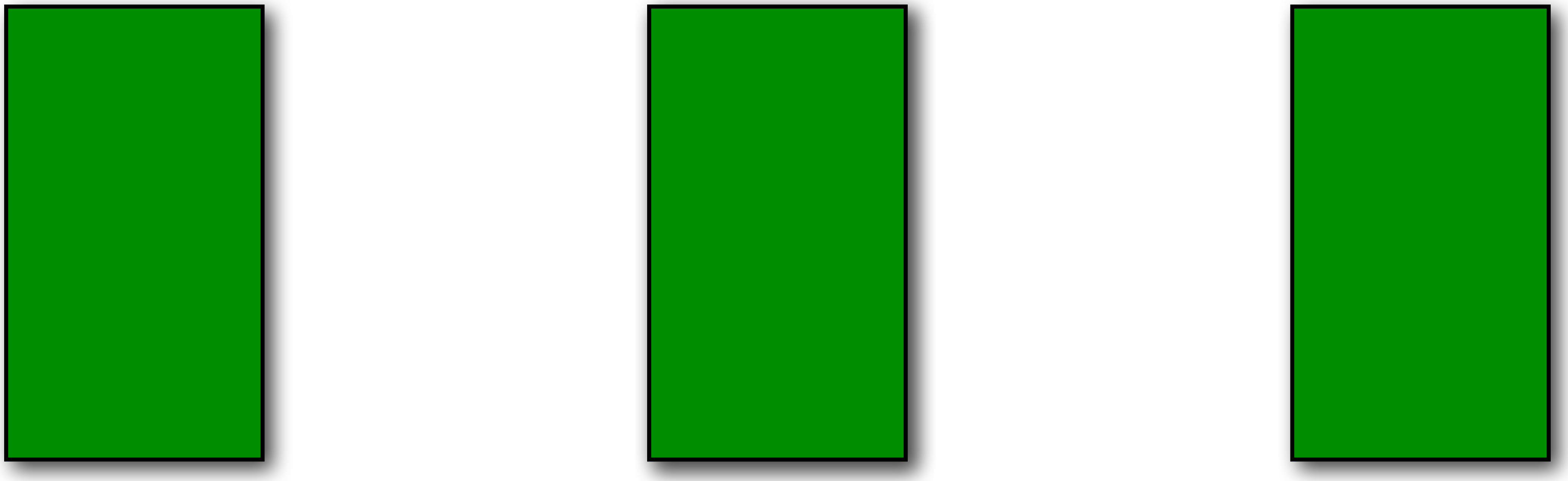
Probabilistic representation

Applied artificial intelligence (EDA132)
Lecture 05
2014-02-04
Elin A. Topp

Show time!

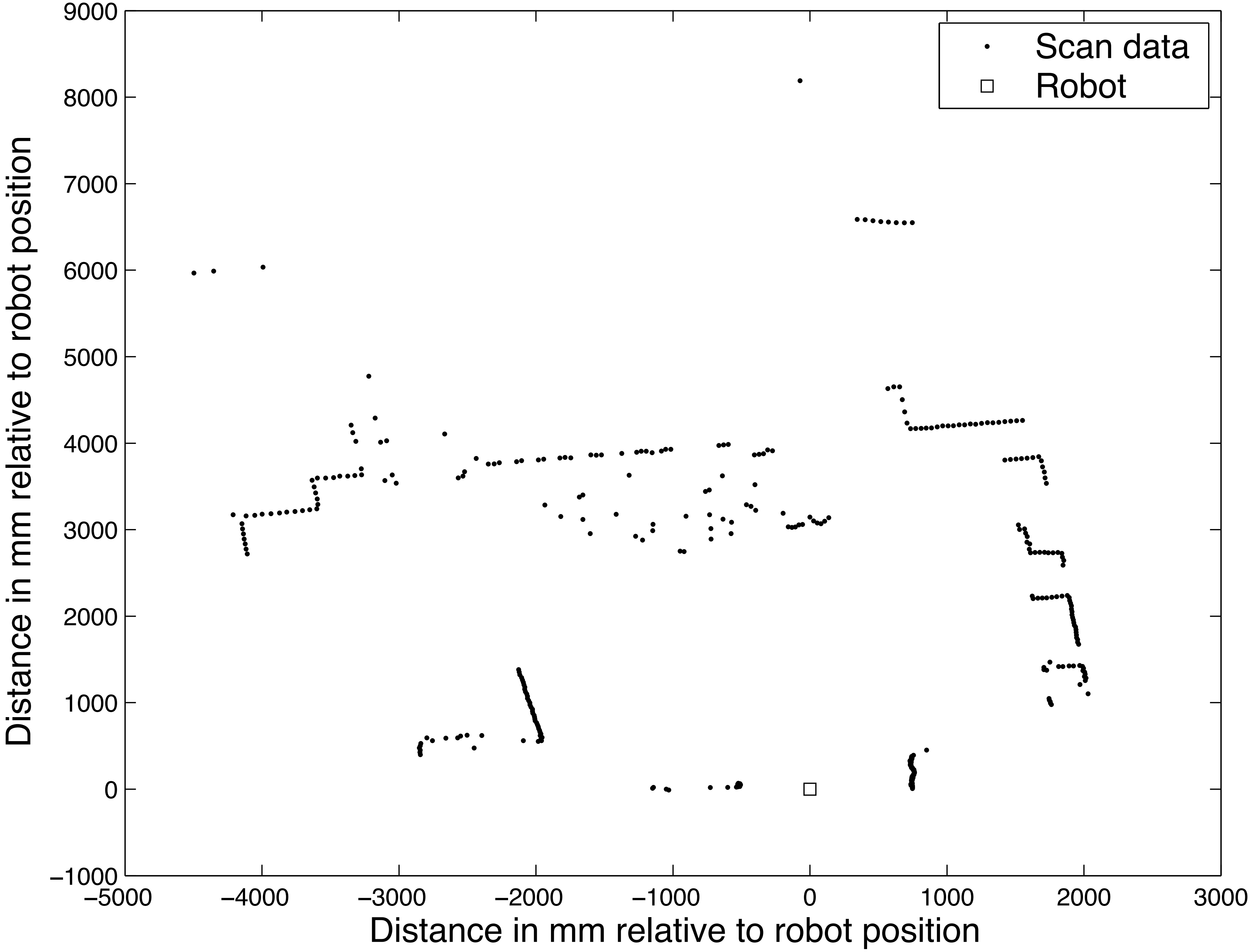
Two boxes of chocolates, one luxury car.
Where is the car?



Philosopher: It does not matter whether I change my choice, I will either get chocolates or a car.

Mathematician: It is more likely to get the car when I change my choice - even though it is not certain!

A robot's view of the world...



Outline

- Uncertainty & probability (chapter 13)
  - Uncertainty
  - Probability
  - Syntax and Semantics
  - Inference
  - Independence and Bayes' Rule
- Bayesian Networks (chapter 14.1-3)
  - Syntax
  - Semantics
  - Efficient representation

Outline

- Uncertainty & probability (chapter 13)
  - Uncertainty
  - Probability
  - Syntax and Semantics
  - Inference
  - Independence and Bayes' Rule
- Bayesian Networks (chapter 14.1-3)
  - Syntax
  - Semantics
  - Efficient representation

Uncertainty

Situation: Get to the airport in time for the flight (by car)

Action A_t := "Leave for airport t minutes before flight departs"

Question: will A_t get me there on time?

Deal with:

1) partial observability (road states, other drivers, ...)

2) noisy sensors (traffic reports)

3) uncertainty in action outcomes (flat tire, car failure, ...)

4) complexity of modeling and predicting traffic

Use pure logic? Well... :

1) risks falsehood: "A_{25} will get me there on time"

or 2) leads to conclusions too weak for decision making:

"A_{25} will get me there on time if there is no accident and it does not rain and my tires hold, and ..."

(A_{1440} would probably hold, but the waiting time would be intolerable, given the quality of airport food...)

Rational decision

A_{25}, A_{90}, A_{180}, A_{1440}, ... what is "the right thing to do?"

Obviously dependent on relative importance of goals (being in time vs minimizing waiting time) AND on their respective likelihood of being achieved.

Uncertain reasoning: diagnosing a patient, i.e., find the CAUSE for the symptoms displayed.

"Diagnostic" rule: Toothache $\Rightarrow$ Cavity ??? No!

Complex rule: Toothache $\Rightarrow$ Cavity $\vee$ GumProblem $\vee$ Abscess $\vee$ ... ??? Too much!

"Causal" rule: Cavity $\Rightarrow$ Toothache ??? Well... not always

X

Using logic?

Fixing such “rules” would mean to make them logically exhaustive, but that is bound to fail due to:

Laziness (too much work to list all options)

Theoretical ignorance (there is simply no complete theory)

Practical ignorance (might be impossible to test exhaustively)

⇒ better use ***probabilities*** to represent certain ***knowledge states***

⇒ Rational decisions (decision theory) combine probability and utility theory

X

Probability

Probabilistic assertions summarise effects of

laziness: failure to enumerate exceptions, qualifications, etc.

ignorance: lack of relevant facts, initial conditions, etc.

Subjective or Bayesian probability:

Probabilities relate propositions to one's state of knowledge

e.g., $P(A_{25} \mid \textit{no reported accidents}) = 0.06$

Not claims of a "probabilistic tendency" in the current situation, but maybe learned from past experience of similar situations.

Probabilities of propositions change with new evidence:

e.g., $P(A_{25} \mid \textit{no reported accidents, it's 5:00 in the morning}) = 0.15$

Making decisions under uncertainty

Suppose the following believes (from past experience):

$$\begin{aligned}
P(\,A_{25} \text{ gets me there on time} \mid \ldots) &= 0.04\\
P(\,A_{90} \text{ gets me there on time} \mid \ldots) &= 0.70\\
P(\,A_{120} \text{ gets me there on time} \mid \ldots) &= 0.95\\
P(\,A_{1440} \text{ gets me there on time} \mid \ldots) &= 0.9999
\end{aligned}$$

Which action to choose?

Depends on my preferences for "missing flight" vs. "waiting (with airport cuisine)", etc.

Utility theory is used to represent and infer preferences

Decision theory = utility theory + probability theory

Probability basics

A set Ω - the sample space, e.g., the 6 possible rolls of a die.

$\omega \in \Omega$ is a sample point / possible world / atomic event

A probability space of probability model is a sample space with an assignment $P(\omega)$ for every $\omega \in \Omega$ so that:

$$0 \leq P(\omega) \leq 1$$

$$\Sigma_{\omega} P(\omega) = 1$$

An event A is any subset of Ω

$$P(A) = \Sigma_{\{\omega \in A\}} P(\omega)$$

E.g., $P(\text{ die roll} < 4) = P(1) + P(2) + P(3) = 1/6 + 1/6 + 1/6 = 1/2$

X

Random variables

A random variable is a function from sample points to some range, e.g., the reals or Booleans,

e.g., $Odd(1) = true$.

P induces a *probability distribution* for any random variable X

$$P(X = x_i) = \Sigma_{\{\omega : X(\omega) = x_i\}} P(\omega)$$

e.g., $P(Odd = true) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$

X

Propositions

A proposition describes the event (set of sample points) where it (the proposition) holds, i.e.,

Given Boolean random variables A and B:

event a = *set of sample points where* $A(\omega) = true$

event $\neg a$ = *set of sample points where* $A(\omega) = false$

event $a \wedge b$ = *points where* $A(\omega) = true$ *and* $B(\omega) = true$

Often in AI applications, the sample points are defined by the values of a set of random variables, i.e., the sample space is the Cartesian product of the ranges of the variables.

Prior probability

Prior or *unconditional probabilities* of propositions

e.g., $P(\mathit{Cavity} = \mathit{true}) = 0.2$ and

$P(\mathit{Weather} = \mathit{sunny}) = 0.72$

correspond to belief *prior to the arrival of any (new) evidence*

Probability distribution gives values for all possible assignments (normalised):

$\mathbf{P}(\mathit{Weather}) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$

Joint probability distribution for a set of (independent) random variables gives the probability of every atomic event on those random variables (i.e., every sample point):

$\mathbf{P}(\mathit{Weather}, \mathit{Cavity})$ = a 4×2 matrix of values:

Weather Cavity	sunny	rain	cloudy	snow
true	0.144	0.02	0.016	0.02
false	0.576	0.08	0.064	0.08

Posterior probability

Most often, there is *some* information, i.e., *evidence*, that one can base their belief on:

e.g., *P(cavity) = 0.2* (prior, no evidence for anything), but

P(cavity | toothache) = 0.6

corresponds to belief *after the arrival of some evidence*
(also: *posterior* or *conditional probability*).

OBS: NOT "if toothache, then 60% chance of cavity"

THINK *"given that toothache is all I know" instead!*

Evidence remains valid after more evidence arrives, but it might become less useful

Evidence may be completely useless, i.e., irrelevant.

P(cavity | toothache, sunny) = P(cavity | toothache)

Domain knowledge lets us do this kind of inference.

Posterior probability (2)

Definition of conditional / posterior probability:

$$P(a \mid b) = \frac{P(a \wedge b)}{P(b)} \quad \text{if } P(b) \neq 0$$

or as *Product rule* (for a and b being true, we need b true and then a true, given b):

$$P(a \wedge b) \quad = \quad P(a \mid b)\, P(b) \quad = \quad P(b \mid a)\, P(a)$$

and in general for whole distributions (e.g.):

$$\mathbf{P}(\mathit{Weather}, \mathit{Cavity}) \quad = \quad \mathbf{P}(\mathit{Weather} \mid \mathit{Cavity})\, \mathbf{P}(\mathit{Cavity})$$

(gives a $4x2$ set of equations)

Chain rule (successive application of product rule):

$$\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_1, \ldots, X_{n-1})\, \mathbf{P}(X_n \mid X_1, \ldots, X_{n-1})$$

$$= \mathbf{P}(X_1, \ldots, X_{n-2})\, \mathbf{P}(X_{n-1} \mid X_1, \ldots, X_{n-1})\, \mathbf{P}(X_n \mid X_1, \ldots, X_{n-1})$$

$$= \ldots = \prod_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$$

Inference

Probabilistic inference:

Computation of posterior probabilities given observed evidence

starting out with the full joint distribution as “knowledge base”:

Inference by enumeration

	toothache		¬ toothache	
	catch	¬ catch	catch	¬ catch
cavity	0.108	0.012	0.072	0.008
¬ cavity	0.016	0.064	0.144	0.576

For any proposition Φ, sum the atomic events where it is true:

$$P(\Phi) = \Sigma_{\omega:\omega\models\Phi} P(\omega)$$

$$P(\text{ cavity} \lor \text{toothache}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

$$P(\text{ toothache}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

Can also compute posterior probabilities:

$$P(\neg\text{cavity} \mid \text{toothache}) = \frac{P(\neg\text{cavity} \land \text{toothache})}{P(\text{ toothache})}$$

$$= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

Normalisation

	toothache		¬ toothache	
	catch	¬ catch	catch	¬ catch
cavity	0.108	0.012	0.072	0.008
¬ cavity	0.016	0.064	0.144	0.576

Denominator can be viewed as a *normalisation constant*:

$$\mathbf{P}(Cavity \mid toothache) = \alpha\, \mathbf{P}(Cavity, toothache)$$

$$= \alpha\,[\mathbf{P}(Cavity, toothache, catch) + \mathbf{P}(Cavity, toothache, \neg catch)]$$

$$= \alpha\,[\langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle]$$

$$= \alpha\, \langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle$$

And the good news:

We can compute $\mathbf{P}(Cavity \mid toothache)$ without knowing the value of $P(toothache)$!

The suicidal student

A young student kills herself. Her diary is found. In the diary she speculates about her childhood and the possibility of her father abusing her during childhood. She had reported headaches to her friends and therapist, and started the diary due to the therapist's recommendation.

The father ends up in court, since

"headaches are caused by PTSD, and PTSD is caused by abuse"

Would you agree?

Psychologist knowing "the math" argues:

P(headache | PTSD) = high (statistics)

P(PTSD | abuse in childhood) = high (statistics)

ok, yes, sure, but:

You did not consider the relevant relations of

P(PTSD | headache) or

P(abuse in childhood | PTSD),

i.e., you mixed up cause and effect in your argumentation!

Bayes' Rule

Recap *product rule*: $P(a \wedge b) = P(a \mid b)\, P(b) = P(b \mid a)\, P(a)$

$\Rightarrow$ Bayes' Rule $P(a \mid b) = \dfrac{P(b \mid a)\, P(a)}{P(b)}$

or in distribution form:

$$\mathbf{P}(Y \mid X) = \frac{\mathbf{P}(X \mid Y)\, \mathbf{P}(Y)}{\mathbf{P}(X)} = \alpha\, \mathbf{P}(X \mid Y)\, \mathbf{P}(Y)$$

Useful for assessing *diagnostic* probability from *causal* probability

$$P(Cause \mid Effect) = \frac{P(Effect \mid Cause)\, P(Cause)}{P(Effect)}$$

E.g., with M "meningitis", S "stiff neck":

$$P(m \mid s) = \frac{P(s \mid m)\, P(m)}{P(s)} = \frac{0.8 * 0.0001}{0.1} = 0.0008$$ (not too bad, really!)

All is well that ends well ...

We can model cause-effect relationships,

we can base our judgement on mathematically sound inference,

we can even do this inference with only partial knowledge on the priors, ...

... but

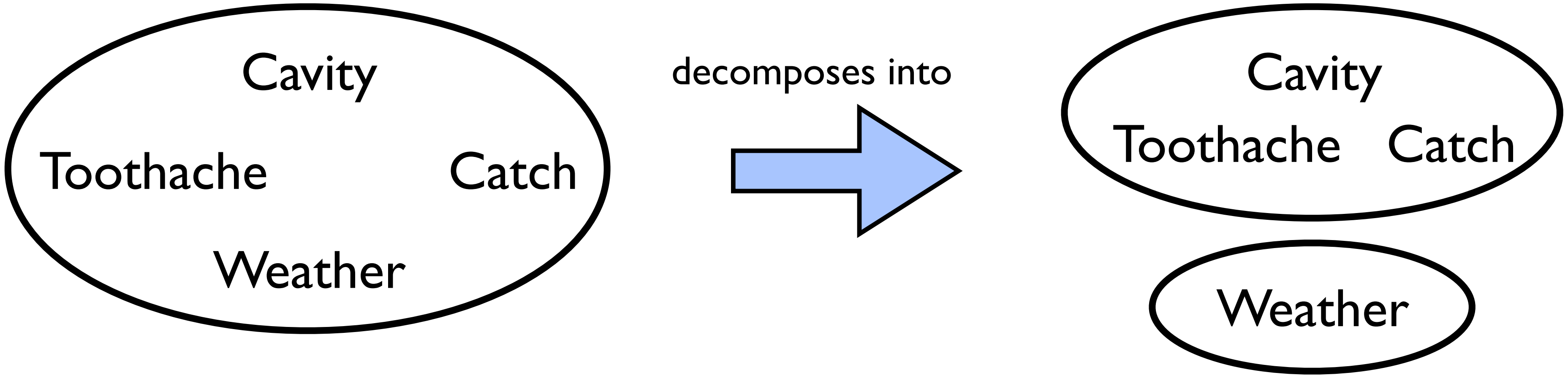
n Boolean variables give us an input table of size $O(2^n)$ ...

(and for non-Booleans it gets even more nasty...)

Independence

A and B are *independent* iff

$\mathbf{P}(A \mid B) = \mathbf{P}(A)$ or $\mathbf{P}(B \mid A) = \mathbf{P}(B)$ or $\mathbf{P}(A, B) = \mathbf{P}(A)\,\mathbf{P}(B)$



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

32 entries reduced to 8 + 4. This absolute independence is powerful but rare!

Some fields (like dentistry) have still a lot, maybe hundreds, of variables, none of them being independent.

What can be done to overcome this mess...?

Conditional independence

$\mathbf{P}(\mathit{Toothache}, \mathit{Cavity}, \mathit{Catch})$ has $2^3 - 1 = 7$ independent entries (must sum up to 1)

But: If there is a cavity, the probability for “catch” does not depend on whether there is a toothache:

(1) $P(\mathit{catch} \mid \mathit{toothache}, \mathit{cavity}) = P(\mathit{catch} \mid \mathit{cavity})$

The same holds when there is no cavity:

(2) $P(\mathit{catch} \mid \mathit{toothache}, \neg \mathit{cavity}) = P(\mathit{catch} \mid \neg \mathit{cavity})$

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

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

Writing out full joint distribution using chain rule:

$$\begin{aligned}
&\mathbf{P}(\mathit{Toothache}, \mathit{Catch}, \mathit{Cavity})\\
&= \mathbf{P}(\mathit{Toothache} \mid \mathit{Catch}, \mathit{Cavity})\, \mathbf{P}(\mathit{Catch}, \mathit{Cavity})\\
&= \mathbf{P}(\mathit{Toothache} \mid \mathit{Catch}, \mathit{Cavity})\, \mathbf{P}(\mathit{Catch} \mid \mathit{Cavity})\, \mathbf{P}(\mathit{Cavity})\\
&= \mathbf{P}(\mathit{Toothache} \mid \mathit{Cavity})\, \mathbf{P}(\mathit{Catch} \mid \mathit{Cavity})\, \mathbf{P}(\mathit{Cavity})
\end{aligned}$$

gives thus $2 + 2 + 1 = 5$ independent entries

Conditional independence (2)

In most cases, the use of conditional independence reduces the size of the representation of the joint distribution from exponential in n to linear in n.

Hence:

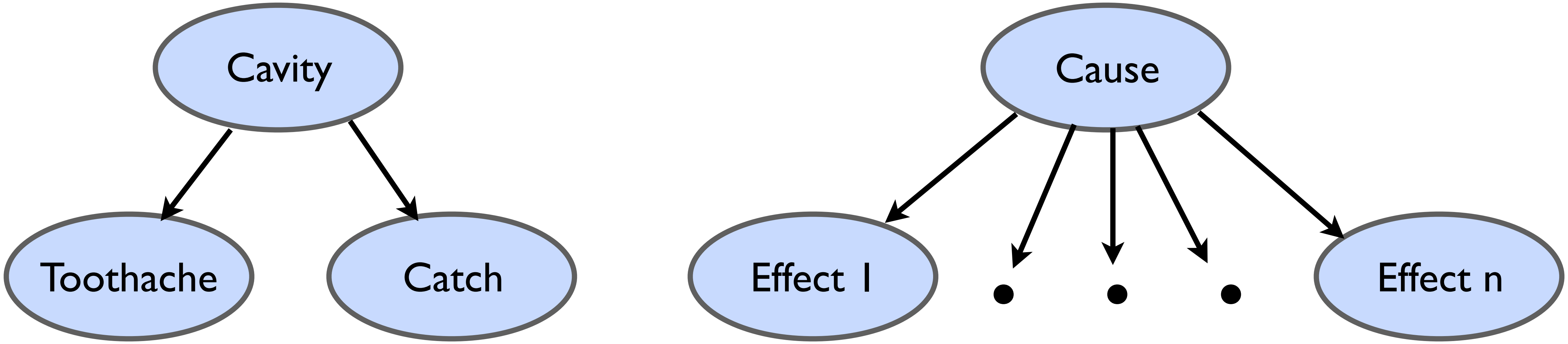
Conditional independence is our most basic and robust form of knowledge about uncertain environments

Bayes' Rule and conditional independence

$\mathbf{P}(\mathit{Cavity} \mid \mathit{toothache} \wedge \mathit{catch})$
$= \alpha\ \mathbf{P}(\mathit{toothache} \wedge \mathit{catch} \mid \mathit{Cavity})\ \mathbf{P}(\mathit{Cavity})$
$= \alpha\ \mathbf{P}(\mathit{toothache} \mid \mathit{Cavity})\ \mathbf{P}(\mathit{catch} \mid \mathit{Cavity})\ \mathbf{P}(\mathit{Cavity})$

An example of a *naive Bayes* model:

$$\mathbf{P}(\mathit{Cause}, \mathit{Effect}_1, \ldots, \mathit{Effect}_n) = \mathbf{P}(\mathit{Cause}) \prod_i \mathbf{P}(\mathit{Effect}_i \mid \mathit{Cause})$$



The total number of parameters is *linear* in *n*

Summary

Probability is a way to formalise and represent uncertain knowledge

The *joint probability distribution* specifies probability over every *atomic event*

Queries can be answered by *summing* over atomic events

Bayes' rule can be applied to compute posterior probabilities so that *diagnostic* probabilities can be assessed from *causal* ones

For *nontrivial* domains, we must find a way to *reduce* the joint size

Independence and *conditional independence* provide the tools

Outline

- Uncertainty & probability (chapter 13)
  - Uncertainty
  - Probability
  - Syntax and Semantics
  - Inference
  - Independence and Bayes' Rule
- Bayesian Networks (chapter 14.1-3)
  - Syntax
  - Semantics
  - Efficient representation

Outline

- Uncertainty (chapter 13)
  - Uncertainty
  - Probability
  - Syntax and Semantics
  - Inference
  - Independence and Bayes' Rule
- Bayesian Networks (chapter 14.1-3)
  - Syntax
  - Semantics
  - Efficient representation

Bayesian networks

A simple, graphical notation for conditional independence assertions and hence for compact specification of full joint distributions

Syntax:

a set of nodes, one per random variable

a directed, acyclic graph (link ≈ "directly influences")

a conditional distribution for each node given its parents:

$\mathbf{P}(X_i \mid Parents(X_i))$

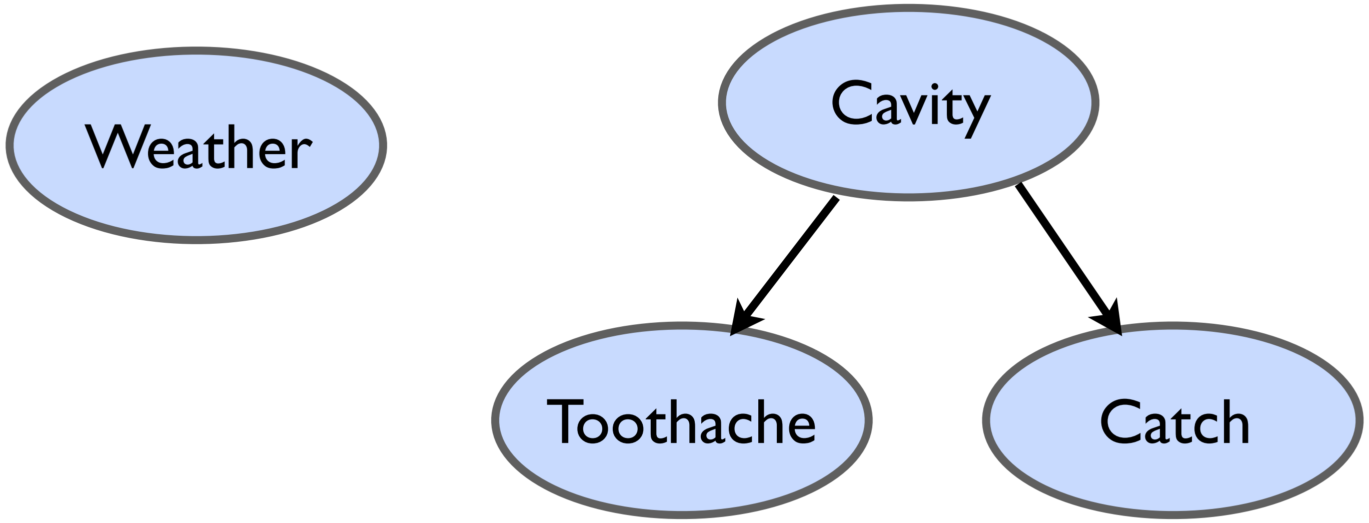
In the simplest case, conditional distribution represented as a

conditional probability table (CPT)

giving the distribution over X_i for each combination of parent values

Example

Topology of network encodes conditional independence assertions:



Weather is independent of the other variables

Toothache and *Catch* are conditionally independent given *Cavity*

Example 2

I am at work, my neighbour John calls to say my alarm is ringing, but neighbour Mary does not call.

Sometimes the alarm is set off by minor earthquakes.

Is there a burglar?

Variables: *Burglar*, *Earthquake*, *Alarm*, *JohnCalls*, *MaryCalls*

Network topology reflects "causal" knowledge:

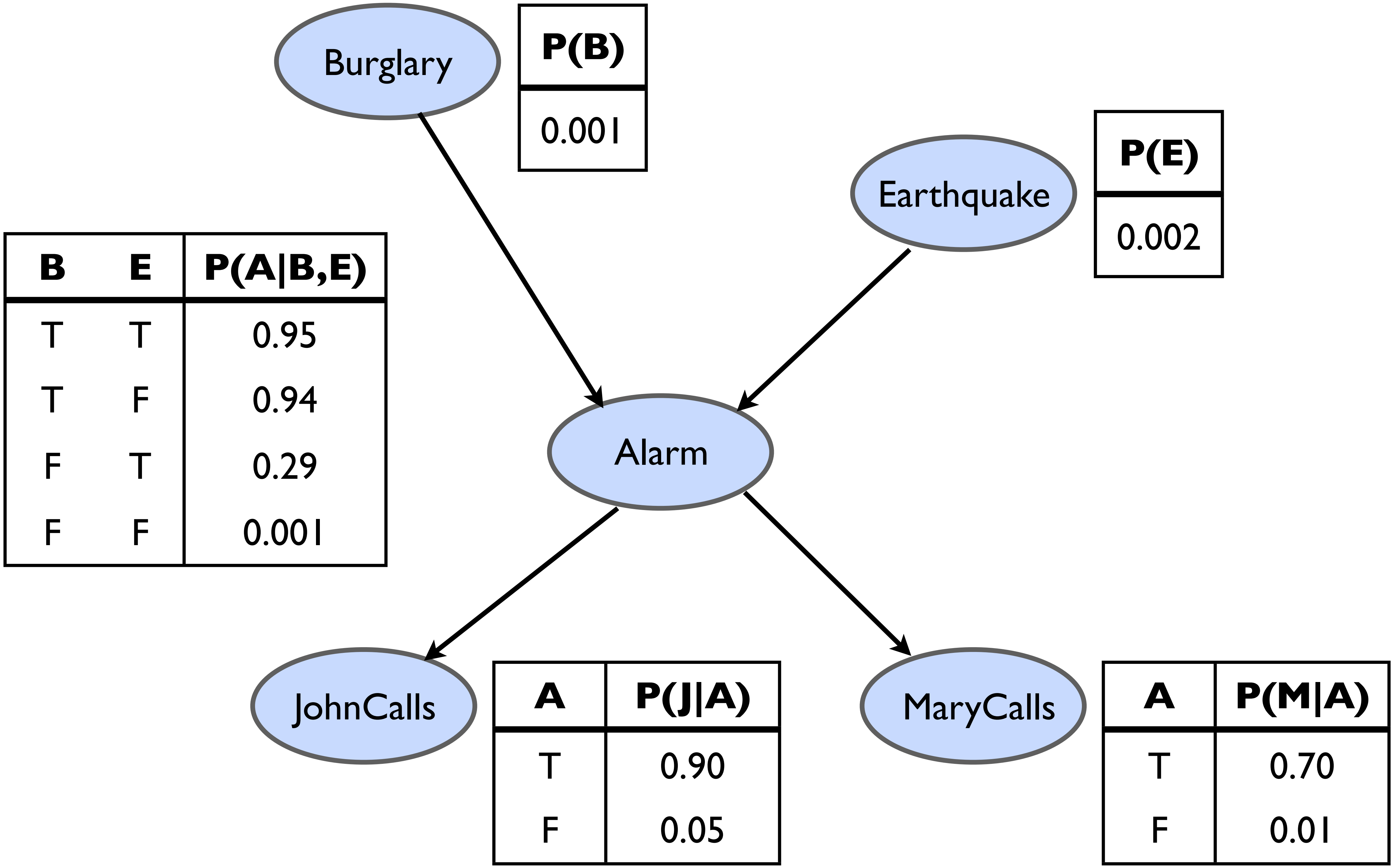
A burglar can set the alarm off

An earthquake can set the alarm off

The alarm can cause John to call

The alarm can cause Mary to call

Example 2 (2)



P(B)
0.001

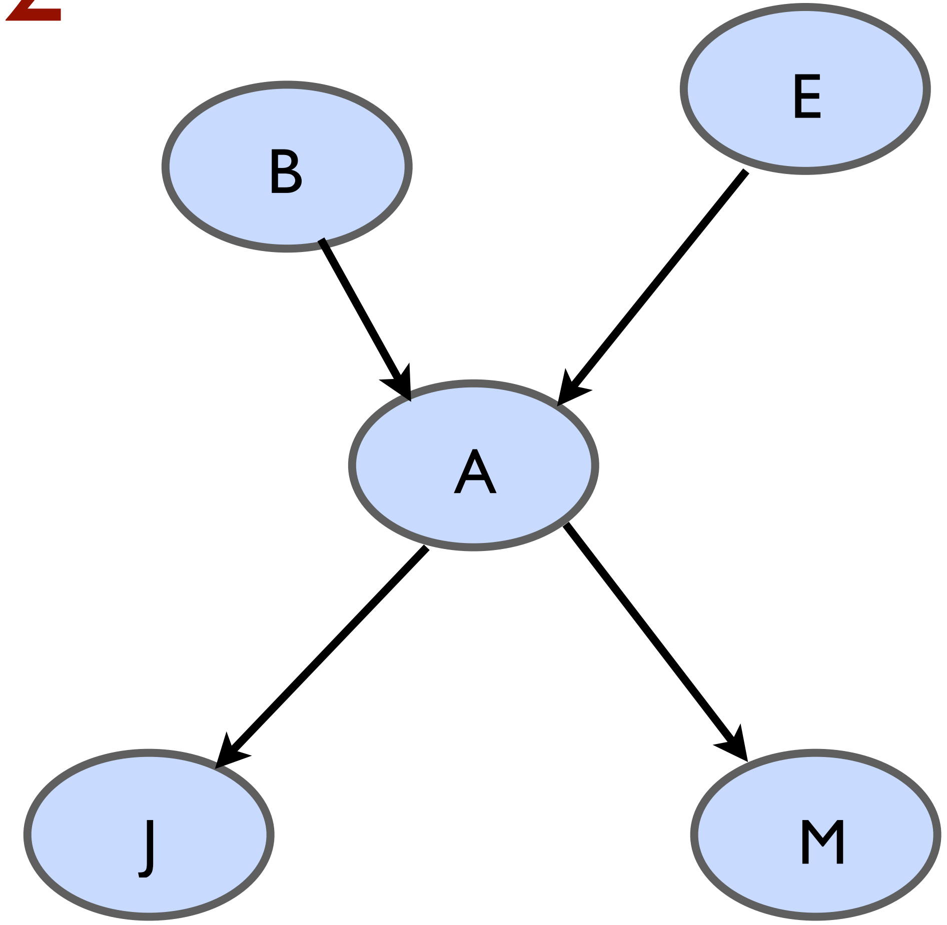
P(E)
0.002

B	E	P(A|B,E)
T	T	0.95
T	F	0.94
F	T	0.29
F	F	0.001

A	P(J|A)
T	0.90
F	0.05

A	P(M|A)
T	0.70
F	0.01

Example 2



A CPT for Boolean X_i with k Boolean parents has 2^k rows for the combinations of parent values

Each row requires one number p for $X_i = true$ (the number for $X_i = false$ is just $1-p$)

If each variable has no more than k parents, the complete network requires $O(n\ 2^k)$ numbers

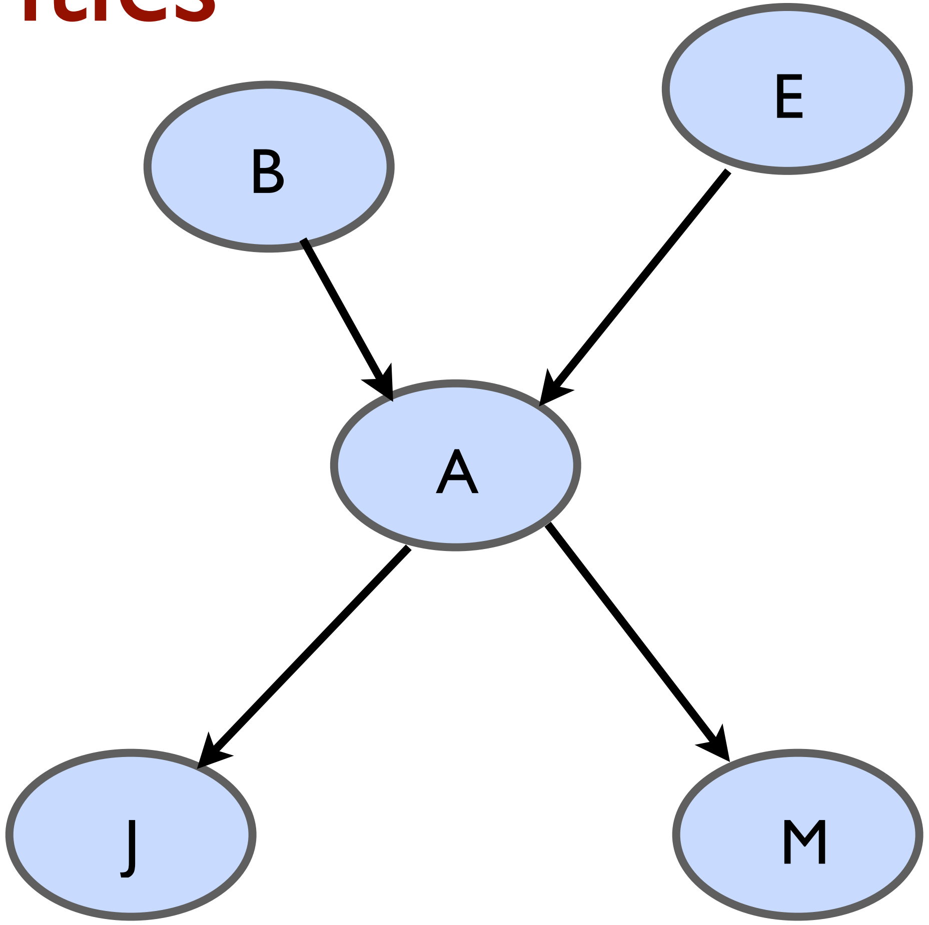
I.e., grows linearly with n, vs. $O(2^n)$ for the full joint distribution

For burglary net, $1 + 1 + 4 + 2 + 2 = 10$ numbers (vs. $2^5 - 1 = 31$)

Outline

- Uncertainty & probability (chapter 13)
  - Uncertainty
  - Probability
  - Syntax and Semantics
  - Inference
  - Independence and Bayes' Rule
- Bayesian Networks (chapter 14.1-3)
  - Syntax
  - Semantics
  - Efficient representation

Global semantics



Global semantics defines the full joint distribution as the product of the local conditional distributions:

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

E.g., $P(j \wedge m \wedge a \wedge \neg b \wedge \neg e)$

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

$$= 0.9 * 0.7 * 0.001 * 0.999 * 0.998$$

$$\approx 0.000628$$

Constructing Bayesian networks

We need a method such that a series of locally testable assertions of conditional independence guarantees the required global semantics.

1. Choose an ordering of variables $X_1, \dots, X_n$

2. For $i = 1$ to n

   add X_i to the network

   select parents from $X_1, \dots, X_{i-1}$ such that

   $\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \dots, X_{i-1})$

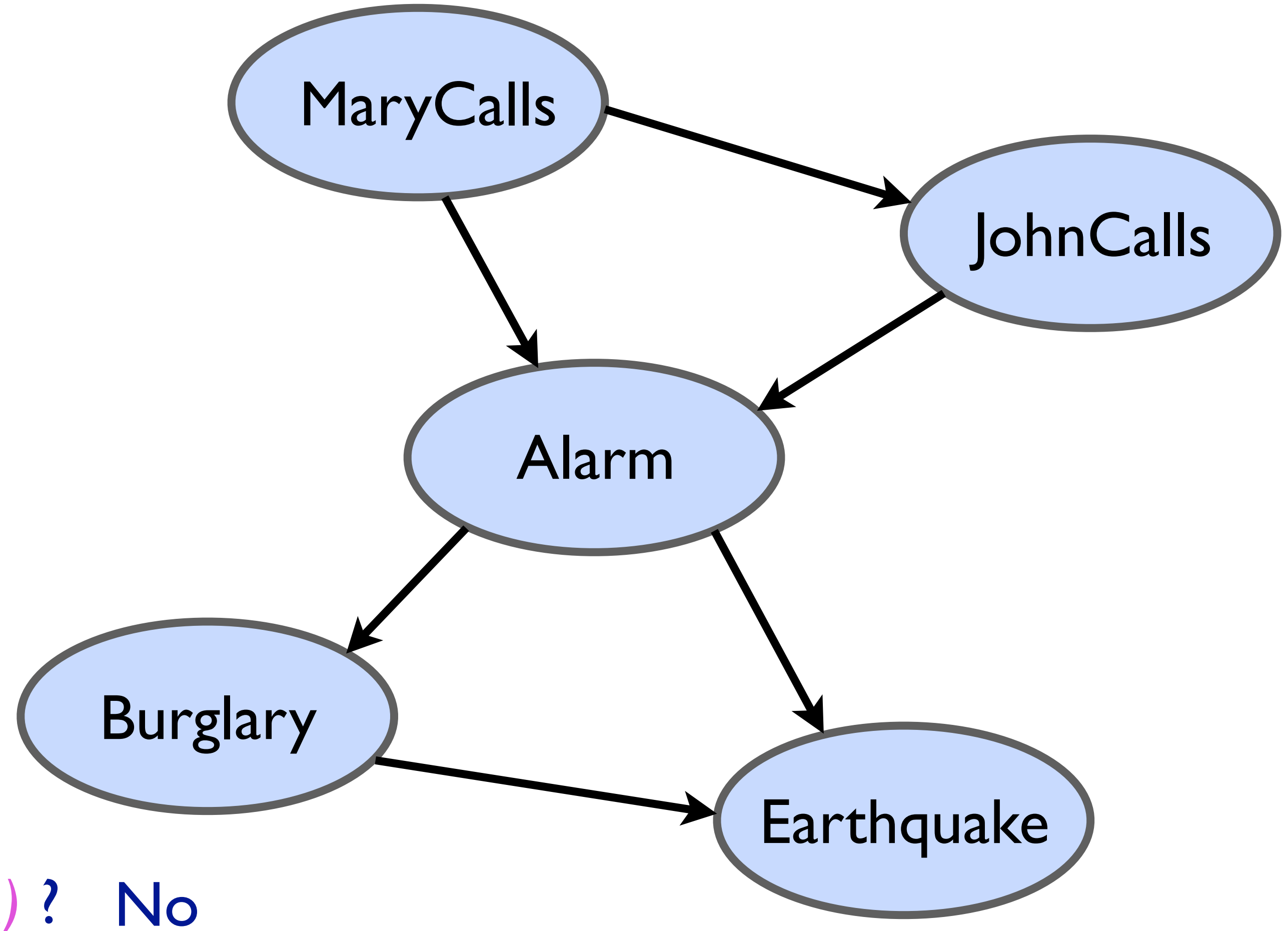
This choice of parents guarantees the global semantics:

$$\mathbf{P}(X_1, \dots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid X_1, \dots, X_{i-1}) \qquad \text{(chain rule)}$$

$$= \prod_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i)) \qquad \text{(by construction)}$$

Construction example

Suppose we choose the ordering M, J, A, B, E



$P(J \mid M) = P(J)$? No

$P(A \mid J, M) = P(A \mid J)$? $P(A \mid J, M) = P(A)$? No

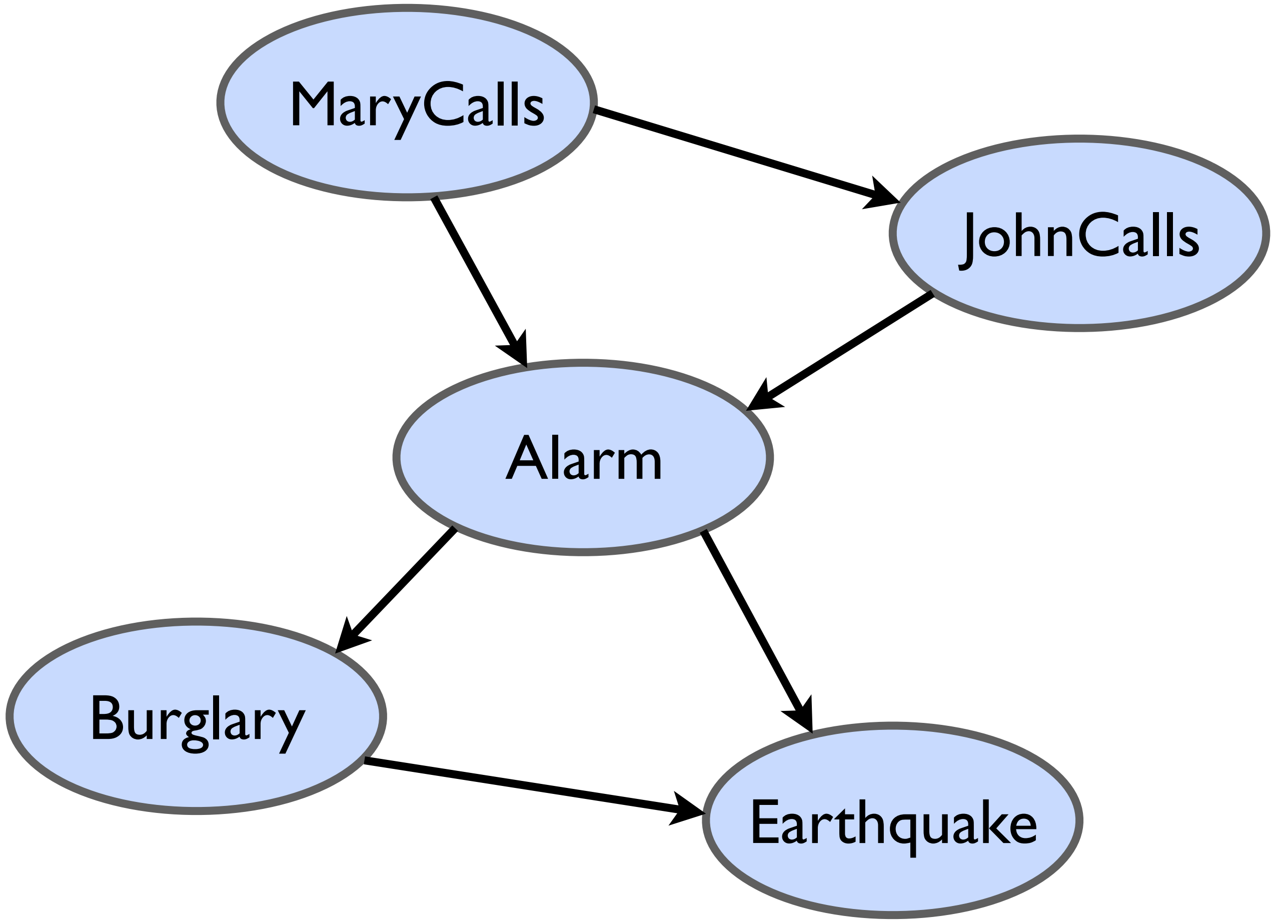
$P(B \mid A, J, M) = P(B \mid A)$? Yes

$P(B \mid A, J, M) = P(B)$? No

$P(E \mid B, A, J, M) = P(E \mid A)$? No

$P(E \mid B, A, J, M) = P(E \mid A, B)$? Yes

Construction example



Deciding conditional independence is hard in noncausal directions

(Causal models and conditional independence seem hardwired for humans!)

Assessing conditional probabilities is hard in noncausal directions

Network is less compact: *1 + 2 + 4 +2 +4 = 13* numbers

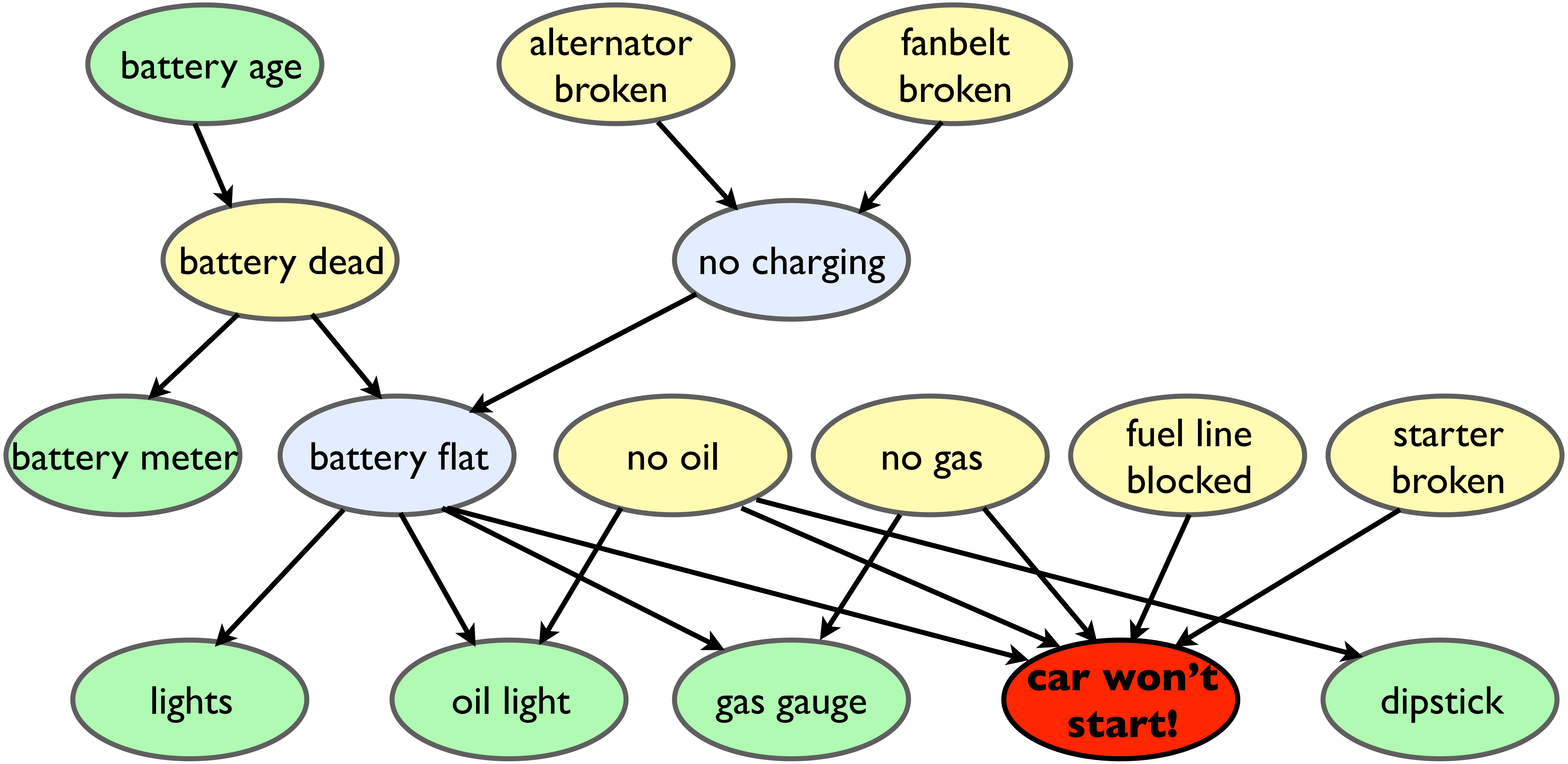
Hence: Choose preferably an order corresponding to the cause → effect "chain"

Locally structured (sparse): Car diagnosis

Initial evidence: The *** car won't start!

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

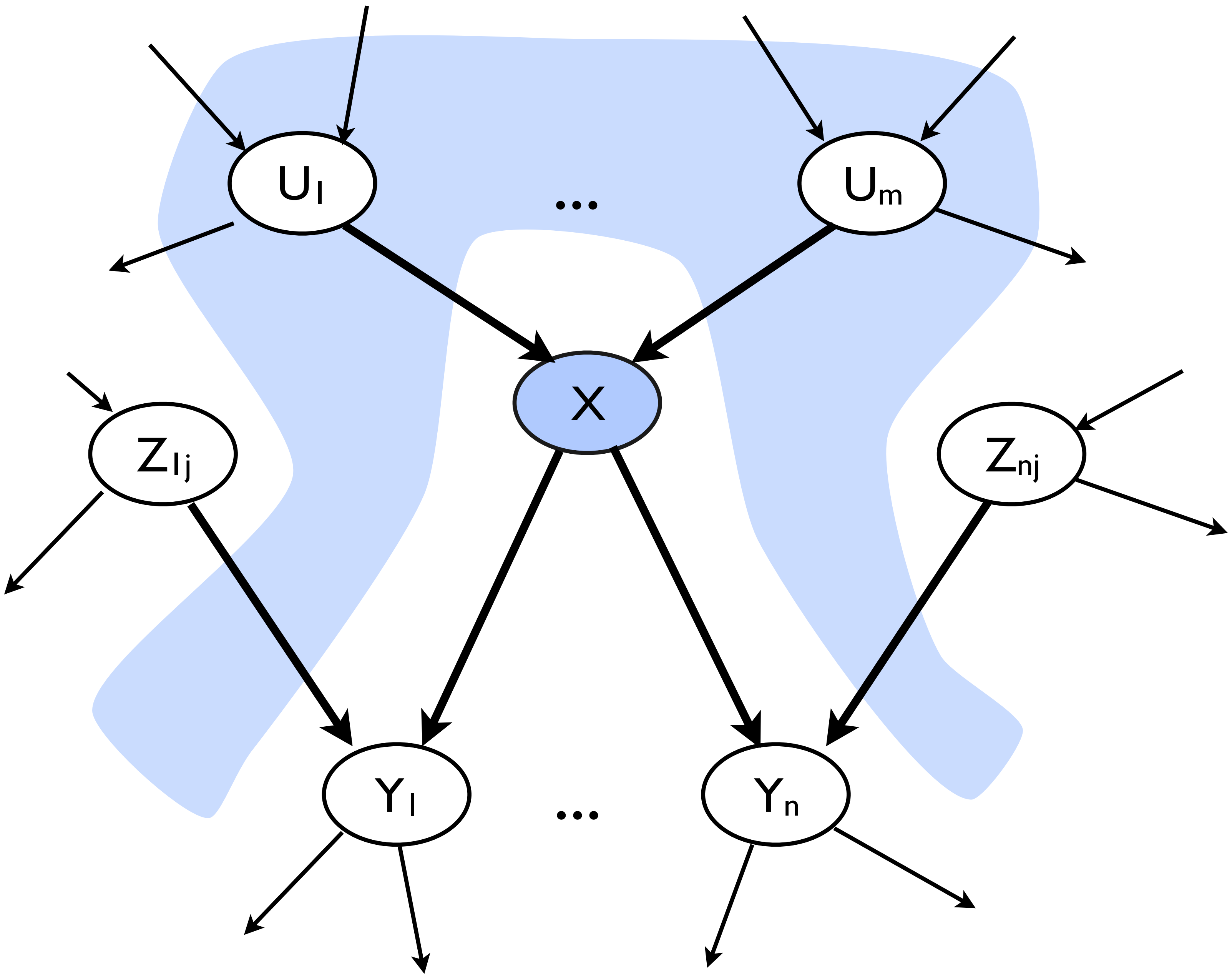
Hidden variables (blue) ensure sparse structure / reduce parameters



X

Local semantics

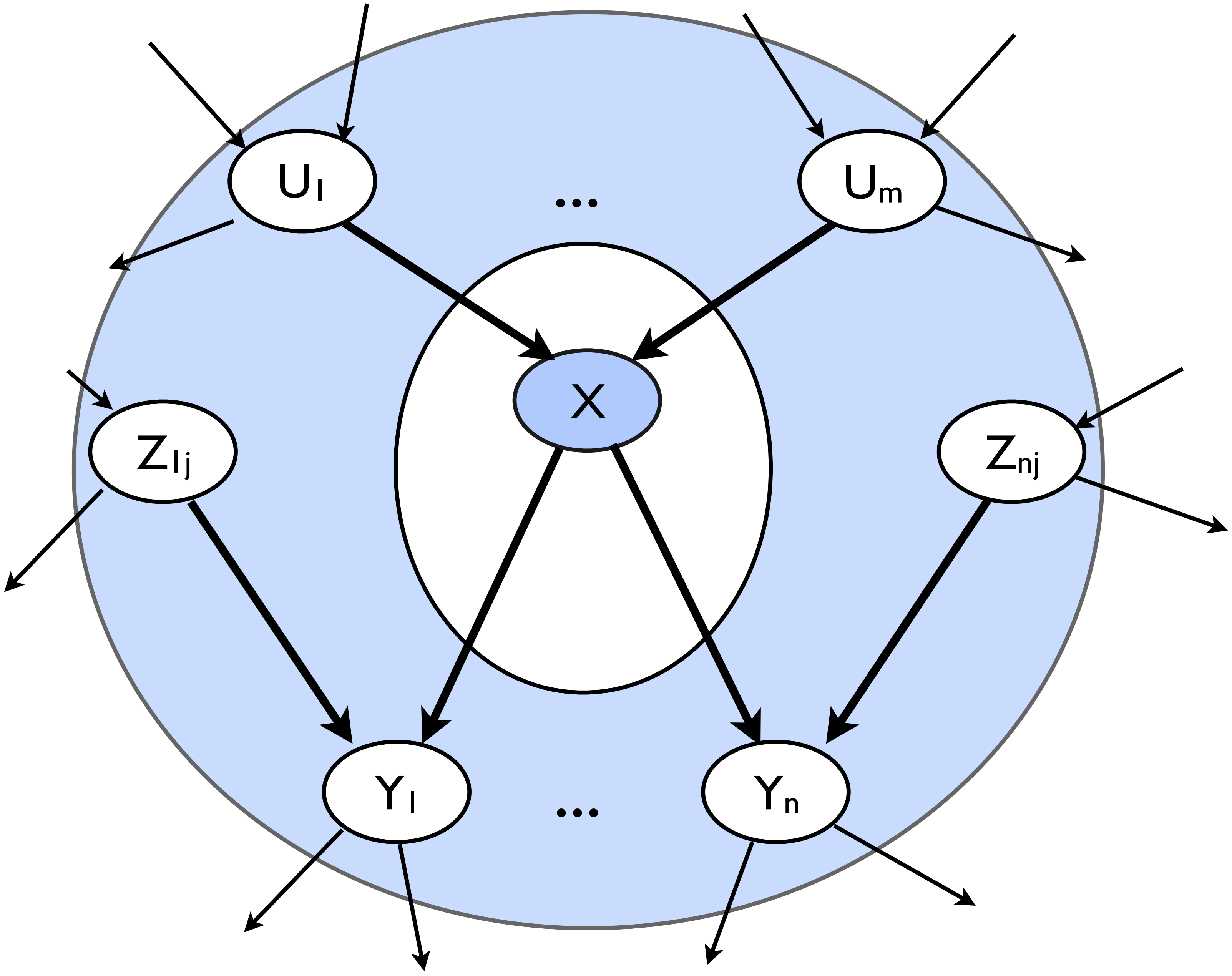
Local semantics: each node is conditionally independent of its non-descendants given its parents



Markov blanket

Each node is conditionally independent of all others given its

Markov blanket: parents + children + children's parents



Outline

- Uncertainty & probability (chapter 13)
  - Uncertainty
  - Probability
  - Syntax and Semantics
  - Inference
  - Independence and Bayes' Rule
- Bayesian Networks (chapter 14.1-3)
  - Syntax
  - Semantics
  - Efficient representation

Compact conditional distributions

CPT grows exponentially with numbers of parents (i.e., causes to the effect)

CPT becomes infinite with continuous-valued parent or child

Solution: *canonical* distributions that are defined compactly

Deterministic nodes are the simplest case:

$X = f(\mathit{Parents}(X))$ for some function f

E.g., Boolean functions

$\mathit{NorthAmerican} \Leftrightarrow \mathit{Canadian} \vee \mathit{US} \vee \mathit{Mexican}$

E.g., numerical relationships among continuous variables

$$\frac{\delta \mathit{Level}}{\delta t} = \mathit{inflow} + \mathit{precipitation} - \mathit{outflow} - \mathit{evaporation}$$

Compact conditional distributions (2)

Noisy-OR distributions model multiple noninteracting causes

1) Parents $U_1 \ldots U_k$ include all causes (add *leak node* for "miscellaneous" ones)

2) Independent failure probability q_i for each cause alone

$$\Rightarrow P(X \mid U_1, \ldots , U_j, \neg U_{j+1}, \ldots , \neg U_k) = 1 - \prod_{i=1}^{j} q_i$$

Cold	*Flu*	*Malaria*	*P(Fever)*	*P(¬Fever)*
F	F	F	***0.0***	*1.0*
F	F	T	*0.9*	***0.1***
F	T	F	*0.8*	***0.2***
F	T	T	*0.98*	*0.02 = 0.2 * 0.1*
T	F	F	*0.4*	***0.6***
T	F	T	*0.94*	*0.06 = 0.6 * 0.1*
T	T	F	*0.88*	*0.12 = 0.6 * 0.2*
T	T	T	*0.988*	*0.012 = 0.6 * 0.2 * 0.1*

Number of parameters linear in number of parents

Summary

Bayesian networks provide a natural representation for (causally induced) conditional independence

Topology + CPTs = compact representation of joint distribution

Generally easy for (non)experts to construct

Canonical distributions (e.g., noisy-OR) = compact representation of CPTs

Continuous variables ⇒ parameterised distributions (e.g., linear Gaussians)