Uncertain Knowledge and Reasoning

By Dr. Sonali Patil

Reasoning Under Uncertainty

- Uncertainty
- Sources of uncertainty
- Methods to handle Uncertainty
- Probability Theory
- Uncertainty and Rational Decisions
- Basic Probability Notations
- Probability Axioms
- Bayes Rule
- Inference using full joint distribution
- Independence
- Bayesian Networks
- Inferencing using Bayesian Networks

Uncertainty

Reasoning under uncertainty.

- There are different types of uncertainties. What are the different ways in which you can deal with that?

- The doorbell problem
  - The doorbell rang at 12 O'clock at midnight.
  - Que to ans
    - was someone there at the door?
    - Mohan was sleeping in the room. Did Mohan wake up when the doorbell rang?
    - My fact is that the doorbell rang at 12 O'clock in the midnight. Therefore if we place the propositions in the logic form

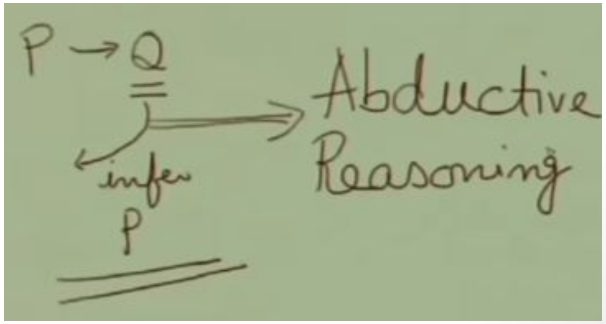
- Proposition 1: $AtDoor(x) \Rightarrow Doorbell$
- Proposition 2: $Doorbell \Rightarrow Wake(Mohan)$

Uncertainty

- Given Doorbell, can we say AtDoor(x), because

AtDoor(x) ➡ Doorbell?

- Can we say that there is some one at Door? We can using the deductive reasoning/normal implication (p implies q, if p true…Q is necessarily true, if p false…q may be or may not be true)
  - May not press the door bell button
- Abductive Reasoning (p implies q and we find q is true then we infer p. Most of the time right, but may not always). Other reasons, though rare
  - Short Circuit
  - Wind
  - Dog or other Animal pressed the button



Uncertainty

- Given Doorbell, can we say Wake(Mohan), because Doorbell ⇒ Wake(Mohan)?
- Using, Deductive Reasoning. Yes, if proposition 2 is always true

  However always this may not be true (May be tired and in sound sleep)
- Similarly, using Abductive Reasoning, Mohan wakes up does it mean that the bell has rang?

  This is also not true always. There may be other reasons for waking up

Hence, we cannot answer either of Questions with certainty.

- Proposition 1 is incomplete. Modifying it as

  AtDoor(x) ∨ ShortCkt ∨ Wind … ⇒ Doorbell

  Doesn't help because the list of possible causes on the left is huge (infinite??)
- Proposition 2 is *often* true, but not a tautology.

Uncertainty

Planning Example

- Let action A(t) denote leaving for the airport t minutes before the flight
  - For a given value of t, will A(t) get me there on time?
- Problems:
  - Partial observability (roads, other drivers' plans, etc.)
  - Noisy sensors (traffic reports)
  - Uncertainty in action outcomes (flat tire, etc.)
  - Immense complexity of modelling and predicting traffic

Uncertainty

- Diagnosis always involves uncertainty.
- Eg:

Dental diagnosis: (toothache)

Toothache $\Rightarrow$ *Cavity*

Its wrong as not all people with toothaches have cavity. It may be due other problems

Toothache $\Rightarrow$ *Cavity V Gum Problem V Abscess…….*

In order to complete the list, we have to add an almost unlimited list of possible problems

The causal rule for this:

Cavity $\Rightarrow$ *Toothache*

This is not also the right one. Not all cavities cause pain

Uncertainty

- Trying to cope up with domains like Medical diagnosis fails for 3 main reasons:

✔ Laziness:

Too hard to list out all antecedence & consequents needed to ensure an exception-less rule and too hard to use such rules.

✔ Theoretical ignorance:

The domains like Medical science has No complete theory for the domain.

✔ Practical ignorance:

Even all rules are known – uncertain about a particular patient. Not all test have been or can be run.

Uncertainty

- The problems like Doorbell/diagnosis are very common in real world
- In AI, we need to reason under such circumstances
- We solve such problems by proper modelling of Uncertainty and impreciseness and developing appropriate reasoning techniques

Sources of uncertainty

- Implications may be weak

$\text{doorbell } (0.8) \Rightarrow \text{wake(Mohan)}$

quantification of frequency with which the rule applies

- Imprecise language like often, rarely, sometimes
  - Need to quantify these terms of frequencies
  - Need to design rules for reasoning with these frequencies

- Precise information (input) may be too complex
  - Too many antecedents or consequents

$\text{AtDoor}(x) \vee \text{ShortCkt} \vee \text{Wind} \ldots \Rightarrow \text{Doorbell}$

- Incomplete Knowledge
  - We may not know or guess all the possible antecedents or consequents
  - The bell rang due to some spooky reason

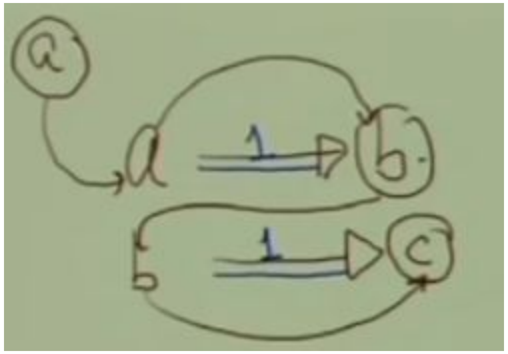
Sources of uncertainty

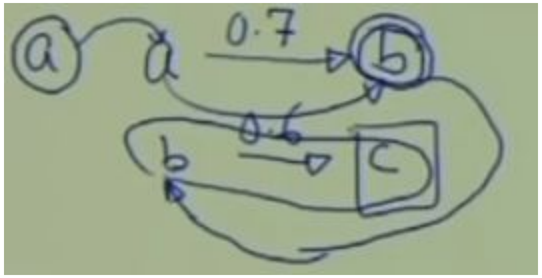
- Conflicting Information
  - Patient-complicated symptoms-two diff doctors-may be possible they differ in there diagnosis if the symptoms do not lead to a vary obvious disease

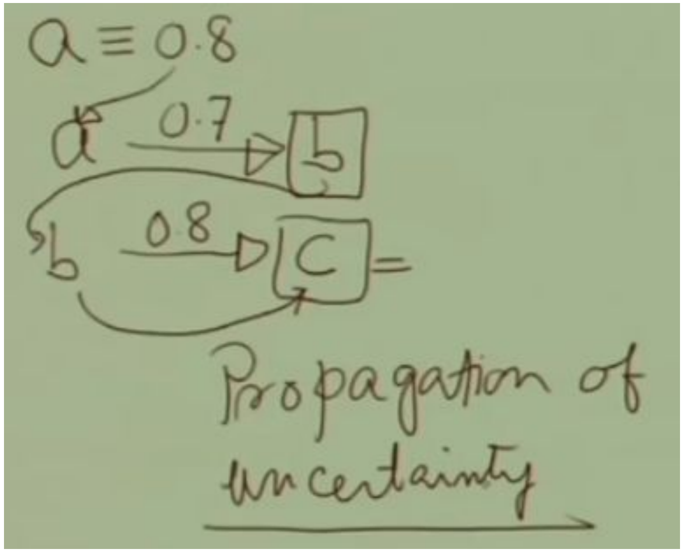
Experts often provide conflicting information:
quantification of measure of belief

- Propagation of Uncertainties
  - In absence of interdependencies of propagation of uncertain knowledge the uncertainty of the conclusions increases

Tomorrow(sunny) [0.6], Tomorrow(warm) [0.8]
Tomorrow(sunny) $\wedge$ Tomorrow(warm) [?]


a
a
1
b
b
1
c


a
a
0.7
b
0.6
b
c


a ≡ 0.8
a
0.7
b
b
0.8
C
Propagation of uncertainty

Methods of handling Uncertainty

- Fuzzy Logic
  - Logic that extends traditional 2-valued logic to be a continuous logic (values from 0 to 1)
    - while this early on was developed to handle natural language ambiguities such as "you are *very* tall" it instead is more successfully applied to device controllers
- Probabilistic Reasoning
  - Using probabilities as part of the data and using Bayes theorem or variants to reason over what is most likely
- Hidden Markov Models
  - A variant of probabilistic reasoning where internal states are not observable (so they are called hidden)
- Certainty Factors and Qualitative Fuzzy Logics
  - More ad hoc approaches (non formal) that might be more flexible or at least more human-like (MYCIN expert system)
- Neural Networks

Uncertainty tradeoffs

- **Bayesian networks:** Nice theoretical properties combined with efficient reasoning make BNs very popular; limited expressiveness, knowledge, engineering challenges may limit uses
- **Non-monotonic logic:** Represent commonsense reasoning, but can be computationally very expensive
- **Certainty factors:** Not semantically well founded
- **Fuzzy reasoning:** Semantics are unclear (fuzzy!), but has proved very useful for commercial applications

Probability Theory

- Deals with degrees of belief
- Provides a way of summarizing the uncertainty that comes from our laziness & ignorance thereby solving the qualification problem (specifying all exceptions)
  - A90 will take us to airport on time, as long as car doesn't break down or run out of gas, does not indulge into accident, no accidents on bridge, plane doesn't live early, no meteorite hits the car, and ….)
- Toothache $\xrightarrow{0.8}$ cavity
  - The probability that the patient has a cavity, given that she has a toothache is 0.8

Probability Theory

- Consider previous statement: “The probability that the patient has a cavity, given that she has a toothache is 0.8”
- If we later learn that patient has a history of gum disease we can say “The probability that the patient has a cavity, given that she has a toothache and a history of gum disease, is 0.4”
- If further we gather evidence, we can say “The probability that the patient has a cavity, given all we know now, is almost zero”
- Above three statements do not contradict each other; each is a separate assertion about a difference knowledge state

Uncertainty and rational decisions

- “Say A90 has 92% chance of catching our flight. Is it rational choice? Not necessarily
- A180 has higher probability of reaching. If its vital to not miss the flight, then its worth risking the longer wait time at airport
- A1440 almost guarantees reaching on time but I’d have to stay overnight in the airport (intolerable wait and may be unpleasant diet of airport food)
- To make choices, an agent must have preferences between different possible outcomes of various plans
- **Utility Theory** is used to represent & reason with preferences

Uncertainty and rational decisions

- **Utility Theory**
  - Every state has a degree of usefulness or utility, to an agent and the agent will prefer states with higher utility
  - The utility state is relative to agent
  - Ex. Consider the state in which White has checkmated Black in chess. Here, Utility is high for agent playing White but low for agent playing Black
  - A Utility function can account for any set of preferences- quirky or typical, noble or perverse

Uncertainty and rational decisions

- **Decision Theory**
  - Preferences as expressed by utilities, are combined with probabilities in general theory of rational decisions

  Decision Theory = Probability Theory + Utility Theory
- **Maximum Expected Utility(MEU)**
  - An agent is rational if and only if it chooses the action that yields the highest expected utility, averaged over all possible outcomes of the action. This is principle of MEU
  - Here the term expected is not vague. Its average or statistical mean of outcomes weighted by the probability of outcome
- The basic difference between A decision-theoretic agent & other agents is that the former's belief state represents not just the possibilities for world states but also their probabilities

Uncertainty and rational decisions

```
function DT-AGENT(percept) returns an action
   persistent: belief_state, probabilistic beliefs about the current state of the world
               action, the agent's action

   update belief_state based on action and percept
   calculate outcome probabilities for actions,
       given action descriptions and current belief_state
   select action with highest expected utility
       given probabilities of outcomes and utility information
   return action
```

A decision-theoretic agent that selects rational actions.

Uncertainty and rational decisions summary

- Rational behavior:
  - For each possible action, identify the possible outcomes
  - Compute the probability of each outcome
  - Compute the utility of each outcome
  - Compute the probability-weighted (expected) utility over possible outcomes for each action
  - Select the action with the highest expected utility (principle of **Maximum Expected Utility**)

Bayesian reasoning

- Probability theory
- Bayesian inference
  - Use probability theory and information about independence
  - Reason diagnostically (from evidence (effects) to conclusions (causes)) or causally (from causes to effects)
- Bayesian networks
  - Compact representation of probability distribution over a set of propositional random variables
  - Take advantage of independence relationships

Basic Probability Notation

- A **random variable** is a variable whose possible values are the numerical outcomes of a random experiment.
  - It is a function which associates a unique numerical value with every outcome of an experiment.
  - Its value varies with every trial of the experiment.
  - It describes an outcome that cannot be determined in advance
  - It is Boolean , Discrete or continuous
  - Ex. Roll of a die, number of emails received in a day etc.
- The **sample space S** of the random variable X is the set of all possible worlds
  - The possible worlds are mutually exclusive & exhaustive (at a time one possible outcome and all possible outcomes are in the S)
  - Tossing a coin: S={H,T}
  - Tossing two coins simultaneously S={HH, HT, TH, TT}
  - Rolling a die: S={ 1,2,3,4,5,6}

Basic Probability Notation

- An atomic event is a complete specification of the state of the world about which the agent is uncertain.
- Eg:

  Cavity & Toothache has four distinct atomic events.

  $Cavity = False \land Toothache = True$

  $Cavity = True \land Toothache = True$

  $Cavity = False \land Toothache = False$

  $Cavity = True \land Toothache = False$

Basic Probability Notation

- The sample space is denoted by Ω (upper case omega) and elements in sample space are denoted by ω (lower case omega)
- P(ω) -> Probability of occurrence of ω

$$0 \leq P(\omega) \leq 1 \text{ for every } \omega \text{ and } \sum_{\omega \in \Omega} P(\omega) = 1 \,.$$

- Probabilistic assertions & queries are not about particular possible worlds, but about sets of them
  - The two dice add upto 11, Doubles are rolled; Picking ace from pack of cards, number of email> 100 in a day, etc.
  - These sets are called **events.** Event is subset of ω
  - Events are described by proposition in common language

Basic Probability Notation

- The probability associated with the proposition is defined to be the sum of the probabilities of the world in which For any proposition ϕ, $P(\phi) = \sum_{\omega \in \phi} P(\omega)$.

- ϕ is getting odd number after rolling the dice

  S={1, 2, 3, 4, 5, 6}, ϕ ={1, 3, 5}

  P(Odd)=P(1)+ P(3) + P(5)= 1/6 + 1/6 + 1/6 = ½

Unconditional or Prior Probabilities

- Degree of belief in proposition in the absence of any other information/evidence
- P(Fever)=0.1
  - The probability that the patient has fever is 0.1(in absence of any other information
- A die is rolled, P(odd), P(even) indicated the probability of getting the even number and the probability of getting the even number on the rolled dice respectively. Both of these are prior probabilities
- When a pair of dice rolled simultaneously, the possible outcomes are 36. P(doubles), P(Total=15) are prior probabilities

Unconditional or Prior Probabilities

- The random variables Fever, Doubles, Odd, Even are **Discrete Random variables** as they take finite number of distinct values
- The **Boolean random** variables have values True or false ex. P(cavity)
- A **continuous random variable** is a random variable that takes infinite number of distinct values
  - EX. $P(Temp=x) = Uniform_{[18C,26C]}(x)$
  - Expresses that the temperature is distributed uniformly between 10 and 26 degrees
  - This is called Probability Density Function

Conditional or Posterior Probabilities

- Let A be an event in the world and B be another event. Suppose that events A and B are not mutually exclusive, but occur conditionally on the occurrence of the other. **The probability that event A will occur if event B occurs is called the conditional probability**. Conditional probability is denoted mathematically as $p(A|B)$ in which the vertical bar represents GIVEN and the complete probability expression is interpreted as "Conditional probability of event A occurring given that event B has occurred".

$$p(A|B) = \frac{\text{the number of times } A \text{ and } B \text{ can occur}}{\text{the number of times } B \text{ can occur}}$$

Conditional or Posterior Probabilities

- The number of times A and B can occur, or the probability that both A and B will occur, is called the **joint probability** of A and B. It is represented mathematically as $p(A \cap B)$. The number of ways B can occur is the probability of B, $p(B)$, and thus

$$p(A|B) = \frac{p(A \cap B)}{p(B)}$$

- The eq. of conditional can also be written in the form

$$p(A \cap B) = p(A|B)\ p(B) \Rightarrow$$ **Product Rule**

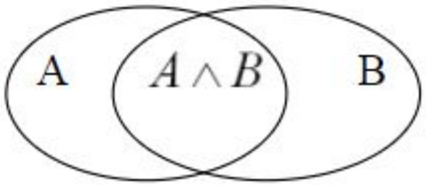
- Similarly, the conditional probability of event B occurring given that event A has occurred equals

$$p(B|A) = \frac{p(B \cap A)}{p(A)}$$

$$p(B \cap A) = p(B|A)\, p(A) \Rightarrow$$ **Product Rule**

Probability Axioms

- All probabilities are between 0 & 1
  - $0 \leq P(A) \leq 1$
- Necessarily True propositions have probability 1 and necessarily false propositions have probability 0
  - ($P(true) = 1$ and $P(false) = 0$)
- Probability of disjunction
  - $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$ ⇨ **Inclusion-Exclusion Principle**



- These axioms often called as Kolmogorov’s axiom

Probability Axioms

- From Axioms we can derive other properties
- $P(A \vee B) = P(A) + P(B) - P(A \wedge B)$ Substitute $B = \neg A$
  - $P(A \vee \neg A) = P(A) + P(\neg A) - P(A \wedge \neg A)$

    $1 \quad = P(A) + P(\neg A) – 0$

    $P(\neg A)= 1- P(A)$

    $P(A) = 1- P(\neg A)$
- A and B mutually exclusive $P(A \vee B) = P(A) + P(B)$

  $P(e_1 \vee e_2 \vee e_3 \vee \ldots e_n) = P(e_1) + P(e_2) + P(e_3) + \ldots + P(e_n)$

  The probability of a proposition **a** is equal to the sum of the probabilities of the atomic events in which **a** holds

  e(a) – the set of atomic events in which **a** holds

  $$P(a) = \sum_{e_i \in e(a)} P(e_i)$$

Bayesian or Bayes Rule

- Let A be an event in the world and B be another event.

  Hence from product rule

  $P(A \wedge B) = P(A|B) * P(B)$

  $P(A \wedge B) = P(B|A) * P(A)$

- LHS are same. Equating the RHS of both equations yields

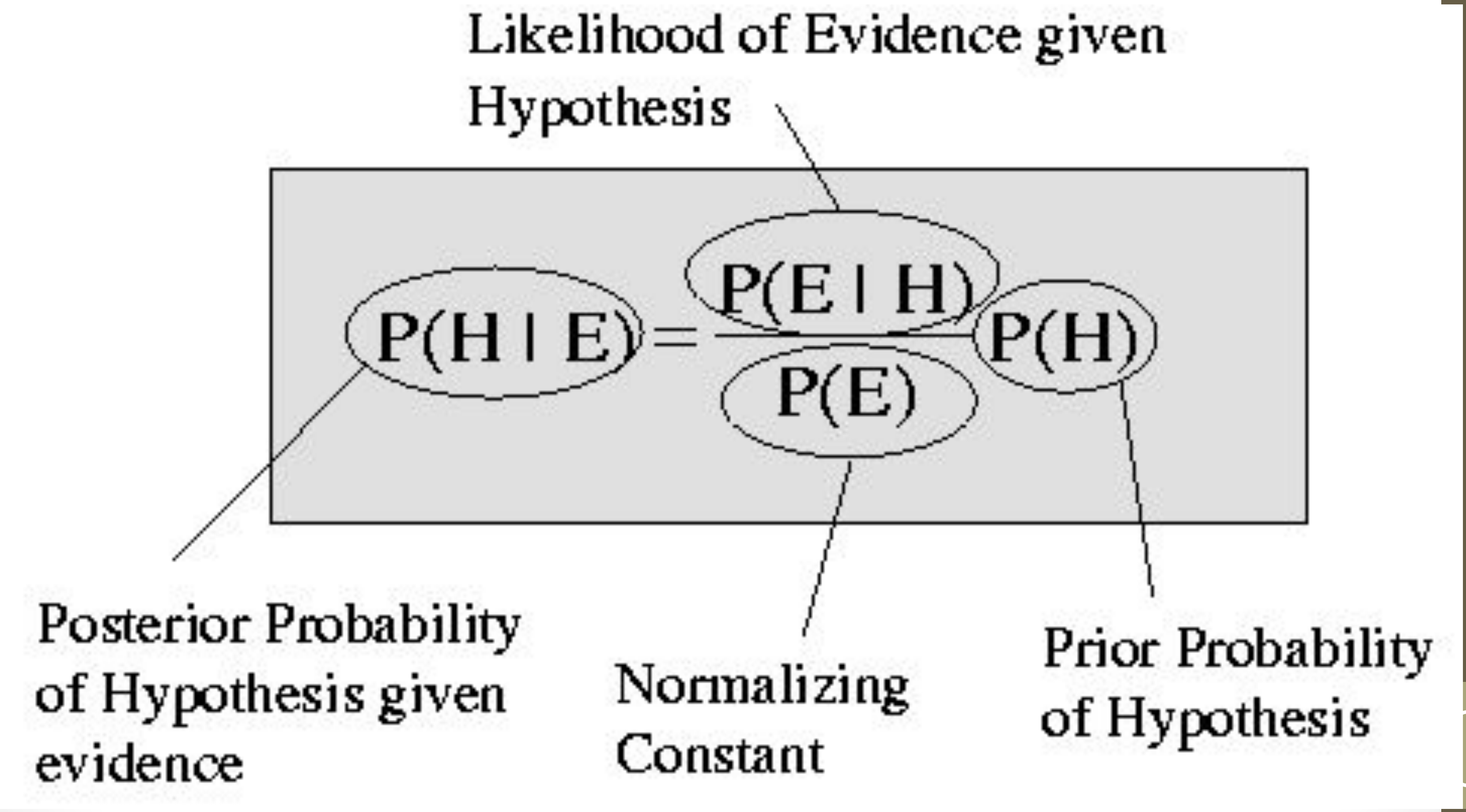
$$P(A|B) = P(B \mid A) * P(A) / P(B)$$

$$P(B|A) = P(A \mid B) * P(B) / P(A)$$

Bayes rule/ Bayesian rule

where:
p(A|B) is the conditional probability that event A occurs given that event B has occurred;
p(B|A) is the conditional probability of event B occurring given that event A has occurred;
p(A) is the probability of event A occurring;
p(B) is the probability of event B occurring.

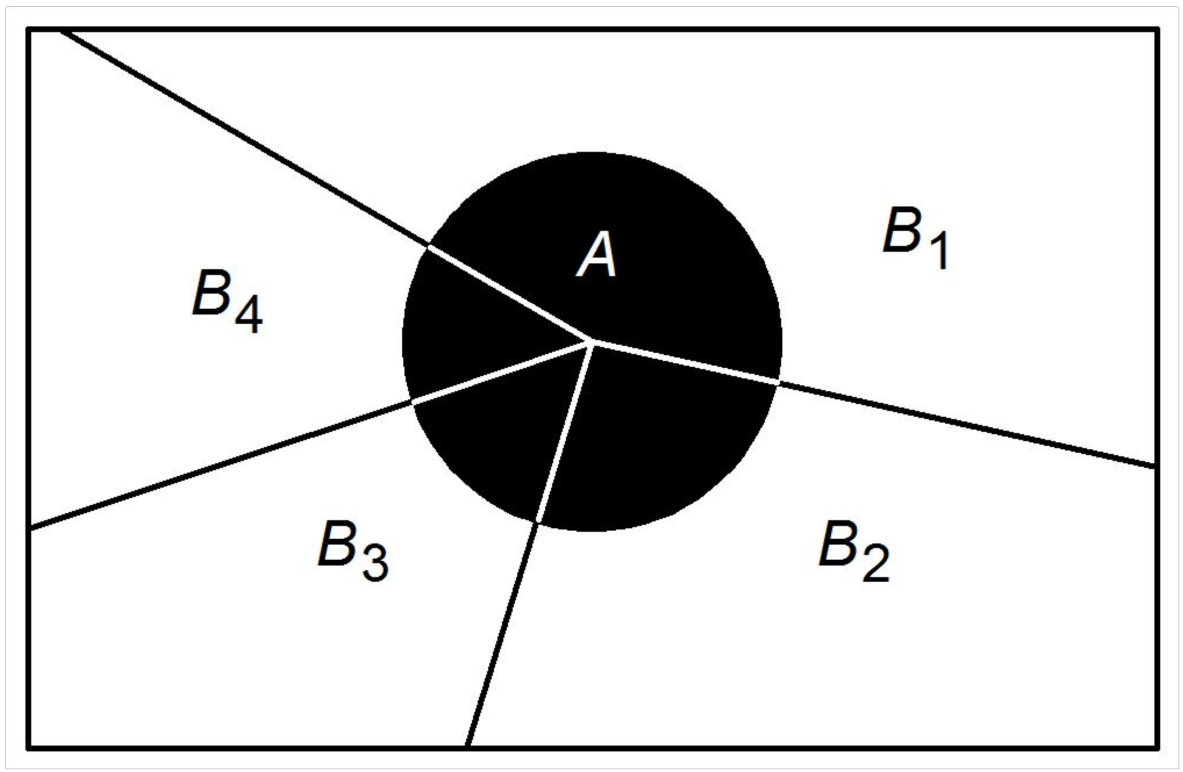
Bayesian or Bayes Rule
(Hypothesis-Evidence)



Bayesian or Bayes Rule

The Joint Probability

$$\sum_{i=1}^{n} p(A \cap B_i) = \sum_{i=1}^{n} p(A|B_i) \times p(B_i)$$



Bayesian or Bayes Rule

- If the occurrence of event A depends on only two mutually exclusive events, B and NOT B , we obtain:

$$p(A) = p(A|B) \times p(B) + p(A|\neg B) \times p(\neg B)$$

where $\neg$ is the logical function NOT.

Similarly,

$$p(B) = p(B|A) \times p(A) + p(B|\neg A) \times p(\neg A)$$

Substituting this equation into the Bayesian rule yields:

$$p(A|B) = \frac{p(B|A) \times p(A)}{p(B|A) \times p(A) + p(B|\neg A) \times p(\neg A)}$$

Bayesian or Bayes Rule

- The Bayesian rule expressed in terms of hypotheses and evidence looks like this:

$$p(H|E) = \frac{p(E|H) \times p(H)}{p(E|H) \times p(H) + p(E|\neg H) \times p(\neg H)}$$

where:

p(H) is the prior probability of hypothesis H being true;

p(E|H) is the probability that hypothesis H being true will result in evidence E; p(H) is the prior probability of hypothesis H being false;

p(E| H) is the probability of finding evidence E even when hypothesis H is false.

Example 2: Bayes' rule Meningitis

Example : Bayes' rule

Problem: Meningitis

- A doctor knows that the disease meningitis causes the patient to have a stiff neck, say, 50% of the time. The doctor also knows some unconditional facts: the prior probability of a patient having meningitis is 1/50,000, and the prior probability of any patient having a stiff neck is 1/20.
- Disease Meningitis:
- It Cause patient to have stiff neck- 50% of the time.
- Prior Probability that Patients has meningitis is 1/50000.
- Prior Probability that patient has stiff neck is 1/20.

- Let s be stiff neck & m be Meningitis.

  $P(s|m) = 0.5$

  $P(m) = 1/50000$

  $P(s) = 1/20$

  $$P(m|s) = \frac{P(s|m)\,P(m)}{P(s)}$$

  $$= \frac{0.5 * 1/50000}{1/20} = 0.0002.$$

- 1 in 5000 patients with stiff neck to have Meningitis.

Example: Bayes Rule

- In the meningitis case, perhaps the doctor knows that 1 out of 5000 patients with stiff necks has meningitis, and therefore has no need to use Bayes' rule.
- Unfortunately, diagnostic knowledge is often more tenuous than causal knowledge.
- If there is a sudden epidemic of meningitis, the unconditional probability of meningitis, P(M), will go up.
- The doctor who derived P(M\S) from statistical observation of patients before the epidemic will have no idea how to update the value, but the doctor who computes P(M\S) from the other three values will see that P(M\S) should go up proportionately to P(M).
- Most importantly, the causal information P(S\M) is unaffected by the epidemic, because it simply reflects the way meningitis works.
- The use of this kind of direct causal or model-based knowledge provides the crucial robustness needed to make probabilistic systems feasible in the real world.

Example: Bayes Rule

Problem: Does Patient has cancer or not?

A patient takes a lab test and the results come back as positive. The test returns a correct positive in only 98% of cases in which the disease is actually present and a correct negative result in only 97% of the cases in which disease is not present. Further more 0.008 of entire population have this cancer.

Needs to find:

Example: Bayes Rule

Problem: Does Patient has cancer or not?

A patient takes a lab test and the **results come back as positive**. The test returns a correct positive in only 98% of cases in which the disease is actually present and a correct negative result in only 97% of the cases in which disease is not present. Further more 0.008 of entire population have this cancer.

Needs to find:

1. Probability of patient having cancer given that the test is positive i.e. P (cancer| test +),
2. Probability of patient not having cancer given that the test is positive i.e. P ($\neg$ cancer| test +)
3. See which is more probable

Example: Bayes Rule

Problem: Does Patient has cancer or not?

A patient takes a lab test and the **results come back as Negative**. The test returns a correct positive in only 98% of cases in which the disease is actually present and a correct negative result in only 97% of the cases in which disease is not present. Further more 0.008 of entire population have this cancer.

Needs to find:

Example: Bayes Rule

$$P(\text{Cancer} \mid \text{Test+}) = \frac{P(\text{Test+} \mid \text{Cancer})P(\text{Cancer})}{P(\text{Test +})}$$

$P(\text{Test+} \mid \text{Cancer}) = 0.9 \quad P(\text{Test}- \mid \text{Cancer}) = 0.1$

$P(\text{Test+} \mid \text{No Cancer}) = 0.01 \quad P(\text{Test}- \mid \text{Cancer}) = 0.99$

$P(\text{Cancer}) = 0.0001$

$$\begin{aligned} P(\text{Test+}) &= P(\text{Test+} \mid \text{Cancer})P(\text{Cancer}) + \\ &\quad P(\text{Test+} \mid \text{No cancer})P(\text{No Cancer}) \\ &= 0.9\text{x}0.0001 + 0.01\text{x}(1-0.0001) \\ &= 0.0010899 \end{aligned}$$

Example: Bayes Rule

$$P(\text{Cancer} \mid \text{Test+}) = \frac{P(\text{Test+} \mid \text{Cancer})P(\text{Cancer})}{P(\text{Test +})}$$

$P(\text{Test+} \mid \text{Cancer}) = 0.9$

$P(\text{Test+} \mid \text{No Cancer}) = 0.01$

$P(\text{Cancer})=0.0001$

$P(\text{Test+})= 0.0010899$

$$P(\text{Cancer} \mid \text{Test+}) = 0.9\text{x}0.0001/0.0010899$$
$$= 0.08$$

Inference using Full Joint Distribution

- The joint probability distribution $P(X_1,\ldots\ldots..,Xn)$ assigns probabilities to all possible atomic events.
- Recall that P(Xi) is a one-dimensional vector of probabilities for the possible values of the variable Xi. Then the joint is an n-dimensional table with a value in every cell giving the probability of that specific state occurring.
- The joint probability distribution for the trivial medical domain consisting of the two Boolean variables Toothache and Cavity:

	Toothache	$\neg$*Toothache*
Cavity	0.04	0.06
$\neg$*Cavity*	0.01	0.89

Inference using Full Joint Distribution

Probability distribution $\mathbf{P}$(Cavity, Toothache)

	Toothache	**~ Toothache**
Cavity	0.04	0.06
~ Cavity	0.01	0.89

Sum of all entries =1

P(Cavity) =

P(Cavity $\vee$ Toothache) =

P(Cavity|Toothache) =

Obtain P(~ cavity), P(Toothache), P(~ Toothache), P(cavity| ~ toothache) P(~ cavity| toothache), P(~ cavity| ~ toothache),

Inference using Full Joint Distribution

Probability distribution **$\mathbf{P}$(Cavity, Toothache)**

	Toothache	**~ Toothache**
Cavity	0.04	0.06
~ Cavity	0.01	0.89

Sum of all entries =1

P(Cavity) = 0.04 + 0.06 = 0.1 (using Axioms)

P(Cavity $\vee$ Toothache) = 0.04 + 0.01 + 0.06 = 0.11

P(Cavity|Toothache) =
(Cavity $\wedge$ Toothache)/P(Toothache)
= 0.04 / (0.04 + 0.01)
= 0.8

- Obtain P(~ cavity), P(Toothache), P(~ Toothache), P(cavity| ~ toothache) P(~ cavity| toothache), P(~ cavity| ~ toothache),

Inference using Full Joint Distribution

Start with the joint distribution $\mathbf{P}(Cavity, Catch, Toothache)$:

	toothache		¬ toothache	
	catch	¬ catch	catch	¬ catch
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

For any proposition ϕ, sum the atomic events where it is true:

$$P(\phi) = \sum_{\omega : \omega \models \phi} P(\omega)$$

	toothache		¬ toothache	
	catch	¬ catch	catch	¬ catch
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

$$P(cavity \lor toothache) = 0.108 + 0.012 + 0.072 + 0.008 + \\ + 0.016 + 0.064 = 0.28$$

	toothache		¬ toothache	
	catch	¬ catch	catch	¬ catch
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

$$P(toothache) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

- Process- Marginalization or summing out.

	toothache		¬ toothache	
	catch	¬ catch	catch	¬ catch
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

You can also compute conditional probabilities:

$$\begin{aligned} P(\neg cavity | toothache) &= \frac{P(\neg cavity \land toothache)}{P(toothache)} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4 \end{aligned}$$

$$P(cavity | toothache) = ?$$

Inference using Full Joint Distribution

P(cavity|toothache)= P(cavity^ Toothache)/ P(Toothache)
= (0.108+0.012)/(0.108+0.012+0.016+0.064)
= 0.6

- Observe, P(cavity|toothache)+P(~ cavity|toothache) =0.6 + 0.4=1 as it should be
- 1/P(toothache) remains constant no matter which value of cavity we calculate. Such constants in probability are called as normalization constant

Inference using Full Joint Distribution

Normalization

	toothache		¬ *toothache*	
	catch	¬ *catch*	*catch*	¬ *catch*
cavity	.108	.012	.072	.008
¬ *cavity*	.016	.064	.144	.576

Denominator can be viewed as a normalization constant α

$$\begin{aligned}
\mathbf{P}(Cavity|toothache) &= \alpha\,\mathbf{P}(Cavity, toothache) \\
&= \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)] \\
&= \alpha\,[\langle 0.108, 0.016\rangle + \langle 0.012, 0.064\rangle] \\
&= \alpha\,\langle 0.12, 0.08\rangle = \langle 0.6, 0.4\rangle
\end{aligned}$$

General idea: compute distribution on query variable by fixing evidence variables and summing over hidden variables

Independence

A and B are independent iff

$\mathbf{P}(A|B) = \mathbf{P}(A)$ or $\mathbf{P}(B|A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\mathbf{P}(B)$

- Independence is simplifying the modelling assumption

Example independence

- N fair, independent coin flips:

$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

H	0.5
T	0.5

$\cdots$

$P(X_n)$

H	0.5
T	0.5

$P(X_1, X_2, \ldots X_n)$

Variable represented for probability are

P(Weather, toothache , catch , cavity)

It can be deduced as

P(weather= cloudy) P(toothache ,catch ,cavity)

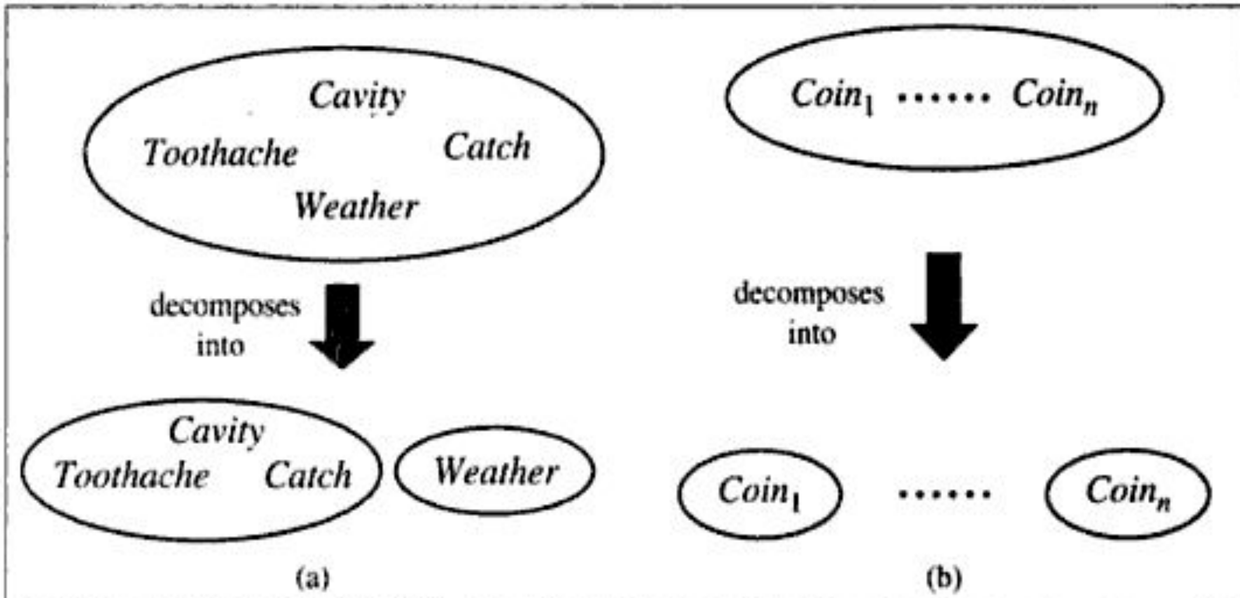


Figure 13.5 Two examples of factoring a large joint distribution into smaller distributions, using absolute independence. (a) Weather and dental problems are independent. (b) Coin flips are independent.

32 entries reduced to 12; for n independent biased coins, $2^n \to n$

Absolute independence powerful but rare

How to verify Independence?

$P_1(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P(T)$

T	P
hot	0.5
cold	0.5

$P(W)$

W	P
sun	0.6
rain	0.4

- Given a joint distribution P_1(T,W) how to verify T and W are independent or not
- Build marginals for each of the variables. Her two variables so two marginals

How to verify Independence?

$P(T)$

T	P
hot	0.5
cold	0.5

$P_1(T, W)$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$P_2(T, W)$

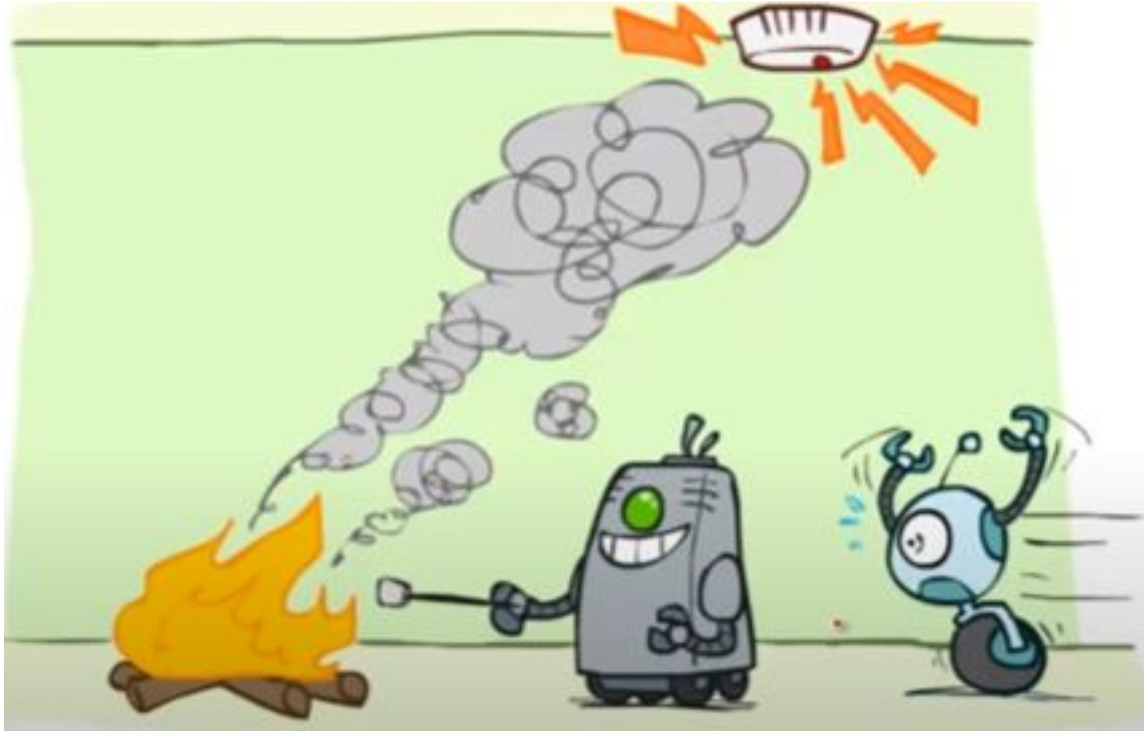
T	W	P
hot	sun	0.3
hot	rain	0.2
cold	sun	0.3
cold	rain	0.2

$P(W)$

W	P
sun	0.6
rain	0.4

- Calculate another distribution P2(T,W) as P(T)*P(W)
- If P1(T,W)= P2(T,W)… T and W are independent

Conditional Independence



Conditional Independence

- Unconditional (absolute) independence very rare (why?)
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.
- X is conditionally independent of Y given Z $X \perp\!\!\!\perp Y | Z$

if and only if:

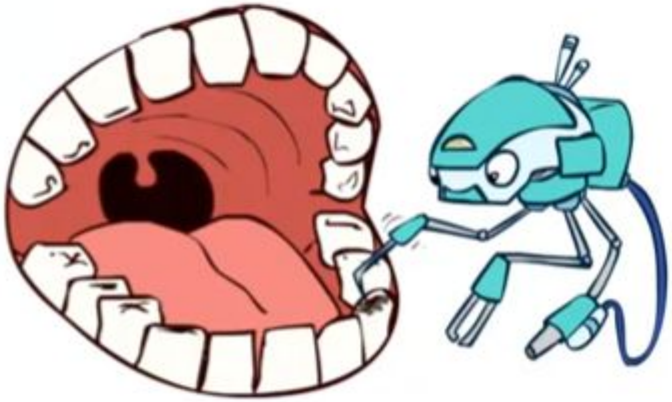
$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently, if and only if

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

Conditional Independence

- P(Toothache, Cavity, Catch)
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - P(+catch | +toothache, +cavity) = P(+catch | +cavity)
- The same independence holds if I don't have a cavity:
  - P(+catch | +toothache, -cavity) = P(+catch| -cavity)
- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)
- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
  - One can be derived from the other easily



Conditional Independence

$\mathbf{P}(Toothache, Cavity, Catch)$ has $2^3 - 1 = 7$ independent entries (or: parameters)

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

(1) $P(catch|toothache, cavity) = P(catch|cavity)$

The same independence holds if I haven't got a cavity:

(2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$

So $Catch$ is conditionally independent of $Toothache$ given $Cavity$:

$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$

Equivalent statements:

$\mathbf{P}(Toothache|Catch, Cavity) = \mathbf{P}(Toothache|Cavity)$

$\mathbf{P}(Toothache, Catch|Cavity) = \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)$

Note that (conditional) independence is symmetric!

Conditional Independence

Write out full joint distribution using chain rule:

$$\mathbf{P}(Toothache, Catch, Cavity)$$
$$= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch, Cavity)$$
$$= \mathbf{P}(Toothache|Catch, Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)$$
$$= \mathbf{P}(Toothache|Cavity)\mathbf{P}(Catch|Cavity)\mathbf{P}(Cavity)$$

I.e., 2 + 2 + 1 = 5 independent numbers (equations 1 and 2 remove 2)

In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.

Conditional independence is our most basic and robust form of knowledge about uncertain environments.

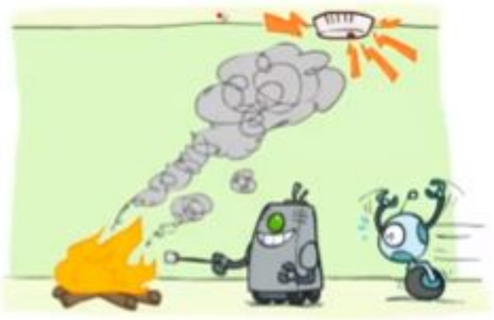
But how then, do we compute e.g. $\mathbf{P}(Cavity|Toothache)$?

Conditional Independence

- What about this domain:
  - Traffic
  - Umbrella
  - Raining

Conditional Independence

- What about this domain:
  - Fire
  - Smoke
  - Alarm



Conditional Independence & Chain Rule

- Chain rule: $P(X_1, X_2, \ldots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\ldots$

- Trivial decomposition:

$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) = P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain}, \text{Traffic})$$

- With assumption of conditional independence:

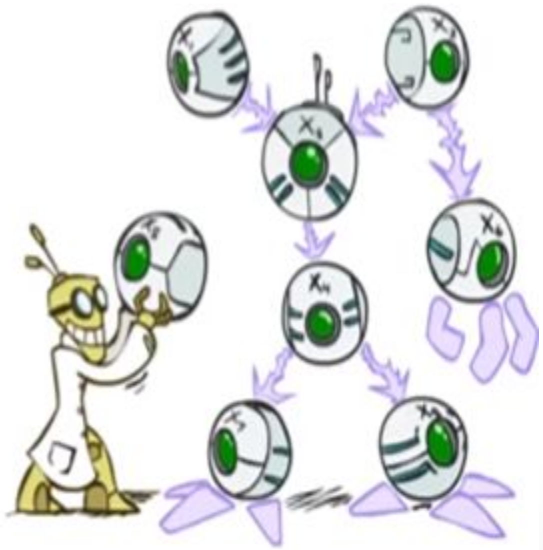
$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) = P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})$$

- Bayes'nets / graphical models help us express conditional independence assumptions

Bayes' Nets: Big Picture

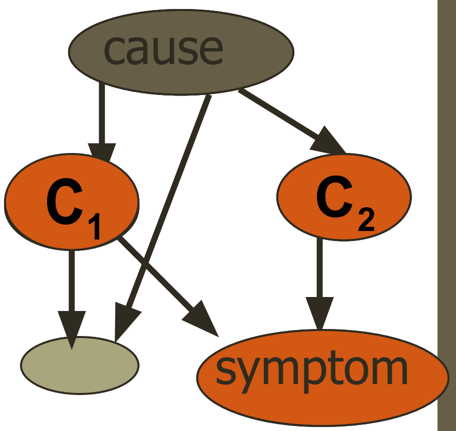
- Two problems with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time

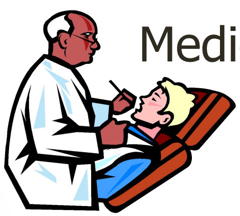
- Bayes' nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - More properly called graphical models
  - We describe how variables locally interact
  - Local interactions chain together to give global, indirect interactions



What Bayesian Networks are good for?

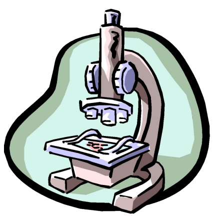
- Diagnosis: P(cause|symptom)=?
- Prediction: P(symptom|cause)=?
- Classification: $\max_{class} P(class|data)$
- Decision-making (given a cost function)





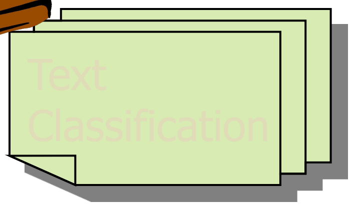
Medicine

Speech recognition

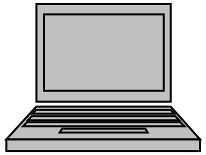


Bio-informatics

Stock market

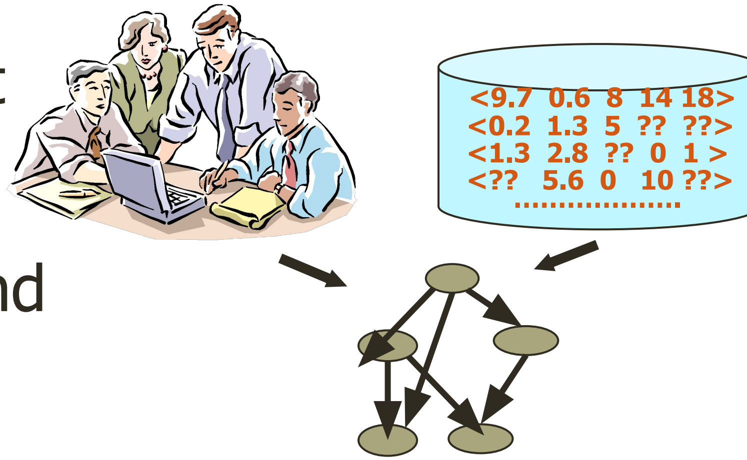


Computer troubleshooting



Why learn Bayesian networks?

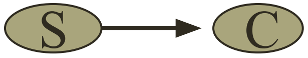
- Combining domain expert knowledge with data



- Efficient representation and inference

- Incremental learning 

- Handling missing data: **<1.3 2.8 ?? 0 1 >**

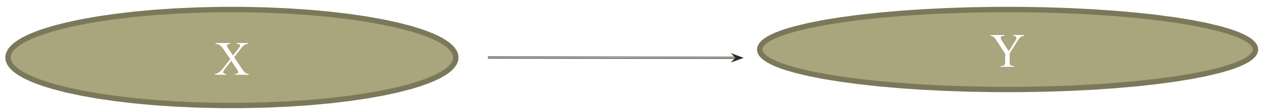
- Learning causal relationships: 

Probabilistic reasoning- Bayesian network

- **Bayesian network is a systematic way to represent independence relationships explicitly.**
- Data structure to represent the dependencies among variables.
- **Directed graph** – each node is annotated with quantitative probability information.

Specification of Bayesian network

1. A set of random variables makes up the node.
2. A set of directed links or arrows connects pair of nodes.

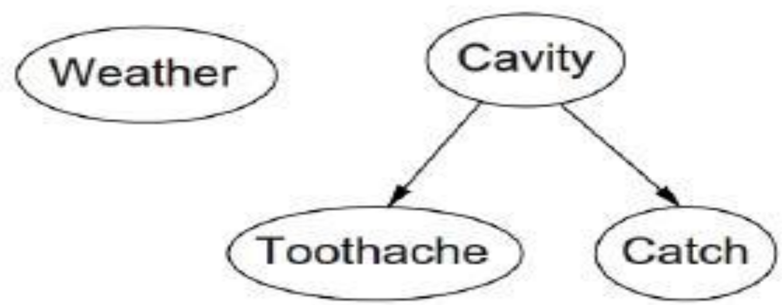


X is the Parent of Y.

3. Each node X_i has a conditional probability distribution $P(X_i|Parent(X_i))$
4. No directed cycles.

Topology

- Nodes & Links – specifies the conditional independence relationships.
- Variables *Weather ,Toothache ,Catch ,Cavity* .



Example: Traffic

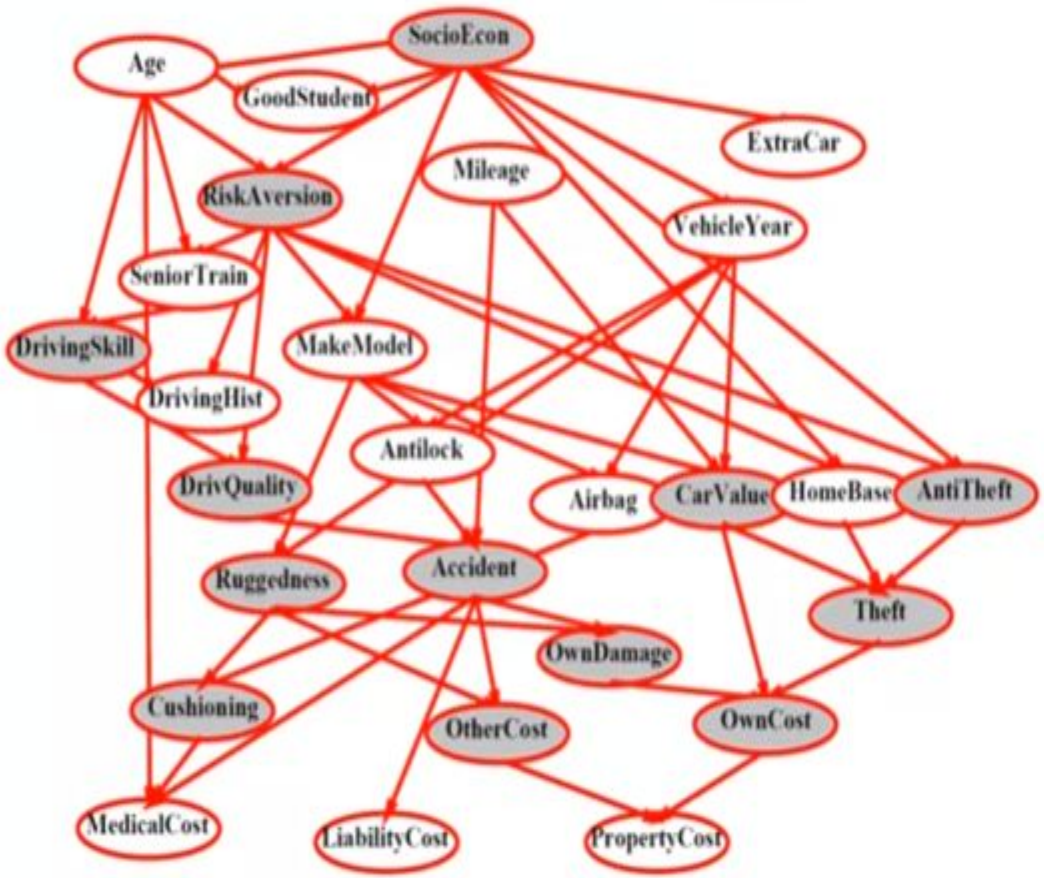
- Variables:
  - R: It rains
  - T: There is traffic

- Model 1: independence

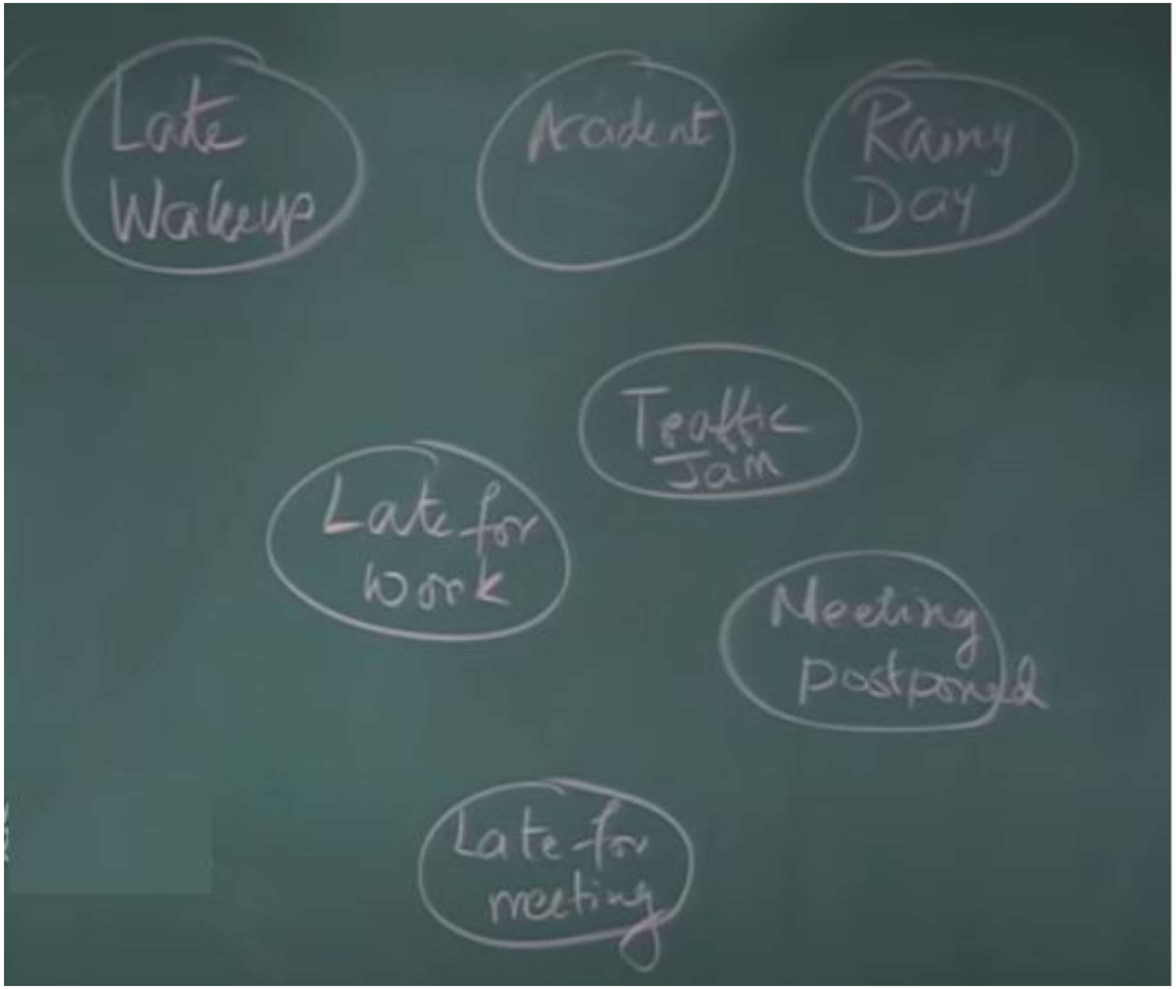
- Model 2: rain causes traffic

- Why is an agent using model 2 better?

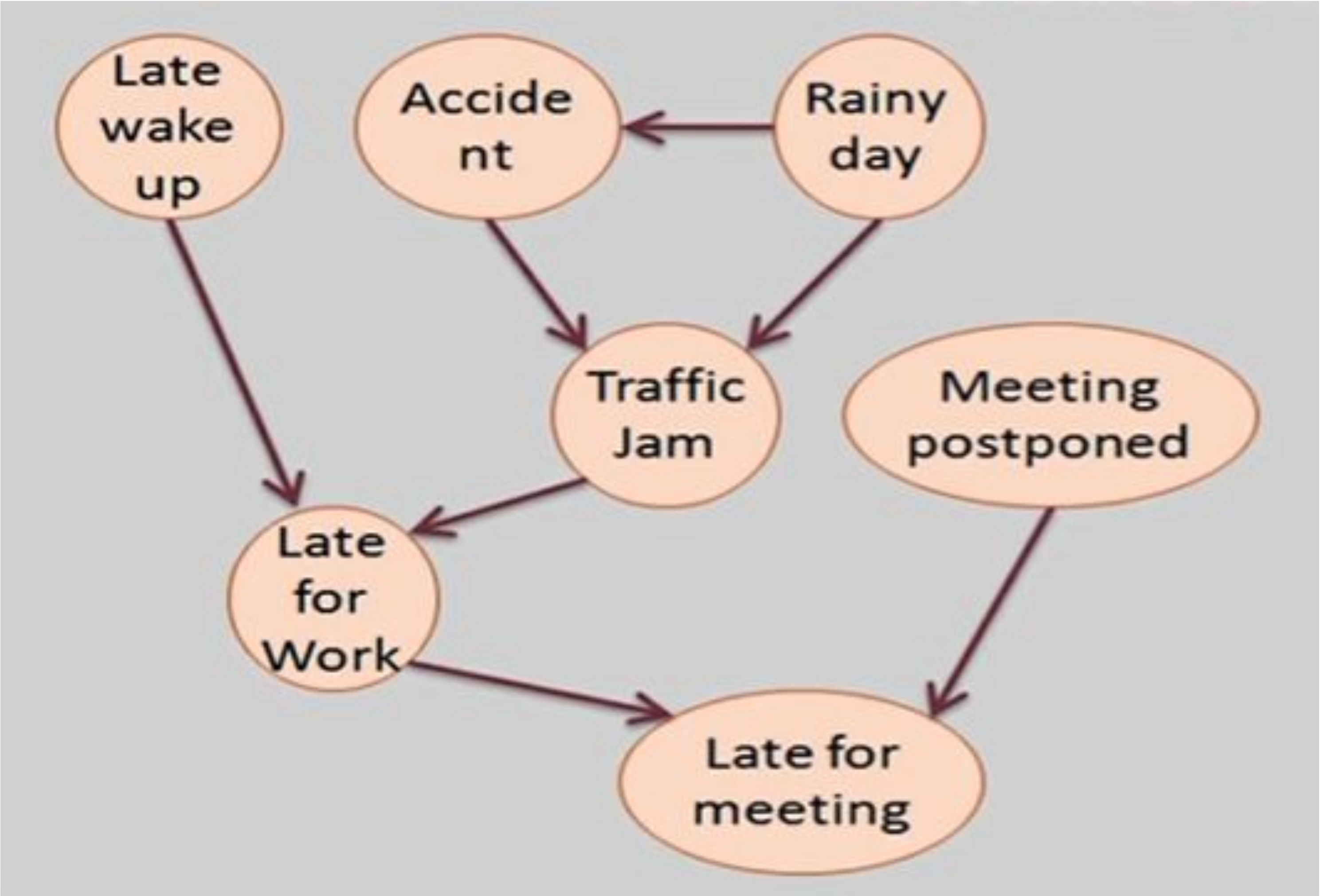
Example Bayes' Net: Insurance

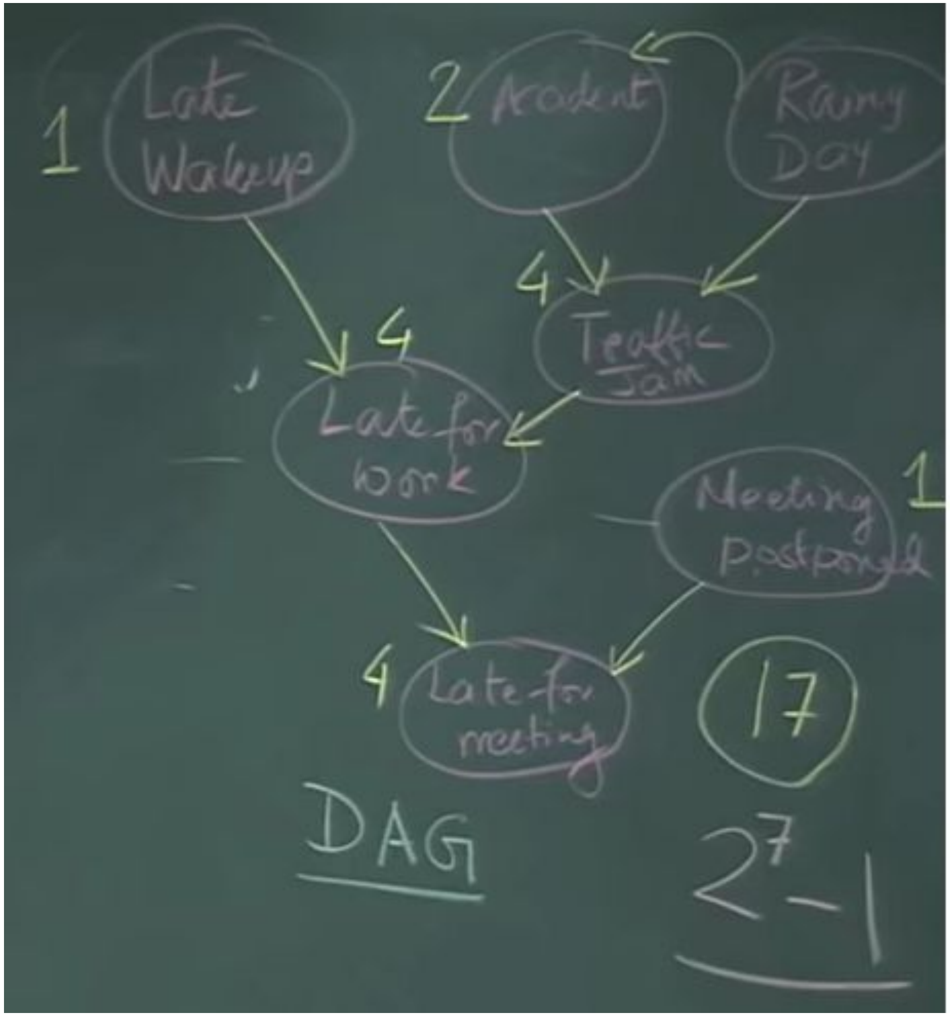


Example Bayes' Net



Example Bayes' Net




1
Late Wakeup
2
Accident
Rainy Day
4
Traffic Jam
4
Late for work
Meeting Postponed
1
4
Late for meeting
17
DAG
2^7-1

Example: Alarm Network

- Variables
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake!

Example: Traffic II

- Let's build a causal graphical model!
- Variables
  - T: Traffic
  - R: It rains
  - L: Low pressure
  - D: Roof drips
  - B: Ballgame
  - C: Cavity

Semantics of Bayesian network

- 2 ways to understand:

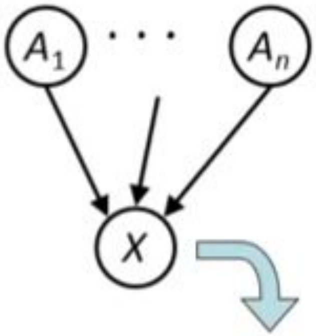
1. To see the network as a representation of the joint probability distribution.
2. View as an encoding of a collection of conditional independence statements.

Semantics of Bayesian network

- A set of nodes, one per variable X
- A directed, acyclic graph
- A conditional distribution for each node
  - A collection of distributions over X, one for each combination of parents' values

$$P(X|a_1 \ldots a_n)$$

  - CPT: conditional probability table
  - Description of a noisy "causal" process



$$P(X|A_1 \ldots A_n)$$

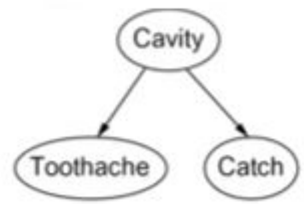
A Bayes net = Topology (graph) + Local Conditional Probabilities

Probabilities in Bayes Nets

- Bayes' nets **implicitly** encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

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

  - Example:



Joint Distribution for this Bayes Net:
*P(cavity, Toothache, Catch) = P(cavity)*P(Toothache | Cavity) * P(Catch | Cavity)*

$P(+cavity, +catch, -toothache)$

Probabilities in Bayes Nets

- Why are we guaranteed that setting

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

  results in a proper joint distribution?

- Chain rule (valid for all distributions): $P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | x_1 \ldots x_{i-1})$

- <u>Assume</u> conditional independences: $P(x_i | x_1, \ldots x_{i-1}) = P(x_i | parents(X_i))$

  → Consequence: $P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | parents(X_i))$

- Not every BN can represent every joint distribution
  - The topology enforces certain conditional independencies

Example: Flip Coins

X_1 X_2 . . . X_n

$P(X_1)$

h	0.5
t	0.5

$P(X_2)$

h	0.5
t	0.5

. . .

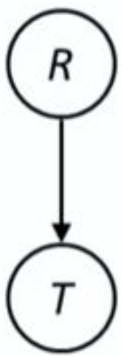
$P(X_n)$

h	0.5
t	0.5

$$P(h, h, t, h) = P(X1=h) * P(X2=h) * P(X3=t) * P(X4=h)$$

Only distributions whose variables are absolutely independent can be represented by a Bayes' net with no arcs.

Example: Traffic



$P(R)$

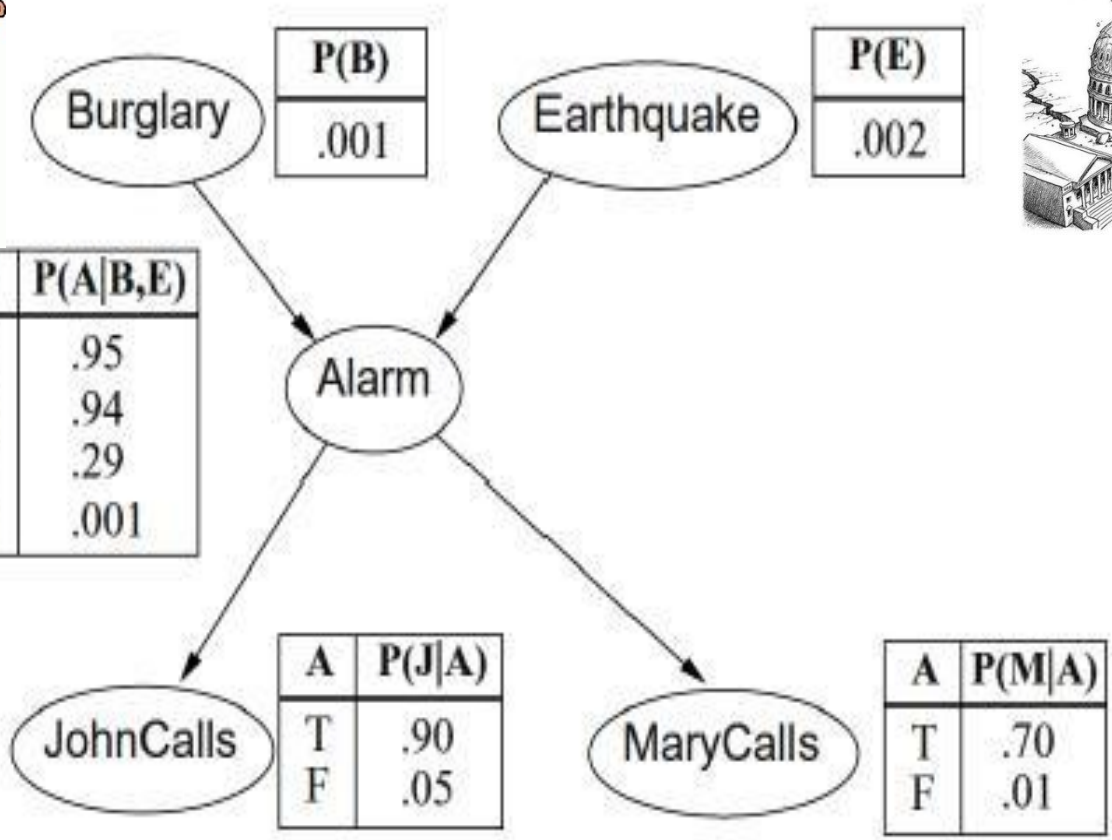
+r	1/4
-r	3/4

$P(T|R)$

+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(+r, -t) = \quad P(+r) * P(-t \mid +r) = \frac{1}{4} * \frac{1}{4}$

Burglar Alarm Example



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

Node ordering

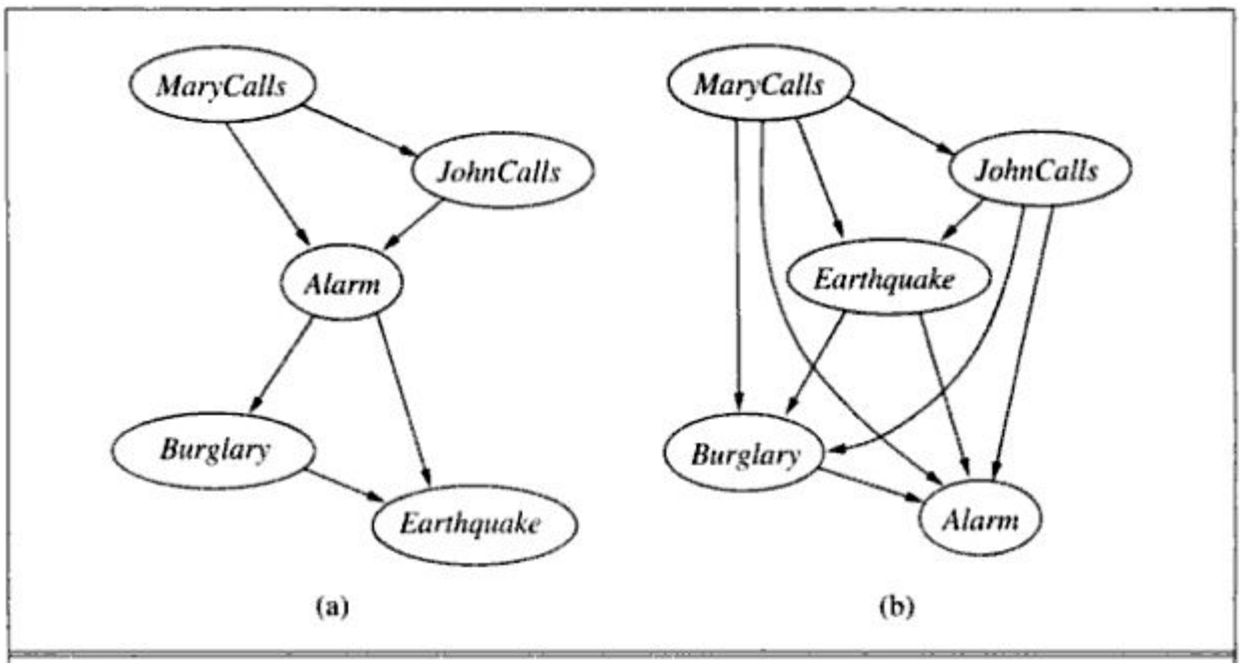
- Adding *MaryCalls*: No parents.
- Adding *JohnCalls*: If Mary calls, that probably means the alarm has gone off, which of course would make it more likely that John calls. Therefore, *JohnCalls* needs *MaryCalls* as a parent
- Adding *Alarm*: Clearly, if both call, it is more likely that the alarm has gone off than if just one or neither call, so we need both *MaryCalls* and *JohnCalls* as parents.
- Adding *Burglary*: If we know the alarm state, then the call from John or Mary might give us information about our phone ringing or Mary's music, but not about burglary:

  $$\mathbf{P}(Burglary|Alarm, JohnCalls, MaryCalls) = \mathbf{P}(Burglary|Alarm) .$$

  Hence we need just *Alarm* as parent.

- Adding *Earthquake*: if the alarm is on, it is more likely that there has been an earthquake. (The alarm is an earthquake detector of sorts.) But if we know that there has been a burglary, then that explains the alarm, and the probability of an earthquake would be only slightly above normal. Hence, we need both *Alarm* and *Burglary* as parents.

Node ordering



Example: Burglar Alarm

- You have a new burglar alarm installed at home.
- It is fairly reliable at detecting burglary, but also sometimes responds to minor earthquakes.
- You have two neighbors, John and Merry , who promised to call you at work whe they hear the alarm.
- John always calls when he hears the alarm, but sometimes confuses telephone ringing with the alarm and calls too.
- Merry likes loud music and sometimes misses the alarm.
- Given the evidence of who has or has not called, we would like to estimate the probability of a burglary.

Example: Burglar Alarm

B	P(B)
+b	0.001
-b	0.999

E	P(E)
+e	0.002
-e	0.998

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

Example: Traffic

- Causal direction

$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

$P(+r, +t) = p(+r) * p(+t \mid +r) = \frac{1}{4} * \frac{3}{4} = 3/16$

Example: Traffic

- Reverse causality?

$P(T)$

+t	9/16
-t	7/16

$P(R|T)$

+t	+r	1/3
	-r	2/3
-t	+r	1/7
	-r	6/7

$P(T, R)$

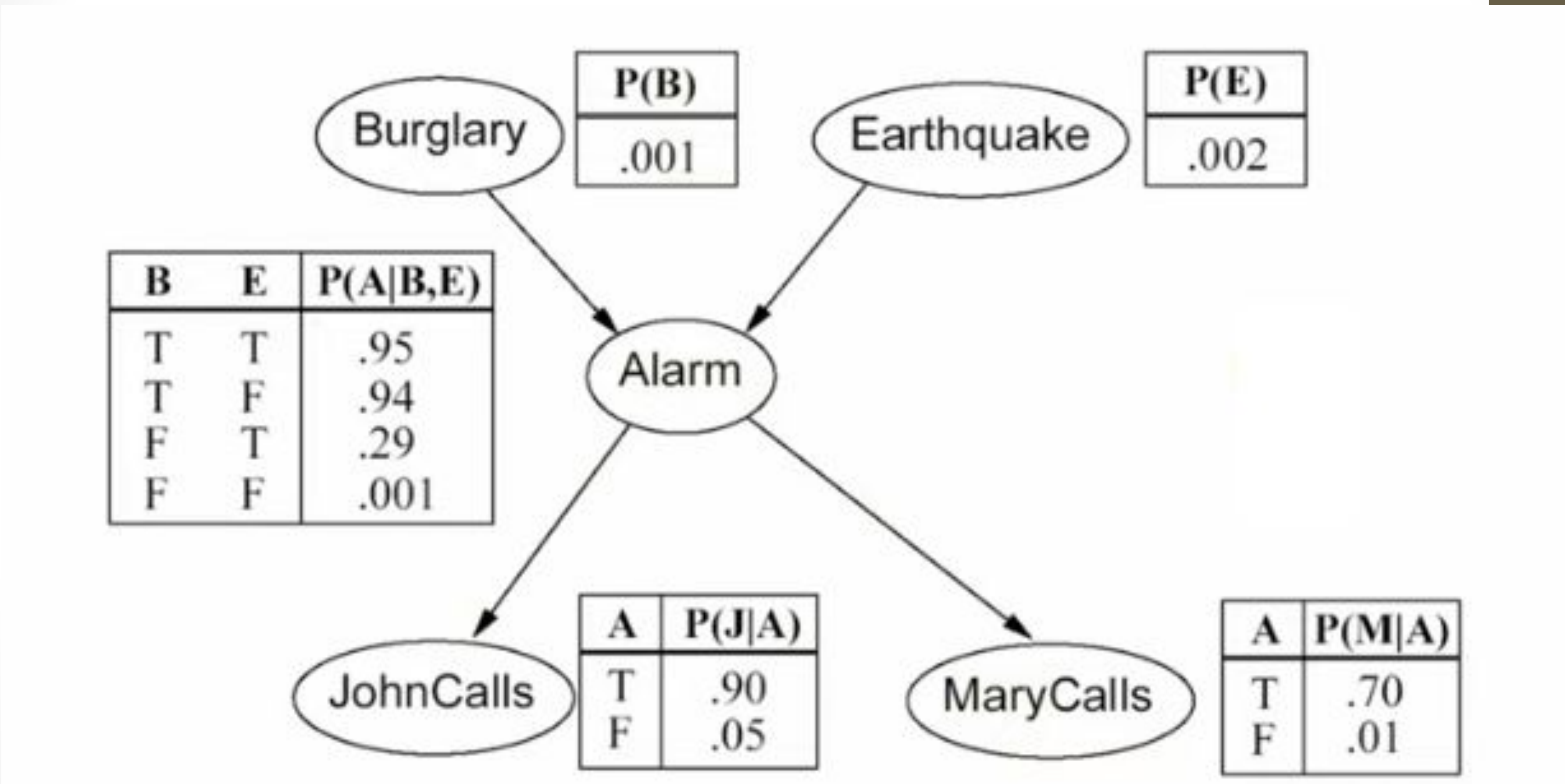
+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

Self learn: Causal Vs diagnostic models. Causal better? Why?

Causality

- When Bayes' nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts
- BNs need not actually be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - E.g. consider the variables *Traffic* and *Drips*
  - End up with arrows that reflect correlation, not causation
- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - Topology really encodes conditional independence

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



P(B)
.001

P(E)
.002

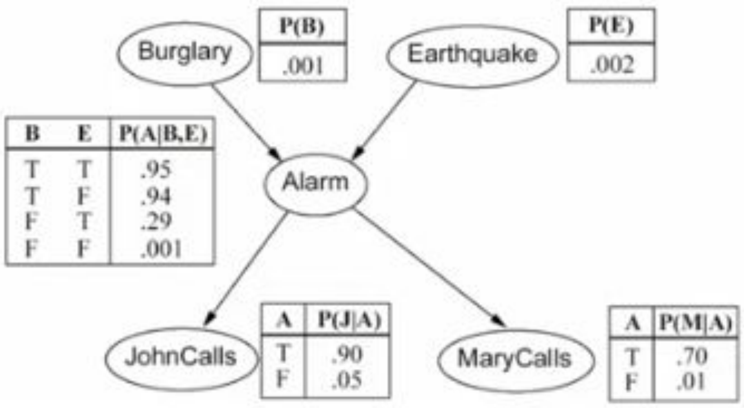
B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

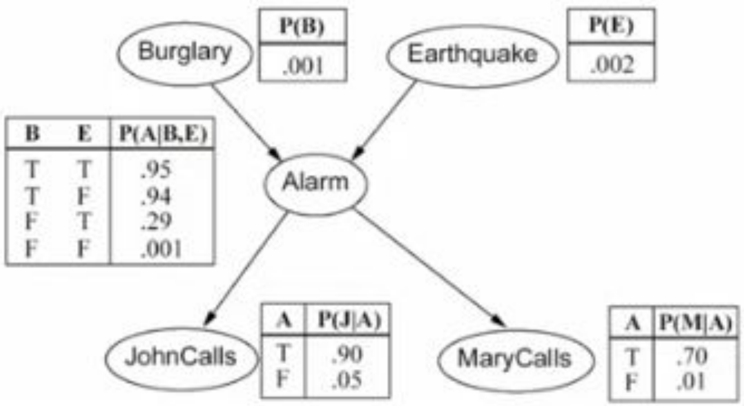
BAYESIAN BELIEF NETWORKS – EXAMPLE – 1

1. What is the probability that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both John and Merry call?



BAYESIAN BELIEF NETWORKS – EXAMPLE – 1

1. What is the probability that the alarm has sounded but neither a burglary nor an earthquake has occurred, and both John and Merry call?



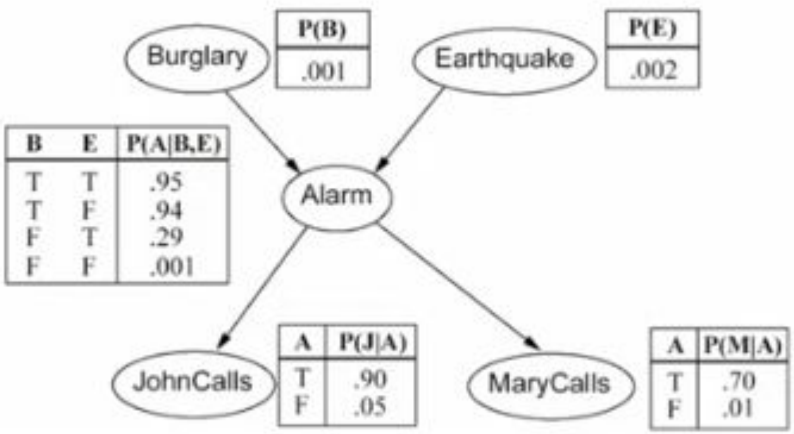
Solution:

$$P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) = P(j \mid a)\, P(m \mid a)\, P(a \mid \neg b, \neg e)\, P(\neg b)\, P(\neg e)$$

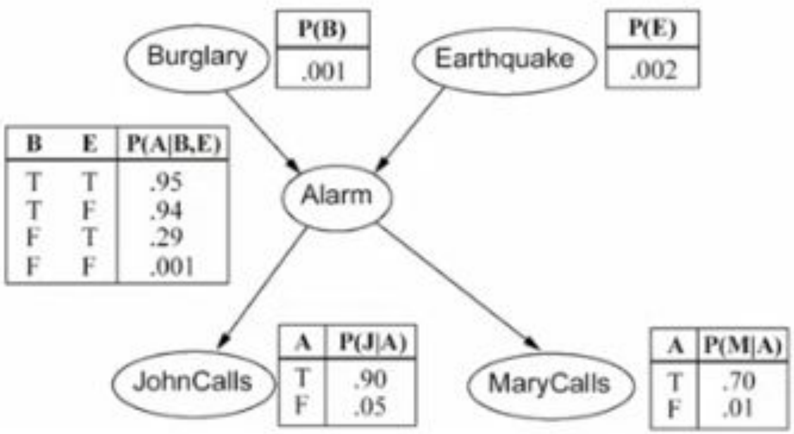
$$= 0.90 \times 0.70 \times 0.001 \times 0.999 \times 0.998$$

$$= 0.00062$$

2. What is the probability that John call?



2. What is the probability that John call?



Solution:

$P(j) = P(j \mid a)\, P(a) + P(j \mid \neg a)\, P(\neg a)$

$= P(j|a)\{P(a|b,e)*P(b,e)+P(a|\neg b,e)*P(\neg b,e)+P(a|b,\neg e)*P(b,\neg e)+P(a|\neg b,\neg e)*P(\neg b,\neg e)\}$

$+ P(j|\neg a)\{P(\neg a|b,e)*P(b,e)+P(\neg a|\neg b,e)*P(\neg b,e)+P(\neg a|b,\neg e)*P(b,\neg e)+P(\neg a|\neg b,\neg e)*P(\neg b,\neg e)\}$

$= 0.90 * 0.00252 + 0.05 * 0.9974 = 0.0521$

Inference in Bayesian networks

- The basic task for any probabilistic inference system is to compute the posterior probability distribution for a set of query nodes, given values for some evidence nodes.
- This task is called belief updating or probabilistic inference.
- Inference simply means to find out probability distribution of some variables which are of your interest given probability distribution of some other variables.
- Inference in Bayesian networks is very flexible, as evidence can be entered about any node while beliefs in any other nodes are updated.
- Major classes of inference algorithms
  - Exact Inference
  - Approximate Inference

Inference in Bayesian networks

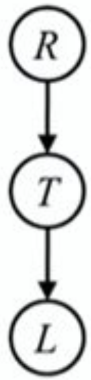
- Exact Inference is intractable in an arbitrary network
  - Enumeration
  - Variable Elimination
- Approximate Inference techniques (Monta Carlo method etc)

Inference by Enumeration

The Traffic Problem:

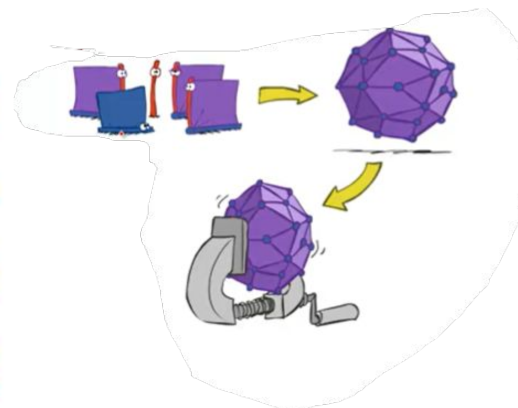
Rain (R), Traffic (T), Late for work (L)

L=?



- Inference by Enumeration

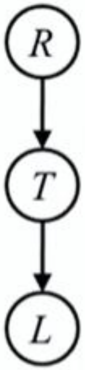
$$= \underbrace{\sum_t \underbrace{\sum_r \underbrace{P(L|t)\underbrace{P(r)P(t|r)}_{\text{Join on r}}}_{\text{Join on t}}}_{\text{Eliminate r}}}_{\text{Eliminate t}}$$



Inference by Enumeration is slow as we join up the whole joint distribution before we sum out the hidden variables

Therefore Variable Elimination

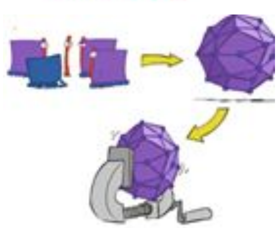
Enumeration Vs Elimination



Rain (R), Traffic (T), Late for work (L)

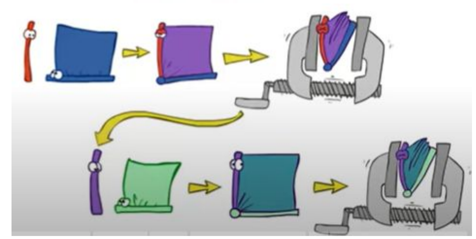
- Inference by Enumeration

$$=\underbrace{\sum_t \underbrace{\sum_r \underbrace{P(L|t)\underbrace{P(r)P(t|r)}_{\text{Join on r}}}_{\text{Join on t}}}_{\text{Eliminate r}}}_{\text{Eliminate t}}$$



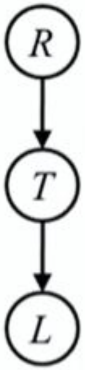
- Variable Elimination

$$=\underbrace{\sum_t \underbrace{P(L|t)\underbrace{\sum_r \underbrace{P(r)P(t|r)}_{\text{Join on r}}}_{\text{Eliminate r}}}_{\text{Join on t}}}_{\text{Eliminate t}}$$



Inference by Elimination: We interleaved joining and marginalizing (elimination). Still NP hard. But faster than enumeration

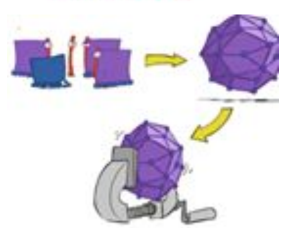
Enumeration Vs Elimination



Rain (R), Traffic (T), Late for work (L)

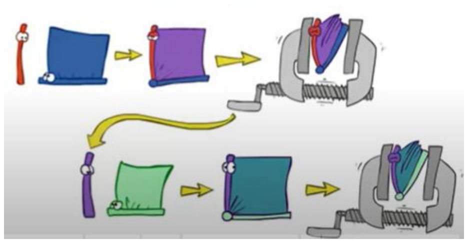
- Inference by Enumeration

$$=\underbrace{\sum_t \underbrace{\sum_r \underbrace{P(L|t)\underbrace{P(r)P(t|r)}_{\text{Join on r}}}_{\text{Join on t}}}_{\text{Eliminate r}}}_{\text{Eliminate t}}$$



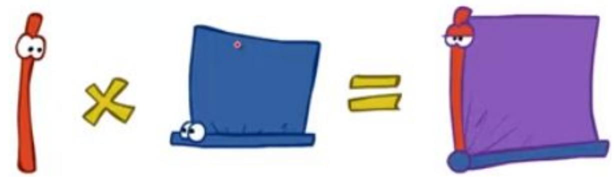
- Variable Elimination

$$=\underbrace{\sum_t \underbrace{P(L|t)\underbrace{\sum_r \underbrace{P(r)P(t|r)}_{\text{Join on r}}}_{\text{Eliminate r}}}_{\text{Join on t}}}_{\text{Eliminate t}}$$



Inference by Elimination: We interleaved joining and marginalizing (elimination). Still NP hard. But faster than enumeration

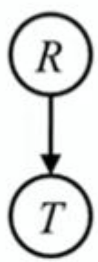
Operations involved: Join and Elimination



- **Join**
  - Similar to database join
  - Get all the factors over joining variable
  - Build a new factor over the union of the variables involved

- Random Variables
  - R: Raining
  - T: Traffic

Example: Join on R



$P(R) \times P(T|R)$ → $P(R,T)$

$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

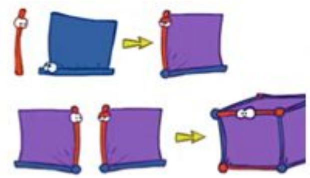
+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

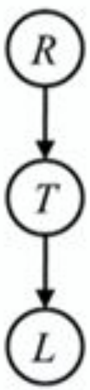
$P(R,T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

Operations involved: Join and Elimination

- **Multiple Joins**





$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

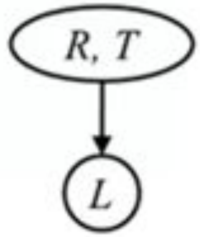
Join R

$P(R,T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9



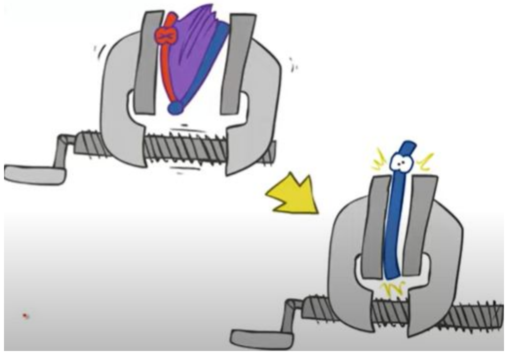
Join T

$P(R,T,L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

Operations involved: Join and Elimination

- Elimination/Marginalization
  - Take out a factor and sum out a variable
    - Shrinks a factor to smaller one
    - It's a projection operation



$P(R, T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

sum R →

$P(T)$

+t	0.17
-t	0.83

Operations involved: Join and Elimination

- **Multiple Elimination**

R, T, L

$P(R, T, L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

Sum out R

T, L

$P(T, L)$

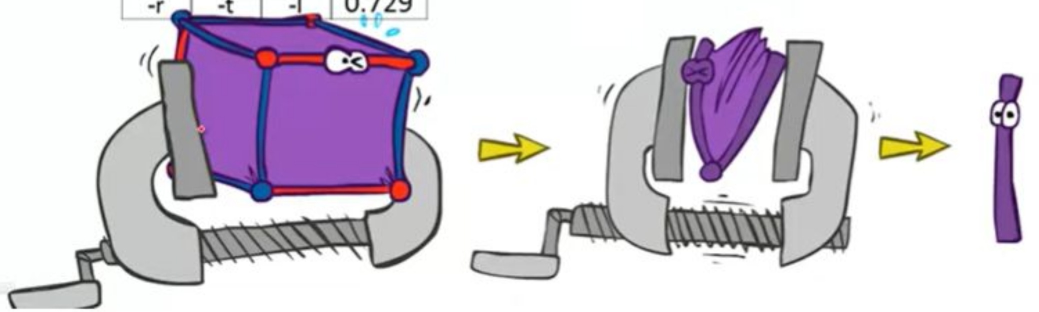
+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

Sum out T

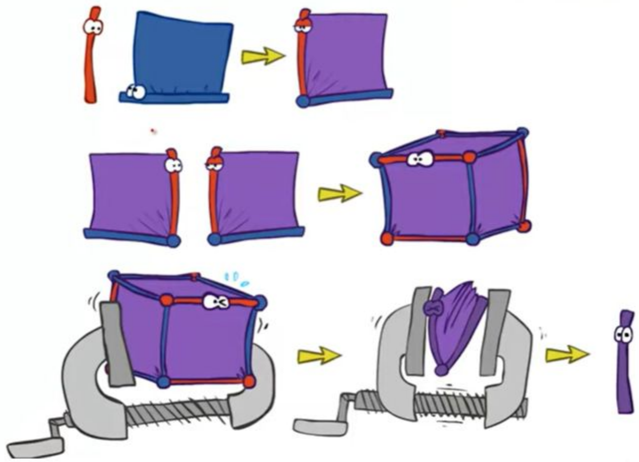
L

$P(L)$

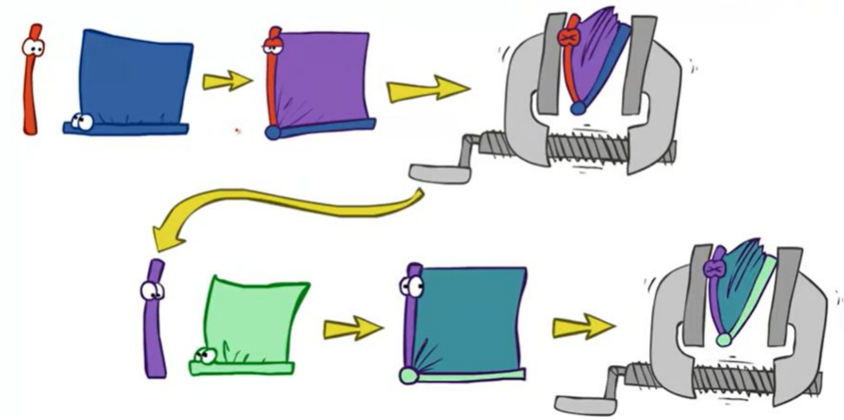
+l	0.134
-l	0.886



Thus Far: Multiple Join, Multiple Eliminate (= Inference by Enumeration)

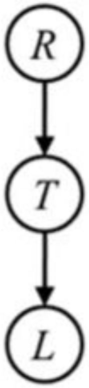


Marginalizing Early (= Variable Elimination)



Traffic Example by Elimination

- Random Variables
  - R: Raining
  - T: Traffic
  - L: Late for class!



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$$P(L) = ?$$

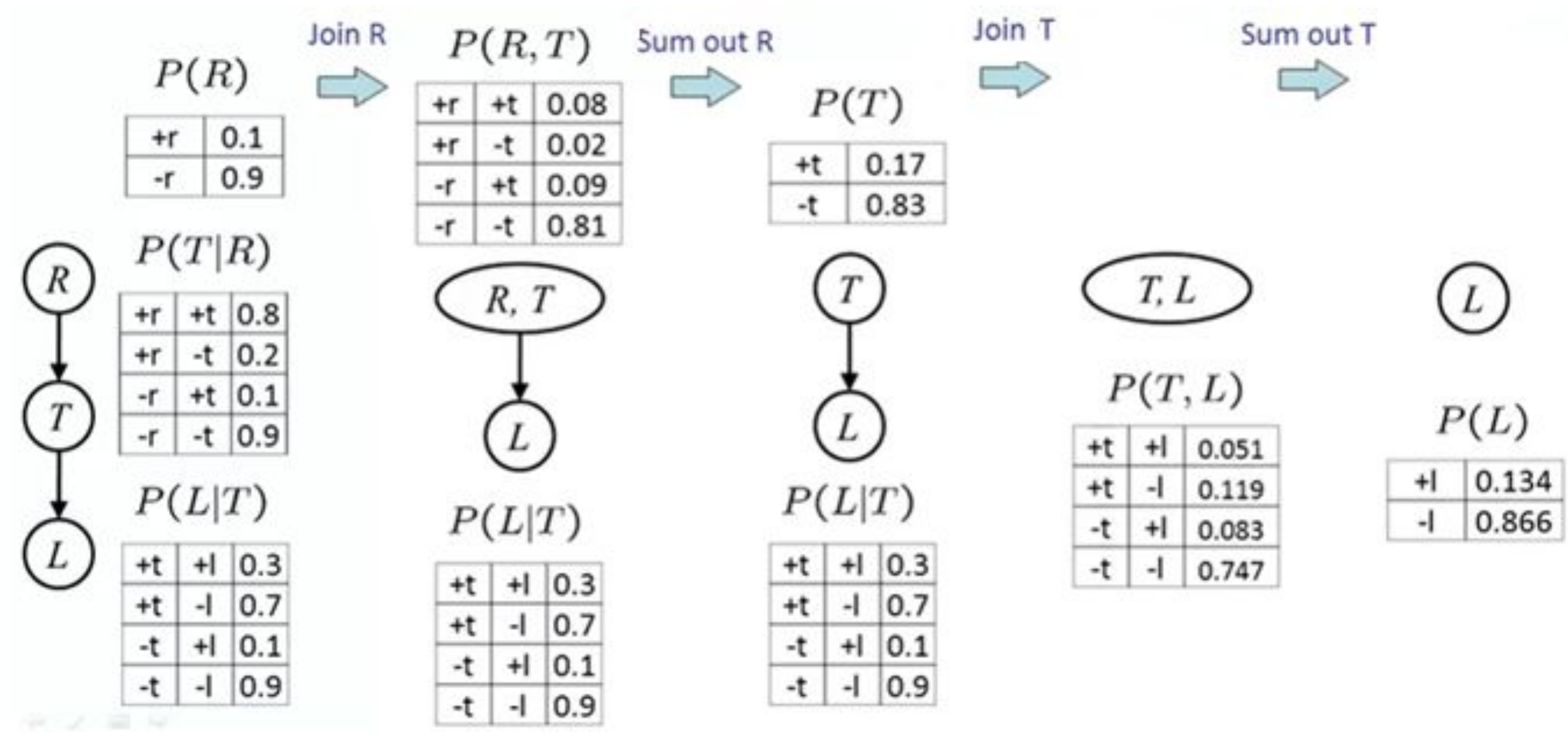
$$= \sum_{r,t} P(r,t,L)$$

$$P(L) = ?$$

$$= \sum_{r,t} P(r,t,L)$$

$$= \sum_{r,t} P(r)P(t|r)P(L|t)$$

Traffic Example by Elimination



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(R,T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(T)$

+t	0.17
-t	0.83

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$P(T,L)$

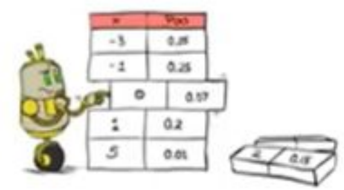
+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

$P(L)$

+l	0.134
-l	0.866

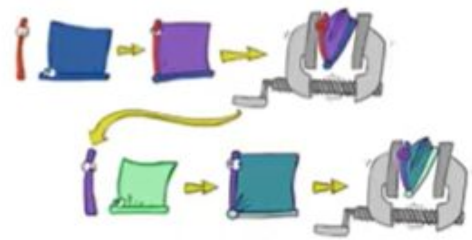
General Variable Elimination

- Query: $P(Q|E_1 = e_1, \ldots E_k = e_k)$



- Start with initial factors:
  - Local CPTs (but instantiated by evidence)

- While there are still hidden variables (not Q or evidence):
  - Pick a hidden variable H
  - Join all factors mentioning H
  - Eliminate (sum out) H



- Join all remaining factors and normalize

 $\times \frac{1}{Z}$