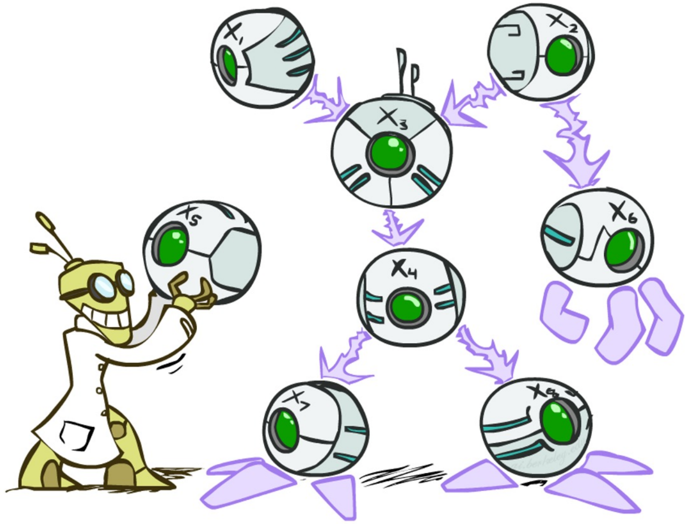
Artificial Intelligence - INFOF311

Bayes nets, basics and representation

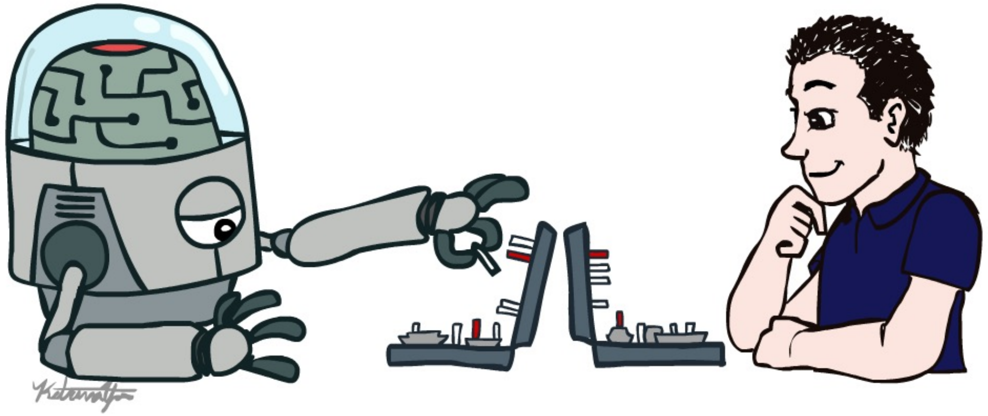


Instructor : Tom Lenaerts

Acknowledgement

We thank Stuart Russell for his generosity in allowing us to use the slide set of the UC Berkeley Course CS188, Introduction to Artificial Intelligence. These slides were created by Dan Klein, Pieter Abbeel and Anca Dragan for CS188 Intro to AI at UC Berkeley. All CS188 materials are available at http://ai.berkeley.edu.

Center for Human-Compatible Artificial Intelligence

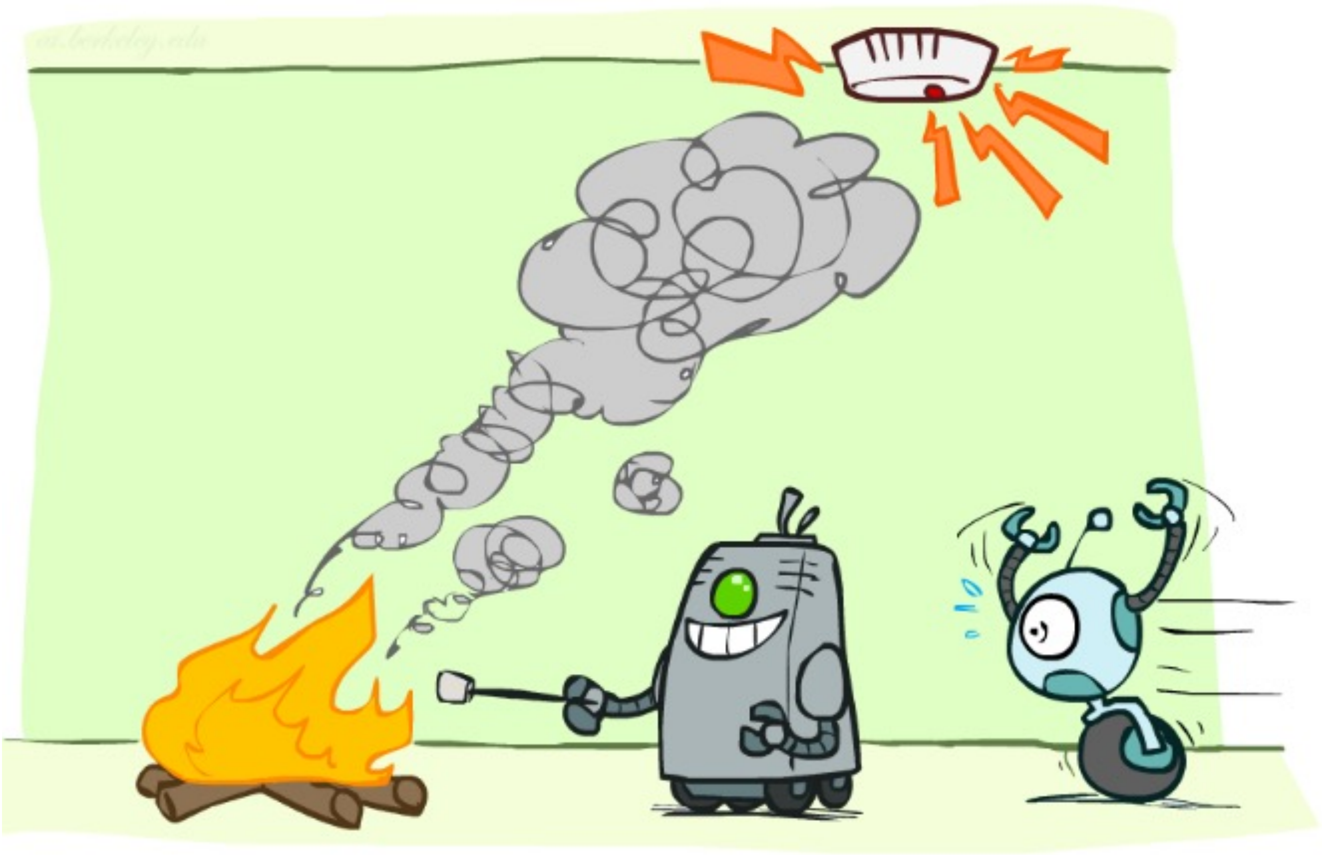


The slides for INFOF311 are slightly modified versions of the slides of the spring and summer CS188 sessions in 2021 and 2022

Reminder: elementary probability

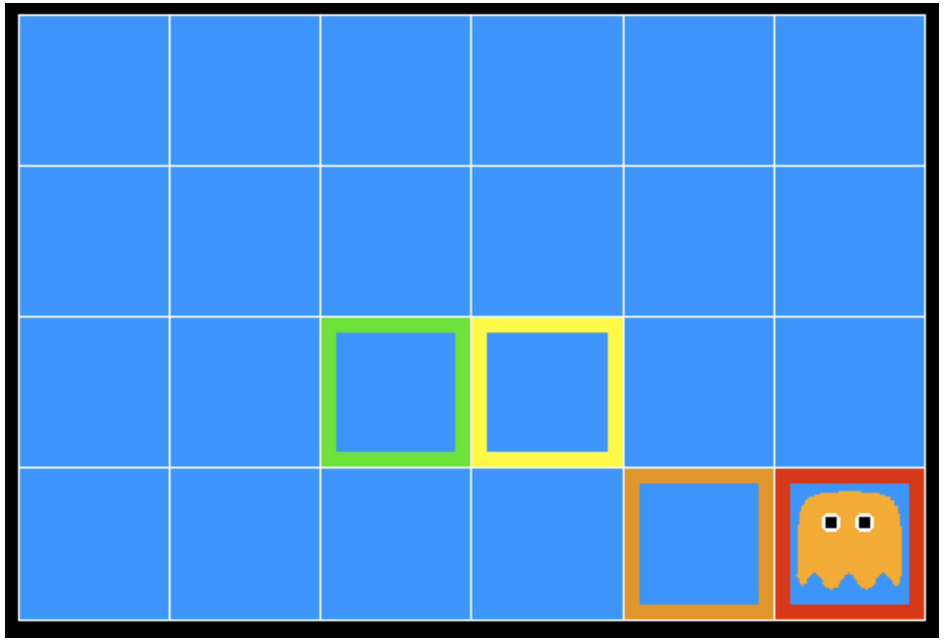
- Basic laws: $0 \leq P(\omega) \leq 1$, $\sum_{\omega \in \Omega} P(\omega) = 1$, $P(A) = \sum_{\omega \in A} P(\omega)$
- Random variable $X(\omega)$ has a value in each ω
  - Distribution $P(X)$ gives probability for each possible value x
  - Joint distribution $P(X,Y)$ gives total probability for each combination x,y
- Summing out/marginalization: $P(X=x) = \sum_y P(X=x,Y=y)$
- Conditional probability: $P(X|Y) = P(X,Y)/P(Y)$
- Product rule: $P(X|Y)P(Y) = P(X,Y) = P(Y|X)P(X)$
  - Generalize to chain rule: $P(X_1,..,X_n) = \prod_i P(X_i \mid X_1,..,X_{i-1})$
- Bayes Rule: $P(X|Y) = P(Y|X)P(X)/P(Y)$

Conditional Independence



Ghostbusters

- A ghost is in the grid somewhere
- Sensor readings tell how close a square is to the ghost
  - On the ghost: usually red
  - 1 or 2 away: mostly orange
  - 3 or 4 away: typically yellow
  - 5+ away: often green
- Click on squares until confident of location, then "***bust***"



Video of Demo Ghostbusters with Probability

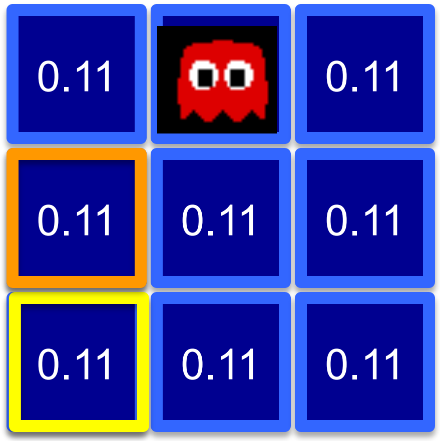
Ghostbusters model

- Variables and ranges:
  - G (ghost location) in $\{(1,1),\ldots,(3,3)\}$
  - $C_{x,y}$ (color measured at square x,y) in {red,orange,yellow,green}

0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11

- Ghostbuster physics:
  - ***Uniform prior distribution*** over ghost location: $P(G)$
  - ***Sensor model***: $P(C_{x,y} \mid G)$ (depends only on distance to G)
    - E.g. $P(C_{1,1} = \text{yellow} \mid G = (1,1)) = 0.1$

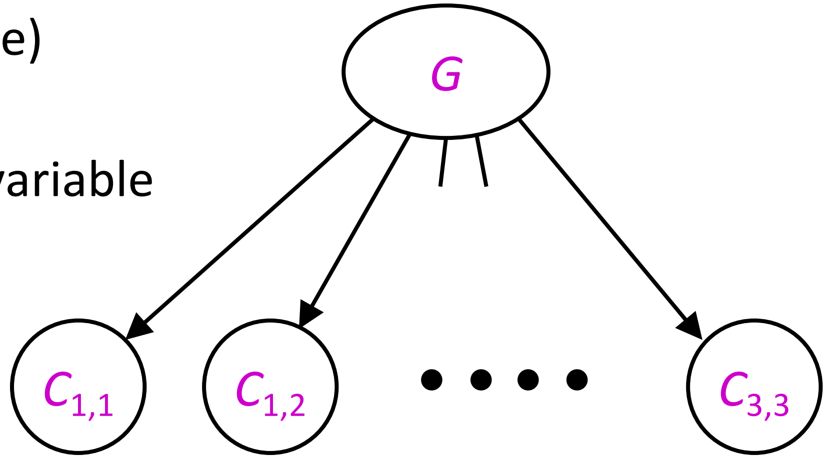
Ghostbusters model, contd.



- $P(G, C_{1,1}, \ldots C_{3,3})$ has $9 \times 4^9 = 2{,}359{,}296$ entries!!!
- Ghostbuster independence:
  - Are $C_{1,1}$ and $C_{1,2}$ independent?
    - E.g., does $P(C_{1,1} = \text{yellow}) = P(C_{1,1} = \text{yellow} \mid C_{1,2} = \text{orange})$?
- Ghostbuster physics again:
  - $P(C_{x,y} \mid G)$ ***depends only on distance to G***
    - So $P(C_{1,1} = \text{yellow} \mid G = (2,3)) = P(C_{1,1} = \text{yellow} \mid G = (2,3), C_{1,2} = \text{orange})$
    - I.e., $C_{1,1}$ is ***conditionally independent*** of $C_{1,2}$ ***given*** G

Ghostbusters model, contd.

- Apply the chain rule to decompose the joint probability model:
- $P(G, C_{1,1}, \ldots C_{3,3}) = P(G)\, P(C_{1,1} \mid G)\, P(C_{1,2} \mid G, C_{1,1})\, P(C_{1,3} \mid G, C_{1,1}, C_{1,2}) \ldots P(C_{3,3} \mid G, C_{1,1}, \ldots, C_{3,2})$
- Now simplify using conditional independence:
- $P(G, C_{1,1}, \ldots C_{3,3}) = P(G)\, P(C_{1,1} \mid G)\, P(C_{1,2} \mid G)\, P(C_{1,3} \mid G) \ldots P(C_{3,3} \mid G)$
- I.e., conditional independence properties of ghostbuster physics simplify the probability model from ***exponential*** to ***quadratic*** in the number of squares
- This is called a ***Naïve Bayes*** model:
  - One discrete query variable (often called the ***class*** or ***category*** variable)
  - All other variables are (potentially) evidence variables
  - Evidence variables are all conditionally independent given the query variable



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:

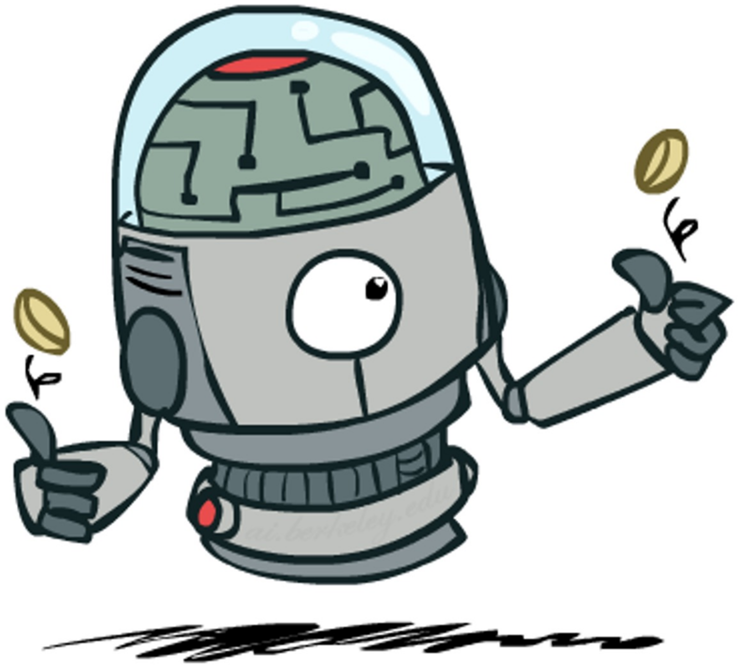
$$\forall x, y : P(x|y) = P(x)$$

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}?



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

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

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

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

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

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

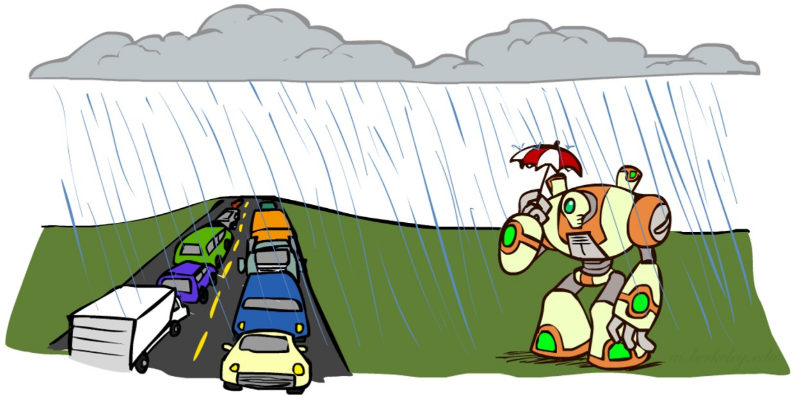
Conditional Independence

- 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:



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

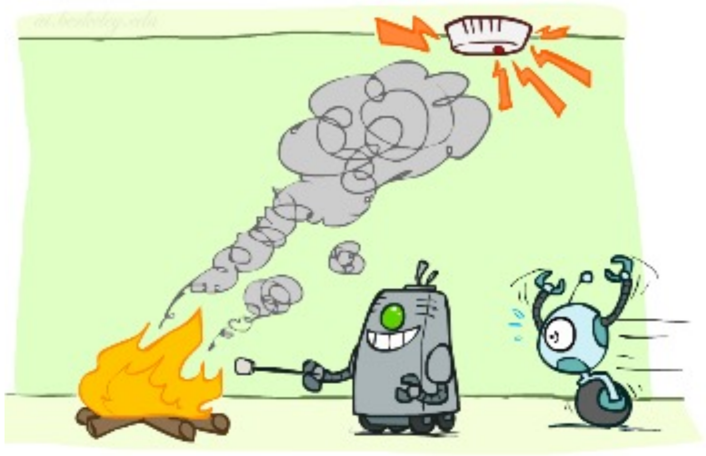
With assumption of conditional independence:

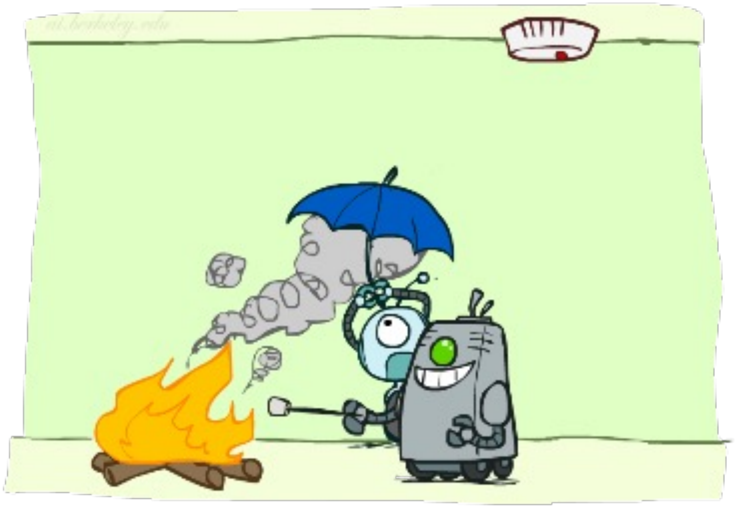
$$P(\text{Traffic}, \text{Rain}, \text{Umbrella}) =$$
$$P(\text{Rain})P(\text{Traffic}|\text{Rain})P(\text{Umbrella}|\text{Rain})$$

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

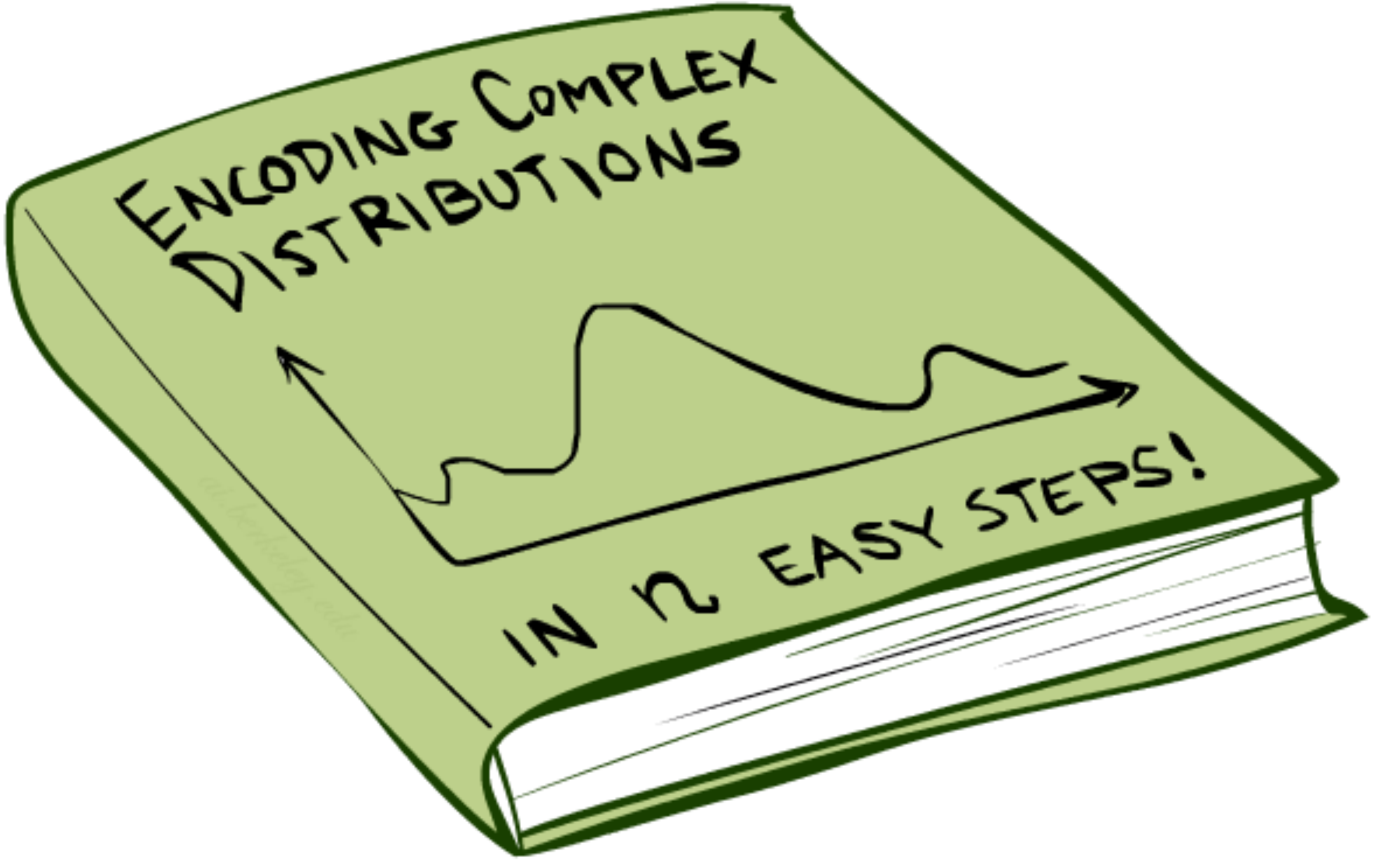
Conditional Independence

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





Bayes Nets: Big Picture



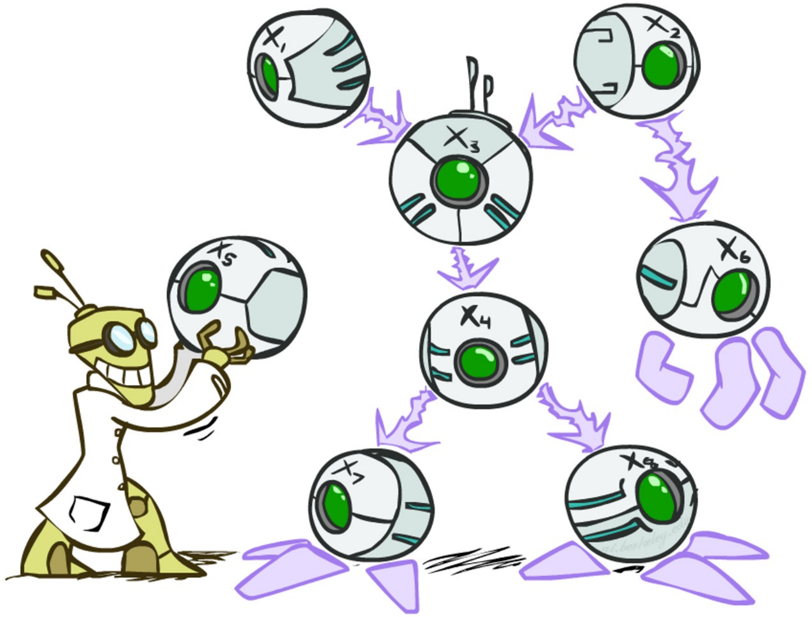
Bayes' Nets: Big Picture

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



Bayes Nets

Part I: Basics, Representation

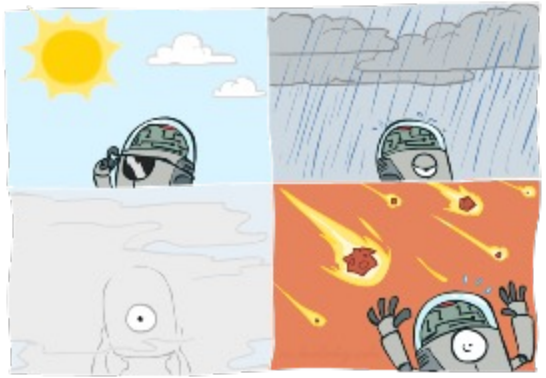
Part II: Exact inference

Part III: Independence

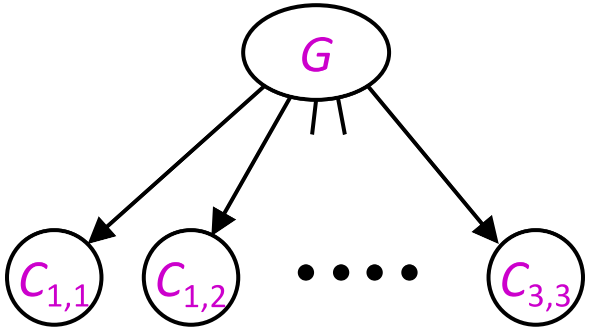
Part IV: Approximate Inference

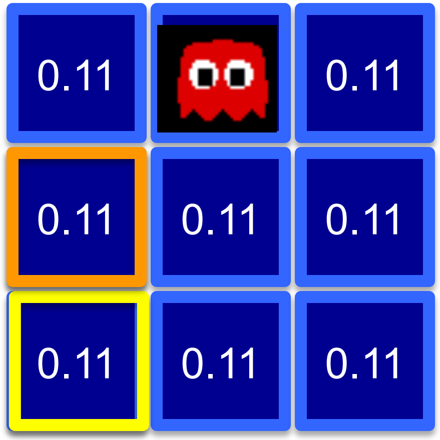
Graphical Model Notation

- Nodes: variables (with domains)
  - Can be assigned (observed) or unassigned (unobserved)



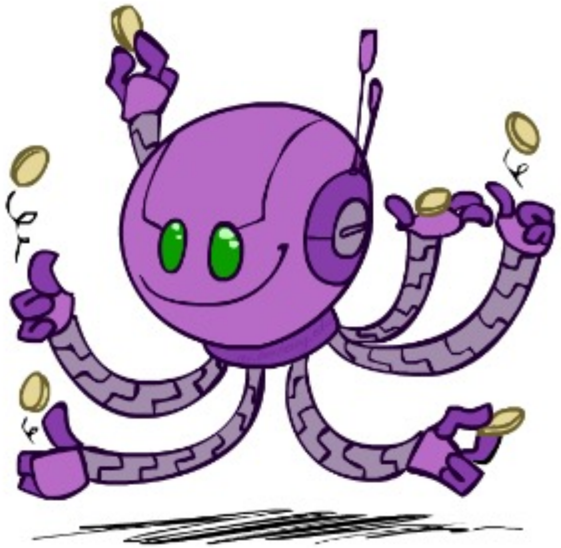
- Arcs: interactions
  - Indicate “direct influence” between variables
  - Formally: absence of arc encodes conditional independence (more later)





Example: Coin Flips

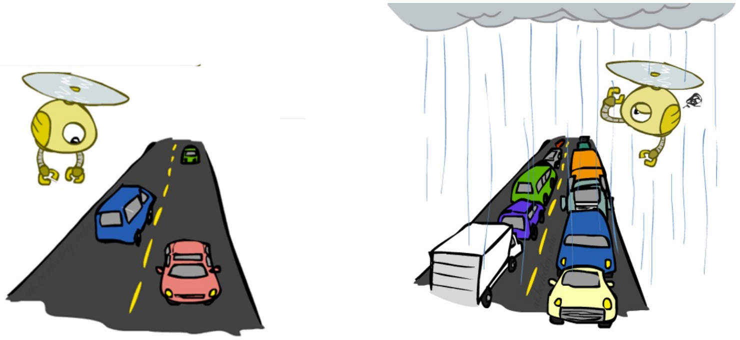
- N independent coin flips



$X_1 \quad X_2 \quad \cdots \quad X_n$

- No interactions between variables: absolute independence

Example: Traffic



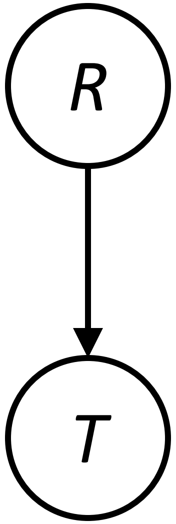
Variables:

R: It rains

T: There is traffic

Model 1: independence

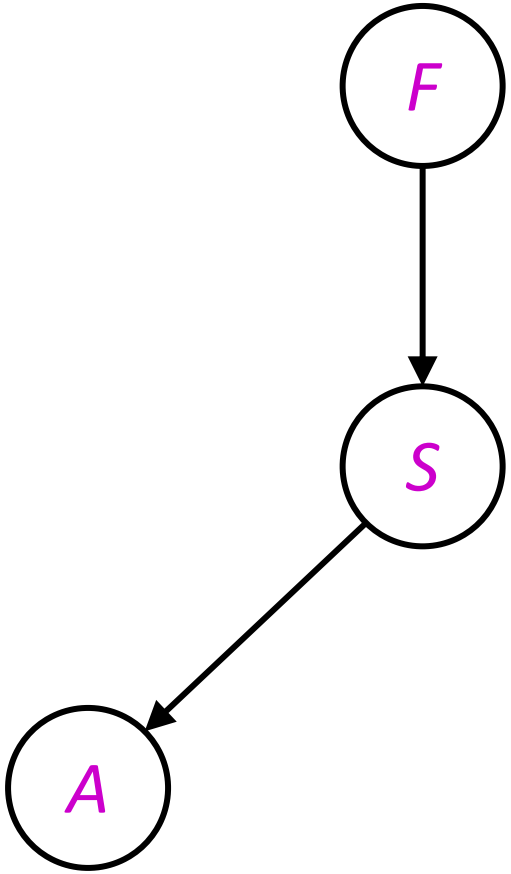
- Model 2: rain causes traffic

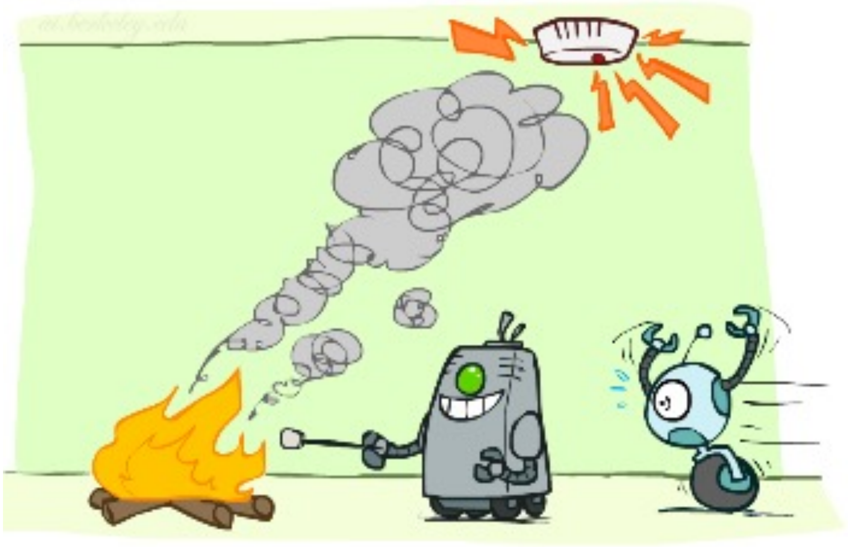


Why is an agent using model 2 better?

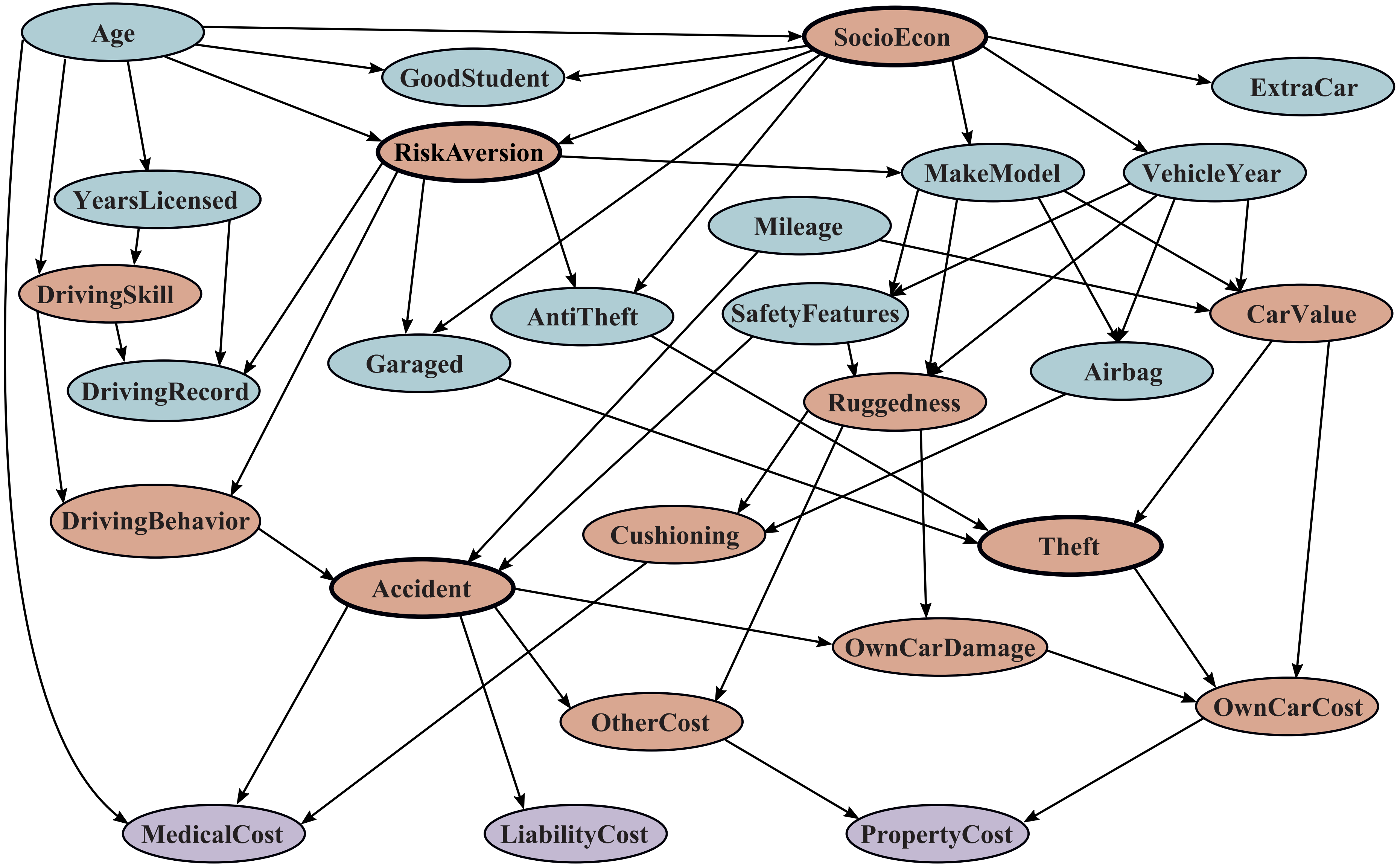
Example: Smoke alarm

- Variables:
  - F: There is fire
  - S: There is smoke
  - A: Alarm sounds

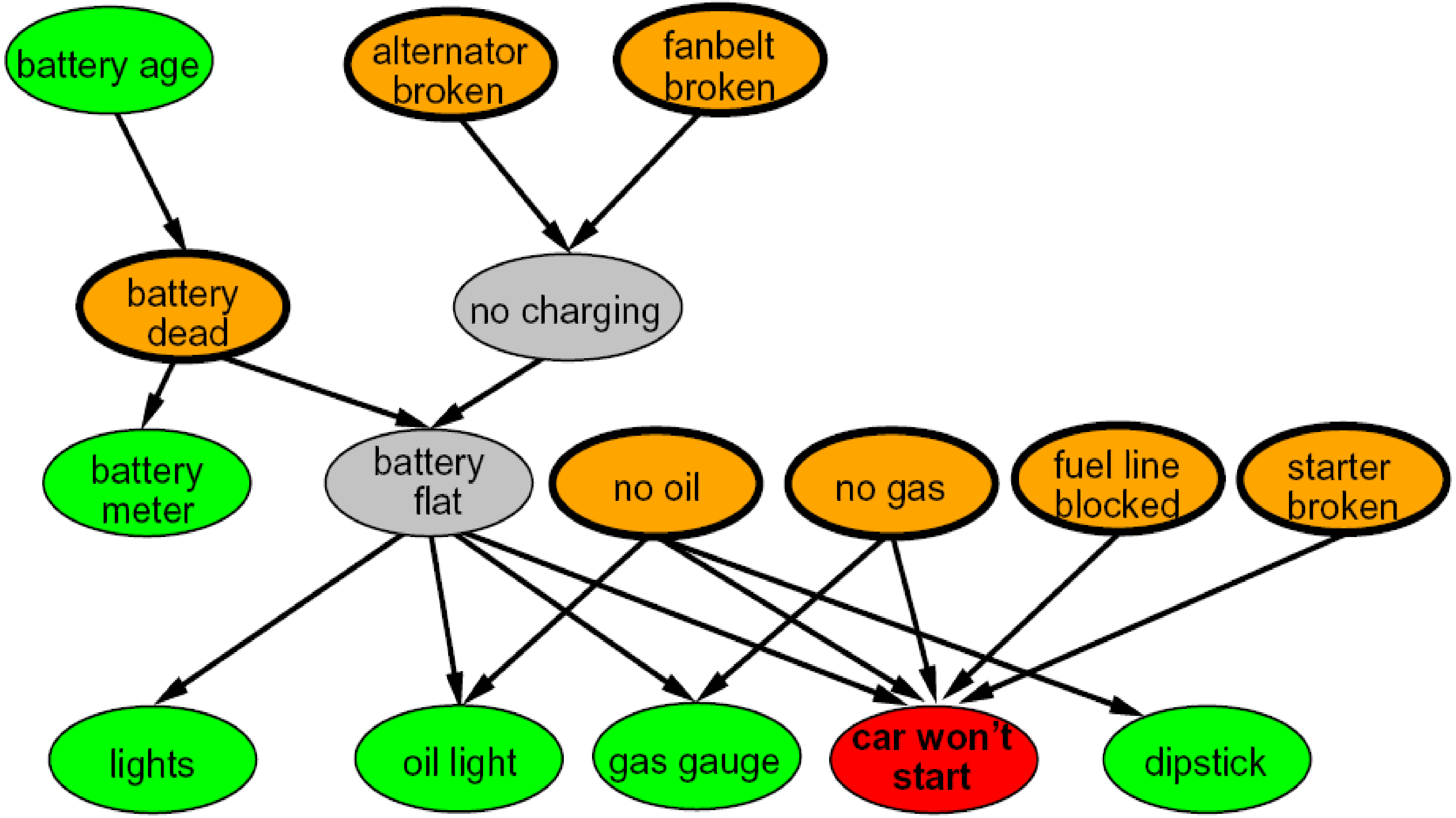




Example Bayes' Net: Car Insurance



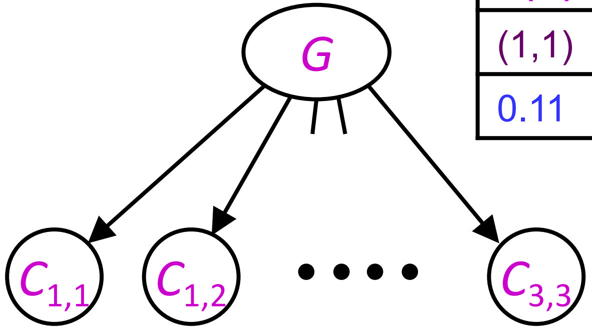
Example Bayes' Net: Car Won't Start



Bayes Net Syntax and Semantics

Bayes Net Syntax

- A set of nodes, one per variable *X*
- A directed, acyclic graph
- A conditional distribution for each node given its ***parent variables*** in the graph
  - ***CPT*** (conditional probability table); each row is a distribution for child given values of its parents



P(*G*)			
(1,1)	(1,2)	(1,3)	…
0.11	0.11	0.11	…

G	**P($C_{1,1}$ | *G*)**			
	g	y	o	r
(1,1)	0.01	0.1	0.3	0.59
(1,2)	0.1	0.3	0.5	0.1
(1,3)	0.3	0.5	0.19	0.01
…				

Bayes net = Topology (graph) + Local Conditional Probabilities

Probabilities in BNs

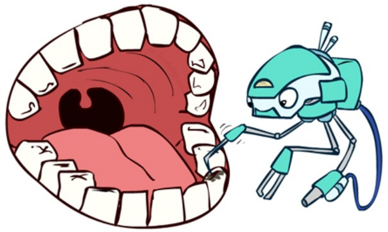
Bayes’ nets implicitly encode joint distributions

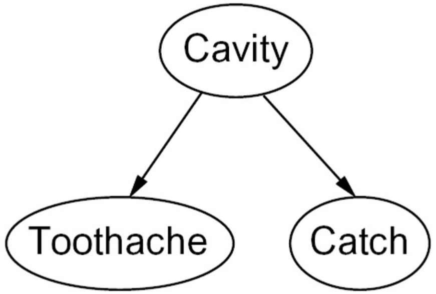
As a product of local conditional distributions

To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

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

Example:





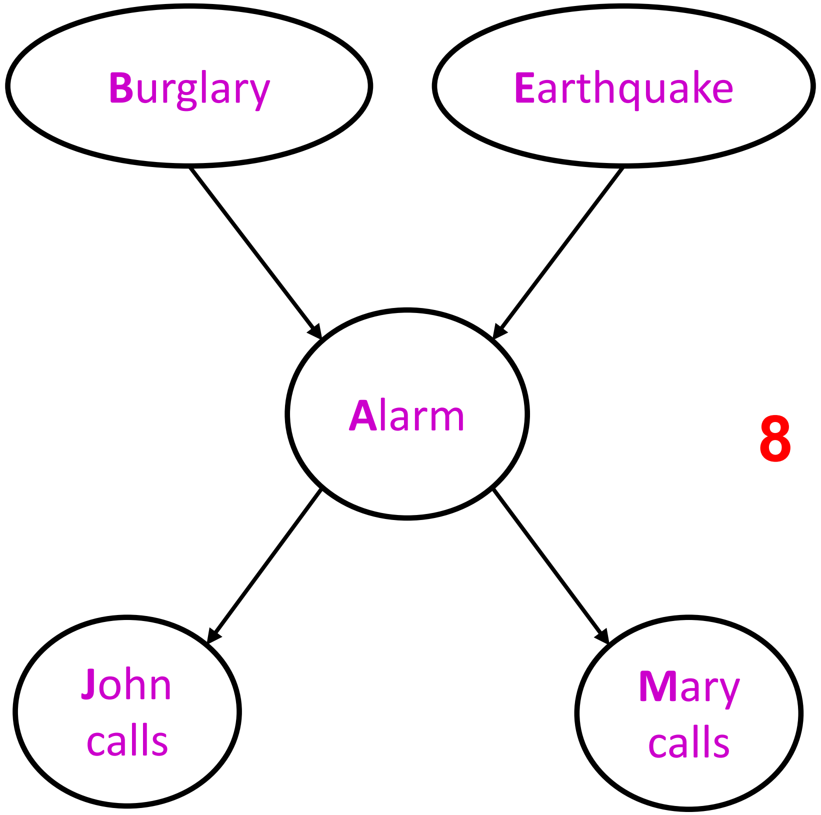
$$P(+\mathit{cavity},\ +\mathit{catch},\ \text{-}\mathit{toothache})$$

=P(-toothache|+cavity)P(+catch|+cavity)P(+cavity)

Example: Alarm Network

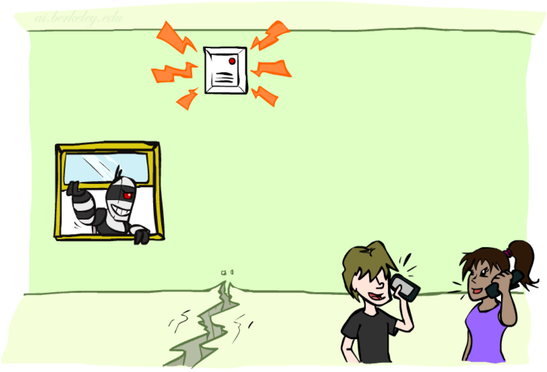
P(B)	
true	false
0.001	0.999

2



2

P(E)	
true	false
0.002	0.998



8

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

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

4

4

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

Number of ***free parameters*** in each CPT:

Parent range sizes $d_1,\ldots,d_k$

Child range size d
Each table **row** must sum to 1

$d \ \Pi_i \ d_i$

General formula for sparse BNs

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

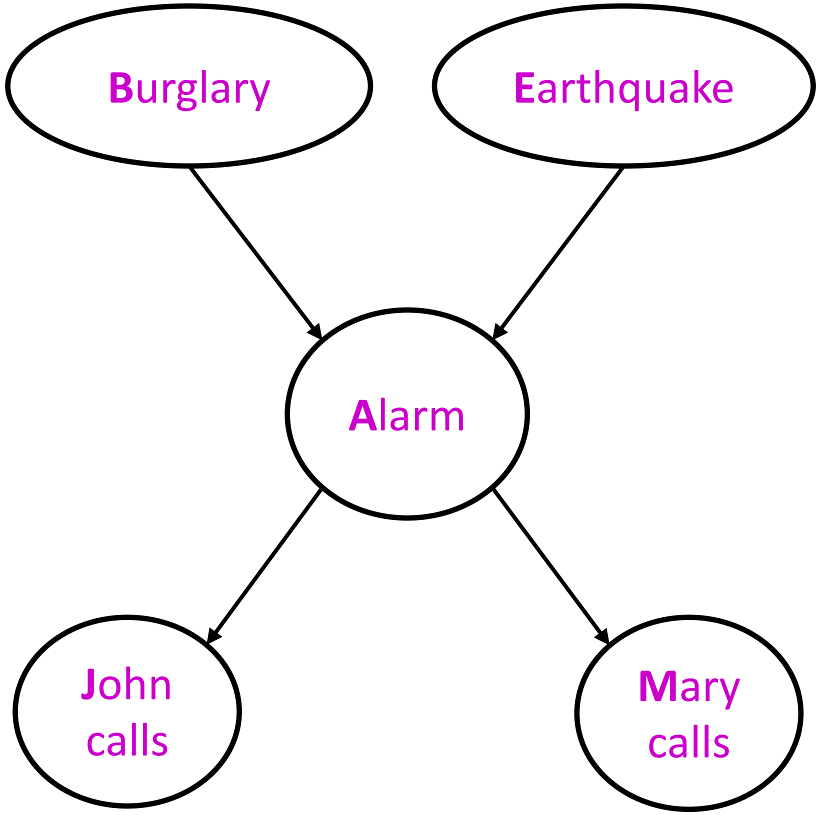
Bayes net global semantics

- Bayes nets encode joint distributions as product of conditional distributions on each variable:

$$P(X_1,..,X_n) = \prod_i P(X_i \mid Parents(X_i))$$

Example

P(B)	
true	false
0.001	0.999



P(E)	
true	false
0.002	0.998

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

$P(b, \neg e, a, \neg j, \neg m) =$

$P(b)\ P(\neg e)\ P(a \mid b, \neg e)\ P(\neg j \mid a)\ P(\neg m \mid a)$

$=.001 \times .998 \times .94 \times .1 \times .3 = .000028$

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

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

Conditional independence in BNs

- Compare the Bayes net global semantics

$$P(X_1,..,X_n) = \prod_i P(X_i \mid Parents(X_i))$$

with the chain rule identity

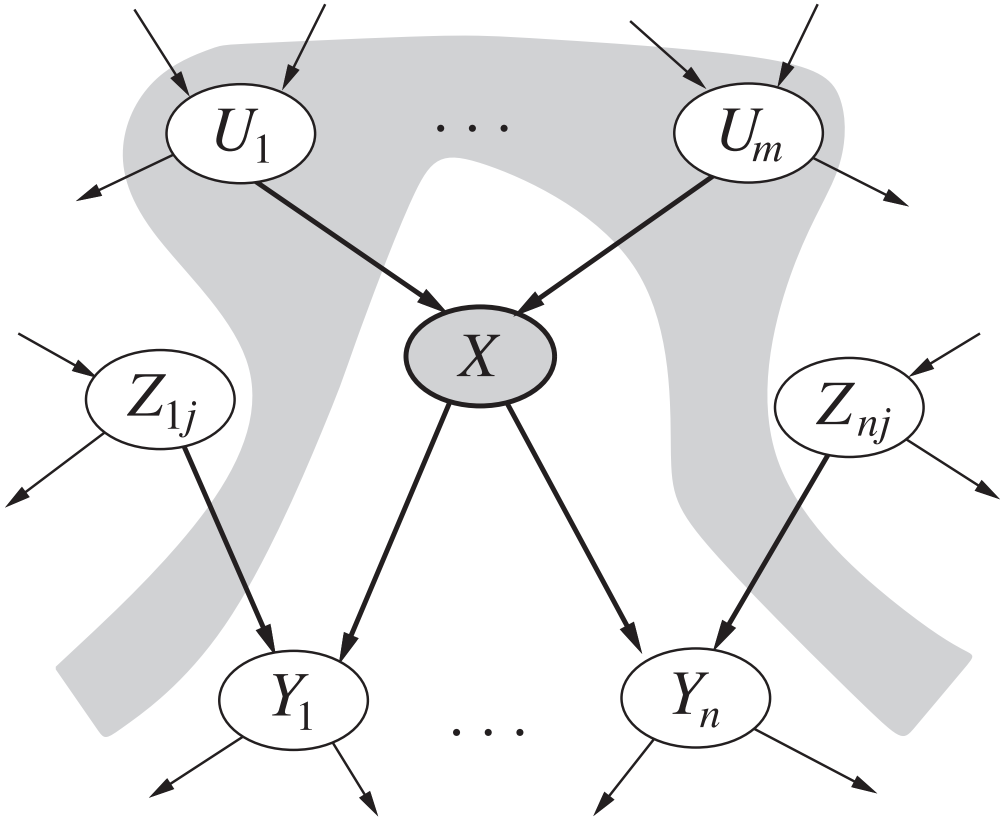
$$P(X_1,..,X_n) = \prod_i P(X_i \mid X_1,\ldots,X_{i-1})$$

Order consistent with graph structure

- Assume (without loss of generality) that $X_1,..,X_n$ sorted in topological order according to the graph (i.e., parents before children), so $Parents(X_i) \subseteq X_1,\ldots,X_{i-1}$
- The Bayes net asserts conditional independences $P(X_i \mid X_1,\ldots,X_{i-1}) = P(X_i \mid Parents(X_i))$
  - To ensure these are valid, choose parents for node X_i that “shield” it from other predecessors

Conditional independence semantics

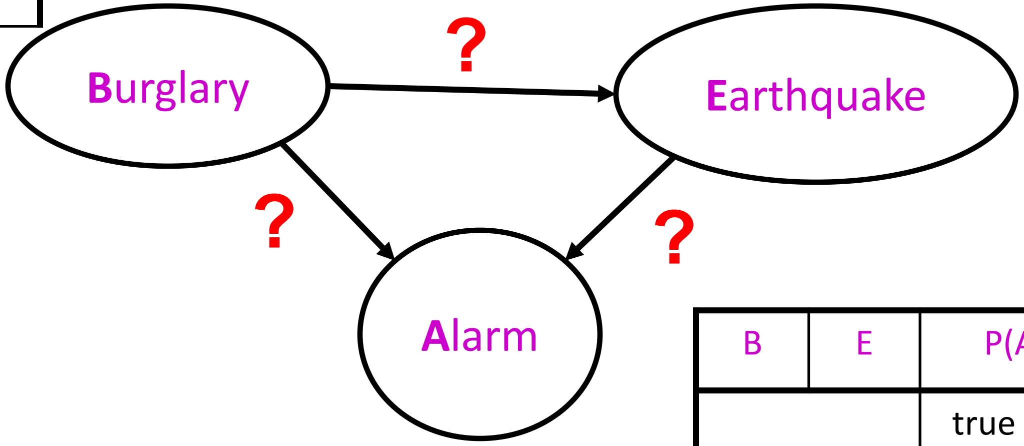
- ***Every variable is conditionally independent of its non-descendants given its parents***
- Conditional independence semantics <=> global semantics



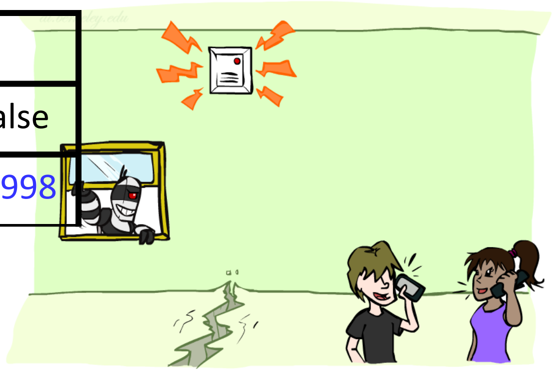
Example: Burglary

- Burglary
- Earthquake
- Alarm

P(B)	
true	false
0.001	0.999



P(E)	
true	false
0.002	0.998

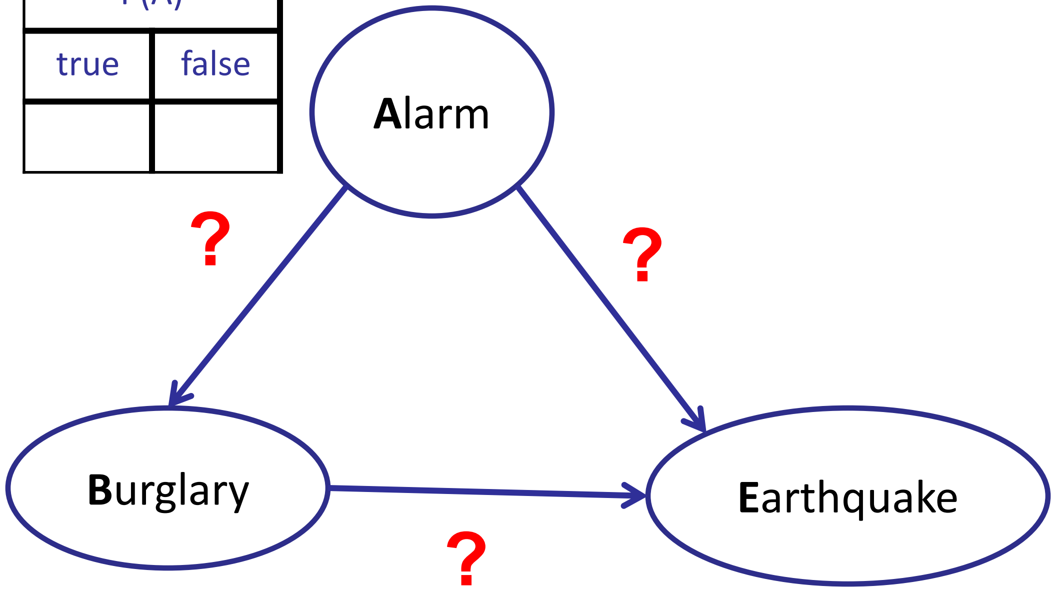


B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

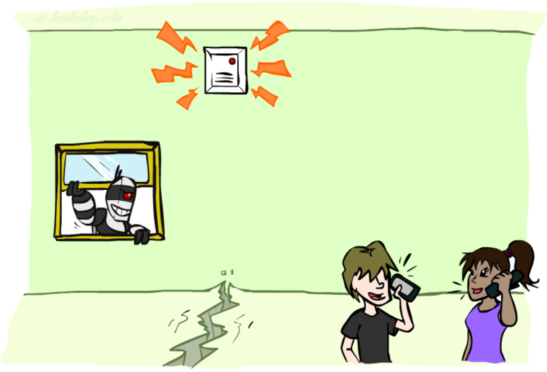
Example: Burglary

- Alarm
- Burglary
- Earthquake

P(A)	
true	false



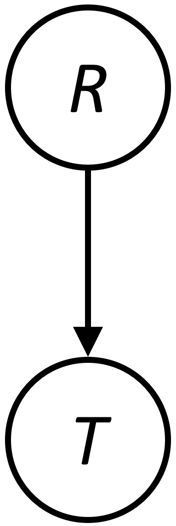
A	P(B|A)	
	true	false
true	?	
false		



A	B	P(E|A,B)	
		true	false
true	true	?	
true	false		
false	true		
false	false		

Causality?

- 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 *Rain*

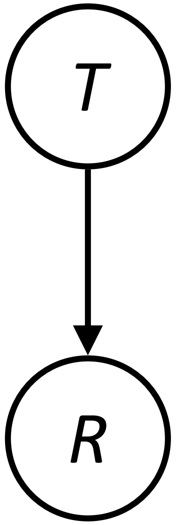


$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

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



$P(T)$

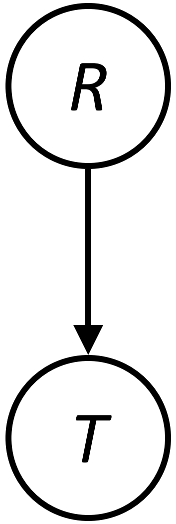
+t	9/16
-t	7/16

$P(R|T)$

+t	+r	1/3
	-r	2/3
-t	+r	1/7
	-r	6/7

Example: Traffic

- Causal direction



$$P(R)$$

+r	1/4
-r	3/4

$$P(T|R)$$

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

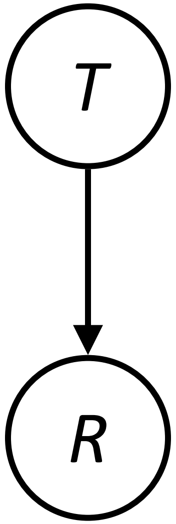
-r	+t	1/2
	-t	1/2

$$P(T,R)$$

+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

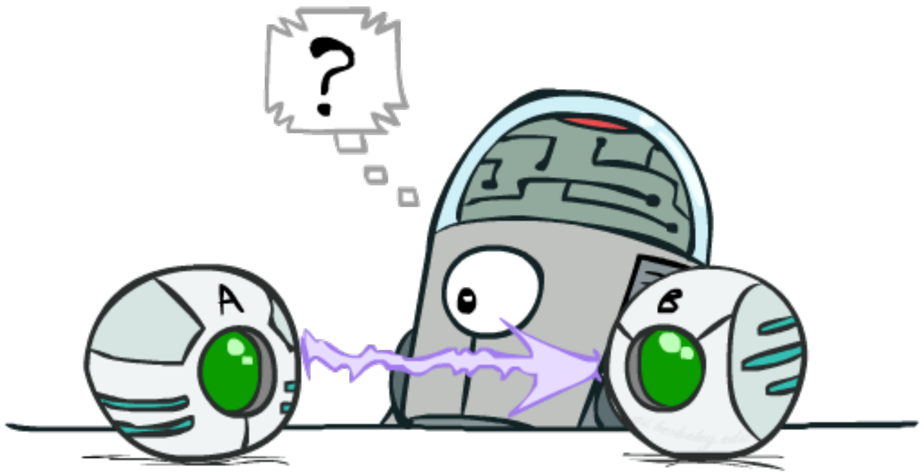
Causality?

- When Bayes' nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to estimate probabilities from data

- 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

$$P(x_i|x_1,\ldots x_{i-1}) = P(x_i|\textit{parents}(X_i))$$



Next Time

✓ Part I: Basics, Representation

Part II: Exact inference

- Enumeration (always exponential complexity)
- Variable elimination (worst-case exponential complexity, often better)

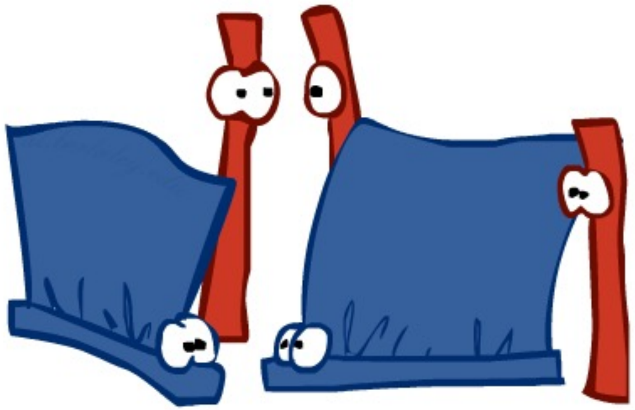
Part III: Independence

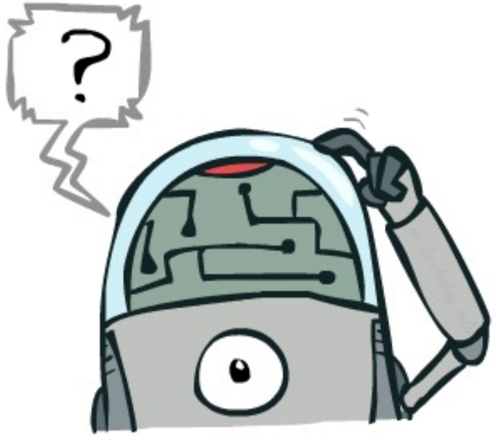
Part IV: Approximate Inference

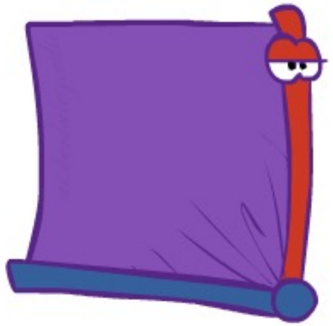
Inference

- Inference: calculating some useful quantity from a probability model (joint probability distribution)

- Examples:
  - Posterior marginal probability
    - $P(Q|e_1,..,e_k)$
    - E.g., what disease might I have?
  - Most likely explanation:
    - $\text{argmax}_{q,r,s}\ P(Q=q,R=r,S=s|e_1,..,e_k)$
    - E.g., what did they say?







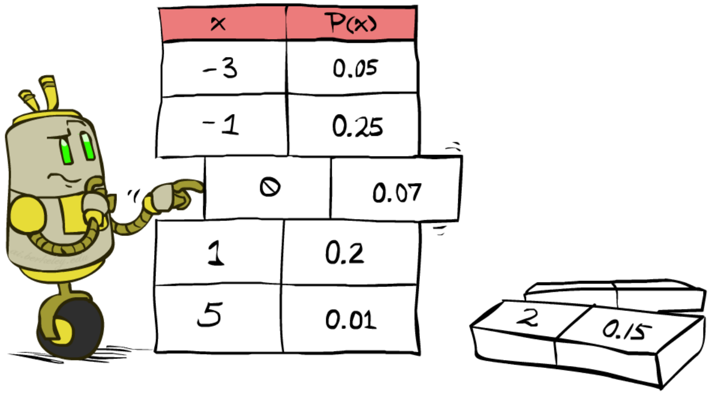
Inference by Enumeration

- Probability model $P(X_1, \ldots, X_n)$ is given
- Partition the variables $X_1, \ldots, X_n$ into sets as follows:
  - Evidence variables: $\boldsymbol{E} = \boldsymbol{e}$
  - Query variables: $\boldsymbol{Q}$
  - Hidden variables: $\boldsymbol{H}$

- We want:

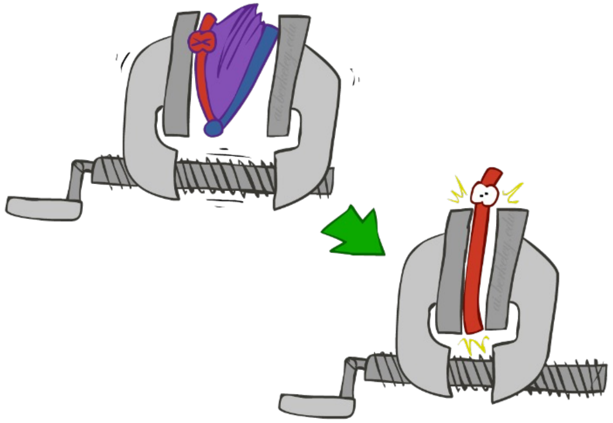
$$P(\boldsymbol{Q} \mid \boldsymbol{e})$$

- Step 1: Select the entries consistent with the evidence



- Step 2: Sum out $\boldsymbol{H}$ from model to get joint of query and evidence

$$P(\boldsymbol{Q}, \boldsymbol{e}) = \sum_h \underbrace{P(\boldsymbol{Q}, \boldsymbol{h}, \boldsymbol{e})}_{X_1, \ldots, X_n}$$

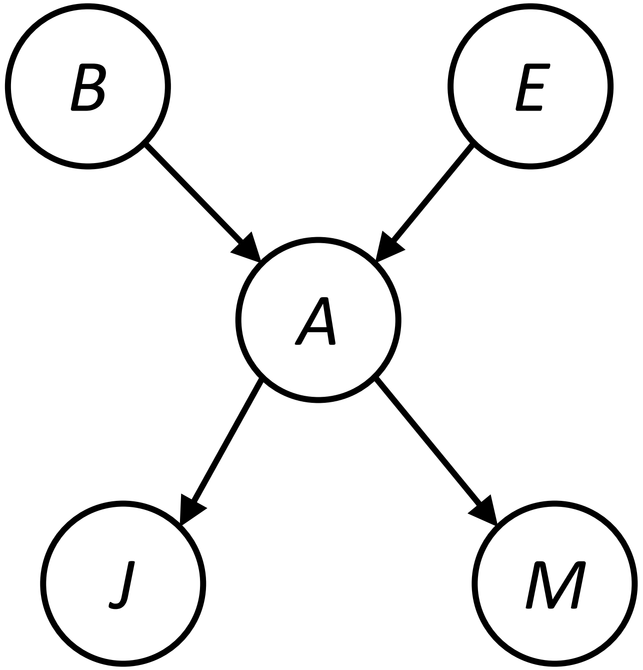


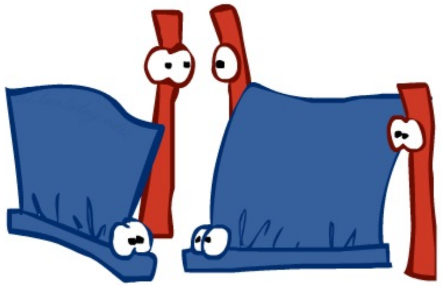
- Step 3: Normalize

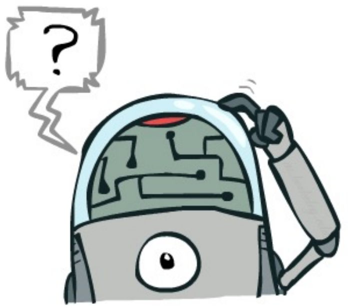
$$P(\boldsymbol{Q} \mid \boldsymbol{e}) = \alpha\, P(\boldsymbol{Q}, \boldsymbol{e})$$

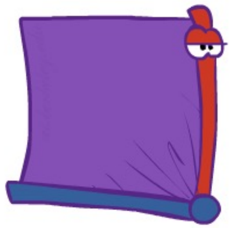
Inference by Enumeration in Bayes Net

- Reminder of inference by enumeration:
  - Any probability of interest can be computed by summing entries from the joint distribution: $P(\boldsymbol{Q} \mid \boldsymbol{e}) = \alpha \sum_{h} P(\boldsymbol{Q}, \boldsymbol{h}, \boldsymbol{e})$
  - Entries from the joint distribution can be obtained from a BN by multiplying the corresponding conditional probabilities
- $P(B \mid j, m) = \alpha \sum_{e,a} P(B, e, a, j, m)$
- $= \alpha \sum_{e,a} P(B)\, P(e)\, P(a \mid B,e)\, P(j \mid a)\, P(m \mid a)$
- So, inference in Bayes nets means computing sums of products of numbers. Sounds easy!!
- Problem: sums of ***exponentially many*** products!









Can we do better?

- Consider **uwy + uwz + uxy + uxz + vwy + vwz + vxy +vxz**
  - 16 multiplies, 7 adds
  - Lots of repeated sub-expressions!
- Rewrite as **(u+v)(w+x)(y+z)**
  - 2 multiplies, 3 adds
- $\sum_{e,a} P(B)\ P(e)\ P(a|B,e)\ P(j|a)\ P(m|a)$
- $= P(B)P(e)P(a|B,e)P(j|a)P(m|a) + P(B)P(\neg e)P(a|B,\neg e)P(j|a)P(m|a)$

  $+ P(B)P(e)P(\neg a|B,e)P(j|\neg a)P(m|\neg a) + P(B)P(\neg e)P(\neg a|B,\neg e)P(j|\neg a)P(m|\neg a)$

  Lots of repeated sub-expressions!

To Summarize …

- Independence and conditional independence are important forms of probabilistic knowledge
- Bayes net encode joint distributions efficiently by taking advantage of conditional independence
  - Global joint probability = product of local conditionals
- Exact inference = sums of products of conditional probabilities from the network