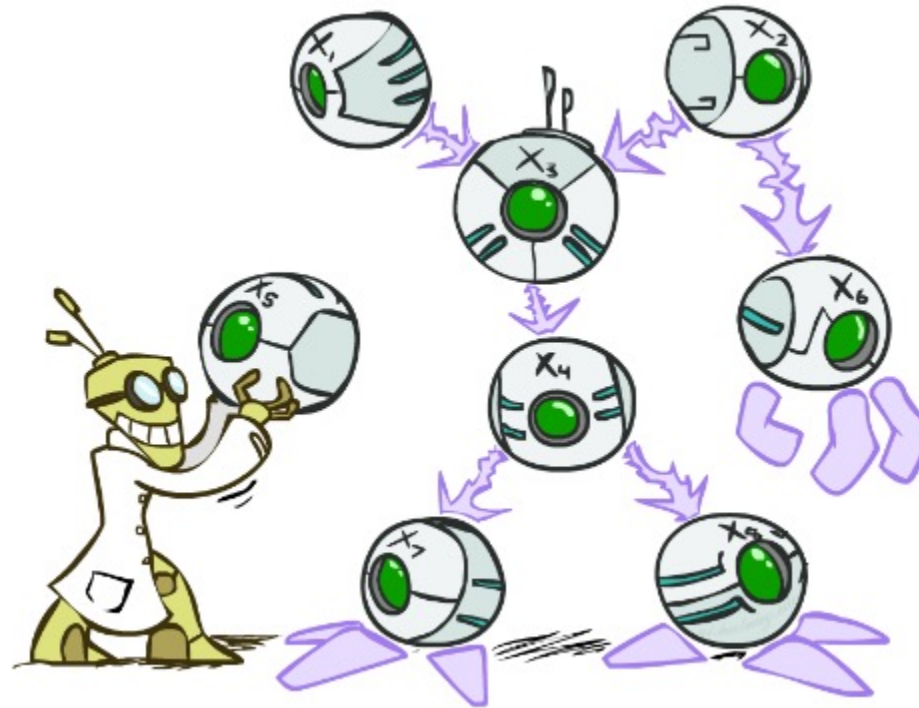


# CS 188: Artificial Intelligence

## Bayes' Nets



FA23 announcement:  
Midterm logistics form is  
on the website! Please  
fill it out ASAP if you  
need an alternate-time  
or remote exam.

Fall 2023

# Review: Probabilistic Inference

- Probabilistic inference: compute a desired probability from other known probabilities (e.g. conditional from joint)
- We generally compute conditional probabilities
  - $P(\text{on time} \mid \text{no reported accidents}) = 0.90$
  - These represent the agent's *beliefs* given the evidence
- Probabilities change with new evidence:
  - $P(\text{on time} \mid \text{no accidents, 5 a.m.}) = 0.95$
  - $P(\text{on time} \mid \text{no accidents, 5 a.m., raining}) = 0.80$
  - Observing new evidence causes *beliefs to be updated*



# Review: Inference by Enumeration

- General case:

- Evidence variables:  $E_1 \dots E_k = e_1 \dots e_k$
  - Query\* variable:  $Q$
  - Hidden variables:  $H_1 \dots H_r$
- $$\left. \begin{array}{l} E_1 \dots E_k = e_1 \dots e_k \\ Q \\ H_1 \dots H_r \end{array} \right\} \begin{array}{l} X_1, X_2, \dots X_n \\ \text{All variables} \end{array}$$

- We want:

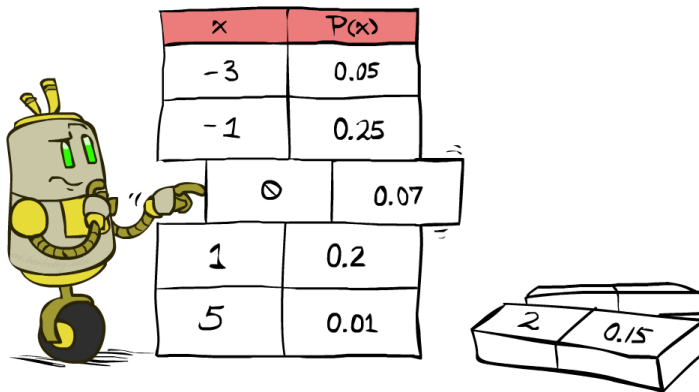
*\* Works fine with multiple query variables, too*

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

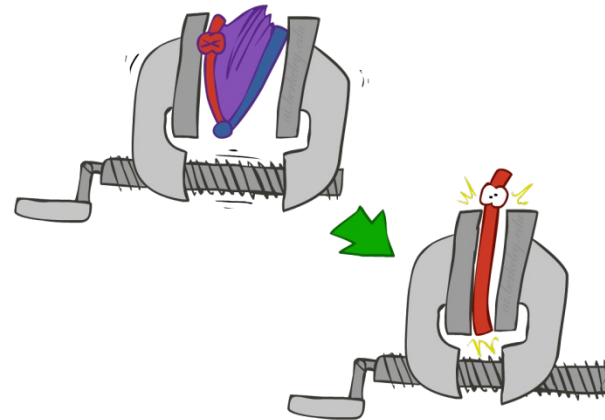
- Step 1: Select the entries consistent with the evidence

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

- Step 3: Normalize



x	P(x)
-3	0.05
-1	0.25
0	0.07
1	0.2
5	0.01



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

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

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

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

# Review: Inference by Enumeration

- $P(W)$ ?
- $P(W \mid \text{winter})$ ?
- $P(W \mid \text{winter, hot})$ ?

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

# Review: Inference by Enumeration

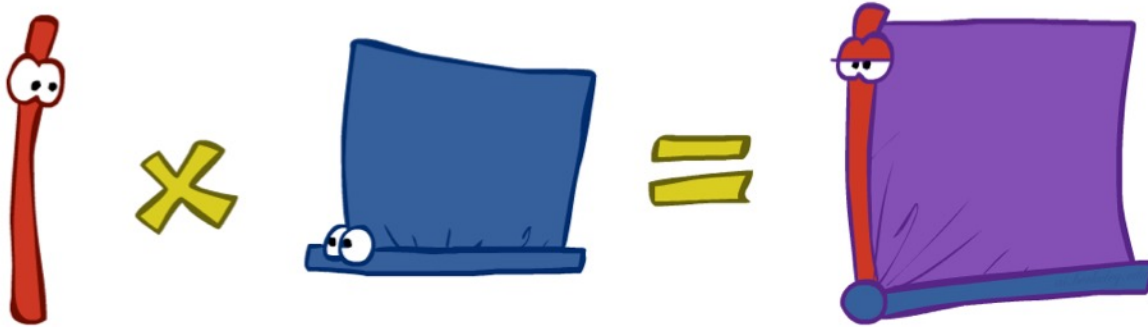
---

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

# Review: The Product Rule

- Sometimes have conditional distributions but want the joint

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



# Review: The Product Rule

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

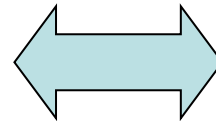
- Example:

$P(W)$

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



$P(D, W)$

D	W	P
wet	sun	0.08
dry	sun	0.72
wet	rain	0.14
dry	rain	0.06

# Review: The Chain Rule

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

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

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

- Why is this always true?

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

product rule.

if  $x_1, x_2$  independent,

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



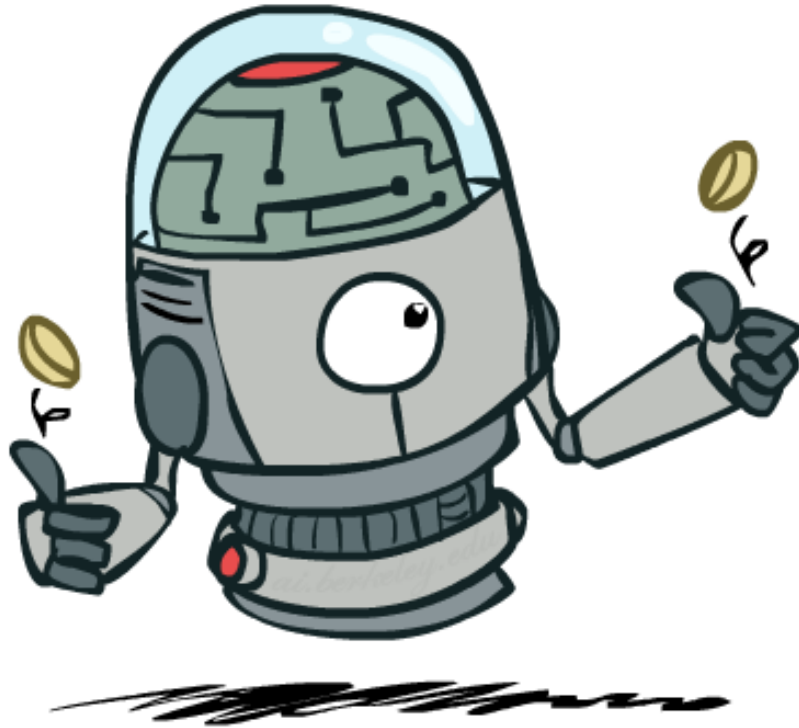
# Probabilistic Models

- 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:

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

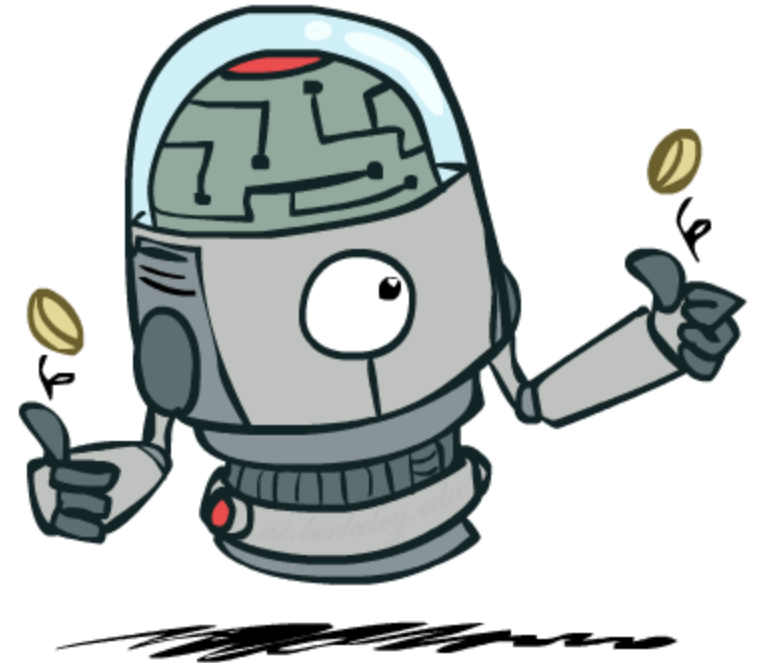
- This says that their joint distribution *factors* into a product two simpler distributions
- Another form:

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

- 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?

$P_1(T, W)$

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

$P(T)$

T	P
hot	0.5
cold	0.5

$P_2(T, W)$

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

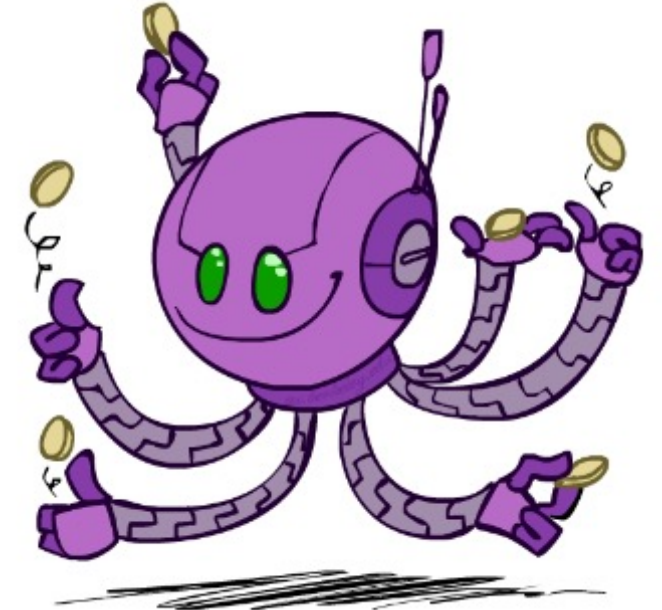
$P(X_2)$

H	0.5
T	0.5

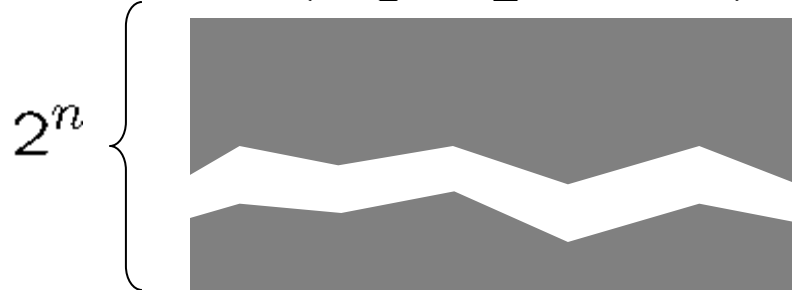
...

$P(X_n)$

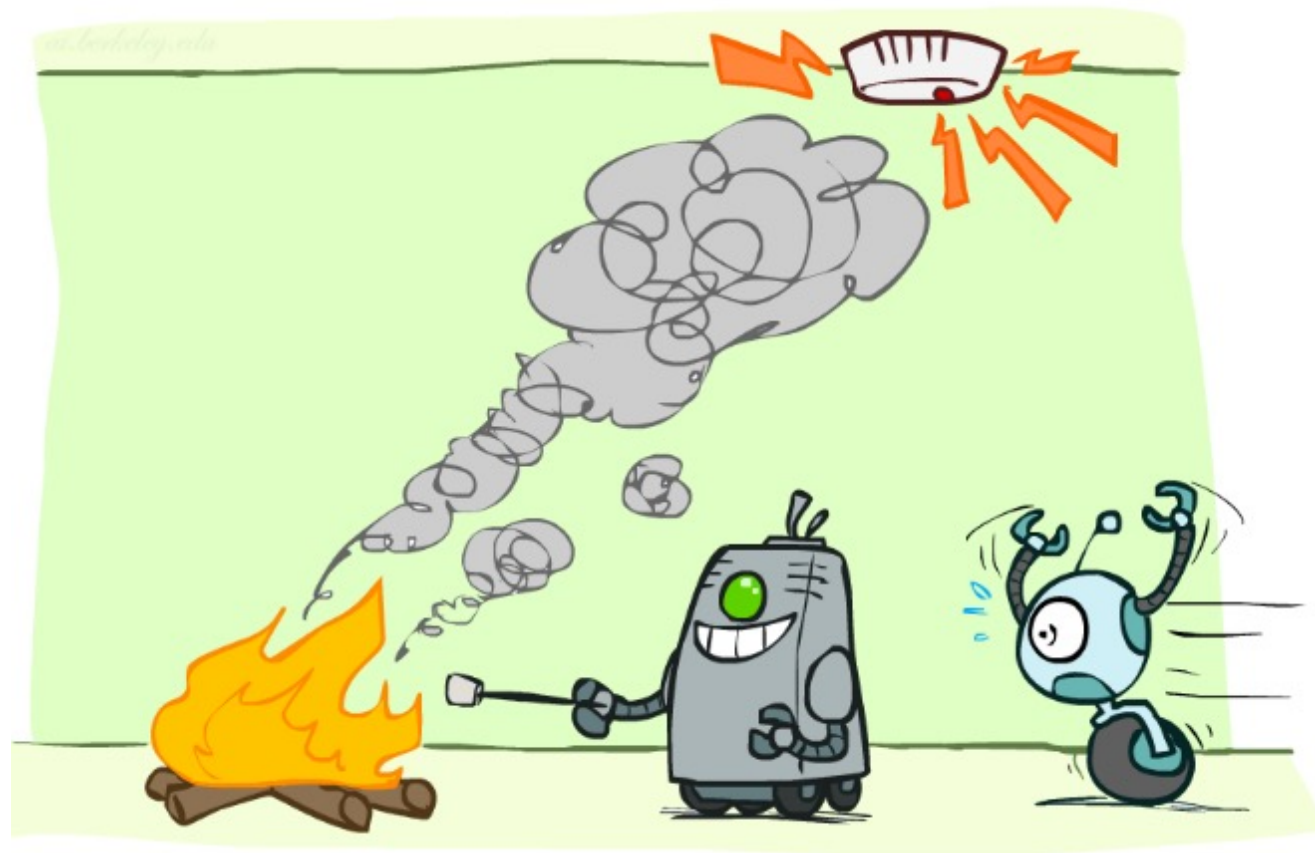
H	0.5
T	0.5



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

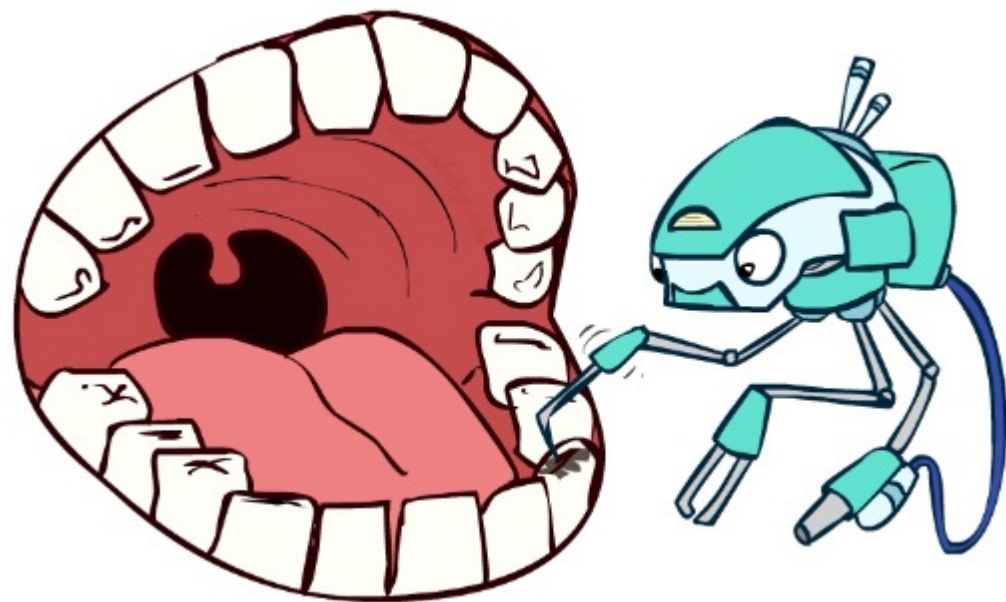


# Conditional Independence



# Conditional Independence

- $P(\text{Toothache}, \text{Cavity}, \text{Catch})$
- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - $P(+\text{catch} \mid +\text{toothache}, +\text{cavity}) = P(+\text{catch} \mid +\text{cavity})$
- The same independence holds if I don't have a cavity:
  - $P(+\text{catch} \mid +\text{toothache}, -\text{cavity}) = P(+\text{catch} \mid -\text{cavity})$
- Catch is *conditionally independent* of Toothache given Cavity:
  - $P(\text{Catch} \mid \text{Toothache}, \text{Cavity}) = P(\text{Catch} \mid \text{Cavity})$
- Equivalent statements:
  - $P(\text{Toothache} \mid \text{Catch}, \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity})$
  - $P(\text{Toothache}, \text{Catch} \mid \text{Cavity}) = P(\text{Toothache} \mid \text{Cavity}) P(\text{Catch} \mid \text{Cavity})$
  - One can be derived from the other easily



# 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

$$X \perp\!\!\!\perp Y | Z$$

if and only if:

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

or, equivalently, if and only if

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

given.



# Conditional Independence

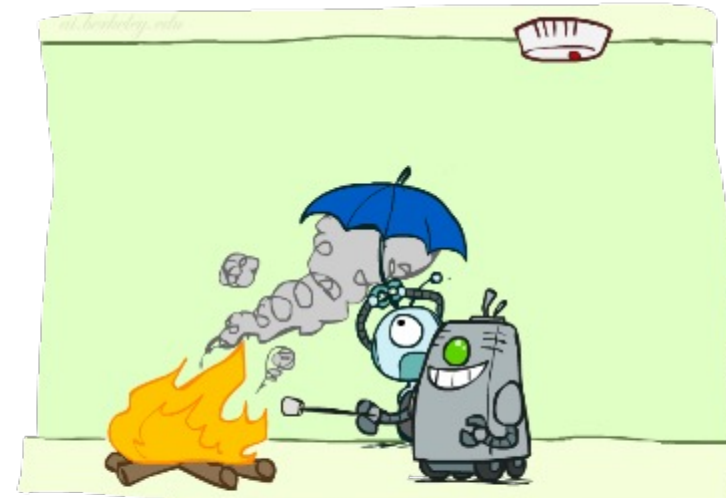
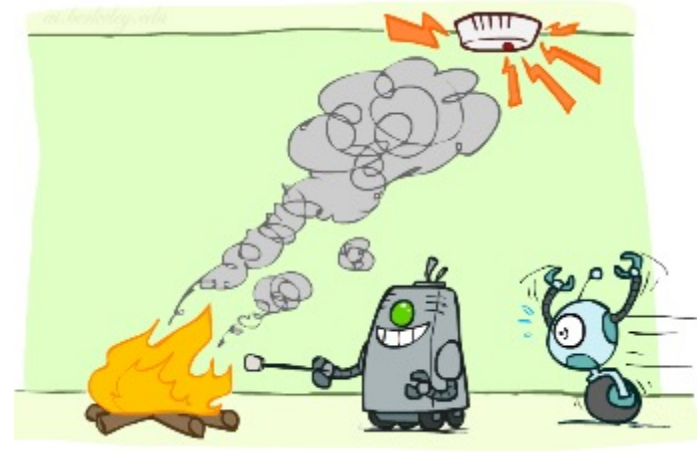
- What about this domain:
  - Traffic
  - Umbrella
  - Raining



# Conditional Independence

- What about this domain:

- Fire
- Smoke
- Alarm



# Conditional Independence and the Chain Rule

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

- Trivial decomposition:

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

$$\frac{P(V, R, T)}{P(R, T)}$$

- With assumption of conditional independence:

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



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

- 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

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

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

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

$$P(+t, +b, +g) = P(g) P(+t|g) P(+b|g)$$

- Givens:

$$P(+g) = 0.5$$

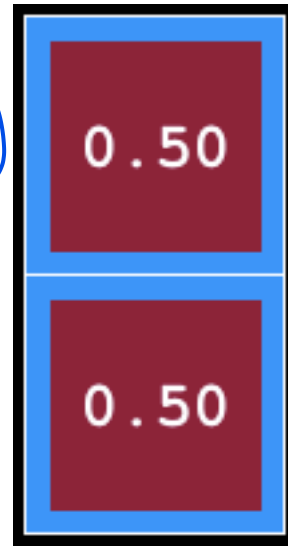
$$P(-g) = 0.5$$

$$P(+t | +g) = 0.8$$

$$P(+t | -g) = 0.4$$

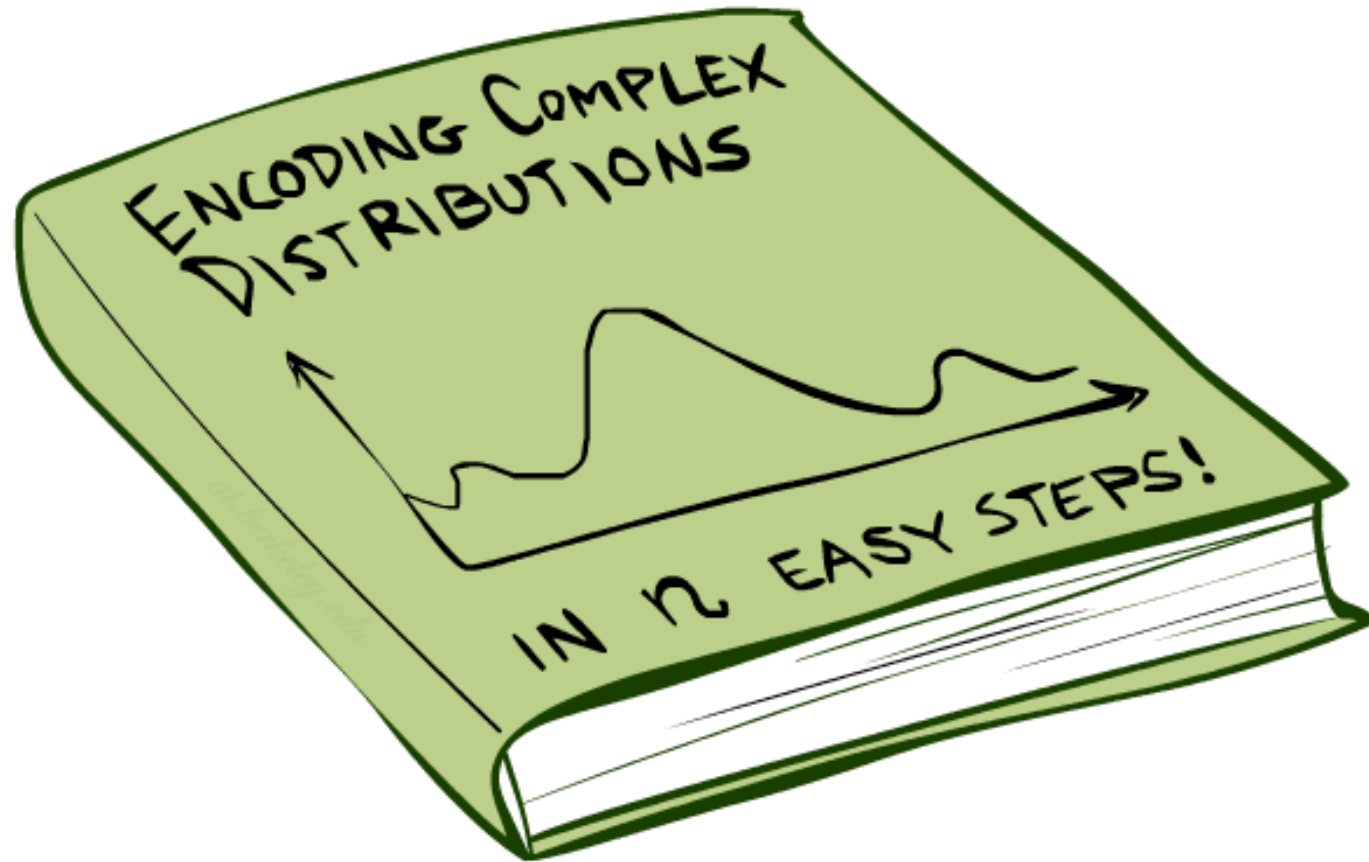
$$P(+b | +g) = 0.4$$

$$P(+b | -g) = 0.8$$



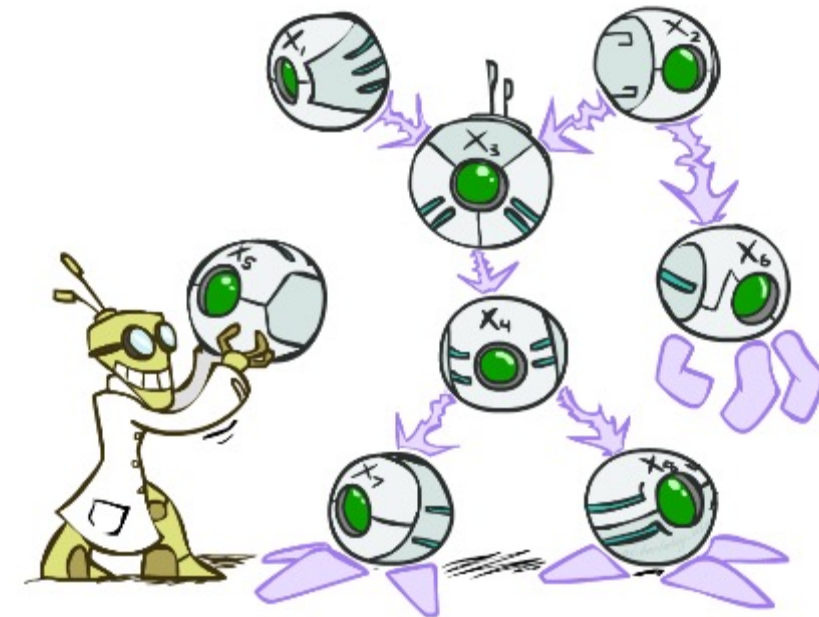
# Bayes' Nets: Big Picture

---



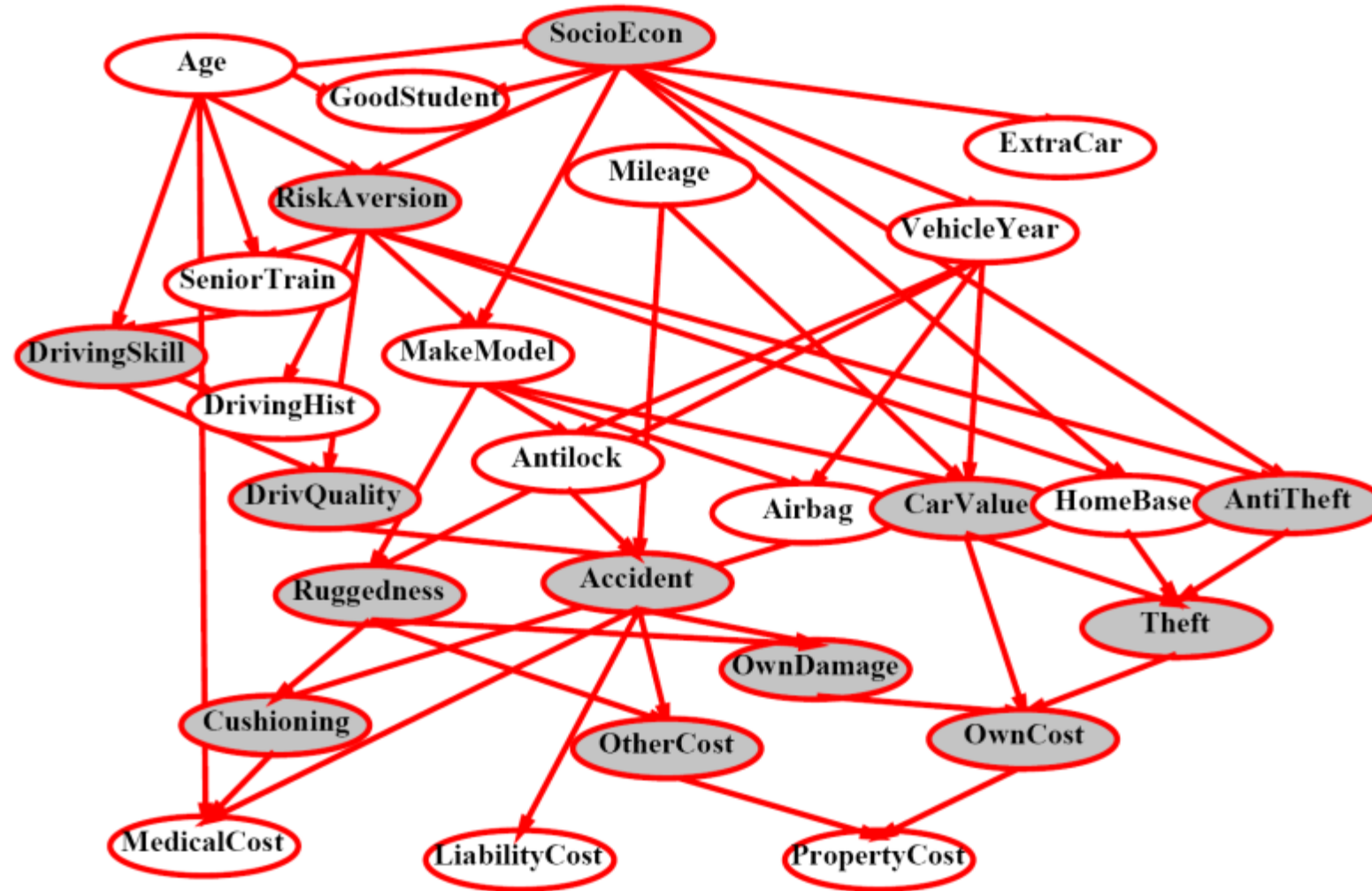
# Bayes' Nets: Big Picture

- Two problems with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time
- **Bayes' nets:** a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - More properly called **graphical models**
  - We describe how variables locally interact
  - Local interactions chain together to give global, indirect interactions
  - For about 10 min, we'll be vague about how these interactions are specified

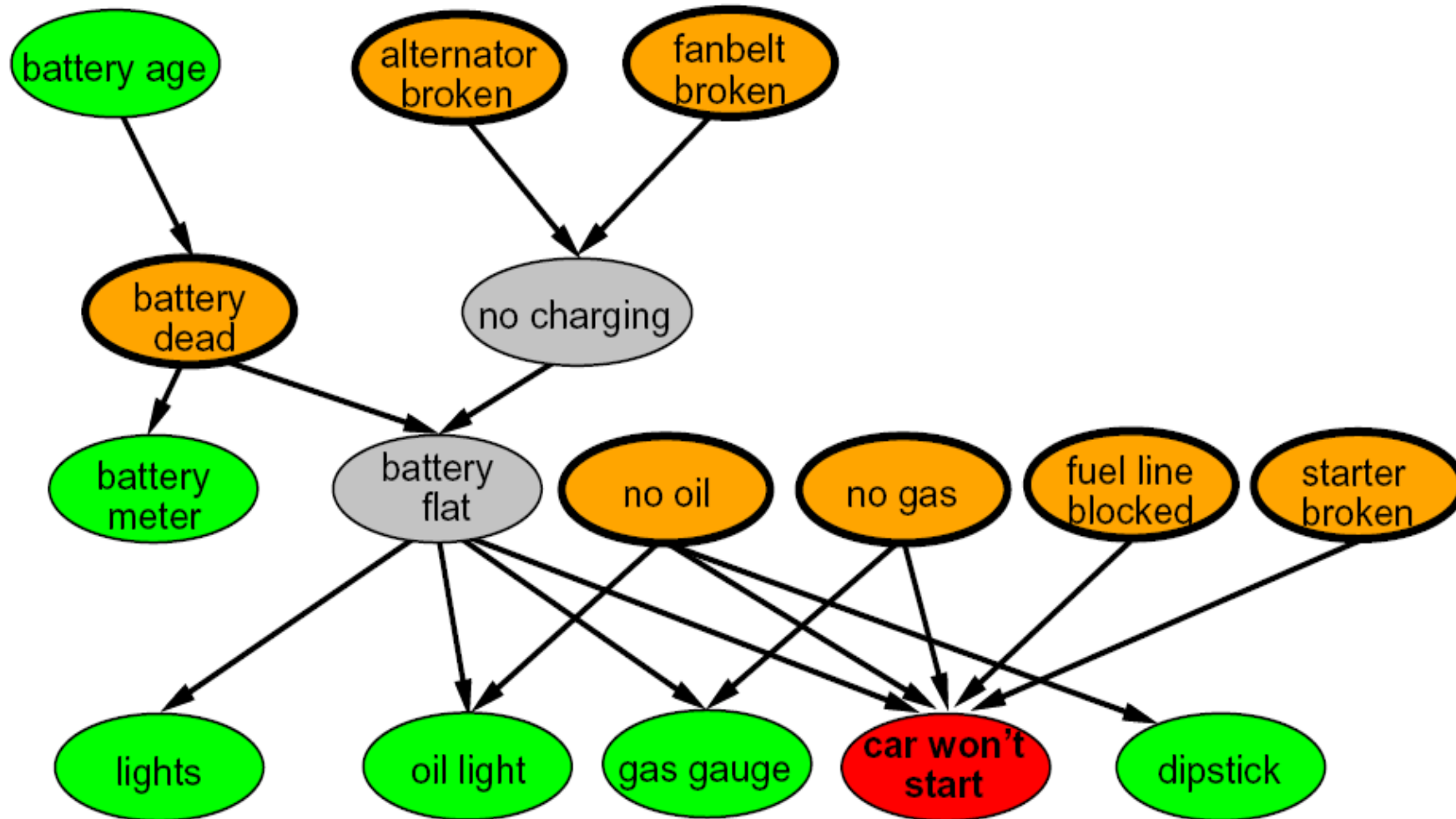




# Example Bayes' Net: Insurance



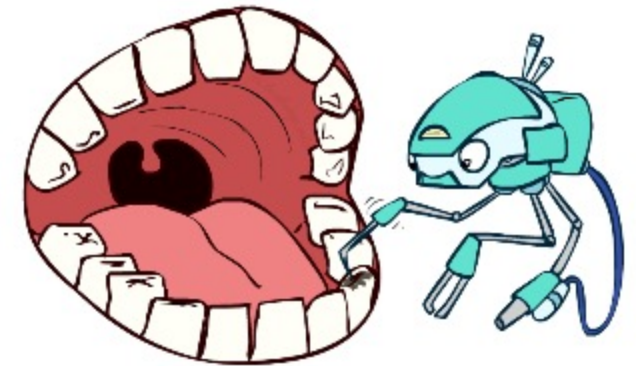
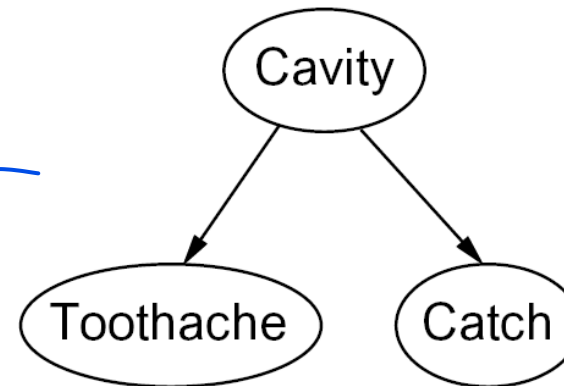
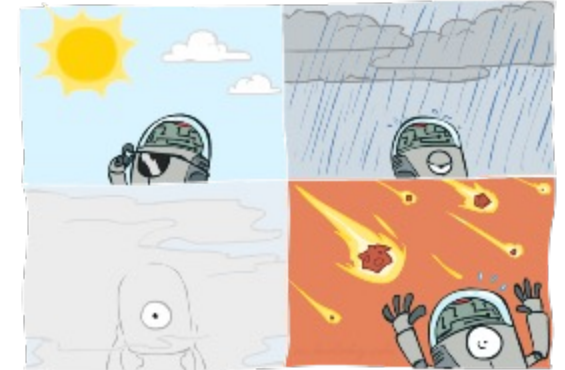
# Example Bayes' Net: Car





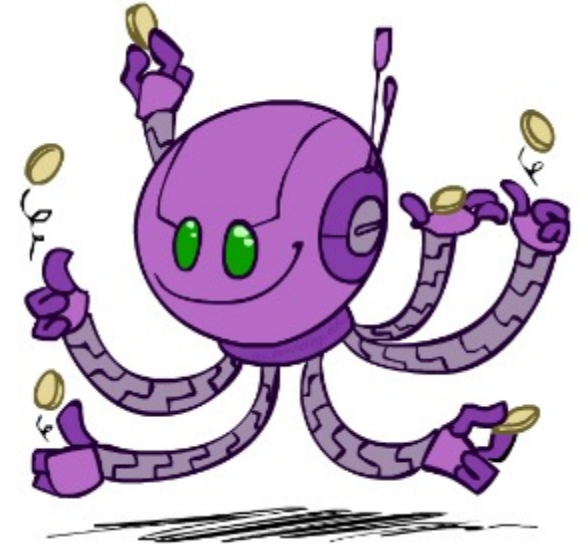
# Graphical Model Notation

- Nodes: variables (with domains)
  - Can be assigned (observed) or unassigned (unobserved)
- Arcs: interactions
  - Similar to CSP constraints
  - Indicate “direct influence” between variables
  - Formally: encode conditional independence (more later)
- For now: imagine that arrows mean direct causation (in general, they don't!)



# Example: Coin Flips

- N independent coin flips



- No interactions between variables: absolute independence

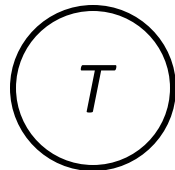
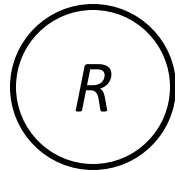
# Example: Traffic

- Variables:

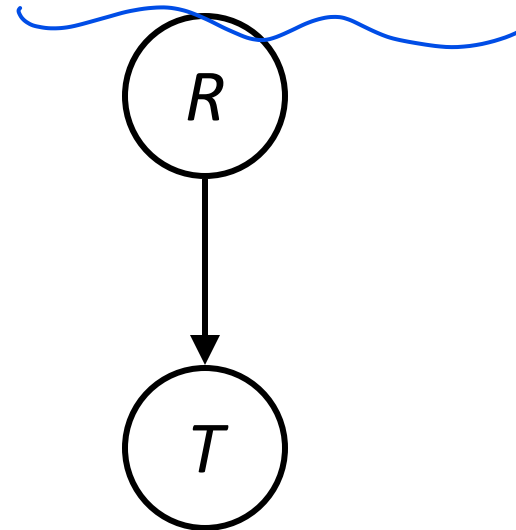
- R: It rains
- T: There is traffic



- Model 1: independence



- Model 2: rain causes traffic



- Why is an agent using model 2 better?

# Example: Traffic II

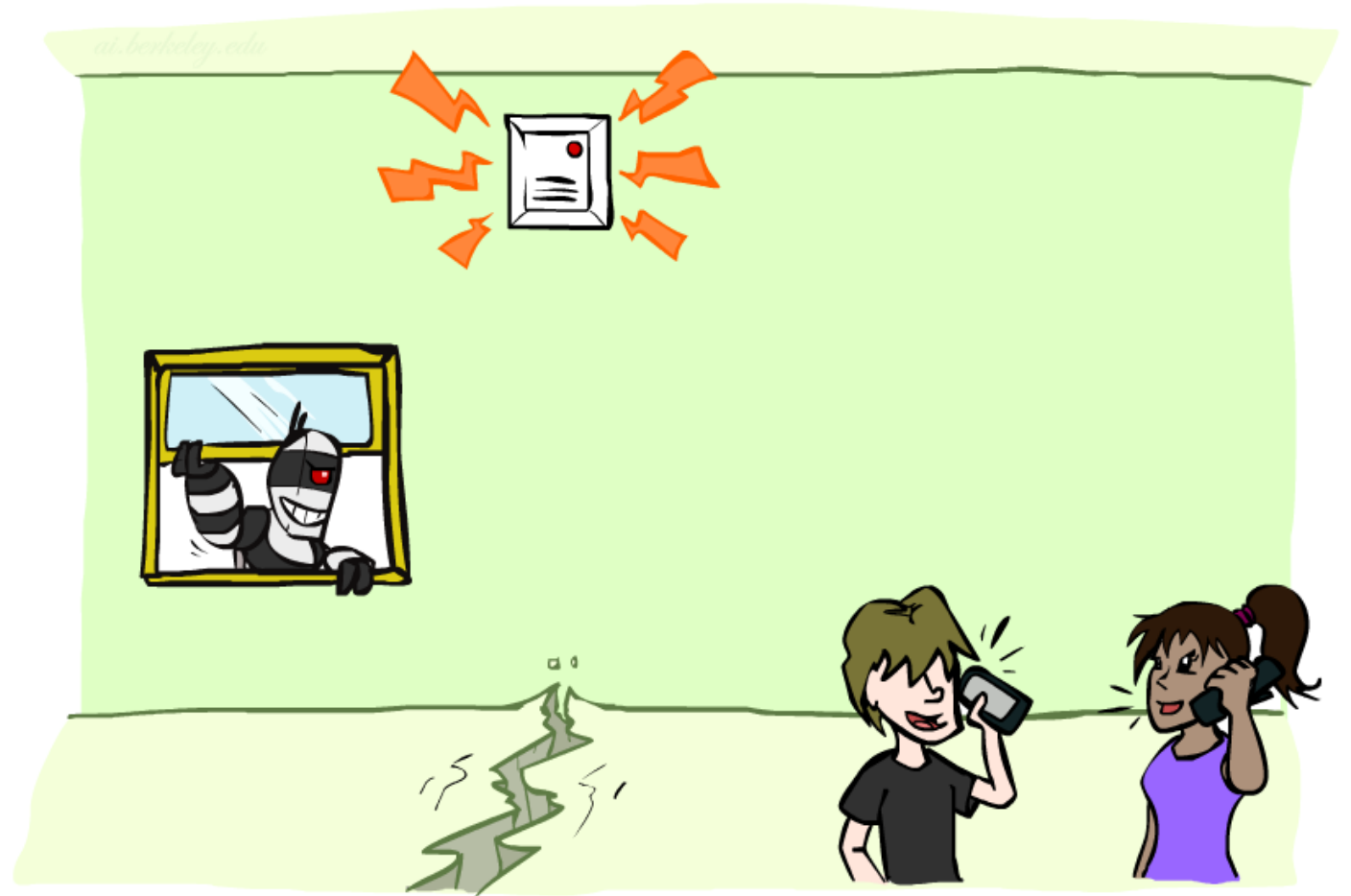
- Let's build a causal graphical model!
- Variables
  - T: Traffic
  - R: It rains
  - L: Low pressure
  - D: Roof drips
  - B: Ballgame
  - C: Cavity



# Example: Alarm Network

- Variables

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



# Bayes' Net Semantics

---



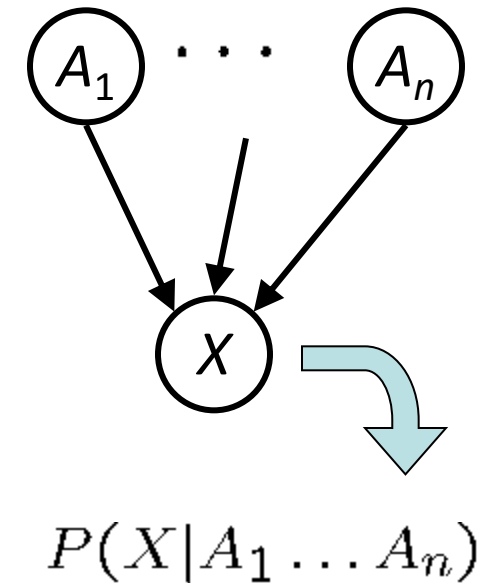
# Bayes' Net Semantics



- A set of nodes, one per variable  $X$
- A directed, acyclic graph
- A conditional distribution for each node
  - A collection of distributions over  $X$ , one for each combination of parents' values

$$P(X|a_1 \dots a_n)$$

- CPT: conditional probability table
- Description of a noisy “causal” process



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

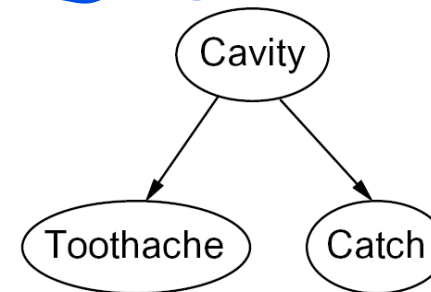
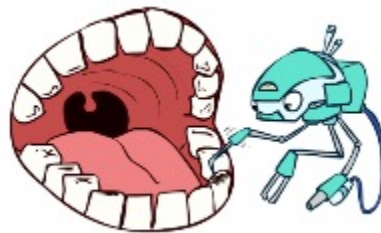
# Probabilities in BNs



- Bayes' nets **implicitly** encode joint distributions
  - As a product of local conditional distributions
  - To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{parents}(X_i))$$

- Example:



$$P(+cavity, +catch, -toothache) = P(x_1) \cdot P(x_2 | x_1) \cdot P(x_3 | x_1)$$



# Probabilities in BNs



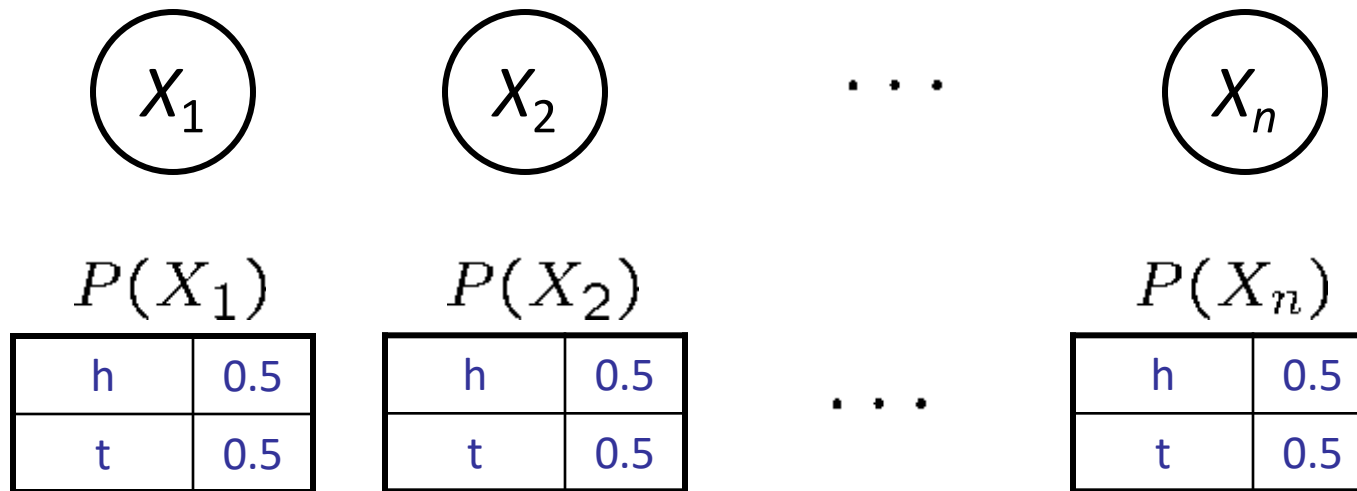
- Why are we guaranteed that setting

$$P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{parents}(X_i))$$

results in a proper joint distribution?

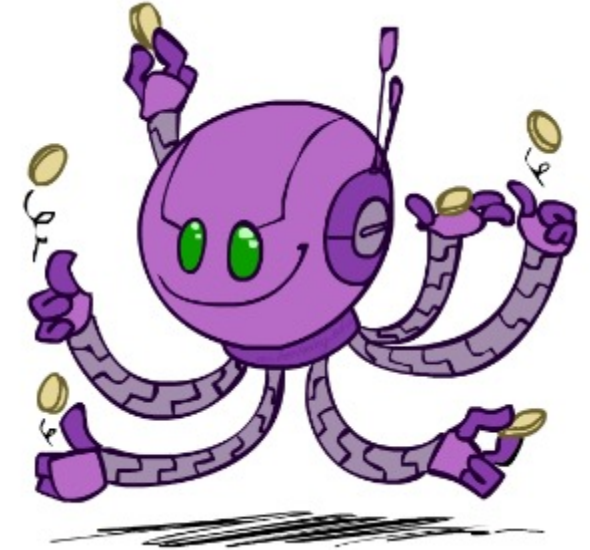
- Chain rule (valid for all distributions):  $P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | x_1 \dots x_{i-1})$
- Assume conditional independences:  $P(x_i | x_1, \dots, x_{i-1}) = P(x_i | \text{parents}(X_i))$ 
  - Consequence:  $P(x_1, x_2, \dots, x_n) = \prod_{i=1}^n P(x_i | \text{parents}(X_i))$
- Not every BN can represent every joint distribution
  - The topology enforces certain conditional independencies

# Example: Coin Flips

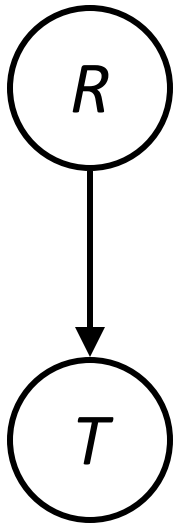


$$P(h, h, t, h) =$$

*Only distributions whose variables are absolutely independent can be represented by a Bayes' net with no arcs.*



# Example: Traffic



$P(R)$	
$+r$	$1/4$
$-r$	$3/4$

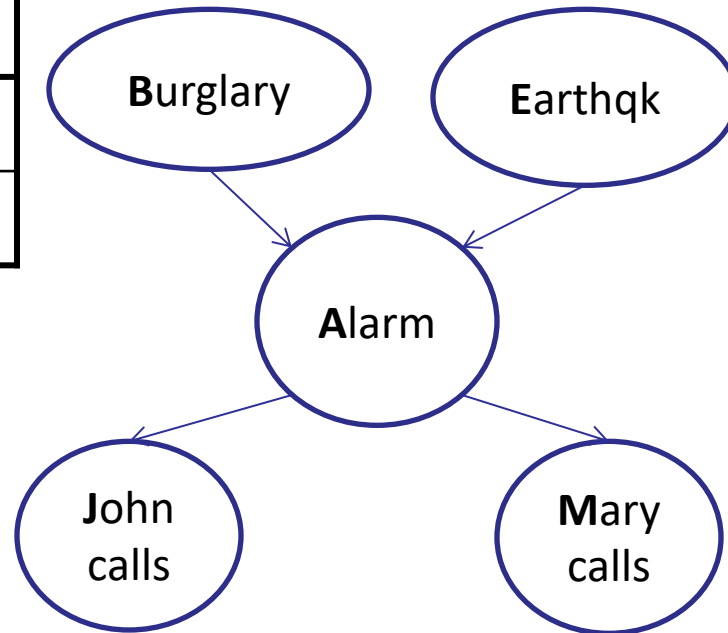
$$P(+r, -t) =$$

$P(T R)$					
$+r$	<table><tr><td><math>+t</math></td><td><math>3/4</math></td></tr><tr><td><math>-t</math></td><td><math>1/4</math></td></tr></table>	$+t$	$3/4$	$-t$	$1/4$
$+t$	$3/4$				
$-t$	$1/4$				
$-r$	<table><tr><td><math>+t</math></td><td><math>1/2</math></td></tr><tr><td><math>-t</math></td><td><math>1/2</math></td></tr></table>	$+t$	$1/2$	$-t$	$1/2$
$+t$	$1/2$				
$-t$	$1/2$				

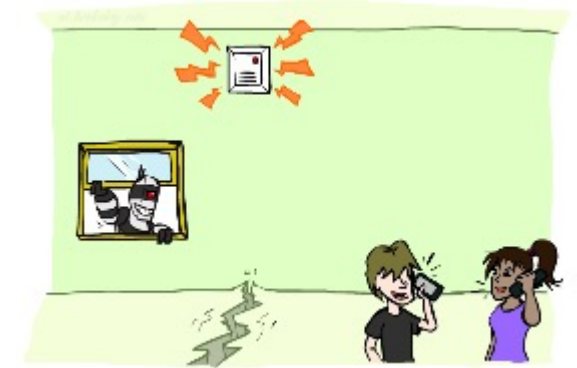


# Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



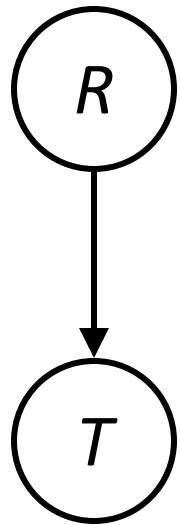
A	J	P(J A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

A	M	P(M A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

# Example: Traffic

- Causal direction



$P(R)$

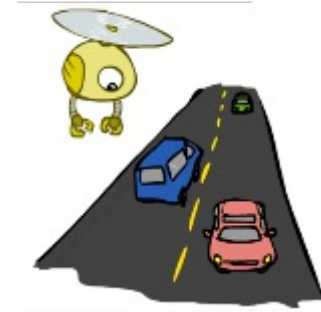
+r	1/4
-r	3/4

$P(T|R)$

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

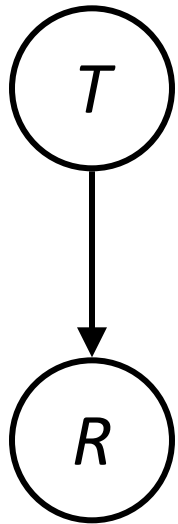
$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16



# Example: Reverse Traffic

- Reverse causality?



$P(T)$

+t	9/16
-t	7/16

$P(R|T)$

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

-t	+r	1/7
	-r	6/7

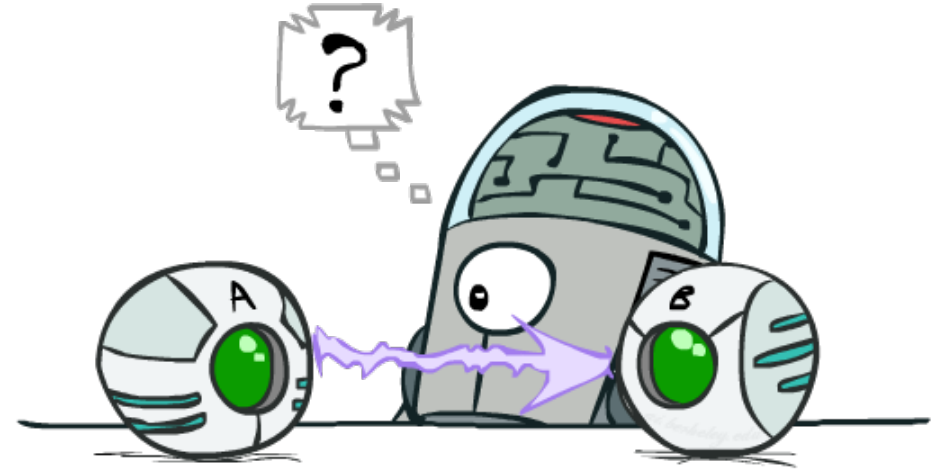


$P(T, R)$

+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

# Causality?

- When Bayes' nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts
- BNs need not actually be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - E.g. consider the variables *Traffic* and *Drips*
  - End up with arrows that reflect correlation, not causation
- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - Topology really encodes conditional independence  
$$P(x_i | x_1, \dots, x_{i-1}) = P(x_i | \text{parents}(X_i))$$



# Bayes' Nets

- So far: how a Bayes' net encodes a joint distribution  
*not every*
- Next: how to answer queries about that distribution
  - Today:
    - First assembled BNs using an intuitive notion of conditional independence as causality
    - Then saw that key property is conditional independence
  - Main goal: answer queries about conditional independence and influence
- After that: how to answer numerical queries (inference)

