

**Computer Learning and the Scientific Method:*
A Proposed Solution to the Information Theoretical
Problem of Meaning**

LEONARD ORNSTEIN, Ph.D.

New York, N. Y.

ABSTRACT

This discussion outlines and implements the theory of an inductive inference technique that automatically discovers classes among large numbers of input patterns, generates operational definitions of class membership with explicit levels of confidence, creates a continuously updated "self-organized" coded hierarchical taxonomic classification of patterns, and recognizes to which already discovered class or classes, if any, a new input belongs in an information-theoretically efficient way. Relationships to the "scientific method" and learning are discussed.

Learning processes which provide the individual and his species with an increased potential to cope with his material, biological and social environment have what evolutionists call "adaptive value" (e.g., see Dobzhansky (1)). Measured on such a scale, the most important aspect of such learning processes is the development of implicit or explicit techniques to accurately estimate the probabilities of future events. And, since the "calculus" of future events from analysis of past experience is the forte of science, it seems reasonable to turn to the methods of science for some useful clues when searching for efficient learning techniques. Modern scientists, and physical scientists in particular, who work regularly with the scientific method and operationalism (2), are keenly aware of some very efficient and sophisticated techniques for learning more about the universe. But if we ask of these who, it seems, should best understand the learning process, "How shall we build—or indeed, is it possible to build—a machine that can learn and think?", the replies are varied (3-8), and, as is indicated in a recent knowledgeable review by Selfridge (9), so far have been relatively unproductive.

In this discussion, I hope to provide a basis for establishing considerably greater confidence in the broad potentials for machine learning via automated inductive processes. This will be attempted by sketching out, in operational terms and against a background of common experience, many theoretical and a few of the practical problems.

* This work represents a revision of "Pattern recognition, morphology and the generation of hypotheses", presented at the Symposium on Machine Methods in Biology, A.A.A.S. convention, New York, 1960, in combination with a revision of part of the final report on P.H.S. Contract #SA-43-ph-3096.

From the Division of Cell Biology and the Cell Research Laboratory of the Department of Pathology of the Mount Sinai Hospital, New York, N. Y.

These will include problems concerning concepts such as “gestalt,” “correspondence of attributes,” “similarity,” “natural groupings,” “classification,” “operational definition,” “discovery,” “hypothesis,” “theory,” “measure of confidence,” “experiment,” and “causality,” all of which I believe can serve useful roles in increasing understanding of learning by inductive processes. Such practical problems as those relevant to the choice of largely “parallel” as opposed to “sequential” computers for implementing such processes will also be briefly examined.

Some solutions to those problems which appear pivotal will be developed and these will be incorporated into a simple set of instructions for an adaptive pattern recognition program for a digital computer which serves as a workable model of the inductive process. This model will be evaluated with some novel information theoretical measures which permit us to maximize the “naturalness” of the “discovered” classes. Whereas the reader should have little difficulty in assimilating much of what is presented in most other sections of this paper, some difficulty with the development of this particular evaluation may be experienced by those who lack some familiarity with Information Theory.

Weaver (10) has defined three levels of Information Theory (also called the Theory of Communication); the Technical Level, the Semantic Level, and the Influential Level. Information Theory has, so far, dealt successfully with some technical problems concerning the processing of information, i.e., problems related to the question, “With what accuracy and efficiency can symbols of communication be transmitted?” and has dealt less successfully with others related to such questions as, “What practical devices can permit both the most accurate and cheapest communication of such symbols?” Until now, such semantic problems as those raised by the questions, “What are the meanings of these symbols?” and “How can the meanings be accurately and economically communicated?” as well as problems of intention raised by such questions as, “How can we most effectively insure that a message will fulfil the purpose of our communication?” have remained as they were when first examined in this context by Weaver, largely untouched. In developing answers to these latter questions, the “information” in Information Theory will begin to have a meaning much closer to the information of every day usage and interest.

Our model will be shown to provide a useful bridge between the technical and semantic levels of Information Theory because it “discovers” “meanings,” it provides a useful approximation to the “ideal coder and decoder,” the possibility of which is predicted by Shannon's Binary Coding Theorem (10), and it seems to begin to answer, at both the technical and semantic levels, the question of how to accurately and economically communicate symbols and meanings. (This model will also demonstrate that the problem of meaning is no longer “irrelevant to the engineering (technical) problem” (see Shannon (10) and Brillouin (11)) .) Connections to the Influential Level of Information Theory will be indicated in our discussion of the concept of “experiment.” Some clues to possible

solutions to most other problems discussed will also be briefly sketched in.

The widespread application of computers to the solution of human problems will remove the individual citizen further and further from crucial decision making processes. Before the solution of such problems is relegated, perhaps irrevocably, to an impersonal but “superior” mechanical intelligence (as some believe is ultimately both desirable and, in any case, inevitable), it will be valuable to be able to convey some understanding of the mechanical decision making process to the average intelligent person. This may be necessary if this new technology is to receive his full support and if those fears which stem from ignorance of the methods of the machine (and the methods of science) are to be lessened. His attention must also be brought to focus on the need for solutions to the newly created problems of “cultural adaptation” of man and machine to one another.

Because this model hews closely to ordinary intuition in many areas of its development, it may also be able to serve such a general educational need. Though this may be of some value in the short run in so far as it might catalyze interest in, and support for, needed technological and social research and development, in the long run, only proven utility of particular machine approaches to inductive processes can provide a satisfactory basis for enduring confidence and support. Hopefully, some elements of this analysis will survive the test of use.

LEARNING AND THE MORPHOLOGICAL SCIENCES

In addition to the elements of the classical scientific method (which will be analyzed in some detail further on), science may have developed some other useful keys to efficient learning. Where in science may we hope to find these keys?

In order to help in developing an answer to this question, let us first examine our personal experiences of learning processes. These can be conveniently resolved into two (not necessarily mutually exclusive) kinds:

a) Learning associated with the raw sensory experiences of the universe*;
and

b) Learning associated with those distillations of the experiences of other human beings which are transmitted by means of social communication.

The wisdom (or folly) of others (which can not be used until experience with a

* Such learning dominates the period from birth to the beginnings of an understanding of language and other social codes (pointing, grimacing and smiling, etc.). Of course, during this period as well as later on, learning is largely dependent upon whatever minimum of heritable “distillations of the experience of the species” are already built or programmed into the nervous system at birth. Such evolutionary “wisdom” is imparted through the particular structure of the nervous system and sense organs and involves both “special purpose” “instinctive” patterns of response to external and internal stimuli as well as more “general purpose” primitive “logical processing routines.” The limits as well as the adaptive value of such “wisdom” are highly dependent upon the built-in bounds to the sensitivity, dynamic range, and kind and magnitude of the “field of view” of each of the sense organs.

minimum of input of raw data has provided rudimentary techniques for handling complex inputs like language) is all ultimately derived from raw environmental experience, even though some of it was gathered back in prehistoric times. This might well lead us to believe that hidden away in the primitive early interactions of the newborn child and his “chaotic” environment lie clues to the fundamentals of the learning process. Unfortunately, direct access to the memories of this period of our lives seems least available to us as individuals, and this kind of interaction is only beginning to be unravelled by careful observation of learning in very young children (12-15).

But among the sciences, the classical morphological sciences have a similar relation to, for example, physics or genetics, as the infant has to the adult. The morphologist is largely concerned with ordering the raw sensory stimuli presented to him by the objects which he studies. The practitioners of the “more advanced” sciences are usually concerned with higher levels of order, i.e., the structuring (in “metalanguages”) of more complex relationships of raw sensory stimuli of the present and past with one another and with such previously recognized order as physical and biological “Laws.” As pointed out by the mathematician-physicist, H. Weyl, (16), “ . . . *the formation of concepts and theories by science . . . is preceded by ordering and classification.* Perhaps more stress should have been laid on this preliminary stage, that still plays a major role in biology while it has become of subordinate importance in physics. The spectacle of the immense variety of plant and animal species displayed by nature has been an early and persistent stimulus for biology to develop to great perfection the *art* of morphological and taxonomic classification. The remarkable fact that the diverse species, notwithstanding the range of variation, mostly exhibit clearly recognizable typical differences, has facilitated the task. *The typical may be elusive in terms of well-defined concepts, and yet we handle it with instinctive certitude, e.g., in recognizing persons.* Nor is it easy to describe in general terms how the process of classification, step by step and *ever more convincingly*, succeeds in separating essential from unessential features” (italics mine) .

Whereas the morphologist may not necessarily be able to present us with the plans for a learning and thinking machine, Weyl’s comment suggests that a serious effort to unravel and explicate the foundations of the taxonomic-morphologic approach may be rewarding. It is my intention to demonstrate that it is relatively “easy to describe in general terms how” a “process of classification, step by step and ever more convincingly, succeeds in separating essential from unessential features,” and that therein lies the foundation of the inductive learning process.

If we ask a herpetologist how he has learned to distinguish one species of frog from another—for example the leopard frog from the pickerel frog—he can hand us a taxonomic key to the Anura of North America, in which some characteristic attributes of each taxonomic group, from Order down to subspecies are collected—all in hierarchical rank (e.g., (17)) . Given a specimen of each of these species—and the key—a novice would have a fair probability of properly identifying these specimens.

Yet appearances are deceiving. The experienced herpetologist himself rarely depends completely on the particular attributes in the key to identify these frogs. In the field—at 40 feet, with only 1/4 of the animal exposed to view—he may easily identify members of these closely related species with a very high degree of confidence. Any two herpetologists are likely to perform about as well in such a circumstance, yet nowhere in herpetological literature are we likely to find a complete set of rules to which each would independently subscribe as the basis for his particular method of identification. Two pathologists studying the same poorly prepared frozen sections of a surgical biopsy of a rare disease may each be able to arrive at an accurate diagnosis. Yet, each, if he had to rely



FIG. 1. Mr. X

wholly on a word description of what the other had seen, would be extremely hesitant to make a diagnosis.

In short, morphologists have canonized some sets of descriptive attributes of the objects which they study. These are not necessarily complete descriptions, and often are not the sets of attributes which they, as experts, actually use. Furthermore, when asked, they are hard put to explicate the basis of their own personal expertise at classification.

To gain some insight into this dilemma, let us examine Weyl's example of the problem of recognizing persons: Most readers will have seen the person in Figure 1 for the first time on reading this article. If, after setting this article down, you were handed a set of one hundred different black and white photographs of caucasian male New Yorkers—including another picture of Mr. X—most of you would probably experience very little difficulty in identifying his photograph from memory. If, on the other hand, after closing these pages, a large group of you readers were all to cooperatively write up a description

of Mr. X which was to be used by another group as the sole basis for identifying his photograph from among the hundred—the frequency of correct identifications would be much much lower. In both tasks you would start with the same mental image or “gestalt.” A new image can usually be compared directly with the “gestalt” with small error. To decode the “gestalt” and translate it into words—to then visualize a new mental image from the word picture—and finally, to compare this mental image to the photograph, involves a large loss of information as well as the generation of considerable error.

A TRANSLATION PROBLEM

Human beings simply do not know how to formulate adequate word pictures, i.e., how to accurately translate the mental record of an optical or acoustic image into the digital symbols we call words (one can recognize a familiar voice almost as readily as a familiar face). That this is a problem of translation seems clear: A caricaturist would be able to translate his mental image of Mr. X's face into a sketch (an analogue output) which many of you would be able to use successfully to select his picture from the set, and a good mimic would do as well in the auditory realm with still another kind of analogue output. However, the “artist's” word picture of face or voice would be much less adequate.

If a key to the learning process resides in the techniques of taxonomic morphology, it should not be too surprising that it has remained hidden from general view. For the morphologists, no less than the rest of us, have marginal verbal access to their mental images, and the taxonomies they have generated are necessarily quite primitive and seem to be more a product of “art,” as Weyl put it, than of science. Their pattern recognition techniques might therefore not have been expected to inspire widespread confidence in their potentials for the construction of a satisfactory model of the learning process.

Cell biologists have been very much concerned with this problem of defining patterns—either explicitly or implicitly, because in one way or another, all their work is related back to the two-dimensional optical images of cells which they see under various kinds of microscopes. During the last 25 years, in order to explore the clearcut early evidence of a relationship between RNA and protein synthesis, and between DNA and the hereditary message, as well as to provide a way down from the growing Tower of Babel on which our two expert pathologists find themselves when discussing their frozen section, cytologists began to try to convert their microscopic observations into the universal language of numbers. In particular, a considerable and successful effort has been expended to quantitate the information about the distribution and concentrations of the various substances of which cells are composed. Microspectrophotometric techniques were devised for measuring the amount of light absorbed as a result of the natural ultraviolet absorption—or the absorption of visible wavelengths by *in situ* colored products produced in parts of cells as the result of specific chemical test reactions (see, for example (18)). It was largely in this particular resurgence of cell biology that I first

cut my scientific teeth. It is from this perspective that I first began to appreciate how central a position pattern recognition, descriptive morphology, and taxonomy hold in relation to an understanding of the general problem of learning.

DISC ELECTROPHORESIS AND A NEW LEAD

A few years ago, as a by-product of our cellular studies, Dr. B. J. Davis and I developed a new, reproducible, high-resolution electrophoretic technique which we call "Disc Electrophoresis" (19, 20). This technique permits the identification of over twenty proteins—and measurement of their concentrations in a sample of as little as 1 microliter of blood serum. Figure 2 shows a photograph of such a run. Position "identifies" the protein; optical density provides a measure of its concentration.

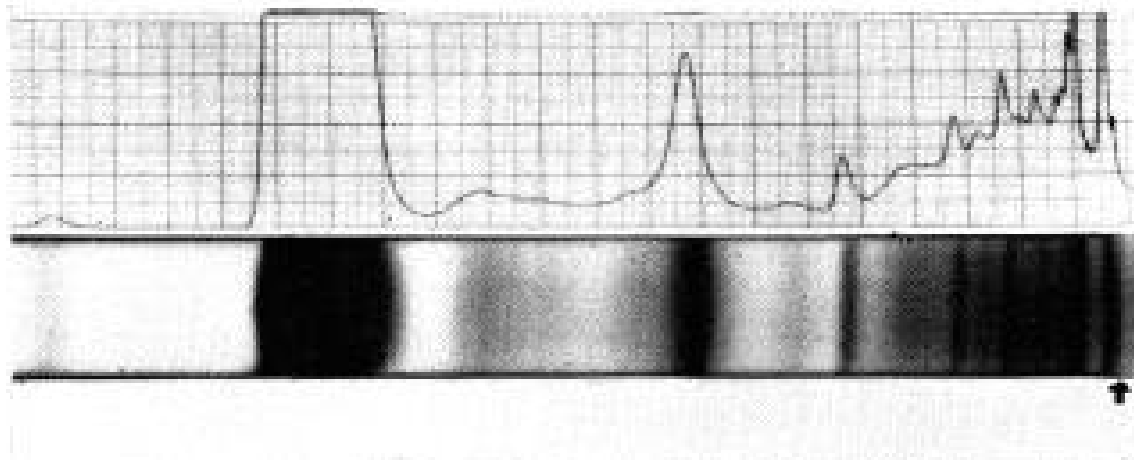


FIG. 2. Proteins of a three microliter sample of human serum, haptoglobin type 2-2. Origin indicated by arrow. Densitometric trace of the sample: abscissa, position; ordinate, optical density (2 *O.D.* full scale). Same run as right hand specimen in Figure 5.

Of the many potentials of this technique, the value, as a clinical diagnostic tool, of this quick quantitative check on over twenty complex substances per sample, seemed obvious, and we have been applying ourselves to the realization of this potential.

The type of analysis which the cytophotometrist performs on optical images of cells is a problem in the recognition of two-dimensional patterns. This new universe of electrophoretic patterns presents us with the one-dimensional homologue of this much more difficult two-dimensional problem. It therefore has provided us with both an opportunity and a compulsion to wade into the problem in both a serious, and, I believe, hopeful way.

In the near future, we expect to be able to collect upwards of 10,000 specimens per week—and to have the data recorded as the logarithms of the optical densities at about 400 positions per sample. These logarithms will be recorded on tape in 7 bit binary code, providing the equivalent of a constant accuracy of about 5% of concentration over a range of concentrations from about 6% (by wt.) of a protein in the blood serum—down to about

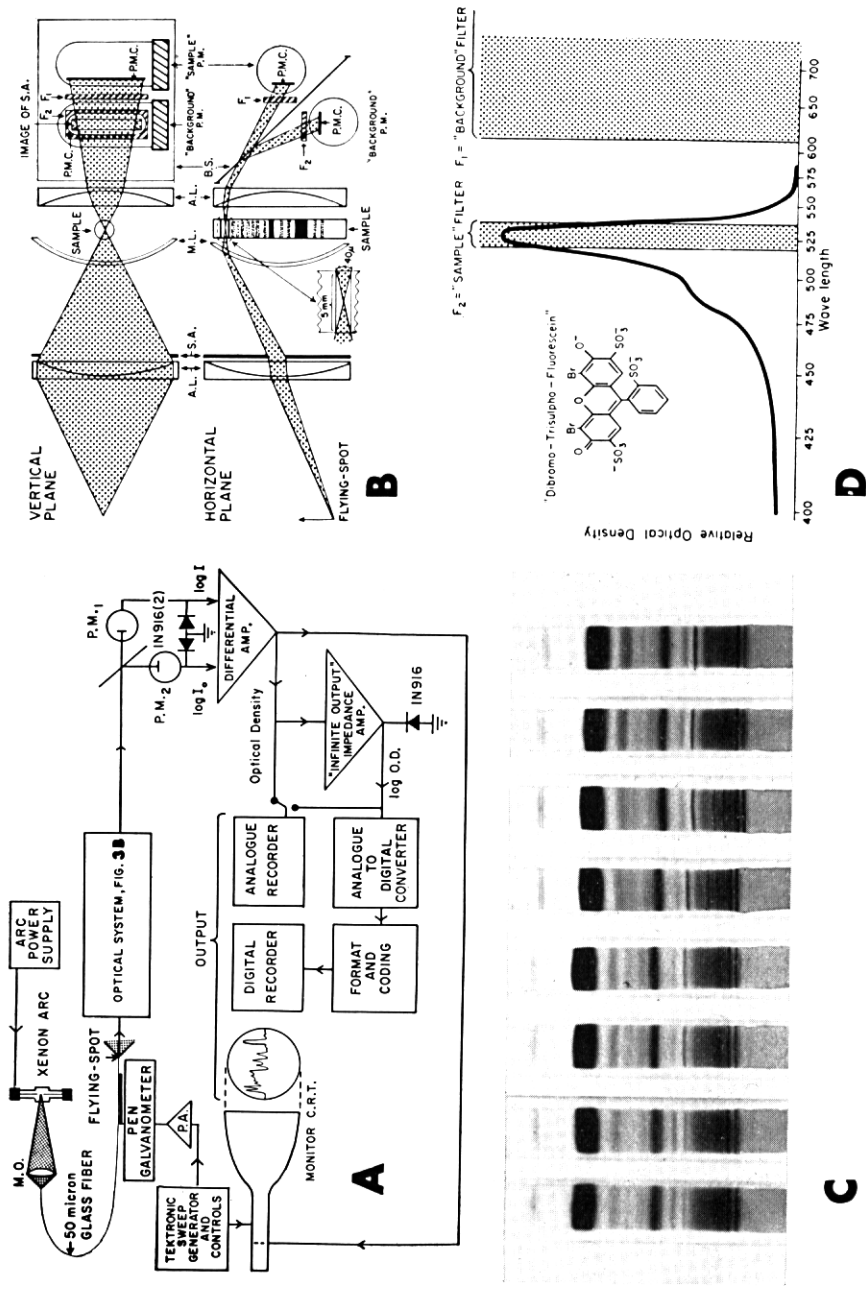


FIG. 3.

0.006% (see Fig. 3). Without a doubt, such data will contain much useful information. The major problem was how to rapidly and efficiently process these data.

We have before us a new, unexplored, enormous but finite universe consisting of the blood sera of all living humans, each consisting of up to some 200 *resolvable* proteins in a bounded concentration range. We have a “sense organ,” the photometric apparatus, which can identify 400 positions in the pattern and measure concentration throughout the range with an essentially fixed resolution. The information about each sample is available as a sequence of 400 7 bit numbers.

How shall we learn from observing this universe?

THE MORPHOLOGICAL APPROACH

The morphological approach suggests that we group all the electrophoretic samples in a comprehensive class; list the attributes which all members of the class share in common; subdivide the main class into subgroups which contain members more similar to one another within each subgroup than to members of other subgroups; list the additional attributes which all members of a subgroup share in common—and continue subdividing subgroups and describing them until the subgroups cease to be easily subdivided. In this way, a taxonomic key would be generated.

Hopefully, the terminal subgroups would contain as members, the samples of sera of individuals who—on inspection, will also be found to share special physiological states in common. Many of the terminal subgroups may “define” certain diseases; some should result in the discovery or resolution of new disease entities within previously poorly defined and understood “disease groups”; and some should discover different normal genetic types. On the other hand, if the comprehensive class contains no “natural” subclasses, or if the “true” values of the data are obscured by excessive noise, for example, by error in measurement, the subgroups will not be useful.

Historically, clearly defined subgroups have generally been discovered before explicit intermediate classes have been well defined—and the taxonomic trees of the

FIG. 3A. Block diagram of a mechanical flying-spot scanner. I_0 , background signal; I , sample signal; $O.D.$, optical density; IN916, “logarithmic diodes.” The flying-spot is generated in the following manner: A very high brightness xenon arc is imaged by a microscope objective, $M.O.$, of numerical aperture (N.A.) 0.3, onto the end of a clad, 50 micron diameter, 20 inch long glass fiber of N.A. 0.6. This end is locked in position. The other end of the fiber is carried on the “pen” of a high-speed, rectilinear pen-galvanometer driven through a power amplifier, $PA.$, by the oscilloscope sweep generator.

FIG. 3B Optical system of scanner: $A.L.$, achromatic lens; $SA.$, slot-shaped aperture; $M.L.$, meniscus lens; $BS.$, beam-splitter; $P.M.$, photomultiplier; $P.M.C.$, photomultiplier photocathode; F_1 and F_2 , filters as indicated in 3D.

FIG. 3C Pairs of Disc Electrophoresis runs of the blood sera of four different individuals showing normal genetic variability.

FIG 3D The absorption spectrum of the anion of a dye that binds strongly to the cationic groups of acid-denatured protein. By measuring the “background” intensity, I_0 , at wavelengths longer than 610 m μ and the intensity transmitted by the sample, I , with a narrow spectral band of green light, a reproducible “double-beam” technique is easily instrumented. Substitution of a narrow band violet filter for F_2 permits measurement of discs of protein that have excessively high $O.D.$ at the absorption peak.

morphological sciences usually have been constructed starting with the specific, and working down to the more general comprehensive class. But for the infant, learning probably starts with general classes within which he discovers special subclasses (e.g., see (4) and (21)). Can we, like the child, in our efforts to learn from observing this new universe, start at the trunk of a tree and work up to the branches? (See Turing's discussion of the "child machine" (3) .)

EFFICIENCY AND TAXONOMIC CLASSIFICATION

Classification is, in fact, a naming procedure. It involves the subdivision of a population of objects, patterns, etc., into subgroups such that individual members of each subgroup are, on the basis of comparison of a set of corresponding attributes, more "similar" to one another than to members of other subgroups. If there are Q unrelated, equally frequent and dissimilar subgroups in a classification, then it will take an average of $Q/2$ comparisons of an unidentified sample to match it to its proper subgroup and "name" it. The names (or code) of unrelated dissimilar subgroups constitute what is called a "nonsignificant code," that is, the code in no way indicates any relationship or degree of "similarity" of one subgroup to another (22).

When a classification takes the form of a taxonomic tree, members of subgroups (branchlets) of the same branch will, in general, be more "similar" to one another than to members of other branches. If there are, on an average, x levels of branching and y branches at each branch point, then $Q = y^x$, but it will take only an average of something like $xy/2$ comparisons of an unidentified object to match it to its proper subgroup and "name" it. (If $y = 2$ and $x = 100$, the identification of a sample by a taxonomic procedure would involve only 100 matching steps as compared to an average of $2^{100}/2$ steps for the same number of nonsignificant pigeon-holes!) If the code for the name of each subgroup is derived from the "numbering" of the branches of the tree on which the subgroup is located, the code will be a "significant" code (22) and simple inspection of this "name" will indicate inter-group similarities. We are interested in the generation of a taxonomic classification and a "significant" code, (especially because a language which uses such a code may also be able to more efficiently convey "meanings" (see page 472)).

CORRESPONDENCE, IDENTITY AND SIMILARITY IN ONE AND TWO DIMENSIONS

In order to usefully compare patterns, it is necessary to design some measure or measures of the "similarity" of one pattern to another. All measures of similarity hinge on definitions of identity. Two objects (patterns, collections, properties, attributes, attribute sets, etc.) are said to be identical if, when "placed in a one to one correspondence," all corresponding parts are equal. For two-dimensional line figures—e.g., triangles, circles, etc.—two objects (occupying different positions in time-space) are said to be identical if, by a combination of translation and rotation they can be made "congruent". For two

dimensional patterns with varying optical densities, the patterns are said to be identical if for every point in one pattern there is a “geometrically corresponding” point in the other and vice-versa, and the optical density of each point is equal to that of its “corresponding” point. An equivalent of the “congruence test” (which provides an operational means for locating “corresponding” points) in this case, involves preparing a “perfect negative” (in the photographic sense*) of one pattern, and, by rotation and translation, the negative and the second pattern are superimposed and “searched” until the minimum amount of light is transmitted through the total area of the superimposed patterns. If the measurements, at this minimum, of optical densities at all points of the superimposed patterns are equal to one another, and in addition, equal to the sum of the optical densities of the point of maximum density and the point of minimum density of the second pattern, then the patterns are identical and the points of the superimposed negative and positive have been placed “in correspondence” at this minimum.† In a similar manner, the concept of identity might be extended to three (or higher) dimensional patterns or objects.

SIMILARITY AS A DISTANCE BETWEEN PATTERNS

Whereas this definition of identity may be quite unambiguous, this is not the case for “similarity.” It is helpful to represent the description of an object by a point in an n -dimensional space (hyper-space), where each attribute of an object (or position in a pattern) is considered to represent a separate “dimension” of the object, and the magnitude of each attribute represents its coordinate (distance from its appropriate coordinate axis) in that dimension of a hyper-space (e.g., see 25, 26, 27). When the problem of recognizing which attribute of one pattern corresponds to that of another is straightforward (as in the case of our one-dimensional electrophoretic patterns—we know the location of the “origin”—and for things such as lists of clinical data where each attribute is identified —e.g., white blood cell count, blood sugar, etc.) then it is easy to specify the total set of coordinates for each point in the particular hyper-space for the members of any group of objects with a corresponding set of attributes. If two objects (or patterns) are identical, they will have the same coordinates (i.e., they will occupy the same point in the particular hyperspace). If the hyper-space is an ordinary Euclidean

* A “perfect negative” is defined as having an optical density for the point corresponding to the point of minimum density in the original, equal to that of the point of maximum optical density in the original, and each other point of the “negative” will have an optical density equal to that maximum optical density minus the optical density of the corresponding point in the original.

† It is of some interest that there is strong evidence that a simple diffusional “random search” permits the “discovery” of the appropriate positions of point to point correspondence between the extremely long (10^4 to 10^7 bit (23)) messages of the complementary “one-dimensional” strands of *DNA* both in the *in vivo* pairing processes that occur in meiosis and somatic pairing as well as in non-living test-tube “nucleic acid hybridization” experiments (24). These experimental results suggest that a random search, using a measure of similarity (see next section) as a figure of merit, might be a surprisingly efficient device, especially if the search were begun at low resolution and finished up at high resolution.

metric space, the “distance” between the points in this space will be defined by the square root of the sum of the squares of the differences between the magnitudes of the corresponding attributes of a pair of objects (the n -dimensional version of the Pythagorean Theorem). The distance between identical objects will be zero. Any non-identical objects will be located at some “distance” from one another. The magnitude of the “Euclidean metric distance” defined above, has often been used as a measure of

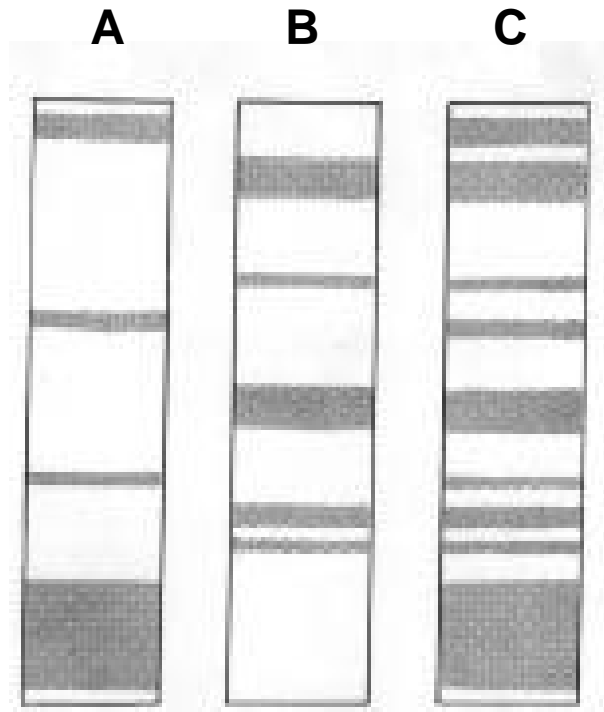


FIG. 4. Pattern **A**; low uniform optical density (0.01) over $3/4$ of its area and optical density 1.0 over the other quarter. Pattern **B** also has $3/4$ of its area almost completely and uniformly transparent but the points conjugate to those which absorb strongly in **A** here have the high transparency, while a different quarter of its area has an optical density of 1.0. Pattern **C** has the identical distribution of the high density regions of both **A** and **B** ($1/2$ of its area) and has high transparency at all other points.

similarity, (e.g., see 28, 29, 30). Zero distance means identity, and infinite distance, ultimate dissimilarity.

Intuitively, this kind of definition of similarity—a “distance” between points in a hyper-space—is quite appealing. However, the use of a “metric space” where distances are defined as, for example, in a Euclidean space, often has certain intuitive as well as computational disadvantages. Consider the three patterns in Figure 4. Pattern **A** has very low uniform optical density (0.01) over $3/4$ of its area and has a distribution of larger finite (1.0) densities over the other $1/4$. Pattern **B** also has $3/4$ of its area almost completely and uniformly transparent but the points conjugate to those which absorb strongly in **A** here have the high transparency while a different $1/4$ of its area has an optical density of 1.0. Pattern **C** has the identical distribution of the high optical density

regions of both patterns **A** and **B** ($1/2$ of its area) and has the high transparency at all other points.

In one intuitive sense, **A** and **B** are extremely dissimilar, but both are similar to **C**. Now the properties of a metric space are such that the distance between **A** and **B** can never be greater than the sum of the distances between **A** and **C**, and **B** and **C**. In fact, **B** will be only $\sqrt{2}$ further from **A** than **C** in Euclidean metric space.

To satisfy our intuitive feelings about similarity for this test case, we must find some different measure (definition) of “distance”—one which will describe **C** as being “close” to both **A** and **B**, while still permitting **A** and **B** to be “very far” from one another.

METRIC AND SEMI-METRIC DISTANCES AS MEASURES OF SIMILARITY

Hyper-spaces are characterized by the kinds of expressions that are used to define the distances between points. The distance defined by the square root of the sum of the squares of the differences between the coordinates of two points characterizes a Euclidean metric space where if **A** and **B** are close to **C**, they can not be too distant from each other. A distance formula, using the n coordinates of each pattern, and which permits **A** and **B** to be close to **C** but far from each other (even infinitely far) defines a more general class of spaces, among which are an infinite number of so-called “semi-metric” spaces. In all such spaces, the distance between identical objects must be zero (as in a metric space—and indeed, all spaces), but the measures of similarity or dissimilarity may be very different from space to space. Clearly then, the choice of a measure of similarity must, in a sense, be arbitrary, but it is reasonable to examine common concepts of similarity to at least see if we can find a useful match between one or more arbitrary “distance” measures and our intuitive feeling for what we usually mean by similar or dissimilar (as in the case of patterns **A**, **B**, and **C**).

Dr. T. T. Tanimoto has proposed (e.g., 31, 32, 33) the following measure: Consider the number of equal pairs of corresponding attributes of two objects. Were these the complete set of attributes of the two objects, the objects would be identical. The presence of even one additional unequal pair of corresponding attributes would destroy the identity relationship, but would leave the objects “quite similar.” More and more additional pairs of unequal attributes would “increase the distance” between these objects. Dr. Tanimoto proposed that the ratio of the number of pairs of non-zero identical corresponding attributes to the total number of pairs of corresponding attributes, of which at least one member is not zero, be defined as the “Similarity Coefficient.” This would have a value of 1 for identical objects and 0 for extremely dissimilar objects and is similar to a probability (in this as well as other respects). A distance formula defining a particularly useful semi-metric space is the negative logarithm to the base 2 of the Tanimoto Similarity Coefficient and is a kind of measure of information (for further discussion see 31). This “Tanimoto distance” is zero for identical objects, increases for less similar objects, and is infinite for extremely dissimilar ones, satisfying, in at least a gross way,

our intuitive feelings for the properties of a reasonable measure of similarity.

The Similarity Coefficient was originally defined by Tanimoto for attributes which had magnitudes of either 0 or 1 (the binary case). He has extended it to cover the continuous case, and here the definition takes the simple form: Similarity Coefficient equals the sum of the smaller of the magnitudes of each pair of corresponding attributes (coordinates) of a pair of objects divided by the sum of the larger of the magnitudes of each pair of corresponding attributes of a pair of objects. For the purposes of illustration, our three patterns can be represented as 3 sets of 4 corresponding attributes: Pattern **A** may be represented as 1.0, 0.01, 0.01, 0.01; **B** as 0.01, 0.01, 1.0, 0.01; and **C** as 1.0, 0.01, 1.0, 0.01.

If we examine this Similarity Coefficient among patterns **A**, **B**, and **C**, we find that the similarity between **A** and **B** is very near 0 (0.02), but the similarity of **A** or **B** to **C** is 0.51. The ratio of the corresponding Tanimoto distances is 5.7 for this case where the transparent background has a non-zero (0.01) optical density. If the background were completely transparent (optical density zero), the ratio of the two distances would be infinite in “Tanimoto Space.” In Euclidean metric space, however, the ratio of the two distances would still be 2. The Similarity Coefficient between **A** or **B** and **C** would equal 0.50 with a completely transparent background. Clearly, “Tanimoto Space” provides a measure of similarity, at least with this model, that is very much closer to our intuitive concepts of similarity and dissimilarity. The Similarity Coefficient is also simpler (therefore cheaper) to compute than a Euclidean metric distance function.

Many other measures of similarity can be devised, some with very special characteristics, such as the classical Correlation Coefficient,

$$C_{xy} = \frac{(\mathbf{x} - \mathbf{x}_i)(\mathbf{y} - \mathbf{y}_i)}{[(\mathbf{x} - \mathbf{x}_i)^2 (\mathbf{y} - \mathbf{y}_i)^2]^{1/2}} .$$

The Correlation Coefficient varies from +1 to -1. Objects which are identical have a Correlation Coefficient of +1 (with the restriction that the Correlation Coefficient between identical or non-identical *uniform* objects—i.e., values of all attributes of an object equal to each other—is indeterminate). Objects which would be identical to each other if all the attributes of one were either multiplied by a single appropriate constant and/or a constant were added to all the attributes of one, also have a Correlation Coefficient of +1. Objects with Correlation Coefficients of -1 are either perfect “negatives” of one another (magnitudes of corresponding attributes vary exactly inversely) or would be perfect negatives if all the attributes of one were multiplied by a single appropriate constant and/or a constant were added to the attributes of one. (In one sense, pairs of objects with a Correlation Coefficient of -1 are the most dissimilar of objects.) To summarize, the Correlation Coefficient gives a measure of similarity, normalizes for proportional and/or additive differences, and provides a measure of the “symmetry of opposites” by so-called “negative correlation.” For our test patterns, the Correlation Coefficient between **A** and **B** is -0.333, and between **A** and **C** or **B** and **C** is +0.577 (and these remain the same if the

background is made completely transparent).

This measure of similarity also has a kind of intuitive appeal—but different from that of Tanimoto's Similarity Coefficient. To convert the Correlation Coefficient to a distance measure for defining another semi-metric space, we can again take the negative logarithm. (The logarithm of -1 , like the $\sqrt{-1}$, is an imaginary number. Thus “Correlation Space” includes two spaces which share their “points at infinity” in common. One is a real semi-metric space and the other is an imaginary space with a corresponding imaginary point for each point in the real space.)

The Correlation Coefficient is a very powerful tool for recognizing certain classes of similarity which might otherwise go unnoticed without an additive or multiplicative normalizing step. However, in cases where the Correlation Coefficient is $+1$ as a result of compensating for both additive and multiplicative differences simultaneously, this result will usually appear to be in complete conflict with an ordinary intuitive sense of similarity or identity; (for example, consider as paired objects, the two sequences of numbers, 9, 6, 12, 3, 15, and 9, 8, 10, 7, 11). It is also, incidentally, about twice as expensive to compute the Correlation Coefficient as the Tanimoto Similarity Coefficient. On the other hand, two patterns in which the magnitudes of the attributes vary independently and at random with respect to one another, will have a Correlation Coefficient of 0 and will be infinitely separated in Correlation Space. This is clearly an attractive feature. Such a pair might occasionally be as close together as one Tanimoto unit in Tanimoto Space (see page 461).

No other measure of similarity of which we are aware seems to approach the Tanimoto Similarity Coefficient or the Correlation Coefficient in both generality and sensitivity, although simple theoretical arguments lead one to believe that there must be a host of as yet undiscovered but useful measures of similarity (distance functions of both metric and semi-metric spaces).

REWEIGHTING AND NORMALIZATION

Two kinds of transformations can be usefully performed on data before measuring the similarity of patterns. These constitute reweighting the data based on the informational content (statistical properties) of the data (e.g., removing redundant data), and normalizing the data—i.e., performing transformations such that two patterns, which before transformation might be quite far apart in a given semi-metric space, will, after transformation, “superimpose” (be identical). Simple transformation may involve translation or biasing (modifying the position or magnitudes of the attributes by an additive constant), expansion or contraction of scale (modifying the magnitudes of the attributes by a multiplicative constant), forming the “negative,” mirror inversion, etc. The purposes of reweighting and normalizing are different although operationally they may be indistinguishable. Thus the removal of redundant information by dividing all values of an attribute of a population by the lowest common denominator among the values of that particular attribute among the members of that population—one kind of reweighting step

—is indistinguishable from a normalizing contraction of scale in the dimension represented by that attribute.

All normalizing or reweighting steps must be of such a nature that when applied to initially identical patterns, the distance between such patterns will remain zero after the normalizing or reweighting procedure. It is also desirable that normalizing or reweighting steps should be of such a nature that no useful information about differences be lost or discarded inadvertently during data processing.

We further distinguish these two processes as follows:

Reweighting; the Removal of Redundancy

The general purpose of reweighting is to reduce the magnitude of a special kind of similarity (e.g., redundancy) in a given population. This is done to increase the sensitivity of a similarity measure for detecting the remaining differences between patterns. It is desirable that reweighting reflect the statistical significance (or information content) of the raw data. Because of its very nature, inadvertent loss of useful information during reweighting is unlikely.

Normalization; a Topological Transformation

The general purpose of normalizing is to increase the sensitivity of a similarity measure to otherwise “hidden similarities” between patterns. Now just what do we mean by “hidden similarities”? Two samples of a single volume of serum, one diluted to half the concentration of the other, would give disc electrophoresis patterns in which the optical density at each band of the undiluted specimen would be twice that of the corresponding band in the other (see Fig. 5). If we arbitrarily decide that such a uniform dilution has little significance for our purposes, but the fact of identity after a change of scale is highly significant, then the application of this transformation is called a normalizing step. Clearly, if we are mistaken about the relative significance of the difference in concentration as compared to the “hidden identity” of these samples, and if we employ a normalizing step which “throws out” the original information about the difference in concentration, not only will we fail to discover our error from the ensuing analysis, but we could never retrace our steps.

Normalization can therefore be hazardous. To eliminate this hazard it might be desirable to compare raw data without any normalization and then to again compare the same data after normalization. If we are in complete ignorance about the information content of the universe of the population being sampled and classified, there is no doubt that this safe procedure is the only reasonable procedure (even though it is also a more laborious one).

When, on the other hand, one has chosen or designed the straight-jacket into which the “raw data” must fit—the particular method of measurement or observation or environmental transducer or “sense organ” used—one has usually already passed the rawest data through a variable number of implicit or explicit normalizing steps. From this frame of reference therefore, when economy is important, it is more realistic to try to

recognize such data-gathering normalizations explicitly, and to add any others which were not conveniently included in the data-gathering process but can be argued to be “reasonable.” If the data are also preserved in something close to their original form, as well as in the normalized forms, it will always be possible to reexamine those original differences between samples which disappear after normalization.

As indicated earlier (page 450), the Correlation Coefficient can automatically provide two kinds of normalization while simultaneously providing a measure of similarity. It is largely because of an “over-abundance” of kinds of normalization that I consider it inappropriate (hazardous) as a primary measure of similarity.

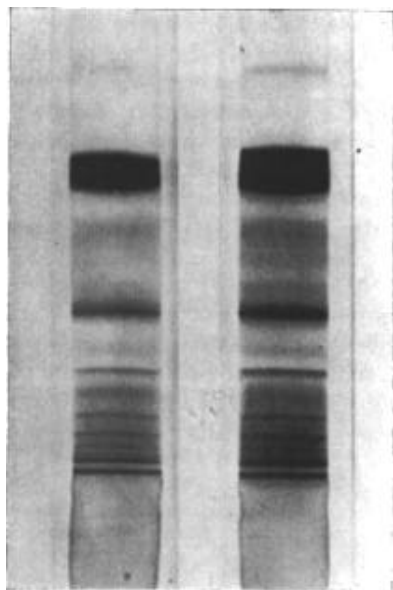


FIG. 5. One of the samples from Figure 3C: 3 microliters of serum in the right-hand run, 1 1/2 microliters in the left-hand run.

In general, normalizing steps should usually precede reweighting steps because it is the redundancy of the normalized population that one wishes to reduce.

A SIMPLE TAXONOMIC SCHEME FOR ONE-DIMENSIONAL PATTERNS

We will now describe a scheme of analysis starting with a matrix of 10,000 columns (the samples) and 400 rows (the positions in each pattern) with 128 numbers from 0 to 1,000 (the anti-logs of the 7 bit logarithms of concentration, accurate to 5 binary places) occupying the positions in the matrix. Given such a matrix, we now apply a normalizing step to insure that samples which are identical, except for a dilution of one relative to the other, will be equal. This is accomplished by dividing the value of each attribute in a

column by the average value for that column.

Next comes a reweighting step. If we locate the smallest non-zero value in each row of the normalized matrix and divide each value in that row by that smallest value, entering such quotients in a new conjugate matrix, we will have reweighted the data so that all corresponding attribute magnitudes shared by all samples will have been removed. With this redundant information eliminated, the measurement of similarity between patterns (columns) will be more sensitive to the magnitudes of residual differences from pattern to pattern. (It is clear that a good deal of sharing of attributes will usually still remain to decrease the sensitivity to differences.) The above procedure is the simplest* for increasing sensitivity to differences without losing information.

The Tanimoto distances between each column and every other column are now computed and are entered into a new matrix, the Similarity Matrix, with 10,000 rows and 10,000 columns. The average Tanimoto distance between each sample and every other sample, i.e., the sum of the entries along a row of the Similarity Matrix divided by 10,000, is computed and samples are ordered in a list on the basis of this average Tanimoto distance. The sample with the smallest average distance from all other samples, (i.e., the “most typical” member of the population) will head this “Hierarchical List” (31).

A histogram of the frequencies of the distances between this most typical member and all other members is plotted. The first “peak” is located. (This relative maximum might rarely occur at the origin.) If it falls more than 1 Tanimoto unit of distance (Similarity Coefficient less than 1/2) from the origin, all specimens within 1 Tanimoto unit of the most typical specimen (as well as the most typical specimen) are transferred to the bottom of the Hierarchical List. The sample now heading the List is selected as “a most typical member” and a histogram of the above type is again plotted, and its first peak located. If the peak falls more than 1 unit of Tanimoto distance from the origin, the Hierarchical List is again rearranged.

If, on the other hand, a first peak is within 1 unit of Tanimoto distance of a most typical member, the first minimum in the distribution curve with a magnitude less than 1/2 of the value of the first peak is located (point **A**) (see Fig. 6). The next point towards the origin with a magnitude equal to 1/2 the value of the first peak is located (point **B**).

* A more balanced reweighting requires more sophisticated data processing. For example, the number of occurrences of each normalized value in each row divided by 10,000 gives us the frequency of occurrence of that normalized concentration of protein at that position in the pattern, $[f(x)]_r =$ frequency of a particular value of x at position (row) r . In so far as the samples making up the set of 10,000 were chosen at random from a given larger population, the set of frequencies approximate the probabilities that a new sample, picked at random from that larger population will have each of these normalized concentrations at each position ($p(x)_r =$ probability of a particular value of x at position r). By multiplying each entry, x , by $-\log_2 p(x)_r$, (the self information (22) of each kind of entry in a row), the magnitude of very rare values of the attributes are emphasized, independent of their original magnitudes. This method of reweighting will be used further on in a more sophisticated scheme.

Then the next point towards the origin with a magnitude equal to $3/4$ of the value of the first peak is also located (point **C**). Twice the distance along the abscissa between the points **B** and **C** is added to the value of the abscissa at **B** to locate a new point on the abscissa (point **D**). The "Boundary of the First Branch" is set by either the location of the minimum, **A**, or the point **D**, whichever is further along the abscissa from the origin. (In this procedure, **D** is the intersection with the abscissa of the line drawn through points **B** and **C**. If the distribution function of the marginal portion of this sub-group were Gaussian, less than 4.5% of the "true" members of that group would lie at Tanimoto distances greater than **D** from the most typical member.)

All specimens within the "radius" (**D** or **A**) of the "semi-metric space hypersphere" surrounding the most typical member or "Archetype" constitute the membership of the

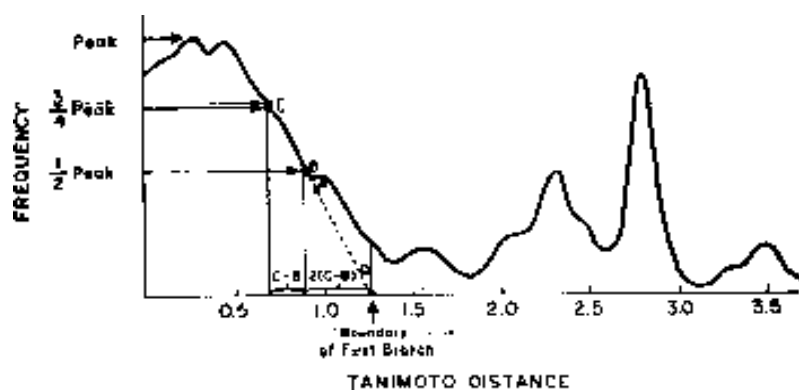


FIG. 6. Model of a frequency distribution (histogram) of Tanimoto distances between a "most typical" sample and all other samples, demonstrating the application of the rules for determining the boundary of the First Branch population.

First Branch population and are entered into a new matrix, the First Branch Matrix.

All members of the First Branch within 1 Tanimoto unit of the Archetype are removed from the Hierarchical List and the highest remaining member on the list is chosen as a candidate for the title, Archetype of the Second Branch. All the steps performed with a most typical sample are repeated with this sample to locate the boundary and members of the Second Branch (which may also include some marginal members of the First Branch). These samples are entered into a Second Branch Matrix, and appropriate members of this Branch are removed from the Hierarchical List, and the whole process is iterated until less than 1,000 members remain on the List. All these remaining samples are then added to the "Residue Branch Matrix of the Trunk."

The numbers of members in each of the Branch Matrices is noted and the Branches are ordered according to their sizes.

A new sample (the ten thousand and first) is now normalized and reweighted with the set of lowest common denominators of the rows from the original Trunk Matrix (as were

all the original samples) and is compared to the Archetype of the Largest Branch (other than the Residue). If the Tanimoto distance between them is less than the radius for that Branch, the sample is added to that Branch Matrix. If the distance also is equal to or exceeds 1 Tanimoto unit, or if it exceeds the radius of that Branch, it is next compared to the Archetype of the Next Largest Branch, and the entire process is iterated until a Branch is located, the Archetype of which is closer than 1 Tanimoto unit to the new sample, or failing that, the new sample is added to the Residue Branch Matrix.

This procedure is repeated with succeeding new samples until the population in a Branch Matrix equals 10,000. At that time, the Branch is normalized, reweighted, and subdivided in exactly the same manner as the original Trunk population. This procedure generates a taxonomic tree and at the same time, provides a means for locating those sets of samples out of the total population of samples studied which are most like a new case. [By examining other attributes (e.g., clinical findings) from the records of the set of patients in a “terminal” Branch, and adding these additional attributes—rows (attributes) may be compared and classified in the same way that patients’ patterns (columns) were compared above, and if the pattern characteristic of a “terminal” Branch is diagnostic of a particular disease entity, then the most significant attributes of the pattern will be grouped with the constellation of clinical findings that are known to characterize the disease.]

JUSTIFICATION OF THE SCHEME

Let us now examine the justification for these steps explicitly, and discuss the amount of computer time required for such data processing. We will then examine a more sophisticated technique which has grown from the above procedure. This technique drastically reduces cost (by a factor of 1,000 or more), reduces “arbitrariness,” and increases flexibility and generality. The refined technique will be listed in a somewhat more formal way, as a set of instructions from which a programmer could more easily and directly work.

GROUP CENTROID, “NATURAL” BOUNDARIES BETWEEN GROUPS, AND HYPER-SPHERES

When one considers subdividing a group of objects, patterns, etc., to generate a taxonomic tree (classification), one usually assumes that some cohesive clusters with easily definable “natural” boundaries exist within the population to be subdivided. This assumption is not necessarily valid. Consider a population of patterns in which the optical densities have a Gaussian density distribution of amplitudes that vary at random from point to point within a pattern and from pattern to pattern. Such a “white noise” source has no “internal structure.” If we examine the distribution in a hyper-space of the points represented by such a group of patterns, we will discover a single cluster with a “bellshaped” density distribution (unimodal) from its “center” out, and no subdivision of such a population would yield “natural subgroups” (Fig. 7).

Consider next, a population of m discrete kinds of patterns which are as distinct as patterns **A** and **B** (page 448). Such a population would provide a set of m discrete points in a hyper-space. Now superimpose low magnitude “white noise” on the individual samples of discrete patterns and examine the distribution of the new noisy patterns. If the magnitude of the noise, relative to the pattern optical densities, is not too great, we will now have a polymodal distribution of m distinguishable clouds or clusters, each centered on the position of the corresponding discrete point in the previous distribution (Fig. 8). If we increase the magnitude of the noise, these clouds will overlap—and if the signal to

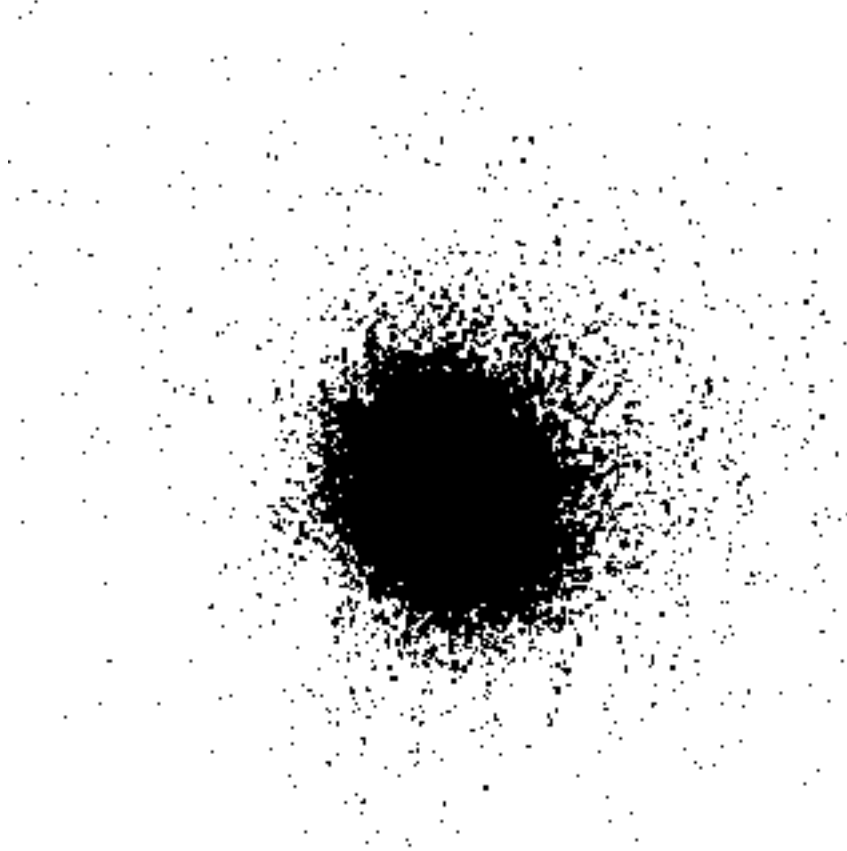


FIG. 7 Model of the distribution of points representing members of a “white noise” population of patterns in a two-dimensional metric space.

noise ratio degenerates sufficiently, the overlapping clouds may become indistinguishable from a single “white noise” distribution.

When there is no “overlap” of clouds, a very large number of kinds of hyper-surfaces (boundaries between n -dimensional volumes) can be generated in the hyper-space, any set of which will sub-divide the population into the same set of m subgroups. Such subgroups will be called natural subgroups. However, as soon as we have “overlap” between two such “noisy” populations, there exist many fewer kinds of surfaces which will yield identically divided populations. Any dividing hyper-surface is then arbitrary, and any claim to “naturalness” for a classification derived from such data can receive its

only support from the arguments used to defend the particular boundary-defining conditions and distance function used. It is convenient to consider the “spherical” surface of the hyper-sphere (surface generated using the distance from the “center of mass” or centroid of an ideal natural cluster to its most distant member as the “radius”) as the

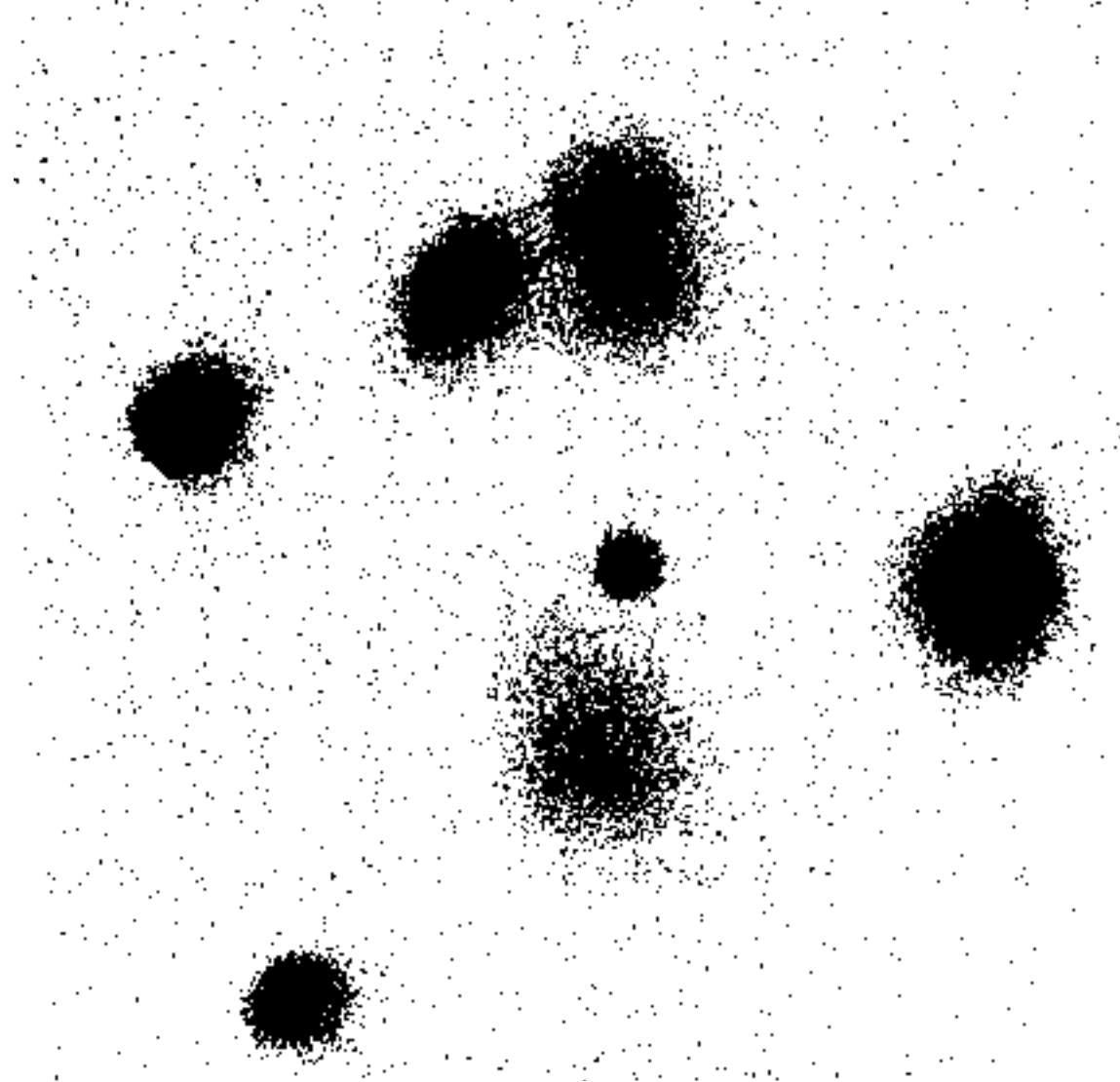


FIG. 8. Model of a polymodal distribution of $m = 7$ different "noisy" populations of patterns in a two-dimensional metric space.

boundary of the cluster (31, 34, 35). (Of course, for the natural subgroup, many other arbitrary hyper-surfaces can usually serve as well, but no other is as conveniently defined. Some infrequently encountered but none the less distinct “natural” populations could never be resolved by a simply connected surface such as a sphere e.g., two interlinked toroids. To discover and separate such populations, less convenient algorithms which search for discontinuities in connectivity must be used.)

Operationally, to generate such a hyper-spherical boundary surface for natural or unnatural clusters, we need two devices:

- 1) We must locate the centroid or a reasonable substitute “central point.”
- 2) We must find a radius measured from that “central point” which generates a hyper-sphere enclosing most of the members of an unnatural cluster, excluding most members of other clusters, and hopefully subdivides most natural populations without any admixture.

RELATIVITY IN METRIC AND SEMI-METRIC HYPER-SPACES

In Euclidean space, if we seat ourselves at any convenient point in the coordinate system and view the relative distances between galaxy clusters, galaxies, star clusters within a galaxy, and individual stars, we can then move to another observation point, and the relative distances between points in this space will be observed to be the same (assuming the use of some rangefinder with infinite resolution). A procedure for locating clusters operating from the origin of the co-ordinate system, or any other fixed point could therefore serve the requirements of device 1). Unfortunately, as pointed out earlier in the discussion of metric spaces, points in n -dimensional Euclidean space will often be quite “crowded” and overlap of subgroup spheres would therefore be expected to make natural subdivision quite rare.

On the other hand, in a semi-metric space such as Tanimoto Space, where cleaner separation of dissimilar clusters can be anticipated, the universe and its galaxies present very different relative views, depending upon the point in space from which one makes his observations. From point **C**, **A** and **B** are both relatively and equally nearby, yet from **A**, **B** is very much further away than nearby **C**.

To locate the centroid of a cluster in a semi-metric space, it is therefore necessary, in principle, to sit down at each and every point in that space and look outwards at all other points, and on that basis locate the centroid of a cluster.

CLASS ARCHETYPE AND THE HIERARCHICAL LIST

If a cluster is “unimodal” and “symmetrical,” the centroid will be closest to the “most typical” member of the population, and the average Tanimoto distance from this member to other members of the cluster will have the minimum value. That is, among the average Tanimoto distances between each member of this cluster and every other member, the “most typical member” will have the lowest value and will therefore head the Hierarchical List (see Tanimoto, 31).

If the cluster is not unimodal—for example, is polymodal, with the centroid of each of the subgroups distributed equally spaced on a “spherical shell,” then the point at the “center” of the shell will be the centroid for this polymodal population. Let us suppose that the spread of each cluster is sufficiently large to extend to the center of the sphere (Figure 9). Then the specimen nearest the center will head the Hierarchical List for this polymodal population.

In the case of the unimodal population, the “most typical” member will have a larger number of closest neighbors than any other member. In contrast, in the case of this spherical shell polymodal population, the member nearest the centroid will have among the fewest number of closest neighbors. If the shell radius is a great number of Tanimoto units in magnitude, then the specimen nearest the centroid will really be a very poor

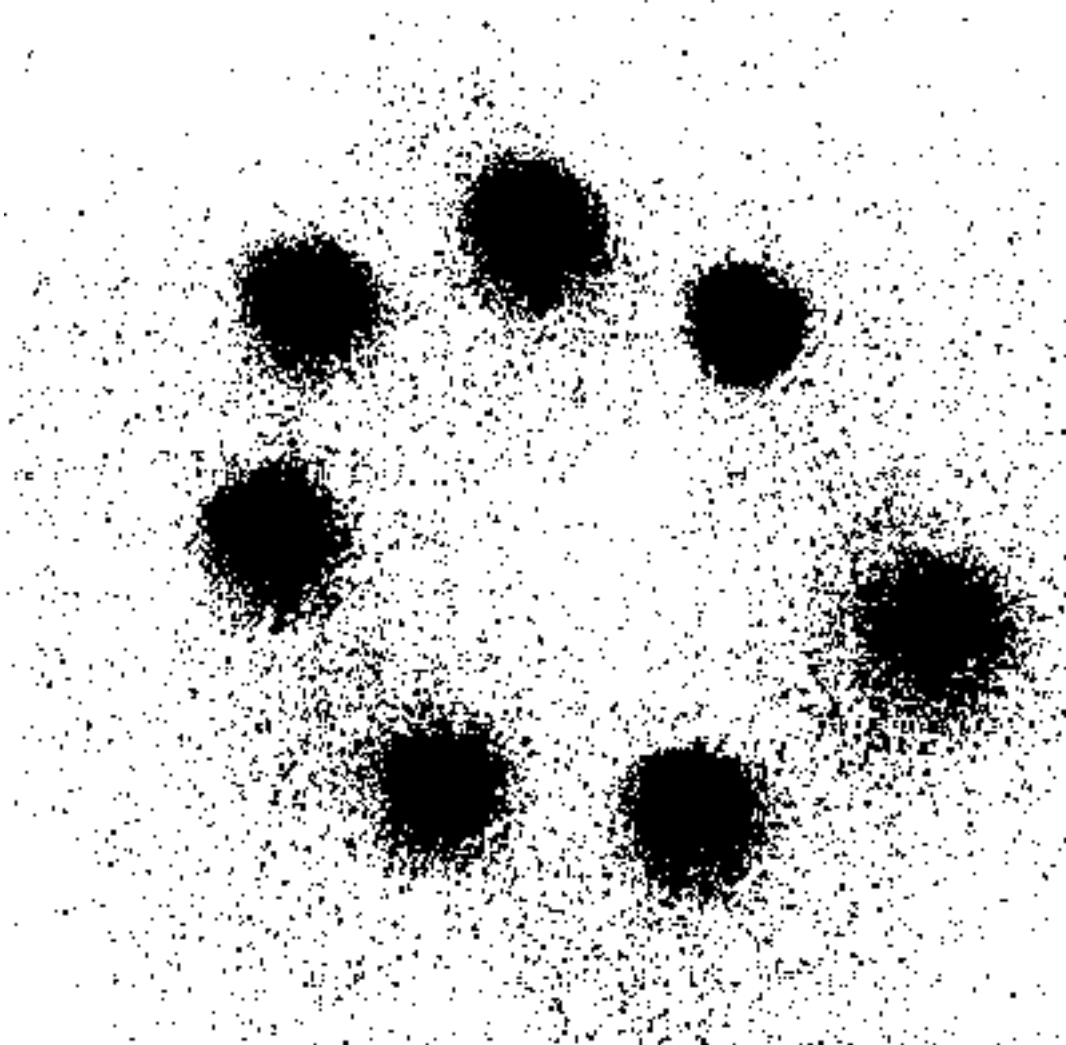


FIG 9. Model of a “spherical shell-distribution” of 7 “noisy” populations of patterns in a two-dimensional metric space.

Archetype. Such a polymodal subgroup may none the less be quite natural in being completely separated from other more distant subgroups (unimodal or polymodal). If we wish the “Archetype” to have the connotation of a “most typical” sample, we must reject such samples near low density centroids as candidates for the title, Archetype of a Branch. Alternatively—and with somewhat less arbitrariness—we might use the designation, Population Centroid, understanding that such a member of a population may be a far cry from what one intuitively thinks of as a “type specimen” or Archetype. (The

two approaches will yield somewhat different branching patterns—(i.e., they will yield different intermediate subgroups) although the number and identity of the “terminal” branches of two such taxonomic trees would be expected to be quite similar.)

The Hierarchical List provides us with the “Population Centroid” of the Trunk. Examination of the frequency distribution of Tanimoto distances between the Population Centroid and all other samples tells us whether it is located in a low density or high density volume of Tanimoto Space. The magnitude of this distribution function and the sign and magnitudes of its first and second derivatives provide all the necessary cues. Clearly, however, the boundary between what one considers to be low density and high density must be arbitrary. We have above proposed using the kinds of criteria that are commonly used in analysis of histograms. [There is obviously considerable room here for the use of more sophisticated criteria which, for example, examine boundaries using assumptions of other than normal distribution (36, 37).]

The following special property of the Similarity Coefficient was also taken into account:

It can be shown that the average Tanimoto distances between a sample of a population, in which all magnitudes of all attributes are equally likely, and all other members have a distribution with the mean location of a peak at 1 Tanimoto unit, and approaches minimum values at both zero and infinite Tanimoto distances. Therefore patterns related only randomly might average separations as small as 1 Tanimoto unit (as will also such clearly related patterns as for example, **A** and **C**).

SUBDIVISION OF THE POPULATION

By transferring a Population Centroid and all its neighbors within 1 Tanimoto unit to the bottom of the Hierarchical List when the first peak is located beyond 1 Tanimoto unit (as would be the case for the polymodal shell model distribution, following the analytical scheme proposed on page 454) we make it more likely that the next highest remaining member on the Hierarchical List will fall near a high density point which is also near a centroid. Transfer of only the centroid itself would usually leave the next nearest member to the centroid at the top of the Hierarchical List. By transferring a sufficient number of closest neighbors of the centroid, we therefore insure a shift to a new center of gravity. In the shell model, this new “central point” would be near the center of one of the overlapping subpopulations on the shell, and the first minimum in the distribution curve would occur beyond the furthest subpopulation on the “opposite side” of the shell, if the populations overlapped sufficiently. If not, an individual subpopulation would be isolated including some marginal members of neighboring subpopulations and excluding some (of the order of 4.5%) of the “proper members” of that first subpopulation.

By using either a Population Centroid, or an Archetype as defined above, we thus locate useful “central points” satisfying the requirements for device 1).

The simple rules for determining the group radii are given on pages 454 and 455 satisfy the requirements for device 2).

Such simple rules can result in “too early” subdivision of a natural subgroup—i.e., two overlapping subgroups may have a finite minimum between them but a zero minimum beyond the second. A set of rules for more natural subdivision would call for preliminary search for the widest or “most significant,” minimum with rules for deciding whether or not to use its location, or that of the first minimum (or any other minimum closer to the origin) as the basis for setting the radius for the subgroup.

The kind of boundary rules described assure that marginal members of groups will appear in more than one Branch of the Tree. This partly compensates for the arbitrariness of “unnatural” boundaries, but does this at the expense of increased computational cost.

Some elements of the analysis presented, though intuitively simple, are none the less relatively novel. We have attempted to resolve the classification problem with methods which have their origins in ordinary intuitions, not because we believe that intuitively appealing methods are necessarily the best, but because experience has often shown that if new and useful methods are to be discovered during an essentially exploratory phase of the development of a field, aimless groping for new directions often results from the loss of perspective that can follow when one casts loose, too early, from intuitive moorings.

Whereas the boundaries of our populations, i.e., hyper-spherical surfaces, are set by a semi-metric distance (such boundaries are now usually referred to as *decision boundaries* in the jargon of decision theory, a new branch of statistics (38, 39) which threatens to engulf the whole), it is more commonly the practice, in decision theory, to carve up a metric decision space with hyper-planes. Less commonly, hyper-spheres and hyper-quadrics have been suggested as promising alternatives (35). The implementation of such techniques will usually be more complicated than the calculation of a simple semi-metric radius, and the complexity of the computations will usually grow very rapidly as one progresses from relatively few simple hyper-planes to the complex but less arbitrary hyper-quadric surfaces.

Because of the “asymmetries” of many interesting semi-metric spaces (including Tanimoto Space and Correlation Space) the simply computed semimetric hyper-spherical boundary can often be expected to give very clean separation of “natural” populations which are themselves asymmetric in a metric space and/or are so crowded that only very complex decision boundaries could hope to separate them cleanly. Thus the use of a simple semi-metric space, in a sense, permits the equivalent of a topological distortion and separation of such natural populations sufficient to permit the use of very easily defined and simply implemented decision boundaries. This results from reversing priorities in setting up the problem. Rather than using a familiar metric space to define a less familiar measure of similarity, a more intuitive measure of similarity is used to define our semi-metric decision space. Tanimoto's Hierarchical List then permits the simple and unambiguous location of those population centroids in semi-metric space which are

necessary to generate the hyper-sphere—but at some expense. We will now examine this added cost to see whether the potential computational simplicity of this semi-metric approach can be preserved while reducing the cost to reasonable levels and at the sacrifice of no more than a modicum of added ambiguity.

COMPUTING COST

If any one kind of operation in such a scheme requires a number of magnitudes more computer-time than all the rest of the operations, examination of the economics of this operation will, in fact, give a reasonable measure of the total cost.

The construction of the Hierarchical List requires the computation of a number of Similarity Coefficients equal to one half the square of the number of samples in the List. The time to compute one Similarity Coefficient between two 400 position samples with 7 bit resolution at each position on the IBM 7090 is about 25×10^{-3} seconds (40). The computation of the Trunk Similarity Matrix (or any other Branch Matrix of 10^4 samples) would therefore require $0.5 \times 10^8 \times 25 \times 10^{-3}$ seconds, or 14.5 days. The computer time per sample would average about two minutes, or at \$500 per hour, about \$17 per sample. Since, as one proceeds up the tree, appreciable numbers of samples will appear in more than one Branch, the cost might rise to a few times this figure.

Since, from the perspective developed to this point in the discussion, the “education” of the computer would require the growth of many levels of branches before the Tree might be diagnostically useful, the investment in computer time might have to become of moon-shot proportions before its utility could be tested. If the number of ultimate Branch Tips is fixed (the population has a finite number of natural indivisible subgroups such that, after addition of members in excess of the number necessary to activate the “subdivide instruction”, all distribution histograms of Tanimoto distances among members of a Branch Tip remain unimodal and drop to zero beyond the peak) then beyond the time when those terminal Branch Tips have been resolved, the “education” of the computer would be complete. From then on, the time required for classifying each additional sample would drop appreciably (for example, for a symmetrical dichotomous tree with 2^{30} Branch Tips, from two minutes per sample to 0.75 seconds). Thus, if one had great confidence in the utility of such a program—sufficient confidence to invest billions of dollars in the education of the computer before it “grew to a productive age,” one might consider going ahead, even with such an enormous initial investment, since the ultimate cost per sample would be quite low.

We did not have *such* confidence. Therefore, it was necessary to examine ways for at least approaching the theoretical performance of such a system by means which would reduce the cost by a number of orders of magnitude.

Since the cost is proportional to the square of the number of samples in a Similarity Matrix, a reduction of the subdivision limit from 10,000 to 1,000 samples would result in a hundred-fold saving. Our original reason for arbitrarily specifying 10,000 samples per

matrix was to try to assure that the lowest common denominators used for reweighting would be representative of the population as a whole with a very high degree of confidence. A more reasonable way to specify when a population should be subdivided would depend upon the observed statistical structure of that population such as might be revealed by continuous monitoring of the parameters of the part of the population sampled. In this way an explicit cue could be automatically provided when the particular population might be subdivided with a specified degree of confidence. If this device would lead to smaller Similarity Matrices, some economy might thereby be achieved. Alternatively, this less arbitrary device might lead to even larger matrices.

HISTOGRAM "CROSSECTIONS" OF A SEMI-METRIC SPACE AND BYPASSING OF THE HIERARCHICAL LIST

More significantly, any device that might bypass the calculation of the Similarity Matrices and Hierarchical List to locate the "central points" of subgroups in Tanimoto Space would permit very substantial savings.

Consider a sample chosen at random from a population of 10,000 patterns. If we compute the Tanimoto distance between it and all other samples and plot a histogram of frequency versus distance, we will, in general, find a distribution curve with a number of maxima and minima. If we search out the highest peak in this histogram we will have located the largest number of specimens at any fixed distance from this first sample. These may either be all quite similar to one another, or as disjoint as our optical density patterns, **A** and **B**, (since patterns infinitely distant from one another may lie at finite and occasionally equal distances from a third pattern in a semi-metric space).

If we use the first histogram to locate a new specimen from this modal peak (any specimen at the distance of the peak from the first specimen), compute the Tanimoto distance between it and all other members and then plot a new histogram, one of the following distributions will be observed:

a) The distribution now peaks very near the origin. This peak is also the largest of the peaks in the histogram and includes approximately the same number of specimens as were included under the peak in the first histogram from which this sample was extracted. This would usually mean that the first peak in fact represents a unimodal and symmetrical group of specimens which were close to one another as well as at nearly equal distances from the first specimen, and this second specimen is therefore very near the "central point" of this largest First Branch population. The process may be repeated and after each iteration, the largest peak should move closer and closer to the origin and each newly extracted modal specimen should approach closer and closer to the specimen which is the "True Population Archetype." For the purposes of useful subdivision a close neighbor of the "True Archetype" will obviously be satisfactory and the degree of closeness required can be explicitly specified. If only ten such sequential "crosssections of this Tanimoto Space" proved to be adequate to locate an "acceptable" Archetype, then the cost of subdivision would have been reduced about 1,000 times. As the Taxonomy grows, the

cost per sample rises to the ultimate cost per sample (of the order of 15¢ on the IBM 7090) with the “fully educated computer” above. (Such a cost is at least tolerable for a useful diagnostic tool.)

b) The distribution may now peak near the origin, but some new peak, far removed from the origin is instead the largest peak. This means that the peak in the first histogram was composed of two or more groups (e.g., like **A** and **B**) which have been separated by choosing one of their members as our new frame of reference. We now proceed to select a sample from the new largest peak and iterate. It will then either provide us with observations a) or b) above, or c).

c) The distribution now provides a set of two or more equal peaks of maximum height. The significance is the same as in b) and our strategy is also the same with any of the equal peaks being used for locating a next candidate for Archetype of the Branch.

Following observations b) or c), a larger number of crossections will in general be required to locate an “acceptable” Archetype for the Branch, but should still be orders of magnitude smaller than for computing an entire Similarity Matrix. The definition of “largest peak” is obviously open to some refinement which would, for example, take both “area under the curve” as well as peak amplitude into account.

This heuristic (41, p. 6) method of sampling the population with trial “crossections” through a semi-metric space to locate the candidates for Archetype of a Branch promises to provide the requisite economy to make this kind of Taxonomic procedure both practical at this time and the leading contender as a model for mechanical induction.

A MORE EFFICIENT PATTERN RECOGNITION PROGRAM

We will now list the steps of an improved Taxonomic procedure.

Definitions

x_{rc} is the magnitude (to 5 significant binary places) of the anti-log of the 7 bit logarithm of the optical density (therefore proportional to the concentration of protein) at position (row) r in sample (column) c .

n_c is the number of samples (columns) in an active memory matrix.

n_r is the number of positions (rows) in a sample, (e.g., 400 in this case).

Then the row and column means are given by,

$$\mathbf{X}_r = \frac{c}{n_c} \begin{matrix} x_{rc} \\ \vdots \\ x_{rc} \end{matrix} ; \quad \mathbf{X}_c = \frac{r}{n_r} \begin{matrix} x_{rc} \\ \vdots \\ x_{rc} \end{matrix}$$

and the matrix mean by,

$$\mathbf{X} = \frac{r}{n_c n_r} \begin{matrix} c \\ \vdots \\ c \end{matrix} \begin{matrix} x_{rc} \\ \vdots \\ x_{rc} \end{matrix} = \frac{r}{n_r} \mathbf{X}_r ,$$

and a “conservatively biased” estimate of the standard deviation of a row is given by,

$$s_{x_r} = \left[\frac{c}{(n_c - 1)} \frac{x_{rc}^2}{n_c} - \mathbf{x}_r^2 \right]^{1/2} \left[\frac{n_c}{(n_c - 1)} \right]^{1/2} \left[\frac{c}{n_c} \frac{x_{rc}^2}{n_c} - \mathbf{x}_r^2 \right]^{1/2} \left[\frac{c}{n_c} \frac{x_{rc}^2}{n_c} - \mathbf{x}_r^2 \right]^{1/2}$$

for $n_c \gg 1$.

The “Signal to Noise Ratio” for a row, which may be used as a *measure of confidence*[†] in the significance of the value, \mathbf{x}_r , is defined as,

$$(S/N)_r = (N/S)_r^{-1} = \frac{\mathbf{x}_r (n_c - 1)^{1/2}}{s_{x_r}} \frac{\mathbf{x}_r n_c^{1/2}}{s_{x_r}} \text{ for } n_c \gg 1.$$

A weighted root-mean-square value of the $(S/N)_r$ ’s is given by,

$$(S/N)_r = n_r \left[\frac{1}{r} (N/S)_r^2 \right]^{-1/2}.$$

All boundary defining steps which explicitly or implicitly affect the level of confidence by setting a limit of acceptable error or acceptable Signal to Noise Ratio will be marked with an asterisk (*). The level of confidence considered acceptable is clearly arbitrary and may be modified from the values suggested below to either decrease or increase the overall confidence that may be placed in the “prediction” of the Taxonomy generated.

[†] If, for each row, the magnitudes of x_{rc} are independent random variables, then from Tchebysheff’s inequality (42), it can be easily shown that the probability that the limiting value of \mathbf{x}_r (the value of \mathbf{x}_r for $n_c = \infty$) will differ from \mathbf{x}_r by less than $\epsilon > 0$, (where $\mathbf{x}_r \pm \epsilon$ is the *confidence interval*) is

$$P_c = 1 - t^{-2}$$

where $t = (\epsilon/\mathbf{x}_r) (S/N)_r$, (ϵ/\mathbf{x}_r) is the maximum tolerable “Fractional Error” and $(S/N)_r$ is the Signal to Noise Ratio as defined above.

If, for each row, the magnitudes of x_{rc} in addition have a unimodal normal density distribution over the columns, then from the Central Limit Theorem (42), for large n_c ,

$$P_c = 2 \int_0^t (2\pi)^{-1/2} e^{-u^2/2} du$$

which is the area under the normal density curve of mean 0 and standard deviation 1, for the confidence interval, $\mathbf{x}_r \pm \epsilon$. For $t = 1.96$, $P_c = 0.95$; $t = 2.58$, $P_c = 0.99$; and $t = 3.30$, $P_c = 0.999$. (For small n_c , t will be distributed approximately according to “Student’s” distribution (43).)

If the distribution is actually multimodal, the probability, P_c , will generally be even larger. It will always be larger for multimodal distributions in which the component unimodal distributions are of Pearson type II (36) which includes the normal density (or Gaussian) distribution and the uniform (or “rectangular”) distribution as the two opposite extremes of this type.

As can be seen from this brief discussion, the *Fractional Error*, which is a measure of the *confidence interval* (that is independent of scale), together with the *Signal to Noise Ratio* (which is also independent of scale), are especially convenient measures of the *confidence level*, P_c .

The greater the confidence required and/or the smaller the acceptable error, the greater the processing delay and the higher the cost.

Crude Instructions for Writing an Improved Taxonomic Program

1) Sample values of x_{rc} and an identification code word for each sample are entered into an active memory sequentially, and running accounts of n_c , $\sum_c x_{rc}$, and $\sum_c x_{rc}^2$ are kept to compute $(S/N)_r$. When $(S/N)_r$ has increased to a value of 1,000*, an artificial average sample consisting of the x_r 's is recorded and the population is subdivided as follows:

2) Each entry, x_r , is normalized by a factor $k_{rc} = x_r^2 / \sum_r x_r x_c$, and the frequencies, $f(k_{rc} x_{rc})_r$, of occurrence of each normalized magnitude, $k_{rc} x_{rc}$, (in 128 uniform* 5 %* intervals) in each row are tabulated in a 358,400* bit memory and each normalized value is reweighted by multiplying by the factor

$$j_{rc} = \frac{-\log_2 f(k_{rc} x_{rc})_r}{\frac{1}{n_r} - f(k_{rc} x_{rc})_r \log_2 f(k_{rc} x_{rc})_r}$$

giving $k_{rc} j_{rc} x_{rc}$. (j_{rc} is the self-information (22) for a particular entry divided by the average self-information per entry.)

3) The Tanimoto distances, i.e., the negative logarithms to the base 2 of the Similarity Coefficient, between a reweighted artificial average sample composed of attributes equal to $x_r j_r$, (where $j_r = j_{rc}$ for $k_{rc} x_{rc} = x_r$ in row r), and all other normalized and reweighted samples are computed; the frequencies of Tanimoto distances, in intervals of 0.01 units* are computed; and a histogram is "plotted."

The "largest peak" is located in the following manner:

The modal frequency of the distribution is located. The two closest minima on each side of the mode with magnitudes *less than* 1/8* of the value of the modal peak are located (points **A** and **A'**). The next points towards the modal peak with magnitudes *equal to* 1/2* of the value of the modal peak are located (points **B** and **B'**). Then the next points towards the modal peak with magnitudes *equal to* 3/4* of the value of the modal peak are also located (points **C** and **C'**). Twice* the distance along the abscissa between points **B** and **C** (on the side nearest the origin) and between **B'** and **C'** (on the side away from the origin) is subtracted in the case of **B-C**, and added in the case of **B'-C'** to the values of the abscissa at **B** and **B'** respectively to locate new points on the abscissa (points **D** and **D'**). The boundaries of the peak population are set by either* the locations of the minima (**A** and **A'**) or* the points **D** and **D'**, whichever are *further* along the abscissa from the modal peak. (In this particular procedure, **D** and **D'** are the intersections with the abscissa of the lines drawn through points **B** and **C**, and **B'** and **C'** respectively. If the distribution function of the marginal portions of this modal peak were Gaussian, less than 4.5 % of the "true" members of that group would lie beyond **D** and **D'**. If the peak were perfectly Gaussian, **A** and **D**, and **A'** and **D'** respectively would almost superimpose, and would be located quite close to the 2 points on the distribution curve.) The total number of

samples included within the boundaries (total area under the curve between the boundaries) is recorded.

The next largest value of frequency *outside of the last defined interval* is located and the entire process is iterated 4* times. The group with the largest number of members is chosen as the “largest peak.” In the event that two or more “equal” peaks satisfy these criteria, the closest* to the origin is arbitrarily chosen as the “largest” peak.

4) A “typical” sample with Tanimoto distance from the reweighted artificial average sample equal to the distance to the peak value of the “largest” peak is selected as a candidate for Archetype of the Branch, and the Tanimoto distance between it and all other normalized and reweighted samples are computed and the procedure in 3) is reiterated until the distance between the most current candidate for Archetype and the “largest” peak drops to less than $0.01 * \text{Tanimoto units}$ (corresponding to a Similarity Coefficient greater than 0.94).

5) The boundary of the “largest” peak is located and as many other minima (rather than 4* other—as specified in 3 above) are located as is necessary until the cumulative population included under the distribution curve out to the most currently located minimum is equal to or greater than 95%* of the population of the histogram. All samples between the origin and the boundary (as defined in 3) associated with the “most significant minimum”—i.e., that minimum such that the distance between \mathbf{D}' on the origin side of a minimum and \mathbf{D} on the side away from the origin, divided by the number of samples under the distribution curve between those two points is maximal—are entered into a different section of active memory as provisional members of the First Main Branch population, and the above most current candidate is elected Archetype of the First Main Branch. The Correlation Coefficient between all the normalized and reweighted provisional members of this Branch at Tanimoto distances from the normalized and reweighted Archetype greater than $0.75*$ units, are computed and all samples yielding values less than $+0.10*$ are excluded from membership in the First Main Branch. All samples at Tanimoto distances equal to or greater than $1*$ unit from the Archetype and less than $1*$ unit from the boundary as well as all excluded in the last operation are entered into a different section of an active memory as members of the First Main Residue Branch. The number of members in the Main Branch and the Residue Branch at the time of division are compared and the Branch with the largest membership is given the binary code name, **1**, and the other the code name, **0**. In the event that two or more “equal” minima satisfy the above criteria, the one which most nearly divides the population into equal branches **1** and **0** is chosen as the “most significant minimum.”

The procedures 3), 4), and 5) are iterated for each of the above Branches (and any created by this instruction) until no further subdivision by these rules (i.e., with the same set of \mathbf{x}_r 's and \mathbf{j}_{rc} 's) is possible. All such branches, though constituting different levels on the dichotomous tree, define the “First Taxonomic Level.”

6) Enter the next “new sample” into the active memory of the Trunk, recomputing $(\mathbf{S}/\mathbf{N})_r$.

a) If $(S/N)_r$ is now less than 999*, continue to add “new samples” until $(S/N)_r$ is equal to or greater than 1,000* and record a new artificial average sample and repeat steps 2) through 5).

b) If the $(S/N)_r$ remains equal to or greater than 1,000* after the addition of the “new sample”, *remove* the “oldest” sample from the active memory of the Trunk. Keep a record of $n_T = n_c$ plus the number of samples removed since the (most recent) subdivision.

Compute the Tanimoto distance between the normalized and reweighted First Main Branch Archetype and the *new sample* also normalized and reweighted with the same set of \mathbf{x}_r 's and appropriate j_{rc} 's as used to normalize and reweight the Archetype. Assign such new sample to the First Branch and/or First Residue Branch on the basis of instructions defining membership in 5), and add the new sample to the appropriate active memory or memories. Iterate with the appropriate sub-Branch(es) to which such sample is assigned *at the First Taxonomic Level* in the order of descending frequency of membership in the respective Branches.

c) Each time n_T increases by an increment equal to n_c^* , compute the Tanimoto distance between the artificial average sample consisting of the \mathbf{x}_r 's for the current contents of the active memory and the last recorded artificial sample from 1) or 6a) (but not 6c)), and if the distance is greater than 0.01* units, then repeat steps 2) through 6).

7) Carry running accounts of subpopulation parameters as in 1) and 6) and reiterate the instructions 1) through 7), now applied to the terminal subpopulations of a Taxonomic Level. The binary code word for the larger of the two First Sub-Branches of **1** will be **11**, and the smaller, **10**, and for the larger of the two First Sub-Branches of Branch **0** will be **01**, and the smaller, **00**, etc. Each time a renormalization and reweighting results in the *resolution* of additional sub-Branches in previously unsubdividable Branches, a new Taxonomic Level is generated. (The Taxonomic Levels may be designated, for example, by indicating the code for members of the First Taxonomic Level in upper case, the Second in lower case, the Third in upper case, etc., to provide this information by direct inspection, however, for the purposes of efficient processing of data, the added coding device is redundant and may be ignored.)

POPULATION STATISTICS AND THE ERGODIC ASSUMPTION

In designing the instructions for constructing the first and less efficient taxonomic classification, the assumption was made (implicitly) that consecutive samples would be picked at random from the parent population, and that the statistics of the population would be stable (i.e., that the probabilities of sampling the different kinds of patterns from the population were fixed for all time, thus defining an ergodic (10) source).

If the sampling process were not random (i.e., knowing the statistics of the source, and having collected one additional sample, we would be in a better position to predict to which Branch the next sample would belong than we were before the sample was collected), then the structure of that taxonomy, which depends upon the relative

frequencies of occurrence of the different patterns, might be a poor representation of the population.

The existence of both long-term evolutionary processes as well as short-term environmental changes guarantees that variations in our sampling cannot, even in principle, be random for all time and that the statistics of the population of all human serum protein patterns must change with time. We must therefore be in a position to detect changes in the sampling process and changes in the population statistics and have practical means of modifying the Taxonomy in a way that follows such changes as occur in order to maintain the usefulness of such a classification.

The $(S/N)_r$ and the intermittent checks on the similarity of artificial average samples provide sensitive indicators of changes in sampling statistics and/or population statistics. With a stable source and random sampling, the \mathbf{x}_r 's, Branch frequencies, $(S/N)_r$, and artificial average samples would all remain fixed and the updating subdivide instructions in 6a) and 6c) would never be activated. With an unstable source, only small $(S/N)_r$ can be used if *any short-term regularities* in the source statistics are even to be recognized.

INFORMATION THEORY AND TAXONOMIC PATTERN RECOGNITION

An examination of this problem from the point of view of Information Theory will be especially useful. If the population statistics are stable and each pattern is sampled independently of all others, each pattern is said to be a *discrete message*, and the parent population of samples is said to be a *discrete message source*. The First Binary Coding Theorem of Shannon (10, 22) states that, given a discrete message source which generates information at an average rate of \mathbf{R} bits per message, and given any δ , it is possible to arrange sequences of binary symbols to represent sequences of messages in such a way that, on the average, *less than* $\mathbf{R} + \delta$ output binary symbols are required to represent the average input message from the source, but this is achieved only at the expense of a coding delay which increases as δ approaches 0. It is not possible to find a complete representation of the source using fewer than an average of \mathbf{R} output binary symbols per message. If all resolved values (in our case 128) of all attributes (in this case 400) appear in the message source and all are "significant" then the message source would contain $2^{7(400)}$ kinds of messages. If all possible messages occurred with *equal* frequency (an "equal likelihood source"), then the maximum value that \mathbf{R} could take for any such message source would equal 2,800 bits. If the different magnitudes of the attributes occur with *unequal* frequencies, $\mathbf{R} < 2,800$ bits. If all messages are sampled at random (i.e., if the parent population is a discrete message source), and if in addition, the words of the messages (the attributes or rows) within a message are independent of one another, then the average self information per message,

$$\mathbf{R} = \sum_{r=1}^{n_r} \sum_{c=1}^{n_c} -f(\mathbf{k}_{rc}\mathbf{x}_{rc}) \log_2 f(\mathbf{k}_{rc}\mathbf{x}_{rc})$$

would be equal to R for large n_c . However, the definition of "message" implies, that words within a message are not independent and that the "meaning" or distinguishing characteristics of the message resides in these correlations. Therefore, while R will be less than 2,800 bits, it will usually be greater than the real value of R because at least some of the words (attributes) of each kind of message (pattern) will show some statistical dependence on one another, and the averaging over rows in computing R leaves out the weighting necessary to correct for this dependence.

The "mutual information," I_m (10, 22), a measure of the complex interrelationships (joint probabilities) among occurrences of words of messages, in principle provides the required measure for correcting the value of R . A measure of the information associated with the joint occurrence of *two* words, \mathbf{a} and \mathbf{b} in their respective places in a message, where the joint occurrence is designated $\mathbf{a,b}$ is $I_{\mathbf{a,b}} = -\log_2 f(\mathbf{a,b})$ and $I_{\mathbf{a,b}} = I_{\mathbf{a}} + I_{\mathbf{b}} - I_{m_{\mathbf{a,b}}}$. The latter equation defines the mutual information, $I_{m_{\mathbf{a,b}}}$, between *two* words, which is equal to zero if the occurrences of the two words in their respective places in the message are uncorrelated; is positive if the joint occurrence is more frequent than predicted in the uncorrelated case; and is negative if the joint occurrence is less frequent than predicted in the uncorrelated case.

Therefore, the correct value of R would only be obtained from R by the subtraction of the average mutual information per message (pattern) associated with the joint occurrence of all pairs, triplets, quadruplets, etc. of words in their respective positions in the population of messages. To calculate this average of the mutual information in a direct way. would require very large numbers of computations and an enormous memory to store all the joint probabilities.

When some prior knowledge about the "structure" of the source leads one to believe that only the correlations between *pairs* of words (or some *other* few combinations) are likely to be significant, the task is somewhat simpler. If some simple function can be shown to permit an approximation of the average mutual information between pairs of rows of such a source, a fairly accurate measure of R may be computed without much difficulty. (Such a technique will be discussed further on.)

In examining a new information source, such prior knowledge will not be available, and the problem of both estimating the value of R and developing the "ideal code" for such a source have appeared to remain beyond present capabilities. The taxonomic approach here presented would appear to change this situation.

SHANNON'S FIRST BINARY CODING THEOREM, AN EFFICIENT CODE AND THE "SEMANTIC PROBLEM"

Whereas the true value of R cannot be computed directly from the raw data, the frequency of membership in a Branch Tip, (n_{T_i} / n_{T_i}), multiplied by the negative logarithm to the base two of the frequency of membership in that Branch Tip, summed over all Branch Tips, converges to R , as n_T approaches infinity, if the Branch Tips are "natural" subgroups. As a result of the structure of his kind of taxonomy and as a result of

the simple instructions for coding its branches, it can be easily shown that the average length of the code words for Branch Tips (averaged on the basis of frequency of Branch Tip membership) also approaches R , the more nearly equal are the subdivisions at each Branching Point. Therefore such a taxonomic procedure is an explicit schema for the “construction” of a useful model of the two “black boxes” constituting the hypothetical “ideal” coder and decoder, the existence of which are predicted by Shannon's First Binary Coding Theorem. (These “black boxes” are also vital parts of the larger “black box” represented by a Turing machine (3).) In addition, the same sets of operations both “discover” or recognize the messages as well as “name” or code them in an efficient taxonomic code book or “dictionary.” This taxonomic procedure provides unambiguous (or almost unambiguous) operational definitions which specify the *meanings* of the class membership designated by the code name. Since the semantic problem of Information Theory is “concerned with the identity, or satisfactorily close approximation, in the interpretation of meaning by the receiver, (of a message) as compared with the intended meaning of the sender” (Weaver, 10), *such a taxonomic procedure also provides a solution to this semantic problem by generating a dictionary of operational definitions.*

If our taxonomic procedure generated a symmetrical dichotomous tree with thirty levels of branching and equal numbers of samples in each terminal branch (Branch Tip), there would be a total of 2^{30} Branch Tips (i.e., kinds of messages) on such a tree, and R would equal 30 bits per message. The code words generated by steps 5) and 7) would be made up of 30 binary symbols (binits) per pattern type. This would, in fact, be a perfect Shannon-Fano Code (44, 22). It is interesting to note that whether the tree were symmetrical or unsymmetrical (i.e., different numbers of branching levels on the way up to different Branch Tips) and independently of the naturalness of the Taxonomy, this code will coincidentally always satisfy the so-called prefix condition, (i.e., no code word for a Branch Tip would constitute a prefix for any longer code word for another Branch Tip), and therefore *continuous* sequences of such code words can be decoded unambiguously. Therefore there would be no need for punctuation code symbols between messages (which necessarily lower the efficiency of a code). Thus we see that even the imperfect Shannon-Fano Code automatically generated by this taxonomic procedure has a number of the desirable properties of one of the “best possible” codes, (e.g., a Huffman Code 45, 22), which satisfy the Shannon Binary Coding Theorem. In this case, the code is generated directly, rather than retrospectively, as in the classical Huffman Coding procedure. If, after resolving the Branch Tips of such a Taxonomic Tree, we find that the variance of the information per Branch Tip is high, (i.e., the average number of symbols per Branch Tip is much larger than R), and it is desired to provide a more efficient code, the Huffman Code for such a message source can then be generated in a straightforward manner. It should, however, be noted that while a Taxonomic Code is a significant code, a Huffman Code is a nonsignificant code (see page 446) Such a Significant Taxonomic Code has the very useful property that each prefix of a given code word, beginning with the longest prefix and ending with the first symbol in the code word, provides the “name”

of a successively more general parent population to which the Branch Tip belongs. In situations where the capacity of a “communication channel” is smaller than the rate of the information source, lopping off appropriate lengths at the ends of code words will permit us to match the source to the channel, with a corresponding but necessary loss in “semantic resolution”, *but without changing “dictionaries.”* By comparison, each such change in a Huffman Code requires the generation of a completely new “dictionary”. A Taxonomic Code will therefore be more convenient and useful than a Huffman Code and *will require a smaller overall delay in recoding and decoding when “semantic resolution” must be sacrificed to the Procrustean bed of channel capacity.* If the measure of code efficiency includes taking the variable delays involved in recoding into account, then a Significant Taxonomic Shannon-Fano Code or similarly derived efficient Taxonomic Codes can permit us to minimize recoding delay. From this more general frame of reference (i.e., taking variable channel capacity into account) meaning is *not* “... irrelevant to the engineering problem” (10).

DECREASE IN INFORMATION AT SUBDIVISION AS A MEASURE OF
“NATURALNESS” OF CLASSES

At each subdivision, it would be desirable to have some independent check as to whether the separation so far achieved, considering the large number of “arbitrary” boundary conditions (the * steps in the instructions), is anywhere near optimal. That is, at one extreme the subdivision may have no more significance than a random division of the parent population. At the other extreme, the subdivision may have separated two distinct “natural” classes of messages. In the first case, the “true” value of each R for each sub-Branch, which we shall designate R_1 and R_0 respectively, would be equal to the unknown R for the parent population. In the second case, R_1 and R_0 would each be smaller than R for the parent population, and $R - ([n_{T_1}R_1/(n_{T_1} + n_{T_0})] + [n_{T_0}R_0/(n_{T_1} + n_{T_0})])$, the difference between the information content of the parent population and the average information of the two Branch populations (which is also the average decrease in information per subdivision), would be equal to one bit if $n_{T_1} = n_{T_0}$, and equal to less than one bit for unequal frequency distributions. *If the subdivision is into natural subclasses*, and if Branches **1** and **0** are considered as Branch Tips, R_1 and R_0 would each equal 0, and R would equal

$$([-n_{T_1}/(n_{T_1} + n_{T_0}) \log_2[n_{T_1}/(n_{T_1} + n_{T_0})] + [-n_{T_0}/(n_{T_1} + n_{T_0}) \log_2[n_{T_0}/(n_{T_1} + n_{T_0})]) .$$

We shall designate this measure as M_R . In this special case, the difference between the information content of the parent population and the average information of the two Branch populations, (see above), will also equal M_R . For a given population, as different boundary conditions, (*), are tested, this value of the decrease in information per subdivision, calculated from the frequency of membership in the sub-Branches, can be compared to

$$M_R = R - ([n_{T_1}R_1/(n_{T_1} + n_{T_0})] + [n_{T_0}R_0/(n_{T_1} + n_{T_0})]),$$

(where the \mathbf{R} 's are calculated as on page 470) and the difference, $\mathbf{M}_R - \mathbf{M}_R$, can be used as a figure of merit. The boundary conditions, (*), which give the smallest difference would, in principle, be the “best” and least arbitrary. Even when $\mathbf{R} > \mathbf{R}$, $\mathbf{R}_1 > \mathbf{R}_1$, and $\mathbf{R}_0 > \mathbf{R}_0$ $\mathbf{M}_R - \mathbf{M}_R$ should still be equal to or less than 1. The redundancy shared by the parent subpopulations which appears in the \mathbf{R} 's (already discussed on page 471) is removed in computing a difference such as \mathbf{M}_R . With two natural subpopulations with greater redundancy among the rows of the individual subpopulation than in the parent population, \mathbf{M}_R often will be greater than 1, clearly identifying the contribution of the redundancy, and the difference, $\mathbf{M}_R - \mathbf{M}_R$, will then be negative. Only in so far as the statistical fluctuations (noise) in the \mathbf{R} 's are large in comparison to the magnitude of \mathbf{M}_R , will it be a poor measure. Therefore, in the limit, in the case of “natural” subdivision, as $n_{T_1} + n_{T_0}$ becomes very large, $\mathbf{M}_R - \mathbf{M}_R$ will assume a stable value, equal to or less than 0.

Using this figure of merit to provide negative feed-back, the “efficiency” of the subdivision process can be continuously monitored, and “reasonable” boundary conditions which minimize this measure can be discovered at the very earliest stages of the construction of the Taxonomy.

This general approach is most closely related to Tanimoto's use of his entropy (Information) measure for *locating* boundaries between groups (see Rogers and Tanimoto (32), and also R. V. Smith (46)).

INFORMATION SPACE, THE INFORMATION TREE AND CONVERGENCE

Among the many possible spaces one may consider for the purposes of taxonomic classification, there is one, which has ideal properties. We call this ideal space *Information Space*. The distance between two messages (columns), i and j , in Information Space, \mathbf{D}_{ij} , we call *Information Distance* ($\mathbf{D}_{ij} = \mathbf{D}_{ji}$).

Suppose that a dichotomous taxonomy of “natural” classes has been generated in some unspecified way using \mathbf{D}_{ij} as our measure of similarity, and that the values of the \mathbf{R} 's computed from Branch frequencies are the “true” values, (i.e., $\mathbf{M}_R = \mathbf{M}_R$). We then define \mathbf{D}_{ij} as follows:

$$\mathbf{D}_{ij} = I_{c_i} + I_{c_j} - 2T_{c_i c_j}$$

where $T_{c_i c_j} = -\log_2(n_{T_F}/n_T)$, and where F is the code for the first branching point backwards towards the Trunk which is common to the sub-Branches $\mathbf{u1}$ and $\mathbf{w1}$ (to which i and j respectively belong), and $I_{c_i} = -\log_2(n_{\mathbf{u1}}/n_T)$ and $I_{c_j} = -\log_2(n_{\mathbf{w1}}/n_T)$

\mathbf{D}_{ij} is therefore twice the difference between an “average self-information of a pair of messages” and $T_{c_i c_j}$ (which is similar to a mutual information in that it corrects this “average information” by a measure of “taxonomic correlation”). For messages for which the branching point, F , is the Main Trunk, $T_{c_i c_j} = -\log_2(n_T/n_T) = 0$.

Consider columns \mathbf{a} and \mathbf{b} which are both members of “natural” Branch $\mathbf{1}$ (i.e., $F = \mathbf{1}$) and \mathbf{a} is a member of “natural”, “noiseless” Branch Tip $\mathbf{11}$ (i.e., $\mathbf{u1} = \mathbf{11}$) and \mathbf{b} , of “natural” “noiseless” Branch Tip $\mathbf{10}$ (i.e., $\mathbf{w1} = \mathbf{10}$) respectively.

$$\mathbf{D}_{\mathbf{a},\mathbf{b}} = I_{c_{\mathbf{a}}} + I_{c_{\mathbf{b}}} + 2\log_2(n_{T_1}/n_T)$$

where $n_{T_1} = n_{T_{11}} + n_{T_{10}}$ and $I_{c_{\mathbf{a}}} = -\log_2(n_{T_{11}}/n_T)$ and $I_{c_{\mathbf{b}}} = -\log_2(n_{T_{10}}/n_T)$, therefore,

$$\mathbf{D}_{\mathbf{a},\mathbf{b}} = -(\log_2[n_{T_{11}}/(n_{T_{11}} + n_{T_{10}})] - \log_2[n_{T_{10}}/(n_{T_{11}} + n_{T_{10}})])$$

We define distance between **1** and **11** (between a Branch and its sub-Branch) in Information Space in this case as

$$\mathbf{D}_{\mathbf{a},\mathbf{ab}} = = \mathbf{D}_{11,1} = -\log_2[n_{T_{11}}/(n_{T_{11}} + n_{T_{10}})]$$

and

$$\mathbf{D}_{\mathbf{b},\mathbf{ab}} = \mathbf{D}_{10,1} = \mathbf{D}_{1,10} = -\log_2[n_{T_{10}}/(n_{T_{11}} + n_{T_{10}})]$$

so that

$$\mathbf{D}_{\mathbf{a},\mathbf{b}} = \mathbf{D}_{11,10} = \mathbf{D}_{11,1} + \mathbf{D}_{10,1} .$$

Generalizing, $\mathbf{D}_{\mathbf{i},\mathbf{j}}$ will be the sum of the distances from branching point to branching point, measured over the shortest path following the branches of the tree from the Branch Tip of **i** to that of **j**. Such a binary *Information Tree* may thus be constructed “to scale” where the segments from branching point to branching point are of length $\mathbf{D}_{\dots 11, \dots 1}$ and width, $n_{T_{\dots 11}}/n_T$ etc.

Because we do not know how to compute Information Distance until after the Taxonomic Tree has been generated, we must use other distances, such as Tanimoto distance and Correlation distance between reweighted samples, as approximations to Information Distance. From this frame of reference, \mathbf{M}_R is the average Information Distance of a Branch from its sub-Branches, and therefore our figure of merit, $\mathbf{M}_R - \mathbf{M}_R$ may be interpreted as a measure of how closely we have approximated Information Space, and can also be shown to be a reasonable basis for establishing convergence to a “most natural” classification, (see page 456).

SOME SPECIAL DEVICES FOR SPECIAL PROBLEMS

The following procedure should increase the confidence that can be placed in the significance of such a figure of merit for this special case of electrophoretic patterns of serum proteins. The principle can be extended to cover sources containing considerably higher orders of complexity of joint probabilities among words, perhaps even the words in English sentences. There are a few known properties of the information source consisting of the disc electrophoresis patterns of human serum proteins which are especially relevant at this point:

1) Protein discs have finite widths, usually of approximately Gaussian “crosssection” and greater than 1/400 of the total pattern; therefore adjacent rows in a pattern will be highly correlated and provide considerable redundancy.

2) In some genetic systems—for example, the case of the haptoglobin allelic genes 1,

and 2—two genotypes, 2-1, and 2-2 produce sets of different proteins. If one of a set is present, all are generally present (47), again providing added redundancy as in 1).

3) In some genetic systems, again using the case of the haptoglobin allelic genes 1 and 2, each of the three possible common genotypes, 1-1, 2-1, and 2-2, produce entirely different serum proteins; therefore if the proteins of one type are present (a particular set of rows is occupied) the others must be absent (another particular set of rows will be—to a first approximation—unoccupied) and, the full set of rows occupied by the haptoglobins are therefore completely redundant, that is, given the fact that one of the set of proteins is present in a pattern, the presence or absence of all the other members of the set is determined.

4) Linked genes, in cases where the linkage has been evolutionarily stabilized (for example, by a chromosome inversion) will produce sets of proteins such that when one is present, the others will also regularly be present, and when one is absent, the others will be absent, thus providing another source of redundancy.

5) In most cases, (e.g., genes like the Transferrin gene) individual alleles result in the production of single proteins, and a heterozygous individual produces both proteins. No more than two proteins of an allelic series can normally be present in any one individual. This results in a more complex kind of redundancy which will only add appreciably to \mathbf{R} if more than two alleles of a gene occur with high frequency.

6) Certain pairs of proteins (e.g., hemoglobin and the haptoglobins, and a post-albumin and some pre-albumins, form complexes under conditions which depend on the concentration of these proteins in the serum. Thus, when both are present, the complex may or may not be present in addition to, or instead of, the individual proteins. The presence of such complexing phenomena also increases the redundancy of \mathbf{R} .

1), 2), and 3) are responsible for the major part of the redundancy from row to row. Clearly, if only their effects on \mathbf{R} can be reduced or eliminated, \mathbf{M}_R will be a very much more sensitive measure of the efficiency of subdivision.

Consider the magnitude of the Correlation Coefficient between two rows. For a pertinent pair of rows in cases 1), 2), 3), or 4), the value will be very close to 1 or -1. Since $\mathbf{I}_{a,b} = \mathbf{I}_a + \mathbf{I}_b - \mathbf{I}_{m_{a,b}}$, if the average self-information per word for a particular row is,

$$\mathbb{I}_{r_i} = \frac{nc}{1} - f(krcxrc)r \log_2 f(krcxrc)r, \text{ then } \mathbf{I}_{m_{r_a, r_b}} = |C_{r_a, r_b}| (\mathbb{I}_{r_a} + \mathbb{I}_{r_b})/2,$$

where C_{r_a, r_b} is the Correlation Coefficient between rows a and b , and therefore,

$$\mathbb{I}_{m_r} = \left[\sum_j |C_{r_i, r_j}| (\mathbb{I}_{r_i} + \mathbb{I}_{r_j})/2 \right] / \left[1 + \sum_j |C_{r_i, r_j}| \right]$$

and

$$\mathbf{R} = \mathbf{R} - \sum_j \left[\sum_j |C_{r_i, r_j}| (\mathbb{I}_{r_i} + \mathbb{I}_{r_j})/2 \right] / \left[1 + \sum_j |C_{r_i, r_j}| \right] = \mathbf{R}'$$

This computation will remove the bulk of the redundancy in this kind of message

source. It also incidentally provides the means for discovering allelic series of genes of the type described in 2) and 3) and for discovering stable linkage groups as in 4). A “three-way” correlation coefficient would permit the discovery of normal allelic series such as 5) and proteins involved in complexing phenomena such as 6), and would provide a means for computing a value still closer than \mathbf{R}' to the true value of \mathbf{R} . Further exploration of such correlation measures (48, 49, 50, 51) as means for computing higher orders of mutual information would seem to hold some promise for the broader application of such taxonomic programs. For clues to ways of equating, by an appropriate transformation, a measure like the Correlation Coefficient, which is an estimate of the extent of linearity of the relationship between the magnitudes of the dimensions of sets of data, to an estimate of information content like the mutual information, see Linfoot (52) and Kolmogorov (53).

INFORMATION THEORETICAL CONCLUSIONS

To summarize the significance of this Information Theoretical interpretation:

a) The “naturalness” and efficiency of the taxonomy generated by this technique, using any approximation to Information Distance as a measure of similarity, can be measured by means of $\mathbf{M}_R - \mathbf{M}_R$ and maximized from the very outset of the subdivision process.

b) A Taxonomy generated in a manner similar to that outlined provides a means for processing raw data in a way that results automatically both in the “discovery” and characterization of the different kinds of messages from the message source and appears to condense the information contained therein into a form approaching the limit set by theory, thereby maximizing the potential rate and accuracy with which such information can be transmitted and used. In this way, the problem of meaning, i.e., the “semantic problem,” is also solved.

c) In so far as the sampling process departs from random sampling of a stable source, the frequencies which ultimately determine the structure of the binary code of the Branch Tips will change, the average information per message will change, and the proper set of normalizing “constants”, the \mathbf{x}_r 's will change. If the original subdivision and coding are not altered in the light of such change, the value of δ will increase and the code will become less and less efficient. From this frame of reference, arbitrary magnitudes of loss of efficiency, beyond which loss of confidence requires that the rules for resubdivision and recoding be activated, prevent our exceeding a maximum acceptable loss in efficiency. Such confidence levels are set by instructions 6a) and 6c).

d) In so far as the source itself changes its statistics, the “naturalness” of Branches may change, requiring “pruning” of some Branches and the growth and further branching of others. The continuous updating of the Tree by addition of new samples as prescribed in the instructions provides for the necessary growth. Redivision and recoding based on updated frequencies and \mathbf{x}_r 's take care of pruning and also are provided for by instructions 6a) and 6c). Thus cues and mechanisms for “reeducation of the computer” are built into this scheme.

For all message sources other than absolutely stable ones, maximum efficiency will

only be approached in proportion to the stability (in time) of the source and sampling procedure. In the limit, for a source which is (or seems to be) continuously, rapidly and *randomly* changing its statistics, “education” is hopeless and little useful analysis of samples (messages) from such a source is possible.

INDUCTION, HYPOTHESIS, THEORY AND LAW OF NATURE

At this point, I would hope that the reader has developed some confidence in some of the potentials of this relatively objective pattern recognition scheme. But the development of a special purpose technique which looks at simple one-dimensional objects may seem to fall far short of my promise to deal with the “broad potentials for machine learning via automated inductive processes” since these should at least include mechanisms for discovering “causal” relationships, scientific theories and laws. In the following sections, I will try to indicate that the solutions offered by this primitive model also promise to help us to unravel most of these problems. I will try to reformulate pertinent familiar concepts within operational frames of reference which place a minimum number of constraints, derived from their more classical formulations, on their structures. Such definitions may initially appear quite unfamiliar and perhaps even in conflict with ordinary usage, but I believe sober reflection will usually result in provisional acceptance.

We will begin by asking, “What connection does the discovery of classes of protein patterns have with the generation of hypotheses?” (See 41, p. 8).

There is a set of processes closely associated with inductive inference; first, speculation, or in its more dignified form, formulation of an hypothesis; second, establishment of an hypothesis or operational definition; third, establishment of a theory, and finally, of a Law of Nature, all of which will be defined below:

a) In its most primitive and fundamental form, formulation of a simple hypothesis requires the selection of a set of objects, events or attributes from a larger population. These objects are then arranged into a class or classes, each of which is usually assumed (explicitly or implicitly) to contain other as yet unobserved members very “similar” to the members of the finite sample. As an article of faith* in the frequency doctrine, (the metaphysical foundation of all inductive processes which has so far usually permitted useful reasoning from “the part to the whole” and from the past to the future), the probability that the sample is representative of the larger parent population is assumed to approach 1, the larger the sample and the more nearly random are any variations in the sampling technique. (A description of the sampling technique itself provides the operational definition and boundaries of that part of the universe which is called the larger, “parent population,” (i.e., a description of what corresponds to the “beginning” and “end” of a message), and therefore defines the message source. Were all possible boundaries in the universe observed, and were the sampling techniques completely

* This is a necessary concession (54, 55) to David Hume's famous critique of inductive inference and is at the heart of all so-called heuristic approaches.

random, then the parent population would be the class of all classes*) It is then guessed that the objects belonging to such a class can be “proven” to be related to each other by objective (but ultimately imprecise) comparison of their attributes. (For conflicting points of view, see for example 8, 58, 59.)

b) Observation of the attributes of a new sample of members drawn from the larger population, together with some estimation of the confidence with which this observation appears to confirm the guess, (often in the form of what we call an experiment), constitutes a test of a simple hypothesis. (The test of a more complex hypothesis may involve drawing the new sample from a still different but “related” population.) If the magnitude of confidence exceeds some arbitrarily agreed upon level, we usually say that the hypothesis has been “confirmed.” The combination of a) plus “confirmation” by b) is equivalent to the tentative establishment of an operational definition.

c) We usually dignify an hypothesis or an operational definition which concerns a very comprehensive class of objects, events or attributes by calling it a theory.

d) A Law is a theory which appears to have been established with a great degree of confidence. †

From this point of view, the generalized method—or methods, for the generation of operational definitions could include the methods for the generation of hypotheses, theories and Laws of Nature. Let us therefore examine the operational definition more closely.

OPERATIONAL DEFINITIONS AND DISCOVERY

An operational definition is usually a list of the attributes common to individuals of a class which is designated by a common symbol or name. These attributes must be able to be observed or conceived by means of prescribed physical or logical operations within the limits of precision set by these operational techniques and/or by the limits set by the

* As pointed out by Wigner (56), science “aims only at the discovery of the laws of nature, that is the *regularities* of events.... We have ceased to expect ... an explanation of *all* events . . .” (italics mine), and it appears that it is the “arbitrary” restriction of our field of view or attention to something less comprehensive than the class of all classes that makes it possible to learn anything at all. Our limited experience of the class of all classes is through what must be a non-random sample since the accumulation of even the very beginnings of a “representative” sample of *this* class must take an infinite amount of time.

R. J. Solomonoff (57) has also addressed himself to the general problem of prediction. He seeks solutions by searching for codes which parse a continuous string of symbols, representing the total of past experience, in a “most efficient way”. Were an efficient technique developed for implementing his procedure, it might then be possible to define all the boundaries of “messages” at all levels, i.e., elementary symbols, “words,” “sentences,” “paragraphs,” “chapters,” “books” and “reas of knowledge” in an unambiguous way. His model may therefore provide a frame of reference for defining what might reasonably be meant by Wigner’s “regularities of events.” This problem, when looked at from the perspective of pattern recognition, turns out to be the general problem of defining “correspondence” (see page 447).

† “Laws of Nature” as used here include relationships such as the Conservation Laws as well as relationships among such relationships, e.g., the geometrical and dynamical “Invariance Principles” (56).

logical structure of our concepts (e.g., by the Heisenberg “Uncertainty Principle” (54), Godel’s “incompleteness proof,” etc. (60)). In so far as some attributes may also be shared by members of other classes, an explicit hierarchy of attributes, in order of significance, is usually part of a definition. Likewise, the hierarchical relationship of the defined class to any more comprehensive class or classes is either stated or implied in a definition.

An operational definition of a class of physical entities can be generated in two ways—each of which has a special relationship to common experience:

The first is closely related to the classical idea of the scientific method. It relies, in part, on previously accumulated knowledge of the universe. It may be identified as a “sharpening” of description or definition. It involves a class with known members which has already been informally described on the basis of subjective or vague criteria—and has been previously named. Further observation of additional members is required to determine the measure of confidence that can be placed in the definition.

The second method for generating an operational definition can be most easily identified with the process of “discovery.” It requires no previous “experience” of the particular class to be defined. A “sorting” operation which discovers the class provides the information equivalent to designation of the named class and sample of known members called for in the first method, and, at the same time, may provide some explicit measure of confidence. In both cases, observation of the attributes of the members of a sample provide the “description”—in one case, usually as a hierarchical list of attributes—in the other, for example, as the maximum allowable spread in magnitude of some measure of similarity to some “typical” member of the class, (i.e., the operational specification of a decision boundary) .

My reasons for focusing on the differences between these two methods of generating operational definitions is related to some apparent misunderstandings about what constitutes a valid scientific test.

For the scientific investigator, with his limited human ability to translate mental images or “visions” into words, the “subjective reality” of certain mental images inspires (rightly or wrongly) a level of confidence which is not directly and unambiguously communicable to others. He searches for a tentative communicable operational word picture or description of this image. This is his hypothesis. The external test of the hypothesis is designed to provide a basis for an unambiguous measure of confidence in this description in the minds of others, (as well as to increase his own confidence).

Because of the difficulty in communicating detailed mental images without error, it has been the classical “anti *ad hoc* argument” that “one cannot test an hypothesis with the identical set of data which was used to originally suggest it.” This position can be supported as follows: In the case of a faulty hypothesis, the commission of identical errors of omission or translation is likely to instill, in the external observers, the same false confidence in the hypothesis as inspired its originator. In addition, it will usually require considerably more verbal or numerical data than is extractable from the original data set alone to provide a communicable level of confidence comparable to the

subjective one (for example, in identifying Mr. X, see page 441; i.e., “one picture is worth a thousand words”).

This classical constraint might lead one to falsely conclude that the same set of data cannot be used both for discovery and for “testing” an operational definition of a class by an objective sorting procedure. But the crux of the problem of “establishing” operational definitions by either the classical or the sorting routes, is the explicit determination of an objective, unambiguous and communicable measure of confidence in the definition. When a sorting procedure simultaneously provides a completely objective measure of confidence from the same set of data that permitted the discovery of the class, then the classical taboo of the scientific method will not apply.

MEASURES OF CONFIDENCE, TESTS OF HYPOTHESES AND EXPERIMENTS

Most tests of hypotheses are designed to increase the level of confidence in the description of the class which is defined by the hypothesis, and the simplest “test” involves nothing more than repeated observations of the kind which provided those original data which led to the discovery of the class. Such “tests” generally result in an increase in confidence at a rate which is a function of the square root of the number of class members observed (see footnote, page 466).

An initially low communicable level of confidence in an hypothesis stems from either one or both of the following:

a) Observation has so far produced so few members that the (S/N) is also very small and therefore the confidence in the description (e.g., the \mathbf{x}_r 's and decision boundary), even within a broad confidence interval, is very low.

b) Even though the number of members belonging to the class which is defined by the hypothesis is moderate, the standard deviation of the magnitudes of the attributes of the members is so large that the (S/N) is still quite small.

If the random sampling of the parent population has resulted in only a small accumulation of samples in the particular class, i , even after a great deal of overall sampling, (i.e., very large n_T , very small n_i/n_j), in general, to increase n_i by some factor, K , by further random sampling of the parent population will require increasing n_T , by approximately the same factor K . This is clearly a slow and uneconomical procedure when the explicit goal is an increase in confidence in the particular class, i .

One resorts to the kind of test called an experiment in order to increase confidence at substantially greater rates.

One class of “experiment” consists in applying a new non-random sampling technique which ideally provides an increase in n_i at an expense no greater than that for an equal increase in n_T . Such experiments may differ from the original random sampling of the parent population by virtue of a sharp restriction in the “field of view” of the environmental transducers (measuring or sensing devices) to a more limited portion of the parent population and may be accomplished by the use of more restricted means of defining “correspondence of attributes” and/or the use of “more selective” kinds of transducers and/or the incorporation of special normalizing techniques as part of the transducer instrumentation. It should be noted that the mechanisms of a taxonomic

pattern recognition technique (normalization, reweighting, measures of similarity and techniques for defining decision boundaries) together may be interpreted as specifying the design of a highly specific, ordered set of “sorters,” “sieves” or “filters” for efficiently separating the members of many classes from the mixture that is the parent population. A well designed experiment may therefore be interpreted as involving taking the relationships already discovered among the attributes of the observed members of the class and “related” classes as well as the statistics of such classes as bases for the design or “invention” of a *special purpose filter*. With such a filter, one hopes to efficiently sieve new members of the particular class from a portion of the parent population—ideally in a single “sweep.”

Another kind of experiment can involve an attempt to observe the parent population or its members at a higher level of resolution, and/or over an expanded range of magnitudes of the already known attributes, and/or with the examination of “new” additional attributes not included in the original observations of the parent population, and/or with new normalizing techniques, new similarity measures and/or new decision boundary criteria.

The first class of experiments is mainly aimed at efficiently acquiring a large increase in n_i . In so far as large standard deviations (in case b) stem from “errors in measurement” rather than “true” variations in the attributes of the members of class i , such an experiment may secondarily also lead rapidly to a high $(S/N)_i$ and a high level of confidence if the special purpose filter introduces smaller errors of measurements.

The second class of experiments is aimed directly at reducing the standard deviation by either reducing the weight of variation among the magnitudes of the original set of attributes in the hope that the new set is both characteristic of the class i and relatively homogeneous, and/or by reducing the weight of errors of measurement, and/or by reducing the magnitude of the errors of measurement.

I believe that analysis of all kinds of scientific experiments (with the exception of one remaining class to be discussed a little further on) can be reduced to either one or a combination of the two classes just discussed. Such analyses suggest that experiments can, in principle, be performed by “aiming” a special set of environmental transducers (“sense organs”) at a restricted subset of the class of all classes and subjecting their outputs to analyses by a special purpose taxonomic pattern recognition program, designed on the basis of past experience with the particular subset and “related” classes. Experimental routines are therefore reasonably compatible with the operations of a general purpose taxonomic pattern recognition machine.

(We have in fact proposed to test the diffuse hypothesis, “The set of soluble, circulating, direct transcriptions of the genetic code that are the serum proteins, will vary in kind and concentration as a function of the physiological states of the human organism,” with an experiment that consists in “aiming” our pattern recognition program at a very restricted subset of the class of all classes.)

The economy gained by this process of narrowing our field of view is bought at the price of the introduction into the inductive scheme of a system of values related to

specific intentions, goals or “purpose” (and elevates us to the Third level of Information Theory (10)). To choose a particular class that we consider warrants the special attention of an experiment from among the large number of classes which have already been “discovered” (by, for example, an “unbiased” pattern recognition device as it selects samples “at random” from that part of the universe most proximal to it), implies that this class has higher intrinsic value (for example, higher adaptive value) than those other classes which have failed to receive such “privileged” treatment.

Therefore, to introduce “experiment” as an automatic part of a general purpose inductive taxonomic program will require providing the machine with an “acceptable” set of operationally defined “goals” or “values.” This can provide a general purpose taxonomic pattern recognition machine with the flexibility to “specialize” and considerably increases its versatility and utility at a relatively early stage of its “education.” But it must be kept foremost in our minds that it is just at this point that the possibility of the choice of “an unfortunate” set of values could pose an even more serious threat to mankind than, for example, the choice of anti-social values as the guiding principles for a “mere” human.

CONFIDENCE IN HIGHLY “STRUCTURED” THEORIES

One important aspect of the scientific method which has so far been intentionally neglected in this treatment concerns what might be considered to be still another kind of scientific hypothesis or theory. Such an hypothesis represents something very close to a mathematical theorem, the “form” of which has been perceived by the investigator and has been expressed in verbal or symbolic form. The “test” of such an hypothesis often depends more upon the logical manipulation of relations in which we may, from past experience, already have a great deal of confidence, (as axioms and fundamental propositions are manipulated logically in the proof of a theorem), than on a new observational test. This “mathematical” kind of hypothesis building and testing is more akin to chess playing and language manipulation. It constitutes a very important element in science and is the area where the application of the computer has so far been most promising (41, 61). To summarize, the building of confidence in such theories or hypotheses properly relies on logical deduction from “established fact” perhaps even more than on new empirical induction. By contrast, beautifully consistent, logical castles built on patently weak foundations (i.e., we have low confidence in the premises) demand independent direct verification before they are even likely to be given any attention by a knowledgeable scientific community. Such theoretical construction plays a major role in the so-called method of strong inference (58, 62). This explains the third function of experiment in the advanced sciences. Such an experiment is designed to provide a new kind of observation which will provide an explicit objective measure of confidence to substitute for and/or supplement logical “confirmation.” (When the premises have been only weakly established, a successful experiment can also provide increased confidence in the “axiomatic elements”.)

Because weakly supported hypotheses are by far the more numerous in our experience, many scientists do not easily recognize that there is, in fact, a distinction to be made and they are therefore regularly more “comfortable” with those hypotheses which are directly supported with empirical as opposed to logical tests or “confirmation.” This is better understood when we remember that “logical confirmation”, so far, usually depends upon a less explicit and therefore more ambiguous measure of confidence acquired by mental manipulation of the joint confidence in sets of earlier observations (e.g., those confirming Laws of Nature). To illustrate my point; 1) Although Einstein replaced a growing mountain of *ad hoc* and often contradictory hypotheses (63) with a comparatively simple and much more comprehensive theoretical structure, many initially had little confidence in his Theory of Relativity because it depended upon “deductions from” fundamental invariance principles which were derived, as a set of “axioms,” from the already empirically established high confidence in Newtonian mechanics, Maxwellian electrodynamics, the measured “constancy” of the velocity of light in “vacuum” (independent of the velocity of the frame of reference relative to other frames of reference), and the measured “equivalence” of inertial and gravitational mass. 2) Since “confirmation” with results from a computer is not yet available, many will have little confidence in a substantial part of what appears in this manuscript because any attempt to derive a comfortable measure of confidence by mental manipulations must be non-numerical and vague. This is necessarily true because some of the component elements must be derived from a) the confidence in the past successes of Information Theory, b) confidence in those more remote empirical bases in experience with the technology of communication and with those analyses of language from which Information Theory has been derived, and c) from the non-numerical personal confidence of the individual readers in, for example, the relevance of his own observations of the way he and others classify and process data to the infant-taxonomist analogy of my introduction.

Yet, in contrast, the explicit numerical levels of confidence which can be recorded for those class descriptions which may be provided by a taxonomic computer program are, in principle, amenable to quantitative manipulation to produce net measures of confidence in the combined observations. For example, were all words of all messages independent of one another, it can be easily shown that the weighted root-mean-square value of the Signal to Noise Ratio for words (rows), $(S/N)_r$, (see page 466), provides a reasonable quantitative basis for ascribing an unambiguous measure of joint confidence in the overall mean value of the \mathbf{X}_{r_i} 's.

Since words of messages are usually correlated, the weight of the individual values of the $(N/S)_r$'s must be reduced by some measure of the average mutual information per word (as on page 476), in order to correct for any redundancy which would otherwise lead to falsely high values of $(S/N)_r$. It can be shown that a more accurate value of $(S/N)_r$

which takes such redundancy into account is,

$$(S/N)r_i \left[\begin{array}{c} \left[\begin{array}{c} 1 - \frac{r_j |C_{r_i, r_j}|}{1 + r_j |C_{r_i, r_j}|} \end{array} \right] \\ r_i \end{array} \right] \cdot \left[\begin{array}{c} \left[\begin{array}{c} 1 - \frac{r_j |C_{r_i, r_j}|}{1 + r_j |C_{r_i, r_j}|} \end{array} \right] \\ r_i \end{array} \right] (N/S)r_i^2 \cdot^{-1/2}$$

Such kinds of manipulations can provide the required explicit measure of joint confidence in multiple observations with varying individual levels of confidence while at the same time correcting for redundancy (64). Such measures could substitute for the at present more ambiguous measures of confidence which must be used to judge those theories which are operational definitions of comprehensive classes (of still more primitive but well characterized classes) that have not yet been independently tested by appropriate new observations (by experiments). Indeed, such derived measures will often provide sufficient confidence to permit the useful by-passing of much (sometimes all) repeated independent testing by new kinds of experiments. This has often been the case with careful and knowledgeable engineering design from principles for which past experience has provided a great deal of confidence. Similarly, some physical theories, such as the Theory of Relativity, are grounded on an even more extensive, comprehensive and therefore more convincing, observational base and have, as a result, required relatively few new observations to establish them quite firmly.

This position contrasts with a rather widely and I believe wrongly held view that, “The idea of determining the numerical value of the probability of scientific theories seems preposterous” (65), because as I have suggested above, such measures of confidence could be provided both from direct relevant observation and/or indirectly through estimation of the joint measure of confidence in those pertinent, “more primitive” observations which are to serve as a base for those operational definitions which function as the fundamental propositions of a “mathematical”, scientific theory.

My efforts on these pages have been largely directed at shoring up confidence in the promising possibilities of machine approaches to inductive processes, albeit, mainly by the use of what, though sketchy, I hope will be found to be logically consistent arguments deduced from observationally rooted premises. In the process, I hope I have incidentally helped to destroy some of the remaining apparent magic in science, and have perhaps provided another strand of understanding for a bridge between the “Two Cultures” (66, 67).

PATTERNS OF HIGHER DIMENSIONALITY

Given the raw data equivalent to the experience which has been the source of previously accumulated knowledge, an appropriate sorting method for generating operational definitions will, in principle, generate definitions for all the classes which have been generated by the classical scientific method—but with a much more

comprehensive and clean-cut taxonomy. To repeat all analyses of the past would be extremely time consuming, but when faster computers become available, the devotion of a number of years to the “education” of a general purpose computer may become reasonable. An appropriate sorting method for generating operational definitions will probably be found to be closely related to the newborn’s learning process; is directly related to Dr. Tanimoto’s “Elementary Mathematical Theory of Classification and Prediction” (31); and I believe is, in fact, something very close to the procedure which I have outlined for the discovery of protein pattern classes. The definitions generated by such a procedure become formally equivalent to classical simple definitions if one uses a measure of similarity between rows of attributes for each class (using the normalization and reweighting appropriate to the next most general Branch population from which it was derived) to generate each appropriate descriptive hierarchical list of attributes.

Each serum sample provides us with a one-dimensional pattern, equivalent, for example, to an electrocardiogram, to an absorption spectrum, to a single line in the raster of a two-dimensional television image, etc. In so far as the techniques outlined are of a fundamental nature, the same approaches should be fruitful in the recognition of two-dimensional images—for example, pictorial representations such as microscopic images of cells (in which I am particularly interested), alpha-numeric symbols and words in printed and written texts, phonograms, etc.

The problems involved in going from 1 to 2 dimensional patterns raise formidable, but I believe solvable, topological complications, (68, 69) . If the solution of these is achieved at a level which keeps the problem of correspondence tractable (e.g., see 70), recognition of n-dimensional patterns looks extremely hopeful.

The visual and auditory pattern recognition apparatus possessed by most mammals appears to permit a kind of recognition and discrimination of, for example, face and voice that closely matches human capabilities. For example, most intelligent dogs can recognize a large number of “friends” by sight or sound (as well as “smell”). The evolution of this complex apparatus (sense organs and neural structures) occurred over a period of more than 100 million years. It appears that the small amount of mutational innovation and evolutionary remodeling that can occur in about a million additional years was sufficient to take man across a new data processing threshold. This gave him his meager ability to translate his “high resolution” mental images into sets of symbols for social communication that are somewhat more sophisticated than those of other mammals. Yet despite its very primitive stage of development, this ability to translate (page 422) has apparently taken him most of the additional distance he has come.

For some people, “thinking” mainly involves mental verbalization of both a problem and its detailed solution, but for at least a few of our most creative thinkers (e.g., see (71), pages 83-99 and Einstein’s letter in Appendix II), most of their thinking apparently involves mental manipulation of non-verbal patterns. One wonders whether the common admonition to “think things out verbally,” by perhaps confusing the translation problem with the “analytical” problem, may not have extensively handicapped the formally educated population of the world by largely restricting the bounds of human imagination to the inefficient and low resolution domain of digital symbol manipulation which

follows the "translation step." We have the very common handicap to rapid and comprehensive reading (i.e., at 1,000 to 2,000 words per minute as compared to 200 to 400 words per minute, see for example (72, 15)), that appears to stem from such poor reading habits as resorting to "sub-vocal" "inner speech" (i.e., "saying words of the text to oneself, one by one") as a striking and possibly analogous case.

Intuition derived from a background of "biological experience" suggests that once a general solution to the pattern recognition problem has been developed, a few comparatively simple innovations in the deductive techniques for the machine translation of natural languages (73, 74, 75, 76) should permit us to rapidly boot-strap the computer far past man, (however see some of the reservations of both Bar-Hillel (77) and Selfridge (9)). Since the structure of the language of a pattern recognition program for a digital computer will be explicit and known, in contrast to the case with the "language" of the gestalt (and in contrast to the case with the "language" of a perceptron (78, 79, 7)), we anticipate no problems related to a "translation barrier." Therefore, even without deductive innovations, the range of application of a satisfactory pattern recognition program may exceed the capacities of most mammals and perhaps even man.

TIME, CAUSALITY AND A MECHANICAL ORACLE OF DELPHI

Except for the inferences that may have been drawn from the occasional suggestion that events might be treated as patterns, such taxonomic procedures may seem irrelevant to the discovery or generation of "causal" hypotheses, theories and laws. However, if an event is observed as (or can be transformed so as to correspond to) an ordering of attributes along a time axis, then when events are compared to one another, those which are found to be very similar will often obey the same "causal laws." The occurrence of high correlations or similarity between the attributes of such events provides information equivalent to the usual kinds of causal statements associated with processes, i.e., the earlier attributes are either the "causes" of the later attributes and/or both earlier and later attributes result from "common causes," (e.g., the "Law" which is the operational definition of the class).

In particular, if useful transformations from the relativistic four-dimensional time space continuum to one dimension can be developed, then all aspects of causation will, in principle, be amenable to such taxonomic pattern analysis. In the meantime, many relatively simple "one-dimensional" processes or events (e.g., electrocardiograms) will be analyzable by even this primitive kind of model.

Since it has been convincingly argued that n-dimensional pattern recognition is analogous, if not homologous to what we generally call thinking (3, 4, 5), we may be close to having the blueprints for a mechanical Oracle of Delphi.

CONCLUSION

It would appear that a primitive frame-work for a process which fulfills most of the functions of the "scientific method" can probably be programmed for digital computers. It provides the mechanism for both "discovery" and for automatic generation of

operational definitions. From these, the symbolism which is the grist of the “more sophisticated part” of the scientific mill can be derived.

Is the more sophisticated aspect of the scientific method, the generation of hypothetical models followed by testing these models by further select observation, really different in kind from the more primitive process?

I have already indicated my belief that it is not. The level of complexity of the attributes of “objects” to be compared may be greater, but the added sophistication comes mainly from the use of additional and more complex “measures of similarity” and/or more complex forms of normalization. For example, the Correlation Coefficient permits us to discover the existence of any linear relationships between two sets of data, x and y , such that $y_i = mx_i + b$ for all values of m and b except $m = 0$. If one performs the non-linear transformation represented by taking the logarithms of the data points, then the Correlation Coefficient, when applied to such transformed (normalized) data will detect any relationship between the two sets of data such that $y_i = cx_i^n$ for all values of n and c except n or c equal to zero.

If one considers equations to which data are usually fit in “scientific models” (e.g., linear and non-linear algebraic and differential equations, etc.), semi-metric measures and transformations probably exist which have a similar relationship to these equations as the Correlation Coefficient in combination with a logarithmic transformation has to simple exponential equations. If such additional similarity measures and transformations were added to this type of pattern recognition program, much of the more elegant aspects of the scientific method might, in principle, be relegated to the computer (especially if some sophisticated heuristic can be developed for determining which data points in one set “correspond” to the data points in another).

Golomb (80) has stated that “it is scarcely an exaggeration to assert that *classification* is the most fundamental objective in mathematics,” (italics mine). It would therefore seem reasonable to expect that efforts of socially motivated mathematicians might be easily and profitably turned in these directions. Many of the voids to be filled seem to be directly related to Topology, (particularly Algebraic Topology), Set Theory, Group Theory and the very broad areas of Analysis.

Let us suppose that the ideal set of analytical techniques had already been produced by the mathematicians, and that we were convinced that such techniques should be applied to the analysis of the widest possible range of problems. What remaining bottlenecks can we envisage?

Present day computer technology has (for good reason) been mainly committed to the development of computers that handle most data in their arithmetic sections sequentially. If an analytical scheme requires the computation of 10,000 similarity coefficients, the time necessary to perform such computations must be approximately 10,000 times the time to compute one similarity coefficient because the same hardware must be used sequentially for each of the individual computations. With the advent of “integrated micrologic circuitry” which occupies much less volume, consumes much less power and is faster and potentially more reliable than existing circuitry (81, 82), industry is

beginning to plan for machines which have considerable parallel capabilities (83). However, only one serious effort seems to be directed towards the ultimate construction of a machine with anything approaching 10,000 parallel arithmetic sections (84). Yet because of the enormous numbers of computations required for such an analytical taxonomic scheme, such machine capabilities plus a high-speed memory 100 or more times the capacity of the IBM 7090, would probably be mandatory to make it economically feasible to use a taxonomic program for the solution of the widest range of problems. It would appear that the remarkable capabilities of the vertebrate brain (and especially the human brain), in spite of the comparatively slow speed of the individual "circuits," are largely attributable to the enormous numbers of "parallel circuits" at all levels of the nervous system, (e.g., see 85, 86, 87, 88). Given such a quantitative change in computer design, we should expect as striking an advance in data processing capabilities as in the evolutionary transition from protochordate to man. Since, in contrast to the mutational part of the evolutionary process, the technological innovation would be goal oriented, this kind of data processing can change man's way of life at a staggering rate.

Let us suppose that the social and industrial motivation were sufficient to propel us rapidly along the road to widespread use of analytical taxonomic programs on large-scale, parallel-circuit, digital computers for the solution of the widest range of human problems. Will we humans be prepared to cope with the intellectual, cultural and emotional problems that are likely to be created by the most severe and extensive case of technological unemployment we have ever had to face? (89) In competition with the machine, most, if not all of us, may be found to be amateurs even at that scientific game of chess that we play with nature—and which we call research (6, 9, 90). Many of the problems posed by this kind of question may demand solutions within the next few generations. Viewed from such a perspective, from among the most knowledgeable estimates of the formal educational needs of our younger generations, (e.g., 91, 92), none comes even remotely close to the mark. In the words of the late Norbert Wiener, father of Cybernetics, who was troubled by similar considerations, "The hour is very late, and the choice of good and evil knocks at our door" (93).

SUMMARY

The purpose of this paper has been to examine some aspects of the general problem of learning in terms of pattern recognition. In our approach, we start with a detailed "image" of each pattern of the set, placing a minimum number of arbitrary constraints on the limits of the raw data, (e.g., setting only a maximum level of resolution and the outer bounds of each image field). The data points of the image or pattern are then represented as the coordinates of each pattern, and the pattern itself, as a point in an n -dimensional space (hyper-space), and a semi-metric measure of "similarity" of pattern to pattern is used to define a distance between samples and characterizes this hyperspace as a semi-metric space (rather than a metric space). The semi-metric distances between patterns are examined and the largest "natural" constellations or clusters are separated from one

another. The data points of patterns within each constellation are reweighted, based on the information content of the subpopulation, and are further divided into subgroups by the same techniques. Those boundary conditions which must be arbitrarily set at the outset can be systematically varied to produce the maximum decrease in information at each subdivision. This serves as the device for converging to the “most natural” subdivision by measuring how closely our semi-metric space approximates a defined, ideal “Information Space”, thus removing a large element of remaining “arbitrariness” from the procedure. In this way, a taxonomy of patterns is automatically generated. By introducing a simple binary coding device which assigns **0**'s and **1**'s to designate the smaller and larger constellations at each subdivision, an efficient Shannon-Fano Code (10, 22) (in terms of Shannon's Binary Coding Theorem (10)) for the “description,” “communication” and processing of such patterns is automatically generated. Such a taxonomic procedure both “discovers” classes of patterns and orders them in a “dictionary” and provides the means of efficiently recognizing whether any new sample of a pattern “belongs” to one or more of these already discovered classes. It thus solves the “semantic problem” of Information Theory (10). Mechanisms for continuously updating the process on the basis of the statistics of the portion of the population of patterns already sampled and analyzed are explicitly introduced into the scheme.

This technique departs widely from the more familiar decision-space model of decision theory, (e.g., see 30, 94, 95, 96, 97, 98, 99, 100), where, at the outset, specific “teleological” constraints are usually placed on a limited synthetic set of pattern parameters which it is hoped will allow the differentiation and recognition of the different kinds of known patterns; where a metric decision-space is usually carved up by sets of “hyper-planes” which are intuitively and computationally more complex and usually more arbitrarily distributed than our decision boundaries; and where no attempt is usually made to structure the resulting set of classes in an efficient “dictionary” or taxonomy.

The relationship of our model to “learning” and to the “scientific method” is discussed. The areas of mathematics which might be further exploited to increase the sophistication and generality of such procedures are noted. The pertinence of computer design to the efficient and widespread application of such a class of techniques is briefly discussed and a change in direction is recommended in this connection. Attention is finally drawn to potential social consequences of the widespread use of such techniques.

ACKNOWLEDGMENTS

This work is dedicated to the memory of Mount Sinai's Paul Klemperer. As a result of my close and continuous association with Dr. Klemperer from 1955 until his death in 1964, this great pathologist's continuing concern with the explication of the historical foundations of modern concepts of medicine and science (e.g., see 101) provided inspiration for a good portion of this inquiry.

It was also through the stimulus provided by frequent discussion with T. T. Tanimoto of the basic ideas in his “Elementary Mathematical Theory of Classification and

Prediction” (31) over the period from 1957 to 1961 that most of the viewpoints presented here first began to take shape. Both Dr. Tanimoto and my colleagues, B. J. Davis and S. Diamond have persevered as constant sounding boards through many changes in direction, and have provided most of the necessary negative feedback which has led to this fairly stable and, hopefully, useful perspective.

REFERENCES

1. Dobzhansky, T.: *Mankind Evolving*, Yale Univ. Press, New Haven, 1962.
2. Bridgman, P. W.: *The Nature of Physical Theory*, Dover, New York, 1936.
3. Turing, A. M.: Computing Machinery and Intelligence. *Mind*, **59**: 433, 1950; also reprinted in (41).
4. Armer, P.: Attitudes toward intelligent machines. Symposium on Bionics, *Wadd Tech. Rep.* **60,600,13**, 1960, also reprinted in (41).
5. Minsky, M.: Steps toward artificial intelligence. *Proc. IRE*, **49**: 8, 1961, also reprinted in (41).
6. Kelley, J. L., and Selfridge, O. G.: Sophistication in computers: A disagreement. *IRE Trans. in Inf. Theory*, **IT-8**: 78, 1962.
7. Minsky, M., and Selfridge, O. G.: Learning in random nets, in *Proc. 4th London Symp. on Inf. Theory*. C. Cherry, Ed., Academic Press, New York, 1961, pp.335-347.
8. Brillouin, L.: Empirical laws in physical theories; the respective roles of information and imagination, in *Self-Organizing Systems*, M. C. Yovitts, G. T. Jacobi, and G. D. Goldstein, Eds. Spartan, Wash., D. C., 1962, pp. 231-242.
9. Selfridge, O. G.: *The organization of organization*, in *Self-Organizing Systems*, M. C. Yovitts, G. T. Jacobi, and G. D. Goldstein, Eds., Spartan, Wash., D. C., 1962, pp. 1-7.
10. Shannon, C. E., and Weaver, W.: *The Mathematical Theory of Communication*. Univ. Ill. Press, Urbana, 1949.
11. Brillouin, L.: *Science and Information Theory*. Academic Press, New York, 1955.
12. Inhelder, B., and Piaget, J.: *The Early Growth of Logic in the Child, Classification and Seriation*. Harper and Row, New York, 1964.
13. Fantz, R. L.: Visual Experience in Infants: Decreased Attention to Familiar Patterns Relative to Novel Ones. *Science*, **146**: 668, 1964.
14. Bruner, J. S.: The course of cognitive growth. *Am. Psychol.*, **19**: 1, 1964.
15. Gibson, E. J.: Learning to read. *Science*, **148**: 1066, 1965.
16. Weyl, H.: *Philosophy of Mathematics and Natural Science*. Atheneum, New York, 1963.
17. Wright, A. H., and Wright, A. A.: *Handbook of Frogs and Toads of the United States and Canada*. Comstock, Ithaca, N. Y., 1949.
18. Pollister, A. W., and Ornstein, L.: The photometric chemical analysis of cells, in *Analytical Cytology*. R. C. Mellors, Ed., McGraw-Hill, New York, 1959, pp. 431-518.
19. Ornstein, L.: Disc Electrophoresis—I, Background and Theory. *Annals N. Y. Acad. Sci.*, **121**: Art. 2, 321, 1964.
20. Davis, B. J.: Disc Electrophoresis—II, Method and Application to Human Serum Proteins. *Annals N. Y. Acad. Sci.*, **121**: Art. 2, 404, 1964.
21. Fantz, R. L.: The origin of form perception. *Scientific American*, **204**: 66, 1961.
22. Elias, P.: Information Theory, in *Handbook of Automation Computation and Control*, Vol. 1. E. M. Grabbe, S. Ramo, and D. E. Wooldridge, Eds. Wiley, New York, 1958, pp. 16-01-16-48.
23. Ornstein, L.: Life on other planets: Some exponential speculations. *Science*, **144**: 614, 1964.

24. Hoyer, B. H., McCarthy, B. J., and Bolton, E. T.: A molecular approach in the systematics of higher organisms. *Science*, 144: 959, 1964.
25. Abramson, N., and Braverman, D.: Learning to recognize patterns in a random environment. *IRE Trans. on Inf. Theory*, **IT-8**: S-58, 1962.
26. Fischler, M., Mattson, R. L., Firschein, O., and Healy, L. D.: An approach to general pattern recognition. *IRE Trans. on Inf. Theory*, **IT-8**: S-64, 1962.
27. Sebestyen, G. S.: Recognition of membership in classes. *IRE Trans. on Inf. Theory*, **IT-7**: 44, 1961.
28. Kalin, T. A.: *Some Metric Considerations in Pattern Recognition*. Res. Lab. of Electronics, M.I.T. Cambridge, Mass., 1960.
29. Sokal, R., and Sneath, P.: *Principles of Numerical Taxonomy*. Freeman, San Francisco, 1963.
30. Sebestyen, G. S.: *Decision Making Processes in Pattern Recognition*. Macmillan, New York, 1962.
31. Tanimoto, T. T.: An Elementary Mathematical Theory of Classification and Prediction. *I.B.M. Program IBCLF*, 1959.
32. Rogers, D. J., and Tanimoto, T. T.: A computer program for classifying plants. *Science*, **132** : 1115, 1960.
33. Tanimoto, T. T.: Non-linear model for a computer assisted medical diagnostic procedure. *Trans. N. Y. Acad. Sci., Ser. 2*, **23**: 576, 1961.
34. Hamming, R. W.: Error detecting and error correcting codes. *Bell Tech. J.*, **29**: 147, 1950.
35. Cooper, P. W.: Hyperplanes, hyperspheres, and hyperquadrics as decision boundaries, in *Computer and Information Sciences*. J. T. Tou and R. H. Wilcox, Eds. Spartan, Wash., D. C. 1964, pp. 111-138.
36. Pearson, K.: *Early Statistical Papers*. Cambridge Univ. Press, London, 1956.
37. Tanimoto, T. T.: A class of exponential distributions and their associated Minkowski geometries. *Notices Amer. Math. Soc.*, **8**: 432, 1961.
38. Wald, A.: *Statistical Decision Functions*. Wiley, New York, 1950.
39. Middleton, D.: *An Introduction to Statistical Communication Theory*. McGraw-Hill, New York, 1960.
40. Loomis, R. G.: Mathematics and Appl. Sect. Data Systems Division, IBM, (personal communication) .
41. Feigenbaum, E. A., and Feldman, J., Eds. *Computers and Thought*. McGraw-Hill, New York, 1963.
42. Copeland, A. H.: Probability, in *Handbook of Automation, Computation and Control*, E. M. Grabbe, S. Ramo, and D. E. Wooldridge, Eds. Wiley, New York, 1958, **Vol. I**, pp. 12-01-12-20.
43. "Student" (Gosset, W. S.): The probable error of a mean. *Biometrika*, **6**: 1, 1908.
44. Fano, R. M.: *Transmission of information; a statistical theory of communication*. Wiley, New York, 1961.
45. Huffman, D. A.: A method for the construction of minimal-redundancy codes. *Proc. IRE*, **40**: 1098, 1952.
46. Smith, R. V.: Similarity and Entropy. *Com. A.C.M.*, **7**: 397, 1964.
47. Smithes, O.: Zone electrophoresis in starch gels and its application to studies of serum proteins. *Advances in Protein Chem.*, **14**: 65, 1959.
48. Watanabe, S.: Information theoretic analysis of multivariate correlation. *IBM Jour. of R. and D.*, **4**: 66, 1960.
49. Watanabe, S.: Une explication mathématique du classement d'objets, in *Information and Prediction in Science*. S. Dockx and P. Bernays, Eds. Academic Press, New York, 1965, pp. 39-76.
50. Solomon, H.: Classification procedures based on dichotomous response vectors, in *Studies in Item Analysis and Prediction*. H. Solomon, Ed. Stanford Univ. Press, Palo Alto, 1961 pp. 177-186.

51. Lazarsfeld, P. F.: The algebra of dichotomous systems, in *Studies in Item Analysis and Prediction*. H. Solomon, Ed. Stanford Univ. Press, Palo Alto, Calif., 1961.
52. Linfoot, E. H.: An informational measure of correlation, *Information and Control*, **1**: 85, 1957.
53. Kolmogorov, A. N.: *Foundations of the Theory of Probability*. Chelsea, New York, 1950.
54. Born, M.: *Natural Philosophy of Cause and Chance*. Dover, New York, 1964.
55. Hobart, R. E.: Hume without skepticism, II. *Mind*, **39**: 409, 1930.
56. Wigner, E. P.: Events, Laws of Nature, and Invariance Principles. *Science*, **145**: 995, 1964
57. Solomonoff, R. J.: A Formal Theory of Inductive Inference, Part I. *Information and Control*, **7**: 1, 1964.
58. Popper, K.: *The Logic of Scientific Discovery*. Basic Books, New York, 1959.
59. Hanson, N. R.: Galileo's discoveries in dynamics. *Science*, **147**: 471, 1965.
60. Henkin, L.: Are logic and mathematics identical?, *Science*, **138**: 788, 1962.
61. Slagle, J. R.: A multipurpose theorem-proving heuristic program that learns, in *Information Processing 1965*: Proc. of IFIP Congress 65. W.A. Kalenich, Ed. Spartan, Wash., D. C., 1965, **Vol. II**, in press.
62. Platt, J.: Strong inference. *Science*, **146**: 347, 1964.
63. Born, M.: *Einstein's Theory of Relativity*. Dover, New York, 1962.
64. Fieller, E. C.: The distribution of the index in a normal bivariate population. *Biometrika*, **24** : 428, 1932.
65. Feigl, H.: The logical character of the principle of induction, in *Reading in Philosophical Analysis*. H. Feigl and W. Sellars, Eds. Appleton-Century-Croft, New York, 1949.
66. Snow, C. P.: *Two Cultures and the Scientific Revolution*. Cambridge Univ. Press, Cambridge, England, 1959.
67. Holton, G.: Modern Science and the Intellectual Tradition. *Science*, **131**: 1187, 1960.
68. Novikoff, A. B. J.: Integral geometry as a tool in pattern perception, in *Principles of Self-Organization*. H. Von Foerster and G. W. Zopf, Sr., Eds. Pergamon, New York, 1962.
69. Singer, J. R.: An electronic analogue of the human recognition system. *J. Opt. Soc. Am.*, **51**: 61, 1961.
70. Pinkerton, R. C.: Scanning and Form. *Astounding Science Fiction*, Dec., 1954, pp. 102-111.
71. Hadamard, J.: *The psychology of invention in the mathematical field*. Dover, New York, 1954.
72. Treisman, A. M.: Reading Rate, Word Information and Auditory Monitoring of Speech. *Nature*, **205**: 1297, 1965.
73. Vygotsky, L. S.: *Thought and Language*. Ed. and trans. by E. Hanfmann and G. Vaker. Wiley, New York, 1962.
74. Miller, G. A.: Some psychological studies of grammar. *Am. Psychol.*, **17**: 748, 1962.
75. Hockett, C. F.: Animal "languages" and human language, in *The Evolution of Man's Capacity for Culture*. J. H. Spuhler, Ed. Wayne State Univ. Press, Detroit, 1959, pp. 32-39.
76. Garvin, P. L., Ed.: *Natural Language and the Computer*. McGraw-Hill, New York, 1963.
77. Bar-Hillel, Y.: The present status of automatic translation of languages, in *Advances in Computers*. F. L. Alt, Ed. Academic Press, New York, 1960, pp. 92-163.
78. Rosenblatt, F.: *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan, Wash, D. C., 1962.
79. Block, H. D.: The perceptron: A Model for Brain Functioning, I. *Rev. Mod. Phys.*, **34**: 123, 1962.

80. Golomb, S. W.: A mathematical theory of discrete classification, in *Proc. 4th London Sym. on Inf. Theory*. C. Cherry, Ed. Academic Press, New York, 1961.
81. Rajchman, J. A.: Integrated magnetic and superconductive memories—A survey of techniques, in *Information Processing 1965: Proc. of IFIP Congress 65*. W. A. Kalenich, Ed. Spartan, Wash., D. C., 1965, **Vol. I**, pp. 123-129.
82. Davis, E. M.: Integrated circuits—Commercial computer applications, in *Information Processing 1965: Proc. of IFIP Congress 65*. W. A. Kalenich, Ed. Spartan, Wash., D. C., 1965, **Vol. II**, in press.
83. Fernbach, S.: Computers in the U.S.A.—Today and tomorrow, in *Information Processing 1965: Proc. of IFIP Congress 65*. W. A. Kalenich, Ed. Spartan, Wash., D. C., 1965, **Vol. I**, pp. 77-91.
84. Carroll, A. B., Gregory, J. G., Leonard, W. H., and Slotnick, D. L.: The SOLOMON II computing system, in *Information Processing 1965: Proc. of IFIP Congress 65*. W. A. Kalenich, Ed. Spartan, Wash., D. C., 1965, **Vol. II**, in press.
85. Lettvin, J. Y., Maturana, H., McCulloch, W. S. and Pitts, W.: What the frog's eye tells the frog's brain. *Proc. IRE*, **47**: 1940, 1959.
86. Miller, G. A.: Decision units in the perception of speech. *IRE Trans. on Inf. Theory*, **IT-8** : 81, 1962.
87. Land, E. H.: The Retinex. *American Scientist*, **52**: 247, 1964.
88. Neisser, U.: Visual search. *Scientific American*, **210**: 94, 1964.
89. Reagan, M. D.: For a guaranteed income. *New York Times Magazine*, June 7, 1964, p. 20.
90. Wiener, N.: *The human use of human beings*. Doubleday and Company, Garden City, New York, 1956.
91. Conant, J. B.: *The American high school today*. McGraw-Hill, New York, 1959.
92. Rickover, H. G.: *American education, a national failure; the problem of our schools and what we can learn from England*. Dutton, New York, 1963.
93. Wiener, N.: Some moral and technological consequences of automation. *Science*, **131**, 1355, 1960.
94. Highleyman, W. K.: The design and analysis of pattern recognition experiments. *Bell System Tech.J.*, **41**(2): 723, 1962.
95. Barus, C.: A scheme for recognizing patterns from an unspecified class, in *Optical Character Recognition*. G. L. Fischer, D. K. Pollock, B. Raddack, and M. E. Stevens, Eds. Spartan, Wash., D. C., 1962, pp. 227-247.
96. Sebestyen, G. S.: Recognition by an adaptive process of sample set construction. *IRE Trans. on Inf. Theory*, **IT-8**: S-82, 1962.
97. Needham, R. M.: A method for using computers in information classification, in *Proc. Intern. Fed. Inform. Processing Congr., 1962*, North-Holland, Amsterdam, 1963.
98. Abramson, N., Braverman, D., and Sebestyen, G. S.: Pattern recognition and machine learning. *IEEE, Trans. on Inf. Theory*, **IT-9**: 257, 1963.
99. Bonner, R. E.: On some clustering techniques. *IBM Jour. of R. and D.*, **8**: 22, 1964.
100. Nilsson, N. J.: *Learning Machines: Foundations of Trainable Pattern-Classifying Systems*. McGraw-Hill, New York, 1965.
101. Klemperer, P.: The growth of physiological knowledge; its historical background. *Bull. N. Y. Acad. Med.*, **39**: 765, 1963.