

Probabilistic Graphical Model: A view from moon

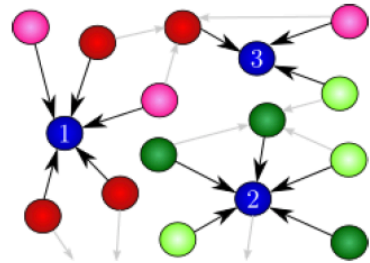
Kayhan Batmanghelich

Logistics

- Class webpage:
 - <https://kayhan.dbmi.pitt.edu/node/38>

10-708 (CMU) Probabilistic Graphical Models

Probabilistic Graphical Models



Overview

Many of the problems in artificial intelligence, statistics, computer systems, computer vision, natural language processing, and computational biology, among many other fields, can be viewed as the search for a coherent global conclusion from local information. The probabilistic graphical model's framework provides a unified view for this wide range of problems, enabling efficient inference, decision-making and learning in problems with a very large number of attributes and huge datasets. This graduate-level course will provide you with a strong foundation for both applying graphical models to complex problems and for addressing core research topics in graphical models. The class will cover three aspects: The core representation, including Bayesian and Markov networks, and dynamic Bayesian networks; probabilistic inference algorithms, both exact and approximate; and, learning methods for both the parameters and the structure of graphical models. Students entering the class should have a pre-existing working knowledge of probability, statistics, and algorithms, though the class has been designed to allow students with a strong numerate background to catch up and fully participate. It is expected that after taking this class, the students should have obtained sufficient working knowledge of multivariate probabilistic modeling and inference for practical applications, should be able to formulate and solve a wide range of problems in their own domain using GM and can advance into more specialized technical literature by themselves. Students are required to have successfully completed 10701 or 10715, or an equivalent class.

Where and When

- **Time:** Tuesday, Thursday 12:00 - 1:20 pm
- **Location:** Gates-Hillman Center 4307
- **Recitations:**

Logistics

- **References:**

- Daphne Koller and Nir Friedman, **Probabilistic Graphical Models**
- M. I. Jordan, **An Introduction to Probabilistic Graphical Models**
- K. Murphy, **Machine Learning: A Probabilistic Perspective**
- C.M. Bishop, **Pattern Recognition and Machine Learning**
- D. Barber, **Bayesian Reasoning and Machine Learning**
- D. J. C. MacKay, **Information Theory, Inference, and Learning Algorithms**

- **Mailing Lists:**

- To contact the instructors: 10708Spring18@gmail.com
- Class announcements list: **send email with title (Add me to the class announcement)**

- **TA:**

Xiongtao	Ruan	xruan@andrew.cmu.edu
Yifeng	Tao	yifengt@andrew.cmu.edu
Yuanning	Li	yuanninl@andrew.cmu.edu

- **Guest Lecturers:**

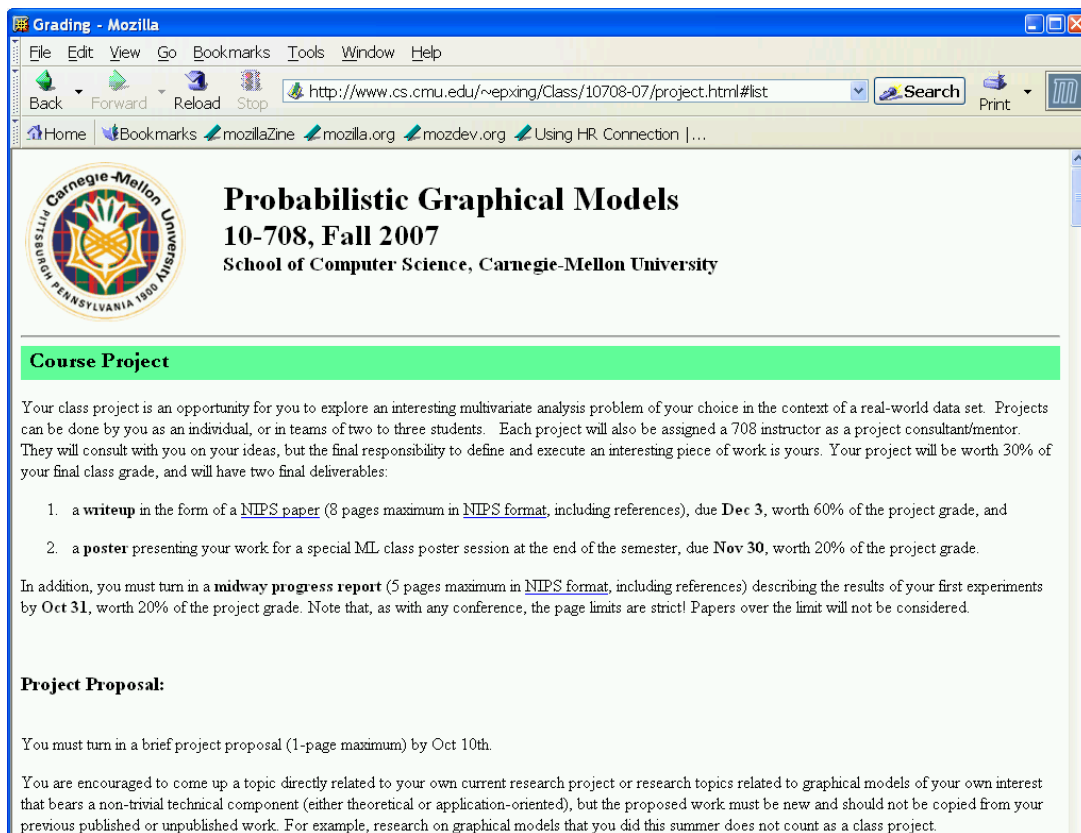
- A few

- **Instruction aids:** Piazza

Logistics

- 5 homework (**HW0 + 4 HWs**) assignments: 45% of grade
 - Theory exercises, Implementation exercises
- Scribe duties: 5% (~once to twice for the whole semester)
- Reading report after every module: 10%
- Final project: 40% of grade
 - Applying PGM to the development of a real, substantial ML system
 - Natural Language Processing: Innovative language alignment methods
 - Computer Vision/Medical Vision: Innovative Image/text captioning, Domain transfer learning
 - Computational Biology applications: Incorporating multi-omic dataset to understand the diseases.
 - Causality: Learning Causal GM with missing data.
 - Theoretical and/or algorithmic work
 - Innovative Inference approach in the intersection of deep learning and Bayesian inference.
 - Analyzing the behavior of the distributed SVI algorithms.
 - 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:




Grading - Mozilla

File Edit View Go Bookmarks Tools Window Help

Back Forward Reload Stop <http://www.cs.cmu.edu/~epxing/Class/10708-07/project.html#list> Search Print

Home Bookmarks mozillaZine mozilla.org mozdev.org Using HR Connection |...

 **Probabilistic Graphical Models**
10-708, Fall 2007
School of Computer Science, Carnegie-Mellon University

Course Project

Your class project is an opportunity for you to explore an interesting multivariate analysis problem of your choice in the context of a real-world data set. Projects can be done by you as an individual, or in teams of two to three students. Each project will also be assigned a 708 instructor as a project consultant/mentor. They will consult with you on your ideas, but the final responsibility to define and execute an interesting piece of work is yours. Your project will be worth 30% of your final class grade, and will have two final deliverables:

1. a **writeup** in the form of a [NIPS paper](#) (8 pages maximum in [NIPS format](#), including references), due **Dec 3**, worth 60% of the project grade, and
2. a **poster** presenting your work for a special ML class poster session at the end of the semester, due **Nov 30**, worth 20% of the project grade.

In addition, you must turn in a **midway progress report** (5 pages maximum in [NIPS format](#), including references) describing the results of your first experiments by **Oct 31**, worth 20% of the project grade. Note that, as with any conference, the page limits are strict! Papers over the limit will not be considered.

Project Proposal:

You must turn in a brief project proposal (1-page maximum) by Oct 10th.

You are encouraged to come up a topic directly related to your own current research project or research topics related to graphical models of your own interest that bears a non-trivial technical component (either theoretical or application-oriented), but the proposed work must be new and should not be copied from your previous published or unpublished work. For example, research on graphical models that you did this summer does not count as a class project.

- 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 Faruqui, 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

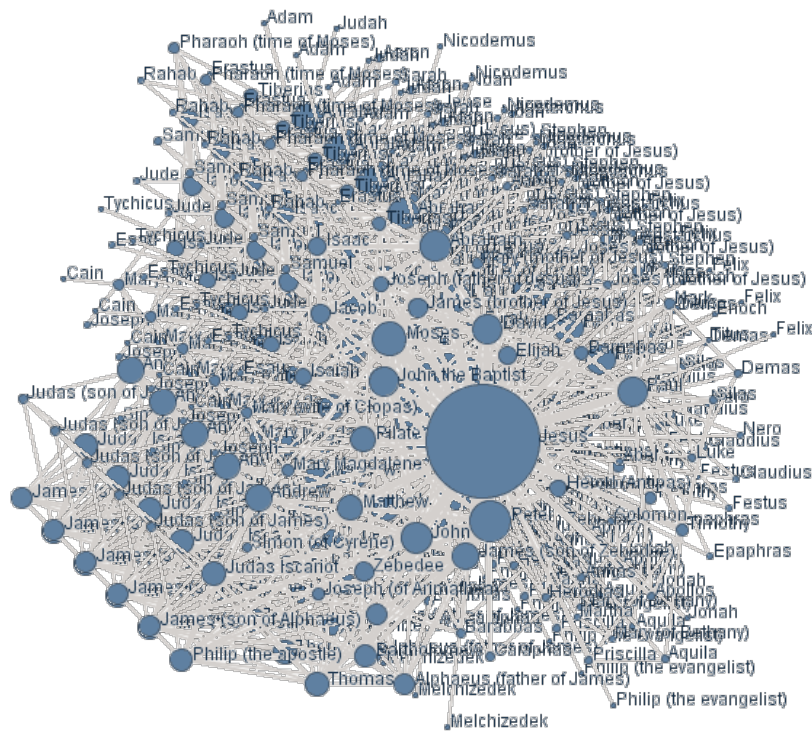
Other projects:

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



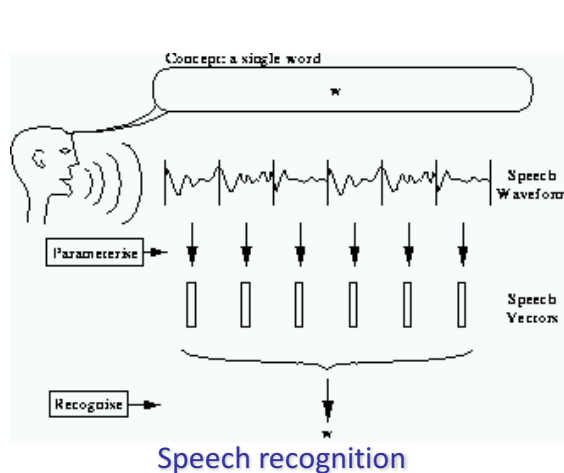
Model

$$\mathcal{M}_G$$

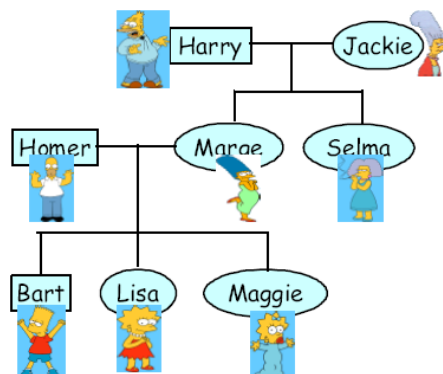
Data

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

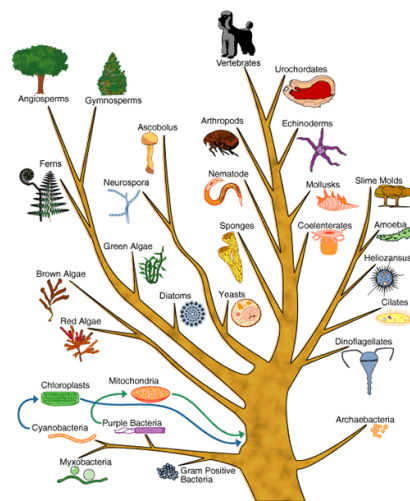
Reasoning under uncertainty!



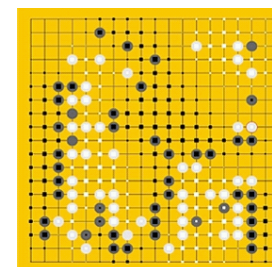
Computer vision



Pedigree



Evolution



Games



Robotic control

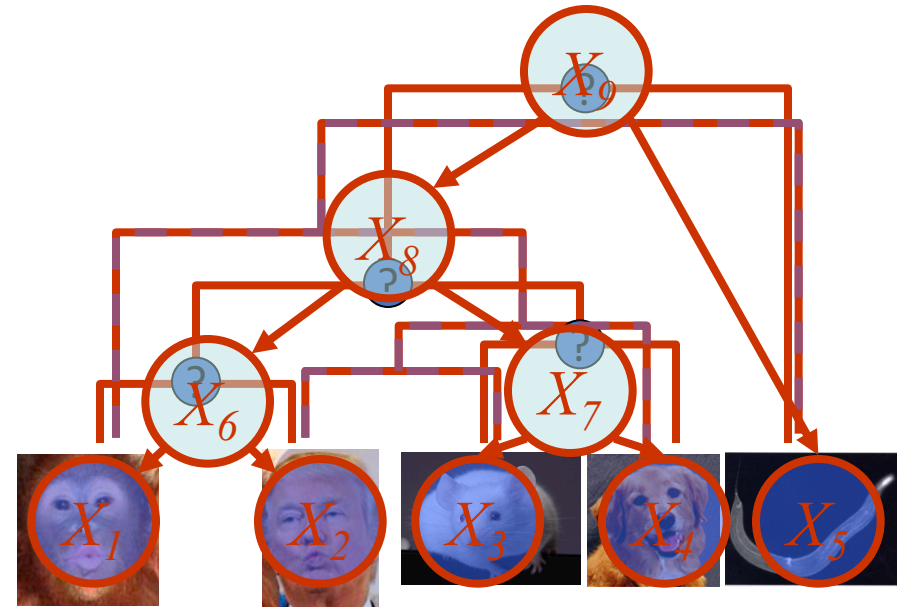


The Fundamental Questions

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

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

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

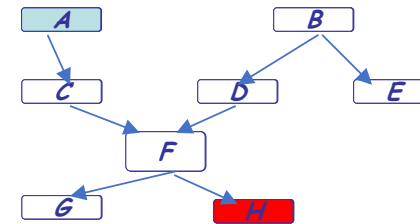


Recap of Basic Prob. Concepts

- 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
- How do we represent that many element? Do we need such a big table?
- How to incorporate scientific/medical insight?

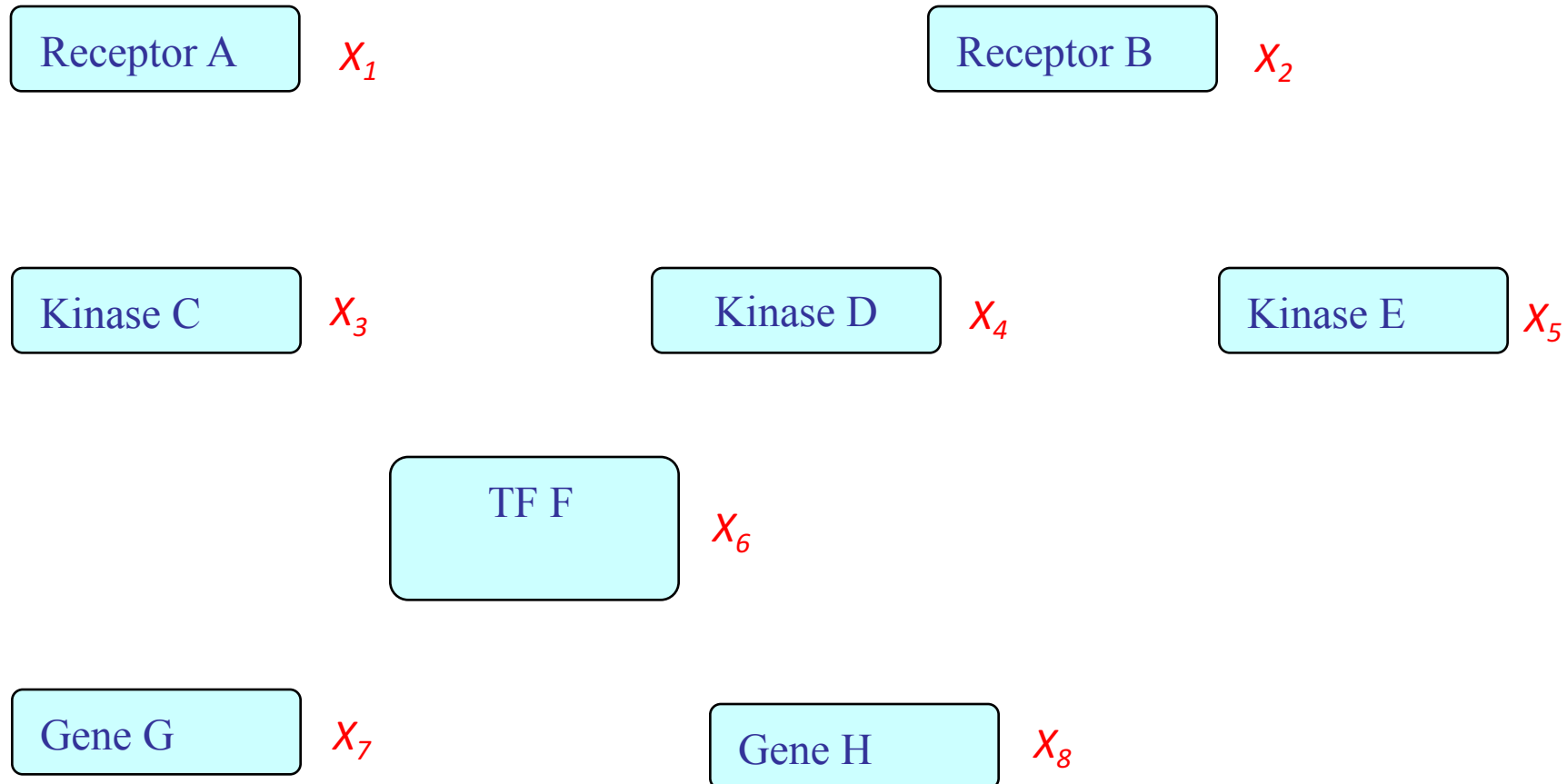


- Inference: If not all variables are observable, how to compute the conditional distribution of latent variables given evidence?
 - Computing $p(H|A)$ would require summing over all 2^6 configurations of the unobserved variables
- Learning: where do we get all this probabilities?
 - Maximal-likelihood estimation? but how many data do we need?
 - Are there other est. principles?
 - What if we just have data and want to **learn** the relationship?

What is a Graphical Model?

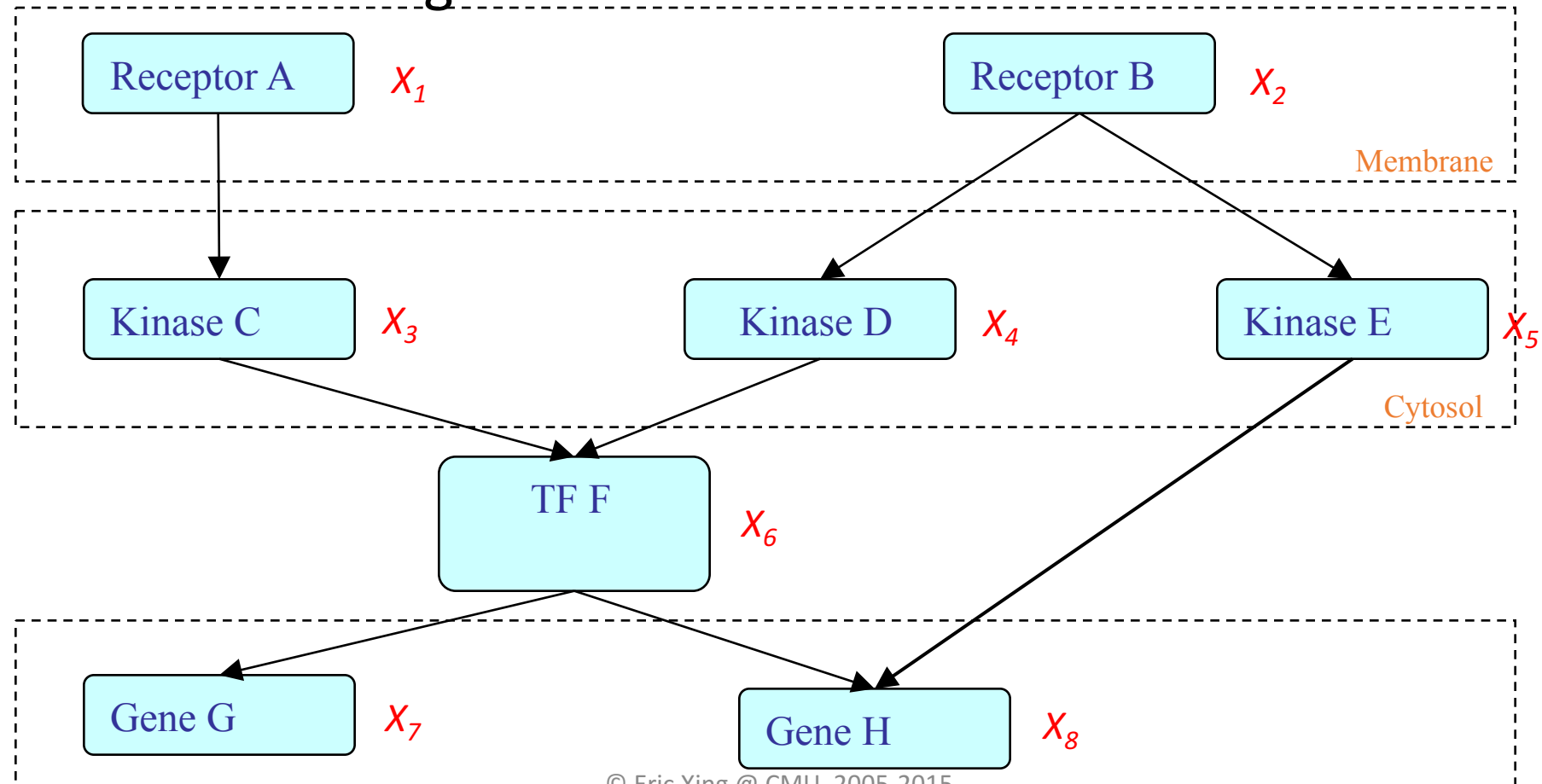
--- Multivariate Distribution in High-D Space

- A possible world for cellular signal transduction:



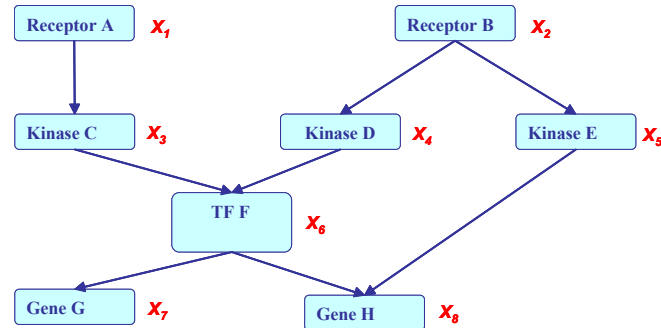
GM: Structure Simplifies Representation

- Dependencies among variables



Why we may favor a PGM?

- 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.,



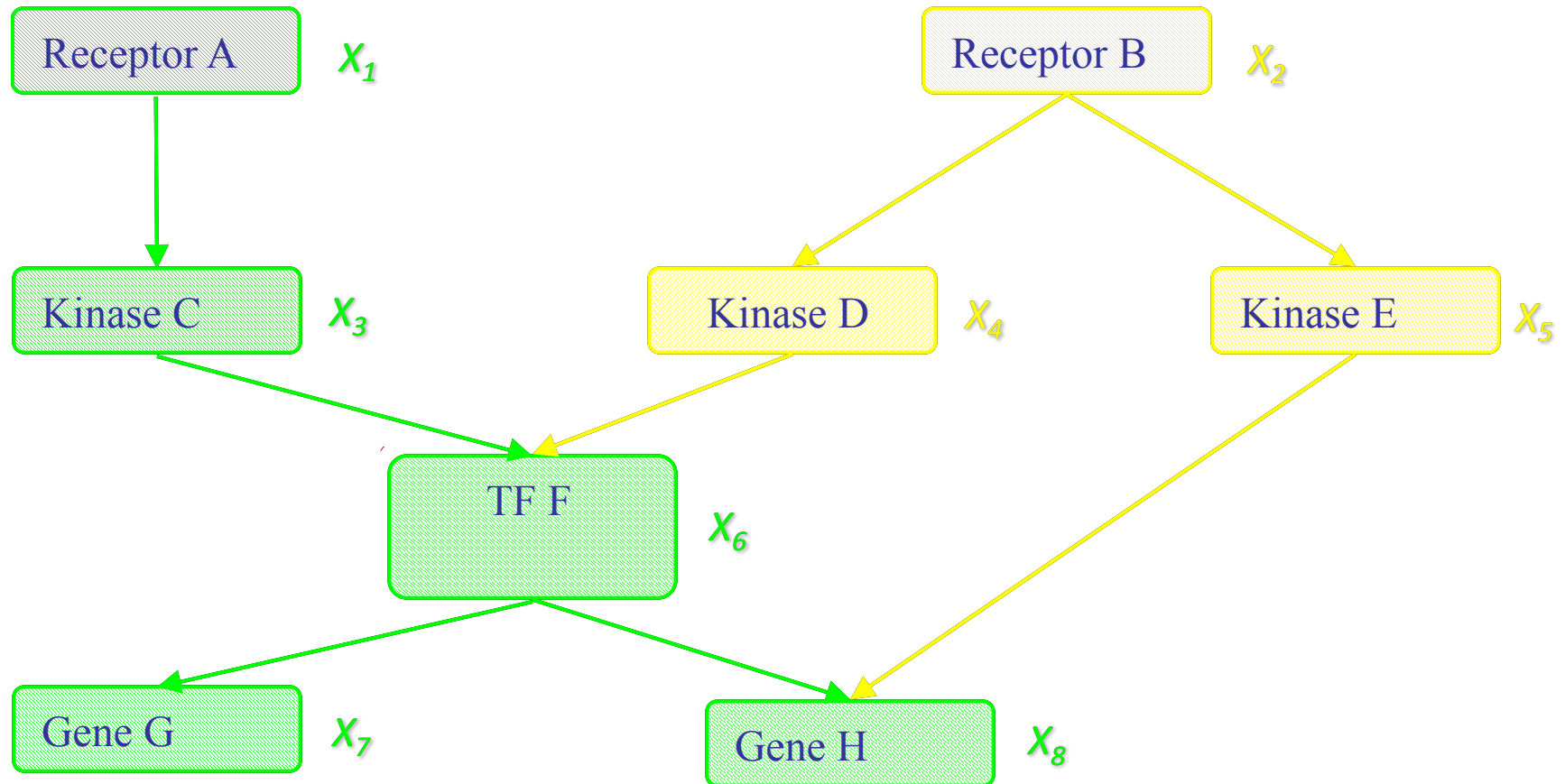
$$\begin{aligned} &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\ &= P(X_1) P(X_2) P(X_3 | X_1) P(X_4 | X_2) P(X_5 | X_2) \\ &\quad P(X_6 | X_3, X_4) P(X_7 | X_6) P(X_8 | X_5, X_6) \end{aligned}$$

Stay tune for what are these independencies!

- 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

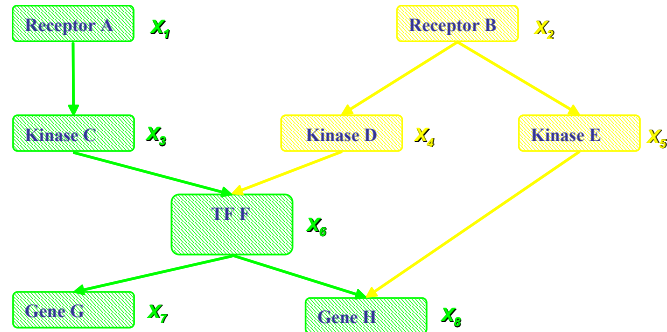


More Data Integration

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

Why we may favor a PGM?

- 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.,



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

- 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:

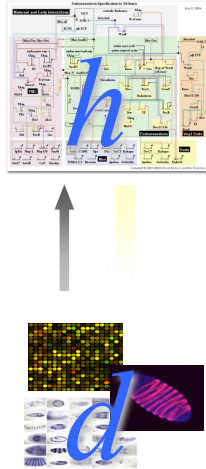
Posterior probability

Likelihood

Prior probability

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

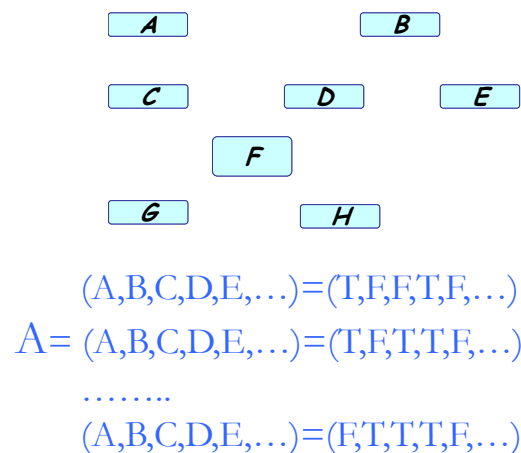
Sum over space of hypotheses



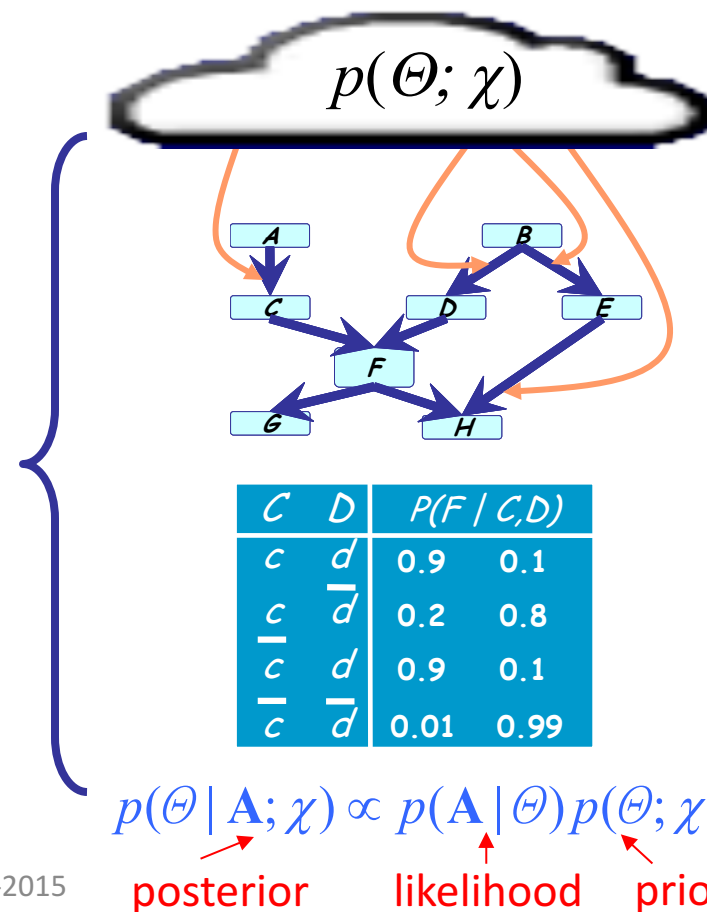
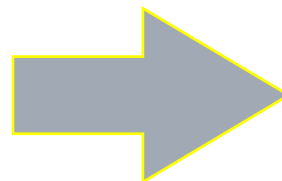
- This allows us to capture uncertainty about the model in a principled way
- But how can we specify and represent a complicated model?
 - Typically the number of genes need to be modeled are in the order of thousands!

GM: MLE and Bayesian Learning

- Probabilistic statements of Θ is conditioned on the values of the observed variables \mathbf{A}_{obs} and prior $p(\theta; \chi)$

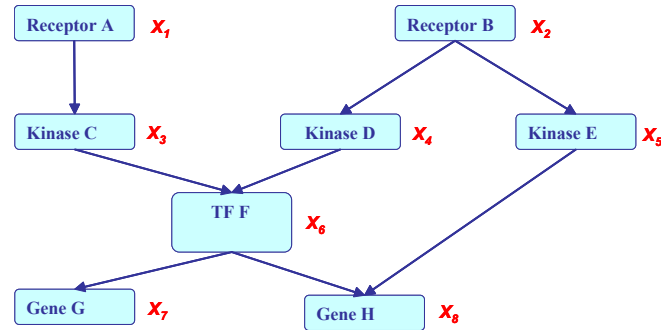


$$\Theta_{\text{Bayes}} = \int \Theta p(\Theta | \mathbf{A}, \chi) d\Theta$$



Why we may favor a PGM?

- 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.,



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

- 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**



5 min break ... and enjoy the video by the imposteriors

Mark Glickman

Senior Lecturer on Statistics
Department of Statistics
Harvard University



Bradley P. Carlin

Professor of Biostatistics

Mayo Professor in Public Health

UNIVERSITY OF MINNESOTA



Jennifer L. Hill

Professor of Applied Statistics and Data Science



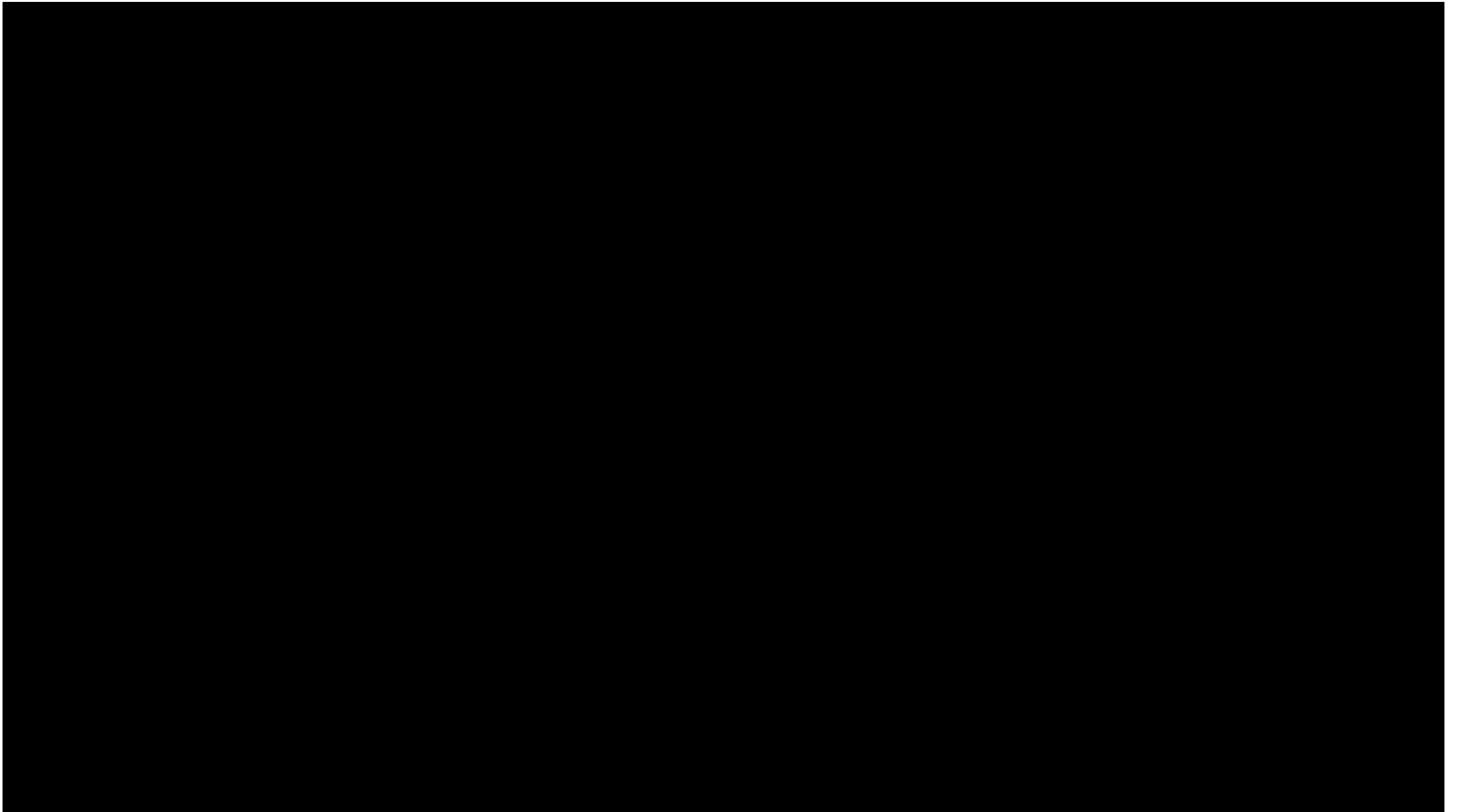
Michael I. Jordan

Pehong Chen Distinguished Professor
Department of EECS
Department of Statistics
AMP Lab
Berkeley AI Research Lab
University of California, Berkeley



Donald Hedeker, PhD

Professor of Biostatistics, University of
Chicago



You don't have to be Bayesian to enjoy the class

So What Is a PGM After All?

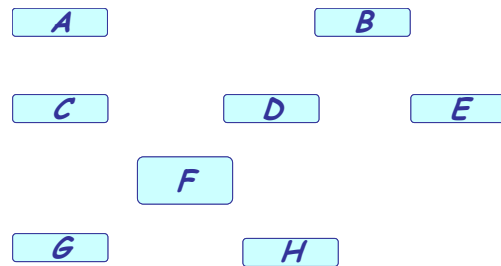
In a nutshell:

PGM = Multivariate Statistics + Structure

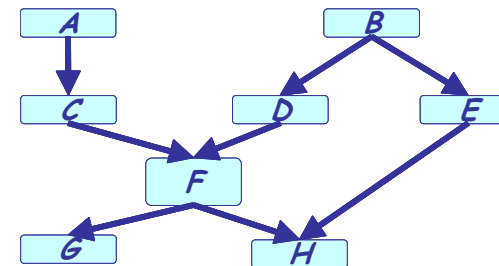
GM = Multivariate Obj. Func. + Structure

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**



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



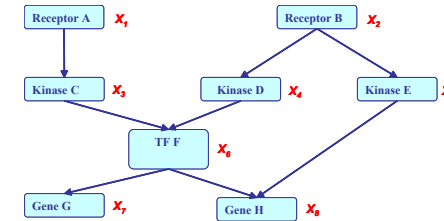
$$P(X_{1:8}) = P(X_1)P(X_2)P(X_3 | X_1X_2)P(X_4 | X_2)P(X_5 | X_2) \\ P(X_6 | X_3, X_4)P(X_7 | X_6)P(X_8 | X_5, X_6)$$

- 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

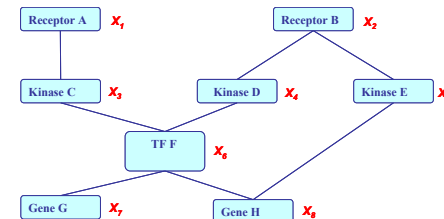
- **Directed edges** give **causality** relationships (Bayesian Network or Directed Graphical Model):

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



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

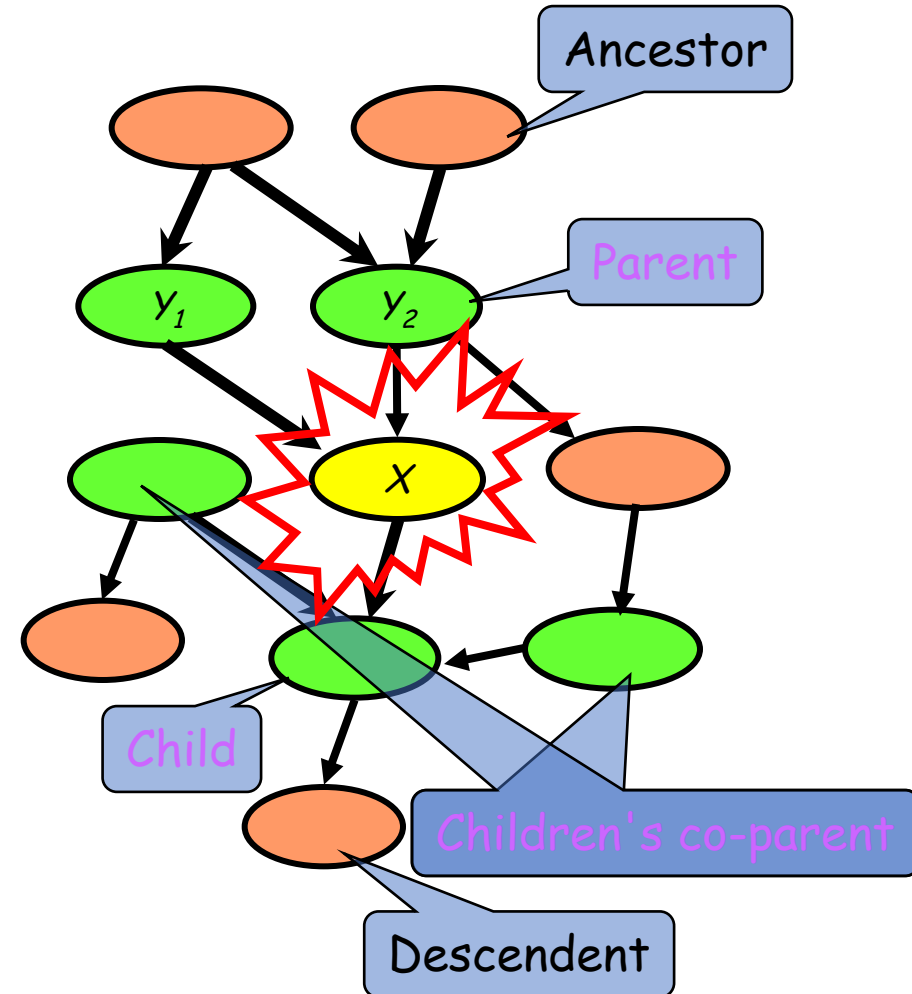
$$\begin{aligned}
 &P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8) \\
 &= \frac{1}{Z} \exp \{E(X_1) + E(X_2) + E(X_3, X_1) + E(X_4, X_2) + E(X_5, X_2) \\
 &\quad + E(X_6, X_3, X_4) + E(X_7, X_6) + E(X_8, X_5, X_6)\}
 \end{aligned}$$



Bayesian Networks

Structure: *DAG*

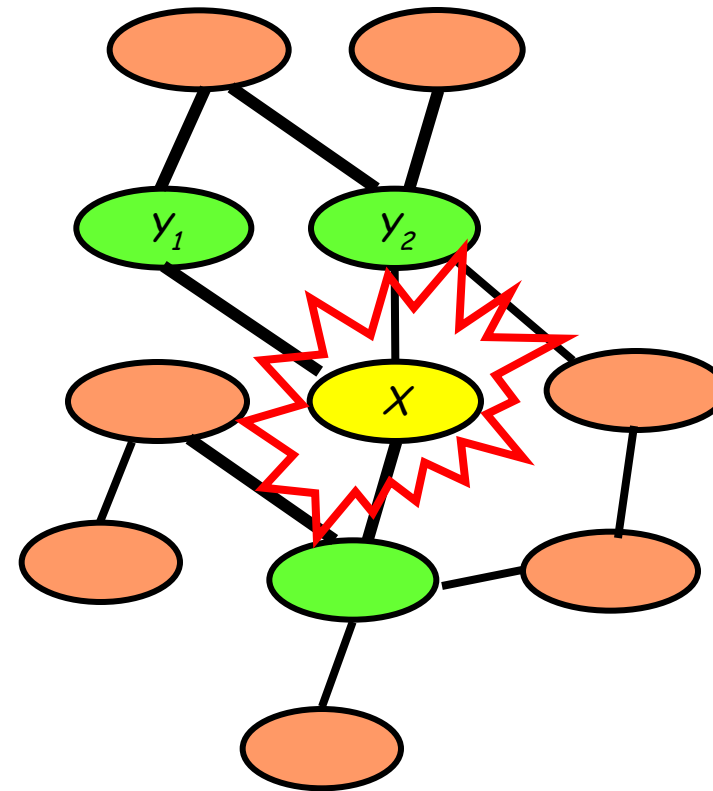
- Meaning: a node is **conditionally independent** of every other node in the network outside its **Markov blanket**
- Local conditional distributions (**CPD**) and the **DAG** completely determine the **joint** dist.
- Give causality relationships, and facilitate a generative process



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 \mathcal{G} ,

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

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

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

GMs are your old friends

Density estimation

Parametric and nonparametric methods

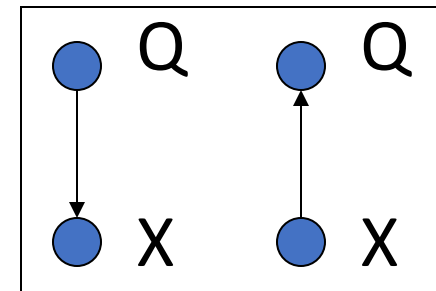
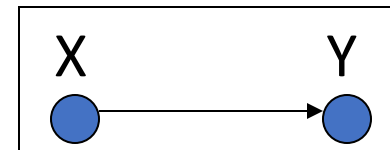
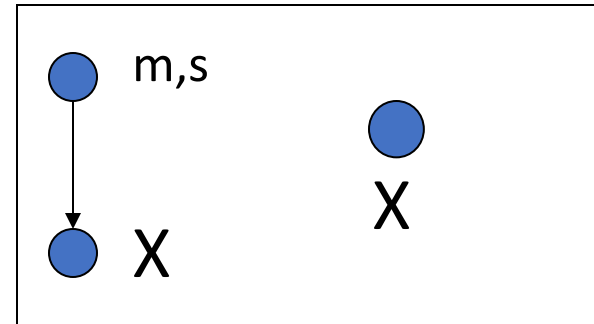
Regression

Linear, conditional mixture, nonparametric

Classification

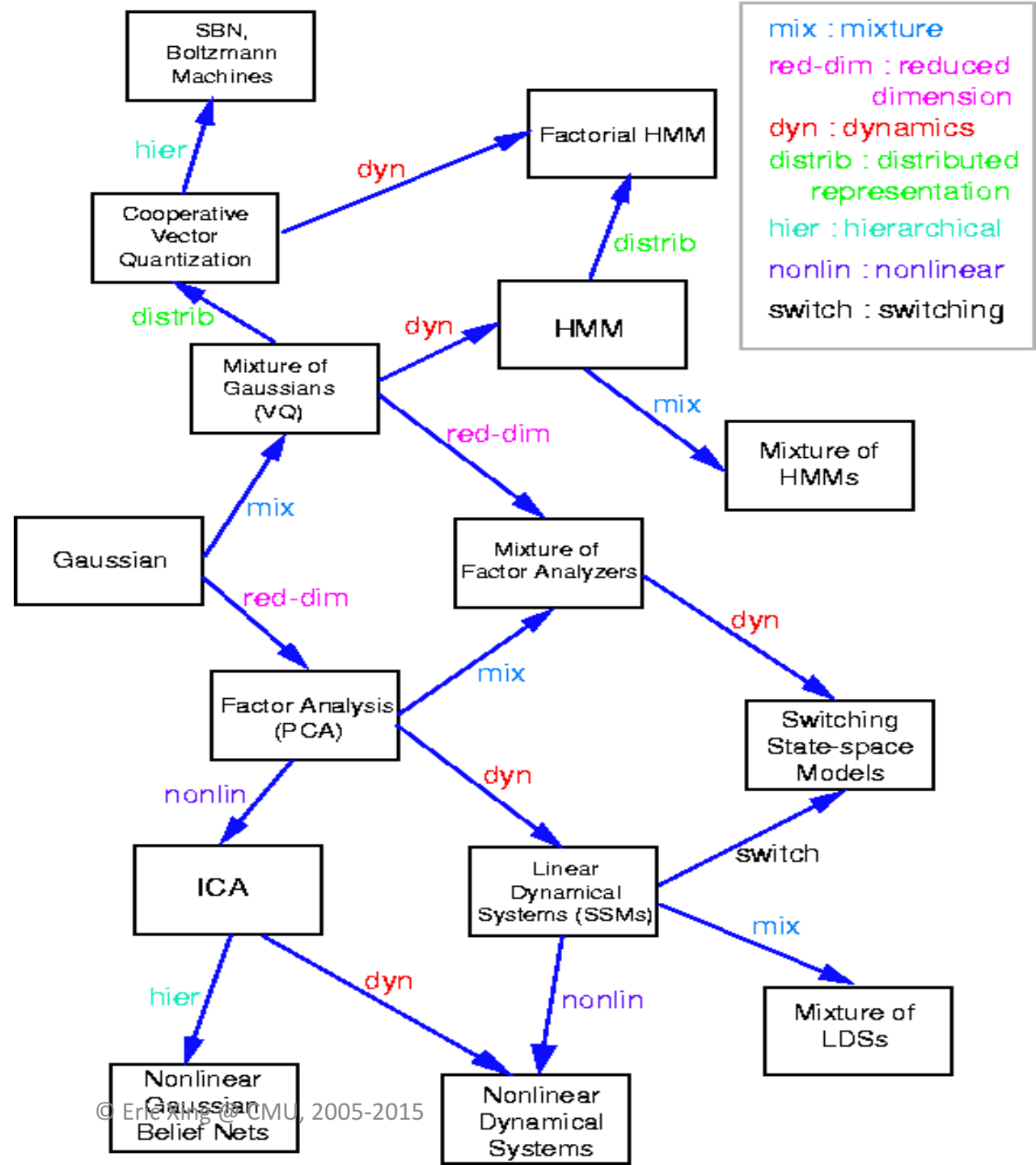
Generative and discriminative approach

Clustering



An (incomplete) genealogy of graphical models

(Picture by Zoubin
Ghahramani and
Sam Roweis)



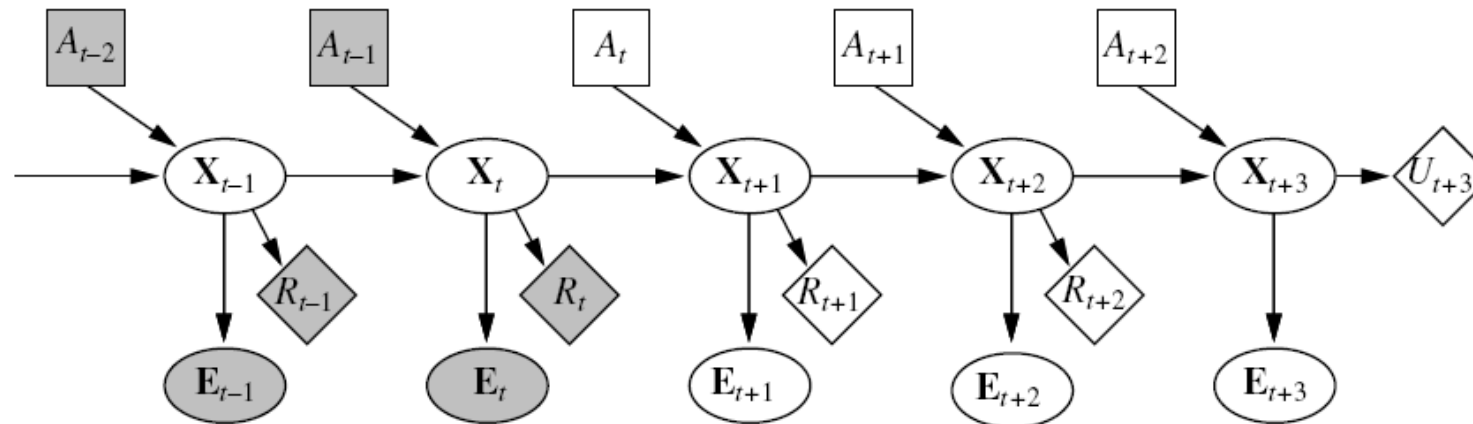
Questions ?

Plan for the Class

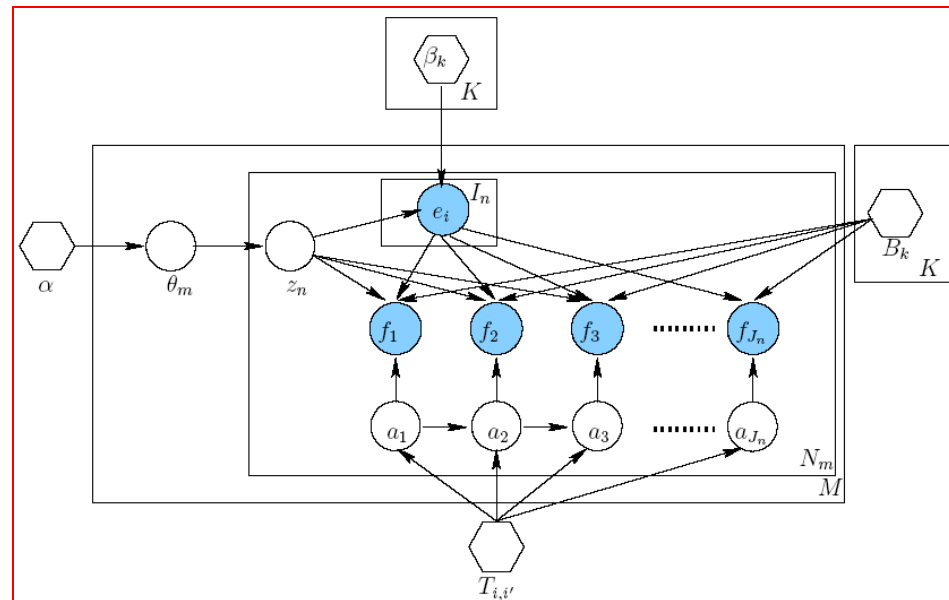
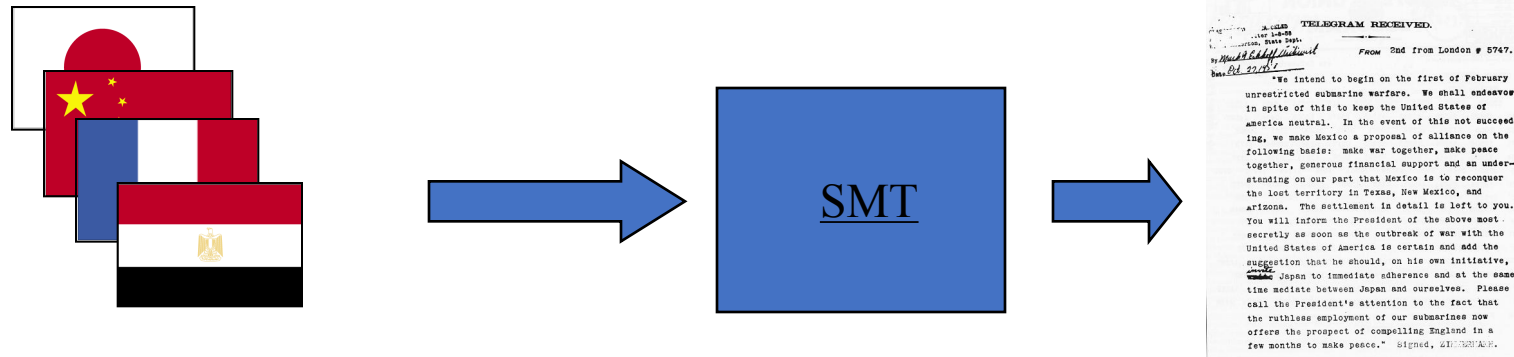
- **Module 1: Representation**
 - Directed and Undirected Graphical model
- **Module 2: Classical Methods of Inference & Learning**
 - Variable Elimination, Factor and message passing, GLM, Learning fully observed un-/directed graphical model, EM
- **Module 3: Graphical Model in Application**
 - HMM, CRF, Topic Modeling, Factor Analysis, Spike and Slab model
- **Module 4: Approximate Inference**
 - LBP, Mean field, Gibbs, MCMC
- **Module 5: Deep Learning and Graphical Models**
 - VAE, GAN, BiGAN and friends
- **Module 6: Scalability and Optimization**
 - SDG, SVI
- **Module 7: Spectral and non-parametric view**
 - GP, DP, IBP, HDP, other spectral approaches

Fancier GMs: reinforcement learning

- Partially observed Markov decision processes (POMDP)

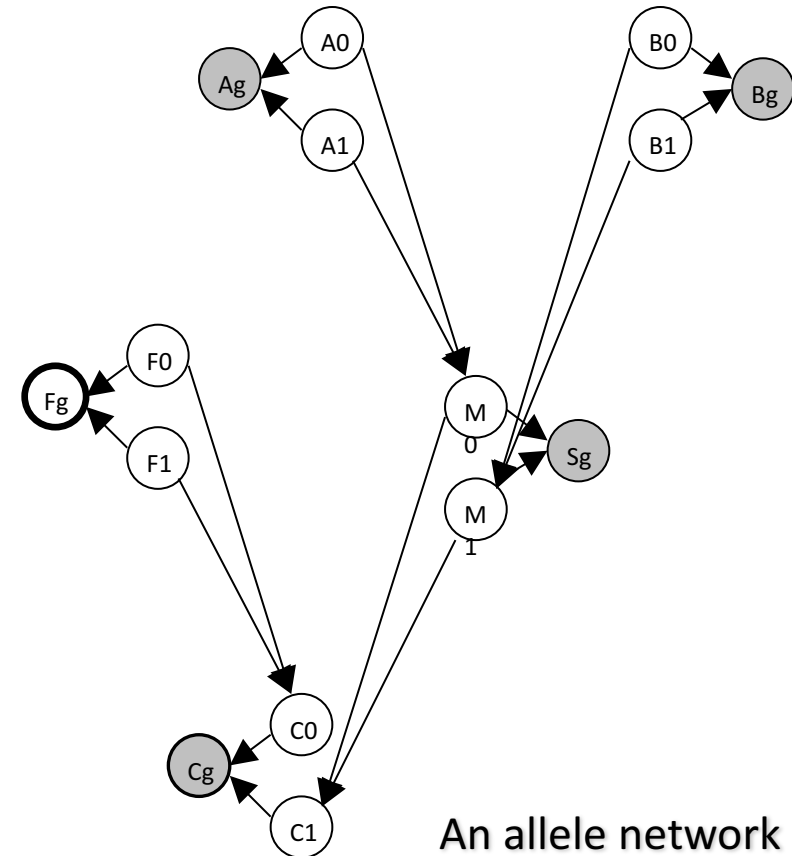
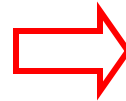
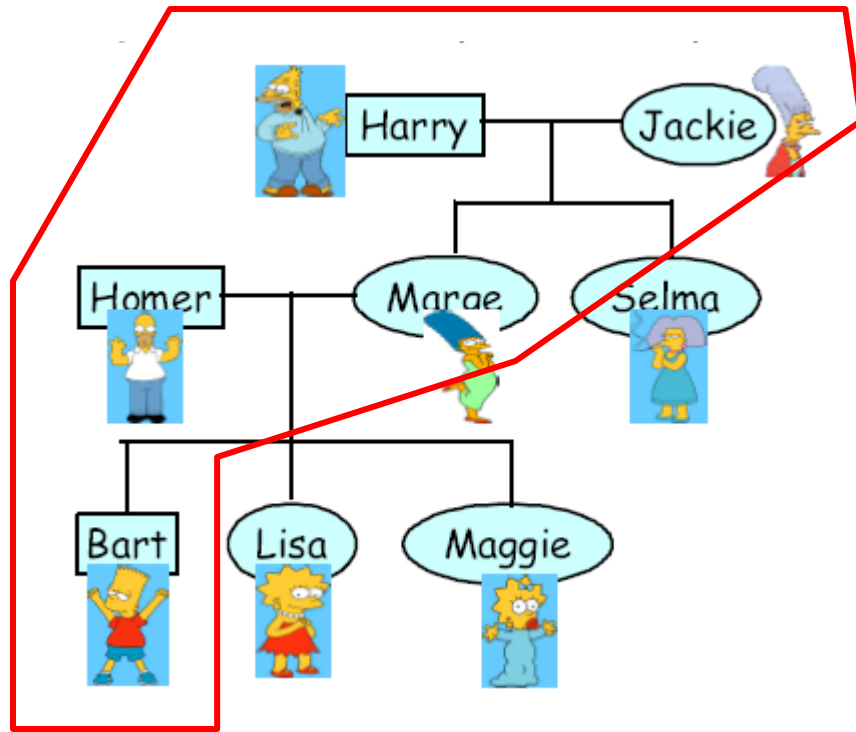


Fancier GMs: machine translation



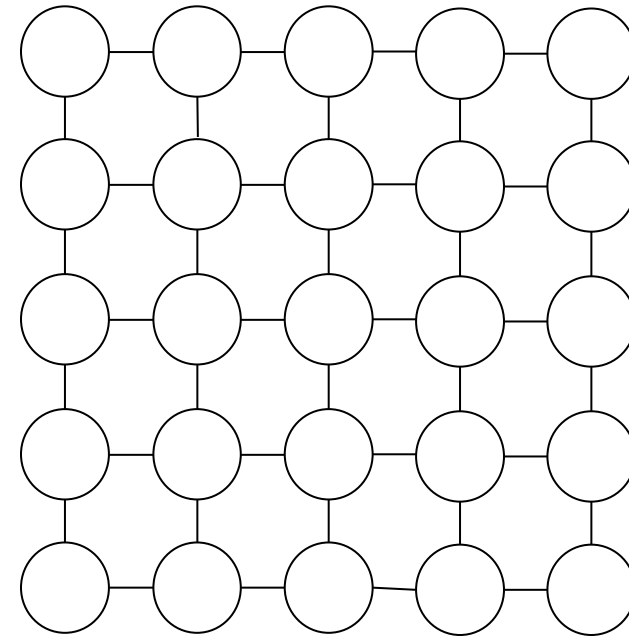
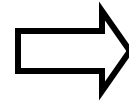
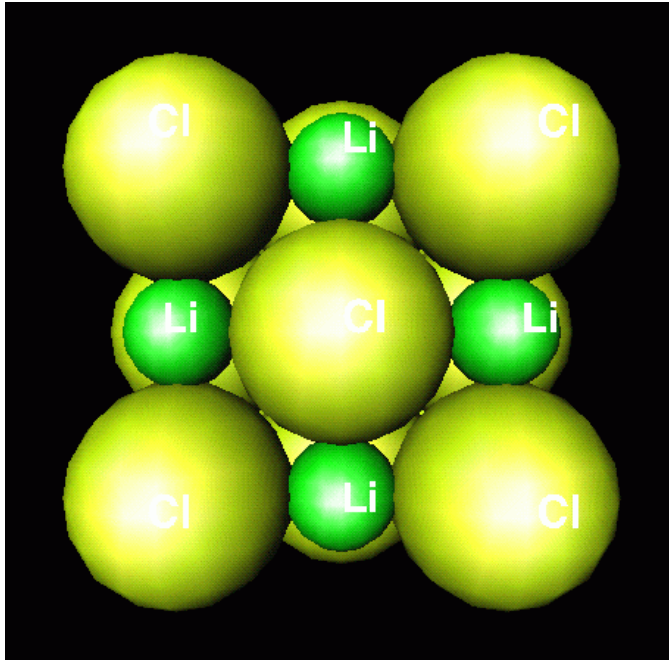
The HM-BiTAM model
(B. Zhao and E.P Xing,
ACL 2006)

Fancier GMs: genetic pedigree



An allele network

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,
 - 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