Information coding in higher sensory and memory areas

Alessandro Treves
SISSA, Cognitive Neuroscience, Trieste, Italy
`http://www.sissa.it/cns/`

January 21, 2000

Running title: Information coding in the CNS

Keywords: information theory, regularization, limited sampling,
neural coding, distributed representations, correlations

1 Understanding neural codes requires information measures

How do you communicate? A moderately bright extra-terrestrial, who were to investigate the codes you use, would reasonably conclude that you mainly communicate verbally. Our E.T. might further describe verbal codes as strings of chunks of variable length, that you appear to call *words*, which can be uttered as collections of phonemes or else written as nearly isomorph collections of letters, and so on with further details. If however E.T. were exceptionally bright, or needed to write a grant application on this investigation, it would probably "discover" that in some situations you also communicate a lot just with your grimace, or with the clothes you have chosen to wear, or in a thousand other ways.

Neurons are vastly simpler than human beings, but the metaphor is not completely silly, because it illustrates the volatility of the notion of neural codes. Nobody in the right of his or her mind would think that nature has designed a unique way for neurons to communicate, and in fact they interact, or affect each other, in a thousand different ways. In certain specific situations neurons may tell each other a lot with the way they compete for peptides, for example, or with the way they couple in ephaptic interactions. Yet a first understanding of the operation of neural networks in the brain requires that we try to describe the main, usual form (or forms) of communication. We should take the approach of the moderately bright investigator, and leave the discovery of exceptional facts for later on. Further, we should try to quantify *how much* is communicated in each situation, because only a quantitative comparison allows to assess different codes, especially if they share part of the content of what is being communicated. Information theory [1] has been developed precisely to quantify communication, and is therefore quintessential to an appraisal of neural codes. Applying information theory to neural activity (rather than to the synthetic communication systems for which it was developed) is however riddled with practical problems and subtleties, which must be clarified before reporting experimental results.

In this chapter, we do not consider other means of neuronal communication than the emission of action potentials, or spikes, and regard them as selfsimilar all-or-none events whose only distinctive features are the time of emission and the identity of the emitting neuron. Thus we restrict ourselves to information being represented, across a given population of neurons, by strings $\{t_{i,k}\}$, where $i = 1, \ldots, C$ labels the emitting neuron and k indexes successive spikes in a prescribed time window. This is still a rather general and potentially very rich language, which in many situations can be reduced to considerably simpler forms. For example, most of the information might be carried simply by the total number of spikes, n_i, emitted by each neuron in the window, irrespective of their timing within the window. Although the spike count is an integer, it becomes a (positive) real number when averaged over several repetitions or even when calculated, in more general terms, by convolving the spike train with a given time kernel. Therefore, it is more convenient to consider, instead, the firing rate, r_i, which, being divided by the time window, or the integral of the kernel, is also relatively invariant across window lengths for quasi stationary processes. The extent to which the firing rates of a population of neurons may or may not carry most of the information represented in the complete list of spike emission times is, of course, a question to be addressed experimentally, in any given situation. This has been done, with some success, mostly at the level of single neurons, as will be discussed later. First, we must consider *what* information we can set out to measure, given the difficulties and subtleties of measuring it in neuronal activity. For simplicity of notation, we shall think of such information as being represented in the rates, although the arguments of the next section apply equally to information represented in the complete list of spike emission times.

2 Sampling with limited populations, correlates and repetitions

The activity of a population A of neurons represents both information and noise. The part that can be considered, in general, to contain information, is the part that varies together with something else, such as the activity of another population B, or some external correlate. It is measured by what is usually called mutual information

$$I(\{r_{i\in A}\};\{r_{j\in B}\}) = \int \prod_{i\in A} \mathrm{d}r_i \int \prod_{j\in B} \mathrm{d}r_j p(\{r_i\},\{r_j\}) \log_2 \left[\frac{p(\{r_i\},\{r_j\})}{p(\{r_i\})p(\{r_j\})}\right]. \tag{1}$$

Note that $\{r_{j\in B}\}$ could stand for the activity of the same population A, but at a different time. Or, it could stand for the parameters of some external correlates. Whatever the case, for the present discussion it is useful to consider that neither set of variables may be easily manipulated by the experimenter. Then, to evaluate the transmission of information from B to A (or viceversa: mutual information is symmetric and does not reflect causality) one should let the coupled system do what it normally does, in "ecological conditions", for the very long time needed to sample accurately the joint probability distribution $p(\{r_{i\in A}\},\{r_{j\in B}\})$. Since we are usually interested in the way the neural *system* operates, and not just in individual neurons, ideally we would need to record the activity of *all* cells in A and B, C_A and C_B. The time required would be exponential in $C_A + C_B$ (times, effectively, the logarithm of the number of discriminable firing levels of each cell), i.e. much longer than the age of this and all preceding universes. Practical limitations on the number of cells that can be recorded simultaneously (a few hundreds, now) make the time required for a single measure less apocaliptic, but still way from affordable. In practice, *a direct measure of mutual information has to be* based on the recording *of the activity of only a handful of cells.* This implies that we are forced to hope that those cells are in some sense "typical", but it also forces us to overlook potentially important coding schemes that could only be revealed, quantitatively, by taking into account the simultaneous activity of many neurons. Substituting computer simulations for real recording experiments only alleviates the constraint by a tiny bit, while an analytical evaluation of mutual information is sometimes possible with formal models of large populations, in which however the result is in-built in the structure of the models. Thus, a first restriction on the applicability of information measures to neural activity is in the size of the population that can be sampled, and effectively in the dimensionality of the codes that can be investigated.

A second restriction emerges when considering the *content* of the information being represented. The content is determined by the set of external (or internal) correlates of the activity in populations A and B. Ideally, they should reproduce the ecological working condition of the neural system being studied. In practice this is hard to do, in the lab, in a reasonable time, also because what this "ecological condition" is may be unknown or ill-determined (except perhaps for peripheral neural systems tightly coupled to specific dimensions of the sensory environment [2]). A common strategy, when studying the CNS, is instead to select a discrete set of elements, $\mathcal{S}$, representative of interesting correlates, and to quantify the mutual information not between A and B, but between A and $\mathcal{S}$

$$I(\{r_{i\in A}\};\{s\in\mathcal{S}\}) = \int \prod_{i\in A} \mathrm{d}r_i \sum_{s\in\mathcal{S}} p(\{r_i\},s)\log_2\left[\frac{p(\{r_i\},s)}{p(\{r_i\})P(s)}\right]. \tag{2}$$

where the capital P denotes a real probability, rather than a probability density. This is conceptually a different quantity, which, since $\mathcal{S}$ is an object of much reduced complexity than the activity in B, is much easier to measure. In particular, if $\mathcal{S}$ includes S equiprobable elements, its entropy $\log_2 S$ will be an upper limit on the mutual information, no matter how large C_A, the population of neurons encoding the set of correlates. Eq. 2 quantifies how much the activity of population A allows us to discriminate between elements in $\mathcal{S}$. Eq. 1 quantifies how much it tells us about the activity of another population B. One may wonder then, what the measurable quantity in Eq. 2 tells us about the impractical-to-measure quantity of Eq. 1. Curiously, this question seems to have been disregarded in the literature, with the exception of Frolov and Murav'ev [3], who consider two quantities analogous to 1 and 2, which they denote as I_2 and I_1, and conclude that the total information that can be extracted from neural activity is the *sum* $I_1 + I_2$. Our recent analyses (Samengo and Treves, to be published) lead to a rather different conclusion. $I(\{r_{i\in A}\};\{s\in\mathcal{S}\})$, far from being a term to be *added* to $I(\{r_{i\in A}\};\{r_{j\in B}\})$, provides a good estimate of it, at least when many different correlates are used, and few enough cells are sampled that one is far from the regime approaching the saturation value at $\log_2 S$. While the numbers, and the quality of the estimate, depend on the exact details of the network to be analysed, our result justifies *a posteriori* the common practice of extracting measures of $I(\{r_{i\in A}\};\{s\in\mathcal{S}\})$. We shall see later on how, when sampling more than a handful of cell, the common practice is to adopt a further simplification along this path, and to extract yet another, distinct information measure, the information about $\mathcal{S}$ recovered by a decoding procedure.

The third major sampling limitation with information measures is intimately (but inversely) related to the other two, and touches directly on a core concern of any scientific measure, that of reproducibility. It is the limitation arising from the limited availability of repetitions of the same observation. Mutual information measures, as they depend on the joint probability of two variables, always require many repetitions. To sample adequately a set of S elements and a firing rate vector which can take of the order of $R =$(max spikes per cell)^C values, one needs of the order of $S \times R$ repetitions [4]. In recording experiments, especially in the CNS of mammals, this requirement is difficult to meet. Since mutual information depends non linearly on joint probabilities, a measure based on insufficient repetitions is not only imprecise, but also, in principle, biased, that is affected by systematic errors. Usually the procedure is to simply substitute observed frequencies for the underlying probabilities (a so called "frequentist" approach), and usually the effect of undersampling the joint probability (upstairs in the logarithm of Eq.2) is much more serious than that of undersampling its marginals (downstairs in the log). Since mutual information is supralinear in the joint probability, its undersampling typically leads to an upward bias, or mean (systematic) error in the measure. Various techniques [5, 6] have been developed to estimate and subtract, or otherwise neutralise, this bias, but their limited efficacy makes limited repetition sampling the most stringent constraint, in practice, on measuring information carried by neural activity in the CNS of mammals. Since the problem is exacerbated when many cells and large sets of correlates are considered, the most reliable measures so far have been obtained with very limited sets of correlates and at the single cell level, and it is to these that we turn next.

3 What code is used by single cells?

Animals interact, via their sensory and motor nervous systems, with a continuously changing world, and it is obvious that also activity in their central nervous system should reflect, in the time dimension, this continuous change. Some investigators have been curiously excited by finding evidence of such strict coupling between the CNS and the outside world. Richmond, Optican and colleagues, instead, have addressed the question of *temporal coding*, at the single cell level, in the correct conceptual framework. They have asked whether individual neurons in the CNS make non-trivial use of the time dimension, by recording the responses of cortical visual neurons to static visual stimuli. A stimulus that is, after its sudden onset, constant in time, may elicit in a given neuron activity that varies in time only in a generic fashion [7], or that varies in time in a way specific to the stimulus itself. In the latter case, the neuron has used time to code something about the static stimulus, something which was not its time dependence. Quantitatively, this would appear as mutual information, between the stimulus used and a descriptor of the response that is sensitive to the timing of spikes, much higher than the information present in a descriptor insensitive to spike timing, like the firing rate or spike count. Note that the quantitative difference would have to be substantial, because a higher-dimensional descriptor will always be able to convey something that any single, prescribed low dimensional descriptor misses out.

The approach taken by Optican and Richmond [8] quickly gained acceptance as a sound basis for revealing temporal codes, and their claim that the time course of single neuron activity carries between 2 and 3 times more information than the spike count had a considerable impact. It was unfortunate for them to discover, in the following years, that this early result was entirely an artifact of the limited number of trials per stimulus they had used. Limited sampling affects differentially the information extracted from descriptors of different dimensionality, and with the time course descriptor it resulted in an upward bias much larger than with the spike count. Having introduced some form of correction for limited sampling [9], the evidence for temporal coding weakened and eventually all but evaporated [10, 11, 12]. A replication with a similar experiment in Edmund Rolls' lab [13] has further suggested that part of the residual difference in mutual information could be due to differential onset latency, which could still be called temporal coding, but of a less interesting nature.

To date, no report has appeared that demonstrates substantial non-trivial usage of time by single cortical neurons [14, 15]. The one apparent exception is the so-called phase precession in rat hippocampal place cells [16]. The firing of these (principal) cells is modulated by the Theta rhythm, which is expressed mainly in the firing of local interneurons. When a rat runs through the place field of a given cell, this cell tends to fire towards the end of a Theta period as it enters its field, and progressively earlier in phase as it goes through it. The effect can be understood as a simple emergent property, whereby a cell that needs more recurrent activation to complement a weakish afferent input, tends to fire later than cells with a stronger extrinsic drive [17]. On a linear track, place fields tend to be directional, that is to be associated with only one of the two directions in which the field can be traversed. Therefore, the phase of firing can be used to extract some additional information on the exact location of the rat, on top of what is available from, say, the number of spikes emitted over a Theta period. In an open field, however, in which the rat can traverse the same place field along an arbitrary trajectory, and elicit firing in the same cell, the phase information cannot be used for absolute localization, independent of the trajectory, and the postulated temporal coding through phase precession reveals itself as a mere epiphenomenon [18].

Nevertheless, the body of experiments addressing temporal coding at the single cell level has stimulated the development of information extraction procedures, among them those addressing limited sampling [9, 19, 11, 5, 6], that turn out to be crucial also in measuring the information conveyed by populations of cells.

4 Is a neuron conveying information only when it fires?

The intuition of many neurophysiologists is that central neurons transmit information simply when they fire. In the extreme, a *spike* is regarded as a *quantum* of information, and even confused with a *bit* (which in fact is just a unit, and implies no quantization at all). No matter how crude, this intuition is reinforced by the lack of evidence for sophisticated coding schemes: cortical neurons appear uninterested in the game of transcribing stationary signals into fancy temporal waveforms. Yet, one could think of other non-trivial coding schemes, which do not involve the time dimension, but just clever manipulation, by the neuron, of its conditional firing probability. For example, certain connectionist models assume that a unit active at its maximum level reports the presence of its own preferred correlate (e.g., the sight of one's grandmother), while any intermediate level of activation would be elicited by other correlates, with partially shared attributes [20]. A neuron behaving according to such model might be expected to reliably fire at top rate, say 10 spikes in 100 ms, when detecting the grandmother, and to fire between 0 and 9 spikes when detecting other senior ladies. Each of them in some of the repetitions of the experiment may resemble more the true grandmother, and thus evoke more spikes, than in other repetitions. Then $P(r|\text{granma})$ would be strongly peaked at 100 Hz, while $P(r|\text{ladyX})$ would be more broadly distributed between 0 and 90 Hz. The *meaning* of, say, 3 spikes in close succession would be rather different depending on whether there are 7 more close by, or just 6.

Nothing of this sort has ever been observed with neurons. Neurons appear to use spikes in a simple-minded fashion. Moreover, neurons can be as informative when they fail to fire as they are when they do fire. Roughly speaking, the only information they provide is in the extent to which their current firing level is above or below their mean firing level. One way to confirm this is to compute the quantity

$$I_1(s) = \int \mathrm{d}r p(r|s) \log_2 \frac{p(r|s)}{p(r)}, \tag{3}$$

which depends on the probability of each firing rate conditional on the correlate s and, when averaged over correlates, yields the mutual information. We have mistakenly called this quantity 'information per stimulus', or 'stimulus-specific information' [21], along with others. Recently DeWeese and Meister [22] have correctly pointed out that $I_1(s)$ is not additive, as any information quantity should be, while the similar quantity

$$I_2(s) = \int \mathrm{d}r p(r|s) \log_2 p(r|s) - \int \mathrm{d}r p(r) \log_2 p(r) \quad (4)$$

in fact is additive. $I_2(s)$ also averages to the mutual information, which is positive definite, but as a function of s $I_2(s)$ takes also negative values. $I_1(s)$ is not additive but positive definite, and should be called the 'stimulus-specific surprise', as proposed by DeWeese and Meister [22]. In any case, the interest in $I_1(s)$ is not so much in quantifying information, but rather in illustrating the simplicity of the firing rate code. This can be appreciated by first taking the limit of a very brief time window Δt [23]. In such a window the cell may emit at most a single spike, with probability $r_s \Delta t = \Delta t \int \mathrm{d}r p(r|s) r$; and $I_1(s)$ reduces to its limit, the 'stimulus-specific surprise per spike' $\chi(s) = (1/\bar{r}) \mathrm{d}I_1(s)/\mathrm{d}t$. Relative to the overall mean rate $\bar{r} = \sum_s P(s) r_s$, $\chi(s)$ is a universal curve,

$$\chi(x) = x \log_2 x + \frac{1}{\log 2}(1-x), \quad (5)$$

where $x = r_s/\bar{r}$ (see Fig. 1 and Ref. [24]). This universality is intimately related to the availability of a single 'symbol', the spike, in the neural alphabet, at least in the limit of short times, when the emission of more spikes has negligible probability. Over a longer time window, instead, $I_1(s)$ is not constrained to follow the universal $\chi(s)$ curve, and a departure from it could reveal a more sophisticated code. For example, DeWeese and Meister [22] remind us that for an optimal code that saturates the channel capacity, the specific surprise should be *constant* across different correlates. This is far from what has been observed in the very few cases when this issue has been probed. The specific surprise appears to follow the universal curve [25, 21, 26], indicating that the firing rate code is likely to remain as simple as it is forced to be for short times. In agreement with this, typically firing rates elicited by repetitions of the same stimulus, or correlate, have a variance monotonic in the mean rate r_s, and a simple distribution around the mean, between Poisson and normal, again not hinting at any clever manipulation of the conditional probabilities $P(r|s)$ [27].

Related evidence, though not in terms of conditional probabilities, comes from the observation of spike count distributions produced by cortical neurons in their normal operating regime. It has been suggested by Levy and Baxter [28] that an *exponential* spike count distribution would reveal optimal coding, subject to a metabolic constraint on the energy consumption associated with each spike. This would be an example of a clever design principle implemented in the brain. An attempt to search for such exponential distributions by subjecting visual neurons to more or less ecological stimulation has only shown exponential *tails* [29], not fully exponential distributions, while it has been shown that the observed distributions can be explained as the result of an elementary random process [30], which has nothing to do with optimising the neural code.

Currently available evidence on single neurons thus indicates that the simple neurophysiologists' intuition is, essentially, accurate. Cortical neurons appear not only unable to make creative use of time, but also unable to alter the mapping between the input they receive and the spikes they produce, on the basis of any coding optimization principle. If this is correct, the equivalent of the old 'tuning curve', that is the distribution of mean firing rates to each correlate, is all that is necessary to characterize adequately the activity of a single cortical neuron. If the relevant correlates are simple one-dimensional parameters, such as orientation in V1, then the tuning curve is simply described by giving e.g. preferred orientation, width, baseline and peak value (and the informational aspects are usually quantified by just the Fisher information, whose relation to mutual information is discussed by Brunel and Nadal [31], see also the chapter by

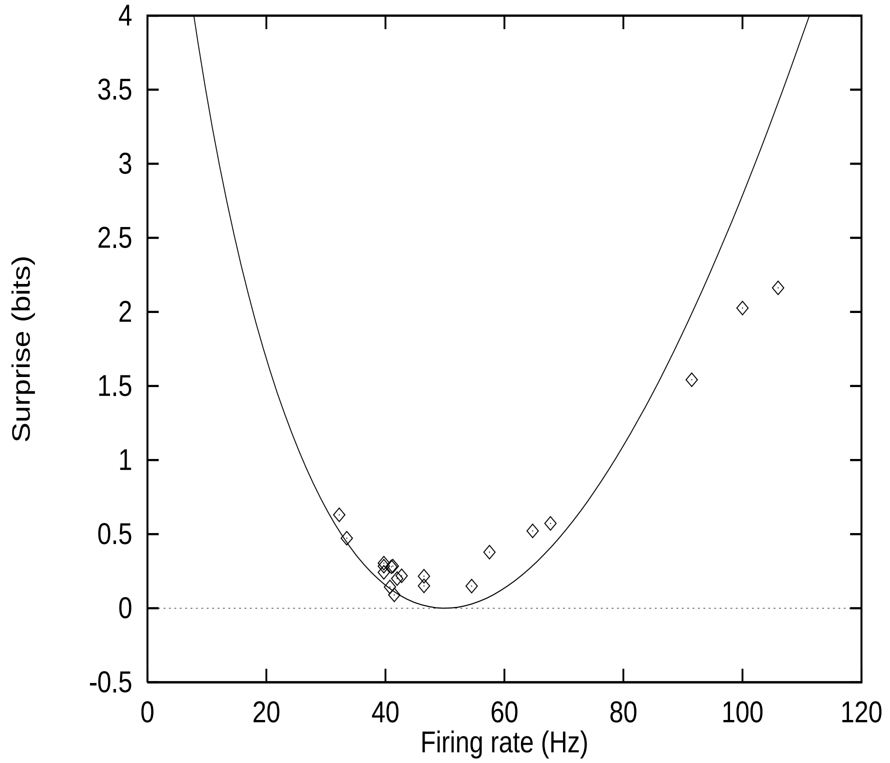

Figure 1: The stimulus-specific surprise from real data follows the universal curve valid in the $t \to 0$ limit. Real data from an inferior temporal cortex cell responding to 20 face stimuli over 500 ms [21]. The curve is the surprise rate, expressed as bits per 100 $msec$, for a mean firing rate of 50 Hz. The main difference between limit curve and real data is just a rescaling, roughly by the factor 5 implicit in the graph.

Fukumizu in this book). If the relevant correlates are embedded in a less transparent domain set, such as faces or fractals [32], then in principle the mean rate r_s to each correlate should be given (the variance being largely determined by the mean [27]). However the gross feature of the distribution of rates can be still conveniently described with fewer parameters, such as overall mean rate, spontaneous rate and sparsity [33] of the distribution. To such parameters we turn at the end of this chapter; before, we should consider the possibilities offered by population coding.

5 Are neighbouring neurons telling the same story?

The studies cited above show that, in roughly ecological conditions, single cortical neurons typically can transmit up to a fraction of a bit, about stationary correlates, over a few hundreds of ms (with instantaneous information rates occasionally a bit higher). This is clearly way below the behavioural discrimination capability of the animal. Therefore, we are brought to consider the transmission of information by populations of neurons. One crucial question is the extent to which the information provided by different neurons is the same, that is, redundant.

This issue has been addressed, perhaps for the first time at a quantitative level, by Gawne and Richmond [34]. Recording from pairs of inferior temporal cortex neurons in the monkey, responding to a set of 32 simple visual stimuli (Walsh patterns), they have compared the information obtained by considering both responses to the sum of that obtained for each response alone. On average across several pairs, they have found an information 'overlap' $y = 0.2$ shared by a pair, e.g. a single cell information $I(1) \simeq 0.23$ bits and for the pair $I(2) \simeq 0.41$ bits $\simeq I(1) + I(1)(1-y)$. This seems to imply that as much as 80% of what the second cell has to say is fresh information, not yet reported by the first cell - not much redundancy. Gawne and Richmond have, however, noted that even such limited redundancy would have drastic effects if it held among arbitrary pairs of cells in a local population. They have considered a simple model, which assumes that if a fraction $1-y$ of the information conveyed by the second cell is novel, then a third cell would on average convey a fraction $(1-y)^2$ of novel information (and a fraction $y(1-y)$ shared with each preceding cell, and y^2 with both); the i^{th} cell recorded would contribute a fraction $(1-y)^{i-1}$ novel information, and adding up all contributions one ends up with $I(\infty) = I(1)/y$, or just 1.15 bits in their experiment. Since 5 bits are necessary to discriminate 32 stimuli, they have concluded that even an infinitely large population of cells with that apparently limited level of redundancy would not be able to code for their small stimulus set; and therefore that the mean redundancy among neurons farther away, than those they recorded from, should decrease considerably, towards zero, to account for the fact that behaviourally the animal is obviously able to discriminate.

A similar warning, that even small redundancies can have drastic effects on the representational capacity of a population, was put forward by Zohary, Shadlen and Newsome [35]. They looked at the correlated discharge of MT neurons to randomly moving dots, where a single (unidimensional) parameter was used as correlate, the average motion of the dots. Their perspective was different from that of Gawne and Richmond, but the result seemed to imply, again, that adding more and more cells adds little to the accuracy of neural codes.

What appeared important, then, was to go beyond what could be extrapolated from the shared information between pairs of cells, and measure directly the information that could be extracted from large populations. Going for large populations requires two changes in the approach. The first, experimental, is that many cells have to be recorded simultaneously, and thus with multiple electrodes. Alternatively, separately recorded cells can be considered, but the results should be later checked against those obtained with simultaneous recordings, because these are needed to record trial-to-trial correlations, and their possible effects on information. The second change, in the analysis, is brought about by the exponential explosion of the 'response space' spanned by $\{r_{i \in A}\}$, when $i = 1, \ldots, C$ and C becomes large. The explosion makes it impossible to sample adequately the space, and thus to measure directly the mutual information in Eq.2. A standard procedure is to use a *decoding* algorithm, that converts the vector $r_{i \in A}$ into a prediction of which correlate s' elicited it, or else assigns probabilities $P(s'|r_{i \in A})$ to each possible correlate. The result is that one measures the decoded information

$$I(\{s \in \mathcal{S}\}; \{s' \in \mathcal{S}\}) = \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{S}} P(s, s') \log_2 \left[\frac{P(s, s')}{P(s)P(s')} \right]. \tag{6}$$

Inasmuch as decoding is done correctly, that is the functions $s'(r_{i \in A})$ or $P(s'|r_{i \in A})$ do not contain any *a priori* knowledge on the actual correlate s, the decoded information is less or equal to the original mutual information; just in the same way that any mapping from a variable (here $r_{i \in A}$) to another, e.g. to a regularized variable, can only degrade, or at most preserve, but not improve, the correlation between the original variable and a third one. Decoding can be done in a variety of ways, and it is not possible to quantify the information loss resulting from each particular algorithm. Still, experience with most commonly used algorithms, and comparisons, when possible, with direct measures, suggest that in many cases the information loss is minor. In particular, very simple decoding algorithms, like those that may conceivably be implemented in the brain, appear often to lose only slighly more information than sophisticated algorithms based on Bayesian models [36].

Using non simultaneous recordings from up to 58 cells in the monkey temporal cortex, $S = 5$ stimuli, and a simple decoding algorithm, Gochin *et al.* [37] proposed to investigate the scaling of $I(\{s \in \mathcal{S}\}; \{s' \in \mathcal{S}\})$ with the number of cells used for decoding. They expressed their result in terms of the *novelty* in the information conveyed by C cells, defined as the ratio of such information with the sum of that provided by each cell alone. They found that the novelty scaled as $1/\sqrt{C}$, intermediate between the $1/C$ behaviour corresponding to no new information being provided by additional cells, and the trend to a constant, if at least a finite part of what each cell contributed, were novel. The $1/\sqrt{C}$ behaviour seemed appealing in that it vaguely matched noise suppression by C independent processes carrying the same signal; unfortunately, it was likely an artefact, generated by determining a curve on the basis of just 3 points, by failing to correct for limited sampling, and most importantly by neglecting to consider the ceiling effect, at $I = \log_2 S = 2.32$ bits, just 9 times above the average single cell information, $I = 0.26$ bits.

Our replications of this type of measure, with decoding algorithms based simply on the firing rates of simultaneously or non simultaneously recorded cells, have exposed a different scaling behaviour, in all cases investigated [38, 36, 26] (cf. Fig. 2). This is a linear increase $I(C) \propto C$, eventually saturating towards the ceiling at $I_{max} = \log_2 S$. The crucial point is that the saturation level depends on the set of correlates used, and mainly on their number, and it has nothing to do, in principle, with the coding capacity of the population of cells. A simple empirical model describing rather well the whole scaling from linear to saturating is an extension of the Gawne and Richmond model, which in addition assumes that their 'overlap' y also represents the average fraction of I_{max} conveyed by single cells (as an overlap it would presumably be lower, when measured across distant pairs, than when measured, as in their experiment, only across pairs of nearby cells). Two cells then overlap over a fraction y of the fraction yI_{max} each conveys, that is the essential assumption of the model is that overlapping areas are randomly distributed across 'information space'. The information carried by C cells

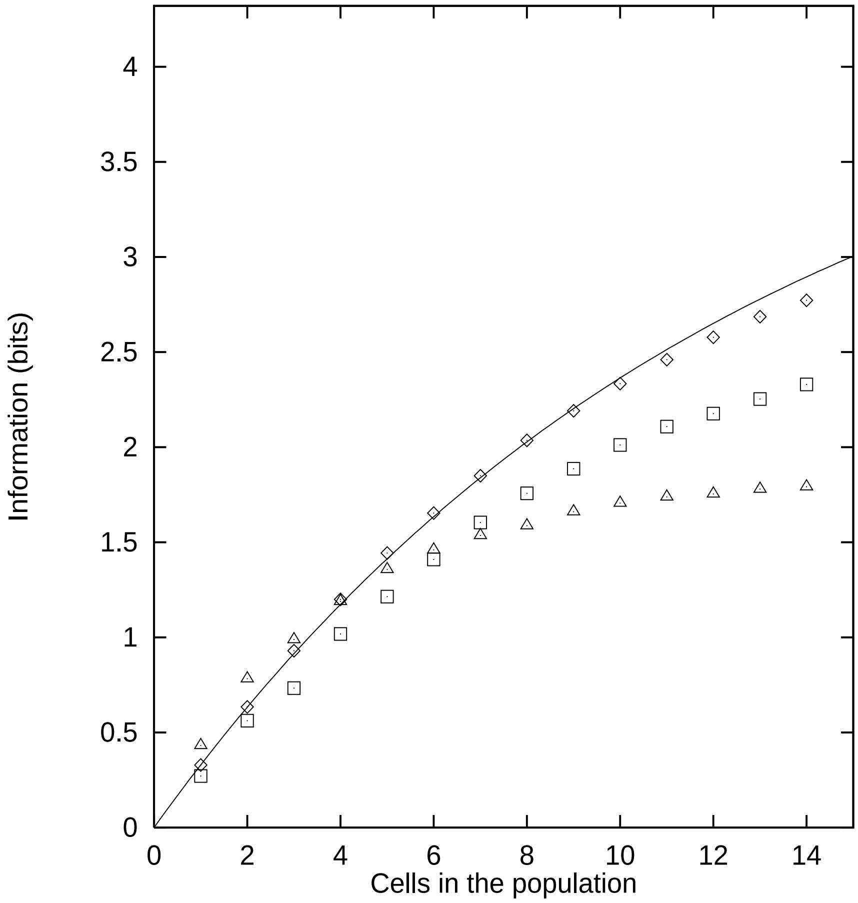

Figure 2: The information extracted from a population of cells saturates at the entropy of the stimulus set. Real data from up to 14 non-simultaneously recorded inferior temporal cortex cells responding to 20 face stimuli over 500 ms, adapted from Ref. [36]. The intermediate and lower data points correspond to reduced sets of 9 and 4 stimuli, respectively. The curve is the simple exponential saturation model of Eq. 7.

is found, with an easy derivation, to be, according to this random model

$$I(C) = I_{max}\left[1-(1-y)^C\right], \tag{7}$$

or, in words, a simple exponential saturation to the ceiling. This simple scaling has now been derived analytically, and found to apply exactly in quite general cases (Samengo and Treves, to be published). The important element, assumed true in the analytical derivation and apparently approximately holding also in the experimental recordings, is the lack of *correlation* in the activity of different cells. Correlations can be of two main types, sometimes referred to as 'signal' and 'noise' correlations: those appearing across repeated trials with the same correlate (denoted $\gamma_{ij}(s)$ in the following section); and those between the average 'tuning curves' of several cells, that is between their activity distribution across correlates, once averaged over many repetitions of each (denoted ν_{ij}). Neither type of correlation is considered in the analytical derivation. In the experiments, signal correlations would indeed have an effect, if substantial, while noise correlations would be unlikely to be ever able to effect a departure from the behaviour described by Eq.7: even with simultaneous recordings, decoding algorithms based just on firing rates would likely miss out any influence of such correlations. What is needed then, in order to go beyond Eq.7 and address the potential role of correlations, is an alternative approach that does not rely on decoding.

6 Can the effect of correlations be quantified?

The role of correlations in producing redundancy, or alternatively synergy, among neural signals has been investigated both outside [39] and within [15] the context of population coding. While redundancy is, in common intuition, the default outcome of correlated signals, it is easy to devise situations in which correlations lead to synergy. Consider the toy case of Fig. 3 with 2 cells responding to 3 stimuli. Synergy may result from a positive noise correlation (in the trial to trial variability), if the mean rates to different stimuli are anticorrelated, and viceversa from a negative noise correlation, if the mean rates to different stimuli are positively correlated. When signal and noise correlation are of the same sign, the result is always redundancy. The impact of correlations on redundancy is probably minimal when the mean responses are weakly correlated across the stimulus set (perhaps the 'natural condition'?). Given this realm of possibilities, it is desirable to take an approach, applicable to real data, that enables separating out the information transmitted by individual spikes, emitted by single neurons within an ensemble, from positive or negative contributions due to correlations in firing activity among neurons. One such approach focuses on short time windows [40, 41].

We shall see now how in the limit of what is transmitted over very short windows, a simple formula quantifies the *corrections* to the instantaneous information rate (determined solely by mean firing rates) which result from correlations in spike emission between pairs of neurons. Positive corrections imply synergy, while negative corrections indicate redundancy. The information carried by the population response can be expanded in a Taylor series [23, 40]

$$I(t) = t\, I_t + \frac{t^2}{2}\, I_{tt} + \ldots \tag{8}$$

The first time derivative depends only on the mean rates:

$$I_t \quad = \quad \sum_{i=1}^{C} \left\langle \overline{r}_i(s) \log_2 \frac{\overline{r}_i(s)}{\langle \overline{r}_i(s')\rangle_{s'}} \right\rangle_s \tag{9}$$

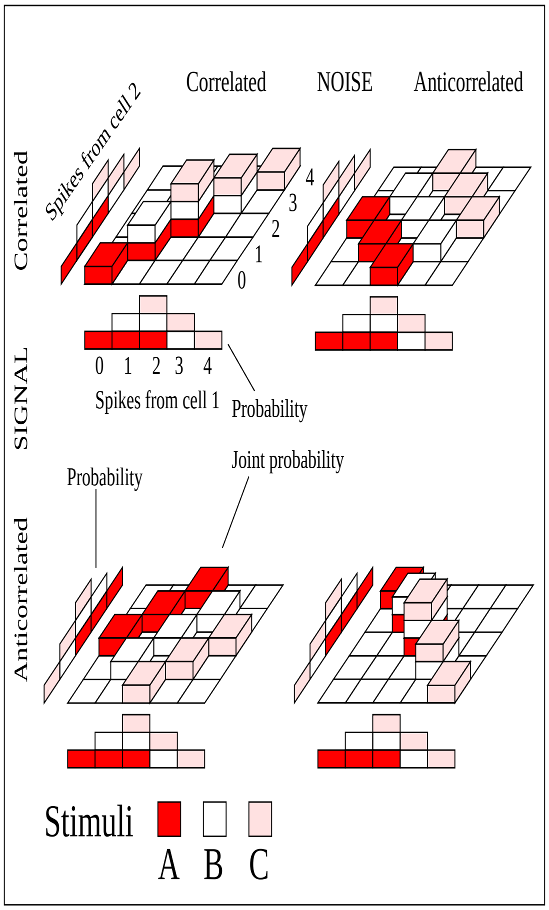

Figure 3: A toy case illustrating possibilities for synergy and redundancy, adapted from Fig. 4 of Ref. [40]. A quick calculation shows that signal and noise both correlated, or both anticorrelated, result in redundancy, while the other two situations produce, given the responses indicated in the figure, synergy.

and it is purely a sum of all single cell contributions, each of the form earlier described by Skaggs and McNaughton [42] and Bialek *et al.* [2] for single cells. The formula clarifies how misguided it is to link a high information *rate* to a high signal-to-noise ratio, which is the conceptual framework tacitly implied in Refs. [35] and [37]. The rate, that is the first derivative of the mutual information, only reflects the extent to which the mean responses of each cell are distributed across stimuli; it does not reflect anything of the variability of those responses, that is of their noisiness, nor anything of the correlations among the mean responses of different cells.

The effect of (pairwise) correlations begins to be felt in the second derivative, instead, and it is convenient then to introduce appropriate measures of such correlations. Pairwise correlations in the response variability ('noise' correlation) can be quantified by

$$\gamma_{ij}(s) = \frac{\overline{r_i(s) r_j(s)}}{\overline{r}_i(s)\overline{r}_j(s)} - 1, \tag{10}$$

i.e. the amount of trial by trial concurrent firing, compared to that expected in the uncorrelated case. The degree of similarity in the mean response profiles of the cells to different stimuli ('signal' correlation) can instead be quantified by

$$\nu_{ij} = \frac{< \overline{r}_i(s)\overline{r}_j(s) >_s}{< \overline{r}_i(s) >_s < \overline{r}_j(s) >_s} - 1. \tag{11}$$

The second derivative, I_{tt}, breaks into 3 components. The first term of I_{tt} depends only on the mean rates and on their correlations:

$$I_{tt}^{(1)} = \frac{1}{\ln 2} \sum_{i=1}^{C} \sum_{j=1}^{C} \langle \overline{r}_i(s) \rangle_s \langle \overline{r}_j(s) \rangle_s \left[\nu_{ij} + (1+\nu_{ij}) \ln(\frac{1}{1+\nu_{ij}}) \right]; \tag{12}$$

the second term is non-zero only when correlations are present in the noise, even if stimulus-independent

$$I_{tt}^{(2)} = \sum_{i=1}^{C} \sum_{j=1}^{C} \left[\langle \overline{r}_i(s)\overline{r}_j(s)\gamma_{ij}(s) \rangle_s \right] \log_2(\frac{1}{1+\nu_{ij}}); \tag{13}$$

the third term contributes only if correlations are stimulus-dependent

$$\begin{aligned} I_{tt}^{(3)} &= \sum_{i=1}^{C} \sum_{j=1}^{C} \langle \overline{r}_i(s)\overline{r}_j(s)(1+\gamma_{ij}(s)) \\ &\times \log_2 \left[\frac{(1+\gamma_{ij}(s)) \langle \overline{r}_i(s')\overline{r}_j(s') \rangle_{s'}}{\langle \overline{r}_i(s')\overline{r}_j(s')(1+\gamma_{ij}(s')) \rangle_{s'}} \right] \Big\rangle_s . \end{aligned} \tag{14}$$

This decomposition still has to be applied extensively to simultaneously recorded neural data. The limited evidence in our hands has not yet revealed a situation in which correlations clearly play a prominent role. Extensive data produced in the laboratories of Eckhorn [43] and Singer [44], which qualitatively point at the importance of correlations, have not, to our knowledge, been analysed in these terms, despite pioneering applications of information theory [45, 46, 47]. A very interesting recent finding [48] could not be quantified properly in terms of information due to the limited sampling available, and it could be reanalysed with the help of this expansion. The expansion has recently being refined in a way that it allow now to assess also the importance of timing relations in pairwise correlations [41]. The crucial question, however, is how soon does the expansion, based on the short time limit, break down. When this occurs, higher order terms in the expansion (starting from those dependent on three-way correlations, and so on) cannot be neglected any longer. The time range of validity of the expansion is thus limited by the requirement that second order terms be small with respect to first order ones, and successive orders be negligible. Since at order n there are C^n terms with C cells, obviously the applicability of the short time limit contracts, in practice, for larger populations. This can be seen in the example from the rat barrel cortex, in Fig. 4. Still, one may ask whether the expansion, at least restricted to second order terms, may afford some insight on neural coding as expressed by large populations of cell.

7 Synergy and redundancy in large populations

Obviously with a few cells, all cases of synergy or redundancy are possible if the correlations are properly engineered – in simulations – or the appropriate special case is recorded – in experiments. The outcome of the information analysis will simply reflect the particularity of each case. With large populations, one may hope to have a better grasp of generic, or typical, cases, more indicative of conditions prevailing at the level of, say, a given cortical module. One may begin by considering a 'null' hypothesis, i.e. that pairwise correlations are purely random, and small in value.

In this null hypothesis, the signal correlations ν_{ij} have zero average, while ν_{ij}^2 could still differ from zero if the ensemble of stimuli used is limited, since a random walk would typically span a range of size $\sqrt{S}$. Then the mean ν_{ij}^2 would decrease with S as $1/S$. The noise correlations might be thought to arise from stimulus independent terms, γ_{ij}, which need not be small, and stimulus dependent contributions $\delta\gamma_{ij}(s)$, which might be expected to get smaller when more trials per stimulus are available, and which on averaging across stimuli would again behave as a random walk.

The effect of such null hypothesis correlations on information transmission can be gauged by further expanding I_{tt} in the small parameters ν_{ij} and $\delta\gamma_{ij}(s)$, i.e. assuming $|\nu_{ij}|^2 << 1$ and $|\delta\gamma_{ij}(s)|^2 << 1$. We consider here a simplified case, in which, for example, all cross terms like $\nu_{ij}\delta\gamma_{ij}$ are taken to vanish; the full derivation will be published elsewhere (Bezzi, Diamond and Treves, to be published). In this case the leading terms in the expansion of $I_{tt}^{(1)} + I_{tt}^{(2)}$ are those quadratic in ν_{ij}

$$I_{tt}^{(1)} + I_{tt}^{(2)} = -\frac{1}{\ln 2}\sum_{i=1}^{C}\sum_{j=1}^{C}(2+\gamma_{ij})\left\langle \overline{r}_i(s)\right\rangle_s \left\langle \overline{r}_j(s)\right\rangle_s \nu_{ij}^2, \tag{15}$$

i.e., contributions to the mutual information which are always negative (indicating *redudancy*). The leading terms in the expansion of $I_{tt}^{(3)}$, if we denote

$$\left\langle f\left[\delta\gamma_{ij}(s)\right]\right\rangle_{\{i,j\},s} = \frac{1}{S}\sum_{i=1}^{S}\frac{\overline{r}_i(s)\overline{r}_j(s)}{\left\langle \overline{r}_i(s)\overline{r}_j(s)\right\rangle_s} f\left[\delta\gamma_{ij}(s)\right] \tag{16}$$

the average over stimuli weighted by the product of the normalized firing rate $\overline{r}_i(s)\overline{r}_j(s)$, and take $\left\langle \delta\gamma_{ij}(s)\right\rangle_{\{i,j\},s}$ to vanish, are

$$I_{tt}^{(3)} = \frac{1}{\ln 2}\sum_{i=1}^{C}\sum_{j=1}^{C}\frac{\nu_{ij}+1}{1+\gamma_{ij}}\left\langle \overline{r}_i(s)\right\rangle_s \left\langle \overline{r}_j(s)\right\rangle_s \left\langle \delta\gamma_{ij}(s)^2\right\rangle_{\{i,j\},s}, \tag{17}$$

that is, contributions to the mutual information which are always positive (indicating *synergy*).

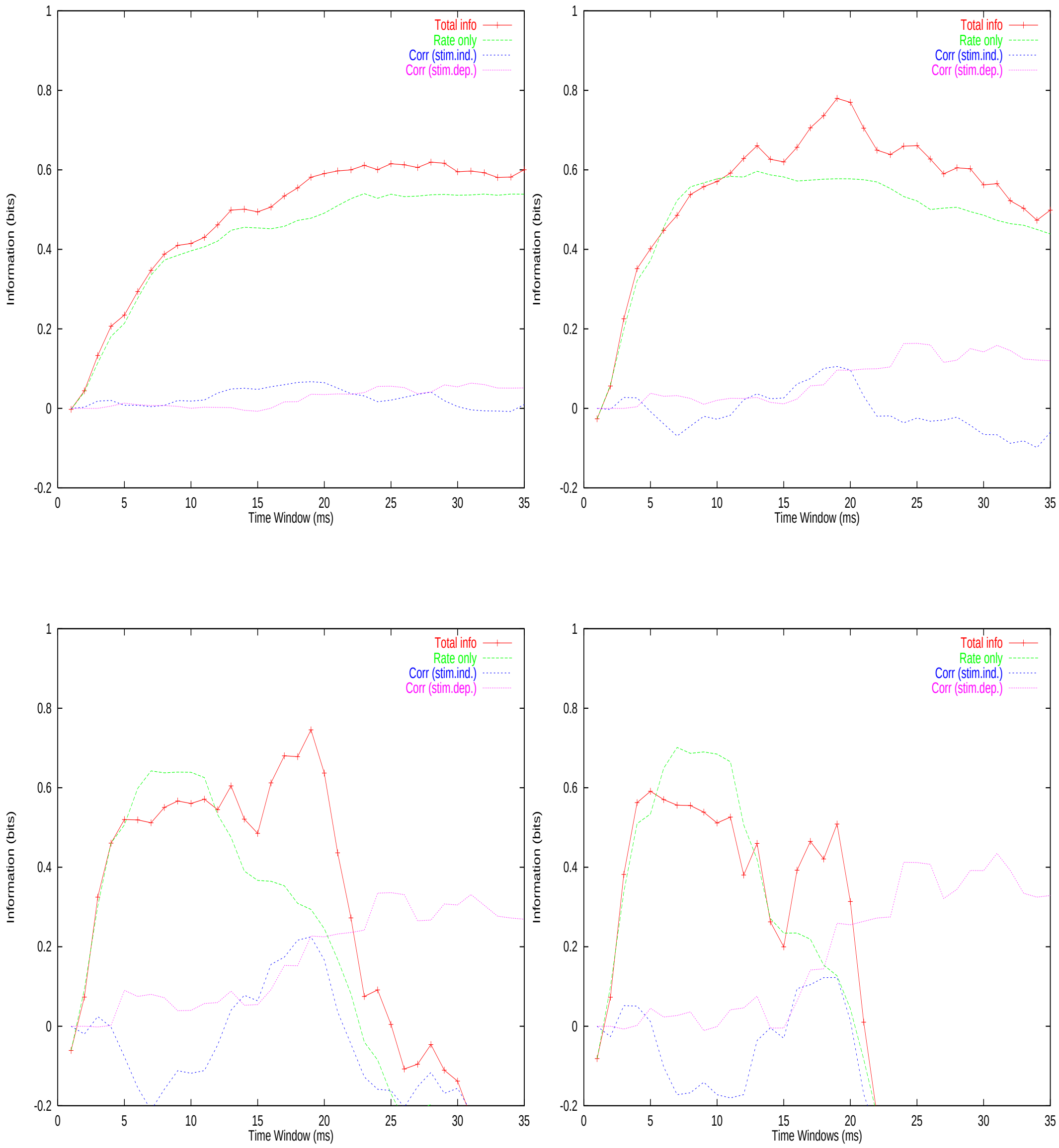

Figure 4: The short-time limit expansion breaks down sooner, the larger the population considered. Cells in rat somatosensory barrel cortex responding to 2 stimuli to the vibrissae. Components of the transmitted information with 3 (top, left), 6 (top, right), 9 (bottom, left) and 12 cells (bottom, right). The first three cases are averaged over 4 sets of cells. Time window: $5-40$ ms. The initial slope (i.e., I_t) is roughly proportional to the number of cells. The effects of the second order terms, quadratic in t, are visible over the brief times between the linear regime and the break down of the expansion. Among several similar data sets analysed, this is close to the worst case, in terms of how soon in time the expansion breaks down.

Thus the leading contributions of the new Taylor expansion are of two types, both coming as $C(C-1)/2$ terms proportional to $\langle \overline{r}_i(s)\rangle_s \langle \overline{r}_j(s)\rangle_s$: the first type, Eq. (15), induces redundancy, and might scale as $1/S$ in our null hypothesis; the second type, Eq. (17), induces synergy, and might scale inversely with the number of trials per stimulus in our null hypothesis. These leading contributions to I_{tt} can be compared to first order contributions to the original Taylor expansion in t (i.e., to the C terms in I_t) in different time ranges. For times $t \approx \mathrm{ISI}/C$, that is $t < \overline{r} > \approx 1/C$, first order terms sum up to be of order one bit, while second order terms are negligible, provided enough stimuli are used and enough trials are available. This occurs however over a time range that becomes shorter as more cells are considered, and the total information conveyed by the population remains of order 1 bit only! For times of the order of the mean interspike interval, $t \simeq \mathrm{ISI}$, first order terms are of order C, while second order ones are of order $C^2 < \nu^2 >$ (with a minus sign, signifying redundancy) and $C^2 < (\delta\gamma)^2 >$ (with a plus sign, signifying synergy), respectively. If $< \nu^2 >$ and $< (\delta\gamma)^2 >$ are not sufficiently small to counteract the additional C factor, these 'random' redundancy and synergy contributions will be substantial. Moreover, over the same time ranges also leading contributions to I_{ttt} and to the next terms in the Taylor expansion in time may be expected to be substantial.

We are therefore led to a surprising conclusion, applying to what is likely the minimum meaningful time range for information transmission, that is the time it takes the typical cell to emit a spike. The conclusion is that a large population of cells, which has not been designed to code stimuli in any particular cooperative manner, may still show large effects of redundancy or synergy, arising simply from random correlations among the firing of the different cells. Such a conclusion reinforces the need for careful experimental studies of the actual correlations prevailing in the neural activity of different parts of the brain. However, it also indicates the importance of considering information decoding along with information encoding: real neurons may not care much for the synergy and redundancy encoded in a multitude of variables they cannot read out, such as the ν_{ij}'s and γ_{ij}'s.

8 Parameters that matter in neuronal representations

What are, then, the variables that real neurons are directly affected by? Clearly, the firing rates r_i, of the neurons they receive inputs from, comprise an important group. Most theoretical analyses of neural networks are grounded on the assumption that the quintessential processing carried out by a single neuron is a dot product operation between the vector of input firing rates and the vector of synaptic weights (cf. [24]). It is the modifiability of individual synaptic weights, and the consequent variance among the synaptic weight vectors of different processing units, that makes individual firing rates important, as even very simplified formal models illustrate. If synaptic weights were taken to be uniform across inputs, the enormous fan-in of cortical connectivity would reduce to a mere device for large sampling. If they were taken to be non-uniform but fixed in time, no new input-output transforms could be established for a given population of cells, so in practice the connectivity would, again, subserve just sampling, except for a few in-built operations. The modifiability of individual synaptic weights, according to so-called Hebbian rules [49] or otherwise, is the cornerstone of the theory of neuron-like parallel distributed processing. Thus, quantitative constraints on memory storage are set by the number of synapses available for individual modification, and in fact they are expressed usually in terms of bits/synapse. In the real brain neurons and synapses operate in vastly more complicated ways than summarized by notions like dot products and synaptic weights. Still, maximising memory storage through maximal synaptic density has been considered by Braitenberg [50] a crucial principle of cortical design.

Individual firing rates are therefore central to neuronal coding because of (long-term) synaptic plasticity, but as there is more to cortical plasticity than synaptic plasticity, so there is likely more to neural codes than individual firing rates. Recently, for example, much attention has been devoted to the exquisite refinement of local inhibitory circuitry in the neocortex (Henri Markram, personal communication, and see [51]). Inhibitory interneurons appear to cluster into some 15 different classes, discriminable on the basis of a combination of morphological, electrophysiological and short-term plasticity properties. Synapses to and from inhibitory interneurons are found to demonstrate long-term plasticity as well. Their connectivity patterns are differentiated also in terms of cortical layers. Although the total numbers of inhibitory neurons and the number and location of their synapse appear unsuitable to make them individually involved in information processing, there is no doubt that they provide for a modulation of cortical dynamics that turns certain collective variables into important parameters of neuronal codes. For example, the average degree of synchronization of an afferent volley to a given cortical patch might be crucial in determining the dynamics of feedforward and feedback inhibition, and consequently the time-course of activation of the pyramidal cells in the patch. This is in contrast to the exact degree of synchronization between any two axons (the dynamical equivalent of a single ν_{ij}), which would itself be relevant only if there were a corresponding modifiable parameter capable of modulating its effects.

At present, our theoretical understanding of neural networks is underdeveloped to deal with such cortical complexities, which are themselves still in the process of being investigated, especially in their dynamical aspects. Despite some promising attempts [52], these are still early days for the elaboration of the appropriate conceptual tools and the identification of the crucial mechanisms and most relevant quantities. At a very basic, and non-dynamical, level, however, it is already clear that the gross statistical features of the distribution of neuronal activity bear a direct import on the efficiency of neuronal codes and of information storage. In the late eighties, a considerable debate between neurophysiologists and modelers centered on the issue of the observed mean level of activity in the cortex, and whether this would make popular models of memory storage totally inappropriate as models of cortical networks [53]. This issue, touching on the first and most basic moment of an abstract 'typical distribution of cortical activity', eventually evaporated when it appeared to be closely linked to the modeling of neurons as binary or sigmoidal units. The *second* moment of such a typical distribution, instead, has a relevance which does not simply stem from crude modelling technology. It was long recognized that the so-called sparseness of the firing, roughly the proportion of cells highly activated at any one time, is a primary determinant of the capacity for memory storage [54, 55]. For non binary units, in particular for real neurons, a generalized measure of the sparseness of their activity can be defined as

$$a = \frac{< r_i >^2}{< (r_i)^2 >} \tag{18}$$

[33, 24]. The more sparse a set of representations expressed by a population of cells ($a \to 0$), the less the representational capacity and the larger the memory capacity of that population, and consistently a is generally found to decrease approaching central memory systems from the sensory periphery [56]. Sparseness is thus a basic and important statistic of neuronal representations, which however does not reflect their interrelationships. To probe the ways in which different representations relate to one another, one must consider other statistics, that go beyond sparseness, and that in fact are linked to information measures.

9 Quantifying the structure of neuronal representations

The structure of neural representations of the outside world has been studied in detail in some simple situations. Typically these are situations in which a well defined correlate of neuronal activity (i.e. a stimulus, a response, or even a behavioural state) is characterized by one or a few parameters that are made to vary continuously or in steps. Examples are the Hubel and Wiesel [57] description of orientation selectivity in cat visual cortex, the O'Keefe [58] finding of place cells in the rat hippocampus, the mitral cell coding of n-aliphatic acid hydrocarbon length in the olfactory system [59], the coding of the direction of movement in 3D-space in the primate motor cortex [60].

In many interesting situations, though, especially in those parts of the brain which are more remote from the periphery, external correlates, or, for simplicity, stimuli, do not vary (either continuously, or in steps) along any obvious physical dimension. Often, in experiments, the set of stimuli used is just a small ensemble of a few disparate individual items, arbitrarily selected and difficult to classify systematically. Examples for the ventral visual system are faces [61], simple or complex [32] abstract patterns, or the schematic objects reached with the reduction procedure of Tanaka *et al* [62]. In such situations, the resulting patterns of neuronal activity across populations of cells can still provide useful insight on the structure of neuronal representations of the outside world, but such insight has to be derived independently of any explicit correlation with a natural, physical structure of the stimulus set.

The only obvious *a priori* metric of the stimulus set, in the general case, is the trivial categorical metric of each element s being equal to itself, and different from any other element in the set. *A posteriori*, the neuronal firing patterns embed the stimulus set into a potentially metric structure defined by the similarities and differences among the patterns, or response vectors, corresponding to the various elements. A truly metric structure can be extracted by quantifying such similarities and differences into a notion of distance (among firing patterns) that satisfies the 3 required relations: positivity, symmetry, the triangle inequality. At a more basic level, though, the overall amount of structure, i.e. the overall importance of relations of similarity and difference among firing patterns, can be quantified even independently of any notion of distance, just from a matrix $Q(s|s')$ characterizing the similarity or confusability of s' with s, a matrix which need not be symmetrical. $Q(s|s')$ can be simply derived from neuronal recordings, after decoding the firing patterns, as the conditional probability $P(s,s')/P(s')$. Whatever the decoding procedure used, $Q(s|s')$ is essentially a measure of the similarity of the current response vector to s' with the mean response vector to s. It is however important to notice that $Q(s|s')$ can also be derived from other measures, for example from behavioural measures of error or confusion in recognition or classification. Behavioural measures of the similarity or confusability of s' with s do not access the representation of the two stimuli directly, but indirectly they reflect the multiplicity of neural representations that are important in generating that particular behaviour. If some of these representations are damaged or lost, as in brain-damaged patients, the resulting behavioural measures can be indicative of the structure of the surviving representations [63].

The amount of structure can be quantified by comparing the mutual information, which in terms of the matrix $Q(s|s')$ reads

$$I = \sum_{s,s' \in \mathcal{S}} Q(s|s')P(s') \log_2 \frac{Q(s|s')}{\sum_{s"} Q(s|s")P(s")} \tag{19}$$

with its minimum and maximum values I_{min} and I_{max} [64] corresponding to a given percent correct $f_{cor} = \sum_s Q(s|s)P(s)$. The lowest information values compatible with a given f_{cor} are those attained when equal probabilities (or equal frequencies of confusion) result for all stimuli

$s \neq s'$. In this case one finds

$$I_{min} = \log_2 S + f_{cor} \log_2 f_{cor} + (1 - f_{cor}) \log_2[(1 - f_{cor})/(S - 1)]. \tag{20}$$

Conversely, maximum information for a given f_{cor} is contained in the confusion matrix when stimuli are confused only within classes of size $1/f_{cor}$, and the individual stimuli within the class are allocated on a purely random basis (for analytical simplicity we consider only unbiased decoding, such that $Q(s|s') \leq Q(s'|s')$, and assume that each class may contain a non integer number of elements). It is easy to see that then

$$I_{max} = \log_2 S + \log_2 f_{cor}. \tag{21}$$

Interpreting the similarity, or probability of confusion, as a monotonically decreasing function of some underlying distance (e.g. as discussed above), the first situation can be taken to correspond to the limit in which the stimuli form an equilateral simplex, or equivalently the stimulus set is drawn from a space of extremely high dimensionality. In the Euclidean $d \to \infty$ limit, points drawn at random from a finite e.g. hyperspherical region tend to be all at the same distance from each other, and from the point of view of the metric of the set this is the *trivial* limit mentioned above. The second situation can be taken to correspond to the *ultrametric* limit, instead, in which all stimuli at distance less than a critical value from each other form clusters such that all distances between members of different classes are above the critical value. This is a non-Euclidean structure (although it could be embedded in a Euclidean space of sufficiently large dimension), and it is a first example of the possible emergence of non-Euclidean aspects from a quantitative analysis that does not rely on *a priori* assumptions.

Intermediate situations between the two extremes are easy to imagine, and can be parametrized in a number of different ways. A convenient parameter that simply quantifies the relative amount of information in excess of the minimum, without having to assume any specific parametrization for the $Q(s|s')$ matrix, is

$$\lambda = \frac{I - I_{min}}{I_{max} - I_{min}} \tag{22}$$

which ranges from 0 to 1 (for unbiased confusion; it can be above 1 if confusion is biased) and can be interpreted as measuring the *metric content* of the matrix. What is quantified by λ can be called the metric content not in the sense that it requires the introduction of a real metric, but simply because it gives the degree to which relationships of being close or different (distant), among stimuli, emerge in the $Q(s|s')$ matrix. For $\lambda = 0$ such relationships are irrelevant, to the point that if confusion occurs, it can be with any (wrong) stimulus. For $\lambda = 1$ close stimuli are so similar as to be fully confused with the correct one, whereas other stimuli are 'maximally distant' and never mistaken for it.

In summary, the metric content index λ quantifies the dispersion in the distribution of 'errors', from maximal, $\lambda = 0$, to minimal, $\lambda = 1$. The 'errors' may be actual behavioural errors in identifying or categorizing stimuli or in producing appropriate responses, or simply calculated from the similarity in the response vectors of a population of cells to different stimuli. Two examples of application of the metric content index, in the second situation, are illustrated in Fig. 5. The analyses summarized in the graphs of Fig. 5 point at two important aspects of the metric content index: its being a relatively intrinsic property of a representation (invariant across the number of cells sampled, within sampling precision) and its variation from one population or cortical area to another. The neuronal recordings are described elsewhere (continuous simultaneous recordings of 42 rat hippocampal CA1 cells, with the rat running a triangular maze, divided in windows 250 *msec* long [38]; and continuous but not simultaneous recordings of 27

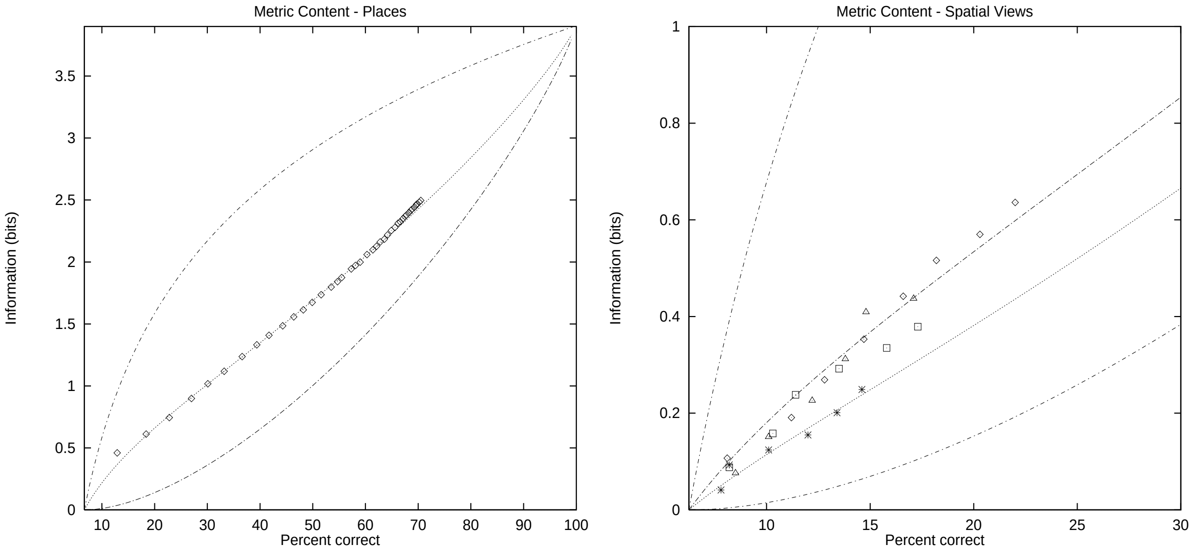

Figure 5: The information decoded from different cell populations *vs.* the corresponding percent correct, in the rat (left) and monkey (right) hippocampus. In both cases different data points with the same symbol correspond to increasing the number of cells included in each population, thus raising percent correct and information. I_{min} and I_{max} are indicated. The rat example illustrate how metric content is a relatively invariant measure (the third curve is for $\lambda = 0.36$) across population sizes. The monkey example indicates quantitative differences among neighbouring populations (the 2 curves are for $\lambda = 0.25$ and $\lambda = 0.15$): datapoints are for populations of CA3 ($*$), CA1 (triangles), parasubiculum (squares) and parahippocampal gyrus cells (diamonds).

monkey cells from the 4 regions indicated, with the monkey freely locomoting in the laboratory, divided in windows 100 $msec$ long [26, 65]).

It should be noted that the similarity matrix is based on response vectors quite different from Georgopoulos' population vectors [60], which live in the physical 3D or 2D movement space rather than in the space of dimensionality equal to the number of cells included, and which correspond to a continuous rather than a discretized correlate. One can see from the figure the extent to which metric content, considering the imprecision with which cells are sampled, their activity is recorded and the information measurss are extracted, is still a relatively stable index. This allows some comparisons to be made even among the metric content characterizing vectors of different dimensionality. For each given cortical area, as more cells are considered, both percent correct and decoded information grow, and the relation between the two, expressed as metric content, varies somewhat, but in a limited band of values characterizing each cortical area. These data, particularly those obtained in the monkey, are not fully adequate, on at least two accounts. First, the number of cells recorded and the number of trials available for each cell and each spatial view were not sufficiently large to safely avoid limited sampling effects. Second, the monkey recordings were not simultaneous. Both inadequacies can be removed with parallel recording from several cells at once, as has become now standard practice in a number of laboratories.

Within these limits, one possible interpretation of the different metric content in the CA3 area, with respect to the other 3 areas sampled, lies in the different pattern of connectivity, whereby in CA3 recurrent collateral connections are the numerically dominant source of inputs to pyramidal cells, and travel relatively long distance, to form an extended network connected by intrinsic circuitry. Considerations based on simplified network models suggest that such a connectivity pattern would express memory representations with a different metric structure from those expressed by networks of different types. The difference could be further related to the qualitative nature of the memory representation, which might be characterized as being more *episodic* in CA3 and more *structured* in the other areas. The metric content depends also on the average sparseness of these representations, though, and further analyses are required to dissociate the effects of connectivity (and of representational structure) from those purely due to changes in sparseness. In particular, it has been shown that in the short-time limit the metric content becomes a transparent function of sparseness [66], and it is possible that even over the 250 $msec$ windows used for the rat, the structure revealed reflects mainly the sparseness of the coding.

The monkey recordings were from neighbouring areas in the temporal lobes, and it is possible that any difference among memory representations will be more striking when more distant areas are compared. In addition, it is possible that any difference may be more striking when the correlate considered does not have its own intrinsic metric, as with spatial views, but instead lives in a high dimensional space, as e.g. with faces, thereby letting more room for arbitrary metric structures to be induced in the neural representations by the learning process. For both reasons, it is interesting to extend this analysis to entirely different experiments, sharing with these only the generic requirement that different populations of cells are recorded in their response to the same set of stimuli, or in general correlates. It is also interesting to deepen the analysis of the structure of representations by looking at subtler aspects, such as the *ultrametric* content [64], that depends on the mutual relations of triplets, rather than pairs, of representations.

Finally, possible changes in the representations that develop with time can examined by recording from the *same* populations – not the same cells – over periods during which some behaviourally relevant phenomenon may have occured, such as new learning, forgetting, or a modulation of the existing representations. One specific such modulation of interest for the case of human patients is the one resulting from localized lesions to another cortical area, which may affect the structure of the representations in surviving areas of the cortex.

Acknowledgments

The analyses and procedures discussed in this chapter have been developed together with several colleagues, as evident from the citations, among them Stefano Panzeri, Edmund Rolls and William Skaggs. Collaborations were supported by the European Commission and the Human Frontier Science Program.

References

[1] C. E. Shannon. A mathematical theory of communication. *AT&T Bell Labs. Tech. J.*, 27:379–423, 1948.

[2] W. Bialek, F. Rieke, R. R. de Ruyter van Steveninck, and D. Warland. Reading a neural code. *Science*, 252:1854–1857, 1991.

[3] A.A. Frolov and I.P. Murav'ev. Informational characteristics of neural networks capable of associative learning based on hebbian plasticity. *Network*, 4:495–536, 1993.

[4] A. Treves and S. Panzeri. The upward bias in measures of information derived from limited data samples. *Neural Comp.*, 7:399–407, 1995.

[5] S. Panzeri and A. Treves. Analytical estimates of limited sampling biases in different information measures. *Network*, 7:87–107, 1996b.

[6] D. Golomb, J.A. Hertz, S. Panzeri, A. Treves, and B.J. Richmond. How well can we estimate the information carried in neuronal responses from limited samples? *Neural Comp.*, 9:649–655, 1997.

[7] M. W. Oram and D. I. Perrett. Time course of neuronal responses discriminating different views of face and head. *J. Neurophysiol.*, 68:70–84, 1992.

[8] L. M. Optican and B. J. Richmond. Temporal encoding of two-dimensional patterns by single units in primate inferior temporal cortex: Iii information theoretic analysis. *J. Neurophysiol.*, 57:162–178, 1987.

[9] L. M. Optican, T. J. Gawne, B. J. Richmond, and P. J. Joseph. Unbiased measures of transmitted information and channel capacity from multivariate neuronal data. *Biological Cybernetics*, 65:305–310, 1991.

[10] E.N. Eskandar, B.J. Richmond, and L.M. Optican. Role of inferior temporal neurons in visual memory: I. temporal encoding of information about visual images, recalled images, and behavioural context. *J. Neurophysiol.*, 68:1277–1295, 1992.

[11] T. W. Kjaer, J. A. Hertz, and B. J. Richmond. Decoding cortical neuronal signals: networks models, information estimation and spatial tuning. *J. Comput. Neurosci.*, 1:109–139, 1994.

[12] J. Heller, J. A. Hertz, T. W. Kjaer, and B. J. Richmond. Information flow and temporal coding in primate pattern vision. *J. Comput. Neurosci.*, 2:175–193, 1995.

[13] M. J. Tovée, E. T. Rolls, A. Treves, and R. J. Bellis. Information encoding and the responses of single neurons in the primate temporal visual cortex. *J. Neurophysiol.*, 70:640–654, 1993.

[14] F. Mechler, J. D. Victor, K. P. Purpura, and R. Shapley. Robust temporal coding of contrast by v1 neurons for transient but not for steady-state stimuli. *J. Neurosci.*, 18:6583–6598, 1998.

[15] M. W. Oram, M. C. Wiener, R. Lestienne, and B. J. M. Richmond. Stochastic nature of precisely timed spike patterns in visual system neuronal responses. *J. Neurophysiol.*, 81:3021–3033, 1999.

[16] J. O'Keefe and M. L. Recce. Phase relationship between hippocampal place units and the eeg theta rhythm. *Hippocampus*, 3:317–330, 1993.

[17] M. V. Tsodyks, W. E. Skaggs, T. J. Sejnowski, and B. L. McNaughton. Population dynamics and theta rhythm phase precession of hippocampal place cell firing: a spiking neuron model. *Hippocampus*, 6:271–280, 1996.

[18] W. E. Skaggs, B. L. McNaughton, M. A. Wilson, and C. A. Barnes. Theta phase precession in hippocampal neuronal populations and the compression of temporal sequences. *Hippocampus*, 6:149–172, 1996.

[19] M.-N. Chee-Orts and L. M. Optican. Cluster method for analysis of transmitted information in multivariate neuronal data. *Biol. Cybern.*, 69:29–35, 1993.

[20] M. Page. Connectionist modeling in psychology: a localist manifesto. *Behavioral Brain Sciences*, 23:in press, 2000.

[21] E. T. Rolls, A. Treves, M. J. Tovée, and S. Panzeri. Information in the neuronal representation of individual stimuli in the primate temporal visual cortex. *J. Comput. Neurosci.*, 4:309–333, 1997b.

[22] M. R. DeWeese and M. Meister. How to measure the information gained from one symbol. *Network*, 11:in press, 2000.

[23] S. Panzeri, G. Biella, E. T. Rolls, W. E. Skaggs, and A. Treves. Speed, noise, information and the graded nature of neuronal responses. *Network*, 7:365–370, 1996a.

[24] E. T. Rolls and A. Treves. *Neural Networks and Brain Function.* Oxford University Press, Oxford, UK, 1998.

[25] E.T Rolls, H.D. Critchley, and A. Treves. Representation of olfactory information in the primate orbitofrontal cortex. *J. Neurophysiol.*, 75:1982–1996, 1996.

[26] E. T. Rolls, A. Treves, R. G. Robertson, P. Georges-Francois, and S. Panzeri. Information about spatial views in an ensemble of primate hippocampal cells. *J. Neurophysiol.*, 79:1797–1813, 1998.

[27] E. D. Gershon, M. C. Wiener, P. E. Latham, and B. J. Richmond. Coding strategies in monkey V1 and inferior temporal cortices. *J. Neurophysiol.*, 79:1135–1144, 1998.

[28] W. B. Levy and R. A. Baxter. Energy efficient neural codes. *Neural Comp.*, 8:531–543, 1996.

[29] R. J. Baddeley, L. F. Abbott, M. Booth, F. Sengpiel, T. Freeman, E. A. Wakeman, and E. T. Rolls. Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc. R. Soc. Lon. Ser. B*, 264:1775–1783, 1997.

[30] Alessandro Treves, Stefano Panzeri, Edmund T Rolls, Michael C A Booth, and Edward A Wakeman. Firing rate distributions and efficiency of information transmission of inferior temporal cortex neurons to natural visual stimuli. *Neural Comp.*, 11:611–641, 1999.

[31] N. Brunel and J. P. Nadal. Mutual information, fisher information and population coding. *Neural Comp.*, 10:1731–1757, 1998.

[32] Y. Miyashita and H. S. Chang. Neuronal correlate of pictorial short-term memory in the primate temporal cortex. *Nature*, 331:68–70, 1988.

[33] A. Treves and E. T. Rolls. What determines the capacity of autoassociative memories in the brain. *Network*, 2:371–397, 1991.

[34] T. J. Gawne and B. J. Richmond. How independent are the messages carried by adjacent inferior temporal cortical neurons? *J. Neurosci.*, 13:2758–2771, 1993.

[35] E. Zohary, M. N. Shadlen, and W. T. Newsome. Correlated neuronal discharge rate and its implication for psychophysical performance. *Nature*, 370:140–143, 1994.

[36] E. T. Rolls, A. Treves, and M. J. Tovée. The representational capacity of the distributed encoding of information provided by populations of neurons in the primate temporal visual cortex. *Exp. Brain Res.*, 114:149–162, 1997a.

[37] P. M. Gochin, M. Colombo, G. A. Dorfman, G. L. Gerstein, and C. G. Gross. Neural ensemble encoding in inferior temporal cortex. *J. Neurophysiol.*, 71:2325–2337, 1994.

[38] Alessandro Treves, William E Skaggs, and Carol A Barnes. How much of the hippocampus can be explained by functional constraints? *Hippocampus*, 6:666–674, 1996b.

[39] N. Brenner, S. P. Strong, R. Koberle, and W. Bialek. Synergy in a neural code. *Neural Comp.*, 12:in press, 2000.

[40] Stefano Panzeri, Simon R Schultz, Alessandro Treves, and Edmund T Rolls. Correlations and the encoding of information in the nervous system. *Proc. Roy. Soc. B*, 266:1001–1012, 1999.

[41] Stefano Panzeri and Simon R Schultz. A unified approach to the study of temporal, correlational and rate coding. *Physics arXiv.org*, page 9908027, 1999.

[42] W E Skaggs and B L McNaughton. Quantification of what it is that hippocampal cell firing encodes. In *Soc. Neurosci. Abstr.*, page 1216, 1992.

[43] R. Eckhorn, R. Bauer, W. Jordan, M. Brosch, W. Kruse, M. Munk, and H.J. Reitboeck. Coherent oscillations: a mechanism of feature linking in the visual cortex? *Biol. Cybern.*, 60:121–130, 1988.

[44] C.M. Gray, P. Konig, A.K. Engel, and W. Singer. Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. *Nature*, 338:334–337, 1989.

[45] R. Eckhorn and B. Pöpel. Rigorous and extended application of information theory to the afferent visual system of the cat. i. basic concepts. *Kybernetik*, 16:191–200, 1974.

[46] R. Eckhorn and B. Pöpel. Rigorous and extended application of information theory to the afferent visual system of the cat. ii. experimental results. *Kybernetik*, 17:7–17, 1975.

[47] R. Eckhorn, O-J. Grüsser, J. Kröller, K. Pellnitz, and B. Pöpel. Efficiency of different neural codes: information transfer calculations for three different neuronal systems. *Biol. Cybern.*, 22:49–60, 1976.

[48] E. M. Maynard, N. G. Hatsopoulos, C. L. Ojakangas, B. D. Acuna, J. N. Sanes, R. A. Normann, and J. P. Donoghue. Neuronal interactions improve cortical population coding of movement direction. *J. Neurosci.*, 19:8083–8093, 1999.

[49] D. O. Hebb. *The Organization of Behavior.* Wiley, New York, 1948.

[50] V. Braitenberg and A. Shüz. *Anatomy of the Cortex: Statistics and Geometry.* Springer Verlag, Berlin, 1991.

[51] Y. Wang, A. Gupta, and H. Markram. Anatomical and functional differentiation of glutamatergic synaptic innervation in the neocortex. *Journal of Physiology (Paris)*, 93:305–317, 1999.

[52] R. J. Douglas and K. A. Martin. A functional microcircuit for cat visual cortex. *Journal of Physiology (London)*, 440:735–769, 1991.

[53] Daniel J Amit and Alessandro Treves. Associative memory neural network with low temporal spiking rates. *Proceedings of the National Academy of Sciences of the USA*, 86:7871–7875, 1989.

[54] M. V. Tsodyks and M. V. Feigel'man. The enhanced storage capacity in neural networks with low activity level. *Europhysics Letters*, 6:101–105, 1988.

[55] J. Buhmann, R. Divko, and K. Schulten. Associative memory with high information content. *Physical Review*, A 39:2689–2692, 1989.

[56] Edmund T Rolls and Alessandro Treves. The relative advantages of sparse versus distributed encoding for associative neuronal networks in the brain. *Network*, 1:407–421, 1990.

[57] D.H. Hubel and T.N. Wiesel. Sequence regularity and geometry of orientation columns in the monkey striate cortex. *Journal of Comparative Neurology*, 1582:267–294, 1974.

[58] J. O'Keefe. A review of the hippocampal place cells. *Progress in Neurobiology*, 13:419–439, 1979.

[59] S.L. Sullivan and L. Dryer. Information processing in mammalian olfactory system. *Journal of Neurobiology*, 30:20–36, 1996.

[60] A. P. Georgopoulos, A. Schwartz, and R. E. Kettner. Neural population coding of movement direction. *Science*, 233:1416–1419, 1986.

[61] E. T. Rolls. Neurophysiological mechanisms underlying face processing within and beyond the temporal cortical visual areas. *Philosophical Transactions of the Royal Society (London)*, B 335:11–21, 1992.

[62] K. Tanaka. Neuronal mechanisms of object recognition. *Science*, 262:685–688, 1993.

[63] Rosapia Lauro-Grotto, Carolina Piccini, Francesca Borgo, and Alessandro Treves. What remains of memories lost in Alzheimer and herpetic encephalitis. *Society for Neuroscience abstract 734.2*, 23:1889, 1997.

[64] A. Treves. On the perceptual structure of face space. *BioSystems*, 40:189–196, 1997.

[65] Alessandro Treves, Pierre Georges-Francois, Stefano Panzeri, Robert G Robertson, and Edmund T Rolls. The metric content of spatial views as represented in the primate hippocampus. In V Torre and J Nicholls, editors, *Neural Circuits and Networks*, NATO Asi Series F, Computer and Systems Sciences, Vol 167, pages 239–247, Berlin, 1998. Springer.

[66] Stefano Panzeri, Alessandro Treves, Simon R Schultz, and Edmund T Rolls. On decoding the responses of a population of neurons from short time windows. *Neural Comp.*, 11:1553–1577, 1999.