Limitations of First-Order Logic

- FOL is very expressive, but...consider how to translate these:
  - "*most* students graduate in 4 years"
    - $\forall$x student(x) $\rightarrow$ duration(undergrad(x))$\leq$years(4) (all???)
  - "*only a few* students switch majors"
    - $\exists$s,m1,m2,t1,t2 student(s)^major(s,m1,t1)$\wedge$major(s,m2,t2) $\wedge$m1$\neq$m2 $\wedge$ t1$\neq$t2 (exists???)
  - "all birds can fly, *except* penguins, stuffed birds, plastic birds, birds with broken wings..."

- The problem(s) with FOL involve expressing:
  - default rules & exceptions
  - degrees of truth
  - strength of rules

Add rule *strengths* or *priorities*

- label each rule with a number indicating its "strength" or "degree of belief"
- stronger rules override conclusions from weaker rules

$$\text{penguin}(x) \rightarrow_{0.9} \neg\text{flies}(x)$$

$$\text{bird}(x) \rightarrow_{0.5} \text{flies}(x)$$

- an old ad-hoc approach (with unclear semantics)
- common approach in early Expert Systems
- "salience" attribute of rules in CLIPS

Probability (Ch. 12)

- an alternative route to encoding default rules like "most birds fly" is to quantify it using probability, $p(fly|bird)$=0.95
- probabilistic reasoning has had a major impact on AI over the years
    - conferences and journals on UAI (Uncertainty in AI)
- probabilistic models has led to major algorithms like:
    - Hidden Markov Models (applications to speech, genomics...)
    - SLAM (simultaneous localization and mapping) for robotics
    - Bayesian networks/graphical models (as knowledge bases)
    - Kalman filters, ICA, POMPDs, ...
    - Reinforcement Learning

Axioms of Probability

- for event e: $0 \leq P(e) \leq 1$
- for mutually exclusive events $e_1..e_n$: $\Sigma_i P(e_i) = 1$
- negation: $P(\neg e) = 1-P(e)$
- Kolmogorov axiom for non-exclusive events:

  $P(a \vee b)=P(a)+P(b)-P(a,b)$

Prior and Conditional Probabilities

- encode knowledge in the form of *prior* probabilities and *conditional* probabilities
  - P(x speaks portugese)=0.012
  - P(x is from Brazil)=0.007
  - P(x speaks portugese|x is from Brazil)=0.9
  - P(x flies|x is a bird)=0.9 (?)

prior probs

conditional probs

- inference is done by calculating *posterior* probabilities given evidence (using Bayes' Rule)
  - compute P(cavity | toothache, flossing, dental history, recent consumption of candy...)
  - compute P(fed will raise interest rate | unemployment=5%, inflation=0.5%, GDP=2%, recent geopolitical events...)

Bayes' Rule

- product rule : *joint prob* P(A,B) = P(A|B)*P(B)
  - P(A|B) is read as "probability of A *given* B"
  - in general, P(A,B)≠P(A)*P(B) (unless A and B are independent)
- Bayes' Rule: convert between causal and diagnostic

$$P(H \mid E) = \frac{P(E \mid H) \cdot P(H)}{P(E)}$$

H = hypothesis (cause, disease)
E = evidence (effect, symptoms)

- joint probabilities: P(E,H), priors: P(H)
- conditional probabilities play role of "rules"
  - people with a toothache are likely to have a cavity
  - p(cavity|toothache) = 0.6

Causal vs. diagnostic knowledge

- *causal*: P(x has a toothache|x has a cavity)=0.9
- *diagnostic*: P(x has a cavity|x has a toothache)=0.6

- typically it is easier to articulate knowledge in the causal direction, but we often want to use it in a diagnostic way to make inferences from observations

- Joint probability table (JPT)
  - you can calculate answer to any question from JPT
  - the problem is there are exponential # of entries (2^N, where N is the number of binary random variables)

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

P(¬*cavity* | *toothache*) = ?

- Joint probability table (JPT)
  - you can calculate answer to any question from JPT
  - the problem is there are exponential # of entries (2^N, where N is the number of binary random variables)

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

$$P(\neg cavity \mid toothache) = P(\neg cavity \wedge toothache) / P(toothache)$$

$$= \frac{0.016+0.064}{(0.108 + 0.012 + 0.016 + 0.064)}$$

$$= 0.4$$

- Joint probability table (JPT)
  - you can calculate answer to any question from JPT
  - the problem is there are exponential # of entries (2^N, where N is the number of binary random variables)

	toothache		*¬ toothache*	
	catch	*¬ catch*	*catch*	*¬ catch*
cavity	.108	.012	.072	.008
¬ cavity	.016	.064	.144	.576

$$P(\neg cavity \mid toothache) = P(\neg cavity \wedge toothache) / P(toothache)$$

$$= \frac{0.016+0.064}{(0.108 + 0.012 + 0.016 + 0.064)}$$

$$= 0.4$$

- **marginalization** - summing out unknown variables

P(cavity) = P(cavity,toothache,catch)+P(cavity,toothache,¬catch)
+P(cavity,¬toothache,catch)+P(cavity,¬toothache,¬catch)

$$P(cavity) = 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

	toothache		*¬ toothache*		
	catch	*¬ catch*	*catch*	*¬ catch*	
cavity	**.108**	**.012**	**.072**	**.008**	=0.2
¬ cavity	**.016**	**.064**	**.144**	**.576**	

- **normalization**
  - suppose we want to compute a *conditional* prob, like $P(X|Y,Z)$
  - using the product rule, we could calculate it using *joint* probs:
    - $P(X|Y,Z) = P(X,Y,Z)/P(Y,Z)$
  - would have to marginalize over X to compute the denominator
    - $P(Y,Z) = P(X,Y,Z)+P(\neg X,Y,Z)$
  - a simpler way to calculate the conditional prob is to compute 2 joint probabilities, $P(X,Y,Z)$ and $P(\neg X,Y,Z)$, and normalize them so they sum up to 1 (X has to be T or F in context of Y and Z)
  - this represents the evidence "for" and "against" X, given Y and Z
    - $P(X|Y,Z) = \alpha P(X,Y,Z)$; $\alpha=1/(P(X,Y,Z)+P(\neg X,Y,Z))$
  - since we have to compute probs both *for* and *against*, it is conventional to represent them as a vector:
    - $<P(X,Y,Z),P(\neg X,Y,Z)>$
  - technically, they don't add up to 1, but we can make them sum to one by dividing by the sum to normalize them
    - $\alpha<P(X,Y,Z),P(\neg X,Y,Z)>$; $\alpha=1/(P(X,Y,Z)+P(\neg X,Y,Z))$
    - $P(X|Y,Z) = P(X,Y,Z)/(P(X,Y,Z)+P(\neg X,Y,Z))$

Conditional Independence

- Applying Bayes' Rule in larger domains has a <u>scalability</u> problem
  - the size of the JPT grows exponentially with the number of variables (2^n for n variables)
- Solution to reduce complexity:
  - employ the <u>Independence Assumption</u>
- Most variables are not strictly independent; most variables are at least partially correlated (but which is cause and which is effect?).
- However, many variables are *conditionally* independent.

A and B are *conditionally independent* <u>given C</u> if:
$P(A,B|C) = P(A|C)P(B|C)$, or equivalently
$P(A|B,C) = P(A|C)$

Conditional Independence

If I have a cavity, the probability that the probe catches in it doesn't depend on whether I have a toothache:

(1) $P(catch|toothache, cavity) = P(catch|cavity)$

The same independence holds if I haven't got a cavity:

(2) $P(catch|toothache, \neg cavity) = P(catch|\neg cavity)$

$Catch$ is conditionally independent of $Toothache$ given $Cavity$:

$\mathbf{P}(Catch|Toothache, Cavity) = \mathbf{P}(Catch|Cavity)$

- conditional independence gives us an efficient way to combine evidence
  - consider P(Cav|toothache,catch)
  - using Bayes' Rule:
    - P(Cav|toothache,catch) $\propto$ P(toothache^catch|Cav)*P(Cav)
    - this requires a mini JPT for all combinations of evidence
  - assuming toothache is conditionally independent of catch given Cavity:
    - P(toothache^catch|Cav) = P(toothache|Cav)*P(catch|Cav)
    - therefore...

    P(Cav|toothache,catch) $\propto$ P(toothache|Cav)*P(catch|Cav)*P(Cav)

Naive Bayes algorithm

- suppose you have a phenomenon that causes several different effects that could be observed
- Cause $\rightarrow$ Effect$_1$, Effect$_2$,..., Effect$_n$
- each effect is probabilistic, but *assume they are all conditionally independent of each other*
- Then an efficient method for detecting or classifying probable causes is:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) \prod_i \mathbf{P}(Effect_i \mid Cause)$$

- if you have some unobserved vars (y), could marginalize them out, but it leads to same Eqn above

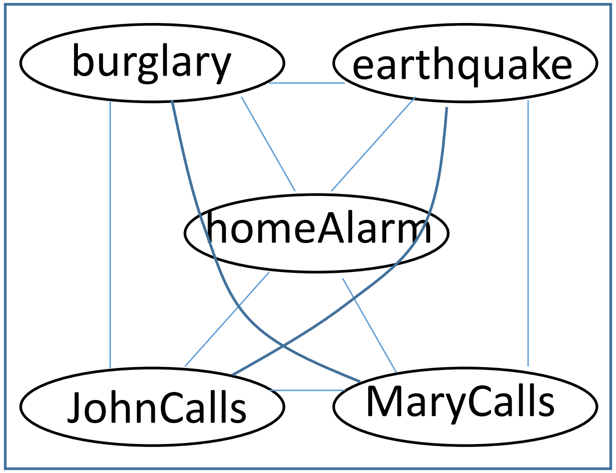
$$\mathbf{P}(Cause \mid \mathbf{e}) = \alpha \sum_{\mathbf{y}} \mathbf{P}(Cause, \mathbf{e}, \mathbf{y})$$

- Example: classifying documents as Bag-of-Words
    - P(doctype=sports|words) = P(sports)*(has "score"|sports)*(has "referee"|sports)*...

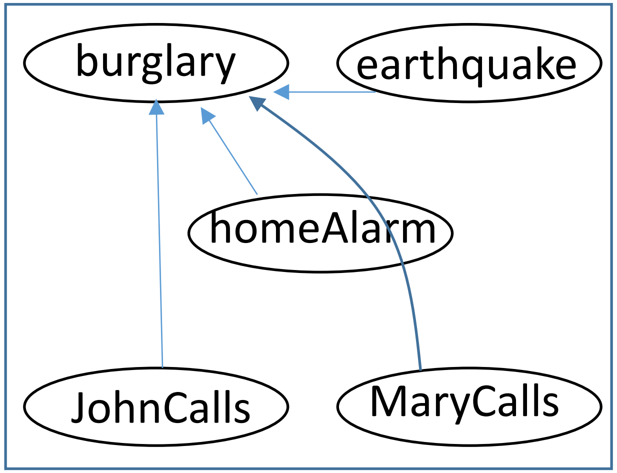
Bayesian Networks (Sec. 13.1 and the first page of Sec 13.2)

- graphical models where *edges represent conditional probabilities*
  - efficient representation because missing edges are assumed to be *conditionally independent* given the nodes in between
- popular for modern AI systems (expert systems)
  - important for handling uncertainty

all vars are correlated, O(n^2) edges, requires full JPT with 2^n rows

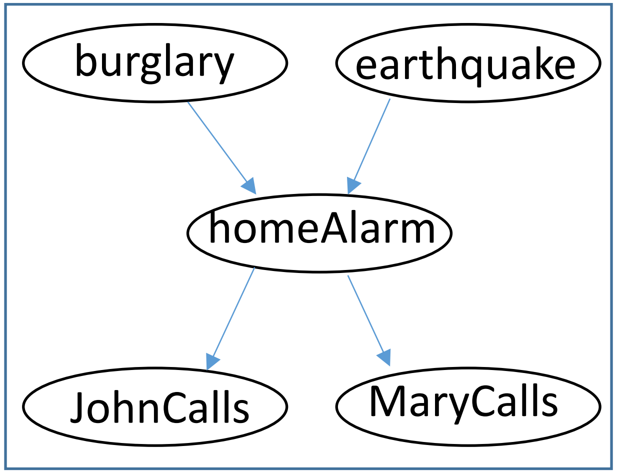


Naive Bayes: compute probability of 1 var depending on all the others (n-1)



requires *independence assumption*

Bayesian Network: selected edges represent conditional dependence

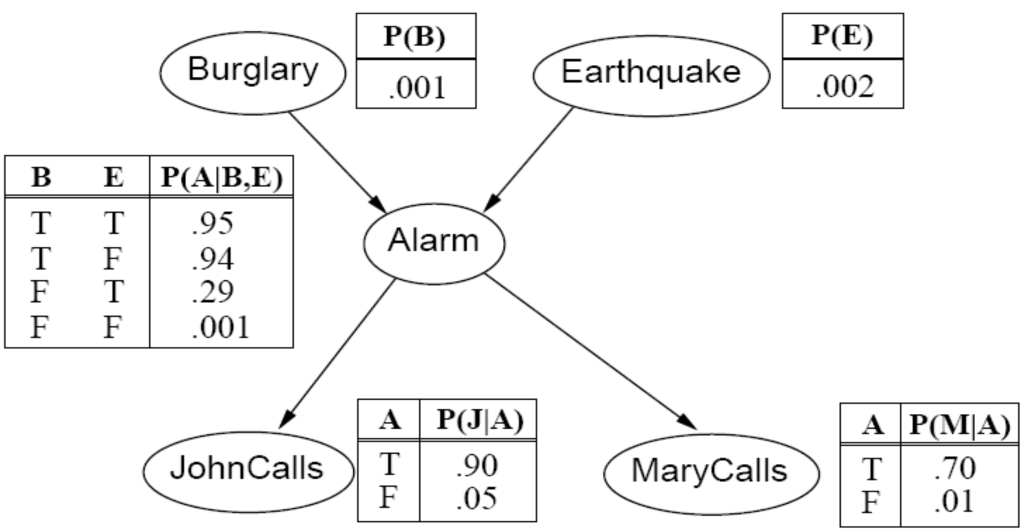


more natural: links follow causality

Bayesian Networks (Sec. 13.1-2)

- prob of each node depends on parents; specify with a mini-JPT
- full JPT has 2^5=32 entries - can answer any query from JPT
- joint prob of full state like <j,m,a,¬b,¬e> is <u>product of prob over all nodes</u>
- prob of <u>each node is conditioned on parents</u>

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



P(B)
.001

P(E)
.002

B	E	P(A|B,E)
T	T	.95
T	F	.94
F	T	.29
F	F	.001

A	P(J|A)
T	.90
F	.05

A	P(M|A)
T	.70
F	.01

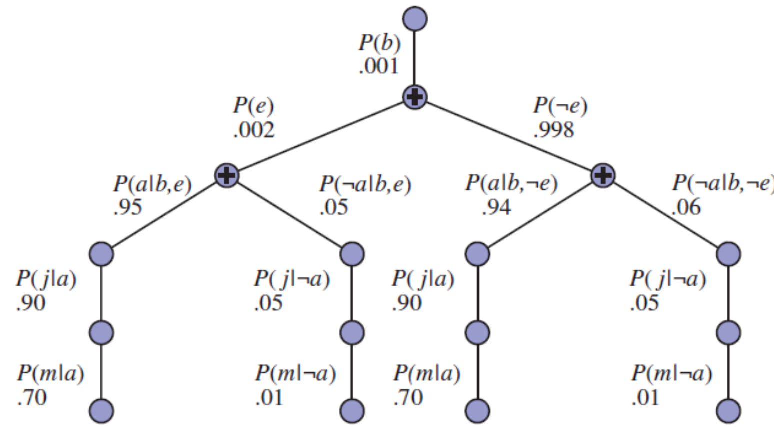
$$\begin{aligned} & P(j \wedge m \wedge a \wedge \neg b \wedge \neg e) \\ = \ & P(j|a)P(m|a)P(a|\neg b, \neg e)P(\neg b)P(\neg e) \\ = \ & 0.9 \times 0.7 \times 0.001 \times 0.999 \times 0.998 \\ \approx \ & 0.00063 \end{aligned}$$

- Efficient algorithms for computing inferences or outcomes conditioned on observations/evidence
  - Variable elimination: factor computations into a tree of products and sums (algebraic calculation from formula)
  - rearrange to minimize number of adds and mults...

$$\mathbf{P}(Burglary \mid JohnCalls = true, MaryCalls = true)$$

$$P(b \mid j,m) = \alpha \sum_e \sum_a P(b)P(e)P(a \mid b,e)P(j \mid a)P(m \mid a)$$

$$P(b \mid j,m) = \alpha P(b) \sum_e P(e) \sum_a P(a \mid b,e)P(j \mid a)P(m \mid a)$$



  - Belief propagation: graph algorithm that updates probs of neighboring nodes when belief of any node changes

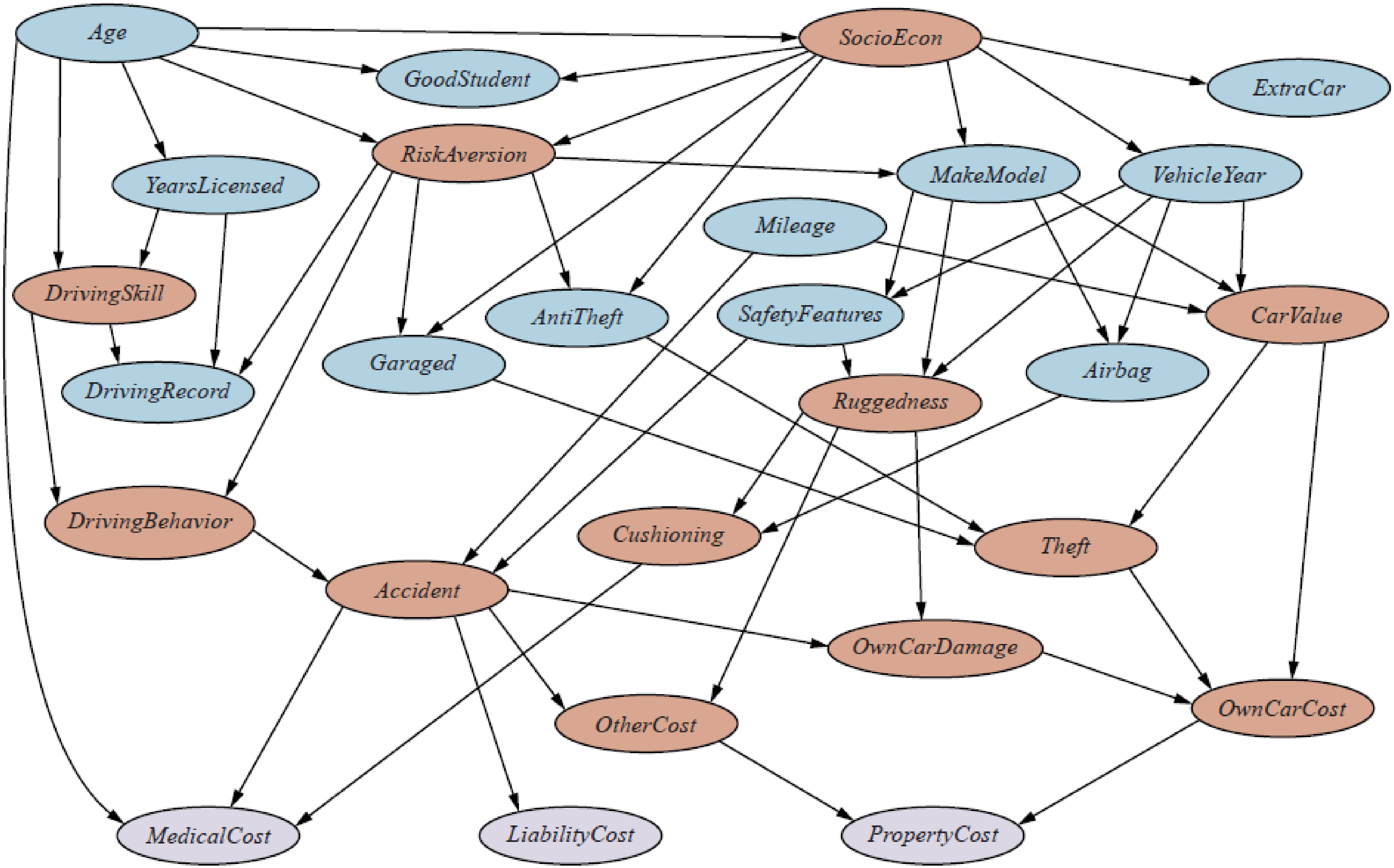
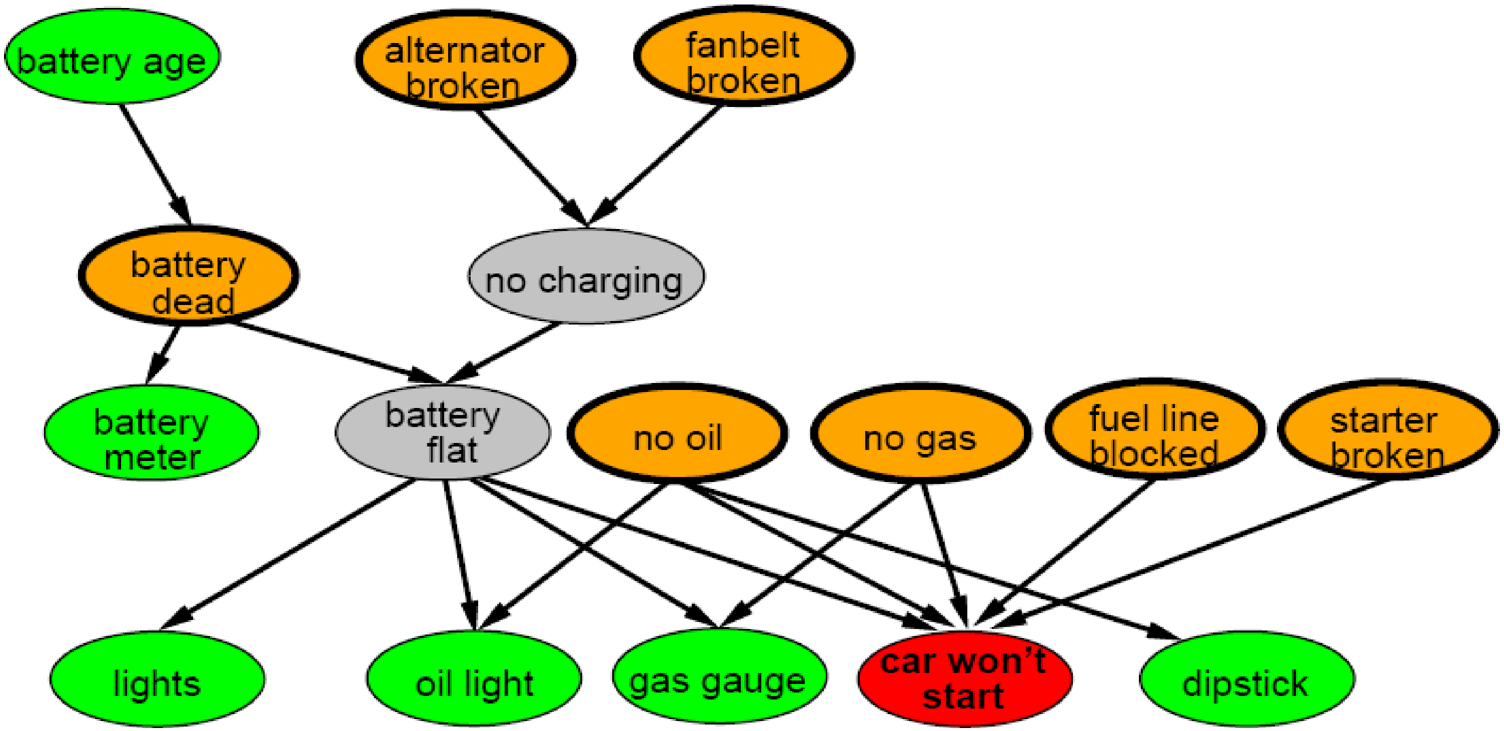


Figure 13.9 A Bayesian network for evaluating car insurance applications.

Initial evidence: car won't start

Testable variables (green), "broken, so fix it" variables (orange)

Hidden variables (gray) ensure sparse structure, reduce parameters



- Many modern knowledge-based systems are based on probabilistic inference
  - including Bayesian networks, Hidden Markov Models, (HMMs), Markov Decision Problems (MDPs)
  - example: Bayesian networks are used for inferring user goals or help needs from actions like mouse clicks in an automated software help system (think 'Clippy')
  - *Decision Theory* combines *utilities* with *probabilities* of outcomes to decide actions to take

- the challenge is capturing all the numbers needed for the prior and conditional probabilities
  - objectivists (frequentists) - probabilities represent outcomes of trials/experiments
  - subjectivists - probabilities are degrees of belief
- probability and statistics is at the core of many Machine Learning algorithms