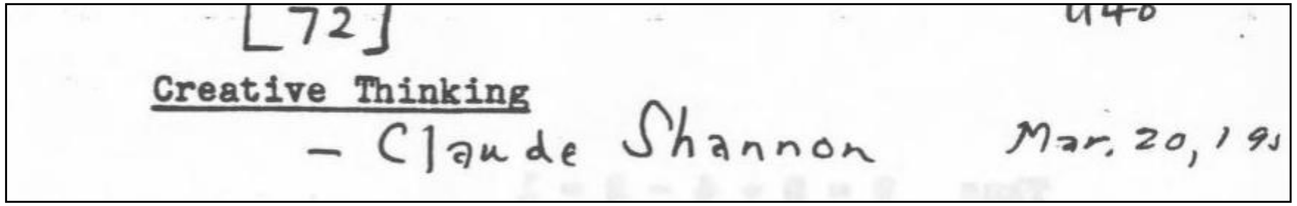
Representation of Information
ECE 598 LV – Lecture 2

Lav R. Varshney

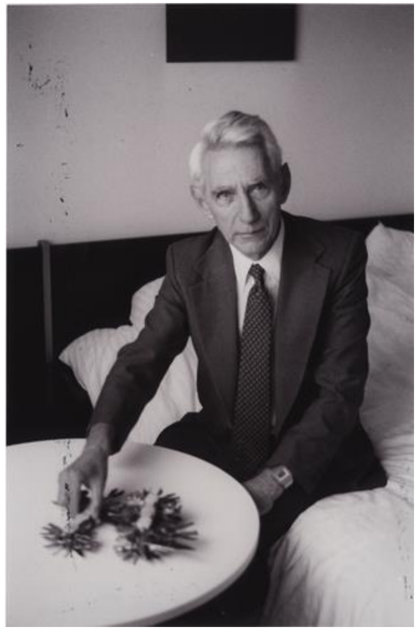
18 January 2024

Tricks for formulating/solving problems (principles of theoretical research)


[72]

Creative Thinking

– Claude Shannon   Mar. 20, 19

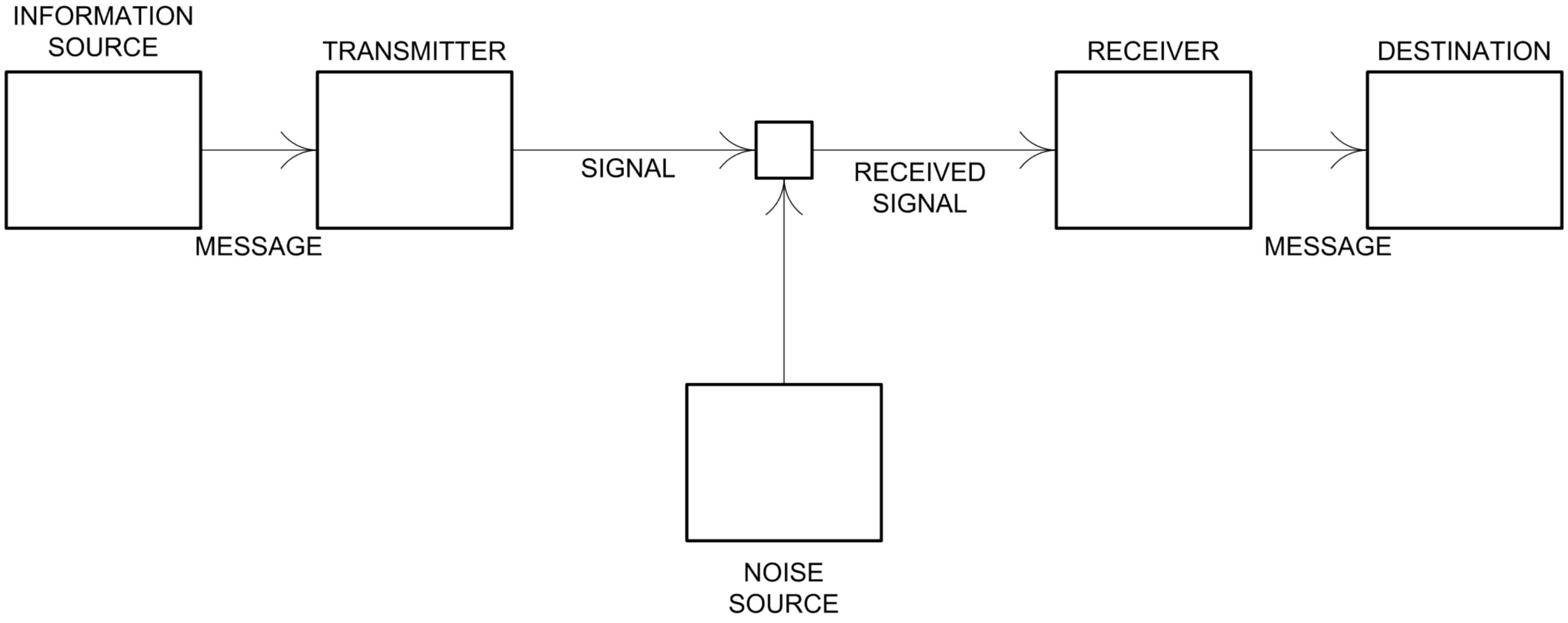


1. Simplification: get rid of enough detail (including practical aspects) for intuitive understanding
2. Similarity to a known problem (experience helps)
3. Reformulate (avoid getting in a rut)
4. Generalize (more than opposite of simplify)
5. Structural analysis (break problem into pieces)
6. Inversion (work back from desired result)

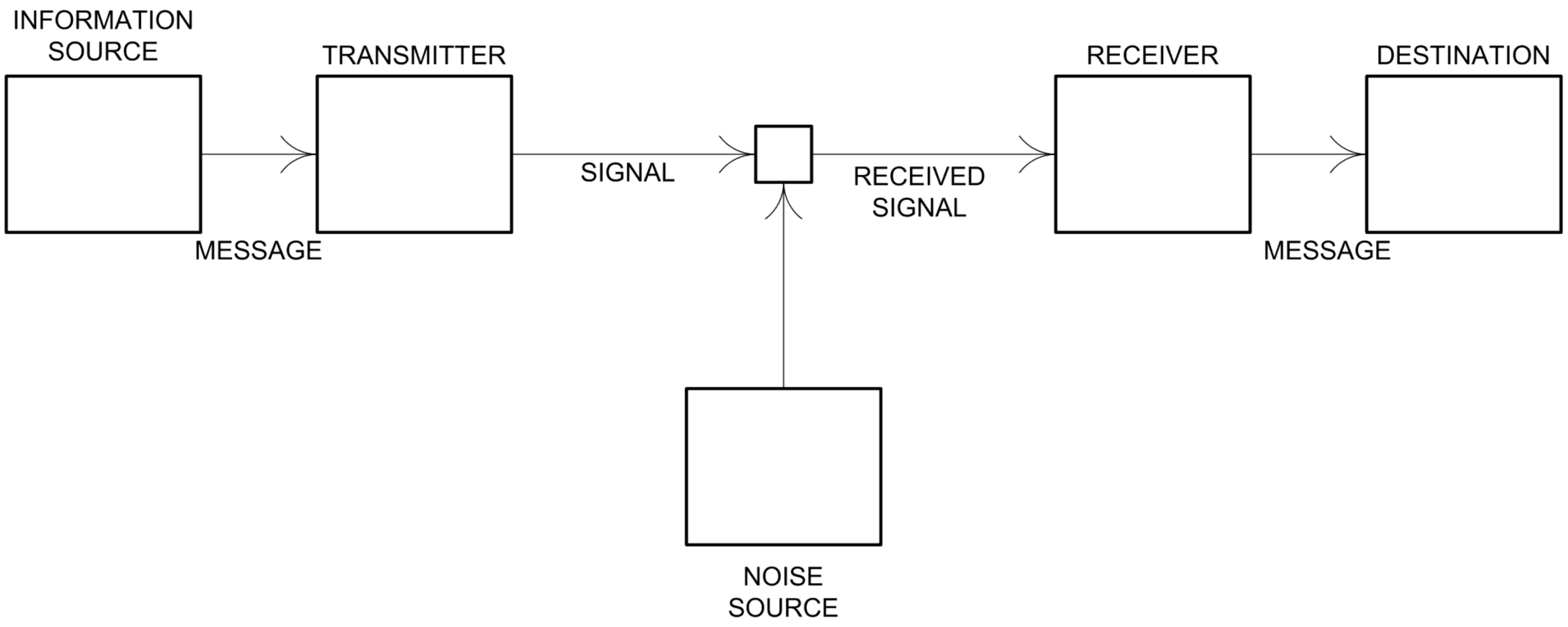
Shannon a la Gallager

- Shannon was almost opposite of applied mathematicians
  - Applied mathematicians solve mathematical models formulated by others (perhaps with minor changes to suit their tools)
- Shannon was a creator of models — his genius lay in determining the core of the problem and removing details that could be reinserted later

- Shannon was interested in several problems at all times
- Shannon studied what was happening in multiple fields, but didn’t work on what many others were working on
- Shannon asked conceptual questions about everyday things

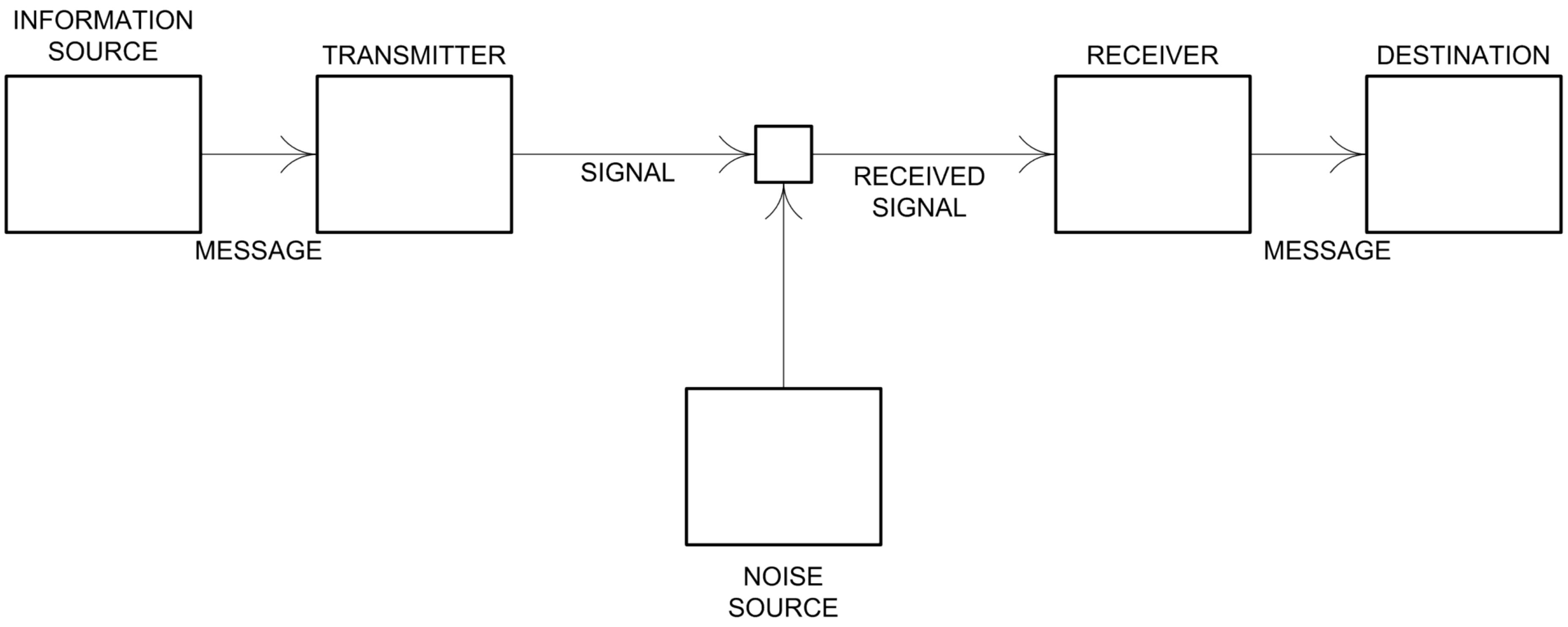


Claude Shannon’s schematic diagram of a general communication system (1948: Figure 1)



Claude Shannon’s schematic diagram of a general communication system (1948: Figure 1)

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point.

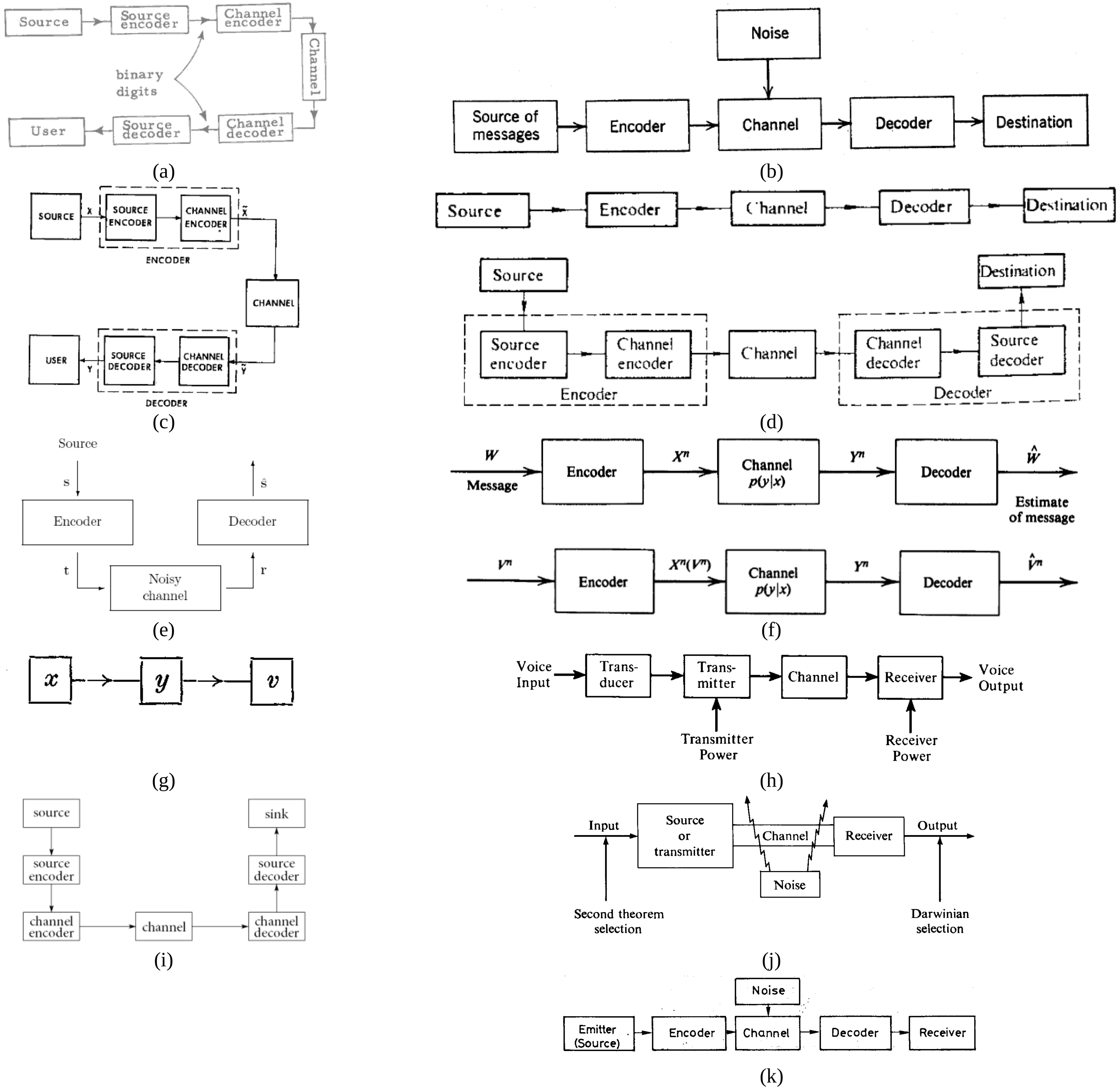


Claude Shannon’s schematic diagram of a general communication system (1948: Figure 1)

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message **selected** at another point.

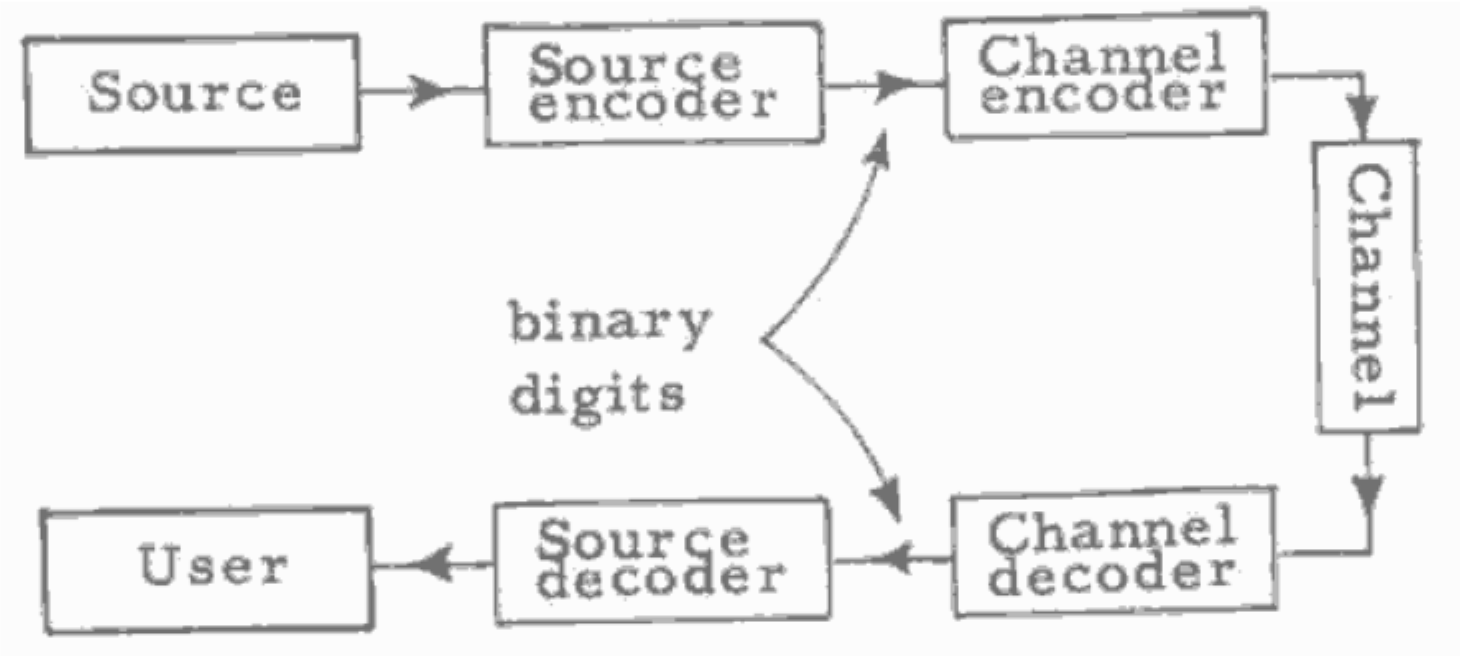
Big idea #1:

Communication is a statistical problem



Various incarnations of Shannon's schematic for a general communication system in textbooks.

(a) Fano (1961: Figure 1.1). (b) Ash (1965: Figure 1.1.1). (c) Berger (1971: Figure 1.2.1). (d) two figures from Csiszár and Körner (1997: Figures 2.1 and 2.2). (e) MacKay (2003: Figure 1.6). (f) two figures from Cover and Thomas (1991: Figures 8.1 and 8.12). (g) Woodward (1953: 58). (h) Hancock (1972: Figure 1.1). (i) Richardson and Urbanke (2008: Figure 1.2). (j) Gatlin (1972: Figure 17). (k) Rényi (1984: 43).



[Fano, *Transmission of Information*, 1961.]

before we had the theory,… we had been dealing with a commodity that we could never see or really define. We were in the situation petroleum engineers would be in if they didn’t have a measuring unit like the gallon. We had intuitive feelings about these matters, but we didn’t have a clear understanding

– Jerome Wiesner (1953)

Certain Factors Affecting Telegraph Speed[1]

By H. NYQUIST

Theoretical Possibilities Using Codes with Different Numbers of Current Values

The speed at which intelligence can be transmitted over a telegraph circuit with a given line speed, *i.e.*, a given rate of sending of signal elements, may be determined approximately by the following formula, the derivation of which is given in Appendix B.

$$W = K \log m$$

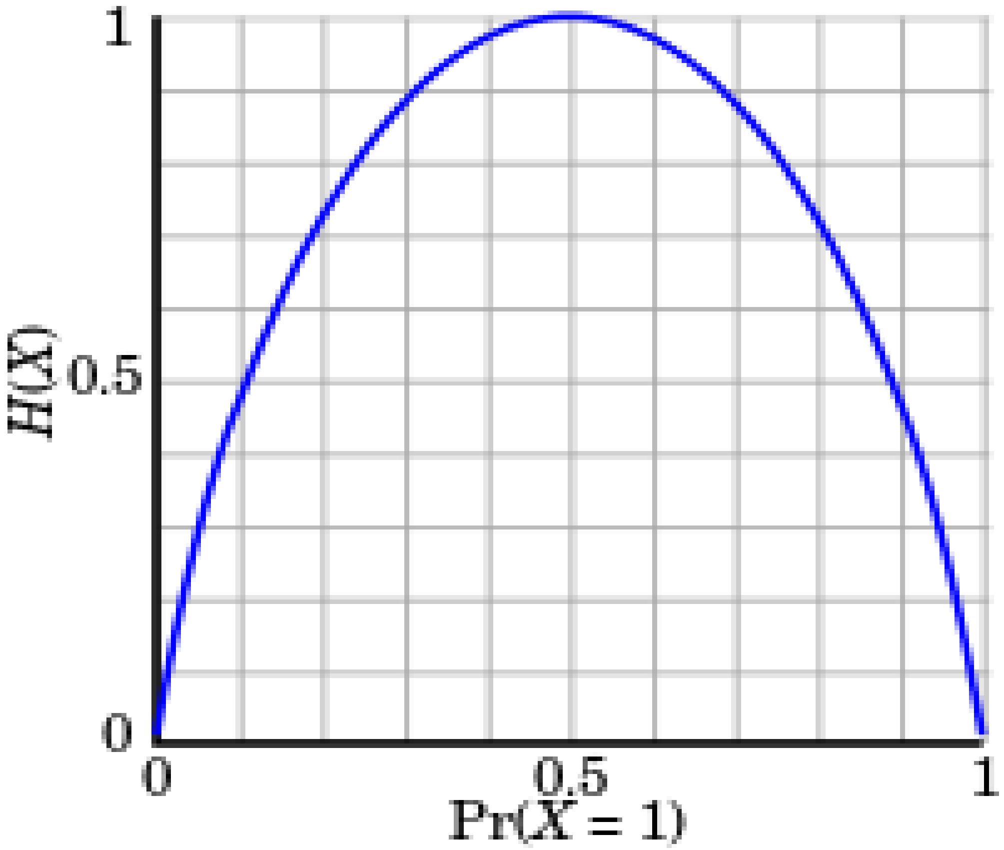
Where W is the speed of transmission of intelligence,
m is the number of current values,
and, K is a constant.

[H. Nyquist, "Certain Factors Affecting Telegraph Speed," *Bell System Technical Journal*, 1924.]

- If the following messages are equally likely, how many bits are being produced?

1. {01101, 11101}
2. {1, 0}
3. {Ẉ, Ẋ}
4. {3333333333, 4444444}
5. {01, 10, 11, 00}
6. {000, 111, 110, 101}

Not just possibilities but probabilities (from Nyquist to Shannon)



[https://en.wikipedia.org/wiki/Binary_entropy_function#/media/File:Binary_entropy_plot.svg]

Big idea #2:

There is a notion of information rate, which can be measured in bits

Axiomatic derivation of mutual information

Let X and Y be discrete random variables with respective alphabets $\mathcal{X}$ and $\mathcal{Y}$. It may help to think of X and Y as representing the input and output of some digital communication system. We are interested in quantifying the amount of information that observation of the occurrence of the event $[Y = y]$ provides about whether or not the event $[X = x]$ also has occurred. We denote this quantity by $I(x, y)$. We assume knowledge of the joint distribution $p(x, y) = Pr[X = x, Y = y]$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. This, of course, provides us with knowledge of the associated marginal distributions $\{p(x), x \in \mathcal{X}\}$ and $\{q(y), y \in \mathcal{Y}\}$ and conditional distributions $\{p(x|y)\}$ and $\{q(y|x)\}$.

We now introduce four postulates, or requirements, that most people consider it reasonable that $I(x, y)$ should obey. After each postulate is introduced, we name it and try to describe the motivation underlying it.

Postulate A. There exists a function $F(\alpha, \beta)$ such that $I(x, y) = F(\alpha, \beta)|_{\alpha=p(x), \beta=p(x|y)}$

The idea behind this postulate is that $[Y = y]$ can convey information about $[X = x]$ only by virtue of the fact that it changes the probability of occurrence of $[X = x]$ from its *apriori* value $p(x)$ to its *aposteriori* value $p(x|y)$. We call Postulate A the **Bayesian Postulate** because it imbeds information into the Bayesian framework for reasoning probabilistically from observations back to their possible causes.

Postulate B. The partial derivatives of $F(\alpha, \beta)$ exist.

That is, $F_1(\alpha, \beta) = \frac{\partial}{\partial \alpha} F(\alpha, \beta)$ and $F_2(\alpha, \beta) = \frac{\partial}{\partial \beta} F(\alpha, \beta)$ exist for $0 \leq \alpha, \beta \leq 1$. We call Postulate A the **Smoothness Postulate**. Since differentiability implies continuity, the Smoothness Postulate implies among other things that an infinitesmal perturbation in the prior or posterior probability of occurrence of $[X = x]$ cannot result in a discontinuous jump in our information measure.

Postulate C. $F(\alpha,\gamma) = F(\alpha,\beta) + F(\beta,\gamma)$, $0 \leq \alpha,\beta,\gamma \leq 1$.

The reasoning underlying Postulate C is that, if y were a vector with two components, say $y = (w,z)$, then the information provided by observing its occurrence would have to be the sum of that provided by observing w and that provided by then observing z. In the first of these two steps the information that $[W = w]$ provides about whether or not $[X = x]$ is $F(p(x), p(x|w))$. Once this information has been provided, the original prior probability $p(x)$ of the event $[X = x]$ is replaced by $p(x|w)$. After $[Z = z]$ subsequently is observed, the posterior probability of occurrence of $[X = x]$ then becomes $p(x|w,z)$, so the additional information provided must be $F(p(x|w), p(x|w,z))$. We conclude that $F(p(x), p(x|w,z)) = F(p(x), p(x|w)) + F(p(x|w), p(x|w,z))$. Since $p(x), p(x|w)$ and $p(x|w,z)$ can range over any numbers in the unit cube in various examples, we are led to Postulate C, which we call the **Successive Revelation Postulate.**

Postulate D. $F(\alpha\gamma, \beta\delta) = F(\alpha, \beta) + F(\gamma, \delta)$, $0 \leq \alpha, \beta, \gamma, \delta \leq 1$.

The motivation behind Postulate D is that, if we have two independent experiments, one with input X and output Y and the other with input U and output V, then the information that observation of the combined output event $[Y = y, V = v]$ provides about whether or not the combined input event $[x = x, U = u]$ occurred should be the sum of that which $[Y = y]$ provides about whether or not $[X = x]$ and that which $[V = v]$ provides about whether or not $[U = u]$. Whenever the (X, Y) and (U, V) experiments are independent, though, the joint prior probability is $p(x, u) = p(x)p(u)$ and the joint posterior probability is $p(x, u|y, v) = p(x|y)p(u|v)$. Hence, we require that

$$F(p(x)p(u), p(x|y)p(u|v)) = F(p(x), p(x|y)) + F(p(u), p(u|v)).$$

Since $p((x), p(x|y), p(u)$ and $p(u|v)$ can assume any values in $[0, 1]^4$ in various examples, we are led to Postulate D, which we call the **Independence Additivity Postulate**.

Postulate A. There exists a function $F(\alpha, \beta)$ such that $I(x, y) = F(\alpha, \beta)|_{\alpha=p(x), \beta=p(x|y)}$

Postulate B. The partial derivatives of $F(\alpha, \beta)$ exist.

Postulate C. $F(\alpha, \gamma) = F(\alpha, \beta) + F(\beta, \gamma)$, $0 \leq \alpha, \beta, \gamma \leq 1$.

Postulate D. $F(\alpha\gamma, \beta\delta) = F(\alpha, \beta) + F(\gamma, \delta)$, $0 \leq \alpha, \beta, \gamma, \delta \leq 1$.

Because of B we may take the partial derivative of both sides of C with respect to β. However, β does not appear on the left hand side of C, so we get

$$0 = F_2(\alpha, \beta) + F_1(\beta, \gamma),$$

or equivalently, $F_2(\alpha, \beta) = -F_1(\beta, \gamma)$. It follows that $F_2(\alpha, \beta)$ cannot vary with α because α does not appear in $F_1(\beta, \gamma)$. That is, $F_2(\alpha, \beta)$ is actually a function only of β which we shall denote by $G'(\beta)$. We have

$$F_2(\alpha, \beta) = -F_1(\beta, \gamma) = G'(\beta).$$

Next observe that if we take the indefinite integral of $F_2(\alpha, \beta)$ with respect to β, we have to get back $F(\alpha, \beta)$ plus a constant of integration, $C = C(\alpha)$, where we have been careful to allow for the fact that the constant may depend on the other argument α in $F(\alpha, \beta)$. That is,

$$\int F_2(\alpha, \beta) d\beta = F(\alpha, \beta) + C(\alpha).$$

[Check this by taking the partial with respect to β and verifying that you get the identity $F_2(\alpha, \beta) = F_2(\alpha, \beta)$.] Hence, we may write

$$\int G'(\beta) d\beta = G(\beta) = F(\alpha, \beta) + C(\alpha),$$

so $F(\alpha, \beta) = G(\beta) - C(\alpha)$. Putting this into C, we get

$$G(\gamma) - C(\alpha) = G(\beta) - C(\alpha) + G(\gamma) - C(\beta),$$

which tells us that $G(\beta) = C(\beta)$. Accordingly,

$$F(\alpha, \beta) = G(\beta) - G(\alpha).$$

Our problem of discovering the functional form of $F(\alpha,\beta)$, a function of two variables, thus has been reduced to that of finding the function $G(\cdot)$ of a single variable.

Now we use Postulate D re-expressed in terms of $G(\cdot)$, namely

$$G(\beta\delta) - G(\alpha\gamma) = G(\beta) - G(\alpha) + G(\delta) - G(\gamma).$$

Taking the derivative of this with respect to δ gives

$$\beta G'(\beta\delta) = G'(\delta).$$

In the limit as $\delta \to 1$ this becomes

$$\beta G'(\beta) = G'(1) = K,$$

where K is a constant. This tells us that

$$G'(\beta) = K/\beta,$$

from which we deduce that

$$G(\beta) = K\ln(\beta) + C,$$

where C is another constant. It follows that

$$F(\alpha,\beta) = K\ln(\beta) + C - K\ln(\alpha) - C,$$

or

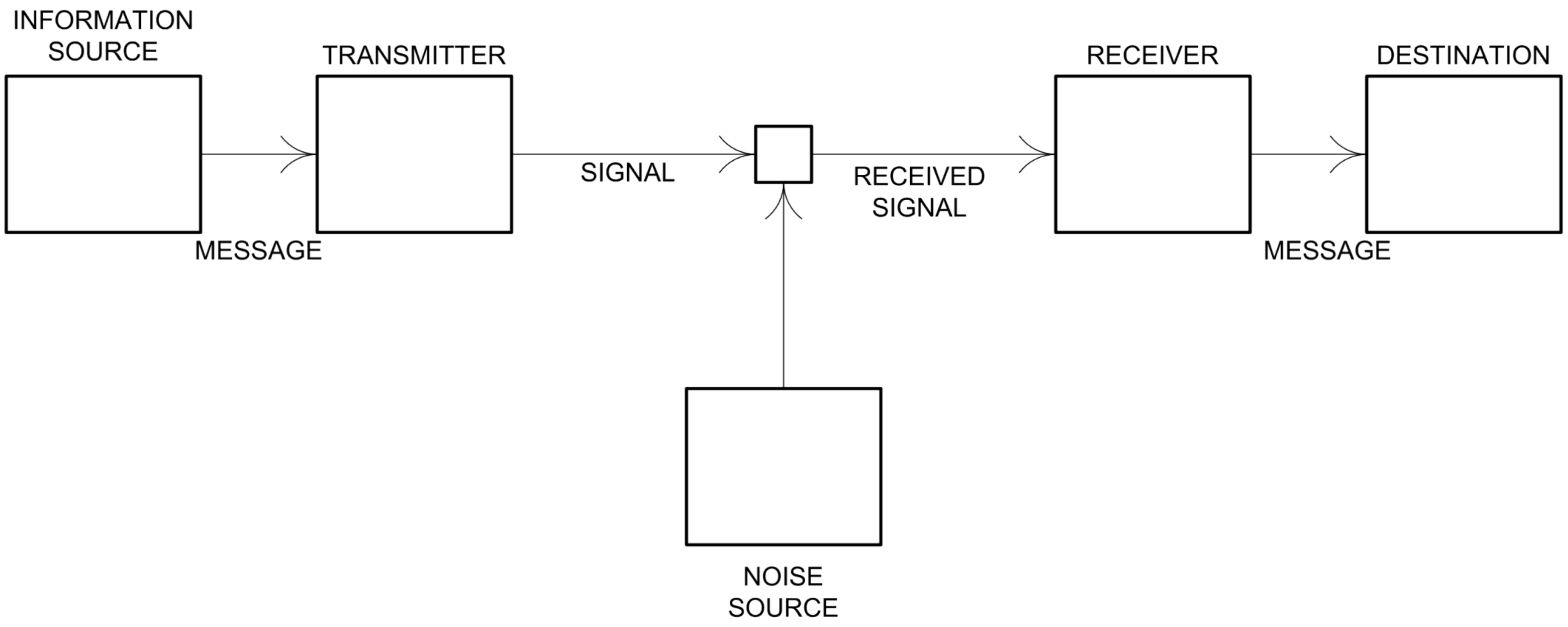
$$F(\alpha,\beta) = K\ln(\frac{\beta}{\alpha}).$$

Referring to Postulate A, we conclude that

$$I(x,y) = K\ln(\frac{p(x|y)}{p(x)}).$$

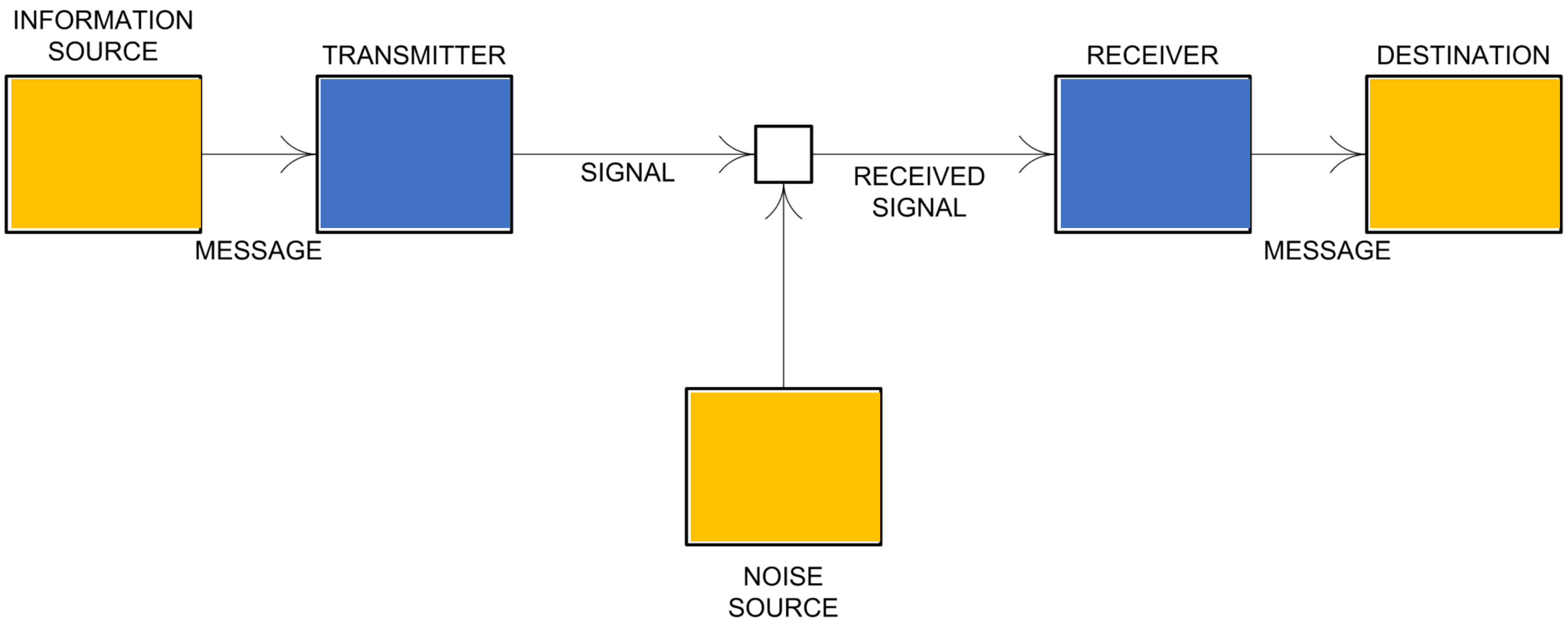
The constant K determines the unit of information. If we set $K = 1$, then I is measured in "nats." It is more common to set $K = \log_2(e) = 1.443$, in which case we say I is measured in "bits" and write

$$I(x,y) = \log_2(\frac{p(x|y)}{p(x)}) \text{ bits}.$$



Claude Shannon’s schematic diagram of a general communication system (1948: Figure 1)

Why is information theory not just applied probability?
What is different from detection and estimation?



Generally thought of as given by nature

Generally up to the design of the engineer

- What is the best that one can do?

- How much can coding help?

Big idea #3:

Coding

Kinds of lossless source codes

- Fixed-to-variable (e.g. Huffman code)
  - Typically want a way to separate codewords without punctuation (unique decodability, e.g. prefix-free)
- Variable-to-fixed (e.g. Tunstall code)
- Variable-to-variable (e.g. concatenation of Tunstall and Huffman)
  - Optimal codes are an open question, whole area largely unstudied

- Fixed-to-(almost)fixed, also called block codes

Consider zero error and arbitrarily small error

Kraft inequality and Shannon-Elias Codes

Block Codes and AEP

Universal source codes

- So far, we assumed that we knew the source distribution in order to design good/optimal codes
- What if we don’t? Learn the probabilities while doing the coding

- Universal source codes, such as Lempel-Ziv (LZ78)

The Beauty of Lempel-Ziv Compression

- https://www.youtube.com/watch?v=RV5aUr8sZD0