

Information Theory for NeuroAI

Study Guide

by Richard J. Young, Ph.D.

richard@deepneuro.ai

Chapter 7 of "The NeuroAI Handbook"

The one currency, the bit, that lets you measure a neuron and a network unit on the same scale.

One-sentence thesis. Information theory turns three fuzzy questions, how surprising, how related, how wrong, into a single unit, the bit, that measures a spiking neuron and a deep network by the same ruler.

PART 1: Surprise, Measured

How Surprised Should You Be?

Start with the feeling, not the formula. "The sun rose this morning" tells you nothing, you already knew it. "I won the lottery" tells you a great deal, because it was so unlikely. Claude Shannon's 1948 insight was that these are the same quantity viewed at different scales: information is the resolution of uncertainty, and a message carries information exactly to the degree it was not expected. Rare events are informative; certain events are not.

Entropy, $H(X)$, is the *average* surprise of a source, how uncertain you are before the outcome arrives. It is measured in **bits**.

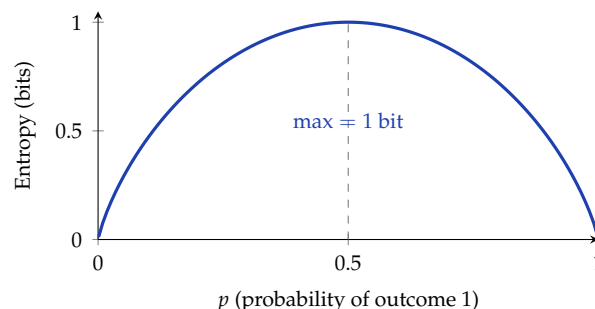
ANALOGY

Entropy is the minimum average number of yes/no questions in a game of 20 Questions. A fair coin needs **1** ("heads?"); a fair 8-sided die needs **3**; an English letter about **4.2**. The fewer questions, the less the source surprises you.

In symbols, sum each outcome's probability against its surprise $\log_2 \frac{1}{p}$:

$$H(X) = - \sum_i p(x_i) \log_2 p(x_i)$$

Two facts do all the work. Entropy is largest when every outcome is equally likely (nothing to predict), and it collapses to zero when one outcome is certain (no surprise possible). A biased coin ($p = 0.9$) has only 0.47 bits; a fair coin has the full 1.



Binary entropy peaks at $p = 0.5$ (maximum uncertainty) and vanishes at $p = 0$ or 1 (certainty). This is why compression works: predictable data carries less than its raw bit-size suggests.

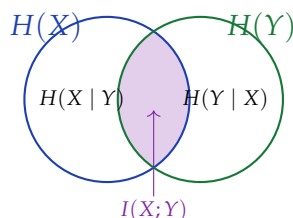
How Much Does One Signal Tell You About Another?

You record a neuron in visual cortex and one in motor cortex. Watching one, how much of your uncertainty about the other disappears? That reduction is **mutual information**, $I(X; Y)$: the knowledge two variables share. Zero means independence, learning X tells you nothing about Y . It is symmetric, and unlike a correlation coefficient it catches nonlinear relationships that a straight-line fit would miss entirely.

The cleanest way to see it is as overlap. Draw one circle for everything X could tell you, $H(X)$, and one for $H(Y)$; their intersection is the shared part.

$$I(X; Y) = H(X) + H(Y) - H(X, Y)$$

Equivalently, $I(X; Y) = H(X) - H(X | Y)$: your starting uncertainty minus what remains after seeing Y .



Mutual information is the overlap of two entropies. The non-overlapping crescents are conditional entropies, what each variable knows that the other does not.

The Cost of Using the Wrong Map

Now suppose reality follows distribution P but you plan using an approximation Q . How much do you pay for the mismatch? **Kullback-Leibler divergence**, $D_{KL}(P \parallel Q)$, is that penalty in bits, the extra surprise you suffer by encoding P 's events with a code built for Q .

$$D_{KL}(P \parallel Q) = \sum_i P(i) \log_2 \frac{P(i)}{Q(i)}$$

It is zero only when $P = Q$, never negative, and,

crucially, **not symmetric**: the cost of a tourist using a local's map differs from a local using the tourist's. **Watch out.** This asymmetry is a modeling choice, not a nuisance, $D_{KL}(P \parallel Q)$ punishes you hard for assigning low probability to events that actually happen.

This is not abstract. **Cross-entropy loss**, used to train nearly every classifier, is KL divergence in disguise: training a network is minimizing the divergence between its predicted distribution and the data's true one. VAEs regularize with a KL term; knowledge distillation transfers a teacher's distribution to a student the same way.

ANALOGY

KL divergence is the tax on a bad weather forecast. If you believe rain comes 10% of the time (Q) but it truly comes 50% (P), you are perpetually caught without an umbrella. The divergence counts, in bits, exactly how much that wrong model costs you on average.

Why the Brain Cares About Bits

A neuron is metabolically expensive; spikes cost energy the brain cannot waste. Horace Barlow's **efficient coding hypothesis** (1961) proposes that sensory systems evolved to represent the natural world at the lowest possible cost, spending bits only where they buy prediction. Two imperatives follow: **reduce redundancy** (do not re-encode what is predictable) and **maximize information** given a fixed bandwidth.

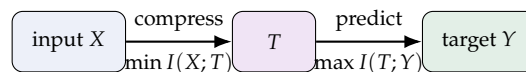
The brain has several tricks for this. **Predictive coding** transmits only the *prediction error*, the part of the input a higher area failed to anticipate, so a fully predicted stimulus travels forward as near silence. **Sparse coding** keeps only a small fraction of neurons active at once, saving energy and easing pattern separation; it is the biological cousin of L1 regularization and

dropout. **Decorrelation** spreads different features across different neurons so their activities stop echoing each other.

There is a hard ceiling on all of this. **Channel capacity** is the maximum reliable rate through a noisy channel, which for a Gaussian channel with signal power P and noise N is:

$$C = \frac{1}{2} \log_2 \left(1 + \frac{P}{N} \right)$$

A single neuron carries only a couple of bits per spike, so the brain beats the ceiling by pooling many noisy neurons into a population code, the same reason the optic nerve's million axons demand ruthless compression upstream.



The information bottleneck (Tishby): squeeze X into a compact code T that throws away everything except what predicts Y . Squeeze too little and you memorize noise; too much and you lose the signal.

PART 4: Information as the Bridge to AI

Same Laws, Silicon or Cells

The point of a common currency is that it lets you put a neuron and a network unit on one balance sheet. Ask a precise question of a spike train, how much does the firing pattern tell you about the stimulus, and it is just mutual information between stimulus s and response r :

$$I(S; R) = \sum_{s,r} p(s, r) \log_2 \frac{p(s, r)}{p(s) p(r)}$$

The same quantity drives modern self-supervised AI. **Contrastive learning** (SimCLR, InfoNCE) trains a model to *maximize* mutual information between two augmented views of one image, forcing it to keep the

invariant essence and discard the incidental, precisely Barlow’s redundancy reduction restated.

Naftali Tishby’s **information bottleneck** frames a deep network as a stack of the squeezes in the PART 3 figure: each layer compresses its input while preserving what the final prediction needs. **Transfer entropy** goes further, measuring *directed* flow, does the past of X help predict the future of Y beyond Y ’s own past, which is how you probe causal information routing in both brain circuits and network layers.

The throughline. Entropy measures surprise, mutual information measures shared structure, KL divergence measures model error, and every one of them applies unchanged whether the substrate is a cortical column or a transformer block. That is the bridge.

Measure	Question it answers	In the brain	In AI
Entropy $H(X)$	How surprising is this source?	Spike-count variability, coding budget	Prediction confidence, compression
Mutual info $I(X; Y)$	How much does X tell about Y ?	Stimulus info in spike trains	Contrastive / self-supervised objectives
KL $D_{KL}(P \ Q)$	How wrong is my model?	Belief update, prior to posterior	Cross-entropy loss, VAE regularizer
Channel capacity	What is the ceiling?	Bits per spike, optic-nerve limit	Layer width, bottleneck design

Quick Self-Test

1. Why does a biased coin ($p = 0.9$) carry less than one bit, while a fair coin carries exactly one?
2. Two variables have a mutual information of zero. What does that tell you, and why is it a stronger statement than “their correlation is zero”?
3. Explain in one sentence why $D_{KL}(P \| Q) \neq D_{KL}(Q \| P)$, and why training a classifier minimizes one of them.
4. What two things does the efficient coding hypothesis say a sensory neuron should do, and how do predictive and sparse coding each achieve one of them?
5. State the information bottleneck objective in words, and name one AI method and one brain phenomenon it describes.

If You Remember One Thing

Entropy (surprise), mutual information (shared knowledge), and KL divergence (the cost of a wrong model) are three questions in one unit, the bit. Because that unit does not care whether the signal comes from a neuron or a network, information theory is the common language in which brains and machines can finally be compared.

