IN PRESS
To appear in Mechanisms of information process ing in the Brain:
Encoding of information in neural populations and networks
Holscher and Munk (Eds) Cambridge University Press 2007

# Measuring distributed properties of neural representations beyond the decoding of local variables — implications for cognition

Adam Johnson,* Jadin C. Jackson,* A. David Redish

22/Dec/2006

## Introduction

Neural representations are distributed. This means that more information can be gleaned from neural ensembles than from single cells. Modern recording technology allows the simultaneous recording of large neural ensembles (of more than 100 cells simultaneously) from awake behaving animals. Historically, the principal means of analyzing representations encoded within large ensembles has been to measure the immediate accuracy of the encoding of behavioral variables ("reconstruction"). In this chapter, we will argue that measuring immediate reconstruction only touches the surface of what can be gleaned from these ensembles.

We will discuss the implications of distributed representation, in particular, the usefulness of measuring self-consistency of the representation within neural ensembles. Because representations are distributed, neurons in a population can agree or disagree on the value being represented. Measuring the extent to which a firing pattern matches expectations can provide an accurate assessment of the self-consistency of a representation. Dynamic changes in the self-consistency of a representation are potentially indicative of cognitive processes. We will also discuss the implications of representation of non-local (non-immediate) values for cognitive processes. Because cognition occurs at fast time scales, changes must be detectable at fast (ms, tens of ms) timescales.

## Representation

As an animal interacts with the world, it encounters various problems for which it must find a solution. The description of the world and the problems encountered within it play a fun-

*Adam Johnson and Jadin Jackson contributed equally to this work.

damental role in how an animal behaves and finds a solution. Sensory and memory processes within the brain provide a description of the world and within that description the brain's decision making processes must select some course of action or behavior.*

The resulting open question is how the information about the world is represented and organized across these brain areas. The use of information from the world in behavior involves two critical processes. The first process is appropriate transformation of information about the world into a representation that is relevant and useful for behavior. The second process is the projection of that representation onto a behavior that allows the animal to interact with its world.

We will call the representation of the world within the brain or any transformation of that representation toward behavior (even if the behavior is not executed) a *neural representation*. This definition is intentionally broad such that the operations underlying directly observable behavior *and* covert mental activities can be considered.

## Encoding (Tuning curves)

What makes a neuron fire? The question can be asked with respect to the neuron's immediate environment — its afferents and ion channels — and with respect to the world beyond. Answering the latter question requires knowing what information is *encoded* by the neuron. An *encoding model* describes a hypothesis relating the information represented by the cell (sensory, perceptual, motivational, motor, etc.) to the activity of a single neuron. The hypothesized relationship between the encoded information $x$ and the neural activity, typically considered in terms of spikes, $s$ can be written as the function $p(s(t)) = T(x(t))$ where $p(s(t))$ probability of a spike at time $s(t)$.† This definition is easily extended to include both preceding experience and planned future behaviors in the encoded information $x$. For simplicity, the present discussion neglects identification of the precise temporal offset in describing the relationship between s(t) and x(t).

These encoding models have classically been found by standard *tuning curves*. More recently, these encoding models have been stated in terms of Shanon information theory, identifying the mutual information between behavioral variables and spike timing (Rieke et al., 1997; Dayan and Abbott, 2001). Other encoding definitions have been based on *linear filter kernels*, which reflect the recent history of variable $x$ in the firing of a cell's spikes (Bialek et al.,

*The decision to not act is still a decision made.

†Of course, the actual activity is also a function of the history of spiking of the neuron (e.g. neurons show refractory periods, bursting, and other history-dependent processes).

1991) (or bursts (Kepecs and Lisman, 2003)).[‡] These encoding models can be measured relative to any available behavioral variable, whether it be immediate sensory input, such as the frequency of an auditory tone, an immediate motor output, such as the target of a saccade or the direction of a reach, or a cognitive variable, such as the location of an animal in the environment.

## Decoding (Reconstruction)

Because the variability of a single cell is usually insufficient to fully describe the entire space of encoded information, information is generally encoded across a population of neurons that differ in their tuning curves (often described by a *family of tuning curves*), such as retinotopic Gaussians or place fields. If information is consistently represented across a population of neurons, then it should be possible to infer the expectation of the variable $x$ by examining the neural activity across the population. This inference can be made using Bayes' rule

$$p(x,s) = p(x|s)p(s) = p(s|x)p(x) \tag{1}$$

where $p(s|x)$ is the probability of observing some set of neural activities given the variable of interest and $p(x|s)$ is the probability of the variable of interest given some set of neural activity. This means that the variable $x$ can be decoded from the neural activity across the population $s$ by

$$p(x|s) = p(s|x)p(x)/p(s) \tag{2}$$

The probability $p(x|s)$ describes how information can be read out or *decoded* from the network. What should be clear from this simple account is that decoding critically depends on the *encoding model*, $p(s|x)$.

The term $s$ in equation 2 reflects the pattern of neural activity across the entire population of cells at time $t$. This analysis thus requires sufficient data to infer the probability density function across an $n$-dimensional space ($n$, where $n$ is the number of cells in the ensemble). For even moderately-sized ensembles, appropriate sampling of $s$ thus requires an inordinate amount of data due to the curse of dimensionality. In many situations, it is convenient to assume that the activity of each cell is conditionally independent, relative to the represented variable $x$ (Zhang et al., 1998; Brown et al., 1998), so that

$$p(x|s) = p(x) \prod_{i \in \text{cells}} \frac{p(s_i|x)}{p(s_i)} \tag{3}$$

[‡]Kernal based methods explain the observed neural activity in terms of both the represented information and the neuronal dynamics of the cell itself. The generative method below effectively extends this perspective to include unobserved variables that can only be determined by examining ensembles with decoding.

However, the validity of this assumption is still controversial (Nirenberg and Latham, 2003; Schneiderman et al., 2003; Averbeck et al., 2006).

Although Bayes' rule (Eq. 2) provides an optimal solution for decoding, even the simplied version (Eq. 3) is often not computationally tractable. As a result, several other non-probability based methods have been developed for decoding (e.g. template matching, Wilson and McNaughton, 1993; Averbeck et al., 2003a; Zhang et al., 1998, linearly weighted averaging, Georgopoulos et al., 1983; Salinas and Abbott, 1994; Zhang et al., 1998). These can be considered as reduced forms of Bayes' rule (Dayan and Abbott, 2001).

## Non-local reconstruction (Memory and cogntition)

While neural activity is typically measured and discussed in terms of an observable external variable $x(t)$: $p(s(t)) = T(x(t))$, a more inclusive statement is that neural activity reflects an internal representation of this variable. That internal representation can potentially deviate from the external world. This point is particularly important when investigating processes in which cognition potentially plays a role; one of the hallmarks of cognitive processing is the connection of the observable world with the animal's externally invisible goals or motivations (Tulving, 1983, 2001, 2002; Suddendorf and Busby, 2003; Gray, 2004; Ferbinteanu et al., 2006; Johnson and Redish, 2006).

During normal navigation, as rats perform active behavioral tasks on an environment, the first order information encoded within hippocampal pyramidal cells is the location of the animal (O'Keefe and Nadel, 1978; Redish, 1999). However, when rats are sleeping or when they pause at feeder sites to eat or groom, the hippocampus changes state, and the hippocampal firing reflects internal dynamics rather than its primary inputs (O'Keefe and Nadel, 1978; Wilson and McNaughton, 1994; Chrobak and Buzsáki, 1994, 1996; Ylinen et al., 1995; Csicsvari et al., 1999; Chrobak et al., 2000). Cell firing during subsequent sleep states reflects recently experienced memories rather than the current location of the animal (Wilson and McNaughton, 1994; Kudrimoti et al., 1999; Nádasdy et al., 1999; Sutherland and McNaughton, 2000; Hoffmann and McNaughton, 2002; Lee and Wilson, 2002). Reconstruction during non-attentive waking states reveals representations of *non-local information* (Jensen and Lisman, 2000; Jackson et al., 2006, see Fig. 1).

[Figure 1 about here.]

The slow dynamics in which reconstruction tracks behavior (Wilson and McNaughton, 1994; Zhang et al., 1998; Brown et al., 1998) and the fast dynamics of replay (e.g. Fig. 1) are examples of different *information processing modes* (Harris-Warrick and Marder, 1991; Hasselmo and Bower, 1993; Buzsáki, 1989; Redish, 1999). These two modes occur in recognizably distinct

brain states, characterized by distinct local field potential frequencies. The neural firing patterns of both projection (pyramidal) cells and interneurons change between modes, as do the neuromodulators present (Vanderwolf, 1971; O'Keefe and Nadel, 1978; Hasselmo and Bower, 1993; Somogyi and Klausberger, 2005). These modes are thought to be differentially involved in learning, storage, and recall (Buzsáki, 1989; Hasselmo and Bower, 1993; Redish, 1999). Later in this chapter, we will discuss the implications of multiple generative models $p_M(s|x)$ for understanding these multiple information processing modes. In order to differentiate between models, we first need to address the question of *self-consistency*.

## Self-consistency (Coherency)

While the development of ensemble based reconstruction methods such as those described above has allowed us to probe more deeply into the brain's processing of behavioral information, we run the risk of assuming that an animal's brain rigidly adheres to representing the present behavioral status of the animal. In doing so, reconstruction errors are viewed as "noise in the system", ignoring the cognitive questions of memory and recall that are fundamental to the brain's inner workings. For instance, what is recall or confusion and how does the brain represent competing values in ambiguous situations? To answer these questions, we need to consider how units within a network function together to form a coherent representation i.e. one that is internally consistent across all units.

A *coherent* or self-consistent representation is one in which the firing of all neurons in a network conforms to some pattern expected from observations during normal (local, baseline) encoding. For instance, if one records from an ensemble of motor cortical cells, one possible model of the network would be to assume that the firing of each neuron is tuned to the direction of movement. This tuning, if it exists, should dictate the distributed pattern of activity across the neurons in the network. If the network is representing a particular direction, all neurons with any tuning to that direction should be firing to some degree specified by their respective tuning curves and neurons that are tuned to very different directions should be firing rarely if at all. In other words, neurons with similar preferred directions should respond similarly if the network is responding in a manner consistent with the data set used to construct the neuronal tuning curves. If this is not true, there is a fundamental difference between the model and the current status of the network. This principle allows for the formulation of a measure of the *coherency* or *self-consistency* of a neural ensemble (Redish et al., 2000; Jackson and Redish, 2003; Jackson, 2006). (See Figure 2.)

[Figure 2 about here.]

Figure 2 B, C, and D show three hypothetical states for a network made up of neurons with tuning curves shaped like the one depicted in Figure 2 A, but centered at even intervals along x. The behavioral variables $\hat{x}_1$ and $\hat{x}_2$ are shown for reference. The pattern in B is consistent with behavioral variable $\hat{x}_2$ but not with $\hat{x}_1$. A reconstruction algorithm would yield value $\hat{x}_2$. If the actual value was $\hat{x}_1$, then reconstruction error $|\hat{x}_2 - \hat{x}_1|$ would be high even as the network state is internally consistent. The left mode of the pattern in C is consistent with behavioral variable $\hat{x}_1$ but neither mode is consistent with $\hat{x}_2$. A vector based reconstruction algorithm would yield value $\hat{x}_2$, while template matching or Bayesian methods would yield $\hat{x}_1$ or the right peak depending on the noise in the system. If the actual value was $\hat{x}_1$, then reconstruction error $|\hat{x}_2 - \hat{x}_1|$ would either be low or high depending on the reconstruction method and the noise in neuronal activity. However, neither reconstruction measure would reveal the underlying representational ambiguity. The state in D is not consistent with either behavioral variable $\hat{x}_1$ or $\hat{x}_2$. However, each reconstruction method would yield a value such as $\hat{x}_1$ or $\hat{x}_2$ even though the underlying state is completely random. Each of these scenarios suggests very different cognitive processes are occurring in this brain network; accessing these processes though an appropriate measure of internal consistency is one primary aim of this chapter.

Redish et al. (2000) first suggested that a mathematical comparison between expected and actual activity patterns could provide useful information about the dynamics of neural processes. They used such a comparison to identify when the hippocampal ensemble transitioned between two spatial reprsentations.

Averbeck and colleagues (Averbeck, 2001; Averbeck et al., 2002, 2003b) recorded from frontal and parietal neural ensembles in a shape-copying task and compared the neural activity patterns during the time monkeys were actually drawing the shapes with the neural activity patterns during the preparatory period. They first calculated the $n$-dimensional tuple of firing rates during each segment of the copying process (e.g. $F_{\text{bottom line of square}} = (f_1, f_2, \ldots, f_n)$, where $f_i$ is the firing rate of cell $i$, and $n$ is number of cells in ensemble). They then measured the Euclidean distance between the firing rate patterns of the ensembles in each 25 ms bin of the preparatory period and the firing rate pattern of the ensemble during each segment of the copying process. They found that at the end of the preparatory period, the first component was most likely to be closest in this distance metric, while the second component was next closest, the third component next, and so forth.

In the limited condition of cosine-tuned neurons (Georgopoulos et al., 1983), one can measure self-consistency by measuring the length of the population vector (Smyrnis et al., 1992, which is a measure of the variance of the circular distribution, Mardia, 1972, Batschelet, 1981). Georgopoulos et al. (1988) used this to measure development of a motor plan during mental

rotation tasks. As the animal developed a motor plan, the length of the population vector increased.

While the comparison process laid out by Jackson and Redish (2003, see below) requires a hypothesized behavioral variable $\hat{x}$ in order to define the expected activity packet $\hat{A}$, the hypothesized behavioral variable does not have to reflect anything about the outside world. It can be based on the estimated representation of a given variable $\hat{x}$. This estimated variable can be found by an internal decoding process (i.e. *reconstruction*). For example, Johnson et al. (2005) recorded neural ensembles of head direction cells from the postsubiculum of rats in a cylinder-foraging task and calculated the coherency of the head direction representation relative to the reconstructed head direction $\hat{\phi}$. Populations that were highly self-consistent were more likely to provide an accurate representation of the world. In other words, actual reconstruction error (e.g. $|\hat{\phi} - \text{actual } \phi|$) was reflected in the self-consistency of the representation, even though the self-consistency could be determined from entirely internal signals. Thus, if down-stream structures used only self-consistent representations for making decisions, then the animal would be more likely to be using accurate representations of the outside world.

### Comparing actual and expected activity patterns

In the following section, we review the results of Jackson and Redish (2003) and show the generality of the results. Further details can be found in the original paper.

Activity packets were defined as the weighted sum of the tuning curves. (Jackson and Redish (2003) showed that normalizing by the average tuning curve made this a linear calculation and simplified subsequent analyses.)

$$A(x,t) = \frac{\sum_k T_k(x) \cdot F_k(t)}{\sum_k T_k(x)} \tag{4}$$

where $k$ ranges over the available cells in the ensemble, $T_k(x)$ is the tuning curve of cell $k$ relative to variable $x$, and $F_k(t)$ is the firing rate of cell $k$ at time $t$. The activity packet is thus a function over both the behavioral (possibly multi-dimensional) variable $x$ and time. The expected activity packet is then defined as the weighted sum of the tuning curves, weighted, not by the actual firing rate of the cells, but rather by the *expected* firing rate of the cells.

$$\hat{A}(x,t) = \frac{\sum_k T_k(x) \cdot E(F_k(t))}{\sum_k T_k(x)} \tag{5}$$

$$= \frac{\sum_k T_k(x) \cdot T_k(\hat{x}(t))}{\sum_k T_k(x)} \tag{6}$$

where $\hat{x}(t)$ is the hypothesized value of variable $x$ at time $t$.

Once the actual activity packet $A(x,t)$ and the expected activity packet $\hat{A}(x,t)$ have been defined, then the self-consistency of the population can be measured by comparing the two packets. We have explored multiple comparison methods, including dot-product (DP, Redish et al., 2000), root-mean-squared-error (RMS, Jackson and Redish, 2003), variance (VAR, Johnson et al., 2005), which can be brought into the same units as RMS by taking the square-root and using the standard-deviation instead (STD, Jackson, 2006). See Jackson (2006) for a review.

$$C_{DP}(t) = \hat{A}(x,t) \cdot A(x,t) \tag{7}$$

$$I_{RMS}(t) = \frac{\sqrt{\int_x (A(x,t) - \hat{A}(x,t))^2 dx}}{\int_x \hat{A}(x,t) dx} \tag{8}$$

$$I_{VAR}(t) = \frac{\mathrm{var}_x(A(x,t) - \hat{A}(x,t))}{\int_x \hat{A}(x,t) dx} \tag{9}$$

$$I_{STD}(t) = \frac{\mathrm{stdev}_x(A(x,t) - \hat{A}(x,t))}{\int_x \hat{A}(x,t) dx} \tag{10}$$

Here, we use a $C$ to denote that the measure $C_{DP}$ measures the consistency, or similarity, between the actual and expected representations; we use $I$ to denote the other measures, which measure inconsistency, or dissimilarity, between the actual and expected activity packets. The integration is done over the entire representational space. We include $C_{DP}$ for completeness, but we have found that $I_{RMS}$ and $I_{STD}$ are the most sensitive and recommend their use for experimental analyses (Jackson, 2006).

Statistically, the self-consistency of the ensemble relative to hypothesized behavioral variable $\hat{x} \in x$ can be defined as the probability of accepting the null hypothesis that the actual and expected activity packets are the same:

$$H_0 : \forall_x \forall_t A(x,t) = \hat{A}(x,t) \tag{11}$$

In practice, we expect the validity of this hypothesis to vary over time and generally measure it as a function of time

$$\text{For a given time } t, H_0(t) : \forall_x A(x,t) = \hat{A}(x,t) \tag{12}$$

The probability of accepting the null hypothesis can be found by empirically determining the probability distribution of the measurement of choice under conditions of stability. *Self-consistency* can then be defined as the deviation from this expected probability distribution. If the measure implemented detects differences between the activity packets, this probability is

equal to the probability of seeing a larger difference between the actual and expected representation given the data in the training set. If this probability is sufficiently small, the actual and expected activity packets are more different than a large majority of the samples in our training set and we can reject the null hypothesis that the actual representation is the same as the expected representation.

### Validation: simulations

Simulations provide a fast, efficient, and, most importantly, controlled means of generating data for the purposes of characterizing ensemble measures. The attractor network used for the simulations was a standard local-excitatory/global-inhibitory network used to model which has extensively studied (Wilson and Cowan, 1973; Amari, 1977; Ermentrout and Cowan, 1979; Kohonen, 1982, 1984; Redish, 1999; Eliasmith and Anderson, 2003; Jackson and Redish, 2003) and has been used to model numerous systems in the brain (Droulez and Berthoz, 1991; Munoz et al., 1991; Arai et al., 1994; Skaggs et al., 1995; Redish et al., 1996; Zhang, 1996; Samsonovich and McNaughton, 1997; Redish and Touretzky, 1998; Redish, 1999; Tsodyks, 1999; Doboli et al., 2000; Laing and Chow, 2001; Goodridge and Touretzky, 2000; Tsodyks, 2005; Wills et al., 2005). Briefly, this network employed symmetric local excitatory connections between neurons with similar preferred directions and global inhibition with periodic boundary conditions. Thus, this network can be thought of as a circular ring of neurons with a stable attractor state consisting of a single mode of active neurons, which could be located anywhere on the ring. This local-excitation/global-inhbition, ring-based attractor network has several useful properties for the study of self-consistency; however, it is important to note that the self-consistency equations above (Eq. 4– 12) make no assumptions about the shape or structure of the tuning curves or the network connectivity (Jackson and Redish, 2003). The only assumption made is that tuning curves are stable over the course of the training and test sets.

**Issue 1: Random Network Firing vs. Stable Activity Mode.** When started from random noise, neurons in a ring attractor will compete until a group of neighbors wins and the network settles to a stable mode of activity at that location (i.e. representing one direction). Neurons with preferred directions near this direction will have higher firing rates than those distant from this direction. Thus, the final mode will be randomly selected given a random input (Wilson and Cowan, 1973; Kohonen, 1977).

In this situation, reconstruction techniques such as the vector mean (Mardia, 1972, also known as the *population vector*, Georgopoulos et al., 1983) always provide an answer and cannot be used to differentiate the random and settled states. In contrast, self-consistency measures differentiate the random and settled states (see Figure 3). Because of the nonlinearities of the measure, $I_{RMS}$ detected the time of settling accurately, displaying a stark

difference between the two states. While in the random state, the self-consistency measurement showed that the random state was significantly different from the expected "bump" of activity ($p < 0.005$). After the network transitioned to a stable representational state, coherency showed a higher probability of match.

[Figure 3 about here.]

**Issue 2: Rotation vs. Jump.** When this system is in a stable state (i.e. representing one direction), and network inputs drive neurons with preferred directions near the represented direction (within 60° in our network), the represented direction will shift toward the input (Redish et al., 1996; Zhang, 1996; Samsonovich and McNaughton, 1997; Redish, 1999). Chaining this extra-network excitation to the represented direction forces the network to rotate continuously. In contrast, when the network inputs drive neurons with preferred directions far from the represented direction (greater than 60° in our network), the system will non-linearly jump to a new direction if the strength of the drive is large enough to overcome the global inhibition (Zhang, 1996; Samsonovich and McNaughton, 1997; Redish, 1999).

Reconstruction showed a smooth transition through intermediate orientations in both the rotation and jump conditions (Figure 4). Reconstruction thus suggested that both of these transitions were simple rotations, yet the dynamics of these two transitions were fundamentally different. $I_{RMS}$, however, detected the difference. In the jump condition, $I_{RMS}$ showed a strong transient increase at the time of transition ($I_{RMS}$: time-steps 562–609, $p < 0.005$), but no corresponding increase during the rotation (time-steps 200–800, $p > 0.005$).

[Figure 4 about here.]

**Issue 3: Ambiguous vs. single valued representations.** If the network is started from a bimodal state (i.e. with inputs at two different directions), the population of neurons representing each input location will compete until the network settles on a single "bump" (Kohonen, 1977, 1982, 1984; Redish, 1999). This can serve as a selection process to resolve ambiguity (Wilson and Cowan, 1973; Kohonen, 1977; Redish and Touretzky, 1998). The location of the specific result will depend on the noise in the network and the difference in direction between the candidate inputs (Redish, 1997).

During the settling process, there will temporarily be more than one bump of activity, one at the center of each input, essentially representing multiple values. Classical reconstruction techniques will be unable to determine whether or not the network has resolved the ambiguity. As shown in Figure 5, self-consistency measures readily identify the resolution of the ambiguity.

[Figure 5 about here.]

### Self-consistency in a Bayesian framework

The self-consistency measures reviewed above are based on observed changes in expected distributions, allowing the generation of statistical $p$-values identifying times when the representation significantly differs from the expected distribution (null hypothesis $H_0$, Eq. 12, above). Recent reconstruction analyses have been based on Bayesian and information measures (Rieke et al., 1997; Zhang et al., 1998; Brown et al., 1998; Zemel et al., 1998; Dayan and Abbott, 2001). It is possible to reinterpret the self-consistency question in terms of Bayesian reconstruction.

In the methods above, the activity packet (Eq. (4)) measures the expected distribution of variable $x$ given the firing rate at time $t$, $F(t)$. In Bayesian terms, the posterior distribution $p(x|s(t))$ (Eq. 2) provides the term analogous to the expected activity packet above. Like the activity packet, this term is a function over the variable $x$ and time. In the methods above, the self-consistency equations (Eqs. 7-10) measure the consistency of this reconstruction process relative to the (implied) model defined by the tuning curves $T(x)$ (defined as the *expected activity packet*, Eq. 6). In the Bayesian formula, after decoding the neural representation, the validity can be determined by estimating how consistent the decoded representation is with the observed neural activity. One of the advantages of using a probabilistic approach in comparison to non-probabilistic estimate based methods is that rather that deriving a single estimate and using that in tandem with the tuning curves to develop the expected activities, the entire posterior distribution $p(x|s(t))$ can be used to generate the expected activity. Recall that the probability distribution (or the derived estimate) is a mapping over the space of $x$. Basic probability theory shows that the product of this probability with the encoding model produces a joint distribution over the spiking activity of the ensemble and the decoded variable. Substituting $p(x|s(t))$ for $p(x)$ in equation 2 gives

$$p(\hat{s}, \hat{x}) = p(\hat{s}|\hat{x})p(\hat{x}|s(t)) \tag{13}$$

where $\hat{x}$ reminds us that $p(\hat{x}|s(t))$ is estimating the distribution of the variable $x$ given the observed spiking patterns, and $\hat{s}$ reminds us that $p(\hat{s}|\hat{x})$ is estimating the firing pattern $s$ given our estimate of the variable $x$. Taking the marginal distribution with respect to the neural activity $s$ (integrating across the variable $x$) provides the probability of a given neural activity set.

$$p_{consistency}(s) = \int_x p(\hat{s}|\hat{x})p(\hat{x}|s) \tag{14}$$

This formulation of consistency indicates the probability that the set of observed neural activity were generated by the decoded neural representation provides a normative method for assessment for competing models, and makes clear that we need to make explicit the genera-

tive model used.

$$p_{consistency}(s|M) = \int_x p_M(s|\hat{x})p(\hat{x}|\hat{s}, M) \tag{15}$$

where $M$ indicates the generative model used. A generative model is *consistent* with the observed neural activity when $p_{consistency}(s|M)$ is high and *inconsistent* when this probability is low. Generative models can be compared using standard methods (such as odds ratios and measured in decibans, Jaynes, 2003).

## Multiple Models in hippocampus

The previous sections have developed the idea that neural activity can represent many types of information – from sensory descriptions of the world to motor planning for behavior and even to the cognitive processes in between – and that the organization of this information is a critical aspect for both interpreting that information from within the system (as downstream neurons) or from outside the system (as experimenters). The suggestion of this approach is that rather than examining a decoded representation with respect to how well it matches an experimentally observed or controlled variable, a decoded representation should be examined on the basis of its intrinsic organization or consistency. In some instances a neural representation may match an observed variable and be well-organized. In other instances, the neural representation may be disorganized. However, in other instances, the neural representation may completely disagree with an observed variable, but remain relatively well-organized. Spatial representations within the hippocampus provide such an example.

Spatial representations in the hippocampus have been explored using a variety of decoding methods (Wilson and McNaughton, 1994; Zhang et al., 1998; Brown et al., 1998; Jensen and Lisman, 2000). Generally, the neural activity of place cells and the decoded spatial representation very well predicts the animal's position within the environment; however, recent studies have shown that place cell activity can remain well-organized even when the decoded representation does not match the animal's position (Skaggs et al., 1996; Lee and Wilson, 2002; Johnson and Redish, 2005, 2006; Foster and Wilson, 2006, see also Figure 1, above).

Within the hippocampus, multiple brain states have been identified based on characteristic local field potential activity (Vanderwolf, 1971; O'Keefe and Nadel, 1978). The hippocampal neural representations of space show different representational dynamics during these multiple brain states. In the theta regime, *phase-precession* describes a dynamic that occurs during each theta cycle in which the spatial representation sweeps through positions recently occupied by the animal to positions that will likely be occupied by the animal (O'Keefe and Recce, 1993; Skaggs et al., 1996). The neural representation during this sweep is time-compressed approximately 10-15 times relative to animal behavior during task performance (Skaggs et al.,

1996). In the large irregular activity (LIA) regime, *route replay* describes a dynamic that occurs during slow wave sleep, following task performance in which neuronal activity present during task performance is replayed (Kudrimoti et al., 1999; Nádasdy et al., 1999; Lee and Wilson, 2002). Spiking activity in sharp-wave replay is time-compressed 40 times relative to animal behavior during the task (Nádasdy et al., 1999; Lee and Wilson, 2002). The observation of both phase precession and sharp wave ripples during awake states (O'Keefe and Nadel, 1978; O'Keefe and Recce, 1993; Foster and Wilson, 2006; Jackson et al., 2006; O'Neill et al., 2006), suggests that the hippocampal representation of space may operate with multiple spatiotemporal dynamics, even during awake behaviors.

Application of self-consistency measures provide a method for examining spatial representations in the hippocampus with respect to multiple spatiotemporal dynamics. Explicit comparison of multiple models of hypothesized representation dynamics allows identification of the underlying dynamical state of the neural representation. A model of the dynamics of a neural representation can be most simply described as a Markov process $p(\hat{x}_t|\hat{x}_{t-1})$, that gives the probability of the representation transitioning from the estimate $\hat{x}_{t-1}$ to a new estimate $\hat{x}_t$. These models can be as simple as a Brownian walk or as complex as a rigidly specified directional flow. Models of representational dynamics are easily added to the Bayes' decoding equation shown above (equation 2) and can be written in terms of a predictive filter.

The term predictive filter refers to the recursive application of a *prediction step* which predicts the temporal evolution of the neural representation given the previous prediction and the proposed dynamical model and a *correction step* which corrects the prediction based on the spikes observed at time $t$. The prediction can be written as

$$p(\hat{x}_t|s_{t-1}) = \int p(\hat{x}_t|\hat{x}_{t-1})p(\hat{x}_{t-1}|s_{t-1})d\hat{x}_{t-1} \tag{16}$$

where $p(\hat{x}_t|\hat{x}_{t-1})$ describes the hypothesized model of representation dynamics and the term $p(\hat{x}_{t-1}|s_{t-1})$ represents the previously predicted neural representation. The correction step can be written as

$$p(\hat{x}_t|s_t) = \frac{p(s_t|\hat{x}_t)p(\hat{x}_t|s_{t-1})}{p(s_t|s_{t-1})} \tag{17}$$

where $p(s_t|s_{t-1})$ is the probability of the neural activity set $s_t$ given the previous set of neural activity $s_{t-1}$ and the term $p(\hat{x}_t|s_{t-1})$ represents a prediction from the previous equation. Predictive filters have been used for decoding neural activity within a variety of brain areas (e.g. Brown et al., 1998; Brockwell et al., 2004; Wu et al., 2006).

After decoding the neural representation for each of the proposed models, the validity of the hypothesized model can be determined by estimating how consistent the decoded representation is with the observed neural activity.

Four generative models were used to perform reconstruction from a hippocampal neural ensemble recorded from the CA1 region of an animal running on a 4T Multiple-T maze (Schmitzer-Torbert and Redish, 2002, 2004). Four generative models were examined, each of which allowed the probability distribution to spread at the prediction step with different rates: $1\times \equiv p(\hat{x}_t|\hat{x}_{t-1})^1$, $15\times \equiv p(\hat{x}_t|\hat{x}_{t-1})^{15}$, $40\times \equiv p(\hat{x}_t|\hat{x}_{t-1})^{40}$, and $99\times \equiv p(\hat{x}_t|\hat{x}_{t-1})^{99}$, where $p(\hat{x}_t|\hat{x}_{t-1})$ was a Gaussian function with $\sigma$ proportional to the average velocity of the rat. The $99\times$ model provided a nearly uniform distribution over the scale of the Multiple-T maze. As can be seen in Figure 6, different models were more consistent at different times, reflecting changes in the neural dynamics.

[Figure 6 about here.]

Model selection was accomplished by calculating an error between the expected spiking activity given the posterior distribution of each filter and observed spiking data as in Equation 15, above. As noted above, the different generative models are hypothesized to reflect different information processing modes. In the hippocampus, these modes are reflected in local field potential signals (O'Keefe and Nadel, 1978; Buzsáki, 1989, 2006; Hasselmo and Bower, 1993; Redish, 1999), thus we hypothesized that the characteristic local field potential power spectrum for each spatiotemporal filter should show similar trends - specifically, the $1\times$ and $15\times$ filters should show increased power within theta frequencies ($7-10$ Hz) while the $40\times$ filter should show increased power within slow wave delta ($2-6$ Hz) and sharp wave ripple ($170-200$ Hz) frequencies. Clear differences were found within slow wave and $\theta$ frequencies. Differences between the characteristic power spectra for each filter were similar between CA1 and CA3. Consistent with previous results (Lee et al., 2004; Leutgeb et al., 2004), subfield analysis found that more dynamic models (e.g. $99\times$) were more often selected in CA1 data sets, relative to the CA3 data sets (see Figure 7).

[Figure 7 about here.]

While generative models have been broadly used to explain and decode neural activity (e.g. Brown et al., 1998; Rao and Ballard, 1999; Lee and Mumford, 2003; Brockwell et al., 2004; Serruya et al., 2004; Wu et al., 2006), one notable distinction should be made between the typical generative model formulation and the present formulation. Because we are concerned with the dynamical regulation of neural representations by cognitive processes, particularly explicit memory retrieval, we suggest that multiple generative models are necessary to explain observed neural activity. A single model is generally not enough because cognition requires the interactive use of dynamical information based on sensory or motor processes *and* planning, motivation or, for lack of another word, cognitive processes. Within each of these types

of representation, cognition modulates an ongoing process. This is precisely the type of modulation that is sought when examining learning and memory or any cognitive processes and mathematically it can be identified as a changes in the model prior $p(x)$. In terms of the generative model, this simply states that there exists a prior non-uniform distribution $p_1(x)$ which better describes the neural activity than a uniform distribution $p_2(x)$. The critical aspect of this formulation is that the goal is to completely generate the observed neural activity. Because of the probabilistic treatment, it becomes straightforward to integrate the contributions of both representation driven aspects of neural activity (e.g. above) and intrinsically driven neural dynamics such as refractory period (Frank et al., 2002).

## Conclusions

A variety of experimental and theoretical results suggest the existence of cognitive processes requiring active memory use in decision making. These processes are non-trivial to assess in human populations using such measures as self-report and are even more difficult to assess in non-human populations. Identifying such cognitive processes in non-human animals will require the development of measures to examine computations underlying these processes. Central to this approach is the development of statistical algorithms for decoding neural representations at multiple time scales and validation or error-assessment methods that allow characterization of cognitive processes related to, but not necessarily mirrored by, directly observable behavior. In this chapter, we have described a method for examining highly dynamic cognitive processes through observation of neural representations with multiple dynamics.

Reconstruction alone cannot be used to infer internal states of an animal's sensory and cognitive networks such as the difference between random firing and well-represented variables. This is particularly important when considering issues of memory and recall. One function of memory is to appropriately link a current experience to a past experience; in the case of the hippocampus, this may mean using the same spatial map as was previously used in an environment. However, a primary usefulness of a memory is in its ability to influence disconnected experiences through recall of past events or episodes. In this case of recall, one would expect that neuronal firing would, by definition, be disconnected from the current behavioral state of the animal. Recall may be detected by reconstruction methods identifying values very different from the current behavioral value. Usually, these values are considered noise to be removed from a reconstruction algorithm. Using a coherency method like those presented here will allow an investigator to judge whether these aberrant reconstructions are truly valid representational events.

## Acknowledgements

This work was primarily supported by R01-MH06829. Additional support was provided by NSF-IGERT training grant #9870633 (JCJ, AJ), by the Center for Cognitive Science at the University of Minnesota (AJ, T32HD007151) and a 3M Graduate Fellowship (AJ). We thank Paul Schrater for helpful discussions, particularly with the development of the generative models and model selection methods.

## References

Amari SI (1977) Dynamics of pattern formation in lateral-inhibition type neural fields. Biological Cybernetics 27:77–87.

Arai K, Keller EL, Edelman JA (1994) Two-dimensional neural network model of the primate saccadic system. Neural Networks 7(6/7):1115–1135.

Averbeck BB (2001) Neural mechanisms of copying Geometrical Shapes. Ph.D. thesis, University of Minnesota.

Averbeck BB, Chafee MV, Crowe DA, Georgopoulos AP (2002) Parallel processing of serial movements in prefrontal cortex. Proceedings of the National Academy of Sciences, USA 99:13172–13177.

Averbeck BB, Chafee MV, Crowe DA, Georgopoulos AP (2003a) Neural activity in prefrontal cortex during copying geometrical shapes I. Single cells encode shape, sequence, and metric parameters. Experimental Brain Research 150:127–141.

Averbeck BB, Crowe DA, Chafee MV, Georgopoulos AP (2003b) Neural activity in prefrontal cortex during copying geometrical shapes II. Decoding shape segments from neural ensembles. Experimental Brain Research 150:142–153.

Averbeck BB, Latham PE, Pouget A (2006) Neural correlations, population coding and computation. Nature Reviews Neuroscience 7:358–366.

Batschelet E (1981) Circular statistics in biology. New York: Academic Press.

Bialek W, Rieke F, de Ruyter van Steveninck RR, Warland D (1991) Reading a neural code. Science 252(5014):1854–1857.

Brockwell AE, Rojas AL, Kass RE (2004) Recursive bayesian decoding of motor cortical signals by particle filtering. Journal of Neurophysiology 91:1899–1907.

Brown EN, Frank LM, Tang D, Quirk MC, Wilson MA (1998) A statistical paradigm for neural spike train decoding applied to position prediction from ensemble firing patterns of rat hippocampal place cells. Journal of Neuroscience 18(18):7411–7425.

Buzsáki G (1989) Two-stage model of memory trace formation: A role for "noisy" brain states. Neuroscience 31(3):551–570.

Buzsáki G (2006) Rhythms of the Brain. Oxford.

Chrobak JJ, Buzsáki G (1994) Selective activation of deep layer (V-VI) retrohippocampal neurons during hippocampal sharp waves in the behaving rat. Journal of Neuroscience 14(10):6160–6170.

Chrobak JJ, Buzsáki G (1996) High-frequency oscillations in the output networks of the hippocampal-entorhinal axis of the freely behaving rat. Journal of Neuroscience 16(9):3056–3066.

Chrobak JJ, Lörincz A, Buzsáki G (2000) Physiological patterns in the hippocampo-entorhinal cortex system. Hippocampus 10:457–465.

Csicsvari J, Hirase H, Czurkó A, Buzsàki G (1999) Fast network oscillations in the hippocampal CA1 region of the behaving rat. Journal of Neuroscience 19(RC20):1–4.

Dayan P, Abbott LF (2001) Theoretical Neuroscience. MIT Press.

Doboli S, Minai AA, Best PJ (2000) Latent attractors: a model for context-dependent place representations in the hippocampus. Neural Computation 12(5):1009–1043.

Droulez J, Berthoz A (1991) A neural network model of sensoritopic maps with predictive short-term memory properties. Proceedings of the National Academy of Sciences, USA 88:9653–9657.

Eliasmith C, Anderson CH (2003) Neural Engineering. Cambridge MA: MIT Press.

Ermentrout B, Cowan J (1979) A mathematical theory of visual hallucination patterns. Biological Cybernetics 34:137–150.

Ferbinteanu J, Kennedy PJ, Shapiro ML (2006) Episodic memory — from brain to mind. Hippocampus 16(9):704–715.

Foster DJ, Wilson MA (2006) Reverse replay of behavioural sequences in hippocampal place cells during the awake state. Nature 440(7084):680–683.

Frank LM, Eden UT, Solo V, Wilson MA, Brown EN (2002) Contrasting patterns of receptive field plasticity in the hippocampus and the entorhinal cortex: An adaptive filtering approach. Journal of Neuroscience 22(9):3817–3830.

Georgopoulos AP, Caminiti R, Kalaska JF, Massey JT (1983) Spatial coding of movement: A hypothesis concerning the coding of movement direction by motor cortical populations. Experimental Brain Research Suppl.(7):327–336.

Georgopoulos AP, Kettner RE, Schwartz AB (1988) Primate motor cortex and free arm movements to visual targets in three-dimensional space. II. Coding of the direction of movement by a neuronal population. Journal of Neuroscience 8(8):2928–2937.

Goodridge JP, Touretzky DS (2000) Modeling attractor deformation in the rodent head-direction system. Journal of Neurophysiology 83:3402–4310.

Gray J (2004) Consciousness: Creeping up on the Hard Problem. Oxford.

Harris-Warrick RM, Marder E (1991) Modulation of neural networks for behavior. Annual Review of Neuroscience 14(1):39–57.

Hasselmo ME, Bower JM (1993) Acetylcholine and memory. Trends in Neurosciences 16(6):218–222.

Hoffmann KL, McNaughton BL (2002) Coordinated Reactivation of Distributed Memory Traces in Primate Neocortex. Science 297(5589):2070–2073.

Jackson J (2006) Network Consistency and Hippocampal Dynamics: Using the properties of cell assemblies to probe the hippocampal representation of space. Ph.D. thesis, University of Minnesota.

Jackson JC, Johnson A, Redish AD (2006) Hippocampal sharp waves and reactivation during awake states depend on repeated sequential experience. Journal of Neuroscience 26:12415–12426.

Jackson JC, Redish AD (2003) Detecting dynamical changes within a simulated neural ensemble using a measure of representational quality. Network: Computation in Neural Systems 14:629–645.

Jaynes ET (2003) Probability Theory. Cambridge.

Jensen O, Lisman JE (2000) Position reconstruction from an ensemble of hippocampal place cells: contribution of theta phase encoding. Journal of Neurophysiology 83(5):2602–2609.

Johnson A, Redish AD (2005) Observation of transient neural dynamics in the rodent hippocampus during behavior of a sequential decision task using predictive filter methods. Acta Neurobiologiae Experimentalis 65(Suppl 2005):103. Poster 245.

Johnson A, Redish AD (2006) Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point: a possible mechanism for the consideration of alternatives. Society for Neuroscience Abstracts .

Johnson A, Seeland KD, Redish AD (2005) Reconstruction of the postsubiculum head direction signal from neural ensembles. Hippocampus 15:86–96.

Kepecs A, Lisman J (2003) Information encoding and computation with spikes and bursts. Network: Computation in Neural Systems 14(1):103–118.

Kohonen T (1977) Associative Memory: A System-Theoretical Approach. New York: Springer.

Kohonen T (1982) Self-organized formation of topologically correct feature maps. Biological Cybernetics 43:59–69.

Kohonen T (1984) Self-Organization and Associative Memory. New York: Springer-Verlag.

Kudrimoti HS, Barnes CA, McNaughton BL (1999) Reactivation of hippocampal cell assemblies: Effects of behavioral state, experience, and EEG dynamics. Journal of Neuroscience 19(10):4090–4101.

Laing CR, Chow CC (2001) Stationary bumps in networks of spiking neurons. Neural Computation 13(7):1473–1494.

Lee AK, Wilson MA (2002) Memory of sequential experience in the hippocampus during slow wave sleep. Neuron 36:1183–1194.

Lee I, Yoganarasimha D, Rao G, , Knierim JJ (2004) Comparison of population coherence of place cells in hippocampal subfields of CA1 and CA3. Nature 430(6998):456–459.

Lee TS, Mumford D (2003) Hierarchical bayesian inference in the visual cortex. Journal of Optical Society of America, A 20(7):1434–1448.

Leutgeb S, Leutgeb JK, Treves A, Moser MB, Moser EI (2004) Distinct Ensemble Codes in Hippocampal Areas CA3 and CA1. Science 305(5688):1295–1298.

Mardia KV (1972) Statistics of Directional Data. New York: Academic Press.

Munoz DP, Pélisson D, Guitton D (1991) Movement of neural activity on the superior colliculus motor map during gaze shifts. Science 251:1358–1360.

Nádasdy Z, Hirase H, Czurkó A, Csicsvari J, Buzsáki G (1999) Replay and time compression of recurring spike sequences in the hippocampus. Journal of Neuroscience 19(2):9497–9507.

Nirenberg S, Latham PE (2003) Decoding neuronal spike trains: How important are correlations? Proceedings of the National Academy of Sciences, USA 100(12):7348–7353.

O'Keefe J, Nadel L (1978) The Hippocampus as a Cognitive Map. Oxford: Clarendon Press.

O'Keefe J, Recce M (1993) Phase relationship between hippocampal place units and the EEG theta rhythm. Hippocampus 3:317–330.

O'Neill J, Senior T, Csicsvari J (2006) Place-selective firing of ca1 pyramidal cells during sharp wave/ripple network patterns in exploratory behavior. Neuron 49:143–155.

Rao RPN, Ballard DH (1999) Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience 2(1):79–87.

Redish AD (1997) Beyond the cognitive map: Contributions to a computational neuroscience theory of rodent navigation. Ph.D. thesis, Carnegie Mellon University.

Redish AD (1999) Beyond the Cognitive Map: From Place Cells to Episodic Memory. Cambridge MA: MIT Press.

Redish AD, Elga AN, Touretzky DS (1996) A coupled attractor model of the rodent head direction system. Network: Computation in Neural Systems 7(4):671–685.

Redish AD, Rosenzweig ES, Bohanick JD, McNaughton BL, Barnes CA (2000) Dynamics of hippocampal ensemble realignment: Time vs. space. Journal of Neuroscience 20(24):9289–9309.

Redish AD, Touretzky DS (1998) The role of the hippocampus in solving the Morris water maze. Neural Computation 10(1):73–111.

Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W (1997) Spikes. Cambridge MA: MIT Press.

Salinas E, Abbott L (1994) Vector reconstruction from firing rates. Journal of Computational Neuroscience 1(1-2):89–107.

Samsonovich AV, McNaughton BL (1997) Path integration and cognitive mapping in a continuous attractor neural network model. Journal of Neuroscience 17(15):5900–5920.

Schmitzer-Torbert NC, Redish AD (2002) Development of path stereotypy in a single day in rats on a multiple-T maze. Archives Italiennes de Biologie 140:295–301.

Schmitzer-Torbert NC, Redish AD (2004) Neuronal activity in the rodent dorsal striatum in sequential navigation: Separation of spatial and reward responses on the multiple-T task. Journal of Neurophysiology 91(5):2259–2272.

Schneiderman E, Bialek W, Berry MJ (2003) Synergy, redundancy, and independence in population codes. Journal of Neuroscience 23(37):11539–11553.

Serruya M, Hatsopoulos N, Fellows M, Paninski L, Donoghue J (2004) Robustness of neuroprosthetic decoding algorithms. Biological Cybernetics 88(3):219–228.

Skaggs WE, Knierim JJ, Kudrimoti HS, McNaughton BL (1995) A model of the neural basis of the rat's sense of direction. In: Advances in Neural Information Processing Systems 7 (Tesauro G, Touretzky DS, Leen TK, eds.), pp. 173–180. Cambridge MA: MIT Press.

Skaggs WE, McNaughton BL, Wilson MA, Barnes CA (1996) Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. Hippocampus 6(2):149–173.

Smyrnis M, Taira M, Ashe J, Georgopoulos AP (1992) Motor cortical activity in a memorized delay task. Experimental Brain Research 92:139–151.

Somogyi P, Klausberger T (2005) Defined types of cortical interneurone structure space and spike timing in the hippocampus. J Physiol (Lond) 562:9–26.

Suddendorf T, Busby J (2003) Mental time travel in animals? Trend in cognitive sciences 7(9):391–396.

Sutherland GR, McNaughton BL (2000) Memory trace reactivation in hippocampal and neocortical neuronal ensembles. Current Opinion in Neurobiology 10(2):180–6.

Tsodyks M (1999) Attractor network models of spatial maps in hippocampus. Hippocampus 9(4):481–489.

Tsodyks M (2005) Attractor neural networks and spatial maps in hippocampus. Neuron 48(2):168–169.

Tulving E (1983) Elements of Episodic Memory. New York: Oxford University Press.

Tulving E (2001) Episodic memory and common sense: how far apart? Philosophical Transactions of the Royal Society B: Biological Sciences 356(1413):1505–1515.

Tulving E (2002) Episodic memory: From mind to brain. Annual Review of Psychology 53(1):1–25.

Vanderwolf CH (1971) Limbic-diencephalic mechanisms of voluntary movement. Psychological Review 78(2):83–113.

Wills TJ, Lever C, Cacucci F, Burgess N, O'Keefe J (2005) Attractor Dynamics in the Hippocampal Representation of the Local Environment. Science 308(5723):873–876.

Wilson HR, Cowan JD (1973) A mathematical theory of the functional dynamics of cortical and thalamic tissue. Kybernetik 13:55–80.

Wilson MA, McNaughton BL (1993) Dynamics of the hippocampal ensemble code for space. Science 261:1055–1058.

Wilson MA, McNaughton BL (1994) Reactivation of hippocampal ensemble memories during sleep. Science 265:676–679.

Wu W, Gao Y, Bienenstock E, Donoghue JP, Black MJ (2006) Bayesian population decoding of motor cortical activity using a kalman filter. Neural Computation 18(1):80–118.

Ylinen A, Bragin A, Nadasdy Z, Jando G, Szabo I, Sik A, Buzsáki G (1995) Sharp wave-associated high-frequency oscillation (200 Hz) in the intact hippocampus: Network and intracellular mechanisms. Journal of Neuroscience 15(1):30–46.

Zemel RS, Dayan P, Pouget A (1998) Probabilistic interpretation of population codes. Neural Computation 10(2):403–430.

Zhang K (1996) Representation of spatial orientation by the intrinsic dynamics of the head-direction cell ensemble: A theory. Journal of Neuroscience 16(6):2112–2126.

Zhang K, Ginzburg I, McNaughton BL, Sejnowski TJ (1998) Interpreting neuronal population activity by reconstruction: Unified framework with application to hippocampal place cells. Journal of Neurophysiology 79:1017–1044.

## List of Figures

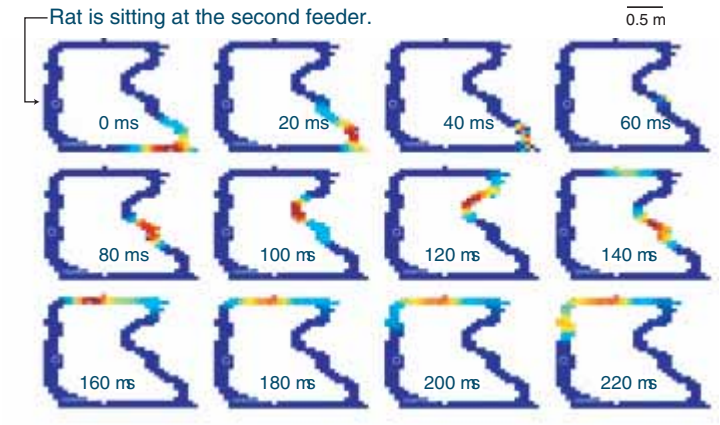


**FIGURE 1:** Replay of experience on the maze during an awake-sharp wave. The rat is sitting at the second feeder throughout the event. The distribution starts at the base of the first T and moves through the full maze in 220 msec (typical behavioral run times through this maze = 10–12 seconds). The reconstructed location is indicated by color (red high probability, blue low probability). Panels arranged from left to right, top to bottom in 20 msec intervals. Note the coherent, but non-local reconstruction of the representation during the sharp wave.

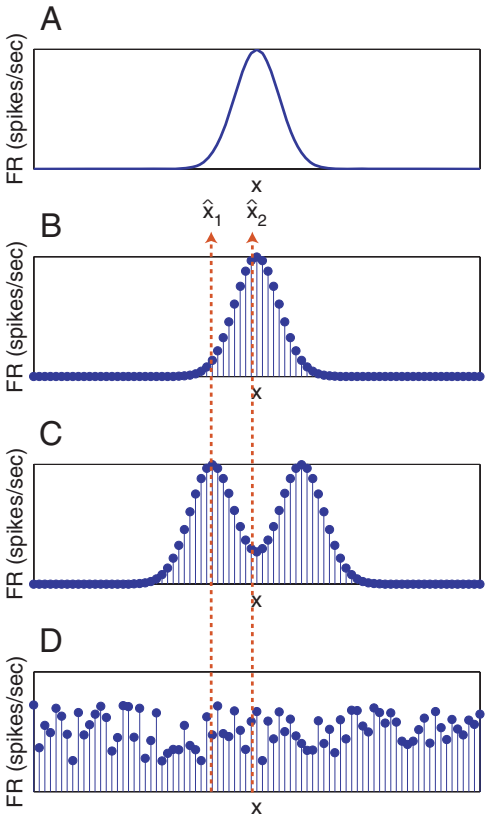


**FIGURE 2:** Self consistency. (A) An example unimodal tuning curve. The stimulus or behavioral variable is on the $x$-axis with firing rate represented along the $y$-axis. (B) A "coherent" network firing pattern. The stimulus or behavioral variable is on the $x$-axis with firing rate of each neuron represented along the $y$-axis. Each line represents the location of a neuron's preferred stimulus, with height equal to the neuron's firing rate. If each neuron in a network had unimodal tuning curves identical to the neuron represented in A but with the peak firing occurring at a different preferred stimulus $x$, then when the preferred stimulus of the neuron in A is presented, this is the expected network firing pattern. This pattern is consistent with behavioral variable $\hat{x}_2$ but not with $\hat{x}_1$. (C) A bimodal representation would represent an ambiguous or incoherent state of the network described in B, since the unimodal tuning curves would predict only one mode of activity should be possible of a single stimulus $x$ and that outside this mode neurons should be silent. One mode is consistent with behavioral variable $\hat{x}_1$ but neither mode is consistent with $\hat{x}_2$. (D) As in C this representation would represent a confused or incoherent state of the network described in B, since the unimodal tuning curves would predict a prominent mode of activity and that outside this mode neurons should be silent. This state is not consistent with either behavioral variable $\hat{x}_1$ or $\hat{x}_2$. From Jackson (2006).

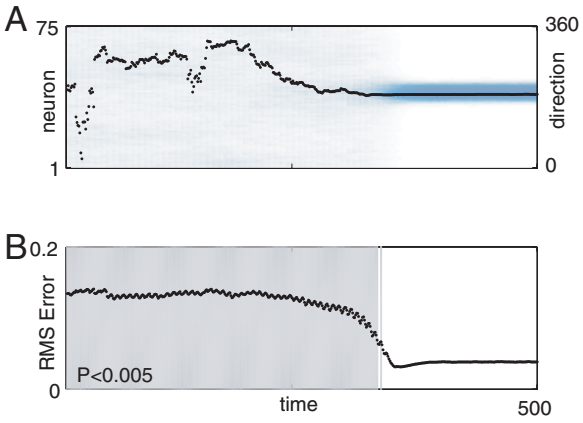


**FIGURE 3:** A simulation started with random input to the network settles to a stable state. (A) The neural activity. Time is shown in time-steps on the $x$-axis. Neurons ordered by their preferred direction ( $0^\circ$ – $360^\circ$) along the $y$-axis, shaded according to their firing rate. Black dots indicate the direction extracted from the population activity using population-vector reconstruction. Note, that the reconstruction algorithm yields a position whether or not there is an actual mode of activity present at that location. (B) The $I_{RMS}$ measure of inconsistency between actual and expected activity packets. During the random state, the discrepancy between the actual and expected activity packets is high ($p < 0.005$, gray zone). Upon reaching the stable state at time-step 342, the difference drops ($p > 0.005$, white zone). From Jackson (2006), see also Jackson and Redish (2003).

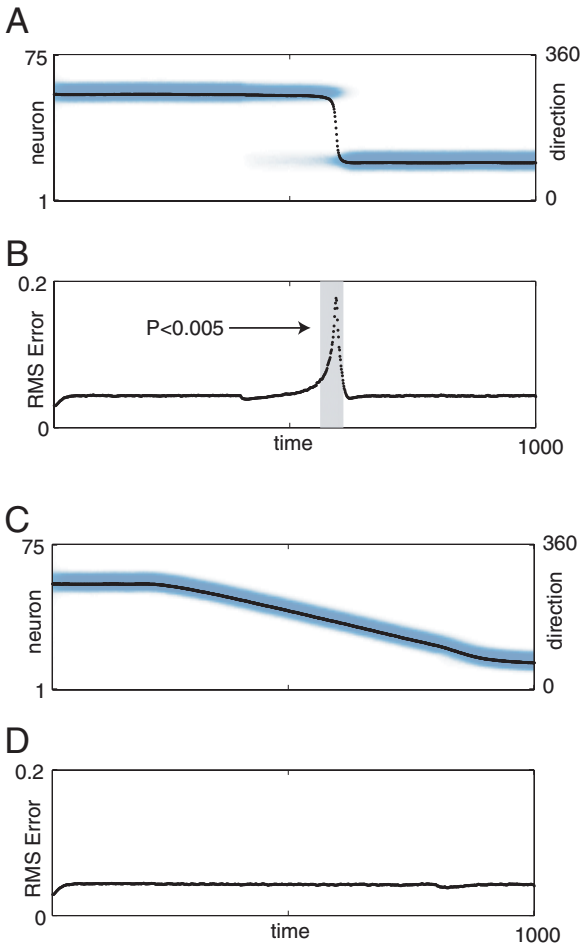


**FIGURE 4:** (A) Offset activity produces a jump in the representation. Layout as in Figure 3. Note that the reconstructed position shows a smooth rotation from the initial position of activity before the jump, through positions where there is no network activity, to the final location of activity after the jump. (B) The $I_{RMS}$ measure of inconsistency between actual and expected activity packets. The discrepancy between the actual and expected activity packets is low during the stable state, before and after the jump ($p > 0.005$, white zone), but high during the transient bimodal activity state at the moment of the jump from time-steps 562–609 ($p < 0.005$, gray zone). C) A smooth rotation induced in the network yields stable results. Layout as above. The reconstructed position follows the activity of the network faithfully. (D) The $I_{RMS}$ measure of inconsistency between actual and expected activity packets. Throughout the rotation, the network maintains a stable state with a small difference between the actual and expected activity packets ($p > 0.005$). From Jackson (2006), see also Jackson and Redish (2003).

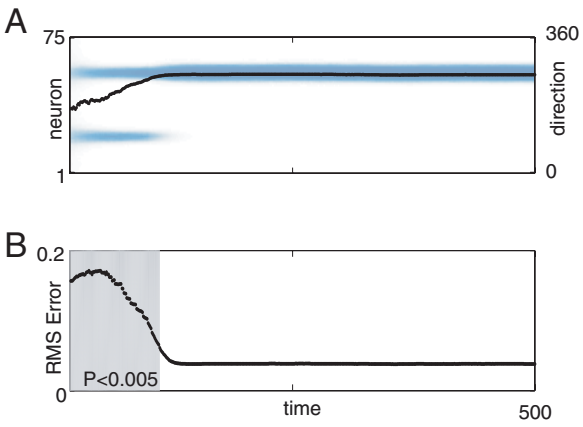


**FIGURE 5:** (A) A simulation started with competing inputs settles to a single mode of activity. Layout as in Figure 3. Note that the reconstructed position shows a smooth rotation from the mean position, where there is no network activity to the winning location. (B) The $I_{RMS}$ measure of inconsistency between actual and expected activity packets. The discrepancy between the actual and expected activity packets is high during the initial bimodal state before the competition is resolved ($p < 0.005$, gray zone), but low afterwards ($p > 0.005$, white zone). From Jackson (2006), see also Jackson and Redish (2003).

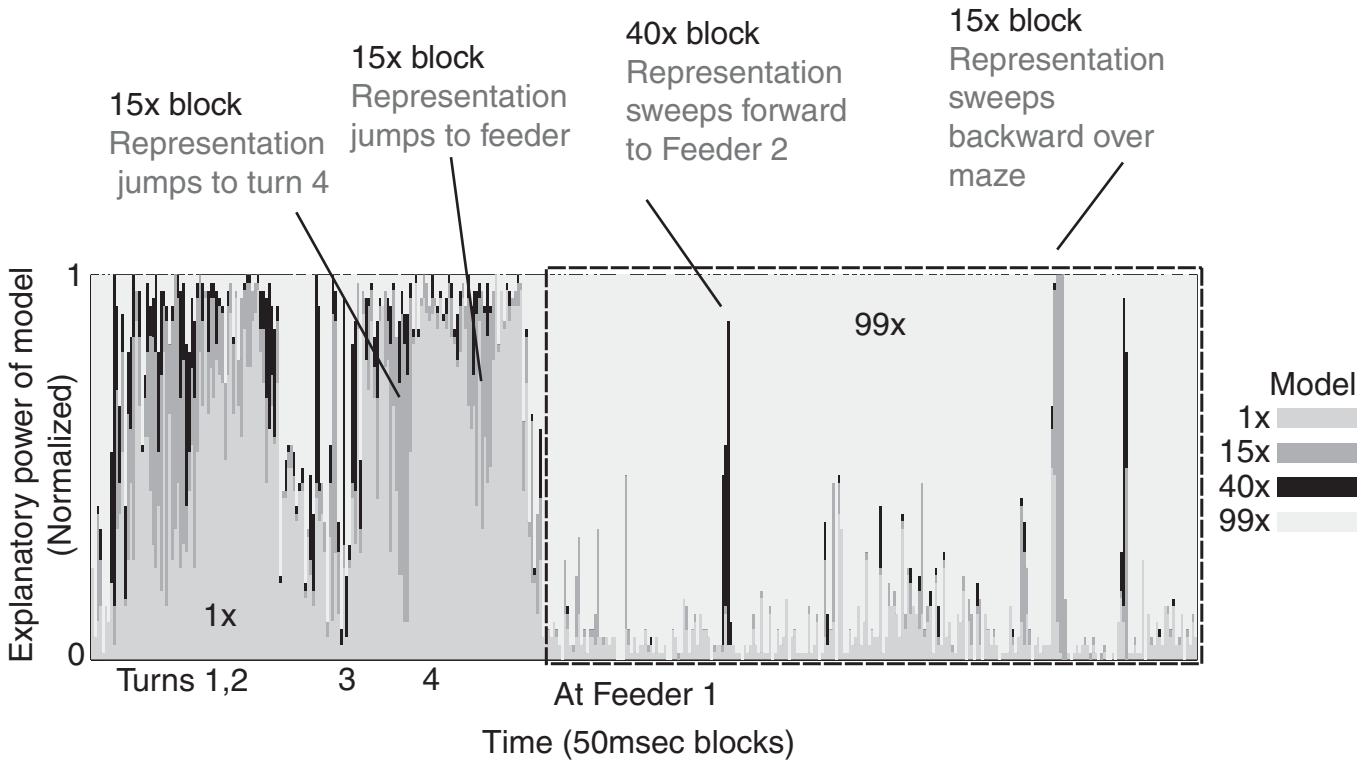


**FIGURE 6:** Multiple generative models in the hippocampus. Four generative models were examined 1×, 15×, 40×, and 99×. During the first portion (Turns 1–4), the animal was running through the maze. During the second portion, the animal paused at the first feeder to rest, groom, and eat.

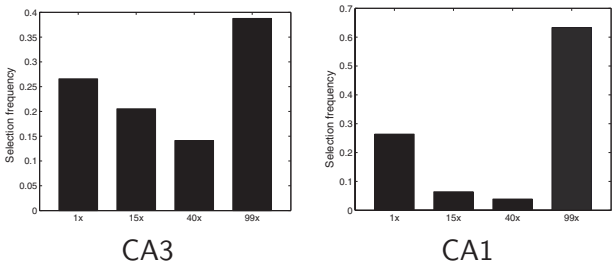


**FIGURE 7:** Percentage of samples in which each model was found to be the most consistent (Eq. (15)). The $99\times$ filter was often selected during jumps or intervals in which few spikes are fired.