We have a folk knowledge of what information is: if I don’t know something, and you tell me, you have provided me “information.” In the quantification of information created by Claude Shannon, when he was working at AT&T (see Ref. [1]), what you tell me—the information—is associated with the outcome of a random, or quasi-random, process. For example, if I flip a coin, the outcome of that toss is something you might want to know.1

Information theory is fundamentally about signals, not the meaning they carry. What we measure thus requires interpretation. Conversely, in its universality, information theory applies just as much to the written and spoken words of humans as to the electronic machines for which it was first developed. And it allows us to compare quite very distant worlds—no more, and no less, exciting than, say, comparing the real income of a English bricklayer in 1350 to one in 1780, the hours worked by a French housewife in 1810 and 1950, or the life expectancy of a hunter-gatherer of the Maasai to that of a child in a Manchester factory of 1840.

1 Mind

That we can quantify information is both intriguing and mysterious. Intriguing, because information is one of the fundamental features of our minds and our social worlds. History, psychology, economics, cognitive science and economics would all grind to a halt were their practitioners forbidden from using the concept of information at will. Mysterious, because information is a fundamentally epistemic property: it is about what one knows, and is, as such, relative to that observer in a way that one’s (real or nominal) salary, height, daily caloric intake, or place and date of birth are not.

Subject-relative facts, of course, abound—facts about trust, say, or allegiance, virtue, belief, love—and they make up a core part of the worlds we want to understand. What we learned in the twentieth century is that at least one such fact, the information one has, can be quantified. The

---

1 This notion of information is a subcase of the more general kinds of information we might want to know. Your Social Security number is (quasi)-randomly assigned; so (let us say) are a multitude of other contingent historical facts, such as the outcome of the Battle of Hastings and the codes found on your DNA. Conversely, there are other kinds of information imparted, often in mathematics classes, that are not well-described as the outcome of something that “could have gone the other way.” Information Theory as described here is not equipped to discuss the latter.
study of information in the social world is in its infancy, but not without some recent successes under its belt (this author claims to have seen a few).\footnote{The economists have tried to quantify another set of subject-relative facts, one’s desires, through utility theory, with somewhat less empirical success. Facts on the edge between reality and perception include those concerning inequality, and we have made a great deal of progress in figuring out both how to measure inequality, and what its implications are [2].}

2 Probabilities

In general, the more uncertain the process is, the more I have to learn about it, and the more information there is “in the process.” If I always say the same thing when I pick up the phone, there is less information when I call you. If I say one of a huge number of different things, there’s much more information. When a process varies a great deal, there’s a lot you’ll need to specify about it, and we can talk about, for example, how some facts about that process might be signaled by outside sources. Or we might imagine two processes, each varying a great deal, and ask how the one might be distinguished from another.

To avoid mathematical complexities, we’ll assume that the processes we are interested in have a discrete, finite number of outcomes. For example, the coin can either be heads or tails; a man talking can choose from only a finite number of possible words for the next part of the sentence; and so forth. The process can then be represented as a probability distribution over outcomes:

\[
\vec{p} = \{p_1, p_2, \ldots, p_n\},
\]  

\[i.e.,\ a\ list\ of\ probabilities\ for\ the\ different\ outcomes;\ this\ list\ is\ supposed\ to\ be\ exhaustive,\ but\ to\ describe\ mutually\ exclusive\ outcomes;\ a\ consequence\ is\ that\ the\ total\ of\ all\ the\ p_i\ is\ one.\]

A series of beautiful theorems, often attributed to the physicist Cox [3], relates the theory probability to that of optimal inductive inference. The rules of how to correctly manipulate probabilities, and how these connect to reliable thought itself, are described in the companion article, Bayesian Reasoning for Intelligent People

3 Entropy Defined

We want a function, call it \(H(\vec{p})\), that takes a list of probabilities and tells us the “information” in that process. Let’s require the function to obey four simple axioms. The first two are so simple that they almost seem trivial. The third is intuitive. The fourth is deep.

1. Continuity (if I only change the probabilities a little, the information of the process should change only a little).

2. Symmetry (if I reorder the list of probabilities I gave you, you should get the same answer).

3. Condition of Maximum Information: \(H(\vec{p})\) is at its maximum value when all the \(p_i\) are equal.

4. Additivity (discussed below).
Continuity is simple, but Symmetry is a bit loaded: it says that the information is the same whether the probability of events I feed it is (for example) \( \{ p_A = 0.2, p_B = 0.8 \} \) or \( \{ p_A = 0.8, p_B = 0.2 \} \). Of course, if the two probabilities represent the dispositions of a jury, option A is “guilty” and option B is “not guilty,” the prisoner will care a great deal! But \( H \) does not care. For this reason, information is often called a “syntactic” theory, concerned only with the properties of symbols in abstraction from their meanings.

The Maximum condition fits with our folk concept: if every possible outcome (from the list of possible outcomes) is equally likely, the process has the maximum information. Note! Do not get confused here. The process itself in this case is very random, and we often associate randomness with lack of information (e.g., if a student starts talking randomly in class, we say he is telling us nothing).

But information is concerned with specifying the outcome of a process; imagine having to describe the behavior of a student to a doctor. If the student’s behavior is very random, you have to have a longer conversation (“on Monday he was taking about cats, but then Tuesday he was on to bus schedules, and Wednesday it was about the filling in his teeth...”) as opposed to a less random, and thus lower information, process (“every day, he says hello and that’s it.”)

The Additivity requirement is profound. Say my students have two properties: gender (male or female) and hair color (red or yellow). I can talk about the entropy (or information) of the class, which means “if I pick a student at random, how uncertain am I about the gender/hair-color outcome”:

\[
H(\{p_{my}, p_{mr}, p_{fy}, p_{fr}\})
\]

We can also talk about the coarse-grained version: “if I pick a student at random, how uncertain am I about the gender”:

\[
H(\{p_m, p_f\})
\]

where note (just to check understanding) \( p_m \) is equal to \( p_{my} + p_{mr} \), \( p_f \) is equal to \( p_{fy} + p_{fr} \), and \( p_m + p_f \) is equal to one.

We want uncertainties to add in a sensible way. In particular, we want the uncertainty about the overall outcome to be the sum of the gender outcome and the weighted sum of the hair-color outcomes for each gender:

\[
H(\{p_{my}, p_{mr}, p_{fy}, p_{fr}\}) = H(\{p_m, p_f\}) + p_m H(\{p'_{my}, p'_{mr}\}) + p_f H(\{p'_{fy}, p'_{fr}\}),
\]

where \( p'_{my} \) is shorthand for the conditional probability: i.e., given that I know it’s a guy, what is the probability of him being red- or yellow-haired. This is simple to compute in the following way:

\[
p_{my} = p_m p'_{my},
\]

i.e., “the probability of being both male and yellow-haired is the probability of being male times the probability of, given that I am male, having yellow hair,” and so elementary algebra gives us

\[
p'_{my} = \frac{p_{my}}{p_m}.
\]

Given this additional thing we want the information measure to have, something magical happens: there is only (up to a constant factor) one (one!) possible mathematical form. It is

\[
H(\{p_1, p_2, \ldots, p_n\}) = - \sum_{i=1}^{n} p_i \log p_i
\]
where the choice of the base of the logarithm is the one freedom. If you chose the base-2 logarithm, 
$H$ has units of bits—the same bits as you buy by the Gigabyte in an iPod. Any other choice of 
function, beyond picking different bases, will violate at least one of the four conditions (and usually 
more than one).

Eq. 4 is deep, and it is worth perhaps thinking about in different ways. You can think of the 
$p_{my}$-like probabilities as being the leaves of a tree, and the $p_m$-like ones being the nodes higher up 
the tree. Then it’s like “the uncertainty as to which branch I’m on, and then the uncertainty about 
which leaf on the different possible branches I might have ended up on.”

That is a reason why these measures are particularly useful for the hierarchical structures we are 
using for the semantic analysis. Entropy breaks up into the “uncertainty about whether I am talking 
about semantics within the Intellect hierarchy vs. the Affections hierarchy” and then, within those 
two hierarchies, uncertainties about the more fine-grained distinctions.

4 Animal–Vegetable–Mineral

The inverse-square law of gravitation describes the motions of the planets just as well as the fall 
of a coffee cup. In a similar way, Eq. 7 turns out to be the answer to such a huge diversity of 
seemingly distinct questions that entire books can (and have) been written about it.

Here is one example. You may have, at some point, on a long car-ride, played the game 
“animal–vegetable–mineral”. Player one thinks of something, and by a series of yes-no questions, 
the other players attempt to guess. “Is it bigger than a breadbox?” No. “Does it have fur?” Yes. 
“Is it a mammal?” No. And so forth.

Some questions are better than others—for example, you usually try to eliminate the big general 
categories first (hence, the name of the game itself—is it an animal?) before going more specific. 
Asking on the first round “is it a carburetor?” is likely to waste time—unless, perhaps, you are 
playing the game on Car Talk.

If a game is lasting a very long time you might wonder what the optimal set of questions is to 
ask. “Could I have gotten the answer sooner, if I had skipped that useless question about the fur?” 
A moment’s reflection shows that, in fact, the optimal set of questions depends upon the player: if 
someone is biased towards material things, you’ll tend to focus on questions that split hairs among 
weights, sizes, and shapes. If someone is biased towards historical figures, you might split hairs 
about eras of birth. You can imagine writing down a script for your friend: “first ask if it’s a person; 
then if yes, ask if they were born before 1900, if no, ask if it’s a country...”; or for another friend: 
“first ask if it’s bigger than a breadbox; if yes, then...”; or for someone else, “first ask if it’s Elsa 
from Frozen. It almost always is.”

For every set of preferences (i.e., for every probability distribution over the things player one 
might choose) there is an optimal script (there may be more than one, but usually not). And for 
each person and optimal script for that person, the game will last five rounds, or ten rounds, or 
seven, or, twenty, depending on what they choose that time.

Remarkably, the number of questions you have to ask on average for a particular person and 
opimal script pair, is given by Eq. 7. In a very real sense, we are measuring information: the 
average number of yes-no questions we’ll need to ask to find out an answer (this is all true, except 
for one small wrinkle, that the answer is good to within one; see, e.g., an explicit version of the 
classic construction, by David Darmon).
5 Mutual Information

We are often interested in the extent which knowledge of one thing informs us about something else. People carrying umbrellas, for example, tells us something about the weather; it is not perfect but (informally) if you tell me something about the weather, you also reduce my uncertainty about umbrella-carrying. How to quantify this? For simplicity, we’ll assume you only get to see one person—your colleague, who walks in the door with or without his umbrella.

Consider $p_{w,i}$, the probability that the weather is of type $i$ ($i$ could label “rain,” “cloudy,” “sleet,” “sun,” and so forth). The uncertainty in the weather is then just $H(w)$, or

$$H(w) = - \sum_{i=1}^{n} p_{w,i} \log p_{w,i}$$

(8)

Now consider the conditional probability of weather of type $i$ given that you see someone carrying an umbrella—$p_{w,i|U}$. Generally, $p_{w,i|U}$ will be higher than $p_{w,i}$ when $i$ is labeling wet weather. You can compute the uncertainty of the weather given that people are carrying umbrellas as

$$H(w|U) = - \sum_{i=1}^{n} p_{w,i|U} \log p_{w,i|U},$$

(9)

or, in words, “the uncertainty about the weather, given that the person who walked in was carrying an umbrella” (you can read the $|$ as “given that” or “conditional on”). Similarly, for the reverse case, what the weather tends to be like when the person who walks in is not carrying an umbrella, $p_{w,i|\neg U}$, and the associated uncertainty

$$H(w|\neg U) = - \sum_{i=1}^{n} p_{w,i|\neg U} \log p_{w,i|\neg U}.$$  

(10)

The uncertainty drop is $H(w) - H(w|U)$ in one case, and $H(w) - H(w|\neg U)$ in the second case.

Note that the drop can be positive or negative—in some climates, seeing your colleague not carrying an umbrella will make you more uncertain about the weather. Consider, for example, an extremely rainy climate; it is either sunny, cloudy, or rainy, but most often rainy. You are generally quite certain about the weather before you see your colleague (it’s raining). So when he walks through the door without his umbrella, you think it’s less likely to be raining, and so you are more uncertain (the options sunny, cloudy, or rainy are now more evenly balanced).

Consider the average drop in uncertainty

$$I(w,u) = (H(w) - H(w|U))p_U + (H(w) - H(w|\neg U))p_{\neg U}$$

$$= H(w) - [H(w|U)p_U + H(w|\neg U)p_{\neg U}],$$

(11)

where $p_U$ is the probability of your colleague walking in with his umbrella, averaged over the weather (if this is confusing, think of it as the number of times that year he’s walked in with his umbrella), and similarly for $p_{\neg U}$, equal to $1 - p_U$.

We give Eq. 11 a special name: the mutual information between $w$ and $u$. It tells us how much less uncertain we are, on average, about $w$ given that we know $u$. 


We note (without proof) two lovely things. Mutual information is always positive—sometimes, the particular value of \( u \) will make you more uncertain (no umbrella in a rainy climate example, above), but “on average, information never hurts”. And, mutual information is always symmetric—the information you get about the weather, given knowledge of umbrella-carrying is equal to the information you get about umbrella-carrying, given the weather.

Mutual information will return soon, when we discuss the Jensen-Shannon Distance. To draw you back to the animal–vegetable–mineral game, mutual information is how many questions are effectively answered for you by learning the signal.

6 Surprisal and Relative Entropy

Let’s define a funny concept: surprisal. The surprisal of an event \( i \) is

\[
S(i) = \log \left( \frac{1}{p_i} \right).
\]

(12)

For example, if \( p_i \) is very small, then \( 1/p_i \) is very large. Then you take the logarithm, and it won’t be as large, but it will still be largeish. So you are surprised. If \( p_i \) is close to unity, and the event \( I \) is thus very likely, then \( 1/p_i \) is also close to unity, and the logarithm of one is zero, so the surprisal is low. Note that \( \log \left( \frac{1}{p_i} \right) \) can just as well be written as \( -\log p_i \), by the properties of logarithms.

From Eq. 7, we can see that the entropy of a process is “the average surprisal,” i.e.,

\[
\sum_{i=1}^{n} p_i S(i) = -\sum_{i=1}^{n} p_i \log p_i,
\]

(13)

the weighted sum of the surprisals is just Eq. 7.

That is cute, but all we did was define some somewhat silly concept, \( S(i) \). However, since we perceive things in terms of relative risk, rather than absolute risk—\( x \) is five times more likely than \( y \), for example—the claim might be made that surprisal is a somewhat psychologically plausible. When something doubles in likelihood, the surprisal increases by one unit, regardless of the starting probability. If \( p_x \) is half as likely as \( p_y \), and \( p_y \) is half as likely as \( p_z \), then the surprisal of \( x \) is one unit more than the surprisal of \( y \), and the surprisal of \( y \) is one unit more than the surprisal of \( z \).

Now ask the following question: say the true probability of an event \( i \) is \( p_i \), but I have a (false) belief that it is \( p'_i \). The true surprisal is \( \log \left( \frac{1}{p_i} \right) \), but I think the surprisal is \( \log \left( \frac{1}{p'_i} \right) \). Call this false surprisal \( S'(i) \).

Then, we can ask “what is the difference in surprisal.”

\[
D(i) = S'(i) - S(i)
\]

(14)

Imagine this as two people watching a soccer match. The star scorer for Chelsea is on a great training regimen (say), and so I know that he’s likely to score (\( p_{\text{score}} \) close to unity). When he does, I’m not surprised. You don’t know this, so you’re totally shocked when Chelsea scores, because you have set \( p'_{\text{score}} \) much closer to zero.

\[3\] This logarithmic scale extends to other aspects of human perception. For example, the just-barely-perceptible difference for many of our senses is a multiplicative factor of the current perception: if you are holding a penny, and someone adds a paperclip, you will notice, but if you are holding a fifteen kilo weight, you will not.
You see what’s coming: now ask, what is the average difference in surprisal—\( i.e. \), the sum over all \( D(i) \), for each value of \( i \), weighted by the probability of seeing event \( i \).

\[
D(\{p\}, \{p'\}) = \sum_{i=1}^{n} p_i D(i) = \sum_{i=1}^{n} p_i \log \left( \frac{p_i}{p'_i} \right),
\]

(15)

where you should verify this just to test your logarithm manipulation abilities. Note that, although it is not obvious, \( D \) is always greater than or equal to zero, with \( D \) equal to zero if and only if \( p_i \) is equal to \( p'_i \) for all \( i \). This is a consequence of Jensen’s inequality, discussed in many places including Ref. [4] (pg. 35).

The function \( D \), that takes two probability distributions, is sometimes called the Kullback-Leibler divergence, or distance (“Leibler” is pronounced as if it were written “Liebler”). It is the measure of distance between probability distributions. There are many ways to get this, we just did a cute argument about the difference in surprisal between someone who has true vs. false beliefs.

Note that \( D \) is not symmetric! If I believe Chelsea has a good coach, and they don’t, is it different from me believing that Chelsea has a bad coach, and they do. We’ll return to this in the next section.

7 From one text to another

One use of \( D \), Eq. 15, is to measure the distance between texts. The tricky thing is that it is not symmetric, so that the distance from \( A \) to \( B \) is not the distance from \( B \) to \( A \). Here is the full argument.

Take the semantic vector for text \( A \), a vector \( \{a_1, a_2, \ldots, a_i, \ldots, a_k\} \), where \( a_i \) is the number of words with semantic category \( i \) found in text \( A \) and \( k \) is the total number of categories, remembering that distributions like these must describe mutually exclusive outcomes, and so containing categories (hyponymic relations) are forbidden.

Normalize this vector, to get a list of probabilities, \( p_{a,i} \). Formally, that says the following: “the text is just a bag of concepts; I reach my hand in and pull out a concept at random.” We have

\[
p_{a,i} = \frac{a_i}{\sum_{i=1}^{n} a_i}.
\]

(16)

We can ask “what is the semantic information of this text.” That means, explicitly,

\[
H_a = -\sum_{i=1}^{n} p_{a,i} \log p_{a,i}.
\]

(17)

and this measure will behave well under coarse-graining (so that if we had a finer-grained semantic, we would get a higher entropy, and it would decompose into the sum of the coarse-grained semantic information plus the weighted sum of the finer-grained specifications).

We can then go on to ask “what is the semantic distance between two texts,” using the Kullback-Leibler divergence defined in the previous section. This (just to be tedious) would be

\[
D(a, b) = \sum_{i=1}^{n} p_{a,i} \log \frac{p_{a,i}}{p_{b,i}},
\]

(18)
where we write $D(a, b)$ as a shortcut for $D(\{p_a\}, \{p_b\})$. Again (to be explicit), this is the average difference in surprisal if you are expecting a text with property $i$ drawn with probabilities $p_{b,i}$, but you are actually looking at a text with property $i$ drawn from the distribution $p_{a,i}$.

Again, $D(a, b)$ is not equal to $D(b, a)$, which makes it not a normal distance. In some cases, even, $D(a, b)$ can be finite, while $D(b, a)$ can be infinite. This turns out to be a nice way to see how $D$ works.

In particular, imagine two books: one is an encyclopedia, E, which has non-zero probability for every semantic category. The other is a book on Molecular Biology, MB. The distance $D(E, MB)$ is infinite: there are some $i$ for which $p_{a,i}$ is non-zero while $p_{b,i}$ is zero (e.g., when $i$ refers to a semantic category that has nothing to do with Molecular Biology). Meanwhile, the distance $D(MB, E)$ is not zero, but could be quite small, and is always finite (note that “0 log 0” is zero, as can be proved by l’Hôpital’s rule from calculus).

Put another way, I can reconstruct the Molecular Biology textbook from the Encyclopedia, but not vice-versa. For this reason, $D(a, b)$ is sometimes called “the distance from $b$ to $a$” (but other people call it the distance from $a$ to $b$, so be careful!)

Just to give another example, say there are two people, me and Einstein, and three states: \{sleeping, eating, revolutionizing physics\}. For me, the probabilities are \{0.2,0.8,0.0\}, while for Einstein, they are \{0.19,0.8,0.01\}. If you are expecting Einstein, and get me, you are usually not more or less surprised by my behavior than someone expecting me: Einstein and I both sleep and eat a lot. If you are expecting me, and get Einstein, there is a small chance (one percent) that you will see the guy revolutionize physics, which will be infinitely surprising if you were expecting me.

8 Distances between texts

In the previous section, we introduced the Kullback-Leibler divergence, which allowed us to quantify “given that I’m expecting X, how surprised am I when I see Y instead.” This divergence is not a distance in the normal sense: for example, it is not symmetric. It turns out that there is a very closely related measure, the Jensen-Shannon Distance (JSD) that has much nicer properties and, at the same time, has a direct epistemic interpretation.

Consider two distributions, as before, $p_a$ and $p_b$. Define the mixture, $p_m$ as

$$p_{m,i} = \frac{1}{2}(p_{a,i} + p_{b,i})$$  \hspace{1cm} (19)

The Jensen-Shannon distance is then defined as

$$JSD(a, b) = \frac{1}{2} \left( D(a, m) + D(b, m) \right),$$  \hspace{1cm} (20)

or, in words, “the average of the divergences from $m$ to $a$ and from $m$ to $b$”. It is always positive, and, if you work in log base two, always between zero and one. If the texts are identical, then it is zero; if the texts are “optimally distinguishable”, then it is one.

Mathematically, the formulation of Eq. 20 is equivalent to

$$JSD(a, b) = H(m) - \frac{1}{2}(H(a) + H(b)),$$  \hspace{1cm} (21)

8
or, in words, the entropy of the mixture $m$, minus the average of the entropies of $a$ and $b$.

If we take $a$ and $b$, as before, to be distributions over words, $H(a)$ is “how uncertain you are about the next word, given that it is being drawn from distribution $a$”. Meanwhile, $H(m)$ is (informally) “how uncertain you are about the next word, given that you don’t know if it is being drawn from $a$ vs. $b$”. Imagine, for example, that distribution $a$ is almost certain to give you the word “cat”, while distribution $b$ is almost certain to give you the word “dog”. Distribution $m$ will give you “cat” with probability one-half, and “dog” with probability one-half; here $H(m)$ is much larger than both $H(a)$ and $H(b)$. $H(m)$ can be larger or smaller than $H(a)$, but it is always greater than (or equal to) the average of $H(a)$ and $H(b)$, which is why the JSD is always greater than (or equal to) zero.

We can interpret JSD, then, in the following way: say you don’t know which distribution you’re using (i.e., you’re drawing from the mixture $m$). How much less uncertain (on average) would you be if you knew whether you were drawing from $a$ vs. $b$. “On average” here means “average the drop in uncertainty for going from $m$ to $a$, with the drop in uncertainty in going from $m$ to $b$,”

$$JSD(a, b) = \frac{1}{2} [(H(m) - H(a)) + (H(m) - H(b))],$$

and by inspection, Eq. 22 and Eq. 21 are identical.

Finally, we can turn this into an identification problem. Say your colleague chooses (with 50-50 probability) either distribution $a$ or $b$, but you do not know which; call the “indicator variable”, i.e., the choice your colleague has made, $z$, and say that she’s using $a$ when $z$ is equal to one, and $b$ when $z$ is equal to two.

Now that you’ve defined $z$, you can write $H(a)$ as $H(m|1)$, and $H(b)$ as $H(m|2)$. Using Eq. 21, we can write

$$JSD(a, b) = H(m) - H(m|z),$$

which gives us the final result: the Jensen-Shannon Distance between two texts is the mutual information between the text identity and a sample from the texts. More informally, if you don’t know which kind of text you’re reading, the Jensen-Shannon Distance tells you how much the uncertainty drops—how much information you get—when you hear a single word. This is the quantity used in both Ref. [5] and [6] to quantify the epistemic structure of discourse.

References