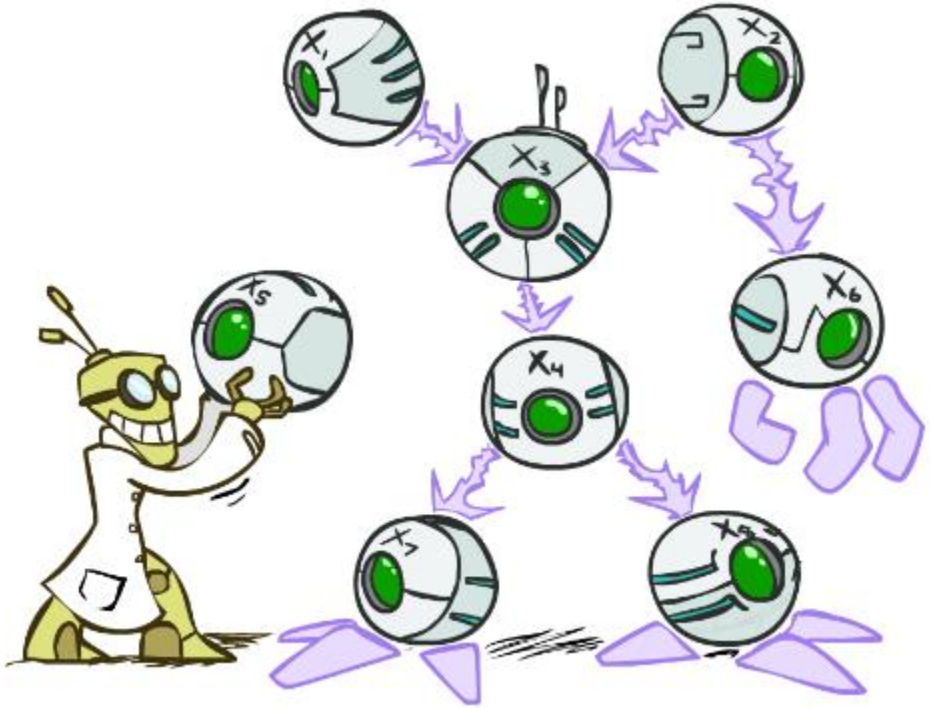
Artificial Intelligence

Bayes' Nets



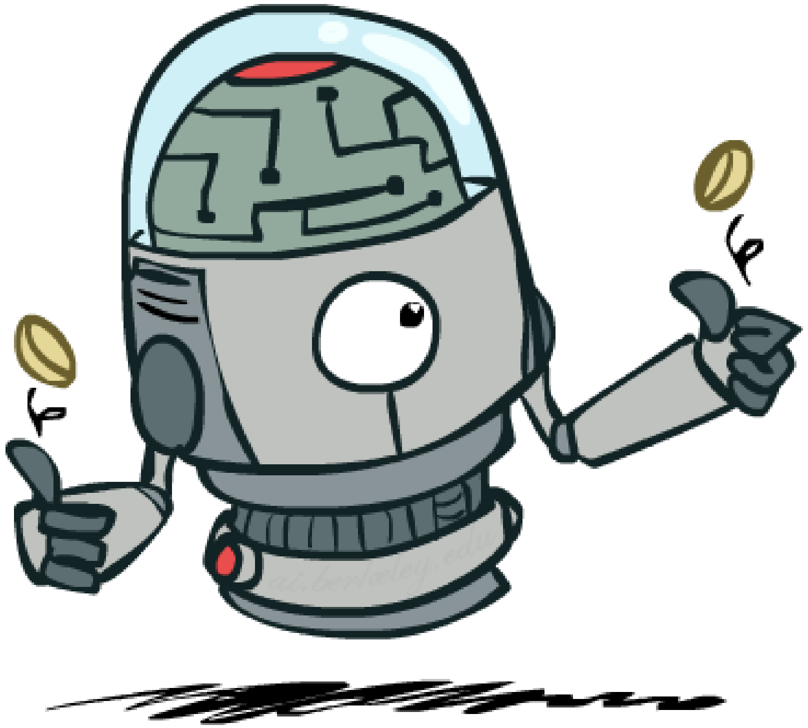
Probabilistic Models

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

- Models are always simplifications
  - May not account for every variable
  - May not account for all interactions between variables
  - "All models are wrong; but some are useful."
    – George E. P. Box

- What do we do with probabilistic models?
  - We (or our agents) need to reason about unknown variables, given evidence
  - Example: explanation (diagnostic reasoning)
  - Example: prediction (causal reasoning)

Independence



Independence

- Two variables are *independent* iff:

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

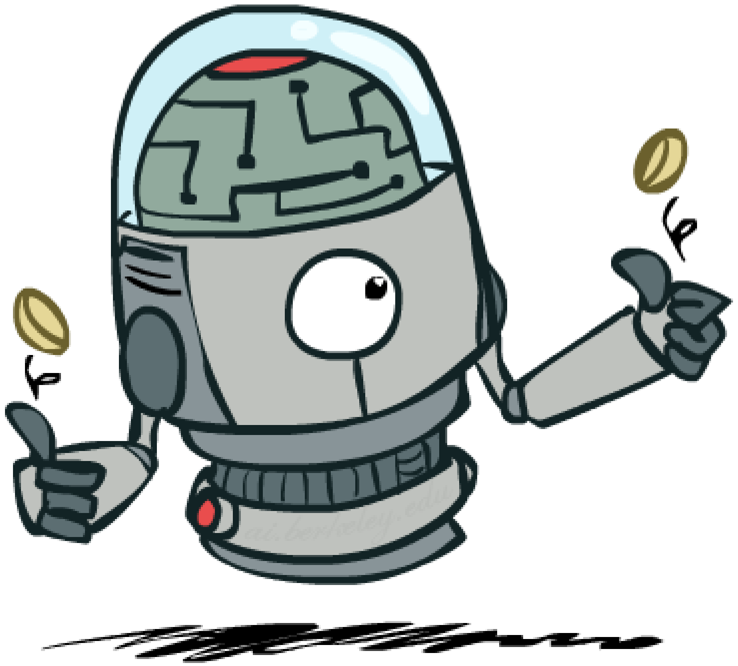
  - This says that their joint distribution *factors* into a product two simpler distributions
  - Another form:

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

$$\boxed{\forall x, y : P(y|x) = P(y)}$$ alternative

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

- Independence is a simplifying *modeling assumption*
  - *Empirical* joint distributions: at best "close" to independent
  - What could we assume for {Weather, Traffic, Cavity, Toothache}?



Example: Independence?

$P(T)$

T	P
hot	0.5
cold	0.5

$P_1(T, W)$

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

$P(W)$

W	P
sun	0.6
rain	0.4

$P_2(T, W) = P(W)P(T)$

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

Example: Independence

- N fair, independent coin flips:

$P(X_1)$

H	0.5
T	0.5

$P(X_2)$

H	0.5
T	0.5

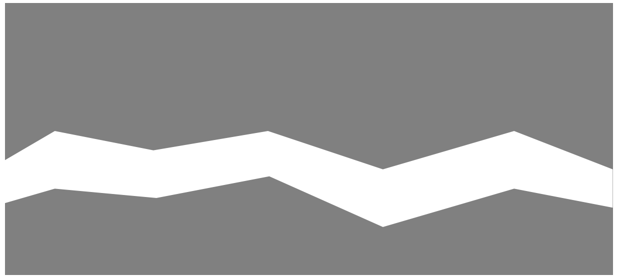
$\ldots$

$P(X_n)$

H	0.5
T	0.5

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

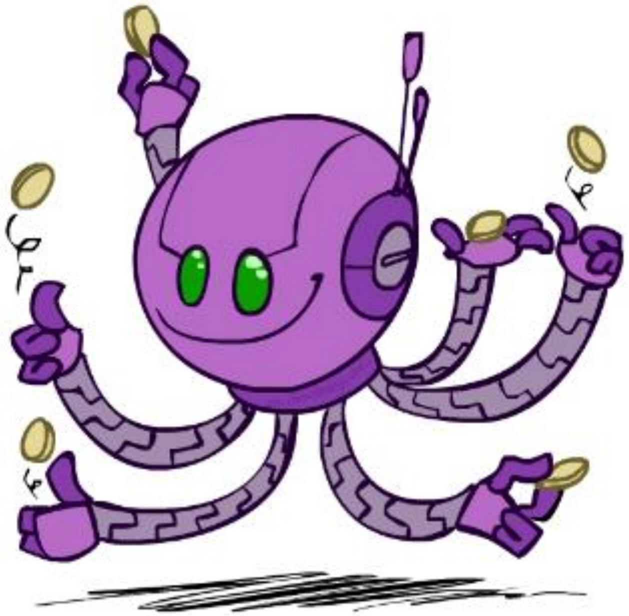
2^n



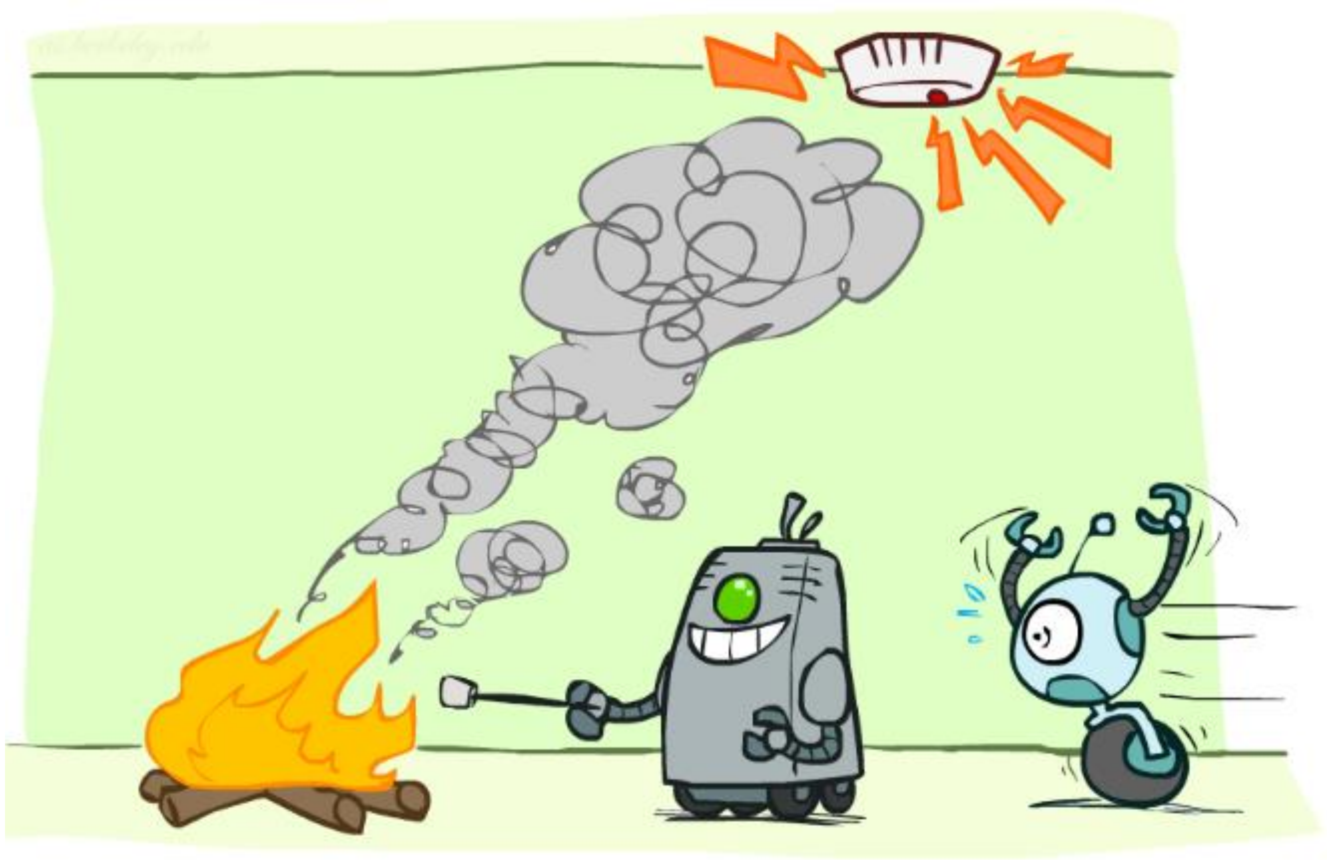
$2n$

$\wedge$

2^n



Conditional Independence



Conditional Independence

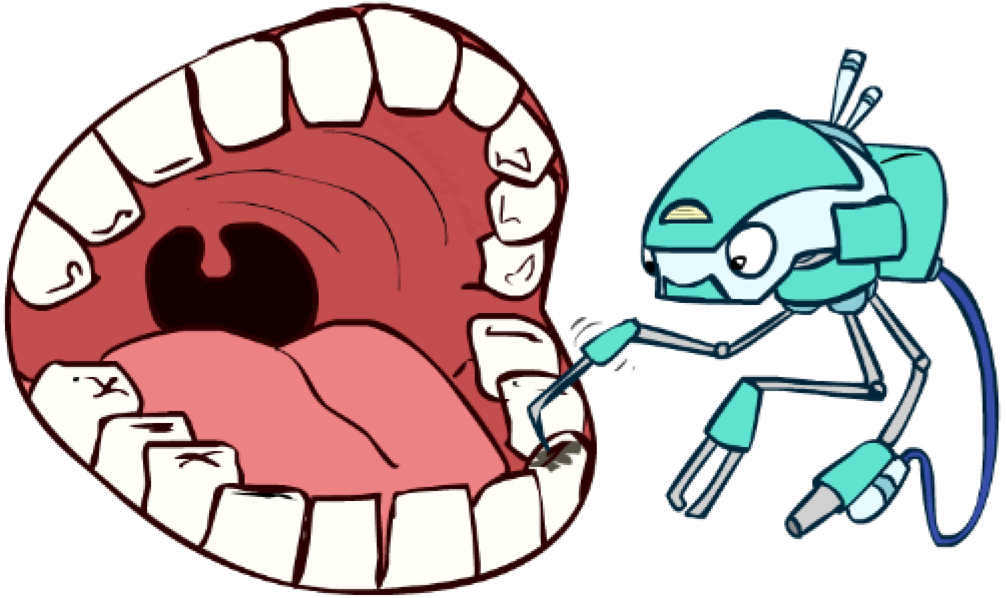
- P(Toothache, Cavity, Catch)

- If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:
  - P(+catch | +toothache, +cavity) = P(+catch | +cavity)

- The same independence holds if I don’t have a cavity:
  - P(+catch | +toothache, -cavity) = P(+catch| -cavity)

- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)

- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
  - One can be derived from the other easily



In analogy with independence

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

Conditional Independence

- Unconditional (absolute) independence very rare (why?)
- *Conditional independence* is our most basic and robust form of knowledge about uncertain environments.
- X is conditionally independent of Y given Z

$$X \perp\!\!\!\perp Y \mid Z$$

if and only if:

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

or, equivalently, if and only if

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

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

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

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

Conditional Independence

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

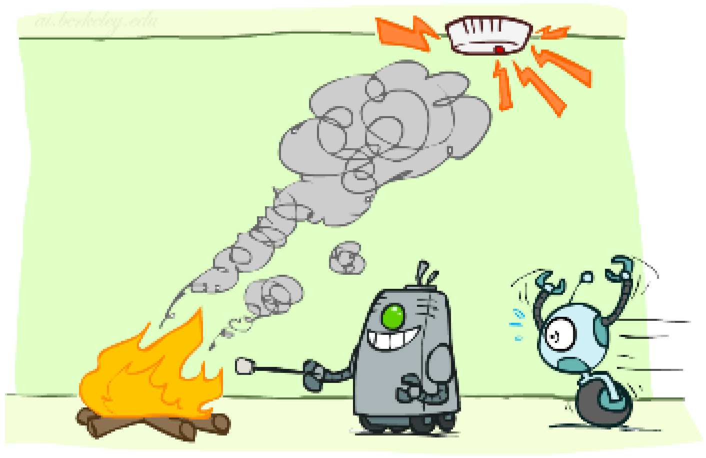
$$T \perp\!\!\!\perp U \mid R$$

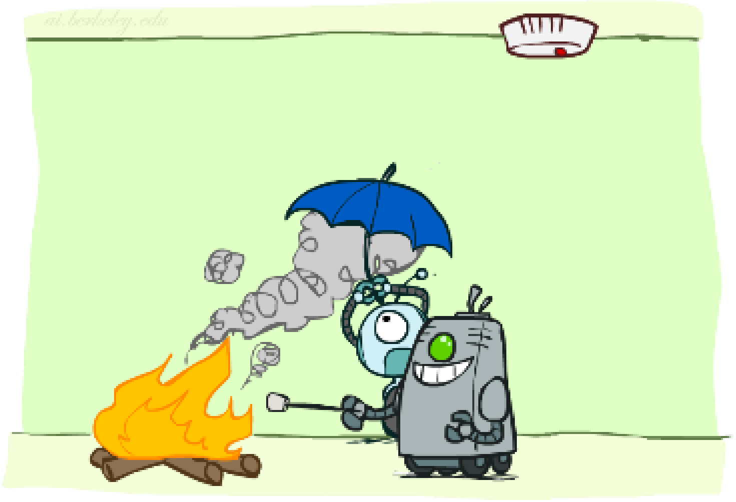
$$U \perp\!\!\!\perp T \mid R$$

Conditional Independence

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

$$F \perp\!\!\!\perp A \mid S$$





Conditional Independence and the Chain Rule

- Chain rule: $$P(X_1, X_2, \dots X_n) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)\dots$$

- Trivial decomposition:

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

- With assumption of conditional independence:

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

$$P(Y|\underbrace{X_1, X_2, \dots, X_K}_{2^K \text{ combinations}})$$

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

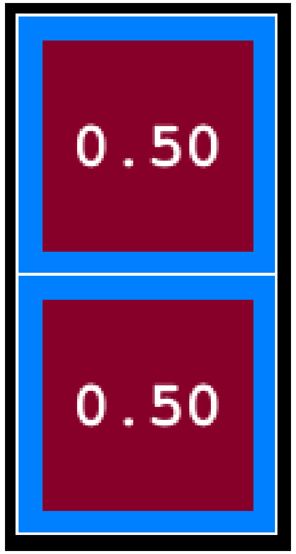
Ghostbusters Chain Rule

- Each sensor depends only on where the ghost is

- That means, the two sensors are conditionally independent, given the ghost position

- T: Top square is red
  B: Bottom square is red
  G: Ghost is in the top

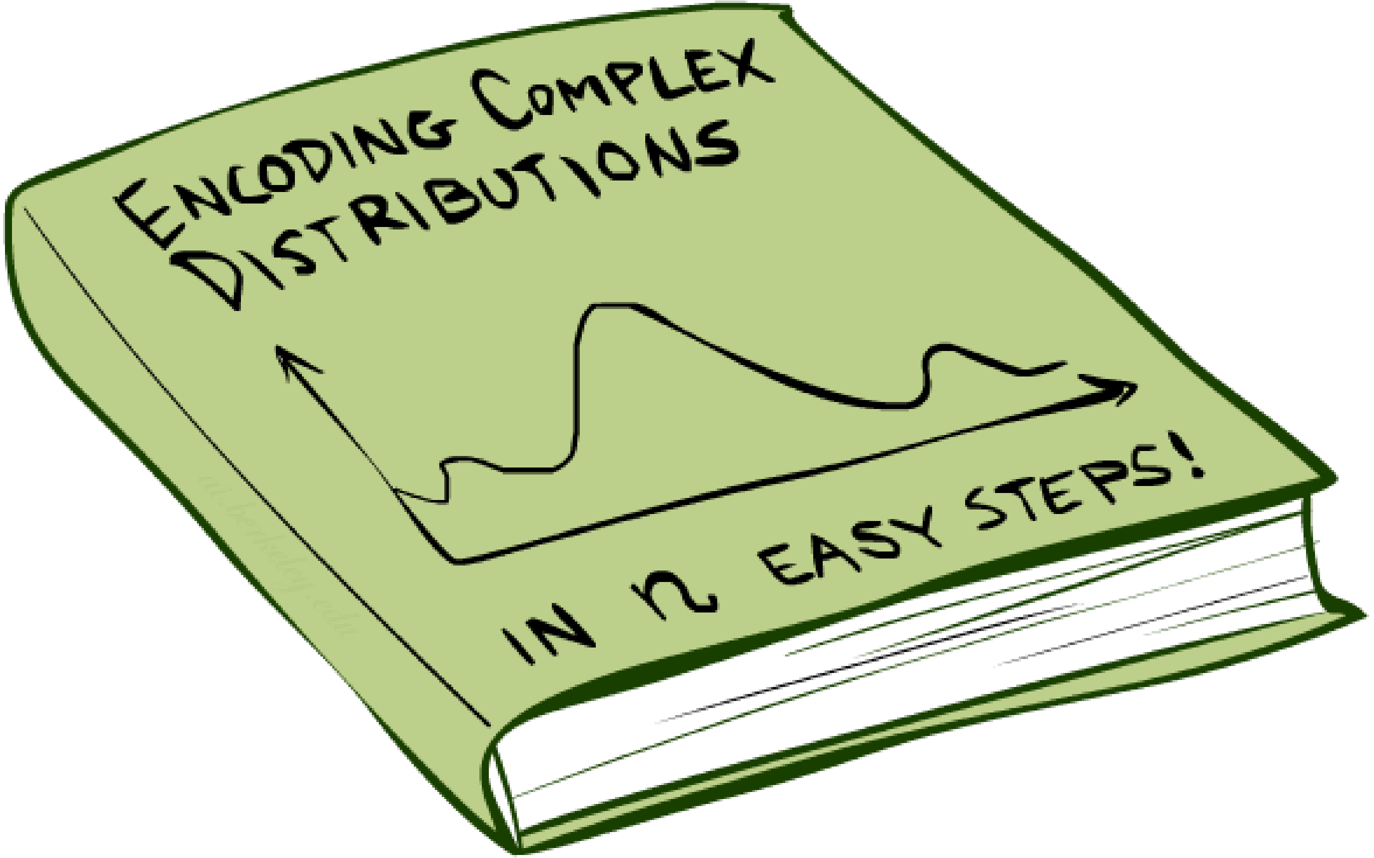
- Givens:
  P(+g) = 0.5
  P(-g) = 0.5
  P(+t | +g) = 0.8
  P(+t | -g) = 0.4
  P(+b | +g) = 0.4
  P(+b | -g) = 0.8



P(T,B,G) = P(G) P(T|G) P(B|G)

T	B	G	P(T,B,G)
+t	+b	+g	0.16
+t	+b	-g	0.16
+t	-b	+g	0.24
+t	-b	-g	0.04
-t	+b	+g	0.04
-t	+b	-g	0.24
-t	-b	+g	0.06
-t	-b	-g	0.06

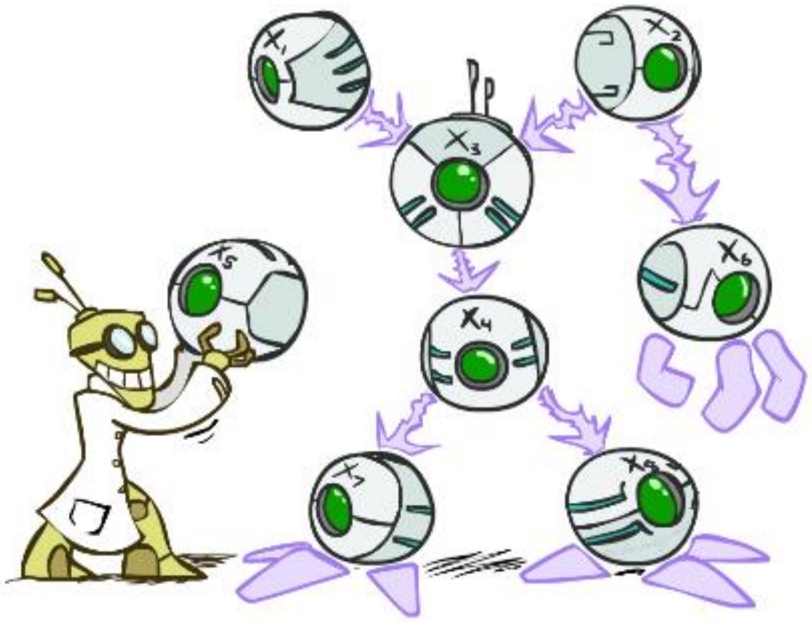
Bayes' Nets: Big Picture



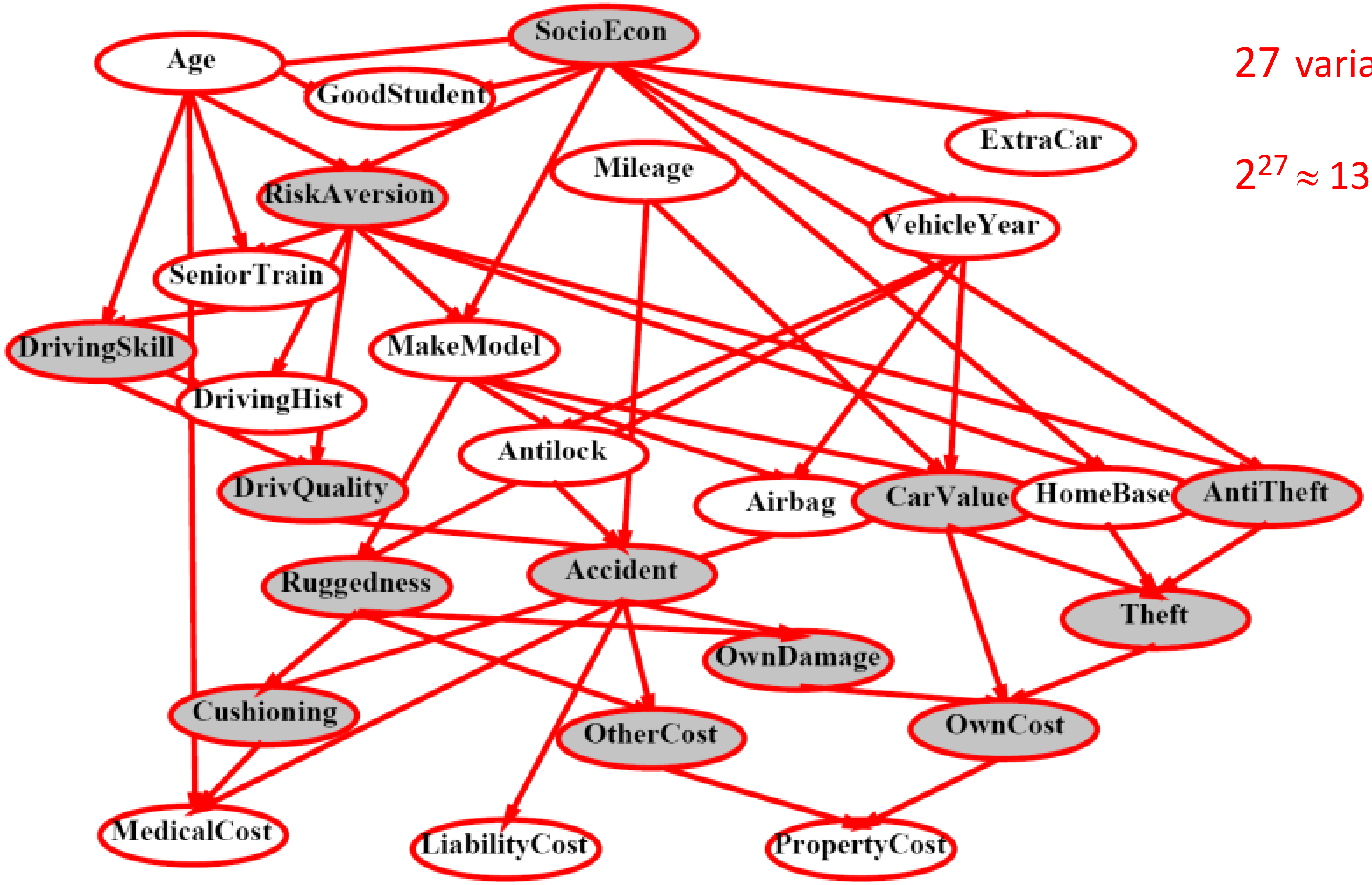
Bayes' Nets: Big Picture

- 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



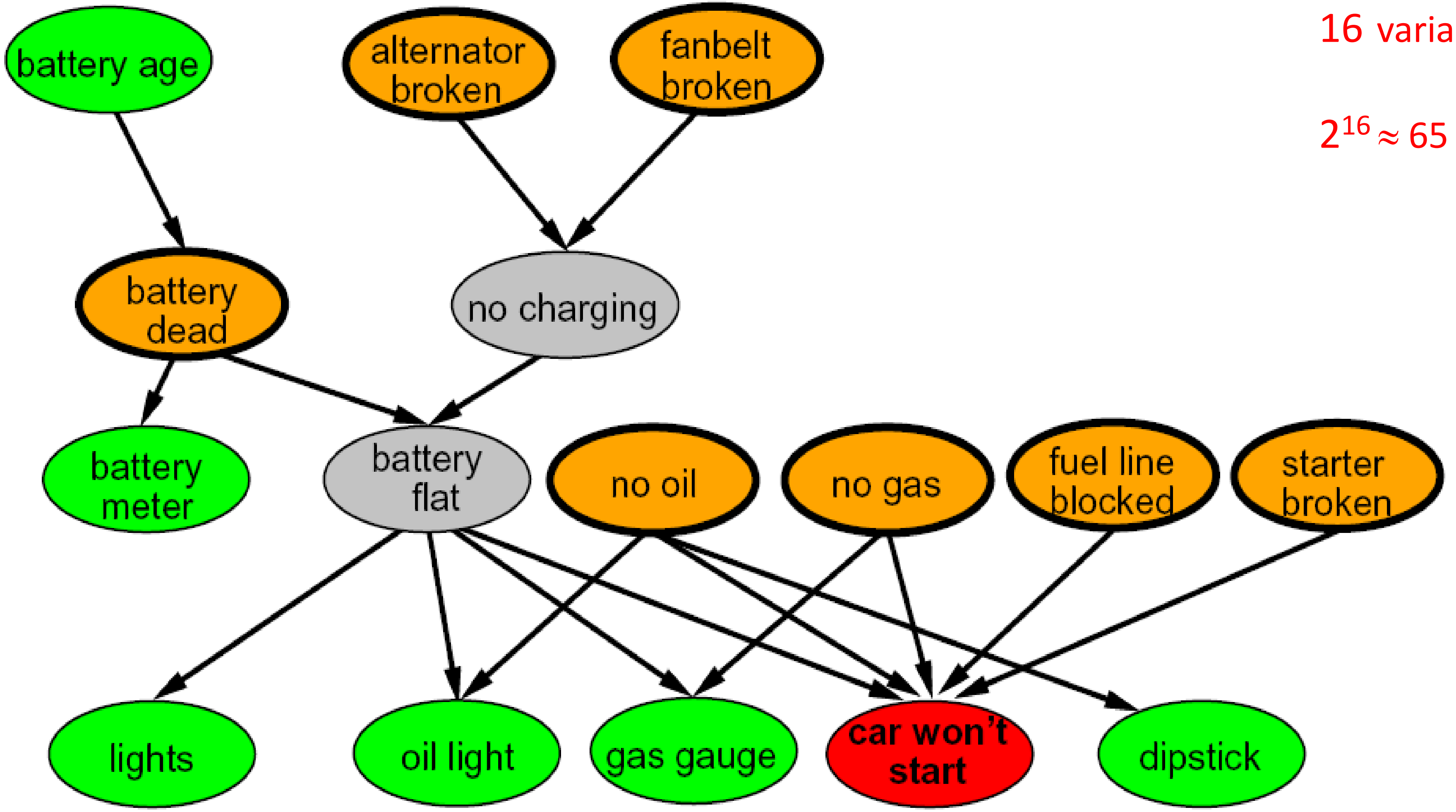
Example Bayes' Net: Insurance



27 variables

$2^{27} \approx 134$ million entries

Example Bayes' Net: Car

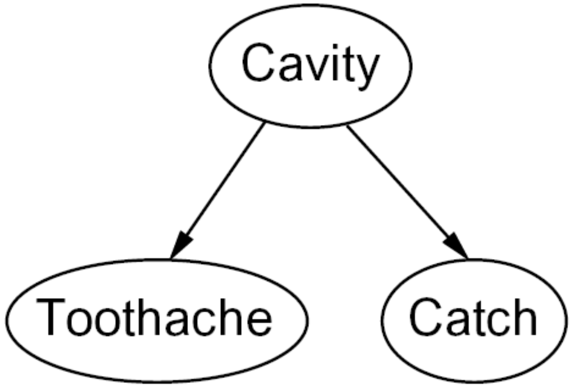


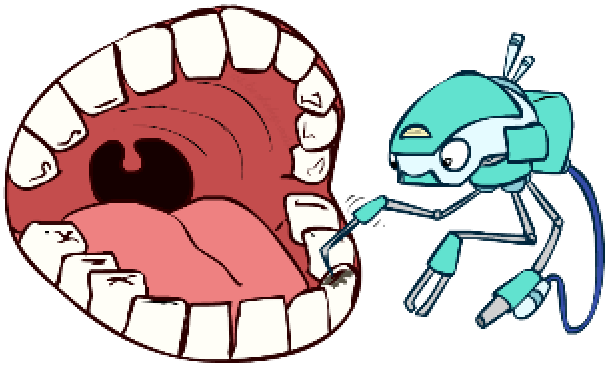
16 variables

$2^{16} \approx 65\ 500$ entries

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!)

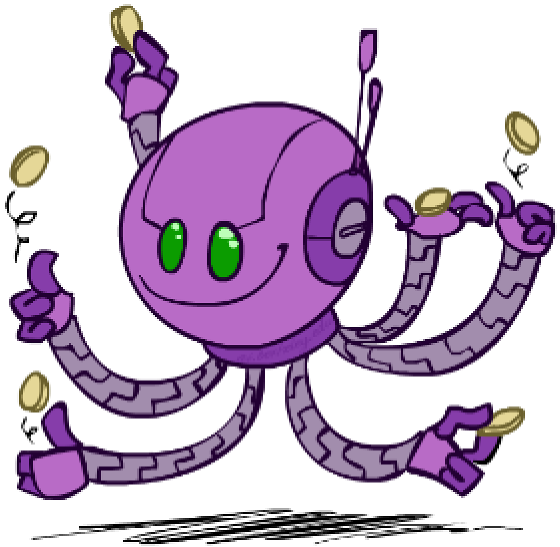




Example: Coin Flips

- N independent coin flips

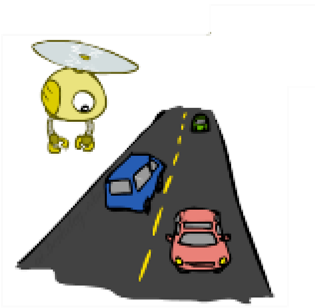
X_1 X_2 $\cdots$ X_n

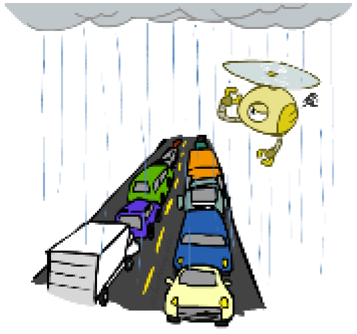


- No interactions between variables: absolute independence

Example: Traffic

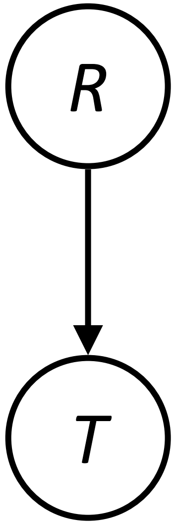
- Variables:
  - R: It rains
  - T: There is traffic





- Model 1: independence

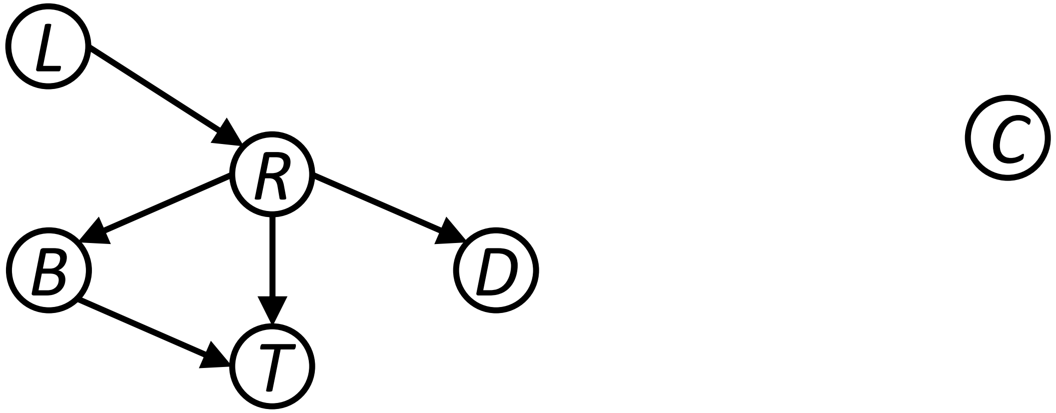
- Model 2: rain causes traffic



- Why is an agent using model 2 better?

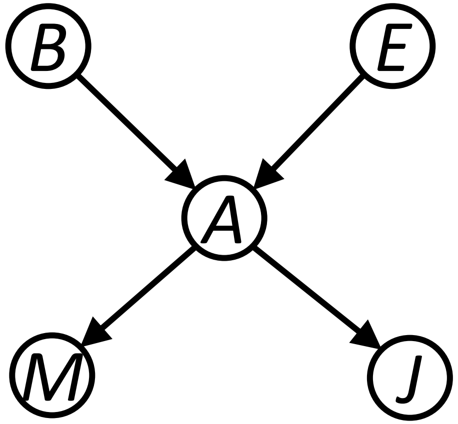
Example: Traffic II

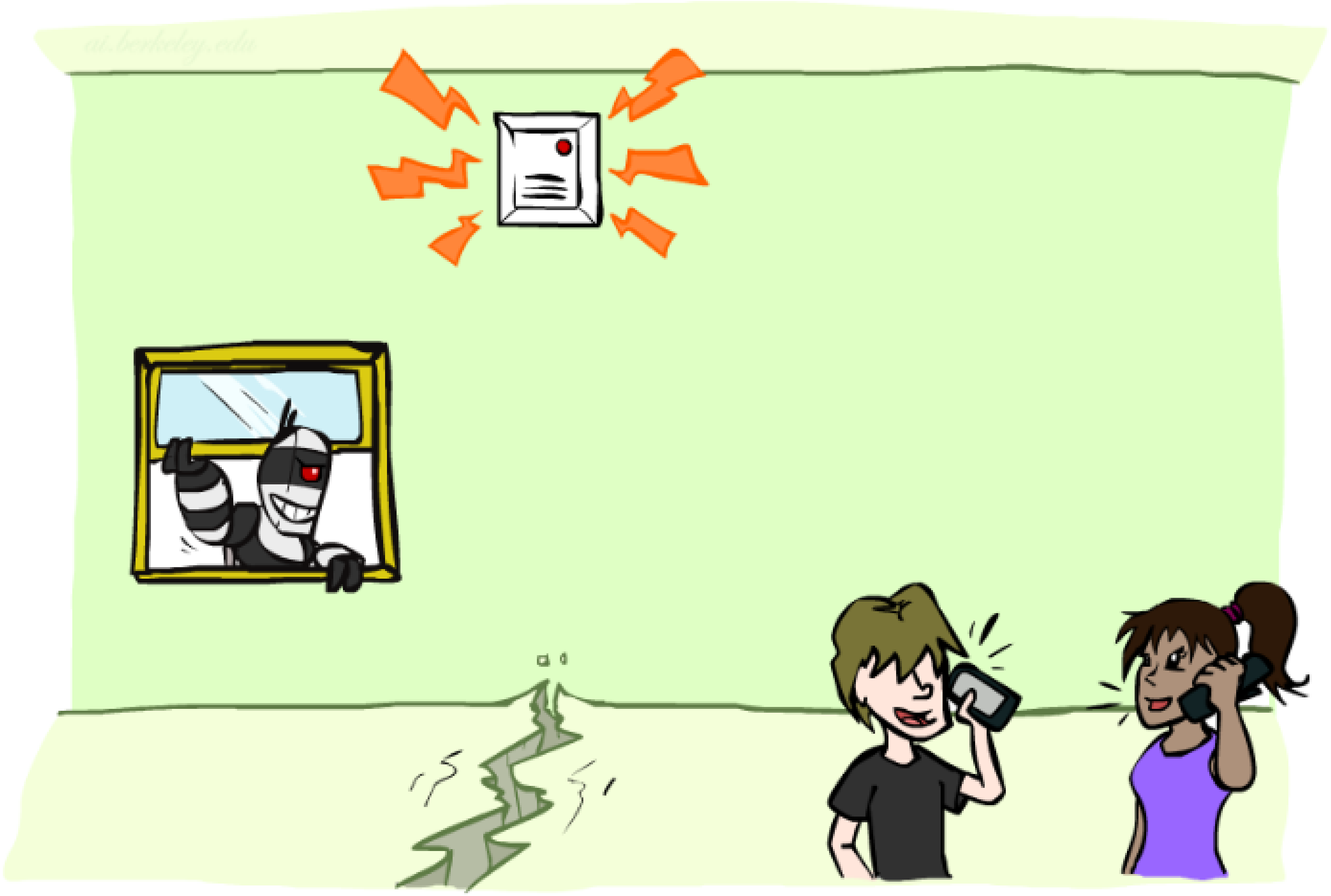
- Let's build a causal graphical model!
- Variables
  - T: Traffic
  - R: It rains
  - L: Low pressure
  - D: Roof drips
  - B: Ballgame
  - C: Cavity



Example: Alarm Network

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





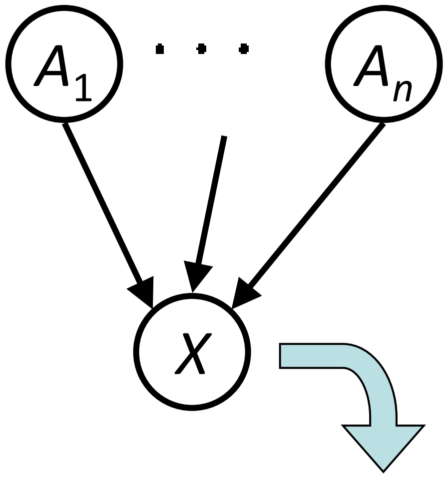
Bayes' Net Semantics

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

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

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



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

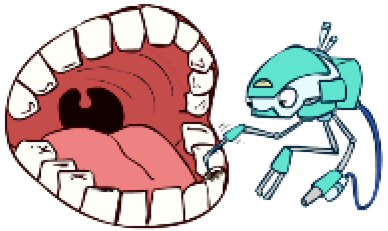
A Bayes net = Topology (graph) + Local Conditional Probabilities

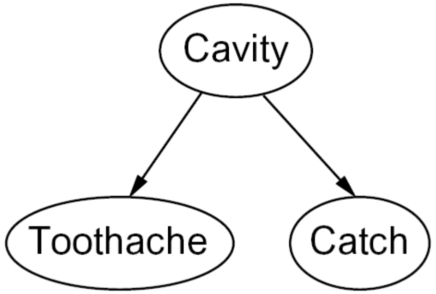
Probabilities in BNs

- 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:

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

  - Example:





$$P(+cavity, +catch, -toothache) \;= P(+cavity)\, P(+catch|+cavity)\, P(-toothache| + cavity)$$

Probabilities in BNs

- Why are we guaranteed that setting

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

results in a proper joint distribution?

- Chain rule (valid for all distributions): $P(x_1, x_2, \ldots x_n) = \prod_{i=1}^{n} P(x_i | x_1 \ldots x_{i-1})$

- <u>Assume</u> conditional independences: $P(x_i | x_1, \ldots x_{i-1}) = P(x_i | parents(X_i))$

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

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

Example: Coin Flips

X_1 X_2 … X_n

$P(X_1)$

h	0.5
t	0.5

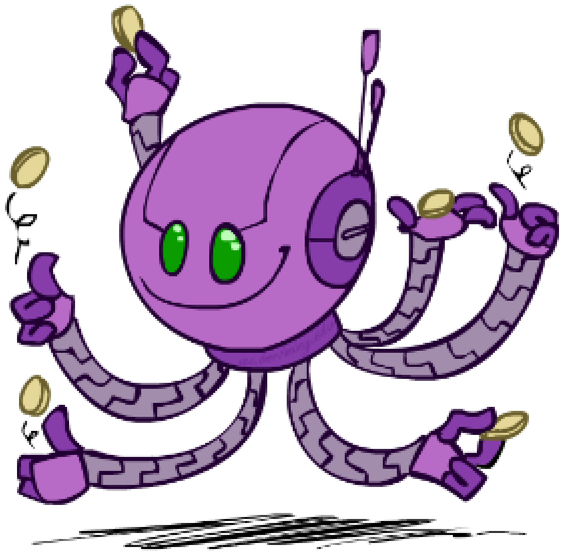
$P(X_2)$

h	0.5
t	0.5

…

$P(X_n)$

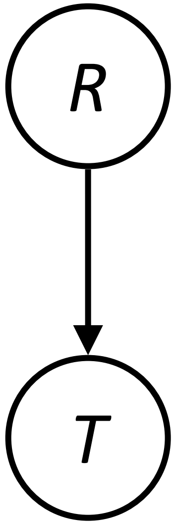
h	0.5
t	0.5



$P(h, h, t, h) =$ 0.5 * 0.5 * 0.5 * 0.5

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

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

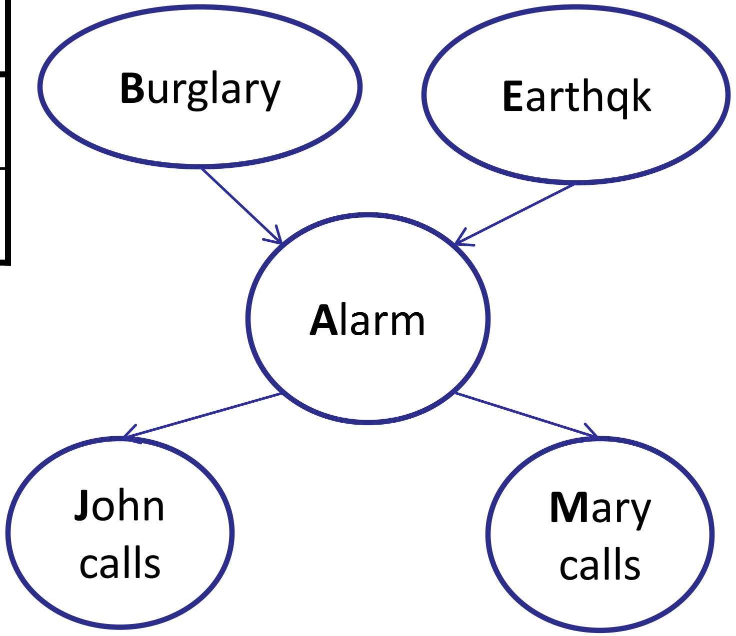
$$= \frac{1}{4} * \frac{1}{4} = \frac{1}{16}$$

Example: Alarm Network

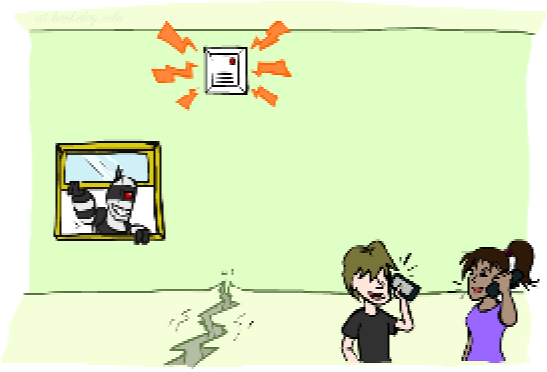
Node ordering:

B, E, A, J, M

B	P(B)
+b	$\boxed{0.001}$
-b	0.999



E	P(E)
+e	0.002
-e	$\boxed{0.998}$



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

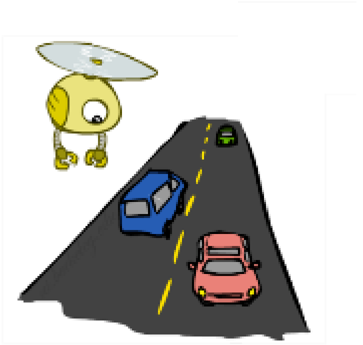
A	J	P(J|A)
+a	+j	0.9
+a	-j	$\boxed{0.1}$
-a	+j	0.05
-a	-j	0.95

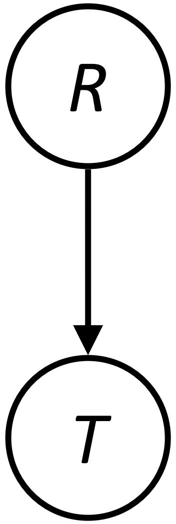
A	M	P(M|A)
+a	+m	$\boxed{0.7}$
+a	-m	0.3
-a	+m	0.01
-a	-m	0.99

B	E	A	P(A|B,E)
+b	+e	+a	0.95
+b	+e	-a	0.05
+b	-e	+a	$\boxed{0.94}$
+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: Traffic

- Causal direction





$P(R)$

+r	1/4
-r	3/4

$P(T|R)$

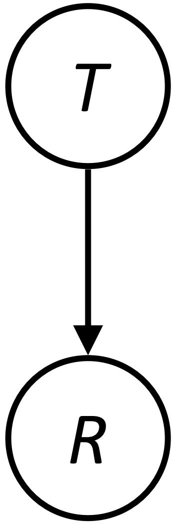
+r	+t	3/4
	-t	1/4
-r	+t	1/2
	-t	1/2

$P(T, R)$

+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

$P(R|T)$

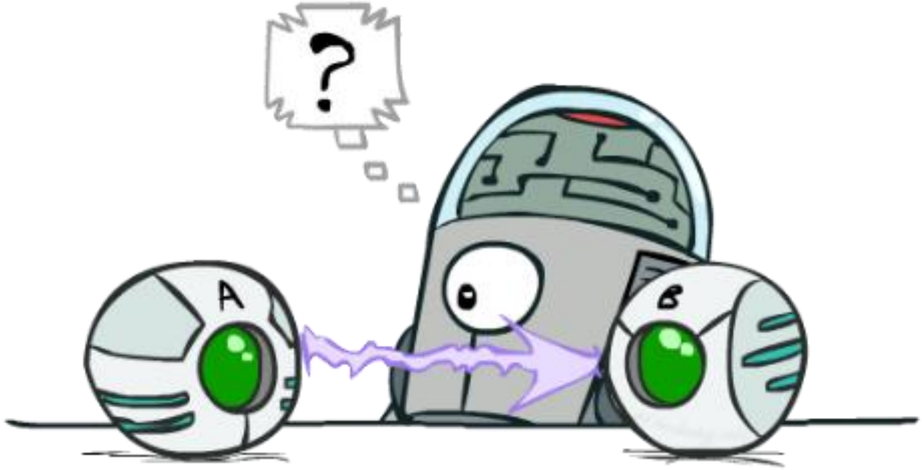
+t	+r	1/3
	-r	2/3
-t	+r	1/7
	-r	6/7

$P(T,R)$

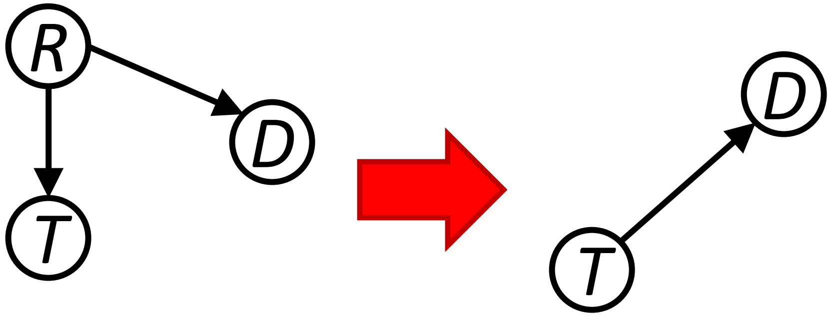
+r	+t	3/16
+r	-t	1/16
-r	+t	6/16
-r	-t	6/16

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



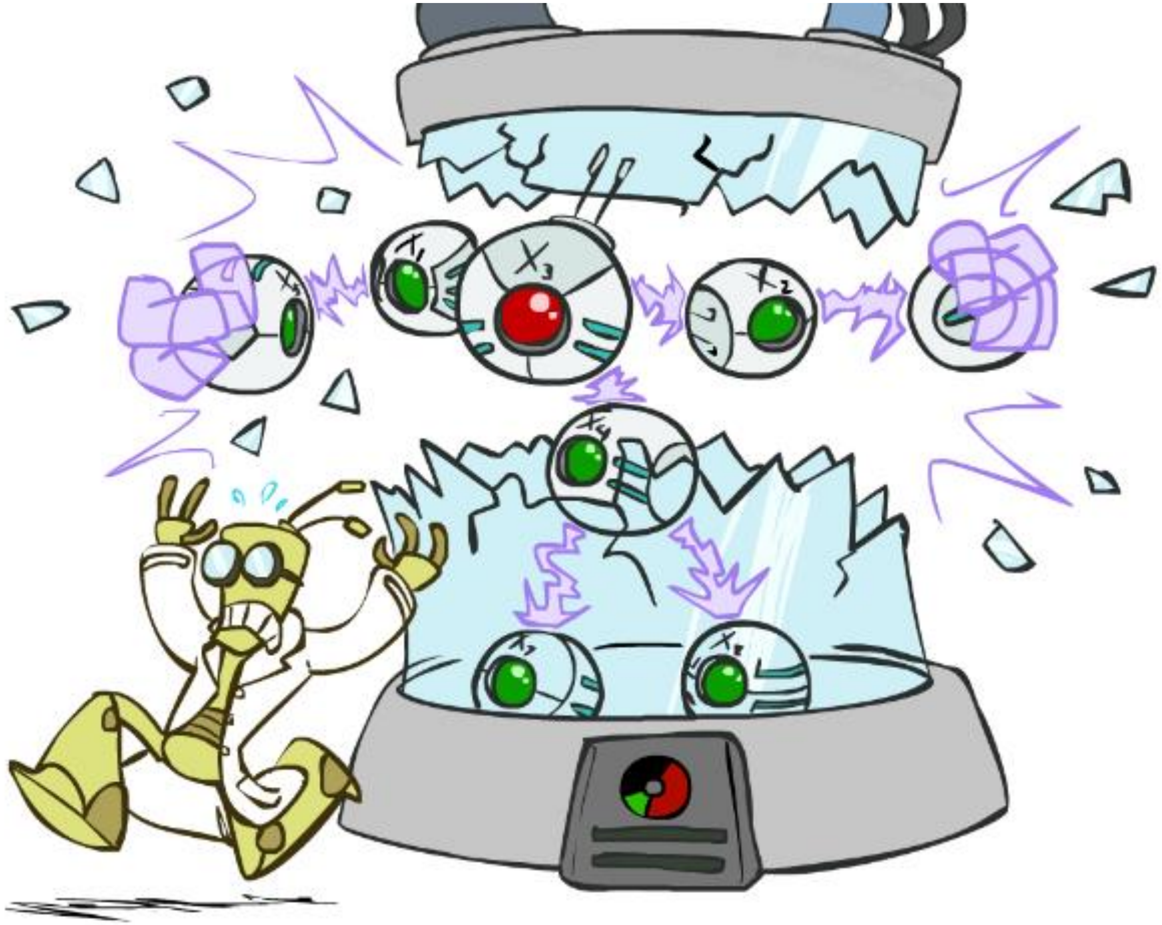
- BNs need not actually be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - E.g. consider the variables *Traffic* and *Drips*
  - End up with arrows that reflect correlation, not causation



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

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

Bayes’ Nets: Independence

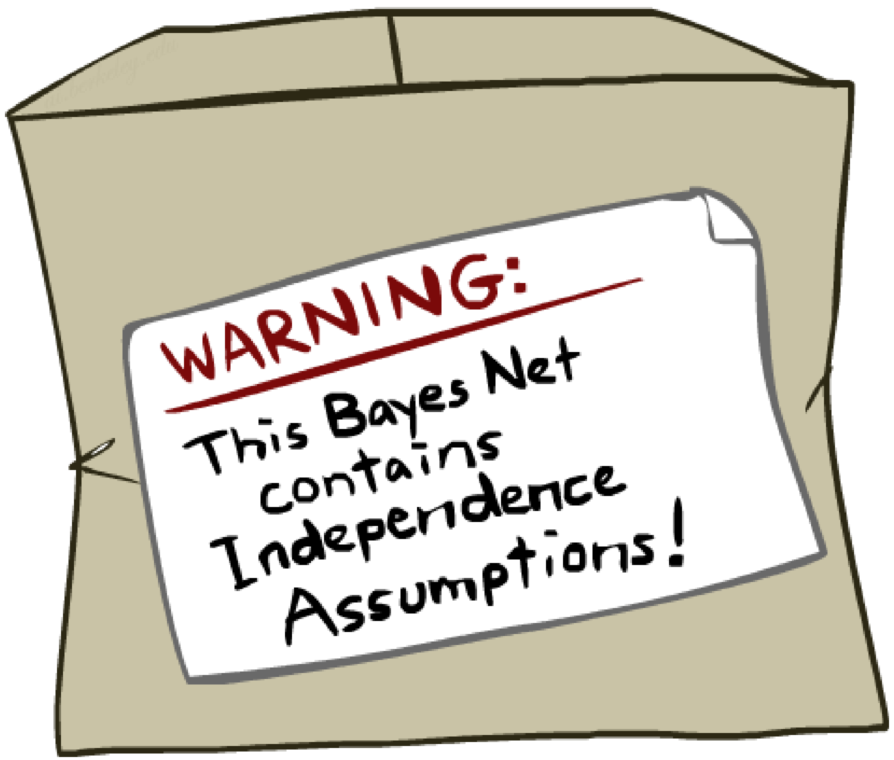


Bayes Nets: Assumptions

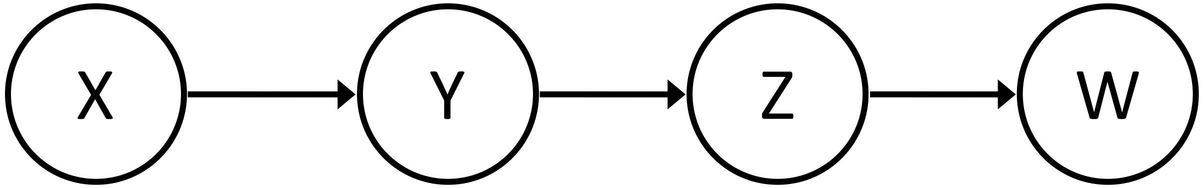
- Assumptions we are required to make to define the Bayes net when given the graph:

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

- Beyond above "chain rule → 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



Example



- Conditional independence assumptions directly from simplifications in chain rule:

$$P(X,Y,Z,W) = P(X)\ P(Y|X)\ P(Z|X,Y)\ P(W|X,Y,Z)$$ Chain rule

$$= P(X)\ P(Y|X)\ P(Z|Y)\ P(W|Z)$$ Bayes net assumptions

$\mathrm{Z} \perp\!\!\!\perp X|Y \quad \mathrm{W} \perp\!\!\!\perp X,Y|Z$

- Additional implied conditional independence assumptions?

$\mathrm{W} \perp\!\!\!\perp X|Y$

$P(w|x,y) \stackrel{?}{=} P(w|y)$

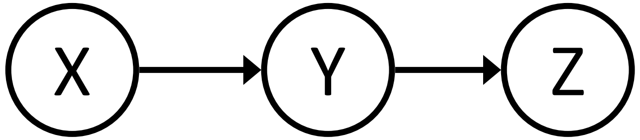
$$P(w|x,y) = \frac{P(w,x,y)}{P(x,y)} = \frac{\sum_z P(w,x,y,z)}{P(x,y)}$$

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

$$= \sum_z P(z|y)P(w|z) = \sum_z P(z|y)P(w|z,y) = \sum_z P(z,w|y) = P(w|y)$$

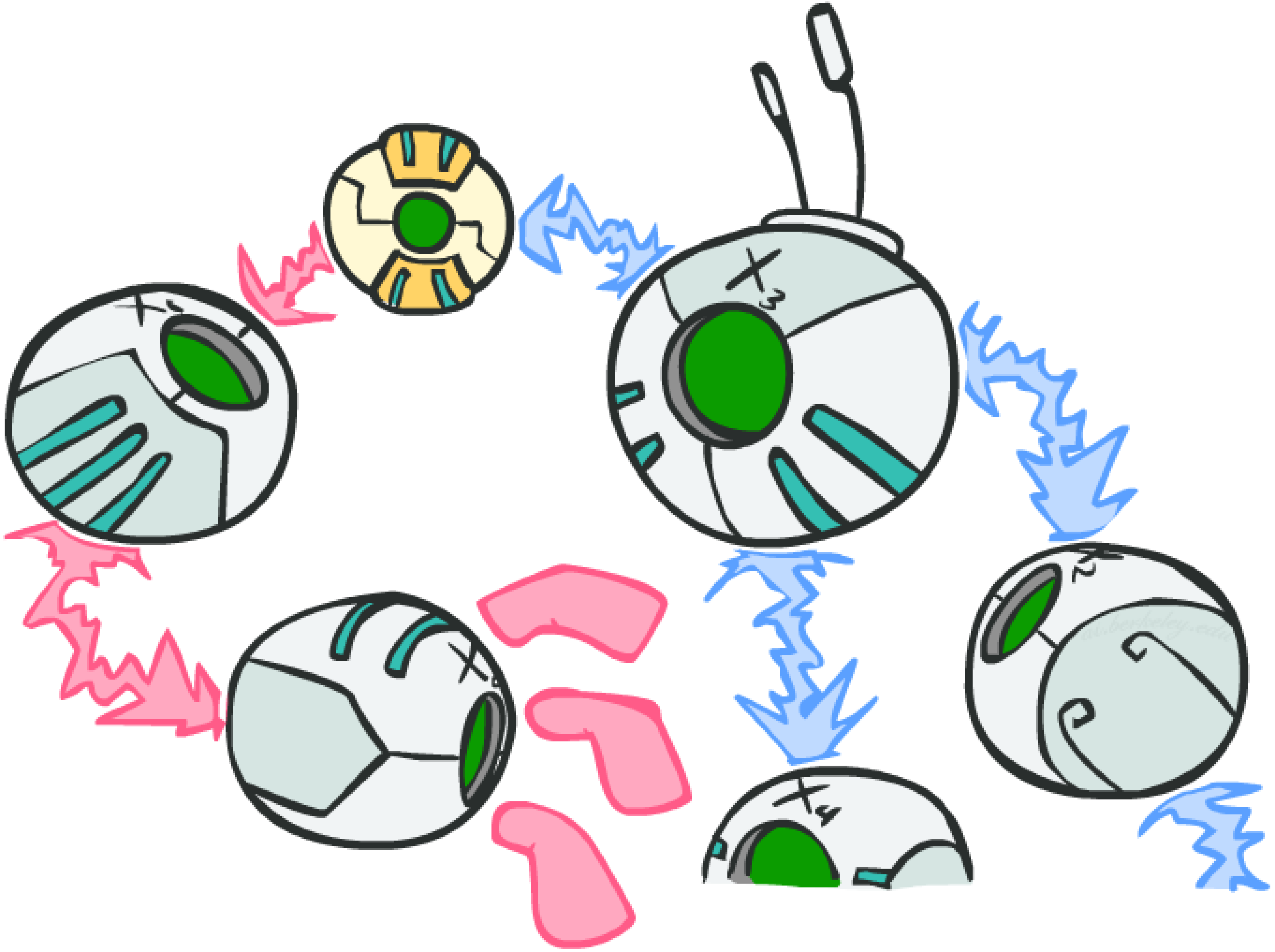
Independence in a BN

- 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:



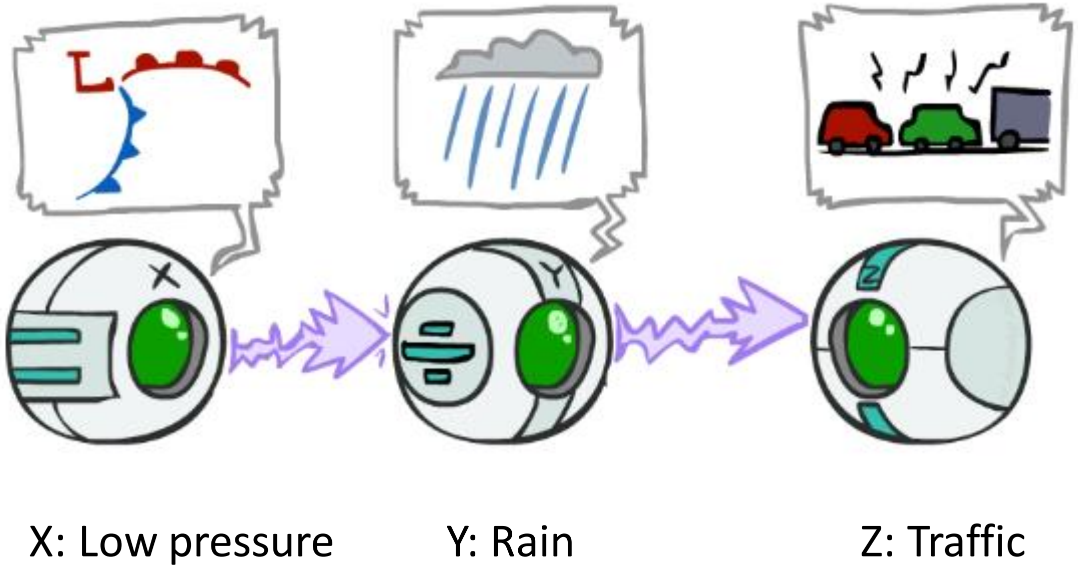
  - Question: are X and Z necessarily independent?
    - Answer: no. Example: low pressure causes rain, which causes traffic.
    - X can influence Z, Z can influence X (via Y)

Independence properties of triples



Causal Chains

- This configuration is a "causal chain"



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

- 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?

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

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

$$= P(z|y)$$

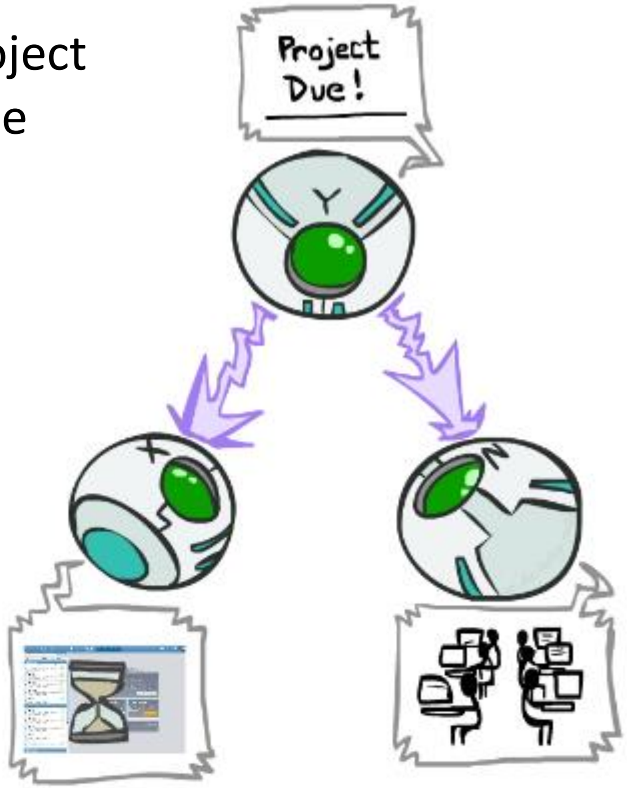
Yes!

- Evidence along the chain "blocks" the influence

Common Cause

- This configuration is a "common cause"

Y: Project due



X: Forums busy

Z: Lab full

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

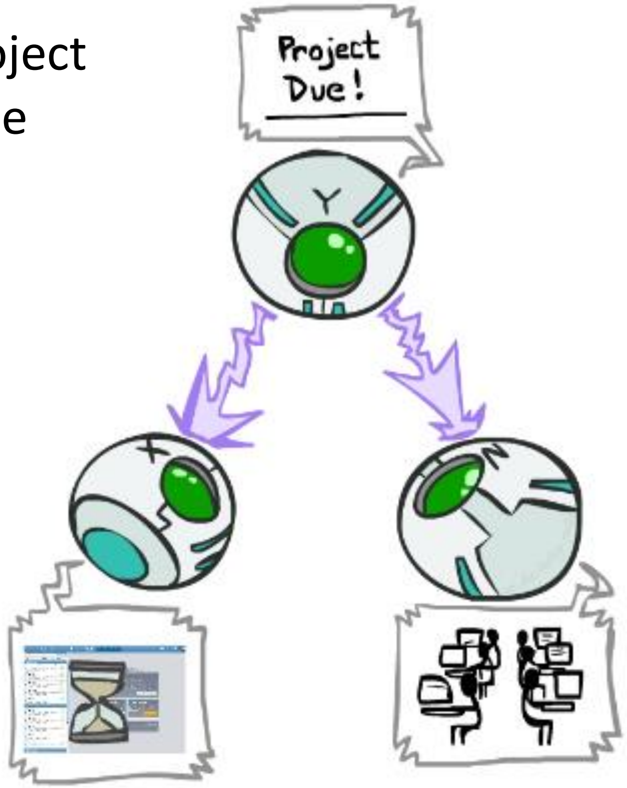
- Guaranteed X independent of Z ? *No!*
  - One example set of CPTs for which X is not independent of Z is sufficient to show this independence is not guaranteed.
  - Example:
    - Project due causes both forums busy and lab full
    - In numbers:

P(+x | +y) = 1, P(-x | -y) = 1,
P(+z | +y) = 1, P(-z | -y) = 1

Common Cause

- This configuration is a "common cause"

Y: Project due



X: Forums busy

Z: Lab full

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

- Guaranteed X and Z independent given Y?

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

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

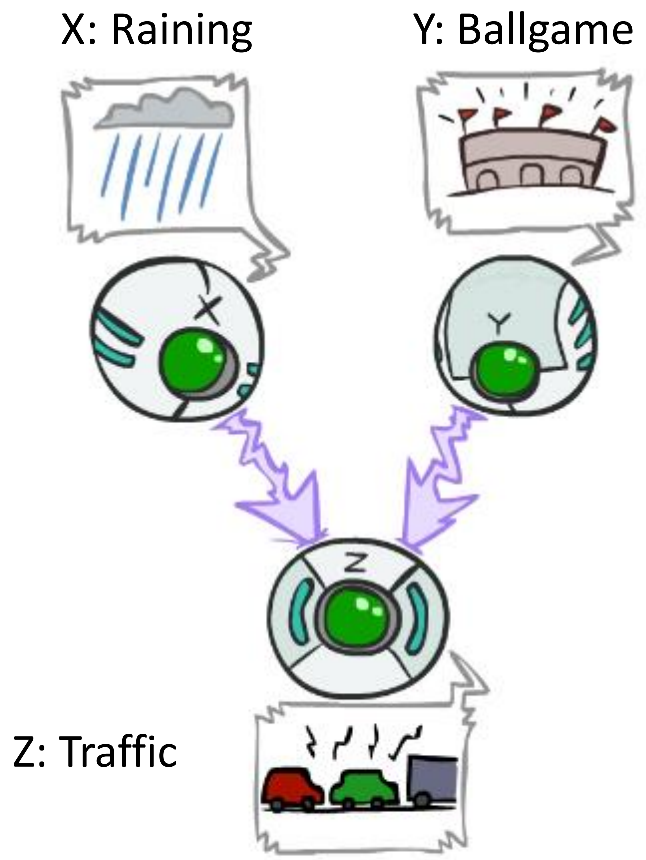
$$= P(z|y)$$

Yes!

- Observing the cause blocks influence between effects.

Common Effect

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



- Are X and Y independent?
  - *Yes*: the ballgame and the rain cause traffic, but they are not correlated

$$P(X,Y,Z) = P(X)P(Y)P(Z|X,Y)$$

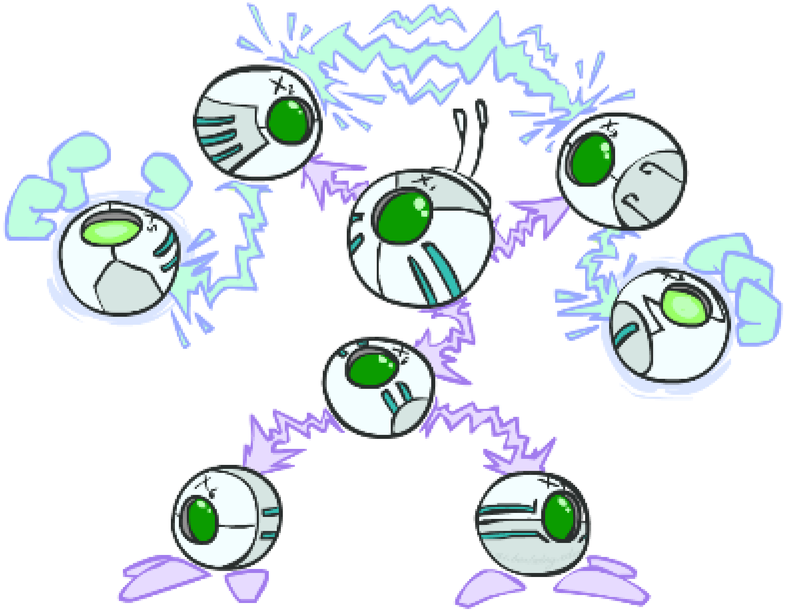
$$P(x,y) = \sum_z P(x,y,z) = \sum_z P(x)P(y)P(z|x,y) =$$

$$= P(x)P(y)\underbrace{\sum_z P(z|x,y)}_{1} = P(x)P(y)$$

- 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.

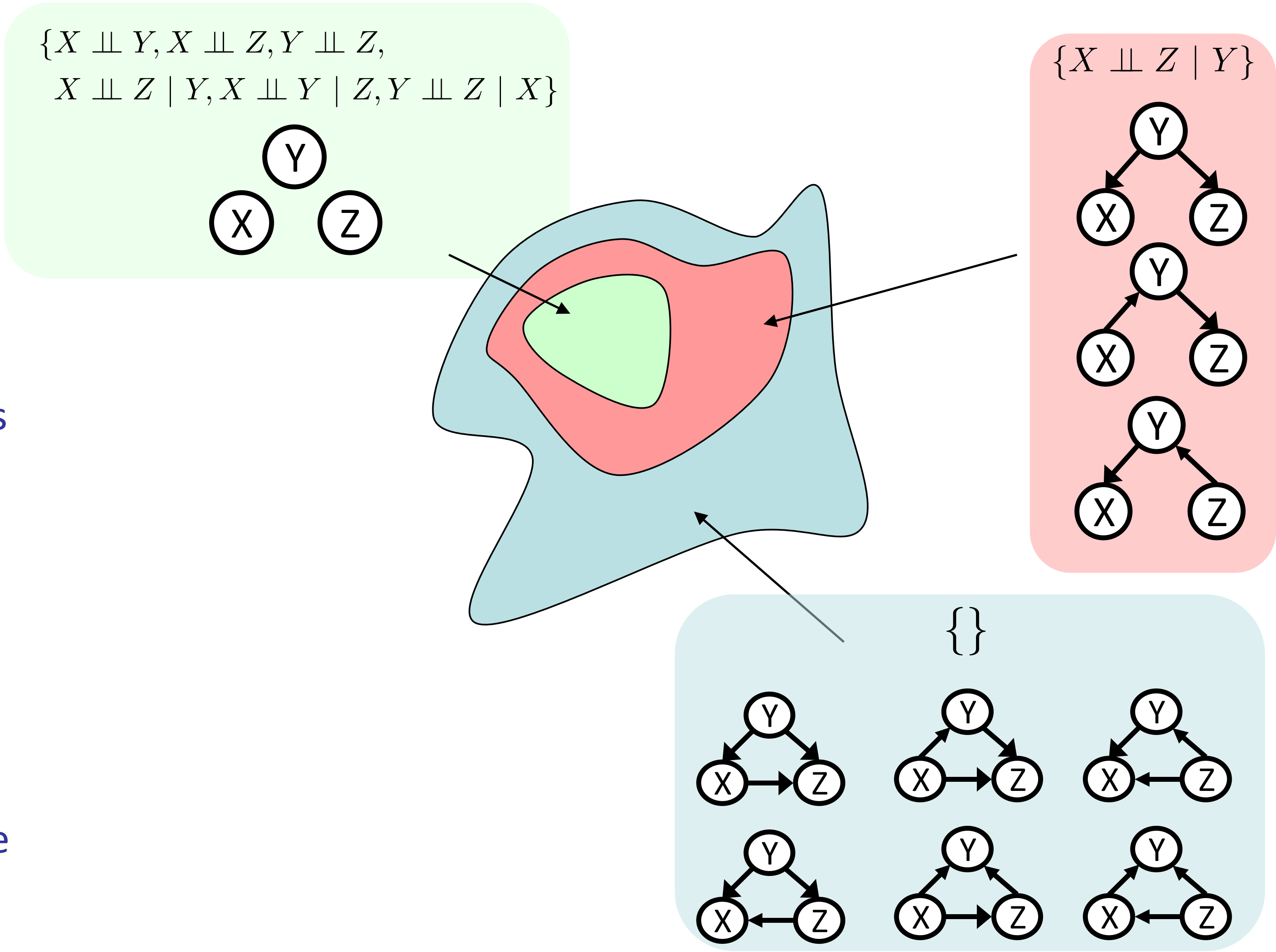
The General Case

- 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



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)
- Adding arcs increases the set of distributions, but has several costs
- Full conditioning can encode any distribution



Summary

- A Bayes' net is an efficient encoding of a probabilistic model of a domain
  - A directed, acyclic graph, one node per random variable
  - A conditional probability table (CPT) for each node
  - Bayes' nets implicitly encode joint distributions
- Questions we can ask:
  - 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?
- 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