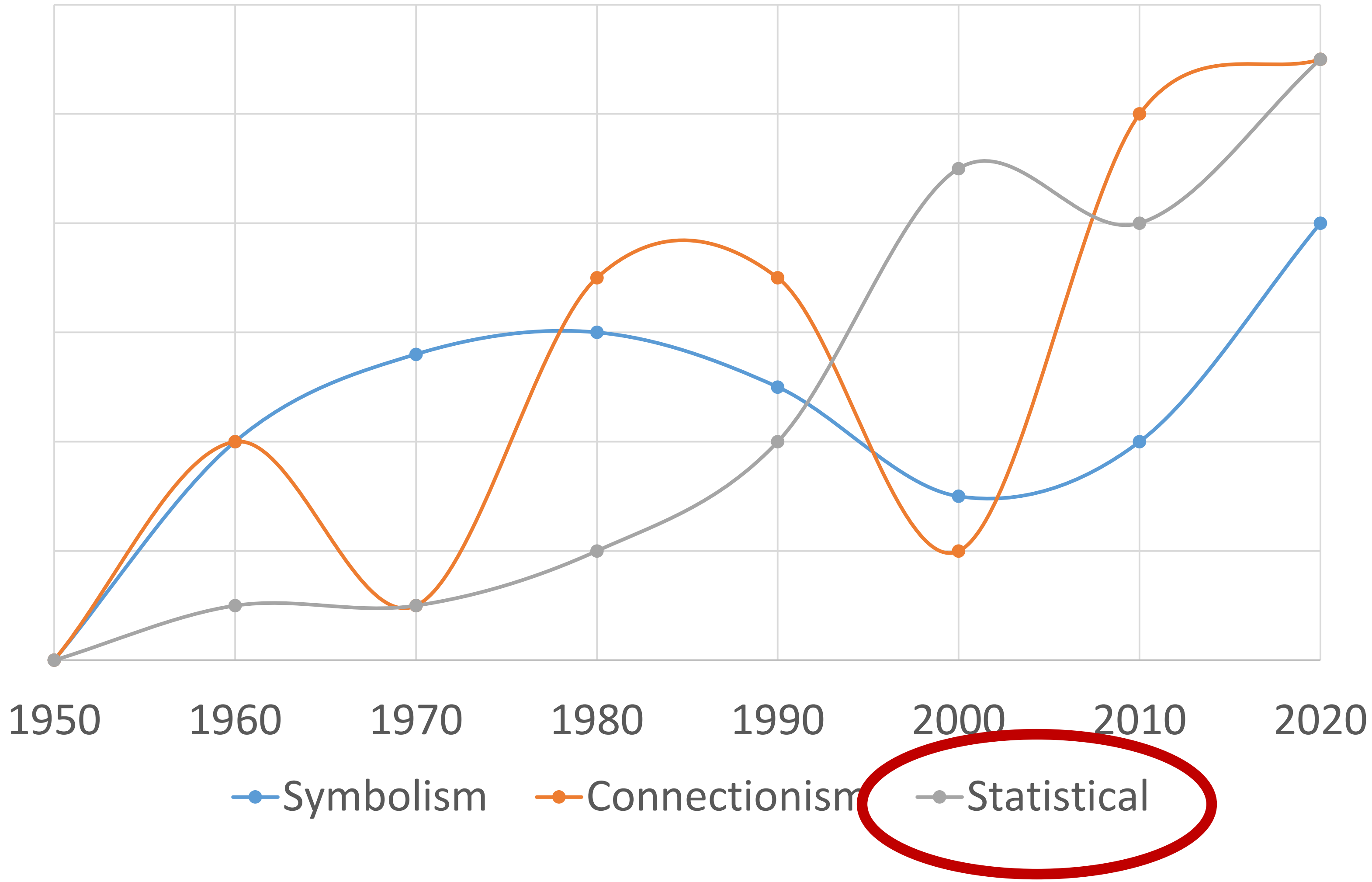
Three types of (strong) AI approaches



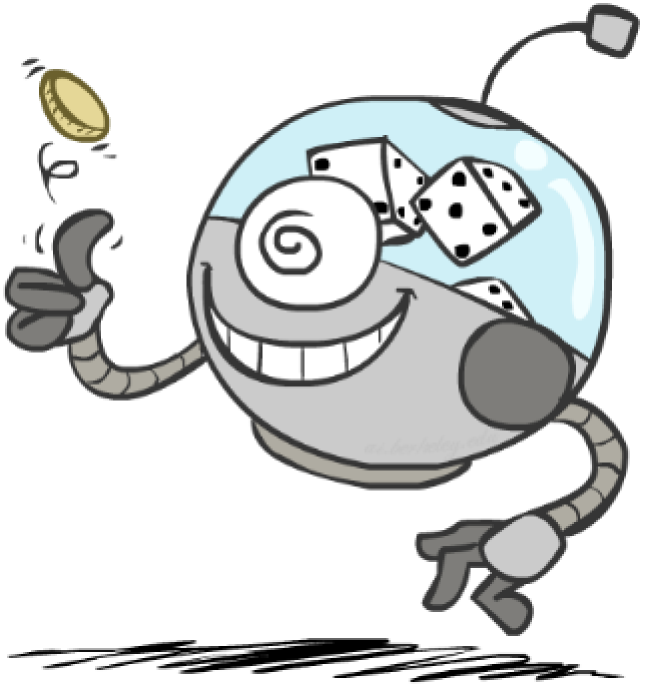
Probability

AIMA Chapter 13

[Adapted from slides by Dan Klein and Pieter Abbeel at UC Berkeley (ai.berkeley.edu).]

Outline

- Probability
  - Random Variables
  - Joint and Marginal Distributions
  - Conditional Distributions
  - Inference
  - Product Rule, Chain Rule, Bayes' Rule



Uncertainty

- My flight to New York is scheduled to leave at 11:25
  - Let action A_t = leave home t minutes before flight and drive to the airport
  - Will A_t ensure that I catch the plane?
- Problems:
  - noisy sensors (radio traffic reports, Google maps)
  - uncertain action outcome (car breaking down, accident, etc.)
  - partial observability (other drivers' plans, etc.)
  - immense complexity of modelling and predicting traffic, security line, etc.

Responses to uncertainty

- Ignore it – map directly from percept stream (known) to actions
  - Hopeless!
- Some sort of softening of logical rules (***fudge factors***)
  - $A_{1440} \rightarrow_{\mathbf{0.9999}} \mathit{CatchPlane}$
  - $\mathit{CatchPlane} \rightarrow_{\mathbf{0.95}} \neg\, \mathit{MajorTrafficJam}$
  - Hence, chaining these together, $A_{1440} \rightarrow_{\mathbf{0.949}} \neg\, \mathit{MajorTrafficJam}$
  - Oops
- Probability (Mahaviracarya (9th C.), Cardamo (1565))
  - Given the available evidence and the choice A_{120}, I will catch the plane with probability 0.92

Probability

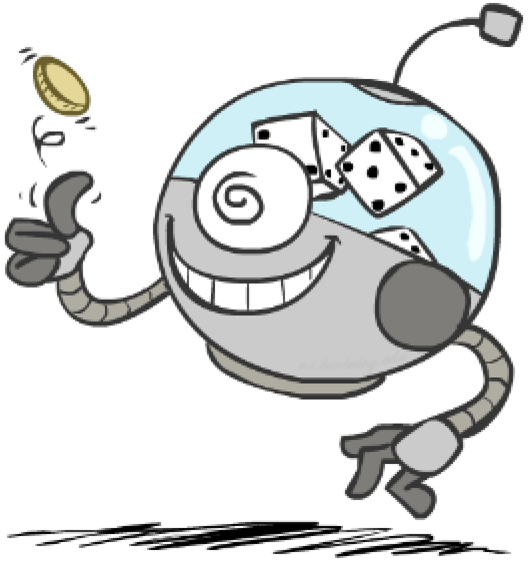
- Probability
  - Given the available evidence and the choice A_{120}, I will catch the plane with probability 0.92
- ***Subjective*** or ***Bayesian*** probability:
  - Probabilities relate propositions to one’s own state of knowledge
    - ignorance: lack of relevant facts, initial conditions, etc.
    - laziness: failure to list all exceptions, compute detailed predictions, etc.
  - Not claiming a “probabilistic tendency” in the actual situation (traffic is not like quantum mechanics)

Decisions

- Suppose I believe
  - $P(CatchPlane \mid A_{60}, \textit{all my evidence…}) = 0.51$
  - $P(CatchPlane \mid A_{120}, \textit{all my evidence…}) = 0.97$
  - $P(CatchPlane \mid A_{1440}, \textit{all my evidence…}) = 0.9999$
- Which action should I choose?
- Depends on my ***preferences*** for, e.g., missing flight, airport food, etc.
- ***Utility theory*** is used to represent and infer preferences
- ***Decision theory*** = utility theory + probability theory
- ***Maximize expected utility*** : $a^* = \mathrm{argmax}_a \sum_s P(s \mid a)\, U(s)$

Random Variables

- A random variable is some aspect of the world (formally a ***deterministic function*** of ω) about which we (may) be uncertain
  - *Odd* = Is the dice roll an odd number?
  - *T* = Is it hot or cold?
  - *D* = How long will it take to get to the airport?
- Random variables have domains
  - *Odd* in {true, false} e.g. *Odd*(1)=true, *Odd*(6) = false
    - often write the *event* Odd=true as odd, Odd=false as $\neg$odd
  - *T* in {hot, cold}
  - *D* in $[0, \infty)$

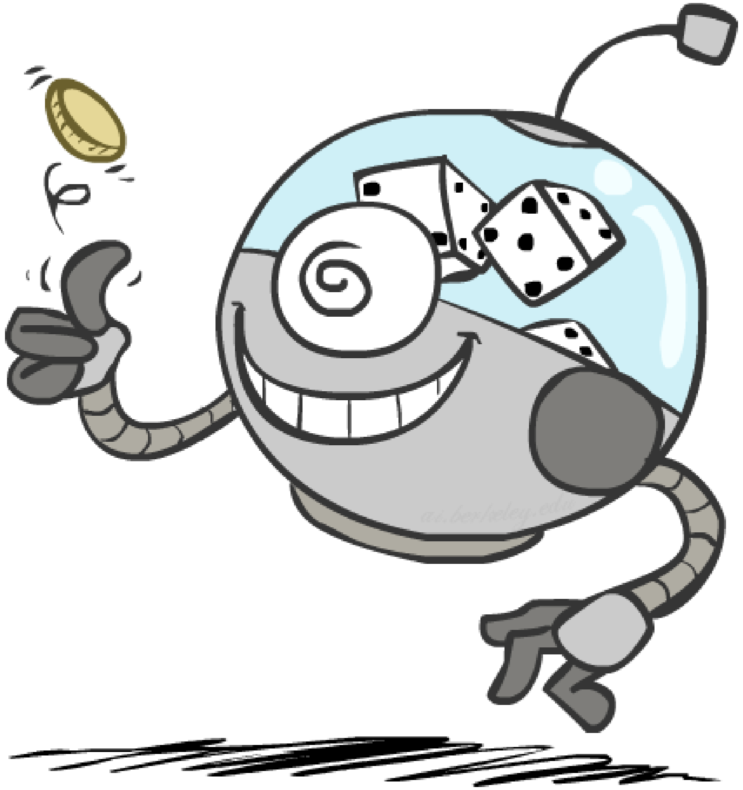


Random Variables

- A random variable is some aspect of the world about which we (may) have uncertainty
    - R = Is it raining?
    - T = Is it hot or cold?
    - D = How long will it take to drive to work?
    - L = Where is the pacman?

- We denote random variables with capital letters

- Like variables in a CSP, random variables have domains
    - R in {true, false}   (often write as {+r, -r})
    - T in {hot, cold}
    - D in [0, ∞)
    - L in possible locations, maybe {(0,0), (0,1), …}



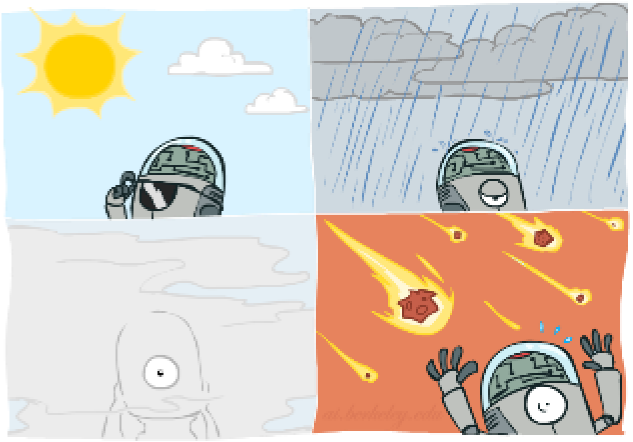
Probability Distributions

- Associate a probability with each value of a random variable
  - Temperature:

$P(T)$

T	P
hot	0.5
cold	0.5

  - Weather:



$P(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0

  - A probability is a single number

$$P(W = rain) = 0.1$$

Shorthand notation: $P(rain) = P(W = rain)$,

- Must have: $\forall x \;\; P(X = x) \geq 0$ and $\sum_x P(X = x) = 1$

Joint Distributions

- A *joint distribution* over a set of random variables: $X_1, X_2, \ldots X_n$ specifies a real number for each assignment (or *outcome*):

$$P(X_1 = x_1, X_2 = x_2, \ldots X_n = x_n)$$

$$P(x_1, x_2, \ldots x_n)$$

  - Must obey: $P(x_1, x_2, \ldots x_n) \geq 0$

$$\sum_{(x_1, x_2, \ldots x_n)} P(x_1, x_2, \ldots x_n) = 1$$

$P(T, W)$

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

- Size of distribution for n variables with domain size d? d^n
  - For all but the smallest distributions, cannot write out by hand!

Probabilistic Models

- A probabilistic model is a joint distribution over a set of random variables

- Probabilistic models:
  - (Random) variables with domains
  - Joint distributions: say whether assignments (outcomes) are likely
  - Ideally: only certain variables directly interact

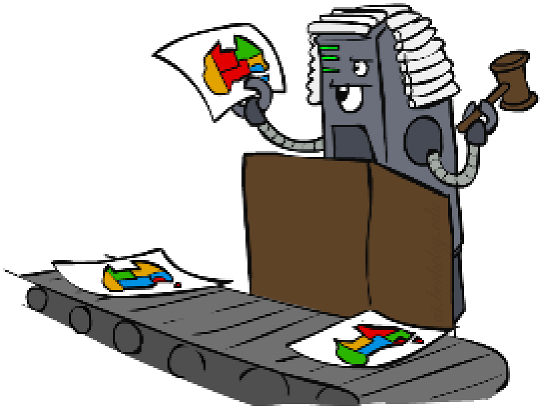
- Constraint satisfaction problems:
  - Variables with domains
  - Constraints: state whether assignments are possible
  - Ideally: only certain variables directly interact

Distribution over T,W

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

Constraint over T,W

T	W	P
hot	sun	T
hot	rain	F
cold	sun	F
cold	rain	T



Probabilities of events

- An *event* is a set E of outcomes

$$P(E) = \sum_{(x_1 \ldots x_n) \in E} P(x_1 \ldots x_n)$$

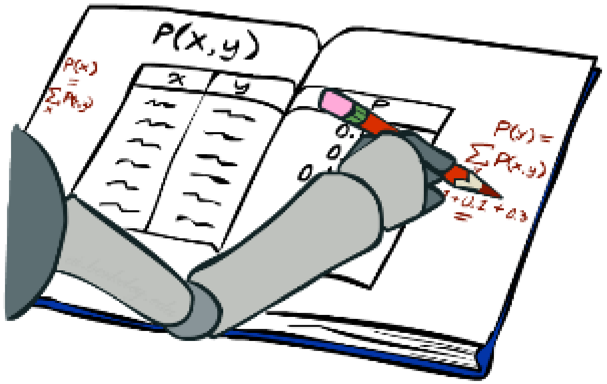
- Given a joint distribution over all variables, we can compute any event probability!
  - Probability that it's hot AND sunny?
  - Probability that it's hot?
  - Probability that it's hot OR sunny?

$P(T, W)$

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

Marginal Distributions

- Marginal distributions are sub-tables which eliminate variables
- Marginalization (summing out): Combine collapsed rows by adding



$P(T,W)$

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

$$P(t) = \sum_w P(t,w)$$

$P(T)$

T	P
hot	0.5
cold	0.5

$$P(w) = \sum_t P(t,w)$$

$P(W)$

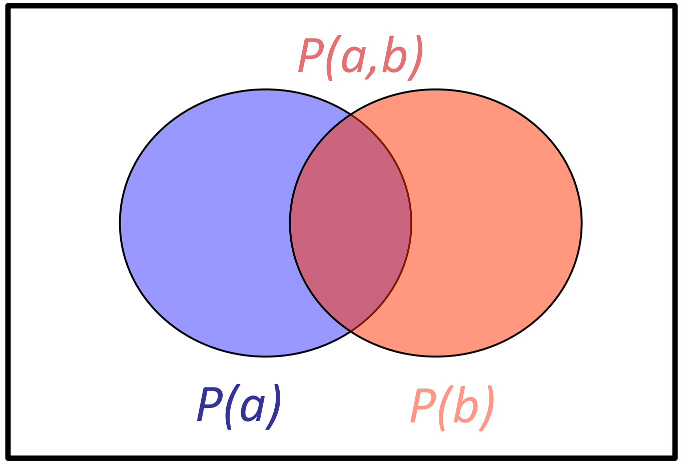
W	P
sun	0.6
rain	0.4

$$P(X_1 = x_1) = \sum_{x_2} P(X_1 = x_1, X_2 = x_2)$$

Conditional Probabilities

- The probability of an event given that another event has occurred

$$P(a|b) = \frac{P(a,b)}{P(b)}$$



$$P(T,W)$$

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

$$P(W=s|T=c) = \frac{P(W=s,T=c)}{P(T=c)} = 0.4$$

$$= P(W=s,T=c) + P(W=r,T=c)$$

$$= 0.2 + 0.3 \;\; = 0.5$$

Conditional Distributions

- Conditional distributions are probability distributions over some variables given fixed values of others

Joint Distribution

$P(T, W)$

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

Conditional Distributions

$P(W|T)$

$P(W|T = hot)$

W	P
sun	0.8
rain	0.2

$P(W|T = cold)$

W	P
sun	0.4
rain	0.6

Normalization Trick

$$P(T, W)$$

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

$$\begin{aligned}
P(W = s|T = c) &= \frac{P(W = s, T = c)}{P(T = c)} \\
&= \frac{P(W = s, T = c)}{P(W = s, T = c) + P(W = r, T = c)} \\
&= \frac{0.2}{0.2 + 0.3} = 0.4
\end{aligned}$$

$$\begin{aligned}
P(W = r|T = c) &= \frac{P(W = r, T = c)}{P(T = c)} \\
&= \frac{P(W = r, T = c)}{P(W = s, T = c) + P(W = r, T = c)} \\
&= \frac{0.3}{0.2 + 0.3} = 0.6
\end{aligned}$$

$$P(W|T = c)$$

W	P
sun	0.4
rain	0.6

Normalization Trick

$$\begin{aligned} P(W=s|T=c) &= \frac{P(W=s,T=c)}{P(T=c)} \\ &= \frac{P(W=s,T=c)}{P(W=s,T=c)+P(W=r,T=c)} \\ &= \frac{0.2}{0.2+0.3} = 0.4 \end{aligned}$$

$P(T,W)$

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

SELECT the joint probabilities matching the evidence

$P(c,W)$

T	W	P
cold	sun	0.2
cold	rain	0.3

NORMALIZE the selection (make it sum to one)

$P(W|T=c)$

W	P
sun	0.4
rain	0.6

$$\begin{aligned} P(W=r|T=c) &= \frac{P(W=r,T=c)}{P(T=c)} \\ &= \frac{P(W=r,T=c)}{P(W=s,T=c)+P(W=r,T=c)} \\ &= \frac{0.3}{0.2+0.3} = 0.6 \end{aligned}$$

Normalization Trick

$P(T, W)$

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

SELECT the joint probabilities matching the evidence

$P(c, W)$

T	W	P
cold	sun	0.2
cold	rain	0.3

NORMALIZE the selection (make it sum to one)

$P(W|T = c)$

W	P
sun	0.4
rain	0.6

- Why does this work? Sum of selection is P(evidence)! (P(T=c), here)

$$P(x_1|x_2) = \frac{P(x_1, x_2)}{P(x_2)} = \frac{P(x_1, x_2)}{\sum_{x_1} P(x_1, x_2)}$$

Probabilistic Inference

- Probabilistic inference
  - compute a desired probability from other known probabilities (e.g. conditional from joint)

- We generally compute conditional probabilities
  - These represent the agent’s beliefs given the evidence
  - P(on time | no reported accidents) = 0.90

Inference by Enumeration

- General case:
  - Evidence variables: $E_1 \ldots E_k = e_1 \ldots e_k$
  - Query variable: Q
  - Hidden variables: $H_1 \ldots H_r$

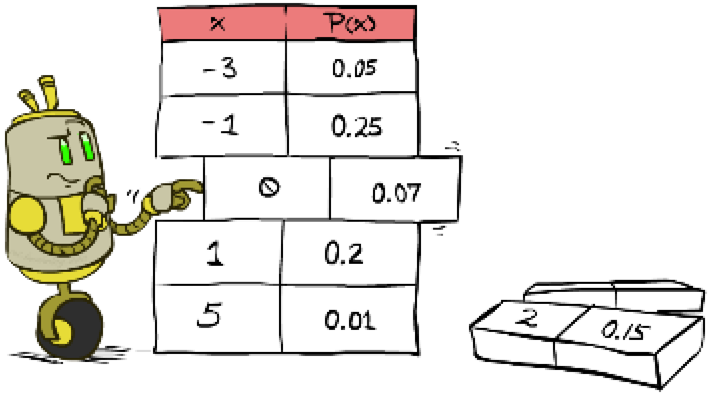
$$\left.\begin{array}{l} E_1 \ldots E_k = e_1 \ldots e_k \\ Q \\ H_1 \ldots H_r \end{array}\right\} X_1, X_2, \ldots X_n$$

All variables

- We want:

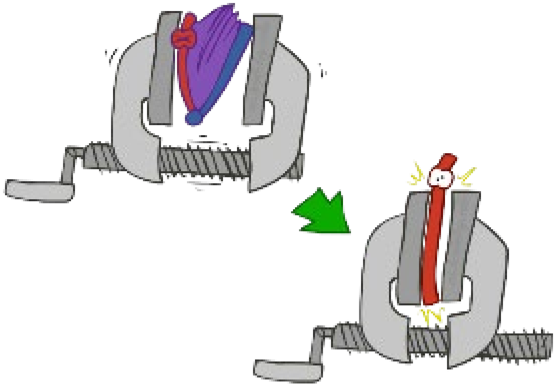
$$P(Q|e_1 \ldots e_k)$$

- Step 1: Select the entries consistent with the evidence



x	P(x)
-3	0.05
-1	0.25
⊘	0.07
1	0.2
5	0.01
2	0.15

- Step 2: Sum out H to get joint of Query and evidence



$$P(Q, e_1 \ldots e_k) = \sum_{h_1 \ldots h_r} \underbrace{P(Q, h_1 \ldots h_r, e_1 \ldots e_k)}_{X_1, X_2, \ldots X_n}$$

- Step 3: Normalize

$$\times \frac{1}{Z}$$

$$Z = \sum_q P(Q, e_1 \cdots e_k)$$

$$P(Q|e_1 \cdots e_k) = \frac{1}{Z} P(Q, e_1 \cdots e_k)$$

Inference by Enumeration

1. Select the entries consistent with the evidence
2. Sum out H to get joint of Query and evidence
3. Normalize

- P(W | winter)? sun: 0.5, rain: 0.5
- P(W | winter, hot)? sun: 0.67, rain: 0.33

S	T	W	P
summer	hot	sun	0.30
summer	hot	rain	0.05
summer	cold	sun	0.10
summer	cold	rain	0.05
winter	hot	sun	0.10
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.20

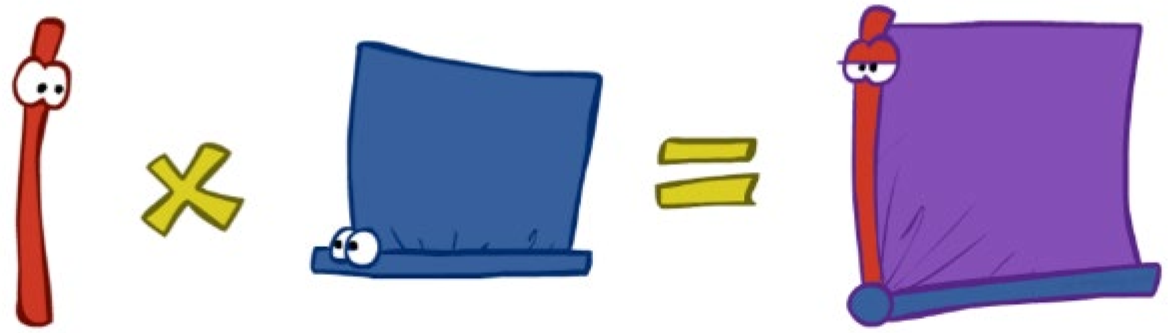
Inference by Enumeration

- Obvious problems:
  - Worst-case time complexity $O(d^n)$
  - Space complexity $O(d^n)$ to store the joint distribution

The Product Rule

- Sometimes have conditional distributions but want the joint

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



The Product Rule

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

- Example:

$P(W)$

W	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

$P(D, W)$

D	W	P
wet	sun	
dry	sun	
wet	rain	
dry	rain	

The Chain Rule

- More generally, can always write any joint distribution as an incremental product of conditional distributions

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, x_2, \dots x_n) = \prod_i P(x_i|x_1 \dots x_{i-1})$$

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

- In the running for most important AI equation!

Inference with Bayes' Rule

- Example: Diagnostic probability from causal probability:

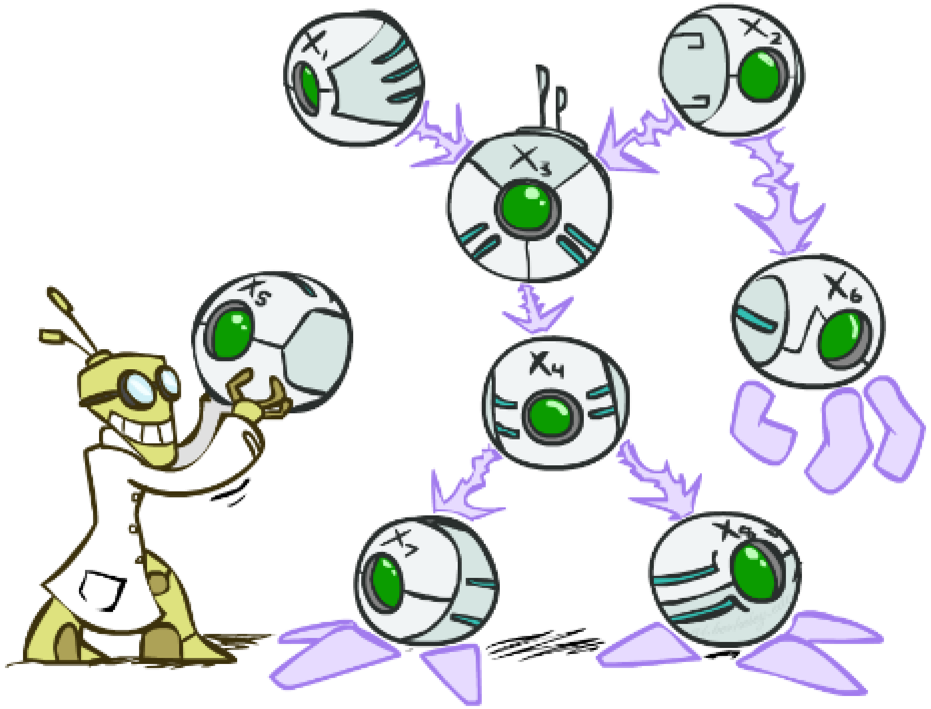
$$P(\text{cause}|\text{effect}) = \frac{P(\text{effect}|\text{cause})P(\text{cause})}{P(\text{effect})}$$

- Example:
  - M: meningitis, S: stiff neck

$$\left.\begin{aligned} P(+m) &= 0.0001 \\ P(+s|+m) &= 0.8 \\ P(+s|-m) &= 0.01 \end{aligned}\right\} \text{Example givens}$$

$$P(+m|+s) = \frac{P(+s|+m)P(+m)}{P(+s)} = \frac{P(+s|+m)P(+m)}{P(+s|+m)P(+m) + P(+s|-m)P(-m)} = \frac{0.8 \times 0.0001}{0.8 \times 0.0001 + 0.01 \times 0.9999}$$

Bayesian Networks



AIMA Chapter 14.1, 14.2

[Adapted from slides by Dan Klein and Pieter Abbeel at UC Berkeley (ai.berkeley.edu)
and from slides by Markus Svensén & Christopher M Bishop]

Additional Reference

- [PRML] Pattern Recognition and Machine Learning, Christopher Bishop, Springer 2006.
    - Chapter 8.1 - 8.3

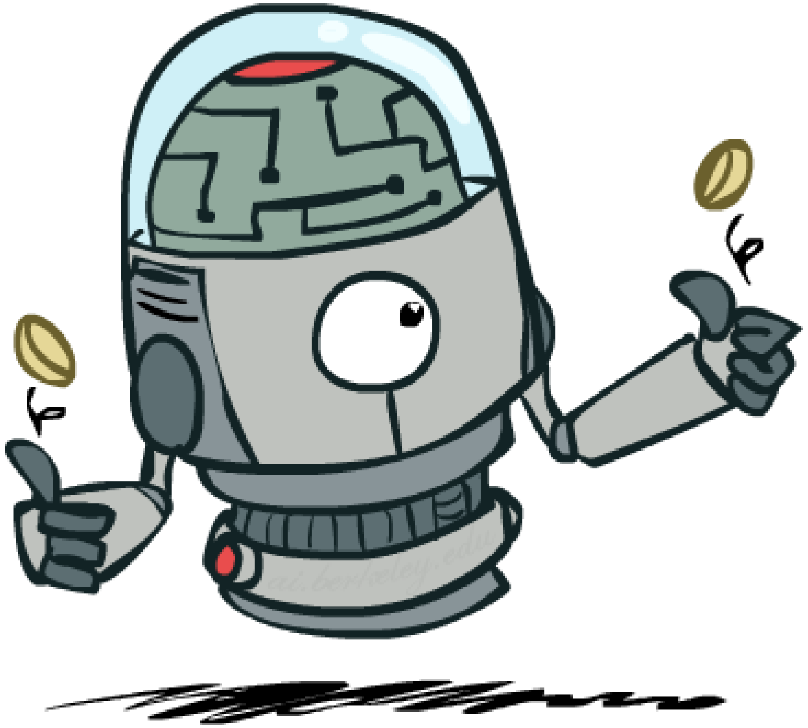
Probabilistic Models

- Models describe how (a portion of) the world works
  - Models are always simplifications
  - May omit some variables and interactions
  - "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: making decisions based on expected utility

- How do we build models, avoiding the d^n blowup?

Independence



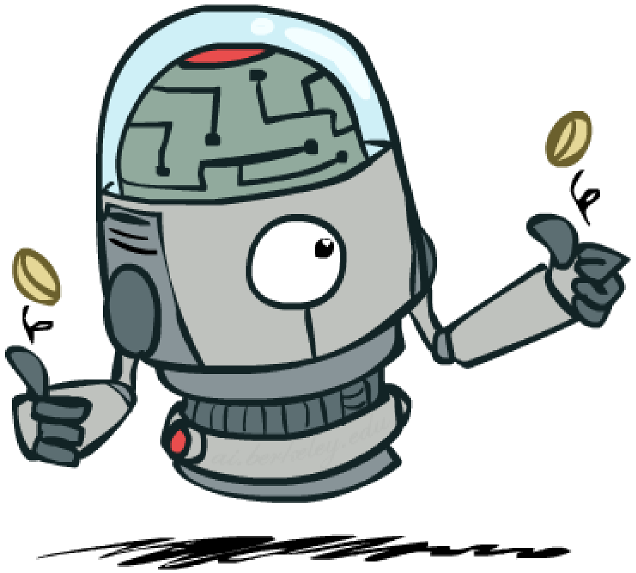
Independence

- Two variables X and Y are (absolutely) ***independent*** if

$$\forall x,y \qquad P(x,y) = P(x)P(y) \qquad\qquad X \perp\!\!\!\perp Y$$

  - This says that their joint distribution ***factors*** into a product of two simpler distributions
  - Combine with product rule $P(x,y) = P(x|y)P(y)$ we obtain another form:

$$\forall x,y\ P(x|y) = P(x) \quad \text{or} \quad \forall x,y\ P(y|x) = P(y)$$



- Example: two dice rolls $Roll_1$ and $Roll_2$
  - $P(Roll_1=5, Roll_2=5) = P(Roll_1=5)P(Roll_2=5) = 1/6 \times 1/6 = 1/36$
  - $P(Roll_2=5 \mid Roll_1=5) = P(Roll_2=5)$

Independence in the real world

- Independence is a simplifying *modeling assumption*
    - Sometimes it’s reasonable for real-world variables
    - What could we assume for {Weather, Temperature, Cavity, Toothache}?
        - Cavity and Toothache are ***not*** independent of each other
            - Ditto for hundreds of dentistry variables
        - Weather and Temperature are ***not*** independent of each other
            - Ditto for hundreds of meteorological variables
        - Cavity and Toothache are ***roughly*** independent of Weather and Temperature

Conditional Independence

- Unconditional (absolute) independence is rare
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.
- X is conditionally independent of Y given Z $X \perp\!\!\!\perp Y | Z$

  if and only if:

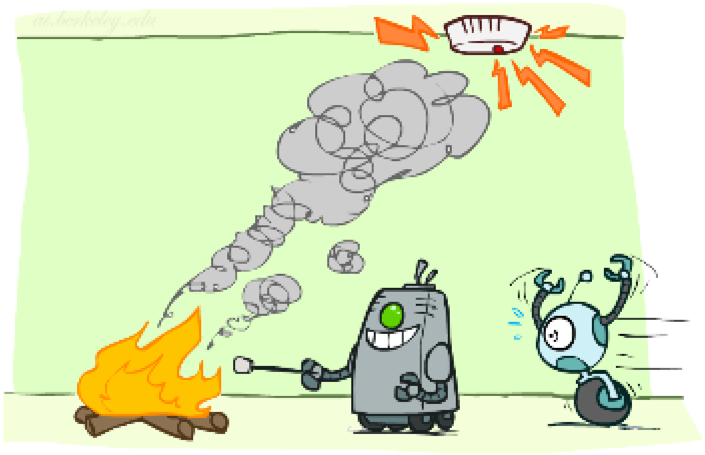
  $\forall x,y,z \qquad P(x \mid y,z) = P(x|z)$

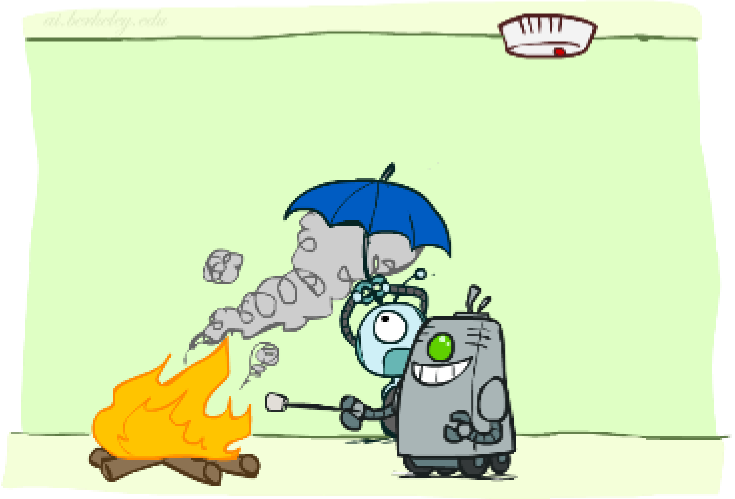
  or, equivalently, if and only if

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

Conditional Independence

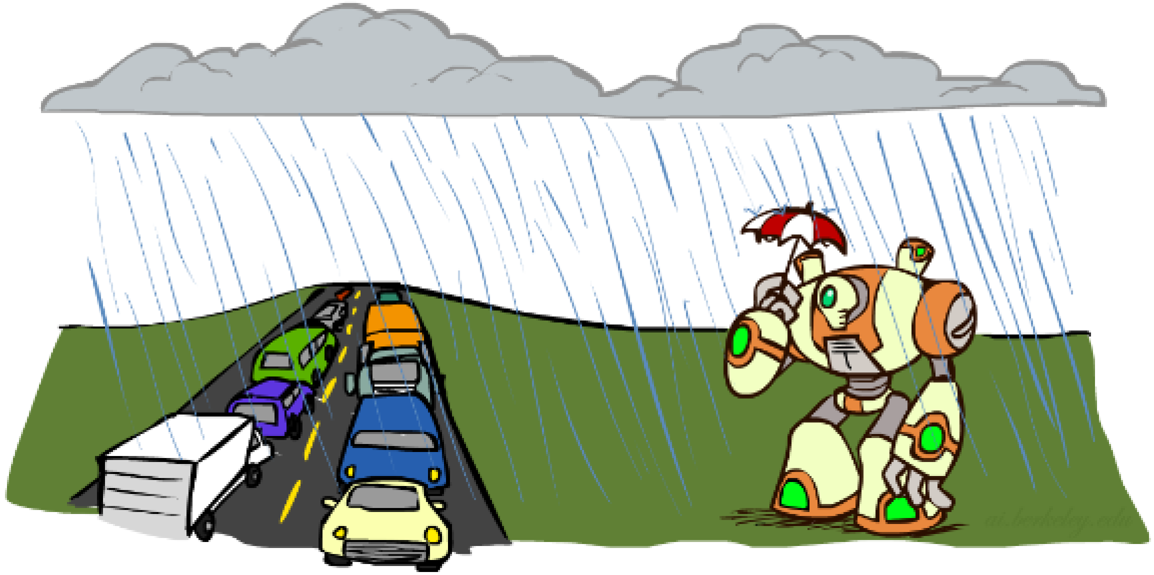
- What about this domain:
  - Fire
  - Smoke
  - Alarm (smoke detector)





Conditional Independence

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

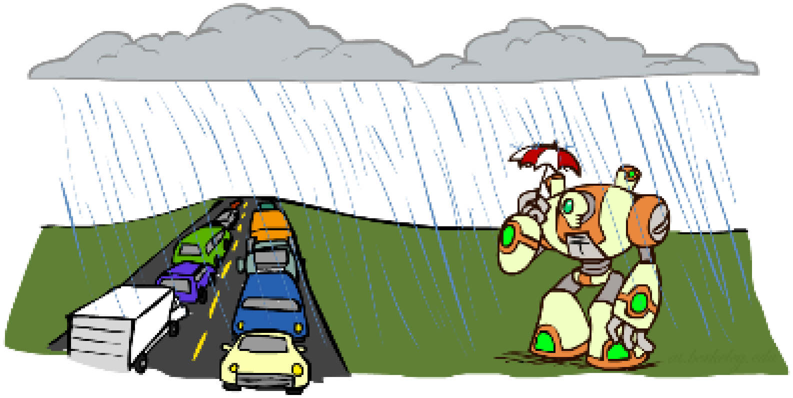


Conditional Independence and the Chain Rule

- 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:

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



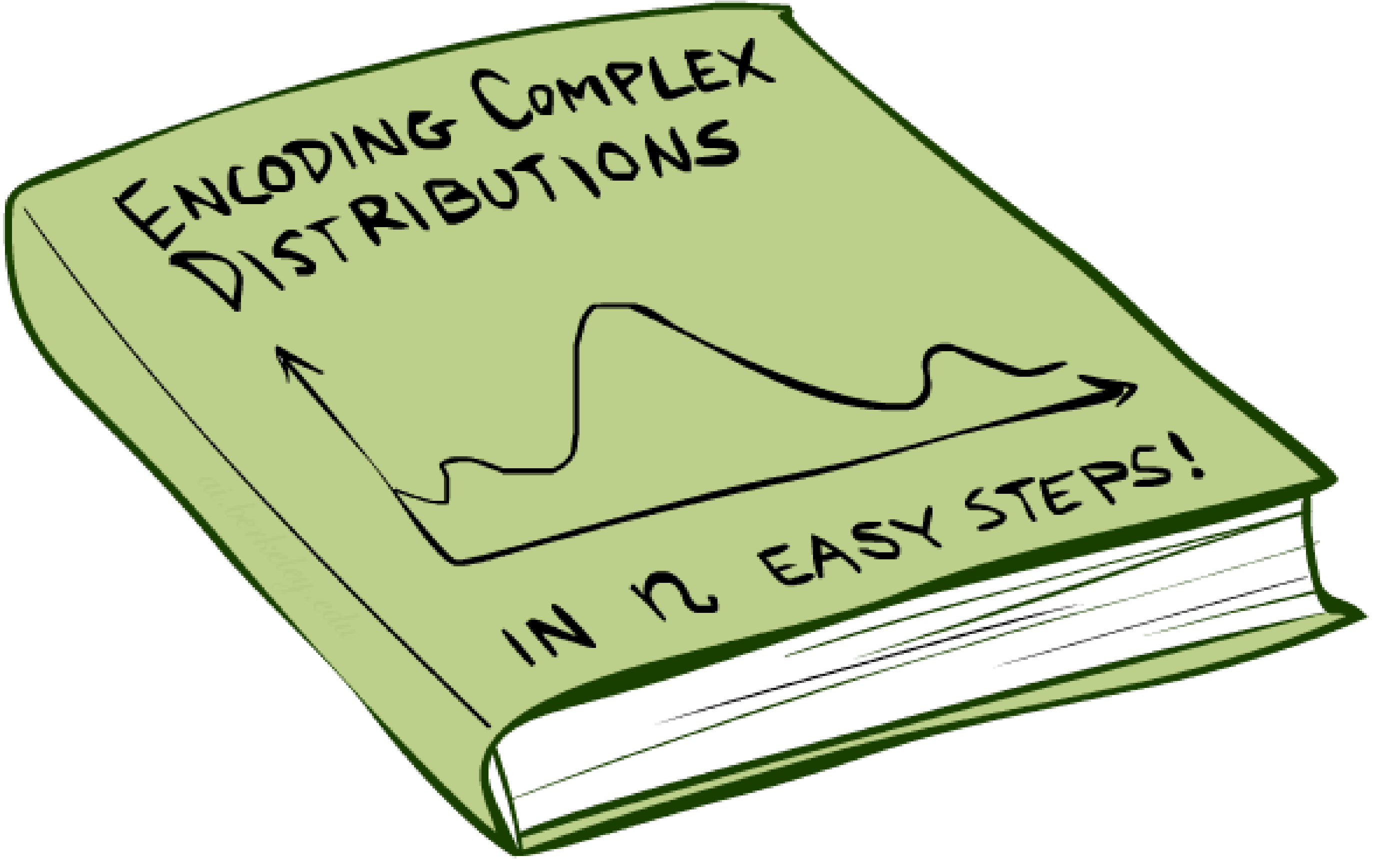
- With assumption of conditional independence:

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

Requires less space to encode!

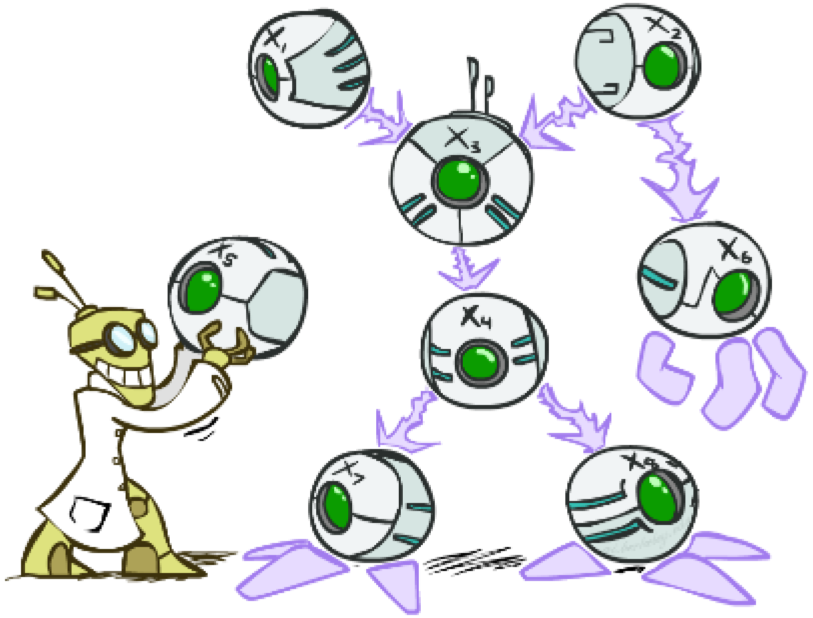
- BayesNets / graphical models help us express conditional independence assumptions

Bayesian Networks: Big Picture

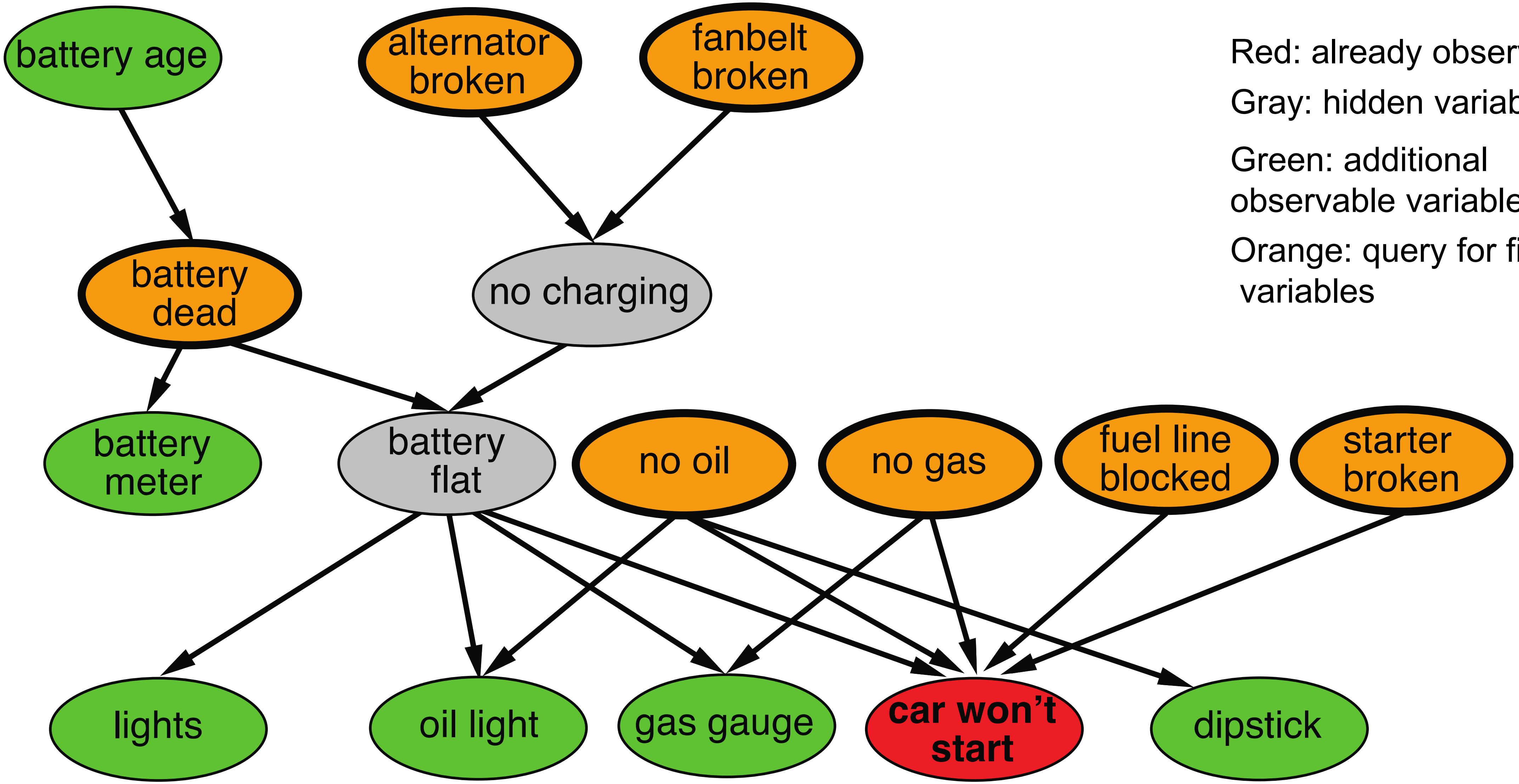


Bayesian Networks: Big Picture

- Full joint distribution tables answer every question, but:
  - Size is exponential in the number of variables
  - Need gazillions of examples to learn the probabilities
  - Inference by enumeration (summing out hiddens) is too slow
- Bayesian networks:
  - Express all the conditional independence relationships in a domain
  - Factor the joint distribution into a product of small conditionals
  - Often reduce size from exponential to linear
  - Faster learning from fewer examples
  - Faster inference (linear time in some important cases)



Example Bayes Net: Amateur Car Mechanic



Red: already observed

Gray: hidden variables

Green: additional observable variables

Orange: query for fixable variables

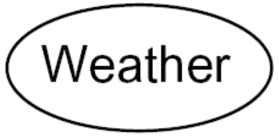
Bayesian Networks Syntax

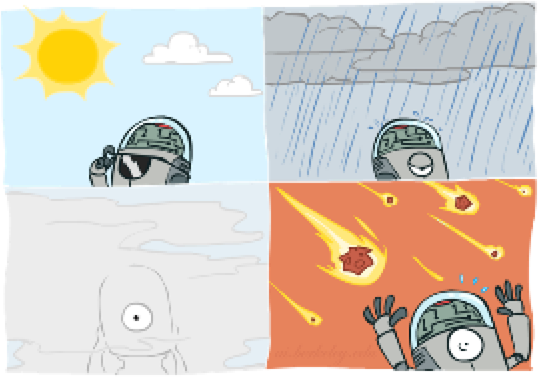
Bayesian Networks Syntax

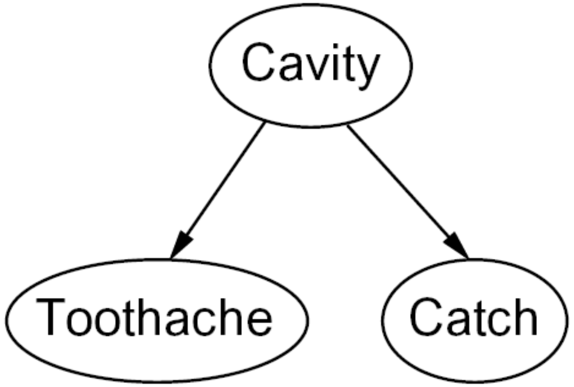
- Nodes: variables (with domains)

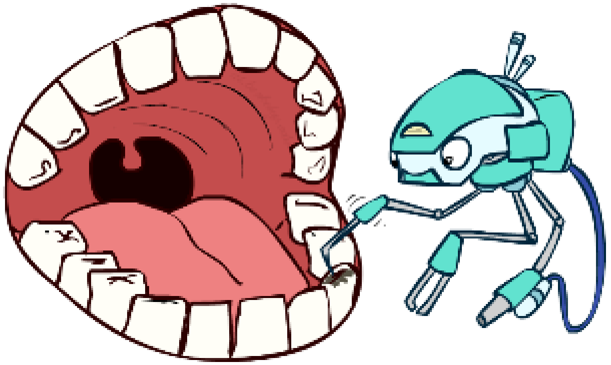
- Arcs: interactions
  - Indicate “direct influence” between variables
  - For now: imagine that arrows mean direct causation (in general, they may not!)
  - Formally: encode conditional independence (more later)

- No cycle is allowed!



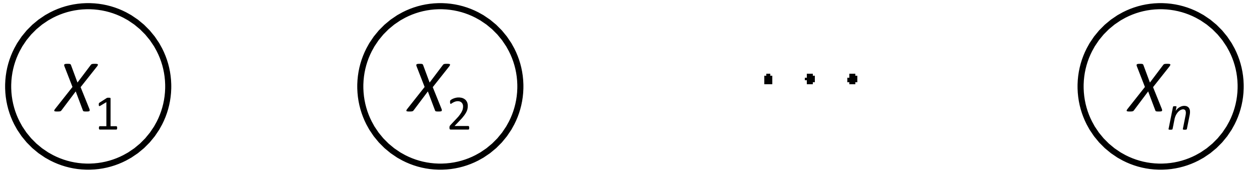




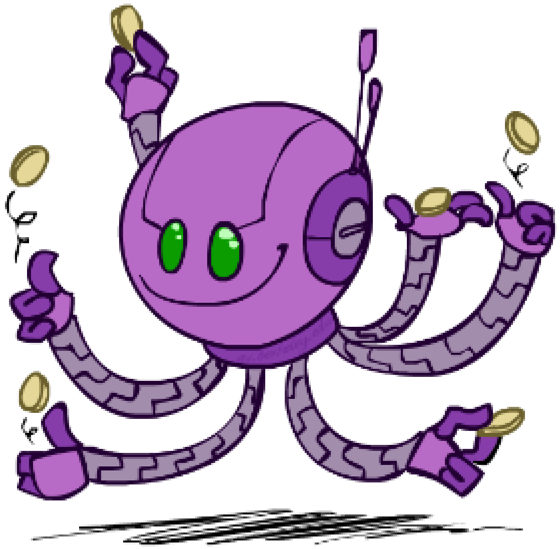


Example: Coin Flips

- N independent coin flips



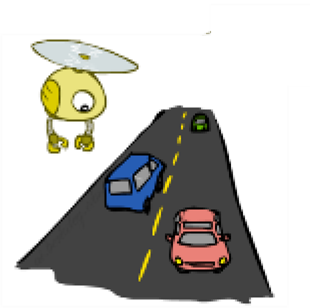
X_1 X_2 $\cdots$ X_n



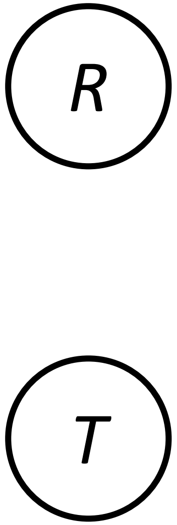
- No interactions between variables: absolute independence

Example: Traffic

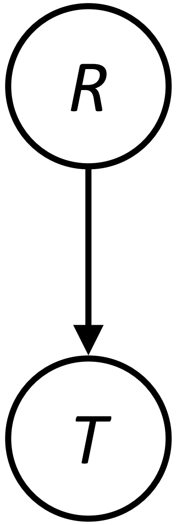
- Variables:
  - R: It rains
  - T: There is traffic



- Model 1: independence

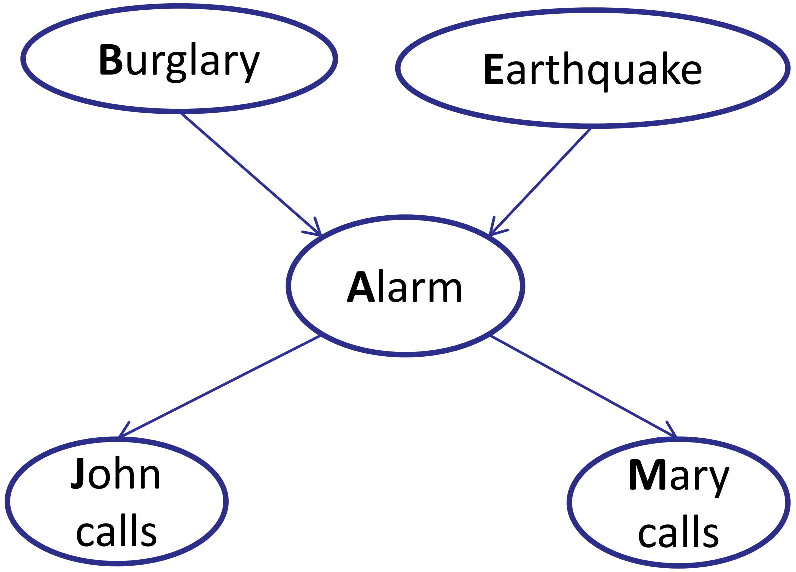


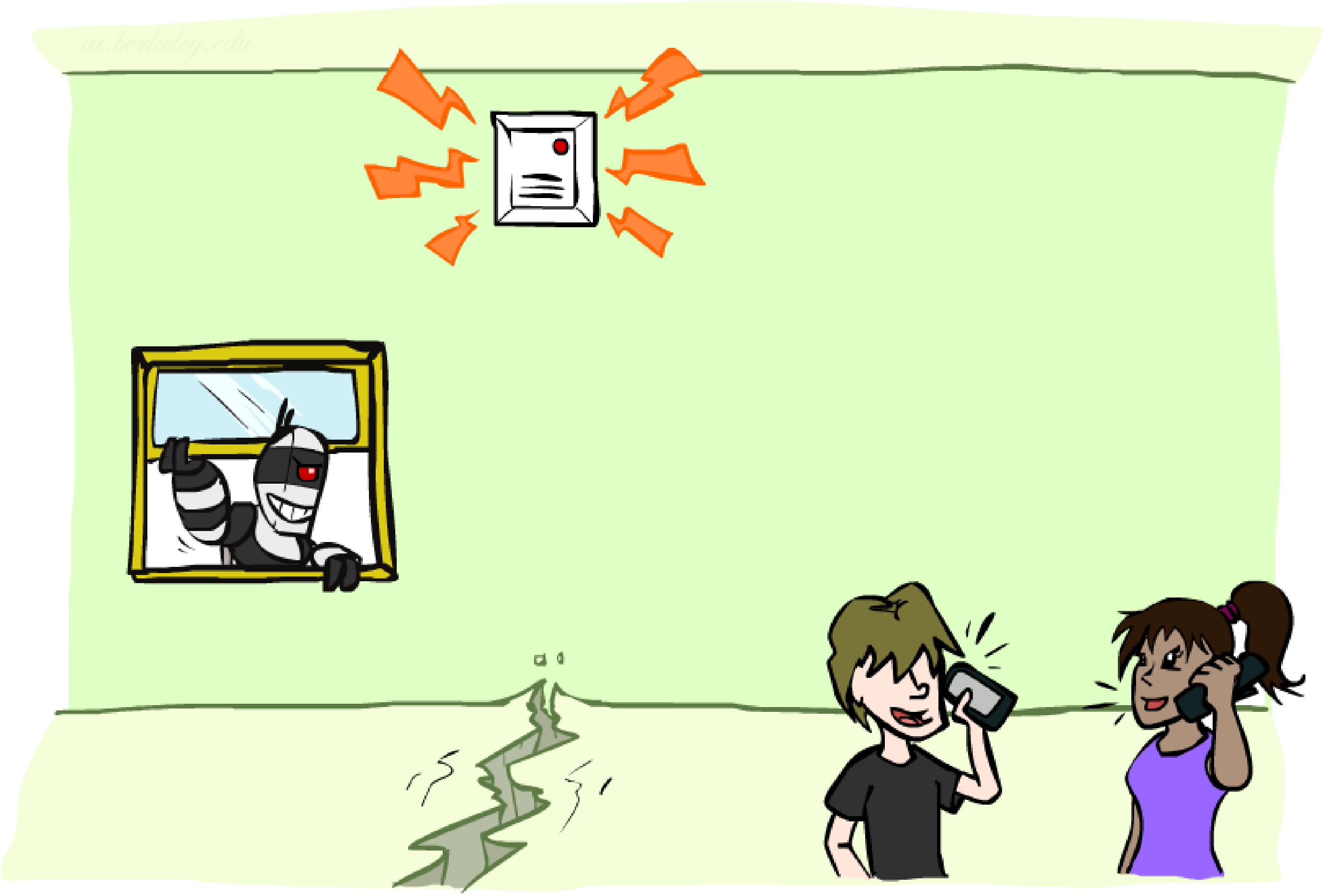
- Model 2: rain causes traffic



Example: Alarm Network

- Variables
  - B: Burglary
  - A: Alarm goes off
  - M: Mary calls
  - J: John calls
  - E: Earthquake!

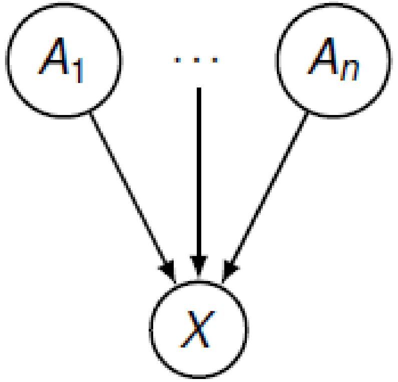




Bayesian Networks Syntax

- A directed, acyclic graph

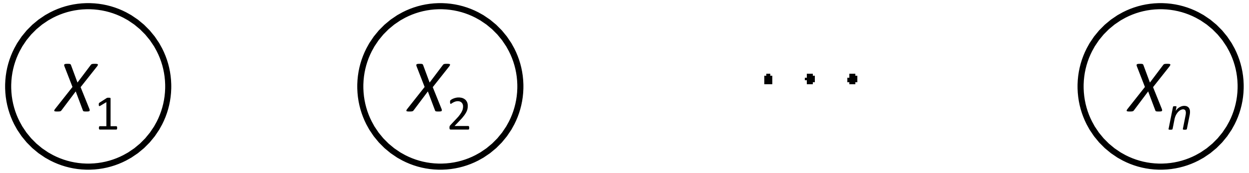
- Conditional distributions for each node given its ***parent variables*** in the graph
    - ***CPT***: conditional probability table: each row is a distribution for child given a configuration of its parents
    - Description of a noisy “causal” process



$$P(X|A_1, \cdots, A_n)$$

A Bayes net = Topology (graph) + Local Conditional Probabilities

Example: Coin Flips



X_1 X_2 . . . X_n

$P(X_1)$

h	0.5
t	0.5

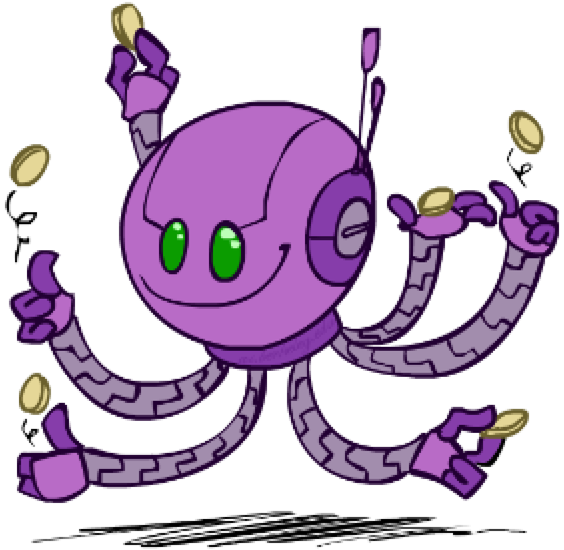
$P(X_2)$

h	0.5
t	0.5

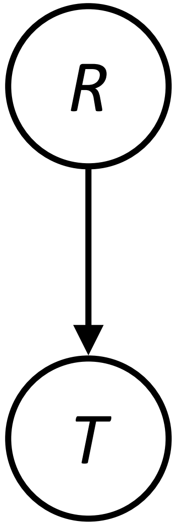
. . .

$P(X_n)$

h	0.5
t	0.5



Example: Traffic

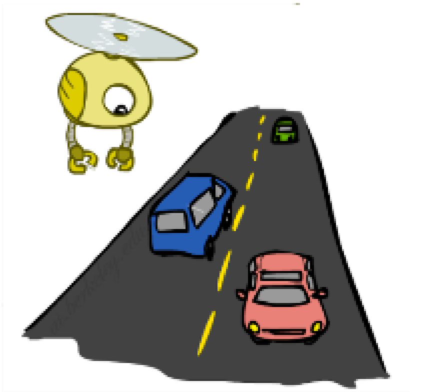


$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

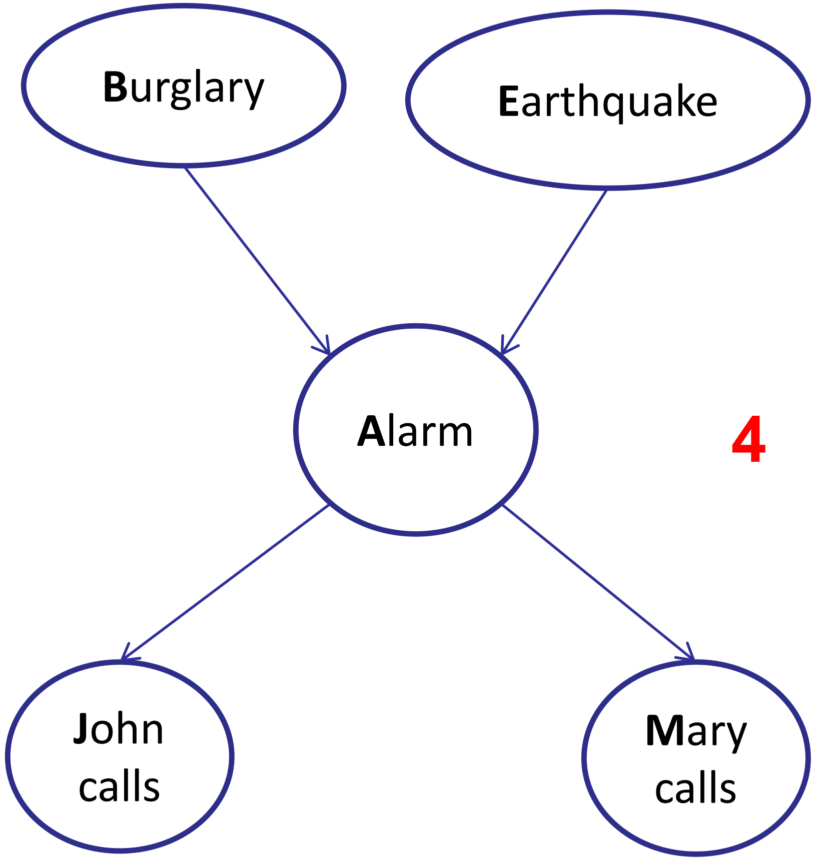
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2



Example: Alarm Network

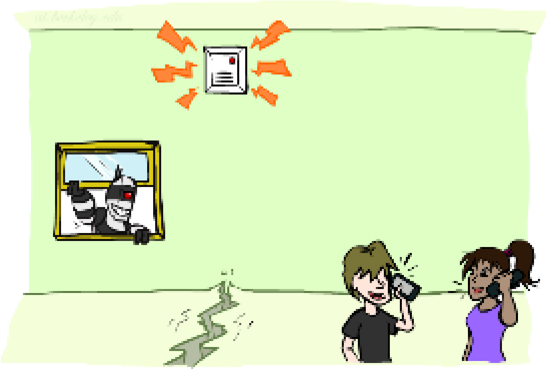
P(B)	
true	false
0.001	0.999

1



1

P(E)	
true	false
0.002	0.998



4

B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

Number of free parameters in each CPT:

- Parent domain sizes $d_1,\ldots,d_k$
- Child domain size d
- Each table row must sum to 1

$(d-1)\ \Pi_i\ d_i$

A	P(J|A)	
	true	false
true	0.9	0.1
false	0.05	0.95

2

2

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

General formula for sparse BNs

- Suppose
  - n variables
  - Maximum domain size is d
  - Maximum number of parents is k
- Full joint distribution has size $O(d^n)$
- Bayes net has size $O(n \cdot d^{k+1})$
  - Linear scaling with n as long as causal structure is local