

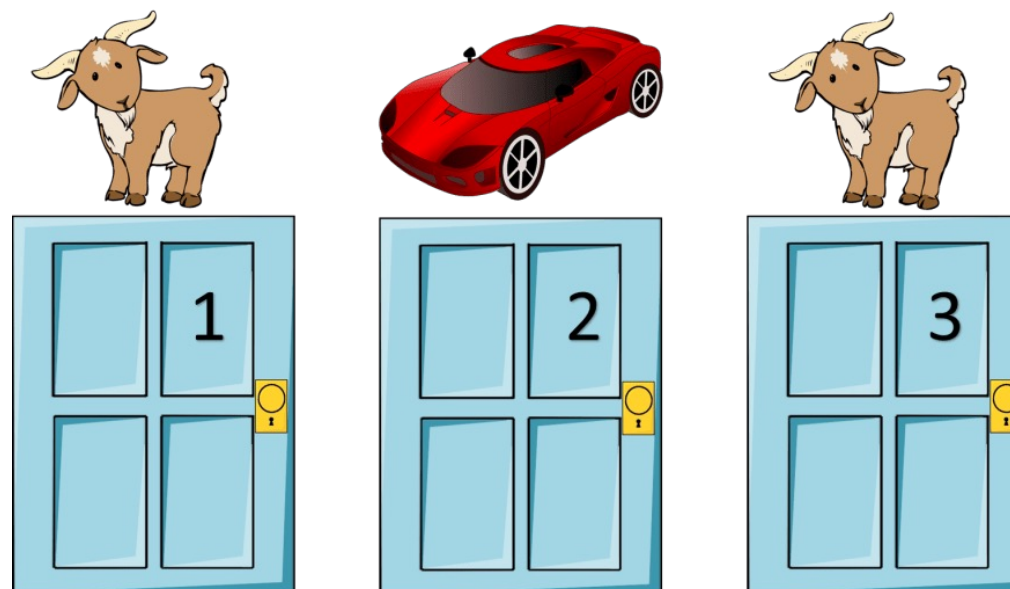
BAYESIAN NETWORKS

Ashish Pujari

Lecture Outline

- Probabilistic Graphical Models (PGMs)
- Probability Review
- Bayesian Networks
- Bayesian Networks Inference

Monty Hall Problem



PROBABILISTIC GRAPHICAL MODELS

Overview

Modeling Complex Systems

- Modeling uncertainty
 - Noisy measurements
 - Incomplete knowledge
 - Partially observed system states
 - Intrinsically stochastic systems
- Missing information
 - Recognizing or reacting to new circumstances
- Explainability
 - Black box models can't explain reasons behind their predictions

Expert Systems

- Computer systems that emulate the decision-making ability of a human expert using *if-then* rules
- Evolution
 - Early 1980s: Rule based expert systems unable handle uncertainty in reasoning.
 - Late 1980s: Bayesian networks - a probabilistic approach
 - Late 1990s: Decline in use of expert systems.

Pathfinder 1
Rule Based System

Pathfinder 2
Naïve Bayes Model

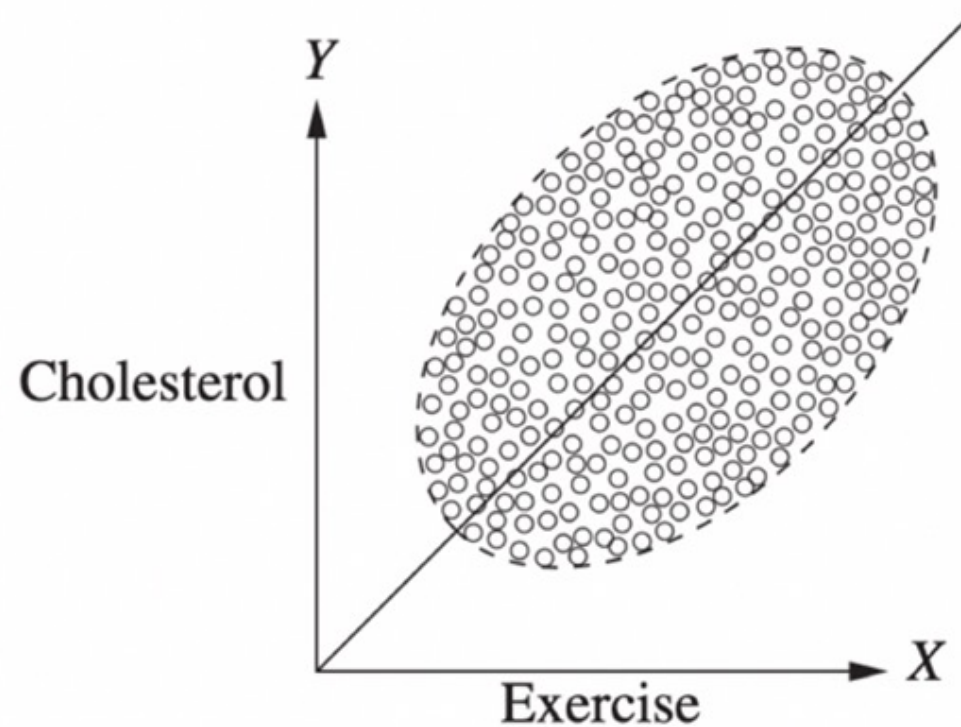
Pathfinder 3
Naïve Bayes with Knowledge Engineering

Pathfinder 4
Bayesian Network

[Pathfinder \(1989\)](#) - Diagnosis of lymph node pathologies.
Heckerman, Nathwani, et al.

Correlation vs Causation

- The logic of association is sometimes insufficient

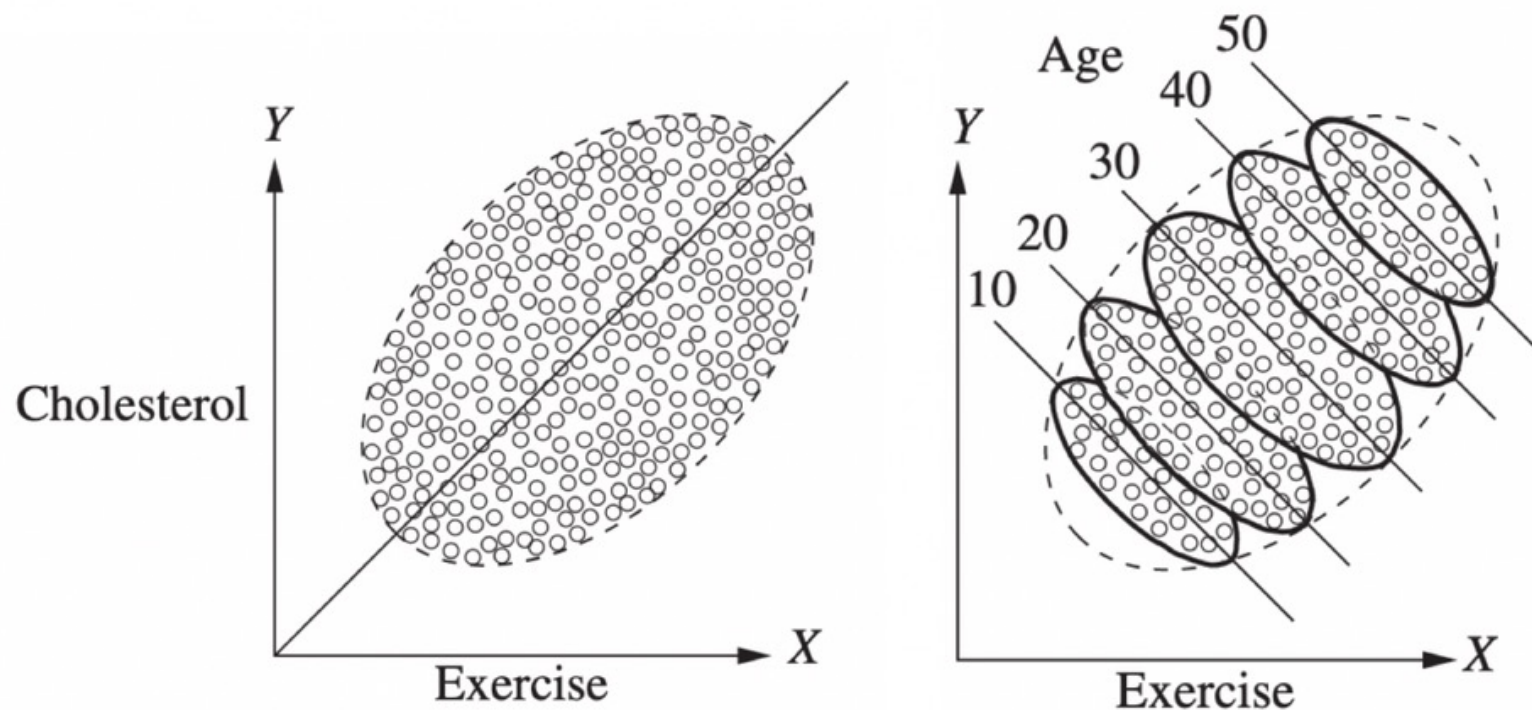


Concrete causal dilemmas with cholesterol

In a study of weekly exercise effects on cholesterol, the population-level results show a positive... [+] (Pearl et al., 2016)

Simpson's Paradox

- Characterizes a reversal or cancellation of a global association between two variables, when conditioned upon a third.

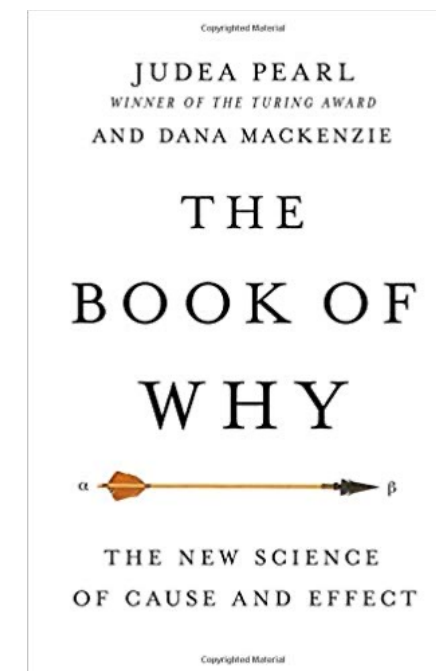


Association between exercise and cholesterol is reversed when conditioned on age.

In a study of weekly exercise effects on cholesterol, the population-level results show a positive... [+] (Pearl et al., 2016)

Causal Reasoning

- Causal questions in research:
 - How effective is a given treatment in preventing a disease ?
 - How likely is it that a component will fail, given current state of the system?
 - What is the relationship between leisure time and mental health ?
 - What happens to the global economy if there is a pandemic ?

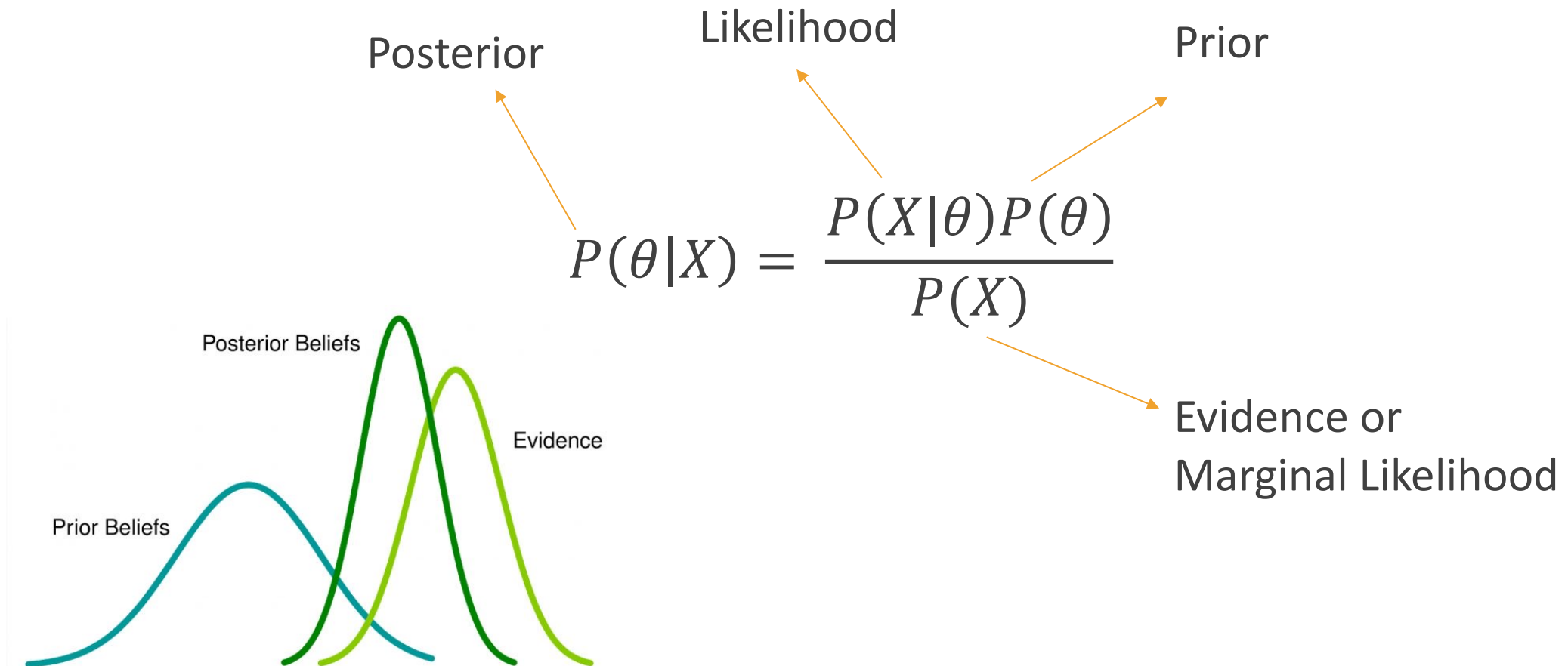


“Causal reasoning is an indispensable component of human thought that should be formalized and algorithmicized toward achieving human-level machine intelligence.” - Judea Pearl

Probabilistic Models

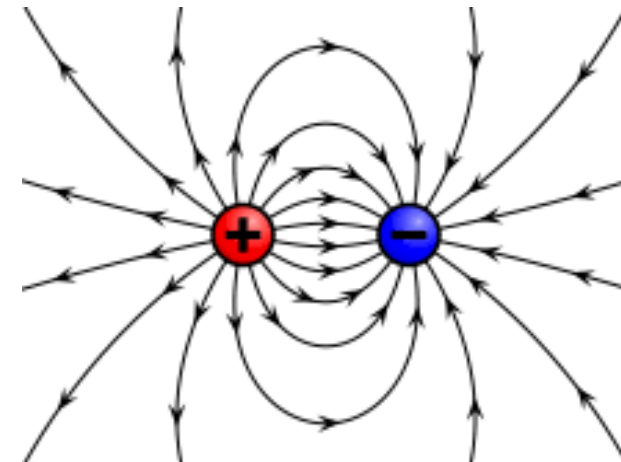
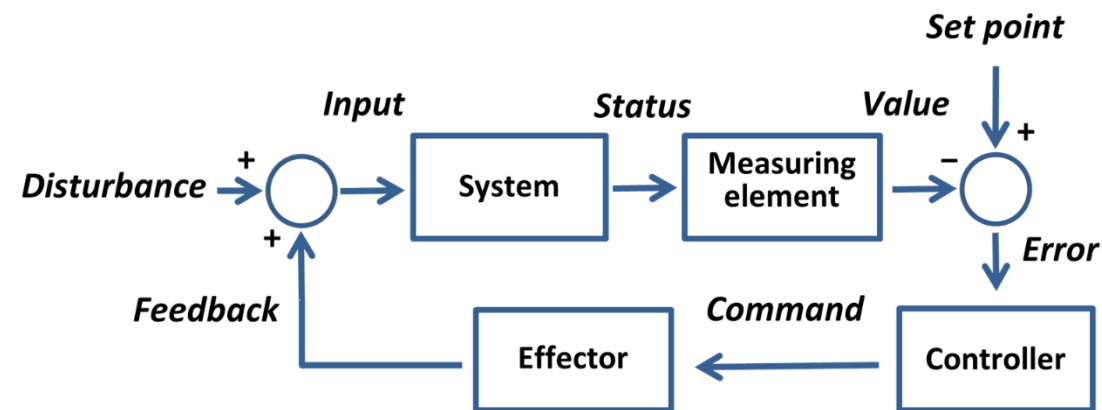
- Mathematical frameworks used to represent uncertainty in data and make predictions based on probabilistic inference.
- Approach
 - Express relationships between variables using probability distributions, allowing for the quantification of uncertainty and the incorporation of prior knowledge.
- Examples
 - Bayesian Networks (BN), Hidden Markov Models (HMM), Markov Random Fields (MRF), Conditional Random Fields (CRF), Gaussian Processes
- Applications
 - Statistical physics, quantum mechanics, and theoretical computer science.

Bayesian Inference



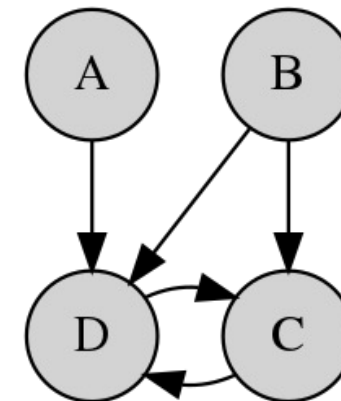
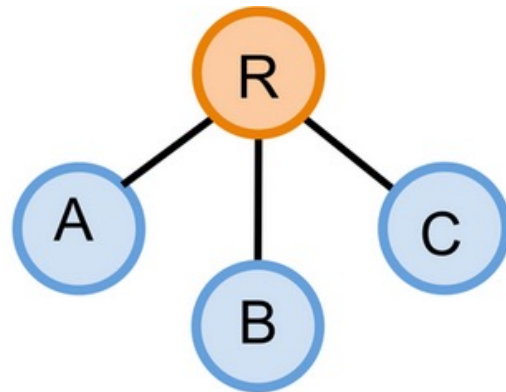
Graphical Representations

- Graphical representations can be very powerful in understanding and predicting behavior of complex systems in engineering, physics and other fields.



Probabilistic Graphical Models (PGMs)

- Provide a graphical representation to encode the complex relationship between a set of random variables.
- Structure
 - The nodes correspond to the random variables
 - Edges (or arcs) correspond to direct probabilistic interactions between them.



Applications

- Medicine
 - medical diagnosis, drug discovery
- Forensics
 - victim identification, kinship analysis
- Genomics
 - Gene Regulatory Network - governs the expression levels of mRNA and protein
- Biomonitoring
 - quantify the concentration of chemicals in blood and tissue
- Autonomous Systems
 - robot localization and mapping
- Natural Language Processing
 - Document classification, speech recognition
- Information Retrieval
 - semantic search, search intent, recommendation systems
- Image Processing
 - Image classification, segmentation, generation
- Information Security
 - spam filtering, intrusion detection

PROBABILITY REVIEW

Marginal Probability Distribution

- Marginal distribution represents our prior knowledge about an event

$$P(x) = \sum_y P(x, y)$$

- Example:

$$P(\textit{Intelligence})$$

- Our prior belief about students' intelligence before learning anything else about a particular student.

Conditional Probability Distribution

- Probability of one event occurring with some relationship to one or more other events.
- $P(y|x)$ is the probability distribution of Y when X is known to be a particular value
- Example:

$$P(\textit{Intelligence} \mid \textit{Grade} = A)$$

- Conditional distribution represents our more informed distribution after learning her grade

Joint Probability Distribution

- Chain Rule

$$p(x_1, x_2, \dots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_1, x_2) \cdots p(x_n|x_1, \dots, x_{n-1})$$

- Despite factorization, the JPD still requires a number of values that grows exponentially with the number n of variables (e.g., we need $2^n - 1$ values if all variables are binary)
- Mutual Independence

$$p(x_1, \dots, x_n) = p(x_1)p(x_2)p(x_3) \cdots p(x_n)$$

Joint Probability Distribution: Example

- Modeling a joint probability distribution over 4 binary variables

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)p(x_4|x_3, x_2, x_1)$$

x_1	x_2	x_3	$p(x_4 x_3, x_2, x_1)$
0	0	0	α_1
0	0	1	α_2
0	1	0	α_3
0	1	1	α_4
1	0	0	α_5
1	0	1	α_6
1	1	0	α_7
1	1	1	α_8

2^{n-1} parameters

Independence

- Two events are mutually exclusive or disjoint if they cannot occur at the same time.
- Two random variables are independent if

$$\forall x, y : P(x, y) = P(x)P(y)$$

- Or, equivalently, If

$$\forall x, y : P(x|y) = P(x) \quad \text{or} \quad \forall x, y : P(y|x) = P(y)$$

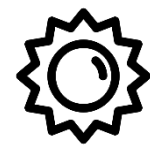
- Notation

$$X \perp Y$$


$$Y \perp X$$


- Example: N fair, independent coin flips
- Note: Independence is a simplifying modeling assumption, and, in many situations, a strong independence assumption can be very limiting

Independence: Example




Temp	Prob
Hot	0.5
Cold	0.5





Weather	Prob
Sun	0.6
Rain	0.4



T	W	$P = P(T)P(W)$
Hot	Sun	$0.5 \times 0.6 = 0.3$
Hot	Rain	$0.5 \times 0.4 = 0.2$
Cold	Sun	$0.5 \times 0.6 = 0.3$
Cold	Rain	$0.5 \times 0.5 = 0.2$

$T \perp W$

T	W	P
Hot	Sun	0.4
Hot	Rain	0.1
Cold	Sun	0.2
Cold	Rain	0.3

$T \not\perp W$

Conditional Independence

- Two random variables X and Y are conditionally independent given another random variable Z if

$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

- Or, equivalently, If and only if

$$\forall x, y, z : P(x|z, y) = P(x|z)$$

or

$$\forall x, y, z : P(y|z, x) = P(y|z)$$

i.e., whenever $Z = z$, the information $Y = y$ does not influence the probability of x

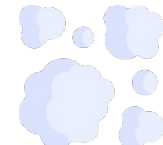
Conditional Independence

- Absolute independence is very rare. Conditional independence is a robust way to represent and model uncertain environments

$$X \perp Y \mid Z$$

$$Y \perp X \mid Z$$

- E.g.,
 - Alarm going off is conditionally independent of the fire given that there is smoke



$$\textit{Alarm} \perp \textit{Fire} \mid \textit{Smoke}$$

Conditional Independence: Example



$$Traffic \perp Umbrella \mid Rain$$

- Chain Rule Decomposition

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_2, x_1)$$

$$P(Rain, Traffic, Umbrella) = P(Rain) P(Traffic \mid Rain) P(Umbrella \mid Rain, Traffic)$$

- Conditional Independence

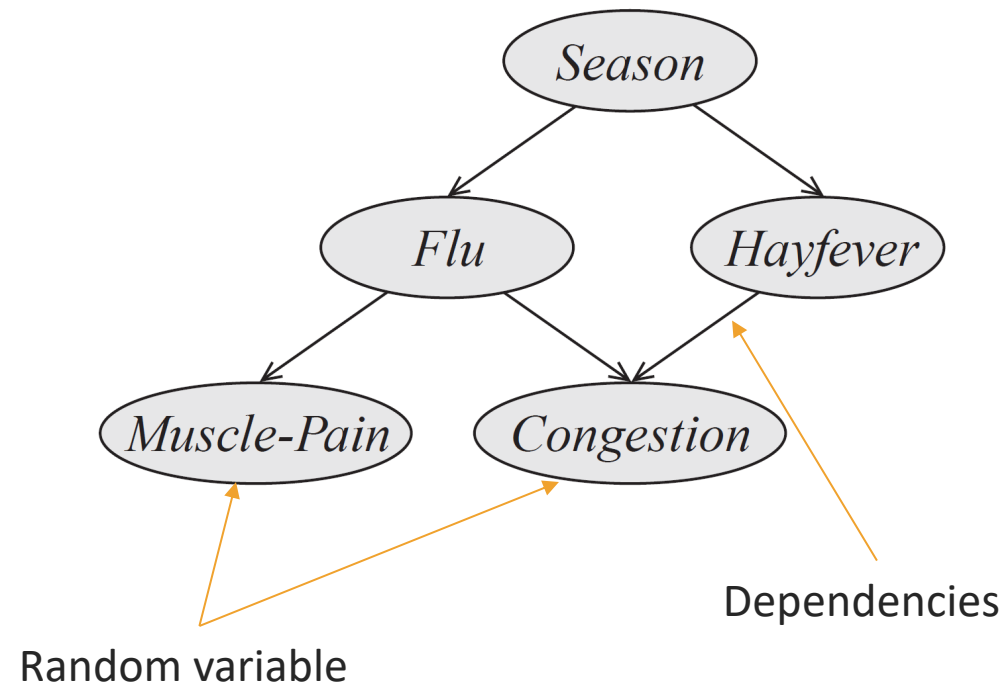
$$P(x_3|x_2, x_1) = P(x_3|x_1)$$

$$P(Rain, Traffic, Umbrella) = P(Rain) P(Traffic \mid Rain) P(Umbrella \mid Rain)$$

BAYESIAN NETWORKS

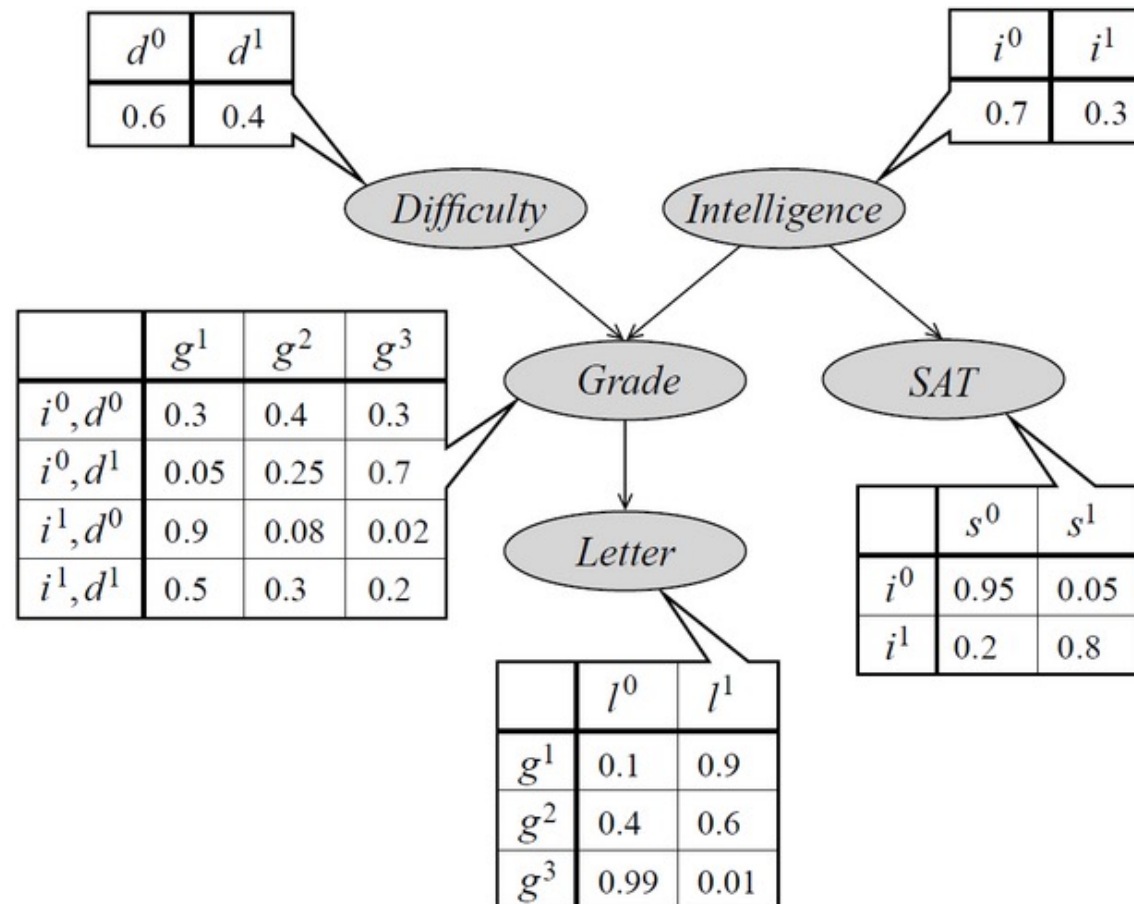
Representation, Inference

Bayesian Networks (Bayes Nets)

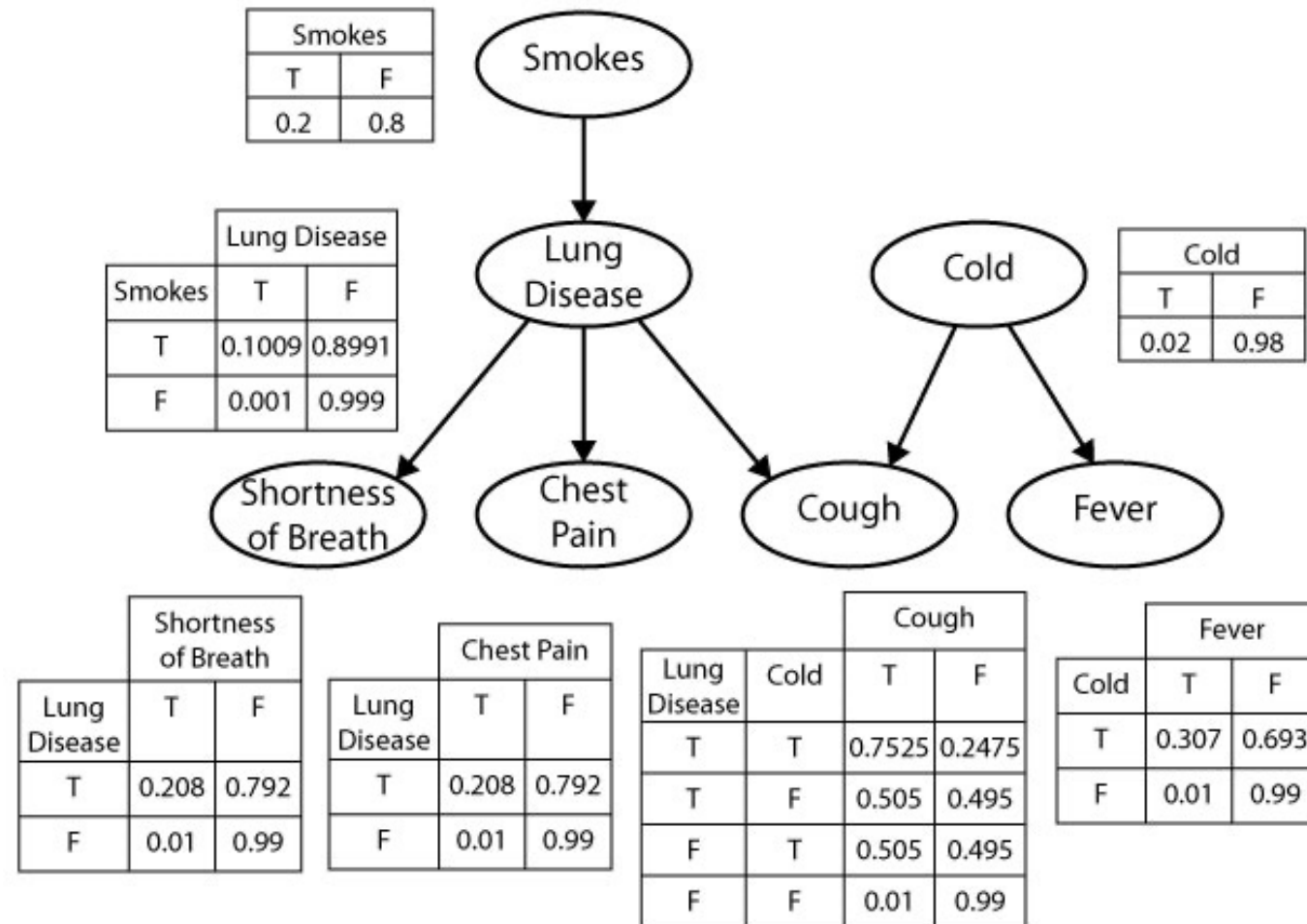


- A directed acyclic graph $G(V, E)$
- Each node $i \in V$ corresponds to a random variable X_i
- X_i has a finite set of mutually exclusive states $\{x_{i_1}, x_{i_2}, \dots, x_{i_n}\}$
- $pa(i)$ denotes the set of parents of node i in the graph
- to each node $i \in V$ corresponds a conditional probability table $P(X_i | (X_j)_{j \in pa(i)})$
- the DAG implies conditional independence relations between $(X_i)_{i \in V}$

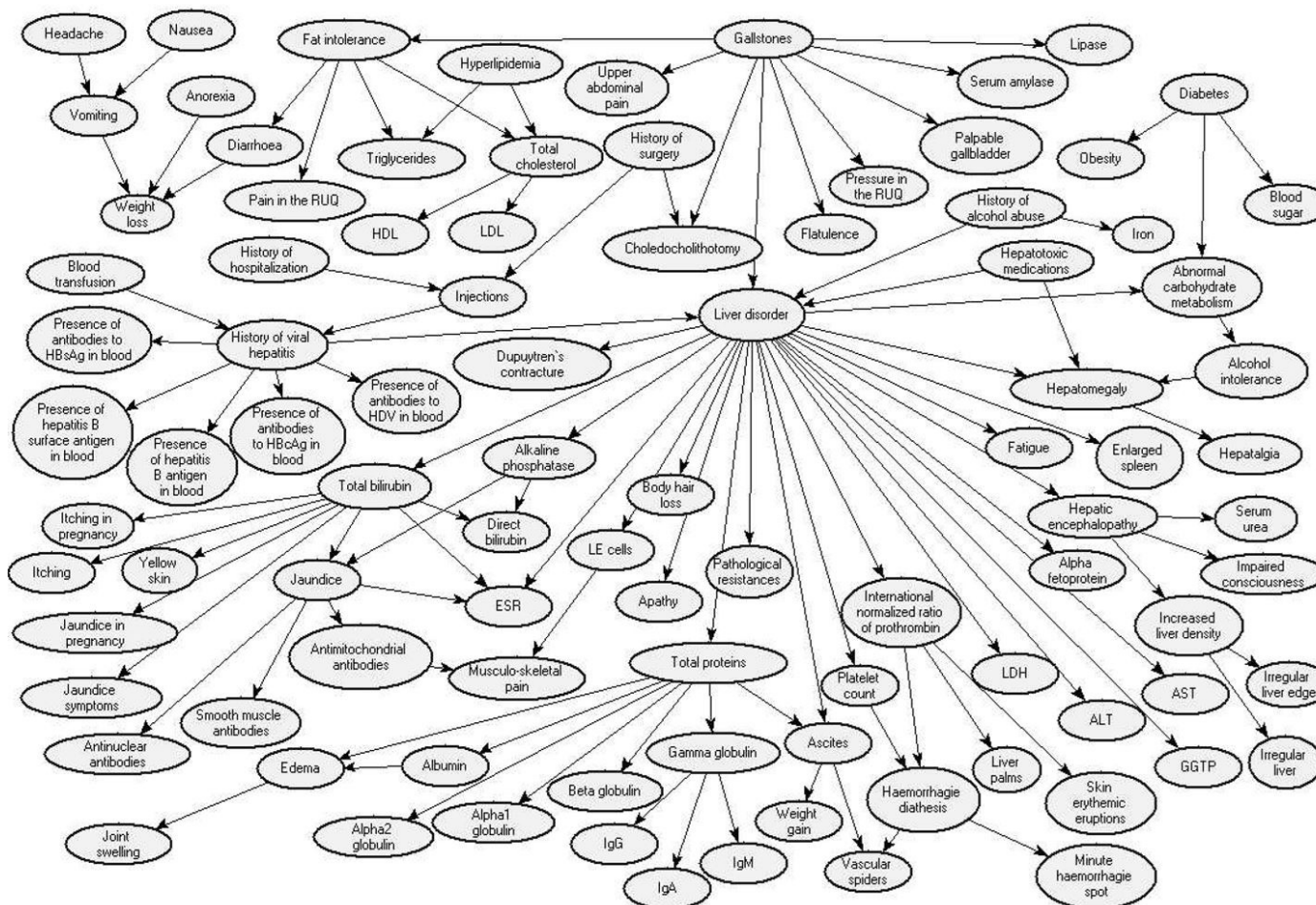
Bayes Nets: Example - Student Net



Bayes Nets: Example - Lung Disease



Bayes Nets: Example - Liver Disorder



The joint probability distribution would be $2^{94}-1$ probability values.

"A Bayesian Network Model for Diagnosis of Liver Disorders" – Agnieszka Onisko, et al.

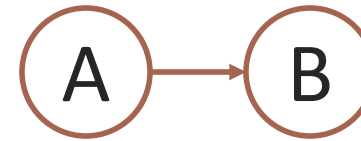
Bayes Nets: Independence

- Independencies
 - Crucial for understanding network behavior
 - Independence properties are important for answering queries
 - Independence can also be exploited to reduce computation of inference
- Important questions about a Bayes Net
 - Are two nodes independent given certain evidence?
 - If yes, can calculate using algebra (really tedious)
 - If no, can prove with a counter example

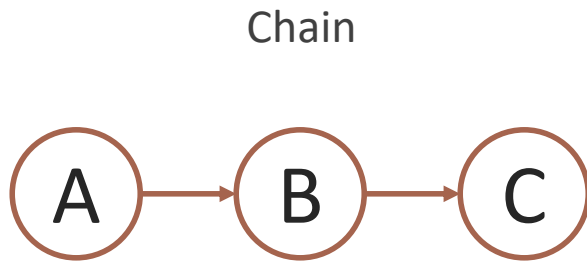
Graph Building Blocks



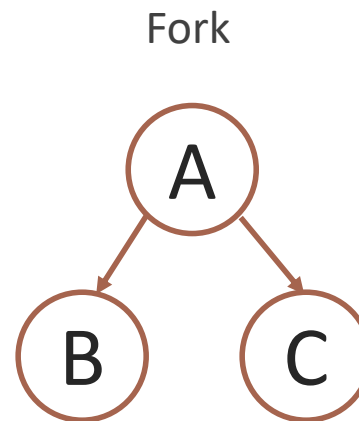
Nodes



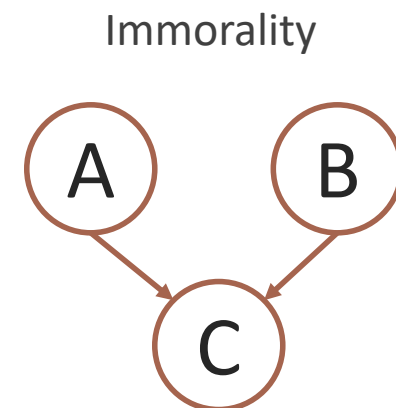
Association



Chain

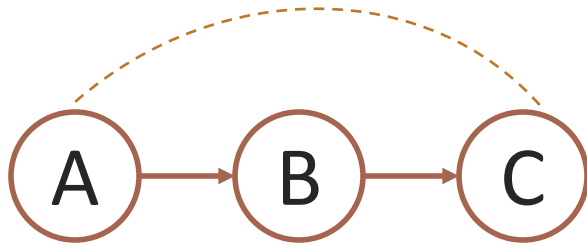


Fork

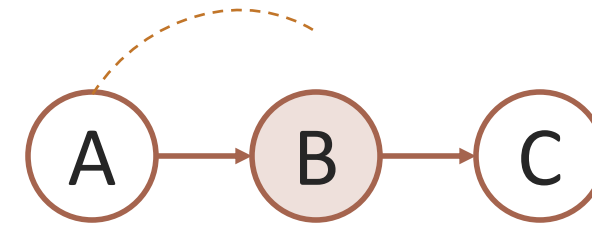
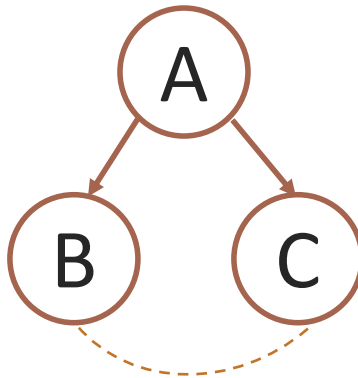


Immorality

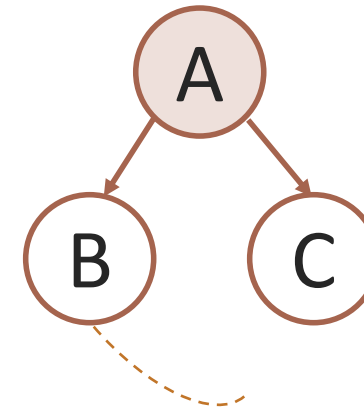
Blocked Nodes



Unblocked Path
(Dependence)



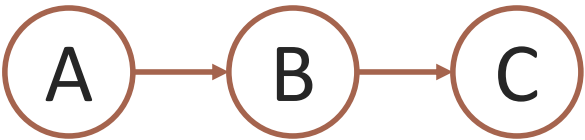
Blocked Path
(Independence)



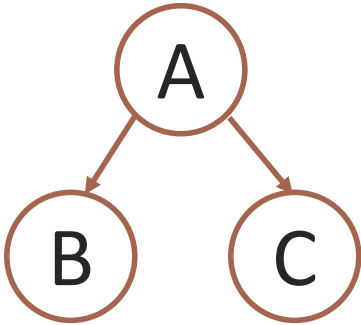
Conditional Independence

- Joint $P(A,B,C)$

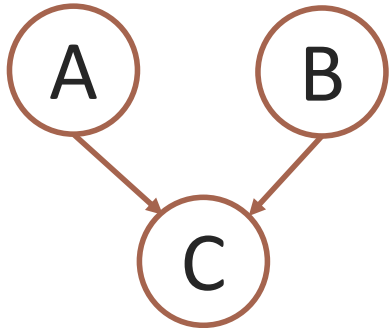
Causal chain



Common Cause



Common Effect



$P(A)P(B A)P(C A, B)$	$P(A)P(B A)P(C A, B)$	$P(A)P(B A)P(C A, B)$
$P(A)P(B A)P(C B)$	$P(A)P(B A)P(C A)$	$P(A)P(B)P(C A, B)$
Assumption $P(C A, B) = P(C B)$ C is independent from A given B	Assumption $P(C A, B) = P(C A)$ C is independent from B given A	Assumption $P(B A) = P(B)$ A is independent from B

Bayes Nets: Conditional Independence

- Assume conditional independences

$$P(x_i | x_1, \dots, x_{i-1}) = P(x_i | \text{Parents}(X_i))$$

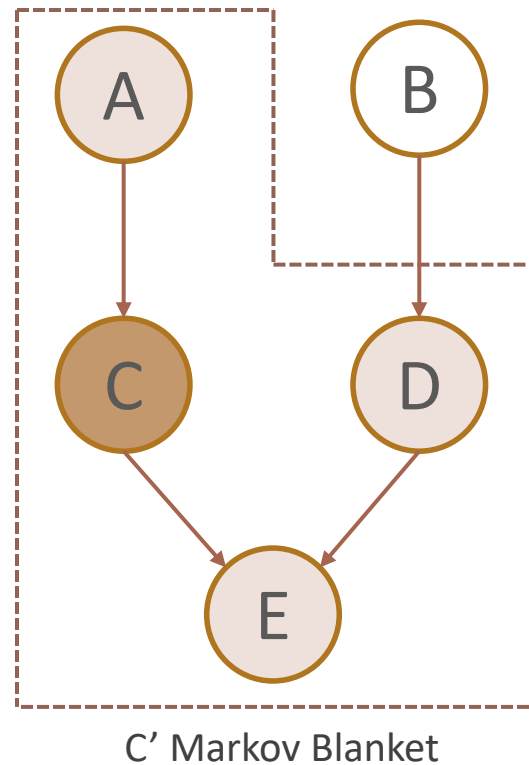
- JPD represented by the Bayesian network :

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{Parents}(X_i))$$

- It can be shown that this is a legal probability distribution $P \geq 0, \sum P = 1$

Markov Blanket

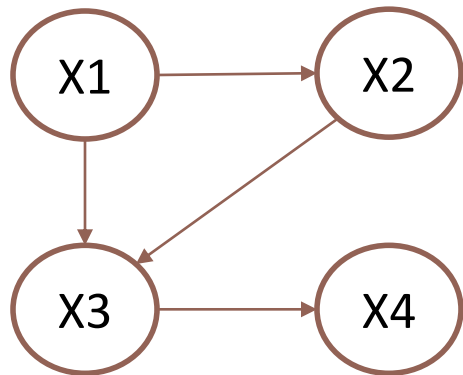
- Each variable is conditionally independent of all other variables given its Markov blanket i.e., its parents, children and children's parents



Bayes Net: JPD Example

- Modeling a joint probability distribution over 4 binary variables

$$p(x_1, x_2, x_3, x_4) = p(x_1)p(x_2|x_1)p(x_3|x_2, x_1)p(x_4|x_3)$$



x_1	x_2	x_3	$p(x_4 x_3)$
0	0	0	α_1
0	0	1	α_2
0	1	0	α_3
0	1	1	α_4
1	0	0	α_1
1	0	1	α_2
1	1	0	α_3
1	1	1	α_4

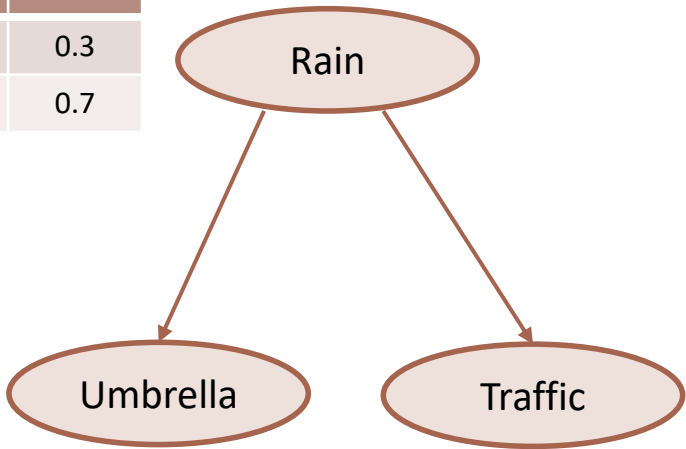
} $\ll 2^{n-1}$ parameters

Bayes Net: JPD Example

$Traffic \perp Umbrella \mid Rain$

$P(R)$

R	P
T	0.3
F	0.7



$P(U|R)$

	Umbrella	
Rain	T	F
T	0.8	0.2
F	0.1	0.9

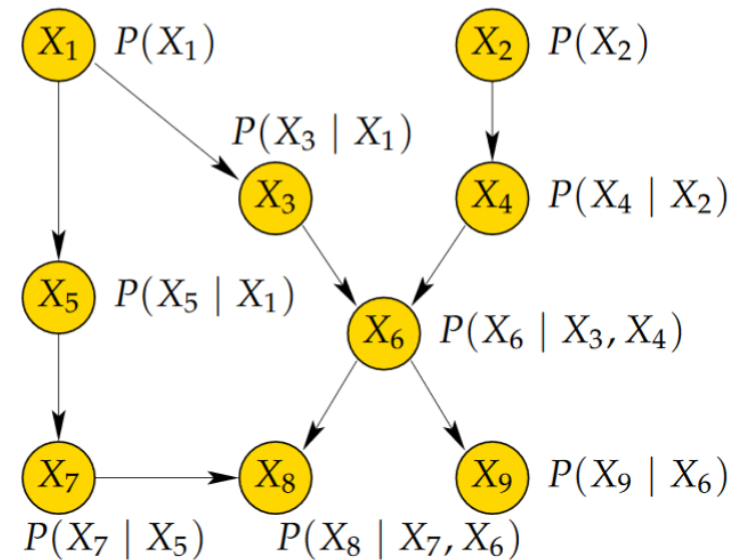
$P(T|R)$

	Traffic	
Rain	T	F
T	0.6	0.4
F	0.3	0.7

$P(U, T, R) = P(R) P(U|R) P(T|R)$

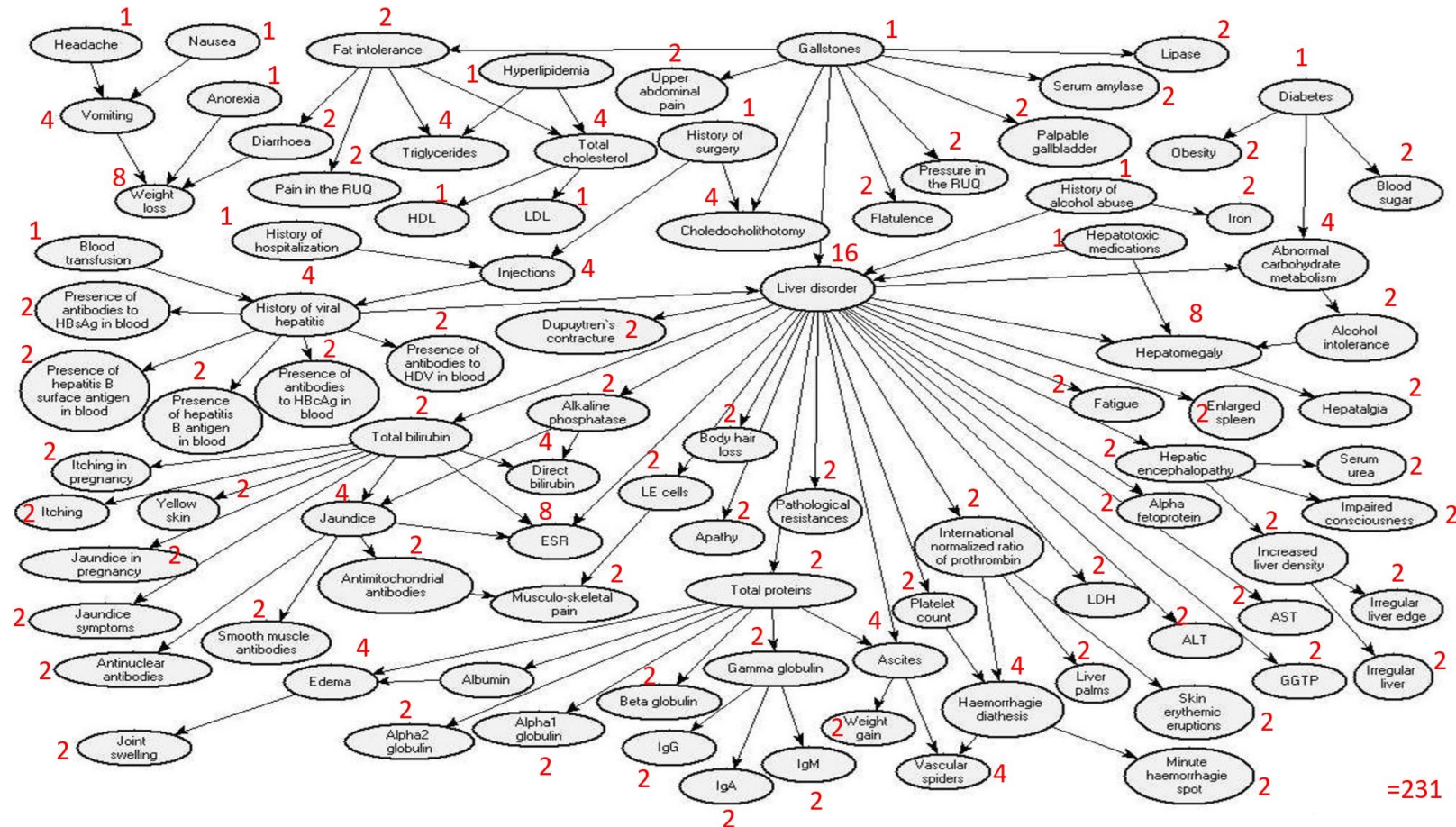
R	U	T	P(U, T, R)
T	T	T	0.144
T	T	F	0.096
T	F	T	0.036
T	F	F	0.024
F	T	T	0.021
F	T	F	0.049
F	F	T	0.189
F	F	F	0.441

Bayes Net: JPD Example



$$\begin{aligned} P(X_1, \dots, X_9) &= \\ &= P(X_9 | X_8, \dots, X_1) \cdot P(X_8 | X_7, \dots, X_1) \cdot \dots \cdot P(X_2 | X_1) \cdot P(X_1) \\ &= P(X_9 | X_6) \cdot P(X_8 | X_7, X_6) \cdot P(X_7 | X_5) \cdot P(X_6 | X_4, X_3) \\ &\quad \cdot P(X_5 | X_1) \cdot P(X_4 | X_2) \cdot P(X_3 | X_1) \cdot P(X_2) \cdot P(X_1) \end{aligned}$$

Bayes Net: JPD Liver Disorder



Through a Bayes Network we need to know about only 231 probability values to specify the joint probability of the liver disorder.

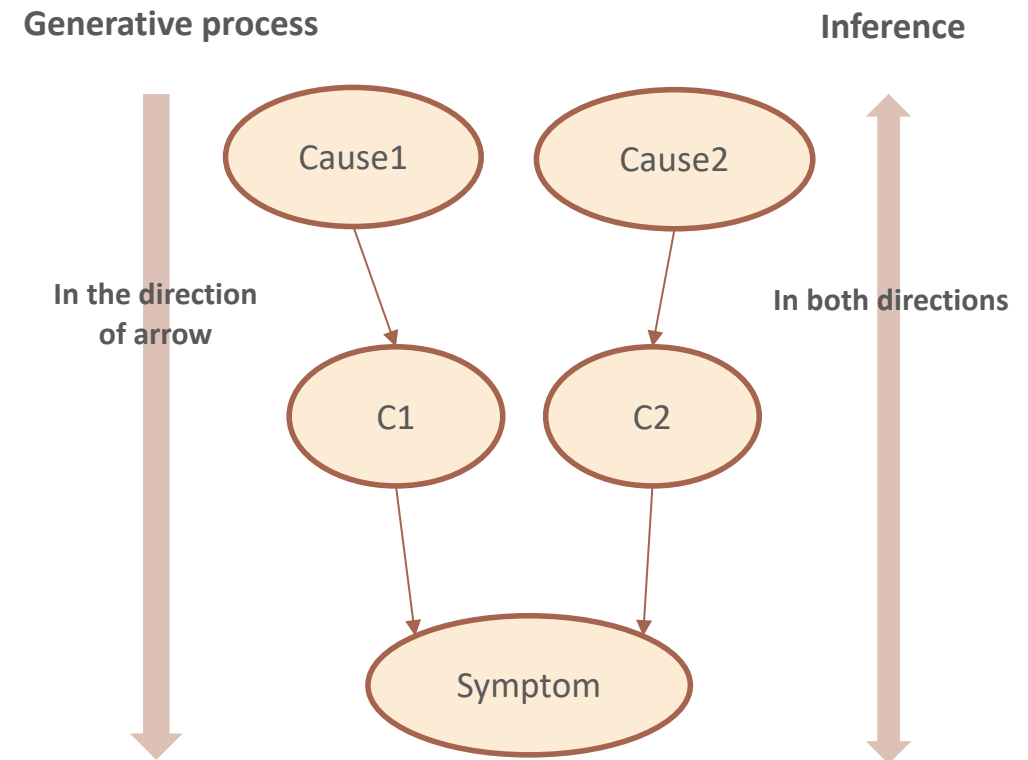
BAYES NETS INFERENCE

Bayes Nets: Inference

- BNs are useful for making predictions, diagnoses and explanations by computing the conditional probability distribution of a variable (or a set of variables) of interest
- Challenges
 - The problem of exact inference in graphical models is NP-hard ([Cooper, 1990](#)) , however efficient methods have been developed to cut down the possibly exponential time taken.
 - One approach is to use the factored representation of the JPD for efficient marginalization.
 - Decision problem associated with Bayesian network inference is NP-complete

Bayes Nets: Uses

- Knowledge representation
 - To model and explain a domain
- Decision-making
 - For domains with uncertainty
- Diagnosis - inferring inputs from outputs
 - $P(Cause|Symptom)$
- Prediction - inferring outputs from inputs
 - $P(Symptom|Cause)$
- Classification
 - $Argmax_{class} P(class|data)$



Bayes Nets: Marginal MAP Queries

- Evidence
 - A subset E of random variables in the model, and an instantiation e to these variables
- Query variables
 - A subset Q of random variables in the network.
- Inference:
 - Compute $P(Q = q \mid E = e)$ i.e., the *posterior probability distribution* over the values q of Q , conditioned on the fact that $E = e$.
 - This expression can also be viewed as the marginal over Q , in the distribution we obtain by conditioning on e .
 - If there is no evidence, probabilities of interest are prior probabilities $p(x_i)$.

$$p(q|e) = \frac{p(q, e)}{p(e)}$$

Bayes Nets: Inference Using JPD

- Approach

- The probability of any variable X_i conditioned on \mathbf{e} , i.e., $p(x_i|\mathbf{e})$:

$$P(x_i|e) = \frac{P(x_i, e)}{P(e)} \propto \sum_u p(x_i, e, u)$$

- The JPD $p(x_i, e, u)$ can be obtained with factorization which uses the information given in the BN, the conditional probabilities of each node given its parents.
 - Using the JPD we can respond to all possible inference queries by marginalization (summing out over irrelevant variables \mathbf{u}).
- Complexity
 - Summing over the JPD takes exponential time due to its exponential size, and more efficient methods have been developed. The key issue is how to exploit the factorization to avoid the exponential complexity. E.g., Variable Elimination

Probabilistic Inference Tasks

Belief Updating	$BEL(X_i) = P(X_i = x_i e)$	X_1, \dots, X_n are network variables
Most Probable Explanation (MPE)	$\bar{x}^* = \arg \max_x \sum P(\bar{x}, e)$	
Maximum A Posteriori Hypothesis (MAP)	$(a_1^*, \dots, a_k^*) = \arg \max_x \sum_{\bar{X}/A} P(\bar{x}, e)$	$A \subseteq X$ Hypothesis variables
Maximum Expected Utility (MEU) Decision	$(d_1^*, \dots, d_k^*) = \arg \max_d \sum_{\bar{X}/D} P(\bar{x}, e) U(\bar{x})$	$D \subseteq X$ Decision variables $U(\bar{x})$ Utility function

Causal Reasoning

- Predicting downstream effects of factors

- Scenario:

- How likely is a student to get a strong letter (knowing nothing else)?

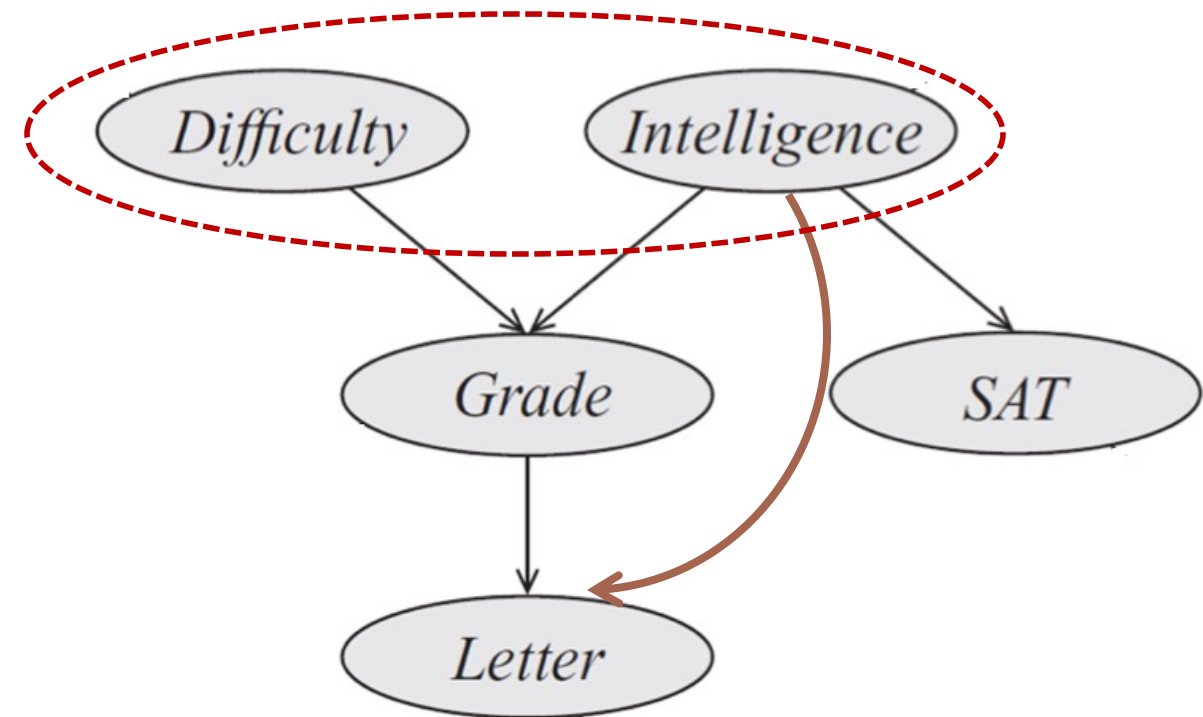
$$P(l^1) = 0.502$$

- But this student is not so intelligent (i^0)

$$P(l^1|i^0) = 0.389$$

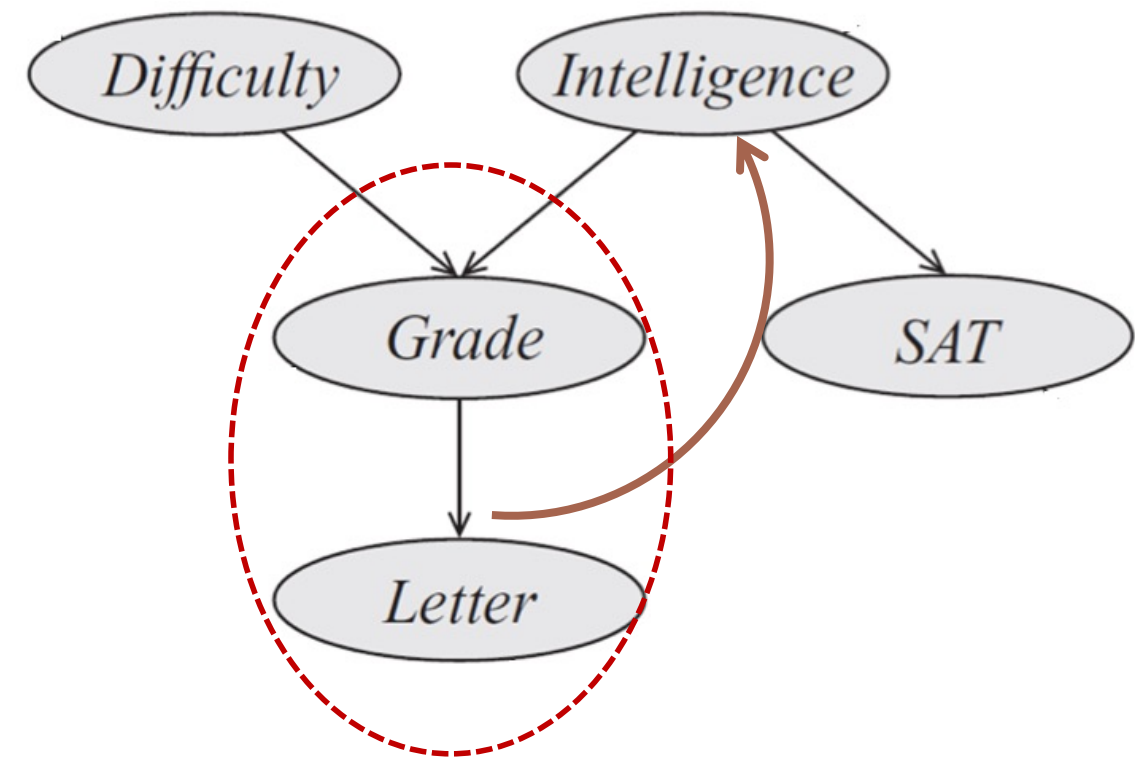
- Then we find out if the course is easy

$$P(l^1|i^0, d^0) = 0.513$$



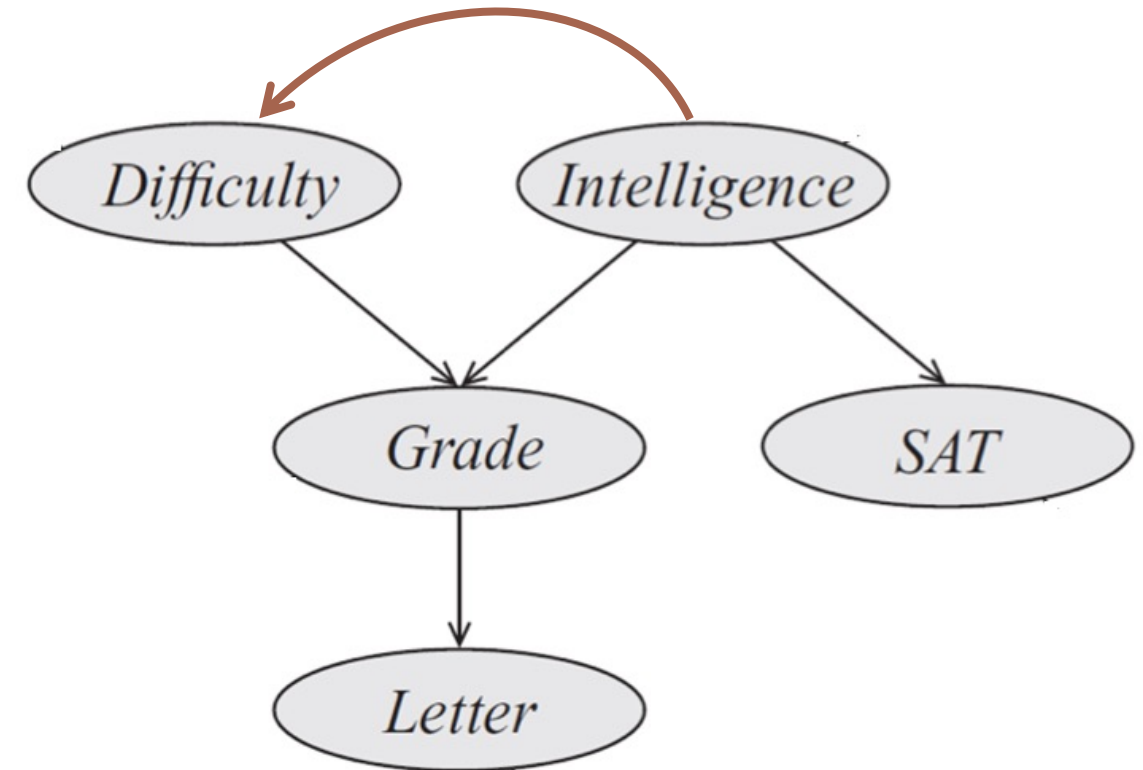
Evidential Reasoning

- Reasoning from effects to causes
- Scenario:
 - Recruiter wants to hire intelligent student
 - A priori this student is 30% likely to be intelligent
 $P(i^1) = 0.3$
 - Finds that the student received grade C
 $P(i^1|g^3) = 0.079$
 - Probability that class is difficult
 $P(d^1|g^3) = 0.629$
 - If recruiter has lost grade but has letter
 $P(i^1|l^0) = 0.14$
 - Recruiter has both grade and letter
 $P(i^1|l^0, g^3) = 0.079$
- Summary
 - Letter is immaterial

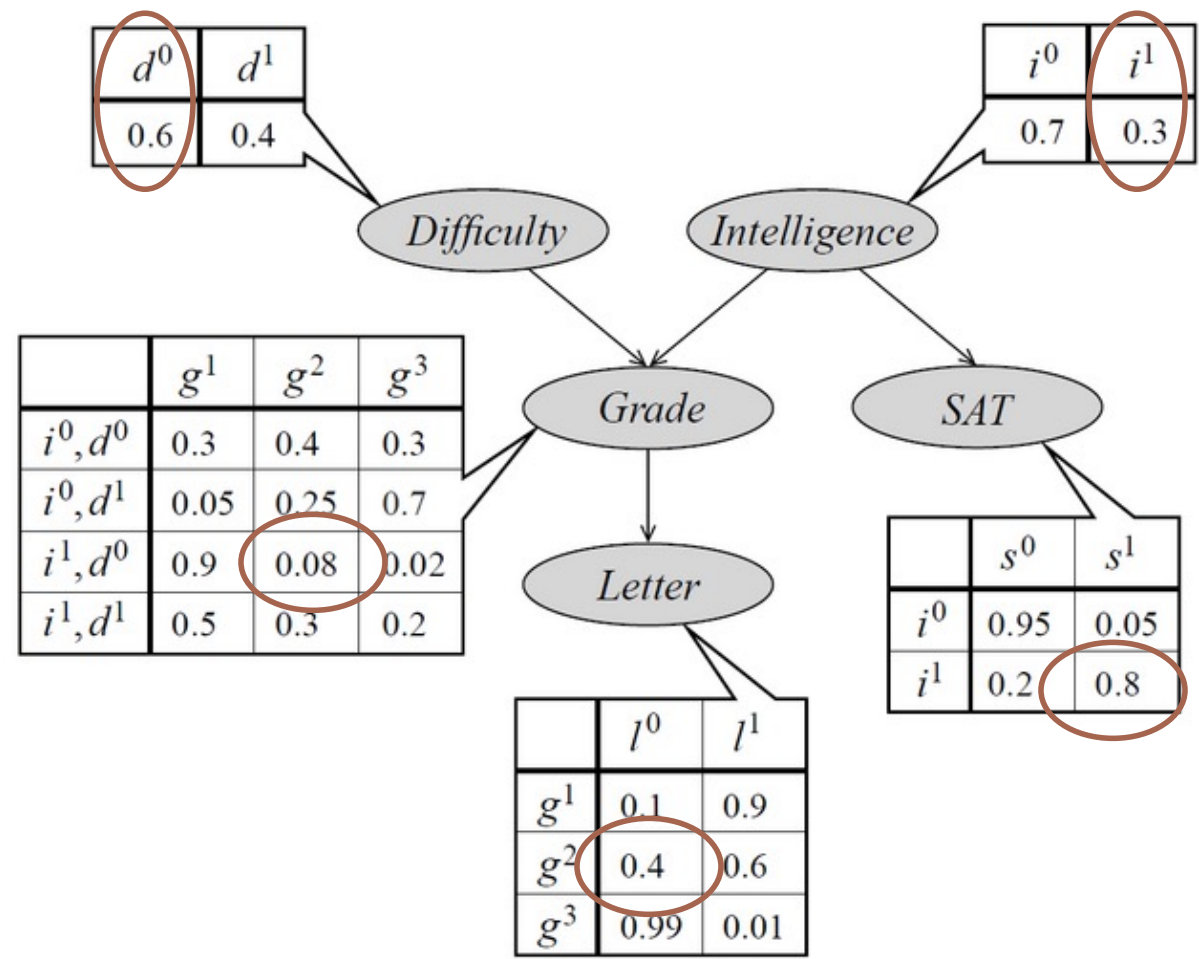


Intercausal Reasoning

- Reasoning from one causal factor to another
- Scenario:
 - Recruiter has grade (letter does not matter)
 $P(i^1|g^3) = P(i^1|l^0, g^3) = 0.079$
 - Recruiter receives high SAT score (leads to dramatic increase)
 $P(i^1|g^3, s^1) = 0.578$
 - Probability of class is difficult also goes up from:
 $P(d^1|g^3) = 0.629$ to $P(d^1|g^3, s^1) = 0.76$
- Summary
 - Information about SAT score gave us information about Intelligence which with Grade told us about difficulty of course



Bayes Net Inference: Example



Question: What is the probability that an intelligent student gets a B in an easy class, a high SAT and a weak letter?

$$P(d^0, i^1, g^2, s^1, l^0)$$
$$= 0.6 * 0.3 * 0.08 * 0.8 * 0.4$$

	Values
Intelligence	i^0 (Low), i^1 (High)
Grade	g^1 (A), g^2 (B), g^3 (C)
Difficulty	d^0 (Easy), d^1 (Hard)
Letter	l^0 (Weak), l^1 (Strong)
SAT	s^0 (Low), s^1 (High)

Bayes Nets: Advantages

- Modeling Uncertainty
 - BNs represent both knowledge and uncertainty of a domain
 - What is known is represented by the causal structure of the domain (graph)
 - What is unknown is materialized by probabilities (the random variables)
- Interpretability
 - BNs are both mathematically rigorous and intuitively understandable
 - Combine principles from graph theory, probability theory, computer science, and statistics.
- Dimensionality
 - Rigid systems might find it necessary to enumerate every possibility; BNs are more compact

Summary

- Uncertainty is an unescapable aspect of most real-world systems.
- To obtain meaningful conclusions, we need to reason not just about what is possible, but also about what is probable
- Bayes Nets provide a formal and efficient framework for reasoning using probability theory considering multiple possible outcomes and their likelihood
- Bayes Nets can be learned from data.