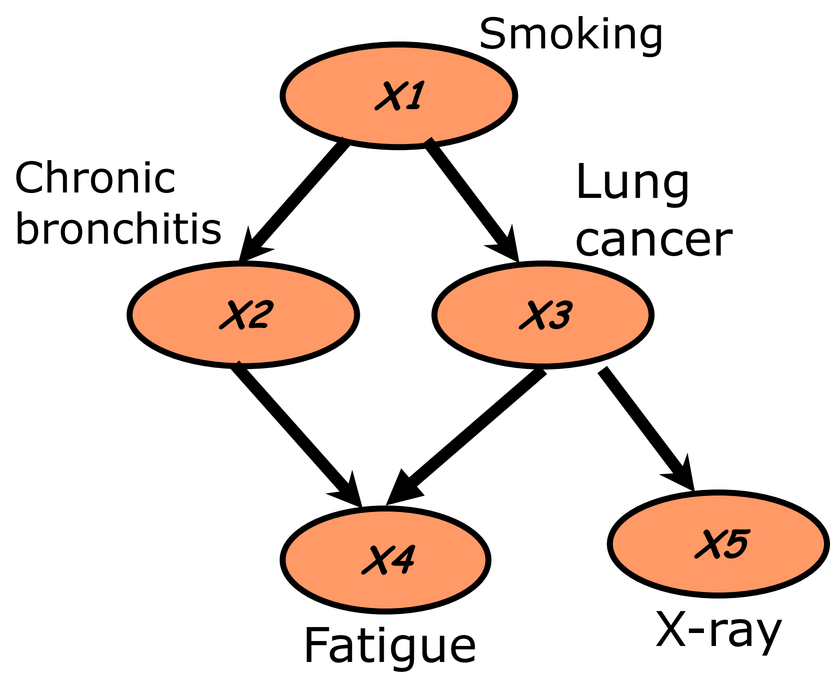
BAYESIAN NETWORKS
&
DECISION MAKING

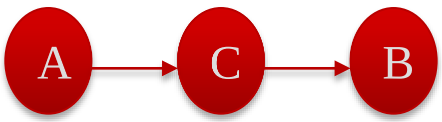
Bayesian Networks (BNs)

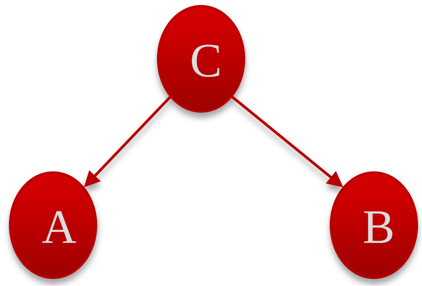
- Directed Acyclic Graphs (DAGs) for joint probability distributions
  - Nodes – random variables
  - Edges – direct dependence

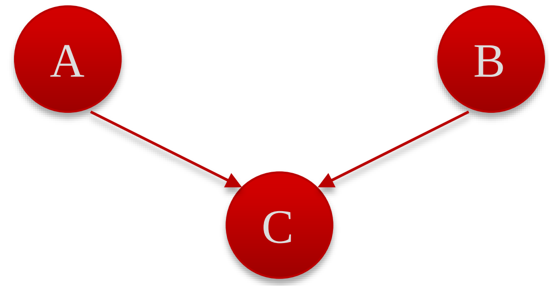


$$P(X1, X2, X3, X4, X5)$$
$$= P(X1)P(X2 \mid X1)P(X3 \mid X1)P(X4 \mid X2, X3)P(X5 \mid X3)$$

Independent Assumptions

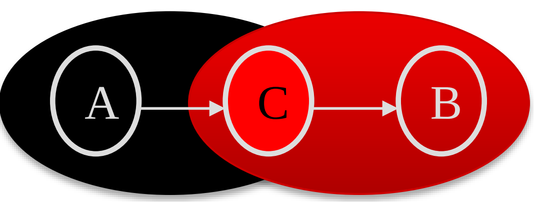


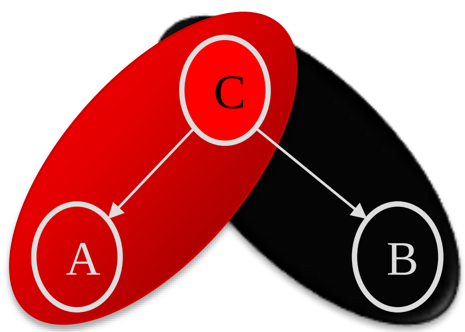


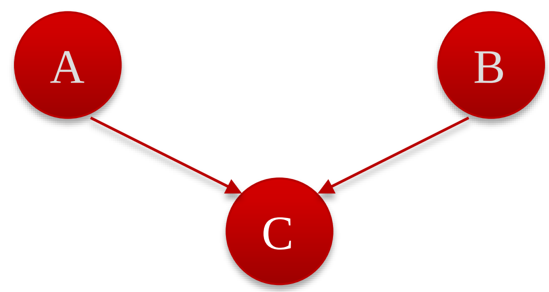


A and B are *marginally* dependent

A and B are *marginally* independent







A and B are *conditionally* independent

A and B are *conditionally* dependent

Inference in Belief Networks

- Recall that belief networks specify conditional independence between nodes (random variables).

- It also specifies the full joint distribution of variables:

$$P(x_n, \cdots, x_2, x_1) \;=\; \Pi_{i=1}^{n} P(x_i | Parents(x_i))$$

- The basic task of a belief network is as follows:

  - Compute the posterior probability for a query variable given an observed event

  - i.e., an **assignment of values** to a set of **evidence variables** while other variables are **not assigned values** (the so-called **hidden variables**).

Inference in Belief Networks

In other words…

- Let **E** denote a set of evidence values E_1, E_2, ⋯ , E_m. Let the observed event be **e** = $(e_1, e_2, \cdots, e_m)$.
- Let **Y** denote a set of non-evidence variables $Y_1, Y_2, \cdots, Y_l$.
- Let **X** denote the query variable.
- Hence, the belief network is composed of the nodes **X [E [Y**.
- The network specifies **P(X ^ E ^ Y)**.
- We want to compute (i.e., answer the query) **P(X | E); i.e., What is the posterior probability of X given the observed evidence E? Is it different than its prior probability?**

Inference in Belief Networks

- We need to note that:

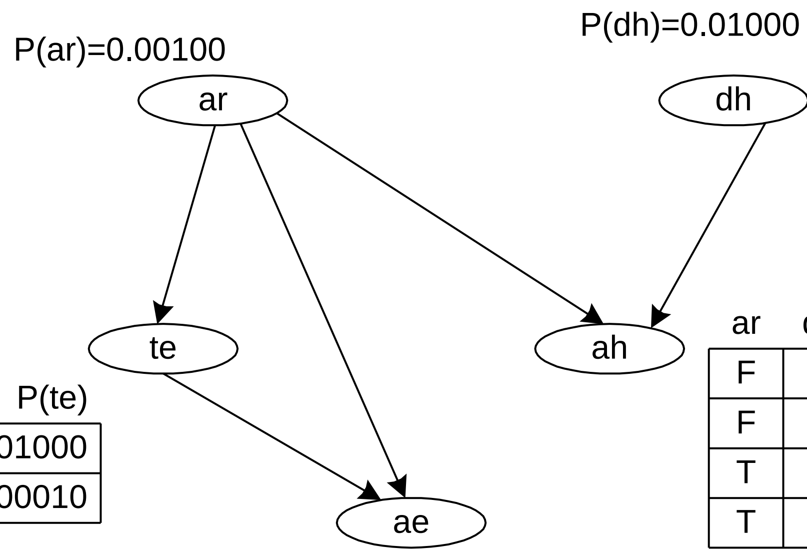
  $P(X|E) = \alpha\, P(X,E)$ where α is a normalization constant (this is just Bayes Theorem…)

- We can write this in terms of the full joint distribution if we sum out the non-evidence variables:

  $$P(X|E) = \alpha\, P(X,E) = \alpha \sum_{y} P(X,E,Y)$$

Example of Inference

- Consider the following belief network.



ar	P(te)
F	0.01000
T	0.00010

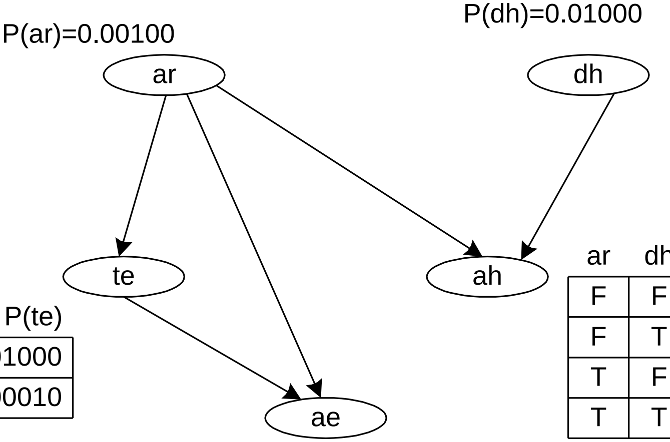
ar	dh	P(ah)
F	F	0.00001
F	T	0.99000
T	F	0.99000
T	T	0.10000

te	ar	P(ae)
F	F	0.00010
F	T	0.99000
T	F	0.99000
T	T	0.10000

Example of Inference

- Without making any observations, we can compute P(te):

$$
\begin{aligned}
P(te) &= P(te \wedge ar) + P(te \wedge \neg ar) \\
&= P(te|ar)P(ar) + P(te|\neg ar)P(\neg ar) \\
&= P(te|ar)P(ar) + P(te|\neg ar)(1 - P(ar)) \\
&= 0.0001(0.001) + 0.01(1 - 0.001) \\
&= 0.00999
\end{aligned}
$$



ar	P(te)
F	0.01000
T	0.00010

ar	dh	P(ah)
F	F	0.00001
F	T	0.99000
T	F	0.99000
T	T	0.10000

te	ar	P(ae)
F	F	0.00010
F	T	0.99000
T	F	0.99000
T	T	0.10000

Example of Inference

- However, say we have observed **ae** to be **TRUE**. We would like to find if **te** becomes more likely given this observation; i.e., we want to find **P(te|ae).**

- Network tells us that:

$$P(ae,te,ah,ar,dh) = P(dh)P(ar)P(ah|ar,dh)P(te|ar)P(ae|te,ar)$$

- So we want to find:

$$P(te|ae) = \alpha \sum_{ar} \sum_{ah} \sum_{dh} P(dh)P(ar)P(ah|ar,dh)P(te|ar)P(ae|te,ar)$$

- Note that we need to find **P(te|ae) and P(: te|ae)** in order to **normalize properly**.

- To find one of these values, we need to sum 8 terms, each involving 5 multiplications. All the information required is in the network.

Example of Inference

- We can compute P(te|ae):

ah	dh	ar	P(dh)	P(ar)	P(ah|ar,dh)	P(te|ar)	P(ae|te,ar)	
T	F	F	0.99000	0.99900	0.00001	0.01000	0.99000	0.00000009791199
T	F	T	0.99000	0.00100	0.99000	0.00010	0.10000	0.00000000980100
T	T	F	0.01000	0.99900	0.99000	0.01000	0.99000	0.00009791199000
T	T	T	0.01000	0.00100	0.10000	0.00010	0.10000	0.00000000001000
F	F	F	0.99000	0.99900	0.99999	0.01000	0.99000	0.00979110108801
F	F	T	0.99000	0.00100	0.01000	0.00010	0.10000	0.00000000009900
F	T	F	0.01000	0.99900	0.01000	0.01000	0.99000	0.00000098901000
F	T	T	0.01000	0.00100	0.90000	0.00010	0.10000	0.00000000009000

P(te|ae) = 0.00989011000000

- Similarly, we can compute P(:te|ae) = 0.000999693198000001.

- Hence, P(te|ae) = α < 0.00989011, 0.000999693198000001> = <0.9082,0.0918> and our belief that te=TRUE given observation of ae has increased to 0.9082 from its prior probability of 0.00999.

Example of Inference

- We could have also calculated **P(ar|ae) = 0.0909** and **P(dh|ae) = 0.01** in a similar fashion.
- Consider now that we observe both **ae** = **TRUE and ah = TRUE.**
- A similar calculation would give that **P(te|ae,ah) = 0.0917** (and the summation would have been over only four terms), **P(ar|ae,ah)=0.908** and **P(dh|ae,ah)=0.0926**.
- Note that the observation of additional evidence **ah** tends to "**explain away**" things.
  - Once **both ae** and **ah** are **observed**, a **single cause ar** that can **explain both the observations becomes very likely**.

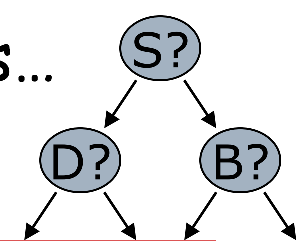
Classifying E-mail as Spam

- Suppose we are building a computer program to classify e-mails as being either "spam" or "ham."
  - "spam" = unwanted e-mails
  - "ham" = good e-mails

SPAM Vs. HAM

Classifying E-mail as Spam

- Some important features of an e-mail message might help us decide if it's spam:
  - Does it contain the word "bargain" ? B=true/false
  - Does it contain the word "discount" ? D=true/false
  - Does it contain the word "sale" ? S=true/false
- Maybe we could build a computer program that could learn the likelihood that an e-mail is 'spam' based on whether it contains these words.
  - We could definitely do it with decision trees…
  - Is there another way?



Classifying E-mail as Spam

- We take a big sample of e-mails:
  - We know that 2000 of them are 'spam'
  - We know that 4000 of them are 'ham'
- For each of these groups, we calculate:
  - Number of times each of the words "bargain," "discount," and "sale" occur in a message.

	Total	# "bargain"	# "discount"	# "sale"
SPAM	2000	800	280	1100
HAM	4000	300	34	500

What do we know?

- We can calculate:

  P(spam) = 2000/6000 = .3333

  P(ham) = 4000/6000 = .6667

  P(B)= (800+300)/6000 = .1833

  P(D)= (280+34)/6000 = .0523

  P(S)= (1100+500)/6000 = .2667

  Remember we had 6000 e-mails in our sample.

What do we know?

- From our table, we can calculate:

  $P(B|spam) = P(B \wedge spam) / P(spam)$ Conditioning Rule

  $P(B|spam) = (800/6000) / P(spam)$ From Table

  $P(B|spam) = (.1333) / (.3333)$ From Previous Slide

  $P(B|spam) = .4000$

  Same for $P(D|spam)$, $P(S|spam)$, $P(B|ham)$, etc.

What do we know?

$P(\text{spam})$, $P(\text{ham})$
$P(B|\text{spam})$, $P(B|\text{ham})$,
$P(D|\text{spam})$, $P(D|\text{ham})$,
$P(S|\text{spam})$, $P(S|\text{ham})$.

We also know $P(B)$, $P(D)$, $P(S)$, but we're not *going* to use these in this problem.

We also figure out values like $P(\neg D)$ or $P(\neg D|\text{spam})$ by subtracting a value from 1.

What will our task be?

- We'll give the program a new e-mail, and it will have to decide the probability it is spam or ham.
  - The computer program can easily check whether the words "bargain," "discount," and "sale" occur inside the new message. So, it will know whether B or $\neg$B, etc.

- For instance some new e-mail might contain "bargain" and "sale" but not "discount."
  - We'll need to decide:

$$P(\text{ spam} \mid B \wedge S \wedge \neg D)$$

$$P(\text{ ham} \mid B \wedge S \wedge \neg D)$$

Actually, we just need one or the other, but let's see how to calc both.

We're going to use Bayes' Rule

$$\frac{P(\ B \wedge S \wedge \neg D \mid \text{spam}\ )\,P(\ \text{spam}\ )}{P(\ B \wedge S \wedge \neg D\ )}$$

So, we'll say that
the e-mail is SPAM,
if this number is bigger.

Compared to…

Same
Denominator

$$\frac{P(\ B \wedge S \wedge \neg D \mid \text{ham}\ )\,P(\ \text{ham}\ )}{P(\ B \wedge S \wedge \neg D\ )}$$

Or, we'll say that
the e-mail is HAM,
if this number is bigger.

We're going to use Bayes' Rule

$$P(B \wedge S \wedge \neg D \mid spam)\, P(spam)$$

We know
these values.

So, we'll say that
the e-mail is SPAM,
if this number is bigger.

Compared to…

$$P(B \wedge S \wedge \neg D \mid ham)\, P(ham)$$

Or, we'll say that
the e-mail is HAM,
if this number is bigger.

We're going to use Bayes' Rule

$P(B \wedge S \wedge \neg D \mid spam) (.3333)$

So, we'll say that
the e-mail is SPAM,
if this number is bigger.

Compared to…

$P(B \wedge S \wedge \neg D \mid ham) (.6667)$

Or, we'll say that
the e-mail is HAM,
if this number is bigger.

Estimation Trick

$P(\, B \wedge S \wedge \neg D \mid \text{spam} \,)$ (.3333)

So, we'll say that
the e-mail is SPAM,
if this number is bigger.

Compared to…

$P(\, B \wedge S \wedge \neg D \mid \text{ham} \,)$ (.6667)

This final step involves a trick we use to calculate a probability of the form above.

Or, we'll say that
the e-mail is HAM,
if this number is bigger.

Estimation Trick

$P(B \wedge S \wedge \neg D \mid spam)$ (.3333)

So, we'll say that
the e-mail is SPAM,
if this number is bigger.

Compared to…

$P(B \wedge S \wedge \neg D \mid ham)$ (.6667)

Or, we'll say that
the e-mail is HAM,
if this number is bigger.

We don't have a way to really calculate this. We could have counted when each of the 8 combinations of B,S,D occurred in our data, but we didn't. We kept records for B, S, and D separately. Now we need to combine them.

Estimation Trick

$P(B \wedge S \wedge \neg D \mid spam)$ (.3333)

So, we'll say that
the e-mail is SPAM,
if this number is bigger.

Compared to…

$P(B \wedge S \wedge \neg D \mid ham)$ (.6667)

Or, we'll say that
the e-mail is HAM,
if this number is bigger.

The Trick:
We can estimate this: $P(B \wedge S \wedge \neg D \mid ham)$
by calculating this: $P(B|ham) * P(S|ham) * (\neg D|ham)$
It's called the **Naïve Bayes Assumption**… We're being naïve by making the assumption that the variables are independent and this estimate is good enough.

What Does Independent Mean?

- It means that for any two variables, the probability of one is not affected by whether or not the other one is true.
  - So, this means $P(A) = P(A|B) = P(A|\neg B)$
  - If our variables are words that often co-occur in sentences, then this is probably not the case, ("George" "Bush"), but we're going to pretend that the probabilities of our words don't affect one another so that we can use this special estimation trick.

Estimation Trick

$P(\, B \wedge S \wedge \neg D \mid spam\,)(\ .3333\ )$

So, we'll say that
the e-mail is SPAM,
if this number is bigger.

Compared to…

$P(\, B \wedge S \wedge \neg D \mid ham\,)\ (\ .6667\ )$

Or, we'll say that
the e-mail is HAM,
if this number is bigger.

Estimates:

$P(B \wedge S \wedge \neg D \mid spam) = P(B|spam)*P(S|spam)*P(\neg D|spam)$
$= 0.4 * 0.55 * 0.86 = \mathbf{0.1892}$

$P(B \wedge S \wedge \neg D \mid ham) = P(B|ham)*P(S|ham)*P(\neg D|ham)$
$= .075 * .125 * .9915 = \mathbf{0.0093}$

Estimation Trick

(.1892) * (.3333)

So, we'll say that
the e-mail is SPAM,
if this number is bigger.

Compared to…

Or, we'll say that
the e-mail is HAM,
if this number is bigger.

(.0093) * (.6667)

Estimates:

$P(B \wedge S \wedge \neg D \mid spam) = P(B|spam)*P(S|spam)*P(\neg D|spam)$
$= 0.4 * 0.55 * 0.86 = \mathbf{0.1892}$

$P(B \wedge S \wedge \neg D \mid ham) = P(B|ham)*P(S|ham)*P(\neg D|ham)$
$= .075 * .125 * .9915 = \mathbf{0.0093}$

Estimation Trick

(.1892) * (.3333) = **.0631**

So, we'll say that
the e-mail is SPAM,
if this number is bigger.

Compared to…

(.0093) * (.6667) = **.0062**

Or, we'll say that
the e-mail is HAM,
if this number is bigger.

Estimates:

$P(B \wedge S \wedge \neg D \mid spam) = P(B|spam)*P(S|spam)*P(\neg D|spam)$
$= 0.4 * 0.55 * 0.86 = \mathbf{0.1892}$
$P(B \wedge S \wedge \neg D \mid ham) = P(B|ham)*P(S|ham)*P(\neg D|ham)$
$= .075 * .125 * .9915 = \mathbf{0.0093}$

Estimation Trick

It's SPAM!

(.1892) * (.3333)= **.0631**

Compared to…

So, we'll say that the e-mail is SPAM, if this number is bigger.

(.0093) * (.6667) = **.0062**

Or, we'll say that the e-mail is HAM, if this number is bigger.

Estimates:

$P(B \wedge S \wedge \neg D \mid spam) = P(B|spam)*P(S|spam)*P(\neg D|spam)$
$= 0.4 * 0.55 * 0.86 = \mathbf{0.1892}$

$P(B \wedge S \wedge \neg D \mid ham) = P(B|ham)*P(S|ham)*P(\neg D|ham)$
$= .075 * .125 * .9915 = \mathbf{0.0093}$

Being Naïve Can Help

- We built a Naïve Bayes Classifer
  - Being naïve and assuming our variables were independent allowed us to build something smart.

The important point…

- Bayes' Rule was at the heart of this e-mail "classifier."
  - It is what helped us flip probabilities like this:
    $$P(\text{spam} \mid B \wedge S \wedge \neg D)$$
  - Into equations like this:
    $$P(B \wedge S \wedge \neg D \mid \text{spam})\, P(\text{spam}) \,/\, P(B \wedge S \wedge \neg D)$$
  - It let us <u>reverse</u> the direction of the conditional.

- Bayes' Rule is central to the working of many other machine learning algorithms in AI.

The Skills You Need

- Probability Calculations
  - Working with Bayes' Rule.
  - How to go from data you collect to: P(D) or P(D|ham).
  - Naïve Bayes Assumption
  - Naïve Bayes Classifier

MYCIN

- Developed at Stanford University in 1972
- Regarded as the first true “expert system”
- **Assist physicians in the treatment of blood infections**
- Many revisions and extensions to MYCIN over the years

The MYCIN Problem

- Physician wishes to specify an "antimicrobial agent" - basically an antibiotic - to kill bacteria or arrest their growth
- Some agents are poisonous!
- No agent is effective against all bacteria
- Most physicians are not expert in the field of antibiotics

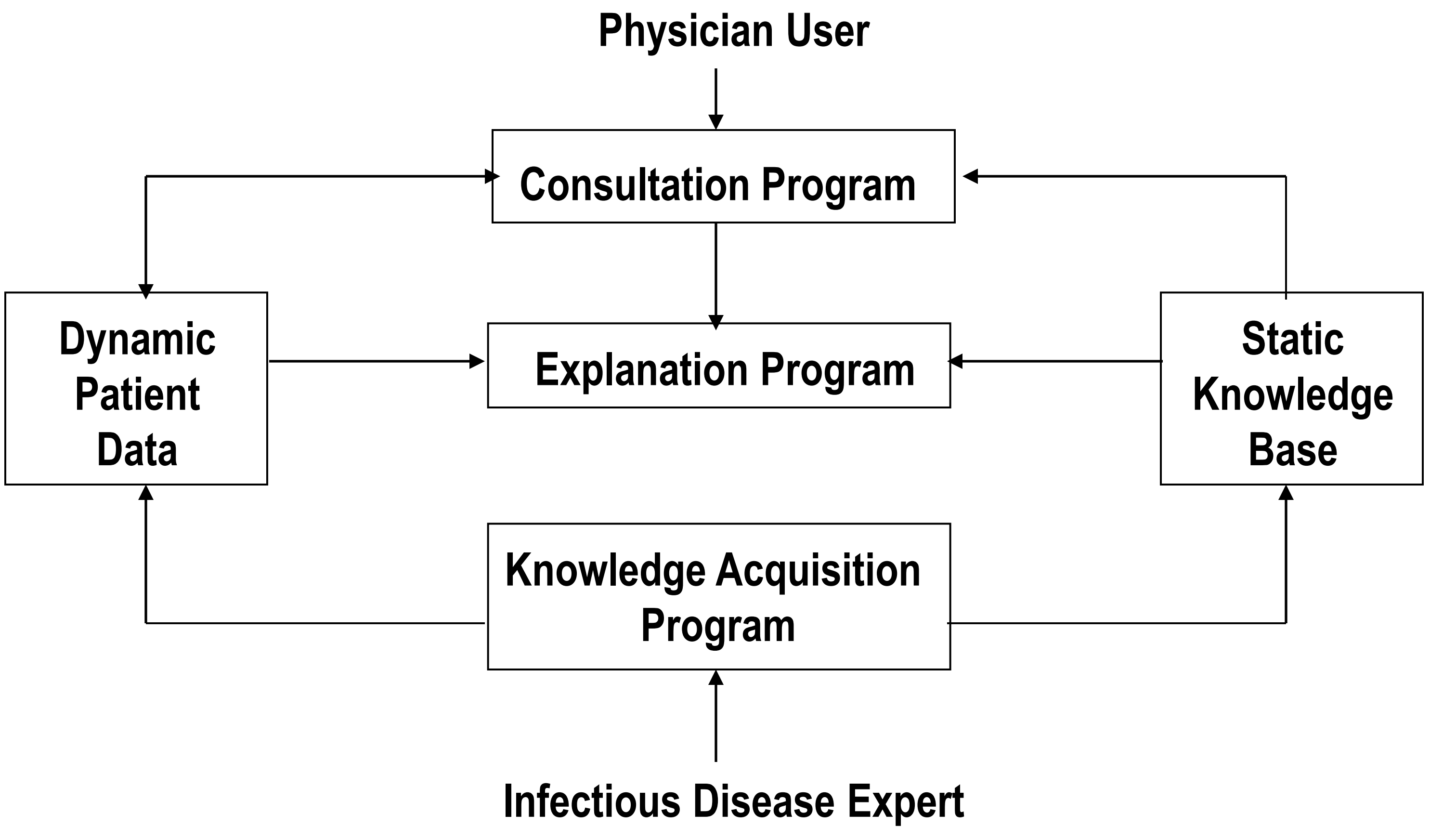
The Decision Process

- There are four questions in the process of deciding on treatment:
  - **Does the patient have a significant infection?**
  - **What are the organism(s) involved?**
  - **What set of drugs might be appropriate to treat the infection?**
  - **What is the best choice of drug or combination of drugs to treat the infection?**

MYCIN Components

- KNOWLEDGE BASE:
  - facts and knowledge about the domain
- DYNAMIC PATIENT DATABASE:
  - information about a particular case
- CONSULTATION PROGRAM:
  - asks questions, gives advice on a particular case
- EXPLANATION PROGRAM:
  - answers questions and justifies advice
- KNOWLEDGE ACQUISITION PROGRAM:
  - adds new rules and changes exisiting rules

Basic MYCIN Structure



The MYCIN Knowledge Base

- Where the rules are held
- Basic rule structure in MYCIN is:

  if **condition$_1$** ***and….and*** **condition$_m$** ***hold***

  then draw **conclusion$_1$** ***and….and*** **condition$_n$**
- Rules written in the LISP programming language
- Rules can include certainty factors to help weight the conclusions drawn

An Example Rule

IF:(1) The stain of the organism is Gram negative, and

(2) The morphology of the organism is rod, and

(3) The aerobicity of the organism is aerobic

THEN:

There is strongly suggestive evidence (0.8) that the class of the organism is *Enterobacteria*

Calculating Certainty

- Rule certainties are regarded as probabilities
- Therefore must apply the rules of probability in combining rules
- Multiplying probabilities which are less than certain results in lower and lower certainty!
- Eg $0.8 \times 0.6 = 0.48$

Other Types of Knowledge

- Facts and definitions such as:
  - lists of all organisms known to the system
  - “knowledge tables” of clinical parameters and the values they can take (eg morphology)
  - classification system for clinical parameters and the context in which they are applied (eg referring to patient or organism)
- Much of MYCIN’s knowledge refers to 65 clinical parameters

Evaluating MCYIN

- Many studies show that MYCIN's recommendations compare favourably with experts for diseases like meningitis
- Study compared on real patients with expert and non-expert physicians:
  - MYCIN matched experts
  - MYCIN was better than non-experts

MYCIN Limitations

- A research tool - never intended for practical application
- Limited knowledge base - only covers a small number of infectious diseases

Pathfinder

- Pathfinder system. (Heckerman 1991, Probabilistic Similarity Networks,
- MIT Press, Cambridge MA).
- Diagnostic system for lymph-node diseases.
- 60 diseases and 100 symptoms and test-results.
- 14,000 probabilities
- Expert consulted to make net.
- 8 hours to determine variables.
- 35 hours for net topology.
- 40 hours for probability table values.
- Apparently, the experts found it quite easy to invent the causal links and probabilities.
- Pathfinder is now outperforming the world experts in diagnosis. Being extended to several dozen other medical domains.

Decision Making

Decision Making

Decision making may be defined as:

- Intentional and reflective choice in response to perceived needs (Kleindorfer *et al.*, 1993)
- Decision maker's (DM's) choice of one alternative or a subset of alternatives among all possible alternatives with respect to her/his goal or goals (Evren and Ülengin, 1992)
- Solving a problem by choosing, ranking, or classifying over the available alternatives that are characterized by multiple criteria (Topcu, 1999)
- Topics:
  - Maximum Expected Utility (MEU)
  - Decision network
  - Making decisions
    - Russell & Norvig, chapter 16

Acting Under Uncertainty

- With no uncertainty, rational decision is to pick action with “best” outcome
  - Two actions
    - #1 leads to great outcome
    - #2 leads to good outcome
  - It’s only rational to pick #1
  - Assumes outcome is 100% certain
- With uncertainty, it’s a little harder
  - Two actions
    - #1 has 1% probability to lead to great outcome
    - #2 has 90% probability to lead to good outcome
  - What is the rational decision?

Acting Under Uncertainty

- Maximum Expected Utility (MEU)
  - Pick action that leads to best outcome averaged over all possible outcomes of the action
- How do we compute the MEU?
  - Easy once we know the probability of each outcome and their utility

Utility

- Value of a state or outcome
- Computed by utility function
  - $U(S)$ = utility of state S
  - $U(S) \in [0,1]$ if normalized

Expected Utility

- Sum of utility of each possible outcome times probability of that outcome
  - Known evidence E about the world
  - Action A has i possible outcomes, with probability $P(Result_i(A)|Do(A),E)$
  - Utility of each outcome is $U(Result_i(A))$
    - Evaluation function of the state of the world given $Result_i(A)$
- $EU(A|E)=\sum_i P(Result_i(A)|Do(A),E)\ U(Result_i(A))$

Maximum Expected Utility

- List all possible actions A_j
- For each action, list all possible outcomes $Result_i(A_j)$
- Compute $EU(A_j|E)$
- Pick action that maximises EU

Utility of Money

- Use money as measure of utility?
- Example
  - A_1 = 100% chance of \$1M
  - A_2 = 50% chance of \$3M or nothing
  - $EU(A_2)$ = \$1.5M > \$1M = $EU(A_1)$
- Is that rational?

Utility of Money

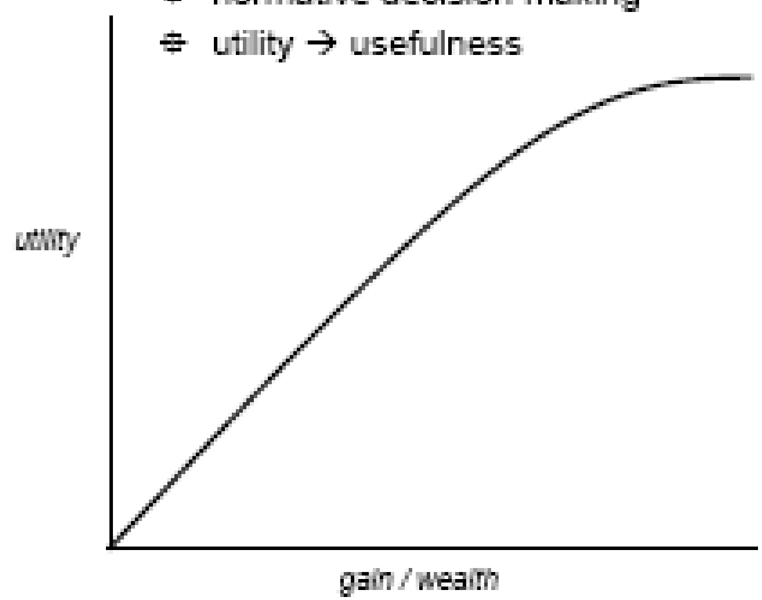
Decreasing Marginal Utility

Daniel Bernoulli (1738) claimed that the utility of money declines with amount won. Utility is the logarithmic function for the money possessed.

n	P	Pr	U	EU
1	1/2	$2	0.30	0.15
2	1/4	$4	0.62	0.15
5	1/32	$32	1.50	0.04
9	1/512	$512	2.71	0.005
10	1/1024	$1024	3.01	0.003

U = log (Pr)

- normative decision making
- utility → usefulness



Axioms

- Given three states A, B, C
- A ❺ B
  - The agent prefers A to B
- A ~ B
  - The agent is indifferent between A and B
- A ❽ B
  - The agent prefers A to B or is indifferent between A and B
- [p_1, A; p_2, B; p_3, C]
  - A can occur with probability p_1, B can occur with probability p_2, C can occur with probability p_3

Axioms

- Orderability
  - (A ❺ B) ∨ (B ❺ A) ∨ (A ~ B)
- Transitivity
  - (A ❺ B) ∧ (B ❺ C) ⇒ (A ❺ C)
- Continuity
  - A ❺ B ❺ C ⇒ ∃p [p, A; 1-p, C] ~ B
- Substituability
  - A ~ B ⇒ [p, A; 1-p, C] ~ [p, B; 1-p, C]
- Monotonicity
  - A ❺ B ⇒ (p ≥ q ⇔ [p, A; 1-p, B] ❽ [q, A; 1-q, B])
- Decomposability
  - [p, A; 1-p, [q, B; 1-q, C]] ~ [p, A; (1-p)q, B; (1-p)(1-q), C]

Axioms

- Utility principle
  - $U(A) > U(B) \Leftrightarrow A$ ❺ B
  - $U(A) = U(B) \Leftrightarrow A \sim B$
- Maximum utility principle
  - $U([p_1, A_1; \ldots ; p_n, A_n]) = \sum_i p_i U(A_i)$
- Given these axioms, MEU is rational!

Decision Network

- Our agent makes decisions given evidence
  - Observed variables and conditional probability tables of hidden variables
- Similar to conditional probability
  - Probability of variables given other variables
  - Relationships represented graphically in Bayesian network
- Could we make a similar graph here?

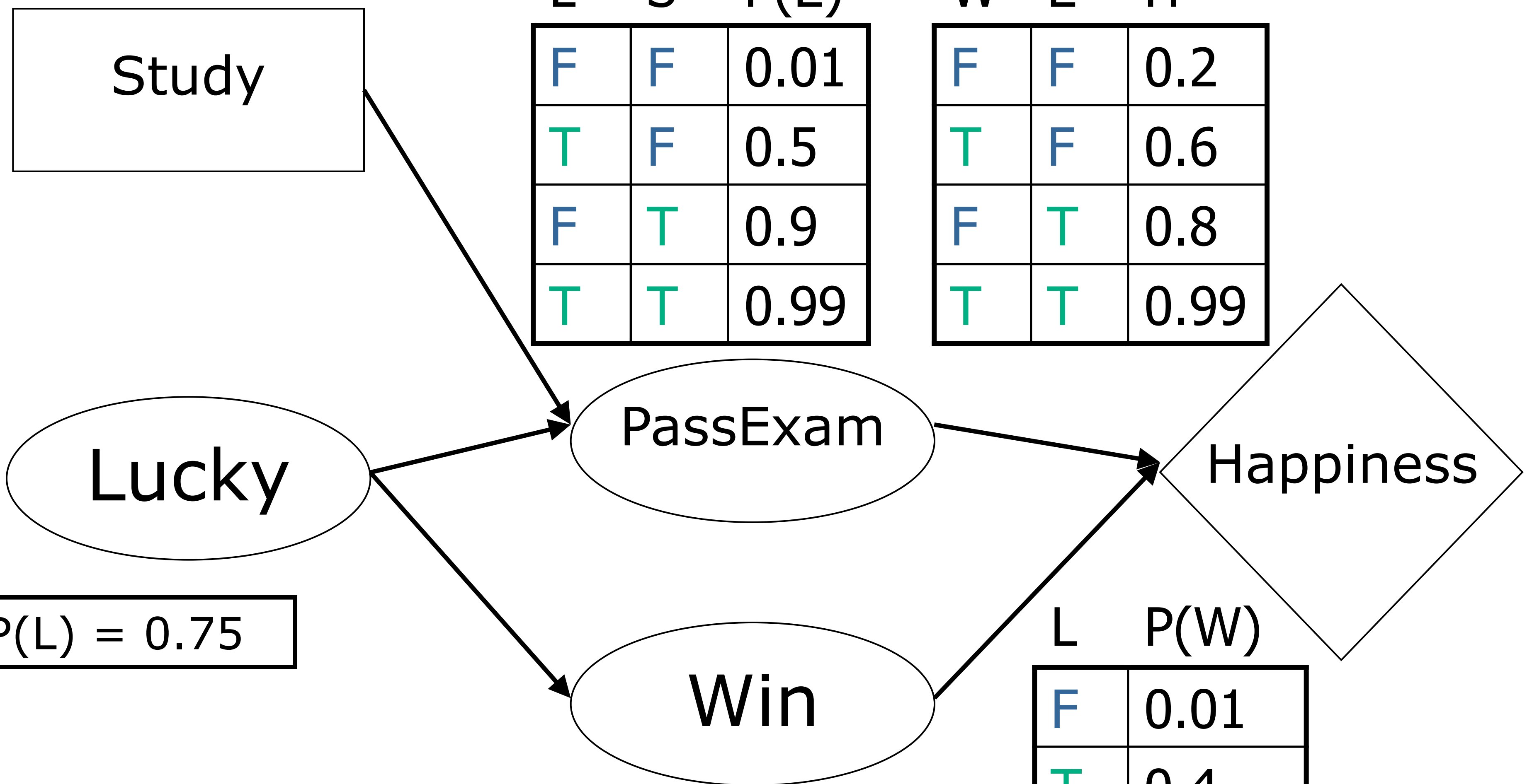
Decision Network

- Sometimes called *influence diagram*
- Like a Bayesian Network for decision making
  - Start with variables of problem
  - Add decision variables that the agent controls
  - Add utility variable that specify how good each state is

Decision Network

- Chance node (oval)
  - Uncertain variable
  - Like in Bayesian network
- Decision node (rectangle)
  - Choice of action
  - Parents: variables affecting decision, evidence
  - Children: variables affected by decision
- Utility node (diamond)
  - Utility function
  - Parents: variables affecting utility
  - Typically only one in network

Decision Network Example



L	S	P(E)
F	F	0.01
T	F	0.5
F	T	0.9
T	T	0.99

W	E	H
F	F	0.2
T	F	0.6
F	T	0.8
T	T	0.99

L	P(W)
F	0.01
T	0.4

P(L) = 0.75

Decision Network Example