Editors' Suggestion

Machine-Learning Optimized Measurements of Chaotic Dynamical Systems via the Information Bottleneck

Kieran A. Murphy[1] and Dani S. Bassett[1,2,3]
[1]Department of Bioengineering, School of Engineering and Applied Science
[2]Department of Electrical and Systems Engineering, School of Engineering and Applied Science; Department of Neurology and Department of Psychiatry, Perelman School of Medicine; Department of Physics and Astronomy, College of Arts and Sciences, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA
[3]The Santa Fe Institute, Santa Fe, New Mexico 87501, USA

(Received 8 November 2023; accepted 9 April 2024; published 10 May 2024)

Deterministic chaos permits a precise notion of a "perfect measurement" as one that, when obtained repeatedly, captures all of the information created by the system's evolution with minimal redundancy. Finding an optimal measurement is challenging and has generally required intimate knowledge of the dynamics in the few cases where it has been done. We establish an equivalence between a perfect measurement and a variant of the information bottleneck. As a consequence, we can employ machine learning to optimize measurement processes that efficiently extract information from trajectory data. We obtain approximately optimal measurements for multiple chaotic maps and lay the necessary groundwork for efficient information extraction from general time series.

DOI: 10.1103/PhysRevLett.132.197201

Encapsulated in deterministic chaos is the fundamental obstruction to predictability that can result from nonlinearity in a system's evolution, even in the absence of randomness [1,2]. Signatures of chaos are found broadly, from weather [3,4] to the brain [5,6], and tools developed in the study of chaos have been applied more broadly still [7,8]. Advancing capabilities to forecast chaotic dynamics thus has marked potential for impact. The challenge may be glimpsed through the relation between precision and predictability: for any predictive model utilizing less than infinite precision, the error of prediction grows exponentially [9]. Given the inherent difficulties and the potential for impact, the field has recently turned to machine learning [10–13].

Machine-learning approaches to forecasting chaotic dynamics generically utilize full-precision states as input to the predictive model. Yet, the elusive determinism of chaos gives rise to a curious fact: beyond a certain precision per state, the ability to forecast given a partial trajectory saturates [1]. There is thus a measurement capacity beyond which resources—whether to acquire the measurement or to record the trajectory—are wasted. Instead, it is sufficient to discretize the continuous-valued states, yielding an analogous system with "symbolic dynamics" that simplifies the statistical analysis of the system [7,14–16] and can be used for various applications [17], including anomaly detection [18] and communication [19]. Here we employ machine learning to optimize the measurement of and, equivalently, the extraction of information from a chaotic system.

The requisite precision per measurement, in the form of a number of bits per state, is a fundamental quantity of the system: the metric entropy [2]. Also known as the Kolmogorov-Sinai (KS) entropy, it corresponds to the rate of information creation, generated from infinitesimal scales by the expansion of nearby points under the dynamics, and is commonly referred to as sensitive dependence on initial conditions [1,20]. For many systems of interest, the metric entropy is equal to the sum of the system's positive Lyapunov exponents [2].

A finite metric entropy—which can be used to define chaos [21–23] and quantify the extent of chaos in a system [9,24]—implies that a system's continuous-valued trajectory through state space has the same information content, in the asymptotic limit, as a corresponding sequence of discrete-valued measurements, although only if the measurement process is optimal. Discrete-valued measurements *color* state space according to a partition, clustering states according to the partition element to which they belong. Optimizing over the enormous space of possible partitions has traditionally been avoided by finding special partitions called generators (or generating partitions) that are known to extract all information created by the dynamics [25]. Generators may be remarkably coarse [26]; one approximate generator for the Ikeda map [27] requires only two colors to partition state space [Fig. 1(a)]. Despite the intricate structure of the attractor, a measurement of the state with the capacity of 1 bit captures all information created by the dynamics.

Finding a generator is challenging and has been accomplished for only a limited number of low-dimensional systems. Traditionally, methods have leveraged intimate knowledge of the dynamics [28–34], reflecting the

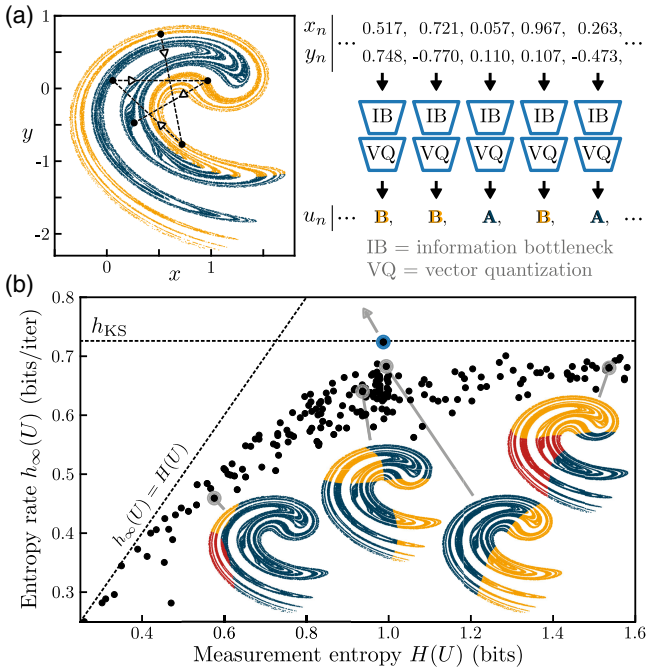

FIG. 1. (a) The attractor of the Ikeda map and a partial trajectory. States colored blue (orange) are measured as outcome A (B). The proposed method optimizes the measurement of a chaotic system using artificial neural networks that apply an IB and vector quantization (VQ) to each continuous-valued state. (b) The space of possible discrete measurements visualized in terms of the entropy of a single measurement $H(U)$ and the rate of entropy production in the infinite duration limit $h_\infty(U)$. The entropy rate of any partition is upper bounded by $H(U)$ and the metric entropy h_{KS} (dashed lines). The blue point corresponds to the partition in (a) that was optimized with the proposed method. The black points correspond to measurements parametrized by neural networks with random weights. Error bars on the entropy rates are within the markers.

fundamental connection between optimal information extraction and structure in the system's state space. As an alternative, purely data-driven approaches can reveal structure in the dynamical system with increased robustness to noise and the potential to scale to higher dimensions [26,35–38]. Whereas previous data-driven approaches have utilized classical methods of clustering with heuristics to simplify the search problem, here we bring the expressivity of deep learning to the task of optimizing information extraction from a chaotic system. Our focus is on partitions that perfectly extract information, of which generators are a subset [2], though in practice we find that the optimized partitions are closely related to previously established generators. Central to our approach is an equivalence between efficient information extraction and an objective from rate-distortion theory known as the distributed information bottleneck [39].

The distributed information bottleneck is a rate-distortion scenario to optimize the lossy compression of multiple sources of information individually so as to maximize collective information about an auxiliary quantity [39,40]. The lossy compression extracts some information and discards the rest and is accomplished by distributing an information bottleneck (IB) [41] to each source. By extracting important bits of information across multiple sources, the distributed IB has been used to decompose information in complex systems [42] and to provide interpretability to black-box machine-learning models by identifying the information utilized for prediction [40]. Here we use the distributed IB to lossily compress each state in a finite trajectory such that the sequence of measurements contains maximal information about a "reference" state taken from the trajectory.

Let a state $x \in \mathcal{M}$ exist in $\mathbb{R}^d$, where d is the dimension of the state space. A map $\mathcal{F}(x)\colon \mathcal{M} \to \mathcal{M}$ propagates a state forward in time by one iteration, i.e., $\mathcal{F}(x_n) = x_{n+1}$. We consider discrete time maps in this Letter; for the purposes of extracting information, a continuous-time flow can be converted to a discrete map by sampling with a fixed interval [2]. The dynamics are fully described by the map $\mathcal{F}$, but a probabilistic view is often more natural [7] and allows us to utilize information theory. For ergodic dynamical systems, which will be our concern in this Letter, the natural probability distribution over states $p(x)$ is an invariant measure, $\mathcal{F}[p(x)] = p(x)$, and can be obtained by iterating forward a long trajectory [2].

Given a probability distribution over states, we can define a random variable X_n for the state at time step n and a random process $\boldsymbol{X}_{n:n+L}$ for a sequence of random variables $X_n, X_{n+1}, \ldots, X_{n+L-1}$. For a stationary process, the start of the trajectory is unimportant, and instead we consider the statistics of subsequences of length L, which we denote $\boldsymbol{X}_L$.

A continuous-valued state can be "measured" and, specifically, discretized through the use of a partition that divides the support of $p(x)$ into disjoint subsets [26]. The outcome of a measurement, a random variable $U = f(X)$, converts the continuous-valued state to the index of the subset to which it belongs. Thus a sequence $\boldsymbol{X}_L$ is replaced by a sequence of discrete measurements $\boldsymbol{U}_L$.

How much information does a measurement U convey about the original state X? Intuitively, information gained reduces uncertainty. The amount of uncertainty about the outcome of a random variable Z may be quantified via the Shannon entropy, $H(Z) = \mathbb{E}_{z\sim p(z)}[-\log p(z)]$ [43]. The mutual information contained in two random variables is given by the reduction of entropy in one variable after finding the value of the other, $I(Z_1;Z_2) = H(Z_1) - H(Z_1|Z_2)$ [44]. Measuring a state—by recording in which subset of a partition it resides—conveys $I(U;X) = H(U)$ bits of information because the mapping from X to U is deterministic [i.e., $H(U|X) = 0$].

As a random process plays out, entropy is generated from the uncertainty about each subsequent outcome. The entropy rate is defined as the average entropy generated per

step in the limit where the sequence length becomes infinite, $h_\infty(U) = \lim_{L\to\infty} H(\boldsymbol{U}_L)/L$ [44]. The largest achievable entropy rate of any partition U is the KS entropy [2],

$$h_{\mathrm{KS}} = \sup_U h_\infty(U). \tag{1}$$

A "trivial" partition that achieves the maximal entropy rate can be approximated by a fine discretization of state space [24], though it is possible for partitions to capture all information produced by the dynamics and also be coarse, i.e., with minimal entropy [26] [Fig. 1(a)].

The objective for a partition with minimal entropy and maximal entropy rate can be cast as a rate-distortion problem [44]. Viewing the measurement as a communication channel, we simultaneously minimize the information transmitted about a single state $I(U;X)$ while maximizing the information transmitted about the trajectory, $I(\boldsymbol{U}_L;\boldsymbol{X}_L)$. For a chaotic system, a single continuous-valued state contains the same information as any portion of its trajectory, $I(\boldsymbol{U}_L;X_{\mathrm{ref}}) = I(\boldsymbol{U}_L;\boldsymbol{X}_L)$. By using a large but finite L, we recover a distributed IB formulation where the states in a sequence serve as the distributed sources of information and the reference state is the auxiliary variable. We minimize the distributed IB Lagrangian [39,40] over possible measurements U,

$$\mathcal{L} = -I(\boldsymbol{U}_L;X_{\mathrm{ref}}) + \beta \sum_{i=0}^{L-1} I(U_i;X_i), \tag{2}$$

where β is a parameter that determines the cost of information transmission relative to information about X_{ref}.

The space of possible partitions is vast, filled with colorings of the attractor that encapsulate suboptimal information [45,46]. Figure 1(b) displays example partitions of the Ikeda map by their measurement entropy $H(U)$ and entropy rate $h_\infty(U)$. By replacing the mutual information terms in Eq. (2) with appropriate bounds—namely, noise contrastive estimation (InfoNCE) [47] as a lower bound on $I(\boldsymbol{U}_L;X_{\mathrm{ref}})$ and the variational upper bound central to variational autoencoders on $I(U_i;X_i)$ [48–50]—we can train artificial neural networks that parametrize the space of partitions.

The lossy compression that maps each state x to its corresponding measurement u presents a difficulty for deep learning because quantization operations are nondifferentiable [51]. By contrast to *hard* measurements, where $H(U|X) = 0$, *soft* measurements contain stochasticity that can be used to facilitate optimization [52]. We trained with soft measurements that were hardened for inference and used a two-step compression process to gradually remove information about the state. Each step was performed by a distinct multilayer perceptron (MLP) [Fig. 1(a)]. The first MLP performed a soft measurement by mapping x to a distribution, $p(\tilde{u}|x)$, in an intermediate representation space where the Kullback-Leibler divergence penalty from variational autoencoders bottlenecked the transmitted information [48,49]. A sample from the distribution $p(\tilde{u}|x)$ was then input to a second MLP—the vector quantizer—whose output could be smoothly transitioned from a soft to a hard measurement. The output u was a probability vector over symbols [e.g., A and B in Fig. 1(a)] parametrized by a temperaturelike parameter that discretizes the assignment in one limit by making all probability distributions a one-hot vector [52]. The magnitude of the information bottleneck penalty β was annealed during training to gradually increase information transmission. For learning binary partitions of the systems studied in this Letter, training was stopped after the information transmitted by the information bottleneck (the first MLP) exceeded 1 bit.

All states $\boldsymbol{x}_L$ were measured and then the corresponding measurements $\boldsymbol{u}_L$ were input to a third MLP that predicted the reference state x_{ref}. To directly maximize information extraction $I(\boldsymbol{U}_L;X_{\mathrm{ref}})$, the InfoNCE loss [47] evaluates the prediction in a shared representation space to which both $\boldsymbol{u}_L$ and x_{ref} are mapped (requiring a fourth MLP that maps x_{ref} to the space). The loss quantifies how similar the representation for the sequence $\boldsymbol{u}_L$ is to that of its corresponding x_{ref}, in comparison to the representations of a batch of states sampled randomly from the entire attractor. Network architectures and training details may be found in the Supplemental Material [53].

To assess optimized measurements, the difference between the known metric entropy and the measurement's entropy rate gives the information per iteration that the measurement fails to capture. Lossless data compression commonly forms the basis of approaches to estimate entropy rate because the same regularities in a sequence that reduce its entropy rate are what data compression is designed to leverage to shrink file size [23,70–72]. We compared several methods on symbolized sequences from chaotic systems with known generators (Supplemental Material [53]) and obtained the most precise estimates with a form of data compression known as context tree weighting [73], whereby we extrapolated from finite size scaling with an ansatz proposed in Schürmann and Grassberger [23].

While strict equivalence between the distributed IB and a coarse partition with entropy rate equal to h_{KS} is valid only in the limit where the sequence length L is infinite, we found that $L \approx 10$ was sufficient to consistently find partitions with an entropy rate at least $0.99h_{\mathrm{KS}}$ for the chaotic maps studied (Fig. 2). Optimization was robust, with the performance of 20 trials tightly clustered for larger L (displayed as the violin plots in Fig. 2). Partitions with the largest entropy rate for $L = 12$ are displayed in the right column of Fig. 2. For the Hénon and logistic maps, the optimized partitions are slight variations of reported generators [23,28], while the partition for the Ikeda map differs from generators reported in prior work [26,30,35,37].

The generator for a chaotic map is not unique [29]. All images and preimages of a generating partition

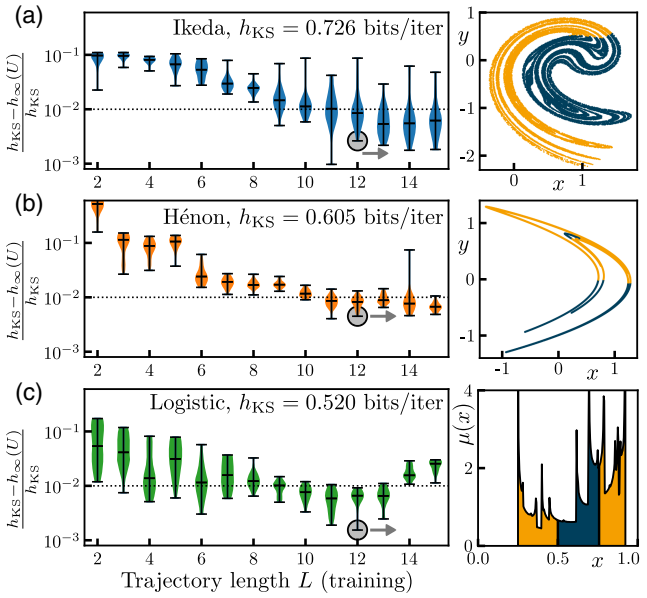

FIG. 2. The fractional deviation of the entropy rate $h_\infty(U)$ from the metric entropy h_{KS} as a function of the length of the trajectory L used for training for (a) Ikeda and (b) Hénon maps with standard parameters and (c) the logistic map with $r = 3.7115$. The distribution of values of the entropy rate over 20 trials for each L are shown as violin plots, with the extrema and median indicated by the black horizontal marks; the dotted line indicates $h_\infty(U) = 0.99 h_{KS}$. For each system, the partition found with the largest entropy rate for $L = 12$ is shown on the right. For the logistic map, the probability density of states $p(x)$ is colored according to the optimized partition.

(i.e., iterating the points in each partition element forward or backward in time while maintaining the assignments) are also generators. Additionally, there can be nontrivial variants that cross through special points called homoclinic tangencies, where the stable and unstable manifolds run parallel to one another [29]. We found that certain details of the training process steered optimization to qualitatively different partitions (Fig. 3). Varying the reference state drove the optimization to different iterates of the same base partition [Figs. 3(a)–3(d); iterates in Supplemental Material [53]], and a slower annealing of the information $\dot{\beta}$ reached deeper into the forward and backward iterates of the base partition [Fig. 3(b)]. Different iterates require different functions to integrate multiple measurements [Fig. 3(e)]; the iterates found through optimization allow for simpler integration functions. The base partition found for the Ikeda map [Fig. 3(a), reference state 5 of $L = 12$] closely resembles what has been found in prior work [30].

In this Letter, we established an equivalence between the distributed information bottleneck and the minimal partition able to capture all information generated by a chaotic system and leveraged the equivalence to optimize measurement of a chaotic dynamical system with machine learning. Operating without knowledge of the dynamics and without reliance on properties specific to generating partitions such as the unique description of periodic orbits [30], homoclinic tangencies [28,29,32,33], or the Koopman operator [34], the method is not restricted to chaotic systems. Instead, the deterministic chaos serves as a test bed, where the notion of an optimal measurement scheme

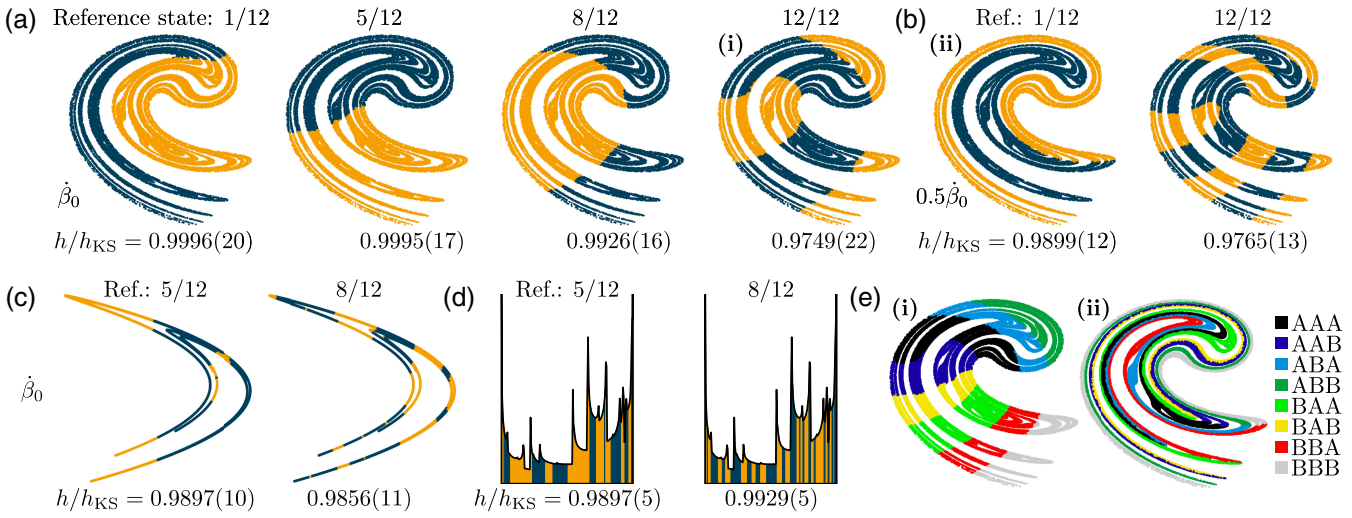

FIG. 3. (a) For optimized partitions of the Ikeda map, trained with sequences of length $L = 12$ and with a base information annealing rate $\dot{\beta} = \dot{\beta}_0$, the reference state (i.e., the one predicted from the sequence of measurements) strongly influences the found partition. Each displayed partition is the one with the largest entropy rate (listed under the partition) after ten trials. Coordinate axes have been suppressed. (b) The same as in (a), but with half the rate of information annealing $\dot{\beta} = 0.5\dot{\beta}_0$ during training. (c) Two of the partitions of the Hénon map, identified with the same annealing rate as in (a). (d) Same as in (c) for the logistic map with $r = 3.7115$. (e) Colorings of the Ikeda attractor where, given a sequence of three measurements $\boldsymbol{u}_{:3}$, the reference state is the final state x_3, utilizing for the measurement U the partition labeled (i) (left) and (ii) (right) in (a) and (b), respectively.

has been made precise and can be quantitatively evaluated. Optimizing a measurement to efficiently extract maximal information about an underlying state is a broad goal in understanding inference in biological and computational systems. Prior research has predominantly focused on a singular compression that extracts maximal information from the present and/or past about the future [74–77]. A notable exception is the recursive information bottleneck [78], in which a sequence of measurements are recursively aggregated and each step's measurement scheme can vary based on what has been previously observed. By contrast, the distributed information bottleneck setup of this Letter optimizes a fixed measurement scheme that is repeatedly applied for maximal aggregate information, a plausible scenario for sensing organisms or sensory devices.

The current study focused exclusively on relatively simple chaotic maps that have been well characterized so as to establish the capabilities of the proposed method. The variegated faces of chaos present exciting opportunities for the optimization of measurement processes and can serve as a rich test bed for machine-learning methods of compression [13,51]. In the large majority of systems where there is no precedent partition with which to compare, the metric entropy can be readily estimated from data [9,24] and other means of quantitatively evaluating partitions can be used [31] to ground the measurement schemes learned by the proposed method.

To our knowledge, the connection between metric entropy and a rate-distortion objective was previously unconsidered in the literature. The transformation $\mathcal{F}$ whose repeated application creates the strange attractor evolves points in state space such that a fixed measurement can continually acquire new information about the continuous-valued trajectory *ad infinitum*. By optimizing the minimally redundant lossy compression of the attractor, the spawned information that serves to limit predictability for any finite model is manifest as a specific coloring of the attractor that divides state space with a remarkably crude, yet highly specific cut.

We gratefully acknowledge Sam Dillavou, Sarah E. Marzen, Kevin A. Mitchell, and Navendu S. Patil for helpful discussions, as well as Jason Z. Kim and Suman Kulkarni for comments on the manuscript.

[1] R. Shaw, Strange attractors, chaotic behavior, and information flow, Z. Naturforsch. A **36**, 80 (1981).
[2] J.-P. Eckmann and D. Ruelle, Ergodic theory of chaos and strange attractors, Rev. Mod. Phys. **57**, 617 (1985).
[3] A. Tsonis and J. Elsner, Chaos, strange attractors, and weather, Bull. Am. Meteorol. Soc. **70**, 14 (1989).
[4] J. Slingo and T. Palmer, Uncertainty in weather and climate prediction, Phil. Trans. R. Soc. A **369**, 4751 (2011).
[5] S. Kargarnovin, C. Hernandez, F. V. Farahani, and W. Karwowski, Evidence of chaos in electroencephalogram signatures of human performance: A systematic review, Brain Sci. **13**, 813 (2023).
[6] J. E. Skinner, M. Molnar, T. Vybiral, and M. Mitra, Application of chaos theory to biology and medicine, Integrative physiological and behavioral science : the official Journal of the Pavlovian Society **27**, 39 (1992).
[7] G. Nicolis and C. Nicolis, *Foundations of Complex Systems: Emergence, Information and Prediction* (World Scientific, Singapore, 2012).
[8] E. Bradley and H. Kantz, Nonlinear time-series analysis revisited, Chaos **25**, 097610 (2015).
[9] D. J. Wales, Calculating the rate of loss of information from chaotic time series by forecasting, Nature (London) **350**, 485 (1991).
[10] J. Pathak, B. Hunt, M. Girvan, Z. Lu, and E. Ott, Model-free prediction of large spatiotemporally chaotic systems from data: A reservoir computing approach, Phys. Rev. Lett. **120**, 024102 (2018).
[11] P. Amil, M. C. Soriano, and C. Masoller, Machine learning algorithms for predicting the amplitude of chaotic laser pulses, Chaos **29**, 113111 (2019).
[12] Y. Tang, J. Kurths, W. Lin, E. Ott, and L. Kocarev, Introduction to focus issue: When machine learning meets complex systems: Networks, chaos, and nonlinear dynamics, Chaos **30**, 063151 (2020).
[13] W. Gilpin, Model scale versus domain knowledge in statistical forecasting of chaotic systems, Phys. Rev. Res. **5**, 043252 (2023).
[14] S. G. Williams, Introduction to symbolic dynamics, in *Proceedings of Symposia in Applied Mathematics* (American Mathematical Society, Providence, 2004), Vol. 60, pp. 1–12.
[15] M.-P. Béal and D. Perrin, Symbolic dynamics and finite automata, in *Handbook of Formal Languages: Volume 2. Linear Modeling: Background and Application* (Springer, New York, 2013), pp. 463–506.
[16] D. Lind and B. Marcus, *An Introduction to Symbolic Dynamics and Coding* (Cambridge University Press, Cambridge, 2021).
[17] Y. Hirata and J. M. Amigó, A review of symbolic dynamics and symbolic reconstruction of dynamical systems, Chaos **33**, 052101 (2023).
[18] A. Ray, Symbolic dynamic analysis of complex systems for anomaly detection, Signal Process. **84**, 1115 (2004).
[19] S. Hayes, C. Grebogi, E. Ott, and A. Mark, Experimental control of chaos for communication, Phys. Rev. Lett. **73**, 1781 (1994).
[20] R. G. James, K. Burke, and J. P. Crutchfield, Chaos forgets and remembers: Measuring information creation, destruction, and storage, Phys. Lett. A **378**, 2124 (2014).
[21] P. Gaspard and X.-J. Wang, Noise, chaos, and (ϵ, τ)-entropy per unit time, Phys. Rep. **235**, 291 (1993).
[22] C. Beck and F. Schögl, *Thermodynamics of chaotic systems*, Cambridge Nonlinear Science Series (Cambridge University Press, Cambridge, 1993).
[23] T. Schürmann and P. Grassberger, Entropy estimation of symbol sequences, Chaos **6**, 414 (1996).
[24] A. Cohen and I. Procaccia, Computing the Kolmogorov entropy from time signals of dissipative and conservative dynamical systems, Phys. Rev. A **31**, 1872 (1985).

[25] R. Badii and A. Politi, *Complexity: Hierarchical Structures and Scaling in Physics*, Cambridge Nonlinear Science Series (Cambridge University Press, Cambridge, 1997).
[26] M. B. Kennel and M. Buhl, Estimating good discrete partitions from observed data: Symbolic false nearest neighbors, Phys. Rev. Lett. **91**, 084102 (2003).
[27] K. Ikeda, Multiple-valued stationary state and its instability of the transmitted light by a ring cavity system, Opt. Commun. **30**, 257 (1979).
[28] P. Grassberger and H. Kantz, Generating partitions for the dissipative Hénon map, Phys. Lett. A **113**, 235 (1985).
[29] L. Jaeger and H. Kantz, Structure of generating partitions for two-dimensional maps, J. Phys. A **30**, L567 (1997).
[30] R. L. Davidchack, Y.-C. Lai, E. M. Bollt, and M. Dhamala, Estimating generating partitions of chaotic systems by unstable periodic orbits, Phys. Rev. E **61**, 1353 (2000).
[31] J. Plumecoq and M. Lefranc, From template analysis to generating partitions II: Characterization of the symbolic encodings, Physica D (Amsterdam) **144**, 259 (2000).
[32] K. A. Mitchell, Partitioning two-dimensional mixed phase spaces, Physica D (Amsterdam) **241**, 1718 (2012).
[33] M. Chai and Y. Lan, Symbolic partition in chaotic maps, Chaos **31**, 033144 (2021).
[34] C. Zhang, H. Li, and Y. Lan, Phase space partition with Koopman analysis, Chaos **32**, 063132 (2022).
[35] Y. Hirata, K. Judd, and D. Kilminster, Estimating a generating partition from observed time series: Symbolic shadowing, Phys. Rev. E **70**, 016215 (2004).
[36] N. S. Patil and J. P. Cusumano, Empirical generating partitions of driven oscillators using optimized symbolic shadowing, Phys. Rev. E **98**, 032211 (2018).
[37] N. F. Ghalyan, D. J. Miller, and A. Ray, A locally optimal algorithm for estimating a generating partition from an observed time series and its application to anomaly detection, Neural Comput. **30**, 2500 (2018).
[38] N. Rubido, C. Grebogi, and M. S. Baptista, Entropy-based generating Markov partitions for complex systems, Chaos **28**, 033611 (2018).
[39] I. E. Aguerri and A. Zaidi, Distributed information bottleneck method for discrete and gaussian sources, in *International Zurich Seminar on Information and Communication (IZS 2018)* (ETH Zurich, Zurich, 2018), pp. 35–39.
[40] K. A. Murphy and D. S. Bassett, Interpretability with full complexity by constraining feature information, in *International Conference on Learning Representations (ICLR)* (2023).
[41] N. Tishby, F. C. Pereira, and W. Bialek, The information bottleneck method, arXiv:physics/0004057.
[42] K. A. Murphy and D. S. Bassett, Information decomposition to identify relevant variation in complex systems with machine learning, Proc. Natl. Acad. Sci. U.S.A. **121**, e2312988121 (2024).
[43] C. E. Shannon, A mathematical theory of communication, Bell Syst. Tech. J. **27**, 379 (1948).
[44] T. M. Cover and J. A. Thomas, *Elements of Information Theory* (John Wiley & Sons, New York, 1999).
[45] E. M. Bollt, T. Stanford, Y.-C. Lai, and K. Życzkowski, What symbolic dynamics do we get with a misplaced partition? On the validity of threshold crossings analysis of chaotic time-series, Physica D (Amsterdam) **154**, 259 (2001).
[46] C. Cafaro, W. M. Lord, J. Sun, and E. M. Bollt, Causation entropy from symbolic representations of dynamical systems, Chaos **25**, 043106 (2015).
[47] A. v. d. Oord, Y. Li, and O. Vinyals, Representation learning with contrastive predictive coding, arXiv:1807.03748.
[48] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, beta-VAE: Learning basic visual concepts with a constrained variational framework, in *International Conference on Learning Representations* (2017).
[49] A. A. Alemi, I. Fischer, J. V. Dillon, and K. Murphy, Deep variational information bottleneck, in *International Conference on Learning Representations* (2017).
[50] I. E. Aguerri and A. Zaidi, Distributed variational representation learning, IEEE Trans. Pattern Anal. Mach. Intell. **43**, 120 (2021).
[51] E. Jang, S. Gu, and B. Poole, Categorical reparameterization with Gumbel-softmax, in *International Conference on Learning Representations* (2017).
[52] E. Agustsson, F. Mentzer, M. Tschannen, L. Cavigelli, R. Timofte, L. Benini, and L. V. Gool, Soft-to-hard vector quantization for end-to-end learning compressible representations, in *Advances in Neural Information Processing Systems* (Curran Associates, NY, 2017), Vol. 30.
[53] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevLett.132.197201 for additional information about the machine -learning implementation and benchmarking results for multiple methods of estimating the entropy rate of sequences, which includes Refs. [54–69].
[54] M. B. Kennel, J. Shlens, H. D. I. Abarbanel, and E. J. Chichilnisky, Estimating entropy rates with Bayesian confidence intervals, Neural Comput. **17**, 1531 (2005).
[55] S. Ro, B. Guo, A. Shih, T. V. Phan, R. H. Austin, D. Levine, P. M. Chaikin, and S. Martiniani, Model-free measurement of local entropy production and extractable work in active matter, Phys. Rev. Lett. **129**, 220601 (2022).
[56] T. Schürmann, A note on entropy estimation, Neural Comput. **27**, 2097 (2015).
[57] D. P. Kingma and M. Welling, Auto-encoding variational Bayes, in *International Conference on Learning Representations (ICLR)* (2014).
[58] A. Kolchinsky, B. D. Tracey, and D. H. Wolpert, Nonlinear information bottleneck, Entropy **21**, 1181 (2019).
[59] B. Poole, S. Ozair, A. Van Den Oord, A. Alemi, and G. Tucker, On variational bounds of mutual information, in *International Conference on Machine Learning* (PMLR, 2019), pp. 5171–5180.
[60] D. Maliniak, R. Powers, and B. F. Walter, The gender citation gap in international relations, Int. Organ. **67**, 889 (2013).
[61] N. Caplar, S. Tacchella, and S. Birrer, Quantitative evaluation of gender bias in astronomical publications from citation counts, Nat. Astron. **1**, 1 (2017).
[62] P. Chakravartty, R. Kuo, V. Grubbs, and C. McIlwain, #CommunicationSoWhite, J. Commun. **68**, 254 (2018).

[63] M. L. Dion, J. L. Sumner, and S. M. Mitchell, Gendered citation patterns across political science and social science methodology fields, Political Anal. **26**, 312 (2018).
[64] J. D. Dworkin, K. A. Linn, E. G. Teich, P. Zurn, R. T. Shinohara, and D. S. Bassett, The extent and drivers of gender imbalance in neuroscience reference lists, Nat. Neurosci. **23**, 918 (2020).
[65] E. G. Teich, J. Z. Kim, C. W. Lynn, S. C. Simon, A. A. Klishin, K. P. Szymula, P. Srivastava, L. C. Bassett, P. Zurn, J. D. Dworkin *et al.*, Citation inequity and gendered citation practices in contemporary physics, Nat. Phys. **18**, 1161 (2022).
[66] P. Zurn, D. S. Bassett, and N. C. Rust, The citation diversity statement: A practice of transparency, a way of life, Trends Cognit. Sci. **24**, 669 (2020).
[67] J. Dworkin, P. Zurn, and D. S. Bassett, (In)citing action to realize an equitable future, Neuron **106**, 890 (2020).
[68] D. Zhou, E. J. Cornblath, J. Stiso, E. G. Teich, J. D. Dworkin, A. S. Blevins, and D. S. Bassett, Gender diversity statement and code notebook v1. 0, Zenodo (CERN, Switzerland, 2020).
[69] Z. Budrikis, Growing citation gender gap, Nat. Rev. Phys. **2**, 346 (2020).
[70] M. B. Kennel and A. I. Mees, Context-tree modeling of observed symbolic dynamics, Phys. Rev. E **66**, 056209 (2002).
[71] Y. Gao, I. Kontoyiannis, and E. Bienenstock, Estimating the entropy of binary time series: Methodology, some theory and a simulation study, Entropy **10**, 71 (2008).
[72] S. Martiniani, P. M. Chaikin, and D. Levine, Quantifying hidden order out of equilibrium, Phys. Rev. X **9**, 011031 (2019).
[73] F. M. Willems, Y. M. Shtarkov, and T. J. Tjalkens, The context-tree weighting method: Basic properties, IEEE Trans. Inf. Theory **41**, 653 (1995).
[74] F. Creutzig, A. Globerson, and N. Tishby, Past-future information bottleneck in dynamical systems, Phys. Rev. E **79**, 041925 (2009).
[75] S. Still, J. P. Crutchfield, and C. J. Ellison, Optimal causal inference: Estimating stored information and approximating causal architecture, Chaos **20**, 037111 (2010).
[76] S. E. Palmer, O. Marre, M. J. Berry, and W. Bialek, Predictive information in a sensory population, Proc. Natl. Acad. Sci. U.S.A. **112**, 6908 (2015).
[77] S. E. Marzen and J. P. Crutchfield, Predictive rate-distortion for infinite-order Markov processes, J. Stat. Phys. **163**, 1312 (2016).
[78] S. Still, Information bottleneck approach to predictive inference, Entropy **16**, 968 (2014).