in **Inside versus Outside**, H. Atmanspacher, editor, Series in Synergetics, Springer, Berlin, (1993) 235 - 272.
SFI Technical Report 93-06-035.

Observing Complexity and The Complexity of Observation

James P. Crutchfield

Physics Department
University of California
Berkeley, California 94720

Abstract

The distortions introduced by the measurement process can lead to drastic consequences for an observer's ability to infer structure in its environment. Several examples illustrate the appearance of infinite complexity and irreducible indeterminacy in classical, deterministic processes. Along the way several notions of complexity and an approach to a general solution — hierarchical machine reconstruction — are reviewed.

Proceedings of the International Workshop on *Endo/Exo-Problems in Physics*,
held at Ringberg Castle, Tegernsee, Germany, 29 March to 2 April 1993.

Contents

List of Figures

1 Appearances May Be Deceiving ... But Just How So?

An observer's notion of what is random and what is complex in its environment depends directly on the quality of measurements and the computational resources available for inference. The resources are coarsely measured by the amount of raw data, of memory, and of the time available for model estimation. Although these play a key role, the discovery of structure in an environment depends more directly and subtlely on the effects of measurement distortion.

The goal here is to give an overview of an inductive framework that addresses the problem of measurement distortion for dynamical systems and stochastic processes. A strong emphasis is placed on classifying processes according to their intrinsic computational capability. The main results indicate the overwhelming effect of the simple fact that measurements are only indirect reflections of the "internal" states of a process. In short, a process with a finite number of internal states — i.e. of finite complexity — can appear to an observer to have infinite complexity. The consequences for inferring regularity are clear: the resources required can be unbounded. This results in an *irreducible* uncertainty in observing classical dynamical systems and stochastic processes.

Illustrative examples will be drawn from the onset of unpredictability in nonlinear systems, spatio-temporal pattern recognition, and finite stochastic nondeterministic processes. These examples demonstrate the crucial role of inductive inference not only to scientific discovery — which is obvious — but even in the simplest of ideal cases. The consequence is that the observational theory of classical physics is nontrivial. And this seems to be somewhat at variance with the implicit philosophical stance of several founders of modern physics that many of the then-new properties of quantum systems were intrinsic and did not appear in classical systems. The examples below make one suspicious of any type of blanket statement along the latter lines.

Before the examples are presented, we need to see why a measurement theory of classical processes is necessary. This is the burden of the next section, which points out that this requirement is the result of various types of dynamical instability in nonlinear systems — of which chaos is just one. The section also identifies the driving force — or *intentionality* — of inductive inference in this framework. This is the search for causality. The result is a relatively new view of the information dynamics of nonlinear processes: complexity and unpredictability are distinct and complementary properties. The examples then follow immediately. A final commentary is given in the last section along with some discussion of how to break out of the problems that have been illustrated.

2 Observing Complexity

2.1 Measurement Theory for Classical Processes*

There are three "laws" of information dynamics that emerge in any description of unpredictable systems. The laws are working assumptions that are used implicitly by many in the application of information theory to chaotic dynamical systems. The laws cannot be proven, but

* A portion of this section is excerpted from section 1.4 "Information Dynamics" in [1]. The discussion of the three laws in the following should be compared with that in [2] and [3].

rather summarize experience and represent useful fundamental concepts. They are hypotheses to be validated and modified in their application to interpreting experiments. I will list them first as a group and then discuss each in turn, mentioning some analogies with equilibrium thermodynamics.

1. Total information is conserved.
2. An observer's information about the state of a chaotic system can only decrease, without additional measurement.
3. An observer cannot obtain infinite information about a system's state.

In the first law total information refers to the entire observer-experiment system. The observer and the system-under-study are subsystems of a larger, encompassing system. This larger system can be called the universe of discourse. (See Figure 1). For the present purposes an observer is a subsystem that (i) interacts with a system-under-study (its environment), (ii) has internal states that are correlated with the system-under-study's instantaneous states, and (iii) has internal states that are correlated with the system-under-study's future states. The latter condition says that the observer attempts to predict the system-under-study's behavior. The second and third conditions will be construed to mean that an observer attempts to model the system-under-study. Conservation within the universe refers to the equality of the amount of information obtained by the observer and lost by the system-under-study during measurement. This section assumes a (Heisenberg) bidirectional flow of information between the observer and the system-under-study during measurement. The remainder of the discussion, however, restricts the interaction to (Einstein) unidirectional flow from the system-under-study to the observer. Finally, as in information theory proper, information content is a basic undefined notion.* Ways to quantify it will be given shortly.

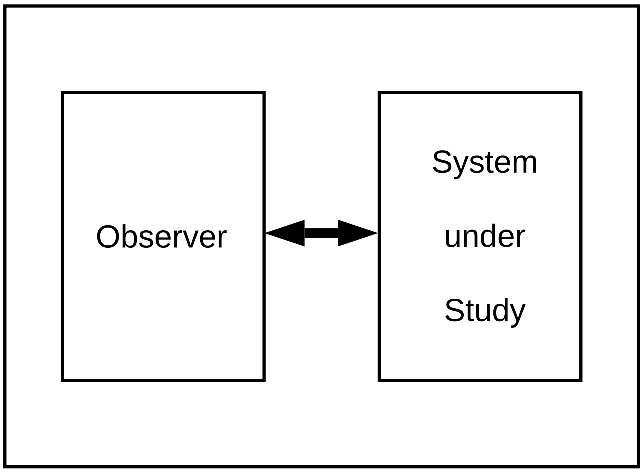

Figure 1 The universe of discourse for information dynamics. (After [1].) The double-headed arrow between the system-under-study and the observer indicates the (Heisenberg) bidirectional flow of information during measurement. In the latter portions of the discussion the interaction will be restricted to (Einstein) unidirectional measurement interaction in which information flows only from the system-under-study to the observer.

* While notable, this is not the least bit unusual. Energy in physics has a similar ontological status.

The total information context is the universe. If information appears not to be conserved, then the boundaries of the universe are not sufficiently large. Thus, like the first law of thermodynamics, information conservation can be made to hold by fiat.

This apparently simple point needs some elaboration. One need not appeal to the first law of thermodynamics — that energy is conserved — and its elevation of energy to a primary concept for all physical systems. Indeed, there are processes for which energy either is not defined or simply is not a primary descriptive concept. Dissipative dynamical systems or, more generally, open systems come to mind. Generally, only information and measurement need be defined. In this framework energy is a derivative concept, appropriate to a subclass of processes. For these it is based on and verified as a property of the information collected by measurement. For a much broader range of processes there is still a first law — conservation of information — for the universe. In some ways it is really not much more than a different way of saying that a probability distribution is normalized.* And ultimately, like energy conservation, it is a principle imposed by us and our formalisms on the world: if violated we invent new concepts — interactions, forces, particles, complexities, what have you — to reestablish its correctness. This is how “physical principles”, like energy conservation, differ from “physical laws”, like the ideal gas law. So, perhaps, one should refer to the above three “laws” as “principles”.

The second law reflects an observer’s inability to accurately predict the future evolution of a chaotic system. A measurement reveals that the system-under-study is in some small region of its state space. As long as there is any error in determining the exact state, however, the observer will be unable to predict the chaotic system’s behavior beyond some finite time, due to the chaotic dynamics. The measurement uncertainty means that any model for prediction cannot be put into a state identical to the system-under-study’s. Even assuming that the model completely describes the system’s dynamics, the deviation of the model’s behavior from the system’s will grow exponentially. When the deviation is as large as the system’s attractor (say) the behavior is no longer predictable.

The second law is analogous to that in thermodynamics which says the entropy of an isolated system increases to a maximum. The thermodynamic entropy can be formulated in terms of “missing information” about the state of a thermodynamic system once a macroscopic state — fixed by the volume, the total energy, and so on — is specified. According to Boltzmann, thermodynamic entropy is the logarithm of the number of microscopic states giving rise to the same value of the macroscopic state function. An observer’s measurement of the system-under-study is a macroscopic act that isolates the system to within some set of indistinguishable microscopic states. Boltzmann’s idea then is that the thermodynamic entropy is proportional to the information *obtained via measurement* that indicates the system is in a given macroscopic state.

Preparing a thermodynamic system in a small ensemble of microscopic states with fixed energy (say) corresponds to having low entropy, or a large amount of information about the system’s macroscopic state. Subject to energy conservation and assuming the microscopic dynamics is in “molecular chaos”, however, the system evolves toward equilibrium visiting larger regions of state space than initially. It appears more disordered, the entropy increases,

* It can be argued that measurement information is more basic than probability.[4]

and the observer’s initial information can be used with decreasing effectiveness to determine the system’s current microscopic state.

From information conservation, the system produces information that “displaces” the observer’s. As time goes on, the observer is less able to predict the system’s microscopic state. The observer “loses” information until the next measurement is made.

The third law is equivalent to the impossibility of the observer making infinitely precise measurements about a system’s state. It is analogous to the third law of thermodynamics that says a system cannot be prepared in states of zero entropy, such as zero temperature. Zero entropy corresponds to perfect knowledge of a system’s state. There is no missing information because an infinitely precise measurement yields a complete determination of the system’s state.

Chaotic systems, although purely classical, obey the above third law because they exponentially amplify fluctuations.[*] The third law applies to the observer’s measurements of a chaotic system as the latter continually receives information from other parts of the universe, *including* the observer. Consequently, the state of the observer must be included in the classical complete determination of the system-under-study.

An estimate of the effect of external fluctuations on chaotic behavior will illustrate their importance when observing chaotic systems. This will provide a quantitative motivation for including the observer and the rest of the universe in a complete (arbitrarily precise) state-determination of a chaotic system. Consider the gravitational effect of an electron at the “edge” of the known universe (˜17 billion light years) on a terrestrial game of billiards. Assume, for simplicity, that during a given shot the game is energy conserving over half an hour and that the balls are hit sufficiently hard to cause a few collisions each second. The unpredictability of the billiards’ state can be conservatively estimated as an information loss rate of approximately 1 bit per second. The uncertainty caused by the existence or nonexistence of the electron at the edge of the universe leads to total unpredictability in about six minutes. An electron at the edge of the solar system does so in four minutes, and if they move around, the billiard players do so in about one half minute. Similarly, an erratic dripping kitchen faucet[6] becomes unpredictable in less than half a minute if the uncertainties are produced by the neighbor’s cat prowling in the garden. Such is the sensitivity of chaos to external influences. This comes about by its exponential amplification of uncertainty, noise, and error. These examples demonstrate the necessity of considering the observer in an arbitrarily accurate determination of a chaotic system’s state. Not only will the occasional measurement perturb the system-under-study, but the system is constantly bathed in and amplifies information from any couplings — gravitational or otherwise — to the observer.

The observer’s attempt at a complete determination of the system-under-study necessarily leads to an attempt at determination of the entire universe of which the observer is part. The observer must observe and measure itself, since the exact state of the system-under-study includes observer-state information. This produces an infinite regression of necessary measurements and requires the storage of information of the universe’s state within a subsystem of it. The process of storage itself involves the measurement and manipulation of further subsystems’ states. The infinite regression thus requires the storage of an infinite amount of information. Regardless

[*] The “fluctuations” are to be thought of as coming not only from extrinsic influences, but also from the deterministic dynamics.[5]

of the size of the universe, this self-observation and internal self-coding is impossible. One concludes that arbitrarily precise measurements of the chaotic system-under-study's state are not possible. The third law of information dynamics seems forced upon us by the sensitivity of chaos to external information sources.

This result and its argument remind one, somewhat loosely, of the self-referential paradoxes associated with Gödel's incompleteness theorem.[7] Godel's results state that in a formal system of sufficient structure there are true statements expressible in the system's syntax that are not provable using the system's rules. The formal system cannot "know" everything implicit in its axioms and rules of deduction. In the present setting, the analog of the "formal system with sufficient structure" is the universe of discourse consisting of an observer that models the system-under-study. The act of state determination plays the role of "true statements". That chaos requires the observer to determine the universe's state is analogous to Gödel's construction of a self-referential statement about the formal system. And the incompleteness result itself appears here as the conclusion that the universe's state cannot be stored within a subsystem of the universe. Indeed, this last statement is more similar to Turing's use of the diagonal construction from Cantor's theory of infinite sets to demonstrate the existence of uncomputable real numbers — i.e. there are entities that exist outside the algorithmic framework.[8] In this way the third law of information dynamics represents a type of "measurement" incompleteness or "observational" uncomputability for chaotic dynamics.

The conclusion drawn from these points is that chaos, and many other types of instability for that matter, force one to consider at the outset the full universe of discourse. As noted, this is due to the basic nature of chaotic behavior: exponential amplification of errors requires a full accounting for the flow of information. It has been pointed out that chaos leads via Kolmogorov's notion of complexity to infinite (universal Turing machine) models, if one requires Laplacian determinism.[9] But this is a trivial extreme case that turns on notions of infinite precision and computational capabilities. In practice, of course, available precision and computational resources are finite. The consequence of this finiteness is that one must make an explicit model of the measurement process for classical (nonlinear) physics. In this the consequences of chaotic behavior for general physical theory are qualitatively different from periodic and purely stochastic behavior — behavior types admitted decades ago as "physics". Accounting for the measurement process, in turn, introduces the study of the subjectivity and model representation dependence inherent in an observer's ability to discover and recognize structure in the system-under-study; this will be the subject of later sections. Chaos thus motivates the study of how measurement distortion complicates the inference of structure.

2.2 Endo contra Exo

In this light, my simple summary of the "endo-exo problem for physics", or at least that classical part which is addressed below,* is the following. The physical properties of the system-under-study that an observer can distinguish differ from those accessible outside the universe of discourse. Endophysics is the collection of theories and facts the observer develops to predict

* There is no direct attempt, for example, to see quantum mechanics in classical physics as suggested by Rössler.[10] But I find the results from the measurement distortion examples highly suggestive. For direct consideration of the quantum context see [11] and [12].

and model the system-under-study. In contrast, exophysics describes the universe of discourse in its entirety. Restated in these terms the working hypothesis then is that endophysics differs from exophysics.

The question is, then, does this hypothesis have any (i) mathematical and (ii) empirical consequences? I believe the answer is affirmative in both cases. So how might we study this?

The following gives several concrete examples of how the interaction of an observer and the system-under-study can lead to significant limitations on what the observer can infer from experimentation. The interaction is called a measurement; the effective dynamics of the interaction is an instrument. I see no way in principle to distinguish in a physical theory the notions in each of these pairwise identifications. The usual distinction is based on measurement (say) having something to do with the intentionality[13] of an observer; namely, the observer intends to model the system-under-study. And this is tantamount to invoking some sort of "intelligence" to describe the capabilities of one part of the universe of discourse. But the point is that, as far as we know at this time, there is no physical basis for identifying intelligent subsystems. Thus, if system S_1 "interacts" with system S_2, then S_1 "measures" some aspect of S_2's state. Information is transferred, possibly both ways. The time-dependent manner in which the information is transferred determines the effective "instrumentation". Until we can define and detect intelligent subsystems in physical terms, we are left with these identifications.

By no means do the examples encompass the entire problem. In fact, one major simplification in all of the following is that during measurement information flows only from the system-under-study to the observer; and not the other way around. Nonetheless, I fail to see how concrete progress on the endo-exo problem can ignore the difficulties indicated by the examples.

2.3 The Search for Causality

The axiom of this entire approach — and perhaps the very explanability of nature in terms of classical physics — is the search for causal states. By this I mean the following. An observer notes at some time that the system-under-study appears to be in some configuration $\mathbf{A}$ and at some (say) later time it is seen to be in configuration $\mathbf{B}$. Then some elements of $\mathbf{A}$ "caused" some elements of $\mathbf{B}$ to occur if and only if varying elements of $\mathbf{A}$ would have led to different elements of $\mathbf{B}$ occurring. The remainder of this section gives this notion of causality a concrete form.

Over the last decade or so there has been a good deal of effort expended to understand how this notion of causality can be formalized and, perhaps more importantly, implemented for measurements of a chaotic classical process. The key notion is that a causal state renders the future conditionally independent of the past. In other words, if the observer knows the system-under-study's current causal state, the observer needs no other information from the past to determine the range of future behavior. In this way, causal states summarize or compress the (possibly infinite) past. But how can this idea be reduced to practice?

The answer to this turns on a generalization of the "reconstructed states" introduced, under the assumption that the process is a continuous-state dynamical system, by Packard *et al.*[14] The contention there was that a single time series necessarily contained all of the information about the dynamics of that time series. The notion of reconstructed state was based on Poincaré's view of the intrinsic dimension of an object.[15] This was defined as the largest number of successive cuts through the object resulting in isolated points. A spherical shell in three dimensions by his method is two dimensional since the first cut typically results in a circle and then a second cut, of that circle, isolates two points. One way Packard *et al.* implemented this used probability distributions conditioned on values of the time series' derivatives. That is, the coordinates of the reconstructed state space were taken to be successive time derivatives and the cuts were specified by setting their values. This was, in fact, an implementation of the differential geometric view of the derivatives as locally spanning the graph of the dynamic.

In this reconstruction procedure a state of the underlying process is identified by increasing the number of conditioning variables, employing successively higher derivatives, until the conditional probability distribution peaks. It was noted shortly thereafter that in the presence of extrinsic noise a number of conditions is reached beyond which the conditional distribution is no longer sharpened. And, as a result, the process's state cannot be further identified. The width of the resulting distribution then gives an estimate of the effective extrinsic noise level and so also an estimate of the maximum amount of information contained in observable states. The minimum number of conditions first leading to this situation is an estimate of the effective dimension.[16]

The method of time-derivative reconstruction gives the key to discovering causal states in a more general setting. It is important first to note that there is a basic flaw in the original formulations of reconstruction. If information important in determining the observed behavior is not "contained" in the time series, then there is an irreducible amount of apparent randomness for the observer. (Explicit examples will be given in the next section.) This reflects dynamics that is not reconstructible. The failure in the first proposals is that they did not properly formalize the effect of extrinsic information, or noise, on the reconstruction process. The problem is unavoidable, naturally. In applications it was typically dealt with in an *ad hoc* manner and not seen as the fundamental and *prior* issue that it is. In the context of continuous time series the problem is addressed systematically with model order estimation methods which balance deterministic structure and apparent randomness.[17] But even this approach is only a partial solution. It confuses several basic difficulties in modeling which can be avoided by recasting reconstruction as the search for causal states, as done here.

To see how the generalization goes, let's restrict consideration to discrete-valued time series. If one is interested in describing continuum-state systems, then this move should be seen as purely pragmatic: an instrument will have some finite accuracy, generically denoted ϵ, and individual measurements, denoted s, will range over an alphabet $\mathcal{A} = \left\{0, 1, 2, \ldots, \epsilon^{-1} - 1\right\}$. For discrete time series a causal state is defined to be the set of subsequences that render the future conditionally independent of the past. Thus, the observer identifies a state at different times in a data stream as being in identical conditions of ignorance about the future.[18] (See Figure 2 for a schematic illustration that ignores probabilistic aspects.)

Now we can begin to formalize the notion of causal state. Consider two parts of a data stream $\mathrm{s} = \ldots s_{-2} s_{-1} s_0 s_1 s_2 \ldots$. The one-sided forward sequence $\mathrm{s}_t^{\rightarrow} = s_t s_{t+1} s_{t+2} s_{t+3} \ldots$ and one-sided reverse sequence $\mathrm{s}_t^{\leftarrow} = \ldots s_{t-3} s_{t-2} s_{t-1} s_t$ are obtained from s by splitting it at time t into the forward- and reverse-time semi-infinite subsequences. They represent the information about the future and past, respectively. Consider the joint distribution of possible forward sequences $\{\mathrm{s}^{\rightarrow}\}$ and reverse sequences $\{\mathrm{s}^{\leftarrow}\}$ over all times t

$$\Pr(\mathrm{s}) = \Pr(\mathrm{s}^{\rightarrow}, \mathrm{s}^{\leftarrow}) = \Pr(\mathrm{s}^{\rightarrow} | \mathrm{s}^{\leftarrow}) \Pr(\mathrm{s}^{\leftarrow}) \tag{1}$$

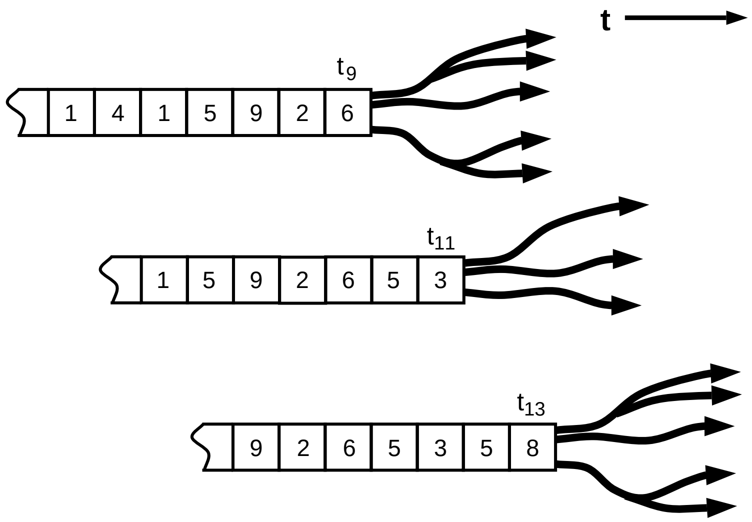

Figure 2 Within a single data stream morph-equivalence induces conditionally-independent states. When the template of future possibilities, i.e. allowed future subsequences and their past-conditioned probabilities, has the same structure then the process is in the same causal state. At t_9 and at t_{13}, the process is in the same causal state; at t_{11} it is in a different causal state. The figure only illustrates the nonprobabilistic aspects of morph-equivalence. (After [19].)

The conditional distribution $\mathrm{Pr}(\mathrm{s}^{\rightarrow}|\omega)$ is to be understood as a function over all possible forward sequences $\{\mathrm{s}^{\rightarrow}\}$ that can follow the particular sequence ω where ever it occurs in s.

Then the same causal state $S \in \mathbf{S}$ is associated with all those times $t, t' \in \{t_{i_1}, t_{i_2}, t_{i_3} \ldots : i_k \in \mathbf{Z}\}$ such that past-conditioned future distributions are the same. That is,

$$t \sim t' \text{ if and only if } \mathrm{Pr}(\mathrm{s}^{\rightarrow}|\mathrm{s}_t^{\leftarrow}) = \mathrm{Pr}(\mathrm{s}^{\rightarrow}|\mathrm{s}_{t'}^{\leftarrow}) \tag{2}$$

If the process generating the data stream is ergodic, then there are several comments that serve to clarify how this relation defines causal states. First, the sequences $\mathrm{s}_t^{\leftarrow}$ and $\mathrm{s}_{t'}^{\leftarrow}$ are typically distinct. If $t \sim t'$, Eq. (2) means that upon having seen different histories one can be, nonetheless, in the same state of knowledge or ignorance about what will happen in the future. Second, $\mathrm{s}_t^{\leftarrow}$ and $\mathrm{s}_{t'}^{\leftarrow}$, when considered as particular symbol sequences, will each occur in s many times other than t and t', respectively. Finally, the conditional distributions $\mathrm{Pr}(\mathrm{s}^{\rightarrow}|\mathrm{s}_t^{\leftarrow})$ and $\mathrm{Pr}(\mathrm{s}^{\rightarrow}|\mathrm{s}_{t'}^{\leftarrow})$ are functions over a nontrivial range of "follower" sequences $\mathrm{s}^{\rightarrow}$.

This gives a formal definition to the set $\mathbf{S}$ of causal states as equivalence classes of future predictability: $\sim$ is the underlying equivalence relation that partitions temporal shifts of the data stream into equivalence classes. In the following the states will be taken simply as the labels for those classes. This does more than simplify the discussion. As integers ranging over $\{0, 1, 2, \ldots, \|\mathbf{S}\| - 1\}$, the states convey all of the required information to render the future conditionally independent of the past. For a given state S the set of future sequences $\{\mathrm{s}_S^{\rightarrow} : S \in \mathbf{S}\}$ that can be observed from it is called its future morph. (Recall Fig. 2.) The set of sequences that lead to S is called its past morph. Note that the state and its morphs are the contexts in which an individual measurement takes on semantic content. Each measurement is anticipated or "understood" by the observer *vis á vis* its model and, in particular, the structure of the states.[20]

Once the causal states are found, the temporal evolution of the process — its symbolic dynamic — is given by a mapping T from states to states $T : \mathbf{S} \rightarrow \mathbf{S}$; that is, $S_{t+1} = TS_t$. The pair $\mathrm{M} = (\mathbf{S}, T)$ is referred to as an ϵ-machine; where ϵ simply reminds us that what we have reconstructed is an approximation and depends on the measuring instrument's characteristics — such as its resolution. The procedure that begins with a data stream and estimates the number of states and their transition structure and probabilities is referred to as ϵ-machine reconstruction.[18]

There are a few points that must be brought out concerning what these reconstructed machines represent. First, by the definition of future-equivalent states, the machines give the minimal information dependency between the morphs. It is in this respect that they represent the causal structure of the morphs considered as events. The machines capture the information flow within the given data stream. If state $\mathbf{B}$ follows state $\mathbf{A}$ then $\mathbf{A}$ is a cause of $\mathbf{B}$ and $\mathbf{B}$ is one effect of $\mathbf{A}$. Second, machine reconstruction produces minimal models up to the given prediction error level. This minimality guarantees that there are no other events (morphs) that intervene, at the given error level, to render $\mathbf{A}$ and $\mathbf{B}$ independent. In this case, we say that information flows from $\mathbf{A}$ to $\mathbf{B}$. The amount of information that flows is the negative logarithm of the connecting transition probability: $-\log_2 p_{\mathbf{A}\rightarrow\mathbf{B}}$. Third, time is the natural ordering captured by machines.[*] Finally, anticipating the fuller definition given later, an ϵ-machine for a process is the minimal causal representation reconstructed using the least powerful computational model class that yields a finite complexity. The motivations for this more elaborate definition will become clearer only after the role of representation is appreciated.

2.4 Prediction or Modeling?

Similar notions of state can be found in many literatures, such as linear stochastic processes,[21] symbolic dynamics,[22] ergodic theory,[23] automata theory,[24], statistical mechanics,[25] and others. I would simply note that my own background — which is strictly irrelevant to the main points, but does inform the discussion — derives from an attempt to understand the puzzle of deterministic chaos as a physical phenomenon.

Examples of the search for causal states are quite numerous and have addressed both temporal[14,17,18,26] and spatio-temporal[17,27–29] "chaotic" processes. The list could be extended quite a bit if one included nonlinear modeling[30] and artificial neural networks.[31] However, in these endeavors there typically is a strong emphasis on statistical parameter estimation within a fixed-size model class, with a corresponding lack of effort in discovering the intrinsic computational structure of processes. Indeed, the notion of causality and the meaning of any measure of a process's complexity requires the search for the smallest model consistent with the data.

These comments bring us to the fundamental differences between prediction and modeling. The distinctions are rather clearly drawn in computational learning theory.[32] But, roughly, the difference is that in prediction the goal is to produce the best guess of future behavior, *by any means whatsoever*; in contrast with modeling the goal is to learn something about the process's structure. Naturally, prediction is aided by means of a good model; but typically

[*] In [20] it is shown that temporal asymmetry can be detected: machines reconstructed in different "time" directions can have different numbers of causal states.

efforts at prediction allow for any sort of representation, as long as it gives good forecasts. And so, for a given process there may be good predictors — such as historical look up tables — that indicate little, if anything, about the process's causal structure. Modeling demands much more and, if successful, it provides much more; certainly more than just good forecasts. The dichotomy, as drawn here, is that modeling is the search for causality, and prediction is the search for determinism.

2.5 Unpredictability versus Complexity

With the modeling methodology laid out, several statistics can be defined that capture how information is generated and processed. A useful coordinate-independent measure of unpredictability is Shannon's entropy rate.[33] If one already knows the process's distribution $\Pr(\omega)$ over infinite sequences ω, then the entropy rate is defined

$$h_\mu = \lim_{L\to\infty} \frac{H\big(\Pr(s^L)\big)}{L} \tag{3}$$

in which $\Pr\big(s^L\big)$ is the marginal distribution, obtained from $\Pr(\omega)$, over the set of length L sequences s^L and H is the average of the self-information, $-\log_2 \Pr\big(s^L\big)$, over $\Pr\big(s^L\big)$. In simple terms, h_μ measures the rate at which the process appears to produce information. Its units are bits per symbol. The higher the entropy rate, the more information produced, and the more unpredictable the process appears to be.

Unfortunately, but not surprisingly, in many situations one does not know $\Pr(\omega)$ and so the definition in Eq. (3) is not directly applicable. A simple rewriting of it will show the important role played by causal states in computing the entropy. The form of Eq. (3) indicates that the entropy rate is the slope of the curve $H(L) = H\big(\Pr(s^L)\big)$. This form, it turns out, is not a particularly good estimator of the slope. The two-point slope definition

$$h_\mu = \lim_{L\to\infty} \{H(L) - H(L-1)\} \tag{4}$$

often converges more quickly. This is equivalent to the conditional entropy form

$$h_\mu = \lim_{L\to\infty} H\Big(\Pr\Big(s|s^{L-1}\Big)\Big) \tag{5}$$

in which $\Pr\big(s|s^{L-1}\big)$ is the conditional distribution of the next symbol s given the past length $L-1$ sequences and H averages $-\log_2 \Pr\big(s|s^{L-1}\big)$ over $\Pr\big(s^L\big)$. Assuming that we have a "typical" data stream s and that the process is ergodic, the entropy becomes

$$h_\mu = H(\Pr(s_{t+1}|\mathrm{s}_t^{\leftarrow})) \tag{6}$$

where $\Pr(s_{t+1}|\mathrm{s}_t^{\leftarrow})$ is the conditional distribution of the next symbol s_{t+1} given the semi-infinite past $\mathrm{s}_t^{\leftarrow}$ and H averages the conditional distribution over $\Pr(s^{\leftarrow})$.* Now, if we know the set of

* This step is somewhat subtle. The L-limit has been replaced by an infinite past; which is no problem. The difficulty comes in realizing that there can be large sets of sequences that remain transient with respect to the asymptotic measure. This is discussed under the heading of "synchronization" in Ref. [20], for example. It is why the restrictions to ergodic processes and "typical" sequences are introduced. See [5] for a more detailed discussion of typicality.

causal states $\mathbf{S}$ and find the one to which $\mathrm{s}_t^{\leftarrow}$ leads, then we can greatly simplify the definition. It becomes

$$h_\mu = H(\Pr(s|S)) \tag{7}$$

in which $\Pr(s|S)$ is the conditional distribution of the next symbol s given the current state $S \in \mathbf{S}$. Aside from casting off the limits in the definition, this new form is a much more compact, and sometimes closed-form, expression for the entropy rate.

It is to be expected then that many methods for exactly computing a process's information production rate require the reduction to some sort of causal representation.[*] By the preceding definition we see that causal states are the only way to get the correct conditional probabilities — and the latter is what is required for entropy in Eq. (7). Perhaps this seems to be something of a tautology. One might object that this is just an artifact of Shannon's notion of entropy. But I know of no other (qualitatively different) measure of unpredictability that is coordinate independent and is not made more efficacious by some notion of causal state. Renyi's generalization of Shannon's entropy, for example, is similarly improved using causal states.[5]

Interestingly, this reformulation also applies to the calculation of the entropy rate for continuous-state dynamical systems. Often, the most direct means for this class of processes is to partition the continuous state space into a finite set of elements that can be associated with a Markov chain or Markov process. Then with an estimate of the probability of these Markovian states and their transition probabilities, the entropy rate follows directly from Eq. (7).[†]

Another useful, and closely related, measure of the range of behavior is the topological entropy h, which simply looks at the growth rate of the total number $N(L)$ of sequences with increasing length L

$$h = \lim_{L \to \infty} \frac{\log_2 N(L)}{L} \tag{8}$$

It follows from Eq. (3) that $h \geq h_\mu$ and that if $\Pr\big(s^L\big)$ is constant over those s^L which occur, then $h = h_\mu$.

Thinking about quantifying unpredictability in these ways suggests there are other, and perhaps more immediate, measures of a process's structure: the complexities. A process's topological complexity C_0 is simply given in terms of the minimal number of causal states

$$C_0 = \log_2 \|\mathbf{S}\| \tag{9}$$

It is an upper bound on the amount of information needed to specify which state a process is in. Following the complementary relation between the entropy rate and the topological entropy, there is a probabilistic version of the "counting" topological complexity. It is formulated as

[*] Indeed, I like to think of each and every algorithm or analytic method for calculating the entropy as effectively defining, perhaps indirectly, a notion of state. Shannon suggested a particularly clever way of estimating the entropy rate of English without recourse to an explicit notion of state.[34] This empirical method used human subjects as the "conditioning" agents; thereby employing the human ability to recognize appropriate conditionally-independent contexts for predicting printed texts.

[†] If one knows the equations of motion, unlike our observer, then there is another approach to calculating the entropy rate as the sum of the positive Lyapunov exponents.[35–37]

follows. The $\|\mathbf{S}\| \times \|\mathbf{S}\|$ transition probability matrix T determines the asymptotic causal state probabilities as its left eigenvector

$$p_{\mathbf{S}} T = p_{\mathbf{S}} \tag{10}$$

in which $p_{\mathbf{S}}$ is normalized in probability: $\sum_{S \in \mathbf{S}} p_S = 1$. From this we have an informational quantity for the machine's size

$$C_\mu = H(p_{\mathbf{S}}) \tag{11}$$

It is called the statistical complexity. If, as provided by machine reconstruction, the machine is minimal, then C_μ is the amount of memory in bits contained in the process.[18]

3 The Complexity of Observing ...

Up to this point the framework of the endo-exo distinction, the basic motivations for a classical measurement theory, and a few quantifiers have been presented. This section gets down to practicalities — What happens to the internal information if an observer does not have direct access to the internal states of the system-under-study? What if the observer has selected a representation that does not match the system-under-study's internal information-processing architecture?

The examples in this section all concern effects induced by the observer viewing the internal structure through some instrument. They show just what can happen and how the consequences force one to consider the general question of mapping between representations — one "inside" the universe of discourse and the other "outside". The inside, or endophysical, view is that of the observer. The representation of interest is the observer's model of the system-under-study. The outside, or exophysical, view is that of omniscience, that is, of the universe of discourse's designer. The exophysical representation is complete by definition. In particular, it contains the exact equations of motion for the system-under-study. The central concern is the difference between the endo- and exophysical representations of the system-under-study, since the latter is the (possibly) common element. The system-under-study itself is a process that can be categorized by the computational class — e.g. stochastic finite-memory processes — required to produce its behavior. Before learning and modeling can begin, though, the observer itself must adopt some model class. This is a range of models from which, given a data stream, it must choose the most appropriate. Naturally, the selected class brings its own descriptional capabilities and limitations to the modeling task. In the spirit of the preceding sections, the main assumption made in the following is that the observer is constrained to causal model classes. In effect, the examples are a study of mapping between endo- and exocomputational complexity classes. And the questions posed above become questions about how the complexity changes, both quantitatively and qualitatively.

To appreciate the approach and its consequences we can delve a little further into these considerations by drawing parallels. Recall that the goal here is for the observer to infer how much of the structure in a data stream can be ascribed to the observer's selected representation, or model class. The selection of a model class induces a set of equivalences in the space of processes. If, for example, one represents binary sources only in terms of the frequency of 0s and 1s, then a uniformly random source and the periodic source producing $\ldots 010101 \ldots$ are indistinguishable. More generally, if the observer assumes a model class that accounts for periodicity and for ideal randomness, then it will be able to finitely represent processes that consist of various combinations of those elementary ones. This decomposes the space of all processes into equivalence classes that only distinguish these components. A quantity, like entropy or statistical complexity, that is constant in each equivalence class is said to be an invariant of the modeling decomposition.

Of course, there are circumstances in which the model class is inadequate; where a more descriptive representation must be used. When infinite representations appear necessary, for example, they hint that the model class should be augmented. And this, in turn, will refine the decomposition and lead to new invariants. A later section addresses this problem directly.

The overall picture is simple: the observer is trying to map the structure of an unknown process onto that describable in terms of its selected representation. An analogous, but restricted type of cross-class representation is also pursued formally in ergodic and computation theories by showing how particular objects — stationary processes or computational tasks — can be mapped onto one another. The motivations there are that the structure of the induced decomposition defines the equivalence concept and, furthermore, the latter can be quantified by an invariant. A classic problem in ergodic theory has been to identify those systems that are isomorphic to Bernoulli processes, which are idealizations of randomness. The associated invariant used for this is the metric entropy, introduced into dynamical systems theory by Kolmogorov[38] and Sinai[39] from Shannon's information theory, i.e. Eq. (3). It turns out that two Bernoulli processes are equivalent if they have the same entropy and satisfy a few other natural constraints.[40] Similarly, in computation theory there has been a continuing effort to establish an equivalence between various hard-to-solve, but easily-verified computational tasks. This is the class of nondeterministic polynomial (NP) problems. If one can guess an answer, it can be determined to be correct or not in polynomial time. The equivalence between NP problems requires that within a polynomial number of computational steps a problem can be reduced to some hardest problem. These hardest problems are called NP-complete.[41] The invariant of this polynomial-time reduction equivalence is a coarsened version of the growth rate, as a function of problem size, of the computation required to solve the problem.

These parallels simply serve to illustrate the venerable tradition of "mapping between classes". The implied methodology is that the fiducial representation is more familiar and so has more semantic content. With a mapping established, one "understands" characteristics of the original problem in terms of their appearance in the fiducial representation.

Inference for an observer is no different in this regard. The strategy in the following is to give examples of the endo-exo problem as a mismatch between the system-under-study's exocomplexity — its intrinsic complexity class — and its endocomplexity — its apparent causal complexity class. The examples come from three rather different disciplines: stochastic finite-memory processes, deterministic spatial automata, and continuum-state dynamical systems. It turns out that, even if an entropy-rate-like quantity is an invariant of the measurement process, the endocomplexity can diverge. At the very minimum it becomes clear that one cannot associate information production, and especially information conservation, solely with Shannon's entropy rate: complexity plays a key role.

3.1 ... Indeterminism

The first example illustrates how a measuring instrument can introduce indeterminism into the observation of a very simple, purely temporal process. The indeterminism vastly increases the apparent complexity. The example following this one then shows how the same phenomenon extends to spatio-temporal processes. The problem there is greatly exacerbated; and so the purely temporal example is something of a prerequisite. In computational terms, this section studies the cost of inferring a causal model of a nondeterministic finite-memory process. The observer assumes the process can be represented with a less powerful class of causal models: stochastic deterministic finite automata. The basic ideas and terminology will be introduced by example and the results summarized in terms of "mapping between classes".

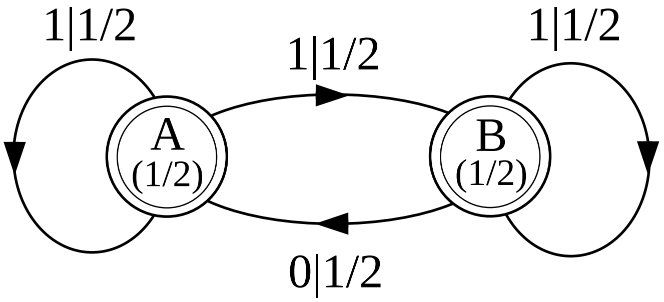

Figure 3 The source is a stochastic nondeterministic finite automaton — a class sometimes referred to as hidden Markov models. The hidden process consists of two states $\{\mathbf{A}, \mathbf{B}\}$ and uniform branching between them — denoted by the fractions p on the edge labels $s|p$. The observer does not have access to the internal state sequences, but instead views the process through the symbols s on the edge labels $s|p$. The inscribed circle in each state indicates that both states are start states. The fractions in parentheses give their asymptotic probabilities, which also will be taken as their initial probabilities.

The system-under-study is the two-state stochastic process shown in Figure 3. There are two internal states $\{\mathbf{A}, \mathbf{B}\}$. Transitions between them are indicated with labeled, directed edges. The labels $s|p$ give the probability p of taking the transition. When the transition is taken the observer receives the measurement symbol $s \in \{0, 1\}$. The association of these symbols with the transitions constitutes the instrument through which the observer views the internal state dynamics. The observer assumes it has no knowledge of the start state and so the process could have started in either $\mathbf{A}$ or $\mathbf{B}$ with equal likelihood.

Figure 4 shows the minimal machine for the process's internal state dynamics. It is the single state Bernoulli process $\mathrm{B}\left(\frac{1}{2}, \frac{1}{2}\right)$ — a fair coin. From Eqs. (7) and (10) it is evident that the metric entropy is $h_\mu = 1$ bit per symbol, as is the topological entropy h. From Eqs. (9), (10), and (11) both the topological complexity and statistical complexities are zero. It is a very random, but simple process.

The goal, of course, is for the observer, using as long a $\{0, 1\}$ data stream as is necessary, to learn the causal structure of this simple process. It has no knowledge of Figure 3, for example. The overall inference procedure is best illustrated in two steps. The first is learning a model of the "topological" process that produces the set of sequences in the data stream, ignoring the

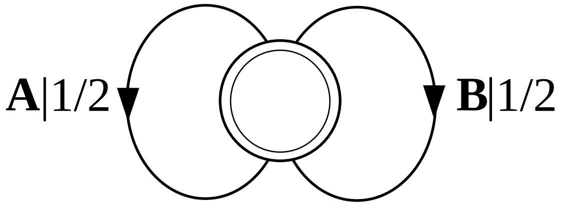

Figure 4 The minimal machine for Figure 3's internal state process. It has a single state and equal branching probabilities. The topological and statistical complexities are zero and the topological and metric entropies are 1 bit per state symbol — a highly unpredictable, but low complexity process. That this is the correct minimal description of the internal state process follows directly from using machine reconstruction, assuming direct access to the internal state sequences $\mathbf{ABABBA}\ldots$. All state sequences are allowed and those of equal length have the same probability.

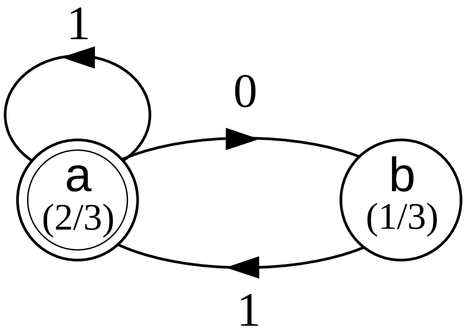

Figure 5 The process's topological structure is given by a deterministic finite automaton — the golden mean machine. The only rule defining the sequences is "no consecutive 0s". The number of sequences of length L is given by the Fibonacci number F_{L+2}; the growth rate or topological entropy h, by the golden mean $\phi = \frac{1}{2}(1 + \sqrt{5})$: $h = \log_2 \phi$. The numbers in parentheses give the states' asymptotic probabilities.

probabilities with which they occur. The second step is to also learn a model that gives the sequences' probabilities.

The first step is relatively straightforward and can be explained briefly in words. (Though, I highly recommend the exercise of finding the morphs via Eq. (2) and their transition structure.) Inspection of the stochastic automaton's output symbols in Figure 3 shows that if $s = 0$ is observed, then $s = 1$ must follow. Further reflection shows that this is the only restriction: consecutive 0s are not produced. All other binary sequences occur.

The automaton, again "topological", that captures this property is shown in Figure 5. This automaton is also what machine reconstruction generates. (It is the answer to the topological portion of the reconstruction exercise.) There are several things to notice. First, the state **a** has a circle inscribed in it. This denotes that **a** is the start state; and it happens to be the unique start state. The reconstructed machine has removed the first element of non-causality in the original process: ignorance of the start state. Second, the automaton is deterministic — a term used here as it is in formal language theory and which does *not* refer to probabilistic elements. Determinism means that from each state a symbol selects a unique successor state. Note that the original process (Figure 3) with its measurement labeling is not deterministic. If the process happens to be in state $\mathbf{A}$ and the observer then sees $s = 1$, then at the next time step the internal process can be in either state $\mathbf{A}$ or $\mathbf{B}$. This ambiguity grows as one looks at longer and longer sequences. Generally, indeterminism leads to a many-to-one association between internal state sequences and measurement sequences. In this example, the observation of 0110 could have been produced from either the internal state sequence $\mathbf{BAABA}$ or $\mathbf{BABBA}$.

The consequences of indeterminism, though, become apparent in the second inference step: learning the observed sequences' probabilities. The machine resulting from full reconstruction is shown in Figure 6. It has an infinite number of causal states. All of their transitions are deterministic. Note that the infinite machine preserves the original process's reset property: when $s = 0$ is observed the machine moves to a unique state and from this state $s = 1$ must be seen. But what happened, in comparison to the finite machine of Figure 5, to produce the infinite machine in Figure 6? The indeterminism mentioned above for state $\mathbf{A}$ has lead to a causal representation that keeps track of the number of consecutive 1s since the last $s = 0$. For example, if 01 has been observed, then $\Pr(s = 0) = \frac{1}{4}$ and $\Pr(s = 1) = \frac{3}{4}$. But if 011 has been observed, $\Pr(s = 0) = \frac{1}{3}$ and $\Pr(s = 1) = \frac{2}{3}$. In this way the causal representation accounts for the observer's uncertainty in each internal states' contribution to producing the next symbol. The result is that as more consecutive 1s are seen the relative probability of seeing $s = 0$ or $s = 1$ continues to change — and eventually converges to a fair coin. This is reflected in the change in transition probabilities down the machine's backbone. Causal machine reconstruction shows exactly what accounting is required in order to correctly predict the transition probabilities. But it gives more than just optimal prediction. It provides an estimate of the process's complexity and a complete representation of the distribution $\Pr(\omega)$ over infinite sequences.

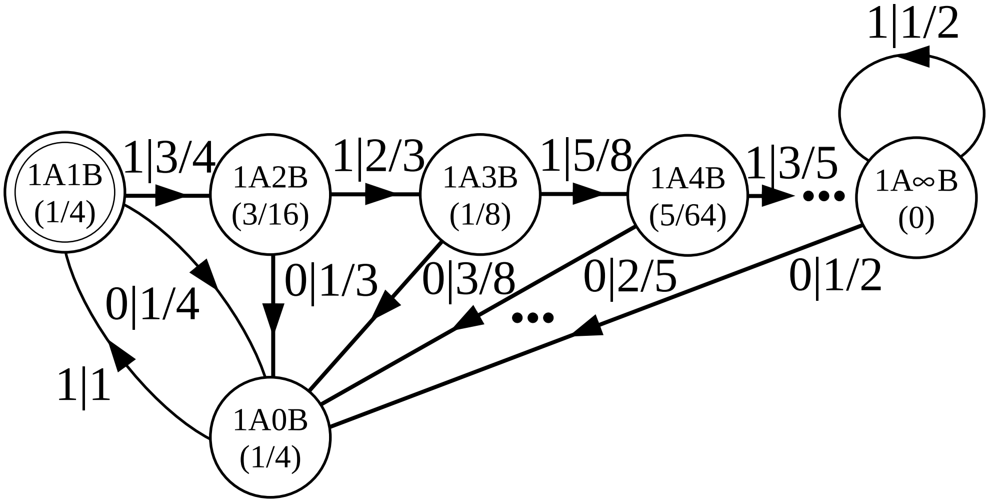

Figure 6 The infinite causal representation of the nondeterministic process of Figure 3. The labels in the states indicate the relative weights of the original internal states $\{\mathbf{A}, \mathbf{B}\}$. The numbers in parentheses are the asymptotic state probabilities: $\Pr(v = 1\mathrm{A}i\mathrm{B}) = (i + 1)2^{-i-2}$.

Interestingly, even if the observer has knowledge of Figure 3, the infinite causal machine of Figure 6 represents in a graphical way the requirements for achieving optimal predictability of the original process. There is no shortcut to computing, for example, the original process's entropy rate and complexities, since the machine in Figure 6, though infinite, is minimal. That is, there is no smaller (causal) machine that correctly gives $\Pr(\omega)$. From the topological machine it follows that the topological entropy is $h = \log_2 \phi \approx 0.694242$ and from Eqs. (7) and (10) that the metric entropy is $h_\mu \approx 0.677867$ bits per symbol. Recall that the original process's topological and statistical complexities were zero. From Eqs. (9), (10), and (11) the causal machine's topological complexity is infinite, $C_0 = \log_2 \|\mathbf{S}\|$, and its statistical complexity is $C_\mu \approx 2.71147$ bits. These are rather large changes in appearance due to the instrumentation.

This example is just one from a rich class of processes called — depending on the field — recurrent hidden Markov models, stochastic nondeterministic finite automata, or functions of Markov chains. The difficulty of finding the entropy rate for these processes was first noted in the late 1950's.[42] It is only recently, however, that a procedure for determining the equivalence of two such processes has been given.[43] That this problem area bears on the complexity of observation and the result that finite complexity processes can appear infinitely complex is also recent.[44]

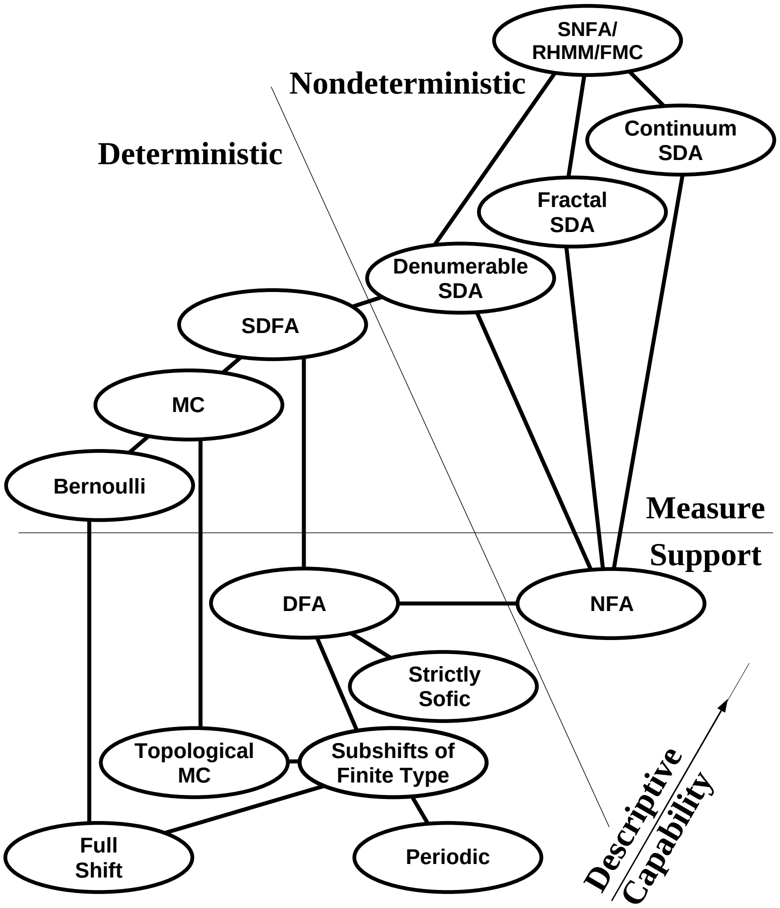

Figure 7 The computational hierarchy for finite-memory nonstochastic (below the Measure-Support line) and stochastic discrete processes (above). Here "Support" refers to the sets of sequences, i.e. what the "topological" machines describe; "Measure" refers to sequence probabilities, i.e. what the "stochastic" machines describe. The abbreviations are: A is automaton, F is finite, D is deterministic, N is nondeterministic, S is stochastic, MC is Markov chain, HMM is hidden Markov model, RHMM is recurrent HMM, and FMC is function of MC.

Getting back to the view of "mapping between classes", the preceding results can be summarized using the computational model hierarchy of Figure 7. In this figure each ellipse denotes a model class. As one moves up the diagram classes become more powerful in the sense that they can finitely describe a wider range of processes than lower classes. A class below and connected to a given one can finitely describe only a subset of the processes finitely described by the higher one. Additionally, the hierarchy is only a partial ordering of representation capability. There can be incomparable classes.

In formal language theory it is well-known that deterministic finite automata (DFA) are as powerful as nondeterministic finite automata (NFA).[24] This is shown in the hierarchy as both classes being connected at the same height. But the equivalence is just topological; that is, it concerns only the descriptive capability of each class for sets of observed sequences. If one augments these two classes, though, to account for probabilistic structure over the sequences, the equivalence is broken in a dramatic way — as the above example demonstrated. This is shown in the figure. The class of SNFA is higher than that of the stochastic deterministic finite automata (SDFA). Crudely speaking, if a DFA has transition probabilities added to its edges, one obtains the single class of SDFA. But if transition probabilities are added to NFA, then the class is qualitatively more powerful and, as it turns out, splits into three distinct classes.[44] Each of these classes is more powerful than the SDFA class. The new causal classes — called stochastic deterministic automata (SDA) — are distinguished by having a countable infinity, a fractional continuum, or a full continuum of causal states.

Initially, the original process (Figure 3) was undistinguished as an SNFA process. Via the analysis outlined above its causal representation showed that it is equivalent to a denumerable stochastic deterministic automaton (DSDA). And, generally, in terms of descriptive power $\mathrm{DSDA} \subset \mathrm{SNFA}$. But recall that we interpret the SNFA as the system-under-study, which is a Markov chain (MC), plus a measuring instrument. So the computational class interpretation of the endocomplexity explosion is that $\mathrm{MC} \rightarrow \mathrm{DSDA}$ under measurement distortion. That is, MC and DSDA are qualitatively different classes and, in particular, $\mathrm{MC} \subset \mathrm{DSDA}$, as shown in the hierarchy. The representational divergence that separates them is characteristic of the transition from a lower to a higher class.

3.2 ... Spatial Distortion

The next example illustrates how spatial measurement distortion — errors introduced in detecting the "local" state in a small region — leads to increased apparent complexity. It turns out that the problem is somewhat more extreme here than in the previous case. But the setting uses two of the simplest model classes for spatial processes — cellular automata (CA) and cellular transducers (CT). Both of these are discrete in space, time, and site value. First, the model classes will be briefly introduced, along with a measure of the degree of reconstructibility that will be used to monitor the success of the modeling effort. Then a series of models of a cellular automaton will be reconstructed from pattern data observed via a local instrument. The results will be interpreted in the larger context of a computational hierarchy for discrete spatial processes — an analog of the one just seen for finitary stochastic processes.

The following assumes the reader is somewhat familiar with deterministic CA.[45] At time t the global state $\mathbf{q}_{\mathrm{t}}$ of a CA is a sequence of symbols in some local state alphabet: $\mathbf{q}_t = q_t^0 q_t^1 \ldots q_t^{N-1}, q_t^i \in \mathcal{Q}$, for an N site lattice. The global state's temporal evolution is specified by the CA's rule table $\phi : \mathcal{Q}^{2\mathrm{r}+1} \rightarrow \mathcal{Q}$ that maps a neighborhood pattern $p = q^{-r} \ldots q^0 \ldots q^r \in \mathcal{Q}^{2\mathrm{r}+1}$ of radius r to the value of the site at the next time. That is, the symbols in the local state at the next time are determined by the equations of motion $q^i_{t+1} = \phi(q^{i-r}_t \ldots q^i_t \ldots q^{i+r}_t)$. The look up table (LUT) ϕ specifies the local space-time dynamics and is identified by an integer index, the CA rule number.[45] A schematic view of a CA's information processing architecture is given in Figure 8.

Now consider a probabilistic generalization of CA.[46] A stochastic CA (SCA) is specified by its probability transition table

$$q^i_{t+1} = \begin{cases} 0 & \text{with } \Pr(0|q^{i-r}_t \ldots q^i_t \ldots q^{i+r}_t) \\ 1 & \text{otherwise} \end{cases} \tag{12}$$

where $\Pr(q|p)$ is the probability of local state q conditioned on seeing neighborhood pattern p at the previous time. The degree of an SCA's nondeterminism is measured by the indeterminacy

$$\Xi(r) = - \sum_{p \in \mathcal{Q}^{2r+1}} Pr(p) \sum_{q \in \mathcal{Q}} Pr(q|p) \log_2 Pr(q|p) \tag{13}$$

$\Xi(r)$ measures the uncertainty of a site's value at the next time step given knowledge of the current neighborhood pattern. Its units are bits per site per unit time. If $\Xi(r) = 0$, then the SCA reduces to a deterministic CA of radius r: there is no choice in the future site values. Note that estimation of the indeterminacy $\Xi(r)$ requires the reconstruction of a radius r SCA since it uses the associated neighborhood conditional probability distribution, Eq. (12).

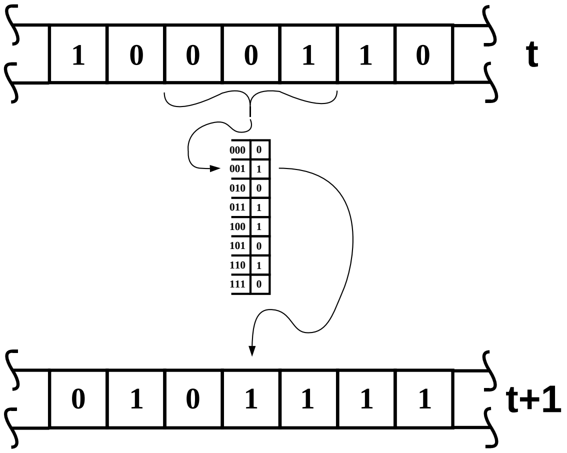

Figure 8 A schematic view of a cellular automaton's information processing architecture. The CA's spatially-local dynamic is shown at the center as a look up table (LUT) that maps from neighborhoods — triplets of sites in this case — to the next value taken by the center site. The LUT shows ECA 90.

A space-time diagram showing the evolution of the binary $r = 1$ elementary CA (ECA) 90 from a random initial condition is shown in Figure 9 for reference. The most notable space-time feature of ECA 90 consists of triangles in which contiguous 0 sequences shrink in length with time. This indicates a local "saturation and reset" dynamics: a contiguous patch of three or more 1s or of 01s resets to a patch of 0s. Then the spatial propagation of information invades the 0-patch from the left and right sides. But other than this feature the space-time diagram appears more or less structureless. The spatial entropy density was estimated to be $h_\mu(\mathbf{q}_t) \approx 1.00$ bits per site — the highest possible — and the spatial statistical complexity was $C_\mu(\mathbf{q}_t) \approx 0.00$ bits

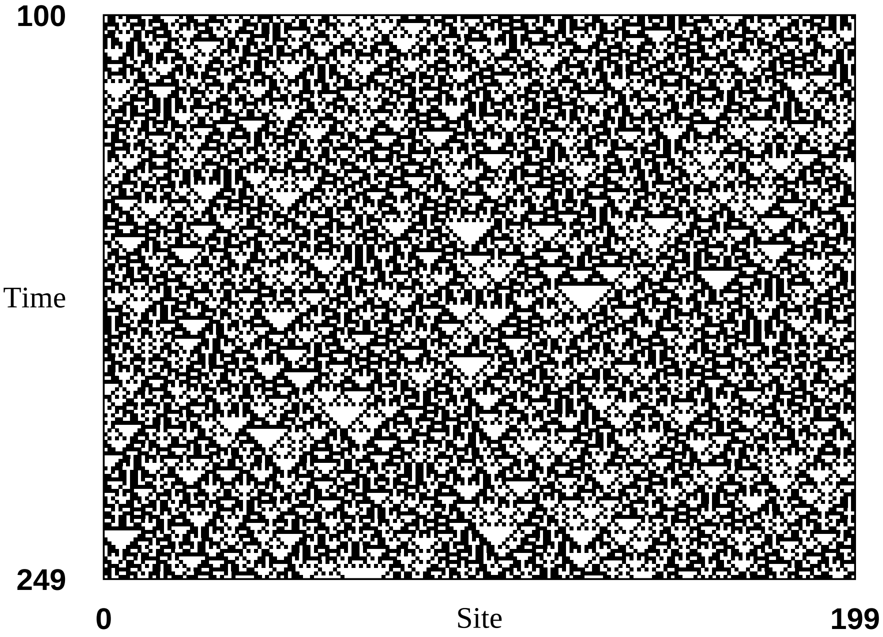

Figure 9 Space-time diagram of elementary CA 90. The horizontal axis gives the spatial site index; the vertical, time increasing downward. $N = 200$ sites are shown for 150 iterations, after 99 transient steps. Black cells denote $q_t^i = 1$; white, $q_t^i = 0$. The initial pattern was arbitrary.

per site — the lowest possible. ECA 90 is a linear CA in the sense that the space-time diagram is the superposition of the evolution from initial patterns with single 1-sites.[47]

In contrast to CA or SCA the (deterministic) cellular transducer (CT) explicitly incorporates the measurement process. There are both internal hidden states and observable symbols at each lattice site. When the number of local internal states is finite, we refer to a finitary CT (FCT). A schematic view of an FCT's information processing architecture is given in Figure 10. When the internal dynamics is governed by a CA and the instrument is also given by a CA LUT, then we have the class of elementary FCT (EFCT). EFCT are "cellular automata with cellular measuring instruments".

Let's define EFCT a little more carefully. An EFCT's local state is a pair $\left(q_t^i, s_t^i\right)$ of symbols, one q_t^i from a finite set $\mathcal{Q}$ of internal states and the other s_t^i from a finite set $\mathcal{A}$ of observed measurements. The global internal state $\mathbf{q}_t$ evolves as in a CA: there is an internal state update rule ϕ that operates on an internal neighborhood pattern to produce the next internal state q_{t+1}^i. In contrast to CA, however, an observer does not have direct access to the internal states, but instead measures symbols that are a spatially-local function of the internal state neighborhood. That is, the observed global state $\mathbf{s}_t = s_t^0 s_t^1 \ldots s_t^{N-1}, s_t^i \in \mathcal{A}$, is determined by an observation function ψ

$$s_{t+1}^i = \psi\left(q_t^{i-r} \ldots q_t^i \ldots q_t^{i+r}\right) \tag{14}$$

Let ϕ and ψ denote the CA rule number associated with the ϕ and ψ rule tables. Then EFCT can be denoted $\phi \backslash \psi$ with the first number identifying the internal state rule table and the second, the observation function. In this notation ECA 90 is equivalent to FCT 90\204 since $\psi = 204$ is the nearest neighbor identity LUT.

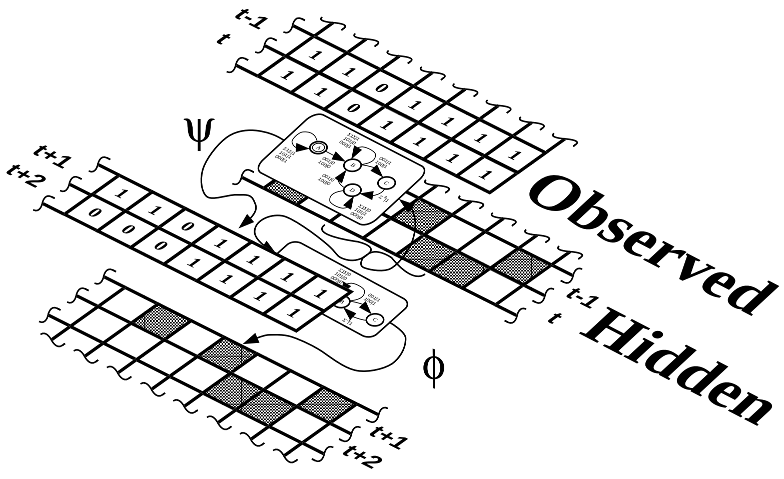

Figure 10 A schematic view of a finitary cellular transducer's information processing architecture. The hidden state and hidden dynamic (lower layer) are governed by a local finite state machine that maps neighborhoods to the center site's next local state. The instrument (upper layer) also maps from neighborhoods, but to the observed symbol at the center site. If both the hidden and instrumentation local state machines are LUTs, then we have the class of elementary FCTs.

The observed patterns $\{\mathrm{s}_t\}$ generated by FCT 90\222 are shown in Figure 11 in the same format as Figure 9. For direct comparison, the same initial state was used in both figures. This is reflected in the spatio-temporal coincidence of the 0-triangles, for example, in the two diagrams. Other than this there is little superficial commonality to the space-time diagrams. For FCT 90\222 there are fewer 0-triangles, a very high proportion of $q_t^i = 1$ sites, and a number of isolated $q_t^i = 0$ sites. The observed entropy and complexity were $h_\mu(\mathrm{s}_t) \approx 0.76$ and $C_\mu(\mathrm{s}_t) \approx 0.13$ bits per site; the comparable internal quantities are given by ECA 90: 1.0 and 0.0, respectively, as noted above. Thus, although the internal and observed patterns are largely unpredictable, more memory must be used to predict the FCT's observed patterns. This example illustrates a property of deterministic instruments: the observed data's unpredictability cannot be larger than the internal process's; but the statistical complexity can be either increased or decreased.

The spatial inference problem of interest can now be stated — Can the observed space-time data from FCT 90\222 be reconstructed as a CA? Or, in statistical terms, how well does an estimated model approximate the probability distribution

$$\mathrm{Pr}\begin{pmatrix} \ddots & & \vdots & & \\ & s_{t-1}^{i-1} & s_{t-1}^{i} & s_{t-1}^{i+1} & \\ \cdots & s_{t}^{i-1} & s_{t}^{i} & s_{t}^{i+1} & \cdots \\ & s_{t+1}^{i-1} & s_{t+1}^{i} & s_{t=1}^{i+1} & \\ & & \vdots & & \ddots \end{pmatrix} \qquad (15)$$

over space-time regions?

Figure 12 shows a space-time diagram generated from the radius $r = 3$ estimated CA — it is actually a stochastic CA. The evolution should be compared with Figures 9 and 11. All three space-time diagrams used the same initial pattern.

Estimating the radius $r = 0$ to $r = 3$ nearest-neighbor conditional transition tables gave SCAs with relatively large indeterminacies: $\Xi(r) = \{0.738, 0.468, 0.442, 0.429\}$ with $r =$

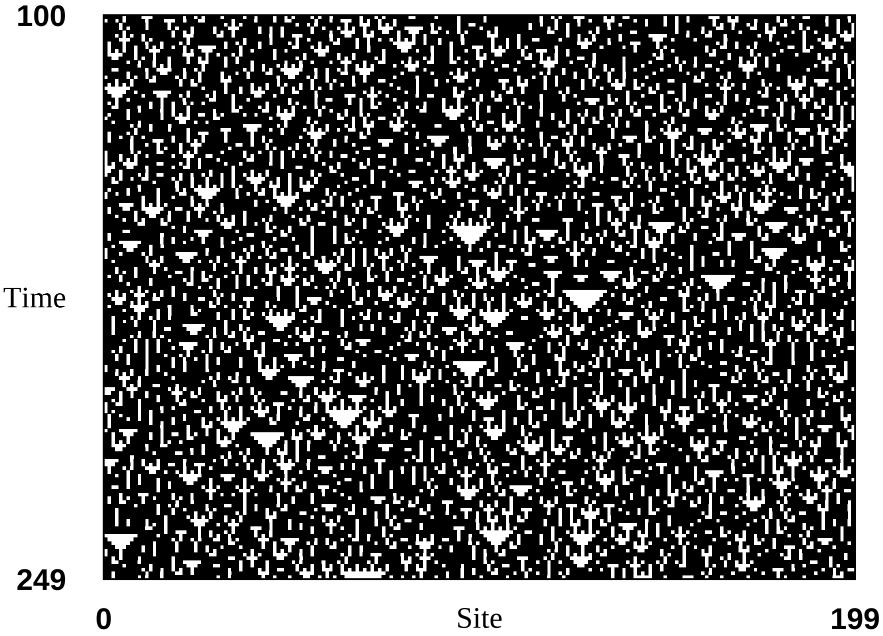

Figure 11 Space-time diagram of elementary FCT 90\222. Same initial pattern and format as in Figure 9.

$0, 1, 2, 3$, respectively. This indicates that the estimated dynamic is moderately stochastic. Figure 12 indicates more graphically that the estimated SCA differs from FCT 90\222. There is, for example, no spatio-temporal coincidence of 0-triangles when compared to Figure 11. Additionally, the size distribution of 0-triangles has shifted to smaller lengths and there are almost no 0-triangles topped with contiguous 1s. These differences are not reflected in the SCA's spatial entropy, $h_\mu(\mathbf{q}_t) \approx 0.78$, which is close to that (0.76) found for FCT 90\222. Though the SCA's patterns are less complex, $C_\mu(\mathbf{q}_t) \approx 0.0$, than those generated by FCT 90\222 (0.13).

Nonetheless, it turns out that the indeterminacy remains relatively high at larger radius and therefore the data series is not well modeled by SCA. It even appears to reach a plateau. On an $N = 500$ site lattice that started from an arbitrary pattern and that was allowed to relax for 10^4 iterations, a radius $r = 10$ indeterminacy of $\Xi(10) \approx 0.39$ bits was estimated over 10^6 iterations. Note that the CA model at $r = 10$ has something like 2×10^6 parameters to be estimated; though, only several thousand require much data for estimation, since they appear to be neither 0.0 nor 1.0. The FCT 90\222 space-time data appear "unreconstructible" with respect to the CA class.

The conclusion from space-time diagrams, indeterminacy, entropy, and statistical complexity, is that even large radius SCA, let alone deterministic CA, do not capture the structures generated by nearest-neighbor EFCT. Large indeterminacy at large radius suggests, erroneously, that the mechanism underlying the FCT data series has, at a minimum, a large spatial radius dynamic coupled to a stochastic process. Thus, even though the FCT considered has radius one, to an observer the patterns appear to be generated by a relatively nonlocal SCA: important structure appears in neighborhoods of 21 sites rather than just 3 sites. It is also noteworthy that when the estimated dynamic is unstable ($h_\mu > 0$) even the smallest indeterminacy leads to significant prediction errors between the given and the simulated data series. This appeared in the lack of absolute time and space correlation between the FCT and reconstructed SCA time histories

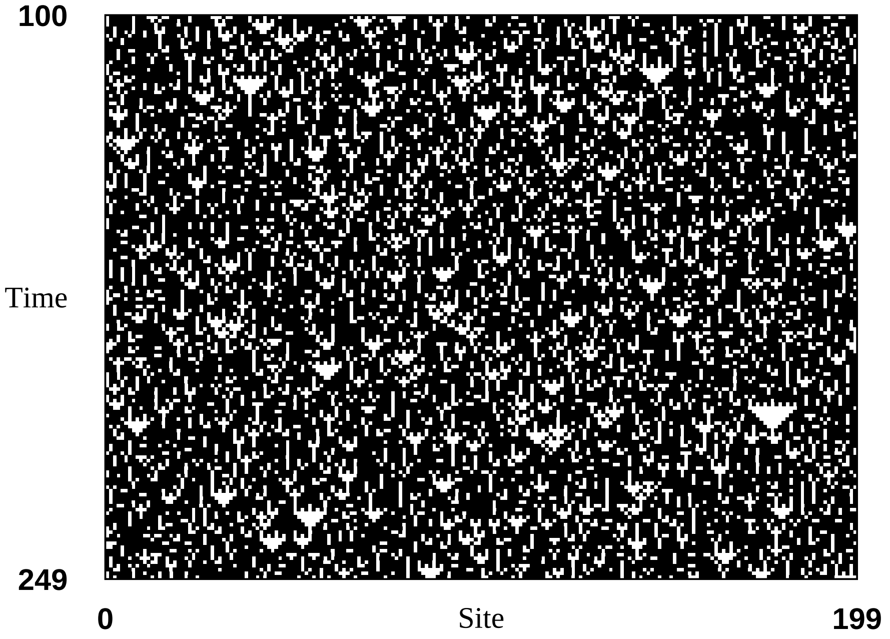

Figure 12 Simulation of the radius $r = 3$ SCA estimated from an FCT 90\222 pattern data series. Same initial pattern and format as in Figure 11. The 128 SCA LUT parameters were estimated using spatial data series collected, after 10^4 transient iterations, over 10^6 iterations on an $N = 500$ site lattice starting from an arbitrary initial state.

in Figures 11 and 12; which was due, of course, to the instability-driven amplification of the SCAs stochastic behavior.

So what, if anything, is troubling about this? First of all, the exponential explosion in model complexity with increasing radius puts severe constraints on the observer's ability to predict space-time patterns and their probabilities of occurrence. The complexity of the SCA class is exponential in radius: there are 2^{2r+1} conditional transition probabilities at radius r that must be estimated. Any observer with finite computational resources — finite in time or in storage, for example — that assumes CA or SCA classes will rapidly run out of these resources when trying to improve predictability by (say) increasing radius. An irreducible randomness is the result of reaching that point of exhaustion. Second of all, as designers of the universe of discourse we know that it simply is not the case that the dynamics is either stochastic — the FCT is purely deterministic — or that it used an infinite amount of computational power to produce the space-time patterns — the FCT used two nearest-neighbor LUTs and stored two binary states, one internal and one external.

As in the previous section, the apparent-complexity explosion can be summarized via a computational hierarchy. (See Figure 13.) This one describes the model classes for spatial systems that are discrete in space, time, and local state. It is concerned with the representability of spatio-temporal patterns with respect to automata that operate in parallel, at every site simultaneously, rather than with respect to automata that serially scan the spatio-temporal pattern into one-dimensional symbol strings. The latter approach to complexity classification is used, for example, in Refs. [29], [45], and [48]. Finally, it is focused on deterministic machine architectures and not, for example, on grammars or production rules.

The endocomplexity explosion is reflected in the spatial discrete computation hierarchy as the inequivalence of elementary FCT and CA. The hierarchy indicates that FCT is a more powerful class. The implication is that the observer's use of a lower level representation leads to an infinite representation for spatio-temporal processes — i.e. systems-under-study — that are strictly produced in a higher level class.

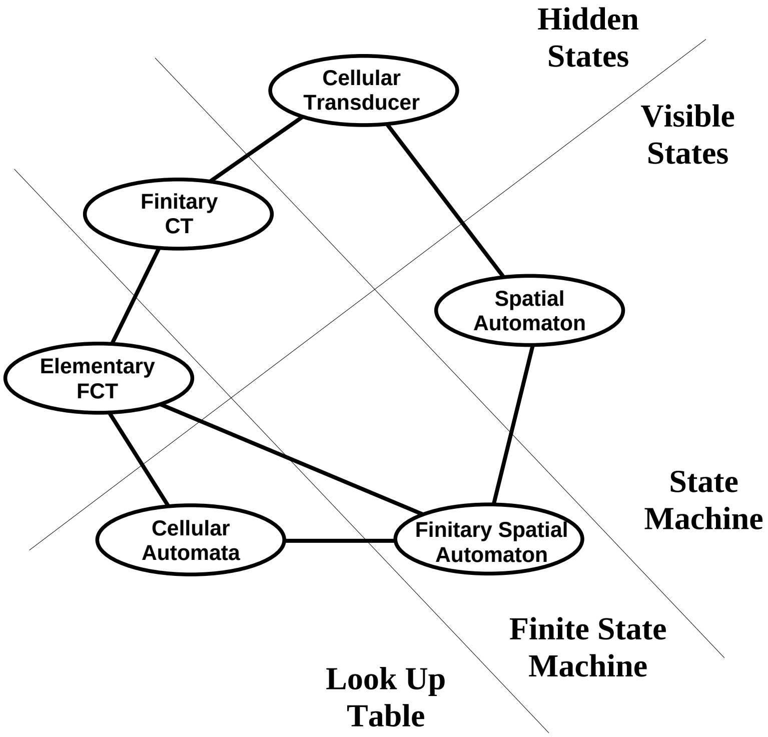

Figure 13 The spatial computation hierarchy for discrete-local-state lattice dynamical systems. Unlike the previous hierarchy of Figure 7 the stochastic analogs are not shown. Little appears to be known about that extension. The classes here represent the minimal parallel (deterministic) automaton architectures necessary for representing spatio-temporal patterns. Their relationship to the recognition complexity classes for spatio-temporal patterns that are scanned into one-dimensional string languages is not yet worked out. The abbreviation CT denotes cellular transducer. The phrases "Look up Table", "Finite State Machine", and "State Machine", refer to the type of local equations of motion.

3.3 ... Chaos

The previous two examples concerned processes over discrete-alphabet strings. As such their complexity and structure could be analyzed in much the same way that contemporary computation theory views formal languages: explicit computational hierarchies could be delineated, for example. But many models used in science deal with processes with continuous states. This section briefly reviews some work along these lines that addresses the measurement problem. The idea is to take iterated maps of the interval — a favorite set of prototypes — as a legitimate real-valued computation class.* First a particular iterated map is introduced and then three examples are presented to illustrate the intrinsic complexity in continuum-state processes and the effect of measurement distortion.

* This should be compared with Ref. [49].

The data streams of interest here are derived from a trajectory of a continuum-state dynamical system, the logistic map, observed with a very coarse measuring instrument. The trajectory is generated by iterating the map

$$x_{n+1} = f(x_n) \tag{16}$$

with the logistic function $f(x) = rx(1 - x)$ in which $r \in [0, 4]$ and $x_0 \in [0, 1]$. The map's maximum occurs at $x_c = \frac{1}{2}$. The trajectory $\mathrm{x} = x_0 x_1 x_2 x_3 \ldots$ is converted to a discrete sequence by observing it via the generating binary partition

$$\mathcal{P} = \{x_n \in [0, x_c) \Rightarrow s = 0, x_n \in [x_c, 1] \Rightarrow s = 1\} \tag{17}$$

The generating property means that sufficiently long binary sequences identify arbitrarily small segments of initial conditions. Due to this, the information processing in the logistic map can be studied using the "coarse" measuring instrument $\mathcal{P}$.

The first example looks at the period-doubling onset of chaos at $r = r_c \approx 3.5699456718695445\ldots$. The data stream produced at the onset of chaos leads to an infinite machine. (For details see Refs. [18] and [50].) This is consonant with the view introduced by Feigenbaum that this onset of chaos can be viewed as a phase transition at which the correlation length diverges.[51] The computational analog, as we have analyzed it, is that the process intrinsically has an infinite memory capacity. There is more that the computational analysis gives, however. For example, the infinite memory is organized in a particular way such that the logistic map is not a universal Turing machine, but a less powerful nested stack automaton. The complexity here is a property not so much of the measuring instrument, but of the internal dynamics; even though the internal continuum states are observed with a coarse (binary) measuring instrument. Infinite complexity also appears for other routes to chaos, such as that found via the frequency-locking of incommensurate oscillators.

The second example examines the logistic map at $r = 3.7$, well into the map's chaotic regime. The reconstructed machine is shown in Figure 14. As in any estimation procedure there are parameters to be set in machine reconstruction. For the machine shown in the figure the two parameters of interest are the tree depth D and the morph depth L, which were set to $(D, L) = (12, 6)$. As the caption notes, however, at larger reconstruction parameters more states are found. The machine size grows very slowly, as the observer attempts to make better models. Unlike the onset of chaos for which there is a linear lower bound growth, at $r = 3.7$ one does not know whether the machine size will become finite or will diverge. In contrast, there are other r-values at which one knows the topological machines are finite; these are characterized by $f^n(x_c)$ becoming asymptotically periodic. But at $r = 3.7$ it simply appears that the topological and metric machines are infinite. If this is the case, then the continuum computational process may be manifesting itself by leading to infinite discrete representations.

The final example of this section shows how to construct an instrument so that the logistic map, at its most random and least complex parameter value, appears equivalent to the simple SNFA of Figure 3.

First, set $r = 4$ — a parameter at which the logistic map's attractor fills the interval and has the maximal entropy rate of $h = h_\mu = 1$. Here the topological and statistical complexities vanish.

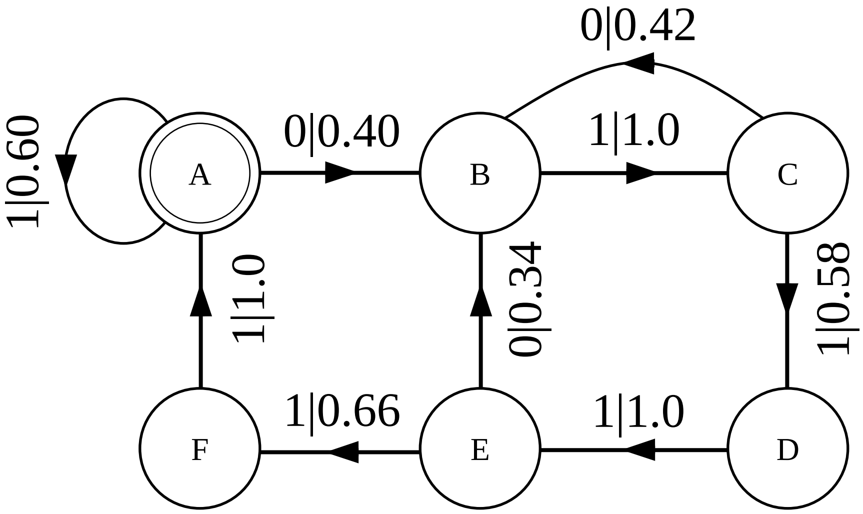

Figure 14 The six state machine reconstructed from a binary partition of the logistic map at $r = 3.7$ using 10^8 iterations, after 300 transient iterations, and using tree depth $D = 12$ and morph depth $L = 6$. If $D = 16$ and $L = 8$ we also find six causal states and nearly the same transition probabilities. But at $D = 20$ and $L = 10$ seven causal states are reconstructed; at $D = 22$ and $L = 11$, there were eight causal states; at $D = 24$ and $L = 12$, there were eleven causal states.

The probability density function of the invariant measure over “internal” states $x \in [0, 1]$ is

$$\Pr(x) = \frac{1}{\pi\sqrt{x - x^2}} \tag{18}$$

Second, associate a state $\mathbf{A}$ with the event $x_t \in [0, x_c)$ and a state $\mathbf{B}$ with the event $x_t \in [x_c, 1]$. Finally, use a sliding-block code on the resulting $\mathbf{A} - \mathbf{B}$ stream that outputs $s = 1$ when the length 2 subsequences $\mathbf{AA}$, $\mathbf{AB}$, or $\mathbf{BB}$ occur, and $s = 0$ when $\mathbf{BA}$ occurs. The binary data stream that is produced is exactly that produced by the SNFA of Figure 3.

This can be seen by noting that the two intermediate states here are the same as the SNFA internal states. They also have the same asymptotic probabilities; that is

$$\Pr(\mathbf{A}) = \int_0^{x_c} dx \Pr(x) = \frac{1}{2} \tag{19}$$

and by symmetry $\Pr(\mathbf{B}) = \frac{1}{2}$. The two inverse iterates of x_c, $x_\pm = x_c \pm \frac{1}{2\sqrt{2}}$, delimit the interval segments corresponding to the occurrence of $\mathbf{A} - \mathbf{B}$ pairs. These are then used to compute the transition probabilities, such as

$$\Pr(\mathbf{A} \rightarrow \mathbf{A}) = \frac{\Pr(\mathbf{AA})}{\Pr(\mathbf{A})} = 2 \int_0^{x_-} dx \Pr(x) \tag{20}$$

It turns out they are all equal to $\frac{1}{2}$.

This construction might seem somewhat contrived with the use of the pairwise $\mathbf{A} - \mathbf{B}$ coding. But it can be reinterpreted without recourse to an intermediate code. It turns out that the $0 - 1$ data stream comes directly from the binary partition

$$\mathcal{P} = \{x_n \in [0, x_+) \Rightarrow s = 1, x_n \in [x_+, 1] \Rightarrow s = 0\} \tag{21}$$

This is a partition that is not much more complicated than the original. The main difference is that the "decision point", originally at x_c, has been moved over to x_+.

The observer can be considered to have simply selected the wrong instrument. The penalty is infinite complexity, as a previous section demonstrated for the simple SNFA of Figure 3. Thus, the logistic map can appear to have an infinite number of causal states and so infinite topological complexity. As in the preceding two sections, and in contrast to the preceding two logistic map examples which illustrated infinite intrinsic complexity, this one illustrates measurement-induced complexity, but for a continuum-state process.

4 Just So

4.1 Complexity of Generation $\neq$ Complexity of Recognition

These examples have shown that there can be several mechanisms responsible for the appearance of infinite complexity. One was that the process, such as the logistic map at the onset of chaos, intrinsically has an infinite amount of memory. The other, and the one that was emphasized, was indeterminism caused by measurement distortion. The operant mechanism in the latter cases was that indeterminism induced by the measurement process mapped the distribution over hidden state sequences onto the measurement sequence distribution in a subtle infinite-to-one way. Information appeared to have been lost since the entropy was reduced, but it reappeared as apparent complexity. Indeed, just throwing information away is not enough, since the infinite-to-one mapping to all 0s produces zero complexity in the measurement sequences. And this as a data stream is eminently reconstructible. To result in an infinite number of causal states, measurement distortion must additionally generate an infinity of conditional measures from the (possibly finite complexity) internal state sequence measure.

I hope it is clear that simple processes can appear to be quite complex. Stated more formally, the examples show how the complexity of recognizing behavior can be substantially higher than the complexity of generating that behavior. This appears as two types of problem: (i) the behavior appears quantitatively more or less complex and (ii) it appears qualitatively more or less complex. The latter refers to the change of computational classes that measurement distortion requires for causal recognition.

4.2 Extrinsic Noise and Limited Resources

The most mundane way the complexity explosion affects the study of endo-exo problems appears when considering observers with finite inference resources — like compute time and storage. As soon as these are limited then infinite complexity leads either to the appearance of effective randomness or to the need to change the observer's current model class. The next two sections address these in turn.

Before this, it is important to briefly remark on the effect of adding additional extrinsic noise by (say) flipping measurement symbols in a way that is uncorrelated with the internal dynamics or by putting the internal states themselves in contact with a heat bath. In turns out that in these cases, the apparent complexity is reduced monotonically with increasing extrinsic noise level.[16] This type of corruption of the internal state throws information away. Moreover, infinite apparent complexity becomes finite, even for infinitesimal noise. The coupling between the effects of extrinsic noise and apparent complexity gives one approximation to the analysis of how finite computational resources for inference affect apparent complexity.

4.3 Irreducible Classical Uncertainty

The preceding discussion turned on a curious problem: locally-deterministic or locally-stochastic behavior viewed with a locally-deterministic instrument can appear substantially more random than it is to all levels of approximation. What might the physical implications be? If local temporal or spatio-temporal states are obscured necessarily by the act of measurement, then microphysical reality would forever appear irreducibly uncertain. This would occur without invoking randomness; it could be a property of a purely deterministic world. Information distortion during measurement could be due to some intrinsic nonlinearity of the measurement act on microphysical scales. This might be analogous to (say) that seen in the stochastic NFA or in the unreconstructible FCT 90\222. Or it could be given by a measurement transducer more general than a LUT; for example, one that was itself dynamic. Apparent randomness, even on the shortest time scales, is consistent with underlying determinism.

Irreducible indeterminacy, as illustrated above, is consistent with internal deterministic dynamics, even though the latter may never be accessible, testable, or identifiable, using "reasonable" representations. Conversely, more sophisticated modeling techniques may be required for the discovery of internal structure than the estimation of local LUT-like statistics that assume strict independence in some form. At the very least, the physical implications point to an important role that a measurement theory of nonlinear chaotic processes can play in basic physical theory. In concert with this, a systematic reevaluation of how accepted model classes preclude the discovery of natural mechanisms appears necessary. An emphasis on discovering causal states would seem especially helpful in this. The two computational hierarchies presented here also go some distance in this direction.

4.4 Towards a Theory of Hierarchical Learning

Is there a way around the problem of distortion-induced complexity? Or, will we always be precluded from discovering simplicity due to our lack of prior knowledge of the structure of nature's processes? Given its determinant role in answering these questions, how can representation dependence be addressed? This last section sketches a solution to these questions — a way to break out of weak model classes, to learn more powerful ones. It is called hierarchical machine reconstruction.[48,52]

First, recall the common aspects of the computational hierarchies in Figures 7 and 13. At each level in a hierarchy there are a number of elements that can be identified, such as the following.

1. **Models** M, in some class $\mathcal{M}$, consisting of states and transitions observed via a measurement function.
2. **Languages** being the ensembles of finitely representable behaviors.

3. **Symmetries** reflecting the observer's assumptions about a process's structure. These determine the semantic content of the model class $\mathcal{M}$, which is defined by equivalence relations $\sim$ corresponding to each symmetry.
4. **Reconstruction** being the procedure for producing estimated models. It factors out a symmetry from a data stream s. Formally, reconstruction of model $\mathrm{M} \in \mathcal{M}$ is denoted as $\mathrm{M} = \mathrm{s}/\sim$.
5. **Complexity** of a process being the size of a reconstructed model M with respect to the given class $\mathcal{M}$: $C(\mathrm{s}|\mathcal{M}) = \|\mathrm{M}\|$.
6. **Predictability** being estimated with reference to the distinguishable states as in Eq. (7).

It is crucial that reconstructed models $\mathrm{M} \in \mathcal{M}$ be minimal. This is so that M contains no more structure than and no additional properties beyond the system-under-study. The simplest explication of this is to note that there are many multiple state representations of an ideal random binary string. But if the size of representation is to have any meaning, such as the amount of memory, only the single state process can be allowed as the model from which it is computed. Additionally, a minimal model maximizes posterior distribution $\Pr(\mathrm{M}|\mathrm{s})$ over $\mathcal{M}$ via Bayesian balancing of the modeling prior $\Pr(\mathrm{M})$ and the sample likelihood $\Pr(\mathrm{s}|\mathrm{M})$. It is important to keep in mind that Bayesian optimization is applied only within a given model class. But as such it does allow one to automatically determine the setting of the reconstruction parameters.

At this level of abstraction, viz. discussing the structure of the hierarchy of model classes, the relativity of information, entropy, and complexity becomes clear. They all depend on the observer's assumed representation. And the representation's properties determine what they can mean.

ϵ-machine reconstruction was introduced above as a way to find causal states. It was also noted that these states appeared to be related to notions of state familiar from other fields. But, it should now be clear that there is an inductive hierarchy delineated by different notions of state.

Finally, sufficient groundwork has been laid in order to formulate the definition of an ϵ-machine. An ϵ-machine is that

minimal model at the
least computationally powerful level yielding a
finite description.

The definition builds in an adaptive notion that the observer initially might not have the correct model class. How does it find a better representation? Moving up the inductive hierarchy can be associated with the innovation of new notions of state and so new representations. One can envision a procedure — call it hierarchical machine reconstruction — that implements this incremental movement up the hierarchy as follows.*

1. At the lowest level, the data stream is its own, rather degenerate and uninformative, model: $\mathrm{M}_0 = \mathrm{s}$. Initially set the hierarchy level indicator to one step higher: $l = 1$.
2. Reconstruct the level l model M_l from the lower level model by factoring out the regularities — equivalence classes — in the state transition structure of the lower level model M_{l-1}: $\mathrm{M}_l = \mathrm{M}_{l-1}/\sim$, where $\sim$ denotes the equivalence relation defining the equivalence classes.

* The procedure assumes a (possibly infinite) collection of symmetries that is complete.

Literally, one looks for regularities in groups of states in M_{l-1}. The groups revealing regularity in M_{l-1} become the states of M_l; the transitions between the M_{l-1}-state groups become the transitions in M_l.

3. Test the parsimony of the l-level class's descriptive capability by estimating successively more accurate models. The degree of approximation is generally denoted ϵ here, with $\epsilon \rightarrow 0$ being the limit of increasingly accurate models. If the model complexity diverges, $\|\mathrm{M}_l\| \underset{\epsilon \rightarrow 0}{\rightarrow} \infty$, then set $l \leftarrow l + 1$ and go back to 2 and move up another level.
4. If $\|\mathrm{M}_l\| \underset{\epsilon \rightarrow 0}{<} \infty$, then the procedure has found the first level that is the least computationally powerful and that gives a finite description. An ϵ-machine has been reconstructed. Quit.

The process of going from step 3 back to step 2 — i.e. of jumping up the hierarchy to a new model class — is called "innovation". A large part of innovating a new model class is simply a reapplication of machine reconstruction. As noted, the central method of discovering structure is to group lower-level states into equivalence classes of the same future morph. These equivalence classes then become the notion of state at the new level. A series of increasingly accurate lower level models are, in this sense, a data stream — $\mathrm{M}_{l-1}(\epsilon), \mathrm{M}_{l-1}\left(\frac{\epsilon}{2}\right), \mathrm{M}_{l-1}\left(\frac{\epsilon}{4}\right), \mathrm{M}_{l-1}\left(\frac{\epsilon}{8}\right), \ldots$ — for reconstruction at the next higher level M_l. For example, at the onset of chaos hierarchical machine reconstruction goes across four levels — data, trees, finite automata, and stack automata — before finding a finite representation.

There is an additional element beyond the grouping of states according to their transition (morph) structure, though. This is seen in the SNFA example as the innovation of a resettable counter for DSDA,[44] at the onset of chaos as the innovation of string productions,[50] and in discrete spatial processes as the innovation of local state machines to break away from cellular automata LUT representations.[28] In each case it was quite straightforward to find the additional structural element riding on top of the higher level causal states. But since, as far as I know, no one has delineated an exhaustive and ordered spectrum of basic computational elements, innovation must contain a component, albeit small, of undetermined discovery.

I still hold out a hope for complete automation of hierarchical machine reconstruction. But I hold no illusions as to its simplicity. The simplest way to say this is that the inductive computational hierarchy, like others, is only a partial ordering. There is no reason to think that it would be a linear order — other than one's belief in the simplicity of nature, perhaps. And, for that matter, there is no reason (yet) to think that the branching degree, as one moves up the inductive hierarchy, will be finite, let alone small.

The meta-reconstruction algorithm results in a hierarchy of computation classes — the ϵ-machine hierarchy. Unlike the generative hierarchy of Chomsky,[24] this is a causal hierarchy for inductive inference. It takes into account the possibility, for example, that causal recognition might be distinct from the complexity of the generating process.

Acknowledgments

This work has benefitted from suggestions by Harald Atmanspacher, Jim Hanson, Melanie Mitchell, Dan Upper, and Karl Young. It was supported by AFOSR 91-0293.

Bibliography

[1] J. P. Crutchfield. *Noisy Chaos*. PhD thesis, University of California, Santa Cruz, 1983. Published by University Microfilms Intl, Ann Arbor, Michigan.

[2] F. W. Kantor. *Information Mechanics*. Wiley, New York, 1977.

[3] H. Atmanspacher. The aspect of information production in the process of measurement. *Found. Phys.*, 19:553, 1989.

[4] J. P. Crutchfield. Information and its metric. In L. Lam and H. C. Morris, editors, *Nonlinear Structures in Physical Systems - Pattern Formation, Chaos and Waves*, page 119, New York, 1990. Springer-Verlag.

[5] K. Young and J. P. Crutchfield. Fluctuation spectroscopy. *Chaos, Solitons, and Fractals*, in press, 1993. Special Issue on Complexity, W. Ebeling, editor, SFI Technical Report 93-05-028.

[6] P. Martien, S. C. Pope, P. L. Scott, and R. S. Shaw. The chaotic behavior of the leaky faucet. *Physics Letters*, 110A:399, 1985.

[7] E. Nagel and J. R. Newman. *Gödel's Proof*. New York University Press, New York, 1968.

[8] A. M. Turing. On computable numbers, with an application to the entsheidungsproblem. *Proc. Lond. Math. Soc. Ser. 2*, 42:230, 1936.

[9] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of one-dimensional maps: Entropies, finite precision, and noise. *Intl. J. Theo. Phys.*, 21:433, 1982.

[10] O. E. Rössler. Endophysics. In J. L. Casti and A. Karlqvist, editors, *Real Brains, Artificial Minds*, pages 25–46, New York, 1987. North-Holland.

[11] H. Primas. Mathematical and philosophical questions in the theory of open and macroscopic quantum systems. In A. I. Miller, editor, *Sixty-Two Years of Uncertainty: Historical, Philosophical and Physical Inquiries into the Foundations of Quantum Mechanics*, pages 233–257, New York, 1990. Plenum.

[12] D. Finkelstein. Finite physics. In R. Herken, editor, *The Universal Turing Machine. A Half-Century Survey*, pages 349–376, Hamburg, 1988. Kammerer & Unverzagt.

[13] D. C. Dennett. *The Intentional Stance*. MIT Press, Cambridge, 1987.

[14] N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw. Geometry from a time series. *Phys. Rev. Let.*, 45:712, 1980.

[15] H. Poincaré. Why space has three dimensions. In J. W. Bolduc, editor, *Mathematics and Science: Last Essays*, pages 25–44. Dover Publications, New York, 1963.

[16] J. P. Crutchfield and N. H. Packard. Symbolic dynamics of noisy chaos. *Physica*, 7D:201, 1983.

[17] J. P. Crutchfield and B. S. McNamara. Equations of motion from a data series. *Complex Systems*, 1:417, 1987.

[18] J. P. Crutchfield and K. Young. Inferring statistical complexity. *Phys. Rev. Let.*, 63:105, 1989.

[19] J. P. Crutchfield. Knowledge and meaning ... chaos and complexity. In L. Lam and V. Naroditsky, editors, *Modeling Complex Phenomena*, page 66, Berlin, 1992. Springer-Verlag.

[20] J. P. Crutchfield. Semantics and thermodynamics. In M. Casdagli and S. Eubank, editors, *Nonlinear Modeling and Forecasting*, volume XII of *Santa Fe Institute Studies in the Sciences of Complexity*, page 317, Reading, Massachusetts, 1992. Addison-Wesley.

[21] P. E. Caines. *Linear Stochastic Systems*. Wiley, New York, 1988.

[22] B. Kitchens and S. Tuncel. Finitary measures for subshifts of finite type and sofic systems. *Memoirs of the AMS*, 58:no. 338, 1985.

[23] I. P. Cornfeld, S. V. Fomin, and Ya. G. Sinai. *Ergodic Theory*. Springer-Verlag, Berlin, 1982.

[24] J. E. Hopcroft and J. D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, Reading, 1979.

[25] O. Penrose. *Foundations of statistical mechanics; a deductive treatment*. Pergamon Press, Oxford, 1970.

[26] F. Takens. Detecting strange attractors in fluid turbulence. In D. A. Rand and L. S. Young, editors, *Symposium on Dynamical Systems and Turbulence*, volume 898, page 366, Berlin, 1981. Springer-Verlag.

[27] T. F. Meyer, F. C. Richards, and N. H. Packard. A learning algorithm for the analysis of complex spatial data. *Phys. Rev. Lett.*, 63, 1989.

[28] J. P. Crutchfield. Unreconstructible at any radius. *Phys. Lett. A*, 171:52 – 60, 1992.

[29] J. P. Crutchfield and J. E. Hanson. Turbulent pattern bases for cellular automata. *Physica D*, in press, December 1993. Santa Fe Institute Report SFI-93-03-010.

[30] M. Casdagli and S. Eubank, editors. *Nonlinear Modeling*, SFI Studies in the Sciences of Complexity, Reading, Massachusetts, 1992. Addison-Wesley.

[31] J. Hertz, A. Krogh, and R. G. Palmer. *An Introduction to the Theory of Neural Networks*, volume 1 of *Lecture Notes, Studies in the Sciences of Complexity*. Addison-Wesley, Redwood City, California, 1991.

[32] D. Angluin and C. H. Smith. Inductive inference: Theory and methods. *Comp. Surveys*, 15:237, 1983.

[33] C. E. Shannon and W. Weaver. *The Mathematical Theory of Communication*. University of Illinois Press, Champaign-Urbana, 1962.

[34] C. E. Shannon. Prediction and entropy of printed english. *Bell Sys. Tech. J.*, 30:50, 1951.

[35] I. Shimada and T. Nagashima. A numerical approach to ergodic problem of dissipative dynamical systems. *Prog. Theo. Phys.*, 61:1605, 1979.

[36] J. P. Crutchfield. Prediction and stability in classical mechanics. University of California, Santa Cruz, 1979. Bachelor's Thesis.

[37] G. Benettin, L. Galgani, A. Giorgilli, and J.-M. Strelcyn. Lyapunov characteristic exponents for smooth dynamical systems and for hamiltonian systems; a method for computing all of them. *Meccanica*, 15:9, 1980.

[38] A. N. Kolmogorov. A new metric invariant of transient dynamical systems and automorphisms in Lebesgue spaces. *Dokl. Akad. Nauk. SSSR*, 119:861, 1958. (Russian) Math. Rev. vol. 21, no. 2035a.

[39] Ja. G. Sinai. On the notion of entropy of a dynamical system. *Dokl. Akad. Nauk. SSSR*, 124:768, 1959.

[40] D. S. Ornstein. Ergodic theory, randomness, and chaos. *Science*, 243:182, 1989.

[41] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, New York, 1979.

[42] D. Blackwell and L. Koopmans. On the identifiability problem for functions of Markov chains. *Ann. Math. Statist.*, 28:1011, 1957.

[43] H. Ito, S.-I. Amari, and K. Kobayashi. Identifiability of hidden Markov information sources and their minimum degrees of freedom. *IEEE Info. Th.*, 38:324, 1992.

[44] J. P. Crutchfield and D. R. Upper. In preparation, 1993.

[45] S. Wolfram. *Theory and Applications of Cellular Automata*. World Scientific Publishers, Singapore, 1986.

[46] L. Shulman and P. Seiden. Statistical mechanics of a dynamical system based on Conway's game of life. *J. Stat. Phys.*, 19:293, 1978.

[47] O. Martin, A. Odlyzko, and S. Wolfram. Algebraic properties of cellular automata. *Commun. Math. Phys.*, 93:219, 1984.

[48] J. E. Hanson and J. P. Crutchfield. The attractor-basin portrait of a cellular automaton. *J. Stat. Phys.*, 66:1415, 1992.

[49] L. Blum, M. Shub, and S. Smale. On a theory of computation over the real numbers. *Bull. AMS*, 21:1, 1989.

[50] J. P. Crutchfield and K. Young. Computation at the onset of chaos. In W. Zurek, editor, *Entropy, Complexity, and the Physics of Information*, volume VIII of *SFI Studies in the Sciences of Complexity*, page 223, Reading, Massachusetts, 1990. Addison-Wesley.

[51] M. J. Feigenbaum. Universal behavior in nonlinear systems. *Physica*, 7D:16, 1983.

[52] J. P. Crutchfield. Reconstructing language hierarchies. In H. A. Atmanspracher and H. Scheingraber, editors, *Information Dynamics*, page 45, New York, 1991. Plenum.