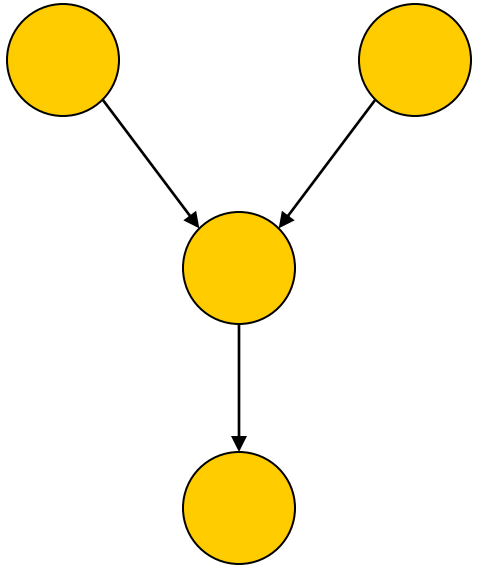


Bayesian Networks
Part I: Probability Primer

Prof. Dr. Peter Haddawy

Faculty of ICT

Mahidol University

Some material adopted from slides by Andrew Moore, Weng-Kee Wong, and Darlene Goldstein.

Introduction

Suppose you are trying to determine if a patient has pneumonia. You observe the following symptoms:

- The patient has a cough
- The patient has a fever
- The patient has difficulty breathing

Introduction

You want to determine how likely it is that the patient has pneumonia given that the patient has a cough, a fever, and difficulty breathing.

We are not 100% certain that the patient has pneumonia because of these symptoms. We are dealing with uncertainty.

Introduction

Now suppose you order a chest x-ray and the results are positive.

Your belief that that the patient has pneumonia is now much higher.

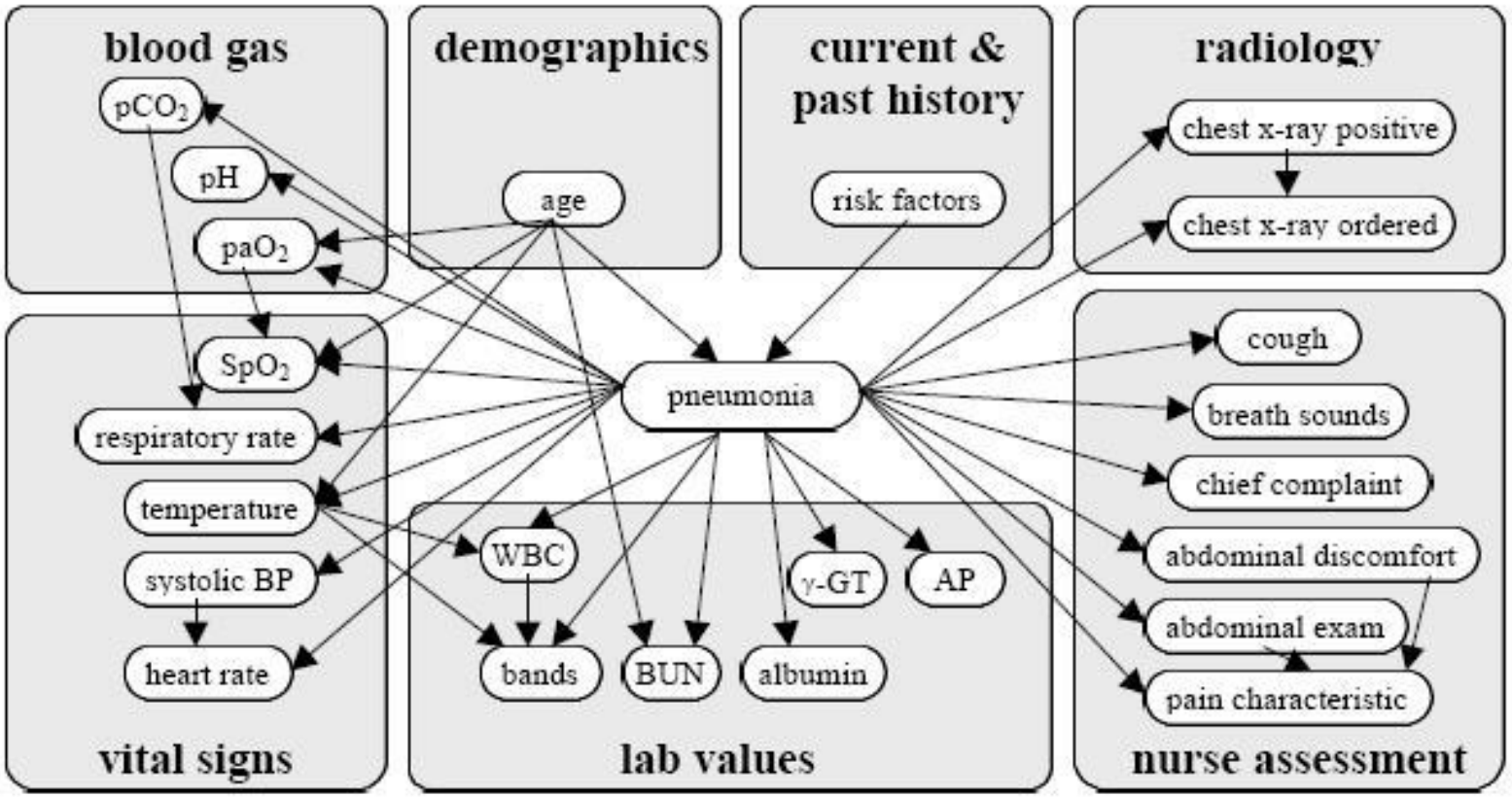
Introduction

- In the previous slides, what you observed affected your belief that the patient has pneumonia
- This is called reasoning with uncertainty
- Wouldn't it be nice if we had some methodology for reasoning with uncertainty? Well, in fact we do...

Bayesian Networks

- Bayesian networks help us reason with uncertainty
- In the opinion of many AI researchers, Bayesian networks are one of the most significant contributions in AI in the last 20 years
- They are used in many applications:
  - Spam filtering / Text mining
  - Speech recognition
  - Robotics
  - Diagnostic systems
  - Disease surveillance
  - Intelligent tutoring

Bayesian Networks (An Example)



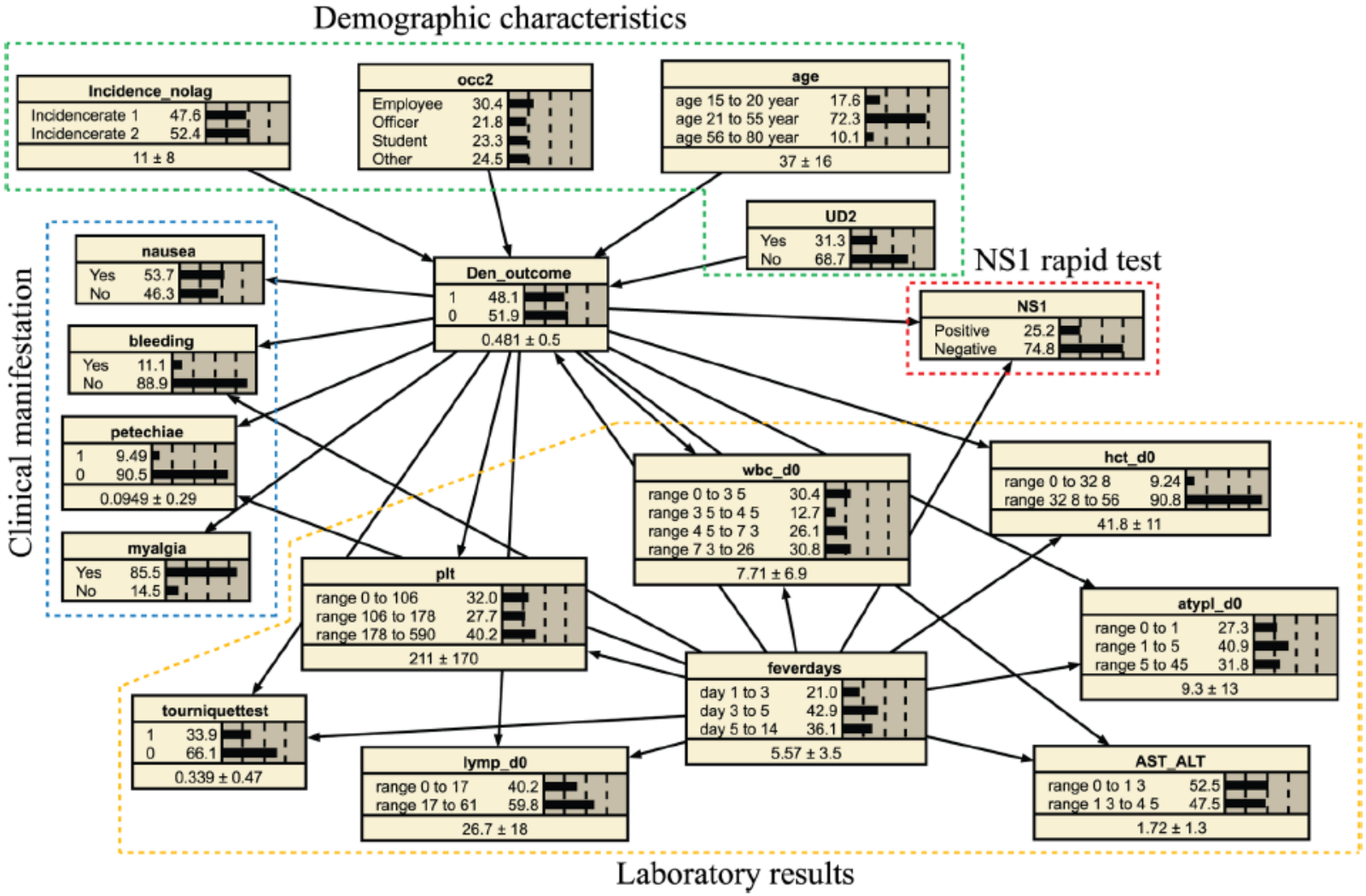


Fig 1. Final Bayesian network model for dengue diagnosis.

C. Sa-ngamuang, P. Haddawy, V. Luvira, W. Piyaphanee, S. Iamsirithaworn, S. Lawpoolsri, Accuracy of Dengue Clinical Diagnosis with and without NS1 Antigen Rapid Test: Comparison between Human and Bayesian Network Model Decision, *PLOS Neglected Tropical Diseases*, 12(6): e0006573, June 2018.

Types of Classification Algorithms

- Memory based
  - Define a distance between samples
  - K Nearest Neighbor
- Decision surface
  - Find best partition of the space
  - Decision trees
  - SVM
- Generative models
  - Induce a model and impose a decision rule
  - Bayesian Networks

Explainable AI (XAI)

- An AI system for which humans can understand the decisions or predictions made.
- In many domains, we need to understand the decisions to build trust in the algorithm
- This is also related to liability
  - medicine, defense, finance, law
- In 2018 the EU introduced a right to explanation in the General Data Protection Right (GDPR)

Types of Models

- Non-interpretable
  - Neural nets
  - Ensemble techniques (e.g. Random Forest)
  - But tend to be the most powerful

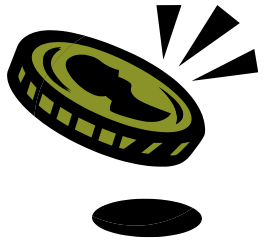
- Interpretable (Explainable)
  - Decision trees
  - KNN
  - Association rules
  - Bayesian Networks

Outline

- Probability basics
- Bayesian network representational concepts
- Building models
- Modeling Techniques
- Evaluating Models
- Explanation
- Case Studies (time permitting)

Probability Basics: Random Variables

- A random variable is the basic element of probability
- Refers to an event for which there is some degree of uncertainty as to its outcome
- For example, the random variable *A* could be the event of getting a heads on a coin flip



Probability Basics

- Describe the **state of the world** with a set of value assignments to a set of random variables.
  - Medicine: status of patient, test results
  - Weather: temperature, humidity, wind direction, rain
- A **random variable** is a variable that can take on exactly one value from a set of mutually exclusive and exhaustive values
  - Temp = {<25, 25 - 35, >35}
  - Humidity = {low, med, high}
  - Wind = {N, S, E, W}
  - Rain = {Yes, No}
  - One possible state is one complete assignment of values to the random variables.
  - (Temp >35, Humidity = med, Wind = N, Rain = Yes)

Probability Basics

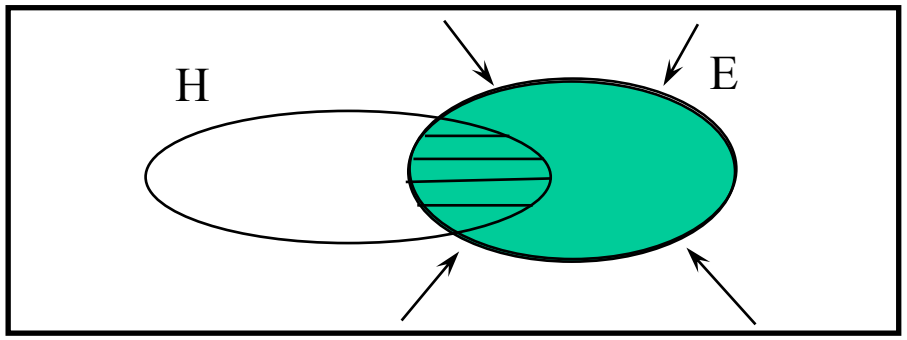
- **Symbols:**
  - We will use capital letters (A, B, …) for random variables and lower-case letters (a, b, …) for their values. E.g. $A = a_1$
  - A,B: A *and* B
  - So P(A,B) means the probability of A *and* B
- We will write A when we mean any value of the random variable A
  - P(Temperature) means the probability of any value of Temperature
  - P(Temperature = <38) means the probability of a particular value

Probability Basics

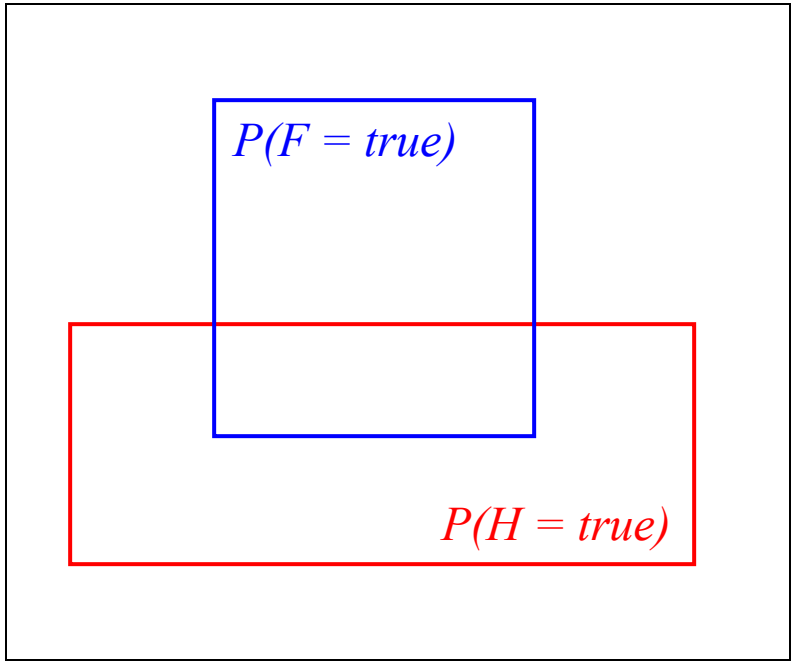
- **Axioms of probability**
  - $P(A) \geq 0$, where A is any proposition
  - $P(T) = 1$
  - $P(A \text{ or } B) = P(A) + P(B)$ if A and B are mutually exclusive

Conditional Probability

- P(*H* = *true* | *E* = *true*) = Out of all the outcomes in which *E* is true, how many also have *H* equal to true
- Read this as: "Probability of *H* conditioned on *E*" or "Probability of *H* given *E*"
- A measure of how relevant E is to H



Conditional Probability



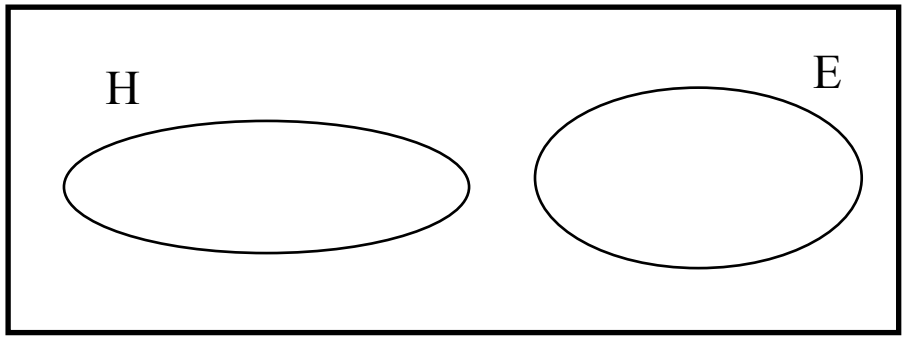
$$P(H=true|F=true)$$

$$= \frac{\text{Area of "H and F" region}}{\text{Area of "F" region}}$$

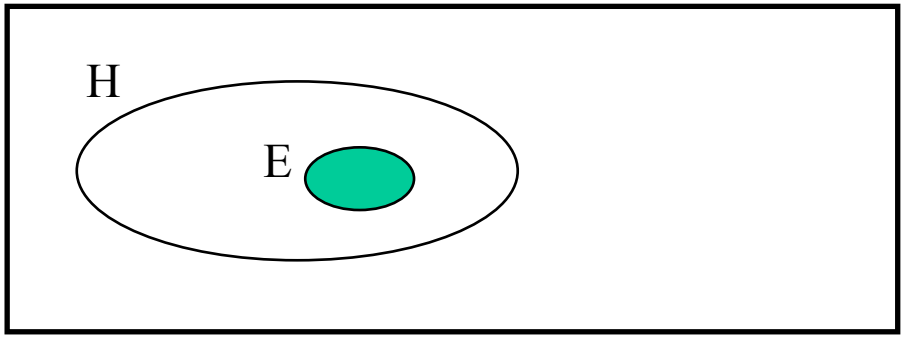
$$= \frac{P(H=true, F=true)}{P(F=true)}$$

In general, $P(X|Y)=P(X,Y)/P(Y)$

Conditional Probability



P(H|E) = 0



P(H|E) = 1

The Joint Probability Distribution

- If we have a probabilistic knowledge base, we would like to be able to query the probability of any combination of variables

  $P(A = true , B = true)$

  $P(A = false \mid B = true, C = false)$

  $P(A = true \mid B = true, C = true)$
- To be able to compute any such query, we need a specification of the full joint distribution
- Notice that each combination is a Possible World
- The probabilities of the possible worlds must sum to 1. Why?

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Sums to 1

The Joint Probability Distribution

- Once you have the joint probability distribution, you can calculate any probability involving *A*, *B*, and *C*

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Examples of things you can compute:

- *P(A=true)* = sum of *P(A,B,C)* in rows with *A=true*
- *P(A=true, B = true | C=true)* =

  P(A = true, B = true, C = true) / P(C = true) = prob of row 8 / sum of P(A,B,C) in rows with C = true

Updating Beliefs

- Beliefs are not static; they change with new information
- If our initial belief in H is P(H) and we observe E, then our new belief in H should be P(H|E).
- Model of an intelligent agent:
  - An agent has some current beliefs about the possible states of the world. "prior probabilities"
  - It makes some observation
  - It updates its by conditioning on that observation
  - Continues updating as it makes more observations

	A	**B**	**C**	**P(A,B,C)**
W1	false	false	false	0.1
W2	false	false	true	0.2
W3	false	true	false	0.05
W4	false	true	true	0.05
W5	true	false	false	0.3
W6	true	false	true	0.1
W7	true	true	false	0.05
W8	true	true	true	0.15

Why this makes sense

- Suppose an agent's beliefs are represented by the table
- So P(A=true) = W5+W6+W7+W8 =

  .3 + .1 + .05 + .15 = .6
- Suppose the agent observes that C = true
- What should the agent's probability of A = true now be? Call it P′.
- The probabilities of the possible worlds consistent with C=true must sum to one.
- Note that the information that C=true does not change the relative probabilities of the remaining possible worlds, e.g. W2 should still be twice as likely as W6
- We can make the probabilities sum to 1 and keep the relative values if we normalize by the sum of the remaining possible worlds

	A	B	C	P(A,B,C)
W1	false	false	false	0.1
W2	false	false	true	0.2
W3	false	true	false	0.05
W4	false	true	true	0.05
W5	true	false	false	0.3
W6	true	false	true	0.1
W7	true	true	false	0.05
W8	true	true	true	0.15

Why this makes sense

So P′(Wi) = P(Wi)/[P(W2) + P(W4) + P(W6) + P(W8)]

For example P′(W2) = .2/(.2 + .05 + .1 + .15) = .2/.5 = .4

And

P′(A=true) = [P(W6) + P(W8)])/[P(W2) + P(W4) + P(W6) + P(W8)] = .5

What do we get by conditioning?

P′(A=true) = P(A=true|C=true) = P(A=true,C=true)/P(C=true) =

[P(W6) + P(W8)])/[P(W2) + P(W4) + P(W6) + P(W8)] = .5

The same result!

	A	B	C	P(A,B,C)
W1	false	false	false	0.1
W2	false	false	true	0.2
W3	false	true	false	0.05
W4	false	true	true	0.05
W5	true	false	false	0.3
W6	true	false	true	0.1
W7	true	true	false	0.05
W8	true	true	true	0.15

The Problem with Joint Distributions

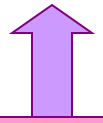
- Lots of entries in the table to fill up!
- For *k* Boolean random variables, you need a table of size 2^k
- 100 variables – impossible!
- How do we use fewer numbers?
- Need the concept of *independence*

A	B	C	P(A,B,C)
false	false	false	0.1
false	false	true	0.2
false	true	false	0.05
false	true	true	0.05
true	false	false	0.3
true	false	true	0.1
true	true	false	0.05
true	true	true	0.15

Independence

Variables A and B are independent if any of the following hold:

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



This says that knowing the outcome of A does not tell me anything new about the outcome of B.

Independence

How is independence useful?

- Suppose you have n coin flips and you want to calculate the joint distribution $P(C_1, \ldots, C_n)$
- If the coin flips are not independent, you need 2^n values in the table
- If the coin flips are independent, then

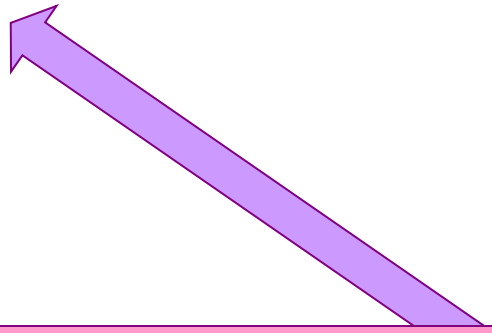
$$P(C_1,\ldots,C_n) = \prod_{i=1}^{n} P(C_i)$$

Each P(C_i) table has 2 entries and there are n of them for a total of $2n$ values

Conditional Independence

Variables A and B are conditionally independent given C if any of the following hold:

- $P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$
- $P(A \mid B, C) = P(A \mid C)$
- $P(B \mid A, C) = P(B \mid C)$



Knowing C tells me everything about A. I don't gain anything by knowing B.

Some Useful Rules

- **Product rule:** $P(A,B) = P(A|B)\,P(B)$
  - Conditioned on C: $P(A,B|C) = P(A|B,C)\,P(B|C)$

- **Marginalizing (summing over a partition):**
  - $P(A) = \Sigma_{bi}\, P(A,B=b_i)$

    $= \Sigma_{bi}\, P(A \mid B=b_i)\, P(B=b_i)$

- **Chain Rule**
  - $P(A,B,C) = P(A|B,C) \times P(B|C) \times P(C)$

Bayes' Rule

$$P(H \mid E) = \frac{P(E \mid H)P(H)}{P(E)}$$

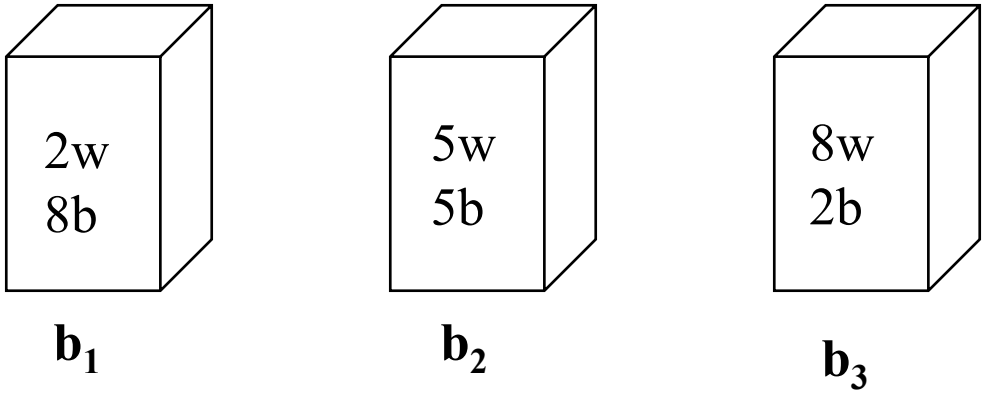
Bayes' rule allows us to express the quantity P(H|E), which people often find hard to assess, in terms of quantities that can be drawn from experience. Can think from cause (H) to effect (E).

Want: P(COVID | temp>38)
Assess: P(temp>38 | COVID)

Want: P(Battery = low | Beam = weak)
Assess: P(Beam = weak | Battery = low)

Bayes' Rule (cont'd)

- **Example:** Choose a ball at random from one of three boxes.



What is $P(B = b_1|C = w)$?

$P(C=w|B=b_1) = .2$ $P(C=w|B=b_2) = .5$ $P(C=w|B=b_3) = .8$

$P(B=b_i) = .33$

Bayes' Rule (cont'd)

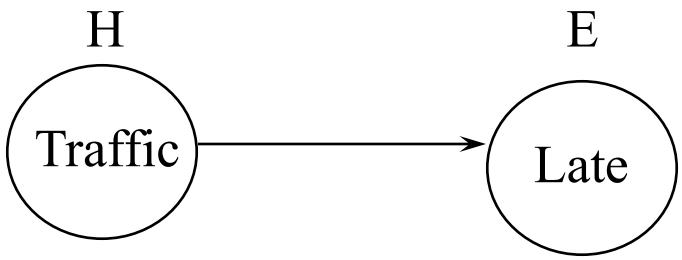
$$P(B = b_1 \mid C = w) = \frac{P(C = w \mid B = b_1)P(B = b_1)}{P(C = w)}$$

$$\begin{aligned} P(C = w) = {} & P(C = w \mid B = b_1)P(B = b_1) + \\ & P(C = w \mid B = b_2)P(B = b_2) + \\ & P(C = w \mid B = b_3)P(B = b_3) \\ = {} & (.2)(1/3) + (.5)(1/3) + (.8)(1/3) \\ = {} & (.15)(1/3) = .5 \end{aligned}$$

$$P(B = b_1 \mid C = w) = \frac{(.2)(1/3)}{(.5)} = .13$$

Confirmation is Symmetric

- Suppose P(E|H) > P(E)

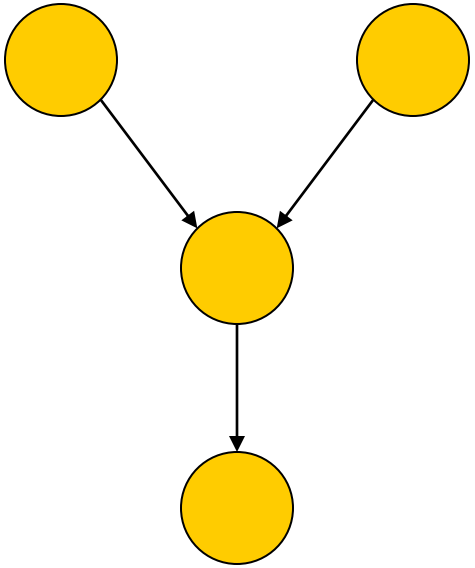


$$P(E \mid H) > P(E)$$

$$\frac{P(E,H)}{P(H)} > P(E)$$

$$P(E,H) > P(E)P(H)$$

$$P(H \mid E) > P(H)$$



Bayesian Networks
Part II

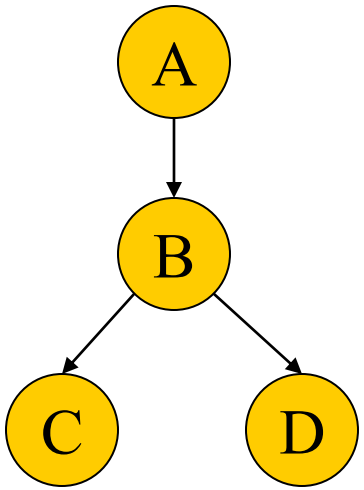
Prof. Dr. Peter Haddawy
Faculty of ICT
Mahidol University

Some material adopted from slides by Andrew Moore, Weng-Kee Wong, and Darlene Goldstein.

A Bayesian Network

A Bayesian network is made up of:

1. A Directed Acyclic Graph



2. A set of tables for each node in the graph

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

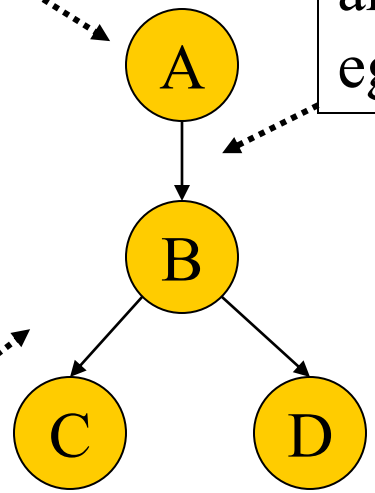
B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

A Directed Acyclic Graph

Each node in the graph is a random variable

A node *X* is a parent of another node *Y* if there is an arrow from node *X* to node *Y* eg. *A* is a parent of *B*



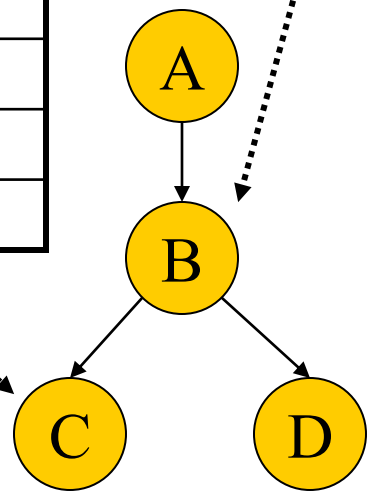
Informally, an arrow from node *X* to node *Y* means *X* has a direct influence on *Y*

A Set of Tables for Each Node

A	P(A)
false	0.6
true	0.4

A	B	P(B|A)
false	false	0.01
false	true	0.99
true	false	0.7
true	true	0.3

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1



B	D	P(D|B)
false	false	0.02
false	true	0.98
true	false	0.05
true	true	0.95

Each node X_i has a conditional probability distribution $P(X_i \mid \text{Parents}(X_i))$ that quantifies the effect of the parents on the node

The parameters are the probabilities in these conditional probability tables (CPTs)

A Set of Tables for Each Node

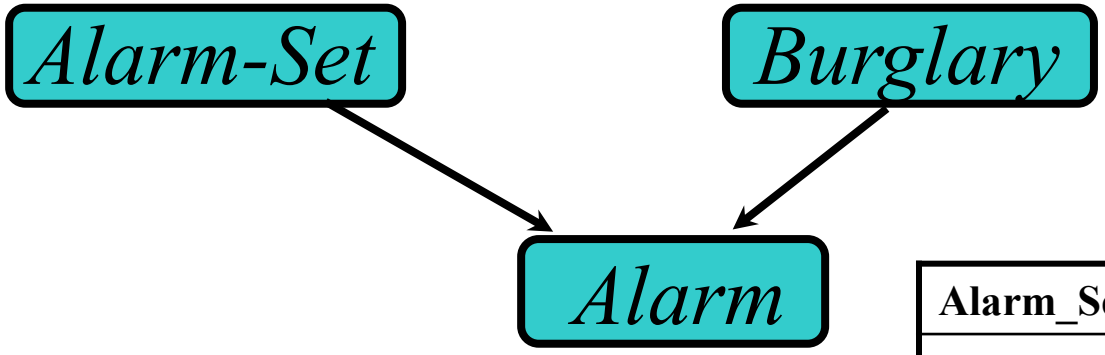
Conditional Probability
Distribution for C given B

B	C	P(C|B)
false	false	0.4
false	true	0.6
true	false	0.9
true	true	0.1

For a given combination of values of the parents (B in this example), the entries for P(C=true | B) and P(C=false | B) must add up to 1
eg. P(C=true | B=false) + P(C=false |B=false)=1

If you have a Boolean variable with k Boolean parents, this table has 2^{k+1} probabilities (but only 2^k need to be stored)

Interacting Causes

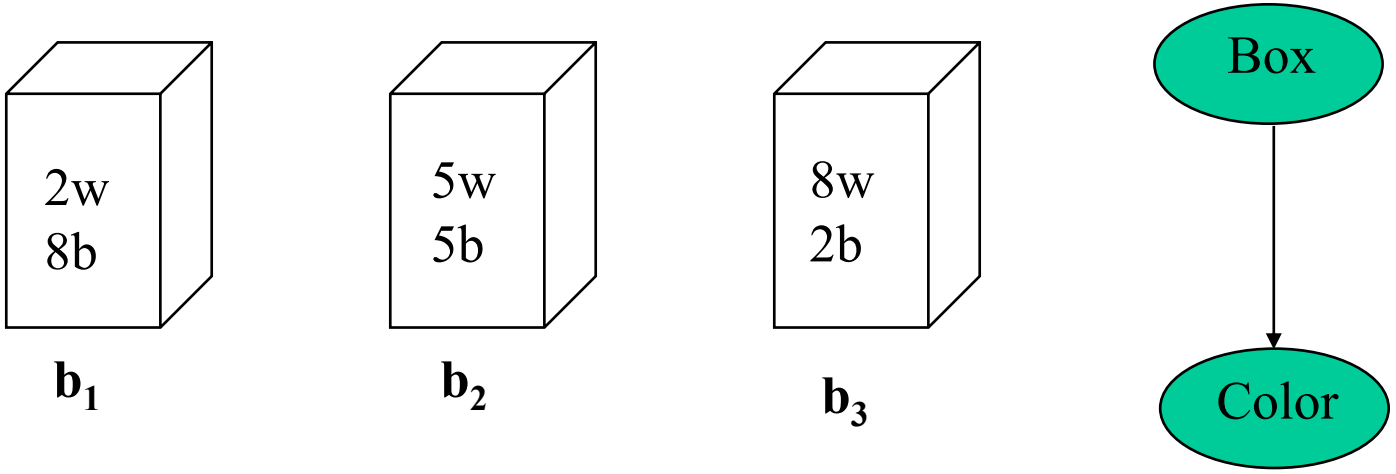


Alarm_Set	Burglary	P(Alarm|AS,B)
false	false	0
false	true	0
true	false	0.01
true	true	0.95

- If a node has multiple parents, those causes may interact in various ways. So we need the probability of the child given all combinations of the parents
- *E.g., Alarm* can go off only if it is *Set*

Bayes' Rule Revisited

Choose a ball at random from one of three boxes.



What is $P(B = b_1|C = w)$?

$P(C=w|B=b_1) = .2$ $P(C=w|B=b_2) = .5$ $P(C=w|B=b_3) = .8$

$P(B=b_i) = .33$

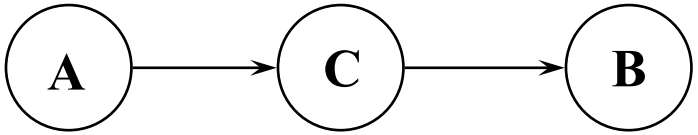
Bayesian Networks

Two important properties:

1. Encodes the conditional independence relationships between the variables in the graph structure
2. Is a compact representation of the joint probability distribution over the variables

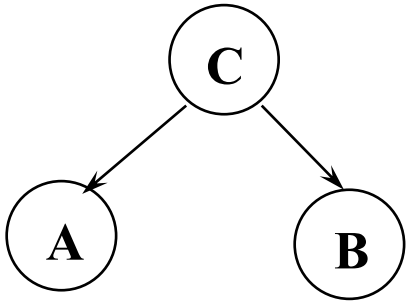
Three Types of Connections

- Serial



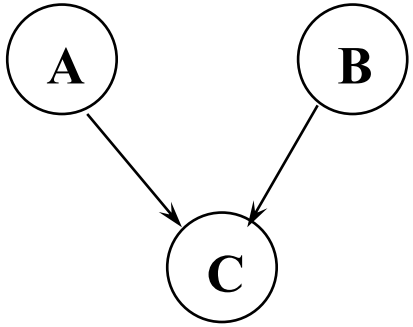
A and B are independent given C

- Diverging



A and B are independent given C

- Converging



A and B are independent but may become dependent when given C

Types of Connections

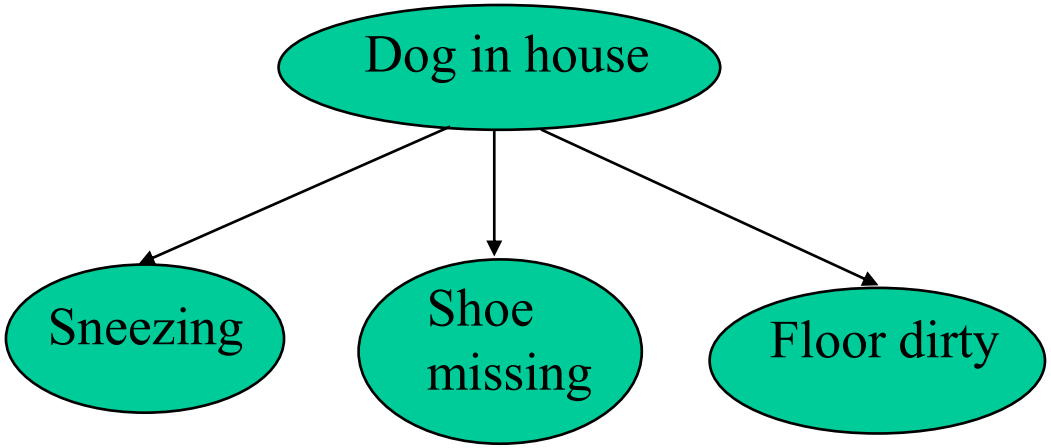
- **Serial connections**

  Evidence may be transmitted through a serial connection unless the state of the variable in the connection is known.

Types of Connections

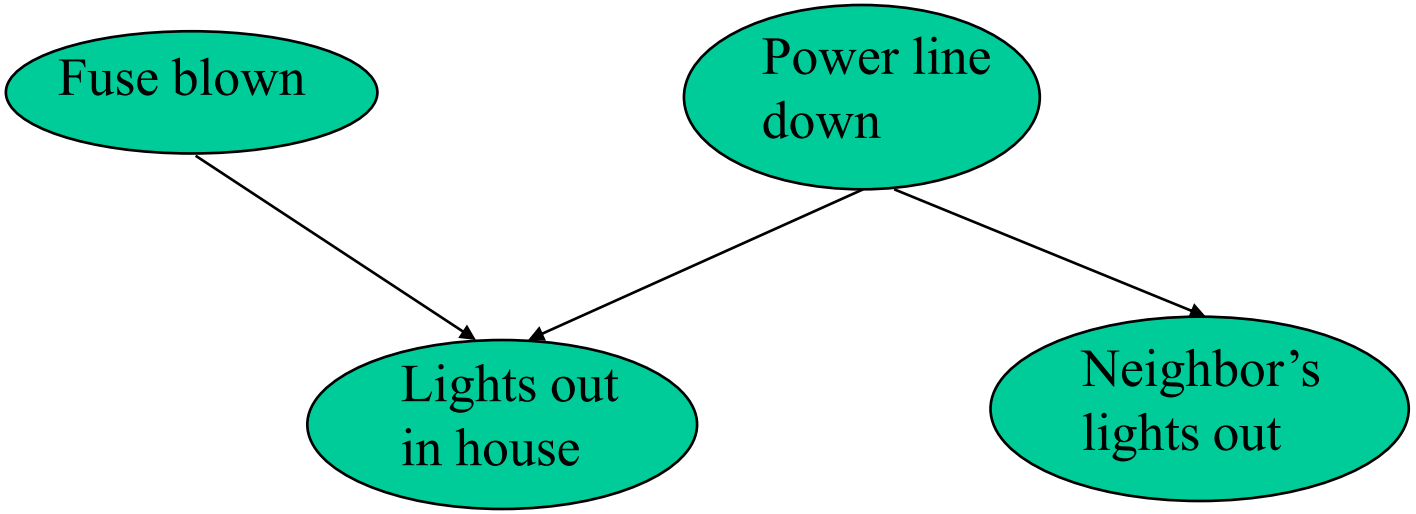
- **Diverging connections**

  Evidence can be transmitted through a diverging connection unless it is instantiated.

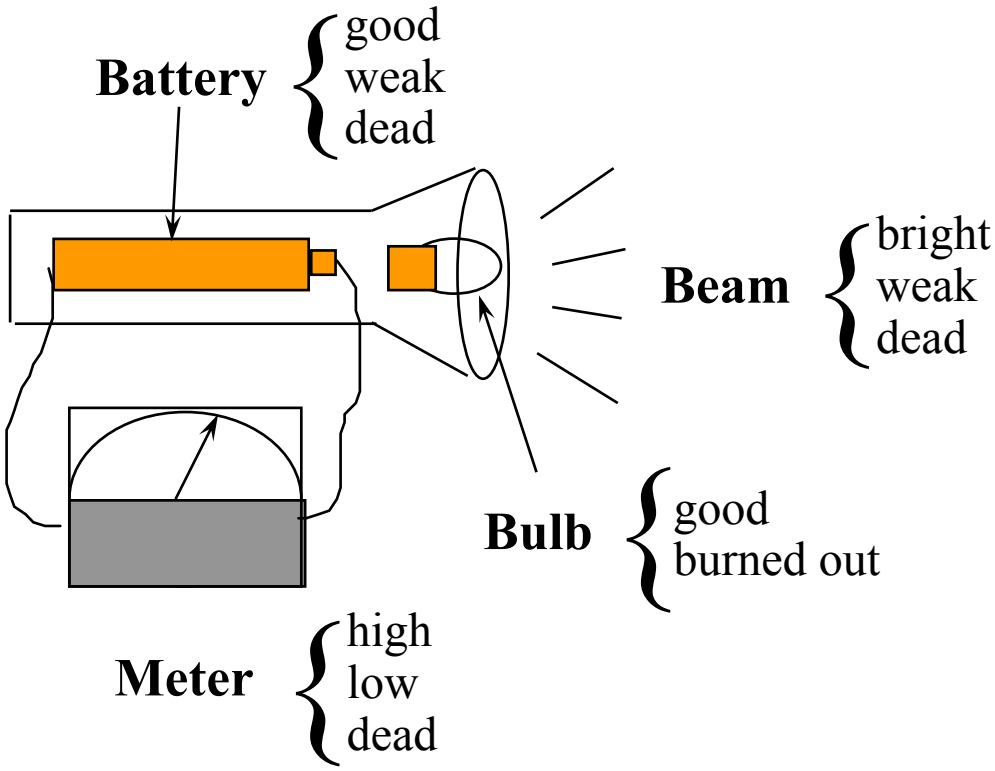


Types of Connections (cont'd)

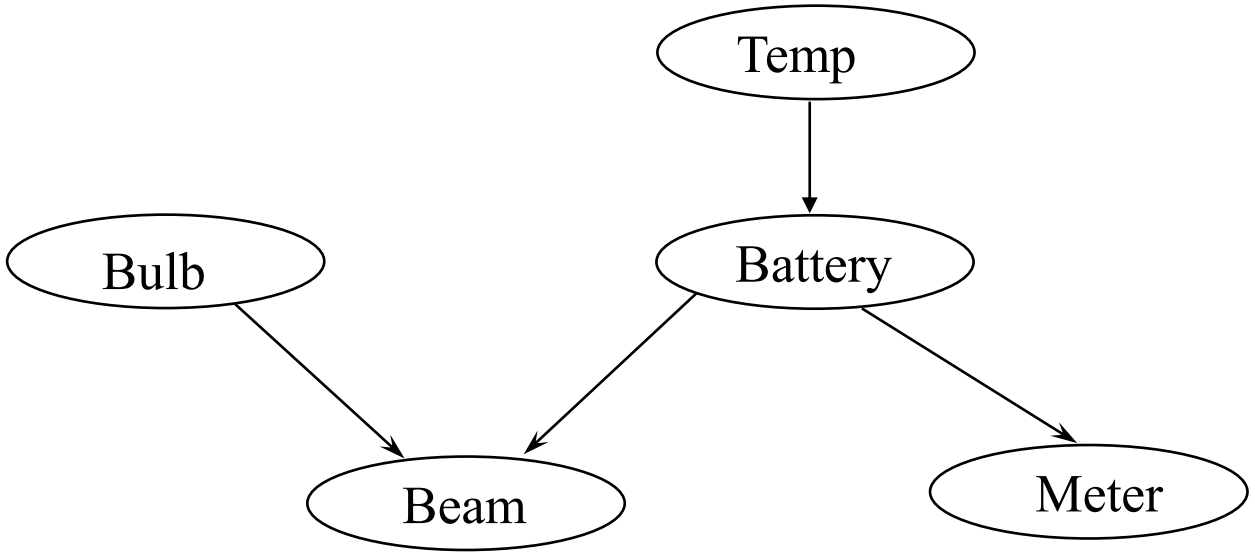
- **Converging connections**



Flashlight Example

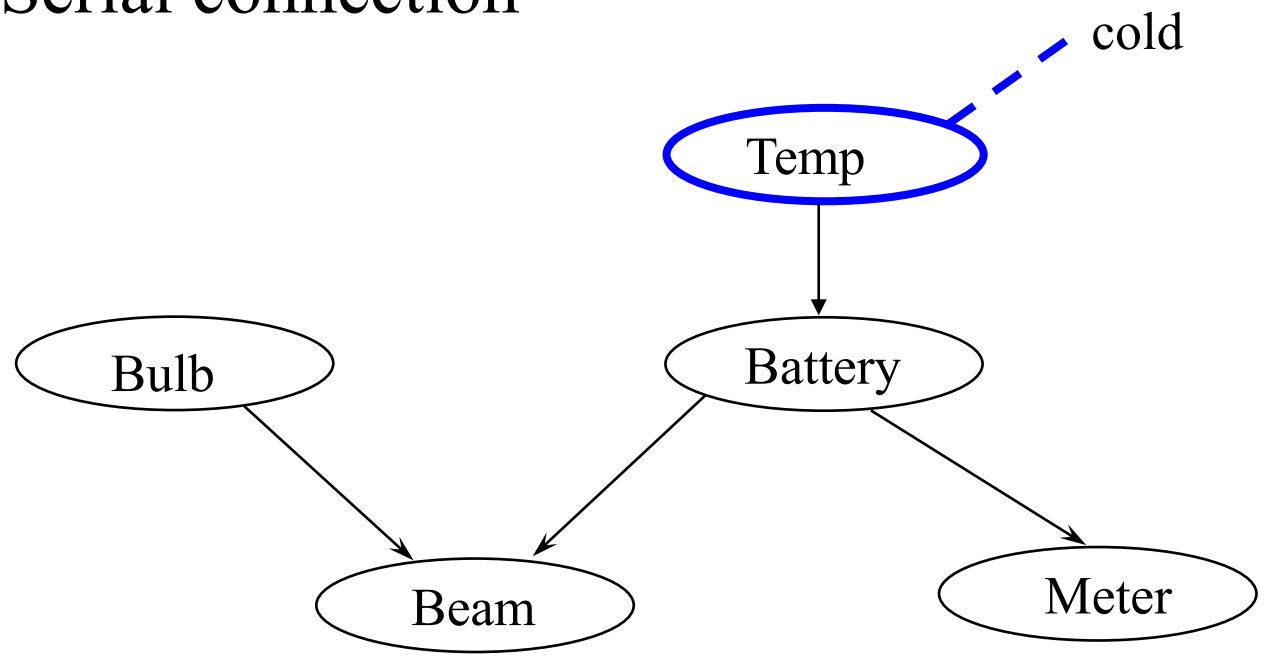


Flashlight Example



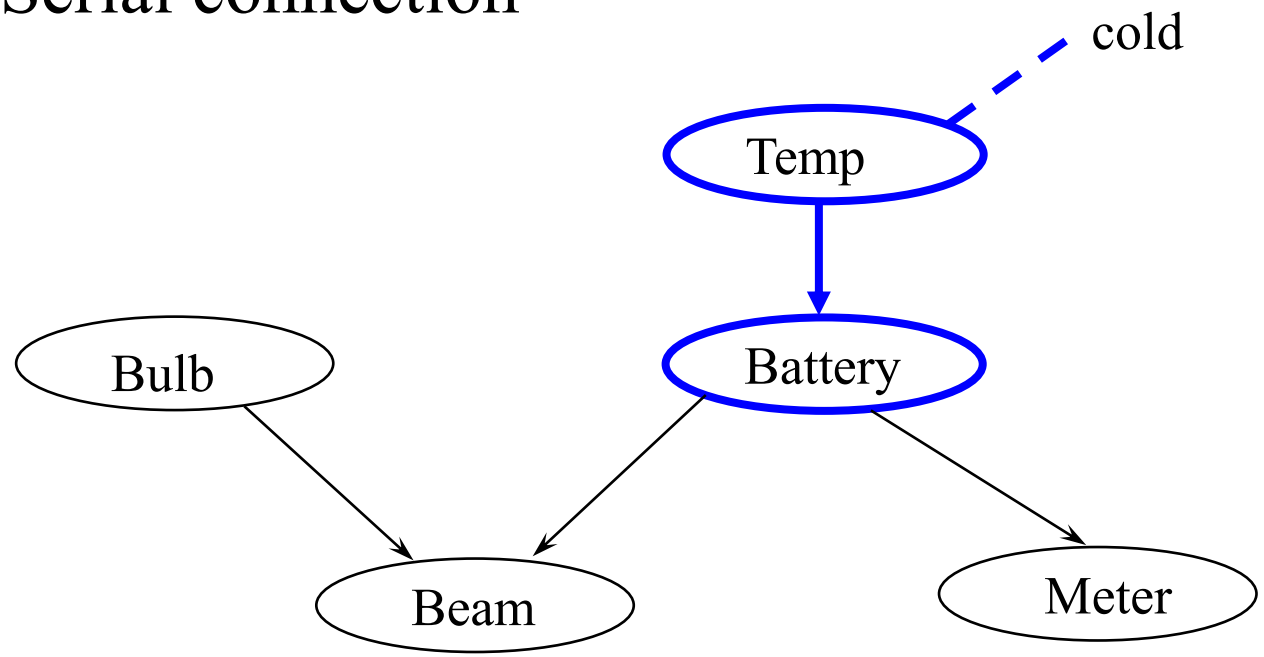
Flashlight Example

- Serial connection



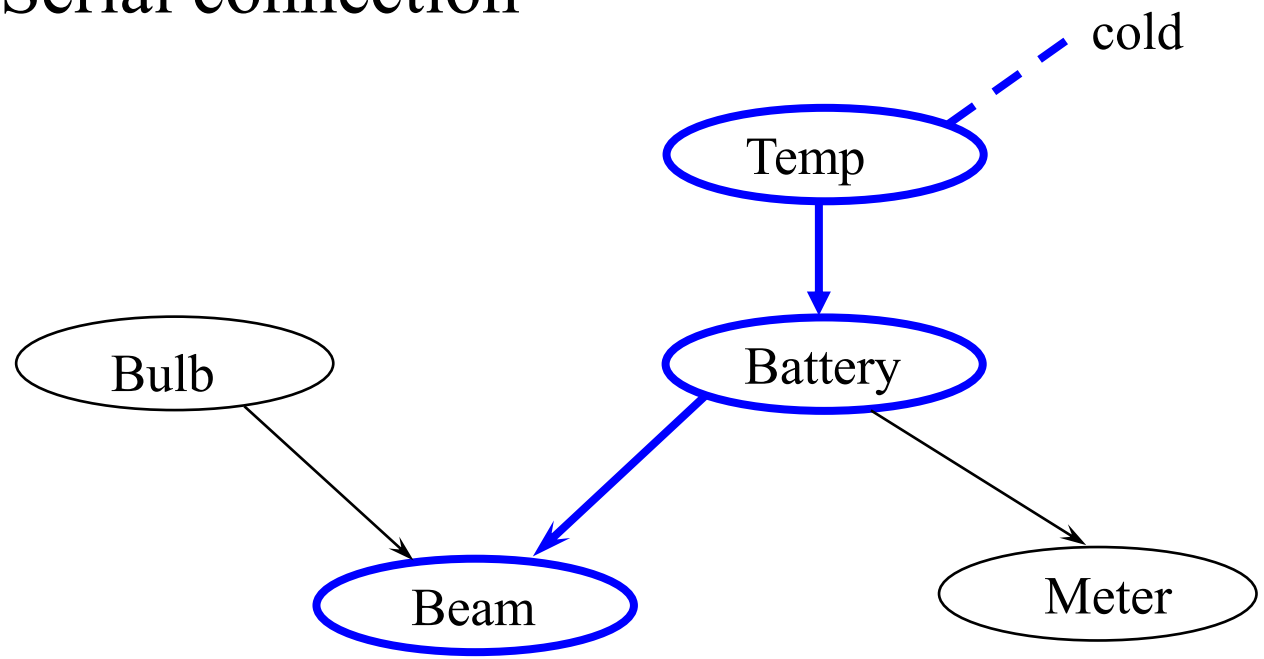
Flashlight Example

- Serial connection



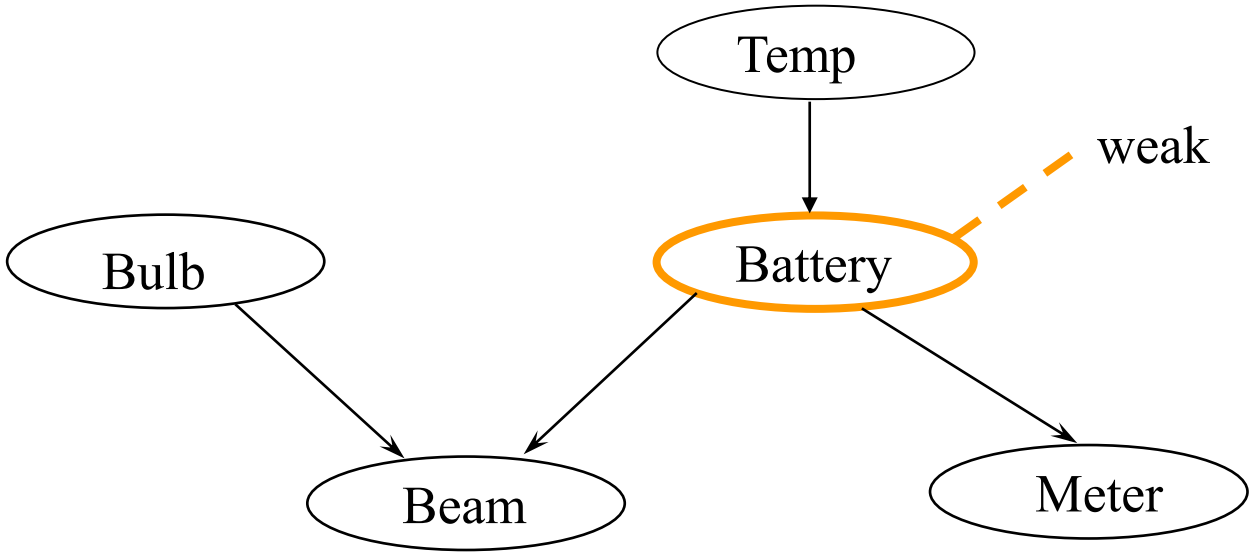
Flashlight Example

- Serial connection



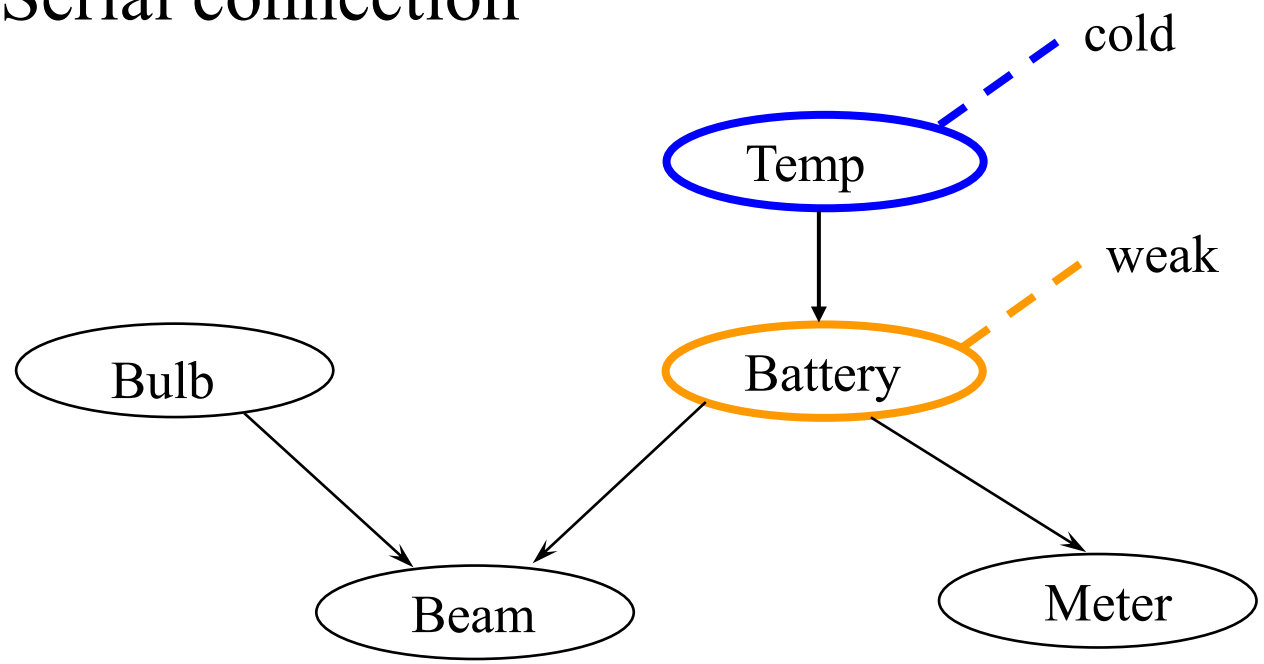
Flashlight Example

- Serial connection



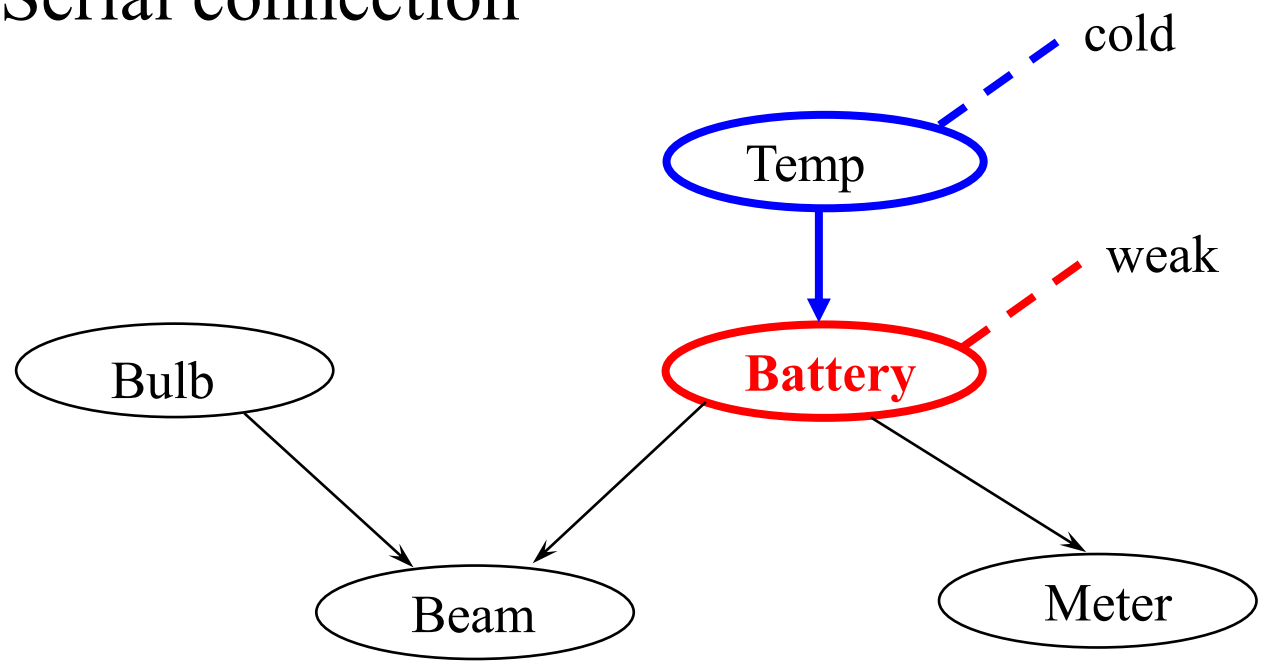
Flashlight Example

- Serial connection



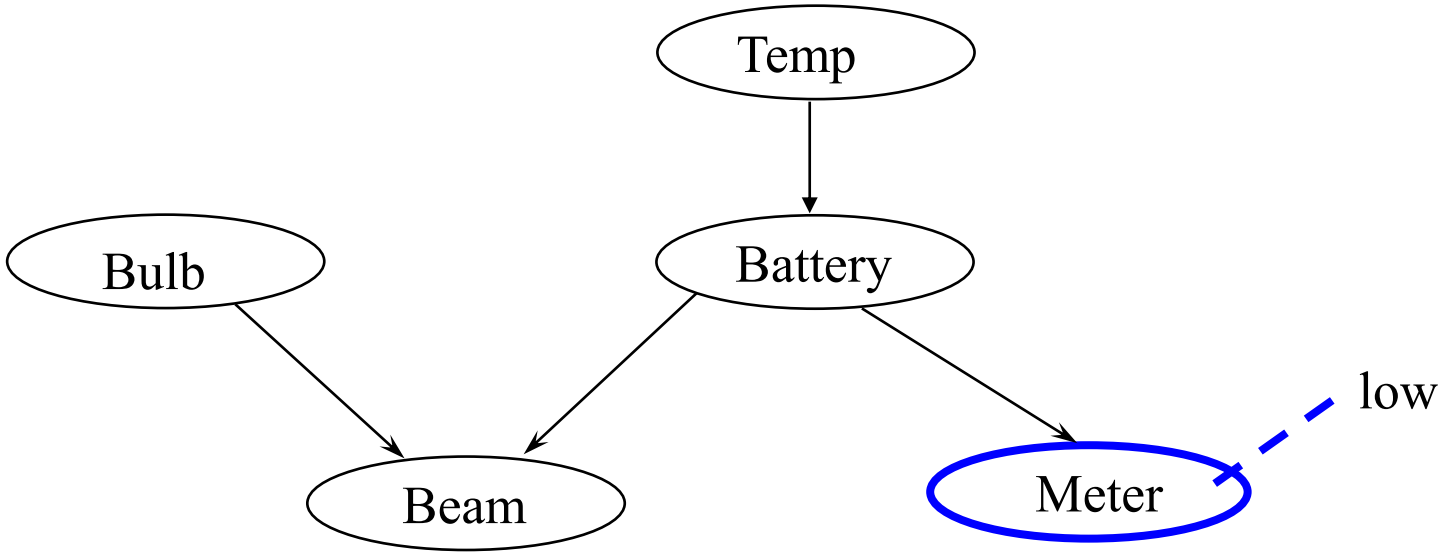
Flashlight Example

- Serial connection



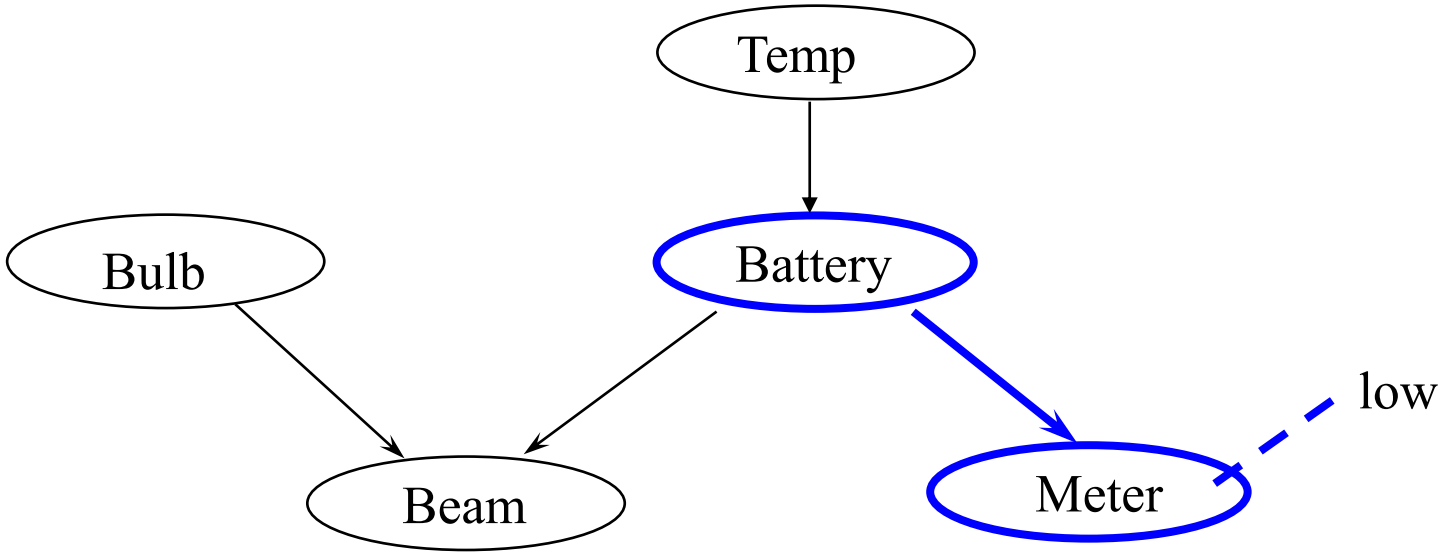
Flashlight Example

- Diverging connection



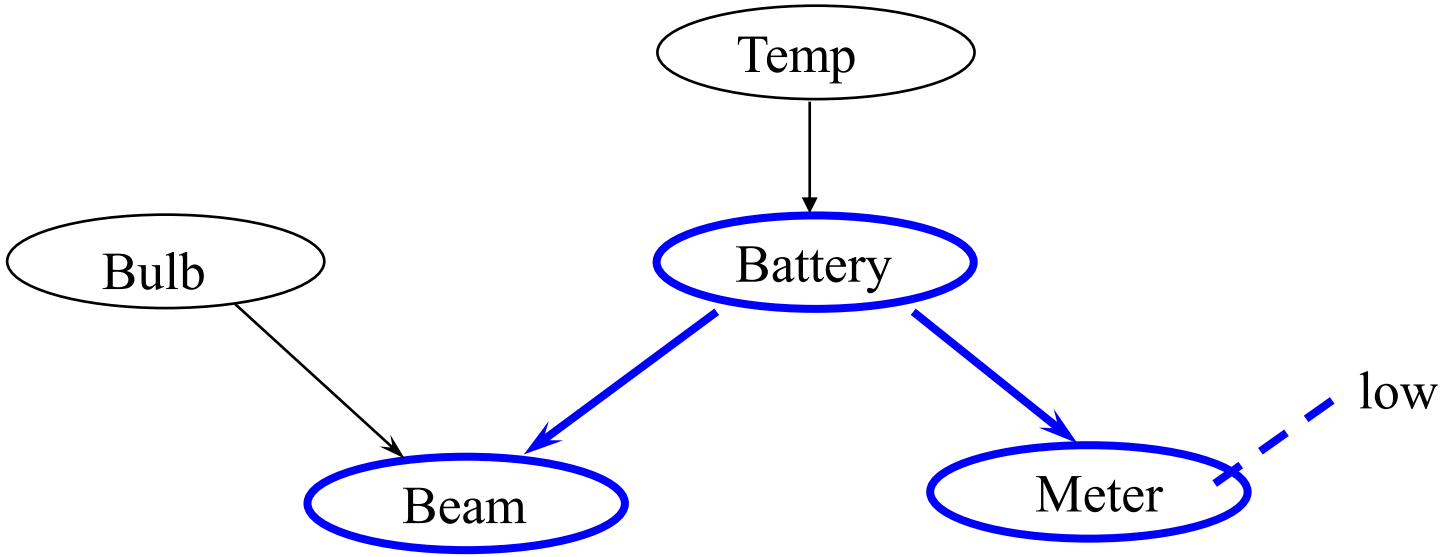
Flashlight Example

- Diverging connection



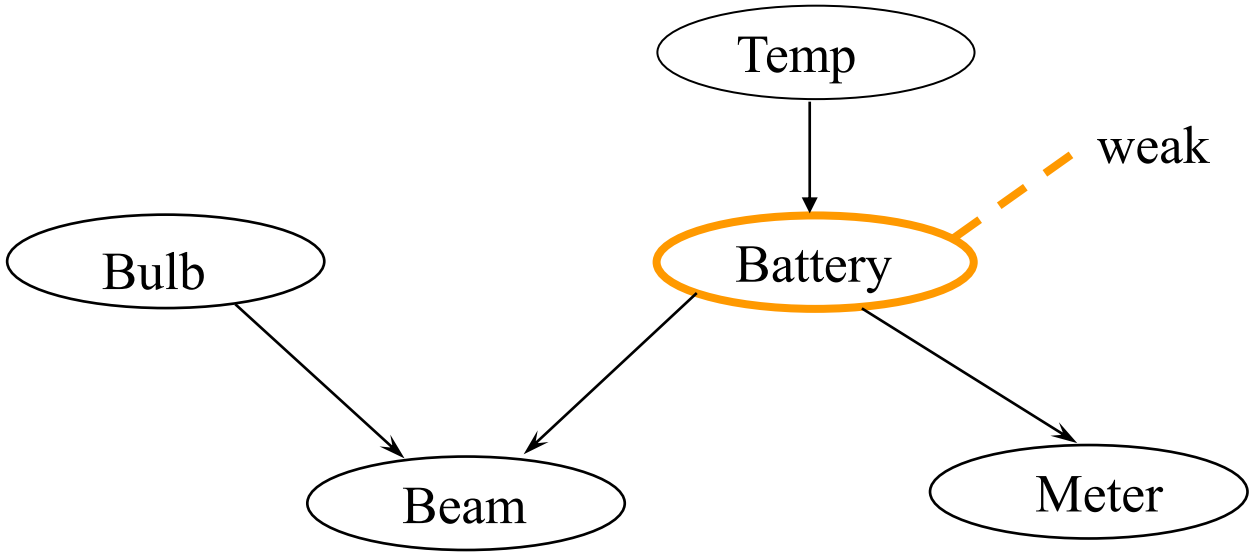
Flashlight Example

- Diverging connection



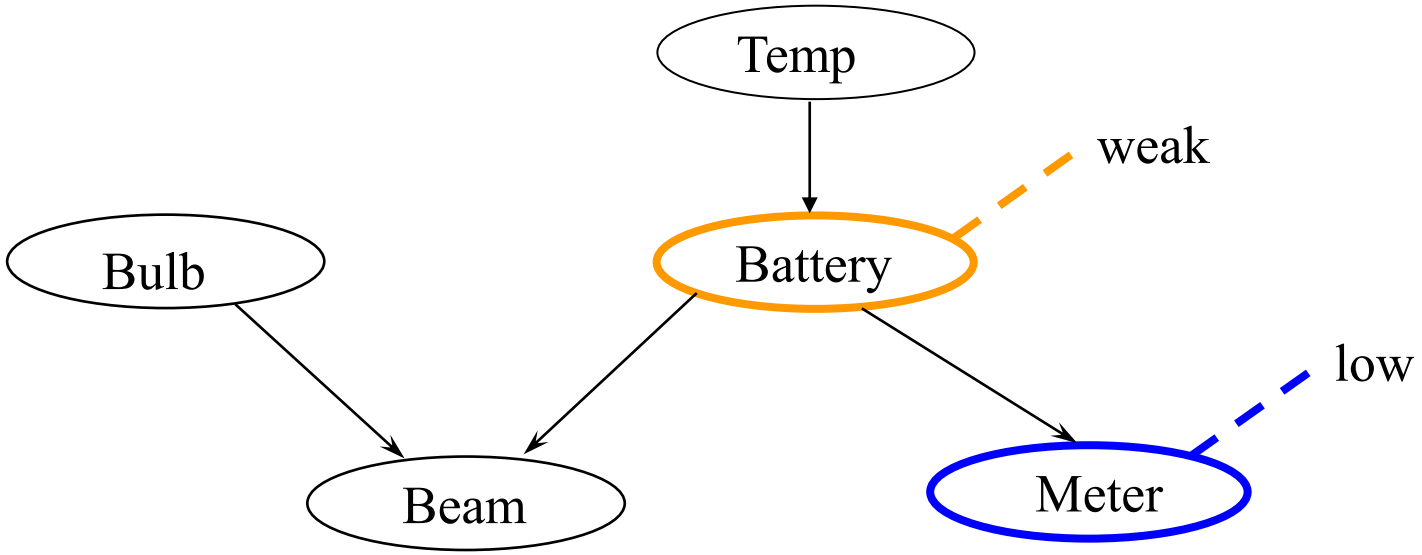
Flashlight Example

- Diverging connection



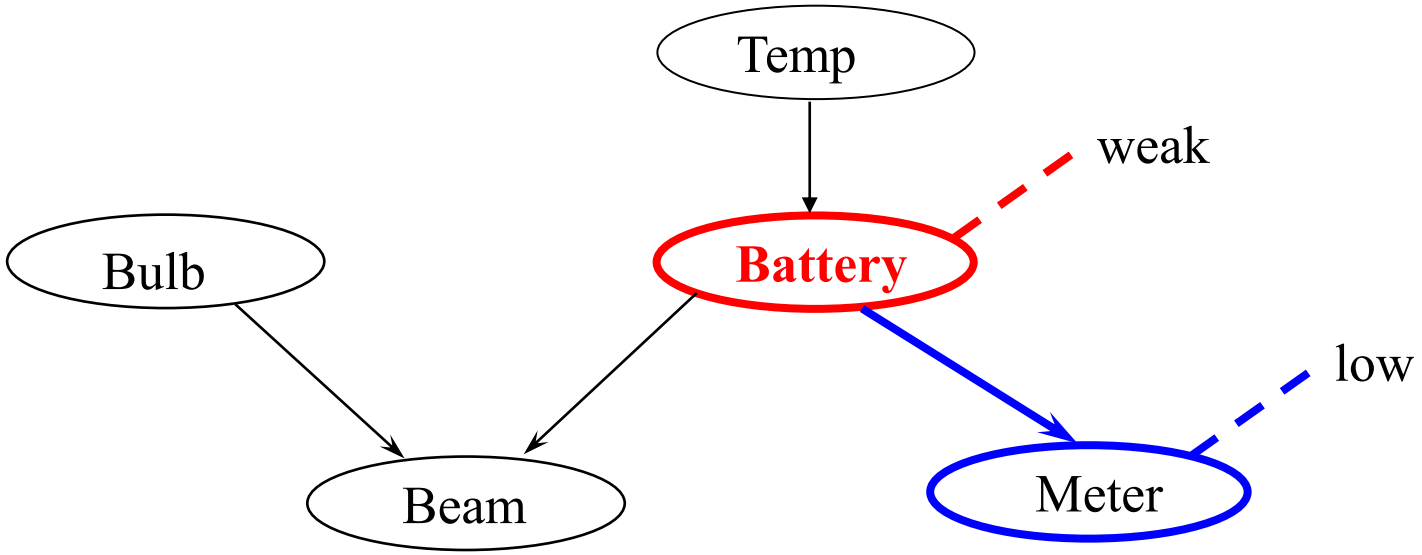
Flashlight Example

- Diverging connection



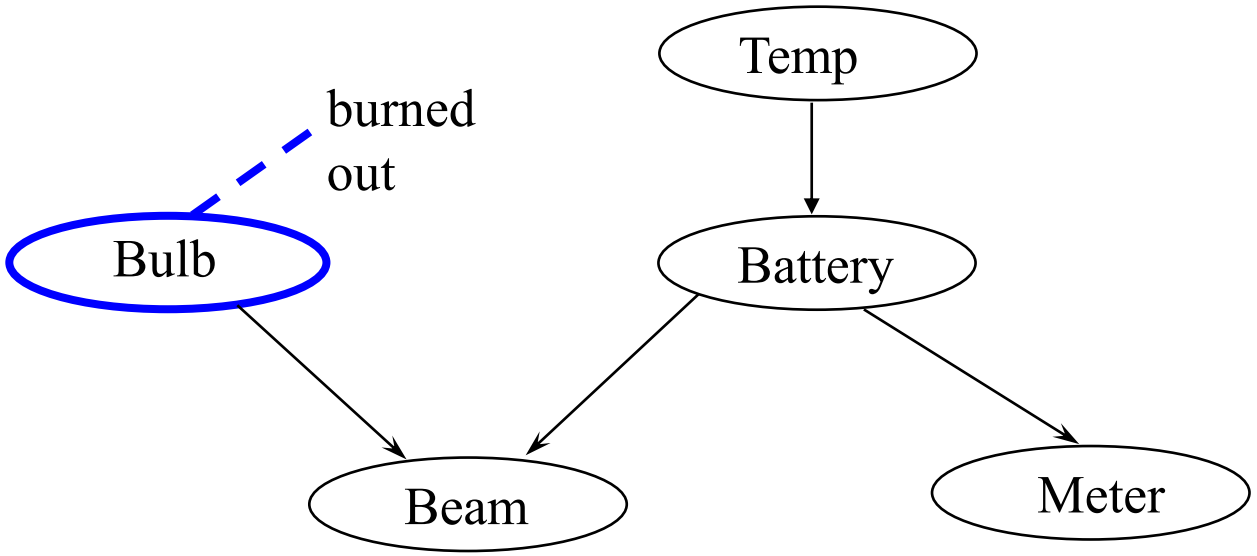
Flashlight Example

- Diverging connection



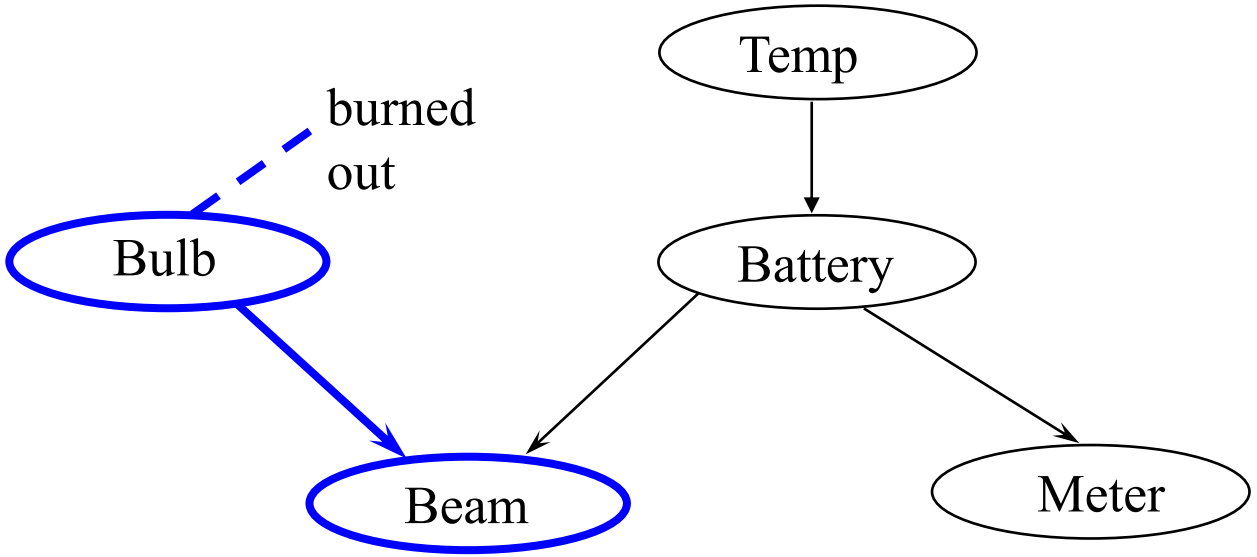
Flashlight Example

- Converging connection



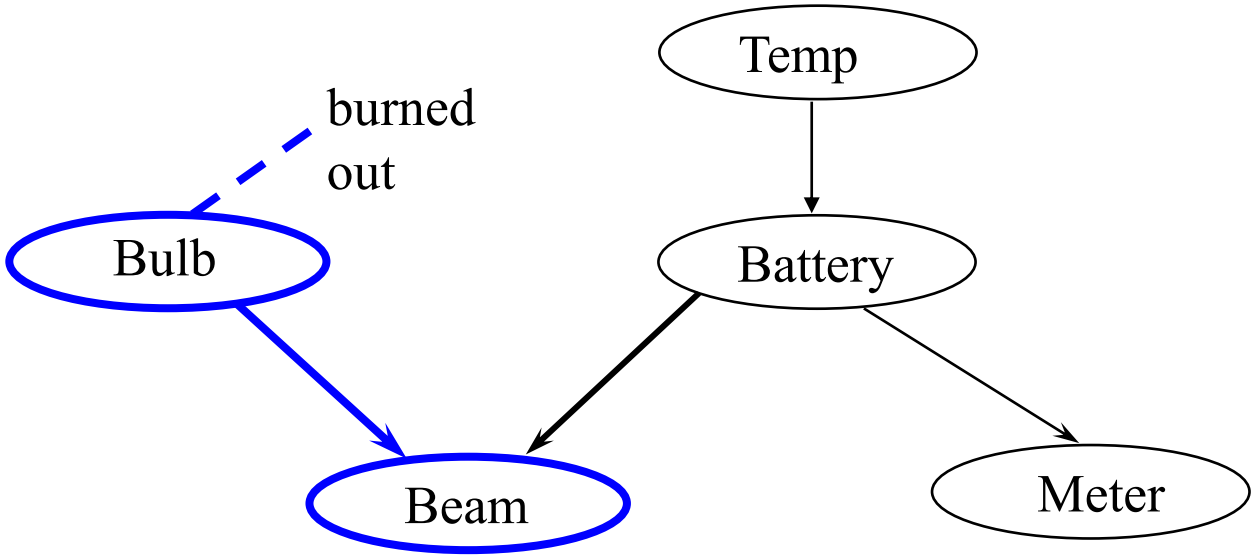
Flashlight Example

- Converging connection



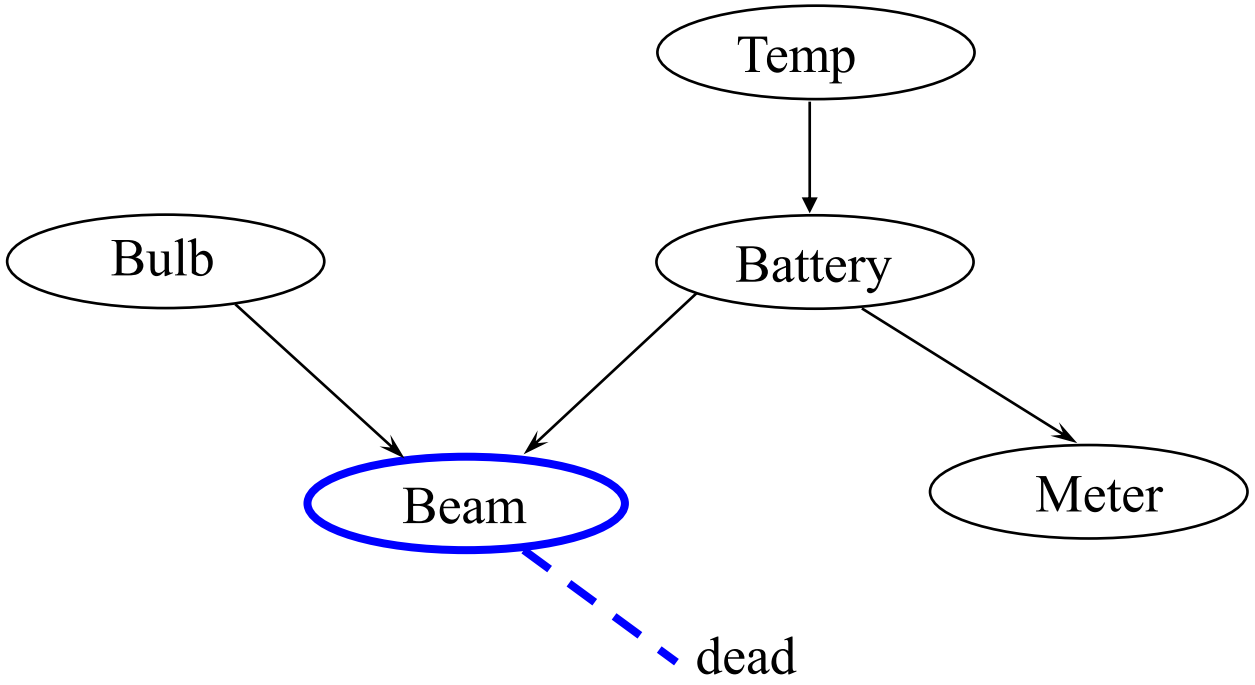
Flashlight Example

- Converging connection



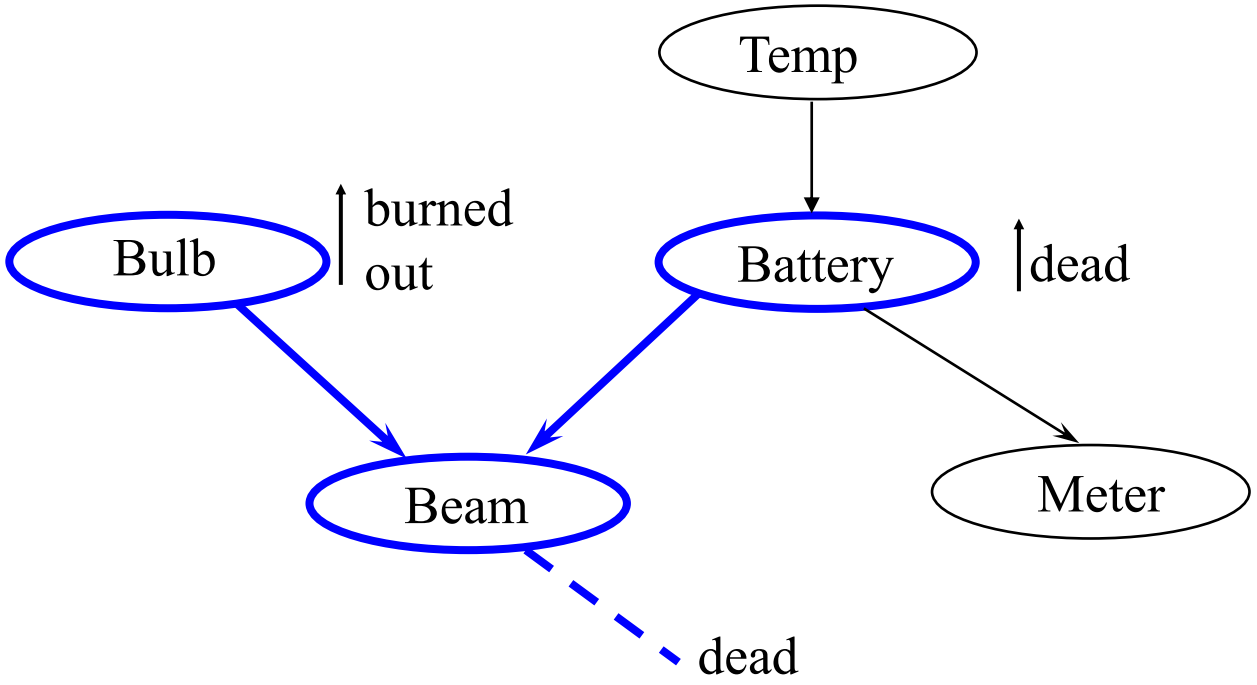
Flashlight Example

- Converging connection



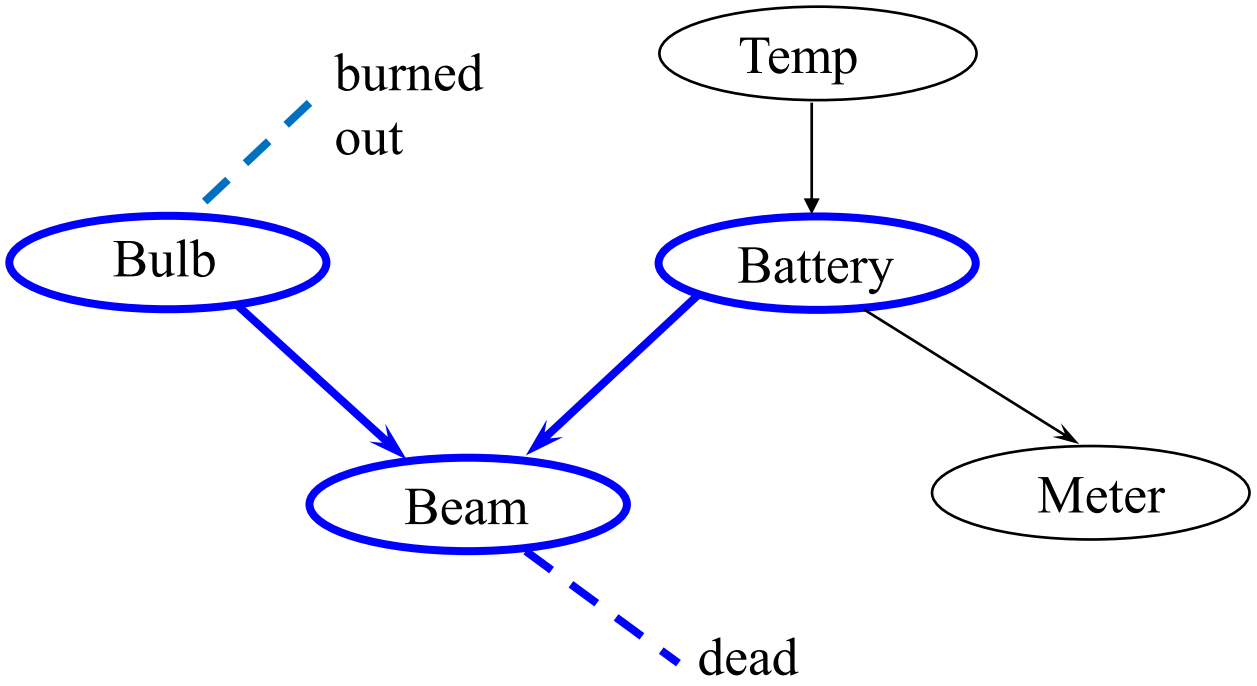
Flashlight Example

- Converging connection



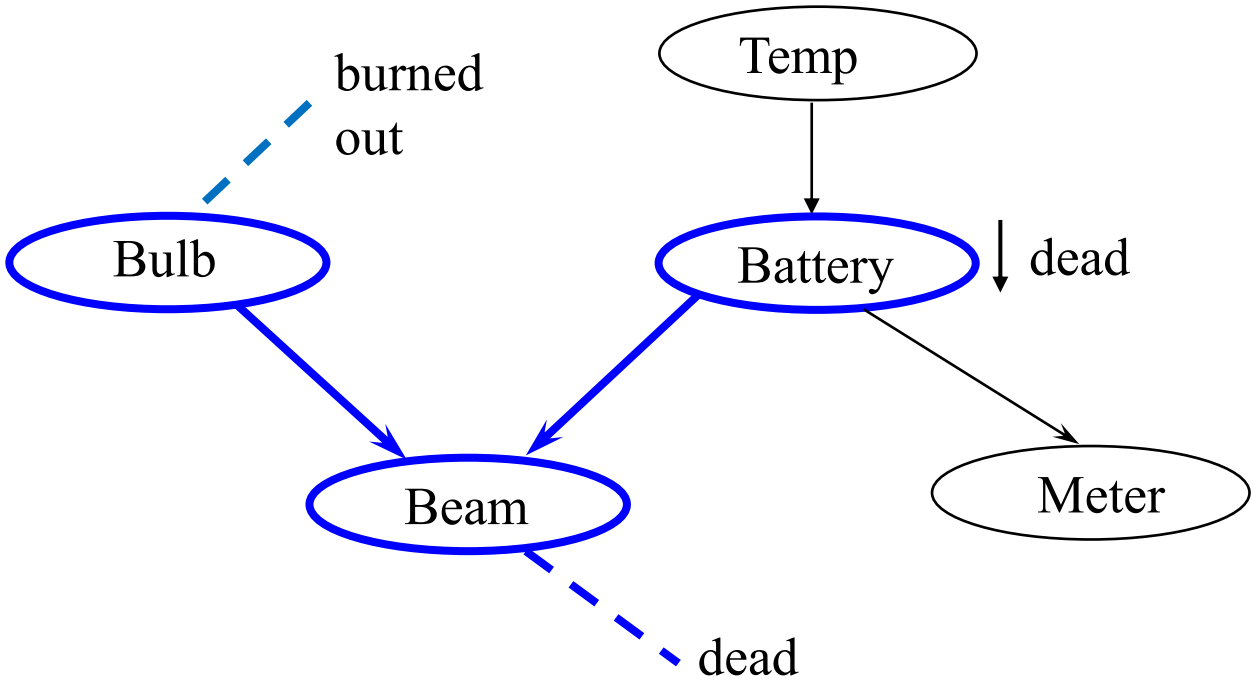
Flashlight Example

- Converging connection



Flashlight Example

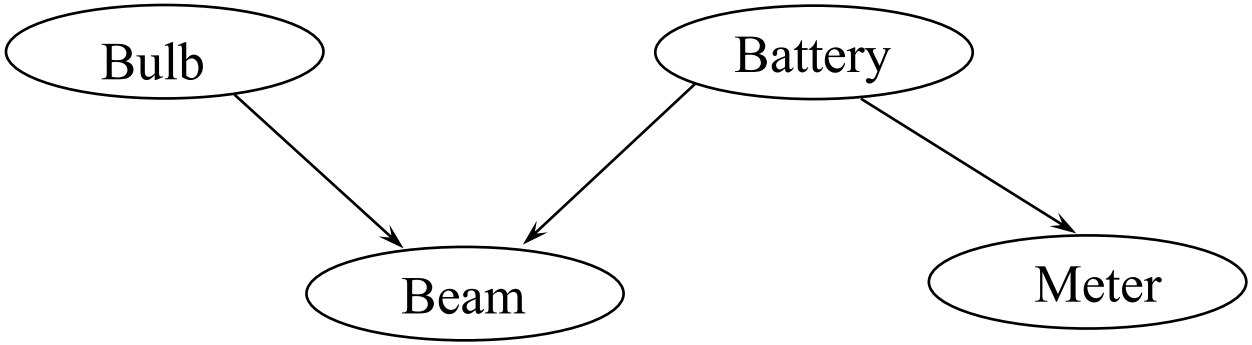
- Converging connection



Building Models

- The purpose of a Bayesian model is often to give estimates on the likelihoods of events which are not observable or only observable at an unacceptably high price. These events are called **hypotheses.**
- Steps in building the structure of a Bayesian network:
  1. Represent the hypotheses as random variables.
  2. Identify observable information which can tell us something about the states of the hypotheses. We call these **observable** or **information variables.**
  3. Determine the causal structure between the variables, i.e., which events have a direct causal impact on which other events.
  4. Quantify the links with CPTs.

Building Models: Flashlight Example



Variable States

Each variable can assume one of a list of mutually exclusive and exhaustive states.

Variable	Possible Values
Bulb	Good
	Burnt out
Battery	Good
	Weak
	Dead
Beam	Bright
	Weak
	Dead
Meter	High
	Low
	Dead

Probabilities

Each variable must be assigned a conditional probability table. Root nodes are assigned unconditional probabilities.

P(Bulb)

Value	Probability
Good	0.9
Burnt out	0.1

P(Battery)

Value	Probability
Good	0.8
Weak	0.1
Dead	0.1

Probabilities (cont'd)

Nodes with parents are assigned probabilities conditioned on all combinations of states of the parents.

P(Meter | Battery)

Conditioners	Value	Probability
Battery = Good	High	0.97
	Low	0.015
	Dead	0.015
Battery = Weak	High	0.02
	Low	0.97
	Dead	0.01
Battery = Dead	High	0.00
	Low	0.00
	Dead	1.00

Probabilities (cont'd)

P(Beam | Bulb,Battery)

Conditioners	Value	Probability
Bulb = Good	Bright	0.99
Battery = Good	Weak	0.005
	Dead	0.005
Bulb = Good	Bright	0.00
Battery = Weak	Weak	0.99
	Dead	0.01
Bulb = Good	Bright	0.00
Battery = Dead	Weak	0.00
	Dead	1.00
Bulb = Burnt out	Bright	0.00
Battery = Good	Weak	0.00
	Dead	1.00 ...

Joint Probability Distribution

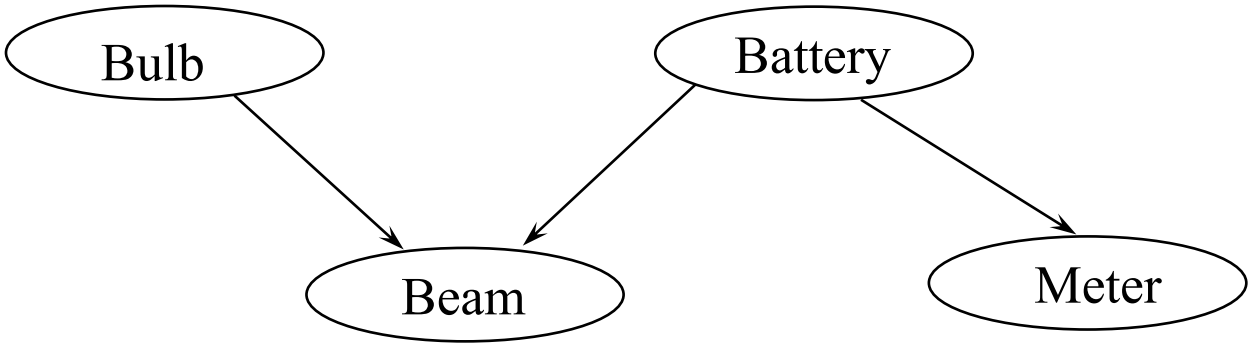
- **Independence lemma**

  A variable is independent of all variables that are not its successors given its direct predecessors.

- **Theorem:** Given a Bayesian network G, with variables $V_1, \ldots, V_n$ there is exactly one probability distribution consistent with the conditional probability tables. That distribution is given by

$$P(V_1,\ldots,V_n) = \prod_i P(V_i \mid pa(V_i))$$

Joint Probability Distribution



P(Bulb, Beam, Battery, Meter) =

P(Bulb) P(Battery) P(Beam | Battery,Bulb)

P(Meter | Battery).

Joint Probability Distribution (cont'd)

- We have 3 random variables (Battery, Beam, Meter) with three states each, and one (Bulb) with two states.
- **Table:** 3 x 3 x 3 x 2 = 54 probabilities
- **Bayesian network:**
  - Bulb = 2
  - Battery = 3
  - Beam = 2 x 3 x 3 = 18
  - Meter = 3 x 3 = 9
  - **Total =** 32 probabilities

Inferences

- Computing posterior probabilities
  - Diagnostic inference
    - What is the most likely cause of beam weak?
  - Predictive inference
    - What would happen to the beam if the battery was weak?
- Determining the most informative test
- Determining the configuration of highest probability
- Determining optimal courses of action
- Explanation

Bayes Nets and Logic

Probability = Generalization of Propositional Logic

Represent the expression

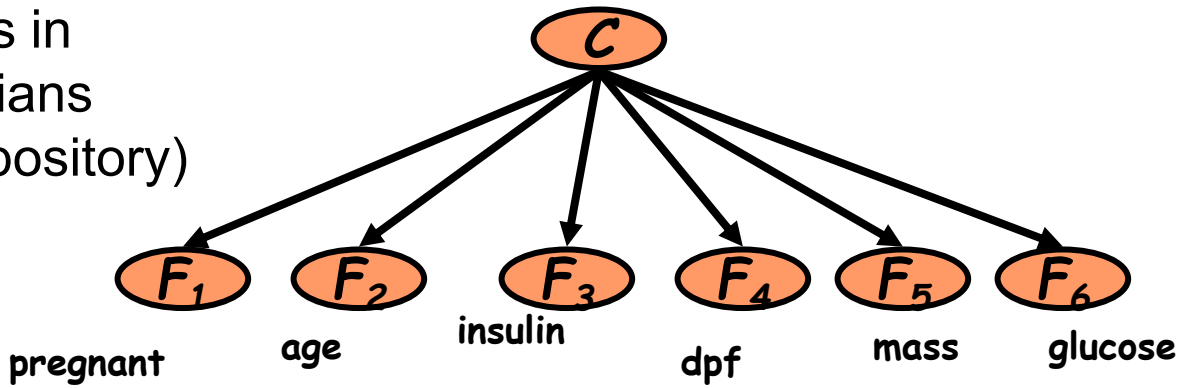
$$X = (A \wedge B) \vee C$$

as a Bayes Net

Some Useful Special Cases of Bayes Nets

The Naïve Bayes Classifier

Diabetes in
Pima Indians
(from UCI repository)

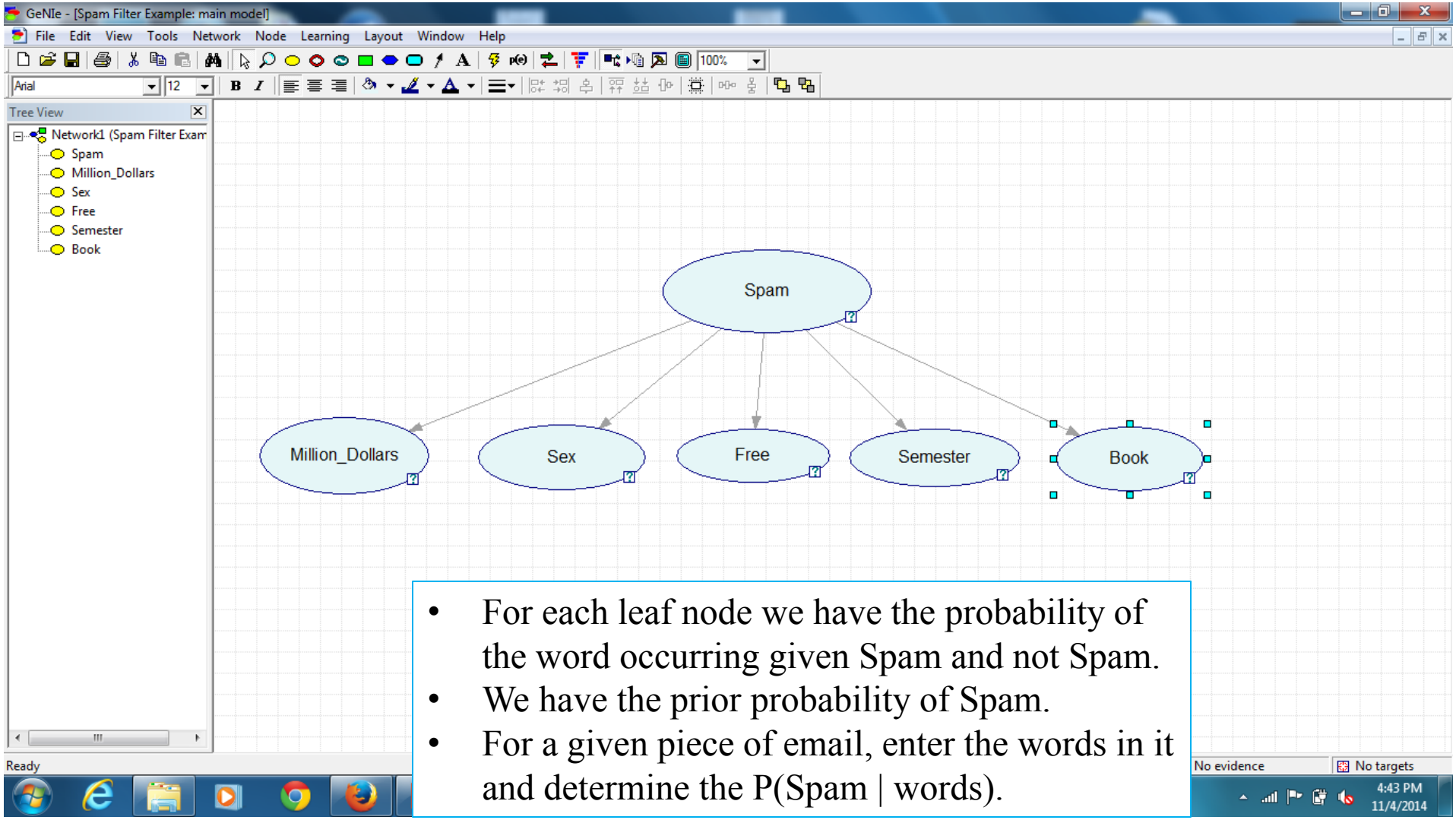


- Fixed structure encoding the assumption that features are independent of each other given the class.

$$P(C \mid F_1, \ldots, F_6) \propto P(F_1 \mid C) \bullet P(F_2 \mid C) \bullet \cdots \bullet P(F_6 \mid C) \bullet P(C)$$

- Learning amounts to estimating the parameters for each $P(F_i|C)$ for each F_i.

Naïve Bayes Spam Filter

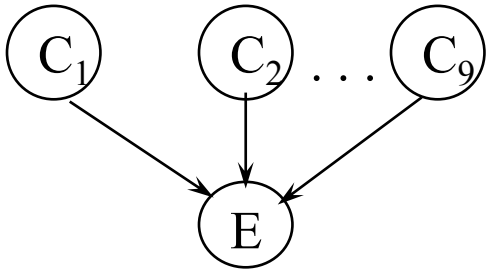


- For each leaf node we have the probability of the word occurring given Spam and not Spam.
- We have the prior probability of Spam.
- For a given piece of email, enter the words in it and determine the P(Spam | words).

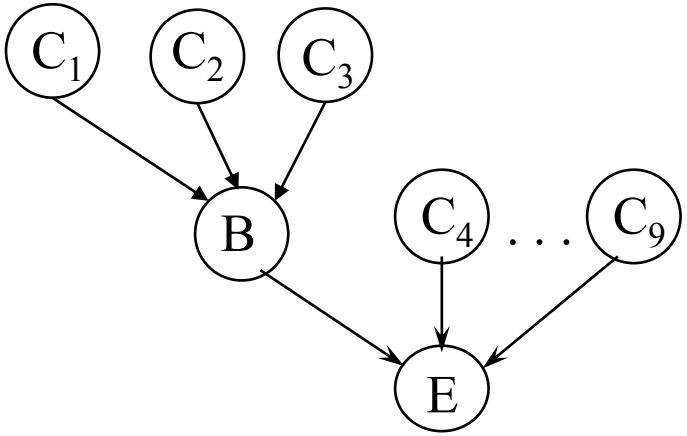
Modeling Techniques for Dealing with Large CPTs

Divorcing

- If a node has many parents, we may have difficulty obtaining the data for the conditional probability table. If the variables in this network are binary, the CPT contains $2^{10} = 1024$ entries.

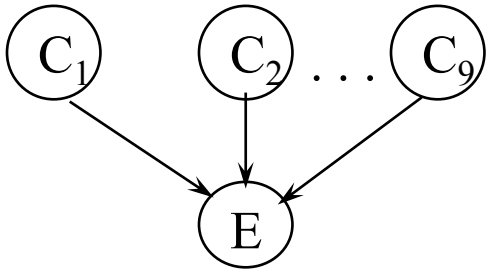


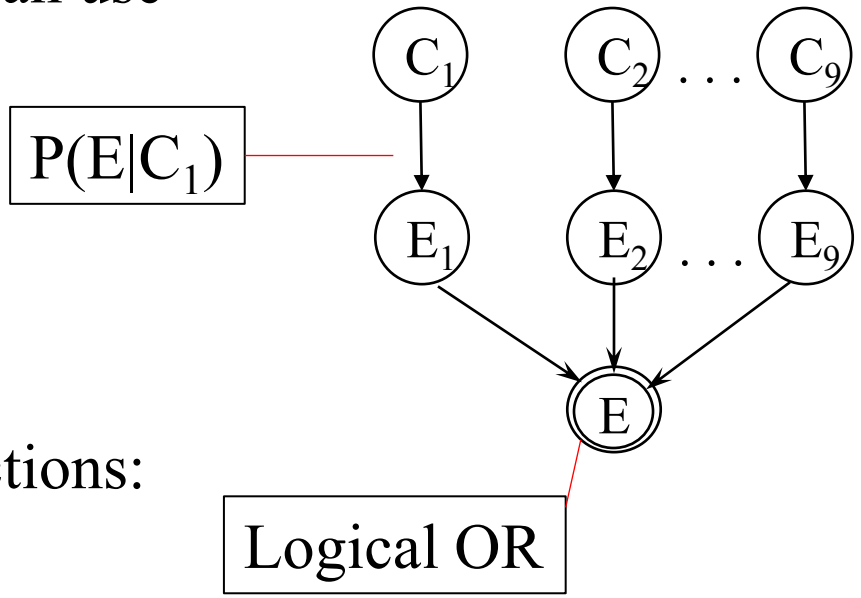
- We can reduce the table size by introducing an intermediate variable.
- The CPTs now have $2^4 + 2^8 = 272$ entries.



Noisy OR

- If the causes act independently and in a regular way, we may use a generic combination function.
- If the C_i act independently and each C_i is sufficient to cause E, we can use Noisy-OR.





- Other commonly used functions: AND, MAX, SUM

Evaluation

Evaluation Challenge

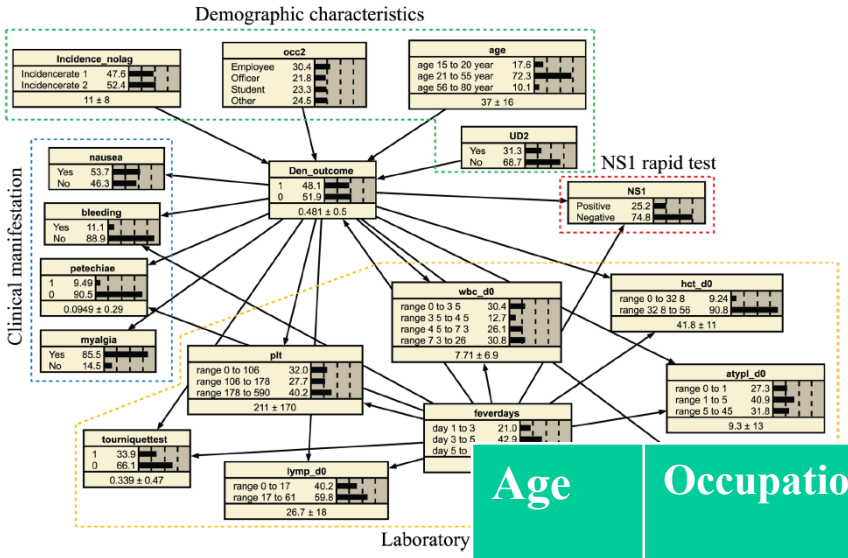


Fig 1. Final Bayesian network model for dengue diagnosis.

How accurate is this Bayes net?

Age	Occupation	Nausea	WBC	Dengue	BN diagnosis
15	Student	Y	3	1	.9
22	Employee	N	5	1	.7
58	Officer	Y	2	0	.3
40	Employee	Y	4	0	.8

Introduction to ROC curves

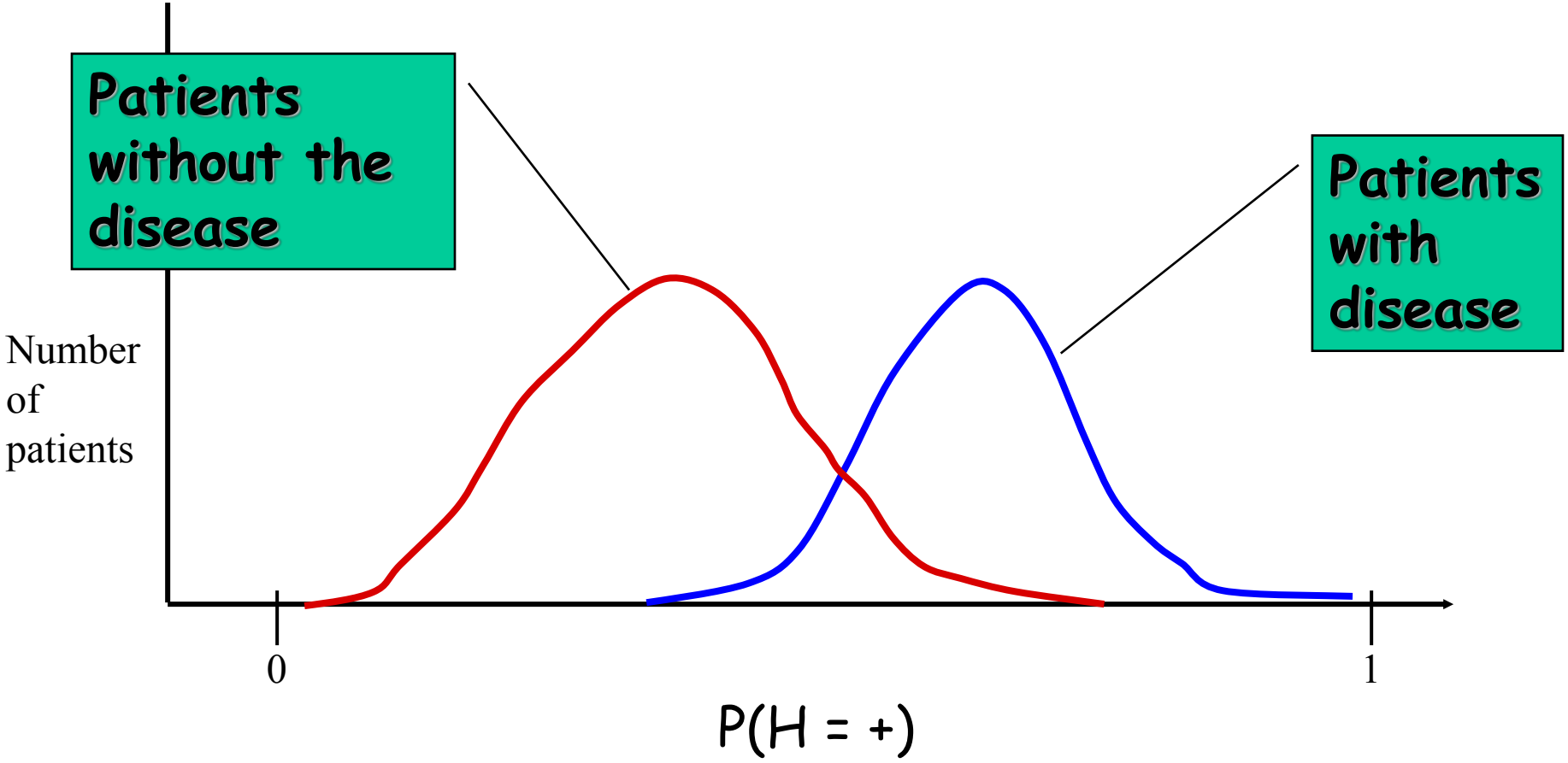
- *ROC = Receiver Operating Characteristic*
- Started in electronic signal detection theory (1940s - 1950s)
- Has become very popular in biomedical applications
- Can be used to compare tests/procedures
- Also used to in AI assess classifiers and diagnostic systems

ROC curves: simplest case

- We will look at analyses for hypotheses that have just two states (it can be generalized to more)
- Consider a Bayesian network with a hypothesis node H that has two states: +,-
- For any case, the network produces P(H = +).

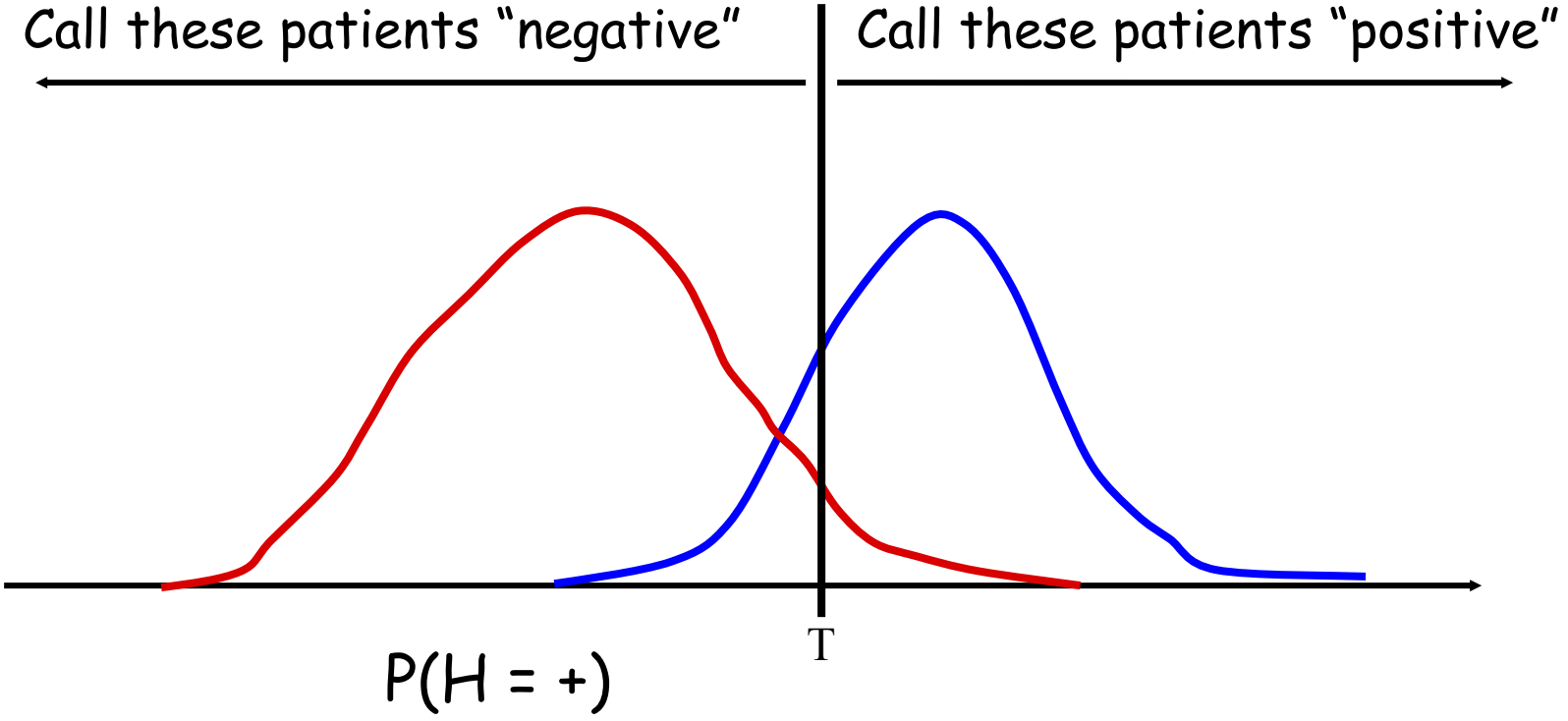
Example

We put 100 patients into the Bayes net: 50 +, 50- and get the following distribution

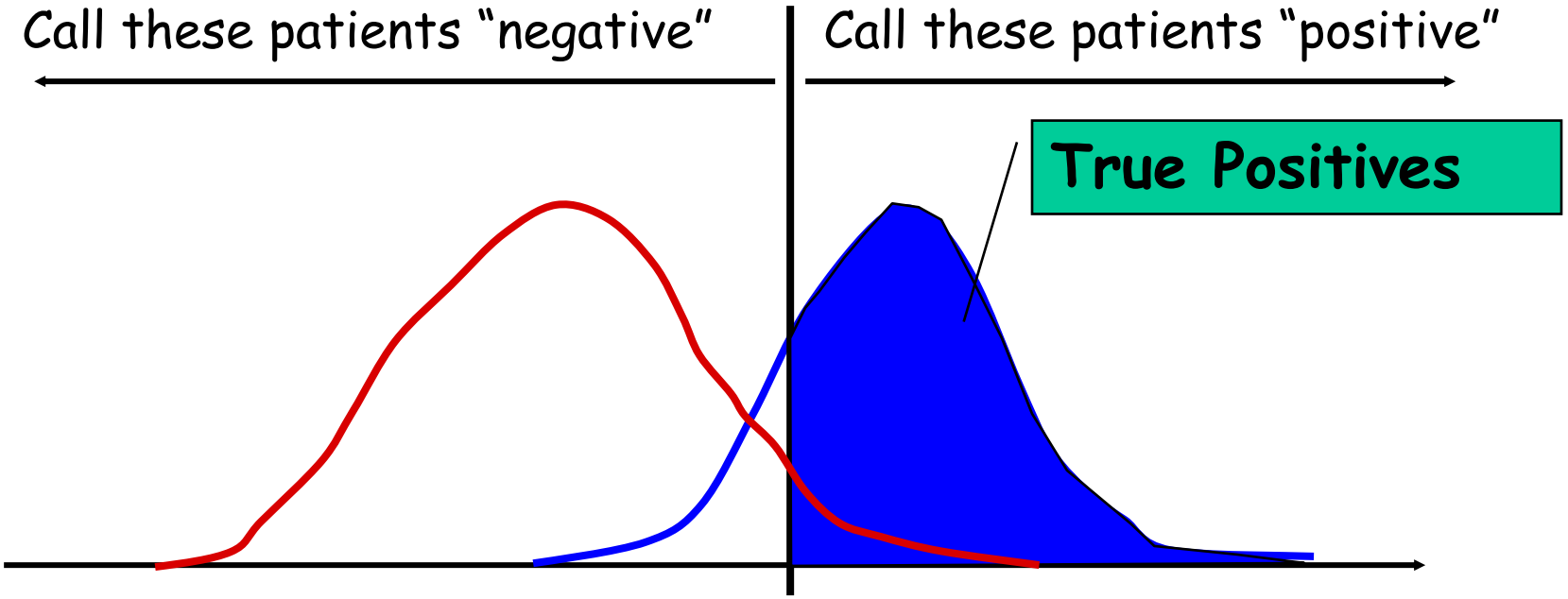


Threshold

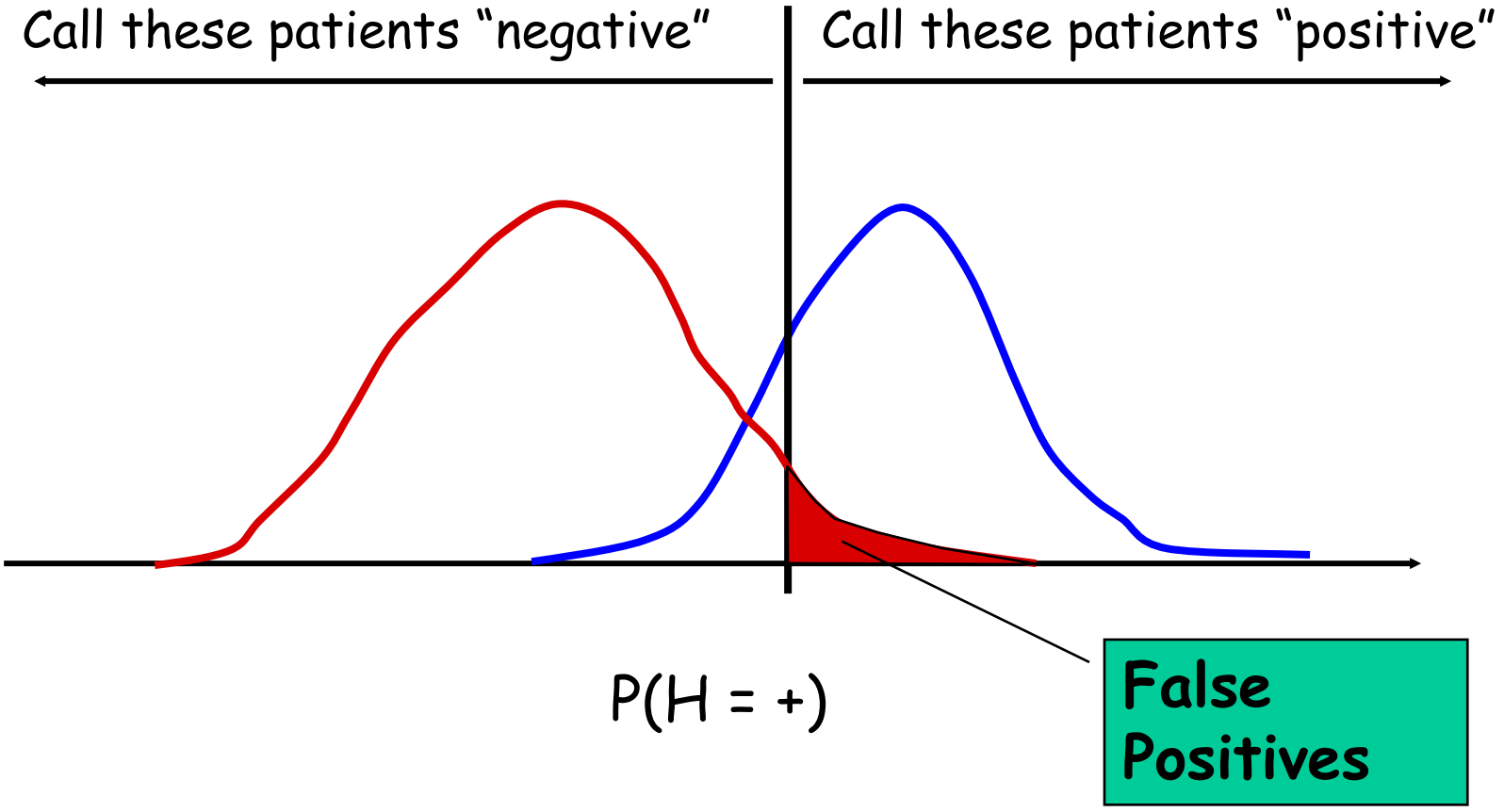
If we wish to use the network to classify cases as + or -, we need to choose a threshold T such that if P(H = +) > T we say it is + and otherwise -

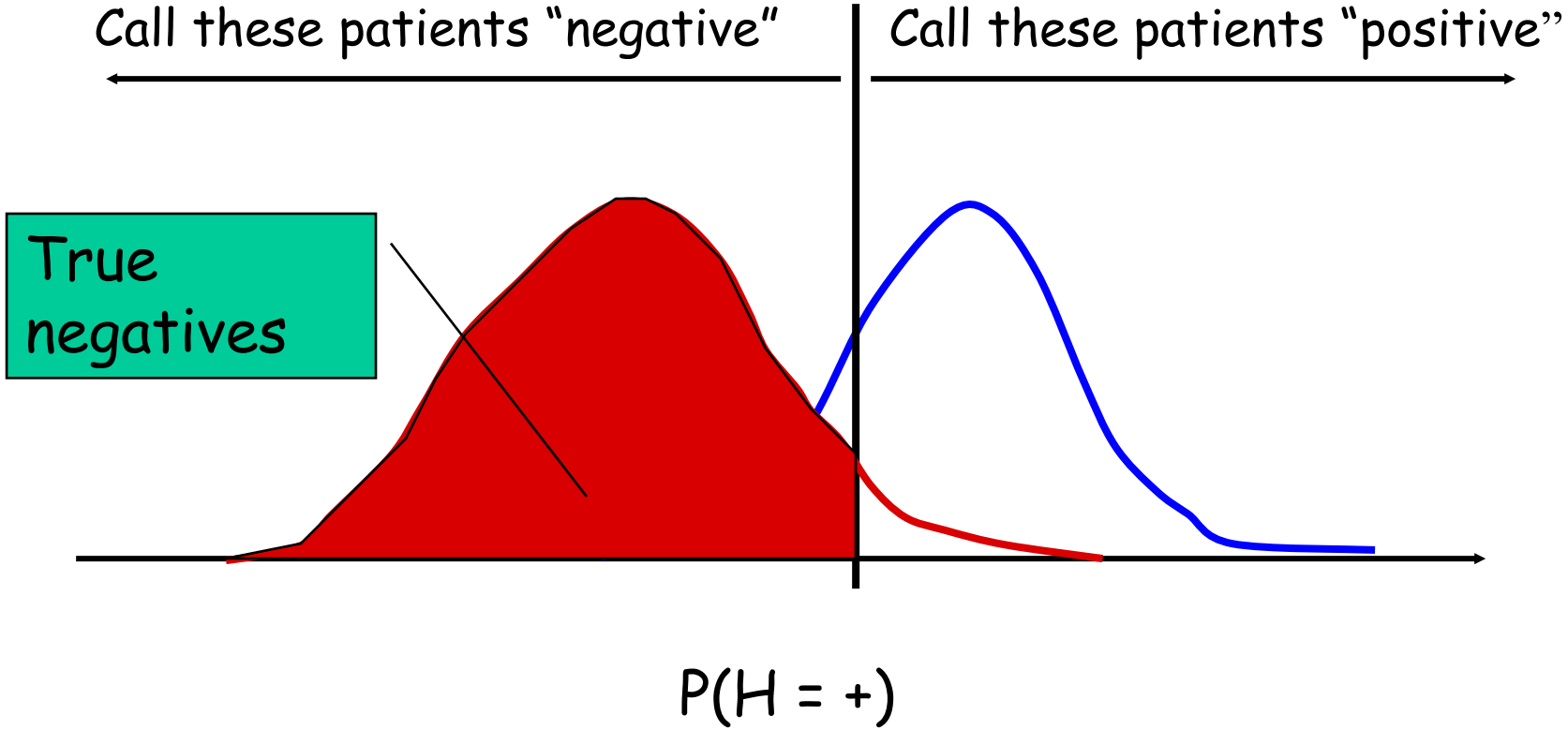


Some definitions ...

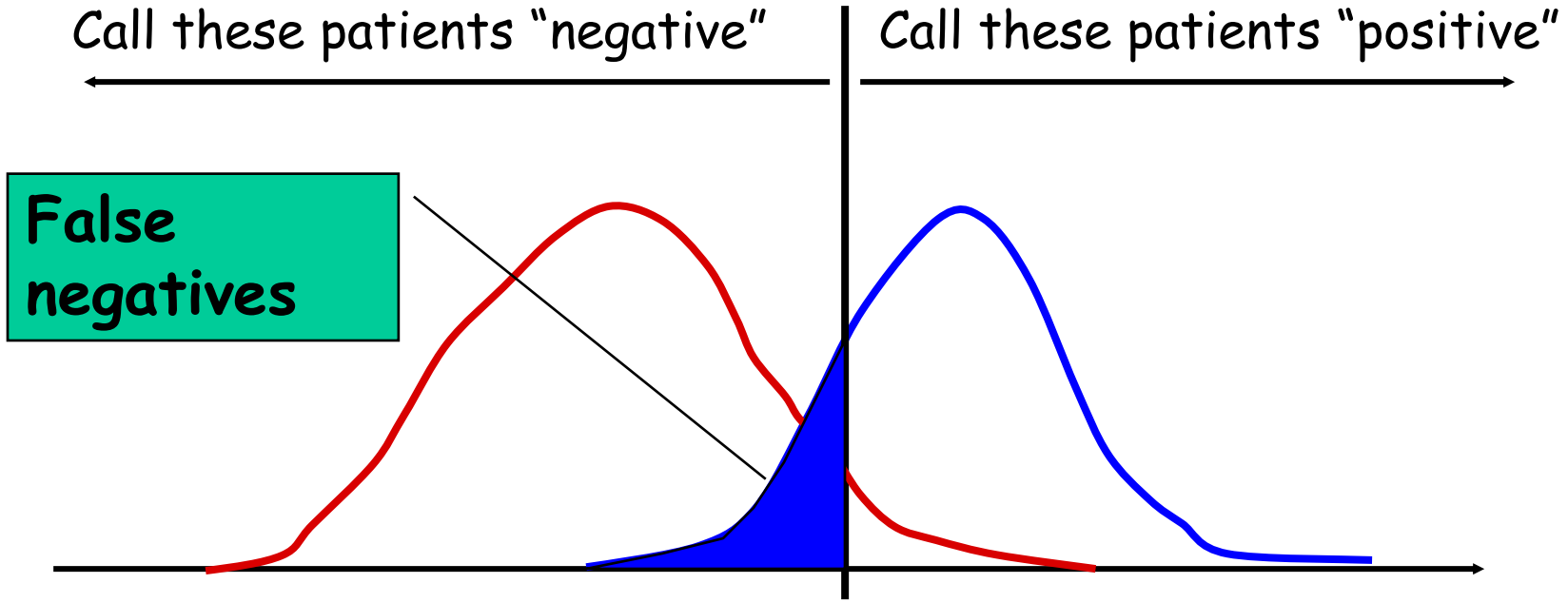


without the disease
with the disease


Call these patients "negative"
Call these patients "positive"
P(H = +)
False Positives
without the disease
with the disease

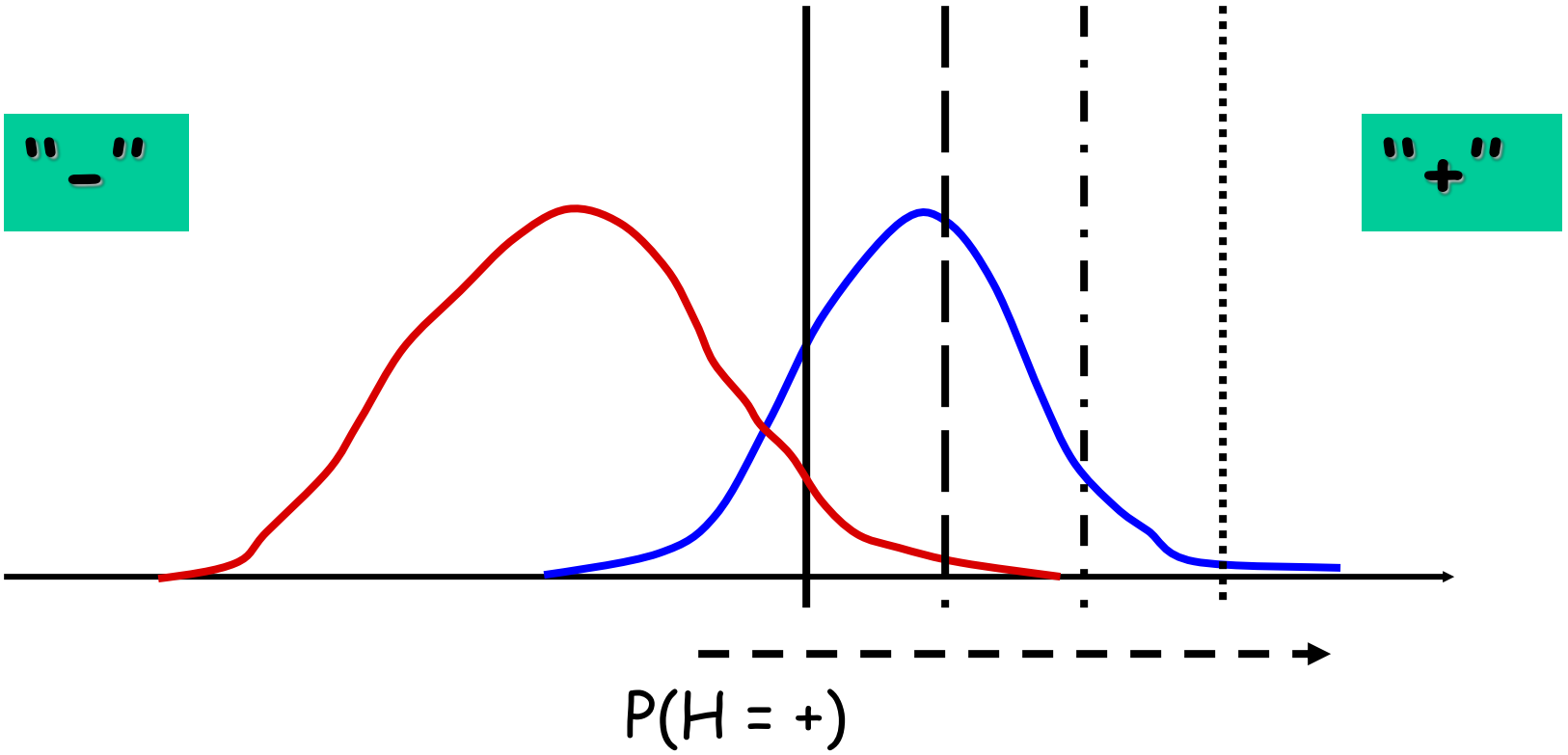


without the disease
with the disease



without the disease
with the disease

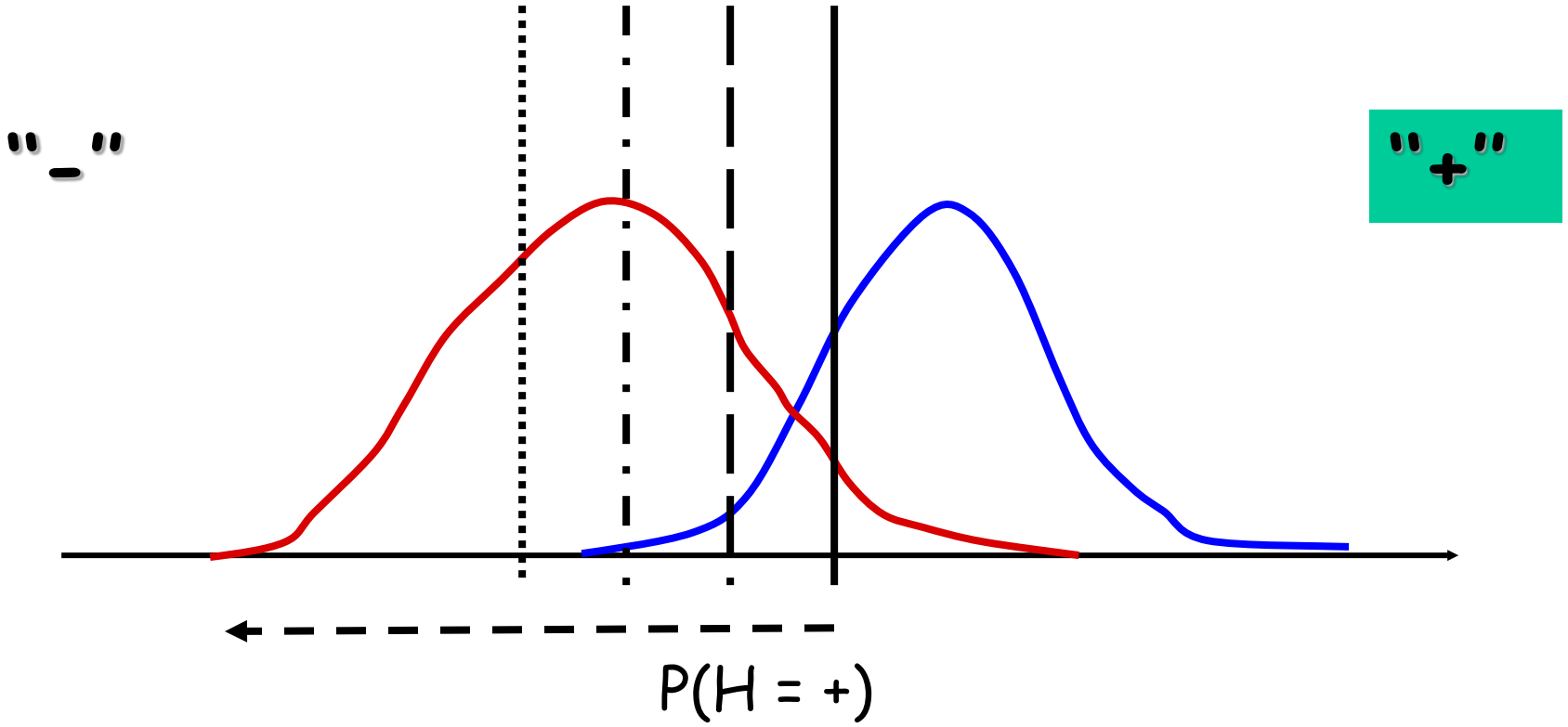
Moving the Threshold: right



without the disease
with the disease

Decreases False Positives
Increases False Negatives

Moving the Threshold: left



without the disease
with the disease

Decreases False Negatives
Increases False Positives

For any given threshold, we can produce a Contingency Table

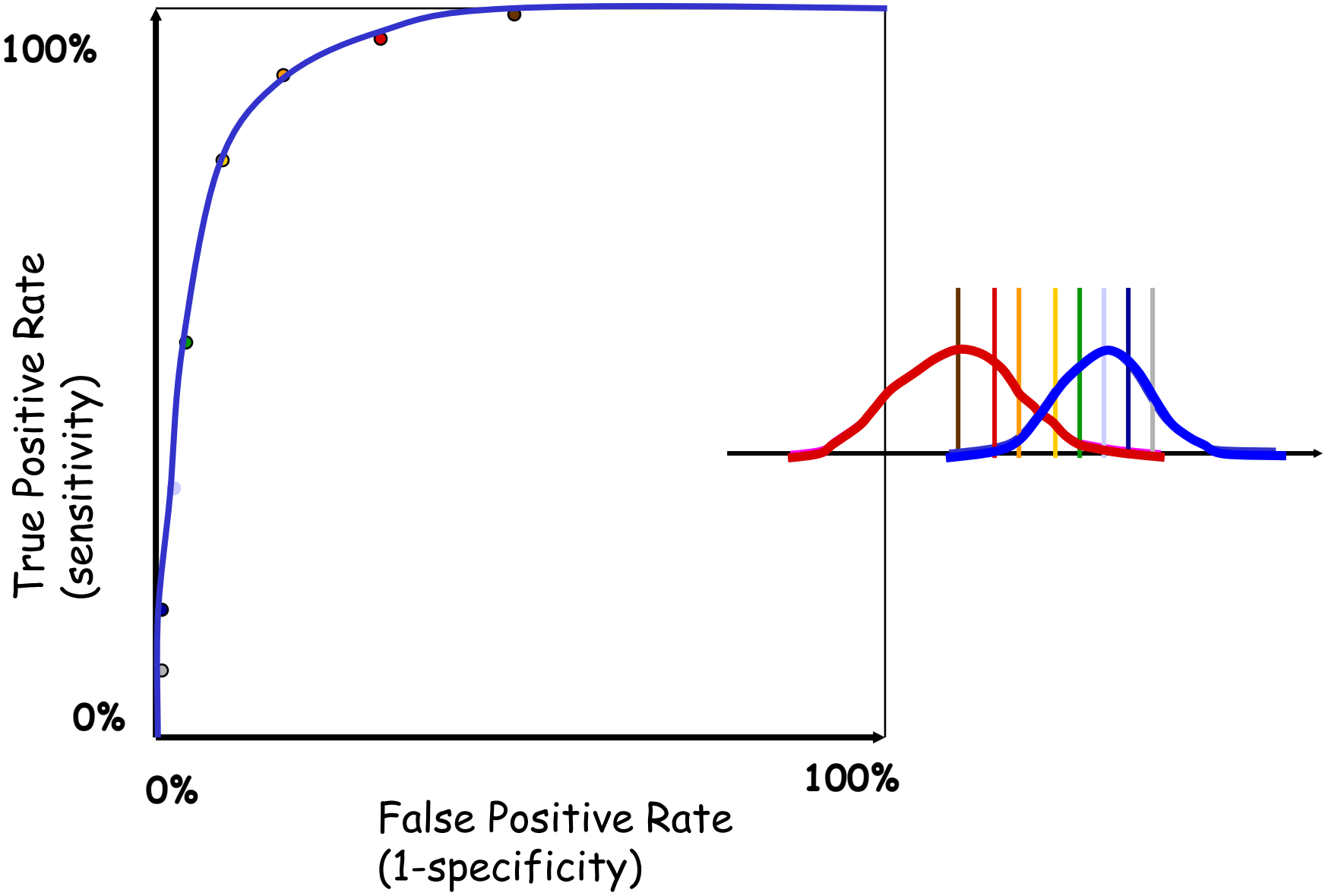
Actual

Predicted

	-	+	Total
-	TN	FN	N’
+	FP	TP	P’
Total	N	P	

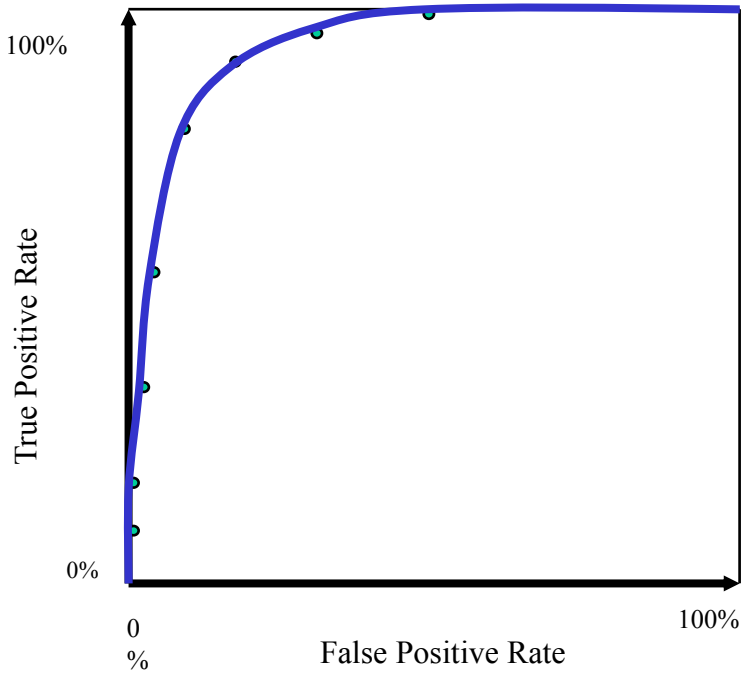
True Positive Rate (TPR) = TP/P = TP/(TP + FN)
False Positive Rate (FPR) = FP/N = FP/(FP + TN)
Sensitivity = TPR
Specificity = 1 – FPR

ROC curve

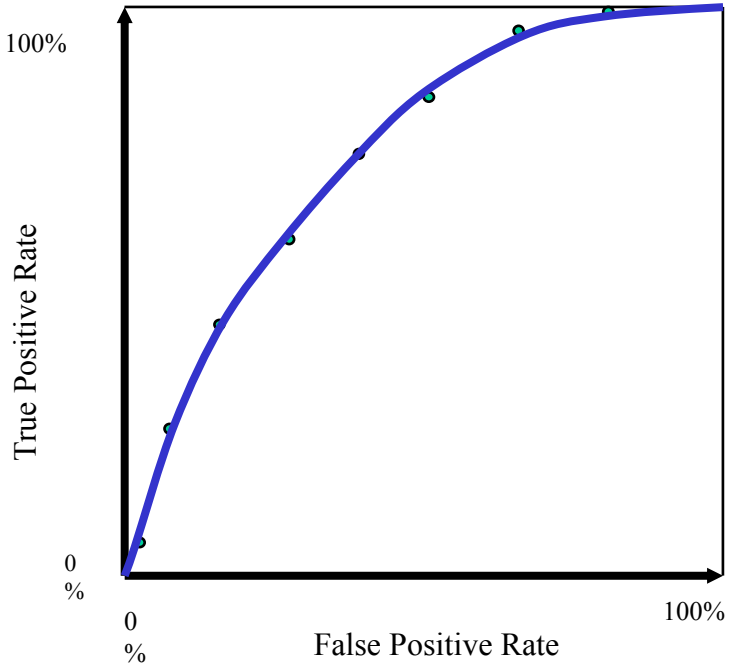

100%
True Positive Rate
(sensitivity)
0%
0%
100%
False Positive Rate
(1-specificity)

ROC curve comparison

A good test:

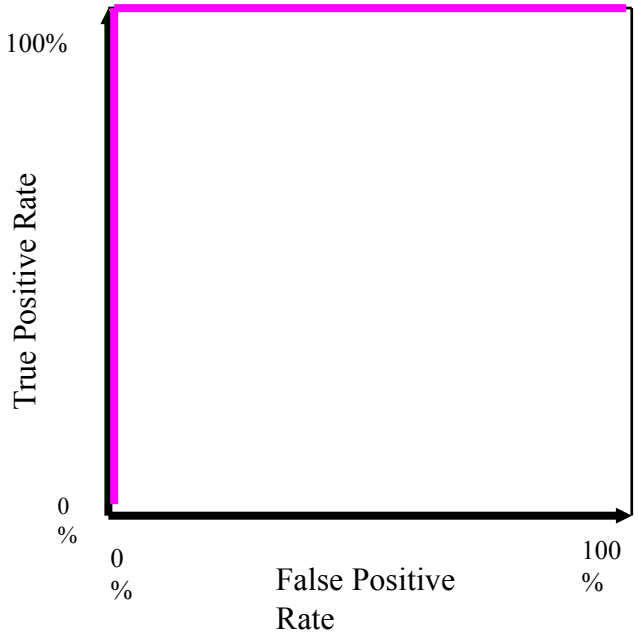


A poor test:



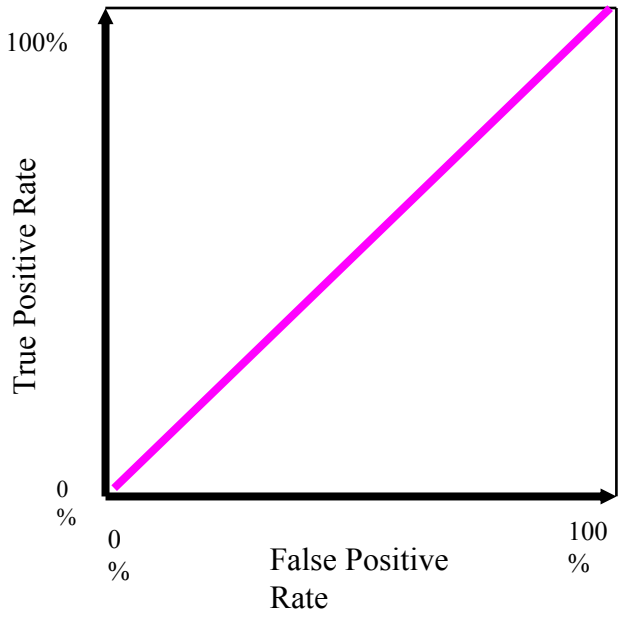
ROC curve extremes

Best Test:



The distributions don't overlap at all

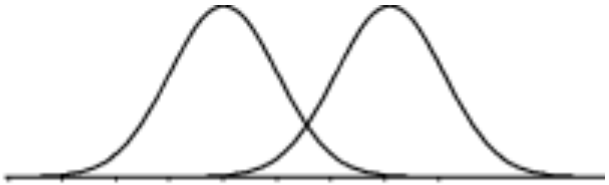
Worst test:



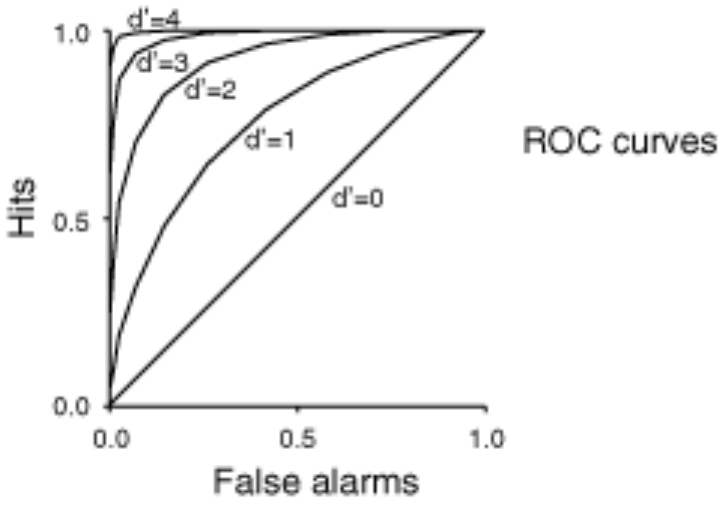
The distributions overlap completely

ROC Analysis

d' = 1 (lots of overlap)



d' = 3 (not much overlap)



ROC curves

The area under the curve (A_z) is an overall performance indicator for the accuracy of a system.

Interpretation of A_Z

- A_Z can be interpreted as the probability that the test result from a randomly chosen diseased individual is more indicative of disease than that from a randomly chosen nondiseased individual:

$$P(X_i \geq X_j \mid D_i = 1, D_j = 0),$$

where $X_i = P(H_i = +)$ and $X_j = P(H_j = +)$

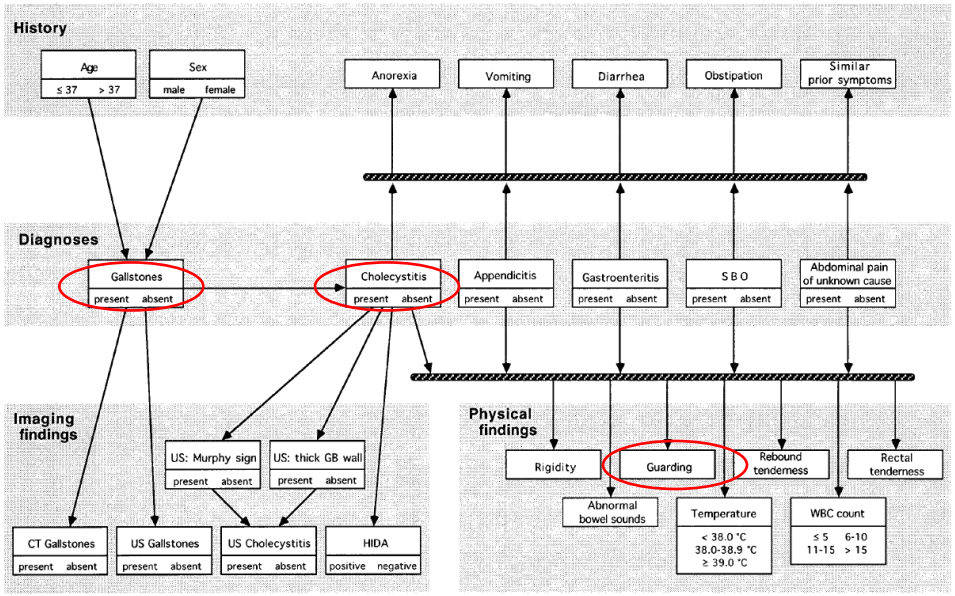
- So can think of this as a nonparametric distance between disease/nondisease test results

Explainability

Explanations in Bayes nets

- Consider a Bayes net with a 2-state hypothesis node H
- We enter a set of evidence E and observe that the probability of the hypothesis increases
- We want to know why it increased
- Two step process
  - List the influential pieces of evidence E_i
  - Show the paths along with the influence flowed

Gallbladder Network



A 41-year-old woman presents with anorexia and acute abdominal pain; she denies vomiting, diarrhea, obstipation or similar previous symptoms. Guarding is present with no rigidity or rebound tenderness. There were no abnormal bowel sounds. Her white blood cell count is 12,600.
We would like to diagnose the presence of gallstones.

Fig. 1. Bayesian-network model of gallbladder disease. The horizontal bars simplify the figure: all five nodes that lead to the bars influence all of the nodes that lead from the bars.

Before presenting any evidence, the probability of GALLSTONES being present is 0.128.
Presentation of the evidence results in a posterior probability of 0.227 for the presence of GALLSTONES.

The following pieces of evidence are considered important (in order of importance):
Presence of GUARDING results in a posterior probability of 0.175 for GALLSTONES.
AGE of 41 results in a posterior probability of 0.172 for GALLSTONES.

Their influence flows along the following paths:
GUARDING is caused by CHOLECYSTITIS, which is caused by GALLSTONES.
AGE influences GALLSTONES.

Influential Evidence

- Influential pieces of evidence
- Consider a hypothesis H and a set of evidence $E=\{E_1, E_2,\ldots, E_n\}$
- We measure the influence of a piece of evidence E_i on the hypothesis H in terms of whether and to what extent the shift from $P(H)$ to $P(H|E_i)$ agrees with the shift from $P(H)$ to $P(H|E)$.

- $influence(H; E; E_j) = \sum_{h_j \in H} I(h_j; E) I(h_j; E_i)$
- Information provided about a by b (mutual information)

  I(a;b) = log(P(a|b)/P(a))

Paths of Influence

1. Identify all paths along with evidence can flow based on d-separation (active paths).
2. Compute the strengths of the paths.
3. Display the top N strongest paths.

- Strength of path – chain is only a strong as its weakest link
- For each node *N* along an active path, from E_i to *H*, compute *impact(N; E_i)*

$$impact(N; E_i) = \sum_{n_j \in H} \left| I(n_j; E_i) \right|$$

- The strength of the path is the minimum of the impacts of the nodes