# Announcements

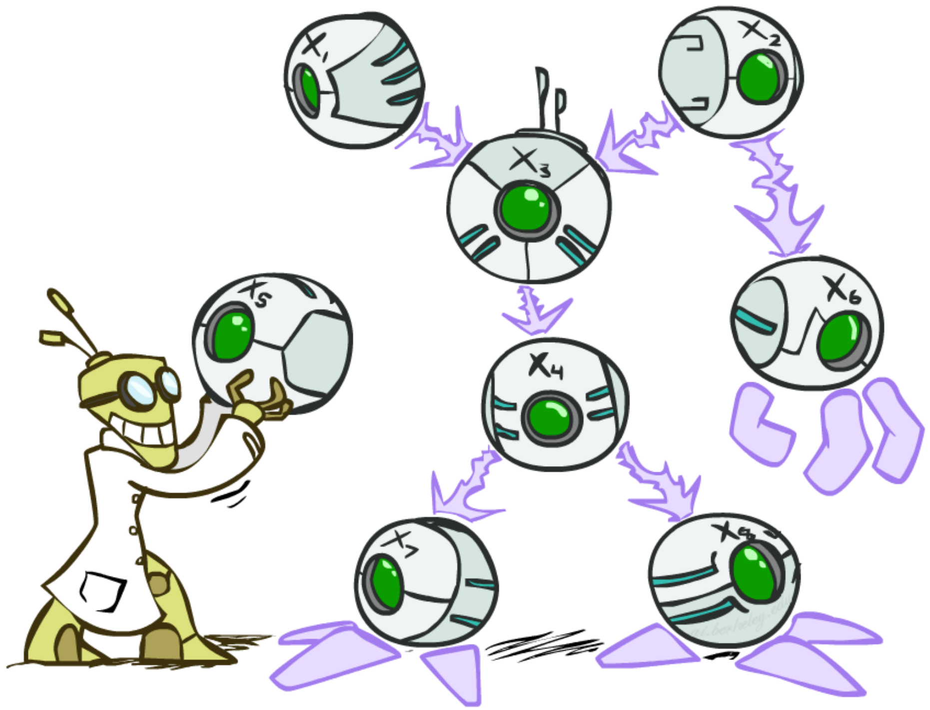
- HW3 is due **today, February 20**, 11:59pm PT
- Project 3 is due **Tuesday, February 27**, 11:59pm PT
- HW4 out later this week; due **Friday, March 1, 11:59pm PT**
- Midterm: **Tuesday, March 5, 7pm PT** (more info on website)

Pre-scan attendance QR code now!

(Password appears later)

# CS 188: Artificial Intelligence

## Bayes Nets



[These slides were created by Dan Klein and Pieter Abbeel for CS188 Intro to AI at UC Berkeley.  All CS188 materials are available at http://ai.berkeley.edu.]

# Recap: Probability

- Basic laws: $0 \leq P(\omega) \leq 1 \qquad \sum_{\omega \in \Omega} P(\omega) = 1$
- Events: subsets of $\Omega$: $P(A) = \sum_{\omega \in A} P(\omega)$
- Random variable $X(\omega)$ has a value in each $\omega$
  - Distribution $P(X)$ gives probability for each possible value $x$
  - Joint distribution $P(X,Y)$ gives total probability for each combination $x,y$
- Summing out/marginalization: $P(X=x) = \sum_y P(X=x, Y=y)$
- Conditional probability: $P(X|Y) = \mathrm{P}(X,Y)/\mathrm{P}(Y)$
- Product rule: $P(X|Y)P(Y) = \mathrm{P}(X,Y) = P(Y|X)P(X)$
  - Generalize to chain rule: $P(X_1,..,X_n) = \prod_i P(X_i \mid X_1,..,X_{i-1})$

# Recap: Strict Independence

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

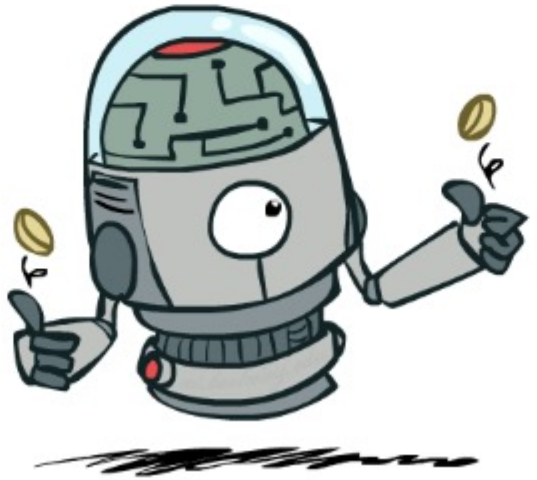
$$\forall x,y \qquad P(x, y) = P(x)\, P(y)$$

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

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

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

- Example: two dice rolls $Roll_1$ and $Roll_2$
  - $P(Roll_1{=}5, Roll_2{=}3) \;=\; P(Roll_1{=}5)\, P(Roll_2{=}3) \;=\; 1/6 \times 1/6 \;=\; 1/36$
  - $P(Roll_2{=}3 \mid Roll_1{=}5) \;=\; P(Roll_2{=}3)$



# Recap: Strict Independence

- *n* fair, independent coin flips:

$P(X_1)$

| | |
|---|---|
| H | 0.5 |
| T | 0.5 |

$P(X_2)$

| | |
|---|---|
| H | 0.5 |
| T | 0.5 |

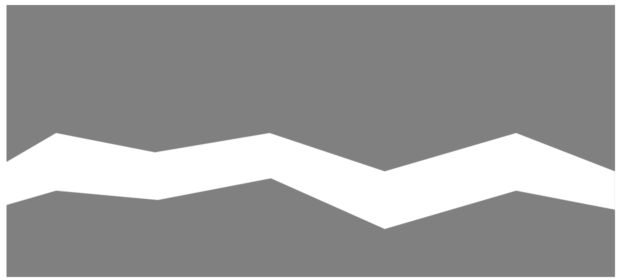
. . .

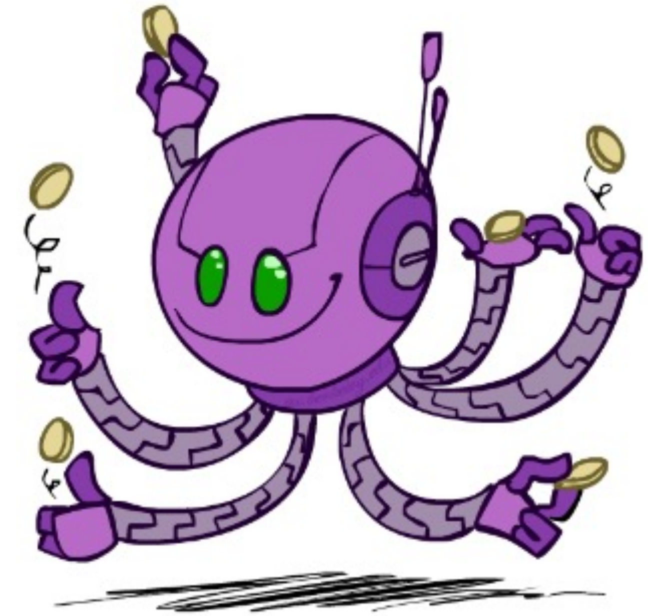
$P(X_n)$

| | |
|---|---|
| H | 0.5 |
| T | 0.5 |

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

$2^n$





# Example: Strict Independence

$P$(Rain, Traffic, Umbrella)

| Rain | Traffic | Umbrella | P | P_indep. |
|---|---|---|---|---|
| F | F | F | 0.504 | 0.314 |
| F | F | T | 0.056 | 0.141 |
| F | T | F | 0.126 | 0.169 |
| F | T | T | 0.014 | 0.076 |
| T | F | F | 0.018 | 0.135 |
| T | F | T | 0.072 | 0.060 |
| T | T | F | 0.042 | 0.072 |
| T | T | T | 0.168 | 0.033 |

✗

$P$(Rain)

| F | T |
|---|---|
| 0.7 | 0.3 |

*

$P$(Traffic)

| F | T |
|---|---|
| 0.65 | 0.35 |

*

$P$(Umbrella)

| F | T |
|---|---|
| 0.69 | 0.31 |

=

# Example: Chain Rule

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

| Rain | Traffic | Umbrella | P |
|---|---|---|---|
| F | F | F | 0.504 |
| F | F | T | 0.056 |
| F | T | F | 0.126 |
| F | T | T | 0.014 |
| T | F | F | 0.018 |
| T | F | T | 0.072 |
| T | T | F | 0.042 |
| T | T | T | 0.168 |

✓

$P(\text{Rain})$

| F | T |
|---|---|
| 0.7 | 0.3 |

*

$P(\text{Traf.}\,|\,\text{Rain})$

| | Traffic | |
|---|---|---|
| Rain | F | T |
| F | 0.8 | 0.2 |
| T | 0.3 | 0.7 |

*

$P(\text{Umbr.}\,|\,\text{Rain, Traf.})$

| | | Umbrella | |
|---|---|---|---|
| Rain | Traffic | F | T |
| F | F | 0.9 | 0.1 |
| F | T | 0.9 | 0.1 |
| T | F | 0.2 | 0.8 |
| T | T | 0.2 | 0.8 |

**conditional independence**

=

# Example: Chain Rule

*P*(Rain, Traffic, Umbrella)

| Rain | Traffic | Umbrella | P |
|---|---|---|---|
| F | F | F | 0.504 |
| F | F | T | 0.056 |
| F | T | F | 0.126 |
| F | T | T | 0.014 |
| T | F | F | 0.018 |
| T | F | T | 0.072 |
| T | T | F | 0.042 |
| T | T | T | 0.168 |

✓

*P*(Rain)

| F | T |
|---|---|
| 0.7 | 0.3 |

*

*P*(Traf.|Rain)

| | Traffic | |
|---|---|---|
| Rain | F | T |
| F | 0.8 | 0.2 |
| T | 0.3 | 0.7 |

*

*P*(Umbr.|Rain)

| | Umbrella | |
|---|---|---|
| Rain | F | T |
| F | 0.9 | 0.1 |
| T | 0.2 | 0.8 |

**conditional independence**

=

# Conditional Independence

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

- $X$ is conditionally independent of $Y$ given $Z$ if and only if:

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

  or, equivalently, if and only if

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

# Conditional Independence

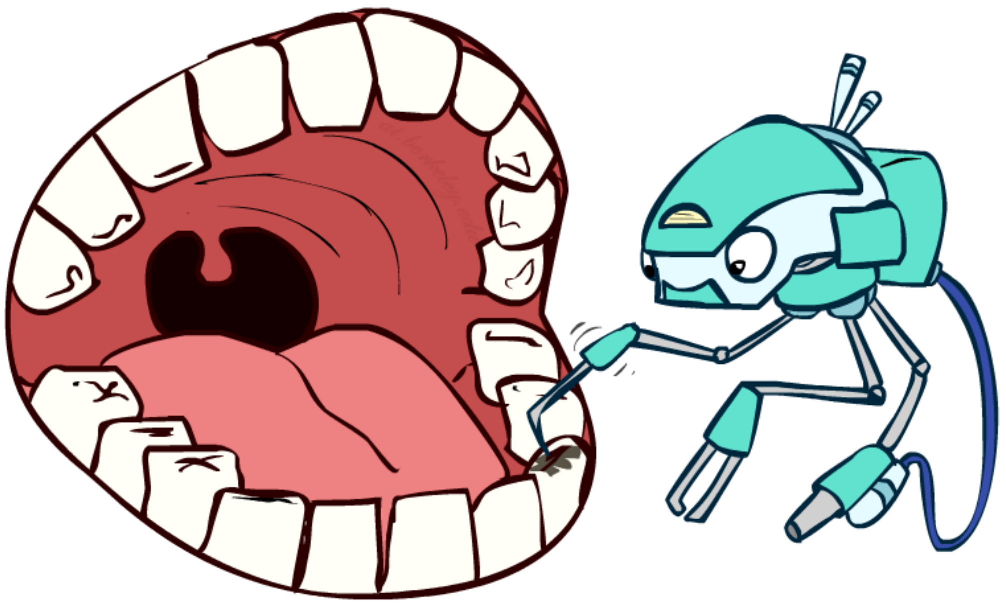
- P(Toothache, Cavity, Catch)

- 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)

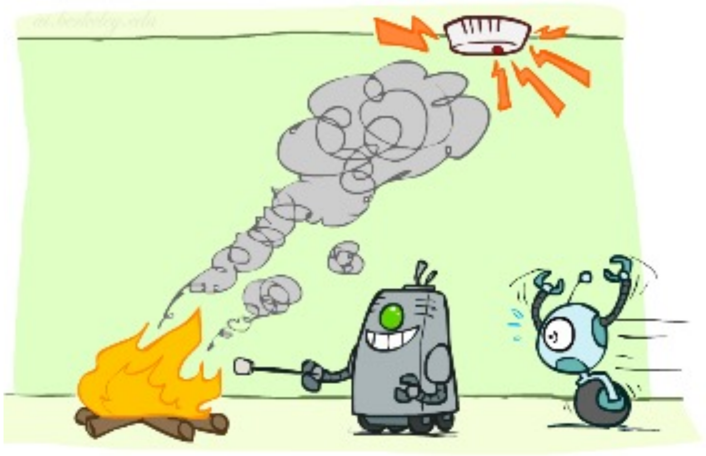
- Catch is *conditionally independent* of Toothache given Cavity:
  - P(Catch | Toothache, Cavity) = P(Catch | Cavity)

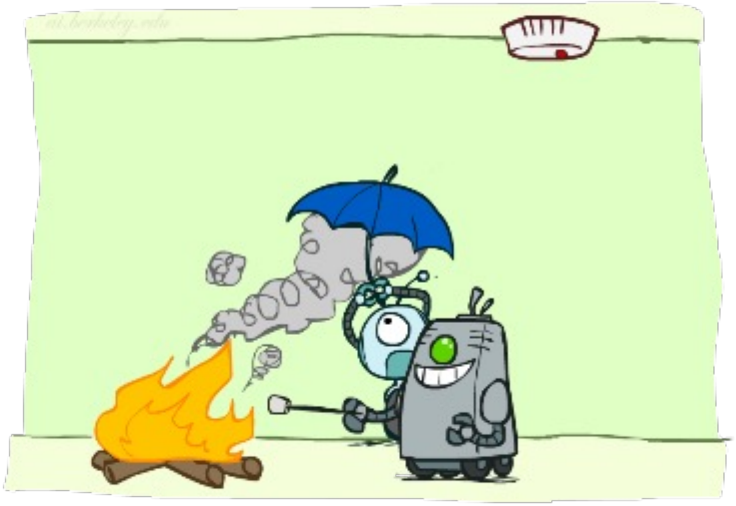
- Equivalent statements:
  - P(Toothache | Catch , Cavity) = P(Toothache | Cavity)
  - P(Toothache, Catch | Cavity) = P(Toothache | Cavity) P(Catch | Cavity)
  - One can be derived from the other easily



# Conditional Independence

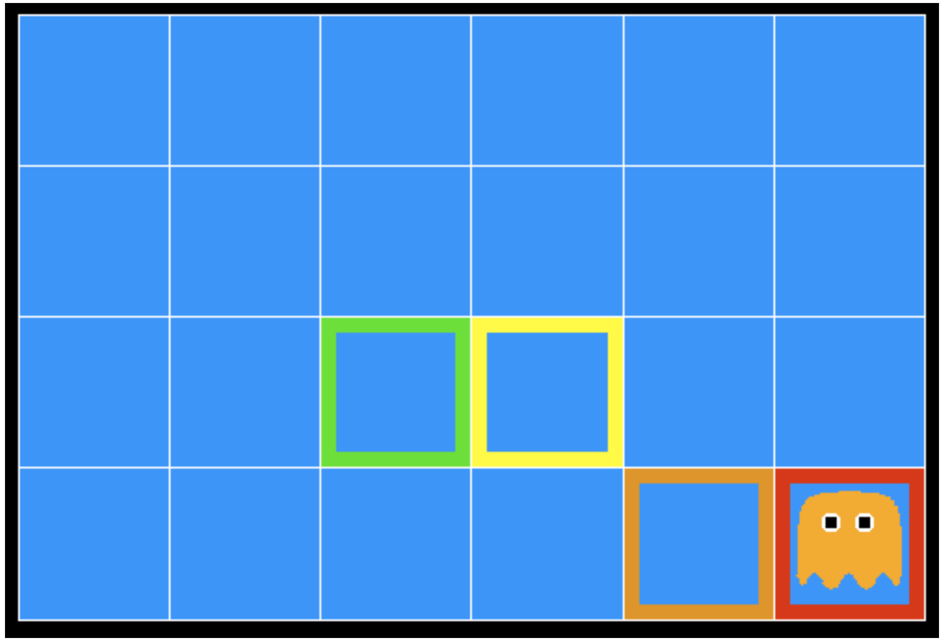
- What about this domain:
  - Fire
  - Smoke
  - Alarm





# Ghostbusters

- A ghost is in the grid somewhere
- Sensor readings tell how close a square is to the ghost
  - On the ghost: usually red
  - 1 or 2 away: usually orange
  - 3 or 4 away: usually yellow
  - 5+ away: usually green
- Click on squares until confident of location, then "***bust***"



# Video of Demo Ghostbusters with Probability

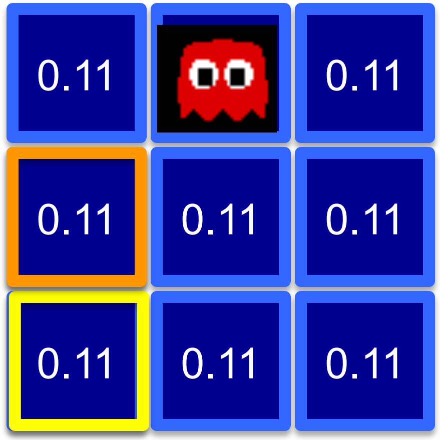
# Ghostbusters model

- Variables and ranges:
  - $G$ (ghost location) in $\{(1,1),\ldots,(3,3)\}$
  - $C_{x,y}$ (color measured at square x,y) in {red,orange,yellow,green}

| | | |
|---|---|---|
| 0.11 | 0.11 | 0.11 |
| 0.11 | 0.11 | 0.11 |
| 0.11 | 0.11 | 0.11 |

- Ghostbuster physics:
  - ***Uniform prior distribution*** over ghost location: $P(G)$
  - ***Sensor model***: $P(C_{x,y} \mid G)$ (depends only on distance to G)
    - E.g. $P(C_{1,1} = \text{yellow} \mid G = (1,1)) = 0.1$

# Ghostbusters model, contd.

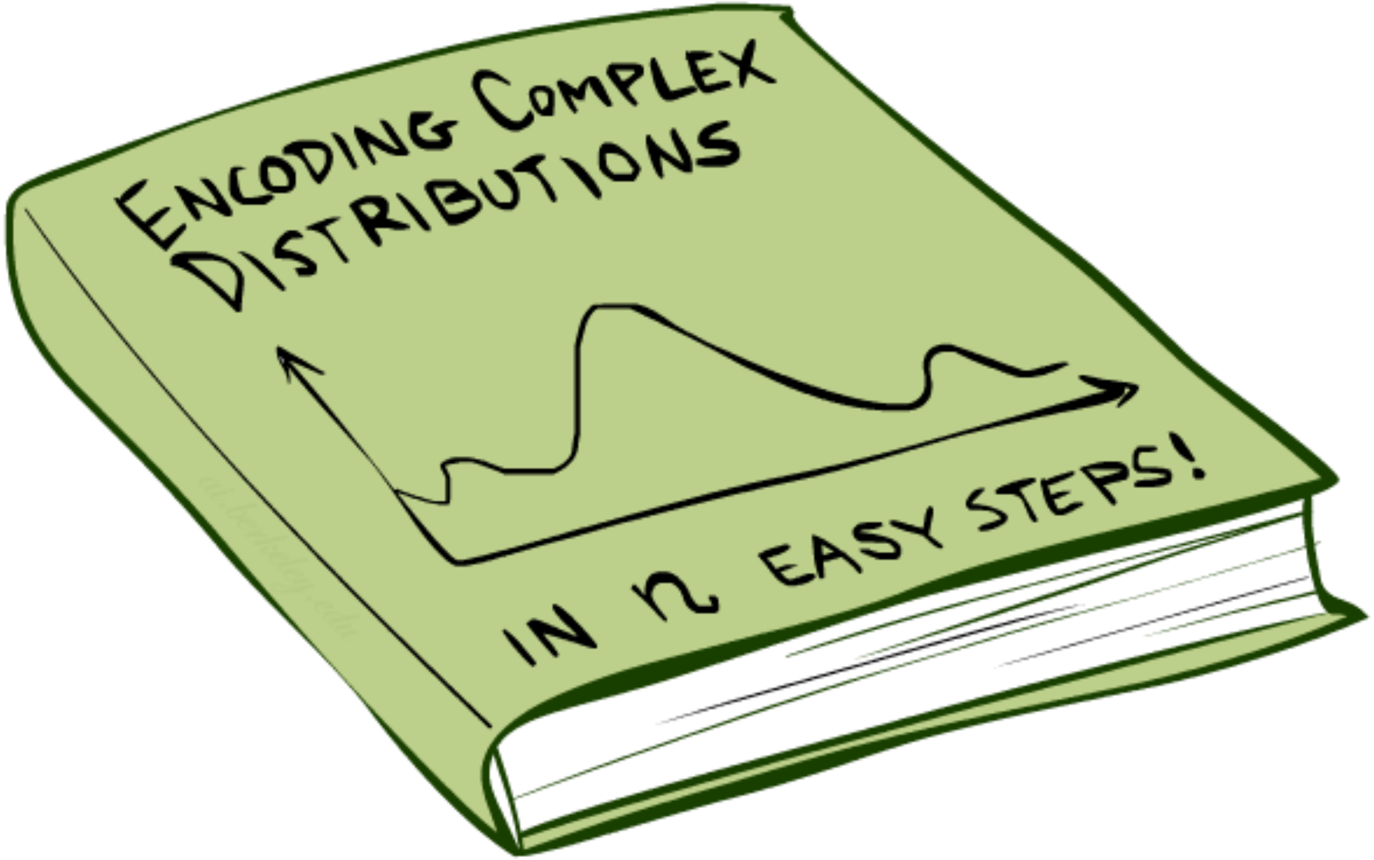



- $P(G, C_{1,1}, \ldots C_{3,3})$ has $9 \times 4^9 = 2{,}359{,}296$ entries!!!
- Ghostbuster independence:
  - Are $C_{1,1}$ and $C_{1,2}$ independent?
    - E.g., does $P(C_{1,1} = \text{yellow}) = P(C_{1,1} = \text{yellow} \mid C_{1,2} = \text{orange})$ ?
- Ghostbuster physics again:
  - $P(C_{x,y} \mid G)$ ***depends <u>only</u> on distance to G***
    - So $P(C_{1,1} = \text{yellow} \mid \underline{G = (2,3)}) = P(C_{1,1} = \text{yellow} \mid \underline{G = (2,3)}, C_{1,2} = \text{orange})$
    - I.e., $C_{1,1}$ is ***conditionally independent*** of $C_{1,2}$ ***given*** $G$

# Ghostbusters model, contd.

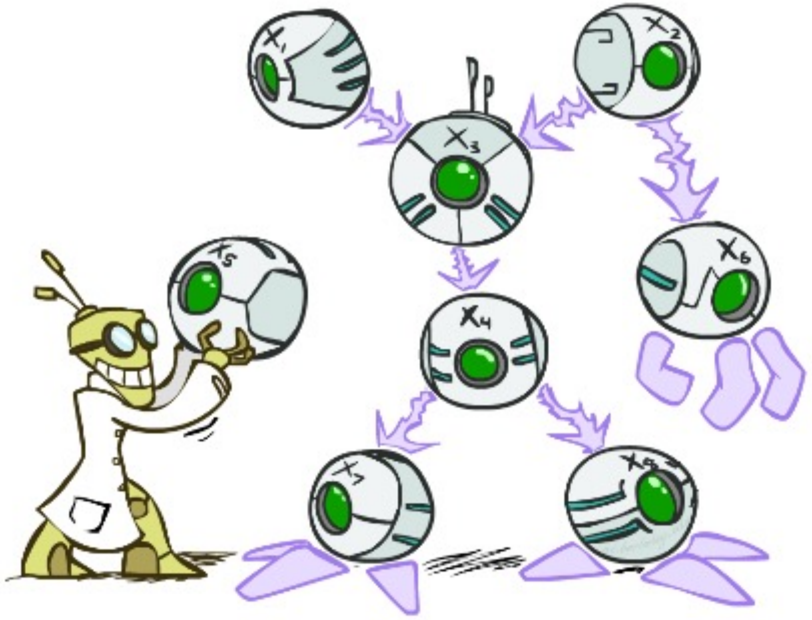
- Apply the chain rule to decompose the joint probability model:
- $\mathrm{P}(G, C_{1,1}, \ldots C_{3,3}) = \mathrm{P}(G)\, \mathrm{P}(C_{1,1} \mid G)\, \mathrm{P}(C_{1,2} \mid G, C_{1,1})\, \mathrm{P}(C_{1,3} \mid G, C_{1,1}, C_{1,2}) \ldots \mathrm{P}(C_{3,3} \mid G, C_{1,1}, \ldots, C_{3,2})$
- Now simplify using conditional independence:
- $\mathrm{P}(G, C_{1,1}, \ldots C_{3,3}) = \mathrm{P}(G)\, \mathrm{P}(C_{1,1} \mid G)\, \mathrm{P}(C_{1,2} \mid G)\, \mathrm{P}(C_{1,3} \mid G) \ldots \mathrm{P}(C_{3,3} \mid G)$
- I.e., conditional independence properties of ghostbuster physics simplify the probability model from ***exponential*** to ***quadratic*** in the number of squares

# Bayes Nets: Big Picture



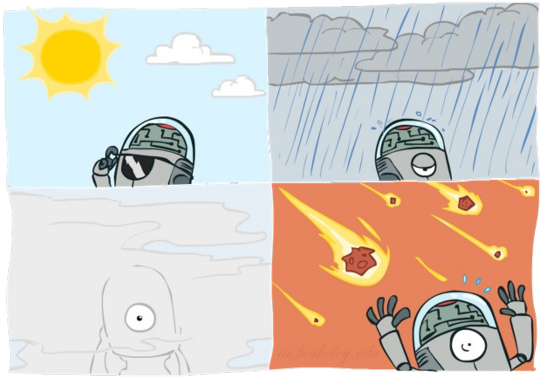
# Bayes Nets: Big Picture

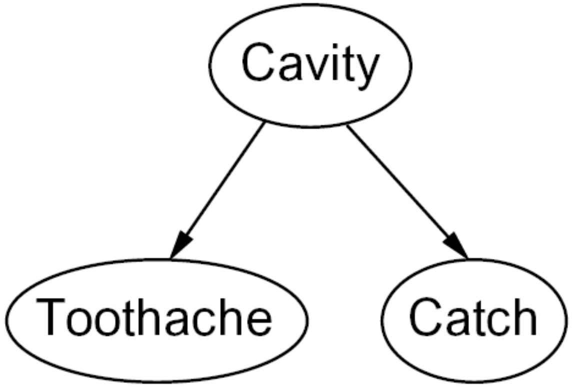
- Bayes nets: a technique for describing complex joint distributions (models) using simple, conditional distributions
  - A subset of the general class of graphical models
- Use local causality/conditional independence:
  - the world is composed of many variables,
  - each interacting locally with a few others
- Outline
  - Representation
  - Exact inference
  - Approximate inference

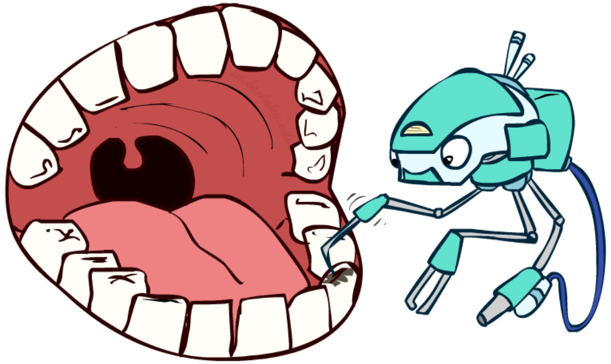


# Graphical Model Notation

- Nodes: variables (with domains)
  - Can be assigned (observed) or unassigned (unobserved)
- Arcs: interactions
  - Indicate “direct influence” between variables
  - Formally: absence of arc encodes conditional independence (more later)
- For now: imagine that arrows mean direct causation (in general, they don’t!)

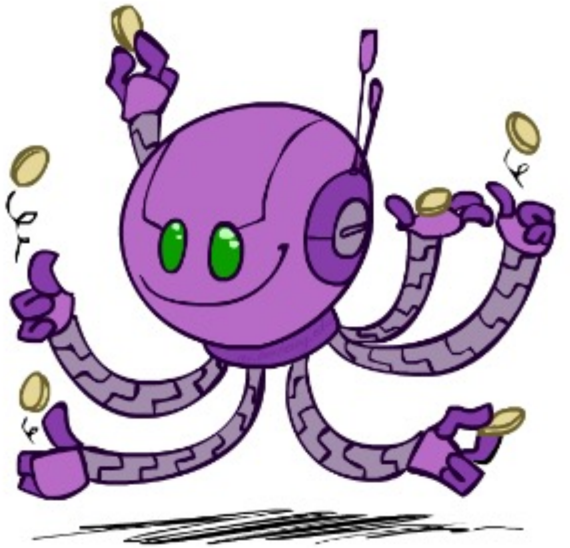





# Example: Coin Flips

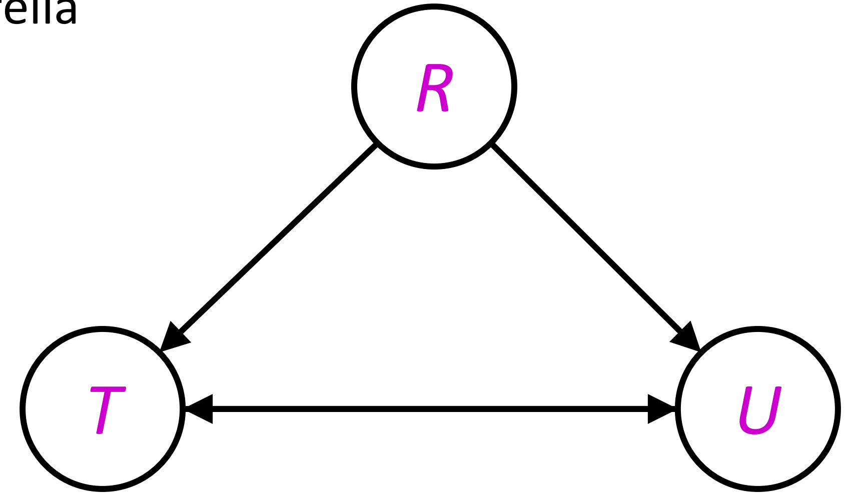
- *n* independent coin flips

- No interactions between variables: strict independence

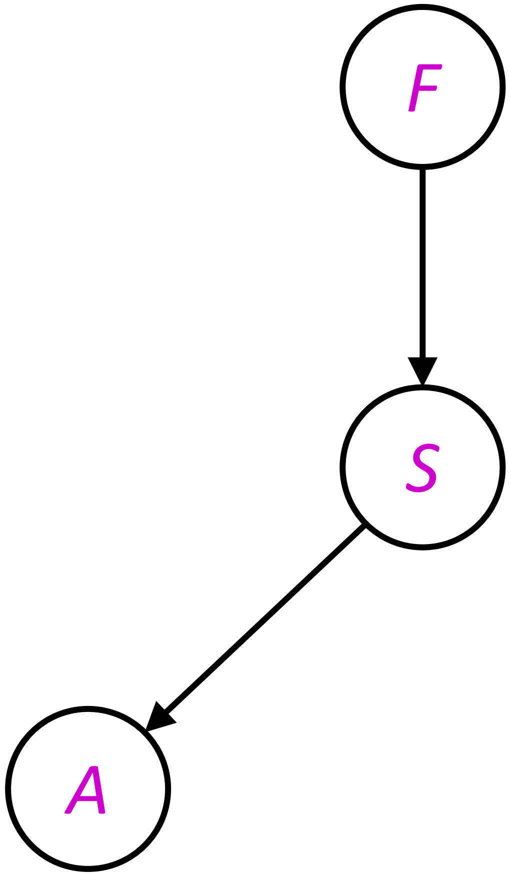
# Example: Traffic

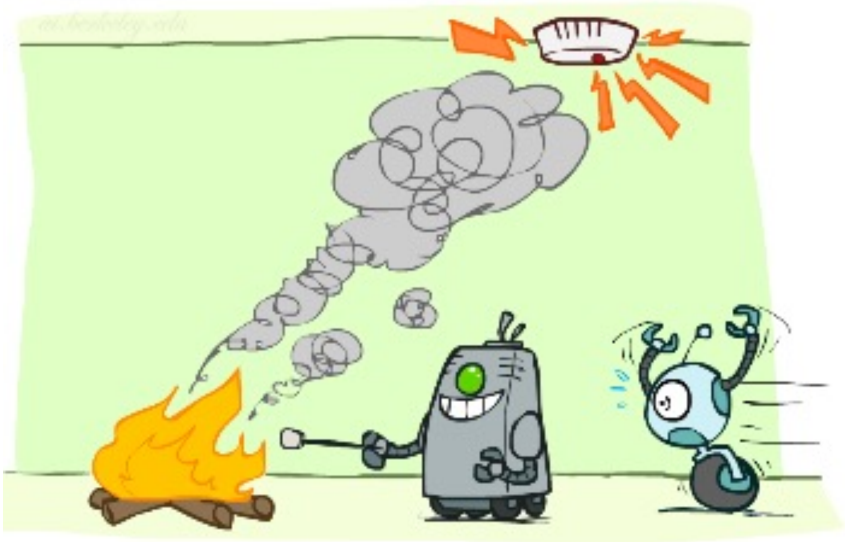
- Variables:
  - ***T***: There is traffic
  - ***U***: I'm holding my umbrella
  - ***R***: It rains

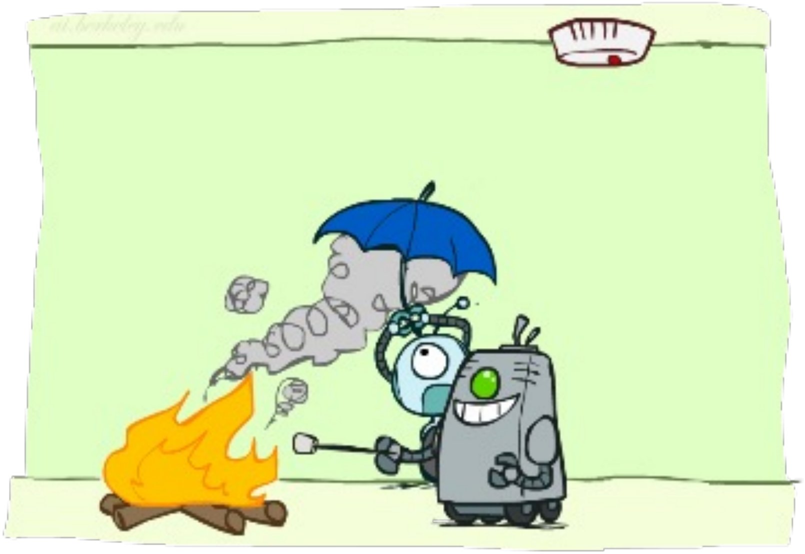



# Example: Smoke alarm

- Variables:
  - ***F***: There is fire
  - ***S***: There is smoke
  - ***A***: Alarm sounds

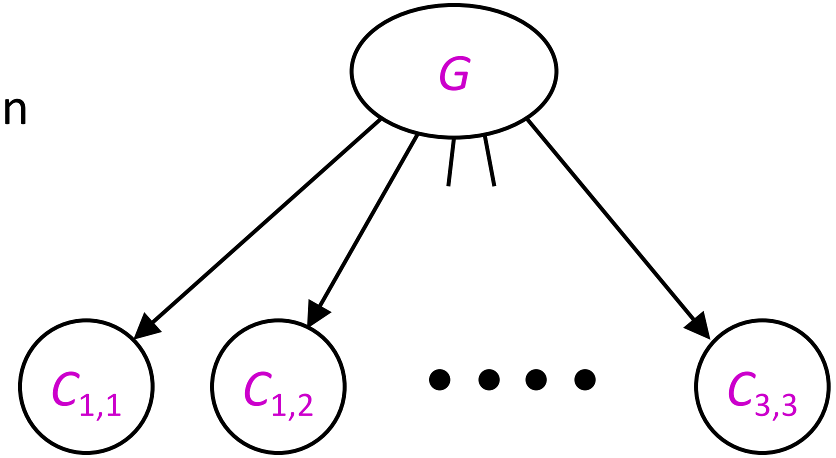


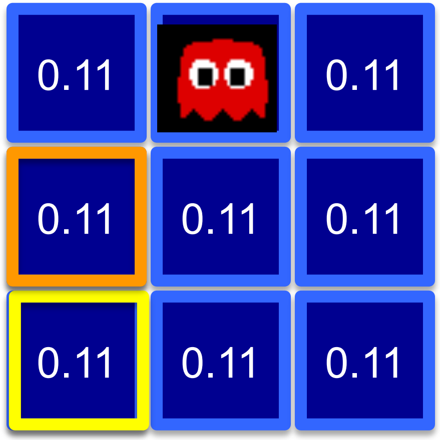





# Example: Ghostbusters

- Variables:
  - $G$: The ghost's location
  - $C_{1,1}$ , … $C_{3,3}$ :
    The observation at each location

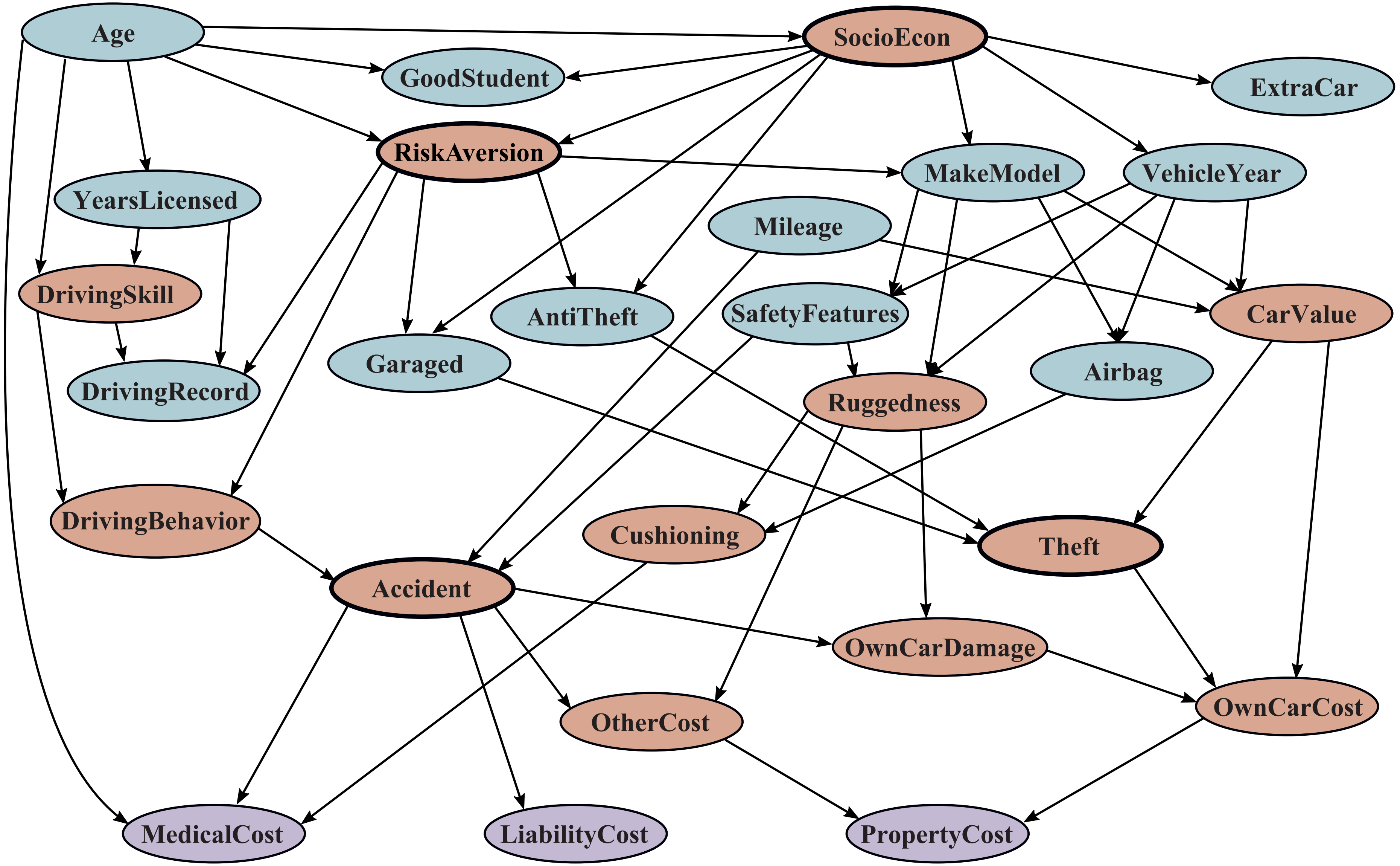


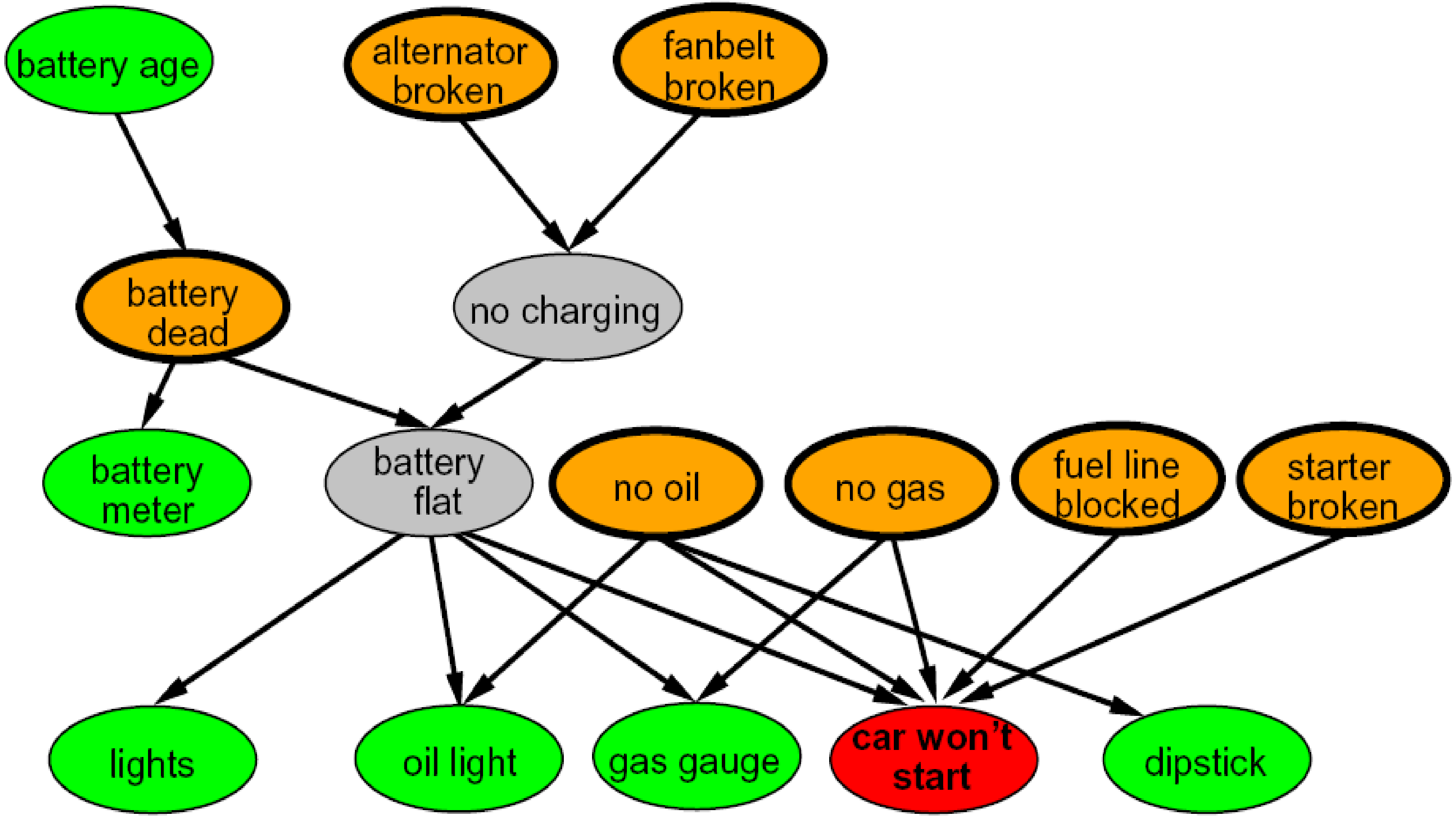

- Want to estimate:
  $P( G \mid C_{1,1} , \dots C_{3,3} )$

- This is called a ***Naïve Bayes*** model:
  - One discrete query variable (often called the ***class*** or ***category*** variable)
  - All other variables are (potentially) evidence variables
  - Evidence variables are all conditionally independent given the query variable

# Example Bayes' Net: Car Insurance



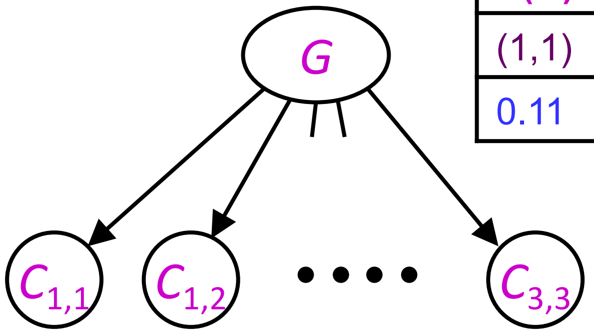
# Example Bayes' Net: Car Won't Start



# Bayes Net Syntax and Semantics

# Bayes Net Syntax

- A set of nodes, one per variable $X_i$
- A directed, acyclic graph
- A conditional distribution for each node given its ***parent variables*** in the graph
  - ***CPT*** (conditional probability table); each row is a distribution for child given values of its parents



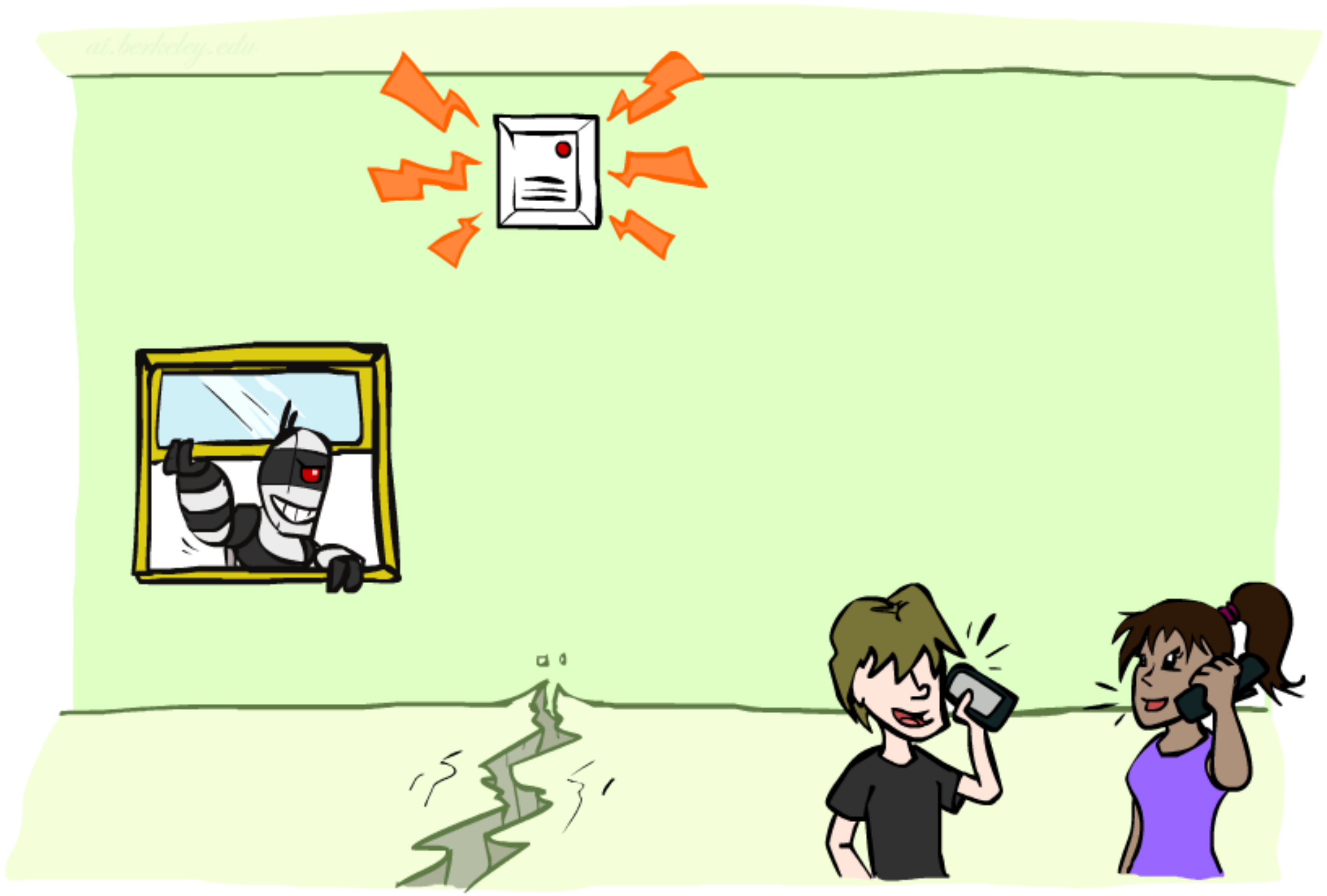

| **P(*G*)** | | | |
|---|---|---|---|
| (1,1) | (1,2) | (1,3) | … |
| 0.11 | 0.11 | 0.11 | … |

| ***G*** | **P(*C*$_{1,1}$ \| *G*)** | | | |
|---|---|---|---|---|
| | g | y | o | r |
| (1,1) | 0.01 | 0.1 | 0.3 | 0.59 |
| (1,2) | 0.1 | 0.3 | 0.5 | 0.1 |
| (1,3) | 0.3 | 0.5 | 0.19 | 0.01 |
| … | | | | |

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

# Example: Alarm Network

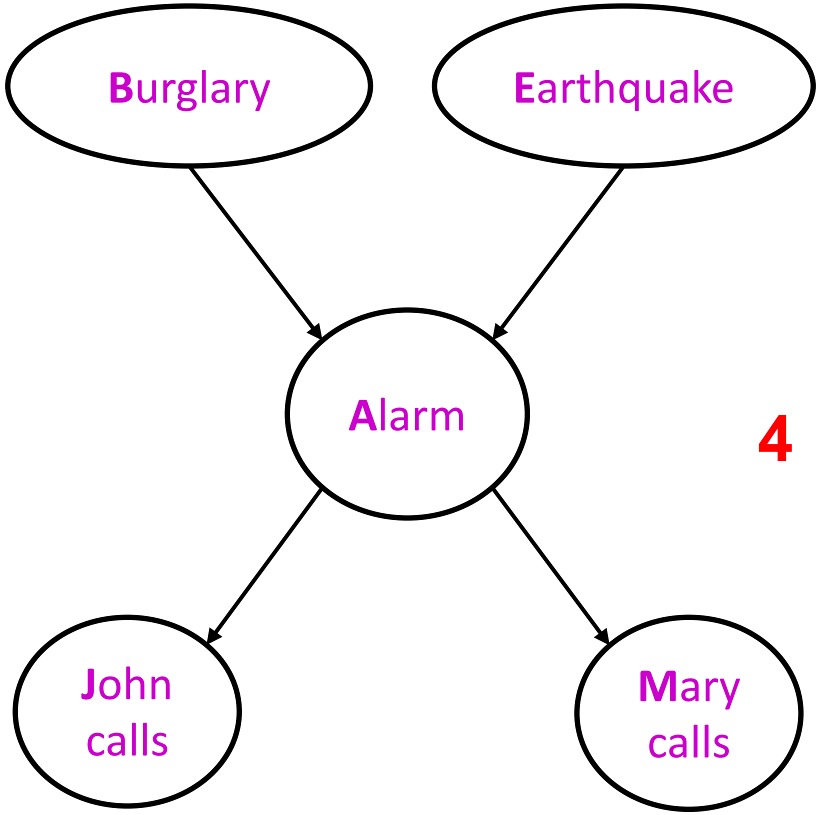
- Variables
  - B: Burglary
  - E: Earthquake
  - A: Alarm goes off
  - J: John calls
  - M: Mary calls



# Example: Alarm Network

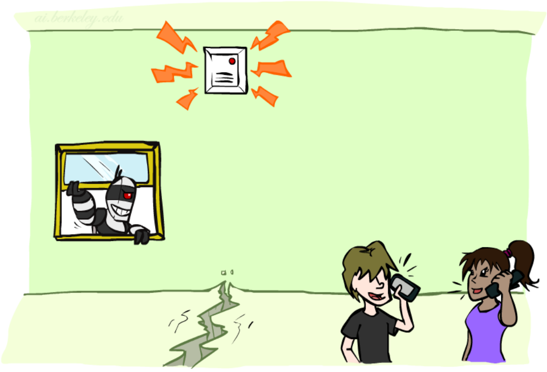
| P(B) | |
|---|---|
| true | false |
| 0.001 | 0.999 |

1




1

| P(E) | |
|---|---|
| true | false |
| 0.002 | 0.998 |



4

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

| A | P(J\|A) | |
|---|---|---|
| | true | false |
| true | 0.9 | 0.1 |
| false | 0.05 | 0.95 |

2

2

| A | P(M\|A) | |
|---|---|---|
| | true | false |
| true | 0.7 | 0.3 |
| false | 0.01 | 0.99 |

Number of ***free parameters*** in each CPT:

Parent range sizes $d_1,\ldots,d_k$

Child range size $d$

Each table row must sum to 1

$$(d-1)\prod_i d_i$$

# Bayes net global semantics


- Bayes nets encode joint distributions as product of conditional distributions on each variable:

$$P(X_1,..,X_n) = \prod_i P(X_i \mid Parents(X_i))$$

- Exploits sparse structure: number of parents is usually small

# Size of a Bayes Net

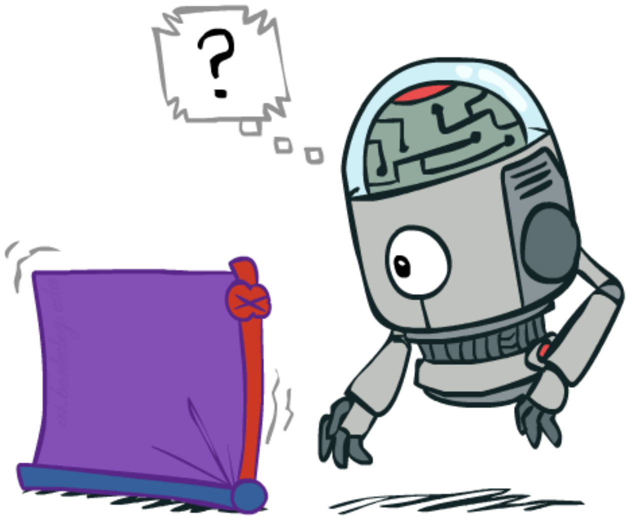
- How big is a joint distribution over $N$ variables, each with $d$ values?

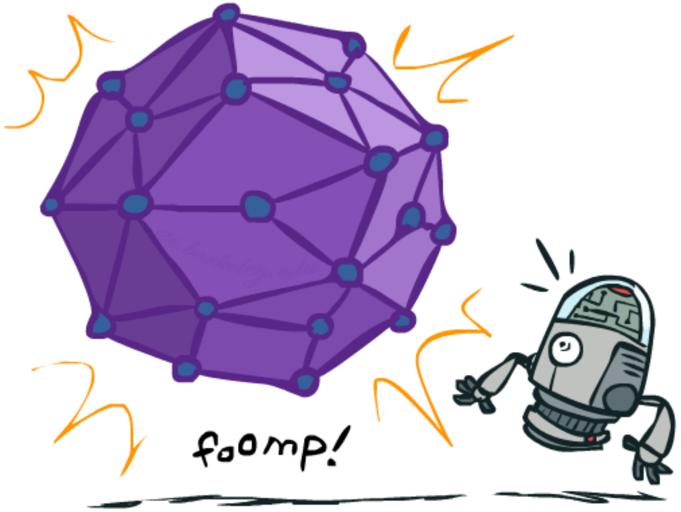
  $d^N$

- How big is an $N$-node net if nodes have at most $k$ parents?

  $O(N * d^k)$

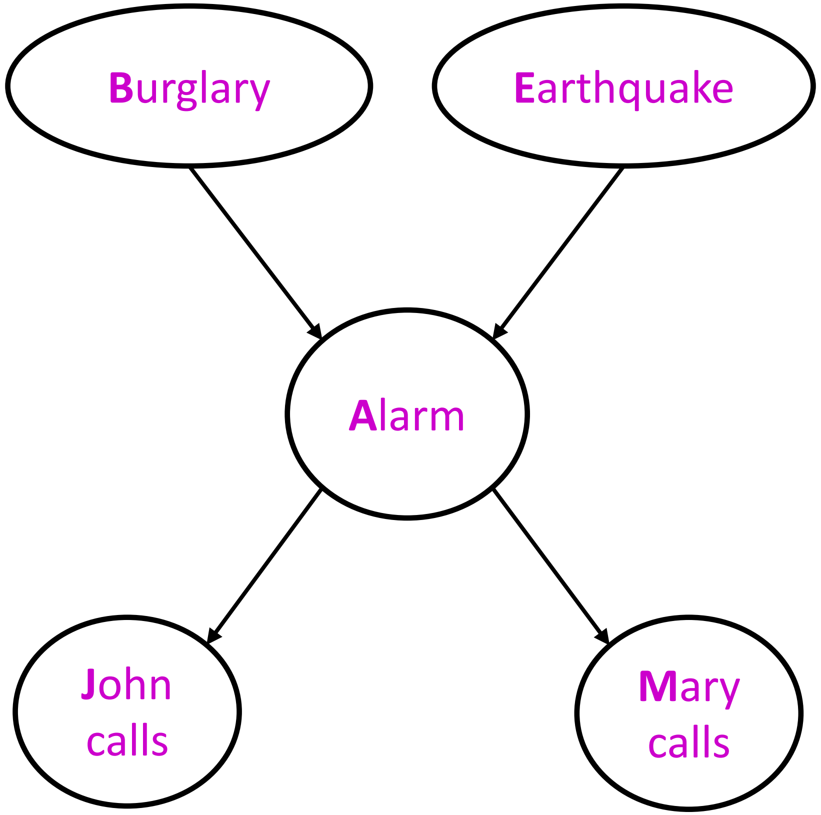
- Both give you the power to calculate $P(X_1, X_2, \ldots, X_N)$
- Bayes Nets: huge space savings with sparsity!
- Also easier to elicit local CPTs
- Also faster to answer queries (coming)





# Example

| P(B) | |
|---|---|
| true | false |
| 0.001 | 0.999 |




| P(E) | |
|---|---|
| true | false |
| 0.002 | 0.998 |

$P(b, \neg e, a, \neg j, \neg m) =$

$P(b)\ P(\neg e)\ P(a|b,\neg e)\ P(\neg j|a)\ P(\neg m|a)$

$=.001x.998x.94x.1x.3=.000028$

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

| A | P(J\|A) | |
|---|---|---|
| | true | false |
| true | 0.9 | 0.1 |
| false | 0.05 | 0.95 |

| A | P(M\|A) | |
|---|---|---|
| | true | false |
| true | 0.7 | 0.3 |
| false | 0.01 | 0.99 |

# Conditional independence in BNs


- Compare the Bayes net global semantics

$$P(X_1,..,X_n) = \prod_i P(X_i \mid Parents(X_i))$$

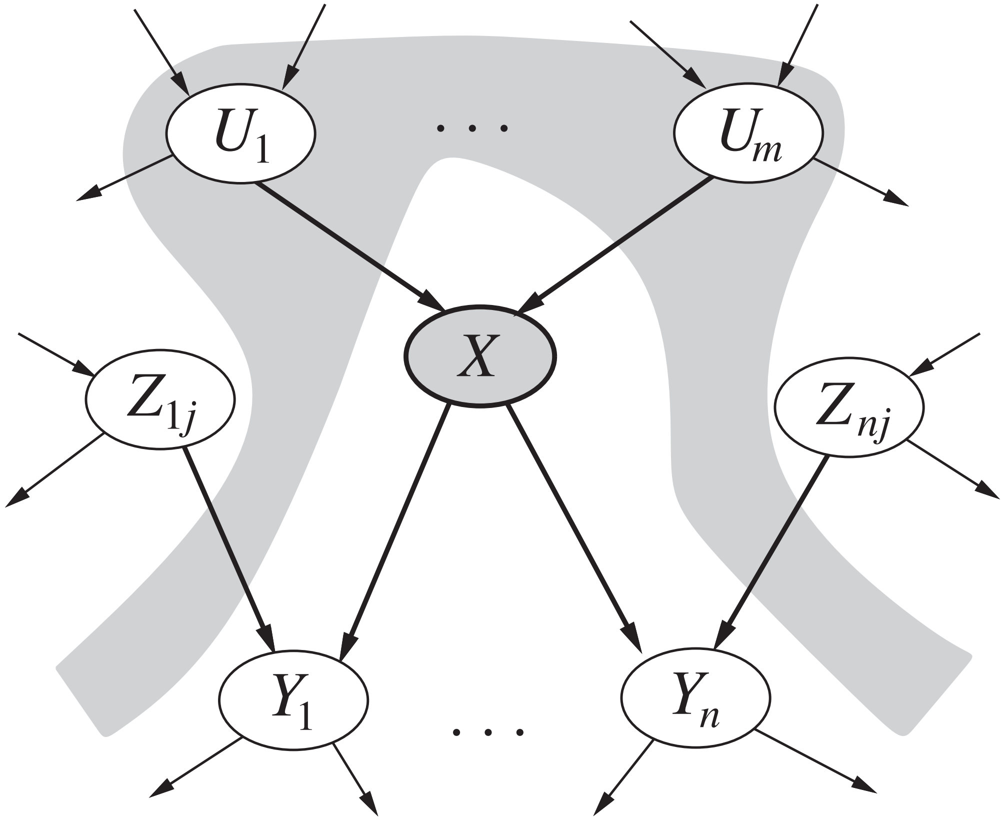
  with the chain rule identity

$$P(X_1,..,X_n) = \prod_i P(X_i \mid X_1,\ldots,X_{i-1})$$

- Assume (without loss of generality) that $X_1,..,X_n$ sorted in topological order according to the graph (i.e., parents before children), so $Parents(X_i) \subseteq X_1,\ldots,X_{i-1}$
- So the Bayes net asserts conditional independences $P(X_i \mid X_1,\ldots,X_{i-1}) = P(X_i \mid Parents(X_i))$
  - To ensure these are valid, choose parents for node $X_i$ that "shield" it from other predecessors

# Conditional independence semantics

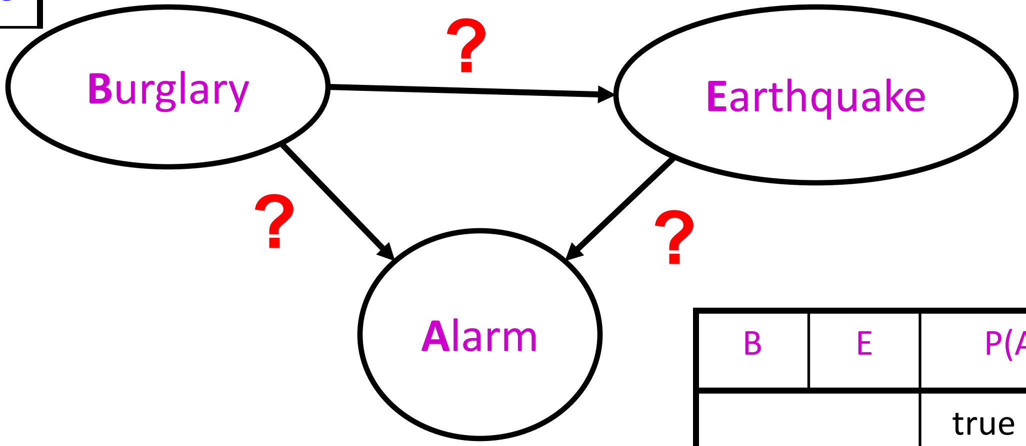
- ***Every variable is conditionally independent of its non-descendants given its parents***
- Conditional independence semantics <=> global semantics



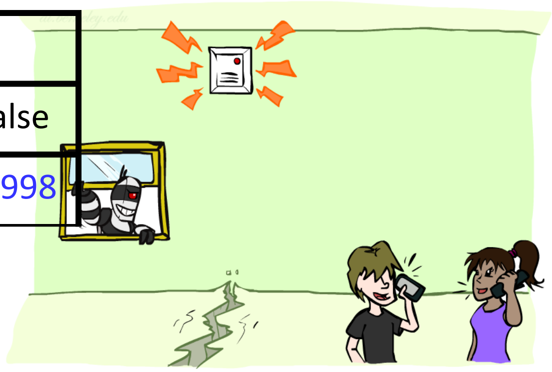
# Example: Burglary

- Burglary
- Earthquake
- Alarm

| P(B) | |
|---|---|
| true | false |
| 0.001 | 0.999 |




| P(E) | |
|---|---|
| true | false |
| 0.002 | 0.998 |

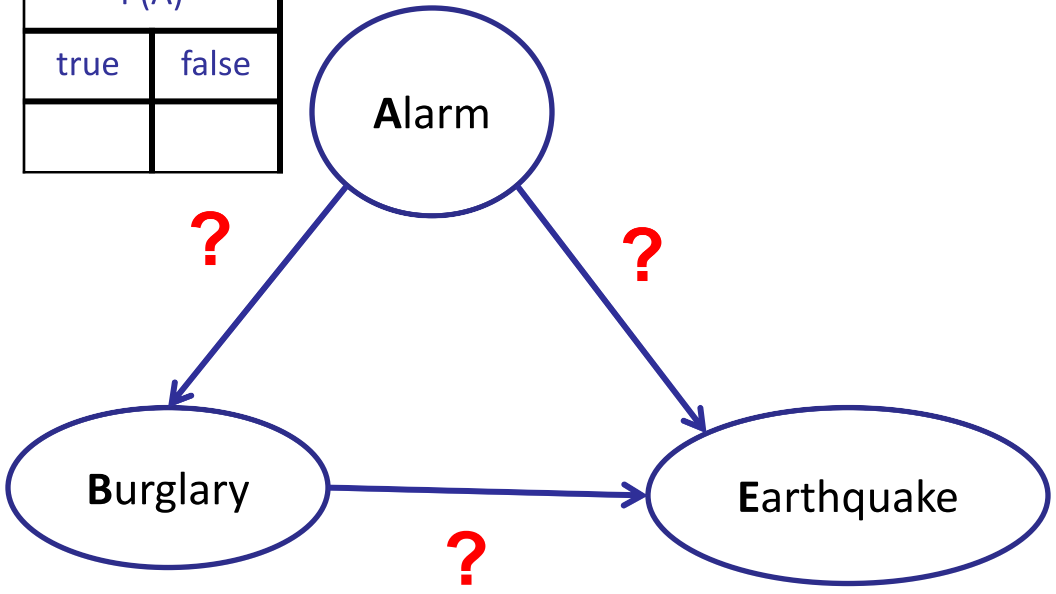


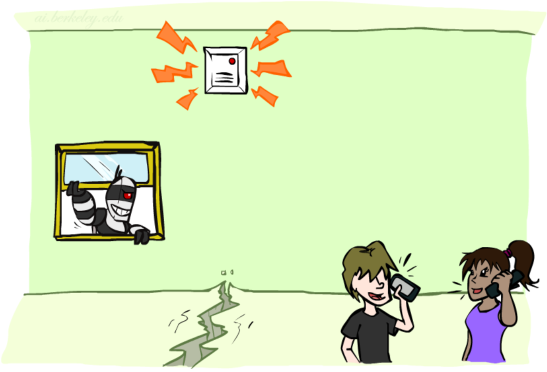
| B | E | P(A\|B,E) | |
|---|---|---|---|
| | | true | false |
| true | true | 0.95 | 0.05 |
| true | false | 0.94 | 0.06 |
| false | true | 0.29 | 0.71 |
| false | false | 0.001 | 0.999 |

# Example: Burglary

- Alarm
- Burglary
- Earthquake

| P(A) | |
|---|---|
| true | false |
| | |

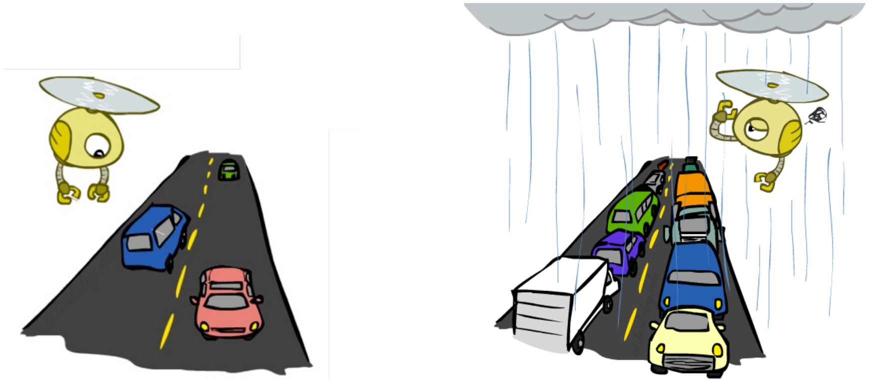


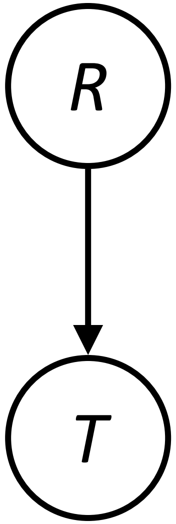



| A | P(B\|A) | |
|---|---|---|
| | true | false |
| true | ? | |
| false | | |

| A | B | P(E\|A,B) | |
|---|---|---|---|
| | | true | false |
| true | true | ? | |
| true | false | | |
| false | true | | |
| false | false | | |

# 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 |

$P(T,R)$

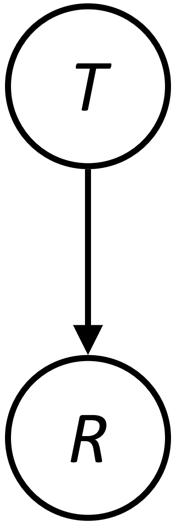
| | | |
|---|---|---|
| +r | +t | 3/16 |
| +r | -t | 1/16 |
| -r | +t | 6/16 |
| -r | -t | 6/16 |

$P(T)$

| | |
|---|---|
| +t | 9/16 |
| -t | 7/16 |

$P(R|T)$

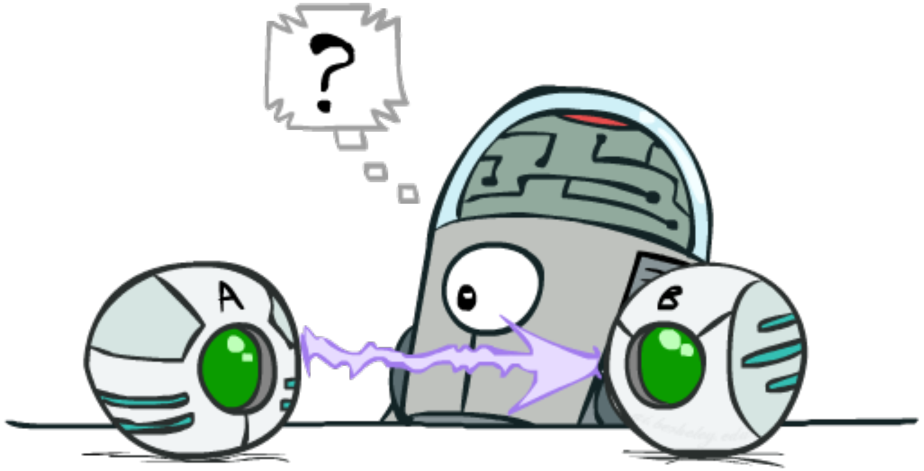
| | | |
|---|---|---|
| +t | +r | 1/3 |
| | -r | 2/3 |
| -t | +r | 1/7 |
| | -r | 6/7 |



# Causality?

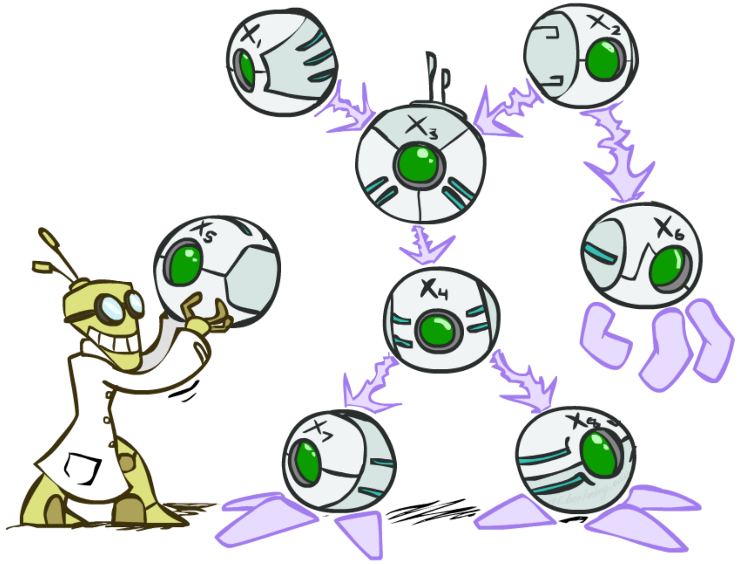
- When Bayes' nets reflect the true causal patterns:
  - Often simpler (nodes have fewer parents)
  - Often easier to think about
  - Often easier to elicit from experts
- BNs need not actually be causal
  - Sometimes no causal net exists over the domain (especially if variables are missing)
  - E.g. consider the variables *Traffic* and *Rain*
  - 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

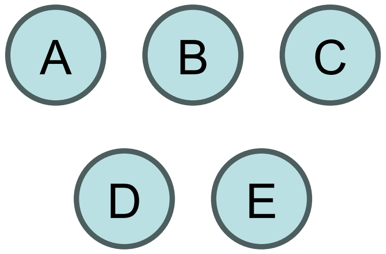
$$P(x_i|x_1,\ldots x_{i-1}) = P(x_i|\mathit{parents}(X_i))$$



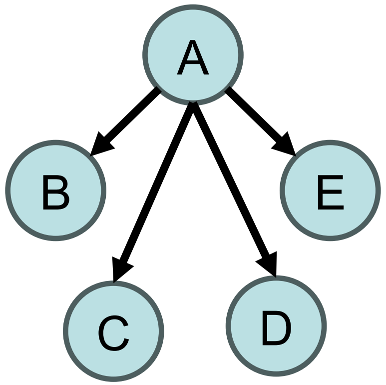
# Summary

- Independence and conditional independence are important forms of probabilistic knowledge
- Bayes net encode joint distributions efficiently by taking advantage of conditional independence
  - Global joint probability = product of local conditionals
- Allows for flexible tradeoff between model accuracy and memory/compute efficiency

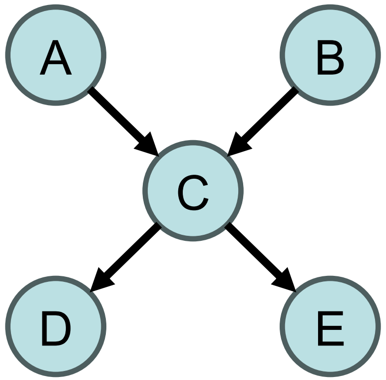




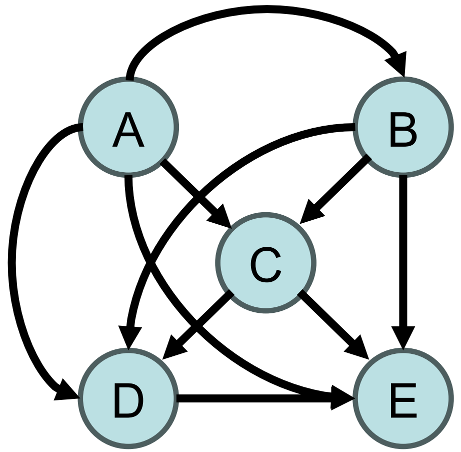
Strict Independence



Naïve Bayes



Sparse Bayes Net



Joint Distribution