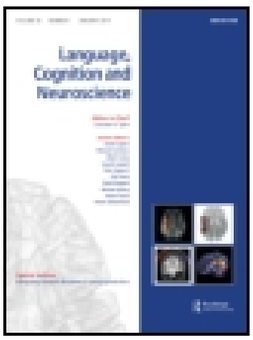

Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

ISSN: 2327-3798 (Print) 2327-3801 (Online) Journal homepage: http://www.tandfonline.com/loi/plcp21

A null model for cortical representations with grandmothers galore

Gabriel Kreiman

To cite this article: Gabriel Kreiman (2016): A null model for cortical representations with grandmothers galore, Language, Cognition and Neuroscience, DOI: 10.1080/23273798.2016.1218033

To link to this article: http://dx.doi.org/10.1080/23273798.2016.1218033

Published online: 09 Aug 2016.

LANGUAGE, COGNITION AND NEUROSCIENCE, 2016
http://dx.doi.org/10.1080/23273798.2016.1218033

Routledge
Taylor & Francis Group

A null model for cortical representations with grandmothers galore

Gabriel Kreiman

Children's Hospital, Harvard Medical School, Boston, MA, USA

ABSTRACT

There has been extensive discussion in the literature about the extent to which cortical representations can be described as localist or distributed. Here, we discuss a simple null model that encompasses a family of related architectures describing the transformation of signals throughout the parts of the visual system involved in object recognition. This family of models constitutes a rigorous first approximation to explain the neurophysiological properties of ventral visual cortex. This null model contains both distributed and local representations throughout the entire hierarchy of computations and the responses of individual units are meaningful and interpretable when encoding is adequately defined for each computational stage.

ARTICLE HISTORY
Received 21 March 2016
Accepted 17 July 2016

KEYWORDS
Computational models; visual recognition; sparse coding; localist representation; human visual cortex

How does the brain represent information? A cartoonish answer was entertainingly and endearingly depicted in a recent Pixar film, Inside Out, where we get to peek inside the mind of a young girl, and find five characters, each representing basic emotions: Joy, Sadness, Fear, Disgust and Anger. The notion of a little character inside the brain to represent each emotion could easily be extended by adding more and more characters to represent other sensations, thoughts, actions, decisions, even grandmothers. There could be one character that gets activated whenever we see a chair, another one for grandma, another one whenever we listen to Beethoven, another one whenever we are hungry and another one when we raise our right arm. Despite the seemingly naïve nature of this scheme, after replacing "characters" by "neurons", this idea seems to be at the heart of how lay people and many Cognitive Scientists and Neuroscientists think about brain representations.

I argue here that this cartoon version of cortical representations is neither completely wrong nor completely right. We need to dig deeper into what we mean by representations and translate those definitions into the biophysical language of neurons. What emerges is an even simpler, and yet far more wonderful and elegant biologically plausible description of how brains store information, *a null model for cortical representations*. I will focus the discussion here on visual information, specifically on how we see and recognise shapes, objects and faces. I restrict the discussion to visual objects because we know more about the visual system and the transformation of inputs along the visual hierarchy than about any other modality or any other aspect of cognition. This should not be interpreted to imply that we fully understand visual object processing; there remain fundamental unanswered questions in the field. Despite these lacunae, we are beginning to develop biologically inspired computational models that provide a reasonable approximation and first sketch of visual processing.

I will argue that the term "grandmother cell" and several derivatives and synonyms have *not* been well defined (Gross, 2002). In the Cognitive Science literature, the question about information representation has elicited significant discussion between so-called "localist" representations (McClelland & Rumelhart, 1981) and "distributed" representations (McClelland, Rumelhart & PDP Research Group, 1986); see Bowers (2009) for a clear and detailed review and discussion of the literature. It seems clear to most investigators that there is no one-to-one map between a single neuron and perception (i.e. it cannot be the case that there is only one neuron in the entire brain that responds to a given photograph of grandma and it cannot be the case that a given neuron responds only to one particular photograph of grandma and to no other possible photograph of grandma or other stimuli). Once we abandon the idea of a one-to-one map, one formulation of the question about how information is represented in the brain asks whether single neurons store "meaningful things" or not. For example, it could be argued that the word *grandmother* constitutes a meaningful thing and we can ask whether there are neurons that respond selectively to *grandmother* and not to other stimuli. Toy computational models have been built where there are units that

CONTACT Gabriel Kreiman gabriel.kreiman@tch.harvard.edu

represent letters or words. Yet, when discussing real brains, this apparently simple question is not sufficiently well defined. Do we mean to imply that there is a single neuron in the entire brain that responds to *grandmother*? What do we mean by "neuronal response"? How selective does the neuron need to be in this definition? Do we mean to imply that the neuron(s) would respond to grandmother above baseline and to no other possible stimulus in the world? Would the neuron(s) respond to the word "grandma" and the misspelled word "grandmotter"? Would the neuron respond to a picture of grandma, to any picture of grandma, to any grandma, to grandma's clothes, to her gait, her voice, her accent, etc.? The universe of related questions is infinite and so is the list of discussions that can be generated around such poor definitions. Instead, I will provide a brief, and admittedly incomplete, description of the cascade of processes that lead to representing visual information, arguing that this involves both *implicit* and *explicit* representations and that ultimately defining these operations is the crux of the problem to understand cortical representations.

Basic tenets of representation

The basic requirements for representing information in the brain have been lucidly articulated by several authors before (e.g. Barlow, 1972; Crick & Koch, 2003; Parker & Newsome, 1998; among others). To constrain the discussion here, we consider an experiment where a visual stimulus is presented (e.g. a picture of grandma), the subject reports his/her perception (e.g. indicating yes/no whether he/she recognised grandma), and we scrutinise the responses of multiple neurons. I will re-state the main postulates by copying, expanding, rephrasing and adding examples to the ideas in Parker and Newsome (1998), to define the basic tenets of representation of visual information in cortex:

1. *Reflection of perception*. The neuronal responses elicited by presentation of the stimulus should be *directly comparable to perception and behaviour* as reported by the subject in terms of *magnitude, timing, duration and specificity*. Even though we obviously cannot see grandma without a retina, this tenet implies that the retinal ganglion cells (RGC) are *not* part of the explicit visual representation of grandma.
2. *Timing*. The neuronal responses should occur *before or at the moment of* perception (and not afterwards). This is a basic and fundamental definition of causality and has implications for how neuronal responses are defined, analysed and interpreted. Visual recognition of an object flashed on the screen takes place within approximately 150 ms of stimulus onset (Kirchner & Thorpe, 2006; Potter & Levy, 1969; Thorpe, Fize, & Marlot, 1996) and therefore the neuronal responses should occur within this time frame.
3. *Selectivity*. The neurons in question should signal *relevant* and *selective* information *when and only when* the organism perceives the stimulus (in the scenario described here, a two-alternative forced choice discrimination of whether grandma is present). There are abundant studies documenting selective responses to stimuli throughout visual cortex, for example, in primary visual cortex (Hubel & Wiesel, 1962), middle temporal (MT) cortex (Baker, Petersen, Newsome, & Allman, 1981) and inferior temporal cortex (ITC) (Desimone, Albright, Gross, & Bruce, 1984). Throughout the brain, neurons fire spontaneously even in the absence of their preferred stimuli. Here, we focus on elevated responses beyond this baseline firing.
4. *Stability*. The neuronal representation should be stable over *prolonged periods of time*. The representation of grandma should be the same today and a month from now. This is not to say that the brain cannot acquire or forget information. Plasticity is a fundamental property of cortical circuits but there should exist a stable representation of information in cortex. Recent studies have shown that the representation of visually selective information can be very stable (e.g. Bansal et al., 2012; Bondar, Leopold, Richmond, Victor, & Logothesis, 2009; McMahon, Jones, Bondar, & Leopold, 2014; Tolias et al., 2007).
5. *Single trial interpretability*. Differences in the firing patterns of the candidate neurons (or a subset thereof) to different external stimuli should be sufficiently reliable in a statistical sense to account for, and be reconciled with, the precision of the organism's responses in single trials. In other words, we should be able to reliably read-out information about the stimulus from the neuronal responses, in single trials, and with a linear read-out mechanism. The imposition of linearity here is important; the retina contains all the information about the picture of grandma but it cannot be linearly read out. Machine learning techniques have been extensively used to extract selective visual information in single trials (e.g. Hung, Kreiman, Poggio, & DiCarlo, 2005). The imposition of single trials interpretability is critical: the brain cannot average across trials to make a decision (the question of noise in neuronal responses is further elaborated upon in the "Discussion" section).

6. *Predictive of perceptual fluctuations*. Fluctuations in the firing of some set of the candidate neurons to repeated presentation of identical external stimuli should be predictive of the observer's judgement on individual stimulus presentations including errors and other behavioural changes (e.g. Newsome, Britten, & Movshon, 1989). For example, in the context of bistable perception, an observer's subjective report for the presence or absence of the stimulus should correlate with the neuronal responses, as observed in ITC (Sheinberg & Logothetis, 1997) but not in primary visual cortex (Leopold & Logothetis, 1996).
7. *Invariance*. The representation should be robust to certain transformations of the input. For example, in the context of recognising an object, recognition is essentially unaffected by a wide range of changes in the object's size, illumination, rotation and other transformations. Robustness is not complete: it is possible to disrupt recognition if the image transformation is sufficiently large. Yet, recognition is remarkably robust to image changes and the neuronal representation should reflect this degree of invariance (e.g. Deco & Rolls, 2004; DiCarlo, Zoccolan, & Rust, 2012; Ito, Tamura, Fujita, & Tanaka, 1995; Riesenhuber & Poggio, 1999; Sary, Vogels, & Orban, 1993).
8. *Task independence*. The firing patterns of the neurons in question should not be affected by how behaviour is reported. Grandma is grandma, regardless of whether we ask subjects to indicate yes/no by pressing buttons, to verbally report her name or to passively view the picture. There are of course plenty of neurons in cortex that are strongly modulated by task demands, but there should exist a stable representation that is task-independent. For example, to a first approximation, neurons in ITC carry information about object shapes that is conveyed to pre-frontal cortex where investigators have described strong modulation depending on the task demands (Freedman, Riesenhuber, Poggio, & Miller, 2001; Meyers, Freedman, Kreiman, Miller, & Poggio, 2008). Task demands can modulate the representation (e.g. visual attention can significantly enhance neuronal activity throughout visual cortex (Reynolds & Chelazzi, 2004)) but it should still be possible to decode the information about grandma in a task-invariant manner.
9. *Susceptibility to stimulation*. Direct interference with the firing patterns of some set of the candidate neurons (e.g. by electrical, chemical or optogenetic stimulation) should lead to measurable changes at the perceptual level. These perceptual changes should be contingent on the spatial and temporal specificity of the external manipulation. Current injection has been shown to bias behavioural responses in monkeys in a content-specific fashion in different parts of visual cortex, (e.g. Afraz, Kiani, & Esteky, 2006; Salzman, Britten, & Newsome, 1990).
10. *Susceptibility to lesions*. Temporary or permanent removal of all or part of the candidate set of neurons should lead to a measurable perceptual deficit (e.g. Afraz, Boyden, & DiCarlo, 2015; Dean, 1976; Holmes, 1918).

Starting from the very beginning

When photons impinge on the eyes, RGC respond vigorously if there is a change in luminosity within their receptive fields. To get right to the point: is this a grandmother-like representation or a distributed one? The question is never even discussed among retinal physiologists and hardly makes sense. We have a rather elaborate understanding of how photons are transformed into electrical signals and we are beginning to elucidate the circuitry that leads to RGC firing at a very fine level of detail (e.g. Field et al., 2010). The picture of grandma is *implicitly* represented by the activity of a large number of RGCs; one could rephrase this statement to argue that the information about a complex visual stimulus such as grandma is distributed over multiple RGCs. At the same time, the representation of light changes in a specific location is *explicitly* represented by RGCs. The RGCs therefore behave like grandmother cells in terms of explicitly representing changes in luminosity within their receptive field. Coarsely speaking, we can represent the firing rate r of a given RGC as $r = g\left(\sum_{i=1}^{N} w_i x_i\right)$, where g is a non-linearity (e.g. a threshold), w_i denote the synaptic weights of the N units that project onto the RGC, and x_i denote the activities of the N inputs. Although admittedly oversimplified, this formulation forms the foundation of the basic intuition of how neurons integrate information and remains at the heart of models of network activity (Dayan & Abbott, 2001).

RGCs project to neurons in the lateral geniculate nucleus (LGN), and these cells in turn project to primary visual cortex (V1). Neurons in primary visual cortex respond vigorously and selectively to oriented bars presented within their receptive field. Neurons in primary visual cortex *do not respond uniquely to a single orientation*. A neuron that shows maximal response to a bar of, say, 45-degree orientation will also respond to bars of 44-degree and 46-degree orientation. There is a tuning curve with a certain width that

describes how selective the neuron is for different orientations. Investigators have described *simple* and *complex* cells in V1, the latter presumably pooling information from multiple simple cells to achieve robustness: whereas the responses of a simple cell are sensitive to the exact location of the oriented bar within its receptive field, a complex cell will yield approximately the same response when the bar is translated to different positions within the receptive field. Again, in the study of primary visual cortex the discussion about grandmother cell-like representations is notably absent. There are multiple computational models that provide a reasonable account of how orientation tuning emerges in V1 (e.g. Carandini et al., 2005; Hubel & Wiesel, 1962; Olshausen & Field, 1996; Priebe & Ferster, 2012). The fundamental questions to completely understand the responses of neurons in layer 4 of primary visual cortex involve fully elucidating the activity of the relevant inputs (x_i in the formulation above), the strength of the relevant connections (w_i in the formulation above) and the specific nature of the computations including thresholds, time constants, synaptic adaptation, synaptic depression and dendritic non-linearities. Information about grandma remains distributed over a large number of V1 cells. We cannot linearly decode the presence of grandma from an ensemble of V1 neurons. However, the presence of an oriented bar at a specific location can be linearly and robustly decoded in single trials from the activity of a small number of V1 neurons.

In sum, in the pathway from the retina to V1, information is represented both in a distributed way and in a grandmother-like way. If we define neuronal response as an enhanced firing rate beyond baseline firing and we define a "meaningful thing" as a stimulus with the right contrast, colour, orientation and spatial frequency localised within the receptive field of a given V1 neuron, then this neuron behaves like a grandmother cell. The representation is redundant because there are multiple similar neurons and is achieved by virtue of integrating the activity over a distributed representation of the same stimulus at the level of the LGN.

The ventral visual cortex

Ascending through the visual hierarchy, we have a decent idea of the structures and anatomical pathways that transform visual information along the so-called *what* pathway involved in object recognition in macaque monkeys (Connor, Brincat, & Pasupathy, 2007; Felleman & Van Essen, 1991; Markov et al., 2012). Yet, we understand much less about the details of the computations that transform information from one visual area to the next.

The ventral visual stream can be described by an approximate hierarchy: V1 neurons project to area V2, V2 neurons in turn project to area V4, V4 neurons project to ITC. In addition to these projections, there are abundant horizontal connections within each area and top-down connections (e.g. V2 projects back to V1). Moving from V1 to ITC, there is a progressive increase in the size of the receptive fields, in the degree of complexity of the type of features represented at each stage and also in the degree of tolerance to image transformations. For example, some neurons in V4 are particularly sensitive to the extent and type of curvature of the visual stimulus (Pasupathy & Connor, 2002). Although the specific circuits that give rise to selectivity along the ventral stream are not clearly understood, it is possible to approximate the type of selective responses to different shapes using the same type of ideas articulated above by non-linearly pooling inputs from simpler responses (Cadieu et al., 2007).

ITC constitutes a vast expanse of the ventral stream. Lesions in ITC typically lead to specific deficits in visual object recognition (e.g. Afraz et al., 2015; Benton & Wav Allen, 1972; Dean, 1976). Neurons in ITC respond to complex shapes including abstract ones like folded paperclips or fractal patterns and real-world objects including faces (Desimone et al., 1984; Logothetis & Sheinberg, 1996; Logothetis, Pauls, & Poggio, 1995; Naya, Sakai, & Miyashita, 1996; Tanaka, 1996; Tsao, Freiwald, Tootell, & Livingstone, 2006). Their responses also show a significant degree of robustness to a large number of stimulus transformations including changes in preferred stimulus position within the receptive field, scale, illumination, rotation, and many others (Ito et al., 1995; Logothetis et al., 1995; Sary et al., 1993; Tovee, Rolls, & Azzopardi, 1994). It is possible to linearly decode in single trials the identity of objects from the activity of small ensembles of ITC neurons (Hung et al., 2005).

Consider an ITC neuron that responds selectively to a complex shape, say a face. To be clear, what we mean in practice is that the investigator records the activity of this neuron and counts the number of spikes in a given time window in response to presentation of a battery of multiple stimuli, say on the order of several hundred different shapes. The stimuli that contain faces elicit not only a response above baseline, but also a response that is statistically stronger than the responses to the other stimuli in the set. If we were able to trace the input connections of this ITC neuron, and one day in the not too distant future we may be able to do this, we would likely observe that the inputs are distributed over multiple V4 cells. As a null hypothesis, we imagine the transformation of V4-like responses into ITC-like

responses to obey the same principles that govern the transformation of multiple LGN inputs converging to give rise to orientation selectivity in V1. In other words, at each stage of processing, there is a transformation from a distributed representation to a grandmother-like representation. Each stage of processing thus connotes a distributed representation and also a localist one, depending on what particular aspects of the stimulus we are considering. Many of the discussions about localist versus distributed representations in the brain have centred on high-level "meaningful things" such as grandmothers, words, or the concept of love or fear. Unfortunately, for most of these high-level things, we still do not really understand the degree of selectivity and invariance of the relevant neurons, let alone the inputs that can give rise to such representations. In the context of visual information, the simplest null model is that we can build a representation of progressively more complex shapes by adequate pooling and filtering signals from the previous processing stage. This leads to an elegant null model for cortical representations, with grandmothers galore as long as we define what is represented at each stage and also with clearly distributed representations.

Computational models of visual recognition

The brief qualitative description of the transformation of visual inputs in the previous section has been formalised in the form of a theoretical framework and quantitatively instantiated into algorithms for visual recognition (DiCarlo et al., 2012; Fukushima, 1980; Mel, 1997; Olshausen, Anderson, & Van Essen, 1993; Perrett, Oram, Hietanen, & Benson, 1994; Riesenhuber & Poggio, 1999; Rolls, 1991; Serre et al., 2007). These biologically inspired algorithms are characterised by a bottom-up hierarchy of linear and non-linear computations that start at the pixel level and progressively build a complex representation of the visual inputs. Ascending through the visual hierarchy, units display larger receptive fields, selectivity to more complex visual features and increased tolerance to changes in those features. The specific implementation details vary across different models but the general principles are the ones articulated in the previous section: units pooling information from previous layers, weighted linear sums followed by thresholding, non-linearities that give rise to tolerance. As a reasonable first approximation, these models assume that "cortex is cortex" and hence that the same mathematical operations are performed at each level (except that each level operates on different inputs and therefore conveys different outputs).

There have been multiple implementations of this type of network architecture (e.g. Miconi, Groomes, & Kreiman, 2015; Riesenhuber & Poggio, 1999; Serre et al., 2007; Wyatte, Curran, & O'Reilly, 2012; Yamins et al., 2014, among many others). These architectures also constitute the basic foundation of deep convolutional networks, which have been successful in many engineering applications, most notably in computer vision (e.g. Krizhevsky, Sutskever, & Hinton, 2012).

This family of models constitutes only a first-order approximation to the complexities of the cortical circuitry; for example, most of these models involve exclusively feed-forward projections and it is well known that there are abundant horizontal and feedback connections throughout neocortex (Douglas & Martin, 2004). Despite this and many other simplifications, these models have been shown to capture basic properties of neuronal responses throughout the ventral visual stream (Riesenhuber & Poggio, 1999; Serre et al., 2007).

An attractive feature of computational models is that we can inspect the activity of every unit and we can explain with perfect accuracy all the inputs and outputs to any possible stimulus. The question of localist versus distributed representations has been discussed before in the context of certain computational models (e.g. Bowers, 2009; McClelland & Rumelhart, 1981; McClelland, et al., 1986) but those computational models have not been directly linked to neurophysiological measurements. Is information in deep hierarchical architectures for visual recognition represented in a localist manner or a distributed one? Each unit in these models is selective and represents a meaningful thing. In the typical computational implementation of these models, every unit shows selectivity. The biological implementation could differ substantially: there are a large number of different types of neurons in every patch of cortex and we still lack a fundamental understanding of the functional properties of these neuronal types and whether all of them show the same type and degree of selectivity. Each unit is part of a small group of units (compared to the total number of units in each area) that codes for that meaningful thing, we can precisely pinpoint what that meaningful thing is for each unit, injecting current into those units leads to an enhanced representation of the features they encode and removing those units from the model leads to a visual scotoma for those features. In other words, every unit in the model is a grandmother unit, as long as we define grandmother in terms of the unit's specific preferences. At the same time, each unit is part of a large ensemble that projects onto the next stage to build a more complex feature. There is no real dichotomy between localist and distributed representations in

these models. All the computations are clearly defined in terms of a few well-defined rules, a simple, yet beautiful null model for cortical representations.

Discussion

It does *not* make sense to discuss distributed versus localist representations in the brain without defining clearly the identity of the neurons in question and the specific type of information that is represented. Accepting the family of models succinctly described above as a plausible first-order description for the functions of ventral visual cortex, one could argue that we have only described the representation of visual features that range from luminance changes to oriented bars to more complex shapes but we have not specifically described how to represent chairs, dogs, grandmothers and other objects. However, these models provide an explicit description of how to represent those shapes. Indeed, those models are routinely tested in object recognition tasks that involve dozens to millions such objects (e.g. Krizhevsky et al., 2012; Serre et al., 2007). In a typical scenario, a set of training images is presented to the models to simulate the activity of every unit in response to every image. Next, a machine learning approach is used to train the weights of a classifier to map the activity of units in one or more layers of the model to the image labels (e.g. label "grandmother" for the corresponding set of pictures that include grandma). Finally, the model is tested with a *different* set of images to evaluate how well it can label those new images. In many cases, a linear kernel is used for the classifier. In this case, a given linear classifier unit takes the distributed input from several units in an earlier layer, multiplies the activity of those units by suitable weights and thresholds the results, finally deciding in a binary way whether the image should be labelled "grandmother" or not. These classifier units are localist par excellence; the inputs to those units contain a distributed representation of visual information. This process is repeated throughout the hierarchical model.

While in some instantiations of these models there may be a single classifier unit for each label, in the brain there is a significant degree of redundancy (Barlow, 1972). The notion of a localist representation does not depend on having a one-to-one map between units and meaningful things. This is a typical misconception of the definition of grandmother cells that has been cleared in multiple earlier discussions. In the original parable by Lettvin, there were 18,000 such cells that responded uniquely to "mother" (Gross, 2002). At no stage of processing do we have only one neuron representing information. There are multiple RGCs that represent luminance changes in a given point in space, multiple V1 neurons that respond to a certain orientation in a given location, and multiple ITC neurons that represent complex visual features at a given location.

There are many important nuances to the representation of information in the brain that deserve further discussion. Once we allow for the notion established in the previous paragraph that the map between neurons and meaningful things is not one-to-one, we may ask how many neurons are involved in representing each feature. Neurophysiological recordings typically show that representations tend to be sparse. Sparseness can be loosely defined as using only a small number of neurons out of the entire set of possible neurons to represent a particular feature (for a more formal definition, see equation 4 in Olshausen and Field (1996)). Sparseness can also be loosely defined by the related but distinct notion that a neuron responds to only a small fraction of all possible stimuli. In computational models, introducing sparseness as a constraint can alter the resulting representation and lead to the development of biologically plausible feature tuning properties. For example, there has been considerable discussion about the idea that information is sparsely represented in primary visual cortex (e.g. Olshausen & Field, 1996). Sparseness is an attractive property because it leads to a more efficient representation that is easier to decode in later stages by virtue of the independence of the inputs, and it may also have energetic efficiency implications (Laughlin, van Steveninck, & Anderson, 1998). Some investigators draw a sharp distinction between localist representations, sparse distributed representations and dense distributed representations (e.g. Bowers, 2009). In the null model outlined here, the distinction is merely a quantitative one. Each visual feature at each location is represented by a relatively small number of units (but more than one unit).

Two essential causal manipulations provide evidence that we are in the right track towards understanding the representation of visual information: disturbing the representation through small lesions or by injecting currents into local circuits. Lesions in V1 lead to localised scotomas in the visual field and lesions in higher visual areas lead to specific visual deficits such as the inability to detect colour, motion or complex shapes (e.g. Dean, 1976; Zeki, 1990, 1991). Eliminating a single neuron is unlikely to lead to clear behavioural manifestations, presumably due to the robustness and redundancy in the representation. We currently do not have the technical ability to selectively lesion or inactivate all the neurons that respond to a specific feature. Yet, a recent study

demonstrated that optogenetic silencing of a relatively small group of neurons in a patch of cortex containing neurons selective for face gender can lead to a small but significant impairment in gender discrimination (Afraz et al., 2015). Advocates of distributed representations would be right to point out that a large number of neurons is removed in these lesion studies; future technological developments will allow a tighter titration of the behavioural consequences of eliminating specific types and different numbers of neurons from cortical circuits. The specificity of the ensuing effects in lesion studies supports the notion that there is a localised, interpretable representation of visual information in cortical circuits. This specificity in the consequences of local lesions is also consistent with the type of effects one would expect from silencing local groups of units in the computational models outlined in the previous section.

Another important causal manipulation that can shed light on the nature of the neural representations involves examining the effects of electrical stimulation. Injecting current through high-impedance microelectrodes placed extracellularly probably alters the activity of hundreds to several thousand neurons in the vicinity of the electrode, depending on the type of electrode, the current intensity, current injection pattern, brain area and tissue excitability (Logothetis et al., 2010; Tehovnik, 1996), in addition to indirect effects on other brain areas. Several studies have demonstrated that perceptual and behavioural effects can be elicited via current injection in this fashion. Stimulating V1 leads to perception of localised phosphenes (e.g. Brindley, Donaldson, Falconer, & Rushton, 1972; Dobelle & Mladejovsky, 1974; Tehovnik, Slocum, Smirnakis, & Tolias, 2009). Similarly, stimulating other parts of visual cortex can lead to specific visually triggered behaviours including motion detection enhancement upon stimulating area MT (Salzman et al., 1990) or face detection enhancement upon stimulating ITC (Afraz et al., 2006). Recently, a series of elegant studies have advanced the possibility of stimulating individual neurons, albeit not in the visual system (e.g. Brecht, Schneider, Sakmann, & Margrie, 2004, reviewed in Doron and Brecht (2015)). Rather surprisingly, these studies have demonstrated that stimulating single neurons can elicit both sensory percepts as well as motor outputs. For example, adding approximately 15 action potentials to baseline activity in the rat barrel cortex, which is involved in the representation of whisker tactile stimulation, led to significant and specific behavioural reports. The results of this type of experiment are typically weak and show a significant degree of variability. Furthermore, the mechanisms of cortical propagation in these so-called nanostimulation experiments are poorly understood; it has been suggested that the behavioural manifestations arise as a consequence of amplification via recurrent connections. Despite these considerations, it is difficult to argue that the activity of individual neurons is uninterpretable given the results of these micro and nanostimulation studies. Rather, the stimulation studies lend further support to the notion of local circuits with interpretable responses that can be activated via stimulation to elicit specific perceptual sensations or motor outputs. The results of these stimulation experiments are also consistent with the consequences of activating local groups of units in the type of computational models outlined in the previous section.

Information from the top layers of the ventral visual stream is conveyed to multiple brain areas, most notably the medial temporal lobe and frontal cortex. Neurophysiological recordings in the human medial temporal lobe have described a plethora of interesting selective responses to places (Ekstrom et al., 2003), object categories (Kreiman, Koch, & Fried, 2000b; Mormann et al., 2011) and specific individuals or landmarks (Quian Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). These responses could be interpreted as additional examples of a localist representation for high-level information. However, as fascinating as those signals are, the neuronal responses in the medial temporal lobe structures do not fulfil all the requirements outlined above required for a representation of visual information, most notably the fact that removing those neurons does not lead to visual impairments (Squire, Stark, & Clark, 2004). Rather, the medial temporal lobe structures play a fundamental role in interpreting the degree of familiarity and novelty of current stimuli, associating multi-modal information in the context of emotional valence and prior knowledge, to form and retrieve memories (Brown & Aggleton, 2001; Cameron, Yashar, Wilson, & Fried, 2001; Eichenbaum, 2004; Gelbard-Sagiv, Mukamel, Harel, Malach, & Fried, 2008; Kreiman, 2007; Kreiman, Koch, & Fried, 2000a; Mormann et al., 2014; Rutishauser, Mamelak, & Schuman, 2006; Rutishauser, Schuman, & Mamelak, 2008; Squire et al., 2004).

Information from the ventral visual cortex is not only conveyed to the medial temporal lobe but also to frontal cortex structures that play a critical role in interpreting the visual inputs in a task-dependent manner to make decisions leading to behavioural outputs (Miller & Cohen, 2001). Neurons in pre-frontal cortex show dynamic tuning to task-dependent variables. Recently, it has been argued that each neuron in pre-frontal cortex is tuned to multiple different aspects of an object sequence memory task rather than a single

one (Rigotti et al., 2013). This so-called mixed-selectivity does *not* imply that the responses of individual units cannot be interpreted. This question goes back to the definition of "meaningful things". If a neuron shows an increased firing rate during the presentation of a specific stimulus and also before a given motor response, it may be challenging to label the neuron as "purely visual" or "purely motor". However, the inadequacy of these labels pre-defined by the investigators does not preclude from decoding the properties of neuronal responses. The neuron in question still shows a reproducible response that can be correlated to the sensory events, decisions and motor outputs during the task. The interpretational challenges arise only because of our attempt to force specific anthropomorphic descriptions based on language to the changes in firing rate. These language-based descriptions of neuronal preferences have been relatively successful in early sensory areas but it seems likely, and far more powerful, that we will require mathematically defined operations to explain the responses in higher aspects of cognition. Even within the complex and fascinating realm of high-level cognitive information represented in frontal cortex, there is still hope for interpreting the activity of individual neurons.

The apparent "mixed" selectivity alluded to in the previous paragraph is not the only conceptual barrier to interpreting the activity of cortical neurons. Another important problem arises when we consider the response dynamics and the dependence on behavioural tasks. Neuronal responses throughout the visual system show complex temporal dynamics that evolve on scales of tens to hundreds of milliseconds (e.g. Hung et al., 2005; Meyers et al., 2008; Richmond, Optican, & Spitzer, 1990; Ringach, Hawken, & Shapley, 1997; Smith, Majaj, & Movshon, 2005; Woloszyn & Sheinberg, 2009). Additionally, the responses throughout visual cortex can be significantly modulated by task demands (e.g. Bansal et al., 2014; Baylis & Rolls, 1987; Gilbert, Li, & Piech, 2009; Ramalingam, McManus, Li, & Gilbert, 2013; Rigotti et al., 2013; Vogels, Sary, & Orban, 1995; Woloszyn & Sheinberg, 2009). The family of models described in the previous section does not really account for neural dynamics (other than the serial passing of signals from one layer to the next) nor does it clearly describe how the same neural representation can be used for multiple different tasks. Furthermore, if the interpretation of spike trains from a given neuron evolves over time due to internal dynamics and/or due to task demands, it is unclear how post-synaptic neurons can decode such time-varying inputs. Changes over long temporal scales (e.g. due to learning) may be easier to decipher since they may be accompanied by concomitant modifications in the synaptic weights, yet we need to be able to also understand changes in the meaning of spike trains that occur within tens to hundreds of milliseconds. In the extreme version, the same firing rate of a given neuron could convey distinct types of information at two different time points. Understanding these dynamics and task-dependent effects opens the doors to interesting questions both for localist and distributed representations and will require significant extensions to the simplified family of models described here.

A prominent example of dynamics in neural network models is illustrated in attractor networks. Attractor networks constitute particularly interesting computational models that are rather distinct form the family of bottom-up architectures described in the previous section (e.g. Hopfield, 1982). Briefly, units are interconnected in an all-to-all fashion and the state of the network is defined by the activity of all units. Each unit participates in representing every possible input and every input is represented by all units. In some sense, each unit in this type of attractor network is *not* representing any meaningful thing in and of itself. This seems to be a major departure from the hierarchical feed-forward models described here as a null model. Even though there has not been clear compelling biological evidence for or against the possibility of an attractor network type of neocortical circuits, given the dense interconnectivity of cortical circuits, it is not inconceivable that these ideas may help us better understand the dynamics of computations in cortex. Still, these networks typically do not explicitly specify how and when information is read out from these models. Brains need a way of converting the representation of visual information into decisions and actions (e.g. to indicate that a particular image does or does not contain a picture of grandma). We need an additional step to peek inside the attractor network, examine its state and determine whether that state corresponds to one of the stored patterns. The simplest way to think about how this read-out mechanism could be implemented with biological circuits is to wire all of the units in the network to another set of units that are capable of discriminating different states. These read-out units would behave as localist units deciphering the distributed representation in the attractor network.

Another important issue that should be discussed is the apparent unreliability of neuronal responses. Upon repeated presentation of an identical stimulus, neurons seem to respond capriciously, and the variance in the spike counts can be as high as the mean spike count. The high degree of trial-to-trial variability in the spike trains elicited in response to a given stimulus has puzzled neuroscientists trying to understand how neurons encode information for decades (Perkel,

Gerstein, & Moore, 1967). This degree of variability seems to pose a challenge to a localist representation where the activity of each neuron in each trial matters and should be interpretable. One possibility is that the activity of each neuron can be described as a localist representation only on average, but this seems like an artificial and convoluted argument; the brain must function in individual trials. A more attractive possibility is that the apparently high noise levels are a consequence of heterogeneity in the experimental conditions. Even when investigators think that the conditions are identical from one trial to the next, there exist multiple internal and small external variables that can be different. In support of this argument, cortical responses can show smaller degrees of trial-to-trial variability when there is better control of experimental conditions (e.g. Bair & Koch, 1996; Churchland et al., 2010). Furthermore, several studies examining responses in primary or early sensory neurons where there may be less fluctuation in internal conditions have demonstrated that variability can be very low, in some cases, as low as the theoretical limits imposed by the discrete nature of spikes (e.g. Berry & Meister, 1998; Kreiman, Krahe, Metzner, Koch, & Gabbiani, 2000; van Steveninck, Lewen, Strong, Koberle, & Bialek, 1997). Additionally, variability can also be significantly reduced in slice recordings (Holt, Softky, Koch, & Douglas, 1996; Mainen & Sejnowski, 1995; Stevens & Zador, 1998). In sum, under conditions where it is possible to exert better control over the experimental conditions, trial-to-trial variability is significantly reduced and the responses of individual neurons become more interpretable in single trials.

Even if we accept the rudimentary sketch proposed as a null model for the cortical representation of visual information, we still have very little to say about a myriad of high-level and abstract ideas such as how the brain encodes "love", or "I would like to talk to grandma today", or "I miss grandma's delicious food". In a completely speculative and admittedly naïve tone, we can imagine that evolution works by duplicating, copying, improving and refining existing circuits. Hence, the basic computational principles uncovered when scrutinising the representation of visual information could well extrapolate to other aspects of cognition. Needless to say, there is a vast and fascinating uncharted territory awaiting for neuroscientists to probe the representation of high-level cognitive information.

Disclosure statement

No potential conflict of interest was reported by the author.

Funding

This work was supported by NSF (1358839, CCF-1231216) and NIH (R01EY026025).

References

Afraz, A., Boyden, E. S., & DiCarlo, J. J. (2015). Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. *Proceedings of the National Academy of Sciences*, *112*(21), 6730–6735. doi:10.1073/pnas.1423328112

Afraz, S. R., Kiani, R., & Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. *Nature*, *442*(7103), 692–695. doi:10.1038/nature04982

Bair, W., & Koch, C. (1996). Temporal precision of spike trains in extrastriate cortex of the behaving macaque monkey. *Neural Computation*, *8*, 1185–1202. doi:10.1152/jn.90980.2008

Baker, J. F., Petersen, S. E., Newsome, W. T., & Allman, J. M. (1981). Visual response properties of neurons in four extrastriate visual areas of the owl monkey (*Aotus trivirgatus*): A quantitative comparison of medial, dorsomedial, dorsolateral, and middle temporal areas. *J Neurophysiol*, *45*(3), 397–416.

Bansal, A. K., Madhavan, R., Agam, Y., Golby, A., Madsen, J. R., & Kreiman, G. (2014). Neural dynamics underlying target detection in the human brain. *Journal of Neuroscience*, *34*(8), 3042–3055. doi:10.1523/JNEUROSCI.3781-13.2014

Bansal, A. K., Singer, J. M., Anderson, W. S., Golby, A., Madsen, J. R., & Kreiman, G. (2012). Temporal stability of visually selective responses in intracranial field potentials recorded from human occipital and temporal lobes. *Journal of Neurophysiology*, *108*, 3073–3086. doi:10.1152/jn.00458.2012

Barlow, H. B. (1972). Single units and sensation: A neuron doctrine for perception. *Perception*, *1*, 371–394. doi:10.1068/p010371

Baylis, G. C., & Rolls, E. T. (1987). Responses of neurons in the inferior temporal cortex in short-term and serial recognition memory tasks. *Experimental Brain Research*, *65*(3), 614–622. doi:10.1007/bf00235984

Benton, A. L., & Wav Allen, M. W. (1972). Prosopagnosia and facial discrimination. *Journal of the Neurological Sciences*, *15*, 167. doi:10.1016/0022-510x(72)90004-4

Berry, M. J., & Meister, M. (1998). Refractoriness and neural precision. *Journal of Neuroscience*, *18*(6), 2200–2211.

Bondar, I. V., Leopold, D. A., Richmond, B. J., Victor, J. D., & Logothesis, N. K. (2009). Long-term stability of visual pattern selective responses of monkey temporal lobe neurons. *PLoS One*, *4*(12), doi:10.1371/journal.pone.0008222

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review*, *116*(1), 220–251. doi:2009-00258-008 [pii] 10.1037/a0014462

Brecht, M., Schneider, M., Sakmann, B., & Margrie, T. W. (2004). Whisker movements evoked by stimulation of single pyramidal cells in rat motor cortex. *Nature*, *427*, 704–710. doi:10.1038/nature02266

Brindley, G. S., Donaldson, P. E., Falconer, M. A., Rushton, D. N. (1972). The extent of the region of occiptal cortex that when stimulated gives phosphenes fixed in the visual field. *Journal of Physiology*, *225*(2), 57–58.

Brown, M. W., & Aggleton, J. P. (2001). Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience*, *2*(1), 51–61. doi:10.1038/35049064

Cadieu, C., Kouh, M., Pasupathy, A., Connor, C., Riesenhuber, M., & Poggio, T. (2007). A model of V4 shape selectivity and invariance. *Journal of Neurophysiology*, *98*(3), 1733–1750. doi:10.1152/jn.01265.2006

Cameron, K. A., Yashar, S., Wilson, C. L., & Fried, I. (2001). Human hippocampal neurons predict how well word pairs will be remembered. *Neuron*, *30*, 289–298. doi:10.1016/S0896-6273(01)00280-X

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., … Rust, N. C. (2005). Do we know what the early visual system does? *Journal of Neuroscience*, *25*(46), 10577–10597. doi:10.1523/JNEUROSCI.3726-05.2005

Churchland, M. M., Yu, B. M., Cunningham, J. P., Sugrue, L. P., Cohen, M. R., Corrado, G. S., … Shenoy, K. V. (2010). Stimulus onset quenches neural variability: A widespread cortical phenomenon. *Nature Neuroscience*, *13*(3), 369–378. doi:nn.2501 [pii] 10.1038/nn.2501

Connor, C. E., Brincat, S. L., & Pasupathy, A. (2007). Transformation of shape information in the ventral pathway. *Current Opinion in Neurobiology*, *17*(2), 140–147. doi:10.1016/j.conb.2007.03.002

Crick, F., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience*, *6*(2), 119–126. doi:10.1038/nn0203-119

Dayan, P., & Abbott, L. (2001). *Theoretical neuroscience*. Cambridge, MA: MIT Press.

Dean, P. (1976). Effects of inferotemporal lesions on the behavior of monkeys. *Psychological Bulletin*, *83*(1), 41–71. doi:10.1037/0033-2909.83.1.41

Deco, G., & Rolls, E. T. (2004). A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, *44*(6), 621–642. doi:10.1016/j.visres.2003.09.037

Desimone, R., Albright, T. D., Gross, C. G., & Bruce, C. (1984). Stimulus-selective properties of inferior temporal neurons in the macaque. *Journal of Neuroscience*, *4*(8), 2051–2062.

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. doi:S0896-6273(12)00092-X [pii] 10.1016/j.neuron.2012.01.010

Dobelle, W. H., & Mladejovsky, M. G. (1974). Phosphenes produced by electrical stimulation of human occipital cortex, and their application to the developement of a prosthesis for the blind. *The Journal of Physiology*, *243*(2), 553. doi:10.1113/jphysiol.1974.sp010766

Doron, G., & Brecht, M. (2015). What single-cell stimulation has told us about neural coding. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1677). doi:10.1098/rstb.2014.0204

Douglas, R. J., & Martin, K. A. (2004). Neuronal circuits of the neocortex. *Annual Review of Neuroscience*, *27*, 419–451. doi:10.1146/annurev.neuro.27.070203.144152

Eichenbaum, H. (2004). Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron*, *44*(1), 109–120. doi:10.1016/j.neuron.2004.08.028

Ekstrom, A. D., Kahana, M. J., Caplan, J. B., Fields, T. A., Isham, E. A., Newman, E. L., & Fried, I. (2003). Cellular networks underlying human spatial navigation. *Nature*, *425*(6954), 184–187. doi:10.1038/nature01964

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*, 1–47. doi:10.1093/cercor/1.1.1

Field, G. D., Gauthier, J. L., Sher, A., Greschner, M., Machado, T. A., Jepson, L. H., … Chichilnisky, E. J. (2010). Functional connectivity in the retina at the resolution of photoreceptors. *Nature*, *467*(7316), 673–677. doi:10.1038/nature09424

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2001). Categorical representation of visual stimuli in the primate prefrontal cortex. *Science*, *291*, 312–316. doi:10.1126/science.291.5502.312

Fukushima, K. (1980). Neocognitron: A self organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202. doi:10.1007/978-3-642-46466-9_18

Gelbard-Sagiv, H., Mukamel, R., Harel, M., Malach, R., & Fried, I. (2008). Internally generated reactivation of single neurons in human Hippocampus during free recall. *Science*, doi:10.1126/science.1164685

Gilbert, C. D., Li, W., & Piech, V. (2009). Perceptual learning and adult cortical plasticity. *The Journal of Physiology*, *587*(Pt 12), 2743–2751. doi:10.1113/jphysiol.2009.171488

Gross, C. G. (2002). Genealogy of the "grandmother cell". *The Neuroscientist*, *8*(5), 512–518. doi:10.1177/107385802237175

Holmes, G. (1918). Disturbances of visual orientation. *British Journal Ophthalmology*, *2*(10), 506–516.

Holt, G. R., Softky, W. R., Koch, C., & Douglas, R. J. (1996). Comparison of discharge variability in vitro and invivo in cat visual cortex neurons. *Journal of Neurophysiology*, *75*(5), 1806–1814.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 2554–2558. doi:10.1073/pnas.79.8.2554

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of Physiology*, *160*, 106–154. doi:10.1113/jphysiol.1962.sp006837

Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast read-out of object identity from Macaque inferior temporal cortex. *Science*, *310*, 863–866. doi:10.1126/science.1117593

Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, *73*(1), 218–226.

Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision Research*, *46*(11), 1762–1776. doi:S0042-6989(05)00511-0 [pii] 10.1016/j.visres.2005.10.002

Kreiman, G. (2007). Single neuron approaches to human vision and memories. *Current Opinion in Neurobiology*, *17*(4), 471–475. doi:10.1016/j.conb.2007.07.005

Kreiman, G., Koch, C., & Fried, I. (2000a). Imagery neurons in the human brain. *Nature*, *408*, 357–361. doi:10.1038/35042575

Kreiman, G., Koch, C., & Fried, I. (2000b). Category-specific visual responses of single neurons in the human medial temporal lobe. *Nature Neuroscience*, *3*(9), 946–953. doi:10.1038/78868

Kreiman, G., Krahe, R., Metzner, W., Koch, C., & Gabbiani, F. (2000). Robustness and variability of neuronal coding by amplitude sensitive afferents in the weakly electric fish *Eigenmannia*. *Journal of Neurophysiology*, *84*(1), 189–204.

Krizhevsky, A., Sutskever, I., & Hinton, G. (2012). *ImageNet Classification with Deep Convolutional Neural Networks*. Paper presented at the NIPS, Montreal.

Laughlin, S. B., van Steveninck, R. R. R., & Anderson, J. C. (1998). The metabolic cost of neural information. *Nature Neuroscience*, *1*(1), 36–41. doi:10.1038/236

Leopold, D. A., & Logothetis, N. K. (1996). Activity changes in early visual cortex reflect monkeys' percepts during binocular rivalry. *Nature*, *379*, 549–553. doi:10.1038/379549a0

Logothetis, N. K., Augath, M., Murayama, Y., Rauch, A., Sultan, F., Goense, J., … Merkle, H. (2010). The effects of electrical microstimulation on cortical signal propagation. *Nature Neuroscience*, *13*(10), 1283–1291. doi:10.1038/nn.2631

Logothetis, N. K., Pauls, J., & Poggio, T. (1995). Shape representation in the inferior temporal cortex of monkeys. *Current Biology*, *5*(5), 552–563. doi:10.1016/s0960-9822(95)00108-4

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual object recognition. *Annual Review of Neuroscience*, *19*, 577–621. doi:10.1146/annurev.ne.19.030196.003045

Mainen, Z. F., & Sejnowski, T. J. (1995). Reliability of spike timing in neocortical neurons. *Science*, *268*, 1503–1506. doi:10.1126/science.7770778

Markov, N. T., Ercsey-Ravasz, M. M., RibeiroGomes, A. R., Lamy, C., Magrou, L., Vezoli, J., … Kennedy, H. (2012). A weighted and directed interareal connectivity matrix for Macaque Cerebral Cortex. *Cerebral Cortex*, doi:bhs270 [pii] 10.1093/cercor/bhs270

McClelland, J. L, & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*, 375–407. doi:10.1037/0033-295x.88.5.375

McClelland, J. L., Rumelhart, D. E., & PDP Research Group. (1986). *Parallel distributed processing: Psychological and biological models (Vol. 2)*. Cambridge, MA: MIT Press.

McMahon, D. B., Jones, A. P., Bondar, I. V., & Leopold, D. A. (2014). Face-selective neurons maintain consistent visual responses across months. *Proceedings of the National Academy of Sciences*, *111*(22), 8251–8256. doi:10.1073/pnas.1318331111.

Mel, B. (1997). SEEMORE: Combining color, shape and texture histogramming in a neurally inspired approach to visual object recognition. *Neural Computation*, *9*, 777. doi:10.1162/neco.1997.9.4.777.

Meyers, E., Freedman, D. J., Kreiman, G., Miller, E. K., & Poggio, T. (2008). Dynamic population coding of category information in ITC and PFC. *Journal of Neurophysiology*, *100*, 1407–1419. doi:10.1152/jn.90248.2008

Miconi, T., Groomes, L., & Kreiman, G. (2015). There's Waldo! A normalization model of visual search predicts single-trial human fixations in an object search task. *Cerebral Cortex*, doi:10.1093/cercor/bhv129

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience*, *24*, 167–202. doi:10.1146/annurev.neuro.24.1.167

Mormann, F., Dubois, J., Kornblith, S., Milosavljevic, M., Cerf, M., Ison, M., … Koch, C. (2011). A category-specific response to animals in the right human amygdala. *Nature Neuroscience*, *14*(10), 1247–1249. doi:10.1038/nn.2899

Mormann, F., Ison, M., Quiroga, R. Q., Koch, C., Fried, I., & Kreiman, G. (2014). Visual cognitive adventures of single neurons in the human medial temporal lobe. In I. Fried, U. Rutishauser, M. Cerf, & G. Kreiman (Eds.), *Single neuron studies of the human brain. Probing cognition* (121–151). Cambridge, MA: MIT Press.

Naya, Y., Sakai, K., & Miyashita, Y. (1996). Activity of primate inferotemporal neurons related to a sought target in a paired-association task. *Proceedings of the National Academy of Sciences*, *93*, 2664–2669. doi:10.1073/pnas.93.7.2664

Newsome, W. T., Britten, K. H., & Movshon, J. A. (1989). Neuronal correlates of a perceptual decision. *Nature*, *341*, 52–54. doi:10.1038/341052a0

Olshausen, B. A., Anderson, C. H., & Van Essen, D. C. (1993). A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information. *The Journal of Neuroscience*, *13*(11), 4700–4719.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609. doi:10.1038/381607a0

Parker, A. J., & Newsome, W. T. (1998). Sense and the single neuron: Probing the physiology of perception. *Annual Review of Neuroscience*, *21*, 227–277. doi:10.1146/annurev.neuro.21.1.227

Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area V4. *Nature Neuroscience*, *5*(12), 1332–1338. doi:10.1038/nn972

Perkel, D. H., Gerstein, G. L., & Moore, G. P. (1967). Neuronal spike trains and stochastic point processes. *Biophysical Journal*, *7*, 391–418. doi:10.1016/S0006-3495(67)86596-2

Perrett, D. I., Oram, M. W., Hietanen, J. K., & Benson, P. J. (1994). Issues of representation in object vision. In M. J. Farah, & G. Ratcliff (Eds.), *The neuropsychology of high-level vision* (pp. 33–61). Hillsdale, NJ: Lawrence Erlbaum Associates.

Potter, M. C., & Levy, E. (1969). Recognition memory for a rapid sequence of pictures. *Journal of Experimental Psychology*, *81*(1), 10–15. doi:10.1037/h0027470

Priebe, N. J., & Ferster, D. (2012). Mechanisms of neuronal computation in mammalian visual cortex. *Neuron*, *75*(2), 194–208. doi:10.1016/j.neuron.2012.06.011

Quian Quiroga, R., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*, 1102–1107. doi:10.1038/nature03687

Ramalingam, N., McManus, J. N., Li, W., & Gilbert, C. D. (2013). Top-down modulation of lateral interactions in visual cortex. *Journal of Neuroscience*, *33*(5), 1773–1789. doi:10.1523/JNEUROSCI.3825-12.2013

Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, *27*, 611–647. doi:10.1146/annurev.neuro.26.041002.131039

Richmond, B. J., Optican, L. M., & Spitzer, H. (1990). Temporal encoding of two-dimensional patterns by single units in primate primary visual cortex. I. Stimulus-response relations. *Journal of Neurophysiology*, *64*(2), 351–369.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025. doi:10.1038/14819

Rigotti, M., Barak, O., Warden, M. R., Wang, X. J., Daw, N. D., Miller, E. K., & Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. *Nature*, *497*(7451), 585–590. doi:10.1038/nature12160

Ringach, D. L., Hawken, M. J., & Shapley, R. (1997). Dynamics of orientation tuning in macaque primary visual cortex. *Nature*, *387*(6630), 281–284. doi:10.1038/387281a0

Rolls, E. T. (1991). Neural organization of higher visual functions. *Current Opinion in Neurobiology*, *1*, 274–278.

Rutishauser, U., Mamelak, A. N., & Schuman, E. M. (2006). Single-trial learning of novel stimuli by individual neurons of the human hippocampus-amygdala complex. *Neuron*, *49*(6), 805–813. doi:10.1016/j.neuron.2006.02.015

Rutishauser, U., Schuman, E. M., & Mamelak, A. N. (2008). Activity of human hippocampal and amygdala neurons during retrieval of declarative memories. *Proceedings of the National Academy of Sciences*, *105*(1), 329–334. doi:10.1073/pnas.0706015105

Salzman, C. D., Britten, K. H., & Newsome, W. T. (1990). Cortical microstimulation influences perceptual judgments of motion direction. *Nature*, *346*, 174–177. doi:10.1038/346174a0

Sary, G., Vogels, R., & Orban, G. A. (1993). Cue-invariant shape selectivity of macaque inferior temporal neurons. *Science*, *260*, 995–997. doi:10.1126/science.8493538

Serre, T., Kreiman, G., Kouh, M., Cadieu, C., Knoblich, U., & Poggio, T. (2007). A quantitative theory of immediate visual recognition. *Progress In Brain Research*, *165C*, 33–56. doi:10.1016/S0079-6123(06)65004-8

Sheinberg, D. L., & Logothetis, N. K. (1997). The role of temporal areas in perceptual organization. *Proceedings of the National Academy of Sciences, USA*, *94*, 3408–3413. doi:10.1073/pnas.94.7.3408

Smith, M. A., Majaj, N. J., & Movshon, J. A. (2005). Dynamics of motion signaling by neurons in macaque area MT. *Nature Neuroscience*, *8*(2), 220–228. doi:10.1038/nn1382

Squire, L. R., Stark, C. E., & Clark, R. E. (2004). The medial temporal lobe. *Annual Review of Neuroscience*, *27*, 279–306. doi:10.1146/annurev.neuro.27.070203.144130

van Steveninck, R., Lewen, G. D., Strong, S. P., Koberle, R., & Bialek, W. (1997). Reproducibility and variability in neural spike trains. *Science*, *275*, 1805–1808. doi:10.1126/science.275.5307.1805

Stevens, C. F., & Zador, A. M. (1998). Input synchrony and the irregular firing of cortical neurons. *Nature Neuroscience*, *1*(3), 210–217. doi:10.1038/659

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience*, *19*, 109–139. doi:10.1146/annurev.ne.19.030196.000545

Tehovnik, E. J. (1996). Electrical stimulation of neural tissue to evoke behavioral responses. *Journal of Neuroscience Methods*, *65*(1), 1–17. doi:10.1016/0165-0270(95)00131-x

Tehovnik, E. J., Slocum, W. M., Smirnakis, S. M., & Tolias, A. S. (2009). Microstimulation of visual cortex to restore vision. *Progress in Brain Research*, *175*, 347–375. doi:10.1016/S0079-6123(09)17524-6

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522. doi:10.1038/381520a0

Tolias, A. D., Ecker, A. S., Siapas, A. G., Hoenselaar, A., Keliris, G. A., & Logothesis, N. (2007). Recording chronically from the same neurons in awake, behaving primates. *Journal of Neurophysiology*. doi:10.1152/jn.00260.2007

Tovee, M. J., Rolls, E. T., & Azzopardi, P. (1994). Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. *Journal of Neurophysiology*, *72*(3), 1049–1060.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*(5761), 670–674. doi:10.1126/science.1119983

Vogels, R., Sary, G., & Orban, G. A. (1995). How task-related are the responses of inferior temporal neurons? *Visual Neuroscience*, *12*(2), 207–214. doi:10.1017/s0952523800007884

Woloszyn, L., & Sheinberg, D. L. (2009). Neural dynamics in inferior temporal cortex during a visual working memory task. *Journal of Neuroscience*, *29*(17), 5494–5507. doi:29/17/5494 [pii] 10.1523/JNEUROSCI.5785-08.2009

Wyatte, D., Curran, T., & O'Reilly, R. (2012). The limits of feedforward vision: Recurrent processing promotes robust object recognition when objects are degraded. *Journal of Cognitive Neuroscience*, *24*(11), 2248–2261. doi:10.1162/jocn_a_00282

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. doi:10.1073/pnas.1403112111

Zeki, S. (1990). A century of cerebral achromatopsia. *Brain*, *113*(Pt 6), 1721–1777. doi:10.1093/brain/113.6.1721

Zeki, S. (1991). Cerebral akinetopsia (visual motion blindness). A review. *Brain*, *114*(Pt 2), 811–824. doi:10.1093/brain/114.4.2021