# APPROACHES TO INFORMATION-THEORETIC ANALYSIS OF NEURAL ACTIVITY

Submitted to *Biological Theory*

Jonathan D. Victor

Department of Neurology and Neuroscience
Weill Medical College of Cornell University
1300 York Avenue

New York City, NY 10021
Tel: 212 746 2343
Fax: 212 746 8984

jdvicto@med.cornell.edu

Abbreviated title: Information-theoretic analysis of neural activity

0 Figures, 1 Table.

**ABSTRACT**

Understanding how neurons represent, process, and manipulate information is one of the main goals of neuroscience. These issues are fundamentally abstract, and information theory plays a key role in formalizing and addressing them. However, application of information theory to experimental data is fraught with many challenges. Meeting these challenges has led to a variety of innovative analytical techniques, with complementary domains of applicability, assumptions, and goals.

## INTRODUCTION

The goal of this review is to identify some of the questions in neuroscience for which information-theoretic techniques provide useful insights and approaches, and to survey the variety of techniques that are applicable to the analysis of neurophysiologic data.

How neurons represent, process, and transmit information is of fundamental interest in neuroscience. The basic biophysics that underlies neuronal action potential generation is well established, as is the biophysics underlying many aspects of synaptic physiology and dendritic information processing. Nevertheless, the features of neuronal activity that convey and manipulate information are not well understood. Among the possibilities are relatively straightforward features, such as the number of spikes fired by a population of neurons (Shadlen & Newsome, 1998), but also more subtle ones, such as, their precise times of occurrence (Berry, Warland & Meister, 1997, Gawne, 2000, Softky, 1994, Théunissen, Roddey, Stufflebeam, Clague & Miller, 1996), the pattern of intervals (Sen, Jorge-Rivera, Marder & Abbott, 1996), the presence or absence of correlations and synchrony (Dan, Alonso, Usrey & Reid, 1998, Meister, Lagnado & Baylor, 1995, Rodriguez, George, Lachaux, Martinerie, Renault & Varela, 1999, Samonds, Zhou, Bernard & Bonds, 2006), oscillations (Gray & Singer, 1989), or other patterns of activity (Abeles & Prut, 1996).

Questions related to neural coding are intrinsically abstract, since, at a minimum, they seek a description of a mapping from events, percepts, and actions to something very different: patterns of neural activity. Though it may be tempting to assume that a common set of principles governs neural coding, it is more reasonable to anticipate that there is a diversity of biological solutions to the coding problem. That is, we anticipate that neural coding will differ greatly

according to the pressures under which a system has evolved. Such "design" criteria likely include minimizing the number of neurons or their connections, minimizing energy utilization, minimizing response latency, maximizing robustness in the face of injury, or maximizing the capacity for learning. We anticipate that coding strategies may differ across brain regions, even within a single "system". For example, cortical regions early in visual processing (V1, V2) are tightly topographically organized, while visual regions at the "top" of the inferotemporal stream, which interact extensively with polysensory areas and the hippocampus, have little topographic organization. Even within the early stages of visual processing, there is a qualitative change between coding in V1, and coding in V2 – with temporal multiplexing of multiple visual submodalities much more prominent in V2 (Victor & Purpura, 1996a). Finally, strategies for representing information, even within a particular cell type, are likely task-dependent and subject to top-down influences. For example, attention modulates firing rate (Luck, Chelazzi, Hillyard & Desimone, 1997, Reynolds, Pasternak & Desimone, 2000) and synchrony (Roelfsema, Lamme & Spekreijse, 2004, van der Togt, Kalitzin, Spekreijse, Lamme & Super, 2006). However, it is as yet unclear what is the primary neural correlate of attention.

*Need for joint experimental and theoretical/computational approach*

A purely experimental approach to these questions is not likely to succeed, in that manipulation of one feature of neural activity (e.g., increasing firing rate by electrical stimulation), is certain to change other aspects as well (e.g., interval structure, and degree of correlation). Thus, while such experiments (Salzman & Newsome, 1994) are critical in demonstrating that a particular brain region is relevant to a particular function, they provide little insight into neural coding.

An appropriate theoretical infrastructure is needed to disentangle these confounds, and also to compare results across a range of modalities, preparations, brain areas, and species. Shannon's groundbreaking work in information and communication theory (Shannon & Weaver, 1949) is the natural starting point for this theoretical infrastructure (Rieke, Warland, de Ruyter van Steveninck & Bialek, 1997). But, while application of Shannon's ideas to man-made communication channels is relatively straightforward, difficulties arise in attempting to apply information measures to biologic systems. Fundamentally, the Shannon theory was designed for characterizing communication systems whose principles were understood, not for the "inverse problem" of determining the principles by which a system works from observations of its behavior.

To make full use of information theory (and to avoid assuming answers to the above questions), one would want to begin with as few assumptions as possible about the nature of the neural code. A minimal assumption is that each possible configuration of neural activity (i.e., each arrangement of spikes across time and a set of neurons) is a candidate for a code word. Ideally, the formalism of information theory would then determine the actual set of words (and hence, the structure of the neural code) from this starting point. Unfortunately, this program rapidly runs into practical difficulties. Experimental estimates of information are biased by finiteness of datasets, and the extent of this bias is directly proportional to the size of the *a priori* set of words (Carlton, 1969). Moreover, the Shannon theory does not attempt to describe the relationship between a sensory or motor domain and neural activity (i.e., the nature of the neural representation), but merely provides an index of how faithful this representation is. As we will see below, these considerations motivate a variety of approaches to the analysis of neural coding. These approaches share the goal of quantification of information. However, they differ

substantially in the scope of the assumptions concerning neural coding, in the extent to which they yield a description of the representation provided by the code, and the kinds of data to which they may be applied.

*Correlation and causation*

Correlation of a behavior or stimulus with a statistical feature of the neural response does not imply that this feature of the neural response is used by the nervous system. Some of the approaches described below, coupled with appropriate experimental design, may be useful in determining causal relationships. For example, a multichannel recording of neural activity (e.g. field potential activity at different locations (Schiff, Kalik & Purpura, 2000) or multiple neurons within a cluster (Reich, Mechler & Victor, 2001b)) can be partitioned into two subsets of channels, one considered as the "input", and one considered as the "output". One can then determine whether statistical features in the "input" activity can predict later activity in the "output" channels. A positive answer demonstrates that the statistical features of the input are indeed used at later times in neural processing, thus going a substantial step beyond merely demonstrating the presence of these features.

Alternatively, because information cannot be created *de novo* within the nervous system, it may be possible to rule out a candidate neural code, by showing that it cannot support the sensory performance of the organism. This strategy has demonstrated the importance of spike timing in retinal coding (Nirenberg, Jacobs, Fridman, Latham, Douglas, Alam & Prusky, 2006).

## INFORMATION-THEORETIC TOOLS APPLICABLE TO NEURAL DATA

*General comments: a wide variety of approaches*

Many strategies for the application of information-theoretic tools to neural data have been proposed (Table 1). As seen in the Table, these strategies have diverse, and to some extent complementary, domains of applicability, limitations, conceptual underpinnings, and questions that can be addressed. We precede our survey by some general comments on these inter-related axes.

### *Experimental design*

A typical experiment in classical sensory neurophysiology consists of recording neural responses to a large number of presentations of a small set of sensory stimuli. The set of sensory stimuli is generally chosen to be "simple", with elements that vary along some perceptually salient parameter, or set of parameters. For example, in characterizing neurons in primary visual cortex, a typical stimulus set consists of gratings of varying contrast and/or orientation. Responses to such stimuli can be analyzed (without information-theoretic tools) to provide measures of neural "tuning" to these parameters. The information-theoretic viewpoint considers the neuron to be a communication channel. The "transmitted information" is a natural measure of to what extent an observer of the neural response can reduce uncertainty about which stimulus was presented. There is no pretense that this kind of experiment can fully characterize the response properties of the neuron. Nor can it hope to determine its information-transmitting capacity, since the set of stimuli is intentionally restricted to a tiny subset of all possible stimuli. Rather, the goal of information-theoretic analysis of this kind of experiment is to determine

which aspects of the response are responsible for coding the some perceptual parameter of interest, and the extent to which this coding is reliable.

An alternative experimental design, especially popular in vision, is based on the rapid presentation of a large set of stimuli (Wu, David & Gallant, 2006), repeated a small number of times if at all. The stimuli might be chosen to test a particular kind of model (e.g., white noise, m-sequences), or in the hope that they represent ethologically important stimuli (e.g., real-world movies). The goal of this kind of experiment is to build a model for the functional relationship between a neuron's input and its output. Such a model can then be tested by its ability to predict responses to other stimuli. Information-theoretic tools can then be applied to determine the information rate for the neuron's output under the conditions of the particular experiment. Moreover, if a believable model for the neuron's behavior can be constructed, then, at least in principle, the maximal information-transmitting capacity of the neuron (across all possible stimuli) can be calculated.

One might argue that the distinction between these two kinds of experiments is not very meaningful, since an information-theoretic analysis method that is intended to be applied to one kind of experiment can be forced to apply to the other. However, such application is unlikely to be practical, or to achieve its intended goal, even though there is nothing in the formalism of these approaches that prevents such attempts. The basic issue is that, like any other application of mathematical concepts to laboratory data (see Slepian (Slepian, 1976) for a beautiful discussion), a rigorous implementation of information-theoretic analyses requires evaluation of limits that cannot be achieved in the laboratory. Short of these limits, there is no *guarantee* that values estimated from laboratory data are close to their values at these limits. This difficulty typically

persists even if one goes through the efforts of analyzing exactly how rapidly the limits are approached – since this analysis is also an asymptotic one.

Thus, although the distinctions between the methods we discuss have clearcut and rigorous theoretical foundations, their practical domains of applicability are distinguished by qualitative terms and fuzzy borders (Table 1). But this should not be taken as an excuse to ignore the philosophical differences between these approaches. At a concrete level, such differences can be recovered by an analysis of how two kinds of procedures differ in simple test cases, whose behavior can be determined analytically. More fundamentally, ignoring the distinctions between these approaches would deny one of the important contributions of the mathematical biologist – namely, creation of formalisms that allow testing, refinement, and extensions of biological intuition.

### *Response types*

All information-theoretic methods discussed here can be applied to experiments in which the responses are the sequences of stereotyped action potentials ("spike trains") produced by a single neuron -- the substrate for information transmission over large distances. Many of the methods are also applicable to neural signals other than action potentials. For example, subthreshold fluctuations of membrane voltage carry information within neurons. Some small neurons, such as the interneurons of the retina, do not generate action potentials, and use these continuously varying voltage fluctuations for transmission of information between neurons. Another signal that is appropriate for information-theoretic analysis is the "local field potential," an extracellularly-recorded voltage that represents a combination of synaptic activity, subthreshold fluctuations of membrane voltage, and, to a lesser extent, summed spiking activity, in a neighborhood of approximately1 mm or less.

A spike train is most naturally represented as a point process, while intracellular and extracellular voltages are most naturally represented as a continuous real-valued function of time. As we will see below, some information-theoretic approaches are directly applicable to the point process itself. Other approaches have functions of time as their primary object of analysis. They can also be applied to spike trains, but only after the latter are converted into functions of time. Methods for making this conversion include convolution with a standard template, such as a Gaussian, or simply considering the spike trains to be a train of delta-functions. The latter approach can only be used for methods that do not require that the signals be continuous. Finally, the methods that are the most directly tied to Shannon's ideas (Shannon & Weaver, 1949) have a discrete sequence of symbols drawn from a finite set, typically {0,1}, as their primary object of analysis. These methods can be applied to spike trains by dividing the data record into narrow time bins, and keeping track of how many spikes occurred in each. They can also be applied to continuous signals, by sampling them in time and discretizing them in amplitude. The utility of these approaches depends critically on how information estimates vary with bin width, which in turn depends on the biological system and the amount of data available.

Understanding neural coding requires not only a characterization of the behavior of individual neurons, but also of their joint activity. Datasets in which many channels of simultaneously recorded neural activity (spikes, continuous signals, and combinations) are increasingly available. All of the methods we will consider have immediate formal extensions from single channels to multiple channels, but these extensions differ widely in practicality. The "multichannel" regime deserves to be broken into two regimes – that of "few" channels and "many" channels. Some methods effectively require estimation of a number of parameters that grows exponentially with the number of channels; these methods are likely to break down in the

"few" channel regime. For other methods, the effective number of parameters to be estimated grows more slowly if at all, but these methods may have computational demands that limit application when many channels are present.

*A survey of methods for information estimation*

*The direct method*

The "direct method" (Ruyter van Steveninck, Lewen, Strong, Koberle & Bialek, 1997, Strong, Koberle, Ruyter van Steveninck & Bialek, 1998) for the estimation of information in spike trains is closest to a literal implementation of Shannon's ideas, and makes only minimal assumptions about the nature of the code. Thus, it provides a rigorous estimate of information, provided that sufficient data are available.

The primary data consist of records of a single neuron's response. These records are first partitioned into segments of length $L$. Each segment is converted into a discrete sequence of symbols (0 or 1) by subdividing it into successive bins of width $\Delta T$, and forming an integer sequence in which each entry indicates the number of spikes within one of these bins. $\Delta T$ is typically taken to be sufficiently short so that each bin contains at most one spike. For each integer sequence $s$, the probability of its occurrence, $p(s)$, is estimated from experimental data. Two entropies are then calculated. The "total entropy", $H_{total} = -\sum p(s) \log_2 p(s)$, expresses the entropy of the entire repertoire of the observed behavior of the neuron, *for all stimuli*. The noise entropy $H_{noise}$ is a corresponding sum but restricted to responses to a single stimulus. The estimated information is $I = H_{total} - H_{noise}$.

The estimated information $I$ depends on the binning parameters $L$ and $\Delta T$. Strong et al. (Strong et al., 1998) provide a procedure for extrapolating to the limits of $\Delta T = 0$ and $L = \infty$, as is required for a rigorous true information estimate.

The direct method has been used at several levels of the visual system, including mammalian retina (Nirenberg, Carcieri, Jacobs & Latham, 2001), lateral geniculate nucleus (Reinagel & Reid, 2000), primary visual cortex (Reich et al., 2001b), and extrastriate visual cortex (Buracas, Zador, DeWeese & Albright, 1998). In each of these settings, the stimulus consisted of a rapidly-varying temporal sequence, often constructed from a pseudorandom sequence but occasionally derived from natural images (London, Schreibman, Hausser, Larkum & Segev, 2002, Nirenberg et al., 2001). However, the method can also be applied to data derived from discrete presentation of a small set of stimuli (Reich, Mechler & Victor, 2001a).

*Limitations*

The main limitation of the direct method is that it is simply not possible to make a *rigorous* extrapolation to the limits of $\Delta T = 0$ and $L = \infty$. These limits of course cannot be attained experimentally, but biologic considerations can provide guidelines for values of $\Delta T$ and $L$ beyond which one can assume that an asymptotic regime is reached. Unfortunately, this regime may be in accessible in practice.

For mammalian cortex, a reasonable choice of $\Delta T$ is 1 ms (an upper limit for the intrinsic precision of a neuron), while a reasonable choice for $L$ is 100 ms (a lower limit for the duration of a response). Consequently, the number of possible sequences whose probabilities must be estimated is very large ($2^{L/\Delta T}$), and the probability distribution is necessarily undersampled by laboratory data. In this regime, entropy estimates are unreliable and highly biased -- the bias is proportional to the number of probabilities that must be estimated, and

inversely proportional to the total number of observations. As described below, debiasing techniques are available, but these procedures are ineffective when most bins are not even sampled at all. Consequently, the direct approach is limited to situations in which responses are highly reproducible, such as insect systems or the retina (so that only a very small number of the possible spike train configurations occur), or, to estimates of instantaneous information rate (artificially limiting $L$).

The direct method may be extended (Johnson, Gruner, Baggerly & Seshagiri, 2001, Nirenberg et al., 2001, Reich et al., 2001b) to simultaneous recordings from multiple neurons. In an $M$-neuron experiment, the response within each bin of length $\Delta T$ is described by an $M$-tuple of bits, in which each bit represents the firing of one neuron. Otherwise, the estimation of information proceeds exactly as for single-neuron responses. However, the undersampling of the space of all possible sequences is even more severe, since the number of possible sequences is given by $2^{ML/\Delta T}$.

In sum, the philosophy that keeps the direct method closest to Shannon's ideas is also its main limitation. Since minimal assumptions are made about the nature of the code, the probability of each response (as represented by a discrete sequence) is an independent quantity to be estimated from data. That is, the tradeoff for an approach that is free of *a priori* assumptions is one that, for rigorous implementation, requires an impracticably large amount of data in many circumstances. Moreover, the direct method provides little insight into *how* information is carried – since how information is carried is explicitly a statement about the relationships among the response sequences.

*Estimators of entropy of a discrete distribution*

A key component of the "direct method", as well as of many of the approaches described below, is that the entropy of a discrete distribution must be estimated from a finite set of observations. This seemingly simple problem is surprisingly subtle. The entropy of a discrete distribution with *J* bins and a probability $p_j$ in each bin is $H = -\sum_{j=1}^{J} p_j \log_2 p_j$. The naïve approach is to estimate this by setting $p_j = n_j / N$, where $n_j$ is the number of times that the *j*th outcome is observed, and *N* is the total number of observations. This "plug-in" estimator is well-known to be biased – fundamentally, because of the curvature of the log function. A standard fix is to add a bias correction(Carlton, 1969, Miller, 1955, Treves & Panzeri, 1995). This bias correction is asymptotically exact for large *N*, but requires knowledge of the number of kinds of categories (or bins), *J*, that occur with nonzero probability. Moreover, typical datasets are not in the "asymptotic" regime, which requires that even the least likely outcome has been observed several times. An alternative correction is the jackknife (Efron, 1982, Efron & Tibshirani, 1998), but this has similar asymptotic behavior. More sophisticated estimators have recently been introduced, with clear advantages in regimes relevant to laboratory data. These include Paninski's estimator (Paninski, 2003), which is provably the least-biased of all polynomial estimators, the "KT" (Krichevsky & Trofimov, 1981) and "SG" (Schurmann & Grassberger, 1996) estimators, which are based on single Dirichlet priors (Wolpert & Wolf, 1995), the "NSB" estimator (Nemenman, Bialek & de Ruyter van Steveninck, 2004), which considers a family of Dirichlet priors, and the Chao-Shen (Chao & Shen, 2003) estimator, recently introduced in ecology. However, none of these estimates succeed in the severely undersampled regime characteristic of cortical datasets.

*Metric space method*

The direct method, though virtually assumption-free, can have prohibitive data requirements, and does not attempt to characterize the manner in which information is represented. The metric space method (Victor, 2005, Victor & Purpura, 1997) represents an alternative viewpoint. By making assumptions as to the nature of a neural code, it can provide useful estimates of information in settings in which the direct method will fail (limited amounts of data, and especially high firing precision but low firing rate).

The metric space method considers several generic families of neural codes, each of which is designed to test a particular hypothesis of *how* information is carried, such as via spike counts, or via the timing of spikes, or via the interval structure of the spike trains. Each of these hypotheses is then formalized in terms of a family of metrics – notions of distance (i.e.., dissimilarity) between spike trains. The metrics have a common structure, which allows comparison of the hypotheses on a level playing field. Since the metrics explicitly recognize that neural responses are point processes and their structure respects the continuity of time, the binning process that limits the use of the direct method is avoided. However, the metric-space method typically underestimates the total information that is present, since only a few stereotyped (but interpretable!) hypotheses for the code are considered. Also, because of the way that information is calculated, the approach is limited to analysis of episodic responses to a discrete set of stimuli.

Many neurons can be considered to behave like coincidence detectors (Bourne & Nicoll, 1993, Cline, 1997, Markram, Lubke, Frotscher & Sakmann, 1997, Mel, 1993, Softky & Koch, 1993, Usrey, Reppas & Reid, 1998). This suggests that the meaning of a spike train is determined by the timing of the individual spikes, since it is those timings that determine how

the multiple inputs onto a dendritic tree interact to determine a postsynaptic neuron's behavior. To assess the extent to which spike times carry information, the approach uses a family of metrics denoted $D^{spike}[q]$, parameterized by a quantity $q$ (see below) that describes the role of temporal pattern. According to the metric $D^{spike}[q]$, the distance between two spike trains is the minimum total "cost" to transform one spike train into the other via any sequence of insertions, deletions, and time-shifts of spikes. The cost of moving a spike by an amount of time $t$ is set at $qt$, and the cost of inserting or deleting a spike is set at 1. Thus, in the sense of $D^{spike}[q]$, spike trains are considered similar if they have approximately the same number of spikes, and these spikes occur at approximately the same times, i.e., within $1/q$ or less. A neuron that behaves like a coincidence detector with temporal precision $1/q$ would see incoming spike trains as similar or different, according to the metric $D^{spike}[q]$.

A second family of metrics, denoted by $D^{interval}[q]$, is motivated by the notion that a synaptic response depends on its recent history, and thus, the intervals between successive spikes may also carry information (Abbott, Varela, Sen & Nelson, 1997, Bliss & Collingridge, 1993, Sen et al., 1996, Usrey et al., 1998). In the metric $D^{interval}[q]$, the distance between two spike trains is defined as the minimum total cost to transform one spike train into the other via any sequence of insertions of spikes, deletions of spikes, and expansions or contractions of interspike intervals. The parameter $q$ specifies the cost $qt$ of changing an interspike interval by an amount $t$. In the limit that $q$ approaches 0, both $D^{spike}[q]$ and $D^{interval}[q]$ approach a metric $D^{count}$, which is sensitive only to the number of spikes, and not to any aspect of their timing.

Each metric is then evaluated by the extent to which it distinguishes the responses to each of the stimuli – namely, the transmitted information between stimulus and response clusters. The

dependence of the transmitted information on $q$ for $D^{spike}[q]$, and $D^{interval}[q]$ characterizes the importance of spike timing and interspike intervals, across a range of temporal precisions.

Applications of this approach to neural data, including visual cortex (Reich, Mechler & Victor, 2000, Samonds & Bonds, 2004, Victor & Purpura, 1996a), chemical senses (Di Lorenzo & Victor, 2003, Stopfer, Bhagavan, Smith & Laurent, 1997), and electric sense (Kreiman, Krahe, Metzner, Koch & Gabbiani, 2000), are reviewed in (Victor, 2005)).

The metric-space approach is readily extended to the multineuronal context. A multiunit recording is a sequence of labeled events, with the label representing the neuron of origin. To assess the importance of *which* neuron fires each spike, multineuronal metrics add an additional transformation between spike trains: changing the label associated with a neuron. The cost of this transformation is assigned the quantity $k$. The extreme $k = 0$ corresponds to a code in which the neuron of origin is irrelevant (since it is free to change the label associated with each spike). The other extreme, $k = 2$, corresponds to a labeled-line code (since it costs as much to change the label on a spike as it does to remove it from one neuron, and insert it into another). The above analyses can then be carried out for the two-parameter family $D^{spike}[q,k]$.

By introducing a *single* parameter to explore the continuum between codes in which neuron of origin is irrelevant and labeled-line codes, the explosion of parameters that might otherwise hobble attempts to analyze multineuronal data is circumvented. We have applied this approach to simultaneously-recorded neural pairs in V1 (Aronov, Reich, Mechler & Victor, 2001), and have found that responses are best decoded by keeping track of which neuron fired which spike, but only a modest amount of information is lost by ignoring the neuron of origin. This is in keeping with our analysis of multineuronal recordings in V1 via the direct method (Reich et al., 2001b), but is complementary to it: the direct method can analyze recordings of up

to 6 neurons (the limits of our recording), but only looks at information rates over brief time intervals (e.g., 15 ms). In contrast, the metric-space method can examine responses over extended periods.

For multineuronal responses, algorithms for the calculation of distances via straightforward extension of the Sellers algorithm (Sellers, 1974) (see below) yield a calculation time proportional to $c^{2M}$, where $M$ is the number of neurons and $c$ is the typical number of spikes in a spike train. An improved dynamic programming algorithm that drops the exponent from $2M$ to $M+1$ was recently found (Aronov, 2003). This dramatic improvement makes calculations on triplets of neurons practical on a desktop, and enables analysis of 4 to 8 neurons (for firing rates typical of cortical neurons) with a parallel processor array.

*Limitations*

One important limitation of the metric-space approach is that there is no guarantee that the manner of information transmission is similar to either of these caricatures. For example, the informative precision of a spike may be greater during the transient part of a response than during a later period in which firing occurs at a lower rate. In the multineuronal situation, it may be appropriate to distinguish among some neurons within the population and not others, rather than to have a single omnibus cost for changing the label of a neuron. One can augment the metric space method by including these (and other) variations. Consequently, the maximal value of the transmitted information obtained with any of the candidate metrics is at best an underestimate of the total amount of information. Since there are also coding strategies that do not readily fit into the metric structure, it is difficult to place rigorous bounds on the extent of this underestimate.

A second major limitation of the metric space method is a consequence of the clustering stage, in which distances between responses to the same stimulus, and distances between responses to different stimuli, are compared. For the clustering stage to be effective, the number of samples collected in response to each stimulus must be somewhat larger than the number of stimuli. This makes it impractical to apply the metric-space method to responses elicited to long, rich sequences of continuously presented stimuli.

*Relation to comparison of genetic sequences*

The above metrics for spike trains have a common structure: distance is defined as the minimum cost of a transformation of one sequence into another, via a sequence of prescribed elementary transformations. This structure is formally identical to that of the distances used to compare genetic sequences (Sellers, 1974). For genetic sequences, the elementary transformations include insertion, deletion, and alteration of a discrete element. The spike train metrics operate on point processes in continuous time, while the distances for genetic sequences operate on discrete sequences. Despite this topological difference, the highly efficient dynamic programming algorithms developed by Sellers (Sellers, 1974) for genetic sequences can be adapted to spike train metrics, so that the calculations described above can be carried out efficiently.

*Not just information*

The metric-space approach, and others to be described below, goes beyond traditional information-theoretic analysis in an important way. One can determine whether the presumptive code provides for a *representation* of the stimulus domain, and not just for faithful discrimination of distinct stimuli. One way to accomplish this is to use the pairwise distances as

the starting point for multidimensional scaling (Aronov et al., 2001, Victor & Purpura, 1997). For example, re-analysis of the auditory data of Middlebrooks et al. (Middlebrooks, Clock, Xu & Green, 1994) demonstrated that the temporal aspects of the spike trains not only identify the azimuth of origin of a sound, but also that these temporal aspects *represent* the azimuth: they map the responses into a circular locus in an abstract response space (Victor & Purpura, 1997). Moreover, the coordinates within the multidimensional scaling space are the temporal features that distinguish and represent the stimuli. Such an analysis of V1 recordings (Aronov et al., 2001) demonstrated a consistent temporal representation of spatial phase for across neurons, with one coordinate consisting of the sustained portion of the response, and a second coordinate consisting of a transient component.

### *Embedding method*

The "embedding method" is an approach that combines many of the advantages of the two approaches discussed above(Victor, 2002). Like the metric space method, it exploits the continuity of time and avoids binning. But in contrast to the metric space method, it makes no assumptions concerning the nature of the code, other than that it respects the continuity of time. Consequently, it is provably unbiased(Kozachenko & Leonenko, 1987) – at least when sufficient data are available. It can be extended to multichannel data, but its behavior is intermediate between that of the metric space method (a single parameter is added) and the direct method (exponential growth in number of parameters to be estimated). While the approach cleanly separates information carried by spike counts from information carried by spike times, it does not provide as detailed a parsing of temporal information as does the metric space method. In contrast to both the metric space method and the direct method, this approach is immediately applicable to continuous responses and spike trains.

The key idea behind this approach is a formalization of a basic attribute that a coding scheme must have in order to be biologically plausible. A sufficiently small change in the time of occurrence of a spike cannot result in a change in the meaning of a spike train, and spike trains that differ by only an infinitesimal change in a spike time must have nearly identical probabilities. Thus, like the metric space method, the continuity of time is used explicitly. But unlike the metric space method, there is no assumption made concerning the relationship of spike trains that differ by large displacements of a spike. Also, in contrast to the metric space method, the approach does not assume a relationship between the two spike trains that differ by insertion or deletion of a spike. These ideas are naturally formalized in terms of the topology of spike trains(McFadden, 1965). That is, the space of spike trains of finite duration can be considered to consist of a discrete set of strata, one for each number of spikes. Spike trains with $n$ spikes form an $n$-dimensional manifold (parameterized by the time of each spike). A neuron's output is a probability distribution on this set of strata. Within each stratum, the probability distribution is assumed to vary smoothly, but between strata no assumptions are made.

Thus, to determine the amount of transmitted information in an experimental dataset, spike trains are stratified according to the number of spikes $n$ in the response. This partitioning generates one component of the information, $I_{count}$, reflecting the extent to which the total number of spikes in the response can distinguish between the stimuli. Since $I_{count}$ is determined from a relatively small number of response categories, a standard discrete calculation may be used, and standard bias corrections are effective. Then, the $n$th stratum is analyzed to determine a contribution of spike timing $I_{timing}(n)$. The total information is $I_{count} + \sum_n I_{timing}(n)$, where the second term is the total information due to spike timing.

The calculation within the *n*th stratum crucially exploits the assumption that the probability distribution is a continuous function of the spike times. To determine $I_{timing}(n)$, the spike trains in the *n*th stratum are embedded into a Euclidean space of dimension $r \le n$. The coordinates assigned to a response are determined by inner products with a set of functions $f_1$, …, $f_r$: a spike train *x* with spikes at times $\tau_1, \tau_2, \ldots, \tau_n$ is mapped to coordintes $c_h(x) = \sum_{k=1}^{n} f_h(\tau_k)$.

For continuous signals, there is no discrete component corresponding to the number of spikes, and all responses are embedded into a space of the same dimension. A reasonable choice for the embedding is the natural extension of the above linear map to continuous signals: a signal $v(t)$ is mapped into the coordinates $c_h(v) = \int f_h(t)v(t)dt$.

As in the direct method, transmitted information is calculated as a difference between a "total entropy" determined from all responses considered together, and a "noise entropy" determined within the responses to each stimulus. However, in contrast to the direct method, these entropies are determined by examining the statistics of the nearest-neighbor distances (Kozachenko & Leonenko, 1987). In particular, the contribution of spike timing to the information within the *n*th stratum is estimated by

$$I_{timing}(n) \approx \frac{r}{N(n)} \sum_{j=1}^{N(n)} \log_2\left(\frac{\lambda_j}{\lambda_j^*}\right) - \sum_{k=1}^{s} \frac{N(n,a_k)}{N(n)} \log_2 \frac{N(n,a_k)-1}{N(n)-1} \ ,$$

where $N(n)$ is the number of spike trains with *n* spikes, $N(n,a_k)$ is the number of spike trains with *n* spikes elicited by the *k*th stimulus, $\lambda_j$ is the distance between the *j*th spike train and its nearest neighbor, and $\lambda_j^*$ is the distance between the *j*th spike train and its nearest neighbor elicited by the same stimulus. For quantities of data typically available in an experiment, this nearest-neighbor estimator (of entropy or of information) is substantially more efficient than binned

methods. Demonstration that this estimator is unbiased (Kozachenko & Leonenko, 1987) relies critically on the assumption of smoothness of the probability distribution.

*Limitations*

The limitations of the embedding approach relate chiefly to the discrete component of the entropy estimate. When the range of the number of spikes in responses is large, there are many discrete partitions. In this regime, the bias estimates for $I_{count}$ may be ineffective. Moreover, at the tails of the distributions of spike counts, there are only a few responses, so that the estimate of $I_{timing}$ may be ineffective. These difficulties may be mitigated by lumping together partitions with similar numbers of spikes, but this compromises the unbiased nature of the estimator. The practical difficulties of the discrete component are exacerbated when the method is applied to multineuronal data, since a separate partition is required for each combination $(n_1, n_2, ..., n_M)$ of spike counts on each of the $M$ neurons. This rate of growth of the number of partitions that must be separately analyzed, though high, is much lower than in the direct method, since it is independent of (rather than exponential in) temporal resolution.

*Relation to general dynamical systems approaches*

Estimation of entropy from the statistics of nearest neighbors is related to estimation of dimension of a dynamical system's trajectory or attractor set. Grassberger and Procaccia (1983) describe several versions of such procedures, wherein dimension is determined from the relationship between the number of points within a given radius, and the radius. When plotted on log-log coordinates, the slope of this relationship is the sought-after dimension. But in the present situation, the slope is known (the dimension of the space in which we have embedded spike trains), and the quantity of interest, the entropy, is essentially the intercept of this line.

Grassberger's (1988) finite-sample debiasing procedure applies specifically to the slope (dimension); the Kozachenko and Lenonenko (1987) estimator debiases the intercept (entropy).

Grassberger and colleagues (Kraskov, Stogbauer & Grassberger, 2004) have recently described a related approach to estimating mutual information via a nearest-neighbor approach that avoids explicit estimates of dimension. However, this approach requires that the response variable has a definite dimension. Thus, for application to spike trains, a procedure such as stratification by spike count is required to obtain an unbiased estimator, as in (Victor, 2002).

*Context tree method*

The context tree method is a promising new approach both for entropy estimation (Kennel, Shlens, Abarbanel & Chichilnisky, 2005, London et al., 2002) and for estimation of mutual information applicable to the "many-presentation" experimental design (Shlens, Kennel, Abarbanel & Chichilnisky, 2006). Like the direct entropy estimator (Ruyter van Steveninck et al., 1997, Strong et al., 1998), it is based on a discrete representation of spike trains, but, it also makes crucial use of the dynamic nature of spike trains – namely, that they a spike train is a temporal sequence in which the recent past influences the probability of spiking. This dynamic process is modeled as a "context tree" (Rissanen, 1989), which differs from a Markov process in that the depth of the history dependence can be non-uniform. This model form is intuitively appealing for neural data, and results in a substantial increase in efficiency compared with approaches (see "Compression method", below) that make use of dynamics, but do not postulate a model form.

In essence, the method has two components: estimation of a context tree model from the spike train data, and then calculation of entropy from the context tree itself (e.g., by a Wolpert-Wolf estimator (Wolpert & Wolf, 1995)). However, rather than choose a single context tree

model (cf. (Hirata & Mees, 2003)), the approach considers many context tree models. Each model's contribution is discounted (Willems, Shtarkov & Tjalkens, 1995) by a factor that considers both the complexity of the model (its "codelength," (Solomonoff, 1964)) and the extent to which the model is a poor fit to the data. An advantage of this approach is that confidence limits on the entropy estimates can be determined via a Monte Carlo method that explores the range of estimates that would result from alternative context tree models (Kennel et al., 2005).

### *Other methods*

Below we describe several other approaches that may be usefully applied to estimation of information in neural data. Our goal is to emphasize the variety of viewpoints that may be taken, rather than to present an exhaustive review.

#### *Principal components*

The procedures used by Richmond and Optican (Chee-Orts & Optican, 1993, Optican & Richmond, 1987, Richmond & Optican, 1987) are based on principal-components analysis of rate functions estimated from single-trial neural response. The hypothesis underlying this approach is that information is coded as a firing rate envelope, and that individual spike trains serve as estimators of this envelope. This approach can also be viewed as a kind of the embedding method, in that the rate coding hypothesis leads to embedding of all responses in a space of the same dimension, regardless of the number of spikes. Within this space, information is estimated by parceling this space into multidimensional bins. A regularization procedure based on an additive noise model and an assumed Gaussian shape of the response cluster were used to improve performance (Chee-Orts & Optican, 1993). To the extent that neural codes

indeed conform to the rate envelope hypothesis, the principal-components approach will provide a good description of the code, with limited sample sets of the size achievable in typical experiments (McClurkin, Optican, Richmond & Gawne, 1991, Optican & Richmond, 1987, Richmond & Optican, 1987). However, by design, it will overlook any other forms of coding. Additionally, the Gaussian regularization for estimation of entropy, rather than the nearest-neighbor estimator used in the embedding method, is tantamount ot adding an assumption about the manner in which responses vary across trials.

*Reconstruction method*

The reconstruction method of Bialek and coworkers (Bialek, Rieke, Ruyter van Steveninck & Warland, 1991) was the first information-theoretic approach successfully applied to decoding dynamic neural activity. It provides another way of avoiding the difficulties associated with estimating a large number of probabilities, as is required by the direct method. The basic strategy is to identify a transformation of the observed neural response that best reproduces the known stimulus sequence. The transmitted information in the neural response is then known to be at least as high as the mutual information between the actual stimulus and the stimulus reproduced by this transformation rule. In some settings, *a priori* calculations allow for an independently calculated upper bound on the amount of information in the neural response, based on the theoretical limits of a sensory system (Bialek et al., 1991). When the upper bound provided by these considerations is close to the lower bound provided by a reconstruction, this approach is particularly powerful and elegant.

To seek a transformation between the neural response and the stimulus, a functional form must be chosen. This functional form is typically linear, though nonlinear extensions via the Volterra formalism (Marmarelis & Marmarelis, 1978) can be used. The kernels that describe

the transformation can then be interpreted as a recipe for "reading" the neural code (Bialek et al., 1991). The approach is typically applied to the spiking activity of single neurons (Théunissen et al., 1996), but the concept readily extends to multiple channels and/or continuously varying data. One limitation of the approach is that that the stimulus must be represented as a time series, rather than as discrete elements of a space. More fundamentally, the approach may be impractical for highly nonlinear transformations, such as are likely to be present within the mammalian central nervous system, since the fitting of second-order (or higher) terms in a Volterra series will not be robust.

*Power series method*

Panzeri and Schultz (Panzeri & Schultz, 2001, Schultz & Panzeri, 2001) introduced another strategy for overcoming many of the shortcomings of the direct method by exploiting the continuity of time. Here, the basic assumption is that information is an analytic function of the length of the analysis interval $L$. Under this assumption, information can be expanded as a power series in $L$. Very short intervals are likely to contain at most one spike. The probability that a pair of spikes occurs within the analysis interval increases with the square of the length of the interval. Thus, an advantage of this approach is that the terms of the Taylor series expansion separate the contributions of firing rate, pairwise correlation between spikes, and higher-order correlations. This parsing of temporal information, which is explicitly order-by-order, is intrinsically limited to spike trains. However, it is distinct from (and more detailed than) the kind of parsing provided by the metric space method. Additionally, this approach bypasses the construction of a response space, so there is no attempt to determine whether stimuli are "represented" by the temporal patterns of activity.

In contrast to the reconstruction method, it is not assumed that the relationship between a spike train and what it represents has a low-order power series expansion. Rather, a power series is used to represent the information content of a spike train as a function of the duration of the interval (i.e., order-by-order in the number of spikes). Thus, the power series method will have no trouble with highly nonlinear transformations such as thresholds and saturations that might lead to difficulties with the reconstruction method.

The power series approach is readily extended to multiple spike trains, but at any fixed order of approximation, the number of cross-terms grows as a polynomial in the number of neurons. The second-order terms can be further separated into auto- and cross-correlation terms, providing insight into how information is coded across a population of neurons. On the other hand, when the spike trains have structure such as regularity or bursts, there is no guarantee that the power series converges rapidly, or even at all. This may prevent successful application to such spike trains, or to large analysis intervals.

This approach has been used successfully to study somatosensory encoding in rat barrel cortex. Temporal analysis of single spike trains demonstrated an important role for timing of the first spike(Panzeri, Petersen, Schultz, Lebedev & Diamond, 2001), with a smaller role for subsequent multispike patterns. Analysis of multichannel data demonstrated the practicality of the approach for studying coding by correlated activity across neurons, initially with a limited temporal analysis (Panzeri, Schultz, Treves & Rolls, 1999) and later with a full temporal analysis (Petersen, Panzeri & Diamond, 2001).

*Compression method*

The entropy of a spike train can be measured by how susceptible it is to lossless data compression, via the Lempel-Ziv algorithm (Farach, Noordewier, Savari, Shepp, Wyner & Ziv,

1995, Kontoyiannis, Algoet, Suhov & Wyner, 1998, Levy, 2000, Wyner & Ziv, 1989). As in the direct approach, spike trains are segmented and discretized into a sequence of symbols, and no assumptions are made as to the nature of the code, or the statistical structure of spike trains.

In essence, the Lempel-Ziv algorithm seeks to compress a sequence of symbols by rewriting the sequence in terms of a hierarchy of repeating substrings. The substrings that occur frequently thus provide a characterization of the statistical structure of the neural activity. Additionally, the behavior of the compression algorithm as a function of bin width could be used to characterize the temporal precision of the code. One anticipates that this approach should be highly adept at dealing with high-order statistical patterns of spikes, such as bursts (or even runs of bursts), because the compression algorithm intrinsically seeks recursive layers of structure. Another consequence of the avoidance of an explicit estimate of spike train probabilities is that multineuronal data *per se* should not be an obstacle.

While in principle this approach is exact, convergence of the entropy estimates is difficult to bound (Levy, 2000) and appears sensitive to the details of the compression algorithm, such as the choice of the initial dictionary of strings. Nevertheless, it can result in efficient, meaningful entropy estimates when applied to neural data (Amigo, Szczepanski, Wajnryb & Sanchez-Vives, 2004). Determination of algorithmic complexity (Rapp, Zimmerman, Vining, Cohen, Albano & Jimenez-Montano, 1994) is a related approach, as are the context-tree methods described above.

*Spectrotemporal methods*

Spectrotemporal (or time-frequency) analysis is a general exploratory method that is particularly suitable for neural data, both spiking and continuous (Mitra & Pesaran, 1999). It is not typically considered an information-theoretic tool, but we mention it here because it also can be used to identify meaningful statistical structure in spike trains.

Spectrotemporal analysis is a natural extension of spectral analysis. Spectral analysis formally requires that the signals to be analyzed are "stationary" (i.e., have statistical properties that do not change in time). Neural signals, especially those influenced by external stimuli, do not have this property; rather, this evolution in time may be specifically of interest. The straightforward way to deal with this problem is simply to segment the data into periods that are sufficiently brief so that within each period, the signals can be assumed stationary. Standard spectral analysis applied to each segment can then reveal how the frequency characteristics of a signal evolve over time. As is well known, the length of the analysis segment and the achievable frequency resolution limit are reciprocally related. Sophisticated spectrotemporal techniques based on multitaper estimates (Mitra & Pesaran, 1999, Thomson, 1982) and wavelets(Quiroga, Rosso, Basar & Schurmann, 2001, Schiff, Aldroubi, Unser & Sato, 1994), while of course unable to circumvent limits on simultaneous resolution in time and frequency, represent a principled way to approach them.

Spectrotemporal analysis can identify stimulus-dependent changes in neural activity that would escape ordinary averaging techniques, such as event-related synchronization and desynchronization (Pfurtscheller & Andrew, 1999). Spectral analysis has a natural extension to the multichannel context: calculation of coherences (or cross-spectra) between channels that characterize their correlations within each frequency band. Spectrotemporal analysis has a directly analogous extension, which provides a description of how the coherence between signals evolves over time. The phase relationships between activity in different channels (e.g., different neurons or field potentials in different brain regions) provide another way to identify the direction of information transfer. The frequency bands at which coherence is present can suggest how information is transferred. For example (Schiff et al., 2000, Schiff, Kalik & Purpura, 2001),

coherence between activity in distant cortical areas and between cortex and thalamus is present at particular frequency bands at specific times during a behavioral task, and is correlated with behavioral performance.

Another contact with information-theoretic approaches is that regions of the time-frequency spectrum can be used as classifiers of the neural response (Jarvis & Mitra, 2001). Under fairly general assumptions, the log(power) in non-overlapping regions of a time-frequency spectrum are approximately independently-distributed Gaussian variables. Thus, reduction of a set of responses into measures of power in multiple time-frequency regions can serve as a first step in calculation of transmitted information. The amount of information, as well as the time-frequency regions that are critical in transmitting it, can thus be readily determined. Note that this approach to estimating information not only exploits the continuity of time, but also the intuition that neural coding is smooth in the frequency domain.

Wavelet methods (Quiroga et al., 2001, Schiff et al., 1994, Tallon, Bertrand, Bouchet & Pernier, 1995) and multitaper methods, in essence, are complementary strategies for parceling the spectrotemporal domain into rectangular tiles. In multitaper methods, the tiles are uniform, and thus optimized for detecting features of a given temporal duration or frequency bandwidth. In contrast, wavelets tile the spectrotemporal domain with regions whose dimensions are reciprocally related, and thus optimized for detecting features whose durations and bandwidths have a given ratio.

*Surrogate datasets*

Since many hypotheses concerning neural coding can be phrased in terms of comparisons between the observed data and surrogate datasets, procedures for surrogate data generation are important adjuncts to the procedures described above. The use of surrogate data sets for testing

hypotheses concerning the dynamics of continuous neurophysiologic data are widely appreciated (Schiff, So, Chang, Burke & Sauer, 1996, Theiler, Galdrikian, Longtin & Farmer, 1991, Theiler & Rapp, 1996). The approach is at least as relevant to testing and refining hypotheses concerning information transmission in spike trains.

### *Shuffling*

Perhaps the simplest hypothesis that one might want to test is whether the amount of information in an experimental dataset is nonzero. As mentioned above, analytic estimates of the bias in information estimates are available. However, these estimates may not be applicable for at least two reasons: the asymptotic regime may not be reached because the dataset size is too small, or, the analysis method (e.g., the metric space approach) does not treat each response independently. But even in these circumstances, use of shuffled datasets can determine whether the estimated amount of information, viewed as a nonparametric measure of correlation between input and output, is greater than chance (Victor & Purpura, 1996b).

For multichannel data sets, additional simple surrogate datasets are useful. To determine whether correlations between responses can be explained on the basis of common driving by a stimulus, rather than neuronal interconnections, the "shift-predictor", or more generally, the "shuffle-correction" (Perkel, Gerstein & Moore, 1967) can be used. Here, the individual channels of the responses to a particular stimulus are re-grouped within that stimulus.

### *Maximum-entropy methods: single neurons*

For continuous signals, it is often of interest to determine whether observed dynamical features of a neural signal are fully explained by its second-order correlation properties. If so, then the signals are consistent with a (perhaps multichannel) Gaussian white noise that has been

linearly filtered. If not, nonlinear dynamics must be present. This kind of question can be addressed by re-analyzing surrogate data that is constrained to have the same second-order correlation structure as the original data, and has higher-order correlations determined by the maximizing the entropy under these constraints. Such surrogate data are conveniently created by randomizing phases but preserving amplitudes (Schiff et al., 1996, Theiler et al., 1991, Theiler & Rapp, 1996).

The maximum-entropy idea is readily extended to spike trains, providing natural "coordinates" for response distributions in an elegant formal framework(Amari, 2001, Nakahara & Amari, 2002).

This approach can be used to formalize questions related to the important notion of "temporal coding" (Théunissen & Miller, 1995). Informally, "temporal coding" means that the time course of neural activity, and not just the number of spikes, carries information. Here, the term "time course" includes not only the time-dependent firing rate, but also more subtle features of the firing pattern, such as interval structure, or highly reproducible "triplets" of spikes (Lestienne & Tuckwell, 1997). These aspects of firing pattern can be distinguished by comparing the information-theoretic analysis of the original data with analysis of surrogate datasets that match the observed responses in terms of the time-dependent firing rate, but are otherwise unconstrained. Such surrogates are inhomogeneous Poisson processes, whose firing rate is determined by the observed post-stimulus histogram, and are thus examples of constrained maximum entropy processes.

Surrogate datasets can be further constrained to match the original data in terms of spike counts on each trial. Such datasets can easily be created by "exchange resampling" (Victor & Purpura, 1996b). A further refinement constrains the interspike interval distribution as well

(Oram, Wiener, Lestienne & Richmond, 1999). These strategies have been used to show that precisely-timed triplets of spikes do not contribute to information transfer (Baker & Lemon, 2000, Oram et al., 1999).

### *Maximum-entropy methods: multiple neurons*

Application of maximum-entropy principles to analysis of multineuronal activity can lead to substantial insights. It is impossible to determine the stimulus-response distribution empirically for an entire neuronal population, since the dimensionality of this distribution is very large. However, a practical approach is to measure the individual stimulus-conditioned response probabilities of each neuron, and to assume that the full stimulus-conditioned population response distribution is its maximum-entropy extension. This approach is equivalent to approximating the stimulus-conditioned population response distribution as a product of individual stimulus-conditioned response distributions. In the retina, an important model system, the error incurred by this approximation appears to be quite small (Nirenberg et al., 2001, Nirenberg & Latham, 2003).

Maximum-entropy methods can provide a compact and comprehensible representation of the correlation structure of the spontaneous activity of neuronal populations. In two recent studies(Schneidman, Berry, Segev & Bialek, 2006, Shlens, Field, Gauthier, Grivich, Petrusca, Sher, Litke & Chichilnisky, 2006), maximum-entropy extension from measured pairwise correlations accounted for the bulk of high-order multineuronal correlations. Combining these strategies (i.e., fashioning maximum-entropy distributions from a combination of stimulus-conditioned single-neuron distributions and low-order response correlations) may provide a powerful way to analyze and understand population coding.

**SUMMARY**

Understanding how neurons and neural populations represent information requires a combined experimental and theoretical approach. Shannon's information theory provides the appropriate theoretical framework. In the Shannon approach, no assumptions are made concerning the relationships of the coding elements to each other, or to the objects being represented. This generality is a fundamental aspect of the strength and elegance of the Shannon approach. However, its generality also engenders challenges to its use in experimental neuroscience, for two reasons. First, neural activity is characterized by a wide range of timescales, from the submillisecond range (e.g., the intrinsic precision of spike generation) to times on the order of a second (e.g., inhibitory synaptic potentials). Thus, without the imposition of additional hypotheses as to the nature of the code, the number of codes that need to be explored are far too great for a direct experimental attack. Second, the relationship of the neural activity to the objects being represented *is* of interest. This relationship is important to understand the mechanism of coding, and because neural activity must not only convey information, but also manipulate it.

These considerations provide both a (retrospective) rationale for, and a unified view of, many approaches that have recently been advanced for the analysis of neural coding. The approaches described here vary in the assumptions made concerning the neural code, ranging from virtually no assumptions, to merely exploiting the continuity of time, to positing very specific forms for the relationship between coding elements. Making such assumptions allows analysis to be carried out on datasets that are typically available from experiments. By assuming that the codes have structure, these approaches also allow for identification of a systematic relationship between the objects and the code – i.e., a representation. However, imposition of

assumptions necessarily increases the risk that the relevant neural codes are simply not being considered. At present, neuroscientists can grapple with this problem by exploring a variety of approaches, each with its own set of assumptions, and hoping that the biological conclusions are relatively independent of the methodology chosen. It remains to be seen whether a more systematic and fundamentally satisfying theoretical approach can be fashioned.

**ACKNOWLEDGEMENTS**

The author thanks Daniel Gardner, Simon Schultz, Chip Levy, Alex Casti, and Keith Purpura and for helpful comments and suggestions. This work is supported in part by NIH EY9314 to JV and MH68012 to Daniel Gardner.

## REFERENCES

Abbott, L.F., Varela, J.A., Sen, K., & Nelson, S.B. (1997). Synaptic depression and cortical gain control. *Science, 275* (5297), 220-224.
Abeles, M., & Prut, Y. (1996). Spatio-temporal firing patterns in the frontal cortex of behaving monkeys. *J Physiol Paris, 90* (3-4), 249-250.
Amari, S.-I. (2001). Information geometry on hierarchy of probability distributions. *IEEE Transactions on Information Theory, 47*, 1701-1711.
Amigo, J.M., Szczepanski, J., Wajnryb, E., & Sanchez-Vives, M. (2004). Estimating the entropy rate of spike trains via Lempel-Ziv complexity. *Neural Comput, 16*, 717-736.
Aronov, D. (2003). Fast algorithm for the metric-space analysis of simultaneous responses of multiple single neurons. *J Neurosci Methods, 124* (2), 175-179.
Aronov, D., Reich, D.S., Mechler, F., & Victor, J.D. (2001). Multidimensional representation of spatial phase in V1. *Invest. Ophth. Vis. Sci., 42*, 405.
Baker, S.N., & Lemon, R.N. (2000). Precise spatiotemporal repeating patterns in monkey primary and supplementary motor areas occur at chance levels. *J Neurophysiol, 84* (4), 1770-1780.
Berry, M.J., Warland, D.K., & Meister, M. (1997). The structure and precision of retinal spike trains. *Proc Natl Acad Sci U S A, 94* (10), 5411-5416.
Bialek, W., Rieke, F., Ruyter van Steveninck, R.R.d., & Warland, D. (1991). Reading a neural code. *Science, 252* (5014), 1854-1857.
Bliss, T.V., & Collingridge, G.L. (1993). A synaptic model of memory: long-term potentiation in the hippocampus. *Nature, 361* (6407), 31-39.
Bourne, H., & Nicoll, R. (1993). Molecular machines integrate coincident synaptic signals. *Cell 72/Neuron 10,* (suppl), 65-85.
Buracas, G.T., Zador, A.M., DeWeese, M.R., & Albright, T.D. (1998). Efficient discrimination of temporal patterns by motion-sensitive neurons in primate visual cortex. *Neuron, 20* (5), 959-969.
Carlton, A.G. (1969). On the bias of information estimates. *Psychol Bull, 71*, 108-109.
Chao, A., & Shen, T.-J. (2003). Nonparametric estimate of Shannon's index of diversity when there are unseen species in a sample. *Environmental and Ecological Statistics, 10*, 429-443.
Chee-Orts, M.N., & Optican, L.M. (1993). Cluster method for analysis of transmitted information in multivariate neuronal data. *Biol Cybern, 69* (1), 29-35.
Cline, H. (1997). Coincidence detection in the nervous system. *Trends Neurosci, 19* (12), 566-567.
Dan, Y., Alonso, J.M., Usrey, W.M., & Reid, R.C. (1998). Coding of visual information by precisely correlated spikes in the lateral geniculate nucleus. *Nat Neurosci, 1* (6), 501-507.
Di Lorenzo, P.M., & Victor, J.D. (2003). Taste response variability and temporal coding in the nucleus of the solitary tract of the rat. *J. Neurophysiol., 90*, 1418-1431.
Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. (Philadelphia: SIAM.
Efron, B., & Tibshirani, R.J. (1998). An Introduction to the Bootstrap. *Monographs on Statistics and Applied Probability*, 57 (p. 436). Boca Raton, FL: Chapman & Hall/CRC Press.

Farach, M., Noordewier, M., Savari, S., Shepp, L., Wyner, A., & Ziv, J. (1995). On the entropy of DNA: Algorithms and measurements based on memory and rapid convergence. *Proc. Sixth Annual ACM-SIAM Symposium on Discrete Algorithms* (pp. 48-57).
Gawne, T.J. (2000). The simultaneous coding of orientation and contrast in the responses of V1 complex cells. *Exp Brain Res, 133* (3), 293-302.
Grassberger, P. (1988). Finite sample corrections to entropy and dimension estimates. *Phys Lett. A, 128*, 369-373.
Grassberger, P., & Procaccia, I. (1983). Measuring the strangeness of strange attractors. *Physica D, 9*, 189-208.
Gray, C.M., & Singer, W. (1989). Stimulus-specific neuronal oscillations in orientation columns of cat visual cortex. *Proc Natl Acad Sci U S A, 86* (5), 1698-1702.
Hirata, Y., & Mees, A.I. (2003). Estimating topological entropy via a symbolic data compression technique. *Phys Rev E Stat Nonlin Soft Matter Phys, 67* (2 Pt 2), 026205.
Jarvis, M.R., & Mitra, P.P. (2001). Sampling properties of the spectrum and coherency of sequences of action potentials. *Neural Comput, 13* (4), 717-749.
Johnson, D.H., Gruner, C.M., Baggerly, K., & Seshagiri, C. (2001). Information-theoretic analysis of neural coding. *J Comput Neurosci, 10* (1), 47-69.
Kennel, M.B., Shlens, J., Abarbanel, H.D., & Chichilnisky, E.J. (2005). Estimating entropy rates with Bayesian confidence intervals. *Neural Comput, 17* (7), 1531-1576.
Kontoyiannis, I., Algoet, P.H., Suhov, Y.M., & Wyner, A.J. (1998). Nonparametric entropy estimation for stationary processes and random fields, with applications to English text. *IEEE Trans. Inform. Theory, 44*, 1319-1327.
Kozachenko, L.F., & Leonenko, N.N. (1987). Sample estimate of the entropy of a random vector. *Problemy Peredachi Informatsii, 23* (2), 9-16.
Kraskov, A., Stogbauer, H., & Grassberger, P. (2004). Estimating mutual information. *Phys Rev E Stat Nonlin Soft Matter Phys, 69* (6 Pt 2), 066138.
Kreiman, G., Krahe, R., Metzner, W., Koch, C., & Gabbiani, F. (2000). Robustness and variability of neuronal coding by amplitude-sensitive afferents in the weakly electric fish eigenmannia. *J Neurophysiol, 84* (1), 189-204.
Krichevsky, R., & Trofimov, V. (1981). The performance of universal coding. *IEEE Trans. Info. Theory, IT-27*, 199-207.
Lestienne, R., & Tuckwell, H.C. (1997). The significance of precisely replicating patterns in mammalian CNS spike trains. *Neurosci, 82* (2), 315-336.
Levy, W.B. (2000). Experiences, thoughts, and conjectures on implementing a Lempel-Ziv-type algorithm to measure information in a spike train. *Neural Information Processing Systems Workshop on Information and Statistical Structure in Spike Trains* (Breckenridge, CO.
London, M., Schreibman, A., Hausser, M., Larkum, M.E., & Segev, I. (2002). The information efficacy of a synapse. *Nat Neurosci, 5* (4), 332-340.
Luck, S.J., Chelazzi, L., Hillyard, S.A., & Desimone, R. (1997). Neural mechanisms of spatial selective attention in areas V1, V2, and V4 of macaque visual cortex. *J Neurophysiol, 77* (1), 24-42.
Markram, H., Lubke, J., Frotscher, M., & Sakmann, B. (1997). Regulation of synaptic efficacy by coincidence of postsynaptic APs and EPSPs. *Science, 275* (5297), 213-215.
Marmarelis, P.Z., & Marmarelis, V.Z. (1978). Analysis of Physiological Systems: The White-Noise Approach. *Computers in Biology and Medicine* (p. 487). New York: Plenum.

McClurkin, J.W., Optican, L.M., Richmond, B.J., & Gawne, T.J. (1991). Concurrent processing and complexity of temporally encoded neuronal messages in visual perception. *Science, 253* (5020), 675-677.
McFadden, J.A. (1965). The entropy of a point process. *Journal of the Society for Industrial and Applied Mathematics, 13* (4), 988-994.
Meister, M., Lagnado, L., & Baylor, D.A. (1995). Concerted signaling by retinal ganglion cells. *Science, 270* (5239), 1207-1210.
Mel, B.W. (1993). Synaptic integration in an excitable dendritic tree. *J Neurophysiol, 70* (3), 1086-1101.
Middlebrooks, J.C., Clock, A.E., Xu, L., & Green, D.M. (1994). A panoramic code for sound location by cortical neurons. *Science, 264* (5160), 842-844.
Miller, G.A. (1955). Note on the bias on information estimates. *Information Theory in Psychology: Problems and Methods, II-B*, 95-100.
Mitra, P.P., & Pesaran, B. (1999). Analysis of dynamic brain imaging data. *Biophys J, 76* (2), 691-708.
Nakahara, H., & Amari, S. (2002). Information-geometric measure for neural spikes. *Neural Comput, 14* (10), 2269-2316.
Nemenman, I., Bialek, W., & de Ruyter van Steveninck, R. (2004). Entropy and information in neural spike trains: progress on the sampling problem. *Phys Rev E Stat Nonlin Soft Matter Phys, 69* (5 Pt 2), 056111.
Nirenberg, S., Carcieri, S.M., Jacobs, A.L., & Latham, P.E. (2001). Retinal ganglion cells act largely as independent encoders. *Nature, 411* (6838), 698-701.
Nirenberg, S., Jacobs, A., Fridman, G., Latham, P., Douglas, R., Alam, N., & Prusky, G. (2006). Ruling out and ruling in neural codes. *Journal of Vision, 6* (6), 889a.
Nirenberg, S., & Latham, P.E. (2003). Decoding neuronal spike trains: how important are correlations? *Proc Natl Acad Sci U S A, 100* (12), 7348-7353.
Optican, L.M., & Richmond, B.J. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. III. Information theoretic analysis. *J Neurophysiol, 57* (1), 162-178.
Oram, M.W., Wiener, M.C., Lestienne, R., & Richmond, B.J. (1999). Stochastic nature of precisely timed spike patterns in visual system neuronal responses. *J Neurophysiol, 81* (6), 3021-3033.
Paninski, L. (2003). Estimation of entropy and mutual information. *Neural Computation, 15*, 1191.
Panzeri, S., Petersen, R.S., Schultz, S.R., Lebedev, M., & Diamond, M.E. (2001). The role of spike timing in the coding of stimulus location in rat somatosensory cortex. *Neuron, 29* (3), 769-777.
Panzeri, S., & Schultz, S.R. (2001). A unified approach to the study of temporal, correlational, and rate coding. *Neural Comput, 13* (6), 1311-1349.
Panzeri, S., Schultz, S.R., Treves, A., & Rolls, E.T. (1999). Correlations and the encoding of information in the nervous system. *Proc R Soc Lond B Biol Sci, 266* (1423), 1001-1012.
Perkel, D.H., Gerstein, G.L., & Moore, G.P. (1967). Neuronal spike trains and stochastic point processes. II. Simultaneous spike trains. *Biophys J, 7* (4), 419-440.
Petersen, R.S., Panzeri, S., & Diamond, M.E. (2001). Population coding of stimulus location in rat somatosensory cortex. *Neuron, 32* (3), 503-514.

Pfurtscheller, G., & Andrew, C. (1999). Event-Related changes of band power and coherence: methodology and interpretation. *J Clin Neurophysiol, 16* (6), 512-519.
Quiroga, R.Q., Rosso, O.A., Basar, E., & Schurmann, M. (2001). Wavelet entropy in event-related potentials: a new method shows ordering of EEG oscillations. *Biol Cybern, 84* (4), 291-299.
Rapp, P.E., Zimmerman, I.D., Vining, E.P., Cohen, N., Albano, A.M., & Jimenez-Montano, M.A. (1994). The algorithmic complexity of neural spike trains increases during focal seizures. *J Neurosci, 14* (8), 4731-4739.
Reich, D.S., Mechler, F., & Victor, J.D. (2000). Temporal coding of contrast in primary visual cortex: when, what, and why. *J Neurophysiol, submitted*
Reich, D.S., Mechler, F., & Victor, J.D. (2001a). Formal and attribute-specific information in primary visual cortex. *J Neurophysiol, 85* (1), 305-318.
Reich, D.S., Mechler, F., & Victor, J.D. (2001b). Independent and redundant information in nearby cortical neurons. *Science, 294* (5551), 2566-2568.
Reinagel, P., & Reid, R.C. (2000). Temporal coding of visual information in the thalamus. *J Neurosci, 20* (14), 5392-5400.
Reynolds, J.H., Pasternak, T., & Desimone, R. (2000). Attention increases sensitivity of V4 neurons. *Neuron, 26* (3), 703-714.
Richmond, B.J., & Optican, L.M. (1987). Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex. II. Quantification of response waveform. *J Neurophysiol, 57* (1), 147-161.
Rieke, F., Warland, D., de Ruyter van Steveninck, R., & Bialek, W. (1997). Spikes: Exploring the Neural Code. (Cambridge, MA: MIT Press.
Rissanen, J. (1989). Stochastic complexity in statistical inquiry. *World Scientific series in computer science ; v. 15* (pp. iii, 177 p.). Singapore ; Teaneck, NJ: World Scientific.
Rodriguez, E., George, N., Lachaux, J.P., Martinerie, J., Renault, B., & Varela, F.J. (1999). Perception's shadow: long-distance synchronization of human brain activity. *Nature, 397* (6718), 430-433.
Roelfsema, P.R., Lamme, V.A., & Spekreijse, H. (2004). Synchrony and covariation of firing rates in the primary visual cortex during contour grouping. *Nat Neurosci, 7* (9), 982-991.
Ruyter van Steveninck, R.R.d., Lewen, G.D., Strong, S.P., Koberle, R., & Bialek, W. (1997). Reproducibility and variability in neural spike trains. *Science, 275* (5307), 1805-1808.
Salzman, C.D., & Newsome, W.T. (1994). Neural mechanisms for forming a perceptual decision. *Science, 264* (5156), 231-237.
Samonds, J.M., & Bonds, A.B. (2004). From another angle: Differences in cortical coding between fine and coarse discrimination of orientation. *J Neurophysiol, 91* (3), 1193-1202.
Samonds, J.M., Zhou, Z., Bernard, M.R., & Bonds, A.B. (2006). Synchronous activity in cat visual cortex encodes collinear and cocircular contours. *J Neurophysiol, 95* (4), 2602-2616.
Schiff, N.D., Kalik, S.F., & Purpura, K.P. (2000). Episodic dynamics of cortical processing in the ventral stream during free-viewing: Analysis of local field potentials in striate/extrastriate and inferotemporal cortices. *Soc Neurosci Abstr, 26*, 1199 (#1448.1193).
Schiff, N.D., Kalik, S.F., & Purpura, K.P. (2001). Sustained activity in the central thalamus and extrastriate areas during attentive visuomotor behavior: correlation of single unit activity and local field potentials. *Society for Neuroscience Abstracts, 27*, 1910.
Schiff, S.J., Aldroubi, A., Unser, M., & Sato, S. (1994). Fast wavelet transformation of EEG. *Electroencephalogr Clin Neurophysiol, 91* (6), 442-455.

Schiff, S.J., So, P., Chang, T., Burke, R.E., & Sauer, T. (1996). Detecting dynamical interdependence and generalized synchrony through mutual prediction in a neural ensemble. *Phys Rev E, 54* (6), 6708-6724.
Schneidman, E., Berry, M.J., 2nd, Segev, R., & Bialek, W. (2006). Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature, 440* (7087), 1007-1012.
Schultz, S.R., & Panzeri, S. (2001). Temporal correlations and neural spike train entropy. *Phys Rev Lett, 86* (25), 5823-5826.
Schurmann, T., & Grassberger, P. (1996). Entropy estimation of symbol sequences. *Chaos, 6* (3), 414-427.
Sellers, P. (1974). On the theory and computation of evolutionary distances. *SIAM J. Appl. Math., 26*, 787-793.
Sen, K., Jorge-Rivera, J.C., Marder, E., & Abbott, L.F. (1996). Decoding synapses. *J Neurosci, 16* (19), 6307-6318.
Shadlen, M.N., & Newsome, W.T. (1998). The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *J Neurosci, 18* (10), 3870-3896.
Shannon, C.E., & Weaver, W. (1949). The Mathematical Theory of Communication. (Urbana: U of Illinois P.
Shlens, J., Field, G.D., Gauthier, J.L., Grivich, M.I., Petrusca, D., Sher, A., Litke, A.M., & Chichilnisky, E.J. (2006). Probing the structure of multi-neuron firing patterns in the primate retina using maximum entropy methods. *CoSyNe* (Salt Lake City, Utah.
Shlens, J., Kennel, M.B., Abarbanel, H.D., & Chichilnisky, E.J. (2006). Estimating information rates with confidence intervals in neural spike trains. *Neural Comput, in press*
Slepian, D. (1976). On bandwidth. *Proceedings of the IEEE, 64*, 292-300.
Softky, W. (1994). Sub-millisecond coincidence detection in active dendritic trees. *Neuroscience, 58* (1), 13-41.
Softky, W.R., & Koch, C. (1993). The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J Neurosci, 13* (1), 334-350.
Solomonoff, R. (1964). A formal theory of inductive inference. Part I. *Information and Control, 7*, 1-22.
Stopfer, M., Bhagavan, S., Smith, B.H., & Laurent, G. (1997). Impaired odour discrimination on desynchronization of odour-encoding neural assemblies. *Nature, 390* (6655), 70-74.
Strong, S.P., Koberle, R., Ruyter van Steveninck, R.R., & Bialek, W. (1998). Entropy and information in neural spike trains. *Phys Rev Lett, 80* (1), 197-200.
Tallon, C., Bertrand, O., Bouchet, P., & Pernier, J. (1995). Gamma-range activity evoked by coherent visual stimuli in humans. *Eur J Neurosci, 7* (6), 1285-1291.
Theiler, J., Galdrikian, B., Longtin, A., & Farmer, J. (1991). Testing for nonlinearity in time series: the method of surrogate data. *Los Alamos National Laboratory Preprint, LA-UR-91-3343*
Theiler, J., & Rapp, P.E. (1996). Re-examination of the evidence for low-dimensional, nonlinear structure in the human electroencephalogram. *Electroencephalogr Clin Neurophysiol, 98* (3), 213-222.
Théunissen, F., & Miller, J.P. (1995). Temporal encoding in nervous systems: a rigorous definition. *J Comput Neurosci, 2* (2), 149-162.
Théunissen, F., Roddey, J.C., Stufflebeam, S., Clague, H., & Miller, J.P. (1996). Information theoretic analysis of dynamical encoding by four identified primary sensory interneurons in the cricket cercal system. *J Neurophysiol, 75* (4), 1345-1364.

Thomson, D.J. (1982). Spectrum estimation and harmonic analysis. *Proc IEEE, 70* (9), 1055-1096.
Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Computation, 7*, 399-407.
Usrey, W.M., Reppas, J.B., & Reid, R.C. (1998). Paired-spike interactions and synaptic efficacy of retinal inputs to the thalamus. *Nature, 395* (6700), 384-387.
van der Togt, C., Kalitzin, S., Spekreijse, H., Lamme, V.A., & Super, H. (2006). Synchrony dynamics in monkey V1 predict success in visual detection. *Cereb Cortex, 16* (1), 136-148.
Victor, J.D. (2002). Binless strategies for estimation of information from neural data. *Phys. Rev. E, 66*, 51903.
Victor, J.D. (2005). Spike train metrics. *Curr Opin Neurobiol, 15* (5), 585-592.
Victor, J.D., & Purpura, K.P. (1996a). Nature and precision of temporal coding in visual cortex: a metric- space analysis. *J Neurophysiol, 76* (2), 1310-1326.
Victor, J.D., & Purpura, K.P. (1996b). Nature and precision of temporal coding in visual cortex: A metric-space analysis. *J Neurophysiol, 76* (2), 1310-1326.
Victor, J.D., & Purpura, K.P. (1997). Metric-space analysis of spike trains: theory, algorithms and application. *Network, 8*, 127-164.
Willems, F.M.J., Shtarkov, Y.M., & Tjalkens, T.J. (1995). The context-tree weighting method: basic properties. *IEEE Trans. Inf. Theory, 41*, 653-664.
Wolpert, D.H., & Wolf, D.R. (1995). Estimating functions of probability distributions from a finite set of samples. *Physical Review. E. Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics, 52* (6), 6841-6854.
Wu, M., David, S.V., & Gallant, J. (2006). Complete functional characterization of sensory neurons by system identification. *Ann. Rev. Neurosci., 29*
Wyner, A.D., & Ziv, J. (1989). Some asymptotic properties of entropy of a stationary ergodic data source with applications to data compression. *IEEE Trans. Inform. Theory, 35*, 1250-1258.

## FIGURE AND TABLE LEGENDS

*Table 1.*

Characteristics of several methods for the information-theoretic analysis of neural data.