Lecture 9
Bayesian Networks

Lei Li and Yu-xiang Wang
UCSB

Some slides adapted from Yexiang Xue, Pat Virtue

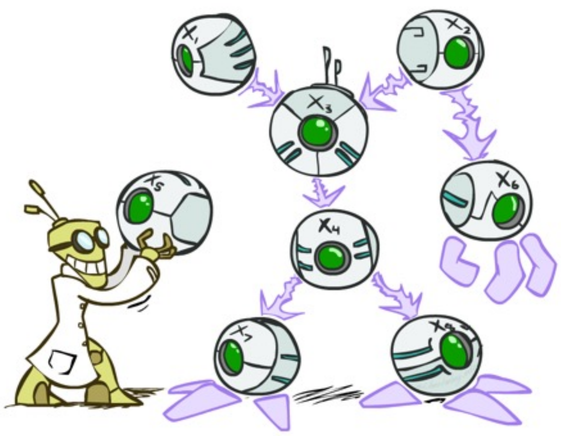
Recap

- Attention mechanism in neural networks
- Transformer
  - Multi-head attention
  - Positional embedding
  - Residual connection
  - Layer norm
  - Cross attention

Representing Probabilistic Dependency

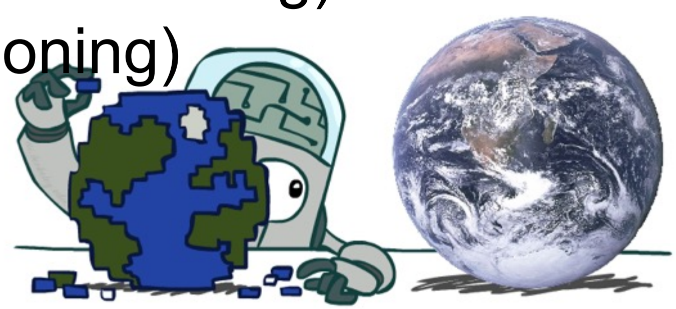
- Two problems with using full joint distribution tables as probabilistic models:
  - Unless there are only a few variables, the joint is WAY too big to represent explicitly
  - Hard to learn (estimate) anything empirically about more than a few variables at a time

- Bayesian networks: a technique for describing complex joint distributions (models) using simple, local distributions (conditional probabilities)
  - More properly called graphical models
  - Describe how variables locally interact; Local interactions chain together to give global, indirect interactions



Probabilistic Graphical Models

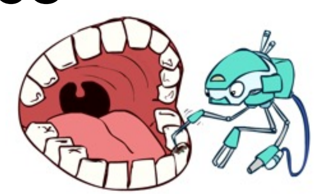
- Models describe how (a portion of) the world works
- Models are always simplifications
  - May not account for every variable
  - May not account for all interactions between variables
  - “All models are wrong; but some are useful.” – George E. P. Box
- What do we do with probabilistic models?
  -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



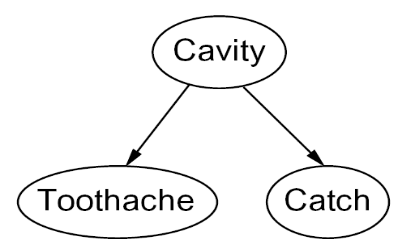
Bayesian Networks: Nodes and Arcs

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

- Arcs: interactions
  - Indicate “direct influence” between variables
  - Formally: encode conditional independence (more later)

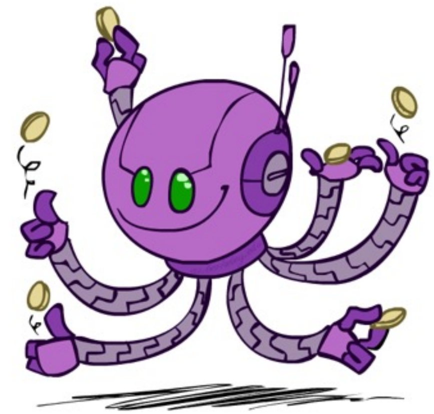


- For now: imagine that arrows mean direct causation (in general, they don’t!)



Example: Coin Flips

- N independent coin flips

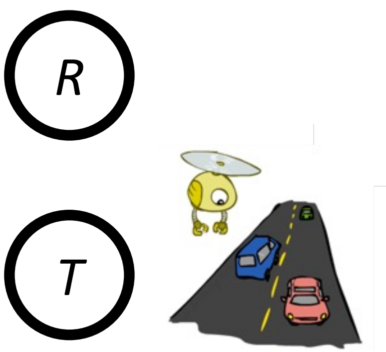


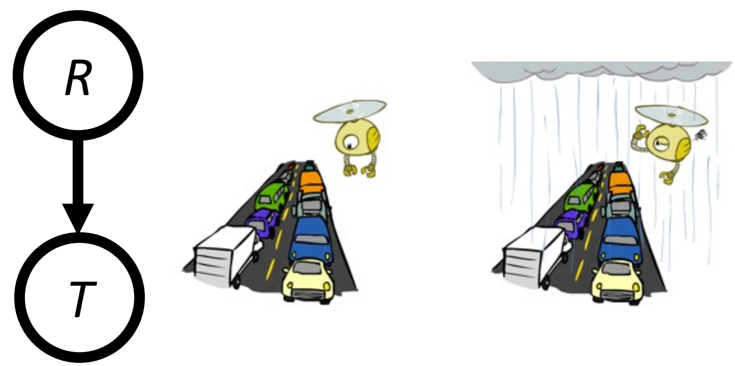
- No interactions between variables: absolute independence

Example: Rain and Traffic

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

Model 2: rain causes traffic

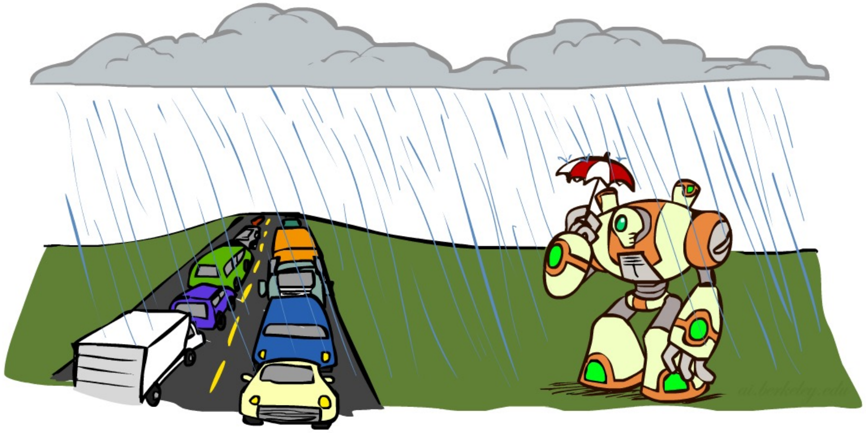


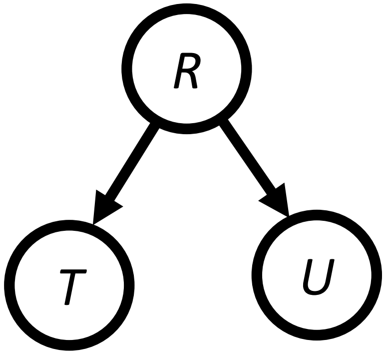


- Why is an agent using model 2 better?

Example: Traffic II

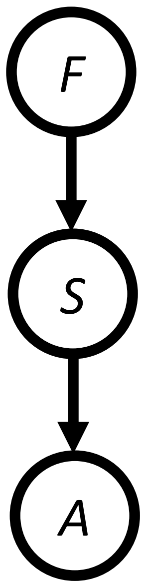
- Let's build a graphical model
- Variables
  - T: Traffic
  - R: It rains
  - U: Umbrella

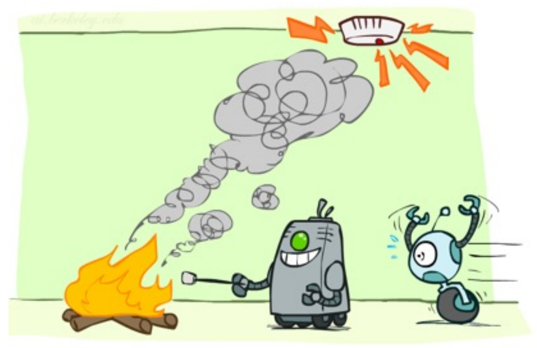


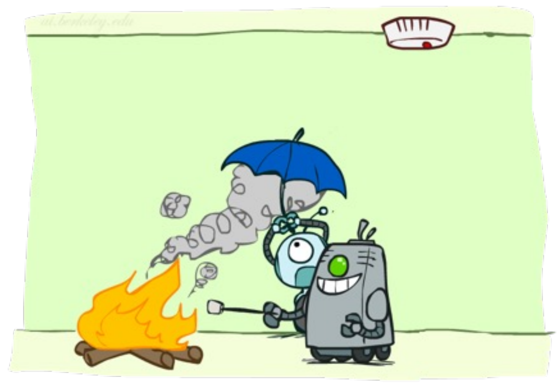


Example: fire, smoke, alarm

- Variables:
  - Fire
  - Smoke
  - Alarm

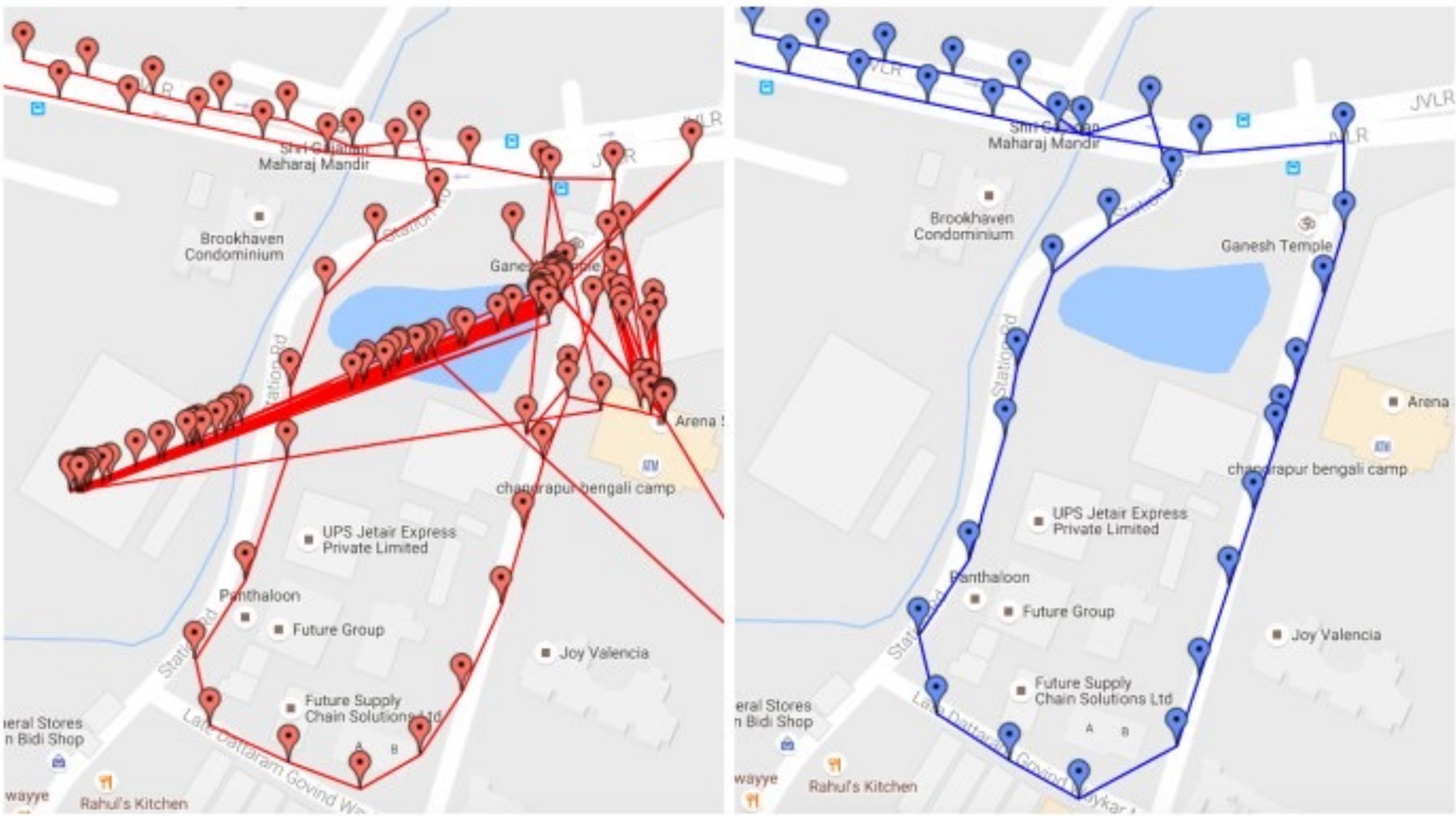






Example: localization

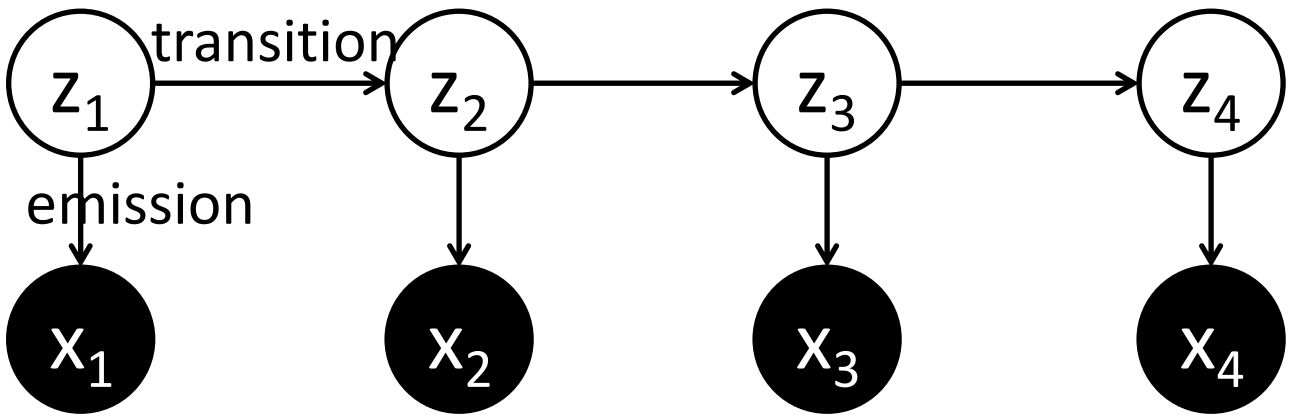
- GPS data can be noisy



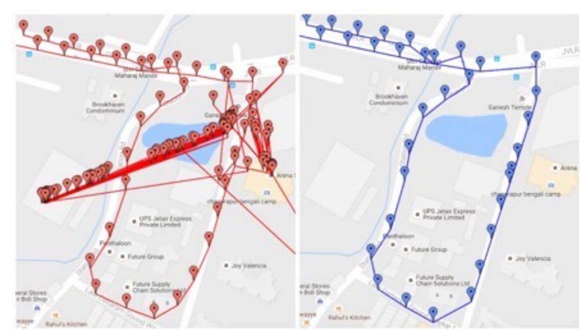
Example: localization

actual location, Velocity, acceleration

Observed location

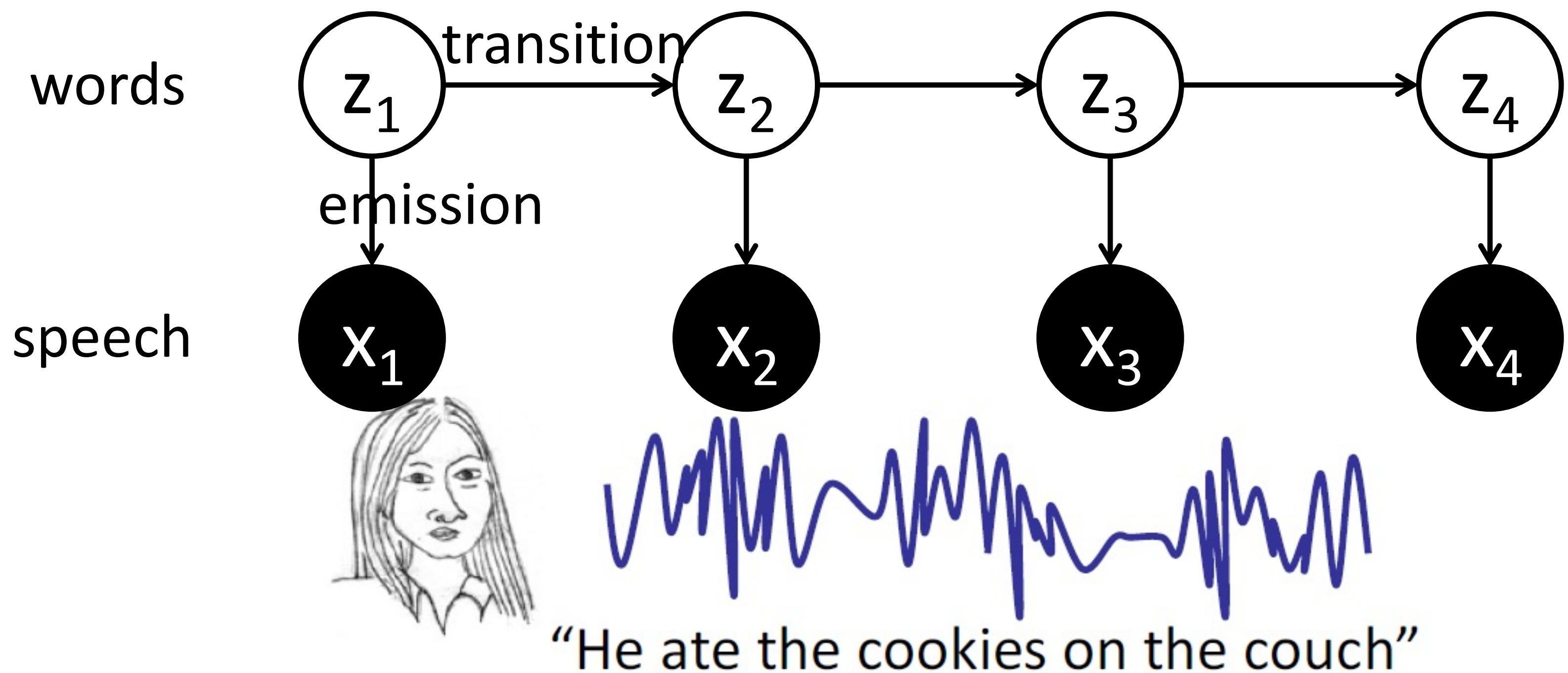


Linear dynamical systems



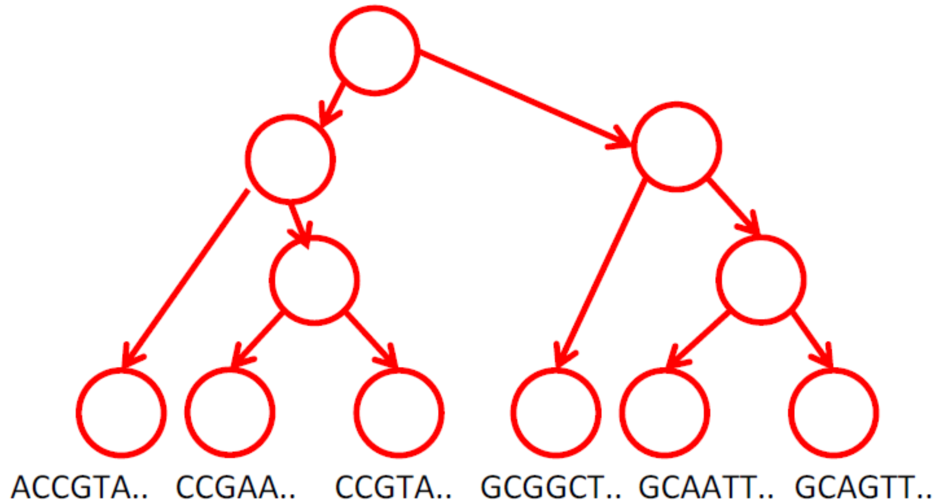
Example: automatic speech recognition (ASR)

- Infer spoken words from audio signals
- Hidden Markov models
- Could also be modeled using RNN/Transformer

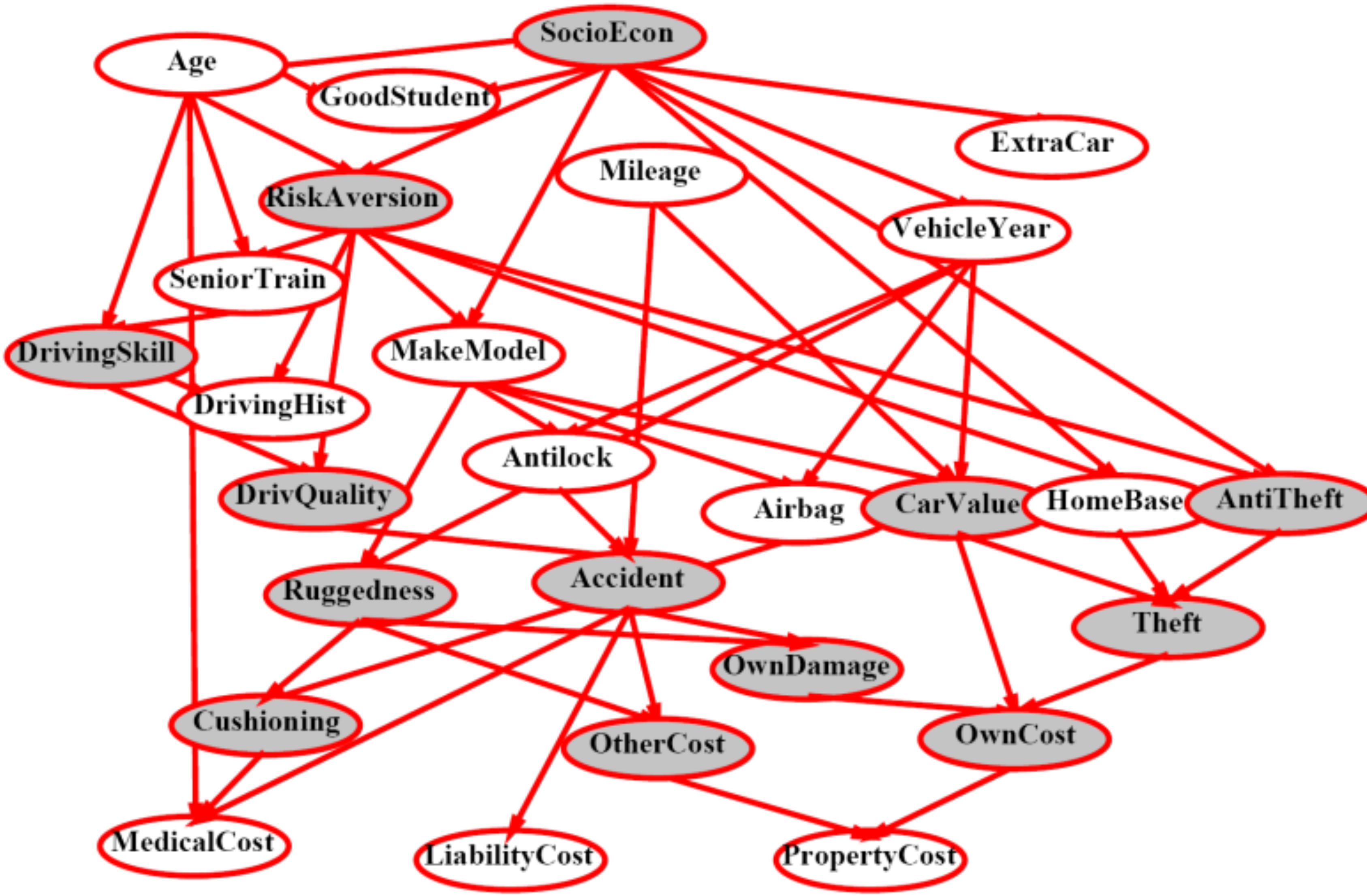


Example: evolutionary biology

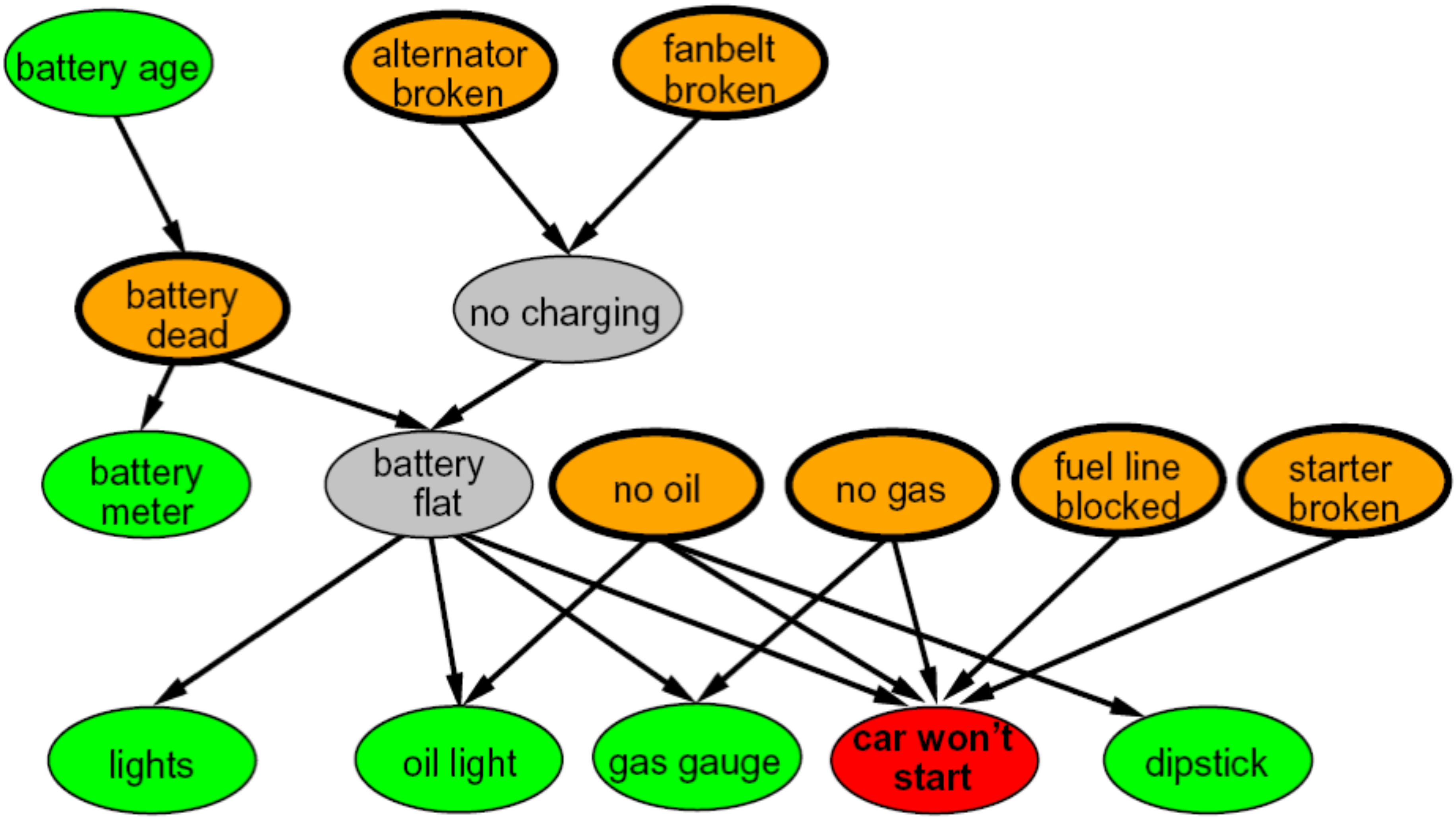
- Reconstruct a phylogenetic tree from DNA sequences of current species (Corvid-19)



Example: Insurance



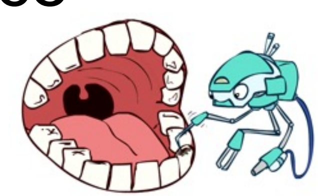
Example: Car diagnosis



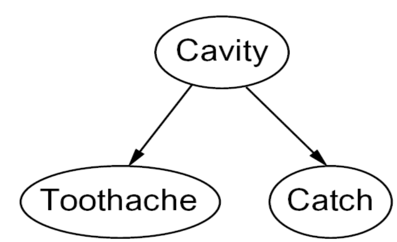
Bayesian Networks: Nodes and Arcs

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

- Arcs: interactions
  - Indicate “direct influence” between variables
  - Formally: encode conditional independence (more later)



- For now: imagine that arrows mean direct causation (in general, they don’t!)

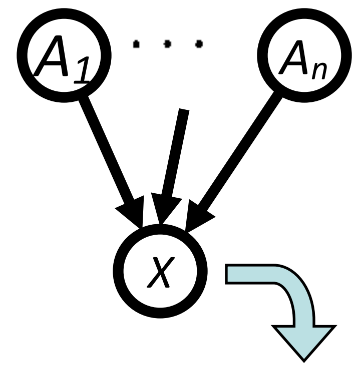


Bayesian network Semantics

- 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

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

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



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

Bayesian network = Topology (graph) + Local Conditional Probabilities

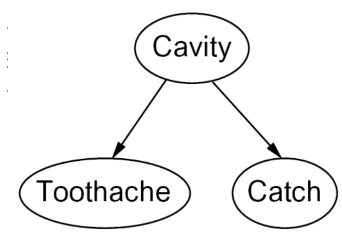
Probabilities in BNs

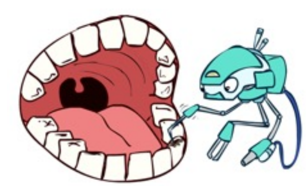
- Bayes' nets implicitly encode the joint distribution
  - As a product of local conditional distributions

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

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

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





Probabilities in BNs

- Why are we guaranteed that the following results in a proper joint distribution?

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

- 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, from topological order:

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

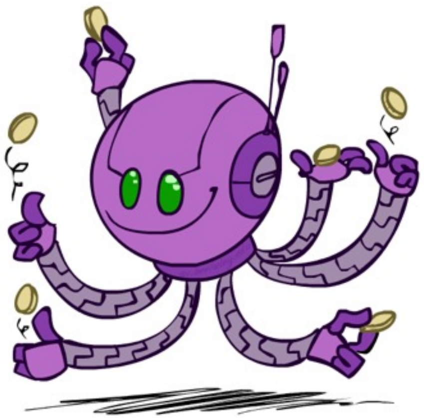
→ Consequence:

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

- Not every BN can represent every joint distribution
  - 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

$\cdots$

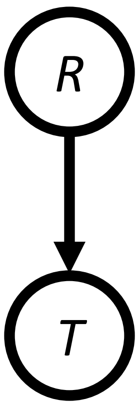
$P(X_n)$

h	0.5
t	0.5

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

Only distributions whose variables are absolutely independent can be represented by a Bayesian network with no arcs.

Example: Traffic



$P(R)$

+r	1/4
-r	3/4

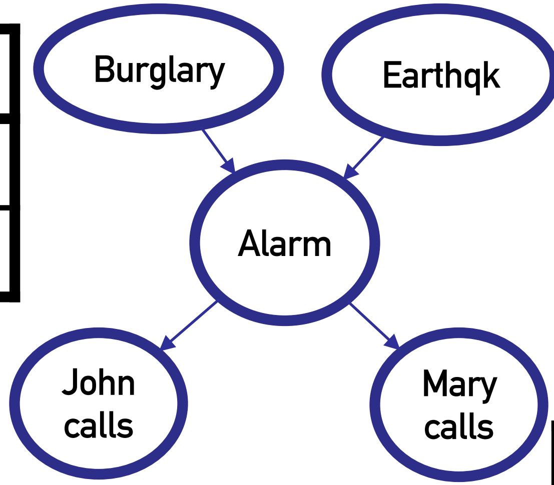
$$P(+r, -t) =$$

$P(T|R)$

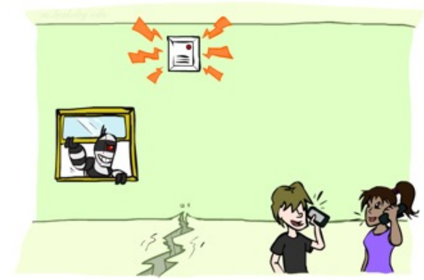
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

Example: Alarm Network

B	P(B)
+b	0.001
-b	0.999



E	P(E)
+e	0.002
-e	0.998



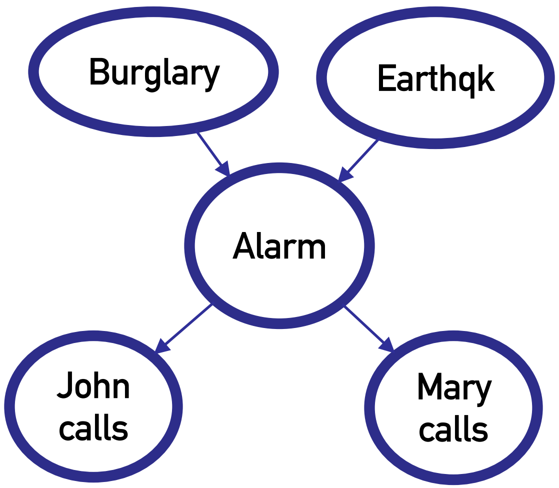
B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	0.94
+b	-e	-a	0.06
-b	+e	+a	0.29
-b	+e	-a	0.71
-b	-e	+a	0.001
-b	-e	-a	0.999

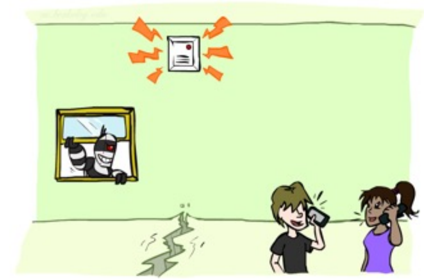
A	J	P(J|A)
+a	+j	0.9
+a	-j	0.1
-a	+j	0.05
-a	-j	0.95

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

B	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

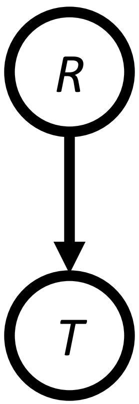
A	M	P(M|A)
+a	+m	0.7
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

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: Traffic

- Causal direction



$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

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

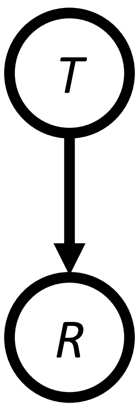
-r	+t	1/2
	-t	1/2

$P(T,R)$

+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 Bayesian networks 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 AirlineDelay
  - 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

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

Bayesian Networks

- So far: how a Bayesian network encodes a joint distribution
- Inference: How to answer numerical queries regarding marginal distribution of a variable given observations
- Learning: How to estimate parameters from data
- Structure learning: how to learn graphs

Inference by Enumeration

- Obvious problems:
  - Worst-case time complexity O(dn)
  - Space complexity O(dn) to store the joint distribution
  - Sample complexity (need many examples to estimate probabilities for full joint)
- Need new way of specifying the joint distribution!

Bayes Nets: Assumptions

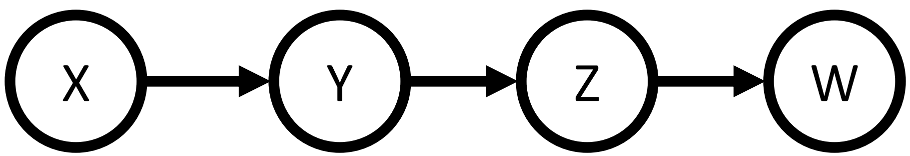
- Definition of Bayes net given the graph,:

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

- This assumes that a node is conditionally independent of other ancestors given its parents
- Often additional conditional independences, which can be read off the graph
- Important for modeling: understand assumptions made when choosing a Bayes net graphical structure

Example

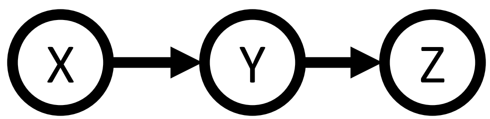
- Consider this chain shaped Bayesian Network:



- What conditional independence structures do we have?

Independence in a BN

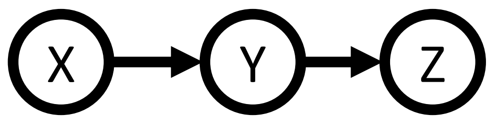
- Important question about a BN:
  - Are two nodes independent given certain evidence?
  - If yes, can prove using algebra (tedious in general). If no, can prove with a counter example
  - Example:

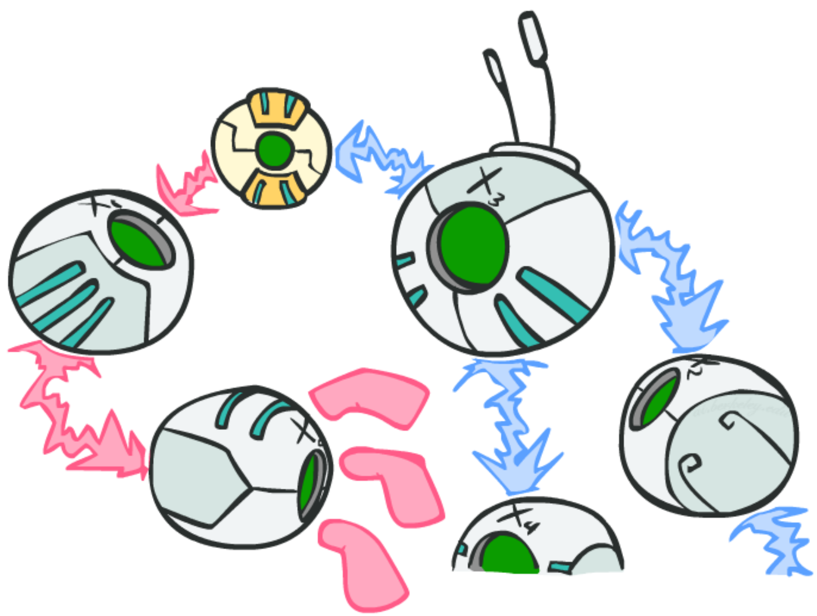


  - Question: are X and Z necessarily independent?
    - Answer: no. Example: low pressure causes rain, which causes traffic.
    - X can influence Z, Z can influence X (via Y)

Determining conditional independence via D-separation

- D-separation: a condition / algorithm for answering queries about independence





Causal Chains

- This configuration is a "causal chain"

X: Low pressure Y: Rain Z: Traffic

$$P(x,y,z) = P(x)P(y|x)P(z|y)$$

- Guaranteed X independent of Z ? **No**

- One example set of CPTs for which X is not independent of Z is sufficient to show this independence is not guaranteed

- Example:
  - Low pressure causes rain causes traffic, high pressure causes no rain causes no traffic
  - In numbers:
    P(+y | +x) = 1, P(-y | - x) = 1,
    P(+z | +y) = 1, P(-z | -y) = 1

Causal Chains

- This configuration is a “causal chain”

X: Low pressure Y: Rain Z: Traffic

$$P(x, y, z) = P(x)P(y|x)P(z|y)$$

- Guaranteed X independent of Z given Y?

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

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

$$= P(z|y)$$

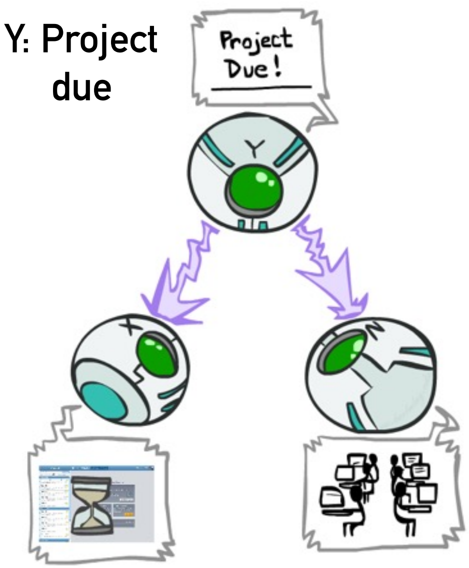
Yes!

- Evidence along the chain “blocks” the influence

Common Cause

- This configuration is a "common cause"

Y: Project due



X: Forums busy

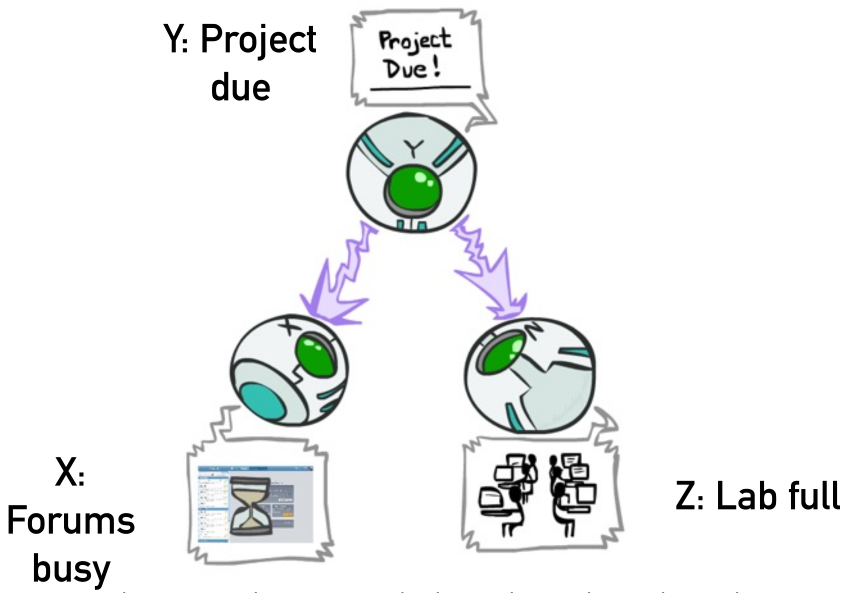
Z: Lab full

$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

- Guaranteed X independent of Z? **No**
  - One example set of CPTs for which X is not independent of Z is sufficient to show this independence is not guaranteed
  - Example:
    - Project due causes both forums busy and lab full
    - In numbers:
      P(+x | +y) = 1, P(-x | -y) = 1,
      P(+z | +y) = 1, P(-z | -y) = 1

Common Cause

- This configuration is a "common cause"



$$P(x, y, z) = P(y)P(x|y)P(z|y)$$

- Guaranteed X and Z independent given Y?

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

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

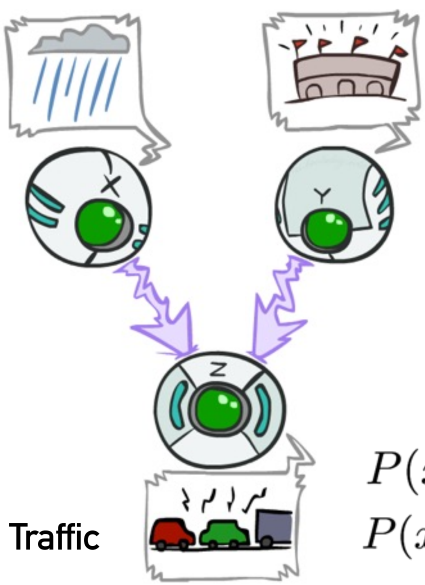
$$= P(z|y)$$

Yes!

- Observing the cause blocks influence between effects

Common Effect

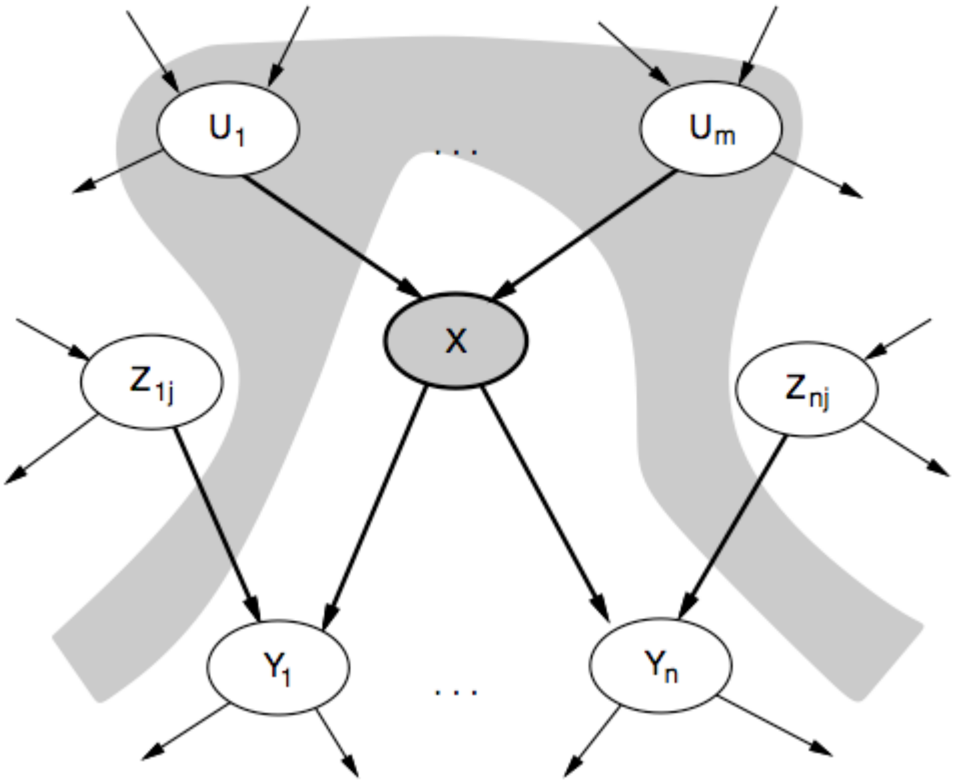
- Last configuration: two causes of one effect (*v-structure*)



$$P(x, y, z) = P(x)P(y)P(z|x, y)$$

- Are X and Y independent?
  - **Yes**: the ballgame and the rain cause traffic, but they are not correlated
    (Still need to prove this from Bayes net)
- Are X and Y independent given Z?
  - **No**: seeing traffic puts the rain and the ballgame in competition as explanation
- This is backwards from the other cases—Observing an effect activates influence between possible causes.

Conditional independence

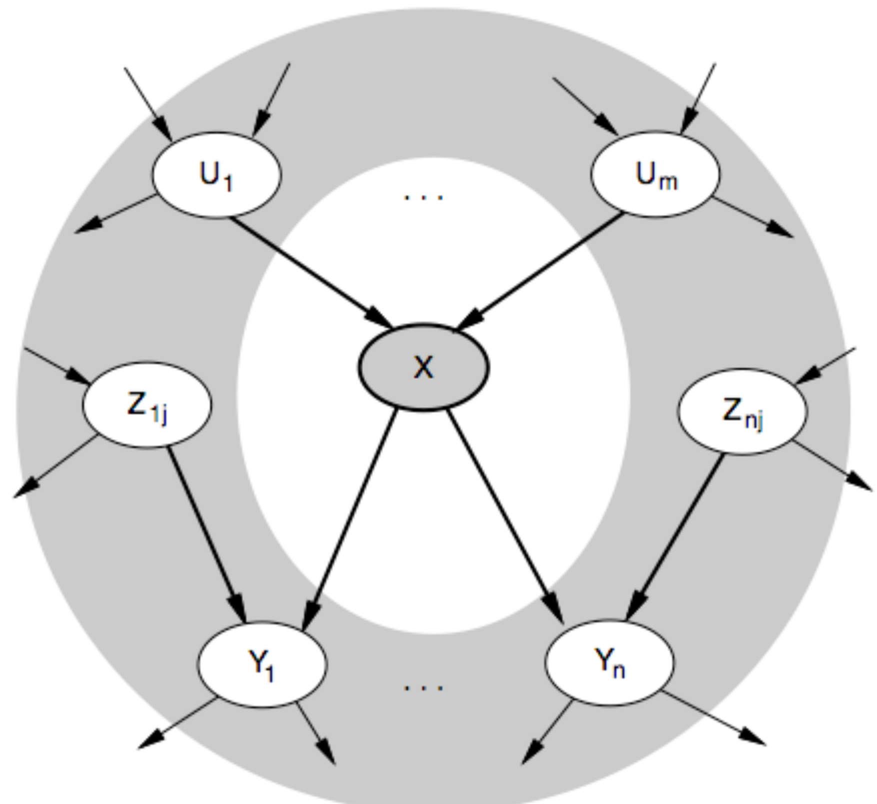


Each node is conditionally independent of its non-descendents given its parents

Markov Blanket

Each node is conditionally independent of the rest of the graph given its Markov Blanket

*The Markov blanket of a node A in a Bayesian network is the set of nodes composed of A's parents, A's children, and A's children's other parents.

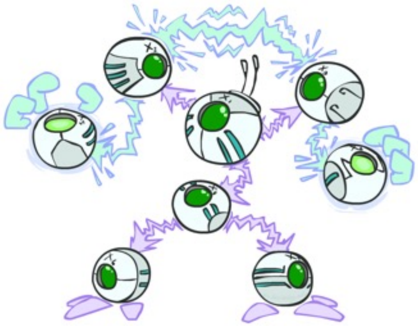


The General Case

- General question: in a given BN, are two variables independent (given evidence)?

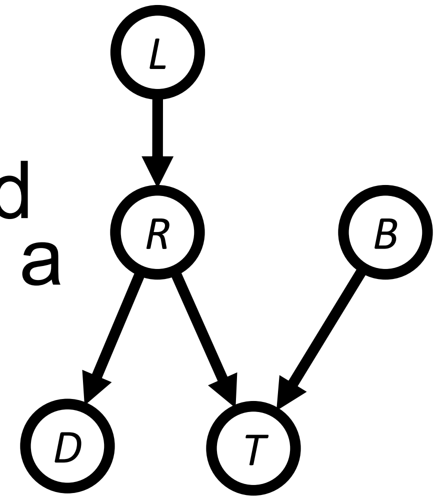
- Solution: analyze the graph

- Any complex example can be broken into repetitions of the three canonical cases



Reachability

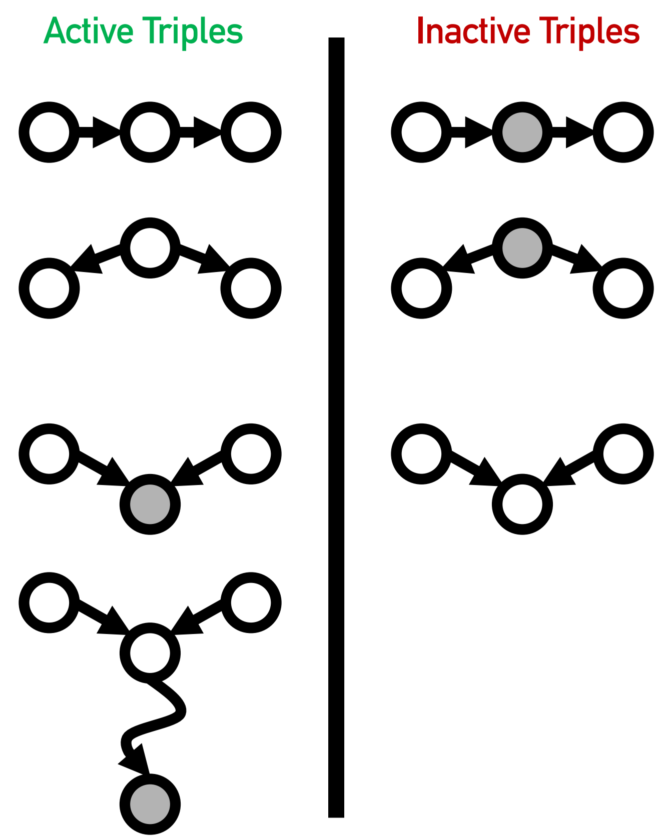
- **Recipe**: shade evidence nodes, look for paths in the resulting graph



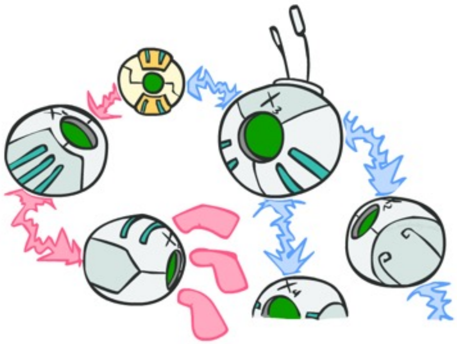
- Attempt 1: if two nodes are connected by an undirected path not blocked by a shaded node, they are conditionally independent
- Almost works, but not quite
  – Where does it break?
  – Answer: the v-structure at T doesn't count as a link in a path unless "active"

Active / Inactive Paths

- Question: Are X and Y conditionally independent given evidence variables {Z}?
  - Yes, if X and Y "d-separated" by Z
  - Consider all (undirected) paths from X to Y
  - No active paths = independence!
- A path is active if each triple is active:
  - Causal chain $A \rightarrow B \rightarrow C$ where B is unobserved (either direction)
  - Common cause $A \leftarrow B \rightarrow C$ where B is unobserved
  - Common effect (aka v-structure) $A \rightarrow B \leftarrow C$ where B *or one of its descendents* is observed
- All it takes to block a path is a single inactive segment



D-Separation



- Query: $X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$?
- Check all (undirected!) paths between X_i and X_j
  - If one or more active, then independence not guaranteed
  - Otherwise (i.e. if all paths are inactive),then independence is guaranteed

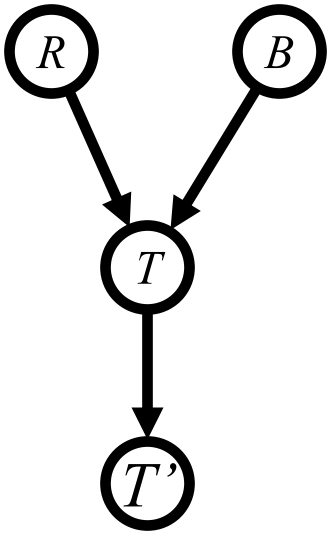
$$X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

Example

$R \perp\!\!\!\perp B$ Yes

$R \perp\!\!\!\perp B | T$

$R \perp\!\!\!\perp B | T'$



Example

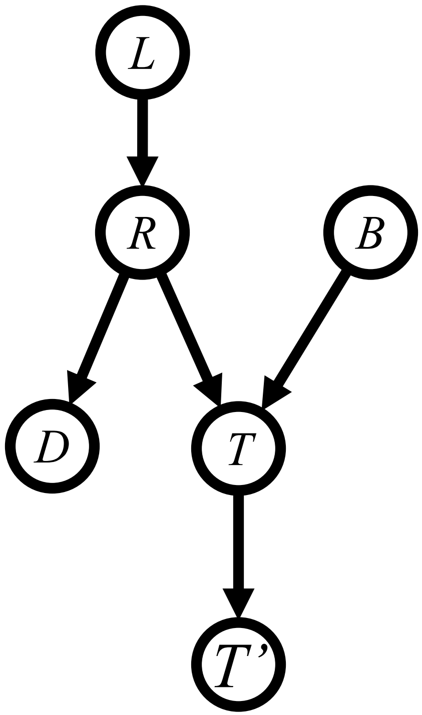
$L \perp\!\!\!\perp T' | T$ **Yes**

$L \perp\!\!\!\perp B$ **Yes**

$L \perp\!\!\!\perp B | T$

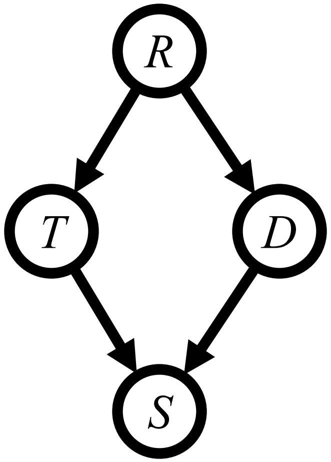
$L \perp\!\!\!\perp B | T'$

$L \perp\!\!\!\perp B | T, R$ **Yes**



Example

- Variables:
  - R: Raining; T: Traffic
  - D: Roof drips; S: I'm sad



- Questions:

$T \perp\!\!\!\perp D$

$T \perp\!\!\!\perp D | R$ **Yes**

$T \perp\!\!\!\perp D | R, S$

Structure Implications

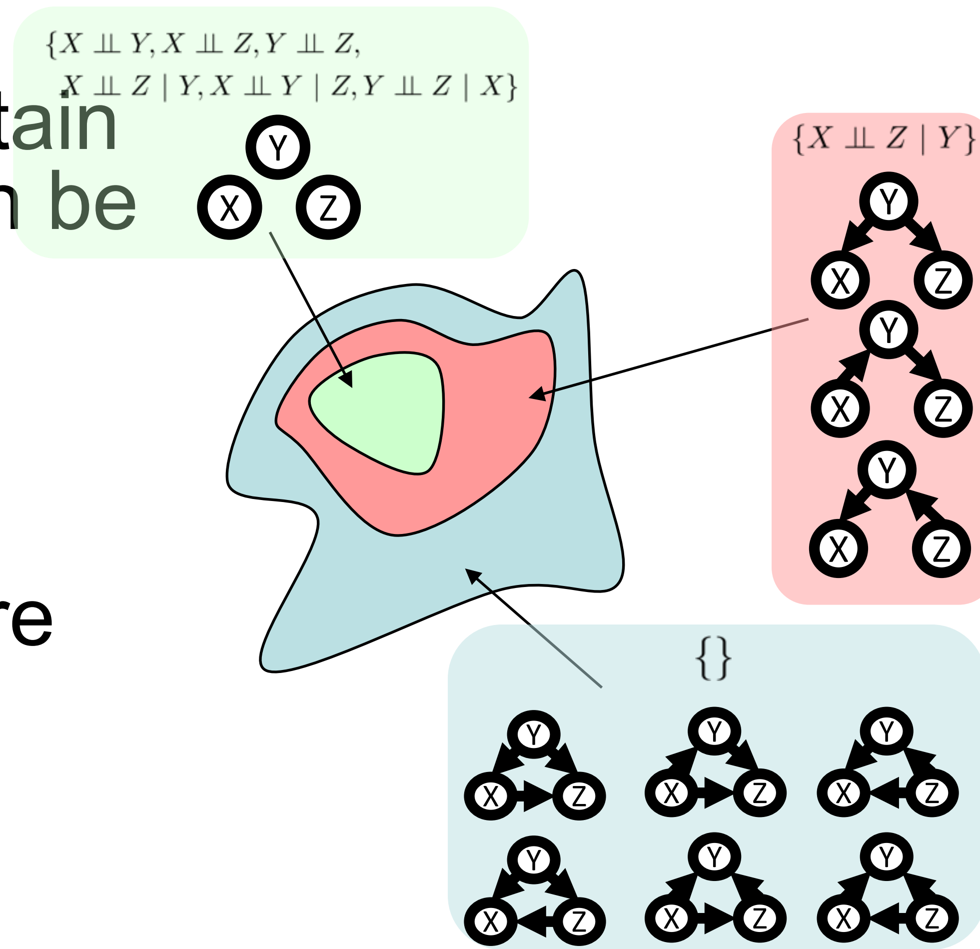
- Given a Bayes net structure, can run d-separation algorithm to build a complete list of conditional independences that are necessarily true of the form

$$X_i \perp\!\!\!\perp X_j | \{X_{k_1}, ..., X_{k_n}\}$$

- This list determines the set of probability distributions that can be represented

Graph Topology Limits Distributions

- Given some graph topology G, only certain joint distributions can be encoded
- The graph structure guarantees certain (conditional) independences (there might be more independence)



Bayes Networks: Summary

- Bayes nets compactly encode joint distributions
- Guaranteed independencies of distributions can be deduced from BN graph structure
- D-separation gives precise conditional independence guarantees from graph alone
- A Bayesian network's joint distribution may have further (conditional) independence that is not detectable until you inspect its specific (quantitative) distribution

Inference

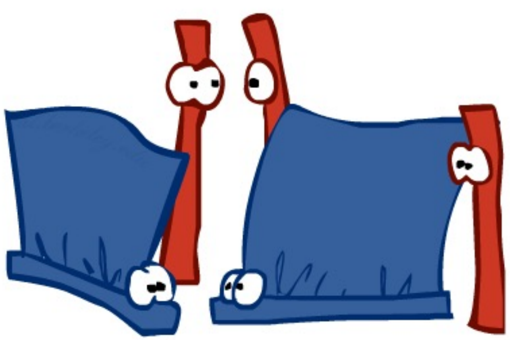
- Inference: calculating some useful quantity from a joint probability distribution

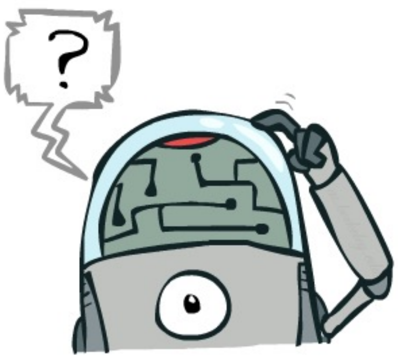
- Examples:
  - Posterior probability

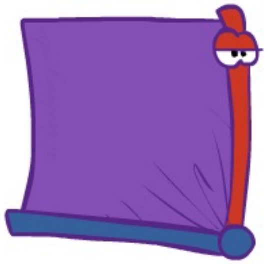
    $$P(Q|E_1 = e_1, \ldots E_k = e_k)$$

  - Most likely explanation:

    $$\text{argmax}_q \; P(Q = q|E_1 = e_1 \ldots)$$



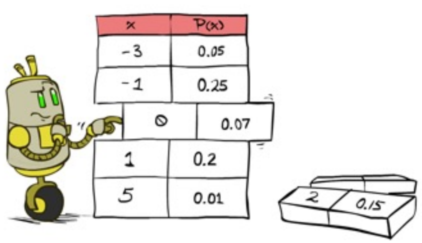




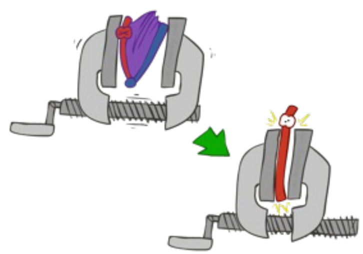
Inference by Enumeration

- General case: $P(Q|e_1 \dots e_k)$
  - Evidence variables: $E_1 \dots E_k = e_1 \dots e_k$
  - Query* variable: Q
  - Hidden variables: $H_1 \dots H_r$

- Step 1: Select the entries consistent with the evidence



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



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

- Step 3: Normalize

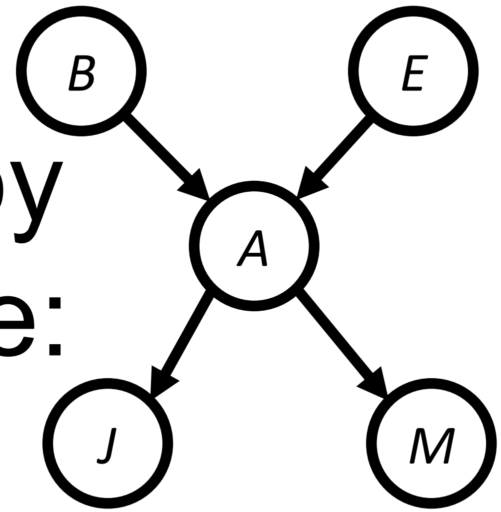
$$\times \frac{1}{Z}$$

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

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

Inference by Enumeration in Bayes Net

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



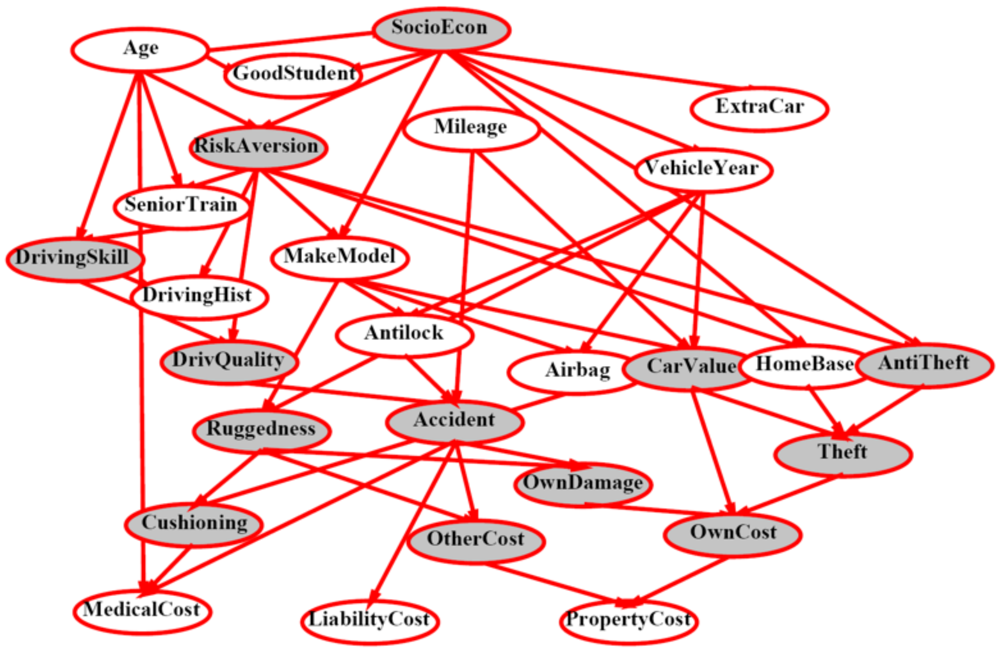
$$P(B \mid +j, +m) \quad \propto_B P(B, +j, +m)$$

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

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

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

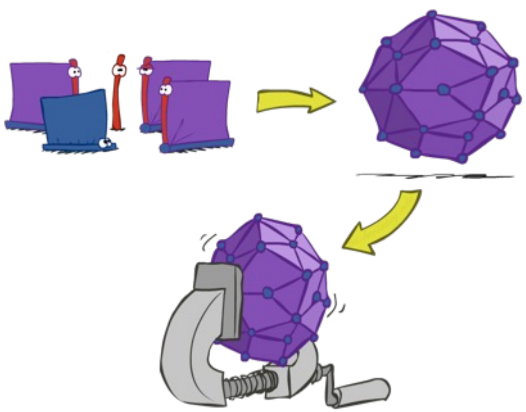
Inference by Enumeration?



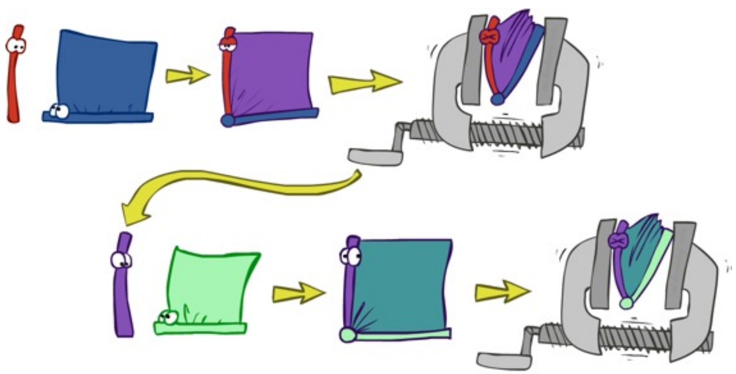
$$P(Antilock|observed\ variables) = ?$$

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



  - First we'll need some new notation: factors

Factor Zoo I

- Joint distribution: P(X,Y)
  - –Entries P(x,y) for all x, y
  - –Sums to 1
- Selected joint: P(x,Y)
  - –A slice of the joint distribution
  - –Entries P(x,y) for fixed x, all y
  - –Sums to P(x)
- Number of capital letters = dimensionality of the table

$P(T, W)$

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

$P(cold, W)$

T	W	P
cold	sun	0.2
cold	rain	0.3

Factor Zoo II

- Single conditional: P(Y | x)
  - Entries P(y | x) for fixed x, all y
  - Sums to 1

$P(W|cold)$

T	W	P
cold	sun	0.4
cold	rain	0.6

- Family of conditionals: P(X |Y)
  - Multiple conditionals
  - Entries P(x | y) for all x, y
  - Sums to |Y|

$P(W|T)$

T	W	P
hot	sun	0.8
hot	rain	0.2
cold	sun	0.4
cold	rain	0.6

$P(W|hot)$

$P(W|cold)$

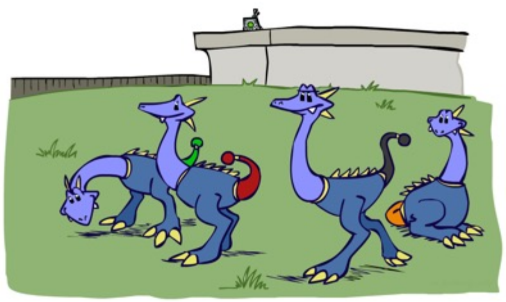
Factor Zoo III

- Specified family: P(y | X)
  - Entries P(y | x) for fixed y, but for all x
  - Sums to … unknown

$P(rain|T)$

T	W	P
hot	rain	0.2
cold	rain	0.6

$\}P(rain|hot)$
$\}P(rain|cold)$

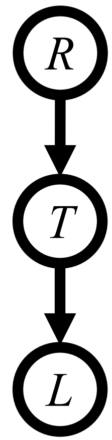


Factor Zoo Summary

- In general, when we write $P(Y_1 \ldots Y_N \mid X_1 \ldots X_M)$
- It is a "factor," a multi-dimensional array
- Its values are $P(y_1 \ldots y_N \mid x_1 \ldots x_M)$
- Any assigned (=lower-case) X or Y is a dimension missing (selected) from the array

Example: Traffic Domain

- Random Variables
  - R: Raining
  - T: Traffic
  - L: Late for class



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

$$P(L) = ?$$

$$= \sum_{r,t} P(r,t,L)$$

$$= \sum_{r,t} P(r)P(t|r)P(L|t)$$

Inference by Enumeration: Procedural Outline

- Track objects called **factors**
- Initial factors are local CPTs (one per node)

$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

- Any known values are selected, e.g. if we know $L = +\ell$, the initial factors are:

$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

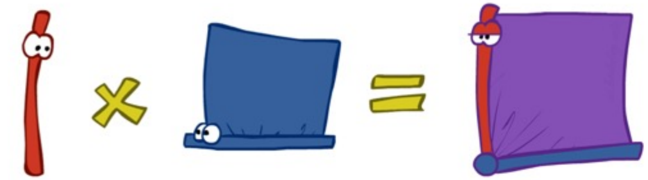
+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(+\ell|T)$

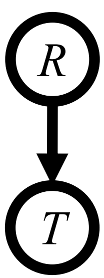
+t	+l	0.3
-t	+l	0.1

- Procedure: Join all factors, then eliminate all hidden variables

Operation 1: Join Factors



- First basic operation: **joining factors**
- Combining factors: (just like a database join)
  - Get all factors over the joining variable; Build a new factor over the union of the variables involved
- Example: Join on R



$P(R)$

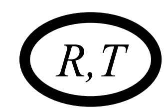
+r	0.1
-r	0.9

$\times$ $P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$\Rightarrow$ $P(R,T)$

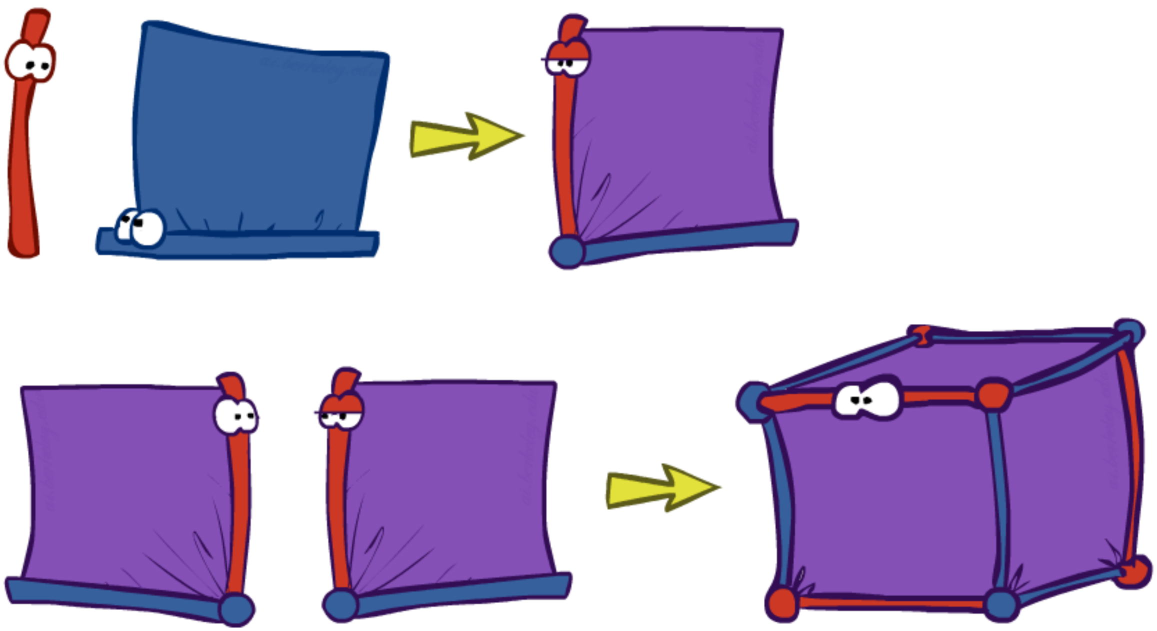
+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81



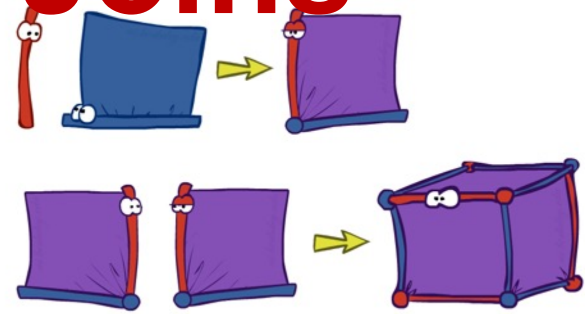
- Computation for each entry: pointwise products

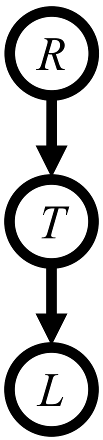
$$\forall r,t: \quad P(r,t) = P(r) \cdot P(t|r)$$

Example: Multiple Joins



Example: Multiple Joins





$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

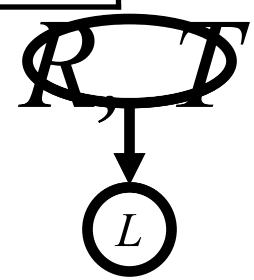
$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Join R

$P(R,T)$

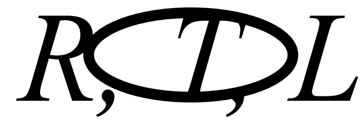
+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81



$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Join T



$P(R,T,L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

Operation 2: Eliminate

- Second basic operation: **marginalization**
- Take a factor and sum out a variable
  - Shrinks a factor to a smaller one
  - A **projection** operation
- Example:

$P(R,T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81

sum R →

$P(T)$

+t	0.17
-t	0.83

Multiple Elimination

R, T, L

$P(R, T, L)$

+r	+t	+l	0.024
+r	+t	-l	0.056
+r	-t	+l	0.002
+r	-t	-l	0.018
-r	+t	+l	0.027
-r	+t	-l	0.063
-r	-t	+l	0.081
-r	-t	-l	0.729

Sum out R

T, L

$P(T, L)$

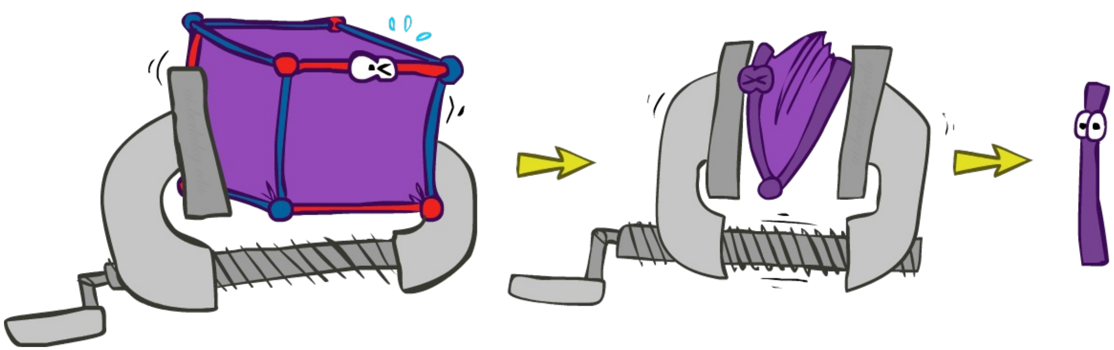
+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

Sum out T

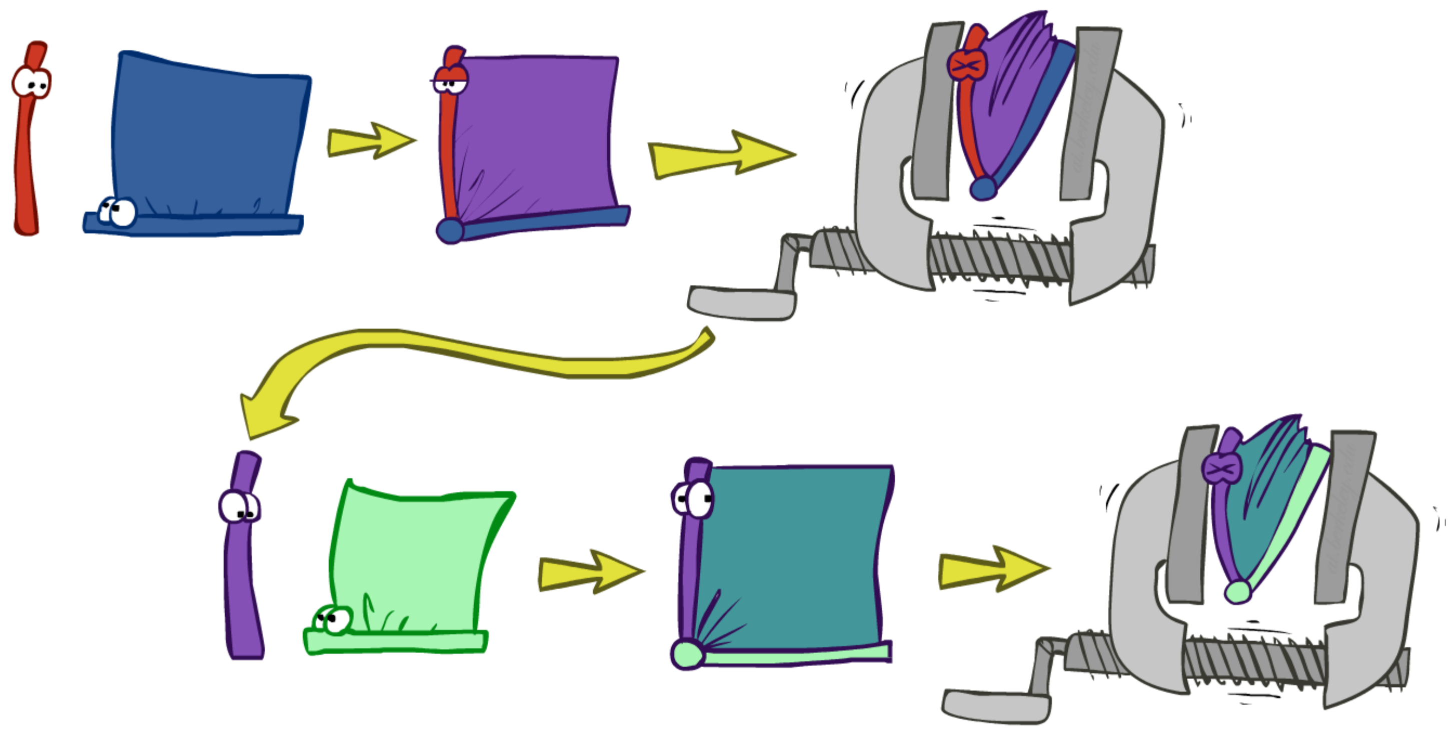
L

$P(L)$

+l	0.134
-l	0.886

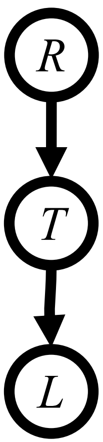


Variable Elimination = Marginalizing Early



Traffic Domain

- Inference by Enumeration



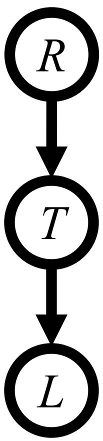
$$P(L) = \, ?$$

$$= \underbrace{\sum_t \underbrace{\sum_r \underbrace{P(L|t)\underbrace{P(r)P(t|r)}_{\text{Join on r}}}_{\text{Join on t}}}_{\text{Eliminate r}}}_{\text{Eliminate t}}$$

- Variable Elimination

$$= \underbrace{\sum_t \underbrace{P(L|t) \underbrace{\sum_r \underbrace{P(r)P(t|r)}_{\text{Join on r}}}_{\text{Eliminate r}}}_{\text{Join on t}}}_{\text{Eliminate t}}$$

Variable Elimination



$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

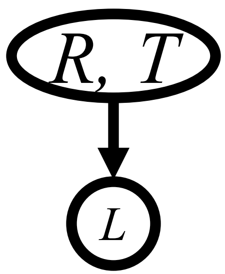
$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Join R

$P(R, T)$

+r	+t	0.08
+r	-t	0.02
-r	+t	0.09
-r	-t	0.81



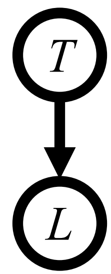
$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

Sum out R

$P(T)$

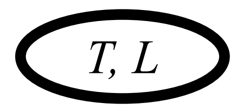
+t	0.17
-t	0.83



$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

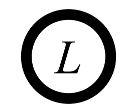
Join T



$P(T, L)$

+t	+l	0.051
+t	-l	0.119
-t	+l	0.083
-t	-l	0.747

Sum out T



$P(L)$

+l	0.134
-l	0.866

Evidence

- If evidence, start with factors that select that evidence
  - No evidence uses these initial factors:

$P(R)$

+r	0.1
-r	0.9

$P(T|R)$

+r	+t	0.8
+r	-t	0.2
-r	+t	0.1
-r	-t	0.9

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

  - To compute $P(L|+r)$, the initial factors become:

$P(+r)$

+r	0.1

$P(T|+r)$

+r	+t	0.8
+r	-t	0.2

$P(L|T)$

+t	+l	0.3
+t	-l	0.7
-t	+l	0.1
-t	-l	0.9

- Then we eliminate all variables other than query + evidence

Evidence

- Result will be a selected joint of query and evidence
  - To get our answer, just normalize — that 's it!
  - E.g. for P(L | +r), we would end up with:

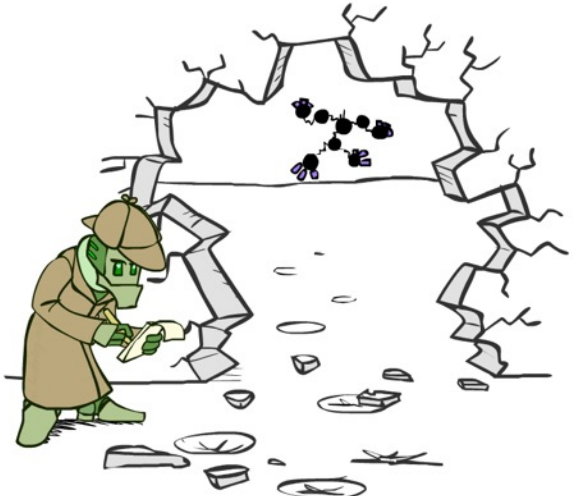
$P(+r, L)$

+r	+l	0.026
+r	-l	0.074

Normalize →

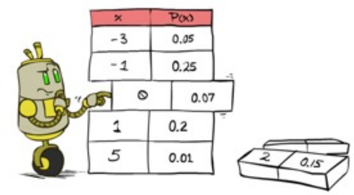
$P(L | +r)$

+l	0.26
-l	0.74

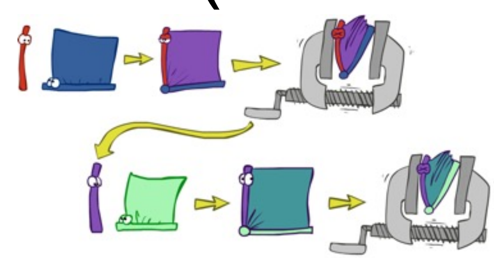


General Variable Elimination

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



- Start with initial factors:
  - Local CPTs (but instantiated by evidence)
- While there are still hidden variables (not Q or evidence):



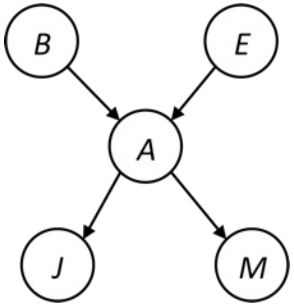
  - Pick a hidden variable H
  - Join all factors mentioning H
  - Eliminate (sum out) H

- Join all remaining factors and normalize $\times \frac{1}{Z}$

Example

$$P(B|j,m) \propto P(B,j,m)$$

$$\boxed{P(B) \qquad P(E) \qquad P(A|B,E) \qquad P(j|A) \qquad P(m|A)}$$



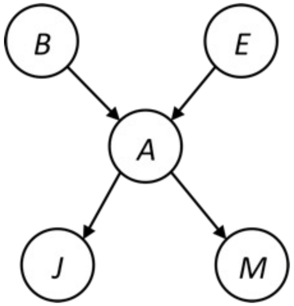
Choose A

$$\left.\begin{array}{l} P(A|B,E) \\ P(j|A) \\ P(m|A) \end{array}\right. \xrightarrow{\times} P(j,m,A|B,E) \xrightarrow{\sum} P(j,m|B,E)$$

$$\boxed{P(B) \qquad P(E) \qquad P(j,m|B,E)}$$

Example

$$P(B) \qquad P(E) \qquad P(j,m|B,E)$$



Choose E

$$P(E)$$
$$P(j,m|B,E)$$

$\times$

$$P(j,m,E|B)$$

$\sum$

$$P(j,m|B)$$

$$P(B) \qquad P(j,m|B)$$

Finish with B

$$P(B)$$
$$P(j,m|B)$$

$\times$

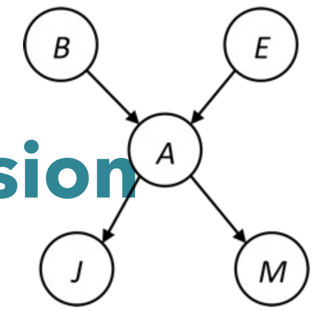
$$P(j,m,B)$$

Normalize

$$P(B|j,m)$$

Same Example in Equations

marginal can be obtained from joint by summing out $P(B|j,m) \propto P(B,j,m)$



use Bayes' net joint distribution expression

$$P(B) \quad P(E) \quad P(A|B,E) \quad P(j|A) \quad P(m|A)$$

use x*(y+z) = xy + xz

joining on a, and the

use x*(y+z) = xy + x:

joining on e, and the

$$\begin{aligned}
P(B|j,m) &\propto P(B,j,m) \\
&= \sum_{e,a} P(B,j,m,e,a) \\
&= \sum_{e,a} P(B)P(e)P(a|B,e)P(j|a)P(m|a) \\
&= \sum_{e} P(B)P(e) \sum_{a} P(a|B,e)P(j|a)P(m|a) \\
&= \sum_{e} P(B)P(e) f_1(B,e,j,m) \\
&= P(B) \sum_{e} P(e) f_1(B,e,j,m) \\
&= P(B) f_2(B,j,m)
\end{aligned}$$

All we are doing is exploiting uwy + uwz + uxy + uxz + vwy + vwz + vxy +vxz = (u+v)(w+x)(y+z) to improve computational efficiency!

Summary

- Bayesian networks:
  - Directed acyclic graph
  - Nodes are random variables
  - arcs are probabilistic dependencies
- Examine dependence of two variables given observation: d-separation
- Inference:
  - Variable elimination for discrete variables

Next up

- Gaussian Mixture Model
- Linear Dynamical Systems
- Learning parameters for BNs