Notes on the "What is Information?" Workshop

Robert L. Constable
Cornell University, Ithaca, NY 14853-7501

Over the course of four and a half days, forty scientists from four countries considered the question, "what is information?" Their backgrounds in biology, chemistry, computer science, information theory, mathematics and physics encouraged interdisciplinary examination; and their common interest in the question stimulated intense discussion. Surrounded by the stark beauty of the desert setting for the Workshop, at Sde Boker in the Negev, the participants were also drawn together in common realization of the special opportunity at hand.[*]

To appreciate the scope of the workshop, consider that the meeting was inspired by a geologist, Professor Arie Issar, who sees messages of information in the primitive but planned flint tools of the early Homo erectus. He and the other participants discussed use of the Bekenstein number of a black hole to bound the potential information content of matter. They looked at how a colony of bacteria communicates as it searches for food, and related its biochemical mechanisms to similar ones used in the embryonic human brain as it grows its network of neurons. They saw how information theoretic techniques for decoding the neural signals of the H1 cell of a fly's eye can also be used in tasks such as discriminating among texts in different natural languages and clustering proteins according to similarity. They looked at writings sent to the meeting by John A. Wheeler from his 1994 book *At Home in the Universe;* these established a framework for understanding how quantum physical reality might arise from qubits.

Professor Bekenstein derived his 1972 method for bounding the possible information content of a material object based on its calculated effect on the surface area of a black hole's event horizon should the object fall into the hole. He brought the participants up to date on results in black hole thermodynamics, discussing Hawking radiation and the idea that information might be left in the *remnants* of an evaporated black hole. He noted that quantum mechanics imposes bounds on information processing, and he showed how to estimate the maximum possible information capacity of a physical object using his measure, noting that a typical paper book can hold at most 10^{37} bits regardless of the scale of storage (characters etched an atom wide or stored in electron spins, wrapped in strings, whatever) and regardless of the technology available to store and read the information.

Bekenstein's method relies on the simple structure of a black hole – a point mass characterized by momentum, charge and angular momentum. The black hole is a limit of the thermodynamic approach to estimating the number of quantum states that a physical system can assume. Ignoring the probability distribution on these states, one defines the system's *entropy* by the logarithm of the number of states (normalized to zero K temperature). Describing a particular state of the system requires a certain number of bits, and that number is a measure of the "possible quantity of physical information" in it (perhaps seen as the content needed to situate the object in a physical theory). Dr. J. Smolin related the concept of information in this sense to the more conventional notions of "information content" by reviewing how accounting for the cost of information processing in computational agents resolves the famous paradox of Maxwell's demon trying to violate the second law of thermodynamics by sorting fast and slow molecules without using energy. Professor Ziv related the physicist's notion of entropy to Shannon's framework for information theory, defining the notion of *Shannon entropy* for a system of sequences of symbols governed by a probability distribution.

Professor Ziv went on to present simple algorithms for approximating Shannon entropy by "parsing" individual strings of characters selected from an ensemble governed by a probability distribution. Parsing generates an efficient code for representing the information in the string as well as approximating the probability distribution; it is the basis of the widely used Lempel-Ziv algorithm for data compression. His

[*] The meeting was proposed by the geologist Professor Arie Issar, who has long considered information as a fourth dimension in his scheme for classifying geological events. Dean Miriam Cohen, dean of the BGU faculty of Natural Science, and Dr. Eitan Bachmat of Computer Science organized the meeting.

methods assign information content to individual strings, and in that way Ziv's approach connects to that of Andrie Kolmogorov and Gregory Chaitin from the 60's, who also assign information content to individual strings, rather than to an ensemble. Dr. Chaitin reviewed his definition of the information in a string as the size of the minimal Turing machine which can generate it. He calls a string *random* if the minimal description is not shorter than the string itself, that is, if the string is "incompressible" using Turing machines. Ziv noted that in this context, his own approach is to use finite state machines in place of Turing machines. Chaitin ended with a discussion of his definition of a single infinite string, omega, which is the halting probability of programs. He noted that nearly every result about the string is unprovable, so facts about it are "true for no reason."

Professor Naftali Tishby suggested another connection between Shannon entropy and information. He said that information is always "about something" and thus is a relation between two entities or processes. He noted that it is possible to present information theory on the fundamental relation, "what does X tell us about Y." By making Y explicit, he takes a step toward a semantic view of information. This approach is adopted in a forthcoming book by David MacKay *Information Theory, Inference & Learning Algorithms*, (Cambridge University Press, 2002). Professor Tishby applied techniques from information theory – the "information bottleneck" method – to show how the H1 cell of a fly's visual system relates horizontal motion in its visual field to neural impulses sent to the ganglia. He isolated four signals being recovered by H1 from the inputs and computed a minimal coding for that information. He also used the method to automatically classify natural language text, say clustering the text into English, German, Italian, French and Russian.

Professor Michael Levitt, a biologist, pointed out how nature takes advantage of the combinatorics of long chains of amino acids. Their state space is astronomical, e.g., a string of 100 acids represents 20^{100} possible sequences. The Shannon entropy is huge. Some of these strings, on the order of a hundred thousand, are naturally occurring proteins on which all life depends, and among them are chains of over 200 amino acids. We already know the structure of thousands of proteins and can figure it out for some others in a few weeks each because they can be crystallized. (But others are very hard to crystallize.) The forces of charge, chemical bonding and so forth cause these chains to fold into geometric structures. Nature uses these *shapes* to carry out all the processes essential to life, such as digesting the lunch the participants had eaten just before his lecture. Digestion is not like "pouring chemicals on the food," because that process would digest our tissues as well. Digestion is an intricate process controlled by information in the form of geometric shapes interacting with molecules. He sees life as arising from a combination of physical laws and information. Professor Levitt noted that one subject in computer science that is critical to understanding life is computational geometry; this observation was confirmed by contributions to biology already made by some of the computational geometers attending the workshop. While the quantitative information potential (entropy) of the space of proteins is huge, its information content is expressed by geometric shapes with only a few of the amino acids being "active chemical sites."

Professor Levitt compared evolution to a large software project: all the programs are written in assembly language without prior design. Natural selection over the course of hundreds of millions of years produces programs that work robustly because faulty ones are discarded. The result is a growing collection of extremely reliable components from which nature assembles new structures. By this means, organisms on earth evolve more and more complexity. The sun's energy is converted into complexity, sucking entropy out of the universe and locally reversing the thermodynamic dispersion of structure into disorder.

Professor Eshel Ben-Jacob dramatically illustrated the way nature reuses structures and mechanisms. He explained the process of chemotaxis that enables a colony of bacteria to locate food. Chemotaxis in bacteria is caused by changes in the combination of forward motion generated when flagella rotate counter-clockwise and their tumbling motion. Bacteria respond to chemicals in the environment to adjust their motion toward food sources and away from toxins. Ben-Jacob noted that chemotaxis is used by the embryonic human brain to weave the growing neurons into a network. We inherit the genes for chemotaxis from the bacteria.

Professor Ben-Jacob is a physicist studying the fundamental life processes in bacteria, including their information processing. By a mixture of physics experiments, video clips, photographs, and overheads, he

made his points. He demonstrated the boundary between organized nonliving structures and the simplest living ones. In a petri dish he created a "singular perturbation" that changed the diffusion process of a drop of food coloring, and he noted that such perturbations in the process of seeking equilibrium (optimization) are common to animate and inanimate matter. For instance, they create snowflakes, and they reorganize a bacterial colony in response to stress. Reorganization allows it to search for food. When a colony is under more severe stress, say starving, it discovers this condition by distributed communication. At a certain point, the colony "decides" to sporulate. That is, each cell wraps its seplicated DNA in a spore, and then dies, leaving the spores to awake when conditions improve, perhaps hundreds of years later. He notes that in a sense the colony is "self aware." (I think it is interesting that distributed computer systems are "aware" in the same sense, and can reach consensus on a course of action by sending messages among processes.)

Ben-Jacob also reported remarkable discoveries from the collaboration of biologist Laura Landweber and computer scientis t Lila Karli in the study of two species of ciliates that appear to compute; they are *Oxytricha, trifallax* and *nova*. These microorganisms have two nuclei; one of them contains the coding DNA (only about two percent of the total DNA) used to make proteins. The other, which is much larger, contains the complete DNA, including the noncoding or so called "junk DNA." When the cell divides, the coding nucleus dissolves and the larger one divides. Then, in the new cell, the complete nucleus constructs a new coding nucleus, throwing out the junk DNA and combining the rest in the proper order. So one nucleus is a program that assembles a coding nucleus capable of making the proteins needed in the current environment. Ben-Jacob speculates that perhaps this computing process can also use information from the environment in which the coding nucleus is assembled. He has discovered that colonies of bacteria can influence the DNA of their foraging parties to elongate their bodies. If most of them are elongated, then their combined chemotaxis action is amplified. Karli has built a computer model that seems to explain the computations of the ciliate nucleus, and she has shown that this model is universal (equivalent to a Turing machine). That model could help explore the question Ben-Jacob raises.

Professor Constable, a computer scientist, examined the progression from data to information to knowledge. He proposed that information is data interpreted by a computing system. Examples include a program executing in a distributed system, a gene expressing itself in a cell, a protein folded by physical forces, and a theory implemented in a theorem proving system. To explain the last example, he noted that interactive theorem provers are guided by user supplied information to bring them into a new state of knowledge. Such provers rely on *logics of information.* The most directly practical of these logics use a *semantics of evidence* to give meaning to propositions and their proofs. Such logics are mechanisms for keeping track of the information content of assertions in order to cause a theorem prover to assert that a proposition is "known" and thus create new knowledge. Computing systems can treat the same object as both data and information, e.g., DNA in ciliates – as data when processed by the complete nucleus and as information in the coding nucleus.

When logics of information are implemented, computers can manipulate the evidence and "animate" theories expressed in the logic. For example, from information coded in the *proof* of an assertion that an object exists, a computer can create a *program* that computes the object. Since their emergence in 1971, these logics have become a useful programming technology. Logics of information will also support Wheeler's interpretation of quantum mechanics as a theory that models reality by relating the evidence from quantum experiments – the "yes" or "no" bits that come from measurement and the resulting decoherence of the quantum state. This is a theory that reconstructs reality from distributed agreement about the meaning of bits – as Wheeler says, the "It from Bit."

Constable noted that when physical theories are implemented, it is possible to automatically extract programs that simulate aspects of reality from proofs of theorems about it. In the context of the physical theory, simulations are knowledge. Smolin discussed the idea that classical computers cannot feasibly simulate quantum reality, and that quantum computers are needed to do that. But Smolin noted that so far, even in principle, quantum algorithms extend classical ones only for a limited class of problems, such as factoring numbers. Professor Krishnamurthy also made this point in his lecture on quantum computing, and he noted that computing based on quantum field theory was limited; for example, he believes that quantum field computers cannot compute Ackermann's function nor take fixed points.

Professor Issar's wish to express within only one set of coordinates the evolution of the form of the hominids (i.e. presented on space-time dimensions) and the evolution of their intelligence (i.e. ability to produce tools of increased sophistication) brought him to the suggestion to add to space-time an additional dimension, which he defined as "information." He suggests that this is the dimension along which all cognition processes and intelligence structures are taking place. Thus while space is a dimension which is measured by a foot or a meter, and time by a pulse or a clock, information is a dimension measurable by a brain or a computer.

The workshop included other lectures on information theory, game theory, quantum computing and so forth. Professor Ayval Ramati presented an evening lecture about the views of Newton and Leibniz on information. A complete program is available at the Ben Gu rion University Computer Science Department web site. A number of topics of this Workshop are also touched on in the book *The Bit and the Pendulum,* by Tom Siegfried (John Wiley & Sons, NY, 2000). Siegfried's work brings him into contact with many top scientists every year, and in the preface he says, "At the foundations of both biological and physical science, specialists today are construing their research in terms of information and information processing." Discussions and lectures at the Workshop confirm that the word "information" in its several related meanings draws scientists together, in a global information optimization process. At some point, there may be a "singular perturbation" that brings into being a new science of information that will inform physics, chemistry, biology, neuroscience, communication and computer science. This Workshop may have moved the research community toward such a creative perturbation. There are plans to publish a proceedings and to continue the meeting on a biannual basis.