WHAT IS INFORMATION?

CHRIS HILLMAN

1. Introduction

This is, by all accounts, the Age of Information. Most people who know that there is such a thing as a theory of information know, or think they know, that in the classic paper [15] Claude Shannon bequeathed to us the One True Theory of information. But anyone who has carefully read this paper knows that Shannon did no such thing.

It must not be forgotten that Shannon called his theory "a general theory of *communication*", not a theory of *information*. The distinction is crucial. As Shannon put it in [15]:

> The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have *meaning*; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one *selected from a set* of possible messages.

It would be impossible to over-stress the fact that *all aspects of "information" other than statistical phenomena are completely irrelevant to communication theory.* Indeed, "information" appears in this theory only to the degree that the successful communication of information may lead to statistical correlations between the behavior of two systems. Any such correlation must presumably reflect some common causal influence upon these systems, but communication theory is emphatically not a theory of causal influence; nor is it a theory of knowledge or meaning. Rather, it is a purely probabilistic theory concerned with those statistical phenomena which are relevant to the following two fundamental problems [2]:

(1) The problem of data storage. Shannon showed how we can compress data by removing redundancy.
(2) The problem of transmission of messages over noisy communication channels. Shannon showed it is possible to add just enough redundancy to combat the degradation of messages due to noise.

In short, Shannon left wide open the problem of devising a theory to study any aspect of "information" other than purely statistical behavior.

It is my contention that there are, in fact, *many* alternative notions of information worthy of study, quite apart from the semantic aspects of information mentioned

Date: July 7, 1995.
1991 *Mathematics Subject Classification.* Primary 94A17; Secondary 94A15.
Key words and phrases. measures of information, general theory of information.

by Shannon. I make my case for this contention in the second section of this paper. So far the number of theories devised to deal with informational problems other than the problems of communication is surprisingly small. For one example of an information theory quite different from (but clearly analogous to) communication theory, see my paper [8], which describes a theory of the information required to describe a geometric notion. For a theory of information as it relates to knowledge and meaning, see [4].

I further contend that various notions of information in fact provide some fundamental guiding principles of modern mathematics. Indeed, information is intimately involved in our twentieth century understanding of what the basic problems of mathematics are and how we should go about studying them. I argue in favor of this viewpoint in the third section of this paper.

2. A Brief Comparative Review of Entropy Theories

In this section I hope to draw attention to a several particularly striking features of the development of the entropy concept in the mathematical literature during the past fifty years.

2.1. **The Variety of Entropies.** Since 1948 there has been a fantastic proliferation in the number of different quantities called "entropy". Specifically: searching the MathSci disk (1940-1995) [13] yields 7101 records which mention the word "entropy". Even if only 1% of these are actually distinct articles which introduce new quantities called the "entropy" of something, this would give an estimate of about one hundred distinct "entropies" so far defined! In fact, spot checking indicates that the true figure could be well over five hundred— and this estimate does not include the physics, chemistry, or engineering literature.

The astonishing variety of these definitions may be indicated by the following examples:

(1) In ergodic theory, the **probabilistic entropy** of a finite measureable partition $\mathcal{A}$ of a probability measure space $(X, \mathcal{M}, \mu)$ is defined as:

$$H_\mu(\mathcal{A}) = -\sum_{i=1}^{m} \mu(A_i) \log \mu(A_i),$$

Here $\mathcal{A}$ is the partition $X = \cup A_j$ and the sets A_j are the **atoms** of $\mathcal{A}$. For a given atom, the number $\mu(A_j)$ represents the probability of the event $x \in A_j$. The entropy of $\mathcal{A}$ with respect to a measure preserving transformation S on X is

$$h_\mu(S, \mathcal{A}) = \lim_{n\to\infty} (1/n)\, H_\mu(\mathcal{A} \vee S^{\#}\mathcal{A} \vee \ldots S^{\#n-1}\mathcal{A})$$

where $\mathcal{A}\vee\mathcal{B}$ is the **join** of $\mathcal{A}, \mathcal{B}$ and $S^{\#}\mathcal{A}$ is the **pullback** of $\mathcal{A}$. The so-called metric entropy of S, $h_\mu(S)$, is the supremum of the numbers $h_\mu(S, \mathcal{A})$ over the set of all finite partitions. (See [6] for a brief overview of the theory of probabilistic entropy. See [18] for an excellent introduction to ergodic theory.)

(2) In topological dynamics, if X is a compact Hausdorff space, then the **topological entropy** of a finite partition of X into Borel sets, $\mathcal{A}$, can be defined as

$$H(\mathcal{A}) = \log |\mathcal{A}|,$$

where $|\mathcal{A}|$ is the number of sets in the partition. The topological entropy of $\mathcal{A}$ with respect to a continuous mapping S on X, and the topological entropy of S itself, are defined the same way as for measure spaces, using the topological entropy instead of the probabilistic entropy. (See [9][18] for more information on topological entropy.)

(3) In algorithmic information theory, the **algorithmic entropy** of a finite binary string x is defined [14][17] as:

$$H_U(x) = \min_{U(s)=x}|s|,$$

where U is a universal Turing machine, and $U(s) = x$ means "s is a string which forms the code for a program which prints out the binary string x". The entropy of a semi-infinite string can be defined [19] as the limit (when it exists)

$$H(x) = \lim_{n\to\infty}(1/n)\cdot H_U(x_n)$$

where x_n is the finite string formed by truncating everything after the n-th symbol.

(4) The sequence space consisting of all binary sequences, thought of as functions $x : \mathbf{Z} \to \{0,1\}$, can be given a metric defined by $d(x,y) = 2^{-n}$, where n is the smallest $|k|$ such that $x(k) \neq y(k)$. Then a **shift space** X is a closed subspace of this metric space which is invariant under the **shift mapping** S defined by $(Sx)(n) = x(n+1)$ for all $n \in \mathbf{Z}$. (Notice that the shift mapping simply shifts every sequence one place to the right.) A **block** $w \in B_n(X)$ is any sequence w of n consecutive symbols which occurs somewhere in some $x \in X$. Then the entropy of X can be defined [12] as

$$h(X) = \lim_{n\to\infty}(1/n)\cdot \log B_n(X)$$

(5) In statistical mechanics, a **microstate** is a function $\varphi : X \to S$ where X, S are certain finite sets of sizes n, r respectively. The partition of X into preimages $X = \cup_{s\in S}\varphi^{-1}(s)$ is called the **kernel** of φ. (This terminology comes from Universal Algebra.) The entropy of φ is defined [16] as:

$$H(\varphi) = \log\left(\begin{array}{ccccc} & & n & & \\ n_1 & n_2 & \ldots & n_{r-1} & n_r \end{array}\right)$$

where the n_j is the size of the j-th set in the kernel. The multinomial coefficient in this expression gives the number of microstates having the same **macrostate** as φ.

(6) In the theory of complexions [8], if G is a finite group acting on a set X, the **galois entropy** of a subset $A \subset X$ is defined by $H_G(A) = \log[G : \triangleleft A]$, where $\triangleleft A$ is the pointwise stabilizer of A. If G is a finite dimensional Lie group acting on one of its homogeneous spaces X, then the entropy of $A \subset X$ is $H_G(A) = \dim G\cdot A$, where $G\cdot A$ is a certain smooth manifold which represents the variety of effects the action by G on X can have on the set A.

(7) In probability theory, if $(X, \mathcal{M}, \mu)$ is a probability measure space, a **density** on X is an L^1 function φ such that $|\varphi| = 1$ and $\varphi(x) \geq 0$ for all $x \in X$. Then the entropy of φ is defined [11] as

$$H_\mu(\varphi) = -\int_X \varphi\log\varphi d\mu$$

(8) Probabilistic and topological entropies often appear as Hausdorff dimensions [5]; for instance the Equipartition Theorem [2] shows that the probabilistic entropies $h_S(\mathcal{A})$ (in the case where S is ergodic) behave like the Hausdorff dimension of the set of S-typical points in X.

The need for some system for organizing these concepts is great and growing daily, if we are to ameliorate the fact that many people studying one of these concepts knows little or nothing about any of the others.

2.2. **Interrelationships Between Different Entropies.** The fact that this is not an impossible task is demonstrated by the existence of numerous subtle and often very surprising relationships between the known "entropies". For example:

(1) If S is a continuous map on a compact Hausdorff space X, the topological entropy $h(S) = \sup h_\mu(S)$, where the supremum is taken over the compact convex set of **regular Borel S-invariant probability measures** (see [18].
(2) If X is a shift, then $h(X)$ agrees with the topological entropy of $S|X$ [12].
(3) If S is an ergodic measure-preserving transformation on the probability space $(X, \mathcal{M}, \mu)$, and $\mathcal{A}$ is a finite partition of X, then the typical sequences of length n produced by the stochastic message source defined by $(X, \mathcal{M}, \mu, S, \mathcal{A})$ have galois entropies under the natural action by the groups S_n which approach $h_\mu(S, \mathcal{A})$ under an appropriate limiting process [8].
(4) There are conditions under which algorithmic entropies $H(x)$, galois entropies $H_G(x)$, metric entropies $h_\mu(S)$, and the topological entropy $h(S)$ are all defined and must agree. (Parts of this claim are proven in [19] and others in [8].)

Warning! There are also some *non-relations* which are insufficiently appreciated. For instance, the "entropy" of a density is analogous not to probabilistic entropy but rather to a related quantity called **divergence** [2].

2.3. **Specifying One of Many Alternatives.** Entropies typically arise in situations where one has a list of alternatives, and exactly one of these alternative choices is to be specified. Obviously, the fewer the alternatives, the less the variety of choice, and this variety is what the appropriate entropy measures. For example:

(1) The entropies $H_\mu(\mathcal{A})$ and $H(\mathcal{A})$ are probabilistic and combinatorial/topological notions, respectively, of the variety of alternative locations for a point $x \in X$, *relative to the partition* $\mathcal{A}$.
(2) The algorithmic entropy $H_U(x)$ measures the variety of alternative strings of the same length as the minimal program which will produce x. That is, it measures the difficulty of distinguishing this program from all other strings of the same length.
(3) The entropy $H(\varphi)$ of a microstate $\varphi : X \to S$ measures the variety of microstates with the same macrostate as φ.
(4) The galois entropies $H_G(A)$ measure the variety of ways the action by G on X can move the points of the set A [8].

In many situations, there is a *dynamical process* which causes the number of alternatives to grow with some parameter (often time, but sometimes distance from an initial point). In such situations, there are **static** entropies measuring the variety of alternatives at a given stage in this process, and **dynamic** entropies measuring the logarithmic growth rate of the variety of alternatives. For example:

(1) $h_\mu(S,\mathcal{A})$ and $h(S,\mathcal{A})$ are the logarithmic growth rates of the number of alternative "typical" messages of length n, with respect to n, in probabilistic and topological senses, respectively.
(2) If X is a shift space, $h(X)$ is the logarithmic growth rate of the number of alternative blocks of a length n, with respect to n.

Notice that in my notation, I have consistently denoted static entropies by the capital letter H, whereas dynamic entropies are denoted by the lower case letter h. This is the unwritten convention observed in much of the mathematical literature.

2.4. **The Relationship Between Information and Entropy.** Defining entropies of individual objects, mappings, functions, systems, or whatever, is not enough; to obtain a notion of "information" you must also define *joint, conditional, and interaction entropies* of pairs of systems. Here the **joint entropy** $H(\mathcal{A}\vee\mathcal{B})$ of $\mathcal{A},\mathcal{B}$ can be interpreted as the variety of alternatives if we consider $\mathcal{A},\mathcal{B}$ together, the **conditional entropy** $H(\mathcal{A}/\mathcal{B})$ of $\mathcal{A}$ given $\mathcal{B}$ is the variety of alternatives left in $\mathcal{A}$ if we fix one alternative in $\mathcal{B}$, and the **interaction entropy** of $\mathcal{A},\mathcal{B}$ is the reduction in the variety of alternatives in $\mathcal{A}$ if we fix one alternative in $\mathcal{B}$, and also the reduction in the variety of alternatives in $\mathcal{B}$ if we fix one alternative in $\mathcal{A}$. (The fact that these are equal is known as **causal symmetry**.)

Actually any one of these three defines the other two:

(1) Suppose $H(\mathcal{A}/\mathcal{B})$ defines the conditional entropy of $\mathcal{A}$ given $\mathcal{B}$. Then the joint entropy must be $H(\mathcal{A}\vee\mathcal{B}) = H(\mathcal{B})+H(\mathcal{A}/\mathcal{B})$, the amount by which the variety of alternatives increases if we consider not just $\mathcal{B}$ but $\mathcal{A}$ as well, and the interaction entropy must be $I(\mathcal{A},\mathcal{B}) = H(\mathcal{A}) - H(\mathcal{A}/\mathcal{B})$, the difference between the variety of alternatives in $\mathcal{A}$ and the variety of alternatives if we fix one alternative in $\mathcal{B}$; that is, the reduction of variety in $\mathcal{A}$ if we fix one alternative in $\mathcal{B}$.
(2) Suppose $H(\mathcal{A}\vee\mathcal{B})$ defines the joint entropy of $\mathcal{A},\mathcal{B}$. Then the conditional entropy of $\mathcal{A}$ given $\mathcal{B}$ must be the difference $H(\mathcal{A}/\mathcal{B}) = H(\mathcal{A}\vee\mathcal{B}) - H(\mathcal{B})$, the amount by which the variety of alternatives in $\mathcal{A},\mathcal{B}$ together is reduced if we fix one alternative in $\mathcal{B}$, and the interaction entropy must be $I(\mathcal{A},\mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A}\vee\mathcal{B})$.
(3) Suppose $I(\mathcal{A},\mathcal{B})$ defines the interaction entropy of $\mathcal{A},\mathcal{B}$. Then the conditional entropy of $\mathcal{A}$ given $\mathcal{B}$ must be $H(\mathcal{A}/\mathcal{B}) = H(\mathcal{A}) - I(\mathcal{A},\mathcal{B})$ and the joint entropy of $\mathcal{A},\mathcal{B}$ must be $H(\mathcal{A}\vee\mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}) - I(\mathcal{A},\mathcal{B})$.

However, while joint and conditional entropies are often the entropy of some derived object, the interaction entropy is a linear combination of "true" entropies which usually cannot be written as the entropy of a derived object. This phenomenom is clearly evident in the classical theory of the probabilistic entropies $H(\mathcal{A})$, where the conditional entropy

$$\begin{aligned} H(\mathcal{A}/\mathcal{B}) &= -\sum_j\sum_k p(A_j\cap B_k)\log\frac{p(A_j\cap B_k)}{p(B_k)} \\ &= \sum_k p(B_k)H(\mathcal{A}\cap B_k) \end{aligned}$$

is the average over the sets B_k of the partition $\mathcal{B}$ of the entropies $H(\mathcal{A}\cap B_k)$ defined by the partitions of each B_k into sets of form $A_j\cap B_k$, taken with respect to the conditional probability measure on B_k. However, the interaction entropy (often

called "mutual information" in this context)

$$I(\mathcal{A},\mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A}\vee\mathcal{B})$$

(apparently) cannot be written as the entropy of any partition derived from $\mathcal{A},\mathcal{B}$. The same thing happens in the very different case of the galois entropies introduced in [8] (with a qualification in the case of Lie groups).

2.5. **Formal Properties of Entropies.** All of the most important entropies share essentially the same characteristic formal properties. By formal properties I mean those which do not depend upon the mathematical context (e.g. probabilistic, algorithmic, topological) in which the entropy in question is defined. These are the properties which are forced simply by the fact that joint, conditional, and interaction entropies must be interpretable as above.

I show in [7] how to unify a good chunk of the theory of entropy by giving an axiomatic development of such formal properties. For the reader's convenience I will briefly review the basic ideas here.

We must first define the objects on which we will define entropies.

Definition 2.1. *A* **joinset** $(\Omega,\vee)$ *is a set* Ω *of elements, which will always be denoted in this paper by calligraphic letters such as* $\mathcal{A}$*, together with a binary operation called* **join** *and written* $\mathcal{A}\vee\mathcal{B}$*, such that the following properties hold:*

(1) **Associative Law:** $(\mathcal{A}\vee\mathcal{B})\vee\mathcal{C} = \mathcal{A}\vee(\mathcal{B}\vee\mathcal{C})$.
(2) **Commutative Law:** $\mathcal{A}\vee\mathcal{B} = \mathcal{B}\vee\mathcal{A}$.
(3) **Idempotent Law:** $\mathcal{A}\vee\mathcal{A} = \mathcal{A}$.
(4) **Zero Element:** *There is an element* $\mathcal{Z}$ *such that for all* $\mathcal{A}$, $\mathcal{A}\vee\mathcal{Z} = \mathcal{A}$.

We interpret the elements of Ω to be collections of alternatives. If $\mathcal{A},\mathcal{B}\in\Omega$, then $\mathcal{A}\vee\mathcal{B}$ is a "merged" collection representing the alternatives available to $\mathcal{A},\mathcal{B}$ jointly. In this case, fixing one alternative in $\mathcal{A}\vee\mathcal{B}$ allows us to specify unique alternatives in each of $\mathcal{A},\mathcal{B}$; conversely, if we have specified unique alternatives in each of $\mathcal{A},\mathcal{B}$, we have also specified a unique alternative in $\mathcal{A}\vee\mathcal{B}$.

Every joinset is a poset with respect to the partial order $\leq$ defined by declaring that $\mathcal{A}\leq\mathcal{B}$ if $\mathcal{A}\vee\mathcal{B} = \mathcal{B}$. We interpret $\mathcal{A}\leq\mathcal{B}$ to mean that specifying an alternative in $\mathcal{B}$ also specifies a unique alternative in $\mathcal{A}$. We can characterize $\mathcal{A}\vee\mathcal{B}$ as the **least upper bound** of $\mathcal{A},\mathcal{B}$ in this poset. The zero element $\mathcal{Z}$ is the minimal element of the joinset, with respect to the partial order $\leq$.

(See the early chapters of [3] for a very readable and thorough discussion of partial orders and posets.)

Definition 2.2. *Let* $(\Omega,\vee)$ *be a joinset. A function* $H:\Omega\to\mathbf{R}$ *is an* **entropy valuation** *on* $(\Omega,\vee)$ *if the following three properties hold:*

(1) **Positivity Axiom:** *For all* $\mathcal{A}\in\Omega$, $H(\mathcal{A})\geq 0$*, with equality if* $\mathcal{A} = \mathcal{Z}$.
(2) **Monotonicity Axiom:** *If* $\mathcal{A}\leq\mathcal{B}$*, then*
$$H(\mathcal{A}) \leq H(\mathcal{B})$$
(3) **Contractivity Axiom:** *If* $\mathcal{A}\leq\mathcal{B}$*, then for all* $\mathcal{C}$*,*
$$H(\mathcal{B}\vee\mathcal{C}) - H(\mathcal{A}\vee\mathcal{C}) \leq H(\mathcal{B}) - H(\mathcal{A})$$

Given two elements $\mathcal{A},\mathcal{B}\in\Omega$*, the number* $H(\mathcal{A})$ *is called the entropy of* $\mathcal{A}$ *and* $H(\mathcal{A}/\mathcal{B})$ *is called the* **conditional entropy** *of* $\mathcal{A}$ *given* $\mathcal{B}$.

Many of the entropies listed above are in fact entropy valuations:

(1) When Ω is the set of finite partitions of $(X, \mathcal{M}, \mu)$, then $H_\mu(\cdot)$ is an entropy valuation. Moreover, whenever S is a measure-preserving transformation on X, $h_\mu(S, \mathcal{A})$ is also an entropy valuation.
(2) Similarly for topological entropies.
(3) When Ω is the set of partitions of X a finite set, then the physical entropy $H(\varphi)$ defined above is an entropy valuation, provided we replace $\varphi : X \to S$ with its kernel.
(4) The galois entropies defined in [8] are entropy valuations.
(5) Hausdorff dimension in an entropy valuation, provided that we take Ω to be the collection of finite dimensional subsets of a metric space.

Given an entropy valuation $H(\cdot)$, we can define conditional and interaction entropies from the joint entropies as indicated earlier. The following are among the more important formal properties which are automatically satisfied by any entropy valuation:

(1) *Quotient Rule.* $H(\mathcal{A} \vee \mathcal{B}/\mathcal{C}) = H(\mathcal{A}/\mathcal{C}) + H(\mathcal{B}/\mathcal{A} \vee \mathcal{C})$.
(2) *Interaction Identities.*

$$\begin{aligned} I(\mathcal{A}, \mathcal{B}) &= H(\mathcal{A}) + H(\mathcal{B}) - H(\mathcal{A} \vee \mathcal{B}) \\ &= H(\mathcal{A}) - H(\mathcal{A}/\mathcal{B}) \\ &= H(\mathcal{B}) - H(\mathcal{B}/\mathcal{A}) \end{aligned}$$

Notice these are precisely the identities relating $H(\mathcal{A} \vee \mathcal{B})$, $H(\mathcal{A}/\mathcal{B})$, $H(\mathcal{B}/\mathcal{A})$, and $I(\mathcal{A}, \mathcal{B})$ which arose in the previous subsection.
(3) *Order Properties.* If $\mathcal{A} \leq \mathcal{B}$, then $H(\mathcal{A}/\mathcal{C}) \leq H(\mathcal{B}/\mathcal{C})$ and $H(\mathcal{C}/\mathcal{A}) \geq H(\mathcal{C}/\mathcal{B})$.
(4) *Redundancy.* $H(\mathcal{A} \vee \mathcal{B}/\mathcal{A} \vee \mathcal{C}) = H(\mathcal{B}/\mathcal{A} \vee \mathcal{C})$.
(5) *Subadditivity.* $H(\mathcal{A} \vee \mathcal{B}/\mathcal{C}) \leq H(\mathcal{A}/\mathcal{C}) + H(\mathcal{B}/\mathcal{C})$.
(6) *Dependence Relation.* $\mathcal{A}$ is **formally dependent** on $\mathcal{B}$, written $\mathcal{A} \ll \mathcal{B}$, if $H(\mathcal{A}/\mathcal{B}) = 0$. We can interpret $\mathcal{A} \ll \mathcal{B}$ to mean that specifying an alternative in $\mathcal{B}$ essentially specifies a unique alternative in $\mathcal{A}$. This relation $\ll$ is respected by the join operation, in the sense that $\mathcal{A}_1 \ll \mathcal{B}_1$ and $\mathcal{A}_2 \ll \mathcal{B}_2$ implies that $\mathcal{A}_1 \vee \mathcal{A}_2 \ll \mathcal{B}_1 \vee \mathcal{B}_2$.
(7) *Codependency Classes.* $\mathcal{A}, \mathcal{B}$ are **formally codependent**, written $\mathcal{A} \approx \mathcal{B}$, if $\mathcal{A} \ll \mathcal{B}$ and $\mathcal{B} \ll \mathcal{A}$. It is easy to see that $\approx$ is an equivalence relation on elements which is respected by the join operation. We can interpret $\mathcal{A} \approx \mathcal{B}$ to mean that $\mathcal{A}, \mathcal{B}$ contain essentially the same alternatives.
(8) *Class Functions.* Entropy is a class function in the sense that if $\mathcal{A} \approx \mathcal{B}$ then $H(\mathcal{A}) = H(\mathcal{B})$; the conditional entropy, interaction entropy, and entropy distance are also class functions. This is essentially trivial, because formal dependency and codependency have been defined to make this true. It is a remarkable fact, however, that these *formal notions* of dependency and codependency usually agree with *natural notions* in particular cases. For instance, in the case where Ω is the set of finite partitions of a probability measure space $(X, \mathcal{M}, \mu)$, formal codependency agrees with the natural notion which says that $\mathcal{A}, \mathcal{B}$ are "indistinguishable" if there is a bijection between their atoms, say $A_i \leftrightarrow B_i$, such that $\mu(A_i \Delta B_i) = 0$; that is, corresponding atoms differ only by a null set (a set of measure zero). Notice that all such natural notions of dependency and codependency are very much context-dependent, whereas the formal notions are context-free.

(9) *Class Joinset.* Because join respects codependency classes in the sense that $\mathcal{A}_1 \approx \mathcal{A}_2$ and $\mathcal{B}_1 \approx \mathcal{B}_2$ implies $\mathcal{A}_1 \vee \mathcal{B}_1 \approx \mathcal{A}_2 \vee \mathcal{B}_2$, the set of codependency classes forms a **quotient joinset** of Ω. Because $H(\cdot)$ is a class function, it defines an entropy valuation on the joinset of codependency classes. We will call this the **class joinset**. The **entropy distance**

$$D(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}/\mathcal{B}) + H(\mathcal{B}/\mathcal{A})$$

is non-negative, symmetric and satisfies the triangle inequality. Moreover, it is positive definite when considered as a function on codependency classes, so the joinset of codependency classes forms a *metric space*. Because $H(\cdot)$, $H(\cdot/\cdot)$ and $I(\cdot, \cdot)$ are class functions, they give well defined entropy, conditional entropy, and mutual information functions on the class joinset. Thus, we can replace the original joinset with the class joinset without loss of anything essential.

The following three items are among the more important formal properties of the geometry on the class joinset induced by the entropy distance.

(10) *Chain Additivity.* If $\mathcal{A} \ll \mathcal{B} \ll \mathcal{C}$, then $D(\mathcal{A}, \mathcal{C}) = D(\mathcal{A}, \mathcal{B}) + D(\mathcal{B}, \mathcal{C})$. For a picture, see Figure 1a.

(11) *Lambda Property.* We have the identity $D(\mathcal{A}, \mathcal{B}) = D(\mathcal{A}, \mathcal{A} \vee \mathcal{B}) + D(\mathcal{A} \vee \mathcal{B})$. For a picture, see Figure 1b.

(12) *Diamond Lemma.*

Suppose $\mathcal{E} \ll \mathcal{A}, \mathcal{B}$. Then $D(\mathcal{E}, \mathcal{A}) \leq D(\mathcal{B}, \mathcal{A} \vee \mathcal{B})$ and $D(\mathcal{E}, \mathcal{B}) \leq D(\mathcal{A}, \mathcal{A} \vee \mathcal{B})$. Moreover,

$$D(\mathcal{E}, \mathcal{A}) + D(\mathcal{A}, \mathcal{A} \vee \mathcal{B}) = D(\mathcal{E}, \mathcal{B}) + D(\mathcal{B}, \mathcal{A} \vee \mathcal{B})$$

For a picture, see Figure 2. This property is an "additive" analogue of a well-known property of group indices. This is no accident; see [7].

(13) *Dependency Criteria.* The following are equivalent:

- $\mathcal{A} \ll \mathcal{B}$.
- $H(\mathcal{A}/\mathcal{B}) = 0$.
- $H(\mathcal{A} \vee \mathcal{B}) = H(\mathcal{B})$.
- $D(\mathcal{A}, \mathcal{B}) = H(\mathcal{B}/\mathcal{A})$.

(14) *Codependency Criteria.* The following are equivalent:

- $\mathcal{A} \approx \mathcal{B}$.
- $H(\mathcal{A}/\mathcal{B}) = H(\mathcal{B}/\mathcal{A}) = 0$.
- $H(\mathcal{A} \vee \mathcal{B}) = H(\mathcal{A}) = H(\mathcal{B})$.
- $D(\mathcal{A}, \mathcal{B}) = 0$.

(15) *Independence Criteria.* The following are equivalent:

- $I(\mathcal{A}, \mathcal{B}) = 0$.
- $H(\mathcal{A}/\mathcal{B}) = H(\mathcal{A})$.
- $H(\mathcal{B}/\mathcal{A}) = H(\mathcal{B})$.
- $H(\mathcal{A} \vee \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B})$.
- $H(\mathcal{A} \vee \mathcal{B}) = D(\mathcal{A}, \mathcal{B})$.
- $D(\mathcal{A}, \mathcal{B}) = H(\mathcal{A}) + H(\mathcal{B})$.

(16) *Independence Relation.* Since $H(\mathcal{A}/\mathcal{B}) = H(\mathcal{A})$ means that $\mathcal{B}$ gives no information about $\mathcal{A}$ and $H(\mathcal{B}/\mathcal{A}) = H(\mathcal{B})$ means that $\mathcal{A}$ gives no information about $\mathcal{B}$, this suggests defining $\mathcal{A}, \mathcal{B}$ to be **formally independent**, written $\mathcal{A} \,?\, \mathcal{B}$, if $I(\mathcal{A}, \mathcal{B}) = 0$. Note that $\mathcal{A} \,?\, \mathcal{B}$ can be interpreted to mean that specifying an alternative in $\mathcal{A}$ does not reduce the variety of alternatives in $\mathcal{B}$ and

FIGURE 1. Left: $D(\mathcal{A},\mathcal{C}) = D(\mathcal{A},\mathcal{B}) + D(\mathcal{B},\mathcal{C}) = p + q$. Right: $D(\mathcal{E},\mathcal{F}) = D(\mathcal{E},\mathcal{E}\vee\mathcal{F}) + D(\mathcal{E}\vee\mathcal{F},\mathcal{F}) = s + t$.

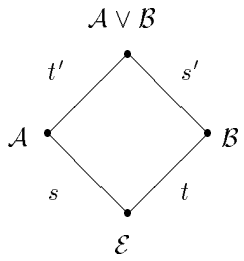

FIGURE 2. The Diamond Lemma says that $s' \leq s$, $t' \leq t$, and $s + t' = t + s'$.

vice versa. Once again, this *formal*, context-free notion of independence often agrees with a *natural* but context-dependent notion of independence. For instance, in the case of probabilistic entropy, $\mathcal{A} ? \mathcal{B}$ iff for all atoms $A \in \mathcal{A}$ and $B \in \mathcal{B}$, $\mu(A \cap B) = \mu(A)\mu(B)$; that is, iff the sets in $\mathcal{A}, \mathcal{B}$ are pairwise independent in the usual sense of probability theory. See [7] for more examples.

(17) *Data Processing Principle.* If $\mathcal{A} \ll \mathcal{B}$, then for all $\mathcal{C}$, $I(\mathcal{A},\mathcal{C}) \leq I(\mathcal{B},\mathcal{C})$.
(18) *Subcancellation Law.* $D(\mathcal{B}\vee\mathcal{A}, \mathcal{C}\vee\mathcal{A}) \leq D(\mathcal{B},\mathcal{C})$.
(19) *Lipschitz Continuity.* Entropy is Lipschitz continuous with respect to the entropy distance, in the sense that $|H(\mathcal{A}) - H(\mathcal{B})| \leq D(\mathcal{A},\mathcal{B})$. Similar formulae hold for conditional and interaction entropies.

2.6. **Entropy as a Numerical Invariant.** Entropy often appears as an *invariant* of the objects in a category.

For example:

(1) The entropies $H_\mu(\mathcal{A})$ are invariants of $\mathcal{A} \in \Omega$ under measure-theoretic automorphisms of $(X, \mathcal{M}, \mu)$. Likewise for $H(\mathcal{A})$ and homeomorphisms.
(2) Similarly, the corresponding dynamical entropies $h_\mu(S)$ and $h(S)$ are invariant under, respectively, measure-theoretic and topological **conjugacies** of transformations $S : X \to X$.
(3) Galois entropies are numerical invariants of subsets of a given set X acted upon by a group G [8].
(4) Hausdorff dimension is invariant under lipeomorphisms [5].

2.7. **Unexplored Territory.** Despite the variety of entropies which have already been defined, there are vast swaths of territory which remain entirely unexplored. I particularly draw attention to the many problems connected with formalizing notions of information and complexity relevant to mathematical biology. For instance: How can we represent information as something which may or may not help a given organism to survive to maturity? How can we quantify the notion of the complexity of a cell? The complexity of an ecosystem? Can we quantify the idea of complexity as an emergent phenomenom?

3. The Role of Information in Modern Mathematics

I contend that notions of information play a fundamental role in formulating the problems of modern mathematics, although this role is not often explicitly recognized. I contend further that taking proper account of the central role of information-theoretical thinking in modern mathematics provides powerful motivation for the formulation of new information theories suited to mathematical problems other than the ones facing Shannon in 1948.

Consider the following problem:

Fundamental Problem 1. *Given category C, how much information is needed to describe the structure of an arbitrary object, X, up to isomorphism? That is, how much information is needed to distinguish a given isomorphism class of objects from the collection of all such classes?*

For example:

(1) *The category of sets:* two sets can be placed in one-to-one correspondence iff their cardinalities agree; therefore, the set-theoretic structure of X is completely described by a single number, $|X|$.
(2) *The category of linear spaces:* the linear structure of X is completely described by a single number, $\dim X$.
(3) *The category of groups:* it suffices to give a presentation by means of generators and relations, but it is known that there is no effective general procedure for finding a minimal presentation. Thus, the problem would appear to be formally unsolvable for this category.
(4) *The category of finite abelian groups:* it is well known that every finite abelian group is isomorphic to a unique group of the form

$$\mathbf{Z}_{n_1} \oplus \mathbf{Z}_{n_2} \ldots \oplus \mathbf{Z}_{n_r}, \text{ where } n_1 | n_2 \ldots | n_r$$

Therefore, it suffices to give the non-negative integers $n_1, n_2, \ldots n_r$.
(5) *The category of finitely generated abelian groups:* it is well known that every such group X is isomorphic to a unique group of form $\mathbf{Z}^n \oplus T$, where T is a finite abelian group (the torsion group of X). Therefore, it suffices to give the non-negative integers $n, n_1, n_2, \ldots n_r$, where $n_1 | n_2 \ldots | n_r$ describes T as above.
(6) *The category of finite dimensional topological manifolds:* the complete solution is known for two-dimensional manifolds, but the three dimensional case is very difficult.
(7) *The category of topological spaces:* this problem is surely very difficult and probably formally unsolvable.

We can draw a number of conclusions from our discussion of these examples:

Lesson 1. *The existence of "mathematical structure" in an object may mean that it can be specified more compactly than would otherwise be the case.*

Lesson 2. *Looking for efficient representations of the structure of an object in some category is essentially the same thing as looking for a "maximally compact" canonical form or a minimal set of invariants, and thus may be regarded as a refinement of the standard problem of classifying the objects of a category up to isomorphism.*

As we saw, Problem 1 may be intractable. In this case, it is common to simplify the problem by transforming the original classification problem into a simpler problem by applying a functor taking the original category into a simpler category whose classification problem has been solved. This is the essential idea behind the linear representations of various categories. The canonical example of such a functor is of course the homology functor from the category of continuous maps between finite dimensional topological manifolds into the category of finitely generated abelian groups. The functorial method works because of the fact that if X, Y are isomorphic in the original category, then their images X^*, Y^* must be isomorphic in the target category; therefore, if X^*, Y^* are *not* isomorphic, neither are X, Y. The functorial method does not completely solve the original problem, but it does provide a powerful tool for telling apart distinct objects in the original category, and thus provides a partial solution whenever it can be applied. We can sum this discussion up as follows:

Lesson 3. *A functor is a device for encoding information about the structure of an object in one category into the structure of an object in a second, simpler, category, while preserving that structure you are most interested in.*

Indeed, a "good" functor is one which picks out precisely that structure relevant to a given problem and "forgets" the irrelevant details. It is a device for stepping back from the trees so that you can see the forest.

Next, consider a related problem:

Fundamental Problem 2. *Given a category C and two objects X, Y in C, how much information is needed to describe a morphism $\varphi : X \to Y$? That is, how much information is needed to distinguish one morphism $\varphi : X \to Y$ from the set of all other such morphisms?*

For example:

(1) *The category of sets:* there is no shortcut: we must name the image of every point in X, because we can always find two distinct set mappings $\varphi, \psi : X \to Y$ which agree on any proper subset $A \subset X$.

(2) *The category of linear spaces:* Given a basis for X and Y, it suffices to describe the image of each basis vector for X in terms of the basis for Y, because two linear maps which agree on a basis must be identical. On the other hand, we cannot get away with giving less information, because we can find always find two distinct linear maps $\varphi, \psi : X \to Y$ which agree on any proper subset of a given basis for X. This gives a complete solution in the category of linear mappings to Problem 2.

Note that the information required consists of $\dim Y$ real numbers for every basic vector of X, and thus of a table of $\dim X \cdot \dim Y$ numbers in all—

in a word, the required information consists of the usual matrix representation of the linear mapping, and the amount of information is most naturally measured in units of "the information required to specify an arbitrary real number".

(3) *The category of groups:* given presentations of the domain and target groups, it is neccessary and sufficient to name the image of each generator in the domain. In this case the essential information is a finite sequence of integers for each generator of the domain.

(4) *The category of finite abelian groups:* in this category there is a well known procedure for finding a minimal generating set, so we can determine the most effective representation of group homs out of any given finite abelian group. This gives a complete solution in the category of finite abelian groups to Problem 2.

Notice that in this case the information is naturally listed by an integer valued matrix, and the amount of information needed can be measured in units of "the information required to specify an arbitrary integer". It is not hard to see how it could also be measured in terms of bits, which would give a theory in which homomorphisms involving small numbers would be easier to specify.

(5) *The category of topological spaces:* it suffices to name the image of every point in some dense subset of X, since two continuous maps $\varphi, \psi : X \to Y$ which agree on a dense subset of X must be identical. However, to my knowledge no-one has considered the question of finding a minimal dense set in a given topological space; this problem is probably formally unsolvable.

We can some more conclusions from our discussion of these examples:

Lesson 4. *The existence of "mathematical structure" in an object X may mean that one can specify a morphism out of X more compactly than one can specify an arbitrary mapping.*

Lesson 5. *Looking for efficient representations of morphisms in a given category will often lead naturally to fundamental concepts for the theory of that category.*

For instance, observe that I mentioned above the concepts of "basis", "generating set", and "dense subset"; these concepts are fundamental to the study of linear spaces, groups, and topological spaces respectively.

Lesson 6. *The appropriate "units of information" may be quite different for two different categories.*

Finally, consider a third problem:

Fundamental Problem 3. *Given an object X in a concrete category C and a list of alternative subobjects, functions, or whatever, associated with X, what is the minimal information needed to specify one of these alternatives?*

The alternatives associated with X might be the subsets of X, or the subobjects of X, or the morphisms from X to a given object, or the real valued functions on X, and so on. The previous two problems are both special cases of this one (there is a standard technique for making a category of whose objects are the morphisms of C and whose morphisms are functors).

Here are just a few examples:

(1) *the category of sets:* to specify a subset A of a given set X, there are in general no shortcuts; you must explicitly name every element of A.
(2) *the category of linear spaces:* to specify a subspace A of a vector space X, it suffices to give a basis for A; conversely, no lesser amount of information will distinguish A from all other subspaces of X. This gives a complete solution of Fundamental Problem 3 in the category of set mappings.
(3) *the category of affine spaces (and convex mappings):* if A happens to be invariant under a specific subgroup H of $\mathrm{Aut}(X)$, the group of affine automorphisms of X, it will have a simpler description than an arbitrary subset. This is true because we need only mention the group H, the fact that A is invariant under H, and then describe a way of picking out A *from among all other sets invariant under* H— and this last is significantly easier than describing how to pick out A from among the collection of all subsets of X!

    For example, if $X = \mathbf{R}^2$ and A is a circle, then A is invariant under a subgroup H of $\mathrm{Aut}\, X = E(2)$ (consisting of all rotations about a given point in $\mathbf{R}^2$, namely the center of A). In this situation, A can be *unambiguously distinguished* from all other circles invariant under H by giving a single non-negative real number, its radius. The group H itself can be described efficiently by giving the coordinates of the center of the circle, so we are simply reinterpreting the obvious fact that a circle can be defined by giving its center and its radius!

Once again we can draw several conclusions:

Lesson 7. *Subobjects can usually be described much more efficiently than arbitrary subsets.*

Lesson 8. *The existence of symmetries in a subset $A \subset X$ usually means that we can describe A much more efficiently than would otherwise be the case, by naming the group of symmetries and then describing how to pick out A from among the other subsets invariant under this group.*

For an elaboration of the idea that more symmetrical subsets are easier to distinguish from their translates (under a given group action) than less symmetrical subsets, see [8].

References

1. J. Aczel and J. Daroczy, *On measures of information and their characterizations.* Mathematics in science and engineering. New York: Academic Press, 1975.
2. Thomas M. Cover and Joy A. Thomas, *Elements of information theory.* Wiley, New York, 1991.
3. B.A. Davey and H.A. Priestly, *Introduction to lattices and order.* Cambridge: Cambridge University Press, 1984.
4. Keith Devlin, *Logic and information.* New York: Cambridge University Press, 1991.
5. Chris Hillman, *What is Hausdorff dimension?*, preprint, 1995. *Note*: all of my currently available preprints may be found at the URL:

    `http://www.math.washington.edu/~hillman/personal.html`
6. —, *An entropy primer*, preprint, 1995.
7. —, *A formal theory of information: statics*, preprint, 1995.
8. —, *Symmetry and information*, forthcoming.
9. Jans Ledet Jensen, "Chaotic dynamical systems with a view towards statistics: a review". In *Networks and chaos : statistical and probabilistic aspects.* Edited by O.E. Barndorff-Nielsen, J.L. Jensen, and W.S Kendall. New York : Chapman & Hall, 1993.

10. Anatole Katok and Boris Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*. Cambridge: Cambridge University Press, 1995.
11. Andrej Lasota and Michael C. Mackey, *Chaos, fractals, and noise: stochastic aspects of dynamics*. New York: Springer, 1994.
12. Douglas Lind and Brian Marcus, *Introduction to Symbolic Dynamics and Coding*, to appear.
13. *MathSci disc* [computer file]. Wellesley Hills, MA : SilverPlatter Information, 1989 and semiannually thereafter.
14. Gregorz Rozenberg and Arto Salomaa, *Cornerstones of undecidability*. New York: Prentice Hall, 1994.
15. C. E. Shannon, "A mathematical theory of communication", in C. E. Shannon and Warren Weaver, *The mathematical theory of communication.* University of Illinois Press, Urbana, 1949.
16. H. C. van Ness, *Understanding thermodynamics.* New York: Dover, 1969.
17. V. V. Vyugin, "Algorithmic entropy (complexity) of finite objects and its application to defining randomness and amount of information", *Selecta Mathematica* (1994) 13 (4), 357–389.
18. Peter Walters, *Introduction to ergodic theory*. New York: Springer, 1981.
19. Homer S. White, "Algorithmic complexity of points in dynamical systems", *Ergodic Th. and Dyn. Systems* (1993) 13, 807–830.

Department of Mathematics, University of Washington, Seattle, Washington 98195
E-mail address: hillman@math.washington.edu