NeuroImage 180 (2018) 88–100

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Ghosts in machine learning for cognitive neuroscience: Moving from data to theory

Thomas Carlson [a,b,*,1], Erin Goddard [b,c,1], David M. Kaplan [b,d,e,1], Colin Klein [b,f,1], J. Brendan Ritchie [g,1]

[a] School of Psychology, The University of Sydney, Sydney, NSW, 2006, Australia
[b] ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, Sydney, NSW, 2109, Australia
[c] McGill Vision Research Group, McGill University, Montreal, QC, Canada
[d] Department of Cognitive Science, Macquarie University, Sydney, NSW, 2109, Australia
[e] Perception in Action Research Centre, Macquarie University, Sydney, NSW, 2109, Australia
[f] Department of Philosophy, Macquarie University, Sydney, NSW 2109, Australia
[g] Laboratory of Biological Psychology, KU Leuven, 3000 Leuven, Flemish Brabant, Belgium

ARTICLE INFO

Keywords
Multivariate pattern analysis
Brain decoding
Exploratory methods
fMRI
Magnetoencephalography

ABSTRACT

The application of machine learning methods to neuroimaging data has fundamentally altered the field of cognitive neuroscience. Future progress in understanding brain function using these methods will require addressing a number of key methodological and interpretive challenges. Because these challenges often remain unseen and metaphorically "haunt" our efforts to use these methods to understand the brain, we refer to them as "ghosts". In this paper, we describe three such ghosts, situate them within a more general framework from philosophy of science, and then describe steps to address them. The first ghost arises from difficulties in determining what information machine learning classifiers use for decoding. The second ghost arises from the interplay of experimental design and the structure of information in the brain – that is, our methods embody implicit assumptions about information processing in the brain, and it is often difficult to determine if those assumptions are satisfied. The third ghost emerges from our limited ability to distinguish information that is merely decodable from the brain from information that is represented and used by the brain. Each of the three ghosts place limits on the interpretability of decoding research in cognitive neuroscience. There are no easy solutions, but facing these issues squarely will provide a clearer path to understanding the nature of representation and computation in the human brain.

1. Introduction: data, pattern, theory

Textbooks present scientific confirmation as a matter of fitting theory to data. Savvy philosophers and scientists have long known better. High-level theories do not make direct predictions about data. To borrow a framework from philosophy of science, scientific inference is not a one-step process from data to theory but a *two-step* process from data to *phenomenon* to theory (Bogen and Woodward, 1988; Suppes, 1962). For example, the Standard Model in physics is not tested directly against the voluminous data from particle colliders. Instead, that collider data is processed to give evidence for some stable, replicable phenomenon – Z^0 decay, for example – and then the Standard Model is checked to see if it can account for that phenomenon. Similarly, plate tectonics did not explain magnetometer readings but rather *the spreading of the mid-Atlantic ridge*. General relativity did not explain a series of telescopic observations but the *precession of Mercury*.

So too with various types of data in cognitive neuroscience. What one typically aims to explain is not raw data itself (e.g., changes in BOLD signal), or even a particular set of results from a single experiment. Rather, the goal is arguably to uncover and explain stable and replicable patterns of activation in response to a stimulus or task. It is of only mild interest that inferior temporal (IT) cortex was activated in this or that experiment. It is, however, of great importance that IT cortex is reliably activated by a wide variety of object recognition tasks.

Many early critiques of neuroimaging focused on these two inferential steps as they applied to univariate analyses of brain activation.

* Corresponding author. 318 Griffith-Taylor Bldg., School of Psychology, University of Sydney, Sydney, Australia, 2006.
E-mail address: thomas.carlson@sydney.edu.au (T. Carlson).
[1] All authors contributed equally to this work.

https://doi.org/10.1016/j.neuroimage.2017.08.019
Received 17 February 2017; Received in revised form 17 July 2017; Accepted 4 August 2017
Available online 6 August 2017

Insofar as simple univariate analyses seemed problematic, it was precisely because of weak links in the inference from data to replicable phenomenon (Klein, 2010; Logothetis et al., 2001; Nair, 2005; Poldrack, 2006). At the same time as the weaknesses in univariate analyses were becoming apparent, developments in machine learning techniques were changing the world of science, technology, medicine, and industry (Jordan and Mitchell, 2015). Perhaps unsurprisingly, machine learning methods have also found their way into cognitive neuroscience, most prominently under the banner of multivariate pattern analysis (MVPA) or "brain decoding". Some uses of machine learning in neuroscience directly address practical problems. For example, machine learning methods can be used to decipher patterns in neural data for clinical diagnosis and rehabilitation purposes including brain-machine interfaces (Hatsopoulos and Donoghue, 2009). Such uses are judged solely by their utility, and are otherwise unconstrained in the data and methods they use. We mention these to put them aside. Our focus will be on the application of decoding methods in the pursuit of basic knowledge about brain function.

Machine learning methods have become popular in part because they do not require many of the problematic auxiliary assumptions that plague univariate analyses. Specifically, MVPA arguably does not require strong commitments about the viability of reverse inference (Poldrack, 2006). Nor does MVPA assume a simple relationship between brain activity and the BOLD response (Logothetis et al., 2001), or the specifics of process decomposition (Sternberg, 2011). Further, MVPA allows researchers to deal with extremely large datasets utilising a wide range of techniques including structural MRI, DTI, fMRI, EEG, and MEG. The combination of large datasets and comparatively fewer assumptions gives machine learning methods an air of objectivity: rather than relying on old assumptions about cognitive architecture, we might simply let the brain tell us which categories provide the best fit (Anderson, 2014).

Yet machine learning does not directly connect theory and data any more than univariate analyses. The primary outcome from machine learning analyses is not (we suggest) a direct test of theory but rather evidence concerning stable patterns of brain activity – phenomena, in the above parlance. Such patterns are typically characterised in terms of a neural population's representational space: that is, how activity in the population activity relates both to the world and to other neural representations. The phenomena thus uncovered are what provide a basis for our tests of theories about cognition and brain function.

Machine learning brings with it its own set of problems. Precisely because it offers up simple patterns, it can be easy to read too much into data – to see phenomena that are not really there. This article outlines three of these metaphorical "ghosts" in machine learning techniques, as applied in cognitive neuroscience. The first involves the source of MVPA data itself, and the need to achieve greater specificity about the information we are measuring in the brain. The second involves the move from data to phenomenon, in particular when using dimensionality reduction techniques to go from complex datasets to simple patterns. The third and final challenge comes in moving from phenomenon to theory, and the difference between measuring information in the brain and inferring how the brain might actually use this information. Each of the three ghosts place limits on the interpretability of decoding research in cognitive neuroscience. Although there are no easy solutions, awareness of these issues will provide a clearer path to understanding the nature of representation and computation in the human brain.

Most will be familiar with some of these challenges, and some will be familiar with all of them. Many researchers have expressed related concerns about the interpretation of MVPA decoding results in cognitive neuroscience, as well as offering similar recommendations that this issue must be handled with care (e.g., Davis and Poldrack, 2014; de-Wit et al., 2016; Dubois et al., 2015; Guest and Love, 2017; Haynes, 2015; Poldrack and Farah, 2015; Ritchie et al., in press). One of our goals in this paper is to show that these problems can be fit into a common framework that connects them to ones faced previously by other, more well-established scientific disciplines. This is not an exercise in pessimism, however. We think that by clarifying the different steps of scientific inference and identifying the points at which problems often arise, we can arrive at useful constraints on the design and interpretation of machine learning studies.

Finally, in highlighting several field-specific challenges facing decoding research, we do not mean to imply that other interpretive and inferential issues associated with neuroimaging in general are somehow irrelevant. Importantly, the inferences licensed by decoding methods – like all neuroimaging methods – are limited by the fact that they are inherently correlational (Poldrack, 2011). Consequently, demonstrating significant decoding in a given brain region during task performance cannot by itself establish that it plays a causal role in that performance. Interventions, which include transcranial magnetic stimulation, reversible inactivation, lesions, and optogenetics, provide essential causal information that complements the evidence supplied by decoding studies (Pearl, 1995; Spirtes et al., 2000; Woodward, 2003). Related general critiques of decoding research based on their reliance on reverse inference (e.g., Poldrack, 2006, 2008) may also be germane, but fall outside the scope of this article to address. Importantly, we are squarely focused on internal steps that decoding researchers can take to overcome the field-specific interpretative and inferential challenges described above – without depending on help from other methods.

2. The ghost of source ambiguity

In science, data is the foundation upon which we discover phenomena and test theories. The same is true in cognitive neuroscience. But what exactly is the nature of the data we rely on in decoding research? Although there is consensus that machine learning methods measure information in the brain, it is quite common for there to be uncertainty about the underlying source of this information. The first ghost arises from the gap between our ability to measure information and our capacity to determine the underlying neural source. The former enables us to tell whether, and perhaps even how much, decodable information is present about the stimulus or task condition in a brain representation. Yet only the latter – identifying the neural source of this information – permits the data to act as a foundation for interpretation and brings us closer to the aim of understanding neural representations and processes.

Ascertaining the true neural source of decodable information, however, is extremely difficult because the mere presence of decodable information is ambiguous between potential sources (Bartels et al., 2008; Naselaris and Kay, 2015; Op de Beeck, 2010). To illustrate this, consider a hypothetical scenario from another branch of science. Suppose a simple linear classifier such as Gaussian Naïve Bayes (GNB) is successfully trained to predict whether a hurricane will form based on data from a large array of meteorological sensors. At this stage, we would have learned that information about hurricanes is present in the multivariate data collected from the sensors. Although this result would be useful for all kinds of practical purposes, we would not have appreciably deepened our understanding of hurricanes. At a minimum, if the classifier is to help us understand hurricanes, we would have to determine what information in the sensor data is driving the classification. To do this, one might inspect the classifier weights. Perhaps one would then find that a combination of dew point and humidity drove the classification. Only now would we begin to understand the relationship between these meteorological variables and hurricanes, and thereby add to our knowledge of hurricanes. Moreover, having identified these variables as important factors for hurricanes puts us in the position to study how these factors interact with other variables (e.g. wind speed, atmospheric pressure, etc.), potentially deepening our knowledge of hurricanes still further. The lesson here is that not all data is equal; even useful and predictive data can fail to give us the sort of information we need for advancing

understanding.

2.1. Case study: source ambiguity in orientation decoding

The most rigorous investigation of the link between decodable information and its underlying neural source involves fMRI orientation decoding in human visual cortex. While the orientation decoding debate can be viewed as a success in terms of rigor, it also illustrates the challenges associated with bridging the gap between decodable information and identifying the underlying neural source.

Early seminal decoding studies demonstrated that the orientation of visual gratings could be decoded from the primary visual cortex using BOLD fMRI (Haynes and Rees, 2005; Kamitani and Tong, 2005). We have known that neurons in early visual cortex explicitly represent orientation information for five decades (Hubel and Wiesel, 1968). One could thus view this result as trivial. What launched an enthusiastic decade long debate was "how" orientation information represented in neurons was accessed by the classifier – i.e. bridging decodable information to its neural source. Making this link was non-trivial because the fMRI scanning resolution in these experiments was 3 mm while orientation information in human primary visual cortex is represented in ~0.5 mm wide columns (Yacoub et al., 2008). It was thus unclear how information represented at the columnar scale could be accessible to the classifier. The stimulus was simple, the experiment was straightforward, and there was a wealth of existing knowledge about how orientation information is represented: this early result was an ideal opportunity to link decodable information to its underlying neural source.

However, demonstrating how decodable information about orientation arises from neural activity measured with fMRI proved to be difficult. Why? Ideas about the neural source of decodable information proved difficult to disambiguate (Fig. 1). An initial proposal was that machine learning classifiers could exploit small biases in the proportion of the orientation tuned neurons within individual voxels. This proposal led to the idea that decoding methods confer "hyperacuity" to fMRI (Boynton, 2005; Kamitani and Tong, 2005). Later, the "biased map" account argued that unequal distributions of orientation-tuned neurons across the cortical map (Furmanski and Engel, 2000; Sasaki et al., 2006) create coarse scale biases; and that these biases are sufficient to account for orientation decoding (Freeman et al., 2011, 2013). Most recently, a model that assumed neither fine scale nor coarse scale biases was used to show the edges of grating patterns could generate distortions in the map of cortical activity, providing another potential source of information for orientation decoding (Carlson, 2014).

Ultimately, some combination of these factors (and perhaps others yet to be discovered), rather than a single factor, accounts for how decodable information about orientation arises from neural activity in visual cortex

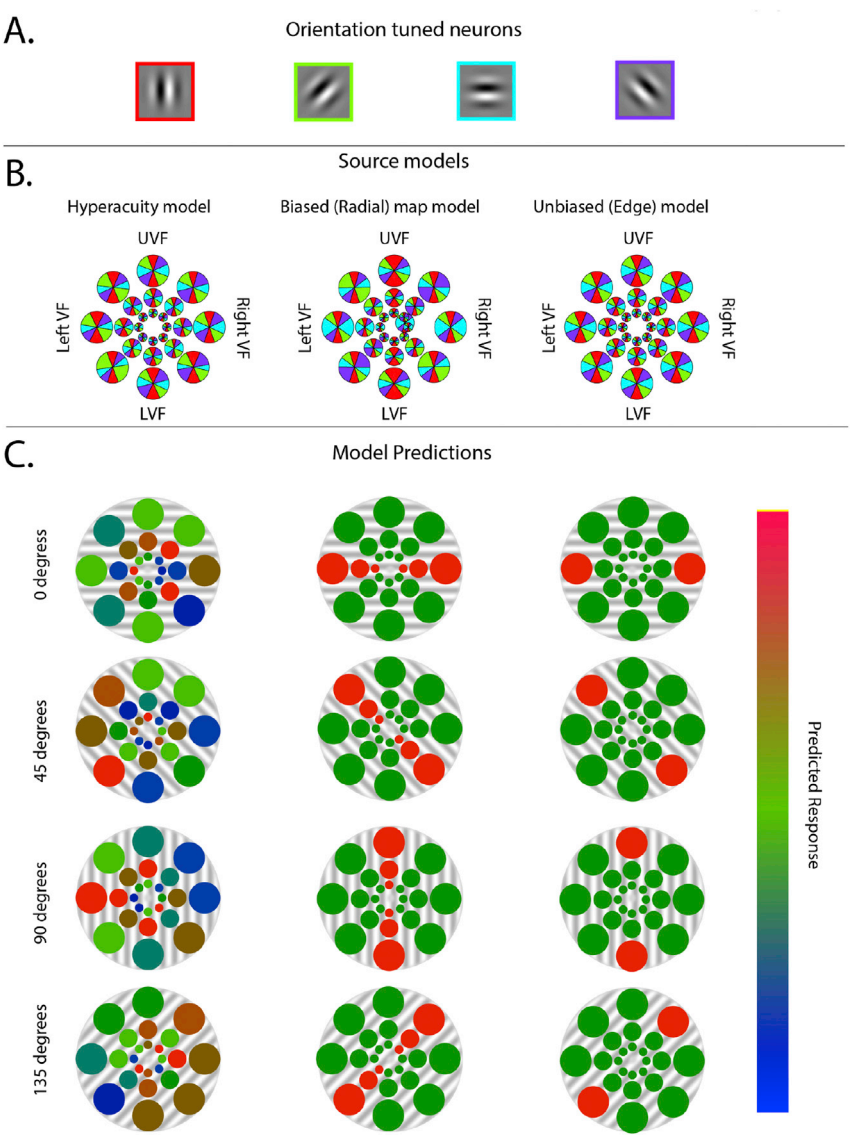

Fig. 1. Orientation decoding source models. **A**: Four groups of hypothetical orientation-tuned neurons with colours indicating the preferred orientation. **B:** Three models about the source of information for orientation decoding each make different assumptions. The pie charts represent fMRI voxels at different locations in the cortical map in visual cortex, where the coloured wedges indicate the proportion of each of the four groups of orientation-tuned neurons in each voxel. The Hyperacuity model assumes that random sampling results in small biases in the proportion of neurons in each voxel. Note the variation in the size of the wedges at different spatial locations. The Biased map model assumes that the distribution of neuron varies systematically across the cortical map. This graphic shows an example of a radial bias, in which there is a larger proportion of neurons pointing towards the fovea. Note the upper visual field (UVF) and lower visual field (LVF) have a greater proportion of vertically tuned neurons; and the left visual field (Left VF) and right visual field (Right VF) have a greater proportion of horizontally tuned neurons. The unbiased model assumes there are no differences in the proportion of neurons across voxels. **C:** The predicted responses of the model voxels in **B** superimposed over four example grating stimuli. The fill colour of each circle indicates the predicted strength of response for that model voxel. The Hyperacuity model predicts variations in each voxel's response that is determined by the random sampling process. Decodable information is assumed to arise from these small biases. The biased map model predicts systematic variations in each voxel's response at each location in the retinotopic map. For the radial bias, the largest response comes from map locations that align with the stimulus orientation. Decodable information is assumed to arise from these map level biases. The unbiased model predicts no variation in the response across voxels, except for voxels at or near the edge of the stimulus (outer ring), where an edge artefact creates a disproportionately large response that mimics the radial bias. This model predicts that the source of decodable information comes from retinotopic map locations corresponding to the edge of the stimulus.

measured with fMRI. Importantly, this debate powerfully demonstrates how identifying the neural source of decodable information can be challenging, even if measuring information in the brain using machine learning methods can be accomplished with relative ease.

2.2. The source of decodable information as a foundation for mechanistic understanding

One might ask, what is really at stake in this debate? Perhaps this is just what vigorous, cutting-edge science looks like. We and others think something more is going on – parties to this debate are grappling with a foundational problem for the field. When decodable information does not have an identified neural source, the scientific conclusions of decoding studies lack a foundation for interpreting subsequent findings. In the orientation decoding debate described above we outlined three candidate sources of decodable information: fine scale biases providing access to orientation information represented in cortical columns, course scale biases accessing differences in the distribution of orientation tuned neurons at the map level, and an edge related stimulus artefact. Decoding studies have also shown that attention enhances the representation of orientation information in visual cortex (Jehee et al., 2011; Kamitani and Tong, 2005) – more specifically attention increases the "decodability" of grating patterns. The ambiguity in the original debate creates a new ambiguity: which candidate source model is attention operating over? We know that stimulus decodability has increased with attention, but how? Is attention enhancing the representation of orientation information in cortical columns? Or is it increasing coarse scale biases at the map level? Or is attention changing biases at the edge representation? Each source provides a different mechanistic explanation of attention's enhancement of the stimulus representation—and ultimately, what we have learned from the finding that attention enhances "decodability".

In the above scenario, we can leverage knowledge from neurophysiology showing that attention influences activity in orientation tuned neurons (Desimone and Duncan, 1995; Kastner and Ungerleider, 2000), thus providing support for the interpretation that attention is operating on orientation information represented in cortical columns. Moreover, in the context of current knowledge, the latter two explanations (attentional enhancement of coarse scale biases and edge representations) can be seen as lacking clear evidential support. Many decoding studies, however, are conducted in research areas where the source of decodable information is either unspecified or unknown (e.g., object recognition, memory and language), and we lack the benefit of foundational knowledge like that we have for early visual cortex. For research in these areas, one needs to be especially vigilant about the distinction between decodable information and the neural source of information driving the classification, as it is the (often unidentified) source that is relevant for constructing mechanistic explanations of brain processes.

2.3. Why is uncovering the neural source so difficult?

As exemplified above, data about decodability are relatively weak on their own. When combined with information about the neural source, however, decodability results can be used as a basis for mechanistic understanding. How then do we get from decodability to a neural source? An efficient way of doing this is to inspect the classifier weights.[2] In the toy hurricane example, this approach would point us back to the sensors driving the classification. From there, details about the sensor type (e.g., whether they were dew point or humidity sensors) would inform us about the source of decodable information. In cognitive neuroscience, our multivariate data (e.g. fMRI voxels) do not come with tidy labels indicating their information-bearing function. Using a voxel's location in the brain, we might be able to infer something broadly about the type of information an fMRI voxel is measuring (e.g., an fMRI voxel in visual cortex is likely measuring activity related to encoding a visual stimulus), but this is rarely sufficient to identify the source of decodable information.

An alternative means of tracing decodable information back to its source in cognitive neuroscience is to leverage theory to generate testable predictions about the spatial distribution of decodable information across recording sites (c.f. Carlson et al., 2003; Haxby et al., 2001). Orientation decoding again provides an illustrative example. Indeed, it is a best-case scenario since not just one but two of the theoretical accounts of the neural source make predictions about the spatial distribution of information (Fig. 1), thus enabling these accounts to be tested explicitly (Carlson, 2014; Freeman et al., 2011, 2013; Wardle et al., 2017). The biased map account precisely predicts that the magnitude of the response across locations in the cortical map will covary with the orientation of the stimulus; and the edge account further predicts this activity will be localised to the edge of the stimulus. The hyperacuity account, in contrast, is based on the assumption that cortical columns in fMRI voxels will be irregularly sampled, resulting in a random spatial distribution of classifier weights across the cortical map. This prediction was qualitatively tested in the initial report of orientation decoding, and the weights indeed appeared random (Kamitani and Tong, 2005). Subsequent research, however, showed that coarse scale biases do contribute to the decodability of the stimulus orientation (Freeman et al., 2011, 2013), calling into question whether the "randomness" test can in fact be interpreted as positive evidence. Furthermore, accounts lacking spatial predictions exhibit a degree of freedom that allows them be superimposed on top of a broad range of empirical results. This makes such accounts difficult to disprove. For example, if the spatial distribution of information were found to align with the biased map account's predictions (Freeman et al., 2011, 2013), one could still argue that fine scale sampling biases provide an additional source of decodable information that goes unseen in the "noise" in the classifier weights.

Note that even though the lack of a spatial prediction is a shortcoming of the hyperacuity account, this does not invalidate the account. In fact, encoding model-based analyses such as voxel-wise modelling and partial receptive field mapping can be viewed as successful extensions of the idea that fMRI voxels contain biases in their proportions of neurons tuned to different orientations (Dumoulin and Wandell, 2008; Kay et al., 2008). Nevertheless, most hypotheses aiming to relate decodable information to a neural source will be "shapeless" in the sense that they do not predict how information will be distributed spatially across recording sites. As such, the approach of using the spatial distribution of decodable information can only be used to adjudicate candidate sources in (relatively rare) cases where the alternative accounts predict spatially organised patterns of information across voxels.

2.4. Identifying a source versus assessing the contribution of multiple sources

Decoding methods leverage any and all information to make accurate classifications. It is therefore a plausible scenario that multiple candidate sources might simultaneously be contributing to the decodability of a given stimulus or condition. Under this scenario, a distinction needs to be drawn between whether or not some source of information contributes to decoding and if so, what the source's contribution is in the context of multiple other sources. This distinction is essential, as it affects the strength of the mechanistic explanations of brain processes that we aim to draw from the data. We again use orientation decoding as an example, except this time focusing on orientation decoding based on magnetoencephalography (MEG).

If the aim is determining whether a particular source of information contributes to successful decoding, one widely accepted approach involves using control experiments and/or control analyses to rule out other sources. One recent study employed this approach to test whether orientation information represented in cortical columns could be

[2] See Haufe et al. (2014) for discussion of practical limitations of this approach for neuroscience applications.

resolved with MEG (Cichy et al., 2015). The authors employed multiple controls to carefully rule out alternative sources including stimulus edges and global biases, and even provided a model demonstrating how orientation columns could be resolved with MEG using MVPA. This study makes a compelling case that orientation tuning in cortical columns is one source of information for orientation decoding in MEG (Cichy et al., 2015; Stokes et al., 2015).

The controls approach is an effective one for determining whether a source is contributing to decoding. However, if the aim is to leverage a source as a foundation for mechanistic interpretation and understanding (see section 2.2), it is necessary to examine the relative contribution of different sources. One approach to this involves examining multiple source models in a competitive context. Following a protocol used previously by Alink et al. (2013) to study hyperacuity in fMRI, Wardle et al. (2016) performed such a study using the representational similarity analysis (RSA) framework (Kriegeskorte et al., 2008a) with MEG. The study tested a wide range of stimuli, including several designed to elucidate the source of decodable information for orientation decoding in MEG. This study also found that grating orientation could be decoded with appropriate controls in place, supporting the previous study's findings (Cichy et al., 2015). The competitive model testing, however, provides a more complete picture of the contribution of different sources. Wardle et al. (2016) showed that multiple models accounted for the decodability of the stimuli indicating that multiple sources were contributing. Furthermore, the orientation model was among the weakest, and even a richer model of primary visual cortex (the first layer of HMAX; Riesenhuber and Poggio, 1999; Serre et al., 2007) provided only a modest account. If orientation information represented in cortical columns were a key source of information for orientation decoding in MEG, we would have expected these models to perform better. Thus, while this study confirmed orientation information represented in cortical columns is one source, it is not a major source of decodable information.

These early studies in orientation decoding in MEG highlight the need to assess the relative contribution of different sources, when multiple candidate sources have been identified. This is especially important when trying to use decodability results as a foundation for mechanistic understanding. How might the result that attention increases grating stimulus *decodability* in the context of MEG be viewed differently? Given that orientation information represented in cortical columns has been found to be small source relative to others, would we still be confident that we are measuring attentional enhancement of orientation information represented in cortical columns?

2.5. Moving beyond the first ghost

This section has focused on a specific instance of a more general problem: when it comes to the hunt for stable phenomena, not all data are created equal. Decodability is first and foremost a metric of classifier performance. Using decoding results as data is unproblematic for applied uses of machine learning. But for studies aiming to tell us something about the brain, the decodability metric leaves the source of decodable information either unspecified or ambiguous. This makes it difficult to move beyond claims about data to claims about stable, repeatable *neural* phenomena. And the latter is what cognitive neuroscientists ultimately require when they seek to test theories.

We emphasise that this is not an insurmountable problem. Rather, our goal has been to suggest that it can take serious and sustained investigation to resolve questions about neural sources. Minimally, those using these methods to understand the brain should continue to internally query "what is the source driving decodability?", in order to move closer to the goal of providing mechanistic explanations.

3. The ghost of perceived neutrality

In the previous section, we discussed challenges associated with data. The next potential point of friction is the move from data to a stable, replicable phenomenon. Recall again that theories are primarily tested against phenomena, rather than individual bits of data gathered by experimenters. Sometimes the move from data to phenomenon is relatively straightforward: a regression analysis shows a clear linear trend, or a *t*-test shows an obvious difference between two groups. But many interesting phenomena are more complex – and brain phenomena are almost certainly to be in this category – and the move from data to phenomenon is consequently more fraught.

A natural way of thinking about brain representations – and one consonant with the presumed population coding used by the brain – is to model each brain region's activity as a high-dimensional state space. The information represented by a region can be represented as points or regions within this space (DiCarlo and Cox, 2007). Exemplar-based decoding approaches characterise these representational spaces by modelling the evoked activity of individual stimulus exemplars as points in the representational space (Kriegeskorte and Kievit, 2013; Kriegeskorte et al., 2008a). The geometric configuration of exemplars within the space can then inform us about the functional significance of the representation. For example, if animate and inanimate object exemplars form separable clusters in the space, the representation could be used by the brain to discriminate these two categories of objects (Kiani et al., 2007; Kriegeskorte et al., 2008b). This state space representation framework exemplifies the transition from data to phenomenon. The representational space is the phenomenon that we aim to understand and test against our theories. For example, is animacy coding a basic organizational principle of inferior temporal cortex? To (re)construct a representational space, we use data. The representational similarity analysis (RSA) framework (Kriegeskorte and Kievit, 2013; Kriegeskorte et al., 2008a), for example, uses the pairwise distance between the evoked response to different exemplars (i.e., the "data") as a proxy for distance between exemplars representation the brain's representational space (i.e., the phenomenon). Importantly, this movement from data to phenomenon embodies assumptions that could potentially warp the space away from its true nature. In this example, the choice of a distance metric that is applied to the data (correlation distance, decodability, etc.) will invariably affect the reconstruction.

There are also interpretive challenges. The high-dimensional nature of state spaces used to model brain representations are often beyond the human visual system's capacity for interpretation (typically limited to 3 or 4 dimensions), so it is difficult to grasp their intrinsic structure. One can leverage model testing to investigate their structure, e.g., using representational similarity analysis (Kriegeskorte and Kievit, 2013; Kriegeskorte et al., 2008a); however, in mainstream applications this yields only a correlation metric that summarizes the relationship between two (very) complex spaces.

Another approach to understanding these representational spaces is to use data-driven methods such as multidimensional scaling (MDS) to reduce the dimensionality of the spaces, and thereby aid human interpretation and comprehension. These techniques are most often used for visualising large datasets, frequently in conjunction with model testing. In such uses, they provide a valuable source of intuition about the underlying structure. However, some make a further inference: that the extracted dimensions reveal intrinsic dimensions of the neural population response itself. For example, Kriegeskorte et al. (2008b) used fMRI in human inferior temporal cortex (IT) and serial single-electrode recordings in monkey IT to study responses to real world objects. Using MDS and hierarchical clustering they found that responses to objects of the same animacy category were grouped together, in both human and monkey IT. They conclude that animacy is a relevant stimulus feature for area IT, and argue that MDS "can reveal the properties that dominate the representation of our stimuli in the population code without any prior hypotheses". Kiani et al. (2007), looking at the same monkey IT recordings later used by Kriegeskorte, report a similar structure in monkey IT when using a stimulus set of >1000 objects. This has led to the view that animacy is an important feature dimension in IT. Similar approaches

have been used to argue that representations of objects in monkey IT have around 100 dimensions (Lehky et al., 2014). Dimensionality reduction has also been used as a tool to investigate a diversity of other brain functions including working memory (Machens et al., 2010), speech production (Bouchard et al., 2013) and semantic knowledge (Huth et al., 2016; Zinszer et al., 2016).

These data-driven approaches have been lauded as potential solutions to the problem of "conceptual baggage … biasing the space of hypotheses that we consider" when we try to characterise neural responses in higher-order visual cortical areas, and further that these methods help to "circumvent these biases by searching for structure in the functional responses … in a hypothesis-neutral fashion" (Kanwisher, 2010). This is a noble goal, but the practice is far from perfect. In particular, even apparently unsupervised data-driven analyses require substantial experimenter input in the step from data to phenomenon. This reflects a deeper problem: even the move from data to phenomenon cannot be entirely hypothesis-free (Hanson, 1958). In the following sections, we further illustrate the assumptions and interpretative issues made in moving from data to phenomenon when using data-driven analyses to understand these representational spaces.

3.1. Data-driven analyses: hypothesis-neutral?

Completely understanding the function of a brain region requires an understanding of how it represents information. As noted previously, one way to conceptualize brain representations is as a high-dimensional state space with individual stimulus exemplars occupying distinct points in the space. Constructing these spaces and interpreting them is far from being hypothesis-neutral. As noted previously, one example of this is choosing a proxy for distance in the transition from data to phenomenon. However, even before we start to analyse data, choices have already been made that will affect the outcome. The act of stimulus selection introduces implicit assumptions both about the stimuli and the underlying representational space. As we cannot measure neural responses to every possible stimulus, we must determine what seem like the most relevant dimensions along which our stimuli will vary. This is a vital yet highly nontrivial task, as the dimensions which seem important to us may not be the ones which are important to the brain. This is particularly relevant for those studying brain responses to complex naturalistic stimuli, where the most appropriate stimulus dimensions can be especially unclear.

Identifying meaningful structure in the results of data-driven analyses is also non-trivial, and is susceptible to the researcher's preconceptions of "sensible" structure. *Pareidolia* is the tendency to see shapes and patterns where there are none. For example, we readily see shapes in clouds, Jesus on burnt toast, or even the Virgin Mary in an MRI scan (Hannan, 2016; Voss et al., 2012). There is no reason to think that scientists are immune to this sort of error: witness the ironic history of the Rorschach test, in which many clinicians were willing to see elaborate diagnostic patterns in patient responses to inkblots (Wood et al., 2003). Pareidolia is arguably only a risk when one fails to appreciate it as a real possibility. Most of the time, good scientists do. Yet, visualisations of data-driven analyses can invite misinterpretation precisely because they can make it seem like no interpretation is necessary. Further, these visualisations are often taken to be intuitive evidence in their own right, which means they be given a false impression of the empirical power of the work itself (Weisberg et al., 2008).

3.2. The (failed) application of data-driven exploratory analyses in a well understood system

We next illustrate some these conceptual points using an example of multi-electrode recordings from marmoset area middle-temporal area (MT). This full details of this illustrative analysis are covered in another paper in this issue (Goddard et al., 2018). Area MT has been extensively studied, and contains a high proportion of cells that are selective for motion direction and speed (Albright, 1984; Britten et al., 1996; Maunsell and Van Essen, 1983; Newsome et al., 1989; Salzman et al., 1990). Moreover, direction and speed are dimensions of visual motion that correspond to perfectly objective physical magnitudes and are behaviourally relevant to these organisms. For these reasons, area MT is generally accepted as playing a key role in visual motion perception. We reasoned that if data-driven techniques are able to extract complex dimensions like "animacy" from higher visual areas, then extraction of these simple stimulus dimensions from MT ought to be straightforward.

The truth, however, is more complex. We analysed multi-unit activity from area MT in anaesthetised marmoset monkeys, while moving dot patterns were presented which systematically varied in speed and direction. We used a linear classifier (LDA) to discriminate between all pairwise stimulus combinations within each dataset. Even with very short time bins, classification performance was high, averaging up to 93% for some stimulus pairs. As expected, classifier accuracy decreased when the stimuli were more similar in direction and/or speed. This pattern of results is consistent with the existing literature on area MT, namely that it encodes both motion speed and direction.

To test the utility of dimensionality reduction approaches for "discovering" important feature dimensions in the neural code we applied these methods to the complete pattern of classifier performance. If the dimension reduction approaches are extracting the most meaningful dimensions for understanding the population response, most of the variability in the data should be captured by just 2 dimensions – a direction dimension that orders stimuli by direction but not speed and a speed dimension that orders stimuli by speed but groups stimuli of the same direction.

Fig. 2 shows the dimensions extracted by MDS. There is no single dimension that clearly maps onto direction or speed. While there certainly is structure and ordering of the responses by stimulus feature, it is not obvious that one could discover that direction and speed were important feature dimensions if this were not already known.

Imagine a researcher naive to the original stimuli dimensions attempting to decipher – on the basis of Fig. 2 alone – what MT is doing. The plots do not lack structure; a careless researcher could hypothesise endlessly about the Rorschach-like patterns and what they mean. As a related example, consider area "OIC R3" in Huth et al.'s (2016) semantic map of the brain, which collects together responses to words about philosophy, science, religion, and spirituality. This either reveals a deep truth or a failure of the method. Without further investigation, one's position would represent only preconceived ideas. Conversely, the plots in Fig. 2 do not lack mystery: a careful researcher could dismiss them as noise. The dismissive researcher would be wrong, and the ambitious researcher would be unlikely – based on such plots alone – to discover an accurate picture of MT's representational space.

At this point, the proponent of data-driven methods faces an uncomfortable dilemma. On the one hand, it could be that direction and speed really are underlying dimensions of neural activity in MT, and that dimensionality-reduction techniques simply cannot extract them (or, more cautiously, the techniques we used cannot necessarily do so). But we have used standard techniques in pedestrian ways: if they fail in this relatively easy case, we ought to be more suspicious of their use in complicated ones.

On the other hand, it could be that the underlying representational space of MT is in fact different than the tested stimulus space. This is not an unreasonable assumption: there is no *a priori* reason why MT could not represent distance and speed by extracting some functions of both. If so, our methods would in fact be accurately extracting the underlying representational structure. Yet in any particular case, this is a difficult line to tread. First, the results strongly depend on which dimensionality reduction approach one uses and which dimensions one plots (Goddard et al., 2018).

Second, the choice of dimensionality reduction approach (PCA, MDS, etc.) makes assumptions about the structure of information in the representation. Hierarchical clustering algorithms, for example, *assume* that there is meaningful clustering in the data, and can return a hierarchical

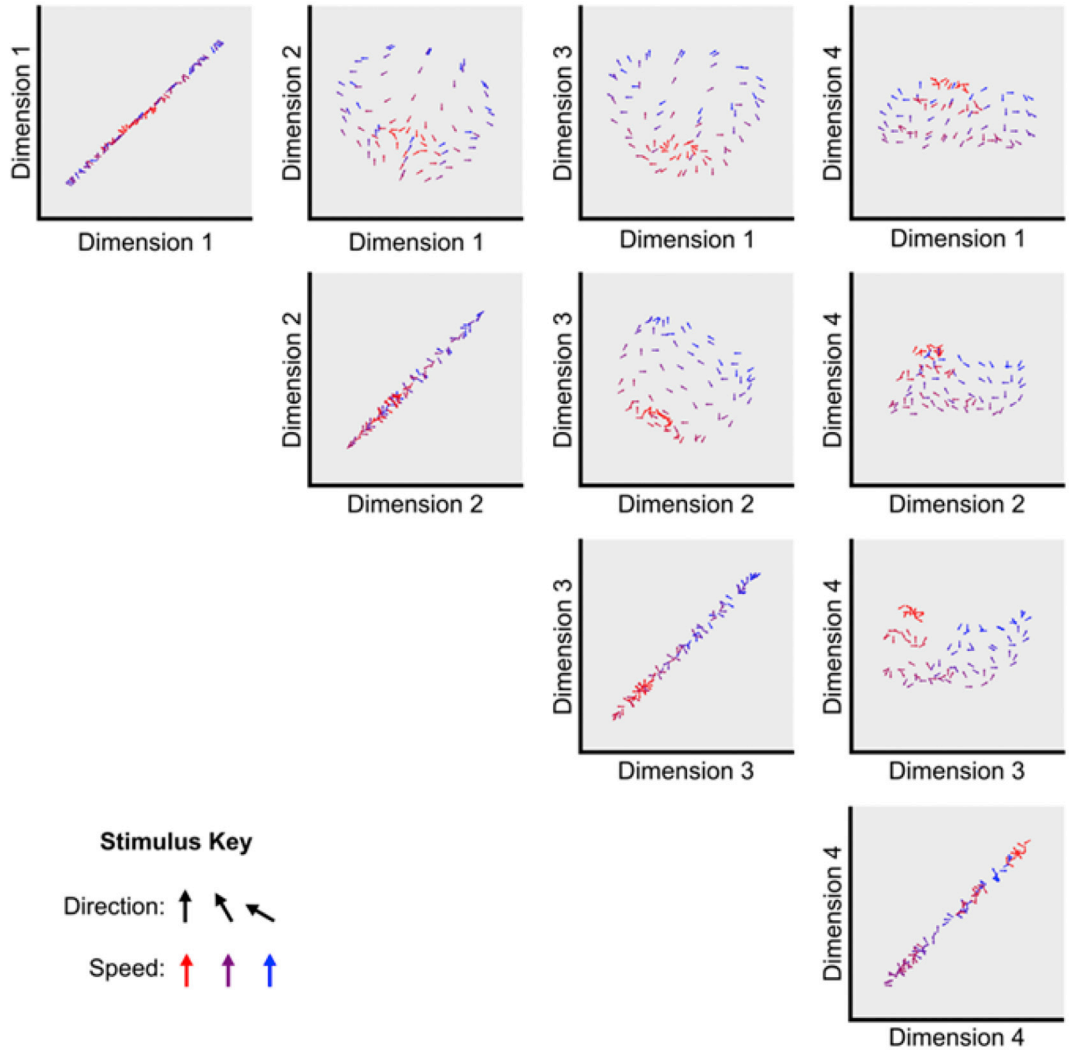

Fig. 2. Summary of dimension reduction by multi-dimensional scaling (MDS) of the classifier performance discriminating multi-unit responses to moving dot fields (84 unique stimuli, of 12 directions and 7 speeds): Data for the 84 unique stimuli projected into spaces defined either by a single dimension or a pair of dimensions from the MDS solution with 4 dimensions. Each moving dot field stimulus is defined by an arrow, where the direction and speed of the stimulus are given by the direction and colour of the arrow respectively (blue = slowest speed, red = fastest speed).

solution even if the target data does not actually form a meaningful hierarchy. Speed and direction of motion are plausibly continuous dimensions: hierarchical clustering on these would merely lead one astray. Or, similarly, k-means clustering requires the experimenter to choose the number of clusters, and will dutifully return the requested number even from a completely homogenous set of noise. Again, failure to realise this could easily lead an experimenter to think they had discovered structure where there was none. On a related note, dimensionality reduction mechanisms often embody basic assumptions (such as linearity of response), which are hard to justify if we think that the stimulus and the way the stimulus is represented might differ in important ways.

Third and finally, these methods require that the experimenter can distinguish real but unexpected dimensions from mere noise or failure of the method. Absent some further story, this means that the researcher must do serious interpretive work on the extracted dimensions. But avoiding the need for such interpretation is lauded as a strength of moving to data-driven methods.

To put the last point a slightly different way, motion and direction were picked as stimulus dimensions in part because they are salient dimensions to us as perceivers. MT clearly supports that salience, but there is no reason why MT must support this by having dimensions which correspond to the perceptually salient ones. A basic principle of psychological research is that one must sample stimulus space appropriately. That is, the stimuli one chooses must be either a systematic or a fully random sample of the full stimulus space (see Judd et al., 2012 for a recent review). But if we do not know the underlying representational dimensions of MT, then we should not have confidence that our intuitive division of stimuli was an unbiased sample of possible stimuli. And if it were a biased sample, dimensionality reduction may tell us more about our own biases in choosing stimuli than it does about the brain itself.

3.3. Dimensionality reduction and exploratory analyses

Where does this leave data-driven approaches as a means of helping us to understand representational spaces? We reiterate that our primary target has been the use of such methods to do pure data-driven extraction of underlying representational dimensions of a brain region. We have been suspicious of such uses, because it is unlikely that experimenters adequately sample the available representational space of the underlying brain region, and because the extracted dimensions are not guaranteed to be intelligible.

That said, two other uses of dimensionality reduction may be more defensible (and these are often confused with data-driven extraction). These uses are considered in greater detail in Goddard et al., 2018. On

the one hand, dimensionality reduction may be a useful visual aid to exploratory analysis. Exploratory analysis seeks to find hypotheses that are worth spending time and resources on exploring. As a hypothesis-generating procedure, exploratory analysis does not necessarily raise the likelihood that the resulting hypotheses will be true or even good, but it may be a valuable first step in understanding an otherwise complex system like the brain.

On the other hand, dimensionality reduction can be used in a limited way to test hypotheses, and hypothesis-driven approaches may be more defensible, so long as they are carefully constrained. By the logic we have just sketched, the results reported by Kriegeskorte et al. (2008) do not provide strong evidence that animacy is an intrinsic dimension of IT representational space. On the other hand, if what you wanted to know is where in the brain has a representational space (with unknown dimensions) with enough information to distinguish animate from inanimate objects, then the same results suggest that IT is a prime candidate. More generally, dimensionality reduction might be useful for generating models that are then tested with new data (Brodersen et al., 2011), although one still has the problem of moving from extracted dimensions to models. The key feature to realise is that an area may carry information about a perceptually or conceptually salient dimension of variation without treating it as a *neural* dimension of variation. To get to the latter, we will probably have to look elsewhere.

3.4. Moving beyond the second ghost

The move to phenomena is challenging as it has embedded assumptions. We think that most researchers are well aware of the dangers of finding patterns where there are none. Many statistical methods are invoked precisely to guard against this. When specific interpretive dangers are made salient, good researchers recognize and avoid them. Yet pareidolia threatens most dangerously when it is subtle – when assumptions are baked into methods which present themselves as assumption-free. Many researchers have an intimate understanding of the tools they used, and so this was arguably less of a worry. Yet as many decoding methods are being built into standard toolboxes, new users can inadvertently load a host of assumptions at the click of a button. All the more reason to remain vigilant at the step from data to phenomenon.

4. The ghost of underconstrained representational interpretation

Phenomena like stable feature spaces are interesting because they bear on theories about how cognitive systems represent the world. Neural decoding, and related methods, have been promoted as a means of investigating both the content (Haynes, 2015; Haynes and Rees, 2006; Norman et al., 2006; O'Toole et al., 2007) and structure (Haxby et al., 2014; Kriegeskorte and Kievit, 2013) of neural representations. An equally important and related question concerns precisely what information is actually *represented* and *used* by the brain for downstream processing and ultimately to guide behaviour, and whether this can be recovered reliably using these methods. This is a move from a stable phenomenon – a feature space – to a theory about how the brain actually works (for more detail on the distinction, see Goddard et al., 2018). As with the previous two steps, the transition from phenomenon to theory can look easier than it actually is.

Does the presence of decodable information in a given brain region provide compelling evidence that this information is represented and used by the brain? We think the question is rarely asked, at least in this fashion. Most decoding papers focus on the hard work of data collection and inference to phenomenon. Questions about neural representations – about where and why and how information is actually used in the brain – require a final inferential step from phenomenon to theory. Conclusions here, if and when they are drawn at all, are often more tentative. Increasingly, researchers are coming to appreciate that this inference is insufficiently constrained and can lead to erroneous claims about the information that is represented and used at the neural level (e.g., Haynes, 2015). We explore this as a way of showing the problems that can affect inference from a reliable phenomenon to a further theoretical claim.

4.1. Neural read-out as a constraint on interpretation

At first glance, this sounds like a highly specific interpretive issue that pertains to neuroimaging research in which MVPA decoding tools are employed. Yet, if one scratches just below the surface, it becomes clear that the problem of interpreting MVPA decoding results is neither particularly new nor special. Instead, it relates to a more general – and more basic – problem that neuroscientists must contend with whenever interpreting neural data in representational or information-processing terms (Eliasmith and Anderson, 2002; Piccinini and Shagrir, 2014). For example, this same issue crops up in earlier debates in neurophysiology about the nature of the neural code (Rieke, 1997). Although the focal issue there was whether the brain might also represent information using temporal codes rather than simply a rate code (deCharms and Zador, 2000), the deeper methodological issue concerns how to justify and effectively constrain our representational gloss on neural systems.

In this earlier context, Rieke (1997) challenged the deep-seated conviction in neurophysiology that the trial-averaged spiking activity of individual neurons – the firing rate – is the vehicle for "what the neuron represents". They argue this assumption is misguided because, strictly speaking, neurons never receive an average firing rate as input or transmit it downstream as output. Instead of assuming that information is carried by firing rates, they argue that the field should focus on deciphering how information is carried by individual spike trains (which can and do serve as input and output signals between cells). From this perspective, average firing rates take on a secondary or derivative importance because they are merely summaries of aggregates of individual spike trains. As such, they might be entirely inconsequential to how neurons carry and transmit information, in the same way that an average of a thousand telephone calls would tell you very little about speech and communication.

Critically, Rieke (1997) describe this shift in perspective as "taking the organism's point of view" because it involves restricting interpretation to signals that are functionally available to, and therefore exploitable by, the system itself. If a given neural response is interpreted as carrying a certain piece of information, such as the orientation of a visual stimulus, then this information must either be used or potentially available for use by "downstream" neurons to justify the claim that the neural signal "represents" stimulus orientation. The lesson for cognitive neuroscience is that the information represented and used by the system itself – the brain's intrinsic decoding procedure – may differ from the information that is extractable in principle, for example, by an arbitrary decoding method.

The connection between representation and downstream use has not gone unnoticed by cognitive scientists (e.g., Kirsh, 1990; Marr, 1982). For example, some attempts along these lines have been made to define the precise conditions under which information is represented in computational systems (Kirsh, 1990, 2003). As a simple illustration, consider a situation in which files or data are no longer accessible in a computer because of a bad read/write head in the hard drive that prevents it from initialising. In principle, this information is still encoded or stored on the drive in the sense that there are methods by which this information may be recovered. Yet, in an important practical sense, the stored information remains inaccessible to the system itself. Without a functioning hard drive, no internal computations can be performed on these data, even if there are external methods by which this information can be recovered. If the information cannot be used, then what justification is there for saying the information is internally represented in the computer?

This distinction between information that is usable by the system itself versus information that is merely recoverable by some arbitrary method is precisely the distinction that is needed to clarify why decoding can sometimes generate problems for inferences about neural

representation. The worry is that our decoding methods – like external data recovery methods – may reveal information that is simply unavailable for uptake by the brain.

"Downstream use" suggests itself as a promising constraint on representation in neural systems (Eliasmith and Anderson, 2002). Under this scheme, to claim that decodable information is represented in a given brain area, it must be used or usable by downstream computational processes in the brain (*neural read-out*) or reflected in the generation of output behaviour (*behavioural read-out*) (Fig. 3). If the brain cannot extract and carry out transformations on this information, or otherwise put it to use, there seems little reason to claim it is neurally "represented".

Despite useful insights about the connection between downstream use and representation from other fields, there is currently no consensus on how to incorporate these ideas into cognitive neuroscience when interpreting the presence of information in the brain revealed through decoding.

4.2. The biological plausibility of MVPA and neural read-out

When applying these concerns to cognitive neuroscience, and particularly the information measured with MVPA, the details differ but availability for use by the brain remains an important constraint.

A working assumption in cognitive neuroscience is that the brain utilises population codes: neural representations are thought to be encoded by patterns of activity distributed over populations of neurons varying from a small number of neurons to in principle the entire brain (Pouget et al., 2000). Under the assumption that the brain's own decoding procedures rely on a linear combination of responses from the different units in the neural population and that neuroimaging techniques coarsely detect these distributed encoding patterns, then linear classifiers have been considered a reasonable proxy for the decoding procedures utilised by the brain (DiCarlo and Cox, 2007; Kamitani and Tong, 2005; Misaki et al., 2010; Pouget et al., 2000; Yamins and DiCarlo, 2016). Thus, the biological plausibility of linear classifiers allows them to serve as a stand-in for the "read-out" processes occurring in the brain (de-Wit et al., 2016).

The biological plausibility of linear classifiers is often held to justify not only conclusions about what information is present in a given brain region, but also to suggest that this information is represented in a format that is functionally usable, if not actively used, by downstream processes. Critically, the inference does not strictly follow, but does provide defeasible evidence that this information is represented in the brain (Cox and Savoy, 2003; Kriegeskorte and Bandettini, 2007). Thus, the biological plausibility of classifiers is supposed to help constrain our interpretation of decoding results.

Even with such hedging, however, it remains unclear whether such inferences are warranted. Results showing that not all decodable information can be read out in behaviour suggests that MVPA provides weak evidence for representation as such (Williams et al., 2007), as do results showing that coding in some discrete regions cannot be read out with linear classifiers (Dubois et al., 2015). These findings challenge the idea that the biological plausibility of classifiers places sufficient constraints on how we interpret decoding findings. Central to the biological plausibility argument is the idea that linear read-out is similar to the operations performed by the brain itself, unlike nonlinear methods that may be overpowered compared to the operations used by the brain (Kamitani and Tong, 2005; Misaki et al., 2010). It is frequently assumed that linear read-out provides the right model, but as a hypothesis it largely remains unexplored (Yamins and DiCarlo, 2016). Ritchie et al. (in press) argue that linear methods can also be overpowered. In this respect, there is little basis for biological plausibility providing an adequate constraint on the representational interpretation of MVPA results.

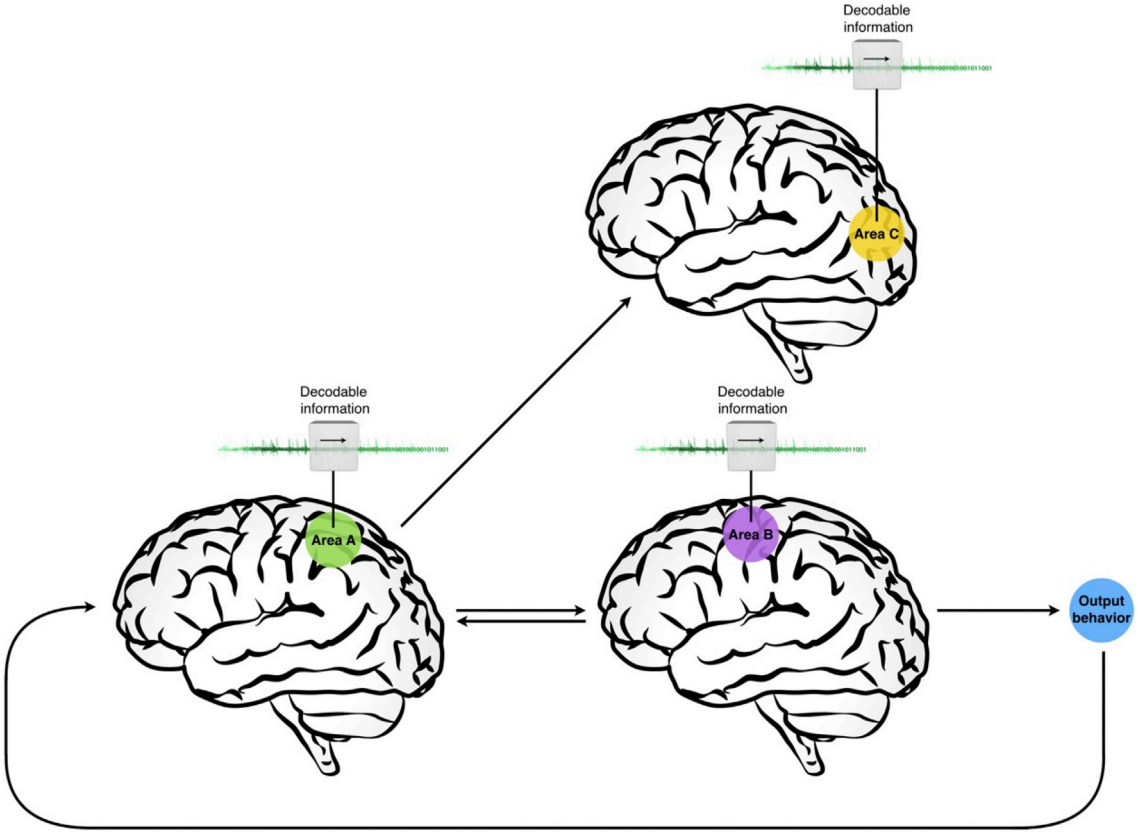

Fig. 3. Downstream use as a constraint on neural representation. Decodable information is represented in a brain area (A), if it is used or usable by downstream areas (B and/or C). Or decodable information is represented in a brain area (B), if it is directly used in the generation of output behaviour. Revealing decodable information in a brain area (C), which is neither used by downstream areas nor reflected in behaviour, provides weak support for claims about neural representation.

Similar issues also arise with respect to model-based MVPA techniques. Representational Similarity Analysis (RSA) is sometimes billed as a technique for investigating the structure of neural representations, by taking a geometric focus (Kriegeskorte et al., 2008a). Thus, by reconstructing the dissimilarity structure from pattern responses, and comparing them to different model dissimilarities, one can estimate how well the relations between the pattern responses are captured by the models. More advanced voxel-wise predictive models try to anticipate the pattern response for stimuli. These techniques have the virtue of directly connecting pattern responses to different hypotheses about how a representation might be structured, but by themselves do not suffice to provide evidence that the structure is exploited (Naselaris and Kay, 2015; Huth et al., 2012).

Another consideration is the spatial scale of "readout". Decodable information can be found throughout the brain, particularly if one relaxes the constraint on how information is read out. However, if one is suggesting that the brain as a whole or some particular brain region actually uses this information, this implies that the readout mechanism has access to all this information. We ourselves are guilty of this error in a recent study (Carlson et al., 2014). In examining a possible readout mechanism, we showed reaction times for object categorisation could be predicted from the structure of object category information decoded from the entirety of inferior temporal (IT) cortex. In interpreting our results, we assumed – without direct evidence – that a readout mechanism existed which could monitor activity over such a large swath of cortex. The important point is that if one aims to argue that the brain represents (and uses) information revealed through decoding, evidence for a readout mechanism capable of operating at approximately the same spatial scale is essential. Without such evidence, it remains unclear whether such a code has any representational significance at all.

4.3. Psychological plausibility as a constraint on interpretation

The most common proposal to reign in the representational interpretation of MVPA findings is via connection to stimulus-directed behaviour. The thinking goes that since observer behaviour is also a downstream effect of neural representations of the stimulus, then showing a connection between some behavioural measure and the decodability of a stimulus should provide stronger evidence that the information uncovered with MVPA is in fact used by some downstream process in a task-related manner (de-Wit et al., 2016; Naselaris et al., 2011; Tong and Pratte, 2012; Williams et al., 2007).

However, mere connection to behaviour is not enough, for at least two reasons. First, a predictive relationship between an MVPA measure (like classifier accuracy) and subject behaviour could be spurious (Ritchie et al., in press). Second, even if the relationship is not spurious, it might not reveal how the information in a brain region is exploited. For example, a common application of RSA is to construct a behavioural dissimilarity matrix from similarity judgments for visual stimuli, which can in turn be used to predict how well these stimuli will be distinguished by a classifier, and compared to a neural dissimilarity matrix constructed from pairwise comparisons of neural activity patterns (Bracci and Op de Beeck, 2016; Connolly et al., 2012; Mur et al., 2013; Wardle et al., 2016). Since a dissimilarity matrix provides a general-format model of the pairwise relationships between points in geometric space, a correlation between the two matrices suggests that there is some correspondence between the psychological space exploited by observers when making the judgments, and the encoding space as reflected in the neural dissimilarity matrix. However, it is important to recognize that such a correspondence does not provide an account of how the neural responses might be driving the participants' judgments. When correlated, the behavioural dissimilarity matrix in effect provides one of potentially many models that can be compared with the neural responses. The focus is therefore not on how neural responses predict observer behaviour, but the reverse: how observer behaviour can be used to model neural responses. So, while the relationship is not spurious, it is focused on the wrong predictive relationship.

While it is clear that observer behaviour is a (far) downstream effect of the sort of read-out procedure utilised the brain, what is missing is a connection to the neural code, as reflected in patterns of measured neural activity, which allows us to take it as a proxy for this procedure. Or to put the point differently: how does one provide evidence that the encoding activation space of a region is what is driving the nuances of observer behaviour? Directly comparing behavioural and neural dissimilarity matrices already points us towards a promising approach, based on distance-based models of observer responses of choice and reaction times for categorisation, which we will briefly describe.

According to distance-based models, observer behaviour on categorisation tasks reflects distance metrics defined over psychological spaces (Ashby and Maddox, 1993, 1994). This is true of many familiar approaches from psychology. Prototype models characterise observer choice as a result of the distance of a stimulus representation in psychological space to the central tendencies of the distribution of different categories within the space, while exemplar models predict choice from the summated distance of a target stimulus representation in comparison to all other individual stimulus representations. Further models, intermediary between these two (e.g. with local prototyping or clustering) are also possible (Briscoe and Feldman, 2011; Vanpaemel and Storms, 2008), as are models that adaptively compare stimulus representations to different category clusters during category learning (Love, Medin, and Gureckis, 2004). Rather than estimating distance between representations themselves, decision boundary models predict observer choice using the distance between a stimulus and a decision boundary through the space. This is the case with the simplest version of such models, signal detection theory (Green and Swets, 1966), as well as its multidimensional extensions (Ashby and Townsend, 1986). Nor are these models restricted to predicting observer choice, but apply to reaction times as well (Ashby, 2000; Nosofsky and Palmeri, 1997; Ratcliff, 1985).

One way to conceptualize these models is that they offer theories of how observers evaluate evidence when performing a categorisation task, where the evidence in question is characterised in terms of a psychological space. If a brain region implements such a space, then we can apply these same models to the neural spaces we reconstruct with MVPA (de Hollander et al., 2016; Forstmann et al., 2011). If we can predict observer behaviour from such a space, in line with a psychological model of this space, then this would provide stronger (albeit far from conclusive) evidence that the information contained in the space may be exploited by downstream processes. The reason is that in this case the models offer a hypothesis of *how* information present in neural activation patterns are related to the evidence used by observers carrying out an experimental task. It is thus the *psychological plausibility* of these models that warrants treating observer behaviour as a proxy for downstream processing, and provides evidence that the information may be formatted in a manner that is used or is usable by the brain in a task relevant manner (Ritchie and Carlson, 2016; Ritchie et al., in press).

The proposed connection between neural and psychological spaces is not a new one (Edelman et al., 1998; Kriegeskorte et al., 2008a), but what we believe has been under appreciated is its potential to help constrain the representational interpretation of MVPA findings. Of course, behaviour is still operating as a proxy for the process of interest, but the virtue of this approach is that behavioural data is easy to collect (either on- or off-line), and there is a rich collection of distance-based models of choice and reaction time that can be directly related to neural data sets. Indeed, many researchers have already adopted approaches similar to what we are suggesting, which we will briefly review.

Some studies have directly related categorisation models to neural activation spaces. For example Op de Beeck et al. (2008) compared exemplar models where the input space was either generated by observer behaviour or from cellular recordings in monkey IT (Op de Beeck et al., 2001). They found that the latter data in fact provided a better characterisation of choice behaviour for a number of categorisation tasks defined for simple parameterised shape stimuli. More recently Mack et al.

(2013) compared exemplar and prototype models to neural activation from across the brain and found that the former model had greater mutual information with the pattern responses from multiple brain regions. And Davis and Poldrack (2014) showed that neural typicality in temporal and occipital brain regions was a strong predictor of observer typicality judgments for novel learned object categories. Finally, Mack et al. (2016) found that attentionally weighting of stimulus dimensions for different categorisation tasks correlated with neural patterns in left anterior hippocampus using a clustering model of category learning.

Another approach based on decision boundary models focuses on reaction times rather than choice behaviour. A key aspect of these models is that since a decision boundary is drawn to separate between the distributions of different stimuli, evidence close to the boundary will be more ambiguous between stimulus types, while evidence far from the boundary will be less ambiguous. Under the assumption that reaction times tend to vary with the quality of evidence, a prediction of these models is that distance a decision boundary negatively correlates with reaction time (Ashby and Maddox, 1994; Pike, 1973). Applying this model to the activation spaces reconstructed using MVPA, recently we have tested the idea that distance from a classifier decision boundary would likewise negatively correlate with reaction times (Ritchie and Carlson, 2016). We found that the prediction was born out based on neural patterns from ventral temporal cortex in humans (Carlson et al., 2014), and with MEG decoding where we observed that the distance-RT correlation tended to reflect the time course of classifier performance (Ritchie et al., 2015).

The general application of these points is that if we want to make inferences about the content and structure of neural representations that are targeted with MVPA methods, then we would be well served to focus on measures that make a connection to the downstream read-out procedures of the brain. Here we have highlighted one promising avenue, which is to connect psychological models to the activation spaces recovered with MVPA. However, this is just one approach, which itself is not necessarily sufficient for inferring that the information latent in any particular brain region is being read-out by later processing. The more general point is that greater focus must be put on how decodable information is used by the brain.

4.4. Moving beyond the third ghost

Theories are tested against stable phenomena. Yet even assuming the existence of an uncontroversial phenomenon, theory testing can be complex in its own right. This is not a new observation, and presented at a general level we think hardly anyone would object. However, we have suggested here that at least in parts of the decoding literature, phenomenon and theory are often run together: that the discovery of decodable information is taken to provide direct evidence about underlying representational capacities of a brain region. This is a problematic inference, even when it happens upon the truth. We have tried to suggest methods by which the inference can be made more reliable, but there is plenty left to explore.

5. Conclusions

Machine learning methods are a valuable addition to the toolkit of cognitive neuroscience. Indeed, it is the very *power* of machine learning that gives rise to the "ghosts" we have identified. First, since these powerful techniques can produce demonstrable results even when the investigators do not fully understand what information is driving the classification, this can mask the importance of determining the underlying neural source of data upon which the classification is based. Second, the nature of these methods raises potential problems for the inference from data to phenomenon. Precisely because decoding methods often make it seem like no interpretation is required, misinterpretation can occur. Finally, problems can arise with the move from phenomenon to theory when the discovery of decodable information is taken to provide direct evidence about underlying representational capacities of a brain region. For each of these identified "ghosts" we have suggested some ways to move beyond them, but much more remains to be explored.

As with any scientific technique, the multi-step nature of the inferences involved in decoding research in cognitive neuroscience creates various interpretative challenges. By separating out the links in the inferential chain, and showing how and why problems arise, we have demonstrated how some of the more obvious pitfalls can be avoided.

Acknowledgments

This project was funded under Australian Research Council Future Fellowships to C.K. and T.A.C. (FT140100422, FT120100816), an ARC Discovery Project to T.A.C. (DP160101300), and an FWO [PEGASUS][2] Marie Sklodowska-Curie Fellowship to J.B.R. (12T9217N). We thank S.S. Solomon, S.K. Cheong, S.C. Chen and A.S. Pietersen for assistance with electrophysiological data collection.

References

Albright, T.D., 1984. Direction and orientation selectivity of neurons in visual area MT of the macaque. J. Neurophysiol. 52, 1106–1130.
Alink, A., Krugliak, A., Walther, A., Kriegeskorte, N., 2013. fMRI orientation decoding in V1 does not require global maps or globally coherent orientation stimuli. Front. Psychol. 4, 493.
Anderson, M.L., 2014. After Phrenology : Neural Reuse and the Interactive Brain.
Ashby, F.G., 2000. A stochastic version of general recognition theory. J. Math. Psychol. 44, 310–329.
Ashby, F.G., Maddox, W.T., 1993. Relations between prototype, exemplar, and decision bound models of categorization. J. Math. Psychol. 37, 372–400.
Ashby, F.G., Maddox, W.T., 1994. A response time theory of separability and integrality in speeded classification. J. Math. Psychol. 38, 423–466.
Ashby, F.G., Townsend, J.T., 1986. Varieties of perceptual independence. Psychol. Rev. 93, 154–179.
Bartels, A., Logothetis, N.K., Moutoussis, K., 2008. fMRI and its interpretations: an illustration on directional selectivity in area V5/MT. Trends Neurosci. 31, 444–453.
Bogen, J., Woodward, J., 1988. Saving the phenomena. Philos. Rev. 97, 303–352.
Bouchard, K.E., Mesgarani, N., Johnson, K., Chang, E.F., 2013. Functional organization of human sensorimotor cortex for speech articulation. Nature 495, 327–332.
Boynton, G.M., 2005. Imaging orientation selectivity: decoding conscious perception in. Nat. Neurosci. 8, 541–542, V1.
Bracci, S., Op de Beeck, H., 2016. Dissociations and associations between shape and category representations in the two visual pathways. J. Neurosci. 36, 432–444.
Briscoe, E., Feldman, J., 2011. Conceptual complexity and the bias/variance tradeoff. Cognition 118, 2–16.
Britten, K.H., Newsome, W.T., Shadlen, M.N., Celebrini, S., Movshon, J.A., 1996. A relationship between behavioral choice and the visual responses of neurons in macaque MT. Vis. Neurosci. 13, 87–100.
Brodersen, K.H., Haiss, F., Ong, C.S., Jung, F., Tittgemeyer, M., Buhmann, J.M., Weber, B., Stephan, K.E., 2011. Model-based feature construction for multivariate decoding. Neuroimage 56, 601–615.
Carlson, T.A., 2014. Orientation decoding in human visual cortex: new insights from an unbiased perspective. J. Neurosci. 34, 8373–8383.
Carlson, T.A., Ritchie, J.B., Kriegeskorte, N., Durvasula, S., Ma, J., 2014. Reaction time for object categorization is predicted by representational distance. J. Cogn. Neurosci. 26, 132–142.
Carlson, T.A., Schrater, P., He, S., 2003. Patterns of activity in the categorical representations of objects. J. Cogn. Neurosci. 15, 704–717.
Cichy, R.M., Ramirez, F.M., Pantazis, D., 2015. Can visual information encoded in cortical columns be decoded from magnetoencephalography data in humans? Neuroimage 121, 193–204.
Connolly, A.C., Guntupalli, J.S., Gors, J., Hanke, M., Halchenko, Y.O., Wu, Y.C., Abdi, H., Haxby, J.V., 2012. The representation of biological classes in the human brain. J. Neurosci. 32, 2608–2618.
Cox, D.D., Savoy, R.L., 2003. Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage 19, 261–270.
Davis, T., Poldrack, R.A., 2014. Quantifying the internal structure of categories using a neural typicality measure. Cereb. Cortex 24, 1720–1737.
de Hollander, G., Forstmann, B.U., Brown, S.D., 2016. Different ways of linking behavioral and neural data via computational cognitive models. Biol. Psychiatry Cognit. Neurosci. Neuroimaging 1, 101–109.
de-Wit, L., Alexander, D., Ekroll, V., Wagemans, J., 2016. Is neuroimaging measuring information in the brain? Psychol. Bull. Rev. 23, 1415–1428.
deCharms, R.C., Zador, A., 2000. Neural representation and the cortical code. Annu. Rev. Neurosci. 23, 613–647.
Desimone, R., Duncan, J., 1995. Neural mechanisms of selective visual attention. Annu. Rev. Neurosci. 18, 193–222.

DiCarlo, J.J., Cox, D.D., 2007. Untangling invariant object recognition. Trends Cogn. Sci. 11, 333–341.
Dubois, J., de Berker, A.O., Tsao, D.Y., 2015. Single-unit recordings in the macaque face patch system reveal limitations of fMRI MVPA. J. Neurosci. 35, 2791–2802.
Dumoulin, S.O., Wandell, B.A., 2008. Population receptive field estimates in human visual cortex. Neuroimage 39, 647–660.
Edelman, S., Grill-Spector, K., Kushnir, T., Malach, R., 1998. Toward direct visualization of the internal shape representation space by fMRI. Psychobiology 26, 309–321.
Eliasmith, C., Anderson, C.H., 2002. Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems. MIT Press, Cambridge, Mass.
Forstmann, B.U., Wagenmakers, E.J., Eichele, T., Brown, S., Serences, J.T., 2011. Reciprocal relations between cognitive neuroscience and formal cognitive models: opposites attract? Trends Cogn. Sci. 15, 272–279.
Freeman, J., Brouwer, G.J., Heeger, D.J., Merriam, E.P., 2011. Orientation decoding depends on maps, not columns. J. Neurosci. 31, 4792–4804.
Freeman, J., Heeger, D.J., Merriam, E.P., 2013. Coarse-scale biases for spirals and orientation in human visual cortex. J. Neurosci. 33, 19695–19703.
Furmanski, C.S., Engel, S.A., 2000. An oblique effect in human primary visual cortex. Nat. Neurosci. 3, 535–536.
Goddard, E., Klein, C., Solomon, S.G., Hogendoorn, H., Carlson, T.A., 2018. Interpreting the dimensions of neural feature representations revealed by dimensionality reduction. Neuroimage 180 (Part A), 41–67.
Green, D.M., Swets, J.A., 1966. Signal Detection Theory and Psychophysics. Wiley, New York.
Guest, O., Love, B.C., 2017. What the success of brain imaging implies about the neural code. Elife 6.
Hannan, T., 2016. Jesus on toast. Australas. Sci. 37, 41.
Hanson, N.R., 1958. Patterns of Discovery; an Inquiry into the Conceptual Foundations of Science. University Press, Cambridge Eng.
Hatsopoulos, N.G., Donoghue, J.P., 2009. The science of neural interface systems. Annu. Rev. Neurosci. 32, 249–266.
Haufe, S., Meinecke, F., Gorgen, K., Dahne, S., Haynes, J.D., Blankertz, B., Biessmann, F., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage 87, 96–110.
Haxby, J.V., Connolly, A.C., Guntupalli, J.S., 2014. Decoding neural representational spaces using multivariate pattern analysis. Annu. Rev. Neurosci. 37, 435–456.
Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.
Haynes, J.D., 2015. A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. Neuron 87, 257–270.
Haynes, J.D., Rees, G., 2005. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. Nat. Neurosci. 8, 686–691.
Haynes, J.D., Rees, G., 2006. Decoding mental states from brain activity in humans. Nat. Rev. Neurosci. 7, 523–534.
Hubel, D.H., Wiesel, T.N., 1968. Receptive fields and functional architecture of monkey striate cortex. J. Physiol. 195, 215–243.
Huth, A.G., de Heer, W.A., Griffiths, T.L., Theunissen, F.E., Gallant, J.L., 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532, 453–458.
Huth, A.G., Nishimoto, S., Vu, A.T., Gallant, J.L., 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron 76, 1210–1224.
Jehee, J.F., Brady, D.K., Tong, F., 2011. Attention improves encoding of task-relevant features in the human visual cortex. J. Neurosci. 31, 8210–8219.
Jordan, M.I., Mitchell, T.M., 2015. Machine learning: trends, perspectives, and prospects. Science 349, 255–260.
Judd, C.M., Westfall, J., Kenny, D.A., 2012. Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. J. Pers. Soc. Psychol. 103, 54–69.
Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8, 679–685.
Kanwisher, N., 2010. Functional specificity in the human brain: a window into the functional architecture of the mind. Proc. Natl. Acad. Sci. U. S. A. 107, 11163–11170.
Kastner, S., Ungerleider, L.G., 2000. Mechanisms of visual attention in the human cortex. Annu. Rev. Neurosci. 23, 315–341.
Kay, K.N., Naselaris, T., Prenger, R.J., Gallant, J.L., 2008. Identifying natural images from human brain activity. Nature 452, 352–355.
Kiani, R., Esteky, H., Mirpour, K., Tanaka, K., 2007. Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. J. Neurophysiol. 97, 4296–4309.
Kirsh, D., 1990. When is information explicitly represented? In: Hanson, P.P. (Ed.), Information, Language, and cognition. University of British Columbia Press, Vancouver, 411 pp.
Kirsh, D., 2003. Implicit and Explicit Representation. In: Nadel, L. (Ed.), Encyclopedia of cognitive Science. Nature Publishing Group, London, pp. 478–481.
Klein, C., 2010. Images are not the evidence of neuroimaging. Br. J. Philos. Sci. 61, 265–278.
Kriegeskorte, N., Bandettini, P., 2007. Analyzing for information, not activation, to exploit high-resolution fMRI. Neuroimage 38, 649–662.
Kriegeskorte, N., Kievit, R.A., 2013. Representational geometry: integrating cognition, computation, and the brain. Trends Cogn. Sci. 17, 401–412.
Kriegeskorte, N., Mur, M., Bandettini, P., 2008. Representational similarity analysis - connecting the branches of systems neuroscience. Front. Syst. Neurosci. 2, 4.
Kriegeskorte, N., Mur, M., Ruff, D.A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., Bandettini, P.A., 2008. Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60, 1126–1141.
Love, B.C., Medin, D.L., Gureckis, T.M., 2004. SUSTAIN: a network model of category learning. Psychol. Rev. 111, 309–332.
Lehky, S.R., Kiani, R., Esteky, H., Tanaka, K., 2014. Dimensionality of object representations in monkey inferotemporal cortex. Neural comput.. 26, 2135–2162.
Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., Oeltermann, A., 2001. Neurophysiological investigation of the basis of the fMRI signal. Nature 412, 150–157.
Machens, C.K., Romo, R., Brody, C.D., 2010. Functional, but not anatomical, separation of "what" and "when" in prefrontal cortex. J. Neurosci. 30, 350–360.
Mack, M.L., Preston, A.R., Love, B.C., 2013. Decoding the brain's algorithm for categorization from its neural implementation. Curr. Biol. 23, 2023–2027.
Marr, D., 1982. Vision: a Computational Investigation into the Human Representation and Processing of Visual Information. W.H. Freeman, New York.
Maunsell, J.H., Van Essen, D.C., 1983. Functional properties of neurons in middle temporal visual area of the macaque monkey. I. Selectivity for stimulus direction, speed, and orientation. J. Neurophysiol. 49, 1127–1147.
Misaki, M., Kim, Y., Bandettini, P.A., Kriegeskorte, N., 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. Neuroimage 53, 103–118.
Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P.A., Kriegeskorte, N., 2013. Human object-similarity judgments reflect and transcend the primate-IT object representation. Front. Psychol. 4, 128.
Nair, D.G., 2005. About being BOLD. Brain Res. Brain Res. Rev. 50, 229–243.
Naselaris, T., Kay, K.N., 2015. Resolving ambiguities of MVPA using explicit models of representation. Trends Cogn. Sci. 19, 551–554.
Naselaris, T., Kay, K.N., Nishimoto, S., Gallant, J.L., 2011. Encoding and decoding in fMRI. Neuroimage 56, 400–410.
Newsome, W.T., Britten, K.H., Movshon, J.A., 1989. Neuronal correlates of a perceptual decision. Nature 341, 52–54.
Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V., 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. 10, 424–430.
Nosofsky, R.M., Palmeri, T.J., 1997. An exemplar-based random walk model of speeded classification. Psychol. Rev. 104, 266–300.
O'Toole, A.J., Jiang, F., Abdi, H., Penard, N., Dunlop, J.P., Parent, M.A., 2007. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. J. Cogn. Neurosci. 19, 1735–1752.
Op de Beeck, H., Wagemans, J., Vogels, R., 2001. Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. Nat. Neurosci. 4, 1244–1252.
Op de Beeck, H.P., 2010. Probing the mysterious underpinnings of multi-voxel fMRI analyses. Neuroimage 50, 567–571.
Op de Beeck, H.P., Wagemans, J., Vogels, R., 2008. The representation of perceived shape similarity and its role for category learning in monkeys: a modeling study. Vis. Res. 48, 598–610.
Pearl, J., 1995. Causal diagrams for empirical research. Biometrika 82, 669–688.
Piccinini, G., Shagrir, O., 2014. Foundations of computational neuroscience. Curr. Opin. Neurobiol. 25, 25–30.
Pike, R., 1973. Response latency models for signal detection. Psychol. Rev. 80, 53–68.
Poldrack, R.A., 2006. Can cognitive processes be inferred from neuroimaging data? Trends Cogn. Sci. 10, 59–63.
Poldrack, R.A., 2008. The role of fMRI in cognitive neuroscience: where do we stand? Curr. Opin. Neurobiol. 18, 223–227.
Poldrack, R.A., 2011. Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron 72, 692–697.
Poldrack, R.A., Farah, M.J., 2015. Progress and challenges in probing the human brain. Nature 526, 371–379.
Pouget, A., Dayan, P., Zemel, R., 2000. Information processing with population codes. Nat. Rev. Neurosci. 1, 125–132.
Ratcliff, R., 1985. Theoretical interpretations of the speed and accuracy of positive and negative responses. Psychol. Rev. 92, 212–225.
Rieke, F., 1997. Spikes: Exploring the Neural Code. MIT Press, Cambridge, Mass.
Riesenhuber, M., Poggio, T., 1999. Hierarchical models of object recognition in cortex. Nat. Neurosci. 2, 1019–1025.
Ritchie, J.B., Carlson, T.A., 2016. Neural decoding and "inner" psychophysics: a distance-to-bound approach for linking mind, brain, and behavior. Front. Neurosci. 10, 190.
Ritchie J.B., Kaplan, D.M., and Klein, C., Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. Br. J. Philos. Sci. in press.
Ritchie, J.B., Tovar, D.A., Carlson, T.A., 2015. Emerging object representations in the visual system predict reaction times for categorization. PLoS Comput. Biol. 11, e1004316.
Salzman, C.D., Britten, K.H., Newsome, W.T., 1990. Cortical microstimulation influences perceptual judgements of motion direction. Nature 346, 174–177.
Sasaki, Y., Rajimehr, R., Kim, B.W., Ekstrom, L.B., Vanduffel, W., Tootell, R.B., 2006. The radial bias: a different slant on visual orientation sensitivity in human and nonhuman primates. Neuron 51, 661–670.
Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., Poggio, T., 2007. Robust object recognition with cortex-like mechanisms. IEEE Trans. Pattern Anal. Mach. Intell. 29, 411–426.
Spirtes, P., Glymour, C.N., Scheines, R., 2000. Causation, Prediction, and Search, second ed. MIT Press, Cambridge, Mass.
Sternberg, S., 2011. Modular processes in mind and brain. Cogn. Neuropsychol. 28, 156–208.
Stokes, M.G., Wolff, M.J., Spaak, E., 2015. Decoding rich spatial information with high temporal resolution. Trends Cogn. Sci. 19, 636–638.

Suppes, P., 1962. Models of data. In: Nagel, E., Suppes, P., Tarski, A. (Eds.), Logic, Methodology, and Philosophy of Science. Stanford University Press, Stanford, CA, pp. 252–261.
Tong, F., Pratte, M.S., 2012. Decoding patterns of human brain activity. Annu. Rev. Psychol. 63, 483–509.
Vanpaemel, W., Storms, G., 2008. In search of abstraction: the varying abstraction model of categorization. Psychol. Bull. Rev. 15, 732–749.
Voss, J.L., Federmeier, K.D., Paller, K.A., 2012. The potato chip really does look like Elvis! Neural hallmarks of conceptual processing associated with finding novel shapes subjectively meaningful. Cereb. Cortex 22, 2354–2364.
Wardle, S.G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S.M., Carlson, T.A., 2016. Perceptual similarity of visual patterns predicts dynamic neural activation patterns measured with MEG. Neuroimage 132, 59–70.
Wardle, S.G., Ritchie, J.B., Seymour, K., Carlson, T.A., 2017. Edge-related activity is not necessary to explain orientation decoding in human visual cortex. J. Neurosci. 37, 1187–1196.
Weisberg, D.S., Keil, F.C., Goodstein, J., Rawson, E., Gray, J.R., 2008. The seductive allure of neuroscience explanations. J. Cogn. Neurosci. 20, 470–477.
Williams, M.A., Dang, S., Kanwisher, N.G., 2007. Only some spatial patterns of fMRI response are read out in task performance. Nat. Neurosci. 10, 685–686.
Wood, J., Nezworski, M., Lilienfeld, S., Garb, H., 2003. What's Wrong with the Rorschach?: Science Confronts the Controversial Inkblot Test. Jossey-Bass, San Francisco, CA.
Woodward, J., 2003. Making Things Happen: a Theory of Causal Explanation. Oxford University Press, Oxford; New York.
Yacoub, E., Harel, N., Ugurbil, K., 2008. High-field fMRI unveils orientation columns in humans. Proc. Natl. Acad. Sci. U. S. A. 105, 10607–10612.
Yamins, D.L., DiCarlo, J.J., 2016. Eight open questions in the computational modeling of higher sensory cortex. Curr. Opin. Neurobiol. 37, 114–120.
Zinszer, B.D., Anderson, A.J., Kang, O., Wheatley, T., Raizada, R.D., 2016. Semantic structural alignment of neural representational spaces enables translation between English and Chinese words. J. Cogn. Neurosci. 28, 1749–1759.