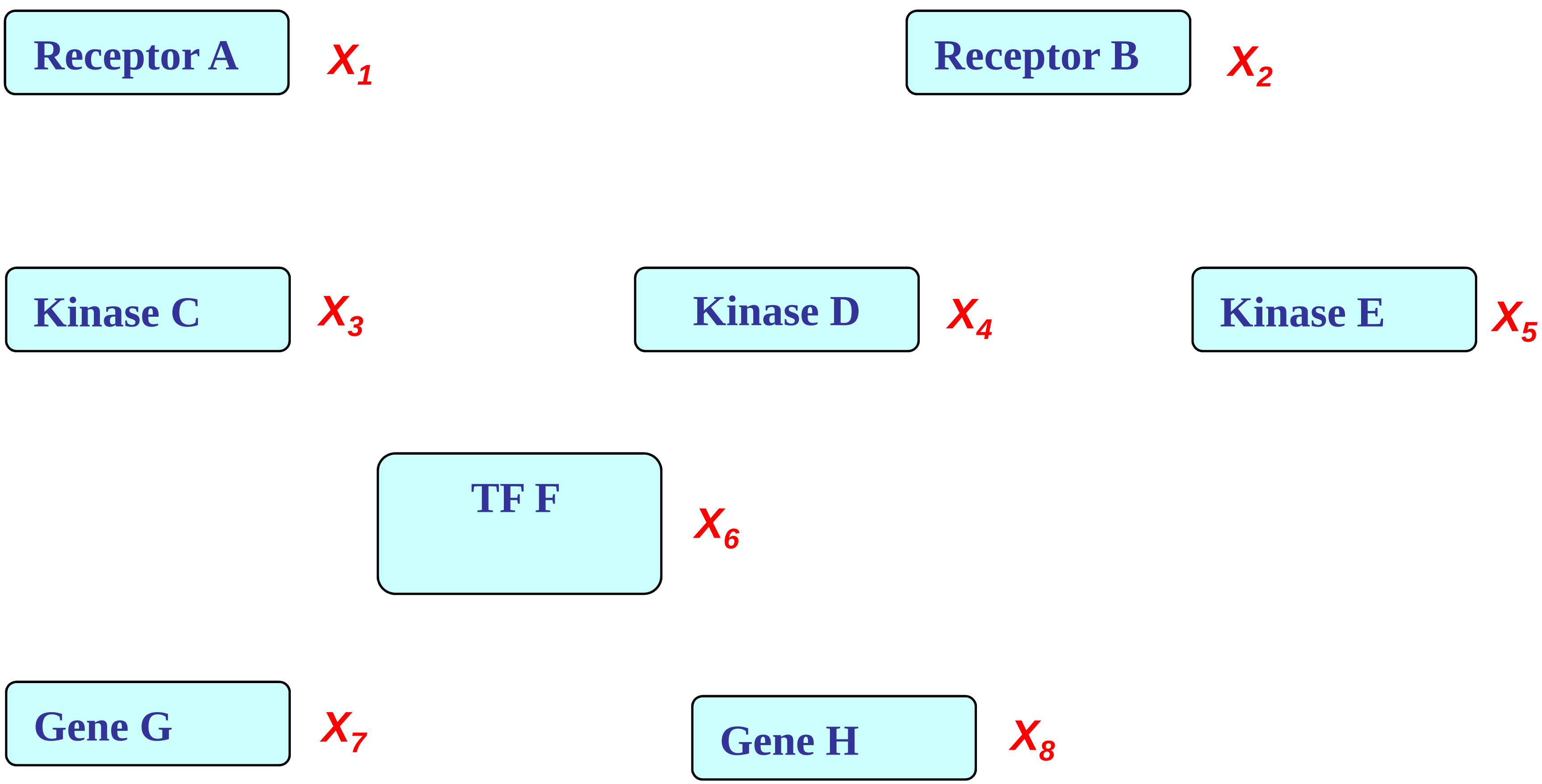
Multivariate Distribution in High-D Space

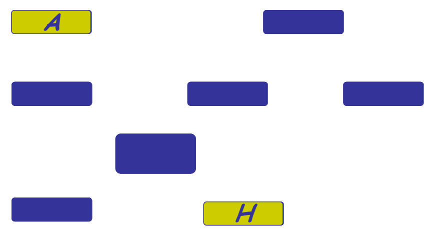
- A possible world for cellular signal transduction:



Recap of Basic Prob. Concepts

- Representation: what is the joint probability dist. on multiple variables?

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



  - How many state configurations in total? --- 2^8
  - Are they all needed to be represented?
  - **Do we get any scientific/medical insight?**

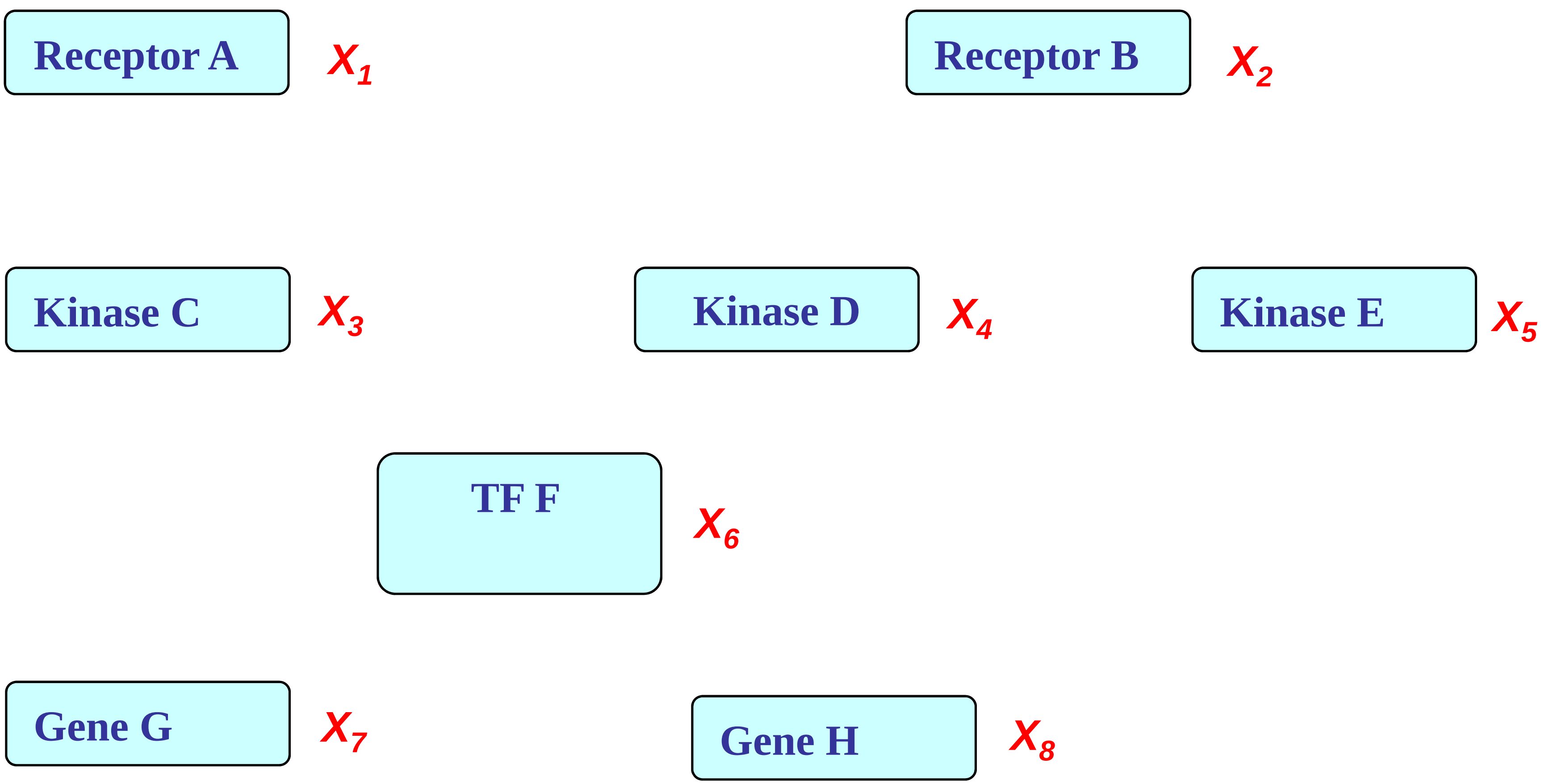
- Learning: where do we get all this probabilities?
  - Maximal-likelihood estimation? but how many data do we need?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?

- Inference: If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?
  - Computing $p(H|A)$ would require summing over all 2^6 configurations of the unobserved variables

What is a Graphical Model?

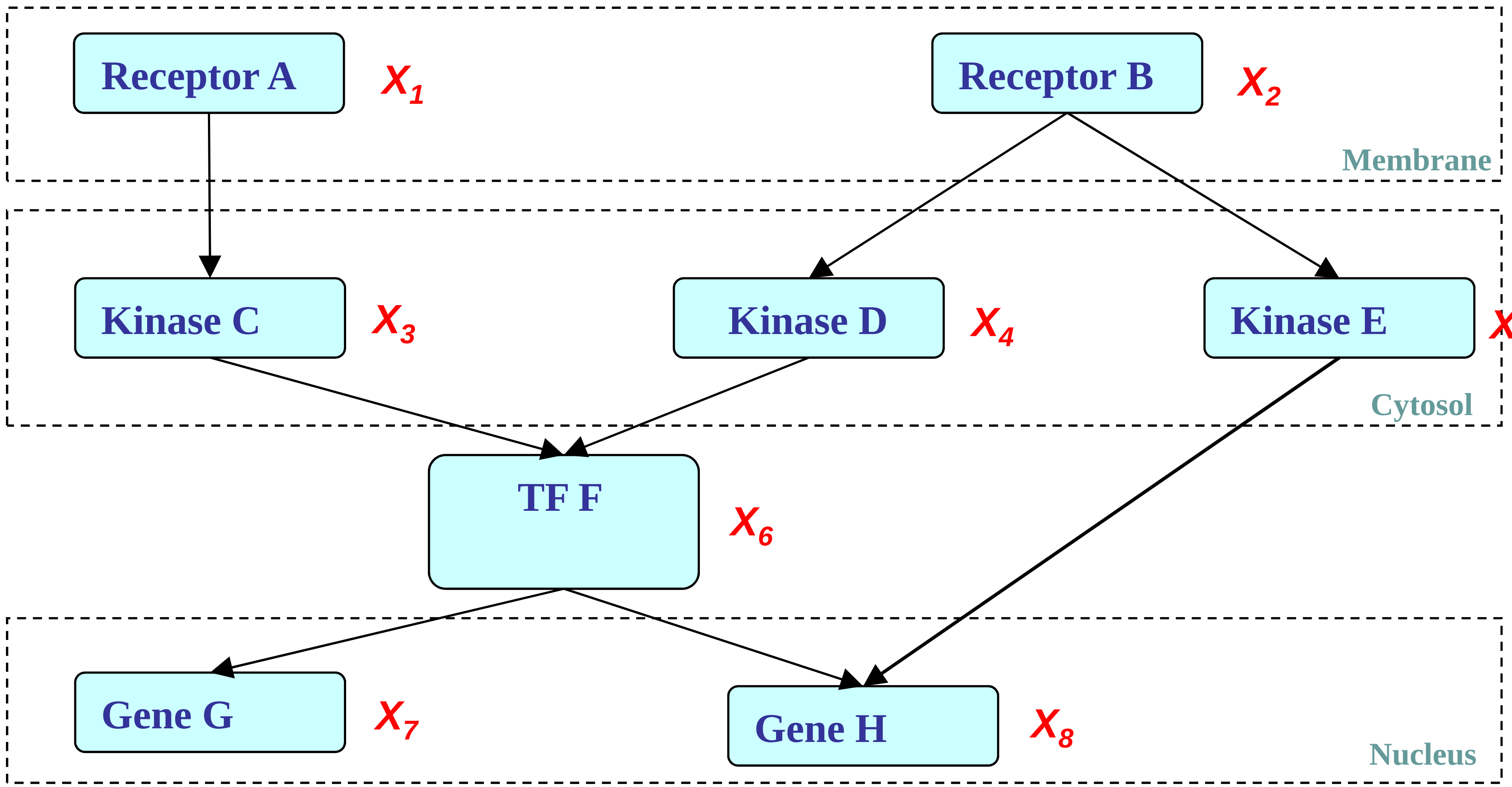
--- example from a signal transduction pathway

- A possible world for cellular signal transduction:



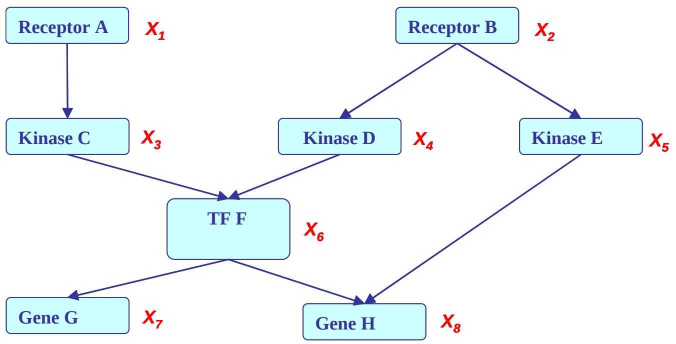
GM: Structure Simplifies Representation

- Dependencies among variables



Probabilistic Graphical Models, con'd

- 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.,



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

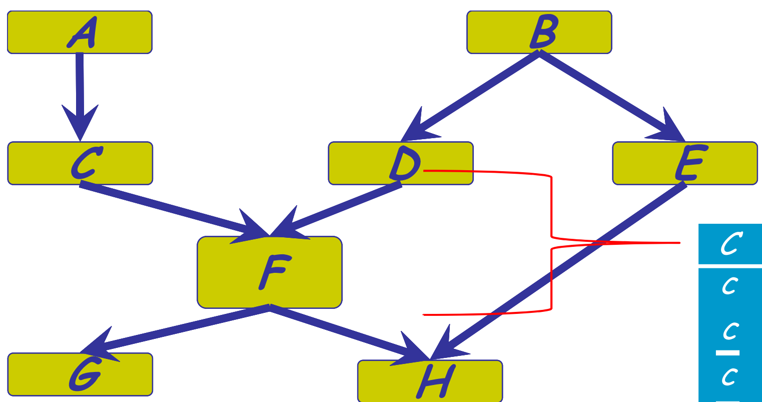
$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

- Why we may favor a PGM?
  - Representation cost: how many probability statements are needed?

    an 8-fold reduction from 2^8!
  - Algorithms for systematic and efficient inference/learning computation
    - Exploring the graph structure and probabilistic (e.g., Bayesian, Markovian) semantics
  - Incorporation of domain knowledge and causal (logical) structures

Specification of a BN

- There are two components to any GM:
  - the *qualitative* specification
  - the *quantitative* specification



C	D	P(F | C,D)	
c	d	0.9	0.1
c	$\bar{d}$	0.2	0.8
$\bar{c}$	d	0.9	0.1
$\bar{c}$	$\bar{d}$	0.01	0.99

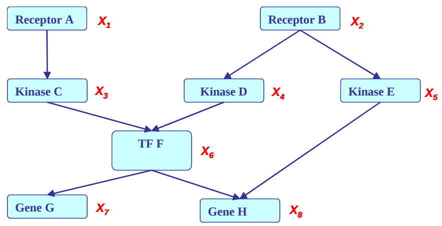
Qualitative Specification

- Where does the qualitative specification come from?
  - Prior knowledge of causal relationships
  - Prior knowledge of modular relationships
  - Assessment from experts
  - Learning from data
  - We simply link a certain architecture (e.g. a layered graph)
  - …

Two types of GMs

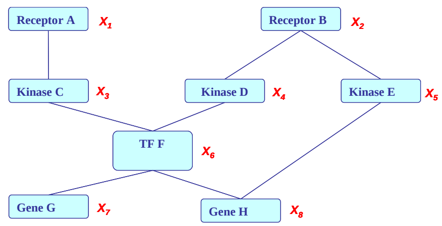
- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):

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



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

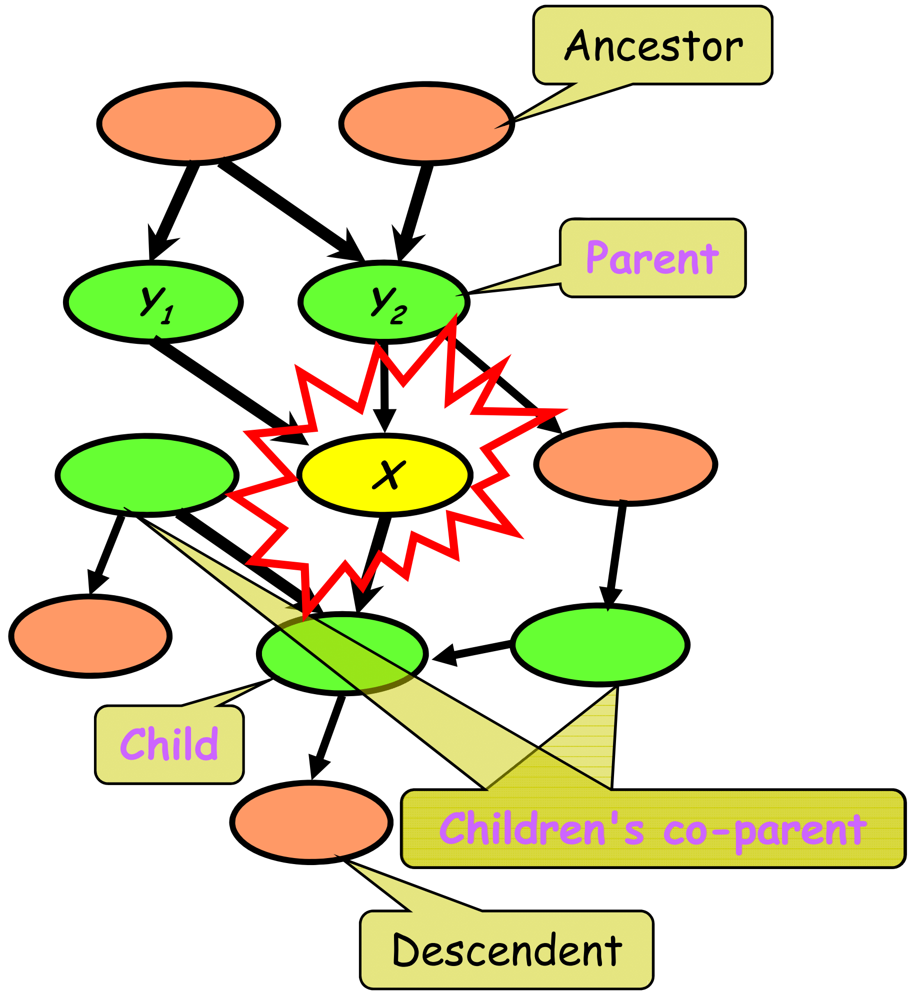
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= 1/Z \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2) + E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}$$



Bayesian Network: Conditional Independence Semantics

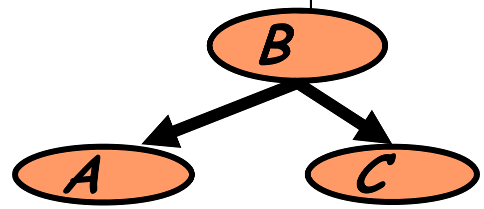
Structure: ***DAG***

- Meaning: a node is **conditionally independent** of every other node in the network outside its **Markov blanket**
- Local conditional distributions (**CPD**) and the **DAG** completely determine the **joint** dist.
- Give **causality** relationships, and facilitate a **generative** process

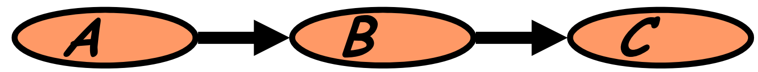


Local Structures & Independencies

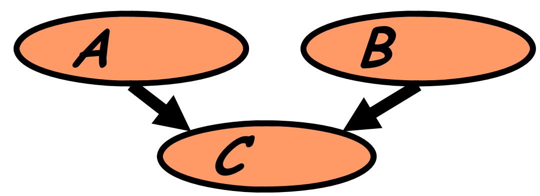
- Common parent
  - Fixing B decouples A and C
    "given the level of gene B, the levels of A and C are independent"



- 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"



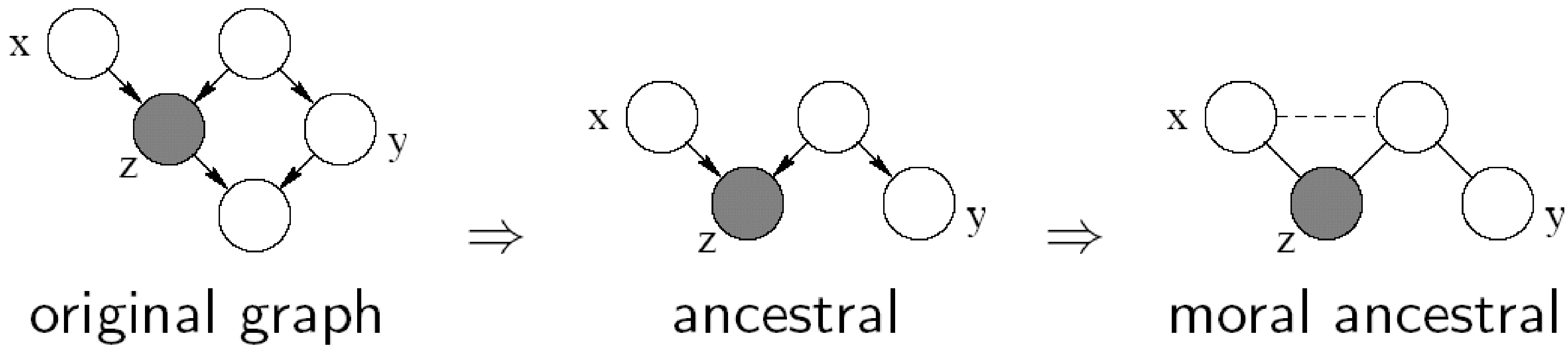
- The language is compact, the concepts are rich!

Graph separation criterion

- D-separation criterion for Bayesian networks (D for Directed edges):

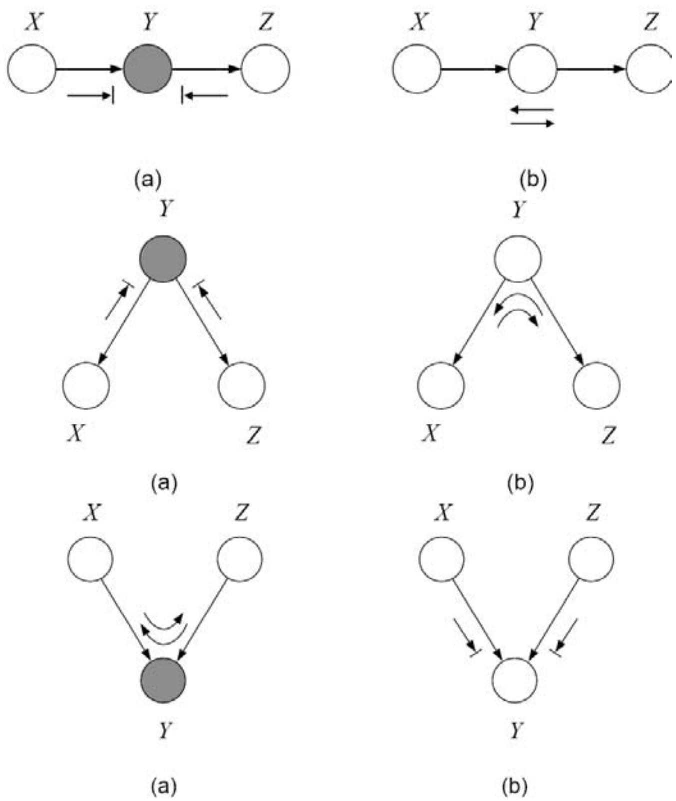
  Definition: variables x and y are *D-separated* (conditionally independent) given z if they are separated in the *moralized* ancestral graph

- Example:



Global Markov properties of DAGs

- 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):

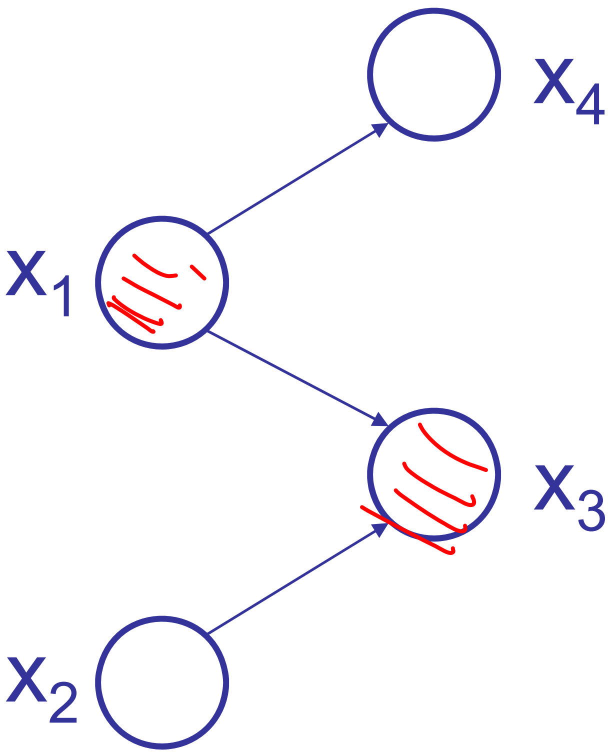


- Defn: $\mathcal{I}(\mathcal{G})$=all independence properties that correspond to d-separation:

$$I(G) = \{X \perp Z | Y : \mathrm{dsep}_G(X; Z | Y)\}$$

- D-separation is sound and complete

Example:



- Complete the I(G) of this graph:

~~X_4~~

$X_1 \perp X_2$

$X_2 \perp X_4$

$X_3 \perp X_4 \mid X_1$

$X_2 \perp X_4 \mid X_3, X_1$

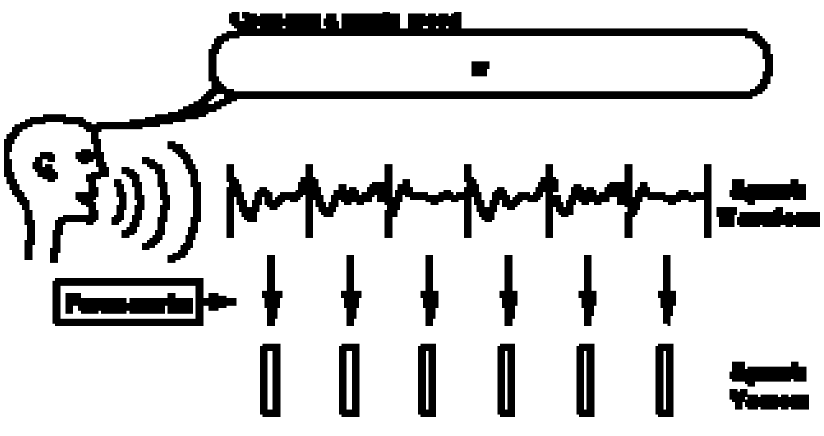
$X_4 \perp X_3 \mid X_1$

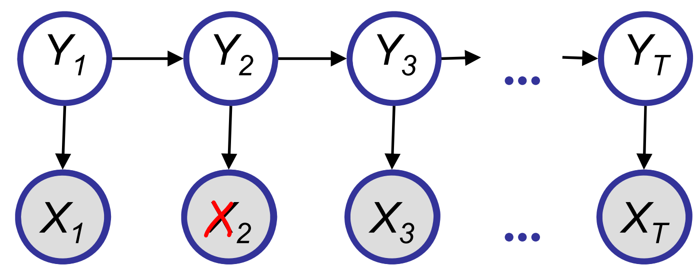
$X_4 \perp X_2 \mid X_1$

$X_4 \perp X_2 X_3 \mid X_1$

Example

- Speech recognition

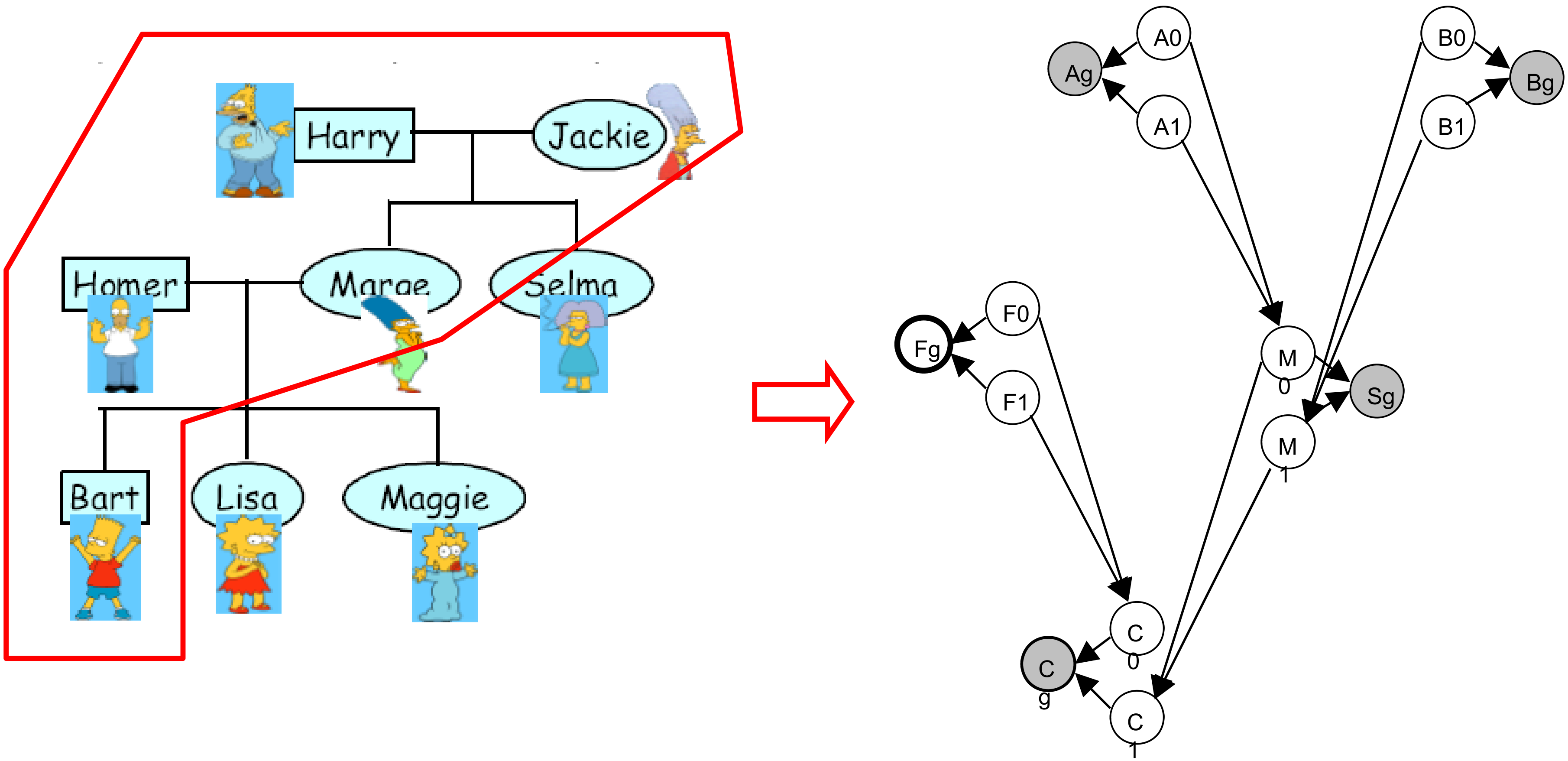




Hidden Markov Model

Example, con'd

- Genetic Pedigree

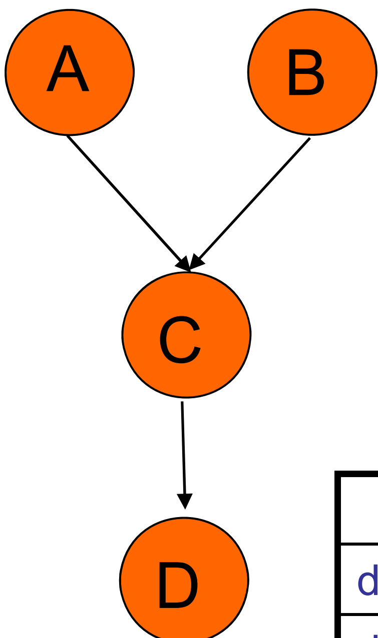


Conditional probability tables (CPTs)

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

$$P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)$$



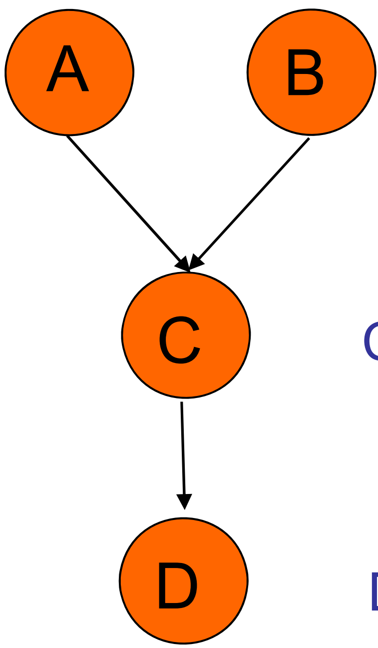
	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

	c^0	c^1
d^0	0.3	0.5
d^1	07	0.5

Conditional probability density func. (CPDs)

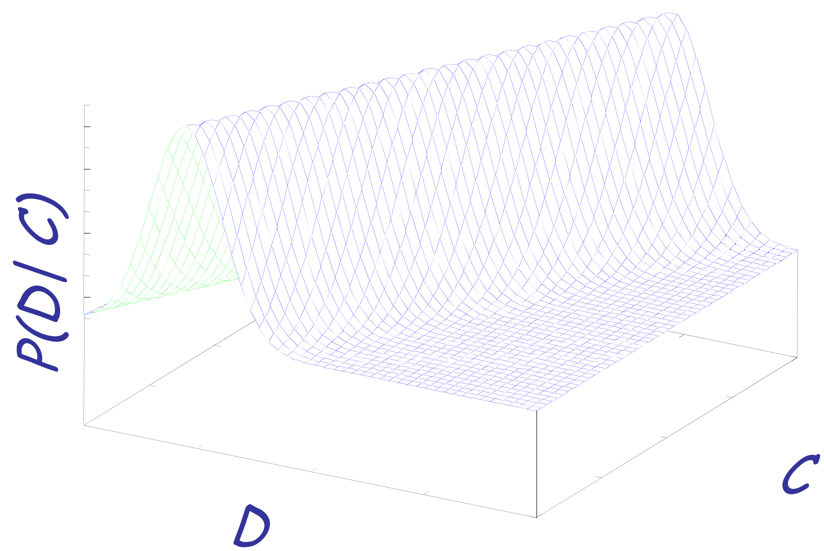
$A \sim N(\mu_a, \Sigma_a)$ $B \sim N(\mu_b, \Sigma_b)$

$P(a,b,c.d) = P(a)P(b)P(c|a,b)P(d|c)$

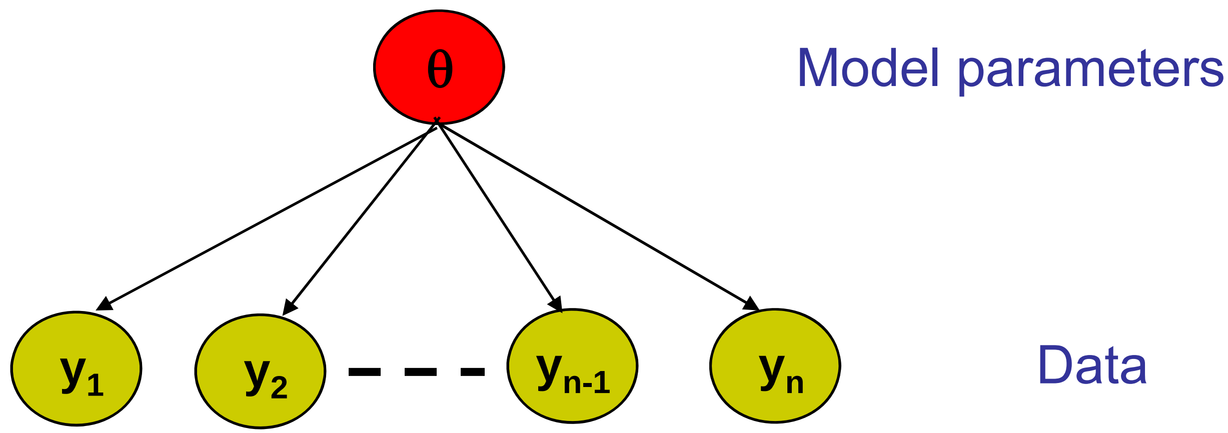


$C \sim N(A+B, \Sigma_c)$

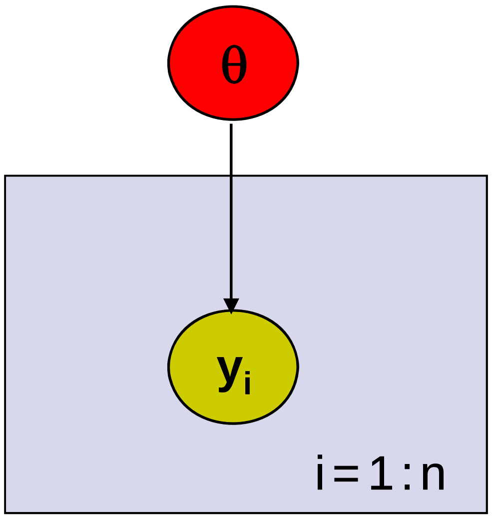
$D \sim N(\mu_a+C, \Sigma_a)$



Conditionally Independent Observations



“Plate” Notation



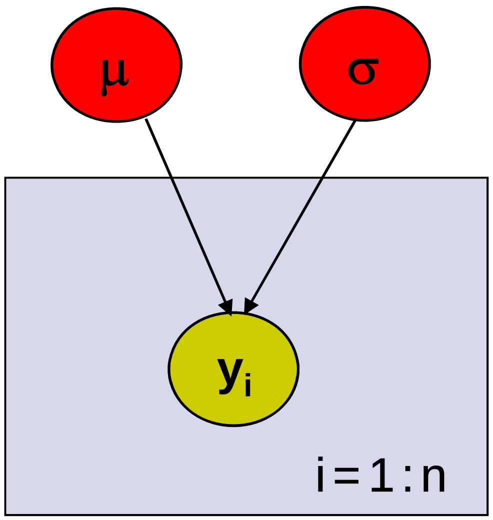
Model parameters

Data = $\{y_1, \ldots y_n\}$

Plate = rectangle in graphical model

variables within a plate are replicated
in a conditionally independent manner

Example: Gaussian Model



Generative model:

$p(y_1,\ldots y_n \mid \mu, \sigma) = P \;\; p(y_i \mid \mu, \sigma)$

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

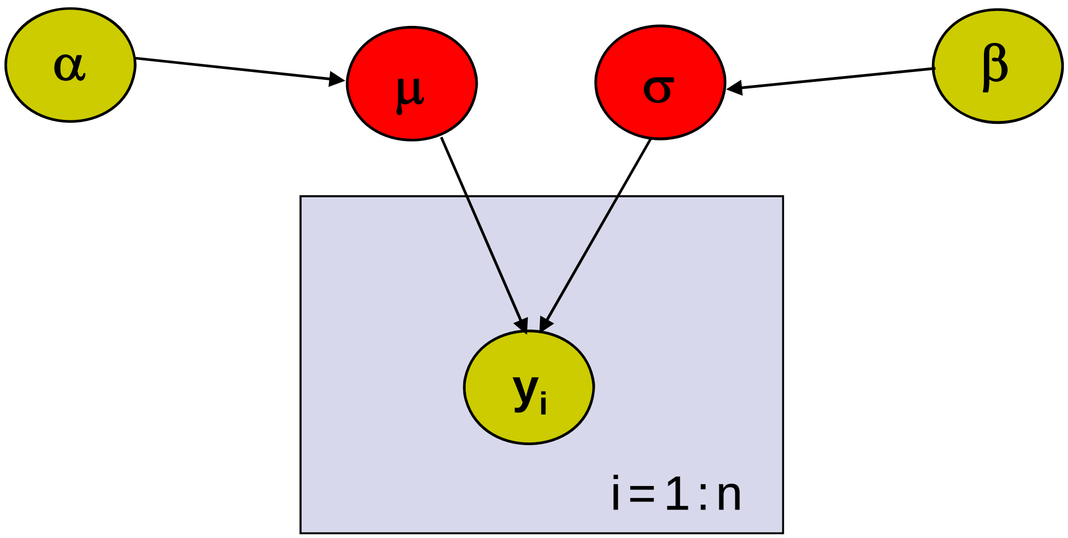
$= \;\; p(D \;\mid \theta)$

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

- Likelihood = p(data | parameters)
  = p(D | θ)
  = L (θ)

- Likelihood tells us how likely the observed data are conditioned on a particular setting of the parameters
  - Often easier to work with log L (θ)

Example: Bayesian Gaussian Model

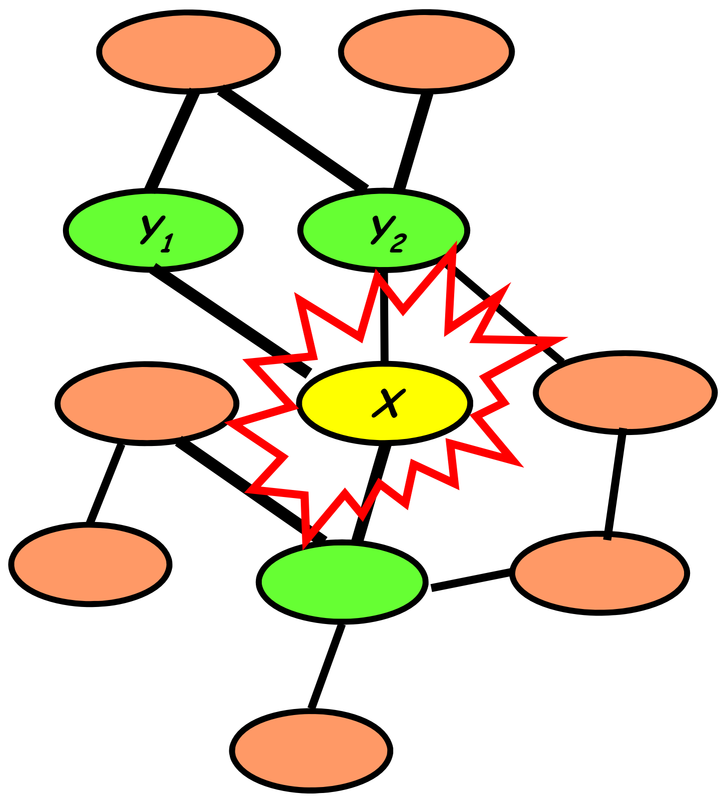


Note: priors and parameters are assumed independent here

Markov Random Fields

Structure: an ***undirected graph***

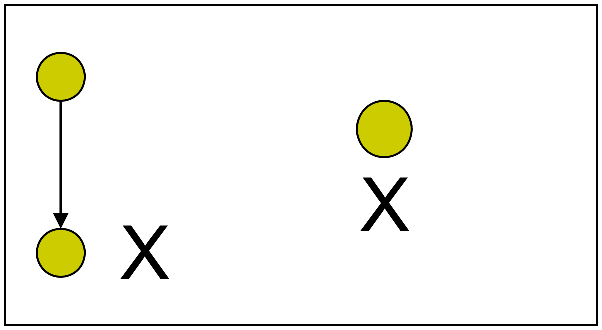
- Meaning: a node is **conditionally independent** of every other node in the network given its **Directed neighbors**
- Local contingency functions (**potentials**) and the **cliques** in the graph completely determine the **joint** dist.
- Give **correlations** between variables, but no explicit way to generate samples



GMs are your old friends

Density estimation

Parametric and nonparametric methods

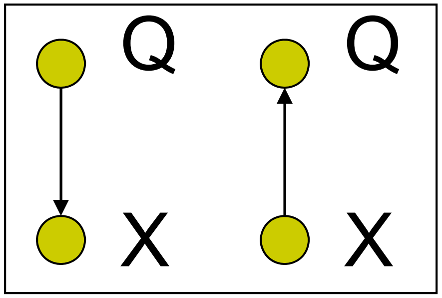


Regression

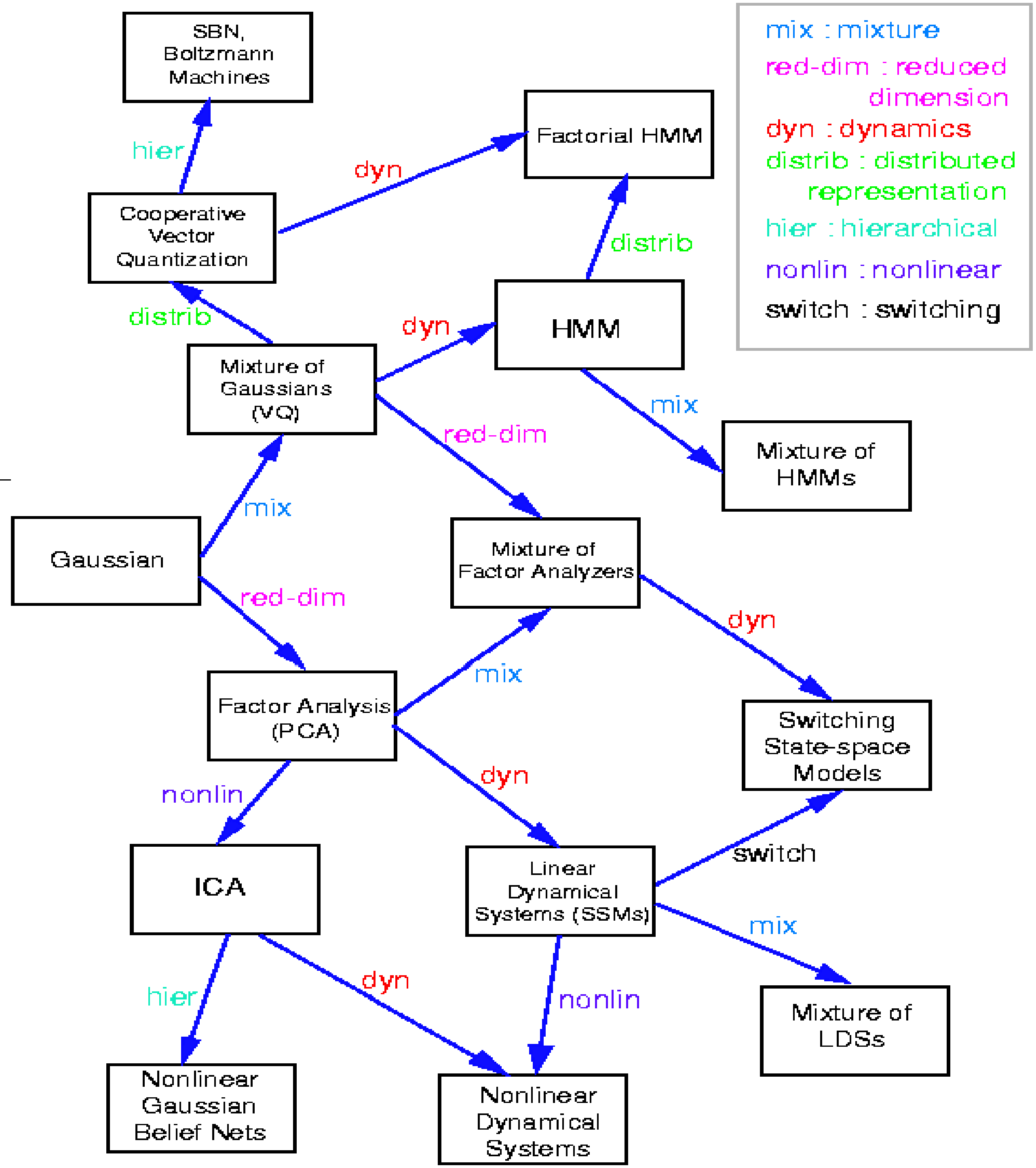
Linear, conditional mixture, nonparametric

Classification

Generative and discriminative approach



An (incomplete) genealogy of graphical models



(Picture by Zoubin Ghahramani and Sam Roweis)

Why graphical models

- **Probability theory** provides the **glue** whereby the parts are combined, ensuring that the system as a whole is consistent, and providing ways to interface models to data.
- The **graph theoretic** side of graphical models provides both an intuitively appealing interface by which humans can model highly-interacting sets of variables as well as a data structure that lends itself naturally to the design of efficient general-purpose algorithms.
- **Many of the classical multivariate probabilistic systems** studied in fields such as statistics, systems engineering, information theory, pattern recognition and statistical mechanics **are special cases of the general graphical model formalism**

  -- examples include mixture models, factor analysis, hidden Markov models, Kalman filters and Ising models.

- The graphical model framework provides a way to view all of these systems as instances of a **common underlying formalism**.

--- M. Jordan