

Probabilistic representation

Applied artificial intelligence (EDAI32)

Lecture 03

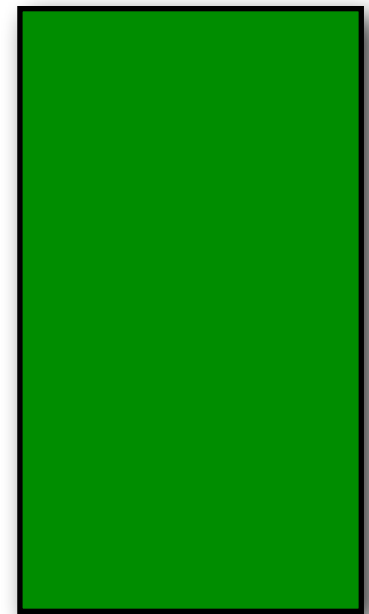
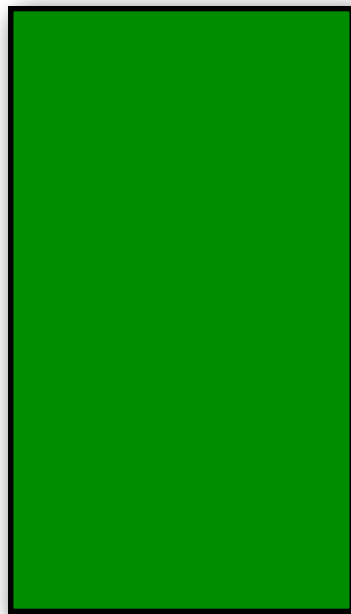
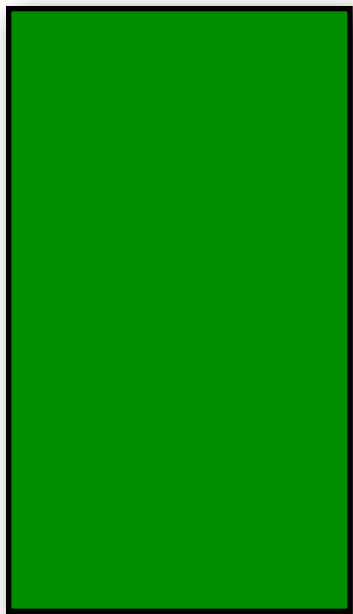
2016-01-26

Elin A. Topp

Material based on course book, chapter 13, 14.1-3

Show time!

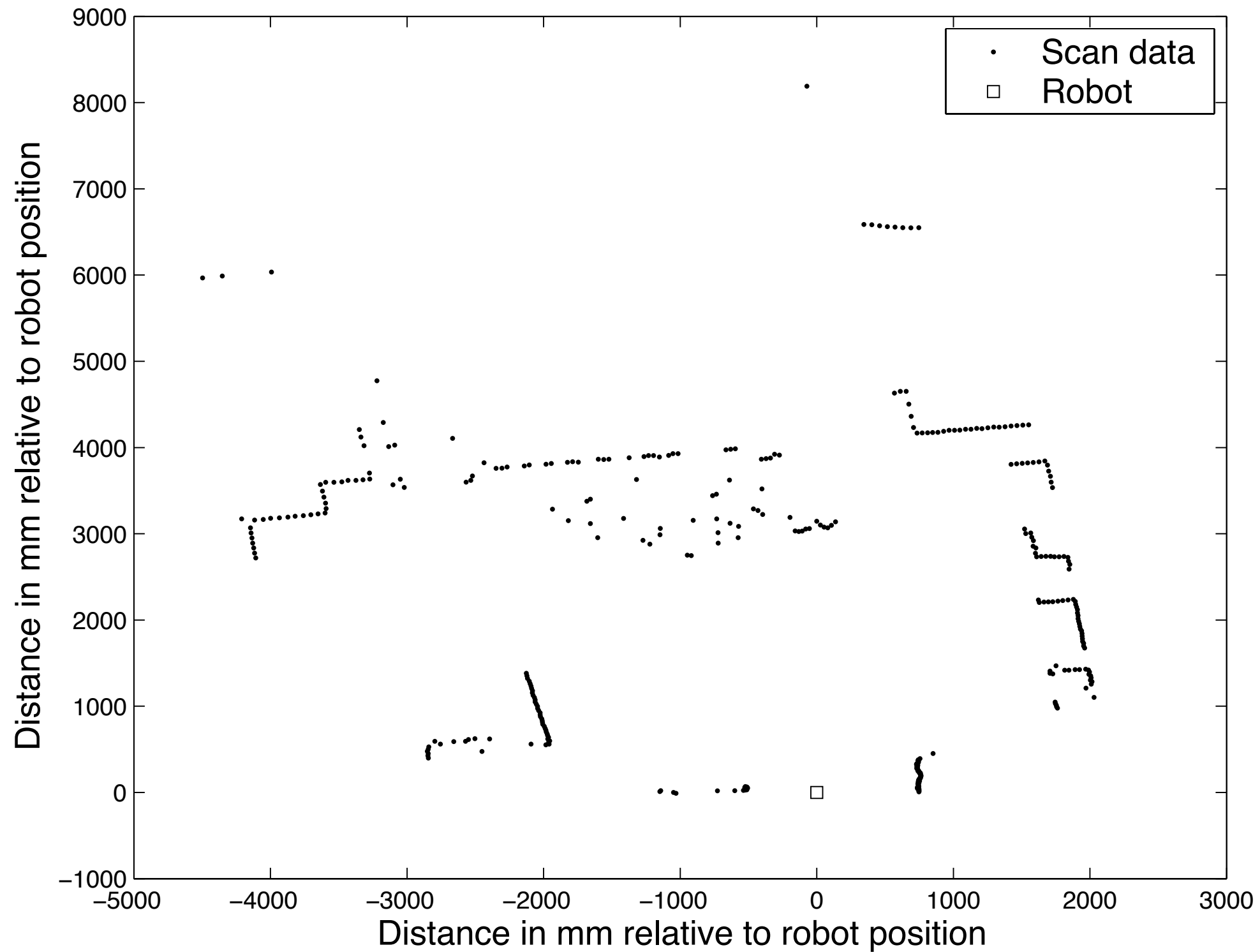
Two boxes of chocolates, one luxury car.
Where is the car?



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

Outline

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

Uncertainty and decisions

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 modelling and predicting traffic

Use rules / conditions? 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,...)

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

Probability basics

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

$\omega \in \Omega$ a sample point / possible world / atomic event, e.g., the outcome “2”.

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

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

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

An event a is any subset of Ω

$$P(a) = \sum_{\{\omega \in A\}} P(\omega)$$

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

Random variables

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

e.g., when rolling a die and looking for odd numbers,

$$\text{Odd}(n) = \text{true, for } n \in \{1, 3, 5\}$$

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 ω where $A(\omega) = \text{true}$

event $\neg a$ = set of sample points ω where $A(\omega) = \text{false}$

event $a \wedge b$ = points ω where $A(\omega) = \text{true}$ and $B(\omega) = \text{true}$

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.

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

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

$$\text{e.g., } P(\text{Odd} = \text{true}) = \sum_{\{n: \text{Odd}(n) = \text{true}\}} P(n) = P(1) + P(3) + P(5) = 1/6 + 1/6 + 1/6 = 1/2$$

Bayesian 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., $P(A_{25}) = 0.04$

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

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

Prior probability

Prior or unconditional probabilities of propositions

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

$P(\text{Weather} = \text{sunny}) = 0.72$ (e.g., known from statistics)

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

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

$\mathbb{P}(\text{Weather}) = \langle 0.72, 0.1, 0.08, 0.1 \rangle$

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):

$\mathbb{P}(\text{Weather}, \text{Cavity}) =$ 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., $P(\text{cavity}) = 0.2$ (prior, no evidence for anything), but

$$P(\text{cavity} \mid \text{toothache}) = 0.6$$

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.

$$P(\text{cavity} \mid \text{toothache, sunny}) = P(\text{cavity} \mid \text{toothache})$$

Domain knowledge lets us do this kind of inference.

Posterior probability (2)

Definition of conditional / posterior probability:

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

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

$$P(a \wedge b) = P(a | b) P(b) = P(b | a) P(a)$$

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

$$\mathbb{P}(\text{Weather}, \text{Cavity}) = \mathbb{P}(\text{Weather} | \text{Cavity}) \mathbb{P}(\text{Cavity})$$

(gives a 4x2 set of equations)

Chain rule (successive application of product rule):

$$\begin{aligned} \mathbb{P}(X_1, \dots, X_n) &= \mathbb{P}(X_1, \dots, X_{n-1}) \mathbb{P}(X_n | X_1, \dots, X_{n-1}) \\ &= \mathbb{P}(X_1, \dots, X_{n-2}) \mathbb{P}(X_{n-1} | X_1, \dots, X_{n-2}) \mathbb{P}(X_n | X_1, \dots, X_{n-1}) \\ &= \dots = \prod_{i=1}^n \mathbb{P}(X_i | X_1, \dots, X_{i-1}) \end{aligned}$$

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:
Can also compute posterior probabilities:

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

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

$$P(\text{cavity} \mid \text{toothache}) = \frac{0,108 + 0,012}{0,108 + 0,012 + 0,016 + 0,064} = \frac{0,12}{0,28} = 0,428$$

$$P(\neg \text{cavity} \mid \text{toothache}) = \frac{0,016 + 0,064}{0,108 + 0,012 + 0,016 + 0,064} = \frac{0,08}{0,28} = 0,286$$

Normalisation

	toothache		\neg toothache	
	catch	\neg catch	catch	\neg catch
cavity	0,108	0,012	0,072	0,008
\neg cavity	0,016	0,064	0,144	0,576

Denominator can be viewed as a *normalisation constant*:

$$\begin{aligned}
 \mathbb{P}(Cavity \mid toothache) &= \alpha \mathbb{P}(Cavity, toothache) \\
 &= \alpha [\mathbb{P}(Cavity, toothache, catch) + \mathbb{P}(Cavity, toothache, \neg catch)] \\
 &= \alpha [\langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle] \\
 &= \alpha \langle 0.12, 0.08 \rangle = \langle 0.6, 0.4 \rangle
 \end{aligned}$$

And the good news:

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

Inference gone bad

A young student ends her own life. In the following attempt to understand the cause, a personal journal is found. In this 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 writing 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:

$P(\text{headache} \mid \text{PTSD}) = \text{high}$ (statistics)

$P(\text{PTSD} \mid \text{abuse in childhood}) = \text{high}$ (statistics)

ok, yes, sure, but:

You did not consider the relevant relations of

$P(\text{PTSD} \mid \text{headache})$ or

$P(\text{abuse in childhood} \mid \text{PTSD})$,

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

Bayes' Rule

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

$$\Rightarrow \text{Bayes' Rule } P(a | b) = \frac{P(b | a) P(a)}{P(b)}$$

or in distribution form:

$$\mathbb{P}(Y | X) = \frac{\mathbb{P}(X | Y) P(Y)}{\mathbb{P}(X)} = \alpha P(X | Y) P(Y)$$

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

$$P(\text{Cause} | \text{Effect}) = \frac{P(\text{Effect} | \text{Cause}) P(\text{Cause})}{P(\text{Effect})}$$

E.g., with *M* “meningitis”, *S* “stiff neck”:

$$P(m | s) = \frac{P(s | m) P(m)}{P(s)} = \frac{0.7 * 0.00002}{0.01} = 0.0014 \quad (\text{not too bad, really!})$$

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

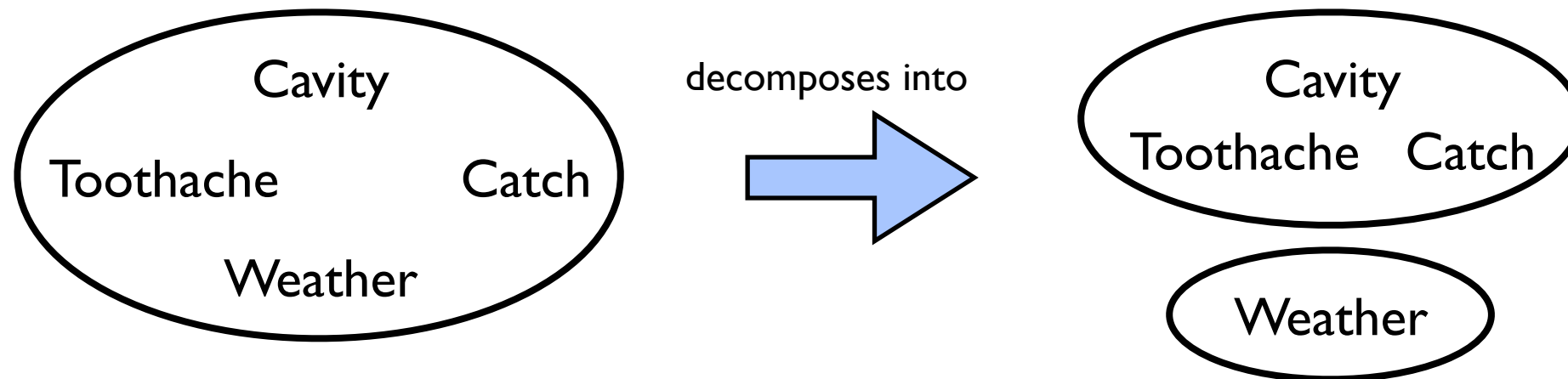
n Boolean variables give us an input table of size $O(2^n)$...

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

Independence

A and B are *independent* iff

$$P(A | B) = P(A) \quad \text{or} \quad P(B | A) = P(B) \quad \text{or} \quad P(A, B) = P(A) P(B)$$



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

32 entries reduced to 8 + 4 (Weather is not Boolean!).

This absolute (*unconditional*) 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

$\mathbb{P}(\textit{Toothache}, \textit{Cavity}, \textit{Catch})$ 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:

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

The same holds when there is no cavity:

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

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

$$\mathbb{P}(\textit{Catch} \mid \textit{Toothache}, \textit{Cavity}) = \mathbb{P}(\textit{Catch} \mid \textit{Cavity})$$

Writing out full joint distribution using chain rule:

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

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

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

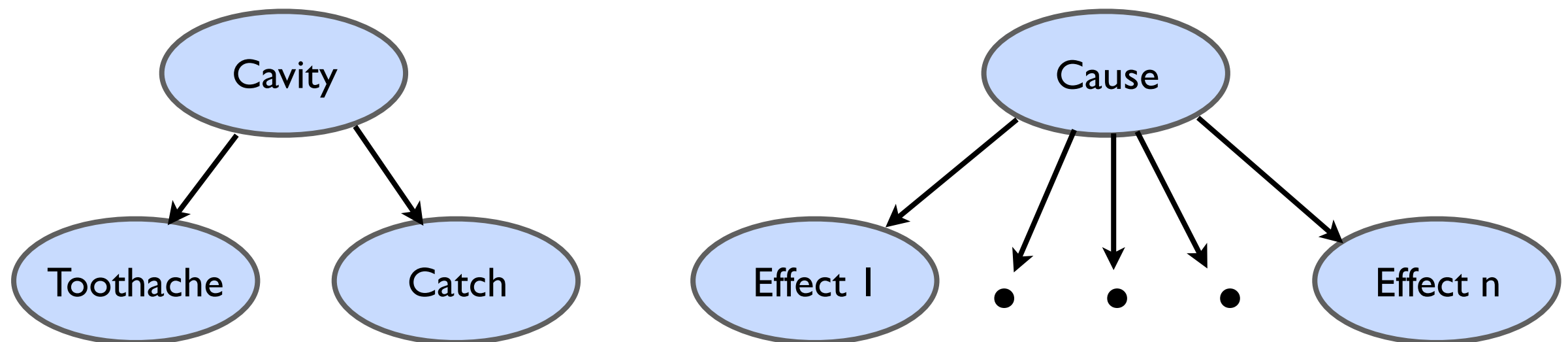
- Uncertainty & probability (chapter 13)
 - Uncertainty
 - Probability
 - Syntax and Semantics
 - Inference
 - Independence and Bayes' Rule
- Bayesian Networks (chapter 14.1-3)
 - Syntax
 - Semantics
 - Efficient representation

Bayes' Rule and conditional independence

$$\begin{aligned} & \mathbb{P}(Cavity \mid toothache \wedge catch) \\ &= \alpha \mathbb{P}(toothache \wedge catch \mid Cavity) \mathbb{P}(Cavity) \\ &= \alpha \mathbb{P}(toothache \mid Cavity) \mathbb{P}(catch \mid Cavity) \mathbb{P}(Cavity) \end{aligned}$$

An example of a *naive Bayes* model:

$$\mathbb{P}(Cause, Effect_1, \dots, Effect_n) = \mathbb{P}(Cause) \prod_i \mathbb{P}(Effect_i \mid Cause)$$



The total number of parameters is *linear* in n

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 \approx “directly influences”)

- a conditional distribution for each node given its parents:

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

In the simplest case, conditional distribution represented as a

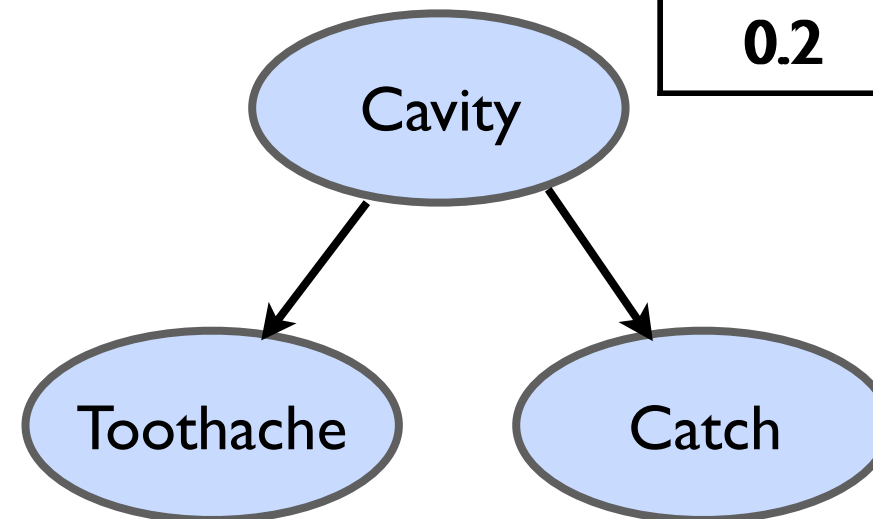
conditional probability table (CPT)

giving the distribution over X_i for each combination of parent values

Example

Topology of network encodes conditional independence assertions:

$P(W=sun)$	$P(W=sun)$	$P(W=rain)$	$P(W=rain)$	$P(W=cloud)$	$P(W=cloud)$	$P(W=snow)$
0.72	0.72	0.1	0.1	0.08	0.08	0.1



$P(Cav)$	$P(\neg Cav)$
0.2	0.8

Cav	$P(T Cav)$	$P(\neg T Cav)$
T	0.6	0.4
F	0.7	0.3

Cav	$P(C Cav)$	$P(\neg C Cav)$
TT	0.9	0.1
FF	0.2	0.8

Weather is (unconditionally, absolutely) independent of the other variables

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

We can skip the dependent columns in the tables to reduce complexity!

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:

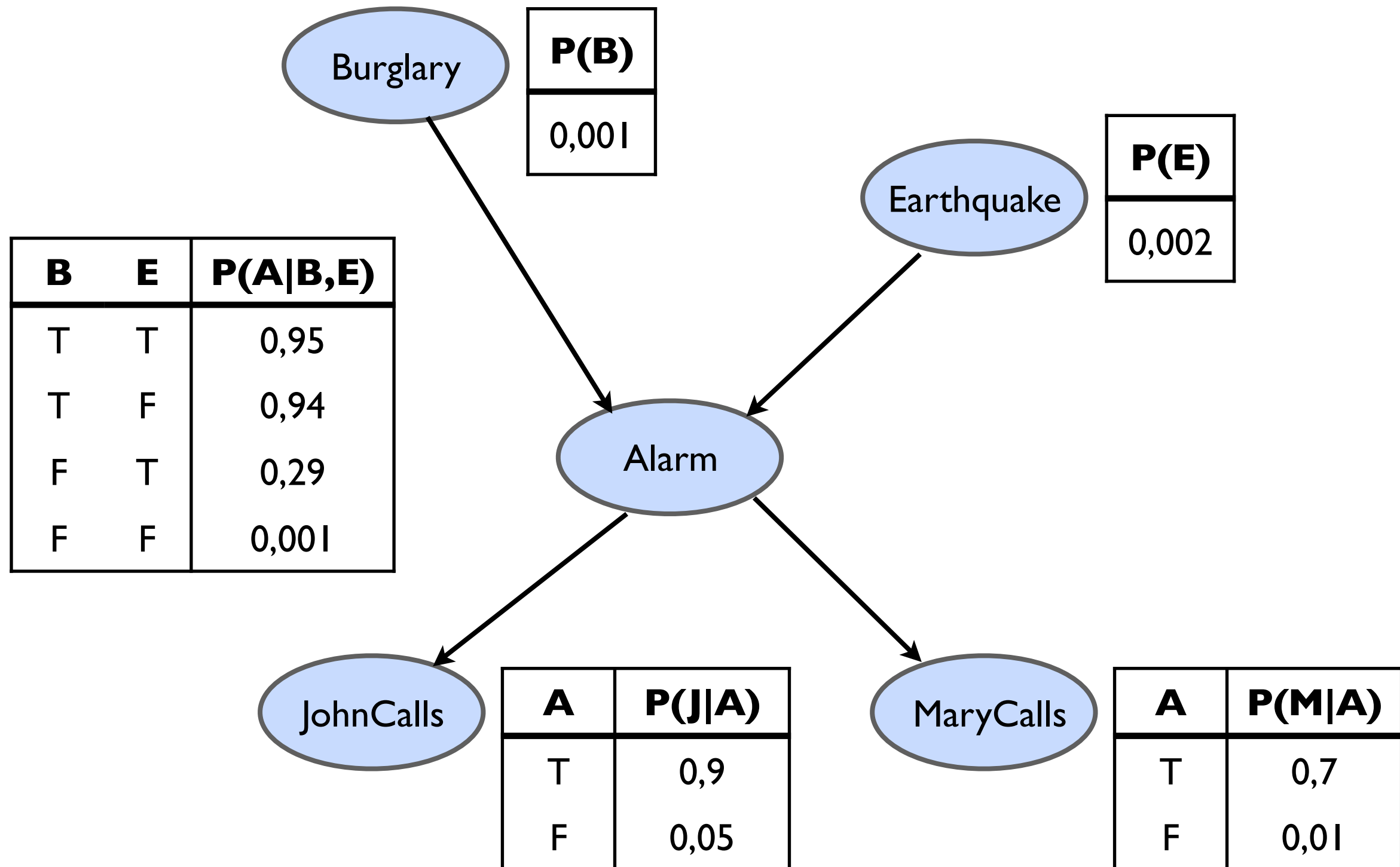
- A burglar can set the alarm off

- An earthquake can set the alarm off

- The alarm can cause John to call

- The alarm can cause Mary to call

Example 2 (2)



Global semantics

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

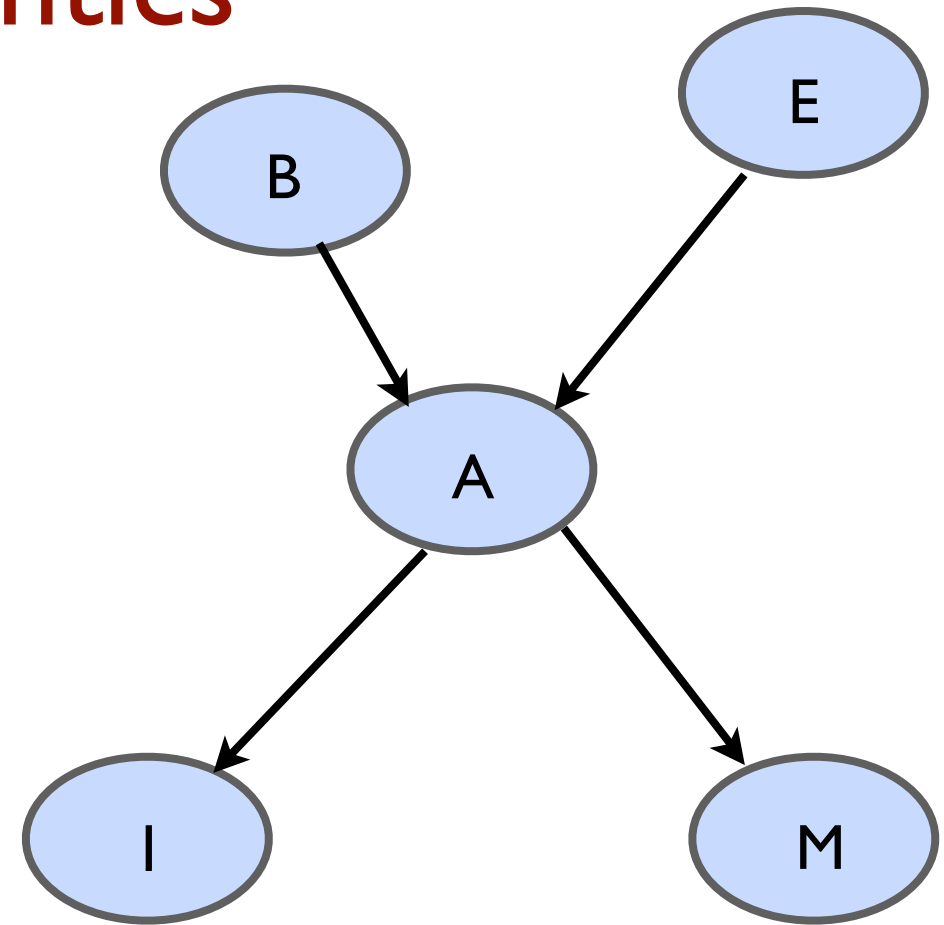
$$P(x_1, \dots, x_n) = \prod_{i=1}^n P(x_i \mid \text{parents}(X_i))$$

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

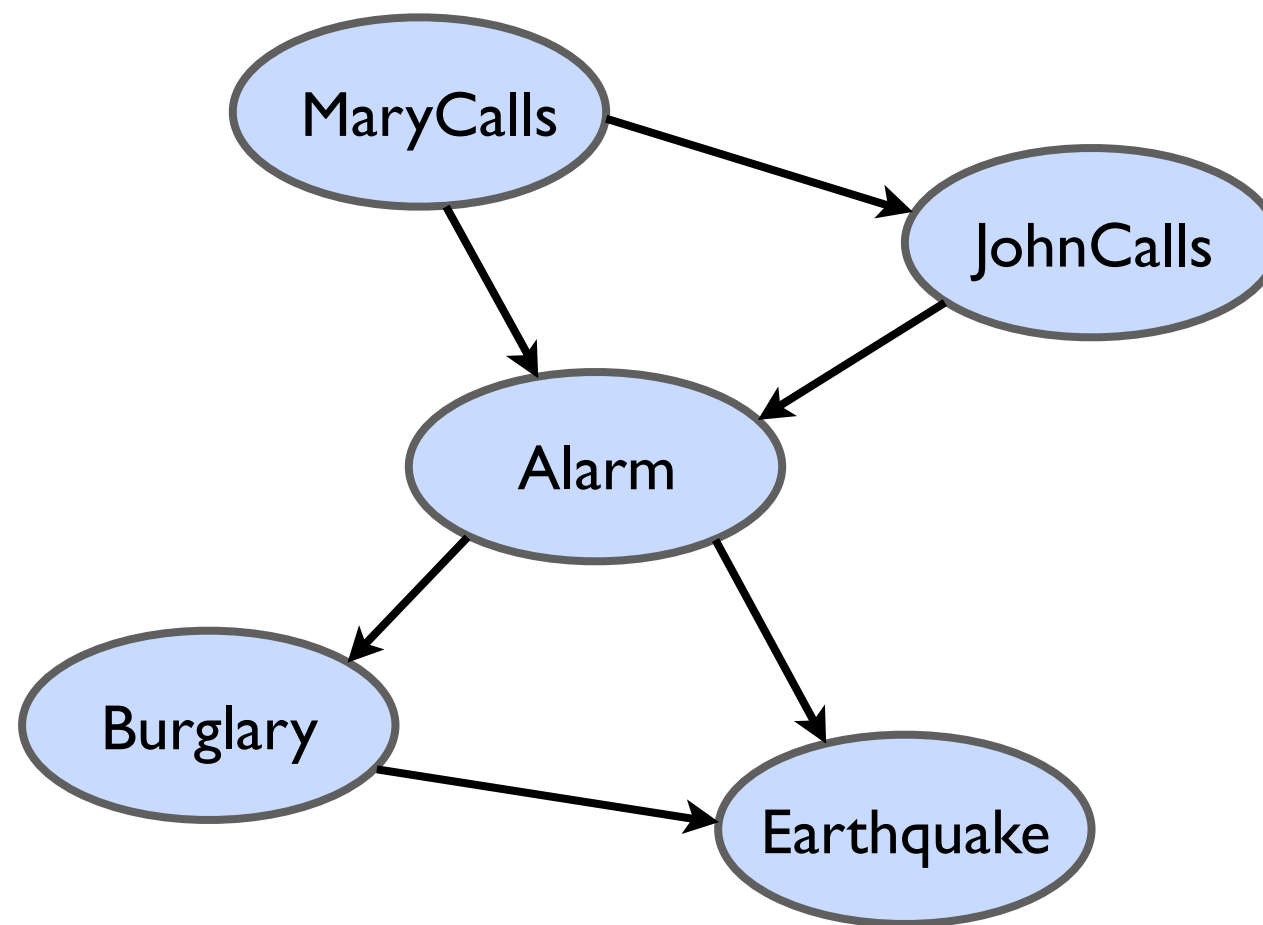
$$P(X_i \mid \text{Parents}(X_i)) = P(X_i \mid X_1, \dots, X_{i-1})$$

This choice of parents guarantees the global semantics:

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

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

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

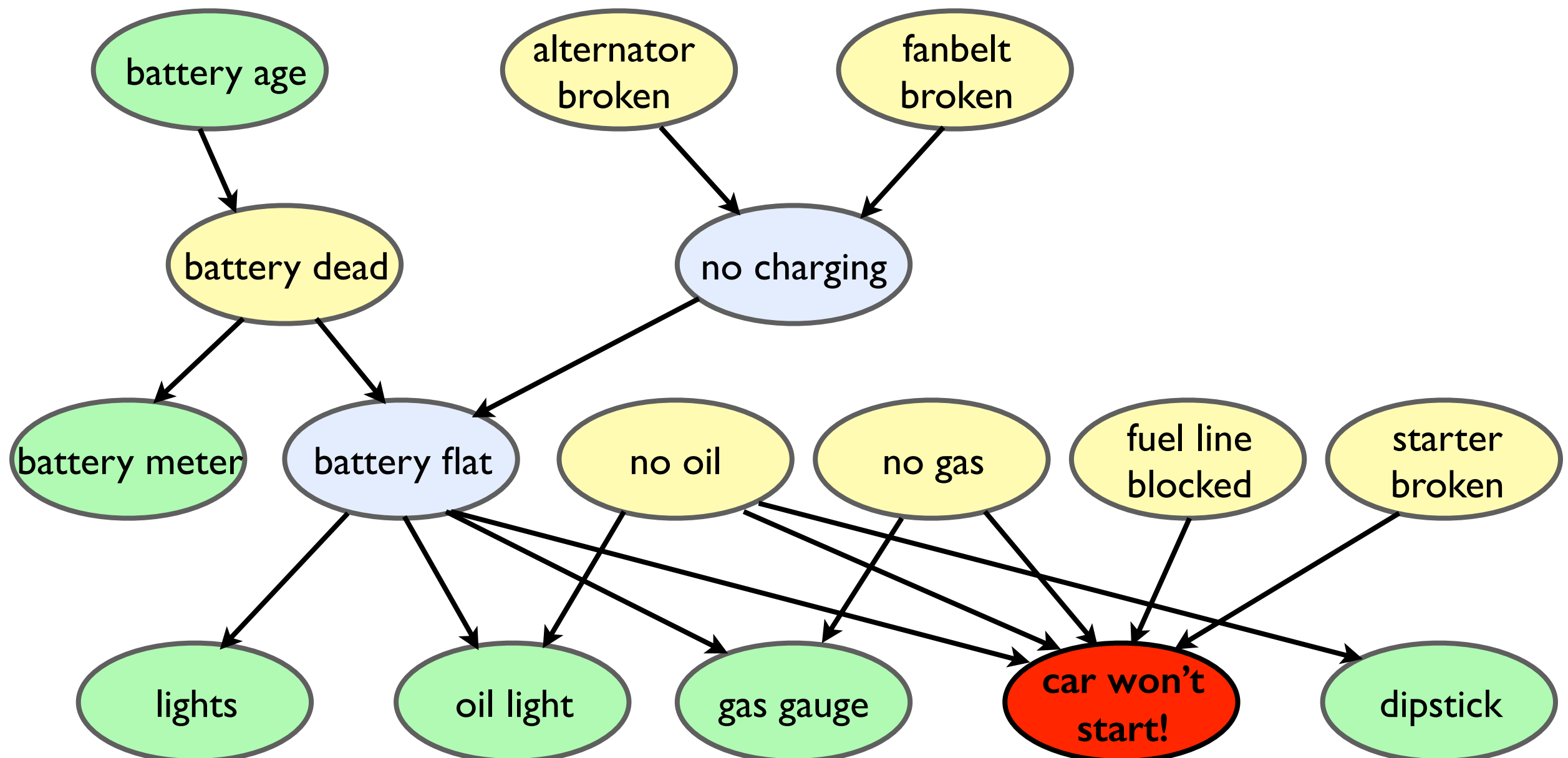
Hence: Choose preferably an order corresponding to the cause \rightarrow effect “chain”

Locally structured (sparse) network

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

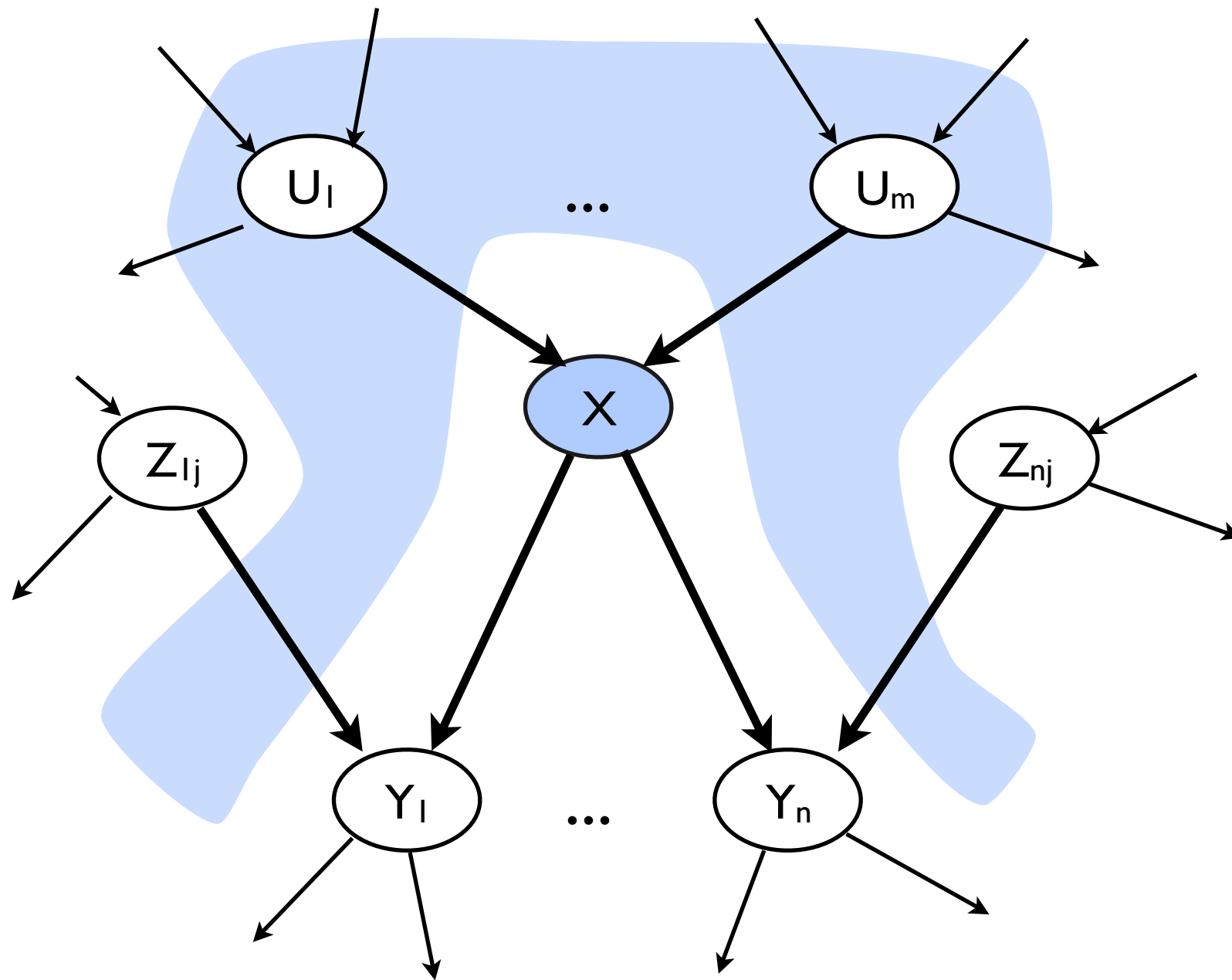
Testable variables (green), “broken, so fix it” variables (yellow)

Hidden variables (blue) ensure sparse structure / reduce parameters



Local semantics

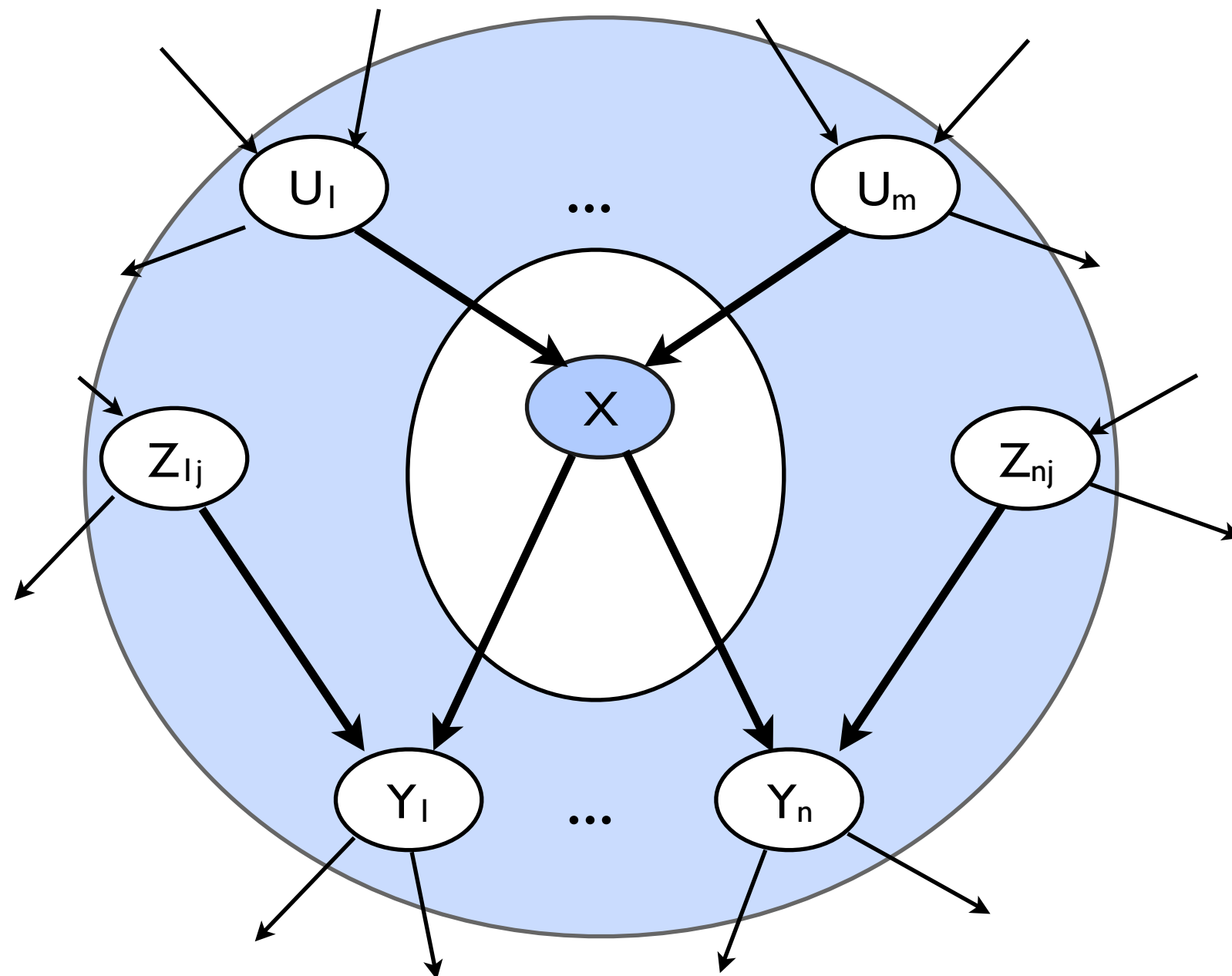
Local semantics: each node is conditionally independent of its non-descendants given its parents



Markov blanket

Each node is conditionally independent of all others given its

Markov blanket: parents + children + children's parents



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

And going further:

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