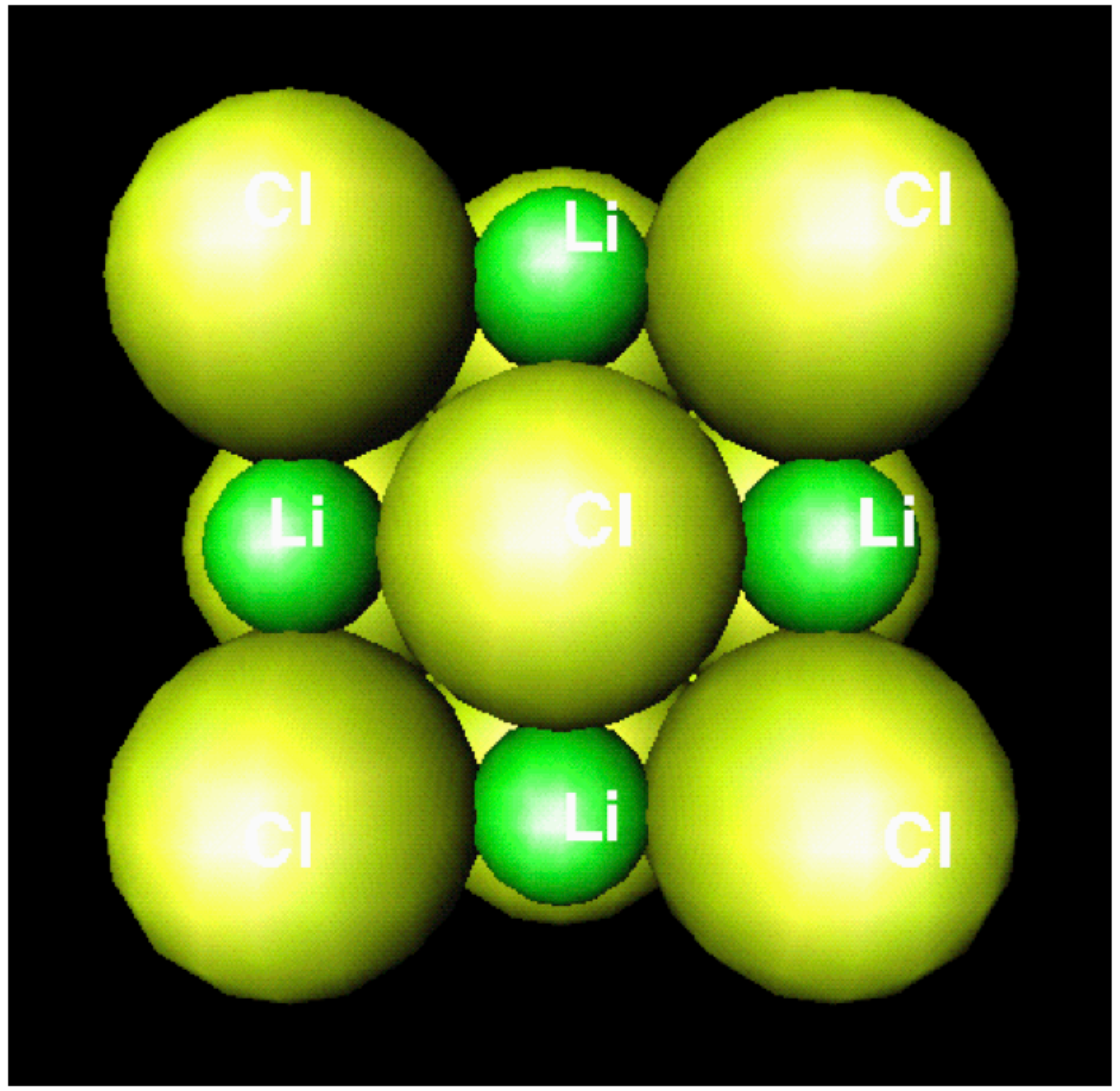
Fancier GMs: Solid State Physics

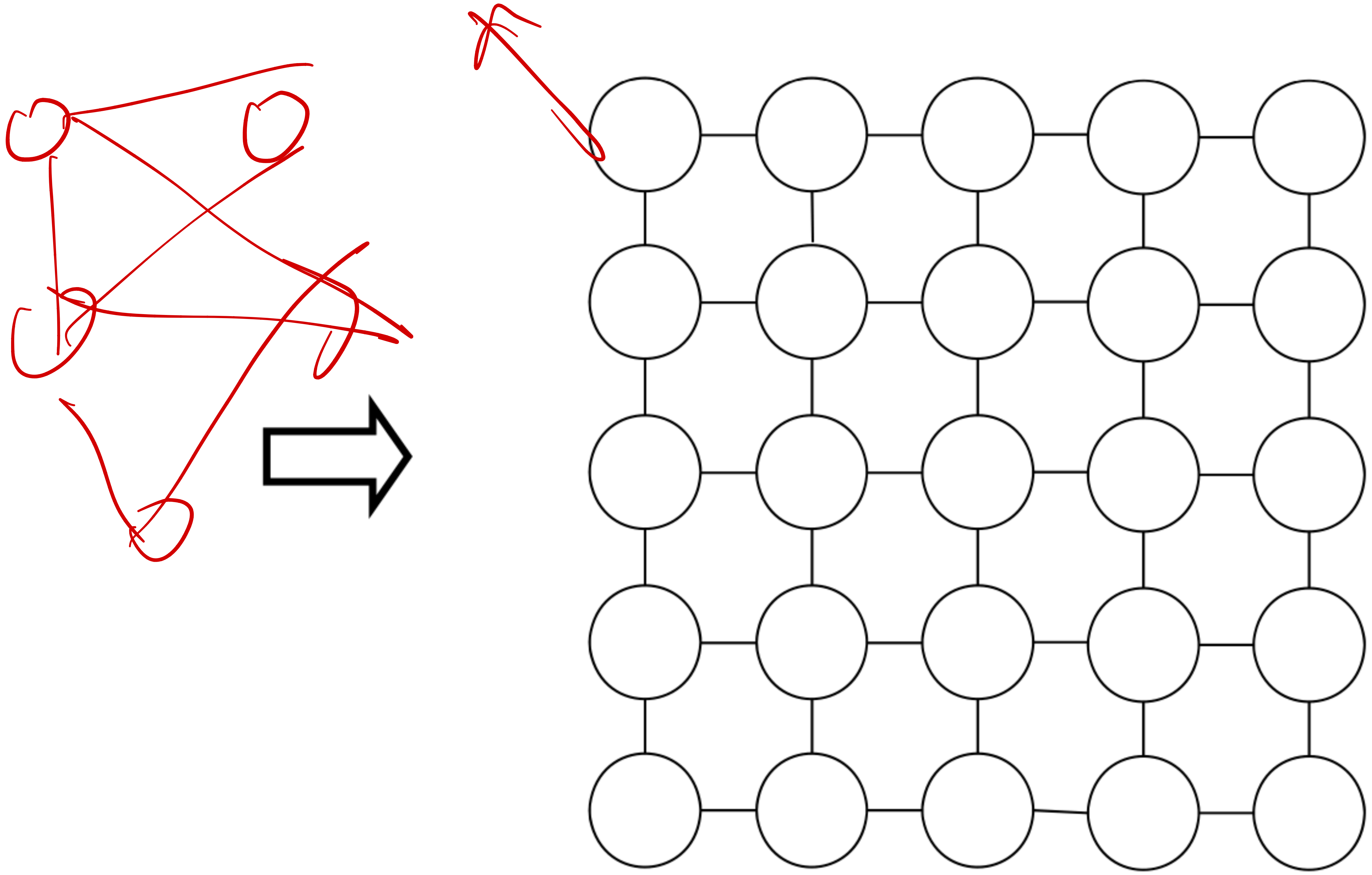




Ising/Potts model

Fancier GMs: Solid State Physics





Ising/Potts model

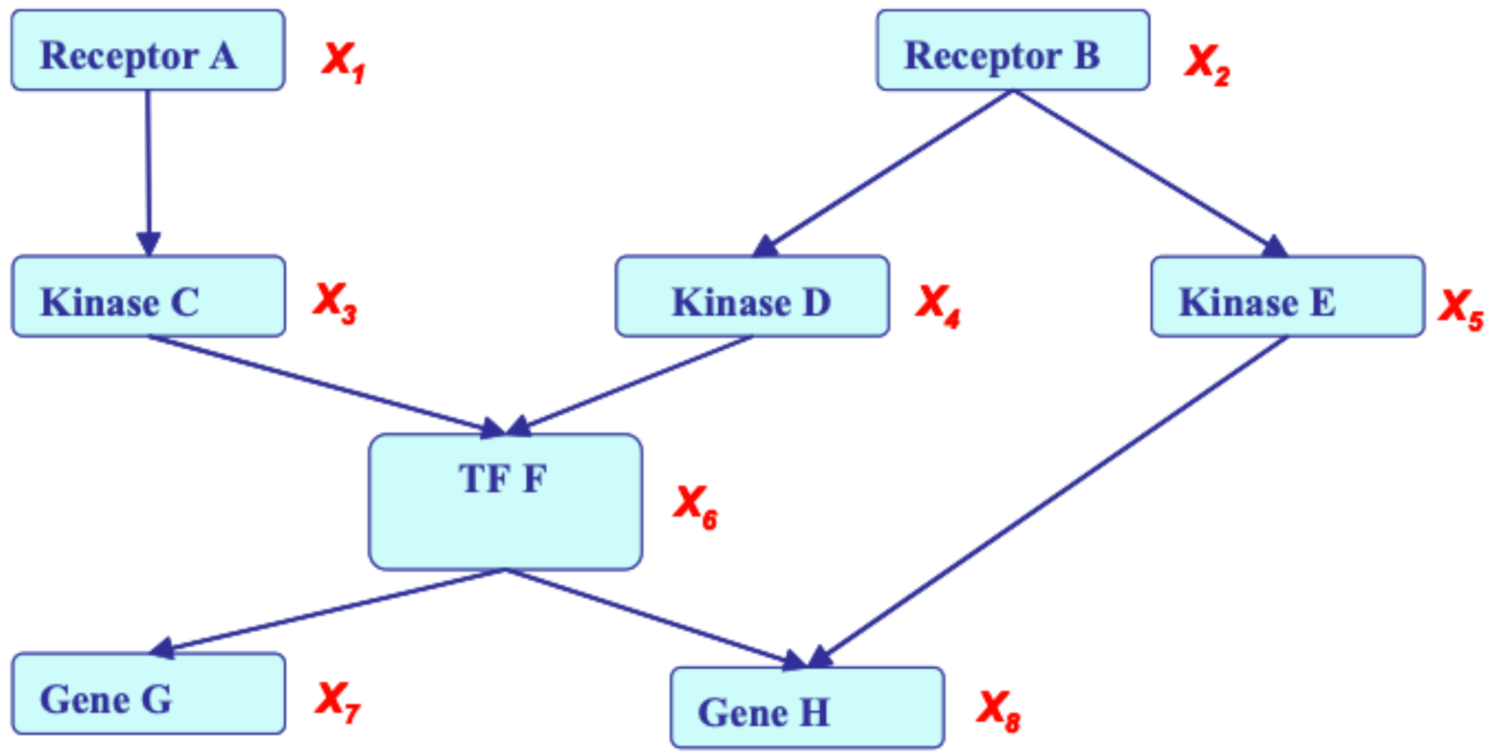
Define the strengths/correlation between different atoms

Why Graphical Models

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

How to Factor a Distribution Given a DAG

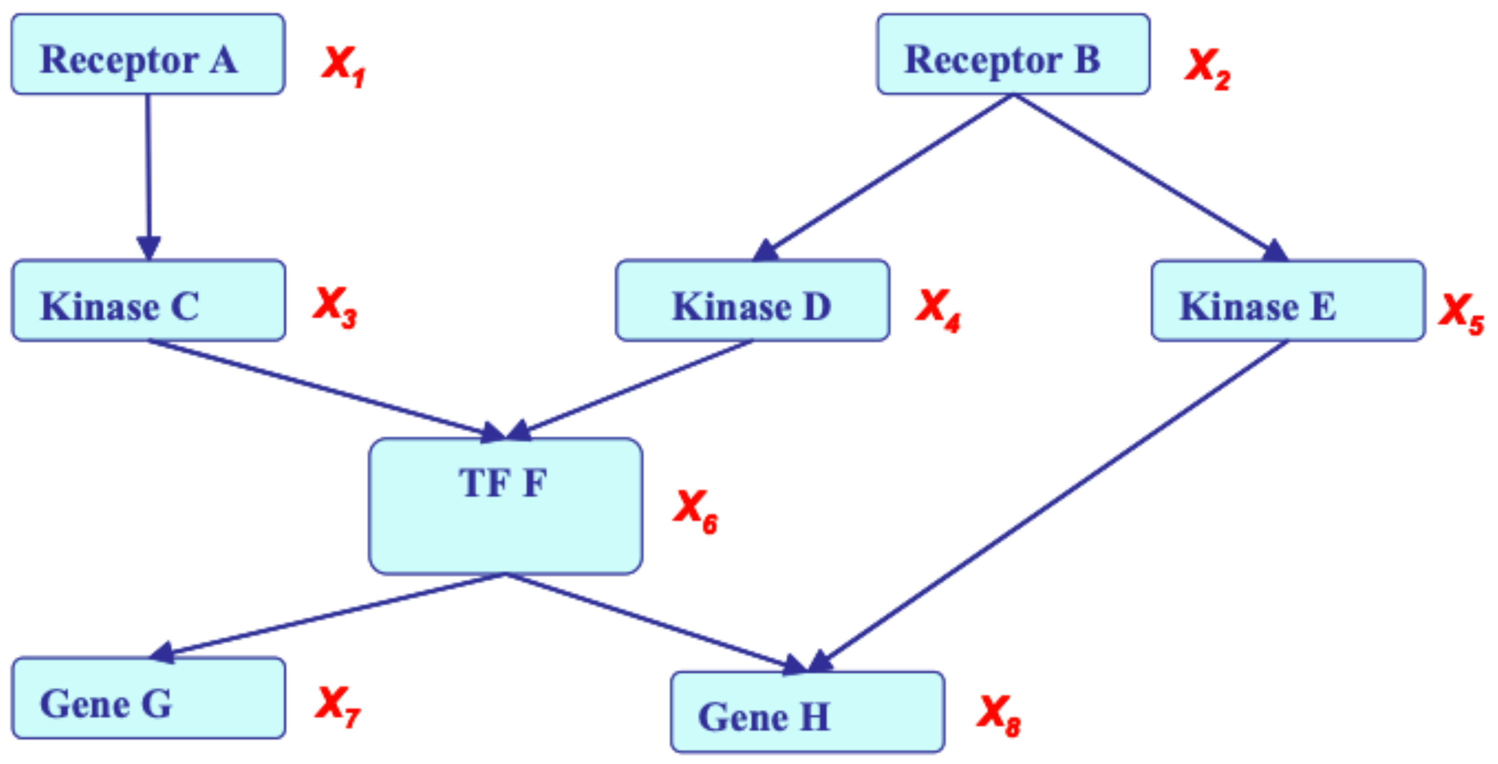
How to Factor a Distribution Given a DAG



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

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

How to Factor a Distribution Given a DAG



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

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

$P(x|$

- **Theorem:**

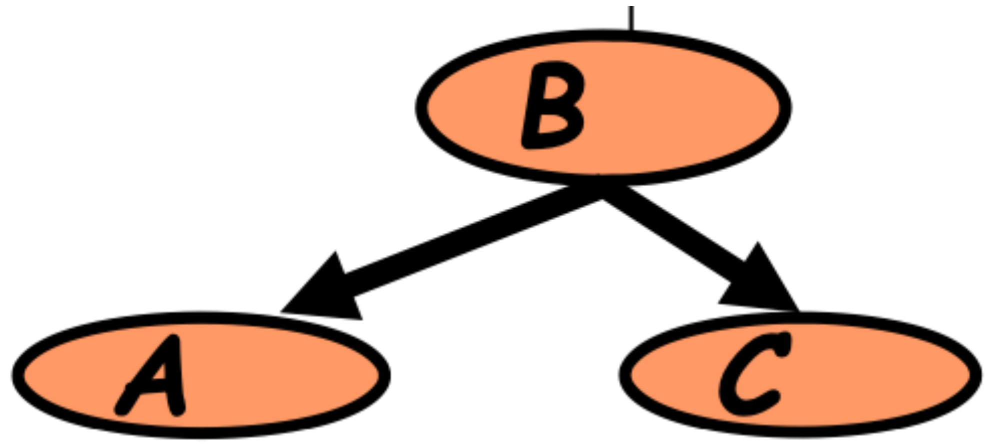
  Given a DAG, The most general form of the probability distribution that is consistent with the (probabilistic independence properties encoded in the) graph factors according to "node given its parents":

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

  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

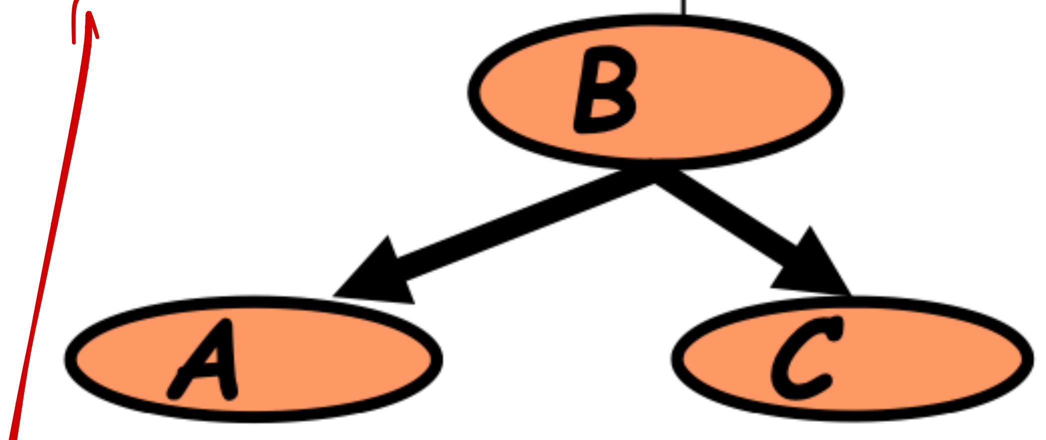
Local Structures & Independence

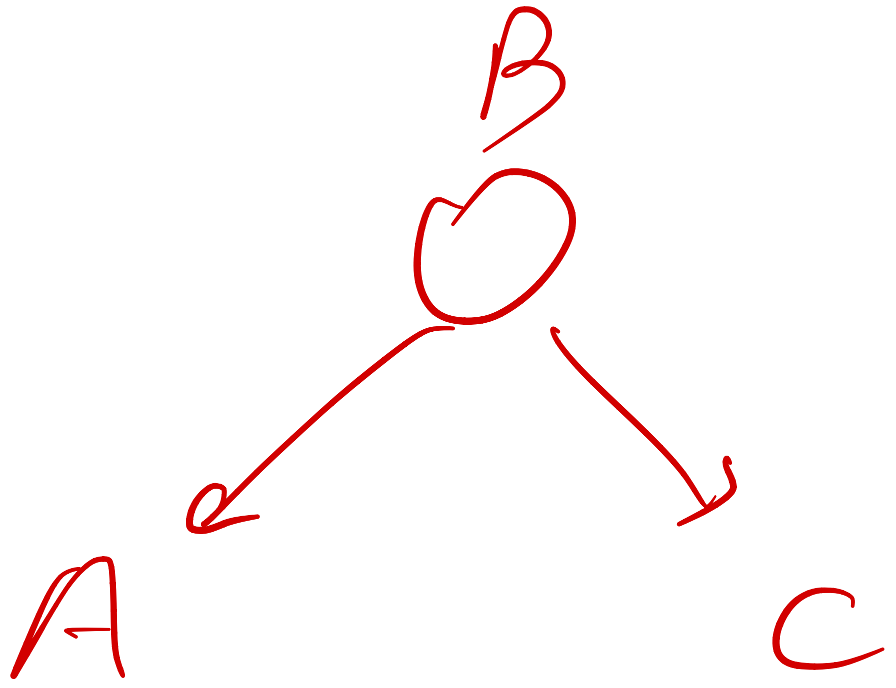


Local Structures & Independence

- Common parent
  - Fixing B decouples A and C ,
    "given the level of gene B, the levels of A and C are independent"

$A \perp C$ given B

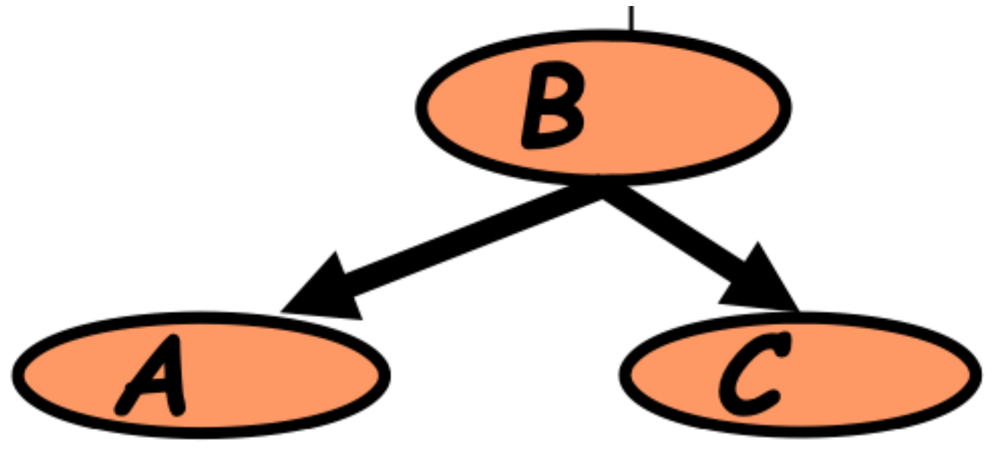


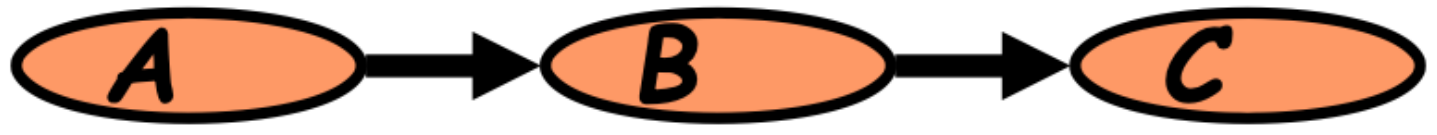


$A \perp C$ given B

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"

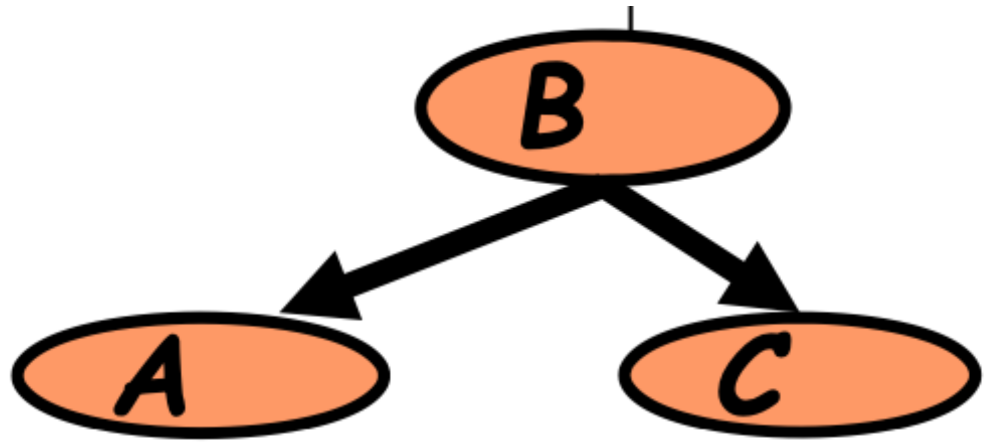




Local Structures & Independence

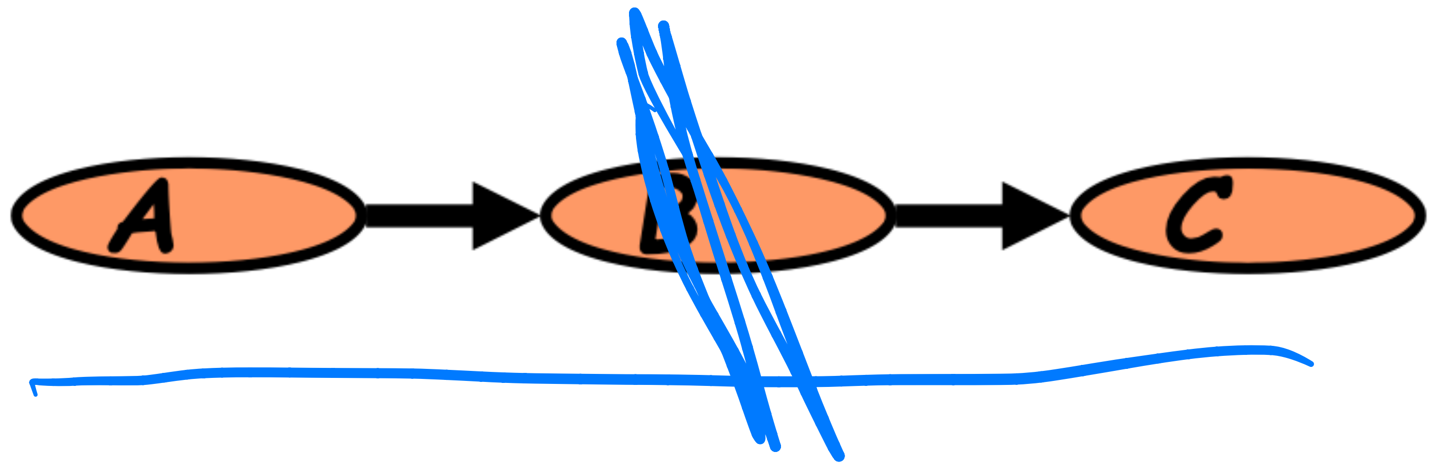
- Common parent
  - Fixing B decouples A and C

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



- Cascade
  - Knowing B decouples A and C

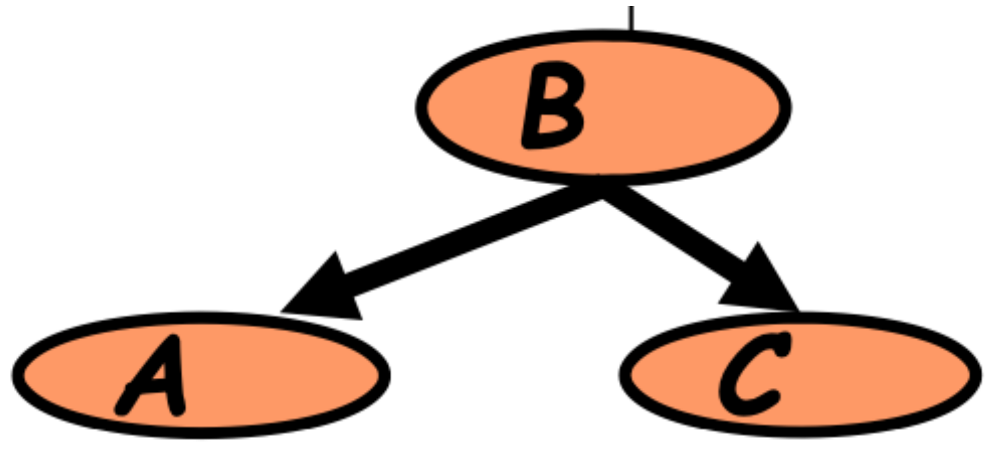
    "given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"



Local Structures & Independence

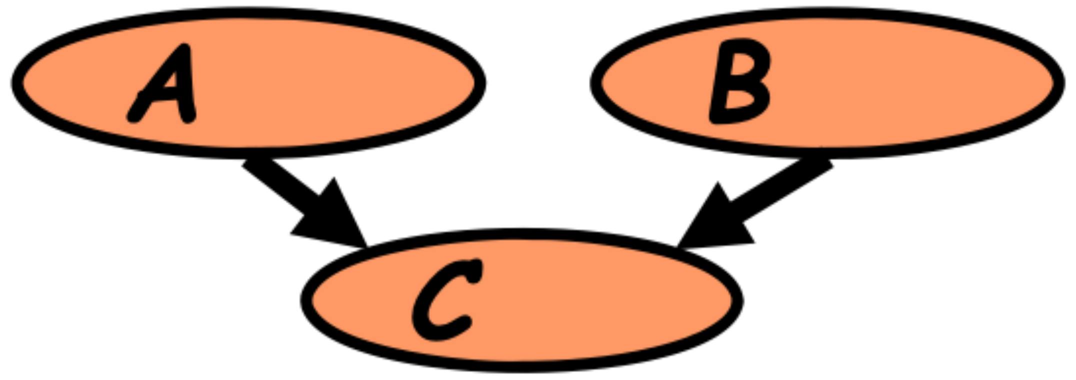
- Common parent
  - Fixing B decouples A and C

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



- Cascade
  - Knowing B decouples A and C

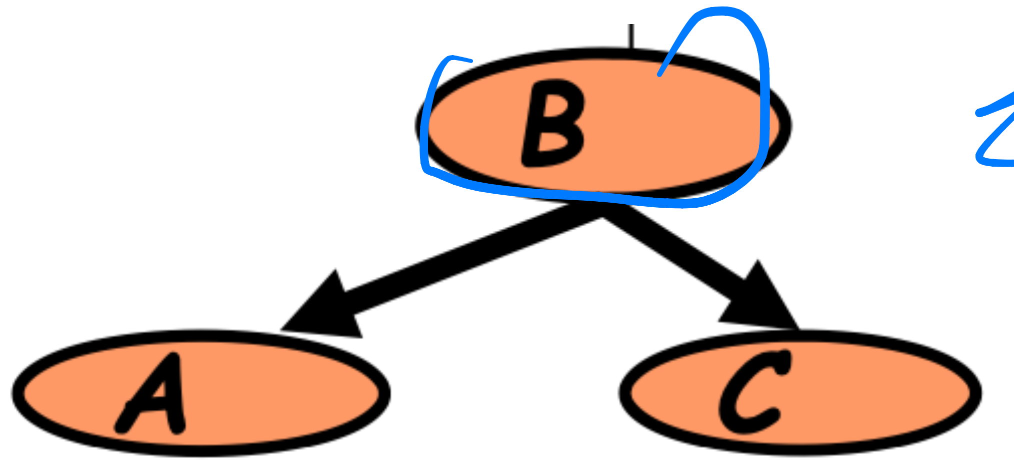
    "given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"



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

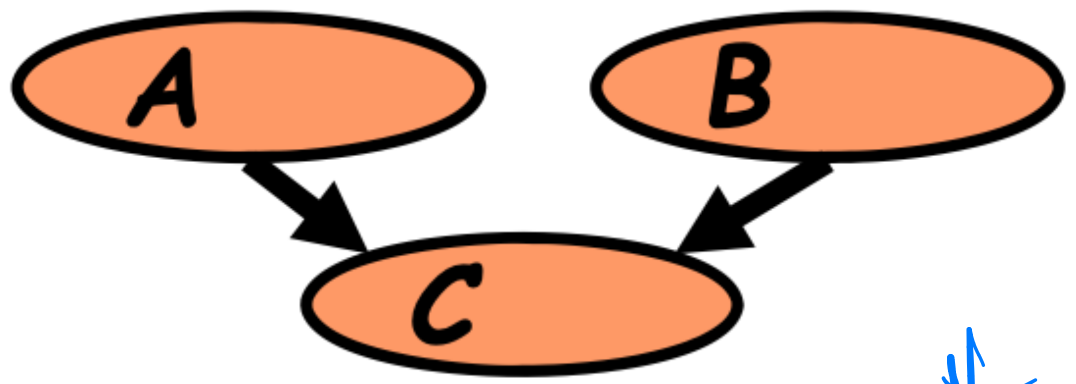
    "given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"

$A \perp B$

- V-structure
  - Knowing C couples A and B

    because A can "explain away" B w.r.t. C

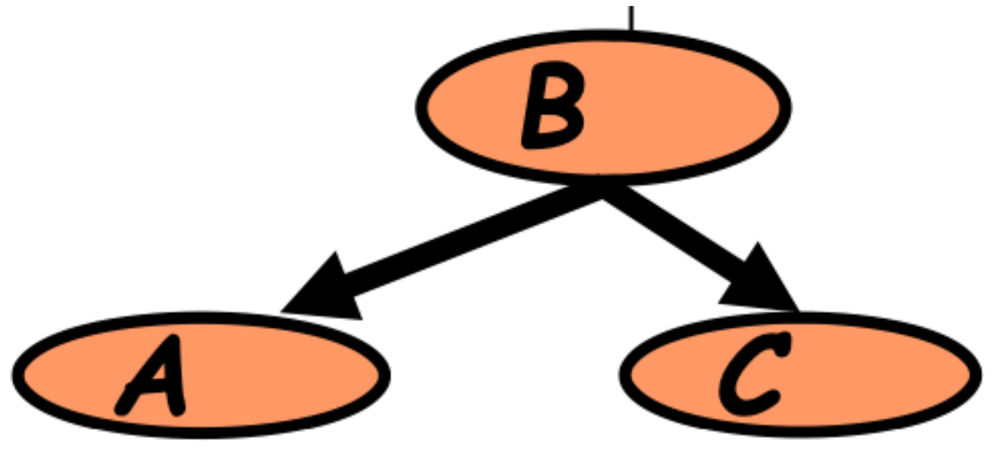
    "If A correlates to C, then chance for B to also correlate to B will decrease"



Local Structures & Independence

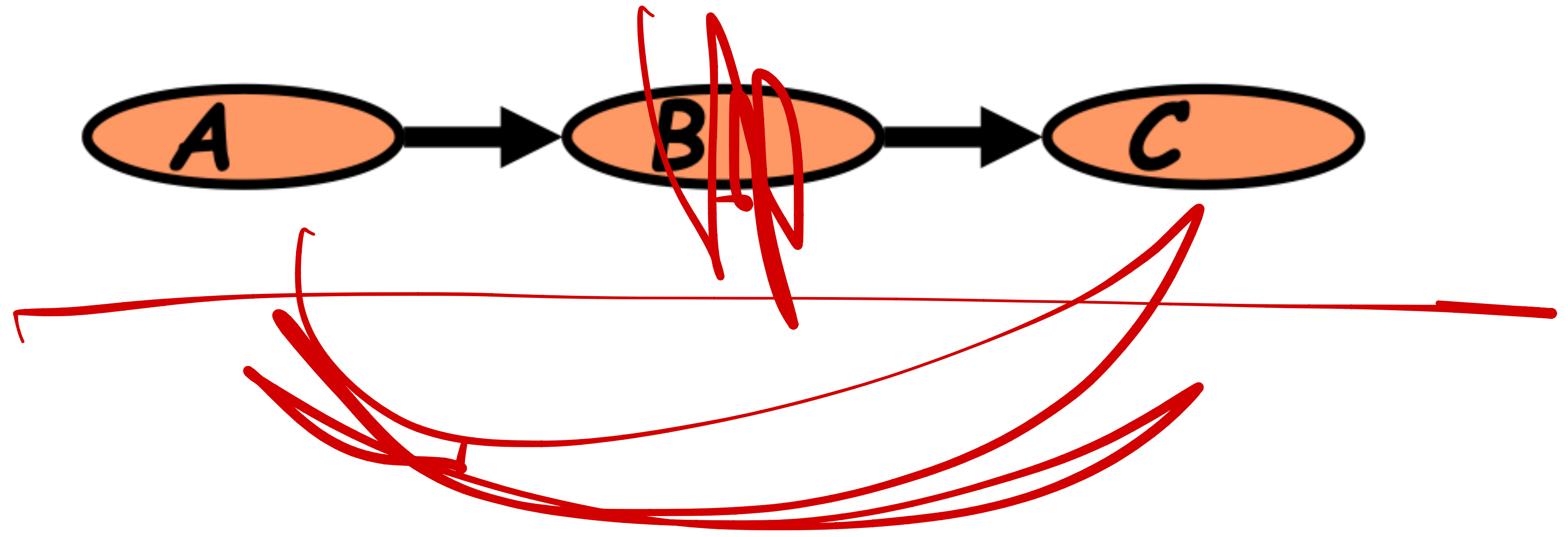
- Common parent
  - Fixing B decouples A and C

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



- Cascade
  - Knowing B decouples A and C

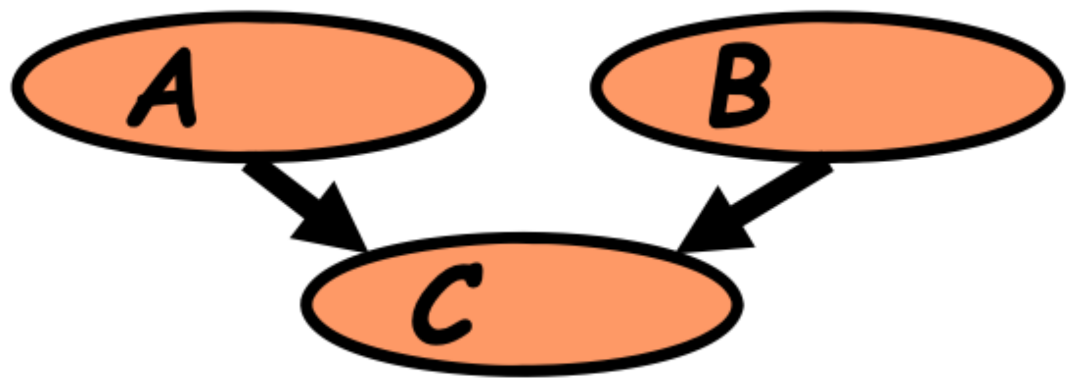
    "given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"



- 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

Global Markov Properties of DAGs

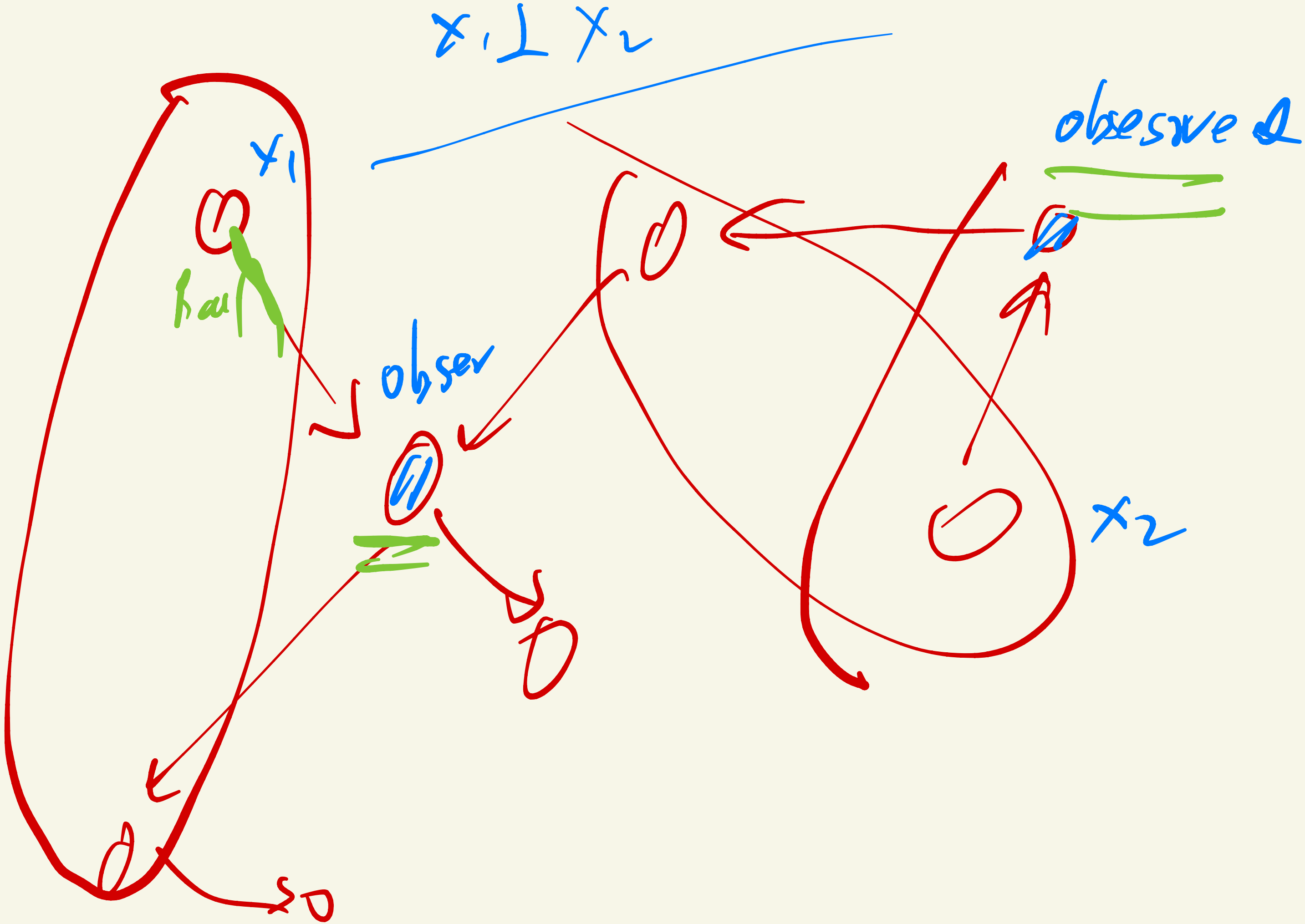
How to determine two variables are conditionally independent given another variable?

Global Markov Properties of DAGs

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

set of

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):

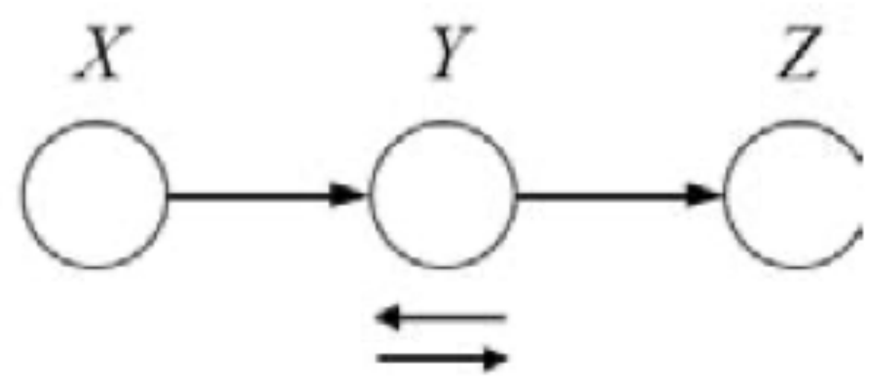


$x_1 \perp x_2$
x_1
fail
obser
observed
x_2

Global Markov Properties of DAGs

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

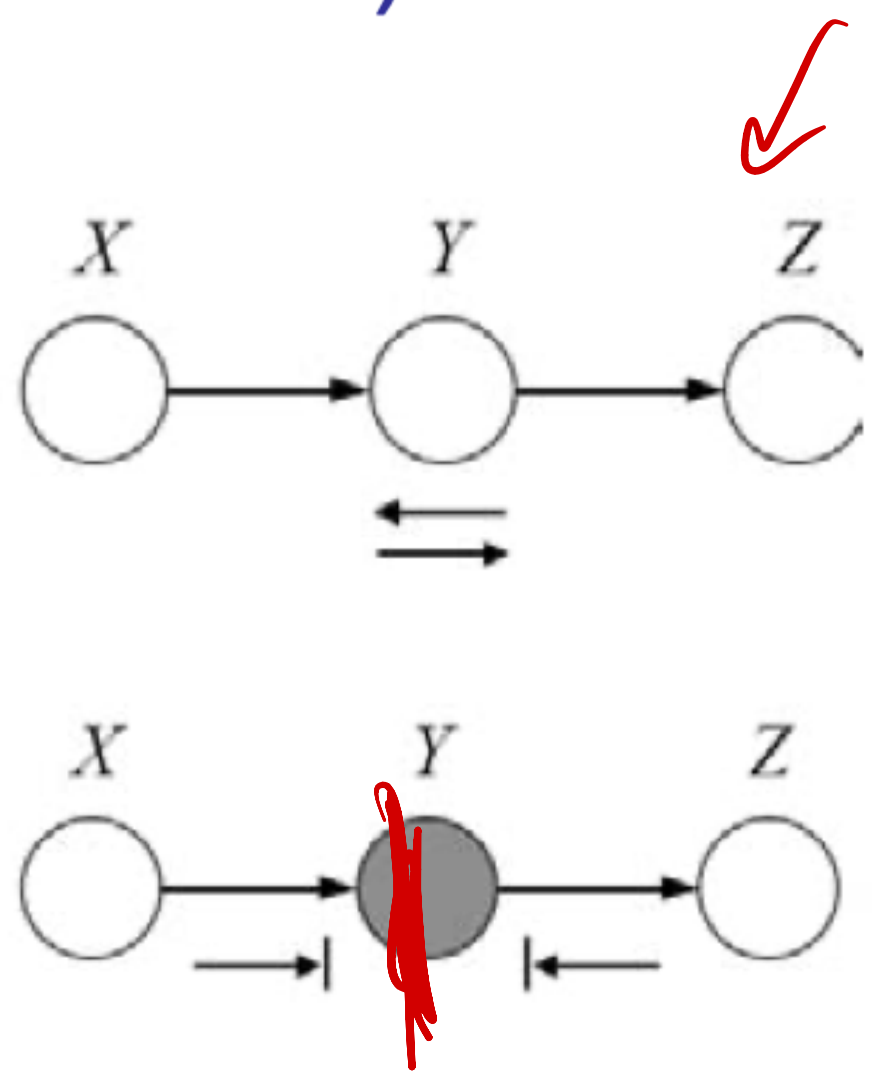
X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "*Bayes-ball*" algorithm illustrated bellow (and plus some boundary conditions):



Global Markov Properties of DAGs

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

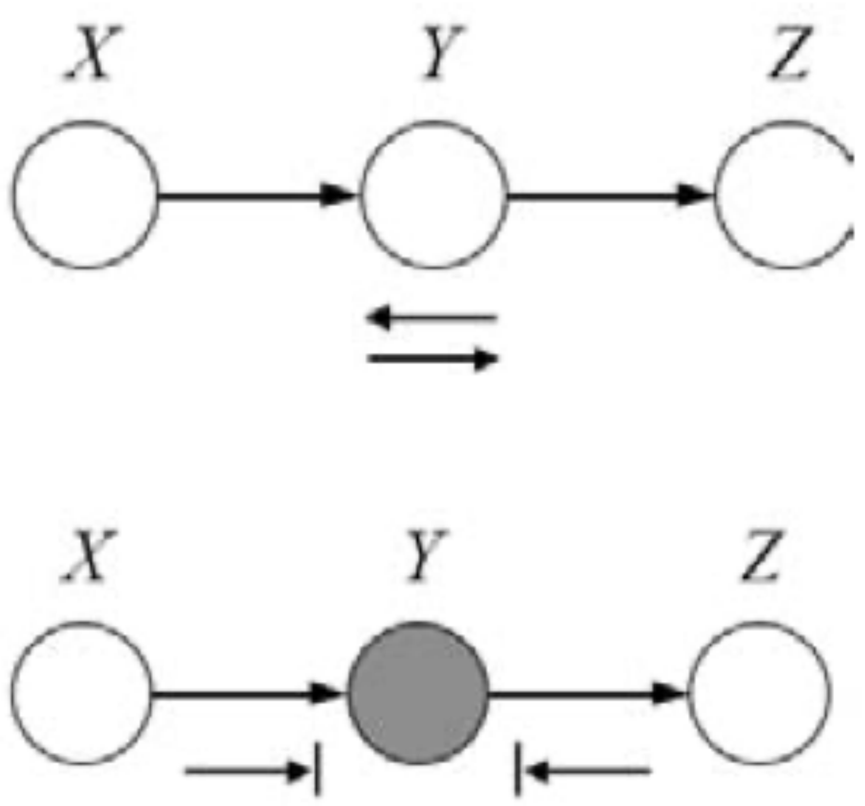
X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "*Bayes-ball*" algorithm illustrated bellow (and plus some boundary conditions):

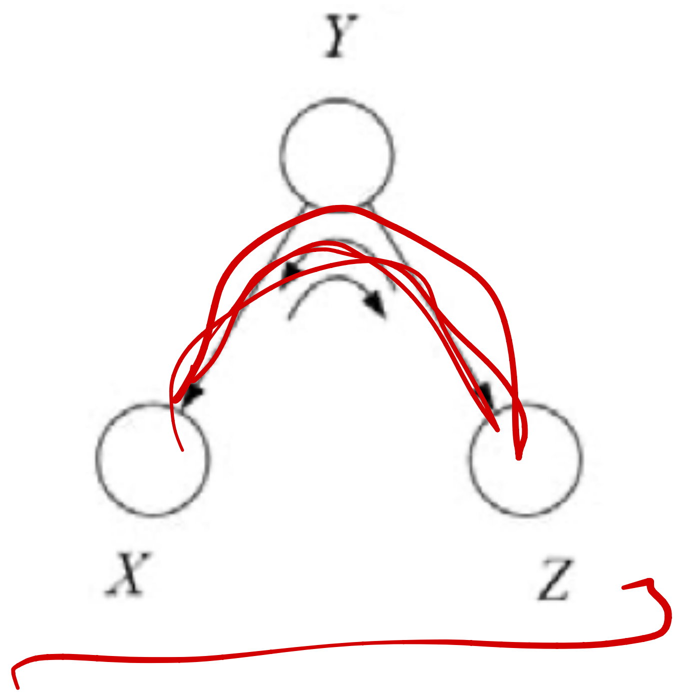


Global Markov Properties of DAGs

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):

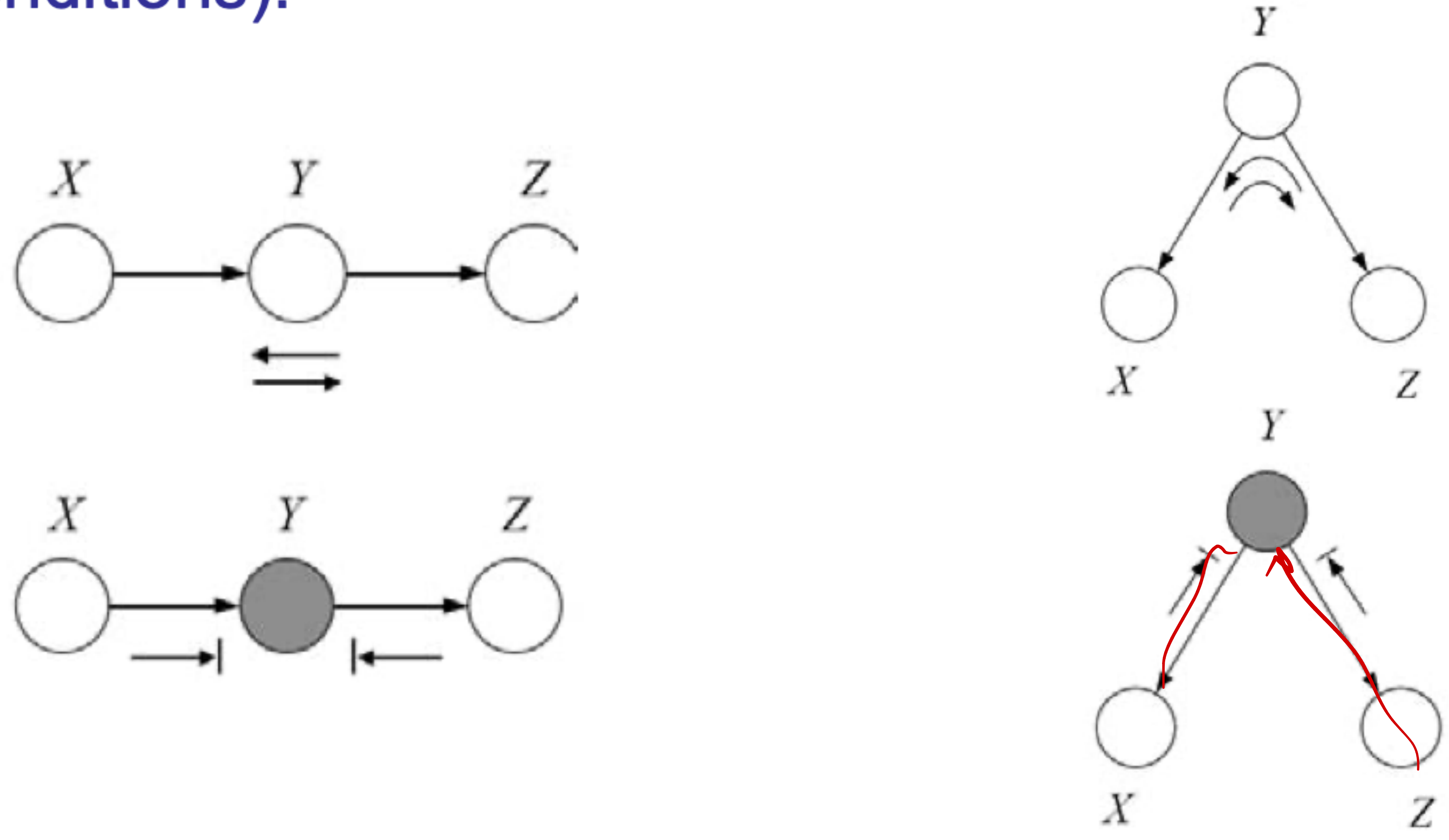




Global Markov Properties of DAGs

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

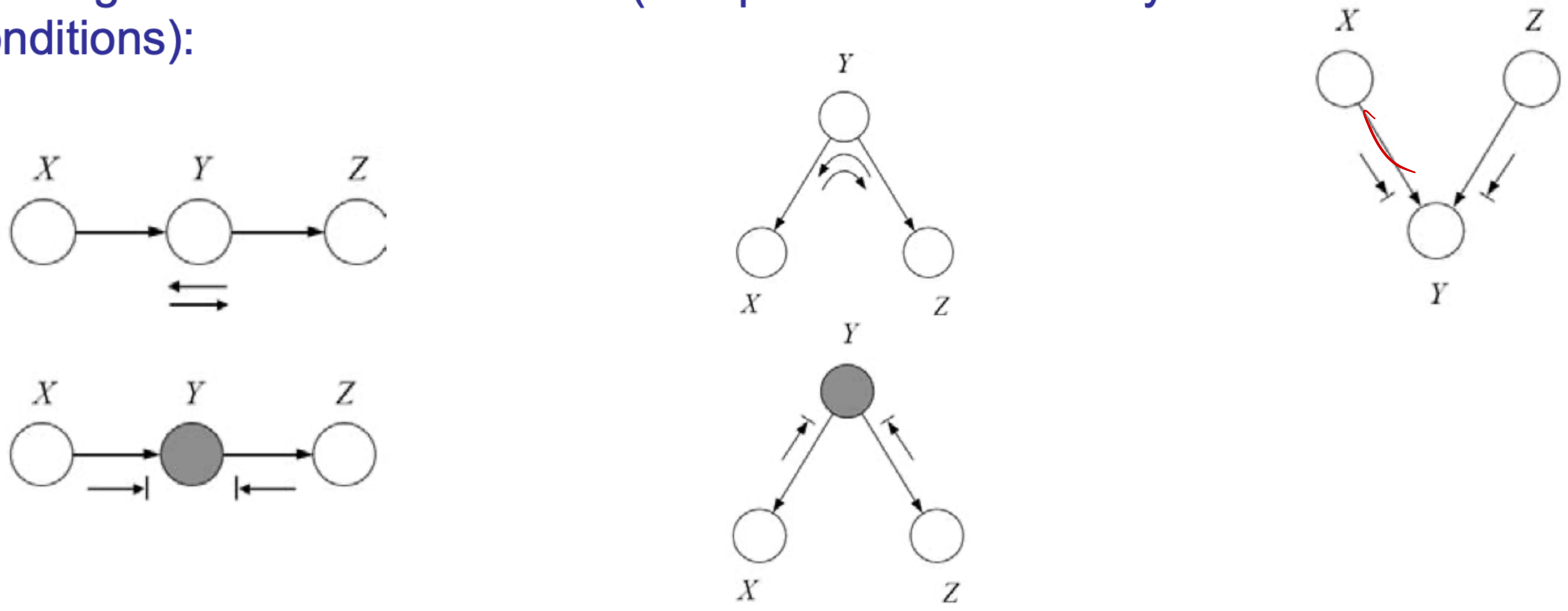
X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "*Bayes-ball*" algorithm illustrated bellow (and plus some boundary conditions):



Global Markov Properties of DAGs

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

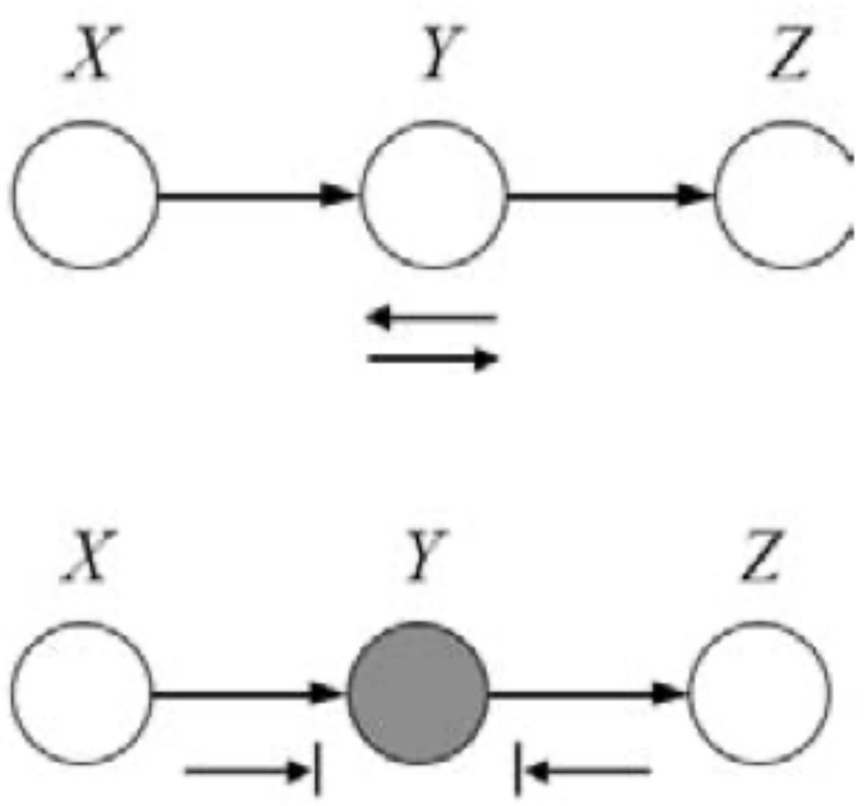
X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "*Bayes-ball*" algorithm illustrated bellow (and plus some boundary conditions):

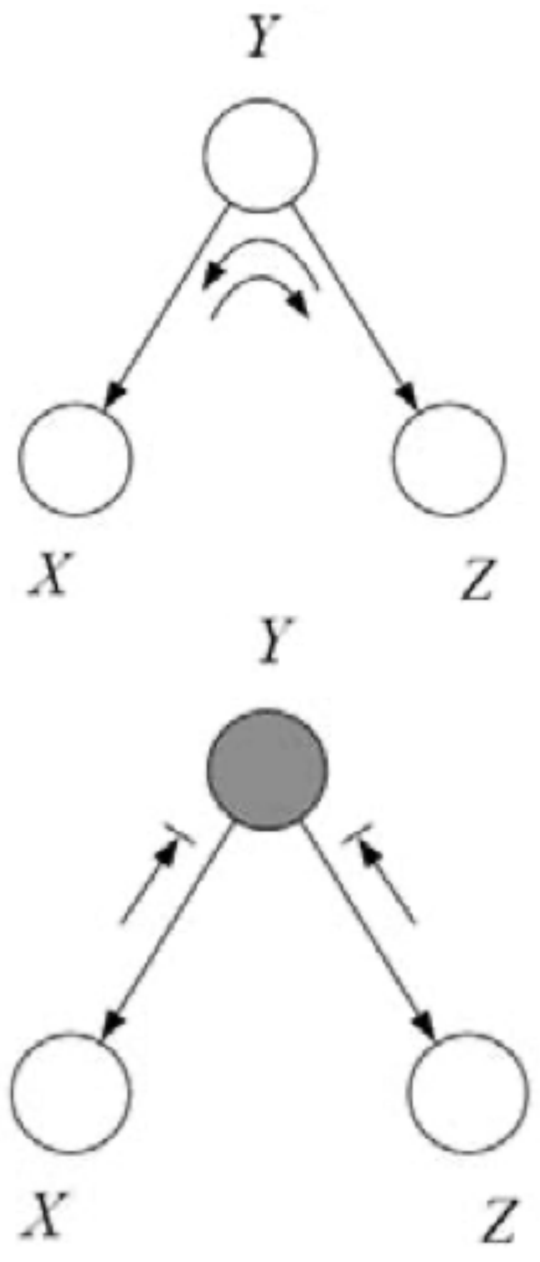


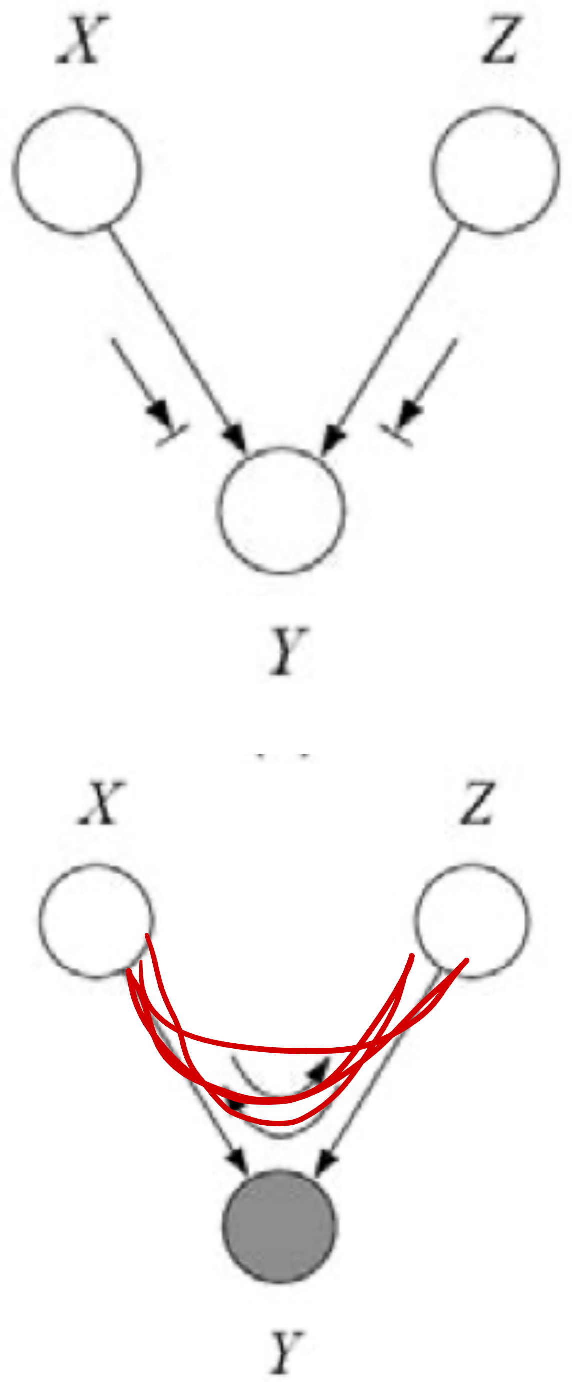
Global Markov Properties of DAGs

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

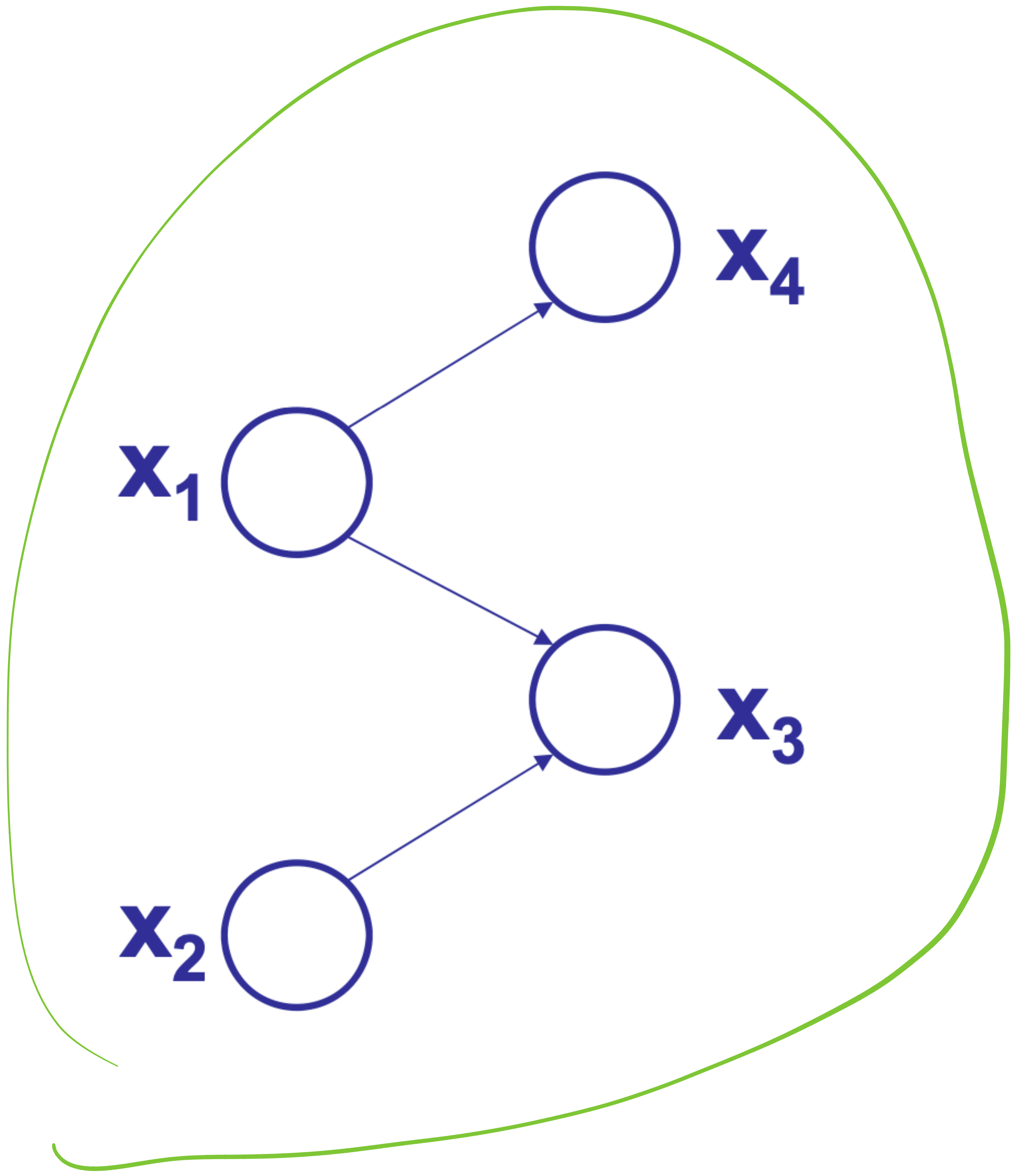
X is **d-separated** (directed-separated) from Z given Y if we can't send a ball from any node in X to any node in Z using the "*Bayes-ball*" algorithm illustrated bellow (and plus some boundary conditions):



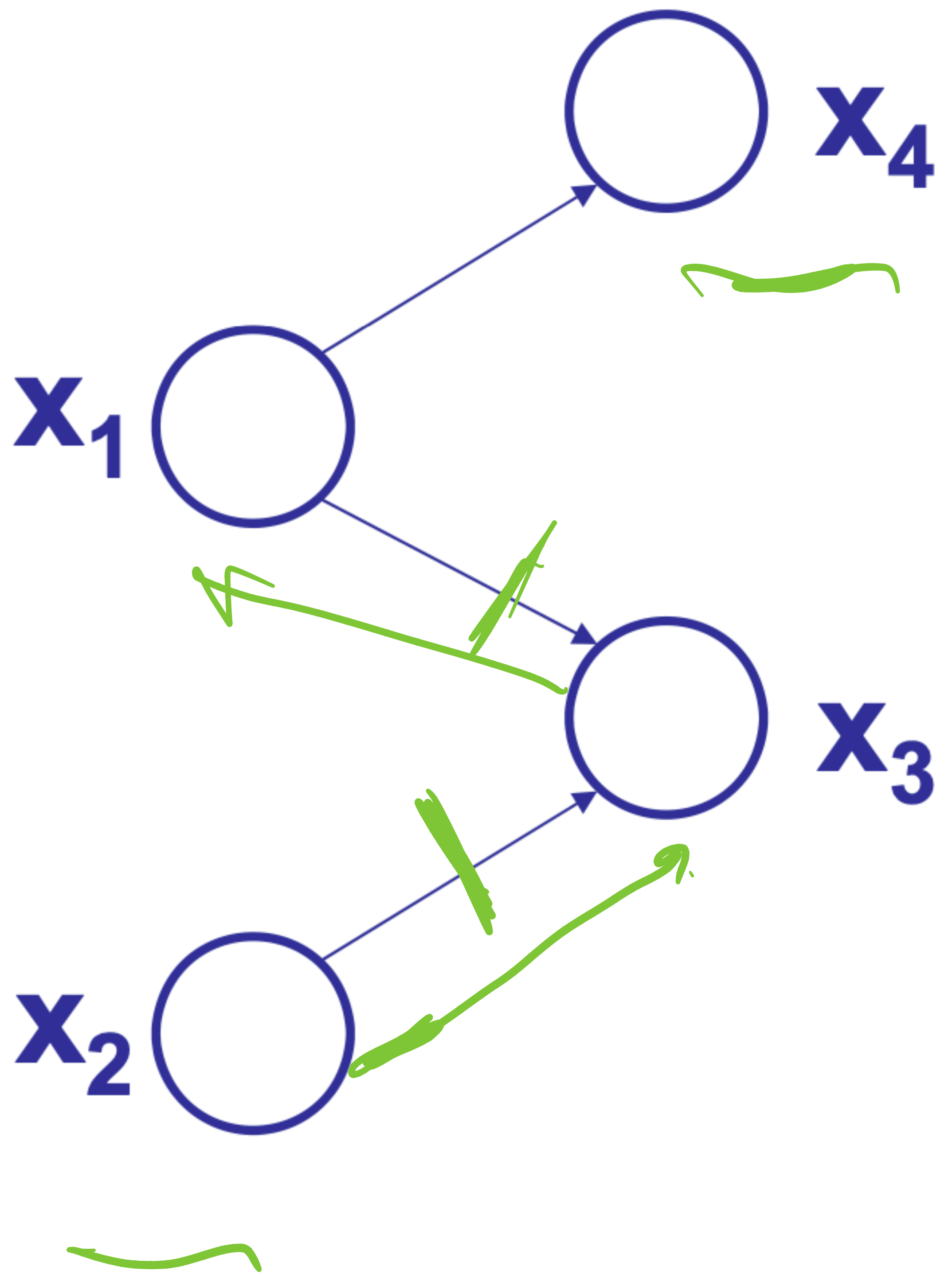




Example

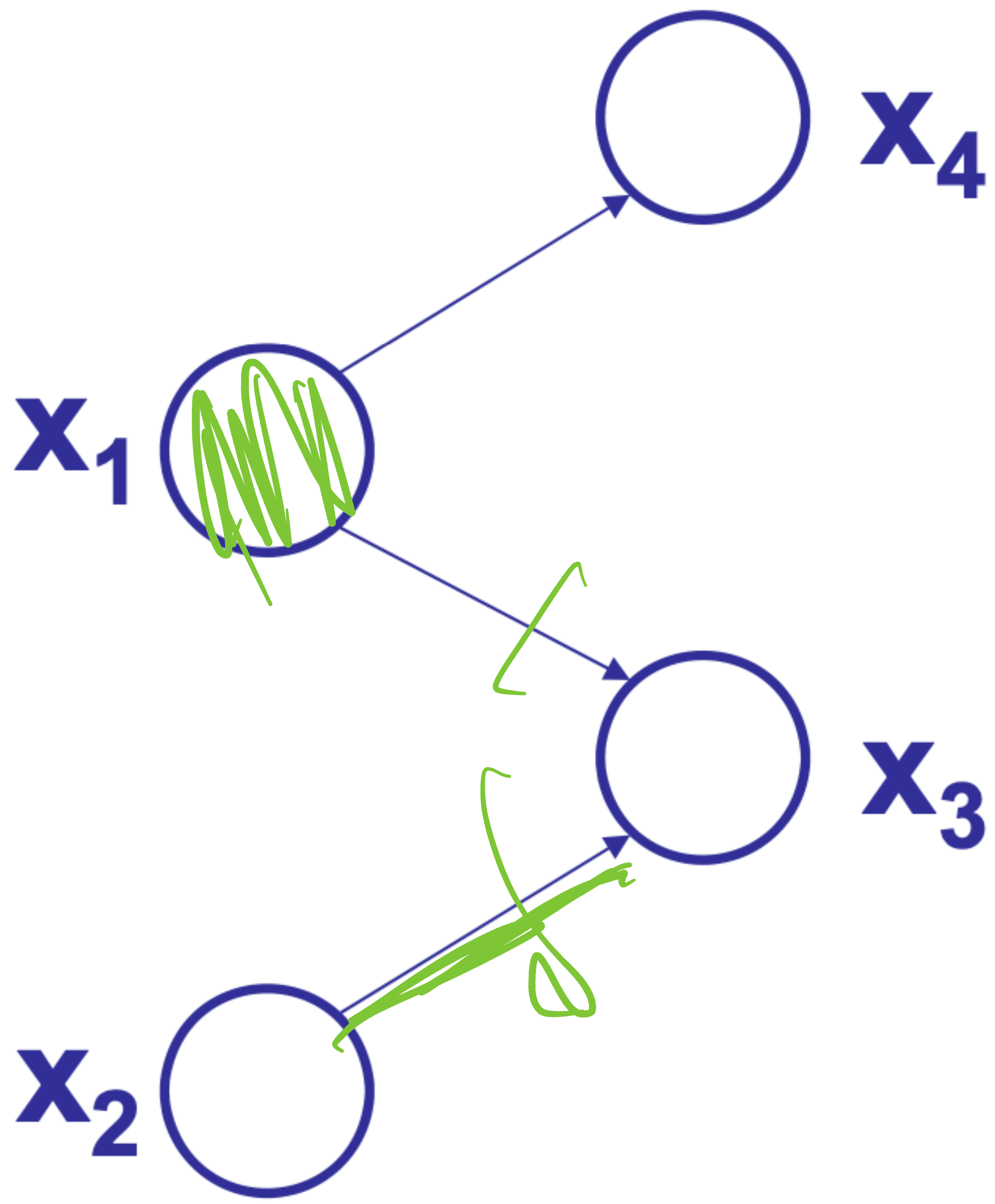


Example



1. Are X2 and X4 independent?

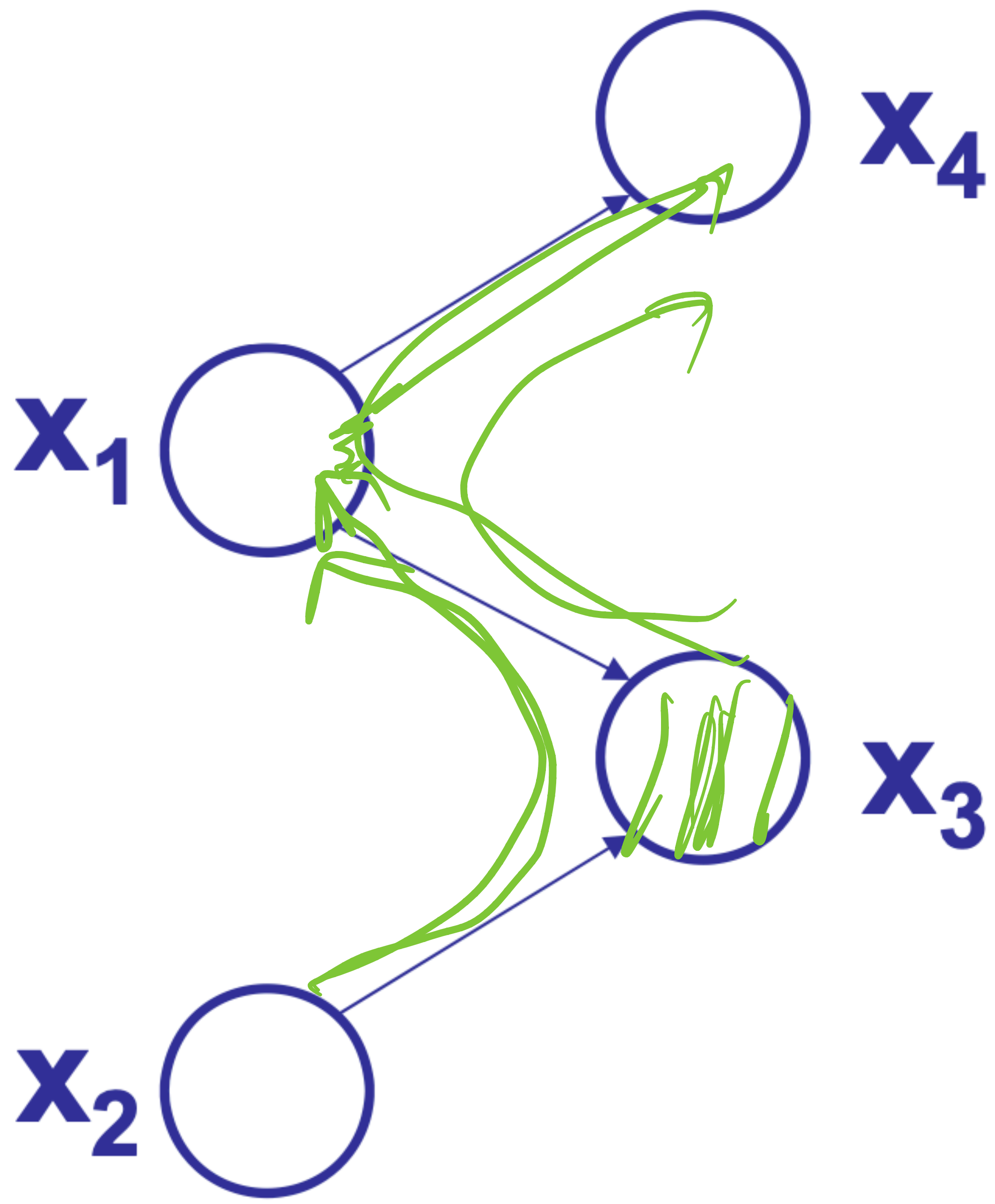
Example



1. Are X2 and X4 independent?

2. Are X2 and X4 conditionally independent given X1?

Example



1. Are X2 and X4 independent?

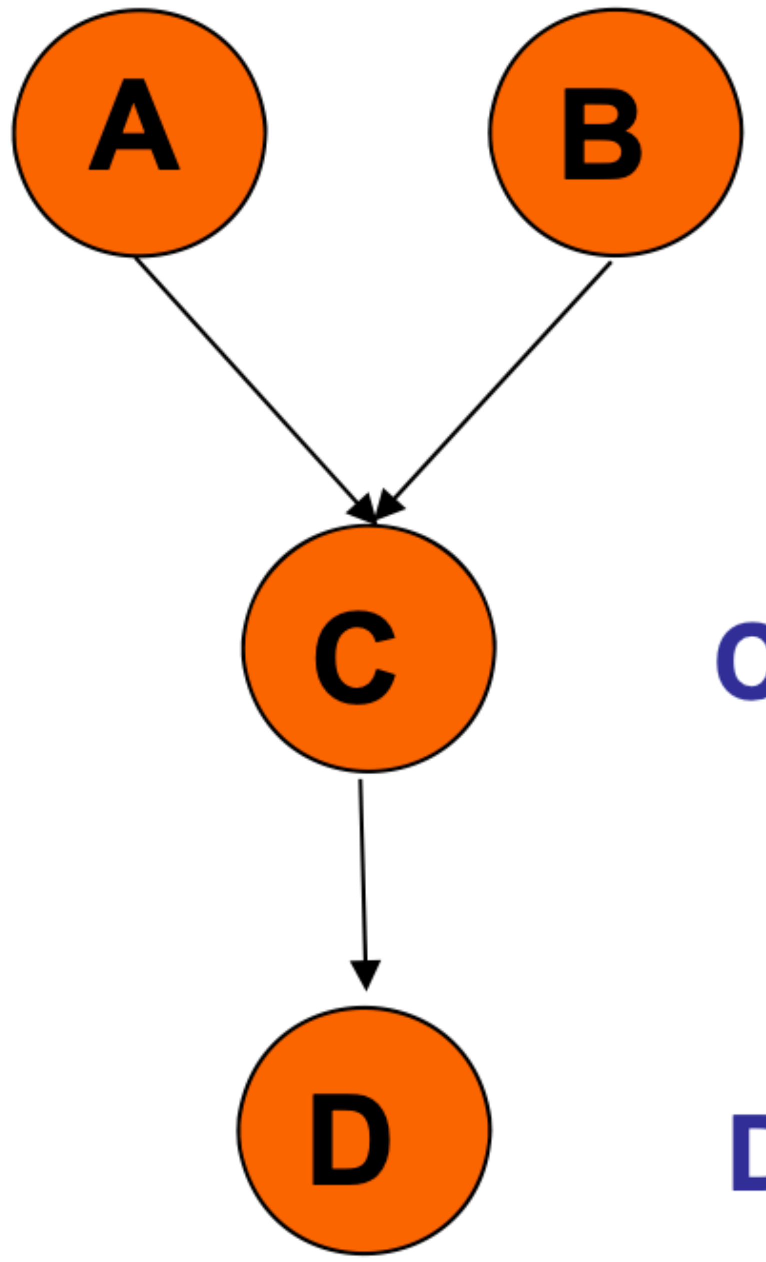
2. Are X2 and X4 conditionally independent given X1?

3. Are X2 and X4 conditionally independent given X3?

Conditional Probability Density Func

Conditional Probability Density Func

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



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

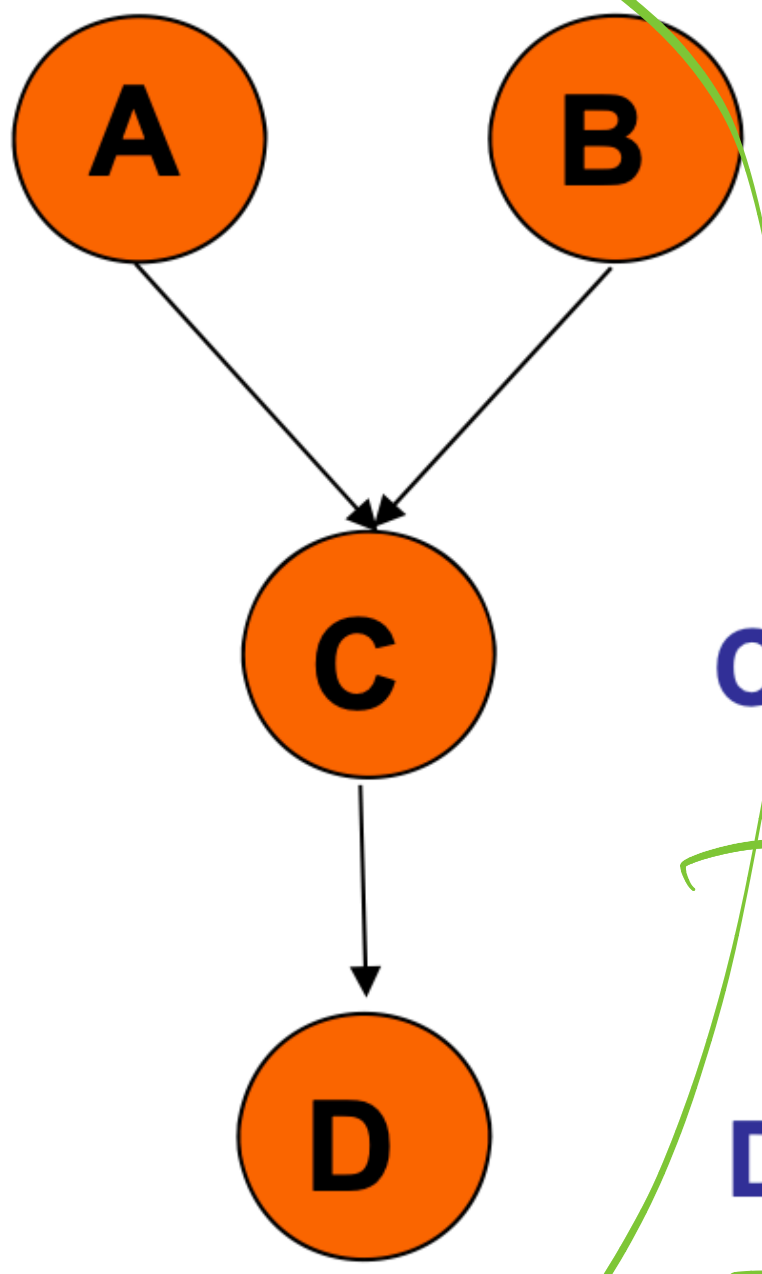
$D \sim N(\mu_a+C, \Sigma_a)$

Conditional Probability Density Func

parameter

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

P(a) P(b) P(c|A,b) P(d|c))



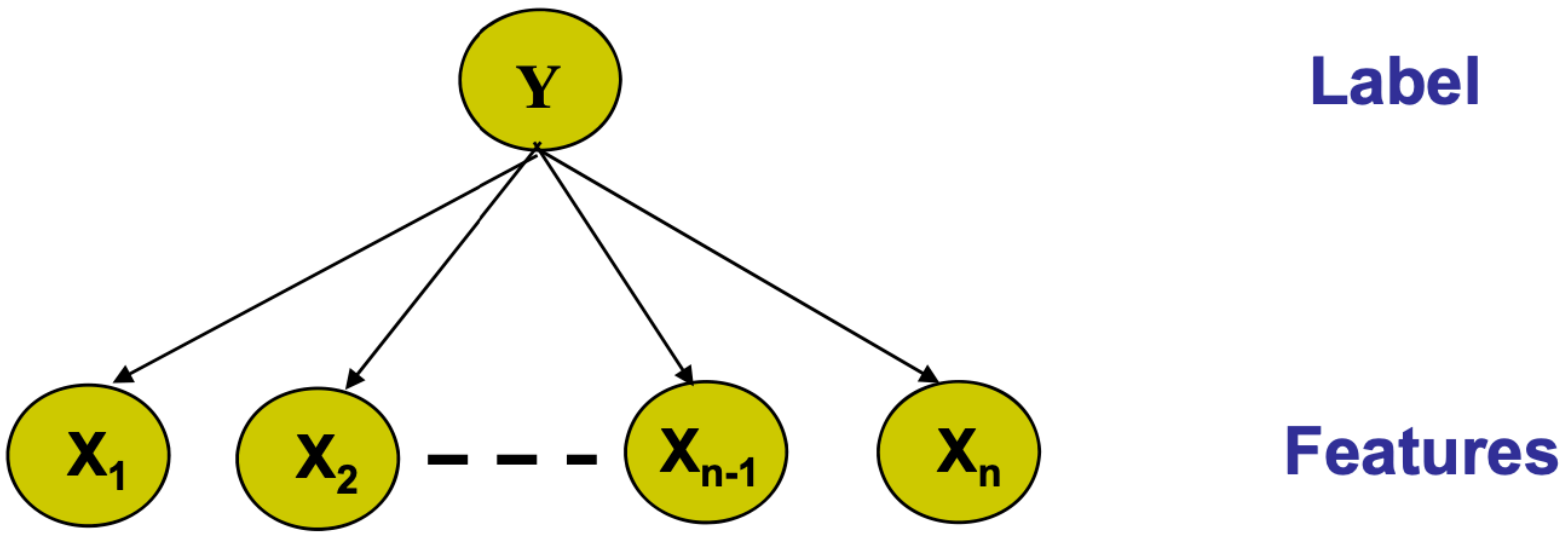
$C \sim N(A+B, \Sigma_c)$

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

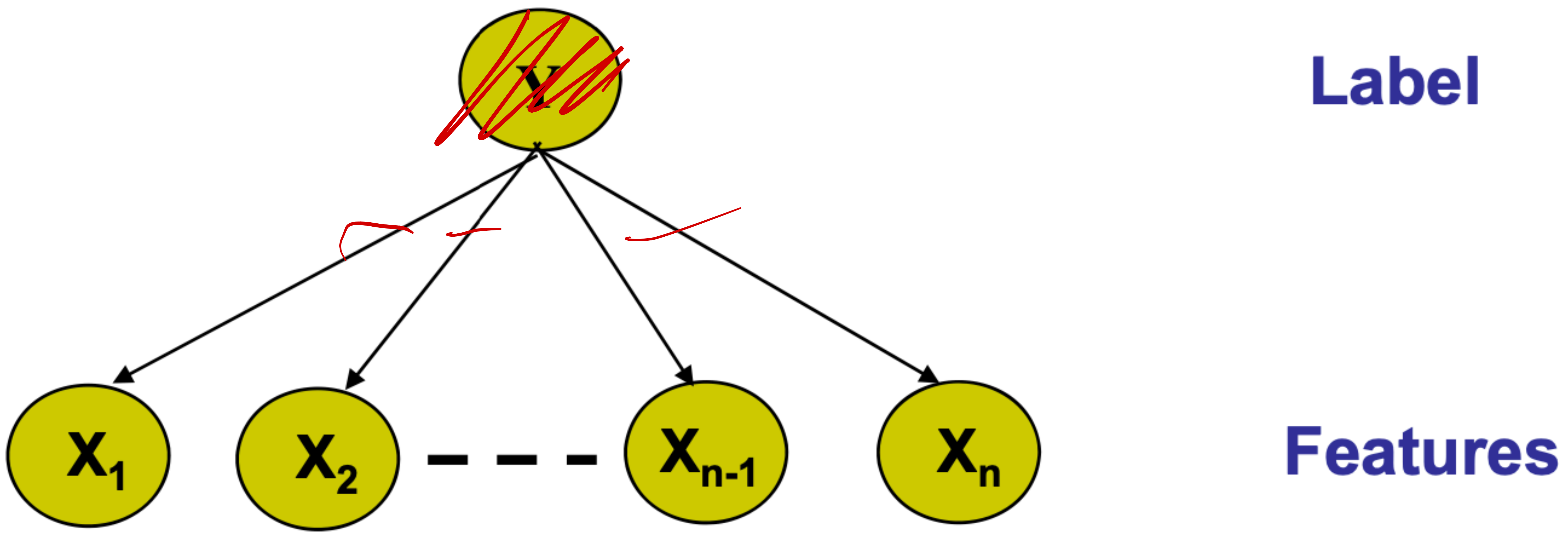
$$\boxed{P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)}$$

Conditional Independencies

Conditional Independencies

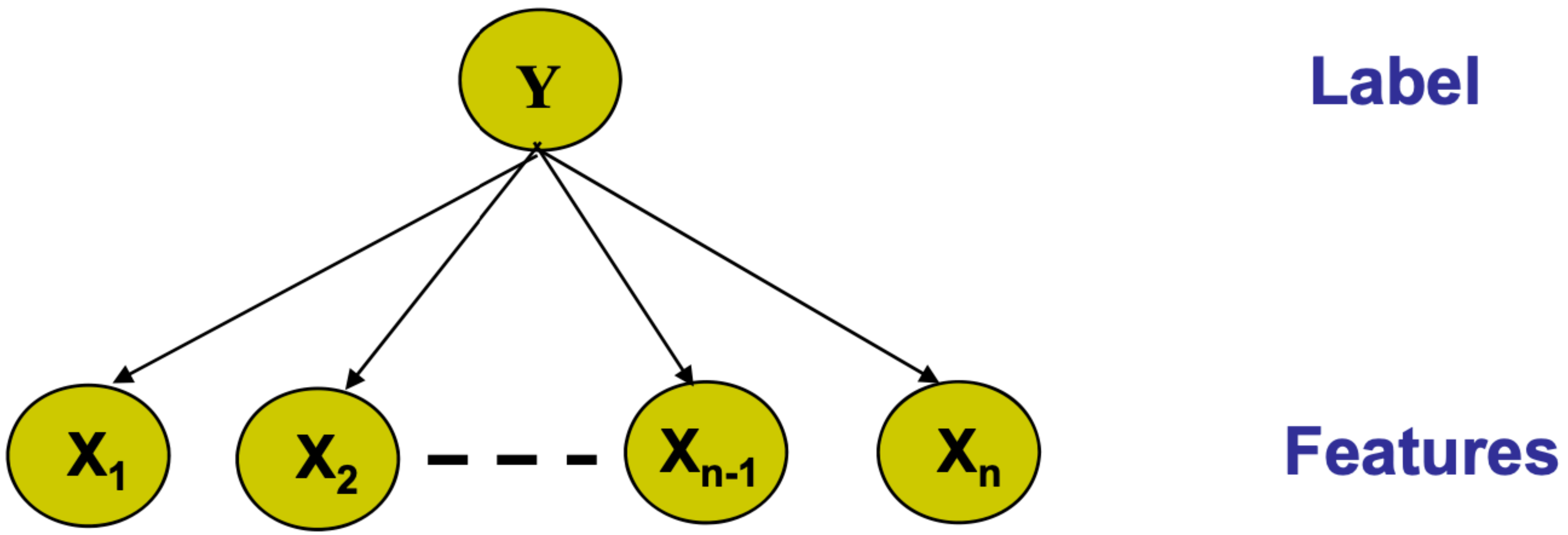


Conditional Independencies



Are Xi D-separated from Xj given Y? Yes

Conditional Independencies

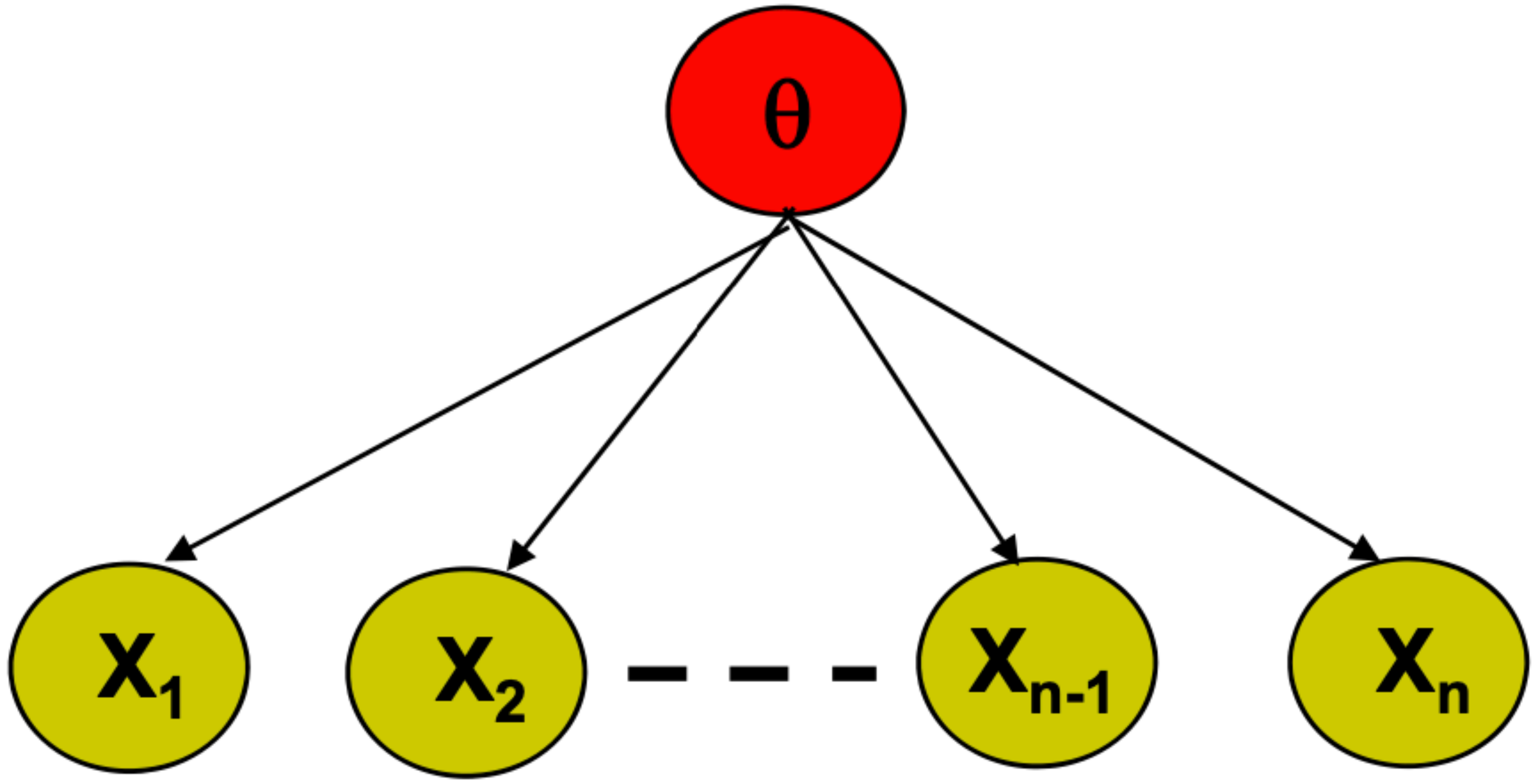


Are Xi D-separated from Xj given Y?

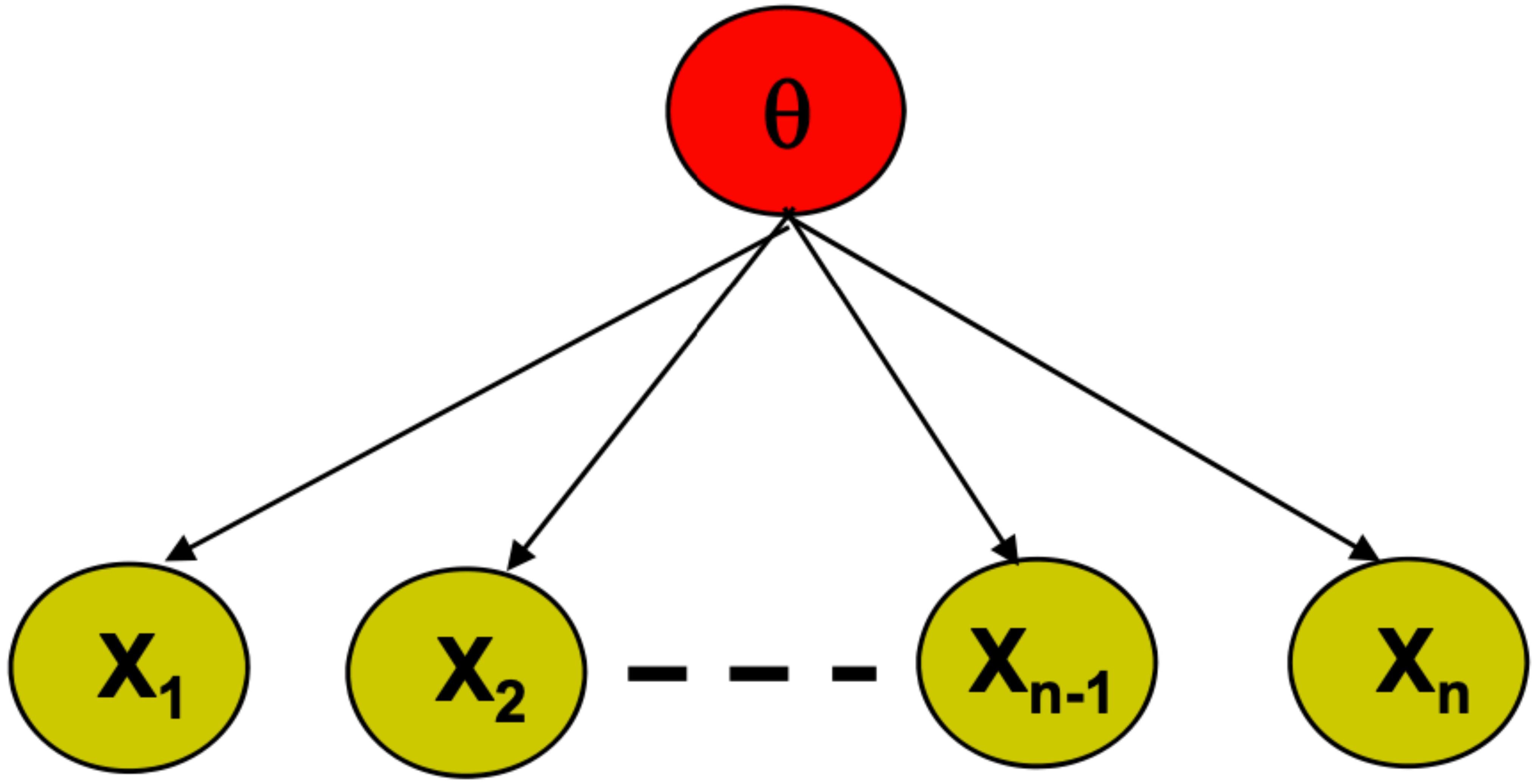
What is this model when Y is observed? Naive Bayes

Conditionally Independent Observations

Conditionally Independent Observations

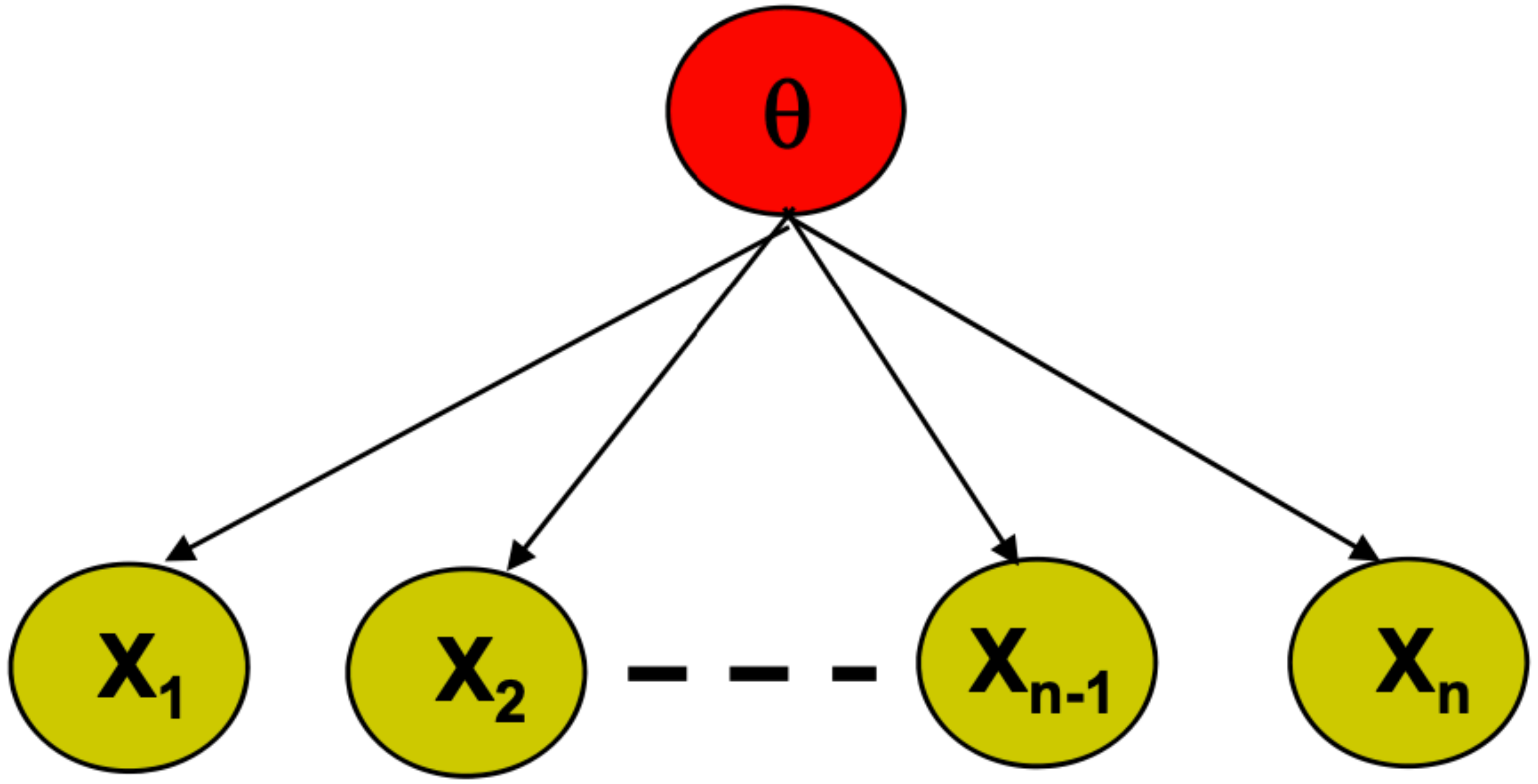


Conditionally Independent Observations



Model parameters

Conditionally Independent Observations



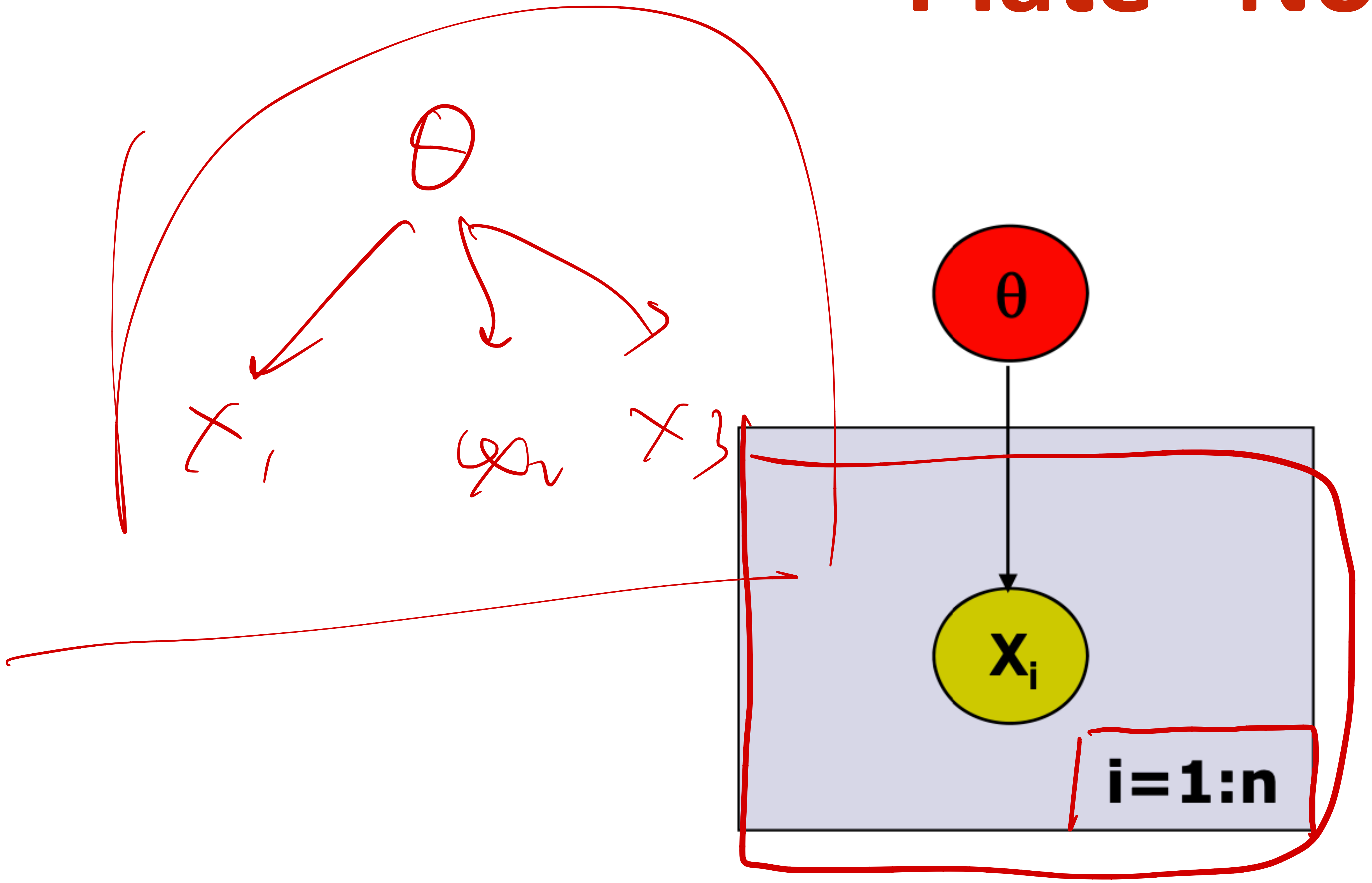
Model parameters

Data {X1, X2 …. Xn}

i.i.d

“Plate” Notation

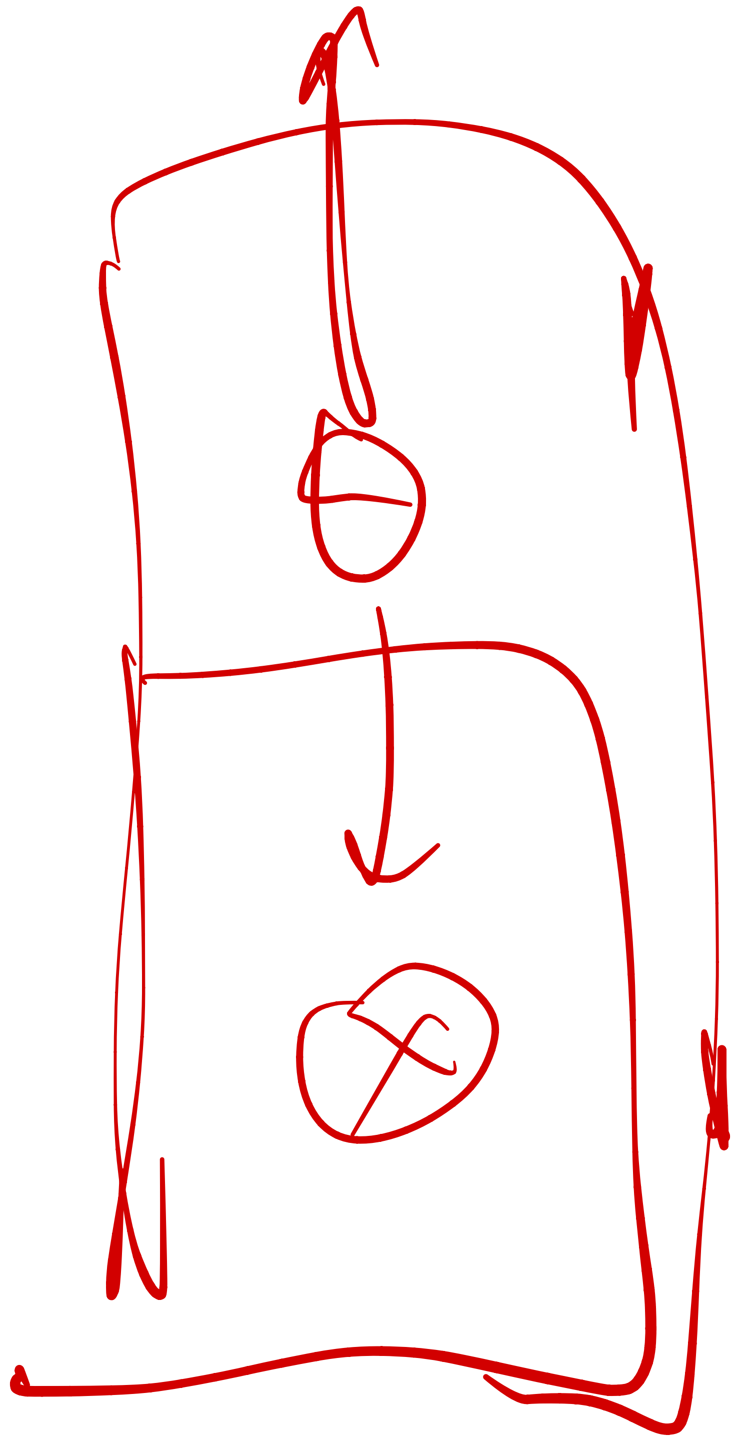
“Plate” Notation

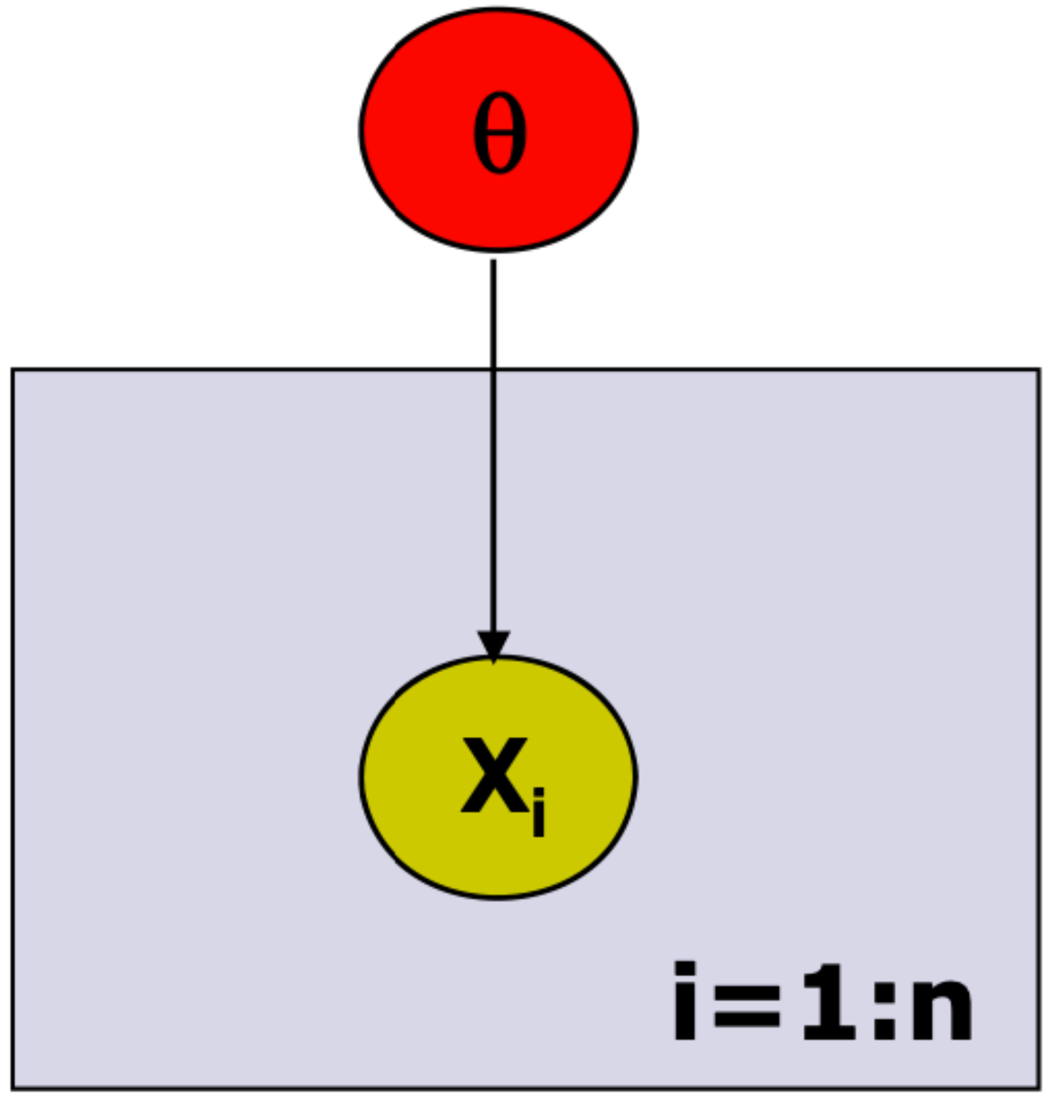


Model parameters

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

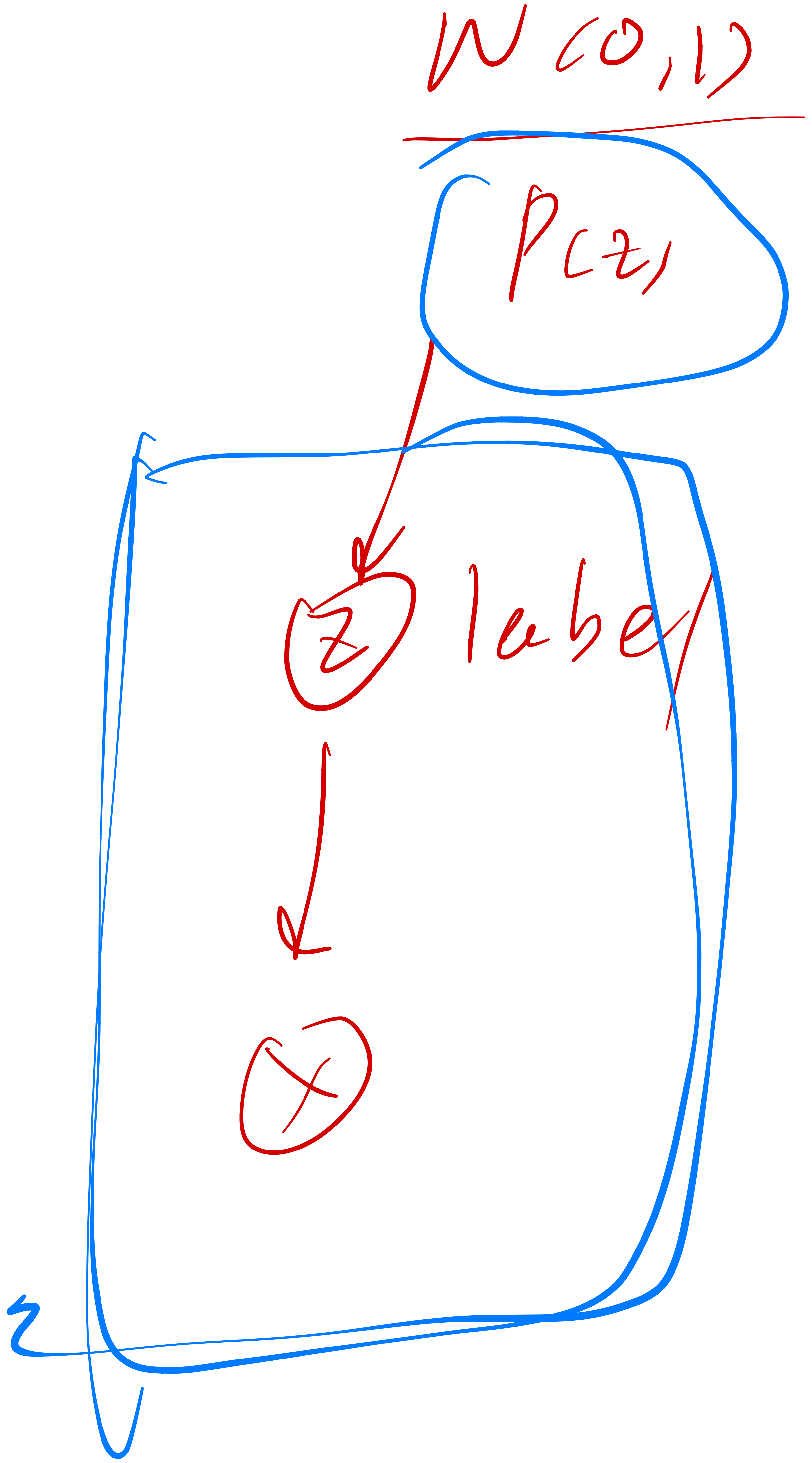
“Plate” Notation



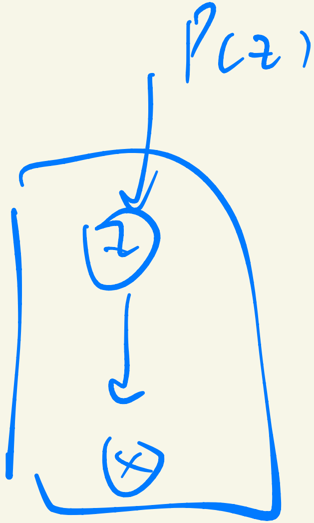


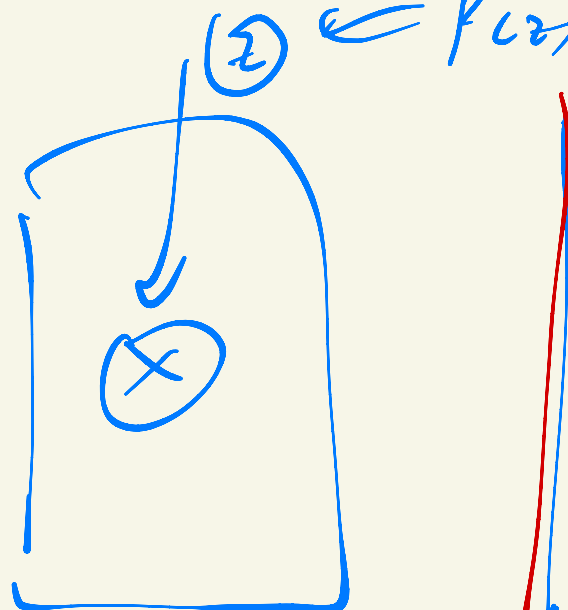
Model parameters

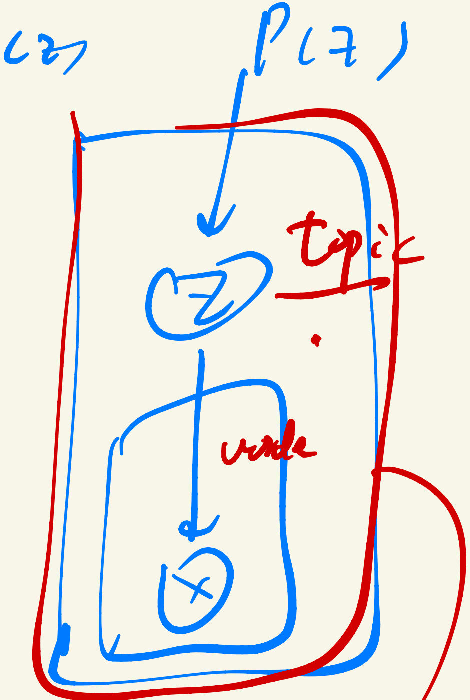
Data = $\{x_1, \ldots x_n\}$



variables within a plate are replicated in a conditionally independent manner







for each document

topic, for each topic, we generate N words

Example: Gaussian Model

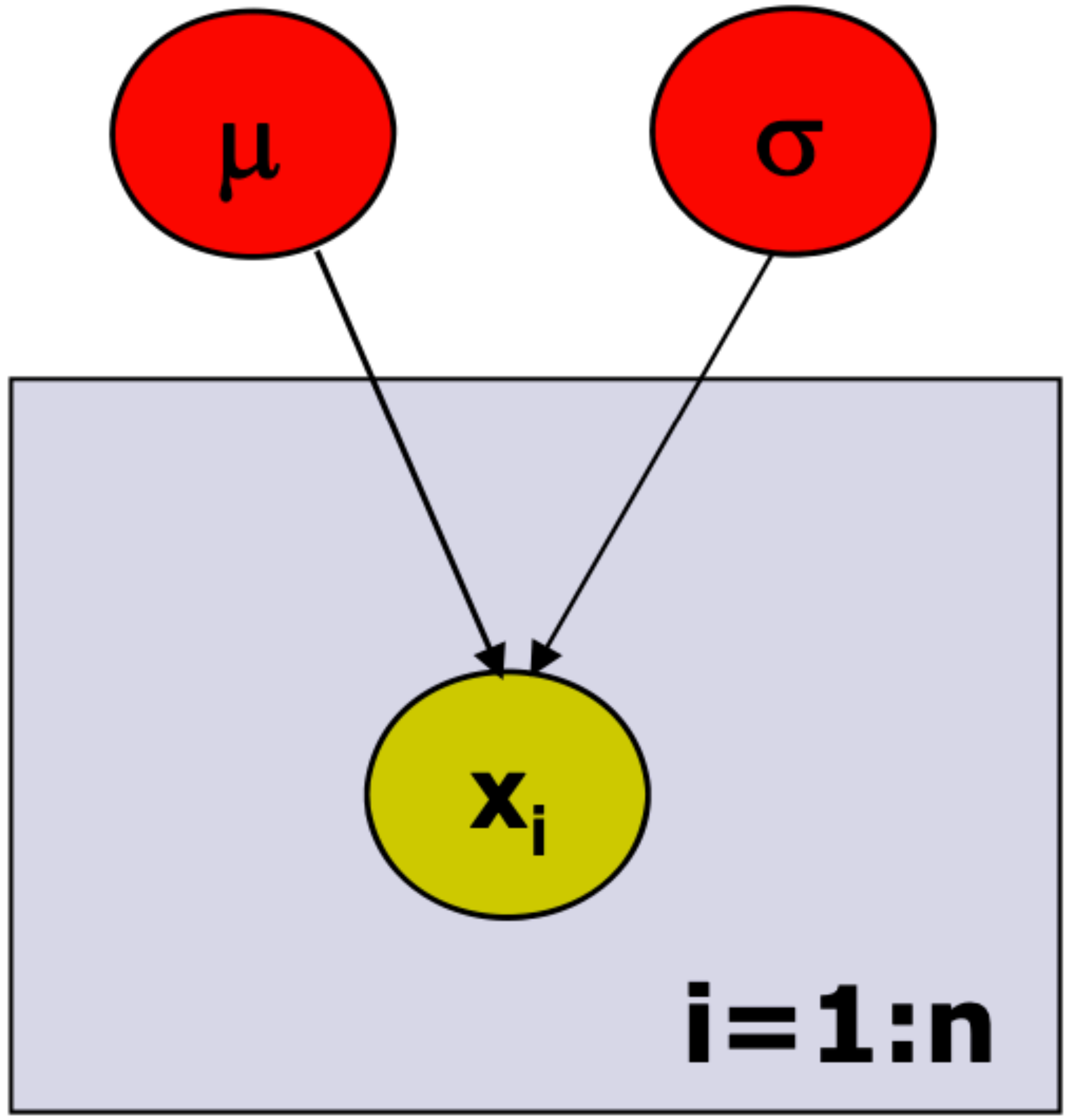
Generative model:

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

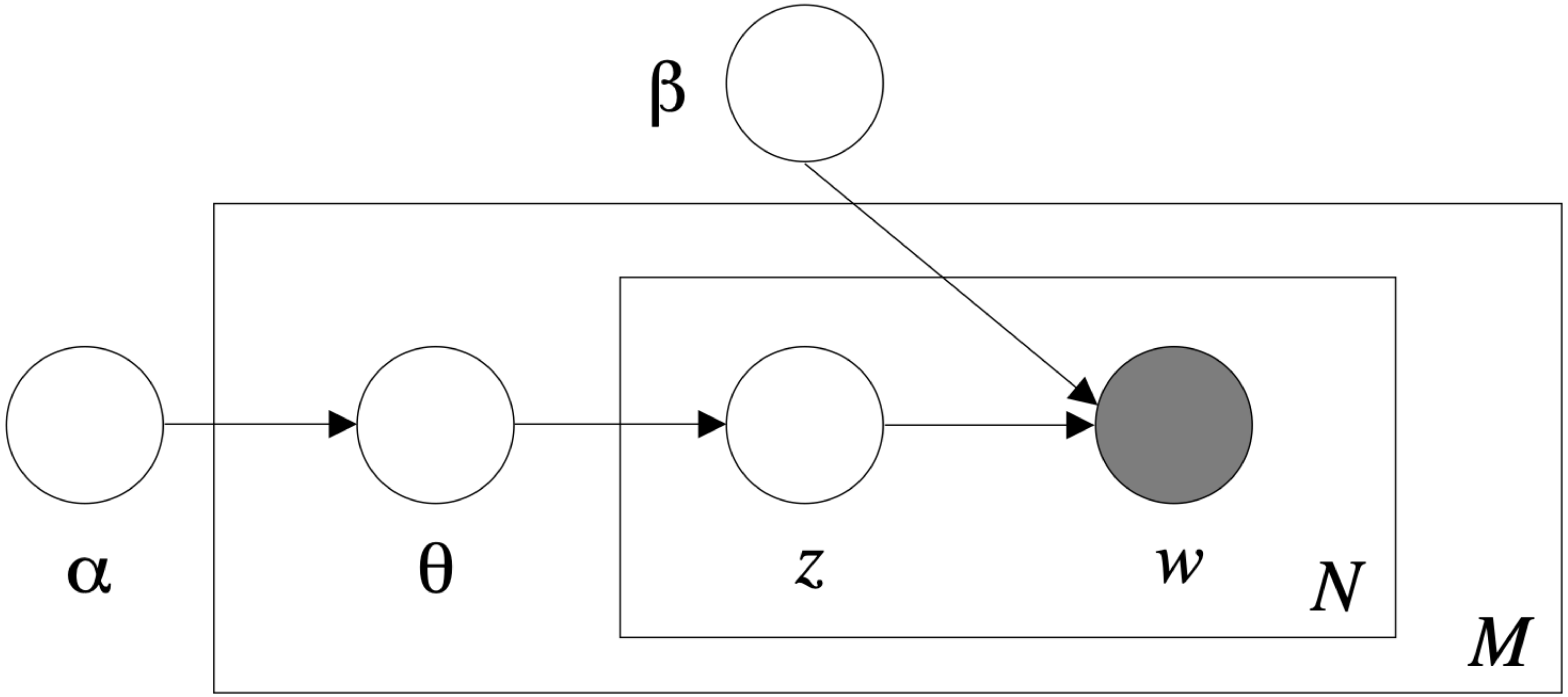
$= \;\; p(\text{data} \mid \text{parameters})$

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

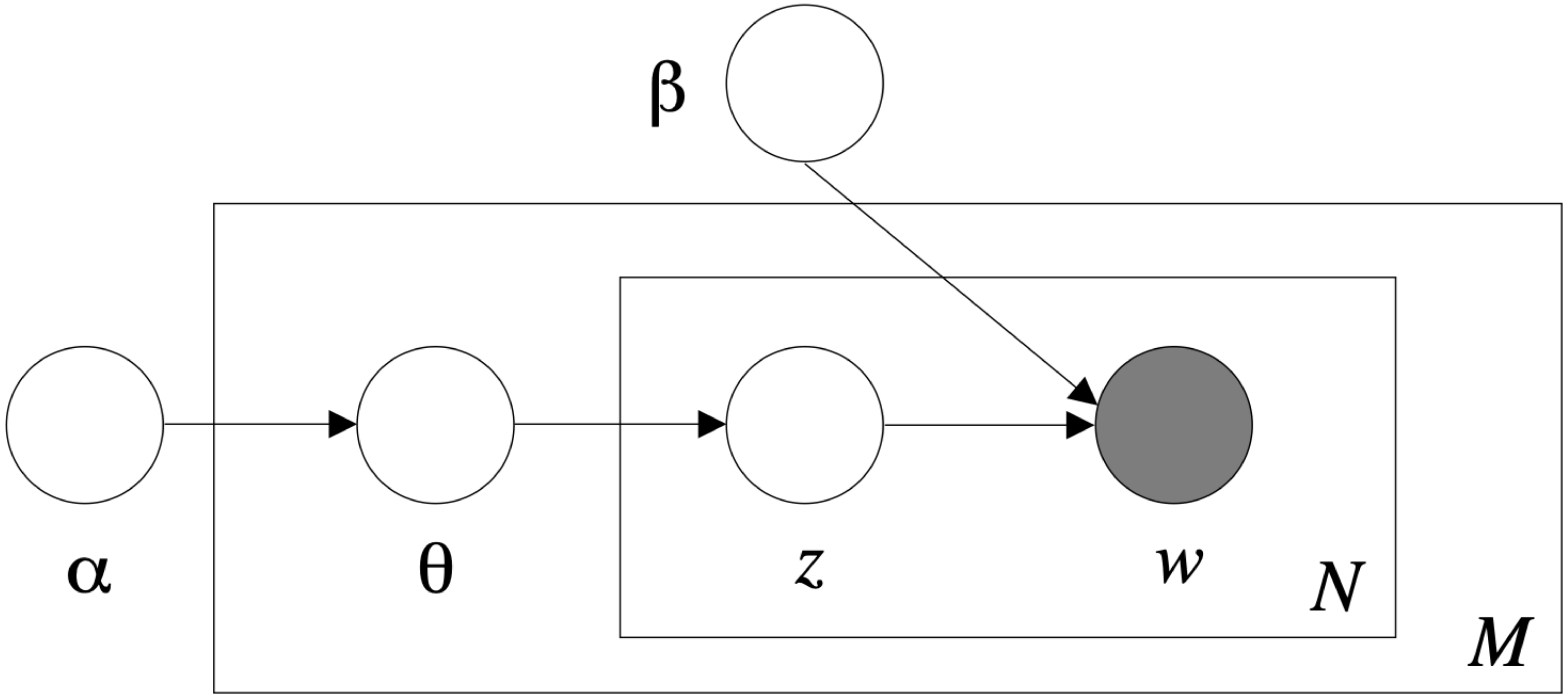
where $\theta = \{\mu, \sigma\}$



Observed Variable and Latent Variable Notations



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

Inference and Learning

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

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

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

Examples

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

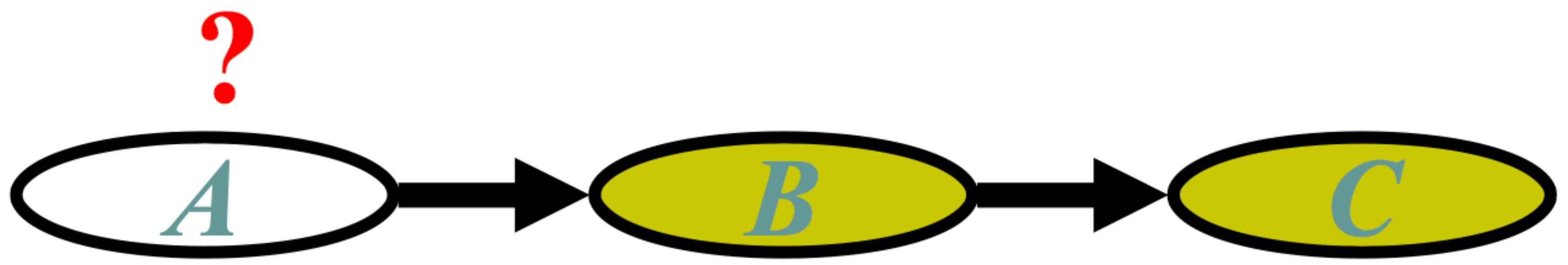
  - the query node is a descendent of the evidence

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



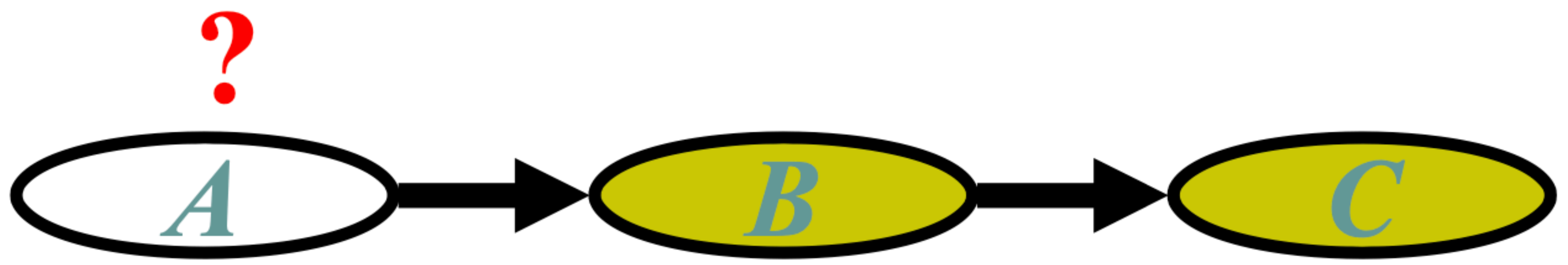
  - the query node an ancestor of the evidence

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

How to Learn the Parameters

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

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

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?

How to do MLE on Latent Variable Models?

Expectation Maximization!

How to do MLE on Latent Variable Models?

Expectation Maximization!

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

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

Approaches to Inference

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

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

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

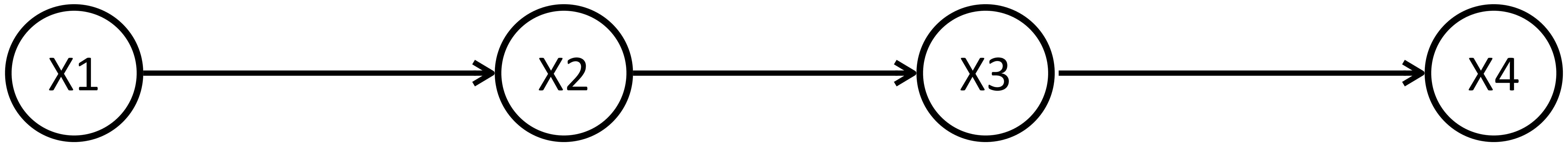
$$P(h) = \sum_g \underbrace{\sum_f \sum_e \sum_d \sum_c \sum_b}_{} \sum_a P(a,b,c,d,e,f,g,h)$$

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

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



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

$$X1 \rightarrow X2 \rightarrow X3 \rightarrow X4$$

What if the random variables follow this chain structure?

Thank You!
Q & A