What Is a Macrostate?
Subjective Observations and Objective Dynamics

Cosma Rohilla Shalizi[1,2,*] and Cristopher Moore[3,4,2,†]
[1] *Center for the Study of Complex Systems, University of Michigan, Ann Arbor, MI 48109*
[2] *Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, NM 87501*
[3] *Computer Science Department, University of New Mexico, Albuquerque, NM 87131*
[4] *Physics and Astronomy Department, University of New Mexico, Albuquerque, NM 87131*

(Dated: Begun 10 July 2000; last revised March 29, 2003)

We consider the question of whether thermodynamic macrostates are objective consequences of dynamics, or subjective reflections of our ignorance of a physical system. We argue that they are both; more specifically, that the set of macrostates forms the unique maximal partition of phase space which 1) is consistent with our observations (a subjective fact about our ability to observe the system) and 2) obeys a Markov process (an objective fact about the system's dynamics). We review the ideas of computational mechanics, an information-theoretic method for finding optimal causal models of stochastic processes, and argue that macrostates coincide with the "causal states" of computational mechanics. Defining a set of macrostates thus consists of an inductive process where we start with a given set of observables, and then refine our partition of phase space until we reach a set of states which predict their own future, i.e. which are Markovian. Macrostates arrived at in this way are provably optimal statistical predictors of the future values of our observables.

Keywords: Statistical mechanics, thermodynamics, macroscopic state, entropy, computational mechanics, causal state, memory effects, Markov processes, hydrodynamics, spin glasses, coarse graining, cellular automata, hysteresis, aging, nonequilibrium behavior, path dependence

Contents

*Electronic address: cshalizi@umich.edu
†Electronic address: moore@santafe.edu

I. WHAT'S STRANGE ABOUT MACROSTATES, OR, IS IT JUST ME?

Almost from the start of statistical mechanics, there has been a tension between subjective or epistemic interpretations of entropy, and objective or physical ones. Many writers, for instance the late E. T. Jaynes (1983), have vigorously asserted that entropy is purely subjective, a quantification of one's lack of knowledge of the molecular state of a system. It is hard to reconcile this story with the many physical processes which are driven by entropy increase, or by competition between maximizing two different kinds of entropy (Fox, 1988). These processes either happen or they don't, and observers, knowledgeable or otherwise, seem completely irrelevant. In a nutshell, the epistemic view of entropy says that an ice-cube melts when I become sufficiently ignorant of it, which is absurd.

These difficulties with entropy are only starker versions of the difficulties afflicting all thermodynamic macroscopic variables. Their interpretation oscillates between a purely epistemic one (they are the variables which we happen to be willing and able to observe) and a purely physical one (they have their own dynamics and have brute physical consequences, e.g. for the amount of work which engines can do). These difficulties are inherited by our definition of macrostates. Standard references define macrostates either as sets of microstates, i.e. subsets of phase space, with given values of a small number of macroscopic observables (Baierlein, 1999; Landau and Lifshitz, 1980; Reichl, 1980), or probability distributions over these (Balian, 1991; Ruelle, 1989). A given set of observables induces a set of macrostates, which form a partition of the phase space;[1] but why is one such partition better than another?

There are generally several different sets of macroscopic variables which can be observed a given system. In some cases, different sets of observables are equivalent, in the sense that they induce the same partition of the phase space, and so their macrostates are in one-to-one correspondence; for instance, for an ideal gas with a constant number of molecules, we obtain the same macrostates by measuring either pressure and volume, or temperature and entropy. In other cases, observing different sets of variables will partition the set of microstates in different ways — producing partitions that are finer, coarser, or incomparable.

Even if we restrict our attention to extensive variables, there is a hierarchy of increasingly disaggregated, fine-grained levels of description, with associated macroscopic variables at each level. At the highest and coarsest level are thermodynamic descriptions, in terms of system-wide extensive variables or bulk averages. Below them are hydrodynamic descriptions, in terms of local densities of extensive quantities. Below them is the Boltzmannian level, described with occupation numbers in cells of single-molecule phase space, or, in the limit, phase-space densities. (Below the Boltzmannian level we get densities over the whole-system phase space, and so statistical mechanics proper.) We can sometimes demonstrate, and in general believe, that we can obtain the coarser descriptions from the finer ones by integration or "contraction" (cf. Keizer, 1987, ch. 9). Thus there are many hydrodynamic macrostates for a given thermodynamic one, i.e. the hydrodynamic partition is much finer.

Clearly there is a problem here if macrostates are *purely* objective. In that case, we should be forced to use *one* level of description. On the other hand, we can formulate and test theories at all levels of description, and we know that, for instance, both thermodynamic and hydrodynamic theories are well-validated for many systems.

We hope to offer a resolution along the following lines. Intelligent creatures (such as, to a small extent, ourselves) start with certain variables which they are able to observe, and which interest them. This collection of variables defines a partition of phase space. This partition may not be an optimal predictor of its own future; it may have non-Markovian dynamics, with unaccounted-for patterns in its variables' time series. In such cases, intelligent observers postulate additional variables, attempt to develop instruments capable of observing them, and thus refine this partition. Our proposal is that good macrostates are precisely the partitions at which this process terminates, i.e. refinements of the observational states whose dynamics are Markovian. One can show that there is a unique coarsest such refinement, given the initial set of observables, and that this refinement is provably optimal in several senses as a statistical predictor of the future. To prove these things we must give a brief summary of the theory of the causal architecture of stochastic processes, known as computational mechanics.

Definitions of Markov processes and some basic information theory can be found in the Appendix.

II. COMPUTATIONAL MECHANICS AND CAUSAL STATES

In the late 1970s and early 1980s, workers in nonlinear dynamics developed methods, called "attractor reconstruction" or "geometry from time series," to reconstruct the vector field of a dynamical system from time series of measurements of some function of the state (Kantz and Schreiber, 1997; Packard, Crutchfield, Farmer and Shaw, 1980; Takens, 1981). Inspired by this, since 1989 Crutchfield et al. (Crutchfield and Feldman, 1997; Crutchfield and Shalizi,

[1] Tolman (1938, p. 77n), while objecting to the *name* "macrostate," regards it as a partition, as above.

1999; Crutchfield and Young, 1989; Feldman and Crutchfield, 1998; Shalizi and Crutchfield, 2001) have formulated a theory, "computational mechanics," which constructs, from observations of a stochastic process, the minimal model capable of generating that process. Put differently, they have developed a technique for discovering and representing all the predictive patterns in a time series.

The minimal model produced by computational mechanics represents the causal architecture of the process, or alternately how it stores and processes information (hence the term *computational* mechanics). If statistical mechanics is a "forward" approach, deriving macro-consequences of micro-dynamics, computational mechanics is an "inverse" approach, finding minimal causal architectures capable of producing the statistics of observed time series.

The key notion in computational mechanics is that of *causal state*, which works like this. We observe a stochastic process, which we break at arbitrary points into "histories" and "futures." Two histories belong to the same causal state if and only if they are equivalent for predicting the future, i.e., if they lead to the same conditional probability distribution for futures.

Formally, consider a discrete-time stochastic process, stretching to infinity in both directions: $\ldots s_{t-1}, s_t, s_{t+1} \ldots = \overleftrightarrow{s}$. Break the process into two parts: one, the "past" or "history," is all the values up to and including time t, which we write $\overleftarrow{s}_t$; the other, the "future" $\overrightarrow{s}_t$, all the values after that time. We write $\overleftarrow{S}$ and $\overrightarrow{S}$ for the set of all possible histories and futures respectively. If the system is in equilibrium and this time series is stationary, we can drop this subscript, and we assume this for now. We wish to predict $\overrightarrow{s}$ on the basis of $\overleftarrow{s}$.

Now, any prediction method treats some histories as equivalent to each other; for instance, if we model the system as only depending on its k previous values, two histories which last differed k time-steps ago are equivalent. Moreover, all we need to predict about the future is which equivalence class we will find ourselves in. Thus any prediction method induces a partition over the phase space, and for any partition there is some probability distribution over the future equivalence classes given which one we are in now.

There are many ways to construct partitions about which we can make correct predictions. For instance, we could lump all of phase space into a single, trivial "macrostate," and announce that given we are in that state now, we will be in the future as well. But obviously such a scheme fails to capture anything important about the system. Computational mechanics seeks a partition which is as coarse as possible, but which captures all available information about the time series of observables. That is, it gives optimal predictions of our observables while attributing as little structure to the system as possible.

With this in mind, we claim that the optimal partition is simply the following. Say that two histories $\overleftarrow{s}$, $\overleftarrow{s}'$ are *causally equivalent* if and only if they give the same conditional distribution for futures:

$$\overleftarrow{s} \sim_\epsilon \overleftarrow{s}' \text{ iff } \Pr[\overrightarrow{s}|\overleftarrow{s}] = \Pr[\overrightarrow{s}|\overleftarrow{s}']$$

This relation $\sim_\epsilon$ is symmetric, reflexive and transitive, and thus divides the set $\overleftarrow{S}$ of all pasts into equivalence classes. We define the *causal state* of a history as its equivalence class under $\sim_\epsilon$:

$$\epsilon(\overleftarrow{s}) \equiv \{\overleftarrow{s}' \,|\, \overleftarrow{s}' \sim_\epsilon \overleftarrow{s}\}$$

Then it is clear that the dependence of the system on its past is completely captured by its causal state,

$$\Pr[s_{t+1}|\overleftarrow{s}_t] = \Pr[s_{t+1}|\epsilon(\overleftarrow{s}_t)]$$

We now formalize our claim that the causal states form an optimal partition of histories. For a given partition $\mathcal{R}$ of histories into equivalence classes, we consider the mutual information $I[\overrightarrow{s};\mathcal{R}]$ between it and the system's future (for definitions of information-theoretic quantities, see Appendix A). This quantity is limited by the mutual information between the system's past and its future, $I[\overrightarrow{s};\mathcal{R}] \leq I[\overrightarrow{s};\overleftarrow{s}]$. We call partitions which attain this limit *prescient*. Shalizi and Crutchfield (2001) showed that the causal states are prescient. Moreover, if two histories are equivalent with respect to some prescient partition, they are almost always in the same causal state as well. Thus, except for a set of measure zero, any prescient partition is a refinement of the causal states, so the causal states form the coarsest possible prescient partition.

Moreover, the causal states are the *least complex* prescient partition in the following sense. Given a partition into equivalence classes $\mathcal{R}$ we define its *statistical complexity* as the entropy $H[\mathcal{R}]$. The statistical complexity is the amount of information which the partition encodes about the past. Normally in statistical mechanics we seek to *maximize* the entropy, but here entropy measures, not the unbiasedness of a distribution, but the complexity of a predictor, and so in minimizing it we are applying Occam's Razor. We can therefore say that the optimal predictor is the prescient predictor of minimal statistical complexity. In fact, for all prescient partitions $\mathcal{R}$ we have $H[\mathcal{R}] \geq H[\epsilon(\overleftarrow{s})]$, so the causal states are the unique prescient states of minimal statistical complexity (Shalizi and Crutchfield (2001), App. D of (Shalizi, 2001)).

Because the causal states are optimal predictors, we can say that the statistical complexity of the process is equal to their statistical complexity; we denote this $C_\mu = H[\epsilon(\overleftarrow{s})]$. The statistical complexity of the process is therefore

the amount of information about its past which is *relevant* to its future, the amount retained internally, so to speak. It also has a nice physical interpretation connected to thermodynamic entropy, which we give in the next section.

A common measure of the complexity of a stochastic process, and of the amount of information stored in it, is the predictive information (Bialek, Nemenman and Tishby, 2001), effective measure complexity (Grassberger, 1986) or "stored information" (Shaw, 1984), $\mathbf{E} \equiv I[\overrightarrow{s}; \overleftarrow{s}]$. Since the causal states are prescient, $I[\overrightarrow{s}; \overleftarrow{s}] = I[\overrightarrow{s}; \epsilon(\overleftarrow{s})] \leq H[\epsilon(\overleftarrow{s})]$ (see Appendix A), so $C_\mu \geq \mathbf{E}$. While a complete knowledge of the causal state reduces the uncertainty in the future by $\mathbf{E}$ bits, one needs C_μ bits to make this prediction. In general (Shalizi and Crutchfield, 2001), $\mathbf{E} \leq C_\mu$, and for the specific case of first-order Markov processes, $C_\mu = \mathbf{E} + h_\mu$, where $h_\mu \equiv H[s_{t+1}|\overleftarrow{s}_t]$ is the *entropy rate* of the process, also equal to $H[s_{t+1}|\epsilon(\overleftarrow{s}_t)]$. These notions will be useful later on, when we propose a definition of emergence based on the ratio between predictive information and statistical complexity.

Readers familiar with the literature on statistical explanation will recognize that the the causal state partition can also be though of as an application of Salmon's notion of a "statistical relevance basis" to stochastic processes (Salmon, 1971, 1984).

If we consider the time series of causal states $\ldots, \epsilon(s_{t-1}), \epsilon(s_t), \epsilon(s_{t+1}), \ldots$, we have a new stochastic process. Moreover, since each causal state $\epsilon(s_t)$ contains all the relevant information about its entire past $\overleftarrow{s}_t$, this is a Markov process, i.e. the probability distribution of futures depends only on the current state (see Appendix B). Thus we have collapsed the original process, regardless of its dependence on its history, into a Markov process — but one which contains all relevant information about the original process.

The observed process is a random function of this Markov process, i.e., a kind of "hidden Markov model". The Markov properties of the causal states justify in part their name, since they are exactly the "screening-off" properties that have long been recognized as essential to causation (Salmon, 1984), and which form the basis of statistical methods of causal inference for non-dynamical systems (Pearl, 2000; Spirtes, Glymour and Scheines, 2001).

If the underlying process $\overleftrightarrow{S}$ is not stationary, all is not lost. Formally, in fact, we can generalize the theory to arbitrary processes. For our purposes here, however, we only need the idea of a conditionally stationary process, which is to say one in which $\Pr(\overrightarrow{s}_t|\overleftarrow{s}_t = \overleftarrow{s}) = \Pr(\overrightarrow{s}_0|\overleftarrow{s}_0 = \overleftarrow{s})$, for all times t and histories $\overleftarrow{s}$. The above theory then carries over directly (stationary processes are all conditionally stationary as well), with the exception that the probability distribution of the causal states, and $H[\mathcal{S}]$, can be a function of time.

We have spoken throughout as we knew all the necessary conditional probabilities exactly. This is sometimes the case with analytical models (Feldman and Crutchfield, 1998), but never with experimental data. However, there are *reconstruction* algorithms which, under mild statistical assumptions, will converge to the correct causal states, given sufficient experimental data (Shalizi *et al.*, 2002). For our present purposes, it is enough to know that causal states can be inferred reliably from observations.

III. CAUSAL STATES FROM COARSE-GRAINED OBSERVATIONS

Consider our favorite statistical-mechanical system. It has a phase space Γ, every point $\mathbf{q}$ of which is a complete specification of the positions, momenta, spins, etc. of all particles. Discretizing time, as is common, we say that the evolution on Γ is governed by an operator T: $\mathbf{q}_{t+1} = \mathrm{T}\mathbf{q}_t$. We do not rule out the possibility that T is stochastic, but we insist that there are no "hidden variables," so that $\{\mathbf{q}_t\}$ form a Markov chain. We are concerned with an ensemble of such systems, so we write the random variable for the current microstate $\mathbf{Q}$. We do *not* assume that the ensemble is any of the usual thermodynamic ensembles, or even that the distribution of $\mathbf{Q}$ is invariant.

The system changes over time. We probe it with observations of limited precision at each time step. What our probes give us are many-one functions of the location in phase space. More formally, the observation process is represented by a function $f : \Gamma \mapsto \Xi$, where Ξ is our favorite (possibly multi-dimensional) space for representing observations[2]. This function f partitions the phase space Γ, i.e., it divides it into mutually exclusive and jointly exhaustive sets, on each of which f takes a unique value. Let the partition of Γ induced by f be $\mathcal{F}$. Then $f(\mathbf{Q}) = S_t$ defines another process, which need not be Markovian. Call it the *observed process*.

We now form the causal states of the observed process. Our observations, as noted, induce a partition of the phase space. Therefore a sequence of observations induces a refinement of that partition. Each observation value x corresponds to a set $\mathcal{F}_x$ of points in phase space. The sequence of observations x, y thus corresponds to the set $\mathcal{F}_{x,y} \equiv \mathcal{F}_y \cap \mathrm{T}\mathcal{F}_x$, i.e., those points at which we observe y now, and at which we would have observed x one step

[2] Strictly, f should map to a distribution of observed values, to represent fluctuations and noise, in which case macroscopic states would be defined by probability distributions of macroscopic variables. This, however, would increase the complexity of our exposition without a corresponding gain in insight, so we will pretend that our observations are exact and noiseless.

back. The sets $\mathcal{F}_{x,y}$ are a refinement of the observed partition $\mathcal{F}$, and we can extend this to countable sequences of observations. The set of causal states, $\boldsymbol{\mathcal{S}}$, is a partition on the set $\overleftarrow{S}$ of observational histories. Therefore it induces a partition on Γ which is a coarsening of the partition induced by infinite-length histories. Call this partition $\boldsymbol{\mathcal{G}}$. Each causal state therefore corresponds to a region of phase space, which in principle is accessible to some coarse-grained observational procedure. Observations of this variable form a new stochastic process $\{\mathcal{S}_t\}$ which is Markovian, and a knowledge of $\mathcal{S}_t$ is all that is needed to predict $\overrightarrow{s}_t$ optimally.

Let us write the partition on $\overleftarrow{S}$ induced by the present observation as $\mathcal{X}$. What is the relationship between the causal partitions, $\boldsymbol{\mathcal{G}}$ and $\boldsymbol{\mathcal{S}}$, and the corresponding observational partitions, $\mathcal{F}$ and $\mathcal{X}$? There are four possibilities:

1. The observational and the causal partitions are the same.
2. The causal partition is a refinements of the observational one.
3. The causal partition is coarser than the observational one.
4. The causal and observational partitions are incomparable.

In the next sections we explain the physical meaning of cases (1–3), and give physical examples of them.

A. The Observables Define a Macrostate

Suppose two observational histories, $\ldots s_{t-2}s_{t-1}s_t$ and $\ldots s'_{t'-2}s'_{t'-1}s'_{t'}$ are causally equivalent iff $s_t = s'_{t'}$. This means that the current causal state is defined by the current values of the macroscopic observables, and conversely any difference in a macroscopic observable means a difference in causal state. In the notation introduced earlier, $\boldsymbol{\mathcal{S}} = \mathcal{X}$ iff $\mathcal{F} = \boldsymbol{\mathcal{G}}$. The macrostates then have all the properties of causal states. Their dynamics are Markovian and statistically reproducible, and no prediction of future values of the macrovariables can be better than one based simply on their present value. An obvious example is the combination of pressure, volume and temperature for an ideal gas near equilibrium.

In such cases, we can give a nice interpretation to the statistical complexity C_μ. Recall that $C_\mu = H[\mathcal{S}]$, the amount of information needed to specify the causal state. Because the causal state and the macrostate S are equivalent, $H[\mathcal{S}] = H[S]$. But $S = f(\mathbf{Q})$, so $H[S|\mathbf{Q}] = 0$ — if we knew the exact microstate, there would be no uncertainty in the macrostate. Now, for any two random variables, $H[X,Y] = H[X] + H[Y|X]$. Let us make both possible decompositions of $H[\mathbf{Q}, S]$.

$$\begin{aligned} H[\mathbf{Q}|S] + H[S] &= H[S|\mathbf{Q}] + H[\mathbf{Q}] \\ H[\mathbf{Q}|S] + C_\mu &= H[\mathbf{Q}] \\ C_\mu &= H[\mathbf{Q}] - H[\mathbf{Q}|S] \\ C_\mu &= I[\mathbf{Q}; S] \end{aligned}$$

That is, the statistical complexity is just the amount of information about the microstate that is contained in the macrovariables.

Since the macrostates form a first-order Markov chain, there is, as we mentioned above, a simple relationship between the statistical complexity, the entropy rate, and the predictive information, viz., $\mathbf{E} = C_\mu - h_\mu$. Since $\mathbf{E} = I[\overrightarrow{s}; \mathcal{S}]$, and $h_\mu = H[S_1|\overleftarrow{s}]$, we have $I[\overrightarrow{s}; \mathcal{S}] = H[\mathcal{S}] - H[\overrightarrow{s}^1|\mathcal{S}]$.

B. The Causal States Are Finer than the Macrostates

Suppose $\boldsymbol{\mathcal{G}}$ is a refinement of $\mathcal{F}$, or, equivalently, $\boldsymbol{\mathcal{S}}$ is a refinement of $\mathcal{X}$. Then, in addition to knowing the current values of the macrovariables, we must know something of their history as well. Or, more exactly, if we do not, we do not have a causally complete set of macrovariables, and the observed dynamics are not only non-Markovian, they cannot be optimally predicted. However, they *can* be optimally predicted from a knowledge of $\mathcal{S}$, whose time-evolution *is* Markovian. Moreover, if we know what cell of $\boldsymbol{\mathcal{G}}$ the system is in, we know the value of $\mathcal{S}$, which suggests that, in principle, there is a observational procedure which will tell us how to optimally predict our original macrovariables.

We can go one step further, however, by invoking a result about refinements of partitions (see the appendix). Suppose A is a partition and B is a refinement of it. Then there exists at least one *minimal factor partition* C such that B is the product of A and C, and this is not true of any partition with fewer cells than C. Since $\boldsymbol{\mathcal{G}}$ is a refinement of $\mathcal{F}$, there is therefore at least one $\mathcal{Z}$ such that $\boldsymbol{\mathcal{G}} = \mathcal{F} \cdot \mathcal{Z}$. If we observe the macrovariable Z corresponding to $\mathcal{Z}$

together with our original macrovariables, it is the same as if we had observed the causal state directly, and so we get a causally complete set of macrovariables and nice macrostates. In other words, we if observe either $\mathcal{S}$ or (S, Z), we reduce the present case to that in the previous subsection.

Generally, there are a very large number of minimal factor partitions which can take the role of $\mathcal{Z}$. Which one we observe is dictated by practical considerations — experimental accessibility, smoothness of the resulting macrovariable over phase space, degree of uncertainty in the macrovariable, etc. This should not be worrisome, however, since there are elementary cases where we can complete a set of macrovariables in more than one way. Given pressure and volume for an ideal gas, for instance, we get the same macrostates from observations of temperature or molecule number.

It is worth noting that the minimal factor partition $\mathcal{Z}$ is incomparable to $\mathcal{F}$, and so in some sense orthogonal or unpredictable from $\mathcal{F}$. Clearly, there is a bijection between $\mathcal{S}$ and (S, Z). Hence $H[\mathcal{S}] = H[S, Z] = H[Z|S] + H[S]$. Since $H[S]$ does not depend on our choice of factor variable Z, it follows that $H[Z|S]$ is the same for all factor variables, i.e., they all have the same degree of uncertainty remaining once we know the original observables. Furthermore, all the information they contain is relevant to the causal state:

$$\begin{aligned} I[\mathcal{S}; Z] &= H[\mathcal{S}] - H[\mathcal{S}|Z] \\ &= H[S, Z] - H[S, Z|Z] \\ &= H[S, Z] - (H[S, Z, Z] - H[Z]) \\ &= H[S, Z] - H[S, Z] + H[Z] \\ &= H[Z] \end{aligned}$$

Similarly, $I[\mathcal{S}; Z|S] = H[Z|S]$, which is independent of the factor partition we use.

Sadly, there is no guarantee that any of the factor partitions are experimentally accessible, still less accessible by practical or easy experimental procedures. In such cases, however, we may still eliminate memory effects from our models by constructing the causal states from observational histories.[3]

For an example of this method in (unwitting) action, consider hysteresis in ferromagnets. The response of a ferromagnetic substance to a magnetic field can be treated, equivalently, either as a function of its past history of applied fields, or as a function of the current applied field and the magnetization. Another example is provided by the study of chaotic dispersion in fluids jets (Cencini *et al.*, 1999). The initial measurement partition here involves the character of the motion of the jet, and shows strong memory effects, significantly complicating the analysis. Recent work has shown how to eliminate these memory effects, by refining the partition of the state space in just the way we suggest above (Abel *et al.*, 2002; Lacorata *et al.*, 2001).

1. An Apparent Counterexample: Disordered Materials

Amorphous solids (Zallen, 1983) and their magnetic equivalents, spin glasses (Fischer and Hertz, 1988) are remarkable not just because of they display slow dynamics, but because they display distinct dynamics on an immense range of time-scales. A crude but graphic illustration is given by ordinary silicate glass. Under mechanical stresses with short characteristic times, it is brittle; under stresses with long characteristic times, it is effectively liquid. Spin glasses, similarly, can display distinct susceptibilities to oscillatory magnetic fields over sixteen orders of magnitude in frequency (Fischer and Hertz, 1988). Since the two cases are basically similar, but the requisite physical theory is easier to grasp for spin glasses, we will concentrate on them.

This hierarchy of time-scales implies that memory effects are very important in disordered materials. Indeed, many of the usual assumptions made in discussions of statistical mechanics, such as having an "aged" ensemble at equilibrium, are simply nonsensical in these cases. At low temperatures, the slowest time-scales can be geologically significant. To work with samples which have aged into equilibrium requires literally inhuman longevity (to say nothing of patience). While technologies which would allow this have been proposed (Dyson, 1979), they are not yet common in laboratories. Have we here found substances where our approach to eliminating memory effects breaks down? And what does one do, if one cannot use properly aged and equilibrated ensembles?

The physical mechanism responsible for the long time-scales actually holds the answer to both questions. Each spin in a spin glass participates in a mixture of ferromagnetic and antiferromagnetic interactions of varying strengths. The result is generally frustration, i.e., no setting of the spins minimizes all interaction energies simultaneously. This leads to the existence of numerous local minima in the energy landscape, generally with widely varying energies, and so widely varying heights of the barriers separating them. One must either flip many spins at once, or equivalently

[3] For more on reducing dialectical or historical explanations to mechanical ones via causal states, see Shalizi (2001).

make many energetically-unfavorable spin flips in succession, to get from one minimum to another. The time it takes to pass between minima will generally be exponential in the height of the energy barrier between them, as one expects from the Arrhenius equation. (The causes and details of frustration in glass are different, but the overall picture is similar.) Thus, on a given time-scale, barriers above a certain height are effectively infinite, i.e., there probability of crossing them is negligible. The spin glass is thus effectively confined to a fixed region of phase space. Within this region, the local minima define metastable states, with characteristic life-spans, and so relative probabilities, that reflect the heights of the barriers surrounding them.

We can thus see the way to eliminating memory effects: one takes as one's macrovariables the occupancy probabilities of the metastable local minima. Those in the effectively-inaccessible region do not contribute. Within the accessible region of phase space, there is a more-or-less gradual leakage of probability from the initial metastable state to the others. To extrapolate this forward, however, we do not need to know the history of that seepage, merely the current distribution over the local minima. In fact, this is a common theoretical ploy, sometimes spoken of as employing "a macroscopic number of macroscopic degrees of freedom" (Fischer and Hertz, 1988). Experimentally, one never studies an equilibrium ensemble, but rather one that is always aging, and it is precisely the aging properties which are of interest!

C. The Macrostates Are Finer than the Causal States

Suppose $\mathcal{X}$ is a refinement of $\boldsymbol{\mathcal{S}}$. Then some distinct values of the macrovariables have exactly the same consequences for the future evolution of the macrovariables. The distinction between those macrostates is meaningless, and some of the details in those macrostates is superfluous. There are several reasons, by no means mutually exclusive, why this might be so.

First, some of our variables could be irrelevant, given the others. More precisely, future events could be statistically independent of the value of variable Y given the present value of other variables X. It is hard to find examples of this in statistical mechanics proper, simply because those variables have been subject to a long process of (informal) selection for relevance, but it is easy to find examples of this in other domains of scientific inquiry. Techniques for identifying, or constructing, combinations of variables which render others irrelevant play a major role in statistical methods of causal inference (Pearl, 2000; Spirtes, Glymour and Scheines, 2001). (Note that if one macrovariable is a deterministic function of the others, then we get the same partition of Γ whether or not we adjoin it to the others. Similarly, the partition of histories we get is the same.)

Second, our observational procedure could encode an "unphysical" distinction. In nematic liquid crystals, for instance, an important role is played by the "director", a local vector indicating the average direction of orientation of the rod-shaped molecules in the neighborhood of a point. However, the molecules in a nematic are symmetric when their long axis is inverted, so the director is not a normal vector, but one in which opposite vectors are identified, i.e., $\mathbf{n} = -\mathbf{n}$. (Collings, 1990; de Gennes and Prost, 1993). If we did not know this, however, and tried to observe the director as an ordinary vector, we would find that which of two opposite observations we got for the director would be a matter of pure chance, i.e., an artifact, and that we would retain full predictive power if we identified opposite director vectors, i.e., if we coarsened our observational partition.

Finally, we may have an unpredictable variable, in the following sense. On the one hand, it takes on significantly different values in regions of the phase space which are visited under the dynamics. On the other hand, given the time scale separating our observations, the dynamics randomizes those values so thoroughly that little or no effective prediction of the variable is possible. In these cases, the variable "washes out" from the partition which maximizes predictive power, namely $\boldsymbol{\mathcal{G}}$. In extreme cases, none of the variables has any predictive power, at the time-scale and resolution available to us, so the observed process becomes a sequence of IID random variables, and $\boldsymbol{\mathcal{G}}$ becomes the trivial partition on Γ. For example, consider a liter of ideal gas at standard temperature and fixed, normal molecule number. If we observe pressure and internal energy (to reasonable precision) at intervals of one year, the dynamics will have so thoroughly mixed phase space that our original observations will have absolutely no predictive value at all. (cf. Shalizi, 2001, ch. 12). In such cases, there is simply no point in making predictions, and one's resources are better used elsewhere.[4]

---

[4] In the real world, it is often not obvious when a variable contains predictive information, at least at the time- and resolution- scale of interest, and great efforts can be devoted to ever-more-elaborate deterministic models of what are, to all intents and purposes, coin-flips. For an example of causal state reconstruction showing that some variables contained no useful information, and how recognition of this led to improve predictions, see Palmer, Fairall and Brewer (2000). Of course, variables which are effectively IID over long times or at coarse resolution can contain a lot of predictive information at finer scales. We will return to this point later; cf. the hierarchical scaling complexities of Badii and Politi (1997).

D. The Physical Meaning of the Causal States

Starting from the observational variables, one can construct the the causal states, and from them the minimal coarse-grained observation which allows for optimal prediction of the original observables. If the two do not coincide, one can profitably replace the original set of observables with a new one, either by adjoining new observations or by eliminating unphysical distinctions or variables without predictive power. In any case, we can construct, from the original macrovariables, a new set of macrovariables whose macrostates are their own causal states.

These well-constructed macrostates have a number of properties it is worth noting. First, their statistical complexity is just the amount of information the macrovariables contain about the microstate — how much our uncertainty about the microstate is reduced by learning the macrostate. Second, the macrostates are Markovian. This means that they will be mixing just when they satisfy the conditions for Markov processes to be mixing. This can be true even when T is *not* mixing. Third, again because the macrostates are Markovian, there is a Gibbs distribution over sequences of macrostates (Brémaud, 1999). We have not had to assume any sort of equilibrium property, however, and this may be part of the reason why Gibbs distributions are still useful out of equilibrium.[5]

We began with certain arbitrary or subjective decisions, about which variables to observe — about what partition $\mathcal{F}$ of Γ to employ. Our desire to have dynamics with good causal properties (Markovianity, etc.) led us first to refine that partition by considering observational histories, and then to group together histories in constructing the causal states. Whether, at that point, we end up adjoining new variables to our original macrovariables, leaving them alone, or even coarsening them, has nothing to do with our experimental decisions or epistemic hankerings, merely the purely mechanical, physical, objective microdynamics. In the causal states we have arrived, so to speak, at objective explanations of subjective quantities. (cf. Crutchfield and Shalizi, 1999)

IV. LEVELS OF DESCRIPTION AND EMERGENCE

Earlier, we raised the puzzle of how different levels of description of the same system can co-exist. The answer, we propose, is that different causal states are induced by different measurement partitions. Consider two measurement partitions, one a coarsening of the other. The coarser measurement is therefore a function of the finer one, and is on a higher, less specific level of description. Suppose the finer, lower-level measurement partition is causal. It is well-known that the Markov property does not generally survive coarsening the states, which means that its coarse-grainings will not, in general, be their own causal states. The causal states of the coarse-grained measurements are well-defined, however, and cannot be any finer than the states of the fine-grained measurement partition. It is possible, however, that one does not need to go all the way back to the original partition to find those causal states — in fact, there is no reason the coarse-grained measurements cannot be identical to their own causal states. We then have two levels of description, and can give a coherent causal account at each level.

In this section, we explore the uses of these ideas in interpreting statistical mechanics, and suggest a definition of emergence. We start by considering the old question of the relationship between molecular dynamics and thermodynamics, in the particularly transparent context of the fluctuations of a gas at equilibrium. This leads us to suggest a definition of "emergence". We then clarify the relationship between generalized hydrodynamics and thermodynamics, and attempt to explain the ubiquity of Gibbs distributions for macroscopic configurations. Finally, we look at the practice of cellular automata and lattice gas modeling for examples of deliberately constructing adequate coarse-grainings.

A. Equilibrium Fluctuations and a Definition of Emergence

Systems prepared "in equilibrium" actually fluctuate continually. If our observations are sufficiently coarse, then we will essentially only see fluctuations about equilibrium which leave us in the linear regime. In that case, the Onsager theory provides the tools to describe the fluctuations, and to do so in terms of the same variables which work at equilibrium (Keizer, 1987).

Consider a cubic centimeter of argon at fixed temperature, pressure and number. (For a more detailed version of what follows, see Shalizi (2001, sec. 11.2.3)). The only macrovariable left to fluctuate is the internal energy. One can calculate, from the Onsager theory, that the Shannon entropy of the internal energy is 33.3 bits. Taking a time-step of one millisecond, the entropy rate, i.e., the rate at which the uncertainty increases, is 4.4 bits. The predictive

[5] We owe this last suggestion to conversation with Erik van Nimwegen, but are pretty sure he disagrees.

information is thus $33.3 - 4.4 = 28.9$ bits. In doing the corresponding calculations for the microstates, we start with the fact that the microstate is its own causal state, since (almost by definition) it is Markovian. Thus $C_\mu = 6.6 \cdot 10^{20}$ bits. If we take the time-step to be one nanosecond, one can estimate h_μ (Gaspard, 1998) to be $3.3 \cdot 10^{20}$ bits, with $\mathbf{E} = 3.3 \cdot 10^{20}$ bits.

Following Palmer (2001), we define the predictive efficiency of a process as the fraction of the information it contains which actually effects the future, i.e., as the ratio $\mathbf{E}/C_\mu$. We then see that the macrostates can be predicted with much higher efficiency (0.87) than the microstates (0.5). Indeed, this comparison is rather unfair to the macrostates, since we are predicting them over a much longer time-scale. If we predicted them at the same time resolution as the microstates, we would find that the efficiency of prediction was essentially one. Conversely, if we tried to predict the microstates at the macro time-scale, we would find an efficiency of prediction of essentially zero. Yet the macrovariables are transparently a function, a coarse-graining, of the microvariables.

This leads us to define a relation of "emergence" between two sets of causal variables if (1) one is a coarse-graining of the other and (2) the coarse-grained variables can be predicted more efficiently. In this sense, we can be precise about the long-standing intuition that thermodynamics emerges from statistical mechanics: thermodynamic variables are more informative about their own dynamics. This also gives us a hint as to what constitutes a *good* set of macrovariables: it should not just be causally complete, but also more predictively efficient than the microvariables. This is not always the case; sometimes coarse-grainings are less efficiently predictable than the original variable, a condition which J. P. Crutchfield (personal communication) has designated "submergence".

B. Hydrodynamics and Levels of Description

One of the more important developments in statistical mechanics and condensed matter physics has been the rise of "generalized hydrodynamics," where description centers on the local densities of extensive quantities and order parameters (Chaikin and Lubensky, 1995; Forster, 1975). Normal hydrodynamics is included as a special case. We are not going to expound this theory, interesting though it is. Rather, we wish to draw out two points.

The first is that many (perhaps all) systems which are adequately described at the hydrodynamic level can also be described, accurately but less precisely, at the thermodynamic level. This is perfectly sensible from our point of view. If one starts with observations of local densities, it is extremely unlikely that these will be adequately predicted from purely global quantities. The causal states one forms remain, therefore, tied to local densities. Conversely, knowledge of the local densities is excessive if all you want to predict are their global averages or sums. The two descriptions coexist, because they are intended to answer different questions — not because one is more objective than the other.

Second, one can show (Shalizi, 2001) that the relationship between the hydrodynamic description and the thermodynamic one is generally one of emergence, in the sense described above. This is comforting, since one can generally "contract" hydrodynamic descriptions into thermodynamic ones (Keizer, 1987). Similarly, the hydrodynamic level itself emerges from the thermodynamic one. Third, when one constructs local causal states (following Shalizi, 2001), one finds that they generally form a Markov random field.[6] Consequently, there is a Gibbs distribution over their configurations. Now, there are many examples of hydrodynamic systems, strongly non-equilibrium in their (standard) thermodynamics, where there are nonetheless important objects which follows Gibbs distributions with various kinds of effective interaction potentials (Cross and Hohenberg, 1993). Perhaps the most striking case of this is vortex lines in turbulent fluids — see Chorin (1994) for a full treatment. For conventional statistical mechanics, this is just so much dumb luck, but from our point of view, it indicates that the vortex lines (or other coherent, structuring objects) are the local causal states. Or rather: if what we find doesn't look Gibbsian, it means we can do better.

C. Building Coarse Grainings in Cellular Automata and Lattice Gases

Cellular automata and lattice gases are fully-discretized classical field theories. That is, time is discrete, space is a discrete regular lattice, and each point or "cell" can take one of a finite number of states at any one time. The state of each cell at time $t+1$ is a fixed, possibly stochastic, function of the state of the cell at time t, along with the states of the cells in a fixed "neighborhood," thus preserving the nice classical property of local interaction. (All cells update in parallel.) Originally introduced to model mechanical self-reproduction (Poundstone, 1984; von Neumann, 1966), cellular automata have proved useful as models of many natural phenomena (Chopard and Droz, 1998; Gutowitz, 1991; Rothman and Zaleski, 1997), as well as mathematically fascinating objects in their own right (Griffeath and

[6] No counter-examples are known. Whether they always form a Markov random field is currently an open question.

Moore, forthcoming). They are important to us here because (1) dynamic models of spin systems are stochastic CAs, and (2) they illustrate the strategy we are advocating, at least as a matter of tacit practice.

When one simulates a CA, one knows, exactly, both the underlying microstate and its dynamics. It can nonetheless be very hard to say what it will do and why it will do it. This is not simply because some CA have high computational complexity (Burks, 1970; Griffeath and Moore, 1996; Moore, 1997). Rather, it is because the raw microstate is too detailed to be of use — C_μ is much too high. One gains understanding by *deliberately* throwing away most of the microscopic information, finding instead coarse-grained observations where the dynamics are simpler to grasp (Crutchfield, 1992). Generally, this means constructing macrostates with well-behaved Markovian dynamics. There are numerous examples of this strategy in the literature, including the many derivations of hydrodynamics in lattice gases (Rothman and Zaleski, 1997), the theory of heat conduction in the Creutz CA (Saito *et al.*, 1999), or the vortex dynamics of the zero-temperature Potts model (Moore, Nordahl, Minar and Shalizi, 1999). The goal, always, is to throw away as much detail as possible, while retaining information relevant to certain aspects of the large-scale dynamics — to find simple but accurate representations. (Simple, inaccurate representations are of course easy to find.) Spatial computational mechanics (Crutchfield and Hanson, 1993; Hanson, 1993; Hanson and Crutchfield, 1997; Hordijk, Shalizi and Crutchfield, 2001) provides tools whereby one can automatically find and filter out low-information patterns, concentrating one's attention on higher-level, information-rich emergent structures.

V. WHAT IS NOT BEING SAID

There are a number of puzzles about macrovariables which our arguments do *not* resolve.

1. Why are so many good macrovariables extensive quantities? It certainly does not seem to follow from the fact that complete sets of macrovariables are causal states. We suspect, however, that something could be made of the following sketch of an argument. Extensive variables, by definition, add across sub-systems. If those sub-systems are independent, or nearly so, then their totals will have large deviation properties (Ellis, 1985, 1999). In other words, they will become increasingly well-behaved, statistically, in large systems. This makes them good candidates for experimental observation. Conceivably, there are many extensive variables, other than the ones we commonly observe, which, while subject to large-deviation principles for their additive fluctuations, are ill-behaved (non-Markovian) over time, and so we ignore them. Conversely, when approximate independence across sub-systems is violated, the good macrovariables are non-extensive, and we need Tsallis statistics (rather than the usual central-limit-theorem statistics).

2. Why are almost all good macrovariables derivatives of thermodynamic potentials? A deflating answer would be that candidate thermodynamic potentials are under intense selection pressure for just this property.

3. Why are some good macrovariables reusable, e.g., why is temperature a good macrovariable for almost everything? It is not simply that we can only observe a few variables and so observe them for almost everything, because order parameters, for instance, are generally not reusable (no sense measuring the nematic director in an antiferromagnet). Moreover, why do these variables typically take similar roles in systems with radically different microphysics? (E.g., temperature again.) To the best of our knowledge, no one has an answer to these questions; we certainly don't.

Finally, our approach does not say why it is legitimate to treat a large, locally unstable mechanical system stochastically in the first place. We have simply assumed that, since we are considering coarse-grained observations, it is legitimate to deal with them statistically. While the maximum entropy principle provides a superficially attractive justification for this, it is open to grave philosophical objections (Guttmann, 1999; Sklar, 1993). Worse, Amari (2001) has shown that maximum entropy distributions are simply those that minimize the degree of statistical dependence between variables.[7] If distributions evolve towards minimal dependency, that is surely just a contingent fact about the dynamics, rather than a universal principle of inference. We believe that the answer lies rather in ergodic theory (Khinchin, 1949; Mackey, 1992), particularly recent developments which emphasize the rapid mixing of low-dimensional projections of high-dimensional smooth dynamics (Dorfman, 1998; Gaspard, 1998; Ruelle, 1999), plus the philosophical assumption that the initial conditions of the world are "generic".

---

[7] Intuitively, this makes sense. The entropy, in bits, is the minimum mean number of binary variables needed to specify a sample drawn from the distribution. Dependencies can be used to shorten the description, hence maximizing the entropy requires minimizing dependencies. Showing this in detail requires an excursion through information geometry, and we refer the interested reader to Amari's paper.

VI. CONCLUSION

Let us recap by way of telling a story. Nameless men in black approach us, as reputable practitioners of statistical mechanics, with a physical system, in this case, a beaker full of gooey, shiny black stuff that sometimes moves spontaneously. We are able to probe certain aspects of it by physically coupling to it — e.g., we can X-ray it, take photographs, attach voltmeters, scatter neutrons through it. The men in black want us to predict certain properties of the black oil, say, what will cause it to quiver in different ways. A mixture of interest and feasibility thus dictates an initial choice of macrovariable. Given this, we attempt to refine our predictive capabilities by considering histories of observations. From them we construct causal states. If the causal states do not coincide with our initial observables, we either supplement them with new variables, in a way which can be determined from the causal states construction; or we eliminate unphysical distinctions and unpredictive variables, again on the basis of the causal states. When we need supplementary variables, we can either devise new experimental methods to observe them, probing new aspects of the physics, or we can merely construct them logically, from histories of our original observables. At the end of this process we have the minimal set of variables from which we can optimally predict the macrovariables of interest; ones which are, moreover, causally complete.

Our initial choice of macrovariables is the product of our ability to observe the system, and our choices about what to predict. Beyond that initial choice, and the requirement that good macrostates have certain causal properties, the causal state we use are completely out of our control, fixed entirely by the objective, microphysical dynamics. A different set of initial variables will, generally, lead to a different set of causal states. Sometimes, but not always, these causal states are related in a hierarchy of emergence. One might put it like this: for every question we ask It, Nature has a definite answer; but Nature has no preferred questions.

Acknowledgments

We thank David Albert, Sven Brueckner, Jim Crutchfield, Michael Lachmann, Erik van Nimwegen, Scott Page, Jay Palmer and Peter Stadler for helpful discussions. CRS thanks Prof. Crutchfield for generous support and encouragement, Kara Kedi for providing a reason to get out of bed each morning, and Kristina Shalizi for being perfect. His work was supported at SFI by the Dynamics of Learning project, under DARPA contractual agreement F30602-00-2-0583, and at the University of Michigan by grant from the James S. McDonnell Foundation. CM thanks Itamar Pitowsky for organizing the International Conference on the Conceptual Foundations of Statistical Mechanics in Jerusalem in May 2000, and Michel Morvan and the Laboratoire d'Informatique Algorithmique Fondements et Applications (LIAFA) at Université Paris 7 for a visit during which some of this paper was written. He is supported by NSF grant PHY-0071139, the Sandia University Research Program, and Los Alamos National Laboratory. He also thanks Tracy Conrad and Spootie the Cat for fleeting glimpses of enlightenment

APPENDIX A: Information Theory

The information contained in a discrete random variable X, also called its *entropy* or *Shannon entropy*, is

$$\begin{aligned} H[X] &\equiv -\sum_x \Pr(X=x)\log_2 \Pr(X=x) \\ &= -\langle \log_2 \Pr(X) \rangle \end{aligned}$$

It is the smallest number of bits (binary distinctions) needed, on average, to specify the value of X. We may think of it as the uncertainty an ideal observer, who knew the true ensemble X is drawn from, would have about X.

The *joint entropy* of two variables X and Y is defined similarly,

$$H[X,Y] \equiv -\sum_{x,y} \Pr(X=x, Y=y)\log_2 \Pr(X=x, Y=y)$$

It is easy to show that $H[X,Y] \leq H[X] + H[Y]$, with equality if and only if X and Y are statistically independent.

The *conditional entropy* of X given Y, is

$$\begin{aligned} H[X|Y] &\equiv H[X,Y] - H[Y] \\ &= -\sum_y \Pr(Y=y) \sum_x \Pr(X=x|Y=y)\log_2 \Pr(X=x|Y=y) \\ &= \sum_y \Pr(Y=y) H[X|Y=y] \end{aligned}$$

Then $H[X|Y=y]$ is the information needed to specify X in the sub-ensemble where Y has the value y, and $H[X|Y]$ is the average information remaining in X given Y.

Finally, the *mutual information* between X and Y is

$$\begin{aligned} I[X;Y] &\equiv H[X] + H[Y] - H[X,Y] \\ &= H[X] - H[X|Y] \\ &= H[Y] - H[Y|X] \;\leq\; H[Y] \;, \end{aligned}$$

that is, the amount by which knowledge of one variable reduces the uncertainty in the other.

APPENDIX B: Markov processes

Suppose a process generates a probability distribution over time series, $\ldots, s_{t-1}, s_t, s_{t+1}, \ldots$. It is a (first-order) *Markov process* if

$$\Pr(s_{t+1} = s | \overleftarrow{s}_t = \ldots, s_{t-1}, s_t) \;=\; \Pr(s_{t+1} = s | s_t) \;.$$

In other words, the only dependence that s_{t+1} has on its entire past history is on its current state s_t, and previous values yield no additional information about its future. Examples of Markov processes include:

- A deterministic dynamical system where $s_{t+1} = f(s_t)$ for some function f
- A series of fair coin flips, where $\Pr(s_{t+1} = s) = 1/2$ for $s \in \{\text{heads}, \text{tails}\}$
- Brownian motion, where $\Pr(s_{t+1} = x | s_t = y) = f(|x-y|)$ for a Gaussian function f

As an example of a non-Markovian process, suppose s_t is the position at time t of a particle which is moving at constant velocity. Here s_{t+1} depends on s_t and s_{t-1}, i.e. its dynamics is second-order, so there are correlations with the past that are not captured by the current state. On the other hand, if we expand our set of observables so that s_t includes both the particle's position and its velocity, then the process becomes first-order Markovian.

APPENDIX C: Partitions

A *partition* P of a set Ω is a set P_i of subsets of Ω which are mutually exclusive and jointly exhaustive. That is, $P_i \cap P_j = \emptyset$ (unless $i = j$), and $\Omega = \bigcup_i P_i$. The sets in P are the *cells* of the partition.

An *equivalence relation* or *equivalence* $\sim$ on Ω is a relation which is reflexive, symmetric and transitive: $a \sim a$, $(a \sim b) \Leftrightarrow (b \sim a)$ and $(a \sim b) \wedge (b \sim c) \Rightarrow (a \sim c)$. The *equivalence class* of a point is the set of all points which are equivalent to it. We write the equivalence class of x as $[x]$; $[x] = \{y | x \sim y\}$. Since every point is equivalent to itself, every point has a non-empty equivalence class, and every point belongs to some equivalence class.

Proposition 1 *Every partition corresponds to an equivalence relation, and vice versa. Equivalence classes are cells of the partition.*

Proof. First, we construct an equivalence relation from a partition. Simply say that $x \sim y$ iff x and y are in the same cell. This is symmetric, reflexive and transitive, hence an equivalence. Now we build a partition from an equivalence relation. We claim that the equivalence classes are mutually exclusive and jointly exhaustive. Mutually exclusive means that either $[x] = [y]$ or $[x] \cap [y] = 0$, for all x and y. To see this, consider any point $z \sim x$. Now, $y \sim z$ if and only if $y \sim x$ — if by transitivity, and only if likewise. Hence $x \sim y$ iff $[x] = [y]$. If $x \not\sim y$, then there cannot exist even one z such that $z \sim x$, and so no point belongs to the intersection of $[x]$ and $[y]$. Since every point has an equivalence class, the set of equivalence classes is exhaustive. QED.

Every function f on Ω induces an equivalence relation $\sim_f$, thus: $a \sim_f b$ iff $f(a) = f(b)$. Similarly every equivalence relation defines an (infinite) class of functions: give each equivalence class a unique label and map points to their equivalence-class labels. Hence every function defines a partition and vice versa.

The *identity* partition is one where each cell contains only a single element of Ω, i.e., where each equivalence class consists of a single point. The *trivial* partition is the one which contains only a single cell, equal to Ω.

One partition, A, is *finer* than another, B, iff for each $a \in A$, there exists a $b \in B$ such that $a \subseteq b$, and, for at least one a, $a \subset b$. Then A is a *refinement* of B, and B is *coarser* than A. Refinement always increases cardinality. If neither A nor B is a refinement of one another, and they are not equal, they are *incomparable*.

Let A and B be two partitions. Construct all the sets formed by taking the intersection of one cell from A with one cell from B. This collection is also a partition, the *product* of A and B. Symbolically, $A \cdot B = \{c | \exists a \in A, b \in b, c = a \cap b\}$. It is a refinement of both A and B.

Proposition 2 (Factoring Refinements) *Let A be a partition and B be any refinement of A. Then there exists a* minimal factor partition *C such that $B = A \cdot C$ and $B \neq A \cdot D$ for any D with fewer cells than C. If A and B are both finite, then the minimal factor partitions are themselves finite, and there is a finite number of them.*

Proof. We construct a minimal factor C. Each cell a of A contains n_a cells from B; $a = \bigcup_{i=1}^{n_a} b_i^a$. Let N be the maximum of n_a over all the cells of A. Define $c_i = \bigcup_{a \in A} b_i^a$. Clearly $C = \{c_i\}$ is a partition, and equally clearly its product with A will be B. Any partition whose product with A is B must have a cardinality of at least N, because it must break (at least) one cell of A into N sub-cells. Hence we have constructed a minimal C whose product with A gives the desired refinement. Moreover any D such that $B = A \cdot D$ must be a refinement of some minimal factor C. The number of minimal factors is at most the number of ways of labeling the subcells of A, viz., $\prod_{a \in A} n_a!$.

Note that minimal factors are incomparable to A. They are not refinements, because each cell of the factor is not entirely contained within a single cell of A. But conversely, there are cells in A which are not entirely contained within a single cell of the factor. Obviously, they are not equal to A. Hence they are incomparable.

References

Abel, M., K. H. Andersen and Guglielmo Lacorata (2002). "Hierarchical Markovian modeling of multi-time scale systems." *Physica D*, **submitted**. URL `arxiv.org/abs/nlin.CD/0201027`.

Amari, Shun-ichi (2001). "Information Geometry on Hierarchy of Probability Distributions." *IEEE Transacttions on Information Theory*, **47**: 1701–11. URL `www.islab.brain.riken.go.jp/~amari/pub/IGHI.pdf`.

Badii, Remo and Antonio Politi (1997). *Complexity: Hierarchical Structures and Scaling in Physics*, vol. 6 of *Cambridge Nonlinear Science Series*. Cambridge: Cambridge University Press.

Baierlein, Ralph (1999). *Thermal Physics*. Cambridge, England: Cambridge University Press.

Balian, Roger (1991). *From Microphysics to Macrophysics: Methods and Applications of Statistical Physics*. Berlin: Springer-Verlag.

Bialek, William, Ilya Nemenman and Naftali Tishby (2001). "Predictability, Complexity and Learning." *Neural Computation*, **13**: 2409–2463. URL `arxiv.org/abs/physics/0007070`.

Brémaud, Pierre (1999). *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*, vol. 31 of *Texts in Applied Mathematics*. Berlin: Springer-Verlag.

Burks, Arthur W. (ed.) (1970). *Essays on Cellular Automata*, Urbana. University of Illinois Press.

Cencini, Massimo, Guglielmo Lacorata, Angelo Vulpiani and Enrico Zambianchi (1999). "Mixing in a Meandering Jet: a Markovian Approximation." *Journal of Physical Oceanography*, **29**: 2578–2594. URL `arxiv.org/abs/chao-dyn/9801027`.

Chaikin, Paul M. and T. C. Lubensky (1995). *Principles of Condensed Matter Physics*. Cambridge, England: Cambridge University Press.

Chopard, Bastien and Michel Droz (1998). *Cellular Automata Modeling of Physical Systems*, vol. 6 of *Collection Aléa Saclay*. Cambridge, England: Cambridge University Press.

Chorin, Alexandre J. (1994). *Vorticity and Turbulence*, vol. 103 of *Applied Mathematical Sciences*. Berlin: Springer Verlag.

Collings, Peter J. (1990). *Liquid Crystals: Nature's Delicate Phase of Matter*. Princeton, New Jersey: Princeton University Press.

Cross, Mark C. and Pierre Hohenberg (1993). "Pattern Formation Out of Equilibrium." *Reviews of Modern Physics*, **65**: 851–1112.

Crutchfield, James P. (1992). "Semantics and Thermodynamics." In *Nonlinear Modeling and Forecasting* (Martin Casdagli and Stephen Eubank, eds.), vol. 12 of *Santa Fe Institute Studies in the Sciences of Complexity*, pp. 317–359. Reading, Massachusetts: Addison-Wesley.

Crutchfield, James P. and David P. Feldman (1997). "Statistical Complexity of Simple One-Dimensional Spin Systems." *Physical Review E*, **55**: 1239R–1243R.

Crutchfield, James P. and James E. Hanson (1993). "Turbulent Pattern Bases for Cellular Automata." *Physica D*, **69**: 279–301.

Crutchfield, James P. and Cosma Rohilla Shalizi (1999). "Thermodynamic Depth of Causal States: Objective Complexity via Minimal Representations." *Physical Review E*, **59**: 275–283. URL `arxiv.org/abs/cond-mat/9808147`.

Crutchfield, James P. and Karl Young (1989). "Inferring Statistical Complexity." *Physical Review Letters*, **63**: 105–108.

de Gennes, Pierre-Gilles and Jacques Prost (1993). *The Physics of Liquid Crystals*, vol. 83 of *International Series of Monographys on Physics*. Oxford: Clarendon Press, 2nd edn.

Dorfman, Jay Robert (1998). *An Introduction to Chaos in Nonequilibrium Statistical Mechanics*, vol. 14 of *Cambridge Lecture Notes in Physics*. Cambridge, England: Cambridge University Press.

Dyson, Freeman J. (1979). "Time without End: Physics and Biology in an Open Universe." *Reviews of Modern Physics*, **51**: 447–460.

Ellis, Richard S. (1985). *Entropy, Large Deviations, and Statistical Mechanics*, vol. 271 of *Grundlehren der mathematischen Wissenschaften*. Berlin: Springer-Verlag.

— (1999). "The Theory of Large Deviations: from Boltzmann's 1877 Calculation to Equilibrium Macrostates in 2D Turbulence." *Physica D*, **133**: 106–136.

Feldman, David P. and James P. Crutchfield (1998). *Discovering Noncritical Organization: Statistical Mechanical, Information Theoretic, and Computational Views of Patterns in Simple One-Dimensional Spin Systems*. Working Paper 98-04-026, Santa Fe Institute. URL `www.santafe.edu/projects/CompMech/papers/DNCO.html`.

Fischer, K. H. and J. A. Hertz (1988). *Spin Glasses*. Cambridge Studies in Magnetism. Cambridge: Cambridge University Press.

Forster, Dieter (1975). *Hydrodynamic Fluctuations, Broken Symmetry, and Correlation Functions*, vol. 47 of *Frontiers in Physics*. Reading, Massachusetts: Benjamin Cummings.

Fox, Ronald F. (1988). *Energy and the Evolution of Life*. New York: W. H. Freeman.

Gaspard, Pierre (1998). *Chaos, Scattering and Statistical Mechanics*. Cambridge Nonlinear Science. Cambridge, England: Cambridge University Press.

Grassberger, Peter (1986). "Toward a Quantitative Theory of Self-Generated Complexity." *International Journal of Theoretical Physics*, **25**: 907–938.

Griffeath, David and Cristopher Moore (1996). "Life Without Death is P-Complete." *Complex Systems*, **10**: 437–447.

Griffeath, David and Cristopher Moore (eds.) (forthcoming). *New Constructions in Cellular Automata*, Oxford. Oxford University Press. Proceedings of the Constructive Cellular Automata Theory conference, Santa Fe Institute, Santa Fe, New Mexico, November 1998.

Gutowitz, Howard (ed.) (1991). *Cellular Automata: Theory and Experiment*, Cambridge, Massachusetts. MIT Press. Also published as *Physica D* **45** (1990), nos. 1–3.

Guttmann, Yair M. (1999). *The Concept of Probability in Statistical Physics*. Cambridge Studies in Probability, Induction, and Decision Theory. Cambridge, England: Cambridge University Press.

Hanson, James E. (1993). *Computational Mechanics of Cellular Automata*. Ph.D. thesis, University of California, Berkeley. URL `www.santafe.edu/projects/CompMech/`.

Hanson, James E. and James P. Crutchfield (1997). "Computational Mechanics of Cellular Automata: An Example." *Physica D*, **103**: 169–189.

Hordijk, Wim, Cosma Rohilla Shalizi and James P. Crutchfield (2001). "An Upper Bound on the Products of Particle Interactions in Cellular Automata." *Physica D*, **154**: 240–258. URL `arxiv.org/abs/nlin.CG/0008038`.

Jaynes, E. T. (1983). *Essays on Probability, Statistics, and Statistical Physics*. London: Reidel.

Kantz, Holger and Thomas Schreiber (1997). *Nonlinear Time Series Analysis*, vol. 7 of *Cambridge Nonlinear Science Series*. Cambridge, England: Cambridge University Press.

Keizer, Joel (1987). *Statistical Thermodynamics of Nonequilibrium Processes*. New York: Springer-Verlag.

Khinchin, Aleksandr Iakovlevich (1949). *Mathematical Foundations of Statistical Mechanics*. New York: Dover Publications. Translated from the Russian by G. Gamow.

Lacorata, Guglielmo, Ruben A. Pasmanter and Angelo Vulpiani (2001). "Markov-chain approach to a process with long-time memory." *Journal of Physical Oceanography*, **submitted**. URL `arxiv.org/abs/nlin.CD/0110010`.

Landau, L. D. and E. M. Lifshitz (1980). *Statistical Physics*. Oxford: Pergamon Press.

Mackey, Michael C. (1992). *Time's Arrow: The Origins of Thermodynamic Behavior*. Berlin: Springer-Verlag.

Moore, Cristopher (1997). "Majority-Vote Cellular Automata, Ising Dynamics, and P-Completeness." *Journal of Statistical Physics*, **88**: 795–805.

Moore, Cristopher, Mats G. Nordahl, Nelson Minar and Cosma Rohilla Shalizi (1999). "Vortex Dynamics and Entropic Forces in Antiferromagnets and Antiferromagnetic Potts Models." *Physical Review E*, **60**: 5433–5351. URL `arxiv.org/abs/cond-mat/9902200`.

Packard, Norman H., James P. Crutchfield, J. Doyne Farmer and Robert S. Shaw (1980). "Geometry from a Time Series." *Physical Review Letters*, **45**: 712–716.

Palmer, A. Jay (2001). "An Efficiency Measure for Agent-Based Modeling." In *Swarmfest 2001 Proceedings*. Santa Fe, New Mexico. URL `www.swarm.org/community-swarmfest-program.html#Palmer`.

Palmer, A. Jay, C. W. Fairall and W. A. Brewer (2000). "Complexity in the Atmosphere." *IEEE Transactions on Geoscience and Remote Sensing*, **38**: 2056–2063.

Pearl, Judea (2000). *Causality: Models, Reasoning, and Inference*. Cambridge, England: Cambridge University Press.

Poundstone, William (1984). *The Recursive Universe: Cosmic Complexity and the Limits of Scientific Knowledge*. New York: William Morrow.

Reichl, L. E. (1980). *A Modern Course in Statistical Physics*. Austin: University of Texas Press, 1st edn.

Rothman, Daniel H. and Stéphane Zaleski (1997). *Lattice-Gas Cellular Automata: Simple Models of Complex Hydrodynamics*, vol. 5 of *Collection Aléa Saclay*. Cambridge, England: Cambridge University Press.

Ruelle, David (1989). *Statistical Mechanics: Rigorous Results*. Reading, Massachusetts: Addison Wesley.

— (1999). "Smooth Dynamics and New Theoretical Ideas in Nonequilibrium Statistical Mechanics." *Journal of Statistical Physics*, **95**: 393–468. URL `arxiv.org/abs/chao-dyn/9812032`.

Saito, K., S. Takesue and S. Miyashita (1999). "Transport Phenomena at a Critical Point: Thermal Conduction in the Creutz Cellular Automaton." *Physical Review E*, **59**: 2783–2794. URL `arxiv.org/abs/cond-mat/9811168`.

Salmon, Wesley C. (1971). *Statistical Explanation and Statistical Relevance*. Pittsburgh: University of Pittsburgh Press. With contributions by Richard C. Jeffrey and James G. Greeno.

— (1984). *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.

Shalizi, Cosma Rohilla (2001). *Causal Architecture, Complexity and Self-Organization in Time Series and Cellular Automata*. Ph.D. thesis, University of Wisconsin-Madison. URL `bactra.org/thesis/`.

Shalizi, Cosma Rohilla and James P. Crutchfield (2001). "Computational Mechanics: Pattern and Prediction, Structure and

Simplicity." *Journal of Statistical Physics*, **104**: 817–879. URL arxiv.org/abs/cond-mat/9907176.

Shalizi, Cosma Rohilla, Kristina Lisa Shalizi and James P. Crutchfield (2002). "An Algorithm for Pattern Discovery in Time Series." *Journal of Machine Learning Research*, **submitted**. URL arxiv.org/abs/cs.LG/0210025.

Shaw, Robert (1984). *The Dripping Faucet as a Model Chaotic System*. Santa Cruz, California: Aerial Press.

Sklar, Lawrence (1993). *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*. Cambridge, England: Cambridge University Press.

Spirtes, Peter, Clark Glymour and Richard Scheines (2001). *Causation, Prediction, and Search*. Adaptive Computation and Machine Learning. Cambridge, Massachusetts: MIT Press, 2nd edn.

Takens, Floris (1981). "Detecting Strange Attractors in Fluid Turbulence." In *Symposium on Dynamical Systems and Turbulence* (D. A. Rand and L. S. Young, eds.), vol. 898 of *Lecture Notes in Mathematics*, p. 366. Berlin: Springer-Verlag.

Tolman, Richard C. (1938). *The Principles of Statistical Mechanics*. The International Series of Monographs on Physics. Oxford: Oxford University Press.

von Neumann, John (1966). *Theory of Self-Reproducing Automata*. Urbana: University of Illinois Press. Edited and completed by Arthur W. Burks.

Zallen, Richard (1983). *The Physics of Amorphous Solids*. New York: Wiley.