Journal of Statistical Physics, Vol. 104, Nos. 3/4, 2001

Computational Mechanics: Pattern and Prediction, Structure and Simplicity

Cosma Rohilla Shalizi[1,2] **and James P. Crutchfield**[1]

Received June 20, 2000; revised February 26, 2001

Computational mechanics, an approach to structural complexity, defines a process's causal states and gives a procedure for finding them. We show that the causal-state representation—an ϵ-machine—is the minimal one consistent with accurate prediction. We establish several results on ϵ-machine optimality and uniqueness and on how ϵ-machines compare to alternative representations. Further results relate measures of randomness and structural complexity obtained from ϵ-machines to those from ergodic and information theories.

KEY WORDS: Complexity; computation; entropy; information; pattern; statistical mechanics; causal state; ϵ-machine.

I. INTRODUCTION

Organized matter is ubiquitous in the natural world, but the branch of physics which ought to handle it—statistical mechanics—lacks a coherent, principled way of describing, quantifying, and detecting the many different kinds of structure nature exhibits. Statistical mechanics has good measures of disorder in thermodynamic entropy and in related quantities, such as the free energies. When augmented with theories of critical phenomena[1] and pattern formation,[2] it also has an extremely successful approach to analyzing patterns formed through symmetry breaking, both in equilibrium[3] and, more recently, outside it.[4] Unfortunately, these successes involve many *ad hoc* procedures—such as guessing relevant order parameters, identifying small parameters for perturbation expansion, and choosing appropriate function bases for spatial decomposition. It is far from

[1] Santa Fe Institute, 1399 Hyde Park Road, Santa Fe, New Mexico 87501; e-mail: shalizi@santafe.edu and chaos@santafe.edu

[2] Physics Department, University of Wisconsin-Madison, Madison, Wisconsin 53706.

0022-4715/01/0800-0817$19.50/0

clear that the present methods can be extended to handle all the many kinds of organization encountered in nature, especially those produced by biological processes.

Computational mechanics[5] is an approach that lets us directly address the issues of pattern, structure, and organization. While keeping concepts and mathematical tools already familiar from statistical mechanics, it is distinct from the latter and complementary to it. In essence, from either empirical data or from a probabilistic description of behavior, it shows how to infer a model of the hidden process that generated the observed behavior. This representation—the ϵ-machine—captures the patterns and regularities in the observations in a way that reflects the causal structure of the process. With this model in hand, one can extrapolate beyond the original observations to predict future behavior. Moreover, in a well defined sense that is the subject of the following, the ϵ-machine is the unique maximally efficient model of the observed data-generating process.

ϵ-Machines themselves reveal, in a very direct way, how the process stores information, and how that stored information is transformed by new inputs and by the passage of time. This, and not using computers for simulations and numerical calculations, is what makes computational mechanics "computational", in the sense of "computation theoretic".

The basic ideas of computational mechanics were introduced a decade ago.[6] Since then they have been used to analyze dynamical systems,[7,8] cellular automata,[9] hidden Markov models,[10] evolved spatial computation,[11] stochastic resonance,[12] globally coupled maps,[13] the dripping faucet experiment,[14] and atmospheric turbulence.[15] Despite this record of successful application, there has been some uncertainty about the mathematical foundations of the subject. In particular, while it seemed evident from construction that an ϵ-machine captured the patterns inherent in a process and did so in a minimal way, no explicit proof of this was published. Moreover, there was no proof that, if the ϵ-machine was optimal in this way, it was the *unique* optimal representation of a process. These gaps have now been filled. Subject to some (reasonable) restrictions on the statistical character of a process, we prove that the ϵ-machine is indeed the unique optimal causal model. The rigorous proof of these results is the main burden of this paper. We gave preliminary versions of the optimality results—but not the uniqueness theorem, which is new here—in ref. 16.

The outline of the exposition is as follows. We begin by showing how computational mechanics relates to other approaches to pattern, randomness, and causality. The upshot of this is to focus our attention on *patterns within a statistical ensemble* and their possible representations. Using ideas from information theory, we state a quantitative version of Occam's Razor for such representations. At that point we define *causal states*,[6] equiv-

alence classes of behaviors, and the structure of transitions between causal states—the ϵ-machine. We then show that the causal states are ideal from the point of view of Occam's Razor, being the simplest way of attaining the maximum possible predictive power. Moreover, we show that the causal states are uniquely optimal. This combination allows us to prove a number of other, related optimality results about ϵ-machines. We examine the assumptions made in deriving these optimality results, and we note that several of them can be lifted without unduly upsetting the theorems. We also establish bounds on a process's *intrinsic computation* as revealed by ϵ-machines and by quantities in information and ergodic theories. Finally, we close by reviewing what has been shown and what seem like promising directions for further work on the mathematical foundations of computational mechanics.

A series of appendices provide supplemental material on information theory, equivalence relations and classes, ϵ-machines for time-reversed processes, technical issues of conditional measures, semi-group theory, and connections and distinctions between computational mechanics and other fields.

To set the stage for the mathematics to follow and to motivate the assumptions used there, we begin now by reviewing prior work on pattern, randomness, and causality. We urge the reader interested only in the mathematical development to skip directly to Section II F—a synopsis of the central goals and assumptions of computational mechanics—and continue from there.

II. PATTERNS

To introduce our approach to—and even to argue that *some* approach is necessary for—discovering and describing patterns in nature we begin by quoting Jorge Luis Borges:

> These ambiguities, redundancies, and deficiencies recall those attributed by Dr. Franz Kuhn to a certain Chinese encyclopedia entitled *Celestial Emporium of Benevolent Knowledge*. On those remote pages it is written that animals are divided into (a) those that belong to the Emperor, (b) embalmed ones, (c) those that are trained, (d) suckling pigs, (e) mermaids, (f) fabulous ones, (g) stray dogs, (h) those that are included in this classification, (i) those that tremble as if they were mad, (j) innumerable ones, (k) those drawn with a very fine camel's hair brush, (l) others, (m) those that have just broken a flower vase, (n) those that resemble flies from a distance.
>
> —J. L. Borges, "The Analytical Language of John Wilkins", in ref. 17, p. 103; see also discussion in ref. 18.

The passage illustrates the profound gulf between patterns, and classifications derived from patterns, that are appropriate to the world and help us to understand it and those patterns which, while perhaps just as legitimate as logical entities, are not at all informative. What makes the *Celestial Emporium's* scheme inherently unsatisfactory, and not just strange, is that it tells us nothing about animals. We want to find patterns in a process that "divide it at the joints, as nature directs, not breaking any limbs in half as a bad carver might" (ref. 19, 265D).

Computational mechanics is not directly concerned with pattern formation *per se*;[4] though we suspect it will ultimately be useful in that domain. Nor is it concerned with pattern recognition as a practical matter as found in, say, neuropsychology,[20] psychophysics and perception,[21] cognitive ethology,[22] computer programming,[23] or signal and image processing.[24,25] Instead, it is concerned with the questions of *what patterns are* and *how patterns should be represented*. One way to highlight the difference is to call this pattern *discovery*, rather than pattern *recognition*.

The bulk of the intellectual discourse on what patterns are has been philosophical. One distinct subset has been conducted under the broad rubric of mathematical logic. Within this there are approaches, on the one hand, that draw on (highly) abstract algebra and the theory of relations; on the other, that approach patterns via the theory of algorithms and effective procedures.

The general idea, in both approaches, is that some object $\mathcal{O}$ has a pattern $\mathcal{P}$—$\mathcal{O}$ has a pattern "represented", "described", "captured", and so on by $\mathcal{P}$—if and only if we can use $\mathcal{P}$ to predict or compress $\mathcal{O}$. Note that the ability to predict implies the ability to compress, but not vice versa; here we stick to prediction. The algebraic and algorithmic strands differ mainly on how $\mathcal{P}$ itself should be represented; that is, they differ in how it is expressed in the vocabulary of some formal scheme.

We should emphasize here that "pattern" in this sense implies a kind of regularity, structure, symmetry, organization, and so on. In contrast, ordinary usage sometimes accepts, for example, speaking about the "pattern" of pixels in a particular slice of between-channels video "snow"; but we prefer to speak of that as the *configuration* of pixels.

A. Algebraic Patterns

Although the problem of pattern discovery appears early, in Plato's *Meno*[26] for example, perhaps the first attempt to make the notion of "pattern" mathematically rigorous was that of Whitehead and Russell in *Principia Mathematica*. They viewed patterns as properties, not of sets, but of relations within or between sets, and accordingly they work out an

elaborate *relation-arithmetic* (ref. 27, vol. II, part IV); cf. ref. 28, ch. 5–6. This starts by defining the *relation-number* of a relation between two sets as the class of all the relations that are equivalent to it under one-to-one, onto mappings of the two sets. In this framework relations share a common pattern or structure if they have the same relation-number. For instance, all square lattices have similar structure since their elements share the same neighborhood relation; as do all hexagonal lattices. Hexagonal and square lattices, however, exhibit different patterns since they have non-isomorphic neighborhood relations—i.e., since they have different relation-numbers. (See also *recoding equivalence* defined in ref. 29.) Less work has been done on this than they—especially Russell[30]—had hoped.

A more recent attempt at developing an algebraic approach to patterns builds on semi-group theory and its Krohn–Rhodes decomposition theorem. Ref. 31 discusses a range of applications of this approach to patterns. Along these lines, Rhodes and Nehaniv have tried to apply semigroup complexity theory to biological evolution.[32] They suggest that the complexity of a biological structure can be measured by the number of subgroups in the decomposition of an automaton that describes the structure.

Yet another algebraic approach has been developed by Grenander and co-workers, primarily for pattern recognition.[33] Essentially, this is a matter of trying to invent a minimal set of *generators* and *bonds* for the pattern in question. Generators can adjoin each other, in a suitable *n*-dimensional space, only if their bonds are compatible. Each pair of compatible bonds specifies at once a binary algebraic operation and an observable element of the configuration built out of the generators. (Our construction in Appendix D, linking an algebraic operation with concatenations of strings, is analogous in a rough way, as are the "observable operator models" of ref. 34.) Probabilities can be attached to these bonds, leading in a natural way to a (Gibbsian) probability distribution over entire configurations. Grenander and his colleagues have used these methods to characterize, *inter alia*, several biological phenomena.[35, 36]

B. Turing Mechanics: Patterns and Effective Procedures

The other path to patterns follows the traditional exploration of the logical foundations of mathematics, as articulated by Frege and Hilbert and pioneered by Church, Gödel, Post, Russell, Turing, and Whitehead. A more recent and relatively more popular approach goes back to Kolmogorov and Chaitin, who were interested in the *exact* reproduction of an individual object;[37–40] in particular, their focus was discrete symbol systems, rather than (say) real numbers or other mathematical objects.

The candidates for expressing the pattern $\mathcal{P}$ were universal Turing machine (UTM) programs—specifically, the shortest UTM program that can exactly produce the object $\mathcal{O}$. This program's length is called $\mathcal{O}$'s *Kolmogorov–Chaitin complexity*. Note that any scheme—automaton, grammar, or what-not—that is Turing equivalent and for which a notion of "length" is well defined will do as a representational scheme. Since we can convert from one such device to another—say, from a Post tag system[41] to a Turing machine—with only a finite description of the first system, such constants are easily assimilated when measuring complexity in this approach.

In particular, consider the first n symbols $\mathcal{O}_n$ of $\mathcal{O}$ and the shortest program $\mathcal{P}_n$ that produces them. We ask, What happens to the limit

$$\lim_{n \to \infty} \frac{|\mathcal{P}_n|}{n}, \tag{1}$$

where $|\mathcal{P}|$ is the length in bits of program $\mathcal{P}$? On the one hand, if there is a fixed-length program $\mathcal{P}$ that generates arbitrarily many digits of $\mathcal{O}$, then this limit vanishes. Most of our interesting numbers, rational or irrational—such as 7, π, e, $\sqrt{2}$—are of this sort. These numbers are eminently compressible: the program $\mathcal{P}$ is the compressed description, and so it captures the pattern obeyed by the sequence describing $\mathcal{O}$. If the limit goes to 1, on the other hand, we have a completely incompressible description and conclude, following Kolmogorov, Chaitin, and others, that $\mathcal{O}$ is random.[37–40, 42, 43] This conclusion is the desired one: the Kolmogorov–Chaitin framework establishes, formally at least, the randomness of an individual object without appeals to probabilistic descriptions or to ensembles of reproducible events. And it does so by referring to a deterministic, algorithmic representation—the UTM.

There are many well-known difficulties with applying Kolmogorov complexity to natural processes. First, as a quantity, it is uncomputable in general, owing to the halting problem.[40] Second, it is maximal for random sequences; this can be construed either as desirable, as just noted, or as a failure to capture structure, depending on one's aims. Third, it only applies to a single sequence; again this is either good or bad. Fourth, it makes no allowance for noise or error, demanding exact reproduction. Finally, $\lim_{n \to \infty} |\mathcal{P}_n|/n$ can vanish, although the computational resources needed to run the program, such as time and storage, grow without bound.

None of these impediments have kept researchers from attempting to use Kolmogorov–Chaitin complexity for practical tasks—such as measuring the complexity of natural objects (e.g., ref. 44), as a basis for theories of inductive inference,[45, 46] and generally as a means of capturing patterns.[47]

As Rissanen (ref. 48, p. 49) says, this is akin to "learn[ing] the properties [of a data set] by writing programs in the hope of finding short ones!"

Various of the difficulties just listed have been addressed by subsequent work. Bennett's *logical depth* accounts for time resources.[49] (In fact, it is the time for the minimal-length program $\mathcal{P}$ to produce $\mathcal{O}$.) Koppel's *sophistication* attempts to separate out the "regularity" portion of the program from the random or instance-specific input data.[50, 51] Ultimately, these extensions and generalizations remain in the UTM, exact-reproduction setting and so inherit inherent uncomputability.

C. Patterns with Error

Motivated by these theoretical difficulties and practical concerns, an obvious next step is to allow our pattern $\mathcal{P}$ some degree of approximation or error, in exchange for shorter descriptions. As a result, we lose perfect reproduction of the original configuration from the pattern. Given the ubiquity of noise in nature, this is a small price to pay. We might also say that sometimes we are willing to accept small deviations from a regularity, without really caring what the precise deviation is. As pointed out in ref. 18's conclusion, this is certainly a prime motivation in thermodynamic descriptions, in which we explicitly throw away, and have no interest in, vast amounts of microscopic detail in order to find a workable description of macroscopic observations.

Some interesting philosophical work on patterns-with-error has been done by Dennett, with reference not just to questions about the nature of patterns and their emergence but also to psychology.[52] The intuition is that truly random processes can be modeled very simply—"to model coin-tossing, toss a coin." Any prediction scheme that is more accurate than assuming complete independence *ipso facto* captures a pattern in the data. There is thus a spectrum of potential pattern-capturers ranging from the assumption of pure noise to the exact reproduction of the data, if that is possible. Dennett notes that there is generally a trade-off between the simplicity of a predictor and its accuracy, and he plausibly describes emergent phenomena[53, 54] as patterns that allow for a large reduction in complexity for only a small reduction in accuracy. Of course, Dennett was not the first to consider predictive schemes that tolerate error and noise; we discuss some of the earlier work in App. 8. However, to our knowledge, he was the first to have made such predictors a central part of an explicit account of *what patterns are*. It must be noted that this account lacks the mathematical detail of the other approaches we have considered so far, and that it relies on the inexact prediction of a single configuration. In fact, it relies on exact predictors that are "fuzzed up" by noise. The introduction

of noise, however, brings in probabilities, and their natural setting is in ensembles. It is in that setting that the ideas we share with Dennett can receive a proper quantitative treatment.

D. Randomness: The Anti-Pattern?

We should at this point say a bit about the relations between *randomness*, *complexity*, and *structure*, at least as we use those words. Ignoring some foundational issues, randomness is actually rather well understood and well handled by classical tools introduced by Boltzmann;[55] Fisher, Neyman, and Pearson;[56] Kolmogorov;[37] and Shannon,[57] among others. One tradition in the study of complexity in fact identifies complexity with randomness and, as we have just seen, this is useful for some purposes. As these purposes are *not* those of analyzing patterns in processes and in real-world data, however, they are not ours. Randomness simply does not correspond to a notion of pattern or structure at all and, by implication, neither Kolmogorov–Chaitin complexity nor any of its spawn measure pattern.

Nonetheless, some approaches to complexity conflate "structure" with the opposite of randomness, as conventionally understood and measured in physics by thermodynamic entropy or a related quantity, such as Shannon entropy. In effect, structure is defined as "one minus disorder". In contrast, we see pattern—structure, organization, regularity, and so on—as describing a coordinate "orthogonal" to a process's degree of randomness. That is, complexity (in our sense) and randomness each capture a useful property necessary to describe how a process manipulates information. This complementarity is even codified by the complexity-entropy diagrams introduced in ref. 6. When we use the word "complexity" we mean "degrees" of pattern, not degrees of randomness.

E. Causation

We want our representations of patterns in dynamical processes to be causal—to say how one state of affairs leads to or produces another. Although a key property, causality enters our development only in an extremely weak sense, the weakest one can use mathematically, which is Hume's:[58] one class of event causes another if the latter always follows the former; the effect invariably succeeds the cause. As good indeterminists, in the following we replace this invariant-succession notion of causality with a more probabilistic one, substituting a homogeneous distribution of successors for the solitary invariable successor. (A precise statement appears in Section IV A's definition of *causal states*.) This approach results in a purely

phenomenological statement of causality, and so it is amenable to experimentation in ways that stronger notions of causality—e.g., that of ref. 59—are not. Ref. 60 independently reaches a concept of causality essentially the same ours via philosophical arguments.

F. Synopsis of Pattern

Our survey leads us to look for an approach to patterns which is at once

1. *Algebraic*, giving us an explicit breakdown or decomposition of the pattern into its parts;
2. *Computational*, showing how the process stores and uses information;
3. *Calculable*, analytically or by systematic approximation;
4. *Causal*, telling us how instances of the pattern are actually produced; and
5. *Naturally stochastic*, not merely tolerant of noise but explicitly formulated in terms of ensembles.

Computational mechanics satisfies all these desiderata.

III. PATTERNS IN ENSEMBLES: PADDLING AROUND OCCAM'S POOL

Here a pattern $\mathcal{P}$ is something knowledge of which lets us predict, at better than chance rates, if possible, the future of sequences drawn from an ensemble $\mathcal{O}$: $\mathcal{P}$ has to be statistically accurate and confer some leverage or advantage as well. Let's fix some notation and state the assumptions that will later let us prove the basic results.

A. Hidden Processes

We restrict ourselves to discrete-valued, discrete-time stationary stochastic processes. (See Section VIIB for discussion of these assumptions.) Intuitively, such processes are sequences of random variables S_i, the values of which are drawn from a countable set $\mathcal{A}$. We let i range over all the integers, and so get a bi-infinite sequence

$$\overleftrightarrow{S} = \ldots S_{-1} S_0 S_1 \ldots \tag{2}$$

In fact, we define a process in terms of the distribution of such sequences; cf. refs. 61 and 62.

Definition 1 (A Process). Let $\mathcal{A}$ be a countable set. Let $\Omega = \mathcal{A}^{\mathbb{Z}}$ be the set of bi-infinite sequences composed from $\mathcal{A}$, $T_i: \Omega \mapsto \mathcal{A}$ be the measurable function that returns the i^{th} element s_i of a bi-infinite sequence $\omega \in \Omega$, and $\mathcal{F}$ the σ-algebra of cylinder sets of Ω. Adding a probability measure P gives us a probability space $(\Omega, \mathcal{F}, \mathrm{P})$, with an associated random variable $\overleftrightarrow{S}$. A *process* is a sequence of random variables $S_i = T_i(\overleftrightarrow{S})$, $i \in \mathbb{Z}$.

Here, and throughout, we follow the convention of using capital letters to denote random variables and lower-case letters their particular values.

It follows from Definition 1 that there are well defined probability distributions for sequences of every finite length. Let $\vec{S}_t^L$ be the sequence of $S_t, S_{t+1}, \ldots, S_{t+L-1}$ of L random variables beginning at S_t. $\vec{S}_t^0 \equiv \lambda$, the null sequence. Likewise, $\overleftarrow{S}_t^L$ denotes the sequence of L random variables going up to S_t, but not including it; $\overleftarrow{S}_t^L = \vec{S}_{t-L}^L$. Both $\vec{S}_t^L$ and $\overleftarrow{S}_t^L$ take values from $s^L \in \mathcal{A}^L$. Similarly, $\vec{S}_t$ and $\overleftarrow{S}_t$ are the semi-infinite sequences starting from and stopping at t and taking values $\vec{s}$ and $\overleftarrow{s}$, respectively.

Intuitively, we can imagine starting with distributions for finite-length sequences and extending them gradually in both directions, until the infinite sequence is reached as a limit. While this can be a useful picture to have in mind, defining a process in this way raises some subtle measure-theoretic issues, such as how distributions over finite-length sequences limit on the infinite-length distribution. To evade these questions, we *start* with the latter, and obtain the former by "marginalization". (Readers will find a particularly clear exposition of this approach in ch. 1 of ref. 62.)

Definition 2 (Stationarity). A process S_i is *stationary* if and only if

$$\mathrm{P}(\vec{S}_t^L = s^L) = \mathrm{P}(\vec{S}_0^L = s^L), \tag{3}$$

for all $t \in \mathbb{Z}$, $L \in \mathbb{Z}^+$, and all $s^L \in \mathcal{A}^L$.

In other words, a stationary process is one that is time-translation invariant. Consequently, $\mathrm{P}(\vec{S}_t = \vec{s}) = \mathrm{P}(\vec{S}_0 = \vec{s})$ and $\mathrm{P}(\overleftarrow{S}_t = \overleftarrow{s}) = \mathrm{P}(\overleftarrow{S}_0 = \overleftarrow{s})$, and so we drop the subscripts from now on.

We will call $\overleftarrow{S}$ and $\overleftarrow{S}^L$ *pasts* or *histories* and $\vec{S}$ and $\vec{S}^L$, *futures*. We will need to refer to the class of all measurable sets of histories; this will be

$\boldsymbol{\mu}(\overleftarrow{S})$.[3] Similarly, the class of all measurable sets of futures is $\boldsymbol{\mu}(\overrightarrow{S})$. It is readily checked[10] that $\boldsymbol{\mu}(\overleftarrow{S}) = \bigcup_{L=1}^{\infty} \boldsymbol{\mu}(\overleftarrow{S}^L)$, and likewise for $\boldsymbol{\mu}(\overrightarrow{S})$.

B. The Pool

Our goal is to predict all or part of $\overrightarrow{S}$ using some function of some part of $\overleftarrow{S}$. We begin by taking the set $\overleftarrow{\mathbf{S}}$ of all pasts and partitioning it into mutually exclusive and jointly comprehensive subsets. That is, we make a class $\mathcal{R}$ of subsets of pasts.[4] (See Fig. 1 for a schematic example.) Each $\rho \in \mathcal{R}$ will be called a *state* or an *effective state*. When the current history $\overleftarrow{S}$ is included in the set ρ, we will speak of the process being in state ρ. Thus, we define a function η from histories to effective states:

$$\eta: \overleftarrow{\mathbf{S}} \mapsto \mathcal{R}. \tag{4}$$

A specific individual history $\overleftarrow{s} \in \overleftarrow{\mathbf{S}}$ maps to a specific state $\rho \in \mathcal{R}$; the random variable $\overleftarrow{S}$ for the past maps to the random variable $\mathcal{R}$ for the effective states. It makes little difference whether we think of η as being a function from a history to a subset of histories or a function from a history to the *label* of that subset. Each interpretation is convenient at different times, and we will use both.

Note that we could use *any* function defined on $\overleftarrow{\mathbf{S}}$ to partition that set, by assigning to the same ρ all the histories $\overleftarrow{s}$ on which the function takes the same value. Similarly, any equivalence relation on $\overleftarrow{\mathbf{S}}$ partitions it. (See Appendix B for more on equivalence relations.) Due to the way we defined a process's distribution, each effective state has a well defined distribution of futures, though not necessarily a unique one.[5] Specifying the effective state thus amounts to making a prediction about the process's future. All the histories belonging to a given effective state are treated as *equivalent for purposes of predicting the future*. (In this way, the framework formally incorporates traditional methods of time-series analysis; see Appendix H1.)

We call the collection of all partitions $\mathcal{R}$ of the set of histories $\overleftarrow{\mathbf{S}}$ *Occam's pool*.

[3] Conventionally, this ought to be $\sigma(\overleftarrow{S})$, but, as the reader will see, that notation would be confusing later on.

[4] At several points our constructions require referring to sets of sets. To help mark the distinction, we call the set of sets of histories a *class*.

[5] This is not true if η is not at least nearly measurable (see Appendix E2b). To paraphrase ref. 140, readers should assume that all our effective-state functions are sufficiently tame, measure-theoretically, that whatever induced distributions we invoke will exist.

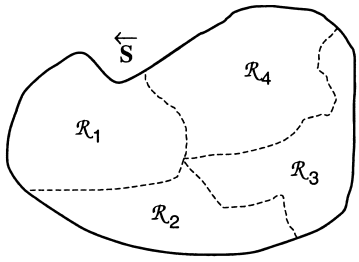

Fig. 1. A schematic picture of a partition of the set $\overleftarrow{\mathbf{S}}$ of all histories into some class of effective states: $\mathcal{R} = \{\mathcal{R}_i : i = 1, 2, 3, 4\}$. Note that the $\mathcal{R}_i$ need not form compact sets; we simply draw them that way for clarity. One should have in mind Cantor sets or other more pathological structures.

C. A Little Information Theory

Since the bulk of the following development will be consumed with notions and results from information theory,[57] we now review several highlights briefly, for the benefit of readers unfamiliar with the theory and to fix notation. Appendix 1 lists a number of useful information-theoretic formulæ, which get called upon in our proofs. Throughout, our notation and style of proof follow those in ref. 63.

1. Entropy Defined

Given a random variable X taking values in a countable set $\mathcal{A}$, the entropy of X is

$$H[X] \equiv -\sum_{x \in \mathcal{A}} \mathrm{P}(X = x) \log_2 \mathrm{P}(X = x), \tag{5}$$

taking $0 \log 0 = 0$. Notice that $H[X]$ is the expectation value of $-\log_2 \mathrm{P}(X = x)$ and is measured in *bits* of information. Caveats of the form "when the sum converges to a finite value" are implicit in all statements about the entropies of infinite countable sets $\mathcal{A}$.

Shannon interpreted $H[X]$ as the *uncertainty in* X. (Those leery of any subjective component in notions like "uncertainty" may read "effective variability" in its place.) He showed, for example, that $H[X]$ is the mean number of yes-or-no questions needed to pick out the value of X on repeated trials, if the questions are chosen to minimize this average.[57]

2. Joint and Conditional Entropies

We define the joint entropy $H[X, Y]$ of two variables X (taking values in $\mathcal{A}$) and Y (taking values in $\mathcal{B}$) in the obvious way,

$$H[X, Y] \equiv -\sum_{(x, y) \in \mathcal{A} \times \mathcal{B}} \mathrm{P}(X = x, Y = y) \log_2 \mathrm{P}(X = x, Y = y).$$

We define the conditional entropy $H[X|Y]$ of one random variable X with respect to another Y from their joint entropy:

$$H[X|Y] \equiv H[X,Y] - H[Y]. \tag{7}$$

This also follows naturally from the definition of conditional probability, since $\mathrm{P}(X=x|Y=y) \equiv \mathrm{P}(X=x, Y=y)/\mathrm{P}(Y=y)$. $H[X|Y]$ measures the mean uncertainty remaining in X once we know Y.[6]

3. Mutual Information

The mutual information $I[X;Y]$ between two variables is defined to be

$$I[X;Y] \equiv H[X] - H[X|Y]. \tag{8}$$

This is the average reduction in uncertainty about X produced by fixing Y. It is non-negative, like all entropies here, and symmetric in the two variables.

D. Patterns in Ensembles

It will be convenient to have a way of talking about the uncertainty of the future. Intuitively, this would just be $H[\vec{S}]$, but in general that quantity is infinite and awkward to manipulate. (The special case in which $H[\vec{S}]$ is finite is dealt with in Appendix G.) Normally, we evade this by considering $H[\vec{S}^L]$, the uncertainty of the next L symbols, treated as a function of L. On occasion, we will refer to the entropy per symbol or *entropy rate*:[57, 63]

$$h[\vec{S}] \equiv \lim_{L\to\infty} \frac{1}{L} H[\vec{S}^L], \tag{9}$$

and the *conditional entropy rate*,

$$h[\vec{S}|X] \equiv \lim_{L\to\infty} \frac{1}{L} H[\vec{S}^L|X], \tag{10}$$

where X is some random variable and the limits exist. For stationary stochastic processes, the limits always exist (ref. 63, Theorem 4.2.1, p. 64).

[6] We can still define the conditional entropy when the conditioning variable is not discrete; in particular, we can still define it when we sometimes need to condition on events of probability zero. All the normal inequalities about conditional entropy we invoke in our proofs still hold good. See Appendix E.

These entropy rates are also always bounded above by $H[S]$; which is a special case of Eq. (A3). Moreover, if $h[\vec{S}] = H[S]$, the process consists of independent variables—independent, identically distributed (IID) variables, in fact, since we are only concerned with stationary processes here.

Definition 3 (Capturing a Pattern). $\mathcal{R}$ *captures a pattern* if and only if there exists an L such that

$$H[\vec{S}^L|\mathcal{R}] < LH[S]. \tag{11}$$

This says that $\mathcal{R}$ captures a pattern when it tells us something about how the distinguishable parts of a process affect each other: $\mathcal{R}$ exhibits their dependence. (We also speak of η, the function associated with pasts, as capturing a pattern, since this is implied by $\mathcal{R}$ capturing a pattern.) Supposing that these parts *do not* affect each other, then we have IID random variables, which is as close to the intuitive notion of "patternless" as one is likely to state mathematically. Note that, because of the independence bound on joint entropies (Eq. (A3)), if the inequality is satisfied for some L, it is also satisfied for every $L' > L$. Thus, we can consider the difference $H[S] - H[\vec{S}^L|\mathcal{R}]/L$, for the smallest L for which it is nonzero, as the *strength of the pattern* captured by $\mathcal{R}$. We will now mark an upper bound (Lemma 1) on the strength of patterns; later we will show how to attain this upper bound (Theorem 1).

E. The Lessons of History

We are now in a position to prove a result about patterns in ensembles that will be useful in connection with our later theorems about causal states.

Lemma 1 (Old Country Lemma). For all $\mathcal{R}$ and for all $L \in \mathbb{Z}^+$,

$$H[\vec{S}^L|\mathcal{R}] \geqslant H[\vec{S}^L|\overleftarrow{S}]. \tag{12}$$

Proof. By construction (Eq. (4)), for all L,

$$H[\vec{S}^L|\mathcal{R}] = H[\vec{S}^L|\eta(\overleftarrow{S})]. \tag{13}$$

But

$$H[\vec{S}^L|\eta(\overleftarrow{S})] \geqslant H[\vec{S}^L|\overleftarrow{S}], \tag{14}$$

since the entropy conditioned on a variable is never more than the entropy conditioned on a function of the variable (Eq. (A14)). QED.

Remark 1. That is, conditioning on the whole of the past reduces the uncertainty in the future to as small a value as possible. Carrying around the whole semi-infinite past is rather bulky and uncomfortable and is a somewhat dismaying prospect. Put a bit differently: we want to forget as much of the past as possible and so reduce its burden. It is the contrast between this desire and the result of Eq. (12) that leads us to call this the *Old Country Lemma*.

Remark 2. Lemma 1 establishes the promised upper bound on the strength of patterns: viz., the strength of the pattern is at most $H[S] - H[\vec{S}^L|\overleftarrow{S}]/L_{past}$, where L_{past} is the least value of L such that $H[\vec{S}^L|\overleftarrow{S}] < LH[S]$.

F. Minimality and Prediction

Let's invoke Occam's Razor: "It is vain to do with more what can be done with less".[64] To use the razor, we need to fix what is to be "done" and what "more" and "less" mean. The job we want done is accurate prediction; i.e., reducing the conditional entropies $H[\vec{S}^L|\mathcal{R}]$ as far as possible, the goal being to attain the bound set by Lemma 1. But we want to do this as simply as possible, with as few resources as possible. On the road to meeting these two constraints—minimal uncertainty and minimal resources—we will need a measure of the second. Since there is a probability measure over pasts, there is an induced measure on the η-states.[7] Accordingly, we define the following measure of resources.

Definition 4 (Complexity of State Classes). The *statistical complexity* of a class $\mathcal{R}$ of states is

$$\begin{aligned} C_\mu(\mathcal{R}) &\equiv H[\mathcal{R}] \\ &= -\sum_{\rho \in \mathcal{R}} \mathrm{P}(\mathcal{R} = \rho) \log_2 \mathrm{P}(\mathcal{R} = \rho), \end{aligned} \tag{15}$$

when the sum converges to a finite value.

[7] Again, this assumes η is at least nearly measurable. See App. E2b.

The μ in C_μ reminds us that it is a measure-theoretic property and depends ultimately on the distribution over the process's sequences, which induces a measure over states.

The statistical complexity of a state class is the average uncertainty (in bits) in the process's current state. This, in turn, is the same as the average amount of memory (in bits) that the process *appears* to retain about the past, given the chosen state class $\mathcal{R}$. (We will later, in Definition 12, see how to define the statistical complexity of a process itself.) The goal is to do with as little of this memory as possible. Restated then, we want to minimize statistical complexity, subject to the constraint of maximally accurate prediction.

The idea behind calling the collection of all partitions of $\tilde{\mathbf{S}}$ Occam's pool should now be clear: One wants to find the shallowest point in the pool. This we now do.

IV. COMPUTATIONAL MECHANICS

> Those who are good at archery learnt from the bow and not from Yi the Archer. Those who know how to manage boats learnt from the boats and not from Wo.
>
> —Anonymous in ref. 65.

The ultimate goal of computational mechanics is to discern the patterns intrinsic to a process. That is, as much as possible, the goal is to let the process describe itself, on its own terms, without appealing to *a priori* assumptions about the process's structure. Here we simply explore the consistency and well-definedness of these goals. In practice, we may be constrained to merely approximate these ideals more or less grossly. Naturally, such problems, which always turn up in implementation, are much easier to address if we start from secure foundations.

Our definitions and constructions in this section rely on conditional probabilities. This is unproblematic so long as we condition on events of nonzero probability. However, we need to condition on events, such as particular histories, whose probability is generally zero. There are well established ways of handling this difficulty, but their attendant technicalities tend to obscure the main lines of our argument. To keep those lines as clear as possible, in this section we state our definitions as though classical conditional probability was adequate, reserving the measure-theoretic treatment of our main concepts for Appendix E, where we note the limitations and caveats required by this stricter approach. Our proofs are constructed so as to be compatible with the proper use of conditional measures, but intelligible (if merely heuristic) without it.

A. Causal States

Definition 5 (A Process's Causal States) The *causal states* of a process are the members of the range of the function ϵ that maps from histories to sets of histories:

$$\begin{aligned} \epsilon: \overleftarrow{\mathbf{S}} &\mapsto 2^{\overleftarrow{\mathbf{S}}} \\ \epsilon(\overleftarrow{s}) &\equiv \{\overleftarrow{s}' \,|\, \mathrm{P}(\vec{S} \in F \,|\, \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\vec{S} \in F \,|\, \overleftarrow{S} = \overleftarrow{s}'), \\ &\qquad \text{for all } F \in \boldsymbol{\mu}(\vec{S}), \overleftarrow{S}' \in \overleftarrow{s}\}, \end{aligned} \tag{16}$$

where $2^{\overleftarrow{\mathbf{S}}}$ is the power set of $\overleftarrow{\mathbf{S}}$ and $\boldsymbol{\mu}(\vec{S})$ is the collection of all measurable future events. We write the i^{th} causal state as $\mathcal{S}_i$ and the set of all causal states as $\boldsymbol{\mathcal{S}}$; the corresponding random variable is denoted $\mathcal{S}$, and its realization σ.

The cardinality and topology of $\boldsymbol{\mathcal{S}}$ are unspecified. $\boldsymbol{\mathcal{S}}$ can be finite, countably infinite, a continuum, a Cantor set, or something stranger still. Examples of these are given in refs. 5 and 10; see especially the examples for hidden Markov models given there.

Alternately and equivalently, we could define an equivalence relation $\sim_\epsilon$ such that two histories are equivalent if and only if they have the same conditional distribution of futures, and then define causal states as the equivalence classes generated by $\sim_\epsilon$. (In fact, this was the original approach.[6]) Either way, the divisions of this partition of $\overleftarrow{\mathbf{S}}$ are made between regions that leave us in different conditions of ignorance about the future.

This last statement suggests another, still equivalent, description of ϵ:

$$\begin{aligned} \epsilon(\overleftarrow{s}) = \{\overleftarrow{s}' \,|\, \mathrm{P}(\vec{S}^L = \vec{s}^L \,|\, \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\vec{S}^L = \vec{s}^L \,|\, \overleftarrow{S} = \overleftarrow{s}'), \\ \vec{s}^L \in \vec{S}^L, \overleftarrow{s}' \in \overleftarrow{S}, L \in \mathbb{Z}^+\}. \end{aligned} \tag{17}$$

Using this we can make the original definition, Eq. (16), more intuitive by picturing a sequence of partitions of the space $\overleftarrow{\mathbf{S}}$ of all histories in which each new partition, induced using $L+1$, is a refinement of the previous one induced using L. At the coarsest level, the first partition ($L = 1$) groups together those histories that have the same distribution for the very next observable. These classes are then subdivided using the distribution of the next two observables, then the next three, four, and so on. The limit of this sequence of partitions—the point at which every member of each class has

the same distribution of futures, of whatever length, as every other member of that class—is the partition of $\overleftarrow{\mathbf{S}}$ induced by $\sim_\epsilon$. See Appendix B for a detailed discussion and review of the equivalence relation $\sim_\epsilon$.

Although they will not be of direct concern in the following, due to the time-asymptotic limits taken, there are transient causal states in addition to those (recurrent) causal states defined above in Eq. (16). Roughly speaking, the transient causal states describe how a lengthening sequence (a history) of observations allows us to identify the recurrent causal states with increasing precision. See the developments in Appendix B and in refs. 10 and 66 for more detail on transient causal states.

Causal states are a particular kind of effective state, and they have all the properties common to effective states (Section III B). In particular, each causal state $\mathcal{S}_i$ has several structures attached:

1. The index i—the state's "name".

2. The set of histories that have brought the process to $\mathcal{S}_i$, which we denote $\{\overleftarrow{s} \in \mathcal{S}_i\}$.

3. A conditional distribution over futures, denoted $\mathrm{P}(\overrightarrow{S} \mid \mathcal{S}_i)$ and equal to $\mathrm{P}(\overrightarrow{S} \mid \overleftarrow{s})$, $\overleftarrow{s} \in \mathcal{S}_i$. Since we refer to this type of distribution frequently and since it is the "shape of the future", we call it the state's *morph*.

Ideally, each of these should be denoted by a different symbol, and there should be distinct functions linking each of these structures to their causal state. To keep the growth of notation under control, however, we shall be tactically vague about these distinctions. Readers may variously picture ϵ as mapping histories to (i) simple indices, (ii) subsets of histories, (iii) distributions over futures, or (iv) ordered triples of indices, subsets, and morphs; or one may even leave ϵ uninterpreted, as preferred, without interfering with the development that follows.

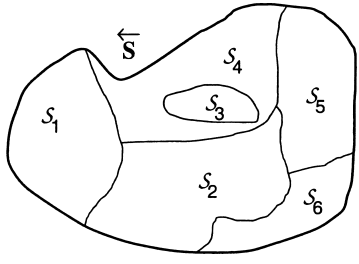

Fig. 2. A schematic representation of the partitioning of the set $\overleftarrow{\mathbf{S}}$ of all histories into causal states $\mathcal{S}_i \in \boldsymbol{\mathcal{S}}$. Within each causal state all the individual histories $\overleftarrow{s}$ have the same morph—the same conditional distribution $\mathrm{P}(\overrightarrow{S} \mid \overleftarrow{s})$ for future observables.

1. Morphs

Each causal state has a unique morph, i.e., no two causal states have the same conditional distribution of futures. This follows directly from Definition 5, and it is not true of effective states in general. Another immediate consequence of that definition is that, for any measurable future event F,

$$\mathrm{P}(\vec{S} \in F \,|\, \mathcal{S} = \epsilon(\overleftarrow{s})) = \mathrm{P}(\vec{S} \in F \,|\, \overleftarrow{S} = \overleftarrow{s}). \tag{18}$$

(Again, this is not generally true of effective states.) This observation lets us prove a useful lemma about the conditional independence of the past $\overleftarrow{S}$ and the future $\vec{S}$.

Lemma 2. The past and the future are independent, conditioning on the causal states.

Proof. By Proposition 4 of Appendix E, $\overleftarrow{S}$ and $\vec{S}$ are independent given $\mathcal{S}$ if and only if, for any measurable set F of futures, $\mathrm{P}(\vec{S} \in F \,|\, \overleftarrow{S} = \overleftarrow{s}, \mathcal{S} = \sigma) = \mathrm{P}(\vec{S} \in F \,|\, \mathcal{S} = \sigma)$. Since $\mathcal{S} = \epsilon(\overleftarrow{S})$, it is automatically true (Eq. (E5)) that $\mathrm{P}(\vec{S} \in F \,|\, \overleftarrow{S} = \overleftarrow{s}, \mathcal{S} = \epsilon(\overleftarrow{s})) = \mathrm{P}(\vec{S} \in F \,|\, \overleftarrow{S} = \overleftarrow{s})$. But then, $\mathrm{P}(\vec{S} \in F \,|\, \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\vec{S} \in F \,|\, \mathcal{S} = \epsilon(\overleftarrow{s}))$, so $\mathrm{P}(\vec{S} \in F \,|\, \overleftarrow{S} = \overleftarrow{s}, \mathcal{S} = \sigma) = \mathrm{P}(\vec{S} \in F \,|\, \mathcal{S} = \sigma)$. QED.

2. Homogeneity

Following ref. 60, we introduce two new definitions and a lemma which are required later on, especially in the proof of Lemma 7 and the theorems depending on that lemma.

Definition 6 (Strict Homogeneity). A set $\mathbf{X}$ is *strictly homogeneous* with respect to a random variable Y when the conditional distribution $\mathrm{P}(Y \,|\, \mathbf{X})$ for Y is the same for all measurable subsets of $\mathbf{X}$.

Definition 7 (Weak Homogeneity). A set $\mathbf{X}$ is *weakly homogeneous* with respect to Y if $\mathbf{X}$ is not strictly homogeneous with respect to Y, but $\mathbf{X} \setminus \mathbf{X}_0$ ($\mathbf{X}$ with $\mathbf{X}_0$ removed) is, where $\mathbf{X}_0$ is a subset of $\mathbf{X}$ of measure 0.

Lemma 3 (Strict Homogeneity of Causal States). A process's causal states are the largest subsets of histories that are all strictly homogeneous with respect to futures of all lengths.

Proof. We must show that, first, the causal states are strictly homogeneous with respect to futures of all lengths and, second, that no larger

strictly homogeneous subsets of histories could be made. The first point, the strict homogeneity of the causal states, is evident from Eq. (17): By construction, all elements of a causal state have the same morph, so any part of a causal state will have the same morph as the whole state. The second point likewise follows from Eq. (17), since the causal state by construction contains *all* the histories with a given morph. Any other set strictly homogeneous with respect to futures must be smaller than a causal state, and any set that includes a causal state as a proper subset cannot be *strictly* homogeneous. QED.

Remark. The statistical explanation literature would say that causal states are the "statistical-relevance basis for causal explanations". The elements of such a basis are, precisely, the largest classes of combinations of independent variables with homogeneous distributions for the dependent variables. See ref. 60 for further discussion along these lines.

B. Causal State-to-State Transitions

The causal state at any given time and the next value of the observed process together determine a new causal state; this is proved shortly in Lemma 5. Thus, there is a natural relation of succession among the causal states; recall the discussion of causality in Section IIE. Moreover, given the current causal state, all the possible next values of the observed sequence ($\vec{S}^1$) have well defined conditional probabilities. In fact, by construction the entire semi-infinite future ($\vec{S}$) does. Thus, there is a well defined probability $T_{ij}^{(s)}$ of the process generating the value $s \in \mathcal{A}$ and going to causal state $\mathcal{S}_j$, if it is in state $\mathcal{S}_i$.

Definition 8 (Causal Transitions). The labeled *transition probability* $T_{ij}^{(s)}$ is the probability of making the transition from state $\mathcal{S}_i$ to state $\mathcal{S}_j$ while emitting the symbol $s \in \mathcal{A}$:

$$T_{ij}^{(s)} \equiv \mathrm{P}(\mathcal{S}' = \mathcal{S}_j, \vec{S}^1 = s \,|\, \mathcal{S} = \mathcal{S}_i), \tag{19}$$

where $\mathcal{S}$ is the current causal state and $\mathcal{S}'$ its successor. We denote the set $\{T_{ij}^{(s)} : s \in \mathcal{A}\}$ by $\mathbf{T}$.

Lemma 4 (Transition Probabilities). $T_{ij}^{(s)}$ is given by

$$T_{ij}^{(s)} = \mathrm{P}(\overleftarrow{S}s \in \mathcal{S}_j \,|\, \overleftarrow{S} \in \mathcal{S}_i), \tag{20}$$

where $\overleftarrow{S}s$ is read as the semi-infinite sequence obtained by concatenating $s \in \mathcal{A}$ onto the end of $\overleftarrow{S}$.

Proof. We show that the events concerned are really the same. That is, we want to show that

$$\{\mathcal{S}' = \mathcal{S}_j, \vec{S}^1 = s, \mathcal{S} = \mathcal{S}_i\} = \{\overleftarrow{S}s \in \mathcal{S}_j, \overleftarrow{S} \in \mathcal{S}\}.$$

Now, that $\mathcal{S} = \mathcal{S}_i$ and $\overleftarrow{S} \in \mathcal{S}_i$ are the same event is clear by construction. So, too, for $\overleftarrow{S}' \in \mathcal{S}_j$ and $\mathcal{S}' = \mathcal{S}_j$. So we can certainly assert that

$$\{\mathcal{S}' = \mathcal{S}_j, \vec{S}^1 = s, \mathcal{S} = \mathcal{S}_i\} = \{\overleftarrow{S}' \in \mathcal{S}_j, \vec{S}^1 = s, \overleftarrow{S} \in \mathcal{S}_i\}.$$

The conjunction of the first and third events implies that, for all $\overleftarrow{s}$, if $\overleftarrow{S} = \overleftarrow{s}$, then $\overleftarrow{S}' = \overleftarrow{s}a$, for some symbol $a \in \mathcal{A}$. But the middle event ensures that $a = s$. Hence,

$$\{\mathcal{S}' = \mathcal{S}_j, \vec{S}^1 = s, \mathcal{S} = \mathcal{S}_i\} = \{\overleftarrow{S}s \in \mathcal{S}_j, \vec{S}^1 = s, \overleftarrow{S} \in \mathcal{S}_i\}.$$

But now the middle event is redundant and can be dropped. Thus,

$$\{\mathcal{S}' = \mathcal{S}_j, \vec{S}^1 = s, \mathcal{S} = \mathcal{S}_i\} = \{\overleftarrow{S}s \in \mathcal{S}_j, \overleftarrow{S} \in \mathcal{S}_i\},$$

as promised. Since the events have the same probability, when conditioned on $\mathcal{S}$, the events $\{\overleftarrow{S}s \in \mathcal{S}_j\}$ and $\{\mathcal{S}' = \mathcal{S}_j, \vec{S}^1 = s\}$ will yield the same conditional probability.[8] QED.

Notice that $T_{ij}^{(\lambda)} = \delta_{ij}$; that is, the transition labeled by the null symbol λ is the identity.

C. ϵ-Machines

The combination of the function ϵ from histories to causal states with the labeled transition probabilities $T_{ij}^{(s)}$ is called the *ϵ-machine* of the process.[5, 6]

Definition 9 (An ϵ-Machine Defined). The *ϵ-machine* of a process is the ordered pair $\{\epsilon, \mathbf{T}\}$, where ϵ is the causal state function and $\mathbf{T}$ is set of the transition matrices for the states defined by ϵ.

[8] Technically, they will yield versions of the same conditional probability, i.e., they will agree with probability 1. See Appendix E.

Equivalently, we may denote an ϵ-machine by $\{\mathcal{S}, \mathbf{T}\}$.

To satisfy the algebraic requirement outlined in Section II F, we make explicit the connection with semi-group theory.

Proposition 1 (ϵ-Machines Are Monoids). The algebra generated by the ϵ-machine $\{\epsilon, \mathbf{T}\}$ is a semi-group with an identity element, i.e., it is a *monoid*.

Proof. See Appendix D.

Remark. Due to this, ϵ-machines can be interpreted as capturing a process's *generalized symmetries*. Any subgroups of an ϵ-machine's semi-group are, in fact, symmetries in the usual sense.

Lemma 5 (ϵ-Machines Are Deterministic). For each $\mathcal{S}_i \in \mathcal{S}$ and each $s \in \mathcal{A}$, there is at most one $\mathcal{S}_j \in \mathcal{S}$ such that, for every history $\overleftarrow{s} \in \mathcal{S}_i$, the history $\overleftarrow{s}s \in \mathcal{S}_j$. If such a $\mathcal{S}_j$ exists, then for all other $\mathcal{S}_k \in \mathcal{S}$, $T_{ik}^{(s)} = 0$. If there is no such $\mathcal{S}_j$, then $T_{ik}^{(s)} = 0$ for all $\mathcal{S}_k \in \mathcal{S}$ whatsoever.

Proof. The first part of the lemma asserts that for all $s \in \mathcal{A}$ and $\overleftarrow{s}, \overleftarrow{s}' \in \overleftarrow{\mathbf{S}}$, if $\epsilon(\overleftarrow{s}) = \epsilon(\overleftarrow{s}')$, then $\epsilon(\overleftarrow{s}s) = \epsilon(\overleftarrow{s}'s)$. ($\overleftarrow{s}s$ is just another history and belongs to one or another causal state.) We show that this follows directly from causal equivalence.

Consider any pair of histories $\overleftarrow{s}, \overleftarrow{s}'$ such that $\epsilon(\overleftarrow{s}) = \epsilon(\overleftarrow{s}')$, any single symbol s, and a (measurable) set F of future events. Let sF denote the set of futures obtained by prefixing the symbol s to each future in F. (sF is also measurable.) By causal equivalence, $\mathrm{P}(\overrightarrow{S} \in sF \,|\, \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\overrightarrow{S} \in sF \,|\, \overleftarrow{S} = \overleftarrow{s}')$. Now, $\overrightarrow{S} \in sF$ can be decomposed into the intersection of two events: $\overrightarrow{S}^1 = s$ and $\overrightarrow{S}_1 \in F$, where $\overrightarrow{S}_1$ is the random variable for the future sequence, ignoring the next symbol. We therefore begin with the following equalities.

$$\mathrm{P}(\overrightarrow{S} \in sF \,|\, \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\overrightarrow{S} \in sF \,|\, \overleftarrow{S} = \overleftarrow{s}')$$

$$\mathrm{P}(\overrightarrow{S}^1 = s, \overrightarrow{S}_1 \in F \,|\, \overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\overrightarrow{S}^1 = s, \overrightarrow{S}_1 \in F \,|\, \overleftarrow{S} = \overleftarrow{s}')$$

For any three random variables X, Y, Z, the conditional probability $\mathrm{P}(Z \in A, Y = y \,|\, X = x)$ can be factored as $\mathrm{P}(Z \in A \,|\, Y = y, X = x) \times \mathrm{P}(Y = y \,|\, X = x)$ (Eq. E4).[9]

$$\mathrm{P}(\overrightarrow{S}_1 \in F \,|\, \overrightarrow{S}^1 = s, \overleftarrow{S} = \overleftarrow{s})\, \mathrm{P}(\overrightarrow{S}^1 = s \,|\, \overleftarrow{S} = \overleftarrow{s})$$
$$= \mathrm{P}(\overrightarrow{S}_1 \in F \,|\, \overrightarrow{S}^1 = s, \overleftarrow{S} = \overleftarrow{s}')\, \mathrm{P}(\overrightarrow{S}^1 = s \,|\, \overleftarrow{S} = \overleftarrow{s}')$$

[9] This assumes the regularity of the conditional probabilities, which is valid for our discrete processes. Again, see Appendix E.

From causal equivalence, the second factors on each side of the equation are equal, so we divide through for them. (We address the case where $P(\vec{S}^1 = s \,|\, \overleftarrow{S} = \overleftarrow{s}) = P(\vec{S}^1 = s \,|\, \overleftarrow{S} = \overleftarrow{s}') = 0$ below.)

$$P(\vec{S}_1 \in F \,|\, \vec{S}^1 = s, \overleftarrow{S} = \overleftarrow{s}) = P(\vec{S}_1 \in F \,|\, \vec{S}^1 = s, \overleftarrow{S} = \overleftarrow{s}')$$
$$P(\vec{S} \in F \,|\, \overleftarrow{S} = \overleftarrow{s}s) = P(\vec{S} \in F \,|\, \overleftarrow{S} = \overleftarrow{s}'s)$$

The last step is justified by stationarity. Since the set F of future events is arbitrary, it follows that $\overleftarrow{s}s \sim_\epsilon \overleftarrow{s}'s$. Consequently, for each $\mathcal{S}_i$ and each s, there is at most one $\mathcal{S}_j$ such that $T_{ij}^{(s)} > 0$.

As remarked, causal equivalence tells us that $P(\vec{S}^1 = s \,|\, \overleftarrow{S} = \overleftarrow{s}) = P(\vec{S}^1 = s \,|\, \overleftarrow{S} = \overleftarrow{s}')$. But they could both be equal to zero, in which case we can't divide through for them. But then, again as promised, it follows that every entry in the transition matrix $T_{ij}^{(s)} = 0$, when $\mathcal{S}_i = \epsilon(\overleftarrow{s})$. Thus, the labeled transition probabilities have the promised form. QED.

Remark 1. In automata theory,[67, 68] a set of states and transitions is said to be *deterministic* if the current state and the next input—here, the next symbol from the original stochastic process—together fix the next state. This use of the word "deterministic" is often confusing, since many stochastic processes (e.g., simple Markov chains) are deterministic in this sense.

Remark 2. Starting from a fixed state, a given symbol always leads to at most one single state. But there can be several transitions from one state to another, each labeled with a different symbol.

Remark 3. Clearly, if $T_{ij}^{(s)} > 0$, then $T_{ij}^{(s)} = P(\vec{S}^1 = s \,|\, \mathcal{S} = \mathcal{S}_i)$. In automata theory the "disallowed" transitions ($T_{ij}^{(s)} = 0$) are sometimes explicitly represented and lead to a "reject" state indicating that the particular history does not occur.

Lemma 6 (ε-Machines Are Markovian) Given the causal state at time $t-1$, the causal state at time t is independent of the causal state at earlier times.

Proof. We start by showing that, writing $\mathcal{S}$, $\mathcal{S}'$, $\mathcal{S}''$ for the sequence of causal states at three successive times, $\mathcal{S}$ and $\mathcal{S}''$ are conditionally independent, given $\mathcal{S}'$.

Let M be a (measurable) set of causal states.

$$P(\mathcal{S}'' \in \mathrm{M} \,|\, \mathcal{S}' = \sigma', \mathcal{S} = \sigma) = P(\vec{S}^1 \in A \,|\, \mathcal{S}' = \sigma', \mathcal{S} = \sigma),$$

where $A \subseteq \mathcal{A}$ is the set of all symbols that lead from σ' to some $\sigma'' \in \mathrm{M}$. This is a well-defined and measurable set, in virtue of Lemma 5 immediately preceding, which also guarantees (see Remark 3 to the Lemma) the equality of conditional probabilities we have used. Invoking Lemma 3, conditioning on $\mathcal{S}$ has no further effect once we have conditioned on $\mathcal{S}'$,

$$\begin{aligned} \mathrm{P}(\vec{S}^1 \in A \,|\, \mathcal{S}' = \sigma', \mathcal{S} = \sigma) &= \mathrm{P}(\vec{S}^1 \in A \,|\, \mathcal{S}' = \sigma') \\ &= \mathrm{P}(\mathcal{S}'' \in \mathrm{M} \,|\, \mathcal{S}' = \sigma') \end{aligned}$$

But (Proposition 4, Appendix E) and Eq. (E3)) this is true if and only if conditional independence holds. Now the lemma follows by straightforward mathematical induction. QED.

Remark 1. This lemma strengthens the claim that the causal states are, in fact, the causally efficacious states: given knowledge of the present state, what has gone before makes no difference. (Again, recall the philosophical preliminaries of Section II E.)

Remark 2. This result indicates that the causal states, considered as a process, define a kind of Markov process. Thus, causal states can be roughly considered to be a generalization of Markovian states. We say "kind of" since the class of ϵ-machines is substantially richer[5, 10] than what one normally associates with Markov processes.[69, 70]

Definition 10 (ϵ-Machine Reconstruction). *ϵ-Machine reconstruction* is any procedure that given a process $\mathrm{P}(\overleftrightarrow{S})$ (respectively an approximation of $\mathrm{P}(\overleftrightarrow{S})$), produces the process's ϵ-machine $\{\mathcal{S}, \mathrm{T}\}$ (respectively an approximation of $\{\mathcal{S}, \mathrm{T}\}$).

Given a mathematical description of a process, one can often calculate analytically its ϵ-machine. (For example, see the computational mechanics analysis of spin systems in ref. 66.) There is also a wide range of algorithms which reconstruct ϵ-machines from empirical estimates of $\mathrm{P}(\overleftrightarrow{S})$. Some, such as those used in refs. 5–7 and 71, operate in "batch" mode, taking the raw data as a whole and producing the ϵ-machine. Others could operate incrementally, in "on-line" mode, taking in individual measurements and re-estimating the set of causal states and their transition probabilities.

V. OPTIMALITIES AND UNIQUENESS

We now show that: causal states are maximally accurate predictors of minimal statistical complexity; they are unique in sharing both properties; and their state-to-state transitions are minimally stochastic. In other words,

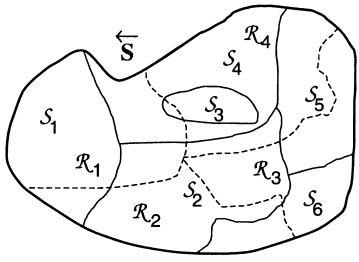

Fig. 3. An alternative class $\mathcal{R}$ of states (delineated by dashed lines) that partition $\overleftarrow{\mathbf{S}}$ overlaid on the causal states $\mathcal{S}$ (outlined by solid lines). Here, for example, $\mathcal{S}_2$ contains parts of $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$ and $\mathcal{R}_4$. The collection of all such alternative partitions form *Occam's pool*. Note again that the $\mathcal{R}_i$ need not be compact nor simply connected, as drawn.

they satisfy both of the constraints borrowed from Occam, and they are the only representations that do so. The overarching moral here is that causal states and ϵ-machines are *the* goals in any learning or modeling scheme. The argument is made by the time-honored means of proving optimality theorems. We address, in our concluding remarks (Section VII), the practicalities involved in attaining these goals.

As part of our strategy, though, we also prove several results that are not optimality results; we call these lemmas to indicate their subordinate status. All of our theorems, and some of our lemmas, will be established by comparing causal states, generated by ϵ, with other rival sets of states, generated by other functions η. In short, none of the rival states—none of the other patterns—can out-perform the causal states.

It is convenient to recall some notation before plunging in. Let $\mathcal{S}$ be the random variable for the current causal state, $\vec{S}^1 \in \mathcal{A}$ the next "observable" we get from the original stochastic process, $\mathcal{S}'$ the next causal state, $\mathcal{R}$ the current state according to η, and $\mathcal{R}'$ the next η-state. σ will stand for a particular value (causal state) of $\mathcal{S}$ and ρ a particular value of $\mathcal{R}$. When we quantify over alternatives to the causal states, we quantify over $\mathcal{R}$.

Theorem 1 (Causal States Are Maximally Prescient)[16]. For all $\mathcal{R}$ and all $L \in \mathbb{Z}^+$,

$$H[\vec{S}^L|\mathcal{R}] \geqslant H[\vec{S}^L|\mathcal{S}]. \tag{21}$$

Proof. We have already seen that $H[\vec{S}^L|\mathcal{R}] \geqslant H[\vec{S}^L|\overleftarrow{S}]$ (Lemma 1). But by construction (Definition 5),

$$\mathrm{P}(\vec{S}^L = \vec{s}^L|\overleftarrow{S} = \overleftarrow{s}) = \mathrm{P}(\vec{S}^L = \vec{s}^L|\mathcal{S} = \epsilon(\overleftarrow{s})). \tag{22}$$

Since entropies depend only on the probability distribution, $H[\vec{S}^L|\mathcal{S}] = H[\vec{S}^L|\overleftarrow{S}]$ for every L. Thus, $H[\vec{S}^L|\mathcal{R}] \geqslant H[\vec{S}^L|\mathcal{S}]$, for all L. QED.

Remark. That is to say, causal states are as good at predicting the future—are as *prescient*—as complete histories. In this, they satisfy the first requirement borrowed from Occam. Since the causal states are well defined and since they can be systematically approximated, we have shown that the upper bound on the strength of patterns (Definition 3 and Lemma 1, Remark) can in fact be reached. Intuitively, the causal states achieve this because, unlike effective states in general, they do not throw away any information about the future which might be contained in $\overleftarrow{S}$. Even more colloquially, to paraphrase the definition of information in ref. 72, the causal states record every difference (about the past) that makes a difference (to the future). We can actually make this intuition quite precise, in an easy corollary to the theorem.

Corollary 1 (Causal States Are Sufficient Statistics). The causal states $\mathcal{S}$ of a process are sufficient statistics for predicting it.

Proof. It follows from Theorem 1 and Eq. (8) that, for all $L \in \mathbb{Z}^+$,

$$I[\vec{S}^L; \mathcal{S}] = I[\vec{S}^L; \overleftarrow{S}], \tag{23}$$

where I was defined in Eq. (8). Consequently, the causal state is a *sufficient statistic*—see refs. 63, p. 37 and 73, section 2.4–2.5—for predicting futures of any length. QED.

All subsequent results concern rival states that are as prescient as the causal states. We call these *prescient rivals* and denote a class of them $\widehat{\mathcal{R}}$.

Definition 11 (Prescient Rivals). *Prescient rivals* $\widehat{\mathcal{R}}$ are states that are as predictive as the causal states; viz., for all $L \in \mathbb{Z}^+$,

$$H[\vec{S}^L|\widehat{\mathcal{R}}] = H[\vec{S}^L|\mathcal{S}]. \tag{24}$$

Remark. Prescient rivals are also sufficient statistics.

Corollary 2 (A Sufficient Condition for Prescience). If $\mathrm{P}(\vec{S}^1 = a|\mathcal{R} = \eta(\overleftarrow{s})) = \mathrm{P}(\vec{S}^1 = a|\mathcal{S} = \epsilon(\overleftarrow{s}))$ for all $a \in \mathcal{A}$, and $\mathcal{R}$ is deterministic (in the sense of Lemma 5), then $\mathcal{R}$ is prescient. That is, deterministic states which get the distribution of the next symbol right are prescient.

Proof. It will be enough to show that, for any L, $\mathrm{P}(\vec{S}^L|\mathcal{R}) = \mathrm{P}(\vec{S}^L|\mathcal{S})$, since then the equality of conditional entropies is obvious. We do this by induction; suppose that the equality of conditional probabilities

holds for all lengths of futures up to some L, and consider futures of length $L+1$.

$$\begin{aligned}
&\mathrm{P}(\vec{S}^{L+1} = s^L a \,|\, \mathcal{R} = \eta(\overleftarrow{s})) \\
&\quad = \mathrm{P}(\vec{S}_{L+1} = a \,|\, \mathcal{R} = \eta(\overleftarrow{s}), \vec{S}^L = s^L)\, \mathrm{P}(\vec{S}^L = s^L \,|\, \mathcal{R} = \eta(\overleftarrow{s})) \\
&\quad = \mathrm{P}(\vec{S}_{L+1} = a \,|\, \mathcal{R} = \eta(\overleftarrow{s}), \vec{S}^L = s^L)\, \mathrm{P}(\vec{S}^L = s^L \,|\, \mathcal{S} = \epsilon(\overleftarrow{s})) \qquad (25)
\end{aligned}$$

where the second equality uses the inductive hypothesis. Since we assume the $\mathcal{R}$ states are deterministic, the combination of the current effective state ($\eta(\overleftarrow{s})$) and the next L symbols (s^L) fixes a unique future effective state, namely $\eta(\overleftarrow{s}s^L)$. Thus, by Proposition 3, Appendix E, we see that $\mathrm{P}(\vec{S}_{L+1} = a \,|\, \mathcal{R} = \eta(\overleftarrow{S}),\ \vec{S}^L = s^L) = \mathrm{P}(\vec{S}^1 = a \,|\, \mathcal{R} = \eta(\overleftarrow{s}s^L))$. Substituting back in,

$$\begin{aligned}
&\mathrm{P}(\vec{S}^{L+1} = s^L a \,|\, \mathcal{R} = \eta(\overleftarrow{s})) \qquad (26) \\
&\quad = \mathrm{P}(\vec{S}^1 = a \,|\, \mathcal{R} = \eta(\overleftarrow{s}s^L))\, \mathrm{P}(\vec{S}^L = s^L \,|\, \mathcal{S} = \epsilon(\overleftarrow{s}) \\
&\quad = \mathrm{P}(\vec{S}^1 = a \,|\, \mathcal{S} = \epsilon(\overleftarrow{s}s^L))\, \mathrm{P}(\vec{S}^L = s^L \,|\, \mathcal{S} = \epsilon(\overleftarrow{s}) \\
&\quad = \mathrm{P}(\vec{S}^{L+1} = s^L a \,|\, \mathcal{S} = \epsilon(\overleftarrow{s})), \qquad (27)
\end{aligned}$$

so the induction is established. Since (by hypothesis) it holds for $L = 1$, it holds for all positive L. QED.

Remark. The causal states satisfy the hypotheses of this proposition. Since, as we shall see (Theorem 2), the causal states are the minimal prescient states, they are also the minimal deterministic states which get the distribution of the next symbol right. This observation is useful in designing ϵ-machine reconstruction procedures.[74]

Lemma 7 (Refinement Lemma). For all prescient rivals $\hat{\mathcal{R}}$ and for each $\hat{\rho} \in \hat{\mathcal{R}}$, there is a $\sigma \in \mathcal{S}$ and a measure-0 subset $\hat{\rho}_0 \subset \hat{\rho}$, possibly empty, such that $\hat{\rho} \setminus \hat{\rho}_0 \subseteq \sigma$, where $\setminus$ is set subtraction.

Proof. We invoke a straightforward extension of Theorem 2.7.3 of ref. 63: If $X_1, X_2, \ldots, X_n$ are random variables over the same set $\mathcal{A}$, each with distinct probability distributions, Θ a random variable over the integers from 1 to n such that $\mathrm{P}(\Theta = i) = \lambda_i$, and Z a random variable over $\mathcal{A}$ such that $Z = X_\Theta$, then

$$\begin{aligned}
H[Z] &= H\left[\sum_{i=1}^{n} \lambda_i X_i\right] \\
&\geqslant \sum_{i=1}^{n} \lambda_i H[X_i]. \qquad (28)
\end{aligned}$$

In words, the entropy of a mixture of distributions is at least the mean of the entropies of those distributions. This follows since H is strictly concave, which in turn follows from $x \log x$ being strictly convex for $x \geqslant 0$. We obtain equality in Eq. (28) if and only if all the λ_i are either 0 or 1, i.e., if and only if Z is at least weakly homogeneous (Definition 7).

The conditional distribution of futures for each rival state ρ can be written as a weighted mixture of the morphs of one or more causal states. (Cf. Fig. 3.) Thus, by Eq. (28), unless every ρ is at least weakly homogeneous with respect to $\vec{S}^L$ (for each L), the entropy of $\vec{S}^L$ conditioned on $\mathcal{R}$ will be higher than the minimum, the entropy conditioned on $\mathcal{S}$. So, in the case of the maximally predictive $\hat{\mathcal{R}}$, every $\hat{\rho} \in \hat{\mathcal{R}}$ must be at least weakly homogeneous with respect to all $\vec{S}^L$. But the causal states are the largest classes that are strictly homogeneous with respect to all $\vec{S}^L$ (Lemma 3). Thus, the strictly homogeneous part of each $\hat{\rho} \in \hat{\mathcal{R}}$ must be a subclass, possibly improper, of some causal state $\sigma \in \boldsymbol{\mathcal{S}}$. QED.

Remark 1. An alternative proof appears in Appendix F.

Remark 2. The content of the lemma can be made quite intuitive, if we ignore for a moment the measure-0 set $\hat{\rho}_0$ of histories mentioned in its statement. It then asserts that any alternative partition $\hat{\mathcal{R}}$ that is as prescient as the causal states must be a refinement of the causal-state partition. That is, each $\hat{\mathcal{R}}_i$ must be a (possibly improper) subset of some $\mathcal{S}_j$. Otherwise, at least one $\hat{\mathcal{R}}_i$ would have to contain parts of at least two causal states. And so, using this $\hat{\mathcal{R}}_i$ to predict the future observables would lead to more uncertainty about $\vec{S}$ than using the causal states. This is illustrated by Fig. 4, which should be contrasted with Fig. 3.

Adding the measure-0 set $\hat{\rho}_0$ of histories to this picture does not change its heuristic content much. Precisely because these histories have zero probability, treating them in an "inappropriate" way makes no discernible difference to predictions, morphs, and so on. There is a problem of terminology, however, since there seems to be no standard name for the relationship between the partitions $\hat{\mathcal{R}}$ and $\boldsymbol{\mathcal{S}}$. We propose to say that the former is a refinement of the latter *almost everywhere* or, simply, a *refinement a.e.*

Remark 3. One cannot work the proof the other way around to show that the causal states have to be a refinement of the equally prescient $\hat{\mathcal{R}}$-states. This is precluded because applying the theorem borrowed from ref. 63, Eq. (28), hinges on being able to reduce uncertainty by specifying from *which* distribution one chooses. Since the causal states are constructed so as to be strictly homogeneous with respect to futures, this is not the case. Lemma 3 and Theorem 1 together protect us.

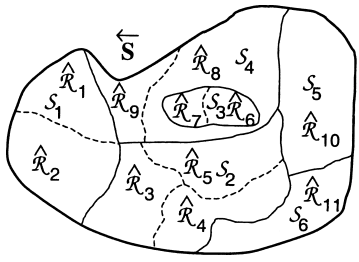

Fig. 4. A prescient rival partition $\hat{\mathcal{R}}$ must be a refinement of the causal-state partition *almost everywhere*. That is, almost all of each $\hat{\mathcal{R}}_i$ must be contained within some $\mathcal{S}_j$; the exceptions, if any, are a set of histories of measure 0. Here for instance $\mathcal{S}_2$ contains the positive-measure parts of $\hat{\mathcal{R}}_3$, $\hat{\mathcal{R}}_4$, and $\hat{\mathcal{R}}_5$. One of these rival states, say $\hat{\mathcal{R}}_3$, could have member-histories in any or all of the other causal states, provided the total measure of such exceptional histories is zero. Cf. Fig. 3.

Remark 4. Because almost all of each prescient rival state is wholly contained within a single causal state, we can construct a function $g: \hat{\mathcal{R}} \mapsto \boldsymbol{\mathcal{S}}$, such that, if $\eta(\overleftarrow{s}) = \hat{\rho}$, then $\epsilon(\vec{S}) = g(\hat{\rho})$ almost always. We can even say that $\mathcal{S} = g(\hat{\mathcal{R}})$ almost always, with the understanding that this means that, for each $\hat{\rho}$, $\mathrm{P}(\mathcal{S} = g(\hat{\rho}) \,|\, \hat{\mathcal{R}} = \hat{\rho}) = 1$.

Theorem 2 (Causal States Are Minimal)[6, 16]. For all prescient rivals $\hat{\mathcal{R}}$,

$$C_\mu(\hat{\boldsymbol{\mathcal{R}}}) \geqslant C_\mu(\mathcal{S}). \tag{29}$$

Proof. By Lemma 7, Remark 4, there is a function g such that $\mathcal{S} = g(\hat{\mathcal{R}})$ almost always. But $H[f(X)] \leqslant H[X]$ (Eq. (A11)) and so

$$H[\mathcal{S}] = H[g(\hat{\mathcal{R}})] \leqslant H[\hat{\mathcal{R}}]. \tag{30}$$

but $C_\mu(\hat{\boldsymbol{\mathcal{R}}}) = H[\hat{\mathcal{R}}]$ (Definition 4). QED.

Remark 1. We have just established that no rival pattern, which is as good at predicting the observations as the causal states, is any simpler, in the sense given by Definition 4, than the causal states. (This is the theorem of ref. 6.) Occam therefore tells us that there is no reason not to use the causal states. The next theorem shows that causal states are uniquely optimal and so that Occam's Razor all but forces us to use them.

Remark 2. Here it becomes important that we are trying to predict the whole of $\vec{S}$ and not just some piece, $\vec{S}^L$. Suppose two histories $\overleftarrow{s}$ and $\overleftarrow{s}'$

have the same conditional distribution for futures of lengths up to L, but differing ones after that. They would then belong to different causal states. An η-state that merged those two causal states, however, would have just as much ability to predict $\vec{S}^L$ as the causal states. More, these $\mathcal{R}$-states would be simpler, in the sense that the uncertainty in the current state would be lower. We conclude that causal states are optimal, but for the hardest job—that of predicting futures of all lengths.

Remark 3. We have already seen (Theorem 1, Remark 2) that causal states are sufficient statistics for predicting futures of all lengths; so are all prescient rivals. A *minimal* sufficient statistic is one that is a function of all other sufficient statistics (ref. 63, p. 38). Since, in the course of the proof of Theorem 2, we have shown that there is a function g from any $\hat{\mathcal{R}}$ to $\mathcal{S}$, we have also shown that causal states are minimal sufficient statistics.

We may now, as promised, define the *statistical complexity of a process.*[5, 6]

Definition 12 (Statistical Complexity of a Process). The *statistical complexity "$C_\mu(\mathcal{O})$" of a process* $\mathcal{O}$ is that of its causal states: $C_\mu(\mathcal{O}) \equiv C_\mu(\boldsymbol{\mathcal{S}})$.

Due to the minimality of causal states we see that the statistical complexity measures the average amount of historical memory stored in the process. Without the minimality theorem, this interpretation would not be possible, since we could trivially elaborate internal states, while still generating the same observed process. C_μ for those states would grow without bound and so be arbitrary and not a characteristic property of the process.[18]

Theorem 3 (Causal States Are Unique). For all prescient rivals $\hat{\mathcal{R}}$, if $C_\mu(\hat{\mathcal{R}}) = C_\mu(\boldsymbol{\mathcal{S}})$, then there exists an invertible function between $\hat{\mathcal{R}}$ and $\boldsymbol{\mathcal{S}}$ that almost always preserves equivalence of state: $\hat{\mathcal{R}}$ and η are the same as $\boldsymbol{\mathcal{S}}$ and ϵ, respectively, except on a set of histories of measure 0.

Proof. From Lemma 7, we know that $\mathcal{S} = g(\hat{\mathcal{R}})$ almost always. We now show that there is a function f such that $\hat{\mathcal{R}} = f(\mathcal{S})$ almost always, implying that $g = f^{-1}$ and that f is the desired relation between the two sets of states. To do this, by Eq. (A12) it is sufficient to show that $H[\hat{\mathcal{R}}|\mathcal{S}] = 0$. Now, it follows from an information-theoretic identity (Eq. (A8)) that

$$H[\mathcal{S}] - H[\mathcal{S}|\hat{\mathcal{R}}] = H[\hat{\mathcal{R}}] - H[\hat{\mathcal{R}}|\mathcal{S}]. \tag{31}$$

Since, by Lemma 7 $H[\mathcal{S}|\widehat{\mathcal{R}}] = 0$, both sides of Eq. (31) are equal to $H[\mathcal{S}]$. But, by hypothesis, $H[\widehat{\mathcal{R}}] = H[\mathcal{S}]$. Thus, $H[\widehat{\mathcal{R}}|\mathcal{S}] = 0$ and so there exists an f such that $\widehat{\mathcal{R}} = f(\mathcal{S})$ almost always. We have then that $f(g(\widehat{\mathcal{R}})) = \widehat{\mathcal{R}}$ and $g(f(\mathcal{S})) = \mathcal{S}$, so $g = f^{-1}$. This implies that f preserves equivalence of states almost always: for almost all $\overleftarrow{s}, \overleftarrow{s}' \in \overleftarrow{\mathbf{S}}$, $\eta(\overleftarrow{s}) = \eta(\overleftarrow{s}')$ if and only if $\epsilon(\overleftarrow{s}) = \epsilon(\overleftarrow{s}')$. QED.

Remark. As in the case of the Refinement Lemma 7, on which the theorem is based, the measure-0 caveats seem unavoidable. A rival that is as predictive and as simple (in the sense of Definition 4) as the causal states, can assign a measure-0 set of histories to different states than the ϵ-machine does, but no more. This makes sense. Such a measure-0 set makes no difference, since its members are never observed, by definition. By the same token, however, nothing prevents a minimal, prescient rival from disagreeing with the ϵ-machine on those histories.

Theorem 4 (ε-Machines Are Minimally Stochastic)[16]. For all prescient rivals $\widehat{\mathcal{R}}$,

$$H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}] \geqslant H[\mathcal{S}'|\mathcal{S}], \tag{32}$$

where $\mathcal{S}'$ and $\widehat{\mathcal{R}}'$ are the next causal state of the process and the next η-state, respectively.

Proof. From Lemma 5, $\mathcal{S}'$ is fixed by $\mathcal{S}$ and $\vec{S}^1$ together, thus $H[\mathcal{S}'|\mathcal{S}, \vec{S}^1] = 0$ by Eq. (A12). Therefore, from the chain rule for entropies Eq. (A6),

$$H[\vec{S}^1|\mathcal{S}] = H[\mathcal{S}', \vec{S}^1|\mathcal{S}]. \tag{33}$$

We have no result like the Determinism Lemma 5 for the rival states $\widehat{\mathcal{R}}$, but entropies are always non-negative: $H[\widehat{\mathcal{R}}'|\widehat{\mathcal{R}}, \vec{S}^1] \geqslant 0$. Since for all L, $H[\vec{S}^L|\widehat{\mathcal{R}}] = H[\vec{S}^L|\mathcal{S}]$ by the definition, Definition 11, of prescient rivals, $H[\vec{S}^1|\widehat{\mathcal{R}}] = H[\vec{S}^1|\mathcal{S}]$. Now we apply the chain rule again,

$$\begin{aligned} H[\widehat{\mathcal{R}}', \vec{S}^1|\widehat{\mathcal{R}}] &= H[\vec{S}^1|\widehat{\mathcal{R}}] + H[\widehat{\mathcal{R}}'|\vec{S}^1, \widehat{\mathcal{R}}] && (34)\\ &\geqslant H[\vec{S}^1|\widehat{\mathcal{R}}] && (35)\\ &= H[\vec{S}^1|\mathcal{S}] && (36)\\ &= H[\mathcal{S}', \vec{S}^1|\mathcal{S}] && (37)\\ &= H[\mathcal{S}'|\mathcal{S}] + H[\vec{S}^1|\mathcal{S}', \mathcal{S}]. && (38) \end{aligned}$$

In going from Eq. (36) to Eq. (37) we have used Eq. (33), and in the last step we have used the chain rule once more.

Using the chain rule one last time, we have

$$H[\hat{\mathcal{R}}', \vec{S}^1 | \hat{\mathcal{R}}] = H[\hat{\mathcal{R}}' | \hat{\mathcal{R}}] + H[\vec{S}^1 | \hat{\mathcal{R}}', \hat{\mathcal{R}}]. \tag{39}$$

Putting these expansions, Eqs. (38) and (39), together we get

$$\begin{aligned} H[\hat{\mathcal{R}}' | \hat{\mathcal{R}}] + H[\vec{S}^1 | \hat{\mathcal{R}}', \hat{\mathcal{R}}] &\geqslant H[\mathcal{S}' | \mathcal{S}] + H[\vec{S}^1 | \mathcal{S}', \mathcal{S}] \\ H[\hat{\mathcal{R}}' | \hat{\mathcal{R}}] - H[\mathcal{S}' | \mathcal{S}] &\geqslant H[\vec{S}^1 | \mathcal{S}', \mathcal{S}] - H[\vec{S}^1 | \hat{\mathcal{R}}', \hat{\mathcal{R}}]. \end{aligned} \tag{40}$$

From Lemma 7, we know that $\mathcal{S} = g(\hat{\mathcal{R}})$, so there is another function g' from ordered pairs of η-states to ordered pairs of causal states: $(\mathcal{S}', \mathcal{S}) = g'(\hat{\mathcal{R}}', \hat{\mathcal{R}})$. Therefore, Eq. (A14) implies

$$H[\vec{S}^1 | \mathcal{S}', \mathcal{S}] \geqslant H[\vec{S}^1 | \hat{\mathcal{R}}', \hat{\mathcal{R}}]. \tag{41}$$

And so, we have that

$$\begin{aligned} H[\vec{S}^1 | \mathcal{S}', \mathcal{S}] - H[\vec{S}^1 | \hat{\mathcal{R}}', \hat{\mathcal{R}}] &\geqslant 0 \\ H[\hat{\mathcal{R}}' | \hat{\mathcal{R}}] - H[\mathcal{S}' | \mathcal{S}] &\geqslant 0 \\ H[\hat{\mathcal{R}}' | \hat{\mathcal{R}}] &\geqslant H[\mathcal{S}' | \mathcal{S}]. \end{aligned} \tag{42}$$

QED.

Remark. What this theorem says is that there is no more uncertainty in transitions between causal states, than there is in the transitions between any other kind of prescient effective states. In other words, the causal states approach as closely to perfect determinism—in the usual physical, non-computation-theoretic sense—as any rival that is as good at predicting the future.

VI. BOUNDS

In this section we develop bounds between measures of structural complexity and entropy derived from ϵ-machines and those from ergodic and information theories, which are perhaps more familiar.

Definition 13 (Excess Entropy). The *excess entropy* $\mathbf{E}$ of a process is the mutual information between its semi-infinite past and its semi-infinite future:

$$\mathbf{E} \equiv I[\vec{S}; \overleftarrow{S}]. \tag{43}$$

The excess entropy is a frequently-used measure of the complexity of stochastic processes and appears under a variety of names; e.g., "predictive information", "stored information", "effective measure complexity", and so on.[75–81] $\mathbf{E}$ measures the amount of *apparent* information stored in the observed behavior about the past. As we now establish, $\mathbf{E}$ is not, in general, the amount of memory that the process stores *internally* about its past; a quantity measured by C_μ.

Theorem 5 (The Bounds of Excess). The statistical complexity C_μ bounds the excess entropy $\mathbf{E}$:

$$\mathbf{E} \leqslant C_\mu, \tag{44}$$

with equality if and only if $H[\mathcal{S}|\vec{S}] = 0$.

Proof. $\mathbf{E} = I[\vec{S}; \overleftarrow{S}] = H[\vec{S}] - H[\vec{S}|\overleftarrow{S}]$ and, by the construction of causal states, $H[\vec{S}|\overleftarrow{S}] = H[\vec{S}|\mathcal{S}]$, so

$$\mathbf{E} = H[\vec{S}] - H[\vec{S}|\mathcal{S}] = I[\vec{S}; \mathcal{S}]. \tag{45}$$

Thus, since the mutual information between two variables is never larger than the self-information of either one of them (Eq. (A9)), $\mathbf{E} \leqslant H[\mathcal{S}] = C_\mu$, with equality if and only if $H[\mathcal{S}|\vec{S}] = 0$. QED.

Remark 1. Note that we have invoked $H[\vec{S}]$, not $H[\vec{S}^L]$, but only while subtracting off quantities like $H[\vec{S}|\overleftarrow{S}]$. We need not worry, therefore, about the existence of a finite $L \to \infty$ limit for $H[\vec{S}^L]$, just that of a finite $L \to \infty$ limit for $I[\vec{S}^L; \overleftarrow{S}]$ and $I[\vec{S}^L; \mathcal{S}]$. There are many elementary cases (e.g., the fair-coin process) where the latter limits exist, while the former do not. (See ref. 62 for details on how to construct such a mutual information with full rigor.)

Remark 2. At first glance, it is tempting to see $\mathbf{E}$ as the amount of information stored in a process. As Theorem 5 shows, this temptation should be resisted. $\mathbf{E}$ is only a lower bound on the true amount of information the process stores about its history, namely C_μ. We can, however, say that $\mathbf{E}$ measures the *apparent* information in the process, since it is defined directly in terms of observed sequences and not in terms of hidden, intrinsic states, as C_μ is.

Remark 3. Perhaps another way to describe what $\mathbf{E}$ measures is to note that, by its implicit assumption of block-Markovian structure, it takes sequence-blocks as states. But even for the class of block-Markovian sources, for which such an assumption is appropriate, excess entropy and

statistical complexity measure different kinds of information storage. Refs. 66 and 82 showed that in the case of one-dimensional range-R spin systems, or any other block-Markovian source where block configurations are isomorphic to causal states:

$$C_\mu = \mathbf{E} + Rh_\mu, \tag{46}$$

for finite R. Only for zero-entropy-rate block-Markovian sources will the excess entropy, a quantity estimated directly from sequence blocks, equal the statistical complexity, the amount of memory stored in the process. Examples of such sources include periodic processes, for which we have $C_\mu = \mathbf{E} = \log_2 p$, where p is the period.

Corollary 3. For all prescient rivals $\hat{\mathcal{R}}$,

$$\mathbf{E} \leqslant H[\hat{\mathcal{R}}]. \tag{47}$$

Proof. This follows directly from Theorem 29, since $H[\hat{\mathcal{R}}] \geqslant C_\mu$. QED.

Lemma 8 (Conditioning Does Not Affect Entropy Rate). For all prescient rivals $\hat{\mathcal{R}}$,

$$h[\vec{S}] = h[\vec{S} \,|\, \hat{\mathcal{R}}], \tag{48}$$

where the entropy rate $h[\vec{S}]$ and the conditional entropy rate $h[\vec{S} \,|\, \hat{\mathcal{R}}]$ were defined in Eq. (9) and Eq. (10), respectively.

Proof. From Theorem 5 (Eq. (44)) and its Corollary 3 (Eq. (47)), we have

$$\lim_{L \to \infty} (H[\vec{S}^L] - H[\vec{S}^L \,|\, \hat{\mathcal{R}}]) \leqslant \lim_{L \to \infty} H[\hat{\mathcal{R}}], \tag{49}$$

or,

$$\lim_{L \to \infty} \frac{H[\vec{S}^L] - H[\vec{S}^L \,|\, \hat{\mathcal{R}}]}{L} \leqslant \lim_{L \to \infty} \frac{H[\hat{\mathcal{R}}]}{L}. \tag{50}$$

Since, by Eq. (A4), $H[\vec{S}^L] - H[\vec{S}^L \,|\, \hat{\mathcal{R}}] \geqslant 0$, we have

$$h[\vec{S}] - h[\vec{S} \,|\, \hat{\mathcal{R}}] = 0. \tag{51}$$

QED.

Remark. Forcing the process into a certain state $\hat{\mathcal{R}} = \hat{\rho}$ is akin to applying a controller, once. But in the infinite-entropy case, $H[\vec{S}^L] \to_{L\to\infty} \infty$, with which we are concerned, the future could contain (or consist of) an infinite sequence of disturbances. In the face of this "grand disturbance", the effects of the finite control are simply washed out.

Another way of viewing this is to reflect on the fact that $h[\vec{S}]$ accounts for the effects of all the dependencies between all the parts of the entire semi-infinite future. This, owing to the time-translation invariance of stationarity, is equivalent to taking account of all the dependencies in the entire process, including those between past and future. But these are what is captured by $h[\vec{S}\,|\,\hat{\mathcal{R}}]$. It is not that conditioning on $\mathcal{R}$ fails to reduce our uncertainty about the future; it does so, for all finite times, and conditioning on $\mathcal{S}$ achieves the maximum possible reduction in uncertainty. Rather, the lemma asserts that such conditioning cannot affect the asymptotic rate at which such uncertainty grows with time.

Theorem 6 (Control Theorem). Given a class $\hat{\mathcal{R}}$ of prescient rivals,

$$H[S] - h[\vec{S}\,|\,\hat{\mathcal{R}}] \leqslant C_\mu, \tag{52}$$

where $H[S]$ is the entropy of a single symbol from the observable stochastic process.

Proof. As is well known (ref. 63, Theorem 4.2.1, p. 64), for any stationary stochastic process,

$$\lim_{L\to\infty} \frac{H[\vec{S}^L]}{L} = \lim_{L\to\infty} H[S_L\,|\,\vec{S}^{L-1}]. \tag{53}$$

Moreover, the limits always exist. Up to this point, we have defined $h[\vec{S}]$ in the manner of the left-hand side; recall Eq. (9). It will be convenient in the following to use that of the right-hand side.

From the definition of conditional entropy, we have

$$\begin{aligned} H[\overleftarrow{S}^L] &= H[\overleftarrow{S}^1\,|\,\overleftarrow{S}^{L-1}] + H[\overleftarrow{S}^{L-1}] \\ &= H[\overleftarrow{S}^{L-1}\,|\,\overleftarrow{S}^1] + H[\overleftarrow{S}^1]. \end{aligned} \tag{54}$$

So we can express the entropy of the last observable the process generated before the present as

$$H[\overleftarrow{S}^1] = H[\overleftarrow{S}^L] - H[\overleftarrow{S}^{L-1}\,|\,\overleftarrow{S}^1] \tag{55}$$

$$= H[\overleftarrow{S}^1\,|\,\overleftarrow{S}^{L-1}] + H[\overleftarrow{S}^{L-1}] - H[\overleftarrow{S}^{L-1}\,|\,\overleftarrow{S}^1] \tag{56}$$

$$= H[\overleftarrow{S}^1\,|\,\overleftarrow{S}^{L-1}] + I[\overleftarrow{S}^{L-1}; \overleftarrow{S}^1]. \tag{57}$$

We go from Eq. (55) to Eq. (56) by substituting the first RHS of Eq. (54) for $H[\overleftarrow{S}^L]$.

Taking the $L \to \infty$ limit has no effect on the LHS,

$$H[\overleftarrow{S}^1] = \lim_{L \to \infty} (H[\overleftarrow{S}^1 | \overleftarrow{S}^{L-1}] + I[\overleftarrow{S}^{L-1}; \overleftarrow{S}^1]). \tag{58}$$

Since the process is stationary, we can move the first term in the limit forward to $H[S_L | \overrightarrow{S}^{L-1}]$. This limit is $h[\overrightarrow{S}]$, by Eq. (53). Furthermore, because of stationarity, $H[\overleftarrow{S}^1] = H[\overrightarrow{S}^1] = H[S]$. Shifting the entropy rate $h[\overrightarrow{S}]$ to the LHS of Eq. (58) and appealing to time-translation once again, we have

$$\begin{aligned} H[S] - h[\overrightarrow{S}] &= \lim_{L \to \infty} I[\overleftarrow{S}^{L-1}; \overleftarrow{S}^1] && (59) \\ &= I[\overleftarrow{S}; \overrightarrow{S}^1] && (60) \\ &= H[\overrightarrow{S}^1] - H[\overrightarrow{S}^1 | \overleftarrow{S}] && (61) \\ &= H[\overrightarrow{S}^1] - H[\overrightarrow{S}^1 | \mathcal{S}] && (62) \\ &= I[\overrightarrow{S}^1; \mathcal{S}] && (63) \\ &\leqslant H[\mathcal{S}] = C_\mu, && (64) \end{aligned}$$

where the last inequality comes from Eq. (A9). QED.

Remark 1. The Control Theorem is inspired by, and is a version of, Ashby's *law of requisite variety* (ref. 83, ch. 11). This states that applying a controller can reduce the uncertainty in the controlled variable by at most the entropy of the control variable. (This result has recently been rediscovered in ref. 84.) Thinking of the controlling variable as the causal state, we have here a limitation on the controller's ability to reduce the entropy *rate*.

Remark 2. This is the only result so far where the difference between the finite-L and the infinite-L cases is important. For the analogous result in the finite case, see Appendix G, Theorem 7.

Remark 3. By applying Theorem 29 and Lemma 48, we could go from the theorem as it stands to $H[S] - h[\overrightarrow{S} | \widehat{\mathcal{R}}] \leqslant H[\widehat{\mathcal{R}}]$. This has a pleasing appearance of symmetry to it, but is actually a weaker limit on the strength of the pattern or, equivalently, on the amount of control that fixing the causal state (or one of its rivals) can exert.

VII. CONCLUDING REMARKS

A. Discussion

Let's review, informally, what we have shown. We began with questions about the nature of patterns and about pattern discovery. Our examination of these issues lead us to want a way of describing patterns that was at once algebraic, computational, intrinsically probabilistic, and causal. We then defined patterns in ensembles, in a very general and abstract sense, as equivalence classes of histories, or sets of hidden states, used for prediction. We defined the strength of such patterns (by their forecasting ability or prescience) and their statistical complexity (by the entropy of the states or the amount of information retained by the process about its history). We showed that there was a limit on how strong such patterns could get for each particular process, given by the predictive ability of the entire past. In this way, we narrowed our goal to finding a predictor of maximum strength and minimum complexity.

Optimal prediction led us to the equivalence relation $\sim_\epsilon$ and the function ϵ and so to representing patterns by causal states and their transitions—the ϵ-machine. Our first theorem showed that the causal states are maximally prescient; our second, that they are the simplest way of representing the pattern of maximum strength; our third theorem, that they are unique in having this double optimality. Further results showed that ϵ-machines are the least stochastic way of capturing maximum-strength patterns and emphasized the need to employ the efficacious but hidden states of the process, rather than just its gross observables, such as sequence blocks.

Why are ϵ-machine states causal? First, ϵ-machine architecture (say, as given by its semi-group algebra) delineates the dependency between the morphs $\mathrm{P}(\vec{S} \mid \overleftarrow{S})$, considered as events in which each new symbol determines the succeeding morph. Thus, if state B follows state A then A is a cause of B and B is an effect of A. Second, ϵ-machine minimality guarantees that there are no other events that intervene to render A and B independent.[18]

The ϵ-machine is thus a causal representation of all the patterns in the process. It is maximally predictive and minimally complex. It is at once computational, since it shows how the process stores information (in the causal states) and transforms that information (in the state-to-state transitions), and algebraic (for details on which see Appendix D). It can be analytically calculated from given distributions and systematically approached from empirical data. It satisfies the basic constraints laid out in Section II F.

These comments suggest that computational mechanics and ϵ-machines are related or may be of interest to a number of fields. Time

series analysis, decision theory, machine learning, and universal coding theory explicitly or implicitly require models of observed processes. The theories of stochastic processes, formal languages and computation, and of measures of physical complexity are all concerned with representations of processes—concerns which also arise in the design of novel forms of computing devices. Very often the motivations of these fields are far removed from computational mechanics. But it is useful, if only by way of contrast, to touch briefly on these areas and highlight one or several connections with computational mechanics, and we do so in Appendix H.

B. Limitations of the Current Results

Let's catalogue the restrictive assumptions we made at the beginning and that were used by our development.

1. We know exact joint probabilities over sequence blocks of all lengths for a process.
2. The observed process takes on discrete values.
3. The process is discrete in time.
4. The process is a pure time series; e.g., without spatial extent.
5. The observed process is stationary.
6. Prediction can only be based on the process's past, not on any outside source of information.

The question arises, Can any be relaxed without much trouble?

One way to lift the first limitation is to develop a statistical error theory for ϵ-machine inference that indicates, say, how much data is required to attain a given level of confidence in an ϵ-machine with a given number of causal states. This program is underway and, given its initial progress, we describe several issues in more detail in the next section.

The second limitation probably can be addressed, but with a corresponding increase in mathematical sophistication. The information-theoretic quantities we have used are also defined for continuous random variables. It is likely that many of the results carry over to the continuous setting.

The third limitation also looks similarly solvable, since continuous-time stochastic process theory is well developed. This may involve sophisticated probability theory or functional analysis.

As for the fourth limitation, there already exist tricks to make spatially extended systems look like time series. Essentially, one looks at all the paths through space-time, treating each one as if it were a time series.

While this works well for data compression,[85] it is not yet clear whether it will be entirely satisfactory for capturing structure.[86] More work needs to be done on this subject.

It is unclear at this time how to relax the assumption of stationarity. One can formally extend most of the results in this paper to nonstationary processes without much trouble. It is, however, unclear how much substantive content these extensions have and, in any case, a systematic classification of nonstationary processes is (at best) in its infant stages.

Finally, one might say that the last restriction is a positive *feature* when it comes to thinking about patterns and the intrinsic structure of a process. "Pattern" is a vague word, of course, but even in ordinary usage it is only supposed to involve things *inside* the process, not the rest of the universe. Given two copies of a document, the contents of one copy can be predicted with an enviable degree of accuracy by looking at the other copy. This tells us that they share a common structure, but says absolutely nothing about what that pattern is, since it is just as true of well-written and tightly-argued scientific papers (which presumably are highly organized) as it is of monkey-at-keyboard pieces of gibberish (which definitely are not).

C. Conclusions and Directions for Future Work

Computational mechanics aims to understand the nature of patterns and pattern discovery. We hope that the foregoing development has convinced the reader that we are neither being rash when we say that we have laid a foundation for those projects, nor that we are being flippant when we say that patterns are what are represented by ϵ-machines, and that we discover them by ϵ-machine reconstruction. We would like to close by marking out two broad avenues for future work.

First, consider the mathematics of ϵ-machines themselves. We have just mentioned possible extensions in the form of lifting assumptions made in this development, but there are many other ways to go. It would be helpful to have a good understanding of the measurement-resolution scaling properties of ϵ-machines for continuous-state processes and of their relation to such ideas in automata theory as the Krohn–Rhodes decomposition.[31] Anyone who manages to absorb Volume II of ref. 27 would probably be in a position to answer interesting questions about the structures that processes preserve, perhaps even to give a purely relation-theoretic account of ϵ-machines. We have alluded in a number of places to the trade-off between prescience and complexity. For a given process there is presumably a sequence of optimal machines connecting the one-state, zero-complexity machine with minimal prescience to the ϵ-machine. Each

member of the path is the maximally prescient machine for a certain level of complexity; it would be very interesting to know what, if anything, we can say in general about the shape of this "prediction frontier".

Second, there is ϵ-machine reconstruction, an activity about which we have said next to nothing. As we mentioned above (p. 24), there are already several algorithms for reconstructing machines from data, even "on-line" ones. It is fairly evident that these algorithms will find the true machine in the limit of infinite time and infinite data. What is needed is an understanding of the *error statistics*[87] of different reconstruction procedures, of the kinds of mistakes these procedures make and the probabilities with which they make them. Ideally, we want to find "confidence regions" for the products of reconstruction. The aim is to calculate (i) the probabilities of different degrees of reconstruction error for a given volume of data, (ii) the amount of data needed to be confident of a fixed bound on the error, or (iii) the rates at which different reconstruction procedures converge on the ϵ-machine. So far, an analytical theory has been developed that predicts the average number of estimated causal states as a function of the amount of data used when reconstructing certain kinds of processes.[88] Once we possess a more complete theory of statistical inference for ϵ-machines, analogous perhaps to what already exists in computational learning theory, we will be in a position to begin analyzing, sensibly and rigorously, the multitude of intriguing patterns and information-processing structures the natural world presents.

APPENDIX A: INFORMATION-THEORETIC FORMULÆ

The following formulæ prove useful in the development. They are relatively intuitive, given our interpretation, and they can all be proved with little more than straight algebra; see ref. 63, ch. 2. Below, f is a nonrandom function.

$$H[X, Y] = H[X] + H[Y|X] \tag{A1}$$

$$H[X, Y] \geqslant H[X] \tag{A2}$$

$$H[X, Y] \leqslant H[X] + H[Y] \tag{A3}$$

$$H[X|Y] \leqslant H[X] \tag{A4}$$

$$H[X|Y] = H[X] \text{ iff } X \text{ is independent of } Y \tag{A5}$$

$$H[X, Y|Z] = H[X|Z] + H[Y|X, Z] \tag{A6}$$

$$H[X, Y|Z] \geqslant H[X|Z] \tag{A7}$$

$$H[X] - H[X|Y] = H[Y] - H[Y|X] \tag{A8}$$

$$I[X;Y] \leqslant H[X] \tag{A9}$$

$$I[X;Y] = H[X] \text{ iff } H[X|Y] = 0 \tag{A10}$$

$$H[f(X)] \leqslant H[X] \tag{A11}$$

$$H[X|Y] = 0 \text{ iff } X = f(Y) \tag{A12}$$

$$H[f(X)|Y] \leqslant H[X|Y] \tag{A13}$$

$$H[X|f(Y)] \geqslant H[X|Y] \tag{A14}$$

Eqs. (A1) and (A6) are called the *chain rules* for entropies. Strictly speaking, the right hand side of Eq. (A12) should read "for each y, $\mathrm{P}(X = x|Y = y) > 0$ for exactly one x".

APPENDIX B: THE EQUIVALENCE RELATION THAT INDUCES CAUSAL STATES

Any relation that is reflexive, symmetric, and transitive is an *equivalence relation*.

Consider the set $\overleftarrow{\mathbf{S}}$ of all past sequences, of any length:

$$\overleftarrow{\mathbf{S}} = \{\overleftarrow{s}^L = s_{L-1}\cdots s_{-1} : s_i \in \mathcal{A}, L \in \mathbb{Z}^+\}. \tag{B1}$$

Recall that $\overleftarrow{s}^0 = \lambda$, the empty string. We define the relation $\sim_\epsilon$ over $\overleftarrow{\mathbf{S}}$ by

$$\overleftarrow{s}_i^K \sim_\epsilon \overleftarrow{s}_j^L \Leftrightarrow \mathrm{P}(\vec{S}|\overleftarrow{s}_i^K) = \mathrm{P}(\vec{S}|\overleftarrow{s}_j^L), \tag{B2}$$

for all semi-infinite $\vec{S} = s_0 s_1 s_2 \cdots$, where $K, L \in \mathbb{Z}^+$. Here we show that $\sim_\epsilon$ is an equivalence relation by reviewing the basic properties of relations, equivalence classes, and partitions. (The proof details are straightforward and are not included. See ref. 89.) We will drop the length variables K and L and denote by $\overleftarrow{s}, \overleftarrow{s}', \overleftarrow{s}'' \in \overleftarrow{\mathbf{S}}$ members of any length in the set $\overleftarrow{\mathbf{S}}$ of Eq. (B1).

First, $\sim_\epsilon$ is a *relation* on $\overleftarrow{\mathbf{S}}$ since we can represent it as a subset of the Cartesian product

$$\overleftarrow{\mathbf{S}} \times \overleftarrow{\mathbf{S}} = \{(\overleftarrow{s}, \overleftarrow{s}') : \overleftarrow{s}, \overleftarrow{s}' \in \overleftarrow{\mathbf{S}}\}. \tag{B3}$$

Second, the relation $\sim_\epsilon$ is an *equivalence relation* on $\overleftarrow{\mathbf{S}}$ since it is

1. reflexive: $\overleftarrow{s} \sim_\epsilon \overleftarrow{s}$, for all $\overleftarrow{s} \in \overleftarrow{\mathbf{S}}$;

2. symmetric: $\overleftarrow{s} \sim_\epsilon \overleftarrow{s}' \Rightarrow \overleftarrow{s}' \sim_\epsilon \overleftarrow{s}$; and
3. transitive: $\overleftarrow{s} \sim_\epsilon \overleftarrow{s}'$ and $\overleftarrow{s}' \sim_\epsilon \overleftarrow{s}'' \Rightarrow \overleftarrow{s} \sim_\epsilon \overleftarrow{s}''$.

Third, if $\overleftarrow{s} \in \overleftarrow{\mathbf{S}}$, the *equivalence class* of $\overleftarrow{s}$ is

$$[\overleftarrow{s}] = \{\overleftarrow{s}' \in \overleftarrow{\mathbf{S}} : \overleftarrow{s}' \sim_\epsilon \overleftarrow{s}\}. \tag{B4}$$

The set of all equivalence classes in $\overleftarrow{\mathbf{S}}$ is denoted $\overleftarrow{\mathbf{S}}/\sim_\epsilon$ and is called the *factor set* of $\overleftarrow{\mathbf{S}}$ with respect to $\sim_\epsilon$. In Section IV A we called the individual equivalence classes *causal states* $\mathcal{S}_i$ and denoted the set of causal states $\boldsymbol{\mathcal{S}} = \{\mathcal{S}_i : i = 0, 1, \ldots, k-1\}$. That is, $\boldsymbol{\mathcal{S}} = \overleftarrow{\mathbf{S}}/\sim_\epsilon$. (We noted in the main development that the cardinality $k = |\boldsymbol{\mathcal{S}}|$ of causal states may or may not be finite.)

Finally, we list several basic properties of the causal-state equivalence classes.

1. $\bigcup_{\overleftarrow{s} \in \overleftarrow{\mathbf{S}}} [\overleftarrow{s}] = \overleftarrow{\mathbf{S}}$.
2. $\bigcup_{i=0}^{k-1} \mathcal{S}_i = \overleftarrow{\mathbf{S}}$.
3. $[\overleftarrow{s}] = [\overleftarrow{s}'] \Leftrightarrow \overleftarrow{s} \sim_\epsilon \overleftarrow{s}'$.
4. If $\overleftarrow{s}, \overleftarrow{s}' \in \overleftarrow{\mathbf{S}}$, either
   (a) $[\overleftarrow{s}] \cap [\overleftarrow{s}'] = \varnothing$ or
   (b) $[\overleftarrow{s}] = [\overleftarrow{s}']$.
5. The causal states $\boldsymbol{\mathcal{S}}$ are a *partition* of $\overleftarrow{\mathbf{S}}$. That is,
   (a) $\mathcal{S}_i \neq \varnothing$ for each i,
   (b) $\bigcup_{i=0}^{k-1} \mathcal{S}_i = \overleftarrow{\mathbf{S}}$, and
   (c) $\mathcal{S}_i \cap \mathcal{S}_j = \varnothing$ for all $i \neq j$.

We denote the start state with $\mathcal{S}_0$. The start state is the causal state associated with $\overleftarrow{s} = \lambda$. That is, $\mathcal{S}_0 = [\lambda]$.

APPENDIX C: TIME REVERSAL

The definitions and properties of the causal states obtained by scanning sequences in the opposite direction, i.e., the causal states $\overleftarrow{\mathbf{S}}/\sim_\epsilon$, follow similarly to those derived just above in Appendix B. In general, $\overleftarrow{\mathbf{S}}/\sim_\epsilon \neq \overrightarrow{\mathbf{S}}/\sim_\epsilon$. That is, *past* causal states are not necessarily the same as *future* causal states; past and future morphs can differ; unlike entropy rate,[16] past and future statistical complexities need not be equal:

$\overleftarrow{C}_\mu \neq \overrightarrow{C}_\mu$;[18] and so on. The presence or lack of this type of time-reversal symmetry, as reflected in these inequalities, is a fundamental property of a process.

APPENDIX D: ε-MACHINES ARE MONOIDS

A *semi-group* is a set of elements closed under an associative binary operator, but without a guarantee that every, or indeed any, element has an inverse.[90] A *monoid* is a semi-group with an identity element. Thus, semi-groups and monoids are generalizations of groups. Just as the algebraic structure of a group is generally interpreted as a symmetry, we propose to interpret the algebraic structure of a semi-group as a *generalized* symmetry. The distinction between monoids and other semi-groups becomes important here: only semi-groups with an identity element—i.e., monoids—can contain subsets that are groups and so represent conventional symmetries.

We claim that the transformations that concatenate strings of symbols from $\mathcal{A}$ onto other such strings form a semi-group G, the generators of which are the transformations that concatenate the elements of $\mathcal{A}$. The identity element is to be provided by concatenating the null symbol λ. The concatenation of string t onto the string s is forbidden if and only if strings of the form st have probability zero in a process. All such concatenations are to be realized by a single semi-group element denoted $\varnothing$. Since if $\mathrm{P}(st) = 0$, then $\mathrm{P}(stu) = \mathrm{P}(ust) = 0$ for any string u, we require that $\varnothing g = g\varnothing = \varnothing$ for all $g \in G$. Can we provide a representation of this semi-group?

Recall that, from our definition of the labeled transition probabilities, $T_{ij}^{(\lambda)} = \delta_{ij}$. Thus, $\mathbf{T}^{(\lambda)}$ is an identity element. This suggests using the labeled transition matrices to form a matrix representation of the semi-group. Accordingly, first define $U_{ij}^{(s)}$ by setting $U_{ij}^{(s)} = 0$ when $T_{ij}^{(s)} = 0$ and $U_{ij}^{(s)} = 1$ otherwise, to remove probabilities. Then define the set of matrices $\mathbf{U} = \{\mathbf{T}^{(\lambda)}\} \cup \{\mathbf{U}^{(\mathbf{s})}, \mathbf{s} \in \mathcal{A}\}$. Finally, define G as the set of all matrices generated from the set $\mathbf{U}$ by recursive multiplication. That is, an element g of G is

$$g^{(ab\ldots cd)} = \mathbf{U}^{(d)}\mathbf{U}^{(c)}\ldots\mathbf{U}^{(b)}\mathbf{U}^{(a)}, \tag{D1}$$

where $a, b, \ldots c, d \in \mathcal{A}$. Clearly, G constitutes a semi-group under matrix multiplication. Moreover, $g^{(a\ldots bc)} = \mathbf{0}$ (the all-zero matrix) if and only if, having emitted the symbols $a\ldots b$ in order, we must arrive in a state from which it is impossible to emit the symbol c. That is, the zero-matrix $\mathbf{0}$ is generated if and only if the concatenation of c onto $a\ldots b$ is forbidden. The element $\varnothing$ is thus the all-zero matrix $\mathbf{0}$, which clearly satisfies the necessary constraints. This completes the proof of Proposition 1.

We call the matrix representation—Eq. (D1) taken over all words in $\mathcal{A}^k$—of G the *semi-group machine* of the ϵ-machine $\{\mathcal{S}, \mathbf{T}\}$. See ref. 91.

APPENDIX E: MEASURE-THEORETIC TREATMENT OF CAUSAL STATES

In Section IV, where we define causal states, ϵ-machines, and their basic properties, we use a great many conditional probabilities. However, there are times when the events on which we condition—particular histories, or particular effective states—have probability zero. Then classical formulæ for conditional probability do not apply, and a more careful and technical treatment, going back to the measure-theoretic basis of probability, is called for. We provide such a treatment in this appendix, showing that the concepts we introduced in Section IV—the causal states, their morphs, and so forth—are well defined measure-theoretically. Our proofs in that section are equally valid whether we interpret the conditional probabilities they invoke classically or measure-theoretically. (The measure-theoretic interpretation raises a few technicalities, which we have flagged with footnotes to those proofs.) And we show here that our methods of proof in subsequent sections are not affected by this change in interpretation.

In what follows, we draw on refs. 61, 62, 92–95. Our notation broadly follows that of ref. 92. A slightly different approach to these issues, and more than slightly different terminology and notation, may be found in chapter 2 of ref. 10.

1. Abstract Definition of Conditional Probability

Definition 14 (Conditional Probability). Consider a probability space $(\Omega, \mathcal{F}, P)$ and a σ-subalgebra $\mathcal{G} \subset \mathcal{F}$. The *conditional probability* of an event $A \in \mathcal{F}$, given the family of events $\mathcal{G}$, is a real-valued random function $P_{A\|\mathcal{G}}(\omega)$, with the following properties:

1. $P_{A\|\mathcal{G}}(\omega)$ is measurable with respect to $\mathcal{G}$; and
2. for any $G \in \mathcal{G}$,

$$\int_G P_{A\|\mathcal{G}}(\omega)\, dP = P(A \cap G) \tag{E1}$$

The latter condition generalizes the classical formula that $P(A \cap G) = \sum_{g \in G} P(A|g)\, P(g)$.

Proposition 2. There always exists a function $P_{A\|\mathcal{G}}(\omega)$ satisfying the just-given conditions. Moreover, if f and g are two functions which both satisfy the above requirements, $f(\omega) = g(\omega)$ for P-almost-all ω.

Proof. The existence of such random variables is vouchsafed to us by the Radon–Nikodym theorem; $P_{A\|\mathcal{G}}(\omega)$ is the Radon–Nikodym derivative of $P(A \cap G)$, which is a measure over $\mathcal{G}$, with respect to P. (The latter is also restricted to the σ-subalgebra $\mathcal{G}$.) The Radon–Nikodym theorem also tells us that any two functions which satisfy the two conditions above agree for P-almost-all points ω. Any such function is called a *version* of the conditional probability. (See any of refs. 61, 92–95 for further details.)

If $\mathcal{G} = \boldsymbol{\mu}(X)$, the σ-algebra generated by the random variable X, then we may write $P_{A\|X=x}(\omega)$ or $P_{A\|X}(\omega)$ in place of $P_{A\|\mathcal{G}}(\omega)$.

It is not always the case that, if we let A vary, while holding ω fixed, we get a proper probability measure. Indeed, there are pathological examples where there are no conditional probability measures, though there are conditional probability functions. A conditional probability function which is a measure for all ω is said to be *regular*. If a regular conditional probability uses as its conditioning σ-algebra that generated by a random variable X, we write $P(\cdot \,|\, X = x)$, as usual.

a. Conditional Expectation

As well as conditional probabilities, we shall need conditional expectations. Their definition is completely analogous to Definition 14. The expectation of the random variable X conditional on the σ-subalgebra $\mathcal{G}$, denoted $\mathbf{E}\{X \,\|\, \mathcal{G}\}$ is an integrable, $\mathcal{G}$-measurable random variable such that $\int_G \mathbf{E}\{X \,\|\, \mathcal{G}\}\, dP = \int_G X dP$ for all $G \in \mathcal{G}$. Conditional probabilities are, of course, the conditional expectations of indicator functions. There is another important relationship between conditional probability and conditional expectation, which we give in the form of another proposition.

Proposition 3 (Coarsening Conditional Probability)[92–95]**.** Consider any two σ-subalgebras $\mathcal{G}$ and $\mathcal{H}$, with $\mathcal{G} \subset \mathcal{H}$. Then

$$\mathbf{P}_{F\|\mathcal{G}}(\omega) = \mathbf{E}\{\mathbf{P}_{F\|\mathcal{H}} \,\|\, \mathcal{G}\}(\omega) \text{ almost surely (a.s.)}, \tag{E2}$$

where we have been explicit about the conditional expectation's dependence on ω.

b. Conditional Independence

Let $\mathcal{G}$ be the conditioning σ-subalgebra, and let $\mathcal{A}$ and $\mathcal{B}$ be two other σ-subalgebras. Then $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent, given

$\mathcal{G}$, just when, for any pair of events A, B, $A \in \mathcal{A}$ and $B \in \mathcal{B}$, $P_{AB\|\mathcal{G}}(\omega) = P_{A\|\mathcal{G}}(\omega)\, P_{B\|\mathcal{G}}(\omega)$ a.s.

Take any two σ-algebras over the same set, $\mathcal{A}$ and $\mathcal{B}$; their product, $\mathcal{AB}$, is the σ-algebra generated by the sets of the form $a \cap b$, where $a \in \mathcal{A}$ and $b \in \mathcal{B}$.

Proposition 4 (Ref. 95, Section 2.5). $\mathcal{A}$ and $\mathcal{B}$ are conditionally independent given $\mathcal{G}$ iff, for all $B \in \mathcal{B}$, $P_{B\|\mathcal{A}G}(\omega) = P_{B\|\mathcal{G}}(\omega)$ a.e., where $\mathcal{A}G$ is defined as above. This is also true if $\mathcal{A}$ and $\mathcal{B}$ are interchanged.

Remark. Assuming regularity of conditional probability, this is equivalent to saying that the random variables Y and Z are independent given X if and only if

$$P(Z \in A \,|\, X = x, Y = y) = P(Z \in A \,|\, X = x) \tag{E3}$$

Proposition 5 (ref. 94, p. 351). Assuming regularity of conditional probability, for any three random variables

$$\begin{aligned} &P(Z \in A, Y = y \,|\, X = x) \\ &\quad = P(Z \in A \,|\, Y = y, X = x)\, P(Y = y \,|\, X = x) \end{aligned} \tag{E4}$$

Let $\mathcal{A} = \mathbf{\mu}(X)$, and $\mathcal{B} = \mathbf{\mu}(f(X))$, for a measurable, nonrandom function f. Then $\mathcal{A}B = \mathbf{\mu}(X, f(X)) = \mathcal{A} = \mathbf{\mu}(X)$. Therefore,

$$P_{A\|X, f(X)}(\omega) = P_{A\|X}(\omega) \text{ a.e.}, \tag{E5}$$

since the conditioning σ-algebras are the same.

2. Restatements and Proofs of the Main Results

We begin by restating the definition of causal equivalence, and so of causal states, in terms adapted to abstract conditional probabilities. We then go through the results of Section IV in order and, where necessary, give alternate proofs of them. (Where new proofs are not needed, we say so.)

a. Definition of Causal States

For us, Ω is the space of two-sided infinite strings over $\mathcal{A}$; $\mathcal{F}$ is the σ-algebra generated by cylinders over such strings; and the probability measure P is simply P (Definition 1).

What we want to do is condition on histories; so we make our conditioning σ-subalgebra $\boldsymbol{\mu}(\overleftarrow{S})$, following the usual convention that $\boldsymbol{\mu}(X)$ is the σ-algebra induced by the random variable X. This contains all finite-length histories, and even all semi-infinite histories, as events. Similarly, designate the σ-subalgebra for futures by $\boldsymbol{\mu}(\overrightarrow{S})$. We want there to be a function $\mathrm{P}_{F \| \boldsymbol{\mu}(\overleftarrow{s})}(\omega)$, at least when $F \in \boldsymbol{\mu}(\overrightarrow{S})$; and we want this to be a probability measure over $\boldsymbol{\mu}(\overrightarrow{S})$, for fixed ω.

As we have seen (Proposition 2), the conditional probability function exists. Moreover, it is regular, since $\boldsymbol{\mu}(\overleftarrow{S})$ is a subalgebra of the σ-algebra of cylinder sets, and S_t always takes its values from a fixed, finite set.[93, 95]

Thus, we *do* have a random variable $\mathrm{P}_{F \| \overleftarrow{S} = \overleftarrow{s}}(\omega)$, which is the probability of the set $F \in \boldsymbol{\mu}(\overrightarrow{S})$, given that $\overleftarrow{S} = \overleftarrow{s}$. We now define causal equivalence thus: $\overleftarrow{s} \sim_\epsilon \overleftarrow{s}'$ iff, for P-almost all pairs ω, ω', if $\omega \in \overleftarrow{s}$ and $\omega' \in \overleftarrow{s}'$, then $\mathrm{P}_{F \| \overleftarrow{S} = \overleftarrow{s}}(\omega) = \mathrm{P}_{F \| \overleftarrow{S} = \overleftarrow{s}'}(\omega')$, for all $F \in \boldsymbol{\mu}(\overrightarrow{S})$. (It is clear that this is an equivalence relation—in particular, that it is transitive.)

It may be comforting to point out (following ref. 10, Section 2.5) that the functions $\mathrm{P}_{F \| \boldsymbol{\mu}(\overleftarrow{S}^L)}(\omega)$, i.e., the probabilities of the fixed future event F conditional on longer and longer histories, almost always converge on $\mathrm{P}_{F \| \boldsymbol{\mu}(\overleftarrow{S})}(\omega)$. This is because of the martingale convergence theorem of Doob (ref. 93, Theorem VII.4.3). For each L, $\boldsymbol{\mu}(\overleftarrow{S}^L) \subset \boldsymbol{\mu}(\overleftarrow{S}^{L+1})$ and the smallest σ-algebra containing them all is $\boldsymbol{\mu}(\overleftarrow{S})$. Thus, for any random variable X with $\mathbf{E}\{|X|\} < \infty$, $\lim_{L \to \infty} \mathbf{E}\{X \| \boldsymbol{\mu}(\overleftarrow{S}^L)\} = \mathbf{E}\{X \| \boldsymbol{\mu}(\overleftarrow{S})\}$ almost surely. Applied to the indicator function 1_F of the future event F, this gives the desired convergence.

Note that if we want only causal equivalence for a finite future, matters are even simpler. Since for finite L every event in $\boldsymbol{\mu}(\overrightarrow{S}^L)$ consists of the union of a finite number of disjoint elementary events (i.e., of a finite number of length-L futures), it suffices if the conditional probability assignments agree for the individual futures. If they agree for every finite L, then we have the alternate definition (Eq. 17) of causal states.

b. Measurability of ϵ

At several points, we need ϵ to be a measurable function, i.e., we need $\boldsymbol{\mu}(\mathcal{S}) \subseteq \boldsymbol{\mu}(\overleftarrow{S})$. This is certainly the case for processes that can be represented as Markov chains, stochastic deterministic finite automata, or conventional hidden Markov models generally. The strongest general result yet obtained is that ϵ is, so to speak, *nearly measurable*.

Proposition 6 (Ref. 10, Prop. 2.5.3). For each causal state $\mathcal{S}_i$, the set $\epsilon^{-1}(\mathcal{S}_i)$ of histories mapping to $\mathcal{S}_i$ is either measurable or the intersection of a measurable set and a set of full measure.

Thus, each $\epsilon^{-1}(\mathcal{S}_i)$ differs from a measurable set in $\boldsymbol{\mu}(\overleftarrow{S})$ by at most a subset of a set of measure zero. This is close enough to complete measurability for our purposes, and we will speak of ϵ *as though* it were always measurable. Finding necessary and sufficient conditions on the process for ϵ to be measurable is an interesting problem.

c. The Morph

We wish to show that the morph of a causal state is well defined, i.e., that the distribution of futures conditional on the entire history is the same as the distribution conditional on the causal state. Start with the fact that, since $\mathcal{S} = \epsilon(\overleftarrow{S})$, and ϵ is nearly measurable, $\boldsymbol{\mu}(\mathcal{S}) \subseteq \boldsymbol{\mu}(\overleftarrow{S})$. This lets us use Proposition E2, and see that $\mathrm{P}_{F \| \mathcal{S} = \mathcal{S}_i}(\omega)$ is the expectation of $\mathrm{P}_{F \| \overleftarrow{S} = \overleftarrow{s}}(\omega)$ over those $\omega \in \mathcal{S}_i$. But, by the construction of causal states, $\mathrm{P}_{F \| \overleftarrow{S} = \overleftarrow{s}}(\omega)$ has the same value for P-almost-all ω. Thus $\mathrm{P}(F \,|\, \mathcal{S} = \mathcal{S}_i) = \mathrm{P}(F \,|\, \overleftarrow{S} = \overleftarrow{s})$ for (almost every) $\overleftarrow{s} \in \mathcal{S}_i$. (We can always find versions of the conditional probabilities which eliminate the "almost-all" and the "almost every" above.) So, since this works for arbitrary future events F, it works in general, and we may say that the distribution of futures is the same whether we condition on the past or on the causal state.

d. Existence of the Conditional Entropy of Futures

As we have seen, $\mathrm{P}_{\overrightarrow{S}^L \| \overleftarrow{S}}(\omega)$ is a probability measure over a finite set, so (ref. 62, Section 5.5), we define the entropy of length-L futures conditional on a particular history $\overleftarrow{s}$ as

$$\begin{aligned} H[\overrightarrow{S}^L \,|\, \overleftarrow{S} = \overleftarrow{s}] & \\ \equiv -\sum_{\{s^L\}} \mathrm{P}(\overrightarrow{S}^L = s^L \,|\, \overleftarrow{S} = \overleftarrow{s}) \log_2 \mathrm{P}(\overrightarrow{S}^L = s^L \,|\, \overleftarrow{S} = \overleftarrow{s}), & \end{aligned} \tag{E6}$$

with the understanding that we omit futures of conditional probability zero from the sum. This is measurable, since $\mathrm{P}(\overrightarrow{S}^L = s^L \,|\, \overleftarrow{S} = \overleftarrow{s})$ is $\boldsymbol{\mu}(\overleftarrow{S})$-measurable for each s^L. Now set

$$H[\overrightarrow{S}^L \,|\, \overleftarrow{S}] \equiv \int H[\overrightarrow{S}^L \,|\, \overleftarrow{S} = \overleftarrow{s}] \, d\mathrm{P}_{\overleftarrow{S}}, \tag{E7}$$

where $\mathrm{P}_{\overleftarrow{S}}$ is the restriction of P to $\boldsymbol{\mu}(\overleftarrow{S})$. (Measurability tells us that the integral exists.)

The procedure for $H[\overrightarrow{S}^L \,|\, \mathcal{R}]$ is similar, but if anything even less problematic.

Note that we do not need to re-do the derivations of Sections V and VI, since those simply exploit standard inequalities of information theory,

which certainly apply to the conditional entropies we have just defined. (Cf. refs. 61 and 62.)

c. The Labeled Transition Probabilities

Recall that we defined the labeled transition probability $T_{ij}^{(s)}$ as the probability of the joint event $\mathcal{S}' = \mathcal{S}_j$ and $\vec{S}^1 = s$, conditional on $\mathcal{S} = \mathcal{S}_i$. Clearly (Proposition 2), the existence of such conditional probabilities is not at issue, nor, as we have seen, is their regularity. We can thus leave Definition 8 alone.

APPENDIX F: ALTERNATE PROOF OF THE REFINEMENT LEMMA

The proof of Lemma 7 carries through verbally, but we do not wish to leave loop-holes. Unfortunately, this means introducing two new bits of mathematics.

First of all, we need the largest classes that are strictly homogeneous (Definition 6) with respect to $\vec{S}^L$ for fixed L; these are, so to speak, truncations of the causal states. Accordingly, we will talk about $\mathcal{S}^L$ and σ^L, which are analogous to $\mathcal{S}$ and σ. We will also need to define the function $\phi_{\sigma\rho}^L \equiv \mathrm{P}(\mathcal{S}^L = \sigma^L | \mathcal{R} = \rho)$.

Putting these together, for every L we have

$$H[\vec{S}^L | \mathcal{R} = \rho] = H\left[\sum_{\{\sigma^L\}} \phi_{\sigma\rho}^L \mathrm{P}(\vec{S}^L | \mathcal{S}^L = \sigma^L) \right] \tag{F1}$$

$$\geqslant \sum_{\{\sigma^L\}} \phi_{\sigma\rho}^L H[\vec{S}^L | \mathcal{S}^L = \sigma^L]. \tag{F2}$$

Thus,

$$H[\vec{S}^L | \mathcal{R}] = \sum_{\{\rho\}} \mathrm{P}(\mathcal{R} = \rho)\, H[\vec{S}^L | \mathcal{R} = \rho] \tag{F3}$$

$$\geqslant \sum_{\{\rho\}} \mathrm{P}(\mathcal{R} = \rho) \sum_{\{\sigma^L\}} \phi_{\sigma\rho}^L H[\vec{S}^L | \mathcal{S}^L = \sigma^L] \tag{F4}$$

$$= \sum_{\{\sigma^L, \rho\}} \mathrm{P}(\mathcal{R} = \rho)\, \phi_{\sigma\rho}^L H[\vec{S}^L | \mathcal{S}^L = \sigma^L] \tag{F5}$$

$$= \sum_{\{\sigma^L, \rho\}} \mathrm{P}(\mathcal{S}^L = \sigma^L, \mathcal{R} = \rho)\, H[\vec{S}^L | \mathcal{S}^L = \sigma^L] \tag{F6}$$

$$= \sum_{\{\sigma^L\}} \mathrm{P}(\mathcal{S}^L = \sigma^L)\, H[\vec{S} | \mathcal{S}^L = \sigma^L] \tag{F7}$$

$$= H[\vec{S}^L | \mathcal{S}^L]. \tag{F8}$$

That is to say,

$$H[\vec{S}^L|\mathcal{R}] \geqslant H[\vec{S}^L|\mathcal{S}^L], \tag{F9}$$

with equality if and only if every $\phi^L_{\sigma\rho}$ is either 0 or 1. Thus, if $H[\vec{S}^L|\hat{\mathcal{R}}] = H[\vec{S}^L|\mathcal{S}^L]$, every $\hat{\rho}$ is entirely contained within some σ^L; except for possible subsets of measure 0. But if this is true for every L—which, in the case of a prescient rival $\hat{\mathcal{R}}$, it is—then every $\hat{\rho}$ is at least weakly homogeneous (Definition 7) with respect to all $\vec{S}^L$. Thus, by Lemma 3, all its members, except for that same subset of measure 0, belong to the same causal state. QED.

APPENDIX G: FINITE ENTROPY FOR THE SEMI-INFINITE FUTURE

While cases where $H[\vec{S}]$ is finite—more exactly, where $\lim_{L\to\infty} H[\vec{S}^L]$ exists and is finite—may be uninteresting for information-theorists, they are of great interest to physicists, since they correspond, among other things, to periodic and limit-cycle behaviors. There are, however, only two substantial differences between what is true of the infinite-entropy processes considered in the main body of the development and the finite-entropy case.

First, we can simply replace statements of the form "for all L, $H[\vec{S}^L]$ … " with $H[\vec{S}]$. For example, the optimal prediction theorem (Theorem 1) for finite-entropy processes becomes for all $\mathcal{R}$, $H[\vec{S}|\mathcal{R}] \geqslant H[\vec{S}|\mathcal{S}]$. The details of the proofs are, however, entirely analogous.

Second, we can prove a substantially stronger version of the Control Theorem (Theorem 6).

Theorem 7 (The Finite-Control Theorem). For all prescient rivals $\hat{\mathcal{R}}$,

$$H[\vec{S}] - H[\vec{S}|\hat{\mathcal{R}}] \leqslant C_\mu. \tag{G1}$$

Proof. By a direct application of Eq. (A9) and the definition of mutual information Eq. (8), we have that

$$H[\vec{S}] - H[\vec{S}|\mathcal{S}] \leqslant H[\mathcal{S}]. \tag{G2}$$

But, by the definition of prescient rivals (Definition 24), $H[\vec{S}|\mathcal{S}] = H[\vec{S}|\hat{\mathcal{R}}]$, and, by definition, $C_\mu = H[\mathcal{S}]$. Substituting equals for equals gives us the theorem. QED.

APPENDIX H: RELATIONS TO OTHER FIELDS

1. Time Series Modeling

The goal of time series modeling is to predict the future of a measurement series on the basis of its past. Broadly speaking, this can be divided into two parts: identify equivalent pasts and then produce a prediction for each class of equivalent pasts. That is, we first pick a function $\eta: \overleftarrow{\mathbf{S}} \mapsto \mathcal{R}$ and then pick another function $p: \mathcal{R} \mapsto \overrightarrow{\mathbf{S}}$. Of course, we can choose for the range of p futures of some finite length (length 1 is popular) or even choose distributions over these. While practical applications often demand a single definite prediction—"You will meet a tall dark stranger", there are obvious advantages to predicting a distribution—"You have a .95 chance of meeting a tall dark stranger and a .05 chance of meeting a tall familiar albino." Clearly, the best choice for p is the actual conditional distribution of futures for each $\rho \in \mathcal{R}$. Given this, the question becomes what the best $\mathcal{R}$ is; i.e., What is the best η? At least in the case of trying to understand the whole of the underlying process, we have shown that the best η is, unambiguously, ϵ. Thus, our discussion has implicitly subsumed that of traditional time series modeling.

Computational mechanics—in its focus on letting the process speak for itself through (possibly impoverished) measurements—follows the spirit that motivated one approach to experimentally testing dynamical systems theory. Specifically, it follows in spirit the methods of reconstructing "geometry from a time series" introduced by refs. 96 and 97. A closer parallel is found, however, in later work on estimating minimal equations of motion from data series.[98]

2. Decision-Theoretic Problems

The classic focus of decision theory is "rules of inductive behavior".[99–101] The problem is to chose functions from observed data to courses of action that possess desirable properties. This task has obvious affinities to considering the properties of ϵ and its rivals η. We can go further and say that what we have done *is* consider a decision problem, in which the available actions consist of predictions about the future of the process. The calculation of the optimum rule of behavior in general faces formidable technicalities, such as providing an estimate of the utility of every different course of action under every different hypothesis about the relevant aspects of the world. On the one hand, it is not hard to concoct time-series tasks where the optimal rule of behavior does not use ϵ at all. On the other hand, if we simply aim to predict the process indefinitely far

into the future, then because the causal states are minimal sufficient statistics for the distribution of futures (Theorem 2 (Eq. (29), Remark 4), the optimal rule of behavior will use ϵ.[100]

3. Stochastic Processes

Clearly, the computational mechanics approach to patterns and pattern discovery involves stochastic processes in an intimate and inextricable way. Probabilists have, of course, long been interested in using information-theoretic tools to analyze stochastic processes, particularly their ergodic behavior.[61, 62, 102, 103] There has also been considerable work in the hidden-Markov-model and optimal-prediction literatures on inferring models of processes from data or from given distributions.[10, 34, 104–106] To the best of our knowledge, however, these two approaches have not been previously combined.

Perhaps the closest approach to the spirit of computational mechanics in the stochastic process literature is, surprisingly, the now-classical theory of optimal prediction and filtering for stationary processes, developed by Wiener and Kolmogorov.[107–111] The two theories share the use of information-theoretic notions, the unification of prediction and structure, and the conviction that "the statistical mechanics of time series" is a "field in which conditions are very remote from those of the statistical mechanics of heat engines and which is thus very well suited to serve as a model of what happens in the living organism" (ref. 111, p. 59). So far as we have been able to learn, however, no one has ever used this theory to explicitly identify causal states and causal structure, leaving these implicit in the mathematical form of the prediction and filtering operators. Moreover, the Wiener-Kolmogorov framework forces us to sharply separate the linear and nonlinear aspects of prediction and filtering, because it has a great deal of trouble calculating nonlinear operators.[109, 110] Computational mechanics is completely indifferent to this issue, since it packs *all* of the process's structure into the ϵ-machine, which is equally calculable in linear or strongly nonlinear situations.

4. Formal Language Theory and Grammatical Inference

A formal language is a set of symbol strings ("words" or "allowed words") drawn from a finite alphabet. Every formal language may be described either by a set of rules (a "grammar") for creating all and only the allowed words, by an abstract automaton which also generates the allowed words, or by an automaton which accepts the allowed words and rejects all "forbidden" words. Our ϵ-machines, stripped of probabilities,

correspond to such automata—generative in the simple case or classificatory, if we add a reject state and move to it when none of the allowed symbols are encountered.

Since Chomsky,[112, 113] it has been known that formal languages can be classified into a hierarchy, the higher levels of which have strictly greater expressive power. The hierarchy is defined by restricting the form of the grammatical rules or, equivalently, by limiting the amount and kind of memory available to the automata. The lowest level of the hierarchy is that of regular languages, which may be familiar to Unix-using readers via regular expressions. These correspond to finite-state machines, for which relatives of our minimality and uniqueness theorems are well known,[67] and the construction of causal states is analogous to "Nerode equivalence classing".[67, 114] Our theorems, however, are *not* restricted to this low-memory, nonstochastic setting; for instance, they apply to hidden Markov models with both finite and infinite numbers of hidden states.[10]

The problem of learning a language from observational data has been extensively studied by linguists and by computer scientists interested in natural-language processing. Unfortunately, well developed learning techniques exist only for the two lowest classes in the Chomsky hierarchy, the regular and the context-free languages. (For a good account of these procedures see ref. 115.) Adapting and extending this work to the reconstruction of ϵ-machines should form a useful area of future research, a point to which we alluded in the concluding remarks.

5. Computational and Statistical Learning Theory

The goal of computational learning theory[116, 117] is to identify algorithms that quickly, reliably, and simply lead to good representations of a target "concept". The latter is typically defined to be a binary dichotomy of a certain feature or input space. Particular attention is paid to results about "probably approximately correct" (PAC) procedures:[118] those having a high probability of finding members of a fixed "representation class" (e.g., neural nets, Boolean functions in disjunctive normal form, or deterministic finite automata). The key word here is "fixed"; as in contemporary time-series analysis, practitioners of this discipline acknowledge the importance of getting the representation class right. (Getting it wrong can make easy problems intractable.) In practice, however, they simply take the representation class as a given, even assuming that we can always count on it having at least one representation which *exactly* captures the target concept. Although this is in line with implicit assumptions in most of mathematical statistics, it seems dubious when analyzing learning in the real world.[5, 119, 120]

In any case, the preceding development made no such assumption. One of the goals of computational mechanics is, exactly, *discovering* the best representation. This is not to say that the results of computational learning theory are not remarkably useful and elegant, nor that one should not take every possible advantage of them in implementing ϵ-machine reconstruction. In our view, though, these theories belong more to statistical inference, particularly to algorithmic parameter estimation, than to foundational questions about the nature of pattern and the dynamics of learning.

Finally, in a sense computational mechanics' focus on causal states is a search for a particular kind of structural decomposition for a process. That decomposition is most directly reflected in the conditional independence of past and future that causal states induce. This decomposition reminds one of the important role that conditional independence plays in contemporary methods for artificial intelligence, both for developing systems that reason in fluctuating environments[121] and the more recently developed algorithmic methods of graphical models.[122, 123]

6. Description-Length Principles and Universal Coding Theory

Rissanen's *minimum description length* (MDL) principle, most fully described in ref. 48, is a procedure for selecting the most concise generative model out of a family of models that are all statistically consistent with given data. The MDL approach starts from Shannon's results on the connection between probability distributions and codes. Rissanen's development follows the inductive framework introduced by Solomonoff.[45]

Suppose we choose a representation that leads to a class $\mathcal{M}$ of models and are given data set X. The MDL principle enjoins us to pick the model $\mathrm{M} \in \mathcal{M}$ that minimizes the sum of the length of the description of X given M, plus the length of description of M given $\mathcal{M}$. The description length of X is taken to be $-\log \mathrm{P}(X \mid \mathrm{M})$; cf. Eq. (5). The description length of M may be regarded as either given by some coding scheme or, equivalently, by some distribution over the members of $\mathcal{M}$. (Despite the similarities to model estimation in a Bayesian framework,[124] Rissanen does not interpret this distribution as a Bayesian prior or regard description length as a measure of evidential support.)

The construction of causal states is somewhat similar to the states estimated in Rissanen's *context* algorithm[48, 125, 126] (and to the "vocabularies" built by universal coding schemes, such as the popular Lempel–Ziv algorithm[127, 128]). Despite the similarities, there are significant differences. For a random source—for which there is a single causal state—the context algorithm estimates a number of states that diverges (at least

logarithmically) with the length of the data stream, rather than inferring a single state, as ϵ-machine reconstruction would. Moreover, we avoid any reference to encodings of rival models or to prior distributions over them; $C_\mu(\mathcal{R})$ is not a description length.

7. Measure Complexity

Ref. 77 proposed that the appropriate measure of the complexity of a process was the "minimal average Shannon information needed" for optimal prediction. This *true measure complexity* was to be taken as the Shannon entropy of the states used by some optimal predictor. The same paper suggested that it could be approximated (from below) by the excess entropy; there called the *effective measure complexity*, as noted in Section VI above. This is a position closely allied to that of computational mechanics, to Rissanen's MDL principle, and to the minimal embeddings introduced by the "geometry of a time series" methods[96] just described.

In contrast to computational mechanics, however, the key notion of "optimal prediction" was left undefined, as were the nature and construction of the states of the optimal predictor. In fact, the predictors used required knowing the process's underlying equations of motion. Moreover, the statistical complexity $C_\mu(\mathcal{S})$ differs from the measure complexities in that it is based on the well defined causal states, whose optimal predictive powers are in turn precisely defined. Thus, computational mechanics is an operational and constructive formalization of the insights expressed in ref. 77.

8. Hierarchical Scaling Complexity

Introduced in ref. 129, ch. 9, this approach seeks, like computational mechanics, to extend certain traditional ideas of statistical physics. In brief, the method is to construct a hierarchy of n^{th}-order Markov models and examine the convergence of their predictions with the real distribution of observables as $n \to \infty$. The discrepancy between prediction and reality is, moreover, defined information theoretically, in terms of the relative entropy or Kullback–Leibler distance.[63, 73] (We have not used this quantity.) The approach implements Weiss's discovery that for finite-state sources there is a structural distinction between block-Markovian sources (*subshifts of finite type*) and *sofic systems*. Weiss showed that, despite their finite memory, sofic systems are the limit of an infinite series of increasingly larger block-Markovian sources.[130]

The hierarchical-scaling-complexity approach has several advantages, particularly its ability to handle issues of scaling in a natural way (see

ref. 129, sec. 9.5). Nonetheless, it does not attain all the goals set in Section II F. Its Markovian predictors are so many black boxes, saying little or nothing about the hidden states of the process, their causal connections, or the intrinsic computation carried on by the process. All of these properties, as we have shown, are manifest from the ϵ-machine. We suggest that a productive line of future work would be to investigate the relationship between hierarchical scaling complexity and computational mechanics, and to see whether they can be synthesized. Along these lines, hierarchical scaling complexity reminds us somewhat of hierarchical ϵ-machine reconstruction described in ref. 5.

9. Continuous Dynamical Computing

Using dynamical systems as computers has become increasingly attractive over the last ten years or so among physicists, computer scientists, and others exploring the physical basis of computation.[131–134] These proposals have ranged from highly abstract ideas about how to embed Turing machines in discrete-time nonlinear continuous maps[7, 135] to, more recently, schemes for specialized numerical computation that could in principle be implemented in current hardware.[136] All of them, however, have been synthetic, in the sense that they concern *designing* dynamical systems that implement a given desired computation or family of computations. In contrast, one of the central questions of computational mechanics is exactly the converse: *given* a dynamical system, how can one detect what it is intrinsically computing?

We believe that having a mathematical basis and a set of tools for answering this question are important to the synthetic, engineering approach to dynamical computing. Using these tools we may be able to discover, for example, novel forms of computation embedded in natural processes that operate at higher speeds, with less energy, and with fewer physical degrees of freedom than currently possible.

ACKNOWLEDGMENTS

We thank Dave Albers, Dave Feldman, Jon Fetter, Rob Haslinger, Wim Hordijk, Amihan Huesmann, Kris Klinkner, Cris Moore, Mitch Porter, Erik van Nimwegen, and Karl Young for advice on the manuscript. We also wish to thank the participants in the 1998 SFI Complex Systems Summer School, the University of Wisconsin-Madison probability seminar, the UW-Madison Physics Department's graduate student mini-colloquium, and the University of Michigan-Ann Arbor Complex Systems seminar for numerous helpful comments on presentations of these results.

This work was supported at the Santa Fe Institute under the Computation, Dynamics, and Inference Program via ONR grant N00014-95-1-0975, NSF grant PHY-9970158, and DARPA contract F30602-00-2-0583.

REFERENCES

1. J. M. Yeomans, *Statistical Mechanics of Phase Transitions* (Clarendon Press, Oxford, 1992).
2. P. Manneville, *Dissipative Structures and Weak Turbulence* (Academic Press, Boston, Massachusetts, 1990).
3. P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics* (Cambridge University Press, Cambridge, England, 1995).
4. M. C. Cross and P. Hohenberg, Pattern Formation Out of Equilibrium, *Reviews of Modern Physics* **65**:851–1112 (1993).
5. J. P. Crutchfield, The calculi of emergence: Computation, dynamics, and induction, *Physica D* **75**:11–54 (1994).
6. J. P. Crutchfield and K. Young, Inferring statistical complexity, *Physical Review Letters* **63**:105–108 (1989).
7. J. P. Crutchfield and K. Young, Computation at the onset of chaos, In Zurek et al.,[137] pages 223–269.
8. N. Perry and P.-M. Binder, Finite statistical complexity for sofic systems, *Physical Review E* **60**:459–463 (1999).
9. J. E. Hanson and J. P. Crutchfield, Computational mechanics of cellular automata: An example, *Physica D* **103**:169–189 (1997).
10. D. R. Upper, *Theory and Algorithms for Hidden Markov Models and Generalized Hidden Markov Models*, PhD thesis (University of California, Berkeley, 1997). Online from `http://www.santafe.edu/projects/CompMech/`.
11. J. P. Crutchfield and M. Mitchell, The evolution of emergent computation, *Proceedings of the National Academy of Sciences* **92**:10742–10746 (1995).
12. A. Witt, A. Neiman, and J. Kurths, Characterizing the dynamics of stochastic bistable systems by measures of complexity, *Physical Review E* **55**:5050–5059 (1997).
13. J. Delgado and R. V. Solé, Collective-induced computation, *Physical Review E* **55**:2338–2344 (1997).
14. W. M. Gonçalves, R. D. Pinto, J. C. Sartorelli, and M. J. de Oliveira, Inferring statistical complexity in the dripping faucet experiment, *Physica A* **257**:385–389 (1998).
15. A. J. Palmer, C. W. Fairall, and W. A. Brewer, Complexity in the atmosphere, *IEEE Transactions on Geoscience and Remote Sensing* **38**:2056–2063 (2000).
16. J. P. Crutchfield and C. R. Shalizi, Thermodynamic depth of causal states: Objective complexity via minimal representations, *Physical Review E* **59**:275–283 (1999). E-print, arxiv.org, cond-mat/9808147.
17. J. L. Borges, *Other Inquisitions, 1937–1952* (University of Texas Press, Austin, 1964). Trans. Ruth L. C. Simms.
18. J. P. Crutchfield, Semantics and thermodynamics, In *Nonlinear Modeling and Forecasting*, M. Casdagli and S. Eubank, eds., volume 12 of *Santa Fe Institute Studies in the Sciences of Complexity*, pages 317–359 (Addison-Wesley, Reading, Massachusetts, 1992).
19. Plato, *Phaedrus*.
20. A. R. Luria, *The Working Brain: An Introduction to Neuropsychology* (Basic Books, New York, 1973).

21. N. V. S. Graham, *Visual Pattern Analyzers*, volume 16 of *Oxford Psychology Series* (Oxford University Press, Oxford, 1989).
22. S. J. Shettleworth, *Cognition, Evolution and Behavior* (Oxford University Press, Oxford, 1998).
23. J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles* (Addison-Wesley, Reading, Massachusetts, 1974).
24. S. P. Banks, *Signal Processing, Image Processing, and Pattern Recognition* (Prentice Hall, New York, 1990).
25. J. S. Lim, *Two-Dimensional Signal and Image Processing* (Prentice Hall, New York, 1990).
26. Plato, *Meno*, In 80D Meno says: "How will you look for it, Socrates, when you do not know at all what it is? How will you aim to search for something you do not know at all? If you should meet it, how will you know that this is the thing that you did not know?" The same difficulty is raised in *Theaetetus*, 197 et seq.
27. A. N. Whitehead and B. Russell, *Principia Mathematica*, 2nd ed. (Cambridge University Press, Cambridge, England), pp. 1925–27.
28. B. Russell, *Introduction to Mathematical Philosophy*, The Muirhead Library of Philosophy, revised ed. (George Allen and Unwin, London, 1920). First edition, 1919. Reprinted New York: Dover Books, 1993.
29. J. P. Crutchfield, Information and its metric, In *Nonlinear Structures in Physical Systems—Pattern Formation, Chaos and Waves*, L. Lam and H. C. Morris, eds. (Springer-Verlag, New York, 1990), pp. 119.
30. B. Russell, *Human Knowledge: Its Scope and Limits* (Simon and Schuster, New York, 1948).
31. J. Rhodes, *Applications of Automata Theory and Algebra via the Mathematical Theory of Complexity to Biology, Physics, Psychology, Philosophy, Games, and Codes* (University of California, Berkeley, California, 1971).
32. C. L. Nehaniv and J. L. Rhodes, Krohn–Rhodes theory, hierarchies, and evolution, In *Mathematical Hierarchies and Biology: DIMACS workshop, November 13–15, 1996*, B. Mirkin, F. R. McMorris, F. S. Roberts, and A. Rzhetsky, eds., volume 37 of *DIMACS Series in Discrete Mathematics and Theoretical Computer Science* (American Mathematical Society, Providence, Rhode Island, 1997).
33. U. Grenander, *Elements of Pattern Theory*, Johns Hopkins Studies in the Mathematical Sciences (Johns Hopkins University Press, Baltimore, Maryland, 1996).
34. H. Jaeger, Observable operator models for discrete stochastic time series, *Neural Computation* **12**:1371–1398 (2000).
35. U. Grenander, Y. Chow, and D. M. Keenan, *Hands: A Pattern Theoretic Study of Biological Shapes*, volume 2 of *Research Notes in Neural Computing* (Springer-Verlag, New York, 1991).
36. U. Grenander and K. Manbeck, A stochastic shape and color model for defect detection in potatoes, *American Statistical Association* **12**:131–151 (1993).
37. A. N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems of Information Transmission* **1**:1–7 (1965).
38. G. Chaitin, On the length of programs for computing finite binary sequences, *Journal of the Association for Computing Machinery* **13**:547–569 (1966).
39. A. N. Kolmogorov, Combinatorial foundations of information theory and the calculus of probabilities, *Russian Mathematical Surveys* **38**:29 (1983).
40. M. Li and P. M. B. Vitanyi, *An Introduction to Kolmogorov Complexity and its Applications* (Springer-Verlag, New York, 1993).
41. M. Minsky, *Computation: Finite and Infinite Machines* (Prentice-Hall, Englewood Cliffs, New Jersey, 1967).

42. P. Martin-Löf, The definition of random sequences, *Information and Control* **9**:602–619 (1966).
43. L. A. Levin, Laws of information conservation (nongrowth) and aspects of the foundation of probability theory, *Problemy Peredachi Informatsii* **10**:30–35 (1974). Translation: *Problems of Information Transmission* **10**:206–210 (1974).
44. V. G. Gurzadyan, Kolmogorov complexity as a descriptor of cosmic microwave background maps, *Europhysics Letters* **46**:114–117 (1999).
45. R. J. Solomonoff, A formal theory of inductive inference, *Information and Control* **7**:1–22 and 224–254 (1964).
46. P. M. B. Vitányi and M. Li, Minimum description length induction, Bayesianism, and Kolmogorov complexity, E-print, arxiv.org, cs.LG/9901014, 1999.
47. G. W. Flake, *The Computational Beauty of Nature: Computer Explorations of Fractals, Chaos, Complex Systems and Adaptation* (MIT Press, Cambridge, Massachusetts, 1998).
48. J. Rissanen, *Stochastic Complexity in Statistical Inquiry* (World Scientific, Singapore, 1989).
49. C. H. Bennett, How to define complexity in physics, and why, In Zurek et al.,[137] pages 137–148.
50. M. Koppel, Complexity, depth, and sophistication, *Complex Systems* **1**:1087–1091 (1987).
51. M. Koppel and H. Atlan, An almost machine-independent theory of program-length complexity, sophistication and induction, *Information Sciences* **56**:23–44 (1991).
52. D. C. Dennett, Real patterns, *Journal of Philosophy* **88**:27–51 (1991). Reprinted in ref. 138.
53. J. P. Crutchfield, Is anything ever new? Considering emergence, In *Complexity: Metaphors, Models, and Reality*, G. Cowan, D. Pines, and D. Melzner, eds., volume 19 of *Santa Fe Institute Studies in the Sciences of Complexity* (Addison-Wesley, Reading, Massachusetts, 1994), pp. 479–497.
54. J. H. Holland, *Emergence: From Chaos to Order* (Addison-Wesley, Reading, Massachusetts, 1998).
55. L. Boltzmann, *Lectures on Gas Theory* (University of California Press, Berkeley, 1964).
56. H. Cramér, *Mathematical Methods of Statistics* (Almqvist and Wiksells, Uppsala, 1945). Republished by Princeton University Press, 1946, as vol. 9 in the Princeton Mathematics Series, and as a paperback, in the Princeton Landmarks in Mathematics and Physics series, 1999.
57. C. E. Shannon, A mathematical theory of communication, *Bell System Technical Journal* **27**:379–423 (1948). Reprinted in ref. 139.
58. D. Hume, *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects* (John Noon, London, 1739). Reprint (Oxford: Clarendon Press, 1951) of original edition, with notes and analytical index.
59. M. Bunge, *Causality: The Place of the Causal Princple in Modern Science* (Harvard University Press, Cambridge, Massachusetts, 1959). Reprinted as *Causality and Modern Science*, NY: Dover Books, 1979.
60. W. C. Salmon, *Scientific Explanation and the Causal Structure of the World* (Princeton University Press, Princeton, 1984).
61. P. Billingsley, *Ergodic Theory and Information*, Tracts on Probablity and Mathematical Statistics (Wiley, New York, 1965).
62. R. M. Gray, *Entropy and Information Theory* (Springer-Verlag, New York, 1990).
63. T. M. Cover and J. A. Thomas, *Elements of Information Theory* (Wiley, New York, 1991).

64. W. O. Ockham, *Philosophical Writings: A Selection, Translated, with an Introduction, by Philotheus Boehner, O.F.M., Late Professor of Philosophy, The Franciscan Institute* (Bobbs-Merrill, Indianapolis, 1964). First pub. various European cities, early 1300s.
65. Anonymous, Kuan Yin Tzu, T'ang Dynasty, Written in China during the T'ang dynasty. Partial translation in Joseph Needham, *Science and Civilisation in China*, vol. II (Cambridge University Press, 1956), p. 73.
66. D. P. Feldman and J. P. Crutchfield, Discovering non-critical organization: Statistical mechanical, information theoretic, and computational views of patterns in simple one-dimensional spin systems, *Journal of Statistical Physics* submitted (1998), Santa Fe Institute Working Paper 98-04-026, http://www.santafe.edu/projects/CompMech/papers/DNCO.html.
67. J. E. Hopcroft and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation* (Addison-Wesley, Reading, 1979), 2nd edition of *Formal Languages and Their Relation to Automata*, 1969.
68. H. R. Lewis and C. H. Papadimitriou, *Elements of the Theory of Computation*, 2nd ed. (Prentice-Hall, Upper Saddle River, New Jersey, 1998).
69. J. G. Kemeny and J. L. Snell, *Finite Markov Chains* (Springer-Verlag, New York, 1976).
70. J. G. Kemeny, J. L. Snell, and A. W. Knapp, *Denumerable Markov Chains*, 2nd ed. (Springer-Verlag, New York, 1976).
71. J. E. Hanson, *Computational Mechanics of Cellular Automata*, PhD thesis (University of California, Berkeley, 1993).
72. G. Bateson, *Mind and Nature: A Necessary Unity* (E. P. Dutton, New York, 1979).
73. S. Kullback, *Information Theory and Statistics*, 2nd ed. (Dover Books, New York, 1968). First edition New York: Wiley, 1959.
74. K. L. Klinkner, C. R. Shalizi, and J. P. Crutchfield, Extensive state estimation: Reconstructing causal states by splitting, Manuscript in preparation, 2001.
75. J. P. Crutchfield and N. H. Packard, Symbolic dynamics of noisy chaos, *Physica D* **7**:201–223 (1983).
76. R. Shaw, *The Dripping Faucet as a Model Chaotic System* (Aerial Press, Santa Cruz, California, 1984).
77. P. Grassberger, Toward a quantitative theory of self-generated complexity, *International Journal of Theoretical Physics* **25**:907–938 (1986).
78. K. Lindgren and M. G. Nordahl, Complexity measures and cellular automata, *Complex Systems* **2**:409–440 (1988).
79. W. Li, On the relationship between complexity and entropy for Markov chains and regular languages, *Complex Systems* **5**:381–399 (1991).
80. D. Arnold, Information-theoretic analysis of phase transitions, *Complex Systems* **10**:143–155 (1996).
81. W. Bialek and N. Tishby, Predictive information, E-print, arxiv.org, cond-mat/9902341, 1999.
82. J. P. Crutchfield and D. P. Feldman, Statistical complexity of simple one-dimensional spin systems, *Physical Review E* **55**:1239R–1243R (1997).
83. W. R. Ashby, *An Introduction to Cybernetics* (Chapman and Hall, London, 1956).
84. H. Touchette and S. Lloyd, Information-theoretic limits of control, *Physical Review Letters* **84**:1156–1159 (1999).
85. A. Lempel and J. Ziv, Compression of two-dimensional data, *IEEE Transactions in Information Theory* **IT-32**:2–8 (1986).
86. D. P. Feldman, *Computational Mechanics of Classical Spin Systems*, PhD thesis (University of California, Davis, 1998). Online at `http://hornacek.coa.edu/dave/Thesis/thesis.html`.

87. D. G. Mayo, *Error and the Growth of Experimental Knowledge* (Science and Its Conceptual Foundations, University of Chicago Press, Chicago, 1996).
88. J. P. Crutchfield and C. Douglas, Imagined complexity: Learning a random process. Manuscript in preparation, 1999.
89. R. Lidl and G. Pilz, *Applied Abstract Algebra* (Springer, New York, 1984).
90. E. S. Ljapin, *Semigroups*, volume 3 of *Translations of Mathematical Monographs* (American Mathematical Society, Providence, Rhode Island, 1963).
91. K. Young, *The Grammar and Statistical Mechanics of Complex Physical Systems*, PhD thesis (University of California, Santa Cruz, 1991).
92. P. Billingsley, *Probability and Measure*, Wiley Series in Probability and Mathematical Statistics (Wiley, New York, 1979).
93. J. L. Doob, *Stochastic Processes*, Wiley Publications in Statistics (Wiley, New York, 1953).
94. M. Loéve, *Probability Theory*, 1st ed. (D. Van Nostrand Company, New York, 1955).
95. M. M. Rao, *Conditional Measures and Applications*, volume 177 of *Monographs and Textbooks in Pure and Applied Mathematics* (Marcel Dekker, New York, 1993).
96. N. H. Packard, J. P. Crutchfield, J. D. Farmer, and R. S. Shaw, Geometry from a time series, *Physical Review Letters* **45**:712–716 (1980).
97. 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 of *Lecture Notes in Mathematics* (Springer-Verlag, Berlin, 1981), pp. 366.
98. J. P. Crutchfield and B. S. McNamara, Equations of motion from a data series, *Complex Systems* **1**:417–452 (1987).
99. J. Neyman, *First Course in Probability and Statistics* (Henry Holt, New York, 1950).
100. D. Blackwell and M. A. Girshick, *Theory of Games and Statistical Decisions* (Wiley, New York, 1954). Reprinted New York: Dover Books, 1979.
101. R. D. Luce and H. Raiffa, *Games and Decisions: Introduction and Critical Survey* (Wiley, New York, 1957).
102. I. M. Gel'fand and A. M. Yaglom, Calculation of the amount of information about a random function contained in another such function, *Uspekhi Matematicheski Nauk* **12**:3–52 (1956). Trans. in *American Mathematical Society Translations* **12**(2):199–246 (1959).
103. P. E. Caines, *Linear Stochastic Systems* (Wiley, New York, 1988).
104. D. Blackwell and L. Koopmans, On the identifiability problem for functions of finite Markov chains, *Annals of Mathematical Statistics* **28**:1011–1015 (1957).
105. H. Ito, S.-I. Amari, and K. Kobayashi, Identifiability of hidden Markov information sources and their minimum degrees of freedom, *IEEE Transactions on Information Theory* **38**:324–333 (1992).
106. P. Algoet, Universal schemes for prediction, gambling and portfolio selection, *The Annals of Probability* **20**:901–941 (1992). See also an important Correction, *The Annals of Probability* **23**:474–478 (1995).
107. A. N. Kolmogorov, Interpolation und extrapolation von stationären zufälligen folgen, *Bull. Acad. Sci. U.S.S.R., Math.* **3**:3–14 (1941). In Russian with German summary.
108. N. Wiener, *Extrapolation, Interpolation, and Smoothing of Stationary Time Series: With Engineering Applications*, The Technology Press of the Massachusetts Institute of Technology, Cambridge, Massachusetts, 1949, "First published during the war as a classifed report to Section D_2, National Defense Research Council".
109. N. Wiener, *Nonlinear Problems in Random Theory* (The Technology Press of the Massachusetts Institute of Technology, Cambridge, Massachusetts, 1958).

110. M. Schetzen, *The Volterra and Wiener Theories of Nonlinear Systems*, 2nd ed. (Robert E. Krieger Publishing Company, Malabar, Florida, 1989). Reprint, with additions, of the first edition, New York: John Wiley, 1980.
111. N. Wiener, *Cybernetics: Or, Control and Communication in the Animal and the Machine*, 2nd ed. (MIT Press, Cambridge, Massachusetts, 1961). First edition New York: Wiley, 1948.
112. N. Chomsky, Three models for the description of language, *IRE Transactions on Information Theory* **2**:113 (1956).
113. N. Chomsky, *Syntactic Structures*, volume 4 of *Janua linguarum, series minor* (Mouton, The Hauge, 1957).
114. B. A. Trakhtenbrot and Y. M. Barzdin, *Finite Automata* (North-Holland, Amsterdam, 1973).
115. E. Charniak, *Statistical Language Learning*, Language, Speech and Communication (MIT Press, Cambridge, Massachusetts, 1993).
116. M. J. Kearns and U. V. Vazirani, *An Introduction to Computational Learning Theory* (MIT Press, Cambridge, Massachusetts, 1994).
117. V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed. (Springer-Verlag, Berlin, 2000).
118. L. G. Valiant, A theory of the learnable, *Communications of the Association for Computing Machinery* **27**:1134–1142 (1984).
119. M. A. Boden, Precis of "The Creative Mind: *Myths* and Mechanisms," *Behaviorial and Brain Sciences* **17**:519–531 (1994).
120. C. Thornton, *Truth from Trash: How Learning Makes Sense*, Complex Adaptive Systems (MIT Press, Cambridge, Massachusetts, 2000).
121. J. Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge University Press, Cambridge, England, 2000).
122. M. I. Jordan, ed., *Learning in Graphical Models*, volume 89 of *NATO Science Series D: Behavioral and Social Sciences* (Dordrecht, 1998).
123. P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*, Adaptive Computation and Machine Learning (MIT Press, Cambridge, Massachusetts, 2000).
124. D. V. Lindley, *Bayesian Statistics, a Review* (Society for Industrial and Applied Mathematics, Philadelphia, 1972).
125. J. Rissanen, A universal data compression system, *IEEE Transactions in Information Theory* **IT-29**:656–664 (1983).
126. P. Bühlmann and A. J. Wyner, *Variable length Markov chains*, Technical Report 497 (UC Berkeley Statistics Department, 1997). Online from `http://www.stat.berkeley.edu/tech-reports/`.
127. A. Lempel and J. Ziv, On the complexity of finite sequences, *IEEE Transactions in Information Theory* **IT-22**:75–81 (1976).
128. J. Ziv and A. Lempel, A universal algorithm for sequential data compression, *IEEE Transactions in Information Theory* **IT-23**:337–343 (1977).
129. R. Badii and A. Politi, *Complexity: Hierarchical Structures and Scaling in Physics*, volume 6 of *Cambridge Nonlinear Science Series* (Cambridge University Press, Cambridge, 1997).
130. B. Weiss, Subshifts of finite type and sofic systems, *Monatshefte für Mathematik* **77**:462–474 (1973).
131. C. Moore, Recursion theory on the reals and continuous-time computation, *Theoretical Computer Science* **162**:23–44 (1996).
132. C. Moore, Dynamical recognizers: Real-time language recognition by analog computers, *Theoretical Computer Science* **201**:99–136 (1998).

133. P. Orponen, A survey of continuous-time computation theory, In D.-Z. Du and K.-I. Ko, editors, *Advances in Algorithms, Languages, and Complexity* (Kluwer Academic, Dordrecht, 1997), pp. 209–224.
134. L. Blum, M. Shub, and S. Smale, On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines, *Bulletin of the American Mathematical Society* **21**:1–46 (1989).
135. C. Moore, Unpredictability and undecidability in dynamical systems, *Physical Review Letters* **64**:2354–2357 (1990).
136. S. Sinha and W. L. Ditto, Dynamics based computation, *Physical Review Letters* **81**:2156–2159 (1998).
137. W. H. Zurek, editor, *Complexity, Entropy, and the Physics of Information*, volume 8 of *Santa Fe Institute Studies in the Sciences of Complexity* (Addison-Wesley, Reading, Massachusetts, 1990).
138. D. C. Dennett, *Brainchildren: Essays on Designing Minds*, Representation and Mind (MIT Press, Cambridge, Massachusetts, 1997).
139. C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication* (University of Illinois Press, Urbana, Illinois, 1963).
140. B. F. Schutz, *Geometrical Methods of Mathematical Physics* (Cambridge University Press, Cambridge, England, 1980).