# TDDC17 LE11 HT2023

## Bayesian Networks

**Fredrik Heintz**

Dept. of Computer Science
Linköping University

fredrik.heintz@liu.se

@FredrikHeintz

Outline:

- Reasoning with Uncertainty
- Bayesian Networks

LINKÖPING
UNIVERSITY

# Seminar Outline

- Basic Probability Theory from a logical perspective
- Bayesian Networks
    - An “efficient” means for doing probabilistic reasoning.
- Bayes’ Rule
- Naive Bayes Model

# Propositional Logic and Models

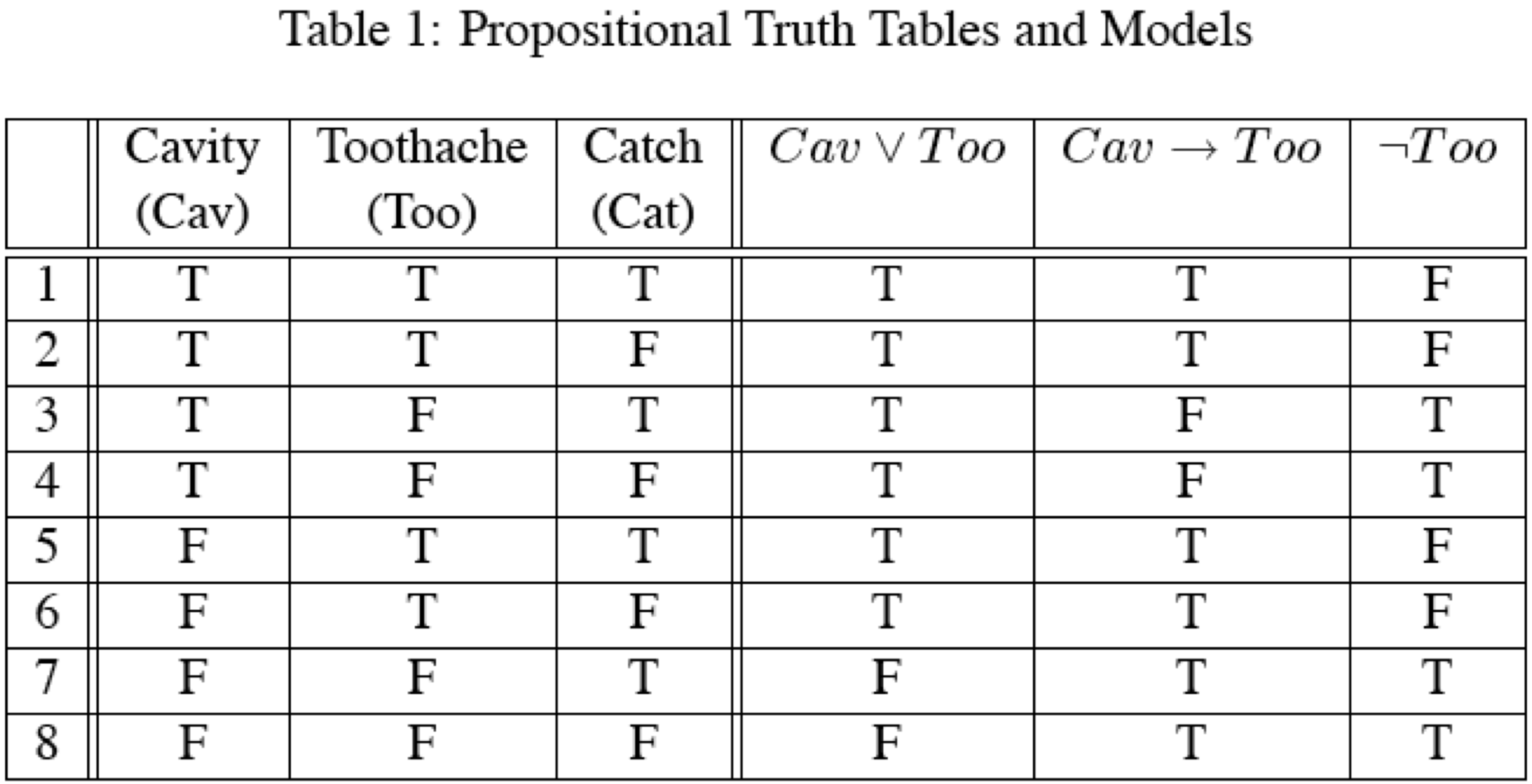


Table 1: Propositional Truth Tables and Models

| | Cavity (Cav) | Toothache (Too) | Catch (Cat) | $Cav \vee Too$ | $Cav \rightarrow Too$ | $\neg Too$ |
|---|---|---|---|---|---|---|
| 1 | T | T | T | T | T | F |
| 2 | T | T | F | T | T | F |
| 3 | T | F | T | T | F | T |
| 4 | T | F | F | T | F | T |
| 5 | F | T | T | T | T | F |
| 6 | F | T | F | T | T | F |
| 7 | F | F | T | F | T | T |
| 8 | F | F | F | F | T | T |

# DNF Characterization of Models

Any propositional formula can be equivalently represented in Disjunctive Normal Form(DNF) based on its truth table characterisation

For example:

$$Cav \vee Tooth \equiv 1 \vee 2 \vee 3 \vee 4 \vee 5 \vee 6$$

The lines in the table that make the formula true

$$\equiv (Cav \wedge Too \wedge Cat) \vee (Cav \wedge Too \wedge \neg Cat) \vee (Cav \wedge \neg Too \wedge Cat) \vee$$
$$(Cav \wedge \neg Too \wedge \neg Cat) \vee (\neg Cav \wedge Too \wedge Cat) \vee (\neg Cav \wedge Too \wedge \neg Cat) \vee$$
$$(\neg Cav \wedge \neg Too \wedge Cat) \vee (\neg Cav \wedge \neg Too \wedge \neg Cat)$$

Observe that:

$$True \equiv 1 \vee 2 \vee 3 \vee 4 \vee 5 \vee 6 \vee 7 \vee 8$$

$$False \equiv \neg True$$

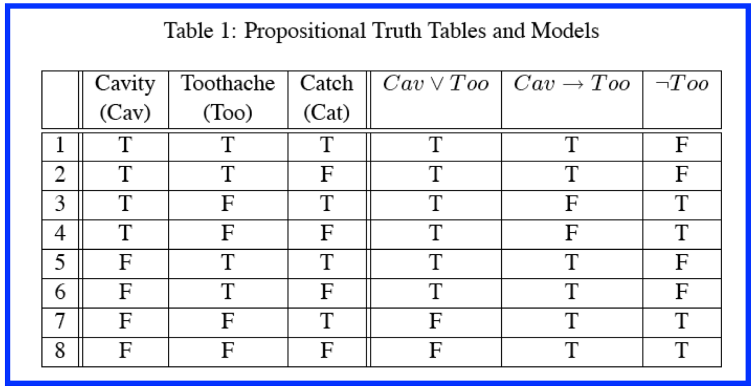
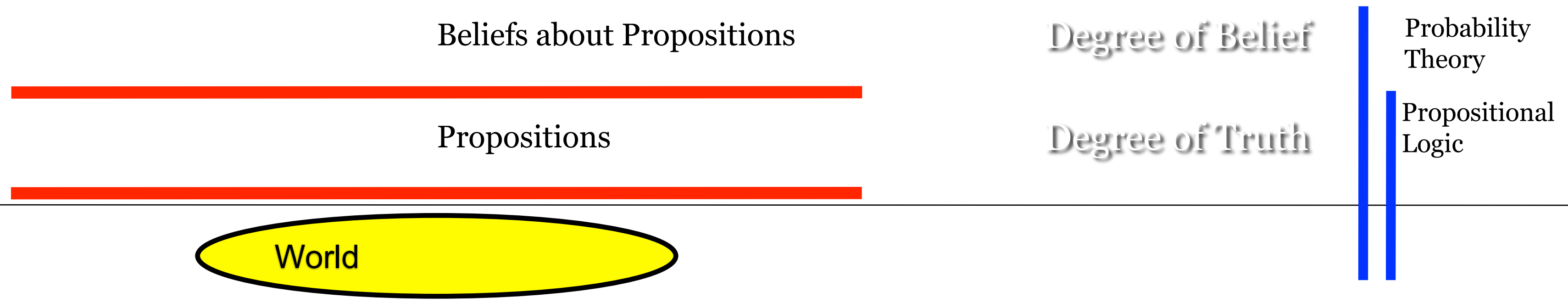


Table 1: Propositional Truth Tables and Models

|  | Cavity (Cav) | Toothache (Too) | Catch (Cat) | $Cav \vee Too$ | $Cav \rightarrow Too$ | $\neg Too$ |
|---|---|---|---|---|---|---|
| 1 | T | T | T | T | T | F |
| 2 | T | T | F | T | T | F |
| 3 | T | F | T | T | F | T |
| 4 | T | F | F | T | F | T |
| 5 | F | T | T | T | T | F |
| 6 | F | T | F | T | T | F |
| 7 | F | F | T | F | T | T |
| 8 | F | F | F | F | T | T |

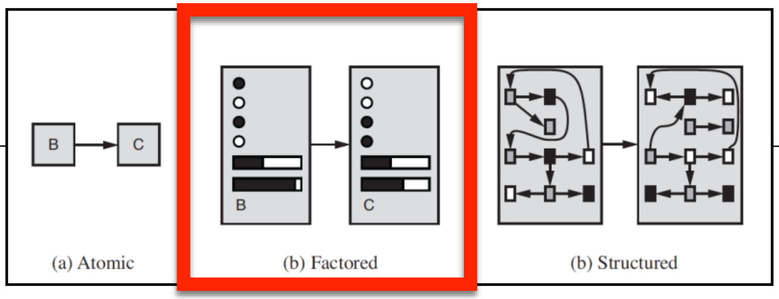
# Degrees of Truth/Belief

- Truth Table Method:
  - Can be used to evaluate the Truth or Falsity of a formula
  - Requires a table with $2^n$ rows, where $n$ is the number of propositional variables in the language
- Propositional logic:
  - Allows the representation of propositions about the world which are True or False
  - In this case, a proposition has a degree of truth, either true or false
- Suppose our knowledge about the truth or falsity of a proposition is uncertain
  - In this case we might want to attach a degree of belief in the propositions truth status
- Observe that the degree of belief is subjective, in the sense that
  - the proposition in question is still considered to be true or false about the world
  - We simply do not have enough information to determine this.
- So, there is a distinction between degrees of truth and degrees of belief



# A Language of Probability

- Just as propositional atoms provide the primitive vocabulary for propositions in propositional logic, random variables will provide the primitive vocabulary for our probabilistic language.
- Random variables:
  - Boolean: $Cavity: \{true, false\}$
  - Discrete: $Weather: \{sunny, rainy, cloudy, snow\}$
  - Continuous: $Temperature: \{x \mid -43.0 \leq x \leq 100.0\}$
- A random variable may be viewed as an aspect/feature of the world that is initially unknown
  - A degree of belief may be attached to a variable/value pair
  - Complex formulas may be formed using Boolean combinations of variable/value pairs



# Probability Distributions

$P(Cavity = true) = P(cavity) = 0.4$

$P(Cavity = false) = P(\neg cavity) = 0.6$

$\mathbf{P}(Cavity) = \langle 0.4,0.6 \rangle$

$\mathbf{P}$ Notation

$P(Weather = sunny) = 0.7$

$P(Weather = rainy) = 0.2$

$P(Weather = cloudy) = 0.08$

$P(Weather = snow) = 0.02$

$\mathbf{P}(Weather) = \langle 0.7,0.2,0.08,0.02 \rangle$

$\mathbf{P}(X)$ is the Probability Distribution (Unconditional or Prior Probability of the random variable $X$

# Joint Probability Distributions

Assume a domain of random variables: $\{X_1, \ldots, X_n\}$

A full joint probability distribution $\mathbf{P}(X_1, \ldots, X_n)$, assigns a probability to each of the possible combinations of variable/value pairs

$$\mathbf{P}(Cavity, Weather) = \langle 0.30, 0.05, 0.145, 0.005, 0.30, 0.05, 0.145, 0.005 \rangle \quad (2 \text{ x } 4)$$

**P** notation: can also mix variables and specific values:

$$\mathbf{P}(cavity, Weather) = \langle 0.30, 0.05, 0.145, 0.005 \rangle \quad (1 \text{ x } 4)$$

$$\mathbf{P}(Cavity, Weather = rainy) = \langle 0.05, 0.05 \rangle \quad (2 \text{ x } 1)$$

# An Example

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

Each logical model is an atomic event

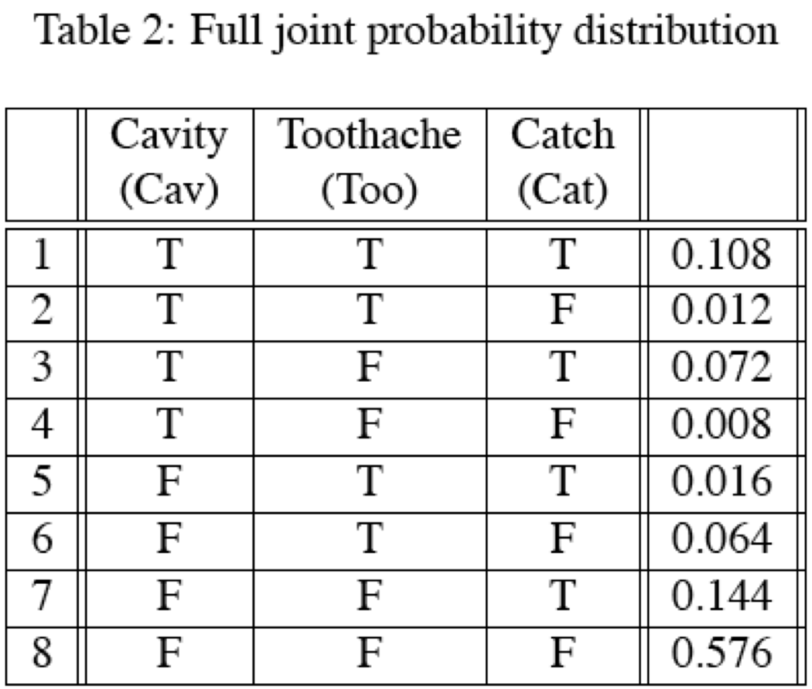


Table 2: Full joint probability distribution

| | Cavity (Cav) | Toothache (Too) | Catch (Cat) | |
|---|---|---|---|---|
| 1 | T | T | T | 0.108 |
| 2 | T | T | F | 0.012 |
| 3 | T | F | T | 0.072 |
| 4 | T | F | F | 0.008 |
| 5 | F | T | T | 0.016 |
| 6 | F | T | F | 0.064 |
| 7 | F | F | T | 0.144 |
| 8 | F | F | F | 0.576 |

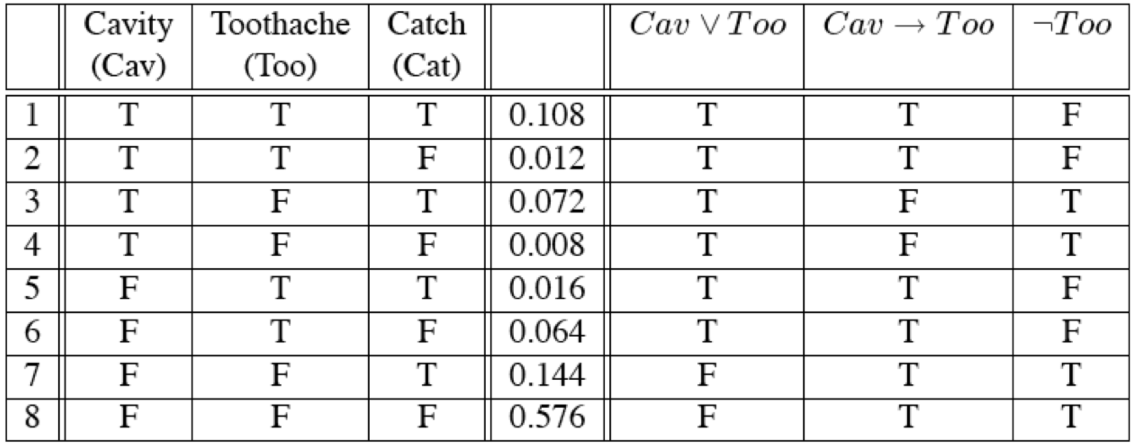
The probabilities for each atomic event (an interpretation) must sum to 1.

# Using a Full Joint Probability Distribution

Using a full joint probability distribution, arbitrary Boolean combinations of variable value pairs can be interpreted by taking the sum of the beliefs attached to each interpretation (atomic event) which satisfies the formula.

Recall our DNF characterisation of logical formulas!

Table 3: Interpreting formulas



| | Cavity<br>(Cav) | Toothache<br>(Too) | Catch<br>(Cat) | | $Cav \vee Too$ | $Cav \rightarrow Too$ | $\neg Too$ |
|---|---|---|---|---|---|---|---|
| 1 | T | T | T | 0.108 | T | T | F |
| 2 | T | T | F | 0.012 | T | T | F |
| 3 | T | F | T | 0.072 | T | F | T |
| 4 | T | F | F | 0.008 | T | F | T |
| 5 | F | T | T | 0.016 | T | T | F |
| 6 | F | T | F | 0.064 | T | T | F |
| 7 | F | F | T | 0.144 | F | T | T |
| 8 | F | F | F | 0.576 | F | T | T |

$$P(cav \vee too) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$$
$$P(cav \rightarrow too) = 0.108 + 0.012 + 0.016 + 0.064 + 0.144 + 0.576 = 0.2$$
$$P(\neg too) = 0.072 + 0.008 + 0.144 + 0.576 = 0.8$$
$$P(\neg too) = 1 - P(too) = 1 - (0.108 + 0.012 + 0.016 + 0.064) = 0.8$$

# Conditional Probability

In classical logic, our main focus is often: $\Gamma \vDash \alpha$

In probability theory, our main focus is often: $P(\mathbf{X} \mid \mathbf{Y})$

Prior probabilities are not adequate once additional evidence concerning previously unknown random variables is introduced:

- One must condition any random variable(s) of interest relative to the new evidence.
- Conditioning is represented using conditional or posterior probabilities.

The probability of $X = x_i$ given $Y = y_j$ is denoted $P(X = x_i \mid Y = y_j)$

$$P(X = x_i \mid Y = y_j) = \frac{P(X = x_i \land Y = y_j)}{P(Y = y_j)}$$

Another way to write this is in the form of the product rule:

$$P(X = x_i \land Y = y_j) = P(X = x_i \mid Y = y_j) * P(Y = y_j)$$

$$P(X = x_i \land Y = y_j) = P(Y = y_j \mid X = x_i) * P(X = x_i)$$

This rule can be generalised using the chain rule

# Some additional $\mathbf{P}$ notation

$\mathbf{P}(X \mid Y)$ denotes the set of equations $P(X = x_i \mid Y = y_j)$ for each possible i, $j$.

For example:

$$\mathbf{P}(X \wedge Y) = \mathbf{P}(X, Y) = \mathbf{P}(X \mid Y) * \mathbf{P}(Y)$$

$$P(X = x_1 \wedge Y = y_1) = P(X = x_1 \mid Y = y_1) * P(Y = y_1)$$
$$P(X = x_1 \wedge Y = y_2) = P(X = x_1 \mid Y = y_2) * P(Y = y_2)$$
$$\vdots$$
$$P(X = x_i \wedge Y = y_j) = P(X = x_i \mid Y = y_j) * P(Y = y_j)$$

Note also that:

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

Conjunction is abbreviated as a “,”

$\mathbf{P}(X, Y)$ is also a distribution so it is equal to a vector if we have the distribution

# Kolmogorov's Axioms

Recall our discussions about logical theories, Δ, consisting of a set of axioms and our interest in $\Delta \models \alpha$

Probability Theory can be built up from three axioms:

1. All probabilities are between 0 and 1.
   - For any proposition $a$, $0 \leq P(a) \leq 1$.
2. Necessarily true (i.e. valid) propositions have probability 1, and necessarily false propositions have probability 0.
   - $P(True) = 1$ and $P(False) = 0$.
3. The probability of a disjunction is given by:
   - $P(a \vee b) = P(a) + P(b) - P(a \wedge b)$

# Some Useful Properties

In probability theory, the set of all possible worlds is called the sample space, $\Omega$. Let $\omega$ refer to elements of the sample space (models/interpretations). Assume $\Omega$ is a discrete countable set of worlds.

$$0 \leq P(\omega) \leq 1, \text{ for all } \omega.$$

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

$$P(True) = 1$$

For any proposition $\phi$, $P(\phi) = \sum_{\omega \in \phi} P(\omega)$

# Marginal Probability & Marginalization

Joint probability distribution $\mathbf{P}(Toothache, Cavity, Catch)$:

0.2

| | $toothache$ | | $\neg toothache$ | |
|---|---|---|---|---|
| | $catch$ | $\neg catch$ | $catch$ | $\neg catch$ |
| $cavity$ | $\underline{0.108}$ | $\underline{0.012}$ | $\underline{0.072}$ | $\underline{0.008}$ |
| $\neg cavity$ | 0.016 | 0.064 | 0.144 | 0.576 |

Marginalization is about extracting the distribution over some subset of variables or a single variable

The marginal probability of $cavity$ is:

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

Let $\mathbf{Y}$ and $\mathbf{Z}$ be sets of variables, and where $\sum_{\mathbf{z}}$ sums over all possible combinations of values of the set of variables $\mathbf{Z}$. Then the general marginalization rule is:

$$\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}, \mathbf{z})$$

# Some Examples

$$\boxed{\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}, \mathbf{z})}$$

Marginal probability of $Cavity \land Catch$

Let $\mathbf{Y} = \{Cavity, Catch\}$ and $\mathbf{Z} = \{Toothache\}$

$$\mathbf{P}(\mathbf{Y}) = \mathbf{P}(\mathbf{Y}, toothache) + \mathbf{P}(\mathbf{Y}, \neg toothache)$$

$$P(cavity, catch) = P(cavity, catch, toothache) + P(cavity, catch, \neg toothache)$$

$$= 0.108 + 0.072 = 0.18$$

Marginal probability of $Cavity$

Let $\mathbf{Y} = \{Cavity\}$ and $\mathbf{Z} = \{Catch, Toothache\}$

$$\mathbf{P}(\mathbf{Y}) = \mathbf{P}(\mathbf{Y}, catch, toothache) + \mathbf{P}(\mathbf{Y}, \neg catch, toothache) + \mathbf{P}(\mathbf{Y}, catch, \neg toothache) + \mathbf{P}(\mathbf{Y}, \neg catch, \neg toothache)$$

$$P(cavity) = P(cavity, catch, toothache) + P(cavity, \neg catch, toothache) + P(cavity, catch, \neg toothache) + P(cavity, \neg catch, \neg toothache)$$

$$= 0.108 + 0.012 + 0.072 + 0.008 = 0.2$$

# Conditionalization

Given the general marginalisation rule:

$$\boxed{\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y}, \mathbf{z})}$$

Applying the product rule to the right hand side results in the conditioning rule:

$$\boxed{\mathbf{P}(\mathbf{Y}) = \sum_{\mathbf{z}} \mathbf{P}(\mathbf{Y} \mid \mathbf{z}) * P(\mathbf{z})}$$

*Both are useful in all kind of derivations of probability expressions*

An example of conditioning:

Let $\mathbf{Y} = \{Cavity\}$ and $\mathbf{Z} = \{Toothache\}$

$$\mathbf{P}(\mathbf{Y}) = \mathbf{P}(\mathbf{Y} \mid toothache) * P(toothache) + \mathbf{P}(\mathbf{Y} \mid \neg toothache) * P(\neg toothache)$$

$$P(cavity) = P(cavity \mid toothache) * P(toothache) + P(cavity \mid \neg toothache) * P(\neg toothache)$$

# Computing Conditional Probabilities

The main form of inference with probabilities is to compute the probability of some variables given evidence of others.

*What is the probability I have a cavity given evidence I have a toothache?*

Unconditional Probabilities

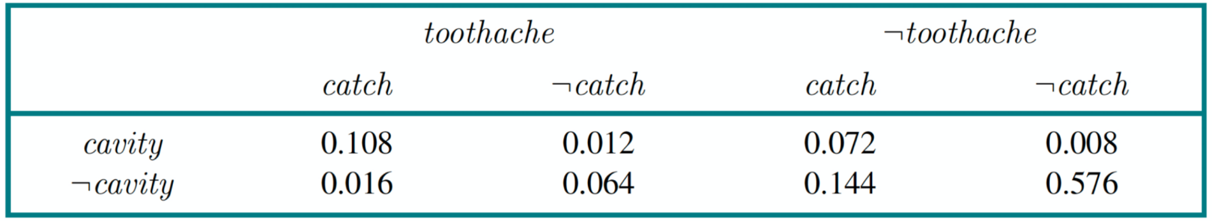
From the joint distribution

$$P(cavity \mid toothache) = \frac{P(cavity \land toothache)}{P(toothache)} = \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6$$

Marginalization

$$P(\neg cavity \mid toothache) = \frac{P(\neg cavity \land toothache)}{P(toothache)} = \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4$$

$$\mathbf{P}(Cavity \mid toothache) = \langle 0.6, 0.4 \rangle$$



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

# Normalization Constants

Given the conditional distribution: $\mathbf{P}(Cavity \mid toothache)$

$P(toothache)$ (in the denominator) can be viewed as a normalization constant to make sure the distribution adds up to 1.

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

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

$$= \alpha[\langle \overset{cavity}{0.108}, \overset{\neg cavity}{0.016}\rangle + \langle \overset{cavity}{0.012}, \overset{\neg cavity}{0.064}\rangle]$$

$$= \alpha\langle \overset{cavity}{0.12}, \overset{\neg cavity}{0.08}\rangle = \langle 0.6, 0.4\rangle$$

$$\alpha = \frac{1}{P(toothache)} = \frac{1}{0.12 + 0.08} = \frac{1}{0.2} = 5$$

Useful shortcut in many probability derivations. Can proceed when the denominator is unknown.

# A General Inference Procedure

Let $X$ be the query variable, $\mathbf{E}$ be the evidence variables, $\mathbf{e}$ be the observed values for them, $\mathbf{Y}$ be the remaining unobserved (hidden) variables and $\mathbf{y}$ be the exhaustive set of sequences of distinct variable/value pairs of the unobserved variables $\mathbf{Y}$.

Note that $\{X\} \cup \mathbf{E} \cup \mathbf{Y}$ is the set of all variables in the full joint distribution.

$$\mathbf{P}(X \mid \mathbf{e}) = \alpha * \mathbf{P}(X, \mathbf{e}) = \alpha * \sum_{\mathbf{y}} \underline{\mathbf{P}(X, \mathbf{e}, \mathbf{y})}$$

Subset of probabilities from the full joint distribution

# An Example

$$\boxed{\mathbf{P}(X \mid \mathbf{e}) = \alpha * \mathbf{P}(X, \mathbf{e}) = \alpha * \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})}$$

$$X = \{Cavity\}, \mathbf{E} = \{Toothache\}, \mathbf{e} = \{toothache\}, \mathbf{Y} = \{Catch\}, \mathbf{y} = \{\{catch\}\{\neg catch\}\}$$

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

$$= \alpha * \sum_{\mathbf{y}} \mathbf{P}(Cavity, toothache, \mathbf{y})$$

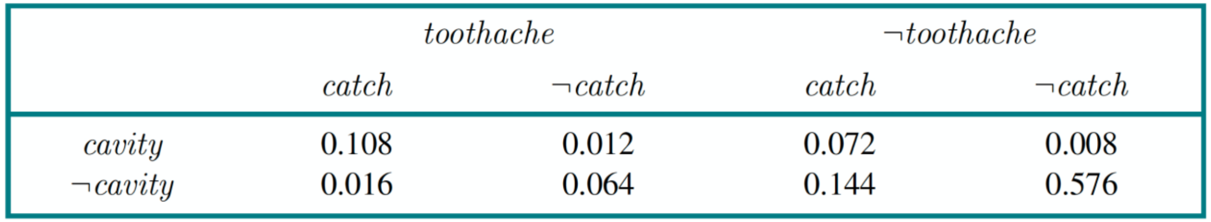
Marginalize

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

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

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

Normalize

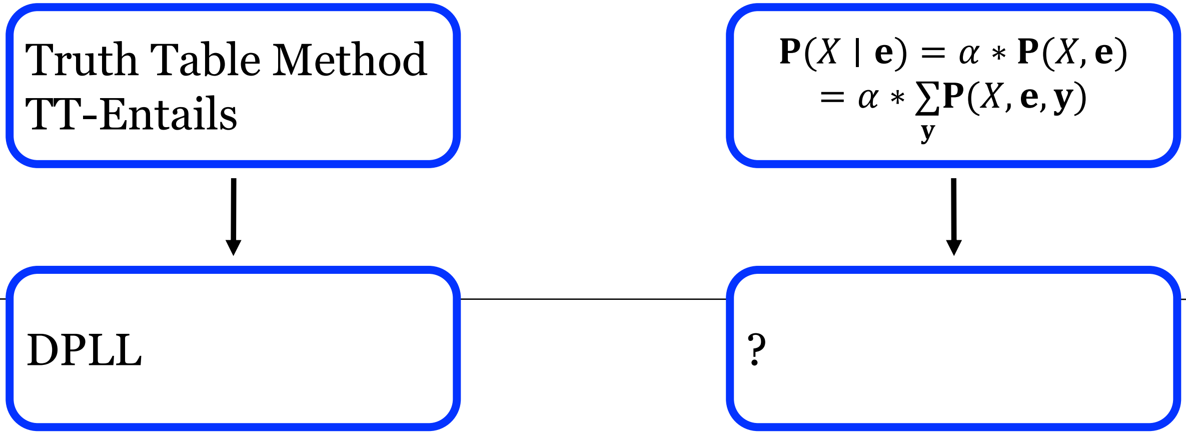


| | *toothache* | | *¬toothache* | |
|---|---|---|---|---|
| | *catch* | *¬catch* | *catch* | *¬catch* |
| *cavity* | 0.108 | 0.012 | 0.072 | 0.008 |
| *¬cavity* | 0.016 | 0.064 | 0.144 | 0.576 |

# Comments

$$\boxed{\mathbf{P}(X \mid \mathbf{e}) = \alpha * \mathbf{P}(X, \mathbf{e}) = \alpha * \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})}$$

- The equation above can serve as a basis for an implementation of an inference procedure.
- Unfortunately, it is not efficient:
  - It requires an input table for the full joint distribution. Assuming $n$ variables, this would require a table size of $\mathbf{O}(2^n)$ and $\mathbf{O}(2^n)$ time to run the algorithm.
- It could be viewed as the theoretical foundation for development of more efficient reasoning techniques.



Truth Table Method
TT-Entails

DPLL

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

?

# Independence

A standard problem-solving heuristic in any area is to break a larger problem up into smaller independent components

Divide and Conquer!

Suppose we extend out joint distribution $\mathbf{P}(Toothache, Catch, Cavity)$ with a new variable: $Weather: \{sunny, rainy, cloudy, snow\}$

$$\mathbf{P}(Toothache, Catch, Cavity, Weather) \qquad (2 * 2 * 2 * 4)$$

This would extend the joint distribution table from 8 to 32 values

Given any values of the 4 variables, the product rule tells us:

$$P(toothach, catch, cavity, Weather = cloudy) =$$

$$P(Weather = cloudy \mid toothach, catch, cavity) * P(toothach, catch, cavity)$$

# Independence

$$P(toothach, catch, cavity, Weather = cloudy) =$$

$$P(Weather = cloudy \mid toothach, catch, cavity) * P(toothach, catch, cavity)$$

It would be intuitively correct to assume that weather has nothing to do with dentistry!

$$P(Weather = cloudy \mid toothach, catch, cavity) = P(Weather = cloudy)$$

From this we can infer:

$$P(toothach, catch, cavity, Weather = cloudy) = P(Weather = cloudy) * P(toothach, catch, cavity)$$

More generally:

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

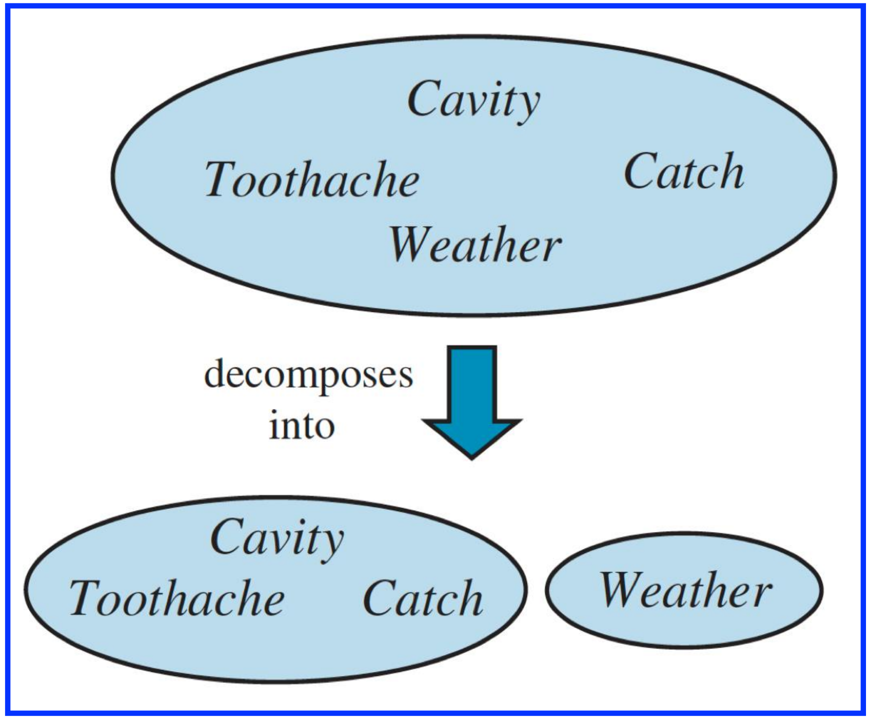
4 element table

8 element table

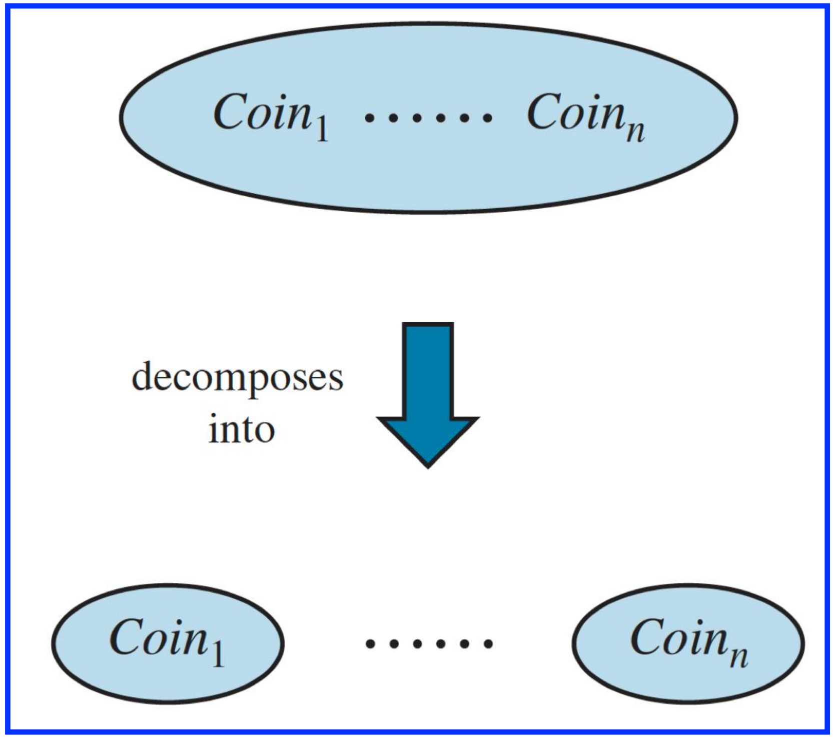
Via partitioning/independence the joint table can be specified using 12 parameters instead of 32.

Independence assumptions might be a basis for more efficient inference techniques!

# Factoring



Dentristry Domain



Coin Flipping Domain

Independence assertions can both reduce the size of the domain representation and make the inferencing problem more efficient.

# Absolute Independence

Independence between variables $X, Y$ can be written as follows:

$$P(X \mid Y) = P(X) \text{ or } P(Y \mid X) = P(Y) \text{ or } P(X, Y) = P(X) * P(Y)$$

- Independence assumptions are domain dependent
  - If the set of variables can be divided into independent subsets, then the full joint probability distribution can be factored into separate distributions on those subsets
- This in turn implies a reduction in the size of the domain representation and in the complexity of the inference problem

# Conditional Independence

The conditional independence of two variables $X$ and $Y$, given a third variable $Z$ is,

$$P(X, Y \mid Z) = P(X \mid Z) * P(Y \mid Z)$$

Equivalently,

$$P(X \mid Y, Z) = P(X \mid Z) \text{ and } P(Y \mid X, Z) = P(Y \mid Z)$$

Suppose $Toothache$ and $Catch$ are independent given $Cavity$, then

$$P(Toothache, Catch \mid Cavity) = P(Toothache \mid Cavity) * P(Catch \mid Cavity)$$

Each is directly caused by $Cavity$, but neither has a direct effect on the other

They are not absolutely independent because if a probe catches in a tooth, it probably has a cavity and that probably causes a toothache.

# More Comments

- Conditional independence assertions allow probabilistic systems to scale up by permitting compact representation of full joint distributions.
- This insight will be used to advantage with **Bayesian Networks**
- The decomposition of large probabilistic domains into weakly connected subsets via conditional independence assumptions is **one of the most important developments before the deep learning breakthrough**.

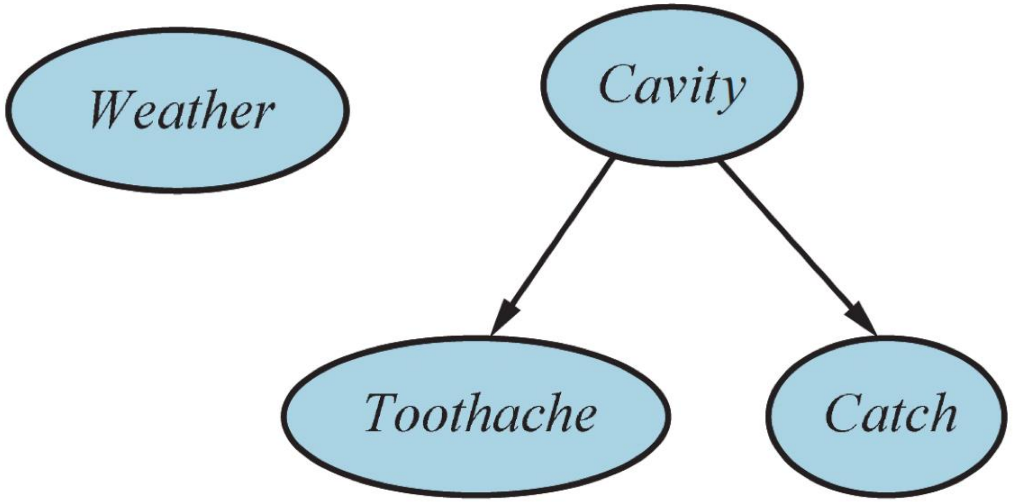
# Bayesian Networks

- Full joint probability distributions can answer any question about a modelled domain.
  - Intractably large as the number of variables grows
  - Specifying probabilities for all atomic events is difficult to do
- Independence and conditional independence assumptions greatly reduce the number of probabilities/parameters needed to be specified in order to define full joint probability distributions
- Bayesian Networks are data structures that represent dependencies among variables and give precise specifications of any full joint probability distribution in a concise manner.

# Bayesian Networks

- A Bayesian Network is a directed graph where each node is annotated with quantitative probability information:
  1. A set of random variables makes up the nodes in the network
  2. A set of directed arrows connects pairs of nodes. If there is an arrow from $X$ to $Y$, $X$ is said to be the parent of $Y$
  3. Each node $X_i$ has a conditional probability distribution $P(X_i \mid Parents(X_i))$ that quantifies the effect of the parents on the node
  4. The graph has no cycles. It is a DAG (directed, acyclic graph)

# An Example

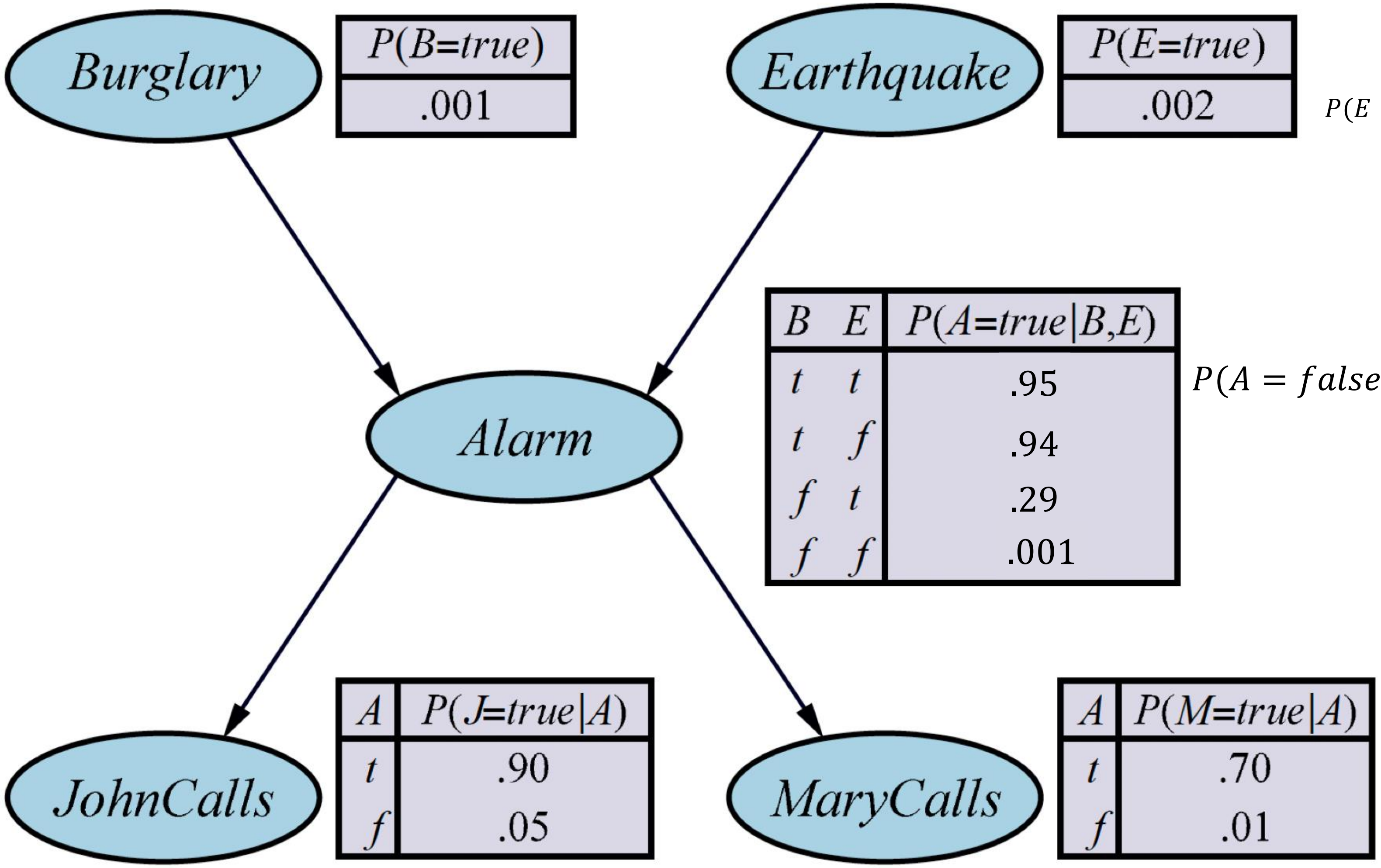


$Toothache$ and $Catch$ are conditionally independent of $Cavity$

$Weather$ is independent of the other three variables

# Another Example (J. Pearl)

- A person installs a new burglar alarm at home. It responds to burglaries, but may also respond to earthquakes on occasion.
- The person has two neighbors, John and Mary, who promise to call you at work when the alarm goes off.
  - John always calls when he hears the alarm, but sometimes confuses the telephone ringing with the alarm sound.
  - Mary, who likes loud music sometimes misses the alarm altogether
- Queries
  - Given evidence of who has or has not called, estimate the probability of a burglary:
    - P(burglary | john, mary)



| $P(B=true)$ |
|---|
| .001 |

| $P(E=true)$ |
|---|
| .002 |

$P(E = false) = 0.998$

| $B$ | $E$ | $P(A=true \mid B,E)$ |
|---|---|---|
| $t$ | $t$ | .95 |
| $t$ | $f$ | .94 |
| $f$ | $t$ | .29 |
| $f$ | $f$ | .001 |

$P(A = false \mid b, e) = 0.05$

| $A$ | $P(J=true \mid A)$ |
|---|---|
| $t$ | .90 |
| $f$ | .05 |

| $A$ | $P(M=true \mid A)$ |
|---|---|
| $t$ | .70 |
| $f$ | .01 |

Note: Conditional table for *Alarm*
is from 4th Ed (Global Edition).
In R&N:4th Ed (standard), It is different.

# Semantics of Bayesian Networks

We are interested in computing entries in the joint probability distribution:

$$P(X_1 = x_1 \land \cdots \land X_n = x_n) \text{ abbreviated } P(x_1, \dots, x_n)$$

This is defined as:

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

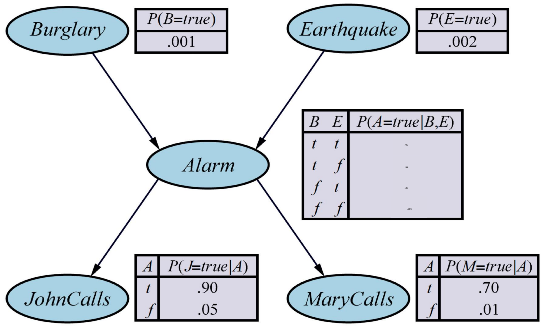
where $parents(X_i)$ denotes the specific values of variables in $Parents(X_i)$

For example, what is the probability that the alarm has sounded, but neither earthquake nor burglary has occurred and both John and Mary call?

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

$$= 0.998 * 0.999 * 0.001 * 0.70 * 0.90$$

$$= 0.00062811126 \approx 0.0006 \approx 0.06\%$$



# Constructing Bayesian Networks

The chain rule can be used to factor a joint distribution into a product of conditional distributions:

$$\mathbf{P}(X_1, \dots, X_n) = \mathbf{P}(X_n \mid X_{n-1}, \dots, X_1) * \mathbf{P}(X_{n-1} \mid X_{n-2}, \dots, X_1) * \cdots * \mathbf{P}(X_2 \mid X_1) * \mathbf{P}(X_1)$$

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

From the semantics of Bayesian Networks, we know:

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

In general:

$$\mathbf{P}(X_1, \dots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$

# Constructing Bayesian Networks

$$\text{Chain Rule: } \mathbf{P}(X_1, \dots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid X_{i-1}, \dots X_1)$$

$$\text{Semantics of BN: } \mathbf{P}(X_1, \dots, X_n) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid Parents(X_i))$$

From the above, for every variable $X_i$ in the network:

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

This is satisfied by ordering the nodes in topological order relative to graph structure:

$$X_1{:}\,Earthquake, X_2{:}\,Burglary, X_3{:}\,Alarm, X_4{:}\,MaryCalls, X_5{:}\,JohnCalls$$

Causes precede Effects

The Bayesian Network is a correct representation of the domain only if each node is conditionally independent of other predecessors in the node ordering, given its parents.

$$\boxed{X_i \perp\!\!\!\perp \{X_{i-1}, \dots X_1\} \setminus Parents(X_i) \mid Parents(X_i)}$$

# Exact Inference in Bayesian Networks

Let $X$ be the query variable, $\mathbf{E}$ be the evidence variables, $\mathbf{e}$ be the observed values for them, $\mathbf{Y}$ be the remaining unobserved (hidden) variables and $\mathbf{y}$ be The exhaustive set of sequences of distinct variable/value pairs of the unobserved variables $\mathbf{Y}$.

Note that $\{X\} \cup \mathbf{E} \cup \mathbf{Y}$ is the set of all variables in the full joint distribution.

$$\mathbf{P}(X \mid \mathbf{e}) = \alpha * \mathbf{P}(X, \mathbf{e}) = \alpha * \sum_{\mathbf{y}} \underline{\mathbf{P}(X, \mathbf{e}, \mathbf{y})}$$

Subset of probabilities from the full joint distribution

We know that the terms $\mathbf{P}(X, \mathbf{e}, \mathbf{y})$ in the joint distribution can be written as products of conditional probabilities from the network. So, a query is answered by computing the sums of products of conditional probabilities from the network.

# An Inference Example

$$\boxed{\mathbf{P}(X \mid \mathbf{e}) = \alpha * \mathbf{P}(X, \mathbf{e}) = \alpha * \sum_{\mathbf{y}} \mathbf{P}(X, \mathbf{e}, \mathbf{y})}$$

$$\boxed{\begin{array}{l} X = \{Burglary\} \\ \mathbf{E} = \{JohnCalls, MaryCalls\} \\ \mathbf{e} = \{johncalls, marycalls\} \\ \mathbf{Y} = \{Earthquake, Alarm\} \end{array}}$$

Query: $\mathbf{P}(Burglary \mid johncalls, marycalls)$

$$\mathbf{P}(Burglary \mid johncalls, marycalls) = \alpha \mathbf{P}(Burglary, johncalls, marycalls)$$

$$= \alpha \sum_{\mathbf{y}} \mathbf{P}(Burglary, johncalls, marycalls, \mathbf{y})$$

$$= \alpha \sum_{\mathbf{e}} \sum_{\mathbf{a}} \mathbf{P}(Burglary, johncalls, marycalls, \mathbf{e}, \mathbf{a})$$

$$= \alpha[\mathbf{P}(B, j, m, e, a) + \mathbf{P}(B, j, m, e, \neg a) + \mathbf{P}(B, j, m, \neg e, a) + \mathbf{P}(B, j, m, \neg e, \neg a)]$$

$$= \alpha \sum_{\mathbf{e}} \sum_{\mathbf{a}} \mathbf{P}(Burglary, johncalls, marycalls, \mathbf{e}, \mathbf{a})$$

$$= \alpha[\mathbf{P}(B, j, m, e, a) + \mathbf{P}(B, j, m, e, \neg a) + \mathbf{P}(B, j, m, \neg e, a) + \mathbf{P}(B, j, m, \neg e, \neg a)]$$

$$\boxed{\mathbf{P}(B, j, m, e, a) = \langle P(b, j, m, e, a), P(\neg b, j, m, e, a)\rangle = \langle 1.197 * 10^{-6}, 0.0003650346\rangle}$$

$$P(b, j, m, e, a) = P(e) * P(b) * P(a \mid e, b) * P(m \mid a) * P(j \mid a)$$
$$= 0.002 * 0.001 * 0.95 * 0.70 * 0.90 = 1.197 * 10^{-6}$$
$$P(\neg b, j, m, e, a) = P(e) * P(\neg b) * P(a \mid e, \neg b) * P(m \mid a) * P(j \mid a)$$
$$= 0.002 * 0.999 * 0.29 * 0.70 * 0.90 = 0.0003650346$$

$$\boxed{\mathbf{P}(B, j, m, e, \neg a) = \langle P(b, j, m, e, \neg a), P(\neg b, j, m, e, \neg a)\rangle = \langle 5 * 10^{-11}, 7.1 * 10^{-10}\rangle}$$

$$P(b, j, m, e, \neg a) = P(e) * P(b) * P(\neg a \mid e, b) * P(m \mid \neg a) * P(j \mid \neg a)$$
$$= 0.002 * 0.001 * 0.05 * 0.01 * 0.05 = 5 * 10^{-11}$$
$$P(\neg b, j, m, e, \neg a) = P(e) * P(\neg b) * P(\neg a \mid e, \neg b) * P(m \mid \neg a) * P(j \mid \neg a)$$
$$= 0.002 * 0.001 * 0.71 * 0.01 * 0.05 = 7.1 * 10^{-10}$$

$$\boxed{\mathbf{P}(B, j, m, \neg e, a) = \langle P(b, j, m, \neg e, a), P(\neg b, j, m, \neg e, a)\rangle = \langle 0.0005910156, 0.00062811126\rangle}$$

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

$$= 0.998 * 0.001 * 0.94 * 0.70 * 0.90 = 0.0005910156$$

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

$$= 0.998 * 0.999 * 0.001 * 0.70 * 0.90 = 0.0062811126$$

$$\boxed{\mathbf{P}(B, j, m, \neg e, \neg a) = \langle P(b, j, m, \neg e, \neg a), P(\neg b, j, m, \neg e, \neg a)\rangle = \langle 2.99 * 10^{-8}, 0.00049351599\rangle}$$

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

$$= 0.998 * 0.001 * 0.06 * 0.01 * 0.05 = 2.99 * 10^{-8}$$

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

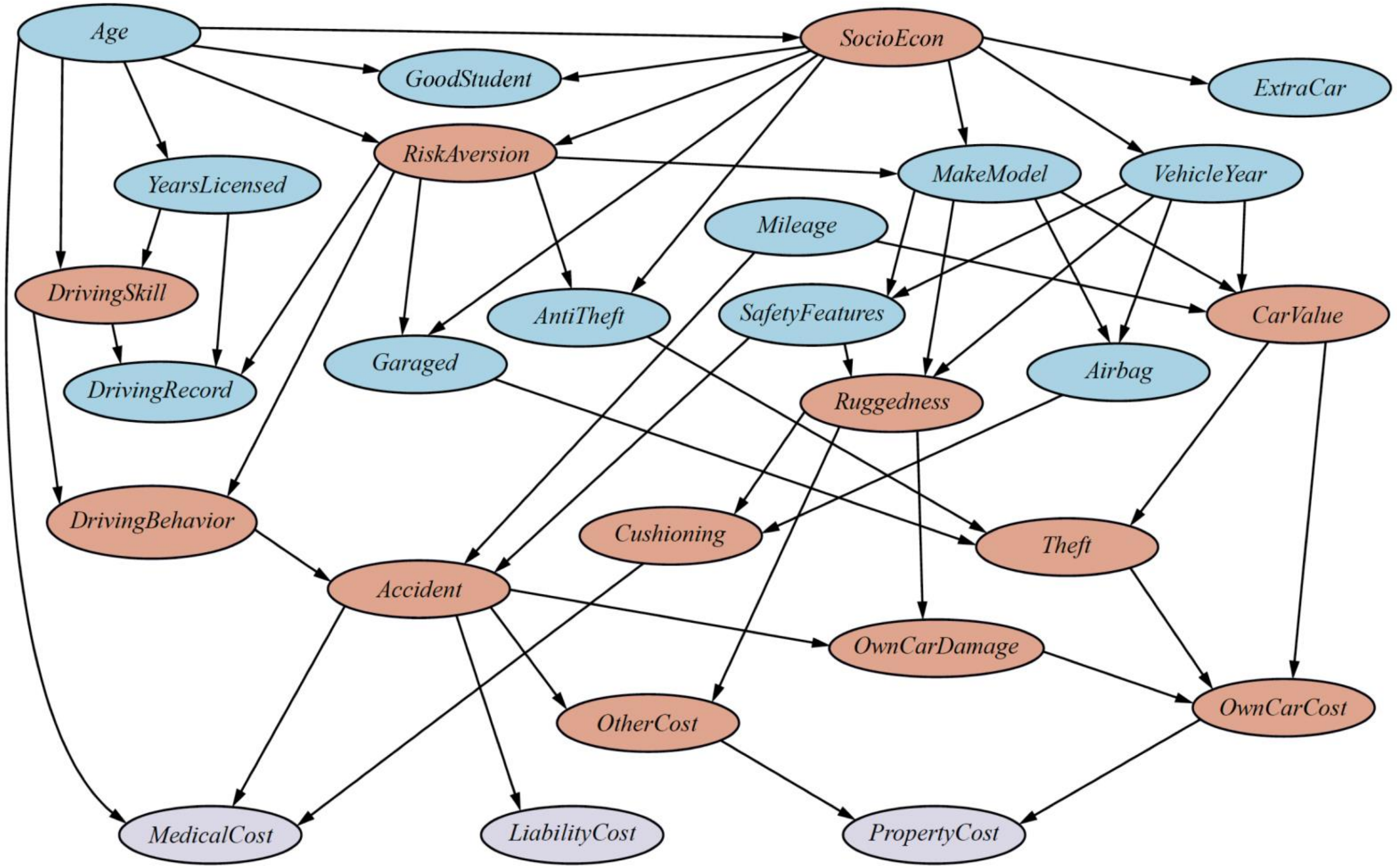
$$= 0.998 * 0.999 * 0.99 * 0.01 * 0.05 = 0.00049351599$$

$$\boxed{\alpha[\langle 1.197 * 10^{-6}, 0.0003650346\rangle + \langle 5 * 10^{-11}, 7.1 * 10^{-10}\rangle + \langle 0.0005910156, 0.00062811126\rangle + \langle 2.99 * 10^{-8}, 0.00049351599\rangle}$$

$$\alpha[\langle 0.0006032851, 0.001486669\rangle] = \langle 0.288659, 0.711340\rangle$$

28.9%chance of a burglary. An increase from the prior chance of 0.1%

# Real World Examples



Evaluating Car Insurance Applications

# Bayes' Rule

The product rule states: $P(x, y) = P(x \mid y) * P(y) = P(y \mid x) * P(x)$

From this we can derive:

$$P(y \mid x) = \frac{P(x \mid y) * P(y)}{P(x)}$$

The more general case for multi-valued variables:

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

A generalized version conditionalized on some evidence $\mathbf{e}$:

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

# Bayes' Rule (Applications)

Bayes' Rule has widespread applications:

Scientific Theories

$$\mathbf{P}(Hypothesis \mid Evidence) = \frac{\mathbf{P}(Evidence \mid Hypothesis) * \mathbf{P}(Hypothesis)}{\mathbf{P}(Evidence)}$$

Causal Reasoning

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

Diagnosis

$$\mathbf{P}(Disease \mid Symptoms) = \frac{\mathbf{P}(Symptoms \mid Disease) * \mathbf{P}(Disease)}{\mathbf{P}(Symptoms)}$$

# Intuitions

$$\mathbf{P}(Hypothesis \mid Evidence) = \frac{\mathbf{P}(Evidence \mid Hypothesis) * \mathbf{P}(Hypothesis)}{\mathbf{P}(Evidence)}$$

Given a prior probability for a hypothesis, $\mathbf{P}(Hypothesis)$, upon receiving new evidence, where its prior probability has already been given, $\mathbf{P}(Evidence)$, what is my revised belief for the hypothesis in the context of the new evidence: $\mathbf{P}(Hypothesis \mid Evidence)$

$\mathbf{P}(Hypothesis)$, is called the prior probability for the hypothesis and
$\mathbf{P}(Hypothesis \mid Evidence)$, is called the posterior probability for the hypothesis

# An Example: Diagnosis

Doctors often know how many patients with a given disease exhibit various symptoms:

$$\mathbf{P}(StiffNeck \mid Meningitis) = 0.5$$

Doctors generally also know some unconditional facts:

$$\mathbf{P}(Meningitis) = \frac{1}{50{,}000}, \mathbf{P}(StiffNeck) = \frac{1}{20}$$

What is the probability a patient has Meningitis given evidence of a stiff neck?

$$\boxed{\mathbf{P}(Disease \mid Symptoms) = \frac{\mathbf{P}(Symptoms \mid Disease) * \mathbf{P}(Disease)}{\mathbf{P}(Symptoms)}}$$

$$\mathbf{P}(Meningitis \mid StiffNeck) = \frac{\mathbf{P}(StiffNeck \mid Meningitis) * \mathbf{P}(Meningitis)}{\mathbf{P}(StiffNeck)}$$

$$= \frac{0.5 * \frac{1}{50{,}000}}{\frac{1}{20}} = 0.0002 = \frac{1}{5000}$$

A marked increase from $\frac{1}{50{,}000}$

# Many Pieces of Evidence/ Normalization

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

$$\mathbf{P}(Meningitis \mid StiffNeck, SwollenBrain) = \frac{\mathbf{P}(StiffNeck \mid Meningitis, SwollenBrain) * \mathbf{P}(Meningitis \mid SwollenBrain)}{\mathbf{P}(StiffNeck \mid SwollenBrain)}$$

Normalized Baye’s Rule

To avoid assessing the evidence (denominator):

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

where $\alpha = \frac{1}{\mathbf{P}(X)} = \frac{1}{\sum_{\mathbf{y}} \mathbf{P}(X|\mathbf{y}) * \mathbf{P}(\mathbf{y})}$

All entries $\mathbf{P}(Y \mid X)$ should sum to 1.

# Naive Bayes Models

Suppose we have a model with single $Cause$ that influences many $Effects$:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n)$$

We *assume* the $Effect's$ variables are independent of each other given $Cause$

or one $Disease$ that has many $Symptoms$:

$$\mathbf{P}(Disease, Symptom_1, \ldots, Symptom_n)$$

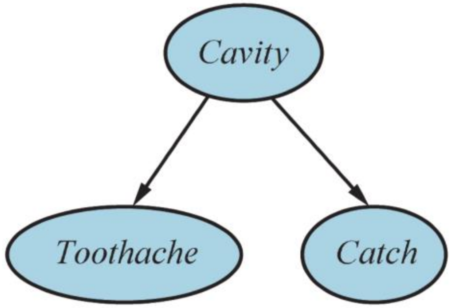
Due to this independence, we can derive:

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) * \prod_{i=1}^{n} \mathbf{P}(Effect_i \mid Cause)$$

# An Example

$$\mathbf{P}(Cause, Effect_1, \ldots, Effect_n) = \mathbf{P}(Cause) * \prod_{i=1}^{n} \mathbf{P}(Effect_i \mid Cause)$$

We know that: $\boxed{Toothache \perp\!\!\!\perp Catch \mid Cavity}$



$$\mathbf{P}(Cavity, Toothache, Catch) = \mathbf{P}(Cavity) * \prod_{i=1}^{n} \mathbf{P}(Effect_i \mid Cavity)$$

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

Naive Bayes modeling is used even when there are dependencies among effects due to its efficiency and correctness of output

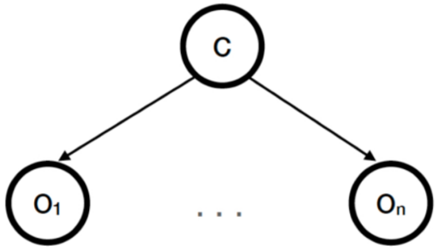
# Naive Bayes: Classification

We are often in situations where we would like to classify something given a set of observations ( features, attributes) about that something.

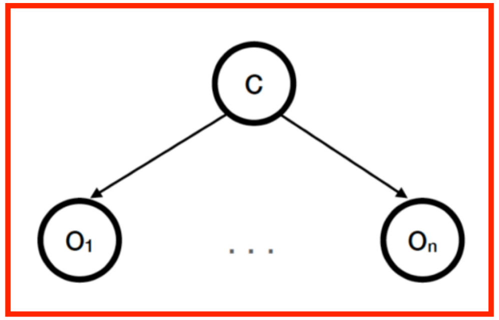
Given a set of random variables $O_1, \ldots, O_n$ representing a set of observations and a random variable $C$ representing classes, we are interested in the joint probability distribution $\mathbf{P}(C, O_1, \ldots, O_n)$, and in particular, a way to compute $\mathbf{P}(C \mid O_{1:n})$

Independence assumption

$\boxed{\forall i, j : i \neq j . O_i \perp\!\!\!\perp O_j \mid C}$



# Naive Bayes: Classification



$$\mathbf{P}(C, O_{1:n}) = \mathbf{P}(C) \prod_{i=1}^{n} \mathbf{P}(O_i \mid C)$$

Naive Bayes

$$\mathbf{P}(C \mid O_{1:n}) = \frac{\mathbf{P}(C, O_{1:n})}{\mathbf{P}(O_{1:n})} = \frac{\mathbf{P}(C, O_{1:n})}{\sum_c \mathbf{P}(c, O_{1:n})} = \alpha \mathbf{P}(C, O_{1:n})$$

Conditional

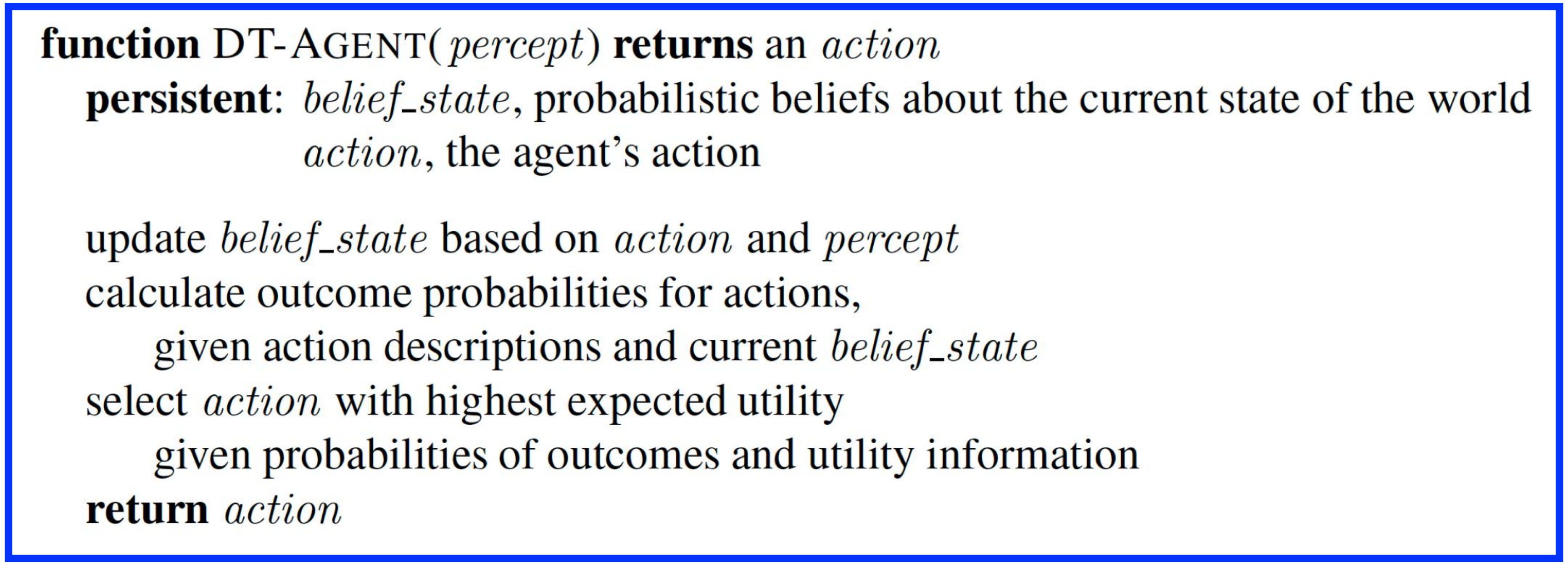
$$\mathbf{P}(C \mid O_{1:n}) = \alpha \mathbf{P}(C, O_{1:n})$$

Normalization

$$\boxed{\mathbf{P}(C, \mid O_{1:n}) = \alpha \mathbf{P}(C) \prod_{i=1}^{n} \mathbf{P}(O_i \mid C)}$$

Substitution

# A Decision Theoretic Agent



```
function DT-AGENT(percept) returns an action
  persistent: belief_state, probabilistic beliefs about the current state of the world
              action, the agent's action

  update belief_state based on action and percept
  calculate outcome probabilities for actions,
     given action descriptions and current belief_state
  select action with highest expected utility
     given probabilities of outcomes and utility information
  return action
```

Belief state is a probability distribution on possible worlds

Principle of Maximum Expected Utility (MEU)

An agent chooses the action that yields the highest expected utility, averaged over all possible outcomes of the action

# Choosing the "right" Action

What is my expected utility/goodness when executing an action?

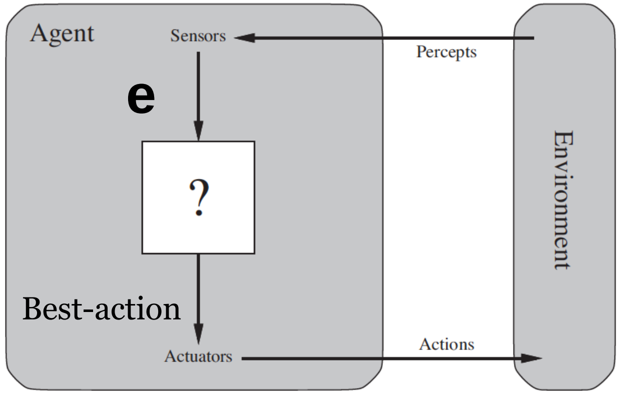
$$EU(\text{action} \mid \mathbf{e}) = \sum_{state'} P(result(\text{action}) = state' \mid \text{action}, \mathbf{e})\, U(state')$$

*Take the weighted average of the utilities for states an action can cause*

Choose: The action that maximises utility in any context!

$$\boxed{best{-}action = argmax_{action} EU(action \mid \mathbf{e})}$$

Decision Theoretic Agent



# TDDC17 AI LE11 HT2023: Reasoning with Uncertainty Bayesian Networks

www.ida.liu.se/~TDDC17

LINKÖPING
UNIVERSITY