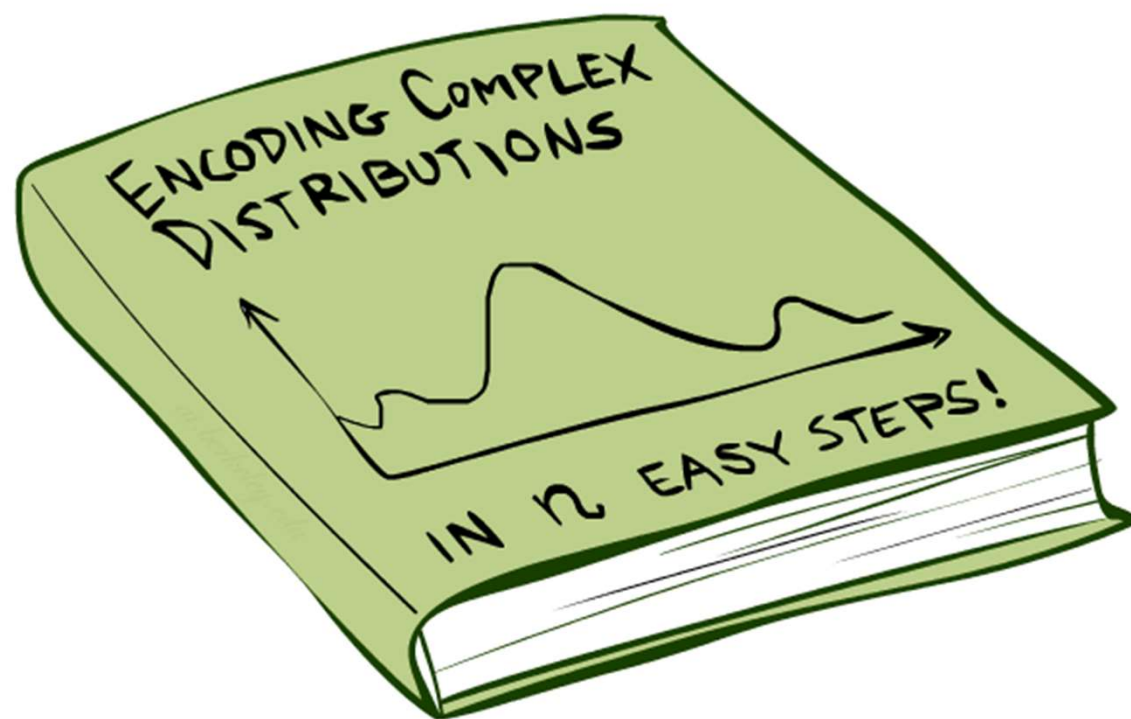


Probabilistic Reasoning

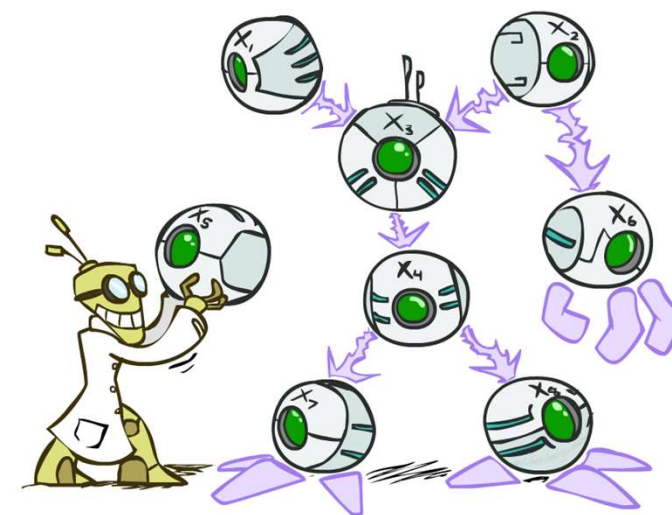
CHAPTER 14 IN THE TEXTBOOK

Bayes' Nets: Big Picture

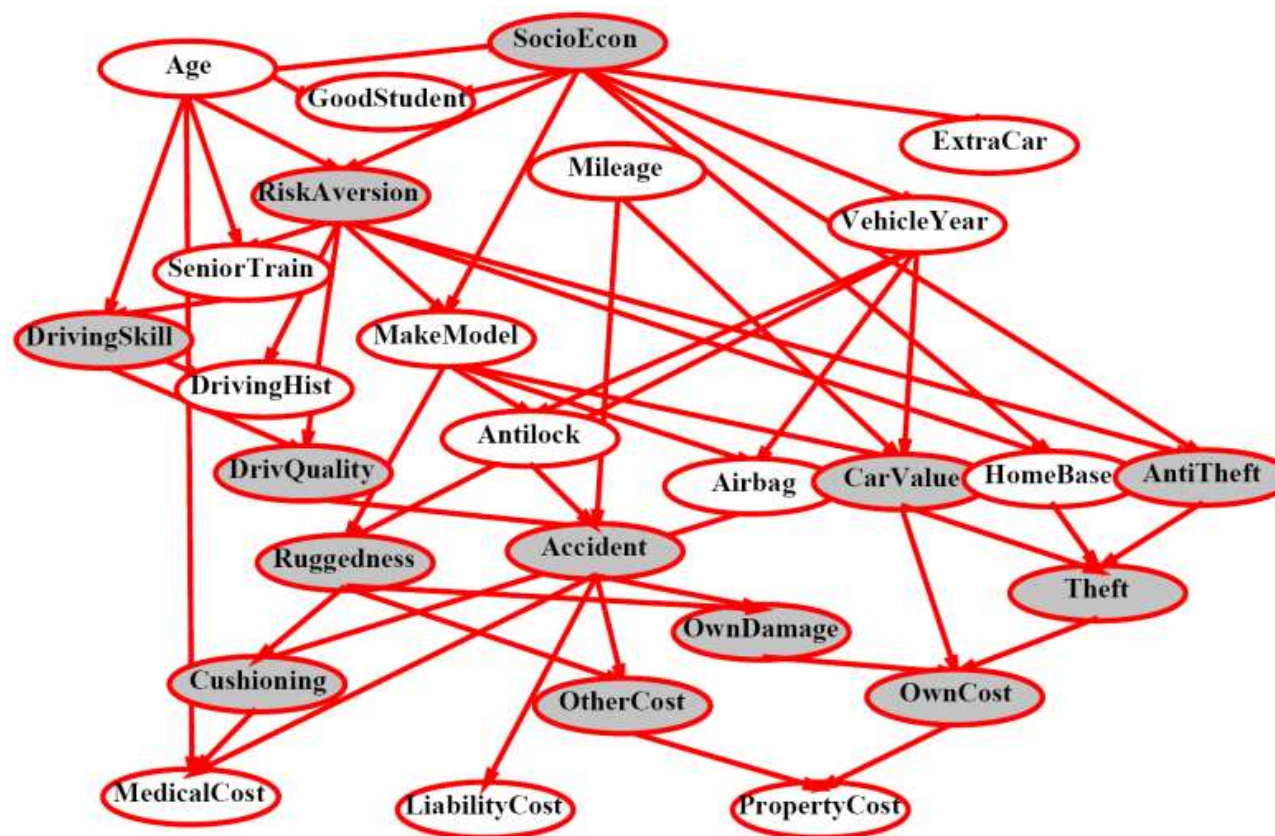


Bayes' Nets: Big Picture

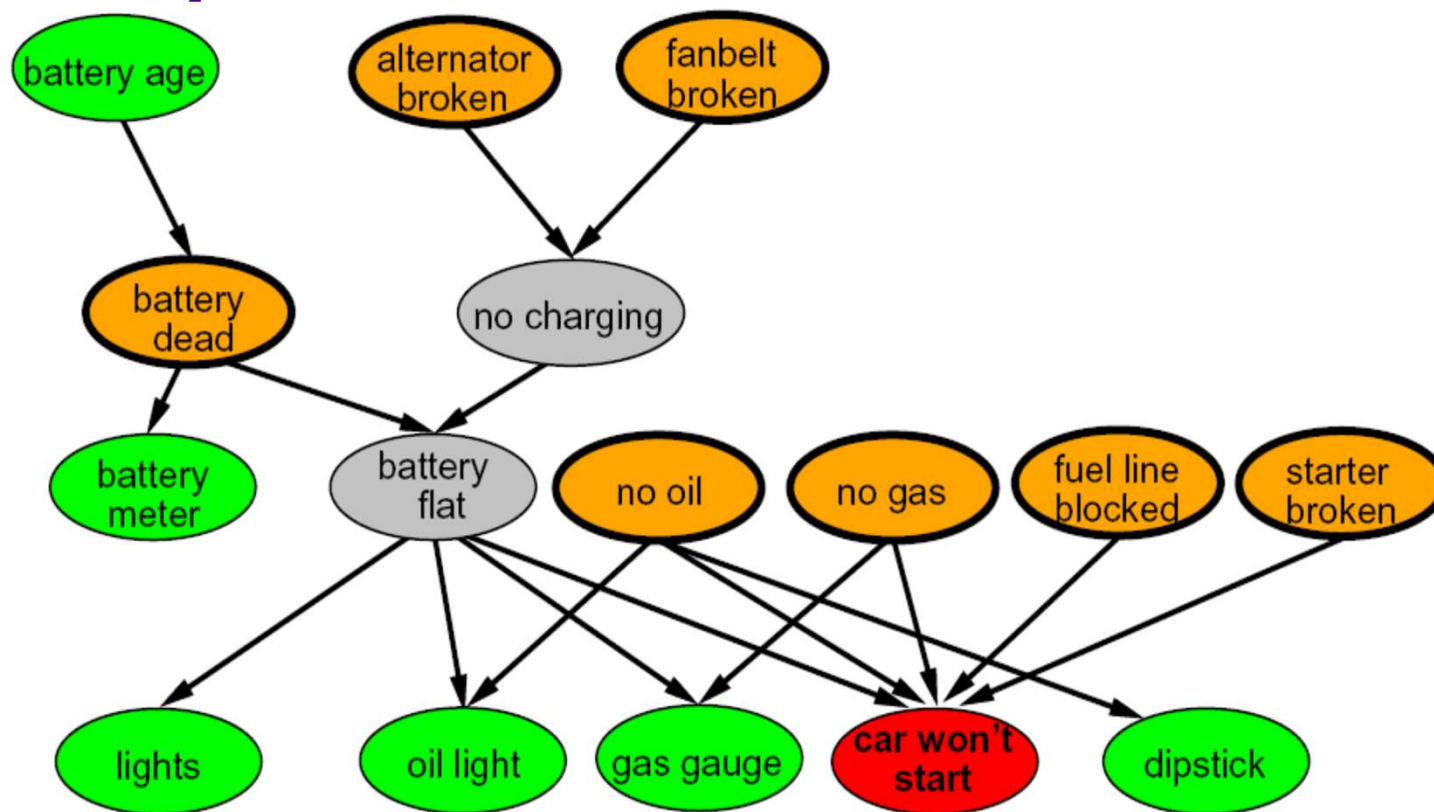
- Two problems with using full joint distribution tables as our probabilistic models:
 - Unless there are only a few variables, the joint is WAY too big to represent explicitly
 - Hard to learn (estimate) anything empirically about more than a few variables at a time
- **Bayes' nets**: a technique for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

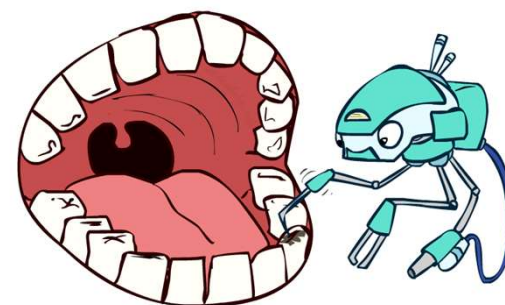
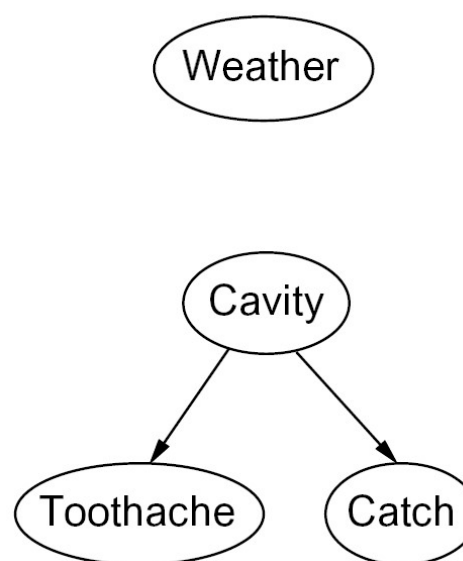


Example Bayes' Net: Car



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!)

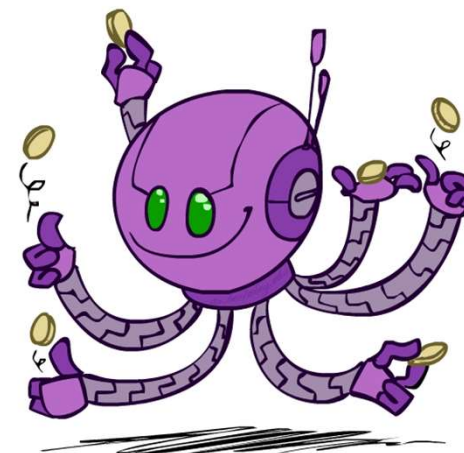


Example: Coin Flips

- N independent coin flips



- No interactions between variables: **absolute independence**



Example: Traffic

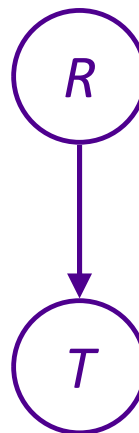
- Variables:
 - R: It rains
 - T: There is traffic
- Model 1: independence



- Why is an agent using model 2 better?



- Model 2: rain causes traffic



Example: Traffic II

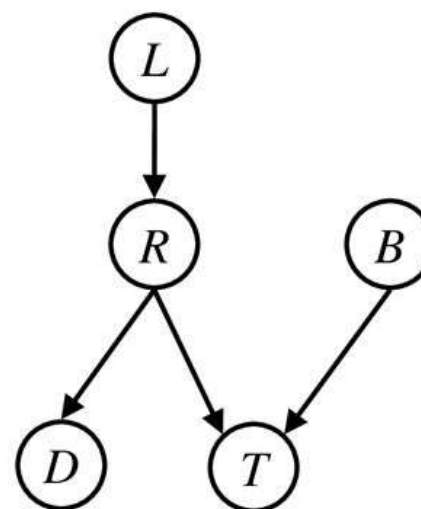
- Let's build a causal graphical model!
- Variables
 - T: Traffic
 - R: It rains
 - L: Low pressure
 - D: Roof drips
 - B: Ballgame
 - C: Cavity



Example: Traffic

- Variables

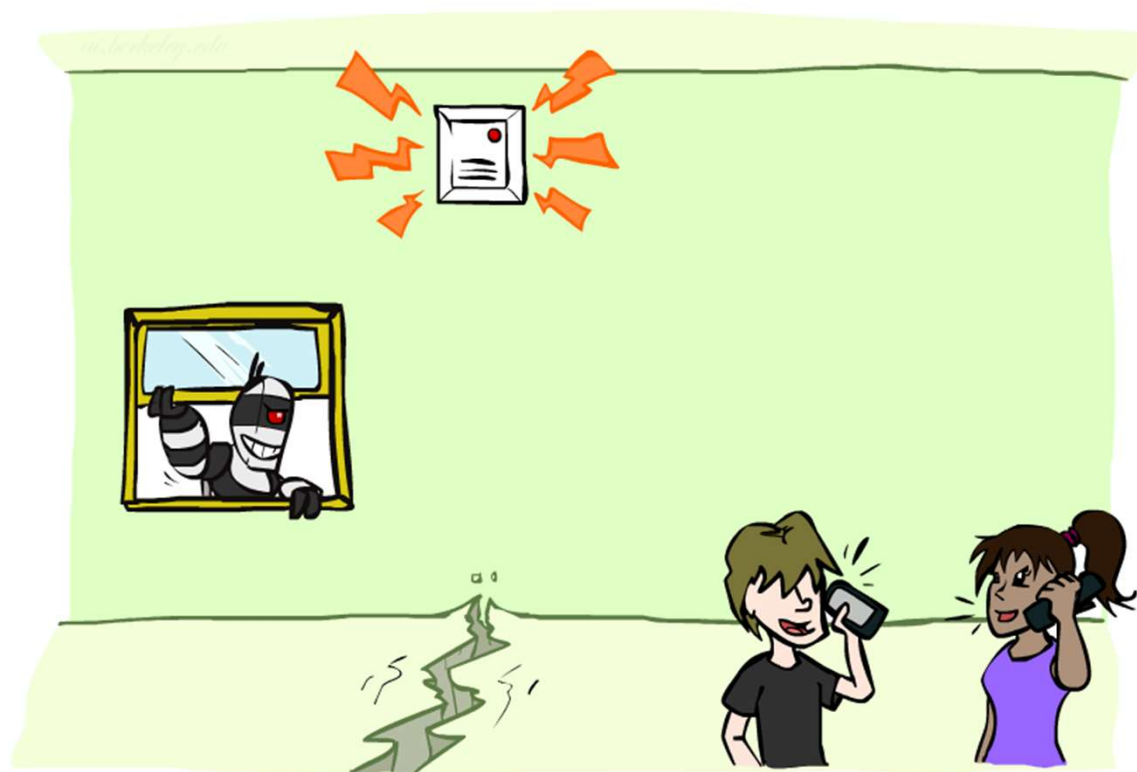
- T: Traffic
- R: It rains
- L: Low pressure
- D: Roof drips
- B: Ballgame

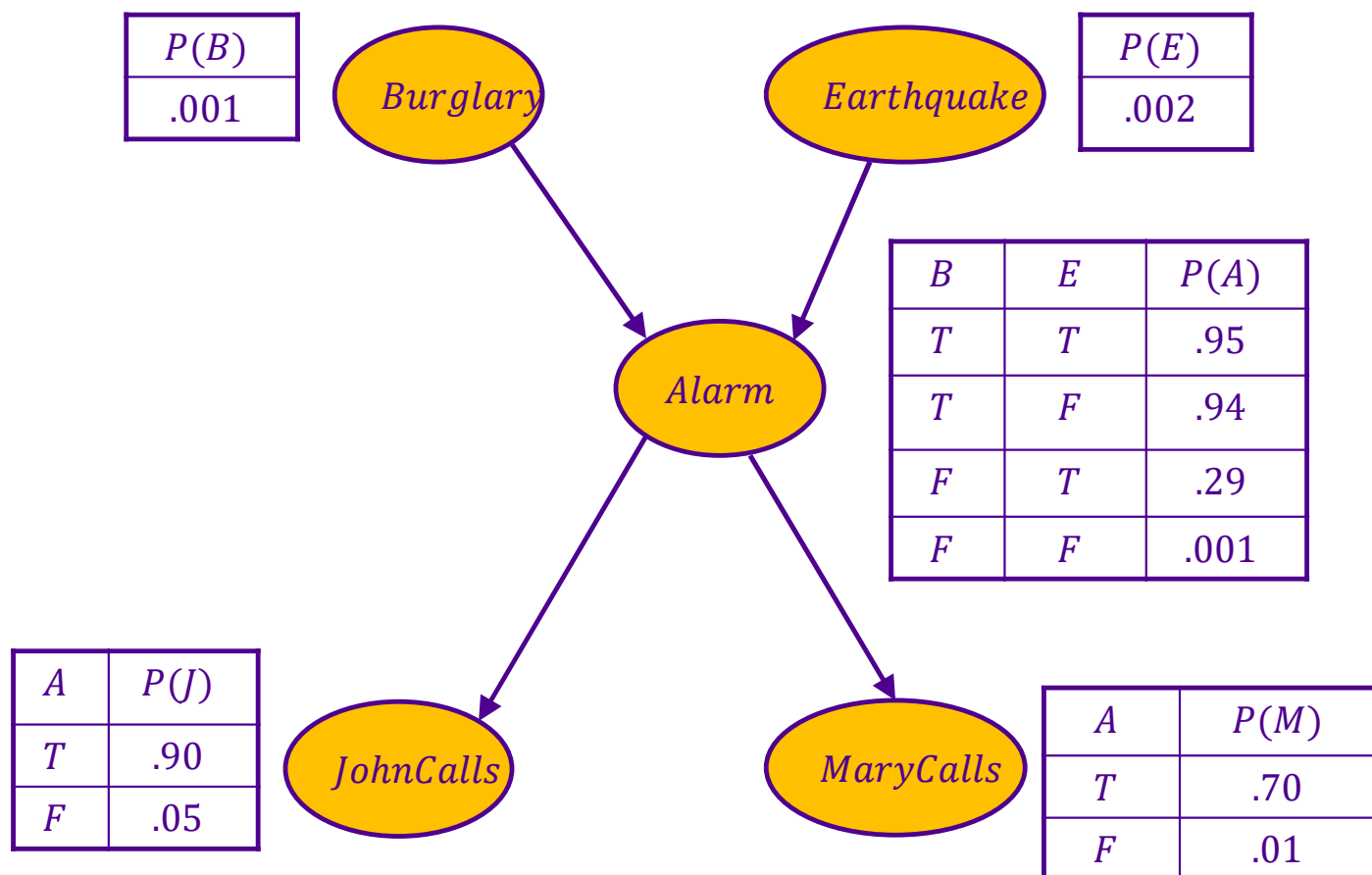


Example: Alarm Network

- Variables

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





Bayes' Net Semantics



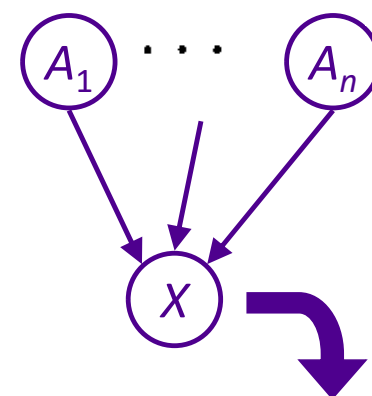
Bayes' Net Semantics



- 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



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

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

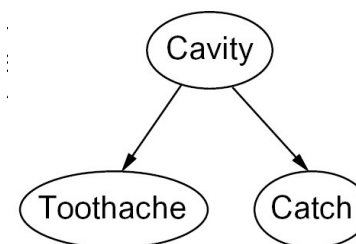
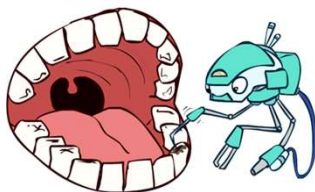


Probabilities in BNs

- 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:



$$P(+cavity, +catch, -toothache)$$



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:

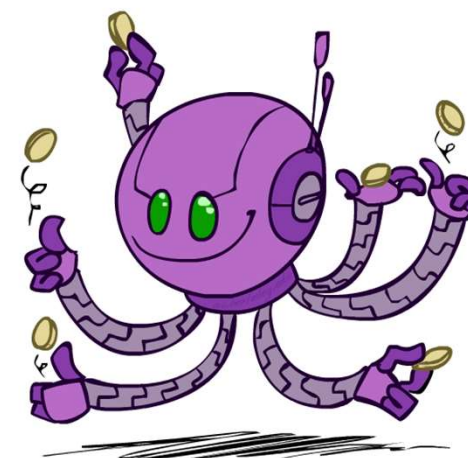
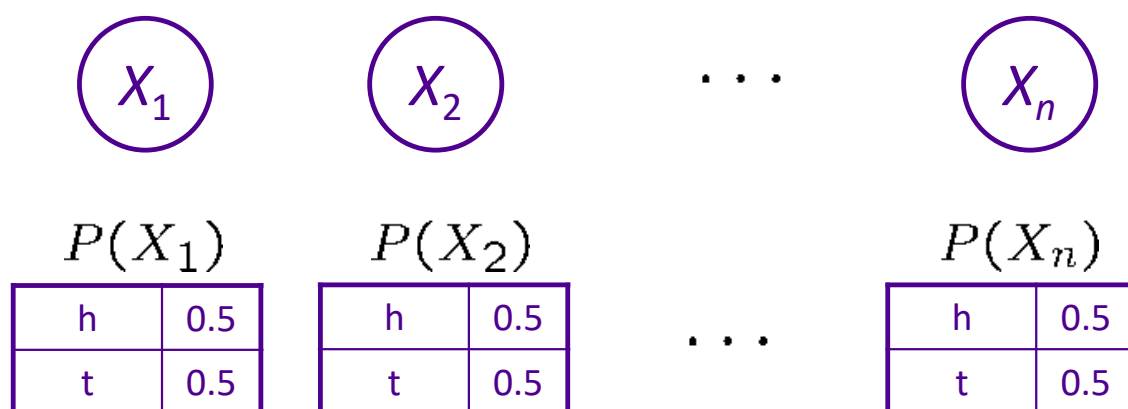
$$P(x_i | x_1, \dots, x_{i-1}) = P(x_i | \text{parents}(X_i))$$

→ 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

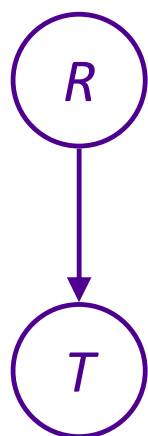
Example: Coin Flips



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

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

Example: Traffic



$$P(R)$$

$+r$	$1/4$
$-r$	$3/4$

$$P(T|R)$$

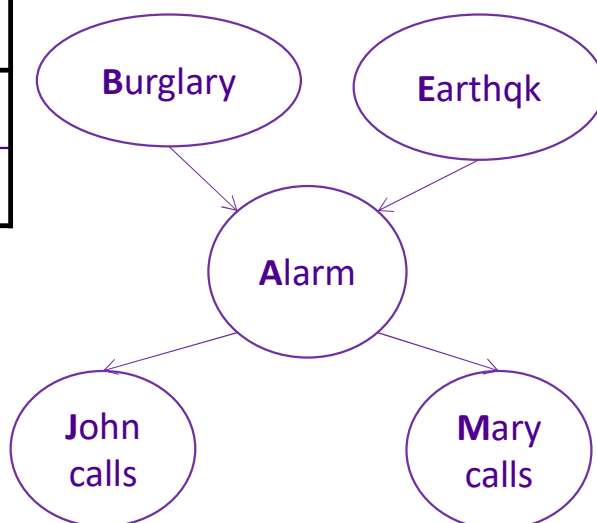
$+r$	$+t$	$3/4$
$+r$	$-t$	$1/4$
$-r$	$+t$	$1/2$
$-r$	$-t$	$1/2$

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



Example: Alarm Network

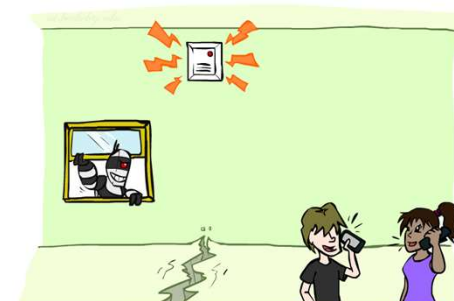
B	$P(B)$
$+b$	0.001
$-b$	0.999



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

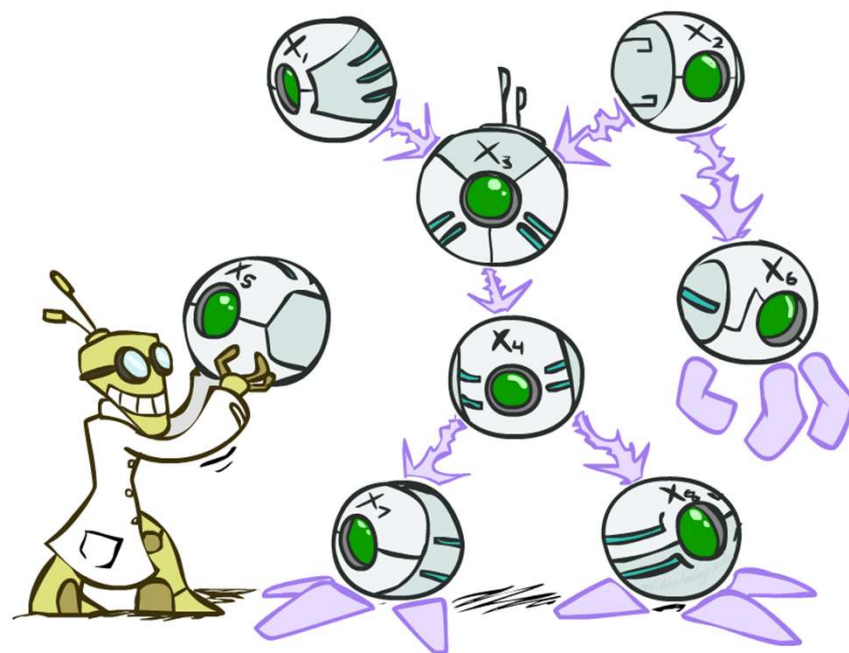
E	$P(E)$
$+e$	0.002
$-e$	0.998



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

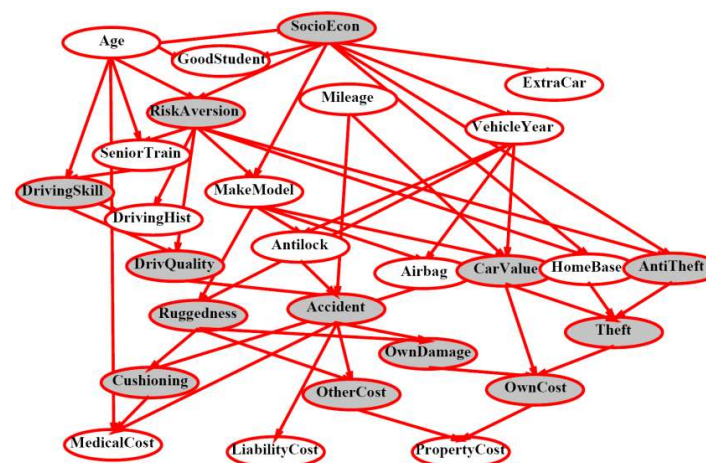
Bayes' Nets

- So far: how a Bayes' net encodes a joint distribution
- 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)



Bayes' Nets

- A Bayes' net is an efficient encoding of a probabilistic model of a domain



- Questions we can ask:
 - Inference: given a fixed BN, what is $P(X \mid e)$?
 - Representation: given a BN graph, what kinds of distributions can it encode?
 - Modeling: what BN is most appropriate for a given domain?

Bayes' Net Semantics

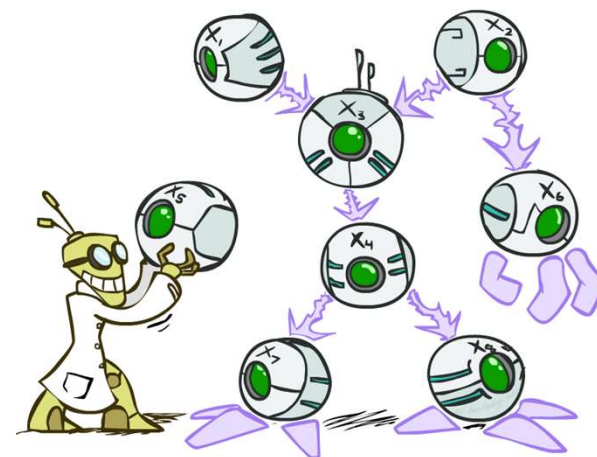
- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
 - A collection of distributions over X , one for each combination of parents' values

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

- Bayes' nets implicitly encode joint distributions

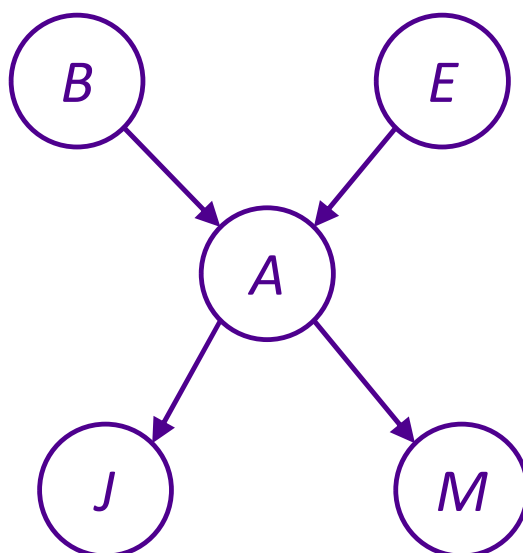
- As a product of local conditional distributions
- To see what probability a BN gives to a full assignment, multiply all the relevant conditionals together:

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



Example: Alarm Network

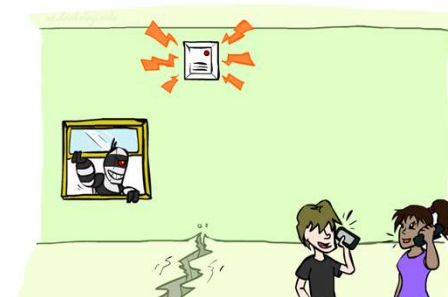
B	P(B)
+b	0.001
-b	0.999



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

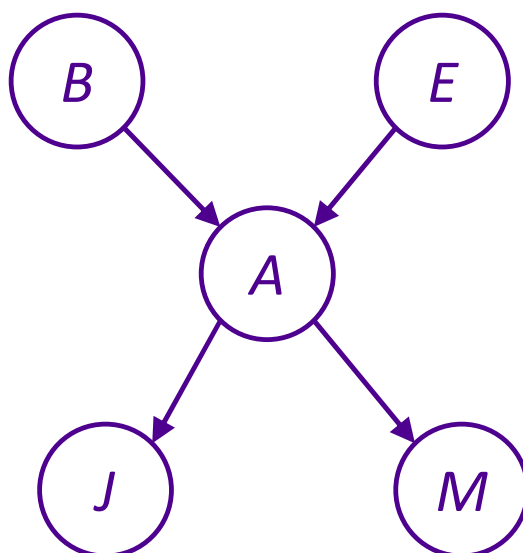


$$P(+b, -e, +a, -j, +m) =$$

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

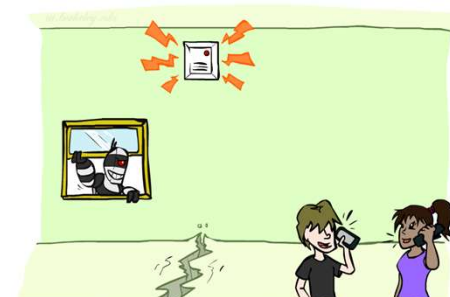
Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998

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



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

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

$$\begin{aligned}
 &P(+b, -e, +a, -j, +m) = \\
 &P(+b)P(-e)P(+a|+b, -e)P(-j|+a)P(+m|+a) = \\
 &0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7
 \end{aligned}$$

Size of a Bayes' Net

- How big is a joint distribution over N Boolean variables?

$$2^N$$

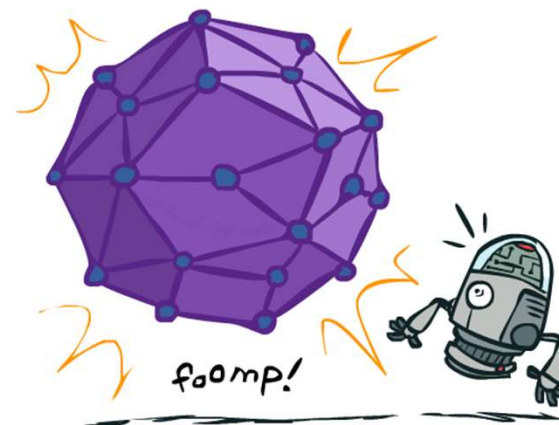
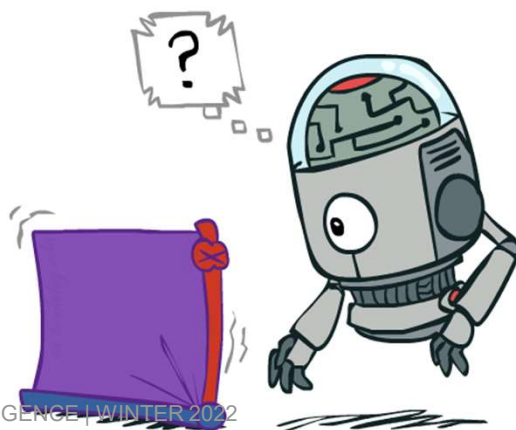
- How big is an N -node net if nodes have up to k parents?

$$O(N * 2^{k+1})$$


- Both give you the power to calculate

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

- BNs: Huge space savings!
- Also easier to elicit local CPTs
- Also faster to answer queries



Bayes' Nets

-  Representation
 - Conditional Independences
 - Probabilistic Inference
 - Learning Bayes' Nets from Data

Conditional Independence

- X and Y are **independent** if

$$\forall x, y \quad P(x, y) = P(x)P(y) \quad \text{---} \rightarrow \quad X \perp\!\!\!\perp Y$$

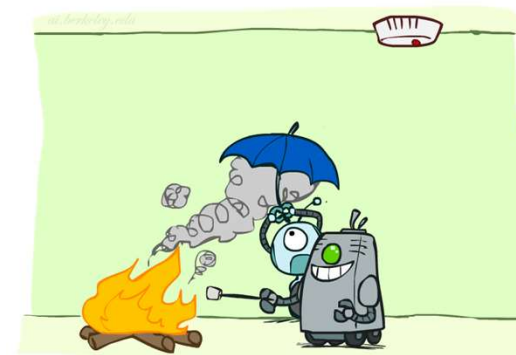
- X and Y are **conditionally independent** given Z

- (Conditional) independence is a property of a distribution

$$\forall x, y, z \quad P(x, y|z) = P(x|z)P(y|z) \quad \text{---} \rightarrow \quad X \perp\!\!\!\perp Y|Z$$

- Example:

$$Alarm \perp\!\!\!\perp Fire|Smoke$$

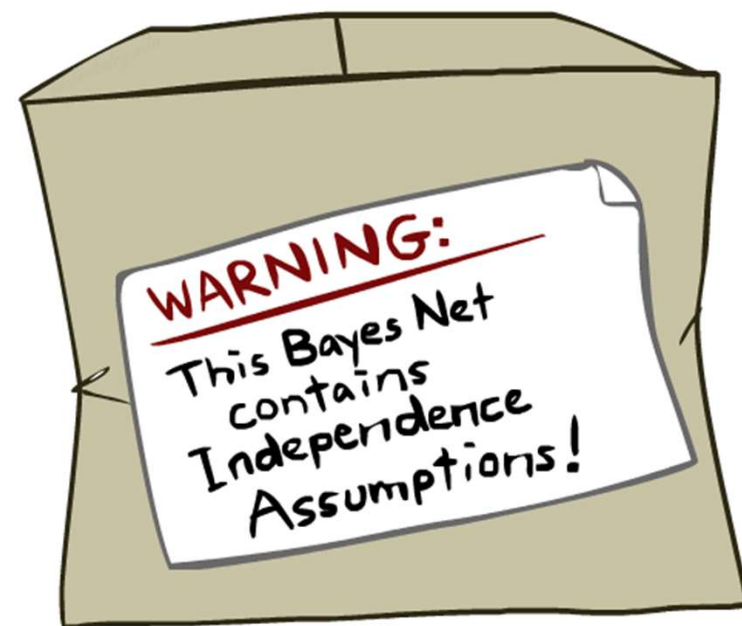


Bayes Nets: Assumptions

- Assumptions we are required to make to define the Bayes net when given the graph:

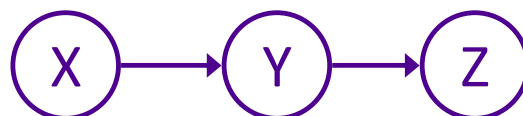
$$P(x_i | x_1 \cdots x_{i-1}) = P(x_i | \text{parents}(X_i))$$

- Beyond above “chain rule \rightarrow Bayes net” conditional independence assumptions
 - Often additional conditional independences
 - They can be read off the graph
- Important for modeling: understand assumptions made when choosing a Bayes net graph



Independence in a BN

- Important question about a BN:
 - Are two nodes independent given certain evidence?
 - If yes, can prove using algebra (tedious in general)
 - If no, can prove with a counter example
 - Example:



- Question: are X and Z necessarily independent?
 - Answer: no. Example: low pressure causes rain, which causes traffic.
 - X can influence Z, Z can influence X (via Y)
 - Addendum: they *could* be independent: how?

Bayes Nets Representation Summary

- Bayes nets compactly encode joint distributions
- Guaranteed independencies of distributions can be deduced from BN graph structure
- A Bayes' net's joint distribution may have further (conditional) independence that is not detectable until you inspect its specific distribution

Bayes' Nets

- Representation
- Conditional Independences
- Probabilistic Inference
 - Enumeration (exact, exponential complexity)
 - Variable elimination (exact, worst-case exponential complexity, often better)
 - Probabilistic inference is NP-complete
 - Sampling (approximate)
- Learning Bayes' Nets from Data

Inference

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

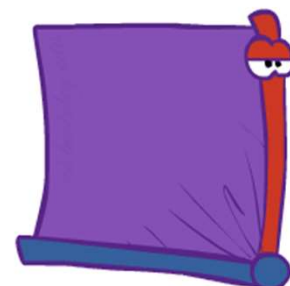
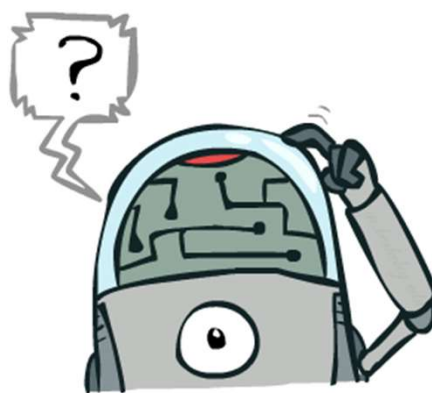
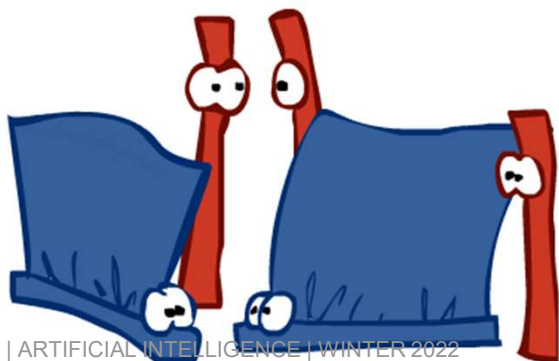
- Examples:

- Posterior probability

$$P(Q|E_1 = e_1, \dots, E_k = e_k)$$

- Most likely explanation:

$$\operatorname{argmax}_q P(Q = q|E_1 = e_1 \dots)$$



Inference by Enumeration

- General case:

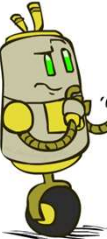
- Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
 - Query* variable: Q
 - Hidden variables: $H_1 \dots H_r$
- X_1, X_2, \dots, X_n
All variables

- We want:

** Works fine with multiple query variables, too*

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

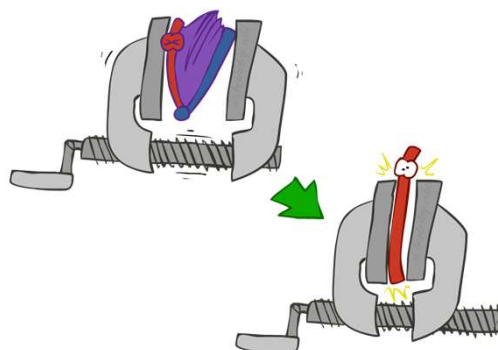
- Step 1: Select the entries consistent with the evidence



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

2 0.15

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



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

- Step 3: Normalize

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

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

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

Inference by Enumeration in Bayes' Net

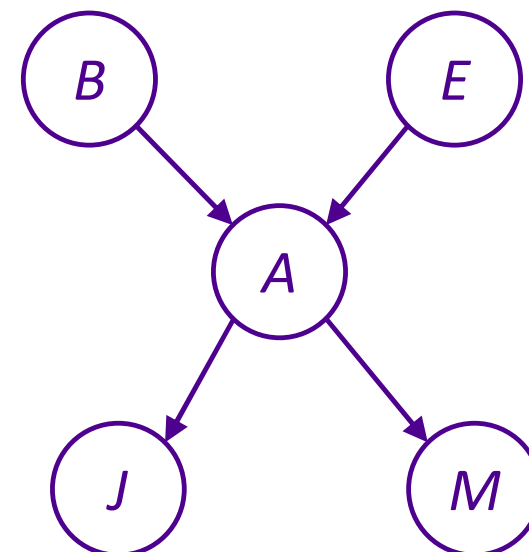
- Given unlimited time, inference in BNs is easy
- Reminder of inference by enumeration by example:

$$P(B \mid +j, +m) \propto_B P(B, +j, +m)$$

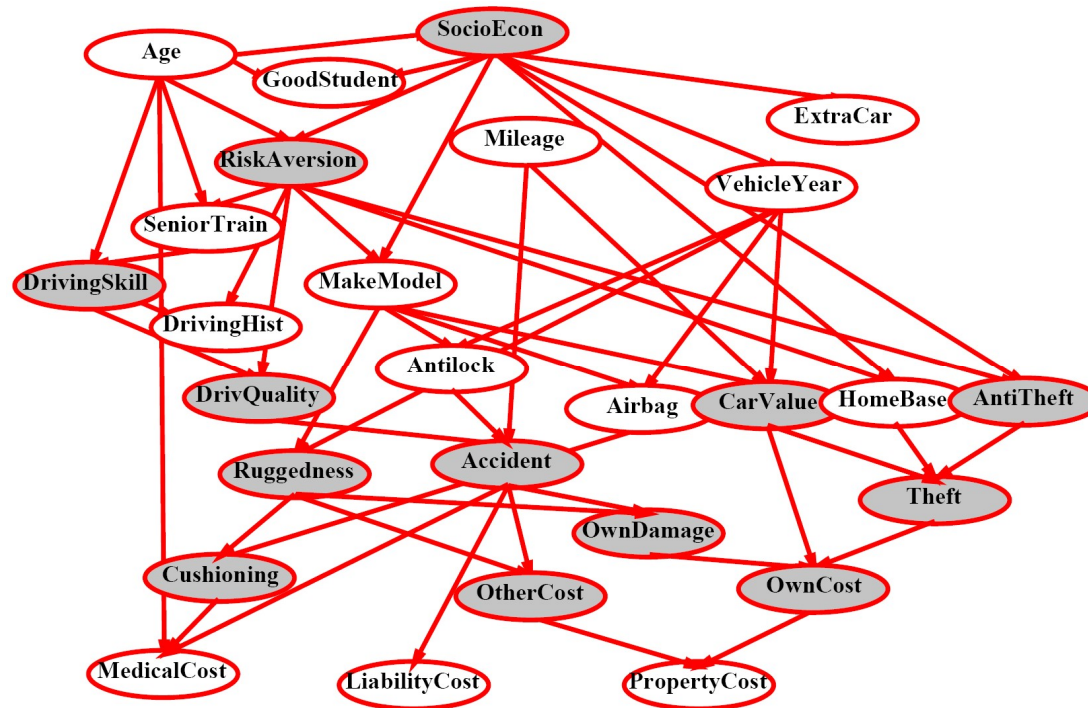
$$= \sum_{e,a} P(B, e, a, +j, +m)$$

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

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



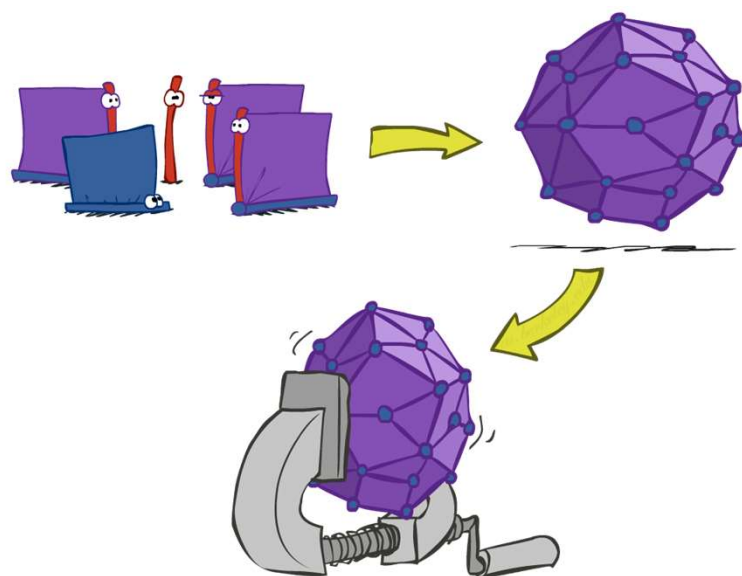
Inference by Enumeration?



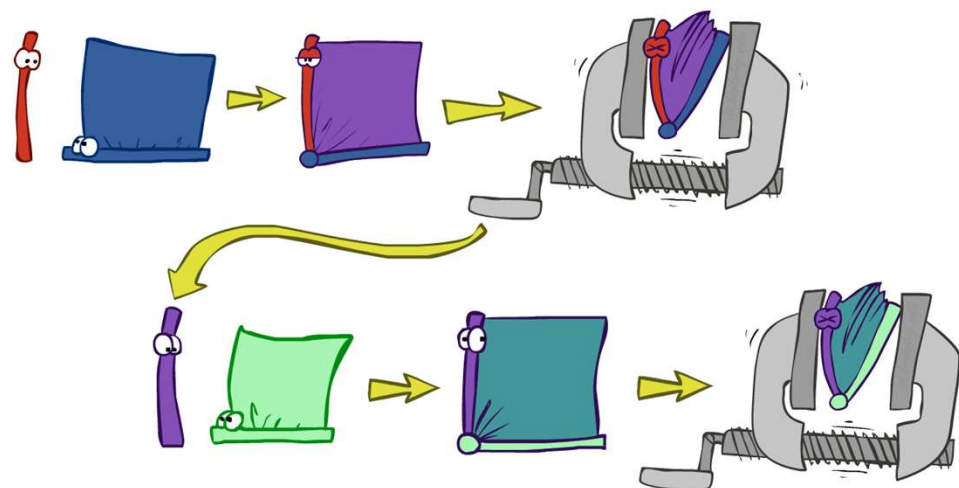
$$P(\textit{Antilock} | \textit{observed variables}) = ?$$

Inference by Enumeration vs. Variable Elimination

- Why is inference by enumeration so slow?
 - You join up the whole joint distribution before you sum out the hidden variables



- Idea: interleave joining and marginalizing!
 - Called “Variable Elimination”
 - Still NP-hard, but usually much faster than inference by enumeration



VE: Computational and Space Complexity

- The computational and space complexity of variable elimination is determined by the largest factor
- The elimination ordering can greatly affect the size of the largest factor.
 - E.g., previous slide's example 2^n vs. 2
- Does there always exist an ordering that only results in small factors?
 - No!

Worst Case Complexity?

- CSP:

$$(x_1 \vee x_2 \vee \neg x_3) \wedge (\neg x_1 \vee x_3 \vee \neg x_4) \wedge (x_2 \vee \neg x_2 \vee x_4) \wedge (\neg x_3 \vee \neg x_4 \vee \neg x_5) \wedge (x_2 \vee x_5 \vee x_7) \wedge (x_4 \vee x_5 \vee x_6) \wedge (\neg x_5 \vee x_6 \vee \neg x_7) \wedge (\neg x_5 \vee \neg x_6 \vee x_7)$$

$$P(X_i = 0) = P(X_i = 1) = 0.5$$

$$Y_1 = X_1 \vee X_2 \vee \neg X_3$$

$$\dots$$

$$Y_8 = \neg X_5 \vee X_6 \vee X_7$$

$$Y_{1,2} = Y_1 \wedge Y_2$$

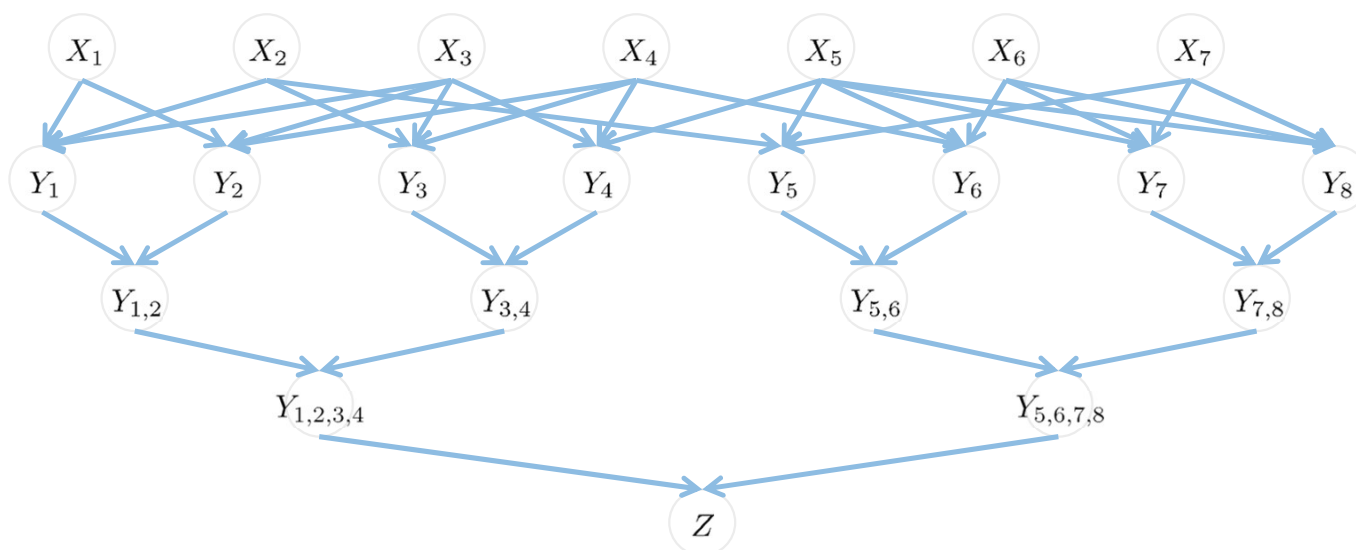
$$\dots$$

$$Y_{7,8} = Y_7 \wedge Y_8$$

$$Y_{1,2,3,4} = Y_{1,2} \wedge Y_{3,4}$$

$$Y_{5,6,7,8} = Y_{5,6} \wedge Y_{7,8}$$

$$Z = Y_{1,2,3,4} \wedge Y_{5,6,7,8}$$








- If we can answer $P(z)$ equal to zero or not, we answered whether the 3-SAT problem has a solution.
- Hence inference in Bayes' nets is NP-hard. No known efficient probabilistic inference in general.

Polytrees

- A polytree is a directed graph with no undirected cycles
- For poly-trees you can always find an ordering that is efficient
 - Try it!!
- Cut-set conditioning for Bayes' net inference
 - Choose set of variables such that if removed only a polytree remains
 - Exercise: Think about how the specifics would work out!

Bayes' Nets

-  Representation
-  Conditional Independences
- Probabilistic Inference
 -  Enumeration (exact, exponential complexity)
 -  Variable elimination (exact, worst-case exponential complexity, often better)
 - Inference is NP-complete
 -  Sampling (approximate)
- Learning Bayes' Nets from Data

Variable Elimination

- Interleave joining and marginalizing
- d^k entries computed for a factor over k variables with domain sizes d
- Ordering of elimination of hidden variables can affect size of factors generated
- Worst case: running time exponential in the size of the Bayes' net

