Probabilistic representation and reasoning

Artificial intelligence (EDAP01)
Lecture 07
2020-02-10
Elin A. Topp

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

Rethinking AIMA

From Stuart Russell's talk at AAAI 2020:

We need systems that have an incentive to be switched off, and to get there, we need to make the systems uncertain about the utility (actual objective) of their action for the human that formulates the task (objective) for the system.

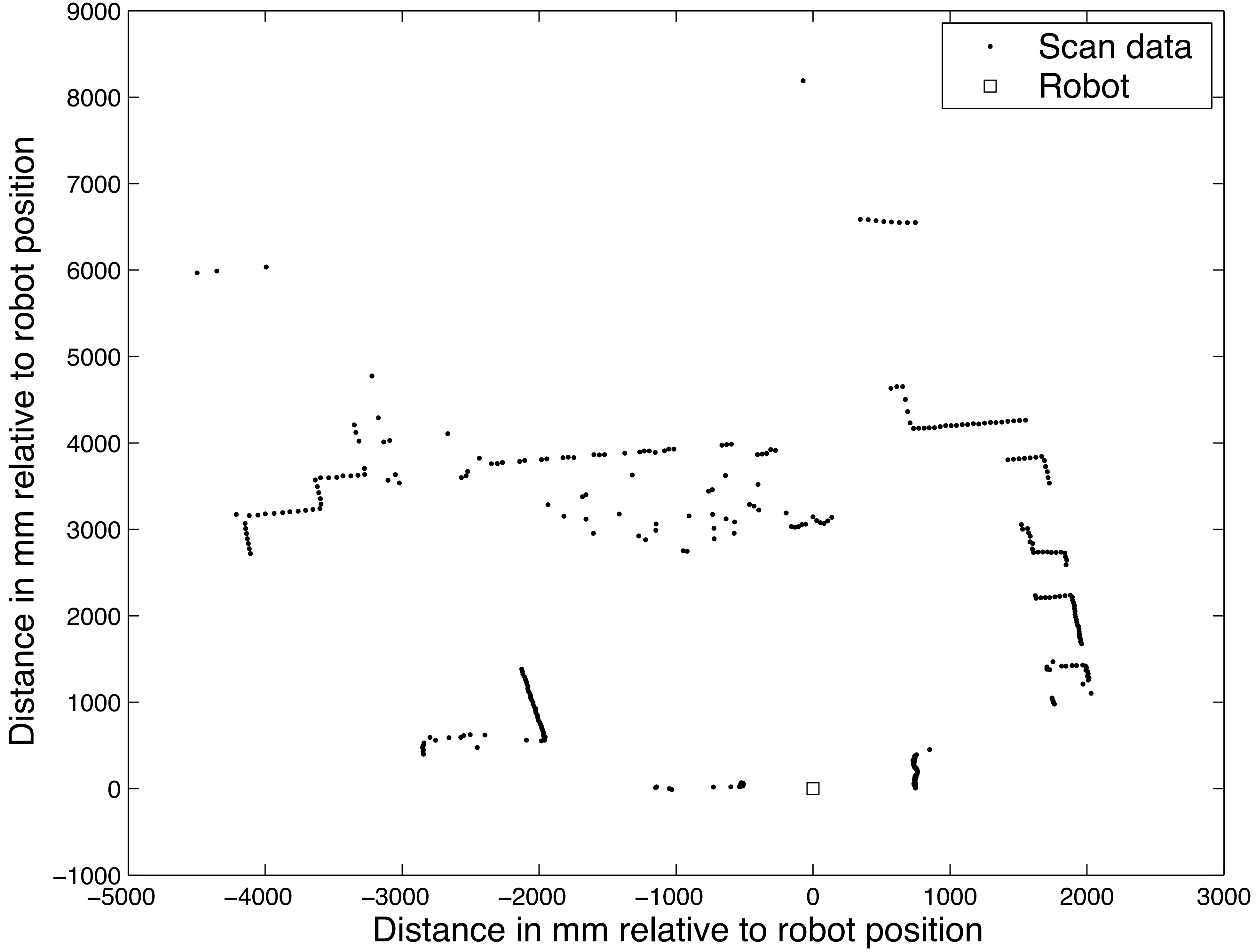
Hence: We need to be able to *model uncertainty* and to *reason under uncertainty*

(Stuart Russell's talk from 28:40 onward, explaining the issues with the book)

Link to streamed talk(s): https://aaai.org/Conferences/AAAI-20/livestreamed-talks/

Recommended reading: The Off-Switch Game (D. Hadfield-Menell, A. Dragan, P. Abbeel, S. Russell)
https://people.eecs.berkeley.edu/~dhm/papers/off_switch_AAAI_ws.pdf

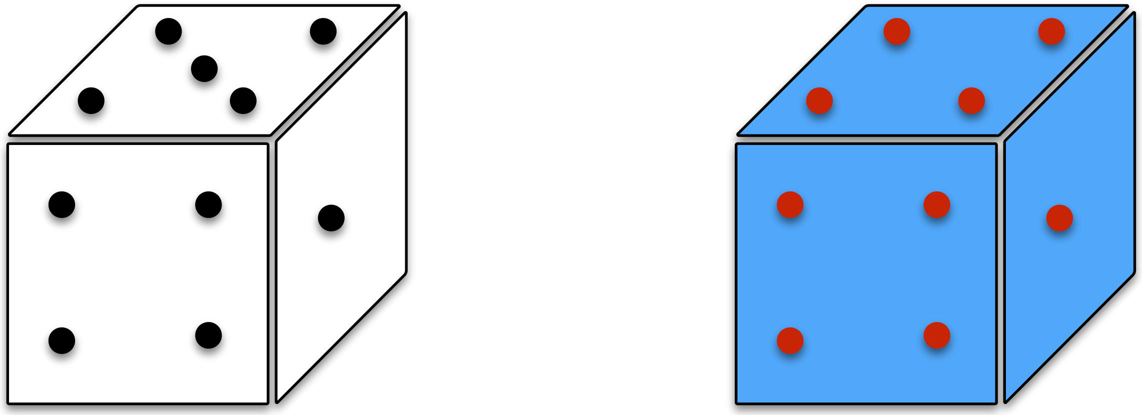
A robot’s view of the world...



What category of “thing” is shown to me?

Object? Workspace? Room? Link to room?
Can we reason about behavioural features and what is causing them?

Drawing and rolling dice - what will I get?



Assume different types of dice, that might have different colours and distribution of values, and even differently coloured dots.

Outline

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

Outline

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

Bayesian Probability

Subjective or Bayesian probability:

Probabilities relate propositions to one's state of knowledge (*A* = "*the observed pattern in the data was caused by a person*")

e.g., *P(A) = 0.2*

e.g., *P(A | there is a ton of "leggy" furniture in the respective room) = 0.1*

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 | ton of furniture, dataset obtained at 7:30 by a bot) = 0.05*

Recommended Reading / Listening: Daniel Kahnemann "Thinking, Fast and Slow"
(good motivation as to why we need the Bayesian view point to understand the world, ch. 15+16)

Notation

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,

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

A *proposition* a describes the *event(s)* for which a variable X takes a specific value, e.g., TRUE

Probability *P* induces a *probability distribution* for any random variable *X* with n possible values:

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

the sum of all probabilities of the atomic events that give X the value x_i

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

Notation 2

Here, we express propositions as the variables taking on certain values directly

We look then for example at

$P(X = x_i), i = 1, \dots n$, for all n values x_i of the Variable X

Thus: $P(X = x_1) = P(X = x_2) = 1/2$

with e.g., x_1 = "dice roll outcome is odd number" and x_2 = "dice roll outcome is even number"

For the *distribution* over the possible values of X we get then:

$$\mathbb{P}(X) = < P(X = x_1), P(X = x_2), \dots, P(X = x_n) >$$

Notation 3

Conditional (posterior) probability:

$P(X = x | Y = y)$ *expresses: "Probability for X being x GIVEN that I know that Y = y"*

Joint (conditional) distributions:

We iterate over (a subset of) the values for the random variables in a computation of a joint distribution, e.g.

$$\mathbf{P}(X,Y) \quad = \quad \mathbf{P}(X | Y)\ \mathbf{P}(Y)$$

describes a set of equations (essentially multiplication of all possible combinations), expressing the joint probability distribution of X and Y as conditional probability distribution of X in dependency of the possible (or specifically given) values of Y with

$< P(X = x_1 | Y = y_1) P(Y = y_1), \ldots, P(X = x_n | Y = y_1) P(Y = y_1)$

$\ldots$

$P(X = x_1 | Y = y_m) P(Y = y_m), \ldots, P(X = x_n | Y = y_m) P(Y = y_m) >$

Prior probability

Prior or *unconditional probabilities* of propositions

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

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

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

Question: What does that mean for a bunch of dice, say 15, which all are of "standard type", but where 10 of them are blue and 5 of them are white?

Joint probabilities

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):

$\mathbf{P}$*(Weather, Person)* = a *4 x 2* matrix of values:

Weather Person	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(person) = 0.2* (prior, no evidence for anything), but

P(person | leg-size) = 0.6

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

OBS: NOT "if leg-size, then 60% chance of person"

THINK *"given that leg-size 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(person | leg-size, sunny) = P(person | leg-size)

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 <u>and</u> b being true, we need b true <u>and</u> 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.):

$$\mathbb{P}(Weather, Person) \quad = \quad \mathbb{P}(Weather \mid Person)\ \mathbb{P}(Person)$$

(a *4x2* set of equations, governed by the chosen (given) value for Person from the array over possible values in each equation)

Question: What does that mean for a bunch of dice, say 15, which all are of "standard type", but where 10 of them are blue and 5 of them are white, with 6 of the blue dice and 4 of the white ones having red dots, and the rest have black dots?

What happens now, if we let 5 of the blue dice have a skewed distribution of values, with each of these special dice only having the values 1, 2, 3, each of them twice?

Chain rule

The Chain rule (successive application of product rule) helps us to disentangle the computation of the joint probability over a larger set of variables:

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

Inference

Probabilistic inference:

Computation of posterior probabilities given observed evidence

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

Inference by enumeration

	leg-size		¬ leg-size	
	curved	¬ curved	curved	¬ curved
person	0.108	0.012	0.072	0.008
¬ person	0.016	0.064	0.144	0.576

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

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

$$P(\text{person} \lor \text{leg-size}) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$

$$P(\text{leg-size}) = 0.108 + 0.012 + 0.016 + 0.064 = 0.2$$

Can also compute posterior probabilities:

$$P(\neg\text{person} \mid \text{leg-size}) = \frac{P(\neg\text{person} \land \text{leg-size})}{P(\text{leg-size})}$$

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

Normalisation

	leg-size		¬ leg-size	
	curved	¬curved	curved	¬ curved
person	0.108	0.012	0.072	0.008
¬ person	0.016	0.064	0.144	0.576

Denominator can be viewed as a *normalisation factor*:

$\mathbb{P}(\mathit{Person} \mid \mathit{leg\text{-}size}) = \alpha\, \mathbb{P}(\mathit{Person}, \mathit{leg\text{-}size})$

$= \alpha\, [\mathbb{P}(\mathit{Person}, \mathit{leg\text{-}size}, \mathit{curved}) + \mathbb{P}(\mathit{Person}, \mathit{leg\text{-}size}, \neg \mathit{curved})]$

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

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

And the good news:

We can compute $\mathbb{P}(\mathit{Person} \mid \mathit{leg\text{-}size})$ without knowing the value of $P(\mathit{leg\text{-}size})$!

Inference gone bad

A young student suffers from depression. In her 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 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(headache | PTSD) = high (statistics)

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

ok, yes, sure, but:

Court folks did not consider the relevant relations of

P(PTSD | headache) or

P(abuse in childhood | PTSD),

i.e., they mixed up cause and effect in their argumentation!

Bayes' Rule

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

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

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":

$$P(m \mid s) = \frac{P(s \mid m)\, P(m)}{P(s)} = \frac{0.7 * 0.00002}{0.01} = 0.0014$$ (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

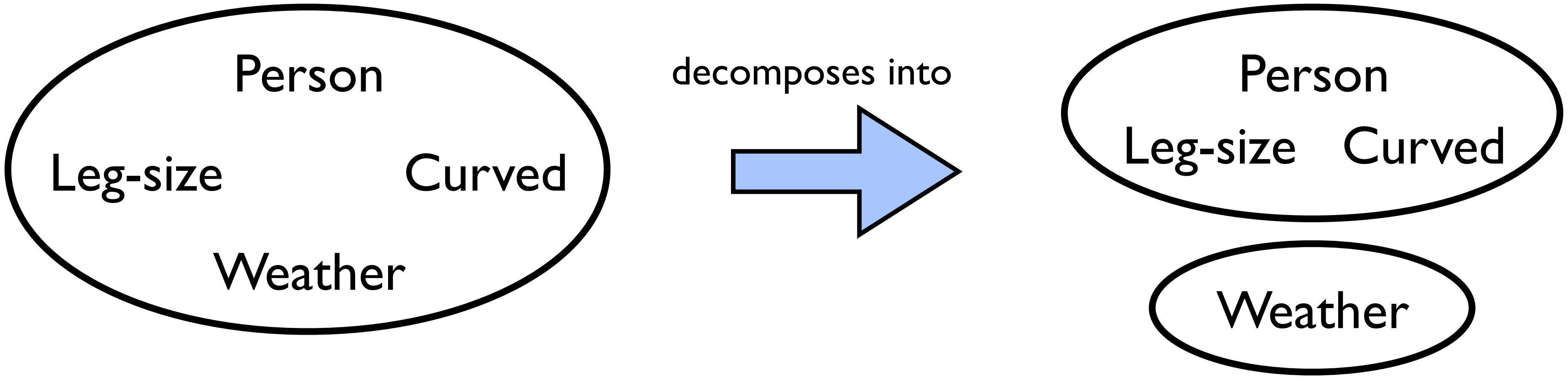
n Boolean variables give us an input table of size $O(2^n)$ ...

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

Independence

A and *B* are *independent* iff

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



Person

Leg-size

Curved

Weather

decomposes into

Person

Leg-size Curved

Weather

$$\mathbb{P}(\textit{Leg-size}, \textit{Curved}, \textit{Person}, \textit{Weather}) = \mathbb{P}(\textit{Leg-size}, \textit{Curved}, \textit{Person})\ \mathbb{P}(\textit{Weather})$$

32 entries reduced to 8 + 4 = 12 (Weather is not Boolean!).
This absolute (*unconditional*) independence is powerful but rare!

Some fields (like robotics and computer vision, or, as used in the book, dentistry) have still a lot, maybe hundreds, of variables, none of them being independent.

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

Conditional independence

$\mathbb{P}(\textit{Leg-size, Person, Curved})$ has $2^3 - 1 = 7$ independent entries (must sum up to 1)

But: If there is a person, the probability for “Curved” does not depend on whether the pattern has leg-size (this dependency is now “implicit” in some sense):

(1) $\mathbb{P}(\textit{Curved} \mid \textit{leg-size, person}) = \mathbb{P}(\textit{Curved} \mid \textit{person})$

The same holds when there is no person:

(2) $\mathbb{P}(\textit{Curved} \mid \textit{leg-size}, \neg\textit{person}) = \mathbb{P}(\textit{Curved} \mid \neg\textit{person})$

Curved is *conditionally independent* of *Leg-size* given *Person*:

$$\mathbb{P}(\textit{Curved} \mid \textit{Leg-size, Person}) = \mathbb{P}(\textit{Curved} \mid \textit{Person})$$

Question: What does that mean in the world of our coloured and partially skewed dice?

Writing out the full joint distribution using chain rule:

$$\begin{aligned}
&\mathbb{P}(\textit{Leg-size, Curved, Person})\\
&= \mathbb{P}(\textit{Leg-size} \mid \textit{Curved, Person})\ \mathbb{P}(\textit{Curved, Person})\\
&= \mathbb{P}(\textit{Leg-size} \mid \textit{Curved, Person})\ \mathbb{P}(\textit{Curved} \mid \textit{Person})\ \mathbb{P}(\textit{Person})\\
&= \mathbb{P}(\textit{Leg-size} \mid \textit{Person})\ \mathbb{P}(\textit{Curved} \mid \textit{Person})\ \mathbb{P}(\textit{Person})
\end{aligned}$$

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* (which is a combination of the relevant values of the random variables)

Queries can be answered by *summing* over atomic events (marginal probability)

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 size of joint distributions

Independence and *conditional independence* provide the tools

Outline

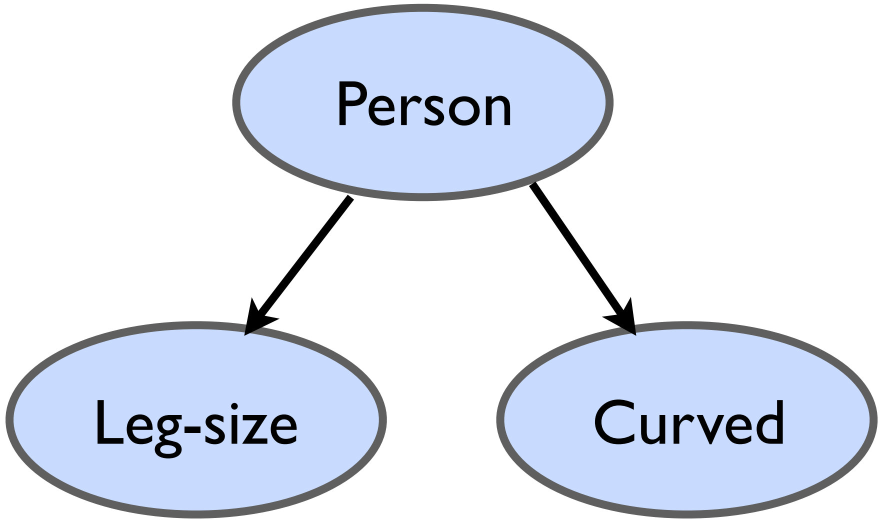
- Uncertainty & probability (chapter 13)
  - Uncertainty
  - Probability
  - Syntax and Semantics
  - Inference
  - Independence and Bayes' Rule
- Bayesian Networks (chapter 14.1-3)
  - Syntax
  - Semantics
  - Efficient representation

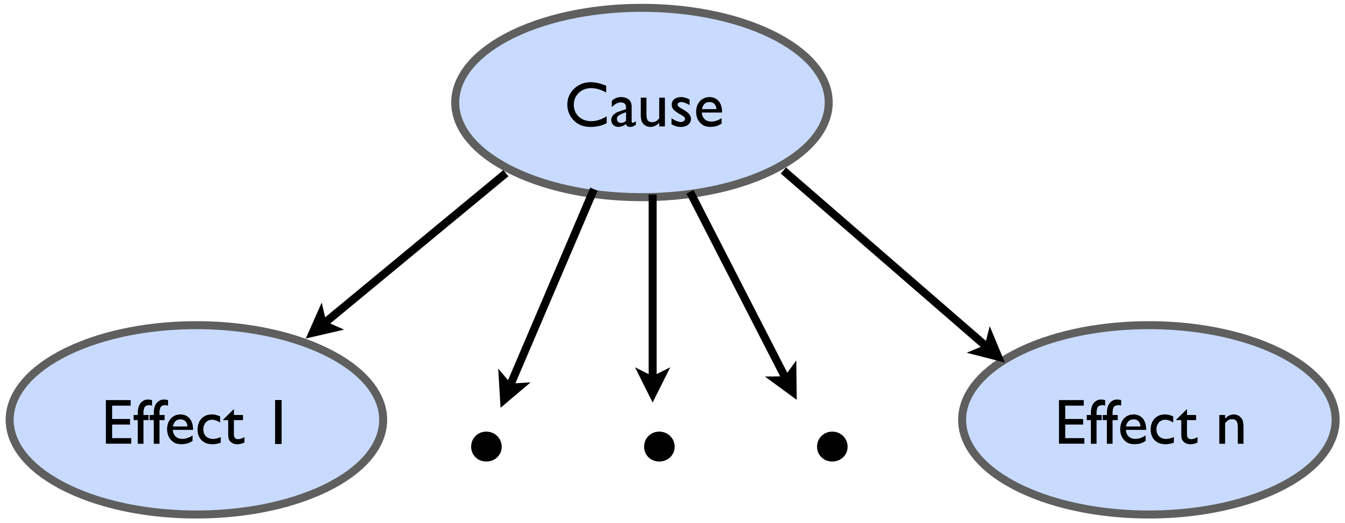
Bayes' Rule and conditional independence

$$\begin{aligned}
&\mathbb{P}(\,\mathit{Person} \mid \mathit{leg\text{-}size} \wedge \mathit{curved})\\
&= \alpha\ \ \mathbb{P}(\,\mathit{leg\text{-}size} \wedge \mathit{curved} \mid \mathit{Person})\ \mathbb{P}(\,\mathit{Person})\\
&= \alpha\ \ \mathbb{P}(\,\mathit{leg\text{-}size} \mid \mathit{Person})\ \mathbb{P}(\,\mathit{curved} \mid \mathit{Person})\ \mathbb{P}(\,\mathit{Person})
\end{aligned}$$

An example of a *naive Bayes* model:

$$\mathbb{P}(\,\mathit{Cause}, \mathit{Effect}_1, \dots, \mathit{Effect}_n) = \ \mathbb{P}(\,\mathit{Cause})\ \textstyle\prod_i \mathbb{P}(\,\mathit{Effect}_i \mid \mathit{Cause})$$





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

Question: How does this "network" look for the dice example?

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:

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

In the simplest case, conditional distribution represented as a

conditional probability table (CPT)

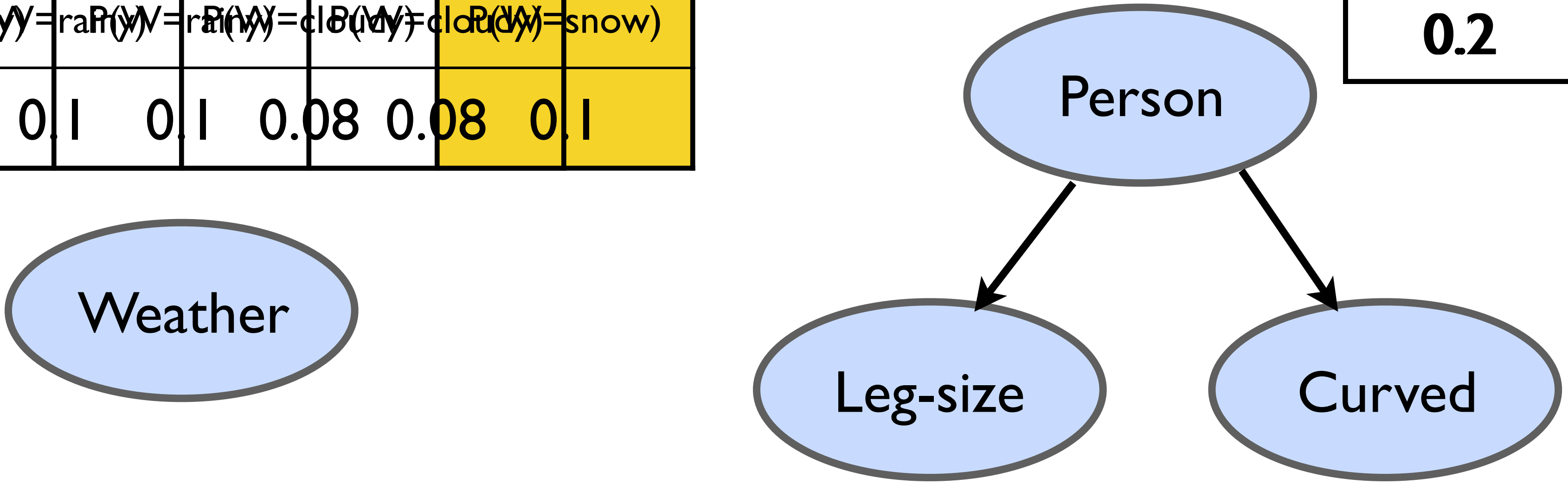
giving the distribution over X_i for each combination of parent values

Example

Topology of network encodes conditional independence assertions:

P(W=sunny)	P(W=rainy)	P(W=cloudy)	P(W=snow)
0.72	0.1	0.08	0.1

P(Per)	P(¬Per)
0.2	0.8



Per	P(L|Per)	P(¬L|Per)
T	0.6	0.4
F	0.1	0.9

Per	P(C|Per)	P(¬C|Per)
T	0.9	0.1
F	0.2	0.8

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

Leg-size and *Curved* are conditionally independent given *Person*

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.

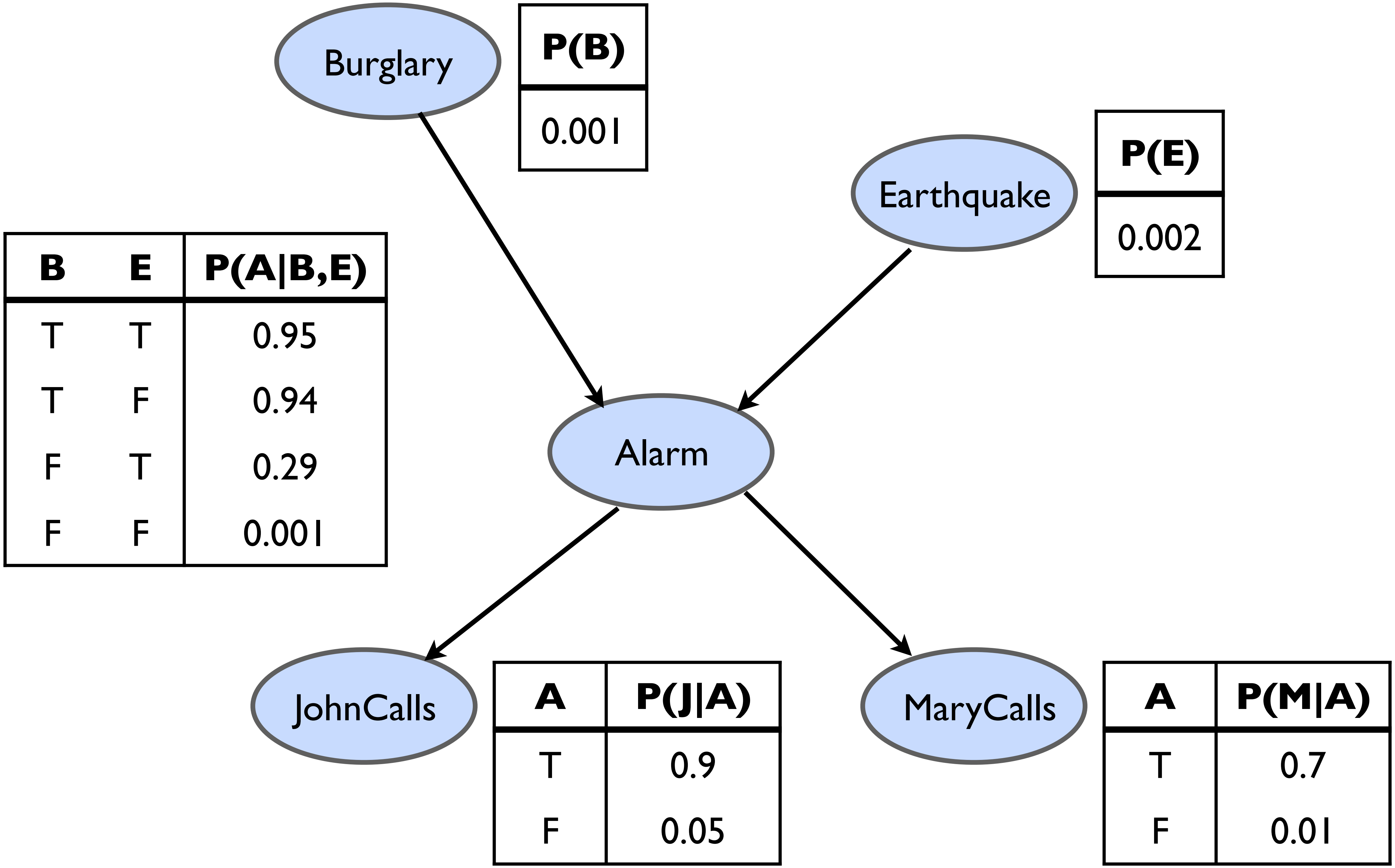
Is there a burglar?

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)



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	0.9
F	0.05

A	P(M|A)
T	0.7
F	0.01

Global semantics

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

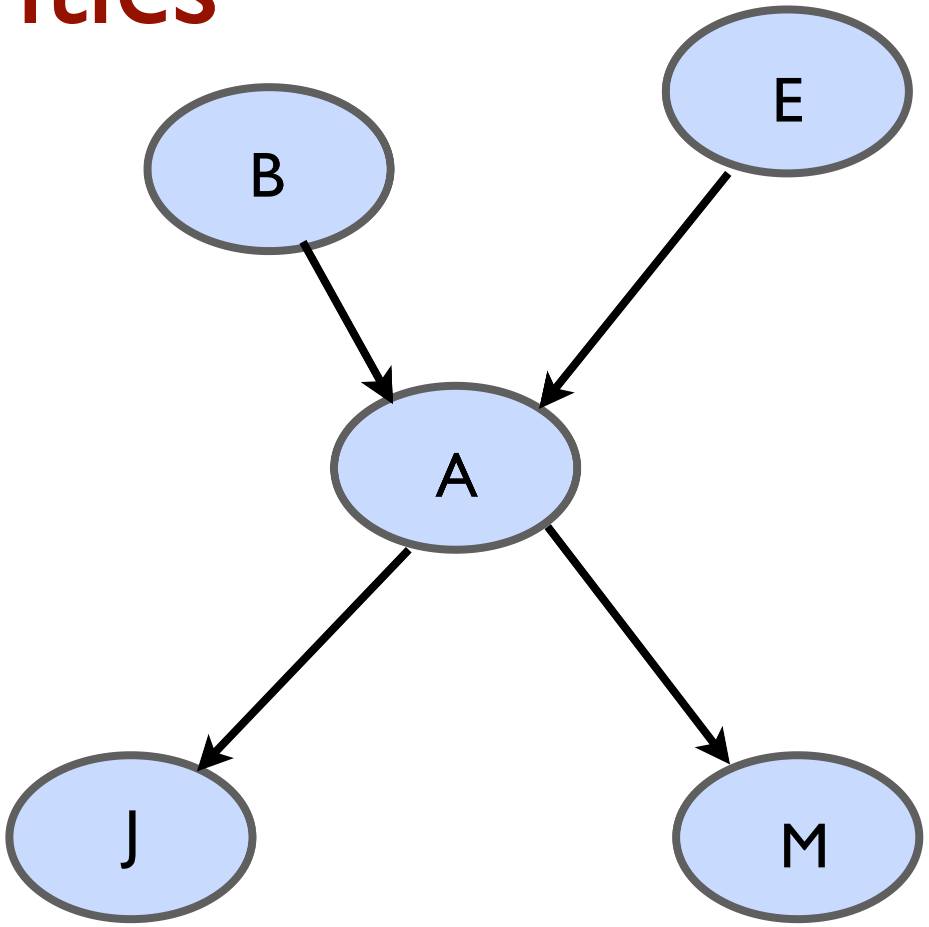
$$P(x_1, \ldots, x_n) = \Pi_{i=1}^{n} P(x_i \mid parents(X_i))$$

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 $X_1, \ldots, X_n$

2. For $i = 1$ to n

    add X_i to the network

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

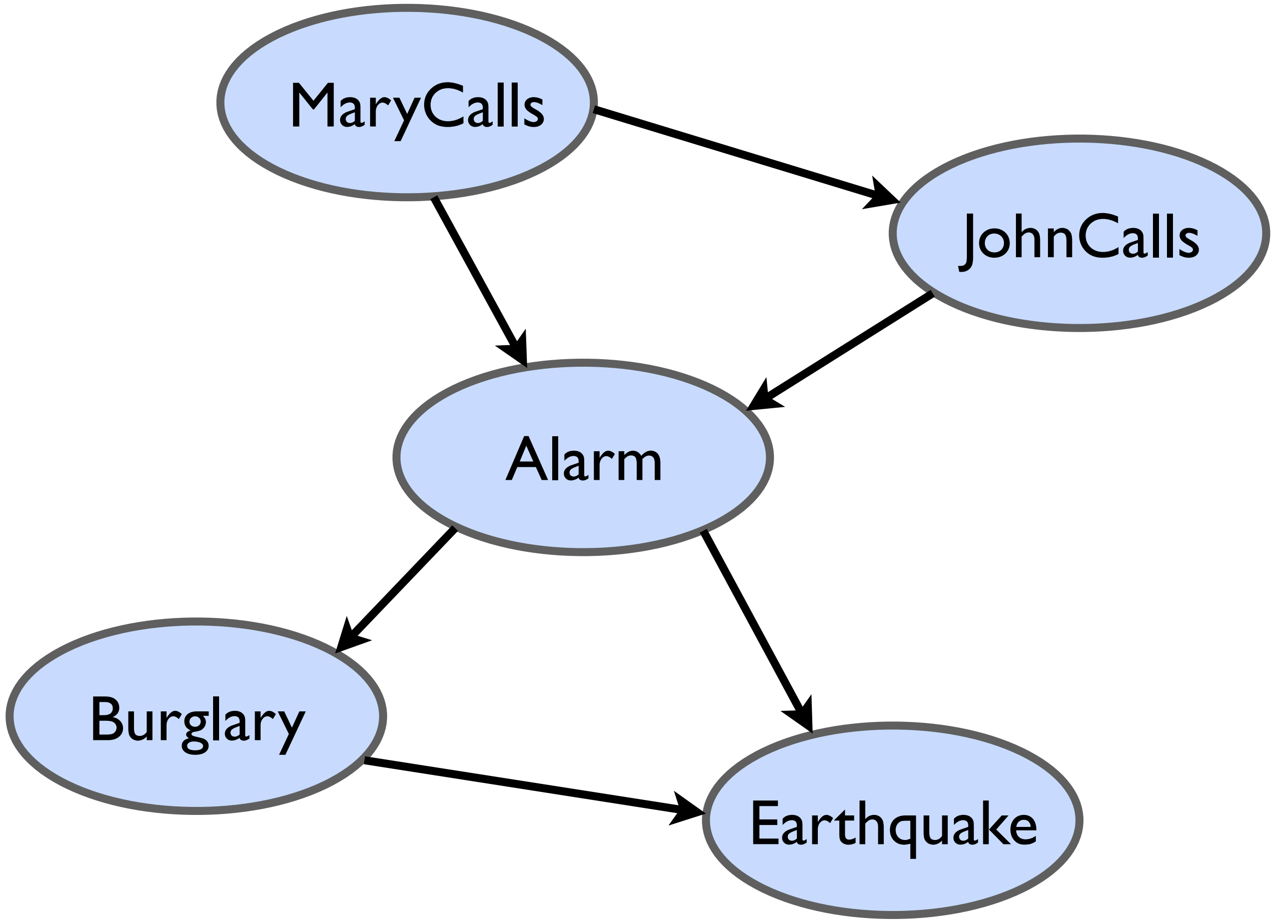
    $\mathbb{P}(X_i \mid Parents(X_i)) = \mathbb{P}(X_i \mid X_1, \ldots, X_{i-1})$

This choice of parents guarantees the global semantics:

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

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

Construction example



Deciding conditional independence is hard in noncausal directions

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

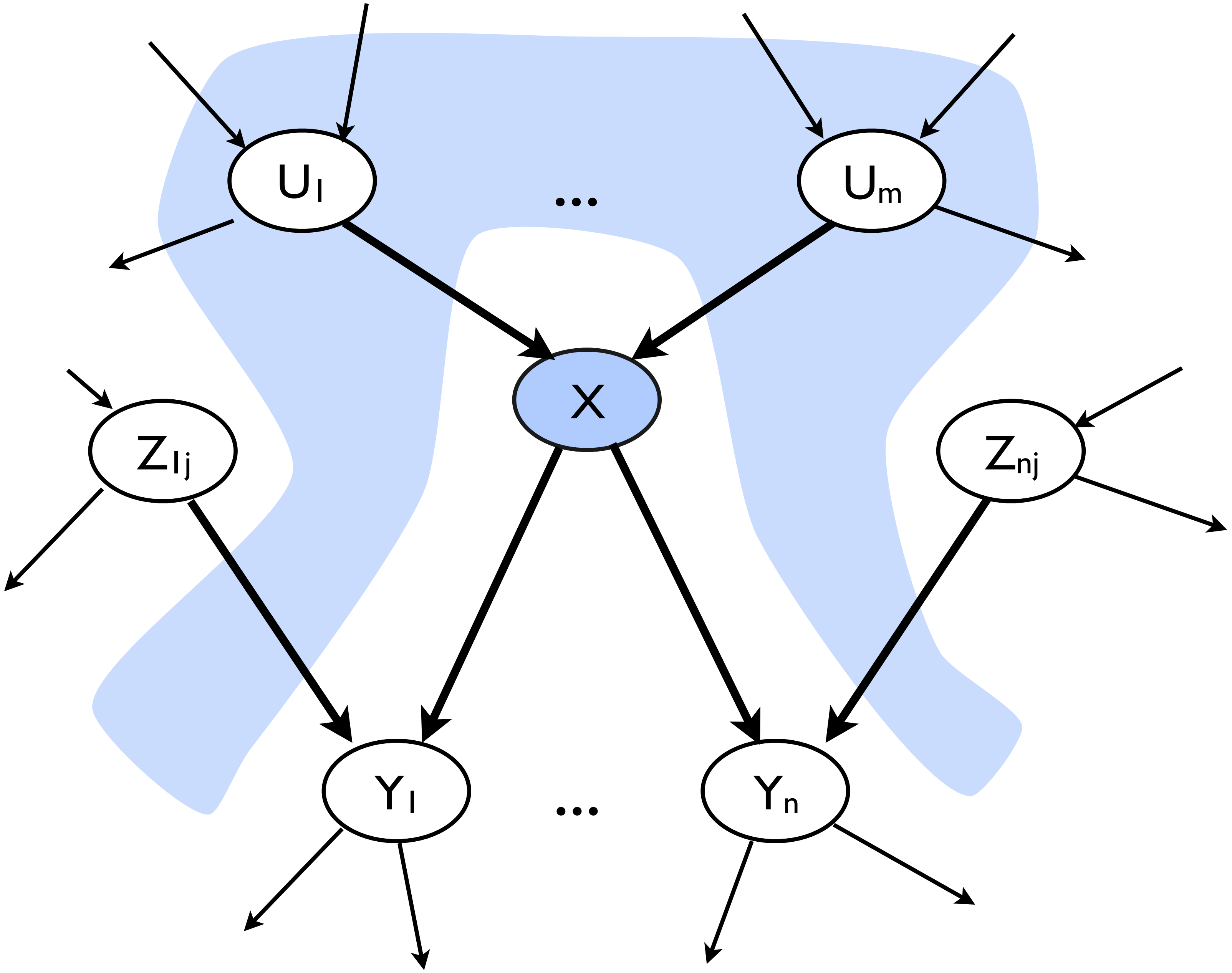
Assessing conditional probabilities is hard in noncausal directions

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

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

Local semantics

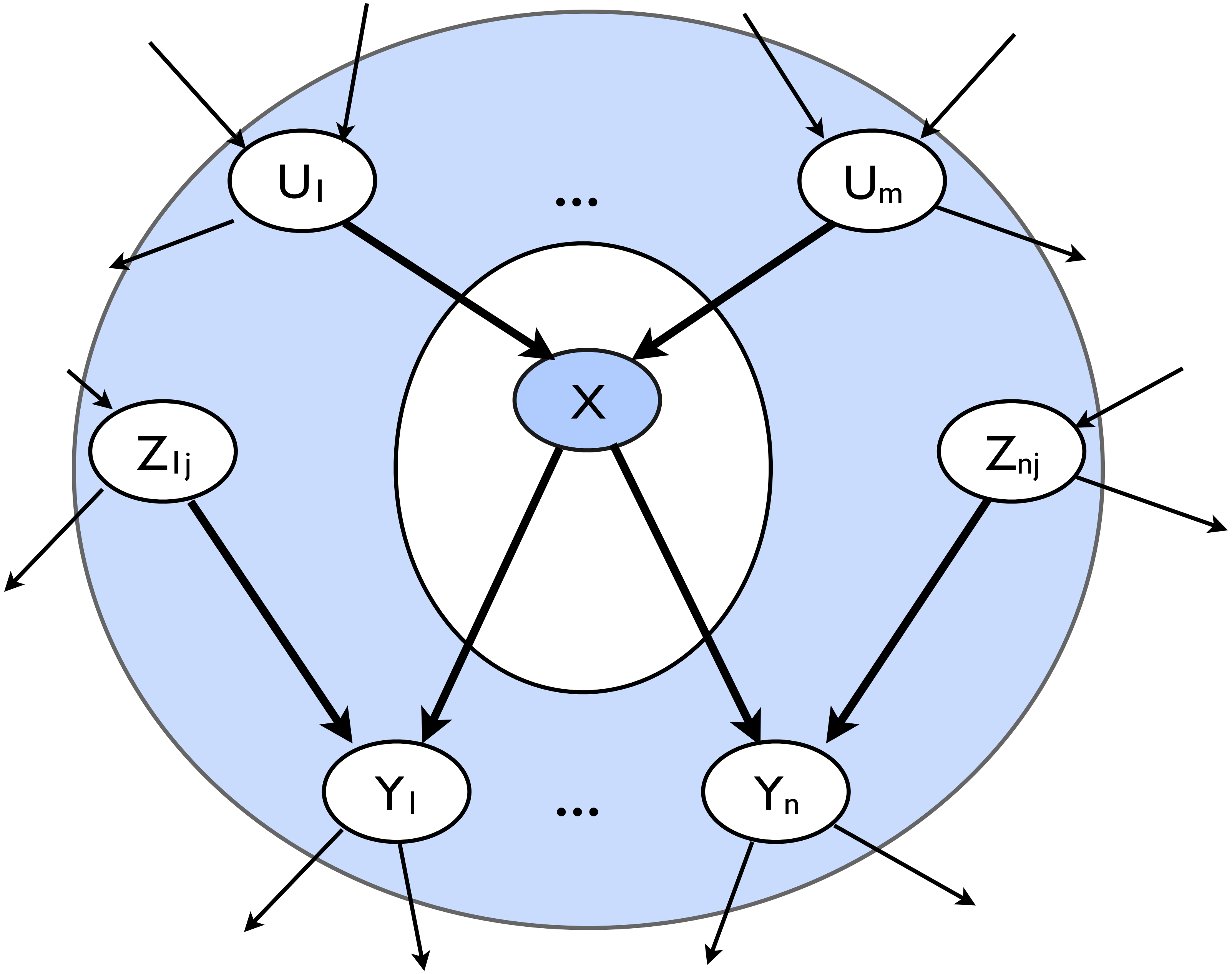
Local semantics: each node is conditionally independent of its non-descendants given its parents



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 ⇒ parameterised distributions (e.g., Gaussian)

Do BNs help for the questions in the beginning?
YES (but that story will be told later …)