Title (8 words): The importance of single-trial analyses in cognitive neuroscience

By Mark Stokes & Eelke Spaak

Abstract (50 words)
Theories of working memory typically assume that information is maintained via persistent neural activity. In contrast, Lundqvist et al. report that single-trial delay activity is actually 'bursty'; the classic profile of persistent activity is an artifact of trialwise averaging. Tackling brain-behaviour relationships at the single-trial level is an important future direction for cognitive neuroscience.

Body (1000 words, 10 references, one fig)
Classic models of working memory assume that information is stored via persistent neural activity [1]. Since the earliest neurophysiological experiments in awake, behaving primates [2], researchers have reported evidence that working memory is maintained via persistent delay activity in prefrontal cortex (PFC). The idea is relatively simple: task-relevant mental representations are kept 'on-line' by maintaining corresponding patterns of neural activity for as long as required.

As in most neurophysiological studies, however, classic evidence for persistent activity was essentially based on the results of many individual trials averaged together to form a putative 'representative' estimate of neural activity. Averaging across trials is important to improve the signal-to-noise ratio (SNR). If we can assume each trial is a noisy sample of the true distribution, the average over trials should represent the signal we could expect from a single noiseless trial. However, the brain does not operate according to this construct of the 'average response'. Real-world interaction requires real-time perception, encoding, and decision making. To understand the neural basis of behaviour, including working memory, we need to understand the neural dynamics that unfold within a single trial.

In a recent paper, Lunqvist et al. [4] developed a novel method for characterising trial-wise neural dynamics during working memory. Specifically, they were interested in whether previous evidence for sustained high-frequency (gamma) activity during a working memory delay, as recorded from local field potentials (LPF) in primate PFC, is actually evident at the single trial, or whether the trialwise dynamics are qualitatively different from the 'average response'. To test this, they developed a 'burstiness' metric to quantify the temporal structure of gamma activity on single trials, i.e., before averaging over trials for statistical analyses. Critically, by averaging over this second-order characteristic (i.e., burstiness), they provide new evidence that the unbroken persistent activity apparent in the trial-averaged representation is actually made up of numerous bursts of gamma activity (see Figure 1a, left panel). The authors conclude that this pattern of stochastic bursting is not consistent with standard models of working memory that posit an unbroken chain of persistent firing [see also

ref 3], but instead favours 'activity-silent' models in which memories are stored in hidden neural states, such as rapid short-term synaptic plasticity [5].

Although these finding are clearly important for models of working memory [6], the trial-wise approach also has important broader implications for cognitive neuroscience. To properly understand brain-behaviour relationships, we need to consider the neural dynamics as they unfold on a single trial. Trial-averaging can misrepresent neural dynamics in many situations. For example, a recent study showed that even stimulus-driven activity in the gamma band is intermittent (or 'bursty') on single trials, despite the robust sustained profile apparent in the trial average [7]. Similarly, in the decision-making literature, single-trial analyses have revealed that activity in parietal neurons actually displays discrete transitions on single trials, rather than a gradual accumulation of decision-related information observed in the trial-average response [8, see Figure 1a, middle panel]. These types of approaches could also be used to address a number of long standing debates, such as graded versus discrete allocation of attention (see Figure 1a, right panel).

Exploring single-trial dynamics is not trivial. Any measure of neural activity is subject to noise, therefore we need multiple samples for any analysis to have sufficient statistical power. We typically achieve statistical power by combining information over trials (which one might call 'vertical power'), but this can easily smooth over important heterogeneity. In [4], the key development was to develop first-level metrics ('burstiness') that can be averaged across trials for statistical inference, without losing the essential structure of interest. But this approach critically depends on an a priori model of the single-trial phenomena (e.g., burstiness in [4], discrete state transitions in [8]). Future methodological developments will allow for much more general approaches to single trial analyses. For example, increasing the number of simultaneous measurements (which we propose to call 'lateral power') will be crucial in more detailed examination of single trial dynamics (Figure 1b). Boosting the number of samples within a single trial will provide qualitatively new insights into the neural dynamics that underpin real-time behaviour.

The prospects of rich single-trial analyses are especially exciting with ever-expanding lateral power in neurophysiology (multi-electrode arrays, calcium imaging, etc.) and with multi-channel methods with high temporal resolution in humans, such as ECoG in patients and MEG in healthy participants. Multivariate decoding techniques from the field of machine learning provide particularly powerful approaches to leverage the lateral power of high- dimensional neural recordings. These allow increasingly detailed characterisation of single-trial information, crucial for not only our understanding of real-time cognition but also for the development of robust brain-computer interfaces. It is becoming increasingly evident that neural activity is inherently high-dimensional, especially in prefrontal cortex where mixed selectivity expands the potential coding space for flexible cognition [9]. Lateral power is

essential to sample dimensionality in real-time, rather than infer coding diversity from the 'representative trial' derived from multiple repetitions of potentially heterogeneous events.

To conclude, a basic assumption in many empirical sciences is that averaging over repeated observations allows us to combine equivalent signals, while cancelling out random noise. However, not all trial-wise variation is noise. In very many circumstances, trial-wise averaging will also cancel out important signals. Consider Galton's [10] bean machine (or 'quincunx'): a marble falls down a board with many pins, and with each hit of a pin the marble has a 50% chance of falling to either side of that pin. At the bottom of the board are several buckets, one of which will finally catch the marble at the end of its fall. By the central limit theorem, the distribution of marbles in the buckets will tend towards a Gaussian distribution. While this is a very useful and important characterization of the data, it does not help us to understand how any particular marble ended up in its particular bucket; the history of the individual marble is lost in the average. For the next big step forward in cognitive neuroscience, we need to focus on the equivalent of individual marbles hitting individual pins: real-time single-trial dynamics.

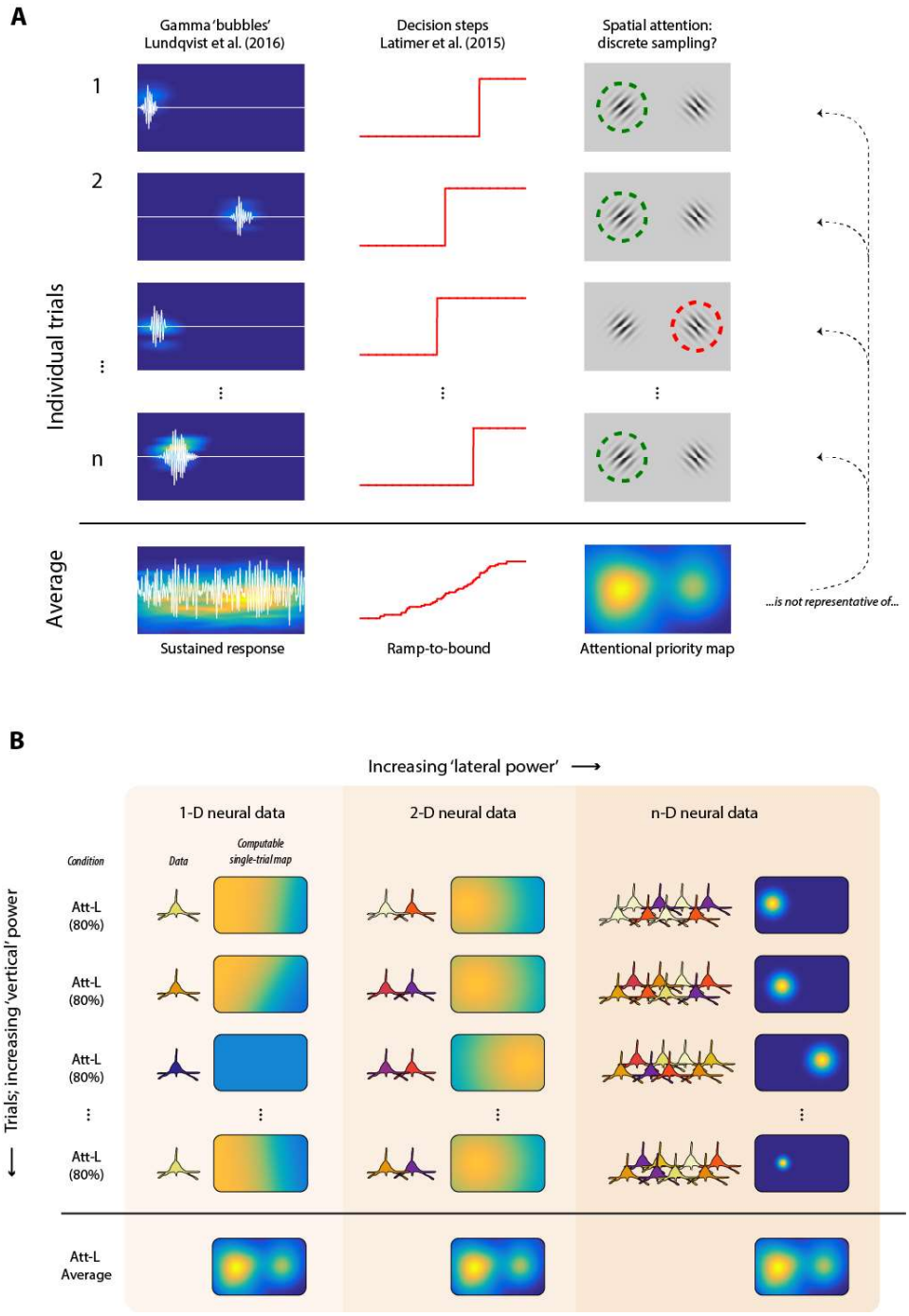

Figure 1

(A) Combining single trials through computing the mean does not necessarily result in a representative response profile. Left: gamma-frequency 'bubbles' [4] (time-frequency representations of power, with traces superimposed) in the prefrontal cortex during individual trials of working memory maintenance activity. The average shows a familiar sustained gamma response, but qualitatively misrepresents the single trial dynamics. Middle: neurons display discrete steps reflecting the time of sensory decisions [8]. The average response shows a classic ramp-to-bound process for the decision. Right: spatial attention might be distributed

in a continuous fashion throughout the visual field (as in the average, bottom), but such an average profile could also be caused by individual trials sampling discretely from visual space (80% of trials on the left, 20% on the right). (B) Typically, statistical power refers to adding extra observations (i.e., trials) over which to average data ('vertical' power; reflected in the vertical dimension here). One can instead think of 'lateral' power: adding more measurements per unit of observation (horizontal dimension). Increasing lateral power is essential for characterizing the neural dynamics on a single-trial level.

1 Sreenivasan, K.K. *et al.* (2014) Revisiting the role of persistent neural activity during working memory. *Trends Cogn. Sci.* 18, 82–89
2 Fuster, J.M. and Alexander, G.E. (1971) Neuron activity related to short-term memory. *Science* 173, 652–654
3 Shafi, M. *et al.* (2007) Variability in neuronal activity in primate cortex during working memory tasks. *Neuroscience* 146, 1082–1108
4 Lundqvist, M. *et al.* (2016) Gamma and Beta Bursts Underlie Working Memory. *Neuron* 0,
5 Sandberg, A. *et al.* (2003) A working memory model based on fast Hebbian learning. *Netw. Bristol Engl.* 14, 789–802
6 Stokes, M.G. (2015) "Activity-silent" working memory in prefrontal cortex: a dynamic coding framework. *Trends Cogn. Sci.* 19, 394–405
7 Lowet, E. *et al.* (2015) Areas V1 and V2 show microsaccade-related 3–4-Hz covariation in gamma power and frequency. *Eur. J. Neurosci.* DOI: 10.1111/ejn.13126
8 Latimer, K.W. *et al.* (2015) Single-trial spike trains in parietal cortex reveal discrete steps during decision-making. *Science* 349, 184–187
9 Rigotti, M. *et al.* (2013) The importance of mixed selectivity in complex cognitive tasks. *Nature* 497, 585–590
10 Galton, Francis (1894) *Natural Inheritance*, Macmillan. 18, 82–89