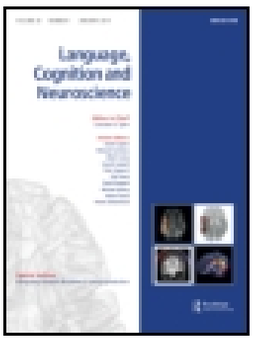

Routledge
Taylor & Francis Group

Language, Cognition and Neuroscience

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

Transient localist representations in critical branching networks

Jeffrey J. Rodny, Timothy M. Shea & Christopher T. Kello

To cite this article: Jeffrey J. Rodny, Timothy M. Shea & Christopher T. Kello (2016): Transient localist representations in critical branching networks, Language, Cognition and Neuroscience, DOI: 10.1080/23273798.2016.1242760

To link to this article: http://dx.doi.org/10.1080/23273798.2016.1242760

LANGUAGE, COGNITION AND NEUROSCIENCE, 2016
http://dx.doi.org/10.1080/23273798.2016.1242760

Routledge
Taylor & Francis Group

Transient localist representations in critical branching networks

Jeffrey J. Rodny, Timothy M. Shea and Christopher T. Kello

Cognitive and Information Sciences, University of California, Merced, CA, USA

ABSTRACT

There is a long-standing debate between localist and distributed representations in neural network research. Different connectionist models have employed one or the other type of representation, but all such models assume that representations are either always stable, or stable after learning converges. The present article explores an alternate possibility, which is that representations continually shift and change, even on relatively fast timescales, and even after learning has stabilised. This possibility of *transient* representation is demonstrated in a spiking neural network model, along with supporting neuroscientific evidence. The model learns localist representations that change over a wide range of timescales. Representations are made transient because synaptic connections are continually enabled and disabled to regulate spike propagation to a homeostasis of *critical branching*. Experiments and models suggest that so-called grandmother cells may shift on relatively fast timescales between different mental representations in different contexts and conditions.

ARTICLE HISTORY
Received 2 March 2016
Accepted 16 September 2016

KEYWORDS
Localist representation; metastability; critical branching; spiking neural network; grandmother cells

Introduction

The relationship between neural and mental activity is central to cognitive science, and it has sparked a number of long-standing debates. One such debate concerns the number of different percepts, concepts, or actions that a given neuron might represent. On the one hand, the activity of a given neuron might contribute to a single representation at some level, that is, *localist representation* (Page, 2000). Such hypothetical neurons are sometimes referred to as *grandmother cells* (Bowers, 2009), based on the illustrative example that the spiking of a given neuron might always contribute to the thought of one's grandmother and nothing else. On the other hand, the activity of a given neuron might contribute to multiple representations, for example, *distributed representation* (Hinton, McClelland, & Rumelhart, 1986). Localist and distributed representations highlight one dimension of the relationship between neural and mental activity, and other dimensions are highlighted by other representations, such as population codes (Erickson, 2001) and sparse codes (Olshausen & Field, 1997).

In the present study, we consider the *stability over time* of relationships between neuronal spikes and mental representations (Durstewitz & Deco, 2008). Stability is somewhat orthogonal to the localist-distributed debate, but we focus on localist representations because the issues are clearer in this context, and because stability has implications for studies of localist representation. The question is simply this – when one asserts that spikes from a given neuron relate to a particular mental representation, over what duration of time does this relationship extend?

The standard position with respect to this question is that the activity of a neuron corresponds with the same mental representation for most or all of its life, that is, a *stable localist representation*. In other words, a neuron will spike in conjunction with the same stimuli or actions if one records from it early in the day, then later in the day, then the next day, week, month, or year. The alternative is that the spikes of a given neuron correspond to different mental representations over time, with correspondences changing possibly on the order of days, hours, minutes, or even seconds, that is, a *transient localist representation*. As with the localist-distributed dimension, there are intermediate possibilities between the extremes of stable and transient representations. In particular, transience must be defined with respect to one or more timescales, and the longer the timescale, the more stable the representation.

In this study, we point out that researchers have tacitly assumed stable representations in the localist-distributed debate, and that most neural network models require stability to support learning and performance. Then we review recent evidence from neuroscience that calls into question the assumption of stability, and

CONTACT Christopher T. Kello ckello@ucmerced.edu

we discuss potential advantages of transient representations for computation and cognitive function. Finally, we review a spiking neural network model that learns transient localist representations while maintaining performance in a simple nonlinear classification task. The model stands as a proof-of-concept that transient representations are viable alternatives to stable representations, and it also leads to different questions and hypotheses about the relationship between neural and mental activity.

Stable representations in neural networks

The idea of stable localist representations can be traced back to early work in neuroscience on sensory and motor systems. It was natural and straightforward to posit that individual neurons contribute to individual percepts and motor commands (i.e. localist representation), and to also assume that the correspondence between neurons and percepts or motor commands remains relatively fixed over a cell's lifetime. For instance, there is evidence that neurons in primary visual cortex have well-defined receptive fields described by a direct correspondence between a neuron's firing rate and the presence of, for example, a luminance contrast of a particular orientation at a particular location in the visual field (Hubel & Wiesel, 1959, 1968). Likewise, there is evidence that the firing of individual motor neurons relates to particular motor activities (Person & Kudina, 1972). In both cases, it is convenient to assume that spikes have stable relationships with sensory and motor activities, because otherwise it is not clear how these spikes could reliably encode information and interact with other neural systems.

The assumption of stable representation was implicit in the earliest empirical studies of neural coding, and also in the earliest models of neural network processing. From Rosenblatt's perceptron (Rosenblatt, 1961) and Selfridge's pandemonium (Selfridge, 1958), to adaptive resonance (Grossberg, 1976) and interactive activation (McClelland & Rumelhart, 1981), it has been nearly universal to formulate units in neural network models with stable representations. In models without learning, each unit is assigned to represent a particular feature or category or concept, such as the letter and word units in an interactive activation model. These representations are stable because they do not change during the "life" of the model. In models with learning, units develop learned localist or distributed representations over time, and these representations stabilise once learning has asymptoted.

For both real and artificial neural networks, the rationale for stable representations is partly grounded in the assumption of relatively stable network connectivity. Learning is typically theorised in terms of changes to the strengths of connections, rather than changes in connectivity (but see constructive learning algorithms as in Tin-Yau & Dit-Yan, 1997). This stable connectivity means that units mostly do not change in terms of their structural relations to other units, which in turn facilitates stable representations (connections can sometimes weaken to the point of being ineffective, but most learning algorithms cannot create new connections). To illustrate, a unit that represents the word "cat" in the interactive activation model does so by virtue of its stable connections to letter units "c", "a", and "t". In cortex, a given neuron in V1 plays a specific role in the hierarchy of visual features by virtue of its connectivity with other layers of the visual system.

Stable versus transient representations

The common assumption of stable representations has proven useful for advancing research on the relationship between neural and mental activity, but the assumption is challenged by evidence for transient representations. To begin with, indirect evidence has been mounting that neural networks can change structurally more often, and on faster timescales, than previously assumed. There are now many known mechanisms of plasticity that modify synaptic strengths, and combined, they can make relatively large changes in short periods of time (Colbran, 2015). Even the connections themselves appear to vary frequently over time due to multiple possible factors. For example, a large proportion of dendritic spines appear to exhibit ongoing variability over time, which means that new spines and synaptic connections are formed, and existing ones are eliminated, on a regular basis (Holtmaat et al., 2005). In addition to structural changes, synaptic transmission can be dynamically modulated by neurotransmitter release at the synapses, causing some synapses to trigger postsynaptic potentials more reliably than others (Branco & Staras, 2009). Also, astrocytes seem to support an additional as yet unknown mechanism of synaptic plasticity, and evidence suggests they can rapidly and dynamically modulate neurotransmitter transmission at synaptic clefts over time (De Pitta et al., 2012).

The extent and degree of plasticity in effective connectivity calls into question the assumption of stable relationships between neuronal spikes and mental representations. As noted earlier, stable connectivity is part of the underlying rationale for assuming stable representations. Indirect evidence about plasticity is informative, but we also need direct tests of stable versus transient representations. Such a test requires a reliable

recording from the same neuron, under consistent stimulus and task conditions, for an extended period of time. Just how much time depends on the degree of stability/transience one wishes to test, but periods of hours or even days are warranted, given that most theories and models assume representations are stable for at least this long. One also needs a way to distinguish transience in recordings due to representational change, versus transience due to noise that may arise from neural processes or from instabilities in recording conditions (e.g. drift in electrode placement).

The challenges of distinguishing stable versus transient representations have limited the number of studies directly testing this distinction, but not surprisingly, there is evidence at both ends of the spectrum. In one illustrative study, researchers found single-cell recording evidence for localist representations in long-term memory (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). Quiroga and colleagues examined the firing rates of neurons in response to various images presented to epileptic patients with depth electrodes placed in their medial temporal lobes, which are brain areas with established relations to long-term memory function. The researchers identified neurons that fired when particular people, buildings, or animals were displayed, across a range of different visual angles and orientations, but did not fire for other stimulus items. They interpreted their results as indicative of a *sparse code* (Quiroga, Kreiman, Koch, & Fried, 2008), which is a coding scheme relatively similar to localist representation with respect to the distributed-localist debate (Quiroga & Kreiman, 2010).

The Quiroga et al. (2005) study is illustrative because the authors assumed stable representations, and any apparent transience was attributed to unwanted changes in recording conditions, such as electrode slippage. Patients were chronically implanted for 7–10 day periods, and data were collected for about 3–4 recording sessions, each about 30 minutes long. Responsive cells were identified in an initial recording session, and then recorded again in later sessions. However, of 993 initially identified cells, only 132 exhibited localist-like responses to the visual stimuli. The article does not contain information about whether cells were tracked across sessions, and if so, how many had consistent response properties across sessions. It is possible that some of the apparently non-localist cells were actually localist but not responsive to the particular stimuli. It is also possible that these cells were localist but transient during the 30 minute sessions. For example, a given cell may have been uniquely responsive to images of Jennifer Aniston during the first 15 minutes, and Halle Berry during the last 15 minutes. In this case, the criteria used by Quiroga et al. would exclude such a cell from analysis as a sparse/localist code.

The assumption of stable representations is common, but it cannot be taken for granted in work on brain–machine interfaces. This point has led to studies that investigate the stability of representations, taking recording conditions into account. The possibility of transient representations is important for brain–machine interfaces because it would complicate the use of machine learning algorithms to learn the mappings between neural activity and machine commands. Stability is simpler because it enables machine learning algorithms to store fixed mappings between neural recordings and machine commands. Some studies have been reassuring in this regard – Ganguly and Carmena (2009) found population codes in macaque primary motor cortex that were fairly stable over multiple week periods, and Chestek et al. (2007) found evidence for stable localist representations over two days in macaque premotor cortex. However, other studies have found evidence for more transient representations – Rokni, Richardson, Bizzi, and Seung (2007) found tuning curves in primary motor cortex to drift over the course of days and even hours (for similar results see János et al., 2013), and similar evidence for transience have been found in studies of sensorimotor maps (Rossini et al., 1994) and olfactory representations (Kato, Chu, Isaacson, & Komiyama, 2012). In sum, currently there is evidence for both stable and transient representations.

Potential benefits of transient representations

Thus far, we have reviewed logical arguments and empirical evidence for the existence of transient representations, but why would brains evolve such a relationship between neural and mental activity? On first pass, it may seem like transience presents difficulties for information processing. If this is true, then transient representations should be weeded out by natural selection, unless they confer some beneficial properties to offset any disadvantages. Here we review three possible benefits of transient representations. Then, as a demonstration, we report a spiking neural network model that overcomes the challenges of instability while learning highly transient, localist representations.

One benefit of transient representations may be to support the context dependency of perceptual and cognitive function (Freeman, 1994). Wittgenstein (1953) pointed out that concepts like "game" are defined not by a necessary and sufficient set of features, but rather, by the interaction of prior knowledge and context. If neuronal spike patterns bear some direct relationship to mental

representations like percepts and concepts, then these spike patterns are predicted to change in a matter of seconds or even faster when the context changes, even when stimuli remain constant. From this perspective, the relationship between neural and mental activity may be transient as a function of context. Evidence for such transience comes from a number of studies, including transient activity in the moth olfactory lobe whose patterns change on short timescales depending on perceptual context (Christensen, Pawlowski, Lei, & Hildebrand, 2000), and analogous findings of hippocampal recordings that change depending on spatial navigation context (Pastalkova, Itskov, Amarasingham, & Buzsáki, 2008).

Another potential benefit of transient representations is their capacity to encode a wider range of percepts, actions, and concepts compared with stable representations. Consider the localist case in particular. A well-known concern about localist encoding is that the size of the representational space is directly and linearly related to the number of neurons, that is, we are limited in the number of representations we can form by the number of neurons in the cortex. This relationship is limiting, but the limitation comes from the underlying assumption of stable localist representation. A localist neuron with transient relationships to mental activity has the potential to encode different percepts or concepts at different times. Realising this potential would increase the total number of mental objects that can be represented, albeit not all representations could be evoked simultaneously. Thus the tradeoff would be that a system of transient localist representations would not be equipped to activate any possible percept or concept at any given time. This tradeoff seems manageable, provided that representations can be dynamically shifted and reallocated as needed.

The ability of a system to dynamically reconfigure relationships between neural and mental activity has been associated with *metastability* (Bressler & Kelso, 2001; Kello, Anderson, Holden, & Van Orden, 2008), which is a computationally adaptive property (and thus potentially beneficial) associated with *critical branching* (Haldeman & Beggs, 2005; Kello, 2013). Critical branching is a principle of homeostatic regulation of spike propagation in a network. Simply put, it states that spikes should be conserved as they propagate, such that there number neither grows nor shrinks over time. This balanced state is theorised to be tenuous, analogous to walking a tight rope when there is positive feedback to either side (i.e. increasing propagation begets further increases, and decreasing propagation begets further decreases). Metastability derives from interactions among neurons that tenuously balance their activities between subcritical and supercritical spike propagation. This balance can be struck in a spiking neural network by tuning the propagation of spikes to their critical branching point, which causes patterns of spike activity to be perpetually in flux as the system teeters between subcritical and supercritical regimes – this perpetual flux in spike patterns is referred to as metastability.

Evidence for critical branching started with a study by Beggs and Plenz (2003), in which the authors recorded intrinsic bursts of spike activity in slice preparations of rat somatosensory cortex. They found spike dynamics to exhibit bursts of activity whose sizes were distributed as an inverse power law, $P \sim 1/S^{\alpha}$ (similar to exponential but with a heavy tail). The exponent parameter was estimated from the data to be consistently near 3/2, and this power law has been referred to as reflecting so-called *neuronal avalanches*. Neuronal avalanches are predicted if spike propagation is effectively a critical branching process (although there are alternative explanations; see Touboul & Destexhe, 2010). Critical branching holds when precisely one spike is propagated for each spike, on average. The observed 3/2 exponent is predicted by virtue of critical branching being associated with a second-order phase transition, and the finding has been replicated across a wide range of neural systems and measurements (for a recent review see Roberts, Boonstra, & Breakspear, 2015).

The associations between neuronal avalanches, critical branching, and metastability may not be readily apparent (see Plenz & Thiagarajan, 2007). Explaining these associations falls outside the present scope, but we note them here because they underlie the upcoming model, and because they reveal a third potential benefit of transient representation. In particular, metastability may maximise the number of different spike patterns that networks can generate, which is analogous to the benefit of greater representational capacity using transient instead of stable codes. The reason for this potential benefit stems from the hypothesis that metastability in neural networks comes from interactions among neurons balanced between mutual independence and interdependence (Kello & Van Orden, 2009). This balance can promote the spontaneous formation of transient spike patterns, and also increase and even maximise the number of different patterns that spontaneously emerge (Tognoli & Kelso, 2014).

A model of transient localist representations

The rationale and evidence for transient representations has prompted some researchers to build computational models that simulate transient patterns of activity. For instance, Haldeman and Beggs (2005) presented a feed-forward, stochastic model of spike propagation that was

simple enough to provide a clear demonstration of how critical branching can maximise pattern formation. The spiking behaviours of neurons were transient, but the model did not represent information about inputs or outputs, and therefore was not representational. Rabinovich, Huerta, Varona, and Afraimovich (2008) presented a mathematical model of transient neural dynamics and showed how transience is compatible with stable, repeatable performance. They described how transient dynamics could be applied to sequential decision-making, which would be representational, but they did not implement such a model. Kwok and Smith (2005) implemented transient representations in a network of threshold gate units and showed that transience can serve to dynamically search a problem space to find multiple solutions, rather than settling on just one. However, the response properties of individual units were not analysed.

When taken together, the modelling studies just reviewed begin to demonstrate how it is possible to simulate transient dynamics in neural network models, and use transient dynamics for information processing. However, none of these models simulated spike dynamics or transient localist representations. Also they did not directly demonstrate how neuronal units can change representations on fast timescales while learning and maintaining performance. Here we present new analyses of a previously published model (Rodny & Kello, 2014) that converges through learning on stable performance in a simple classification task, yet ongoing regulation of network connectivity is shown to produce continual, sometimes abrupt changes to localist representations. These changes are driven by a simple mechanism that adjusts connectivity to balance spike propagation towards its critical branching point (Kello, 2013). As a result, a unit may represent one learned category for an extended period of time, and then switch abruptly to represent another learned category, or switch to become uncorrelated with learned categories. But despite this transience in representation, performance is maintained over output units.

In the next section, we review the model in detail for the interested reader, and further details can be found in the original paper (Rodny & Kello, 2014). We then conduct new analyses on model performance to show directly how the spiking of some units exhibited transient localist representations of output targets. We end with implications for future investigations of stable versus transient representations in neural networks.

Review of critical branching network with learning

Rodny and Kello (2014) presented a spiking reservoir network (Lukoševičius & Jaeger, 2009) with three groups of units: A group of excitatory source units, a group of excitatory and inhibitory reservoir units, and a group of sink units. Source units were activated by external inputs, and their spikes propagated into the reservoir. Reservoir spikes either propagated within reservoir via recurrent synapses, or they exited the reservoir by triggering spikes on sink units. Sink units did not have outgoing connections so their spikes did not propagate further. Units were standard leaky integrate-and-fire, and connectivity was sparse and random, with random axonal transmission delays. Importantly, the synaptic strength of each connection could take on only one of two values (see Branco & Staras, 2009): *disabled* (zero) or *enabled* (a strong + or – value for excitatory or inhibitory units, respectively). Synaptic strengths were switched between two values to create the possibility of relatively rapid changes in connectivity (i.e. with just a small number of switches) that cause transience in representation, as reported later.

The synaptic switching algorithm was designed to enable excitatory synapses when the local branching ratio was subcritical (i.e. <1), and disable excitatory synapses when local estimates were supercritical (>1). Enabling and disabling synapses increased and decreased spike propagation to regulate around the critical branching point. Kello (2013) showed that the algorithm can bring a spiking network to its critical branching point, and also produce neuronal avalanches as predicted. However, that study focused on the intrinsic dynamics of critical branching networks, which means that spikes were not associated with stimuli or responses. Such associations are necessary to investigate transient localist representations.

Rodny and Kello (2014) extended the critical branching algorithm to incorporate a simple learning mechanism, and thereby make spikes "representational" in the sense that each one has a corresponding effect on learned responses to stimuli. In the model, spikes produced by sink units received rewards or punishments, and a trace was added to each synapse that kept track of correlations between signals transmitted across the synapse, and associated rewards or punishments. Synaptic trace values were used to guide the enabling and disabling of synapses. In particular, when the critical branching algorithm signalled a synaptic modification, synapses with high reward correlations were selectively enabled, and synapses with low reward correlations (or high punishment correlations) were selectively disabled. These traces effectively "corralled" transient spike dynamics to generate patterns that increase the probability of generating spikes correlated with rewards.

The nonlinear classification task used by Rodny and Kello (2014) was the exclusive-or (XOR) function, which

is a standard assay of nonlinear classification performance in reservoir computing. In particular, each output spike represented XOR = 0 or XOR = 1, and spikes representing each value were counted during each response period. On each trial the network responded with the value that elicited the greater number of spikes. The sink units that carried output spikes had stable representations because their values were assigned during network initialisation, and these assignments remained fixed for the duration of a simulation. However, reservoir units were not assigned values, so their representations were free to be transient.

The critical branching mechanism continually enabled and disabled recurrent synapses in the reservoir as well as output synapses projecting onto sink units. These synaptic modifications continued even after reward probabilities reached their asymptotic values, because critical branching was never perfectly nor permanently achieved. This inability to achieve perfect critical branching made reservoir spike dynamics transient, because connectivity never became fixed. The effect of dynamic connectivity on spike propagation and performance can be seen in Figure 1, which shows that even after performance reached asymptote at 0.91 mean proportion correct, the model continued to exhibit large fluctuations due to ongoing changes in connectivity.

Continually changing connectivity also resulted in metastable spike patterns. To examine metastability, Rodny and Kello (2014) analysed reservoir spike dynamics as time series of spike patterns, where each pattern was a vector of unit spike counts over a given window of time (also see Sasaki, Matsuki, & Ikegaya, 2007). Pattern overlap was computed using correlation, and an example autocorrelation matrix is shown in the upper left graph of Figure 2. One can see local correlations among spike patterns nearby in time along the diagonal, contrasted with a relative lack of correlation at longer time lags away from the diagonal. This lack of correlation indicates that spike patterns were constantly shifting over time, which can also be visualised by computing the first two principal components of the spike pattern time series. The resulting pattern trajectory in PCA coordinates is plotted in the upper right panel of Figure 2. The transience of pattern dynamics can be seen as a wandering path through PCA space. For the sake of comparison, the same analyses were conducted while the critical branching mechanism was disengaged (after reaching asymptote), thereby freezing connectivity. The bottom two panels show that spike patterns were highly correlated over time while connectivity was frozen, and pattern dynamics moved randomly within a small region of PCA space.

Analyses of transient localist representations

Results reviewed thus far show that a simple mechanism of critical branching can lead to transient spike patterns in terms of metastability. Next, in order to examine the transience of representation in Rodny and Kello's (2014) critical branching network, we need to measure the relationship between reservoir spikes and their effects on output spikes generated by sink units. Specifically, for each reservoir spike, we measured whether its signal arrived at the sink while XOR = 0 was the correct output, or XOR = 1 was the correct output. By this measure, a purely stable, localist representation would correspond to a reservoir unit that spiked only when XOR = 0 or XOR = 1 was correct. A transient localist representation would exhibit runs of XOR = 0 interleaved with XOR = 1. To be distinguished from a unit whose runs occur by chance, run lengths would need to be greater than expected by chance.

We first tested for stable localist representations by examining the long-term biases of reservoir units to spike more when XOR = 0 or XOR = 1. Figure 3 shows two bias distributions using two different measures of bias, plus a third distribution of spike counts per unit that serves as a reference. One bias distribution is the number of spikes produced by each reservoir unit while XOR = 1, minus the number of spikes it produced while XOR = 0. The other is the distribution of proportion of spikes produced by each reservoir unit while XOR = 1, divided by total number of spikes it produced. The bimodality of these two bias distributions is clear evidence that a greater-than-chance number of reservoir units exhibited at least somewhat stable localist representations, in that they exhibited long-term biases towards spiking selectively when XOR = 0 or XOR = 1. The histograms also show an asymmetric bias towards XOR = 1, which means that idiosyncratic factors (e.g. due to initialisation and asynchronous updating) were amplified over the course of learning.

While some reservoir units exhibited somewhat stable localist representations, the histograms show that many units exhibit little or no bias, leaving open the possibility that the model also learned transient localist representations. For instance, only 28% of reservoir units had a localist bias, as defined by greater than 0.8 probability of spiking when XOR = 0 or XOR = 1, exclusively. However, less than 2% of the neurons had a strong localist bias, as defined by a greater than 0.98 probability of spiking when XOR = 0 or XOR = 1, exclusively. Nearly 40% of the units had no strong bias either way, with their spikes roughly evenly divided between the two XOR outputs (proportions between 0.4 and 0.6). Thus

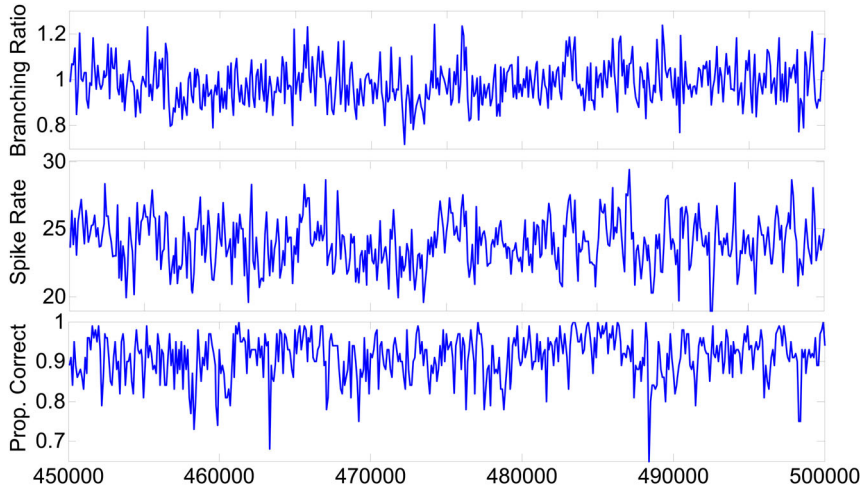

Figure 1. Time series of mean branching ratio estimates, mean reservoir spike rate, and mean proportion of correct responses over time. The X axis is simulation time and data points were averaged every 100 simulation time steps.

we found that very few units converged on localist representations stable over the whole course of asymptotic performance. However, neuroscience experiments record for limited periods of time, so we need to analyse representations on shorter timescales in order to test for relatively stable versus transient representation. For the latter, we need to look for switches between runs of XOR = 0 and XOR = 1 biases in the time series of reservoir spike trains.

We start by showing in Figure 4 some example spike trains that represent the different types of runs observed in reservoir units. In particular, we searched through reservoir activity to find spike trains that exemplify both stable and transient localist representation, as well as spike trains that exhibit no clear relationship with output values. Figure 4 shows raster plots for spike trains exhibiting (1) stable localist representations corresponding to XOR = 0 or XOR = 1; (2) transient localist representations that switch between longer-than-typical runs of XOR = 0 and XOR = 1; and (3) no consistent representation with respect to output values. These spike trains illustrate the diversity of reservoir unit activity in the network. They show that reservoir representations can be unstable while learned classification performance is stable.

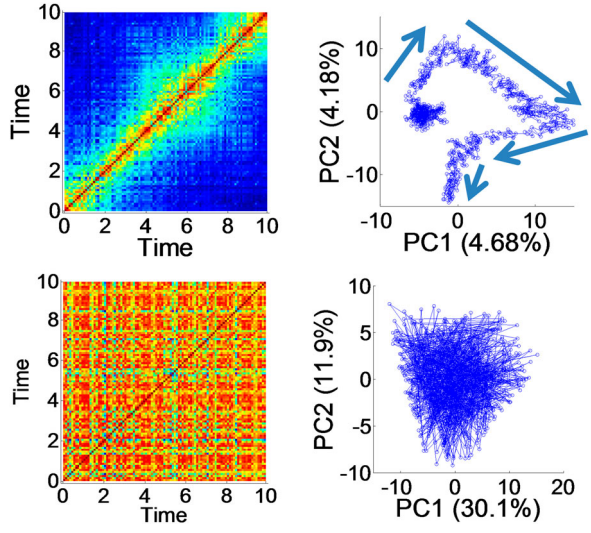

Figure 2. Autocorrelation (left) and PCA (right) analyses of spike pattern time series for the reservoir spiking network while spike propagation was regulated to be critical branching (top), and while regulation was disengaged and connectivity was frozen (bottom). For the autocorrelation plots, red and yellow indicate stronger correlations, light and dark blue indicates little to no correlation.

The examples of localist transient representation shown in Figure 4 are highly unlikely to occur by chance. We further investigate this point by analysing the distribution of observed run lengths as a measure of transience in localist representation. To ensure spike trains of sufficient length for distributional analysis, we selected all reservoir units that spiked at least 5000 times after learning stabilised. For each unit, we counted the number of runs of consecutive spikes corresponding to XOR = 0 or XOR = 1, for each run length *N*. For instance, if a given reservoir unit produce 5 consecutive spikes while XOR = 1 before a 6th spike while XOR = 0, we marked one run of length 5. There are numerous possible distributions, but the evidence for metastability due to critical branching leads to a prediction. Specifically, metastable dynamics are associated with power law distributions in the durations of spike patterns, and a simple proxy for spike pattern durations is the lengths of transient localist runs as just defined. A power law distribution in this case

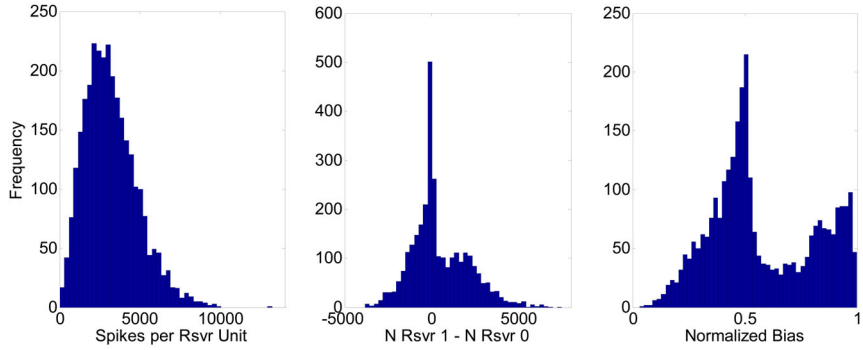

Figure 3. Three histograms based on reservoir spikes produced after learning stabilised. *Left:* number of spikes per unit over an entire simulation (excluding initial transient) *Middle*: Per unit difference between the number of spikes corresponding to XOR = 1 minus XOR = 0. *Right*: Per unit proportion of spikes corresponding to XOR = 1 divided by total number of spikes.

would mean that longer runs occur more often than expected by chance. That is, the tail of the observed distribution should be heavier than an exponential (Poisson) distribution.

Figure 5 plots the distribution of localist run lengths for 10 example reservoir units chosen at random, and also for the aggregate distribution. The distributions are plotted on a log-log scale because power laws appear as linear relations in these coordinates. One can see that run lengths were approximately distributed as a power law because the distribution is roughly linear in logarithmic coordinates (the exponent was approximately two as measured by a regression line fit). This power law relationship means that localist runs occurred more often than expected by chance. Therefore learning and regulation in the critical branching network resulted in transient localist representations for reservoir units.

Another way to investigate transience in localist representation is to simulate the placement of an electrode tapping a random neuron at a random time point, and quantify the probability of finding a localist representation that is stable for a given number of spikes. Figure 6 plots the probability of finding a consistent run of spikes corresponding to XOR = 0 or XOR = 1 for runs at least 10–1000 spikes in length. Results show that stable representations are relatively rare. For instance, there is only about a 10% chance of finding a neuron with a consistent relationship to the target output for 200 or more consecutive spikes. This probability of finding a stable localist representation is

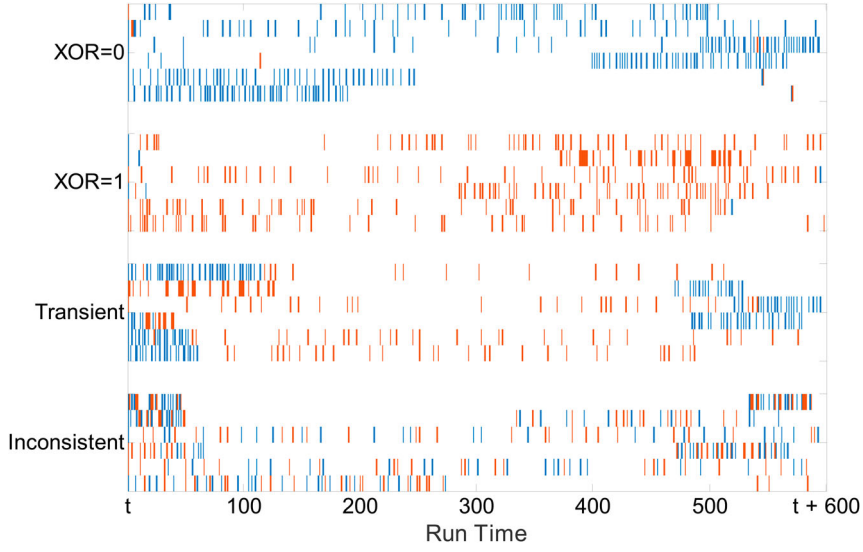

Figure 4. Raster plot of spike trains representing different relationships between reservoir activity and XOR outputs after learning stabilised. Each row contains a sequence of 50 spikes from an example reservoir unit in a period of 600 time units. Spikes are coloured blue or red when they occurred while XOR = 0 or 1, respectively. The top two groups of units show spike trains stably representing XOR = 0 or XOR = 1. The transient group shows spike trains switching between the two different output representations, and the inconsistent group shows spike trains with no consistent relationship to output values.

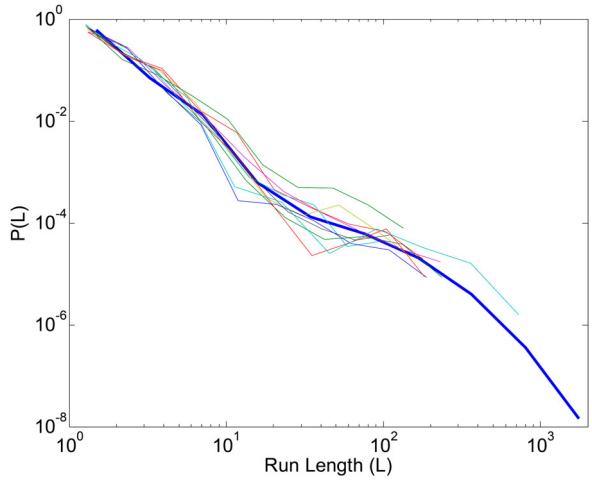

Figure 5. Distribution of localist run lengths for ten randomly selected reservoir units with at least 5000 spikes (thin lines), overlaid by the run length distribution for all reservoir units with at least 5000 spikes (thick blue line).

comparable to neuroscientific findings, albeit it is important to note that recording conditions are perfectly constant in the simulation, whereas real recording conditions are noisy.

Discussion

The debate between distributed and localist representations has been ongoing for many years, and both sides have their strengths and weaknesses. In the present study, we aimed to inform this debate by highlighting an often overlooked aspect of the relationship between neural and mental activity. We presented logical arguments, empirical evidence, and a neural network model that together lead to the proposition of *transient localist representations*. These representations are localist in that the spikes of a given neuron are hypothesised to correspond with individual percepts, concepts, or actions on relatively short timescales (e.g. hours, days, even minutes and seconds). But over a sufficiently long period of time, model neurons were shown to shift among different representations. Transience was localist, but representations shared at least one feature with distributed representation – a given neuron can play a role in multiple representations, with the difference being that a transient localist neuron can only represent a single percept, concept, or action at any one time.

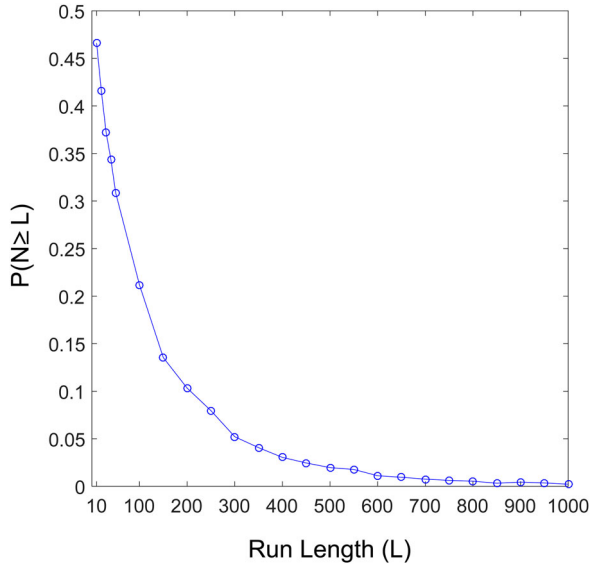

Figure 6. Probability of finding a neuron that exhibits a run of *N* spikes at least *L* long that consistently corresponded to either XOR = 0 or XOR = 1.

The critical branching network demonstrated how transient localist representations can emerge when network connectivity is continually changing due to homeostatic and learning plasticity. The model is primarily a proof-of-concept in this regard, and further work is needed to relate it directly to neuroscience data. For instance, the stimulation timescale is arbitrary, so the model makes no claims about the timescale of transience in real neural systems. Also the model is a generic, randomly connected recurrent network, so it makes no claims about which types of neurons or which brain areas are more likely to exhibit transient representations.

Caveats aside, the model suggest one factor that may contribute to observations of apparently arbitrary relationships between spikes on the one hand, and stimuli, tasks or motor commands on the other. Such neurons may actually be switching their representations among different percepts, concepts, or actions, which might appear arbitrary without examining their time series. We demonstrated that one way to distinguish transient representations from noise is to analyse the distribution of run lengths of localist representation. Analyses like these should be feasible, at least in some single-cell recording studies, and pursued to further test the hypothesis of transient localist representation.

Another future direction is to further investigate the adaptive features of transience for neural and cognitive systems. The increase in representational space over time is an implicit benefit, but a predicted explicit benefit is the responsiveness of representations to adapt to context and changing conditions. We did not test for this benefit in the critical branching model, and we may not be able to without further developing the relationship between synaptic changes driven by homeostasis versus rewards and punishments or input statistics. It would also be helpful to investigate tasks in which contexts and conditions change on the fly, to

see whether transient representations serve to help process perceptual inputs, hold information in memory, or make associations, predictions, and inferences.

The critical branching model may prove useful towards investigating the adaptive benefits of transient representation, localist or otherwise, but there are a number of differences between the model and real neural systems that must be considered. One basic difference is that there were only two categories of output in our model (XOR = 0 versus XOR = 1). Real neural systems deal with many more distinctions among many more categories, and much more variability within categories, especially when it comes to human language and conceptual knowledge. It would be interesting and important to investigate transient representations in models that learn more categories, and more variations within them. It may turn out that the prevalence of transient localist representations depends on factors like the number of categories being learned.

Another basic difference is that the mechanisms and dynamics of units and synapses were greatly simplified in our model compared with real neural systems. Of course all neural network models are greatly simplified, but one must ask whether any of our simplifications played a critical role in our findings of transient localist representations. As already noted, the critical branching mechanism was responsible for metastability, and hence also responsible for the appearance of transient localist representations. Kello (2013) argued that the mechanism and its enabling/disabling of synapses is at least consistent with current neuroscientific data, and its effects explain a number of different power laws observed in neural activity. Therefore the mechanism and its effects appear to provide some insight into metastable dynamics in real spiking neurons, but it is likely that further research is necessary to better capture the relevant principles of homeostatic regulation and plasticity.

Another question raised earlier is whether transient representations are compatible with long-term, stable learning. The critical branching model demonstrates that transient localist representations are compatible with stable classification learning over an extended period of simulation time, and we can be confident that learning would continue to be stable as long as simulation conditions remained the same. However, in the present study we focused on the condition in which rewards and punishments are administered throughout the simulations. Under more realistic conditions, rewards and punishments may not be available for extended periods of time, and nonetheless neural systems continue to function based on past learning. Rodny and Kello (2014) reported a condition in which rewards and punishments were completely removed after asymptotic learning. They found that classification performance continued to be high and stable, even though critical branching continued to enable and disable synapses with rewards or punishments. The reason for maintained performance in the absence of rewards and punishments is that synaptic traces remained fixed and stable, and thus continued to guide the enabling and disabling of synapses.

Finally, it should be noted that the classic stability/plasticity dilemma (Grossberg, 1980) is a potential issue for our model, as it is for other neural network models. If the model is trained on one task, and then the task changes along with rewards and punishments, the first task may be unlearned. This phenomenon is known as "catastrophic interference" (Lewandowsky & Li, 1995), and it is a long-standing issue in neural network research with many hypotheses on how it can be addressed and mitigated. In a critical branching network, catastrophic interference may not be a problem if spike activity is sparse enough and tasks elicit distinct enough reservoir spike patterns to minimise overlap. It is also possible that reservoir dynamics can be generic enough to support projections to multiple output layers representing multiple tasks, as is the premise of reservoir computing. Or, the critical branching mechanism could be modified to avoid changes to synapses as they become more consistently enabled or disabled, as proposed by Kello (2013). Further work is needed to test all these possibilities.

To conclude, our critical branching model does not provide evidence for or against localist or distributed representations, nor was it intended to. Instead the model highlights the possibility that evidence for "grandmother cells" and other types of localist representations may belie neural systems in which representations are more transient than often assumed. Grandmother cells may emerge under a variety of contexts and conditions, but they may also change into grandfather cells sometimes, or redwood cells or who-knows-what cells, and they might even transition from sparse to dense to distributed to localist representations, and back again. Together, the evidence, theories, and models covered herein suggest a need for further investigation into the transience of neural representations and their potentially beneficial properties.

Disclosure statement

No potential conflict of interest was reported by the authors.

References

Beggs, J. M., & Plenz, D. (2003). Neuronal avalanches in neocortical circuits. *The Journal of Neuroscience*, *23*, 11167–11177.

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

Branco, T., & Staras, K. (2009). The probability of neurotransmitter release: Variability and feedback control at single synapses. *Nature Reviews Neuroscience*, *10*(5), 373–383. doi:10.1038/nrn2634

Bressler, S. L., & Kelso, J. A. S. (2001). Cortical coordination dynamics and cognition. *Trends in Cognitive Sciences*, *5*(1), 26–36.

Chestek, C. A., Batista, A. P., Santhanam, G., Yu, B. M., Afshar, A., Cunningham, J. P., … Shenoy, K. V. (2007). Single-neuron stability during repeated reaching in Macaque Premotor Cortex. *The Journal of Neuroscience*, *27*(40), 10742–10750. doi:10.1523/JNEUROSCI.0959-07.2007

Christensen, T. A., Pawlowski, V. M., Lei, H., & Hildebrand, J. G. (2000). Multi-unit recordings reveal context-dependent modulation of synchrony in odor-specific neural ensembles. *Nature Neuroscience*, *3*(9), 927–931.

Colbran, R. J. (2015). Thematic Minireview series: Molecular mechanisms of synaptic plasticity. *Journal of Biological Chemistry*, *290*(48), 28594–28595. doi:10.1074/jbc.R115.696468

De Pitta, M., Volman, V., Berry, H., Parpura, V., Volterra, A., & Ben-Jacob, E. (2012). Computational quest for understanding the role of astrocyte signaling in synaptic transmission and plasticity. *Frontiers in Computational Neuroscience*, *6*. doi:10.3389/fncom.2012.00098

Durstewitz, D., & Deco, G. (2008). Computational significance of transient dynamics in cortical networks. *European Journal of Neuroscience*, *27*(1), 217–227. doi:10.1111/j.1460-9568.2007.05976.x

Erickson, R. P. (2001). The evolution and implications of population and modular neural coding ideas. *Progress in Brain Research*, *130*, 9–29.

Freeman, W. J. (1994). Characterization of state transitions in spatially distributed, chaotic, nonlinear, dynamical systems in cerebral cortex. *Integrative Physiological and Behavioral Science*, *29*(3), 294–306.

Ganguly, K., & Carmena, J. M. (2009). Emergence of a stable cortical map for neuroprosthetic control. *PLoS Biology*, *7*(7), e1000153. doi:10.1371/journal.pbio.1000153

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*(3), 121–134.

Grossberg, S. (1980). How does a brain build a cognitive code? *Psychological Review*, *87*, 1–51.

Haldeman, C., & Beggs, J. M. (2005). Critical branching captures activity in living neural networks and maximizes the number of metastable states. *Physical Review Letters*, *94*(5), 058101. doi:10.1103/PhysRevLett.94.058101

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 77–109). Cambridge, MA: MIT Press.

Holtmaat, A. J. G. D., Trachtenberg, J. T., Wilbrecht, L., Shepherd, G. M., Zhang, X., Knott, G. W., & Svoboda, K. (2005). Transient and persistent dendritic spines in the neocortex in vivo. *Neuron*, *45*(2), 279–291. doi:10.1016/j.neuron.2005.01.003

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591.

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243.

János, A. P., Mark, L. H., Wasim, Q. M., Sydney, C., Emad, E., Gerhard, F., … Leigh, R. H. (2013). Intra-day signal instabilities affect decoding performance in an intracortical neural interface system. *Journal of Neural Engineering*, *10*(3), 036004. doi:10.1088/1741-2560/10/3/036004

Kato, H. K., Chu, M. W., Isaacson, J. S., & Komiyama, T. (2012). Dynamic sensory representations in the olfactory bulb: Modulation by wakefulness and experience. *Neuron*, *76*(5), 962–975. doi:10.1016/j.neuron.2012.09.037

Kello, C. T. (2013). Critical branching neural networks. *Psychological Review*, *120*(1), 230–254.

Kello, C. T., Anderson, G. G., Holden, J. G., & Van Orden, G. C. (2008). The pervasiveness of 1/f scaling in speech reflects the metastable basis of cognition. *Cognitive Science*, *32*(7), 1217–1231. doi:10.1080/03640210801944898

Kello, C. T., & Van Orden, G. C. (2009). Soft-assembly of sensorimotor function. *Nonlinear Dynamics, Psychology, and Life Sciences*, *13*(1), 57–78.

Kwok, T., & Smith, K. A. (2005). Optimization via intermittency with a self-organizing neural network. *Neural Computation*, *17*(11), 2454–2481. doi:10.1162/0899766054796860

Lewandowsky, S., & Li, S.-C. (1995). Catastrophic interference in neural networks: Causes, solutions, and data. In F. N. D. C. J. Brainerd (Ed.), *Interference and inhibition in cognition* (pp. 329–361). San Diego, CA: Academic Press.

Lukoševičius, M., & Jaeger, H. (2009). Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, *3*, 127–149. doi:10.1016/j.cosrev.2009.03.005

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*(5), 375–407.

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, *37*(23), 3311–3325.

Page, M. (2000). Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences*, *23*, 476–477. doi:10.1017/S0140525X00003356

Pastalkova, E., Itskov, V., Amarasingham, A., & Buzsáki, G. (2008). Internally generated cell assembly sequences in the rat hippocampus. *Science*, *321*(5894), 1322–1327. doi:10.1126/science.1159775

Person, R. S., & Kudina, L. P. (1972). Discharge frequency and discharge pattern of human motor units during voluntary contraction of muscle. *Electroencephalography and Clinical Neurophysiology*, *32*(5), 471–483.

Plenz, D., & Thiagarajan, T. C. (2007). The organizing principles of neuronal avalanches: Cell assemblies in the cortex? *Trends in Neurosciences*, *30*(3), 101–110. doi:10.1016/j.tins.2007.01.005

Quiroga, R. Q., & Kreiman, G. (2010). Measuring sparseness in the brain: Comment on Bowers (2009). *Psychological Review*, *117*(1), 291–297. doi:10.1037/a0016917

Quiroga, R. Q., Kreiman, G., Koch, C., & Fried, I. (2008). Sparse but not 'Grandmother-cell' coding in the medial temporal lobe.

Trends in Cognitive Sciences, *12*(3), 87–91. doi:10.1016/j.tics.2007.12.003

Quiroga, R. Q., 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

Rabinovich, M. I., Huerta, R., Varona, P., & Afraimovich, V. S. (2008). Transient cognitive dynamics, metastability, and decision making. *Plos Computational Biology*, *4*(5), e1000072. doi:10.1371/journal.pcbi.1000072

Roberts, J. A., Boonstra, T. W., & Breakspear, M. (2015). The heavy tail of the human brain. *Current Opinion in Neurobiology*, *31*, 164–172. doi:10.1016/j.conb.2014.10.014

Rodny, J., & Kello, C. T. (2014). Learning and variability in spiking neural networks. In P. Bello, M. Guarini, M. McShane, & B. Scassellati (Eds.), *Proceedings of the 37th annual meeting of the Cognitive Science Society* (pp. 1305–1310). Quebec City: Cognitive Science Society.

Rokni, U., Richardson, A. G., Bizzi, E., & Seung, H. S. (2007). Motor learning with unstable neural representations. *Neuron*, *54*(4), 653–666. doi:10.1016/j.neuron.2007.04.030

Rosenblatt, F. (1961). *Principles of neurodynamics: Perceptions and the theory of brain mechanisms*. Washington, DC: Spartan Books.

Rossini, P. M., Martino, G., Narici, L., Pasquarelli, A., Peresson, M., Pizzella, V., … Romani, G. L. (1994). Short-term brain 'plasticity' in humans: Transient finger representation changes in sensory cortex somatotopy following ischemic anesthesia. *Brain Research*, *642*(1–2), 169–177.

Sasaki, T., Matsuki, N., & Ikegaya, Y. (2007). Metastability of active CA3 networks. *The Journal of Neuroscience*, *27*(3), 517–528. doi:10.1523/JNEUROSCI.4514-06.2007

Selfridge, O. G. (1958). *Pandemonium: A paradigm for learning in mechanisation of thought processes.* Paper presented at the Proceedings of a Symposium Held at the National Physical Laboratory (pp. 513–526). London: HMSO.

Tin-Yau, K., & Dit-Yan, Y. (1997). Constructive algorithms for structure learning in feedforward neural networks for regression problems. *IEEE Transactions on Neural Networks*, *8*(3), 630–645.

Tognoli, E., & Kelso, J. A. S. (2014). The metastable brain. *Neuron*, *81*(1), 35–48. doi:10.1016/j.neuron.2013.12.022

Touboul, J., & Destexhe, A. (2010). Can power-law scaling and neuronal avalanches arise from stochastic dynamics? *PLoS ONE*, *5*(2), e8982. doi:10.1371/journal.pone.0008982

Wittgenstein, L. (1953). *Philosophical investigations*. New York, NY: MacMillan.