Inference Engine

The inference engine is a generic control mechanism for navigating through and manipulating knowledge and deduce results in an organized manner.
The inference engine's generic control mechanism applies the axiomatic (self-evident) knowledge present in the knowledge base to the task-specific data to arrive at some conclusion.

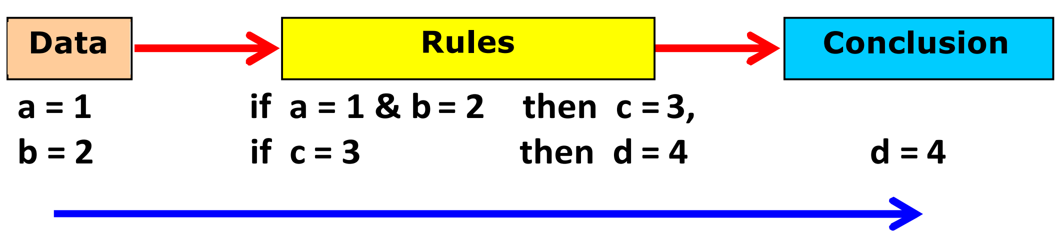
- Inference engine the other key component of all expert systems.
- Just a knowledge base alone is not of much use if there are no facilities for navigating through and manipulating the knowledge to deduce something from knowledge base.
- A knowledge base is usually very large, it is necessary to have inferencing mechanisms that search through the database and deduce results in an organized manner.

The Forward chaining, Backward chaining and Tree searches are some of the techniques used for drawing inferences from the knowledge base.

Forward Chaining Algorithm

Forward chaining is a techniques for drawing inferences from Rule base. Forward-chaining inference is often called data driven.

- The algorithm proceeds from a given situation to a desired goal, adding new assertions (facts) found.
- A forward-chaining, system compares data in the working memory against the conditions in the IF parts of the rules and determines which rule to fire.
- Data Driven



- Example : Forward Channing

  - Given : A Rule base contains following Rule set

Rule 1: If A and C Then F
Rule 2: If A and E Then G

Rule 3: If B Then E

Rule 4: If G Then D

- Problem : Prove

  If A and B true Then D is true

- **Solution** :

(i) ‡ Start with input given **A, B** is true and then

‡ start at **Rule 1** and go forward/down till a rule "fires" is found.

First iteration :

(ii) ‡ **Rule 3** fires : conclusion **E** is true

‡ new knowledge found

(iii) ‡ No other rule fires;

‡ end of first iteration.

(iv) ‡ Goal not found;

‡ new knowledge found at (ii);

‡ go for second iteration

Second iteration:

(v) ‡ **Rule 2** fires : conclusion **G** is true

‡ new knowledge found

(vi) ‡ **Rule 4** fires : conclusion **D** is true

‡ Goal found;

‡ Proved

Backward Chaining Algorithm

Backward chaining is a techniques for drawing inferences from Rule base. Backward-chaining inference is often called goal driven.

‡ The algorithm proceeds from desired goal, adding new assertions found.

‡ A backward-chaining, system looks for the action in the THEN clause of the rules that matches the specified goal.

‡ Goal Driven

b = 2 **if c = 3** **then d = 4** **d = 4**

‡ Example : Backward Channing

- Given : Rule base contains following Rule set

Rule 1:	If A and C	Then	F
Rule 2:	If A and E	Then	G
Rule 3:	If B	Then	E
Rule 4:	If G	Then	D

- Problem : Prove

  If A and B true Then D is true

[Continued from previous slide]

- **Solution** :

(i) ‡ Start with goal ie **D** is true
‡ go backward/up till a rule "fires'' is

First iteration :

(ii) ‡ **Rule 4** fires :
‡ new sub goal to prove **G** is true
‡ go backward

(iii) ‡ **Rule 2** "fires''; conclusion: **A** is true
‡ new sub goal to prove **E** is true
‡ go backward;

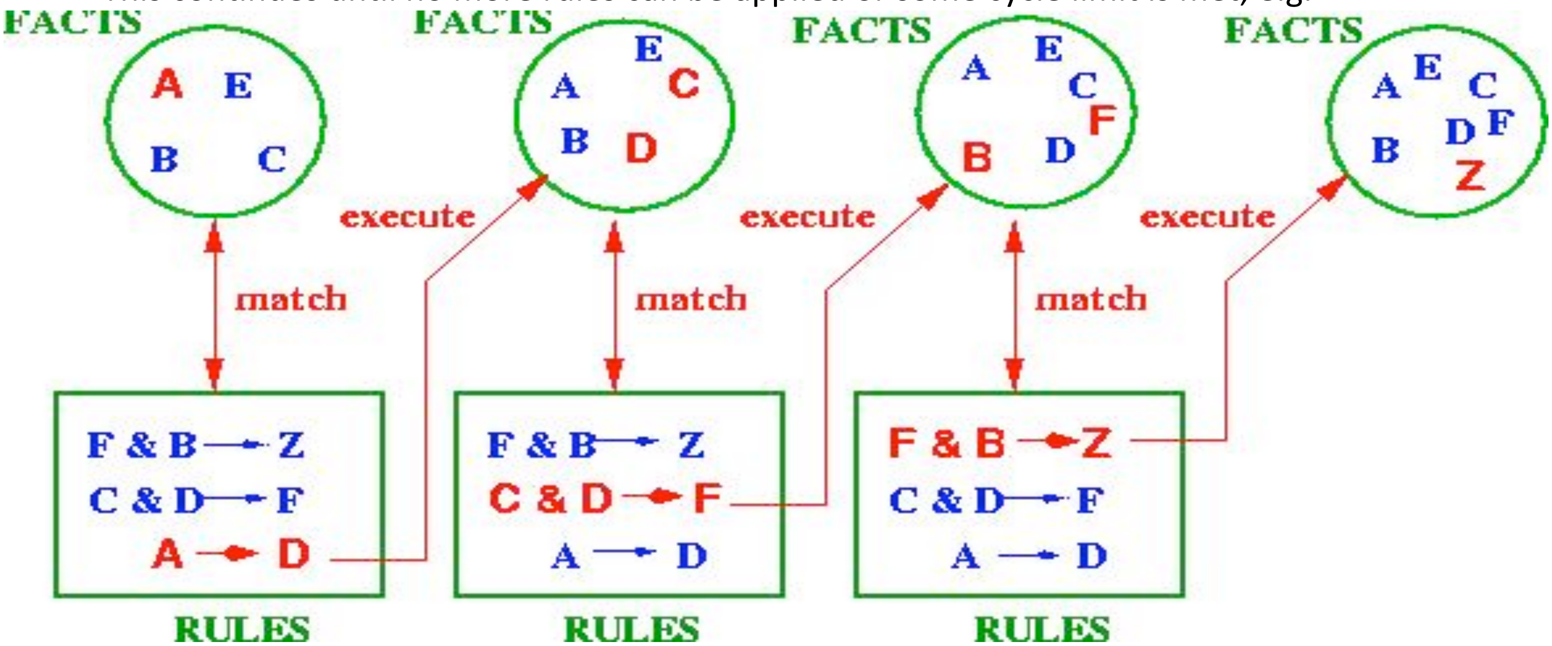
(iv) ‡ no other rule fires; end of first iteration.
‡ new sub goal found at (iii);
‡ go for second iteration

Second iteration:

(v) ‡ **Rule 3** fires :
‡ conclusion **B** is true (2nd input found)
‡ both inputs **A** and **B** ascertained
‡ Proved

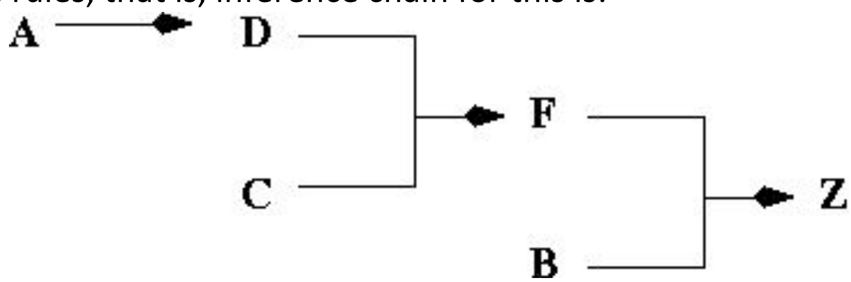
Forward Chaining

premises of the rules (IF-part), perform the actions (THEN-part), possibly updating the knowledge base or working memory.
This continues until no more rules can be applied or some cycle limit is met, e.g.



Forward Chaining (Cont'd)

In example: no more rules, that is, inference chain for this is:



Problem with forward chaining:

many rules may be applicable. The whole process is **not directed** towards a **goal**.

Backward Chaining

Backward chaining or **goal-driven** inference works towards a final state, and by looking at the working memory to see if goal already there. If not look at the actions (THEN-parts) of rules that will establish goal, and set up subgoals for achieving premises of the rules (IF-part).
This continues until some rule can be applied, apply to achieve goal state.

Advantage of backward chaining:

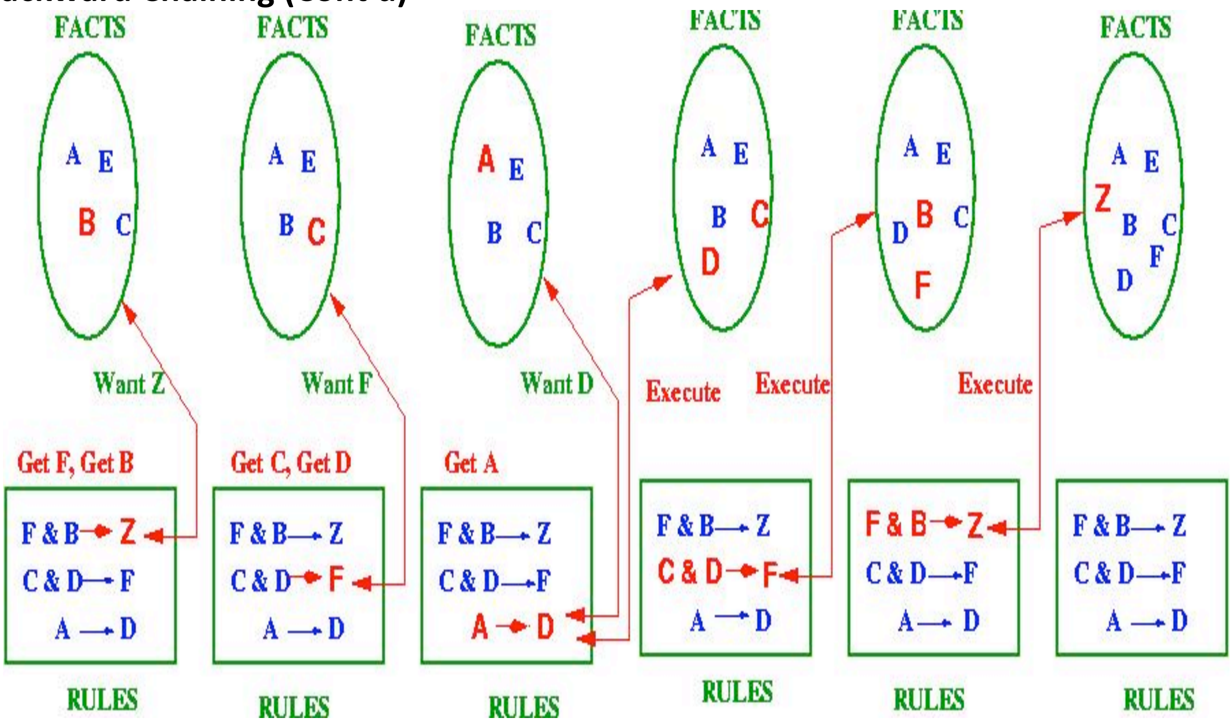
search is directed

Disadvantage of backward chaining:

goal has to be known

Now look at the example from above with backward chaining

Backward Chaining (Cont'd)



Probability and Bayes' Theorem

Bayes' theorem is named after Thomas Bayes, who did early work in probability and decisic theory during the 18th century.

An important goal for many problem solving system is

- To collect evidence as the system goes.
- To modify its behavior on the basis of the evidence.

To model this behavior,

- Statistical theory of evidence is needed
- Bayesian statistics is used

The fundamental notion of Bayes' Theorem is conditional probability:

$$P(H \setminus E)$$

Means the probability that the hypothesis *H* holds given that observed "evidence" is E.

To compute this , we need to take into account:

$P(H_i \setminus E)$ = the probability that hypothesis H_i is true given evidence E

$P(E \setminus H_i)$ = the probability that we will observe evidence E given that hypothesis i is true

$P(H_i)$ = the *a priori* probability that hypothesis i is true in the absence of any specific evidence. These probabilities are called prior probabilities or *prlors*.

k = the number of possible hypotheses

Bayes' theorem then states that

$$P(H_i \backslash E) = \frac{P(E \mid H_i) \cdot P(H_i)}{\sum_{n=1}^{k} P(E \mid H_n) \cdot P(H_n)}$$

Probabilistic reasoning is used when outcomes are uncertain.
Bayes' theorem is basis for uncertain reasoning where the outcomes are unpredictable.

Eg for uncertain outcomes are:

- Physician examines a patient based on patient history, symptoms, test results with physician's experience predicting the unknown disease, but still there is still uncertainity in most diagnosis.

- **Probability :** The Probabilities are numeric values between **0** and **1** (both inclusive) that represent ideal uncertainties (not beliefs).
  - **Probability of event A is P(A)**

$$P(A) = \frac{\textbf{instances of the event A}}{\textbf{total instances}}$$

P(A) = 0 indicates total uncertainty in **A**,

P(A) = 1 indicates total certainty and

0< P(A) < 1 values in between tells degree of uncertainty

Example 1 : A single 6-sided die is rolled.

What is the probability of each outcome?

What is the probability of rolling an even number?

What is the probability of rolling an odd number?

The possible outcomes of this experiment are 1, 2, 3, 4, 5, 6.

The Probabilities are :

P(1) =No of ways to roll 1 / total no of sides = 1/6

P(2) =No of ways to roll 2 / total no of sides = 1/6

P(3) =No of ways to roll 3 / total no of sides = 1/6

P(4) =No of ways to roll 4 / total no of sides = 1/6

P(5)=No of ways to roll 5 / total no of sides = 1/6

P(6) =No of ways to roll 6 / total no of sides = 1/6

P(even) =ways to roll even no / total no of sides = 3/6 = 1/2 P(odd)

=ways to roll odd no / total no of sides = 3/6 = 1/2

- **Conditional probability** **P(A|B)**

A conditional probability is the probability of an event given that another event has occurred.

Example : Roll two dices.

What is the probability that the total of two dice will be greater than 8 given that the first die is a 6 ?

First List of the **joint possibilities** for the two dices are:

(1, 1)	(1, 2)	(1, 3)	(1, 4)	(1, 5)	(1, 6)
(2, 1)	(2, 2)	(2, 3)	(2, 4)	(2, 5)	(2, 6)
(3, 1)	(3, 2)	(3, 3)	(3, 4)	(3, 5)	(3, 6)
(4, 1)	(4, 2)	(4, 3)	(4, 4)	(4, 5)	(4, 6)
(5, 1)	(5, 2)	(5, 3)	(5, 4)	(5, 5)	(5, 6)
(6, 1)	(6, 2)	**(6, 3)**	**(6, 4)**	**(6, 5)**	**(6, 6)**

There are 6 outcomes for which the first die is a 6, and of these, there are 4 outcomes that total more than 8 are (6,3; 6,4; 6,5; 6,6).

The probability of a total > 8 given that first die is 6 is therefore 4/6 = 2/3 .

This probability is written as: $P(\underbrace{\text{total}>8}_{\textbf{event}} \mid \underbrace{\text{1st die} = 6}_{\textbf{condition}}) = 2/3$

Read as "The probability that the total is > 8 given that die one is 6 is 2/3."

- **Probability of A and B is P(A and B)**

The probability that events **A** and **B** both occur.
Written as **P(A|B)** , is the probability of event A given that the event B has occurred.

Two events are independent if the occurrence of one is unrelated to the probability of the occurrence of the other.

‡ **If A and B are independent**

then probability that events **A** and **B** both occur is:

P(A and B) = P(A) x P(B)

ie product of probability of **A** and probability of **B**.

‡ **If A and B are not independent**

then probability that events **A** and **B** both occur is:

P(A and B) = P(A) x P(B|A) where

P(B|A) is **conditional probability** of **B** given **A**

Example 1: P(A and B) if events A and B are independent

- Draw a card from a deck , then replace it, draw another card.
- Find probability that 1st card is Ace of clubs (event A) and 2nd card is any Club (event B).
- Since there is only one Ace of Clubs, therefore probability P(A) = 1/52.
- Since there are 13 Clubs, the probability P(B) = 13/52 = 1/4.
- Therefore, **P(A and B) =** p(A) x p(B) = 1/52 x 1/4 = 1/208.

Example 2: P(A and B) if events A and B are not independent

- Draw a card from a deck, not replacing it, draw another card.
- Find probability that both cards are Aces ie the 1st card is Ace (event A) and the 2nd card is also Ace (event B).
- Since 4 of 52 cards are Aces, therefore probability P(A) = 4/52 = 1/13.
- Of the 51 remaining cards, 3 are aces. so, probability of 2nd card is also Ace (event B) is P(B|A) = 3/51 = 1/17.

■ **Probability of A or B is P(A or B)**

The probability of either event **A** or event **B** occur.

Two events are mutually exclusive if they cannot occur at same time.

‡ **If A and B are mutually exclusive**

then probability that events **A** or **B** occur is:

P(A or B) = p(A) + p(B)

ie sum of probability of **A** and probability of **B**

‡ **If A and B are not mutually exclusive**

then probability that events **A** and **B** both occur is:

P(A or B) = P(A) x P(B|A) – P(A and B) where

P(A and B) is probability that events **A** and **B** both occur while events **A** and **B** are independent and **P(B|A)** is conditional probability of **B** given **A**.

Example 1: P(A or B) if events A or B are mutually exclusive

- Rolling a die.
- Find probability of getting either, event A as 1 or event B as 6?
- Since it is impossible to get both, the event A as 1 and event B as 6 in same roll, these two events are mutually exclusive.
- The probability P(A) = P(1) = 1/6 and P(B) = P(6) = 1/6
- Hence probability of either event A or event B is :
- **P(A or B)** = p(A) + p(B) = 1/6 + 1/6 = 1/3

Example 2: P(A or B) if events A or B are not mutually exclusive

- Find probability that a card from a deck will be either an Ace or a Spade?
- probability P(A) is P(Ace) = 4/52 and P(B) is P(spade) = 13/52.
- Only way in a single draw to be Ace and Spade is Ace of Spade; which is only one, so probability P(A and B) is P(Ace and Spade) = 1/52.
- Therefore, the probability of event A or B is :

P(A or B) = P(A) + P(B) – P(A and B)
= P(ace) + P(spade) - P(Ace and Spade)
= 4/52 + 13/52 - 1/52 = 16/52 = 4/13

Bayesian view of probability is related to degree of belief.

It is a measure of the plausibility of an event given incomplete knowledge.

Bayes' theorem is also known as Bayes' rule or Bayes' law, or called Bayesian reasoning.

The probability of an event **A** conditional on another event **B** ie **P(A|B)** is generally different from probability of **B** conditional on **A** ie **P(B|A)**.

- There is a definite relationship between the two, P(A|B) and P(B|A), and Bayes' theorem is the statement of that relationship.
- Bayes theorem is a way to calculate P(A|B) from a knowledge of P(B|A).
- Bayes' Theorem is a result that allows new information to be used to update the conditional probability of an event.

- **Bayes' Theorem**

Let S be a sample space.

Let $A1, A2, \dots, An$ be a set of mutually exclusive events from S.

Let B be any event from the same S, such that $P(B) > 0$.

Then Bayes' Theorem describes following two probabilities :

$$P(A_k|B) = \frac{P(A_k \cap B)}{P(A1 \cap B) + P(A_2 \cap B) + \dots + P(An \cap B)}$$ and

by invoking the fact $P(A_k \cap B) = P(A_k).P(B|A_k)$ the probability

$$P(A_k|B) = \frac{P(A_k).P(B|A_k)}{P(A_1).P(B|A_1) + P(A2).P(B|A2) + \dots + P(A_n).P(B|A_n)}$$

Applying Bayes' Theorem :

Bayes' theorem is applied while following conditions exist.

‡ the sample space S is partitioned into a set of mutually exclusive events $\{A1, A2, \dots, An\}$.

‡ within S, there exists an event B, for which $P(B) > 0$.

‡ the goal is to compute a conditional probability of the form : $P(A_k|B)$.

‡ you know at least one of the two sets of probabilities described below

- ◇ $P(A_k \cap B)$ for each A_k
- ◇ $P(A_k)$ and $P(B|A_k)$ for each A_k

The Bayes' theorem is best understood through an example below.

Problem :

- Marie's marriage is tomorrow.
  - in recent years, each year it has rained only 5 days.
  - the weatherman has predicted rain for tomorrow.
  - when it actually rains, the weatherman correctly forecasts rain 90% of the time.
  - when it doesn't rain, the weatherman incorrectly forecasts rain 10% of the time.

The question : What is the probability that it will rain on the day of Marie's wedding?

Solution : The sample space is defined by two mutually exclusive events

– "it rains" or "it does not rain". Additionally, a third event occurs when the "weatherman predicts rain".

The events and probabilities are stated below.

- Event A1 : rains on Marie's wedding.
- Event A2 : does not rain on Marie's wedding
- Event B : weatherman predicts rain.
- P(A1) = 5/365 =0.0136985 [Rains 5 days in a year.]
- P(A2) = 360/365 = 0.9863014 [Does not rain 360 days in a year.]
- P(B|A1) = 0.9 [When it rains, the weatherman predicts rain 90% time.]
- P(B|A2) = 0.1 [When it does not rain, weatherman predicts rain 10% time.]

We want to know **P(A1|B)**, the probability that it will rain on the day of Marie's wedding, given a forecast for rain by the weatherman. The answer can be determined from Bayes' theorem, shown below.

$$P(A1|B) = \frac{P(A1).P(B|A1)}{P(A1).P(B|A1)+P(A2).P(B|A2)} = \frac{(0.014)(0.9)}{[(0.014)(0.9)+(0.986)(0.1)]}$$

$$= 0.111$$

So, despite the weatherman's prediction, there is a good chance that Marie will not get rain on at her wedding.

Thus Bayes theorem is used to calculate conditional probabilities.

Suppose for example, for solving medical diagnosis problem, consider the following assertions:

- S: patient has spots
- M: patient has measles
- F: patient has high fever

- ✓ Without any additional evidence, the presence of spots serves as evidence in favor of measles.
- ✓ Spot is also serves as a evidence for fever since measles would also cause fever.
- ✓ Suppose that the patient already has measles, then the additional evidence that he has spots actually tells us about the likelihood of fever.
- ✓ Alternatively either spots alone or fever alone cause evidence in favor of
- ✓ measles.
- ✓ If both are present then we need to find total weight of evidence.
- ✓ But, since spots and fever are not independent events, we need to represent conditional probability that arises from their conjunction.
- ✓ Foe eg. Given a prior body of evidence e and some new observation E, we need to compute

$$P(H \backslash E, e) = P(H \mid E) \cdot \frac{P(e \mid E, H)}{P(e \mid E)}$$

Using Bayes' theorem is intractable for several reasons. They are

- The knowledge acquisition problem is challenging, too many probabilities have to be provided.
- People are very poor probability estimators.
- Space required to store all the probabilities is too large.
- Time required to compute all the probabilities is too large.

Certainty Factors and Rule-Based Systems

Certainty Factor: It is a measure of the extent to which the evidence that is described by the antecedent of the rule supports the conclusion that is given in the rule's consequent.

MYCIN represents most of its diagnostic knowledge as a set of rules. Typical MYCIN rule looks like:

```
If:    (1) the stain of the organism is gram-positive, and
       (2) the morphology of the organism is coccus, and
       (3) the growth conformation of the organism is clumps,
           then there is suggestive evidence (0.7) that
           the identity of the organism is staphylococcus.
```

These rules are internally represented using easy to manipulate LISP list structure

```
PREMISE:   ($AND    (SAME CNTXT GRAM GRAMPOS)
                    (SAME CNTXT MORPH COCCUS)
                    (SAME CNTXT CONFORM CLUMPS))
ACTION:             (CONCLUDE CNTXT IDENT STAPHYLOCOCCUS TALLY 0.7)
```

MYCIN uses backward reasoning to find the disease causing organism. Once it finds the organism, it then attempts to select a therapy by which the disease may be treated.

To understand how MYCIN exploits uncertain information, we need to answer:

- What do certainty factors mean?
- How does MYCIN combine the estimates of certainty in each of its rules to produce a final estimate of the certainty of its conclusions.

A certainty factor (CF[h,e]) is defined in terms of two components:

- *MB*[*h*, *e*]—a measure (between 0 and 1) of belief in hypothesis *h* given the evidence *e*. *MB* measures the extent to which the evidence supports the hypothesis. It is zero if the evidence fails to support the hypothesis.
- *MD*[*h*,*e*]—a measure (between 0 and 1) of disbelief in hypothesis *h* given the evidence *e*. *MD* measures the extent to which the evidence supports the negation of the hypothesis. It is zero if the evidence supports the hypothesis.

From these two measures, we can define the CF as

$$CF[h, e] = MB[h, e] - MD[h, e]$$

- Any particular piece of evidence either supports or denies a hypothesis, and since each MYCIN rule corresponds to one piece of evidence, a single no. suffices each rule to define both the MB and MD and thus the CF.
- As MYCIN reasons, CF's need to be combined to reflect the operation of multiple pieces of evidence and multiple rules applied to a problem.
- Fig. illustrates three combination scenarios that we need to consider.

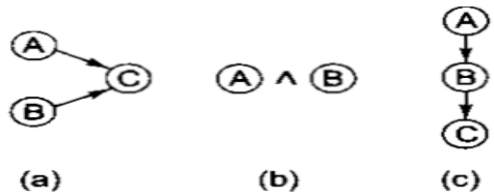


Fig. *Combining Uncertain Rules*

Fig illustrates three combination scenarios that we need to consider
Fig(a) – several rules all provide evidence that relates to a single hypothesis
Fig(b) – considers the belief in collection of several propositions taken together.
Fig(c) - output of one rule provides input to another

Properties to be satisfied to combine these functions in some order are:

- Since the order in which evidence is collected is arbitrary, the combining functions should be commutative and associative.
- Until certainty is reached, additional confirming evidence should increase *MB* (and similarly for disconfirming evidence and *MD*).
- If uncertain inferences are chained together, then the result should be less certain than either of the inferences alone.
  - Consider the scenario one shown in Fig(a) in which several pieces of evidence are combined to determine the CF of one hypothesis
    The measure of belief and disbelief of a hypothesis given two observations s1 and s2 are computed from:

$$MB[h,s_1 \wedge s_2] = \begin{cases} 0 & \text{if } MD[h, s_1 \wedge s_2] = 1 \\ MB[h, s_1] + MB[h, s_2] \cdot (1 - MB[h, s_1]) & \text{otherwise} \end{cases}$$

$$MD[h, s_1 \wedge s_2] = \begin{cases} 0 & \text{if } MB[h, s_1 \wedge s_2] = 1 \\ MD[h, s_1] + MD[h, s_2] \cdot (1 - MD[h, s_1]) & \text{otherwise} \end{cases}$$

A simple example shows how these functions operate. Suppose we make an initial observation that confirms our belief in h with $MB = 0.3$. Then $MD[h,s_1] = 0$ and $CF[h, s_1] = 0.3$. Now we make a second observation, which also confirms h, with $MB[h,s_2] = 0.2$. Now:

$$\begin{aligned} MB[h,s_1 \wedge s_2] &= 0.3 + 0.2 \cdot 0.7 \\ &= 0.44 \\ MD[h,s_1 \wedge s_2] &= 0.0 \\ CF[h,s_1 \wedge s_2] &= 0.44 \end{aligned}$$

- Consider the scenario two shown in Fig(b), in which we need to find certainty factor of combination of hypothesis.
  - The combination of certainty factor can be computed from its MB and MD.
  - The formula MYCIN uses for the MB of the conjunction and the disjunction of two hypothesis are:

$$MB[h_1 \wedge h_2, e] = \min(MB[h_1, e], MB[h_2, e])$$

$$MB[h_1 \wedge h_2, e] = max(MB[h_1,e], MB[h_2,e])$$

MD can be computed analogously

- Consider the scenario two shown in Fig(c), in which rules are chained together with the result that the uncertain outcome of rule must provide the input to another.

$$MB[h, e] = \begin{cases} 1 & \text{if } P(h) = 1 \\ \dfrac{\max[P(h \mid e), P(h)] - P(h)}{1 - P(h)} & \text{otherwise} \end{cases}$$

Similarly, the MD is the proportionate decrease in belief in h as a result of e:

$$MD[h, e] = \begin{cases} 1 & \text{if } P(h) = 0 \\ \dfrac{\min[P(h \mid e), P(h)] - P(h)}{-P(h)} \end{cases}$$

Bayesian Network

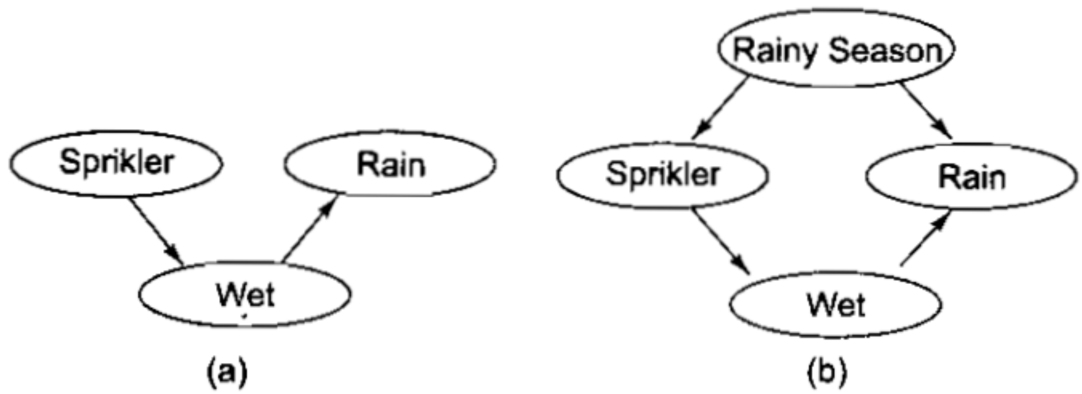
A Bayesian network (or a belief network) is a probabilistic graphical model that represents a set of variables and their probabilistic independencies. For example, a Bayesian network could represent the **probabilistic relationships** between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases.

Bayesian Networks are also called : Bayes nets, Bayesian Belief Networks (BBNs) or simply Belief Networks. Causal Probabilistic Networks (CPNs).

A Bayesian network consists of :

- a set of nodes and a set of directed edges between nodes.
- the edges reflect cause-effect relations within the domain.
- The effects are not completely deterministic (e.g. disease -> symptom).

- the strength of an effect is modeled as a probability.



- ✓ Bayesian belief networks specify joint conditional probability distributions.
- ✓ They allow class conditional independencies to be defined between subsets of variables.
- ✓ They provide a graphical model of causal relationships, on which learning can be performed.
- ✓ Trained Bayesian belief networks can be used for classification.
- ✓ Bayesian belief networks are also known as belief networks, Bayesian networks, and probabilistic networks.

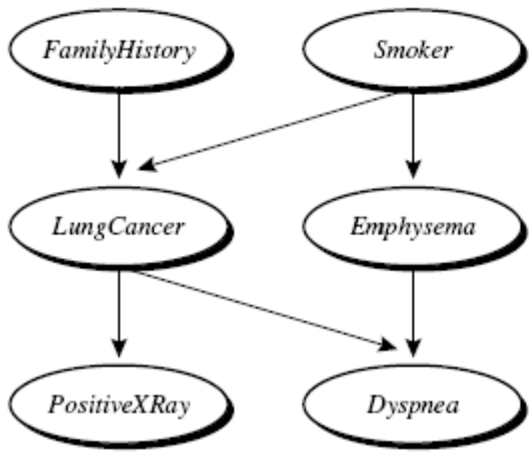
A belief network is defined by two components—

- ✓ A *directed acyclic graph*
- ✓ a set of *conditional probability tables.*

Each node in the directed acyclic graph represents a random variable. The variables may be discrete or continuous-valued. They may correspond to actual attributes given in the data or to "hidden variables" believed to form a relationship (e.g., in the case of medical data, a hidden variable may indicate a syndrome, representing a number of symptoms that, together, characterize a specific disease).

Each arc represents a probabilistic dependence. If an arc is drawn from a node *Y* to a node *Z*, then *Y* is a parent or immediate predecessor of *Z*, and *Z* is a descendant of *Y*. *Each variable is conditionally independent of its non descendants in the graph, given its parents.*

(a)



(b)

	FH, S	*FH, ~S*	*~FH, S*	*~FH, ~S*
LC	0.8	0.5	0.7	0.1
~LC	0.2	0.5	0.3	0.9

A simple Bayesian belief network: (a) A proposed causal model, represented by a directed acyclic graph. (b) The conditional probability table for the values of the variable *LungCancer (LC)* showing each possible combination of the values of its parent nodes, *FamilyHistory (FH)* and *Smoker (S)*. Figure is adapted from [RBKK95].

A belief network has one conditional probability table (CPT) for each variable. The CPT for a variable *Y* specifies the conditional distribution *P*(*Y*j*Parents*(*Y*)), where *Parents*(*Y*) are the parents of *Y*. Figure(b) shows a CPT for the variable *LungCancer*. The conditional probability for each known value of *LungCancer* is given for each possible combination of values of its parents. For instance, from the upper leftmost and bottom rightmost entries, respectively, we see that

$$P(LungCancer = yes \mid FamilyHistory = yes, Smoker = yes) = 0.8$$
$$P(LungCancer = no \mid FamilyHistory = no, Smoker = no) = 0.9$$