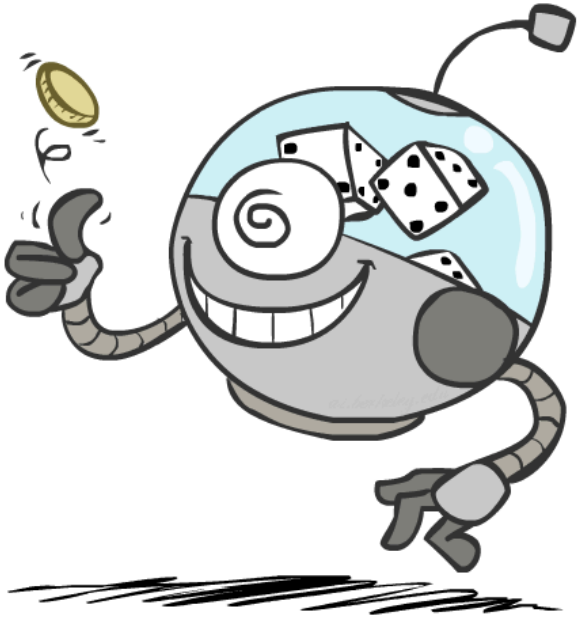
Today

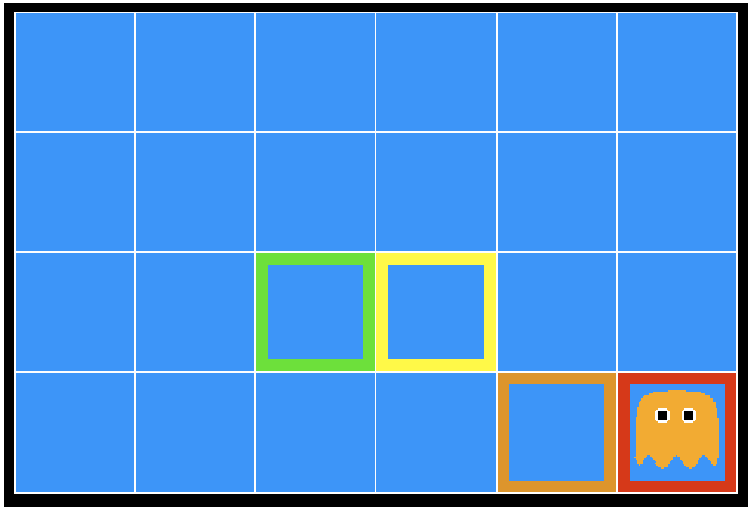
- Random variables
- Probability distributions
- Inference
- Independence
- The Bayes' rule



Do not overlook this lecture!

Quantifying uncertainty

A ghost is hidden in the grid somewhere.

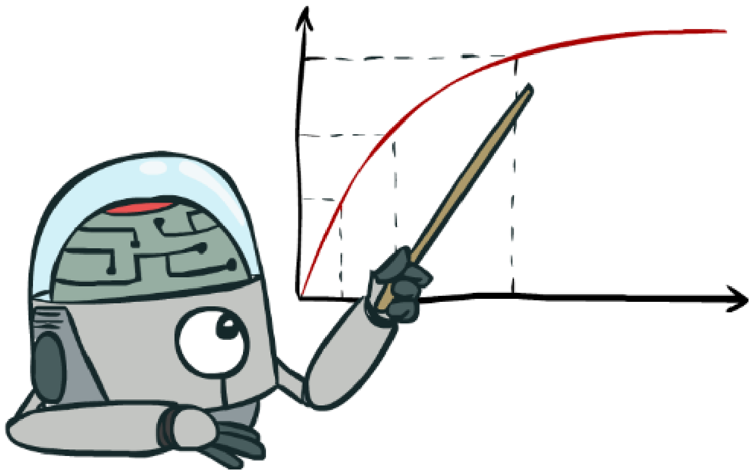


Sensor readings tell how close a square is to the ghost:

- On the ghost: red
- 1 or 2 away: orange
- 3 away: yellow
- 4+ away green

Sensors are noisy, but we know the probability values $P(\mathrm{color}|\mathrm{distance})$, for all colors and all distances.


GHOSTS REMAINING: 1
BUSTS REMAINING: 1
SCORE: 0
MESSAGES:
BUST
TIME+1



Principle of maximum expected utility

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

What does "expected" mean exactly?

Uncertainty

General setup:

- Observed variables or evidence: agent knows certain things about the state of the world (e.g., senso(it is too much work or irrealistic to define an exact model)r readings).
- Unobserved variables: agent needs to reason about other aspects that are uncertain (e.g., where the ghost is).
- (Probabilistic) model: agent knows or believes something about how the observed variables relate to the unobserved variables.

Probabilistic reasoning provides a framework for managing our knowledge and beliefs.

Probabilistic assertions

Probabilistic assertions express the agent's inability to reach a definite decision regarding the truth of a proposition.

- Probability values summarize effects of
  - ignorance (theoretical, practical)
  - laziness (lack of time, resources)
- Probabilities relate propositions to one's own state of knowledge (or lack thereof).
  - e.g., $P(\text{ghost in cell } [3,2]) = 0.02$

Frequentism vs. Bayesianism

What do probability values represent?

- The objectivist frequentist view is that probabilities are real aspects of the universe.
  - i.e., propensities of objects to behave in certain ways.
  - e.g., the fact that a fair coin comes up heads with probability 0.5 is a propensity of the coin itself.
- The subjectivist Bayesian view is that probabilities are a way of characterizing an agent's beliefs or uncertainty.
  - i.e., probabilities do not have external physical significance.
  - This is the interpretation of probabilities that we will use!

How shall we assign numerical values to beliefs?

Kolmogorov's axioms

Begin with a set Ω, the sample space.

$\omega \in \Omega$ is a sample point or possible world.

A probability space is a sample space equipped with a probability function, i.e. an assignment $P : \mathcal{P}(\Omega) \to \mathbb{R}$ such that:

- 1st axiom: $P(\omega) \in \mathbb{R}$, $0 \leq P(\omega)$ for all $\omega \in \Omega$
- 2nd axiom: $P(\Omega) = 1$
- 3rd axiom: $P(\{\omega_1, \ldots, \omega_n\}) = \sum_{i=1}^{n} P(\omega_i)$ for any set of samples

where $\mathcal{P}(\Omega)$ the power set of Ω.

Example

- Ω = the 6 possible rolls of a die.
- ω_i (for $i = 1, \ldots, 6$) are the sample points, each corresponding to an outcome of the die.
- Assignment P for a fair die:

$$P(1) = P(2) = P(3) = P(4) = P(5) = P(6) = \frac{1}{6}$$

Random variables

- A random variable is a function $X : \Omega \rightarrow D_X$ from the sample space to some domain defining its outcomes.
  - e.g., $\text{Odd} : \Omega \rightarrow \{\text{true}, \text{false}\}$ such that $\text{Odd}(\omega) = (\omega \bmod 2 = 1)$.
- P induces a probability distribution for any random variable X.
  - $P(X = x_i) = \sum_{\{\omega : X(\omega) = x_i\}} P(\omega)$
  - e.g., $P(\text{Odd} = \text{true}) = P(1) + P(3) + P(5) = \frac{1}{2}$.
- An event E is a set of outcomes $\{(x_1, ..., x_n), ...\}$ of the variables $X_1, ..., X_n$, such that

$$P(E) = \sum_{(x_1, ..., x_n) \in E} P(X_1 = x_1, ..., X_n = x_n).$$

Notations

- Random variables are written in upper roman letters: X, Y, etc.
- Realizations of a random variable are written in corresponding lower case letters. E.g., x_1, x_2, ..., x_n could be of outcomes of the random variable X.
- The probability value of the realization x is written as $P(X = x)$.
- When clear from context, this will be abbreviated as $P(x)$.
- The probability distribution of the (discrete) random variable X is denoted as $\mathbf{P}(X)$. This corresponds e.g. to a vector of numbers, one for each of the probability values $P(X = x_i)$ (and not to a single scalar value!).

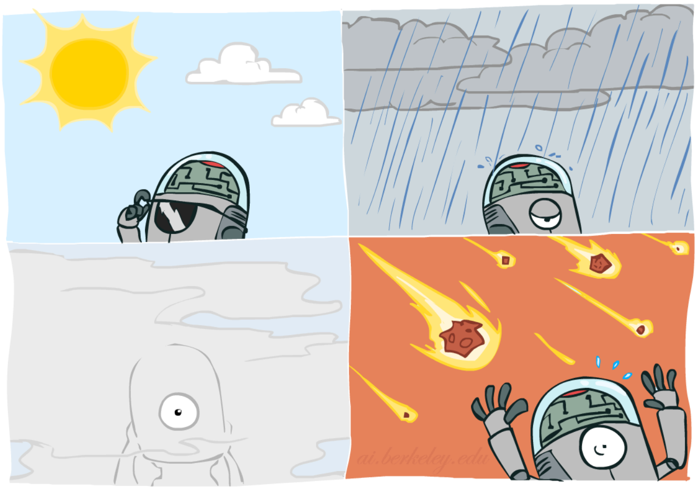
Probability distributions

For discrete variables, the probability distribution can be encoded by a discrete list of the probabilities of the outcomes, known as the probability mass function.

One can think of the probability distribution as a table that associates a probability value to each outcome of the variable.

$\mathbf{P}(W)$

W	P
sun	0.6
rain	0.1
fog	0.3
meteor	0.0



Joint distributions

A joint probability distribution over a set of random variables $X_1, ..., X_n$ specifies the probability of each (combined) outcome:

$$P(X_1 = x_1, ..., X_n = x_n) = \sum_{\{\omega : X_1(\omega) = x_1, ..., X_n(\omega) = x_n\}} P(\omega)$$

$$\mathbf{P}(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

Marginal distributions

The marginal distribution of a subset of a collection of random variables is the joint probability distribution of the variables contained in the subset.

$$\mathbf{P}(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$\mathbf{P}(T)$$

T	P
hot	0.5
cold	0.5

$$P(t) = \sum_w P(t, w)$$

$$\mathbf{P}(W)$$

W	P
sun	0.6
rain	0.4

$$P(w) = \sum_t P(t, w)$$

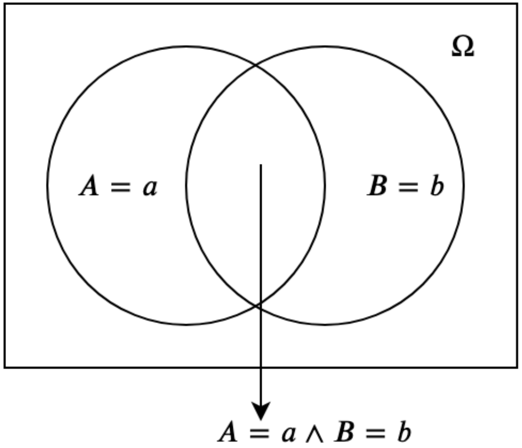
Intuitively, marginal distributions are sub-tables which eliminate variables.

Conditional distributions

The conditional probability of a realization a given the realization b is defined as the ratio of the probability of the joint realization a and b, and the probability of b:

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

Indeed, observing $B = b$ rules out all those possible worlds where $B \neq b$, leaving a set whose total probability is just $P(b)$. Within that set, the worlds for which $A = a$ satisfy $A = a \wedge B = b$ and constitute a fraction $P(a, b)/P(b)$.



Conditional distributions are probability distributions over some variables, given fixed values for others.

$$\mathbf{P}(T, W)$$

T	W	P
hot	sun	0.4
hot	rain	0.1
cold	sun	0.2
cold	rain	0.3

$$\mathbf{P}(W|T = \text{hot})$$

W	P
sun	0.8
rain	0.2

$$\mathbf{P}(W|T = \text{cold})$$

W	P
sun	0.4
rain	0.6

Probabilistic inference

Probabilistic inference is the problem of computing a desired probability from other known probabilities (e.g., conditional from joint).

- We generally compute conditional probabilities.
  - e.g., $P(\text{on time}|\text{no reported accidents}) = 0.9$
  - These represent the agent's beliefs given the evidence.
- Probabilities change with new evidence:
  - e.g., $P(\text{on time}|\text{no reported accidents}, 5\text{AM}) = 0.95$
  - e.g., $P(\text{on time}|\text{no reported accidents}, \text{rain}) = 0.8$
  - e.g., $P(\text{ghost in } [3,2]|\text{red in } [3,2]) = 0.99$
  - Observing new evidence causes beliefs to be updated.

General case

- Evidence variables: $E_1, ..., E_k = e_1, ..., e_k$
- Query variables: Q
- Hidden variables: $H_1, ..., H_r$
- $(Q \cup E_1, ..., E_k \cup H_1, ..., H_r)$ = all variables $X_1, ..., X_n$

Inference is the problem of computing $\mathbf{P}(Q|e_1, ..., e_k)$.

Inference by enumeration

Start from the joint distribution $\mathbf{P}(Q, E_1, ..., E_k, H_1, ..., H_r)$.

1. Select the entries consistent with the evidence $E_1, ..., E_k = e_1, ..., e_k$.
2. Marginalize out the hidden variables to obtain the joint of the query and the evidence variables:

$$\mathbf{P}(Q, e_1, ..., e_k) = \sum_{h_1, ..., h_r} \mathbf{P}(Q, h_1, ..., h_r, e_1, ..., e_k).$$

3. Normalize:

$$Z = \sum_q P(q, e_1, ..., e_k)$$

$$\mathbf{P}(Q|e_1, ..., e_k) = \frac{1}{Z}\mathbf{P}(Q, e_1, ..., e_k)$$

Example

- $\mathbf{P}(W)$?
- $\mathbf{P}(W|\text{winter})$?
- $\mathbf{P}(W|\text{winter}, \text{hot})$?

S	T	W	P
summer	hot	sun	0.3
summer	hot	rain	0.05
summer	cold	sun	0.1
summer	cold	rain	0.05
winter	hot	sun	0.1
winter	hot	rain	0.05
winter	cold	sun	0.15
winter	cold	rain	0.2

Complexity

- Inference by enumeration can be used to answer probabilistic queries for discrete variables (i.e., with a finite number of values).
- However, enumeration does not scale!
  - Assume a domain described by n variables taking at most d values.
  - Space complexity: $O(d^n)$
  - Time complexity: $O(d^n)$

Can we reduce the size of the representation of the joint distribution?

Product rule

$$P(a, b) = P(b)P(a|b)$$

Example

$\mathbf{P}(W)$

W	P
sun	0.8
rain	0.2

$\mathbf{P}(D|W)$

D	W	P
wet	sun	0.1
dry	sun	0.9
wet	rain	0.7
dry	rain	0.3

$\mathbf{P}(D, W)$

D	W	P
wet	sun	?
dry	sun	?
wet	rain	?
dry	rain	?

Chain rule

More generally, any joint distribution can always be written as an incremental product of conditional distributions:

$$P(x_1, x_2, x_3) = P(x_1)P(x_2|x_1)P(x_3|x_1, x_2)$$

$$P(x_1, ..., x_n) = \prod_{i=1}^{n} P(x_i|x_1, ..., x_{i-1})$$

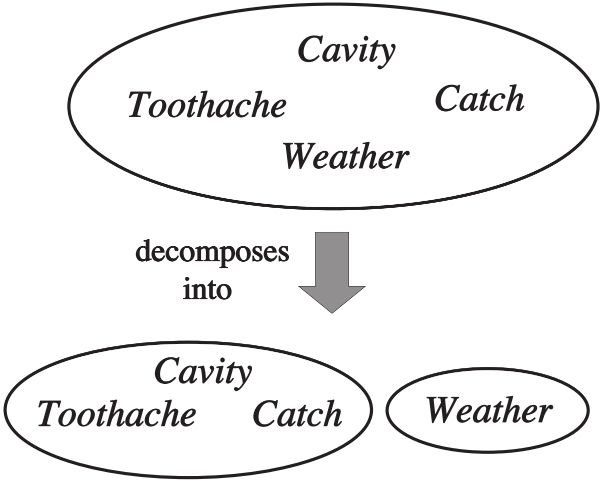
Independence

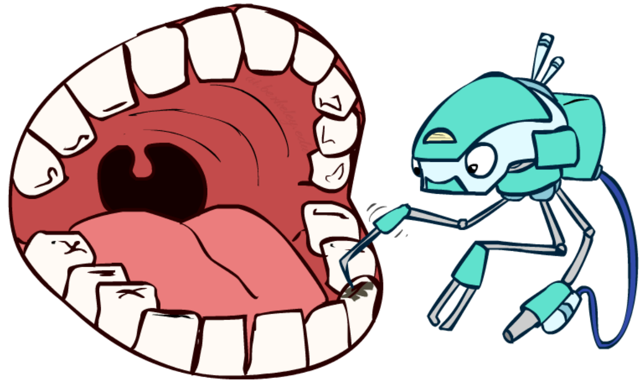
A and B are independent iff, for all $a \in D_A$ and $b \in D_B$,

- $P(a|b) = P(a)$, or
- $P(b|a) = P(b)$, or
- $P(a, b) = P(a)P(b)$

Independence is denoted as $A \perp B$.

Example 1





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

The original 32-entry table reduces to one 8-entry and one 4-entry table (assuming 4 values for Weather and boolean values otherwise).

Example 2

For n independent coin flips, the joint distribution can be fully factored and represented as the product of n 1-entry tables.

- $2^n \rightarrow n$

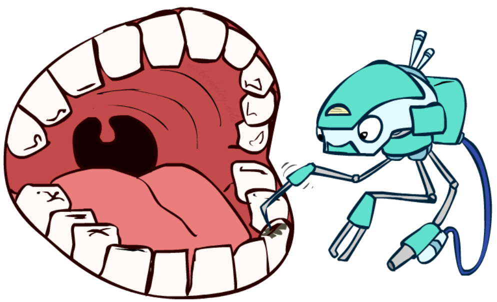
Conditional independence

A and B are conditionally independent given C iff, for all $a \in D_A$, $b \in D_B$ and $c \in D_C$,

- $P(a|b,c) = P(a|c)$, or
- $P(b|a,c) = P(b|c)$, or
- $P(a,b|c) = P(a|c)P(b|c)$

Conditional independence is denoted as $A \perp B|C$.

- Using the chain rule, the join distribution can be factored as a product of conditional distributions.
- Each conditional distribution may potentially be simplified by conditional independence.
- Conditional independence assertions allow probabilistic models to scale up.



Example 1

Assume three random variables Toothache, Catch and Cavity.

Catch is conditionally independent of Toothache, given Cavity. Therefore, we can write:

$$\begin{aligned}&P(\text{toothache}, \text{catch}, \text{cavity}) \\ &= P(\text{toothache}|\text{catch}, \text{cavity}) P(\text{catch}|\text{cavity}) P(\text{cavity}) \\ &= P(\text{toothache}|\text{cavity}) P(\text{catch}|\text{cavity}) P(\text{cavity})\end{aligned}$$

In this case, the representation of the joint distribution reduces to $2+2+1$ independent numbers (instead of $2^n - 1$).

Example 2 (Naive Bayes)

More generally, from the product rule, we have

$$P(\text{cause}, \text{effect}_1, ..., \text{effect}_n) = P(\text{effect}_1, ..., \text{effect}_n | \text{cause}) P(\text{cause})$$

Assuming pairwise conditional independence between the effects given the cause, it comes:

$$P(\text{cause}, \text{effect}_1, ..., \text{effect}_n) = P(\text{cause}) \prod_i P(\text{effect}_i | \text{cause})$$

This probabilistic model is called a naive Bayes model.

- The complexity of this model is $O(n)$ instead of $O(2^n)$ without the conditional independence assumptions.
- Naive Bayes can work surprisingly well in practice, even when the assumptions are wrong.

Study the next slide. **Twice**.

The Bayes' rule

The product rule defines two ways to factor the joint distribution of two random variables.

$$P(a, b) = P(a|b)P(b) = P(b|a)P(a)$$

Therefore,

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

- $P(a)$ is the prior belief on a.
- $P(b)$ is the probability of the evidence b.
- $P(a|b)$ is the posterior belief on a, given the evidence b.
- $P(b|a)$ is the conditional probability of b given a. Depending on the context, this term is called the likelihood.

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

The Bayes' rule is the foundation of many AI systems.

Example 1: diagnostic probability from causal probability.

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

where

- $P(\text{effect}|\text{cause})$ quantifies the relationship in the causal direction.
- $P(\text{cause}|\text{effect})$ describes the diagnostic direction.

Let S=stiff neck and M=meningitis. Given $P(s|m) = 0.7$, $P(m) = 1/50000$, $P(s) = 0.01$, it comes

$$P(m|s) = \frac{P(s|m)P(m)}{P(s)} = \frac{0.7 \times 1/50000}{0.01} = 0.0014.$$

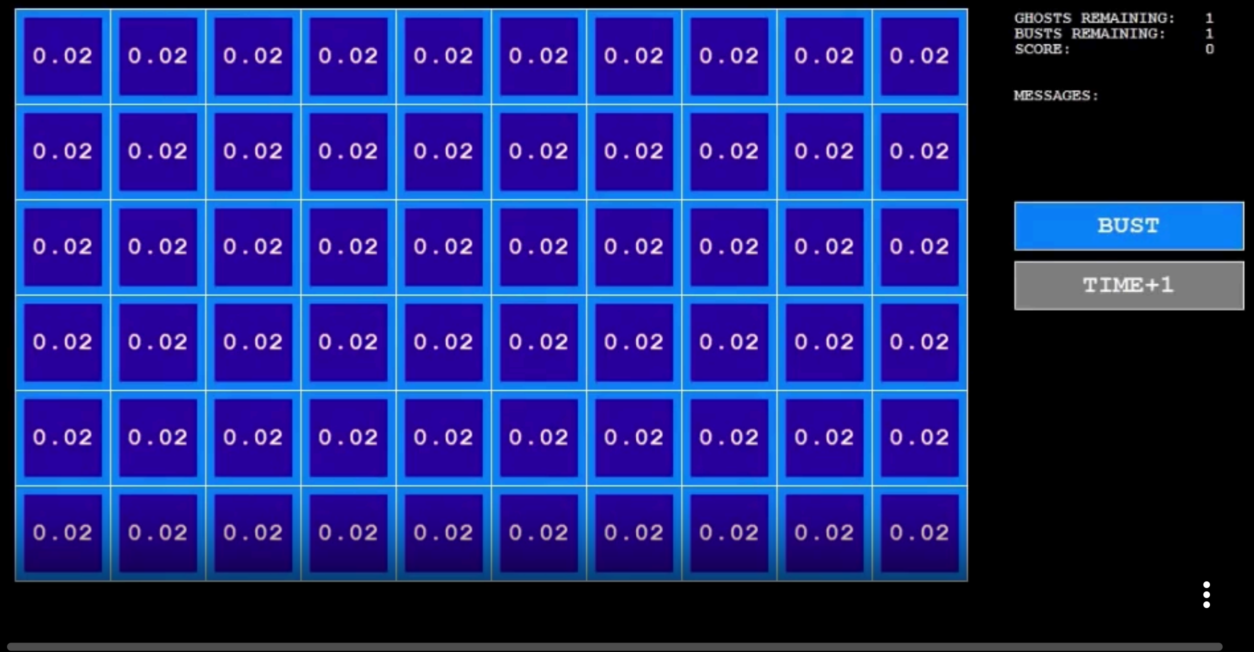
Example 2: Ghostbusters, revisited

- Let us assume a random variable G for the ghost location and a set of random variables $R_{i,j}$ for the individual readings.
- We start with a uniform prior distribution $\mathbf{P}(G)$ over ghost locations.
- We assume a sensor reading model $\mathbf{P}(R_{i,j}|G)$.
    - That is, we know what the sensors do.
    - $R_{i,j}$ = reading color measured at $[i, j]$
        - e.g., $P(R_{1,1} = \text{yellow}|G = [1, 1]) = 0.1$
    - Two readings are conditionally independent, given the ghost position.

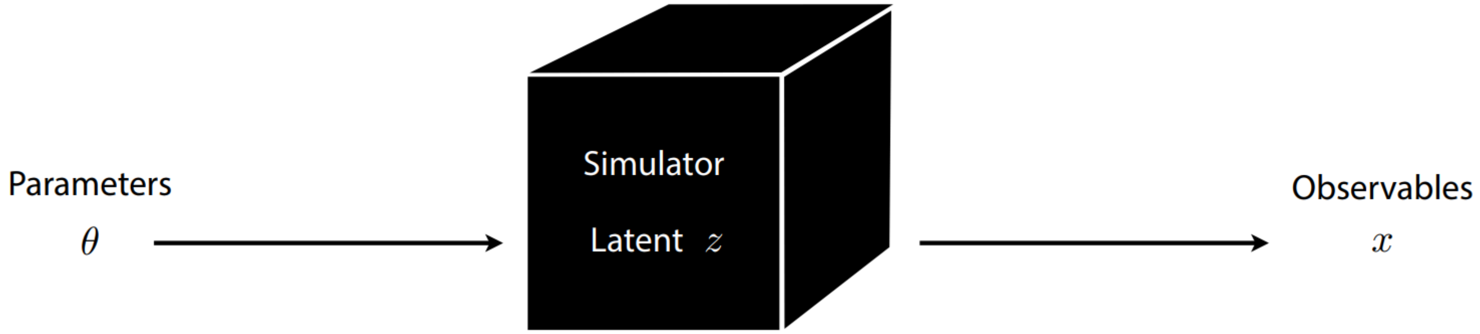
- We can calculate the posterior distribution $\mathbf{P}(G|R_{i,j})$ using Bayes' rule:

$$\mathbf{P}(G|R_{i,j}) = \frac{\mathbf{P}(R_{i,j}|G)\mathbf{P}(G)}{\mathbf{P}(R_{i,j})}.$$

- For the next reading $R_{i',j'}$, this posterior distribution becomes the prior distribution over ghost locations, which we update similarly.


0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02
GHOSTS REMAINING: 1
BUSTS REMAINING: 1
SCORE: 0
MESSAGES:
BUST
TIME+1

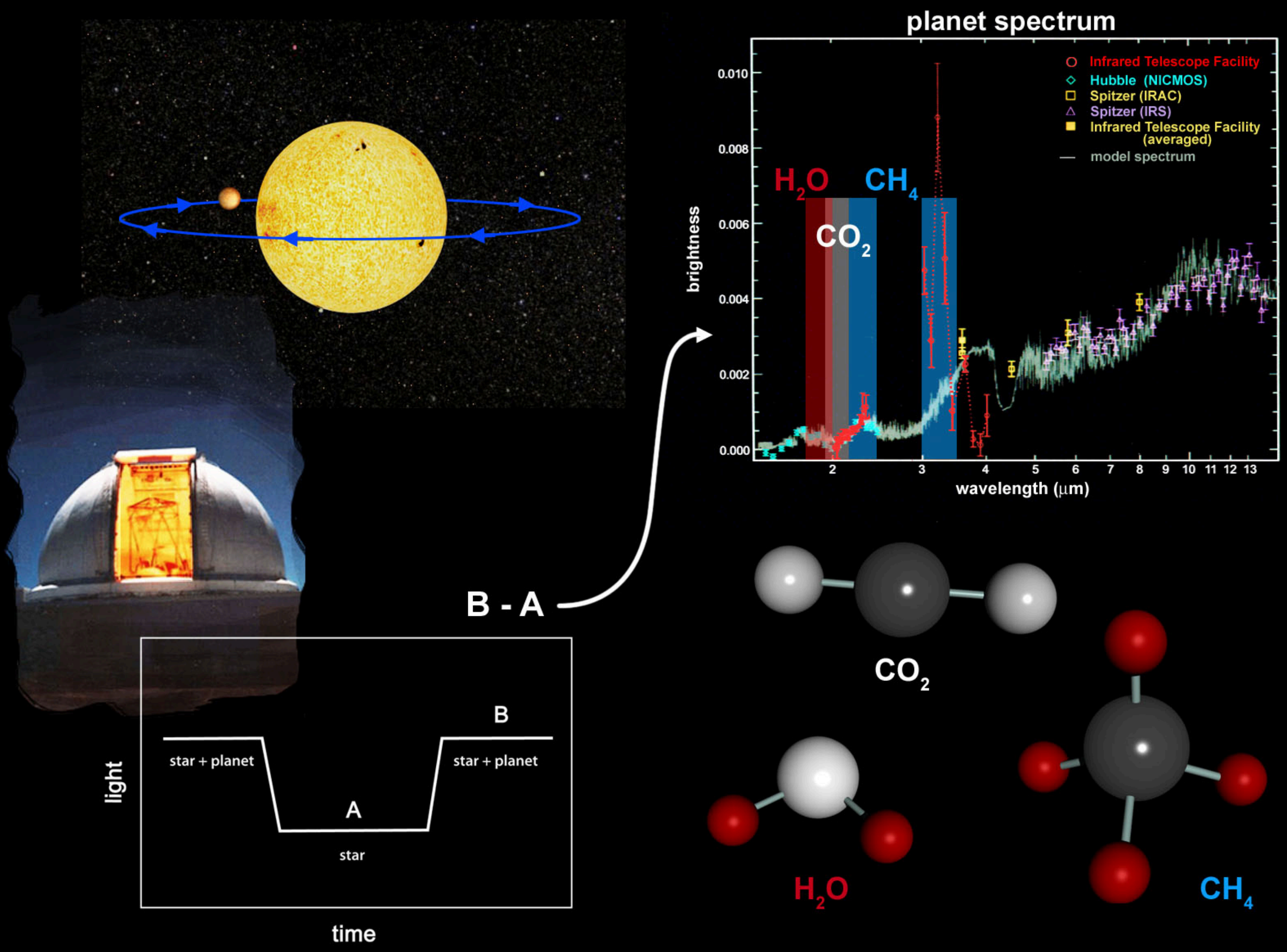
Example 3: AI for Science

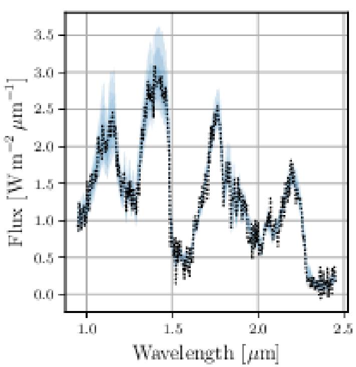


Given some observation x and prior beliefs $p(\theta)$, science is about updating one's knowledge, which may be framed as computing

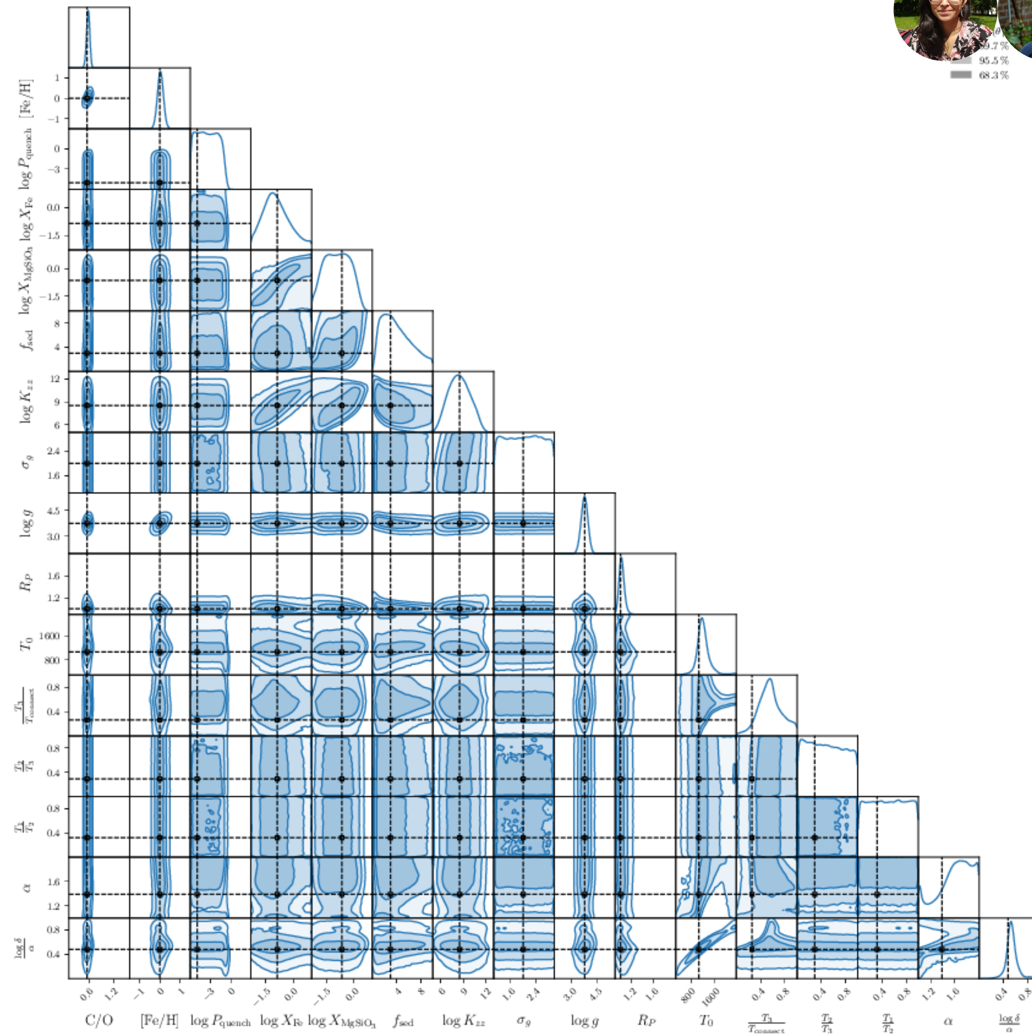
$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}.$$

Exoplanet atmosphere characterization





Flux [W m^{-2} μm^{-1}]
Wavelength [μm]



C/O
[Fe/H]
$\log P_{\rm quench}$
$\log X_{\rm Fe}$
$\log X_{\rm MgSiO_3}$
$f_{\rm sed}$
$\log K_{zz}$
σ_g
$\log g$
R_P
T_0
$\frac{T_3}{T_{\rm connect}}$
$\frac{T_2}{T_3}$
$\frac{T_1}{T_2}$
α
$\frac{\log \delta}{\alpha}$
95.5 %
68.3 %

Summary

- Uncertainty arises because of laziness and ignorance. It is inescapable in complex non-deterministic or partially observable environments.
- Probabilistic reasoning provides a framework for managing our knowledge and beliefs, with the Bayes' rule acting as the workhorse for inference.