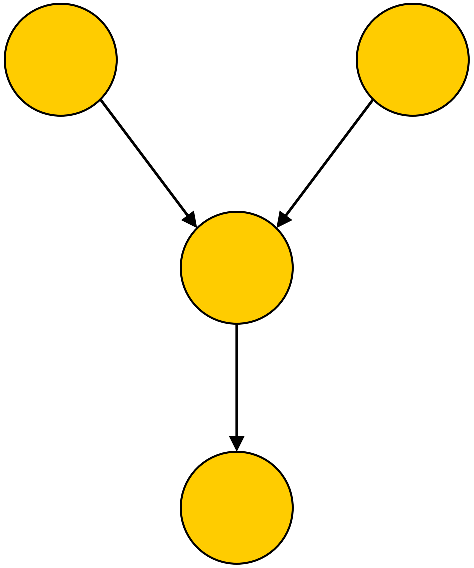


Bayesian Networks
Part I: Probability Primer

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.

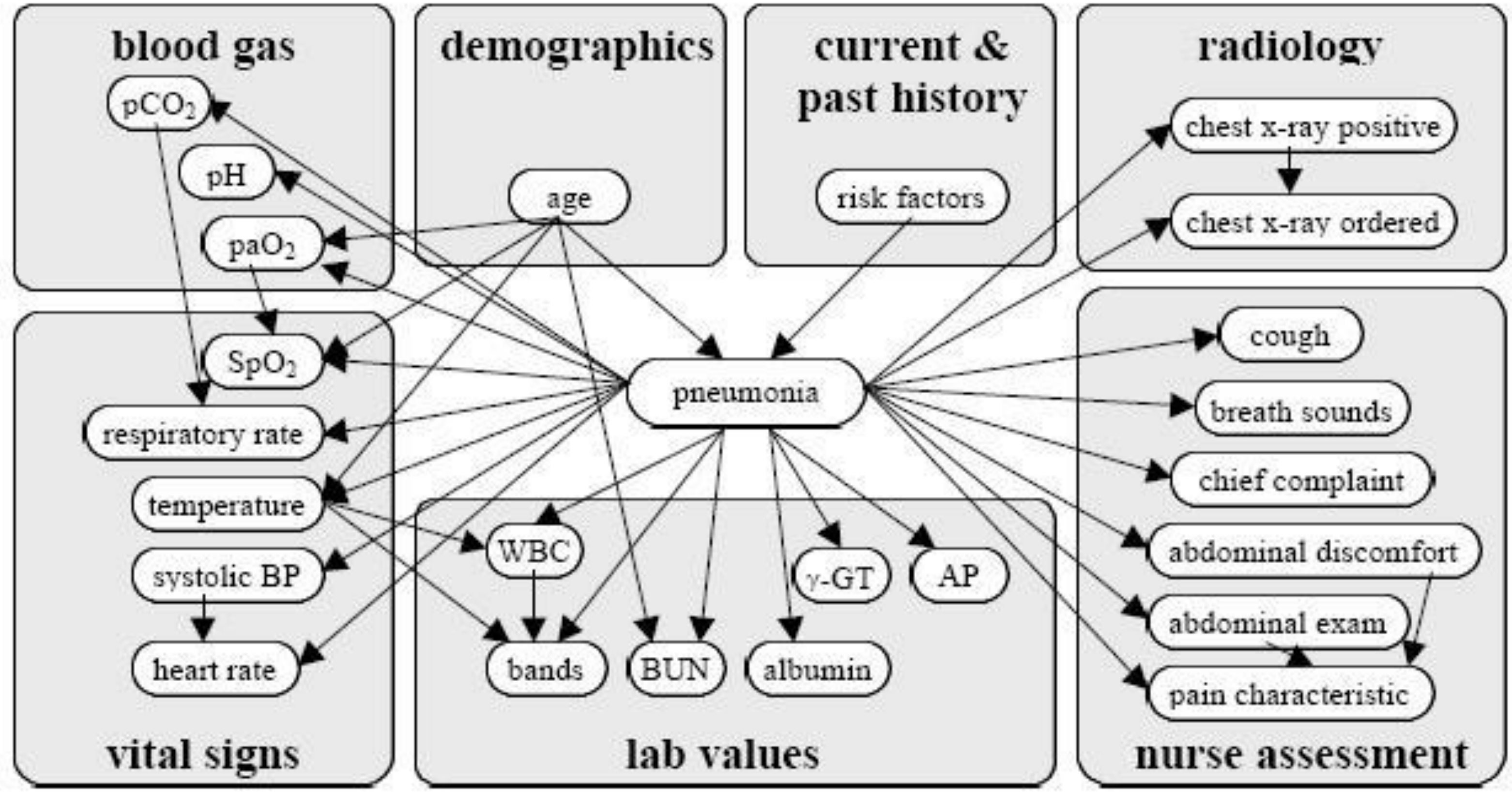
Introduction

- 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)



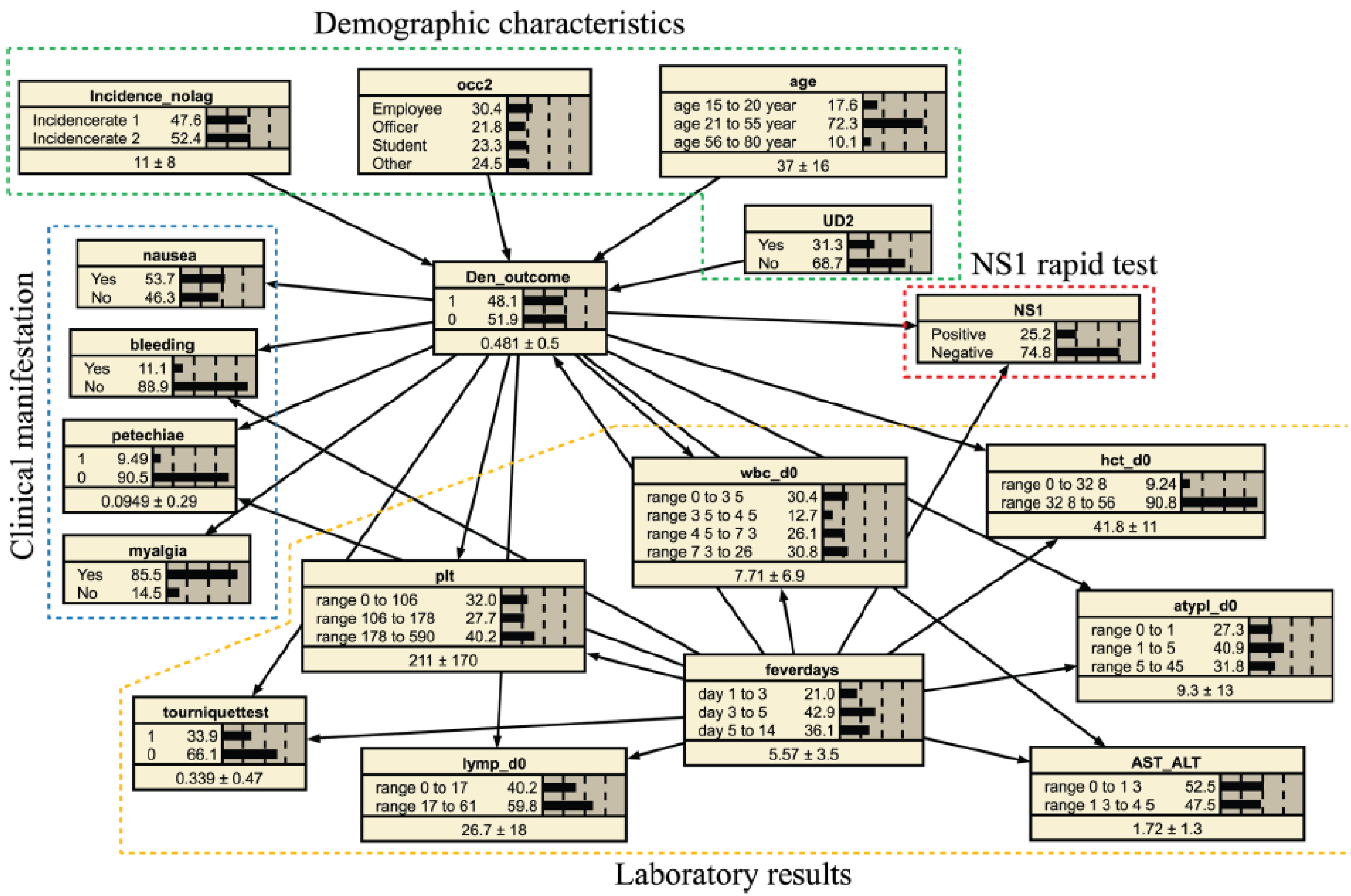


Fig 1. Final Bayesian network model for dengue diagnosis.

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

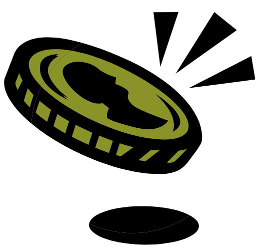
- Interpretable (Explainable)
  - Decision trees
  - KNN
  - Association rules
  - Bayesian Networks

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

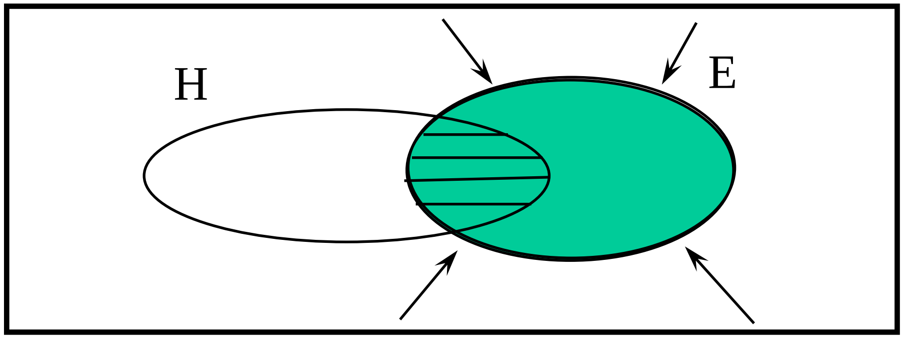
- **Symbols:**
  - We will use capital letters (A, B, …) for random variables and lower-case letters (a, b, …) for their values. E.g. $A = a_1$
  - A,B: A *and* B
  - So P(A,B) means the probability of A *and* B
- We will write A when we mean any value of the random variable A
  - P(Temperature) means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

Probability Basics

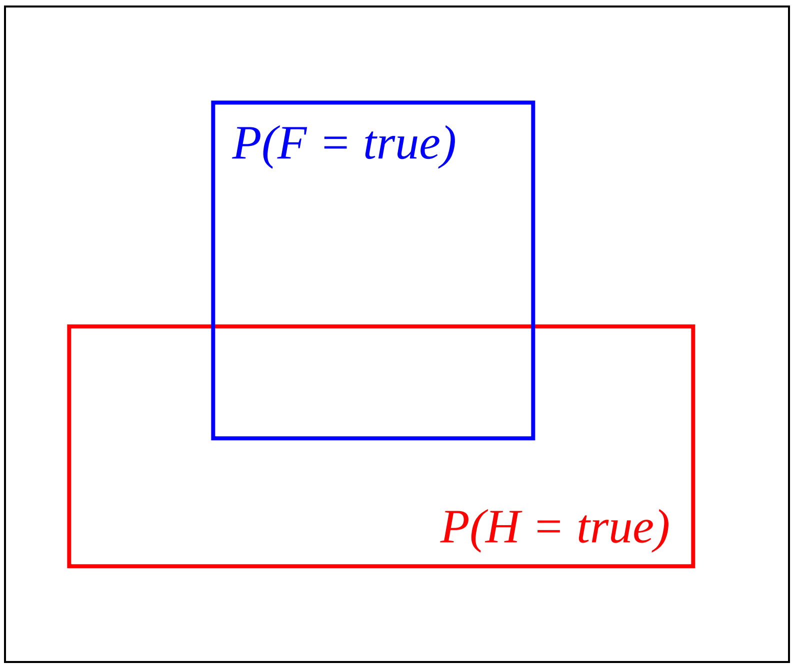
- **Axioms of probability**
  - $P(A) \geq 0$, where A is any proposition
  - $P(T) = 1$
  - $P(A \text{ or } B) = P(A) + P(B)$ if A and B are mutually exclusive

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



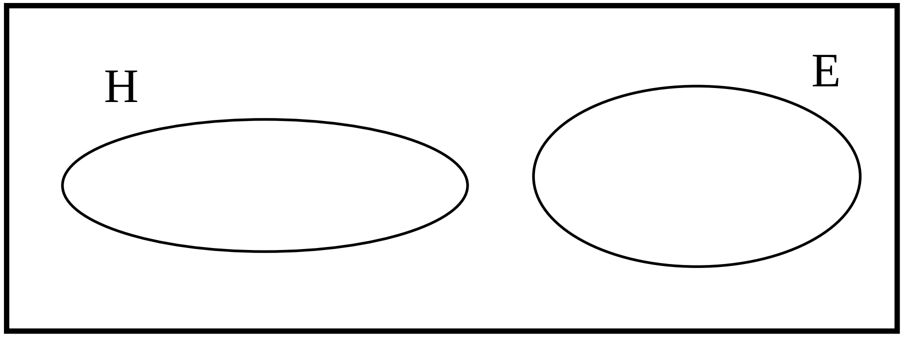
$$P(H=true \mid F=true)$$

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

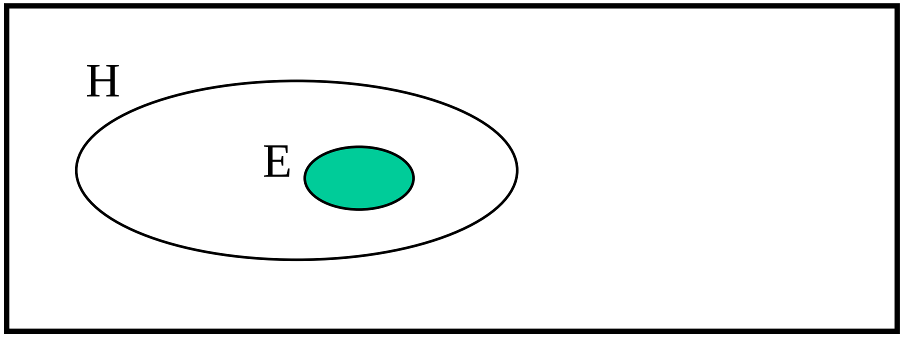
$$= \frac{P(H = true, F = true)}{P(F = true)}$$

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

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

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

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 2*n* 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

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

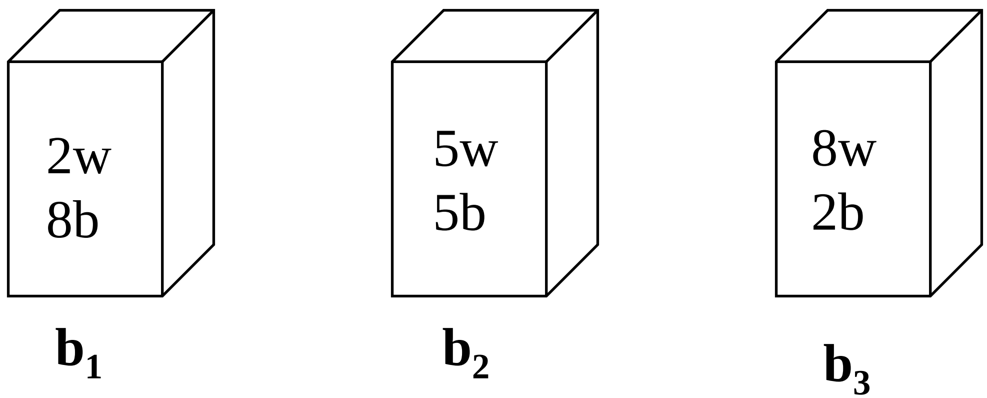
Bayes' rule allows us to express the quantity P(H|E), which people often find hard to assess, in terms of quantities that can be drawn from experience. Can think from cause (H) to effect (E).

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)

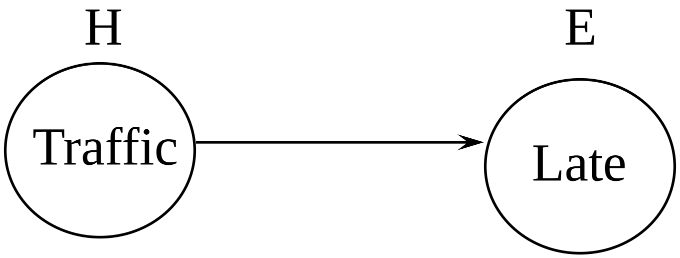
$$P(B=b_1 \mid C=w) = \frac{P(C=w \mid B=b_1)P(B=b_1)}{P(C=w)}$$

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

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

Confirmation is Symmetric

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



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

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

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

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