بسم الله الرحمن الرحیم

Deep Learning

دانشگاه صنعتی مالک اشتر

Mohammad Ali Keyvanrad

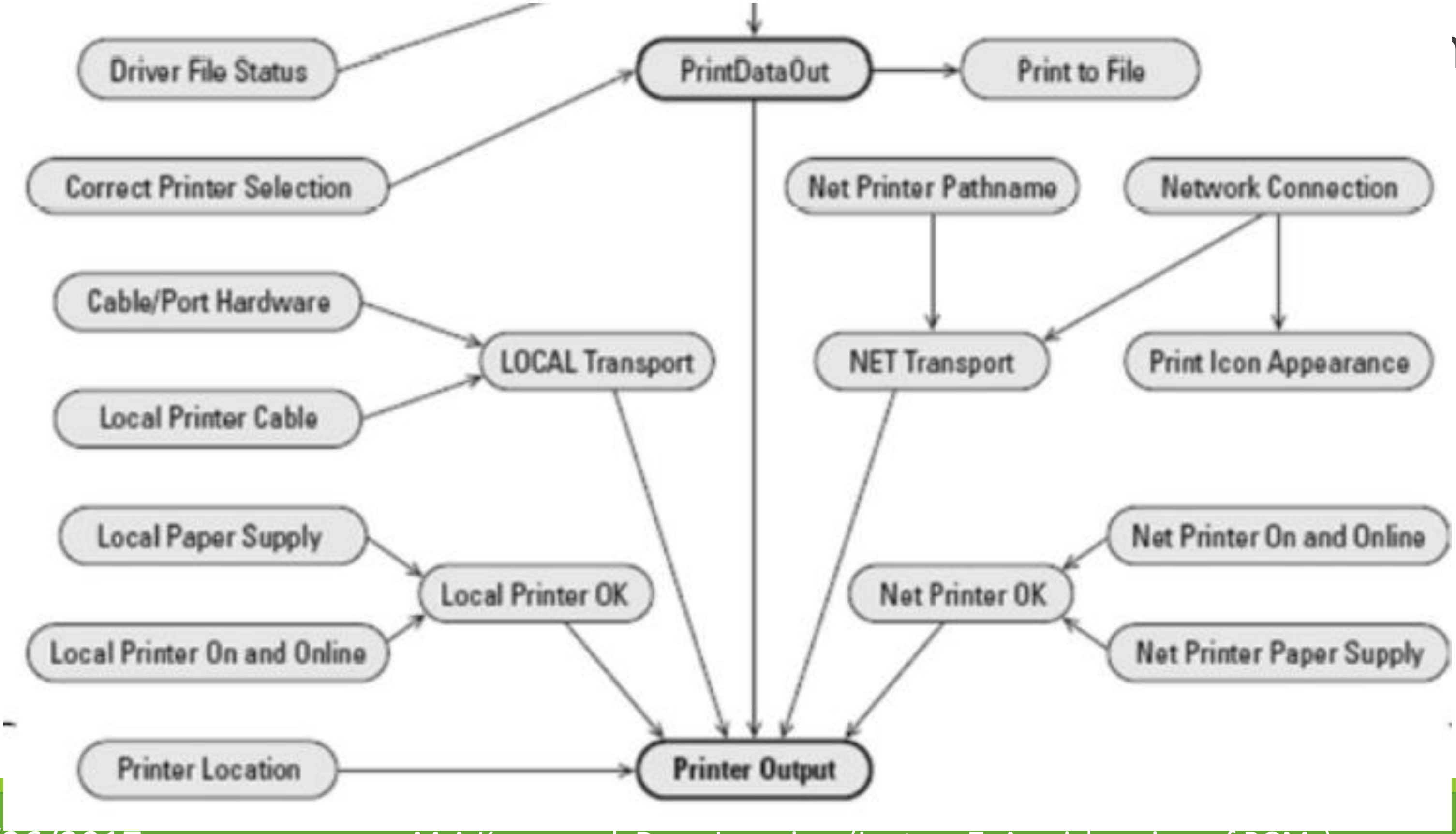
Lecture 7: A quick review of Probabilistic Graphical Models

OUTLINE

- Probabilistic Graphical Models
- Bayesian Networks
  - Generative Modeling
  - General Factorization Property
  - Student Example, CPDs
  - Inference
  - Reasoning Patterns
  - Conditional Independence
- Dynamic Bayesian Networks
- Markov Random Fields

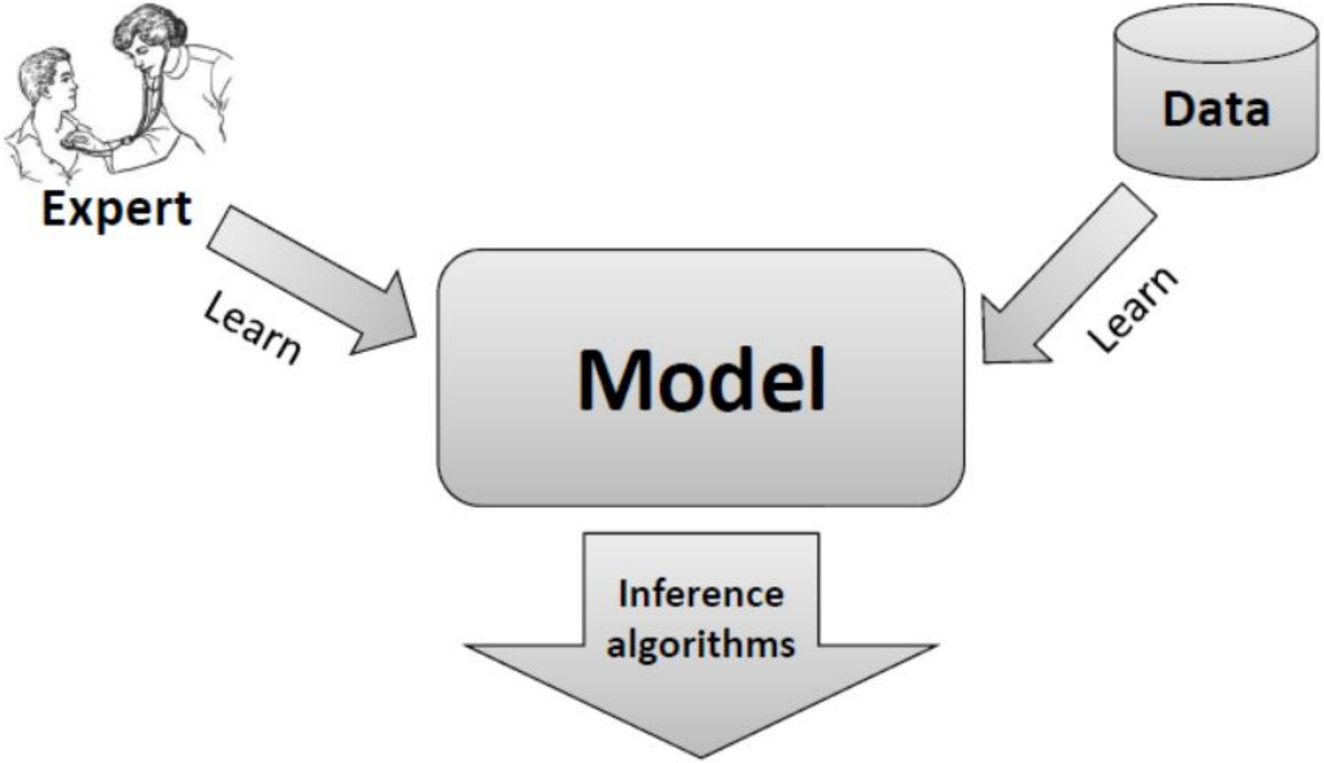
Probabilistic Graphical Models

- PGMs are declarative representation of our understanding of the world



ıs)

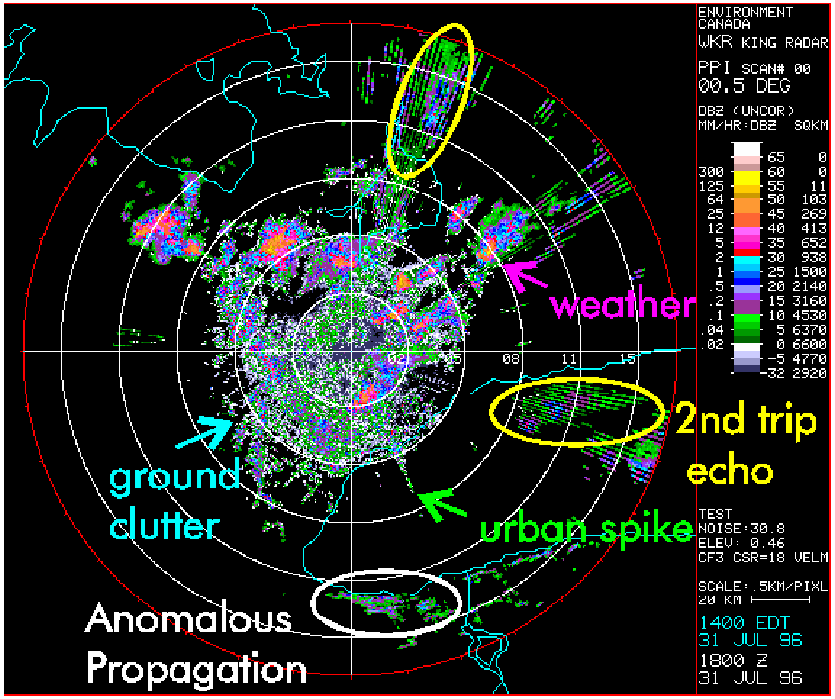
Probabilistic Graphical Models



Probabilistic Graphical Models, instructor: Dr. Ahmad Nickabadi

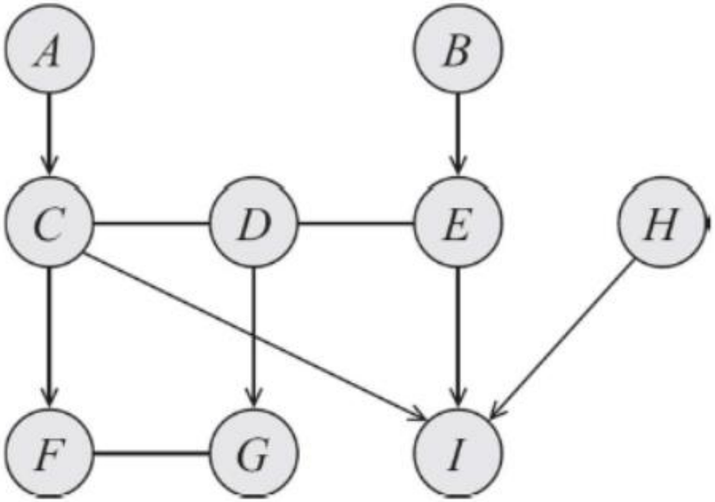
Probabilistic Graphical Models

- PGMs can handle uncertainty
  - Partial knowledge of state of the world
  - Partial and noisy observations
  - Phenomena not covered by our model
  - Inherent non-determinism of the world



Probabilistic Graphical Models

Node, Edge, Directed/Undirected edge, Parent-Child, Neighbor, Node degree, Indegree, Subgraph, Complete subgraph (clique)
Maximal clique, Path, trail,
Cycle, DAG, Loop, Tree,
Triangulated graph



Probabilistic Graphical Models, instructor: Dr. Ahmad Nickabadi

Probabilistic Graphical Models

- Representation
  - Directed
  - Undirected
- Inference
  - Exact
  - Approximate
- Learning
  - Parameters
  - Structure

- Applications:
  - Medical diagnosis
  - Fault diagnosis
  - Natural language processing
  - Traffic analysis
  - Computer vision
  - Speech recognition
  - Robot localization and mapping

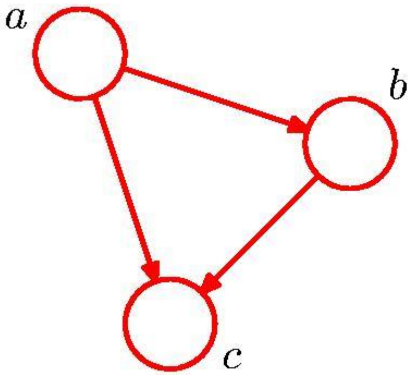
Probabilistic Graphical Models, instructor: Dr. Ahmad Nickabadi

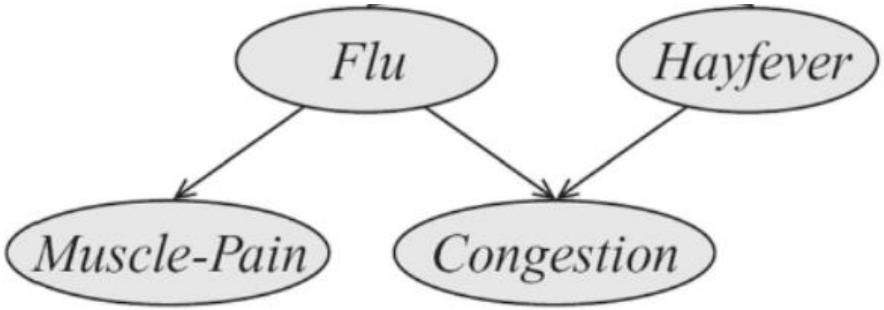
OUTLINE

- Probabilistic Graphical Models
- Bayesian Networks
  - Generative Modeling
  - General Factorization Property
  - Student Example, CPDs
  - Inference
  - Reasoning Patterns
  - Conditional Independence
- Dynamic Bayesian Networks
- Markov Random Fields

Bayesian Networks

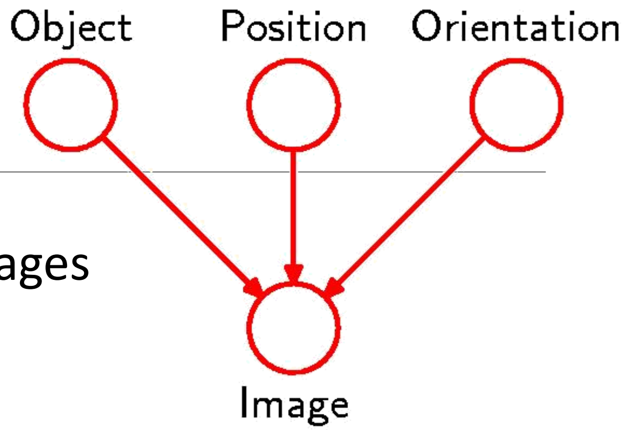
- Bayesian Network is a Directed Acyclic Graph, DAG.
- Provides a compact factorized representation of a joint distribution
- Nodes: random variables
- Edges: direct influences (causality)
- Generative Modeling



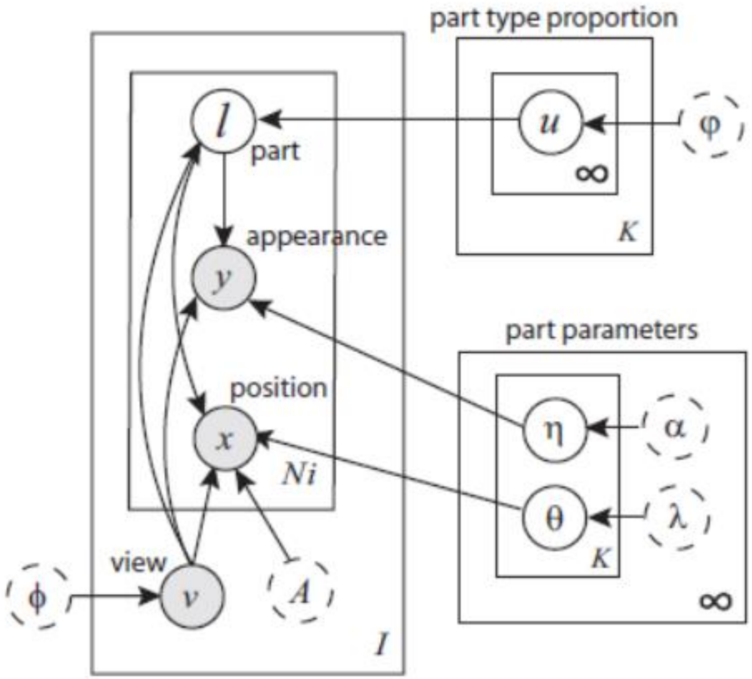


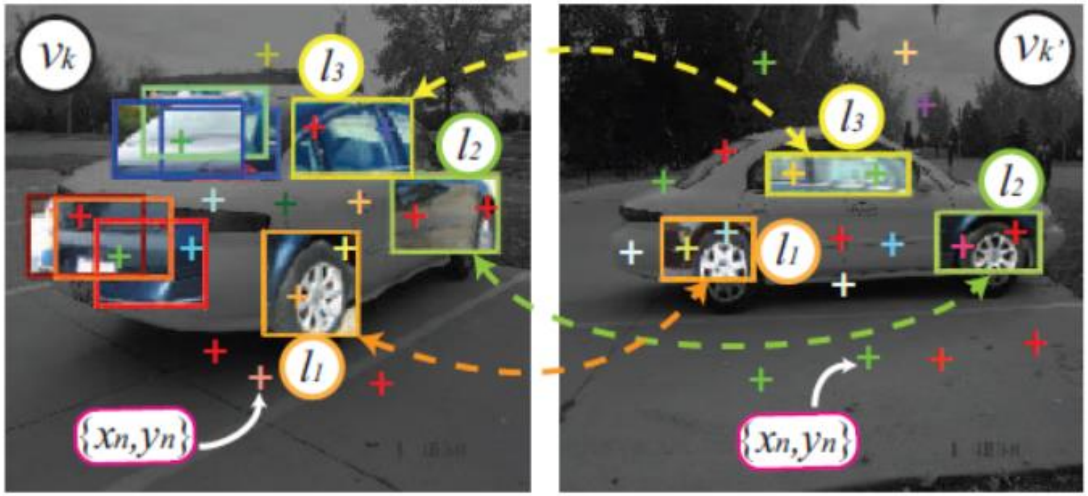
Bayesian Networks

An example: Causal Process for generating images



PRML, C. Bishop



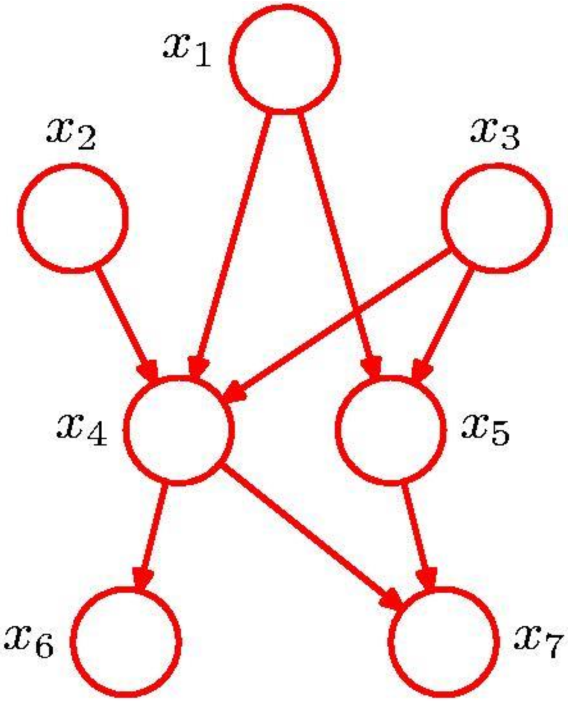


Sun, Min, Hao Su, Silvio Savarese, and Li Fei-Fei, A Multi-View Probabilistic Model for 3D Object Classes, 2009

BNs, General Factorization Property

PRML, C. Bishop

$$p(x_1, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)\\ p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$



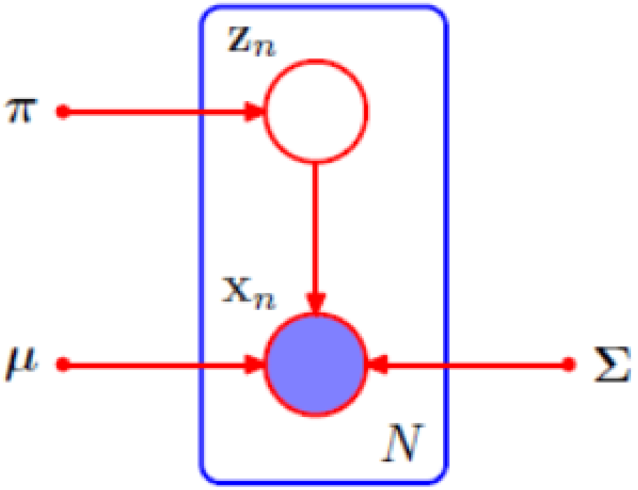
BNs, General Factorization

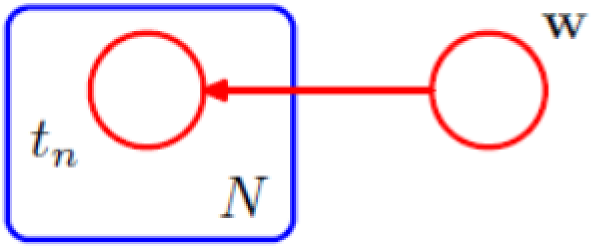
$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

Chain Rule: $p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \ldots p(x_2|x_1)p(x_1)$

Some Conventions

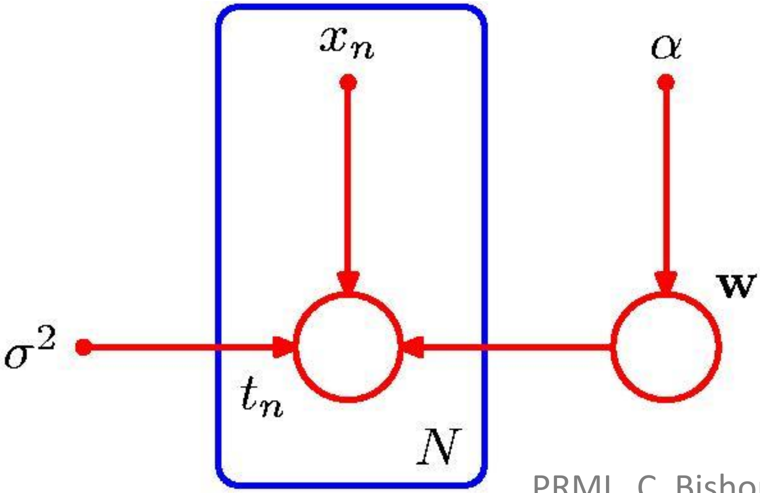
- Plate
- Observable Variables
- Parameters





$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w}).$$

$$p(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n|\mathbf{w}, x_n, \sigma^2).$$



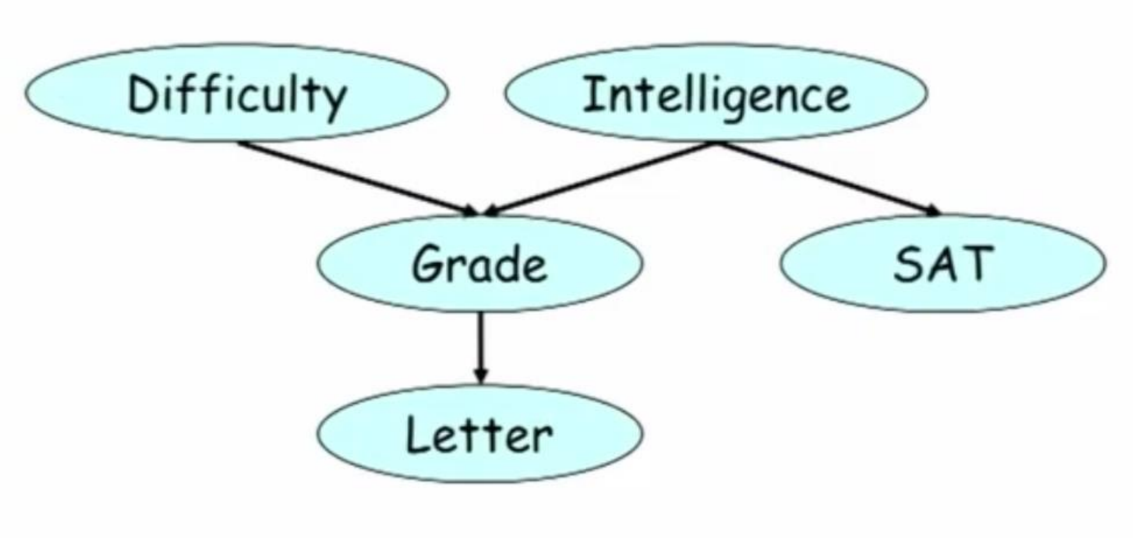
PRML, C. Bishop

OUTLINE

- Probabilistic Graphical Models
- Bayesian Networks
  - Generative Modeling
  - General Factorization Property
  - Student Example, CPDs
  - Inference
  - Reasoning Patterns
  - Conditional Independence
- Dynamic Bayesian Networks
- Markov Random Fields

Student Example

- A BN related to a student and an specific course
- **G**rade
- Course **D**ifficulty
- Student **I**ntelligence
- Student **S**AT
- Reference **L**etter



PGM, D. Koller

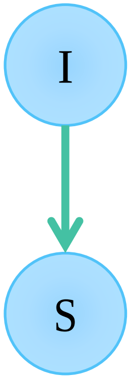
Conditional Probability Distribution | Table, CPD(T)

$$P(I,S) = P(I)P(S \mid I)$$

I	S	$P(I,S)$
i^0	s^0	0.665
i^0	s^1	0.035
i^1	s^0	0.06
i^1	s^1	0.24.

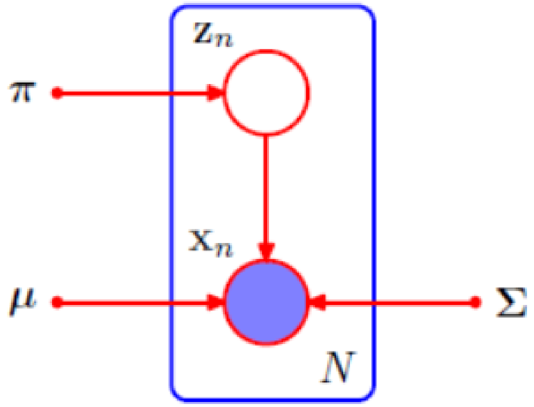
i^0	i^1
0.7	0.3

I	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

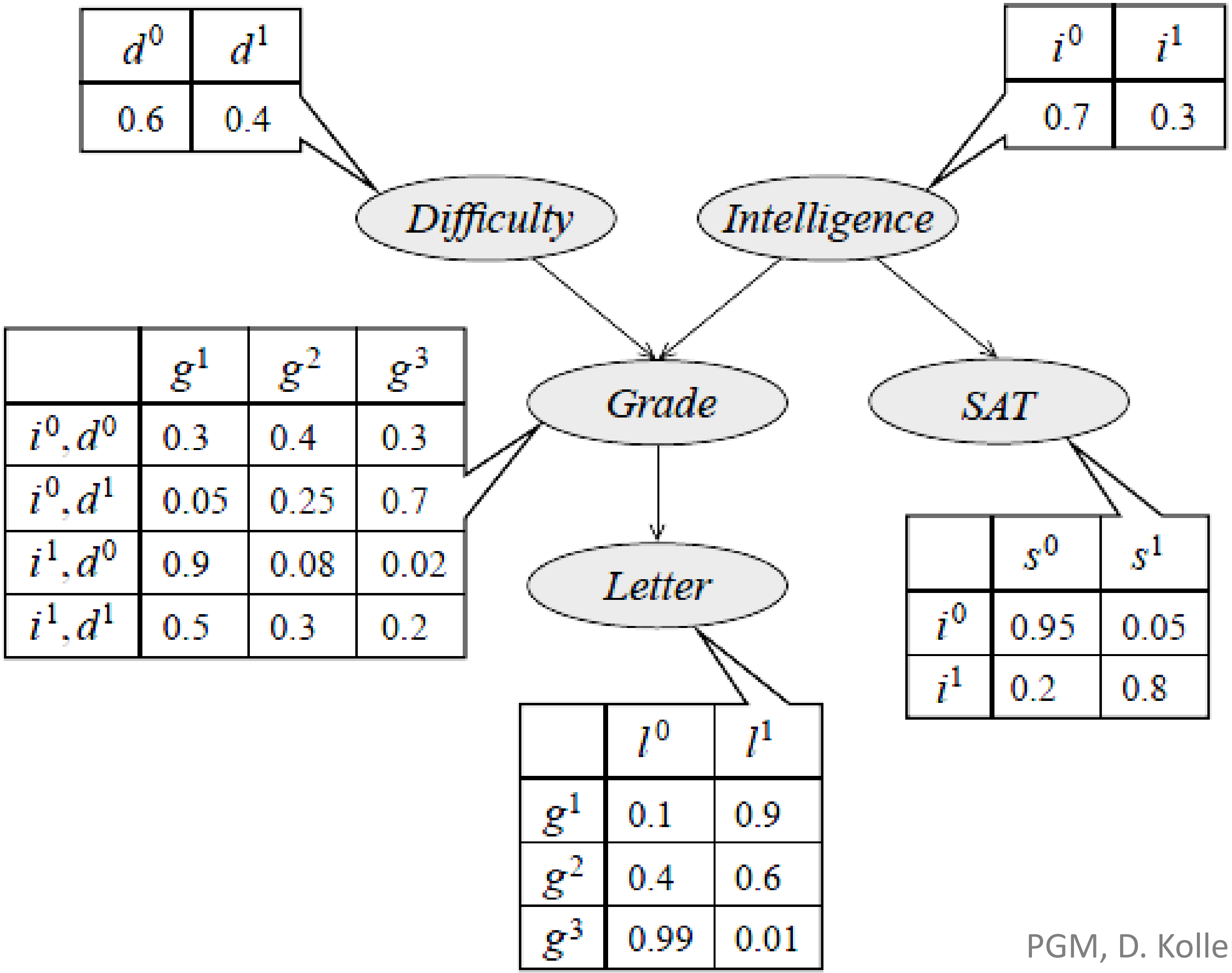


$$p(\boldsymbol{x}_n|z_n = k) = p(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$$

$$p(\boldsymbol{x}_n|z_n = k)p(z_n = k) = p(\boldsymbol{x}_n; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)\pi_k$$



Student Example



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

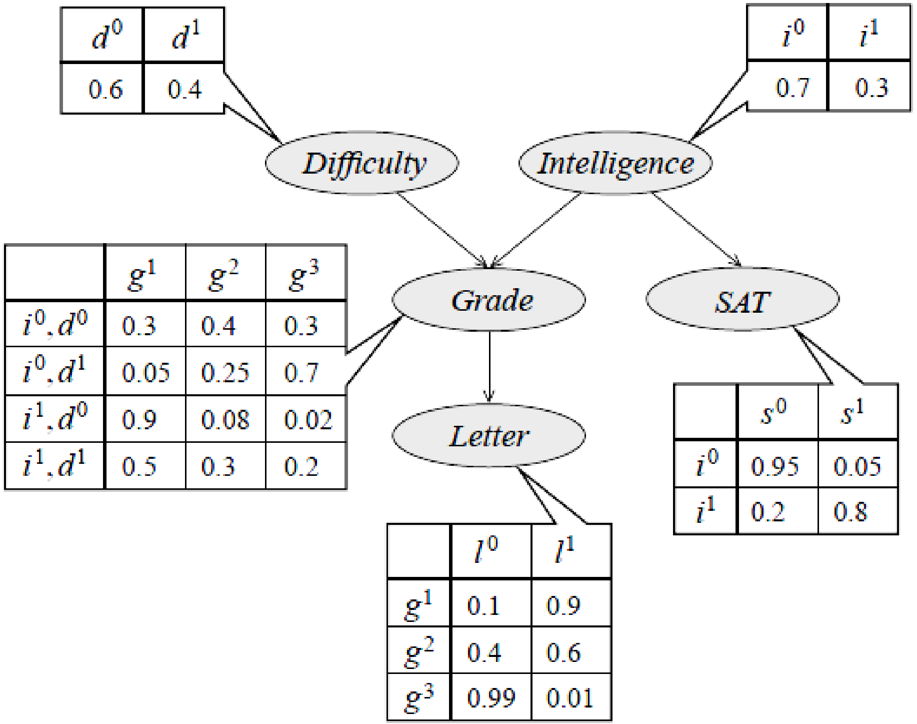
Let’s check all CPTs!

PGM, D. Koller

Student Example
Factorization simplifies joint representation

D	I	G	S	L	P_B
d^0	i^0	g^1	s^0	l^0	0.01197
d^0	i^0	g^1	s^0	l^1	0.10773
d^0	i^0	g^1	s^1	l^0	0.00063
d^0	i^0	g^1	s^1	l^1	0.00567
d^0	i^0	g^2	s^0	l^0	…
d^0	i^0	g^2	s^0	l^1	…
d^0	i^0	g^2	s^1	l^0	…
d^0	i^0	g^2	s^1	l^1	…
d^0	i^0	g^3	s^0	l^0	…
d^0	i^0	g^3	s^0	l^1	…
d^0	i^0	g^3	s^1	l^0	…
d^0	i^0	g^3	s^1	l^1	…
…	…	…	…	…	…

Lets' check some of this table rows



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

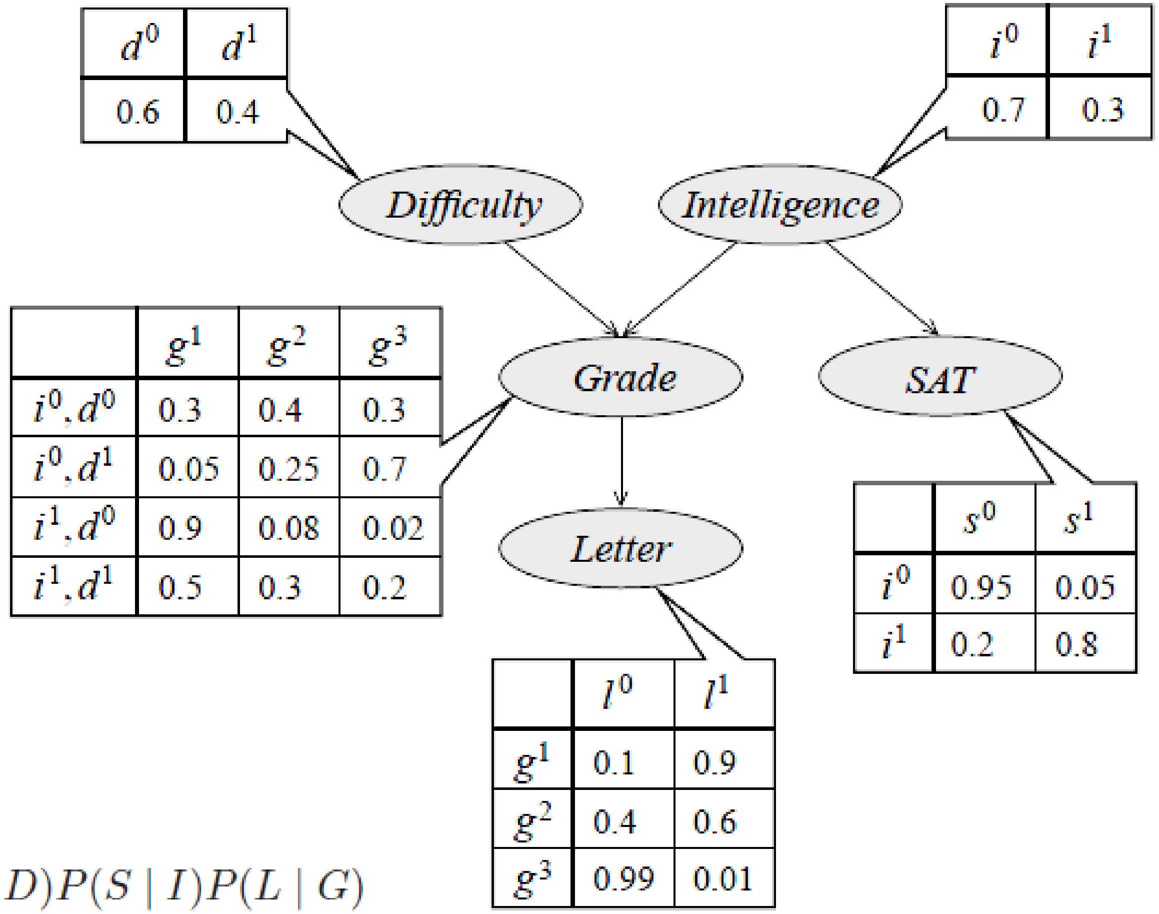
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

OUTLINE

- Probabilistic Graphical Models
- Bayesian Networks
  - Generative Modeling
  - General Factorization Property
  - Student Example, CPDs
  - Inference
  - Reasoning Patterns
  - Conditional Independence
- Dynamic Bayesian Networks
- Markov Random Fields

Student Example, Inference

- $P_B(Y = y|E = e)$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

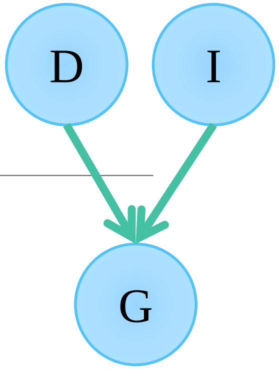
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

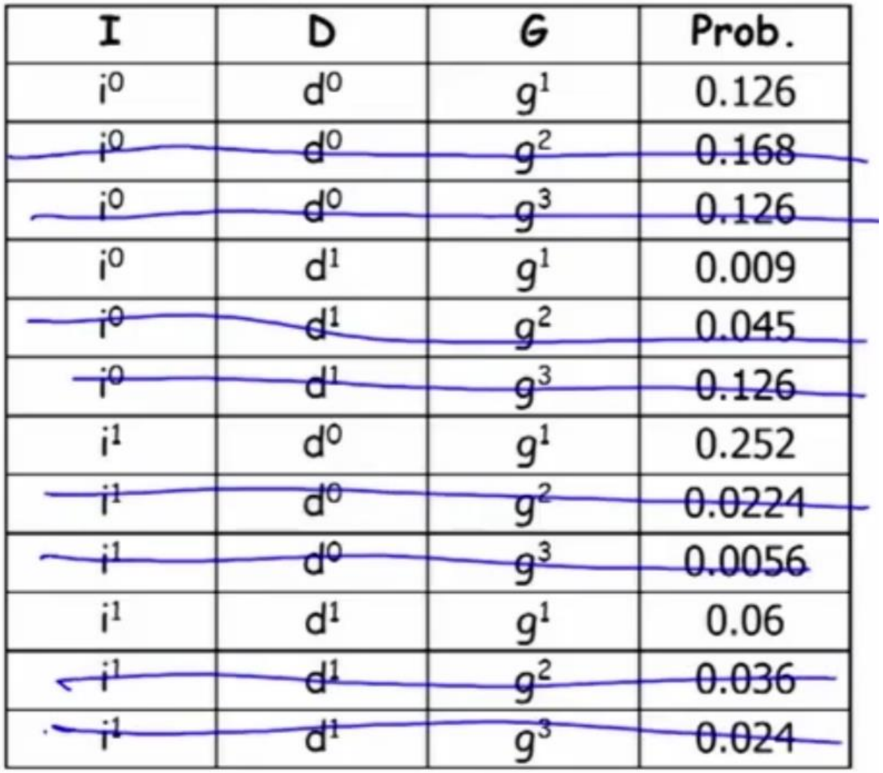
$$P(I, D, G, S, L) = P(I)P(D)P(G \mid I, D)P(S \mid I)P(L \mid G)$$

PGM, D. Koller

Student Example, Inference

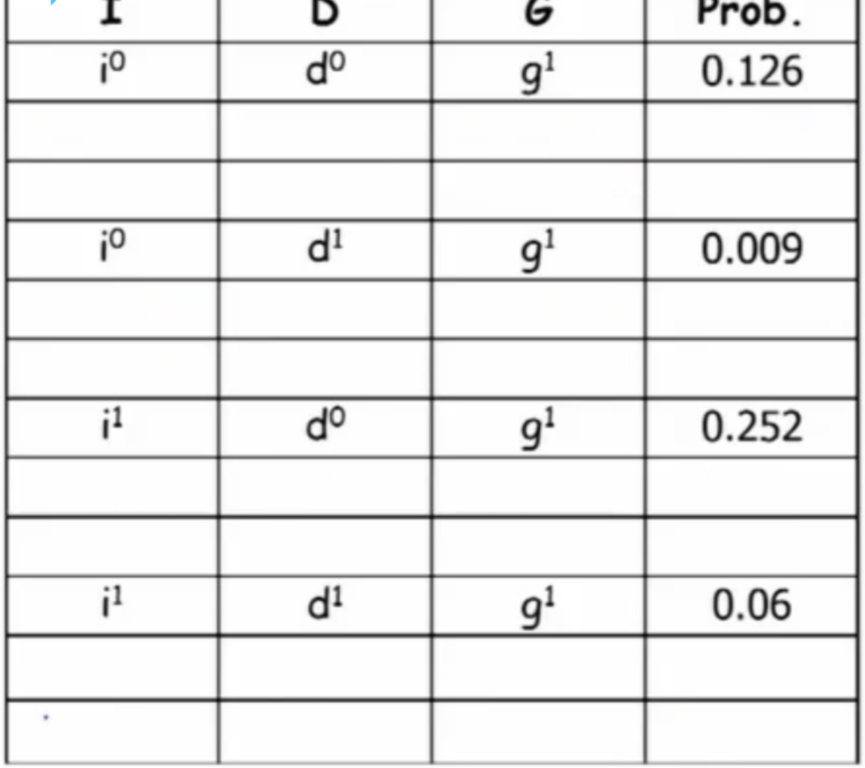


- Conditioning on g^1, $P_B(I, D | g^1)$



I	D	G	Prob.
i^0	d^0	g^1	0.126
~~i^0~~	~~d^0~~	~~g^2~~	~~0.168~~
~~i^0~~	~~d^0~~	~~g^3~~	~~0.126~~
i^0	d^1	g^1	0.009
~~i^0~~	~~d^1~~	~~g^2~~	~~0.045~~
~~i^0~~	~~d^1~~	~~g^3~~	~~0.126~~
i^1	d^0	g^1	0.252
~~i^1~~	~~d^0~~	~~g^2~~	~~0.0224~~
~~i^1~~	~~d^0~~	~~g^3~~	~~0.0056~~
i^1	d^1	g^1	0.06
~~i^1~~	~~d^1~~	~~g^2~~	~~0.036~~
~~i^1~~	~~d^1~~	~~g^3~~	~~0.024~~

Reduction

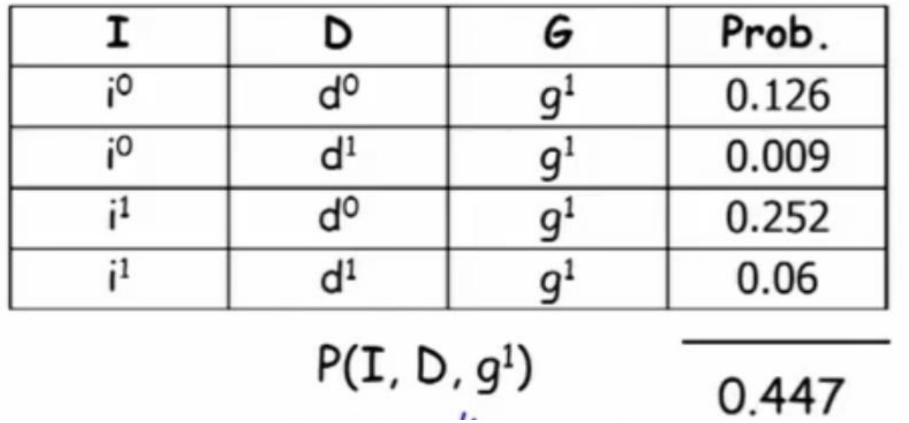


I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.06

Student Example, Inference

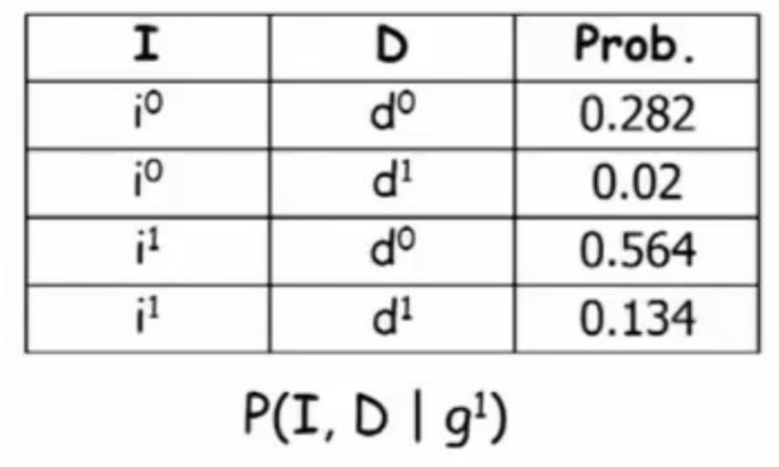
- Conditioning on g^1, $P_B(I, D|g^1)$

Re-normalization



I	D	G	Prob.
i^0	d^0	g^1	0.126
i^0	d^1	g^1	0.009
i^1	d^0	g^1	0.252
i^1	d^1	g^1	0.06

$P(I, D, g^1)$ 0.447



I	D	Prob.
i^0	d^0	0.282
i^0	d^1	0.02
i^1	d^0	0.564
i^1	d^1	0.134

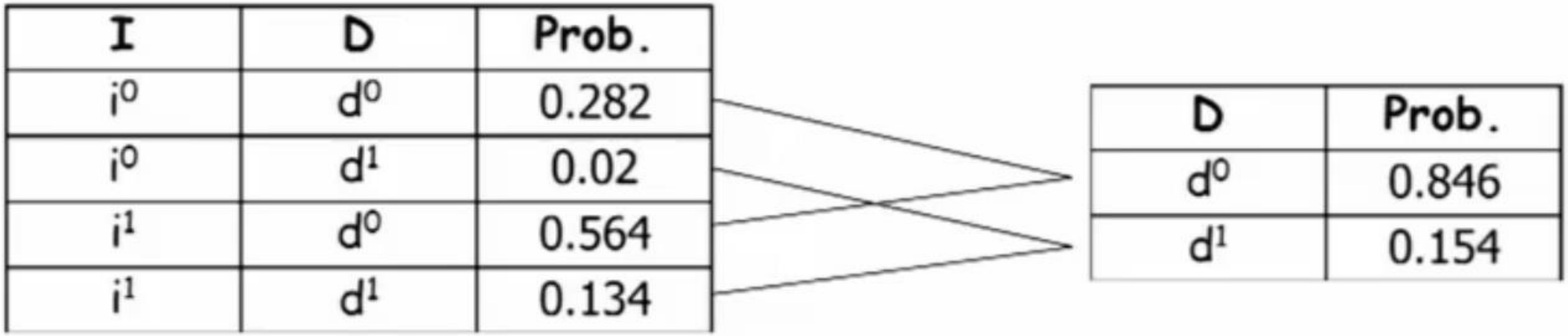
$P(I, D \mid g^1)$

$$P_B(Y = y|E = e) = \frac{P_B(y, e)}{P_B(e) = \sum_y P_B(y, e)}$$

PGM, D. Koller

Student Example, Inference

Marginalization



I	D	Prob.
i^0	d^0	0.282
i^0	d^1	0.02
i^1	d^0	0.564
i^1	d^1	0.134

$P(I, D \mid g^1)$

D	Prob.
d^0	0.846
d^1	0.154

$P(D|g^1)$

$$P(D|g^1) = \sum_I P(I, D|g^1)$$

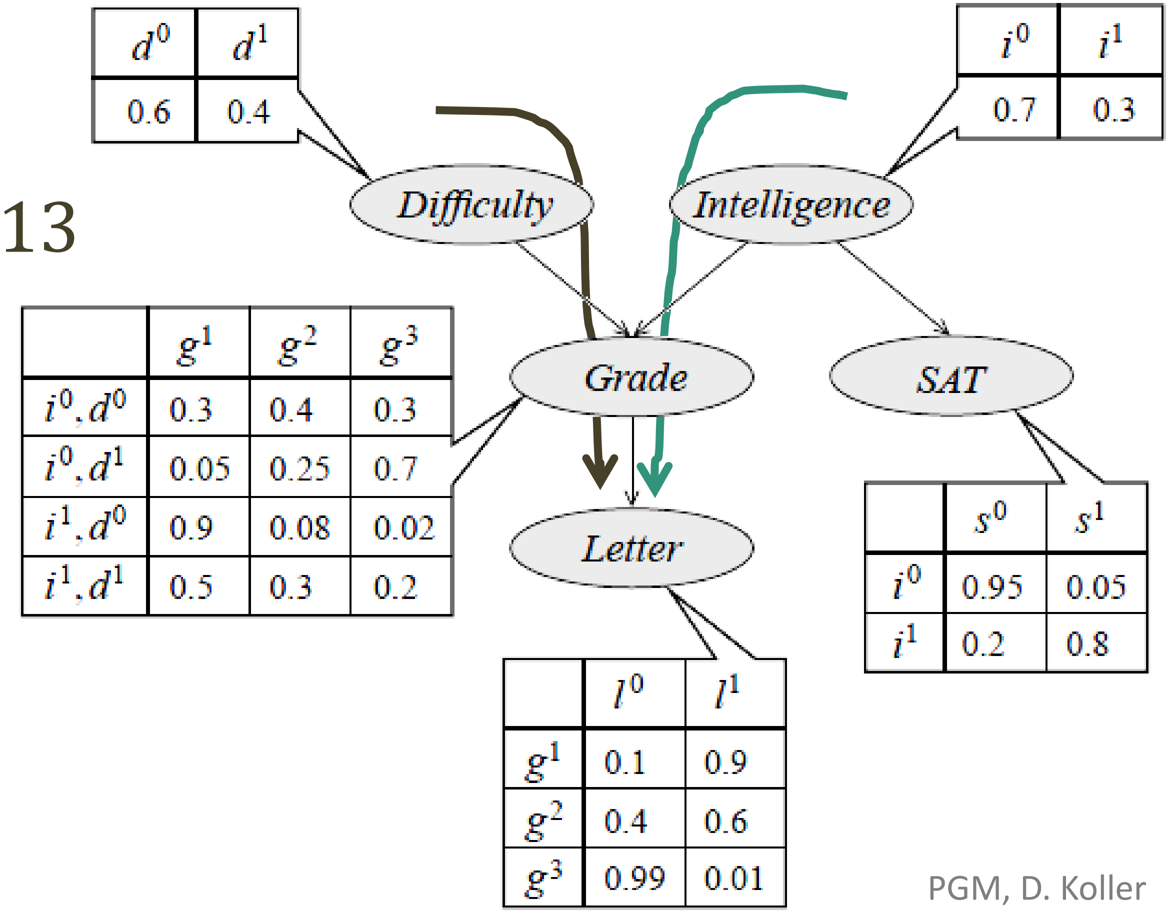
PGM, D. Koller

OUTLINE

- Probabilistic Graphical Models
- Bayesian Networks
  - Generative Modeling
  - General Factorization Property
  - Student Example, CPDs
  - Inference
  - Reasoning Patterns
  - Conditional Independence
- Dynamic Bayesian Networks
- Markov Random Fields

Student Example:
Causal reasoning, Prediction

- $P_B(l^1) = 0.502$
- $P_B(l^1|i^0) = 0.389$
- $P_B(l^1|i^0, d^0) = 0.513$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0,d^0	0.3	0.4	0.3
i^0,d^1	0.05	0.25	0.7
i^1,d^0	0.9	0.08	0.02
i^1,d^1	0.5	0.3	0.2

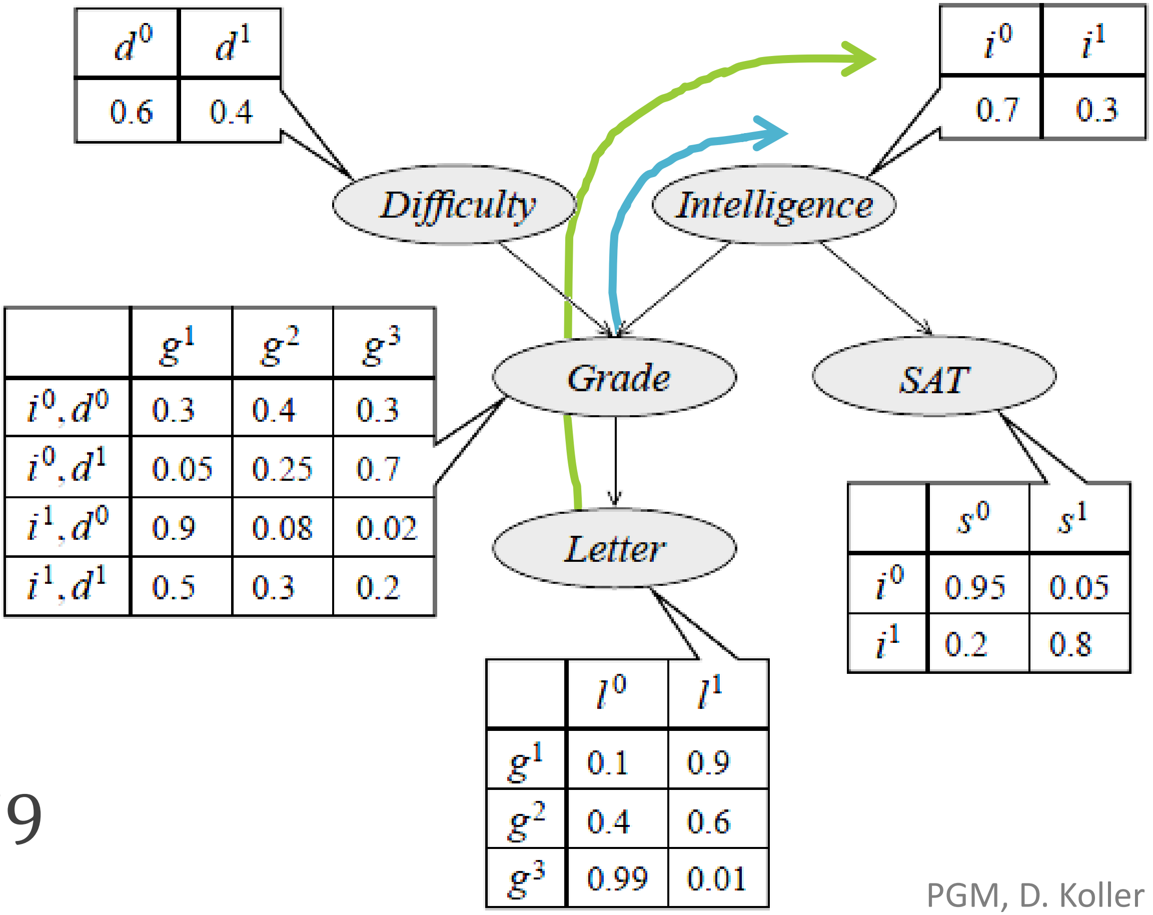
	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

PGM, D. Koller

Student Example: Evidential reasoning, Explanation

- $P_B(i^1) = 0.3$
- $P_B(i^1|g^3) = 0.079$
- $P_B(i^1|l^0) = 0.14$
- $P_B(d^1) = 0.4$
- $P_B(d^1|g^3) = 0.629$
- $P_B(i^1|g^3, l^0) = 0.079$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

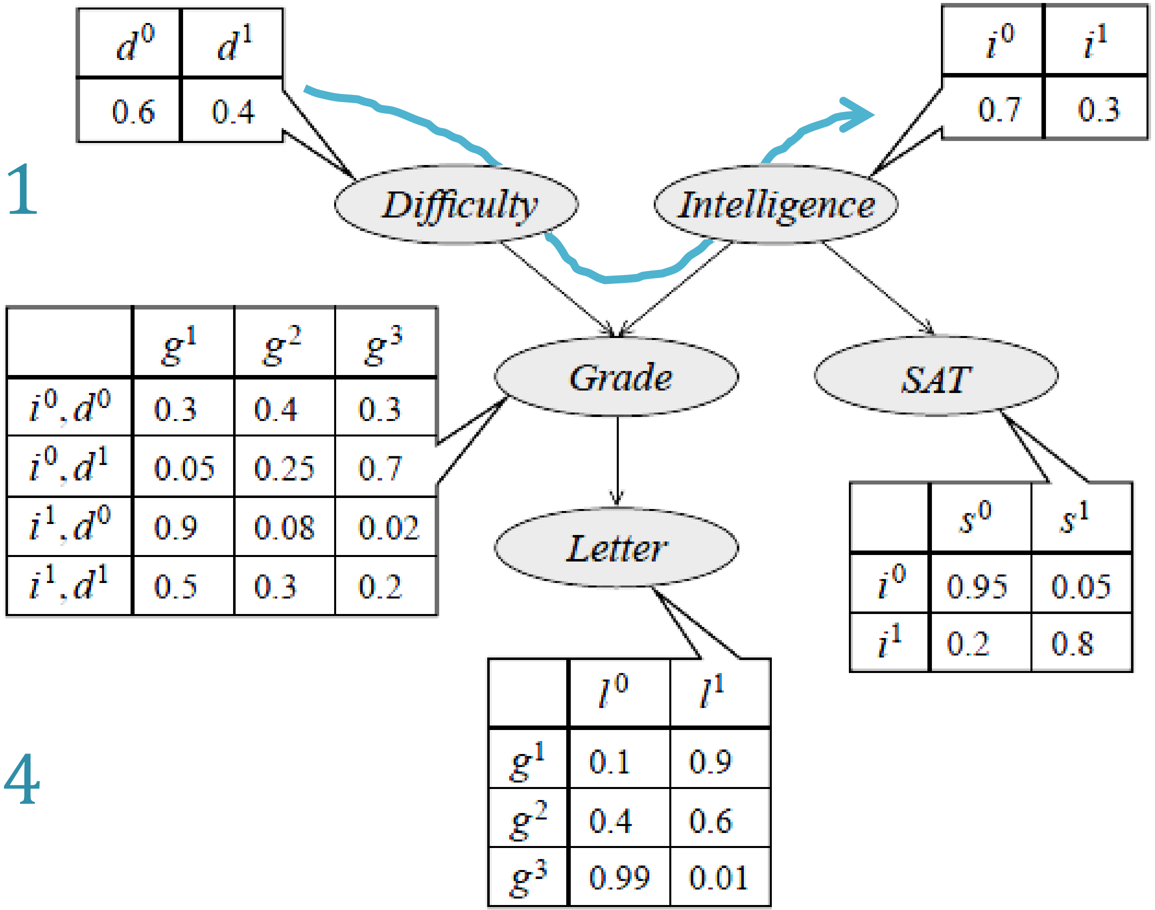
	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

PGM, D. Koller

Student Example: Inter-causal reasoning

PGM, D. Koller

- $P_B(i^1|g^3\ ) = 0.079$
- $P_B(i^1|g^3, d^1) = 0.11$



d^0	d^1
0.6	0.4

i^0	i^1
0.7	0.3

	g^1	g^2	g^3
i^0, d^0	0.3	0.4	0.3
i^0, d^1	0.05	0.25	0.7
i^1, d^0	0.9	0.08	0.02
i^1, d^1	0.5	0.3	0.2

	s^0	s^1
i^0	0.95	0.05
i^1	0.2	0.8

	l^0	l^1
g^1	0.1	0.9
g^2	0.4	0.6
g^3	0.99	0.01

- $P_B(i^1|g^2\ ) = 0.175$
- $P_B(i^1|g^2, d^1) = 0.34$

We have explained away the poor grade via the difficulty of class

OUTLINE

- Probabilistic Graphical Models
- Bayesian Networks
  - Generative Modeling
  - General Factorization Property
  - Student Example, CPDs
  - Inference
  - Reasoning Patterns
  - Conditional Independence
- Dynamic Bayesian Networks
- Markov Random Fields

Conditional Independence

- a is independent of b given c

$$p(a|b,c) = p(a|c)$$

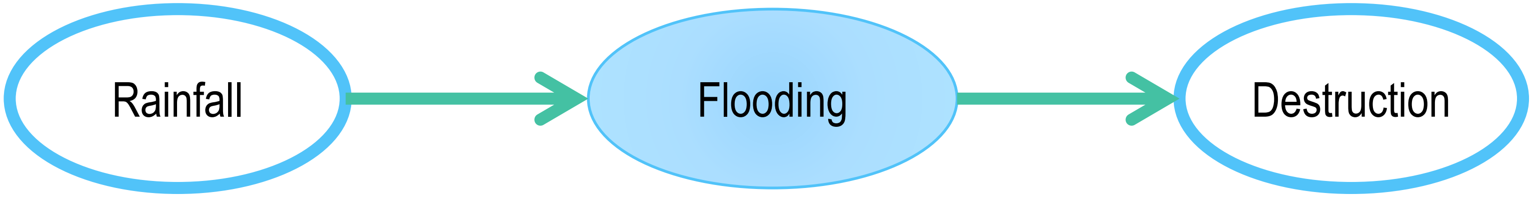
- Equivalently

$$\begin{aligned} p(a,b|c) &= p(a|b,c)p(b|c) \\ &= p(a|c)p(b|c) \end{aligned}$$

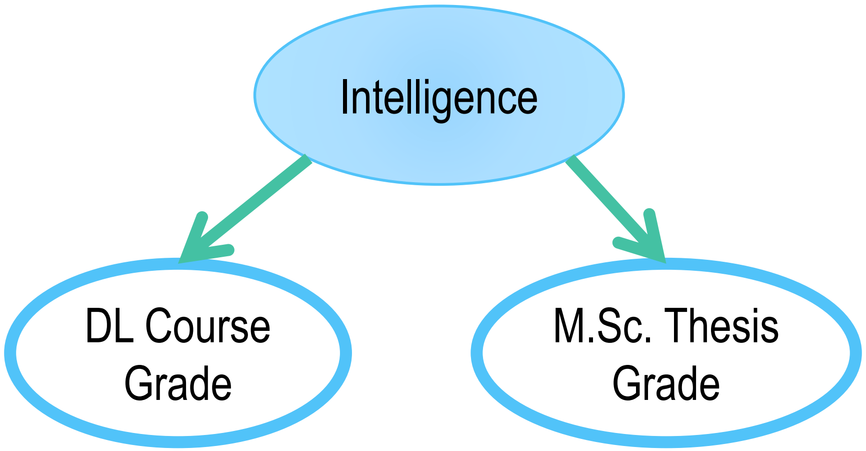
- Notation

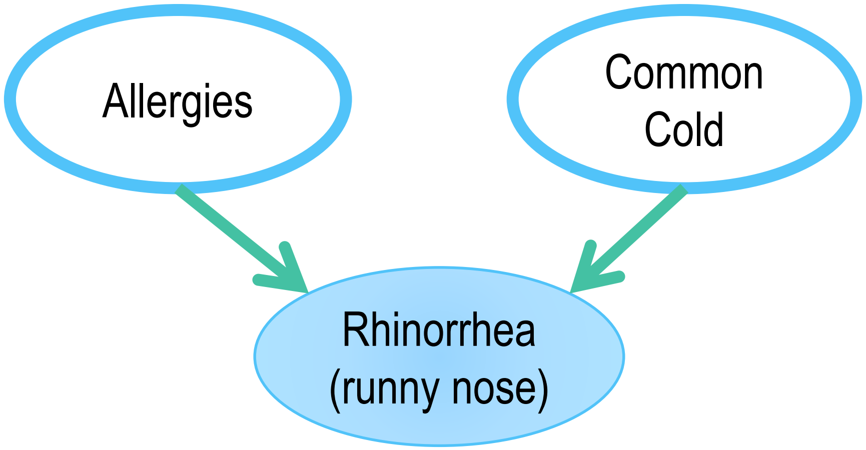
$$a \perp\!\!\!\perp b \mid c$$

Causal Trail, Evidential Trail
Common Cause, Common Effect



Causal Trail & Evidential Trail are active if and only if “Flooding” is not observed



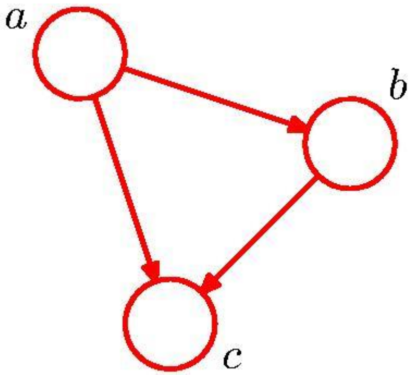


Common cause trail is active if and only if “Intelligence” is not observed

Common Effect trail is active if and only if either “Rhinorrhea” or one of its parents is observed

Bayesian Networks

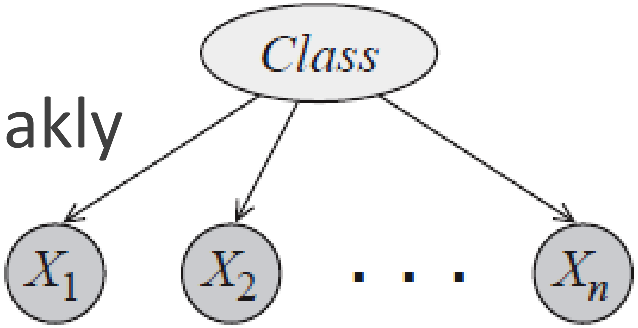
- BN is a DAG.
- Generative Modeling
- General Factorization Property
- BN is a legal distribution $P \geq 0$
  - P is product of CPDs
- BN is a legal distribution $\sum P = 1$
  - Each CPD is legal in this sense
- BN captures independent assumptions about variables
  - BN simplifies the joint using these assumptions



Bayesian Networks:
An example: Naïve Bayes Model

- A simple model for classification
- Class variable is a discrete variable
- X_is are feature variables
- Reasoning pattern: evidential reasoning
- $P(C, X_1, \ldots, X_n) = P(C) \prod_{i=1}^{n} P(X_i \mid C)$
- $X_i \perp X_j \mid C, \quad \forall\, i \neq j$
- Effective in domains with weakly relevant features

PGM, D. Koller



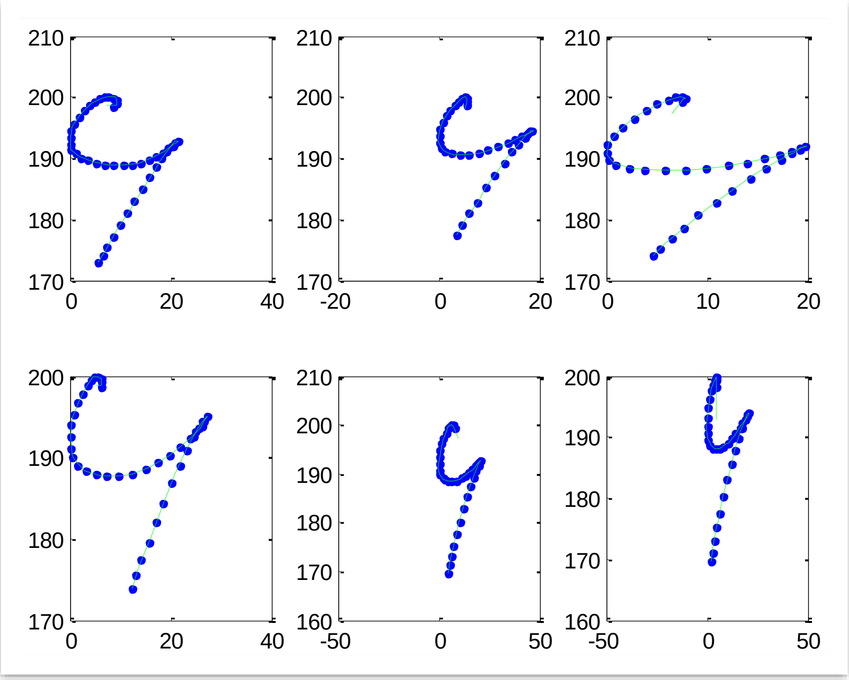
OUTLINE

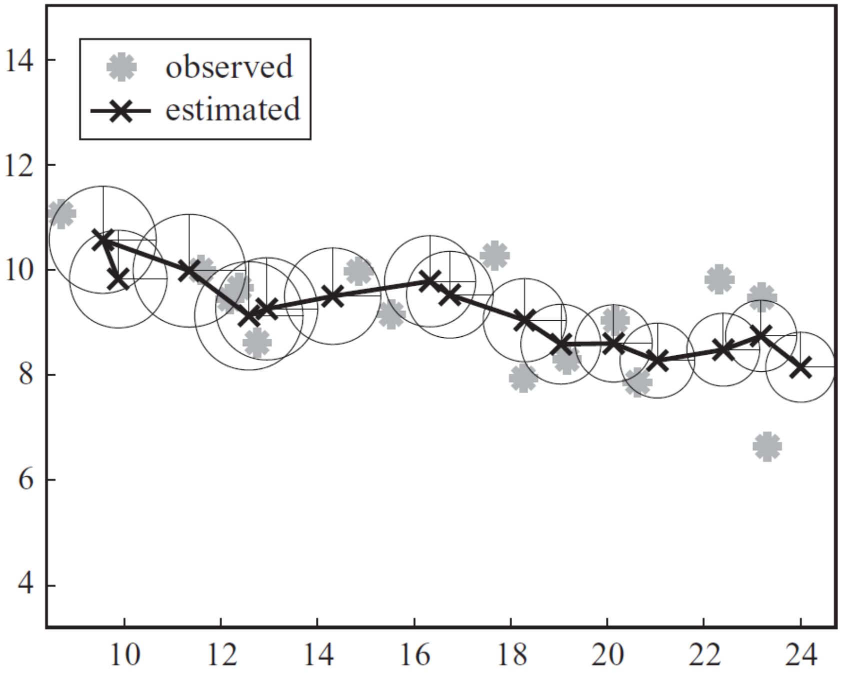
- Probabilistic Graphical Models
- Bayesian Networks
- Dynamic Bayesian Networks
  - Time-series, Stochastic Processes
  - 2-TBN, DBN
  - State-space models, HMMs, KFMs
  - Inference Patterns
- Markov Random Fields

Distribution over trajectories

- Time or space stochastic process

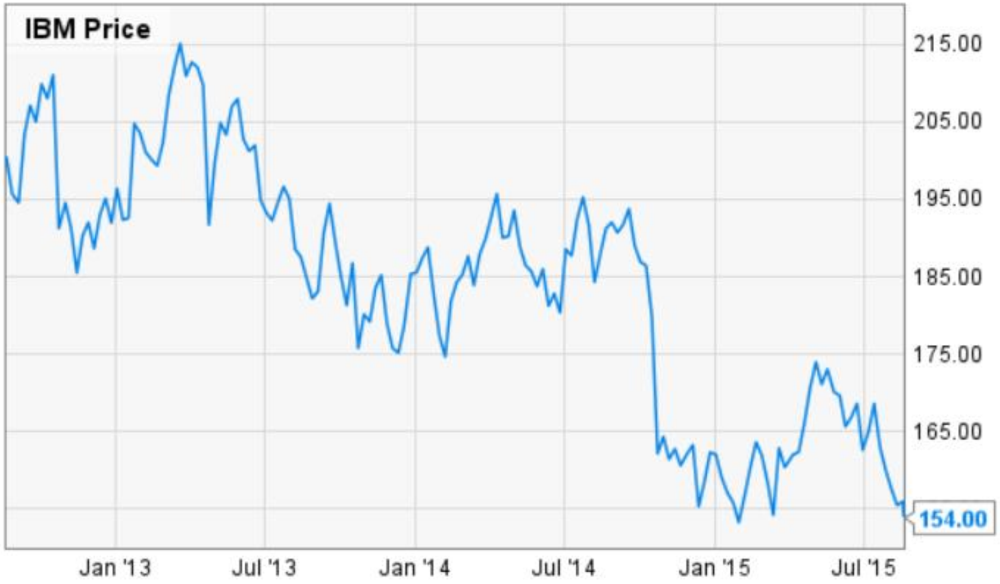
K. P. Murphy





Simplifying Assumptions

- Select a time granularity, Δ
- X_t variable at time t
- $X_{1:t}$ variables from time 1 to t
- Objective: Model $X_{1:T}$



Simplifying Assumptions

- Objective: Model $X_{1:T}$
- Chain rule:
  - $P(X_{1:T}) = P(X_1)\prod_{t=2}^{T} P(X_t|X_{1:t-1})$
- Markov Assumption:
  - $X_{t+1} \perp X_{1:t-1} \mid X_t$
  - $P(X_{1:T}) = P(X_1)\prod_{t=2}^{T} P(X_t|X_{t-1})$
- Time Invariance:
  - $P(X_t|X_{t-1}) = P(X'|X)$

PGM, D. Koller

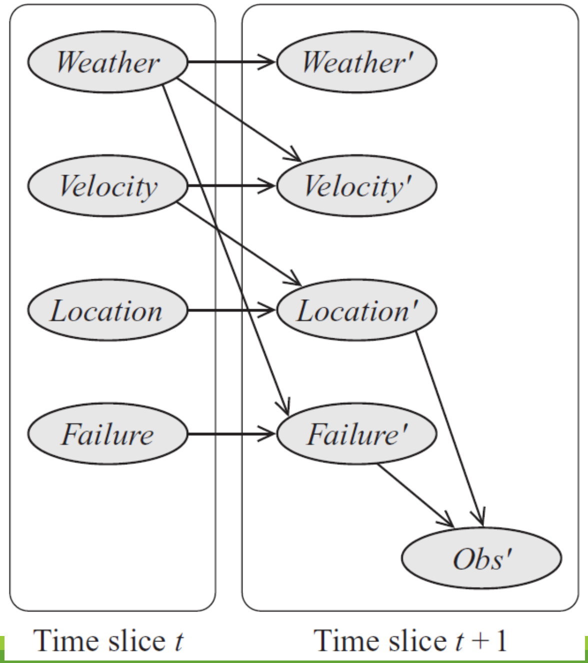
OUTLINE

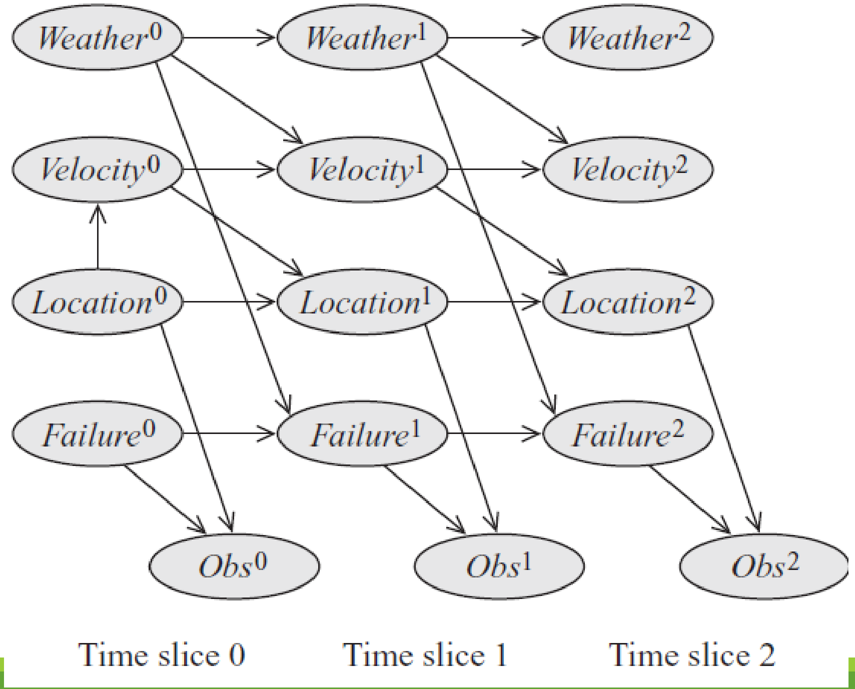
- Probabilistic Graphical Models
- Bayesian Networks
- Dynamic Bayesian Networks
  - Time-series, Stochastic Processes
  - 2-TBN, DBN
  - State-space models, HMMs, KFMs
  - Inference Patterns
- Markov Random Fields

2 Time-Slice Bayesian Network

- $\mathcal{B}_{\rightarrow}$ is a **2-TBN** for the **process**
  - It defines $P(X_t|X_{t-1})$
  - Using a DAG as $P(X_t|X_{t-1}) = \prod_{i=1}^{N} P\left(X_t^i|Pa(X_t^i)\right)$

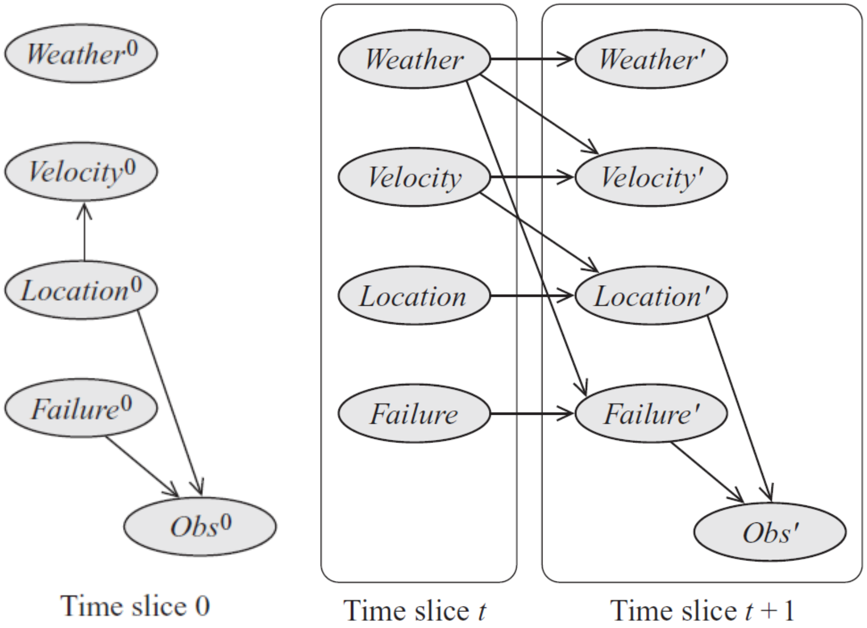
PGM, D. Koller





Dynamic Bayesian Network

- A DBN is a pair $\langle\mathcal{B}_1, \mathcal{B}_\rightarrow\rangle$
- $\mathcal{B}_1$ is a Bayesian network over X_1
  - defines prior $P(X_1)$ or **initial distribution** over states
- $\mathcal{B}_\rightarrow$ is a **2-TBN for the process**



PGM, D. Koller

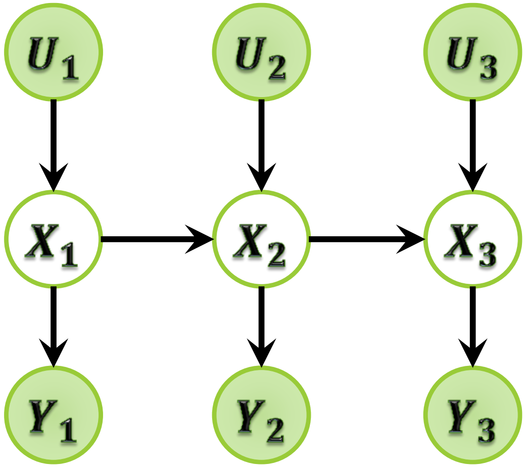
OUTLINE

- Probabilistic Graphical Models
- Bayesian Networks
- Dynamic Bayesian Networks
  - Time-series, Stochastic Processes
  - 2-TBN, DBN
  - State-space models, HMMs, KFMs
  - Inference Patterns
- Markov Random Fields

State-space models

K. P. Murphy

- we assume that there is some underlying hidden state of the world
  - in the controlled case, the hidden state is a function of our inputs
  - the hidden state evolves in time
  - the hidden state generates observations
- In other word: A state-space model is a model of how X_t generates or “causes” Y_t and X_{t+1}
- Mainly: the goal of inference is to invert this mapping
  - i.e.: to infer $X_{1:t}$ given $Y_{1:t}$



State-space models

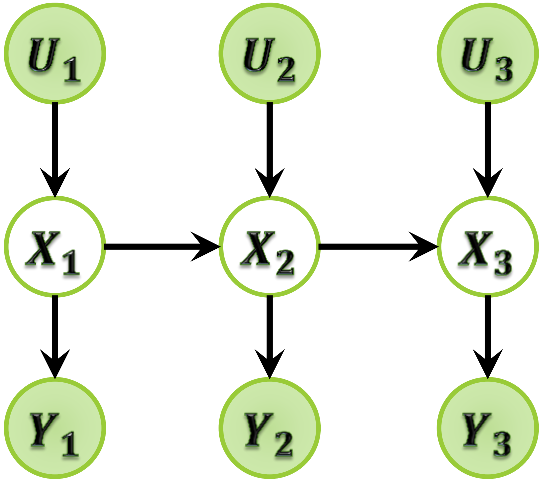
K. P. Murphy

- Any state-space model must define
  - a prior over states, $P(X_1)$
  - state-transition function, $P(X_t|X_{t-1})$
  - observation function, $P(Y_t|X_t)$

system model

Observation model

- In the controlled case, these become
  - $P(X_t|X_{t-1}, U_t)$
  - $P(Y_t|X_t, U_t)$ or $P(Y_t|X_t\ )$



State-space models HMMs, KFMs

K. P. Murphy

- the most common ways of representing state-space models are
  - Hidden Markov Models (HMMs)
    - HMMs assume X_t is a discrete random variable, $X_t \in \{1, \ldots, K\}$
    - There is no other restrictions on the transition or observation function
  - Kalman Filter Models (KFMs)
    - KFMs assume X_t is a vector of continuous random variables

$$X_t \in \mathbb{R}^N$$

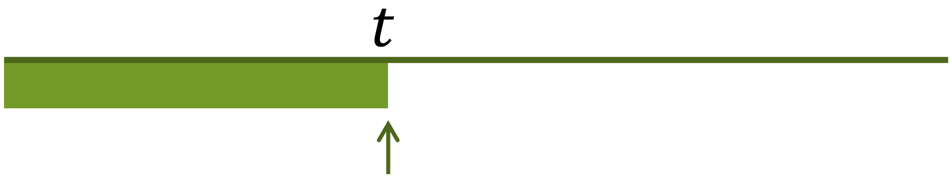
    - $X_{1:T}$ and $Y_{1:T}$ are jointly Gaussian

OUTLINE

- Probabilistic Graphical Models
- Bayesian Networks
- Dynamic Bayesian Networks
  - Time-series, Stochastic Processes
  - 2-TBN, DBN
  - State-space models, HMMs, KFMs
  - Inference Patterns
- Markov Random Fields

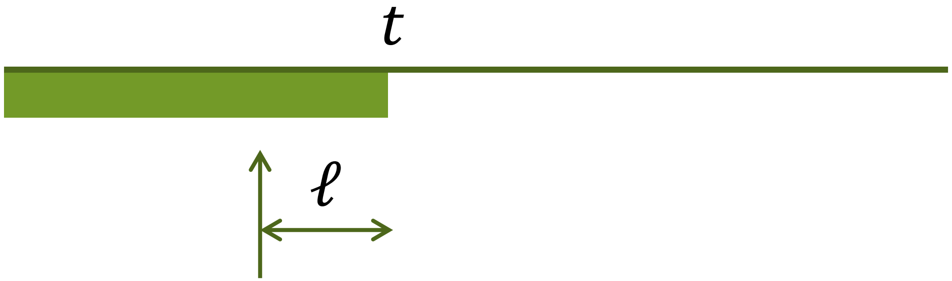
Inference Patterns: Filtering

- Filtering is common inference problem in **online analysis**
- recursively estimate the **belief state** $P(X_t|y_{1:t})$ using Bayes' rule
- $\hat{X}_{t|t-1} = P(X_t|y_{1:t-1})$
  - $\hat{X}_{t|t-1}$ is called prior belief state at time t
- $\hat{X}_{t|t} = P(X_t|y_{1:t-1}, y_t)$= $P(X_t|y_{1:t})$
- This task is traditionally called "filtering"
  - because we are filtering out the noise from the observations



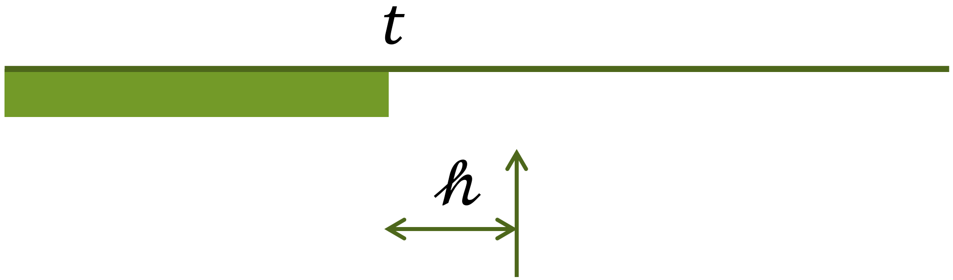
Inference Patterns: Smoothing

- sometimes we want to estimate the state of the past, given all the evidence up to the current time
- $P(X_{t-l}|y_{1:t}), \ell > 0$, ℓ is called lag
  - This is traditionally called "**fixed-lag smoothing**"
- (fixed interval) Smoothing:
  - in the offline case, we can compute:
  - $P(X_t|y_{1:T});\ \forall\ 1 \leq t \leq T$



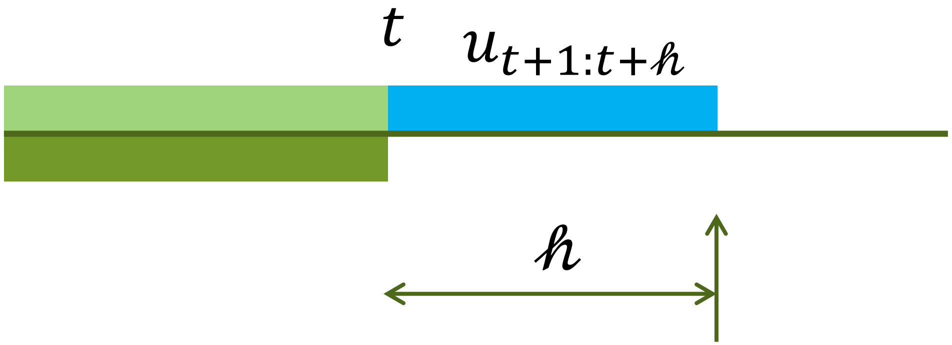
Inference Patterns: Prediction

- we might want to predict the future
- $P(Y_{t+h} = y|y_{1:t}), \mathscr{h} > 0$
  - $\mathscr{h}$ is how far we want to look-ahead
- once we have predicted the future hidden state
  - we can easily convert this into a prediction about the future observations
  - by marginalizing out $X_{t+\mathscr{h}}$



Inference Patterns: Control

- We might want to achieve to some desired output in the future
- Y_{t+h} is the desired output value
- Find the best control parameters over u_t



Inference Patterns: Decoding

- The goal is to compute the **most likely sequence of hidden states given the data**
  - computing the “most probable explanation”
- $x^*_{1:T} = arg \max_{x_{1:T}} P(x_{1:T} | y_{1:T})$

Inference Patterns: Classification

- likelihood of a model, M, is $P(y_{1:t}|M)$:
- we can classify a sequence as follows:
- $C^*(y_{1:T}) = arg\max_{C} P(y_{1:T}|M_C)P(M_C)$
  - $P(y_{1:T}|M_C)$ is the likelihood according to the model for class C
  - $P(M_C)$ is the prior for class C
- This method has the advantage of being able to handle sequences of variable-length

OUTLINE

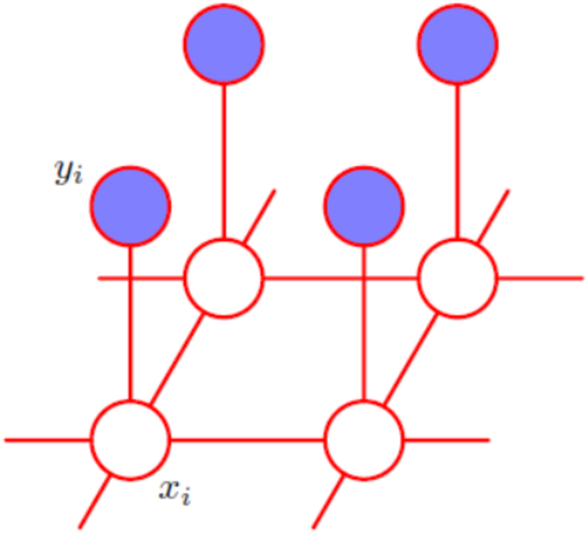
- Probabilistic Graphical Models
- Bayesian Networks
- Dynamic Bayesian Networks
- Markov Random Fields
  - Factorization property, cliques
  - The misconception example
  - Energy functions, Log-linear models
  - Image de-noising example
  - RBMs

Markov Networks or Markov Random Fields

- undirected graphs
- The joint distribution of an MRF is defined by:

$$P(\mathcal{X}) = \frac{1}{Z}\prod_{c\in\mathcal{C}} \phi_c(\mathcal{X}_c)$$

  - $\mathcal{C}$ is the set of maximal cliques
  - $\phi_c(\mathcal{X}_c)$ are potential functions over cliques ($c \in \mathcal{C}$)
  - $\mathcal{X}_c$ is the set of clique variables
  - Z in the normalization factor: $Z = \sum_{\mathcal{X}}\prod_{c\in\mathcal{C}} \phi_c(\mathcal{X}_c)$

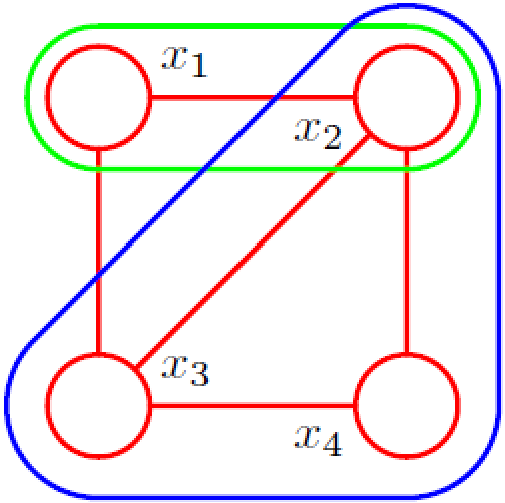


C. Bishop

K. P. Murphy

Cliques and Maximal Cliques

- Clique is a subset of a graph in which all nodes are connected together
- In the following example:
  - Cliques are: $\{x_1, x_2\}$, $\{x_2, x_4\}$, $\{x_3, x_4\}$, $\{x_1, x_3\}$, $\{x_2, x_3\}$, $\{x_1, x_2, x_3\}$, $\{x_2, x_3, x_4\}$
- In maximal cliques we can not add any new node to the clique without it ceasing to be a clique
  - Maximal cliques are: $\{x_1, x_2, x_3\}$, $\{x_2, x_3, x_4\}$



C. Bishop

OUTLINE

- Probabilistic Graphical Models
- Bayesian Networks
- Dynamic Bayesian Networks
- Markov Random Fields
  - Factorization property, cliques
  - The misconception example
  - Energy functions, Log-linear models
  - Image de-noising example
  - RBMs

The misconception example

MMK	MFD	ψ
0	0	100
0	1	1
1	0	1
1	1	100

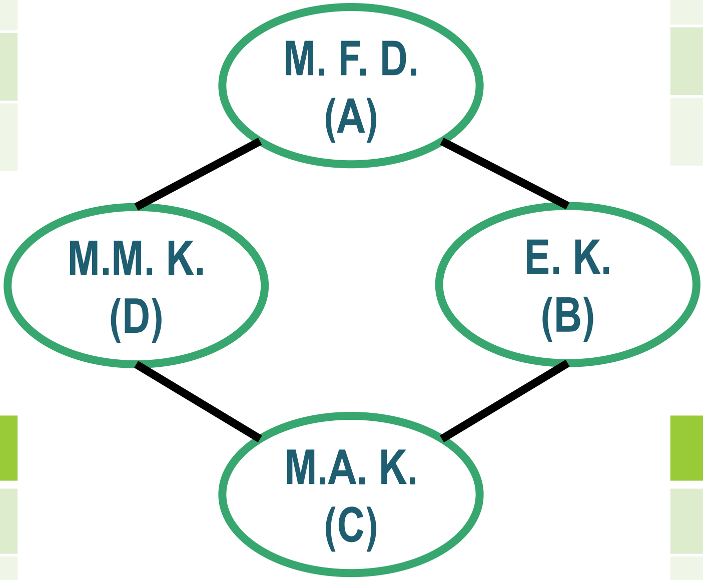
neither of two have the misconception

MFD	EK	ψ
0	0	30
0	1	5
1	0	1
1	1	10

Affinity between values

Like to agree

Like to disagree



MAK	MMK	ψ
0	0	1
0	1	100
1	0	100
1	1	1

EK	MAK	ψ
0	0	100
0	1	1
1	0	1
1	1	100

The joint distribution

Assignment				Unnormalized	Normalized
a^0	b^0	c^0	d^0	$300,000$	0.04
a^0	b^0	c^0	d^1	$300,000$	0.04
a^0	b^0	c^1	d^0	$300,000$	0.04
a^0	b^0	c^1	d^1	30	$4.1 \cdot 10^{-6}$
a^0	b^1	c^0	d^0	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^0	d^1	500	$6.9 \cdot 10^{-5}$
a^0	b^1	c^1	d^0	$5,000,000$	0.69
a^0	b^1	c^1	d^1	500	$6.9 \cdot 10^{-5}$
a^1	b^0	c^0	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^0	d^1	$1,000,000$	0.14
a^1	b^0	c^1	d^0	100	$1.4 \cdot 10^{-5}$
a^1	b^0	c^1	d^1	100	$1.4 \cdot 10^{-5}$
a^1	b^1	c^0	d^0	10	$1.4 \cdot 10^{-6}$
a^1	b^1	c^0	d^1	$100,000$	0.014
a^1	b^1	c^1	d^0	$100,000$	0.014
a^1	b^1	c^1	d^1	$100,000$	0.014

$$Z = \sum_{\mathcal{X}} \tilde{P}(\mathcal{X})$$

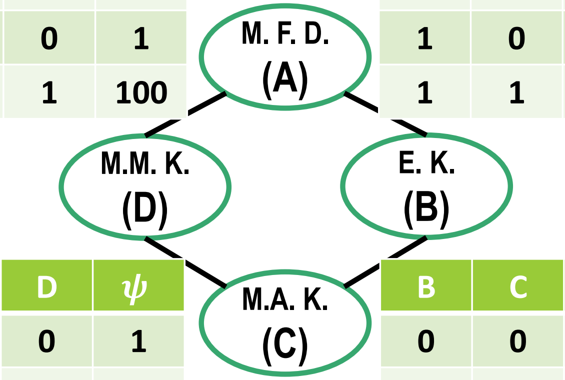
D. Koller

$$\tilde{P}(\mathcal{X}) = \prod_{c\in\mathcal{C}} \phi_c(\mathcal{X}_c)$$

$$P(\mathcal{X}) = \frac{1}{Z}\prod_{c\in\mathcal{C}} \phi_c(\mathcal{X}_c)$$

D	A	ψ
0	0	100
0	1	1
1	0	1
1	1	100

A	B	ψ
0	0	30
0	1	5
1	0	1
1	1	10



C	D	ψ
0	0	1
0	1	100
1	0	100
1	1	1

B	C	ψ
0	0	100
0	1	1
1	0	1
1	1	100

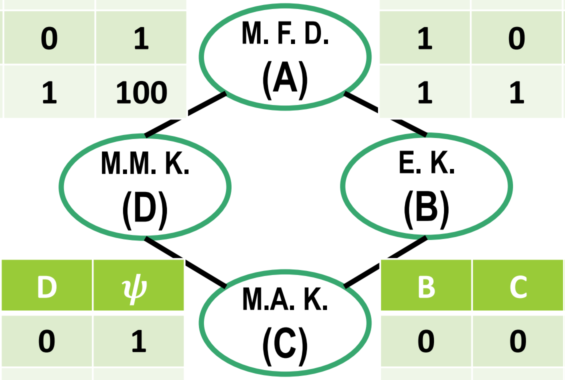
So, what do factors means?

- In your opinion, the factor $\phi_1(A,B)$ is proportional to:
  $\propto$
  - The marginal probability $\mathrm{P}(A,B)$
  - The conditional probability $\mathrm{P}(A|B)$
  - The conditional probability $\mathrm{P}(A,B|C,D)$

$$P(\mathcal{X})=\frac{1}{Z}\prod_{c\in\mathcal{C}}\phi_c(\mathcal{X}_c)$$

D	A	ψ
0	0	100
0	1	1
1	0	1
1	1	100

A	B	ψ
0	0	30
0	1	5
1	0	1
1	1	10



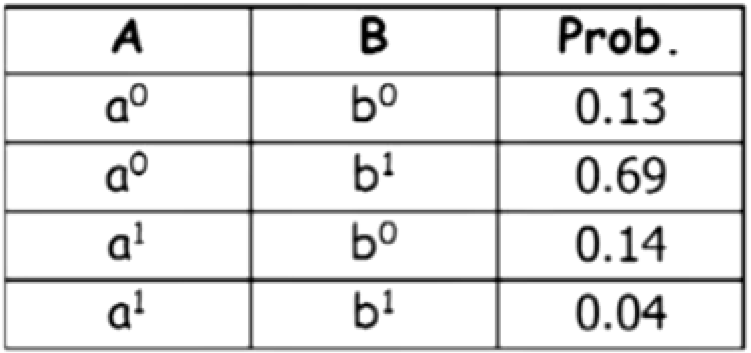
C	D	ψ
0	0	1
0	1	100
1	0	100
1	1	1

B	C	ψ
0	0	100
0	1	1
1	0	1
1	1	100

So, what do factors means?

- $P_{\Phi}(A, B), \Phi = \{\phi_1, \phi_2, \phi_3, \phi_4\}$

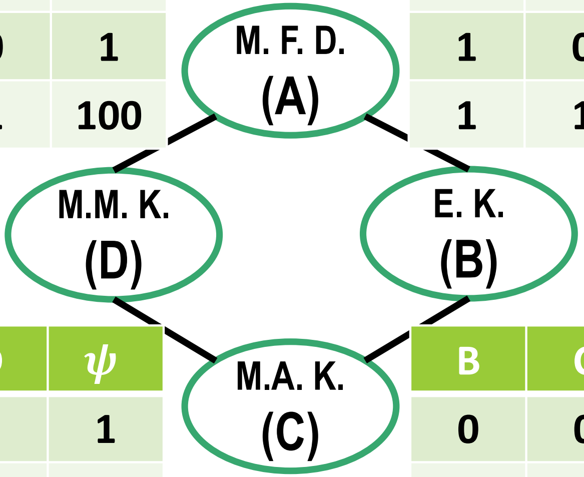
D. Koller



A	B	Prob.
a^0	b^0	0.13
a^0	b^1	0.69
a^1	b^0	0.14
a^1	b^1	0.04

D	A	ψ
0	0	100
0	1	1
1	0	1
1	1	100

A	B	ψ
0	0	30
0	1	5
1	0	1
1	1	10



C	D	ψ
0	0	1
0	1	100
1	0	100
1	1	1

B	C	ψ
0	0	100
0	1	1
1	0	1
1	1	100

In the MRFs, there is not a natural mapping between the probability distribution and the factors that are used to compose it.

OUTLINE

- Probabilistic Graphical Models
- Bayesian Networks
- Dynamic Bayesian Networks
- Markov Random Fields
  - Factorization property, cliques
  - The misconception example
  - Energy functions, Log-linear models
  - Image de-noising example
  - RBMs

Equivalent representation using energy functions

$$P(\mathcal{X}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(\mathcal{X}_c)$$

Gibbs distribution

- Energy function:

$$E(\mathcal{X}_c) = -\log\left(\phi_c(\mathcal{X}_c)\right)$$

- Equivalent representation:

$$P(\mathcal{X}) \propto \exp\left[-\sum_{c \in \mathcal{C}} E(\mathcal{X}_c)\right]$$

Boltzmann distribution

$$\prod_{c \in \mathcal{C}} \exp[-E(\mathcal{X}_c)]$$

Log linear models

- A log linear model is defined by:
  - a set of features $\mathcal{F}\{f_1(\mathcal{X}_1), \dots, f_k(\mathcal{X}_k)\}$
  - a set of weights $w_1, \dots, w_k$

again $\mathcal{X}_i$s are maximal cliques

- such that:

$$P(\mathcal{X}) \propto \exp\left[-\sum_{i=1}^{k} w_i f_i(\mathcal{X}_i)\right]$$

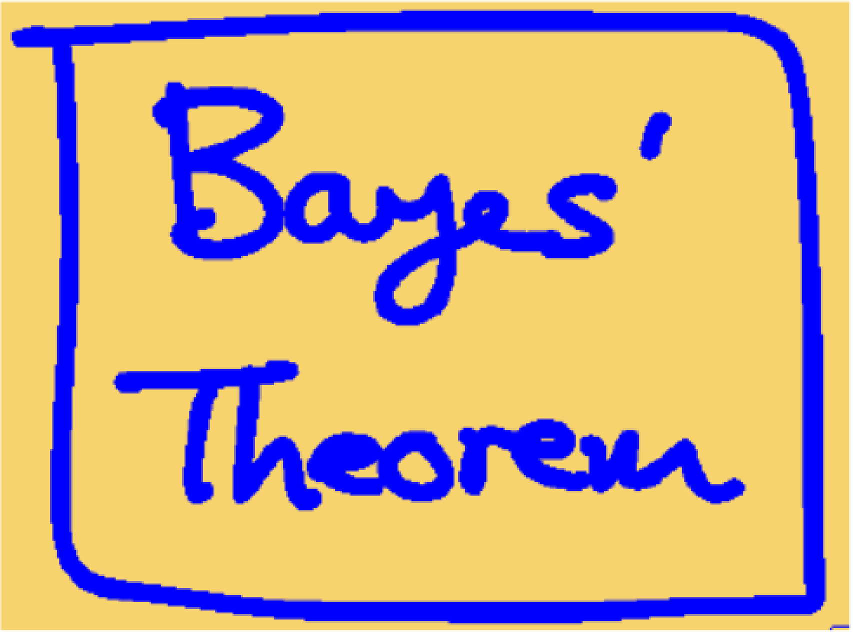
OUTLINE

- Probabilistic Graphical Models
- Bayesian Networks
- Dynamic Bayesian Networks
- Markov Random Fields
  - Factorization property, cliques
  - The misconception example
  - Energy functions, Log-linear models
  - Image de-noising example
  - RBMs

Image de-noising example

C. Bishop

- Flipping pixel color prob. is 10%
- We have an array of noisy image pixels (y_is)
- We want to infer original image (x_is)



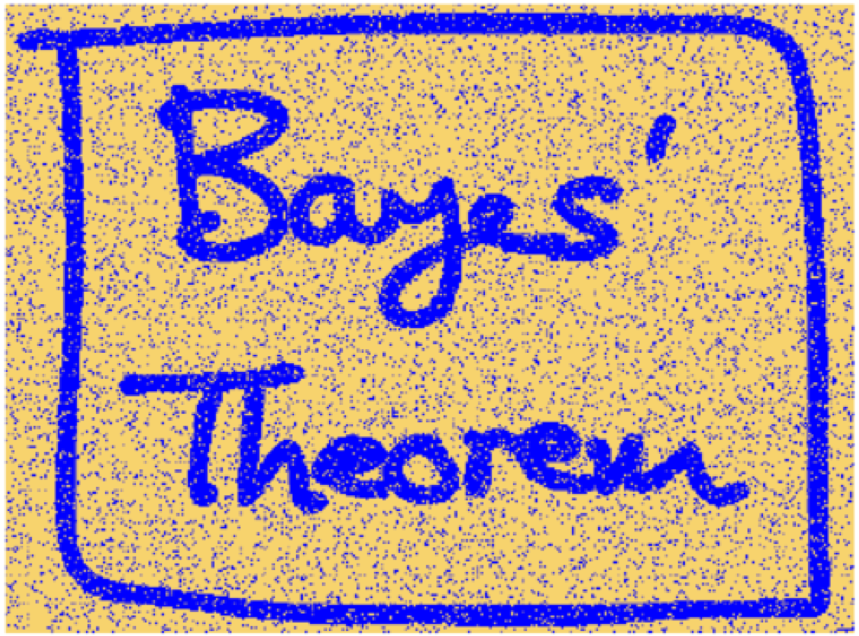
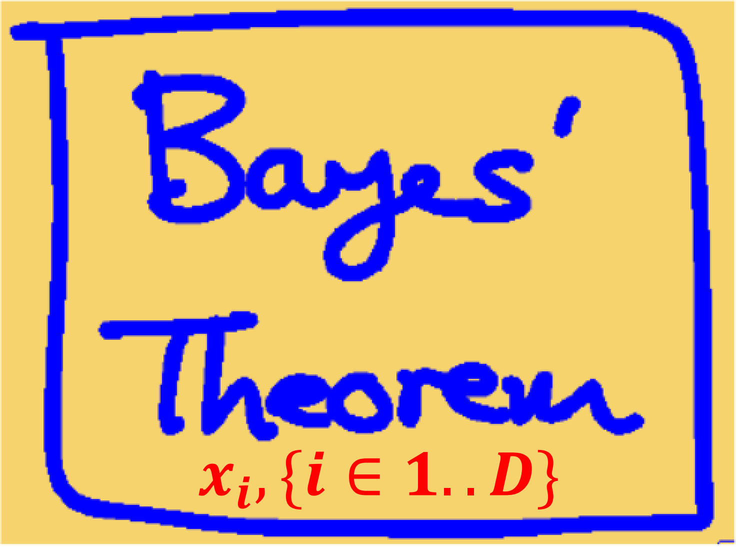
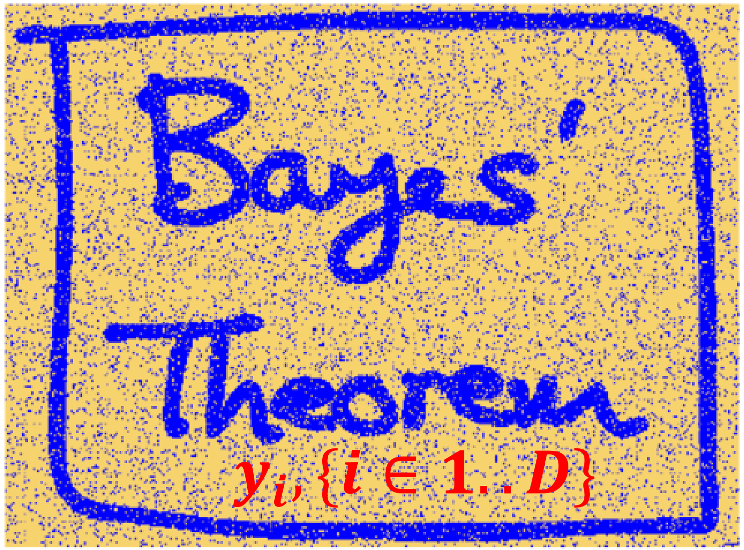


Image de-noising example, embedding our prior knowledge

- y_i and x_is are strongly correlated
  - (sine noise level is small)
- that neighboring pixels x_i and x_js in an image are strongly correlated
- Construct an MRF using this prior knowledge



$\boldsymbol{x_i, \{i \in 1..D\}}$

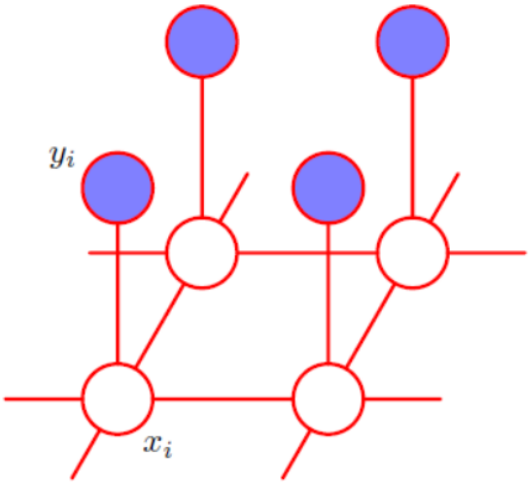


$\boldsymbol{y_i, \{i \in 1..D\}}$

Image de-noising example, model as a pairwise MRF

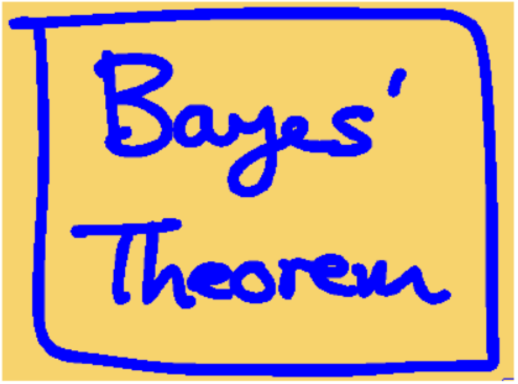
- The graph has two types of cliques:
  - each of which contains two variables

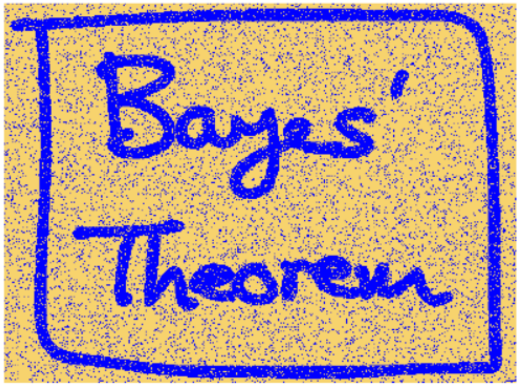
    (a pairwise MRF)
  - $\{x_i, y_i\}$ and $\{x_i, x_j\}$
  - $x_i \in \{-1, 1\}, y_i \in \{-1, 1\}$



C. Bishop

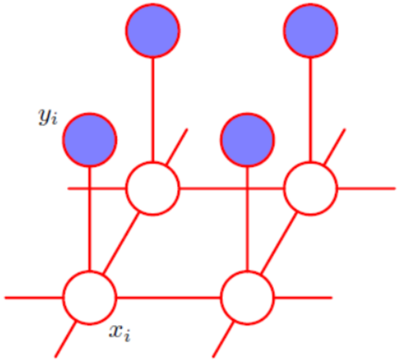
- $-\eta x_i y_i \quad \eta > 0$
- $-\beta x_i x_j \quad \beta > 0$





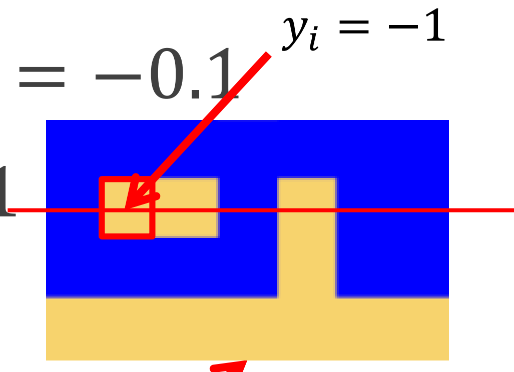
- $E(X, Y) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$

Image de-noising example, inference using ICM



- $E(X,Y) = h\sum_i x_i - \beta\sum_{\{i,j\}} x_i x_j - \eta\sum_i x_i y_i$
  - $h = 0, \beta = 1.0, \eta = 2.1$
- $E(x_i = -1) = -2.1 - (-1 + 1 - 1 - 1) = -0.1$
- $E(x_i = 1) = 2.1 - (1 - 1 + 1 + 1) = 0.1$
- $p \propto \exp[-E]$

Now, what if $\beta = 0$?



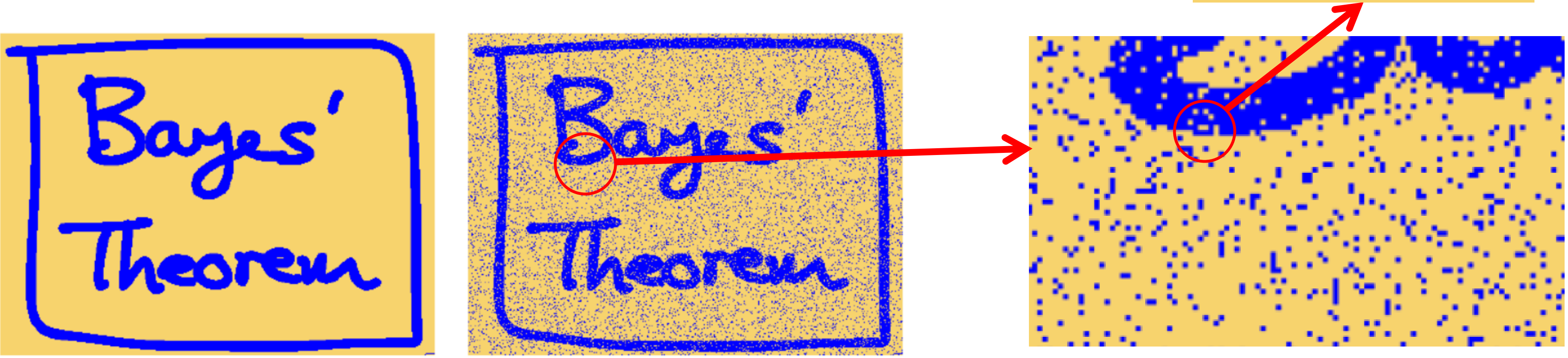
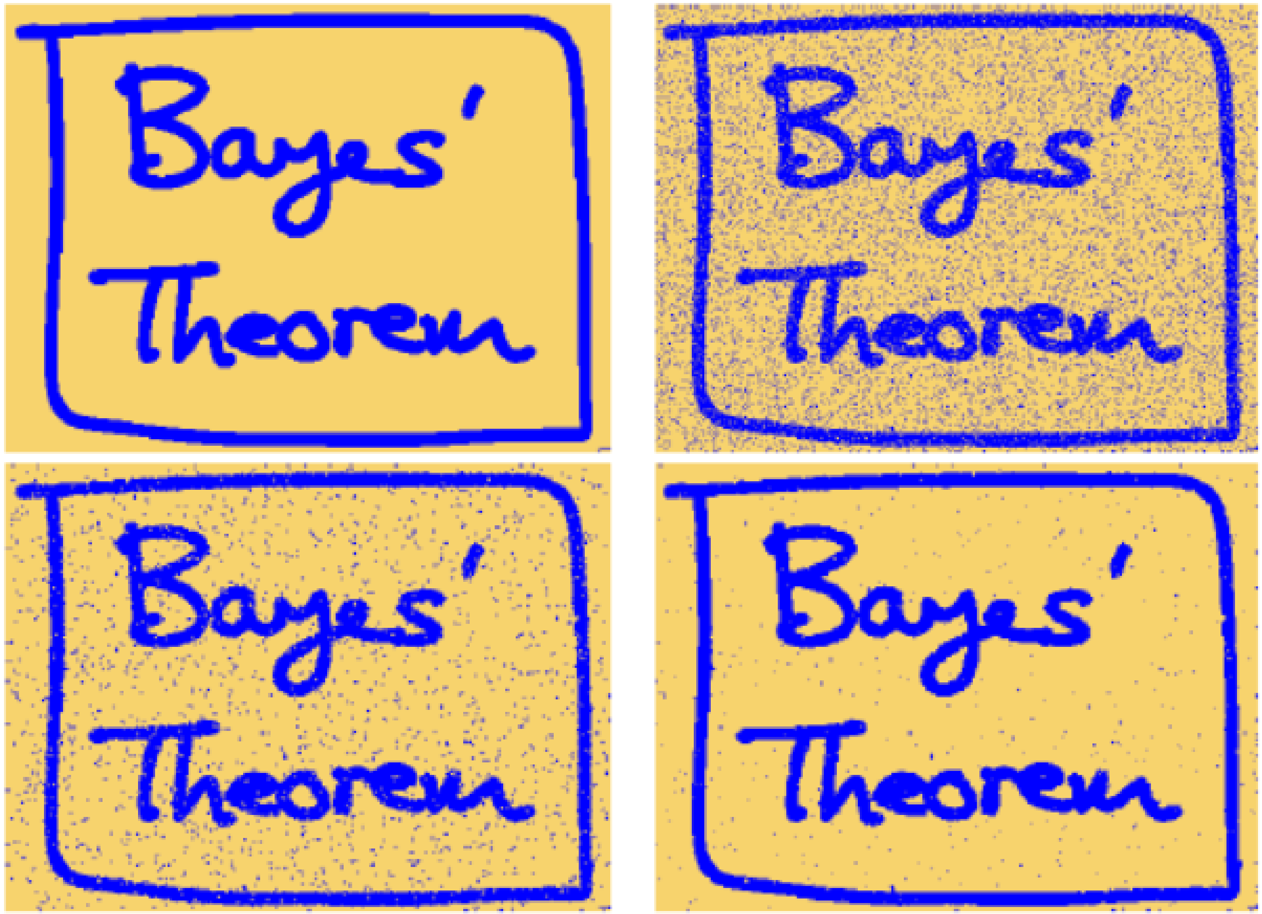
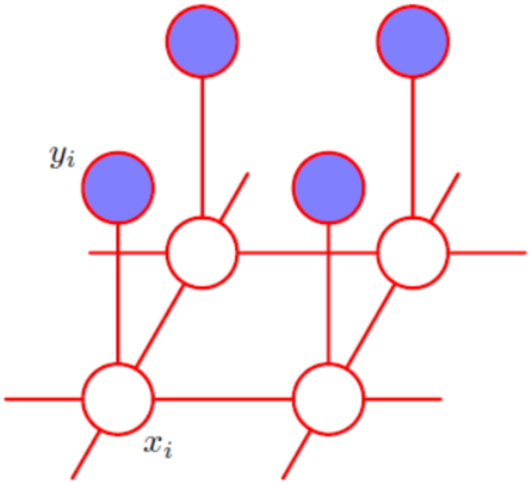


Image de-noising example, de-noising results



Graph cut algorithm

ICM algorithm



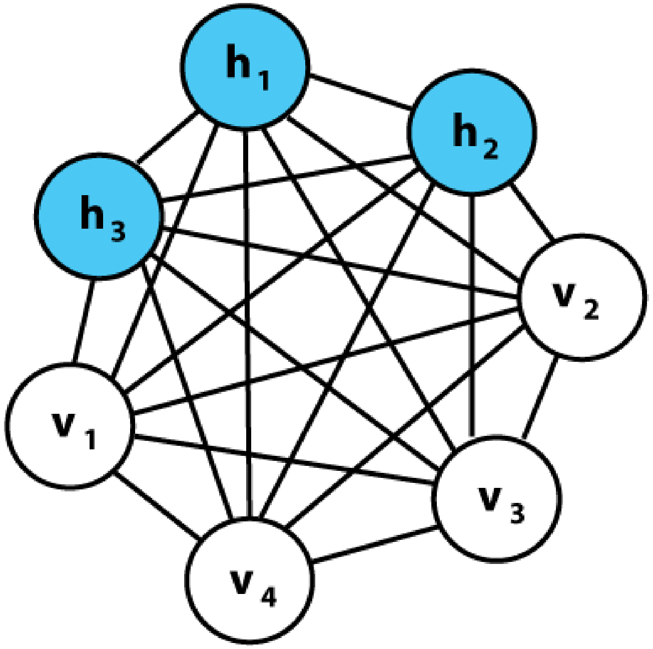
C. Bishop

OUTLINE

- Probabilistic Graphical Models
- Bayesian Networks
- Dynamic Bayesian Networks
- Markov Random Fields
  - Factorization property, cliques
  - The misconception example
  - Energy functions, Log-linear models
  - Image de-noising example
  - RBMs

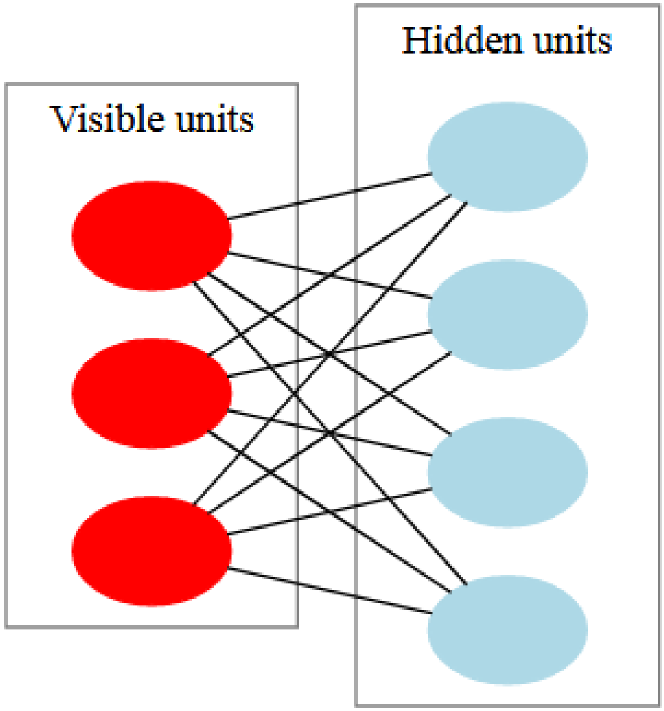
MRF Examples: Boltzmann Machines

Boltzmann Machine



$$E = -\left(\sum_{i<j} w_{ij} x_i x_j + \sum_i \theta_i x_i\right)$$

Restricted Boltzmann Machine



$$E(v,h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i w_{i,j} h_j$$

Wikipedia

References

- BISHOP, CHRISTOPHER M.: *Pattern Recognition and Machine Learning*. Bd. 4. New York : Springer, 2007 — ISBN 978-0-387-31073-2
- KOLLER, D.; FRIEDMAN, N.: *Probabilistic graphical models: principles and techniques* : The MIT Press, 2009
- MURPHY, K. P: *Dynamic Bayesian networks: representation, inference and learning*, University of California, Ph.D. Thesis, 2002
- https://www.coursera.org/learn/probabilistic-graphical-models
- Probabilistic Graphical Models, Instructor: Dr. Ahmad Nickabadi

امام علی (ع):

لا يُدْرَکُ الْعِلْمُ بِراحَةِ الْجِسْمِ.

دانش، با تن آسایی به دست نمی‌آید.

Acquiring knowledge is not possible by laziness.

غرر الحکم، ص ۳۴۸