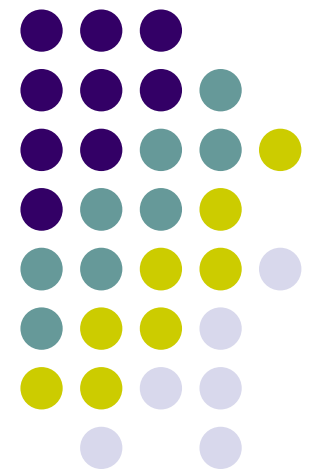




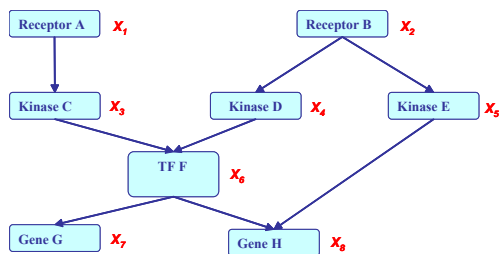
# Probabilistic Graphical Models

## Introduction to GM



Eric Xing

Lecture 1, January 11, 2016



Reading: see class homepage

# Logistics



- Class webpage:
  - <http://www.cs.cmu.edu/~epxing/Class/10708-16/>

The screenshot shows a Mozilla Firefox browser window displaying the class webpage. The address bar shows the URL [www.cs.cmu.edu/~epxing/Class/10708-16/](http://www.cs.cmu.edu/~epxing/Class/10708-16/). The page has a navigation menu with links: Home, Description, People, Lectures, Recitations, Homework, Project, and Previous. The main content area features the Carnegie Mellon University logo, the title "Probabilistic Graphical Models", and the semester "10-708, Spring 2016". Below this, the instructors "Eric Xing" and "Matthew Gormley" are listed, along with their affiliation: "School of Computer Science, Carnegie Mellon University". A blue horizontal bar separates the header from the class details. The details include:

- **Time:** Monday, Wednesday 12:00-1:20 pm
- **Location:** GHC 4307
- **Recitations:** TBD

A link is provided for lecture videos: "Lecture videos of PGM (Spring 2014) can be found [here](#)." Another blue horizontal bar separates the details from the "Announcements" section. The announcements include:

- Class begins on Monday, 01-11-16. See you in class!
- If you are on the waiting list and have not been granted registration, please come to the class on the first day to learn more about your chance of getting enrolled in the class. We are working on getting everyone registered but can not guarantee that at this point.
- Class discussions will be run through [Plazza](#).
- The class mailing list is [10708-students@cs.cmu.edu](mailto:10708-students@cs.cmu.edu). The instructor mailing list is [10708-instructor@cs.cmu.edu](mailto:10708-instructor@cs.cmu.edu).

The footer contains the copyright notice: "© 2016 Eric Xing @ School of Computer Science, Carnegie Mellon University" and the date "Last updated 01/09/2016" with a link to "validate xhtml".

# Logistics



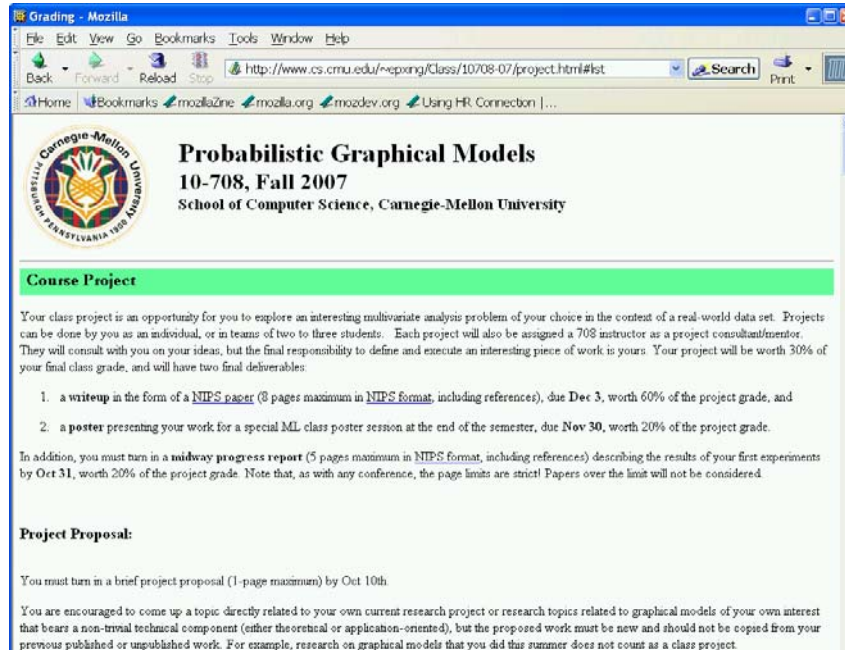
- Text books:
  - Daphne Koller and Nir Friedman, **Probabilistic Graphical Models**
  - M. I. Jordan, **An Introduction to Probabilistic Graphical Models**
- Mailing Lists:
  - To contact the instructors: [10708-instructor@cs.cmu.edu](mailto:10708-instructor@cs.cmu.edu)
  - Class announcements list: [10708-students@cs.cmu.edu](mailto:10708-students@cs.cmu.edu).
- TA:
  - Calvin McCarter, GHC 8008, Office hours: TBA
  - Jin Kyu Kim, GHC 6219, Office hours: TBA
  - Yuntian Deng, GHC , Office hours: TBA
  - Zhiting Hu, GHC 8116, Office hours: TBA
- Lecturers: Eric Xing, Matt Gormley
- Class Assistant:
  - Mallory Deptola, GHC 8001, x8-5527
- Instruction aids: Piazza



# Logistics

- 5 homework assignments: 40% of grade
  - Theory exercises, Implementation exercises
- Scribe duties: 10% (~once to twice for the whole semester)
- Short reading summary: 10% (due at the beginning of every lecture)
- Final project: 40% of grade
  - Applying PGM to the development of a real, substantial ML system
    - Design and Implement a (record-breaking) distributed Logistic Regression, Gradient Boosted Tree, Deep Network, or Topic model on Petuum and apply to ImageNet, Wikipedia, and/or other data
    - Build a web-scale topic or story line tracking system for news media, or a paper recommendation system for conference review matching
    - An online car or people or event detector for web-images and webcam
    - An automatic “what’s up here?” or “photo album” service on iPhone
  - Theoretical and/or algorithmic work
    - a more efficient approximate inference or optimization algorithm, e.g., based on stochastic approximation, proximal average, or other new techniques
    - a distributed sampling scheme with convergence guarantee
  - 3-member team to be formed in the first three weeks, proposal, mid-way report, oral presentation & demo, final report, peer review → possibly conference submission !

# Past projects:



- We will have a prize for the best project(s) ...

- **Award Winning Projects:**

J. Yang, Y. Liu, E. P. Xing and A. Hauptmann, [Harmonium-Based Models for Semantic Video Representation and Classification](#), *Proceedings of The Seventh SIAM International Conference on Data Mining (SDM 2007 best paper)*

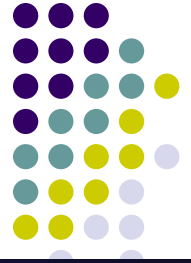
Manaal Faruqi, Jesse Dodge, Sujay Kumar Jauhar, Chris Dyer, Eduard Hovy, Noah A. Smith, [Retrofitting Word Vectors to Semantic Lexicons](#), NAACL 2015 best paper

Others ... such as KDD 2014 best paper

- **Other projects:**

Andreas Krause, Jure Leskovec and Carlos Guestrin, [Data Association for Topic Intensity Tracking](#), *23rd International Conference on Machine Learning (ICML 2006)*.

M. Sachan, A. Dubey, S. Srivastava, E. P. Xing and Eduard Hovy, [Spatial Compactness meets Topical Consistency: Jointly modeling Links and Content for Community Detection](#), *Proceedings of The 7th ACM International Conference on Web Search and Data Mining (WSDM 2014)*.

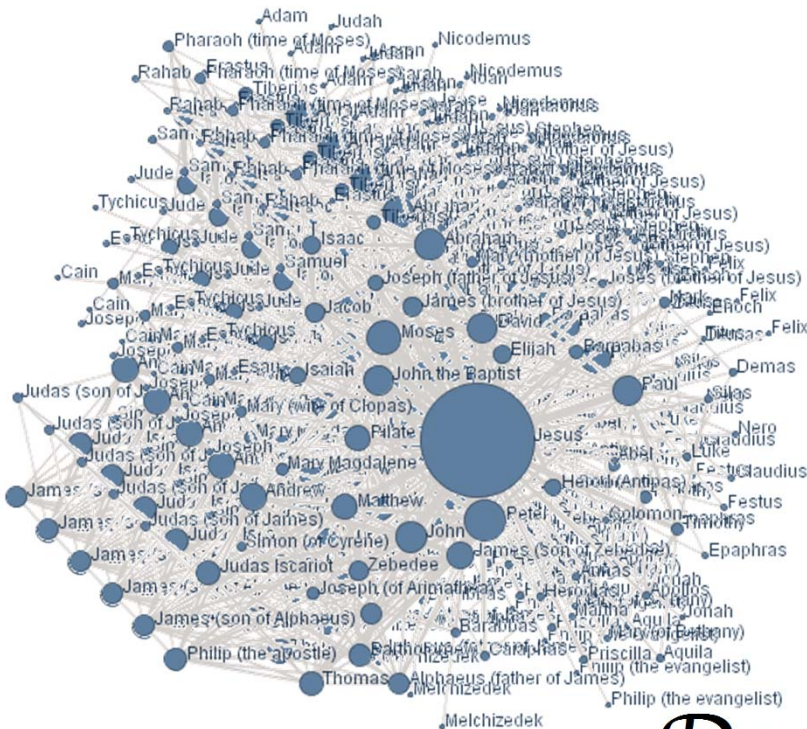


# What Are Graphical Models?

**Graph**

**Model**

$M_G$



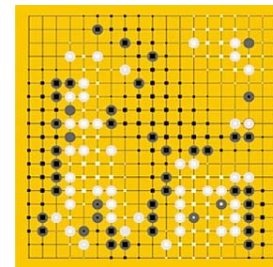
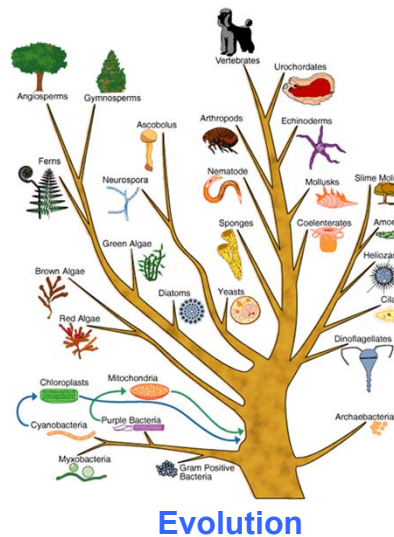
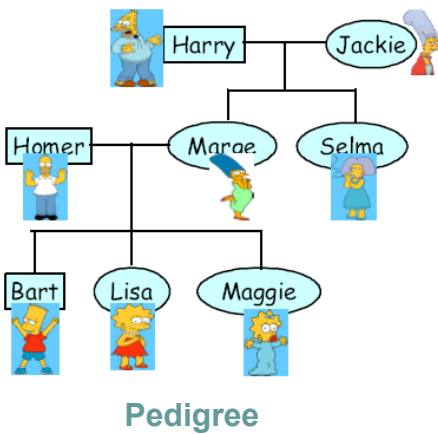
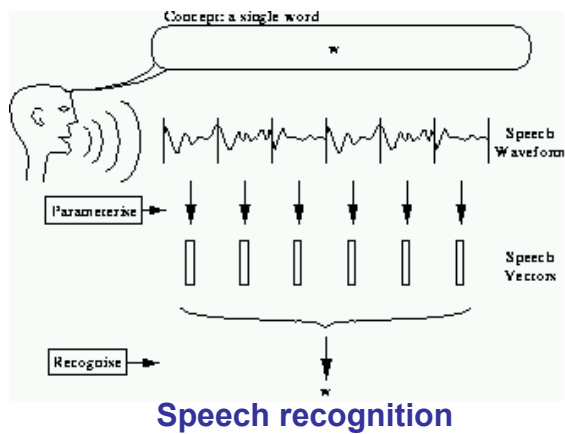
**Data**

$$\mathcal{D} \equiv \{X_1^{(i)}, X_2^{(i)}, \dots, X_m^{(i)}\}_{i=1}^N$$





# Reasoning under uncertainty!





# The Fundamental Questions

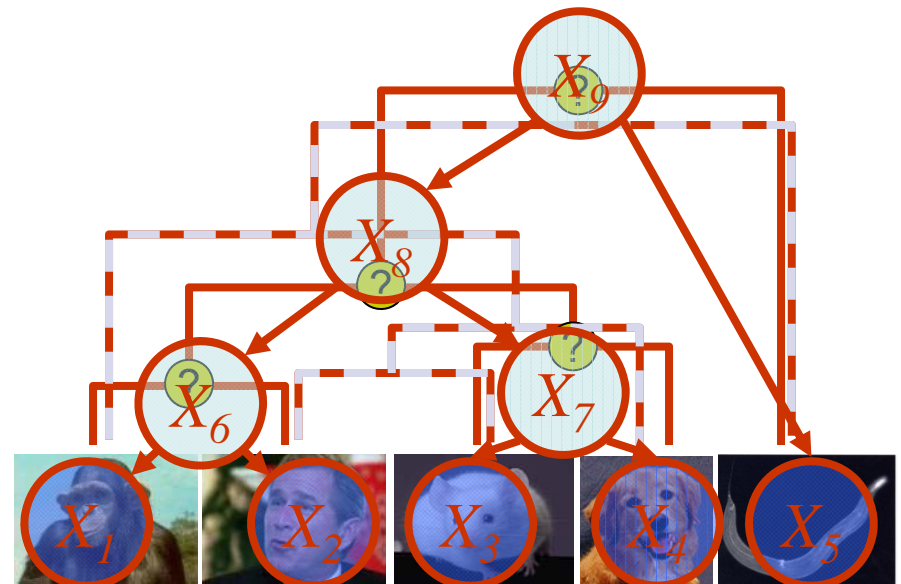
- Representation
  - How to capture/model uncertainties in possible worlds?
  - How to encode our domain knowledge/assumptions/constraints?

- Inference
  - How do I answers questions/queries according to my model and/or based given data?

e.g.:  $P(X_i | \mathbf{D})$

- Learning
  - What model is "right" for my data?

e.g.:  $\mathcal{M} = \arg \max_{\mathcal{M} \in \mathcal{M}} F(\mathbf{D}; \mathcal{M})$





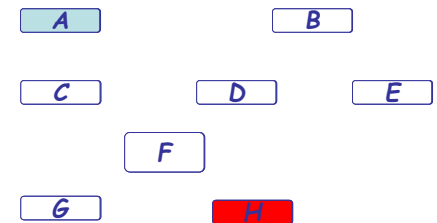


# Recap of Basic Prob. Concepts

- 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?**



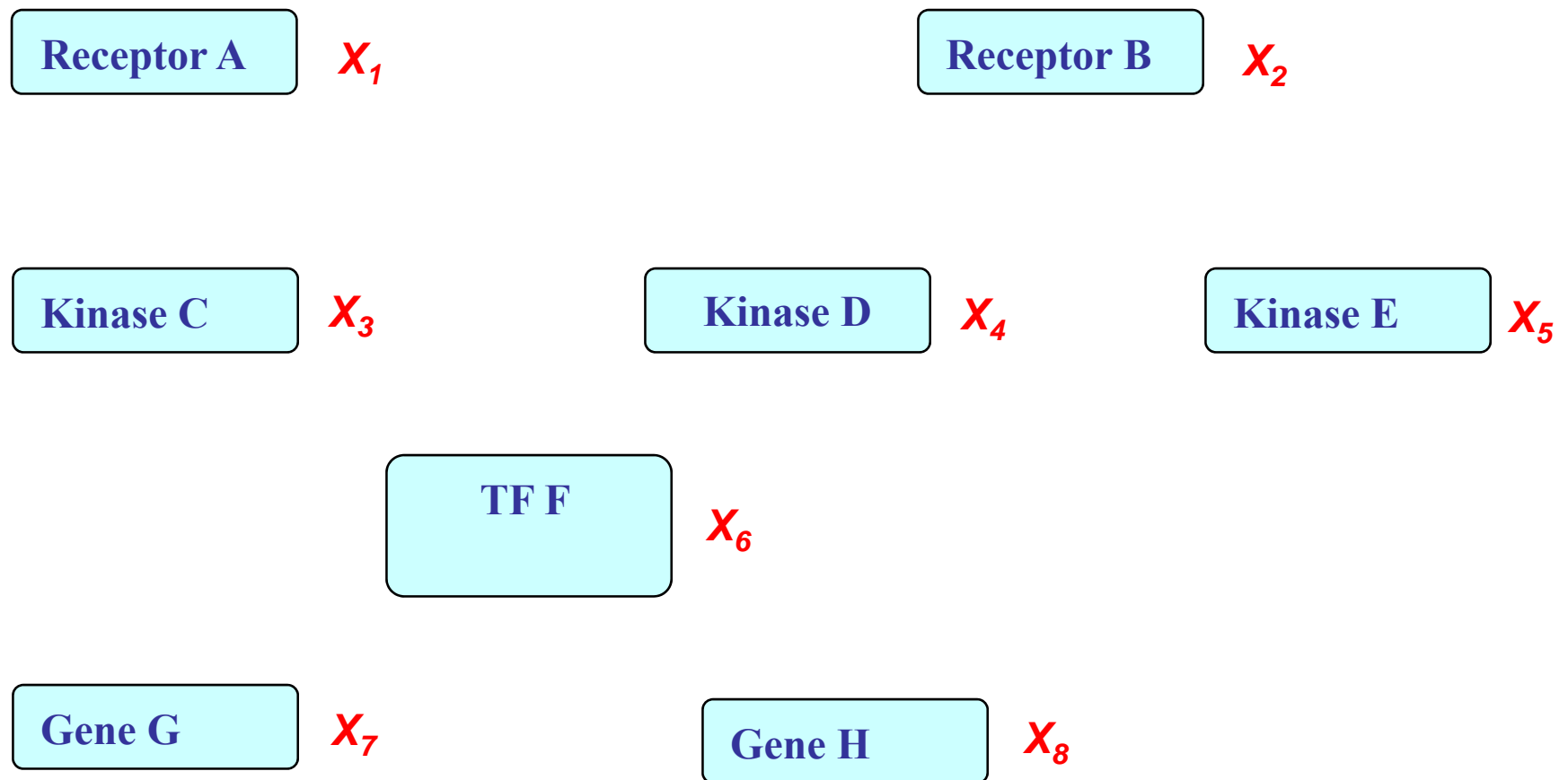
- Learning: where do we get all this probabilities?
  - Maximal-likelihood estimation? but how many data do we need?
  - Are there other est. principles?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?
- Inference: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

--- Multivariate Distribution in High-D Space



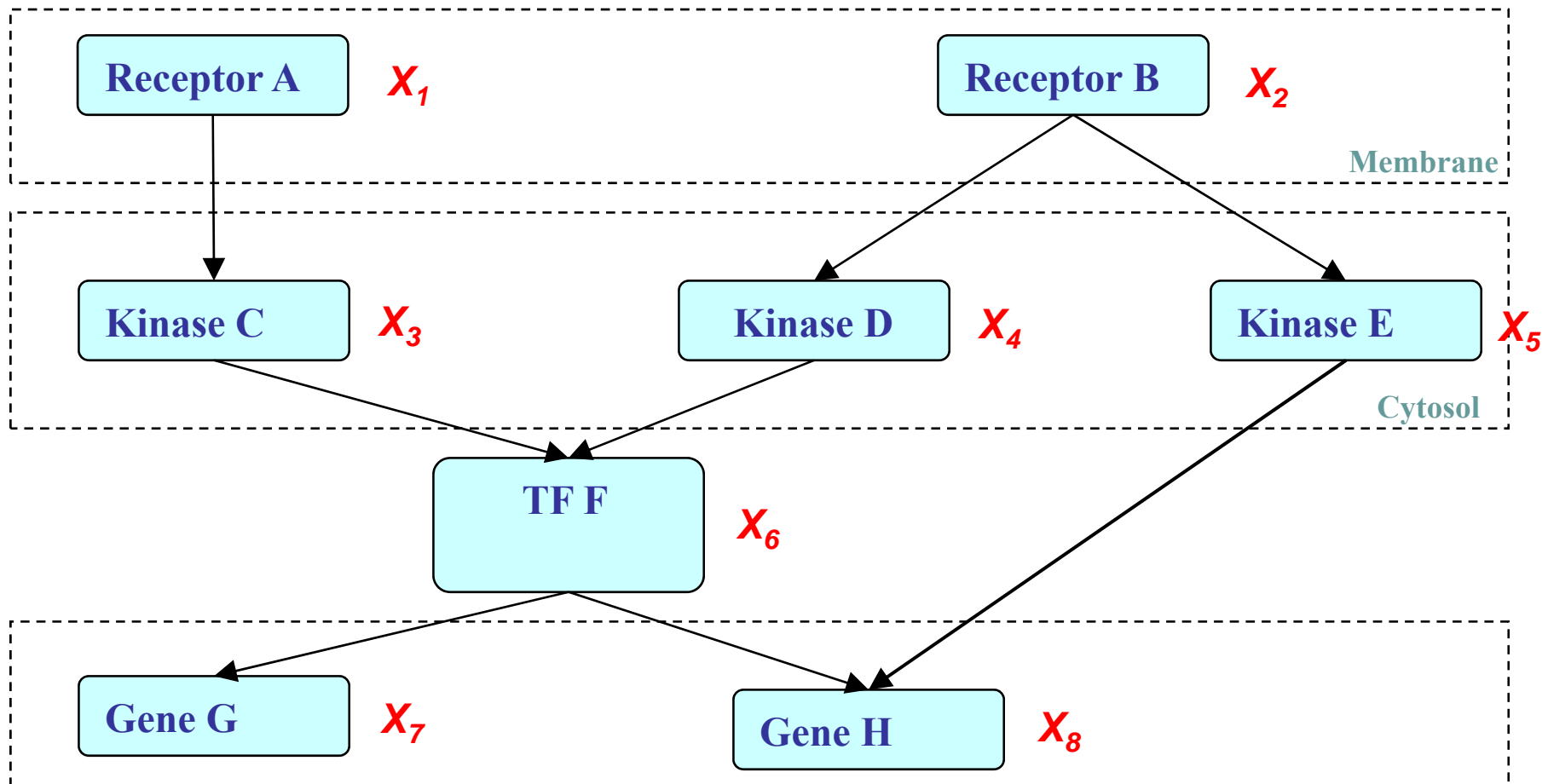
- A possible world for cellular signal transduction:



# GM: Structure Simplifies Representation



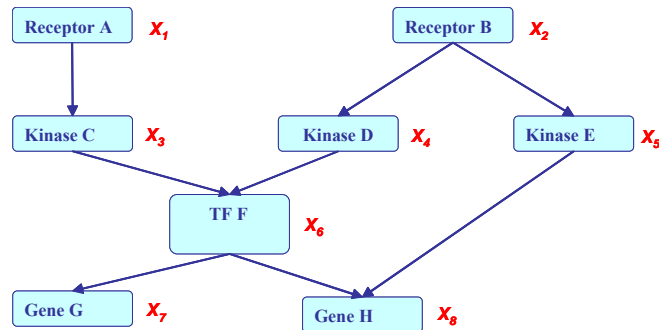
- Dependencies among variables





# 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.,



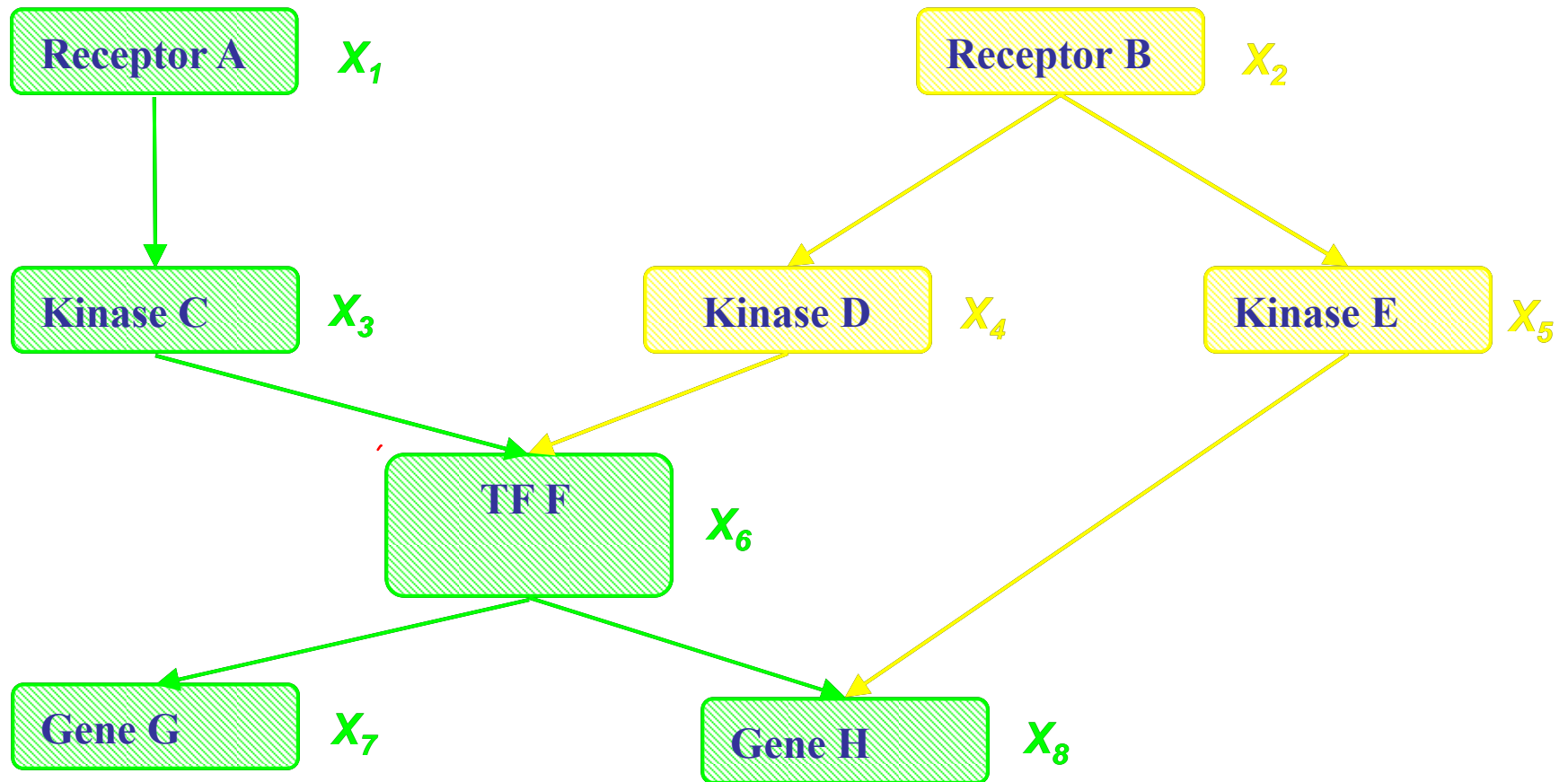
$$\begin{aligned} &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\ &= P(X_1) P(X_2) P(X_3/X_1) P(X_4/X_2) P(X_5/X_2) \\ &P(X_6/X_3, X_4) P(X_7/X_6) P(X_8/X_5, X_6) \end{aligned}$$

Stay tune for what are these independencies!

- Why we may favor a PGM?
  - Incorporation of domain knowledge and causal (logical) structures  
1+1+2+2+2+4+2+4=18, a 16-fold reduction from  $2^8$  in representation cost !



# GM: Data Integration



# More Data Integration

---



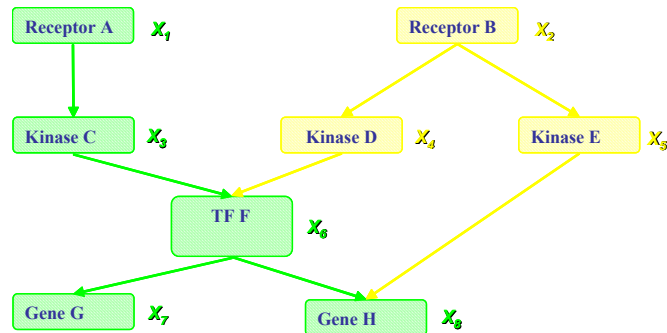
- Text + Image + Network → Holistic Social Media
  
- Genome + Proteome + Transcriptome + Phenome + ... → PanOmic Biology





# 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.,



$$\begin{aligned} &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\ &= P(X_2) P(X_4 | X_2) P(X_5 | X_2) P(X_1) P(X_3 | X_1) \\ &P(X_6 | X_3, X_4) P(X_7 | X_6) P(X_8 | X_5, X_6) \end{aligned}$$

- Why we may favor a PGM?
  - Incorporation of domain knowledge and causal (logical) structures  
 $2+2+4+4+4+8+4+8=36$ , an 8-fold reduction from  $2^8$  in representation cost !
  - Modular combination of heterogeneous parts – data fusion



# Rational Statistical Inference

## The Bayes Theorem:

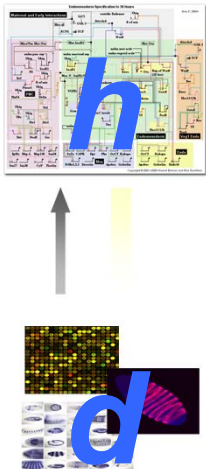
Posterior probability

Likelihood

Prior probability

$$p(h | d) = \frac{p(d | h) p(h)}{\sum_{h' \in H} p(d | h') p(h')}$$

Sum over space of hypotheses

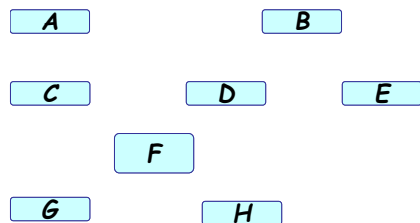


- This allows us to capture uncertainty about the model in a principled way
- But how can we specify and represent a complicated model?
  - Typically the number of genes need to be modeled are in the order of thousands!



# GM: MLE and Bayesian Learning

- Probabilistic statements of  $\Theta$  is conditioned on the values of the observed variables  $\mathbf{A}_{\text{obs}}$  and prior  $p(\cdot | \chi)$

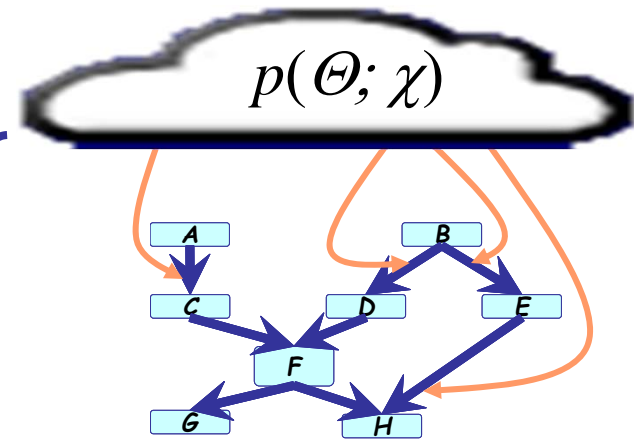
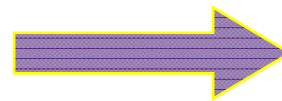


$(A,B,C,D,E,\dots)=(T,F,F,T,F,\dots)$

$\mathbf{A} = (A,B,C,D,E,\dots)=(T,F,T,T,F,\dots)$

.....

$(A,B,C,D,E,\dots)=(F,T,T,T,F,\dots)$



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

$$\Theta_{\text{Bayes}} = \int \Theta p(\Theta | \mathbf{A}, \chi) d\Theta$$

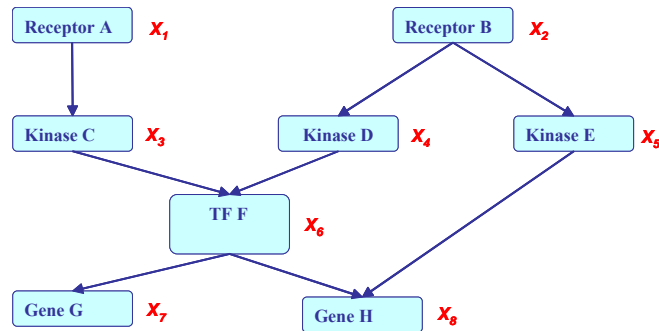
$$p(\Theta | \mathbf{A}; \chi) \propto p(\mathbf{A} | \Theta) p(\Theta; \chi)$$

posterior
likelihood
prior



# 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.,



$$\begin{aligned}
 &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\
 &= P(X_1) P(X_2) P(X_3/X_1) P(X_4/X_2) P(X_5/X_2) \\
 &P(X_6/X_3, X_4) P(X_7/X_6) P(X_8/X_5, X_6)
 \end{aligned}$$

- Why we may favor a PGM?
  - Incorporation of domain knowledge and causal (logical) structures  
 $2+2+4+4+4+8+4+8=36$ , an 8-fold reduction from  $2^8$  in representation cost !
  - Modular combination of heterogeneous parts – data fusion
  - Bayesian Philosophy

- Knowledge meets data





# So What Is a PGM After All?

---

**In a nutshell:**

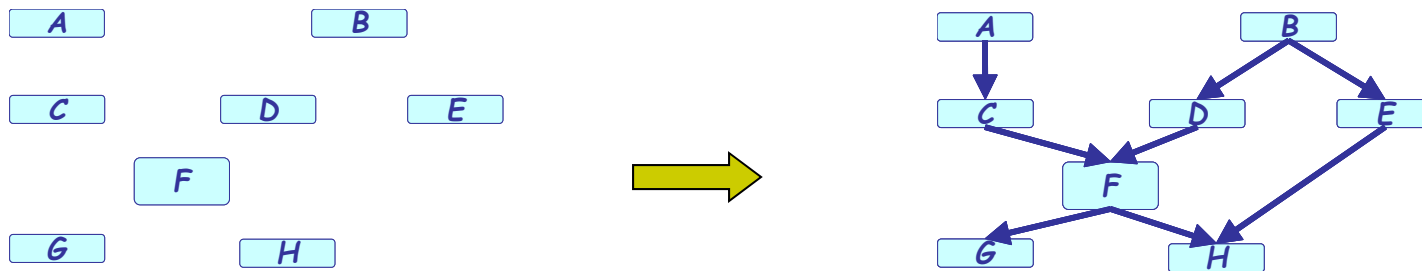
**PGM = Multivariate Statistics + Structure**

**GM = Multivariate Obj. Func. + Structure**



# So What Is a PGM After All?

- The informal blurb:
  - 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**



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

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

- A more formal description:
  - 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

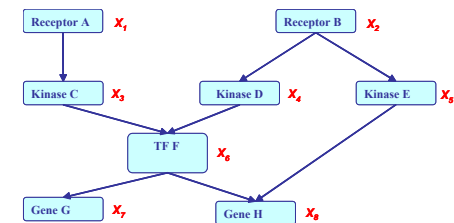




# Two types of GMs

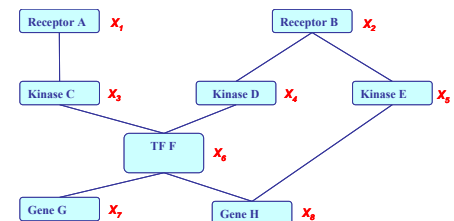
- Directed edges give causality relationships (Bayesian Network or Directed Graphical Model):

$$\begin{aligned}
 &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\
 &= P(X_1) P(X_2) P(X_3/X_1) P(X_4/X_2) P(X_5/X_2) \\
 &\quad P(X_6/X_3, X_4) P(X_7/X_6) P(X_8/X_5, X_6)
 \end{aligned}$$

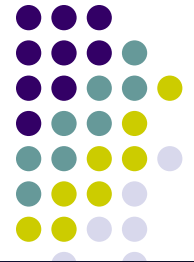


- Undirected edges simply give correlations between variables (Markov Random Field or Undirected Graphical model):

$$\begin{aligned}
 &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\
 &= \frac{1}{Z} \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2) \\
 &\quad + E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}
 \end{aligned}$$

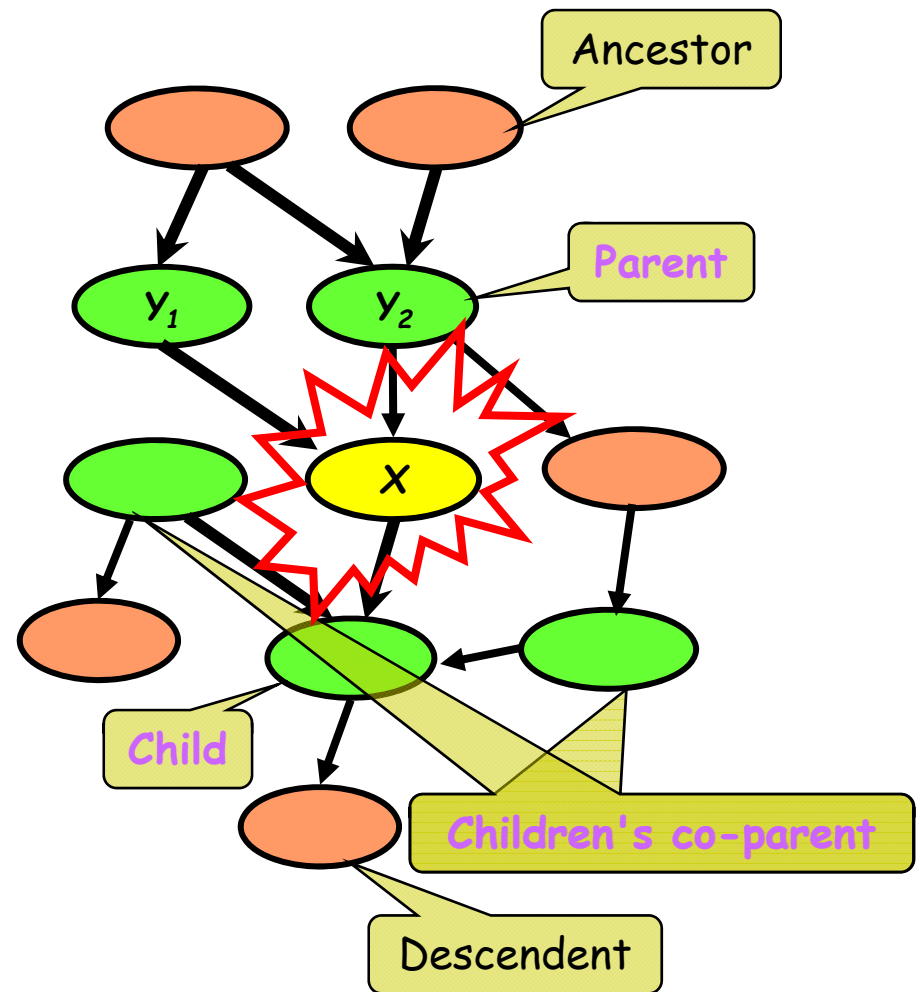


# Bayesian Networks



## 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

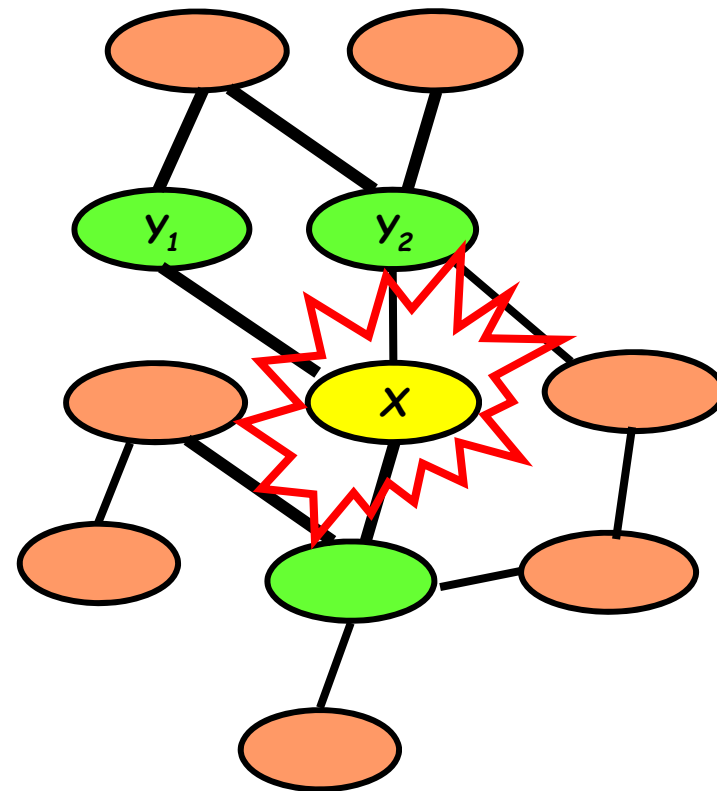




# Markov Random Fields

Structure: *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



# Towards 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

- **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  $G$ ,

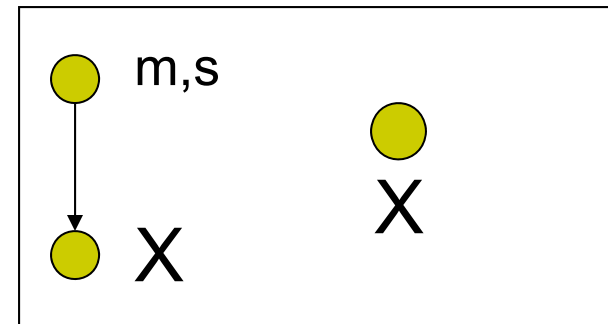
Then  $\mathcal{D}_1 \equiv \mathcal{D}_2$ .



# GMs are your old friends

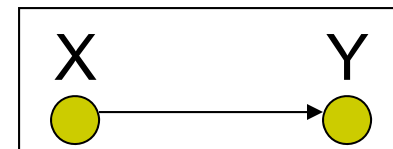
## Density estimation

Parametric and nonparametric methods



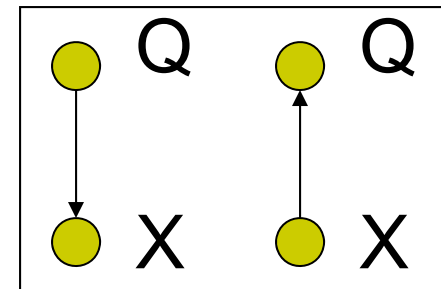
## Regression

Linear, conditional mixture, nonparametric



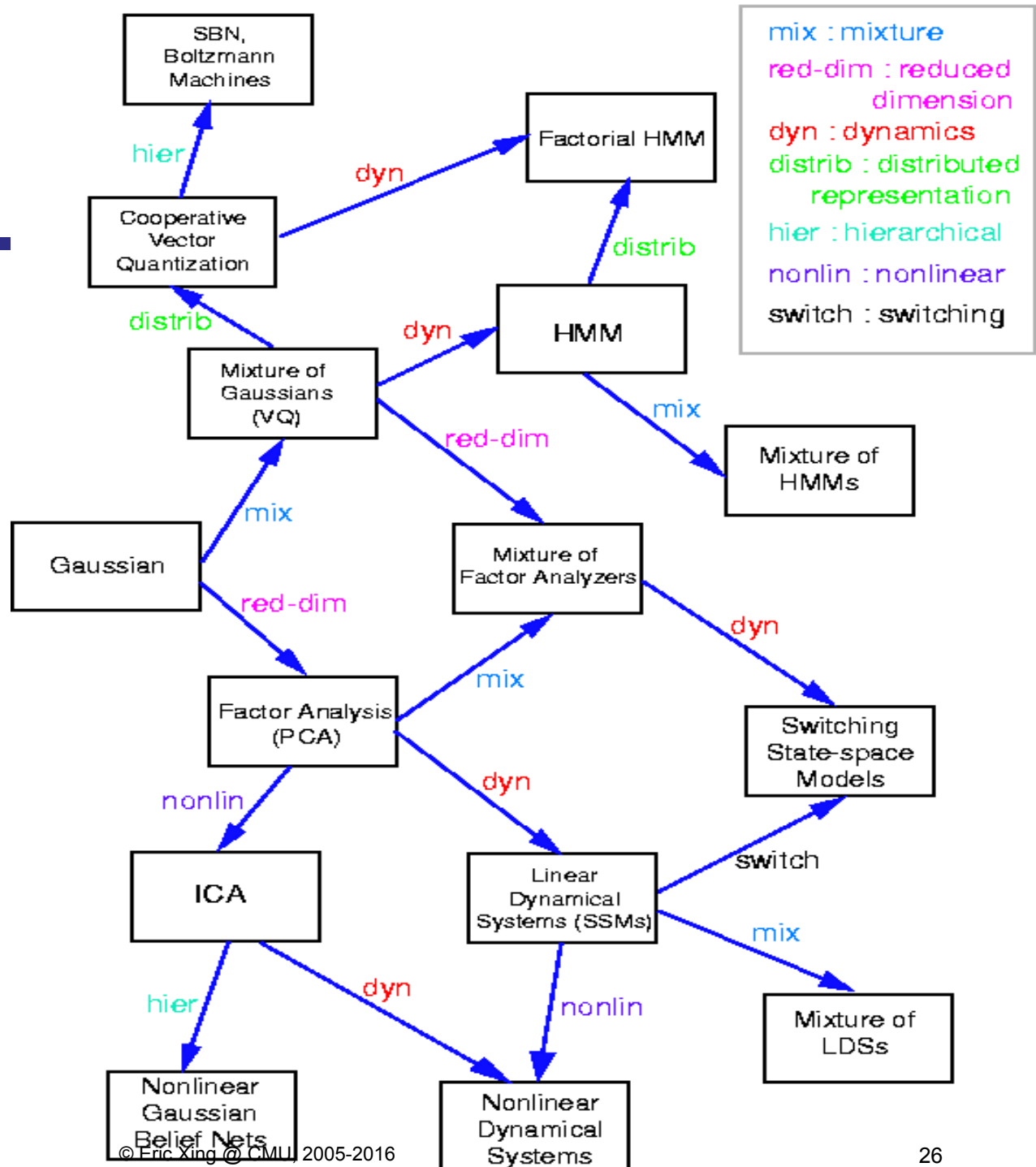
## Classification

Generative and discriminative approach



## Clustering

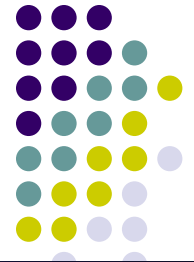
# An (incomplete) genealogy of graphical models



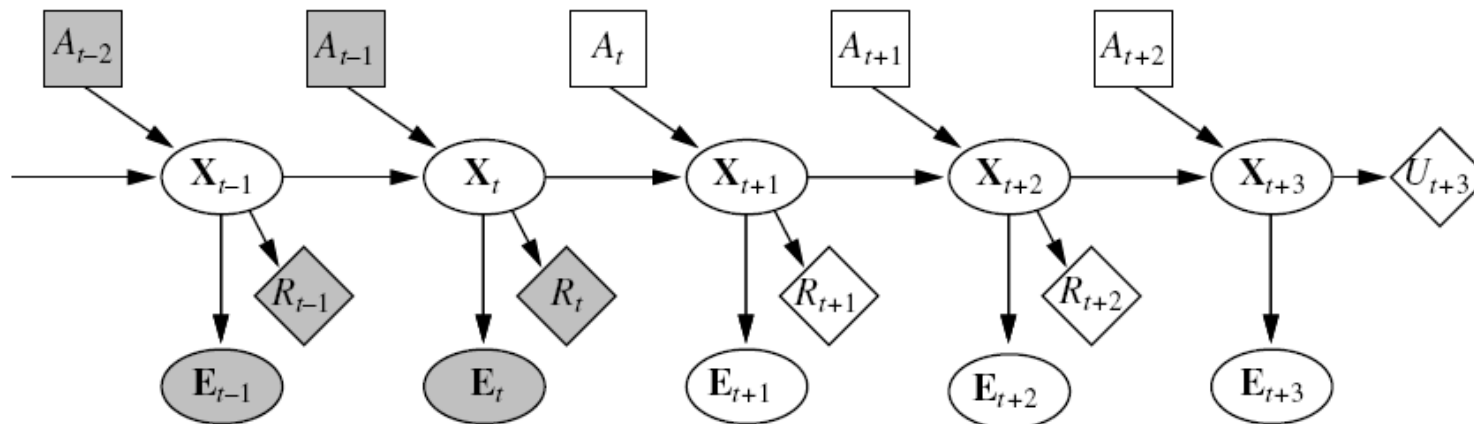
(Picture by Zoubin Ghahramani and Sam Roweis)



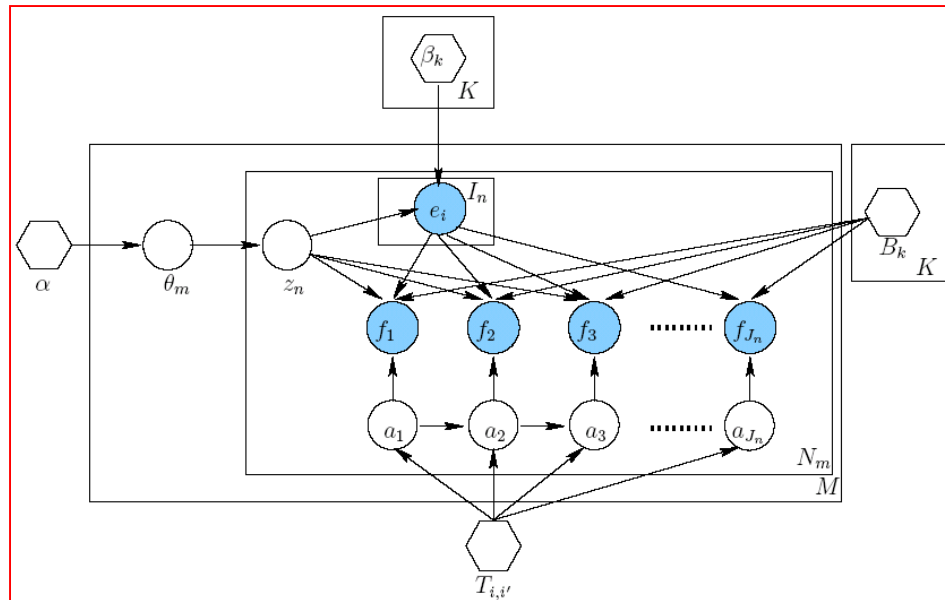
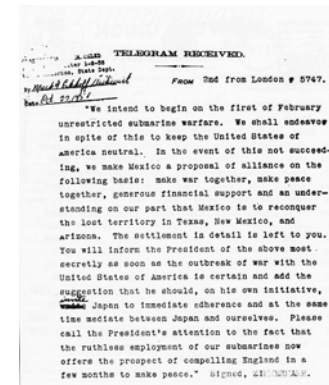
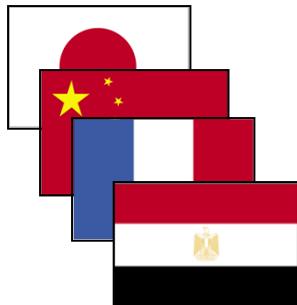
# Fancier GMs: reinforcement learning



- Partially observed Markov decision processes (POMDP)

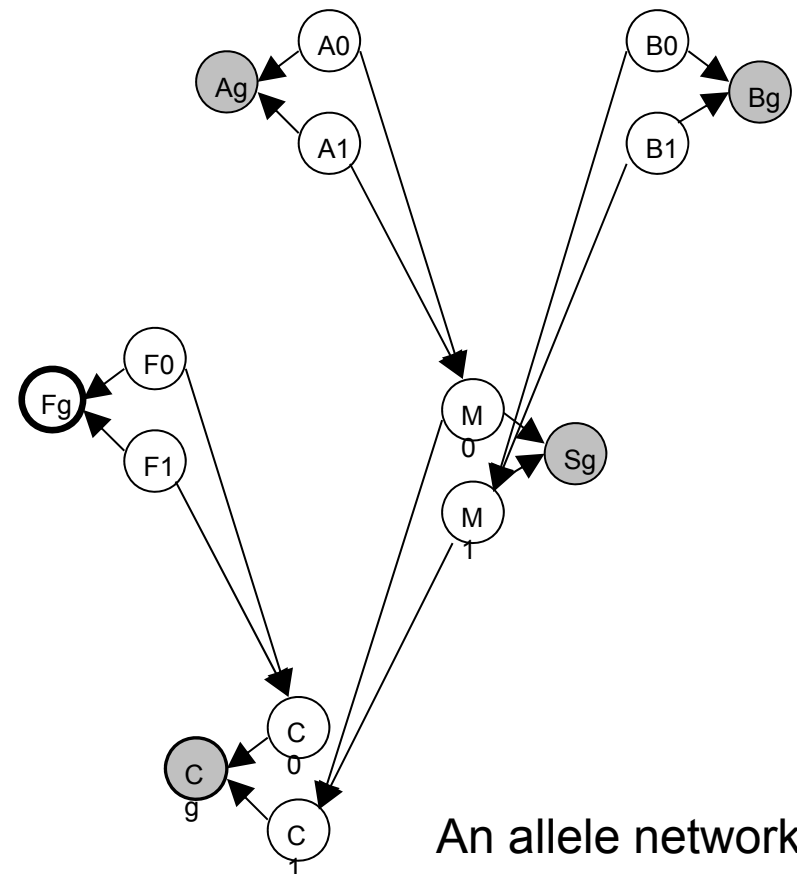
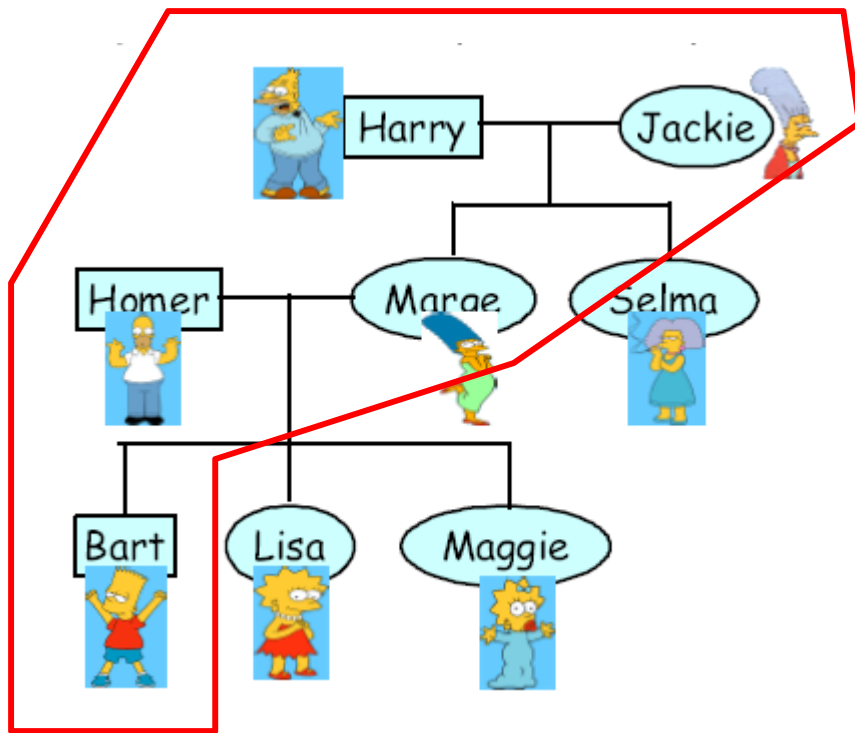


# Fancier GMs: machine translation

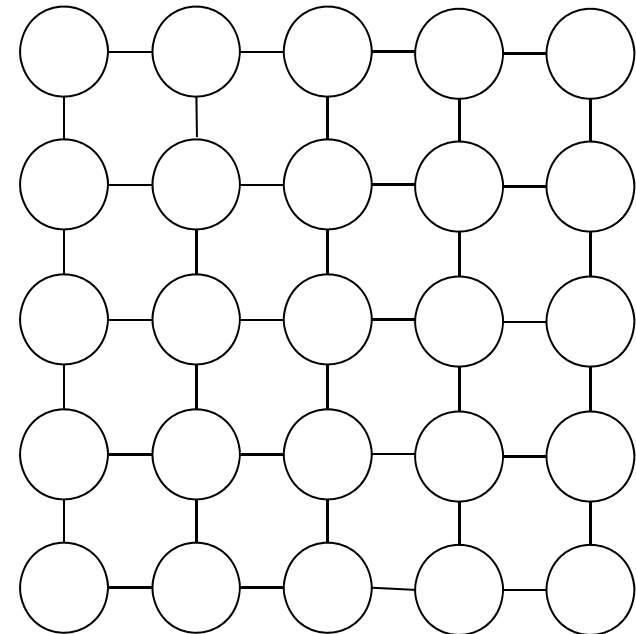
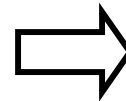
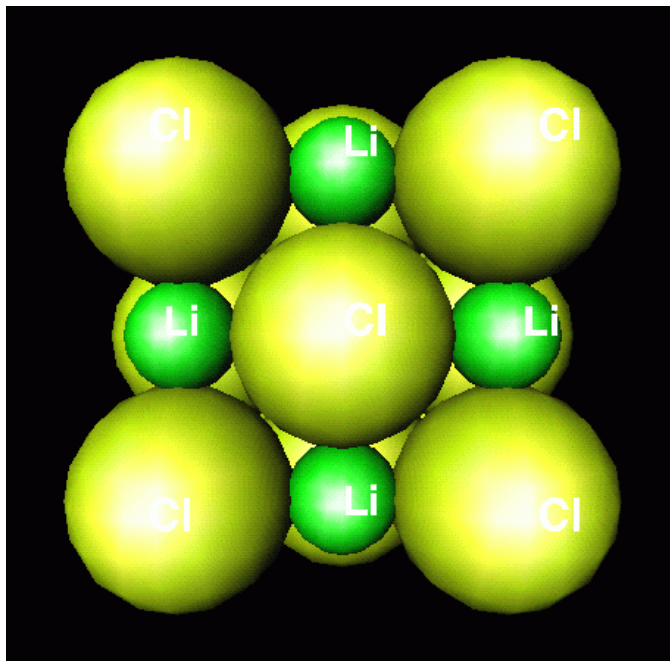


The HM-BiTAM model  
(B. Zhao and E.P Xing,  
ACL 2006)

# Fancier GMs: genetic pedigree



# Fancier GMs: solid state physics



Ising/Potts model



# Application of GMs

---

- Machine Learning
- Computational statistics
  
- Computer vision and graphics
- Natural language processing
- Informational retrieval
- Robotic control
- Decision making under uncertainty
- Error-control codes
- Computational biology
- Genetics and medical diagnosis/prognosis
- Finance and economics
- Etc.



# Why graphical models

---

- A language for communication
  - A language for computation
  - A language for development
- 
- Origins:
    - Wright 1920's
    - Independently developed by Spiegelhalter and Lauritzen in statistics and Pearl in computer science in the late 1980's





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



# Plan for the Class

- Fundamentals of Graphical Models:
  - Bayesian Network and Markov Random Fields
  - Discrete, Continuous and Hybrid models, exponential family, GLIM
  - Basic representation, inference, and learning
  - ...
- Advanced topics and latest developments
  - Approximate inference
    - Monte Carlo algorithms
    - Variational methods and theories
  - “Infinite” GMs: nonparametric Bayesian models
  - Optimization-theoretic formulations for GMs, e.g., Structured sparsity
  - Nonparametric and spectral graphical models, where GM meets kernels and matrix algebra
  - Alternative GM learning paradigms,
    - e.g., Margin-based learning of GMs (where GM meets SVM)
    - e.g. Regularized Bayes: where GM meets SVM, and meets Bayesian, and meets NB ...
- Case studies: popular GMs and applications
  - Multivariate Gaussian Models
  - Conditional random fields
  - Mixed-membership, aka, Topic models

# Questions ?

---

