3
Inferring the Logic of Collective Information Processors

Bryan C. Daniels

1. Introduction

Biology is full of examples of complicated, carefully regulated, and often slightly mysterious examples of information processing: A housefly senses a gust of air and moves its wings, evading my flyswatter; a cascade of signaling proteins translates a hormone detected at a cell's surface into the production of specific genes in the nucleus; a colony of ants finds a new suitable rock crevice to use as a nest after its old one is destroyed. In each case, large sets of individual components must coordinate to carry out different actions depending on an environmental input.

A major challenge for modern science is to connect the small-scale dynamics of these individual components to the information processing consequences at the larger scale of the aggregate whole. If we think of biological systems as performing computations, transforming sensory input into coordinated and adaptive output behavior (Figure 3.1), the goal is to comprehend the logic of these distributed computers.

In this chapter, I summarize a new approach for understanding collective information processing that is emerging at the interface of machine learning, statistical physics, information theory, and more traditional biological and social science. This approach can be viewed as expanding on existing notions of collective computation and distributed computing, which in the past focused mainly on theoretical results in cognitive science and neuroscience (Rumelhart and McClelland, 1996), in order to make them more data-rich and broaden them to include other systems such as collections of fish or ants or people or proteins (Figure 3.1; Couzin, 2009; Solé et al., 2016). Using extensive data sets, we are able to focus on how *specific* collective systems operate, going beyond generalized theory.

The challenge is formidable for at least three reasons: (1) the large number of interacting parts in each system means there are many potential contributors to

Bryan C. Daniels, *Inferring the Logic of Collective Information Processors* In: *Advances in Info-Metrics: Information and Information Processing across Disciplines.* Edited by: Min Chen, J. Michael Dunn, Amos Golan, and Aman Ullah, Oxford University Press (2021).
DOI: 10.1093/oso/9780190636685.003.0003

	Individual scale input		Aggregate output
Ants:	Nest quality	⟶	Nest choice
Neurons:	Sensory signals	⟶	Decision, action
Proteins:	Chemical signal	⟶	Gene expression
Fish:	Predator location	⟶	Group direction
Politicians:	Individual opinions	⟶	Law
Genes:	Developmental cues	⟶	Cell type

Figure 3.1 Interpreting adaptive collective behavior as computation. Across diverse biological and social systems, individuals detect information (input), such as the quality of a potential nest site visited by individual ants, that is combined to determine a single behavior of the aggregate (output), such as the nest choice of an entire ant colony. In each case, the aggregate output is the outcome of complicated interactions among a large number of individuals. Our goal is to understand how the aggregate behavior is produced by the behavior of individual components.

the final output, leading to large, unwieldy models; (2) the number and nonlinearity of interactions means model outputs are dependent on their component-level parameters in highly nontrivial ways, making parameter estimation difficult even if the structure of interactions is known; and (3) different examples of the same adaptive system often vary in their particular structure, making it necessary to build a new model for every new case.

In this chapter, I argue for a two-step process to best extract the logic of collective information processors (Figure 3.2). First, a suite of machine-learning approaches are used to infer a detailed model from data. This step gathers data into a predictive framework that encompasses the full complexity of the system at the level at which data is available. Second, and just as challenging, dimensionality reduction techniques are applied to the full model in order to produce simplified explanations. This abstraction step prunes away detail to more intuitively explain low-dimensional behavior at the aggregate scale.

Information measures are key to this perspective. At the most fundamental level, information processing provides a basic conceptual framework for what living systems are doing. In the case of collective systems, measures such as the Fisher information can be used to connect individual components to aggregate scales. Bayesian inference techniques benefit from concepts in information geometry, and Bayesian model selection can be interpreted as matching the amount of information encoded in the data to the models that describe the data. Finally, information compression is a natural interpretation of dimension reduction methods that lead to simplified understanding.

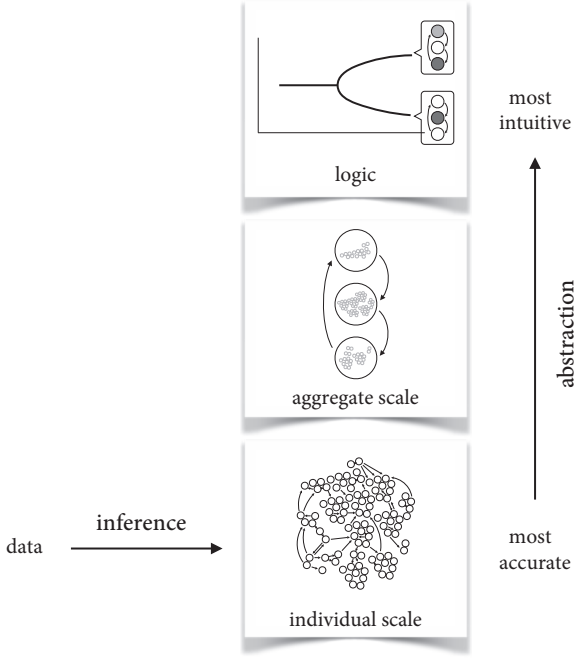

Figure 3.2 A two-step strategy for extracting the logic of biological collective information processors. First, a detailed model is inferred from data taken at the level of individual components. Next, successively more abstract representations allow a simplified understanding of the most important drivers of aggregate output.

The goals of machine learning, artificial intelligence, big data, and particularly data science partially overlap with what we seek in the science of collective behavior: predictive and simplified models of complex data sets. Yet, we are primarily motivated here by fundamental questions about living systems: How do collective systems remain adaptive? How do individual components manage to successfully regulate behavior that involves many other components? What is carefully tuned by evolutionary selection, and what is compensated for through active adjustment? What general strategies do distributed biological systems use to create adaptive logic?

Following the two-step framework of Figure 3.2, let us consider in more detail how to accomplish the tasks of inference and abstraction. We will look at each task in turn, reviewing the challenges that arise and the corresponding state-of-the-art methodologies being developed to address them.

2. First Task: Infer Individual-to-Aggregate Mapping

In biological collectives, the mapping from individual components to aggregate behavior is often unknown, difficult, intricate, degenerate, and nonlinear. The first challenge, then, is to transform data into a predictive model that connects the individual to the aggregate scale.

Performing inference, we search for the most accurate model we can find, evaluating our success by how well we can predict everything we can measure about the system's behavior. With the goal of encompassing all measured variables, often large in number in collective systems, this inference step is likely to produce a model that is particularly complex. This is represented at the bottom of Figure 3.2: a complicated model that acts as an accurate summary of our knowledge, a gathering of all relevant information.

Three major difficulties arise in the inference step:

- Heterogeneity and a large amount of potentially relevant detail at the individual scale (approached using big computers, big data, and a nuanced perspective on the relationship between modeling and theory)
- Unknown interaction structure and limited data (approached using maximum entropy and other effective modeling techniques, making use of model selection and regularization)
- Parameter compensation and emergence (approached using the concept of sloppiness and learning to live with parameter uncertainty)

Each of these issues has been the subject of extensive effort in the past few decades, which we review in this section.

2.1 Inference Challenge 1: An Abundance of Potentially Relevant Detail—Solved by Large-Scale Reverse Engineering

The fundamental challenge in understanding collective systems is clearly that they involve lots of parts. It is often impossible to measure and daunting to reason about the large number of individual-level properties that could be important.

In some cases, this complexity can be cleverly circumvented. The field of statistical physics produces simple mathematical explanations of material properties in terms of the collective behavior of atoms and molecules, using techniques like the renormalization group (Wilson, 1979; Goldenfeld, 1992). But the cleanest explanations rely on multiple assumptions that often do not hold in biological and social systems. First, explanations in statistical physics often

assume that we know the interactions between individual components and that these interactions are uniform across the system.[1] Second, in materials science, we are interested in scales that are extremely large compared to individual components—humans typically care about collective properties involving trillions of trillions of molecules. In short, we know how to parsimoniously describe systems with many contributing parts when interactions are simple and when we zoom out to include a huge number of them.

But the individual neurons in my brain are wired up differently than yours; the likely votes by Congress depend on changing individuals and changing opinions; a large number of distinct and overlapping gene-regulatory networks exist in a given cell. A typical number of individuals in these systems is hundreds to millions or at most billions, always vastly less than a trillion trillion. So what happens when individual components are diverse and changing, when there are a large but not huge number of them, and when each specific system involves myriad contingencies and historical contexts?

Such challenges are not new to science. We approach them as usual with experimentation and hypotheses, building models using carefully reasoned intuition and then testing them and gradually refining them. What is uniquely challenging about these systems is the volume of potentially important detail, the richness of information. It is difficult to hold all the important variables in one's head at once, and given the number of diverse systems we might want to understand, it takes too long to model each specific system anew.

Much effort has been aimed at the problem of systematically producing predictive models of such complicated collective behavior. The dominant conceptual framework is that of the network (Newman, 2010; Natale, Hoffmann, Hernández, and Nemenman, 2018). The proliferation of network explanations for collective biological behavior has ranged in scale from gene-regulatory systems (e.g., Bonneau, 2006; Peter and Davidson, 2017) and neurons (e.g., Bassett et al., 2011) to groups of organisms (e.g., Rosenthal et al., 2015). The implicit and reasonable assumption has been that the best way to proceed is simply to enumerate all the complicated details, pinning down the behavior of every individual and how it interacts with others. When this is tractable, the attitude is to be agnostic as to which details are most important to the aggregate behavior.

Network science is a large and successful scientific enterprise, and a huge number of methods have been developed to "reverse-engineer a network" from data (for a few representative cases and reviews, see Natale et al., 2018, and Bonneau et al., 2006). The zoo of existing methods embody a range of

[1] Or, if we do not know specific individual interactions, we assume any non-uniformities average out in such a way that we can understand the system in terms of a typical average individual.

typical modeling assumptions. For instance, we may assume that individuals are characterized by binary, discrete, or continuous states; that individuals either have dynamic states that update in discrete or continuous time, or that their joint states are described by an equilibrium function; and that interactions among individuals are described by linear or nonlinear functions. These modeling choices define the space of models over which an inference routine must search, which typically proceeds using a minimization algorithm that matches the model to statistics of the data. The results are often interpreted statistically in a Bayesian framework.

If one starts with abundant individual scale data, there is not much to decide in setting up network inference besides these initial modeling assumptions, and the strength of predictions will be determined by the veracity of the assumptions. Fast computers with large memories allow for inferring models of collective behavior with unprecedented detail.

In contrast to inferring detailed networks for specific systems, much of the initial theory of adaptive distributed systems relied on simplifying assumptions that did not require knowing all individual-level details. For instance, we may assume that gene-regulatory networks or neural networks are randomly connected, and then we may ask about the properties of these random networks (e.g., Kauffman, 1969; Amit, Gutfreund, and Sompolinsky, 1985). Classic results in parallel distributed computing ("neural networks") have shown that *arbitrary* computations can be carried out by abstracted neural units, if we are allowed to impose specific interactions.

Automated network inference is reaching an exciting point at which we can begin to theorize about collective behavior without having to rely on these assumptions. We instead can infer and use as our starting point a model of the full, messy, heterogeneous system, using data from, say, simultaneously measured neurons or simultaneously measured genes, as is now becoming routine. We may even study an ensemble representing typical cases of these networks, as are presently being accumulated in online repositories (Daniels et al., 2018).

Data and knowledge of the space of possible interactions are often limited, however, requiring new concepts, which we discuss in the following inference sections. Machine learning has faced this same problem and, in a limited sense, has already solved the problem of getting predictive power in complicated heterogeneous systems. The trade-off for gathering such extensive knowledge is that, in the extreme case, we must resign ourselves to being unable to have a human check all the details. Machine learning has already surpassed the speed at which humans can construct predictive models: automated language translation or image recognition involves incomprehensibly large sets of data, variables, and interactions, and the resulting learned models are represented in a way

that would not be easily understood or checked by a human. The potential usefulness to the science of collectives is that machine-learning techniques are already forming predictive models that capture important aggregate properties. These inscrutable but predictive descriptions provide a useful starting point for constructing understandable scientific theories of collective behavior.

Thus, the increasing speeds of machine learning and statistical inference give us the opportunity to model a large number of distinct systems. But beyond predictions and parsimonious descriptions of the aggregate scale, we also want to understand adaptive computations in terms of the actual individual-level interactions in each system. We want to know how a brain classifies images *using neurons*. How do we incorporate important but limited knowledge about mechanistic interactions at the scale of individual components?

2.2 Inference Challenge 2: Structural Uncertainty Due to Limited Data—Solved by Hierarchical Model Selection and Regularization

In our network inference problem, we want to know how individuals influence each other, yet often, even with detailed measurements, we do not have enough data to pick out the correct interactions from the space of all possible interactions. Even after selecting a particular class of models, the space of all possible models grows combinatorially with the number of individuals N—that is, it becomes huge. This means that even with a large amount of data, when N is moderately large, the space of models can easily overrun the space of all possible data. In this underdetermined regime, in which a model can perfectly fit any possible data, the model becomes useless. The challenge is then to efficiently make use of the detail present in the data while avoiding overfitting.

A particularly productive approach to this problem is to use models that match their level of complexity to the data and questions at hand. Complicated models can be produced if necessary, but when data is limited the model stays simple, in a way that produces better predictions. Intuitively, this works by ignoring smaller signals in the data that are likely caused by nonsystematic noise.

In this way, we match the level of detail, or dimensionality, that is supported by the data. Machine learning may be interpreted in this way more generally, where the process of restricting a model to lie in or close to a lower dimensional subspace is called regularization. This is dramatically realized, for instance, in reservoir computing, where input is nonlinearly transformed into a high-dimensional space and only the most predictive low-dimensional linear combination is retained (Lukoševicius and Jaeger, 2009). Note that such regularization

can also be interpreted as compression or model simplification (explored in section 3 of this chapter). Regularization can be used to find the appropriate level of complexity that produces the maximally accurate model (bottom of Figure 3.2), and similar techniques can then be used to throw away more detail and further reduce the dimensionality (moving toward the top of this figure).

Here, I highlight two approaches that use this conceptual structure. First, in a stochastic equilibrium modeling framework, we can construct a model with output that is as random as possible, adding interactions until statistics of the data are sufficiently well fit. This leads to maximum entropy approaches. Second, in a setting of deterministic dynamics, complexity in the form of nonlinear interactions and "hidden" unmeasured dynamical variables can be added until the system produces dynamics that fit time series data sufficiently well.

As a side note, one might object that putting a lot of effort into dealing with limited data is silly in that we should instead emphasize simply taking more or better data. Though a new experiment is often a good option, having "limited data" can sometimes be interpreted not as a problem with the experiment but a fact of life in the system.[2] Some systems, like a macaque society, have a stable structure only over a limited timescale. Taking more data is not an option, and asking for the "true" structure is not a well-defined question. Yet we may still want to characterize the interactions that are strong enough to have predictable effects.

2.2.1 Maximum Entropy Modeling

One powerful modeling approach in collective behavior is to treat observed states of the system as independent snapshots and then infer the probabilities with which all possible states of the system arise. This makes the most sense when dynamics are fast compared to the phenomena we are interested in and therefore ignorable. This type of model is common in equilibrium statistical physics.

A typical case starts with a system with N individuals, each of which can be either active or inactive—for instance, neurons that are firing or silent, or fish that are startled or calm. A static model produces the probability $p(\vec{x})$ of any given N-dimensional binary state $\vec{x}$ of active and inactive individuals. With enough data, we could estimate these probabilities by simply using the frequency with which every possible aggregate state occurs:

$$p(\vec{x}) \approx \frac{\text{number of observations of state } \vec{x}}{\text{total number of observations}}. \tag{3.1}$$

[2] This is also related to the problem of sloppiness discussed in section 2.3.

The glaring problem for large N is that there are 2^N possible states, so that getting an accurate estimate of these probabilities requires more than 2^N observations. The idea of the maximum entropy approach is to instead force the model to reproduce only statistics that we *can* measure accurately. Specifically, the entropy of $p(\vec{x})$ is maximized given the constraint of fitting some given statistics.

If we can make many measurements of the simultaneous states of individuals, then natural, easily observed statistics are the frequencies with which individuals are active, the frequencies with which pairs of individuals are jointly active, and so on. For instance, the probability that individuals i and j are jointly active puts a specific constraint on a marginal of $p(\vec{x})$:

$$p(i \text{ and } j \text{ active}) = \sum_{\vec{x} \text{ with } x_i=1, x_j=1} p(\vec{x}) \approx \frac{\text{number of observations of } i \text{ and } j \text{ jointly active}}{\text{total number of observations}}. \quad (3.2)$$

Typically, pairwise correlations will be most accurately captured by data, and higher-order correlations will require progressively more data. This motivates the typical form for a maximum entropy expansion, the form of which turns out to be straightforward to derive (Schneidman, Berry, Segev, and Bialek, 2006; Mora and Bialek, 2011; Daniels, Krakauer, and Flack, 2012, 2017):

$$p(\vec{x}) \propto \exp\left(-\sum_i h_i x_i - \sum_{ij} J_{ij} x_i x_j - \sum_{ijk} K_{ijk} x_i x_j x_k + \ldots\right). \quad (3.3)$$

The parameters J_{ij}, K_{ijk}, ... represent effective interactions between individuals that make specific subgroups more or less likely to be simultaneously active. While the form of the maximum entropy distribution (Eq. [3.3]) can be written down analytically, finding the parameters that match the statistics for a particular data set is a difficult inverse problem. Many approaches have been proposed for solving these inverse problems efficiently and in various approximations (Lee and Daniels, 2019).

The expansion after adding each term in Eq. (3.3) is the distribution with maximum entropy that fits those correlations. This is a form of hierarchical model selection: we add degrees of freedom (interactions) to the model until we fit the data well enough, but we do not go back to remove weak or unimportant lower-order interactions. Then each successive model in the list includes strictly more structure than the last, implying that the entropy monotonically decreases as we include more terms and thereby incorporate more information from the data. In cases for which we can estimate the entropy of the full distribution, we can track how much of the information the model captures as we add more

terms (Schneidman et al., 2006). The information captured by pairwise models is in many cases large, with not much left to be fit by higher-order interactions (Schneidman et al., 2006; Merchan and Nemenman, 2016).

Including all possible pairwise interactions can go too far and lead to overfitting if, for instance, some individuals are rarely active, meaning that joint activations are prohibitively rare. More sophisticated versions of the maximum entropy approach instead include only the most important interactions (Ganmor, Segev, and Schneidman, 2011) or use a cluster expansion that incorporates pairwise statistics only among clusters that contribute most to the joint entropy (Cocco and Monasson, 2012).

The maximum entropy approach is also useful in collective behavior in more general contexts than binary states and correlations. In principle, any state space and set of constrained statistics can be incorporated, such as the mean, variance, and correlations of the velocities of flocking birds (Bialek et al., 2014).

Maximum entropy models have been successfully applied to the collective behavior of multiple biological systems, including neurons (Schneidman et al., 2006), flocking birds (Bialek et al., 2014), and animal conflict (Daniels et al., 2012) (though see also warnings about extrapolating pairwise maximum entropy results to larger systems [Roudi, Nirenberg, and Latham, 2009] and the dangers of inferring interaction structure in the case of common input [Schwab, Nemenman, and Mehta, 2014]).

2.2.2 Dynamical Inference

In inferring dynamical systems, too, model selection can be used to adapt to the amount of information in the data. Imagine starting with data from a system that responds to an input with reproducible dynamics. This could be a cellular signal transduction cascade (Daniels et al., 2008), metabolic oscillations (Daniels and Nemenman, 2015), or a worm responding to sensory input (Daniels, Ryu, and Nemenman, 2019). The goal will be to represent the observed dynamics using a set of differential equations. In a very general form,

$$\begin{aligned}\frac{d}{dt}\vec{x}(t) &= \vec{f}\left(\vec{x}(t),\vec{y}(t),\vec{\theta}_x\right)\\ \frac{d}{dt}\vec{y}(t) &= \vec{g}\left(\vec{x}(t),\vec{y}(t),\vec{\theta}_y\right),\end{aligned} \tag{3.4}$$

where $\vec{x}$ is the vector of observed dynamical variables, $\vec{y}$ represents unobserved "hidden" variables, and $\vec{\theta}$ contains parameters controlling the dynamics of individual variables and their interactions. Of course, the important modeling

decisions arise in defining the forms of f and g, setting the space of possible models over which the inference scheme should search.

The general dynamical inference problem can quickly become unwieldy in that the space of possible models is enormous and impossible to search comprehensively. One possibility is to use no hidden variables, adding descriptive power (information) by increasing only the complexity of the function $\vec{f}$. The search space of possible functions can be constrained to a set of preset functions that combine to form $\vec{f}$, either using mathematically convenient functions (x, x^2, y, xy, $\cos(x)$, ...) (Schmidt and Lipson, 2009; Brunton, Proctor, and Kutz, 2016) or functions with biologically inspired nonlinearities ($\tanh(x)$, $(1+\exp(-x))^{-1}$, ...) (Daniels and Nemenman, 2015).

Another possibility is to add complexity both in the form of the right-hand sides of Eq. (3.4) and the addition of hidden dynamical variables $\vec{y}$, providing the opportunity to predict the existence of important unmeasured system components (Daniels and Nemenman, 2015). This approach is particularly useful when data are only available at an aggregate scale (as in Figure 3.4) or some individual-level details are missing. Unfortunately, this only adds to the enormity of the potential search space. Hierarchical model selection can be used as a counteraction: analogously to the maximum entropy expansion in Eq. (3.3), we make the model selection process more efficient by predefining a set of models of increasing complexity that can eventually fit any data, stopping the search when the model fits the data within experimental uncertainty (Daniels and Nemenman, 2015).

These dynamical inference approaches have been successful in producing predictive models from time series in physical systems (Schmidt and Lipson, 2009), simulated data from glycolysis oscillations (Daniels and Nemenman, 2015), and animal locomotion (Daniels et al., 2018).

2.3 Inference Challenge 3: Parameter Uncertainty Due to Scale Separation and Sloppiness—Solved by Bayes and not Focusing on Individual Parameters

Besides the difficulties encountered in the last section in constraining the interaction structure due to having a limited number of datapoints, there is another fundamental challenge to the inference of predictive models in collective systems: a problem that arises with data at the aggregate scale even if we have a huge amount of it. We think of models here as mappings from individual-level mechanisms to aggregate-scale consequences (Figure 3.3). The problem lies in

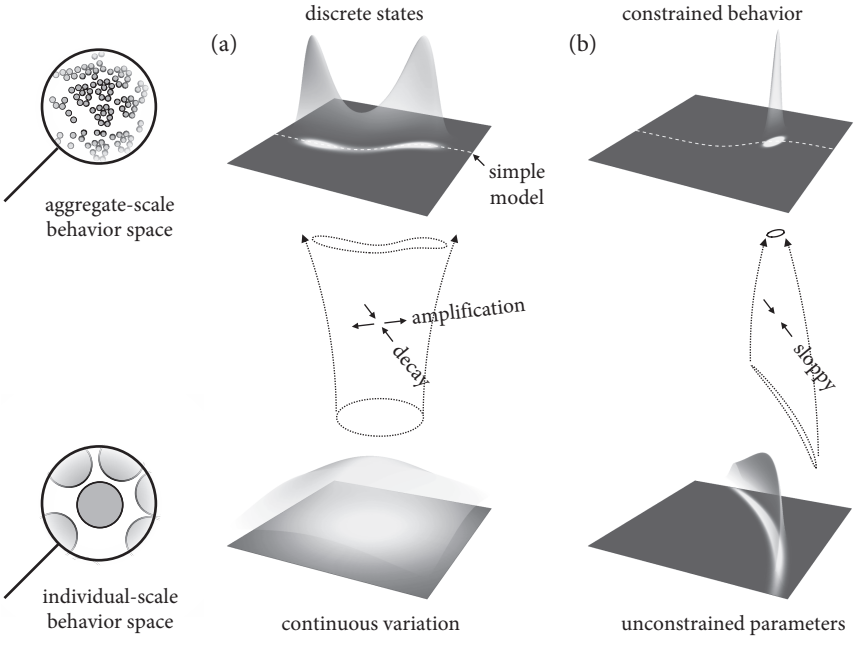

Figure 3.3 Parameter space compression, sloppiness, and simplified emergent models. A schematic representing the idea that some information is lost and some gets amplified in coarse-graining. The 3D plots represent probability distributions over behavior, both at the scale of individuals (bottom) and the scale of the aggregate (top). a) The large space of possible behaviors at the individual scale (bottom) typically maps onto a lower-dimensional output space at the aggregate scale (top). This can be understood as parameter space compression, the idea that moving to larger scales (up in this diagram) causes some effective parameter directions to decay. The dashed line on the top plot represents a simpler, lower-dimensional model that ignores the decaying parameter direction, but still well represents the collective behavior. This case also depicts a collective instability leading to amplification away from the center point. This creates two distinct possible aggregate behaviors, and relates to the idea of phases in statistical physics. See Section 3.2.1. b) For the same system, we now focus on a case in which we make a precise measurement at the aggregate scale. The same parameter space compression implies that even highly constrained aggregate behavior (top) typically corresponds to large regions of possible individual-scale parameters (bottom), leading to the phenomenon of sloppiness. See Section 2.3.

the fact that the typical properties of these mappings make it difficult to constrain parameters at the individual scale.

Not surprisingly, changes at the individual scale can have nonobvious effects—this is what makes complex systems interesting. Perturbing an individual fish may do nothing, while perturbing three fish simultaneously has huge effects (synergy). Removing a protein from a system entirely may modify an existing signaling pathway to take over and give similar output (compensation and robustness). These effects clearly make parameter fitting more challenging by virtue of being highly nonlinear.

What may be surprising, however, is that (1) there is some statistical regularity to the types of nonlinearities we find in these mappings in real systems (the phenomenon of sloppiness); and (2) the ubiquity of these phenomena point us toward an approach that deemphasizes finding the "correct" individual-scale parameters and instead uses Bayesian ensembles over parameters.

The idea is represented schematically in Figure 3.3. Parameters at the individual scale may be very unconstrained (wide probability distribution representing a huge space of possibilities at the bottom of Figure 3.3a), while they get mapped onto only a small space of possible aggregate behaviors (effectively lower-dimensional, thin-wedge probability distribution at the top of Figure 3.3a). This means that some directions in the individual-level parameter space are unimportant or "sloppy," while only a few are important or "stiff." The effect can be dramatic in that even taking lots of measurements to highly constrain the aggregate behavior (Figure 3.3b top) still leaves the possibility of huge swaths of parameter space at the individual level (Figure 3.3b bottom).

Properties of this mapping can be interpreted information-theoretically in terms of the Fisher information matrix. The Fisher information is a generalized measure of sensitivity of model outputs (what we call aggregate level properties) on model parameters. The Fisher information can be interpreted as the curvature of the Kullback-Leibler divergence of the distribution of model behavior as a model parameter is varied. With units of bits divided by the parameter's units squared, it answers the question of how quickly the output behavior becomes distinguishable when that parameter is varied. The Fisher information for parameter μ and aggregate state x is the average squared derivative of the log-likelihood function:

$$\mathcal{I}_x(\mu) = \int \left(\frac{\partial \log p(x)}{\partial \mu} \right)^2 p(x) dx. \tag{3.5}$$

The more general Fisher information matrix describes sensitivity to individual parameters and to simultaneous changes to pairs of parameters:

$$\mathcal{I}_x(\mu,\nu) = \int \frac{\partial \log p(x)}{\partial \mu} \frac{\partial \log p(x)}{\partial \nu} p(x) dx. \tag{3.6}$$

The Fisher information is also useful (see section 3.2.1) in quantifying functional sensitivity. Within model inference, it is a convenient means of describing parameter uncertainty.

The phenomenon of sloppiness describes properties of the Fisher information matrix that are common across a large number of complex dynamical models from systems biology and elsewhere (particularly those with many interacting components, the exact sort we encounter in modeling collective behavior). In most large nonlinear models, eigenvalues of the Fisher information matrix span many orders of magnitude and are roughly evenly spaced in log space. There are typically a few large "stiff" eigenvalues, with corresponding eigenvectors that describe directions in parameter space that are tightly constrained by experimental data. Remaining are a plentiful number of extremely small eigenvalues deemed "sloppy," corresponding to directions in parameter space that can be varied a large amount without changing the aggregate behavior (see Figure 3.3b) (Transtrum et al., 2015).

Roughly speaking, sloppiness is caused by nonlinear compensation: in large systems, it is usually the case that the aggregate scale effects of varying one parameter can be approximately canceled by varying some combination of other parameters. Typically, this sloppy compensation is highly nonlinear, so it is not easy to redefine parameters in such a way to remove sloppiness (as it would be if the mapping were linear).

This is important to inferring models of collective behavior because, even if we know the individual-scale interaction topology, typically many related parameters (rate constants, etc.) are unmeasured. We are then forced to fit them to the data, often using measurements at the aggregate scale. Sloppiness then implies that we will be unable to tightly constrain many directions in parameter space. The practical implication is that the aggregate-level data will be compatible with large swaths of parameter space (Figure 3.3b), leaving certain details of the individual scale unknown. Because sloppiness usually becomes extreme in large systems, this parameter uncertainty persists *even as we take lots of data* at the aggregate scale, becoming a fundamental problem for fitting models of collective behavior.

The solution: stop worrying about fitting a precise set of parameters. (See Daniels, Dobrzynski, and Fey, 2018) for a more detailed discussion of parameter estimation in systems biology.) One tactic is simply to choose a specific set of

parameters that sufficiently well fits the data. Surprisingly, this often produces predictive models (Transtrum et al., 2015) but can be dangerous: it is possible that predictions other than those used to constrain the model depend strongly on the position within the sloppy subspace of possible parameters. Safer is a Bayesian approach, which characterizes the entire sloppy subspace of parameters that fit the data within statistical uncertainty. The most straightforward way to do this is to use Monte Carlo methods to sample from the posterior over parameters; measuring a model output of interest using each member of the parameter ensemble then estimates the posterior distribution of the output.

More broadly, sloppiness suggests that lower-dimensional descriptions should be possible that succinctly capture the behavior controlled by stiff parameter directions. That brings us to our second task: finding simplified descriptions of collective behavior.

3. Second Task: Find Abstract System Logic

We don't model the trillion trillion molecules in a glass of water to predict what it does; instead we use effective models at a different scale. In the past one hundred years, statistical physics and high-energy physics have made great strides in understanding how this works in simple systems. Much of the excitement in the field of biosocial collective behavior is in learning how to analogously build simple but relevant effective models in more complicated systems.

What we get from inference procedures described in section 2 are explanatory models that may be arbitrarily complicated and not easily interpreted. This is the "black box" problem. Generalized methods from machine learning, such as neural networks and reservoir computing, are most prone to this problem, as they tend to intentionally overcompensate with the amount of included detail (Denker et al., 1987. But even approaches that explicitly favor simplicity (as in section 2.2) end up looking like a spaghetti tangle[3] when there is enough data to support it. That is, even a well-characterized system can be hard to understand. The recent excitement in neuroscience about ever-larger and more detailed data sets provides a good motivating example: Suppose we will someday be able to simultaneously measure and faithfully model every neuron in the human brain. Then what next?

Abstraction and simplification are crucial elements in the story of collective behavior. Across all the examples presented in Figure 3.1, the magic of collective behavior lies in large collections of individuals producing coherent

[3] Also known as a hairball (Lander, 2010) or ridiculogram (attributed to Marc Vidal).

low-dimensional dynamics. Our aim is to discover what drives this low-dimensional aggregate behavior and connect it to what we know about the complicated behavior of individuals. The challenge, then, is to find ways to compress existing detailed explanations into simplified understanding.

3.0.1 Why Do We Want to Do This? Advantages of Coarse-Grained, Dimensionality-Reduced Description

We put effort into simplifying models because systems become easier to understand when one finds the right coarse-grained representation. In some cases, we can put this in explicit information-theoretic terms: how many bits do I need to remember to predict system behavior (Daniels et al., 2012)? In some cases, it is a loose qualitative statement that, for instance, categorizing a system as implementing a Hopf bifurcation is easier to conceptualize than a detailed network representation. At a different level, simplified models are also important to understand in that compression is often happening in the system itself, with individual components cognitively adapting to their collective environment (Daniels et al., 2012; Flack, 2017).

This is part of a broader epistemological stance that recognizes "effective" models as legitimate and at times preferable to models defined at a detailed scale (Shalizi and Moore, 2003; Wolpert, Groschow, Libby, and DeDeo, 2015). It is also part of an ongoing debate about how to understand evolutionary forces acting at aggregate scales different than the familiar individual genome scale.[4] In the context of other complex, multiscale systems—for example, modeling whole cells (Babtie and Strumpf, 2017), animal behavior (Stephens, Osborne, and Bialek, 2011), or even abstract systems like cellular automata (Crutchfield and Mitchell, 1995)—this point is well appreciated: the goal is not only to encapsulate all of our knowledge in the most detailed model possible, but also to create approximations that are easier to work with, analytically and intuitively. There is a tension between our "best current understanding" or most accurate model and the model that gives the best intuition.

[4] This can turn into a lofty philosophical argument about epistemology and ontology—is our effective understanding ontologically "real" or just "an accurate description of our pathetic thinking about nature" (Gunawardens, 2014)? Is there an objectively "true" level at which aggregate objects and phenomena exist (Shalizi and Moore, 2003; Hoel, Albantakis, and Tononi, 2013)? Here we will instead focus pragmatically on predictive modeling—the best description is the one that makes the best predictions (which can depend on the question being asked). As was famously quipped by George Box: "All models are wrong but some are useful" (Box, 1979).

3.0.2 Do We Expect to Be Able to Compress? What Does "Logic" Look Like?

A modern understanding of why effective models work (Transtrum et al., 2015; Machta, Chachra, Transtrum, and Sethna, 2013) stems from renormalization group ideas used in statistical physics to characterize phases of matter.[5] The basic idea is to track how different types of interactions become more or less important as we "zoom out" from a system. Analogously to the Central Limit Theorem, the effects of some interactions are washed out, whereas other "relevant" interactions become more important. The relevant interactions are kept in the effective model, and we forget about the irrelevant ones, making for a simpler model. In this way, we have explicitly constructed the model so that it predicts the aspects that we deem most important. In physics, the most important aspects are typically those that occur at large spatial scales or low-energy scales.

In contrast to defining aggregate states in terms of space or energy, in adaptive collective behavior, the most important aspects are those that define informational properties. The key first question is then: What are the important aggregate states that we think of the system as computing? In the theory of computation, these aggregate logic states that define the computation are known as "information-bearing degrees of freedom" (Landauer, 1961). We refer to them here as informational states.

In aiming for simpler representations of adaptive systems, we can therefore be more focused in that we need not care about the simplest explanation of the system in general, but the simplest one that captures the informational states. In this sense, we want a parsimonious description of the "logic" or "algorithm" being implemented by the system. In neuroscience, cognitive science, and systems biology, it is common to talk about computations being performed by a system at the aggregate scale (Flack, 2017; Marr and Poggio, 1976; Dennett, 2014). The "logic" or "algorithm" that we want is precisely a simplified, compressed model of the mapping from the information contained in individuals to the information contained in the aggregate state (Marr and Poggio, 1976; Flack and Krakauer, 2011), one that might be used for control (Tomlin and Axelrod, 2005).

The focus here is on information and computation because, by definition, adaptive systems use relevant information about the state of the world to behave appropriately. Many aspects of adaptive systems are best understood in terms of maximizing relevant information (Nemenman, 2012; Sharpee, 2017), and biology is commonly conceptualized as being fundamentally informational

[5] In high-energy physics, renormalization explains how the laws of physics appear different when average energies are much different, as in, for instance, cosmological epochs just after the Big Bang.

(Krakauer et al., 2011; Davies and Walker, 2016). As one specific example, the visual system has been shown in multiple ways to be informationally optimized: the retina adapts to the statistics of incoming stimuli in a way that maximizes information transfer (Smirnakis et al., 1997), and the properties of retinal cell types produce optimal information transfer for images with statistics found in natural scenes (Kastner, Baccus, and Sharpee, 2015).

Essentially, our goals are the same as those of rate-distortion theory. We want to throw away information that is least important for the computation (lowering the "rate") and then to measure how well the reduced model performs (measuring the "distortion"). In the ideal case, the compressed representation exactly preserves all the aggregate properties that we care about predicting. Doing this efficiently—getting the most power for predicting aggregate properties given a limited amount of retained model information—is precisely the aim of rate-distortion theory and the closely related information bottleneck framework (Still, 2014).

In general, it is not guaranteed that model compression will work. We can imagine situations in which we are unlucky and predictions are impossible without knowing the precise state of every element of the system.[6] Happily for scientists, a (somewhat mysterious) property of our universe is that many details are often unimportant to what we care about (Transtrum, 2015). Compression techniques rely on systems being low-dimensional in some representation, and the trick is to find the right representation. As an example, the JPEG compression format preserves information that is most salient to human observers, and it does a good job for the typical sorts of images we encounter. Similarly, compression techniques for models need to know what aggregate-level features are important and cannot be "one size fits all."

Three broad approaches in particular have been useful in the case of collective behavior in living systems:

- Grouping into modules
- Focusing on aggregate-scale transitions: bifurcations, instability, and criticality
- Explicit model reduction

Note, however, that in general this endeavor of approximation and simplification is rather hopelessly broad and all-encompassing. A huge number of other related approaches exist, more or less specific to particular models and

[6] For instance, in computer engineering, a hash function produces output that changes dramatically with any small change to the input.

systems (e.g., searching through a set of possible structural forms in cognitive science (Kemp and Tenenbaum, 2008) or groupings of species (Feret et al., 2009) in biochemical signaling networks). There is no single correct way to do compression—this is the art of science.

3.1 Logic Approach 1: Emergent Grouped Logic Processors: Clustering, Modularity, Sparse Coding, and Motifs

One simple way to reduce dimensionality in models of collective behavior is to group components into distinct modules. This grouping can be accomplished using a variety of closely related concepts, including clustering, modularity, sparse coding, and community detection. The notion that groups can have distinct collective properties is well understood in, for instance, solid-state physics, where a phonon has a definite identity and physical effects but cannot be understood in terms of any individual molecule. Similarly, in many collective systems explanations formulated in terms of individual components are not well posed. For example: Which gene causes disease X? Which neuron causes decision Y? Which senator was responsible for passage of bill Z?

The phenomenon of modularity—whereby biological systems often consist of relatively independent subgroups—suggests that discovering these groups will produce more parsimonious descriptions. This can also be viewed as finding the important or natural scales of a system (Daniels, Ellison, Krakauer, and Flack, 2016). Particularly in neuroscience (e.g., Bassett et al., 2013) and genomics (e.g., Segal et al., 2003), clustering or searching for modules is used to interpret high-dimensional networks. Often, network inference procedures explicitly start with clustering before doing inference, effectively forcing the inference step to happen at a higher-order scale (Bonneau et al., 2006). Clustering can also be useful when performed at the higher level of dynamics and transitions among aggregate states. For instance, an inferred model of fly motion partitions behaviors into a hierarchical set of stereotypical movements (Berman, Bialek, and Shaevitz, 2016).

Most basic is clustering based on some intuitive notion of similarity—for instance, finding groups of individuals whose behavior is most correlated. This is also called community detection in networks. Similarly to network inference, a huge number of methods have been developed, and the best performing method will depend on the question being asked. Broadly, "hard clustering" methods separate components into nonoverlapping sets, and "soft clustering" allows components to be part of multiple groups. Some common general-purpose methods include k-means, hierarchical clustering, and multivariate Gaussians.

These basic clustering methods find intuitive groupings, but this does not yet necessarily give insight into the logic that produces a system's output. For instance, a simple clustering method will ask for the number of desired clusters k, but without further specifying the problem, we can only choose k arbitrarily. Instead, we want to choose k based on the aspects of the system we care most about, which in this case is its informational output. This leads us toward techniques written in the language of information theory, such as sparse coding (Daniels et al., 2012) or rate-distortion clustering methods (Slonim, Atwal, Tkacik, and Bialek, 2005). For instance, sparse coding can describe a system parsimoniously in terms of commonly appearing active subgroups, and then information theory can measure how much this reduces the information we need to remember in order to best predict future co-occurring individuals. The general idea is to specify a "distortion function," defining what information we want to retain,[7] and then to vary a single parameter λ that determines the complexity of the representation. For small λ, we favor more accuracy and more clusters, in the extreme case putting each component into its own cluster (at the bottom of Figure 3.2). As λ increases, we group components into larger clusters in a way that retains the most information about the system's output (moving toward the top of Figure 3.2).

Another approach, currently less automated, looks for patterns in network connectivity and dynamics that have known informational or logical functionality. These are known as functional motifs (Alon, 2007) or logical subcircuits (Peter and Davidson, 2017). The overabundance of some types of motifs is suggestive that they are more useful for information processing, and the computational properties of these motifs has been explored at length (e.g., Alon, 2007; Payne and Wagner, 2015). In this way, we can think of clusters and motifs as intuitive parts from which more complicated computations are built.

3.2 Logic Approach 2: Instability, Bifurcations, and Criticality

Grouping individual components is a useful step, but it may not tell us much about the system's logic. Another useful approach is to describe the system's behavior in terms of collective transitions or instabilities that control changes among aggregate informational states.

[7] Note that the optimal clusters depend on the distortion function. This captures the fact that the best representation of a system depends on which aspects of the system we consider to be relevant (Shalizi and Moore, 2003). In the case of systems for which we wish to understand the origins of a particular aggregate function, we can take this aggregate "output" to define the relevant informational states.

This focus on higher-level logic is also advantageous in that it sidesteps issues of the inability to constrain parameters (section 2.3), especially when there are unmeasured but important components (as is often the case in, for instance, neuroscience and cell biology). As one example, the identity of stable functional states in a gene-regulatory network has been shown to be robust to parameter changes (Jia, Jolly, and Levine, 2018).

Informational states often correspond to those that include many components or those seen at long timescales. These lead to two mathematical limits and two domains of theory: (1) the limit of many components leads to transitions studied in statistical physics using the language of phases and criticality, and (2) the limit of large time leads to transitions among attractors studied in dynamical systems using the language of bifurcations.

3.2.1 Fisher Information and Criticality

In an equilibrium system, we can make an analogy with statistical physics and talk about a system's phases: What are the coarse-grained aggregate states that characterize the system? As in Figure 3.3, information that washes out at the aggregate scale allows us to ignore some individual-scale details. Information that grows can produce well-separated, distinct aggregate states. These states are primed to become important at the aggregate scale as they carry specific information about the individual scale—they become informational states.

What causes these emergent phases? In physics and in collective behavior more generally, we think about this by considering how the system changes as we move away from the individual level—more atoms, more molecules, more people. Intuitively, whether information is amplified or decays depends on how perturbations spread through a system (Daniels, Krakauer, and Flack, in preparation). A perturbation may die out or be overwhelmed by noise, becoming smaller as it spreads to more individuals. This corresponds to an irrelevant, compressed direction in parameter space in Figure 3.3. Or the perturbation can be amplified, becoming larger in magnitude as it spreads, corresponding to a relevant, growing direction in Figure 3.3. It is this latter case that corresponds to a collective instability that can produce distinct aggregate states (Daniels et al., 2017; Daniels, in preparation). In renormalization group flows, this instability comes from an effective parameter value being amplified as the scale increases.[8] Importantly, these instabilities can be connected to computational functions such as consensus formation and decision making (Daniels, Flack, and Krakauer, 2017). This process through which instabilities create distinct

[8] More general cases of collective behavior are often trickier to represent formally in the renormalization group language because the aggregate states we are interested in are not always simple sums over individuals.

aggregate states, the transition between components behaving as if "anything goes" versus "we have all settled on this particular arrangement," is known as symmetry breaking in physics (Sharpee, 2017; Anderson, 1972; Sethna, 2006). The notion that symmetry breaking and the creation of distinct attractors are fundamental to defining meaning in biology has been explored by numerous authors (Solé et al., 2016; Anderson, 1972; Brender, 2012). Cell types are explained as attractors of gene-regulatory networks (Lang, Li, Collins, and Mehtr, 2014), and swarming and milling states of fish schools are interpreted as collective phases (Tunstrom et al., 2013).

As inputs or individual behavior change, how do they control changes in these aggregate states? This is key to describing information processing and can be measured using the same Fisher information that we used to measure sensitivity to parameters in section 2.3 (Eq. [3.5]). A crucial insight in making a connection to statistical physics is that phase transitions are defined by extreme system sensitivity (Daniels et al., 2017; Daniels et al., in preparation). This is intuitively clear in that small changes in control parameters lead to systemwide changes in behavior at a phase transition: changing your freezer's temperature from –1 degree C to +1 degree C creates very different behavior of the water molecules in the ice cube tray. This intuition is made sharp by the Fisher information, which has been shown to become infinite precisely at phase transitions (Prokopenko, Lizier, Obst, and Wang, 2011).[9]

We can think of the Fisher information as measuring amplification: the degree to which information at the small scale has large, aggregate-scale effects. This can measure the sensitivity of the structure in a social animal group to changes in individual bias toward conflict (Daniels et al., 2017) or the sensitivity of a group of fish to an individual who detects a predator (Sosna et al., 2019). In this way, the informational perspective is useful for framing the idea of phase transitions in biology. Even in finite systems (away from the limit of an infinite number of components that produces sharp "true" phase transitions), the Fisher information connects with biological function as a generalized measure of functional sensitivity.

When viewing biological collectives as computers whose output must be sensitive to changing input, it is perhaps unsurprising that many are found to lie near such instabilities. "Nearness to criticality" has been found across many collective systems (Mora and Bialek, 2011), ranging from neurons (Cocchi, Gollo, Zalesky, and Breakspear, 2017) to flocks (Bialek et al., 2014) to societies

[9] Technically speaking, this is true only at continuous-phase transitions because an energy barrier is associated with discontinuous-phase transitions (leading to hysteresis) that prevents a given (symmetry broken) state from being easily poked into a new aggregate state. Still, the long-time equilibrium state becomes infinitely sensitive to perturbations at discontinuous transitions.

(Daniels et al., 2017). Studies demonstrating criticality typically start with an inferred model and show that small changes to parameters can bring the system to a peak in sensitivity, or measure other indications of self-similarity arising from the system being at a marginal point between information amplification and decay (Daniels et al., in preparation).

It is not necessarily the case that maximal sensitivity is best, and it may be advantageous for adaptive systems to move closer or further from criticality. There is some evidence, for instance, that distance from criticality in the brain varies over the sleep–wake cycle (Priesemann, Valderrama, Wibral, and van Quyen, 2013). Also, fish schools change their density to respond to the perceived level of threat of a predator (Sosna et al., 2019). This can be interpreted as moving closer to criticality when fish feel threatened (letting changes spread more quickly through the school) and staying further from criticality otherwise (to avoid responding to random uninformative changes in individual behavior). In a social system, this distance has been measured in biologically meaningful units as the number of individuals whose behavior would have to change to get to a point of maximal sensitivity (Daniels et al., 2017).

3.2.2 Dynamical Systems and Bifurcations

If our model is deterministic and dynamic (as in dynamical inference, section 2.2.2), it often makes sense to think of the stable attractors as defining logic states of the system. This viewpoint, asking about the system's behavior in the limit as time $t \to \infty$, is analogous to the large number limit that defines phases in equilibrium models. And analogously to the mathematics of phase transitions, here the machinery of nonlinear dynamics (Strogatz, 1994) gives us the language of bifurcations for describing the changes of state that describe a system's logic. (This view of thinking of dynamics as computation has been explored and debated at length in cognitive science; e.g., Michell, 1994; Beer, 2014.)

As an example, consider a system that performs a reproducible dynamic response to a stimulus, such as a worm responding to the application of heat by changing its speed and direction of motion (Daniels, Ryu, and Nemenman, 2019). We are certain that this behavior is controlled by neurons and so we could attempt to infer a model at the level of individual neurons, but we will consider a case in which data at this level of detail is not available. Instead, as shown in Figure 3.4, we have data describing the speed of the worm as a function of time for trials with varying heat intensity, which we treat as input to the system. The adaptive model selection approaches described in section 2.2 can then be used to infer an effective model that uses the types of saturating interactions that are common to biological systems, capturing the dynamics at a coarse-grained level.

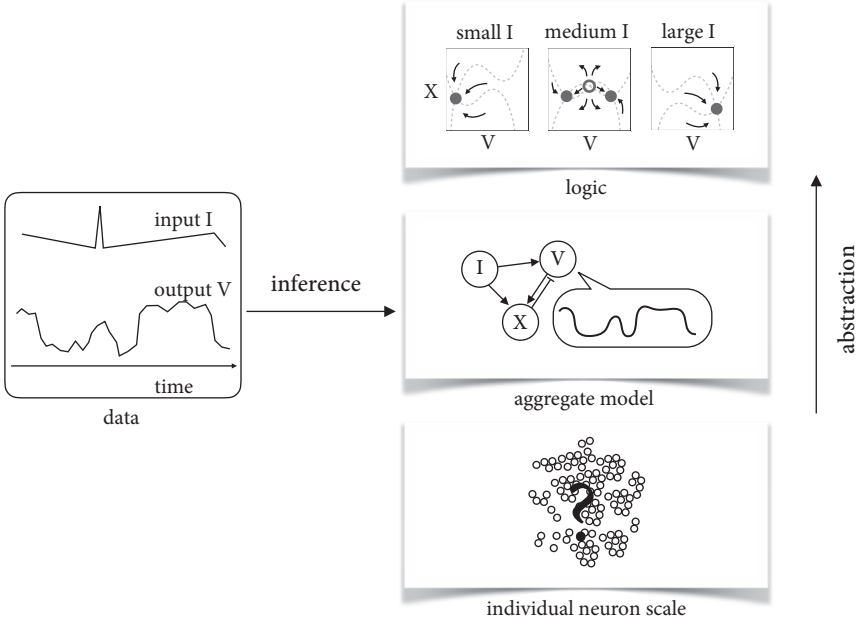

Figure 3.4 Phase space structure as logic. In this caricatured example, time series data leads to an effective dynamical model (represented as nodes and arrows in the middle row) that can predict the result of arbitrary dynamical input. The inference procedure makes use of effective dynamical variables, which can be interpreted as encapsulating the behavior of groups of individual neurons, and in this case includes an additional hidden variable *X*. Examining the phase space structure (top) produces a simple logic in terms of steady-state fixed points (filled dots, with an unstable fixed point as an unfilled dot, nullclines as dotted lines, and dynamical flow lines as solid arrows): the structure of the dynamics can be traced to a pair of saddle-node bifurcations.

Even this coarse-grained representation can be used as a starting point to describe the logic of the system. We abstract to the level of logic by examining the phase space structure of the inferred model, shown in the top row of Figure 3.4.[10] In this scenario, the heat input induces a bifurcation that switches the system between distinct patterns of motion. Even if the system does not ever

[10] In the case shown in Figure 3.4, the inferred model is two-dimensional. This case is particularly simple because, using Morse–Smale theory (Palis and de Melo, 1982), it is possible to uniquely classify almost all (compact) two-dimensional-phase portraits according to the number and types of attractors. It is not always possible to perform such a classification of dynamical systems in higher-dimensional cases. Particularly in cases of deterministic chaos, simpler abstract descriptions may not be possible.

saturate to fully reach one of the fixed-state attractors, examining the model in this way provides a succinct explanation for the switch-like behavior: it arises in this imagined example from a pair of saddle–node bifurcations. Furthermore, to the extent that the effective model is a sufficiently realistic representation of the behavior of coarse-grained groups of neurons, any future explanation in terms of individual neurons will be consistent with the inferred logic.

3.3 Logic Approach 3: Explicit Model Reduction

In the previous sections, we looked for specific reduced representations: grouping individual components, characterizing the system using sensitivity with respect to certain directions in parameter space, or deriving the structure of attractors. In a more general context, it has long been a dream to produce automated approximations and model simplifications that begin with a complicated model and produce a simpler model of the same type. The hope is to find explicit approximations of known detailed models.

Dynamical models written in the form of a Markov process can be analyzed using the ε machine formalism (Shalizi and Crutchfield, 2001). Starting with a known Markov process, this formalism defines the minimally complex Markov model that exactly reproduces the behavior of the original process. Recent developments have generalized this reasoning to the more typical case in which we cannot produce a smaller model that produces the exact same predictions, but instead we look to maximize predictive power while restricting the model size (Marzen and Crutchfield, 2017).

As we saw in section 2.3, sloppiness in models suggests that lower-dimensional representations should be good approximations. Another particularly elegant method uses the same information geometry that defines sloppiness to find such approximate models. Treating the Fisher information matrix as a metric tensor, following geodesics corresponds to mapping out the model manifold, the space of possible outputs of the model. Following the sloppiest direction corresponds to changing parameters in a way that minimally affects the measurable outputs and often approaches boundaries on the model manifold, places where taking combinations of parameters to infinity does not change the model output (Transtrum and Qiu, 2016). In this way, mathematical limits corresponding to simplifying approximations can be found in a semiautomated way. For example, starting with a more complicated mechanistic model for enzyme kinetics, the method can automatically discover the approximations that lead to the widely used Michaelis–Menten model, which assumes that the substrate is in instantaneous chemical equilibrium (Transtrum and Qiu, 2016).

4. The Future of the Science of Living Systems

We observe that proteins, neurons, ants, fish, politicians, and scientists each create structures that process information and perform impressive collective feats. Obtaining a deep understanding of how parts come together to act as an adaptive whole is a worthy challenge for modern science.

The relationship between simplified theories, statistical physics, and machine learning has been emphasized for many decades and continues to be fruitful (Denker et al., 1987; Seung, Sompolinsky, and Tishby, 1992; Mehta and Schwab, 2014). The intent here is to use these ideas to understand many specific cases of collective behavior observed in nature. Since each case is different and involves so many details, it will likely be necessary to construct a new model for every system. Automated methods will be key to expedite this process.

We might dream of a time in the not-too-distant future when all tasks in inferring abstracted models can be accomplished by an automated machine. Such a machine might use fish school data to locate a critical instability and demonstrate how information about predators is maximized when the school is closer to the instability. With spiking data from millions of neurons, it might group them into functional clusters and show how they create an aggregate-level gated sensory classifier. Using insulin expression data from millions of patients, it might produce both patient-specific predictions and an abstract dynamical framework for how interventions control the phase and frequency of oscillations. In each case, the two-step process of Figure 3.2 means that we get both the accuracy of the full messy predictive model and the parsimonious abstracted theory describing the aggregate level logic.

This is not an outrageous goal. It will require conceptual ingenuity both on the side of efficient inference and model selection and on paring down and interpreting predictive models once they are built. The payoff to practitioners is highly predictive models, and the payoff to science is a much improved position to understand overarching principles in biology.

What are we expecting to find? There are already hints that broader principles or strategies for collective information processing are at work. For instance, a two-phase picture for robust collective decision making, moving from a distributed uncertain phase to a redundant consensus phase, arises naturally when using the criticality framework of section 3.2.1. This two-phase implementation of decision-making corresponds to a logical structure that could be tested across a variety of biological and social systems (Daniels et al., 2017; Arehart, Jin, and Daniels, 2018). We are beginning to see how a variety of systems regulate distance from criticality at the individual scale, and how to classify behavioral dynamics based on whether they include bifurcations or instabilities that induce

qualitative changes to the phase space structure based on sensory input (e.g., Daniels et al., 2018).

What might this approach be missing? First, inferring a full generative, predictive model before using it to ask questions about logic and mechanism may be overkill. In at least one case of dynamical inference, some qualitative results about mechanisms that map from microscopic to macroscopic can be obtained without using a specific generative model (Barzel, Liu, and Barabási, 2015). For instance, aggregate states and distance from a symmetry-breaking transition can be identified and tracked using only observed variance in the states of individual players (Mojtahedi et al., 2016). A second related point is that this approach treats inference and abstraction as two separate steps. This creates a clean caricature, with the two parts having different (and competing) goals: inference favors predictability even if it means more complexity, and abstraction favors simplicity even if it means less accuracy. Advances on these two fronts can in some respects be made independently, with inference searching for details to add to make a model more predictive (a common perspective in biology and machine learning) and abstraction searching for ways to throw details away (a common perspective in information theory and statistical physics). Yet the two perspectives are not always clearly separable. We often have to find the right abstract level to do inference at all, as in regularized models, and we may be able to describe logic without having to infer lower-level mechanisms. Ultimately, it is in the tension of combining these two perspectives that science progresses.

5 Acknowledgments

The ideas presented here have been influenced by numerous fruitful discussions with the C4 Collective Computation Group, particularly Jessica Flack, David Krakauer, Chris Ellison, and Eddie Lee. Thanks to Ken Aiello for helpful comments on an earlier draft.

References

Alon, U. (2017). "Network Motifs: Theory and Experimental Approaches." *National Reviews Genetics*, 8(6): 450–461.

Amit, D. J., H. Gutfreund, and H. Sompolinsky. (1985). "Spin-Glass Models of Neural Networks." *Physical Review A*, 32(2): 1007–1018.

Anderson, P. (1972). "More Is Different." *Science,* 177.4047 (1972), 393–396.

Arehart, E., T. Jin, and B. C. Daniels. (2018). "Locating Decision-Making Circuits in a Heterogeneous Neural Network." *Frontiers in Applied Mathematics and Statistics*, 4: 11.

Babtie, A. C., and M. P. H. Stumpf. (2017). "How to Deal with Parameters for Whole-Cell Modelling." *Journal of the Royal Society Interface*, 14(133): 20170237.

Barzel, B., Y.-Y. Liu, and A.-L. Barabási. (2015). "Constructing Minimal Models for Complex System Dynamics." *Nature Communications*, 6: 7186.

Bassett, D. S., et al. (2011). "Dynamic Reconfiguration of Human Brain Networks during Learning." Proceedings of the National Academy of Sciences, 108(18): 7641–7646.

Bassett, D. S., et al. (2013). "Robust Detection of Dynamic Community Structure in Networks." *Chaos*, 23: 013142.

Beer, R. D. (2014). "Dynamical Systems and Embedded Cognition." In K. Frankish and W. Ramsey (eds.), *The Cambridge Handbook of Artificial Intelligence*, (pp. 128–148). Cambridge: Cambridge University Press, 2014.

Berman, G. J., W. Bialek, and J.W. Shaevitz. (2016). "Hierarchy and Predictability in *Drosophila* Behavior." *Proceedings of the National Academy of Sciences*, 113(42): 11943.

Bialek, W., et al. (2014). "Social Interactions Dominate Speed Control in Poising Natural Flocks near Criticality." *Proceedings of the National Academy of Sciences*, 111: 7212–7217.

Bonneau, R., et al. (2006). "The Inferelator: An Algorithm for Learning Parsimonious Regulatory Networks from Systems-Biology Data Sets De Novo." *Genome Biology*, 7(5): 1.

Box, G. E. P. (1979). "Robustness in the Strategy of Scientific Model Building." Army Research Office Workshop on Robustness in Statistics, pp. 201–236.

Brender, N. M. (2012). "Sense-Making and Symmetry-Breaking: Merleau-Ponty, Cognitive Science, and Dynamic Systems Theory." *Symposium: Canadian Journal of Continental Philosophy*, 17(2): 246–270.

Brunton, S. L., J. L. Proctor, and J. N. Kutz. (2016). "Discovering Governing Equations from Data by Sparse Identification of Nonlinear Dynamical Systems." *Proceedings of the National Academy of Sciences*, 113(15): 3932–3937.

Cocchi, L., L. L. Gollo, A. Zalesky, and M. Breakspear. (2017). "Criticality in the Brain." *Progress in Neurobiology*, 158: 132–152.

Cocco, S., and R. Monasson. (2012). "Adaptive Cluster Expansion for the Inverse Ising Problem: Convergence, Algorithm and Tests." *Journal of Statistical Physics*, 147: 252.

Couzin, I. D. (2009). "Collective Cognition in Animal Groups." *Trends in Cognitive Sciences*, 13(1): 36–43.

Crutchfield, J. P., and M. Mitchell. (1995). "The Evolution of Emergent Computation." *Proceedings of the National Academy of Science*, 92: 10742–10746.

Daniels, B. C., et al. (2018). "Criticality Distinguishes the Ensemble of Biological Regulatory Networks." *Physical Review Letters*, 121(13): 138102.

Daniels, B. C., and I. Nemenman. (2015). "Automated Adaptive Inference of Phenomenological Dynamical Models." *Nature Communications*, 6: 8133.

Daniels, B. C., C. J. Ellison, D. C. Krakauer, and J. C. Flack. (2016). "Quantifying Collectivity." *Current Opinion in Neurobiology*, 37: 106–113.

Daniels, B. C., D. C. Krakauer, and J. C. Flack. (2017). "Control of Finite Critical Behavior in a Small-Scale Social System." *Nature Communications*, 8: 14301.

Daniels, B. C., D. C. Krakauer, and J. C. Flack. (in preparation). "Distance from Criticality in Adaptive Collective Behavior."

Daniels, B. C., D. C. Krakauer, and J. C. Flack. (2012). "Sparse Code of Conflict in a Primate Society." *Proceedings of the National Academy of Sciences*, 109(35): 14259.

Daniels, B. C., et al. (2008). "Sloppiness, Robustness, and Evolvability in Systems Biology." *Current Opinion in Biotechnology*, 19(4): 389–395.

Daniels, B. C., J. C. Flack, and D. C. Krakauer. (2017). "Dual Coding Theory Explains Biphasic Collective Computation in Neural Decision-Making." *Frontiers in Neuroscience*, 11: 313.

Daniels, B. C., M. Dobrzynski, and D. Fey. (2018). "Parameter Estimation, Sloppiness, and Model Identifiability." In B. Munsky, L. Tsimring, and W. Hlavacek (eds.), *Quantitative Biology: Theory, Computational Methods, and Models*. Cambridge, MA: MIT Press, 271.

Daniels, B. C., W. S. Ryu, and I. Nemenman. (2019). "Automated, Predictive, and Interpretable Inference of *Caenorhabditis Elegans* Escape Dynamics." *Proceedings of the National Academy of Sciences*, 116(15), 7226–7231.

Davies, P. C. W., and S. I. Walker. (2016). "The Hidden Simplicity of Biology." *Reports on Progress in Physics*, 79: 102601.

Denker, J., et al. (1987). "Large Automatic Learning, Rule Extraction and Generalization." *Complex Systems*, 1: 877–922.

Dennett, D. (2014). "The Software/Wetware Distinction: Comment on 'Unifying Approaches From Cognitive Neuroscience And Comparative Cognition' by W. Tecumseh Fitch." *Physics of Life Reviews*, 11: 367–368.

Feret, J., et al. (2009). "Internal Coarse-Graining of Molecular Systems." *Proceedings of the National Academy of Sciences*, 106(16): 6453–6458.

Flack, J. (2017). "Life's Information Hierarchy." In S. I. Walker, P. C. W. Davies, and G. F. R. Ellis (eds.), *From Matter to Life: Information and Causality*. Cambridge: Cambridge University Press, 283.

Flack, J. C., and D. C. Krakauer. (2011). "Challenges for Complexity Measures: A Perspective from Social Dynamics and Collective Social Computation." *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 21(3): 037108.

Ganmor, E., R. Segev, and E. Schneidman. (2011). "Sparse Low-Order Interaction Network Underlies a Highly Correlated and Learnable Neural Population Code." *Proceedings of the National Academy of Sciences*, 108(23): 9679–9684.

Goldenfeld, N. (1992). *Lectures on Phase Transitions and the Renormalization Group*. New York: Westview Press.

Gunawardena, J. (2014). "Models in Biology: 'Accurate Descriptions of Our Pathetic Thinking'." *BMC Biology*, 12: 29.

Hoel, E. P., L. Albantakis, and G. Tononi. (2013). "Quantifying Causal Emergence Shows That Macro Can Beat Micro." *Proceedings of the National Academy of Sciences*, 110(49): 19790–19795.

Jia, D., M. K. Jolly, and H. Levine. (2018). Uses of Bifurcation Analysis in Understanding Cellular Decision-Making." In B. Munsky, L. Tsimring, and W. Hlavacek (eds.), *Quantitative Biology: Theory, Computational Methods, and Models*. Cambridge, MA: MIT Press, 357.

Kastner, D. B., S. A. Baccus, and T. O. Sharpee. (2015). "Critical and Maximally Informative Encoding between Neural Populations in the Retina." *Proceedings of the National Academy of Science*, 112: 2533–2538.

Kauffman, S. (1969). "Metabolic Stability and Epigenesis In Randomly Constructed Genetic Nets." *Journal of Theoretical Biology*, 22(3): 437–467.

Kemp, C., and J. B. Tenenbaum. (2008). "The Discovery of Structural Form." *Proceedings of the National Academy of Sciences*, 105(31): 10687–10692.

Krakauer, D. C., et al. (2011). "The Challenges and Scope of Theoretical Biology." *Journal of Theoretical Biology*, 276(1): 269–276.

Landauer, R. (1961, July). "Irreversibility and Heat Generation in the Computational Process." *IBM Journal of Research and Development*, 5: 183–191.

Lander, A. D. (2010). "The Edges of Understanding." *BMC Biology* 8: 40.

Lang, A. H., H. Li, J. J. Collins, and P. Mehta. (2014). "Epigenetic Landscapes Explain Partially Reprogrammed Cells and Identify Key Reprogramming Genes." *PLoS Computational Biology*, 10(8):

Lee, E. D., and B. C. Daniels. (2019). "Convenient Interface to Inverse Ising (ConIII): A Python 3 Package for Solving Ising-Type Maximum Entropy Models." *Journal of Open Research Software*, 7: 3.

Lukoševicius, M., and H. Jaeger. (2009). "Reservoir Computing Approaches to Recurrent Neural Network Training." *Computer Science Review*, 3(3): 127–149.

Machta, B. B., R. Chachra, M. K. Transtrum, and J. P. Sethna. (2013). "Parameter Space Compression Underlies Emergent Theories and Predictive Models." *Science*, 342 (6158): 604–607.

Marr, D. C., and T. Poggio. (1976). "From Understanding Computation to Understanding Neural Circuitry." Massachusetts Institute of Technology Artificial Intelligence Laboratory A.I. Memo 357.

Marzen, S. E., and J. P. Crutchfield. (2017). "Nearly Maximally Predictive Features and Their Dimensions." *Physical Review*, E 95 (2017), 051301(R).

Mehta, P., and D. J. Schwab. (2014). "An Exact Mapping between the Variational Renormalization Group and Deep Learning." arXiv preprint 1410.3831.

Merchan, L., and I. Nemenman. (2016). "On the Sufficiency of Pairwise Interactions in Maximum Entropy Models of Networks." *Journal of Statistical Physics*, 162: 1294–1308.

Mitchell, M. (1998). "A Complex-Systems Perspective on the 'Computation vs. Dynamics' Debate in Cognitive Science." *Proceedings of the Twentieth Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Lawrence Erlbaum Associates, 710.

Mojtahedi, M., et al. (2016). "Cell Fate Decision as High-Dimensional Critical State Transition." *PLoS Biology*, 14(12): (2016), e2000640.

Mora, T., and W. Bialek. (2011). "Are Biological Systems Poised at Criticality?" *Journal of Statistical Physics*, 144 (2011), 268–302.

Natale, J. L., D. Hofmann, D. G. Hernández, and I. Nemenman. (2018). "Reverse-Engineering Biological Networks from Large Data Sets." In B. Munsky, W. Hlavacek, and L. Tsimring (eds,), *Quantitative Biology: Theory, Computational Methods, and Models*. Cambridge, MA: MIT Press, 213.

Nemenman, I. (2012). "Information Theory and Adaptation." In M. E. Wall (ed.), *Quantitative Biology: From Molecular to Cellular Systems*, (Chapter 5). Boca Raton, FL: Taylor and Francis.

Newman, M. (2010). *Networks: An Introduction*. New York: Oxford University Press, 2010.

Palis, J., Jr., and W. de Melo. (1982). *Geometric Theory of Dynamical Systems*. Berlin: Springer-Verlag.

Payne, J. L., and A. Wagner. (2015). "Function Does not Follow Form in Gene Regulatory Circuits." Scientific Reports 5, 13015.

Peter, I. S., and E. H. Davidson. (2017). "Assessing Regulatory Information in Developmental Gene Regulatory Networks." *Proceedings of the National Academy of Sciences*, 114(23): 5862–5869.

Priesemann, V., M. Valderrama, M. Wibral, and M. Le Van Quyen. (2013). "Neuronal Avalanches Differ from Wakefulness to Deep Sleep." *PLoS Computational Biology*, 9(3): e1002985.

Prokopenko, M., J. T. Lizier, O. Obst, and X. R. Wang. (2011). "Relating Fisher Information to Order Parameters." *Physical Review E*, 84(4): 41116.

Rosenthal, S. B., et al. (2015). "Revealing the Hidden Networks of Interaction in Mobile Animal Groups Allows Prediction of Complex Behavioral Contagion." *Proceedings of the National Academy of Sciences*, 112(15): 4690–4695.

Roudi, Y., S. Nirenberg, and P. E. Latham. (2009). "Pairwise Maximum Entropy Models for Studying Large Biological Systems: When They Can Work and When They Can't." *PLoS Computational Biology*, 5(5): e1000380.

Rumelhart, D. E., and J. L. McClelland. (1986): *Parallel Distributed Processing*. Vol. 1. Cambridge, MA: MIT Press.

Schmidt, M., and H. Lipson. (2009). "Distilling Free-Form Natural Laws from Experimental Data." *Science*, 324(5923): 81–85.

Schneidman, E., M. J. Berry II, R. Segev, and W. Bialek. (2006). "Weak Pairwise Correlations Imply Strongly Correlated Network States in a Neural Population." *Nature*, 440: 1007.

Schwab, D. J., I. Nemenman, and P. Mehta. (2014). "Zipf's Law and Criticality in Multivariate Data without Fine-Tuning." *Physical Review Letters*, 113(6): 068102.

Segal, E., et al. (2003). "Module Networks: Identifying Regulatory Modules and Their Condition-Specific Regulators from Gene Expression Data." *Nature Genetics*, 34(2): 166–176.

Sethna, J. (2006). *Entropy, Order Parameters, and Complexity*. New York: Oxford University Press.

Seung, H., H. Sompolinsky, and N. Tishby. (1992). "Statistical Mechanics of Learning from Examples." *Physical Review A*, 45(8): 6056.

Shalizi, C. R. and C. Moore. (2003). "What Is a Macrostate? Subjective Observations and Objective Dynamics." arXiv preprint cond-mat/0303625.

Shalizi, C. R., and J. P. Crutchfield. (2001). "Computational Mechanics: Pattern, Prediction Strucutre and Simplicity." *Journal of Statistical Physics*, 104: 817–879.

Sharpee, T. O. (2017). "Optimizing Neural Information Capacity through Discretization." *Neuron*, 94(5): 954–960.

Slonim, N., G. S. Atwal, G. Tkacik, and W. Bialek. (2005). "Information-Based Clustering." *Proceedings of the National Academy of Sciences*, 102(51): 18297–18302.

Smirnakis, S. M., et al. (1997). "Adaptation of Retinal Processing to Image Contrast and Spatial Scale." *Nature*, 386(6620): 69–73.

Solé, R., et al. (2016). "Synthetic Collective Intelligence." *BioSystems*, 148: 47–61.

Sosna, M. M. G., et al. (2019). "Individual and collective encoding of risk in animal groups." *Proceedings of the National Academy of Sciences*, 116(41): 20556–20561.

Stephens, G. J., L. C. Osborne, and W. Bialek. (2011). "Searching for Simplicity in the Analysis of Neurons and Behavior." *Proceedings of the National Academy of Sciences*, 108: 15565–15571.

Still, S. (2014). "Information Bottleneck Approach to Predictive Inference." *Entropy*, 16(2): 968–989.

Strogatz, S. H. (1994). *Nonlinear Dynamics and Chaos*. New York: Perseus.

Tomlin, J. C., and J. D. Axelrod. (2005), "Understanding Biology by Reverse Engineering the Control." *Proceedings of the National Academy of Sciences*, 102(8): 4219–4220.

Transtrum, M. K., and P. Qiu. (2016). "Bridging Mechanistic and Phenomenological Models of Complex Biological Systems." *PLoS Computational Biology*, 12(5): 1–34.

Transtrum, M. K., et al. (2015). "Sloppiness and Emergent Theories in Physics, Biology, and Beyond." *Journal of Chemical Physics*, 143(1): 010901.

Tunstrom, K., et al. (2013). "Collective States, Multistability, and Transitional Behavior In Schooling Fish." *PLoS Computational Biology*, 9: e1002915.

Wilson, K. G. (1979). "Problems in Physics with Many Scales of Length." *Scientific American*, 241(2): 158–179.

Wolpert, D. H., J. A. Groschow, E. Libby, and S. DeDeo. (2015). "Optimal High-Level Descriptions Of Dynamical Systems." arXiv preprint 1409.7403v2.