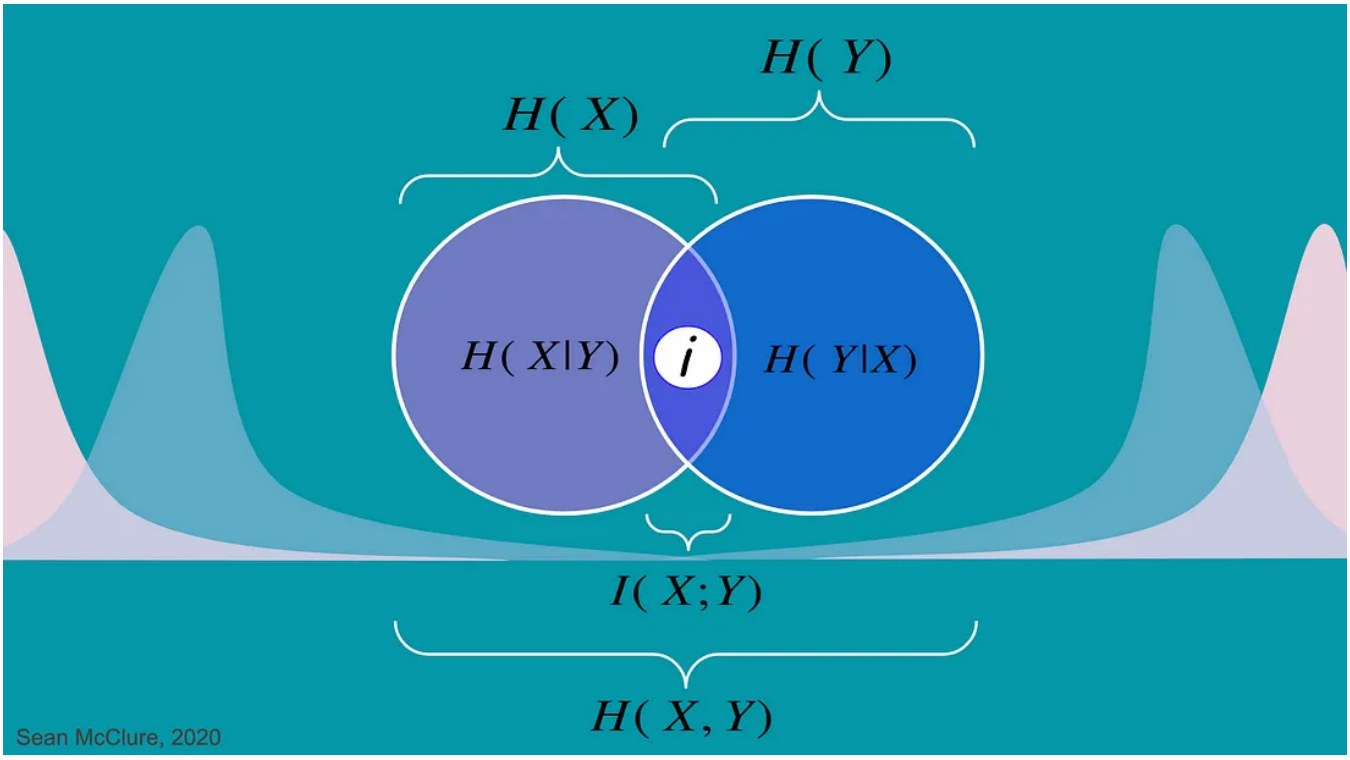
A Deep Conceptual Guide to Mutual Information

Embracing the “Correlation of the 21st Century.”

Sean McClure · Follow

Published in The Startup

36 min read · Nov 7, 2020



The logical relations between various concepts underlying Mutual Information.

Dependency

Causality is a central concept in our lives. It refers to the idea that one event influences another event. Our perception of causality underlies many of our personal opinions. If I believe the internet makes people dumber, or that the President has made things worse, I am suggesting a causal connection; real or not.

In business we look to understand what makes a good hire, a good decision, a good product. In government we create new legislation and policies based on evidence

for social and economic causes. In short, causality has much to do with how we try to make sense of the world.

Of course science rests largely on the idea of causality. We attempt to *interpret* our observations by making causal statements. If we believe we know the event, process or state that contributes to another event, process or state we say we know something about the underlying phenomenon. We also often make *predictions* with whatever causal structure has been uncovered by the models we build.

The statistical concept for dealing with causality is called *dependence*, more commonly called **correlation**. Correlation represents *any statistical association between 2 or more variables*. Of course we all know correlation alone is not some guarantee of causal relationship but it can act as a signpost of a potential relationship between 2 things.

The most common approach to quantifying correlation between variables is the Pearson product-moment correlation coefficient (PPMCC). I will simply refer to this approach as **Pearson's correlation**. Pearson's correlation is the workhorse of dependence and considered an industry standard, inside and out of academia.

The ubiquity of Pearson's correlation means it should be subject to the highest amount of scrutiny compared to other methods. Unfortunately the amount of criticism put to Pearson's correlation is far from that deserved. The reason is its ease of use and extremely intuitive interpretation. Pearson's correlation tells a straightforward story about 2 things "moving" together and can make that story look reasonably technical. A researcher hoping to support some narrative around how things are related can call upon Pearson's correlation to appear "scientific."

But Pearson's correlation makes some very simple assumptions about how 2 things might be related. Most real-world situations are nontrivial and don't lend themselves to such simplistic descriptions of dependence. It's too easy to promote false narratives with techniques like Pearson's correlation since data can always be "tortured" into submission. A scientific justification for how prevalent Pearson's correlation is appears scant at best.

The popularity of Pearson's correlation does however present us with the best **basis of comparison** for contrasting other methods. By looking at more scientifically valid techniques we can better understand where Pearson's correlation falls short.

Specifically, we can see how Pearson's correlation is a very weak proxy to a more rigorous approach to understanding how 2 or more variables share *information*.

We will unpack just how important the concept of information is throughout this article. Our discussion will center around what some have called a "correlation for the 21st Century." The starring role is played by a measure known as **Mutual Information**; the topic of this article. Mutual Information digs much deeper into the notion of dependence by focusing directly on information itself.

Pearson's Shortcomings

The core equations behind Pearson's correlation are as follows:

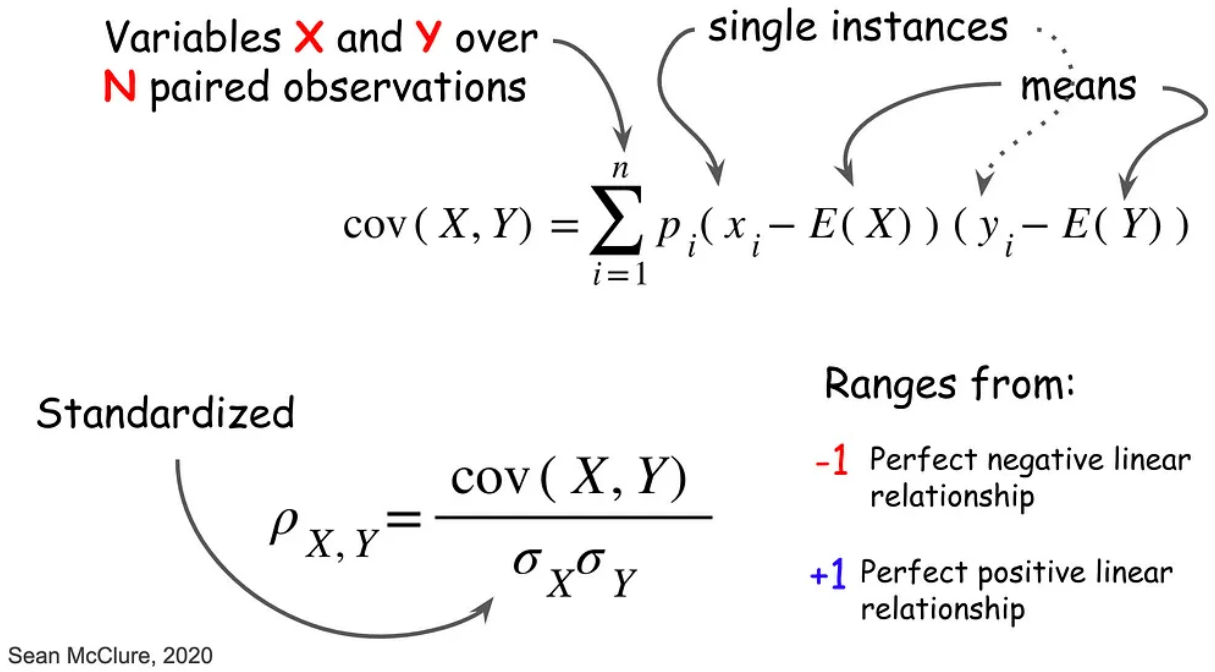


Figure 1 The math behind Pearson's correlation.

Pearson's correlation is the 2nd equation at the bottom, and is calculated using the *normalized covariance* shown at the top. Normalization is done whenever we wish to make data comparable, by adjusting values that are on different scales to a common scale. In Figure 1 we are dividing the covariance of 2 random variables by the product of both their standard deviations to achieve our normalization. Think of this step as "squashing" the possible correlation values between **-1 and 1**, which makes the coefficient easier to work with.

A more intuitive way to think of Pearson's correlation is in terms of its **geometric interpretation**. Any variable (column in a dataset) can be thought of as a *vector* in

vector space. Correlation is then simply calculating the **cosine** of the **angle** θ between two observed vectors:

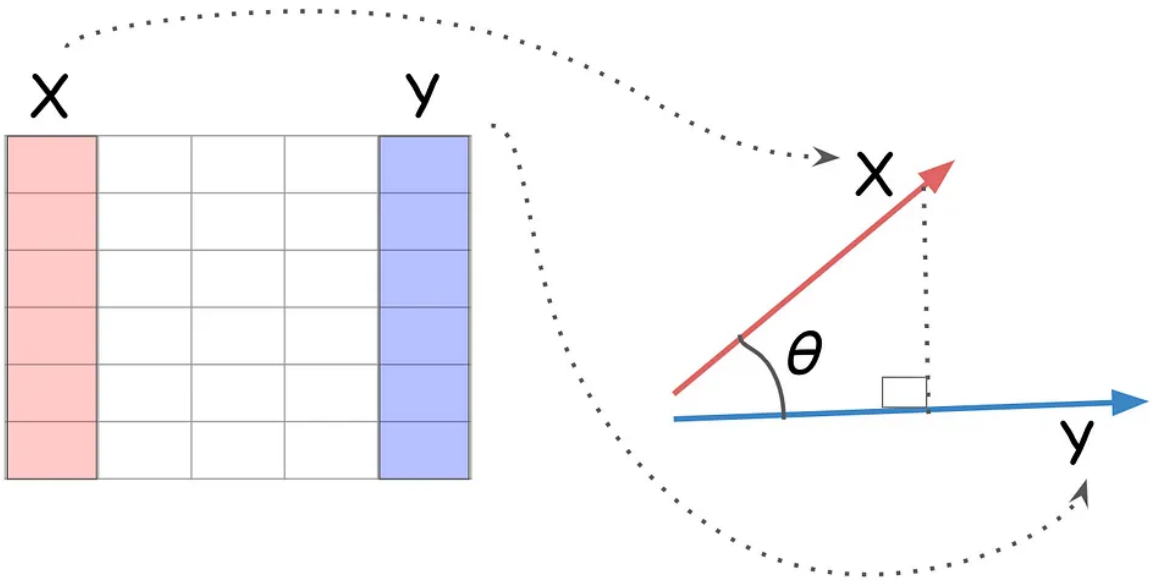


Figure 2 An **intuitive explanation** of Pearson's correlation is the cosine angle between 2 vectors representing each variable.

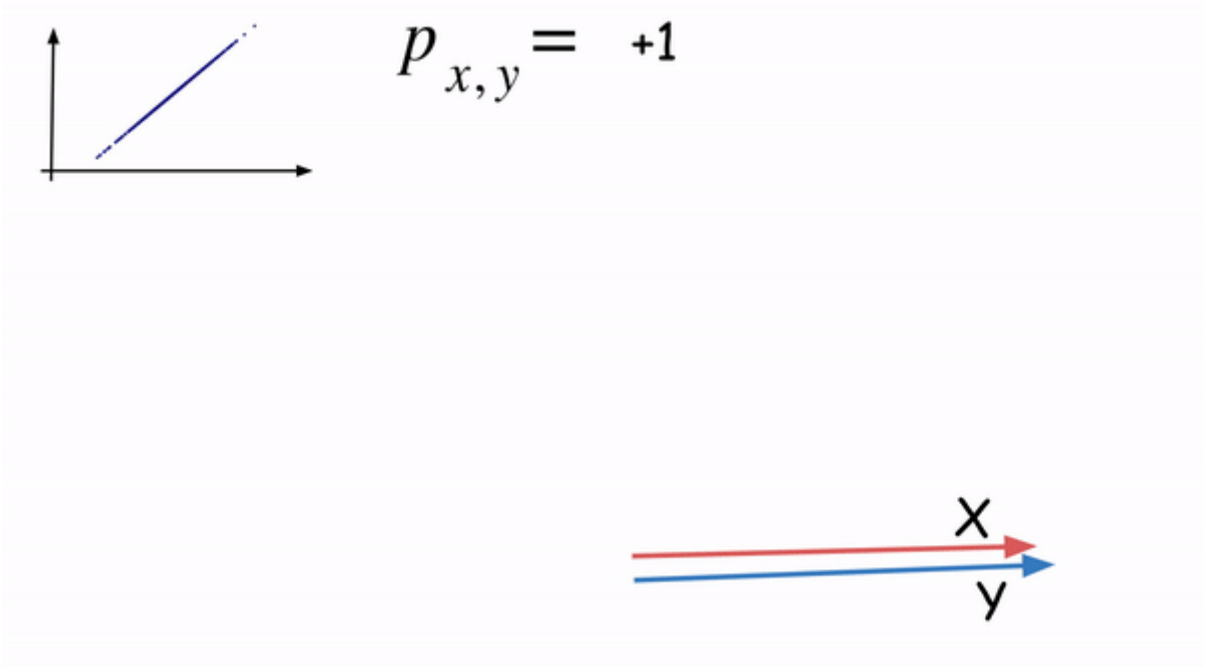


Figure 3 If we **increase the angle** between 2 vectors (variables) we decrease Pearson's correlation. Note the change in variance in the upper left.

> *While covariance gives us the direction of the relationship, correlation gives* both ***direction*** *and* ***strength****. This is because the magnitude of covariance is arbitrary (it depends on the units), meaning if we change the units we change the magnitude for the*

> *exact same phenomenon. Thus correlation normalizes the covariance as discussed above to make the strength of the relationship non-arbitrary.*

Correlation is best suited to *continuous, normally distributed* data and is thus easily swayed by extreme values. As such, correlation will **misrepresent relationships that are not linear**. This occurs VERY OFTEN in practice since much of our world is nonlinear and not normally distributed (your stats class notwithstanding).

Despite being the goto measure for association in practice Pearson's correlation is **not a general measure of dependence.** We say that the *information* given by a correlation coefficient is not enough to define the dependence structure between random variables. The fact that correlation only looks for *linear* dependence means it cannot suggest there is no general correlation when it measures 0 (a correlation of 0 does *not* mean variables are independent).

Only in situations where things are simple (e.g. linear) will Pearson's correlation offer some potential insight into causality (although there are issues with this idea as well). It relies on covariance, which suggests that greater values of one variable mainly correspond with greater values of another variable. While this sounds obvious, it is a simplistic notion of things being related. Variables can be related in all kinds of ways.

To tease out the dependence structure between variables requires we do more than just compare vectors of data. We need to somehow tap directly into the idea of *information*. After all, that is what we're doing when we make a measurement. The outstanding question in any measurement is how much information is being *shared* between the 2 or more things.

Mutual information, as its name suggests, looks to find how much information is shared between 2 variables rather than just noting their commensurate "movement." To grasp such a technique requires we understand information itself, and this brings us to the bulk of this article.

Information

Information is defined as the resolution of uncertainty. This means an appropriate approach would account for the uncertainty inherent in any measurement, and this demands probability.

The most obvious way forward would be to use the **multiplication rule for independent events:**

```
IF P(A)*P(B) = P(A and B) THEN A and B are independent events,
otherwise, they are dependent events.
```

This tells us that variables A and B are to be considered independent if the product of their marginals equals their joint probability.

> *Recall that a **marginal distribution** gives the probabilities of different values of a variable contained within some subset, without reference to values of any other variables. A **joint distribution** gives the probability that 2 or more variables fall within a particular range.*

Why would 2 variables be independent if the product of their marginals equal their joint probability? Imagine a square matrix filled with values that show the combination of rows and columns as shown in Figure 4. A casual look at these values reveals that almost half of them are *redundant*; they are symmetric about the diagonal.

	1	2	3	4	5	6	7	8	9	10
1	**1 x 1**	1 x 2	1 x 3	1 x 4	1 x 5	1 x 6	1 x 7	1 x 8	1 x 9	1 x 10
2	2 x 1	**2 x 2**	2 x 3	2 x 4	2 x 5	2 x 6	2 x 7	2 x 8	2 x 9	2 x 10
3	3 x 1	3 x 2	**3 x 3**	3 x 4	3 x 5	3 x 6	3 x 7	3 x 8	3 x 9	3 x 10
4	4 x 1	4 x 2	4 x 3	**4 x 4**	4 x 5	4 x 6	4 x 7	4 x 8	4 x 9	4 x 10
5	5 x 1	5 x 2	5 x 3	5 x 4	**5 x 5**	5 x 6	5 x 7	5 x 8	5 x 9	5 x 10
6	6 x 1	6 x 2	6 x 3	6 x 4	6 x 5	**6 x 6**	6 x 7	6 x 8	6 x 9	6 x 10
7	7 x 1	7 x 2	7 x 3	7 x 4	7 x 5	7 x 6	**7 x 7**	7 x 8	7 x 9	7 x 10
8	8 x 1	8 x 2	8 x 3	8 x 4	8 x 5	8 x 6	8 x 7	**8 x 8**	8 x 9	8 x 10
9	9 x 1	9 x 2	9 x 3	9 x 4	9 x 5	9 x 6	9 x 7	9 x 8	**9 x 9**	9 x 10
10	10 x 1	10 x 2	0 x 3	10 x 4	10 x 5	10 x 6	10 x 7	10 x 8	10 x 9	**10 x 10**

Figure 4 The sharing of information between 2 variables, visualized as a matrix of values. Symmetry in data signifies redundancy.

This symmetry points to a sharing of information, since we see the same outputs produced by different axes.

Of course this can occur by coincidence, but only to a point. Any 2 variables from a reasonably sized phenomenon will not produce the same outputs by coincidence. Compare this to what Pearson's correlation does: it simply compares variances, meaning as long as 2 variables grow or shrink in unison the variables are deemed dependent.

In other words, whereas Pearson's uses a well-defined *moment* (variance) of an assumed distribution (Gaussian), Mutual Information instead tallies up actual values between all variables considered. The chances of coincidence are far lower with MI than Pearson's.

While the product of 2 distributions contains *all* the information brought by both variables (everything inside the matrix), the joint distribution is more like the pink area above. If outcomes are shared between 2 variables then there is less information in the joint than in the product of marginals. This is why any deviation between the joint distribution and the product of marginals indicates a sharing of information, and thus dependence.

If we wanted to create a method that detects when information is shared we would look to leverage the above concept, and this is precisely what Mutual Information does. Let's look at how Mutual Information is constructed:

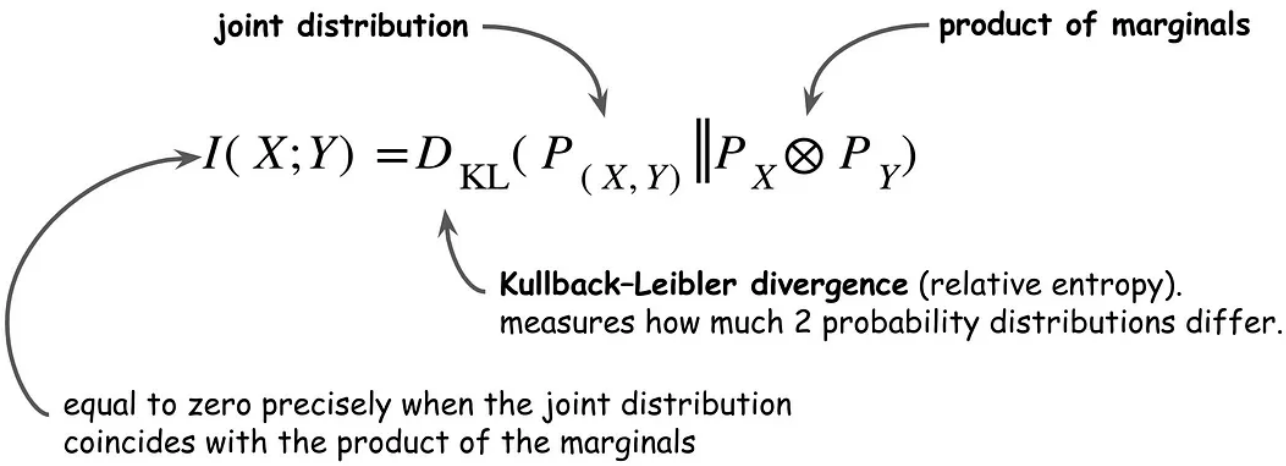


Figure 5 The definition of Mutual Information.

Mutual Information uses something known as Kullback–Leibler divergence (**KL divergence**), which is what we see on the right-hand side of the equation. KL

divergence is a measure of how one probability distribution is different from a second, reference probability distribution.

Let's look more closely at how KL-divergence is defined mathematically using 2 ordinary distributions:

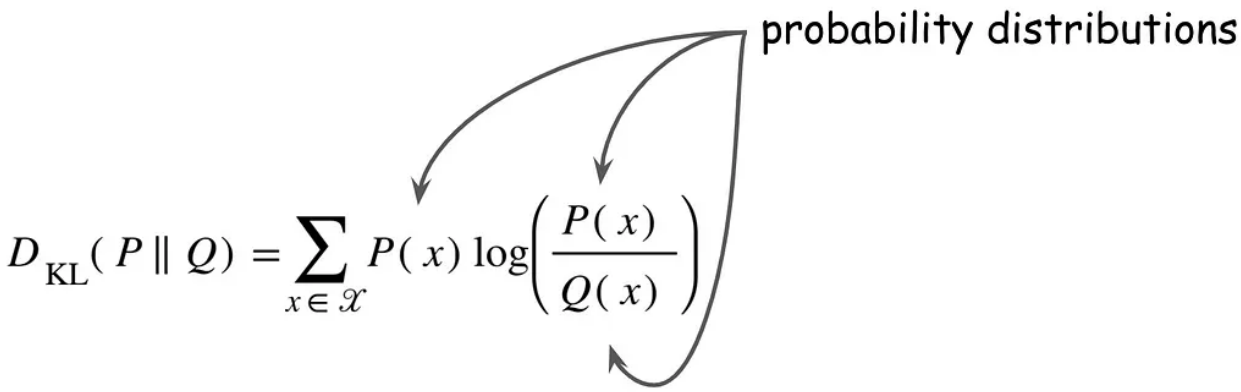
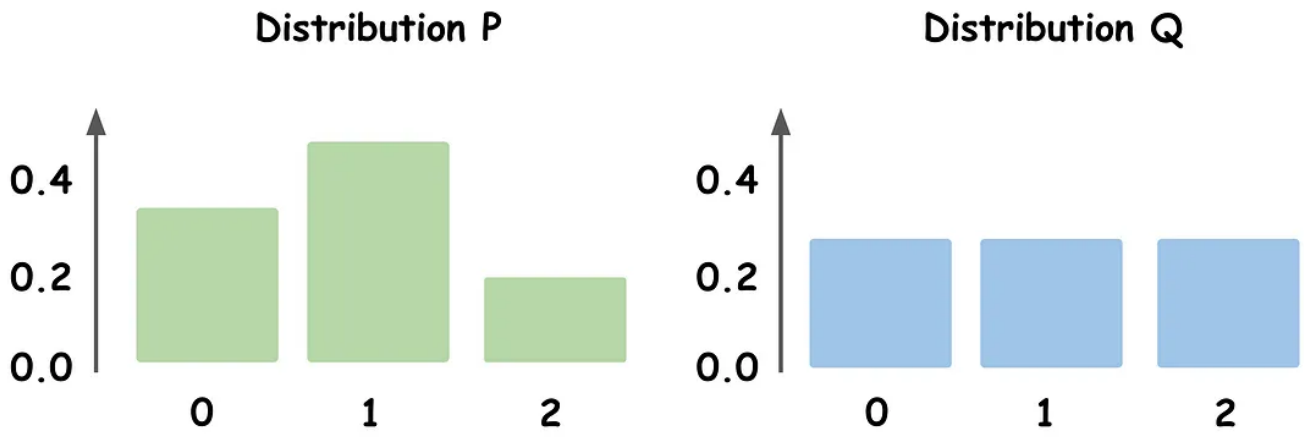


Figure 6 KL-Divergence for discrete variables.

We can see that KL-divergence is the expectation of the logarithmic difference between the probabilities P and Q. The expectation just means we're calculating the average value, hence the summation sign summing over all values. We can see the expression is also logarithmic (more on this later). A critical piece is the use of a ratio between probability distributions. This is how the "*difference*" between distributions is taken into account.

> *Figure 6 shows the discrete form, but we can just as easily express KL-divergence for continuous variables (where* ***p*** *and* ***q*** *are now probability densities instead of mass functions).*

Let's plug values into the KL-divergence formula to see how it works. Figure 7 shows 2 distribution with their respective values.

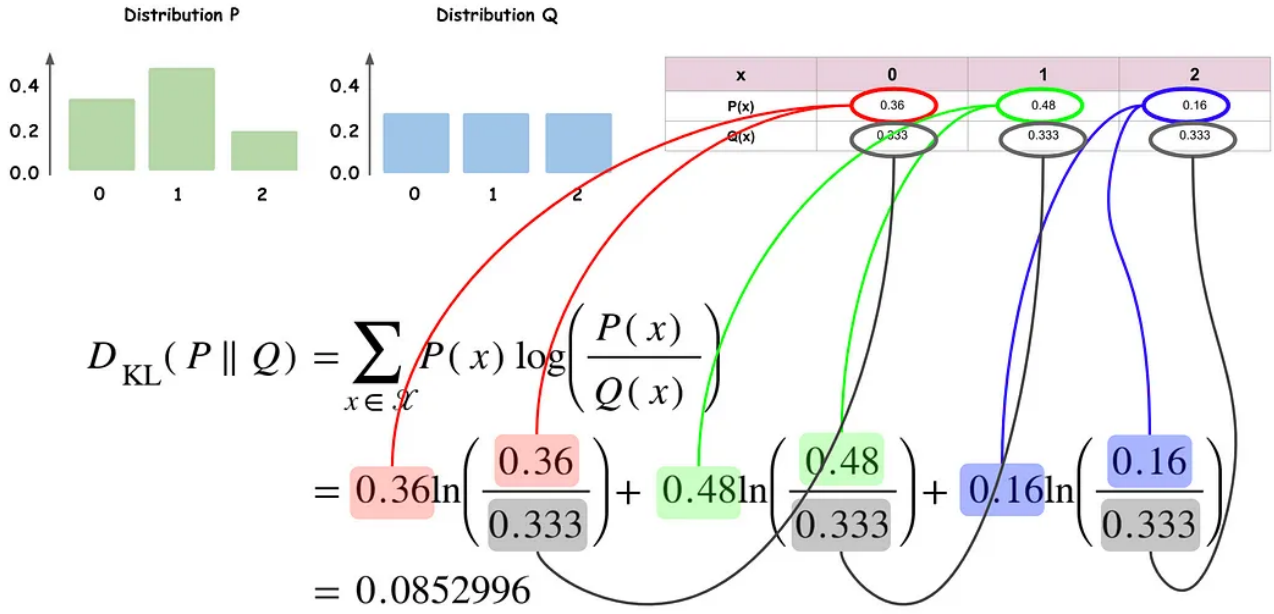


x	0	1	2
P(x)	0.36	0.48	0.16
Q(x)	0.333	0.333	0.333

Sean McClure, 2020

Figure 7 2 different probability distributions and their tabulated values. Adapted from Wikipedia.

Plugging these values into the KL-divergence equation we get:



Sean McClure, 2020

Figure 8 Calculating the KL-Divergence from the tabulated values of 2 distributions.

A KL-divergence of 0 would indicate that the two distributions in question are identical. We can see in our example that the distributions *diverge* from one another (a non-zero result).

Think of P as representing the **true distribution** of the data ("reality") while Q is our theory, **model**, or approximation of P. The intuition behind KL-divergence is that it looks at how different our model of reality is *from* reality. In many applications we are thus looking to *minimize* the KL-divergence between P and Q in order to find the distribution Q (model) that is closest to the distribution P (reality).

> *We often say KL divergence represents the divergence between P **and** Q, but this isn't quite right. There is a fundamental asymmetry in the relation. We should actually say it describes the divergence of P **from** Q, or the divergence from P **to** Q. While this sounds pedantic it better reflects the asymmetry in Bayesian inference; we start from a prior (Q), which updates to the posterior (P).*

There are various ways to use KL-divergence. When talking in terms of Bayesian inference we can think of it as **a measure of the information gained by revising one's beliefs** from the prior probability distribution (the amount of information lost when Q is used to approximate P). In machine learning we call this **information gain**, while in coding theory it refers to the extra bits needed to code samples from P using a code optimized for Q.

We can see that the general definition of KL-divergence doesn't quite look like the one used in Mutual Information. This is because instead of using plain distributions Mutual Information uses the joint and the product of marginals. This is valid since a joint distribution is itself a single distribution, and the product of marginals is also itself a single distribution. In other words, we can look for the KL-divergence between a single joint distribution and a single product distribution to measure the dependence between 2 variables.

Since we are looking for how different the joint is from the product of marginals we are doing what we saw in Figure 4. We already know that if the joint differs from the product of marginals there is some **dependence** between the variables in question.

So we understand how KL-divergence can be formulated in a way that captures shared information. But to truly understand Mutual Information we need to grasp *what information is*. A hint towards doing this is the fact that KL-divergence is also called **relative entropy**. If entropy is being used in the formulation of Mutual Information then it must have something to do with information itself.

Entropy and Information

When we observe something we are being exposed to some source of information, and a model is exploiting that information to explain and/or predict something more general.

Figure 9 shows the path from an observation to the use of a model. Making an observation is done for the sake of being surprised; we are not studying things to observe the obvious rather we wish to notice something we have not seen before. The word "surprise" might seem too colloquial to be useful in a scientific context but it actually has a specific meaning tied to information theory. Information can be thought of as the amount of surprise contained in an observation. We call this *surprisal*.

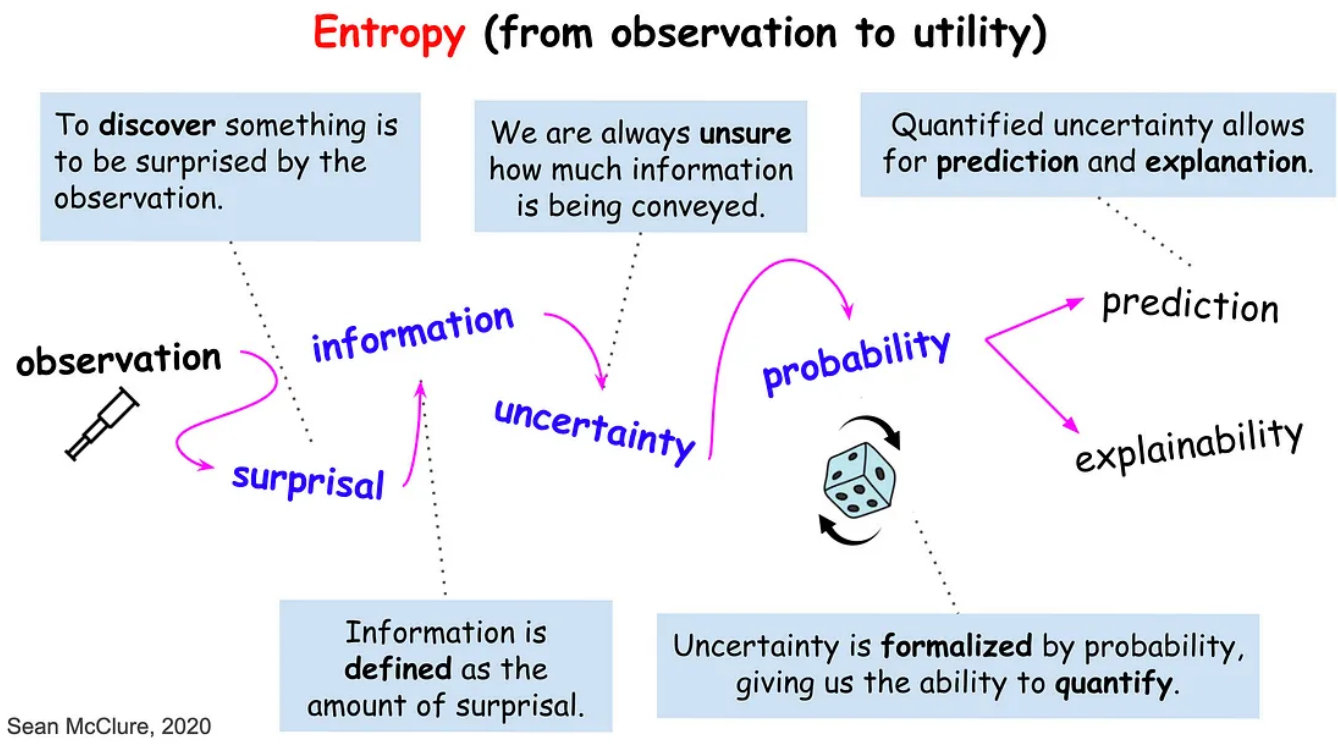


Figure 9 The path from an observation to the use of a model. Entropy oversees all these steps since they all relate back to the idea of surprisal. Die icon from icons8.

The more surprisal associated with a variable the more *information* that variable contains. This makes intuitive sense. If I find out that the details of my observation are things I already knew then it isn't useful to consider these details as information. So we can see the connection between an observation, surprisal and information.

But calling something information doesn't lend our endeavor to measurement. We need to **quantify** the amount of information in the variables we observe. This is possible by interpreting the observation — surprisal — information connection as *uncertainty*. Uncertainty connects us to *probability* since probability counts things in

terms of likely and unlikely outcomes. Probability is what gives us the mathematical framework to quantify what we observe. Finally, if we are in the realm of probability we are squarely aligned with science since probability theory can be considered the "logic of science."

But how does entropy relate all these concepts together? To answer this question let's do a simple experiment with a coin flip and use the equation of entropy to quantify the outcome we observe.

A common expression for entropy looks like this:

$$H(X) = -\sum_{i=1}^{N} P(x_i) \log_b P(x_i)$$

It is a logarithmic measure of the total microscopic states that correspond to a particular macroscopic state. We'll understand exactly what that means shortly but for now, let's see the difference in entropy between a *fair* and *biased* coin toss.

Figure 10 shows how entropy is calculated for a fair coin. We know the probabilities of heads and tails are both 0.5. A fair coin has *uniform probability* since it is equally likely to get either outcome (heads or tails). Uniform probability leads to an entropy representing the *most* we can get from a binary event.

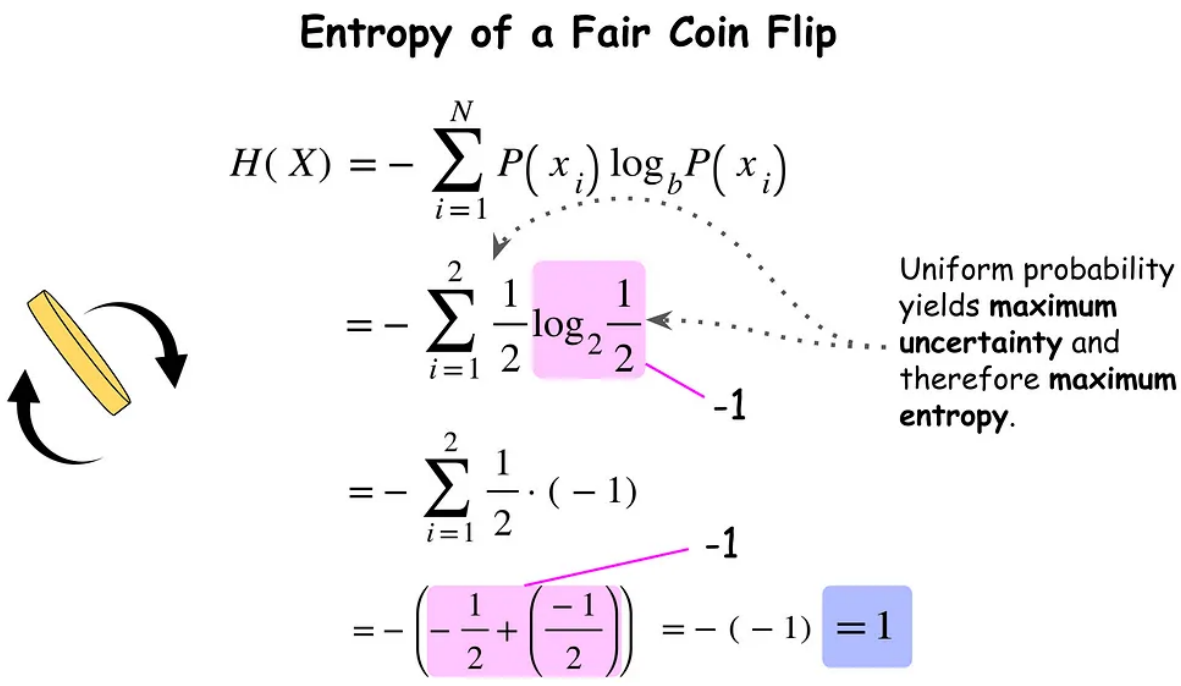


Figure 10 Calculating the entropy of a fair coin toss, using the base 2 logarithm.

> *We are using a **logarithm to the base 2**, which is standard practice when dealing with calculations involving information. There is nothing stopping you from using other bases, however this would give numbers that are less easy to work with for things like coin flips.*

What happens if we bias the coin? Let's *bend* the coin and assume[6] this leads to a *nonuniform probability* in outcome, plugging these probabilities into the entropy equation. We'll say bending the coin gives **heads** a probability of **0.7** and **tails** a probability of **0.3**:

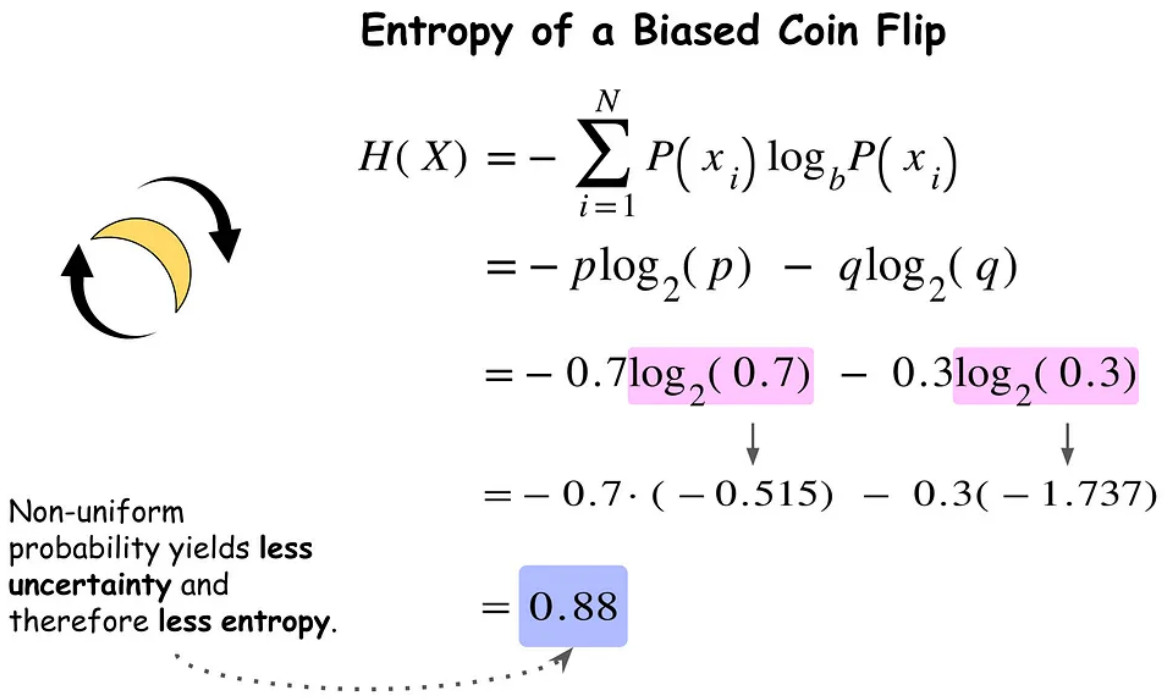


Figure 11 Biasing a coin by bending it, leading to nonuniform probabilities.

Biasing gave us nonuniform probabilities since there is now a higher chance of getting heads than tails. This affected the calculated result of entropy by lowering the value with respect to what we saw with a fair coin:

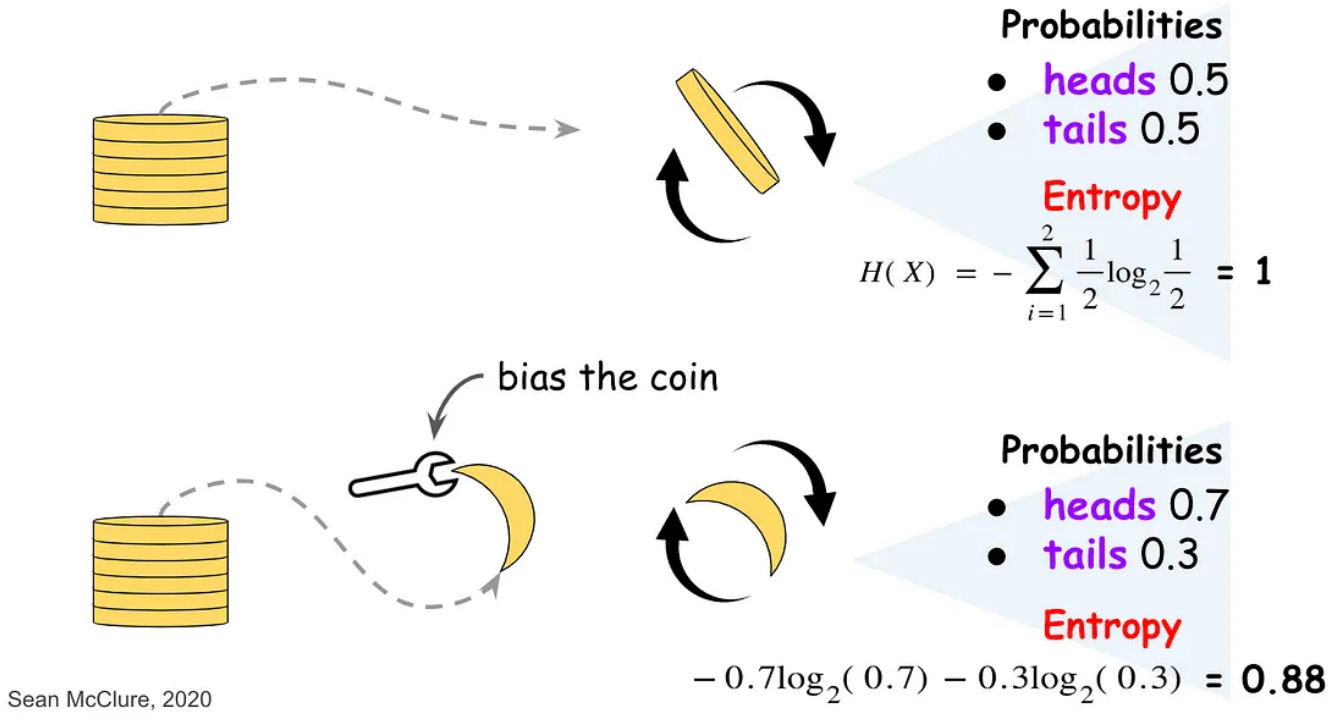


Figure 12 Calculating entropy for a fair and biased coin toss.

What does the lower entropy value with the biased coin mean intuitively? It means there is less uncertainty in the outcome. Plotting the amount of entropy associated with a single coin flip versus the amount of bias shows the following:

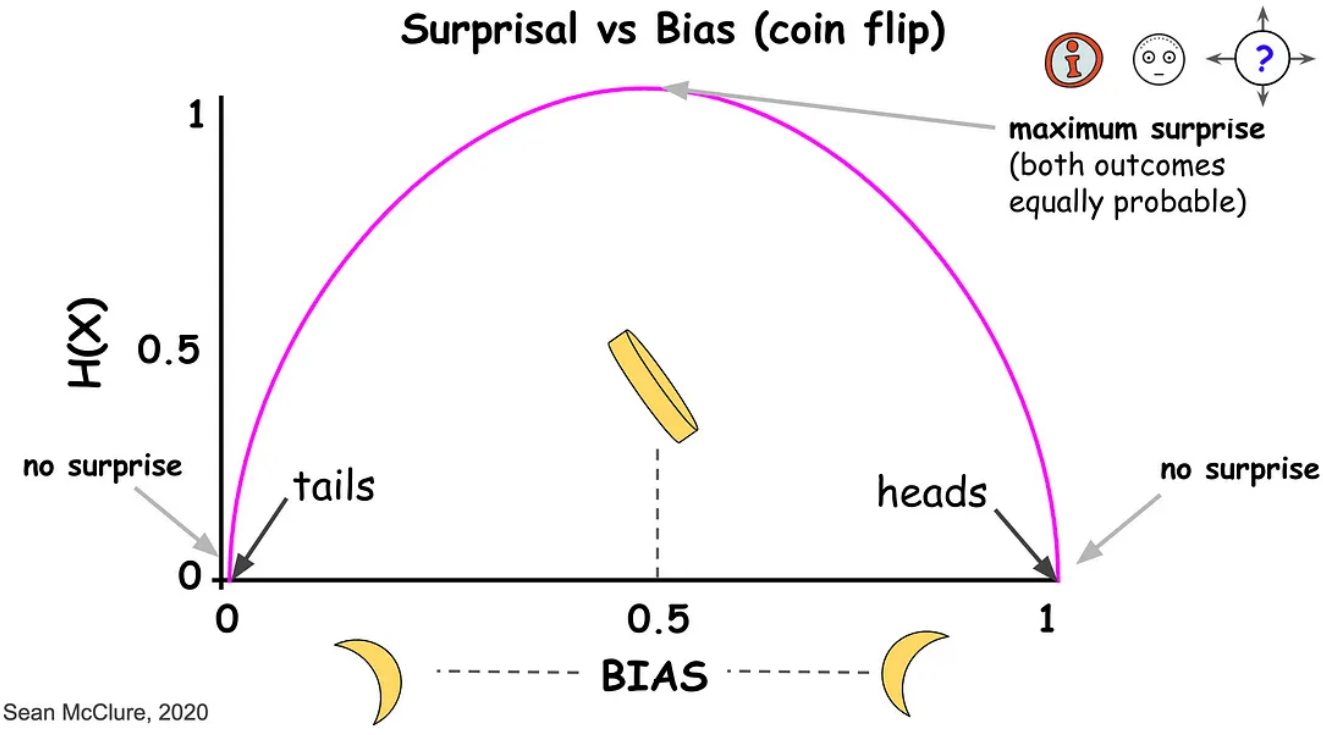


Figure 13 Change in entropy based on the bias applied to a coin. (information and face icons from icons8)

Notice the relationship between bias and surprisal. The more we bend the coin, the less entropy, the less surprised we will be of the outcome.

Let's walk through the steps of Figure 9 to relate entropy to our observation. We can see that a fair coin toss must have the most surprise, whereas a biased coin toss has less surprise. Since there is no "favoritism" when the coin is fair we have no idea to which side the coin will land, but if we bias one side there will be less **surprisal**.

We can also say that observing the result of a fair coin toss provides the maximum amount of **information**. We had no idea of the outcome prior to flipping the fair coin, so we learned more from this outcome than we would from a biased coin.

It becomes obvious that fair coins also contain the most amount of **uncertainty**. We are more uncertain about a fair coin than a biased coin. Biasing a coin reduces the amount of uncertainty in proportion to the amount of bias applied.

Remember that **probability** is how we bring math to uncertainty. We saw above that the fair coin toss has equal probabilities between heads and tails. It is these probabilities that drive the entropy value we calculate.

The result of a fair coin toss is also the least **predictable**, since there is no increased likelihood of having one event over the other. But once we introduce bias to the coin we are creating a situation of nonuniform probabilities between the potential outcomes. *We can see that a prerequisite for being able to make a prediction is the existence of nonuniform probabilities that underly the possible events*.

To wrap up our tour through Figure 9, the **explanation** or interpretation of a measurement is when we attempt to give reasons to our observation ("why do we see this"), which means finding one or more *causal* agents that contribute to what we observe. Bias underlies interpretation as much as it does prediction. A fair coin toss has nothing to interpret since both outcomes are equiprobable. With bias however there is the potential to explain our observation since there is a *reason* for the event (the predominance of one outcome over the other).

2 things worth noting with respect to probability and entropy:

1. **nonuniform probabilities** *decrease* the surprise/information/uncertainty;
2. **smaller probabilities** attached to each possible outcome increases the entropy.

The first point is apparent with the coin toss. We can understand the second point by comparing a rolled die to a tossed coin. A rolled die has higher entropy than a tossed coin since each outcome of a die toss has a smaller probability (p=1/6) than

each outcome of a coin toss (p=1/2). Remember that *all* probabilities must sum to 1 (the outcome was *realized*) thus smaller probabilities attached to each outcome means *more possible outcomes* and thus more surprise upon learning the actual outcome.

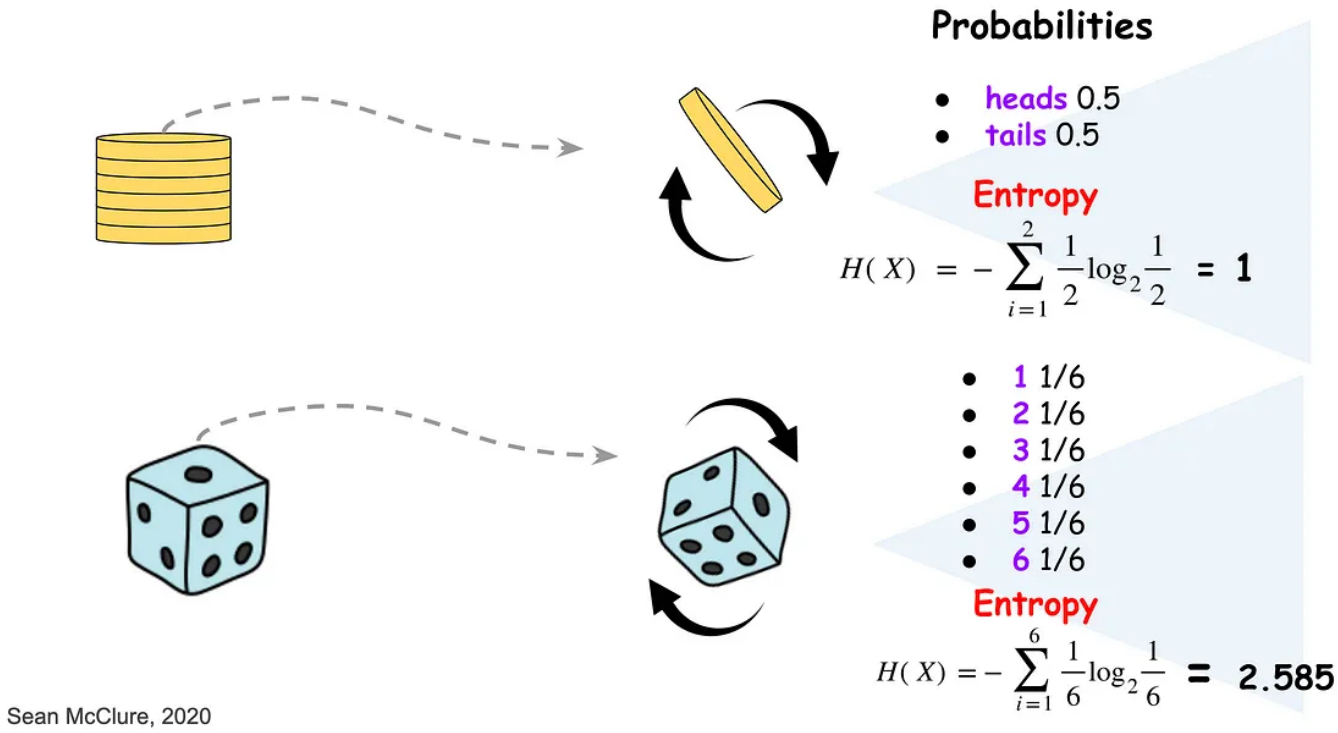


Figure 14 Difference in entropies between tossing a fair coin and rolling a fair die.

Our coin toss example showed how each of the topics in Figure 9 (surprisal, information, uncertainty, probability, prediction, and explanation) fall under the purview of **entropy**.

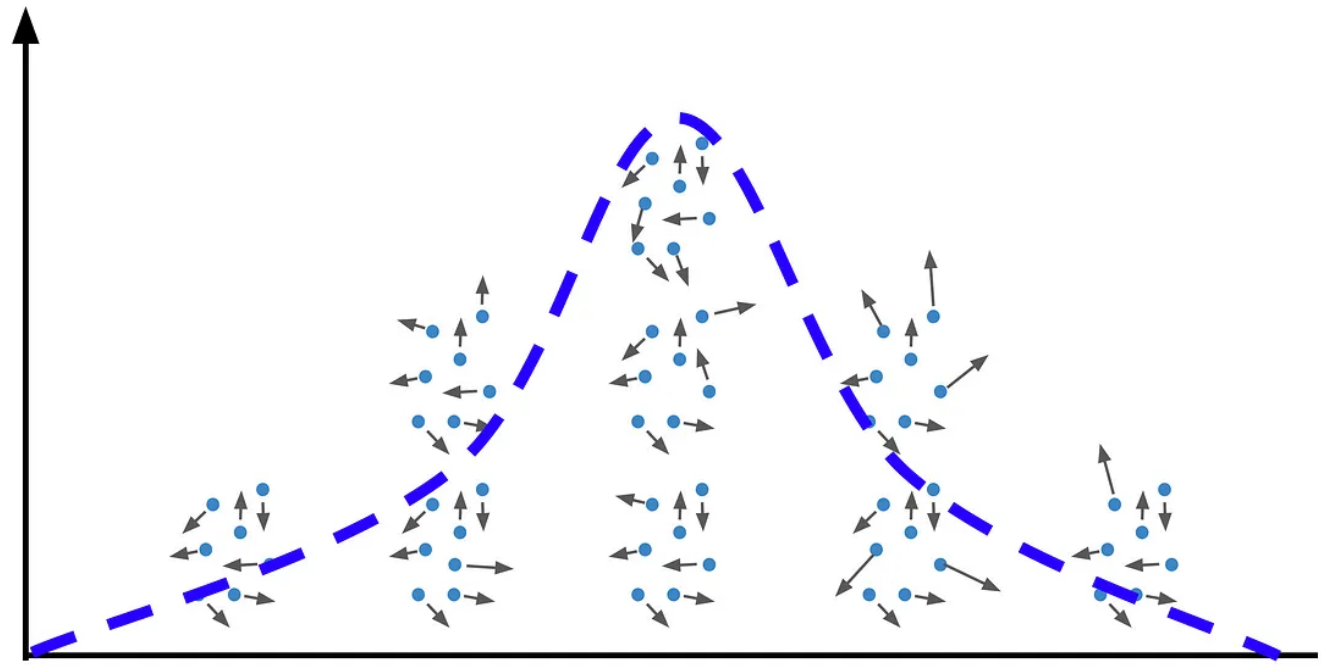
What we just covered was an information-theoretic view of entropy. While this is the most general way to think about entropy it isn't the only way. The original take on entropy was thermodynamic. To truly grasp what entropy (and information) is it's worth thinking about it both in thermodynamic and information-theoretic terms.

Thermodynamic Entropy

In 1803 Lazare Carnot realized that any natural process has an inherent tendency towards the dissipation of useful energy. Lazare's son Sadi Carnot showed that work can be produced via a temperature difference. In the 1850s and 60s physicist Rudolf Clausius provided a mathematical interpretation for the change that occurs with temperature difference, attempting to formulate *how* usable heat was lost whenever work was performed. Clausius named this idea *entropy* and considered it "the differential of a quantity which depends on the configuration of the system."

But it was scientists such as Ludwig Boltzmann, Josiah Willard Gibbs, and James Clerk Maxwell who gave entropy its statistical foundation. Importantly, Boltzmann visualized a way to measure the entropy of an ensemble of ideal gas particles using *probability*.

Figure 15 shows how we interpret an observation of a physical system probabilistically. At any given instant in time a system will have a configuration, specified by the positions and momenta of its components. We can imagine the 7 components (small blue balls) of the system in Figure 15 tumbling about every which way. Now imagine we wanted to measure some macroscopic property of the system such as its temperature. Temperature is a measure of the average kinetic energy of all the molecules in a gas. The temperature we measure is *expected* to be the peak of the probability distribution.



Sean McClure, 2020

Figure 15 Any physical system is composed of many possible configurations, with some most probable set of configurations determining what we observe.

Keep in mind, anything we observe comes from some underlying random process. This means whatever we observe when taking a measurement comes from a *distribution* rather than a specific value. The most probable outcome corresponds to the largest number of configurations that can be "stacked up." In Figure 15 there are 3 configurations that lead to the most probable temperature we measure. Adding heat to a system increases its entropy because it increases the number of possible microscopic states that are consistent with the macrostate.

The statistical mechanical interpretation of entropy concerns itself with the relationship between ***micro*states** and ***macro*states.** A *micro*state is a specific microscopic configuration of a thermodynamic system that the system may occupy in the course of its thermal fluctuations. A *macro*state is defined by the macroscopic properties of the system, such as temperature, pressure, volume, etc. In Figure 15 there are 3 *micro*states that correspond to the most probable *macro*state.

Entropy measures the degree to which the probability of the system is spread out over different possible microstates. With respect to Figure 15 entropy would attempt to quantify how many possible microstates make up the total distribution.

Back to Information-Theoretic Entropy

We first introduced entropy in terms of information, looking at coin flips as our example. There is a striking similarity between the information and thermodynamic interpretations of entropy. Both quantify the uncertainty by considering *all* possible outcomes that could lead to the event.

We can think of our coin flipping example in a similar way to thermodynamic entropy, where the *micro*states are all the possible outcomes and the *macro*state is the event we observe. This is more obvious when we flip at least 2 coins as shown in Figure 16.

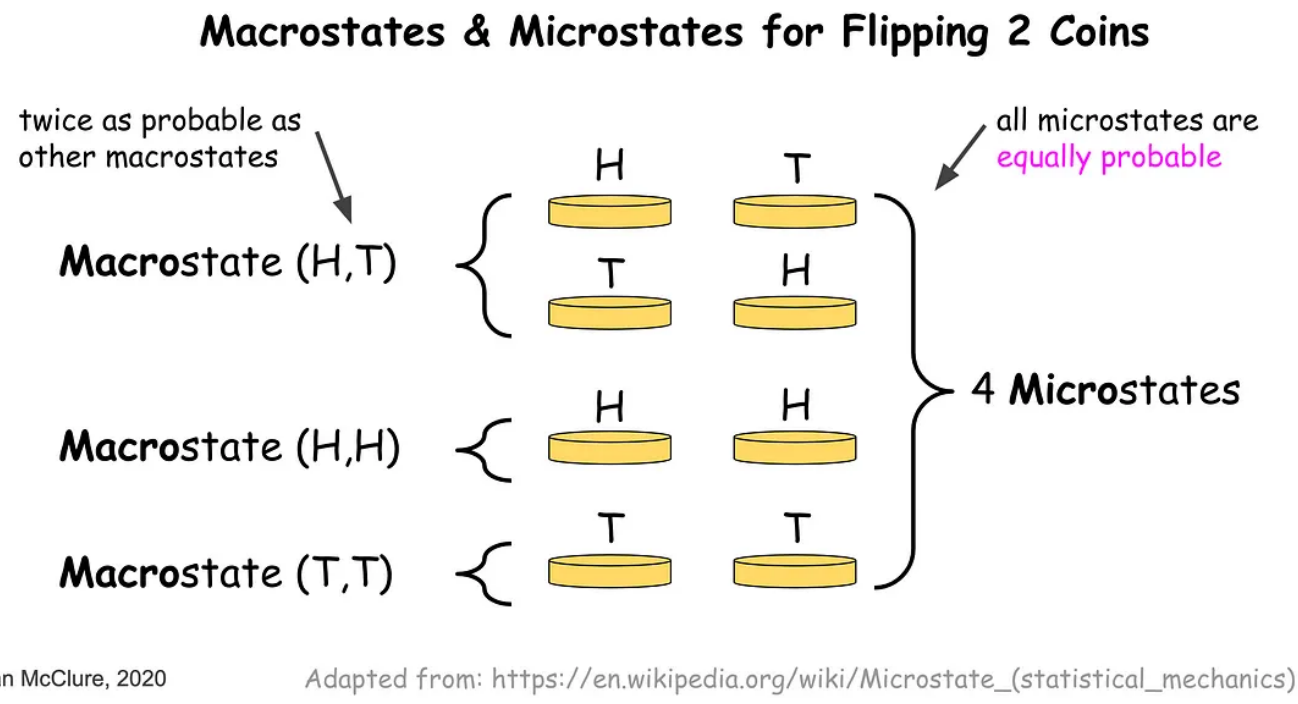
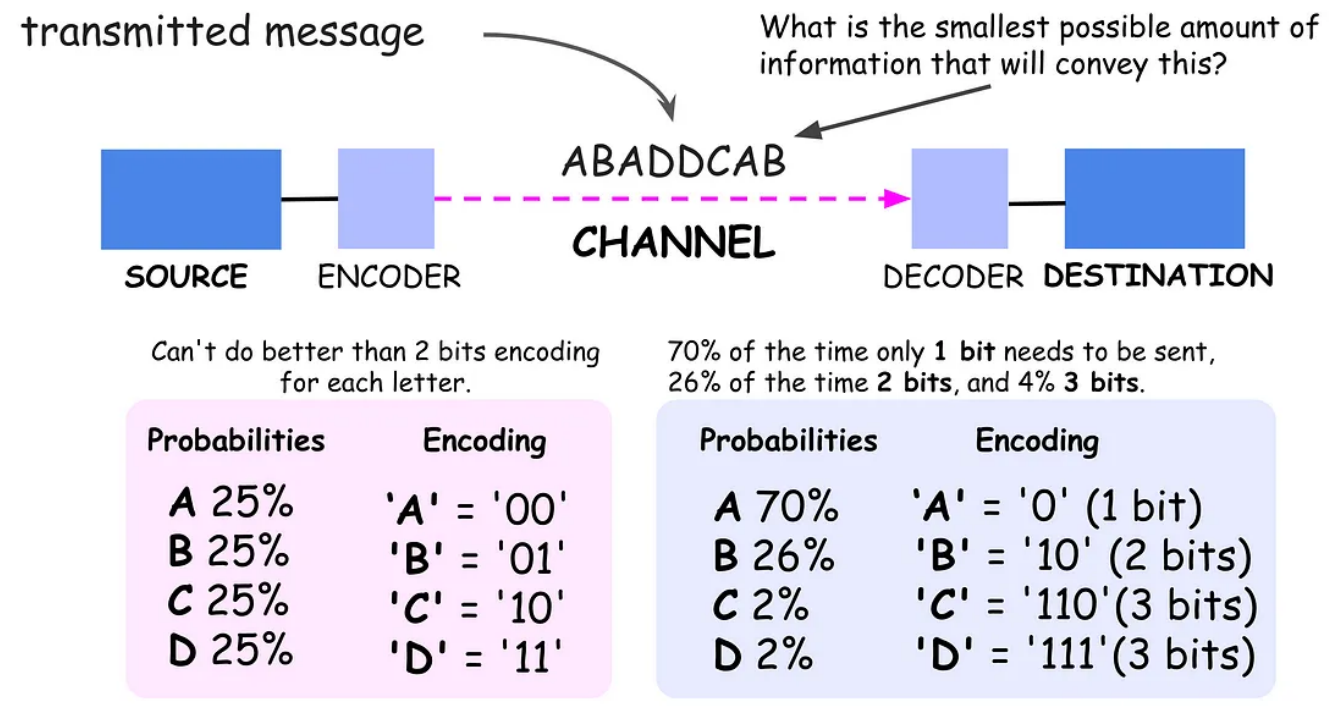


Figure 16 Thinking of coins in terms of microstates and macrostates.

If you were asked to guess which combination of heads and tails is most likely in a fair coin toss (with 2 coins) you could confidently answer "1 heads and 1 tails." The reason is there are 2 *micro*states that lead to the (H,T) *macro*state, whereas (H,H) and (T,T) *macro*states have only 1 each. This is directly analogous to what we saw with thermodynamic entropy. It's merely a matter of realizing that *the most likely outcome for an event is the largest set of configurations that lead to the same observation*.

Our examples in Figures 15 and 16 are simple. But imagine a weather system with pressure gradients, temperature fluctuations, precipitation, etc. We are talking orders of magnitude more complexity than our examples above. But weather is still some set of possible configurations between matter and energy, and what we experience is expected to be *the largest set of configurations that lead to the same observation*.

Let's look at an example more inline with how we think about information. **Compression** is a technology used to reduce the memory footprint of a file. Our compression is "lossless" if we can always recover the entire original message by decompression. The following image shows the well-known example of communicating over a channel with a message passed between source and destination. Encoders are used to convert the original message into another, usually shorter representation to be transmitted. A decoder can then convert the encoded message into its original form.



Sean McClure, 2020

Figure 17 Communicating a message over a channel.

Entropy rears its head again in this situation. In our previous example we saw how entropy quantifies the amount of uncertainty involved in the value of a random variable (outcome of a random process). This idea was made concrete in 1948 by Claude Shannon in his paper "A Mathematical Theory of Communication." Shannon was interested in finding a way to send messages over a noisy channel such that they could be reconstructed at the other end with a low probability of error.

Let's look at 2 different scenarios of data compression and compare them with respect to entropy.

Say we want to know the smallest amount of information that will convey our message ABADDCAB assuming we can only use 1s and 0s. Let's come up with 2 possible encodings. One option is to use 2 bits for each letter, so **A** is 00, **B** is 01, **C** is 10, and **D** is 11. This would work. Another option would be to have **A** coded as 0, **B** as 10, **C** as 110 and **D** as 111. This would also work.

If we tally the total bits we see that our first option uses less bits (8 bits) than our second option (9 bits). At first blush it seems as though our first option is better. But what happens if the letter **A** in our message occurs with 70% probability, **B** with 26%, and **C** and **D** with 2%? In this case we could assign 1 bit to A, which would lower the average number of bits required to send (compress) the message because 70% of the time only one bit needs to be sent, 26% of the time two bits, and only 4% of the time 3 bits. The entropy is lower in our second scenario because it requires less than 2 bits *on average*, whereas the first scenario requires 2 full bits on average (8 bits / 4 letters).

So we can see how the existence of nonuniform probabilities determines how well we can compress a message since it changes its overall entropy. This is just another version of what we stated previously with respect to the relationship between probability and entropy; nonuniform probabilities decrease the surprise/information/uncertainty. If **A** occurs 70% of the time there is obviously *less surprise* in the message, which we now know means less information.

> *We have to be careful here with what we mean when we say less information. We are not removing information from the original message. Rather we are determining how much* actual *information was in the original message and transmitting that using less information. If we treated each letter in the message as equally probable we would be*

> *adding **redundancy** into the transmitted message and transmitting it with more information than necessary.*

To be clear, our compressed message has the same quantity of information as the original just communicated in fewer characters. It has more information (higher entropy) per character. Data compression algorithms find ways to store the original amount of information using less bits, and they do so by removing redundancies from the original message.

Going back to our path from observation to prediction (Figure 9) we can again see how each step plays out under entropy. A message with nonuniform **probabilities** in outcomes (A, B, C, D) decreases the **uncertainty**, which means less **surprisal**, which means less **information** required to transmit the message. Imagine trying to predict the message in Figure 12. Obviously we could **predict** better under the 2nd scenario since A occurs 70% of the time. And just as a biased coin allowed for an **explanation** of what was observed so too does the bias present in communication with a preponderance of one letter over the others.

The Deeper Connection

A big question regarding entropy is whether its thermodynamic and information-theoretic interpretations are different versions of the same thing. The parallels between the two are undeniable, and many regard the information-theoretic version of entropy as the more *general* case, with the thermodynamic version being the *special* case.

> *Information-theoretic entropy is thought of as more* general *since it applies to any probability distribution whereas thermodynamic entropy applies only to thermodynamic probabilities. It can be argued that information-theoretic entropy is much more universal since any probability distribution can be approximated arbitrarily closely by a thermodynamic system[2].*

Regardless where you stand on the debate there is an advantage to thinking of *all* entropy in terms of its information-theoretic interpretation. It allows us to see any physical system as having information content, and any physical process as the transfer of information. This interpretation will help us understand how Mutual Information gets to the heart of dependency and what it means for information to be shared.

To begin our deeper dive into this connection let's look at the original statistical mechanical equation found by Boltzmann, shown in Figure 18.

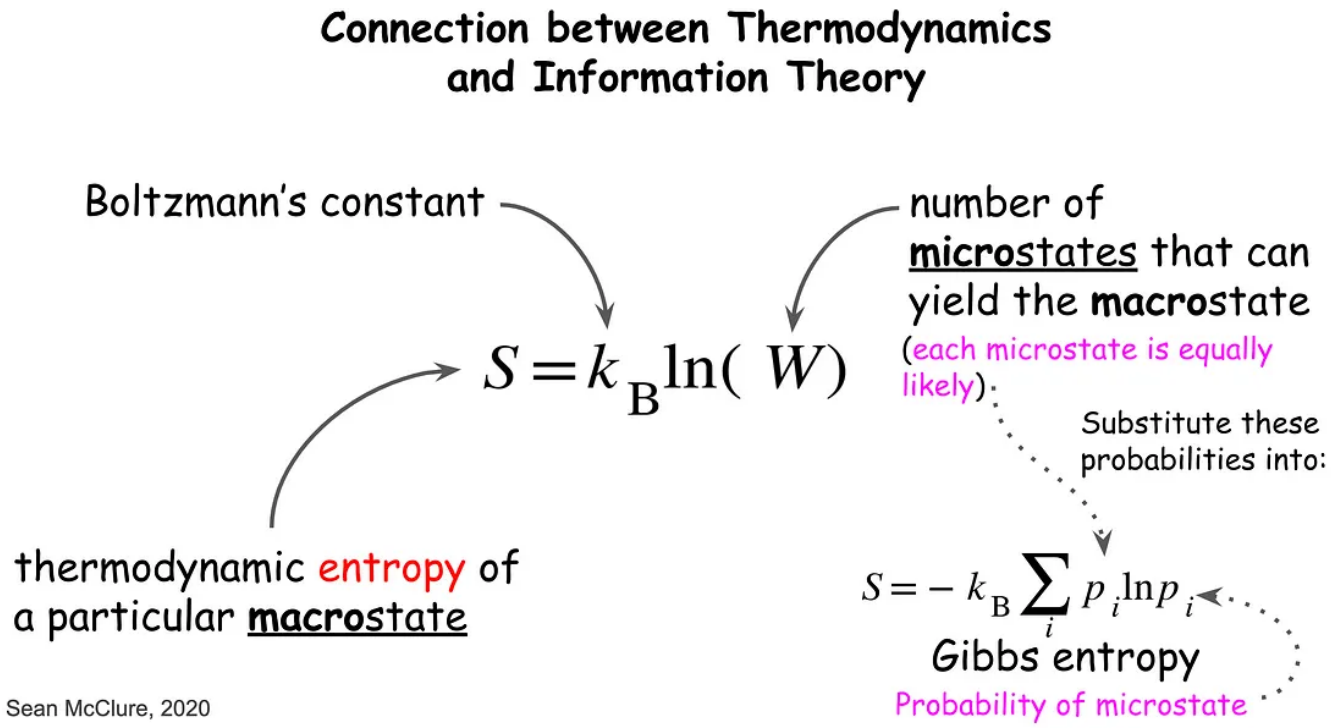


Figure 18 The original thermodynamic expression for entropy.

The first equation is the one on Boltzmann's tombstone, except above written with the natural logarithm. The type of logarithm used just determines the units entropy will be reported in, while Boltzmann's constant **k** simply relates everything to the conventional units of temperature. The important part is the W, which is the number of microstates that lead to the macrostate. In thermodynamic terms the microstates might be the individual atoms or molecules that make up some physical system. This form assumes all the microstates are equiprobable (a *microcanonical ensemble*), where W is the number of microstates.

Gibbs extends the simple equation of Boltzmann that uses equiprobable probabilities to account for nonuniform probabilities. You can see how this is accounted for in the Gibbs entropy equation in the bottom right. By using *individual probabilities* and summing them we allow for distinct contributions from different microstates.

The use of the logarithm is fundamental to understanding why Boltzmann came up with this expression, so let's explore that now.

The Intuition Behind Logarithms

Entropy is routinely measured in the laboratory and is simply heat divided by temperature. It has a well-known one-way behavior corresponding to the 2nd Law of Thermodynamics, which states that the total entropy of an isolated system can never decrease over time. This was all known prior to Boltzmann as practitioners had a strong intuitive appreciation of entropy's behavior[5].

But Boltzmann wasn't satisfied with the definition of entropy as simply heat over temperature since this doesn't tell us what entropy *is*. Some kind of model was needed that aligned with what was known about the motion of atoms and molecules.

If you fill a container with gas and connect it to another container the gas will fill both containers. The only thing changed here is the arrangement of matter (not temperature, pressure, etc.). But the system still exhibits the one-way behavior associated with entropy, so something is still increasing. Boltzmann realized that entropy must have something to do with the *arrangement* of matter rather than some usual thermodynamic quantity.

Boltzmann realized that *in any random process the final arrangement of matter was more **likely** than the initial arrangement*. This thinking allowed Boltzmann to connect the known concept of entropy to **probability**. This was a major shift in science. It meant the usual approach to understanding reality by defining properties like position, speed, weight, size, etc. of physical things *could now be recast purely in terms of uncertainty, and thus information*.

But how does the arrangement of matter get formalized in an equation? We already saw from Figure 15 that a physical system is composed of many possible configurations with some most probable set of configurations determining what we observe. Another way of saying this is that the probability of finding a particular arrangement is related to the **number of ways** a system can be prepared[5].

If I asked you which outcome from rolling a pair of dice was more likely, a 12 or a 7, what would you say? There is only 1 way to roll a 12 (1 in 36 chance) but 6 ways to roll a 7 (6 in 36 change), so obviously we are more likely to roll a 7. This is the same reasoning we applied above to the most likely outcome between heads and tails in a 2-coin flip. This property gave Boltzmann a path towards describing the entropy of a physical system in terms of the number of ways it can be arranged.

There is a problem though. Whereas the entropy used by scientists is **additive**, the number of ways we can arrange things is **multiplicative**. For example, a single die can fall 6 ways (6 sides) but 2 dice can fall 36 ways (6 X 6). How can entropy, which is additive (a system twice as large has double the entropy), be described by a multiplicative process?

This is where the logarithm comes in. **Logarithms** convert multiplicative quantities into additive ones. Entropy is simply the log of the number of ways a system can be arranged, converting an underlying multiplicative phenomenon (things combining) into an additive tool. Of course the number of ways we can arrange physical systems is very (very!) large, estimated around 10^{23}. To bring entropy back into the kind of magnitude's we're used to using in everyday situations Boltzmann multiplied the expression by a "fudge factor", which we now call Boltzmann's constant ($1.38064852 \times 10^{-23}$).

Figure 19 shows the difference between exponential growth viewed on a linear scale and the same growth viewed on a logarithmic scale. Whereas the linear scale shows explosive behavior the logarithmic scale "tames" this into a simple straight line.

> *A common approach to detecting exponential growth is to see if it forms a straight line on a logarithmic scale.*

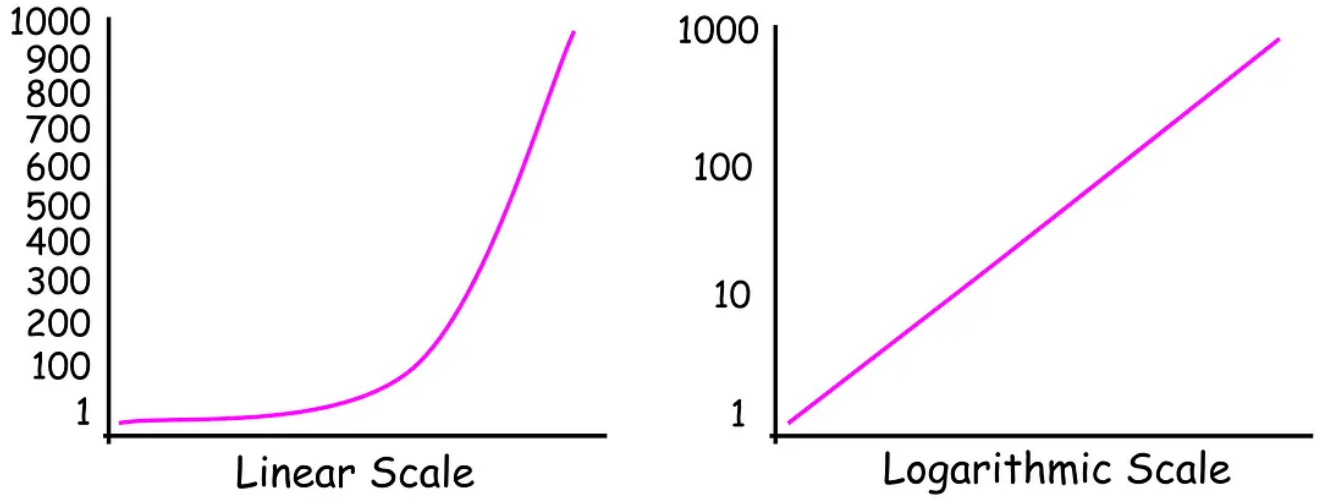


Figure 19 Logarithms allow us to view an exponential process as if it were an additive (linear) process.

Exponential growth is associated with multiplicative processes. Graphically, think of the logarithm as *offloading the explosive growth from a multiplicative process onto the axes* such that the function itself no longer has to bear that growth. If you prefer algebraic interpretations, recall that the logarithm is the inverse function to

exponentiation. This means the logarithm of a given number x is the exponent to which another fixed number (the base) must be raised to produce x. If we keep taking the logarithm of a massive number we will get back a smaller number (the exponent) that counts by 1.

Figure 20 depicts the use of a logarithm as the representation of some combinatorial explosion in terms of simple linear growth. Just as 2 dice "explode" in the number of possible combinations, so too do the various configurations a system can take on. Logarithms are what make working with massive numbers (like the number of ways atoms can be arranged in a material) more manageable.

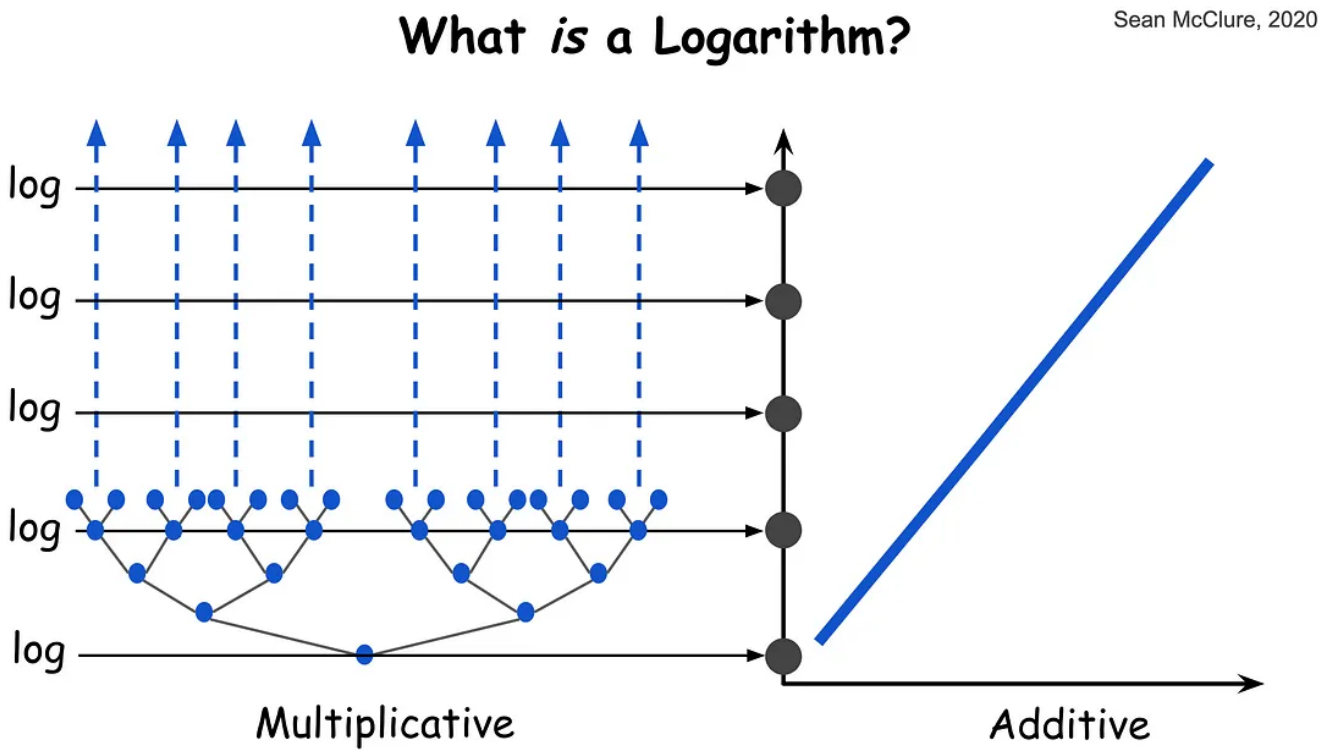


Figure 20 The relationship between a combinatorial explosion and the logarithm used to represent the process as simple linear growth.

A critical realization here is that the use of logarithms takes into account the complexity of the problem. In other words, we don't toss away the combinatorial behavior of systems when we simplify them with logarithms. They are "brought along for the ride" despite our simplification of the problem. Contrast this to something like Pearson's correlation. Pearson's correlation also simplifies a system by depicting it as simple vectors pointing in similar directions, BUT it does so by ignoring the complexity of the problem.

> *The goal in science is not to reduce phenomena down to the simplest description, it is to reduce phenomena down to the simplest description that retains the core properties of the*

> *system.*

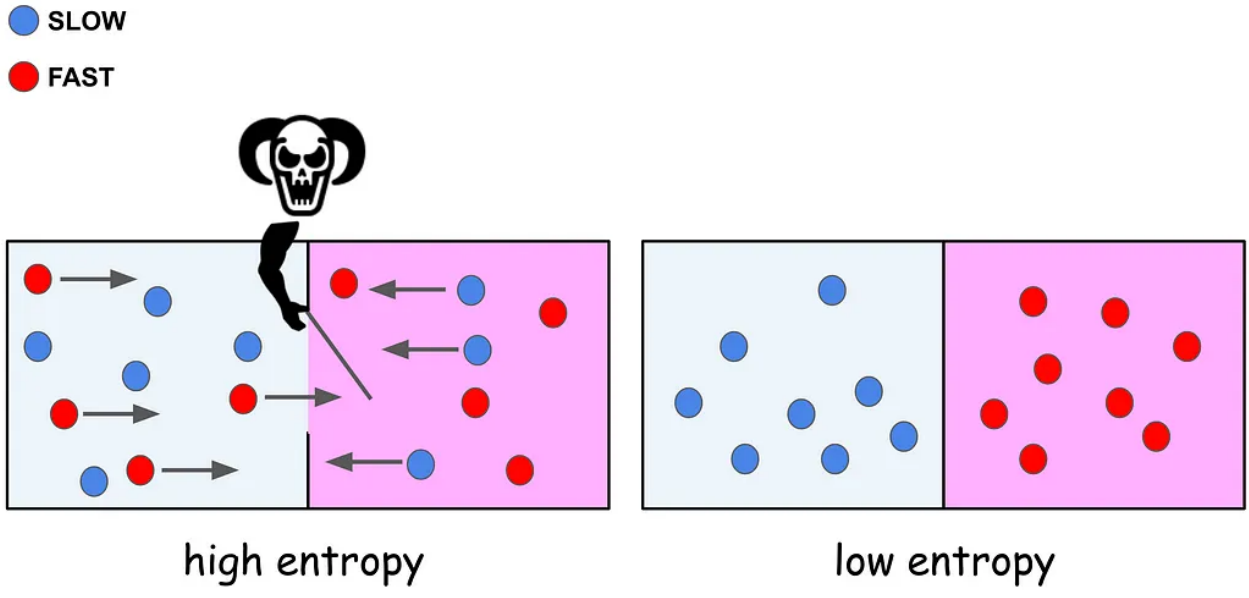
Is Information Physical?

By showing the fundamental connection between the physical arrangement of matter and information we can better understand Mutual Information. To build this case let's look at some well-known **thought experiments**.

The first is **Maxwell's demon**. Maxwell's demon was put forward by physicist James Clerk Maxwell to suggest how the second law of thermodynamics might be violated. He argued it might be possible to create a temperature difference in a gas *without expending work*.

In this thought experiment Maxwell imagines a demon who operates a trap door between 2 compartments of a closed box. While average particle velocities are fairly constant, individual gas molecules travel at fast (red) and slow (blue) speeds. When a given molecule approaches the trap door the demon opens and shuts the door such that all fast molecules end up in one chamber and slow molecules in the other. Note that the shutter is frictionless and thus no work is performed by the demon.

This scenario leads to one of the chambers being hotter than the other. We now have a situation where we have decreased entropy (since there is more order). This results in a temperature difference that could be exploited in a heat engine, and thus we have apparently violated the second law of thermodynamics (entropy only increases).



Sean McClure, 2020

Figure 21 Maxwell's Demon, showing how thermodynamics might hypothetically be violated by demonstrating the possibility of a gas evolving from a higher to a lower entropy state. (demon and arm icons from icons8)

Maxwell's thought experiment was a reasonable challenge to the idea that thermodynamics depends fundamentally on atomic physics (and as such could be treated probabilistically). After all, if it was just a matter of *knowing* more details about a system (individual particle velocities) then isn't it possible we could *arrange for a decrease in entropy to occur?*

This all suggests that small particles might be open to exploitation, making Maxwell's Demon realizable. Importantly, Maxwell's argument suggests entropy has a *subjective* quality to it. The value of entropy one measures depends on the amount of information one has about the system.

We don't regard this thought experiment as a true demonstration of the violation of the 2nd law of thermodynamics since it's expected the demon would in fact increase entropy by segregating the molecules. But the argument stands; theoretically a sufficiently intelligent being could arrange for a perpetual device to be constructed, and so more insight is needed to bring the 2nd law back into focus.

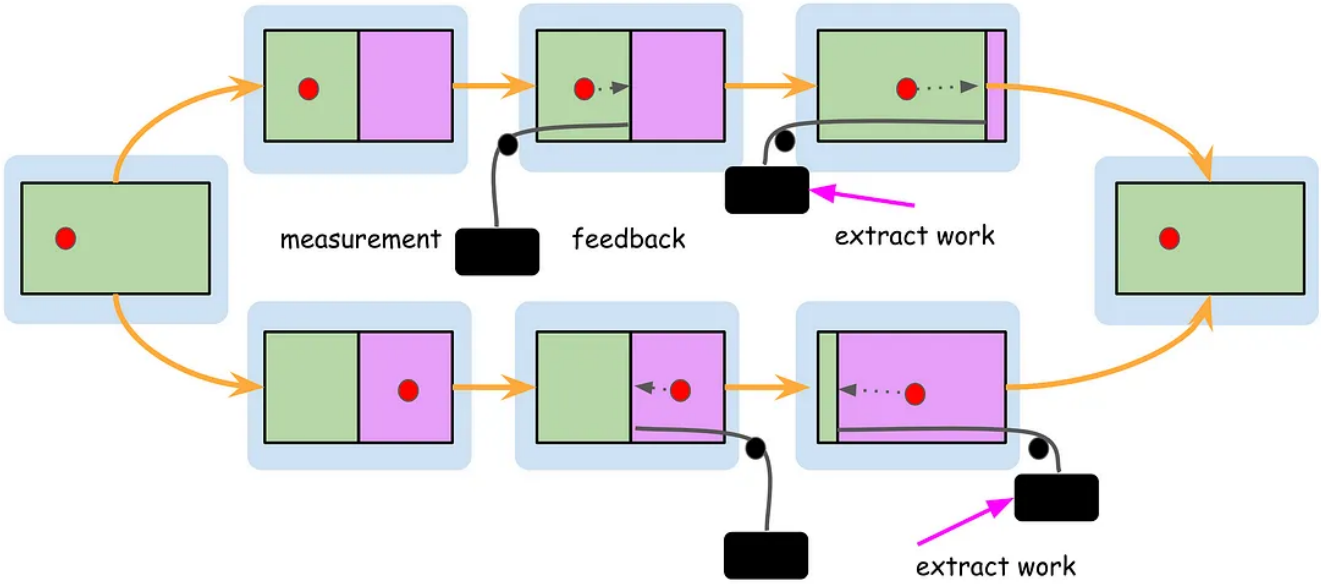
We now know there is some fundamental connection between obtained knowledge and entropy. For non-demon entities like us the acquisition of information means *measurement*. Given Maxwell's argument we can say that measurement is required for entropy reduction to take place (to gain knowledge of the system's configuration). The outstanding question is this: assuming the 2nd law cannot be violated, **what is the compensating entropic cost of this measurement**?

This brings us to **Leo Szilárd**, who had his own version of the Maxwell's thought experiment using a single molecule. Imagine an empty box with no partition in contact with a heat bath (e.g. surrounding container of water). A partition can then be inserted into the box, which divides the container into 2 volumes. The partition is also capable of sliding without friction, left or right. As soon as the partition is added to the container collisions between the molecule and the partition exert a pressure on the partition. Since the partition moves after being inserted we could theoretically add a pulley to the partition that lifts a weight and thus extracts work (molecule hits partition -> exerts pressure -> partition moves -> pulley moves -> extract work).

If we added the partition to the middle of the container we then create a situation where the single molecule would be on either side with *equal probability*.

Now imagine we *knew* which side of the box the molecule was and we *then* introduced a partition in the middle. We could hook up a pulley and weight system and extract work a shown in Figure 22. The work extracted is drawn from the heat bath due to thermal contact. When the partition reaches all the way to one side of the container it can be removed, representing one full cycle. This process can be repeated indefinitely with work continually extracted.

The above description is known as **Szilard's Engine**:



Sean McClure, 2020

Figure 22 Szilard's Engine is a modification of Maxwell's Demon thought experiment showing how work could theoretically be extracted using only the acquisition of information.

In Szilard's single molecule experiment the possession of a single bit of information corresponds to a reduction in the entropy of the physical system (remember, if something moves towards equilibrium then it must have less order, more uncertainty, more entropy). Szilard's engine is thus a case of *information to free (useful) energy conversion* since we had to have possession of information (where the particle was) in order to position a piston such that work could be extracted. Without possession of *information* we wouldn't know how to hook up the pulley system to extract work. This means that in order for the 2nd law to not be violated **the acquisition of knowledge must be accompanied by some entropic cost.**

Szilárd solidifies the relationship of Maxwell's demon to information by connecting the observer via *measurement*. Rather than some supernatural demon being in control of the information, *we* are in control of the information, and can use this to affect the entropy situation. The take-home message here is that Szilárd showed how **the possession of information can have thermodynamic consequences.**

The work of Maxwell and Szilard opens up a new set of questions regarding the **physical limitations of computation**. Between Szilard and some other key players (Brillouin, Gabor, Rothstein) it became apparent that the acquisition of information, via measurement, required a dissipation of energy for every bit of information gathered. More generally, as suggested by von Neumann, every act of information *processing* was necessarily accompanied by this level of energy dissipation[2].

This brings us to Rolf Landauer who in 1961 used his version of a thought experiment. Imagine the Szilard's engine starting in a thermalized state (equilibrium). There would be a reduction in entropy if this were re-set to a *known* state. The only way this would be possible is under *information-preserving microscopically deterministic dynamics*, and such that the uncertainty was "dumped" somewhere else. This "somewhere else" could be the surrounding environment or other degrees of freedom that are non information-bearing. In other words, the environment would increase in heat, and again, the 2nd law of thermodynamics is preserved.

There must be some physical substrate on which computation occurs. Landauer argued that any physical system designed to implement logical operations must have physical states that correspond to the logical states.

To understand this we need to appreciate the distinction between *logically reversible* and *logically irreversible* operations.

> *An operation is **logically reversible** if the input state can be uniquely identified from the output state[2].*

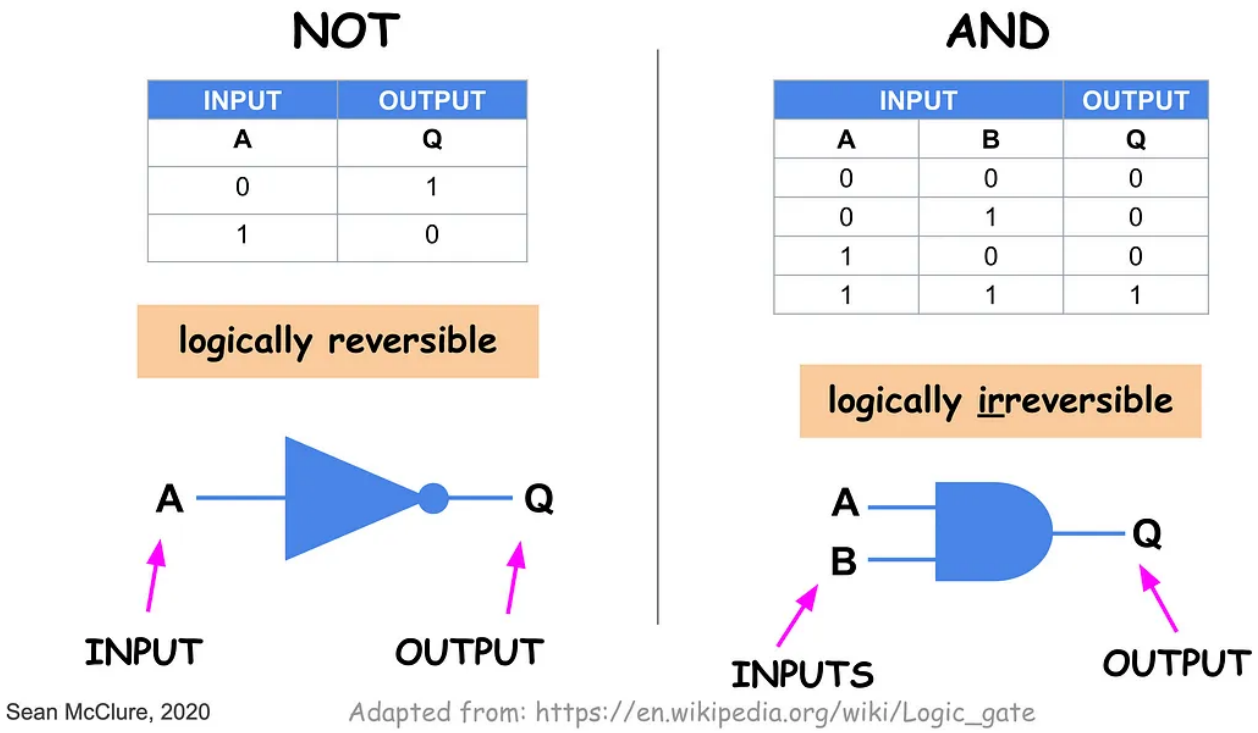


Figure 23 The difference between logically reversible and logically irreversible operations.

The **NOT** operation is an example of a **logically reversible** operation. This is because if the output is 1 then the input *must* be 0, and vice versa. The **AND** operation is an example of a **logically irreversible** operation since if the output is 0 there are *multiple possible inputs* (3 of them in this case), as shown in the above figure. Note that logically reversible operations must have a 1-to-1 mapping.

Reversible operations do not compress the physical state space (the set of all possible configurations of a system). On the other hand, irreversible operations compress the logical state space (look at the table in the above figure; 3 combinations of input led to a single output) and thus *do* compress the physical state space. Landauer argued that a **compression of the physical state space must be accompanied by a corresponding entropy increase in the environment** (via heat dissipation).

It turns out most logical operations are irreversible and thus, according to Landauer, must generate heat[2]. The most basic logically irreversible operation is **resetting** a bit. Landauer used this idea to *quantify* the heat generation described above.

Take as an example 2 input states (0 and 1) that outputs to 0 (a reduction in logical state space) as shown in Figure 24. Similar to the Szilard engine we have a container in thermal contact with a heat bath, with a single molecule and a partition in the

middle. We will say that the molecule on the **left** is logical state **0** and the molecule on the **right** is logical state **1**.

Now remove the partition and allow the single molecule to roam freely throughout the container. Now *reintroduce* the partition on the far right side and push it towards the center. We know from our previous discussion that the single molecule will exert a pressure on the partition, which requires **work** to be performed (again, the energy from this work is transferred to the heat bath).

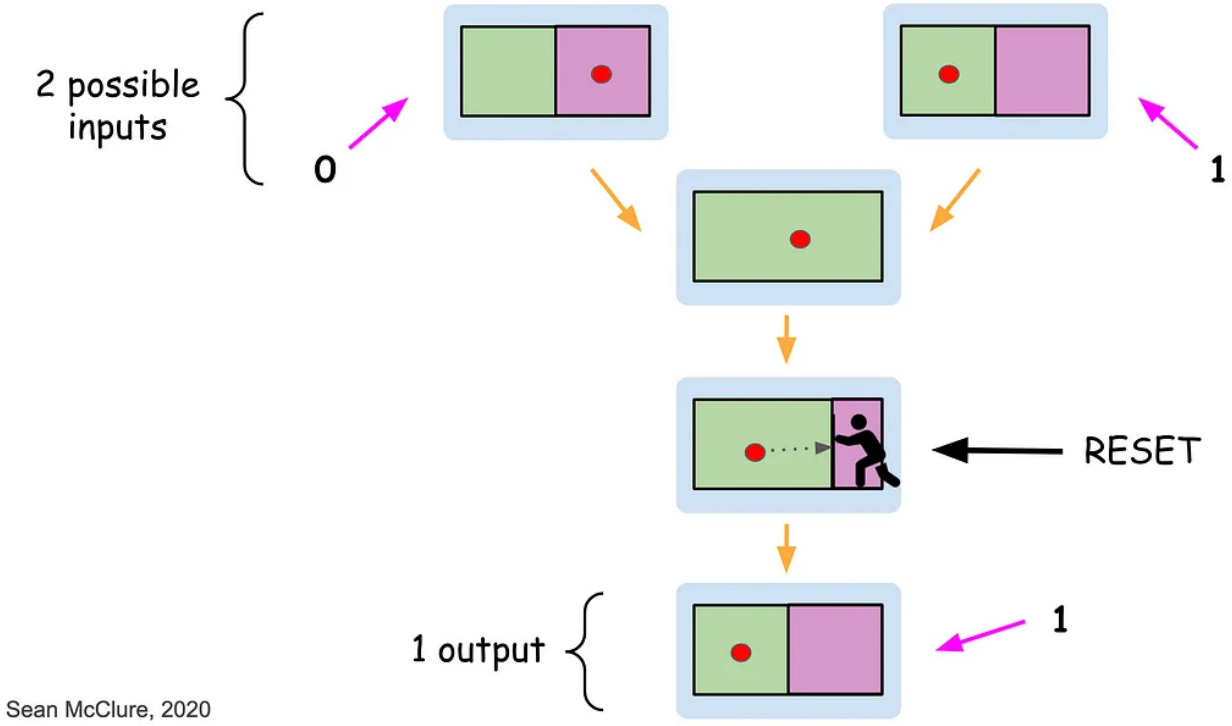


Figure 24 Landauer's thought experiment showing an irreversible operation that moves from 2 possible inputs to one definite output.

We have arrived at something called **Landauer's principle**, which states *there are no possible physical implementations of the resetting operation that can do better than this* (reset a bit to zero converting less than a given amount of work into heat). This amount turns out to be $kT \ln 2$.

This is known as **resetting**, also called **erasure**.

Landauer argued that by only referring to the abstract properties of a logical operation we can deduce a thermodynamic constraint upon *any* physical system that performs a logical operation. *There is a definite connection between abstract logical functions and their physical implementation*. Landauer's Principle makes a strong case for the idea that **information is physical**.

Information as the Number of Yes-No Questions

This is where the real connection between the information-theoretic and thermodynamic views of entropy come into play. We saw from the thought experiments of Maxwell, Szilard and Landauer that entropy can be viewed as the amount of information needed to define the detailed microscopic state of the system, given what is known at the macroscale. Specifically, it is the minimum number of yes/no questions that need to be answered in order to fully specify the microstate, given that we know the macrostate.

At this point we realize that whatever we observe at the macroscopic level does not communicate the detailed information about a microscopic state. And from Landauer we realize that if information has a physical representation then the information must somehow be *embedded* in the statistical mechanical degrees of freedom of that physical system.

What Exactly is Dependency?

Our tour through the physical interpretation of entropy wasn't for historical interest or for forcing some intuitive "analogy." The physical is needed to grasp what dependency is. We already know the *physical* biasing of a probability distribution changes the entropy, and thus the amount of information contained in a variable. We first saw this when we bent the coin, and then with Szilard's engine, placing the partition at some arbitrary location (rather than the middle). What we are left with is the following realization:

> Dependency, at its core, is the coincidence of physical configurations between 2 or more things being measured.

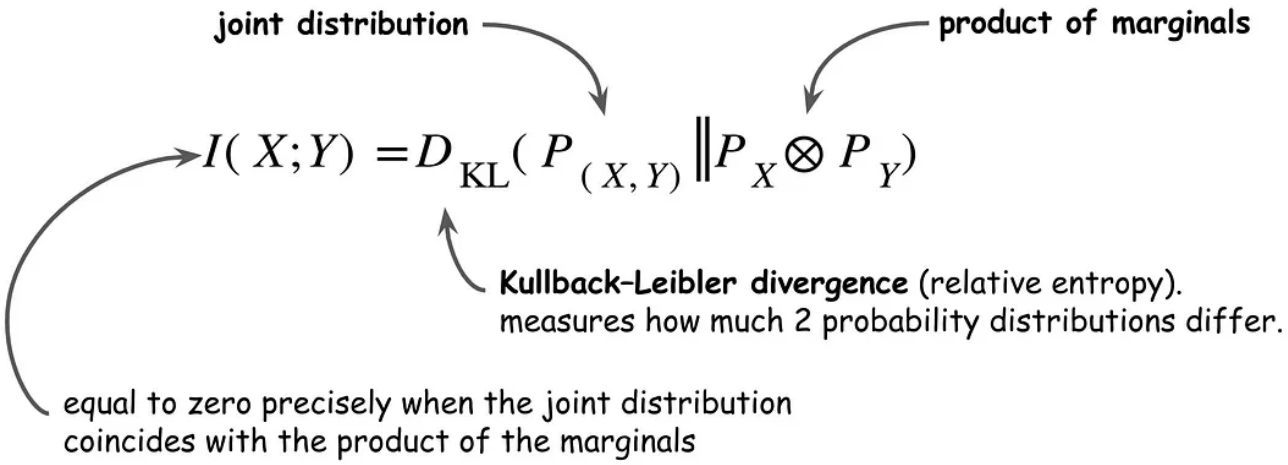
To shirk the physical underpinnings of information allows one to get away with any kind of statistical measure of correlation. But when a physical substrate is accounted for such simplistic notions of dependence are not possible. *Pearson's correlation doesn't look at the configurational difference between variables*. It instead uses a summary (variance) of an assumed distribution (Gaussian) and compares these using vectors of data points.

But mutual information *does* look at the physical configurational coincidence between variables since it looks at how physical outcomes of a random event differ.

These outcomes are captured by a probability measure untethered to any specific distribution, and this is what makes MI a *general* measure of dependence.

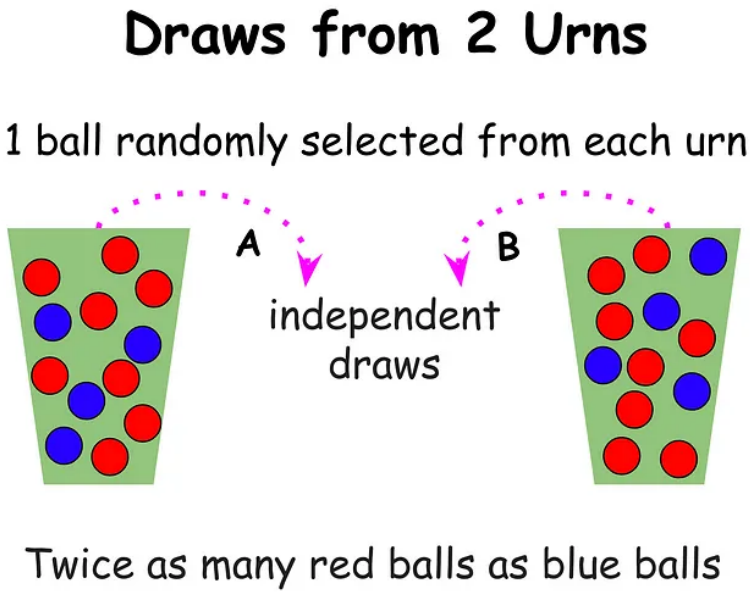
Mutual Information in the Wild

Let's bring our conceptual tour full circle by revisiting the definition of mutual information, originally displayed in Figure 5:



We know MI uses KL-divergence, meaning it uses the discrepancy between a joint distribution and the product of marginals to measure information. Let's build our intuition around **joint** and **marginal** probabilities using the classic urn example.

The following figure shows blue and red marbles being drawn from 2 urns:



Sean McClure, 2020

Figure 25 Independent draws from 2 runs.

Consider **A** and **B** as discrete *random variables* associated with the outcomes of the draw from each urn. The probability of drawing a red marble from either of the urns is **2/3** since there are 8 red marbles out of 12 total marbles (in each urn).

In probability we typically write out the possible counts in a table as shown below:

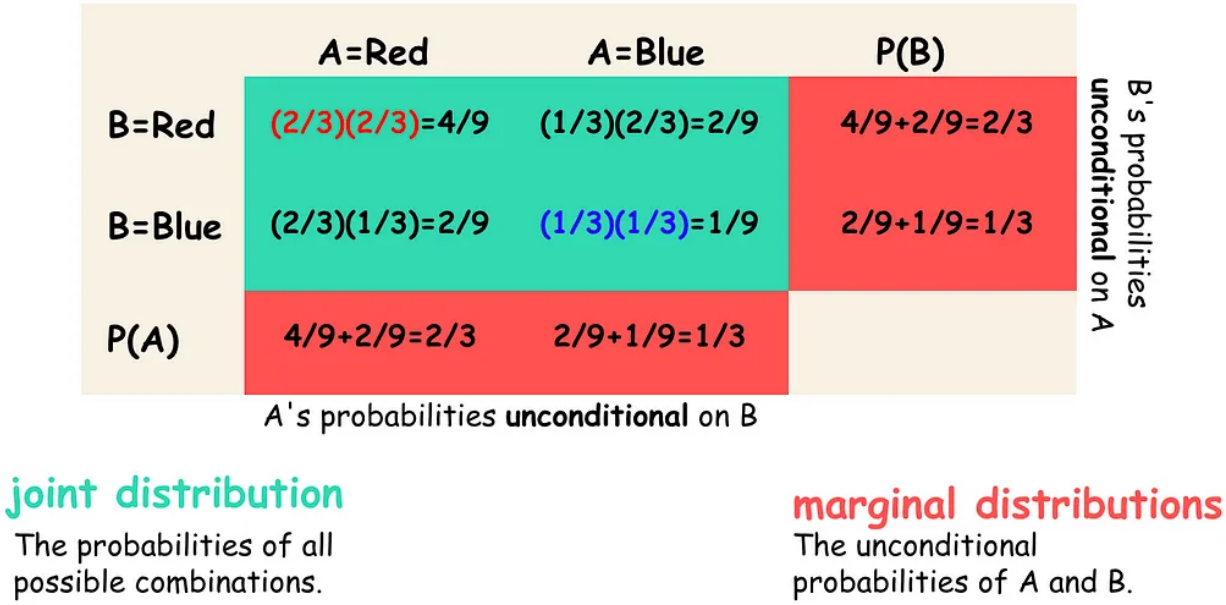


Figure 26 Table showing the various probabilities resulting from 2 independent draws (discrete random variables).

This shows the various probabilities depending on the outcome (the possible combinations from the 2 draws).We have drawing a red marble from both urns, drawing a blue marble from both urns, and drawing one red and one blue marble. The center cells (green) is the joint probability. The red cells are the marginal probabilities.

If we were using machine learning to classify cats in images containing either cats or dogs we would be comparing a vector of predicted labels to a vector of actual labels in a test set. Both of these vectors have distributions, which are either similar to each other or not. Thus a practical use of KL-divergence would be to use it as a *loss function* in a machine learning algorithm, where internal model parameters are adjusted until the delta between predicted and actual labelling is minimized.

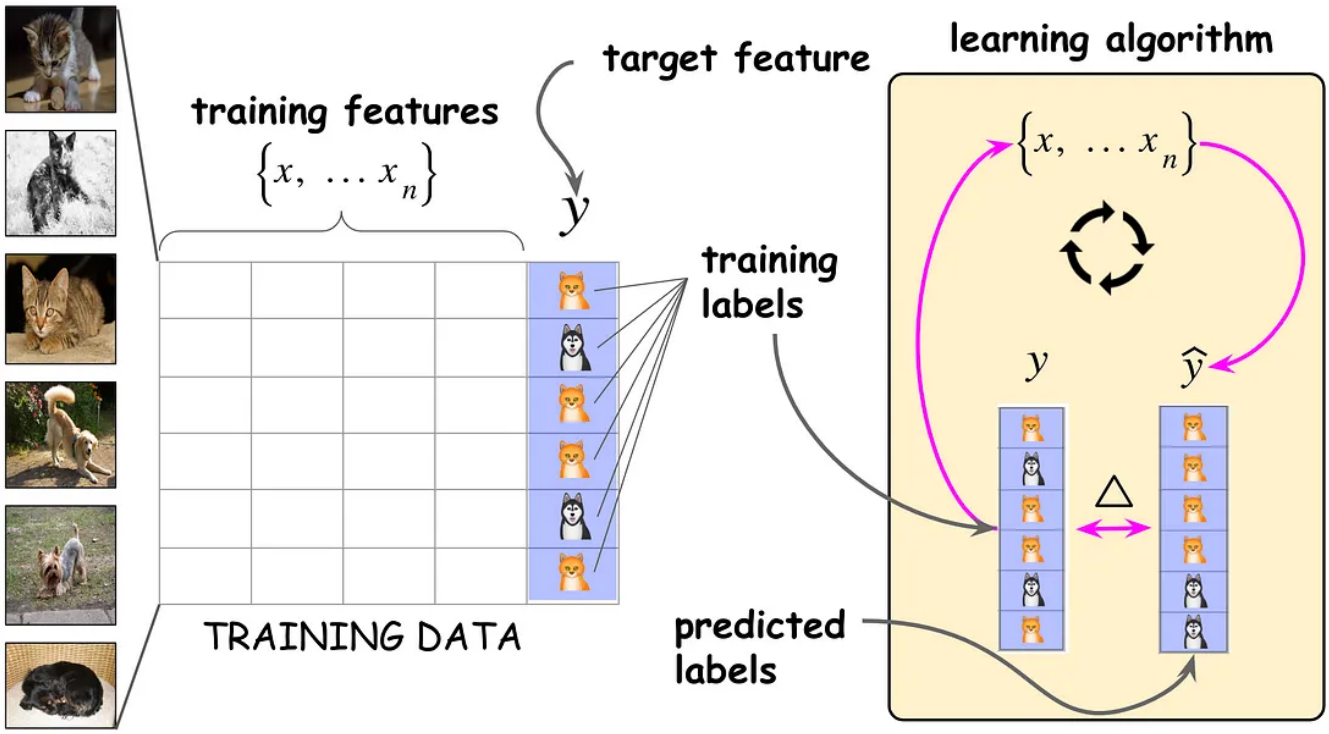


Figure 27 Supervised machine learning involves comparing a predicted vector to a test vector using a loss function as a measure of "distance."

To connect this to the concept of joint and marginal probabilities let's reframe the comparison between predicted and actual vectors in terms of the urn example we looked at previously.

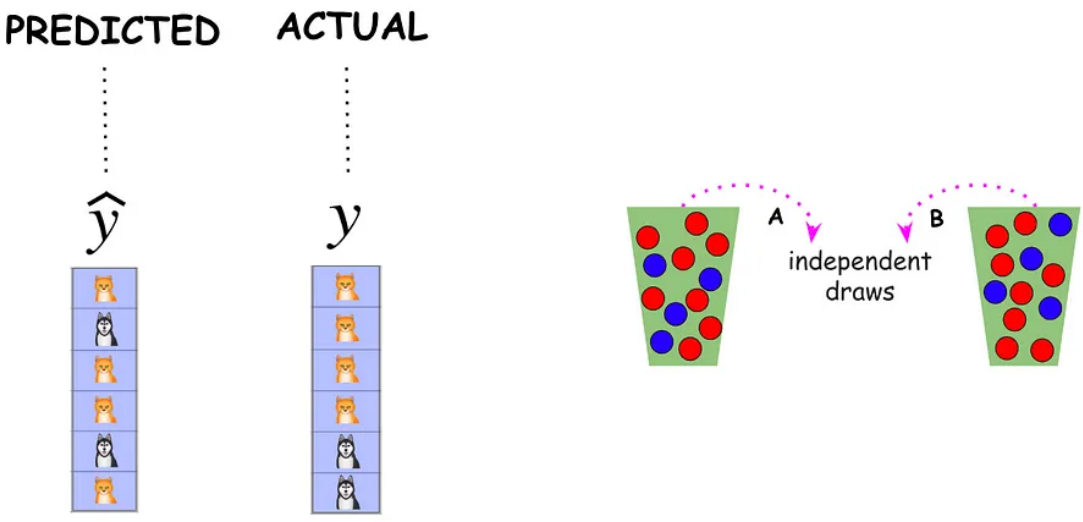


Figure 28 Looking at the similarity between vector comparison (in ML) and independent draws from 2 urns.

Each predicted label is compared to an actual label, which is analogous to a single draw from 2 urns. The learning algorithm is "observing" the draws from the "urns" as 1 of 4 possibilities; cat-cat, cat-dog, dog-dog, dog-cat. Just as with the urn example we can use the number of each of these possibilities to fill a probability table.

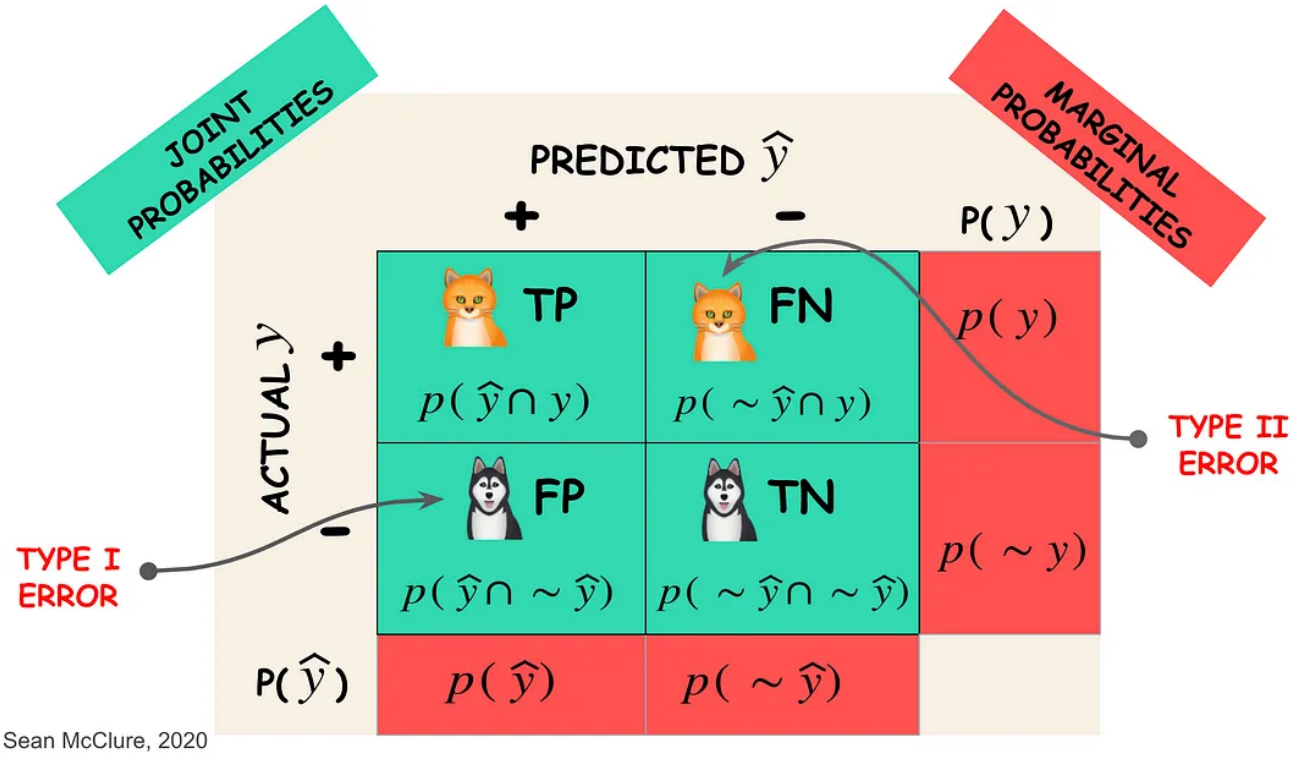


Figure 29 The confusion matrix is nothing more than a probability table counted as independent "draws" from labelled vectors.

Anyone familiar with machine learning will immediately recognize this as a *confusion matrix*. Confusion matrices are used to assess how well our classifier is performing. It is a matrix of the 4 possible outcomes between 2 vectors containing 2 labels. We can thus view the confusion matrix in terms of probability by realizing that the center rectangle are the joint probabilities, while the outer row/column are the marginal probabilities.

Now we can see where the 2 distributions that are being compared in MI come from:

$$I(X;Y) = D_{\mathrm{KL}}(P_{(X,Y)} \| P_X \otimes P_Y)$$

Let's wrap up by looking at an alternative expression for MI, compared to what we saw previously in Figure 5.

$$\text{Mutual Information: } I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Here MI is depicted in terms of the *additive and subtractive relationships* of various information measures associated with variables ***X*** and ***Y***. **H(X)** and **H(Y)** are the *marginal entropies*, while **H(X,Y)** is the joint entropy of X and Y. This tell us that mutual information is a joint entropy subtracted off marginal entropies. We can see this visually in the following Venn diagram:

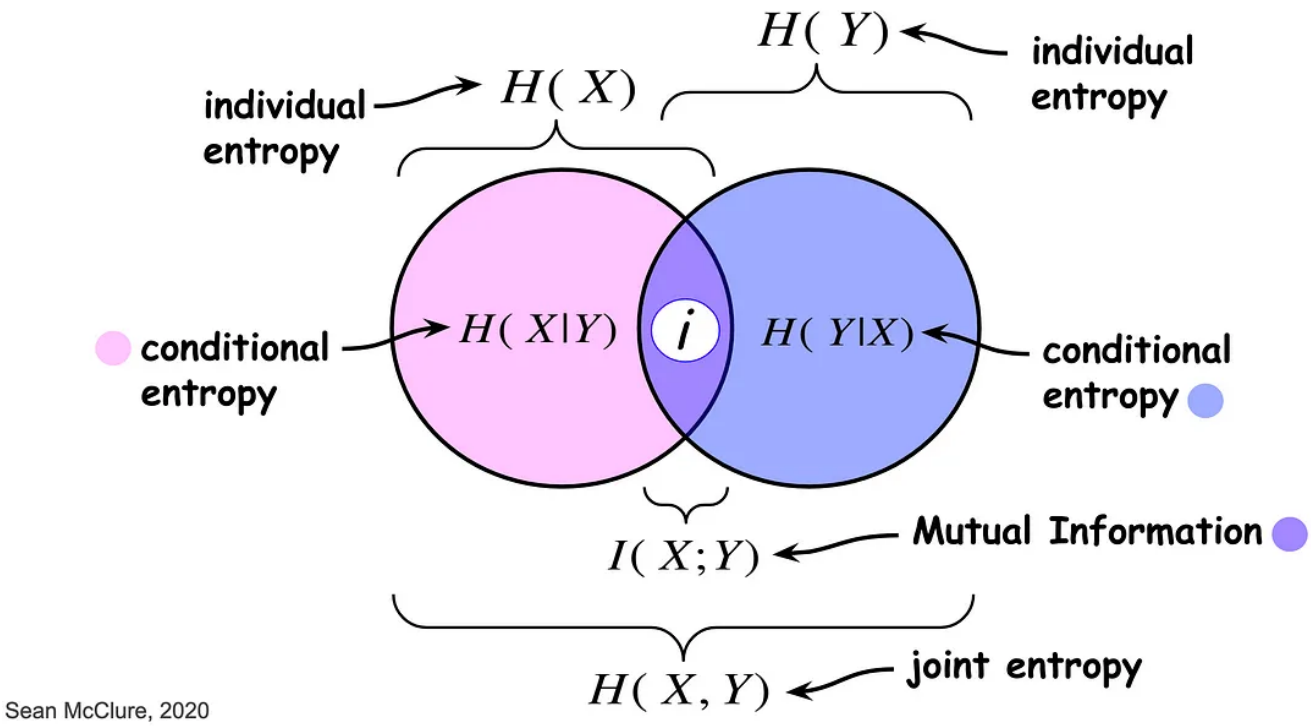


Figure 30 Venn diagram showing **Mutual Information** as the additive and subtractive relationships of information measures associated with correlated variables X and Y. (adapted from Wikipedia)

The area contained by *both* circles is the **joint entropy** *H(X,Y)*. Joint entropy is a measure of the *uncertainty associated with a set of variables*.

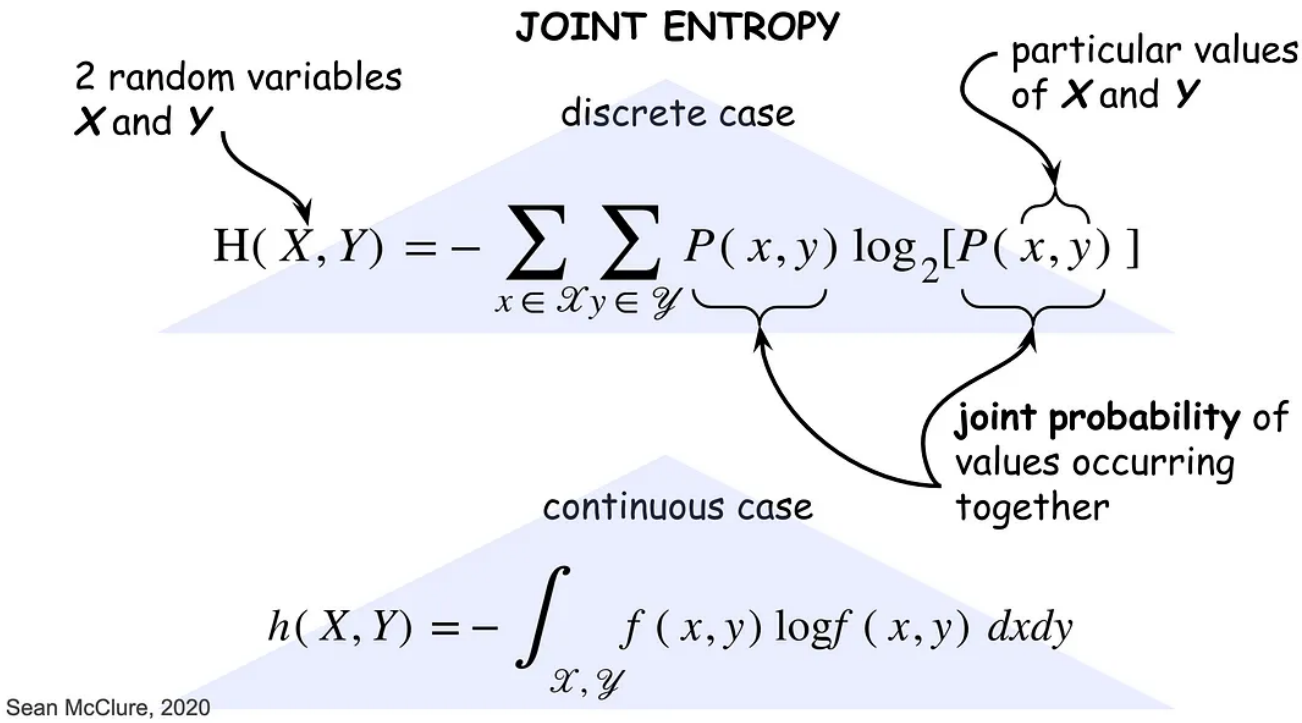


Figure 31 Joint Entropy

Recall the thermodynamic entropy we saw in Figure 18 (Gibbs entropy). Notice the similarity between that equation and the equations above. The approach to capture the uncertainty (entropy) of a system is by **sandwiching probabilities around a log**.

In the case of Gibbs entropy we took the probabilities of the individual microstates and summed them. By using *individual probabilities* and summing them we allowed for distinct contributions from different microstates. Here, we use individual joint probabilities and sum them, so as to allow for distinct contributions from different random variables.

In other words, if entropy is calculated by sandwiching individual probabilities around a log, then it makes sense that joint entropy would be the same thing but with joint probabilities. It thus conceptually makes sense that joint entropy is a measure of the uncertainty associated with a **set** of variable*s* since entropy is how we calculate uncertainty, and we are taking into account how 2 variables probabilistically act together.

> *Of course we can generalize this out to more than just 2 variables. Imagine more summation signs and more variables in the above equation. Still the same approach. We are merely measuring the uncertainty associated with many variables.*

What about the **individual entropies** (full circles on the left and right)? These are the entropy we already know from our discussion on information-theoretic entropy:

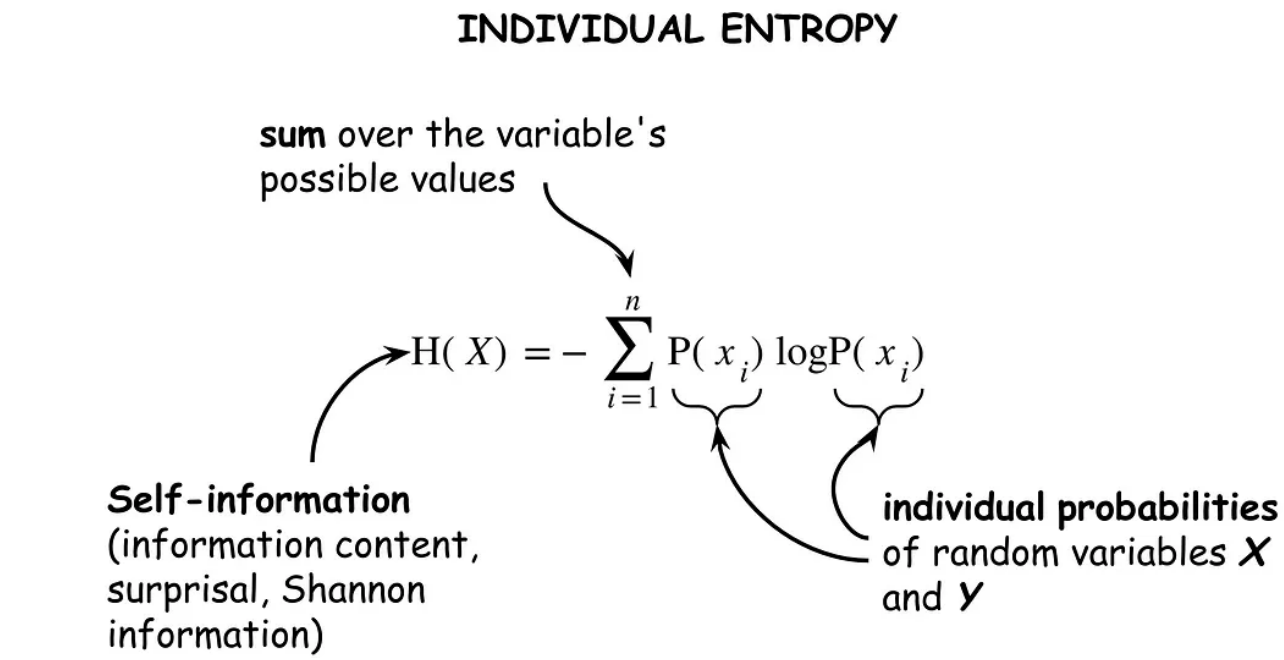


Figure 32 Individual entropies are the self-information contained within a variable.

We can think of these individual entropies as the expected value of the "self-information" contained within the variable. In other words, each variables has some level of "surprisal" (recall earlier) contained within it that we have yet to access. Upon learning the outcome (of the event) we will gain access to this information.

What about the **conditional entropies**, which are the parts of the circles that are *not* overlapping in Figure 27? Conditional entropy is defined as follows:

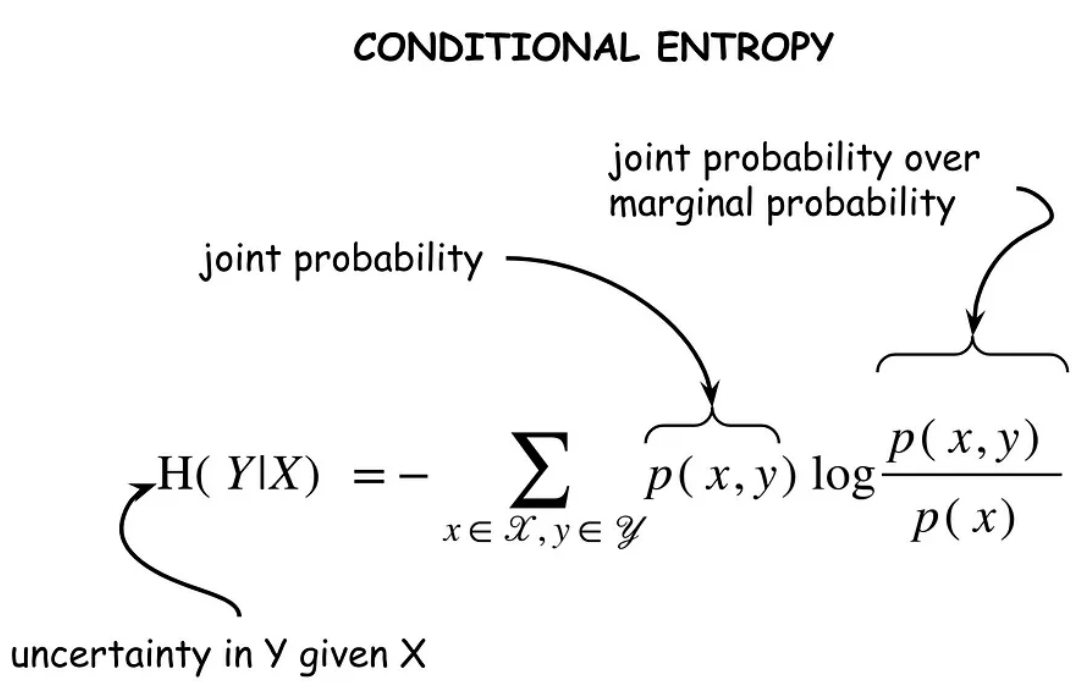


Figure 33 Conditional entropy of a random variable.

> *It should be noted that the Venn diagram is considered somewhat misleading. Refer here https://www.inference.org.uk/itprnn/book.pdf*

Thinking of mutual information in terms of adding and subtracting entropies gives us an additional view on how MI is calculated.

Summary

Causality is a central concept in our lives, underlying virtually every decision we make. The industry standard approach to uncovering causality is Pearson's correlation. There are a number of severe limitations baked into Pearson's, relating to its assumptions regarding how random variables might be related. Mutual Information digs much deeper into the notion of dependence by focusing directly on information itself. Further, by understanding information in terms of its "physicality" we can see how both thermodynamic and information-theoretic interpretations of entropy make any entropic measure of dependency, like MI, truly fundamental.

Further Reading

1. *Entropy* (Wikipedia)
2. Information Processing and Thermodynamic Entropy (Stanford.edu)
3. Ludwig Boltzmann (Wikipedia)
4. Mutual information (Scholarpedia)
5. *Information: The New Language of Science* by Hans Christian von Baeyer
6. How to create an unfair coin and prove it with math
7. You Can Load a Die, But You Can't Bias a Coin

Science | Machine Learning | Correlation | Probability | Statistics