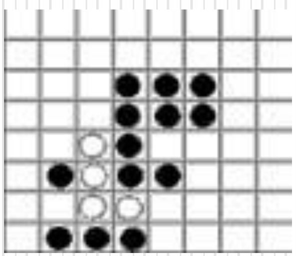
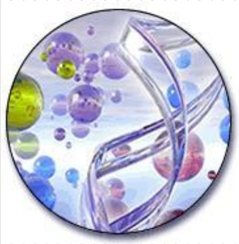
Welcome to

Introduction to Machine Learning!





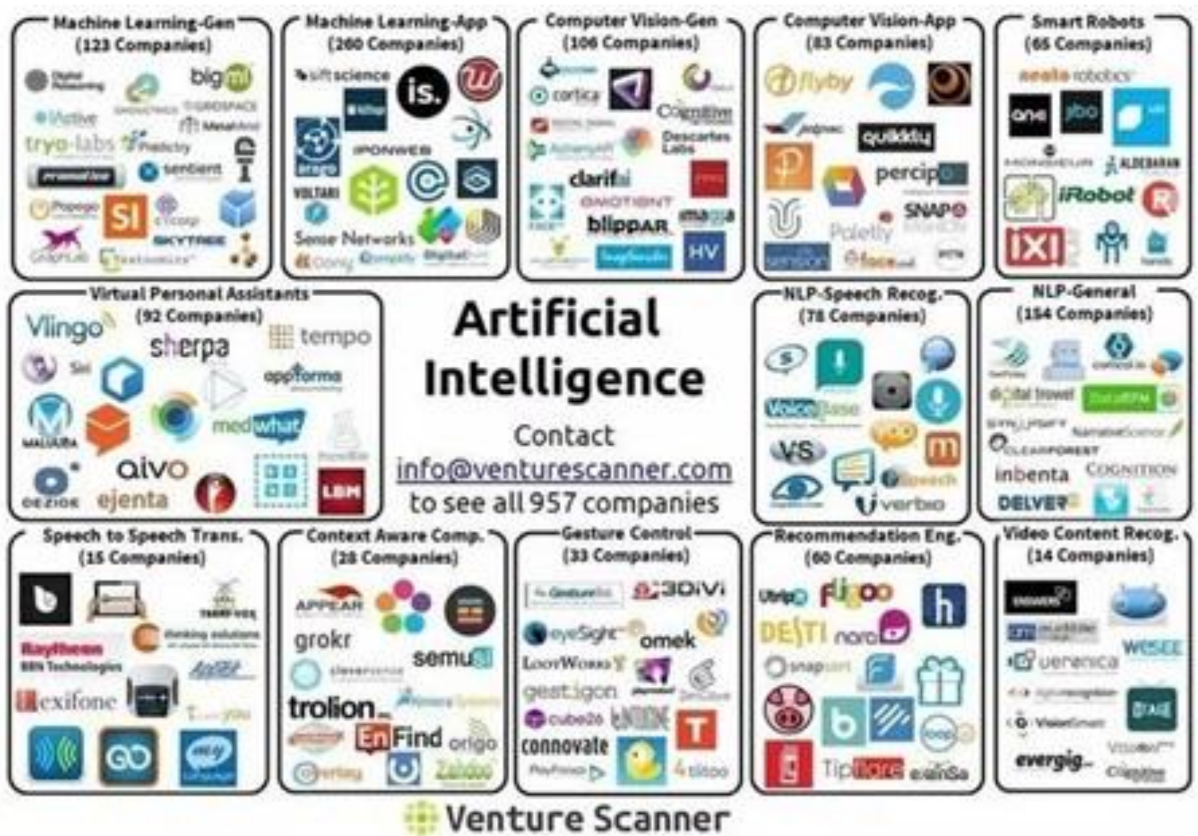
Coffee Time

April 28, 2017

Xiaolin Hu
xlhu@tsinghua.edu.cn

人工智能的市场

新智元



Venture Scanner追踪了957个人工智能公司，横跨13种类，总共融资额达到了47亿美元

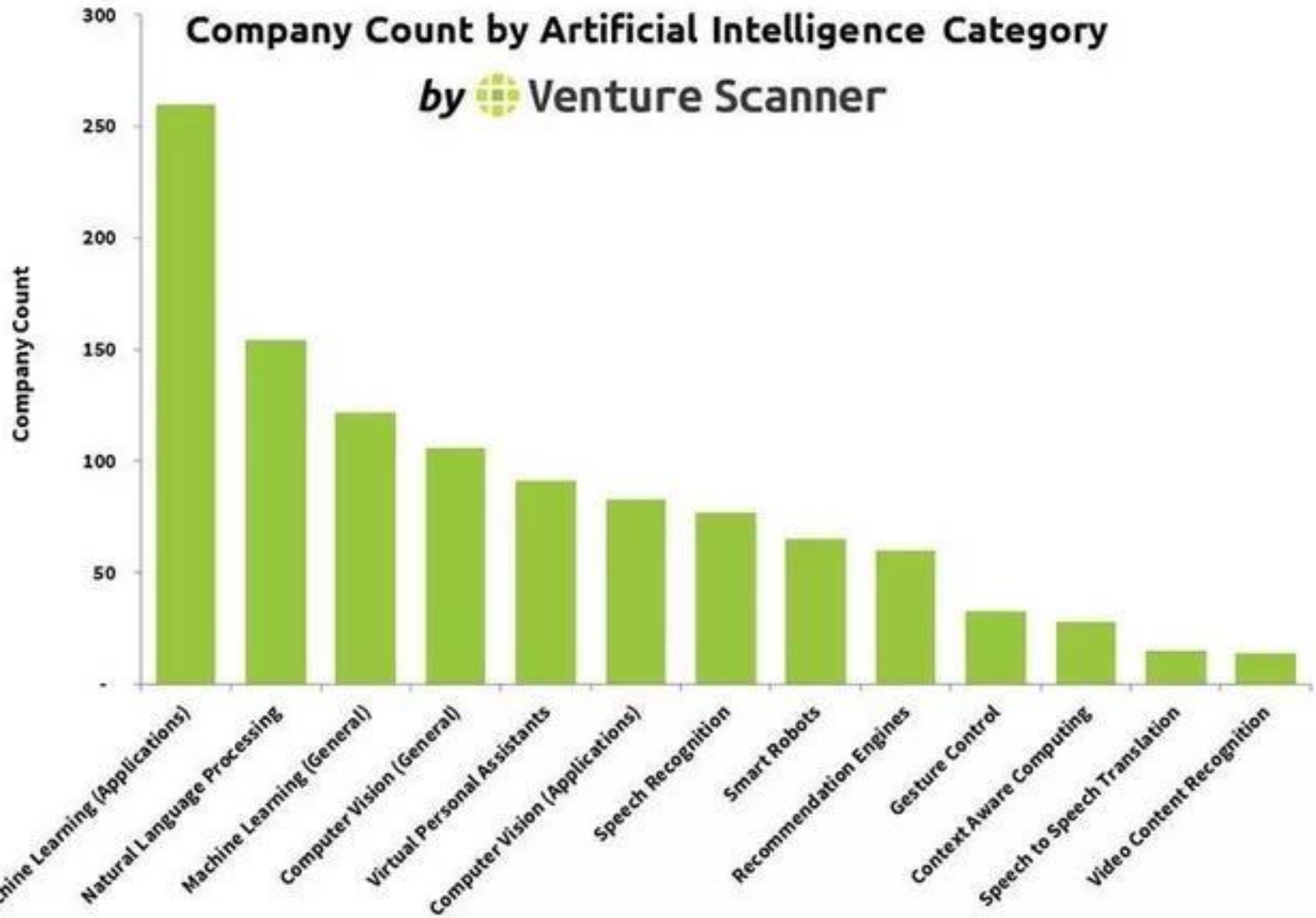
人工智能市场总览

- 深度学习/机器学习（通用）：
  - 通过在已有数据的基础上学习，建立计算机程序。例如包括了预测数据模型和软件平台，分析行为数据。
- 深度学习/机器学习（应用）：
  - 在特定领域已有数据的学习基础上，建立计算机程序。例如包括使用机器学习技术来检测银行错误，或者识别出最好的零售线索。
- 自然语言理解（通用）：
  - 建立计算机算法，能够把人类的语言输入转化成能够理解的表示。例如自动生成叙述文，并且挖掘文本数据。

- 自然语言理解（语音识别）：
  - 处理语音的片段，确定准确的单词，并从中得到含义。例如检测语音命令、并将其转化为可操作数据的软件。
- 计算机视觉/图像识别（通用）：
  - 处理和分析图片，并从中识别出物体，得到相关的信息。例如视觉搜索平台和图片标记的API
- 计算机视觉/图像识别（应用）：
  - 在垂直领域使用图片处理的技术。例如识别人脸或者通过拍照搜索零售产品的软件。
- 手势控制：
  - 通过手势和计算机交互和通信。例如一些软件，可以通过身体的移动来控制电子游戏，或者通过手势独自操作电脑和电视。
- 虚拟个人助理：
  - 根据反馈和命令，执行日常任务和服务。例如一些网站和App，能够帮助人们管理日历。

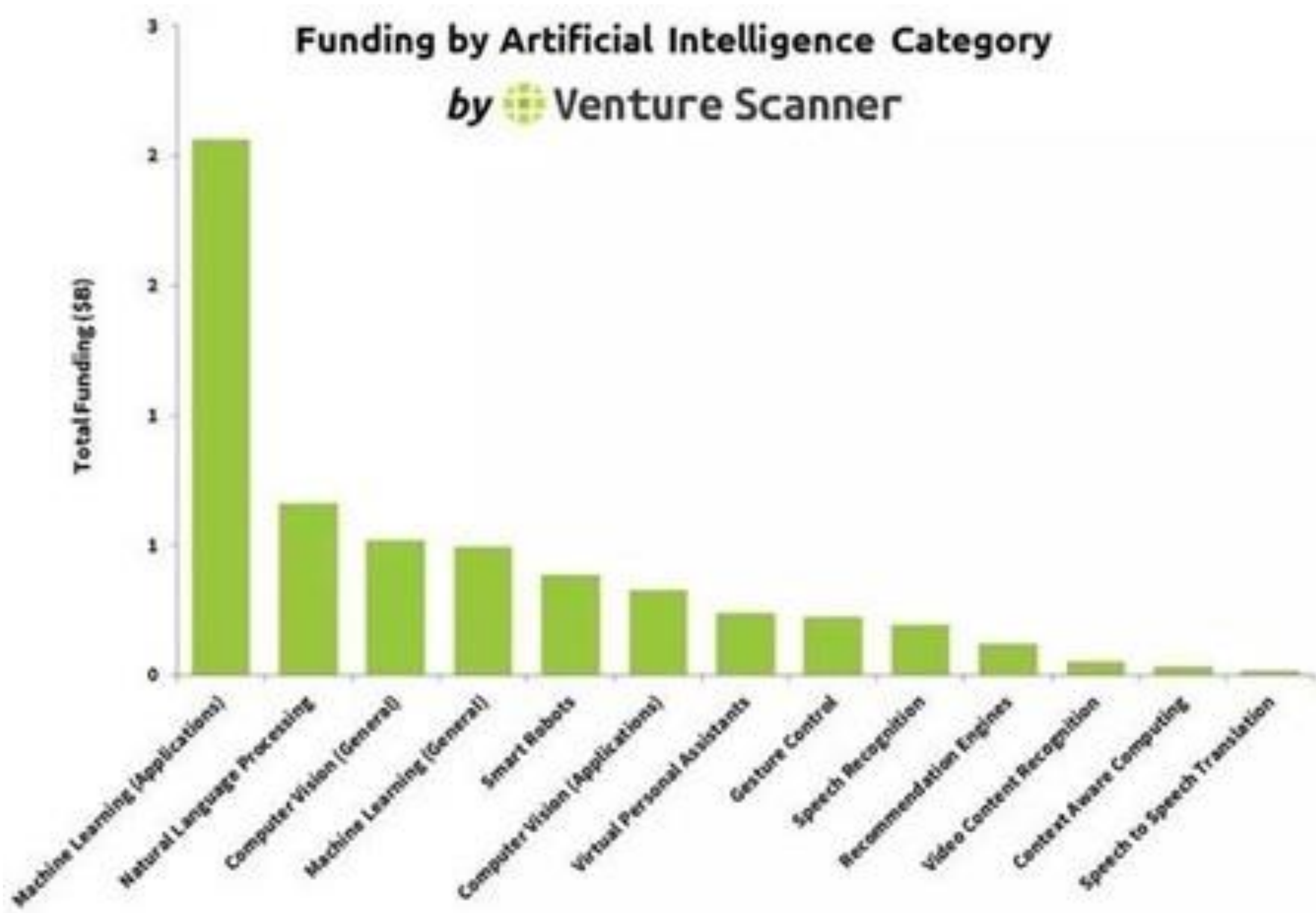
- 智能机器人：
  - 可以从他们的经验中学习，并且根据条件和环境反馈自主行动。例如家庭机器人，可以根据人们的情绪进行反应。
- 推荐引擎和协同过滤：
  - 预测用户对一些项目，例如电影和餐厅的偏好和兴趣，并提供个性化的推荐建议。
- 上下文感知计算：
  - 可以自动察觉它的背景环境. 例子还包括检测到环境黑暗的时候，灯光自动亮起来。
- 语音到语音的翻译：
  - 识别出一个人的语音，并且马上自动翻译成另一种语言。
- 视频自动内容识别：
  - 通过把采样的视频内容和视频库的文件对比，通过该视频的独特性识别出内容。例如在用户上传视频的时候，通过对它采样并和视频库对比，识别出是否盗版。

不同类别公司的数量

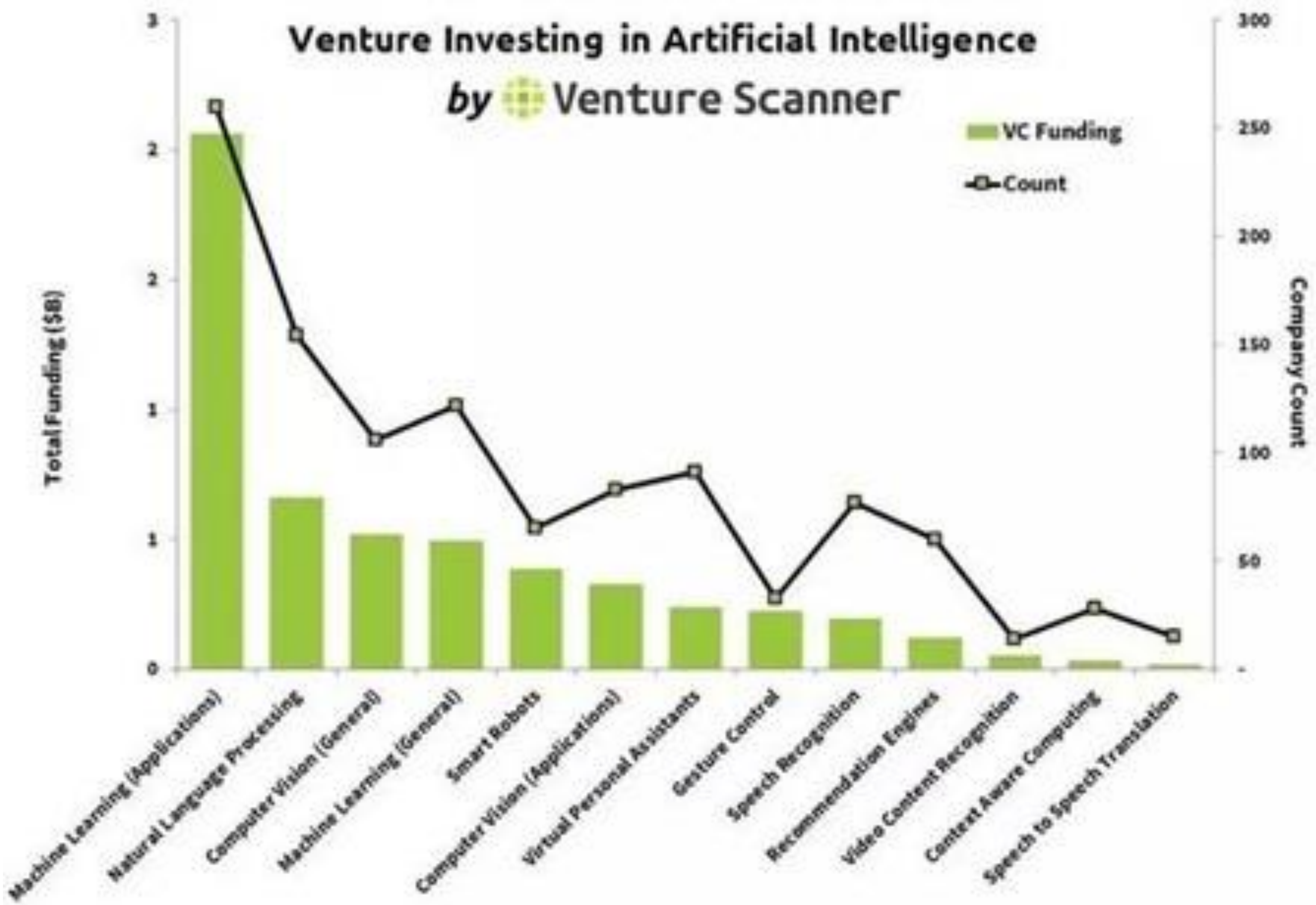


Company Count by Artificial Intelligence Category
by Venture Scanner
Company Count
300
250
200
150
100
50
-
Machine Learning (Applications)
Natural Language Processing
Machine Learning (General)
Computer Vision (General)
Virtual Personal Assistants
Computer Vision (Applications)
Speech Recognition
Smart Robots
Recommendation Engines
Gesture Control
Context Aware Computing
Speech to Speech Translation
Video Content Recognition

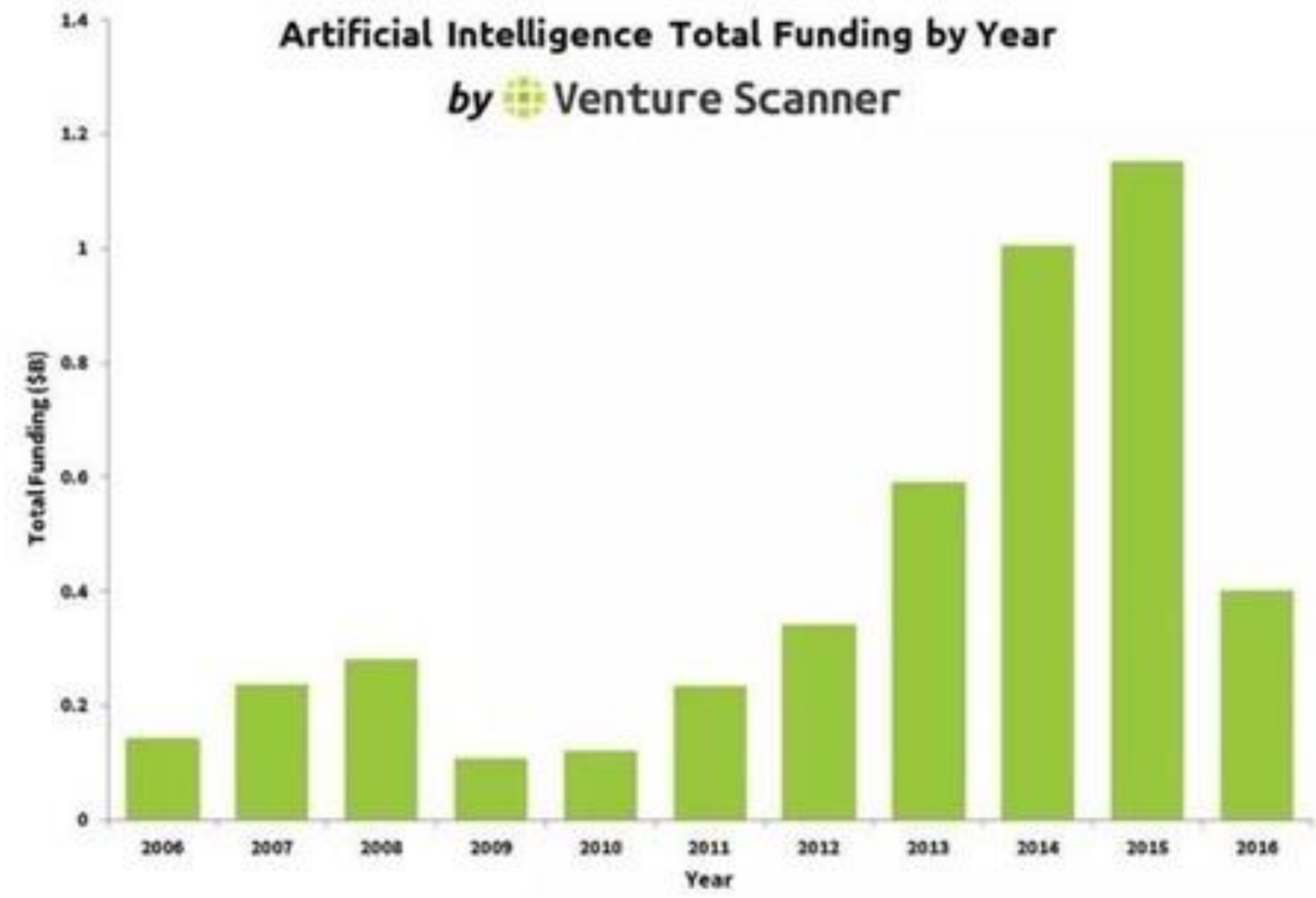
不同类别公司的融资情况



不同公司的风险投资情况



人工智能历年总投资额



Artificial Intelligence Total Funding by Year
by Venture Scanner
Total Funding ($B)
1.4
1.2
1
0.8
0.6
0.4
0.2
0
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
Year

人工智能公司数量，按国家计算

这是人工智能和机器学习
最好的时代！

Topic 11. Probabilistic Graphical Models

Xiaolin Hu

xlhu@tsinghua.edu.cn

Updated on April 28, 2017

Materials from “Pattern Recognition and Machine Learning" by Bishop (2006)

Outline

- Motivation
- Bayesian networks
  - Generative model
  - Conditional independence and D-separation
- Markov random fields
  - Conditional independence and graph separation
  - Joint distribution factorization

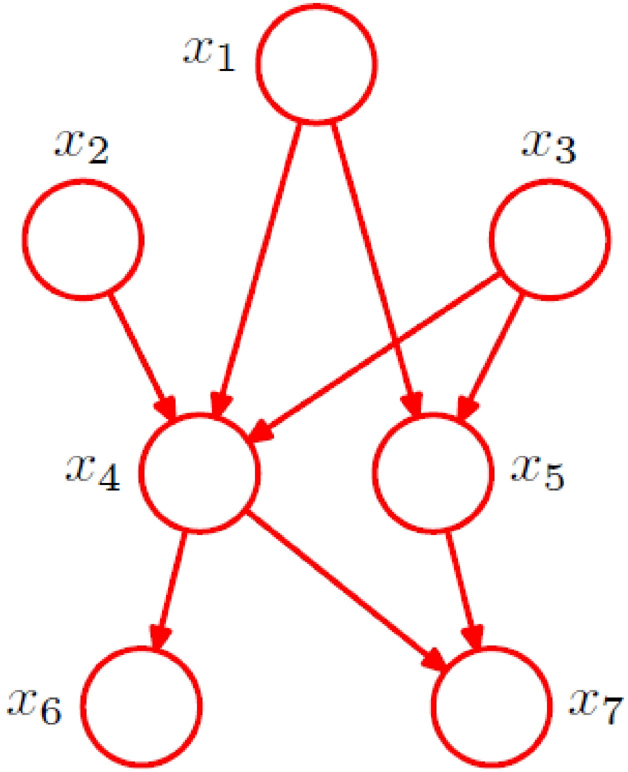
Motivation

- Many things have correlated factors which may constitute a complicated probabilistic model
- Seven variables
  - x_1, x_2, x_3 are independent to each other
  - x_4 depends on x_1, x_2, x_3
  - x_5 depends on x_1, x_3
  - x_6 depends on x_4
  - x_7 depends on x_4, x_5
  - What's the joint distribution?

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

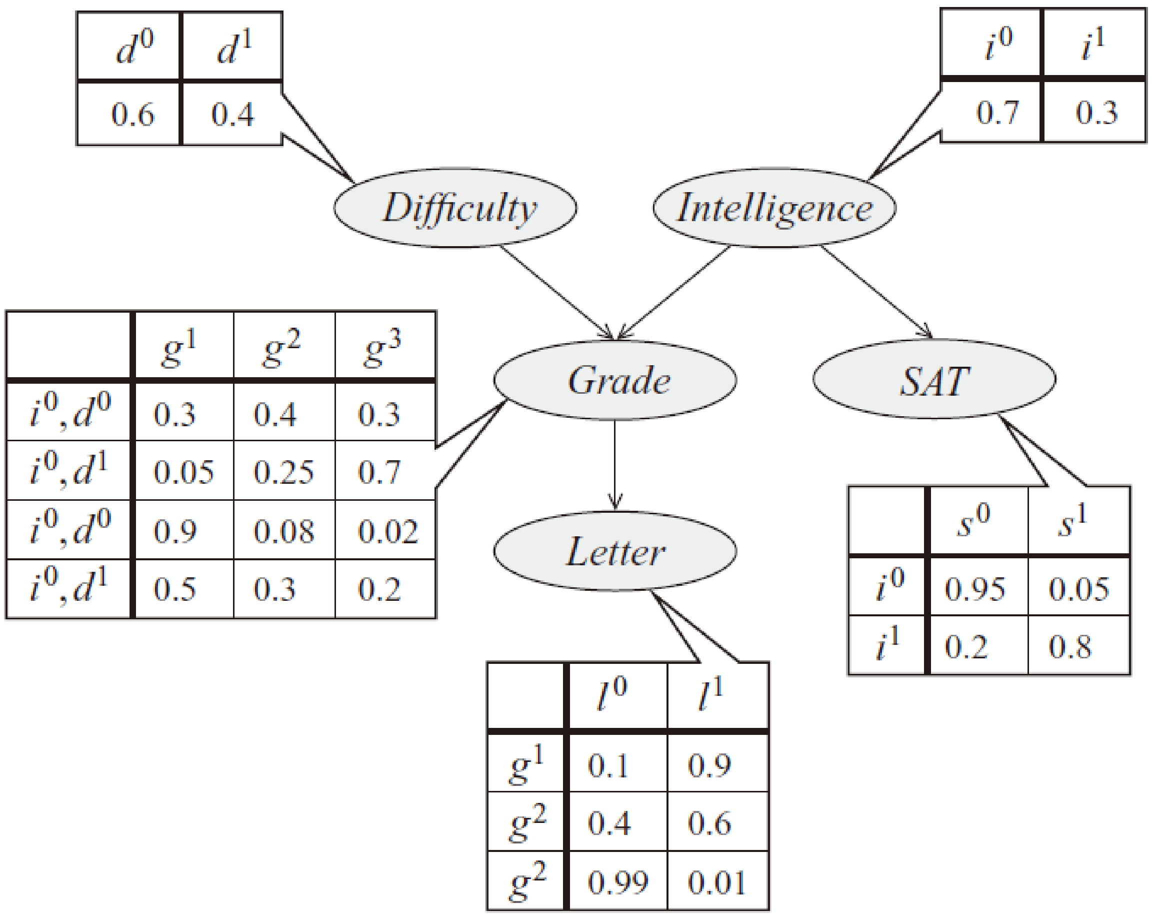
Motivation

A concise representation



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

A problem in reality



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^0, d^0	0.9	0.08	0.02
i^0, 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^2	0.99	0.01

Probabilistic graphical models

- Advantages
  - They provide a simple way to visualize the structure of a probabilistic model
  - Conditional independence properties and other properties can be obtained by inspection of the graph
  - Complex computations can be expressed in terms of graphical manipulations
- Types
  - Bayesian networks – directed graphical models
  - Markov random fields – undirected graphical models

Probabilistic graphical models

- Basic problems
  - Representation
  - Inference
  - Parameter estimation

Outline

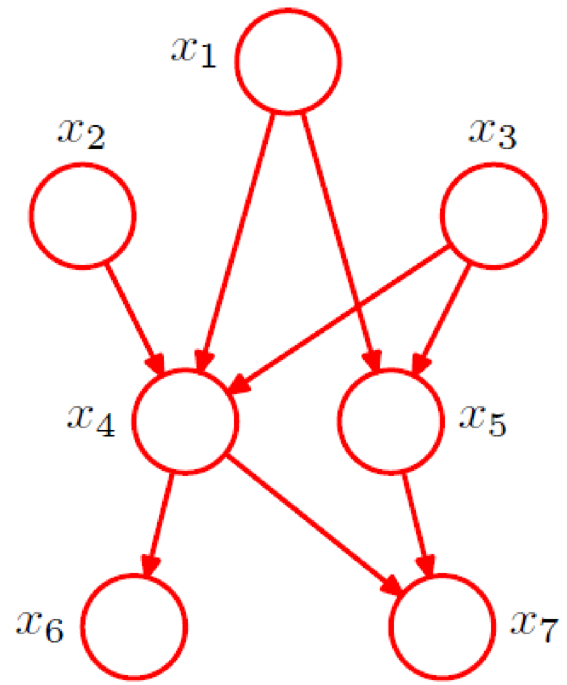
- Motivation
- Bayesian networks
  - Generative model
  - Conditional independence and D-separation
- Markov random fields
  - Conditional independence and graph separation
  - Joint distribution factorization

Bayesian networks

- Directed acyclic graphical models
  - Nodes: variables
  - Arrows: conditional distribution
- The joint distribution for a graph with K nodes

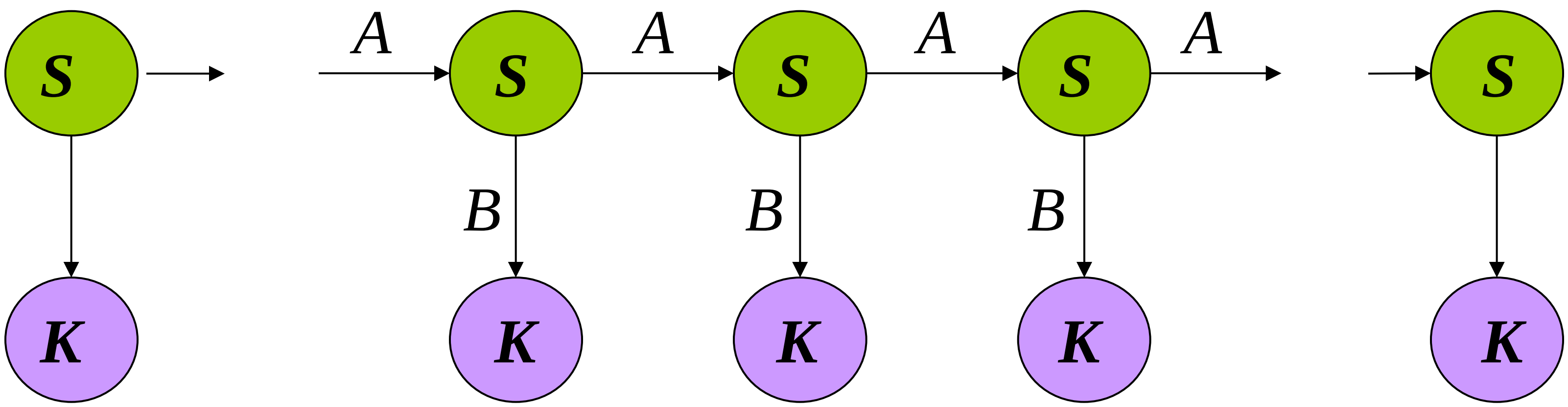
$$p(\mathbf{x}) = \prod_{k=1}^{K} p(x_k|\mathrm{pa}_k)$$

where $\mathbf{pa}_k$ stands for parent nodes of x_k



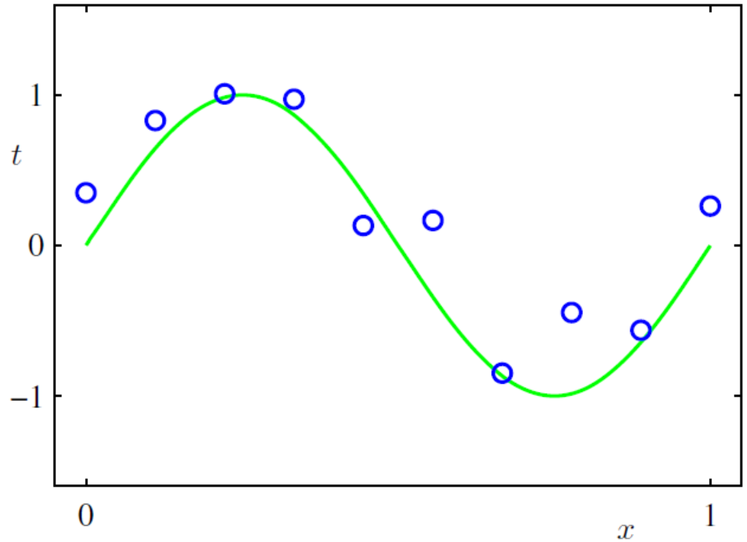
$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$

Hidden Markov model



- A Bayesian network with special structure
- The simplest temporal model
  - There are other temporal models, e.g., the linear dynamical systems (LDS)
- These models are also called dynamic Bayesian networks (DBN)

Example: polynomial regression



- N training samples: $(x_1, t_1), \ldots, (x_N, t_N)$
- Polynomial fit: $y(x, w) = \sum_{j=0}^{M} w_j x^j$

- Assume the error $t - y$ has a zero mean Gaussian distribution, i.e.,

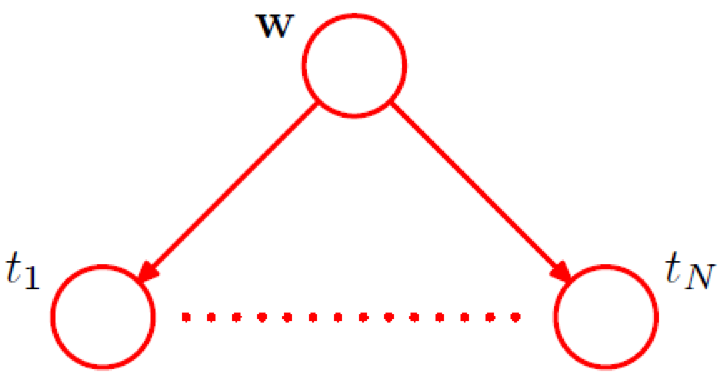
$$p(t|x, w, \beta) = \mathcal{N}(t|y(x, w), \sigma^2)$$

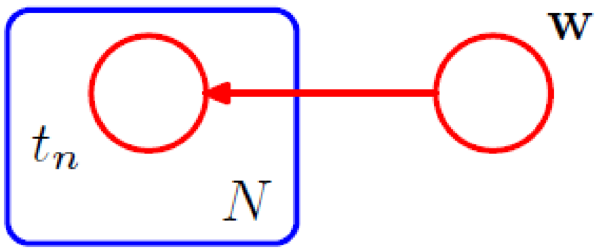
Conditional independence

- The joint distribution of t and w

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

Example: polynomial regression





Graphical representation

Use a plate to represent multiple nodes

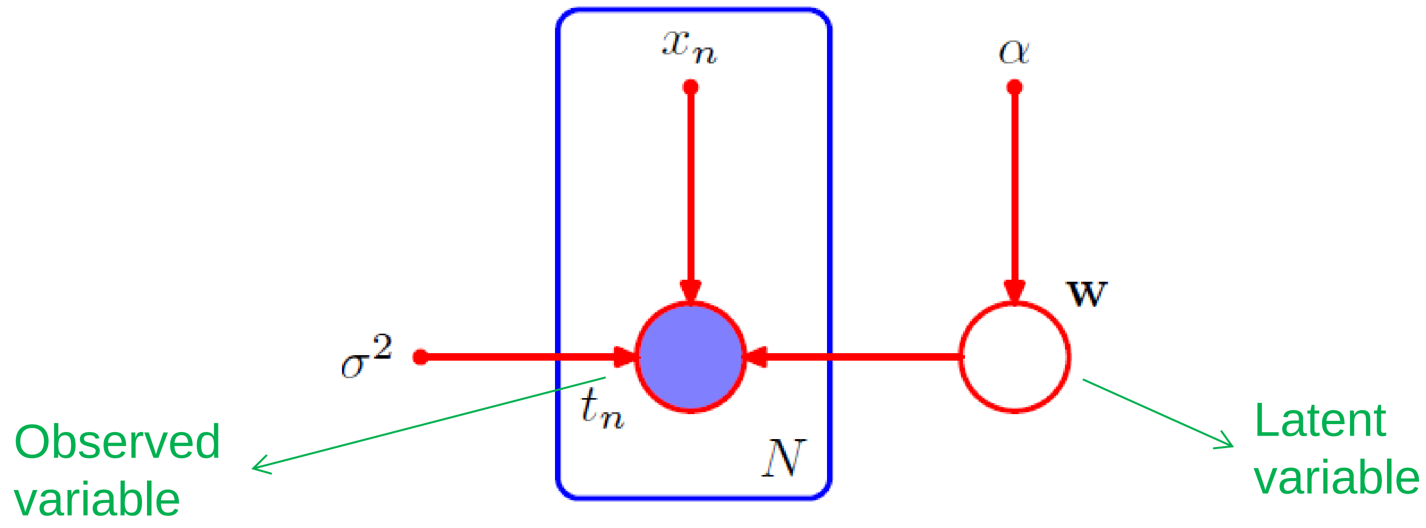
$$p(t|x, w, \beta) = \mathcal{N}(t|y(x, w), \sigma^2)$$

Example: polynomial regression

- Make the parameters and variables explicit

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

where α is a parameter controlling the prior distribution



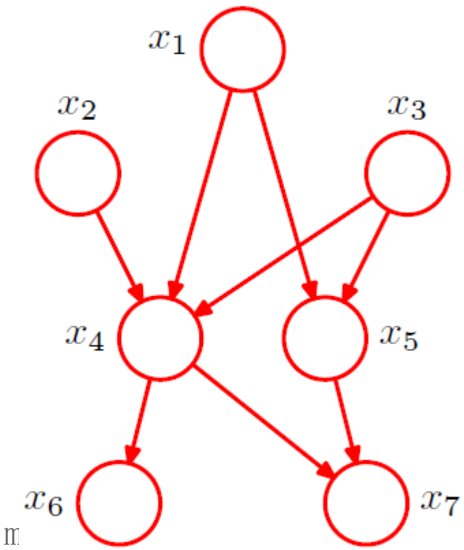
Ancestral sampling

- How to draw a sample from the joint distribution of K variables

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

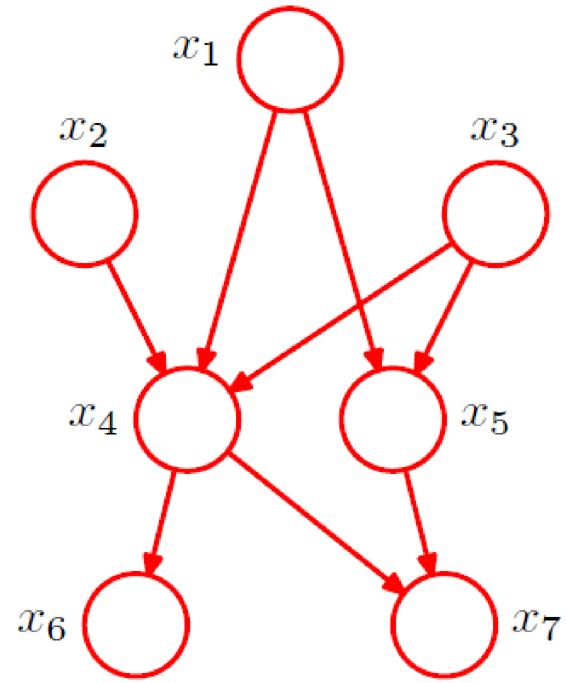
- Suppose the variables have been ordered such that each node has a higher number than any of its parents

Start with the lowest-numbered node and draw a sample from $p(x_n|\mathrm{pa}_n)$ in which the parent variables have been set to their sampled values



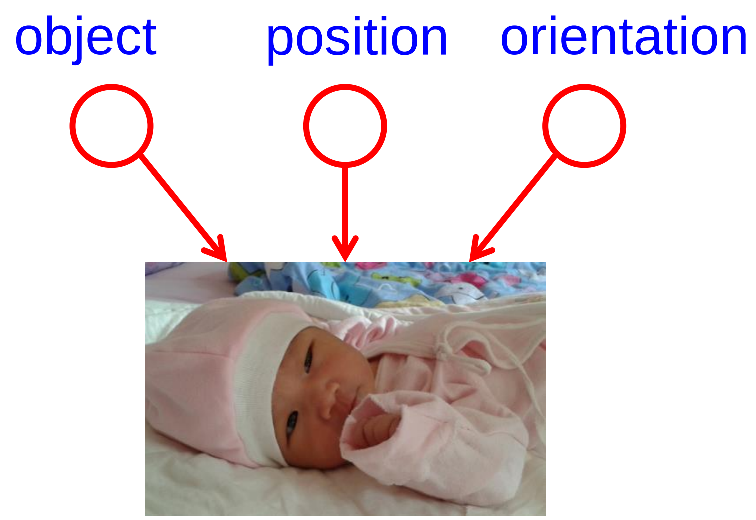
Ancestral sampling

- Illustration
  - Step 1: draw $\hat{x}_1, \hat{x}_2, \hat{x}_3$
  - Step 2: draw $\hat{x}_4, \hat{x}_5$
  - Step 3: draw $\hat{x}_6, \hat{x}_7$
- Then a sample $(\hat{x}_1, \ldots, \hat{x}_7)$ is obtained
- How to draw a sample from some marginal distribution, e.g., $p(x_2, x_4)$?
  - Draw a sample from the full joint distribution then discard $\{\hat{x}_{j \neq 2,4}\}$



Generative models

- In practical applications
  - higher numbered nodes – observed variables
  - lower numbered nodes – latent variables (needn’t have any physical interpretations)
- Graphical models express the processes by which the observed data are generated



Conditional independence

- Definition: suppose the conditional distribution of a, given b and c, does not depend on b, so that

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

  - We say a is conditionally independent of b given c.
  - Equivalently

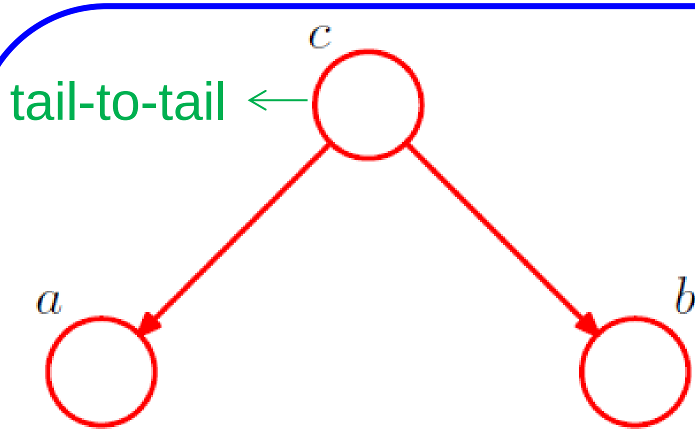
$$\begin{aligned} p(a,b|c) &= p(a|b,c)p(b|c) \\ &= p(a|c)p(b|c). \end{aligned}$$

  - Or simply $a \perp\!\!\!\perp b \mid c$

Basic graph I

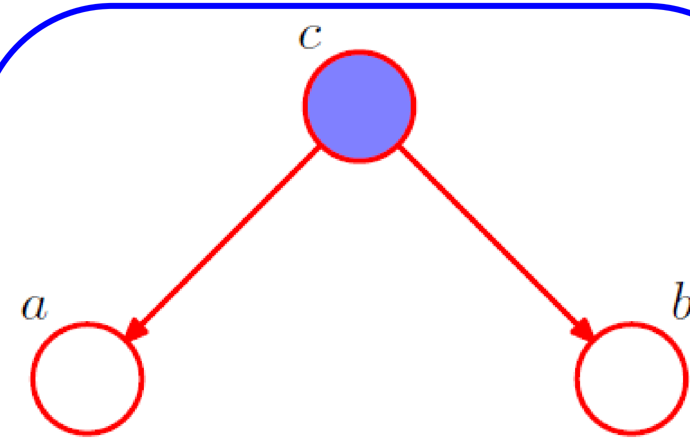
$$p(a,b,c) = p(a|c)p(b|c)p(c)$$

If *c* is observed, the path is blocked!



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

$$a \not\perp\!\!\!\perp b \mid \emptyset$$



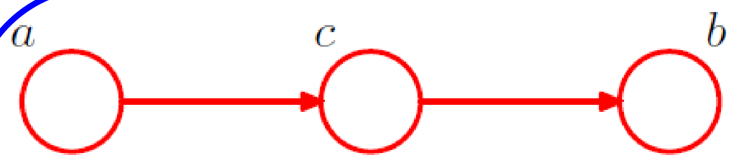
$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

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

Basic graph II

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

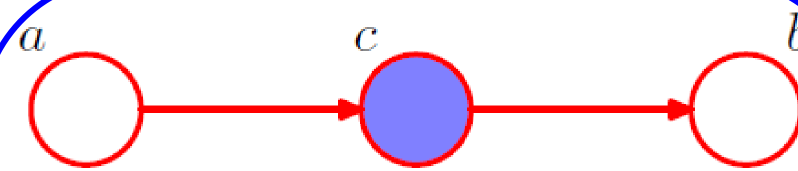
If c is observed, the path is blocked!



head-to-tail

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

$$a \not\perp\!\!\!\perp b \mid \emptyset$$



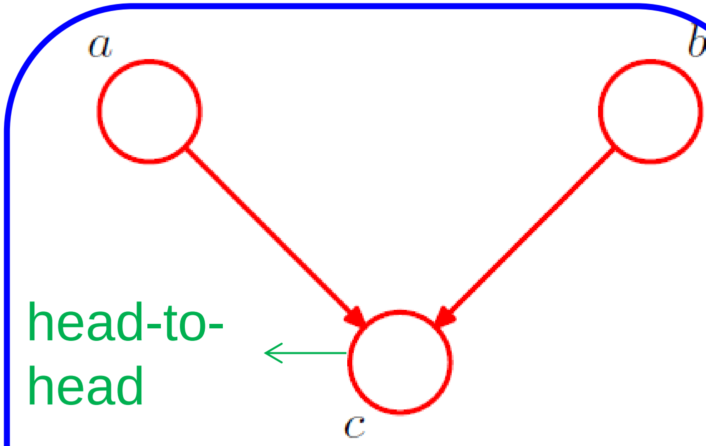
$$\begin{aligned} p(a,b|c) &= \frac{p(a,b,c)}{p(c)} \\ &= \frac{p(a)p(c|a)p(b|c)}{p(c)} \\ &= p(a|c)p(b|c) \end{aligned}$$

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

Basic graph III

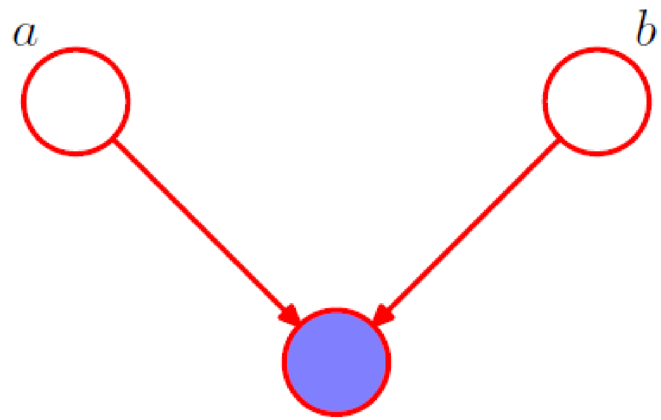
$$p(a,b,c) = p(a)p(b)p(c|a,b)$$

If c is observed, the path is unblocked!



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

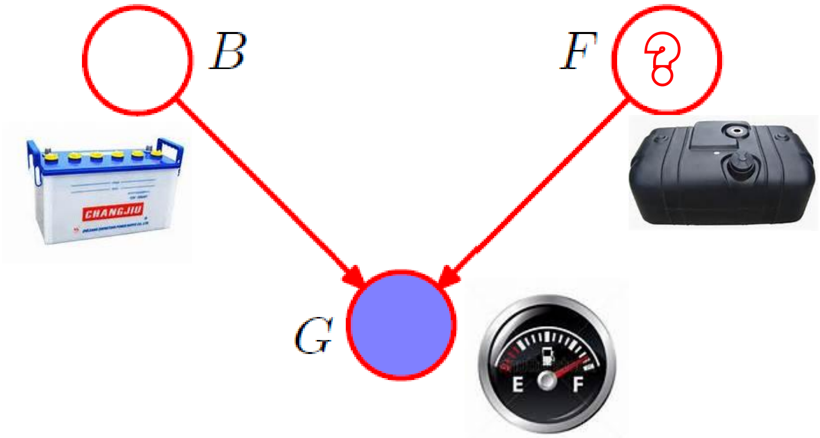
$$a \perp\!\!\!\perp b \mid \emptyset$$



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

$$a \not\!\perp\!\!\!\perp b \mid c$$

"Explaining away"



$$
\begin{aligned}
p(B=1) &= 0.9 \\
p(F=1) &= 0.9 \\
p(G=1|B=1,F=1) &= 0.8 \\
p(G=1|B=1,F=0) &= 0.2 \\
p(G=1|B=0,F=1) &= 0.2 \\
p(G=1|B=0,F=0) &= 0.1
\end{aligned}
$$

G is observed to be 0.

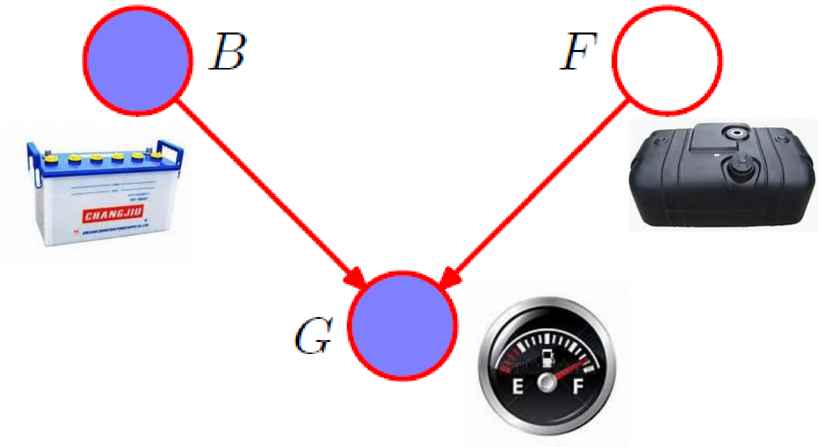
$$p(G=0) = \sum_{B\in\{0,1\}} \sum_{F\in\{0,1\}} p(G=0|B,F)p(B)p(F) = 0.315$$

$$p(G=0|F=0) = \sum_{B\in\{0,1\}} p(G=0|B,F=0)p(B) = 0.81$$

$$p(F=0|G=0) = \frac{p(G=0|F=0)p(F=0)}{p(G=0)} \simeq 0.257$$

The prob. of *F*=0 increases from 0.1 to 0.257 after observing *G*=0

"Explaining away"



G is observed to be 0.
If B is also observed to be 0, then

$$p(F=0|G=0,B=0)$$
$$= \frac{p(G=0|B=0,F=0)p(F=0)}{\sum_{F\in\{0,1\}} p(G=0|B=0,F)p(F)}$$
$$\simeq 0.111$$

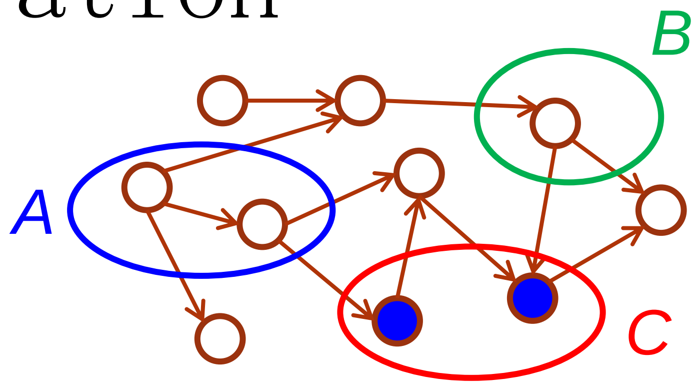
The prob. of F=0 decreases from 0.257 to 0.111 after observing B=0

The battery is flat explains away the observation that the fuel gauge reads empty.

If G is observed, F depends on B!

This is also true if any descendant of G is observed!

D-separation

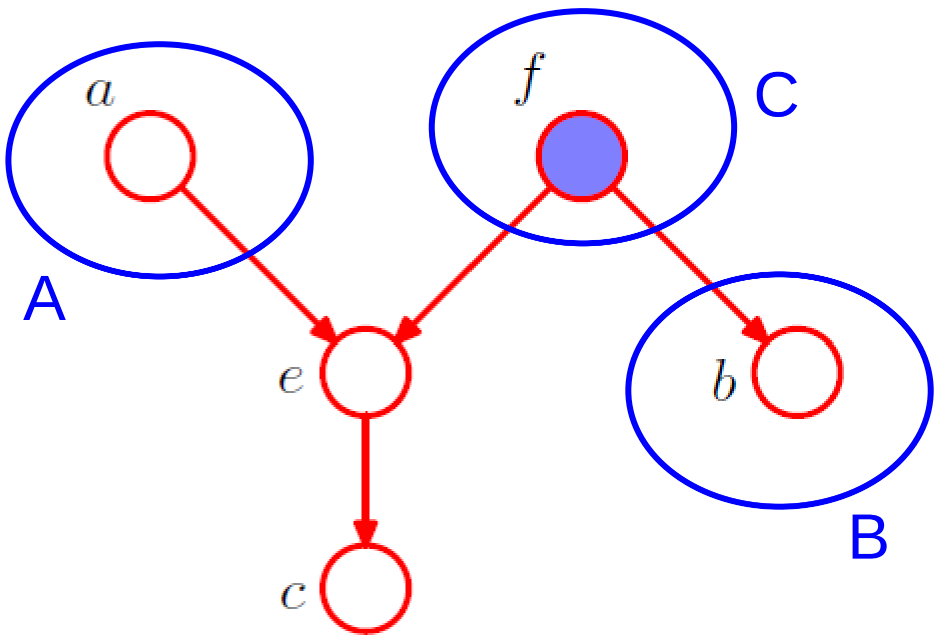


- Suppose A, B, C are arbitrary non-intersecting sets of nodes in a graph
- We want to know if $A \perp\!\!\!\perp B|C$

If all paths from any node in A to any node in B are blocked, then A is said to be d-separated from B by C, and $A \perp\!\!\!\perp B|C$

A path from any node in A to any node in B is blocked if it includes a node such that either

- the node is a head-to-tail or tail-to-tail node and it is in C
- the node is a head-to-head node, and neither the node nor any of its descendants is in C



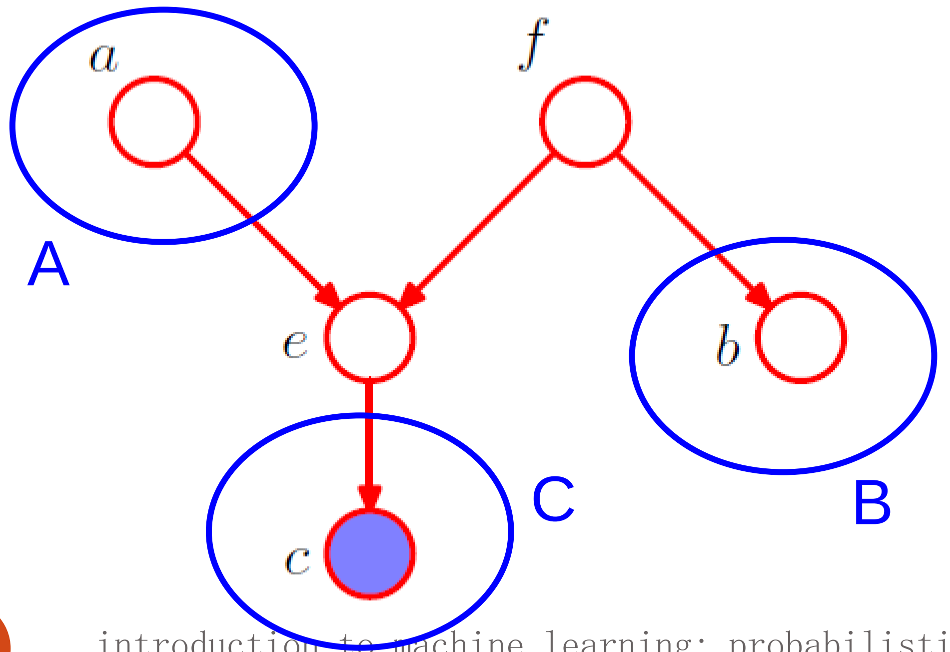
a to b blocked by f?
a to b blocked by e?

YES!

$$A \perp\!\!\!\perp B | C$$

A path from any node in A to any node in B is blocked if it includes a node such that either

- the node is a head-to-tail or tail-to-tail node and it is in C
- the node is a head-to-head node, and neither the node nor any of its descendants is in C

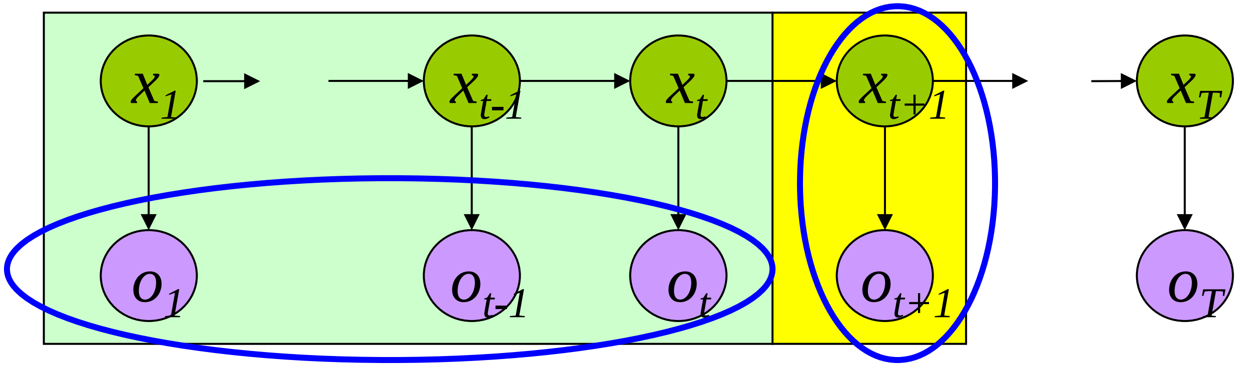


a to b blocked by f?
a to b blocked by e?

NO!

$$A \not\perp\!\!\!\perp B \mid C$$

Review of HMM forward algorithm



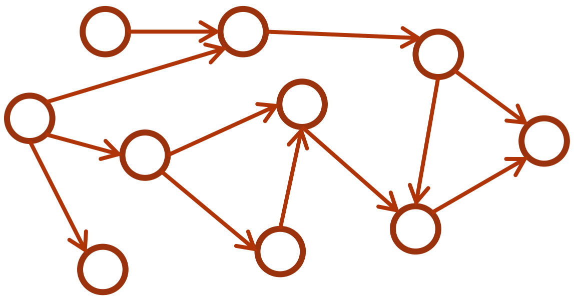
$$\alpha_{t+1}(j) = \sum_{i=1\ldots N} P(o_1 \ldots o_t, o_{t+1}, x_{t+1} = j \mid x_t = i) P(x_t = i)$$

$$= \sum_{i=1\ldots N} P(o_1 \ldots o_t \mid x_t = i) P(o_{t+1}, x_{t+1} = j \mid x_t = i) P(x_t = i)$$

This step can follows from d-separation

Theoretical foundations

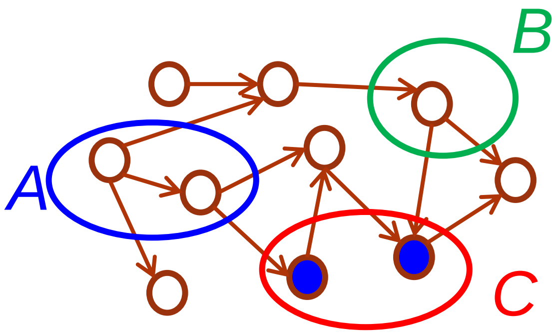
A directed graph



- represents a factorization of the joint probability distribution

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

- expresses conditional independence obtained by d-separation criterion



The two properties are equivalent!

All distributions satisfying the factorization property are those meet the d-separation criterion; vice versa

Outline

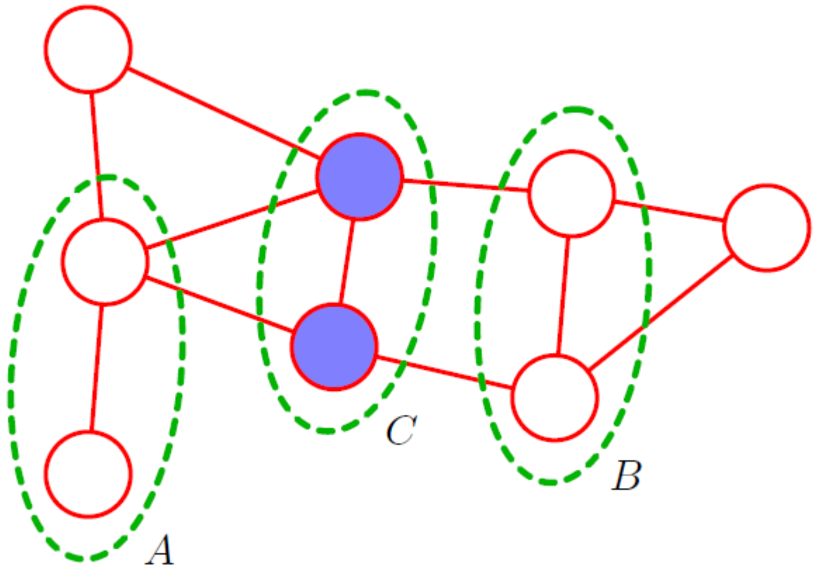
- Motivation
- Bayesian networks
  - Generative model
  - Conditional independence and D-separation
- Markov random fields
  - Conditional independence and graph separation
  - Joint distribution factorization

Markov Random Fields

- Also known as Markov networks or undirected graphical models
- One motivation:
  - Due the presence of head-to-head nodes in directed graph, the conditional independence is inconvenient to be captured
  - Can we define a graph in which the conditional independence is determined by simple graph separation?
  - How about removing the arrows?

Conditional independence

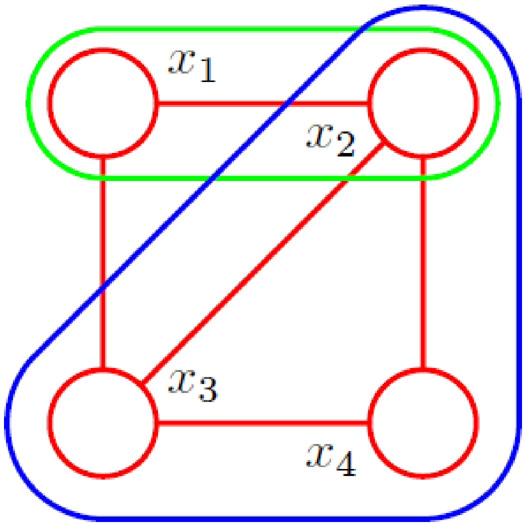
graph separation



- If all paths from A to B pass through one or more nodes in set C, then

$$A \perp\!\!\!\perp B | C$$

Maximum clique



Cliques: $\{x_1, x_2\}, \{x_1, x_3\}, \{x_2, x_3\}, \{x_2, x_4\}, \{x_3, x_4\}$

Maximum cliques: $\{x_1, x_2, x_3\}, \{x_2, x_3, x_4\}$

Clique: a subset of nodes in which there is a link between all pairs of nodes

Maximum clique: a clique such that it is not possible to include any other nodes to form a new clique

Factorization

- The joint distribution

$$p(\mathbf{x}) = \frac{1}{Z}\prod_C \psi_C(\mathbf{x}_C)$$

  - x_C: the nodes in a maximum clique C
  - $\psi_C(x_C)$: potential function which is always positive
  - Z: partition function $Z = \sum_{\mathbf{x}}\prod_C \psi_C(\mathbf{x}_C)$

Exponential functions are often used as the potential function

$$\psi_C(x_C) = \exp\{-E(x_C)\}$$

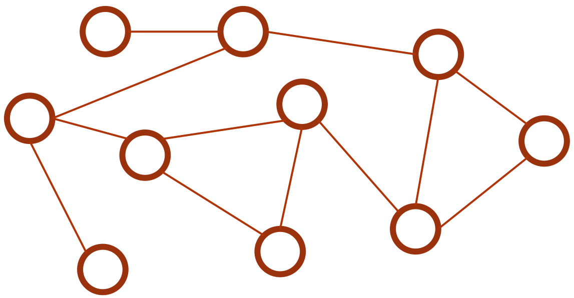
where $E(x_C)$ is called an energy function

Lower energy,
higher prob.

Theoretical foundations

Hammersley-Clifford theorem (Clifford, 1990)

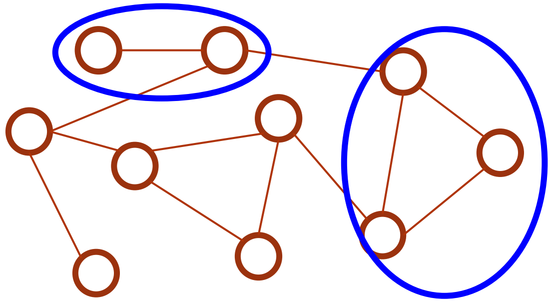
An undirected graph



- represents a factorization of the joint probability distribution

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C} \psi_C(\mathbf{x}_C)$$

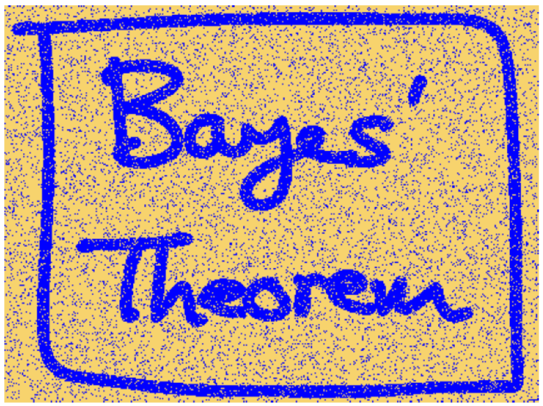
- expresses conditional independence obtained by graph separation criterion



The two properties are equivalent!

All distributions satisfying the factorization property are those meet the graph separation criterion; vice versa

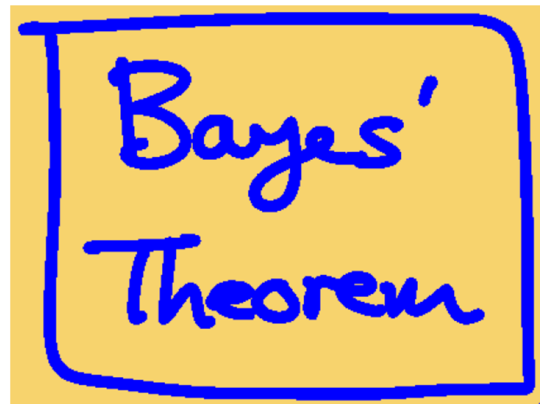
Image de-noising



Corrupted

$y_i \in \{-1, +1\}$

=



Original

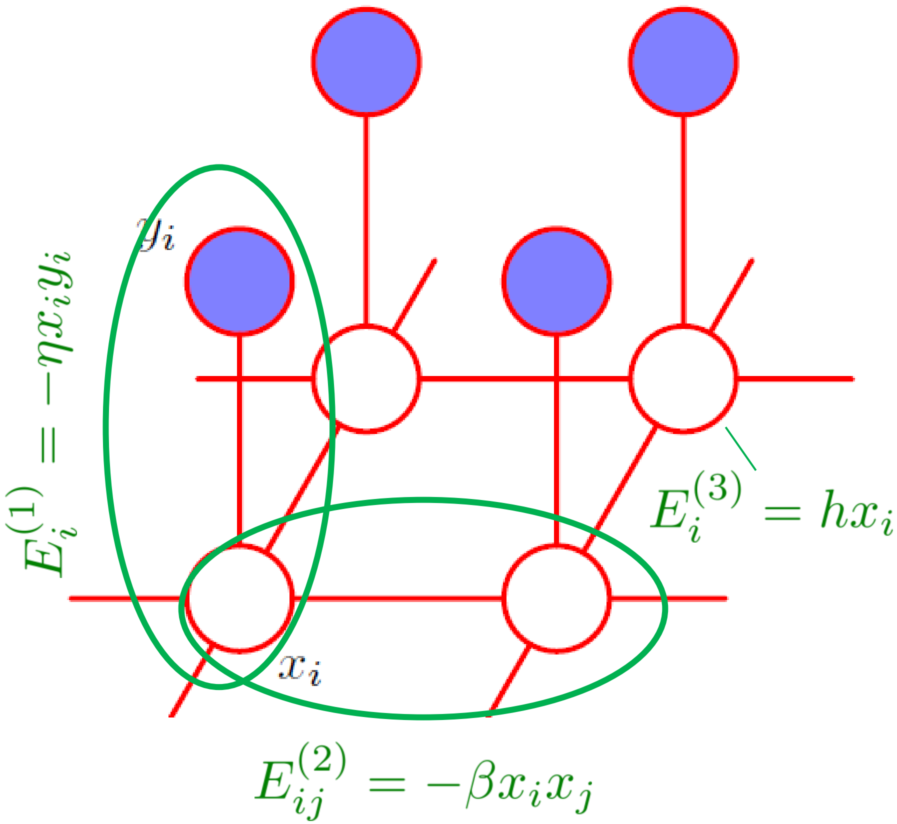
$x_i \in \{-1, +1\}$

+ 10% noise

(random flipping)

- Prior knowledge
  - Low level noise $\rightarrow x_i$ and y_i are correlated
  - Neighboring pixels x_i and x_j in the original image are strongly correlated

Image de-noising



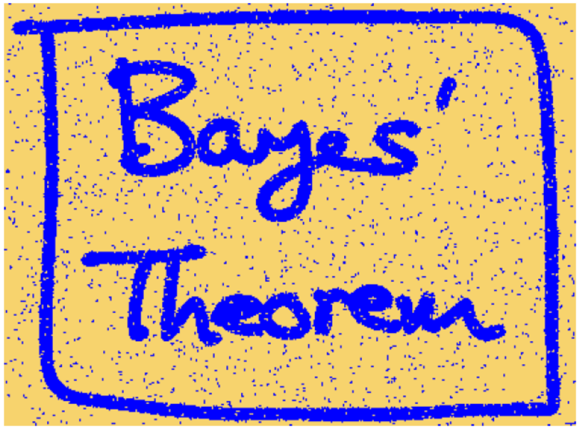
- Two types of maximum cliques
- For each clique, same pixel values imply lower energy ($\beta, \eta, h > 0$)
- A bias term is added to encourage particular sign in preference to the other

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}$$

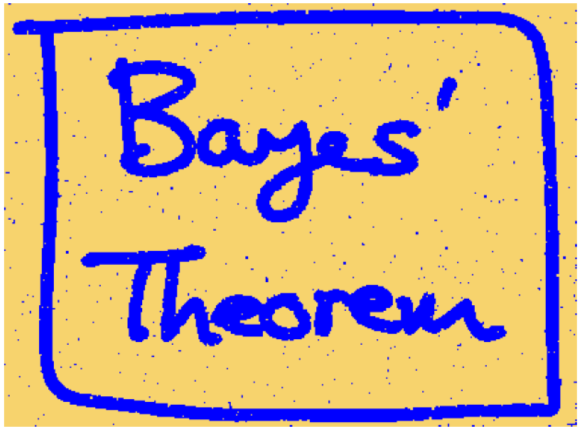
$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$

Image de-noising

- y_i is observed
- Suppose the parameters β, η, h are fixed
- We want to know x_i which minimizes the total energy $\rightarrow$ Inference
  - Iterated conditional modes (ICM)
  - Graph cut



ICM



Graph cut

Overview

- Motivation
- Bayesian networks
  - Generative model
  - Conditional independence and D-separation
- Markov random fields
  - Conditional independence and graph separation
  - Joint distribution factorization

Homework Deadline May 12 (Friday)

- For the linear SVM in the non-separable case

$$\min_{w,b} \frac{1}{2}\langle w, w\rangle + C\sum_i \varepsilon_i$$

$$\text{s.t. } y_i\left(\langle w, x_i\rangle + b\right) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0$$

derive its dual problem and express the optimal hyperplane $f(x) = \langle w^*, x\rangle + b^*$ with respect to the solution of the dual problem (i.e., the contents in slides 38&39 of Topic 8)

- Consider the directed graph shown on the right in which none of the variables is observed.
  - Show that $a \perp\!\!\!\perp b | \emptyset$
  - Suppose we now observe the variable d. Show that in general $a \not\perp\!\!\!\perp b | d$