Bayesian Network

Julyedu: Johnson

2018/07/31

Bayesian Network

- What is Bayesian Network?

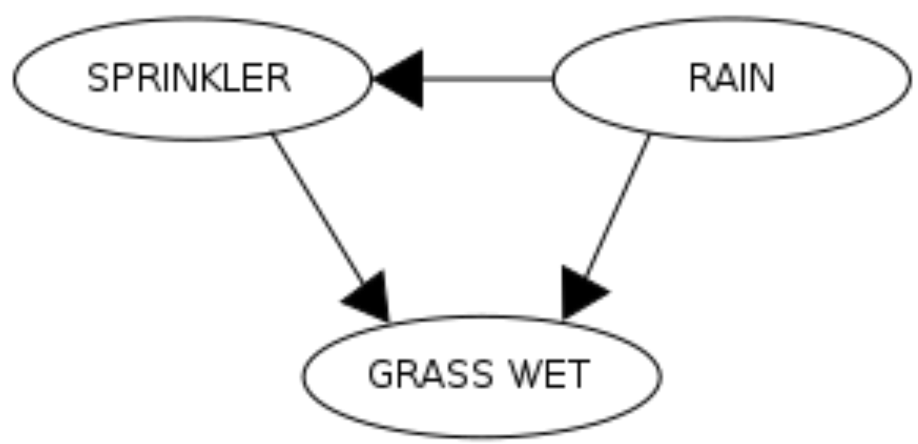
- Examples:
  - Naive Bayes (朴素贝叶斯)
  - Hidden Markov Model (隐马尔可夫模型)
  - Latent Dirichlet Allocation (LDA，主题模型)

- Main Topic:
  - Probability theory + Graph theory
  - Probabilistic graphical model

Examples

- Use directed graph to represent conditional independency:

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



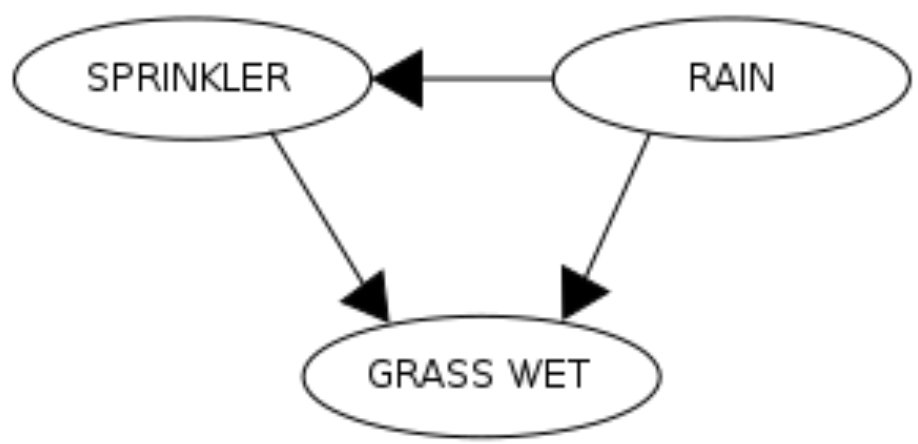
RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Examples

- Use directed graph to represent conditional independency:

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

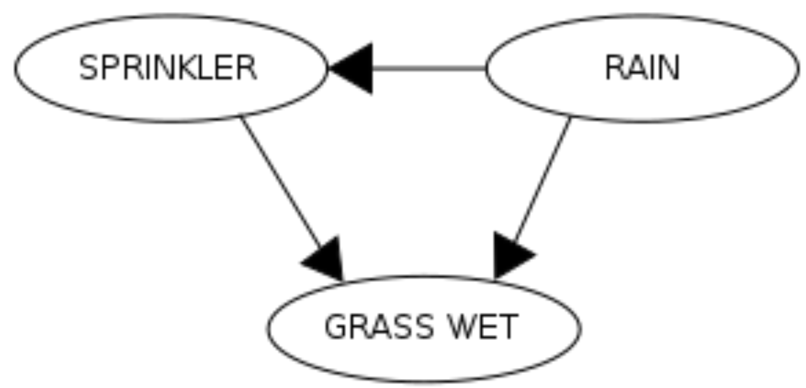
		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

- Each node corresponds to a random variable.
- Each edge corresponds to a conditional dependency between variables.

Definition

- A **Bayesian network** is specified by a directed acyclic graph:
  - G = (V, E) where:
    - One node i for each random variable V_i
    - One conditional probability distribution (CPD) per node: $\Pr(V_i \mid V_{pa(i)})$

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Definition

- A **Bayesian network** is specified by a directed acyclic graph:
  - G = (V, E) where:
    - One node i for each random variable V_i
    - One conditional probability distribution (CPD) per node: $\Pr(V_i \mid V_{pa(i)})$
- Graph structure specifies the factorization of the joint distribution:

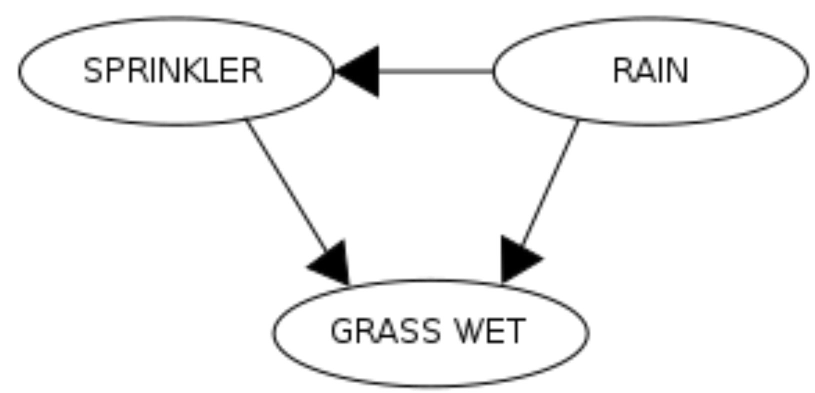
$$\Pr(V_1, \ldots, V_n) = \prod_{i \in [n]} \Pr(V_i \mid V_{pa(i)})$$

Definition

- Graph structure specifies the factorization of the joint distribution:

$$\Pr(V_1, \ldots, V_n) = \prod_{i \in [n]} \Pr(V_i \mid V_{pa(i)})$$

	SPRINKLER	
RAIN	T	F
F	0.4	0.6
T	0.01	0.99



RAIN	
T	F
0.2	0.8

		GRASS WET	
SPRINKLER	RAIN	T	F
F	F	0.0	1.0
F	T	0.8	0.2
T	F	0.9	0.1
T	T	0.99	0.01

Definition

- Graph structure specifies the factorization of the joint distribution:

$$\Pr(V_1, \ldots, V_n) = \prod_{i \in [n]} \Pr(V_i \mid V_{pa(i)})$$

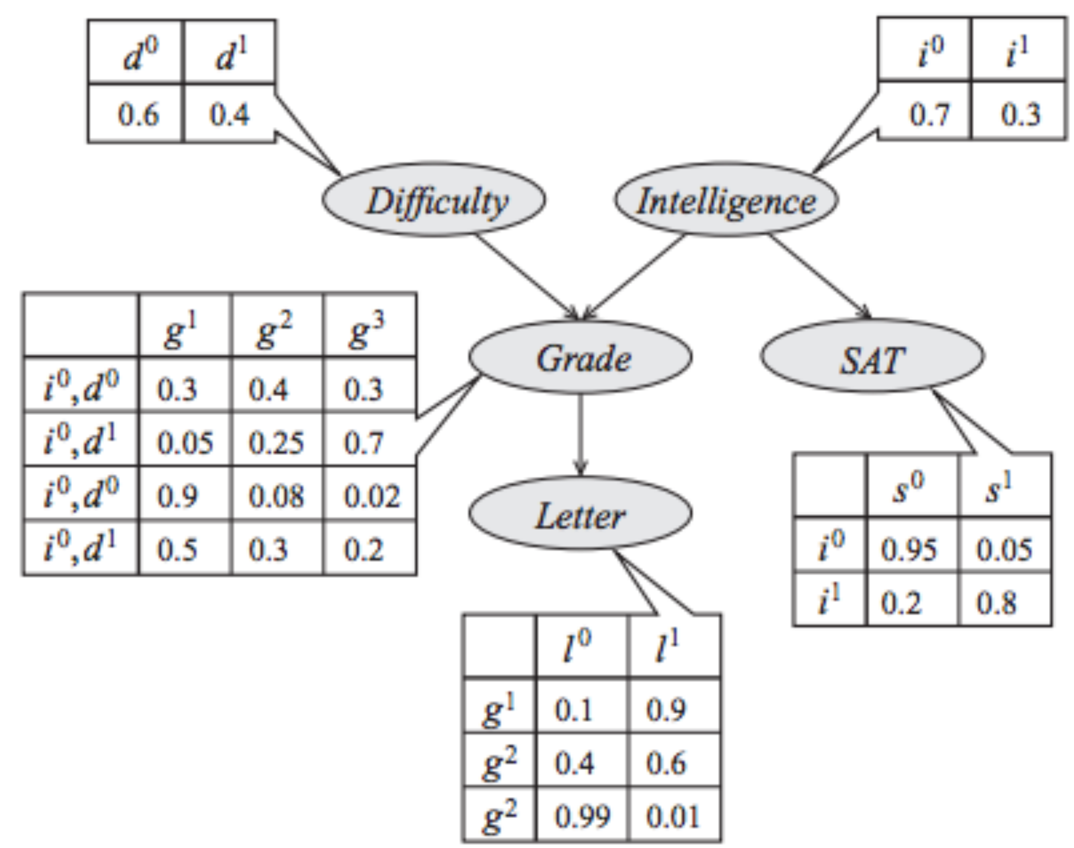
- What's the advantages of Bayesian networks?
  - Compact representation of a probability distribution
  - Clear probabilistic semantics
  - Direct graph can be used to represent causal relationship

More Examples

- Graph structure specifies the factorization of the joint distribution:

$$\Pr(V_1, \ldots, V_n) = \prod_{i \in [n]} \Pr(V_i \mid V_{pa(i)})$$

- What's the joint distribution?

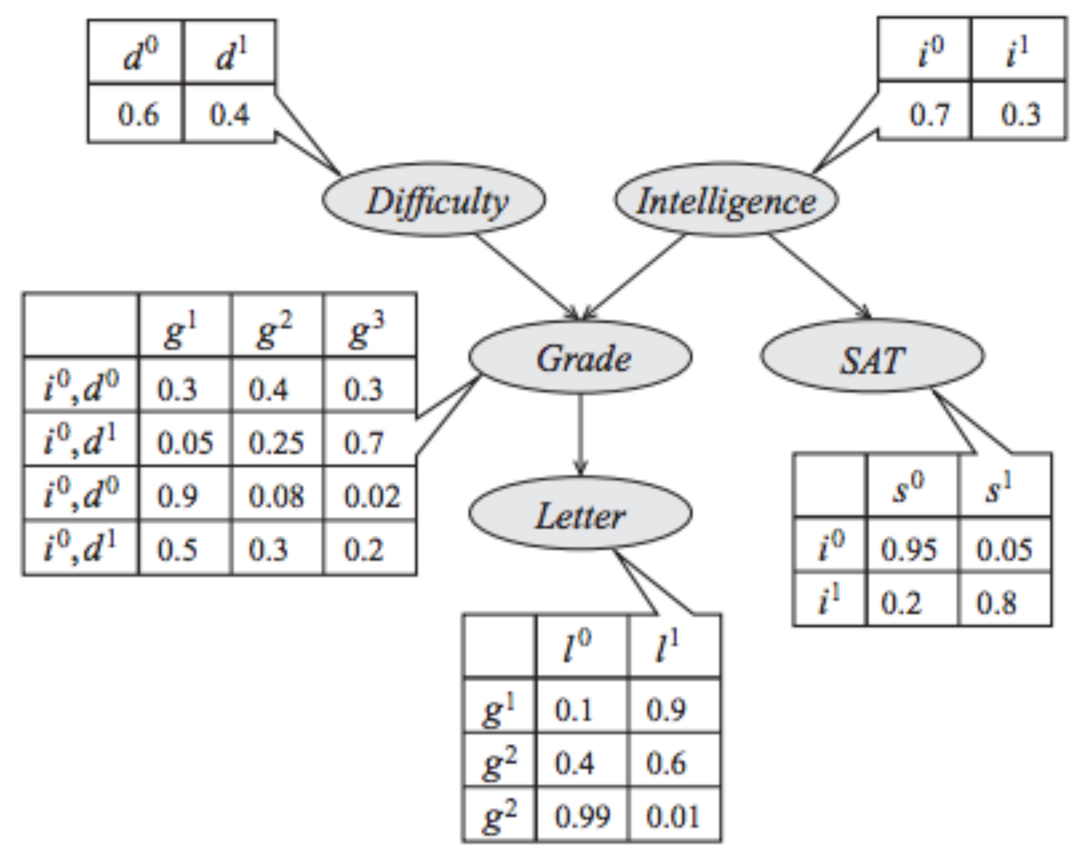


More Examples

- Graph structure specifies the factorization of the joint distribution:

$$\Pr(V_1, \ldots, V_n) = \prod_{i \in [n]} \Pr(V_i \mid V_{pa(i)})$$

- What's the joint distribution?



$$\Pr(d, i, g, s, l) = \Pr(d) \cdot \Pr(i) \cdot \Pr(g \mid d, i) \cdot \Pr(s \mid i) \cdot \Pr(l \mid g)$$

Conditional Independence

- Graph structure implies conditional independency
- If two variables are conditionally independent, graph has no edge between them:

$D \perp I$

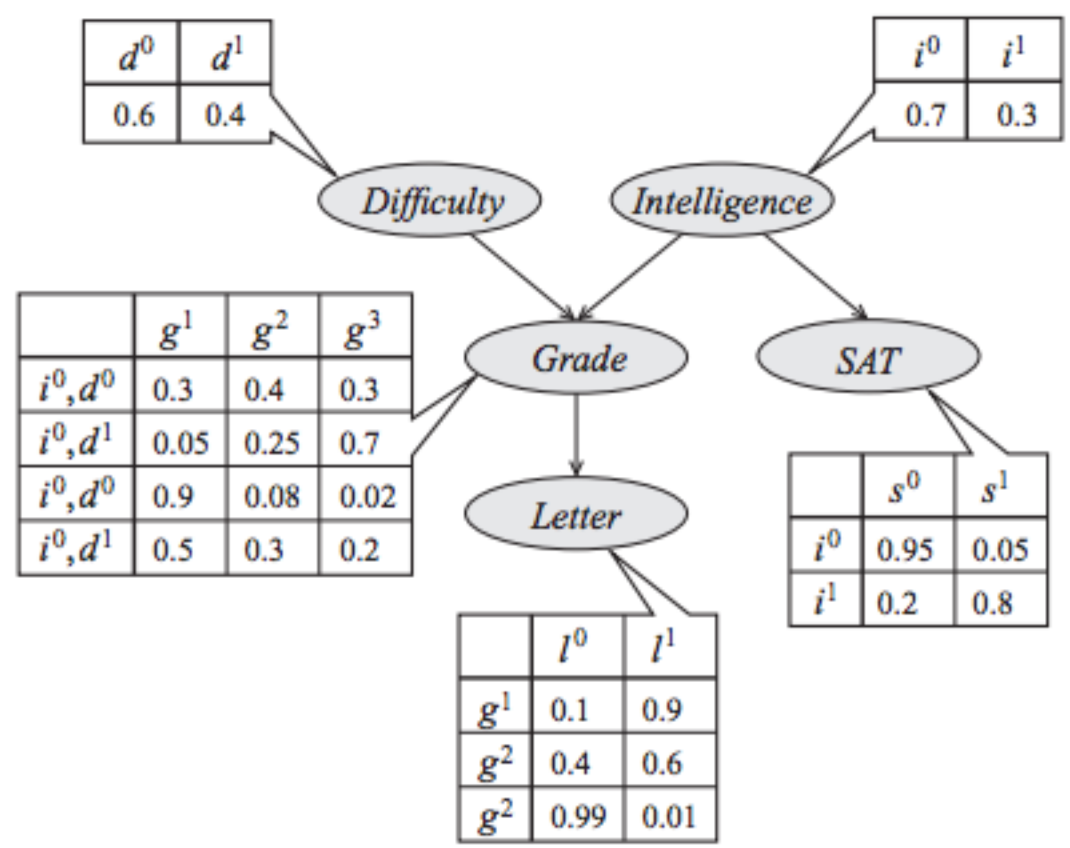
$G \perp S \mid I$

$D \perp L \mid G$

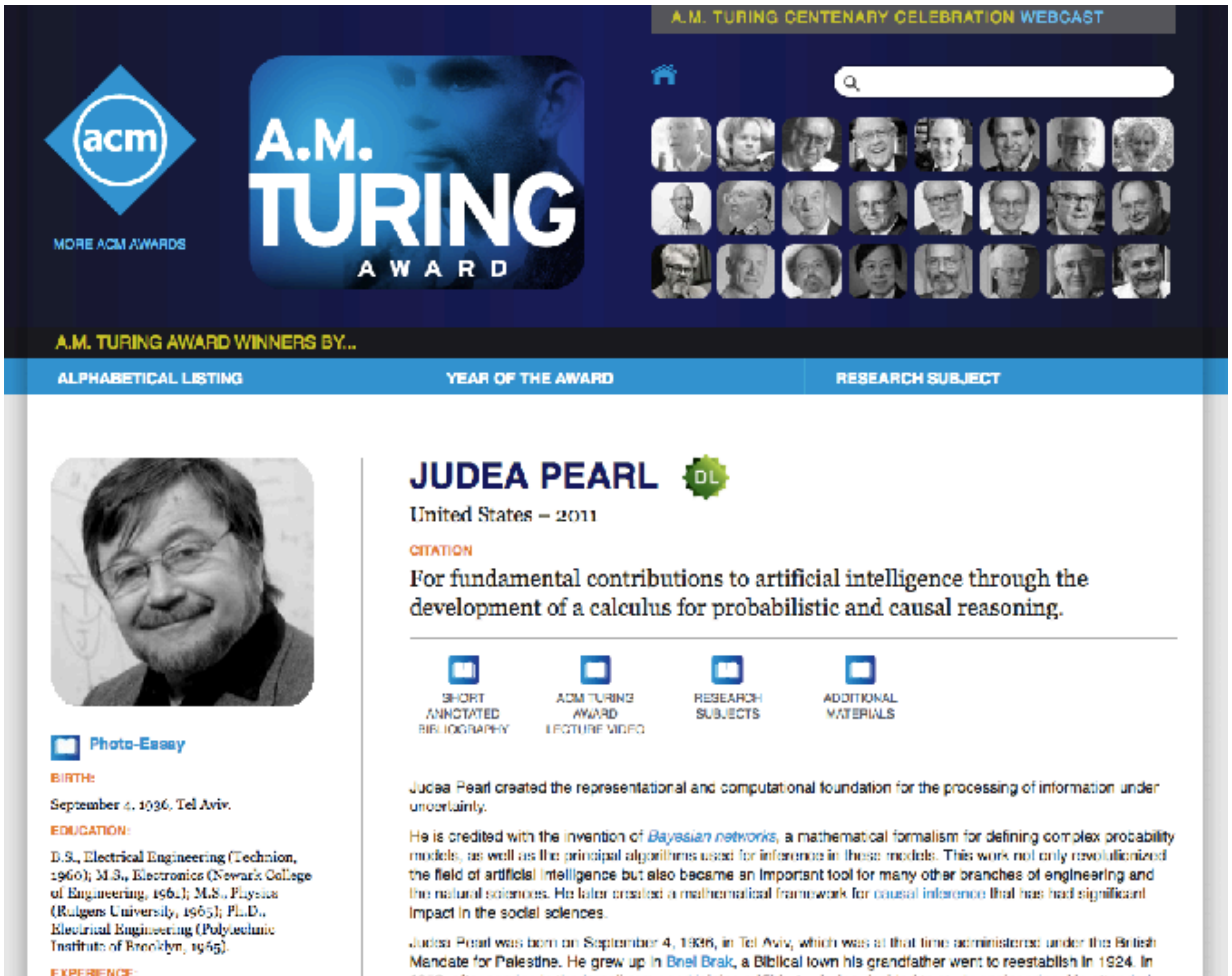
$L \perp S \mid G$

$L \perp S \mid I$

$D \perp S$

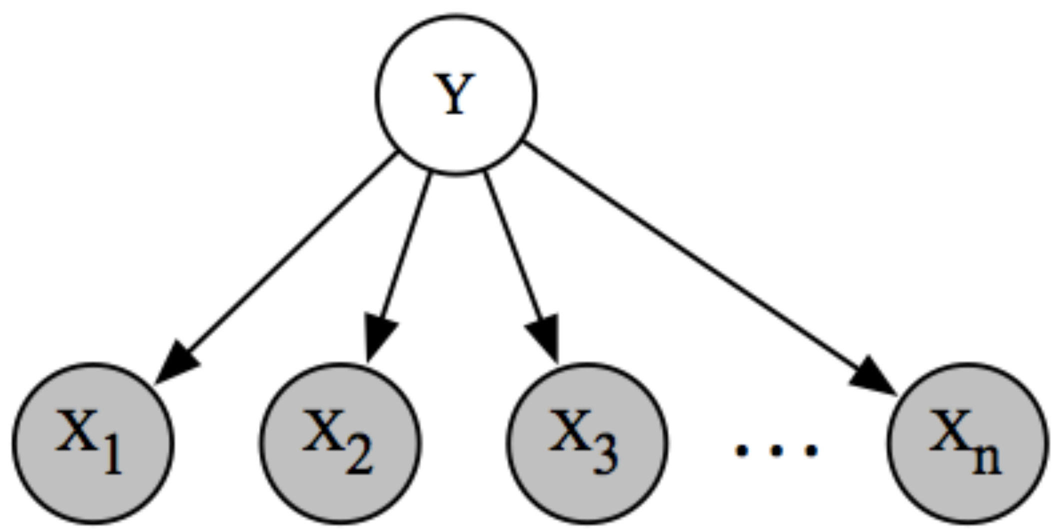


Probabilistic Graphical Models



More Examples

- Naive Bayes as a Bayesian network:

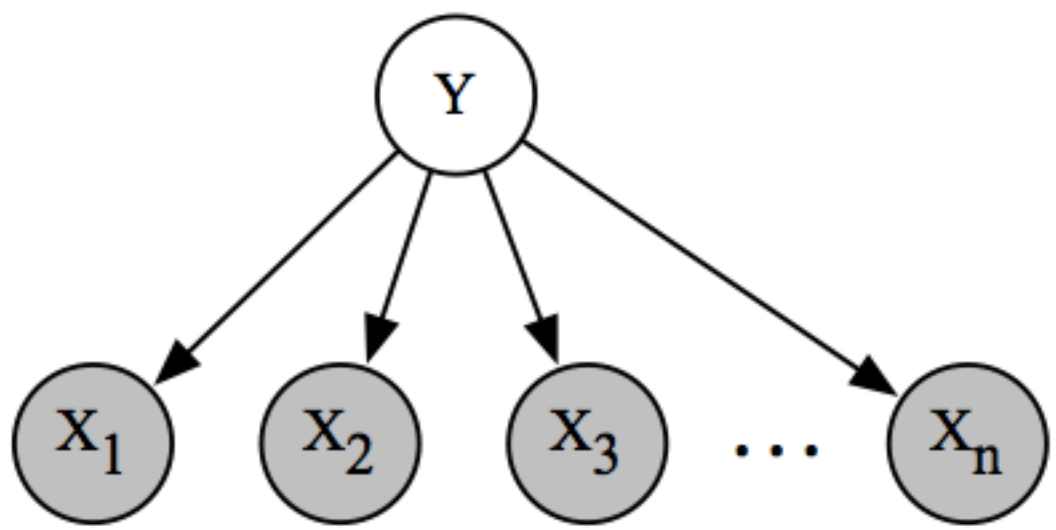


$$\Pr(Y, X_1, \ldots, X_n) = \Pr(Y) \prod_{i \in [n]} \Pr(X_i \mid Y)$$

- Shadowed nodes are **observed**
- White nodes are **hidden**

More Examples

- Naive Bayes as a Bayesian network:

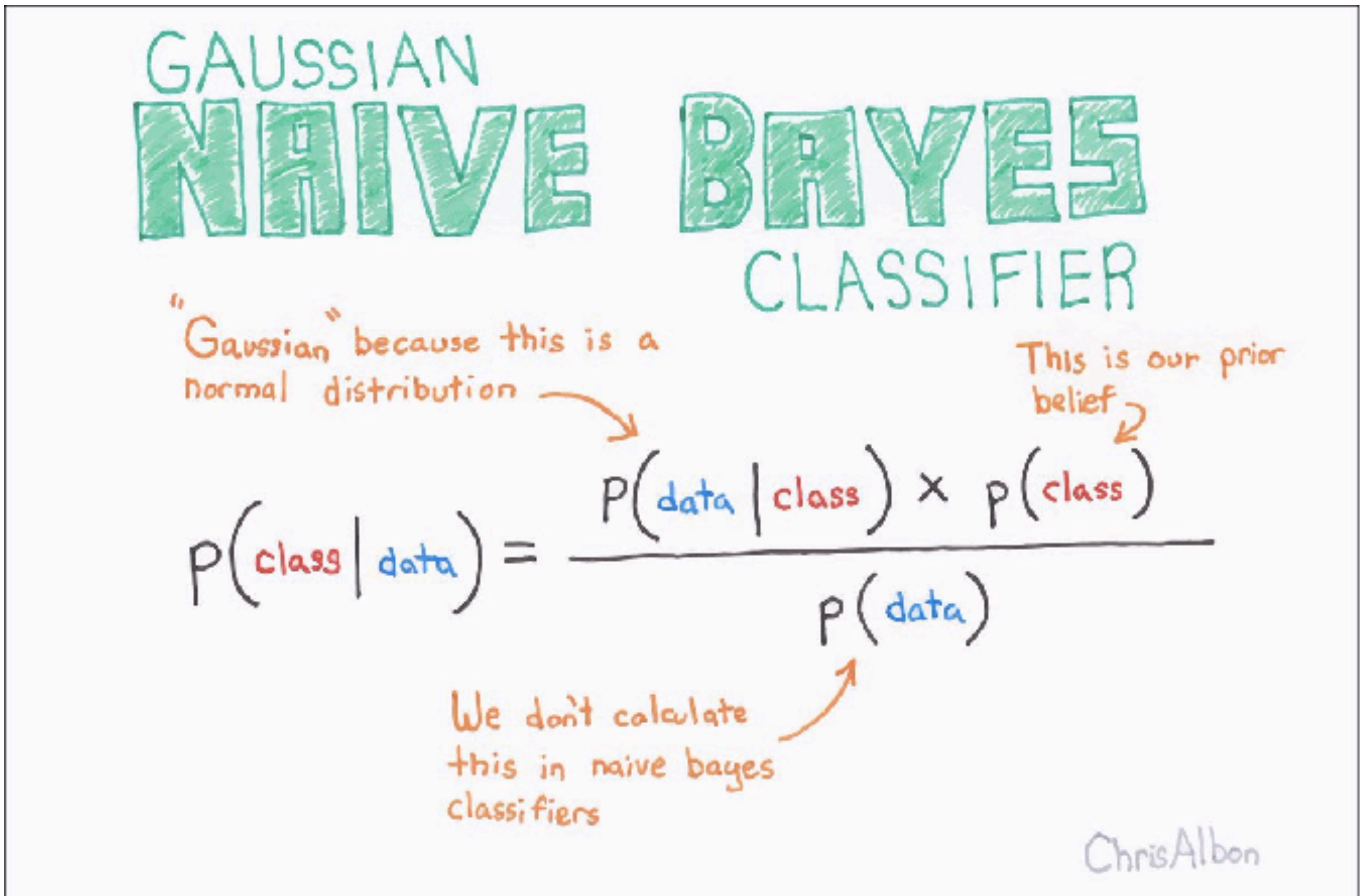


$$\Pr(Y, X_1, \ldots, X_n) = \Pr(Y) \prod_{i \in [n]} \Pr(X_i \mid Y)$$

- Problem: given **observed**, infer **hidden** ?

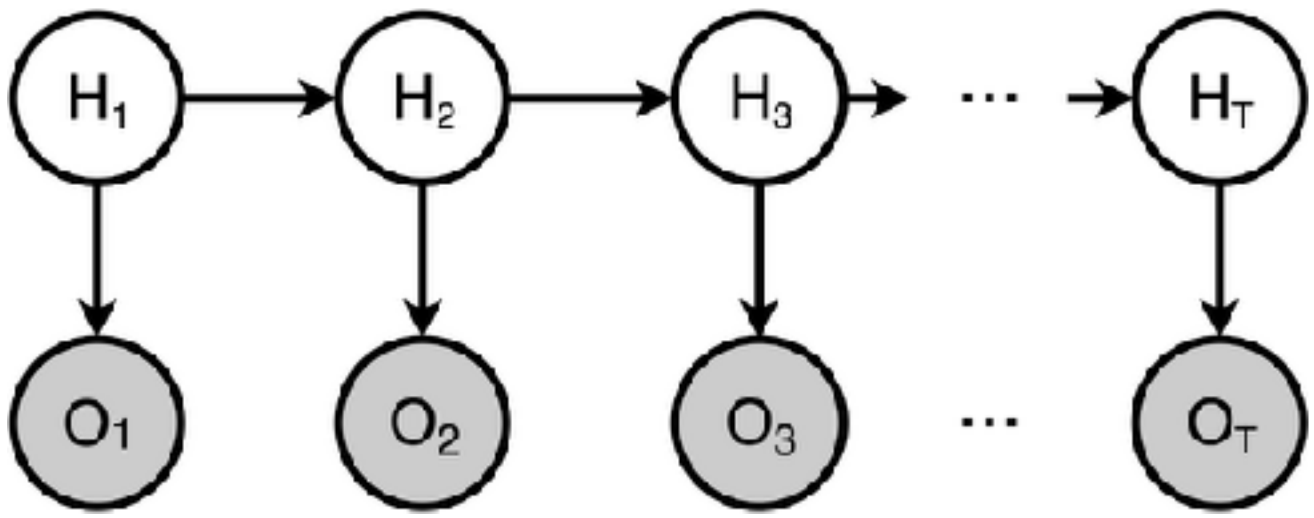
More Examples

- Problem: given **observed**, infer **hidden** ?



More Examples

- Hidden Markov model as a Bayesian network:

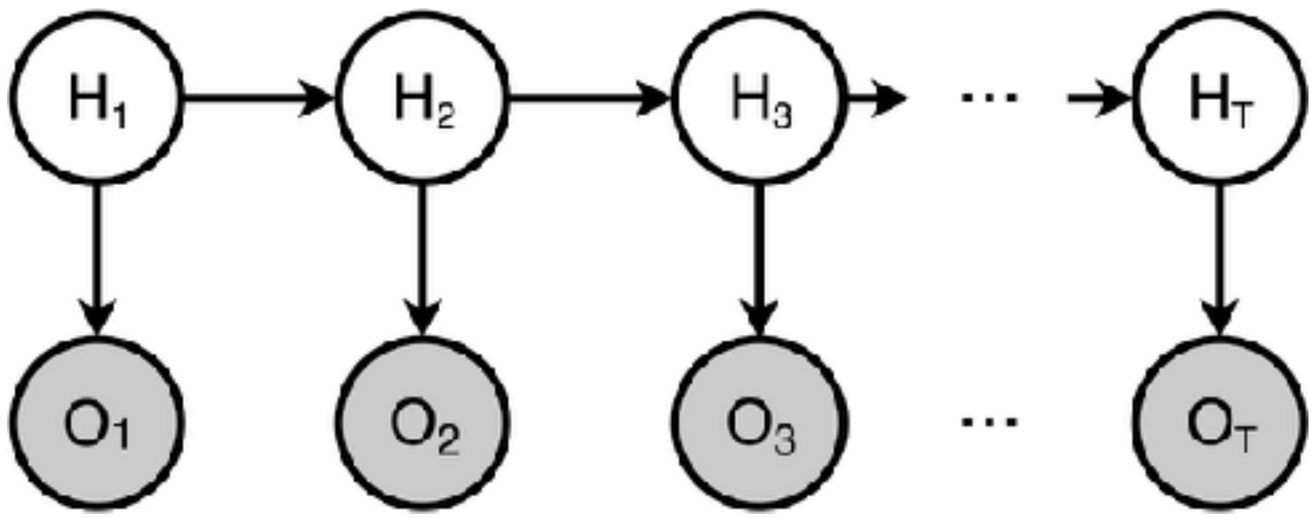


$$\Pr(H_1, \ldots, H_T, O_1, \ldots, O_T) = \Pr(H_1) \prod_{i \in [T]} \Pr(O_i \mid H_i) \prod_{i \in [T-1]} \Pr(H_{i+1} \mid H_i)$$

- Shadowed nodes are **observed**
- White nodes are **hidden**

More Examples

- Hidden Markov model as a Bayesian network:

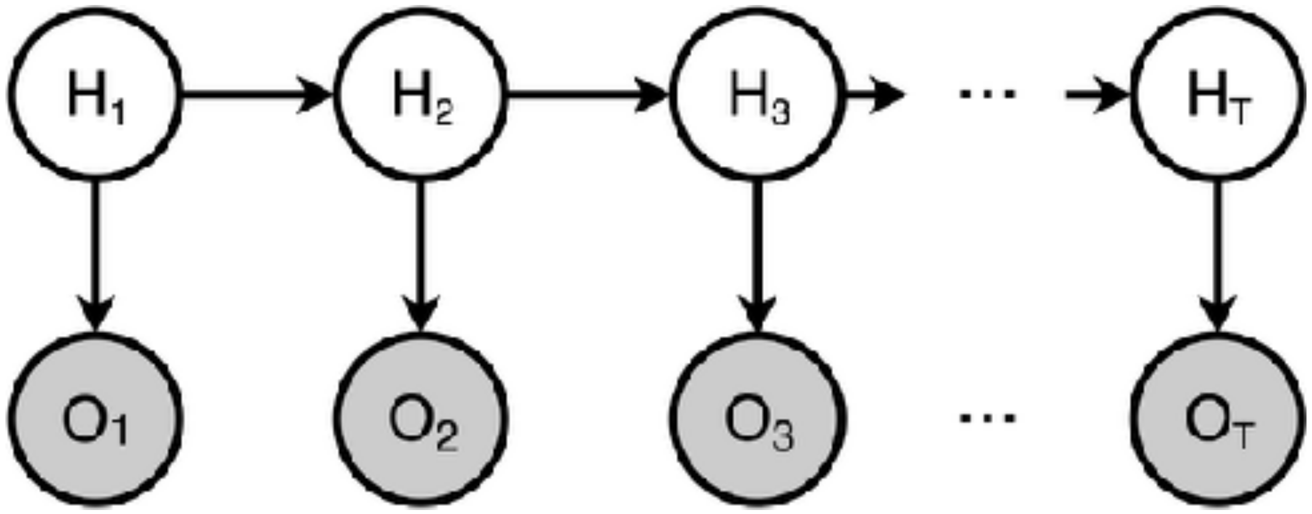


$$\Pr(H_1, \ldots, H_T, O_1, \ldots, O_T) = \Pr(H_1) \prod_{i \in [T]} \Pr(O_i \mid H_i) \prod_{i \in [T-1]} \Pr(H_{i+1} \mid H_i)$$

- Applications:
  - POS Tagging (词性标注)
  - Speech recognition (语音识别)
  - Word segmentation (中文分词)
  - ……

More Examples

- Hidden Markov model as a Bayesian network:



$$\Pr(H_1, \ldots, H_T, O_1, \ldots, O_T) = \Pr(H_1) \prod_{i \in [T]} \Pr(O_i \mid H_i) \prod_{i \in [T-1]} \Pr(H_{i+1} \mid H_i)$$

- Algorithms:
  - Forward-Backward algorithm (inference)
  - Expectation maximization algorithm (learning)

Conditional Independence

- Review:
  - Independence:

$$\Pr(A, B) = \Pr(A) \cdot \Pr(B)$$

$$\Pr(A \mid B) = \Pr(A)$$

$$\Pr(B \mid A) = \Pr(B)$$

How to show the above three formulas are equivalent?

Conditional Independence

- Review:
  - Conditional independence:

$$\Pr(A, B \mid C) = \Pr(A \mid C) \cdot \Pr(B \mid C)$$

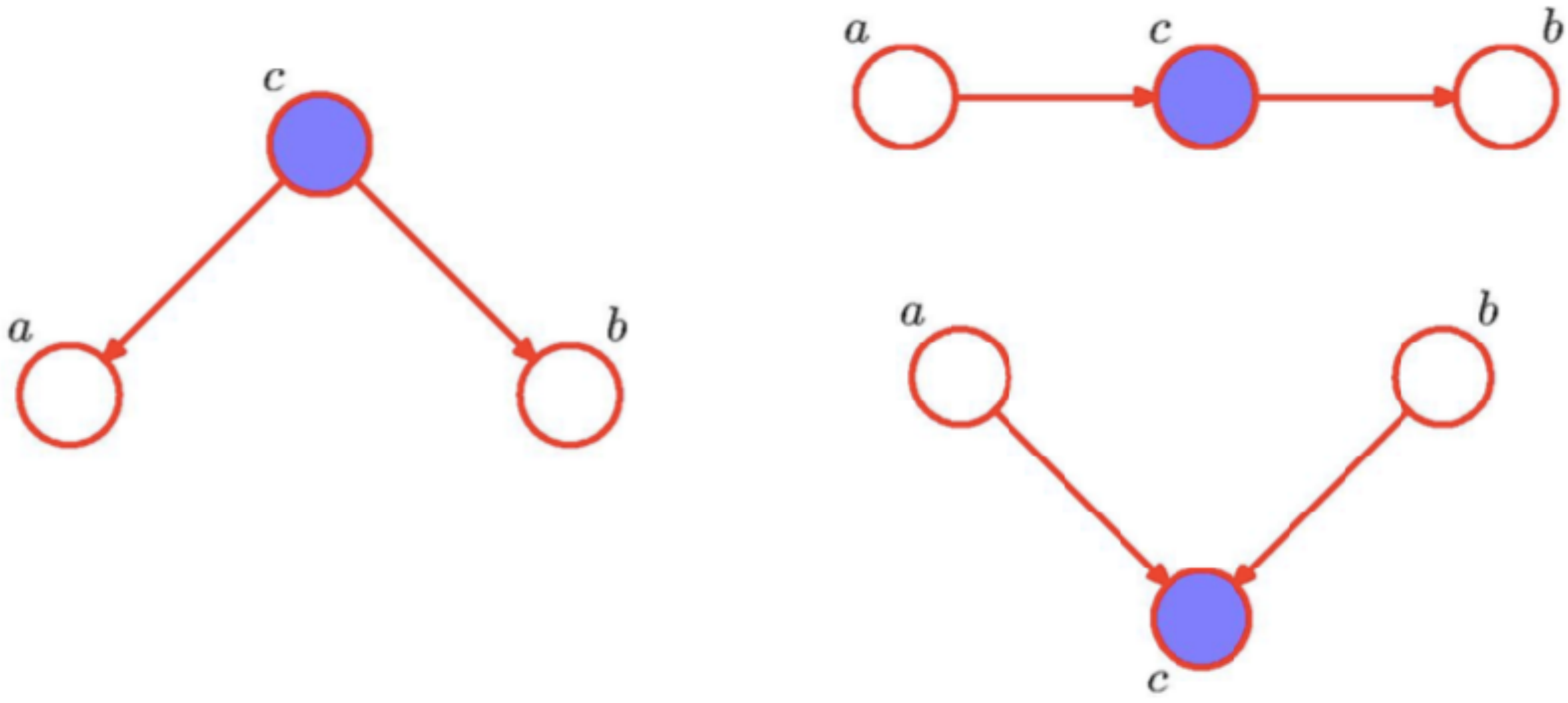
$$\Pr(A \mid B, C) = \Pr(A \mid C)$$

$$\Pr(B \mid A, C) = \Pr(B \mid C)$$

How to show the above three formulas are equivalent?

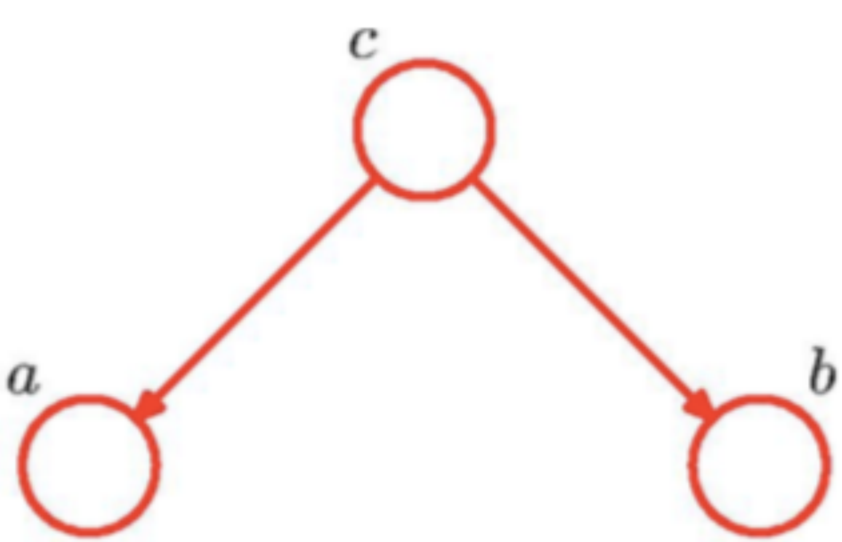
Conditional Independence

- **d-separation**:
  - A method to quickly judge conditional independence given the graph
  - 3 difference cases

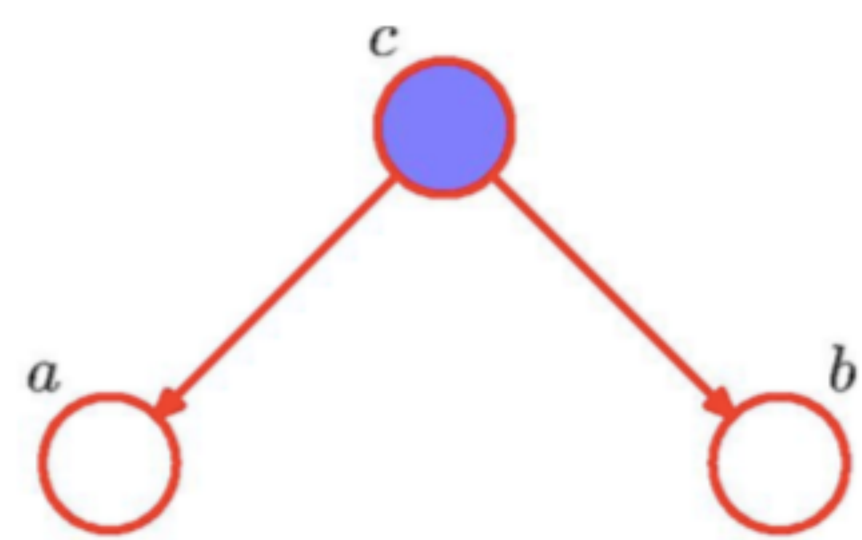


Conditional Independence

- **Tail-to-Tail:**



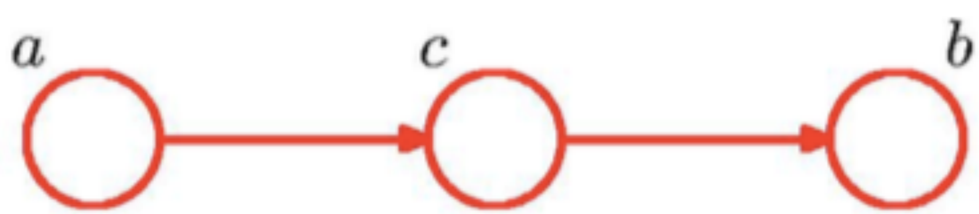
$$a \not\perp b$$

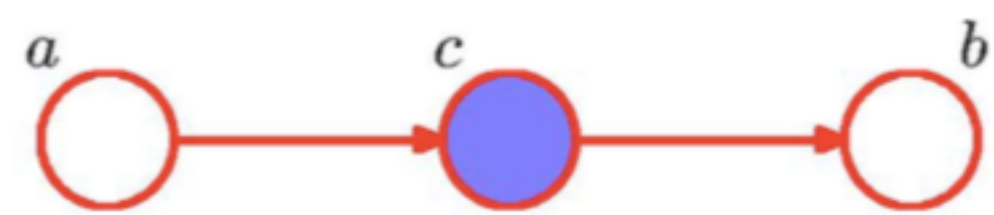


$$a \perp b \mid c$$

Conditional Independence

- **Head-to-Tail:**



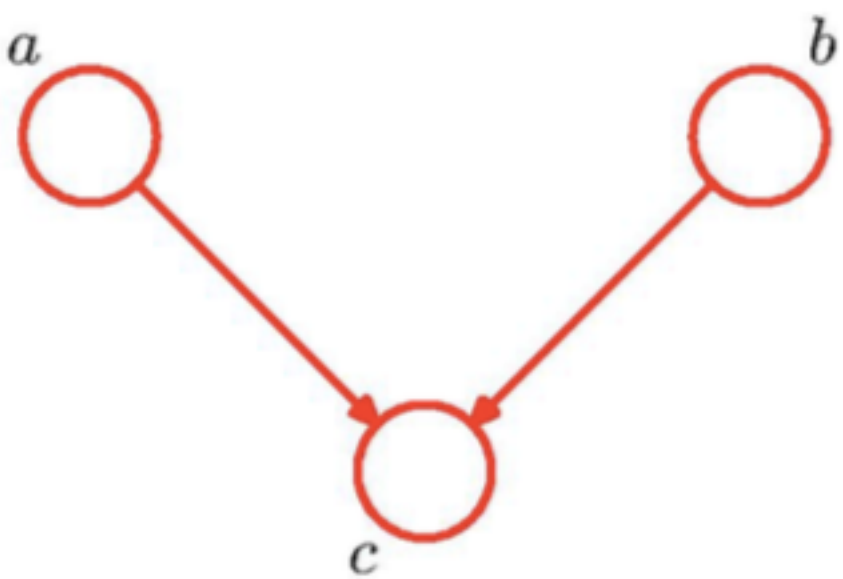


$$a \not\perp b$$

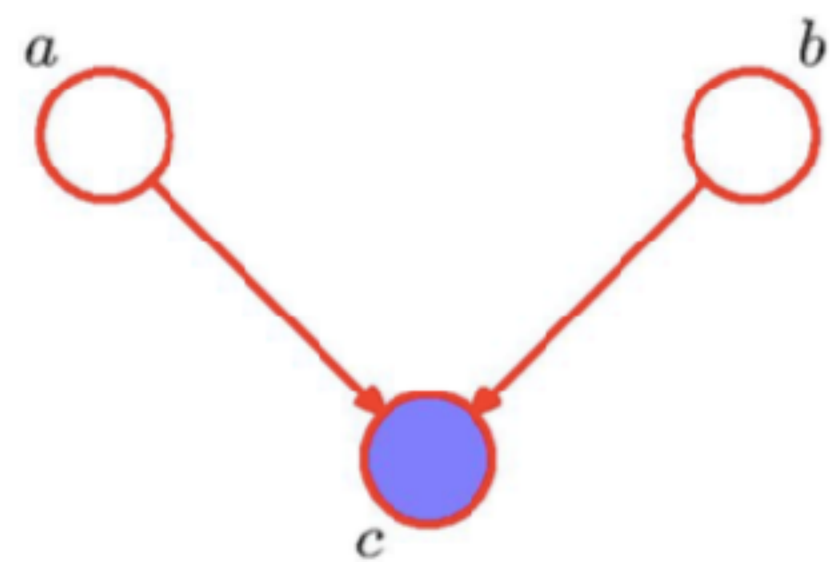
$$a \perp b \mid c$$

Conditional Independence

- **Head-to-Head:**



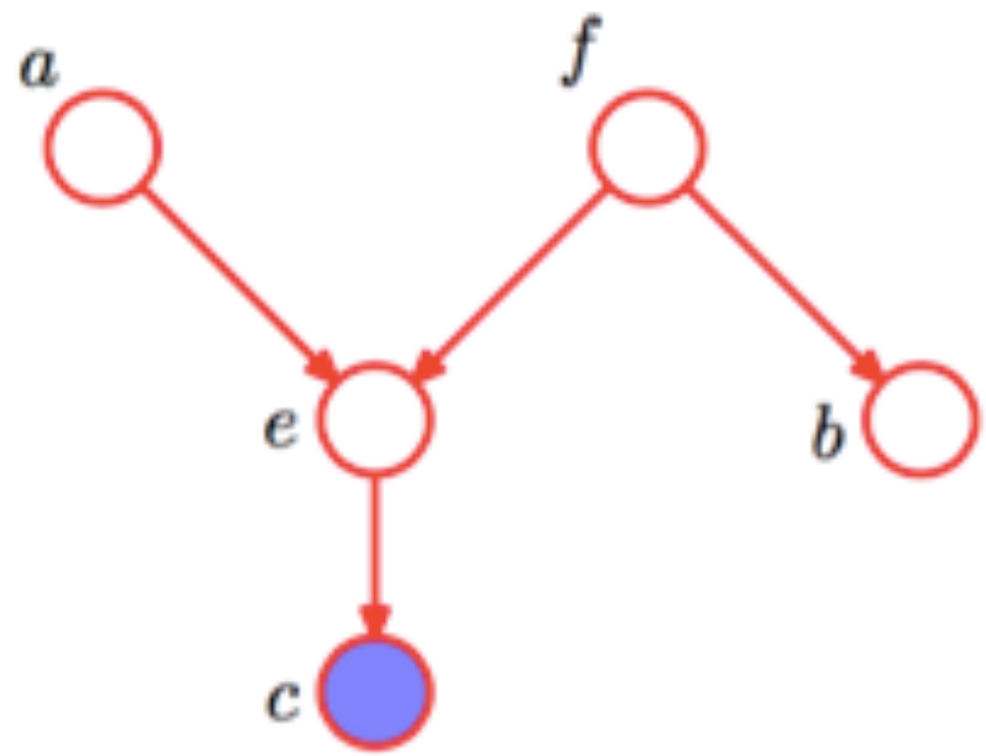
$$a \perp b$$



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

Conditional Independence

- Examples:



(a)

- Is:

  $a \perp b \mid c?$

  $a \perp f \mid c?$

  $e \perp b?$

Thanks and Questions