

# Artificial Intelligence

## Uncertainty

Christopher Simpkins

Kennesaw State University



# Acting Under Uncertainty

Logical agents maintain a belief state and generate contingency plans that account for every possibility. This breaks down for nontrivial problems due to:

- ▶ Exhaustivity of belief state, which must contain all possible states, even unlikely states.
- ▶ Exhaustivity of contingency plan, which must account for every possible, however unlikely, action outcome.
- ▶ Unsatisfiability. There may be no guaranteed plan to achieve the goal. If we must act anyway, how do we choose the best plan?
- ▶ Qualification problem. The closed-world assumption allows us to simplify logical environment specifications for simple domains, but real-world domains contain far more detail which must be accounted for.

# From Logic to Probability Theory

Consider:

$$\textit{Toothache} \implies \textit{Cavity}$$

Not all patients with toothaches have cavities. How about:

$$\textit{Toothache} \implies \textit{Cavity} \vee \textit{GumProblem} \vee \textit{Abscess} \dots$$

We'd need large list after the dots. Try causal direction:

$$\textit{Cavity} \implies \textit{Toothache}$$

But not all cavities hurt. Logic fails to deal with complex domains like medical diagnosis due to:

- ▶ **Exhaustivity** (laziness). Too many antecedents or consequents to list.
- ▶ **Theoretical ignorance**. Rare to have complete logical theory of nontrivial domain.
- ▶ **Practical ignorance**. Even with a complete domain theory, hard to measure all the necessary inputs (medical tests, etc.)

We need to deal with uncertain knowledge, with **degrees of belief**. For that, we use probability theory.

# Probability Theory as Knowledge Representation

- ▶ Ontological commitments of logic and probability theory same:
  - ▶ World composed of facts that do or do not hold in any particular case.
- ▶ Epistemological commitments differ:
  - ▶ Logic assigns true, false, or no opinion to each sentence.
  - ▶ Probability theory assigns numerical degree of belief between 0 (for sentences that are certainly false) and 1 (certainly true) to each sentence.

Probability theory solves the qualification problem by summarizing the uncertainty stemming from our laziness and ignorance.

# What do probabilities mean?



## Probabilities are about our uncertain knowledge.

“80% chance (probability 0.8) patient with a toothache has a cavity.”

- ▶ Out of all the situations that are indistinguishable from the current situation *as far as our knowledge goes*, the patient will have a cavity in 80% of them.
- ▶ Belief could come from statistical data, or domain theory, or combination of sources.

*There is no uncertainty in the actual world. The patient either has a cavity or does not. Probabilities refer to our knowledge of the world state, not the actual world state.*

If our knowledge changes, e.g., we find out patient has history of gum disease, we make a different statement.

“Given patient has a history of gum disease and a toothache, 40% chance patient has gum disease.”

# Rational Decisions

How do we choose between different plans that achieve the goal?

- ▶ Each plan has a likelihood of success, e.g.:
  - ▶ Plan A has a 95% chance of achieving a goal state.
- ▶ Each plan may lead to goal states with different outcomes, each of which is a goal state.
  - ▶ An **outcome** of an action is a completely specified state, which includes elements that are not part of the goal.

Elements of outcomes can be more or less desirable – more or less *useful* – to a given agent.

- ▶ Utility is the quality of being useful.
- ▶ That “usefulness” may simply be pleasure, or even altruism.

Utility theory associates a utility value with each possible outcome, which induces a preference ordering over outcomes.

*An agent is rational if and only if it chooses the action that yields the highest expected utility, averaged over all the possible outcomes of the action.*

# Decision-Theoretic Agents

*Decision theory = probability theory + utility theory*

**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*

Each action,  $a$ , leads to a probability distribution over outcomes, or “result states”:

$$P(\text{RESULT}(a) = s') = \sum_{s \in S} P(s)P(s' | s, a)$$

And each outcome state has a utility. A rational decision-theoretic agent chooses the action that maximize expected utility, that is:

$$\operatorname{argmax}_a \sum_{s' \in S} P(\text{RESULT}(a) = s')U(s')$$

# Probability Models

In logic we have satisfiability:  $M(\alpha)$  is the set of all possible worlds in which sentence  $\alpha$  is true.

In probability theory the set of possible worlds is a **sample space**,  $\Omega$ , which must be mutually exclusive and exhaustive (one and only one possible world must be the case).

$$0 \leq P(\omega) \leq 1 \text{ for every } \omega \text{ and } \sum_{\omega \in \Omega} P(\omega) = 1$$

Each  $\omega$  is a possible world. Probability textbooks typically call these outcomes. For example, each of these pairs is a  $\omega \in \Omega$  for the roll of two dice:

(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 1)	(6, 1)
(1, 2)	(2, 2)	(3, 2)	(4, 2)	(5, 2)	(6, 2)
(1, 3)	(2, 3)	(3, 3)	(4, 3)	(5, 3)	(6, 3)
(1, 4)	(2, 4)	(3, 4)	(4, 4)	(5, 4)	(6, 4)
(1, 5)	(2, 5)	(3, 5)	(4, 5)	(5, 5)	(6, 5)
(1, 6)	(2, 6)	(3, 6)	(4, 6)	(5, 6)	(6, 6)

In probability textbooks, an **experiment** leads to a sample space. Here, the set of possible worlds comes from the task environment.

## Propositions and Events

An **event**, which we denote with  $\phi$  here, is a set of possible worlds, some subset of  $\Omega$ , or set of  $\omega$ , e.g., the event “doubles” contains the 6 boxed elements  $\omega$ :

$(1, 1)$	$(2, 1)$	$(3, 1)$	$(4, 1)$	$(5, 1)$	$(6, 1)$
$(1, 2)$	$(2, 2)$	$(3, 2)$	$(4, 2)$	$(5, 2)$	$(6, 2)$
$(1, 3)$	$(2, 3)$	$(3, 3)$	$(4, 3)$	$(5, 3)$	$(6, 3)$
$(1, 4)$	$(2, 4)$	$(3, 4)$	$(4, 4)$	$(5, 4)$	$(6, 4)$
$(1, 5)$	$(2, 5)$	$(3, 5)$	$(4, 5)$	$(5, 5)$	$(6, 5)$
$(1, 6)$	$(2, 6)$	$(3, 6)$	$(4, 6)$	$(5, 6)$	$(6, 6)$

A **proposition** is an event expressed in a formal language; specifically, for each proposition, the corresponding set contains just those possible worlds in which the proposition holds.

The probability associated with a proposition is defined to be the sum of the probabilities of the worlds in which it holds:

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

$$\text{Example: } P(\text{Doubles}) = P((1, 1)) + \dots + P((6, 6)) = \frac{1}{36} + \dots + \frac{1}{36} = \frac{1}{6}$$

## Prior and Conditional Probabilities

An **unconditional** or **prior** probability is a degree of belief in a proposition with no other information.

A **conditional** or **posterior** probability is a degree of belief in a proposition given some relevant information.

Definition of conditional probability:

$$P(a | b) = \frac{P(a \wedge b)}{P(b)}$$

From that definition we get the **product rule**:

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

Note that  $P(a \wedge b)$  can also be written  $P(a, b)$  or  $P(ab)$ .

## Conditional probability is not implication.

The assertion

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

does not mean “Whenever toothache is true, conclude that cavity is true with probability 0.6”

It means “Whenever toothache is true *and we have no further information*, conclude that cavity is true with probability 0.6.”

If we had the further information that the dentist found no cavities, certainly not the case that cavity is true with probability 0.6; instead we have

$$P(\text{cavity} \mid \text{toothache} \wedge \neg \text{cavity}) = 0$$

## Probability Assertions

A **random variable**, which begins with capital letter, is a function mapping from possible worlds  $\Omega$  to a set of possible values the variable can take.

- ▶  $Die_1 : \{1, \dots, 6\} \rightarrow \{1, \dots, 6\}$
- ▶  $Doubles : \{1, \dots, 6\} \times \{1, \dots, 6\} \rightarrow \{true, false\}$
- ▶  $Weather : \{sun, rain, cloud, snow\} \rightarrow \{sun, rain, cloud, snow\}$

Individual values are written in lowercase and often abbreviated.

- ▶  $P(sun)$  stands for  $P(Weather = sun)$ .

A **probability distribution** is an assignment of probabilities to all values of a random variable, e.g.:

$$P(Weather = sun) = 0.6$$

$$P(Weather = rain) = 0.1$$

$$P(Weather = cloud) = 0.29$$

$$P(Weather = snow) = 0.01$$

Which can also be written  $P(Weather) = \langle 0.6, 0.1, 0.29, 0.01 \rangle$ .

## Joint Probability Distributions

$P(\textit{Weather}, \textit{Cavity})$  denotes the probabilities of all combinations of *Weather* and *Cavity*. This notation is a compact representation of a **joint probability distribution**. The notation

$$P(\textit{Weather}, \textit{Cavity}) = P(\textit{Weather} \mid \textit{Cavity})P(\textit{Cavity})$$

stands for the  $4 \times 2 = 8$  equations:

$$P(W = \textit{sun} \wedge C = \textit{true}) = P(W = \textit{sun} \mid C = \textit{true})P(C = \textit{true})$$

$$P(W = \textit{rain} \wedge C = \textit{true}) = P(W = \textit{rain} \mid C = \textit{true})P(C = \textit{true})$$

$$P(W = \textit{cloud} \wedge C = \textit{true}) = P(W = \textit{cloud} \mid C = \textit{true})P(C = \textit{true})$$

$$P(W = \textit{snow} \wedge C = \textit{true}) = P(W = \textit{snow} \mid C = \textit{true})P(C = \textit{true})$$

$$P(W = \textit{sun} \wedge C = \textit{false}) = P(W = \textit{sun} \mid C = \textit{false})P(C = \textit{false})$$

$$P(W = \textit{rain} \wedge C = \textit{false}) = P(W = \textit{rain} \mid C = \textit{false})P(C = \textit{false})$$

$$P(W = \textit{cloud} \wedge C = \textit{false}) = P(W = \textit{cloud} \mid C = \textit{false})P(C = \textit{false})$$

$$P(W = \textit{snow} \wedge C = \textit{false}) = P(W = \textit{snow} \mid C = \textit{false})P(C = \textit{false})$$

# Probability Axioms

All of probability theory can be built from **Kolmogorov's axioms**.

Law of normalization:

$$0 \leq P(\omega) \leq 1 \text{ for every } \omega \text{ and } \sum_{\omega \in \Omega} P(\omega) = 1$$

Probability of a disjunction, **inclusion-exclusion principle**:

$$P(a \vee b) = P(a) + P(b) - P(a \wedge b)$$

# Why is probability theory a valid basis for rational behavior?

De Finetti's Theorem:

*If Agent 1 expresses a set of degrees of belief that violate the axioms of probability theory then there is a combination of bets by Agent 2 that guarantees that Agent 1 will lose money every time.*

Example:

Proposition	Agent 1's belief	Agent 2 bets	Agent 1 bets	Agent 1 payoffs for each outcome			
				$a, b$	$a, \neg b$	$\neg a, b$	$\neg a, \neg b$
$a$	0.4	\$4 on $a$	\$6 on $\neg a$	-\$6	-\$6	\$4	\$4
$b$	0.3	\$3 on $b$	\$7 on $\neg b$	-\$7	\$3	-\$7	\$3
$a \vee b$	0.8	\$2 on $\neg(a \vee b)$	\$8 on $a \vee b$	\$2	\$2	\$2	-\$8
				-\$11	-\$1	-\$1	-\$1

## Inference Using Full Joint Distributions

Knowledge base is full joint distribution of boolean random variables  $Toothache, Cavity, Catch$ .

	$toothache$		$\neg toothache$	
	$catch$	$\neg catch$	$catch$	$\neg catch$
$cavity$	0.108	0.012	0.072	0.008
$\neg cavity$	0.016	0.064	0.144	0.576

- ▶ All probabilities above sum to 1.
- ▶  $P(cavity \vee toothache) = 0.108 + 0.012 + 0.072 + 0.008 + 0.016 + 0.064 = 0.28$ .

Unconditional or **marginal probability** of cavity:

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

## Marginalization and Conditioning

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	0.108	0.012	0.072	0.008
$\neg$ <i>cavity</i>	0.016	0.064	0.144	0.576

**Marginalization** means “summing out” the non-query variables, which we usually denote with  $Z$ :

$$P(Y) = \sum_z P(Y, Z = z)$$

For example, let  $Y = \text{Cavity}$ :

$$\begin{aligned} P(\text{Cavity}) &= P(\text{Cavity}, \text{toothache}, \text{catch}) + P(\text{Cavity}, \text{toothache}, \neg\text{catch}) \\ &\quad + P(\text{Cavity}, \neg\text{toothache}, \text{catch}) + P(\text{Cavity}, \neg\text{toothache}, \neg\text{catch}) \\ &= \langle 0.108, 0.016 \rangle + \langle 0.012, 0.064 \rangle + \langle 0.072, 0.144 \rangle + \langle 0.008, 0.576 \rangle \\ &= \langle 0.2, 0.8 \rangle \end{aligned}$$

Using the product rule in  $P(Y) = \sum_z P(Y, Z = z)$  we obtain the **conditioning** rule:

$$P(Y) = \sum_z P(Y | z)P(z)$$

These two rules are used widely in probabilistic reasoning.

## Conditional Probability Distributions

We're usually interested in conditional probabilities. To get the conditional probabilities, we use the definition:

$$\begin{aligned}P(\text{cavity} \mid \text{toothache}) &= \frac{P(\text{cavity} \wedge \text{toothache})}{P(\text{toothache})} \\ &= \frac{0.108 + 0.012}{0.108 + 0.012 + 0.016 + 0.064} = 0.6\end{aligned}$$

$$\begin{aligned}P(\neg\text{cavity} \mid \text{toothache}) &= \frac{P(\neg\text{cavity} \wedge \text{toothache})}{P(\text{toothache})} \\ &= \frac{0.016 + 0.064}{0.108 + 0.012 + 0.016 + 0.064} = 0.4\end{aligned}$$

## Normalization

Notice that  $P(\textit{toothache})$  appears in denominator in both equations in the preceding conditional probability calculations. We can view it as a **normalization constant**, denoted  $\alpha$ , that ensures that the conditional probability distribution  $P(\textit{Cavity} \mid \textit{toothache})$  sums to 1. Then we can write:

$$\begin{aligned}P(\textit{Cavity} \mid \textit{toothache}) &= \alpha P(\textit{Cavity}, \textit{toothache}) \\ &= \alpha [P(\textit{Cavity}, \textit{toothache}, \textit{catch}) + P(\textit{Cavity}, \textit{toothache}, \neg \textit{catch})] \\ &= \alpha \langle 0.108, 0.16 \rangle + \langle 0.012, 0.064 \rangle \\ &= \alpha \langle 0.12, 0.08 \rangle \\ &= \langle 0.6, 0.4 \rangle\end{aligned}$$

The cool thing is that we don't even have to know  $P(\textit{toothache})$  to calculate the conditional probability distribution  $P(\textit{Cavity} \mid \textit{toothache})$ . To make this point clearer, imagine the penultimate line above is:

$$P(\textit{Cavity} \mid \textit{toothache}) = \alpha \langle 0.3, 0.2 \rangle$$

What's  $\alpha$ ?

## Inference Using the Full Joint Distribution

All of the preceding can be summarized in a general inference procedure using full joint probability distributions. Let  $E$  be a list of evidence variables,  $e$  be a list of observed values for the evidence variables, and  $Y$  be the remaining unobserved variables. Then:

$$P(X | e) = \alpha P(X, e) = \alpha \sum_y P(X, e, \mathbf{y}) \quad (12.9)$$

Great, so we're done! All we need is a full joint probability distribution and we can answer any query. Unfortunately, not practical.

- ▶ For a domain of  $n$  boolean variables, we have a table of size  $O(2^n)$

So we need different approaches, which we cover next.

# Independence

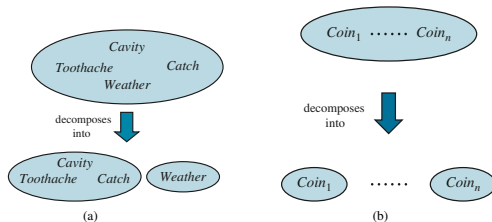
Weather is not affected by teeth. So we can assert:

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

In general, if  $a$  and  $b$  are independent:

$$P(a \mid b) = P(a) \text{ or } P(b \mid a) = P(b) \text{ or } P(a \wedge b) = P(a)P(b)$$

Can be a big help. For example, if you have  $n$  independent coin flips, then instead of  $2^n$  full joint table, you have a product of  $n$  distributions  $P(C_i)$ .



Unfortunately, independence rarely holds in the real world.

## Bayes' Rule

Using the symmetry  $P(X, Y) = P(Y, X)$  and the product rule we can derive **Bayes' rule**:

$$\begin{aligned}P(X, Y) &= P(Y, X) \\P(Y | X)P(X) &= P(X | Y)P(Y) \\P(Y | X) &= \frac{P(X | Y)P(Y)}{P(X)}\end{aligned}$$

The usefulness of Bayes' rule becomes apparent if we consider  $X$  as an effect and  $Y$  as a cause and we want to determine the cause of some effect (evidence) we observe:

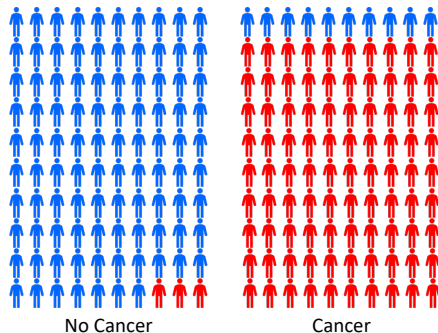
$$P(\text{cause} | \text{effect}) = \frac{P(\text{effect} | \text{cause})P(\text{cause})}{P(\text{effect})}$$

- ▶  $P(\text{effect} | \text{cause})$  quantifies the **causal** direction.
- ▶  $P(\text{cause} | \text{effect})$  quantifies the **diagnostic** direction.

Reasoning from effects to causes is also called **abductive reasoning**. (Is Sherlock Holmes truly employing deduction?)

## A Medical Screening Example

A cancer with occurrence rate of 1% (.01) has a “90% accurate” test, and:



False positive rate: .03, False negative rate: 0.10

Questions:

- ▶ If we screen someone, what is the probability that they test positive?
- ▶ If someone tests positive, what is the probability that they have cancer?

The test is an effect. Cancer is the cause.

## Analysis of Medical Screening Example

With our probabilistic machinery we can analyze this cancer screening example. First, we model the problem in the language of Bayesian probability theory:

$$\begin{aligned}P(C = 1) &= \frac{1}{100} && \text{(Prior probability that someone has cancer)} \\P(C = 0) &= \frac{99}{100} && \text{(Prior probability that someone has no cancer)} \\P(T = 1 \mid C = 1) &= \frac{90}{100} && \text{(Conditional probability of positive test given cancer)} \\P(T = 0 \mid C = 1) &= \frac{10}{100} && \text{(Conditional probability of negative test given cancer)} \\P(T = 1 \mid C = 0) &= \frac{3}{100} && \text{(Conditional probability of positive test given no cancer)} \\P(T = 0 \mid C = 0) &= \frac{97}{100} && \text{(Conditional probability of negative test given no cancer)}\end{aligned}$$

Now we can answer the two questions we posed at the outset:

- ▶ If we screen someone, what is the probability that they test positive?
- ▶ If someone tests positive, what is the probability that they have cancer?

## Analysis of Medical Screening Example

If we screen someone, probability that they test positive (notice use of product rule:  $P(a, b) = P(a | b)P(b)$  and rule 2nd Kolmogorov axiom  $P(x \vee y) = P(x) + P(y) - P(x \wedge y)$ ):

$$P(C = 1) = \frac{1}{100}$$

$$P(C = 0) = \frac{99}{100}$$

$$P(T = 1 | C = 1) = \frac{90}{100}$$

$$P(T = 0 | C = 1) = \frac{10}{100}$$

$$P(T = 1 | C = 0) = \frac{3}{100}$$

$$P(T = 0 | C = 0) = \frac{97}{100}$$

$$\begin{aligned} P(T = 1) &= P(T = 1 | C = 0)P(C = 0) + P(T = 1 | C = 1)P(C = 1) \\ &= \frac{3}{100} \times \frac{99}{100} + \frac{90}{100} \times \frac{1}{100} \\ &= \frac{387}{10,000} \\ &= .0387 \end{aligned}$$

If someone tests positive, probability they have cancer :

$$\begin{aligned} P(C = 1 | T = 1) &= \frac{P(T = 1 | C = 1)P(C = 1)}{P(T = 1)} \\ &= \frac{90}{100} \times \frac{1}{100} \times \frac{10,000}{387} \\ &= \frac{90}{387} \\ &\approx 0.23 \end{aligned}$$

## Bayes' Rule and Combining Evidence

If dentist's probe catches and patient has a toothache, then using the full joint:

	<i>toothache</i>		$\neg$ <i>toothache</i>	
	<i>catch</i>	$\neg$ <i>catch</i>	<i>catch</i>	$\neg$ <i>catch</i>
<i>cavity</i>	0.108	0.012	0.072	0.008
$\neg$ <i>cavity</i>	0.016	0.064	0.144	0.576

we can simply read off the answer:

$$P(\text{Cavity} \mid \text{toothache} \wedge \text{catch}) = \alpha \langle 0.108, 0.016 \rangle \equiv \langle 0.871, 0.129 \rangle$$

But we know this approach doesn't scale. With  $n$  evidence variables we have  $O(2^n)$  possible combinations of observed values.

We could reformulate the problem using Bayes' rule:

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

But, again, we have  $O(2^n)$  combinations of observed evidence. We need some additional domain knowledge.

## Conditional Independence

Toothache and Catch are not independent: if the probe catches in the tooth, then it is likely that the tooth has a cavity and that the cavity causes a toothache.

However, Toothache and Catch are independent given the presence or the absence of a cavity. Each is directly caused by the cavity, but neither has a direct effect on the other. Mathematically, we write this fact as:

$$P(\text{toothache} \wedge \text{catch} \mid \text{Cavity}) = P(\text{toothache} \mid \text{Cavity})P(\text{catch} \mid \text{Cavity})$$

In general, the **conditional independence** of two variables  $X$  and  $Y$ , given a third variable  $Z$ , is defined as:

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

Given these independence assertions we can say  $P(X \mid Y, Z) = P(X \mid Z)$  and  $P(Y \mid X, Z) = P(Y \mid Z)$ .

# Factoring a Joint Distribution using Conditional Independence

Given the conditional independence assertion

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

We can decompose the full joint for *Toothache*, *Catch*, *Cavity*:

$$\begin{aligned} P(\textit{Toothache}, \textit{Catch}, \textit{Cavity}) &= P(\textit{Toothache}, \textit{Catch} \mid \textit{Cavity})P(\textit{Cavity}) && \text{(product rule)} \\ &= P(\textit{Toothache} \mid \textit{Cavity})P(\textit{Catch} \mid \textit{Cavity})P(\textit{Cavity}) && \text{(cond. ind. assertion above)} \end{aligned}$$

This decomposes the large table smaller tables. In general, this technique turns representation that grows as  $O(2^n)$  to one that grows as  $O(n)$ . Conditional independence assertions:

- ▶ allow probabilistic systems to scale, and
- ▶ are much more commonly available than absolute independence assertions.

Conceptually, we say that *Cavity* **separates** *Toothache* and *Catch* because it is a direct cause of both of them.

## Closing Thoughts

The world is uncertain, more precisely, our *knowledge* of the world is uncertain.

- ▶ Probability gives us a language for expressing degrees of belief.
- ▶ Probability theory gives us the analytic tools to construct probabilistic reasoning systems.

In the next lecture we'll begin using these tools to build our first probabilistic reasoning system:  
**Bayesian (Belief) Networks.**