THE RESEARCH FOUNDATION

OF STATE UNIVERSITY OF NEW YORK

P.O. Box 7126

Albany, New York 12224

DEVELOPMENT OF GUIDELINES FOR THE DEFINITION OF THE RELEVANT INFORMATION CONTENT IN DATA CLASSES

ERICH SCHMITT

Principal Investigator
Department of Electrical Engineering
State University of New York at Buffalo
Buffalo, New York 14214

FINAL REPORT CONTRACT NO. NAS8-28264

PREPARED FOR

GEORGE C. MARSHALL SPACE FLIGHT CENTER

Huntsville, Alabama 35812

April 1973

THE RESEARCH FOUNDATION
OF STATE UNIVERSITY OF NEW YORK
P.O. Box 7126
Albany, New York 12224

DEVELOPMENT OF GUIDELINES FOR THE DEFINITION OF THE RELEVANT INFORMATION CONTENT IN DATA CLASSES

ERICH SCHMITT
Principal Investigator
Department of Electrical Engineering
State University of New York at Buffalo
Buffalo, New York 14214

FINAL REPORT CONTRACT NO. NAS8-28264

PREPARED FOR
GEORGE C. MARSHALL SPACE FLIGHT CENTER
Huntsville, Alabama 35812

April 1973

TABLE OF CONTENTS

Page Number

1. Introduction 1

2. Mathematical Model of Scientific Experiment Information Systems 5

   2.1 Information Source 7

   2.2 Measurement Part 10

   2.3 Analysis and Data Usage 13

3. Guidelines for the Definition of Experiment Requirements Concerning the Information Aspects 16

4. Future Improvements 19

5. References 20

Abstract

The problem of experiment design is defined as an information system consisting of information source, measurement unit, environmental disturbances, data handling and storage, and the mathematical analysis and usage of data. Based on today's concept of effective computability, general guidelines for the definition of the relevant information content in data classes are derived. The lack of a universally applicable information theory and corresponding mathematical or system structure is restricting the solvable problem classes to a small set. It is expected that a new relativity theory of information, generally described by a universal algebra of relations will lead to new mathematical models and system structures capable of modeling any well defined practical problem isomorphic to an equivalence relation at any corresponding level of abstractness.

1. Introduction

Any attempt to formalize both the process of information measurements and the analysis of observations requires that the two main parts in scientific experiments be based on a rigorous mathematical concept of information. Therefore, we first discuss the following questions:

What concepts of information definition are known? How can they be applied for the improvement of experiment design?

Today's knowledge of rigorous information definition concepts is based on Shannon's [7] "mathematical theory of communication". This theory assumes that information is a statistical quantity. The application of this theory requires a complete probabilistic description of the process to be analyzed. Clearly, Shannon's concept carries over to the problem of experiment design by using the notion of "information in an experiment," rather than in a message. This has been recognized by various experiment designers in the past, as the references [8], [9], [10] indicate. However, no applicable concept has resulted from this work, although theoretical existence proofs of very surprising phenomena are of great importance. For instance, the translation of Shannon's theoretical results into the area of experiment design implies the existence of the following surprising fact: The noise effect in a measurement can be reduced to an arbitrary small value by appropriate design of the parameters of the measurement and proper sequences of related measurements, even if the signal to noise ratio of each single measurement is much less than one.

Unfortunately, a practical concept of implementation capable of achieving this theoretical bound has not been found for any area. Moreover, it has been recognized by well known information theorists (see for instance references [3], [4[, [5]) that the statistical concept of information definition is not well suited for a universal application to practical problems.

This fact has also been the motivation for us to search for a universal deterministic definition of information applicable to any practical problem. Our belief in the existence of such a concept had been supported in the last few years by the results in the development of a general system theory of discrete systems. The application of these results to today's information measuring and processing principles and structures brings out the fact that we are now operating on a 'primitive' concept level even with the most sophisticated computer systems. One can also show that today's concept of mathematical algorithms, as they are for instance implemented in the form of computer programs, can only be effective for the solution of a very small part of the problems of interest. Further, there can be no "effective computable algorithm" (programmed on the largest computer) to solve the problem of information diagnostics, the optimization of measurement parameters, and the optimization of experiment design in a general way. (The term: "effective computable algorithm" is used in the mathematical world to express the fact that a problem with this algorithmic property is generally solvable by a Turing machine. In simple terms: a Turing machine can be thought of as an idealized computer with an infinite memory space.)

The model on which the discrete system theory is based (developed outside this contract) implied the feasibility to overcome today's limitations on effective computability thru the introduction of abstract universal system concepts which are not restricted to finite sets, but rather are capable of relating transfinite entities in a universal way. This universal system model shows further the possible way of extension to a family of homological structures which are isomorphic to all families of practical problem processes. Expressing this in simple terms would mean that any string of measurement data of an experiment corresponds to an isomorphic universal system structure, namely the equivalence class to which the experiment corresponds. If we now assume that means exist to express the equivalence class by some kind of complexity measure we would not only be able to evaluate relatively the variation of measurement parameters for the purpose of optimization but we could also analyze the data universally, since the equivalence class concept allows the transition into all levels of abstractness such that any functional relationship between arbitrary variables or substrings of the measurement data string can be expressed by the corresponding subclass of the complete data string class.

It should be understood by now that such a universal system must have an infinite number of elements, or to be more exact: it must have a cardinality which is greater than the cardinality of the largest equivalence class of the problem process under analysis.

The question of realizibility of such a system is clearly of main concern since the theoretical concept would not help to solve our practical problems. It has been hoped to develop the complete mathematical structure and its detailed description for an immediate application for all storage, measurement and analysis processes under this contract. However, the task is too immense and requires the development of new mathematical concepts which go far beyond the present status of the most advanced mathematical field of "model theory".

All the requirements of the system have been formulated such as:

The system must be a non-communitative abstract group with a number of generators which is larger than the cardinality of the set of all relations of the input data string of any length. In other words, the system implements universal algebras of relations which is the generalized concept of functions.

Using this universal system structure a general relativity theory of information can be derived in the sense that information of a data string is definable relative to the analysis interpretation rule a user asks. Redundancy is defined universally, namely because for each equivalence class one representative generator element is sufficient. All other relations belonging to the same class are redundant.

In the following sections we will represent the general mathematical model of scientific experiment information systems in a simple way. A small example in the analysis part (section 2.3) will indicate that only a universal approach can solve our today's problems of data "explosion". The cost savings (already in data storage) would be immense.

In section 3 very general guidelines for the present day approach will be stated. They serve the purpose to make experimenters aware of the fact that an experiment must be designed from a complete system standpoint including information source, measurement unit, data handling, data storage, and data usage.

Finally, in section 4 a list of future improvements based on the universal system concept indicates the high importance of the relativity theory of information.

2. Mathematical Model of Scientific Experiment Information Systems.

In general, any experiment is first concerned with the observation of information from a well determined observation space. These involve the decisions of:

(a) What to measure?

(b) Where to measure?

(c) How to measure?

These decisions are not independent of the second part of questions any experiment involves, namely:

(d) Why to measure?

(e) How to analyse the observations?

(f) What to expect as conclusive results?

Question a) refers to observable events which must originate at locations we consider to be relevant for the experiment under investigation. Such a "vague" definition is not satisfactory for any mathematical treatment. Therefore, we introduce the fact thay any measureable event must be generated by a so-called information source. We must consider this origin closely together with all the other questions in order to be able to design an experiment optimally. Any knowledge we desire to gain from an experiment is dependent upon the information source. Sometimes, we also like to observe the effect of some information source on its immediate environment. In certain cases a number of various information sourses might contribute to the events of the observation space under consideration. For example, we might measure the composition of pollutants in the atmosphere which originate from various sources such as industrial plants, cars etc. over a wide area. In section 2.1, therefore, we introduce a mathematical model which allows the structure of any information source or sets of information sources and their interaction to any desired level of accuracy and completeness.

This subsequently gives the key to formulate the remaining questions more clearly. The questions b) Where to measure? and c) How to measure? are in addition very much connected with the selection of the measurement device(s) and the connected parameters such as sensors, sensitivity, accuracy, calibration, scaling, dynamical characteristics, stability etc. These parameters also effect strongly the analysis process under question e).

In order to be able to express all parts and problems of the measurement part clearly we introduce under section 2.2 an appropriate general mathematical model. This model will allow us to structure any practical problem to any desired level of accuracy and completeness. Furthermore, it serves in serial composition with the model of the information source and environmental disturbances for the evaluation of the questions e) How to analyse the observations? and f) What to expect as conclusive results?

These questions will be further treated in section 2.3. Clearly, in addition, the problem of formatting, editing, transmitting, storing, and retrieving the measurement data must be considered. These problem areas will be included in the guidelines under section 3.

2.1 Information Source

The model of an information source has to be complete and general in order to be applicable to any problem area. Not only should the concept allow the model structure to subdivide (decompose) into substructures, but we should also be able to express the behavior on any level of abstraction. For example, considering a radiating source, we should be able to model (in principle) each single atom and the interaction of the whole set of atoms. For other purposes it might be desirable to model the whole set of atoms as a single structure. The latter would be of interest when the emission to the environment is the object of investigation.

It is obvious that for all these requirements we need the concept of sets to represent the collection of the defined elements. Define Z to be the set of internal states of the information source.

$$Z = \left\{ z_1 , z_2 , \dots z_\gamma \dots , z_g \right\}$$

where the elements z_γ represent the various possible states of the information source. The element z_γ can represent in ordered form a catenation of states of substructures. For example,

$$z_\gamma = (z_{\gamma_1} , z_{\gamma_2} , \dots , z_{\gamma_s}) .$$

Any coding for the various elements and subelements is allowed.

Clearly, without loss of generality the elements can be simply coded as sequences of binary digits from the alphabet $\{0, 1\}$. Hence, the elements z_γ are catenated binary strings, and the internal state set is described by the binary sequences $z_1, z_2, \ldots, z_g$. Similarly, we can define the set of input elements X to the information source (if such an input exists).

$$X = \left\{x_1, x_2, \ldots, x_\alpha, \ldots, x_a\right\},$$

where the x_α's are binary sequences. The set of output elements of the information source is denoted by

$$Y = \left\{y_1, y_2, \ldots, y_\lambda, \ldots, y_b\right\}.$$

For simple representation, we assume again the y_λ's to be sequences of binary numbers. Of importance is the capability of the model to express the functional relationship between the sets. In certain cases the relationship might be expressible by the concept of an algebraic function. However, in general, we have to consider the concept of algebraic relations. The two concepts are briefly discussed. For a more detailed understanding, the reader is referred to the literature, for instance, [1]. For given sets S and T, a function f with domain S and codomain T is a rule which assigns to each element $s \in S$ of the domain a single element of the codomain T, called the value of f at s, denoted by f(s). It is also common to call f a map or transformation on S to T, written f : S → T. The binary relation concept includes the function concept as a special case. By a binary relation between two sets S and T we mean a rule α which decides, for any elements $s \in S$ and $t \in T$, whether or not s is in the relation α to t. If it is, we write $s \alpha t$; if it is not, we write $s \overline{\alpha} t$. The main difference between a binary relation and a function is:

binary relation: a certain element s can be related (through various rules) to several elements t:

function: a certain element s has only a single element t assigned.

The relationships between the sets of the information source are then:

$$\delta : (Z \times X)^t \dashrightarrow Z^{t+\tau_1} \qquad (1)$$

$$\omega : (Z \times X)^t \dashrightarrow Y^t \qquad (2)$$

δ is a function or binary relation defined on the cartesian product Z x X to the state set Z. ω is a function or binary relation on Z x X to the output Y. We normally are concerned with systems which are time varying. For this case we can introduce the superscripts in (1) and (2). The complete system and its dynamicsis then described by a sequence of the triples

$$(X, Z, Y)^t, (X, Z, Y)^{t+\tau_1}, (X, Z, Y)^{t+\tau_2}, ...$$

where at each time interval an element out of each set occurs. Since a binary relation can describe a whole set of functions, the model also covers the case that the individual relations (z_γ, x_α) $\delta(z)$ and (z_γ, x_α) $\omega(y)$ change in time. The time dependent rules in the relations δ, ω can then be controlled by the time variable, if the modelling problem is of nonstationary nature. The dynamic process would then be described by a sequence of quintuples

$$(X, Z, Y, \delta, \omega)^t, (X, Z, Y, \delta, \omega)^{t+\tau_1}, \ldots\ldots$$

Note: each of the elements including δ, ω, t can be finally represented as a sequence of binary numbers.

2.2 Measurement Part

Depending on the problem the output Y of the information source might not be completely observable at the input of the measurement device. In addition, some other (unwanted) information sources might interfere. They can be considered as noise sources. Hence, at the input of the measurement device the set of elements

$$M = \{ m_1, m_2, \ldots, m , \ldots, m_k \}$$

is defined. Denoting the noise sources by N we then have the function or binary relation ν to consider

$$\nu : Y \times N \dashrightarrow M \qquad (3)$$

Of course, we are only interested in the relation on Y. The open question is how to eliminate the noise. Usually it is assumed that the noise is additive and arithmetic averages are taken. Before such a procedure is applied careful consideration must be given to the fact that averaging might disturb relevant dynamic information of the measured information source. Further, if the noise acts multiplicatively or generally nonlinearly as expressed in (3), the averaging procedure disturbs more than it helps. The solution to this problem must be shifted to the analysis part.

If the measurement device would be ideal the experimenter would finally obtain either an element of M or a subset of elements of M, depending on how the data are edited. Because the devices are normally not ideal, the set M is further transformed by the instrument error to

$$\mu : M \times I \dashrightarrow D \tag{4}$$

where μ can again be a function or generally a binary relation. The sets I and D are defined by

$$I = \left\{ i_1, i_2, \ldots, i_\zeta, \ldots, i_\gamma \right\}$$

$$D = \left\{ d_1, d_2, \ldots, d_\xi, \ldots, d_e \right\} \quad .$$

(4) allows the description of arbitrary instrument error characteristics.

D is the set of final data sequences edited with all the necessary additional information such as time, location, positioning, etc. The result of an experiment might be a single element d_ξ represented as a sequence of binary numbers. Thus, we have defined the cascade of relations

$$(Z \times X)\ \omega\ Y$$
$$(Y \times N)\ \nu\ M$$
$$(M \times I)\ \mu\ D$$

Note: D contains the data which should contain as much information as possible on the desired set Y or Z of the information source. Any analysis must necessarily use the concept of converse binary relations or the concept of inverse functions, in order to be able to trace backwards from D to Y or Z.

2.3 Analysis and Data Usage

Any string of measurement data is only of worth when it is of further subject of analysis. To measure billions of data bits is useless if nobody has means and finances to analyze it. Therefore, only those measurements should be made which have been shown "a priori" to be effectively analyzable. The experimenter has to outline the analysis approach in all details in order to be able to decide whether his concept is logically consistent, the method is effectively implementable (for instance thru a computer program), and how much effort is involved such as man hours, material, computer time etc.

In order to realize that any "trial and error" analysis method (including all heuristic procedures) is useless, because of combinatorial reasons, a simple example of the analysis of a data string follows.

Assume as an information source (1) a radiation source which emits one of ten different frequencies at a time. As environment to this information source acts, another radiation source (2) ~~evolves~~, which also emits one of ten different frequencies at a time. The behavior of (1) may de dependent on (2) but not vice versa for simplification. For this simple example there exist already 10^{10} possible functions and since the behavior might be non-deterministic we have to consider relations; there are $2^{10^{10}}$ possible relations. If the relation is very complex a large measurement sample would be necessary because of the statistics. Almost no conclusion could be drawn from such an analysis because of the ammumption a priori that the process is stationary. If the latter is not certain the whole measurement is useless and any conclusion false.

Interesting space experiments are normally of much larger complexity and a nontrivial analysis of the data is prohibitive as the combinatorics show. On the other hand, trivial analysis approaches such as computation of averages, variances and the like might turn out a completely false result because of the intrinsic assumptions. One should be very careful in weighing the effort with respect to the possible result for such ill-conditioned experiments.

At this point it is of interest to discuss in general the feasibility of obtaining useful conclusive results thru appropriate optimization of the complete information measurement and analysis system. We learned in the previous sections that each part of the system under discussion has some parts of "uncertainties," such as:

> We normally cannot measure completely all events of an information source, the noise from the environment cannot be pred~~e~~ic~~i~~ted, the accuracy of the measurement device is limited, etc.

In order to model these uncertainties, today's approaches of analysis and optimization use stochastic concepts. Since the stochastic parameters of the various system parts are generally not completely _a priori_ definable, the concepts use estimation principles which require restrictive (unrealistic) assumptions. Mostly, only trivial (linear) mathematical models of this kind are applicable, and still require large amounts of data and analysis time (see, for instance Report NAS 8-26210 under reference 11). Clearly, the obtained results are not conclusive. It has been mentioned in section 1 that the solvability of problems using today's mathematical concepts can be exactly answered with the concept of "effectively computable algorithms". The fact exists that only relatively simple well-defined problems are effectively computable. Already the so-called word problem in associative calculi is unsolvable (for an exhaustive representation of this problem see, for example, reference 12).

But it can be shown that any analysis problem of an experiment from which conclusive results are required is at least as complex as this word problem. So is also the problem of parameter optimization of an experiment. In other words, there is the mathematical proof that the "experimental" problems are not solvable with the mathematical concepts in use today and still large financial effort is put into projects which continuously violate this logical fact. Unfortunately, not only large financial losses result from this, but also wrong conclusions are drawn and published (unrecognizable).

The mathematical fundamentals introduced in section 2 as modeling tools for any experiment design and analysis problem have the new aspect of being completely deterministic in that all problems and parts of uncertainties are uniquely described by algebraic relations. The unique solution of this mathematical structure by a new universal relativity concept of information (see sections 1 and 4) is capable of solving our today's problems in various areas.

3. Guidelines for the Definition of Experiment Requirements Concerning the Information Aspect

The design of scientific experiments requires the careful consideration of all system components and their interaction. These consist of:

Object of experiment (information source)

External influences on the information source

External influences on the measurements

Measurement principles and devices

Data collection, handling and storage

Data analysis concept

The experimenter

Specialists in the various component areas

Subsequent users of measurement data

The whole system is a complicated information structure which has to be considered as an entity for the purpose of efficient synthesis. The experimenter as initiator is the main person at the start. He must formulate the objectives of the experiment in all details. These include: The theoretical study of the information source and all knowledge obtained at previous investigations and experiments by other scientists. This knowledge should be used to improve the system in the various parts, such as measurement requirements, down to the final methods of analysis. He has to establish a model for the data flow: for instance, the sampling rate for the measurement device, and the number system for the representation of the measured data with lower and upper bound. This must be consistent with the expected error in the measurement. The latter requires a careful consideration of all noise sources. The very important part is the usage of the data. He must outline the analysis concept in all details, showing competence in the mathematical procedures involved. It is for instance not sufficient to state: the strings of data are treated as vectors over a linear vector space for which mathematical treatment is well known, because it depends on the size of the arrays of data and on the error whether a solution is feasible at all by today's computation principles. In summary, the experimenter has to show that he considered all important steps and has based on this preliminary investigation all the requirements concerning the various parts of the system.

Since the experimenter normally does not have proficiency in all areas involved during this initial phase of the experiment design he should consult competent specialists to formulate his preliminary concept.

At this point the main work of the experiment design begins as team effort consisting of various specialists in general:

Instrumentation Engineers,
Physists,
Communication Engineers,
Computer Engineers,
Systems Analysts,
Computer Softwave Specialists,
Mathematicians,
Principal Investigators,
Information System Specialists

The Mathematicians and Information System Specialists should guide and control the logical consistency and the solvability of the whole process in the various design phases. They have to be able to understand and translate the problems of the team at all points of design into the appropriate mathematical model, thus, pointing to the weak points during the system design,

The various topics discussed in the previous sections have to be considered by the team of specialists in all details. The list of important points consists of:

Locations of measurements;

Selection of appropriate sensors:

characteristics, sensitivity, bandwidth, functional properties;

Accuracy of instruments:

calibration, scaling, dynamical characteristics, stability, error description by mathematical model;

Data handling:

formatting, editing, collecting, transmitting, storing, file structure for easy retrieving, user groups have to be consulted to consider their needs;

Data analysis:

solvability of mathematical concept must be clearly shown,

Analysis cost estimation:

programming, computer time, amount of data, storage requirements, effort of principal investigator;

Definition of expected objective conclusions from observations (conditioned on final experiment design).

The final design of the experiment has to be critically checked in all parts by a competent information system specialist.

4. Future Improvements

It becomes obvious today that the amount of data which is produced at various institutions exceeds the capacity of computability. Thus, instead of collecting information we indeed obtain"mountains of redundancy," because the information in the data is not extractable. Computers cannot provide further improvements on this for reasons stated in section 1 and 2.3. It is our strong belief that completely different information processing concepts are necessary to overcome these problems. There is now strong evidence that information system structures exist which measure, store, and process data in a similar way as biological systems do. In other words, systems of small physical size can be expected to be developed which represent more computation power than the largest computer today. Such systems can immediately also be used as measurement units by providing the appropriate sensor inputs. The sensing device does not need to be very accurate, since the system implementing algebras of relations is capable of eliminating the noise to any arbitrary small level by appropriate length of the measurement. The measurements are analyzed simultaneously such that whenever the measurement is terminated the classification process into the corresponding equivalence class is also completed and is coded by a universal language the user is free to define without restrictions other than logical consistency.

Such a system cannot be structured like a computer since it works asynchronously relative to the input. Especially this feature will make it appropriate for advanced experiment design for remote locations and for large or complicated information/data tasks.

5. References

1. Birkoff, G. "Modern Applied Algebra" Bartee, T.
   McGraw-Hill Book Company 1970, New York

2. Ash, R. "Information Theory", Interscience Publishers,
   New York 1967

3. Kolmogorov, A. "Logical Basis for Information Theory and
   Probability Theory", IEEE Transactions IT-14, No. 5
   September 1968, pp. 662-664

4. Fine, T. "On the Apparent Convergeuce of Relative
   Frequency and Its Implications", IEEE Transactions
   IT-16, No. 3, May 1970, pp. 251-257

5. Solomonoff, R.J. "A Formal Theory of Inductive Inference",
   Information and Control 7, 1-22 (1964)

6. Chaitin, G.J. "On the Length of Programs for Computing
   Finite Binary Sequences", Journal ACM, Vol. 13,
   No. 4, Oct. 1966, pp. 547-569

7. Shannon, C.E. "A Mathematical Theory of Communication",
   Bell System Tech. J, 27, pp. 379-423, 623-656, 1948

8. Friedlander, S.K. "The Characterization of Aerosols
   Distributed with Respect to Site and Chemical
   Composition".

9. Artem'er, V.V. "Reactivity Measurements and Information",
   Translated from Radiokhimiya, Vol. 4, No. 2, pp. 211-
   215, March - April, 1962

10. Lindley, D.V. "On a Measure of the Information Provided
    by an Experiment". Papers presented at the Chapel
    Hill and Berkeley meetings of the Institute of
    Mathematical Statistics in April and July, 1955.

11. Boulliou, T. L. et al.,"Statistical Design and Data Analysis
    Techniques for Space Station Applications," NASA Contract
    NAS 8-26210, June 1971, MFC.

12. Azierman et al.,"Logic and Automata," Academic Press (1971).