Recap: Probability Basics

- Random variable X with a range of values
  - Some aspect of the world with uncertainty
- Distribution $P(X)$ *(with outcome unobserved)*
  - gives probability for each possible value x
- Joint distribution $P(X,Y)$ *(multiple interacting variables)*
  - gives probability for each combination of values x,y
- Basic laws: $0 \leq P(x) \leq 1$ $\quad \sum_{x \in X} P(x) = 1$
- Events: subsets of outcomes: $P(E) = \sum_{x \in E} P(x)$
- Marginal distribution: $P(x) = \sum_y P(x,y)$ *(marginalization/summing out)*
- Conditional distribution: $P(x|y) = P(x,y)/P(y)$
- Product rule: $P(x|y)P(y) = P(x,y) = P(y|x)P(x)$
- Chain rule: $P(x_1,..,x_n) = \prod_i P(x_i \mid x_1,..,x_{i-1})$ *(repeated application of product rule)*

Independence

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

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

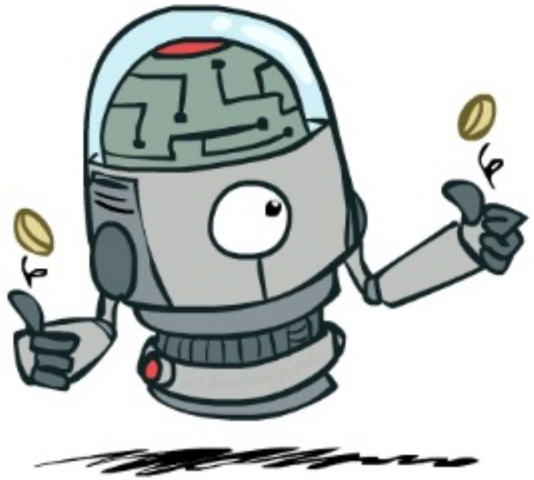
  - I.e., the joint distribution *factors* into a product of two simpler distributions

- Equivalently, via the product rule $P(x,y) = P(x|y)P(y)$,

$$P(x \mid y) = P(x) \quad \text{or} \quad P(y \mid x) = P(y)$$

- We write: $X \perp\!\!\!\perp Y$

- Independence is a simplifying *modeling assumption*
  - In real: joints are at best “close” to independent
  - What could we assume for {Weather, Traffic, Cavity, Toothache}?



Quiz: Independence?

$\forall t, w \qquad P(t, w) = P(t)\, P(w)$

$P(T)$

T	P
hot	0.5
cold	0.5

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

$P_1(T, W)$

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

$P_2(T, W)$

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

$P(W)$

W	P
sun	0.6
rain	0.4

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

Example: Independence

- *n* fair, independent coin flips:

$P(X_1)$

H	0.5
T	0.5

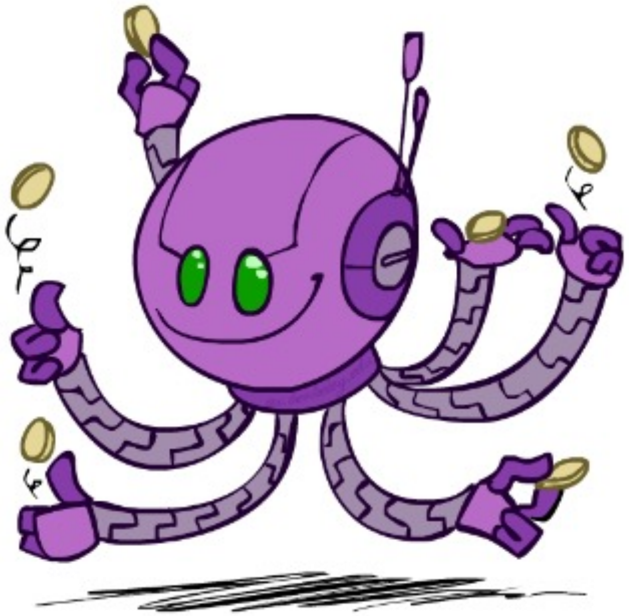
$P(X_2)$

H	0.5
T	0.5

. . .

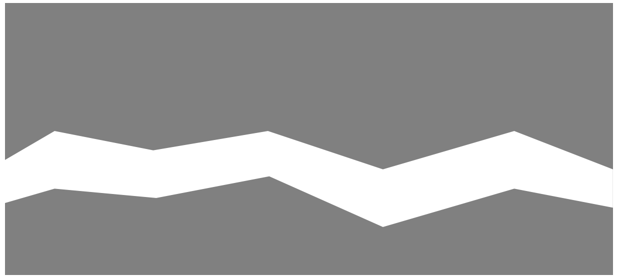
$P(X_n)$

H	0.5
T	0.5



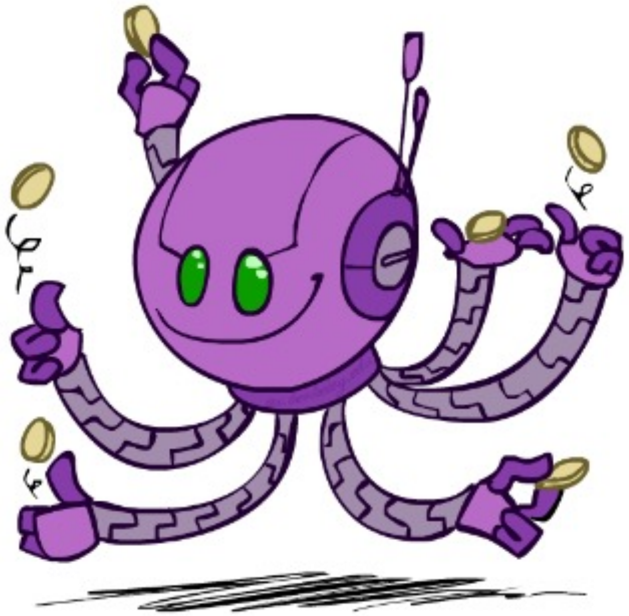
$P(X_1, X_2, \ldots, X_n)$

$2^n \left\{ \right.$

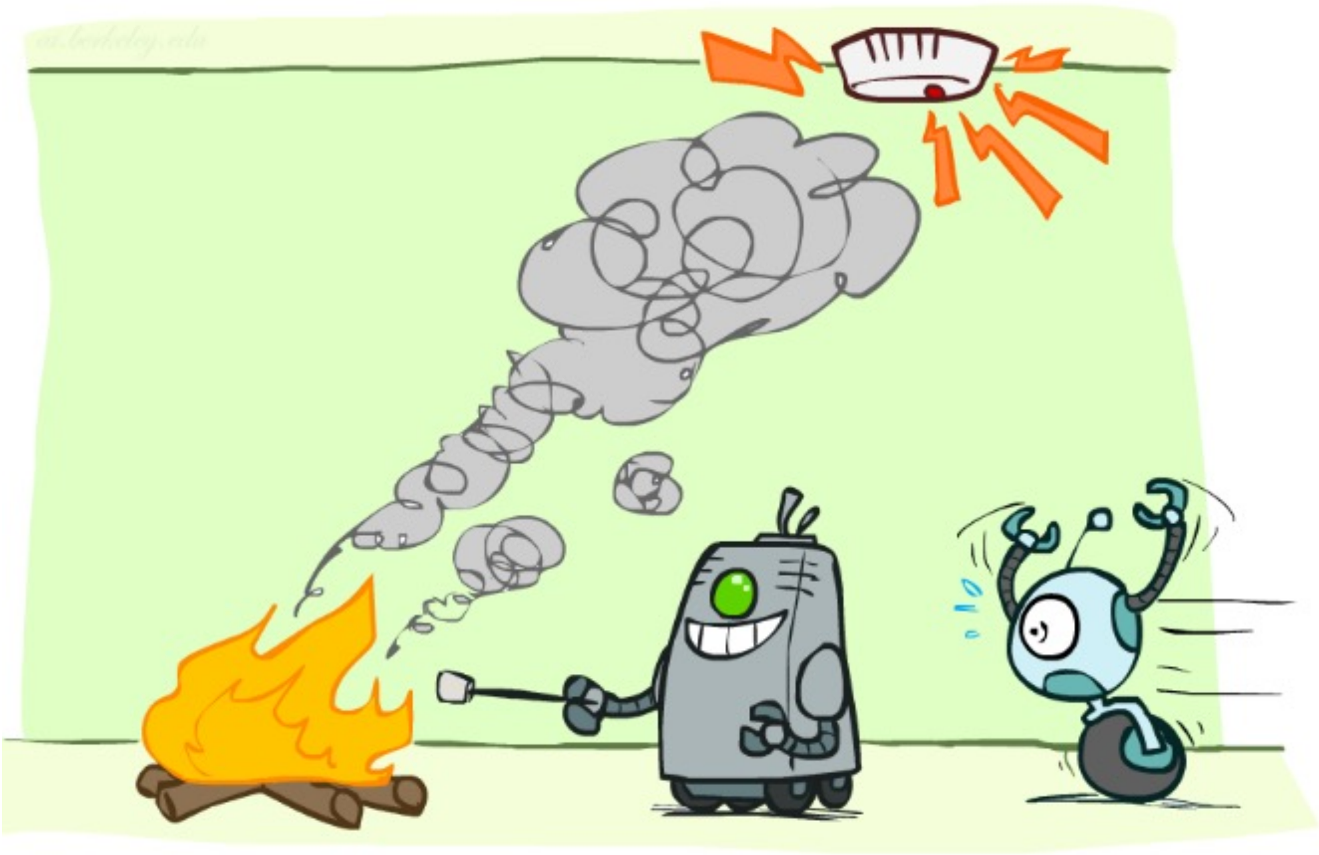


Independence, contd.

- Independence is incredibly powerful
  - Exponential reduction in representation size
- Independence is extremely rare!
- *Conditional* independence is much more common!!

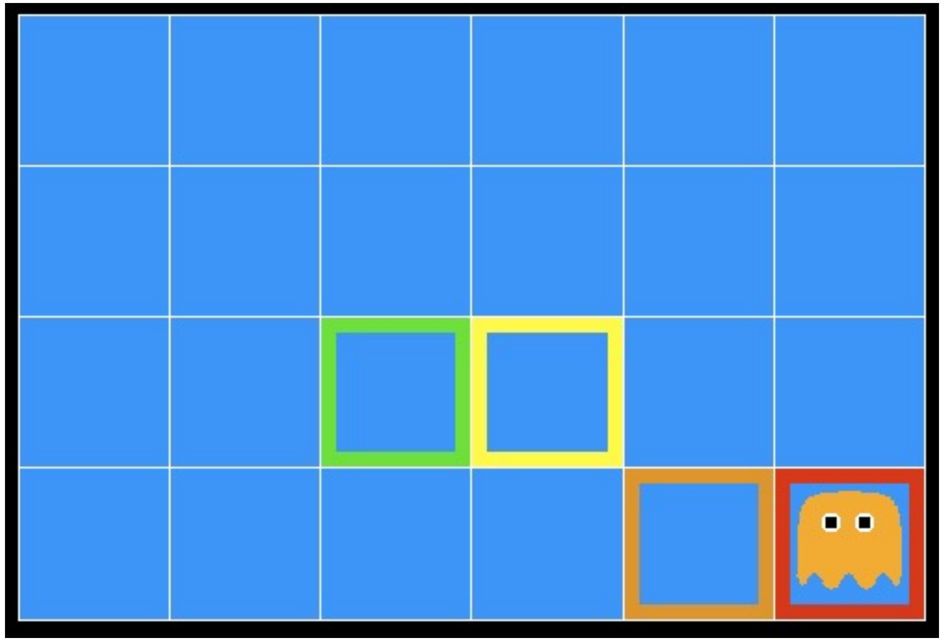


Conditional Independence



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***"



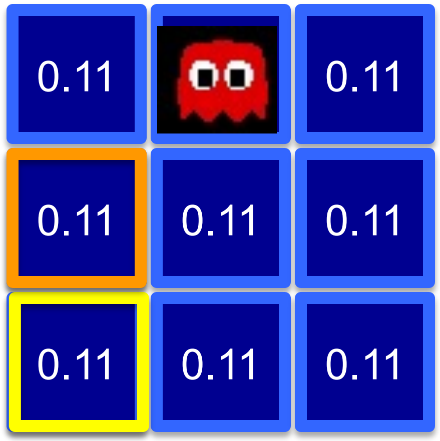
Video of Demo Ghostbusters with Probability

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}
- We have two distributions at hand:
  - *Prior distribution* over ghost location: $P(G)$
    - Let's say this is uniform
  - *Sensor model*: $P(C_{x,y} \mid G)$
    - Let's say it depends only on distance to G
      - E.g. $P(C_{1,1} = \text{red} \mid G = (1,1)) = 0.6$
      - $P(C_{1,1} = \text{orange} \mid G = (1,1)) = 0.25$
      - …

0.11	0.11	0.11
0.11	0.11	0.11
0.11	0.11	0.11

Ghostbusters model, contd.



- Joint distribution $P(G, C_{1,1}, \ldots C_{3,3})$
  - has $9 \times 4^9 = 2{,}359{,}296$ entries!!!
- Ghostbuster independence:
  - Are $C_{1,1}$ and $C_{1,2}$ independent?
    - i.e., does $P(C_{1,1} = \text{yellow}) = P(C_{1,1} = \text{yellow} \mid C_{1,2} = \text{orange})$?
  - No.
- What if G is known?
  - Are $C_{1,1}$ and $C_{1,2}$ still dependent?
  - Sensor model $P(C_{x,y} \mid G)$ depends *only* on distance to G
    - So $P(C_{1,1} = \text{yellow} \mid \underline{G = (2,3)}\,) = P(C_{1,1} = \text{yellow} \mid \underline{G = (2,3)}, C_{1,2} = \text{orange})$
    - I.e., $C_{1,1}$ is *conditionally independent* of $C_{1,2}$ *given G*

Ghostbusters model, contd.

- Simplify the model using the conditional independence?
- 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

9×4^9 $\qquad$ $9 + 4*9^2$

Conditional Independence

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

- Formally, *X* is conditionally independent of *Y* given *Z* if and only if:

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

or, equivalently, if and only if

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

we write: $X \perp\!\!\!\perp Y | Z$

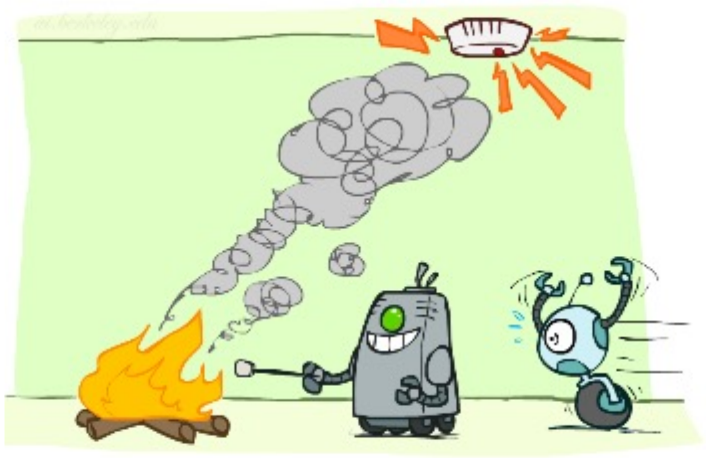
Example: Conditional Independence

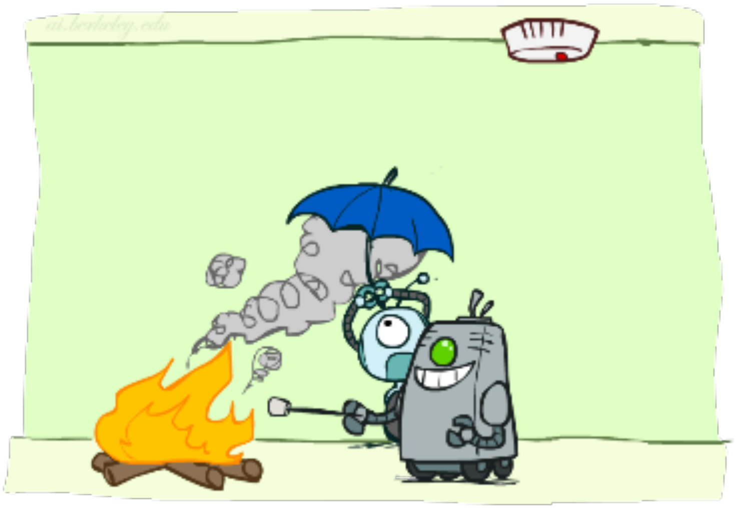
- Domain:
  - **Traffic**: there is heavy traffic
  - **Umbrella**: someone holding umbrella
  - **Raining**: it is raining
- ***Any conditional independence to spot out?***

Conditional Independence

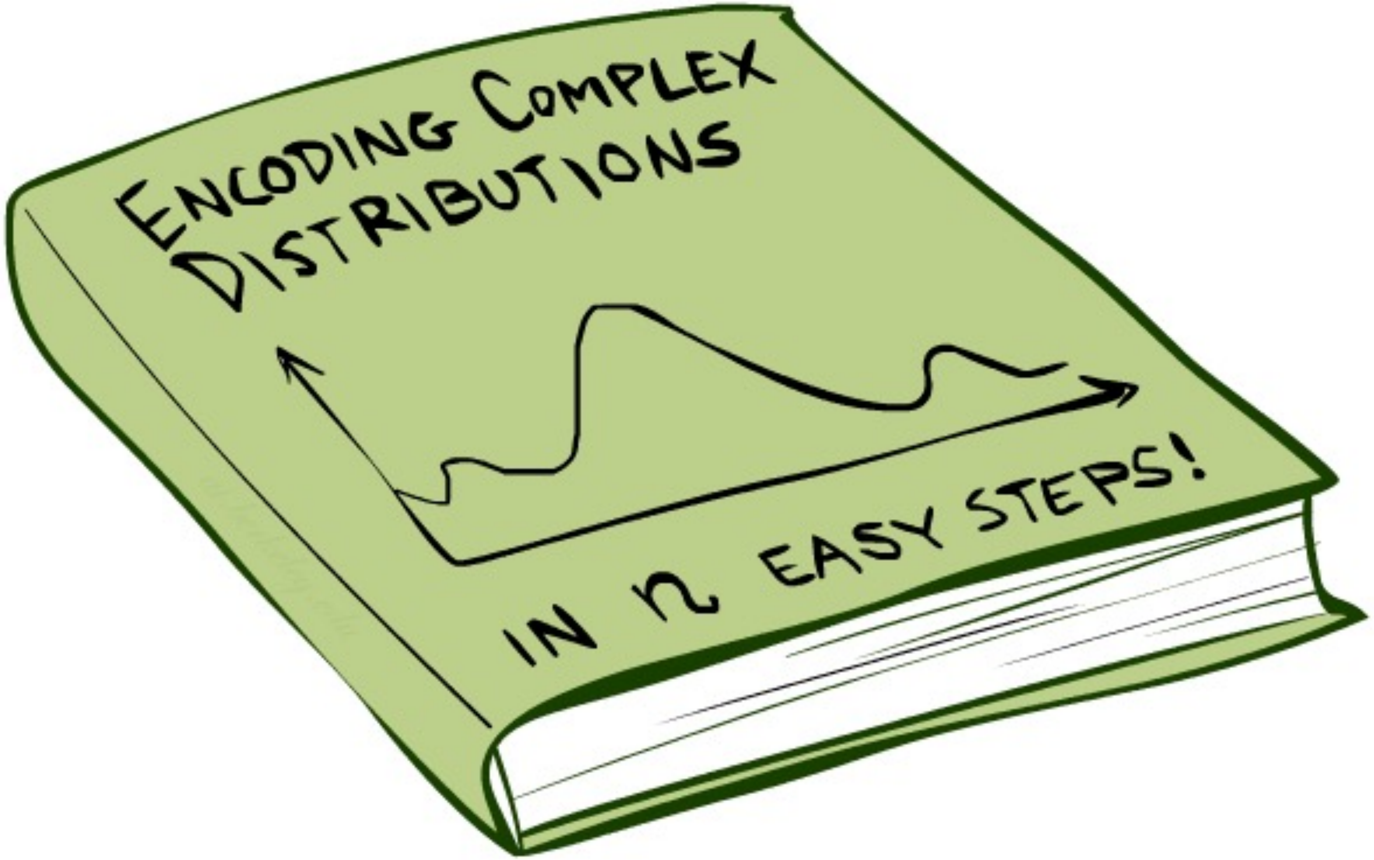
- Domain:
  - **Fire**: there is fire
  - **Smoke**: there is smoke
  - **Alarm**: alarm rings

- ***Any conditional independence to spot out?***





Bayes Nets

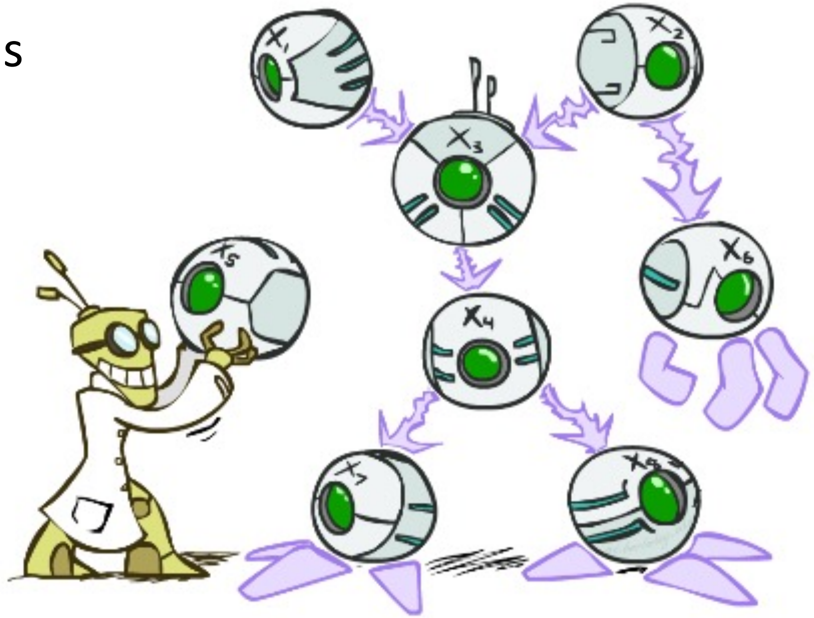


Bayes Nets: Big Picture

- **Bayes nets:** a technique for describing *complex joint distributions* (models) using *simple, conditional distributions*

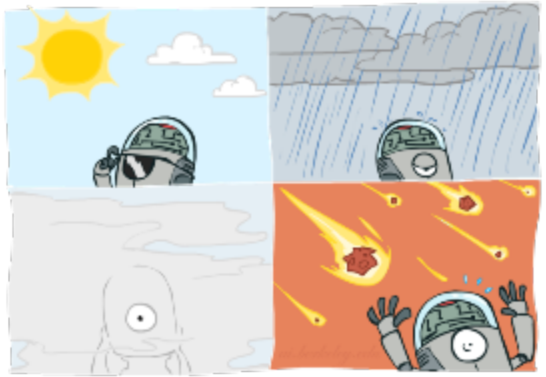
- More properly called graphical models
  - The world is composed of many variables
  - Each variable only interacts *locally* with a few others
  - Local interactions chain together to give *global*, *indirect* interactions
  - Use *local* conditional distributions to represent *global* joint distributions

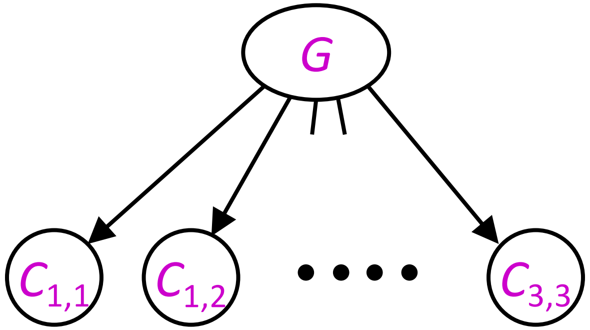
- In coming sessions:
  - Representation
  - Exact inference
  - Approximate inference

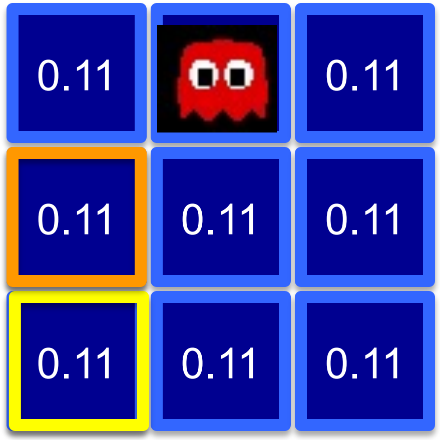


Graphical Model Notation

- Nodes: variables (with ranges)
  - Can be assigned (observed) or unassigned (unobserved)
- Arcs: interactions
  - *Presence* of arcs indicate “direct influence” between variables
  - Formally: *absence* of arcs encodes conditional independence (more later)

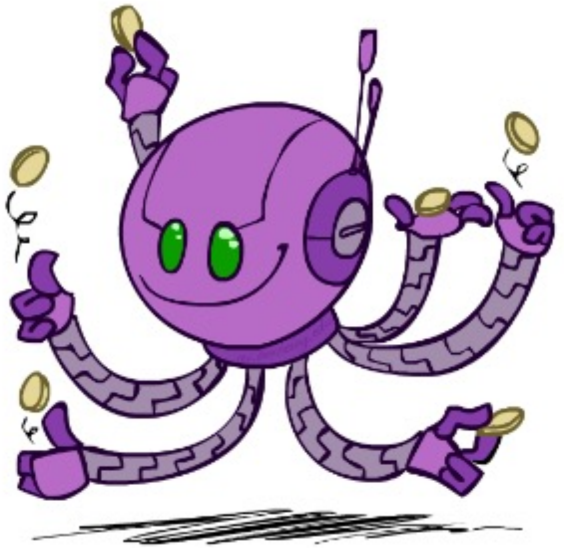






Example: Coin Flips

- *n* independent coin flips

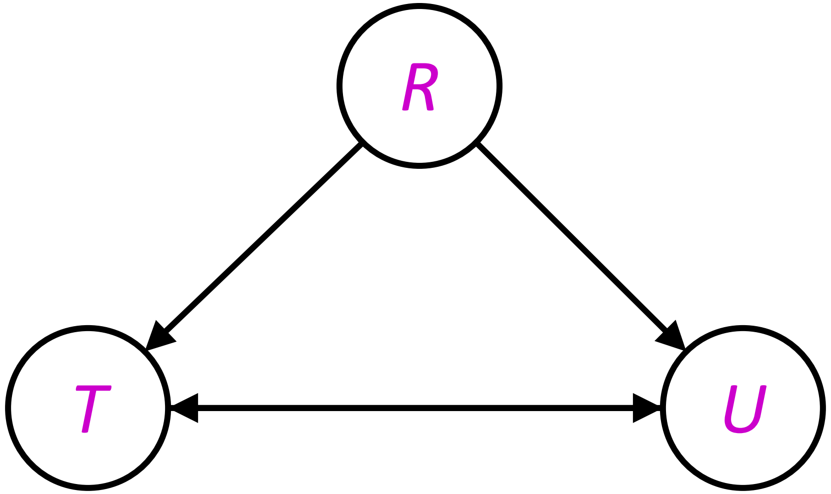


X_1 X_2 $\cdots$ X_n

- No interactions between variables: absolute independence

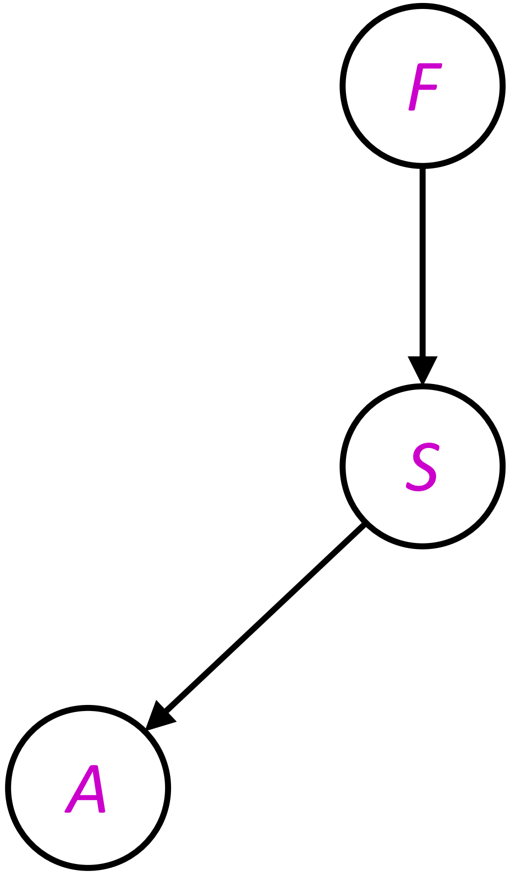
Example: Traffic

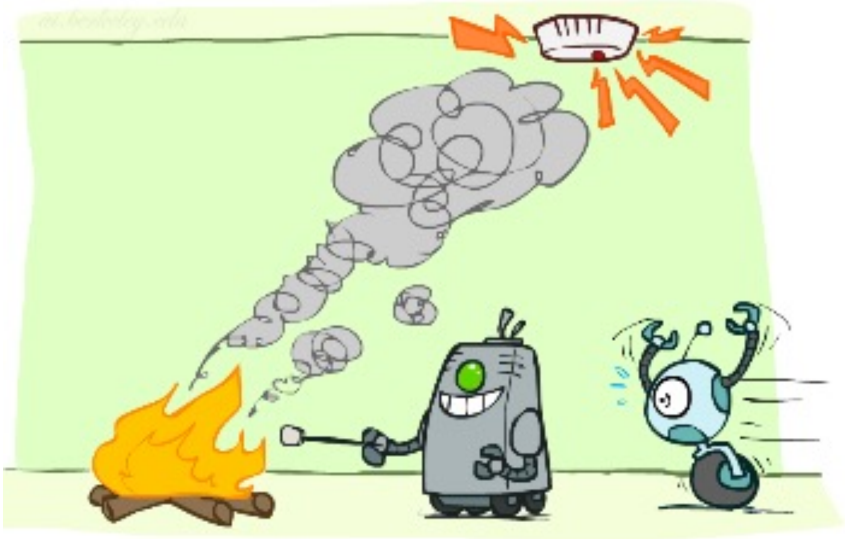
- Variables:
  - T: There is heavy traffic
  - U: I’m holding my umbrella
  - R: It rains



Example: Smoke alarm

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

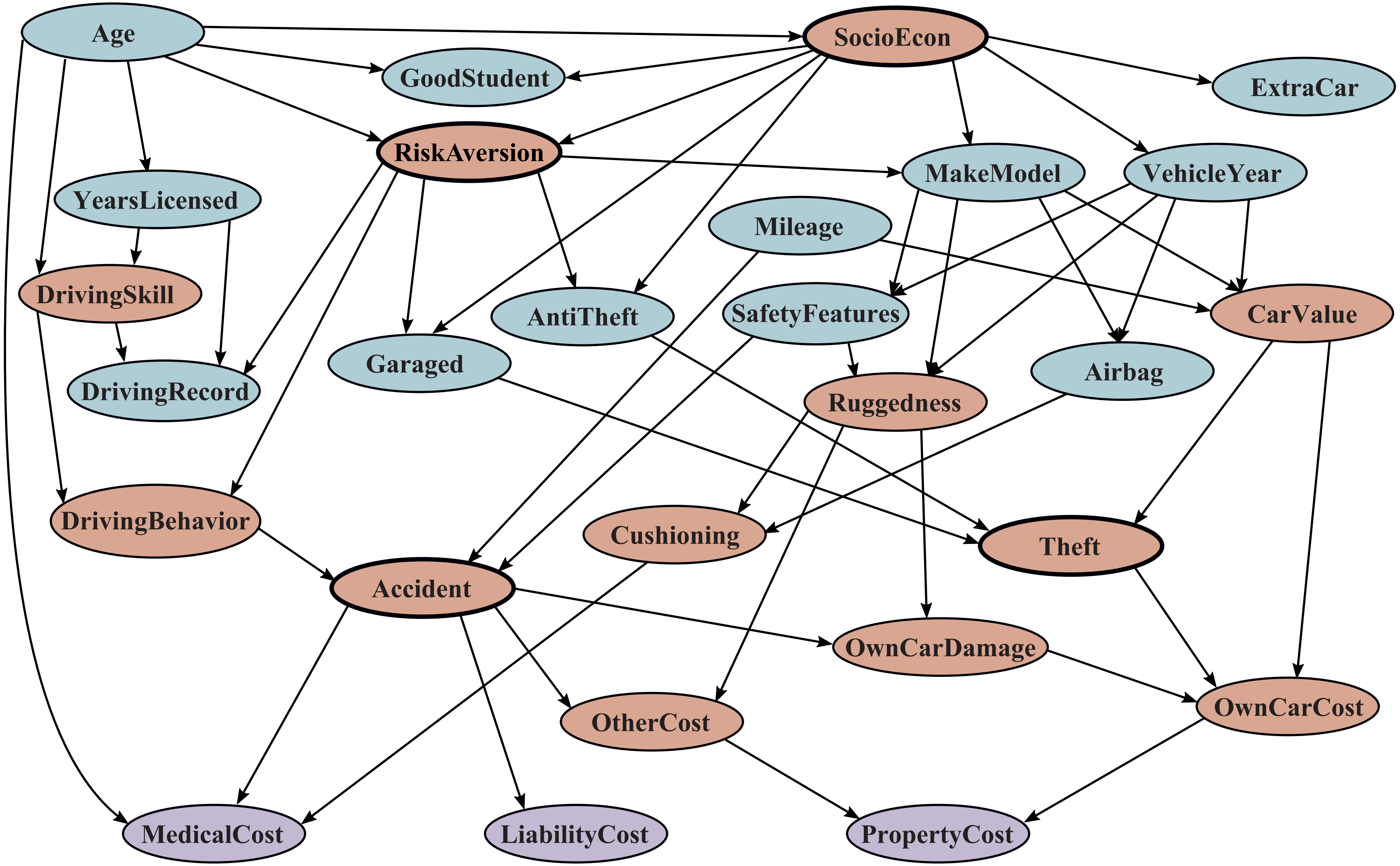




Quiz: Traffic II

- Variables
  - T: Traffic
  - R: It rains
  - L: Low pressure
  - D: Roof drips
  - B: Ballgame
  - C: Cavity
- Can you design a graphical model?

Example: Car Insurance



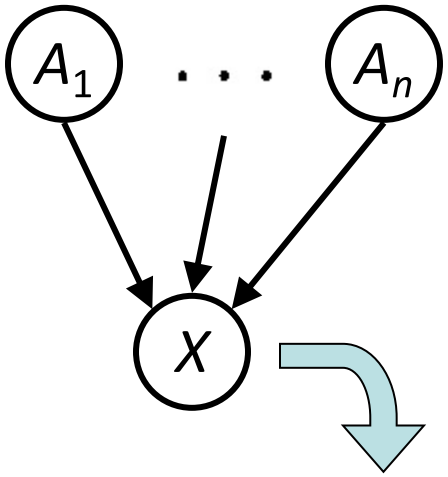
Bayes Net Semantics

Bayes Net Semantics

- A set of nodes, one per variable X
- A directed, acyclic graph
- Conditional distributions for each node
  - A collection of distributions over X, one conditioned on each combination of parents' values

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

  - **CPT**: conditional probability table
    - each row is a conditional distribution over X given a specific combination of parent values
  - Description of a potentially “causal” process

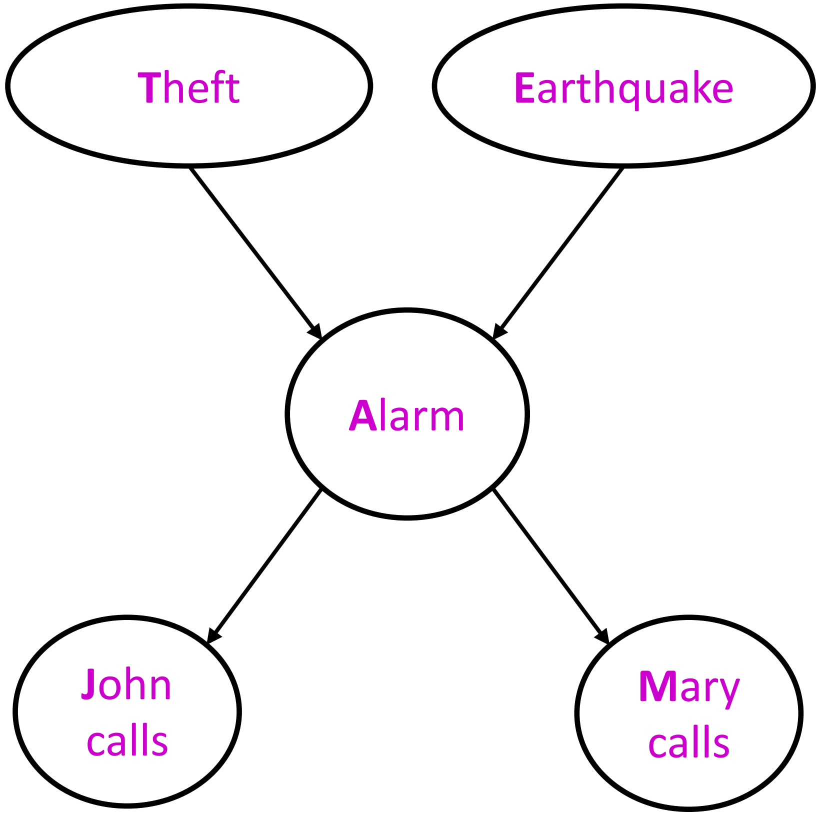


$$P(X|A_1 \ldots A_n)$$

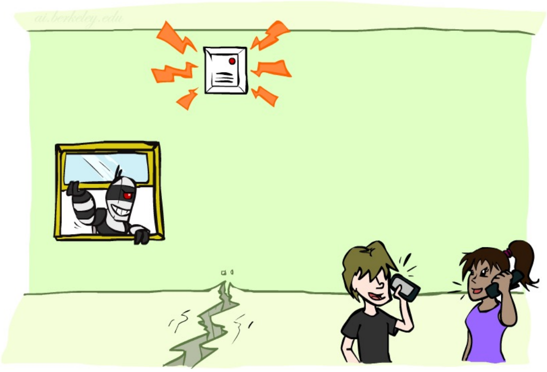
Bayes net = Topology (graph) + Local Conditional Probabilities

Example: Alarm Network

P(T)	
true	false
0.001	0.999



P(E)	
true	false
0.002	0.998



T	E	P(A|T,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

Bayes Net Size

- 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 $O(n \cdot d^{k+1})$
  - Scales linearly with n as long as connections are local (k is small)

Probabilities in BNs

- Recap: a BN describes a *complex, joint distribution* using *simple, conditional distributions*

- How to recover the joint distribution?
  - Take a product of local conditional distributions:

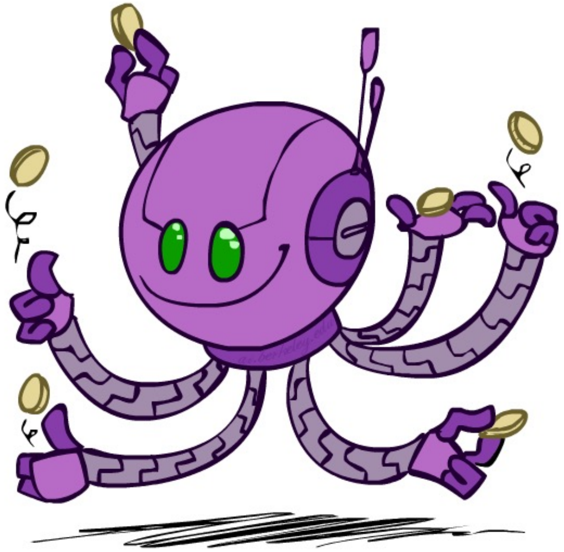
$$P(x_1,\ldots,x_n) = \prod_i P(x_i \mid parents(X_i))$$

  - i.e., to compute the probability of a full assignment, you need to multiply all the relevant conditionals together.

Example: Coin Flips

$$P(x_1,..,x_n) = \prod_i P(x_i \mid parents(X_i))$$

X_1 X_2 . . . X_n



$P(X_1)$

h	0.5
t	0.5

$P(X_2)$

h	0.5
t	0.5

. . .

$P(X_n)$

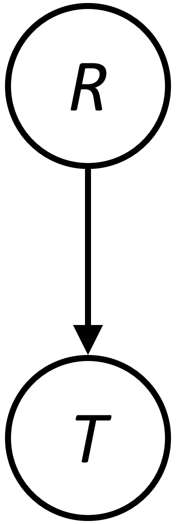
h	0.5
t	0.5

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

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

Example: Traffic

$$P(x_1,..,x_n) = \prod_i P(x_i \mid parents(X_i))$$



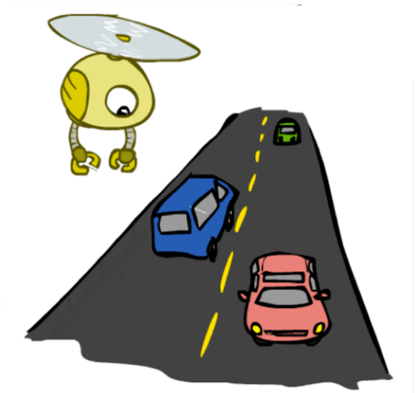
P(R)

+r	1/4
-r	3/4

P(T|R)

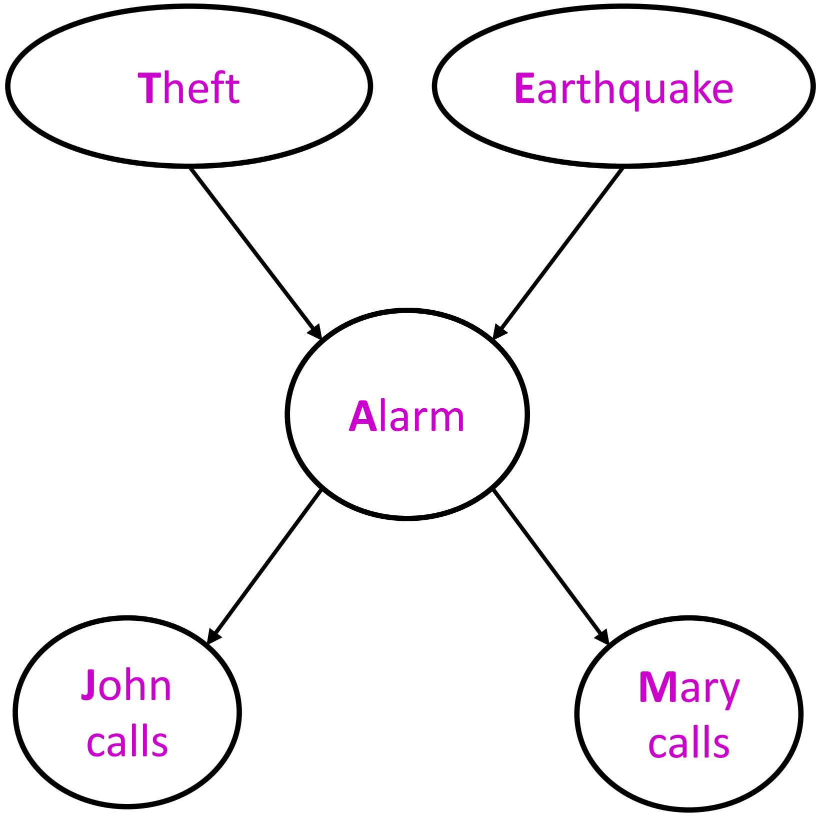
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

P(+r, -t) = P(+r)P(-t|+r) = ¼* ¼



Example: Theft $P(x_1,..,x_n) = \prod_i P(x_i \mid parents(X_i))$

P(T)	
true	false
0.001	0.999



P(E)	
true	false
0.002	0.998

P(+t, -e, +a, -j, -m) = ?

P(+t) P(-e) P(+a|+t,-e) P(-j|+a) P(-m|+a)

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

=.000028

T	E	P(A|T,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

Conditional independence in BNs

- Compare the Bayes net probabilities

$$P(x_1,..,x_n) = \prod_i P(x_i \mid parents(X_i))$$

with the chain rule (valid for all distributions)

$$P(x_1,..,x_n) = \prod_i P(x_i \mid x_1,\ldots,x_{i-1})$$

- **Are they identical? Why?**
  - Assume (without loss of generality) that $X_1,..,X_n$ sorted in topological order according to the graph (i.e., parents before children), so:

    $$Parents(X_i) \subseteq X_1,\ldots,X_{i-1}$$

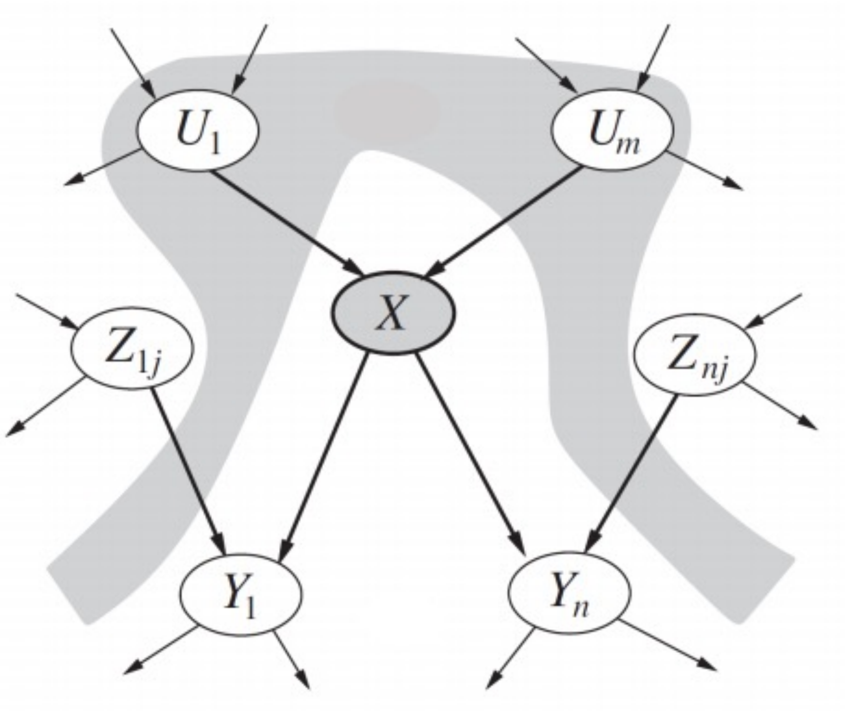
  - <u>Assume</u> conditional independences:

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

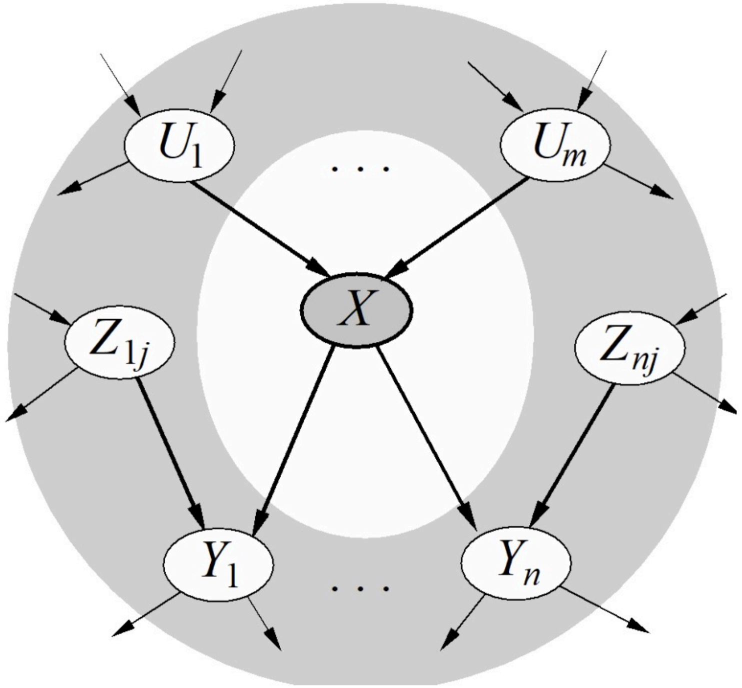
- Not every BN can represent every joint distribution
  - The topology *enforces* certain conditional independencies

Conditional Independence Assumptions

- Each node, given its parents, is conditionally independent of all its non-descendants in the graph



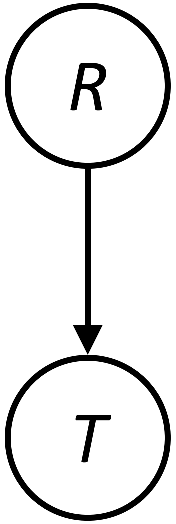
Each node, given its *MarkovBlanket*, is conditionally independent of all other nodes in the graph



MarkovBlanket refers to the parents, children, and children's other parents.

BN Design: Traffic

- Causal direction
  - **Rain** causes **Traffic** to be heavy



P(R)

+r	1/4
-r	3/4

P(T|R)

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

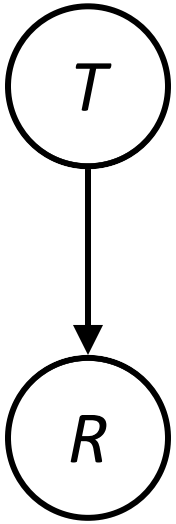
-r	+t	1/2
	-t	1/2

P(R,T)

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

BN Design: 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(R,T)

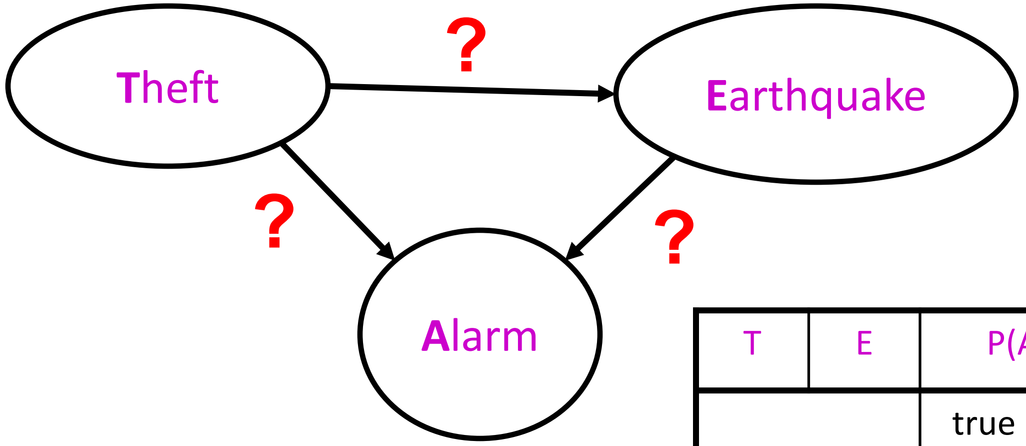
+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

BN Design: Theft

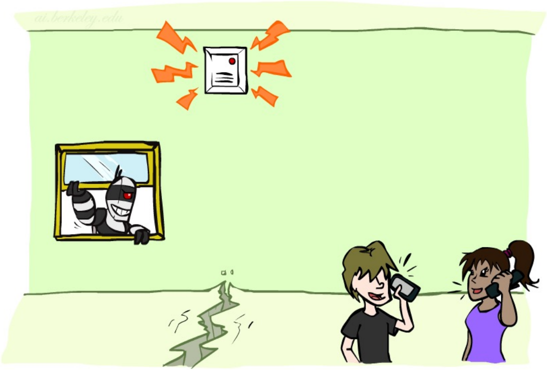
- Causal:
  - Theft
  - Earthquake
  - Alarm

P(T)	
true	false
0.001	0.999

P(E)	
true	false
0.002	0.998



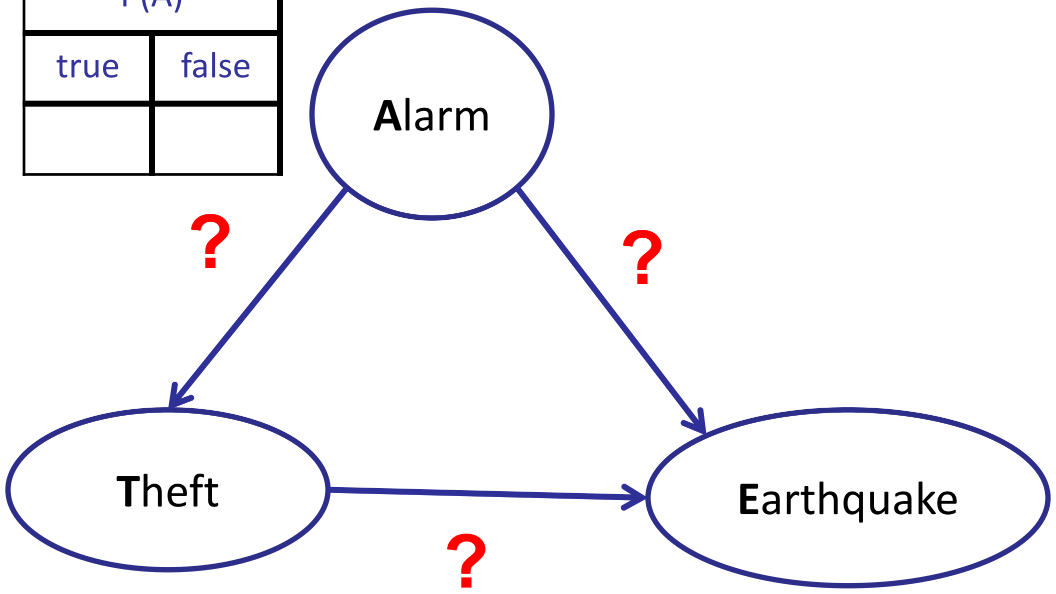
T	E	P(A|T,E)	
		true	false
true	true	0.95	0.05
true	false	0.94	0.06
false	true	0.29	0.71
false	false	0.001	0.999



BN Design: Theft

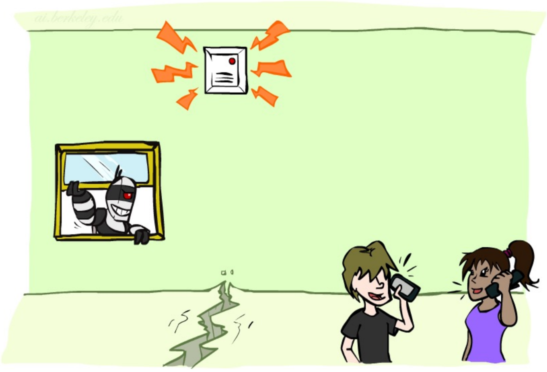
- Non-causal:
  - Alarm
  - Theft
  - Earthquake

P(A)	
true	false



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

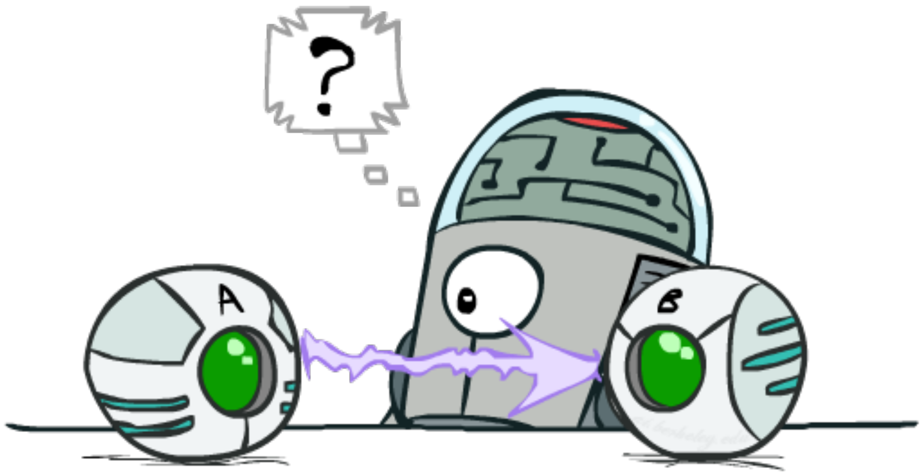
A	T	P(E|A,T)	
		true	false
true	true	?	
true	false		
false	true		
false	false		



Causality?

- When Bayes nets reflect the true causal patterns:
  - Often simpler (more sparse)
  - Often easier to think about
  - Often easier to aquire from experts
- But, BNs not actually need to be causal
  - Sometimes, there is no causal net existing over a domain (especially if variables are missing)
    - E.g. consider the variables *Traffic* and *Drips*
  - End up with arcs that reflect *correlation*, not causation
- What do the arcs really mean?
  - Topology may happen to encode causal structure
  - Topology really encodes conditional independence

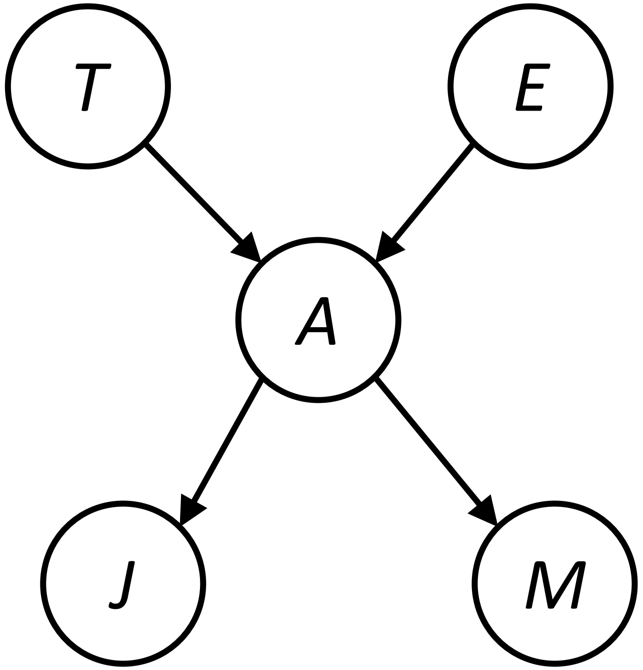
$$P(x_i|x_1,\ldots x_{i-1}) = P(x_i|\mathit{parents}(X_i))$$



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: $P(\boldsymbol{Q} \mid \boldsymbol{e}) = \alpha \sum_h P(\boldsymbol{Q}, \boldsymbol{h}, \boldsymbol{e})$
  - With a BN: can obtain entries of the joint distribution by multiplying the corresponding conditional probabilities
- $P(T \mid j, m) = \alpha \sum_{e,a} P(T, e, a, j, m)$

  $= \alpha \sum_{e,a} P(T)\, P(e)\, P(a \mid T,e)\, P(j \mid a)\, P(m \mid a)$
- So inference in Bayes nets means computing sums of products of numbers:
  - sounds easy!!
- **Problem**: sums of *exponentially many* products!
  - Exponential to the number of hidden variables



Can we do better?

- $\sum_{e,a} P(T)\ \mathrm{P(e)}\ \mathrm{P(a|}T,e)\ P(j|a)\ P(m|a)$
- $= P(T)P(+e)P(\mathrm{+a}|T,+e)P(j|+a)P(m|+a)$

  $+ P(T)P(-e)P(\mathrm{+a}|T,-e)P(j|+a)P(m|+a)$

  $+ P(T)P(+e)P(\mathrm{-a}|T,e)P(j|-a)P(m|-a)$

  $+ P(T)P(-e)P(\mathrm{-a}|T,-e)P(j|-a)P(m|-a)$

Lots of repeated subexpressions!

Next Week: Variable Elimination