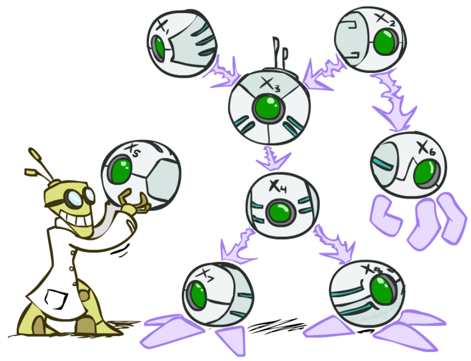
CSE 473: Artificial Intelligence

Bayes' Nets

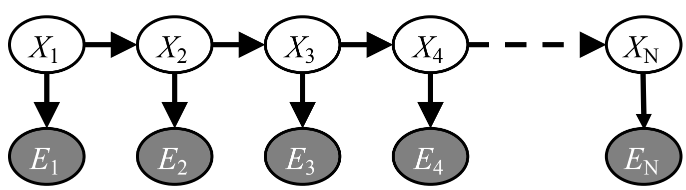


Daniel Weld

Hidden Markov Models

Two random variable at each time step

- Hidden state, X_i
- Observation, E_i



Conditional Independences

Dynamics don't change

- E.g., $P(X_2 \mid X_1) = P(X_{18} \mid X_{17})$

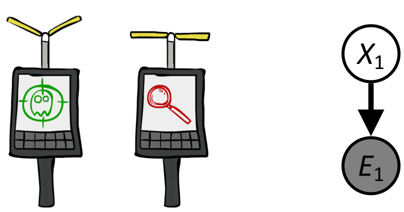
HMM Computations

- Given
  - Parameters
  - Evidence $E_{1:n} = e_{1:n}$
- Inference problems include:
  - Filtering, find $P(X_t|e_{1:t})$ for all t
    - Exact Inference
    - Particle Filter
  - Smoothing, find $P(X_t|e_{1:n})$ for all t
  - Most probable explanation, find

    $x^*_{1:n} = \text{argmax}_{x_{1:n}} P(x_{1:n}|e_{1:n})$

Exact Inference: Forward Algorithm

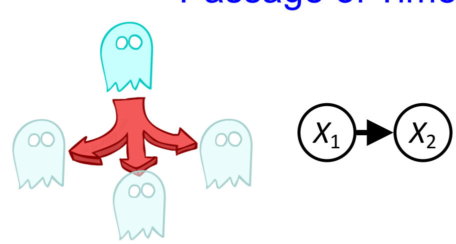
"Observation"



$$P(X_1|e_1)$$

$$\begin{aligned} P(x_1|e_1) &= P(x_1, e_1)/P(e_1) \\ &\propto_{X_1} P(x_1, e_1) \\ &= P(x_1)P(e_1|x_1) \end{aligned}$$

"Passage of Time"



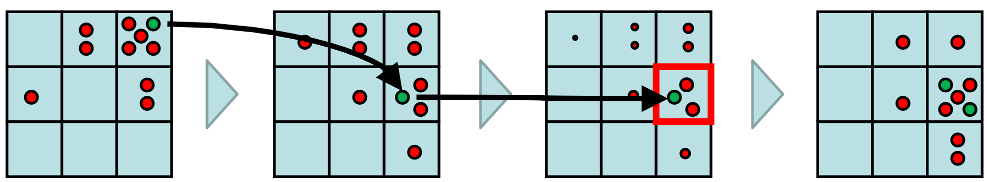
$$P(X_2)$$

$$\begin{aligned} P(x_2) &= \sum_{x_1} P(x_1, x_2) \\ &= \sum_{x_1} P(x_1)P(x_2|x_1) \end{aligned}$$

Particle Filtering: Summary

Particles: track samples of states rather than an explicit distribution

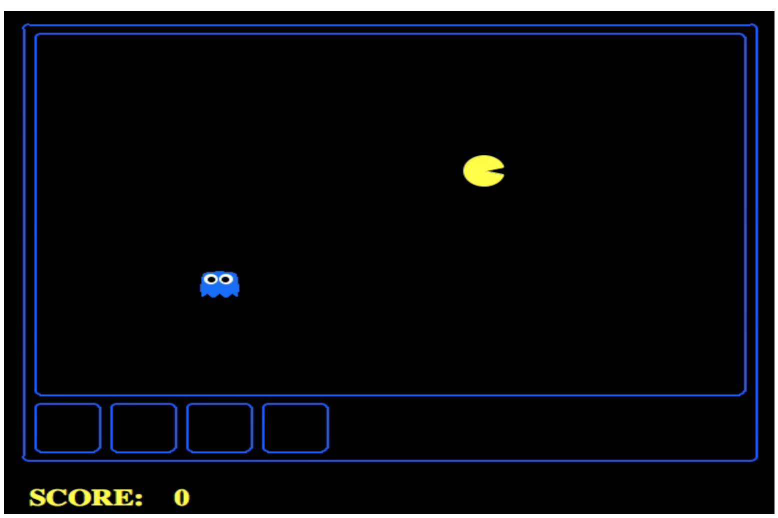
Elapse Weight Resample



Particles:	Particles:	Particles:	(New) Particles:
(3,3)	(3,2)	(3,2) w=.9	(3,2)
(2,3)	(2,3)	(2,3) w=.2	(2,2)
(3,3)	(3,2)	(3,2) w=.9	(3,2)
(3,2)	(3,1)	(3,1) w=.4	(2,3)
(3,3)	(3,3)	(3,3) w=.4	(3,3)
(3,2)	(3,2)	(3,2) w=.9	(3,2)
(1,2)	(1,3)	(1,3) w=.1	(1,3)
(3,3)	(2,3)	(2,3) w=.2	(2,3)
(3,3)	(3,2)	(3,2) w=.9	(3,2)
(2,3)	(2,2)	(2,2) w=.4	(3,2)

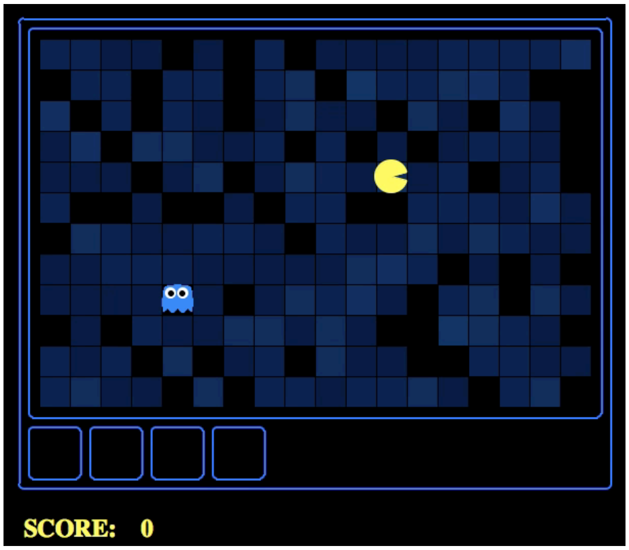
Which Algorithm?

Particle filter, uniform initial beliefs, 25 particles



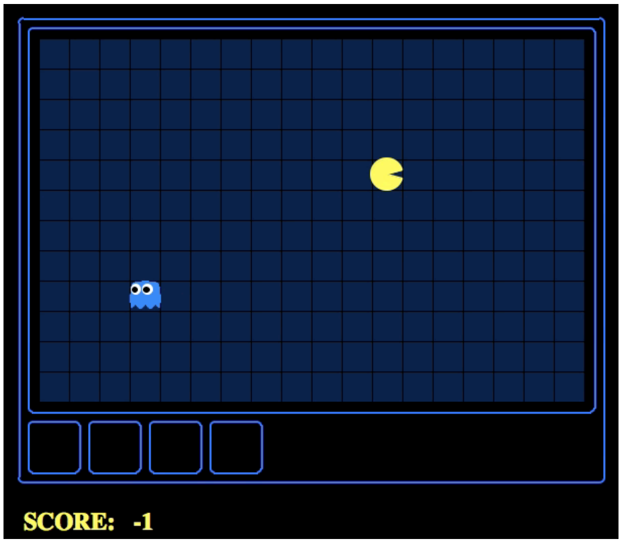
Which Algorithm?

Particle filter, uniform initial beliefs, 300 particles



Which Algorithm?

Exact filter, uniform initial beliefs



Complexity of the Forward Algorithm?

- We are given evidence at each time and want to know

$$B_t(X) = P(X_t|e_{1:t})$$

If only need P(x|e) at the end, only normalize there

- We use the single (time-passage+observation) updates:

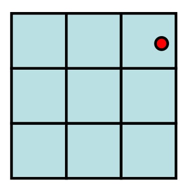
$$P(x_t|e_{1:t}) \propto_X P(e_t|x_t) \sum_{x_{t-1}} P(x_t|x_{t-1}) P(x_{t-1}, e_{1:t-1})$$

- Complexity? $O(|X|^2)$ time & O(X) space

But |X| is ***exponential*** in the number of state variables ☹

Why Does |X| Grow?

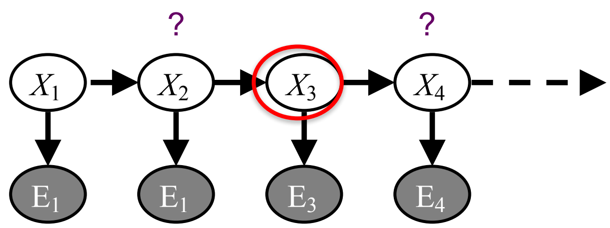
- 1 Ghost: k (eg 9) possible positions in maze
- 2 Ghosts: k^2 combinations
- N Ghosts: k^N combinations



11

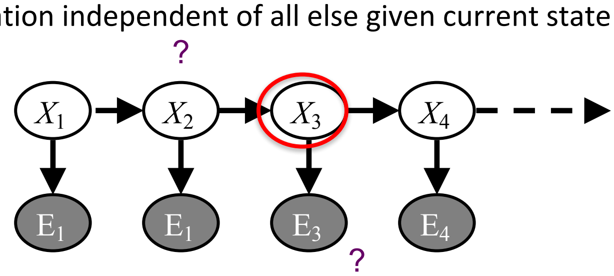
HMM Conditional Independence

- HMMs have two important independence properties:
  - Markov hidden process, future depends on past via the present



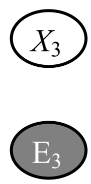
HMM Conditional Independence

- HMMs have two important independence properties:
  - Markov hidden process, future depends on past via the present
  - Current observation independent of all else given current state



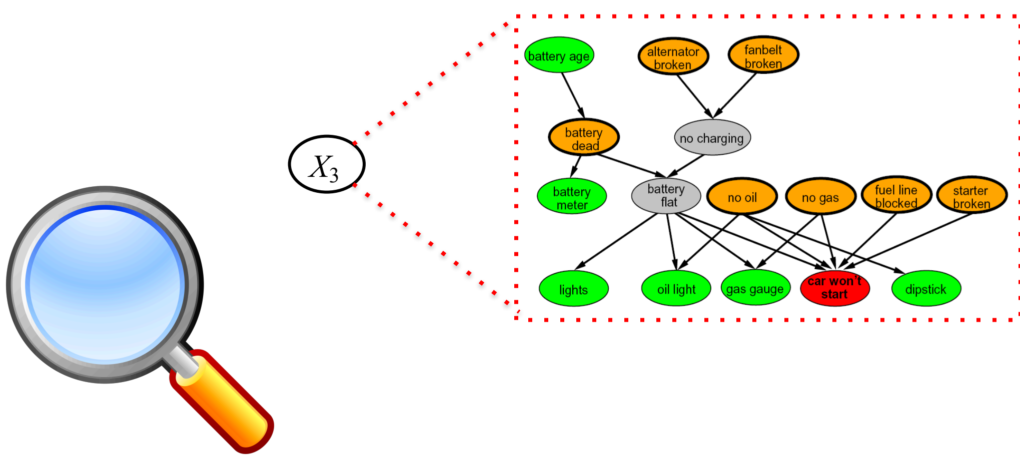
What about Conditional Independence ***in Snapshot***

- Can we do something here?
- Factor X into product of (conditionally) independent random vars?

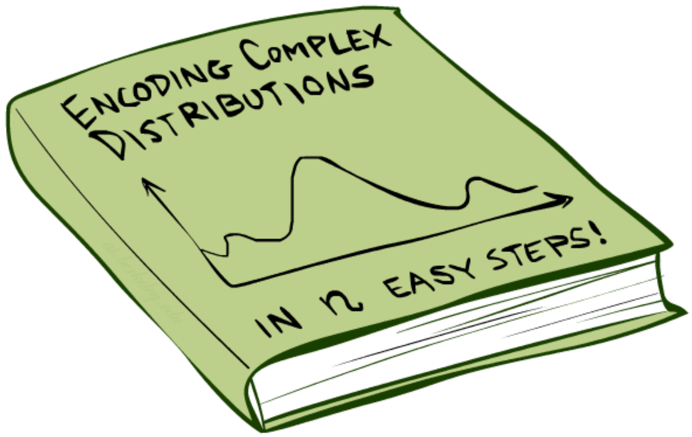


- Maybe also factor E

Yes! with Bayes Nets



Bayes'Nets: Big Picture

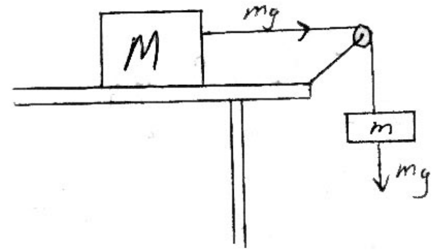


Bayes' Nets

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

Bayes Nets = a Kind of Probabilistic Graphical ***Model***

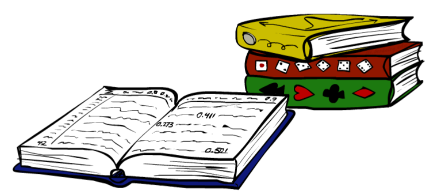
- Models describe how (a portion of) the world works
- Models are always simplifications
  - May not account for every variable
  - May not account for all interactions between variables
  - **"All models are wrong; but some are useful."**
    – George E. P. Box
- What do we do with probabilistic models?
  - We (or our agents) need to reason about unknown variables, given evidence
  - Example: explanation (diagnostic reasoning)
  - Example: prediction (causal reasoning)
  - Example: value of information

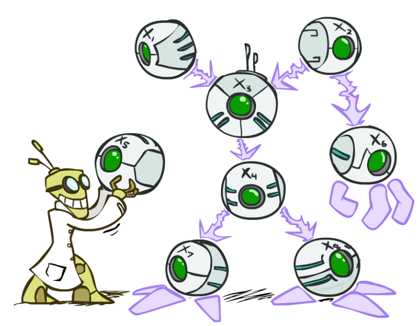


Friction,
Air friction,
Mass of pulley,
Inelastic string, …

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 … aka probabilistic graphical model
  - 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





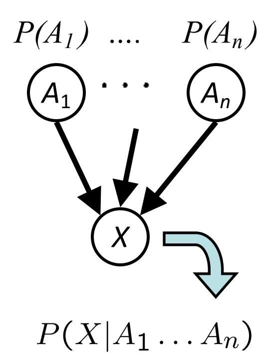
Bayes' Net Semantics

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

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

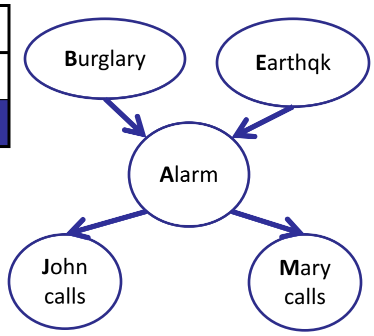
  - CPT: conditional probability table
  - Description of a noisy "causal" process



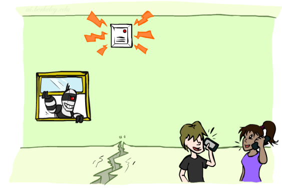
A Bayes net = Topology (graph) + Local Conditional Probabilities

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

Joint Probabilities from BNs

- Why are we guaranteed that setting $P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | \textit{parents}(X_i))$ results in a proper joint distribution?
- Chain rule (valid for all distributions): $P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | x_1 \ldots x_{i-1})$
- Assume conditional independences: $P(x_i | x_1, \ldots x_{i-1}) = P(x_i | \textit{parents}(X_i))$

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

- Every BN represents a joint distribution, but
- Not every distribution can be represented by a specific BN
  - The topology enforces certain conditional independencies

Example: Coin Flips

X_1 X_2 $\cdots$ X_n

$P(X_1)$

h	0.5
t	0.5

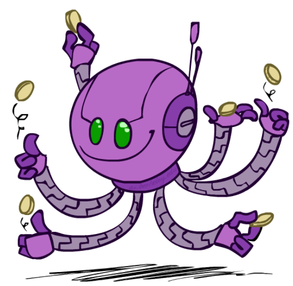
$P(X_2)$

h	0.5
t	0.5

$\cdots$

$P(X_n)$

h	0.5
t	0.5



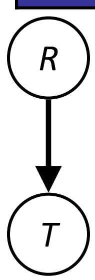
$$P(h, h, t, h) =$$

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



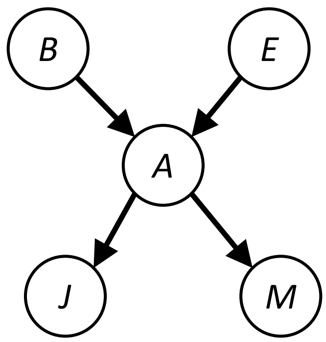
$P(T|R)$

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

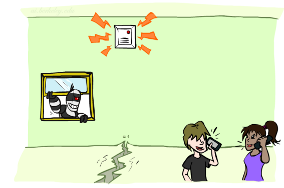
$P(+r, -t) =$ ¼ * ¼ = 1/16

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

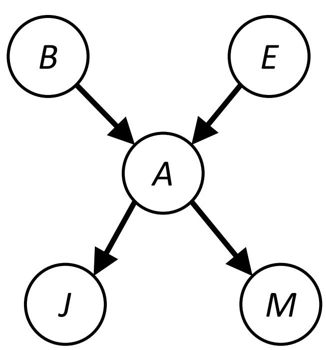
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

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

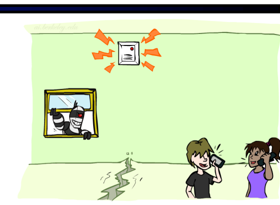
$$P(+b, -e, +a, -j, +m) =$$

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



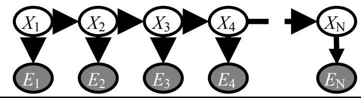
A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

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

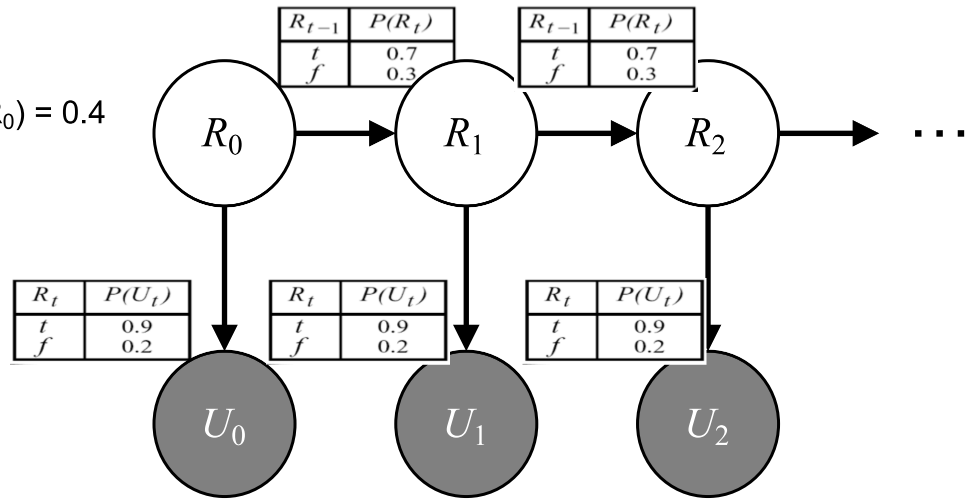
B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

$$P(+b, -e, +a, -j, +m) =$$
$$P(+b)P(-e)P(+a|+b,-e)P(-j|+a)P(+m|+a) =$$
$$0.001 \times 0.998 \times 0.94 \times 0.1 \times 0.7$$

Example: Hidden Markov Models

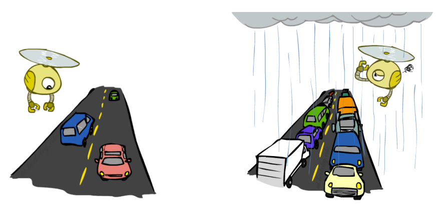


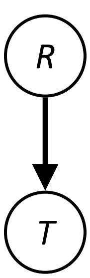
$P(R_0) = 0.4$



What Causes Bad Traffic?

- Causal direction





$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

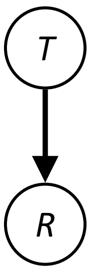
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(T, R)$

+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

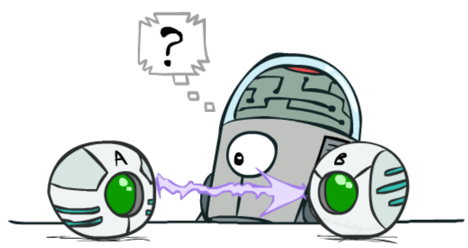
$P(T, R)$

+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

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



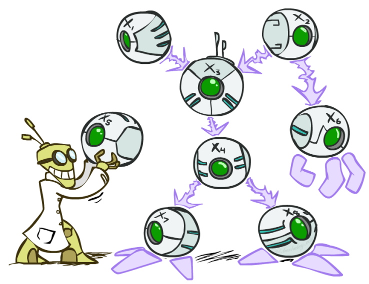
Summary: Bayes' Net Semantics

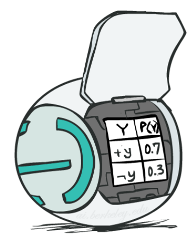
- A directed, acyclic graph, one node per random variable
- A conditional probability table (CPT) for each node
  - A collection of distributions over X, one for each combination of parents' values

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

- Bayes' nets ***compactly*** 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:

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





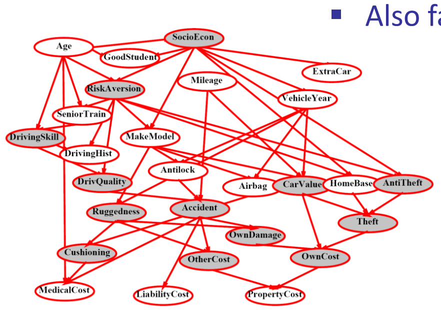
Size of a Bayes' Net

- 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?

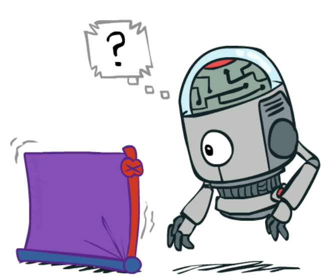
  $O(N * 2^k)$



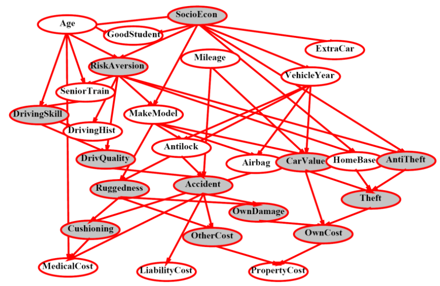
- Both give you the power to calculate

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

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



What's Next with Bayes' Nets



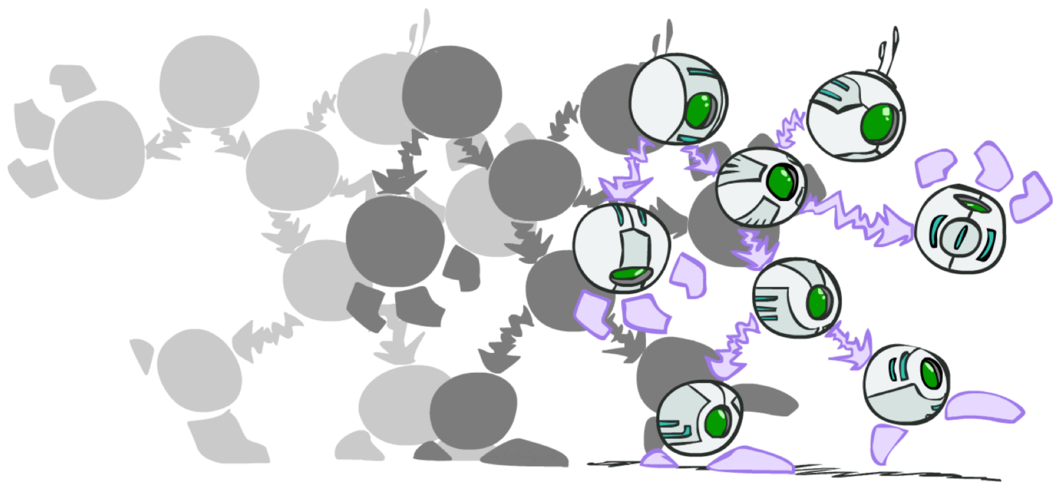
Questions we can ask:

- Definition: P(X = x)
- Inference: given a fixed BN, what is P(X | e)?
- Representation: given a BN graph, what kinds of distributions can it encode?
- Modeling: what BN is most appropriate for a given domain?
- Learning: Given data, what is best BN encoding?

Bayes' Nets

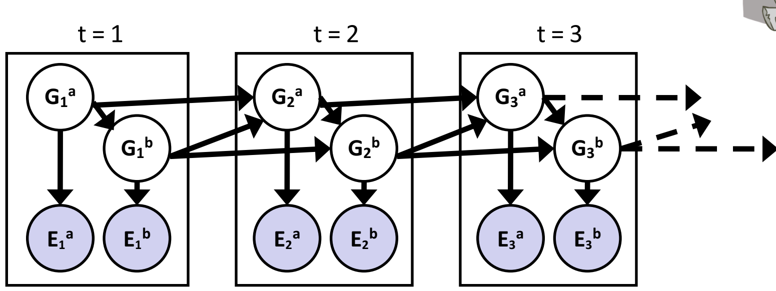
- ✓ Representation
  - Special case: HMMs & DBNs
- Conditional Independences
- Probabilistic Inference
- Learning Bayes' Nets from Data

Dynamic Bayes Nets



Dynamic Bayes Nets (DBNs)

- We want to track multiple variables over time, using multiple sources of evidence
- Idea: Repeat a fixed Bayes net structure at each time
- Variables from time *t* can condition on those from *t-1*



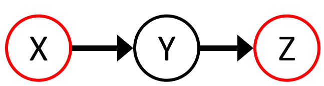
- Dynamic Bayes nets are a generalization of HMMs

DBN Particle Filters

- A particle is a complete sample for a time step
- **Initialize**: Generate prior samples for the t=1 Bayes net
  - Example particle: $\mathbf{G_1^a}$ = (3,3) $\mathbf{G_1^b}$ = (5,3)
- **Elapse time**: Sample a successor for each particle
  - Example successor: $\mathbf{G_2^a}$ = (2,3) $\mathbf{G_2^b}$ = (6,3)
- **Observe**: Weight each *entire* sample by the likelihood of the evidence conditioned on the sample
  - Likelihood: $P(\mathbf{E_1^a} \mid \mathbf{G_1^a}) * P(\mathbf{E_1^b} \mid \mathbf{G_1^b})$
- **Resample:** Select prior samples (tuples of values) in proportion to their likelihood

Conditional 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 1: 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)