BAYESIAN NETWORKS

Ashish Pujari

Lecture Outline

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

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 in AI

- Expert Systems
  - Computer systems that emulates the decision-making ability of a human expert using *if–then* rules
- Evolution
  - Until the early 1980s, AI was represented by rule based expert systems. They were unable to take into account uncertainty in reasoning.
  - Judea Pearl proposed Bayesian networks (late 1980s) - a probabilistic approach that aimed overcome limits of expert systems.
  - The use of expert systems quickly began to decline in 1990s.

Pathfinder 1
Rule Based System

Pathfinder 2
Naïve Bayes Model

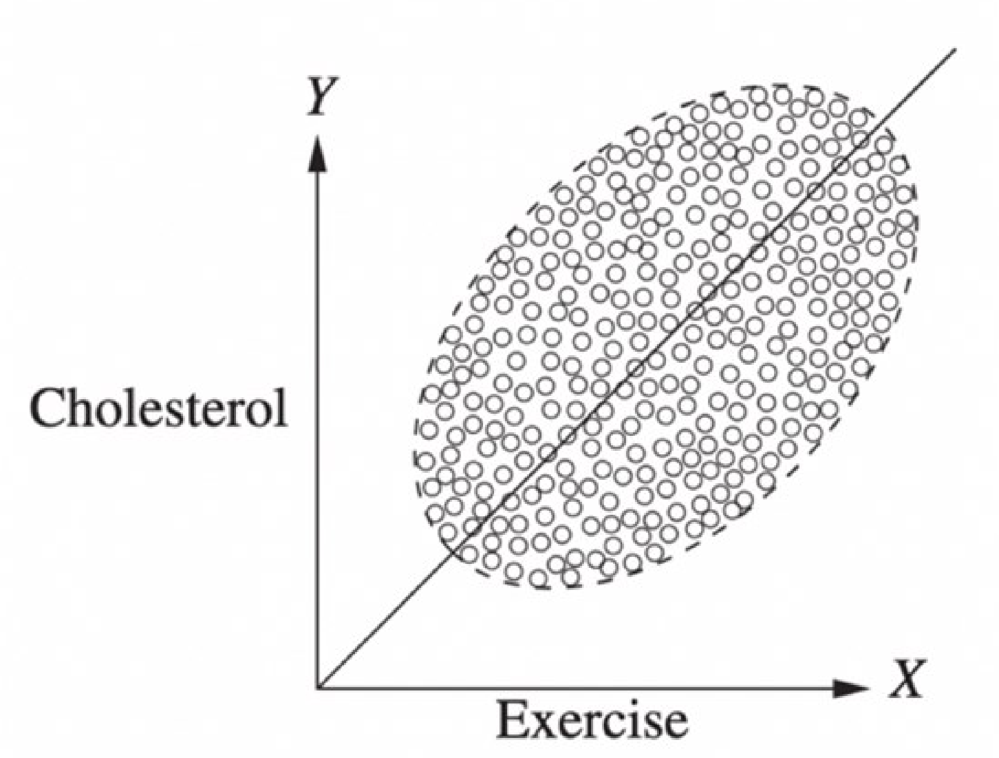
Pathfinder 3
Naïve Bayes with Knowledge Engineering

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

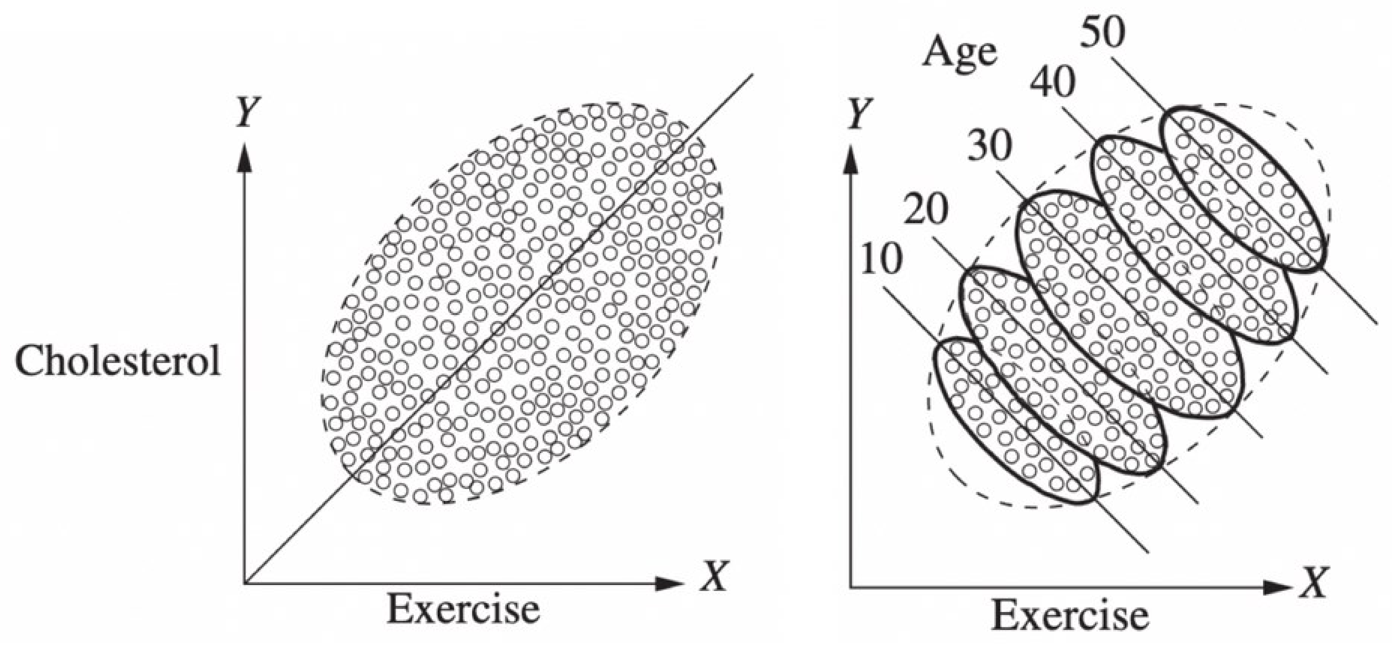


Concrete causal dilemmas with cholesterol

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. E.g., association between exercise and cholesterol is reversed when conditioned on age.

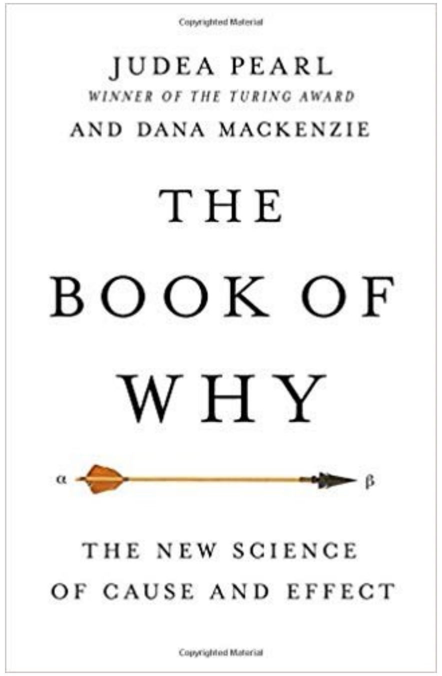


Concrete causal dilemmas with cholesterol

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

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
  - *"Causal reasoning is an indispensable component of human thought that should be formalized and algorithmicized toward achieving human-level machine intelligence."* - Judea Pearl
  - May be used to direct scientific research and eliminate unlikely hypotheses for understanding certain phenomenon. However, it cannot be treated as proof.



Probabilistic Models

- Probabilistic Models incorporate random variables and probability distributions into the model of an event or phenomenon
- While a deterministic model gives a single possible outcome for an event, a probabilistic model gives a probability distribution as a solution.
- Applications
  - Statistical physics, quantum mechanics, and theoretical computer science.

Bayesian Inference

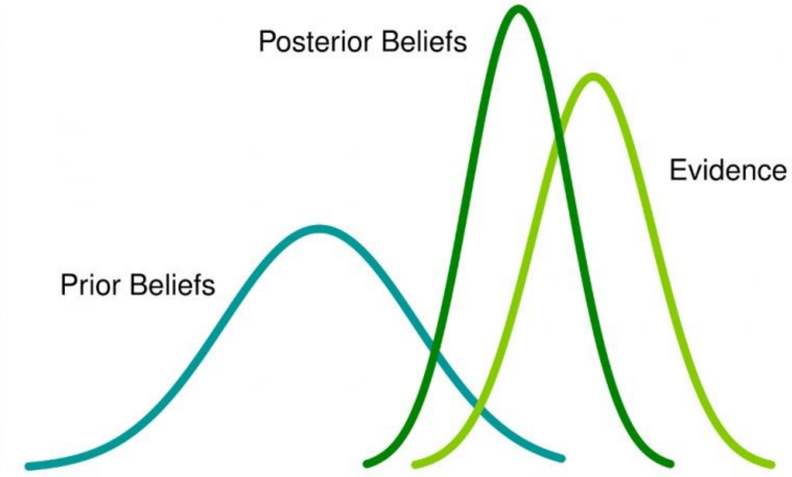
Posterior Probability

Likelihood of observations

Prior Probability

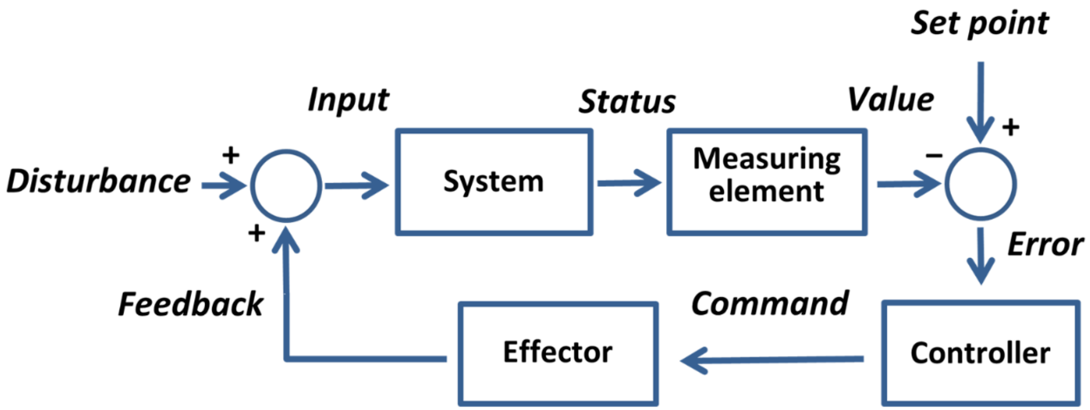
$$P(\theta|X) = \frac{P(X|\theta)P(\theta)}{P(X)}$$

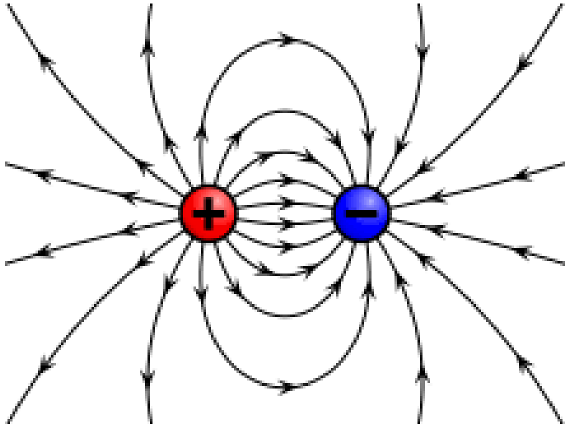
Evidence
Marginal Likelihood



Graphical Representations

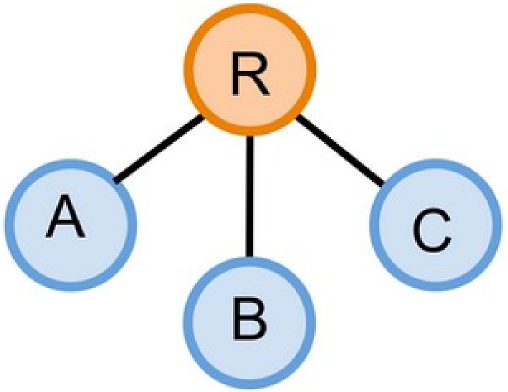
- Graphical representations can be very powerful in understanding and predicting behavior of complex systems in engineering, physics and other fields.

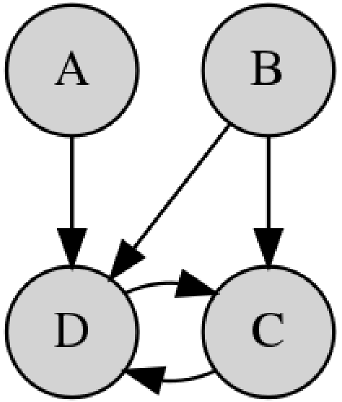




Probabilistic Graphical Models (PGMs)

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





Types of PGMs

- Bayesian Networks (BN)
- Hidden Markov Models (HMM)
- Markov Random Fields (MRF)
- Conditional Random Fields (CRF)

PROBABILITY REVIEW

Definitions

- Random variable
  - A variable generated by a random phenomenon $X : \{x_1, x_2, \dots, x_n\}$
- Expected value
  - $E(X)$ is a generalization of the concept of "average" to included "weighted averages"
- Sample Space
  - Entire collection of possible outcomes from an experiment is termed the sample space, indicated as Ω .
  - Examples $\Omega = \{x_1, x_2, \dots, x_n\}$, $\Omega = \{H, T\}$, $\Omega = \{A\}$
- Events
  - An event is some subset of outcomes from the sample space
- Probability
  - Likelihood that an event will occur under a set of given conditions
  - E.g., $P(A) = \frac{N_A}{N}$ where N_A is the number of times event A occurred from N experiments

Marginal Probability Distribution

- Marginal distribution represents our prior knowledge about an event

$$P(x) = \sum_{y} P(x, y)$$

- Example:

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

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

Conditional Probability Distribution

- Conditional probability is the 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:

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

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

Probability Rules

- Intersection Rule

$$P(A \cap B) = P(A)\, P(B|A)$$

- Union Rule

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = P(A) + P(B) - P(A)P(\, B \mid A\,)$$

- Chain Rule

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

$$P(x_1, x_2, \dots, x_n) = P(x_1)P(x_2|x_1)P(x_3|x_2, x_1) \dots = \prod_{i=1}^{n} P(x_i\, |x_{i-1}, \dots, x_1)$$

- Bayes Rule

$$P(x|y) = \frac{P(y|x)P(x)}{P(y)}$$

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

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

Joint Probability Distribution: Example

- Modeling a joint probability distribution over 4 binary variables

$$p(x1, x2, x3, x4) = p(x1)p(x2|x1)p(x3|x2, x1)p(x4|x3, x2, x1)$$

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

$$\boldsymbol{X} \perp \boldsymbol{Y}$$

$$\boldsymbol{Y} \perp \boldsymbol{X}$$

- 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

$P(T)$

Temp	Prob
Hot	0.5
Cold	0.5

$P(W)$

Weather	Prob
Sun	0.6
Rain	0.4

$P_1(T,W)$

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

$\boldsymbol{T \perp W}$

$P_2(T,W)$

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

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

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

- Notation

$$\boldsymbol{X} \perp \boldsymbol{Y} \mid \boldsymbol{Z}$$

$$\boldsymbol{Y} \perp \boldsymbol{X} \mid \boldsymbol{Z}$$

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

Conditional Independence: Example

- Alarm going off is conditionally independent of the fire given that there is smoke

$$Alarm \perp Fire \mid Smoke$$

Conditional Independence: Example

$$Traffic \perp Umbrella \quad | \quad Rain$$

- Chain Rule Decomposition

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

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

- Conditional Independence

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

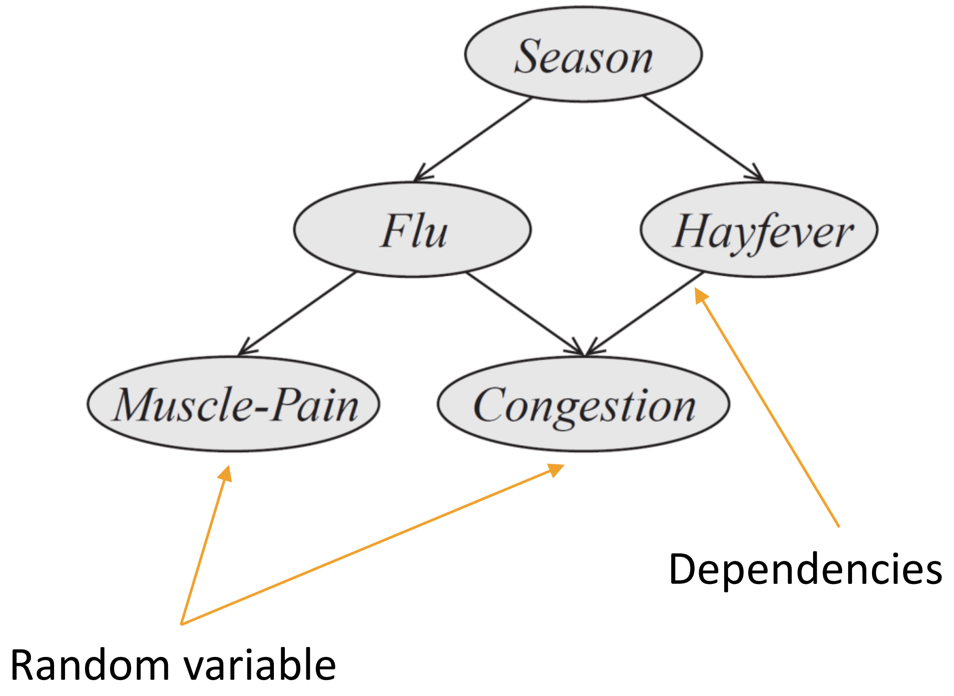
$$P(Rain, Traffic, Umbrella) = P(Rain)\ P(Traffic|Rain)\ P(Umbrella\ |Rain)$$

BAYESIAN NETWORKS

Representation, Inference

Bayesian Networks (Bayes Nets)

- PGMs that consists of a Directed Acyclic Graph (DAG) and a set of local conditional probability distributions
- They exploit domain structure to allow compact representations of complex models
- Since each variable is considered random, its distribution i.e., the probabilities of the different outcomes must be defined



Bayes Nets: Applications

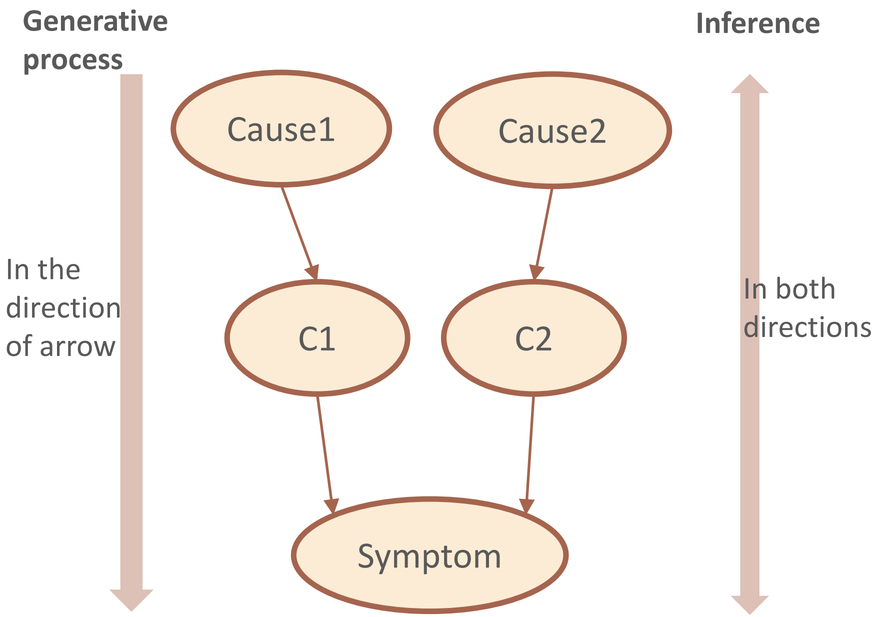
- Medicine
  - medical diagnosis, drug discovery
- Forensics
  - victim identification, kinship analysis
- Genomics
  - Gene Regulatory Network - governs the expression levels of mRNA and protein
- Biomonitoring
  - quantify the concentration of chemicals in blood and tissue
- Autonomous Systems
  - 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

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
    - i.e., the graph
  - What is unknown is materialized by probabilities or the uncertain part of this domain
    - i.e., the nodes which represent the random variable.
- 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

Bayes Nets: Uses

- Knowledge representation
  - To model and explain a domain
- Decision-making
  - Support decision making in domains with uncertainty
  - Find good strategies for solving tasks while maximizing utility
- Diagnosis - inferring inputs from outputs
  - $P(Cause|\ Symptom)$
- Prediction - inferring outputs from inputs
  - $P(Symptom|Cause)$
- Classification
  - $\underset{class}{Argmax}\ P(class|data)$



Bayes Nets: Definition

- 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 $\{x_{i_1}, x_{i_2}, .. x_{i_n}\}$
- $pa(i)$ denotes the set of parents of node i in the graph
- to each node $i \in V$ corresponds a conditional probability table $\mathrm{P}\ (X_i \mid (X_j)_{j \in pa(i)}\ )$
- the DAG implies conditional independence relations between $(X_i)_{i \in V}$

Bayes Nets: Nodes

- Each node represents a discrete random variable, that is, a variable with at least two possible outcomes.

- Random Variables
  - A random variable denotes an attribute, feature, or hypothesis about which we may be uncertain.
  - Each random variable has a set of mutually exclusive and collectively exhaustive possible values.
  - Bayes nets can in principle deal with continuous variables, but discrete nodes are more common. Using continuous variables imposes restrictions for the architecture and usable algorithms.

Bayes Nets: Dependencies

- A link in a Bayes net is interpreted as a causal relationship
- $A \rightarrow B$ means that A is one of the causes of B i.e., B can change due to changes in A

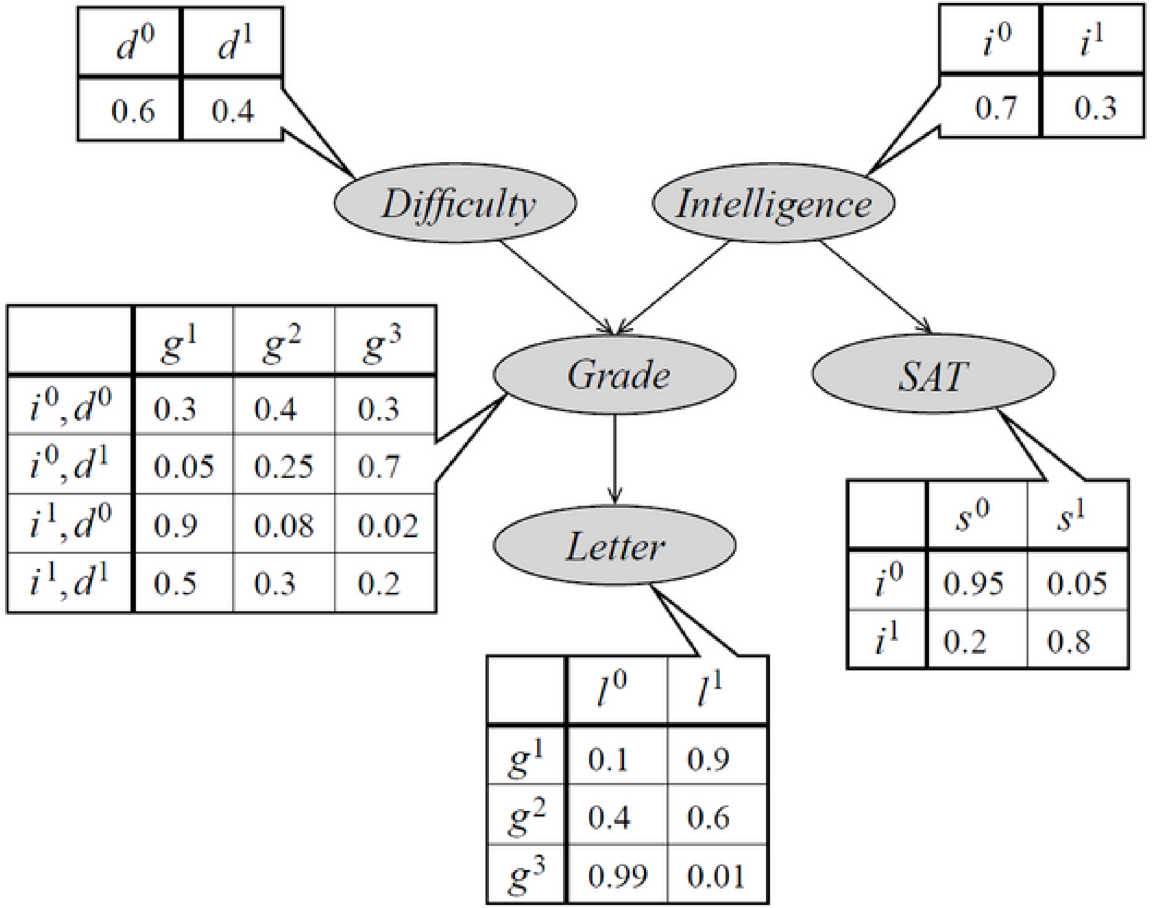
- Guidelines
  - Introducing a dependency should be considered only if you can quantify it.
  - Introducing a dependency requires more explicit knowledge and has a certain cost.
  - In some situations, you may be positive about a qualitative dependency, but unable to quantify it.
  - In other situations, you may have evidence supporting a dependency, but no way to use.

Bayes Nets: Components

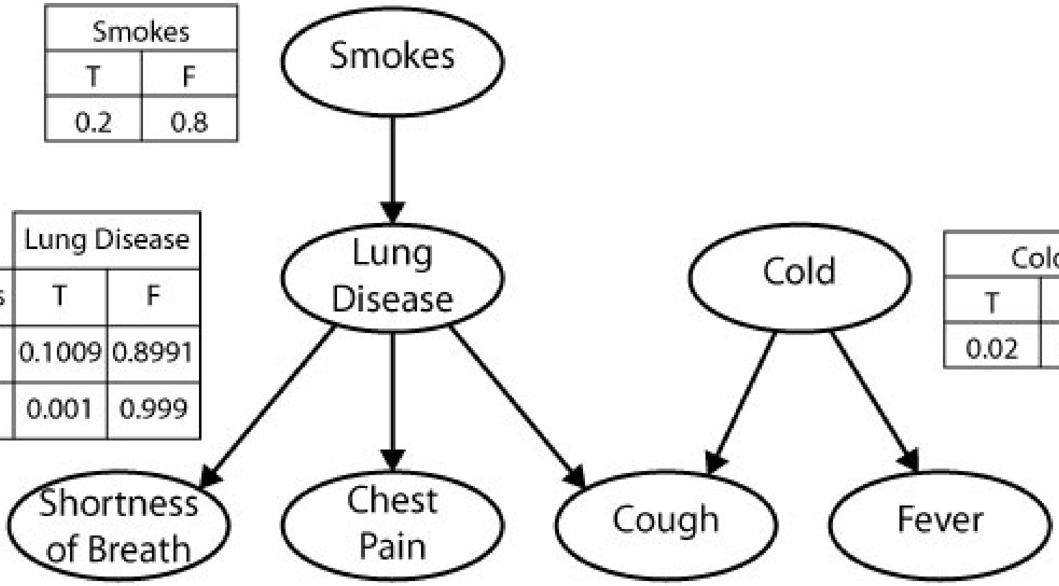
- Qualitative
  - Node: a variable X, with a set of states $\{x_{i_1}, x_{i_2}, .. x_{i_n}\}$
  - Arc: a dependency of a variable X on its parents P.

- Quantitative
  - Conditional probability tables (CPDs), attached to each node, expressing the probability of the variable at the node conditioned on its parents.
  - The distributions of a variable X given each combination of states p_i of its parents P
  - This distribution entails enough information to attribute a probability to any event expressed with the variables of the network.

Bayes Nets: Example - Student Net



Bayes Nets: Example - Lung Disease



Smokes	
T	F
0.2	0.8

	Lung Disease	
Smokes	T	F
T	0.1009	0.8991
F	0.001	0.999

Cold	
T	F
0.02	0.98

	Shortness of Breath	
Lung Disease	T	F
T	0.208	0.792
F	0.01	0.99

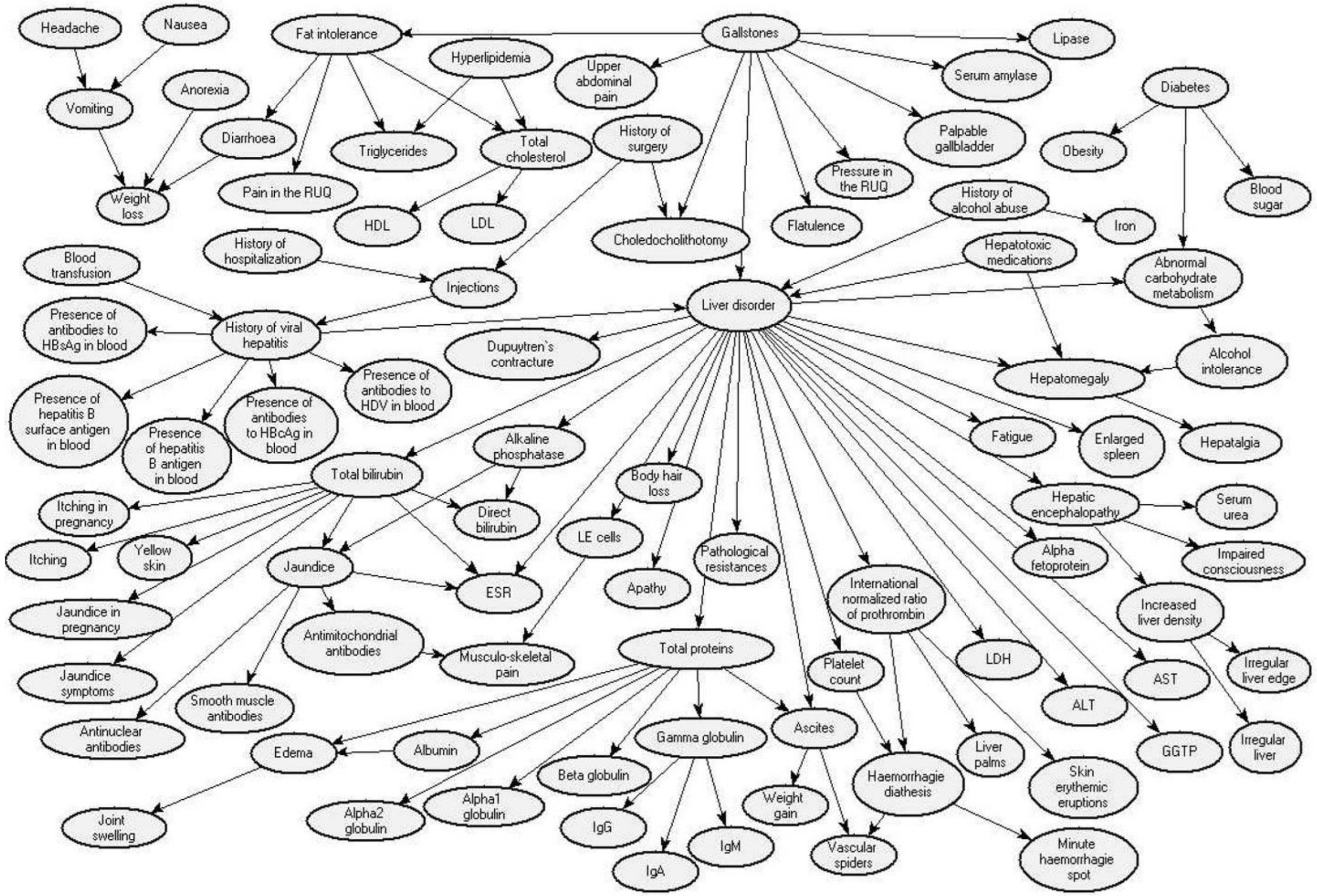
	Chest Pain	
Lung Disease	T	F
T	0.208	0.792
F	0.01	0.99

		Cough	
Lung Disease	Cold	T	F
T	T	0.7525	0.2475
T	F	0.505	0.495
F	T	0.505	0.495
F	F	0.01	0.99

	Fever	
Cold	T	F
T	0.307	0.693
F	0.01	0.99

https://probmods.org/chapters/dependence.html

Bayes Nets: Example - Liver Disorder



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

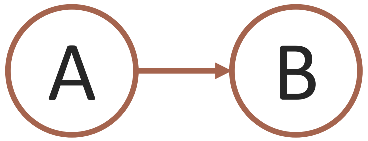
“A Bayesian Network Model for Diagnosis of Liver Disorders” – Agnieszka Onisko, et al.

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

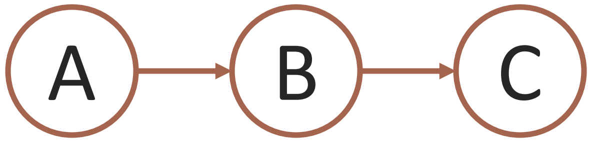
Graph Building Blocks

Nodes

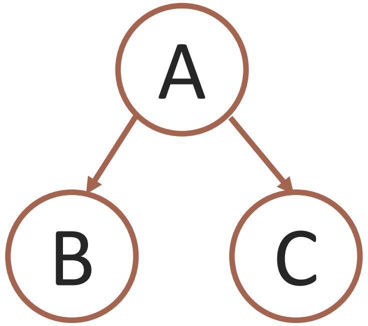


Association

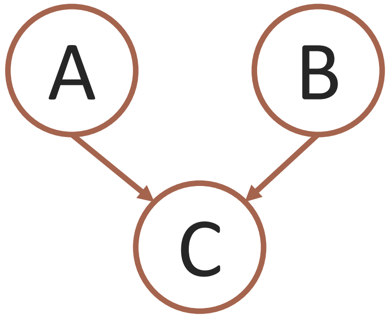
Chain



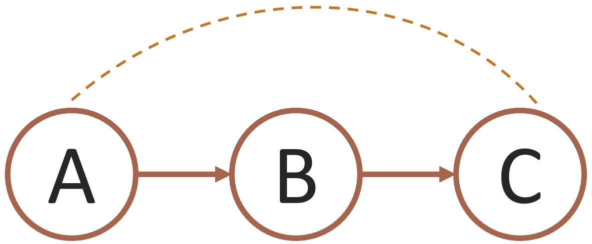
Fork



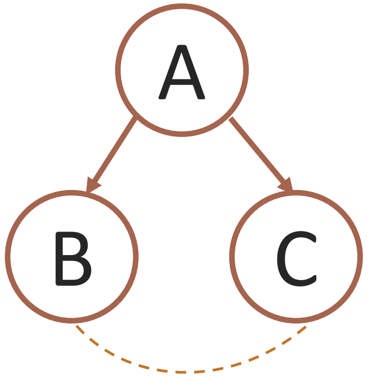
Immorality

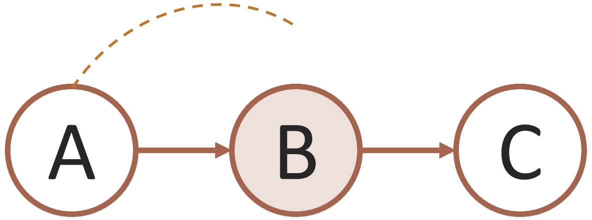


Blocked Nodes

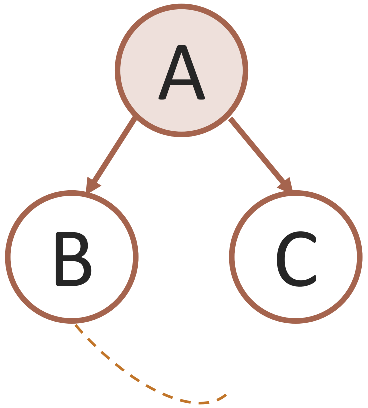


Unblocked Path
(Dependence)



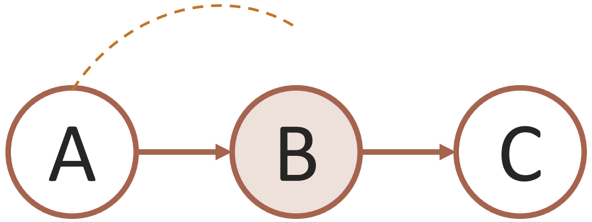


Blocked Path
(Independence)



Proof of Conditional Independence (Chains)

- Show that $P(a, c|b) = P(a|b)P(c|b)$



- Bayesian Network Factorization

$$P(a, b, c) = P(a)P(b|a)\,P(c|a)$$

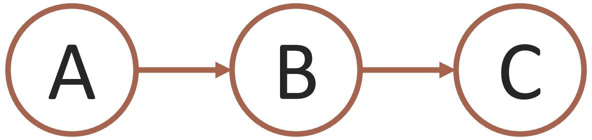
- Bayes Rule

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

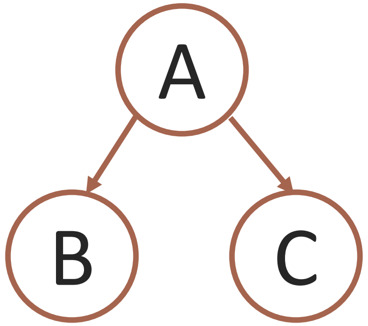
Conditional Independence

- Joint P(A,B,C)

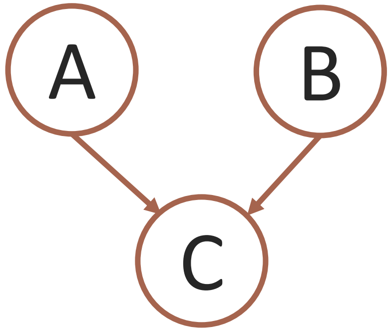
Causal chain



Common Cause



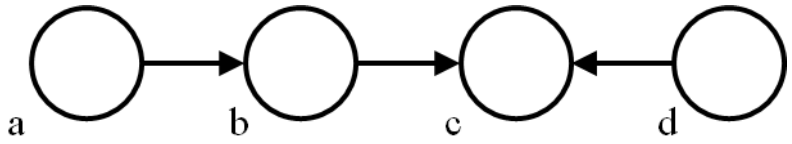
Common Effect



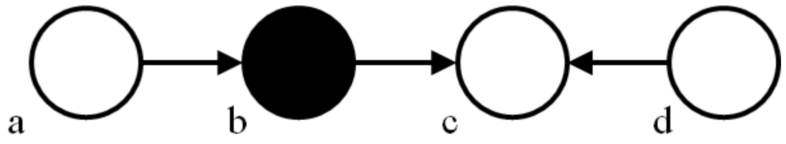
$P(A)P(B\|A)P(C\|A,B)$	$P(A)P(B\|A)P(C\|A,B)$	$P(A)P(B\|A)P(C\|A,B)$
$P(A)P(B\|A)P(C\|B)$	$P(A)P(B\|A)P(C\|A)$	$P(A)P(B)P(C\|A,B)$
Assumption $P(C\|A,B) = P(C\|B)$ C is independent from A given B	Assumption $P(C\|A,B) = P(C\|A)$ C is independent from B given A	Assumption $P(B\|A) = P(B)$ A is independent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

D-Separation Rules

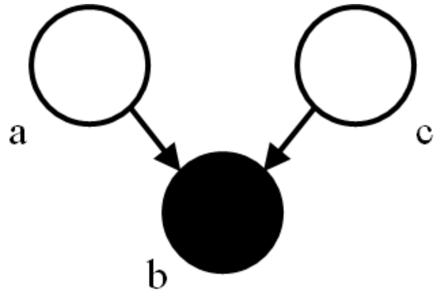
- **d-separation** is a criterion for deciding, from a given a causal graph, whether a set X of variables is independent of another set Y, given a third set Z
- Rule 1
  - *X* and *Y* are d-separated if there is no unidirectional path (any sequence of edges regardless of their directionality) between them. E.g. Nodes *a* and *c* are d-connected whereas *a* and *d* or *b* and *d* are d-separated, since the *b* and *d* are both parents of *c*.



- Rule 2
  - *X* and *Y* are d-separated given another set of nodes *Z* if *Z* "blocks" any unidirectional path between them. E.g. Node *b* blocks the path between *a* and *c*, causing them to be d-separated.



- Rule 3
  - If a collider (node with 2 or more parents) or its descendants is in the set Z, it breaks the d-separation of its parents. *E.g. Node b* breaks the d-separation between *a* and *c*.



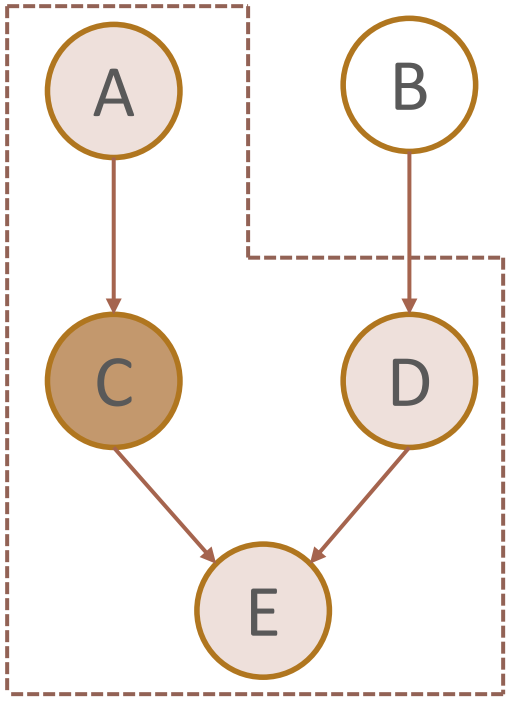
Bayes Nets: Conditional Independence

- *Every variable is conditionally independent of its non-descendants given its parents.*

- Given the values of the direct parents of a variable X_i, this variable X_i is independent of all its other predecessing variables in the graph.
- This property is used to significantly reduce the number of parameters that are required to characterize the JPD of the variables.
- This reduction provides an efficient way to compute the posterior probabilities given the evidence.

Markov Blanket

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



Markov Blanket

Bayes Nets: Conditional Independence

- Conditional independence is halfway between the intractable complete dependence and the infrequent and unrealistic case of mutual independence
- Assume conditional independences

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

- JPD represented by the Bayesian network :

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

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

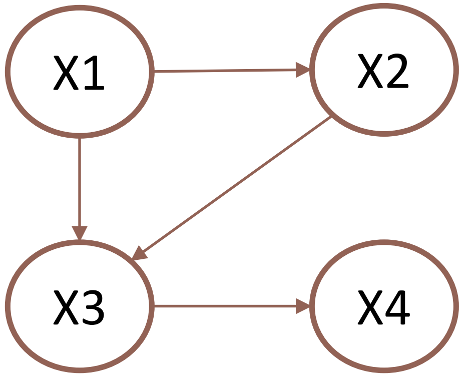
Bayes Net: Joint Probability Distribution (JPD)

- BNs implicitly encode joint distributions as product of local conditional distributions
- JPD over all variables is computed as the product of all these conditional probabilities.
- There are efficient algorithms for computing probabilities without having to generate the underlying JPD which would be unfeasible in many cases.

Bayes Net: JPD Example

- Modeling a joint probability distribution over 4 binary variables

$$p(x1, x2, x3, x4) \; = \; p(x1)p(x2|x1)p(x3|x2, x1)p(x4|x3)$$



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

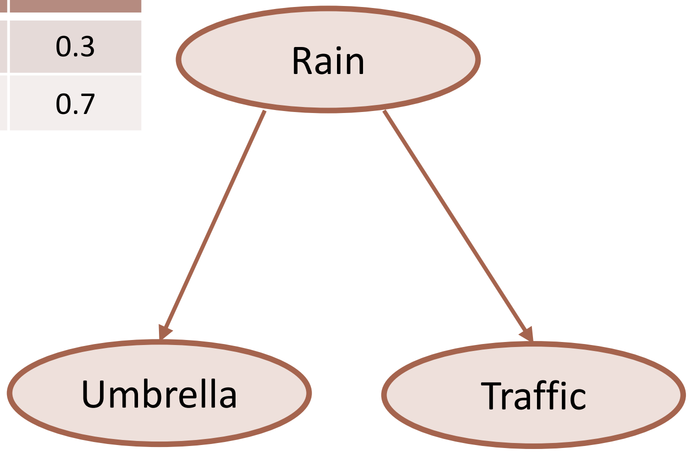
$\ll 2^{n-1}$ parameters

Bayes Net: JPD Example

$$Traffic \perp Umbrella \;\;|\;\; Rain$$

$P(R)$

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

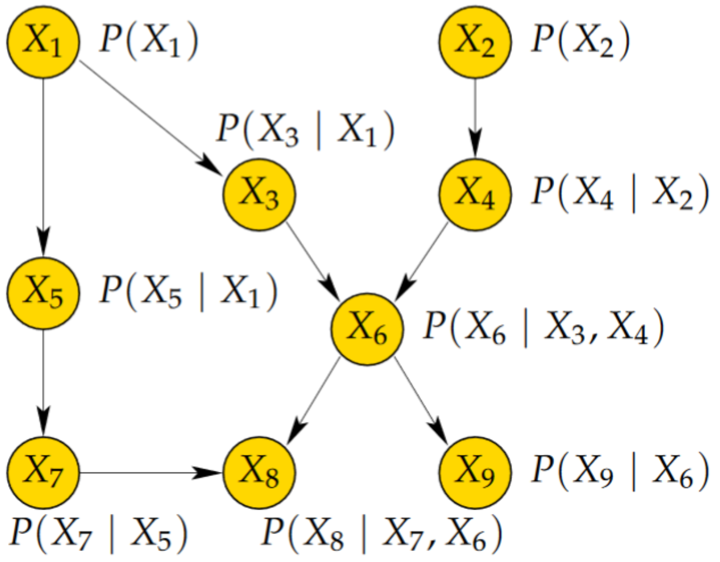
$P(T|R)$

	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

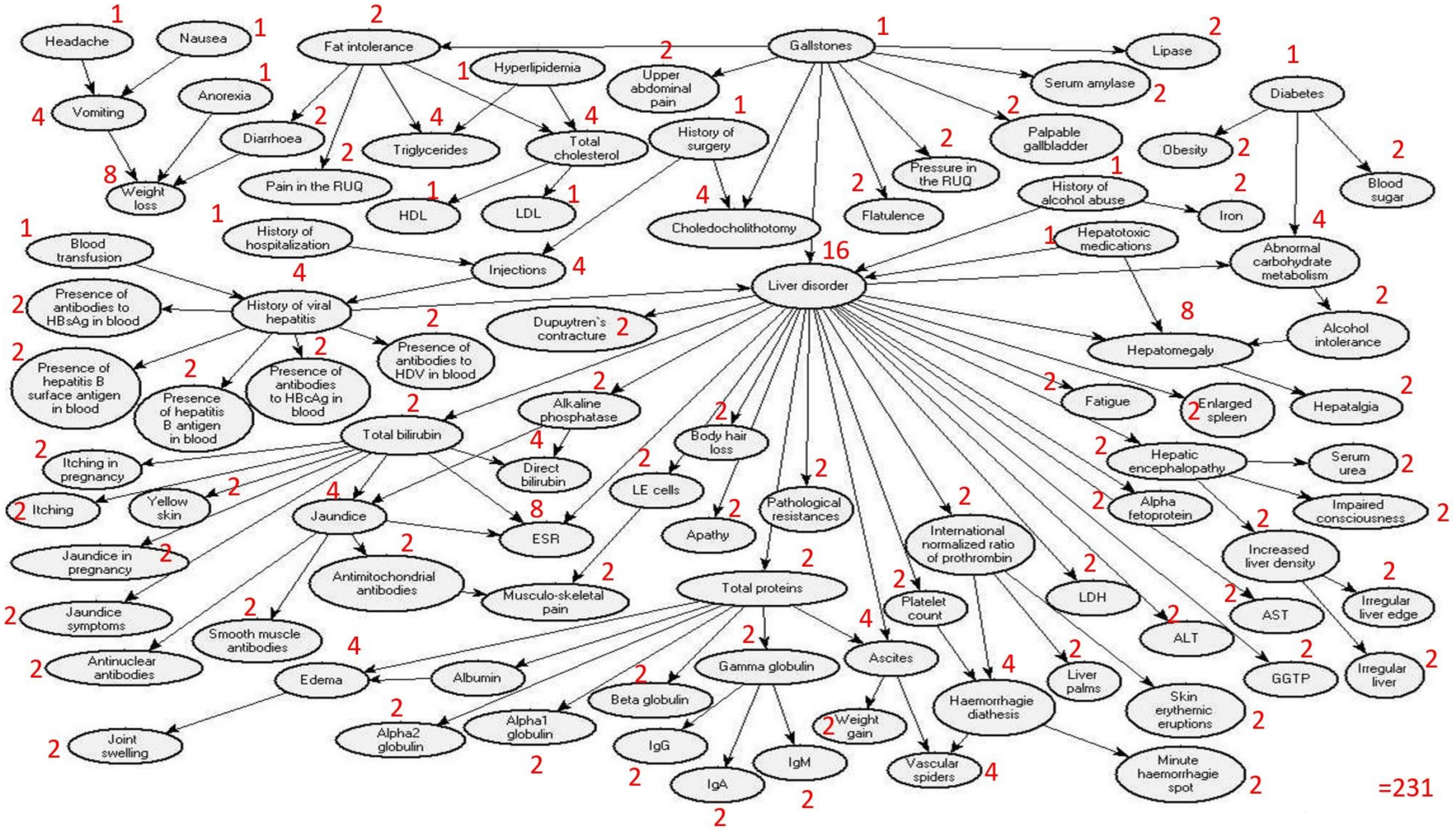
Bayes Net: JPD Example



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

http://staff.utia.cas.cz/vomlel/slides/presentace-karny.pdf

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.

https://www.norwegiancreations.com/2018/09/artificial-intelligence-bayes-network/

Bayes Nets: Inference

- Besides visualizing the relationships between variables, 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

https://www.frontiersin.org/articles/10.3389/fncom.2014.00131/full

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 **e**, i.e., $p(x_i | \mathbf{e})$:

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

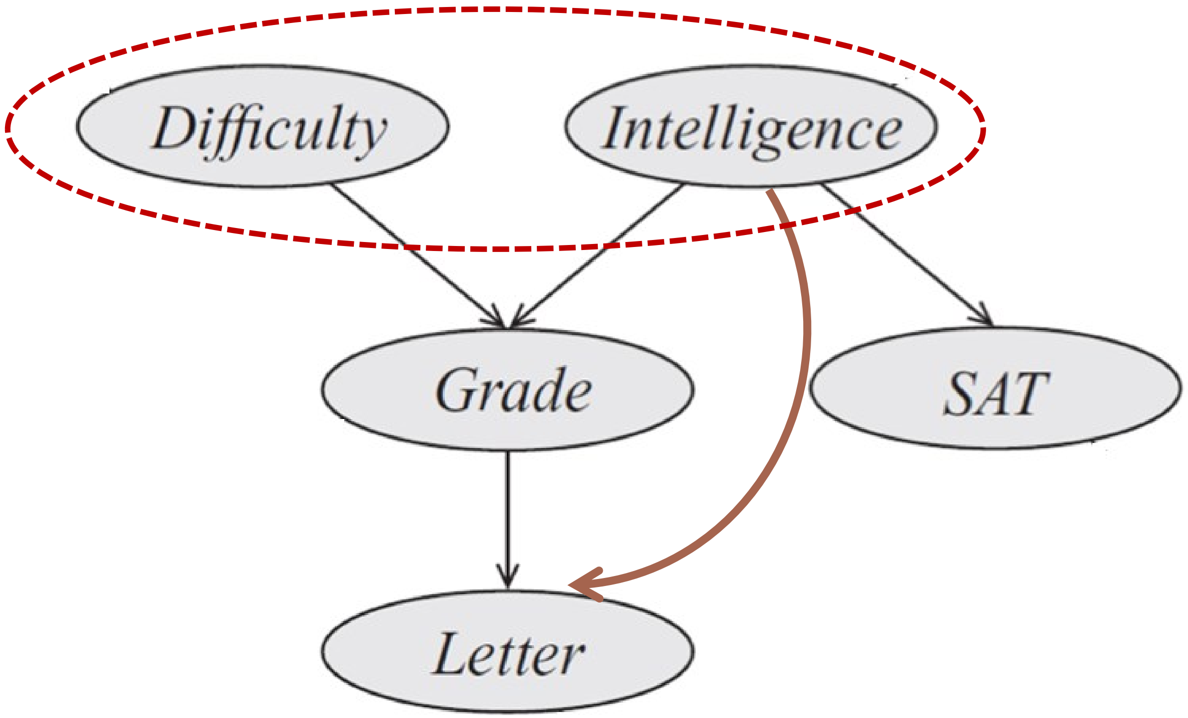
  - The JPD $p(x_i, e, u)$ can be obtained with factorization which uses the information given in the BN, the conditional probabilities of each node given its parents.
  - Using the JPD we can respond to all possible inference queries by marginalization (summing out over irrelevant variables $\boldsymbol{u}$).
- 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	$BEL(X_i) = P(X_i = x_i \mid \mathrm{e})$	$X_1, \ldots, X_n$ are network variables
Finding Most Probable Explanation (MPE)	$\bar{x}^* = \arg\max_x \sum P(\bar{x}, \mathrm{e})$	
Finding Maximum A Posteriori Hypothesis (MAP)	$(a_1^*, \ldots, a_k^*) = \arg\max_x \sum_{X/A} P(\bar{x}, \mathrm{e})$	$A \subseteq X$ Hypothesis variables
Finding Maximum Expected Utility (MEU) Decision	$(d_1^*, \ldots, d_k^*) = \arg\max_d \sum_{X/D} P(\bar{x}, \mathrm{e}) U(\bar{x})$	$D \subseteq X$ Decision variables $U(\bar{x})$ Utility function

Causal Reasoning

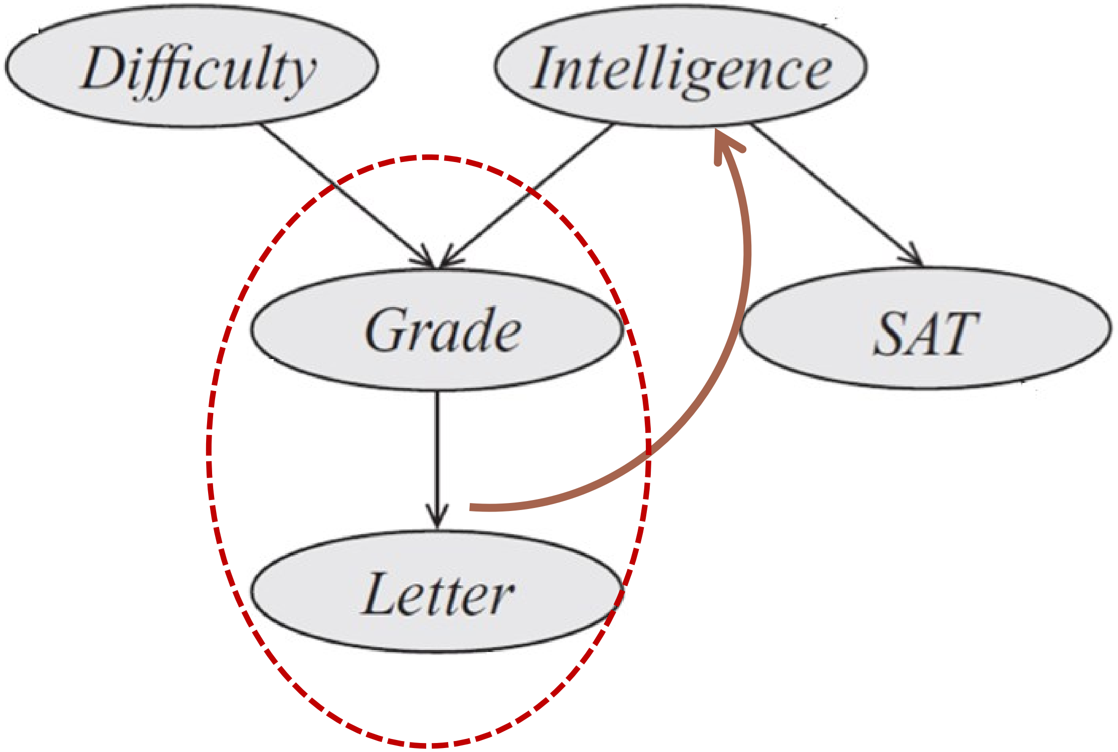
- Predicting downstream effects of factors
- Scenario:
  - How likely is a student to get a strong letter (knowing nothing else)?
    $$P(l^1) = 0.502$$
  - But this student is not so intelligent (i^0)
    $$P(l^1|i^0) = 0.389$$
  - Then we find out if the course is easy
    $$P(l^1|i^0, d^0) = 0.513$$



https://courses.csail.mit.edu/6.034s

Evidential Reasoning

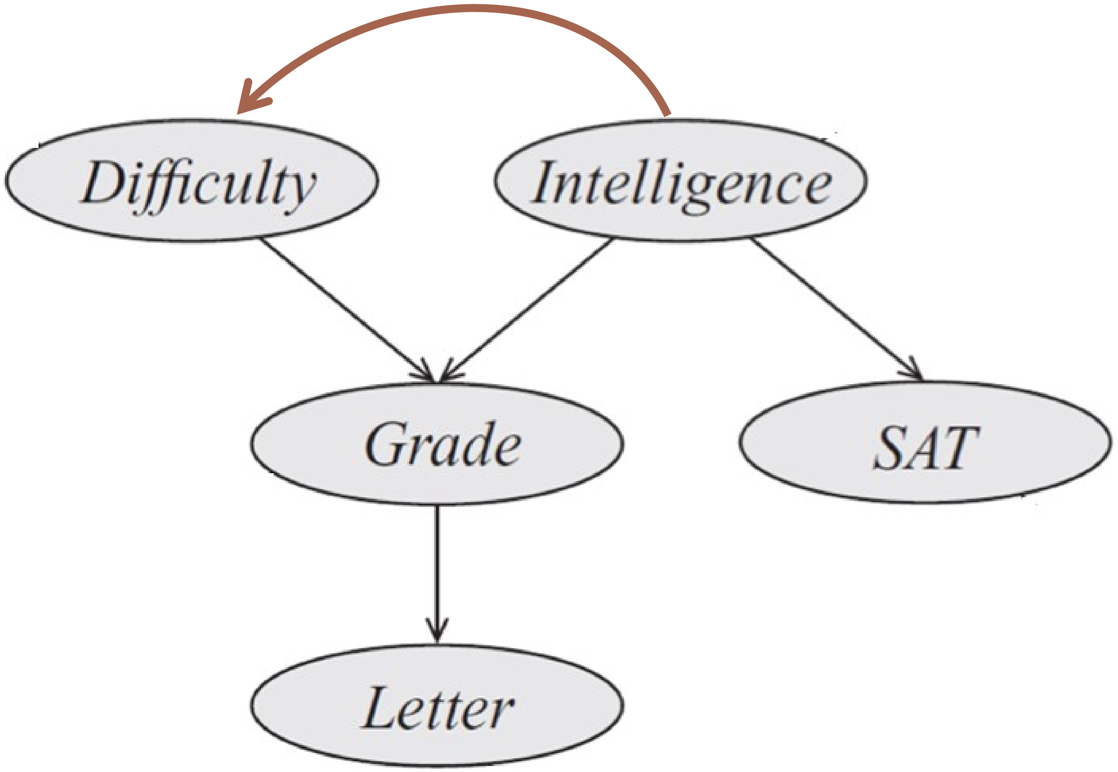
- Reasoning from effects to causes is called evidential reasoning
- Scenario:
  - Recruiter wants to hire intelligent student
  - A priori this student is 30% likely to be intelligent
    $$P(i^1) = 0.3$$
  - Finds that the student received grade C
    $$P(i^1|g^3) = 0.079$$
  - Probability that class is difficult
    $$P(\mathrm{d}^1|g^3) = 0.629$$
  - 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



https://courses.csail.mit.edu/6.034s

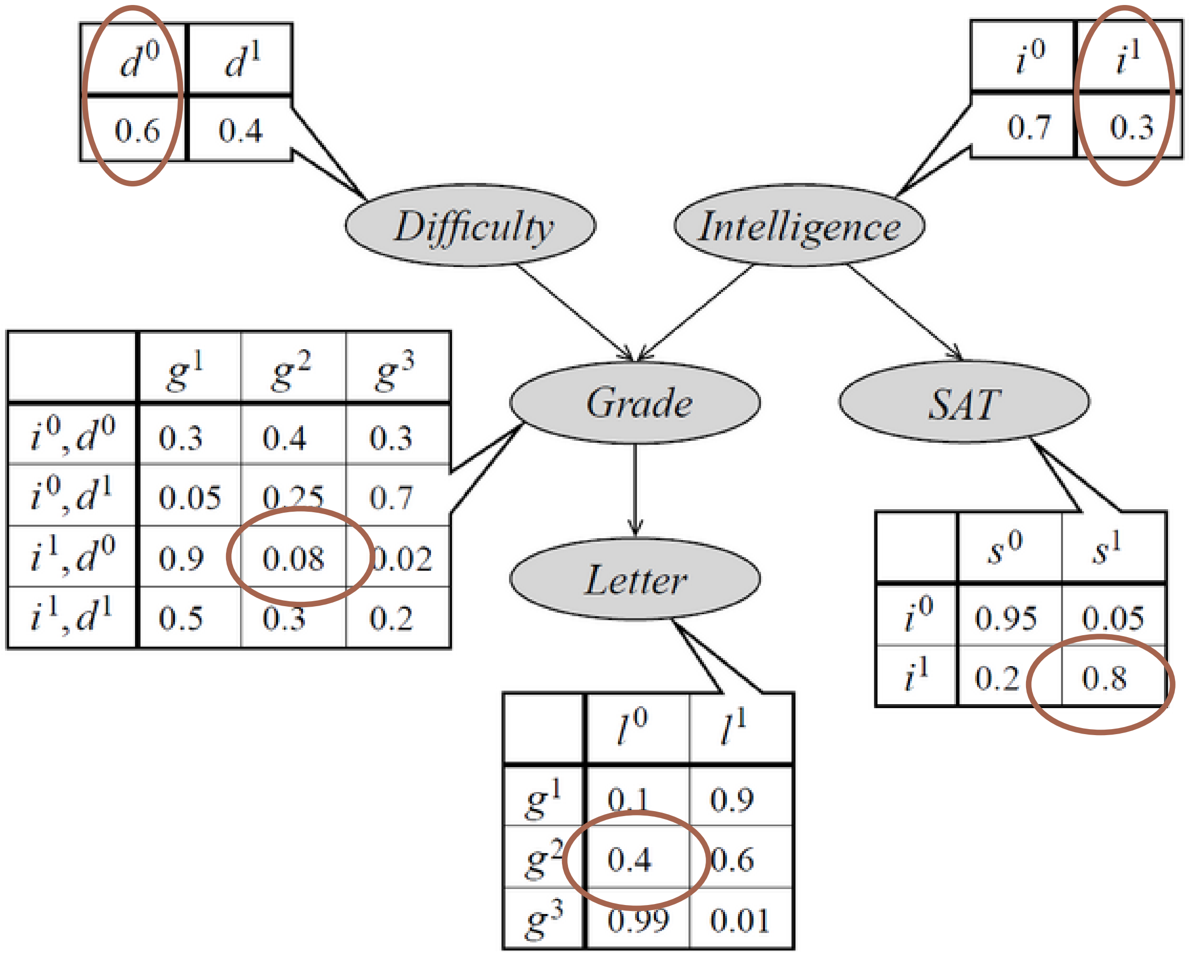
Intercausal Reasoning

- 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:
    $P(\mathrm{d}^1|g^3) = 0.629$ to $P(\mathrm{d}^1|g^3, s^1) = 0.76$
- Summary
  - Information about SAT score gave us information about Intelligence which with Grade told us about difficulty of course



https://courses.csail.mit.edu/6.034s

Bayes Net Inference: Example

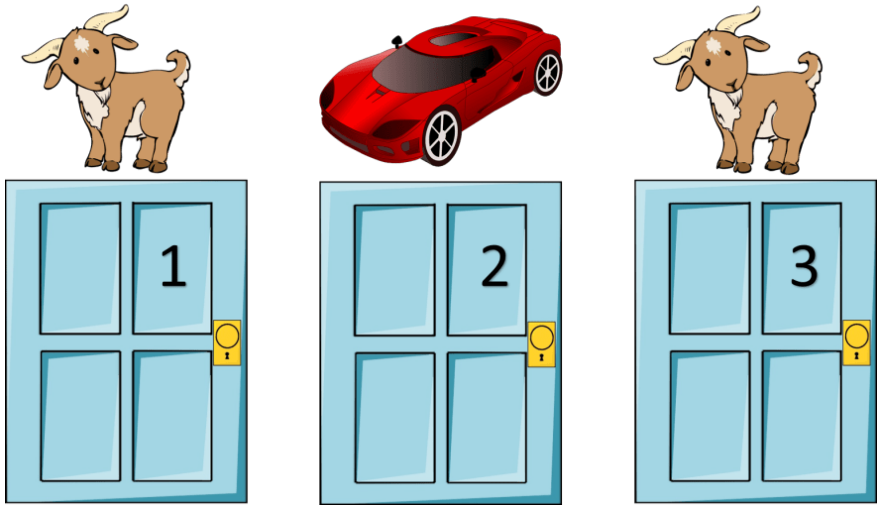


Question: What is the probability that an intelligent student gets a B in an easy class, a high SAT and a weak letter?

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

	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)

Monty Hall Problem



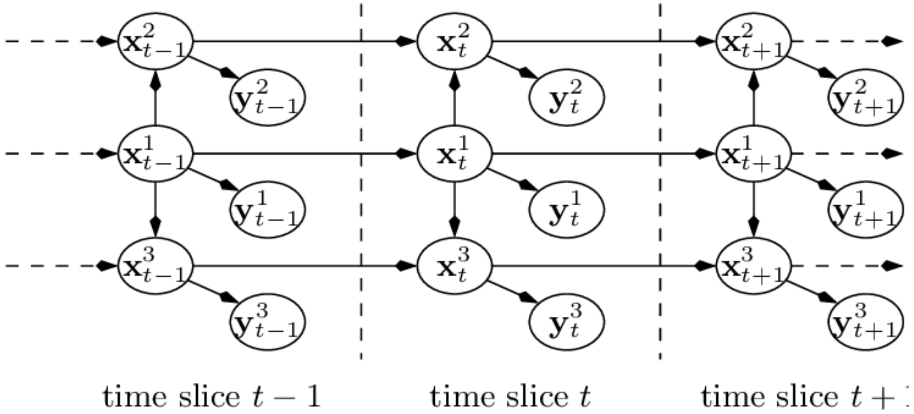
Summary

- 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
- BNs provide a formal and efficient framework for reasoning using probability theory considering multiple possible outcomes and their likelihood
- Bayesian networks can be learned from data.
- Hidden variables and not measurable quantities are major obstacles to causal discovery.

APPENDIX

Dynamic Bayesian Networks

- Dynamic BNs are used in domains that evolve over time
- Approach
  - A discrete time-stamp is introduced, and the same local model (time slice) is repeated for each unit of time. The local model represents a snapshot of the underlying evolving temporal process.
  - The nodes within time slice *t* can be connected to other nodes within the same slice. Also, time slices are inter-connected through temporal or transition arcs that specify how variables change from one time point to another.



https://www.frontiersin.org/articles/10.3389/fncom.2014.00131/full
https://www.researchgate.net/publication/220785551_Simultaneous_Partitioned_Sampling_for_Articulated_Object_Tracking

Decision Graphs

- Decision Graphs (or Influence Diagrams) extend Bayesian Networks by adding Utilities and Decisions
- They are used to find an optimal strategy (set of decisions) which maximizes the overall utility (e.g., profit/loss) of the network.
- Cost based decisions based on models learned from data and/or built from expert opinion.

https://www.bayesserver.com

https://www.bayesserver.com/docs/introduction/decision-graphs