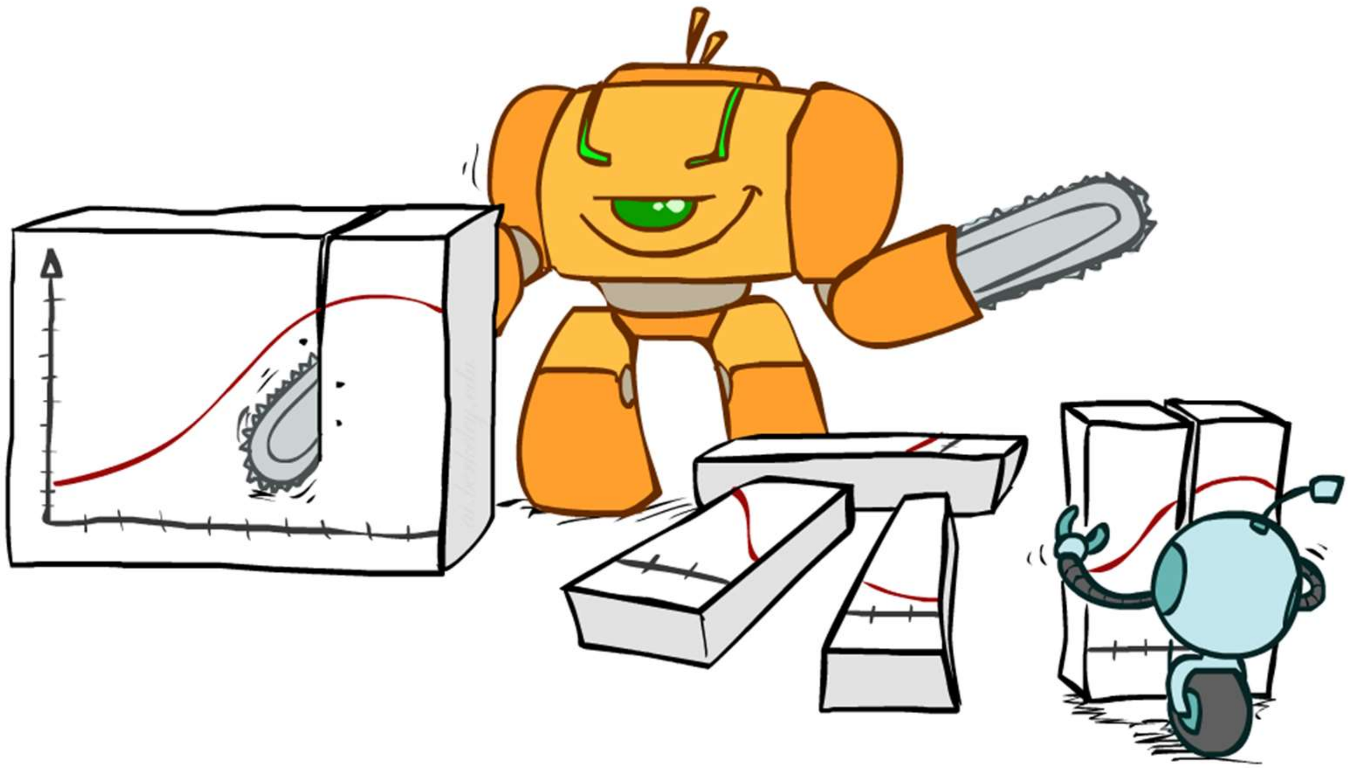
Bayes Rule



Bayes' Rule

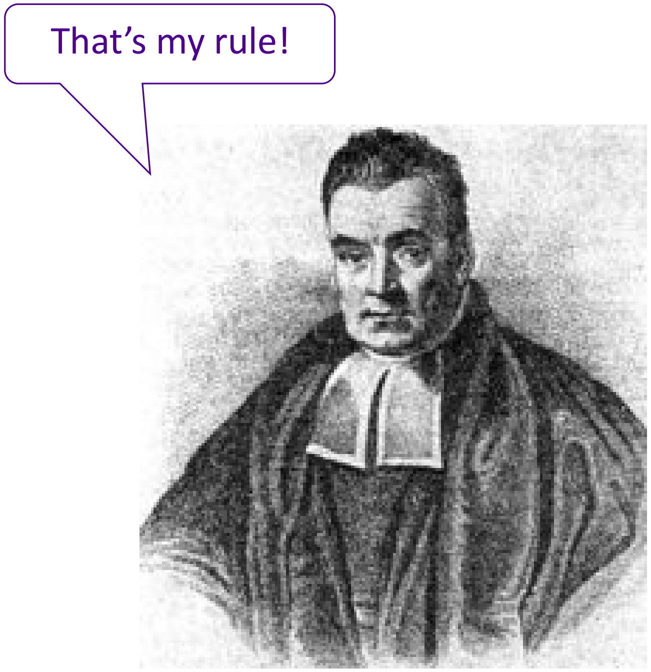
- Two ways to factor a joint distribution over two variables:

$$P(x,y) = P(x|y)P(y) = P(y|x)P(x)$$

- Dividing, we get:

$$P(x|y) = \frac{P(y|x)}{P(y)}P(x)$$

- Why is this at all helpful?
  - Lets us build one conditional from its reverse
  - Often one conditional is tricky but the other one is simple
  - Foundation of many systems we'll see later (e.g. ASR, MT)
- In the running for most important AI equation!



Inference with Bayes' Rule

- Example: Diagnostic probability from causal probability:

$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- Example:
  - C: Coronavirus, F: fever

$$\left.\begin{aligned} P(+c) &= 0.00001 \\ P(+f|+c) &= 0.8 \\ P(+f|-c) &= 0.01 \end{aligned}\right\} \text{Example givens}$$

$$P(+c|+f) = \frac{P(+f|+c)P(+c)}{P(+f)} = \frac{P(+f|+c)P(+c)}{P(+f|+c)P(+c) + P(+f|-c)P(-c)} = \frac{0.8 \times 0.00001}{0.8 \times 0.00001 + 0.01 \times 0.9999} \approx 0.0008$$

  - Note: posterior probability of coronavirus still very small
  - Note: you should still get fevers checked out! Why?

Tampere University

Quiz: Bayes' Rule

- Given:

$P(W)$

R	P
sun	0.8
rain	0.2

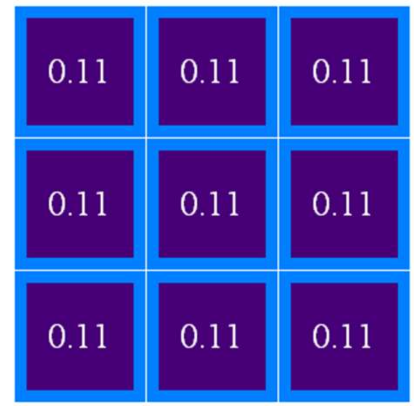
$P(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

- What is P(W | dry) ?

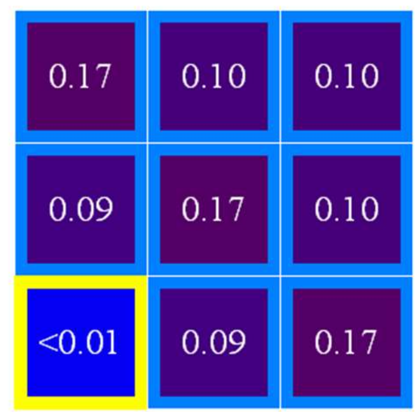
Ghostbusters, Revisited

- Let's say we have two distributions:
  - Prior distribution over ghost location: $P(G)$
    - Let's say this is uniform
  - Sensor reading model: $P(R \mid G)$
    - Given: we know what our sensors do
    - R = reading color measured at (1,1)
    - E.g., $P(R = \text{yellow} \mid G = (1,1)) = 0.1$



- We can calculate the posterior distribution $P(G|r)$ over ghost locations given a reading using Bayes' rule:

$$P(g|r) \propto P(r|g)P(g)$$



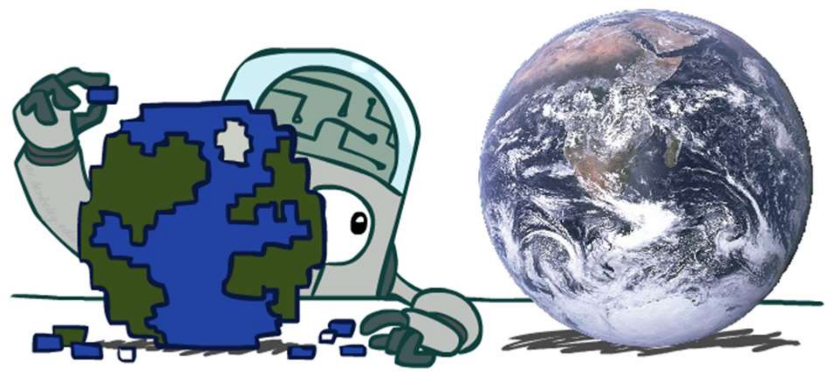
Video of Demo Ghostbusters with Probability

Tampere University

Probabilistic Models

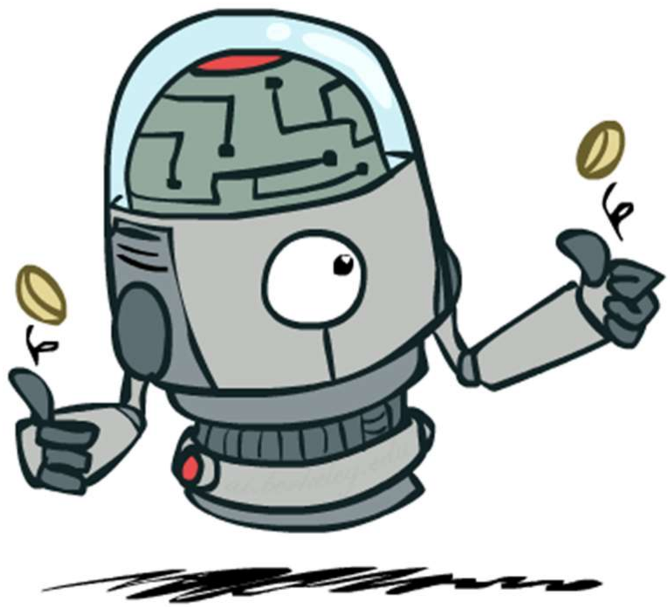
- Models describe how (a portion of) the world works

- Models are always simplifications
  - May not account for every variable
  - May not account for all interactions between variables
  - "All models are wrong; but some are useful."
    – George E. P. Box



- What do we do with probabilistic models?
  - We (or our agents) need to reason about unknown variables, given evidence
  - Example: explanation (diagnostic reasoning)
  - Example: prediction (causal reasoning)
  - Example: value of information

Independence



Independence

- Two variables are *independent* if:

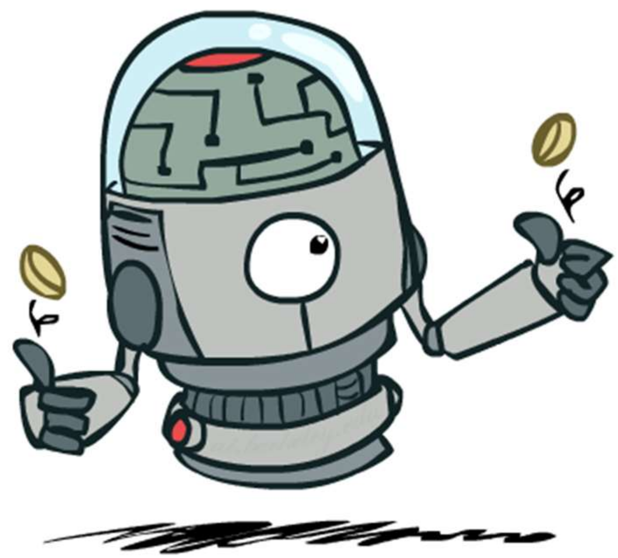
$$\forall x, y : P(x, y) = P(x)P(y)$$

  - This says that their joint distribution *factors* into a product two simpler distributions
  - Another form:

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

  - We write:

$$X \perp\!\!\!\perp Y$$



- Independence is a simplifying *modeling assumption*
  - *Empirical* joint distributions: at best "close" to independent
  - What could we assume for {Weather, Traffic, Cavity, Toothache}?

Example: Independence?

$P(T)$

T	P
hot	0.5
cold	0.5

$P_1(T,W)$

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

$P_2(T,W)$

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

$P(W)$

W	P
sun	0.6
rain	0.4

Example: Independence

- N fair, independent coin flips:

$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

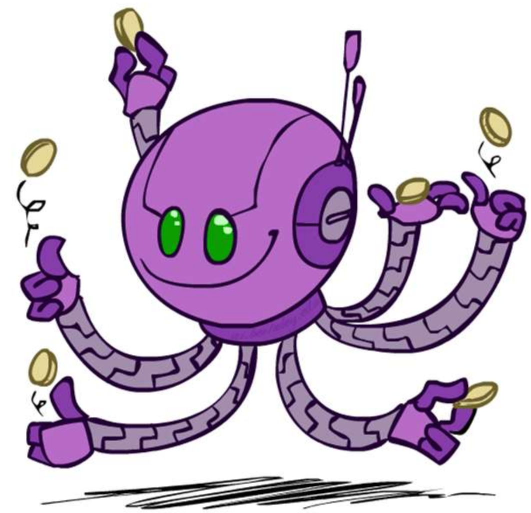
H	0.5
T	0.5

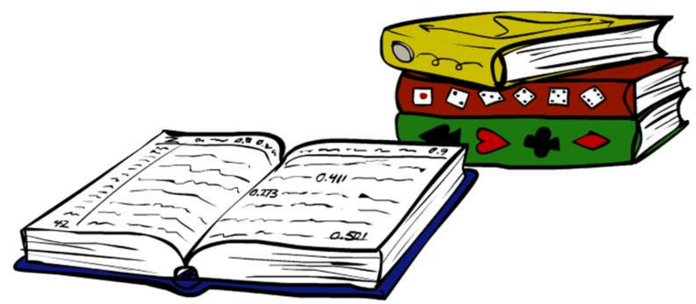
$\ldots$

$P(X_n)$

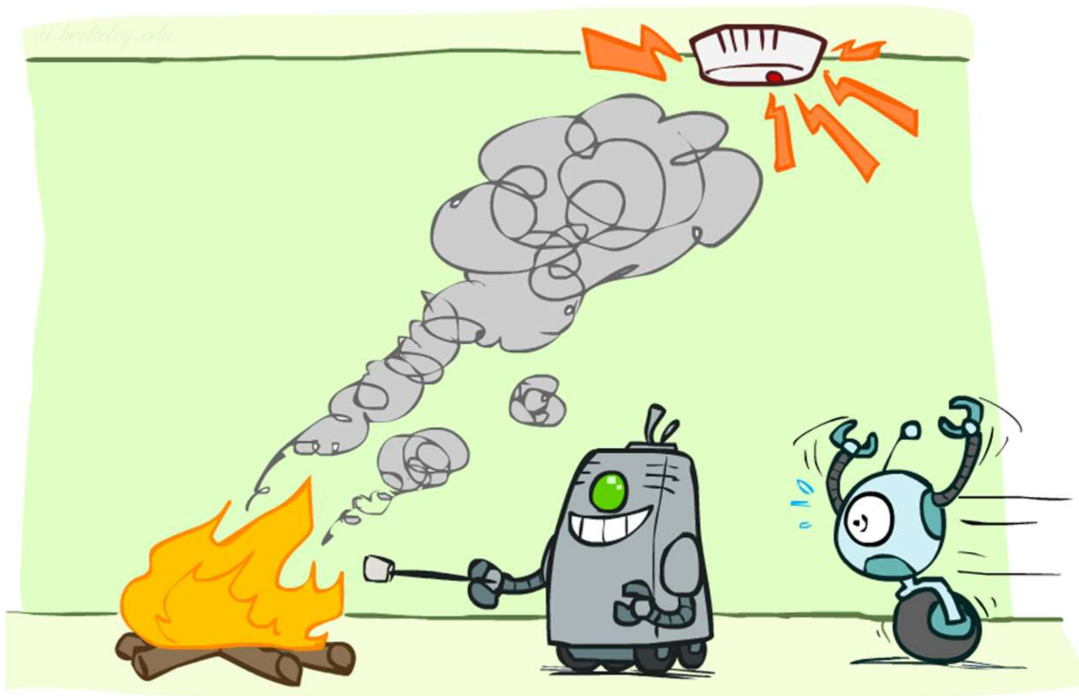
H	0.5
T	0.5

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





Conditional Independence



Tampere University

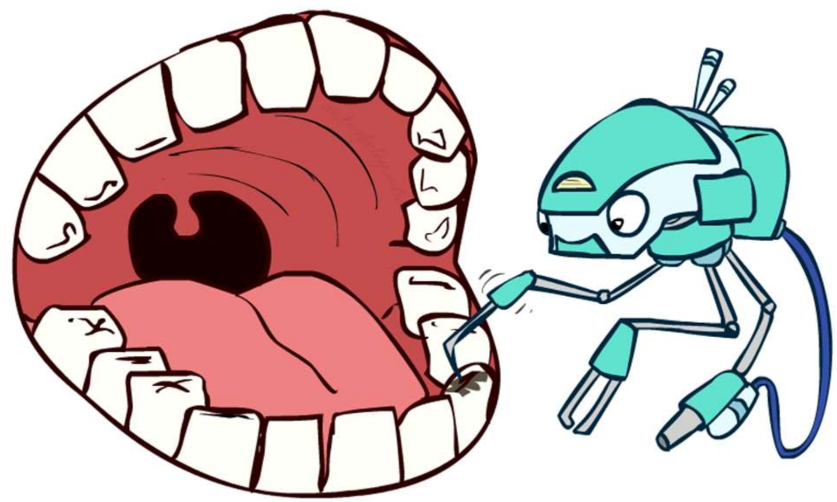
Conditional Independence

- P(Toothache, Cavity, Catch)

- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - P(+catch | +toothache, +cavity) = P(+catch | +cavity)

- The same independence holds if I don't have a cavity:
  - P(+catch | +toothache, -cavity) = P(+catch| -cavity)

- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)



- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
  - One can be derived from the other easily

Tampere University

Conditional Independence

- Unconditional (absolute) independence very rare (why?)

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

- X is conditionally independent of Y given Z $\quad X \perp\!\!\!\perp Y | Z$

  if and only if:

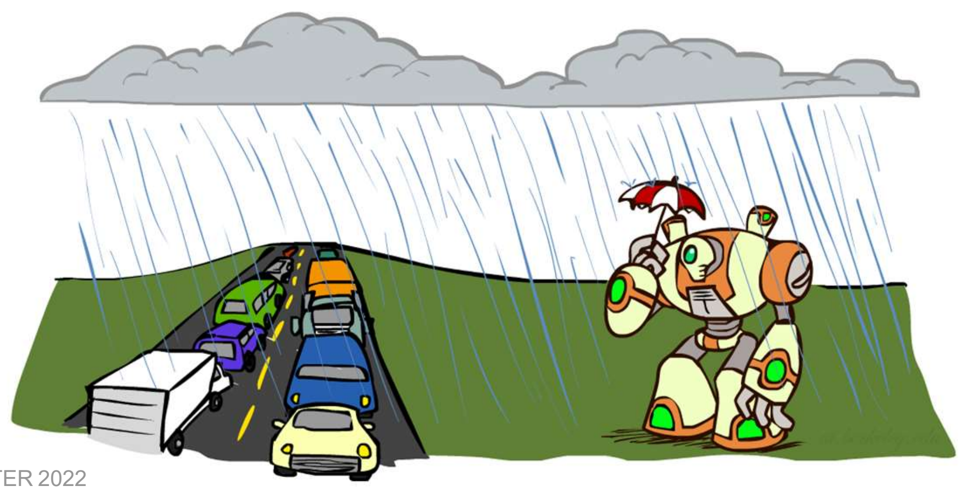
$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

  or, equivalently, if and only if

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

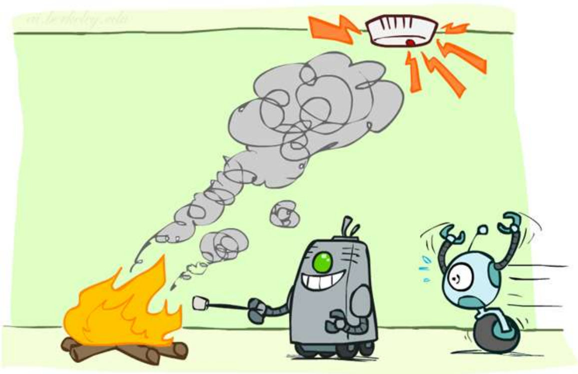
Conditional Independence

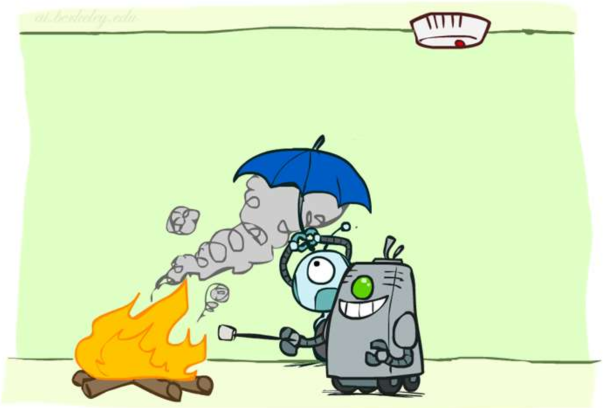
- What about this domain:
  - Traffic
  - Umbrella
  - Raining



Conditional Independence

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





Tampere University

Conditional Independence and the Chain Rule

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

- Trivial decomposition:

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

- With assumption of conditional independence:

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

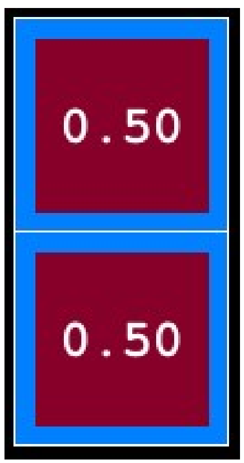
- Bayes' nets / graphical models help us express conditional independence assumptions

Ghostbusters Chain Rule

- Each sensor depends only on where the ghost is
- That means, the two sensors are conditionally independent, given the ghost position

- T: Top square is red
  B: Bottom square is red
  G: Ghost is in the top

- Givens:
  $P(+g) = 0.5$
  $P(-g) = 0.5$
  $P(+t \mid +g) = 0.8$
  $P(+t \mid -g) = 0.4$
  $P(+b \mid +g) = 0.4$
  $P(+b \mid -g) = 0.8$



$$P(T,B,G) = P(G)\,P(T|G)\,P(B|G)$$

T	B	G	$P(T,B,G)$
$+t$	$+b$	$+g$	0.16
$+t$	$+b$	$-g$	0.16
$+t$	$-b$	$+g$	0.24
$+t$	$-b$	$-g$	0.04
$-t$	$+b$	$+g$	0.04
$-t$	$+b$	$-g$	0.24
$-t$	$-b$	$+g$	0.06
$-t$	$-b$	$-g$	0.06

Naïve Bayes

- If all effects are conditionally independent given a single cause, the exponential size of knowledge representation is cut to linear
- A probability distribution is called a **naïve Bayes** (NB) model if all effects $E_1, \dots, E_n$ are conditionally independent, given a single cause C
- The full joint probability distribution can be written as

$$\mathbf{P}(C, E_1, \dots, E_n) = \mathbf{P}(C) \prod_i \mathbf{P}(E_i \mid C)$$

- A simplifying assumption even in cases where the effect variables are not conditionally independent given the cause variable
- In practice, NB systems can work surprisingly well, even when the independence assumption is not true