RESEARCH ARTICLE

Infinitely large, randomly wired sensors cannot predict their input unless they are close to deterministic

Sarah Marzen*

Department of Physics, Physics of Living Systems Group, Massachusetts Institute of Technology, Cambridge, MA, United States of America

* semarzen@mit.edu

Citation: Marzen S (2018) Infinitely large, randomly wired sensors cannot predict their input unless they are close to deterministic. PLoS ONE 13(8): e0202333. https://doi.org/10.1371/journal.pone.0202333

Editor: Bryan C Daniels, Arizona State University & Santa Fe Institute, UNITED STATES

Received: March 7, 2018

Accepted: August 1, 2018

Published: August 29, 2018

Data Availability Statement: All data is available through protocols.io at dx.doi.org/10.17504/protocols.io.p3kdqkw.

Funding: The author was funded by an MIT Physics of Living Systems Fellowship. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The author has declared that no competing interests exist.

Abstract

Building predictive sensors is of paramount importance in science. Can we make a randomly wired sensor “good enough” at predicting its input simply by making it larger? We show that infinitely large, randomly wired sensors are nonspecific for their input, and therefore nonpredictive of future input, unless they are close to deterministic. Nearly deterministic, randomly wired sensors can capture $\sim$ 10% of the predictive information of their inputs for “typical” environments.

Introduction

Prediction is thought to be fundamental to organism functioning, e.g. see Ref. [1] and references therein. By better predicting the world around them, organisms can better choose actions that will maximize their reaped reward. On the other hand, prediction of the future of a time series from its past drives much of science, e.g. the realization that one could predict future particle positions from some parameters (such as particle masses and charges) and current particle positions and velocities.

So how should one build predictive time series models? One common trick is to feed the data we wish to predict into a recurrent network, the state of which can contain enough memory of the past in order to be predictive of the future. Sometimes, these networks are trained so that the parameters defining the network’s dynamics yield optimal predictions of the input time series, e.g. as in Ref. [2] and references therein. However, much utility has been gained from using randomly connected networks (or reservoirs) [3–6], where one merely trains the readout of the network. Such networks can have nearly maximal predictive power if the networks are large enough [7]. Such a finding might also imply that perhaps, evolution need not work so hard to build predictive networks (or sensors) in organisms; rather, randomly wiring large biological sensors could lead to sufficiently good predictive performance.

Here, we identify a key sensor property without which the sensor has little predictive power: determinism. Determinism means that the present sensor state and the present input state uniquely determine the future sensor state. To be clear, the sensors studied in Refs. [3, 4, 6] are deterministic as long as there is no added noise, and so the effect of nondeterminism on predictive capabilities of recurrent networks/reservoirs is somewhat unstudied. Interestingly, we find that the detrimental effects of nondeterminism are compounded rather than mitigated by large sensor size when the sensor has recurrent connections.

We find numerical and analytical evidence that nondeterminism greatly limits the ability of a (recurrent) randomly wired sensor to be predictive of its inputs, due to the weak law of large numbers. For some nondeterministic, randomly wired sensors, there is a finite optimal sensor size at which predictive power is maximized. This optimal sensor size seems to balance the larger predictive capacity of larger sensors against a trend towards nonspecificity for input demanded by the weak law of large numbers.

When the sensors connections are nearly deterministic, then larger sensors are on average more predictive. Large, deterministic, randomly wired sensors can capture $\sim$ 10% of the total predictive information possible. This is comparable to the $\sim$ 20% of predictive information captured by sensors whose weights have been (locally) optimized via the BFGS algorithm.

Setup

First, we discuss the model of the environment, which is given by the output of a unifilar Hidden Markov model. Then, we discuss the model of the sensor, specifying notation for the dynamics of a conditionally Markovian discrete-valued sensor. Finally, we describe metrics for sensor performance: memory; and predictive information captured.

Throughout what follows, we characterize time series and the relations between them via entropy and mutual information. The entropy of a random variable X with realizations x and probability distribution $Pr(X = x)$ is given by

$$H[X] = -\sum_{x} Pr(X = x) \log Pr(X = x), \tag{1}$$

while the mutual or shared information between a random variable X and a random variable Y with realizations x and y, respectively, and joint probability distribution $Pr(X = x, Y = y)$ is given by

$$I[X; Y] = \sum_{x,y} Pr(X = x, Y = y) \log \frac{Pr(X = x, Y = y)}{Pr(X = x)Pr(Y = y)}. \tag{2}$$

One can think of entropy as a measure of the uncertainty of a random variable. Maximally uncertain random variables have a uniform distribution over values, and this maximizes entropy; minimally uncertain random variables are singly supported, and this minimizes entropy. One can think of mutual information as a measure of the nonlinear dependency of two random variables. From the identity $I[X;Y] = H[X] - H[X|Y]$, mutual information is the reduction in uncertainty about a random variable X that comes from knowing a potentially related random variable Y. Operational meanings of entropy and mutual information come from Shannon's source coding and noisy channel coding theorems [8, 9].

Model of environment

To test the predictive capabilities of sensors, we wish to construct a non-Markovian environment that is at least somewhat predictable. The non-Markovianity of the environment will force a predictive sensor to remember longer and longer pasts, while the predictability of the input will guarantee that remembering the "right" things about environmental pasts leads to measurable predictive gains. Ideally, it would also be difficult to infer the right predictive features, so that the environments provide a challenge for sensors that desire to be predictive of their input. Environments generated by random minimal unifilar Hidden Markov models (or ϵMs [10]) as specified below turn out to satisfy all these requirements.

We consider randomly generated binary-alphabet environments, that is, environments generated by a randomly-drawn unifilar Hidden Markov model. Here, $\mathcal{S}_t$ is the random variable for the hidden state at time t, while X_t is the random variable for the observed symbol at time t. An edge-emitting discrete-time Hidden Markov model is specified by a set of internal "hidden" states $\sigma \in \mathbf{S}$, a set of possible observables $x \in \mathcal{X}$, and a labeled transition dynamic $Pr(\mathcal{S}_{t+1} = \sigma', X_t = x | \mathcal{S}_t = \sigma)$. For the purposes of this paper, $\mathcal{X} = \{0, 1\}$. When $\mathbf{S}$ is countable, we can represent the labeled transition dynamic by a set of labeled transition matrices $T^{(x)}$ with elements

$$T^{(x)}_{\sigma',\sigma} := Pr(\mathcal{S}_{t+1} = \sigma', X_t = x | \mathcal{S}_t = \sigma). \tag{3}$$

A unifilar Hidden Markov model is one such that $Pr(\mathcal{S}_{t+1} | X_t = x, \mathcal{S}_t = \sigma)$ is singly supported, i.e. one for which the current state and symbol emitted uniquely specify the next state. This is a version of determinism in that the next state is uniquely determined, but the next symbol is not uniquely determined.

To randomly generate environments, we randomly generate labeled transition matrices $T^{(x)}$ as follows. In each hidden state σ, we choose $\sum_{\sigma'} T^{(0)}_{\sigma',\sigma}$ (the probability of emitting a 0) uniformly at random from the unit interval; this specifies, for binary-alphabet processes, $\sum_{\sigma'} T^{(1)}_{\sigma',\sigma}$ (the probability of emitting a 1). We then randomly choose the state to which one transitions after seeing a 0 and the state to which one transitions after seeing a 1. The fact that there is only one state to which one transitions after seeing a 0 or a 1 implies unifilarity of the resulting Hidden Markov model. The states of a minimal unifilar Hidden Markov model are called causal states [10].

We wish to understand how much memory is required to predict the output of these Hidden Markov models as well as possible and their predictability. Let $\overleftarrow{X}_t$ stand for the past environmental inputs, . . ., X_{t-2}, X_{t-1}, X_t. The memory required is characterized by statistical complexity $C_\mu = H[\mathcal{S}_t]$ [10], while the predictability is characterized by the total correlation rate $\rho_\mu = I[\overleftarrow{X}_t; X_{t+1}]$ [11] also known as a particular value of the predictive information [12, 13].

As the underlying model grows larger, ρ_μ seems to tend in probability to $\sim$ 0.2 nats. These unifilar Hidden Markov models generate infinite-order Markov processes– that is, a process for which the next symbol depends to some extent on all previous symbols. However, a process that is technically infinite-order Markov can still be approximately Markovian [14]. A sensor which perfectly stores the present observed symbol and nothing else (a Markov model) would capture a predictive information of $I[X_t;X_{t+1}]$. A typical value of $I[X_t;X_{t+1}]$ for these environments is 0.002 nats when the environment is of size $|\mathbf{S}| = 30$, and so a typical value of $I[X_t;X_{t+1}]/\rho_\mu$ for these environments is 0.01. In other words, these environments tend to be strongly non-Markovian. And finally, the statistical complexity C_μ of these environments tends to be between $2 - 3$ nats, indicating that these environments provide a predictive challenge for sensors.

Model of sensor

Fix a realization of the environment, x_1, x_2, ... Let R_t be the random variable denoting the sensor's state at time t. We consider conditionally Markovian sensors with a finite number of states and with state space s whose probability evolves according to

$$Pr(R_{t+1} = r') = \sum_r Pr(R_{t+1} = r' | X_t = x_t, R_t = r) Pr(R_t = r), \tag{4}$$

which we represent in matrix-vector notation as

$$p(r_{t+1}) = M^{(x_t)} p(r_t). \tag{5}$$

Given its relation to the conditional probability distribution, $M^{(x_t)}$ is a transition matrix whose columns must sum to 1. See Fig 1, in which the arrows between sensor states indicate transition probabilities and in which the arrow from input to sensor states indicates a dependence of transition probabilities on input.

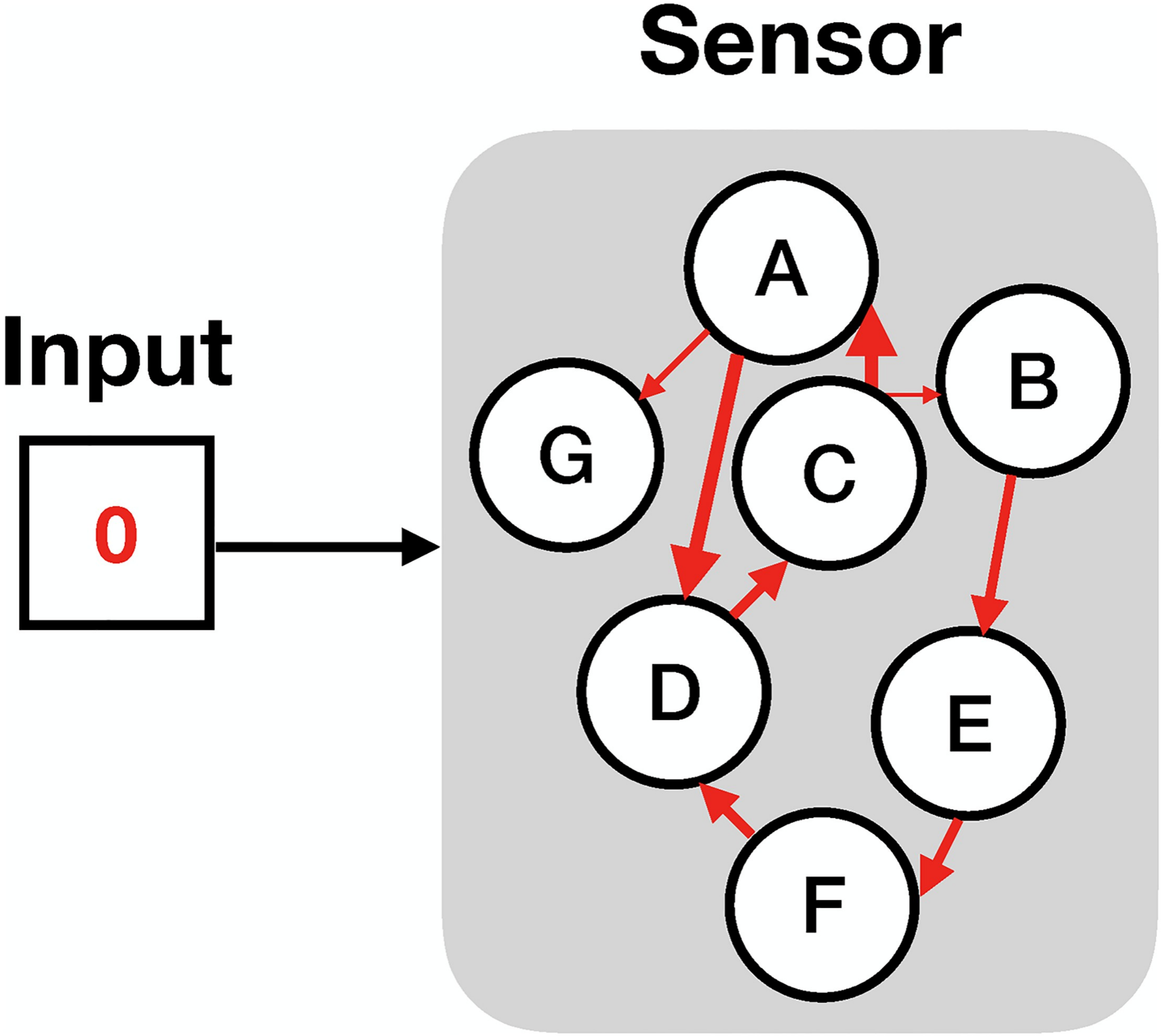

Fig 1. The model of the sensor. The input here is binary– either 0 or 1– and the sensor states comprise the set {A, B, C, D, E, F, G}. Different environmental inputs trigger different transition probabilities between the states of the sensor. These transition probabilities can be large (fat arrows) or small (thin arrows). The sensor shown here is nearly deterministic given an input of 0, in that each sensor state only transitions to either one or two other sensor states.

https://doi.org/10.1371/journal.pone.0202333.g001

The present sensor state depends on both the previous environmental symbol and the previous sensor state, and the recursion leads to the present sensor state depending on arbitrarily long environmental pasts. By construction, the sensor only understands the future of the environment through the past of the environment. In other words, the Markov relation $R_t \rightarrow \overleftarrow{X}_t \rightarrow X_{t+1}$ holds [9], meaning that $Pr(X_{t+1} = x, R_t = r|\overleftarrow{X}_t = \overleftarrow{x}) = Pr(X_{t+1} = x|\overleftarrow{X}_t = \overleftarrow{x})Pr(R_t = r|\overleftarrow{X}_t = \overleftarrow{x})$.

Sensor metrics: Memory and predictive information

We define memory I_{mem} as

$$I_{mem} := I[R_t; \mathcal{S}_t], \tag{6}$$

which is an achievable minimal coding cost of pasts– that is, the minimal amount of space needed to write down information about said pasts– to retain information about the future [15]. In terms of the steady-state probability distribution over input causal states and sensor states $p_{ss}(r, \sigma)$, we have

$$I_{mem} = \sum_{r,\sigma} p_{ss}(r,\sigma) \log \frac{p_{ss}(r,\sigma)}{p_{ss}(r)p_{ss}(\sigma)}, \tag{7}$$

where $p_{ss}(r) = \sum_\sigma p_{ss}(r, \sigma)$ and $p_{ss}(\sigma) = \sum_r p_{ss}(r, \sigma)$. The memory I_{mem} is 0 only when $p_{ss}(r, \sigma) = p_{ss}(r)p_{ss}(\sigma)$– that is, when the sensor is nonspecific for the predictive features of its input. Note, from a standard information theory identity $I[X;Y] \leq H[X]$, that

$$I_{mem} \leq C_\mu, \tag{8}$$

and so one's memory is upper-bounded by the statistical complexity, which is calculable from $p_{ss}(\sigma)$.

We define instantaneous predictive information (which we call predictive information for brevity) I_{pred} as

$$I_{pred} := I[R_t; X_{t+1}]. \tag{9}$$

This predictive information corresponds (under some assumptions) to the increase in expected log growth rate of an asexually reproducing population attainable with a given memory, and is always an upper bound on the increase in expected log growth rate attainable with a given memory [16]. In terms of the steady-state distribution $p_{ss}(r, x)$ over sensor states and future inputs, we have

$$I_{pred} = \sum_{r,x} p_{ss}(r,x) \log \frac{p_{ss}(r,x)}{p_{ss}(r)p_{ss}(x)}. \tag{10}$$

Earlier, causality gave us $R_t \rightarrow \overleftarrow{X}_t \rightarrow X_{t+1}$; and as the hidden states of a unifilar Hidden Markov model are minimal sufficient statistics of prediction [10], we also have $\overleftarrow{X}_t \rightarrow \mathcal{S}_t \rightarrow X_{t+1}$. Together, this gives $R_t \rightarrow \mathcal{S}_t \rightarrow X_{t+1}$, and the Data Processing Inequality [9] therefore reveals

$$I_{pred} \leq I_{mem}. \tag{11}$$

That is, the predictive information captured is always less than one's memory. Another application of the Data Processing Inequality reveals

$$I_{pred} \leq \rho_\mu, \tag{12}$$

and so the total correlation rate is an achievable upper bound on the predictive information. To achieve this upper bound, one needs the sensor states to uniquely determine all causal states σ.

We also define, for reasons that become apparent later on,

$$I'_{mem} := D_{KL}\left[p_{ss}(r,\sigma)||\frac{1}{N}p_{ss}(\sigma)\right]. \tag{13}$$

This is a measure of the deviation between the steady state distribution over sensor states and input causal states, $p_{ss}(r, \sigma)$, from the distribution over sensor states and input causal states were the sensor to be nonspecific for its input *and* were each sensor state to be equally likely. If the sensor is nonspecific for its input but if the distribution over sensor states is nonuniform, then I'_{mem} will be nonzero. As discussed in Section B in S1 File,

$$I_{mem} \leq I'_{mem}. \tag{14}$$

As a result, if we can argue that I'_{mem} tends to 0, we will also have argued that I_{mem} and thus I_{pred} tend to 0.

Methods

We wish to calculate the aforementioned sensor metrics. One could simulate sequences of σ_t, r_t, and x_{t+1} by using $T^{(x)}$, $M^{(x)}$ to randomly choose future sensor and environmental states, and then count the number of occurrences of a particular combination of σ_t, r_t, x_{t+1}. In the case of a large sensor state space, one would ideally use an entropy estimator such as the NSB entropy estimator [17] to calculate the predictive information, so as to avoid prohibitively long simulations.

However, we pursue a different approach that leads to an easier calculation of sensor metrics and to plausibility arguments that underscore the generality of our results. To calculate I_{mem}, we wish to find $p(r_t, \sigma_t)$. To calculate instantaneous predictive information I_{pred}, we wish to find $p(r_t, x_{t+1})$. We can get the latter probability distribution, $p(r_t, x_{t+1})$, from the former probability distribution, $p(r_t, \sigma_t)$, by exploiting the Markov chain $R_t \rightarrow \mathcal{S}_t \rightarrow X_{t+1}$:

$$p(r_t, x_{t+1}) = \sum_{\sigma_t} p(r_t, \sigma_t, x_{t+1}) \tag{15}$$

$$= \sum_{\sigma_t} p(r_t, \sigma_t) p(x_{t+1}|\sigma_t). \tag{16}$$

To find $p(r_t, \sigma_t)$, we can set up a Chapman-Kolmogorov equation:

$$p(r_{t+1}, \sigma_{t+1}) = \sum_{r_t,\sigma_t,x_t} p(r_{t+1}, \sigma_{t+1}|r_t, \sigma_t, x_t) \times p(r_t, \sigma_t, x_t) \tag{17}$$

$$= \sum_{r_t,\sigma_t,x_t} p(r_{t+1}|x_t, r_t) p(\sigma_{t+1}|x_t, \sigma_t) p(x_t|\sigma_t) \times p(r_t, \sigma_t) \tag{18}$$

$$= \sum_{r_t,\sigma_t,x_t} M^{(x_t)}_{r_{t+1},r_t} \frac{T^{(x_t)}_{\sigma_{t+1},\sigma_t}}{1^\top T^{(x_t)}_{:,\sigma_t}} \left(1^\top T^{(x_t)}_{:,\sigma_t}\right) \times p(r_t,\sigma_t) \tag{19}$$

$$= \sum_{r_t,\sigma_t} \left(\sum_{x_t} M^{(x_t)}_{r_{t+1},r_t} T^{(x_t)}_{\sigma_{t+1},\sigma_t} \right) p(r_t,\sigma_t). \tag{20}$$

Eq 20 defines a vector with $|R||\mathbf{S}|$ elements and a corresponding transition matrix. The normalized eigenvector of eigenvalue 1 corresponds to the desired $p_{ss}(r,\sigma)$, and then one can calculate the instantaneous predictive information directly from $p(r_t, x_{t+1})$ via

$$I_{pred} = \sum_{r_t,x_{t+1}} p(r_t,x_{t+1}) \log \frac{p(r_t,x_{t+1})}{p(r_t)p(x_{t+1})}. \tag{21}$$

This serves as an alternative to calculating $p_{ss}(r,\sigma)$, and thus I_{mem} and I_{pred}, through simulation.

Results

There are two parameters that one can play with when designing our randomly wired sensors: sensor size, given by the number of sensor states N; and the method by which connections between sensor states are randomly generated.

Using a sensor of size N corresponds to clustering pasts into N clusters: an input past $\overleftarrow{x}$ leads to sensor state r with probability $p(r|\overleftarrow{x})$. When the sensor is not deterministic, these clusters are *soft clusters*– that is, pasts are assigned probabilistically to clusters. When sensor connections are closer to deterministic, these clusters "harden". From Section A in S1 File, we then expect the memory and predictive information captured by sensors to increase (but saturate) with increasing sensor size and increasing determinism.

In other words, specification of a randomly wired sensor corresponds to a random soft clustering of pasts into predictive features, though the mapping from the sensor specification to the clustering of pasts is nontrivial. Despite the presence of such a mapping, specification of a sensor has one important advantage over specification of a probabilistic clustering of pasts: it is a finite description of a potentially complicated clustering. More concretely, it is more economical to specify the optimal predictor (the ϵM [10]) by input-dependent state transitions, $M^{(x)}$, instead of the conditional probability distribution of hidden states given input histories, $p(r|\overleftarrow{x})$.

First, we consider fully nondeterministic, randomly wired sensors for which the columns of $M^{(x)}$ are independent and identically distributed (i.i.d) draws from a Dirichlet distribution with concentration parameter $\vec{\alpha} = \alpha\vec{1}$. We start with a strange numerical fact. As such sensors grow in size, when α is sufficiently large, both memory and predictive information tend to decrease. See Fig 2. In fact, as N grows, $p(r|\sigma)$ appears to tend to $\frac{1}{N}$ – that is, the sensor appears to become nonspecific for its input. In other words, the capacity of a sensor to remember its input increases as the sensor increases in size, but this capacity is not at all used by fully connected, randomly wired sensors.

We give a plausibility argument for this nonspecificity, which comes down to a statement about the eigenvector of eigenvalue 1 of the transition matrix between sets of causal states and sensor states, (σ,r), with elements $\sum_{x_t} M^{(x_t)}_{r_{t+1},r_t} T^{(x_t)}_{\sigma_{t+1},\sigma_t}$ as described in Methods. This argument is generalized and expanded upon in Section B in S1 File. There is a unique eigenvector of eigenvalue 1 for these transition matrices by the Perron-Frobenius theorem. We will argue that this

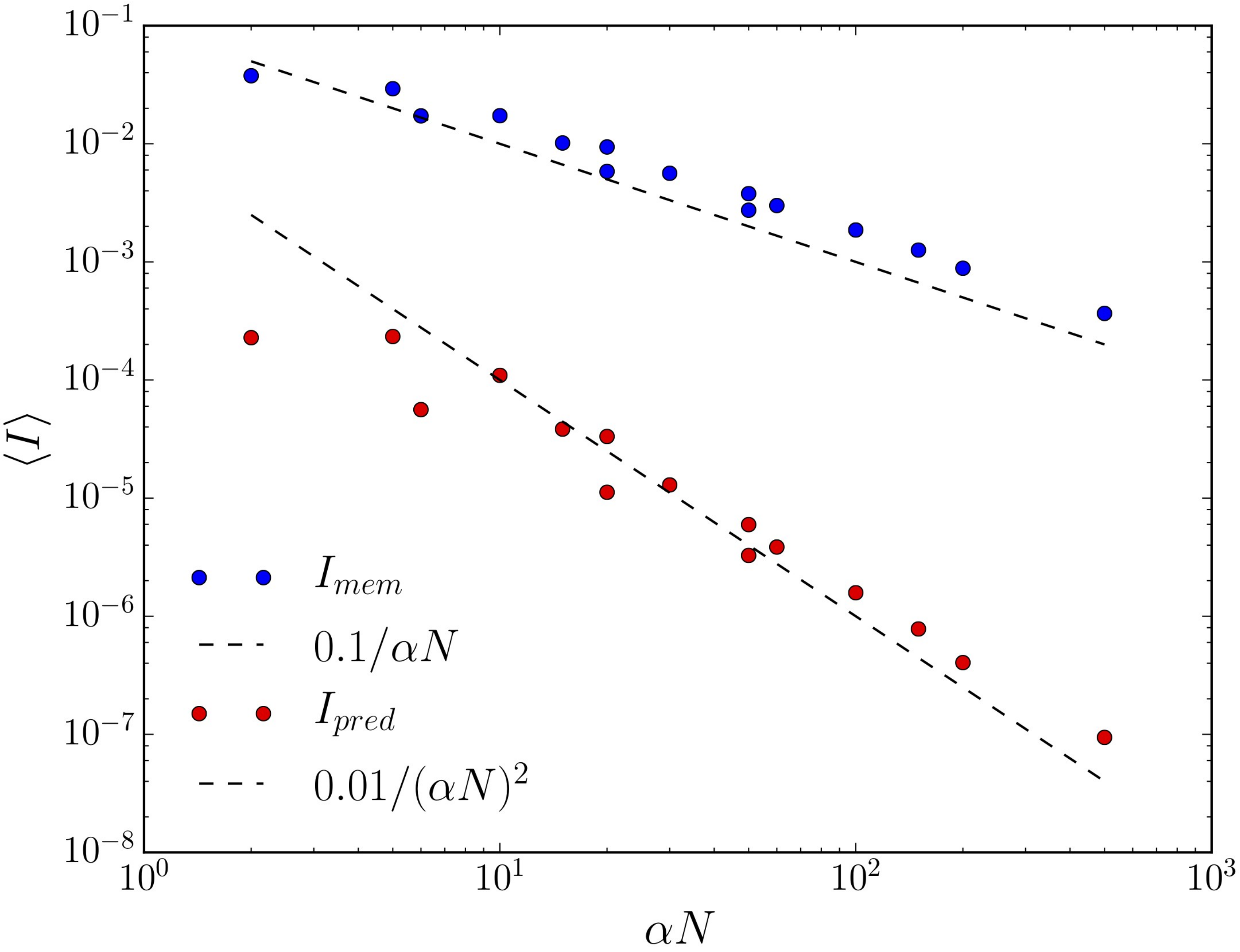

Fig 2. Large, fully nondeterministic, randomly wired sensors are nonspecific for their inputs. As described in the main text, the concentration parameter α controls the distribution from which transition probabilities in the sensor are drawn. We show $\langle I_{mem}\rangle$ (blue) and $\langle I_{pred}\rangle$ (red) for three values of α– 1, 3, and 10– as a function of αN. Both information quantities decrease at varying rates with N and α: roughly $0.1/\alpha N$ for $\langle I_{mem}\rangle$, and roughly $0.01/(\alpha N)^2$ for $\langle I_{pred}\rangle$. Corresponding lines in black dashes are drawn to guide the eye. The environment has total correlation rate $\rho_\mu = 0.198$ nats and statistical complexity $C_\mu = 2.205$ nats, so the total memory and predictive information captured by the sensor is maximally three orders of magnitude smaller than the total possible memory and predictive information captured.

https://doi.org/10.1371/journal.pone.0202333.g002

eigenvector is given in the large N limit by $p(\sigma, r) = \frac{1}{N} p_{ss}(\sigma)$ where $p_{ss}(\sigma) = \text{eig}_1(\sum_x T^{(x)})_\sigma$. If $p(\sigma, r) = \frac{1}{N} p_{ss}(\sigma)$, then

$$\frac{1}{N} p_{ss}(\sigma_{t+1}) \quad = \quad \sum_{\sigma_t, x_t} \left(\frac{1}{N} \sum_{r_t} M^{(x_t)}_{r_{t+1}, r_t} \right) T^{(x_t)}_{\sigma_{t+1}, \sigma_t} p_{ss}(\sigma_t).$$

To understand whether or not this is plausible, we focus on simplifying the right-hand side of this equation. Recall that a Dirichlet distribution in which the concentration parameter vector takes the form $\vec{\alpha} = \alpha\vec{1}$ can be generated by drawing realizations of identical and independently distributed (i.i.d.) Gamma random variables and normalizing them by their sum. Hence, we can write $M^{(x_t)}_{r_{t+1},r_t} = \frac{Y^{(x_t)}_{r_{t+1},r_t}}{\sum_{r'} Y^{(x_t)}_{r',r_t}}$ where $Y^{(x_t)}_{r',r_t}$ have been drawn i.i.d. from a distribution with probability density function $\frac{1}{\Gamma(\alpha)} y^{\alpha-1} e^{-y}$. Then, the sum $\sum_{r_t} M^{(x_t)}_{r_{t+1},r_t} = \sum_{r_t} \frac{Y^{(x_t)}_{r_{t+1},r_t}}{\sum_{r'} Y^{(x_t)}_{r',r_t}}$ is the ratio of two roughly independent Gamma distributions, both with shape parameter $N\alpha$ and scale 1. The ratio of two such Gamma distributions is highly peaked at 1 for large $N\alpha$. Thus $\sum_{r_t} M^{(x_t)}_{r_{t+1},r_t}$ tends to 1, which then implies that

$$\sum_{\sigma_t,x_t} \left(\frac{1}{N} \sum_{r_t} M^{(x_t)}_{r_{t+1},r_t} \right) T^{(x_t)}_{\sigma_{t+1},\sigma_t} p_{ss}(\sigma_t) \quad \rightarrow \quad \frac{1}{N} p_{ss}(\sigma_{t+1})$$

as desired, using $\sum_{\sigma_t} T^{(x_t)}_{\sigma_{t+1},\sigma_t} p_{ss}(\sigma_t) = p_{ss}(\sigma_{t+1})$. In other words, when $N\alpha \gg 1$, $p(r|\sigma) = \frac{1}{N}$ appears to be a reasonable guess for the steady state distribution $p_{ss}(r|\sigma)$. This, in turn, suggests that I_{mem} tends to 0 with probability 1, which would give $I_{pred} \rightarrow 0$ from $0 \leq I_{pred} \leq I_{mem}$. An additional argument given in Section B in S1 File suggests that I'_{mem} is $O\left(\frac{1}{N\alpha}\right)$, though we emphasize that this is a plausibility argument rather than a sketch of a proof.

In other words, infinitely large fully-connected randomly-wired sensors are nonspecific for their inputs, *no matter the input*. This is true even when concentration parameter $\vec{\alpha}$ defining the sensor stochasticity is input-dependent and state-dependent, as an extension of the plausibility argument given above holds as detailed in Section B in S1 File. For numerical evidence, see Fig 3.

More generally, we find that the memory and predictive information captured by random sensors is governed by two trends. Both trends can be described in terms of their effect on the clusters $p(r|\overleftarrow{x})$ that describe how specifically one can determine an input past $\overleftarrow{x}$ from a sensor state r and vice versa. According to the first trend, information captured tends to increase with the number of clusters, as detailed by a null model in Section A in S1 File. According to the second trend, the exponential explosion in possible paths between sensor states given any particular input past $\overleftarrow{x}$ yields increasing nonspecificity. For small enough α, as shown in Fig 4, the behavior of memory and predictive information captured by randomly wired sensors appears to be a balance of these two trends. However, the latter trend for fully nondeterministic sensors always wins, and so infinitely large, fully nondeterministic, randomly wired sensors are completely nonspecific for their inputs even though the sensor dynamics are input-dependent.

How might a large system with unavoidable randomness in its connections avoid the seemingly inevitable march towards nonspecificity for its inputs? After all, large sensors have a large capacity to predict, in principle, harnessed in Refs. [3, 4]. Is there no way in which randomness in wiring can be constrained so that this capacity can be harvested?

A clue is provided by consideration of the minimal optimal predictive sensor of the inputs [10]– the ϵM, of size $N = |\mathcal{R}| = |\mathbf{S}|$. This optimal sensor is constructed so that each r corresponds to a different causal state σ, with transitions $M^{(x)}_{r,r'} = Pr(\mathcal{S}_{t+1} = r'|X_t = x, \mathcal{S}_t = r)$. (When $Pr(X_t = x|\mathcal{S}_t = r)$ is zero, a particular input word is forbidden, and any $M^{(x)}_{r,r'}$ can be chosen.) Note that for this optimally predictive sensor, given a particular input x, a particular sensor state r can only transition to one other sensor state r'. Minimal optimal predictive sensors, therefore, have a great deal of structure: they are deterministic. Many transitions between sensor states are forbidden.

Indeed, sensors optimized so as to maximize predictive information using the L-BFGS algorithm are also nearly deterministic. These (locally) optimal sensors tend to make Markov

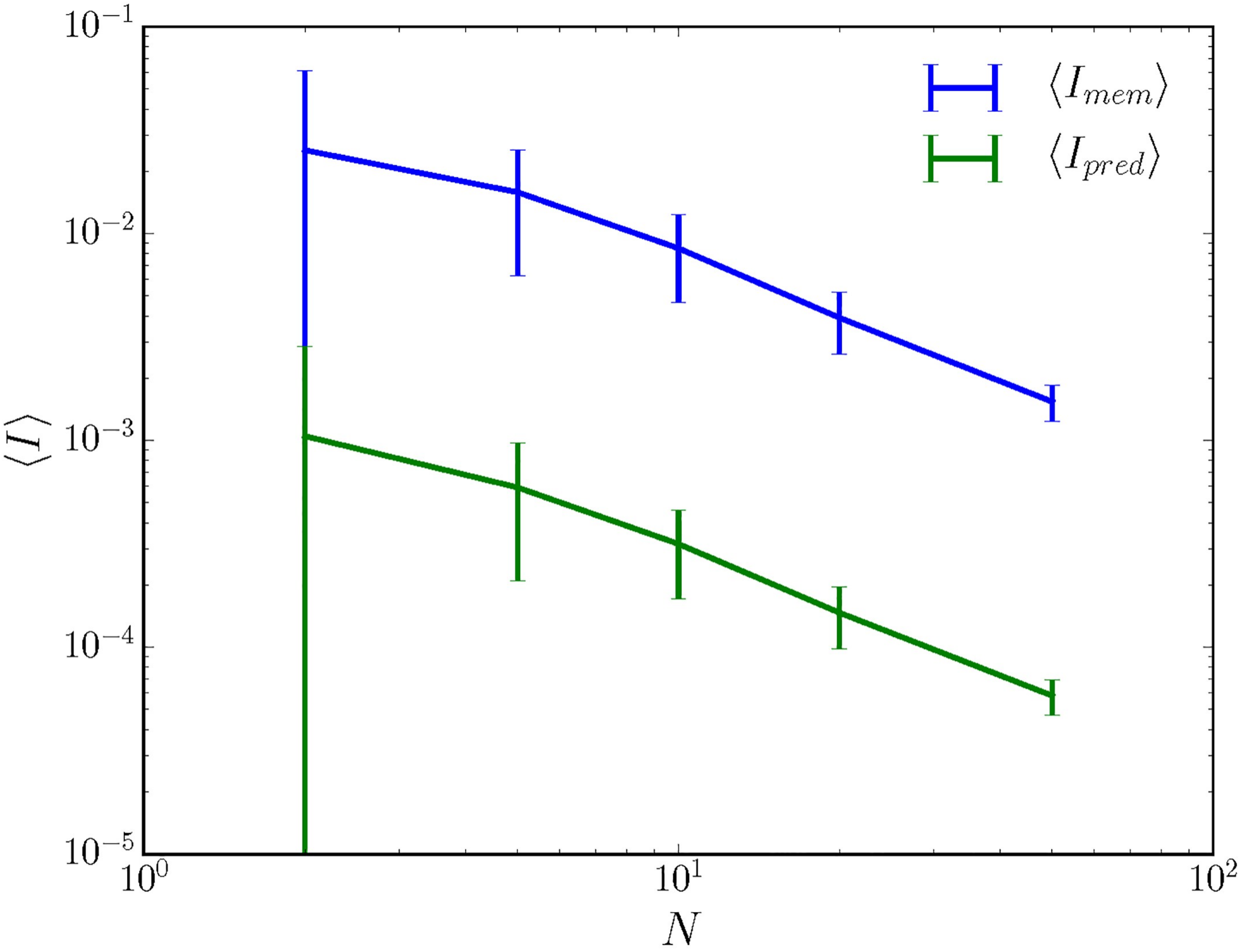

Fig 3. Large, nondeterministic, randomly wired sensors are nonspecific for their inputs even when the stochasticity is input- and state-dependent. As described in the main text, the concentration parameter $\vec{\alpha}$ describes the random wiring of the sensors. Here, $\alpha(x, r)$ is drawn uniformly at random in the interval [0, 10] for each x. Both average memory and average predictive information decrease with sensor size N. The environment again has $|\mathbf{S}| = 30$ but with $\rho_\mu = 0.19$ nats and $C_\mu = 3.1$ nats. The 50% confidence intervals in memory and predictive information decrease with increasing sensor size N, implying that the larger sensors are increasingly likely to be nonspecific for their input.

https://doi.org/10.1371/journal.pone.0202333.g003

models of the input, capturing 0.04 nats or 20% of the total predictive information capturable with $|\mathcal{R}| = 10$ states.

Perhaps unsurprisingly, then, large randomly wired sensors turn out to be specific for their inputs when the sensor is very close to deterministic. Then, $\sum_{r_t} M^{(x_t)}_{r_{t+1},r_t}$ will not be highly concentrated about 1, and the plausibility arguments given above for nonspecificity will fail. To illustrate this, we consider the case of a sensor in which $M^{(x)}_{r,r'}$ is, for each initial sensor state r', nonzero only for k randomly drawn sensor states r; and in which the probability distribution over the k nonzero $M^{(x)}_{r,r'}$ is Dirichlet with concentration parameter α. We see from Fig 5 that the greater the determinism (i.e. the smaller the k), the higher the average predictive

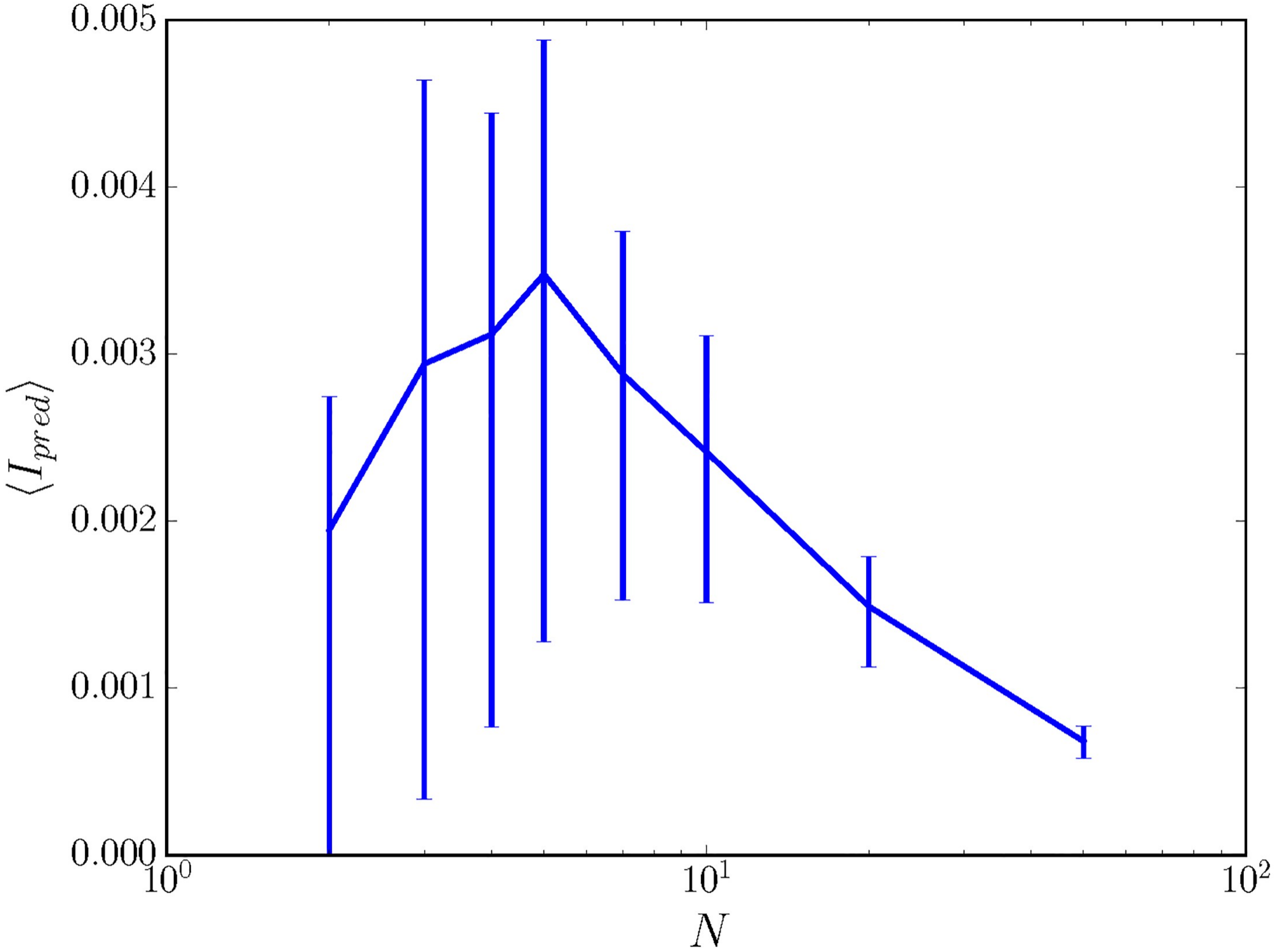

Fig 4. Predictive information can increase or decrease with sensor size. We show average predictive information $\langle I_{pred} \rangle$ as a function of N for $\alpha = 0.1$, where $\langle I_{pred} \rangle$ is estimated from 100 random draws of the sensor. The environment has $|\mathbf{S}| = 30$, and the non-monotonicity of $\langle I_{pred} \rangle$ with N for this value of α seems to hold regardless of environment. Error bars indicate 50% confidence intervals.

https://doi.org/10.1371/journal.pone.0202333.g004

information; and that, no matter the k, the predictive information captured actually increases with the size of the sensor. Large, nearly deterministic, randomly wired sensors have variable values of predictive information, unlike the large, fully nondeterministic, randomly wired sensors as discussed in Section C in S1 File. The average predictive information captured appears to saturate with increasing sensor size, and as such, the quality of the sensor grows with N but is fundamentally limited by k. The same trends hold for average memory $\langle I_{mem} \rangle$, shown in Fig 6.

Note, however, that if k grows with N such that $\lim_{N\to\infty} k$ is infinite, then the plausibility arguments given before will apply and the infinitely large, randomly wired sensor will be nonspecific for its input. Hence, the sparsity of sensor connections– the number of connections divided by the number of possible connections– must tend to 0 as the sensor size increases if an infinitely large randomly-wired sensor is to be at all specific for its input.

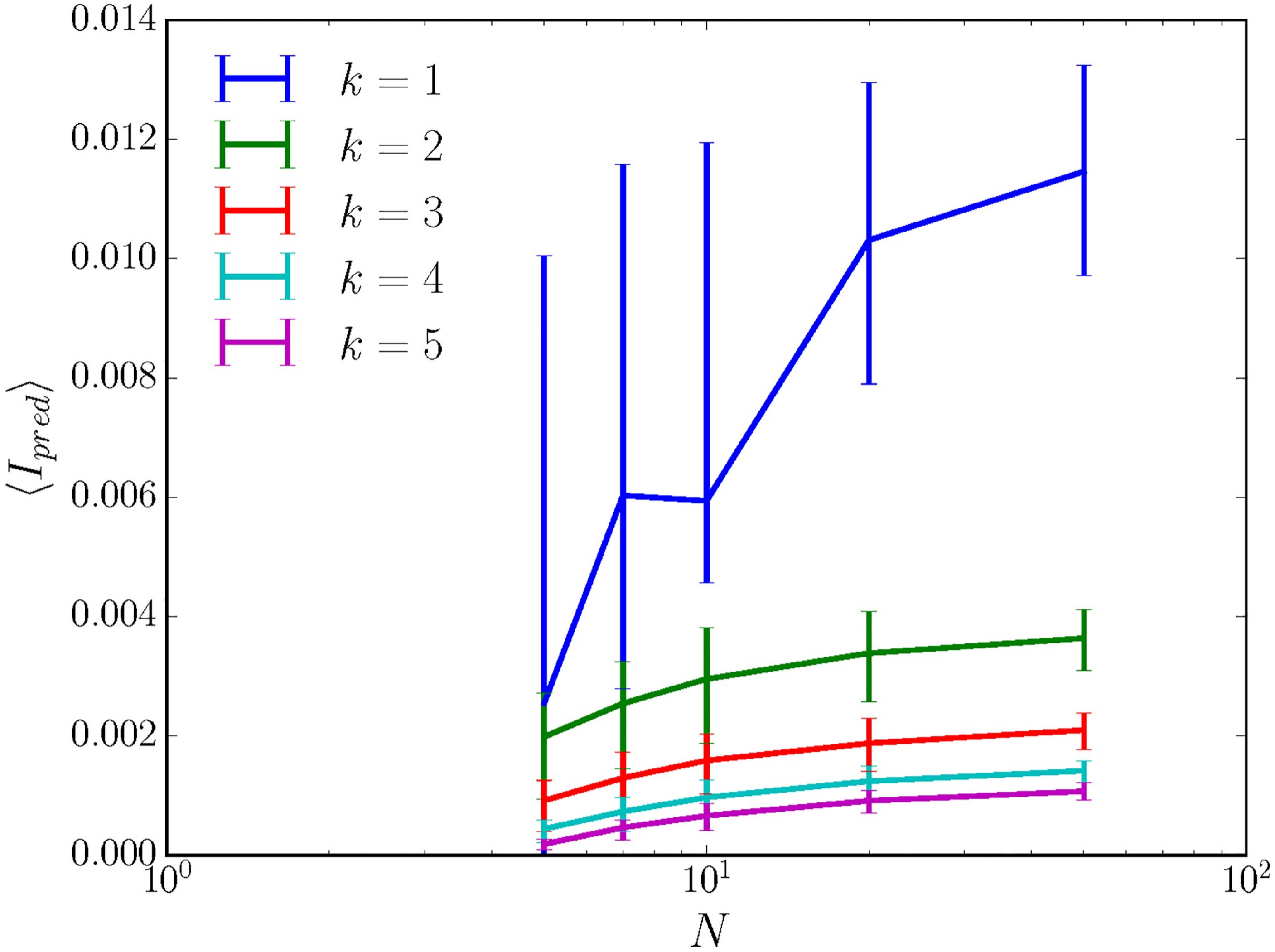

Fig 5. Nearly deterministic, randomly wired sensors capture more predictive information than nondeterministic randomly wired sensors. Nearly deterministic sensors are randomly generated as described in the main text with $\alpha = 3$, k as given in the legend, and N as given by the x-axis. 100 sensors are generated at each possible sensor size $N = |\mathcal{R}|$, and their predictive informations I_{pred} are averaged to give $\langle I_{pred} \rangle$. 50% confidence intervals in $\langle I_{pred} \rangle$ are given by the error bars. The environment has $|\mathbf{S}| = 30$, $\rho_\mu = 0.23$ nats, and $C_\mu = 2.3$ nats, and so these randomly-wired sensors are still only capturing $\sim$ 10% of the total predictive information possible.

https://doi.org/10.1371/journal.pone.0202333.g005

Conclusion

Evolution and scientists have a difficult task: that of creating predictive sensors of time series input that may contain long-range correlations. One can approach this problem by designing learning rules that adjust sensor connection weights so as to optimize sensor predictive capabilities. Nondeterministic sensors can result. Alternatively, one can approach this problem by randomly wiring a sensor and optimizing the sensor readout. We have shown that success of the latter approach requires determinism or something close to it, in that given a present sensor state and present input, one should be able to transition to only a few sensor states in the next time step.

One way to conceptualize the sensor's predictive computation is to view each sensor state as a cluster of pasts– that is, given a particular input past $\overleftarrow{x}$, there is a probability $p(r|\overleftarrow{x})$ of

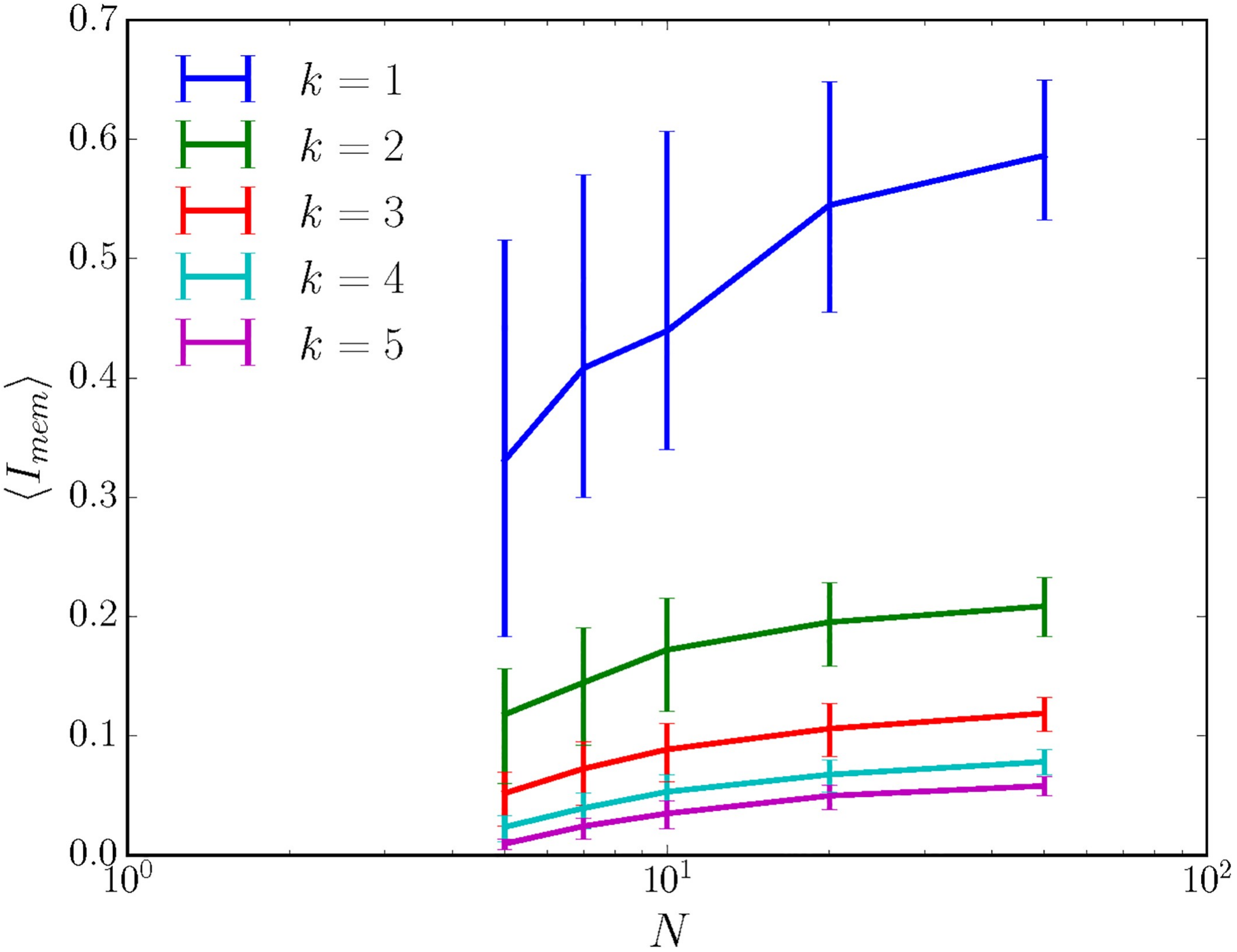

Fig 6. Nearly deterministic, randomly wired sensors have more memory than nondeterministic randomly wired sensors. Nearly deterministic sensors are randomly generated as described in the main text with $\alpha = 3$, k as given in the legend, and N as given by the x-axis. 100 sensors are generated at each possible sensor size $N = |\mathcal{R}|$, and their predictive informations I_{mem} are averaged to give $\langle I_{mem} \rangle$. 50% confidence intervals in $\langle I_{mem} \rangle$ are given by the error bars. The environment has $|\mathbf{S}| = 30$, $\rho_\mu = 0.23$ nats, and $C_\mu = 2.3$ nats, and so these randomly-wired sensors are only capturing $\sim$ 25% of the total memory possible.

https://doi.org/10.1371/journal.pone.0202333.g006

ending up in sensor state r determined in some nonlinear fashion by the dynamics of the input and the dynamics of the sensor. When a sensor is nondeterministic, these clusters are soft clusters, meaning that a number of sensor states r are possible given a particular input past $\overleftarrow{x}$. More sensor states correspond to more clusters, which as shown in Section A in S1 File, tends to increase memory and predictive information captured. However, an increase in the number of sensor states provides one with exponentially more pathways from one sensor state to the next for a particular input past, which on average have equivalent total weights for each input past. These two competing trends can lead to a nonmonotonic dependence of memory and predictive information on sensor size. At large enough sensor sizes, the latter trend always dominates. Hence, infinitely large, fully nondeterministic, randomly wired sensors are increasingly nonspecific for their input. To break the trend towards nonspecificity, one needs to restrict the number of possible paths between sensor states with determinism.

One may well ask what we have learned about optimal sensors from the model presented here. After all, it is well-known in signal estimation problems that sensor stochasticity is detrimental. The interesting twist in our story is that when discussing *recurrent* sensors, the detrimental effects of sensor stochasticity are compounded rather than mitigated by increasing sensor size. Feedforward sensors correspond to the null model studied in Section A in S1 File, while recurrent sensors correspond to the eigenvector analysis of the main text.

In essence, we have asked what kinds of random sensor ensembles yield greater predictive power. This is certainly not the first time that such a question has been asked, e.g. Refs. [7, 18, 19]. Previously studied reservoirs were all deterministic– or rather, the variable that evolved in a conditionally Markovian manner (e.g., membrane voltages as opposed to neural activities) in the reservoir computing applications evolved deterministically. As such, our report on the extremely detrimental effects of nondeterminism on recurrent sensors is new, if perhaps unsurprising. When sensors are deterministic, one can study the effect of the spectral radius of the weight matrix used to evolve the sensor state, the sparsity of aforementioned weight matrix, the function applied to the weight matrix multiplied by the sensor state, and sensor size (as studied here), among other things.

True determinism in physical systems is impossible [20, 21]. Sometimes, noise can be beneficial to the functioning of biological systems [16, 20, 22, 23], but our work here suggests that this noise must be tightly controlled when one wants to remember or predict input. For instance, in a chemical reaction network, too many different possible reactions for a given environmental forcing (nondeterminism) will lead to a nonspecific response, e.g. sparsity in random reaction networks was key to the results of Ref. [24]. Our results suggest that given a particular inherent sensor stochasticity, there is an optimal finite sensor size at which functionality (prediction) is maximized. The size of sensors in biological systems, then, might not always be governed by resource constraints [25–27] but instead governed by degradation of functionality due to unavoidable noise.

Supporting information

S1 File. Section A: Random soft clusters in the information bottleneck method. Section B: Plausibility argument for nonspecificity of large randomly-wired sensors. Section C: Fluctuations in memory and prediction.
(PDF)

S1 Fig. A null model for the effect of sensor size on predictive power predicts that larger sensors capture more information. The average information obtained about the relevant variable Y. The various lines correspond to various values of α, as indicated in the legends, and the x-axis corresponds to variation in the number of clusters N. We chose $M = 30$.
(TIFF)

S2 Fig. Variability in predictive information decreases with sensor size for fully nondeterministic sensors. On the x-axis is $|\mathcal{R}|$, or N, and on the y-axis is the interquartile range (IQR) of I_{pred}. The environment has $\rho_\mu = 0.147$ nats and $C_\mu = 2.36$ nats, but these results seemed to hold qualitatively regardless of particular environment.
(TIFF)

Acknowledgments

The author thanks J. P. Crutchfield and J. England for helpful discussions. SEM was funded by an MIT Physics of Living Systems Fellowship.

Author Contributions

Conceptualization: Sarah Marzen.

Formal analysis: Sarah Marzen.

Investigation: Sarah Marzen.

Resources: Sarah Marzen.

Software: Sarah Marzen.

Validation: Sarah Marzen.

Visualization: Sarah Marzen.

Writing – original draft: Sarah Marzen.

Writing – review & editing: Sarah Marzen.

References

1. Bubic A, Von Cramon DY, Schubotz RI. Prediction, cognition and the brain. Frontiers in human neuroscience. 2010; 4:25. https://doi.org/10.3389/fnhum.2010.00025 PMID: 20631856
2. Martens J, Sutskever I. Learning recurrent neural networks with hessian-free optimization. In: Proceedings of the 28th International Conference on Machine Learning (ICML-11). Citeseer; 2011. p. 1033–1040.
3. Jaeger H. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. Bonn, Germany: German National Research Center for Information Technology GMD Technical Report. 2001; 148(34):13.
4. Maass W, Natschläger T, Markram H. Real-time computing without stable states: A new framework for neural computation based on perturbations. Neural computation. 2002; 14(11):2531–2560. https://doi.org/10.1162/089976602760407955 PMID: 12433288
5. Maass W. Liquid state machines: motivation, theory, and applications. In: Computability in context: computation and logic in the real world. World Scientific; 2011. p. 275–296.
6. Lukoševičius M, Jaeger H, Schrauwen B. Reservoir computing trends. KI-Künstliche Intelligenz. 2012; 26(4):365–371. https://doi.org/10.1007/s13218-012-0204-5
7. Marzen SE. Difference between memory and prediction in linear recurrent networks. Phys Rev E. 2017; 96(3):032308. https://doi.org/10.1103/PhysRevE.96.032308 PMID: 29346995
8. Shannon CE. A Mathematical Theory of Communication. Bell Sys Tech J. 1948; 27:379–423, 623–656. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
9. Cover TM, Thomas JA. Elements of Information Theory. 2nd ed. New York: Wiley-Interscience; 2006.
10. Shalizi CR, Crutchfield JP. Computational Mechanics: Pattern and Prediction, Structure and Simplicity. J Stat Phys. 2001; 104:817–879. https://doi.org/10.1023/A:1010388907793
11. James RG, Ellison CJ, Crutchfield JP. Anatomy of a Bit: Information in a Time Series Observation. CHAOS. 2011; 21(3):037109. https://doi.org/10.1063/1.3637494 PMID: 21974672
12. Bialek W, Nemenman I, Tishby N. Predictability, Complexity, and Learning. Neural Computation. 2001; 13:2409–2463. https://doi.org/10.1162/089976601753195969 PMID: 11674845
13. Bialek W, Nemenman I, Tishby N. Complexity through Nonextensivity. Physica A. 2001; 302:89–99. https://doi.org/10.1016/S0378-4371(01)00444-7
14. Ara PM, James RG, Crutchfield JP. The Elusive Present: Hidden Past and Future Dependence and Why We Build Models. Phys Rev E. 2016; 93(2):022143.
15. Marzen SE, Crutchfield JP. Predictive rate-distortion for infinite-order Markov processes. J Stat Phys. 2016; 163(6):1312–1338. https://doi.org/10.1007/s10955-016-1520-1
16. Marzen SE, Crutchfield JP. Optimized Bacteria are Environmental Prediction Engines. Phys. Rev. E 2018; 98(1): 012408. https://doi.org/10.1103/PhysRevE.98.012408 PMID: 30110764
17. Nemenman I, Shafee F, Bialek W. Entropy and inference, revisited. In: Advances in neural information processing systems; 2002. p. 471–478.
18. Schrauwen B, Büsing L, Legenstein RA. On computational power and the order-chaos phase transition in reservoir computing. In: Advances in Neural Information Processing Systems; 2009. p. 1425–1432.

19. Verstraeten D, Schrauwen B, d'Haene M, Stroobandt D. An experimental unification of reservoir computing methods. Neural networks. 2007; 20(3):391–403. https://doi.org/10.1016/j.neunet.2007.04.003 PMID: 17517492

20. McDonnell MD, Abbott D. What is stochastic resonance? Definitions, misconceptions, debates, and its relevance to biology. PLoS computational biology. 2009; 5(5):e1000348. https://doi.org/10.1371/journal.pcbi.1000348 PMID: 19562010

21. Faisal AA, Selen LP, Wolpert DM. Noise in the nervous system. Nature reviews neuroscience. 2008; 9(4):292. https://doi.org/10.1038/nrn2258 PMID: 18319728

22. Ma WJ, Beck JM, Latham PE, Pouget A. Bayesian inference with probabilistic population codes. Nature neuroscience. 2006; 9(11):1432. https://doi.org/10.1038/nn1790 PMID: 17057707

23. McDonnell MD, Ward LM. The benefits of noise in neural systems: bridging theory and experiment. Nature Reviews Neuroscience. 2011; 12(7):415. https://doi.org/10.1038/nrn3061 PMID: 21685932

24. Horowitz JM, England JL. Spontaneous fine-tuning to environment in many-species chemical reaction networks. Proceedings of the National Academy of Sciences. 2017; 114(29):7565–7570. https://doi.org/10.1073/pnas.1700617114

25. Barlow HB. Possible principles underlying the transformations of sensory messages. 1961;.

26. Marzen SE, DeDeo S. Weak universality in sensory tradeoffs. Phys Rev E. 2016; 94(6):060101. https://doi.org/10.1103/PhysRevE.94.060101 PMID: 28085476

27. Marzen SE, DeDeo S. The evolution of lossy compression. Journal of The Royal Society Interface. 2017; 14(130):20170166. https://doi.org/10.1098/rsif.2017.0166