The Dripping Faucet as a Model Chaotic System

by Robert Shaw

Abstract

Water drops falling from an orifice present a system which is both easily accessible to experiment and common in everyday life. As the flow rate is varied, many features of the phenomenology of nonlinear systems can be seen, including chaotic transitions, familiar and unfamiliar bifurcation sequences, hysteresis, and multiple basins of attraction.

Observation of a physical system in a chaotic regime raises general questions concerning the modeling process. Given a stream of data from an experiment, how does one construct a representation of the deterministic aspects of the system?

Elementary information theory provides a basis for quantifying the predictability of noisy dynamical systems. Examples are given from the experimental data of computations of the two dynamical invariants: a) the information stored in a system, and b) the entropy, or rate of loss of this information.

Contents

In a lecture entitled, "Is the end in sight for theoretical physics?",
Steven Hawking considers the possibility that all fundamental physical laws may
soon be known[1]. He comments: "It is a tribute to how far we have come already
in theoretical physics that it now takes enormous machines and a great deal of
money to perform an experiment whose results we cannot predict." If a descrip-
tion of Nature is to consist solely of elementary particles and the forces
between them, this statement might be true. However systems made up of a
number, even a small number, of subunits can behave in ways which completely
transcend our present understanding. The motion of fluids provides a good exam-
ple of how a vast phenomenology can arise which is to a considerable degree
independent of microscopic physics. We live in a whirl of moving structures,
swept by social, economic, and personal currents whose dominant theme is one of
unpredictability. Yet laws, constraints of some sort seem to be operating, as
evinced by our ability to function. The central issue of physics, that of pred-
ictability, is in fact addressed as a practical matter by each newborn infant:
How do we construct a model from a stream of experimental data which we have not
seen before? How do we use the model to make predictions? What are the limits
of our predictive ability? Simple experiments, as well as the experience of
daily living, still have much to teach us. Here, as a case in point, is an
experimental study of a dripping faucet.

1. Introduction

A common complaint of insomniacs is a leaking faucet. No matter how
severely the tap is wrenched shut, water squeezes through, the steady, clocklike
sound of the falling drops often seems just loud enough to preclude sleep. If
the leak happens to be worse, the patter of the drops can be more rapid, and
irregular. A dripping faucet is an example of a system capable of a chaotic
transition, the same system can change from a periodic and predictable to an

aperiodic, quasi-random pattern of behavior, as a single parameter (in this case, the flow rate) is varied. Such a transition can readily be seen by eye in many faucets, and is an experiment well worth performing in the privacy of one's own kitchen. If you slowly turn up the flow rate, you can often find a regime where the drops, while still separate and distinct, fall in an irregular, never-repeating pattern. The pipe is fixed, the pressure is fixed; what is the source of the irregularity?

Only recently has it been generally realized that simple mechanical oscillators can undergo a transition from predictable to unpredictable behavior analogous to the transition from laminar to turbulent flow in a fluid. A random component in a physical variable is not necessarily due to the interaction of that variable with a great many degrees of freedom, but can be due to "chaotic dynamics" among just a few. A system in such a regime is characterized by short-term predictability, but long-term unpredictability. The system state at one instant of time is causally disconnected with its state far enough into the future. The study of simple nonlinear models capable of chaotic behavior has become an active area of physics and mathematics research, see references [2] for an introduction.

The existence of dynamical chaotic behavior places severe limits on our ability to predict the future of even completely deterministic systems. The question soon arises, if a system is unpredictable, how unpredictable is it? Can we quantify a degree of predictability? Also, can we tell the difference between randomness arising from the interaction of many degrees of freedom, and randomness due to the chaotic dynamics of only a few?

Limits on predictability imply a loss of information, and motivates a look at the formalism and concepts of information theory. Information theory, as developed largely by Shannon, often provides a natural and satisfying framework

for the study of predictability. A main goal of this paper is to demonstrate, by example, how one can take a stream of data from a physical experiment and compute quantities such as a) the amount of information a system is capable of storing, or transmitting from one instant of time to the next, and b) the rate of loss of this information. The paper will conclude with a more general discussion of the modeling process.
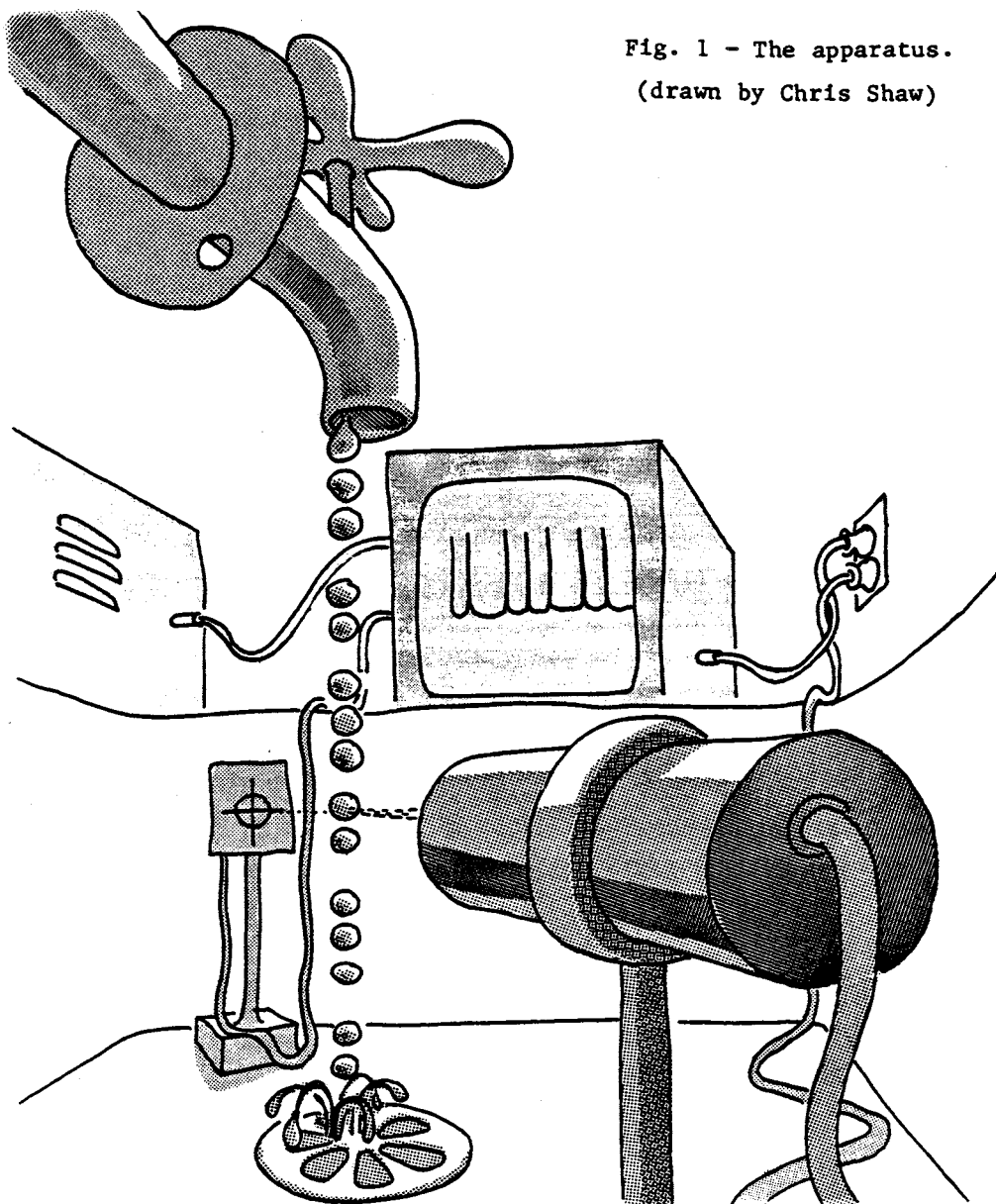

Fig. 1 - The apparatus.
(drawn by Chris Shaw)

2. The experiment

The dripping faucet has been used as an example of an everyday chaotic system in lectures by Rössler[3], Ruelle, myself, and others, but to my knowledge, this is the first experimental study. Experimental work is being performed at UCSC, in collaboration with Peter Scott. Figure 1 illustrates the simple apparatus.

Water from a large tank is metered through a valve to a brass nozzle with about a 1 mm. orifice. Drops falling from the orifice break a light beam, producing pulses in a photocell signal. The pulses operate a timer, yielding the basic data of this experiment, time intervals between successive drops. A Z80 based microcomputer, built by Jim Crutchfield, controls the valve via a stepper motor, reads and resets the timer, and collects data. Timer resolution is 4 microseconds.

Chief sources of external noise are vibration and air currents, to which the system is quite sensitive, and drifts in the flow rate, probably due to the strong temperature dependence of both viscosity and surface tension. However the system is easy to operate, with a little care, and produced interesting data almost immediately. Drop rates of interest were in the range of 1 to 10 per second, corresponding to flow rates of about 30 to 300 gpf [4].

3. Modeling strategy, some expectations about the data

A traditional physics approach to modeling the dripping faucet might be to try and write down the equations of motion, coupled differential equations describing fluid flow and the forces of surface tension, and solve them with appropriate boundary conditions. But this is an exceedingly difficult problem. The calculation of the shape of even a static hanging drop is at present a state-of-the-art computer calculation[5]. A numerical foray into the largely

unexplored world of three-dimensional coupled nonlinear PDE´s is probably
overambitious.

Another, more tractable, approach is to try and build a model directly from
the data, without attempting a model from Newton´s equations and first princi-
ples. Here we forget the physics entirely, and view the system as a black box
producing a stream of numbers T_1, T_2, T_3, ... in this case the drop intervals.
All our questions about predictability are posed abstractly, in terms of these
numbers.

Is there a causal relation between T_n and T_{n+1}, between one drop interval
and the next? A common method for testing a stream of numbers for determinism
is to construct a "return map", the numbers are taken in pairs (T_n, T_{n+1}), and a
point is plotted for each n.

Figure 2 shows how such a plot would appear for two extreme types of data.
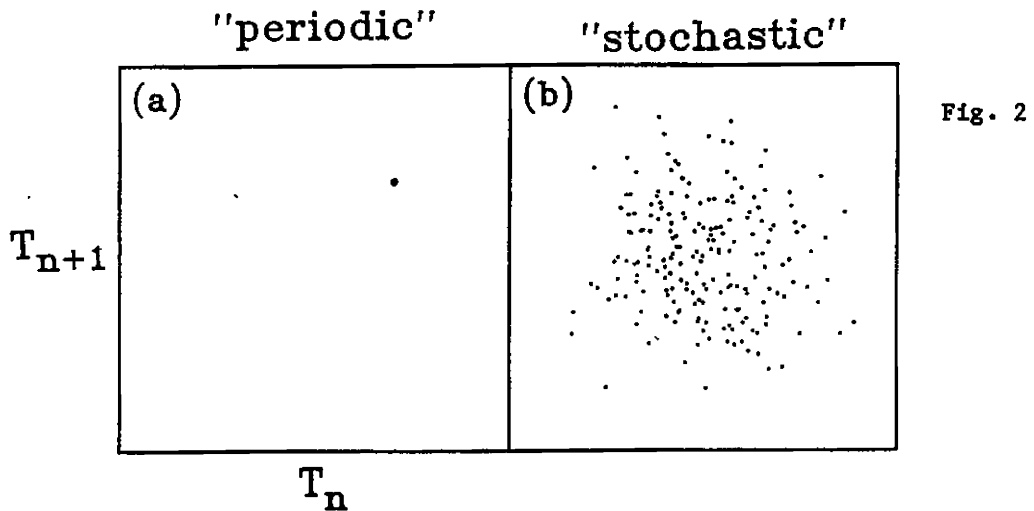

"periodic"   "stochastic"

Fig. 2

T_{n+1}

T_n

a) If the drops are absolutely periodic, then $T_1 = T_2 = T_3$, etc, and each pair of
numbers plots the same point.

b) Suppose each number was selected at random from some distribution of possi-
ble intervals $P(T_n)$, with no correlation between one number and the next.

In the latter case, the plotted pairs would produce a scatter plot, reflecting the absence of any "history" in the number stream. The distribution of points on the scatter plot would be uniquely determined by the rule for statistical independence, $P(T_n, T_{n+1}) = P(T_n)P(T_{n+1})$. This provides the definition for a "purely stochastic", or Markov, number stream.

In general we expect, on physical grounds, something in-between. Systems are usually causal, with the information stored in the dynamical variables providing continuity through time, but we also know that every physical system is subject to noise, which limits the total amount of information it can carry.

Of course, we cannot expect to construct a model of the entire drop dynamics from a series of drop intervals. First, the discretization of the continuous system effected by looking only at drop intervals performs, roughly, a "Poincaré cross-section" of the flow in the system state space. For example, a regular oscillatory dropping motion becomes a single point in the return map. One could imagine many different oscillations with the same period. Even worse, we are projecting the presumably infinite-dimensional dynamics onto a single time variable. The fact that this procedure can yield any structure at all makes a statement about the ability of continuous systems to behave in a "low-dimensional" fashion.

Nevertheless we can hope, if the system is "low-dimensional" enough, to compute the same measures of predictability that we would obtain if we had available much more complete data, for example the position of the whole surface of the drop as a function of time. In order to be useful, quantities such as the "information stored in a system" must be properties of the system, and not the particular type of measurement. This implies an <u>invariance</u> <u>under</u> <u>coordinate</u> <u>transformations</u>, a property which appropriately defined measures of information possess. Thus we can hope that, in some cases, a "good enough" measurement of

<u>any</u> dynamical variable will serve to characterize the predictability.

4. Some phenomenology

The behavior of our dripping faucet system exhibits a rich phenomenology as the flow rate is varied, one which we have not yet systematically investigated. The qualitative type of behavior at a given flow rate, e.g. periodic or chaotic, can depend on initial conditions, indicating multiple "basins of attraction", and changes in behavior as a function of flow rate are often sudden and hysteretic. The drop rate is not a monotone function of the flow rate. An overview of the faucet behavior will appear[6], but at the moment only preliminary results will be presented, consisting of examples of the different types of behavior which can occur.

At low flow rates, the system is in its "water clock" regime, quite strict periodicity is typical. At somewhat increased flow rates, corresponding to about 5 drops per second with our nozzle, variations in the drop intervals on the order of 5-10% start to appear. These can be difficult to detect by eye, the drops still appearing periodic, but the time vs. time maps show that variations are present, and quite structured.

Figure 3 illustrates one type of behavior, every other drop interval is somewhat longer than its predecessor, corresponding physically to a pairing up of the drops, as schematically illustrated to the right of the figure. This is an example of a "period doubling" bifurcation, <u>two</u> drops have to fall before the system behavior repeats itself, and successive points on the time vs. time map alternate between two locations.
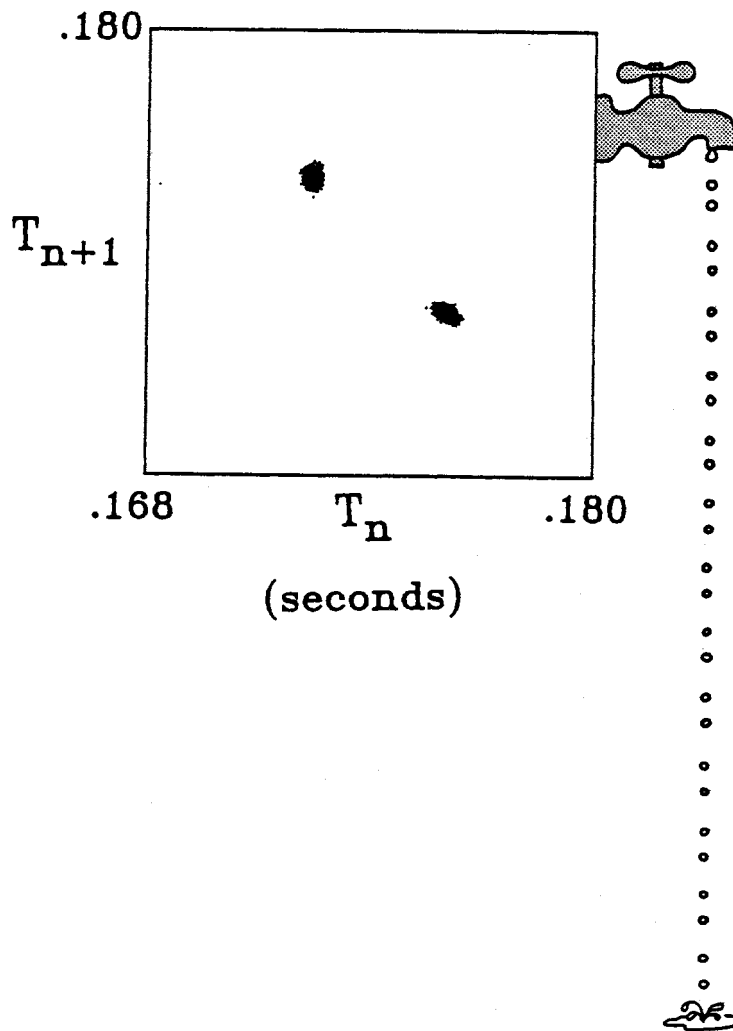
Fig. 3 – Period-two drop regime.

Another flow rate might yield a picture such as Fig. 4a. Here no periodicity is apparent, but the system clearly is not totally random either, a considerable degree of determinism connects one drop interval with the next. The time vs. time map could be approximated by a parabola with some "pure stochastic" noise added. In fact, considerable study has been devoted lately to simple difference equations of the form $X_{n+1} = f(X_n) + \xi$, where $f(X_n)$ is a parabola, and ξ is a noise term, hopefully modeling a system with mixed deterministic and random elements[7]. For comparison, the plot resulting from iteration of such a one-dimensional map with added noise is shown in Fig 4b.
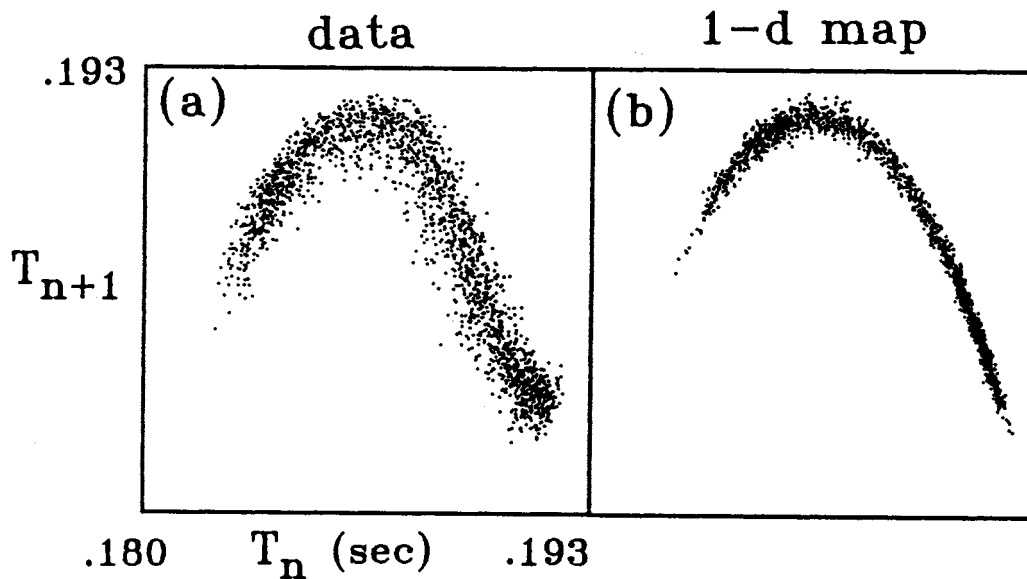

Fig. 4 – Data modeled by the 1-d map with noise:

$$x' = 3.6x(1-x) + .02\,\xi$$

Still another type of behavior is shown in Fig. 5a. This is a very familiar picture to those of us with experience with driven nonlinear oscillators, it is a Poincaré cross-section of a "two-band" attractor, which occurs in the driven Van der Pol oscillator, the driven Duffing oscillator, and many other systems. The two-band attractor is part of the complete period-doubling bifurcation sequence, as described originally by Grossman and Thomae in the context of one-dimensional maps[8], and Shimada and Nagashima[9], Lorenz[10], and Crutchfield et al[11] in the context of continuous flows. One should imagine trajectories moving along a ribbon which loops through the cutting plane twice, the ribbon folding into itself at some point along its length. This is an example of a system with a mixture of periodic and chaotic aspects, points strictly alternate between the two islands in the time vs. time map, but move chaotically within each island. This situation has been termed "noisy periodicity" by May[12], and "semiperiodicity" by Lorenz[10]. Similar pictures occur in iterated maps of the plane such as the Hénon mapping[13], as illustrated in Fig. 5b.
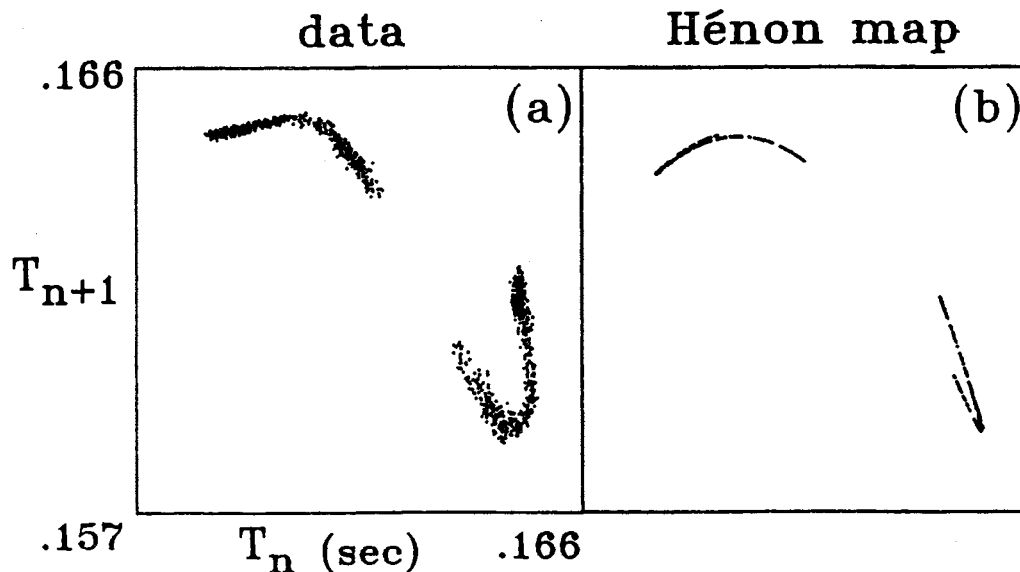

Fig. 5 - Data modeled by the Hénon map:

$$x' = y + 1 - 2.07x^2$$

$$y' = -.33x$$

At a certain flow rate, one starts to see pictures like Fig. 6. The data still has a one-dimensional character, but the map is tangled, one drop interval no longer uniquely determines the next. There apparently still exists a single coordinate which more or less specifies the state of the system, but this coordinate is no longer projected in a one-to-one fashion onto the drop interval. One could resolve the ambiguities in the time vs. time map with various conditional probability schemes, but a more entertaining method is to add another coordinate, namely T_{n+2}, to the display, and view the resulting time vs. time vs. time map as a stereo plot. In this fashion, we let the visual cortex do the work, and those able to view stereo pictures with crossed eyes can verify that the strings which appear to cross in the two-dimensional figure are in fact distinct, and one could in principle lay more reasonable system coordinates along them.
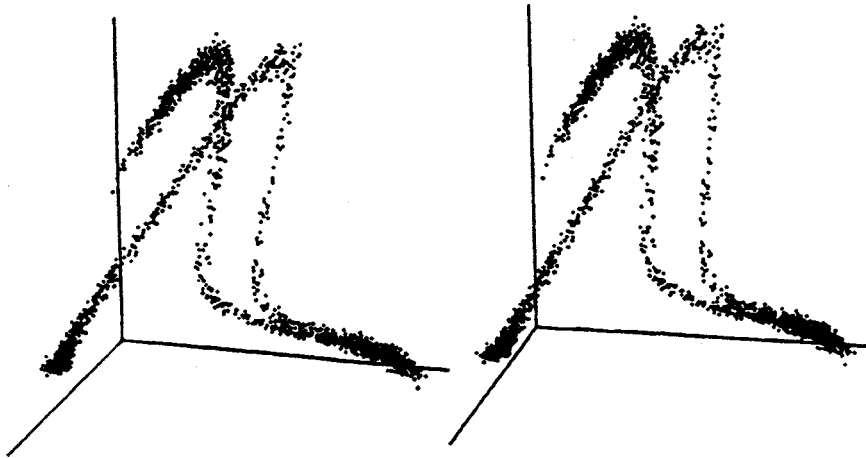
Fig. 6

12

All of the examples presented so far display a nearly one-dimensional character, that is, the value of a single coordinate, often just the drop interval, determines to within fairly narrow limits the value of the next interval. As far as the data is concerned, the state of the system can be specified by a value along a single coordinate. Furthermore, the behavior is familiar from the phenomenology of simple low-dimensional nonlinear systems, despite the fact that the dripping faucet is a continuous fluid system, with, in theory, an infinite number of degrees of freedom. We can only expect that eventually more of those "degrees of freedom" will assert themselves.

Figure 7 illustrates a type of behavior which appears at somewhat increased flow rates. Much of the data falls along one-dimensional curves, but there are regions of the time vs. time map which appear to be higher dimensional. Appendix 2 discusses the interesting possibility of the common occurrence of attractors of what might be called "mixed dimensionality"; in some regions of the state space, the state of the system can be optimally specified by the value of a single coordinate, but in other regions, more coordinates are required.
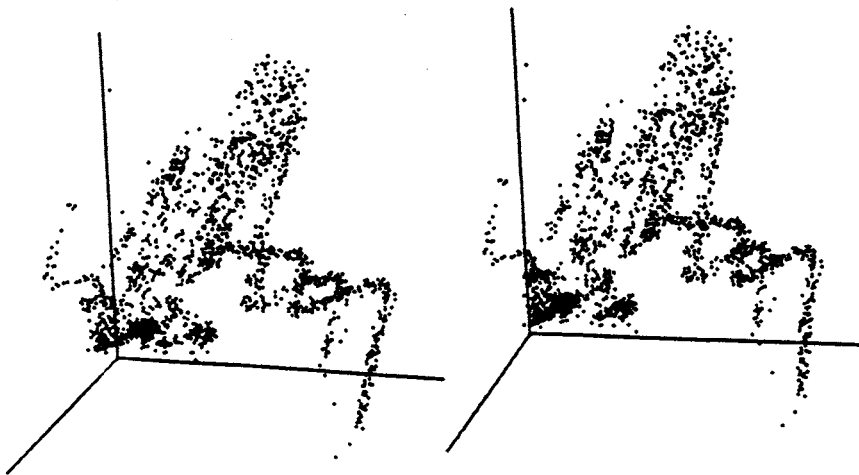
Fig. 7

At still higher flow rates, higher dimensional behavior becomes more and more evident. The variations in drop interval increase to 20% and more, and one obtains pictures like Fig. 8. There is still plenty of structure present, but, as the phenomenology of higher dimensional nonlinear systems is nearly completely unknown, studies are postponed to the future.

85   (msec)   104

32   (msec)   88

29   (msec)   35

28   (msec)   33
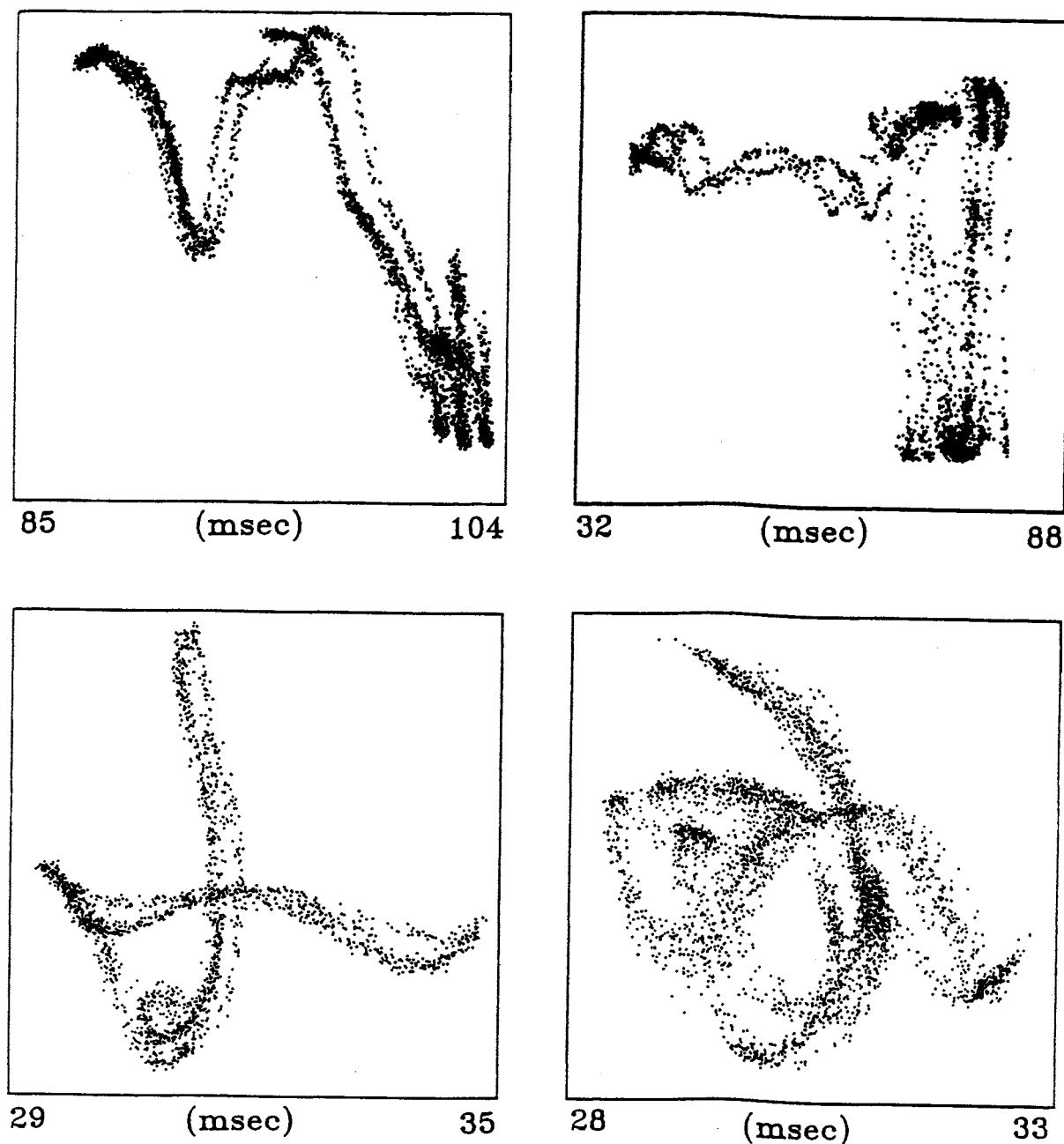
Fig. 8

Further work will be directed toward illuminating the new nonlinear phenomenology which is lurking in this system. A comforting feature of the non-linear jungle is that certain landscapes tend to reappear; whereas dynamical detail can vary endlessly from system to system, certain themes become familiar. For example, in the dripping faucet system we have recognized forms and bifurcation sequences common to low-dimensional models. Perhaps by studying faucet phenomena that is less familiar, we can build metaphors which will be of use in comprehending more complex behavior in other systems.

5. A naive analog model

The term "model" is commonly used in physics to describe a representation of a physical system not involving all of its relevant variables, but rather a much simpler set of variables and a dynamic between them which still manages to preserve some aspect of the qualitative behavior of the complete system. Clearly some degree of predictability will usually be lost in the simplification, but a simple model often has value in its conciseness and explanative power, as well as as a starting point for a more complete description.

Several years ago, I constructed a model of this type for a dripping faucet, and implemented it on an analog computer[14]. This simple, one-dimensional model is so crude as to have perhaps little or no predictive power, but it may be useful in describing to first order why the dripping faucet behaves as it does. When the water drop data collection system was constructed, it became a matter of curiosity to plug the analog computer in place of the physical system, and see if any comparisons could be made.

The model is sketched in Fig. 9. A mass, representing the drop, grows linearly in time, stretching a spring, representing the force of surface tension. When the spring reaches a certain length the mass is suddenly reduced,

representing a drop detaching, by an amount dependent on the speed of the mass when it reaches the critical distance. We thus have a driven nonlinear oscillator, the nonlinearity arising from the sudden change in mass, and with position, velocity, and mass providing the three variables required for the occurence of chaotic behavior in a system evolving in continuous time.

$$F = ma$$

$$mg - kx \approx m\ddot{x}$$

$$\dot{m} = \text{const.}$$

$$\Delta m \propto \dot{x}\Big|_{x=x_o}$$
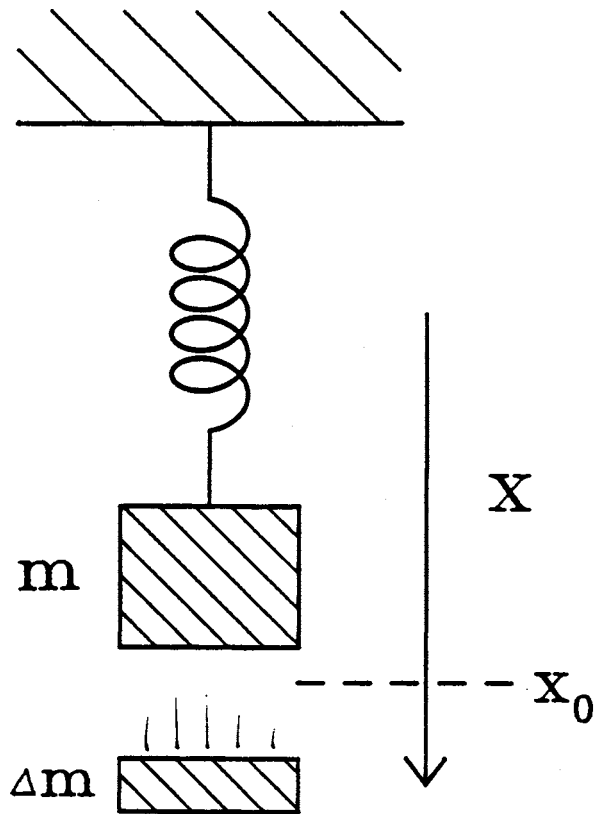
Fig. 9 – Analog model for water drop at low flow rates.
A drop detaches at a distance x_o indicated by
the dotted line.

Sure enough, as a parameter is increased, the analog model enters into periodic motion, then period-doubles, then goes chaotic. The time between "drops", or the sudden reductions in mass of the analog simulation, is the discrete model variable corresponding to the time intervals observed in the physical system. Physical faucet data can be found which closely resembles time vs. time maps obtained from the analog simulation, see for example Fig. 10.

data   analog model
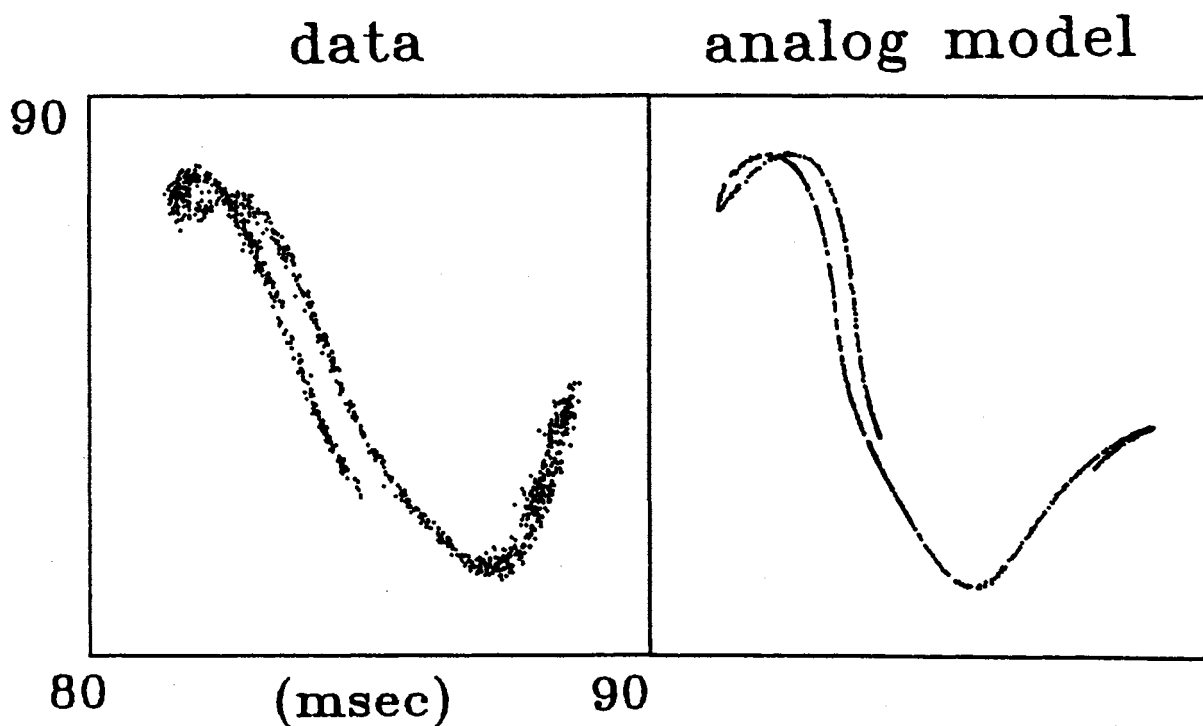
Fig. 10 – Comparison of experiment and "theory".

The analog simulation makes available continuous variables, which are not obtained from the physical experiment as it is presently conducted. By plotting the variables of the analog simulation, we can get an idea of the geometry of the "attractor" describing the motion of the fluid system. Fig. 11 plots two of the three continuous model variables, for the same parameter values yielding the time vs. time map of Fig. 10. This structure can be recognized as a Rössler attractor in its "screw type" or "funnel" parameter regime[15]. The close correspondence of model and experiment of Fig. 10 argues that such a structure is imbedded in the infinite-dimensional state space of the fluid system. Fig. 12 plots the position of the model drop as a function of time. The interaction of the phase of the natural drop oscillations with the discrete events of the drops detaching give rise to varying degrees of "rebound" of the remaining mass, and the resulting motion can be nonperiodic.
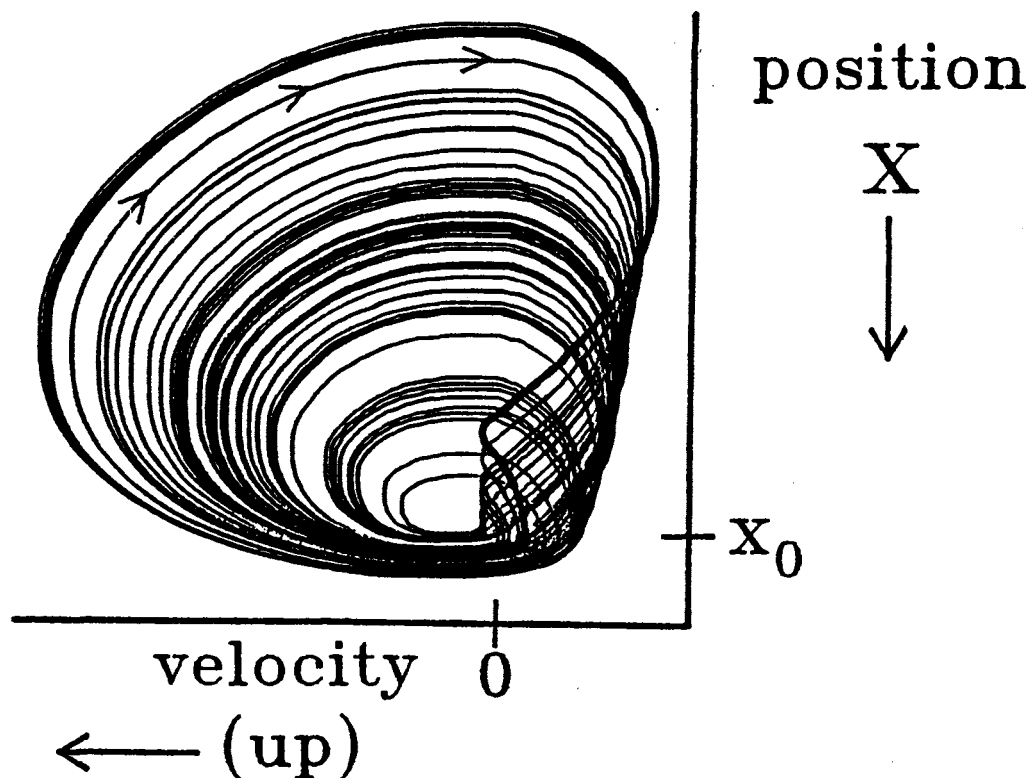
Fig. 11 - Attractor reconstructed from the analog model. Note the rapid rebound when a drop detaches at x_o.
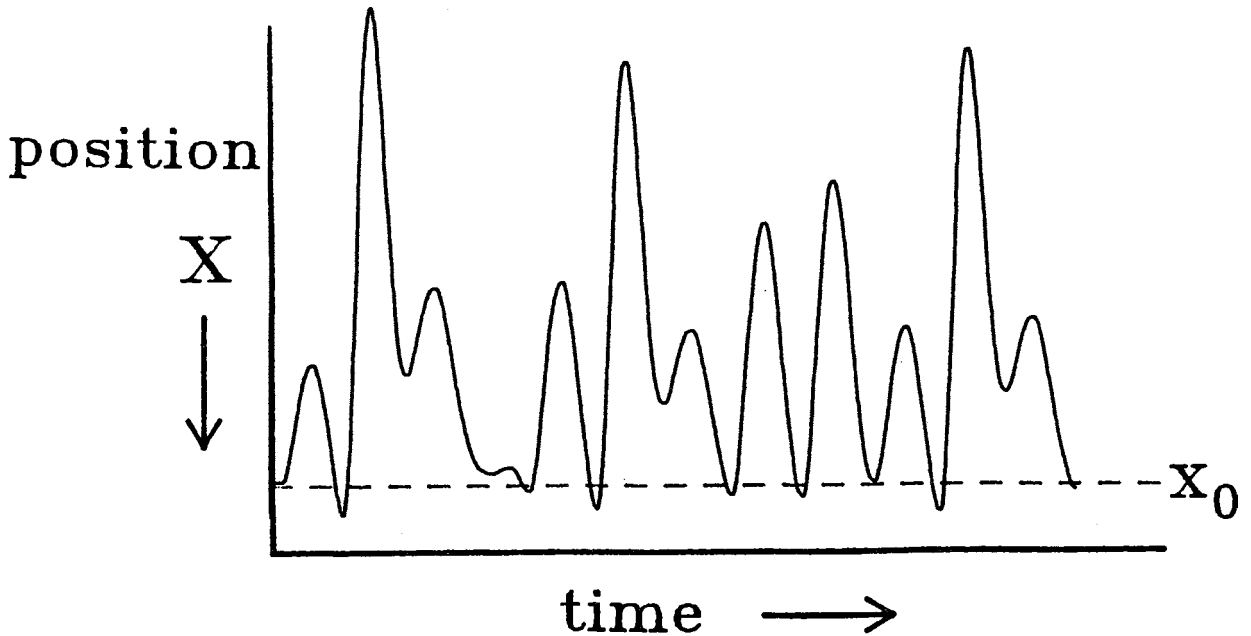
Fig. 12 - Analog model in the nonperiodic regime corresponding to Fig. 11.

No strong claims are made about the accuracy of this particular model, as nearly any driven nonlinear oscillator will behave this way in some parameter range. Nevertheless, a few points are well illustrated by this example.

a) A model of only a few dimensions can sometimes adequately describe the chaotic behavior of a continuum system, the high dimensionality of the fluid system is not required. This is in line with the original discussions of Lorenz[16], Ruelle and Takens[17], and Gollub and Swinney[18].

b) There exist fundamental geometrical structures which reappear in many different nonlinear systems. The equations of the original Rössler system are quite different from those of this model, yet the same attractors appear. Furthermore, Rössler's original work, as well as work in this laboratory and elsewhere, indicate that the changes in topology of the various Rössler

attractors as system parameters are varied is much more general than any one set of equations. Yet the parameter space of even the Rossler system, which produces the simplest chaotic attractor, is largely unexplored. Much work remains to be done to determine "typical" paths through parameter space, and the resulting changes in the behavior of a physical system.

. c) Analog and special purpose hardware could play a leading role in the development of our understanding of nonlinear systems. General purpose digital machines are admirably suited for accurate calculations in particular cases, but often do not serve well for rapid scans of large classes of equations where qualitative understanding is totally lacking. The model-data correspondence of Fig. 10 was obtained by a search through a three-dimensional parameter space which took only a few moments on an analog machine, parameters are varied simply by turning knobs, and feedback is immediate. The time and expense required for the same job on a standard digital machine is horrendous to contemplate. The radical reduction in price and increase in capability of integrated digital circuitry in the past decade enable the construction at low cost of special purpose and hybrid devices specifically designed to rapidly increase our general knowledge of dynamical systems theory. Alas, we have found this viewpoint difficult to communicate to scientific officialdom[19].

Illegitimi
non
Carborundum

6. Information theory and dynamics

We now turn to the question of describing and quantifying the predictability of a stream of numbers obtained from repeated measurements of a physical system. Before treating the water drop data, a brief general discussion of the relevance of information theory to studies of dynamical systems is perhaps in order. Although only well-known results in information theory will be described, a general appreciation has been lacking of their usefulness in describing the predictability of physical systems, particularly in the presence of noise. The viewpoint and vocabulary used have been developed over the past few years at Santa Cruz in collaboration with Jim Crutchfield, Doyne Farmer, and Norman Packard.

The concept of "information" as a physical entity, with a p log p type measure, was perhaps first appreciated by Boltzmann[20]. Gibbs, Szilard, and others addressed the concept with varying degrees of directness, but it was not until Shannon's work "The Mathematical Theory of Communication" that a clear and explicit discussion of information and its properties appeared[21]. Shannon carefully restricted his discussion to the transmission of information from a "transmitter" to a "receiver" through a "channel", each with known statistical properties. Some version of the following figure appears in the early pages of most books on information theory:
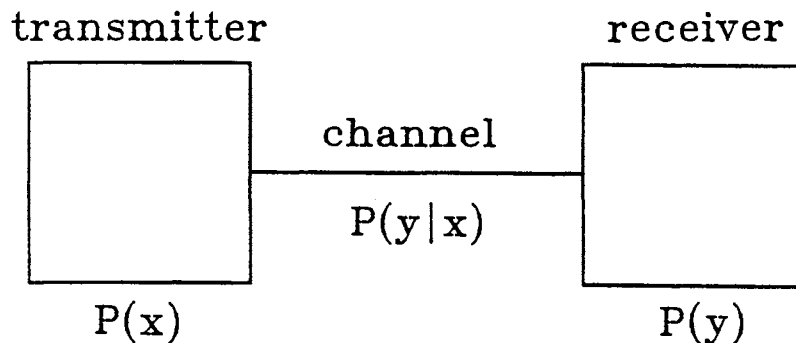
transmitter   receiver

channel

$P(y|x)$

Fig. 13

$P(x)$ $P(y)$

P(x) describes the distribution of possible transmitted messages x, P(y) the distribution of possible received messages y, and the properties of the "channel" connecting the two distributions is contained in the conditional distribution $P(y|x)$, the probability of receiving message y given that message x was transmitted.

The set of possible transmitted and received messages labeled by x and y may assume a number of forms. If the messages are selected from a finite set, for example the letters of the alphabet, then we can label them with an integer, and the conditional distribution becomes a finite matrix $P_{ji} = P(y_j|x_i)$, connecting $\{x_i\}$ with $\{y_j\}$. The set of x and y might be continuous, describing for example voltage levels, or they may be multidimensional.

If the distributions P(x), P(y), and $P(y|x)$ are available, quantities such as the "information" passing through the channel are well-defined and easily computable. In the case of communication systems, these distributions are, as a practical matter, readily available: transmitter, channel, and receiver hardware are man-made devices with known properties, and the first-order statistical properties of the English or other language used (which partially determine the "transmitter" distribution) can be estimated by constructing histograms from a large amount of text, as demonstrated by Shannon. From an economic point of view, Shannon's analysis was required at that point in history, to enable the development of modern communication networks.

The mapping of Shannon's communication metaphor onto the problem of predictability in dynamical systems theory is a small conceptual step. A dynamical system "communicates" some, but not necessarily all, information about its past state into the future:

past   **system dynamics**   **future**

$P(x'|x)$

$P(x)$ $P(x')$
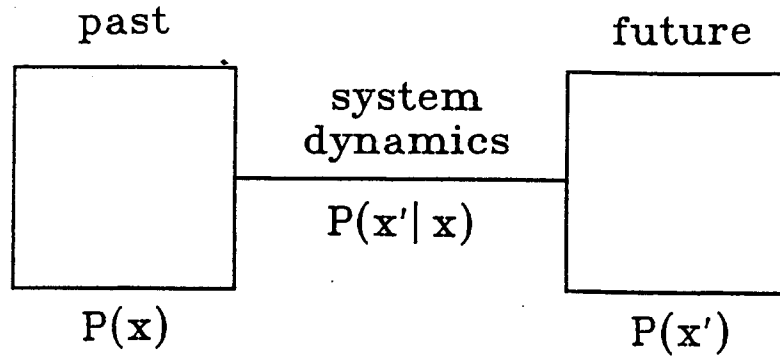
Fig. 14

Observable past and future system variables are described by x and x´, all the other time independent system properties are lumped into the conditional distribution $P(x´|x)$ which describes the causal connection between past and future given by the system dynamics.

Present theory usually treats only autonomous, non-evolving systems, hence x and x´ describe the <u>same</u> set of coordinates at two different times, $x = x_1(t_1), x_2(t_1), \ldots$ and $x´ = x_1(t_2), x_2(t_2), \ldots$ etc. This provides a simplification over the more general case in communication theory, where the transmitted and received symbols may be different. Again, the symbols x and x´ should be thought of as general labels for the coordinates, which might be discrete, continuous, or multidimensional.

<u>Knowledge</u> of the past and future system states is represented by probability distributions $P(x)$ and $P(x´)$ over the variables. If this correspondence can be carried out, the problem of prediction, formally at least, becomes simple, given $P(x)$ and $P(x´|x)$, compute $P(x´)$:

$$P(x') = \int P(x'|x) \, P(x) \, dx$$

In this picture, the "system dynamics" $P(x´|x)$ defines a <u>map</u> from distributions to distributions:
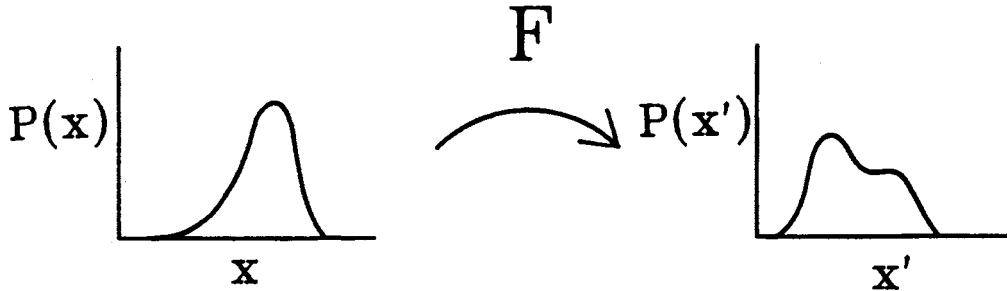
Fig. 15 - Distributions to distributions.

To study the behavior of the system farther into the future than t_2, we simply reapply the mapping F to the distribution $P(x')$, producing a new predicted distribution at t_3. Repeating this process models changes in knowledge of the system variables as it moves through time. Much of the modern work in dynamics involves studies of these sorts of iterated mappings. Classically, we can discuss systems evolving in continuous time by considering mappings which move the system forward in time by an infinitesimal amount.

In the case of the water drop system, the measured system variable is the drop interval T_n, and the dynamics are given by the conditional distributions $P(T_{n+1}|T_n)$ which relate one drop interval to the next. The time variable is discrete, and is labeled by the drop number n.

Intuitively, we expect highly peaked probability distributions to represent fairly definite knowledge about the variables, and broad distributions relative ignorance. Thus, widening of a distribution under the dynamics will indicate a loss of knowledge, or predictive ability. Our task will be to quantify this notion.

24

7. The minimum information distribution $\overline{P}(x)$

Of all the possible distributions $P(x)$ over the variables x representing various guesses about their positions, there is one special distribution $\overline{P}(x)$ representing a __minimum knowledge of the system state__. This is our best guess about the variable positions assuming that we have __total__ knowledge of the time-independent part of the system, the "equations of motion" $P(x^{\prime}|x)$, but __zero__ additional knowledge about the variables.

Usually, the best representation of minimum knowledge of the position of some variable is a flat distribution over the range of that variable, but if the variable is part of a known system, our expectation may be otherwise. For example, consider an observation of the position of a harmonic oscillator with known energy but unknown phase. Our expectation is that we are more likely to find the oscillator near the ends of its travel, as it moves slower there. The probability of finding the oscillator at some position is proportional to the inverse of the velocity at that position.
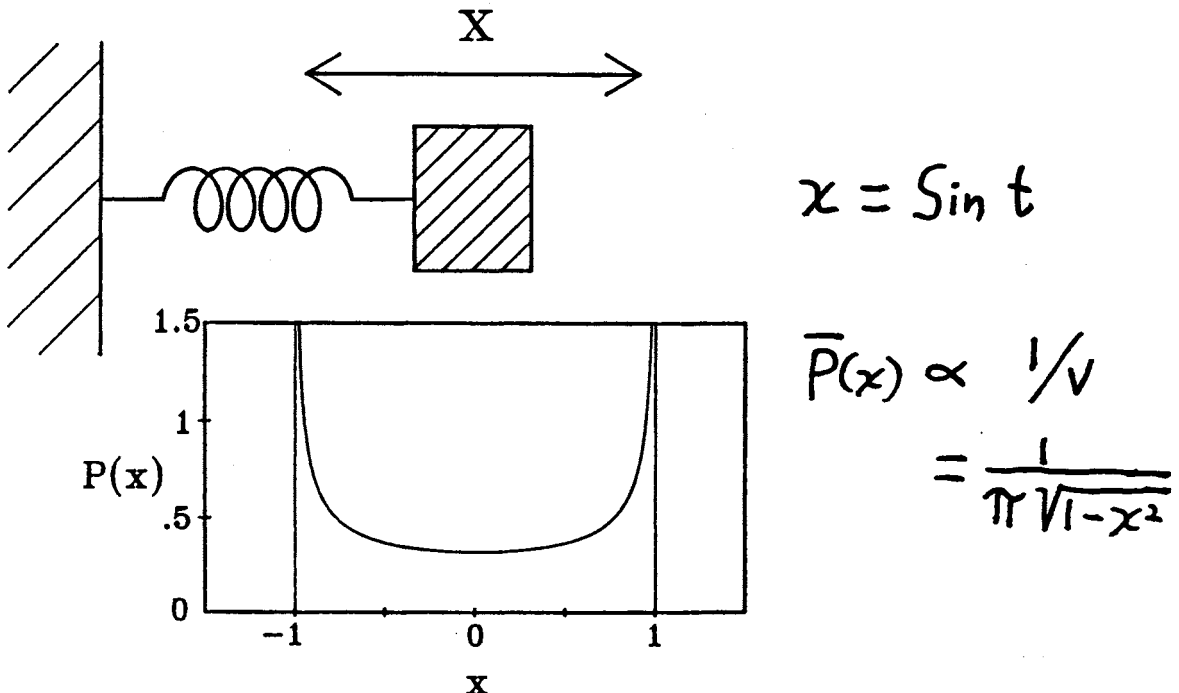

$$x = \sin t$$

$$\overline{P}(x) \propto 1/v = \frac{1}{\pi \sqrt{1-x^2}}$$

Fig. 16 - Minimum information distribution for the simple harmonic oscillator.

The minimum information distribution $\overline{P}(x)$ is thus a property of the system $P(x'|x)$, or of a particular "basin" of the system, if the system state space can be decomposed into independent dynamical regions. The distribution $\overline{P}(x)$ is the baseline against which all more sharply peaked distributions will be measured. As we shall see, if a system has any tendency to lose information, all input distributions $P(x)$ will move toward $\overline{P}(x)$ in a quantifiable sense.

If the system is incapable of transmitting information into the indefinite future, then any initial distribution $P(x)$ will relax in time to the asymptotic "equilibrium distribution" $\overline{P}(x)$. The final minimum state of knowledge will be independent of any initial information. This is the original and intuitively correct concept Boltzmann described by the word "ergodic". The harmonic oscillator is clearly not ergodic in this sense, as it can preserve phase information indefinitely.

$\overline{P}(x)$ also has the property that it is unchanged by the system dynamics, that is, it is invariant under the map F. Clearly $\overline{P}(x)$ must be the same distribution as $\overline{P}(x')$. Otherwise, a prediction could be made about a change of state of the system at a particular time, and our knowledge is not at a minimum. Hence the term for $\overline{P}(x)$ found in the mathematics literature, "invariant measure".

If a system can assume only a finite number of states x_i, the dynamical rule connecting states becomes a finite transition matrix P_{ji}. In this context, the unique asymptotic state vector $\overline{P}_i$ is described by a classical theorem of Frobenius: A matrix consisting of all positive elements has only one positive eigenvalue, furthermore, any positive vector will approach the direction of the unique positive eigenvector under repeated application of the matrix P_{ji}. This fact allows an easy computation of the invariant distribution of some map, one simply starts with any distribution P_i and repetitively applies the map until

the "power method" for finding the largest eigenvalue of a matrix, and is dis-
cussed in the literature under the label "Frobenius-Perron operator". Its use
will be demonstrated later in this paper, and it was explicitly applied in the
dynamical systems context in Ref. [22].

The division of seamless reality into "dynamics" $P(x'|x)$ and "variables" x
is to some extent arbitrary, no physical system is truly independent of time.
But once such a model has been constructed, a __single__ $\overline{P}(x)$ is uniquely defined
for any time-invariant system by the requirement that it describe a minimum
knowledge of the system state.

8. Quantification of predictability

The appropriate measure to quantify an amount of knowledge contained in a
distribution $P(x)$ is the Boltzmann integral

$$I = \int P(x) \log P(x)\, dx$$

This measure yields zero when applied to a flat distribution $P(x) = 1$, and
becomes increasingly positive as the distribution peaks around particular values
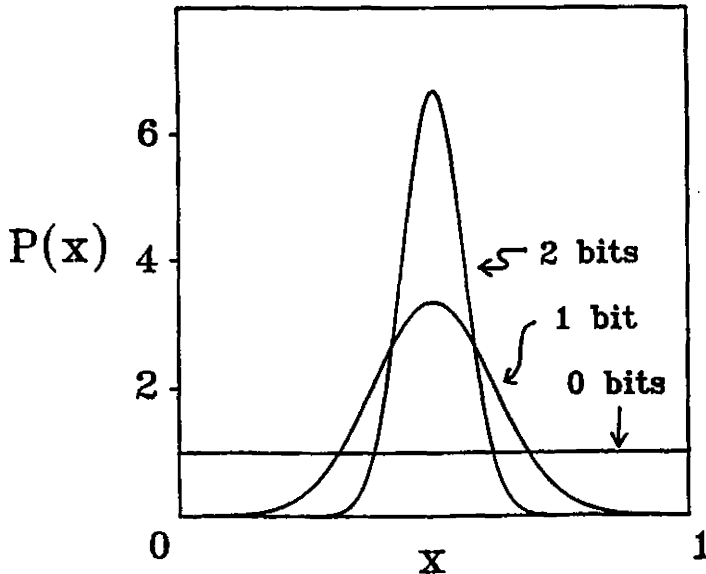in x:

Fig. 17

If the logarithm is taken to the base two, the units of this quantity are "bits".

The number resulting from this integral quantifies the amount of information gained when we learn, for example by experiment, that a variable is distributed as P(x), relative to a flat distribution, representing no particular knowledge of x before the experiment. It is important to realize that "information" in this sense is a <u>relative</u> concept, the information in a distribution P(x) is always measured relative to an a priori expectation $\overline{P}(x)$. The information measure is thus a function from a <u>pair</u> of distributions to the real numbers. If we know something about the distribution of x even before we learn P(x), our a priori expectation $\overline{P}(x)$ is not flat, and the appropriate formula is

$$I(P, \widetilde{P}) = \int P(x) \log \frac{P(x)}{\overline{P}(x)} \, dx$$

This is the "information gain", a sort of distance from the expectation $\overline{P}(x)$ to the new information P(x), which is independent of the particular coordinates used to describe the variables x. The information gain is then <u>invariant</u> under smooth coordinate changes on x.

The information gain is a concept familiar in the context of statistical mechanics[23]. Here the distribution $\overline{P}(x)$ characterizing minimum information, or "thermal equilibrium", is the Boltzmann distribution. If the states of a system have an energy distribution differing from the Boltzmann, then there is a "free energy" which in principle can be extracted from the system without lowering its temperature. The <u>knowledge</u> of the system which allows us to extract "work" from it is the distance from the actual distribution of energies to the Boltzmann, measured by the information gain. The available free energy is simply the information gain multiplied by kT, $\Delta F = I(P,\overline{P})$ kT ln2.

The informational description of dynamical systems is somewhat less straightforward than the static situation of statistical mechanics. For example, the Boltzmann and other familiar distributions characterizing static

equilibria are obtained directly by minimizing the information integral P log P subject to appropriate constraints[21]. But in the dynamical case, the "constraints" are given by equations of motion, and $\overline{P}(x)$ is not so easy to determine. Typically, $\overline{P}(x)$ is experimentally measured, or computed from equations of motion via an iterative procedure, examples will be given in a later section.

Any actual measurement of a physical variable x, corresponding to a sample from the distribution $\overline{P}(x)$, must be to finite precision, typically we know only that x is within some interval x_0 to $x_0 + dx$. This motivates modeling the measurement process by creating a "partition", that is, breaking up a continuous variable x into a finite number of subintervals x_i. Repeated measurements, or samples from $\overline{P}(x)$, gives a string of whatever symbols we are using to label the $\{x_i\}$, with probabilities p_i determined by how much of the probability density $\overline{P}(x)$ falls in each little bin x_i.

The degree of unpredictability, or entropy of the sequence of symbols is likewise quantifiable:

$$H = -\sum_i P_i \log P_i$$

This measure increases with the degree of randomness, if the sequence simply repeats a single symbol x_0, then $p_{x0} = 1$ and all other probabilities are zero, the sequence is totally predictable and the entropy per symbol is zero. The maximum value is obtained when all possible symbols are equally likely to occur.

The entropy, as simply defined above, is clearly not a coordinate invariant quantity. It varies as the partition changes, and generally will increase as the partition is made finer, creating more possible symbols. To see the relation between the information $I(P,\overline{P})$ of P(x) with respect to $\overline{P}(x)$, and the entropy of a finite partition $\{x_i\}$ of the continuous variable x, consider the special partition of $\overline{P}(x)$ which cuts it into N equiprobable elements:
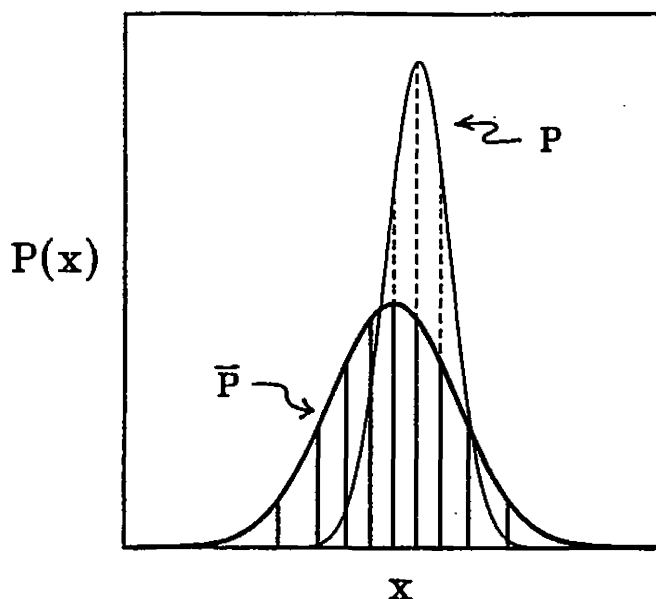
Fig. 18

The entropy of $\overline{P}(x)$ with respect to this partition is simply log(N). Now the entropy of the narrower distribution P(x) with respect to the same partition is approximately $H = log(N) - I(P,\overline{P})$. That is, the information provides a constraint on the entropy, lowering it from the maximum value it could have with a given number of elements in the partition. The equality of this relation becomes more exact as the number of elements grows. Here the entropy measures the degree of unpredictability of a series of discrete symbols, and information the degree of predictability of events arising from a continuous distribution, with respect to a background expectation. Entropy and information defined this way are both positive quantities.

9. Entropy of purely deterministic systems

The most common treatment of classical mechanics has been purely deterministic. One considers the past state of a system as a "point" x, and computes its future position, another "point" x´, using a dynamical rule $x´ = F(x,t)$, usually obtained from Newton´s equations. Physical systems governed by linear forces, central forces, and a few others respond well to this treatment, eclipses are predictable far into the future, and the world is full of clocks

and gyroscopes. But generalizing from this experience contributes to a false world view. Poincaré realized that even in celestial mechanics determinism did not imply unlimited predictability, and the lesson of recent work in nonlinear dynamics is that in even the simplest of "purely deterministic" dynamical systems our predictability may be, as a practical matter, extremely limited.

Difficulties in predictability arise because, in many systems, nearby "points" are spread apart exponentially fast by the deterministic dynamics, "errors" grow uncontrollably. The motion of the macroscopic variables becomes dominated by microscopic fluctuations, supplied by the heat bath for any physical system, and the system displays "chaotic behavior". The average rate of spreading becomes an important parameter of a given system, its "entropy", governing its degree of unpredictability. The existence of this quantity, and the utility of Shannon's information theory in dynamics, was shown by Kolmogorov in a 1959 article, "Entropy per unit time as a metric invariant of automorphisms"[25].

Although the variables of a purely deterministic description are usually considered to be continuous, attempts to compute measures of predictability by performing continuous integrations quickly run into a practical problem. Here the function governing the dynamics takes "points" to "points":
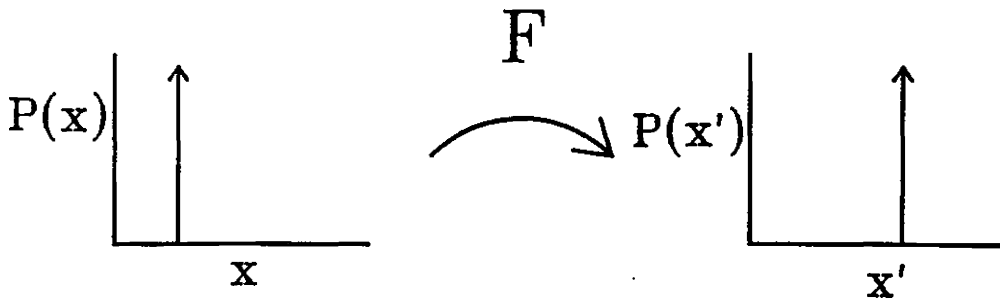

Fig. 19 – Points to points.

The conditional distributions $P(x^-|x)$ are delta functions, and terms such as log $P(x^-|x)$ appearing in the integrals diverge. This corresponds to the fact that a "point" represents an <u>infinite</u> amount of information, to specify its position completely would require an infinite number of digits. The concept of "point", while useful, is thoroughly unphysical, being a convenient name for "arbitrarily sharp probability distribution".

Kolmogorov and Sinai found a way around this difficulty, while partially preserving the fiction of "pure determinism". They applied a "partition" to the domain of the continuous variables x and x^-, breaking it up into small elements. The deterministic dynamics can now map an image of each little block, conserving the underlying probability:
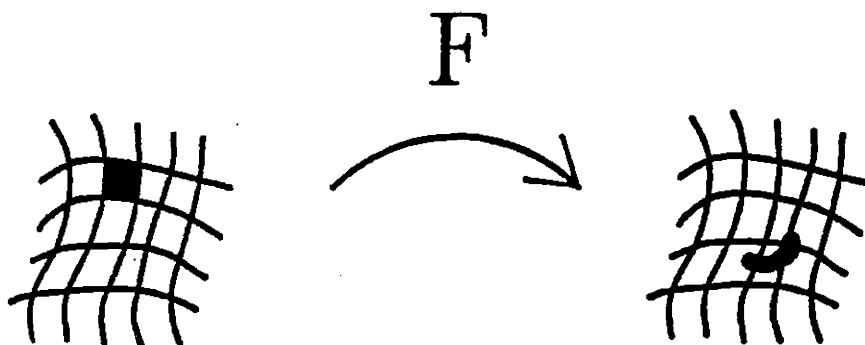
Fig. 20

As the above familiar picture shows, the image of a little block in x may fall over several blocks in x^-. With a certain probability a "point" starting in the block in x may arrive at any of these. By labeling the little blocks $\{x_i\}$ and $\{x^-_j\}$, one can construct a finite matrix of transition probabilities P_{ji}.

To compute the entropy, or degree of unpredictability, of a given system with respect to this partition, one now uses directly Shannon's formula for the information received from a discrete channel when the input is known, $H(x^-|x)$. Anything coming out of the channel that could be predicted from the input is not "information" in this sense, there is no "surprise" associated with it.

The quantity $H(x'|x)$ correctly measures the unpredictability. To compute $H(x'|x)$ from the finite matrix P_{ji} one first considers the entropy of the possible symbols x'_j received given that a particular symbol x_0 has been transmitted, as given by the matrix entries P_{j0}:

$$H(x'|x_0) = -\sum_j P_{j0} \log P_{j0}$$

The <u>average</u> information produced by the system is obtained by weighting this sum by the probability $P(x_i)$ that the various possible x_i actually occur:

$$H(x'|x) = -\sum_{ij} \overline{P_i} P_{ji} \log P_{ji}$$

The number so computed will depend on the partition chosen, but one can now, conceptually at least, pass to the limit of an infinitely fine partition, or the "refinement" of the partition obtained by considering higher and higher interates of the map. Kolmogorov, Sinai, and others were able to show that, if one considers the set of all possible partitions, and selects the one which gives the largest numerical value for the entropy, the number one obtains is a topological invariant of the system, independent of the coordinate system used to describe the original continuous variables x. It should not be too surprising that the optimum partition is one whose number density of microscopic partition elements is proportional to the minimum information probability distribution $\overline{P}(x)$.

A very simple deterministic map which exhibits the main features of a deterministic system with a positive entropy is graphed below (for the nth time):
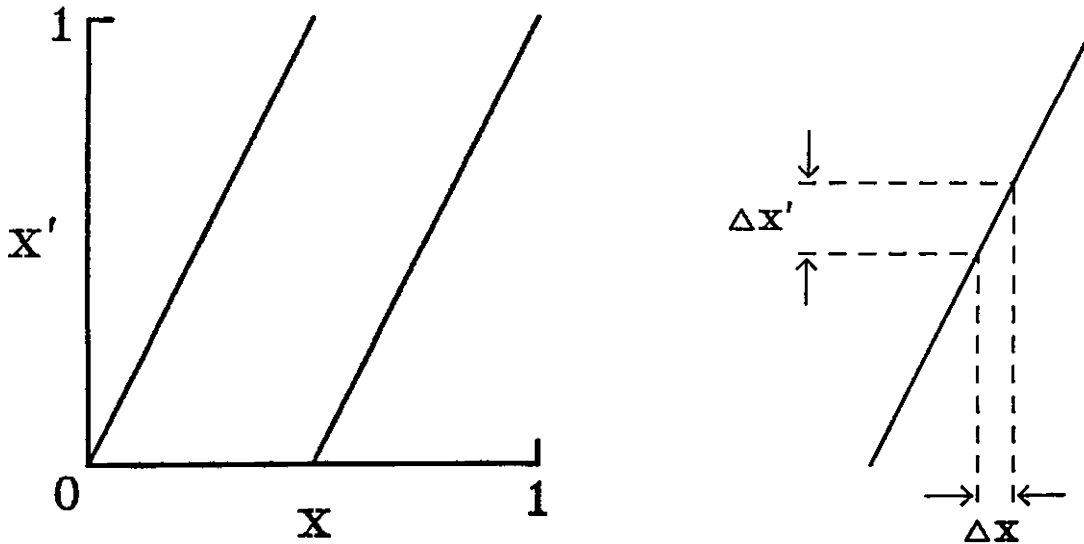
Fig. 21

When viewed from a "point to point" perspective, the map is purely deterministic, each "point" x yields a unique image "point" x´ = F(x). However, if one partitions the interval into small subintervals, representing the uncertainty which must exist in any measurement of x and x´, one finds that, because the slope of the map is two, any element in x will shadow two on the average in y, as indicated on the right-hand part of the figure. Because the uncertainty is doubled, there is an unpredictable element present in x´ of log(2)= 1 bit, regardless of how accurately we might know x. Thus the "purely deterministic" map acts as an information source[22]. It should be noted that the partitioning, no matter how fine, is a conceptual necessity.

The positive entropy of this map can also be seen by considering a point in x´ and trying to determine its preimage point in x. Because the map is two-onto-one, the inverse is non-unique, to go backwards in time, we must replace at each step the one bit which the system produced in the forward time direction. Entropies for these sort of deterministic "one-dimensional maps" can be computed this way by looking backwards in time, see work by Oono[26]. Considerable computer studies of the properties of one-dimensional maps have been performed at Santa Cruz[27].

Entropy measures are nothing more than the result of a simple counting procedure, in this case we are determining the ratio of "paths" to "states":

Fig. 22

$$X_i \longrightarrow X_i'$$
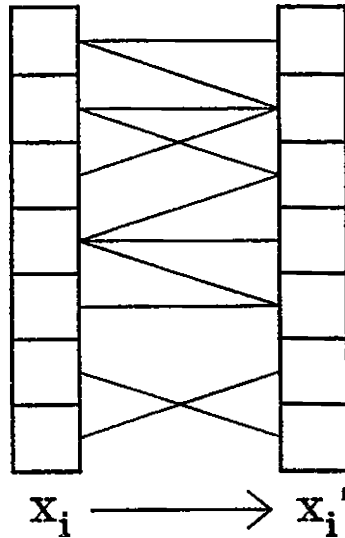
As the above figure, adapted from Shannon, illustrates, a partition divides the variable domain into a number of "states", and the deterministic dynamics connects them with a certain number of "paths". The entropy is simply the log of the ratio of "paths" to "states", where both are understood to be appropriately weighted by their probability of occurence. It is a property of the "purely deterministic" approximation that this ratio converges, regardless of how many "states" are considered.

A more physical interpretation of the entropy was suggested in ref.[22]. Any physical system will be imbedded in a "heat bath", producing random microscopic variations of the variables describing the system. If the deterministic approximation to the system dynamics has a positive entropy, these perturbations will be systematically amplified. The entropy describes the rate at which information flows from the microscopic variables up to the macroscopic. From this point of view, "chaotic" motion of macroscopic variables is not surprising, as it reflects directly the chaotic motion of the heat bath.

10. The noise problem

Difficulties arise in the definition of entropy discussed above when
"noise", or a stochastic element, is allowed to enter a purely deterministic
description of a system.   Typically the noise smears out the dynamics at some
length scale.   The function governing the dynamics can now be described as tak-
ing "points" to "distributions":
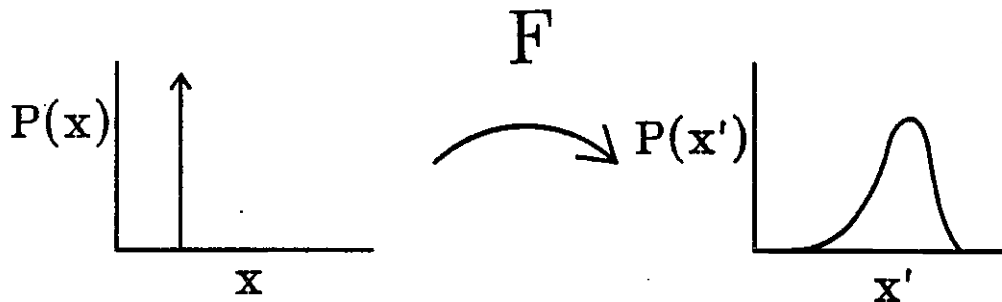
Fig. 23 — Points to distributions.

If we consider a sequence of input functions $P(x)$ of increasing narrowness cen-
tered around some value of x, we will reach a size below which the output dis-
tribution $P(x')$ won't change, we are "in the noise". We can, if we like, then
pass to the limit where $P(x)$ is a delta function, or "point", with no further
change in $P(x')$. It might be noted that this reasonable physical assumption
implies that the map F describing a system with noise is a linear functional.
The image of any $P(x)$ must be the same as that obtained by breaking up $P(x)$ into
weighted "points" x_i, and superposing the resulting weighted distributions
$P(x'|x_i)$.

if we try to apply the "partition" definition of entropy in this situation,
we will have trouble. As, during the limiting process, the partition becomes
finer than the length scale defined by the noise, the number of elements in
$\{x'_i\}$ accessible to a given x_i increases without bound:

$$X_i \longrightarrow X_i'$$
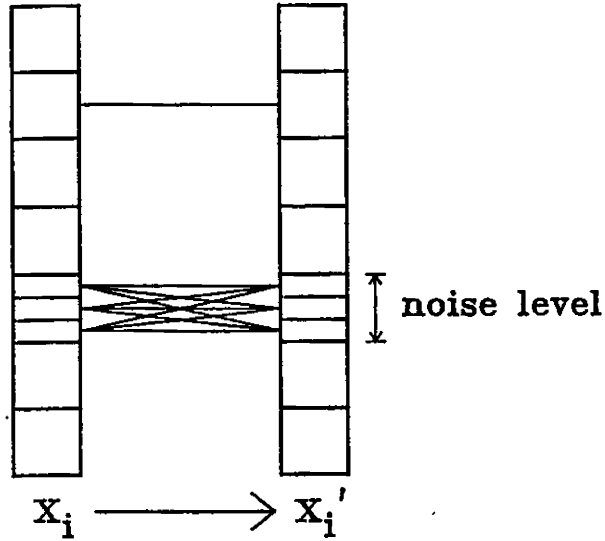
Fig. 24

The ratio of "paths" to "states" does not converge, and the entropy diverges with the log of the number of elements in the partition.

This in fact makes physical sense. If an experimenter with a good amplifier turns the gain way up, he will see plenty of "unpredictable information", in the form of thermal noise, no matter how well-behaved the input signal is. Further increasing the gain will not help him to resolve the input signal better, nor will it help him characterize the amplifier.

Apparently the quantity $H(x'|x)$ is not an appropriate measure of the predictability of a system with a noise component, as it actually measures the unpredictability, which will diverge as resolution is increased. The correct measure is obtained by an even more direct application of Shannon's communication channel image.

11. Information stored in a system

A "system" might be considered a body of information propagating through time. If there is any causal connection between past and future system states, then information can be said to be "stored" in the state variables, or communicated to the immediate future. This stored information places constraints on future states, and enables a degree of predictability. Again, if the system dynamics can be represented by a set of transition probabilities $P(x^-|x)$, this amount of predictability is directly quantifiable, and corresponds, in Shannon's vocabulary, to a particular rate of transmission of information through the "channel" defined by $P(x^-|x)$.

Imagine two observers with access to the same system, at two different instants of time. Suppose they both have complete knowledge of the system dynamics $P(x^-|x)$, but that the later observer knows nothing of the system state at the earlier time. The earlier observer could attempt to "communicate" to the later by putting the system in a particular one of its allowed configurations x. The dynamics would then carry the system through time to the later observer who would read off the resulting system state x^- and, perhaps with the help of a code book, discover the intended message.

Figure 25 illustrates two possibilities. Suppose the earlier observer attempts to communicate by restricting the system state to a narrow interval in x, described by $P(x)$, but that, at the later time, the distribution has relaxed back to the minimum knowledge state $\overline{P}(x^-)$. If this is true regardless of the input distribution then no communication is possible, and no information is stored in the system. If, on the other hand, the output distribution varies with the input, some information is passing into the future.

F

P(x)

x

P(x')

x'

no stored information

F

P(x)

x

P(x')

x'

some information stored
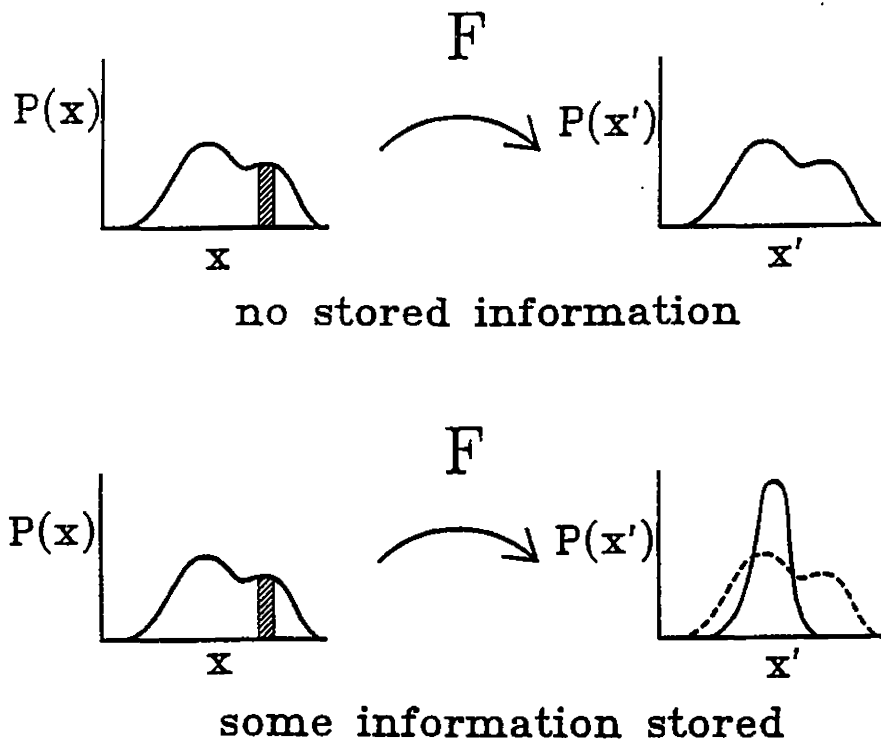
Fig. 25

Notice that, under the rules of this game, an unchanging and completely known "variable" communicates no information into the future. Although a static structure is certainly a predictable element, establishing continuity through time, it becomes a part of the fixed time-independent system description. The "information" measures described here are a property of the system dynamics. In mechanics, as in the human sphere, the transmission of information requires the possibility of change.

To quantify the average amount of information passing from past to future, first consider the earlier observer transmitting a particular x_0 and the later receiving some x'_0 from the distribution $P(x'|x_0)$. The information, or "surprise" associated with this event to the later observer will be measured only with respect to the asymptotic distribution $\overline{P}(x')$, as he does not know x_0. The mean over all x' emanating from some x_0 is

$$I(x'|x_0) = \int P(x'|x_0) \log \frac{P(x'|x_0)}{\overline{P}(x')} dx'$$

and, finally, the average information transmitted using an ensemble of input messages $P(x)$ is:

$$I(x'|x) = \int P(x) \int P(x'|x) \log \frac{P(x'|x)}{\overline{P}(x')} dx' dx$$

This quantity Shannon calls the <u>rate</u> of transmission of information for a continuous channel. He comments that this rate can vary up or down depending on the statistics of the input messages $P(x)$, some will be more suitable than others to the particular properties of the channel. The <u>channel</u> <u>capacity</u> he defines as the maximum rate one can find by varying $P(x)$ over all possible input ensembles, and using the optimum.

In the dynamical systems context, a particular input ensemble is selected by the properties of the system itself, namely the minimum information distribution $\overline{P}(x)$. As was mentioned earlier, this distribution has the property that an input $\overline{P}(x)$ produces an identical output distribution $\overline{P}(x')$, and it will automatically be generated by an autonomous dynamical system operating recursively. We can thus define the <u>information</u> <u>stored</u> <u>in</u> <u>a</u> <u>system</u> as the Shannon channel rate for an input (and output) distribution given by the equilibrium distribution for the system:

$$\text{Stored Information} \equiv \int \bar{P}(x) \int P(x'|x) \log \frac{P(x'|x)}{\bar{P}(x')} \, dx' dx$$

This formula quantifies the average increase in our ability to predict the future of a system when we learn its past. It appears in a number of guises and interpretations, written with joint probability distributions $P(x,x') = P(x)P(x'|x)$ it assumes the symmetric form

$$I(x'|x) = \iint P(x,x') \log \frac{P(x,x')}{\bar{P}(x)\,\bar{P}(x')} \, dx \, dx'$$

In the more general case of a communication channel described by a map $P(y|x)$, where the output symbols y need not be the same as the input symbols x, the channel rate retains its symmetry, $I(y|x) = I(x|y)$. This is not true of the conditional entropies, $H(y|x) \neq H(x|y)$ if $H(x) \neq H(y)$.

If the dynamics $P(x'|x)$ is specified in the form of a finite-dimensional transition matrix P_{ji} the stored information appears as

$$I(x_j'|x_i) = \sum_{ij} \bar{P}_i \, P_{ji} \, \log P_{ji}/\bar{P}_j$$

and terms of entropies

$$I(x'|x) = H(x') - H(x'|x)$$

The latter form makes it clear that the stored information is the <u>difference</u> in the expected randomness of a system with and without knowledge of its past. Note that both $H(x')$ and $H(x'|x)$ will diverge as we proceed to the continuum limit with a noise element present, but that their difference remains finite.

The stored information is a function of the dynamics $P(x^-|x)$ and the invariant distribution $\overline{P}(x)$ which is also determined by the dynamics. Thus the stored information is a map which takes a dynamical system, or a particular basin of a dynamical system, to a positive real number. It is, as Shannon notes, another invariant quantity, independent of the particular coordinate system used to describe the dynamical variables. It correctly goes to zero when $P(x^-|x) = P(x^-)$ meaning the future and past are statistically independent, also notice that it diverges in the case of "pure determinism". A perfectly deterministic system, if such a thing existed, could store an infinite amount of information, even if it had positive entropy, because it would be able to propagate a "point" from the past into the future. If we rewrite the relation between stored information and the entropies:

$$H(x'|x) = H(x') - I(x'|x)$$

we see how Kolmogorov was able to define a finite conditional entropy in the noise-free continuum limit. Both terms on the right diverge as the partition is taken to be indefinitely fine, but their difference $H(x^-|x)$ remains finite. However, the measured information storage capacity $I(x^-|x)$ of any real system will be finite.

12. Examples from the data

At last we are able to return to the dripping faucet, and attempt some data analysis. Figs. 26 and 27 display two sets of data which represent, to coarse resolution, periodic drops. The upper right-hand panels show the data on a greatly expanded scale. We can see that the drops are in fact not strictly periodic, there is what appears to be stochastic noise admixed. It might be of interest to know whether there are correlations in the fluctuations from drop to drop or whether the noise is "purely stochastic". To determine this quantitatively we compute the drop-to-drop stored information.
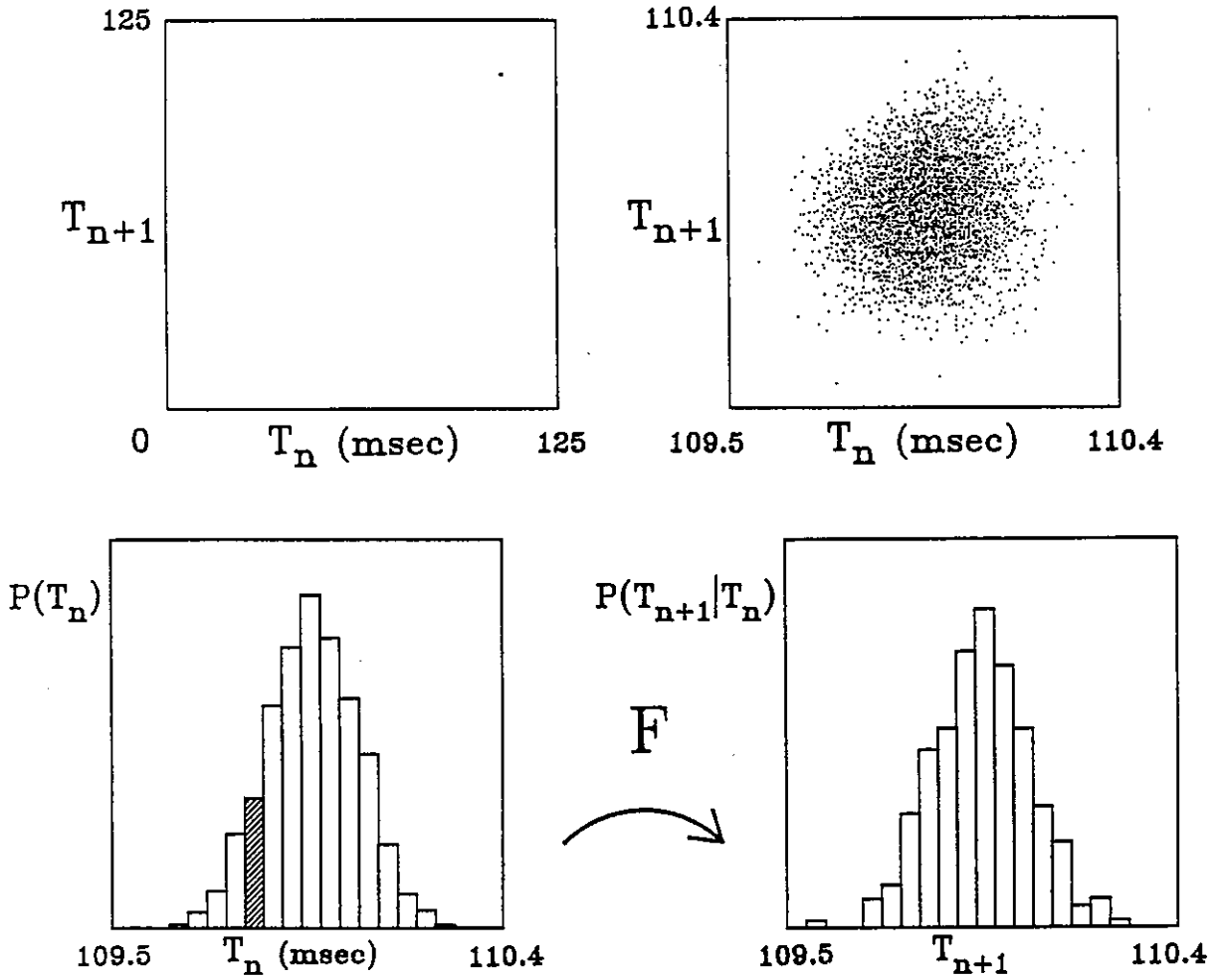
Fig. 26 – "Fixed point" data. The lower right panel displays
the forward image of the shaded histogram bin of
the lower left panel. In this case, the stored information
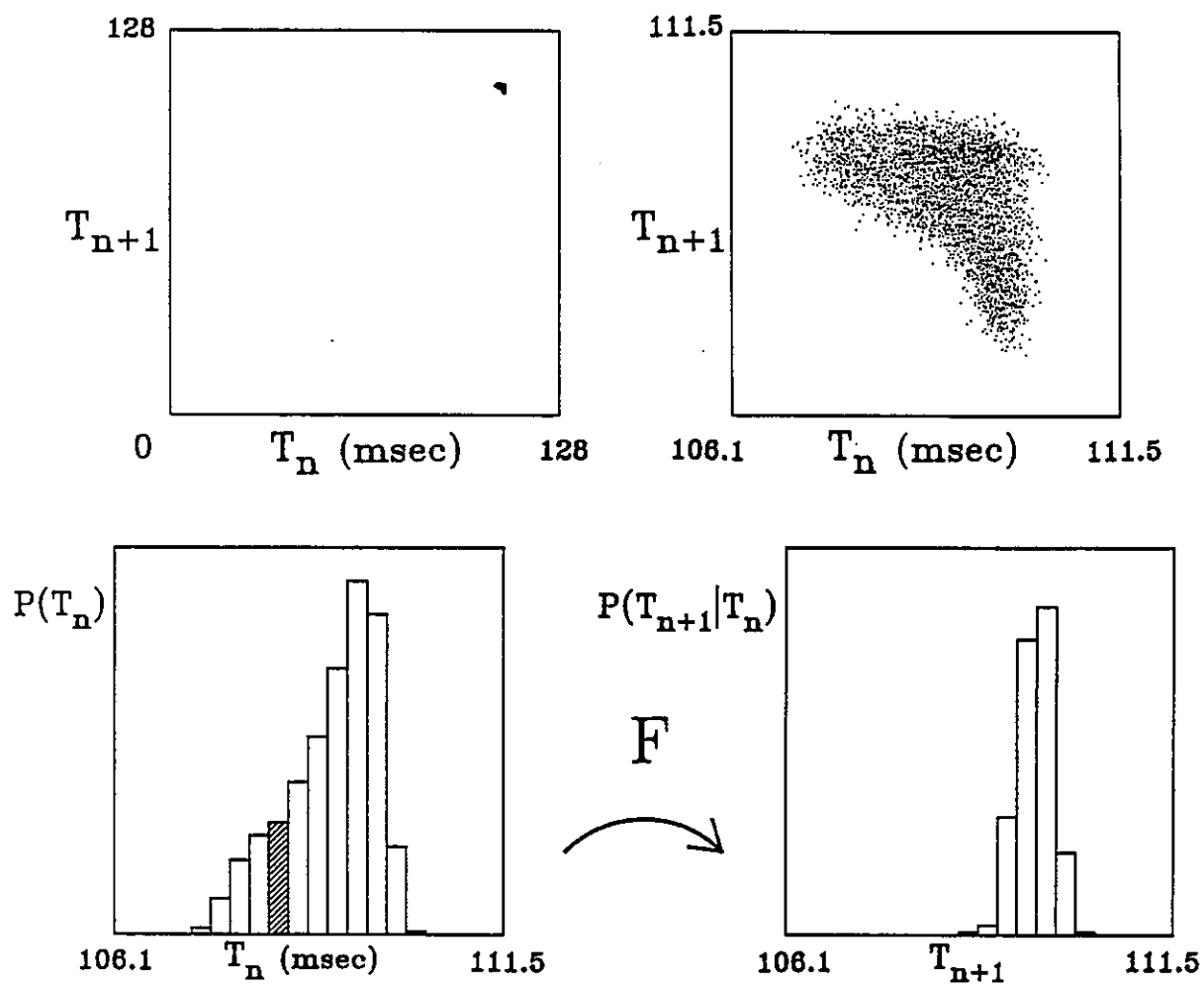
$$I(T_{n+1}|T_n) \approx 0$$

Fig. 27 – Same display as Fig. 26, for a "fixed point" with internal structure. Here

$$I(T_{n+1}|T_n) \approx .3 \text{ bit.}$$

An approximate model of the dynamics of the observed variable T_n is constructed by accumulating the transition matrix $P(T_{n+1}|T_n)$ as a histogram. The observed variation in the drop interval is split into fifteen bins, so each pair of intervals (T_n, T_{n+1}) will increment one of about 225 matrix entries. The only other quantity required to compute the stored information is the equilibrium probability density $\overline{P}(T)$ which is also accumulated as a histogram.

The first "periodic" regime of Fig. 26 is seen to store virtually zero information, fluctuations in the drop intervals are indeed "purely stochastic". The forward image of one bin in the histogram reproduces the shape of the entire equilibrium distribution, and this is true for every bin. The second string of data, shown in Fig. 27, stores a small but significant amount of information from drop to drop, interval variations are not completely independent. This can be seen qualitatively by the pattern of points in the time vs. time map, the system might be said to be in a "period two" regime obscured by noise.

A third example, illustrated in Fig. 28, is supplied by the system in a clear period-two regime. Here long and short drop intervals strictly alternate, but an observer who had just started to look at the system would not be able to tell which would be next. Thus, knowledge of the system past is worth one bit, allowing him to correctly predict whether a long or short drop interval will be next. A check for further structure in the pattern can be performed by accumulating histograms as above, the results are negative. The system stores only a single bit, the rest of the variation is "stochastic".

However, this system differs from the one described in Fig. 27 above in that the information stored perseveres <u>indefinitely</u>, allowing predictions to be made about times arbitrarily far into the future. The period-two drop regime is thus not ergodic, a localized probability distribution does not relax to the distribution $\overline{P}(T)$ representing minimum knowledge of the system state. A single observation determines the long-short sequence for all time.
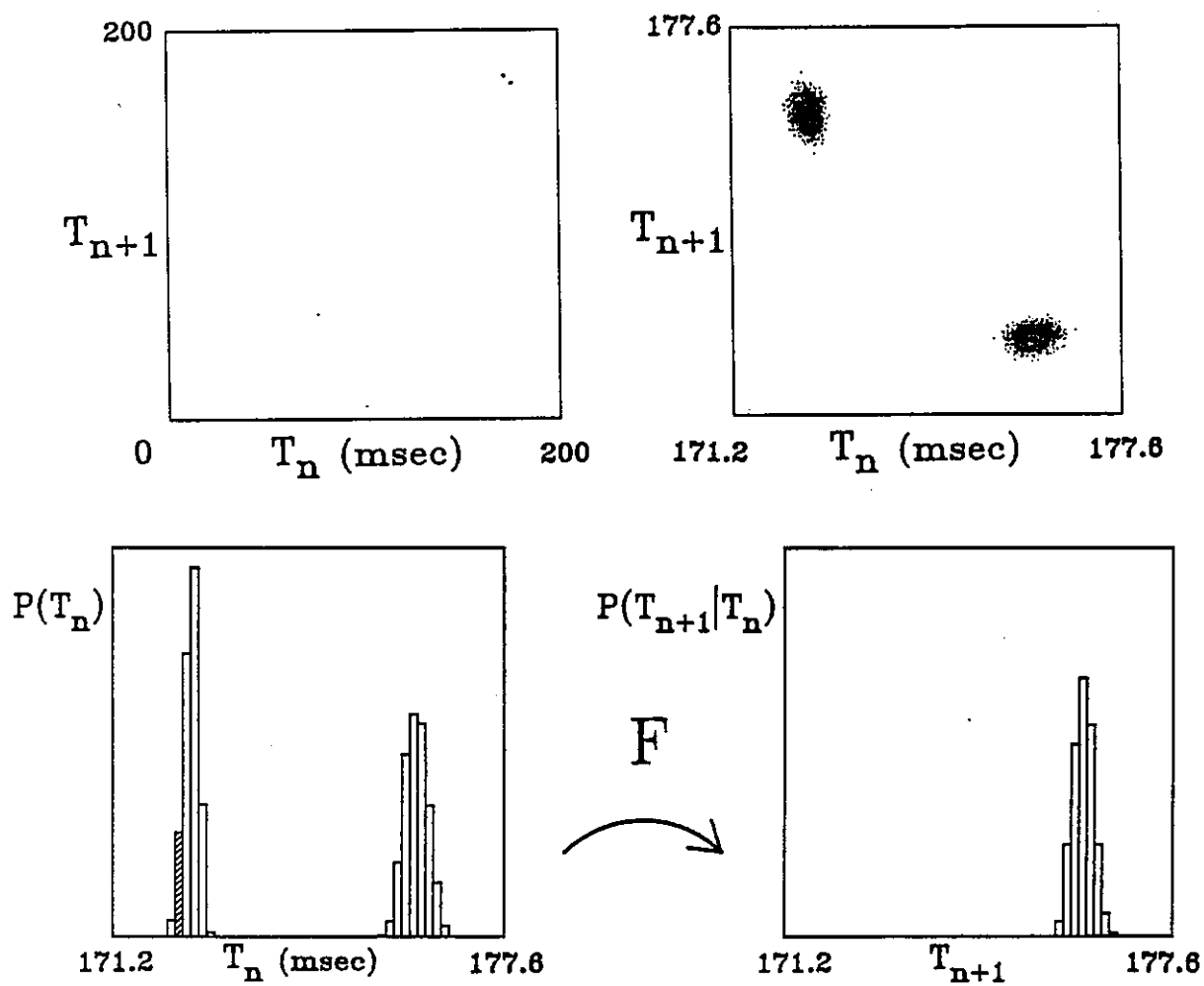
Fig. 28 — "Period two" orbit. Here

$$I(T_{n+1}|T_n) \approx 1 \text{ bit.}$$

A careful reader may have objected to the statement that the nearly periodic example of Fig. 26 contains no stored information. What about the phase? The system is a rather good clock, with scatter in the drop intervals of well under 1%. An early observer could clearly communicate to a later one via the system by adjusting the phase degree of freedom, which the later could read off by comparison with a reference clock.

This situation is an artifact of our description of the dripping faucet with a <u>single</u> state variable, the drop interval, in <u>discrete</u> <u>time</u>, parameterized by the drop number. This description correctly describes the predictive value of the knowledge of one drop interval to the determination of the next, but neglects the imbedding of the system in continuous time. Within the discrete description the "purely deterministic" and "purely stochastic" examples of Fig. 2a,b are identical in that they store zero information.

A complete description of the state of the dripping faucet at an "instant" of time requires (at least) <u>two</u> state variables, the current drop interval T, and a <u>phase</u> variable t, which could be, for example, the time since the preceding drop. States of knowledge will be represented by two-dimensional distributions $P(T,t)$.

Computations of the stored information are still performed with the same formula, though, using the higher dimensional distributions. The conditional distributions involving both T and t can, in this case, be computed from distributions involving only T, the details are relegated to an appendix. The situation, though more complicated, is straightforward. So, for clarity of presentation, much of the discussion of the next sections will consider only the predictability of one drop interval from the previous, that is, a one-dimensional system in discrete time.

The total stored information of the nearly-periodic example in continuous time can be estimated simply by considering the ratio of the scatter in drop intervals to the average interval length. The clock can be "set" to about this accuracy, and the informational value of this knowledge is about

$$I \approx \log_2\left(\frac{110 \text{ ms}}{.5 \text{ ms}}\right) \approx 8 \text{ bits.}$$

This number is more in line with estimates of the "signal-to-noise ratio" one makes from, say, the widths of lines in Fig. 10a. The dripping faucet is a very "clean" system, with a high degree of determinism. The total stored information varies considerably and not monotonically, however, with the flow rate. This reflects presumably changing relationships between the few measured variables and the background continuum. Studies of these interesting questions remain for the future.

13. Rate of loss of stored information

We have been describing the time evolution of the variables of an autonomous system by a <u>map</u> which takes distributions over the variables at time t_1 to time t_2. We have seen that, under a few reasonable assumptions, associated with each such map is an <u>invariant</u> which can be identified physically as the average amount of information stored in the dynamical variables. This number measures the degree to which an observer at t_1 can predict events at t_2.

Predictions farther in the future than t_2, without intervening observation, can be studied by reapplying the same map to the distributions at t_2, yielding the distributions at time t_3. Associated with this compound map is the invariant information stored between t_1 and t_3. Further iteration of the map produces a sequence of numbers I_1, I_2, I_3, ... describing the ability to predict farther and farther into the future. We expect that often our ability to say anything

about the future system state from an observation at t_1 will decline as we con-
sider the more distant future, and that, if the map is ergodic, the sequence I_n
will converge to zero, as all distributions eventually map to a single minimum
information distribution $\overline{P}(x)$. The effect of having made a particular observa-
tion at t_1 is lost.

Now we will examine properties of the <u>rate</u> of loss of the stored informa-
tion, that is, the difference between succeeding terms in the series
I_1, I_2, ... This quantity $h_n = I_n - I_{n+1}$ will be seen to correspond to the
"entropy" of a deterministic map, although there are difficulties in passing to
the deterministic limit.

The expected general form of the curve $I(t)$ describing the stored informa-
tion remaining at time t from a measurement at t = 0 is schematically illus-
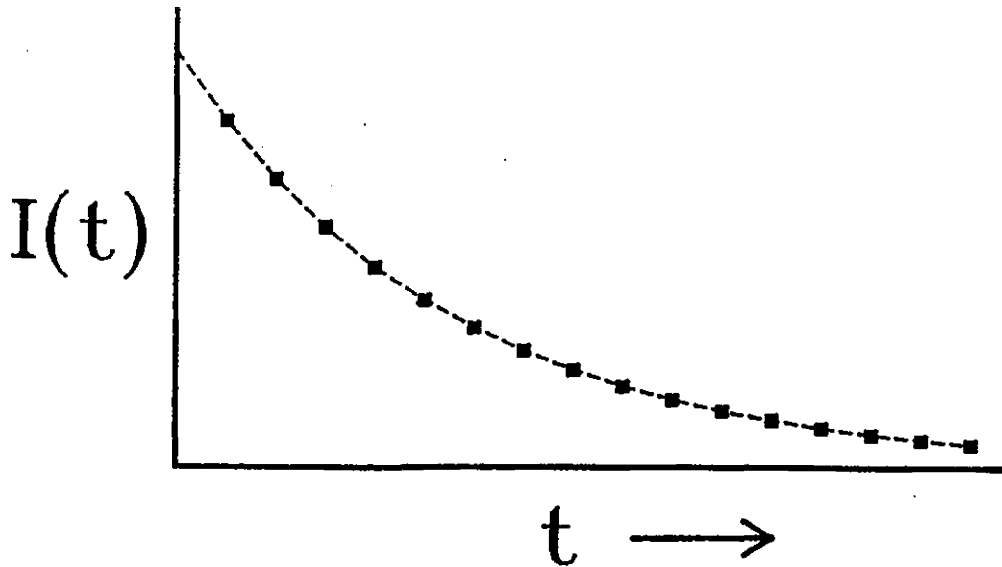trated in Fig. 29.

Fig. 29 - Small squares represent possible measurements.

Two properties of this curve can be shown from "convexity" properties of logarithmic measures of information:

a)   I(t) monotonically decreases, $I_{n+1} < I_n$. This property is intuitively obvious, our information about a system cannot increase in the absence of observation.

b)   The curve I(t) is concave, $h_{n+1} < h_n$. This is also intuitively clear, sharper distributions representing greater stored information will spread faster than broad distributions.

The maximum entropy, or loss of predictability, will thus occur when the stored information is at a maximum. In the case of a discrete map, taking the variables over an interval Δt, a reasonable definition of the entropy is then the loss of information between the first and second iterations of the map. Under this definition both the stored information and entropy of a "purely stochastic" map will be zero.

It is tempting to extend this definition to the continuous case by considering the slope of the curve I(t) at t = 0. But we should remember that an "instant" of time is just as unphysical a concept as a "point" in space. As the intervals between measurements become shorter, the rate of information obtained by observation from the system is increased. In general we can expect that the rate of loss of predictability may depend on the measurement process, the more "severe" the measurement, the greater the perturbation of the original state. Thus the limit process necessary to define a slope cannot be performed. A more complete discussion awaits a quantum mechanical treatment.

14. Example from the data

Figure 30 displays the time vs. time map for some water drop data in a "fuzzy hump" regime, together with an approximation to the minimum knowledge distribution $\bar{P}(T)$ constructed by distributing 4000 drop intervals into 50 histogram bins. A 50 x 50 transition matrix $P(T_{n+1},T_n)$ is likewise assembled. Fig. 31 shows what happens when we place all the "probability fluid" initially in the bin whose location is marked with the arrow in panel (a), and then apply the matrix $P(T_{n+1},T_n)$ several times.

After the first iteration, the initial "delta function" assumes a width characteristic of the stochastic noise level of the data. We can now compute the stored information associated with the map. Its value for this particular initial bin is 2.3 bits, and the average over all bins, 1.8 bits, approximates the invariant I_1. Under only a few applications of the matrix, nearly all the stored information is lost, and the equilibrium distribution closely approaches the experimentally observed distribution of Fig. 30b. The entropy of this map is the maximum loss of stored information per iteration, which occurs between iterations one and two, about .83 bits per drop.

To summarize, we collect our experience with the water drop data as a transition matrix, which becomes our model for its dynamics. This model of course gives good predictions for additional data, if the system is autonomous. The invariant stored information associated with this matrix and its iterates is then computed, this measures the predictability of the system.

As discussed earlier, this computation of the stored information considers only the predictability of the length of one drop interval given knowledge of the length of the preceding interval, without considering that the actual state space is much larger, as the intervals are imbedded in continuous time.
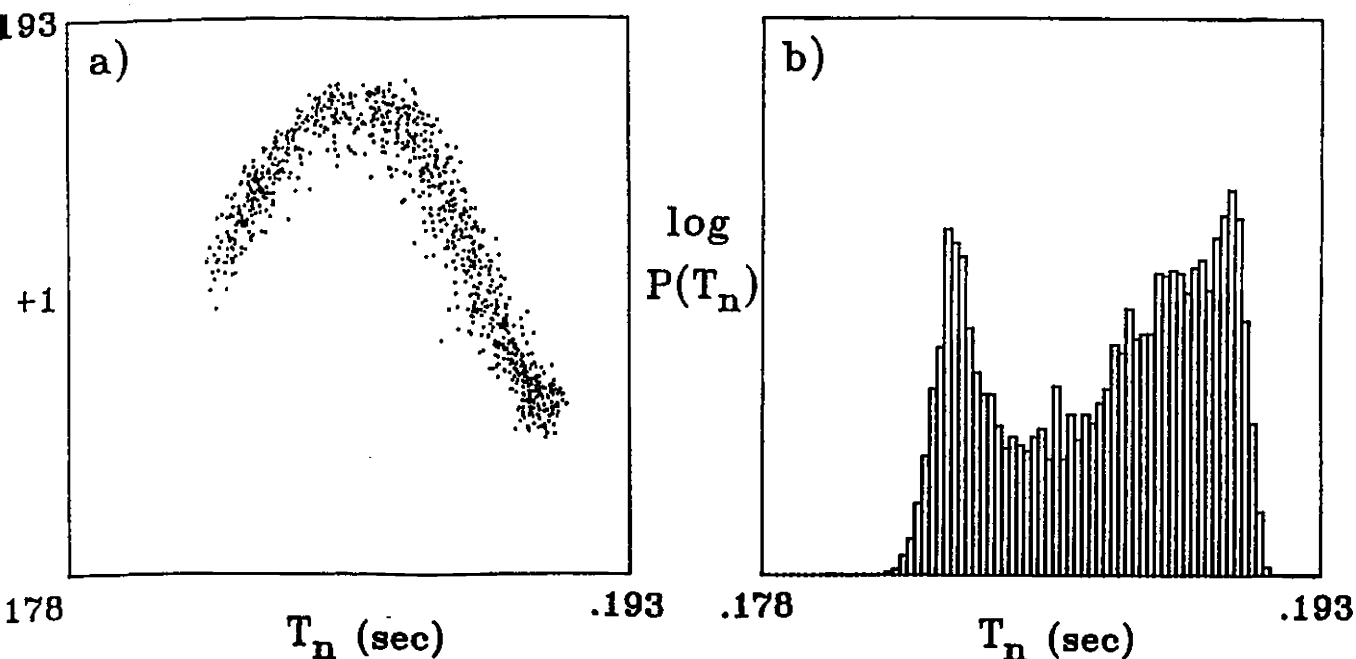
Fig. 30 – Typical "fuzzy hump" data set and corresponding invariant distribution.
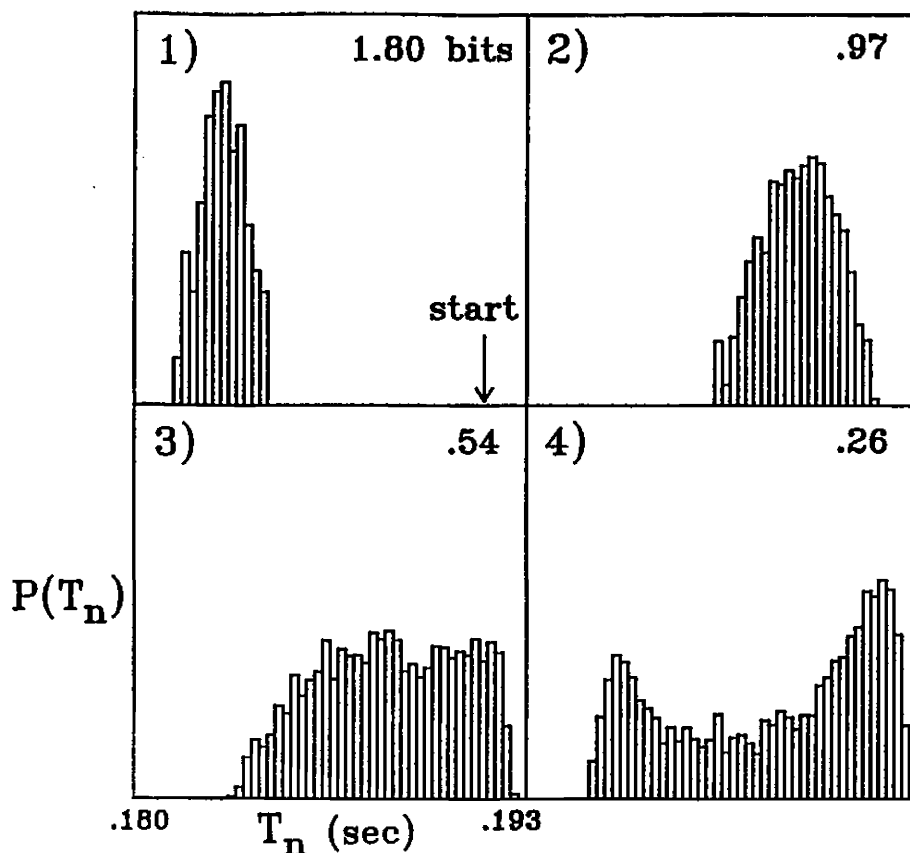

Fig. 31 – Evolution of the image of a single bin through four iterations.
Average stored information remaining after each iteration
is listed in the top right corner of the panels.

When the total stored information is computed, as described in the appendix, the number obtained is much larger, some 9.7 bits. As Fig. 32 illustrates, the first few bits are lost rapidly, but the remaining phase information persists longer.
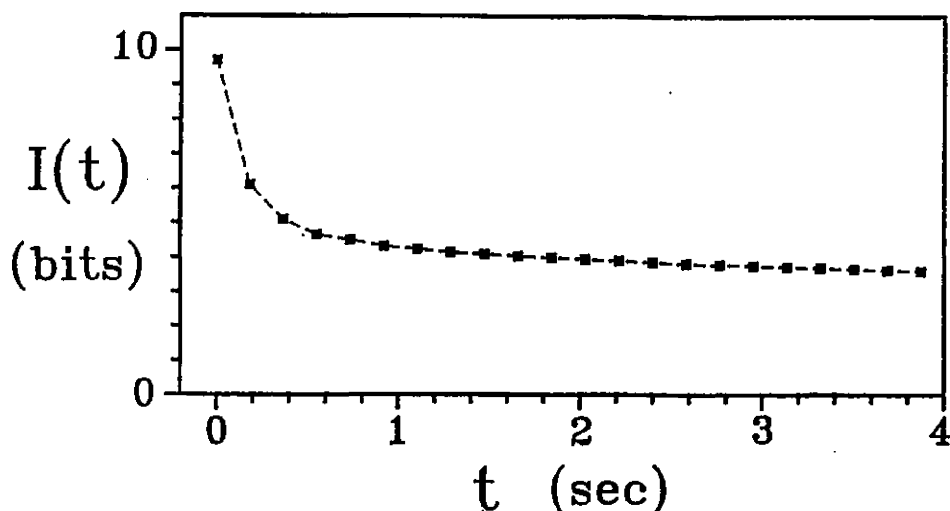
Fig. 32 – Stored information as a function of continuous time
for the data set of Fig. 30a. Computation was performed
for multiples of the average period, as indicated by
the small squares.

In this connection Fig. 33 is instructive. Plotted is the probability of observing a drop fall at some "instant" as a function of time. In the absence of any observation, this distribution is flat, as indicated by the dotted line. But now let us suppose that we know that a drop has fallen at t = 0. Our probability of a drop falling at time t, in the absence of any further knowledge, is shown, for the data of Figs. 30-32. Because the variation of drop intervals is only about 5% of the average interval, considerable phase information is preserved. Successive distributions spread roughly only linearly, as would be expected from repeated convolutions of a narrow gaussian distribution, and the stored information declines at a slow logarithmic rate[28].

Fig. 33 — (top) Time vs. time plot for a data set whose variation
is a small fraction of the average period.
(center) Relative probability of observing a drop
given that one has occurred at t = 0, for drops 1 − 3.
(bottom) Same, for drops 52 − 54.

This statistical effect, resulting in the long-term preservation of information in a chaotic system, has been termed "phase coherence", for more discussion, see refs. [10,29]. For comparison, consider the same plot for data at a much higher flow rate, where interval variations are comparable to the average interval. Here the knowledge that there was a drop at t = 0 is of predictive value for only a short time into the future.
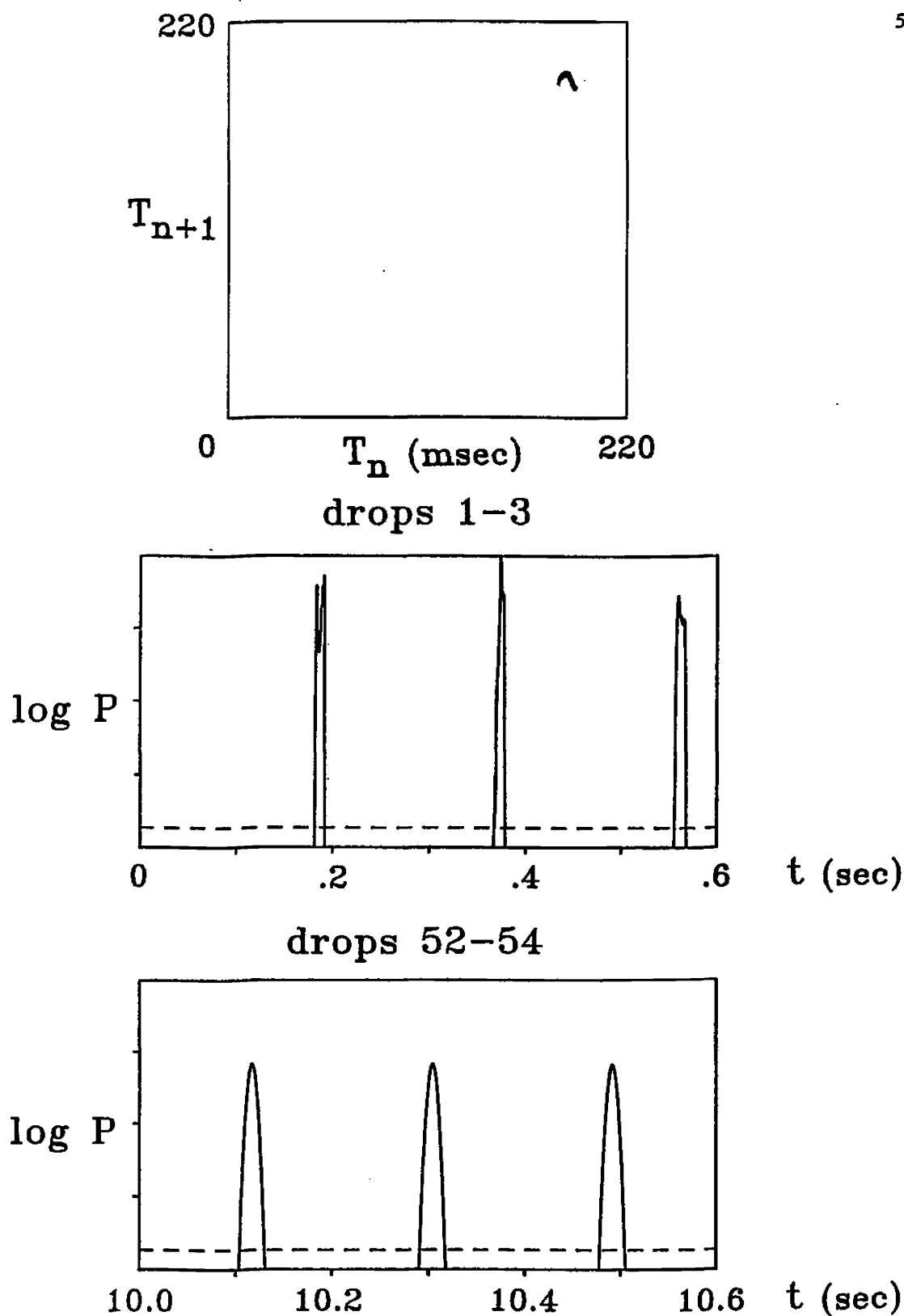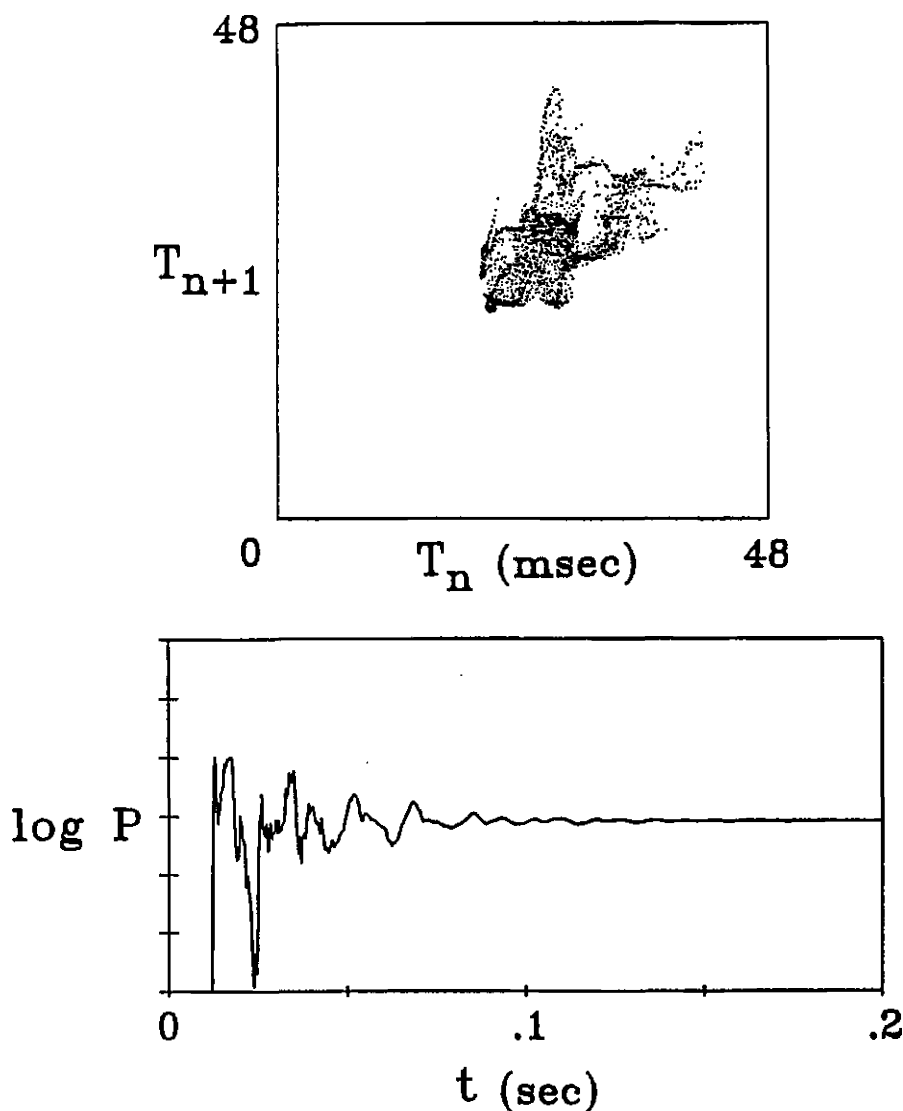

Fig. 34 — (top) Time vs. time map for a data set with variations a large fraction of the average period.
(bottom) Relative probability of observing a drop at time t given that one occurred at t = 0.

15. The entropy in the limit of "pure determinism"

In section 9, the entropy of a purely deterministic system was discussed in terms of a limit process involving the partitioning of the domain into infinitesimal elements. In the preceding section, the entropy of a map with a noise component was defined as a maximum rate of loss of stored information. Here it will be shown that the singular noise-free case <u>cannot</u> in general be approached smoothly as the limit of a mapping with a stochastic element, as the stochastic element tends to zero. The property corresponding to the deterministic entropy of a mapping with a very small stochastic element will be made clear.

The example we will study is the parabolic map with noise added, $x' = rx(1-x) + \zeta$, which can generate a string of numbers resembling faucet data, as illustrated in Fig. 4. Again, ζ is a random variable of some small width, modeling, for example, thermal noise.

Figure 35 shows the invariant distribution $\overline{P}(x)$ of this map at $r = 3.7$ for $\zeta = 0$, the no-noise case, and for $\zeta = .1\%$.

Fig. 35 – Invariant distribution of the parabolic map
with and without added noise.

Here the "noise" uncertainty is modeled for convenience by a simple truncation of the accuracy of the computation. (The experience of Crutchfield and Packard and others has been that the 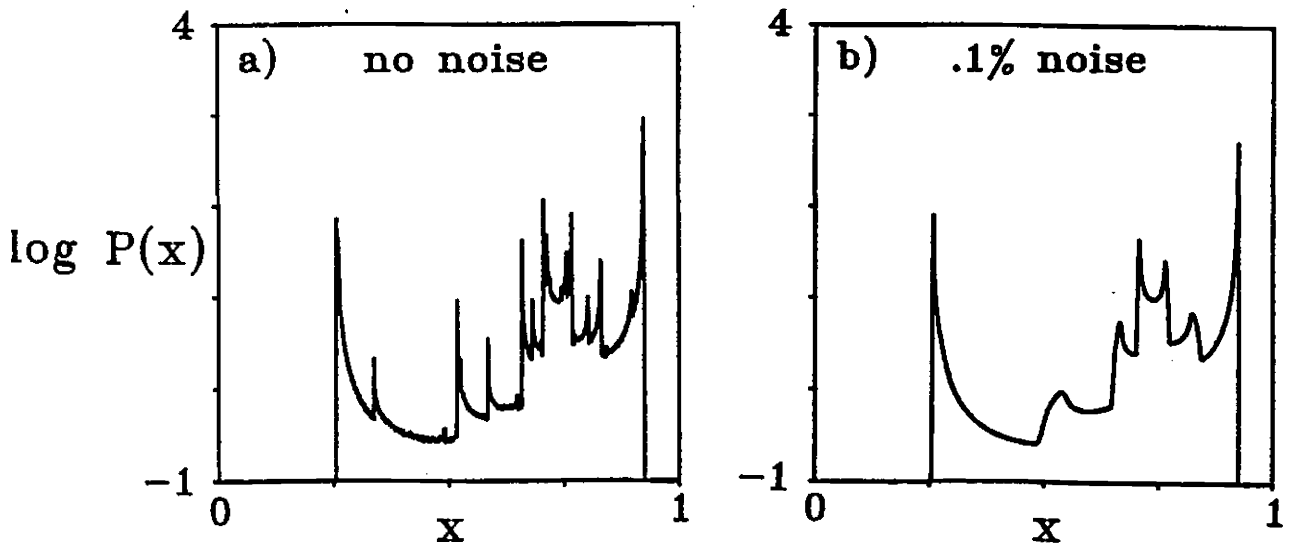outcome of numerical experiments such as this is usually unaffected whether one takes ζ to be of gaussian, uniform, or other form[27]. Only the effective width of ζ is important.) Note that this is a rather small noise level, much smaller than that illustrated in Fig 4b. The noise would be invisible on the scale of that figure, although finer features of the $\overline{P}(x)$ distribution do become smeared out.

In the deterministic case, $\zeta = 0$, the information storage capacity of the map diverges, as mentioned earlier. But it becomes finite with the addition of noise, and the series I_1, I_2, ... of the information stored by higher iterates of the map has the concave behavior discussed in section 13. This series is graphed for three small noise levels in Fig. 36. As expected, the initial value of the stored information becomes higher as the noise is reduced.

The rate of loss of stored information, $h_n = I_n - I_{n+1}$, is graphed for the three cases in Fig. 37. The rate of loss starts off in each case at the same value, then falls off to a plateau. The length of the plateau becomes longer as the noise level decreases. The value of the rate of loss of information in the plateau region is near the value of the entropy of the map in the no-noise case, which is indicated by the dotted horizontal line.

The fact that the entropy of a map with even a low noise level is initially higher than the corresponding deterministic map is due to partition effects, and to regions of the mapping where intervals are contracted, that is, where the slope is less than one. In a purely deterministic description, a small interval representing an uncertainty can be shrunk and then expanded back to the same level without any further loss of information, as schematically illustrated in Fig. 38a. But if there is noise present in the mapping, no interval can be
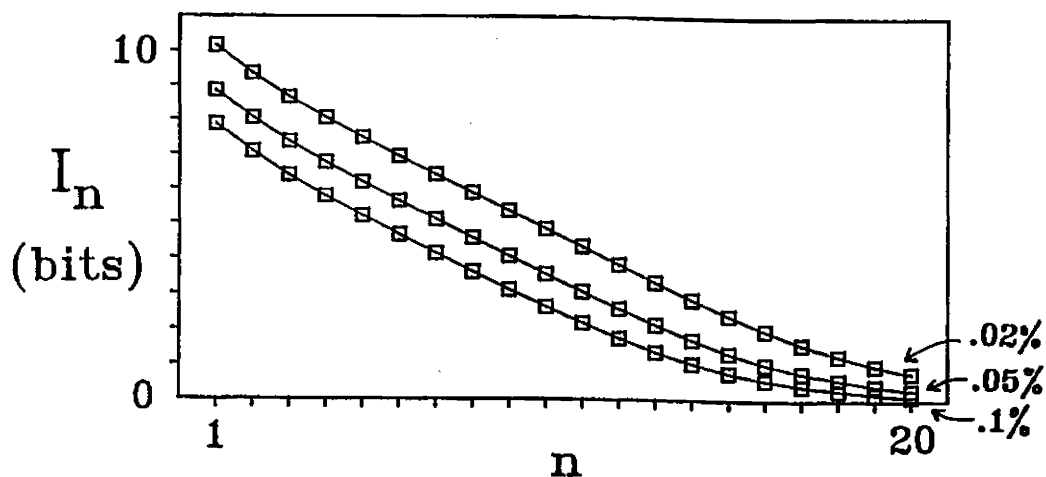
Fig. 36 – Stored information for iterates of the parabolic map
at three different noise levels.
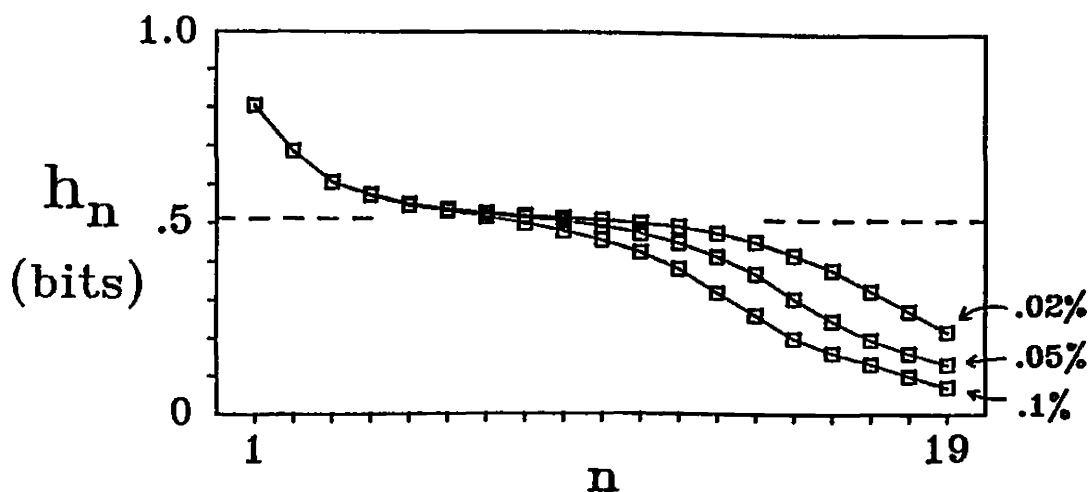
Fig. 37 – "Entropy", or rate of change of stored information.
Dotted line indicates metric entropy of map without noise.

Fig. 38a

Fig. 38b
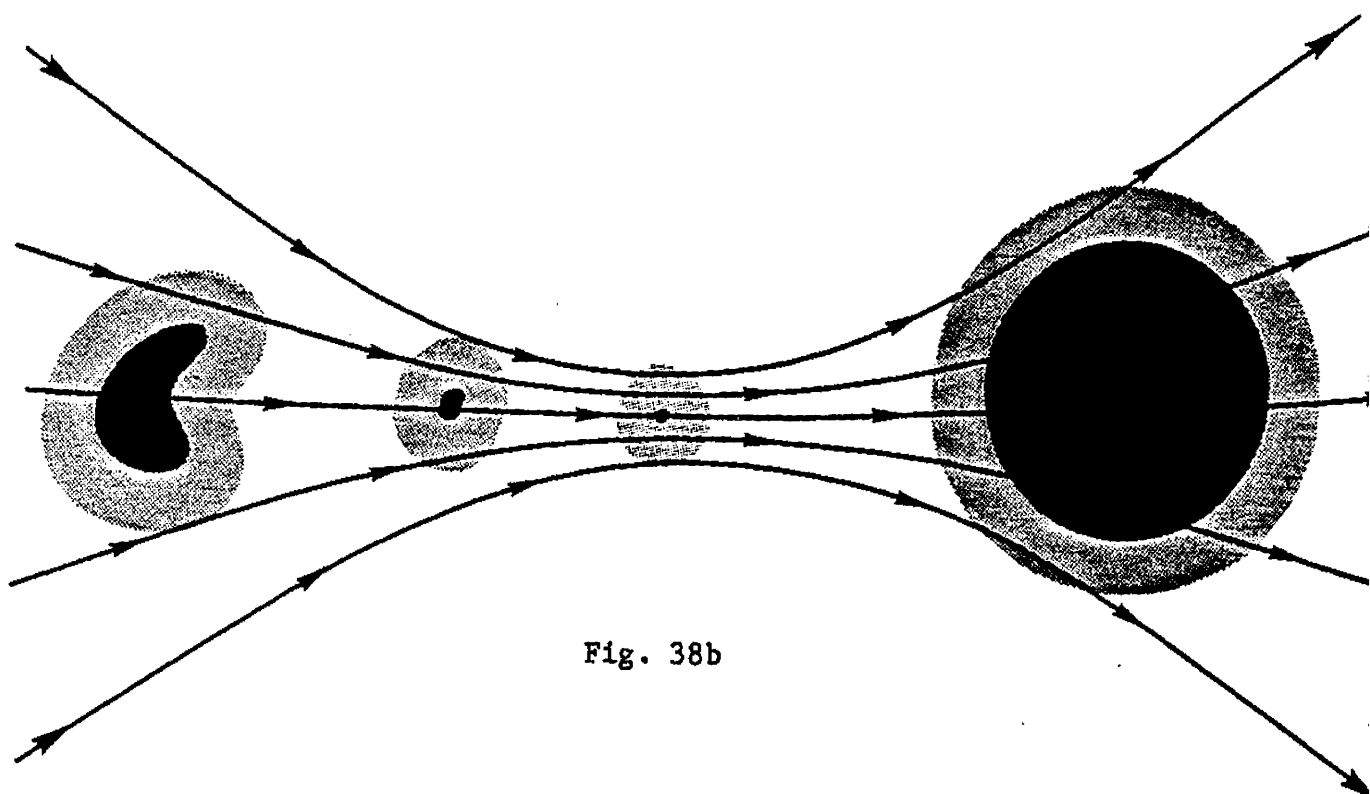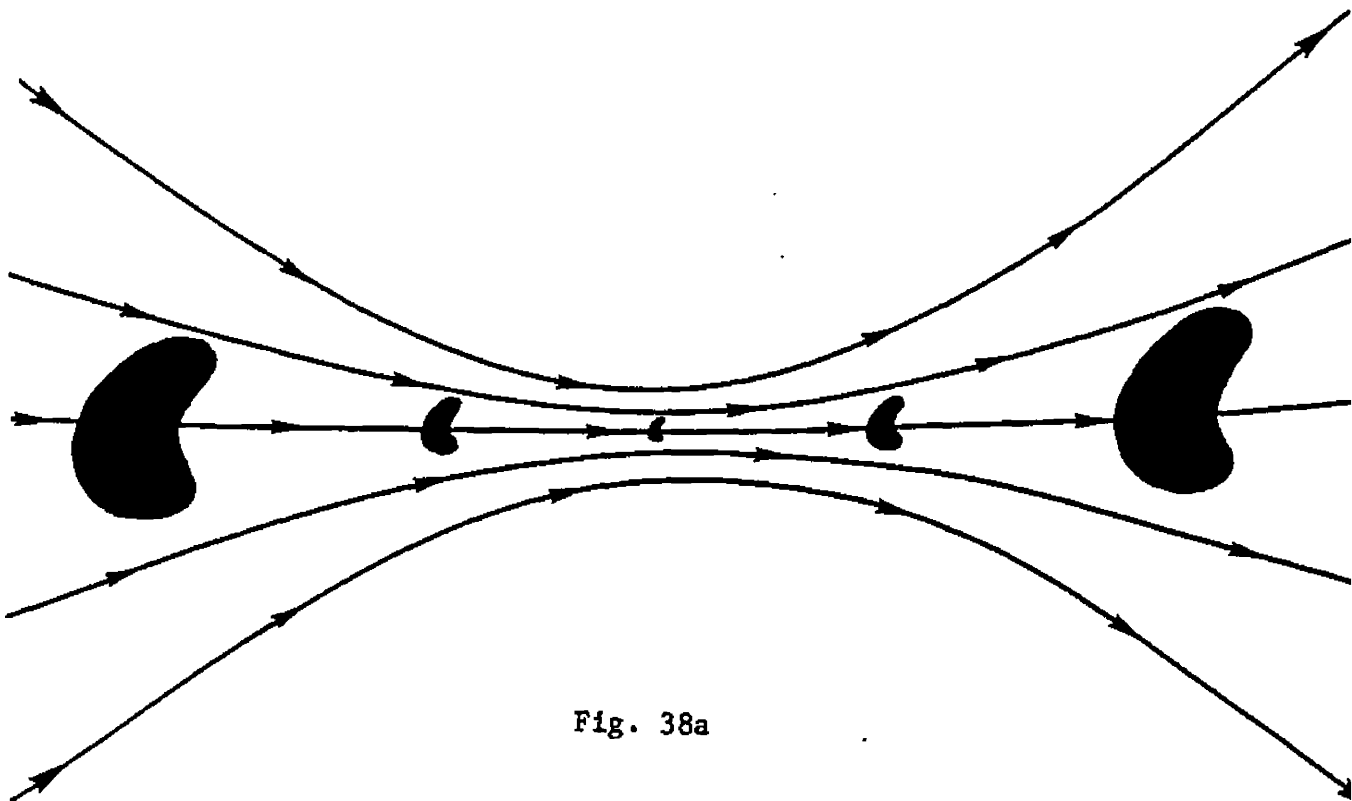
(schematic) - A "reversible" flow can have irreversible effects
        if features are contracted to near the base noise length scale.

contracted to below the noise level. An attempt by the deterministic part of the dynamics to contract and then reexpand a small uncertainty interval will result in a larger uncertainty, and loss of information, as indicated in Fig. 38b.

As the noise level is decreased, the stored information will increase without bound, but the initial part of the entropy curve will remain the same, even at arbitrarily low noise levels. Thus the entropy of the deterministic approximation to a physical system will be relevant only to observations conducted at intermediate length scales, that is, coarse enough not to resolve the noise level, but not so coarse as to smear out large-scale structures. Although more careful studies might be needed to clarify the effect of paritioning on the entropy calculation, it seems that the initial values of the entropy can be estimated from the deterministic map simply by ignoring the contracting parts of the map.

In summary, this definition of entropy, in terms of a rate of loss of stored information, does not suffer the difficulties which the standard definition in terms of partitions finds in the presence of noise. However, the two definitions are equivalent only in a particular intermediate set of length scales.

16. "Observational noise"

In the discussion up to this point, we have assumed that the variables x and x´, representing the results of measurements of a system before and after a time interval Δt, are the "true" variables, that is, knowledge of these variables will give the best possible account of the predictability of the system. We have relied on the coordinate invariance of the various measures of information to justify the use of an arbitrary measurement scheme, we hope that a "good enough" measurement of <u>any</u> set of dynamical variables will equivalently characterize the predictability.

More generally, though, we might expect that a measurement procedure often will <u>not</u> reveal the complete state of the system. The measuring instrument may suffer from a lack of resolution, or there may be "noise" in the measuring instrument independent of any noise in the system itself. Or perhaps the system possesses more "degrees of freedom" than the measurement apparatus, and ambiguities are introduced in the projection onto the measured variables. A more general situation is sketched below:

"true" variables

measured variables

$P(x'|x)$

$P(x|y)$

$P(y'|x')$

$P(y'|y)$

x   x´

y   y´
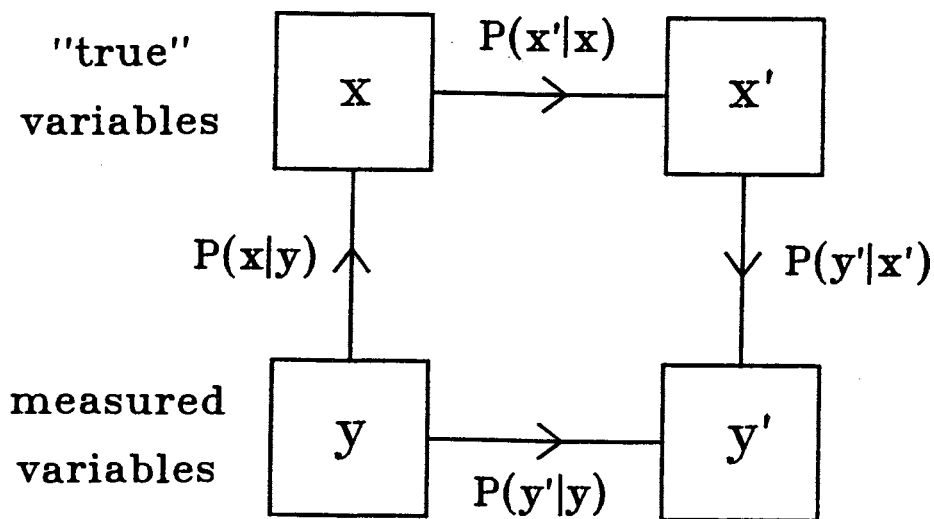
Fig. 39

The "true" before-and-after variables x and x⁻ and their causal connection P(x⁻|x) are not observed directly. Instead, the states y and y⁻ of the <u>measuring instrument</u> are recorded, and only the causal relation P(y⁻|y) between these variables is directly accessible.

The properties of the measuring instrument are specified by conditional distributions connecting the "true" and observed variables. The vertical data paths of Fig. 39 illustrate two interpretations of the effect of a measurement. The distribution P(x|y) describes a state of knowledge of the "true" variables P(x) given an observed distribution P(y). This is a key step, P(x|y) is a "best guess", which, acting on $\overline{P}(y)$, yields $\overline{P}(x)$. The inverse of this conditional distribution, P(y⁻|x⁻), can be interpreted as the projection onto the observed variables P(y⁻) of the "true" distribution P(x⁻).

The process whereby the complete system dynamics P(x⁻|x) generates the observed causal connection P(y⁻|y) might be described as follows: An observed distribution P(y) constrains the possible values of the "true" variables x to a distribution P(x). This distribution is acted upon by the "true" dynamics P(x⁻|x), yielding the distribution P(x⁻), which can be interpreted as a state of knowledge of the "true" variables after a time Δt, given an earlier measurement of the accessible variables y. Finally, this state of knowledge is projected onto the observable variables by a second measurement, yielding the distribution P(y⁻).

If the "true" system dynamics P(x⁻|x) happen to be available, and the properties of the measuring instrument are known, the resulting observed dynamics can easily be computed by composition of the conditional distributions:

$$P(y'|y) = \iint P(y'|x') P(x'|x) P(x|y) \, dx \, dx'$$

This expresses the equivalence of the two paths between y and y⁻ in Fig. 39.

This description of the measurement process might be given by an omniscient observer, possessing a total overview, including the true dynamics $P(x'|x)$. However, the usual case is that we know only the observed dynamics $P(y'|y)$, and perhaps something about the properties of the measuring instrument. The static image of Fig. 39 cannot describe the mysterious process whereby we generate successively better models of a presumably unique true dynamics from limited sense data. Yet hopefully communication theory can be useful, providing both a framework for describing imperfect measurements, and a formalism for quantifying the "degree of imperfection". The communication metaphor suggests a few basic concepts which may survive the future development of our understanding of the modeling process.

a) Measuring instrument as channel:

Objective information about the state of a system is obtained via measurement. The measuring instrument is thus a _channel_ between the system and the observer, and if its properties can be represented by conditional distributions $P(x|y)$ and $P(y'|x')$, the channel can transmit information at a particular _rate_, given by $I(x|y)$ and $I(y'|x')$. This rate places limitations on what can be known about the system.

We are assuming that both the system and the measurement process are time-independent, and that the same instrument is used in both the before and after measurements. Thus there will be identical equilibrium distributions $\overline{P}(y)$ and $\overline{P}(y')$ defined on the measured variables, and the transmission rate both to and from the primed variables will be the same:

$$I(x|y) = I(y|x) = I(y'|x') = I(x'|y')$$

The domain of the measured and "true" variables may well be different, and the measurement may not be being conducted in an optimum fashion, thus this number

is a channel rate, and not a channel capacity. Its units are "bits per observation".

If the measuring instrument is "perfect", there will be a one-to-one mapping between the measured and "true" variables, and clearly the distributions $P(x^-|x)$ and $P(y^-|y)$ will have the same stored information and entropy. Here the rate of transmission through the channel is infinite, if the variables are considered continuous. For any physical measurement, though, the rate is finite, which implies that the stored information associated with the observed model is necessarily less than that of the true dynamic:

$$I(y^-|y) < I(x^-|x)$$

The observed mapping $P(y^-|y)$ is in this picture a <u>composite</u> of the maps associated with the "true" dynamic and the measurements, and it can be shown mathematically (exercise) that the rate associated with any composite mapping is less than or at best equal to the capacity or rate of the <u>smallest</u> of its parts, equality holding only in the singular case where everything else is one-to-one. This is clear physically, the information flow through a series of devices is restricted by the slowest among them.

The measurement instrument acts as a "filter" between the observer and the system, and both the stored information and the entropy associated with the observed dynamic $P(y^-|y)$ will be <u>less</u> <u>than</u> <u>or</u> <u>equal</u> <u>to</u> the stored information and entropy of the "true" dynamic $P(x^-|x)$, for predictions any distance into the future. Thus both the absolute value and the slope of the curve I(t) will be less for the observed variables, as sketched below:
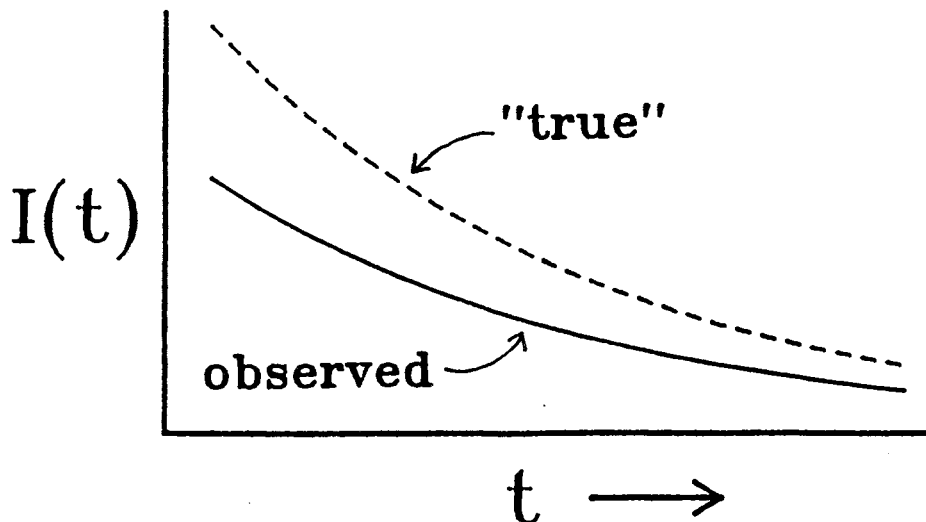
Fig. 40 – Stored information and entropy reduced
by passage through measurement channel.

Although in general the true dynamics $P(x'|x)$ cannot be fully known, usually we have a greater or lesser confidence in the measurements. For example, the water drop dynamics takes place on a time scale of tenths of seconds, whereas measurements are performed with a crystal clock, with a resolution of a few microseconds. Thus we expect that, if the dynamics is indeed low dimensional, we will be able to construct a good model of the true dynamics. This intuitive feeling can be numerically justified, for the "fuzzy hump" data of Fig. 30, by comparing the information the system is capable of storing, about 10 bits, with the measurement channel rate. The latter can be estimated by considering that the timer can resolve the true drop time to 4 microseconds out of an a priori expectation that is roughly flat over the average drop interval, about .1 second. Thus about 14 bits are available for measurement. The fact that the measurement precision is much greater than the measured system storage capacity provides a good basis for the belief that the system is well characterized by the measuring instrument, but whether this is actually the case or not is an undecidable question. This point will be discussed later.

b) Observation rate:

A single observation of a physical system will yield a certain amount of information about the state of the system, the amount is as we have seen quantifiable in bits and limited by the resolution of the measuring instrument. By considering a series of observations, repeated at a given rate, we can define an average "observation rate", the product of the average bits per observation and observations per second. The units of this quantity are thus "bits per second", the same as the entropy.

One task of an experimenter studying a new system is to characterize the determinism of the system, which can be described as the information stored through time in the dynamical variables. If the system has a positive entropy, some or all of this causal connection will be lost in the passage of time. If succeeding measurements are too far apart, they will not be able to resolve the determinism, even if they are of high precision, this situation is sketched below:
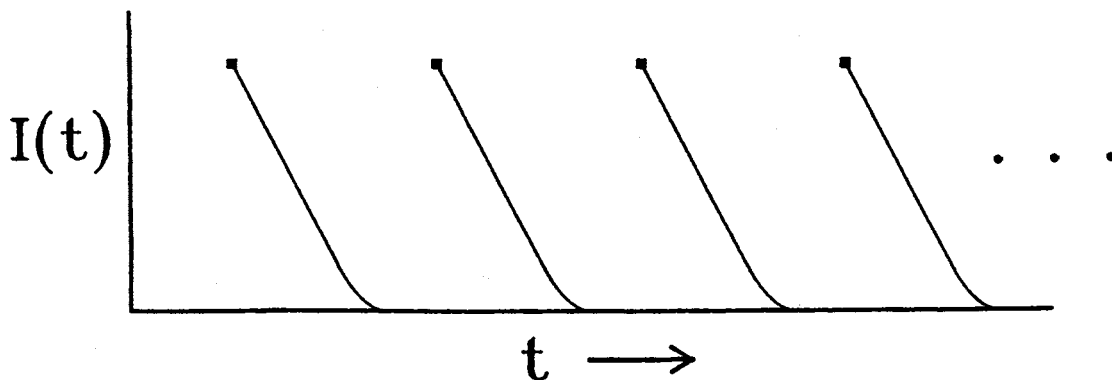

Fig. 41 - Causally disconnected measurements.

Here the observer is only taking random samples out of the minimum information distribution $\overline{P}(x)$. In order to resolve a given degree of determinism, or stored

information, the experimenter must <u>observe</u> at a <u>rate</u> greater than the entropy at that level of stored information. If we assume that the average rate of loss of information is roughly independent of the amount of information stored, as in the figure, then a given degree of determinism can be resolved with either high precision measurements at a low rate, or coarse measurements at a high rate[30].

c) Complete and incomplete measurements:

An autonomous system can be described as a dynamical rule acting on some space of states $\{X\}$, producing a sequence of states X_n, X_{n+1}, X_{n+2}, ... The dynamical rule which we have been writing $P(x^-|x)$ thus will also appear below as $P(X_{n+1}|X_n)$.

If a system is completely specified by its "state" X_n, then there will be no increase in predictive ability upon learning its history X_{n-1}, X_{n-2}, etc. Thus we can define a "complete measurement" as one which allows the maximum possible predictability with a single measurement. With a "complete" set of variables, the dynamics becomes a first-order Markov process, the evolution of the system so described is independent of its history, that is, $P(X_{n+1}|X_n) = P(X_{n+1}|X_n,X_{n-1},...)$. This is like a "nearest-neighbor interaction" between successive observations X_n, X_{n+1}.

The "true" variables are by definition "complete", but this property can be lost in an imperfect projection down to the measured variables. Some unmeasured aspect of the system may persevere through time to be recorded by a later observation. Then a better prediction can be made by considering the past history of measurements, we have, in the language of communication theory, the more complicated situation of a "channel with memory".

There are at least two general ways in which this situation can arise. First, the "true" dynamic might be capable of storing considerably more

information than the measurement channel can transmit. Examples of this case

will be presented in the next section. Second, the dimensionality of the

"true" dynamic might be higher than that of the measured variables, or that the

two sets of variables are not mapped in a one-to-one fashion. This case is

exemplified by Fig. 6, even simple functions can become multivalued. Note that

in the second case, the measured variables cannot be made complete by just

increasing the accuracy of the measurements, an actual change in dimensionality

or mapping of the observation channel is required. A complete set of variables

can of course be constructed by assembling a larger dimensional variable con-

sisting of the present measurement plus as many past measurements that increase

predictability. The process of reconstructing a higher dimensional state space

from lower dimensional observations has been termed "Geometry from a time

series" by Packard et al [31].

17. Symbolic Dynamics

The result of a physical observation is typically recorded in the form of a

number of finite length. Thus, from a measurement point of view, an experiment

can have only a finite number of outcomes. Each outcome can be labeled by a

symbol, and the recorded time behavior of the system then consists of transi-

tions between these symbols. Measurement, in principle, reduces a continous

system to a discrete one. If the underlying geometry of the dynamics is simple,

a rather coarse measurement with a small number of symbols may extract a consid-

erable portion of the determinism of the system. The description of a continu-

ous system by a small number of symbols is often referred to in the mathematics

literature as "symbolic dynamics" [32].

Repeated coarse measurements of a chaotic dynamical system typically will

produce a partially correlated string of numbers. We will begin by discussing

the predictability and entropy of such a data stream. Then we will consider the

case where a measurement can have one of only <u>two</u> outcomes. This provides a particularly simple arena to examine, in a preliminary way, a few rather general questions. Given a data stream, how much can we infer about the system that produced it? What sort of measurements should we make to optimize the predictability of a system? Do we use the same criteria to best determine the past states of a system? This section builds on the recent work of Crutchfield and Packard [33].

First a bit of notation, a string of symbols generated by repeated measurement of some system will be written s_1, s_2, s_3, ... and a <u>history</u> of n such symbols s_1, s_2, ... , s_n will be denoted S^n. Now because the measurement process has produced a discrete set of symbols, we can directly apply the definition of entropy, for a single symbol we have

$$H(s) = -\sum_{s} P(s) \log P(s)$$

where the sum is taken over the possible values each symbol s can take on. The entropy of a string of length n is given by

$$H(S^n) = -\sum_{S^n} P(S^n) \log P(S^n)$$

where the sum is taken over the much larger set of all possible patterns of symbols of length n. Already we can see that practical problems will arise in computing directly the entropy of long strings of symbols.

If successive symbols are totally uncorrelated, as are for example the outcomes of a repeated coin toss, then the entropy of a string of n events is simply n times the entropy or "surprise" of the single event:

$$H(S^n) = nH(s)$$

But if the events are correlated, the "surprise" of later events is <u>reduced</u> by

expectations build up by the earlier history. The classic example, due to Shannon, is written language. The longer a string of text, the easier it is to guess the next lette.

The expected general form of the total entropy as a function of the number of events or symbols is illustrated in Fig. 42 below. The total entropy increases monotonically (we never get "unsurprised"), but the rate slows as expectations are built up, leading to the convex curve shown.
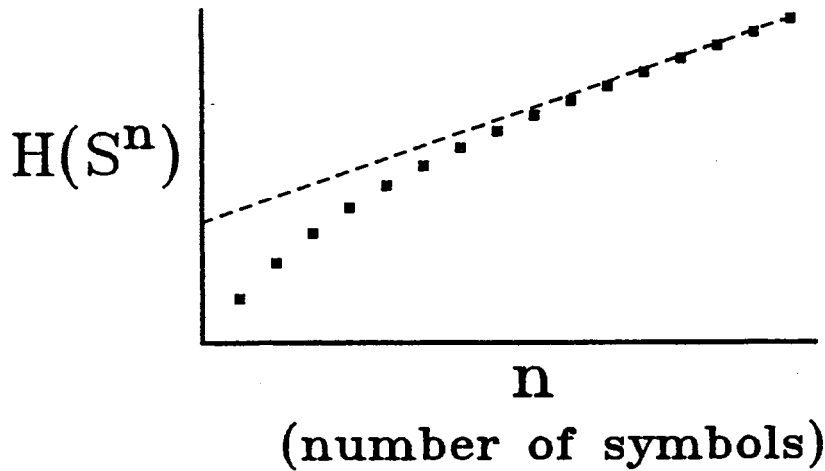
Fig. 42 – Total entropy of a string of symbols as a function of its length. The asymptotic slope of this curve is the entropy per symbol, and the y intercept of this asymptote, indicated by the dotted line, is the stored information.

In a real chaotic system with noise present, initial data has predictive value for only a limited time, events which are too far apart are causally disconnected, as argued in Fig. 41. Correlations extend over only a finite length of the string of symbols, old experience ceases to have predictive value. The effect of this on the graph of Fig. 42 is that the curve approaches a constant slope, that is

$$H(S^{n+1}) - H(S^n) = H(s|S^n) = const. \cong h$$

for large enough n. Shannon refers to the asymptotic value h as the "entropy per symbol" of a discrete signal source.

Now how does one compute the stored information associated with this string of symbols? We know that this number will be finite, if correlations extend only over finite lengths of the string. We also know that, if the curve of Fig. 42 is initially convex, the measurements which produced the symbol string were "incomplete", in the sense discussed in the preceding section. Predictability of the next symbol is increased, up to a point, by considering longer and longer past symbol histories.

A reasonable way to approach this computation is to consider the total predictability of the future symbol string given the past symbol history, $I(s_{n+1}, s_{n+2}, \ldots | s_n, s_{n-1}, \ldots)$, or $I(S^n | S^m)$ if we indicate past and future strings by S^m and S^n. We can now argue that this number will be independent of m and n, if both are larger than the correlation length. A few manipulations of the definitions of entropies and stored information given earlier yield:

$$I(S^n | S^m) = H(S^m) + H(S^n) - H(S^{m+n})$$

This can be seen to be the y intercept of the tangent to the flat part of the curve of Fig. 42, if m and n are large enough. This nice geometric picture tells us, for example, that

$$I = H(S^n) - nh$$

for large n, where h is the asymptotic entropy $H(S^{n+1}) - H(S^n)$. It is clear as well that the y intercept of the tangent to any portion of the curve gives a lower bound to the stored information, and that a criterion for the maximum correlation length is that the y intercept no longer increase with increasing string length n.

If the resolution of the measurement process which produces the symbol stream is increased, the number of values each symbol s can assume increases, and the number of possible patterns S^n increases exponentially. As the resolution approaches the noise level, the asymptotic slope of Fig. 42 will increase without bound. But the y intercept indicated will remain bounded, presumably approaching the stored information of the underlying dynamical system. Please note here the distinction between the entropy of a particular number stream obtained by measurement, and the entropy of a map or system, defined in section 13 to be the rate of decrease of stored information or predictability as future times are considered. The latter "entropy" remains bounded as resolution is increased.

Now we will take a look at the effect of coarse measurements on the familiar parabolic one-dimensional map. This example, the writer hopes, will provide a useful and not overly technical illustration of the ideas suggested above.

The single dynamical variable in a one-dimensional map is a value along the real line between (usually) zero and one. A simple way to model a coarse measurement of this variable is to divide the interval in two, and to report as the measured variable only whether the value of the "true" dynamical variable falls on the left or right half of the interval. This measurement scheme is illustrated for the map $X_{n+1}=3.7X_n(1-X_n)$ in Fig. 43a, where the dividing line is placed at X = .5. Fig. 43b shows the minimum information distribution $\overline{P}(X)$ for this map.
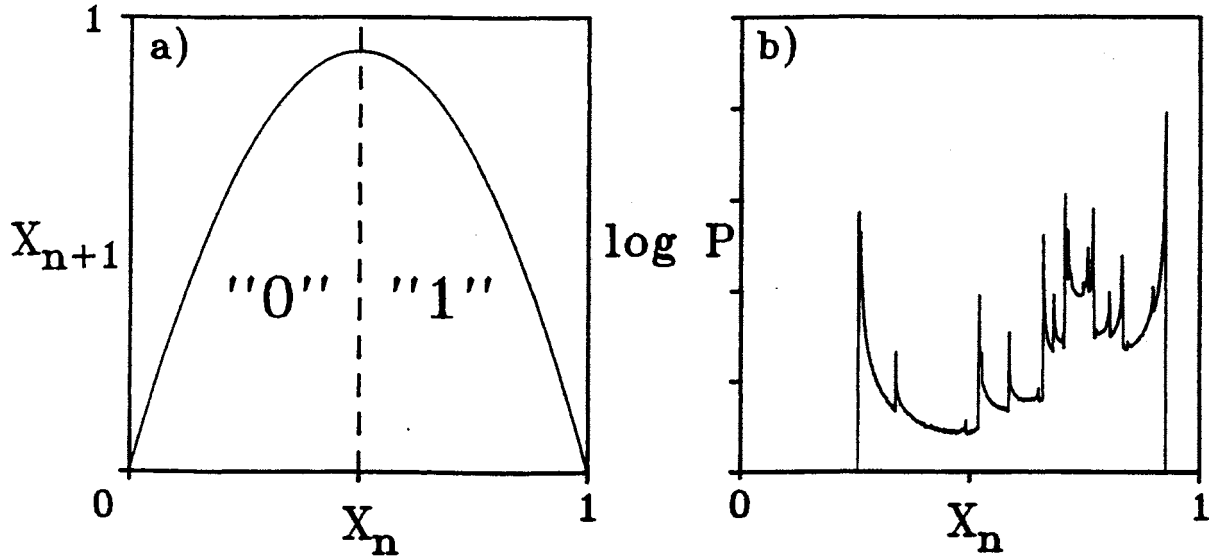
Fig. 43 - Graph and invariant distribution for the map $X_{n+1} = 3.7X_n(1 - X_n)$.

The results of a sequence of repeated measurements of the system are no longer a string of high-precision real numbers between zero and one, but rather a string of the two symbols "0" and "1", labeling the left and right side of the interval. If we denote the string of symbols so generated as s_1, s_2, s_3, ... then the transition matrix of measured quantities $P(y'|y)$ can equivalently be written $P(s_{n+1}|s_n)$.

The information stored in the continuous variable over an iteration of the map is arbitrarily large, as discussed in section 11. But under the dynamic induced by the single symbol measurement, clearly no more than one bit can be stored over a single iteration, as only one bit is being observed. We will now, in Fig. 44, trace the information stored around the schematic data paths of Fig. 39, where $P(x'|x)$ is the "purely deterministic" dynamics defined by the one-dimensional map, and $P(y'|y)$ is the dynamic induced by the coarse measurement.

Fig. 44 – The symbol y = "1" is assumed to have been observed,
corresponding to the coarse measurement .5 < x < 1,
illustrated are the changing probability distributions
as we pass around the data paths of Fig. 39.

As mentioned above, $I(x^{\prime}|x)$ is infinite, as are also $H(x)$ and $H(x^{\prime})$, but $I(x|y)$ and $I(y^{\prime}|x^{\prime})$ can be no more than one bit, the maximum capacity of the measurement channel. $I(x|y)$ is in fact so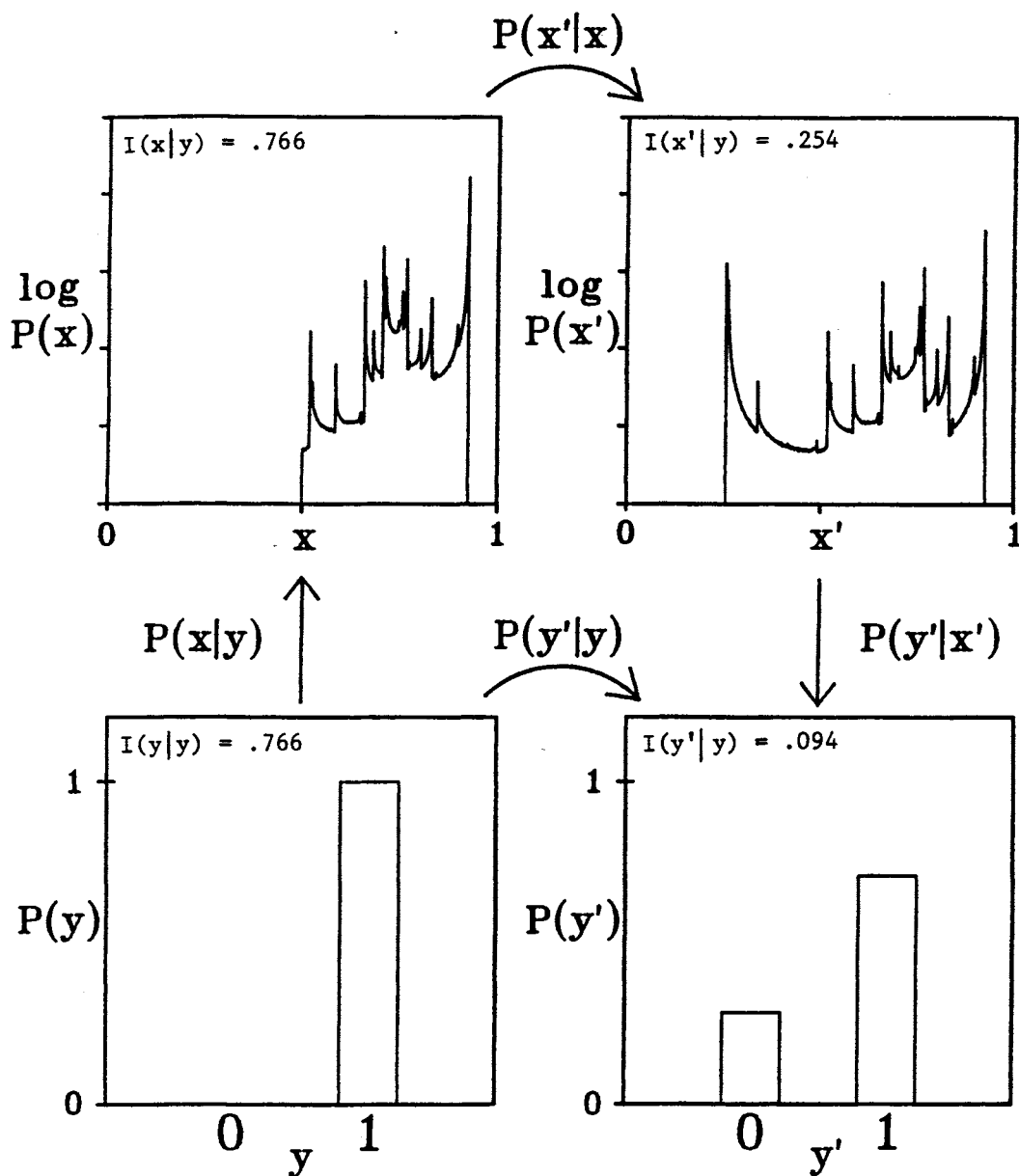mewhat less, as the partition at $X = .5$ does not divide $\overline{P}(x)$ evenly, the iterates of the map fall on the right about 78% of the time. Thus the symbol "1" occurs more often, this constitutes prior knowledge which reduce the average new information obtained in each single measurement to .766 bit. A "best guess" of the state of knowledge of the "true" variable x given that the symbol "1" has been observed is illustrated in the top-left panel of Fig. 44, it is simply the right side of the minimum distribution $\overline{P}(x)$. Given that the channel is noise-free, the knowledge we gain about the system is simply equal to the entropy or "surprise" of the measurement, that is,

$$I(x|y) = H(y)$$

This is the informational value of the inference we make about the "true" system given the limited measurement. This state of knowledge $P(x)$ is now iterated under the "true" dynamics $P(x^{\prime}|x)$, yielding the distribution $P(x^{\prime})$ illustrated in the top-right panel of Fig. 44. This distribution represents our knowledge of the variable after an iteration of the map, given that we knew only that the variable was to the right of .5 in the interval before the iteration. The information stored to this point (averaged over both possible initial measured symbols "0" and "1") has been reduced to .254 bit. The difference between $I(x|y)$ and $I(x^{\backprime}|y)$ is, for this partition position, $H(x^{\prime}|x)$, the entropy of the deterministic map, as computed for example in Ref. [22]. Finally, the distribution $P(x^{\prime})$ is projected down onto the measured variables "0" and "1", as shown in the lower-right panel, further reducing the stored information to .094 bit. Now we have completed our trip around the diagram, and this last number is the stored information associated with the induced dynamic between symbols, $P(y^{\prime}|y)$.

Thus by transforming to and from the "true" variables, we can compute the transition matrix between the symbols "0" and "1", and the stored information associated with this matrix. However, in writing the induced dynamic $P(y'|y)$ as a two-by-two matrix $P(s_{n+1}|s_n)$ we assume that the observationally accessible state of the system is <u>completely</u> <u>specified</u> by a <u>single</u> measurement of the primed variables. That is, the single symbol "0" or "1" describes all we can know about the system. But in this case we can make a better prediction as to the next measured symbol than is possible with just the transition matrix by considering the <u>history</u> of observed symbols, not just the single prior symbol. A single symbol measurement is thus not "complete", that is,

$$P(s_{n+1}|s_n, s_{n-1}, \ldots) \neq P(s_{n+1}|s_n).$$

The stored information associated with the single-symbol matrix $P(y'|y)$, $I(y'|y) = .094$ bit can be improved upon by constructing higher-order transition matrices, considering the previous two, three, or more measured symbols to predict the next. The more of the past which is considered, the better the prediction of the next symbol; the stored information as a function of the number of past symbols considered is graphed below in Fig. 45 for the map of Fig. 43.
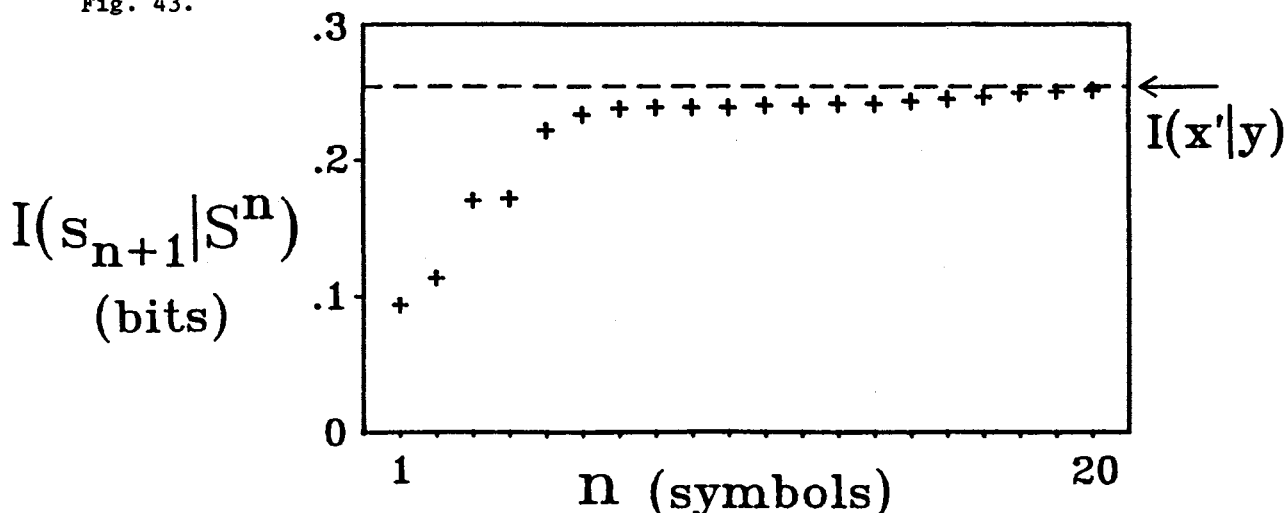

Fig. 45 - Predictability of the next single symbol given a history of length n.

This curve is simply related to the <u>slope</u> of the total entropy curve through the relation:

$$I(s|S^n) = H(s) - H(s|S^n)$$

Note the extremely slow convergence of this curve, even after we take a history of twenty symbols, there is more predictability to be gained by looking back to the twenty-first. This is because our model is noise-free, with correlations extending over indefinite lengths. The tangent to the asymptotic curve becomes, at least, difficult to compute. As we will see in the next section, real data behaves more reasonably.

An upper limit to the predictability of s_{n+1} obtainable with <u>any</u> amount of knowledge of the past is supplied by the value $I(y'|x)$, or $I(s_{n+1}|X_n)$. This is the predictability of the next single symbol s_{n+1} given knowledge of the complete continuous variable X_n, clearly the knowledge resulting from a set of past measured symbols can only approach this value. Knowledge of X_n reduces the next symbol s_{n+1} to a certainty, that is, $H(y'|x) = 0$, and $I(y'|x) = I(y'|x')$ because $I(x'|x)$ is infinite, and $P(y'|x')$ is a projection. However, the entropy of the deterministic map itself remains finite, here $H(x'|x) = .512$ bit.

A second and stricter limit on the predictability of a series of coarse measurements is obtained by considering the value $I(x'|y)$, or $I(X_{n+1}|s_n)$, indicated by the dotted line in Fig. 45. This is the amount of information an observer with full ability to observe the complete continuous variable gains when he learns which side of the interval the variable originated from, corresponding to the upper-right of Fig. 44. This quantity limits the increase in our ability to predict the future symbols s_{n+1}, s_{n+2}, ... given the single symbol s_n, that is, the average increase in our ability to predict the future we obtain with each new measurement.

Note that although $I(x'|y)$ and $I(y'|x)$ describe the information stored along symmetric paths in Fig. 39, they are not equal. The measurement process has introduced a <u>time</u> <u>asymmetry</u> into an autonomous system. In this case, a single measurement gives us more information about the past system state than the future. Because the system is presumed to be autonomous, any number stream derived from it by measurement must be equally predictable forward and backward in time, that is, $I(y'|y)$ must equal $I(y|y')$, and likewise for higher order combinations of symbols. But once we <u>assume</u> a <u>model</u> $P(x'|x)$, in this case explicitly noninvertable, a single measurement may allow us to infer more about the state of the <u>model</u> in one time direction.

For the case of the noise-free one-dimensional map of Fig. 43a, the complete state of the continuous variable X_n can be determined by considering the sequence of the coarsely measured variable s_n, s_{n+1}, etc. it generates. This however is a very special case, and is only possible because i) the system is "perfectly deterministic", with an infinite storage capacity, thus the present state of the variable X determines all future states of the system, and ii) the partition dividing the interval is in exactly the right place to allow a one-to-one mapping between points on the continuous variable X and symbol histories.

We will now study the effect of moving the partition from $x = .5$, changing the projection of the system onto the observed variables y and y'. Figure 46 demonstrates that the effect depends on which informational measure is being considered. $H(s)$, the simple one-symbol entropy ignoring symbol to symbol correlations, is a maximum at the partition position which makes the two possible symbol values equiprobable. $H(s|S^n)$, the "surprise" of the next symbol given a symbol history, is a maximum at $x = .5$, the peak of the map. Its value, in a noise-free model, is limited by the deterministic entropy of the underlying map. $I(s|S^n)$, the predictability of the next symbol given a symbol history, has
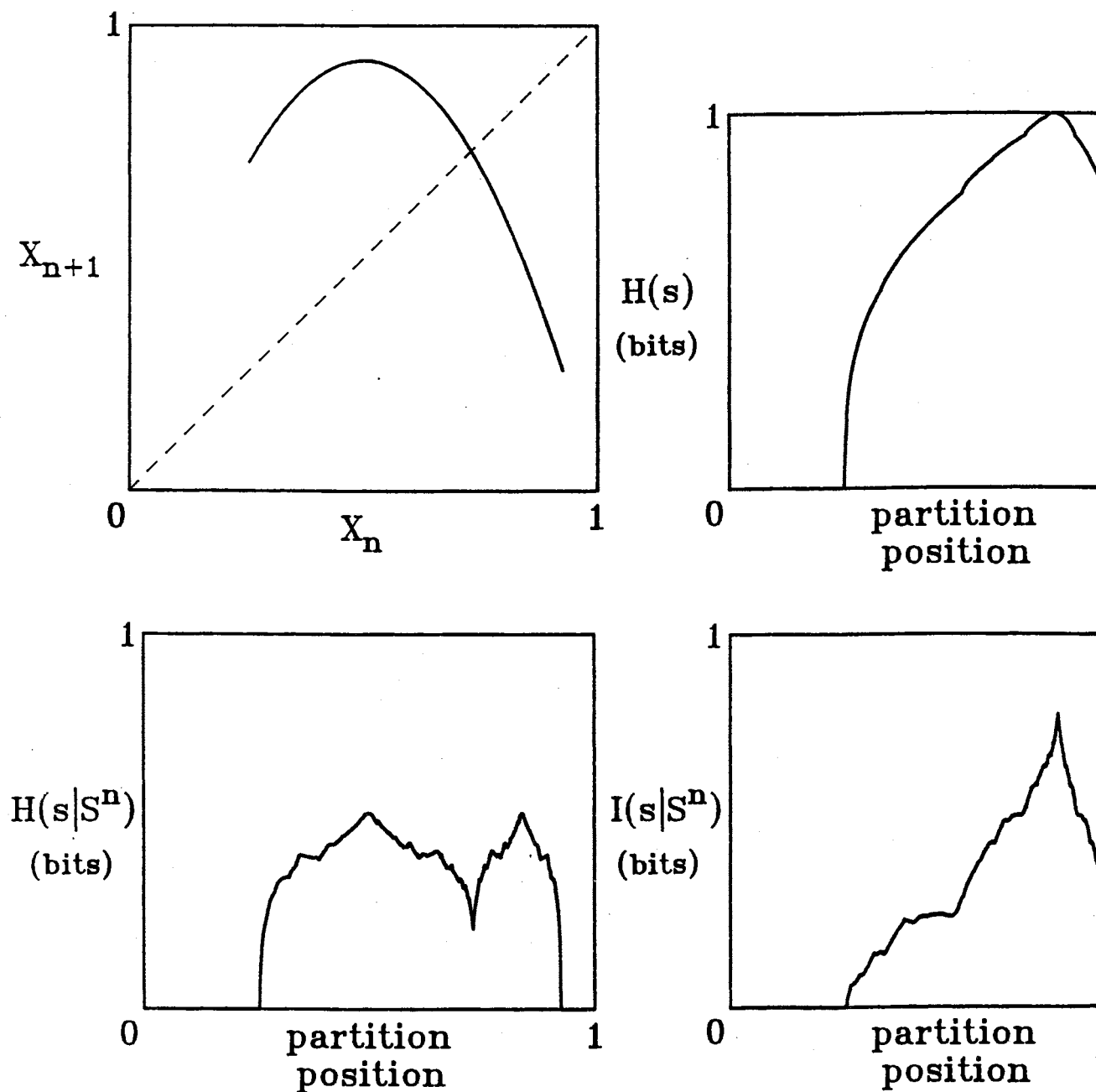
Fig. 46 - The top left panel shows the attracting part of the graph
for the map $x' = 3.7x(1 - x)$. The identity point
of the map is the point in x where the identity line crosses
the graph. In the other panels appear the single symbol
entropy, the entropy of a symbol given its history, and
the predictability of a symbol given its history, all as a
function of partition position. A history of 15 symbols
was used in these calculations.

a sharp maximum at the identity point of the map, x = x´, here simple alternating sequences like 010101 are most likely.

The lesson of Fig. 46 seems to be that, if we have to perform a measurement with constraints, the "best" measurement of a system will depend on the use to which the system is put. If we want to use the system as a recording device, and determine the past states, we may want to optimize the entropy $H(y´|y)$ coming from the system. If we want to send signals into the future, and exercise control over future measurements, we will maximize $I(y´|y)$. An extreme example of the difference between these two measures is given by the full two-onto-one parabolic map x´ = 4x(1-x), or by the binary shift map of Fig. 21, with the measurement partition at x = .5 in both cases. Here the entropy generated by the map, $H(x´|x)$, becomes equal to one bit per iteration, which is the maximum measurement channel rate $I(y´|x´)$. Thus no predictability of the future is possible, yet an initial condition in the continuous variable x can be read off to arbitrary accuracy if we wait long enough.

The considerations guiding optimal partitions in one dimension seem straightforward, but even in two dimensions the situation is unclear [34]. Also, the problem of fuzzy partition boundaries, or noise in the measurement channel, seems open. The whole question of the matching of a measurement channel to a particular system seems largely unexamined in a physics context. "Here be dragons", and a lot of interesting geometry.

18. Example from the data

The effect of a coarse two-symbol measurement on the stored information and entropy will now be illustrated, for the "canonical" data of Fig. 30. As must be clear by now, this data set was chosen as an example due to its similarity to the one-dimensional map with noise added of Fig. 4.

The stored information associated with the "true" dynamic $P(T_{n+1}|T_n)$ is 1.80 bits, as was shown in section 14. The maximum capacity of the observation channel is only one bit, leaving us to suspect that the coarse measurement is not "complete", that is, the quantity $I(y^-|y)$ can be improved on by considering the past history of the primed variables.

Taking a cue from Fig. 43a, we define the two symbols of the coarse measurement by placing the partition at the peak of the distribution. Fig. 47 below shows that there is indeed a slight convexity in the total entropy curve as the number of past symbols is increased. Here, though, the curve reaches an asymptotic slope after only a few past symbols are taken into account. This is understandable in view of Fig. 31, the stored information associated with the "true" dynamic $P(x^-|x)$ falls to near zero after only 4 iterations. There is no point in looking farther back into the past than this, the system state at that point is causally disconnected from the present state.
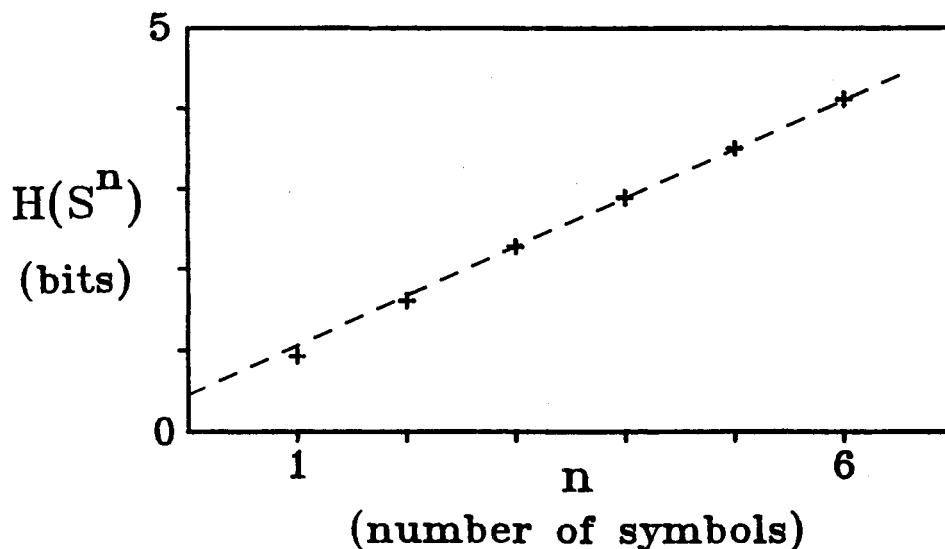
Fig. 47 – Entropy as a function of number of symbols of the "fuzzy hump" data set, for a partition taken through the peak of the hump. The asymptotic slope is maximal for this partition, but not the y intercept.
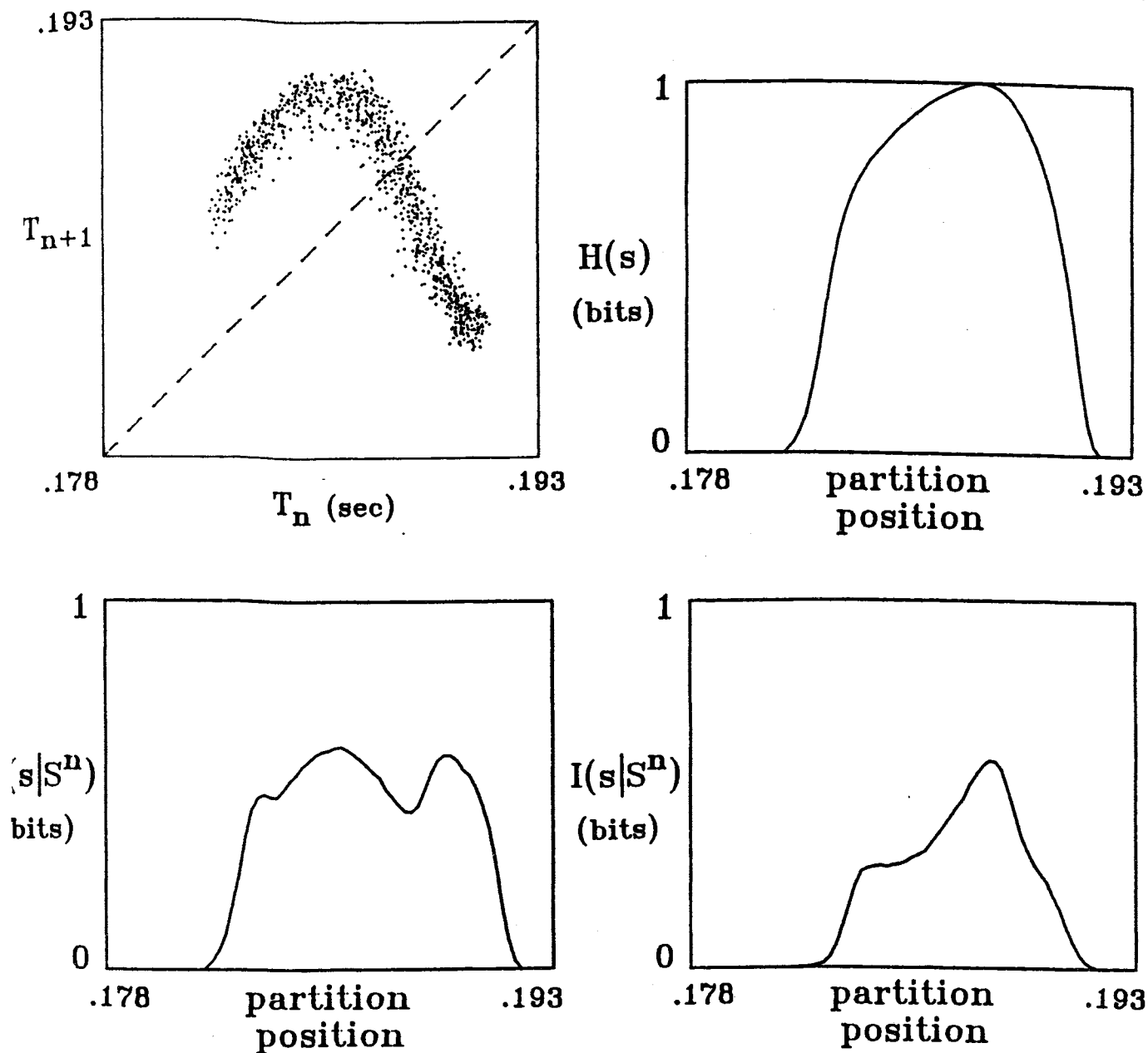
Fig. 48 – Figure corresponding to Fig. 46, for real data. Note
the general correspondence in form of the various
informational measures. A history of 7 symbols was used,
longer than the correlation length of about 4 symbols.

The effect of varying the partition position is shown in Fig. 48. The predictability and surprise of the next symbol given a past symbol history vary pretty much as in the deterministic parabolic map of Fig. 46. Apparently, some of the geometrical considerations governing the selection of optimum partitions carry over from the deterministic to the noisy case.

The reduction in predictability resulting from the coarse measurement is quantified by comparing the stored information of our best idea of the "true" variables, $I(T_{n+1}|T_n)$, with the total predictability of the future symbol string given the past symbol history, $I(s_{n+1}, s_{n+2}, \ldots | s_n, s_{n-1}, \ldots)$. Because correlations extend only over a few measurements, a calculation of the latter quantity is feasible, and is shown as a function of partition position in Fig. 49. The optimum position for best total predictability is, as for the best prediction of the single next symbol, somewhere near the identity. This best value is about .7 bits, thus the reduction due to the coarse resolution of the measurement is at least $1.8 - .7 = 1.1$ bits.
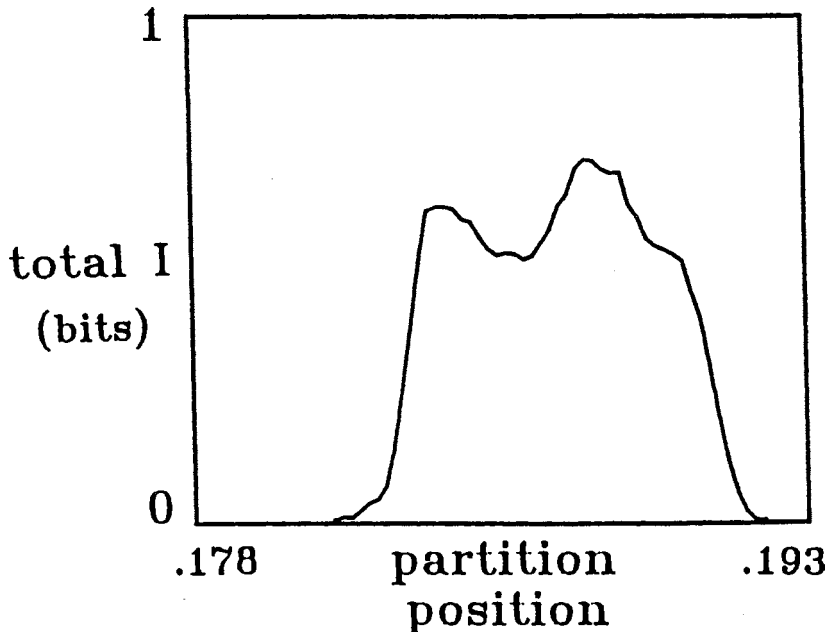

Fig. 49 – Total stored information of the data string as a function of partition position. This figure corresponds to the movement of the y intercept in Fig. 47 as the partition is varied.

We have demonstrated, at perhaps excessive length, methods of characterizing data originating from a chaotic system, with a stochastic noise element, passing through a limited measurement channel. The next section mentions some of the difficulties in applying the methods presented here.

19. Limitations of this modeling procedure

The treatment of experimental data discussed in this paper can be summarized as follows. First, an appropriate set of variables x are selected, and their values discretized into a finite number of possibilities. Next, the transition matrix $P(x'|x)$ is constructed by taking a large quantity of data and accumulating the individual matrix elements as histograms. Finally, numerical measures such as the stored information are computed, to characterize the predictability which this empirically constructed model enables.

An immediate practical difficulty with this procedure is: how do we know that we have selected an "appropriate" set of variables to describe the state of the system? For example, the number of observed variables might be less than the actual "dimensionality" of the system. Measurements will then be "incomplete" in the sense discussed in section 16, and a reconstruction process will be required to generate a more complete characterization of the system state.

Even the question of a practical definition of the "dimension" of a dynamical system is a research problem, this question is briefly discussed in appendix 2. Low-dimensional model systems, as well as experimental data presented here, can have complicated geometrical structure, with what corresponds to an intuitive notion of "dimension" varying locally. The selection of variables in this situation clearly requires the judgement of the observer, implying the prior existence of a model of a more general sort in his mind.

The computation of the stored information associated with a model $P(x'|x)$ amounts to a simple <u>counting</u> procedure, all the dimensionality and topology of the system is projected down to a single number. Much more than this will be required to describe even the gross qualitative geometric features of a system dynamic. In fact, the predictability of the immediate future may vary radically from point to point on the system attractor, this property would not be reflected in the <u>average</u> quantities we have discussed here [35]. Nevertheless, the stored information will be very useful as it does provide an objective measure of the <u>improvement</u> of an observation; as the model or the measuring instrument is improved, the numerical value of the stored information will increase.

Another serious practical problem is the amount of data required to construct a good model. The dynamic transition probabilities $P(x'|x)$ are a physical property of the autonomous system which can only be approximated by a data string of finite length. If there is not enough data, statistical fluctuations in the histogram approximation to $P(x'|x)$ will add false structure to the model, resulting in an artificially high value of the stored information.

The approach to the data taken in this paper was to use histogram bins wide enough to get good statistics, but narrow enough to be below the noise level, and thus capture all the deterministic dynamics. The data records taken so far, typically of lengths of four to eight thousand drop intervals, are in fact only marginally able to satisfy both these conditions simultaneously, even for the simple attractor used as the main example. Also, the noise level below which no causal structure exists is to an extent an <u>assumption,</u> a judgement on the part of the experimenter.

The accumulation of histograms is actually often a poor way to measure a physical probability. Statistical convergence is slow, proceeding as the square root of the number of samples, and the sheer size of the transition matrix for a

high-dimensional system renders the histogram method completely infeasible. Yet there must exist more efficient methods for constructing models, or representing expectations, as demonstrated by the simple fact of our ability to function successfully in a multidimensional world.

Bayesian methods may aid in treating some of the problems of insufficient statistics. Here typically the physical probability distribution is <u>estimated</u> on the basis of a limited amount of data, using certain <u>assumptions</u> about some properties of the distribution to be measured. The consistency of the observed data with these assumptions is checked, and the assumed prior properties of the distribution can then be modified to provide a better fit with succeeding data. The problem of estimating the size of a space of N states from limited data provides a simple example of this approach. If a random process selects out the states at a rate of one per time unit, we would have to wait a time on the order of N log N to see all the states. But <u>duplications</u> in observed states will start to occur at times on the order of the square root of N, and this can be the basis of an estimate of the size of N. This situation is analogous to the "birthday problem": given a room full of people, how many have the same birthday? See work by Erber[36], Ma[37], and others. To improve our ability to construct models with limited data, further research is indicated into describing the expected properties of the probability distributions of various sorts of systems, and the effects of insufficient statistics on entropy-like measures. Some studies of this sort are underway[38].

20. More general models, how "good" is a model?

There is a large conceptual difference between the simple physical model described in section 5 and the histograms accumulated from the data. The latter is just a representation of the data, every transition is explicitly described (although as mentioned above some thought enters into the very act of setting up the measurement channel). The data histograms in a sense can't be wrong, but despite their exhaustive description of the data they in themselves have no explanative value.

Physical models such as the mass-spring example are not constructed directly from the data, but rather by analogy. They consist of algorithms to predict transition matrix entries, in the absence of the data. There are two considerable advantages of this sort of model over the complete transition matrix. First it is much more concise, only desired matrix elements need be actually generated. Second because it was generated by an analogy, it often suggests generalizations beyond the particular data, allowing for example predictions about the effect of varying system parameters.

The raw histograms, despite their limitations, have the nice feature that there exists a measure of quality, the stored information, which is independent of external auxiliary measures such as "utility functions". This measure is a property of any string of numbers derived from a time-invariant process. But once a model of the second type has been constructed by the observer, there are two streams of numbers produced, one by the model and one by the reality, and a measure of the "goodness" of the model will be some sort of "distance" between the two number strings (Fig. 50).

The value of a model in the real world depends upon its use, and thus in general rests on an external evaluation function. Shannon, in the context of
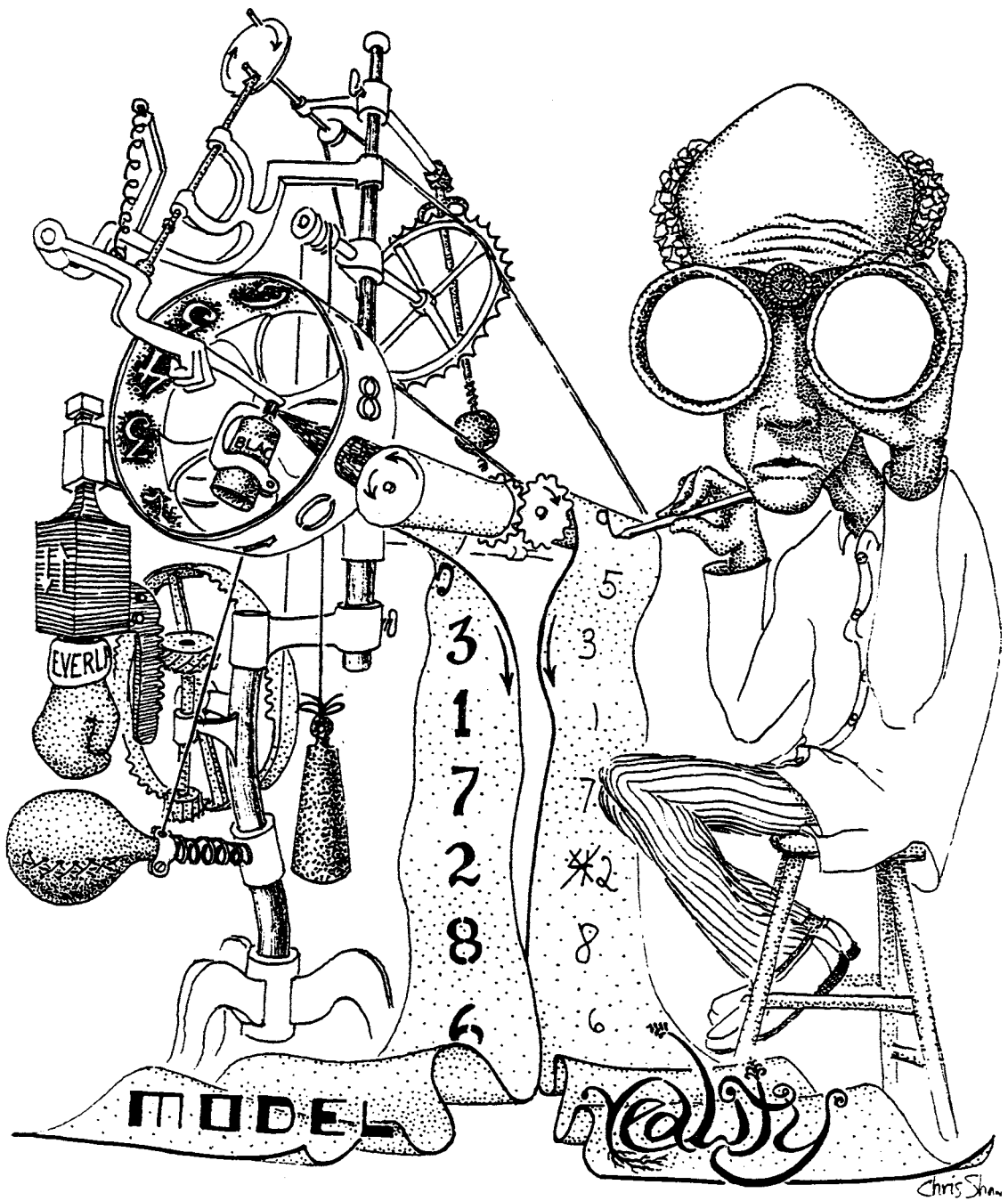
Fig. 50 - The two data streams.

communication theory, introduces a "fidelity criterion" of fairly general form to measure a distance between an input and output number string. This measure, dependent on the value to the user of a particular degree of accuracy in channel transmission, determines an effective rate of transmission through a noisy channel. In the biological world, the effective "evaluation function" judging the utility of a model is often severely binary. A fox and a hare carry with them a rather detailed model of each other's behavior, yet in their interaction all the subtlety is collapsed to two outcomes, the hare is eaten or the fox goes hungry.

If one is willing to simply count distinguishable states and transitions between them, without judging the value of particular items of information, and if a large amount of data is available, a "distance" between model and observed data strings can be defined, independent of an evaluation function. This can be done by computing the predictability of one string given the other. Typically a physically observed initial condition is supplied to the model, and the subsequent physical and model data streams compared. A certain number of digits will correspond between the two number streams, and a certain number will differ, see Fig. 51. The _difference_ between the two number streams is, to the observer, an _unpredictable_ element, and thus represents a _source of information_. For a noisy system (or model), the amount of information so generated will diverge for measurements of arbitrary precision, as discussed earlier. However, the average number of matching digits, and the average rate of loss of matching digits as predictions farther into the future are attempted, can be recorded. It is unclear to the writer whether such measures are nicely coordinate independent, but they seem operationally well-defined in a given setting, and capable of measuring _improvement_ of a model. An improved model will result in a lower rate of divergence between the two number strings, less "surprise" per unit time to an observer making continuous measurements at fixed resolution.

```
        "model"        "observed data"

       .0110101         .0110011

       .1011001         .1011010

       .1101110         .1101001

       .0010010         .0010100          time              Fig. 51

       .0101011         .0100111

       .1000110         .1001010

       .1110101         .1111000
           .                .
           .                .
           .                .
```

These ideas have already appeared in practical use. In a recent talk[40], Cecil Leith discussed a large computer model used for weather prediction. An observed weather pattern is given to the model as an initial condition, and the evolution of the model and the actual weather system compared. The <u>exponential rate of divergence</u> of the two paths through the space of variables is computed, and used to indicate improvement in the model.

It should be noted that predictive errors of fixed constants, or fixed drift rates do not reduce the "goodness" of a model under this sort of measure. Eclipse predictions which are always late by exactly twenty-four hours are of equal value as correct predictions, once the fixed delay has been recognized. In principle, a "table of corrections" can be prepared over a short time to compensate for fixed errors, and linear and even polynomial tendencies for the two strings to drift apart. Only <u>exponential</u> rates of separation will give finite contributions to the logarithmic measures of information.

A second sort of measure of the "goodness" of a model arises from considerations of the <u>conciseness</u> of a model. To actually generate a prediction from a model requires a computation, and one can give a length in bits for a computer program to implement the calculation. Or one can measure the length of the message which communicates the model to another person. If the same "distance" under some measure between model and reality can be achieved with a shorter model, we usually feel that the shorter model is "better". These ideas have been developed in the field known as "algorithmic information theory", by Kolmogorov, Chaitin, and others. A very readable introduction is Chaitin's Scientific American article[41], as well as Joe Ford's article, "How Random is a Coin Toss?"[42]. Chaitin and Kolmogorov define the "randomness" of a number string to be the length of the computer program required to generate it. A "completely random" number can only be specified by explicitly listing each digit, thus the corresponding computer program must be of at least equal length. The asymptotic ratio of program length to output length will be one "bit per bit" for a completely random number, and zero for a completely determined number string, e.g. 1010101010... .

A highly concise model is useless if says nothing about the real data string it is supposed to predict. Thus a reasonable measure of the "goodness" of a model should compare the length of the model to the predictability it enables. For example, one might measure the total size of the body of information comprising the model, including the length of the algorithm, the size of any "table of corrections", and the amount of initial data used, and compare it to the total predictability it enables, that is, the total number of nonredundant "matching digits" integrated over the future.

The chief limitation of the modeling schemes described in this paper, and indeed of all of information theory as it presently exists, is the limitation to

autonomous systems. In any hydrodynamic system of reasonable size, as well as in daily life, the space of possibilities is so huge that all that can be observed are <u>unique</u> <u>events</u>. Only a tiny fraction of the possible distinguishable states are visited in a reasonable time, and the accumulation of average transition rates between individual states is out of the question. But we know that our brains can organize experience, and represent expectations about reality in a way that allows considerable prediction. Despite our constant use of this facility, we have virtually no exterior understanding of it, as demonstrated by our inability to incorporate even a rudimentary form of this process in any man-made machine.

A "good model" in physics participates in this unknown interior organizing process, and is said to be <u>explanative</u>. At this point a "good explanation" can only be defined in terms of the subjective feeling of satisfaction it produces, but before subjective sensation is deemed out of place in a discussion of the construction of models, it should be remembered that such feelings or instincts guide all creative work, scientific or otherwise. A model can have "explanative value" even when it has no predictive power, the outstanding example being the theory of evolution.

21. The <u>undecidability</u> <u>of</u> <u>optimum</u> <u>modeling</u>

As a model is improved, the discrepancy between its predictions and data from the system under study becomes less and less. Under an older view of things, this process could be continued indefinitely; successive improvements to the model could bring the two number strings into coincidence to as great an accuracy as desired. However, if the given system is chaotic, a minimum rate of divergence between the two number strings is guaranteed. Even if the model were an <u>exact</u> <u>physical</u> <u>duplicate</u> of the original system, and was set to the same initial condition, the two would diverge at a rate given by the entropy of the

system. In this situation it becomes reasonable to ask, how are we to know when
we have reached this intrinsic limit to predictability? How are we to know when
we have constructed the optimum model?

A similar question arises in an observational context. It is a common
experience for those of us who conduct physical experiments that data improves
over time, often for no apparent reason. For example, the faucet which produced
the data of Fig. 52a below might, several weeks later, yield the data of Fig.
52b at the same flow rate. The improvement in the data is quantifiable as a
measured increase in the information storage capacity of the system, computed as
described in this paper. How are we to know that further improvements in equip-
ment or technique will not result in even better data, and increased predicta-
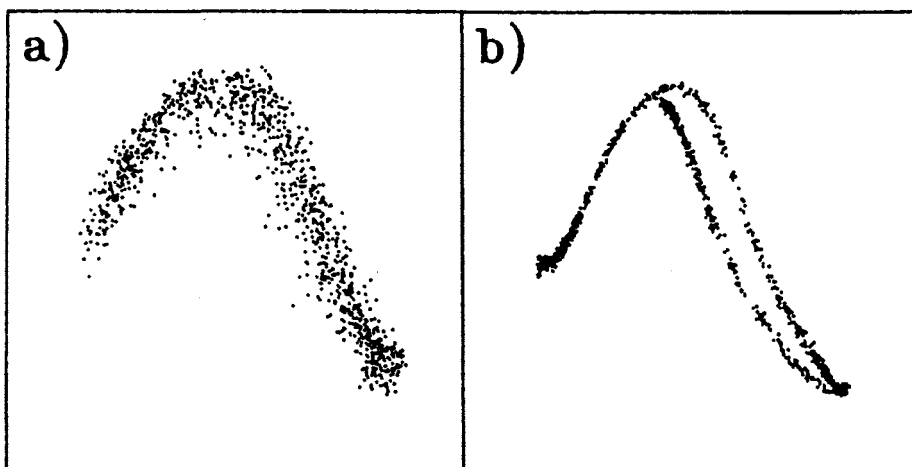bility of the system?

Fig. 52

The purpose of this section is to make the fairly obvious point that we
don't. Furthermore, it will be argued, the question of whether a model is
optimum is undecidable in the sense of Gödel.

From a classical perspective, it is clear that an experimenter's claim that he is "in the noise", and that there is no point in attempting to increase resolution, is an assumption. If the observation rate is less than the entropy of the system at some resolution level, the system will appear stochastic, even if it were "completely deterministic". Boltzmann recognized the logical necessity for this sort of assumption, it appears as an "assumption of molecular chaos" in his treatment of the kinetic theory of gases. In the more modern view this assumption becomes labeled a fact of Nature, and the experimenter will point to "thermal kT noise", or even "basic quantum mechanical indeterminacy" as justification for not expending more effort on resolution. Still, the question of ultimate resolution remains a practical one for chaotic systems with measured noise levels far above the kT level, as is probably the case for the water drop system, and certainly for weather systems.

The concept of "undecidability" has been developed in the context of discrete state logical systems and computing machines, and has not, to my knowledge, been generalized to continuous systems or systems with a noise element. Nevertheless let us usurp some vocabulary and results for use in dynamical systems theory, noting that a large discrete state system can accurately mimic the behavior of a continuous system, and that the addition of noise to a system is unlikely to decrease any "undecidability" it might have.

The discrepancy between the predictions of a model and observations of a physical system is, to the observer, a random element. If the model embodies the observer's complete knowledge of a system (including perhaps the addition of a noise element), he will be unable to distinguish the two resulting number streams, should he get them mixed up. A "proof" that the model was the best possible would involve a proof that the random element, or discrepancy, was intrinsic to the system, or "truly random", not susceptable to any further

deterministic description. However, as Chaitin and others argue in the context of the theory of algorithms, one cannot "prove" a number to be random, to possess such a proof would imply a logical contradiction. Thus, given a set of seemingly random numbers, there is no way in general to show that there does not exist an algorithm shorter than the list itself to generate the numbers, and thus reduce the randomness.

This result makes a certain amount of intuitive sense. A "proof" is a manipulation of the fixed axioms of a closed logical system, whereas "random numbers" are exactly those entities which come from outside the logical system, that is, cannot be generated within the system. To "prove" the number random, the system would have to recognize the number as coming from outside. As there is no representation of the number inside the system, the comparison required by a recognition process cannot occur.

The process of divining a deterministic aspect to a string of data, thus reducing its randomness, might be likened to "breaking a code". Man-made codes, where the apparent randomness is deliberately maximized, illustrate difficulties which in principle can arise in attempts to model natural phenomena. For example, a type of cipher called an "open key" code has been studied recently[43]. Here, even when the algorithm for generating encrypted text is known, the text cannot be decoded by a simple algorithm (it is believed). Only an exhaustive search of impractical length can find the determinism and reduce the randomness of the text string. So we can at least imagine a situation where even if we had guessed a correct model for a physical process, we could not demonstrate that the observed data came from that process.

It is on this basis that one might argue that a description of theoretical physics in terms of particles and interactions between them can only be partial. Even if one finds the elemental rules, determining their qualitative

consequences on a macroscopic scale might require a computation of indefinite length, with a non-unique outcome. Oh, well. Maybe Gödel's theorem has nothing to do with chaos, and there is only a desert where these two questions meet. Then again, study here might address one of the larger issues; what is the limit to man's ability to predict and thus control his own evolution?

22. Conclusion, "ideas" as models

This paper has presented, by means of a physical example, preliminary work on the problem of predictability in the presence of noise. Computation of the stored information and rate of loss of stored information for data generated by the water drop system provided examples of the usefulness of these statistics to describe a degree of predictability, and to distinguish a difference between unpredictability arising from deterministic and stochastic sources. A more general discussion of the problem of modeling was attempted.

The limitations of the procedures illustrated above point out the nearly total lack of understanding of the neural modeling system heired to each of us as our evolutionary birthright. Even an ant possesses capabilities which we are unable to incorporate into our machines. Any general theory of modeling must explain our ability to move easily through a high dimensional world, and, above all, to generate new ideas. Incremental learning, which occurs for example in acquiring a skill, might be described by some sort of conditional distribution which is continually updated in a Bayesian fashion, but an idea, as Drake notes[44], seems a more abrupt occasion. An "idea" can be described as the sudden construction of a model for sense data. Some kind of connection is realized, often increasing the predictability of the perceived world.

The occasion of "having an idea" seems a curious combination of the active and the passive. One can actively accumulate knowledge, and actively exclude

from the local environment influences which deaden creative thought, but one cannot actively make an idea. Instead, one has an idea, a passive process which often involves some amount of waiting around. The delay is probably due to a process more akin to a random search than a computation, as the brain's ordinary recall is usually rapid, taking only a few seconds.

This suggests that there are parts of the brain optimized for making rapid searches through a very large space of possibilities, together with a signal to the 'operator' when a connection has been made. This signal takes the form of a pleasant subjective sensation, often described in the psychological literature as the "aha" experience. It can even happen that one experiences the sensation of an idea about to occur, before the logical content of the idea becomes explicit.

To the extent that we are materialists, we must believe that underlying such events is a physical process. How is it that we can accumulate, and even, to use Jim Crutchfield's phrase, anticipate knowledge? What are the operating principles of the dizzying jewel from which the natural world turns back to gaze upon itself? Perhaps more can be learned from a study of this mechanism from an interior, as well as an exterior perspective. Perhaps there is experimental physics still to be done without even getting out of bed.

Ludwig Boltzmann wrote the entry entitled "model" for the 1902 edition of the Encyclopaedia Britannica. Here he states:

> "...Long ago philosophy perceived the essence of our process of thought to lie in the fact that we attach to the various real objects around us par-ticular physical attributes - our concepts - and by means of these try to represent the objects to our minds... On this view our thoughts stand to things in the same relation as models to the objects they represent. The essence of the process is the attachment of one concept having a definite content to each thing, but without implying complete similarity between thing and thought; for naturally we can know but little of the resemblance of our thoughts to the things to which we attach them. What resemblance there is lies principally in the nature of the connexion..."

Boltzmann lived too early to learn the physical and topological "nature of the connexion". We may be more lucky.

I hope the reader can forgive the loose tone of some of this discussion, and that if it is not science, it is at least entertainment. Perhaps, though, contemplation of a dripping faucet can supply a few rungs in the ladder of metaphor which must be built to understand modeling in a larger sense. When the falling drop is perceived, it falls in the brain as well, outlined in sparkling surreal intensity. The world swarms with life forms, many bearing an individual neural net. Each thin web is capable of directing a complex entity, and in each is mirrored the passion of the world. Yet this instrument sprang from the slime!

Finally, I hope this work suggests that one does not need a particle accelerator to step to the frontiers of physics. The tapestry of reality flutters with moving causal structures, both preserving and generating information at <u>all</u> length scales. This brew of freedom and constraint generates a complexity that renders a simple faucet an enigma, let alone consciousness and evolution. Thus even if the particle physicists were to succeed, and all the laws governing the fundamental building blocks of matter were known, we would still walk down the street without knowing how or why. The mystery remains, as always, as close as our forehead.

Acknowledgements

Appendix 1 – Continuous time calculation

Section 12 describes the computation of the predictability of a time interval T_{n+1} obtained from a measured time interval T_n, given the conditional distribution $P(T_{n+1}|T_n)$ and the equilibrium distribution of intervals $\overline{P}(T)$. In fact, although the water drop system produces data in the form of discrete time intervals, these intervals are imbedded in continuous time. A complete description of the measureable state of the system would include the time elapsed since the preceding drop. The additional phase degree of freedom increases the storage capacity of the system. This larger amount of stored information can be computed from the same data set used to compute the interval-to-interval stored information.

$$T_1 \qquad T_2 \qquad T_3$$

$$t_1 \quad t_2$$
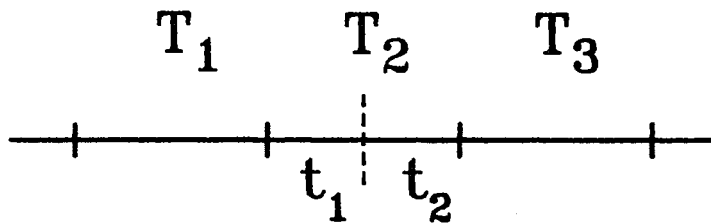
The figure above schematically represents the situation at a particular instant of time, indicated by the dotted line, during a sequence of drop intervals. The time elapsed since the preceding drop is t_1, the time until the next is t_2. Knowledge of the state of the system requires knowledge of the length of some drop interval, and its phase. For an observer at the dotted line to be informed of the system state, he must be told the numbers T_1 and t_1. In the absence of prior observation, he will have to wait until the end of the interval T_3 before a system state is determined. His a priori expectation of observing a pair of numbers t_2 and T_3 is described by a distribution $\overline{P}(t_2,T_3)$ which, by time symmetry, is the same as the distribution of prior values $\overline{P}(t_1,T_1)$.

The stored information is defined in this paper to be the value of the knowledge of the immediate past state in determining the immediate future state. Thus, we must compute the value of the knowledge of the numbers T_1 and t_1 in predicting t_2 and T_3. These four numbers are connected by a conditional distribution $P(t_2, T_3 | t_1, T_1)$, and the full stored information is given by the expression

$$I = \iint P(t_1, T_1) \iint P(t_2, T_3 | t_1, T_1) \log \frac{P(t_2, T_3 | t_1, T_1)}{\overline{P}(t_2, T_3)} \, dt_2 \, dT_3 \, dt_1 \, dT_1$$

Now in this case, we do not need to compute the full four-dimensional distribution, because

$$I(t_2, T_3 | t_1, T_1) = I(t_2 | t_1, T_1)$$

This can be shown algebraically, using $P(T_3 | t_1, t_2, T_1) = P(T_3 | t_1, t_2)$, or appreciated by noting that the predictability of T_3 is completely determined by $T_2 = t_1 + t_2$. Thus there is no additional stored information associated with higher order distributions involving T_3. So the full stored information simplifies to

$$I = \iint \overline{P}(t_1, T_1) \int P(t_2 | t_1, T_1) \log \frac{P(t_2 | t_1, T_1)}{\overline{P}(t_2)} \, dt_2 \, dt_1 \, dT_1$$

The problem then is to express $\overline{P}(t_1, T_1)$, $\overline{P}(t_2)$ and $P(t_2 | t_1, T_1)$ in terms of the measured distributions $\overline{P}(T)$ and $P_o(T_{n+1} | T_n)$. The subscript has been added to $P(T_{n+1} | T_n)$ to distinguish this particular measured distribution in what follows.

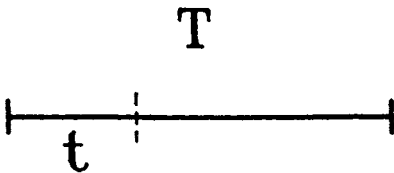First, we will determine the equilibrium distribution $\overline{P}(t_1, T_1)$. Now

$$\overline{P}(t_1, T_1) = P(t_1 | T_1) \, \overline{P}(T_1)$$

and

$$P(t_1 | T_1) = \int P(t_1 | T_2)\, P_0(T_2 | T_1)\, dT_2$$

$$= \int_{t_1}^{\infty} \frac{P_0(T_2 | T_1)}{T_2}\, dT_2$$

This last step is made clear by examining the figure below:

$$P(t | T) = \begin{cases} 1/T & t \leq T \\ 0 & t > T \end{cases}$$

If we divide T at random, all subintervals from zero up to the full length of T are equiprobable. By a similar argument,

$$\overline{P}(t) = \int_{t}^{\infty} \frac{\overline{P}(T)}{T}\, dT$$

The remaining step is to observe that

$$P(t_2 | t_1, T_1) = P_0(t_1 + t_2 | T_1)$$

Thus all parts of the full continuous information integral can be reduced to sums over the measured distributions, as claimed. This integral is approximated by a discrete sum via a binning procedure as before.

In the case where the variation in drop intervals is only a small fraction of the average drop interval, as is the case in the "fuzzy hump" data set, the computation of the stored information simplifies further.
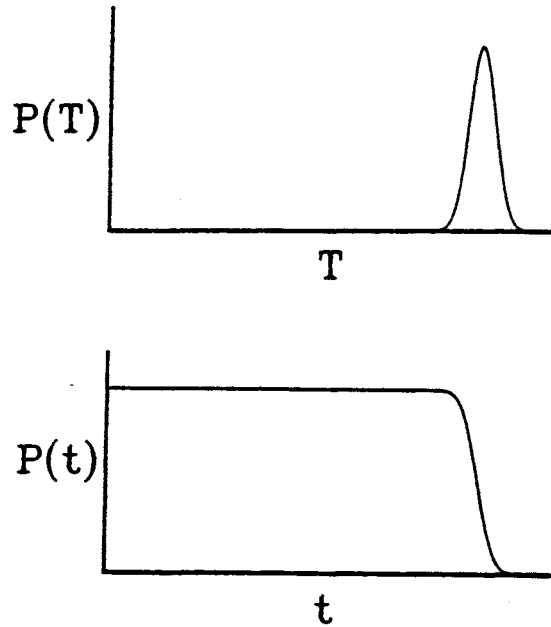
As the above figures suggest, $\overline{P}(t)$ is well approximated by a constant. Also,

$$P(t_1|T_1) \approx \overline{P}(t_1)$$

again because of the small relative variation in T. The two approximations

$$\overline{P}(t_2) \approx 1/T_{average} \quad \text{and} \quad P(t_1|T_1) \approx 1/T_{average}$$

were used in the full stored information calculation illustrated in Fig. 32.

Appendix 2 - The dimension question

The central idea behind the Lorenz, Ruelle-Takens picture of turbulence is that the random element of a turbulent flow can be described, at least in some limited situations, by a model with only a few "dimensions", or "degrees of freedom". It is clear that a large amount of information is required to describe the state of a fluid system in "fully developed" turbulence, but the experimental evidence indicates that near the transition to aperiodic behavior, a "low-dimensional" model often adequately describes the behavior. But how does one define the "dimensionality", "number of modes", or "degrees of freedom" of a physical system? Although the concept of "dimension" is intuitive, operational definitions are less clear. The dimension of a mathematical model is defined in its construction, but in a physical system, the "dimension" becomes a quantity to be determined by observation. One approach, employed by a number of workers, is to examine the dimensionality of an attractor in a state space reconstructed from the time series data of a physical experiment. For example, if some system is undergoing a stable oscillation, its motion can be represented by a limit cycle in an appropriate state space, a one-dimensional closed curve.

There has been considerable recent work in an effort to extract statistical measures from experimental data, and relate them to formally defined measures of dimension. A review cannot be attempted here, suggested references are [45], especially the article of Farmer, Ott, and Yorke. Here will be given only a beginner's dimension calculation for some water drop data, a short general discussion, and a few criticisms and suggestions for future experiments. The writer has benefited from discussions with some of the principles, including Doyne Farmer, John Guckenheimer, and Jim Yorke.

The dimension of an object should describe how its volume scales with increasing linear size. Proceeding informally, let's think of an attractor as a compact object with a "volume" of N equiprobable distinguishable states, localized in state space within a region of radius R. If the "radius" of a single distinguishable state is r, then the dimension of the region containing N states is given by:

$$N = (R/r)^d.$$

Taking the log of this relation, we have:

$$\log(N) = d \log(R/r).$$

The left-hand side is simply the stored information, or number of bits resolvable on the attractor, while $\log(R/r)$ is the resolution obtainable from a measurement along a single linear attractor coordinate. Thus the dimension of the attractor, roughly speaking, is the minimum number of independent measurements required to localize a single system state. If our view of the attractor is limited by the measurement, that is, the stored information intrinsic to the attractor is much larger than the measurement channel rate, or in the notation of this paper:

$$I(x'|x) \gg I(x|y)$$

then we can expect an improved total $\log(N)$ as measurements are improved. The dimension d could then be independent of resolution over some range, and its value would indicate the ratio of increased predictability to increased resolution (as pointed out by Farmer et al).

The number of states "nearby" a particular state should be a function of the dimension of an attractor, and this provides the basis for most of the dimension calculations. Typically, the number of states within some length L is determined as a function of L, $N(L) = L^d$ where d is the dimension. This approach will now be used to attempt a "dimension" measurement for three sets of water drop data. The algorithm below was chosen on the basis of simplicity, it is I believe closest to the "correlation exponent" of Grassberger and Procaccia, and the "pointwise dimension" of Farmer et al, and Guckenheimer[46].

First, a point is picked at random out of the data and its "distance" from all other data points along the one-dimensional drop interval coordinate determined. The resulting list of distances is then sorted in order of distance to nearest neighbor, distance to next-nearest neighbor, etc. This process is repeated for a number of other randomly selected data points, then a _mean_ distance to nearest neighbor, next-nearest neighbor, etc. is computed. This distribution of distances is inverted, and plotted as log(N) vs. log(L), where N is the number of data points within a distance L [47].

Next the data is taken in _pairs_ of successive points, and distances to all successive pairs in the resulting two-dimensional state space computed using, for example, a standard euclidian metric. Mean distances to nearest neighbor, next-nearest neighbor and so forth are computed as before, and log(N) vs. log(L) plotted.

Successive curves are generated by taking the data in successive triplets, 4´s, 5´s, and 6´s. Each step corresponds to increasing the dimension of the state space by one. In principle the observed dimensionality of the attractor generated by the data should also increase, until the dimensionality of the imbedding space is larger than the actual dimensionality of the attractor. Once there is room enough in the imbedding space for the full attractor, the observed

attractor dimension should arrive at is "true" value. Thus, in theory, all one has to do is measure the <u>slope</u> of successive curves, when it stops changing as a function of imbedding dimension, one has determined the dimension of the physical system.

In practice, things are not so simple. Figures A1 - A3 below displays the results of applying the above program to three data sets. For the first data set, the dimension calculation seems rather successful. The attractor of this water drop regime seems to be a gauzy assemblage of two-dimensional ribbons and sheets, if the data is viewed as a stereo plot. The calculated curves become more or less parallel after an imbedding dimension of two is reached, and a fit to the slope of the later curves near their center yields a value surprisingly (to the writer) close to two. However, the second data set demonstrates the problems which quickly arise if the probability distribution on the attractor is not particularly uniform. This set is generated by a somewhat noisy period-two drop regime. The purely stochastic part of the data generates dimension curves whose slopes continuously increase with imbedding dimension, as expected. But there is a radical break in the curves at length scales corresponding to the distance between the two periodic points.

One might learn to usefully interpret a set of curves such as these, but the more usual situation seems to be exemplified by the third data set. To the eye, the attractor seems simple enough, it's a bent one-dimensional string with a little noise on it. The calculated dimension curves, however, seem rather uninformative, there is no clear place to fit a slope, and no apparent way to see that these curves were generated by a quasi one-dimensional object.
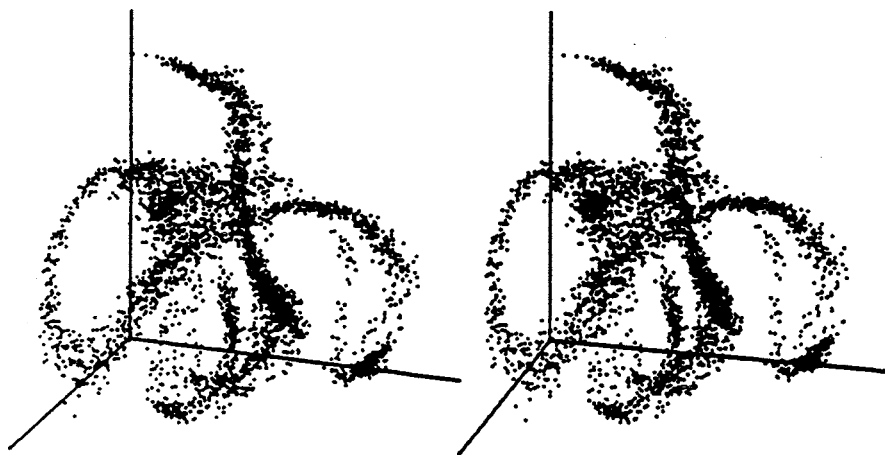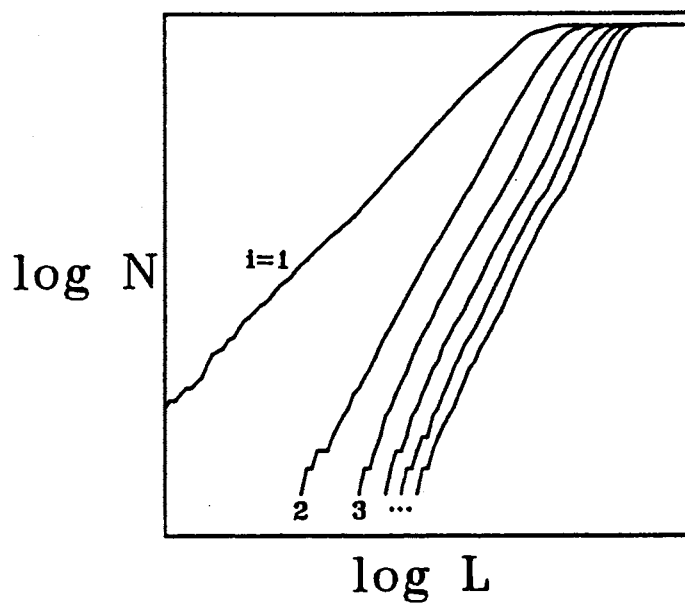
Fig. Al — Dimension curves and attractor. For an embedding dimension
of one, the curve has a slope near one, for all higher
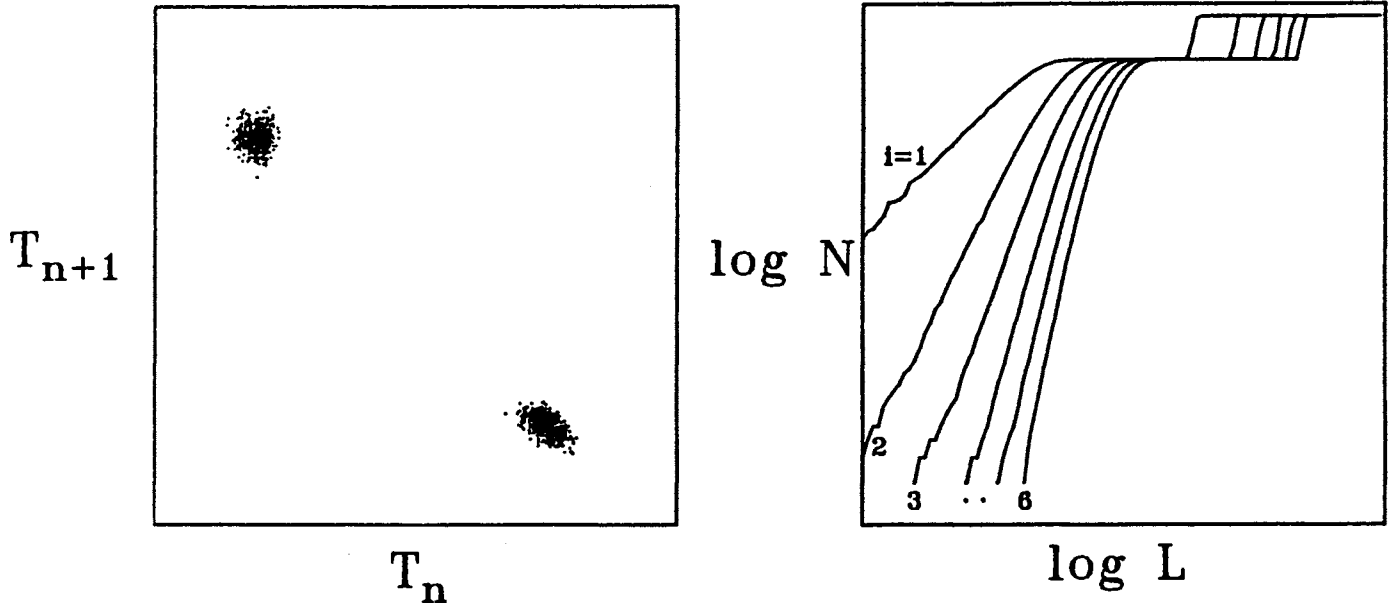embedding dimensions, the slope is near two.

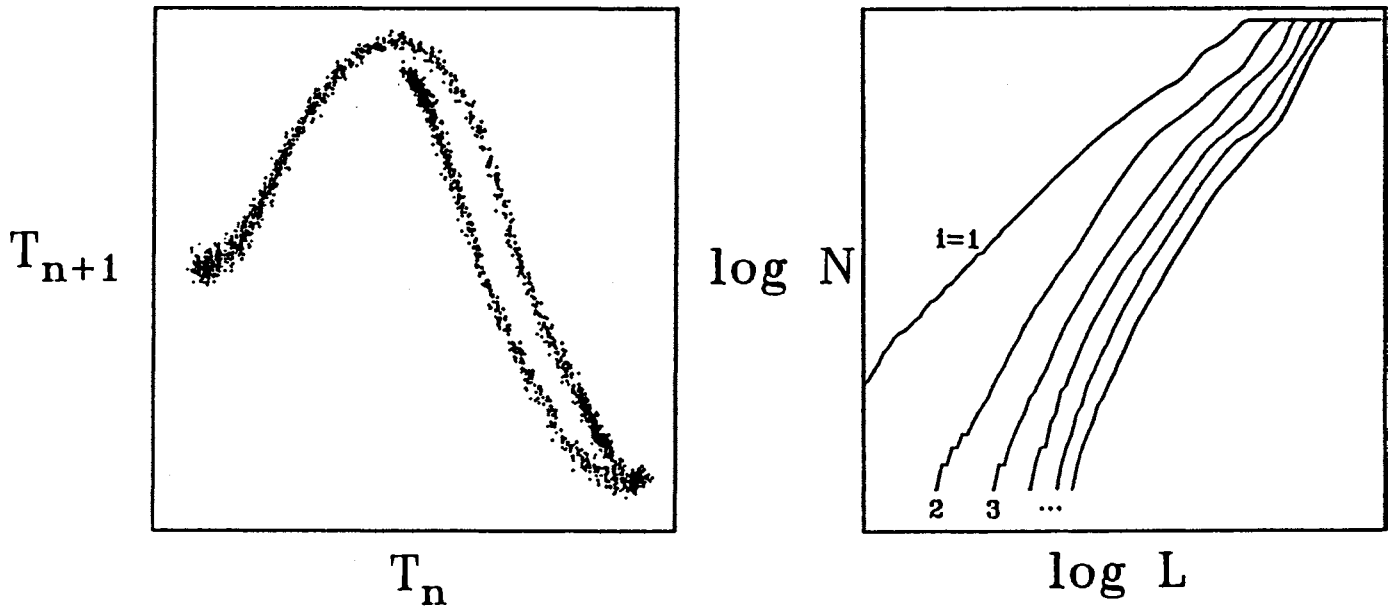Fig. A2 - Dimension curves for period-two regime.

Fig. A3 - Dimension curves for quasi one-dimensional object.

The good news is that sets of dimension curves, even if not particularly useful, are apparently coordinate-independent. The writer tried three different metrics in computing "distances" in the embedding spaces: euclidian, sup norm, and the "manhattan" metric (sum of absolute values of coordinate differences). The dimension curves were translated horizontally in the figures by varying amounts depending on metric, but their general form even in detail was well-preserved.

Armed with this limited experience, we will now present a few criticisms of the dimension industry, as it currently operates. Implicit in most of the recent work on dimension is the assumption that there in fact exists a single number which characterizes the dimension of a chaotic attractor. Discussions often begin by considering Cantor sets, and other self-similar "fractal" objects, where the "dimension" describes scaling relationships as finer and finer length scales are considered. In a purely deterministic treatment, chaotic attractors possess such a Cantor set structure; as the information storage capacity is infinite, the complete past and future history of the dynamics is determined by the infinitesimal structure to be found at any part of the attractor. Thus a dimension calculation using any part of the attractor would yield the same number.

But the base length scale which all physical systems, including digital computers, possess limits the degree of structure the underlying attractor can have, and likewise limits the history it can embody to a finite and computable length of time. So it is to be expected that the "dimension" will usually vary across the attractor, if by "dimension" we mean the number of coordinates required to maximally localize a state. As a simple example, consider a system which in its Poincaré section periodically changes from a fixed point to a limit cycle and back again:
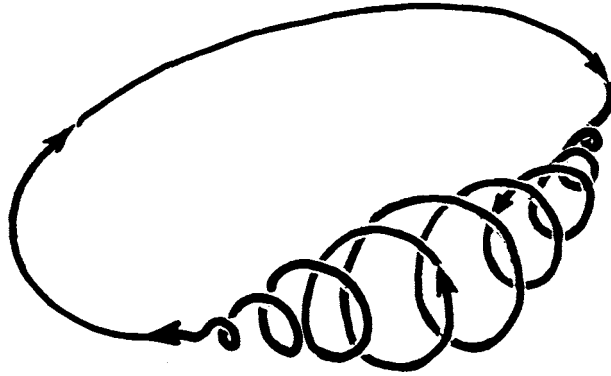
Each time the limit cycle shrinks to a "point" beneath the noise length scale, its phase is lost, and each time it reemerges, its phase is a random variable, determined by the noise. Thus two coordinates are required to achieve maximum predictabilty in the lower part of the figure, and only one in the upper. Objects such as this whose actual structure is supported by noise form an interesting new class, see recent work by Farmer[35] and Deissler[48].

The usefulness of a "dimension" number as a scaling exponent depends on the number of orders of magnitude between the smallest and largest length scales. This range is in fact severely limited for most physical data. The scaling structures which have received so much recent theoretical attention are notably absent in most of the fluid systems so far examined, due to the limited dynamical range, and high dissipation. The experience in this laboratory is that the more low-dimensional structure there is in the reconstructed attractor, the less well-defined is a single "dimension" number.

This is not to say that the set of dimension curves may not have a use as indicating an overall character of a data set, much as a power spectra curve does. But there seems little reason at this point to prefer one dimension algorithm over another as "more fundamental".

The most serious limitation of existing "dimension" measurements of physical systems is their total lack of any consideration of the effect of <u>spatial</u> variation in the system. In the work described here, and in all other work of which the writer is aware, data is supplied by only a <u>single</u> probe of the fluid, and an attractor supposedly characterizing the entire system is obtained by the time delay reconstruction method. Even in the best of circumstances, such a reconstruction necessarily takes place over an interval of time, and for a system with a positive entropy, the causal connection between successive measurements is reduced. In this case, two measurements at different points in the fluid will always yield greater predictability than one. This is intuitively obvious, and corresponds to an increase in the "rate of measurement".

Any discussion of the "dimension" or "degrees of freedom" of a fluid flow must address the flow of causality in a spatial sense. Fluid motion is constrained on the smallest scales by viscosity, and at the largest, by boundary conditions. A fluid moving nonperiodically has managed to gain a degree of independence from these constraints. Experiments have indicated that this can happen in several different ways. If boundaries are close, compared to the typical roll or wave length scale, the fluid motion tends to be collective. The same attractor is obtained by reconstruction of measurements made at any point in the fluid, subject to the resolution limitations discussed above. If the boundaries are far apart, dislocations or defects in regular flow patterns tend to occur (Donnelly's "turbators"), and the transition to turbulence is more akin to the melting of a crystal. At the onset of nonperiodicity in the latter case, the space of possible fluid flow patterns is apparently impossibly large, with a complicated connectivity, and the path of the fluid system through this state space tortuous, even if the entropy, or rate of divergence from any one path is small[49].

It should be possible to distinguish collective coherent nonperiodic fluid motion from non-coherent motion using the same stored information statistic discussed earlier in this paper. Two measurements taken simultaneously at points as close as possible in space will give the same stored information, have the same degree of causal connection, as two measurements close in time taken from a single probe. Thus the loss of information as it flows through the fluid can be measured, at least in an average sense, by mapping the information common to both probes as a function of distance. For a collective fluid motion, we would expect a rather slow drop-off of the mutual information as two probes are separated, with significant correlations remaining throughout the fluid. A quicker drop-off might well be associated with certain fluid flow features passing between the two probes, and might indicate that information transmission through such features is reduced. It may also be possible to quantitatively characterize more fully developed turbulence, measuring the size of typical coherent volumes, as well as the "entropy per unit volume", or "degrees of freedom per unit volume", or "number of positive characteristic exponents per unit volume"[50].

Lack of funding has precluded the early experimental test of these ideas.

References

1) S. Hawking, "Is the End in Sight for Theoretical Physics?"
   Cambridge University Press, 1980
2) R. Abraham and C.D. Shaw, "Dynamics: The Geometry of Behavior",
   Vol 1: Periodic Behavior, Vol. 2: Chaotic Behavior
   Aerial Press, Box 1360, Santa Cruz, Cal. 95061
   Annals of NY Academy of Sciences, Vol. 316 (1978) and Vol. 357 (1980)
   Physica D, Vol. 7 (1983)
   A.J. Lichtenberg and M.A. Lieberman, "Regular and Stochastic
   Motion", Springer-Verlag, 1983.
   J. Guckenheimer and P. Holmes, "Nonlinear Oscillations,
   Dynamical Systems, and Bifurcation of Vector Fields",
   Springer-Verlag, 1983.
   "Synergetics, A Workshop", ed. H. Haken, Springer-Verlag 1977.
   See also Refs. 23 and 30.
3) O. Rössler, in "Synergetics: A Workshop", H. Haken, Ed.
   Springer-Verlag 1977.
4) (gallons per fortnite)
5) S. Hartland and R.W. Hartley, "Axisymmetric Fluid-Liquid
   Interfaces", Elsevier, 1976.
   M. Hozawa et al, J. Chem. Eng. Japan $\underline{14}$, 358 (1981)
6) P. Martien, S. Pope, P. Scott, and R. Shaw, in preparation,
   to appear, and to be published, 1984.
   P. Martien, Senior Thesis, Santa Cruz, 1982
   S. Pope, Senior Thesis, Santa Cruz, 1984
7) J.P. Crutchfield and B.A. Huberman, Phys. Lett. $\underline{77A}$, 407 (1980)
   G. Mayer-Kress and H. Haken, J. Stat. Phys. $\underline{26}$, $\overline{149}$ (1981)
8) S. Grossman and S. Thomae, Z. Naturforsch. $\overline{32a}$, 1353 (1977)
9) I. Shimada and T. Nagashima, Prog. Theo. Phys. $\underline{61}$, 1605 (1979)
10) E.N. Lorenz, Ann. N.Y. Acad. Sci. $\underline{357}$, 282 (198$\overline{0)}$
11) J.P. Crutchfield et al, Phys. Lett. $\overline{76A}$, 1 (1980)
12) R.M. May, Nature $\underline{261}$, 459 (1976)
13) M. Hénon, Comm. Math. Phys. $\underline{50}$, 69 (1976)
14) "Strange Attractors and their Physical Significance",
    J.P. Crutchfield and R. Shaw, videotape, 1978.
15) O. Rössler, Ann. N.Y. Acad. Sci. $\underline{316}$, 376 (1978)
16) E.N. Lorenz, J. Atmos. Sci. $\underline{20}$, 1$\overline{30}$ (1963)
17) D. Ruelle, F. Takens, and T.T.Trés, Comm. Math. Phys $\underline{20}$, 167 (1971)
    D. Ruelle, Ann. N.Y. Acad. Sci. $\underline{316}$, 408 (1978)
18) J.P. Gollub and H.L. Swinney, Phys. Rev. Lett. $\underline{35}$, 927 (1975)
19) "Strange Attractors and their Possible Relation to Turbulence",
    NSF Grant Proposal (Material Science Division) 1980,1981,1982,1983
    "Machine for Simulation of 2-d Partial Differential Equations",
    NSF Grant Proposal (Computer Science Division) 1981
    (copies available on request)
20) L. Boltzmann, personal communication
21) C.E. Shannon and W. Weaver, "The Mathematical Theory of Communication",
    University of Illinois Press, 1962
22) R.S. Shaw, "Strange Attractors, Chaotic Behavior, and Information Flow",
    Santa Cruz, 1978, published in Z. Naturforsch. $\underline{36a}$, 80 (1981)

23) E.T. Jaynes, in "Delaware Seminar in the Foundations of Physics",
        Vol. 1, Springer-Verlag, 1967.
24) L. Boltzmann, 1872. See also Shannon, Ref. 21.
25) A.N. Kolmogorov, Dokl. Akad. Nauk. SSSR 124, 754 (1959)
    Ya. Sinai, Dokl. Akad. Nauk. SSSR 124, 768 (1959)
    A.N. Kolmogorov, Problems of Information Transmission
        1, 3 (1965) and 5, 3 (1969).
26) Y. Oono and M. Osikawa, Prog. Theo. Phys. 64, 54 (1980)
27) J.P. Crutchfield and N.H. Packard, Int. J. of Theo. Phys. 21, 433 (1982)
28) J.D. Farmer, PhD thesis, Santa Cruz (1981)
29) J.D. Farmer, Ann. N.Y. Acad. Sci. 357, 453 (1980)
30) R. Shaw, "Modeling Chaotic Systems", in "Chaos and Order in Nature",
        ed. H. Haken, Springer-Verlag 1981.
31) N.H. Packard et al, Phys. Rev. Lett. 45, 712 (1980)
32) R. Bowen, "Equilibrium States and the Ergodic Theory of
        Anosov Diffeomorphisms", Lect. Notes in Math. 470,
        Springer-Verlag 1975.
    J. Guckenheimer, Inv. Math. 39, 165 (1977)
    N.H. Packard, PhD Thesis, Santa Cruz (1982)
33) J.P. Crutchfield and N.H. Packard, "Symbolic Dynamics
        of Noisy Chaos", in "Order and Chaos" proceedings,
        North-Holland Amsterdam 1983.
34) J. Curry, J. Stat. Phys. 26, 683 (1981)
35) J.D. Farmer, "Sensitive Dependence to Noise Without Sensitive
        Dependence to Initial Conditions", J. Unpub. Results 1, 1 (1983)
36) T. Erber et al, J. Comp. Phys. 49, 394 (1983)
37) S. Ma, J. Stat. Phys. 26, 221 (1981)
38) E. Jen, J.D. Farmer, unpublished
39) L. Wittgenstein, "On Certainty", Harper & Row 1972.
40) C. Leith, "Order in Chaos" Conference, Los Alamos, 1982
41) G.J. Chaitin, Scientific American, May 1975
42) J. Ford, Physics Today, April 1983
43) J. Smith, Byte Magazine, January 1983
44) A.W. Drake, "Fundamentals of Applied Probability Theory",
        McGraw-Hill, 1967.
45) J.D. Farmer, E. Ott, and J.A. Yorke, Physica 7D, 153 (1983)
    J.D. Farmer, Z. Naturforsch. 37a, 1304 (1982)
    P. Grassberger, Phys. Lett. 97A, 227 (1983)
46) P. Grassberger and I. Procaccia, Phys. Rev. Lett. 50, 346 (1983)
    J. Guckenheimer and G. Buzyna, Phys. Rev. Lett. 51, 1438 (1983)
    A. Brandstäter et al, Phys. Rev. Lett. 51, 1442 (1983)
47) If the logarithms are taken before the mean, the measure is
    closest to the "pointwise dimension", otherwise, the measure
    is the "correlation exponent" of Grassberger and Procaccia.
    This change made no difference insofar as the qualitative
    features described here, although numerical values were changed
    by amounts on the order of ten or twenty percent.
48) R.J. Deissler, Phys. Lett. 100A, 451 (1984)
49) R.J. Donnelly et al, Phys. Rev. Lett. 44, 987 (1980)
    G. Ahlers and R.W. Walden, Phys. Rev. Lett. 44, 445 (1980)
    P. Bergé, in "Chaos and Order in Nature", ed. H. Haken,
        Springer-Verlag 1981.
50) D. Ruelle, Phys. Lett. 72A, 81 (1979)