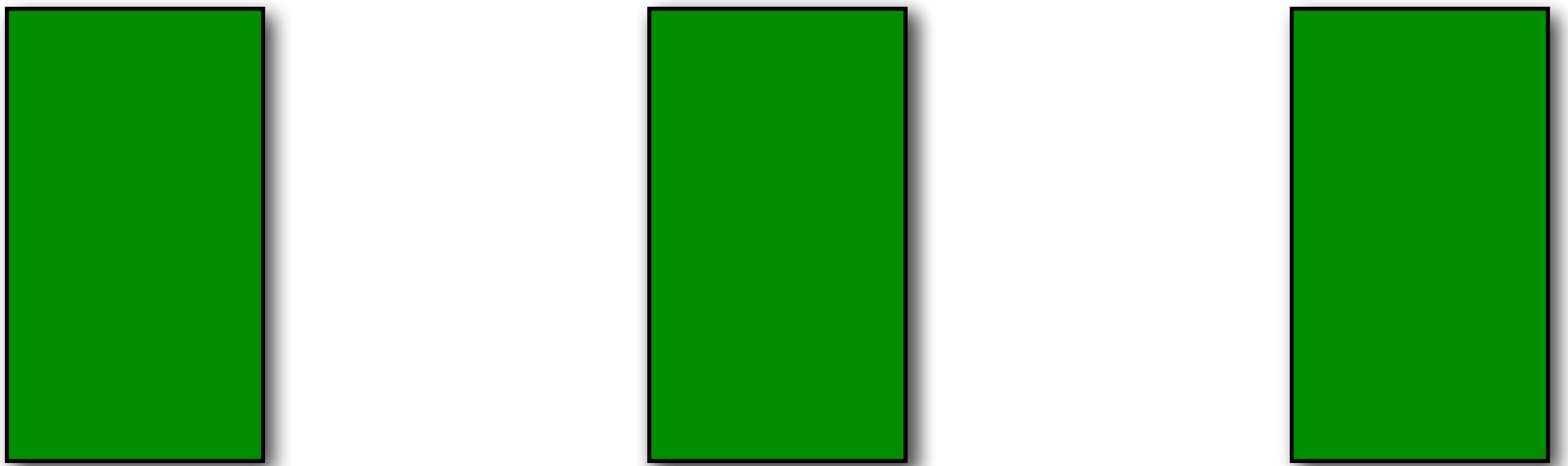
Probabilistic representation and reasoning

Applied artificial intelligence (EDAF70)
Lecture 04
2019-02-01
Elin A. Topp

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

Show time!

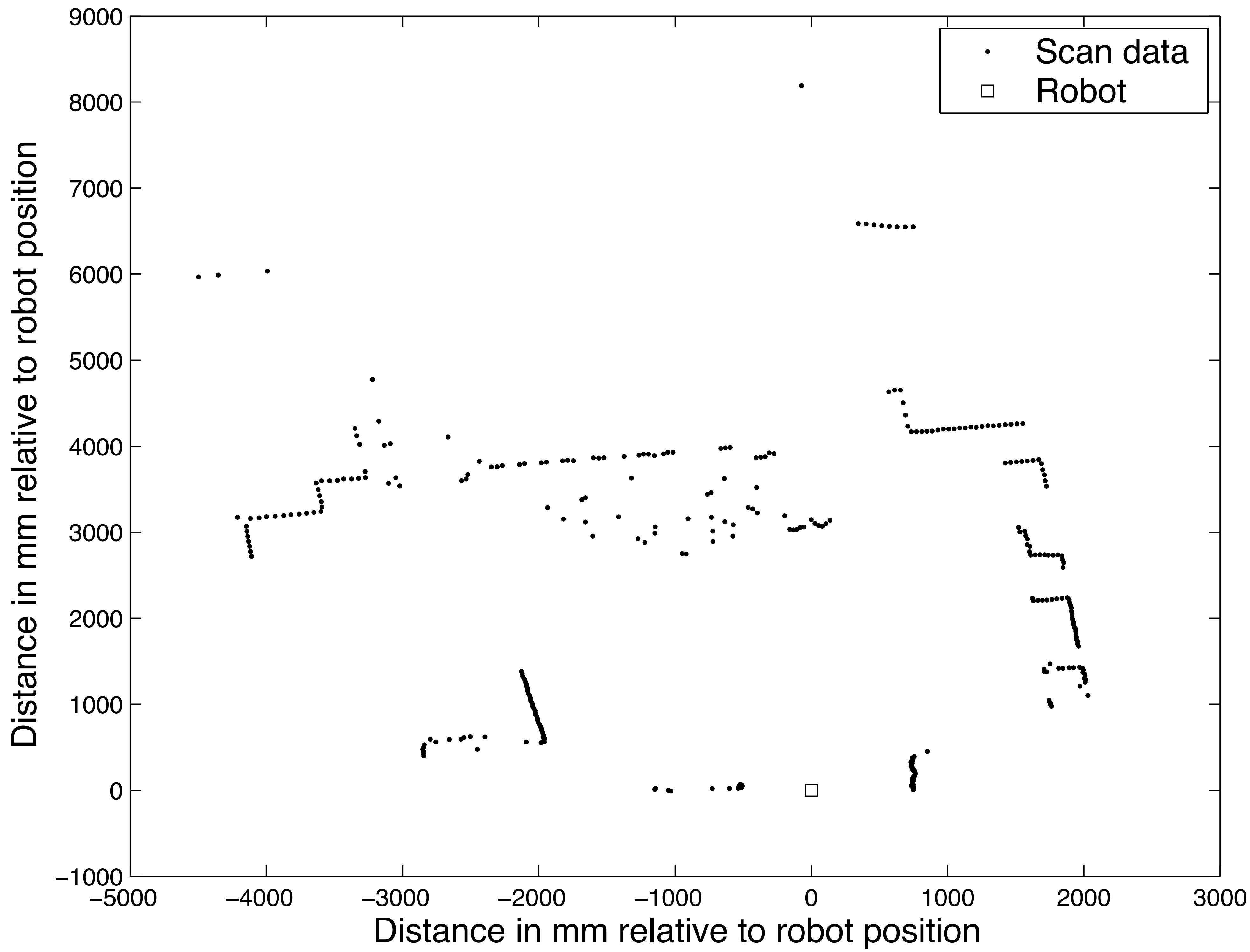
Two boxes of chocolates, one luxury car.
Where is the car?



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 alter my choice - even though it is not certain!

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?

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

Using logic in an uncertain world?

Can we find rules to describe every possible outcome, even when we cannot observe everything? (Chess, Go - and then there was Poker)

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

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 (*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*

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

$P(X = x_i), i = 1,\ldots n$, 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:

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

*and we use vector notation **P**(X) to indicate that we iterate over a subset of the values for X in a computation of a joint distribution, e.g.*

*$\mathbb{P}(X,Y) = \mathbb{P}(X | Y)$ **P**(Y) describes a set of equations, 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*

Prior probability

Prior or *unconditional probabilities* of propositions

e.g., $P(\text{Person} = \text{true}) = 0.2$ 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):

$\mathbb{P}(\text{Weather}, \text{Person})$ = a 4×2 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 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.):

$\mathbb{P}(Weather, Person) \quad = \quad \mathbb{P}(Weather \mid Person)\ \boldsymbol{P}(Person)$
(a *4x2* set of equations, governed by the chosen (given) value for Person from the array over possible values, hence ***P***)

Chain rule (successive application of product rule):

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

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

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

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:

Can also compute posterior probabilities:

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

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

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

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

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

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 constant*:

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

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

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

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

And the good news:

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

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 (vector notation to express, that for the distribution, we normally look at all possible outcomes for Y that govern P(X)):

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

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

$$P(\mathit{cause} \mid \mathit{effect}) = \frac{P(\mathit{effect} \mid \mathit{cause})\, P(\mathit{cause})}{P(\mathit{effect})}$$

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

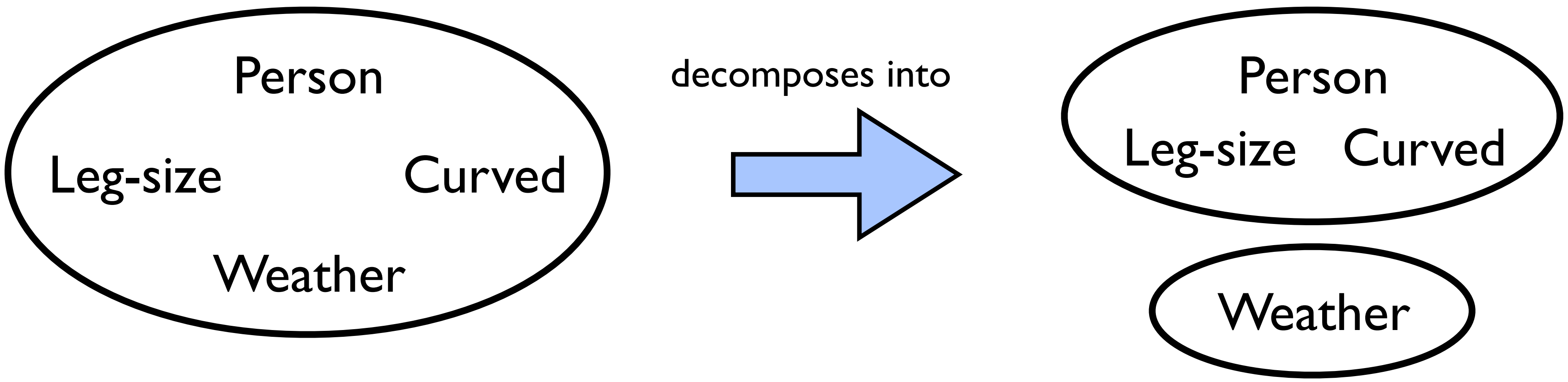
n Boolean variables give us an input table of size $O(2^n)$ ...

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

Independence

A and *B* are *independent* iff

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



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

32 entries reduced to 8 + 4 (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*:

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

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*

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

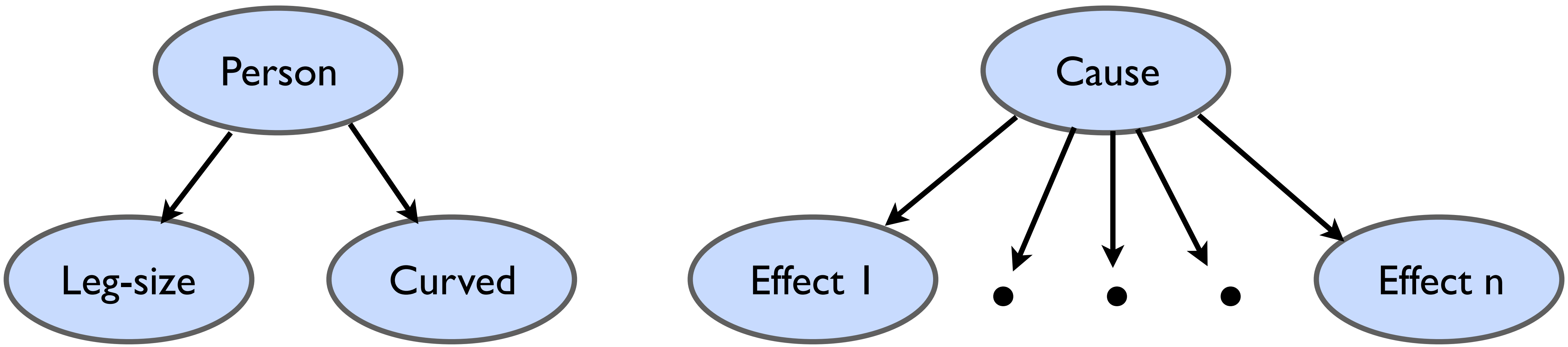
- Uncertainty & probability (chapter 13)
  - Uncertainty
  - Probability
  - Syntax and Semantics
  - Inference
  - Independence and Bayes' Rule
- Bayesian Networks (chapter 14.1-3)
  - Syntax
  - Semantics
  - Efficient representation

Bayes' Rule and conditional independence

$\mathbb{P}(\textit{Person} \mid \textit{leg-size} \wedge \textit{curved})$
$= \alpha\ \mathbb{P}(\textit{leg-size} \wedge \textit{curved} \mid \textit{Person})\ \mathbb{P}(\textit{Person})$
$= \alpha\ \mathbb{P}(\textit{leg-size} \mid \textit{Person})\ \mathbb{P}(\textit{curved} \mid \textit{Person})\ \mathbb{P}(\textit{Person})$

An example of a *naive Bayes* model:

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



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 ≈ "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)

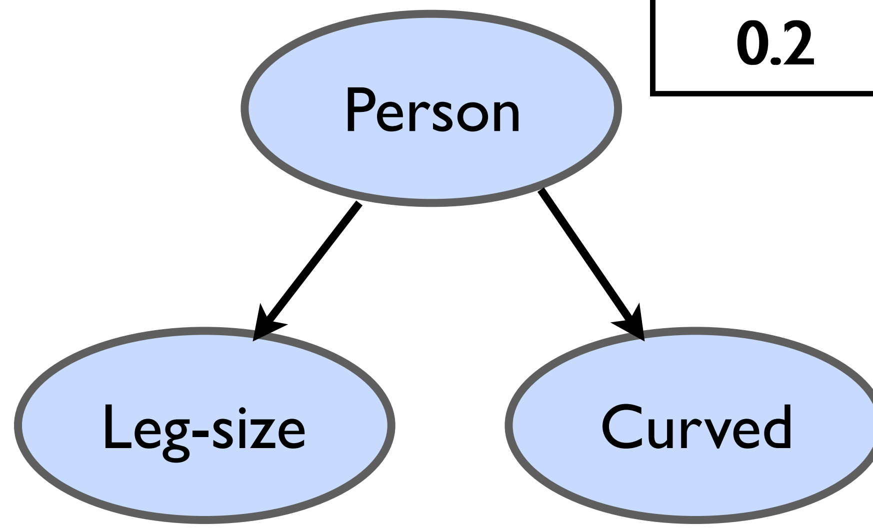
giving the distribution over X_i for each combination of parent values

Example

Topology of network encodes conditional independence assertions:

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

P(Per)	P(¬Per)
0.2	0.8



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

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

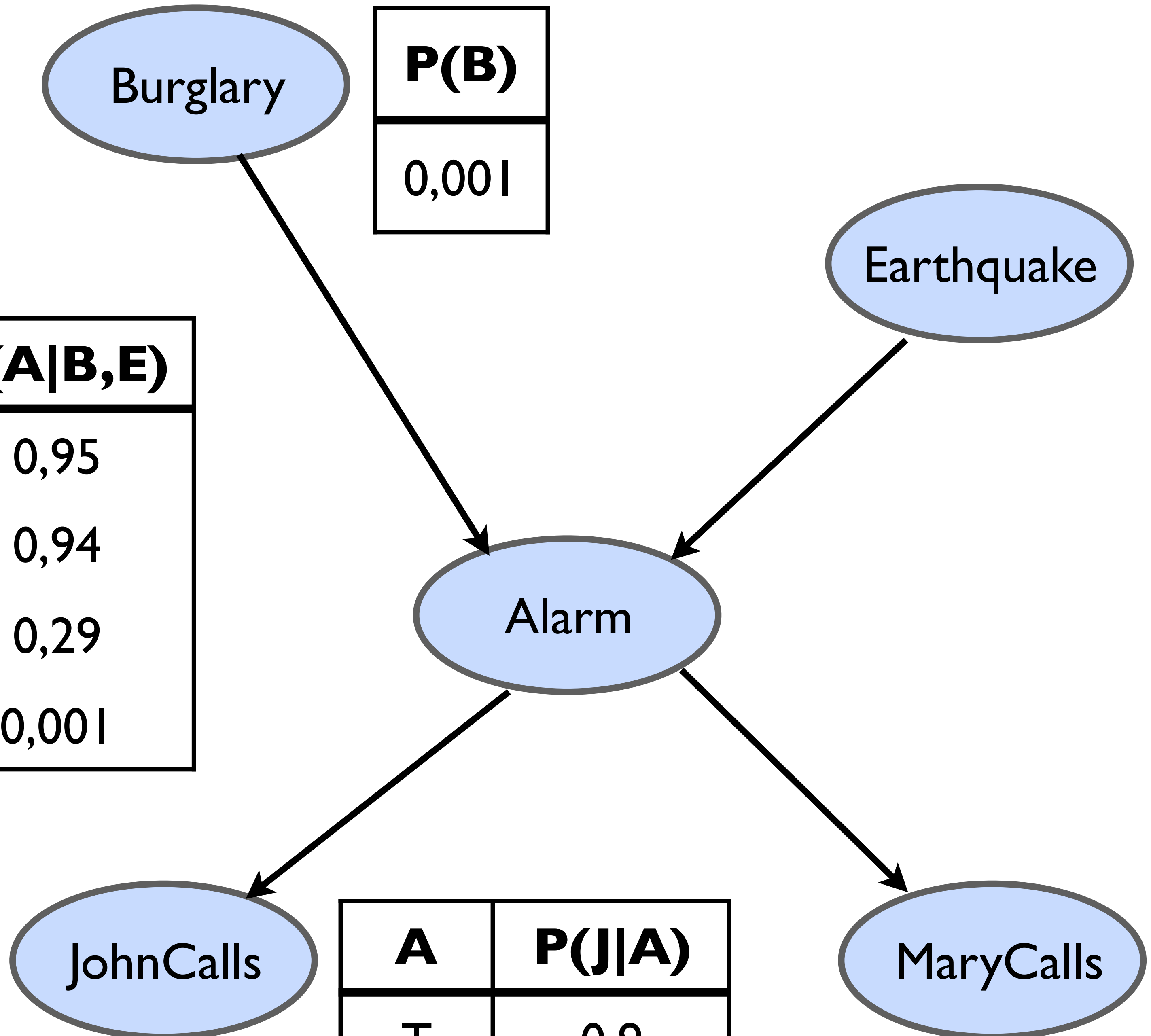
A burglar can set the alarm off

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:

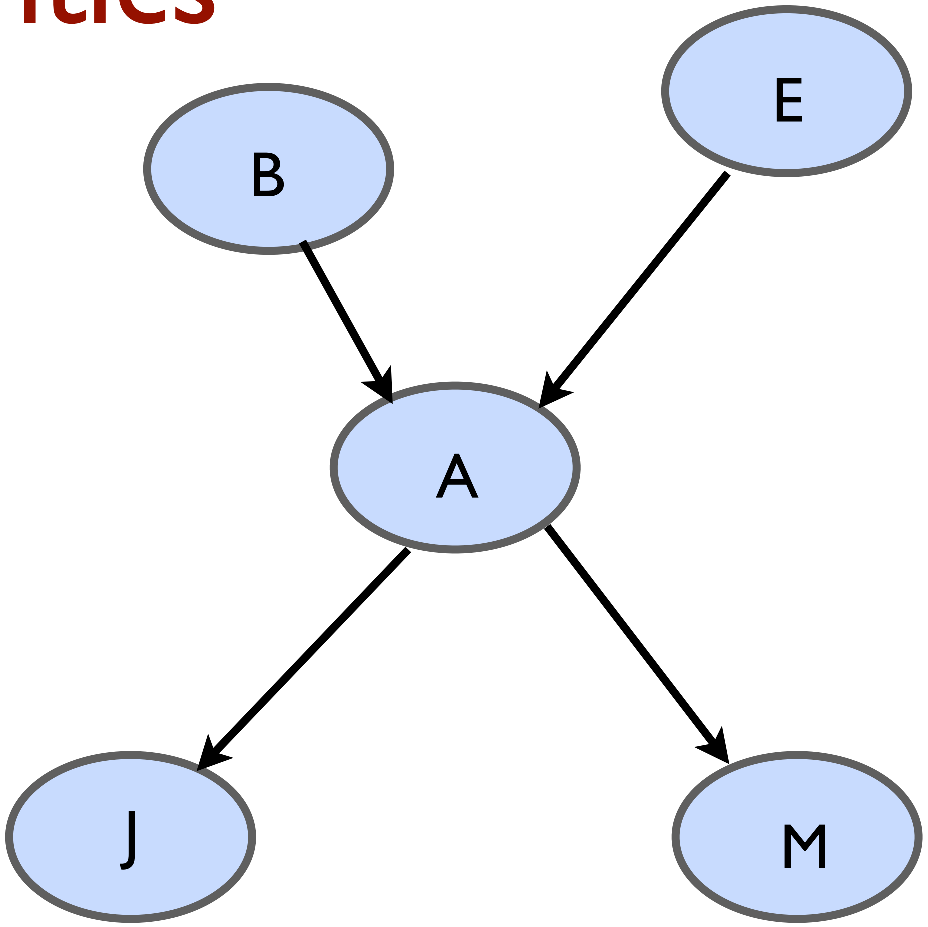
$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i \mid parents(X_i))$$

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

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

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

≈ 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

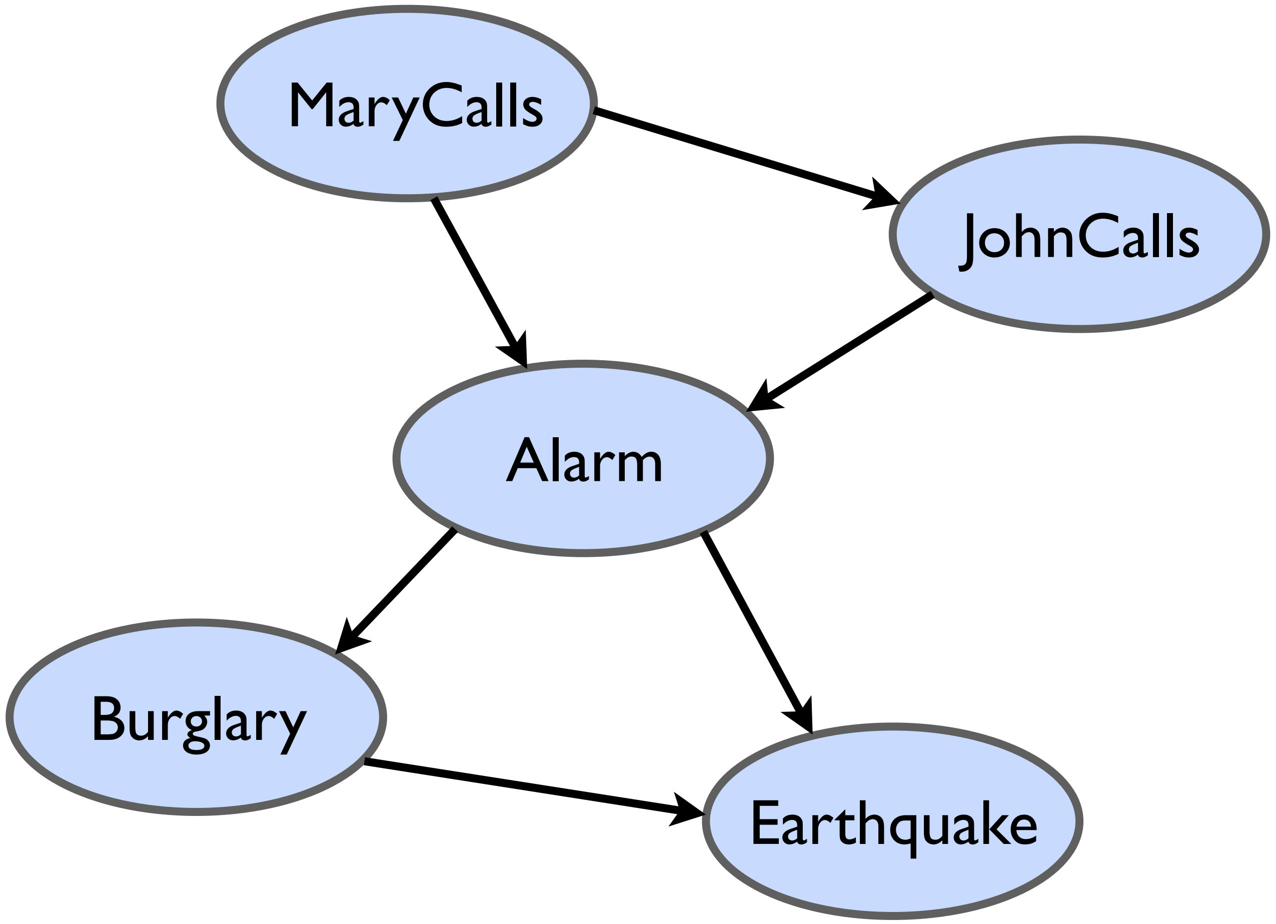
   $$\mathbf{P}(X_i \mid Parents(X_i)) = \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$$

This choice of parents guarantees the global semantics:

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

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

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

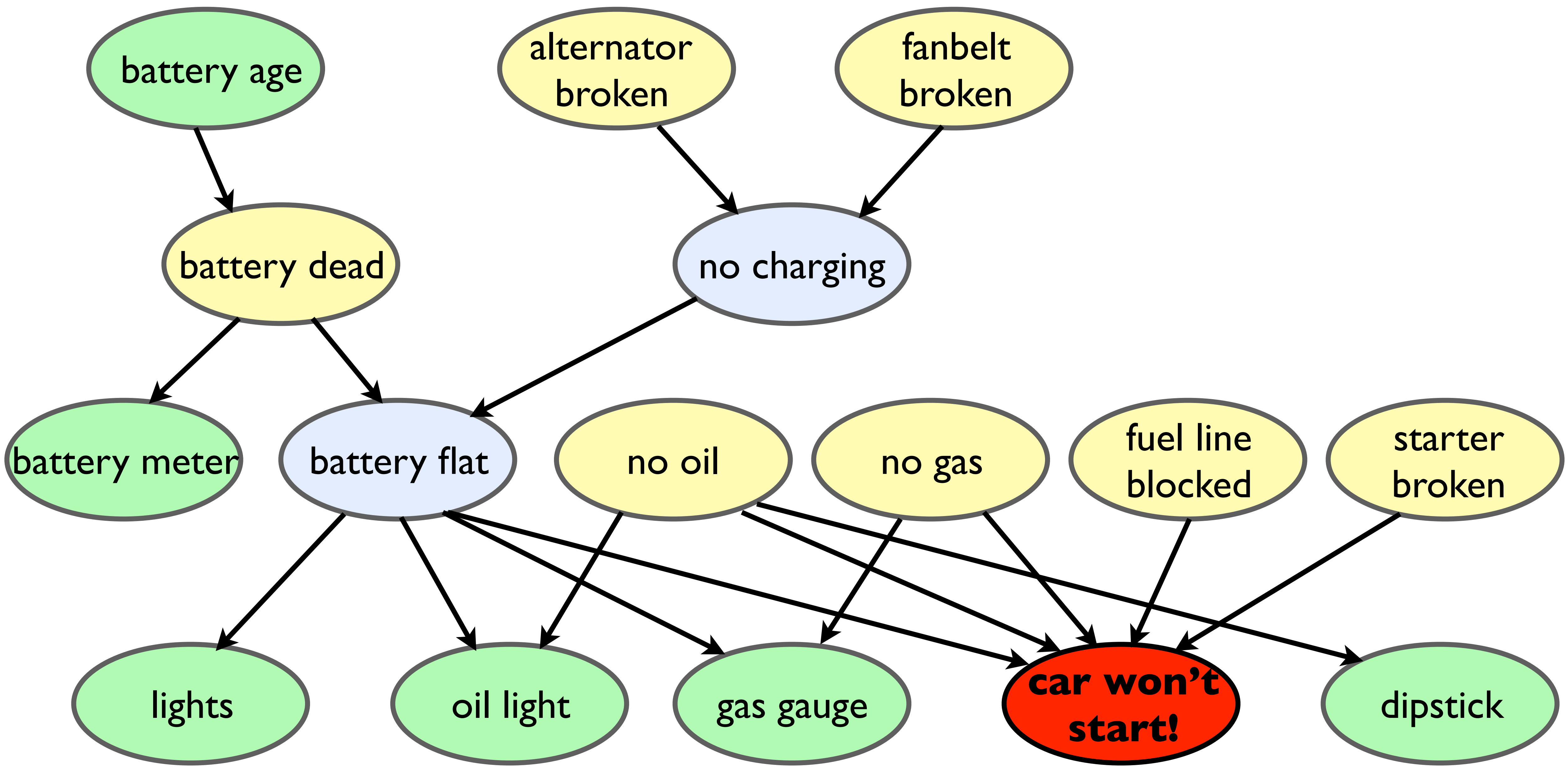
Hence: Choose preferably an order corresponding to the cause → effect "chain"

Locally structured (sparse) network

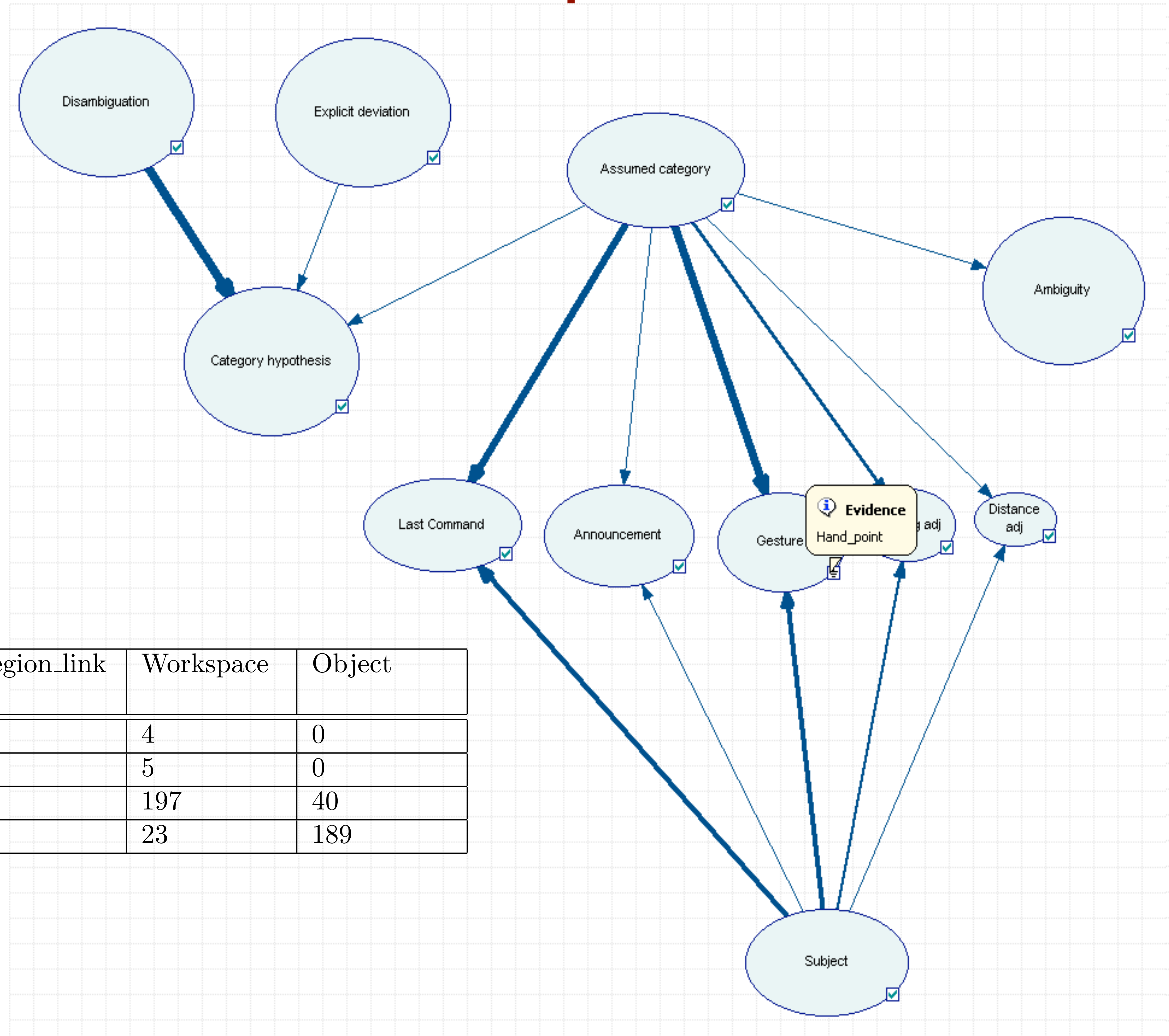
Initial evidence: The *** car won't start!

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

Hidden variables (blue) ensure sparse structure / reduce parameters



BNs for interaction patterns



Prediction Definition	Region	Region_link	Workspace	Object
Region	62	0	4	0
Region_link	16	3	5	0
Workspace	5	0	197	40
Object	0	0	23	189

Elin A. Topp, "Interaction Patterns in Human Augmented Mapping"
Special Issue on Spatial Interaction and Reasoning for Real-World Robotics, RSJ Advanced Robotics, vol 5, issue 31, March 2017

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)

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