School of Computer Science
Carnegie Mellon

Probabilistic Graphical Models

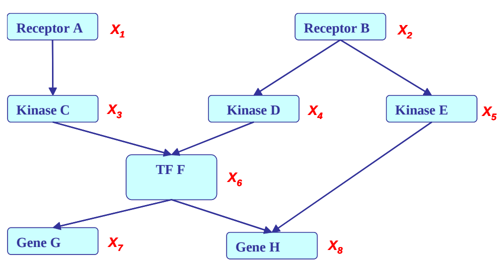
Introduction to GM

and

Directed GMs: Bayesian Networks

Eric Xing

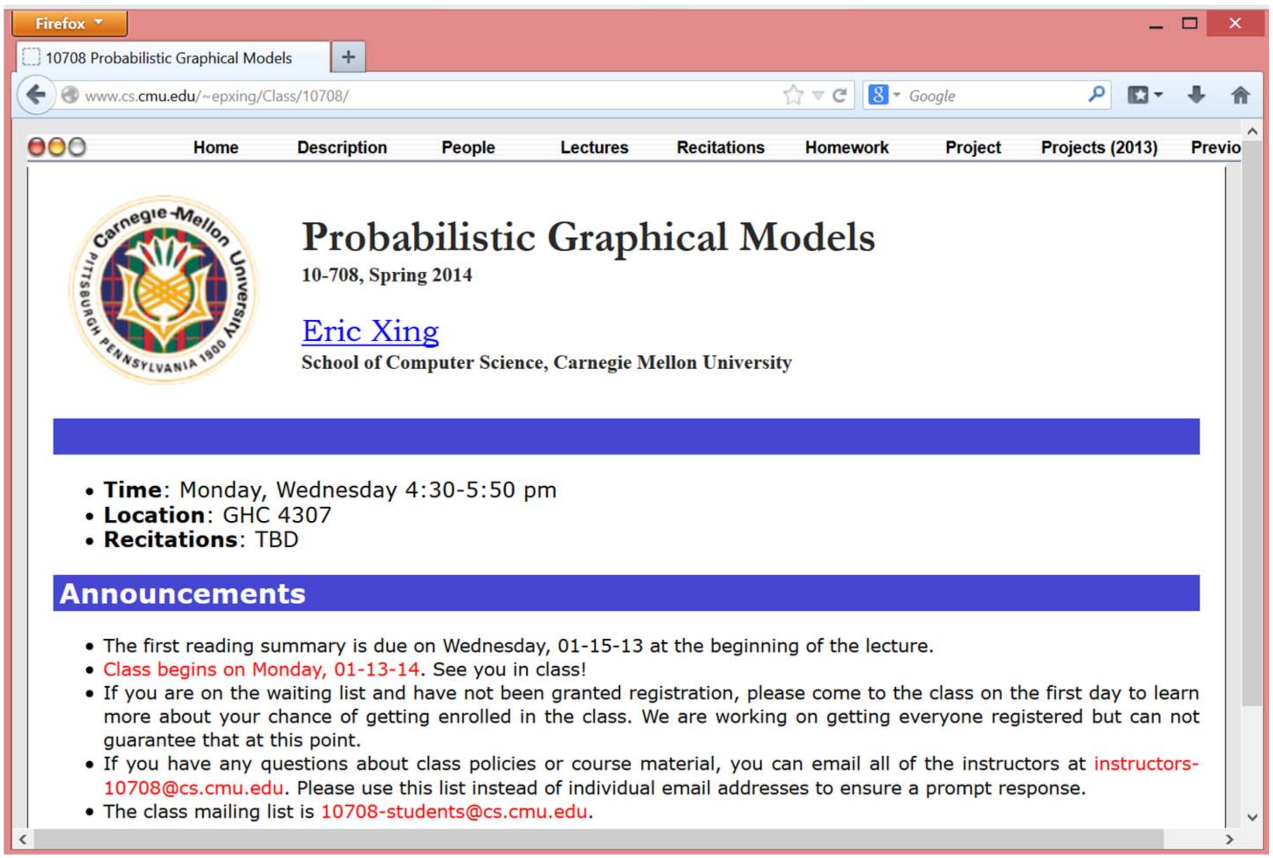
Lecture 1, January 13, 2014



Reading: see class homepage

Logistics

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



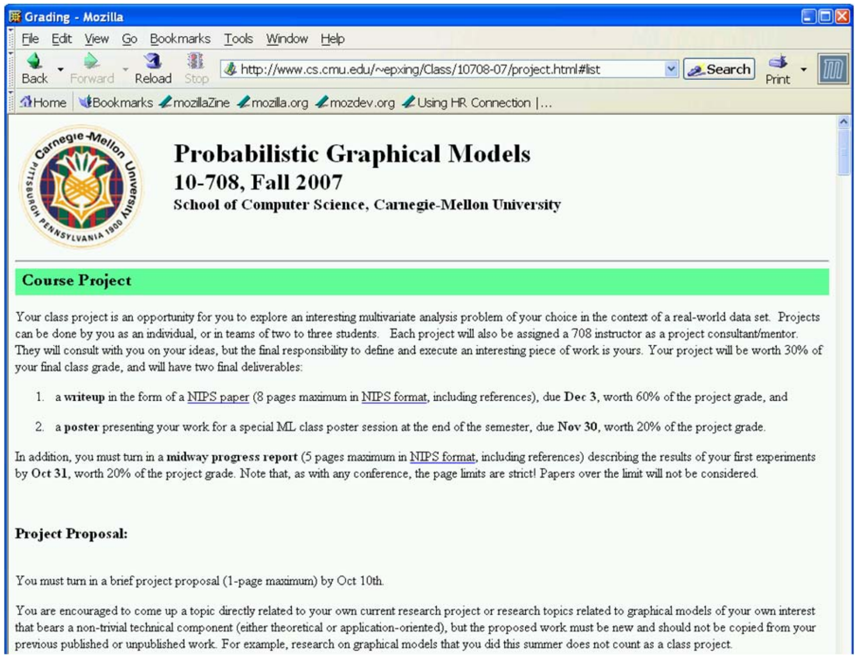
Logistics

- 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: instructor-10708@cs.cmu.edu
  - Class announcements list: 10708-students@cs.cmu.edu.
- TA:
  - Willie Neiswanger, GHC 8011, Office hours: TBA
  - Micol Marchetti-Bowick, GHC 8003, Office hours: TBA
  - Dai Wei, GHC 8011, Office hours: TBA
- Guest Lecturers:
  - TBA
- Class Assistant:
  - Michael Martins, GHC 8001, x8-5527
- Instruction aids: Canvas

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 (rocord-breaking) distributed Deep Network on Petuum and apply to ImageNet 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
    - a distributed sampling scheme with convergence guarantee
  - 3-member team to be formed in the first two weeks, proposal, mid-way presentation, poster & demo, final report, peer review → possibly conference submission !

Past projects:



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

- **Winner of the 2005 project:**

  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)**. (**Recipient of the BEST PAPER Award**)

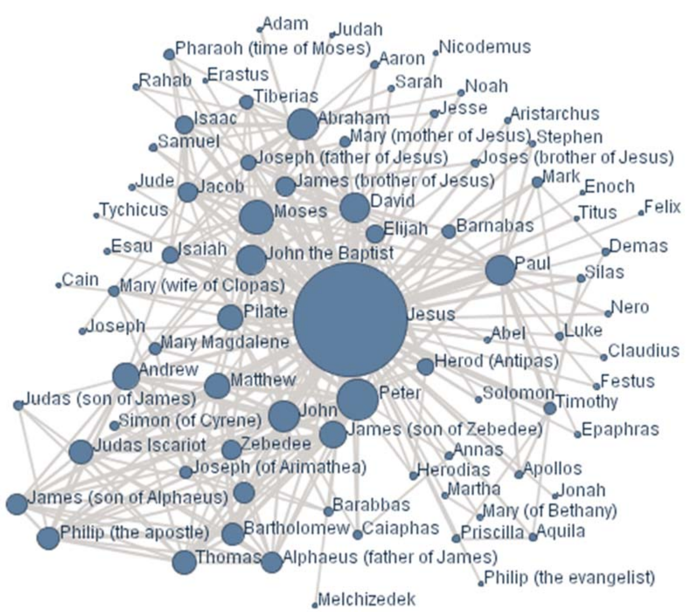
- **Other projects:**

  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



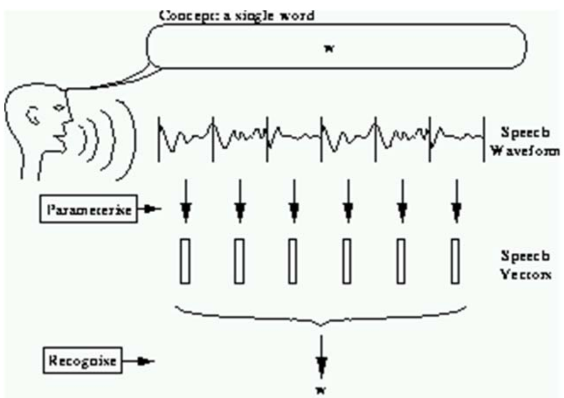
Model

$$\mathcal{M}$$

Data

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

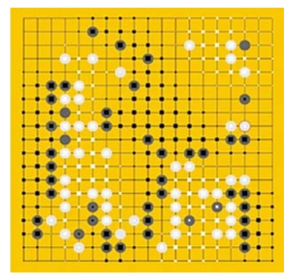
Reasoning under uncertainty!



Speech recognition

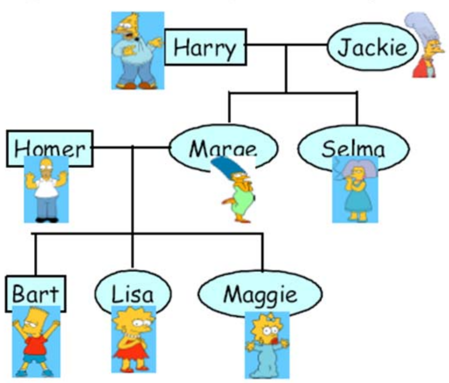
Information retrieval

Computer vision

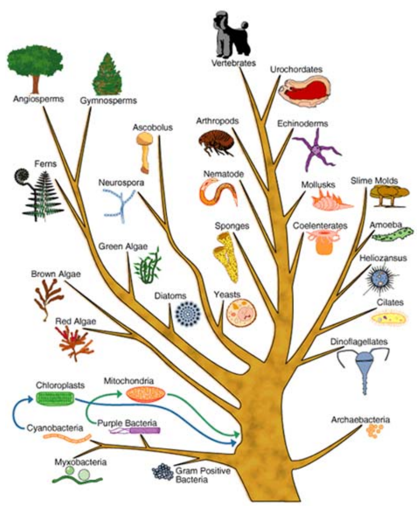


Games

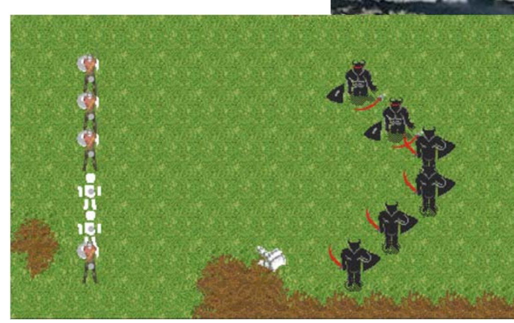
Robotic control



Pedigree



Evolution



Planning

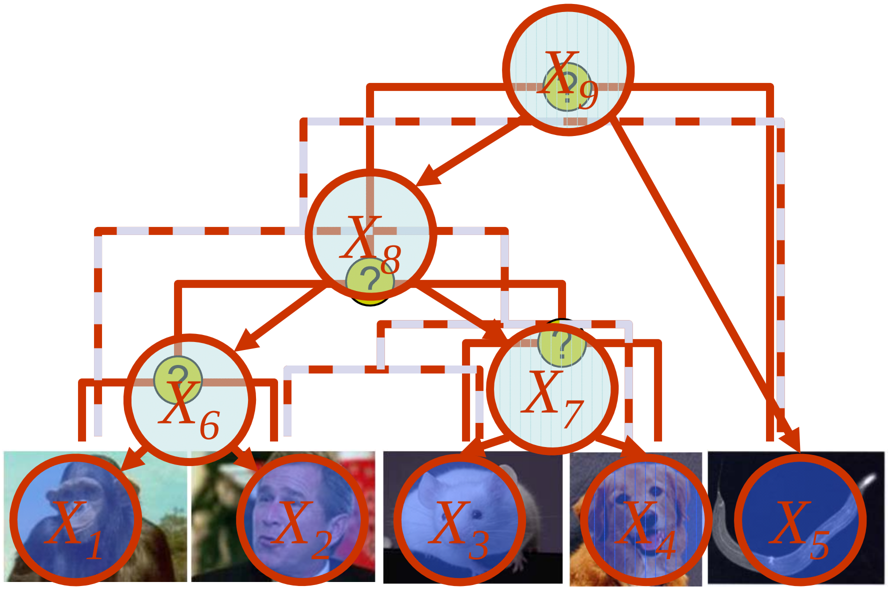
The Fundamental Questions

- 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.: $P(X_i \mid \mathbf{D})$

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

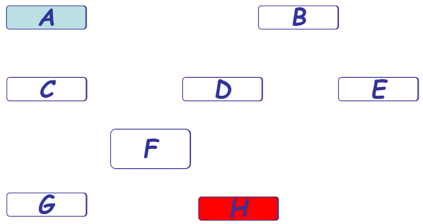
e.g.: $\mathcal{M} = \arg\max_{\mathcal{M} \in \mathrm{M}} F(\mathbf{D}; \mathcal{M})$



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?**

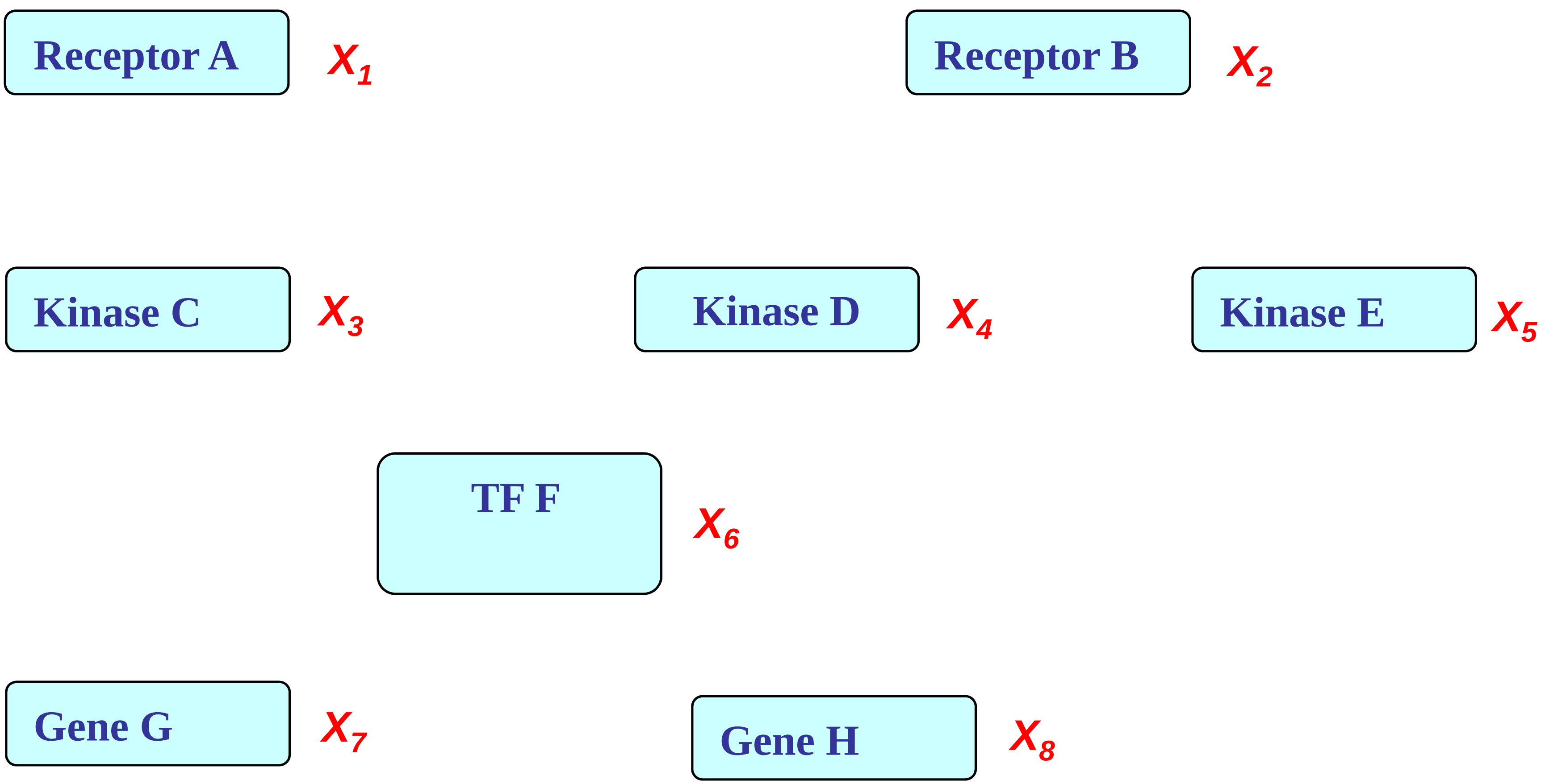
- Learning: where do we get all this probabilities?
  - Maximal-likelihood estimation? but how many data do we need?
  - Are there other est. principles?
  - Where do we put domain knowledge in terms of plausible relationships between variables, and plausible values of the probabilities?

- 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?

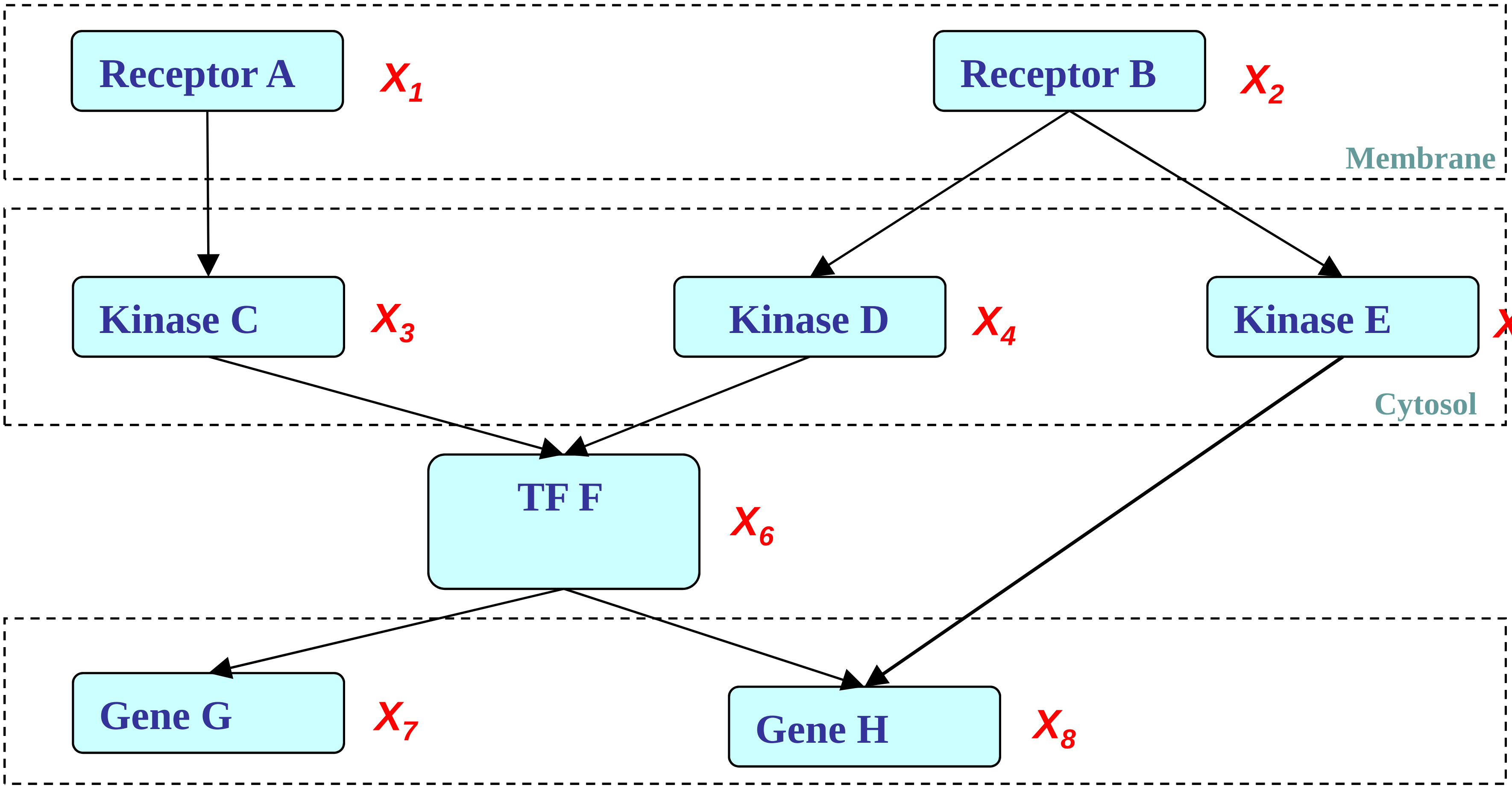
--- Multivariate Distribution in High-D Space

- A possible world for cellular signal transduction:



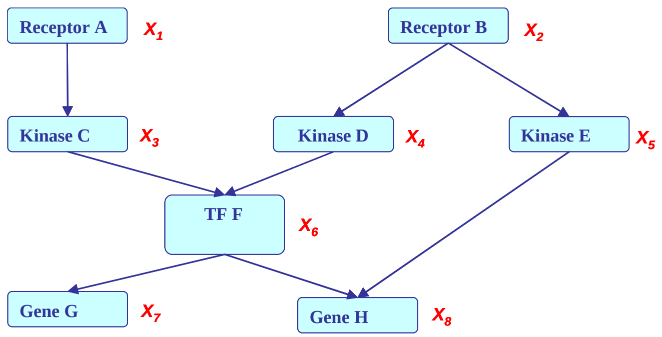
GM: Structure Simplifies Representation

- 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.,



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

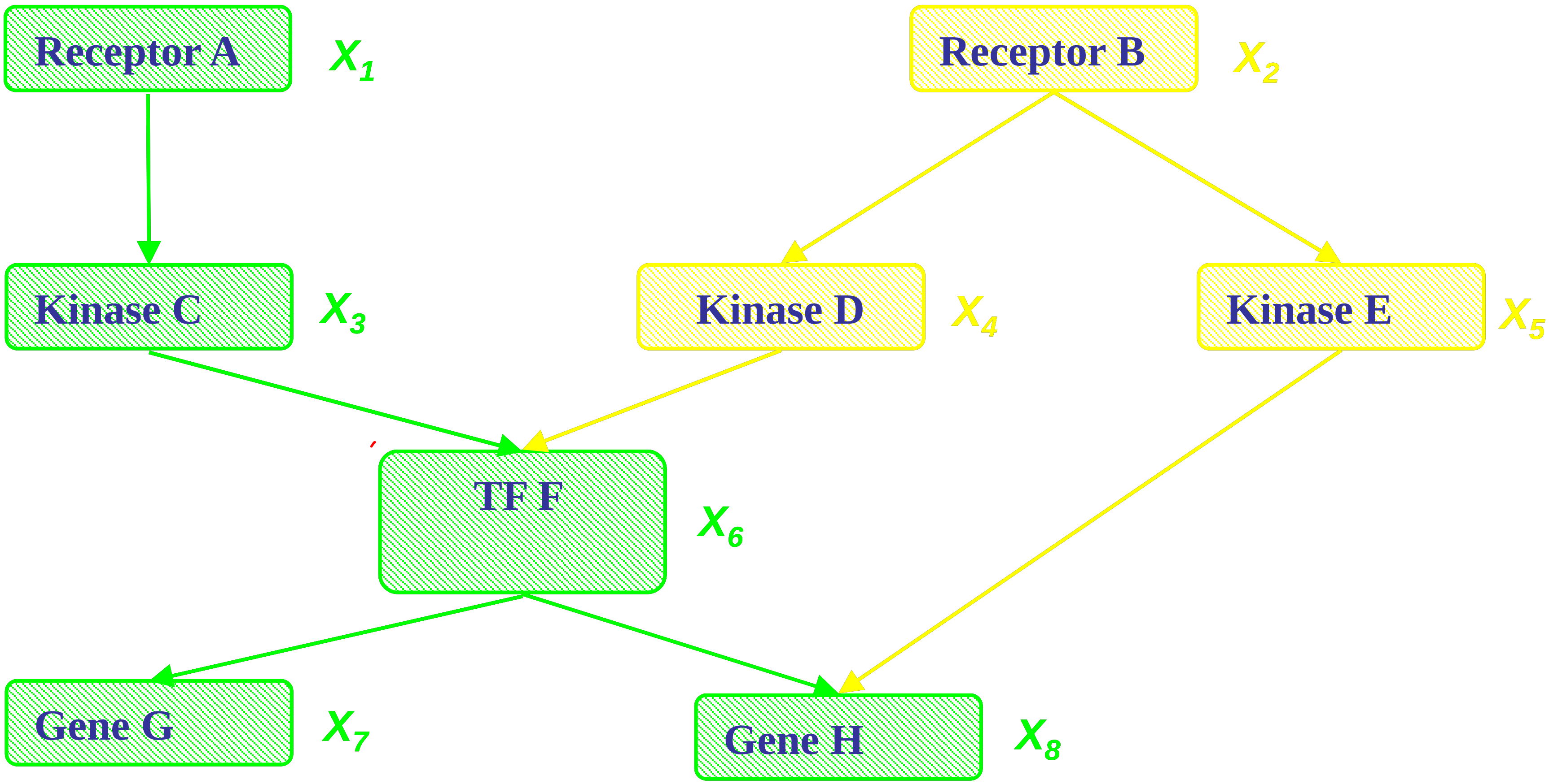
$$= P(X_1)\, P(X_2)\, P(X_3 | X_1)\, P(X_4 | X_2)\, P(X_5 | X_2)\, P(X_6 | X_3, X_4)\, P(X_7 | X_6)\, P(X_8 | X_5, X_6)$$

Stay tune for what are these independencies!

- 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

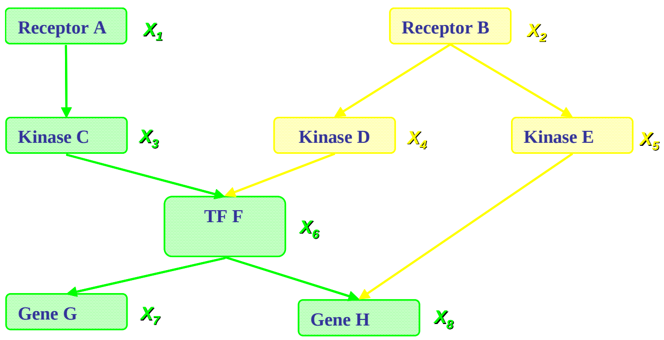


More Data Integration

- Text + Image + Network ➔ Holistic Social Media
- Genome + Proteome + Transcritome + 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.,



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

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

- 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:

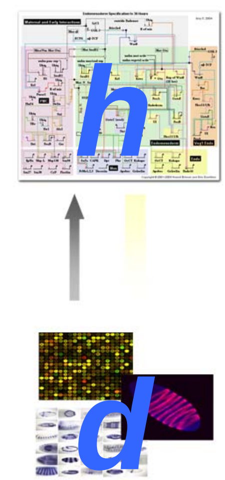
Posterior probability

Likelihood

Prior probability

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

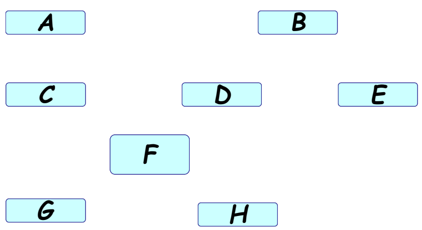
Sum over space of hypotheses



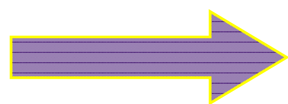
- This allows us to capture uncertainty about the model in a principled way
- But how can we specify and represent a complicated model?
  - **Typically the number of genes need to be modeled are in the order of thousands!**

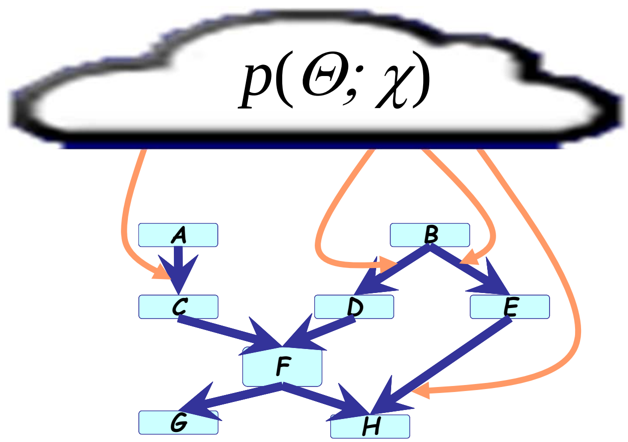
GM: MLE and Bayesian Learning

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



$$\mathbf{A} = \begin{array}{l} (A,B,C,D,E,\ldots)=(T,F,F,T,F,\ldots) \\ (A,B,C,D,E,\ldots)=(T,F,T,T,F,\ldots) \\ \ldots\ldots\ldots \\ (A,B,C,D,E,\ldots)=(F,T,T,T,F,\ldots) \end{array}$$





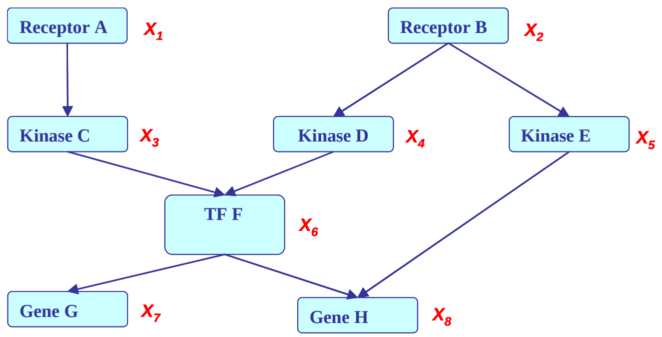
C	D	P(F | C,D)	
c	d	0.9	0.1
c	$\bar{d}$	0.2	0.8
$\bar{c}$	d	0.9	0.1
$\bar{c}$	$\bar{d}$	0.01	0.99

$$\underbrace{p(\Theta \mid \mathbf{A}; \chi)}_{\text{posterior}} \propto \underbrace{p(\mathbf{A} \mid \Theta)}_{\text{likelihood}} \underbrace{p(\Theta; \chi)}_{\text{prior}}$$

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

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.,

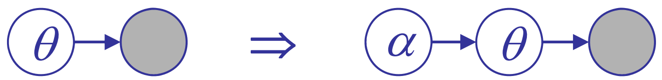


$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= P(X_1)\, P(X_2)\, P(X_3| X_1)\, P(X_4| X_2)\, P(X_5| X_2)\, P(X_6| X_3, X_4)\, P(X_7| X_6)\, P(X_8| X_5, X_6)$$

- Why we may favor a PGM?
  - 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



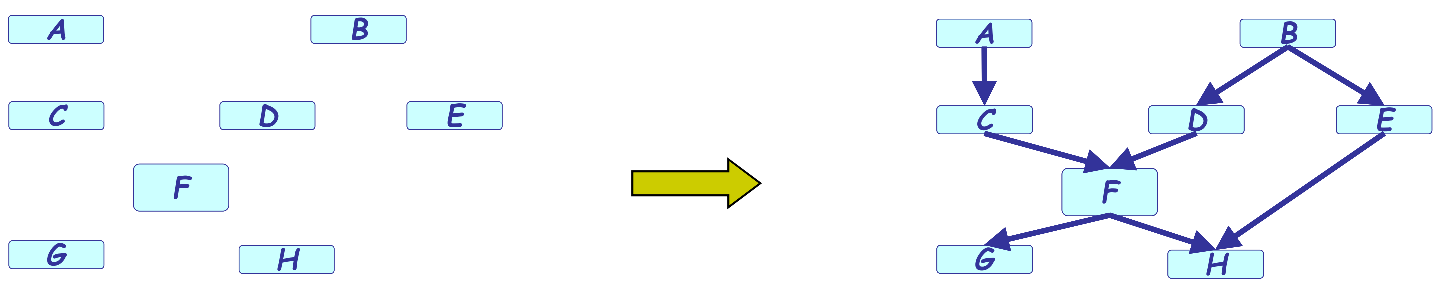
So What is a Graphical Model?

In a nutshell:

GM = Multivariate Statistics + Structure

What is a Graphical Model?

- 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***



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

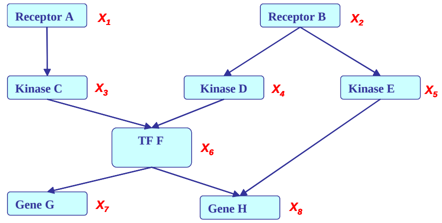
$$P(X_{1:8}) = P(X_1)P(X_2)P(X_3 \mid X_1X_2)P(X_4 \mid X_2)P(X_5 \mid X_2)$$
$$P(X_6 \mid X_3,X_4)P(X_7 \mid X_6)P(X_8 \mid X_5,X_6)$$

- 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

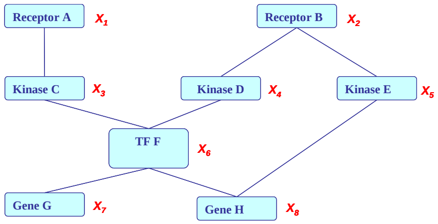
- Directed edges give causality relationships (**Bayesian Network** or **Directed Graphical Model**):

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



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

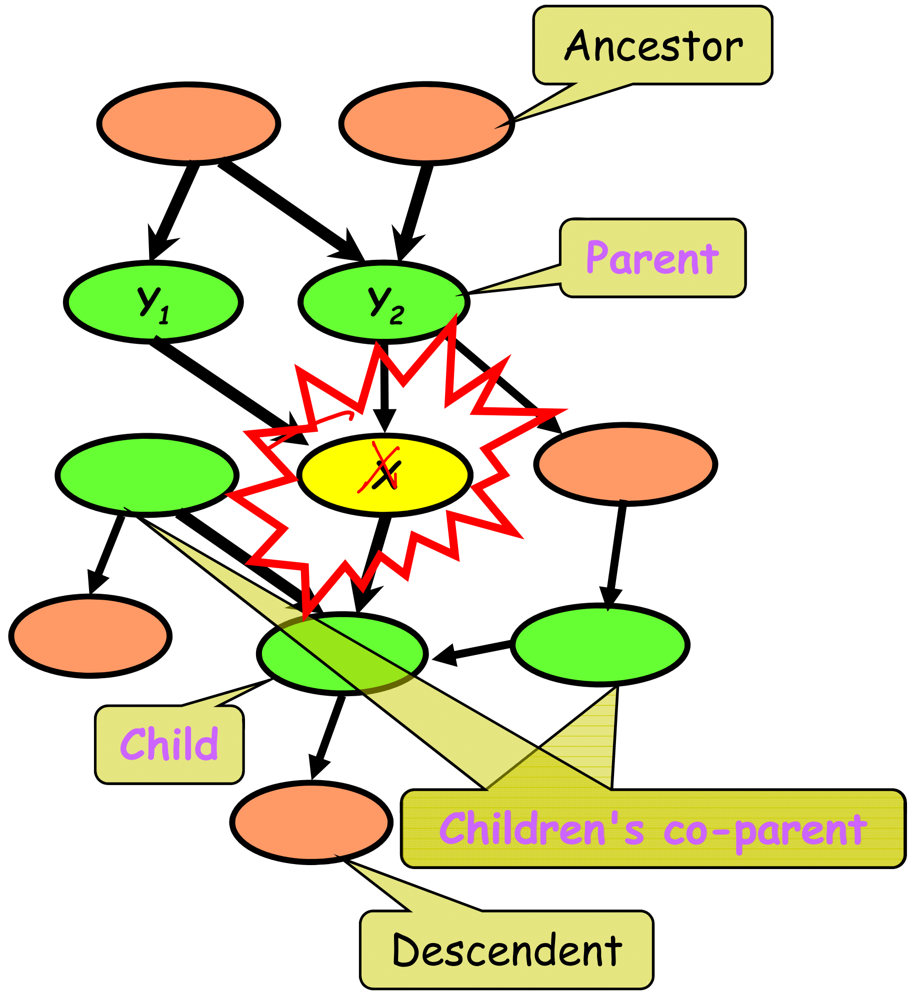
$$P(X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8)$$
$$= 1/Z \exp\{E(X_1)+E(X_2)+E(X_3, X_1)+E(X_4, X_2)+E(X_5, X_2) + E(X_6, X_3, X_4)+E(X_7, X_6)+E(X_8, X_5, X_6)\}$$



Bayesian 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



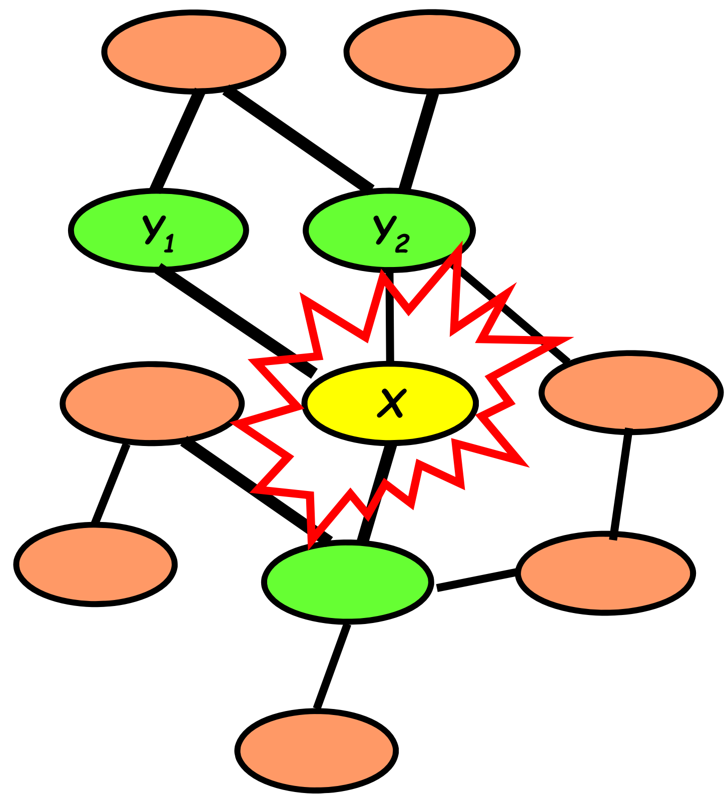
Markov Random Fields

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),

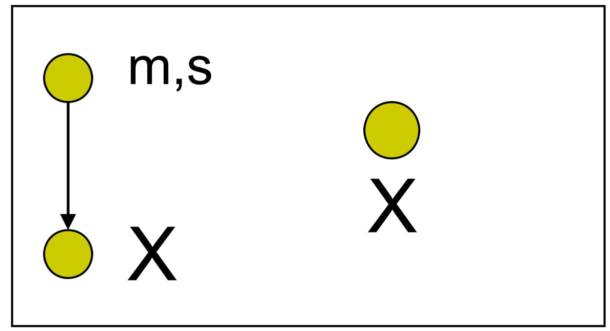
  Let $\mathcal{D}_2$ denote the family of all distributions that factor according to G,

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

GMs are your old friends

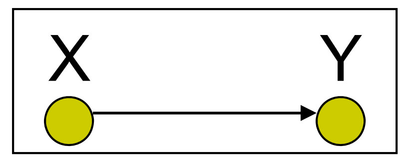
Density estimation

Parametric and nonparametric methods



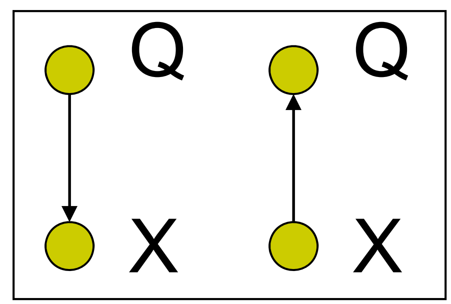
Regression

Linear, conditional mixture, nonparametric



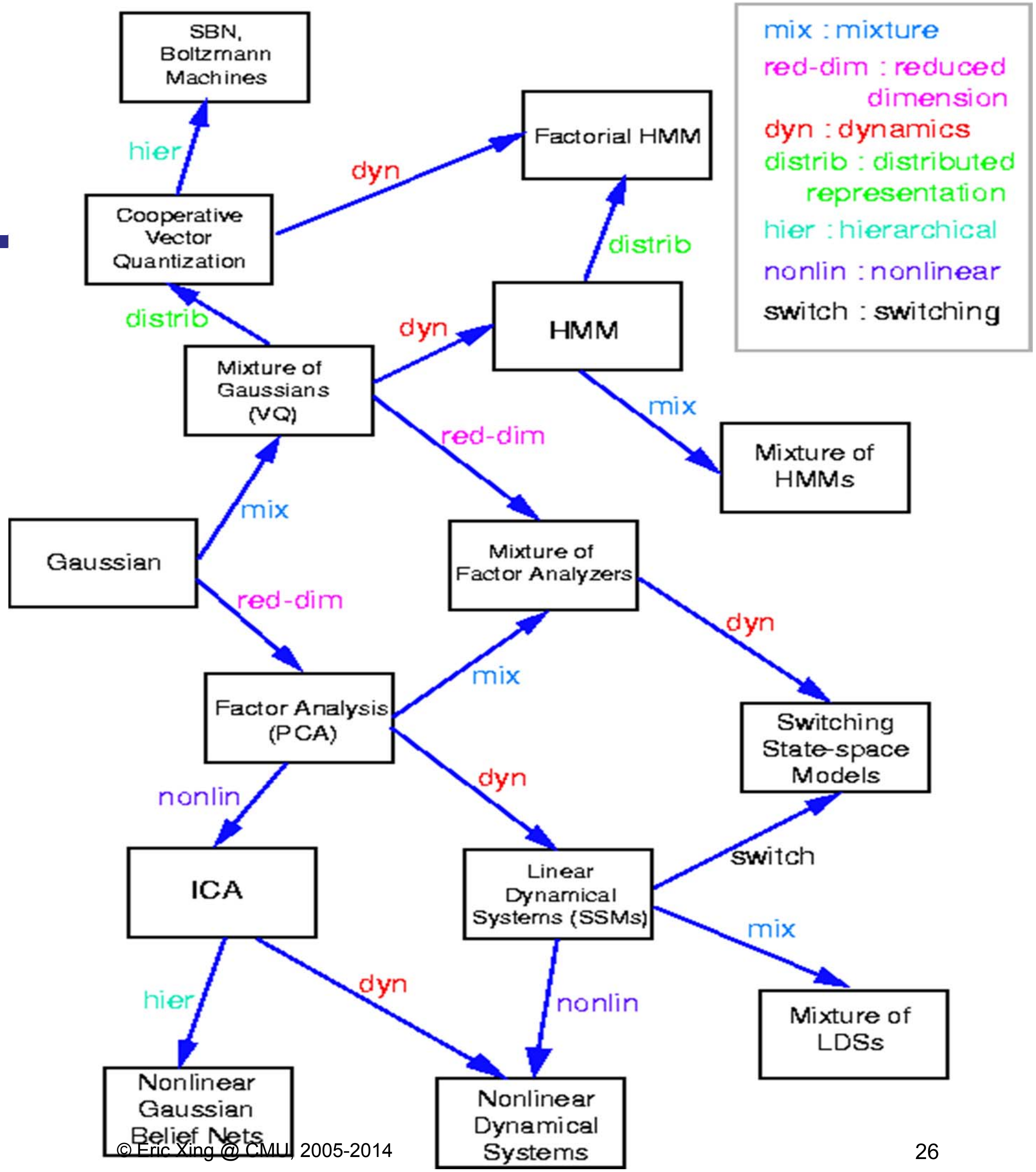
Classification

Generative and discriminative approach



Clustering

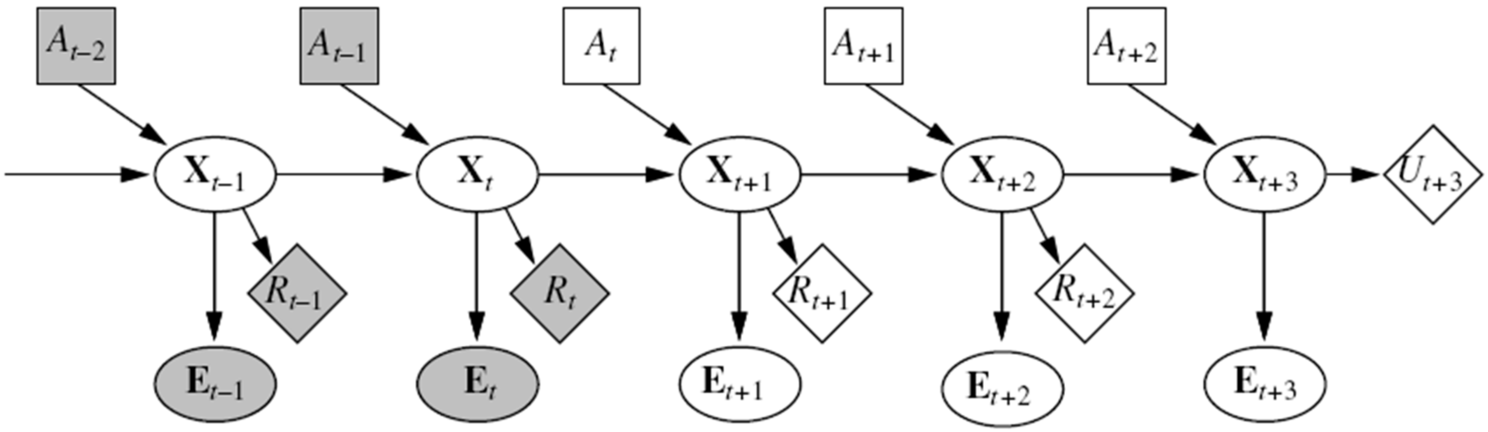
An (incomplete) genealogy of graphical models



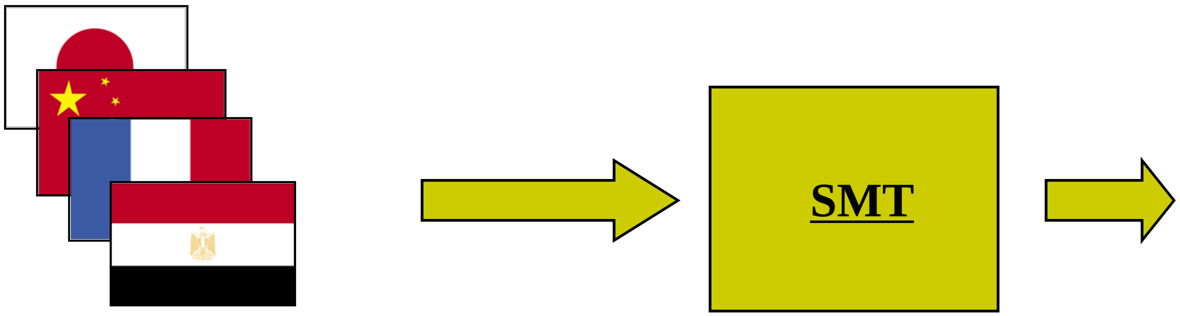
(Picture by Zoubin Ghahramani and Sam Roweis)

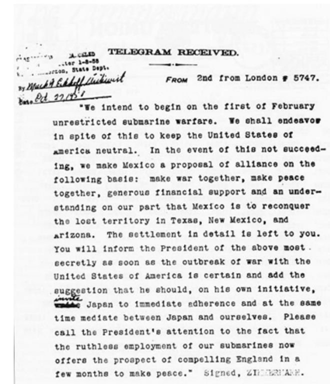
Fancier GMs: reinforcement learning

- Partially observed Markov decision processes (POMDP)



Fancier GMs: machine translation

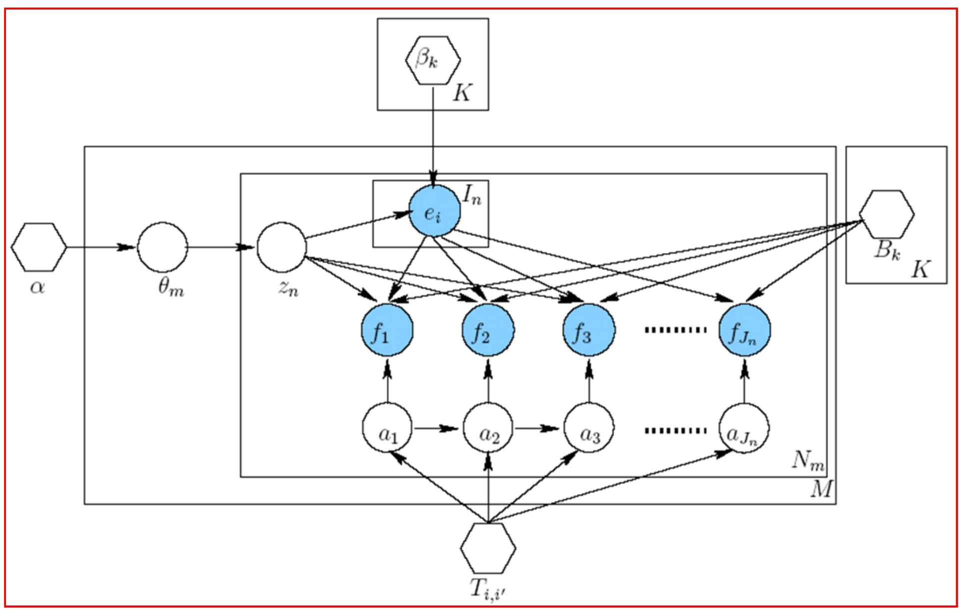




TELEGRAM RECEIVED.

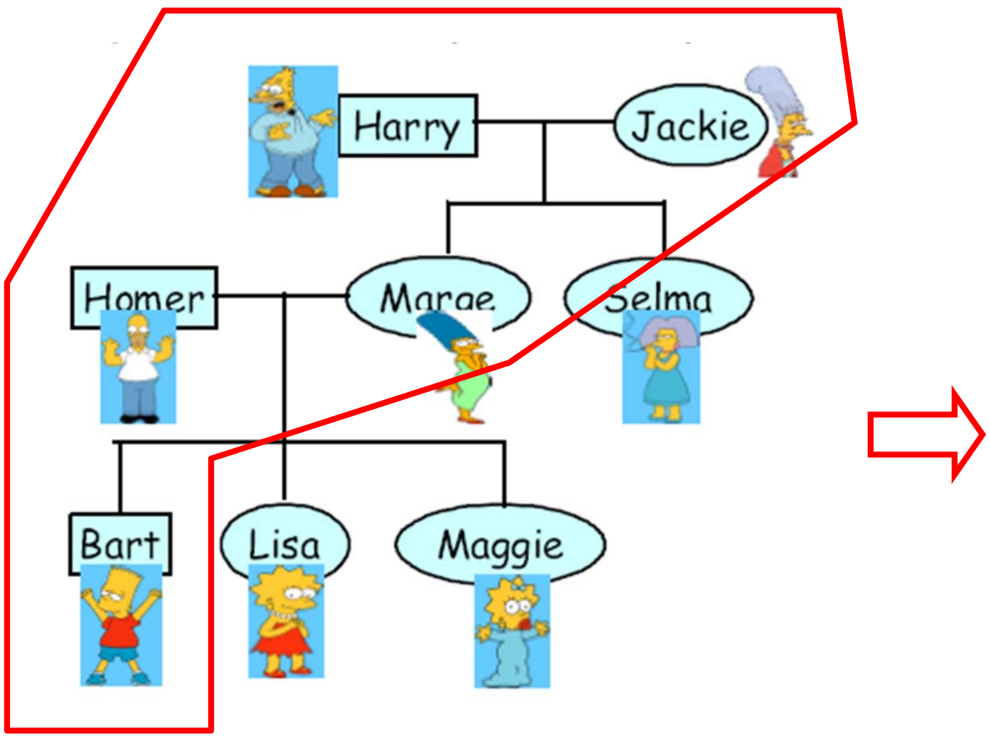
FROM 2nd from London # 5747.

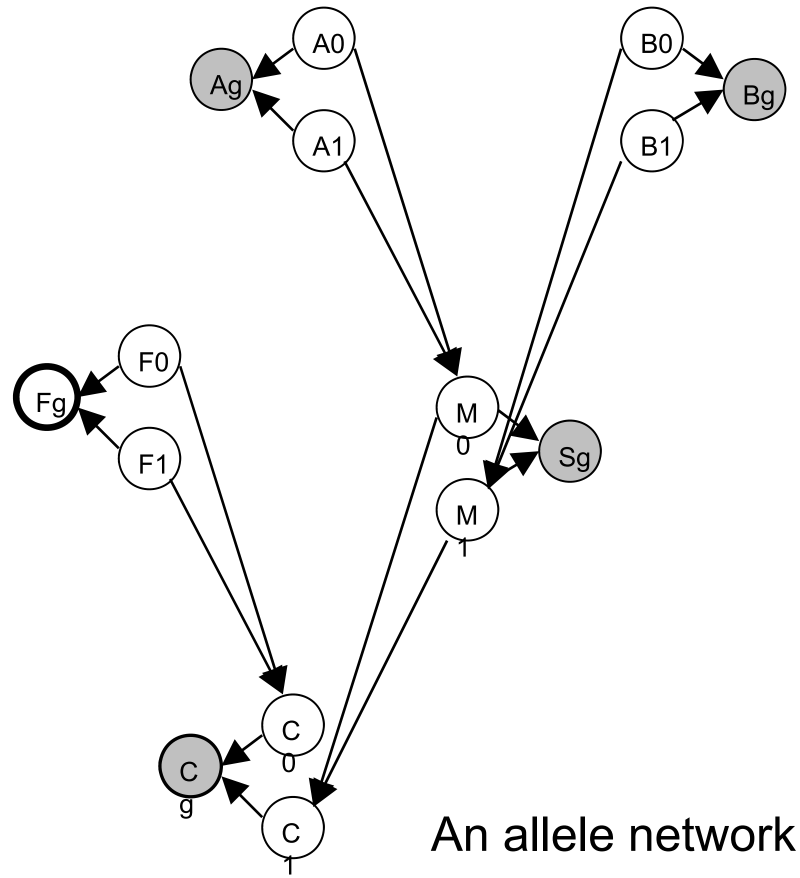
"We intend to begin on the first of February unrestricted submarine warfare. We shall endeavor in spite of this to keep the United States of America neutral. In the event of this not succeeding, we make Mexico a proposal of alliance on the following basis: make war together, make peace together, generous financial support and an understanding on our part that Mexico is to reconquer the lost territory in Texas, New Mexico, and Arizona. The settlement in detail is left to you. You will inform the President of the above most secretly as soon as the outbreak of war with the United States of America is certain and add the suggestion that he should, on his own initiative, invite Japan to immediate adherence and at the same time mediate between Japan and ourselves. Please call the President's attention to the fact that the ruthless employment of our submarines now offers the prospect of compelling England in a few months to make peace." Signed, ZI[illegible].



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

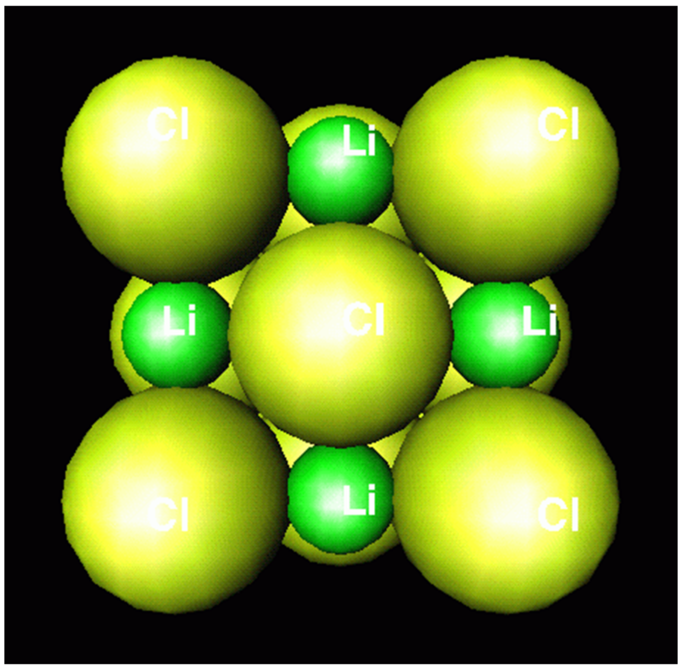
Fancier GMs: genetic pedigree

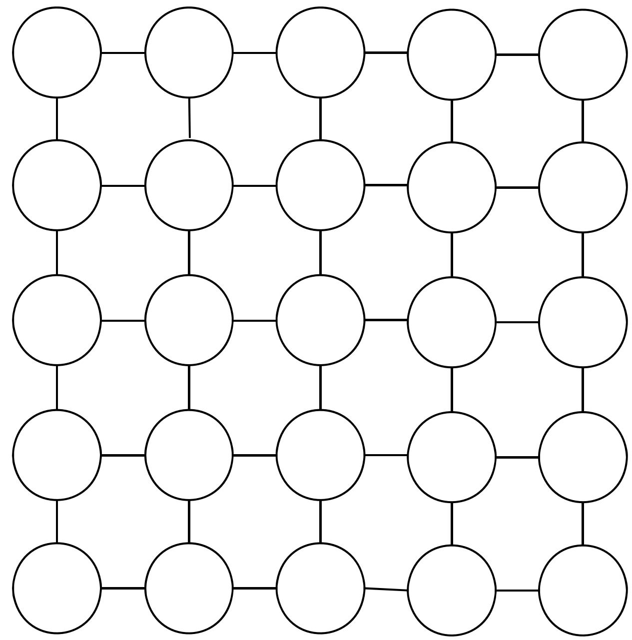




An allele network

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**.

--- M. Jordan

A few myths about graphical models

- They require a localist semantics for the nodes √
- They require a causal semantics for the edges ×
- They are necessarily Bayesian ×
- They are intractable ~~√~~

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
  - Case studies: Popular Bayesian networks and MRFs
    - Multivariate Gaussian Models
    - Hidden Markov Models
    - Mixed-membership, aka, Topic models
    - …
- Advanced topics and latest developments
  - Approximate inference
    - Monte Carlo algorithms
    - Vatiational methods and theories
    - Stochastic algorithms
  - Nonparametric and spectral graphical models, where GM meets kernels and matrix algebra
  - “Infinite” GMs: nonparametric Bayesian models
  - Structured sparsity
  - Margin-based learning of GMs: where GM meets SVM
  - Regularized Bayes: where GM meets SVM, and meets Bayesian, and meets NB …
- Applications