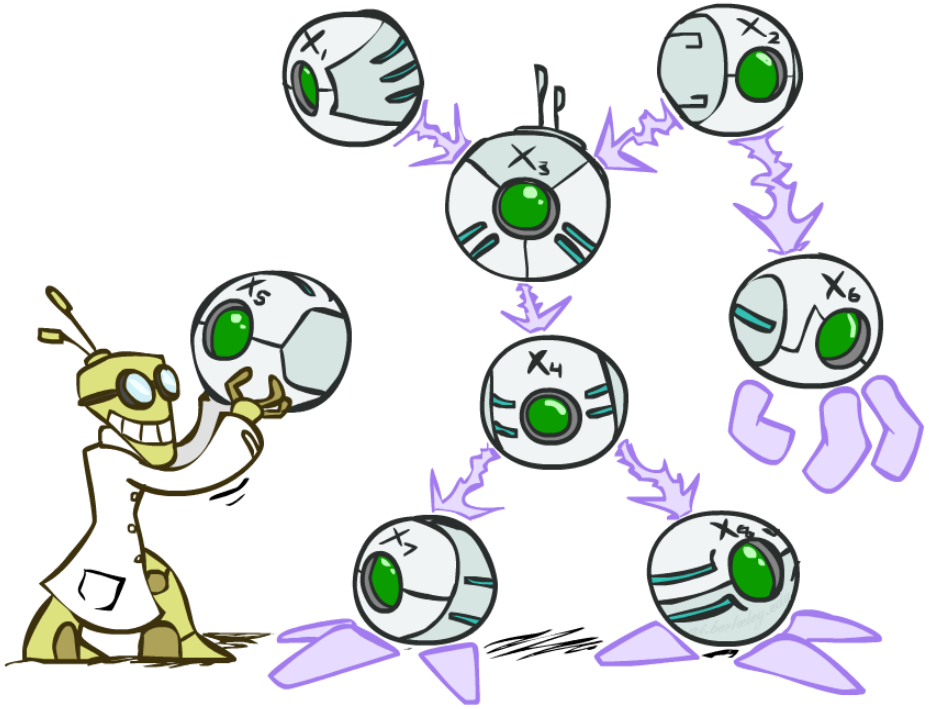
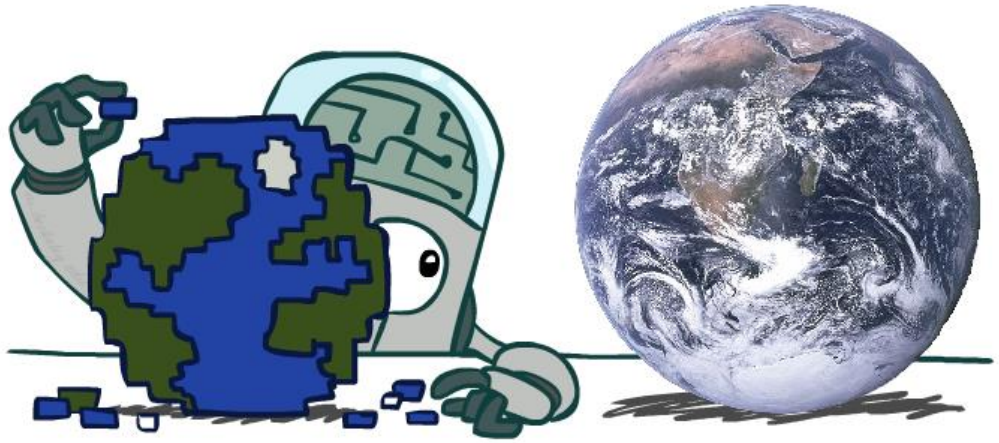
Bayes' Nets



Probabilistic Models

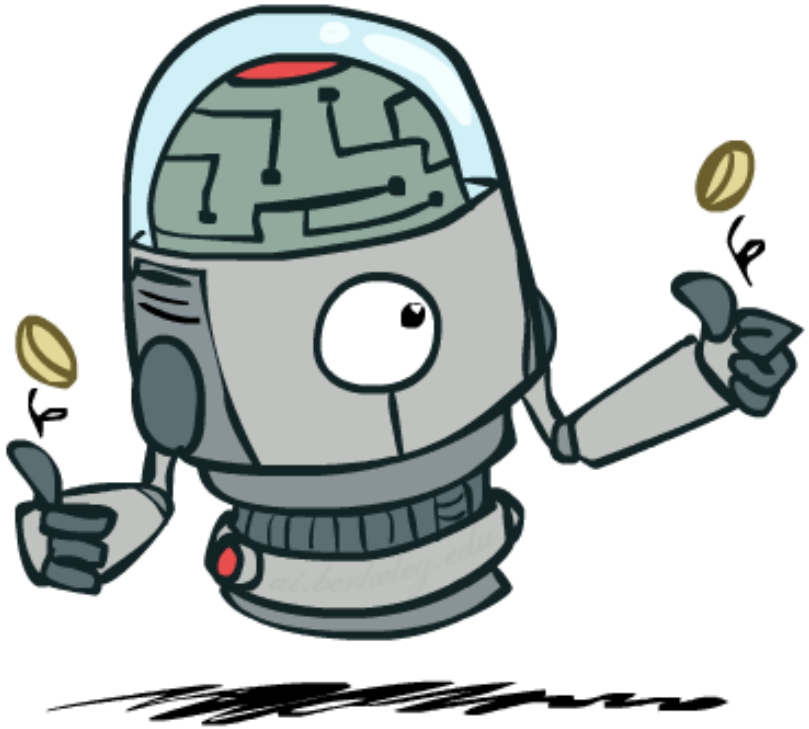
- Models describe how (a portion of) the world works

- 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:

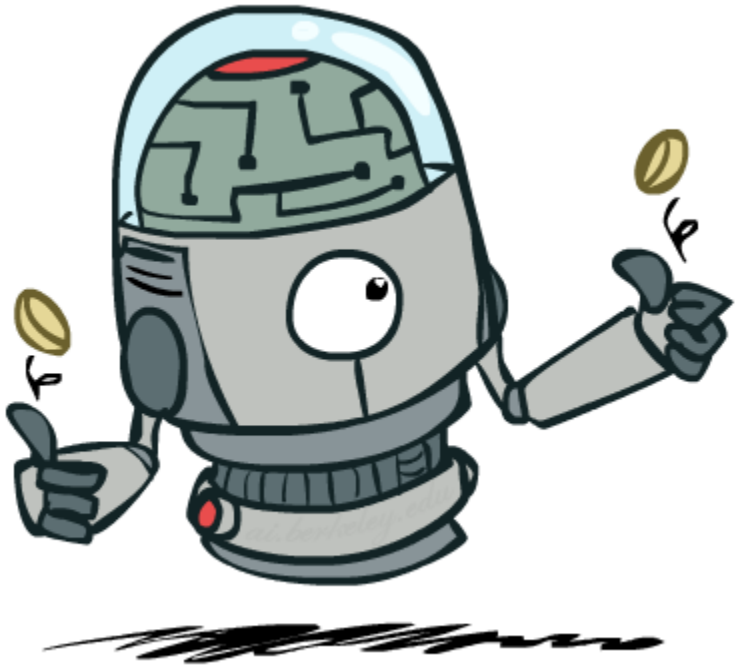
$$\forall x, y : P(x, y) = P(x)P(y)$$

  - 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

$P(W)$

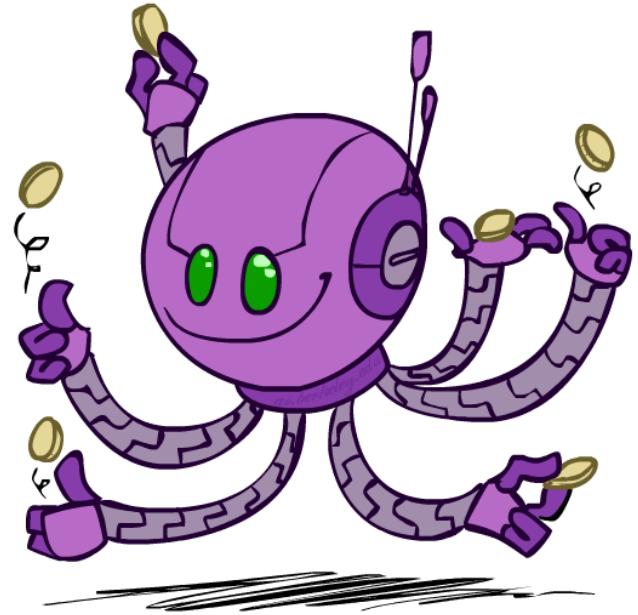
W	P
sun	0.6
rain	0.4

$P_2(T, W)$

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

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

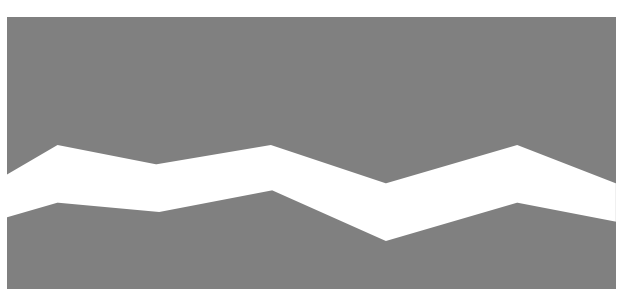
$\ldots$

$P(X_n)$

H	0.5
T	0.5

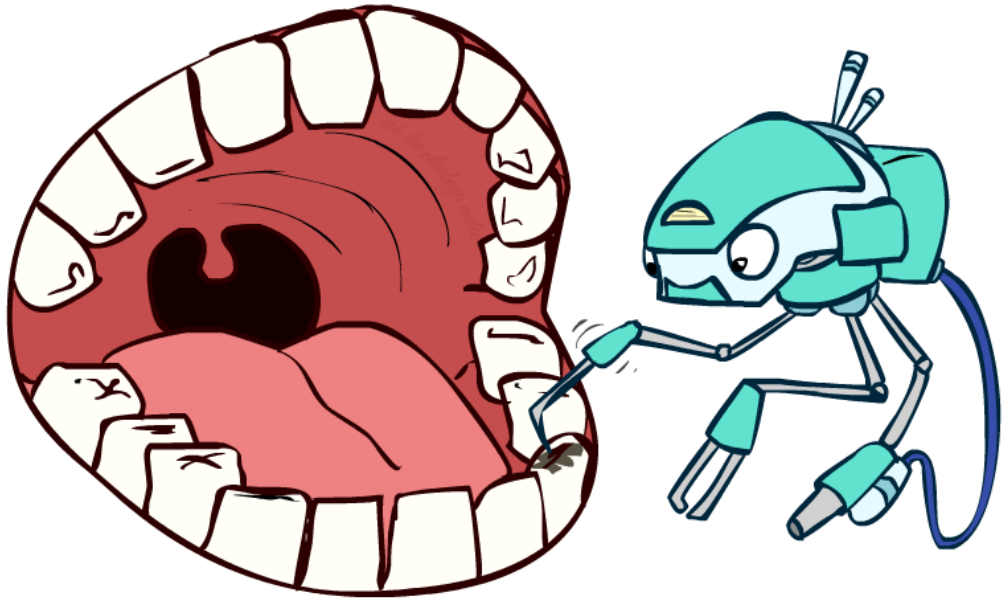
$P(X_1, X_2, \ldots X_n)$

2^n



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



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:

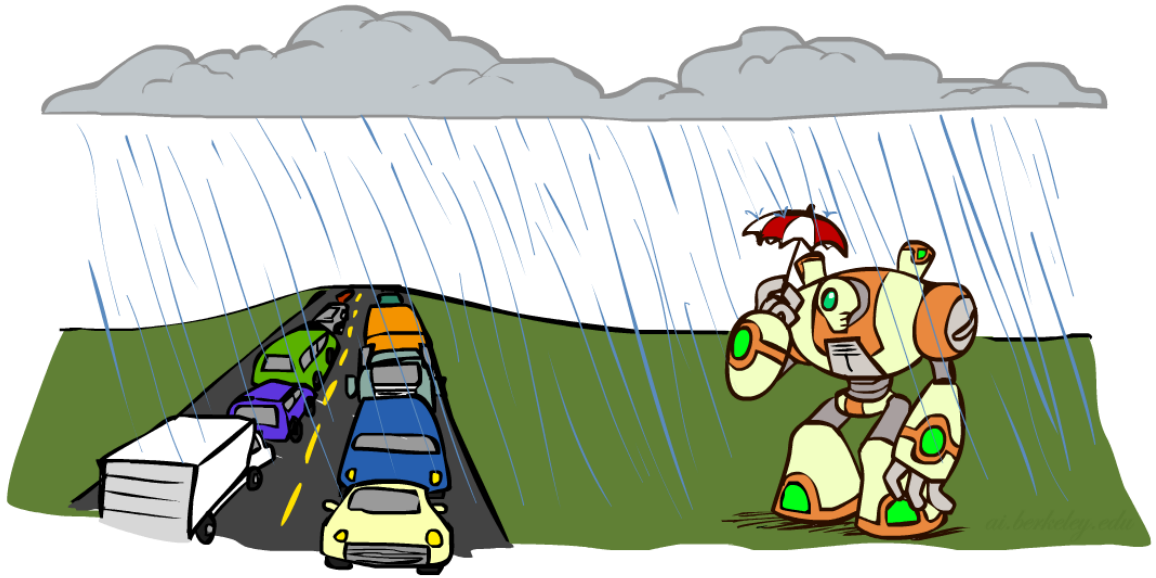
$$\forall x, y, z : P(x, y|z) = P(x|z)P(y|z)$$

or, equivalently, if and only if

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

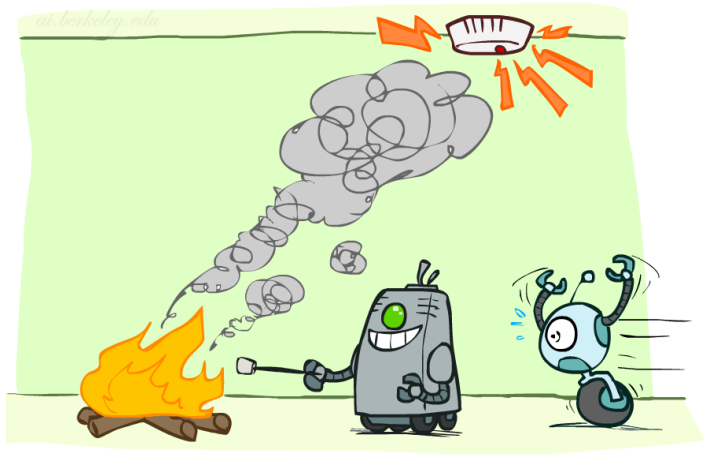
Conditional Independence

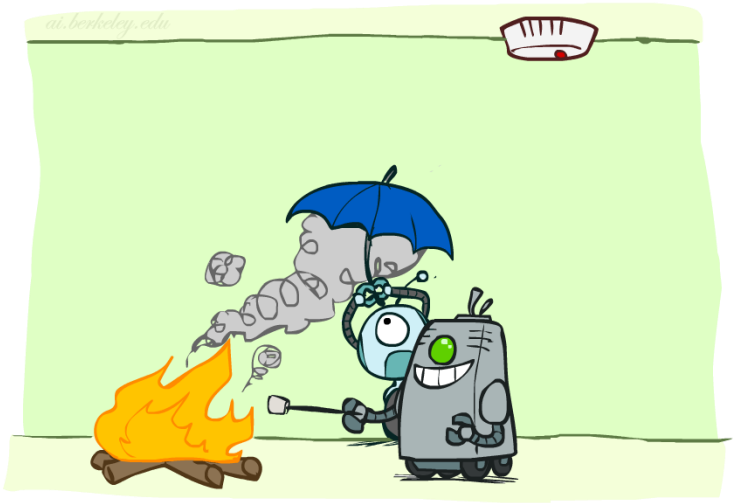
- What about this domain:
  - Traffic
  - Umbrella
  - Raining



Conditional Independence

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



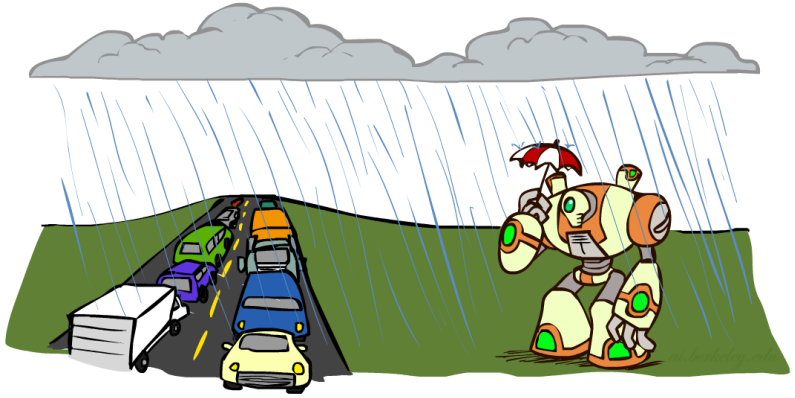


Conditional Independence and the Chain Rule

- 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:

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



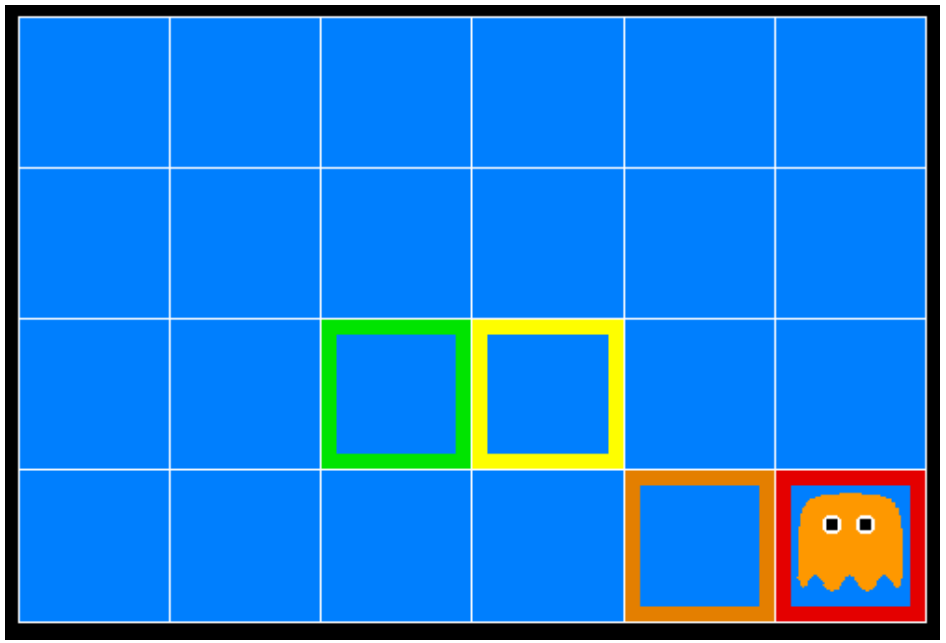
- With assumption of conditional independence:

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

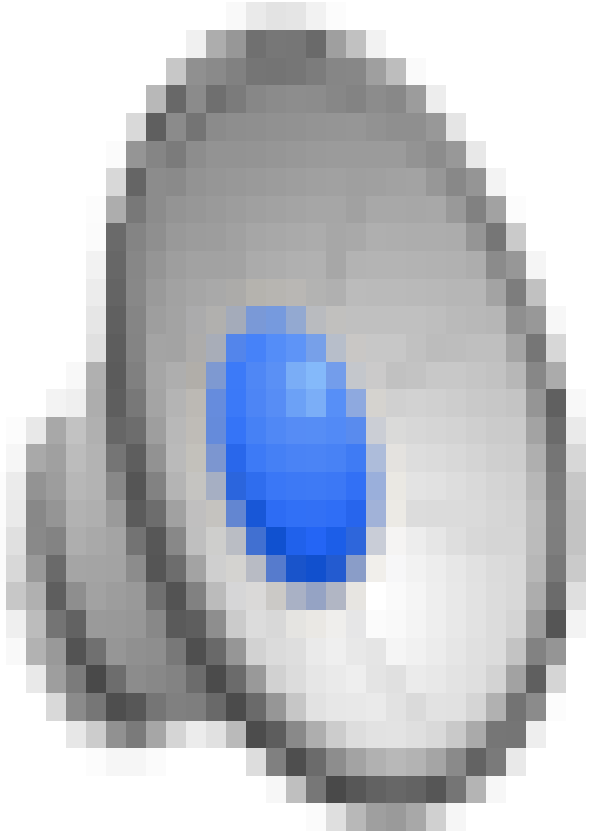
- Bayes'nets / graphical models help us express conditional independence assumptions

Inference in Ghostbusters

- A ghost is in the grid somewhere
- Sensor readings tell how close a square is to the ghost
  - On the ghost: red
  - 1 or 2 away: orange
  - 3 or 4 away: yellow
  - 5+ away: green



Video of Demo Ghostbusters with Probability

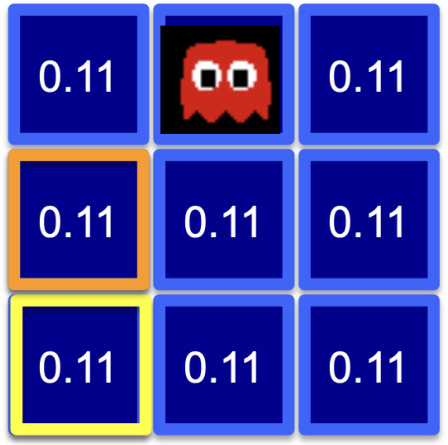


Ghostbusters model

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

- 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.

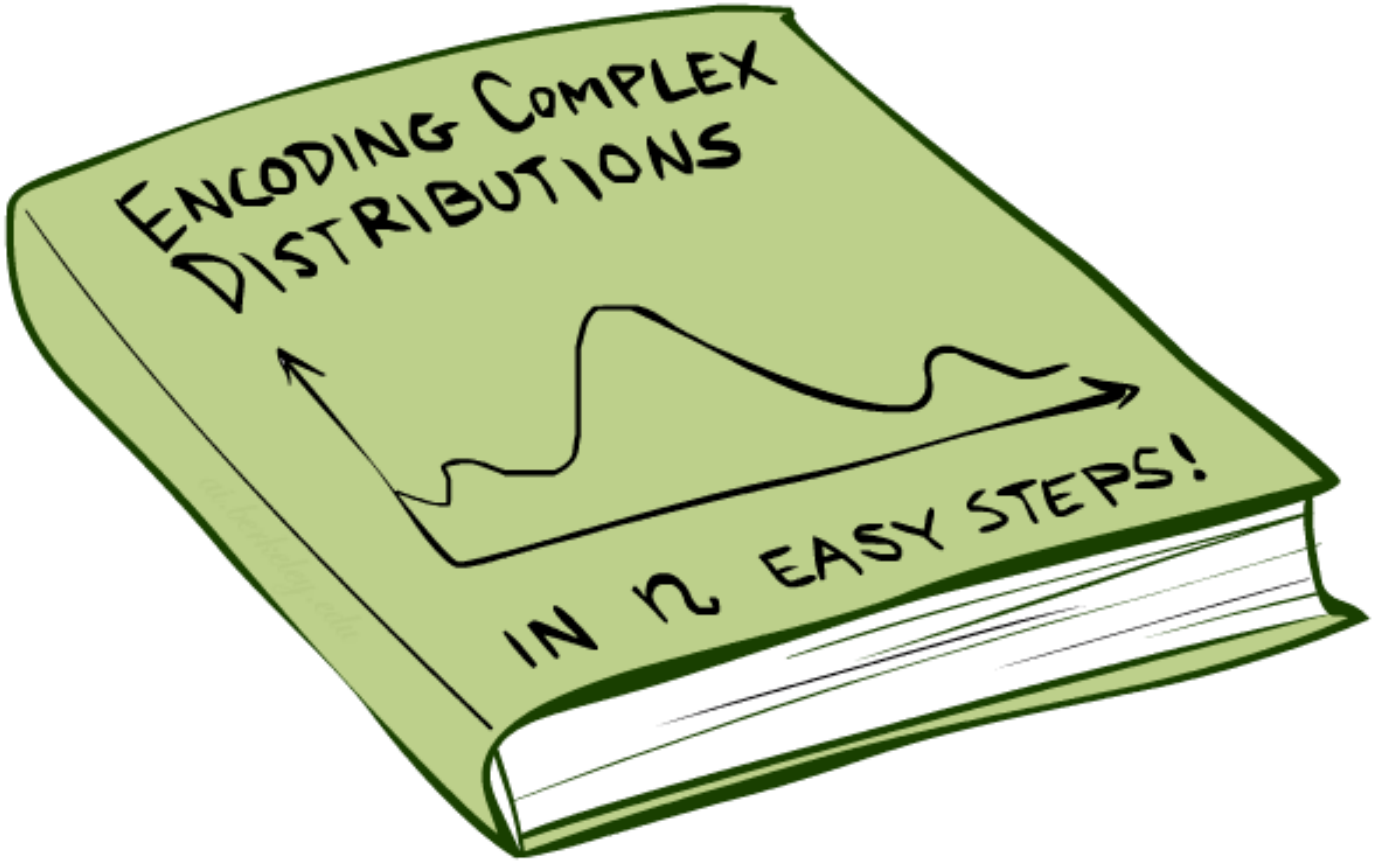


- $P(G, C_{1,1}, \ldots C_{3,3})$ has $9 \times 4^9 = 2{,}359{,}296$ entries!!!
- Ghostbuster independence:
  - Are $C_{1,1}$ and $C_{1,2}$ independent?
    - E.g., does $P(C_{1,1} = \text{yellow}) = P(C_{1,1} = \text{yellow} \mid C_{1,2} = \text{orange})$?
- Ghostbuster physics again:
  - $P(C_{x,y} \mid G)$ ***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

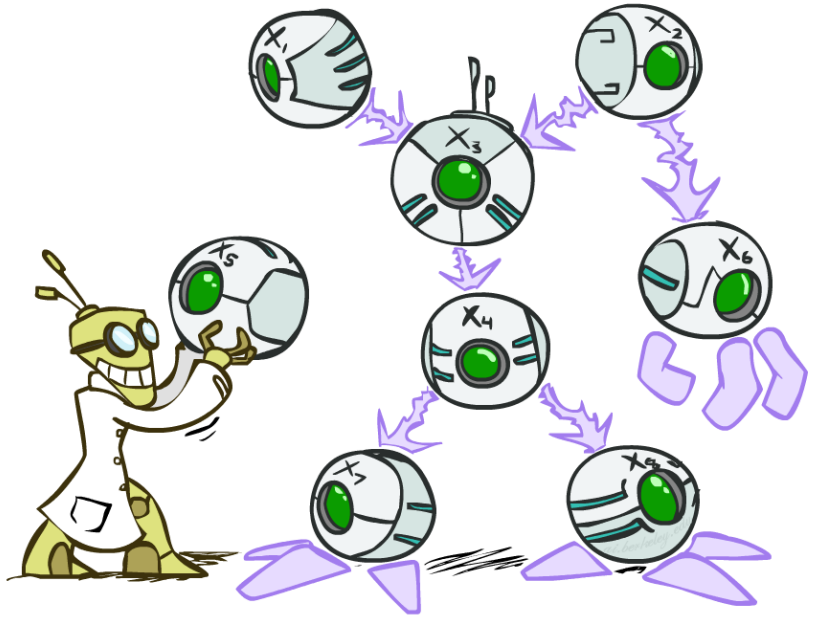
Bayes'Nets: Big Picture



Bayes' Nets: Big Picture

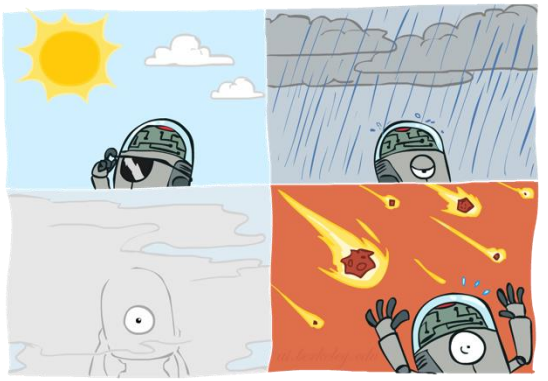
- Two problems with using full joint distribution tables as our probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time

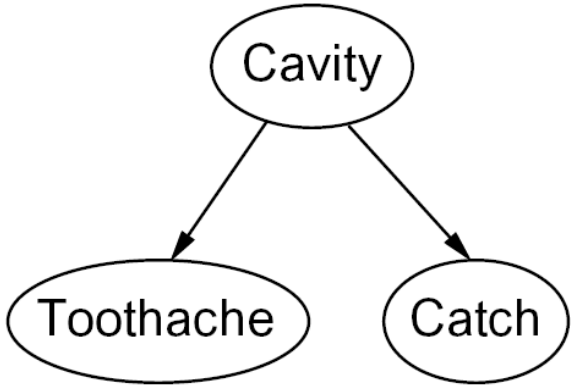
- Bayes' nets: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - More properly called graphical models
  - We describe how variables locally interact
  - Local interactions chain together to give global, indirect interactions
  - For about 10 min, we'll be vague about how these interactions are specified

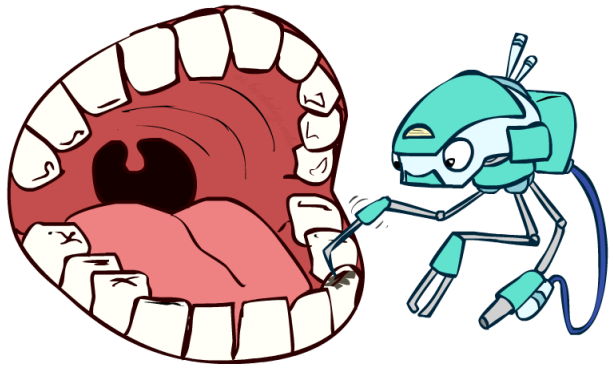


Graphical Model Notation

- 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!)

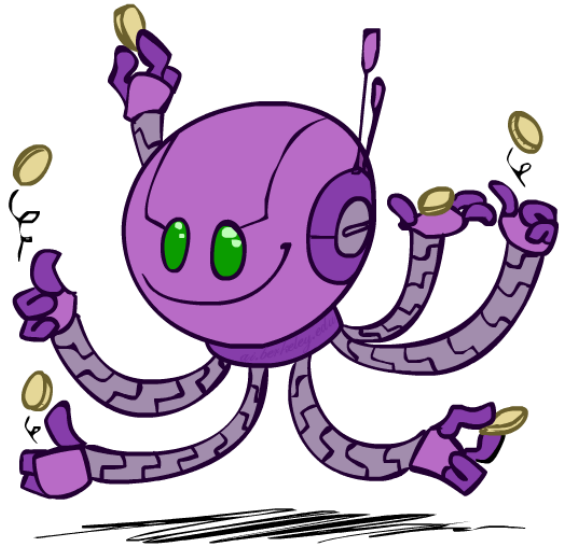






Example: Coin Flips

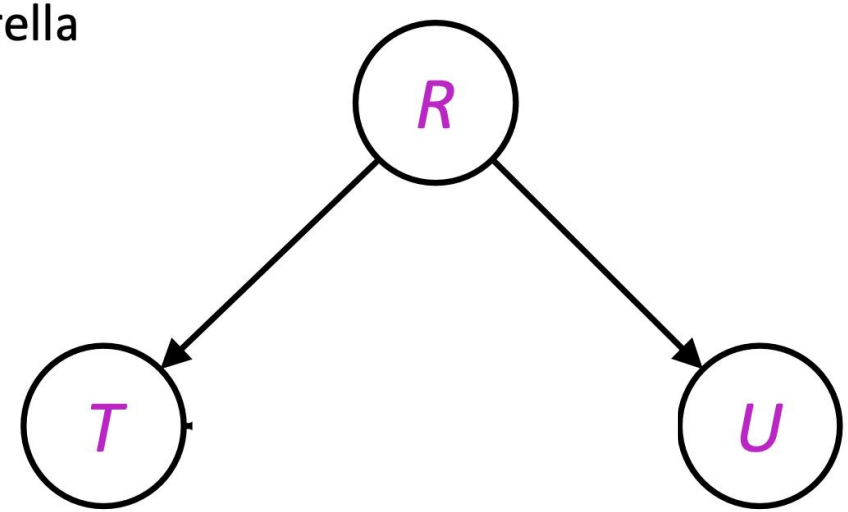
- N independent coin flips



- No interactions between variables: absolute independence

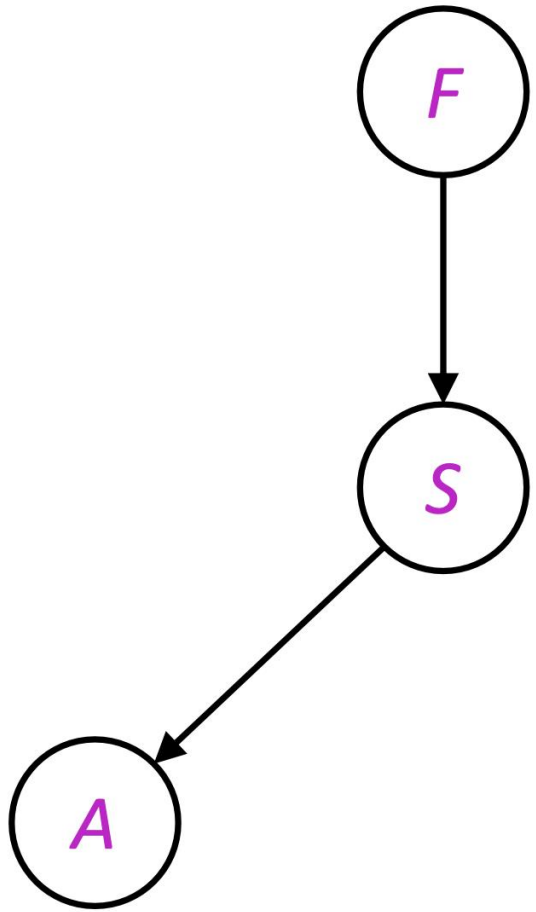
Example: Traffic

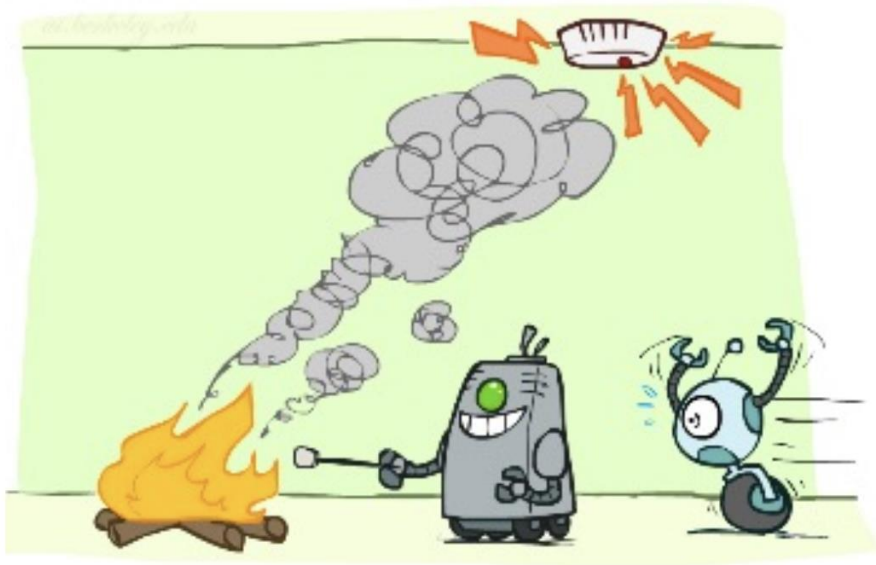
- Variables:
  - ***T***: There is traffic
  - ***U***: I'm holding my umbrella
  - ***R***: It rains



Example: Smoke alarm

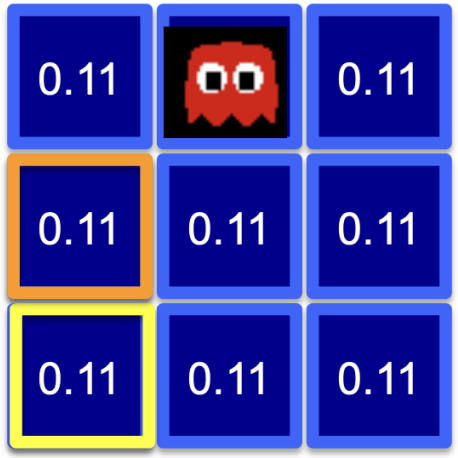
- Variables:
  - ***F***: There is fire
  - ***S***: There is smoke
  - ***A***: Alarm sounds



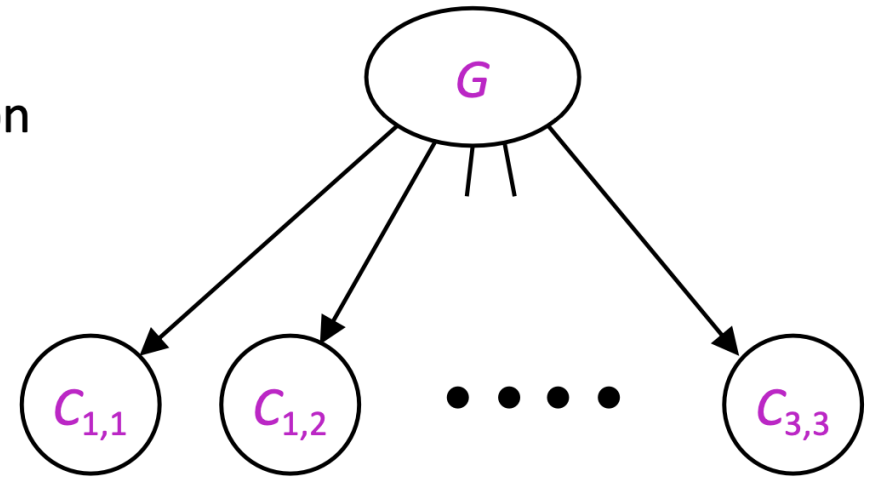


Example: Ghostbusters

- Variables:
  - G: The ghost's location
  - $C_{1,1}$, … $C_{3,3}$:
    The observation at each location

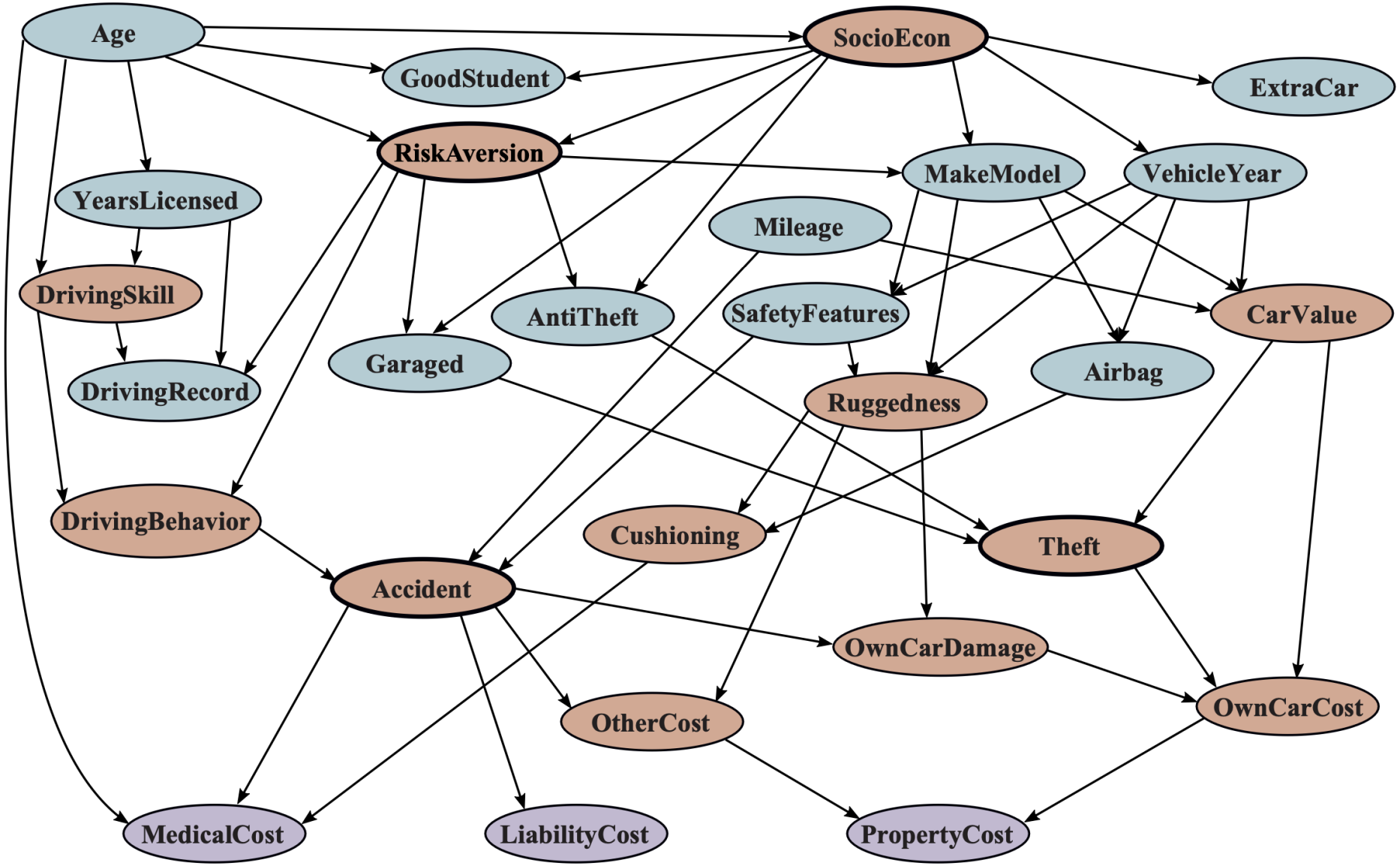


- Want to estimate:
  $P(G \mid C_{1,1} , \ldots C_{3,3})$

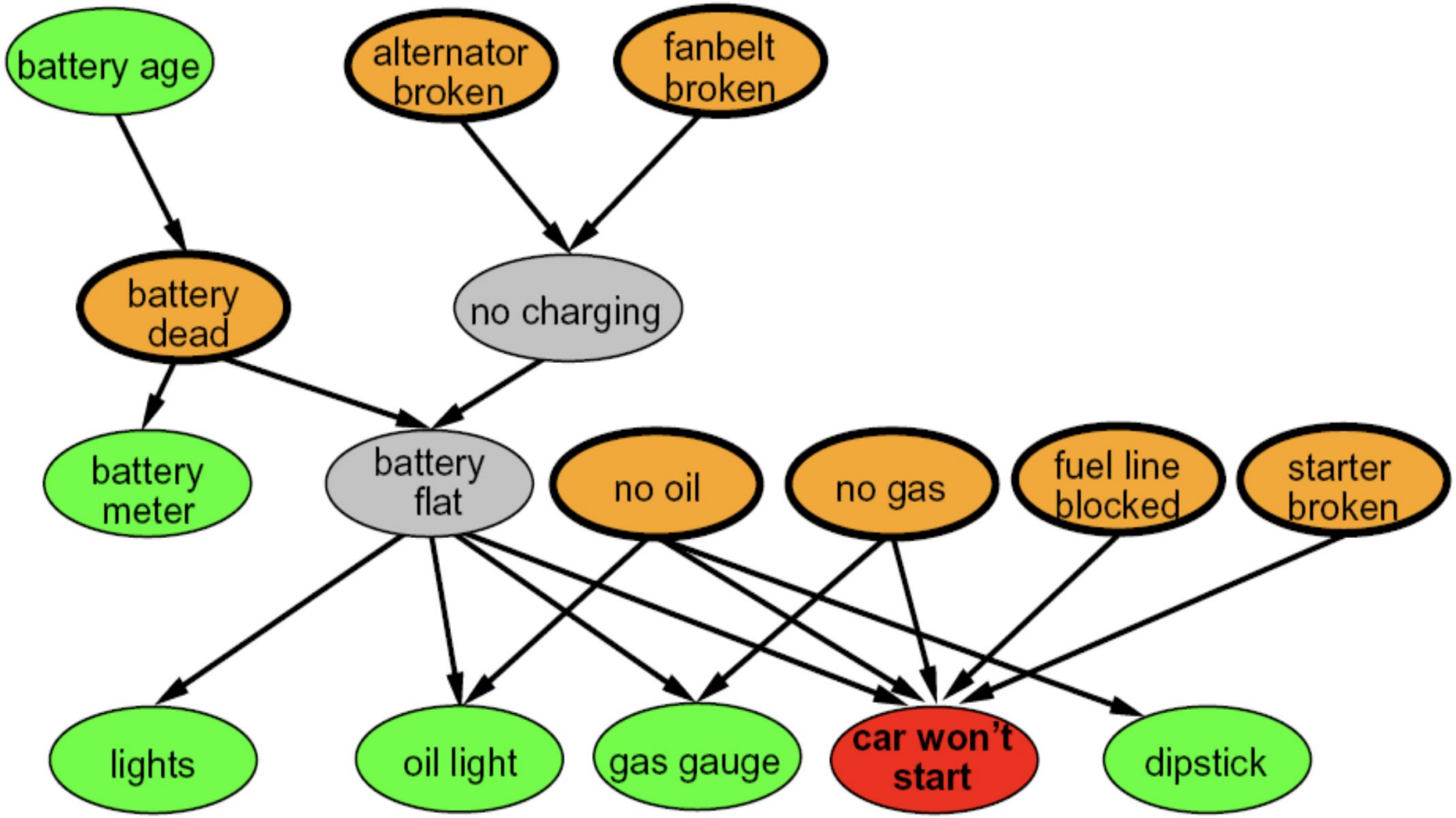


- This is called a ***Naïve Bayes*** model:
  - One discrete query variable (often called the ***class*** or ***category*** variable)
  - All other variables are (potentially) evidence variables
  - Evidence variables are all conditionally independent given the query variable

Example Bayes' Net: Car Insurance



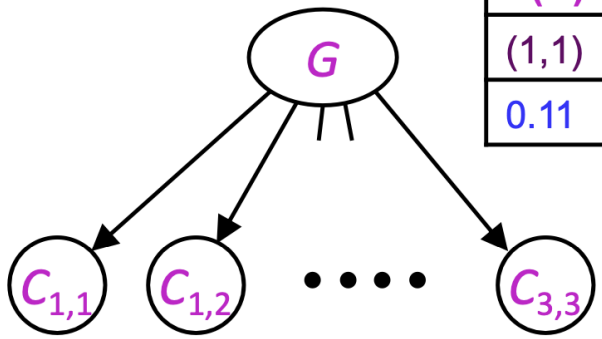
Example Bayes' Net: Car Won't Start



Bayes' Net Semantics

Bayes Net Syntax

- A set of nodes, one per variable X_i
- 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



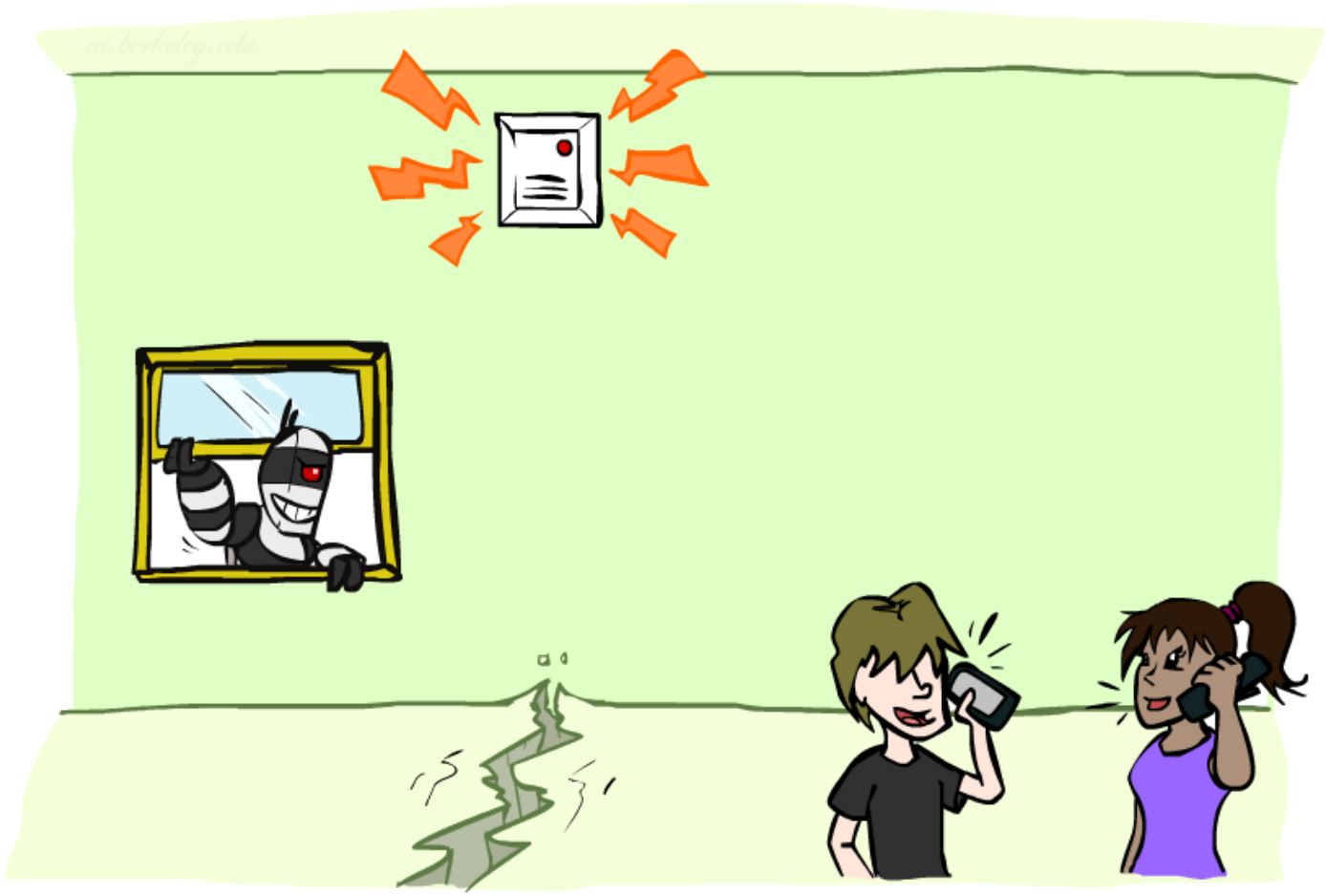
P(G)			
(1,1)	(1,2)	(1,3)	…
0.11	0.11	0.11	…

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
…				

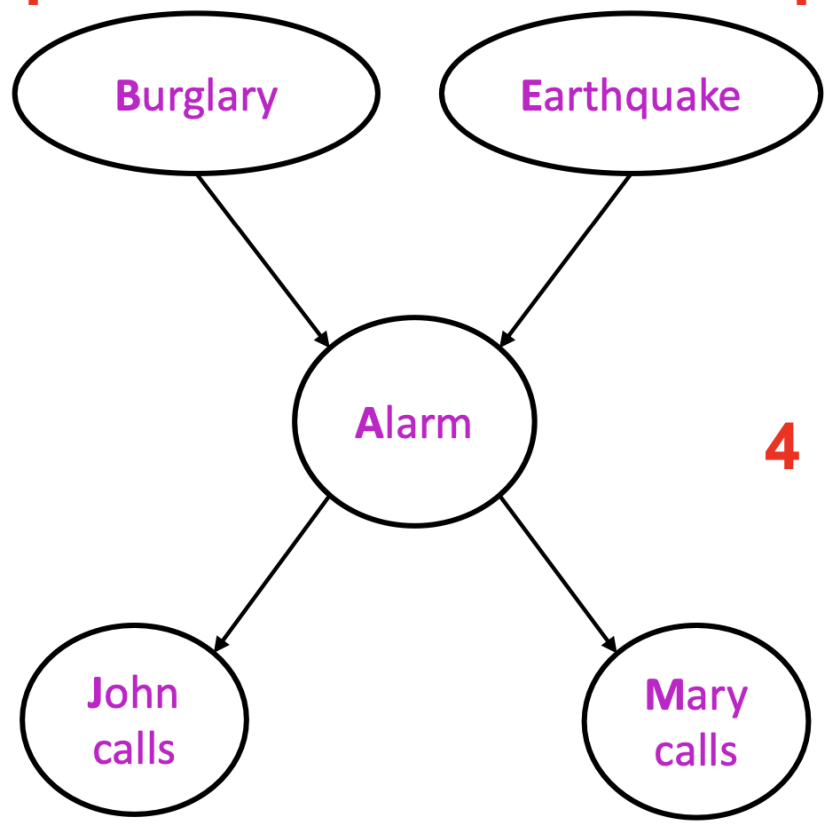
Bayes net = Topology (graph) + Local Conditional Probabilities

Example: Alarm Network

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



Example: Alarm Network



P(B)	
true	false
0.001	0.999

1

P(E)	
true	false
0.002	0.998

1

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

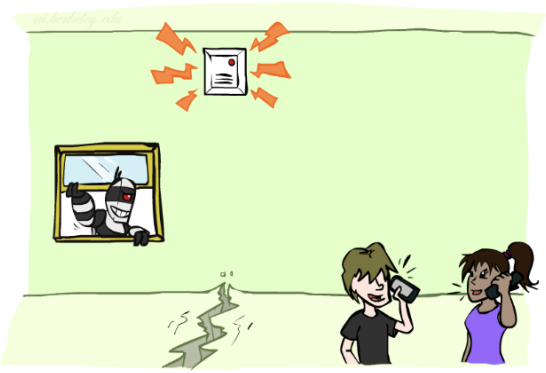
4

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

2

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

2



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

$(d-1)\ \prod_i d_i$

Bayes net global semantics

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

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

- Exploits sparse structure: number of parents is usually small

Size of a Bayes Net

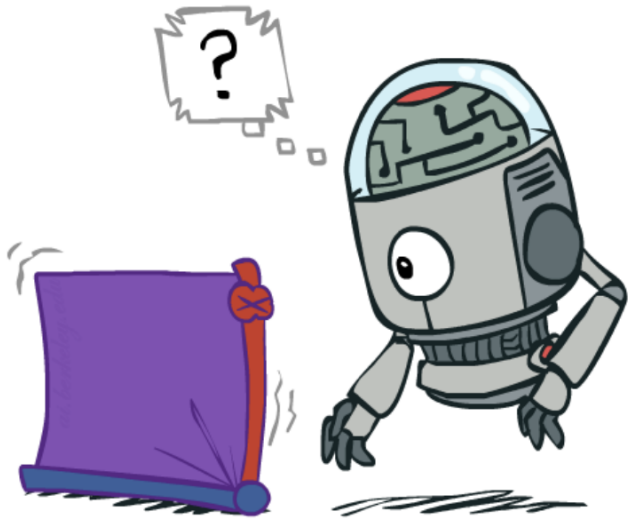
- How big is a joint distribution over N variables, each with d values?

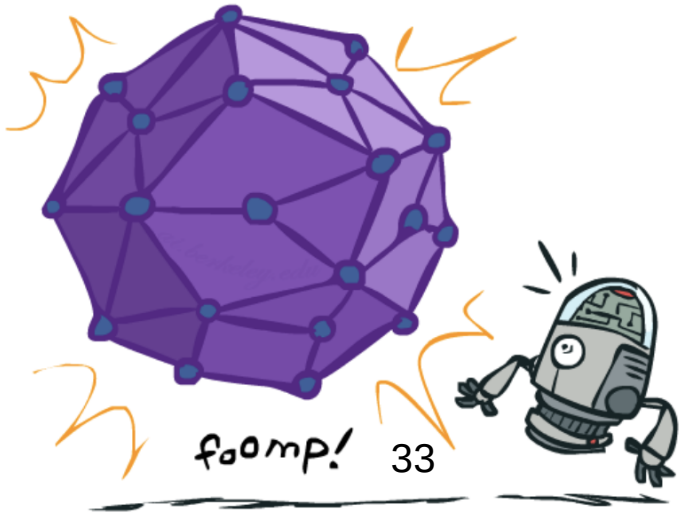
  d^N

- How big is an N-node net if nodes have at most k parents?

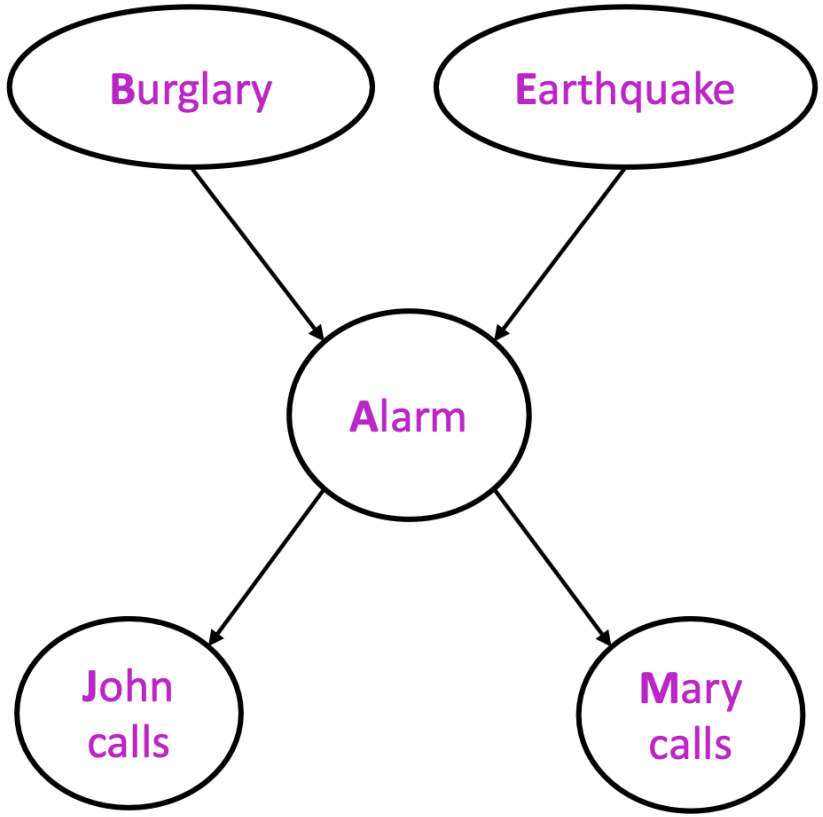
  $O(N * d^k)$

- Both give you the power to calculate $P(X_1, X_2, \ldots, X_N)$
- Bayes Nets: huge space savings with sparsity!
- Also easier to elicit local CPTs
- Also faster to answer queries (coming)





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

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

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

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

$=.001x.998x.94x.1x.3=.000028$

Conditional independence in BNs

- Compare the Bayes net global semantics

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

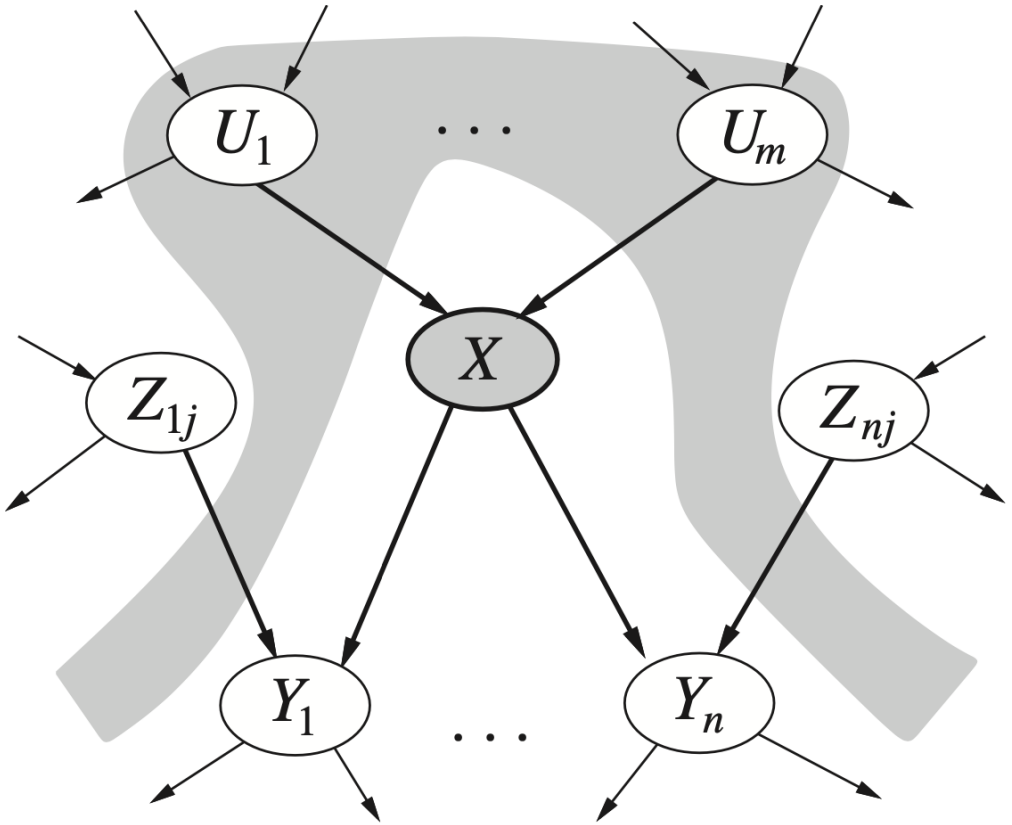
with the chain rule identity

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

- Assume (without loss of generality) that $X_1,..,X_n$ sorted in topological order according to the graph (i.e., parents before children), so $Parents(X_i) \subseteq X_1,\ldots,X_{i-1}$
- So 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

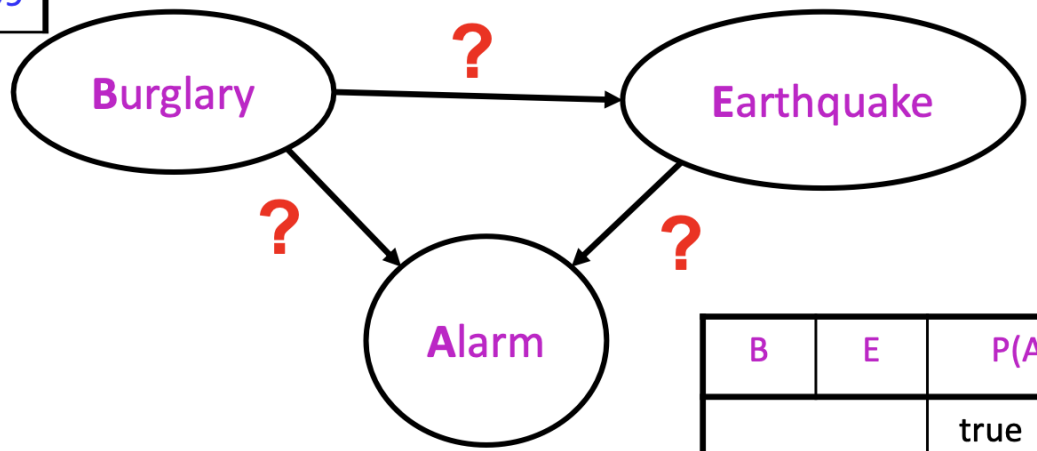
- ***Every variable is conditionally independent of its non-descendants given its parents***
- Conditional independence semantics <=> global semantics



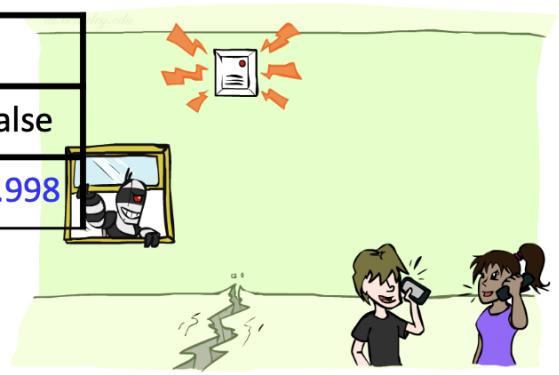
Example: Burglary

- Burglary
- Earthquake
- Alarm

P(B)	
true	false
0.001	0.999



P(E)	
true	false
0.002	0.998

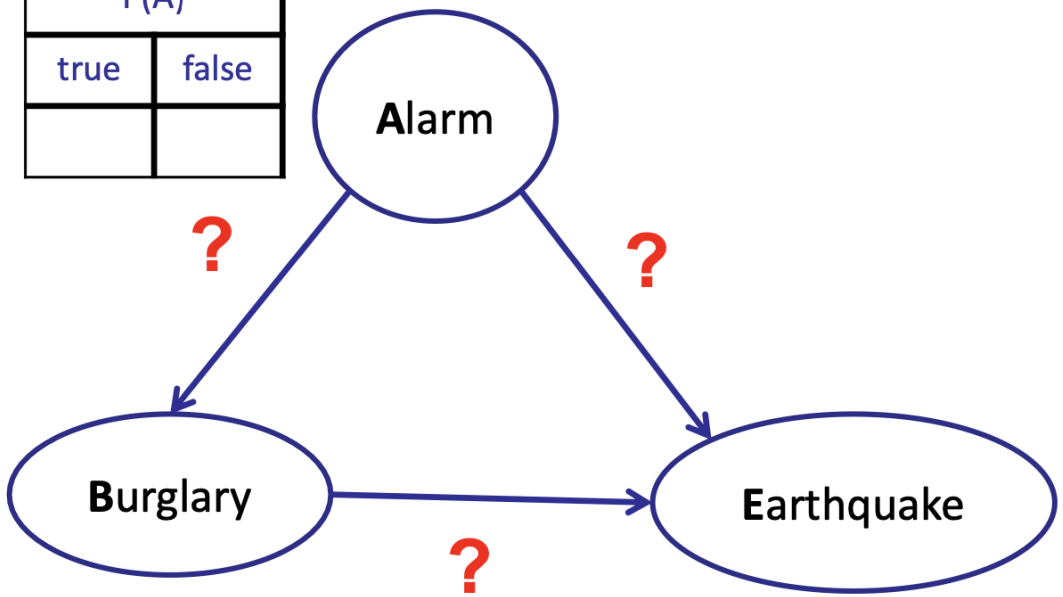


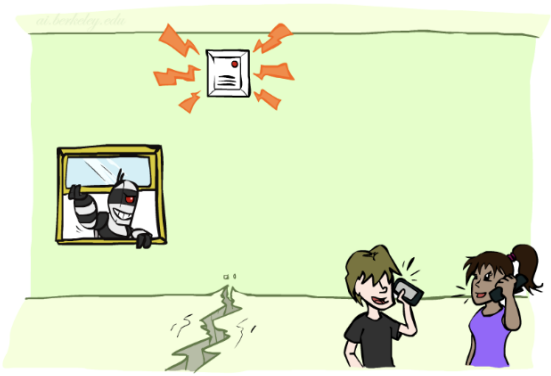
B	E	P(A|B,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999

Example: Burglary

- Alarm
- Burglary
- Earthquake

P(A)	
true	false

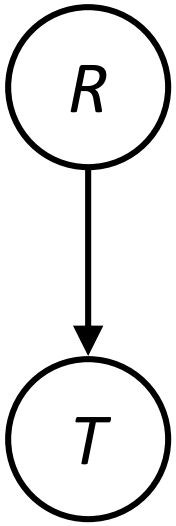




A	P(B|A)	
	true	false
true	?	
false		

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

Example: Traffic



$P(R)$

+r	1/4
-r	3/4

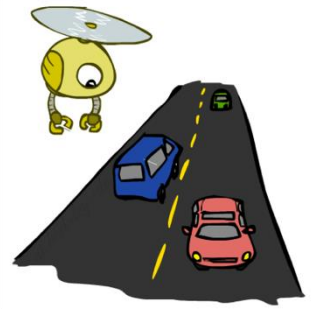
$P(T|R)$

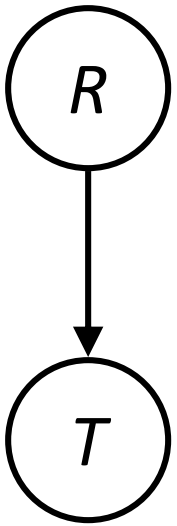
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

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

Example: Traffic

- Causal direction





$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

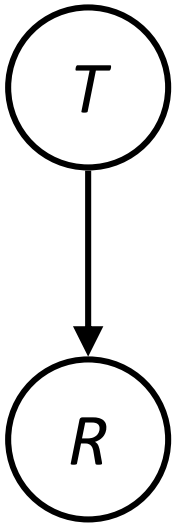
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(T, R)$

+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

$P(R|T)$

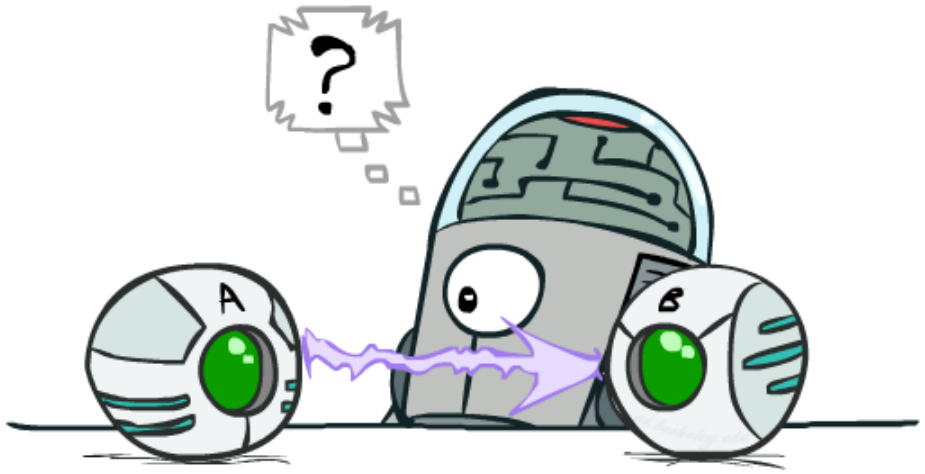
+t	+r	1/3
	-r	2/3
-t	+r	1/7
	-r	6/7

$P(T, R)$

+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



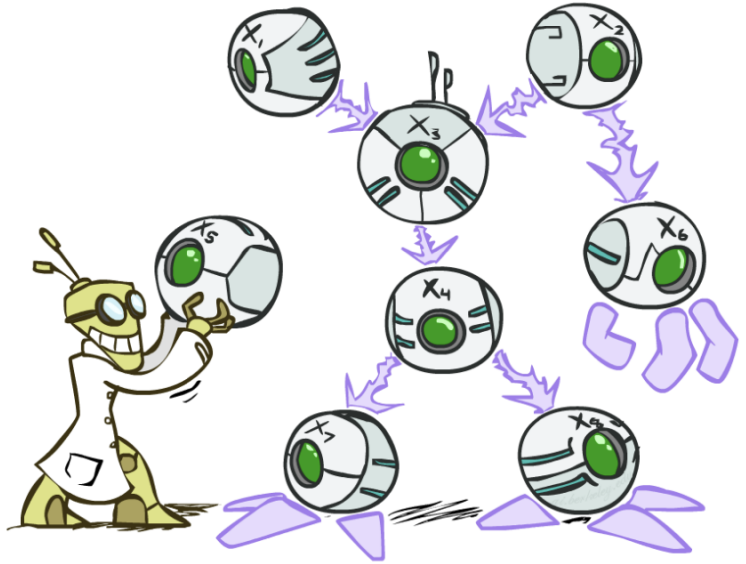
- BNs need not actually be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - E.g. consider the variables *Traffic* and *Drips*
  - End up with arrows that reflect correlation, not causation

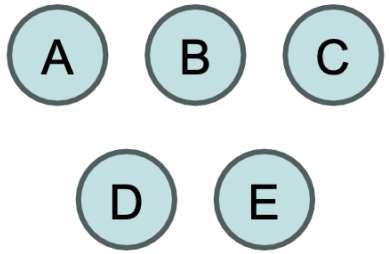
- What do the arrows really mean?
  - Topology may happen to encode causal structure
  - Topology really encodes conditional independence

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

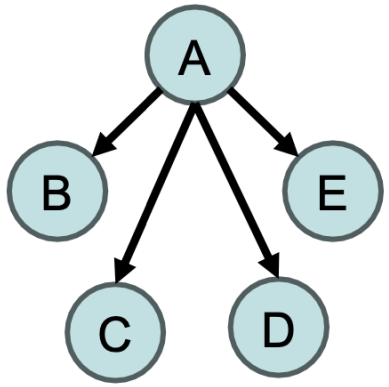
Summary

- 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
- Allows for flexible tradeoff between model accuracy and memory/compute efficiency

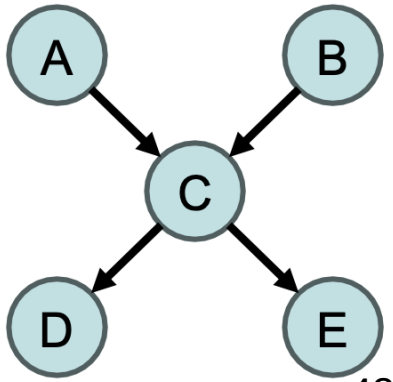




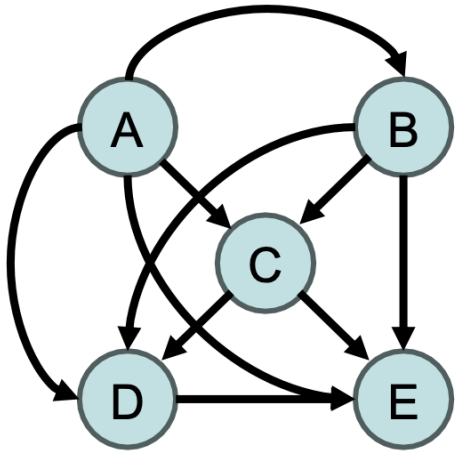
Strict Independence



Naïve Bayes



Sparse Bayes Net



Joint Distribution