Large nuisance modulation has little impact on IT target match performance

Noam Roth, Nicole C. Rust[1]
Department of Psychology, University of Pennsylvania, Philadelphia, PA

[1]Corresponding author.

Summary:

Many everyday tasks require us to extract a specific type of information from our environment while ignoring other things. When the neurons in our brains that carry task-relevant signals are also modulated by task-irrelevant "nuisance" information, nuisance modulation is expected to act as performance-limiting noise. To investigate the impact of nuisance modulation on neural task performance, we recorded responses in inferotemporal cortex (IT) as monkeys performed a task in which they were rewarded for indicating when a target object appeared amid considerable nuisance variation. Within IT, we found a robust, behaviorally-relevant target match signal that was mixed with large nuisance modulations in individual neurons. Unexpectedly, we also found that these nuisance modulations had little impact on performance, either within individual IT neurons or across the IT population. We demonstrate how these results follow from fast processing in IT, which placed IT in a low spike count regime where the impact of nuisance variability was blunted by Poisson-like trial variability. These results demonstrate that some basic intuitions about neural coding are misguided in the context of a fast-processing, low spike count regime.

Introduction:

Task performance is determined not only by the amount of task-relevant signal present in our brains, but also by the presence of noise, which can arise from multiple sources. Internal noise, or "trial variability" manifests as trial-by-trial variations in neural responses under seemingly identical conditions (Fig 1a). External factors can also translate into noise, particularly when a task requires extracting a particular type of information from our environment amid changes in other task-irrelevant, nuisance parameters (Fig 1b; Haefner and Bethge, 2010; Kim et al., 2016). Stated differently, for any given task, neurons in a brain area may be modulated by multiple experimental variables, but when viewed from the perspective of task performance, one type of modulation reflects the task-relevant signal, whereas other types of modulations act as noise.

Despite notions that mixing different types of signals within the responses of individual neurons should be detrimental for task performance (Fig 1b), growing evidence suggests that the brain does often mix them, both at the locus at which task-relevant solutions are computed as well as downstream (Freedman and Assad, 2009; Kobak et al., 2016; Mante et al., 2013; Meister et al., 2013; Raposo et al., 2014; Rigotti et al., 2013; Rishel et al., 2013; Zoccolan et al., 2007). One example is visual target search, which requires the brain to compare incoming visual information with a remembered representation of a target to create a signal that reports when a target match is in view. When considered across changes in target identity (e.g. looking for your car keys and then your wallet), target search can be envisioned as differentiating the same images presented as target matches versus as distractors (e.g. when looking for your car keys, your

wallet is a distractor; when looking for your wallet, your car keys are distractor and your wallet is a target match). Consequently, other types of modulation, such as visual modulation (e.g. signals that differentiate wallets and car keys regardless of what you are searching for), act as noise. A number of lines of evidence suggest that target match information emerges in the ventral visual pathway as early as V4 (Kosai et al., 2014; Maunsell et al., 1991) and inferotemporal cortex (IT, Chelazzi et al., 1993; Eskandar et al., 1992; Leuschow et al., 1994; Miller and Desimone, 1994; Pagan et al., 2013), where nuisance modulation, including visual modulation, is expected to be large. This suggests that nuisance modulation may place strong limitations on neural target match performance in these ventral visual pathway brain areas.

Understanding how nuisance modulation affects neural task performance requires considering its impact in individual neurons as well as across the population. Investigations, focused in part on view-invariant object recognition, have demonstrated the means by which individual neurons can multiplex different types of signals such that each type of signal can be extracted from the population with a simple linear decoder (DiCarlo and Cox, 2007; Hong et al., 2016; Hung et al., 2005; Li et al., 2009). But little attention has been directed toward understanding how signal mixing impacts population performance within the context of these linearly separable representations – e.g. under what conditions would a linearly separable population be better off to parse different signals into non-overlapping subpopulations versus mix them? Some insight into these issues can be gained from work focused on how correlated interactions between neurons impacts population performance within a linear decoding scheme (reviewed by Averbeck et al., 2006; Cohen and Kohn, 2011; Kohn et al., 2016). However, this work has focused nearly exclusively on correlated trial (as opposed to nuisance) variability (but see Kim et al., 2016). Understanding how nuisance modulation impacts neural task performance will thus require extending these population-based approaches to incorporate considerations about nuisance modulation.

To investigate the impact of nuisance modulation on IT target match performance, we recorded neural signals in IT as monkeys performed a modified delayed-match-to-sample task in which they were rewarded for indicating when a target object appeared across changes in the objects' position, size and background context (Fig 2). Our study was motivated by the simple intuition that when neurons are modulated not only by task-relevant signals but also by large, task-irrelevant nuisance modulations, nuisance modulations should act as a source of noise that limits task performance (Fig 1b). As described in more detail below, we did indeed find a robust, behaviorally-relevant target match signal in IT that was in fact mixed with even larger nuisance modulations within the responses of individual IT units. Unexpectedly, we also found that these large nuisance modulations had little impact on neural task performance. We reconciled these seemingly discrepant results by also considering another source of noise, trial variability, which was considerably larger than both signal and nuisance modulations. As we demonstrate below, large trial variability acts to blunt the impact of nuisance modulation on task performance, both within individual IT units and across the IT population.

Results:

The invariant delayed-match-to-sample task (IDMS)

To investigate the degree to which nuisance modulation impacts neural task performance, we trained two monkeys to perform an "invariant delayed-match-to-sample" (IDMS) task that required them to report when target objects appeared across variation in the objects' positions, sizes and background contexts. In this task, the target object was held fixed for short blocks of trials (~3 minutes on average) and each block began with a cue trial indicating the target for that block (Fig 2a, "Cue trial"). Subsequent test trials always began with the presentation of a distractor and on most trials this was followed by 0-5 additional distractors (for a total of 1-6 distractor images) and then an image containing the target match (Fig 2a, "Test trial"). The monkeys' task required them to fixate during the presentation of distractors and make a saccade to a response dot on the screen following target match onset to receive a reward. To minimize the predictability of the match appearing as a trial progressed, on a small subset of the trials the match did not appear and the monkey was rewarded for maintaining fixation. Our experimental design differs from other classic DMS tasks (Chelazzi et al., 1993; Eskandar et al., 1992; Leuschow et al., 1994; Miller and Desimone, 1994; Pagan et al., 2013) in that it does not incorporate a cue at the beginning of each test trial, to better mimic real-world object search conditions in which target matches are not repeats of the same image presented shortly before.

Our experiment included a fixed set of 20 images, broken down into 4 objects presented at each of 5 transformations (Fig 2b). Our goal in selecting these specific images was to make the task of classifying object identity challenging for the IT population and these specific transformations were built on findings from our previous work (Rust and DiCarlo, 2010). In any given block (e.g. a squirrel target block), a subset of 5 of the images would be considered target matches and the remaining 15 would be distractors (Fig 2b). Our full experimental design amounted to 20 images (4 objects presented at 5 identity-preserving transformations), all viewed in the context of each of the 4 objects as a target, resulting in 80 experimental conditions (Fig 2c). In this design, "target matches" fall along the diagonals of each looking at / looking for matrix slice (where "slice" refers to a fixed transformation; Fig 2c, gray). For each condition, we collected at least 20 repeats on correct trials. Monkeys generally performed well on this task (Fig 2d). Their mean reaction times (computed as the time their eyes left the fixation window relative to the target match stimulus onset) were 366 ms and 332 ms (Fig 2e).

As two monkeys performed this task, we recorded neural activity from small populations in inferotemporal cortex (IT) using 24-channel probes. We performed two types of analyses on these data. The first type of analysis was performed on the data recorded simultaneously across units within a single recording session (n=21 sessions). The second type of analysis was performed on data that was concatenated across different sessions to create a pseudopopulation after screening for units based on their stability, isolation, and task modulation (see Methods; n=235 units). For all but one of our analyses (Fig 4d), we counted spikes in a window that started 80 ms following stimulus onset (to allow stimulus-evoked responses time to reach IT) and ended at 250 ms, which was always before the monkeys' reaction times on these trials. For all but one of our analyses (Fig 3c), the data are extracted from trials with correct responses.

IT reflects behaviorally-relevant target match information

The primary focus of this report is the impact of mixing signal and nuisance modulation on neural task performance. Before exploring the consequences of nuisance modulation, we begin by demonstrating that behaviorally-relevant target match information is in fact reflected in IT during the IDMS task.

The IDMS task required monkeys to determine whether each condition (an image viewed in the context of a particular target) was a target match or a distractor. This task ultimately maps all the target match conditions onto one behavioral response (a saccade) and all the distractor conditions onto another (maintain fixation), and as such, this task can be envisioned as a two-way classification that must be performed invariant to changes in other nuisance parameters, including changes in target and image identity (Fig 3a). To quantify the amount and format of target match information within IT, we began by quantifying cross-validated performance of this two-way classification with a linear population decoder (a Fisher Linear Discriminant, FLD). Linear decoder performance began near chance and grew as a function of population size, consistent with a robust IT target match representation (Fig 3b, black). To determine the degree to which a component of IT target match information might be present in a nonlinear format that could not be accessed by a linear decoder, we measured the performance of a maximum likelihood decoder designed to extract target match information regardless of its format (combined linear and nonlinear, Pagan et al., 2013, see Methods). Performance of this nonlinear decoder (Fig 3b, gray) was slightly higher and significantly better than linear decoder performance, suggesting that while the majority of IT target match information is reflected in a linearly separable format, a smaller nonlinear component exists as well.

Upon establishing the format of target match information on correct trials, we were interested in determining the degree to which behavioral confusions were reflected in the IT neural data. To measure this, we focused on the data recorded simultaneously across multiple units within each session, where all units observed the same errors. With this data, we trained the linear decoder to perform the same target match versus distractor classification described for Fig 3b using data from correct trials, and we measured cross-validated performance on pairs of condition-matched trials: one for which the monkey answered correctly, and the other for which the monkey made an error. On correct trials, target match decoder performance grew with population size and reached levels significantly above chance in populations of 24 units (Fig 3c, black). On error trials, decoder performance fell significantly below chance, and these results replicated across each monkey individually (Fig 3c, white). These results establish that IT reflects behaviorally-relevant target match information insofar as this measure co-varies with the monkeys' behavior.

We were also interested in understanding how target match modulation was reflected in individual units. Target match modulation, by definition, requires a differential response to the same images presented as matches versus as distractors - to what degree is this modulation reflected by firing rate increases versus decreases? To measure this, we computed a target match modulation index for each unit as the average difference between the responses to the same images presented as target matches versus as distractors, divided by the sum of those two quantities. This index (Fig 3d) was shifted toward target match preferring units, with a mean value of 1.0 (monkey 1 = 1.21; monkey 2 = 0.70). These results are consistent with a target match signal that is largely reflected in most IT neurons via increased responses to target matches as compared to distractors.

During the IDMS task, nuisance modulation is prominent

As described above, we were interested in understanding whether and how nuisance modulation impacted IT target match performance. As a first step toward addressing this question, we wanted to quantify the relative amounts of target match and nuisance modulation present within individual units. To quantify the different types of modulation reflected in IT, we applied a bias-corrected procedure that quantified different types of modulation in terms of the number of standard deviations around each unit's grand mean spike count (Pagan and Rust, 2014b). Modulation types were grouped into intuitive sets, including modulation that could be attributed to whether each condition was a target match or a distractor (the "target match" signal), modulation due to changes in the identity of the visual stimulus ("visual"), modulation due to changes in the identity of the target ("target"), and "residual" modulations attributed to nonlinear interactions between the visual stimulus and target that were not captured by target match modulation (e.g. specific distractor conditions). We also combined all the different types of "nuisance" modulation into one measure for each neuron.

Our measure of modulation is similar to a multi-way ANOVA, with important extensions. Specifically, a two-way ANOVA applied to a unit's responses (configured into a matrix of 4 targets * 20 images * 20 trials for each condition) would parse the total response variance into two linear terms, a nonlinear interaction term, and an error term. We make 3 extensions to the ANOVA analysis. First, an ANOVA returns measures of variance (in units of spike counts squared) whereas we compute measures of standard deviation (in units of spike count) such that our measures of modulation are intuitive (e.g., doubling firing rates causes signals to double as opposed to quadruple). Second, while the linear terms of the ANOVA map onto our "visual" and "target" modulations (after squaring), we split the ANOVA nonlinear interaction term into two terms, including target match modulation (i.e. Fig 2c gray versus white) and all other nonlinear "residual" modulation. This parsing is essential, as target match modulation corresponds to the signal for the IDMS task whereas residual modulation acts as noise (described in more detail below, Fig 4b). Finally, raw ANOVA values are biased by trial-by-trial variability (which the ANOVA addresses by computing the probability that each term is higher than chance given this noise) whereas our measures of modulation are bias-corrected to provide an unbiased estimate of modulation magnitude (see Methods).

Across the 235 IT units, we found that total nuisance modulation was larger than target match modulation in most cases (Fig 4a), and that average nuisance modulation was 2.7x the average target match signal (Fig 4b). A more detailed parsing of the total nuisance modulation into different subtypes revealed that the largest type of nuisance modulation could be attributed to "visual" modulations (on average 2.5x the target match signal; Fig 4b). Other types of modulation were also prominent, including "target" modulations (on average 0.8x the target match signal; Fig 4b), and "residual" modulation (on average 0.6x the target match modulation; Fig 4b). These results reveal that within IT, nuisance modulations are prominent and they are mixed with the target match signal in individual units.

In sum, the results presented thus far verify the existence of a robust, behaviorally-relevant target match signal in IT, and they confirm our predictions that IT target match signals are mixed with large nuisance modulations within individual IT units. Together, these results support assertions that the activity of IT units during visual target search should be an effective test of the impact that nuisance modulation has on neural task performance.

Unexpectedly, the impact of nuisance modulation on single-unit performance is modest

Ultimately, understanding the impact of nuisance modulation on linearly decoded task performance requires considering both the responses of individual units as well as their population interactions. Here we begin by quantifying the impact of nuisance modulation on individual units, the results of which were quite unexpected.

As a measure of linearly decoded target match performance for individual units, we focus on single-unit d' (Fig 1b). Single-unit d' is determined by the separation between the spike count responses of a unit to the set of all images presented as target matches versus the same images presented as distractors, and is quantified as the ratio between the distance between the means over the average standard deviation of the two distributions (Fig 1b). Single-unit d' is thus proportional to the amount of "target match signal", equivalent to the distance between the means of the responses to target matches and to distractors (Fig 1b, cyan). Conversely, single-unit d' is inversely proportional to the spread within each distribution, where spread is determined by two factors. The first contributor to this spread is the variability in the spike count responses across repeated trials of the same condition, or "trial variability" (Fig 1b, purple). The second contributor to this spread is the dispersion between different conditions within each set, equivalent to all types of modulation that are not the target match signal ("nuisance" modulation; Fig 1b, red). This is why signal mixing is predicted to be detrimental to single-unit task performance – because any nuisance modulation that exists within a unit is predicted to increase the overlap between target matches and distractors.

In a previous report, we formalized these intuitions into a mathematical relationship between the single-unit modulation magnitudes as measured in Fig 4a-b and single-unit d' (Pagan and Rust, 2014b). This derivation can be applied here with minor extensions. To summarize that approach, d' is a measure of the ratio between signal and noise, where signal is proportional to the amount of target match modulation (Fig 4b, cyan) and noise is parsed into one component proportional to total nuisance modulation (Fig 4b, red) and another component proportional to trial variability (Fig 4b, purple):

$$|\mathrm{d}'| = \sqrt{\frac{k_1 * target\ match\ modulation^2}{k_2 * nuisance\ modulation^2 + trial\ variability^2}}$$

where k_1 and k_2 are constants (see Methods). With this formulation, the impact of nuisance modulation on d' can be determined by considering the increase in d' when nuisance modulation is incorporated into the calculation (i.e. for the intact data) compared to when it is not (i.e. a hypothetical scenario in which nuisance modulation does not exist, analogous to the increase in d' in Fig 1a relative to 1b). Fig 4c shows the result of this analysis, which reveals that removing nuisance only results in a modest increase in d' across units, with an average increase of 8.6%. Focusing on the most informative units (i.e. those with the highest d'), did not change the qualitative nature of the result (average impact for the top 25%, 15%, 10% of units = 9.8%, 9.3% and 9.3% respectively).

This modest increase was surprising in light of the fact that nuisance modulations were 2.7x the target match signal (Fig 4b, compare cyan and dark red bars), coupled with the intuition that large nuisance modulation should be highly detrimental to task performance (Fig 1b). However, this result can be understood by examining the trial variability component of the noise, which was even larger than the nuisance component (Fig 4b, compare red and purple bars) and as a result, dominated the denominator of the d' derivation. As an illustrative example, compare

ratios of the numbers 5/(10+100)=0.045 versus 5/(0+100)=0.05; while the first component of the denominator, 10, is 2-fold the size of the numerator (5), including versus excluding it only leads to a change in the total ratio of 10% because the denominator is dominated by the second entry, 100. Consequently, although the amount of nuisance modulation is large relative to the size of the target match signal, its impact is blunted by the existence of trial variability, which is even larger. Stated differently, while IT nuisance modulations are larger than the IT target match signal, both are small relative to the size of trial variability. Because trial variability is so much larger than nuisance variability, the existence of nuisance modulation has little consequence for d'.

Large trial variability in IT is a consequence of fast processing

Why is trial variability so much larger than nuisance modulation (and signal modulation) in our data? During the IDMS task, spike count windows were short, as a consequence of terminating the count window before the monkeys' reaction times, which were fast (Fig 2e; total counting window duration 170 ms, 80-250 ms following stimulus onset). Within these short spike count windows, the average grand mean spike count was 1 spike per condition per trial, and the average peak spike count across the 80 conditions was 2.77 spikes (which translates into mean and peak firing rates of 5.8 spikes/sec and 16.3 spikes/sec, respectively). We also found that, consistent with earlier reports, IT trial variability was approximately Poisson (average variance-to-mean ratio across units = 1.22, relative to the Poisson benchmark of 1.0). Simple simulations confirm that within a low spike count, Poisson regime, trial variability is much larger than signal modulation. Large trial variability in IT thus does not arrive from exotic mechanisms, rather, it is a natural consequence of the low spike counts that follow from fast processing, coupled with Poisson-like trial variability.

To illustrate how the impact of nuisance modulation depends on overall spike count, we recalculated the impact of nuisance modulation as a function of increasing window size. In this analysis, we always started the spike count window for each unit at 80 ms following stimulus onset, and we ended the count window at different times up to 170 ms total duration (equivalent to the count window for the analyses presented in Fig 4b-c). These results illustrate a systematic increase in the impact of nuisance modulation on task performance as a function of spike count window duration (Fig 4d), consistent with the interpretation that the impact of nuisance modulation is inversely proportional to the overall spike count.

To illustrate that the amount of signal mixing we observed would have impacted task performance at higher spike counts than we recorded in our data (e.g. if counting windows were longer and/or firing rates were higher), we performed a simulation in which we rescaled the responses for each unit in our data (after noise correction, see Methods). Specifically, we kept the proportions and types of signal and nuisance modulation for each unit intact, but rescaled the trial-averaged spike count responses for each unit by different factors of N, followed by the reintroduction of Poisson trial variability. We then recomputed the impact of nuisance modulation on single-unit d' as described for Fig 4c-d. We found that the impact of nuisance on d' grew substantially with rescaling (Fig 5c). For example, with a 6-fold rescaling, which roughly translates into a 1 second counting window (under the assumption that the response properties are constant with time), eliminating nuisance resulted in a 50.2% increase in d' (as compared to the 11.7% increase in simulation with no rescaling). The increased impact of nuisance with rescaling cannot not be attributed to changes in the relative amounts of signal and nuisance modulation, as these remained fixed with rescaling (compare Fig 4b and 5b, cyan, red). Rather,

the increased impact of nuisance with rescaling is due to a decrease in magnitude of trial variability relative to the magnitudes of signal and nuisance modulation (compare Fig 4b and 5b, purple).

Together, these results indicate that mixing signals in a fast processing regime (where spike counts are low) has the unexpected consequence that nuisance modulation is largely inconsequential for task performance. In contrast, our simulations reveal that mixing signals in the same proportions but in regime where spike counts are high (e.g. with long integration windows and/or higher firing rates) would be highly detrimental. These results thus suggest that within IT during the IDMS task, the potentially deleterious impact of nuisance modulation is blunted by virtue of a fast processing, low spike count regime.

The impact of nuisance modulation on population performance is also modest:

As we demonstrate in this section, the impact of nuisance modulation on IT performance described above for single units (Fig 4c), remains modest even when population factors are considered. To address population considerations, we begin with a data-based "pseudosimulation" approach that allows us to compute important benchmarks for our results. However, because these simulations require assumptions about the data, we also verify our results with analyses applied directly to neural data.

To estimate the impact of nuisance modulation on IT population performance, we applied approach similar in concept to the single-unit analysis presented in Fig 4c, where we estimated the impact of nuisance by comparing the intact data with a hypothetical version of our data with nuisance removed. However, in the case of the population, we did not have an analytical solution and we thus performed pseudosimulations to determine it. To perform this analysis, we simulated the responses of two versions of each unit: an intact version with the same number and types of signals as well as the same grand mean spike count (after noise correction, see Methods), and a version in which the nuisance modulation was removed. In both cases, we simulated trial variability for each unit with independent, Poisson process. Cross-validated linear decoder performance, measured in units of population d', grew within increasing population size for the intact and nuisance-removed populations with an approximately fixed ratio (Fig 6a). Consequently, the proportional impact of nuisance modulation as a function of population size took on a value of ~15%, regardless of population size (Fig 6b, solid). These results suggest that the impact of nuisance modulation in our data is modest and does not depend on population size (under the assumption that trial variability is Poisson and is independent between units).

Our simulation-based approach allowed us to estimate the impact of nuisance modulation on population performance relative to a benchmark of the same population but without nuisance. However, our pseudosimulations incorporate the assumption that trial variability is independent (i.e. uncorrelated) between units, whereas we do in fact expect it to be weakly correlated (e.g. Cohen and Maunsell, 2009). How might the existence of weakly correlated variability impact our results? To summarize the well-established framework for thinking about correlated trial variability (reviewed by Averbeck et al., 2006; Cohen and Kohn, 2011; Kohn et al., 2016), when the component of trial variability that falls along a linear decoding axis is uncorrelated between neurons, it will average away as a function of population size. Relative to this benchmark, correlated trial variability has the potential to either be beneficial or detrimental to performance

(Fig 7a). We have determined that nuisance modulation is similar insofar as the component of nuisance modulation that falls along a linear decoding axis that is uncorrelated between neurons will average away as a function of population size. Relative to this benchmark, interactions between neurons can configure nuisance modulation to have beneficial or detrimental consequences (Fig 7b).

When a task does not include nuisance variability (e.g. a two-way discrimination between exactly two conditions), the impact of correlated trial variability on population performance can be measured by comparing performance for the simultaneously recorded, intact data with performance when the trials are independently shuffled for each unit to destroy correlations (Averbeck and Lee, 2006). Increases in performance with shuffling indicate that noise correlations are detrimental (Fig 7a, left) whereas decreases in performance indicate that noise correlations are beneficial (Fig 7a, right). This shuffling procedure can be extended for tasks that incorporate a nuisance component by comparing population performance for the intact data with performance when the experimental conditions are shuffled independently for each unit within each class (i.e. shuffling conditions within the set of target matches and within the set of distractors).

To assess the impact of both correlated trial and nuisance variability on IT population performance, we analyzed the raw, simultaneously recorded data within each session. Here we present the results only for populations of size 24 (to simplify the data, given the number of comparisons of interest). Relative to the intact data, shuffling trial variability resulted in a small increase in performance (Fig 7c, “Intact” versus “Shuffle TV”; proportional increase with shuffling = 12%), indicating that correlated trial variability is aligned along the target match decoding axis in a manner that is weakly detrimental. Next we computed performance when both trial and nuisance variability were shuffled, and found that it was slightly higher than shuffling trial variability alone (Fig 7c, “Shuffle TV&NV”; proportional increase = 6%). This suggests that like trial variability, nuisance variability is correlated in a manner weakly detrimental to performance.

How does the existence of weakly detrimental correlated trial and nuisance variability impact the results presented in Fig 6? First, note that the analysis presented in Fig 6 is not impacted by the existence of correlated trial variability (because any correlations were destroyed in the pseudosimulation process). Second, note that Fig 6 presents an estimate of the “total” impact of nuisance variability that captures contributions arising from both the existence of nuisance modulations as well as any detrimental correlations that fall along the decoding axis. To parse their relative contributions, we returned to the pseudosimulation and applied the nuisance shuffling procedure. Shuffling nuisance variability lead to a small proportional increase (relative to shuffling trial variability alone; Fig 7d; 5.4%) that was similar to the value measured for the intact data (6%, as described above). The proportional impact of removing nuisance variability altogether, after shuffling, was 9.4% (Fig 7d; “Shuffle&NV” vs. “Shuffle TV, Remove NV”).

To summarize these results, we measured the impact of nuisance modulation on population performance in simulation by comparing performance of an intact population (with independent trial variability) with a simulation of the same population with nuisance variability removed. In our data, the impact of nuisance modulation was small (~15%) and approximately flat as a function of population size. An analysis targeted at understanding how correlated trial and nuisance variability between units impacts task performance revealed that their contributions to task performance were also measurable but modest, and did not change the interpretation that while

nuisance modulation is large in IT, its impact on task performance (both for single units and for the population) is small.

Discussion:

In many everyday situations, we are faced with the challenge of extracting one type of information from our environment while ignoring many other things that are going on around us. This study was inspired by a very simple intuition: when the neurons involved in computing the solutions for these tasks are modulated by both task-relevant signals as well as task-irrelevant nuisance information, nuisance modulation should be a source of noise that limits our ability to perform these tasks. Unexpectedly, we found that this simple intuition was largely wrong in IT. During a visual target search task, we found that nuisance modulations in IT were indeed large and that they were mixed with task-relevant signals in the responses of individual units, however, their consequences for task performance were modest. This result could be explained by the existence of another noise source, trial variability, which was larger than nuisance variability and blunted its impact on performance. Large trial variability in IT could, in turn, be accounted for by fast processing (implied by fast reaction times), which positioned IT within a low spike count regime, coupled with Poisson trial variability. We found that these results applied not only to individual units but also to the performance of the IT population. Our results thus reveal that when the brain operates in a regime where signals are small relative to the size of trial variability, nuisance modulations are of very little consequence to task performance.

Many of our intuitions about neural coding have been developed within the context of a high spike count regime, largely following on foundational work in early and mid-level visual brain areas in primates (e.g. V1, MT) where firing rates are high. Notably, recent work has called into question whether even in those brain areas, high spike counts do in fact translate into a high signal-to-noise ratio, due to supra-Poisson trial variability that begins to dominate when spike counts are large (Goris et al., 2014). Moreover, the low spike count regime that we present here is likely to be representative of the operating regime in many brain areas during many real-world tasks. The unexpected nature of our results highlights the fact that in this low spike count regime, some of the basic intuitions that we have constructed about neural coding may not hold.

Our results shed insight into why the brain might continue to “mix” modulations for different task-relevant parameters within individual neurons, even at the highest stages. Growing evidence suggests that the brain does not seek to produce neurons with increasingly “pure selectivity” at higher stages of processing, but rather that the brain continues to mix modulations for different task-relevant parameters within individual neurons, both at the locus at which task-relevant solutions are computed, as well as downstream (Freedman and Assad, 2009; Kobak et al., 2016; Mante et al., 2013; Meister et al., 2013; Raposo et al., 2014; Rigotti et al., 2013; Rishel et al., 2013; Zoccolan et al., 2007). A number of explanations have been proposed to account for mixed selectivity. Some studies have documented situations in which signal mixing is an inevitable consequence of the computations required for certain tasks, such as identifying objects invariant to the view in which they appear (Zoccolan et al., 2007). Others have suggested that mixed selectivity may be an essential component of the substrate required to maintain a representation that can rapidly and flexibly switch with changing task demands (Raposo et al., 2014; Rigotti et al., 2013). Still others have maintained that broad tuning across different types of parameters is important for learning new associations (Barak et al., 2013). When viewed from the perspective that signal mixing introduces noise in the form of nuisance

modulation, one might suspect that one or more of these benefits outweigh the performance costs associated with mixed selectivity. However as we demonstrate here, within the fast processing, low spike count regime that most of these high-level brain areas are likely to operate in, large nuisance modulations are expected to have only a modest impact on task performance.

The framework with which we explore how nuisance interactions between different neurons impact population performance builds on foundational work focused on correlated trial variability between units, or "noise correlations" (Averbeck et al., 2006; Cohen and Kohn, 2011; Kohn et al., 2016). Recent work has emphasized the importance of not just measuring the degree to which neurons are correlated, but how those correlations align with a decoding axis and thus how they impact performance (Moreno-Bote et al., 2014). In the visual search task we present here, we found that correlations between units in both trial and nuisance variability had a small, detrimental impact on performance. In other tasks, nuisance interactions along a decoding axis may be much more impactful – such as in the case of dissociating self versus object motion (Kim et al., 2016), and in those cases, other decoding schemes may be required to disambiguate signal from nuisance modulation.

Our results support the existence of a robust target match representation in IT during this task that reflects confusions on trials in which the monkeys make errors (Fig 3c); this result has not been reported previously. One earlier study also explored the responses of IT neurons in the context of a DMS task in which, like ours, the objects could appear at different identity-preserving transformations (Leuschow et al., 1994), but this study did not sort neural responses based on behavior. Target match signals have been investigated most extensively in IT via a classic version of the delayed-match-to-sample (DMS) paradigm where each trial begins with a visual cue indicating the identity of the target object, and this cue is often the same image as the target match. In this paradigm, approximately half of all IT neurons that differentiate target matches from distractors do so with enhanced responses to matches whereas the other half are match suppressed (Miller and Desimone, 1994; Pagan et al., 2013). Because match suppressed responses are thought to arise as the result of passive, stimulus repetition of the target match following the cue, some have speculated that the match enhanced neurons alone carry behaviorally-relevant target match information (Miller and Desimone, 1994). Conversely, others have argued that a representation comprised exclusively of match enhanced neurons would likely confuse the presence of a match with nuisance modulations that evoke changes in overall firing rate, such as changes in stimulus contrast (Engel and Wang, 2011). Additionally, these authors have proposed that matched suppressed neurons could be used in these cases to disambiguate target match versus nuisance modulation. Our results reveal that when target matches do not follow the presentation of the same visual image at a time short time before (as is the case for natural object search), match suppression is very weak (Fig 3e), and consequently, in these cases, this specific disambiguation strategy cannot be employed. Our results also suggest that for the types of nuisance modulation that we have investigated here (changes in position, size and background context), its impact is modest and in these cases, such a strategy is not necessary.

In a previous series of reports (Pagan et al., 2013; Pagan and Rust, 2014a; Pagan et al., 2016), we investigated target match signals in the context of the classic DMS design in which target matches were repeats of cues presented earlier in the trial and each object was presented on a gray background. One of our main findings from that work was that the IT target match representation was reflected in a partially nonlinearly separable format, whereas an IT

downstream projection area, perirhinal cortex, contained the same amount of target match information but in a format that was largely linearly separable. In the data we present here, we also found evidence for a nonlinear component of the IT target match representation, reflected in the significantly higher performance of a maximum likelihood as compared to linear decoder (Fig 3b). However, in this study, a larger proportion of the IT target match representation was linear as compared to our previous DMS results. The source of these quantitative differences is unclear. They could arise from the fact that the IDMS task requires an "invariant" visual representation of object identity, which first emerges in a linearly separable format in the brain area that we are recording from (IT), whereas the DMS task could rely on the visual representation at an earlier stage. Alternatively, these differences could arise from the fact that during IDMS, images are not repeated within a trial, and the stronger nonlinear component revealed in DMS may be produced by stimulus repetition. Our current data cannot distinguish between these alternatives.

Acknowledgments

We thank Marino Pagan for valuable insights, and Margot P. Wohl and Krystal Henderson for their technical contributions. This work was supported by the National Eye Institute of the US National Institutes of Health (award number R01EY020851).

Competing interests

N.C.R. serves as a Board Review Editor for eLife.

METHODS

Experiments were performed on two adult male rhesus macaque monkeys (*Macaca mulatta)* with implanted head posts and recording chambers. All procedures were performed in accordance with the guidelines of the University of Pennsylvania Institutional Animal Care and Use Committee.

The invariant delayed-match-to-sample (IDMS) task:

All behavioral training and testing was performed using standard operant conditioning (juice reward), head stabilization, and high-accuracy, infrared video eye tracking. Stimuli were presented on an LCD monitor with an 85 Hz refresh rate using customized software (http://mworks-project.org).

As an overview, the monkeys' task required an eye movement response to a specific location when a target object appeared within a sequence of distractor images (Fig 2a). Objects were presented across variation in the objects' position, size and background context (Fig 2b). Monkeys viewed a fixed set of 20 images across switches in the identity of 4 target objects, each presented at 5 identity-preserving transformations (Fig 2c). We ran the task in short blocks (~3 min) with a fixed target before another target was pseudorandomly selected. Our design included two types of trials: cue trials and test trials (Fig 2a). Only test trails were analyzed for this report.

Trials were initiated by the monkey fixating on a red dot (0.15°) in the center of a gray screen, within a square window of ±1.5°, followed by a 250 ms delay before a stimulus appeared. Cue trials, which indicated the current target object, were presented at the beginning of each block

and after three subsequent trials with incorrect responses. To minimize confusion, cue trials were designed to be distinct from test trials and began with the presentation of an image of each object that was distinct from the images used on test trials (a large version of the object presented at the center of gaze on a gray background; Fig 2a). Test trials, which are the focus of this report, always began with a distractor image, and neural responses to this image were discarded to minimize non-stationarities such as stimulus onset effects. Distractors were drawn randomly from a pool of 15 possible images within each block without replacement until each distractor was presented once on a correct trial, and the images were then re-randomized. On most trials, a random number of 1-6 distractors were presented, followed by a target match (Fig 2a). On a small fraction of trials, 7 distractors were shown, and the monkey was rewarded for fixating through all distractors. Each stimulus was presented for 400 ms (or until the monkeys' eyes left the fixation window) and was immediately followed by the presentation of the next stimulus. Following the onset of a target match image, monkeys were rewarded for making a saccade to a response target within a window of 75 – 600 ms to receive a juice reward. In monkey 1 this target was positioned 10 degrees below fixation; in monkey 2 it was 10 degrees above fixation. If 400 ms following target onset had elapsed and the monkey had not moved its eyes, a distractor stimulus was immediately presented. If the monkey continued fixating beyond the required reaction time, the trial was considered a "miss". False alarms were differentiated from fixation breaks via a comparison of the monkeys' eye movements with the characteristic pattern of eye movements on correct trials: false alarms were characterized by the eyes leaving the fixation window via its bottom (monkey 1) or top (monkey 2) outside the allowable correct response period and traveling more than 0.5 degrees whereas fixation breaks were characterized by the eyes leaving the fixation window in any other way. Within each block, 4 repeated presentations of the 20 images were collected, and a new target object was then pseudorandomly selected. Following the presentation of all 4 objects as targets, the targets were re-randomized. At least 20 repeats of each condition were collected. Overall, monkeys performed this task with high accuracy. Disregarding fixation breaks (monkey 1: 8% of trials, monkey 2: 12% of trials), percent correct on the remaining trials was as follows: monkey 1: 87% correct, 3% false alarms, and 10% misses; monkey 2: 97% correct, <1% false alarms, and 3% misses.

Neural recording:

The activity of neurons in IT was recorded via a single recording chamber in each monkey. Chamber placement was guided by anatomical magnetic resonance images in both monkeys, and in one monkey, Brainsight neuronavigation (https://www.rogue-research.com/). The region of IT recorded was located on the ventral surface of the brain, over an area that spanned 4 mm lateral to the anterior middle temporal sulcus and 15-19 mm anterior to the ear canals. Neural activity was largely recorded with 24-channel U probes (Plexon, Inc) with linearly arranged recording sites spaced with 100 μm intervals, with a handful of units recorded with single electrodes (Alpha Omega, glass-coated tungsten). Continuous, wideband neural signals were amplified, digitized at 40 kHz and stored using the OmniPlex Data Acquisition System (Plexon). Spike sorting was done manually offline (Plexon Offline Sorter). At least one candidate unit was identified on each recording channel, and 2-3 units were occasionally identified on the same channel. Spike sorting was performed blind to any experimental conditions to avoid bias. The sample size (number of units recorded) was chosen to approximately match our previous work (Pagan et al., 2013; Pagan and Rust, 2014a; Pagan et al., 2016).

For all the analyses presented in this paper, we measured neural responses by counting spikes

in a window that began 80 ms after stimulus onset. For all analyses but Fig 4d, the spike count window ended at 250 ms. On 2.0% of all correct target match presentations, the monkeys had reaction times faster than 250 ms, and those instances were excluded from analysis such that spikes were only counted during periods of fixation. When combining the units recorded across sessions into a larger pseudopopulation, we screened for units that met three criteria. First, units had to be modulated by our task, as quantified by a one-way ANOVA applied to our neural responses (80 conditions * 20 repeats) with $p < 0.01$. Second, we applied a loose criterion on recording stability, as quantified by calculating the variance-to-mean for each unit (computed by fitting the relationship between the mean and variance of spike count across the 80 conditions), and eliminating units with a variance-to-mean ratio > 5. Finally, we applied a loose criterion on unit recording isolation, quantified by calculating the signal-to-noise ratio (SNR) of the waveform (as the difference between the maximum and minimum points of the average waveform, divided by twice the standard deviation across the differences between each waveform and the mean waveform), and excluding (multi)units with an SNR < 2. This yielded a pseudopopulation of 235 units (of 589 possible units), including 99 units from monkey 1 and 126 units from monkey 2.

Population performance:

To determine the performance of the IT population at classifying target matches versus distractors, we applied two types of decoders: a Fisher Linear Discriminant (a linear decoder) and Maximum Likelihood decoder (a nonlinear decoder) using approaches described previously in detail (Pagan et al., 2013) and are summarized here.

When applied to the pseudopopulation data (Fig 3b, Fig 6a), all decoders were cross-validated with the same resampling procedure. On each iteration of the resampling, we randomly shuffled the trials for each condition and for each unit, and (for numbers of neurons less than the full population size) randomly selected units. On each iteration, 18 trials from each condition were used for training the decoder, 1 trial was used to determine a value for regularization, and 1 trial from each condition was used for cross-validated measurement of performance.

To ensure that decoder performance was not biased by unequal numbers of target matches and distractors, on each iteration of the resampling we included 20 target match conditions and 20 (of 60 possible) distractor conditions. Each set of 20 distractors was selected to span all possible combinations of mismatched object and target identities (e.g. objects 1, 2, 3, 4 paired with targets 4, 3, 2, 1), of which there are 9 possible sets. When computing proportion correct (Fig 3b), a mean performance value was computed on each resampling iteration by averaging binary performance outcomes across the 9 possible sets of target matches and distractors, each which contained 40 test trials. Mean and standard error of performance was computed as the mean and standard deviation of performance across 2000 resampling iterations. When computing population d' (Fig 6a), d' was computed on each resampling iteration for the 20 target match conditions and 20 distractor conditions, separately for each set of 9 match/distractor combinations, and then averaged across the 9 sets. Mean and standard error of population d' was computed as the mean and standard deviation of d' across 2000 resampling iterations. For both measures, standard error thus reflected the variability due to the specific trials assigned to training and testing and, for populations smaller than the full size, the specific units chosen.

Fisher Linear Discriminant:

The general form of a linear decoding axis is:

(1) $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b,$

where **w** is an N-dimensional vector (where N is the number of units) containing the linear weights applied to each unit, and b is a scalar value. We fit these parameters using a Fisher Linear Discriminant (FLD), where the vector of linear weights was calculated as:

(2) $\boldsymbol{w} = \Sigma^{-\mathbf{1}}(\mu_1 - \mu_2)$

and b was calculated as:

(3) $b = \boldsymbol{w} \cdot \frac{1}{2}(\mu_1 + \mu_2) = \frac{1}{2}{\mu_1}^T\Sigma^{-\mathbf{1}}\mu_1 - \frac{1}{2}{\mu_2}^T\Sigma^{-\mathbf{1}}\mu_2$

Here μ_1 and μ_2 are the means of the two classes (target matches and distractors, respectively) and the mean covariance matrix is calculated as:

(4) $\Sigma = \frac{\Sigma_1+\Sigma_2}{2}$

where Σ_1 and Σ_2 are the regularized covariance matrices of the two classes. These covariance matrices were computed using a regularized estimate equal to a linear combination of the sample covariance and the identity matrix I (Pagan et al., 2016):

(5) $\Sigma_i = \gamma\,\Sigma_i + (1-\gamma)\cdot I$

To maintain consistency across the different analyses presented throughout the paper, we fixed γ to be 0.9 for all analyses. We determined this value by exploring a range of different γ, and selecting the value that maximized performance globally. Results were qualitatively similar across a range of γ. In this calculation of the FLD parameters, both the information conveyed by individual units as well as their pairwise interactions are considered.

We computed two measures of performance: proportion correct (Fig 3b-c), and population d' (Fig 6a). Each calculation began by computing the dot product of the test data and the linear weights **w**, adjusted by *b* (Eq. 1). Proportion correct was computed as the fraction of test trials that were correctly assigned as target matches and distractors, according to their true labels. Population d' was computed for the distributions of these values across the 20 different images presented as target matches versus as distractors:

(6) $d' = \frac{|\mu_{Match} - \mu_{Distractor}|}{\sigma_{pooled}},$

where μ_{Match} and $\mu_{Distractor}$ correspond to the mean across the set of matches and distractors, $\sigma_{pooled} = \sqrt{\frac{\sigma^2_{Match}+\sigma^2_{Distractor}}{2}}$, and σ_{Match} and $\sigma_{Distractor}$ correspond to the standard deviation across the set of matches and distractors, respectively.

To compare FLD performance on correct versus error trials (Fig 3c), we used the same methods described above with the following modifications. First, the analysis was applied to the simultaneously recorded data within each session, and the correlation structure on each trial was kept intact on each resampling iteration. Second, when more than 24 units were available, a subset of 24 units were selected as those with the most task modulation, quantified via the p-value of a one-way ANOVA applied to each unit's responses (80 conditions * 20 repeats). Finally, on each resampling iteration, each error trial was randomly paired with a correct trial of the same condition and cross-validated performance was performed exclusively for these pairs of correct and error responses. As was the case for the pseudopopulation analysis, training was performed exclusively on correct trials. A mean performance value was computed on each resampling iteration by averaging binary performance outcomes across all possible error trials and their condition-matched correct trial pairs, and averaging across different recording sessions. Mean and standard error of performance was computed as the mean and standard deviation of performance across 2000 resampling iterations. Standard error thus reflected error in a manner similar to the pseudopopulation analysis - the variability due to the specific trials assigned to training and testing and, for populations smaller than the full size, the specific units chosen.

To determine the impact of correlated trial and nuisance variability on IT population performance (Fig 7c-d), we compared the FLD applied to the simultaneously recorded data as described above where the correlation structure on each trial was kept intact on each resampling iteration (Fig 7c, "intact"), with two different shuffling procedures. In the first, we randomly shuffled the trials within each condition, for each unit, on each iteration of the bootstrap (Fig 7c, "Shuffle TV"). In the second, we randomly shuffled both trial variability as well as the order of assignment of each the 20 distractor conditions and 20 target match conditions on each bootstrap iteration (Fig 7c & 7d, "Shuffle TV & NV").

Maximum likelihood decoder:

As a measure of total IT target match information (combined linear and nonlinear), we implemented a maximum likelihood decoder (Fig 3b). We began by using the set of training trials to compute the average response r_{uc} of each unit u to each of the 40 conditions c. We then computed the likelihood that a test response k was generated from a particular condition as a Poisson-distributed variable:

$$(7)\ lik_{u,c}(k) = \frac{(r_{uc})^{k} \cdot e^{-r_{uc}}}{k!}$$

The likelihood that a population response vector was generated in response to each condition was then computed as the product of the likelihoods of the individual neurons. Next, the likelihood that each test vector arose from the category target match as compared to the category distractor as the product of the likelihoods across the conditions within each category. We assigned the population response to the category with the maximum likelihood, and we computed performance as the fraction of trials in which the classification was correct based on the true labels of the test data.

Quantifying single-unit modulation magnitudes:

To quantify the degree to which the firing rates of individual units were modulated by whether an image was presented as a target match versus as a distractor (Fig 3d), we calculated a target match modulation index for each unit by computing its mean spike count response to target matches and to distractors, and computing the ratio of their difference and their sum.
To quantify the degree to which individual units were modulated by different types of task parameters, we applied a bias-corrected, ANOVA-like procedure described in detail by (Pagan et al., 2016) and summarized here. As an overview, this procedure considers the total variance in the spike count responses for each unit across conditions (n=80) and trials for each condition (m=20), and parses this total variance into the variance that can be attributed to each type of experimental parameter and variance attributed to trial variability. Like an ANOVA, the procedure is designed to parse response variance in an intuitive way, such as the variance that can be attributed to changes in the identity of the visual image, the identity of the target object and whether each condition was a target match or a distractor. These variances are converted into measures of spike count modulation (i.e. standard deviation around each unit's grand mean spike count) via a procedure that includes bias correction for over-estimates in modulation due to noise.

The procedure begins by developing an orthonormal basis of 80 vectors designed to capture all types of modulation with intuitive groupings. The number of each type is imposed by the experimental design. This basis $\boldsymbol{b}$ included vectors $\boldsymbol{b}_i$ that reflected 1) the grand mean spike count across all conditions ($\boldsymbol{b}_1$, 1 dimension), 2) whether the object in view was a target or a distractor ($\boldsymbol{b}_2$, 1 dimension), 3) visual image identity ($\boldsymbol{b}_3 - \boldsymbol{b}_{21}$, 19 dimensions), 4) target object identity ($\boldsymbol{b}_{22} - \boldsymbol{b}_{24}$, 3 dimensions), and 5) "residual", nonlinear interactions between target and object identity not captured by target match modulation ($\boldsymbol{b}_{25} - \boldsymbol{b}_{80}$, 56 dimensions). A Gram-Schmidt process was used to convert an initially designed set of vectors into an orthonormal basis.

Because this basis spans the space of all possible responses for our task, each trial-averaged vector of spike count responses to the 80 experimental conditions $\boldsymbol{R}$ can be re-expressed as a weighted sum of these basis vectors. To quantify the amounts of each type of modulation reflected by each unit, we began by computing the squared projection of each basis vector $\boldsymbol{b}_i$ and $\boldsymbol{R}$. An analytical bias correction was then subtracted from this value (Pagan and Rust, 2014b):

(8) $$w_i^2 = (\boldsymbol{R} \cdot \boldsymbol{b}_i^T)^2 - \frac{\sigma_t^2 \cdot (\boldsymbol{b}_i^T)^2}{m}$$

where σ_t^2 indicates the trial variance, averaged across conditions (n=80), and where m indicates the number of trials (m=20). When more than one dimension existed for a type of modulation, we summed values of the same type. Next, we applied a normalization factor (1/(n-1) where n=80) to convert these summed values into variances. Finally, we computed the square root of these quantities to convert them into modulation units that reflected the number of spike count standard deviations around each unit's grand mean spike count. Target match modulation was thus computed as:

(9) $$\sigma_{TM} = \sqrt{\frac{1}{n-1} \cdot w_2^2}$$

and nuisance modulation was computed as:

(10) $\sigma_{Nui} = \sqrt{\frac{1}{n-1} \cdot \sum_{i=3}^{80} w_i^2}$

Similarly, to compute the different subtypes of nuisance modulation, we replaced the weights w_i^2 in Eq. 10 with the weights that corresponded to the orthonormal basis vectors corresponding to each subtype, including visual modulation ($i = 3\ to\ 21$), target modulation ($i = 22\ to\ 24$), and 3) residual modulation ($i = 25\ to\ 80$), as described above.

We computed the trial variability for each unit ($\sigma_{Trial},$) in an comparable manner as the square root of the average (across conditions) variance across trials:

(11) $\sigma_{Trial} = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} \frac{1}{m-1} \cdot \sum_{t=1}^{m} (s_{it} - s_i)^2}$

where the spike count response for a particular trial t of condition i was s_{it}, and the mean spike count response across all trials of condition i was s_i.

When estimating modulation for individual units, (Fig 4a), the bias-corrected squared values were rectified for each unit before taking the square root. When estimating modulation population means (Fig 4b, 5b), the bias-corrected squared values were averaged across units before taking the square root. Because these measures were not normally distributed, standard error about the mean was computed via a bootstrap procedure. On each iteration of the bootstrap (across 1000 iterations), we randomly sampled values from the modulation values for each unit in the population, with replacement. Standard error was computed as the standard deviation across the means of these newly created populations.

Relating modulation magnitudes and single unit performance (d'):

To determine the impact of nuisance modulation on single unit task performance (Fig 4c-d, Fig 5a) we re-expressed d' (Eq. 6) as a function of the different types of signal modulations described above (Eqs. 8-10):

(12) $d' = \frac{|\mu_{Match} - \mu_{Distractor}|}{\sigma_{pooled}} = \sqrt{\frac{a \cdot \sigma_{TM}^2}{b \cdot \sigma_{Nui}^2 + \sigma_{Trial}^2}}$ where $a = \frac{n-1}{3}$, and $b = \frac{n-1}{n}$

This derivation is described in detail in Pagan & Rust (2014).

To quantify the impact of nuisance modulation on single unit performance (d'), we compared each unit's d' in the presence of nuisance modulation (Eq. 12) versus d' when the nuisance modulation term σ_{Nui} was set to zero (d'_{NoNui}). We then calculated the impact of nuisance modulation as the percent increase in d' without nuisance:

(13) $Impact = \left(\frac{d'}{d'_{NoNui}} - 1\right) \cdot 100\%$

Simulations:

To better understand our results, we performed a number of data-inspired simulations. Each simulation began by computing the bias-corrected weights for each unit as described above (Eq. 8).

To explore how rescaling the spike counts by different factors of N influenced the impact of nuisance modulation (Fig 5), we rectified bias-corrected modulations that fell below zero, recomputed the noise-corrected mean spike count responses for each condition, rescaled the mean spike counts by N, and generated trial variability with an independent Poisson process.

To estimate the impact of nuisance modulation on population performance, we simulated two versions of each of our recorded units (Fig 6a compare “Intact” to “Nuisance removed”; Fig 7d compare “Shuffle TV” and “Shuffle TV & NV” to “Shuffle TV, remove NV”). In the “Intact” version, we computed each unit’s responses as described for the rescaling simulation but with a rescale factor N = 1. In the “Nuisance removed” version, we used a similar procedure but set the modulations corresponding to all nuisance dimensions to zero. The responses were thus computed based on the grand mean spike count response as well as the target match modulation alone.

References:

Averbeck, B.B., Latham, P.E., and Pouget, A. (2006). Neural correlations, population coding and computation. Nat Rev Neurosci *7*, 358-366.

Averbeck, B.B., and Lee, D. (2006). Effects of noise correlations on information encoding and decoding. J Neurophysiol *95*, 3633-3644.

Barak, O., Rigotti, M., and Fusi, S. (2013). The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. J Neurosci *33*, 3844-3856.

Chelazzi, L., Miller, E.K., Duncan, J., and Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. Nature *363*, 345-347.

Cohen, M.R., and Kohn, A. (2011). Measuring and interpreting neuronal correlations. Nat Neurosci *14*, 811-819.

Cohen, M.R., and Maunsell, J.H. (2009). Attention improves performance primarily by reducing interneuronal correlations. Nat Neurosci *12*, 1594-1600.

DiCarlo, J.J., and Cox, D.D. (2007). Untangling invariant object recognition. Trends Cogn Sci *11*, 333-341.

Engel, T.A., and Wang, X.J. (2011). Same or different? A neural circuit mechanism of similarity-based pattern match decision making. J Neurosci *31*, 6982-6996.

Eskandar, E.N., Optican, L.M., and Richmond, B.J. (1992). Role of inferior temporal neurons in visual memory. II. Multiplying temporal waveforms related to vision and memory. J Neurophysiol *68*, 1296-1306.

Freedman, D.J., and Assad, J.A. (2009). Distinct encoding of spatial and nonspatial visual information in parietal cortex. J Neurosci *29*, 5671-5680.

Goris, R.L., Movshon, J.A., and Simoncelli, E.P. (2014). Partitioning neuronal variability. Nat Neurosci *17*, 858-865.

Haefner, R.M., and Bethge, M. (2010). Evalutating neural codes for inference using Fisher Information. Paper presented at: Advances in Information Processing Systems.

Hong, H., Yamins, D.L., Majaj, N.J., and DiCarlo, J.J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. Nat Neurosci *19*, 613-622.

Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863-866.

Kim, H.R., Pitkow, X., Angelaki, D.E., and DeAngelis, G.C. (2016). A simple approach to ignoring irrelevant variables by population decoding based on multisensory neurons. J Neurophysiol *116*, 1449-1467.

913 Kobak, D., Brendel, W., Constantinidis, C., Feierstein, C.E., Kepecs, A., Mainen, Z.F., Qi, X.L.,
914 Romo, R., Uchida, N., and Machens, C.K. (2016). Demixed principal component analysis of
915 neural population data. Elife *5*.

916 Kohn, A., Coen-Cagli, R., Kanitscheider, I., and Pouget, A. (2016). Correlations and Neuronal
917 Population Information. Annu Rev Neurosci *39*, 237-256.

918 Kosai, Y., El-Shamayleh, Y., Fyall, A.M., and Pasupathy, A. (2014). The role of visual area V4 in
919 the discrimination of partially occluded shapes. J Neurosci *34*, 8570-8584.

920 Leuschow, A., Miller, E.K., and Desimone, R. (1994). Inferior temporal mechanisms for invariant
921 object recognition. Cerebral Cortex *5*, 523-531.

922 Li, N., Cox, D.D., Zoccolan, D., and DiCarlo, J.J. (2009). What response properties do individual
923 neurons need to underlie position and clutter "invariant" object recognition? J Neurophysiol *102*,
924 360-376.

925 Mante, V., Sussillo, D., Shenoy, K.V., and Newsome, W.T. (2013). Context-dependent
926 computation by recurrent dynamics in prefrontal cortex. Nature *503*, 78-84.

927 Maunsell, J.H., Sclar, G., Nealey, T.A., and DePriest, D.D. (1991). Extraretinal representations
928 in area V4 in the macaque monkey. Vis Neurosci *7*, 561-573.

929 Meister, M.L., Hennig, J.A., and Huk, A.C. (2013). Signal multiplexing and single-neuron
930 computations in lateral intraparietal area during decision-making. J Neurosci *33*, 2254-2267.

931 Miller, E.K., and Desimone, R. (1994). Parallel Neuronal Mechanisms for Short-Term-Memory.
932 Science *263*, 520-522.

933 Moreno-Bote, R., Beck, J., Kanitscheider, I., Pitkow, X., Latham, P., and Pouget, A. (2014).
934 Information-limiting correlations. Nat Neurosci *17*, 1410-1417.

935 Pagan, M., L.S., U., M.P., W., and Rust, N.C. (2013). Signals in inferotemporal cortex and
936 perirhinal cortex suggest an untangling of visual target information. Nature Neuroscience *16*,
937 1132-1139.

938 Pagan, M., and Rust, N.C. (2014a). Dynamic target match signals in perirhinal cortex can be
939 explained by instantaneous computations that act on dynamic input from inferotemporal cortex.
940 J Neurosci *34*, 11067-11084.

941 Pagan, M., and Rust, N.C. (2014b). Quantifying the signals contained in heterogeneous neural
942 responses and determining their relationships with task performance. J Neurophysiol *112*, 1584-
943 1598.

944 Pagan, M., Simoncelli, E.P., and Rust, N.C. (2016). Neural Quadratic Discriminant Analysis:
945 Nonlinear Decoding with V1-Like Computation. Neural Comput, 1-29.

946 Raposo, D., Kaufman, M.T., and Churchland, A.K. (2014). A category-free neural population
947 supports evolving demands during decision-making. Nat Neurosci *17*, 1784-1792.

948 Rigotti, M., Barak, O., Warden, M.R., Wang, X.J., Daw, N.D., Miller, E.K., and Fusi, S. (2013).
949 The importance of mixed selectivity in complex cognitive tasks. Nature *497*, 585-590.

950 Rishel, C.A., Huang, G., and Freedman, D.J. (2013). Independent category and spatial
951 encoding in parietal cortex. Neuron *77*, 969-979.

952 Rust, N.C., and DiCarlo, J.J. (2010). Selectivity and tolerance ("invariance") both increase as
953 visual information propagates from cortical area V4 to IT. J Neurosci *30*, 12978-12995.

954 Zoccolan, D., Kouh, M., Poggio, T., and DiCarlo, J.J. (2007). Trade-off between object
955 selectivity and tolerance in monkey inferotemporal cortex. J Neurosci *27*, 12292-12307.
956

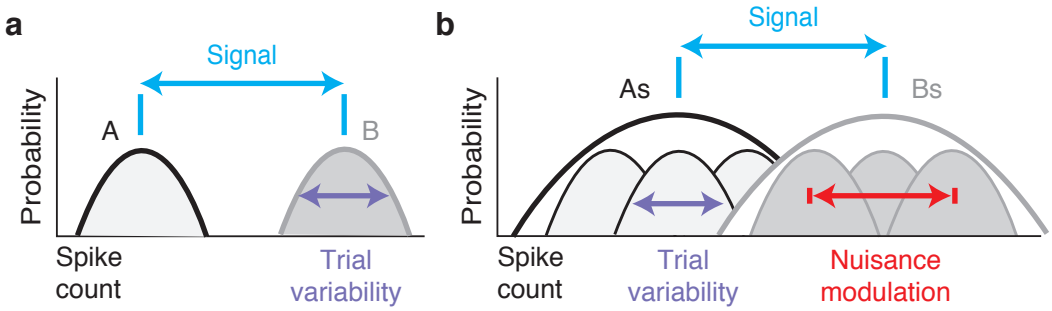

Figure 1. *Nuisance modulation limits task performance.* **a)** Schematic of single unit task performance (d') for a classic, two-way discrimination task in which a subject is asked to label different conditions as "A" or "B" across repeated trials. Shown are hypothetical distributions of spike count responses for the two conditions. Neural task performance, d', is measured as the separation of the two spike count distributions in units of the number of standard deviations separating their means. d' is proportional to the amount of signal, which determines the separation between the means of the distributions (cyan), and d' is inversely proportional to spread within each distribution, which arises as a result of variability across repeated trials within each condition ("trial variability"; purple). **b)** Schematic of single unit task performance (d') for the same discrimination task, but extended to require grouping multiple conditions into each of two sets, "As" and "Bs" (e.g. an object identification task where two objects are presented in multiple background contexts). In this case, "nuisance" modulations (e.g. firing modulations by the background context), increase the spread of the responses within each condition and thus lower d'.

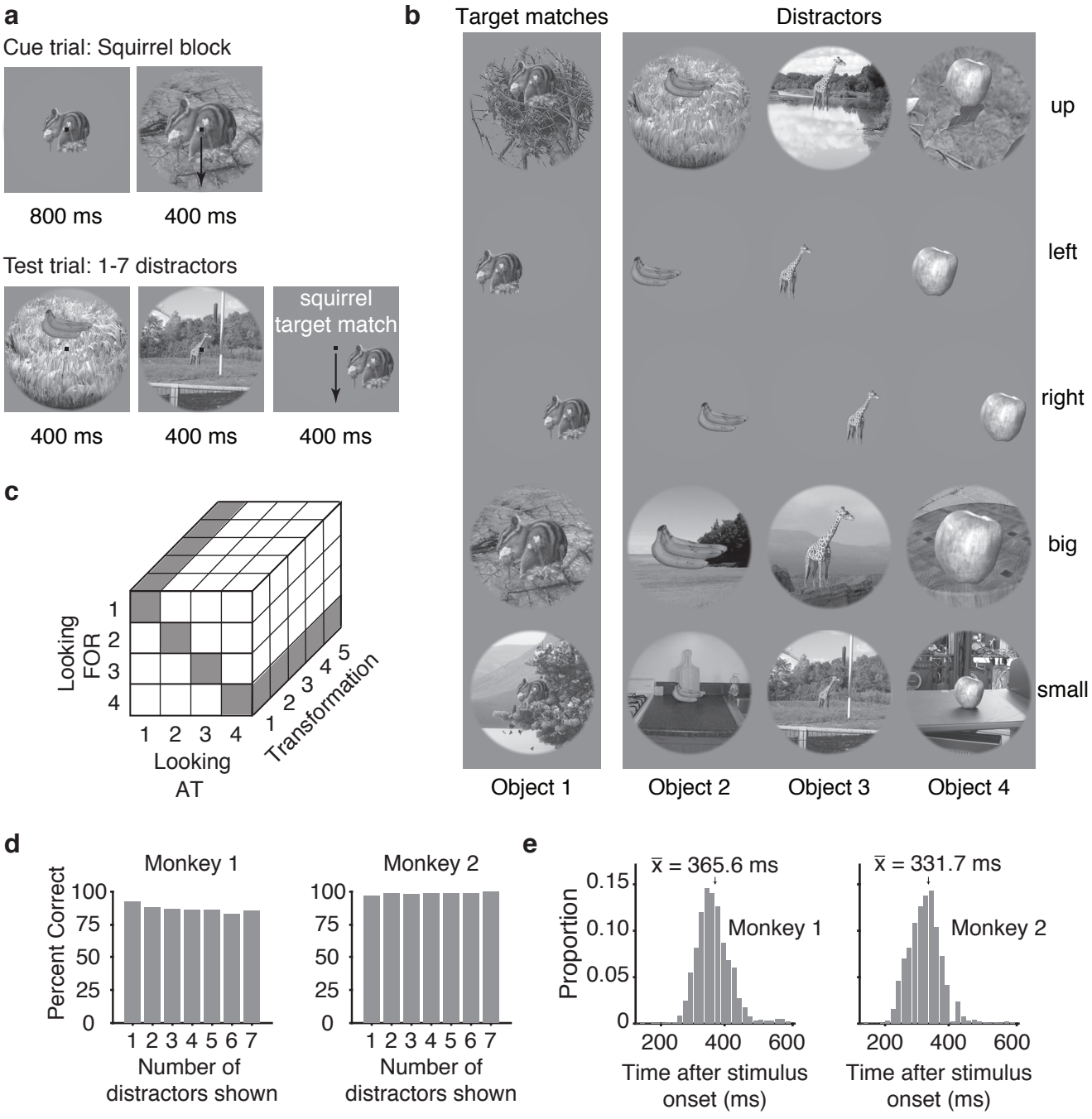

Figure 2. *The invariant delayed-match-to-sample task.* **a)** Monkeys performed an invariant delayed-match-to-sample task. Each block (~3 minutes in duration) began with a cue trial indicating the target object for that block. On subsequent trials, monkeys initiated a trial by fixating on a small dot. After a 250 ms delay, a random number (1-7) of distractors were presented, and on most trials, this was followed by the target match. Monkeys were required to maintain fixation throughout the distractors and make a saccade to a response dot within a window 75 - 600 ms following the onset of the target match to receive a reward. In cases where the target match was presented for 400 ms and the monkey had still not broken fixation, a distractor stimulus was immediately presented. **b)** The experiment included 4 objects presented at each of 5 identity-preserving transformations ("up", "right", "left", "big", "small"), for 20 images in total. In any given block, 5 of the images were presented as target matches and 15 were distractors. **c)** The complete experimental design included looking "at" each of 4 objects, each presented at 5 identity-preserving transformations (for 20 images in total), viewed in the context of looking "for" each object as a target. In this design, target matches (highlighted in gray) fall along the diagonal of each "looking at" / "looking for" transformation slice. **d)** Percent correct for each monkey, calculated based on both misses and false alarms (but disregarding fixation breaks). Percent correct is plotted as a function of the number of distractors shown. **e)** Histograms of reaction times during correct trials (ms after stimulus onset) during the IDMS task for each monkey, with means indicated by arrows and labeled.

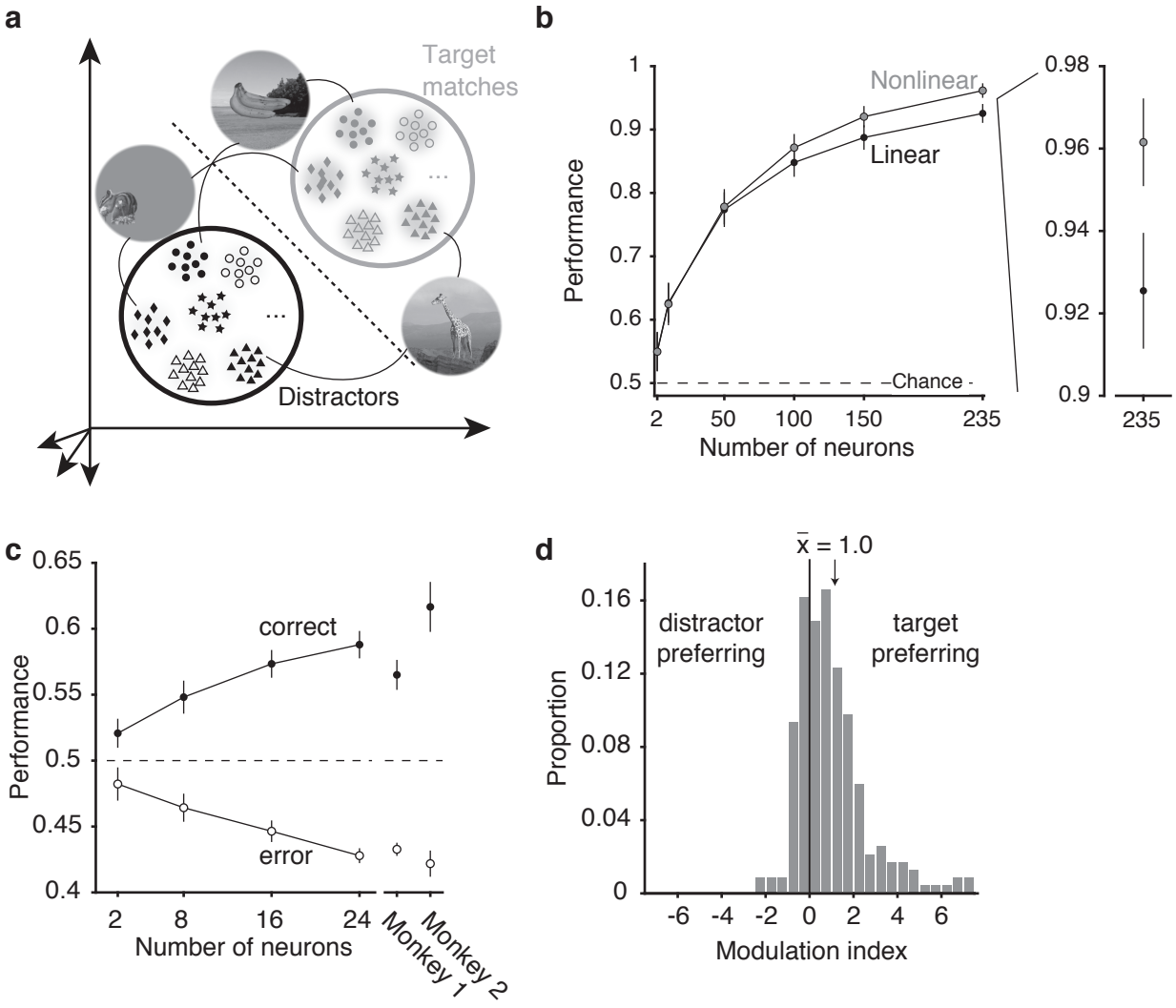

Figure 3. *IT reflects behaviorally-relevant target match information during the IDMS task.* **a)** The target search task can be envisioned as a two-way classification of the same images presented as target matches versus as distractors. Shown are cartoon depictions where each point depicts a hypothetical population response for a population of two neurons on a single trial, and clusters of points depict the dispersion of responses across repeated trials for the same condition. Included are responses to the same images presented as target matches and as distractors - here only 6 images are depicted but 20 images were used in the actual analysis. The dotted line depicts a hypothetical linear decision. **b)** Linear (FLD) and nonlinear (maximum likelihood) decoder performance as a function of population size for a pseudopopulation of 235 units. Error bars (SEM) reflect the variability that can be attributed to the random selection of units (for populations smaller than the full dataset) and the random assignment of training and testing trials in cross-validation. **c)** Linear decoder performance, for simultaneously recorded data, after training on correct trials and cross-validating on pairs of correct and error trials matched for condition. Error bars (SEM) reflect the variability that can be attributed to the random selection of units (for populations smaller than the full dataset) and the random assignment of training and testing trials in cross-validation. **d)** A match modulation index, computed for each unit by calculating the mean spike count response to target matches and to distractors, and computing the ratio of the difference and the sum of these two values. Arrow indicates the distribution mean.

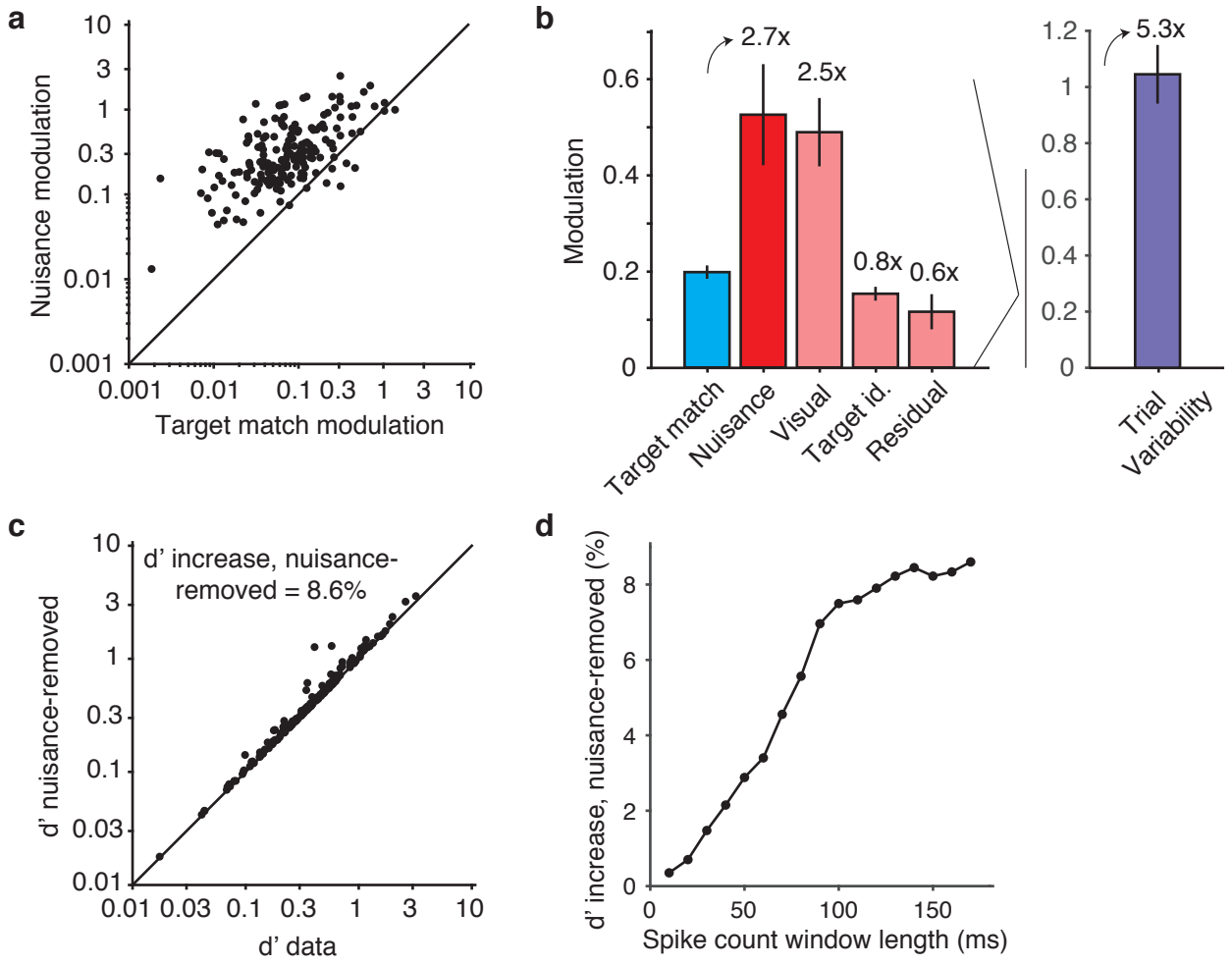

Figure 4. *The impact of nuisance modulation on single-unit d'.* Modulations were computed for each type of experimental parameter, in units of the standard deviations around each unit's grand mean spike count (see Results). **a**) Total nuisance modulation plotted against target match modulation for each unit. **b**) Average modulation magnitudes across units, parsed into target match modulation (cyan), combined nuisance modulation (dark red), and different nuisance modulation subtypes (light red) including visual, target, and residual. The right subpanel indicates the size of trial variability, computed with the same units. Error bars represent standard error across units. Numbers above each type of nuisance modulation indicate its size relative to the target match signal. **c)** Single-unit d' computed on the intact data and with the nuisance-term set to zero. The average proportional impact of nuisance was computed as the average proportional increase in performance when nuisance was removed. **d)** The impact of nuisance modulation on single-unit d' (computed as described in panel c), applied to spike count windows of increasing size.

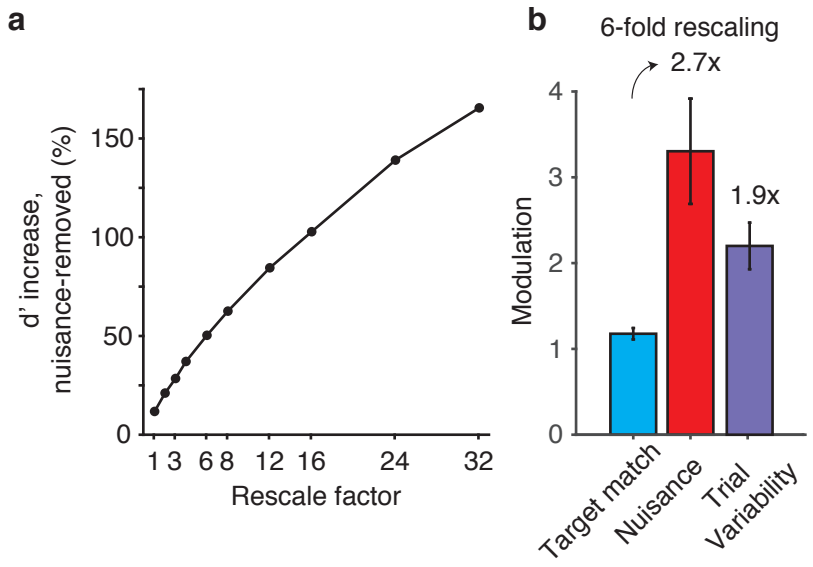

Figure 5. *Nuisance modulation is predicted to be detrimental for higher spike counts.* **a)** The simulated impact of nuisance modulation on single-unit d' as a function of rescaling the spike counts for each unit. **b)** Average modulation magnitudes across simulated units, for the 6-fold spike count rescaling data point in subpanel a.

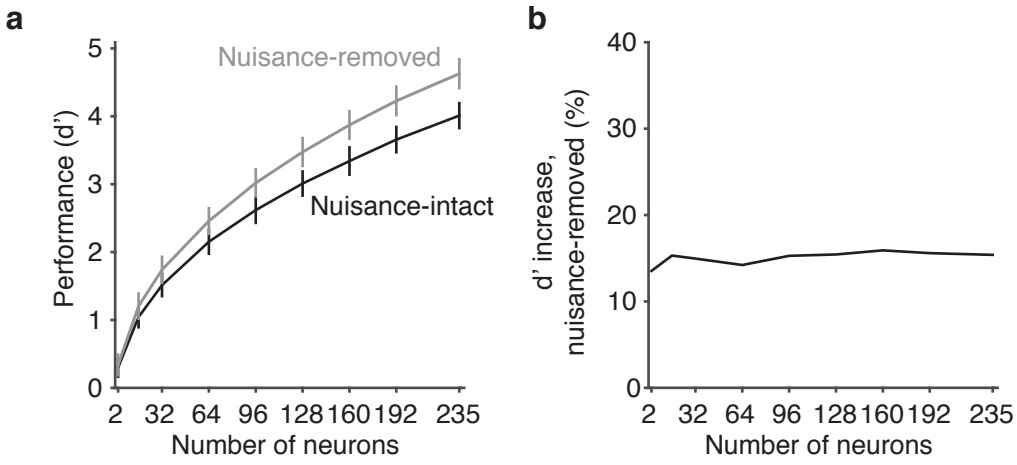

Figure 6. *Estimating the impact of nuisance modulation on population performance.* **a)** Linear decoder performance, shown in units of population d', as a function of population size for two simulated populations: "Nuisance-Intact": a version of our data in which the responses of each unit are replicated (after noise-correction), coupled with independent, Poisson trial variability; "Nuisance-removed": a similar version of our data, but with the nuisance modulations for each unit removed. Error bars (SEM) reflect the variability that can be attributed to the random selection of units (for populations smaller than the full dataset) and the random assignment of training and testing trials in cross-validation. **b)** The proportional impact of nuisance (computed as the proportional increase in performance when nuisance was removed), plotted as a function of population size, computed for the data shown in panel a.

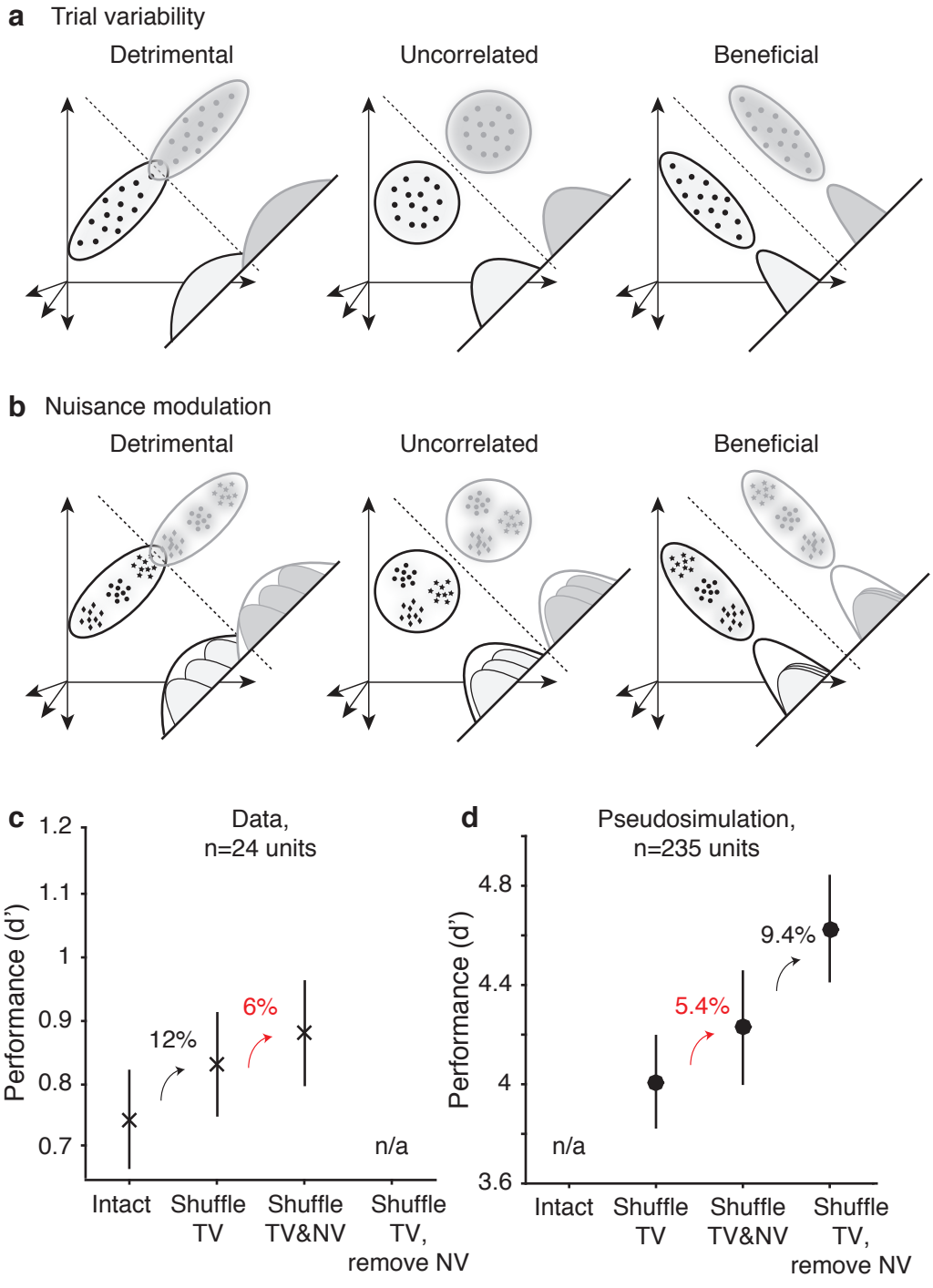

Figure 7. *Understanding how correlated trial variability and correlated nuisance modulations impact task performance.* **a)** Shown are cartoon depictions of the "beneficial" and "detrimental" impact that correlated trial variability can have on task performance relative to the "uncorrelated" benchmark. Each point depicts a hypothetical population response for a population of two neurons on a single trial, and clusters of points depict the dispersion of responses across repeated trials. Dotted lines depict the linear decision boundary optimized for a two-way classification. Population performance is determined by projecting each class onto an axis perpendicular to the decision boundary. Correlated trial variability between units can be configured to increase or decrease the variance of the projected population response relative to benchmark of uncorrelated trial variability, and thus have a detrimental or beneficial impact on performance. **b)** Same as in a, but expanded to incorporate correlated nuisance variability. Included are 3 experimental conditions within each set (clusters of points). Like trial variability, correlated nuisance variability between units can be configured to increase or decrease the variance of the projected population response, relative to benchmark of uncorrelated nuisance variability. **c)** To assess the impact that correlated trial and nuisance variability between units has on population performance, we applied shuffling procedures to the raw data recorded within each session. Shown is linearly decoded population performance (d') for populations of size 24 for: "Intact" – without shuffling; "Shuffle TV" – shuffled trial variability while maintaining nuisance variability correlations intact; and "Shuffle TV&NV" – shuffling both trial and nuisance variability. This analysis cannot be performed in a manner that assess what happens when nuisance variability is removed, thus the placeholder "n/a" for comparison with subpanel d. **d)** The same pseudosimulation data presented in Fig 6 (n = 235). Because that data is simulated as independent between units, the "Intact" condition cannot be computed as indicated by n/a. Shown is "Shuffle TV": linear decoding performance when trial variability is shuffled and nuisance variability is kept intact; "Shuffle TV&NV": performance when trial variability and nuisance variability are shuffled; "Shuffle TV, remove NV": performance when trial variability is shuffled and nuisance variability is removed. In both c and d, numbers above the arrows indicate the proportional increase in d'. Error bars (SEM) reflect the variability that can be attributed to the random assignment of training and testing trials in cross-validation.