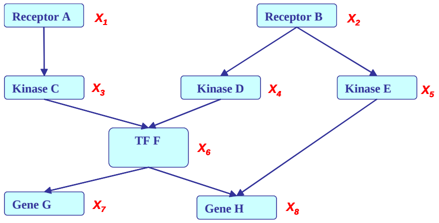
Machine Learning

Graphical Models

Eric Xing

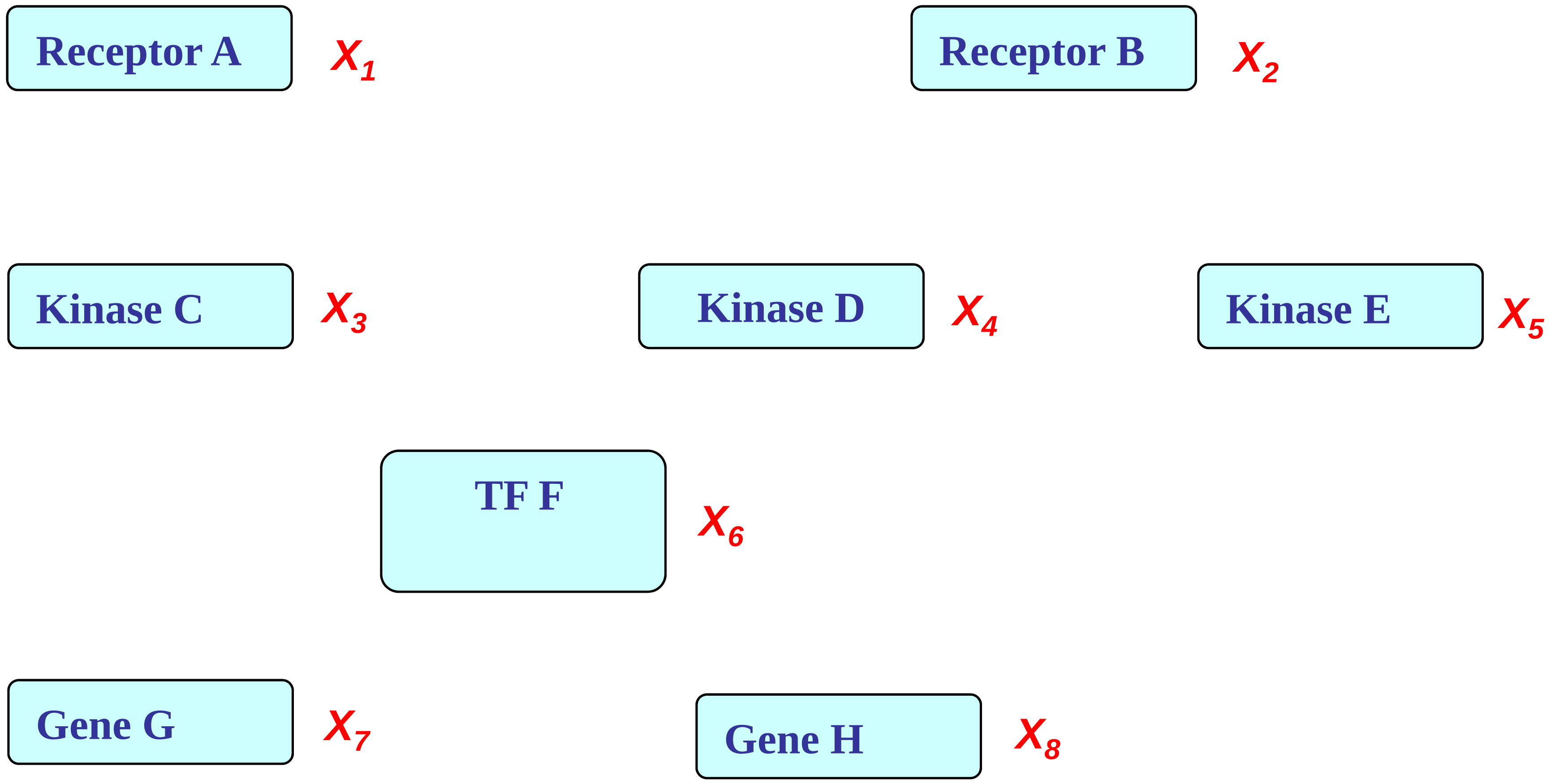
Lecture 10, August 14, 2010



Reading:

Multivariate Distribution in High-D Space

- A possible world for cellular signal transduction:



Receptor A X_1

Receptor B X_2

Kinase C X_3

Kinase D X_4

Kinase E X_5

TF F X_6

Gene G X_7

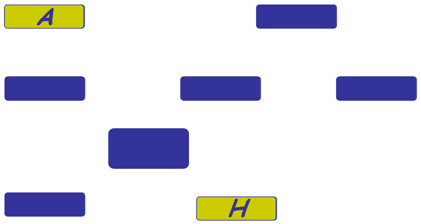
Gene H X_8

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?**



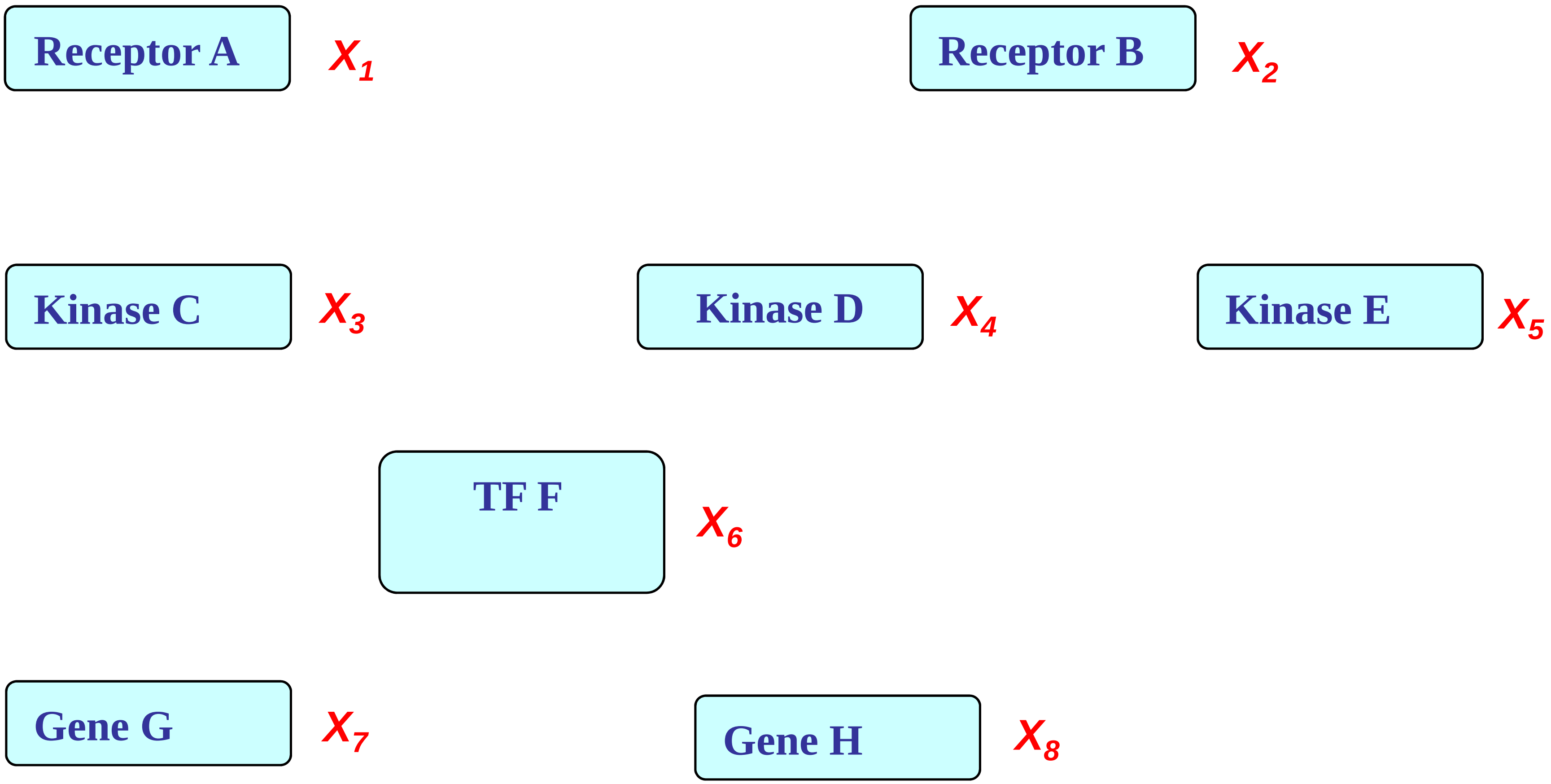
- Learning: where do we get all this probabilities?
  - Maximal-likelihood estimation? but how many data do we need?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?

- 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:



Receptor A X_1

Receptor B X_2

Kinase C X_3

Kinase D X_4

Kinase E X_5

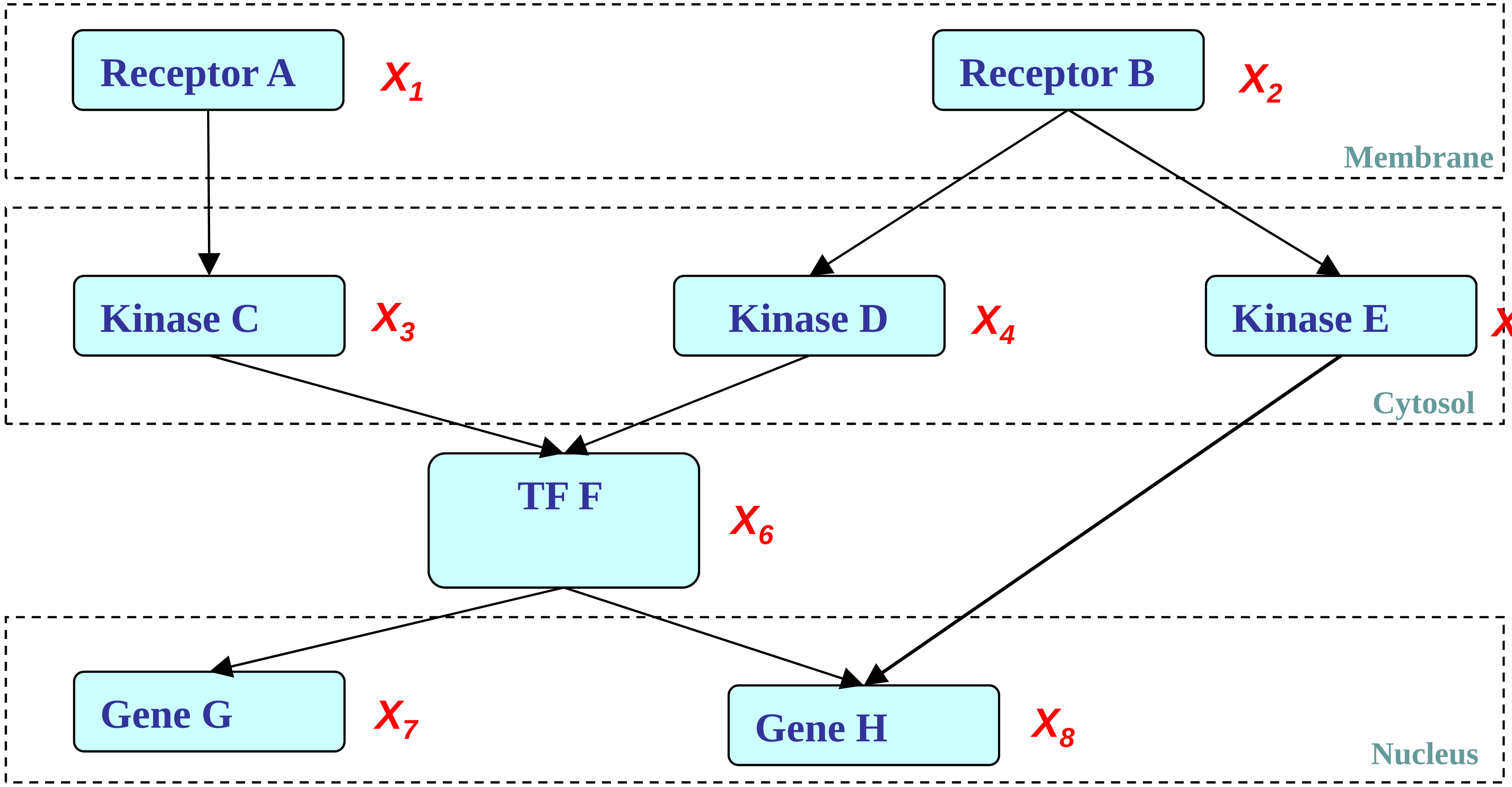
TF F X_6

Gene G X_7

Gene H X_8

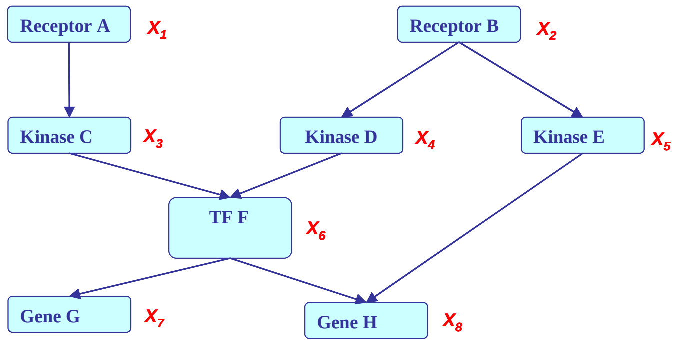
GM: Structure Simplifies Representation

- 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.,



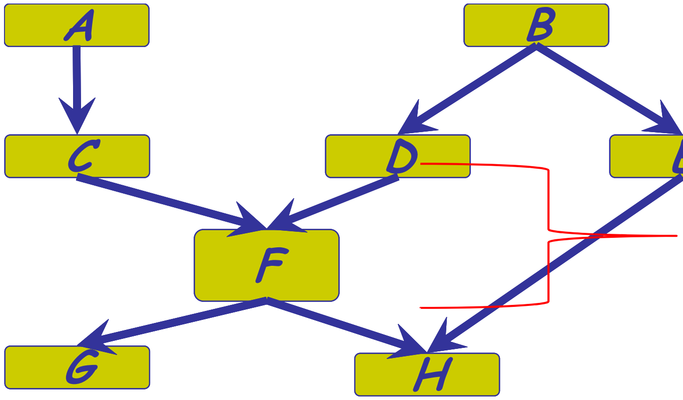
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

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

    2+2+4+4+4+8+4+8=36, 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	$P(F \mid C,D)$	
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

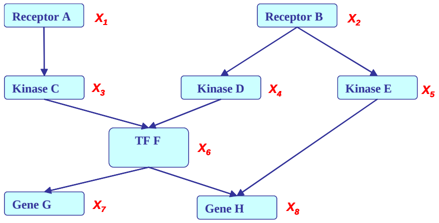
Qualitative Specification

- 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

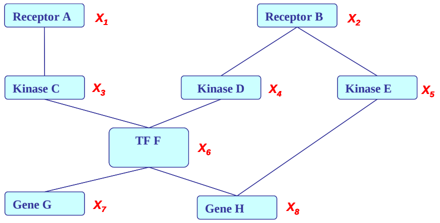
- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):

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



- 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:

- A BN is a directed graph whose nodes represent the random variables and whose edges represent direct influence of one variable on another.

- It is a data structure that provides the skeleton for representing **a joint distribution** compactly in a **factorized** way;

- It offers a compact representation for **a set of conditional independence assumptions** about a distribution;

- We can view the graph as encoding a generative sampling process executed by nature, where the value for each variable is selected by nature using a distribution that depends only on its parents. In other words, each variable is a stochastic function of its parents.

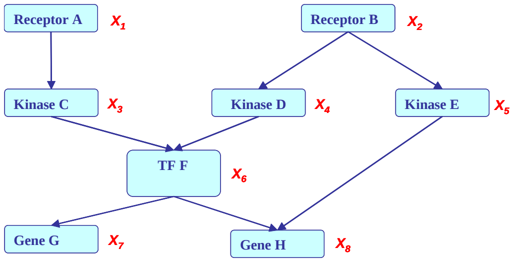
Bayesian Network: Factorization Theorem

- **Theorem:**

  Given a DAG, The most general form of the probability distribution that is consistent with the graph factors according to “node given its parents”:

$$P(\mathbf{X}) = \prod_{i=1:d} P(X_i \mid \mathbf{X}_{\pi_i})$$

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



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

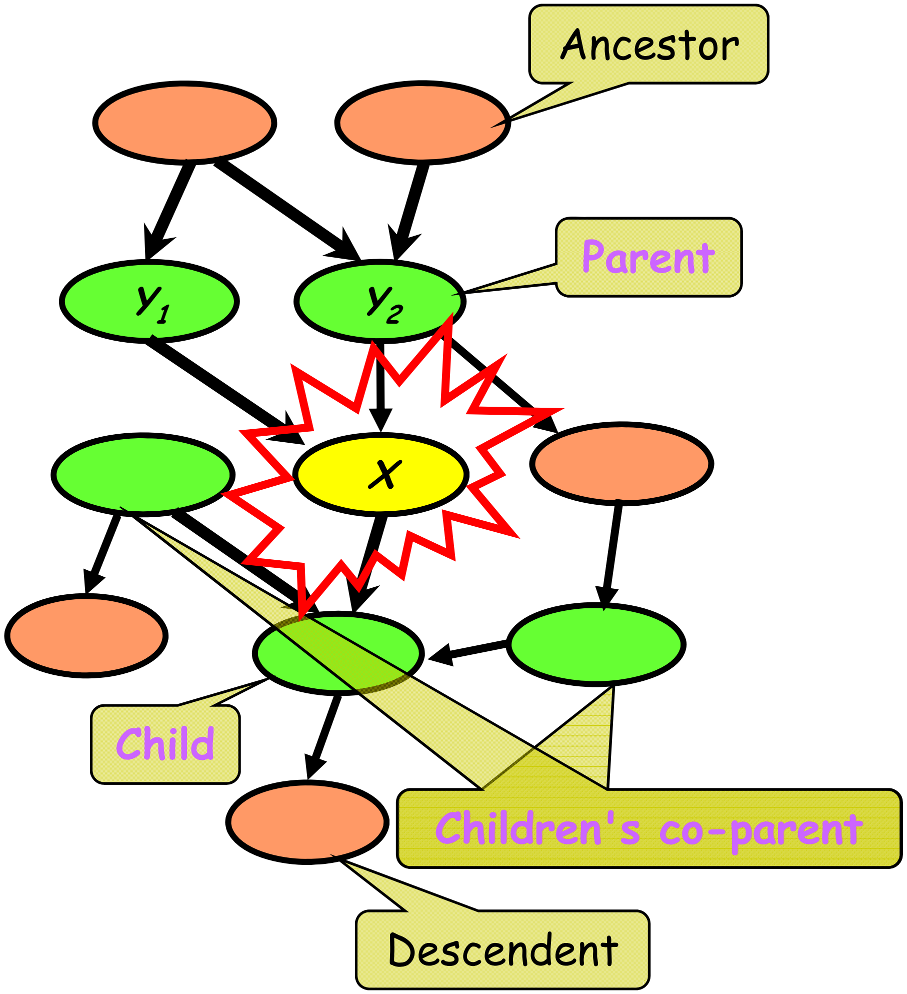
Bayesian Network: Conditional Independence Semantics

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.

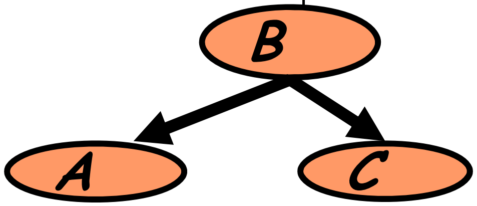
- Give **causality** relationships, and facilitate a **generative** process



Local Structures & Independencies

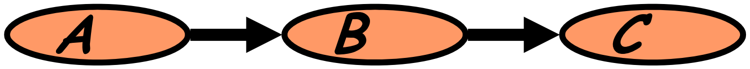
- Common parent
  - Fixing B decouples A and C

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



- Cascade
  - Knowing B decouples A and C

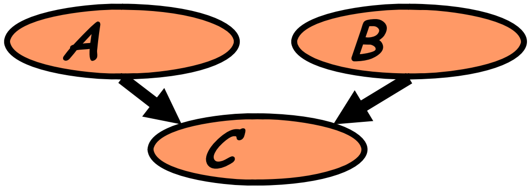
    "given the level of gene B, the level gene A provides no extra prediction value for the level of gene C"



- V-structure
  - Knowing C couples A and B

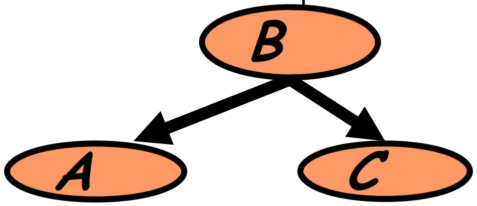
    because A can "explain away" B w.r.t. C

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



- The language is compact, the concepts are rich!

A simple justification

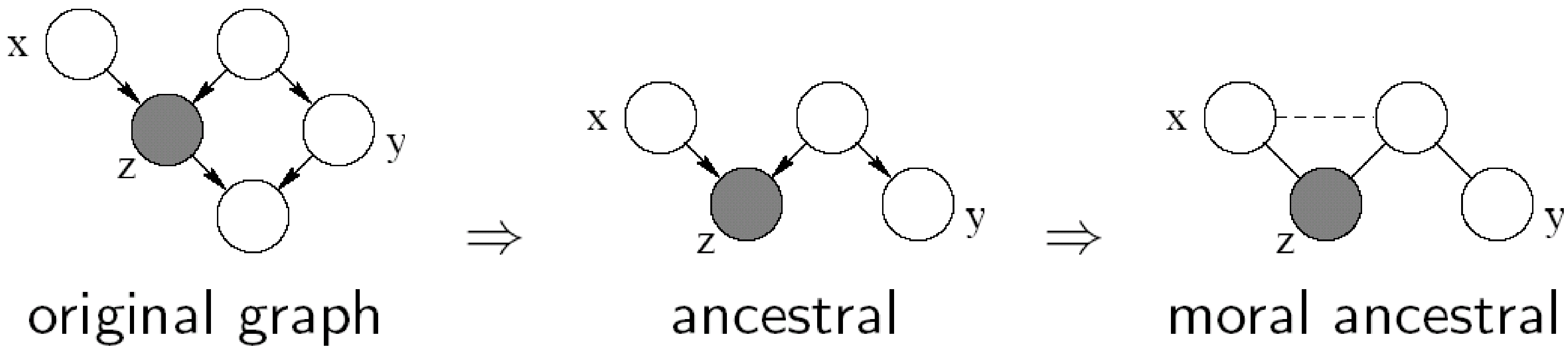


Graph separation criterion

- 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:



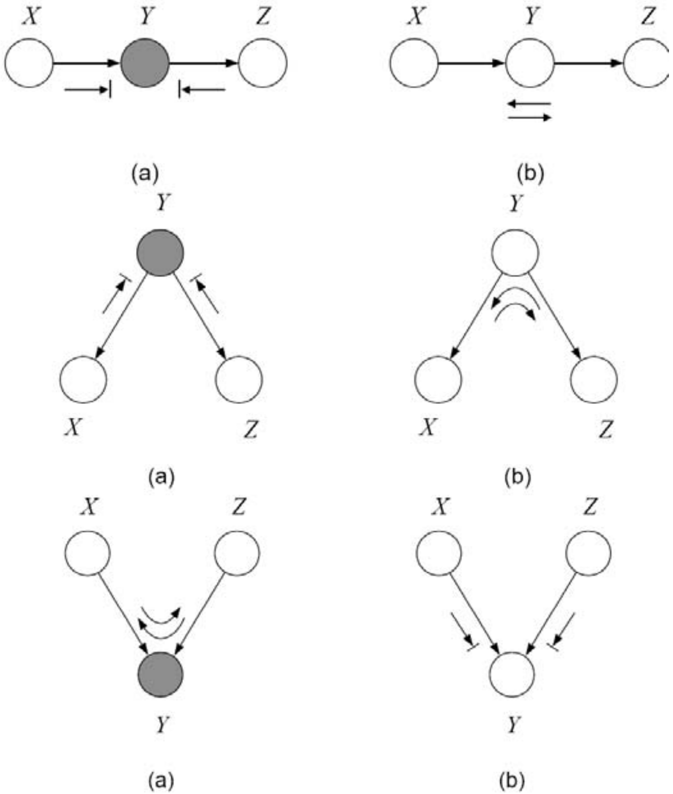
original graph

ancestral

moral ancestral

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):

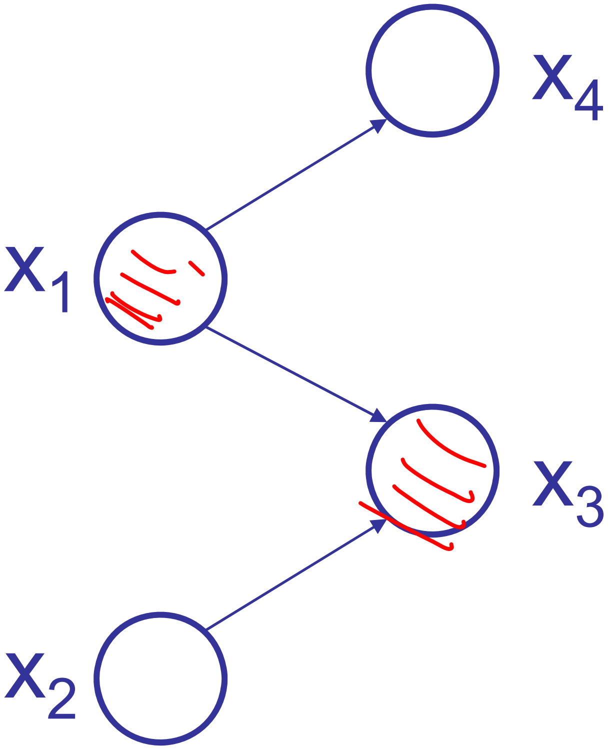


- Defn: $\mathcal{I}(\mathcal{G})$=all independence properties that correspond to d-separation:

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

- D-separation is sound and complete

Example:



- Complete the I(G) of this graph:

X_4.

$X_1 \perp X_2$.

$X_2 \perp X_4$

$X_3 \perp X_4 \mid X_1$

$X_2 \perp X_4 \mid X_3, X_1$.

$X_4 \perp X_3 \mid X_1$

$X_4 \perp X_2 \mid X_1$.

$X_4 \perp X_2 X_3 \mid X_1$

Towards quantitative 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

- **The Equivalence Theorem**

  For a graph G,

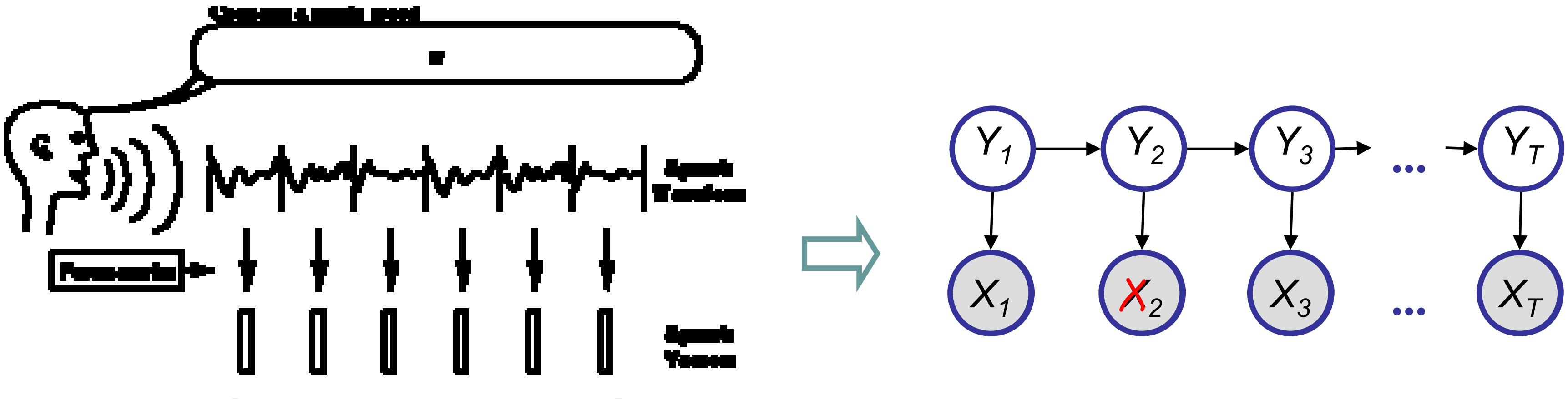
  Let $\mathcal{D}_1$ denote the family of all distributions that satisfy I(G),

  Let $\mathcal{D}_2$ denote the family of all distributions that factor according to G,

  Then $\mathcal{D}_1 \equiv \mathcal{D}_2$.

Example

- Speech recognition



Hidden Markov Model

Knowledge Engineering

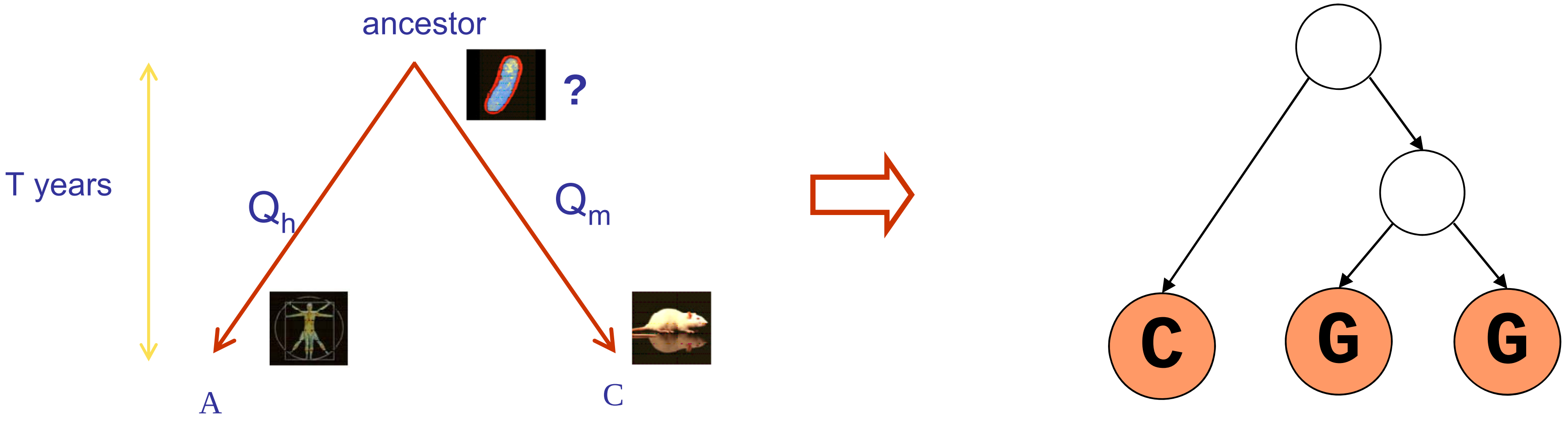
- Picking variables
  - Observed
  - Hidden

- Picking structure
  - CAUSAL
  - Generative

- Picking Probabilities
  - Zero probabilities
  - Orders of magnitudes
  - Relative values

Example, con'd

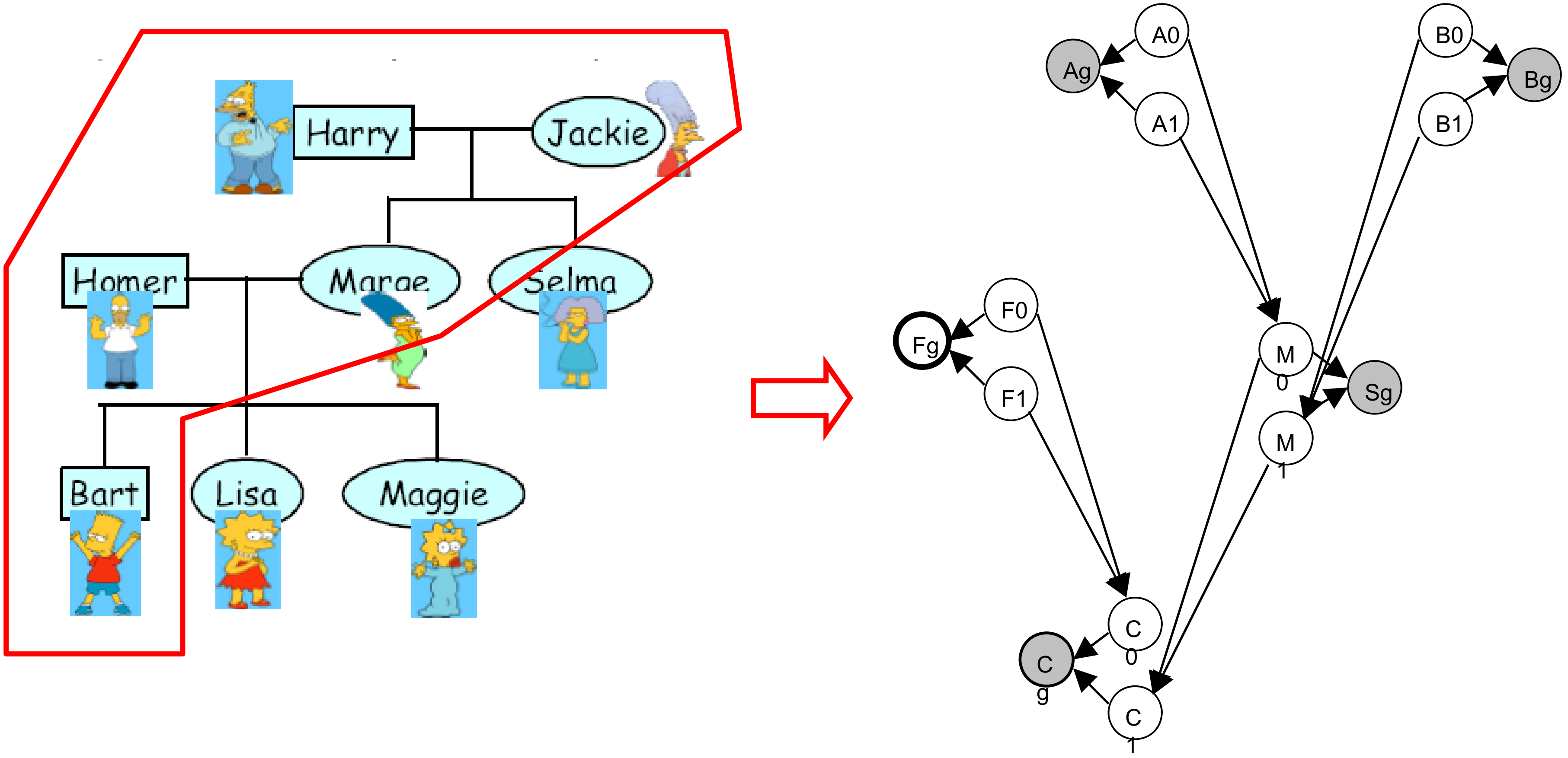
- Evolution



Tree Model

Example, con'd

- Genetic Pedigree

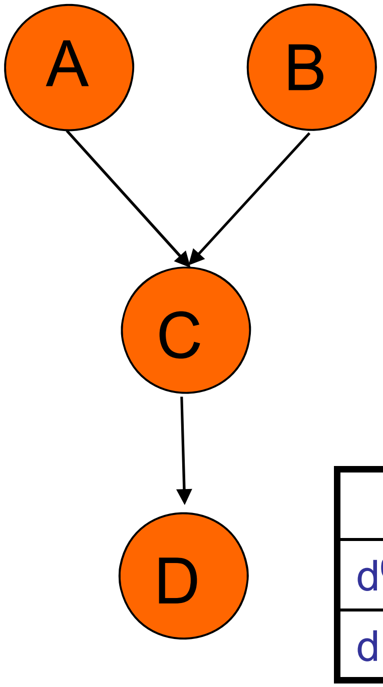


Conditional probability tables (CPTs)

a^0	0.75
a^1	0.25

b^0	0.33
b^1	0.67

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



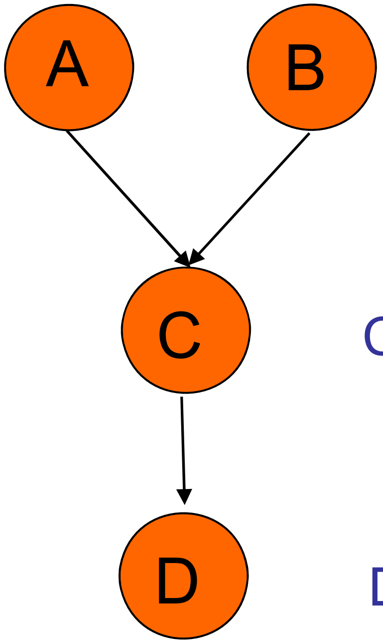
	a^0b^0	a^0b^1	a^1b^0	a^1b^1
c^0	0.45	1	0.9	0.7
c^1	0.55	0	0.1	0.3

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

Conditional probability density func. (CPDs)

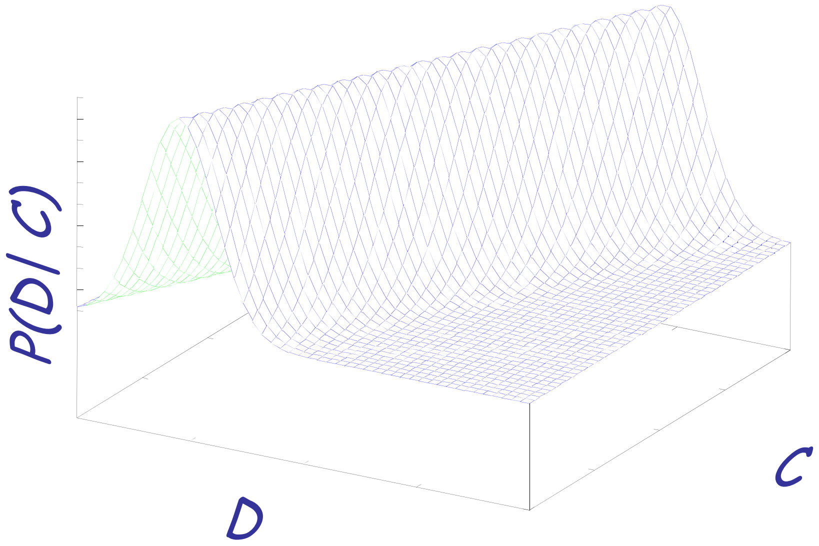
$A \sim N(\mu_a, \Sigma_a)$ $B \sim N(\mu_b, \Sigma_b)$

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

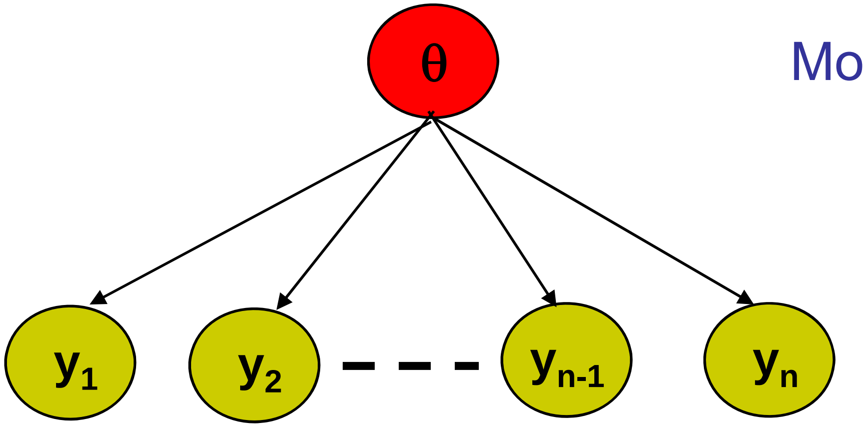


$C \sim N(A+B, \Sigma_c)$

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



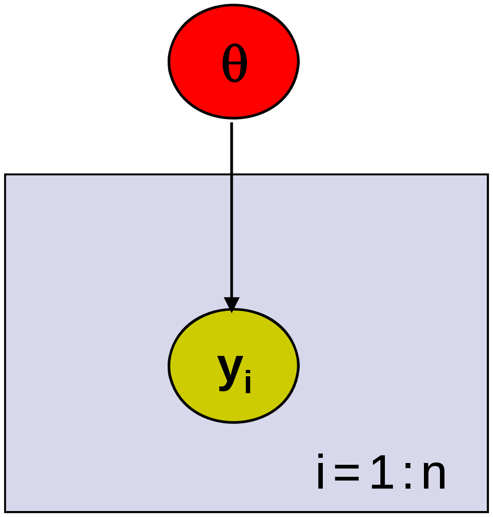
Conditionally Independent Observations



Model parameters

Data

“Plate” Notation



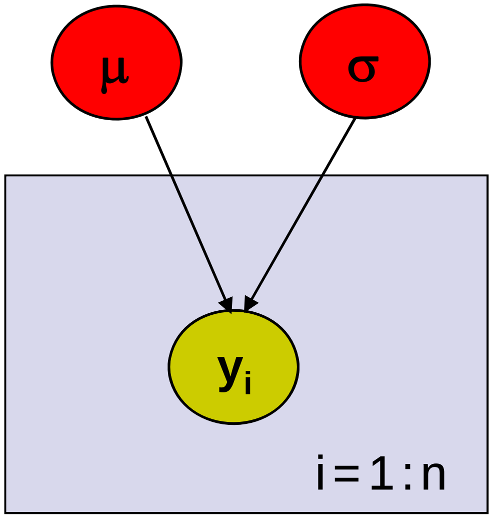
Model parameters

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



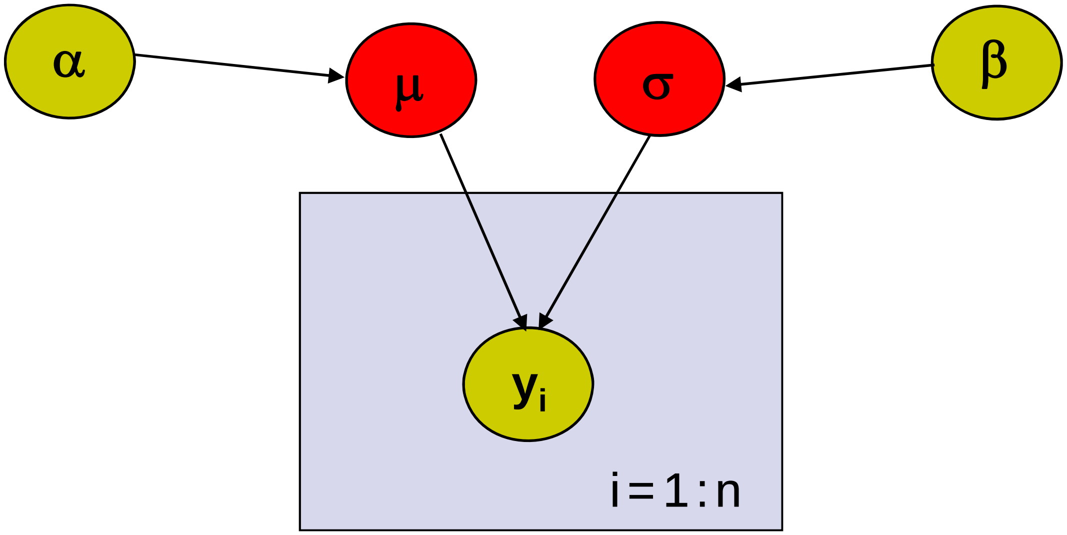
Generative model:

$$p(y_1,\ldots y_n \mid \mu, \sigma) = P \ \ p(y_i \mid \mu, \sigma)$$
$$= \ \ p(\text{data} \mid \text{parameters})$$
$$= \ \ p(D \mid \theta)$$

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

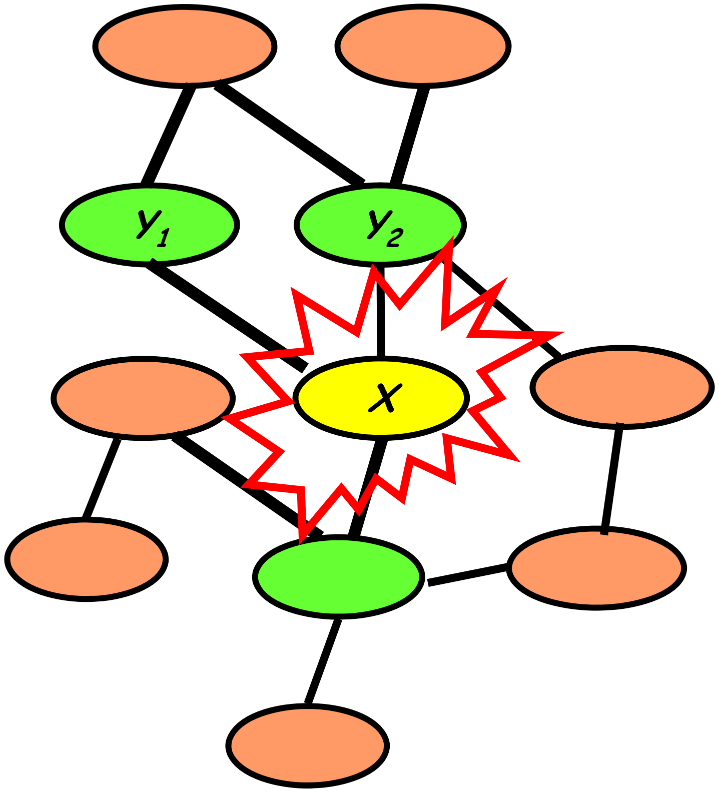
Markov Random Fields

Structure: an ***undirected graph***

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

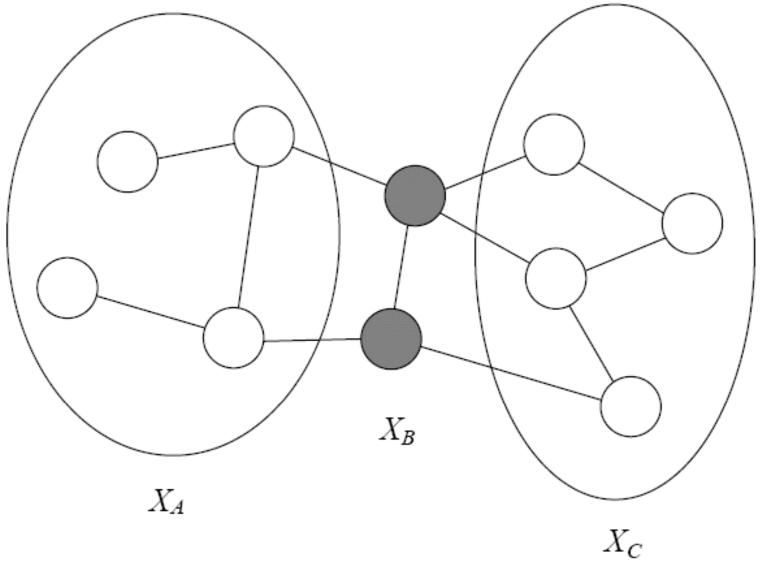
- Local contingency functions (**potentials**) and the **cliques** in the graph completely determine the **joint** dist.

- Give **correlations** between variables, but no explicit way to generate samples



Global Markov property

- Let *H* be an undirected graph:



- *B* ***separates*** *A* and *C* if every path from a node in *A* to a node in *C* passes through a node in *B*: $\mathrm{sep}_H(A;C|B)$
- A probability distribution satisfies the ***global Markov property*** if for any disjoint *A*, *B*, *C*, such that *B* separates *A* and *C*, *A* is independent of *C* given *B:* $\mathrm{I}(H)=\{A\perp C|B):\mathrm{sep}_H(A;C|B)\}$

Soundness and completeness of global Markov property

- Defn: An UG H is an I-map for a distribution P if $I(H) \subseteq I(P)$, i.e., P entails $I(H)$.
- Defn: P is a Gibbs distribution over H if it can be represented as

$$P(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{c \in C} \psi_c(\mathbf{x}_c)$$

- Thm (soundness): If P is a Gibbs distribution over H, then H is an I-map of P.
- Thm (completeness): If $\neg\mathrm{sep}_H(X; Z \mid Y)$, then $X \not\perp_P Z \mid Y$ in ***some*** P that factorizes over H.

Representation

- Defn: an undirected graphical model represents a distribution $P(X_1 ,\ldots,X_n)$ defined by an undirected graph H, and a set of positive ***potential functions*** y_c associated with cliques of H, s.t.

$$P(x_1,\ldots,x_n) = \frac{1}{Z}\prod_{c\in C}\psi_c(\mathbf{x}_c)$$

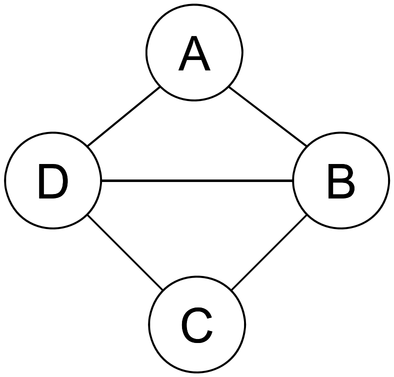
where Z is known as the partition function:

$$Z = \sum_{x_1,\ldots,x_n}\prod_{c\in C}\psi_c(\mathbf{x}_c)$$

- Also known as Markov Random Fields, Markov networks …
- The ***potential function*** can be understood as an contingency function of its arguments assigning "pre-probabilistic" score of their joint configuration.

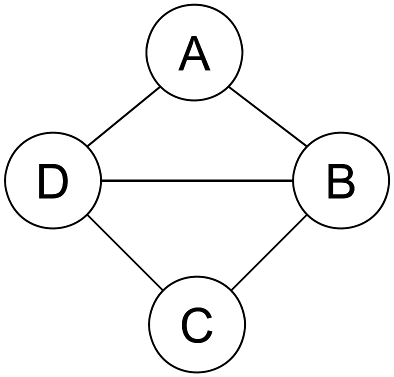
Cliques

- For *G*={*V*,*E*}, a complete subgraph (clique) is a subgraph *G'*={*V'*Í*V*,*E'*Í*E*} such that nodes in *V'* are fully interconnected
- A (maximal) clique is a complete subgraph s.t. any superset *V''*É*V'* is not complete.
- A sub-clique is a not-necessarily-maximal clique.



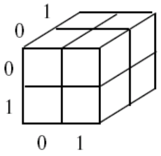
- Example:
  - max-cliques = {*A*,*B*,*D*}, {*B*,*C*,*D*},
  - sub-cliques = {*A*,*B*}, {*C*,*D*}, …→ all edges and singletons

Example UGM – using max cliques



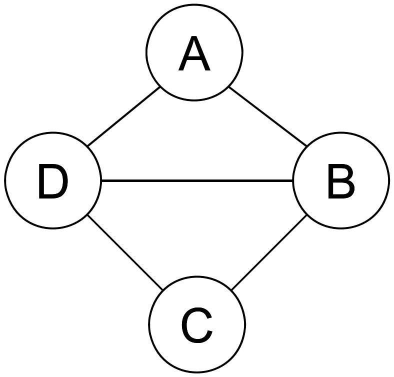
$$P(x_1,x_2,x_3,x_4)=\frac{1}{Z}\psi_c(\mathbf{x}_{124})\times\psi_c(\mathbf{x}_{234})$$

$$Z=\sum_{x_1,x_2,x_3,x_4}\psi_c(\mathbf{x}_{124})\times\psi_c(\mathbf{x}_{234})$$



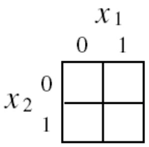
- For discrete nodes, we can represent $P(X_{1:4})$ as two 3D tables instead of one 4D table

Example UGM – using subcliques



$$P(x_1, x_2, x_3, x_4) = \frac{1}{Z} \prod_{ij} \psi_{ij}(\mathbf{x}_{ij})$$

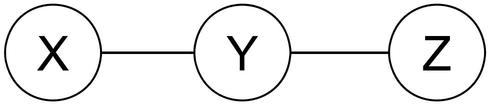
$$= \frac{1}{Z} \psi_{12}(\mathbf{x}_{12}) \psi_{14}(\mathbf{x}_{14}) \psi_{23}(\mathbf{x}_{23}) \psi_{24}(\mathbf{x}_{24}) \psi_{34}(\mathbf{x}_{34})$$



$$Z = \sum_{x_1, x_2, x_3, x_4} \prod_{ij} \psi_{ij}(\mathbf{x}_{ij})$$

- For discrete nodes, we can represent *P*(*X*$_{1:4}$) as 5 2D tables instead of one 4D table

Interpretation of Clique Potentials



- The model implies *X*⊥*Z*|*Y*. This independence statement implies (by definition) that the joint must factorize as:

$$p(x,y,z) = p(y)p(x \mid y)p(z \mid y)$$

- We can write this as: $p(x,y,z) = p(x,y)p(z \mid y)$, but $p(x,y,z) = p(x \mid y)p(z,y)$
  - **cannot** have all potentials be marginals
  - **cannot** have all potentials be conditionals

- The positive clique potentials can only be thought of as general "compatibility", "goodness" or "happiness" functions over their variables, but not as probability distributions.

Exponential Form

- Constraining clique potentials to be positive could be inconvenient (e.g., the interactions between a pair of atoms can be either attractive or repulsive). We represent a clique potential $\psi_c(\mathbf{x}_c)$ in an unconstrained form using a real-value "energy" function $\phi_c(\mathbf{x}_c)$:

$$\psi_c(\mathbf{x}_c) = \exp\{-\phi_c(\mathbf{x}_c)\}$$

For convenience, we will call $\phi_c(\mathbf{x}_c)$ a potential when no confusion arises from the context.

- This gives the joint a nice additive strcuture

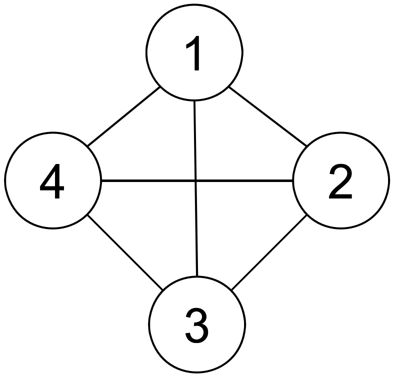
$$p(\mathbf{x}) = \frac{1}{Z}\exp\left\{-\sum_{c\in C}\phi_c(\mathbf{x}_c)\right\} = \frac{1}{Z}\exp\{-H(\mathbf{x})\}$$

where the sum in the exponent is called the "free energy":

$$H(\mathbf{x}) = \sum_{c\in C}\phi_c(\mathbf{x}_c)$$

- In physics, this is called the "Boltzmann distribution".
- In statistics, this is called a log-linear model.

Example: Boltzmann machines



- A fully connected graph with pairwise (edge) potentials on binary-valued nodes (for $x_i \in \{-1,+1\}$ or $x_i \in \{0,1\}$) is called a Boltzmann machine

$$P(x_1,x_2,x_3,x_4) = \frac{1}{Z}\exp\left\{\sum_{ij}\phi_{ij}(x_{i,}x_j)\right\}$$

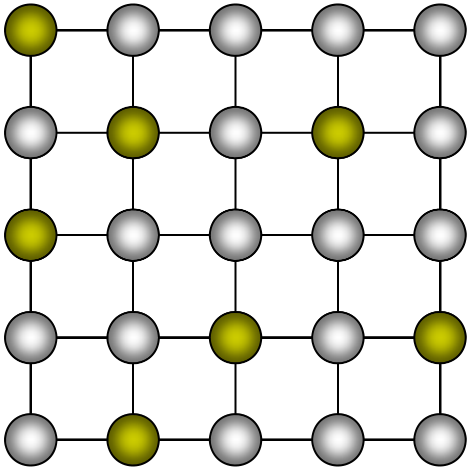
$$= \frac{1}{Z}\exp\left\{\sum_{ij}\theta_{ij}x_ix_j + \sum_i \alpha_i x_i + C\right\}$$

- Hence the overall energy function has the form:

$$H(x) = \sum_{ij}(x_i-\mu)\Theta_{ij}(x_j-\mu) = (x-\mu)^T\Theta(x-\mu)$$

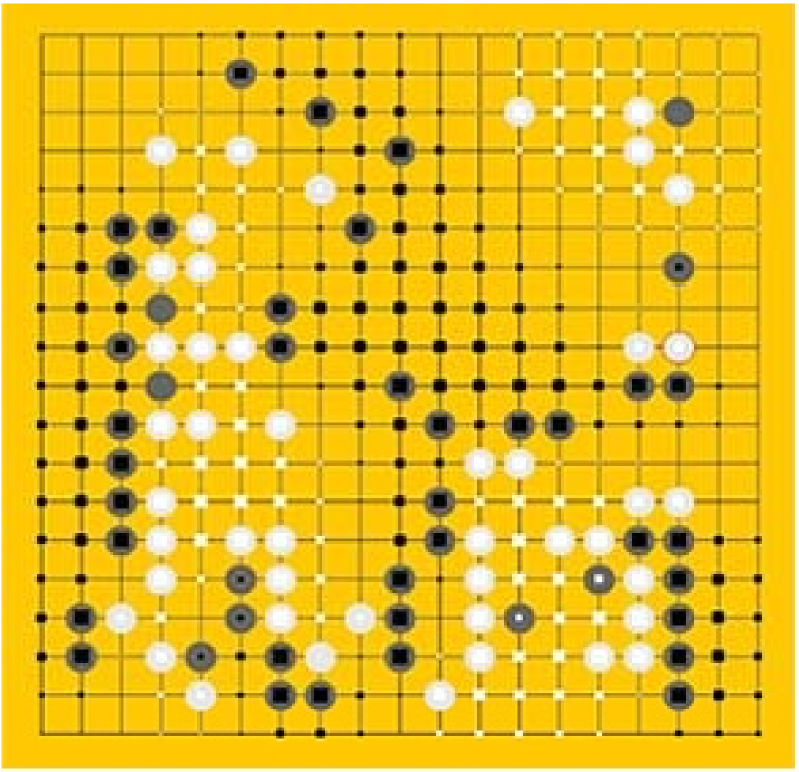
Example: Ising (spin-glass) models

- Nodes are arranged in a regular topology (often a regular packing grid) and connected only to their geometric neighbors.



- Same as sparse Boltzmann machine, where $\theta_{ij} \neq 0$ iff *i,j* are neighbors.
  - e.g., nodes are pixels, potential function encourages nearby pixels to have similar intensities.
- Potts model: multi-state Ising model.

Example: Modeling Go

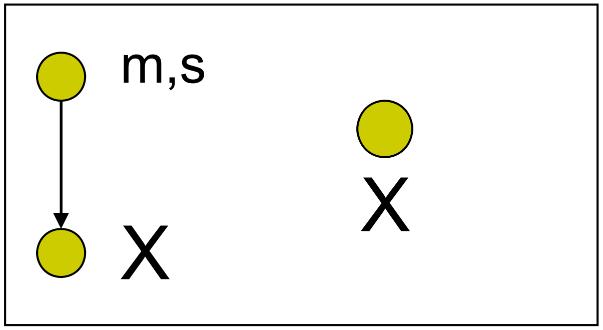


This is the middle position of a Go game. Overlaid is the estimate for the probability of becoming black or white for every intersection. Large squares mean the probability is higher.

GMs are your old friends

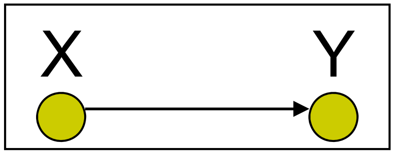
Density estimation

Parametric and nonparametric methods



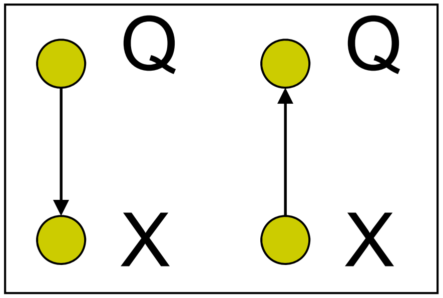
Regression

Linear, conditional mixture, nonparametric

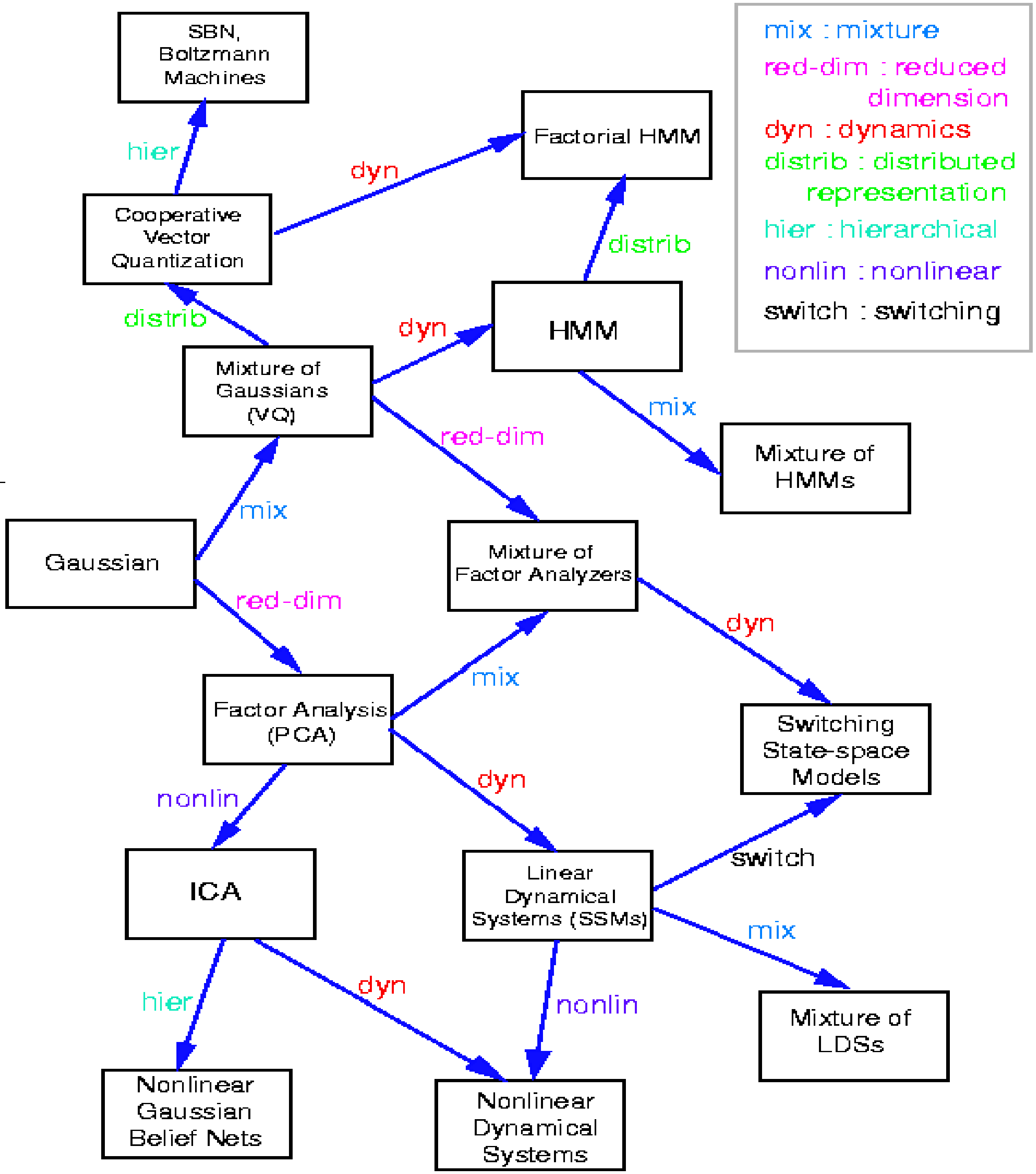


Classification

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