The emergence of Information Theory and its modern implications

Louis Wehenkel

Main sources:

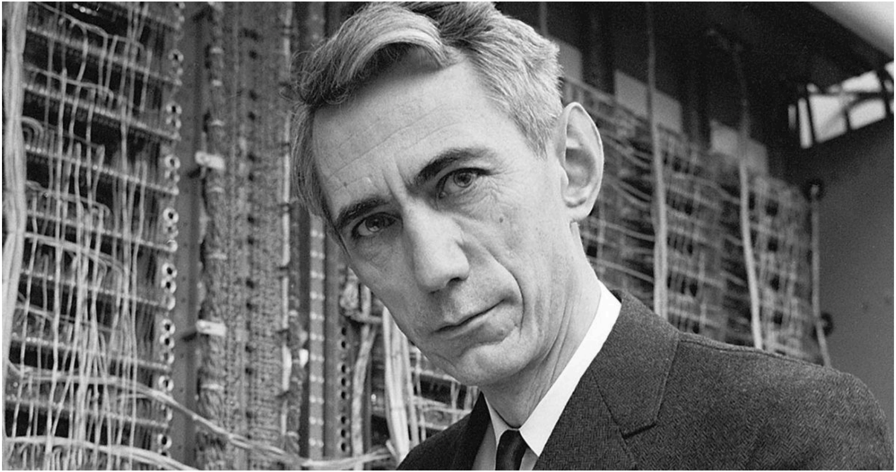
i) the landmark paper "A Mathematical Theory of Communication", by C.E. Shannon in 1948 in the Bell System Technical Journal,

ii) the book "Claude Elwood Shannon - Collected Papers", published by the IEEE Information Society in 1991.

ELEN0060-2: 6/02/2020

LIÈGE université

Montefiore Institute



Claude Elwood Shannon (1916-2001)

Claude Shannon's Background

- 1936: Basic education at the University of Michigan
  Bachelors degrees in EE and Maths
- 1938: Master's Thesis at MIT,
  "A Symbolic Analysis of Relay and Switching Circuits"
- 1940: PhD Thesis at MIT,
  "An Algebra for Theoretical Genetics"
- In 1941 C. Shannon joined Bell Labs (New Jersey).

  During the WW-II period, he worked on various problems, including communications, cryptography, fire control etc., together with many renowned scientists.

  At Bell Labs, he also met with his future wife Betty. They married in 1949 and had three children.

Observed facts (at the time)

O1. Very-long-distance communication of analog signals did not work, because noise accumulation eventually makes the signal impossible to distinguish from the noise.

O2. Among human beings, communication in natural language (written or spoken) shows to be very robust to noise

O3. Human communication is essentially the communication of sequences of words from a finite set of possible words.

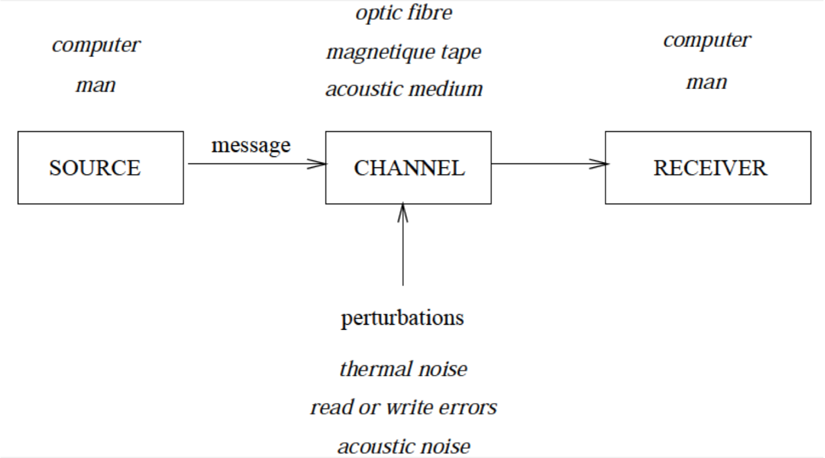
O4. Analog signals can be represented in digital form while controlling the loss of ‘information’, e.g. by sampling and quantization.

O5. Digital signals can be transmitted over analog channels, e.g. via proper modulation techniques.

Motivation (of Shannon ?)

- It was believed that in order to effectively reach error-free communication over noisy channels one would either need an infinite amount of power or an infinite amount of time.
- This belief is in contradiction with observation "O2": humans can communicate 'super efficiently' with bounded resources.
- To resolve such a contradiction in a convincing way, a mathematical theory of "Communication" is required.
- Shannon started to work on this objective in 1941... and published his opinion about this question in 1948.
- His 1948 paper was on the one hand immediately criticised (by outside mathematicians and by some politicians), but on the other hand immediately recognised as a landmark (by the concerned engineering community). 6000+ citations/yr

Shannon's paradigm for modelling communication



Shannon's scientific strategy in 1948

1. Let's first forget about noise: under the "perfect channel" assumption, what does it mean for a source to communicate information, and what can we achieve in these circumstances
2. Next, let's consider the noisy channel case.
3. To develop the theory let's
   1. First consider the discrete case (digital communication)
   2. Then the continuous case (analog communication)
   3. Then wrap up and consider mixed analog-digital cases

NB: he did all this in his 79 pages long landmark 1948 paper; where he defined the mathematical model and provided the proofs of his "3 Theorems" while also sketching novel techniques for proving mathematical theorems.

Step 1: Modelling a simple discrete source

1. **THE DISCRETE MEMORYLESS SOURCE MODEL:**
   A device $\mathcal{S}$ that emits messages in the form of sequences of 'symbols' $\mathcal{X}_i$ chosen randomly and independently from a finite alphabet $\mathcal{A}$ according to a probability mass function $P(\cdot)$.

2. **DEF1: THE ENTROPY RATE of $\mathcal{S}$:**

$$H_{\mathcal{S}} = -\sum_{x \in \mathcal{A}} P(x) \log_2 P(x) \quad \text{bit/symbol} \tag{1}$$

   measures the information produced per symbol emitted by such a source.

3. **THEOREM 1: By using source codes, it is possible to reversibly re-encode (i.e. compress) long messages of a source $\mathcal{S}$ so that on average you need only $H_{\mathcal{S}}$ bits per source symbol, and this is the best you can achieve.**

Step 2: Modelling a simple noisy channel

1. **THE MEMORYLESS BINARY SYMMETRIC CHANNEL MODEL:**
   A device $\mathcal{C}$ receiving at each time step a bit at its input and producing the same bit with prob. $1 - p_e$ at its output.
2. **DEF2: THE COMMUNICATION CAPACITY of $\mathcal{C}$:**

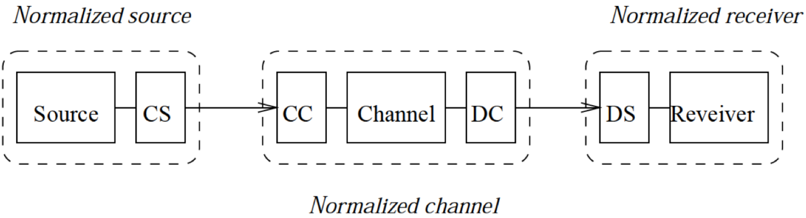
$$C_{\mathcal{C}} = 1 - H_2(p_e) \ \text{ bits/channel usage} \tag{2}$$

   where

$$H_2(p_e) = -[p_e \log_2 p_e + (1 - p_e) \log_2 (1 - p_e)] \tag{3}$$

3. **THEOREM 2: By using channel codes, it is possible to reliably transmit long messages over the channel $\mathcal{C}$ at a rate of $C_{\mathcal{C}}$ source bits per channel usage, and <u>this is the best you can achieve.</u>**

Resulting “Engineered” Communication system



- The source coding/decoding blocks 'CS' and 'DS' only depend on the source properties
- The channel coding/decoding blocks 'CC' and 'DC' only depend on the channel properties
- Optimal performances can be reached by designing these two sub-systems independently!

Numerical examples of discrete sources

1. Order-zero White&Black 'fax' model:
   - $\mathcal{A} = \{0, 1\}$, described by $P(\mathcal{X}_i = 1) = p_1$
   - $H_{\mathcal{S}} = -p_1 \log_2 p_1 - (1 - p_1) \log_2 (1 - p_1)$ bit/pixel
   - e.g. if $p_1 = 0.1$, we have $H_{\mathcal{S}} \approx 0.47$, so that a message of $1G$pixels emitted by such a source can be encoded reversibly with on the average only $0.47G$bits

2. Written English language (Markov chain model):
   - $\mathcal{A} = \{\text{blank}, A, B, \ldots, Z\}$ (27 "letters")
   - $H_{\mathcal{S}} \leq \log_2 27 = 4.75$ bit/letter (would be reached if letters were independent and equally likely to occur)
   - In reality $H_{\mathcal{S}} \approx 1$ bit/letter, because some subsequences of letters are less frequent than others.
   - Thus English text has about $1 - \frac{1}{\log_2 27} = 0.79$ redundancy.

Numerical example of discrete channel

- Consider a binary symmetric channel with $p_e = 0.1$
- We have $H_2(p_e = 0.1) = 0.47$
- Hence $C_{0.1} = 1 - H_2 = 0.53$
- It means that it is possible to communicate in an essentially error-free fashion over such a channel, by using a channel code that on average transmits 2 bits for every source bit.
- WRAPPING up: we can transmit our $1G$bit fax in one second over such a $1G$bit/second channel while ensuring 'zero'-errors.

Subsequent steps of the landmark paper

They consisted of relaxing the "memoryless" assumptions and to bridge the gap from digital to analog and vice versa:

1. Study of English language via Markov models: estimation of information rate to about 1 bit per letter (using a proper generalization of the notion of source entropy rate).

2. Investigation of what it means to reliably communicate analog signals over non noisy channels: led to the so-called rate-distorsion theory and vector quantization.

3. Study of physical (analog) channels: led to the extension of entropies and capacity measures over continuous spaces, and the quantification of the information transmission capacity of physical channels under power limitations and accounting for the noise level in the channel.

Impact post 1950 in communications

- Source coding problem: Huffman (another MIT student), solved the source coding problem in 1952; further work on reversible/irreversible data compression continued.
- Notion of ergodic systems and their asymptotic equipartion property: lots of insights were gained by other scientists by adopting this viewpoint.
- Channel coding problem: many brilliant minds worked on this topic; partial solutions enabled space exploration from 1960 onwards; more recent work yielded near-Shannon limit error-correcting codes (e.g. Turbo-codes in the 1990's).
- Multi-user/multi-channel information theory has been a main topic of research, with further theoretical and practical results.
- Joint source-channel coding is another active area of research.

Impact post 1950 in other fields

- In the early 1950's the field of cybernetics yielded hot debates both in science and in politics... The word was coined by Norbert Wiener, to model natural agents with the principles of information and control theories...
- The field of artificial intelligence emerged as a discipline in the mid 1950's, building on Turing's, Shannon's and others works...
- In the early days (right after the publication of the landmark paper), many other fields (physics, sociology, psychology etc) imported the ideas of Shannon in their way of thinking...
- In 1956, C. Shannon published his Bandwagon paper: he was uncomfortable with the mood that information theory would be a kind of universal panacea in almost every field...

Modern implications

- Machine learning and data science: information theory provides the ground principles of quantitative and qualitative thinking, modelling, and problem solving in these fields.
- Computers and communication: information theory has been and remains a main enabler of reliable communication with non ideal devices, and it most probably shall be fundamental to enable reliable computation over networks of massively connected non ideal devices.
- Cybernetics: the current focuses on neuroscience and neuromorphic engineered systems aim at reproducing the human central nervous systems on computers; information theory expresses some of the fundamental constraints that such system will have to comply with.

Further reading & listening material

- A super-cool historical video: C Shannon - Father of the Information Age: https://www.youtube.com/watch?v=z2Whj_nL-x8
- A more academic but extremely well done video: D MacKay - Introduction to Information Theory: http://videolectures.net/mackay_course_01/
- A pointer for further reading: Leon Brillouin - Science and Information Theory: https://books.google.be/books/about/Science_and_Information_Theory.html?id=DWo7lVRVnhcC&redir_esc=y
- Shannon's Bandwagon paper: https://www.monoskop.org/images/2/2f/Shannon_Claude_E_1956_The_Bandwagon.pdf