Entropy (information theory) | Network Science | Information Theory | Intuitive Explanations | Entropy (physic

What is an intuitive explanation of the concept of entropy in information theory?

Would greatly welcome more detail too, and not just the most general idea.

15 ANSWERS

Andrea Klein, B.S. in Physics from Stanford

Updated Jan 21, 2014 · Upvoted by Mario Alemi, studied at Physics and Tejal Bhamre, PhD candidate in theoretical physics at Princeton

The next step to developing intuition for information entropy is probably to confront the corresponding *equation* -

$$\mathcal{H}(S) = \sum_i p_i \log_2 \frac{1}{p_i}$$

- where it comes from, why it "works," and what it means. You should be in good shape if you understand the following:

1. **What historical context motivated the invention of the "information" concept (and an associated mathematical theory)?**

Here's the slightly (okay, considerably) simplified backstory. In the 1940's, a fellow named Claude Shannon was thinking about electronic communication: telegraphs, TV signals, and so forth. He and his colleagues were quite interested in how best to encode a digital message. On the one hand, it would save resources to make the encodings as concise as possible. On the other hand, real communication isn't perfect, and if even a single bit of a such an encoding somehow got flipped, it might become impossible to reconstruct the original message. You could protect against this by making the encoding very redundant, but then it would be long and expensive to send.

There were some associated problems with understanding the capacity of a system to carry messages, deal with noise, and so on. But it's very hard to reason about this stuff, or prove anything useful, without some kind of abstract model of what's going on. So, like any good scientist, Shannon invented one! He imagined an arbitrary source S capable of transmitting one of a fixed set of messages, where message i has a probability p_i of being transmitted. The next step to developing this theory was to quantify the amount of "information" being transmitted by S. (This eventually made the theory powerful enough to start proving useful formal limits, such as how concisely you can compress a message.)

2. **Why is the information entropy equation a good representation of this concept?**

To see why Shannon's equation works, let's try coming up with a better one - or at least a different one - on our own. So far, "information" is a bit of a vague idea, but we have some ideas for how we want to use it. So, one way to proceed is to think about all the properties we'd like an information equation to have, and then just guess a function that satisfies them. After all, perhaps there's no right answer - we just need something more precise to work with.

Here's an example of what I mean: suppose we know what the source S is, i.e. we know the values of all the transmission probabilities. If some message has probability 1 of being transmitted, then we learn absolutely nothing when we observe that message - we already knew what it was going to be! - so it seems reasonable that the information associated with this S should be 0. Sure enough, if one probability is 1 and the rest are 0, plugging into Shannon's H(S) gives you 0.

Consider another edge case: if all of the messages are equally probable, we expect $\mathcal{H}(S)$ to be maximized. This is, of course, the opposite of the case above. It's a bit less obvious, but if you think about it, the more lopsided the probability distribution is, the less we "learn" by observing a message, and vice versa. (This is what Debidatta-Dwibedi meant in his answer when he associated surprising events with high amounts of information.) Sure enough, Shannon's formula is maximized when all the probabilities are equal. This is easiest to see if you assume there are two events, one with probability p and one with probability $1 - p$, and plot $\mathcal{H}(S)$ - you'll see that it's maximized at $p = 1 - p = .5$..

There are, of course, a ton of other properties we might demand of this equation. The Wikipedia page for information entropy lists a whole bunch more. In fact, Shannon himself had a list, and he proved that only an equation of his form could satisfy all of them! So it seems there was a "right" answer after all - at least, up to the choice of leading constant and logarithmic base. The choice of the log base is in principle somewhat arbitrary, but it does mean something: base n corresponds to a source that uses n symbols to encode messages. So, the version of the equation pictured above (base 2) makes sense when the source is speaking in binary and you want to say that one bit carries one unit of information.

Shannon may have been aware that sometime earlier, physicists ran into a similar challenge trying to pin down the elusive concept of "disorder" in thermodynamic systems. The properties they wanted for their quantity are essentially identical to those needed for Shannon's, so they give rise to an equation of the same form, known in statistical mechanics as an "entropy" equation - hence the term "information entropy."

3. **What new insights can we develop from studying the equation, now that we have it?**

There are certainly plenty, but here's one to get you started: I claim that $\mathcal{H}(S)$ measures the average number of bits needed by S to transmit a single message, if it is encoding them as "smartly" as possible (more on that in a moment). To see this, notice that $\mathcal{H}(S)$ has the form of an expectation value, or weighted sum - there's a probability p_1 that it will need to encode message 1, p_2 of message 2, and so on. So if $\log \frac{1}{p_i}$ is the number of bits needed to encode message i, then my claim makes sense.

Now let's think about how S should do the encoding. Suppose S always transmits a chunk of three of the same letter, i.e. 'AAA' or 'BBB' or ... or 'ZZZ.' Suppose also that there's a 99% probability it will

pick AAA. Now let's say that once an hour, S picks a string (according to its probability distribution) and texts you to tell you the result. (Here, we're assuming no noise / information loss.) S doesn't have to text the string, i.e. message, literally - it could send you any encoding of it, as long as you're able to interpret the result. For example, upon picking AAA, S could text you, "Hey guess what - I picked AAA!" But this is, of course, gratuitous, and very wasteful of your data plan, especially considering that S was almost certain to pick that string anyway.

In general, the more probable a message is, the more sense it makes to pick a super short encoding for that message. In the example above, why not just have S text you "1" if it picks AAA? It's fine to have longer codes for the other possibilities, since they occur so much less often. This way, the average amount of bits used by S to communicate its messages (and the corresponding impact on your phone bill) will be minimized.

Returning to my claim above, it seems that any sensible encoding scheme will use fewer bits for messages with high probability and more bits for messages with low probability. We can see that $\log \frac{1}{p_i}$ gets smaller as the probability gets bigger, and vice versa. In fact, this turns out to be the optimal number of bits to use for a message with transmission probability p_i, and you can also think of it as the amount of information *inherent* to that message. So, if S is encoding its messages optimally, $\mathcal{H}(S)$ is precisely the expected number of bits it will need to encode a given message, or, equivalently, the average number of bits it will use per message over many messages.

For more information on information entropy, I highly recommend Shannon's classic paper on the subject, **"A Mathematical Theory of Communication" (1948)**, as well as the relevant Wikipedia pages.

Is entropy the best explanation of causality?
1,044 Views

What is entropy in information theory? A simple explanation will be very helpful
1,107 Views

What is a good explanation of Information Theory?
828 Views

What is the physical significance of entropy in Information theory?
1,195 Views

What is the best explanation of 'Entropy' so as to get a 'feel' of the concept?
323 Views

Is there an intuitive explanation of why Mutual Information is a symmetric quantity?
581 Views

Can anyone give me an intuitive understanding on entropy?
880 Views

What is the relationship between entropy and information?
2,798 Views

What's an intuitive way to understand entropy?
171,712 Views

What is an intuitive explanation for why change in entropy (dS) is equal to dq/T?
4,172 Views

What is the intuitive explanation of pi?
943 Views

OTHER ANSWERS

Ed Caruthers, PhD and post-doc work in Physics at UT Austin. Published papers in electronic properties of metals, semiconductors, p...

Updated Oct 27, 2014 · Upvoted by Shankar Iyer, Ph.D. in Condensed Matter Theory from Caltech

As others have said, Shannon's definition of *Information* is about prediction vs. uniqueness. Very

predictable answers score low on Information content. Unpredictable answers score high.

Let's use Quora to understand Shannon's definition. Suppose there's a question about politics, and you see that some familiar Quoran has written an answer. Try to predict the answer before you read it. If you find out that you got everything in the answer right, even its length, then you didn't learn anything new. Even if the answer is correct, even if it is well written, even if everyone who sees it understands it perfectly, it didn't contain any information that wasn't already present in that familiar Quoran's previous writings. The Shannon *Information* score would be near zero.

As a second example, if there have already been 5-10 answers to a question, and they all pretty much say the same thing, a new answer that says something different - even if it's wrong! - would get a higher *Information* score than another statement of something we already know.

I think most people find Shannon's measure of *Information* counter-intuitive, at least at first. In ordinary conversation, when we talk about information, we also talk about *Meaning, Correctness,* or at least *Comprehensibility*. And Shannon's equating of *Information* with Entropy gives the highest scores to the most random strings of symbols. So

> aadCibkk dJz15;0)]\ 2 BFKL = qaxguilpj.>9777@#$*&^

would get a high score for Information. But it carries no meaning. And it's not even wrong.

The key is to separate Shannon's definition of Information from Meaning, Correctness, or Comprehensibility. These are separate concepts. And Shannon's concept is useful. I won't repeat what the other answers have already said about that ;-)

Debidatta Dwibedi, works at Carnegie Mellon University

Written Nov 28, 2012 · Upvoted by Mario Alemi, studied at Physics and Ed Caruthers, PhD and post-doc work in Physics at UT Austin. Published pa...

Intuitively, entropy can be thought of a measure of how interesting an event is.

For example, if you tried unlocking the door of your room with your key and (voila!) it opened. Would you consider this an "interesting" event? Would you talk about it with your friends? Even if you did talk about this, how much could you possibly talk about it! However, if you fail to unlock the door, you will have something interesting to share.

So, entropy is the average of how interesting an event(unlocking your door) is. The more interesting outcome(that is your failing to open the door) is rarer. You have a longer story which means more information. But the more probable outcome isn't that interesting and there is not much information that you can possibly give out.

Entropy tells you on an average how much information is contained in an event.

Peter Gribble, plays too much Kerbal Space Program
Written Jul 21, 2014

Lots of great answers here, which is good, because it saves me having to go remind myself of all the details about Shannon and put them here. So I'm just going to put the description that was most useful to me in terms of understanding information theory:

Entropy is the number of bits of **information** in a message, which is different from the number of bits of **data**.

Let's say you have an information source which is an unending sine wave, that's been transmitting forever and will never stop transmitting. It's putting out a vast amount of data, depending how you choose to read it - if you measure its level every microsecond you'll get huge amounts of numbers back.

But none of it will tell you anything, because the source isn't actually changing at all. If you were told to build a receiver to listen to this source, you wouldn't even have to connect it, because you know it will never stop, so you know exactly what it's going to do. It's not sending you anything you can't build into your receiver's design - you can tell it to display a sine wave at the right frequency and phase, and it would always be correct.

Now let's say you're connecting to another source, which you know in advance will either be an unending sine wave or an empty box, with a 50% chance of each. Now, you can't be sure what you're going to receive - you have to listen and determine which of the two answers you're going to get. You know for sure, though, that it's going to be one of those two answers, so you can assign one of them as a 0, and one of them as a 1. Therefore, this source is sending you 1 bit of information.

Our next source is different. It's going to send you one letter of the alphabet, once, at a specific time. The rest of the time, you know it will send you nothing, so that has no information for you, but you don't know what letter it's going to send. But you do know that it has an equal probability of sending each letter, so you assign each a number from 0-25, and you represent that in binary:
a=00000
b=00001
...
Z=11001

It took us five bits to represent all that, so you could say that the message you've received has five bits of information. There are a few numbers that haven't been used, and in some situations you might be able to be clever with fractional entropy and bring that down to four-and-a-bit bits, but in this case there isn't anything you can do, because the probabilities are all the same.

When the probabilities aren't the same, it starts getting more complicated, but that's the essence of entropy in information theory: it's the smallest number of bits that you could use to represent a signal you got sent, if you knew everything about the source in advance and could build your

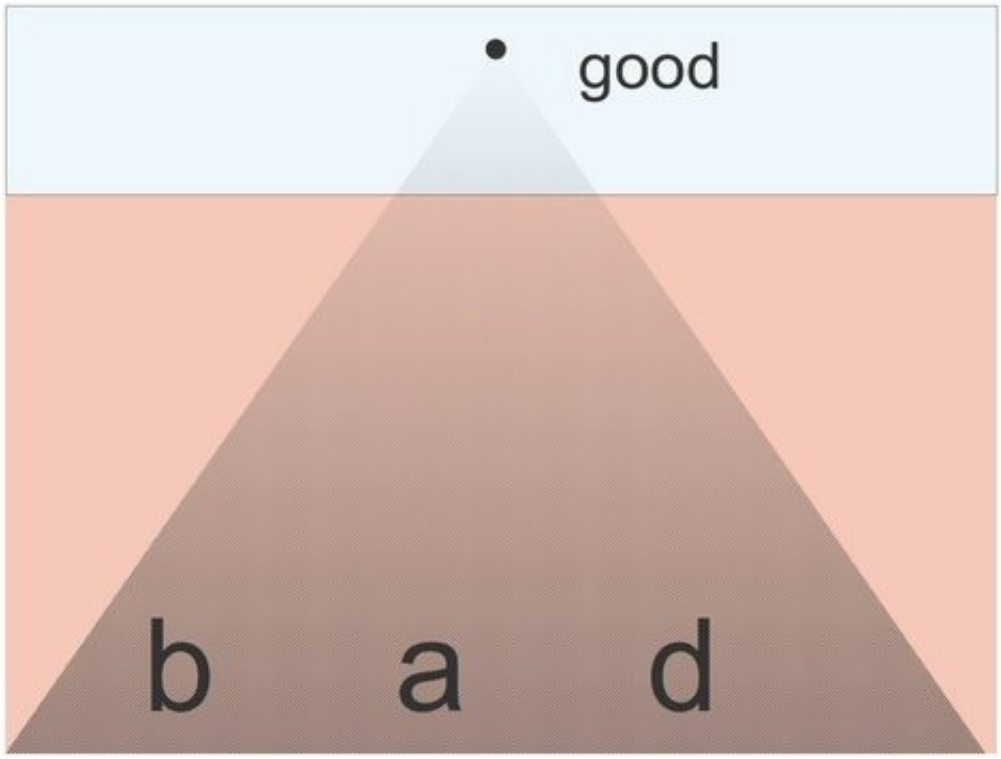
transmission scheme accordingly.

I found that thinking about it in these terms helped me get to grips with information entropy - I hope it helps you, too.

Lors Soren, http://isomorphismes.tumblr.com/tagged/entropy

Written Oct 9, 2013

Think about the Rubik's Cube: 43,252,003,274,489,856,000 possible answers and only one is correct.



The maximum distance is 20—that's the height of my triangle. 99.75% of the probability mass is contained 16–20 steps away, so technically the triangle's walls should curve inward. (So pretend the horizontal scale is logarithmic.) And there are obviously only a few 1-away and 3-away configurations: those are the ones that would lose me as I start with a solved cube (0-away) and quickly descend into chaos.

In this view entropy is the reason I can't just try random stuff and solve the cube. The odds are like a quintillion to one.

You can relate that to a fart dissipating in a room because there are combinatorially more combo-states where the fart particles dissipate away from each other than where they stay together.

Diwakar Tundlam, Thought Engineer; Software Experimenter.

Written Aug 31, 2015

Diwakar Tundlam's answer to What are the best examples of Entropy?

The concept of thermodynamic entropy is a measure of randomness or uncertainty in a collection of items. If a set of clothes is neatly folded and stashed in the closet, one can find exactly what they want with little or no searching. Entropy is low for such an ensemble.

For a haphazard collection, is harder to find an exact specific item. Entropy is high.

Now consider you have to give instructions to someone to find an item from either of these two piles. You don't need to give to many instructions for the orderly pile, but need to give more specific and detailed instructions for the disorderly pile. For example, in the orderly pile, you can just say 'fifth item from the top' and get what you want, since you already know the sorted order. To get the same item from the disorderly pile, you have to give a detailed description of the item.

Higher entropy requires more information to specify than if it were lower. Conversely, higher entropy collections also store more information.

The log scale chosen to represent entropy is arbitrary, but convenient to represent wide variation in the scale of entropy. It maps nicely with number of bits needed to represent that information and is hence preferred to a linear scale.

Keith Allpress

Updated Sep 1, 2015

I think that most scientists would consider Floridi to be fairly uncontroversial. He is careful to construct well considered definitions of information and data, and how they are used.

Floridi simply clarifies all the various potentially confusing jargon and misdirection in the science by focussing ruthlessly on the concept of ***data deficit*** involved in a communicative exchange between an informer and an informee. The quantity of information produced by a device corresponds to the amount of ***data deficit*** erased. It is a function of the average informativeness of the (potentially unlimited) string of symbols produced by the device. And so on. You can find it here.

Page on philosophyofinformation.net

I could rewrite it as a student edition if you like. If I can find the time.

Chris Honsinger, Entrepreneur

Written Apr 16, 2013

Entropy is a measure of how many bits it would take to represent something using a base 2 number system in their most efficient ways.

In it's most elemental form, no model is used to represent the information outside the simple histogram of the probability of occurrence of each symbol of the thing you are representing. This is called 0th order entropy.

Higher "orders" of entropy reflect what may be known apriori. Text documents for example, repeat the symbol of "white" very frequently. In this case, the histogram may include an additional symbol for *repeats* of the "white" symbol. This is often called a 1st Order Entropy. Generally, when a-priori information is known, it can be used to reduce the calculated entropy when compared to 0th order entropy.

The entropy measure is very closely mirrors the number of bits after a data compression algorithm is applied. Entropy measurements and the associated theory of information are the cornerstone of data compression.

Charles H Martin, PhD & NSF Fellow, Theoretical Chemical Physics, University of Chicago

Updated Nov 8, 2013

Entropy is simply a measure of the number of ways of expressing an idea or some information:

A simple example: say we have a 10000 balls and 100,000 boxes. How many ways can we toss the balls in the boxes?
The entropy measures the number of ways of doing this.

Now say the balls are identical, but the boxes are different sizes.
The larger boxes are more likely to catch a ball than the smaller ones.
The entropy tells us something about the most likely way all the balls will end up

Mehul Singh, Data hustler, powerpoint pimp

Written Feb 25, 2014

Originally Answered: What is entropy in information theory?

I am not sure how apt this is but one analogy i can think of is more the variable transforms, the more variance is killed and hence more noise is introduced to the system.

Share

Dima Korolev, Principal Maintainer at C5T/Current (2014-present)

Written Nov 8, 2013

+1 to Andrea Klein's answer.

Would like to add the word "uncertainty".

Both philosophically and mathematically entropy is the measure of how uncertain someone is about some events within some domain -- locally or globally.

On a conceptual level, the more uncertain about some event we are, the more information one needs to carry over to deliver it to us.

If you say it's a sunny day in Portland, that's information. Since one can't be certain about it, transmitting this information requires some "amount of knowledge": in the message itself and in change of my "grand total" of "knowledge about the world".

Without prior knowledge, "it is a sunny day in Hawaii" transmits less information, since Hawaii has more sunny days. "It is a sunny day in Seattle" transmits more.

Shannon has put this together as a simple formula that measures this "level of uncertainty". And called in entropy.

(Btw, the way he set up the experiment of guessing the next letter of an unknown book by reading it letter-by-letter is priceless: exactly the type of experiment that one had to conduct back then!)

The notion of entropy is uniform. So we can put together multiple disconnected facts ("It is raining and Portland" and "The conference starts at 1pm") and measure the amount of information in this joint message together.

This way, Shannon connected the mathematical term "bit", that we now use to create bytes, megabytes and terabytes, to the informational concept of knowledge.

That was the huge part!

Measuring the shortest message length is just one application of the idea, there are many, many more.

Information Theory: I have a random time series and I am calculating Shannon entropy and permutation entropy for that series. How do they differ?
1,271 Views

Where can I find the most elaborate explanation of entropy?
1,524 Views

What is an intuitive explanation of the concept of "strangeness" in particle physics?
1,100 Views

Speech Processing: What is the intuitive explanation of Shannon's information theory?
831 Views

I read that one of the major results of information theory is that information and entropy are roughly the same thing. In what respect is this true?
1,827 Views

What is an intuitive explanation of torque?
1,911 Views

Is there any simple way I can use information theory or any form of entropy in reinforcement learning or Q learning?
1,034 Views

What is an intuitive explanation of Renyi entropy? And how is it different from KL divergence?
3,941 Views

Information Theory: Is there a way to estimate the Shannon entropy for human DNA based on real population statistics?
617 Views

What is the relationship between entropy in information theory and entropy in thermodynamics?
1,098 Views

Entropy (information theory): Is negentropy a respected concept in scientific circles?
1,258 Views

In layman's terms, what is the Loschmidt's Paradox?
96,230 Views

What is an intuitive explanation of the spectral theorem?
86,045 Views

applications in both statistical thermodynamics and information theory? If so, what is it?
2,187 Views

What is the most intuitive explanation for the concept of poles and zeroes?
10,061 Views

If information is negative entropy, then what is positive entropy in terms of information?
888 Views

What is an intuitive explanation of the D'Alembertian?
2,409 Views

Does entropy apply to information?
1,503 Views

What is an intuitive explanation of Functoriality?
731 Views

What is an intuitive explanation for the role of Information Theory in Cryptography?
714 Views