Notes

Dr Greig Russell

09 February, 2020

Contents

List of Figures

1 Introduction

This research will focus on information and its relationship to data. It will also consider the relationship between information and knowledge or semantic meaning.

Far from an academic historical journey through the arcane theories of information, this research aims to understand the distinctions between data, information and knowledge or meaning. The differences between these three entities have become critical in this age of big data and claims that the owners of such vast troves of information now know you and can manipulate you for their benefit as a consequence.

News can now be fake as opposed to being false or merely propaganda and lies. Conspiracy theories abound despite the internet supposedly providing free access to information and knowledge. The emergence of anti-vaccination theories is associated with a resurgence of fatal conditions once thought near eradication.

Equally, data science releases hidden potential benefit from within vast data store and information technology booms as an industry. My dishwasher is now part of the internet of things allowing me to control it from my phone even if I am still not quite sure why I need to be able to control my dishwasher from my phone.

2 The naive perspective on information

From a naive perspective, information has two common usages. The first is one can have “information about” something and the other is “informed by” something.

Being informed-by refers to the idea that some knowledge informs a belief. This claim makes the associated belief stronger than a simple belief but maybe not as strong as a justified belief. There is still the sense that it is probably true as far as it goes, but that a justified belief requires something more.

This sense is when comments like “I can see why you think that, but. . . ” are found in a conversation. Similarly, when information serves as the basis for subsequent actions. This usage occurs in court evidence, where prior knowledge justifies some following steps. The Police might describe that they were “informed by the defendants previous offending”, obtained a search warrant to look for stolen goods.

The latter gives that sense of an informed belief being stronger than just a guess, “the suspect was acting oddly”, but still not being a justified belief that the suspect was the offender because the stolen goods were in his possession. Almost in this sense, that a belief when informed-by something is more of a justified guess than a justified belief so is more of an opinion.

In contrast, having information-about something is to know about that thing and its properties. Being based on observation, the teachings of others or previous experience serves as the justification. Maybe having information-about something is strong enough evidence on which to build justified beliefs about the thing. Both the Courts and Science certainly think so.

Eye witness accounts of events are, traditionally, seen as damming evidence against the accused in court proceedings. Side-stepping the more nuanced recent literature on the subject, the defence against a charge is undermined by a claim of the defendant was witnessed committing the act.

In science, this usage is even more apparent. A zoologist learns information-about an organism and its behaviour. In an MRI machine, the spin properties of the hydrogen atom depend on the other molecules nearby. An analytic ultracentrifuge causes a large macromolecule to fall apart so revealing the molecule’s quaternary and tertiary structure.

The question though is how does the data derived from such observations become information about that subject? So I might observe the properties of a lion, but at what point does lion become a semantic concept about which I have a justified belief.

3 The sceptical perspective on information

Historically there are key two questions. Is there a difference between data and information? How is the semantic meaning attached to the information associated with it?

Consider the Chest X-ray. Once taken, it contains all the data that it will ever provide as to the patient’s chest. There is no extra data available to the medical student, the attending doctor or the specialist radiologist. Yet all three observers might obtain quite different information from the same film (data source). Indeed, the three observers could be the same observer looking at the same chest x-ray film at three different times in their career.

Even take a slightly broader view. Science repeatedly refines the information to be gained from a given set of observations as the underlying theory evolves. The Copernican revolution, where his heliocentric model of the solar system replaced the previous Ptolemy’s geocentric model, was not based on new data. Instead, the data was re-interpreted to provide a more explanatory model.

Indeed, the rise of the flat earth movement on YouTube in the last five years arises in part due to doubts about this paradigm shift and advocates for a return to geocentrism. This movement disputes the authenticity of some more recent observations (post-1573) and favours a naive approach of gazing vaguely out to sea as a defence against some pan-global conspiracy.

Both are examples of how one set of data can inform three quite different beliefs. Yet this is described as the information age [Castells, 1996]. Where “big data” is collected and used to enlighten or control us, depending on the observer’s perspective. Within this information paradigm, data is a physical object that persists over time. It needs to be collected, stored, and then it can be further processed, creating subsequent products, quite analogous to wheat and describing wheat’s journey from the field to the serving as the basis for various bread and cakes.

In the information age, data is information and information is knowledge. A sceptic may be less persuaded of the equivalence.

4 Key Questions

This research will explore two possibly related questions from a historical and sceptical perspective.

1. What is the ontology of information and by default, how is information different from data, using Dretske’s theory of knowledge?
2. How does information cause belief and hence either knowledge or the development of semantic meaning from the perspective of Lewis’ counterfactual theory of causation?

The research will also consider alternative theories of information based on the perspectives gained through the primary analysis.

It will (may) argue that the former falls short, while the latter may be possible but only within a very circumscribed universe of discourse. As a vehicle for discussion, the assumption will be made that information and data are distinct entities, although any encoded data may contain a subset of the original information.

5 Shannon

5.1 Data is not information, unless it is surprising.

In 1948, Claude Shannon first published his Mathematical Theory of Communication [Shannon, 1948], commonly abbreviated to MTC. The development of MTC occurred within a broader historical context, which is crucial to understanding Shannon’s theory.

The principle communication of the era was the telex, which was essential to Shannon’s employer Bell Laboratories was a part of AT&T (The American Telephone & Telegraph Company) [1]. Despite holding the original patent on the telephone, first awarded to AT&T’s founder Alexander Graham Bell, by 1894, the phone was off patent. AT&T went from being a monopoly to having over 6000 competitors by 1904. Unable to compete at the local level, AT&T focused on long-distance integration of these local systems and long-distance business communication via first the telegraph and then the telex.

Paying per word (with a minimum of 10 words) meant that the telegram needed to optimised for information density [Ross, 1928]. To increase this information density beyond prose optimisation and to provide commercial privacy, manual codes, which were then encrypted by machines, were developed [Davies, 1997]. The most famous of these machines was the German Enigma machine first sold in 1923, which became the focus of the cryptographic war by the Poles lead by Marian Rejwski at first, then the English under Alan Turing and later the Americans. The Engima machine applied a series of mathematical transformations to the encoded (or compressed) messages to provide the cryptographic layer.

These principles still exist today. The airline industry uses codes to increase the information density with examples being the Meteorological Terminal Aviation Routine Weather Report (METAR), providing the weather forecast at the airport. Concerns about internet privacy have led to the proliferation of Virtual

[1] A history of AT&T downloaded from http://www.winlab.rutgers.edu/~narayan/Course/Wireless_Revolution/LL1-%20Lecture%201%20reading-%20ATT%20History.doc downloaded Dec.2019

Private Network (VPN) providers. VPN's encrypt the user's web requests and distribute them from a remote server, which may even be in a different country — hiding the original identity, location and content of the messages.

Shannon, as a Bell Labs employee, was a contractor for the US Government and worked on cryptography, including with Alan Turing during the latter's time in the US, especially on decyphering messages encoded with the Enigma machine. Publication of a declassified version of Shannon's paper on "Communication Theory of secrecy systems" was delayed from 1945 until 1949 [Shannon, 1949].

Unsurprising then, from this historical context, are the critical elements of Shannon's MTC;

1. There is a clear separation between the message and the information contained within it.
2. Only the former is encoded and transmitted as data.
3. Communication of data is a mathematical and statistical process.
4. A key focus of the MTC is optimising the communication channel, especially with regards to signal noise degrading the original message.

Nevertheless, the publication of the MTC caused a paradigm shift. MTC describes the language of communication and information adopted by the field subsequently. However, whether MTC discusses the transmission of data or the dissemination of information with semantic meaning remains unclear.

To be clear, Shannon explicitly states that semantic meaning is irrelevant to the engineering problem. The engineering problem is to be able to send every possible option from a finite list of possible messages [Shannon, 1948, Introduction]. Despite Shannon's injunction to the contrary, many authors in discussing the MTC confabulate message with the information within the message [Stone, 2015, section 1.4].

To labour the point to avoid later confusion image a couple faced with a decision between two choices, who decide to settle the matter by the toss of a coin. A coin is a fixed random choice generator; there are a head side and a non-head or tail side, each of which occurs precisely 50% of the time. The nature of a coin does not change despite the specific choices or their subsequent real-world ramifications. Heads may represent that "you will drop the kids at sport" or "we will eat out tonight", but this situational representation (the semantic meaning) is independent of the nature of a coin and its properties.

As will be discussed later, some authors will talk about semantic meaning as being encoded within the message options. This claim of equivalence between the messenger and the contents of a message is dubious. Without a knowledge of the underlying code that links the message with the messenger only knowing the latter does not reveal the former. If a visitor was looking for the person who lost the coin flip about who would take the children to sport, seeing a coin with tails up on the table, conveys no information to them unless they knew the code and can therefore also decode the message.

The general framework of Shannon's communication system is

Underpinning this schematic model is the representation of data by its encoding [Stone, 2015]. In our simplistic example, a single coin differentiates between two equally likely options. Two coins would enable the use of four different options for the message.

One coin represents one bit of information (one binary digit), while two coins represent two bits of information (two binary digits). More formally, to differentiate between m messages we need n bits of information (coins) [Stone, 2015, section 1.3].

$$m = 2^n \tag{1}$$

Alternatively, n bits of information can describe m messages [**?**, information, section 1.3].

$$n = log_2 m \tag{2}$$

The underlying assumption is that all events are equally likely. More realistically, each message option will occur with a different probability. Where for a given system the probabilities will sum to 1;

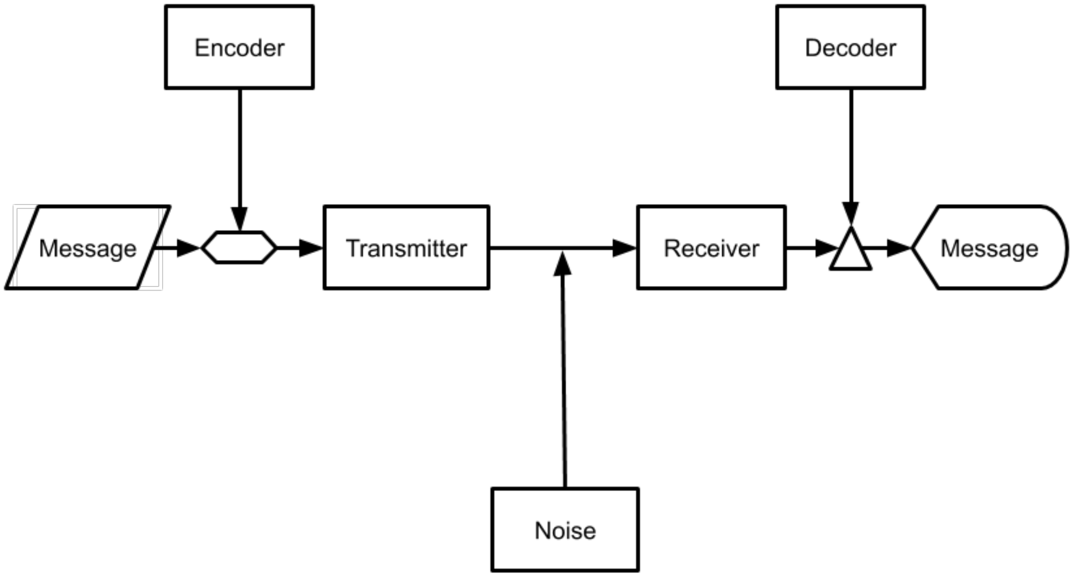

Figure 1: General Communication system From Fig.1 [Shannon, 1948]

$$p(X) = \sum_{i=0}^{n}(pX_i) \tag{3}$$

The Shannon information from receiving a particular message option is given by;

$$I(X) = -log_2(pX) \tag{4}$$

For Shannon, Entropy is a measure of choice within a given system of messages [Shannon and Weaver, 1963]. Although this will be discussed in greater detail later, Weaver [1953] describes how the average Entropy of a message is related to the total entropy within a given system of messages.

$$meanH(X) \simeq \frac{1}{n}\sum_{i=1}^{n} pX_i.log_2\frac{1}{pX_i} \tag{5}$$

Giving rise to information quantity being how unlikely that message is to be received. Rare messages are more critical in a given closed system than common messages. There is an equivalence with “click-bait” journalism where the headline claims the article some surprising and vital message.

Hence the Entropy of a system is also a measure of the systems (Shannon’s) information. The term Entropy occurs in mechanical engineering and relates to the increase in energy states within a closed system like the steam engine. Shannon borrowed the term as his information concept was analogous to the mechanical engineering original. (note to self find the reference).

The other key concept is “equivocation” or loss of information through the communication channel [Shannon and Weaver, 1963]. For Stone [2015] (section 4), equivocation is the average uncertainty about the original message after the signal was received.

$$C = H(X) - H(X|Y) \tag{6}$$

To unpack this equation, the term H(X) is the spread of possible messages that you could receive from the channel (C). Less the uncertainty that the message packages are X given that Y was received. If the channel is noiseless, then H(X|Y) will be zero, and no packets can be lost. The noisier the channel, the higher the probability that some message packages will be lost increases, as will our doubts.

This perspective on information is an engineering perspective and related to the communication system itself. Consider this from the perspective of the population of messages the engineer needs to deliver to the customer. If a particular message occurs frequently, then the unique information available to the customer in that one message is small. If the message infrequently occurs, then the individual information contained in the payload is of higher importance to the communication system and maybe the customer.

Consider a system to inform a technician if a critical system is on-line. If they receive one message every minute, hopefully, most of which will tell them the system is operating normally. From Shannon’s perspective, the unique information to the customer is small in each message and has a short latency before being updated a minute later. Once a week or month, some error state message is sent. Regardless of the importance to the customer, the rarity means this error message is vital to the communications engineer. There will not be an opportunity to update the payload for many 100s or 1000s of cycles. If lost (dropped in technical terms) the potential consequences are much more significant.

Such events are a real-world problem for the IT and communications industry. Because systems usually work for long periods, technicians stop checking the error logs. In the example of back-up and recovery systems, not infrequently their failure, months prior, is only noted when the primary system fails as well.

This analysis assumes that messages have a normal distribution. This assumption may not be correct. A heavily skewed distribution will have many unlikely events in the long tail, many of which will be unlikely to occur events or messages. The difference between the probability of adjacent or near neighbour events is small, but from Shannon's information, being inverse to the message probability artificially magnifies these small differences into more substantial differences in the information contained within them. The other primary assumption is that this is a closed system, which is correct from Shannon's perspective as a communication engineer and the type of problems he needed to address. The real world is an open system with an infinite set of rare events. It can not be encoded, and sporadic events, like gravitational waves, are seldom of any consequence even if indeed astonishing.

For a communication engineer, this model allows us to understand the spread of possible messages within the system and the impact of message loss in the communication channel (think overhead telephone wire for the era).

In considering Shannon's theory, we need to distinguish between the uniqueness of the message and its relevance to the customer. Frequent events may also contain important information. That a spouse loves their partner today, is deeply valued by then. To Shannon, these everyday events have a little surprise, hence information, but many couples may not agree.

That the cat finally knocked over the ugly gift may be quite surprising. Several nights of effort were needed to put the gift in a vulnerable position with a suitable cat treat behind it. The cat showed extraordinary malice and skill by eating the treat without breaking the gift. Its final success is surprising, and in terms of the distribution of events of the household's usual activities, it may be indeed astonishing. The event contains very little actual information, even though Shannon's information is high due to its low probability of occurring.

Shannon, himself, was very clear, that he was describing a theory of communication, not a description of the payload and its semantic meaning to the customer. Others took up this challenge, but from within Shannon's paradigm.

6 Weaver

6.1 Data is information because Shannon's theory is awesome

Weaver [1953] took a broader view of the concept of communication than Shannon [Shannon and Weaver, 1963]. For Weaver, communication was anything that one mind could produce to influence another.

Three problems comprise the challenge of understanding communication.

1. The underlying technical problems.
2. The semantic problems or how is meaning conveyed.
3. The effectiveness problem or how does the message make the intended change in behaviour.

For Weaver, Shannon [1963] had addressed the technical challenges [Weaver, 1953]. Despite Shannon's specific injection that his theory did not solve the latter two problems, Weaver argued the contrary "at least to a significant degree" [Weaver, 1953].

Weaver's perspective on information within communication theory is not on what was said, but what could be said [Weaver, 1953, Section 2.2]. He splits meaning from information, where a single message contains meaning, but the system of possible messages conveys information. So Weaver's unit for measuring total possible information is the base 2 log of the number of choices between message options [Weaver, 1953].

Degrees of freedom is a statistical concept, which means that the antecedent confines the consequent. In communication, this extends to either words or letters within a given language and grammer[Weaver, 1953]. A Markoff chain represents a particular case of such a system. Not only are the possibilities of future choices constrained by the probabilities of previous decisions, but that a random collection of such options also describes the average statistical properties of the whole system. Such highly confined or ergodynic systems are conventional in computer science, particularly algorithms like "parsers", which translate very stylised human-readable code into machine-executable equivalents.

The entropy or total spread of the probabilities is the total amount of information held within a given system of messages where information is the freedom or range of choices available [Weaver, 1953].

$$H(X) \simeq \sum_{i=1}^{n} pX_i.log_2 \frac{1}{pX_i} \tag{7}$$

For a novel speaker of a language, the broader their vocabulary, the more detailed their descriptions can be. Weaver [1953] describes how the redundancy of English is 50%, or that only half the words in a given passage are free choices of the author and not confined by grammar.

Weaver then argues that the concepts within the MTC could apply equally to the problem of semantics through the mechanism of a semantic receiver, which like its communication equivalent, decodes the semantic content with similar issues with semantic noise, provoking equivication[Weaver, 1953]. Although not stated, Weaver appears to suggest that the effectiveness problem is a consequent of the communication and the semantic problem. So if the message arrives intact and is sufficiently understood, it is acted upon correctly.

Weaver's defence or extension of the underlying MTC seems lame. It does not justify why such ergodynic systems like a Markoff chain should be the norm or be generalizable as a concept. The extension of the solution to the technical solution to solving the semantic and effectiveness problems was more of equivalence of the later to the former hence the technical problem solution should also apply.

7 Carnap & Bar-Hillel

7.1 Information has the same properties as data, hence the two are equivalent

Carnap et al. [1952] developed another approach to consider information and particularly semantic information from within Shannon's paradigm.

This approach was from the perspective of inductive logic, with the base unit of information was the sentence. So Pa means that the individual or object a has the property P. More complex sentences are constructed primitive sentences through the use of classical predicate calculus operators such as v, $\hat{}$, or ~ with their usual meanings [Carnap et al., 1952, p.4]. Hence, a language system L_n^π, which describes n objects and π primitive predictors that define the base properties of the objects [Carnap et al., 1952, p.4]. Likewise, sentences can be logically correct or false, while relationships between objects have a logical definition.

From this basis, Carnap et al. [1952] derived that the content of semantic information in a message payload i is;

$$cont(i) = \frac{m}{2^n} \tag{8}$$

Where *cont(i)* is the content, m is an absolute probability function of the content not existing or having no evidence for its existence (p.14) and n is the number of objects in the system.

Carnap et al. [1952] describes their second formulation, considering the measure of information in a sentence.

$$inf(i) = Log \frac{1}{1 - cont(i)} \tag{9}$$

These two elegantly described formulations formulate semantic information or payload of a message as analogous to Shannon's theory of communication and its consideration of the message package itself. The more frequent the information in the sentence, the less the sentence contributes to the conversation on its own. Conversely, the less frequent the message is, the more impact it will have on a conversation.

Yet this also formulation of semantic content creates an evitable paradox namely, that a sentence which is a tautology contains no information. While a sentence that describes a contradiction and must be false, conversely contains the maximal amounts of information possible for the system. So the sentence *17 x 19 = 323* contains no information because it is never false [Carnap et al., 1952].

This thesis would counter-argue that Carnap et al. [1952] describe not a system of semantic information contained within the message payload but a method for understanding the significance of individual messages in the communication channel irrespective of their actual payload.

So to answer the paradox. In a closed communication system with a finite arrange of streamed messages, each iteration of the most common message types tells the operator little. For an operator, silence or no information is a message of infinite and immediate importance. The communication system has broken down!

The Carnap et al. [1952] formulation is parallel to Shannon's communication paradigm. The importance of a message for Shannon is its uniqueness and for Carnap & Bar-Hillel significance is the probability of being false. Neither describes the semantic meaning for either the source or the recipient of the message.

References

Rudolf Carnap, Yehoshua Bar-Hillel, et al. An outline of a theory of semantic information. 1952.

Manuel Castells. *The information age*, volume 98. Oxford Blackwell Publishers, 1996.

Donald Davies. A brief history of cryptography. *Information Security Technical Report*, 2(2):14–17, 1997.

Nelson E Ross. *How to write telegrams properly.* Haldeman-Julius Publications, 1928.

Claude E Shannon. communication theory of secrecy systems. *Bell system technical journal*, 28(4):656–715, 1949.

Claude Elwood Shannon. A mathematical theory of communication. *Bell system technical journal*, 27(3): 379–423, 1948.

Claude Elwood Shannon and Warren Weaver. The mathematical theory of communication–univ. *Illinois press, Urbana, I*, 11:117, 1963.

James V Stone. *Information theory: a tutorial introduction.* Sebtel Press, 2015.

Warren Weaver. Recent contributions to the mathematical theory of communication. *ETC: a review of general semantics*, pages 261–281, 1953.