Deepening Representations

Phillip Hintikka Kieval

Abstract

There is a vast literature in cognitive science and empirical philosophy of mind concerning whether connectionist models can provide explanations in terms of representational content. Critics have claimed that connectionist networks cannot provide properly psychological explanations, since there is no principled way to group tokened representations into types that share the same content. Despite this once vigorous debate, relatively little has been said about representational content in state-of-the-art deep neural networks (DNNs). While these networks can be viewed as extensions of early connectionist models, their added complexity introduces a number of complications. This paper brings modern advances in machine learning and computer vision to bear on the question of content in connectionist networks. I consider what I call the *cluster approach*, which I take to be the strongest account of content in connectionist systems. I show that technical and methodological problems arise when generalizing this approach to DNNs. I then provide a much-needed update to the cluster approach by incorporating the tools of representational similarity analysis (RSA) used in cognitive neuroscience. I argue that RSA should be understood as form of hypothesis testing in search of real, causal patterns in the underlying neural data. This provides important confirmatory evidence about the presences of category representations in an underlying neural activation space.

1 Introduction

In the results of the 14th biennial Critical Assessment of protein Structure Prediction (CASP), AlphaFold 2—a deep learning system designed by Google's DeepMind research group—was recognized as a solution to the "protein folding problem" (Jumper et al., 2020). This achievement was hailed as a major scientific advance for both biology and artificial intelligence (AI). The problem of predicting a protein's structure from its amino acid chain has stood as a grand challenge in biology for the last fifty years. There are an estimated 10^{300} possible conformations of a typical protein. This means that it would take longer than the age of the known universe to enumerate each conformation by brute force calculation (Levinthal, 1969). Yet, AlphaFold 2 can accurately determine structures in a matter of days. The successful prediction of spontaneous folding depends on AlphaFold's capacity to acquire and exploit representations of amino acid residue pairs to generate and update an abstract spatial graph model of a protein's possible structure.

State-of-the-art machine learning systems like AlphaFold 2 are ubiquitous in modern life. These systems find applications in everything from automated decision assistance in medicine and criminal justice to playing complex board games like chess and Go (Silver et al., 2016). Our epistemic networks are increasingly intertwined with machine learning algorithms, and scientists rely on them in complex computational models. Understanding how information is encoded in abstract representations is crucial for understanding the intelligent capacities of these systems. This imperative suggests revisiting connectionist approaches that explain intelligence in terms of the capacities to learn abstract representations from low-level perceptual inputs (Hassabis et al., 2017; Buckner, 2018).

Connectionists hope to explain cognition using artificial neural networks as models (Rumelhart et al., 1986). Connectionist networks consist in large, parallel collections of artificial neurons called units, or nodes, together with weights that measure the strength of connections between nodes. We can view these networks as models of the neurons and synaptic links in the brain at some level of abstraction. Mounting empirical evidence suggests that such models are useful tools for learning about regions of interest in the brain. Even so, connectionism is not without its skeptics (Fodor and Pylyshyn, 1988; Fodor and Lepore, 1999; Marcus, 2018). One standard complaint against connectionism charges that connectionist networks cannot provide a psychological explanation of behavior, since there is no principled way to separate internal representations into groups which share the same content. If connectionism is to succeed as a representational theory of cognition, we must have a plausible theory of representation that preserves sameness of content in the face of individual variance in physical implementation.

This paper scrutinizes an explicit proposal about the type of entity to be considered as content-bearing in connectionist systems, namely that the vehicles of content in connectionist systems are clusters of neural activity distributed across a model's internal mechanism (Tiffany, 1999; Rupert, 2001; Gardenfors, 2000; Shea, 2007). Call this proposal the *cluster*

approach to content. This approach adapts the idea of a state space semantics deployed in earlier accounts of semantic similarity in connectionist networks (Horgan and Tienson, 1996; Churchland, 1998). The basic idea of what Horgan and Tienson (1996) call *dynamical cognition* is that we can use the tools of dynamical systems theory to provide a mathematical characterization of a cognitive system (see also van Gelder, 1995). The set of all possible states of a system defines that system's state space. Applied to connectionist networks, state space conceives of the internal, representational landscape of a network as a mathematically defined space whose axes are constituted by the possible activation levels of the neurons distributed across the network. Individual points in state space correspond to a global pattern of activity in the network. So, points in this space can be construed as token representations. We can then identify clusters of points, or regions in state space, as corresponding to representation types (Horgan and Tienson, 1996; Gardenfors, 2000; Shea, 2007). The cluster approach has become the *de facto* solution to the problem of fixing the semantic content of representations for connectionists over the course of the last two decades. This approach comes with considerable advantages, so it merits careful consideration. As such, this paper begins by first describing the strengths of the approach before revealing pressing problems for generalizing its results. I close with a proposal for moving forward with representational explanations of connectionist networks inspired by recent work in cognitive neuroscience.

Despite its successes, the cluster approach builds on an outdated picture of neural networks. The growing ubiquity of AI has brought increased philosophical attention to deep learning (Lecun et al., 2015; Buckner, 2019). Since deep neural networks (hereafter, DNNs) are extensions of early connectionist networks, the cluster approach to state space semantics would seem the obvious choice for analyzing representational content in these models. Yet, there is a glaring problem with applying the cluster approach to DNNs that has thus far gone unnoticed in the literature on content. The complexity of these networks means they process representations in many hundreds or even thousands of dimensions. The "curse of dimensionality" refers to a set of problems that emerge when analyzing data in high-dimensional spaces that otherwise would not occur in low-dimensional spaces like those of ordinary three-dimensional experience. One such phenomenon concerns the problem of measuring relative distance in high-dimensional space. As dimensionality increases, state space becomes increasingly sparse and nearest neighbor searches from a given starting point become unstable. Sparseness means that measures of distance between points in the space expand and the ratios of relative distance between different points converge. As the number of dimensions approaches infinity, the proportional difference between the nearest and farthest point distances from an arbitrary starting point in the state space approaches zero. This means that the relevant contrast between points in the state space of a model vanishes as dimensionality increases. The curse of dimensionality suggests that the concept of proximity may not even be qualitatively meaningful in the high-dimensional state spaces that are characteristic of DNNs. In practice, modelers must make choices about how to structure their raw data to make inferences about encoded representations. But these choices threaten

to bias conclusions about phenomena drawn from data.

Through the course of this paper I aim to convince the reader that a naive cluster approach faces serious difficulties with high-dimensional state spaces. I then offer a much-needed update to the cluster approach that builds on the tools of representational similarity analysis (RSA) used in cognitive neuroscience. The structure of the paper is as follows: §2 introduces the cluster approach to content in connectionist networks. In §3, I introduce a host of technical problems for generalizing the cluster approach to DNNs posed by the curse of dimensionality. Finally, §4 argues that we can look to empirical work in neuroscience that identifies predictive patterns in the representational structure of computational models as a guide to testing hypotheses about neural representations.

2 The Cluster Approach to Content

Connectionism broadly-construed takes artificial neural networks as idealized models of cognition. A neural network consists in multiple, interconnected layers of units joined together by a pattern of weights which determine the strength of activity passed from one unit to the next. The layers of a network are divided into three classes: an input layer that receives encoded information for processing, an output layer that produces the result of processing, and a hidden layer (or layers) that lies in between. The process of training a simple classifier network involves introducing a large number of antecedently labeled example inputs and using a supervised algorithm called error backpropagation to fine-tune the associative links between units such that the network learns to successfully generalize to novel inputs.

Computational explanations of neural networks involve identifying representational vehicles and a criterion for fixing representational content. The vehicles of content are physical particulars over which the network performs computations and which carry information that can be read out by various downstream processes. Identifying representations in a connectionist model usually involves measuring the distributed patterns of activation across the hidden units of a network associated with different input samples. Each token representation in a neural network will be one such pattern of neural activity distributed across the network's hidden layer(s). Such a pattern, or activation vector, can be characterized as an n-tuple, where n is the number of neurons in the population. We can conceptualize these activations as points in a geometric space for the purpose of comparing informational content. *State space* refers to an abstract, n-dimensional space whose axes are constituted by the possible activation of units in the hidden layers of the model such that any activation vector corresponds to a point in that space. A point in state space corresponds with a token representation.

The *cluster approach* says that sameness of content can be measured by analyzing the internal structure of clusters of points in the model's state space. Rather than identifying individual hidden units or the relational structure between points in state space as the vehicles of content, the cluster approach identifies clusters of activity in state space as the

correct vehicles of content. On this view, representations can be grouped into identical semantic types in virtue of their falling within discrete regions in state space where contentful clusters of neural activity emerge through training. Clusters of points are identified relative to the overall volume filled by the activation points produced by all training samples. Since the size of a cluster is relative to the overall size of the representational space, clusters can be realized in nets with different numbers of hidden units. The idea is just that a set of training samples that produces a cluster of points in state space A might likewise produce a cluster of points in state space B. In such a case, two different state space clusters in two different neural networks can represent the very same property to their respective networks.

Gardenfors (2000) provides a detailed account of how information can be represented using geometric structures in a *conceptual space*. A conceptual space (or feature space) is just an abstract, metric space whose axes are constituted by a number of quality dimensions, which represent the various qualities of objects. Natural properties are represented by convex regions in a conceptual space.[1] A conceptual space is distinct from a representational state space, since the former is just a way of describing the possible properties of a set of stimuli. A conceptual space instantiates a representational state state space when it corresponds to how a system represents a set of stimuli under the relevant task conditions such that neural activity reliably varies along the dimensions of the defined conceptual space (Goddard et al., 2018). Since many feature dimensions are determined perceptually (e.g. color, shape, or spatial properties), the criterion that natural properties correspond to convex regions in a conceptual space turns out to be perceptually grounded (Gardenfors, 2000, p. 77). This makes it serviceable for modeling dynamical cognition when the dimensions of neural variation can be properly identified for a particular task. This framework can be fruitfully applied to a connectionist network to characterize the representational structure of its state space in terms of convex regions, or clusters.

So, neural networks can be described by clusters in state space, but why should we ascribe content to them? Shea (2007), who provides among the clearest articulations of the cluster approach, argues that we should do so "if clusters are to be invoked, as is common empirical practice, in an explanation of a network's ability to generalize its correct performance to new samples; thus that clusters are vehicles of content to the extent that they form the basis of generalization" (Shea, 2007, p. 12-13). If those clusters play a role in the mechanism of operation that results in successful generalization, then those clusters are content bearing.

This appears to be born out by empirical evidence. Training tends to cause networks to develop tight clusters in state space. This is because the goal of training can be characterized as producing a function for correctly sorting inputs into clusters at the output layer. The development of hidden layer clusters serves as an important intermediary in achieving this goal. These clusters can be individuated by taking a trained network and plotting the

[1] A subset C of a conceptual space S is said to be convex if, for all points x and y in C, all points between x and y are also in C (Gardenfors, 2000, p. 66-71).

distribution of activation points in its state space. Convex regions in state space containing clusters of activity can then be identified by measuring and comparing the relative distances of groups of proximal points in relation to the overall volume filled by activity in the space.[2]

According to Shea, content ascriptions are always made on the basis of the successful performance of a network. Regions in state space do not themselves antecedently represent anything. The argument for taking clusters as the vehicles of content is based on giving an explanation for successful generalization. It is only after a network has been trained and can generalize performance to novel inputs that it represents (Shea, 2007). The process of learning transforms the state space of a model from content-less, mathematically defined space to one that tracks properties that are relevant for classification. The process of learning is one by which the model learns to represent fundamental patterns in the distal features of its environment and generate successful behavior on the basis of those patterns.

The cluster approach says that a tokened representation in a neural network associated with a novel input stimulus belongs to a contentful type just in case it falls into some hidden layer cluster formed by samples in the training set that share a property that is causally or constitutively relevant to the classificatory task at the output layer (Shea, 2007). Such a cluster can be identified with a convex region in state space. Belonging to such a region indicates that membership in a representation type that tracks a natural property of perceptual inputs (Gardenfors, 2000). If this is right, then we can use this kind of cluster analysis to verify that the relations between our putative vehicles of content match our intuitions about concept similarity. Furthermore, if clusters are vehicles of content, then different networks can have semantically identical representations. This is empirically testable since we can verify whether a group of sample inputs which activate a cluster in some network A also activate a cluster in another network B. By dispensing with an overly fine-grained notion content, the cluster approach avoids the worries plaguing earlier theories of content in connectionist systems. Instead, clusters in state space group many different states of a network's internal mechanism into types, and a straightforward, representational explanation of neural network performance is in hand.

3 Dimensionality and Category Representations

The field of AI has shifted rapidly since the introduction of the cluster approach. Over the course of the last decade, deep learning has become by far the most successful approach to machine learning (ML) (Lecun et al., 2015). Compared to earlier connectionist networks, state-of-the-art deep neural networks contain many more layers with different kinds of pro-

[2]This can also be achieved using Voronoi tessellation (Gardenfors, 2000). Given a set of sites in a space $P = \{p_1, \ldots, p_n\}$, Voronoi tessellation partitions a space such that an arbitrary point q belongs to a region i when bisector $d(q, p_i)$ is less than $d(q, p_j)$. However, when it comes to a representational space, this explicitly assumes that we have information about prototypical representations, where each site represents a concept prototype (Gardenfors, 2000, p. 87).

cessing nodes that deploy varying activation functions. The deepening of models affords enormous gains in their computational power, efficiency, and ability to solve complex decision problems when compared to shallower networks. DNNs have achieved human-level benchmarks in novel image classification tasks. Their success underwrites advances in autonomous vehicles, facial recognition, and assisted medical diagnoses. Moreover, DNNs are among the strongest of our current candidates for a partial explanatory model of the human visual system (Yamins et al., 2014; Yamins and DiCarlo, 2016). Yamins and co-authors focus their attention primarily on hierarchically structured deep convolutional neural networks (DCNNs) as explanatory models of the primate ventral visual pathway (Cao and Yamins, 2021). These networks are not only appealing for their impressive performance in object categorization tasks. They also incorporate several hallmark features of neural computation, such as nonlinear transduction and max pooling of inputs (Yamins and DiCarlo, 2016). Attention to these networks is essential to a healthy synthesis of work in AI and cognitive neuroscience (Hassabis et al., 2017; Botvinick et al., 2017; Buckner, 2019).

The pioneering work that undergirds the cluster approach builds on exclusively shallow classifier networks. For instance, the networks Shea cites as empirical support for his proposal all had a single hidden layer with only a few units. Conversely, state-of-the-art deep learning models used in modern ML tasks compute over many thousands of individual nodes across many more hidden layers. Even a relatively old five-layer deep convolutional neural network deployed in Krizhevsky et al. (2012) contains 650,000 individual nodes with over 60 million parameters. These models will have state spaces that are orders of magnitude more complex than those in a simple three-layer network. Though it would seem natural to extend the cluster approach to representational explanations of deep learning models, their added complexity introduces pressing technical and methodological problems.

A major impediment to making good sense of the role internal representations play in explanations of decisions made by DNNs arises from the fact that their object recognition tasks operate in a high-dimensional state space. Category clusters in such a space can be characterized instead as multidimensional manifolds embedded in this space. A neural object manifold is a subspace of lower-dimensionality embedded in a high-dimensional activation space. There are good reasons to think that manifolds play a representational role. Typical explanations of perceptual similarity in such models *do* invoke notions similar to those employed by the cluster approach, albeit with additional complexity. The goal of training an image classifier can be understood as locating a global output function for transforming and drawing a discrete boundary—in the form of a linear hyperplane—between category manifolds embedded in the network's state space (DiCarlo and Cox, 2007). But the high-dimensional nature of DNNs poses a significant difficulty for visualization and defies easy understanding. Dimensionality raises important methodological concerns for the cluster approach, since manifolds will not be so well-behaved as their low-dimensional equivalents.

Our ordinary intuitions often fail when dealing with high dimensionality. In practice, manifolds in highly complex models become hopelessly entangled. This entanglement suggest

that category manifolds do not satisfy the convexity criterion (Gardenfors, 2000), since, for many category manifolds, there will be a pair of members x and y such that some point along the bisector $d(x, y)$ is not a member of the manifold. Moreover, general problems with working with high-dimensional data draws into question whether independently measuring the representational similarity of category manifolds remains tenable in our ever deepening models. For example, Gardenfors (2000) relies on Voronoi tessellation as a procedure for identifying convex regions in a feature space corresponding to natural properties. However, Voronoi tessellation is not guaranteed to partition a feature space into convex regions when using a non-Euclidean distance metric. Worse yet, constructing Voronoi diagrams can often become prohibitively expensive in even moderately high-dimensional spaces. This raises a potential concern for the cluster approach, since existing accounts are not attuned to the counter-intuitive behavior of high-dimensional feature spaces.

I should note here that the problem of dimensionality is not limited to DNNs, since dimensionality is strictly speaking a feature of layers within a network, while depth is a structural feature of the network as a whole. Dimensionality could pose a problem even for shallow networks with a single densely packed layer. Under the assumption that distributed patterns of activity are representations, the problem of dimensionality must be dealt with regardless of depth since each activity pattern will be a point in a high-dimensional neural space. For instance, the computational vision model, HMAX (Riesenhuber and Poggio, 1999), is a simple hierarchical feedforward network with four layers where the final layer contains 484 nodes, resulting in a moderately high dimensionality. Similarly, any ImageNET trained network will have a high-dimensional image space, regardless of depth, due to the 224 x 224 pixel images encoded at the input layer. But why any modeler would ever want to design a shallow network with many, many nodes is mysterious given the enormous advantages of increased depth. The focus on depth here is justified by the pragmatic success of task-optimized DNNs as both impressive feats of engineering and as potential explanatory models of regions of interest in the brain.

Empirical work in high-dimensional data analysis has repeatedly drawn attention to what has become known as "the curse of dimensionality." The curse of dimensionality refers to a collection of difficulties posed for similarity measurements, nearest neighbor searches, outlier detection, and other important data analysis tasks that arise in high-dimensional vector spaces. A related problem concerns the general behavior of distance measurements in high-dimensional spaces: it may turn out that such measurements become less and less qualitatively meaningful as dimensionality increases.

High-dimensional multivariate data analysis requires a notion of distance that generalizes to many dimensions. Such distance measurements typically rely on what is called the L_p-norm (or Minkowski metric). Here L_p is a norm picking out a function from a real vector

space to the nonnegative real numbers that satisfies certain properties. Formally,

$$L_p(x, y) = \sum_{i=1}^{n} (\left\| x^i - y^i \right\|^p)^{1/p}$$

This function expresses that, for the points x and y in the n-dimensional real vector space R^n, and for the value p in the set of real numbers greater than or equal to 1, L_p of x and y equals the p^{th}-root of the sum of the differences between each variable of x and y raised to the power of p. For instance, when p takes a value of 2, L_p is the Euclidean norm and distance between points in the vector space is measured using Euclidean distance. When measuring Euclidean distance, this amounts to using the Pythagorean theorem generalized to a space of arbitrarily many dimensions. We can imagine drawing a bisector between points x and y in R^n and then triangulating the length of that line by first summing the squares of the difference between each dimensional variable of x and y and then taking the square root of that total. The choice of norm constrains similarity analyses by effecting how modelers carve up the high-dimensional space. While Euclidean distance is the most commonly used metric norm, p can take a wide range of real values.

This is where the curse of dimensionality rears its head. Making inferences about similarity in ML models becomes difficult in high dimensions, since a training set of fixed size occupies a shrinking fraction of state space as dimensionality increases (Domingos, 2012). Moreover, high-dimensional feature space is sparse. The vast majority of the volume of a multivariate normal distribution lies not near the mean, but in a very small, increasingly distant fraction of the space around the mean. Formal results show that, for a wide range of norms, the proportional difference between the nearest and farthest points from a given sample point in a space vanishes as dimensionality increases (Beyer et al., 1999; Pestov, 2000). This would be a non-issue if there were only two elements to search through. But there are many irrelevant features in most ML tasks, and the high noise-to-signal ratio makes the search unstable. This makes similarity comparisons on the basis of relative distance in state space dubious, since nearest neighbor searches become effectively random as models become more complex. The relative contrast between proximal and distal points tends to become less and less meaningful in high-dimensional spaces. Essentially, the space expands, and the points of interest grow farther and farther apart from one another such that the relative contrast between them becomes qualitatively meaningless. The distances between points which intuitively should be nearby become indistinguishable from the distances between points farther away from each other.

Zimek et al. (2012) demonstrate that using distance measurements to identify clusters becomes increasingly difficult in high-dimensional spaces. Their results show that, for Euclidean distances, the volume of a hypersphere exhibits unusual behavior that makes similarity judgments less meaningful in high-dimensional state spaces. They observed that small changes in its radius could decide whether every point or no point fell within the volume of

a hypersphere (Zimek et al., 2012, p. 369-370). This would be as if, while trying to visualize clusters in a two dimensional scatter plot, very minuscule changes in the radius of a cluster made the difference between every single activation in the network falling within the same cluster and no activations falling within that same cluster.

Zimek et al. (2012) point out that this concentration effect is not always observed when data is multiply distributed. Nearest-neighbor searches can be both theoretically and practically meaningful when the search is limited to objects from the same cluster as the query point, and it is assumed that other clusters are well separated from the cluster in question (Zimek et al., 2012, p. 367). They appeal here to the concept of pairwise stability of clusters, which holds whenever the mean distance between points of different clusters dominates the mean distance between points belonging to the same cluster (see also Bennett et al., 1999). When clusters are pairwise stable, the nearest neighbor to any point belonging to a given cluster will tend to also be a member of that same cluster (Zimek et al., 2012; Bennett et al., 1999). But this is not necessarily helpful when category cluster membership is precisely what is at issue. Additionally, this assumes that all dimensions of a dataset carry information that is relevant for clustering. However, this is often not the case, and irrelevant features in the data can impede the separation of distributions.

Durrant and Kabán (2009) identify the prevalence of irrelevant attributes in feature space as a major contributing factor to the curse of dimensionality. The presence of irrelevant attributes adds noise that masks attributes relevant for categorization. The ratio of noise to correlated attributes in the dataset are a determinative factor in whether distance concentration occurs (Durrant and Kabán, 2009; Zimek et al., 2012). Durrant and Kabán (2009) provide accordingly a negative requirement for avoiding the concentration of distance measures in high-dimensional Euclidean space: the contribution of relevant attributes must grow no slower than that of noise (Durrant and Kabán, 2009, p. 390). Successful category representation acquisition requires parsing through relevant and irrelevant attributes and learning to abstract away from features of category exemplars that do not contribute to successful categorization. When inputs have many irrelevant features, the distance concentration effect poses a problem for category clustering in high dimensions. This points to a connection between the curse of dimensionality and the problem of "nuisance variation" which plagues perceptual similarity tasks in ML. Nuisance factors are repeated and systematic sources of variance that do not contribute to successful decisions. ML systems and computer vision models need to learn how to adjust to nuisance variation to solve perceptual similarity tasks. DNNs' resilience in the face of nuisance variation is frequently explained in terms of their capacity to draw exponentially more linear regions in state space compared to shallower networks to identify far more complex decision boundaries (Montúfar et al., 2014). The question is how can we effectively interpret the representational structure underlying these complex decision boundaries?

One suggestion is to opt for alternative measures of similarity. This is reflected in the common practice in cognitive neuroscience of using correlation distance to measure similar-

ity between points in state space associated with different brain states (Kriegeskorte et al., 2008a; Kriegeskorte and Kievit, 2013; Nili et al., 2014). Correlation distance computes similarity by subtracting the Pearson's correlation coefficient ρ between two activation vectors from 1. Pearson's correlation is a measure of normalized covariance between two variables. Correlation can be given a geometric interpretation due to its relation to cosine distance. The cosine of the angle between two vectors can be obtained by normalizing them by their L_2-norms and then calculating their inner product. The correlation distance between two n-dimensional vectors is equivalent to their cosine distance after subtracting the mean value from each activity pattern.[3] Unlike Euclidean distance, normalization makes correlation distance scale invariant. Correlation distance is 0 when the normalized patterns of activity are identical and normally takes a real value between 0 and 1 where lower values indicate that the two patterns of activity are closer together in state space. Representational similarity analysis (RSA) uses these correlation measures to then construct a simplified model of the underlying representational space (Kriegeskorte et al., 2008a; Kriegeskorte and Kievit, 2013). RSA takes the pairwise distance between activation vectors evoked by different experimental conditions and arranges them into a two-dimensional representational dissimilarity matrix. This same method of analyzing the representational geometry of state space has been used to compare representations in the primate visual system to those in DNNs (Khaligh-Razavi and Kriegeskorte, 2014; Yamins et al., 2014). But if the underlying space is not well-structured (e.g. it suffers from the distance concentration effect), then these problems with structure are likely to manifest in the reconstructed space as well. Moreover, as Carlson et al. (2018) point out, by moving in this way from data to phenomenon modelers make certain choices in reconstructing the brain's representational space. These choices embody assumptions that could potentially distort the true nature of that space (Carlson et al., 2018, p. 92). In this case, the decision to use correlation distance as a similarity measure effects how representations are reconstructed from the data.

The above point is made clear by Bobadilla-Suarez et al. (2020), who suggest that the choice of similarity measure can have significant implications for what the underlying representational space might look like. They consider three distinct families of common similarity measures and analyze how they impact the structure of the underlying representational space. They evaluated the performance of these different similarity measures across different regions of interest in the brain and for different categorization tasks. They did this using a decoding approach, relying on confusability in a ML classifier as a proxy for baseline neural similarity. This is motivated by the intuition that brain states that are similar should be more likely to be confusable in decoding. They then tested the predicted similarity relations generated by each measure against a confusability matrix generated from classifier decisions.

[3]The cosine distance between X and Y can be stated as: $d_{cos}(X,Y) = 1 - \frac{\langle X,Y\rangle}{\|X\|\|Y\|} = 1 - cos(\angle X,Y)$. If $\bar{V} = V - \overline{V}$ represents the difference between some vector and its sample mean, then correlation distance can be defined as $d_\rho(X,Y) = 1 - \frac{\langle \bar{X},\bar{Y}\rangle}{\|\bar{X}\|\|\bar{Y}\|} = 1 - cos(\angle \bar{X},\bar{Y})$.

Their results indicated that "similarity profiles differed between studies, suggesting that the operable neural similarity measures can change as a function of task or stimuli" (Bobadilla-Suarez et al., 2020, p. 377). Moreover, competing metrics tended to outperform correlation distance.

Another suggestion is to use some kind of feature selection to reduce the dimensionality of the representations under investigation. Dimensionality reduction involves reducing the number of variables required to accurately describe high-dimensional patterns of activity. There are a range of different techniques used for dimensionality reduction. Methods of dimensionality reduction that are optimized for feature extraction and cluster analysis may provide an efficient way of visualizing high-dimensional representations. This is reflected in the common practice of using principle component analysis (PCA) or multidimensional scaling (MDS) to analyze and interpret neural manifolds in cognitive neuroscience (see for instance Kriegeskorte et al., 2008b; Mante et al., 2013) PCA reduces the axes of the underlying space to the dimensions with the highest variation with change in stimuli. These highly variable dimensions are interpreted as those which are functionally relevant for the task at hand. This makes it essentially a kind of data-driven functional decomposition. However, neural activity is noisy and highly correlated, and PCA runs the risk of excluding functionally relevant dimensions, or overemphasizing irrelevant ones. In general, dimensionality reduction can introduce similar interpretive difficulties, like overlap in visualized points where none exists in the native dimensionality of state space. Perhaps more problematically, these choices also embody implicit assumptions on the part of the modeler about the content and structure of neural representations. These assumptions threaten to introduce bias in the transition from data to phenomena (Carlson et al., 2018).

Collectively, these issues present sufficient cause for concern that traditional methods for identifying category clusters in state space as the vehicles of content may be beyond reach in DNNs with many thousands of dimensions. Shea (2007) explicitly calls for identifying clusters by first plotting the distribution of points in state space before identifying regions containing clusters of proximal points relative to the overall volume filled by points produced by all training samples (Shea, 2007, p. 9-10). Similarly, Gardenfors (2000) uses Voronoi tessellation with Euclidean distance to partition a feature space into convex regions. But in high-dimensional space, *all* of the points in a distribution will tend to be relatively distant from one another. The intuition-breaking behavior of high-dimensional space means that attempting to identify the border of a cluster in these ways might be effectively impossible since small variations in the radius of the cluster could mean the difference between including or excluding every single activation in the sample set. In practice, modelers make a host of choices to structure the data that underlies a state space to analyze and interpret the representational structure of a neural population or computational model. However, these choices embody implicit assumptions of scientists. These assumptions constrain the range of hypotheses tested against the data, threaten to warp the underlying structure of state space, and possibly bias conclusions drawn from data.

Nevertheless, I think that the general thrust of the cluster approach seems on the right track, since manifolds are invoked in quasi-representational explanations of the mechanism of operation of DNNs. But if the cluster approach is to provide the kind of theory of representation connectionists desire, then we will need a method for measuring representational similarity that is generalizable to high-dimensional models, else the prospect of an adequate state space semantics collapses. Clearly, there is representational structure latent in object manifolds. But the question remains, how can we make the task of interpreting that structure tractable without introducing undesirable conceptual baggage that biases the range of hypotheses modelers consider? The following section looks to the philosophy of science for answers.

4 Content in Deep Neural Networks

The difficulties presented above suggest that we cannot make meaningful, hypothesis-neutral judgments about cluster membership in high-dimensional space. Nevertheless, manifolds—the high-dimensional equivalents of clusters—are typically invoked in existing explanations of the successful performance of DNNs. This suggests that representational explanations of DNNs are still within reach. However, we still need a method for individuating type-identical content when dimensionality precludes grouping representations by directly plotting points in state space. Fortunately, conceptual tools from the philosophy of science can help us move beyond these methodological worries.

4.1 Representational vehicles exploit real patterns in nature

The most influential accounts of representational content—especially Fodor (1990), Dretske (1988), and Millikan (1984)—focus primarily on the intentionality of sensory-perceptual states. Their hope was to incorporate intelligent behavior into the scientific image of the world in a way that preserved the behavior-guiding role played by the contents of internal states. In seeing that my dog Frasier is now sitting beside me, there is something that my accompanying perceptual state is directed at—my furry, tail-wagging companion. The idea goes that there is some internal state that represents Frasier to the mechanism in my brain responsible for processing sensory inputs and producing the right behavior, namely petting him. However, the sense of representation in play here is far more deflationary and cashed out in terms of encoded information that can be read out by downstream processes to produce behavior. First, this involves identifying physical constituents of a mechanism that we can group into types. It also requires a principled criterion for fixing the meaning of these groups according to the information which they purport to encode. To do this we need to say how a system could manage to carve out features of its environment that are alike in kind and represent them for processing in a physical mechanism. Historically, this has meant showing that representations reliably instantiate the right kind of causal relation

with what they represent such that they function to track the relevant kind of phenomena (Dretske, 1988). We can say that a state of a system tracks a kind when the state has been causally selected to control that system's interactions with phenomena of that kind in virtue of the information that it encodes.

We can think of tokened representations as type-identical when they function to track stimuli of the same kind. I offer that we can think of these kinds as stable property clusters bound together by causal-informational relations (Slater, 2015). We can bootstrap Slater's (2015) stable property cluster account of kinds with Andersen's information-theoretic revitalization of Dennett's "real patterns" (see also Potochnik, 2017) to get some purchase on how a neural network could come to track such a kind (Andersen, 2017; Dennett, 1991; Stinson, 2020). According to Andersen (2017), a real pattern is one that "can be reliably picked out and tracked through time and which allows one to make predictions that are better than chance." Real patterns are counterfactually robust; the microphysical state underlying a tokening of a pattern could have been different, while still tokening the same pattern. The basic idea is that kinds or patterns can be reliably picked out and tracked by information-theoretic means and make useful predictions. This ensures that patterns are not met with jury-rigged kinds. Since real patterns make useful predictions and are stable under counterfactual perturbation, a collection of phenomena that manifest a real pattern will constitute a kind.

To get a handle on what constitutes a real pattern, consider a digital chess program in which two computational engines are playing each other in a game of chess. The state of this game at any one moment in time can be described by a complex array of pixels, or a bit map. The bit map gives a complete description of the board-state at any instant in the game. In principle, we could compute the entirety of the current board-state using nothing but the bit map. If we know enough details about the algorithms our chess-playing engines implement, then we could use the bit map to predict future board-states. But that would be extremely computationally costly. It is much more efficient to characterize our program at a higher-level of description in terms of chess positions. At this level of description, familiar patterns emerge from the complicated array of flashing pixels. We can identify them as knights, rooks, pawns, and all of the recognizable features that constitute a board-state in a game of chess. Once recognized as a game of chess, enormously more efficient ways of predicting future board-states become available to us with little loss of accuracy (depending on how adept you are at chess). Recognizing these real patterns means the difference between computing millions of pixels and merely inferring in your head what is likely to be the best move in an ongoing game of chess.

Applying real patterns to representations requires some disambiguation. In doing so, I do not mean to endorse Dennett's instrumentalism about content. Real patterns refer to *causal* patterns that structure our natural world (Andersen, 2017; Potochnik, 2017). These patterns are idealized descriptions of underlying phenomena that allow us to successfully navigate the messy complexities inherent in nature. When thinking about the content of

encoded representations, we can think of them as tracking stable patterns in our environment. Our perceptual mechanisms can exploit these patterns, allowing for more efficient neural computations. We can say that a tokened representation belongs to a type when a system's internal mechanism has learned to exploit a real pattern in nature for the purpose of representing and controlling its behavior with respect to that associated kind of phenomena. This is compatible with the commitment to some non-semantic way of individuating the physical particulars instantiated in a system that encode representations of patterns embodied by distal features of the environment. These representations are veridical when they allow for the successful prediction and manipulation of the features of the environment they function to track. Like other causal-informational patterns, representational content is counterfactually robust. The internal state of a mechanism underlying a tokened representation could have been different while still tokening the same content, or type. This yields a modest picture of concept acquisition in terms of subjective category representations, or "conceptualizations" (Gauker, 2011, p.6), which nevertheless play a useful role in explaining how different neural networks can learn the same concepts. Different networks may have slightly different conceptualizations based on differences in their initialization conditions or input encoding (Mehrer et al., 2020), but the category groupings correspond to the same real patterns with only slight variations or exceptions in how they are represented to the network. This modest realism about the content and structure of representations is also compatible with understanding representations themselves as causal patterns that arise from the transformations of input signals into output signals and thus the legitimate subjects of idealized models (Kriegeskorte and Kievit, 2013; Diedrichsen and Kriegeskorte, 2017).

There are many ways to describe a single pattern. There are often trade-offs—though they needn't be strict ones—between computational efficiency and accuracy. Whether we prefer efficient but noisy patterns or more computationally expensive patterns with lower noise tolerance may depend on how easily and reliably we can identify a pattern and the costs to us for getting things wrong. Dennet refers to specifications of such preferences as "design decisions" (Dennett, 1991). While these design decisions are not available to us when it comes our sensory-perceptual system, they are incorporated into its phylogeny. When it comes to neural networks, many of these design choices are plausibly assimilated through training. The learning process can be interpreted as searching through the feature space to identify patterns with the goal of maximizing the ratio of computational efficiency to error function much in the same way that evolutionary competition selects for efficiency gains in biological systems (Buckner, 2019, p. 10). Since these patterns are considered real features of the natural world, representations manage to carve out real joints in nature through this selective process. As such, the vehicles of content latch on to efficient, stable patterns that allow them to successfully control behavior. These patterns also help scientists make predictions and sound inductive inferences about the behavior of a system.

Understanding representational content in terms of a system learning to exploit real patterns seems to vindicate the cluster approach. After all, the cluster approach understands

training as transforming state space into a mechanism that causally tracks properties relevant for classification. Clusters get their content in virtue of the causal role they play in producing successful behavior. But this understanding also opens new avenues for ascribing content to models where the curse of dimensionality precludes directly mapping out vehicles of content in state space. We just need to identify patterns that are represented to a network's mechanism of operation that allow us to make accurate predictions and inductive inferences about the network's successful performance. In practice, this involves examining idealized models that make patterns in the geometric structure of state space more salient.

Idealization has enjoyed significant philosophical attention, especially in the context of scientific models (Cartwright, 1994; Weisberg, 2007; Rohwer and Rice, 2013; Elliott-Graves and Weisberg, 2014; Potochnik, 2017). Idealizations are assumptions made without regard for whether they are true and often with full knowledge that they are false (Cartwright, 1994; Potochnik, 2017, p. 42). Specifying real patterns that are productive for cognitive neuroscience motivates idealization. Idealizations make a cognitively valuable contribution to scientific models. They do so by representing a system as if it were some way that it is not. The positive content of idealizations is to center the relevance of some real pattern. Potochnik (2017) suggests that "idealizations contribute to understanding by representing as-if to the end of depicting a causal pattern, thereby highlighting certain aspects of that phenomenon (to the exclusion of others) and revealing connections with other, possibly disparate phenomena that embody the same pattern or, in some cases, that are exceptions to that pattern" (Potochnik, 2017, p. 97). I argue that this should be seen as the explanatory goal of representational models in cognitive neuroscience, which compare the representational geometry of different activation spaces. Moreover, what causal pattern occupies the focal point of a particular scientific explanation is sensitive to the interests and goals of scientists. Therefore, it should not come as a surprise that even more data-driven methods of functionally decomposing a high-dimensional state space are not hypothesis-neutral, nor should this be seen as a devastating flaw. Real patterns that are of fundamental interests to the explanatory goals of scientists already constrain the hypothesis-space. This is legitimate so long as those interests are well-motivated by empirical and theoretical considerations. Carlson et al. (2018) make a similar remark that hypothesis-driven approaches that use dimensionality reduction can be defensible, "so long as they are carefully constrained" (Carlson et al., 2018, p. 95). As long as these conditions are satisfied, different ways of decomposing state space to make real patterns salient to scientists are sound methods of testing whether a target phenomenon embodies the causal pattern of interest.

With this on board, the actual methods and practices of cognitive neuroscience and ML can serve as a guide to representational explanations. Recent work in these areas has uncovered a surprising fact about deep convolutional neural networks (DCNNs). Models optimized merely to classify images predict spiking responses in the highest level of the ventral stream, the inferior temporal cortex (IT) (Yamins et al., 2014). That such task-optimized models manage to predict something about the brain supports the notion that these neural

networks form partial explanatory models of the brain (Cao and Yamins, 2021). These developments point in the direction of progress for cognitive science. The quantitative methods used to establish predictive relationships between DCNNs and primate brain activity can be assimilated for the purpose of ascribing content to high-dimensional connectionist models.

4.2 Representational models and hypothesis testing

We want to understand the structure and content of representations encoded in DNNs, at least in part, to better understand how these models can teach us about neural representations in the brain. In service of this goal it is helpful to distinguish between decoding and encoding models that are used to analyze neural representations. Multivariate pattern analysis (MVPA) refers to a set of techniques for analyzing patterns in neural data (Haxby, 2012; Ritchie et al., 2019). MVPA can be separated into decoding and encoding methods (Naselaris et al., 2011; Naselaris and Kay, 2015; Kriegeskorte and Douglas, 2019). Researchers often appeal to linear decodability to reveal the content of neural representations (Eliasmith, 2013). Neuroscientists rely on computational models to make comparisons or assist in decoding subsets of neural populations, or regions of interest, in the brain. The idea is that a biologically plausible linear classifier can stand in as a proxy for the downstream processes that read out information encoded by neural representations. However, decodability does not always license inferences about representations, especially when decoding is itself assisted by ML techniques (Ritchie et al., 2019). Both linear and non-linear classifiers are unconstrained by the information that the brain actually exploits. While brute force classifiers may prove effective at distinguishing activity patterns associated with different experimental conditions, we need an additional guarantee that a decoding exploits the same information encoded by neural representations.

Encoding models aim to predict the response patterns of neural activity from descriptions of the experimental conditions (Kriegeskorte and Douglas, 2019). Encoding models work in the opposite direction of decoding models. Whereas decoding models aim to detect the presence of specific information in the brain, encoding models map sensory stimuli to sensory regions of the brain and make predictions about the representational space. Encoding models thus work in the same direction as the flow of information in the brain to make comprehensive predictions about the representational space (Naselaris and Kay, 2015; Yamins and DiCarlo, 2016; Diedrichsen and Kriegeskorte, 2017). An encoding model can consist in a DNN model trained on some sensory-perceptual task (Kriegeskorte and Douglas, 2019). Since there is no one-to-one correspondence between the model and measurement channels in the brain, we need to apply some level of idealizations to test the predictions of an encoding model. This can either involve testing predictions of raw measurements using a linear transform or by testing predictions about the similarity structure in a representational space (Naselaris et al., 2011; Kriegeskorte and Douglas, 2019; Cao and Yamins, 2021). The latter case is well suited to the goal of analyzing representational content. We can use the techniques of

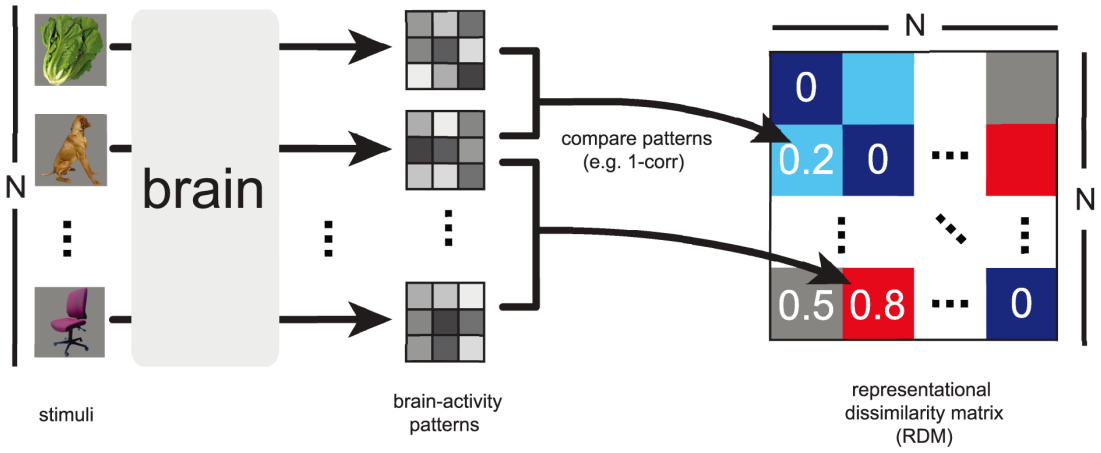

Figure 1: Computing representational dissimilarity matrices (RDMs) (reproduced from Nili et al. (2014)).

representational similarity analysis (RSA) to make hypotheses and test predictions about the similarity relations of representations associated with different experimental conditions. RSA provides a framework for comparing represented information that seeks out such predictive relations by establishing a second-order isomorphism between the representational structure of a model and its target brain region (Kriegeskorte et al., 2008a).

RSA examines the representational structure of an activation space using representational dissimilarity matrices (RDMs). RDMs visualize pairs of neural activations in a square, symmetric matrix of pairwise stimulus correlations. An RDM contains a cell for each unique pair of experimental conditions. Each cell contains a value measuring the similarity between the activation vectors associated with two stimuli. For a given set of stimuli, the matrix describes how similar or dissimilar the representations are according to a chosen similarity measure (e.g correlation distance, Mahalanobis distance) for each pair of stimuli. Entries along the diagonal represent comparisons between identical stimuli and take a value of 0. The value of each off-diagonal entry represents the dissimilarity between the activation patterns corresponding with two different stimuli. Lower value entries indicate that a pair of stimuli produce more similar representations, while a value of 1 indicates no correlation whatsoever. We then construct the matrix by arranging each stimuli into an order, usually according to an observers' intuitive similarity judgments, and assigning the computed (dis)similarity value to its corresponding cell (Figure 1) (Kriegeskorte et al., 2008a; Nili et al., 2014). The result is a two-dimensional map of the similarity relations between a set of activation vectors. RDMs measuring IT neural population responses exhibit a clear block-diagonal structure characteristic of the IT's high performance at object categorization. Predictively adequate

connectionist models will naturally exhibit a similar block-diagonal structure in their own RDMs. When cells are arranged according to observer similarity judgments, strong structural correlation provides evidence that the activation space implements a representational space (Kriegeskorte et al., 2008a). Thus, RDMs can be interpreted as a simplified description of the representational geometry of a given activation space.

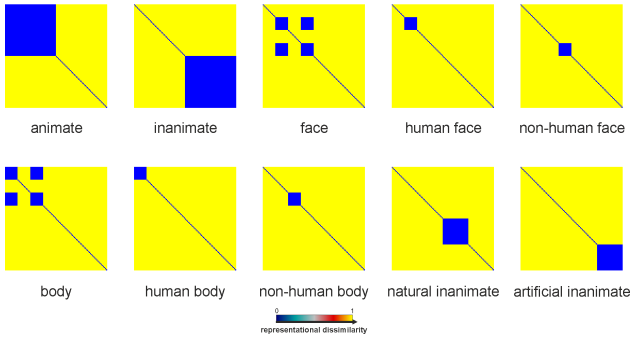

Figure 2: Khaligh-Razavi and Kriegeskorte (2014) created ten different category-cluster RDMs as predictors of clustering. These prediction RDMs were then fit to each computational model using a linear regression to model the semantic structure of the their representations (reproduced from Khaligh-Razavi and Kriegeskorte (2014)).

A clear example of this framework in action can be found in Khaligh-Razavi and Kriegeskorte (2014). Khaligh-Razavi and Kriegeskorte (2014) analyzed brain responses in both monkey IT and human IT for a set of color images of objects spanning a range of animate and inanimate categories. They then used RDMs to compare these representations to those generated by 37 different computational models of varying designs. To measure the strength of clustering they created ten category-cluster RDMs as predictors, which they fit to each IT and model RDM (Figure 2). These grouped the set of experimental conditions according to a number of intuitive categories. The category-clusters represented animate, inanimate, face, human face, non-human face, body, human body, non-human body, natural inanimate, and artificial inanimate (Khaligh-Razavi and Kriegeskorte, 2014). Though models designed to emulate the structure of the ventral stream (such as HMAX and VisNet) were included among the 37 computational models, they found that these were outperformed in predictive accuracy by a task-optimized DCNN. The representations generated by the supervised DCNN best predicted the the category clustering found in the IT RDMs (Figure 3).

One way of interpreting this procedure is that the modelers identified patterns in their experimental conditions and constructed a model of a conceptual space around these patterns to generate the category-cluster RDMs. By fitting this categorical model to an encoding model, we can identify the real patterns that arise as plausible candidates for the content of representations. The modelers then used these patterns to predict the representational structure of IT RDMs. What we find is that—much like the primate visual system—higher

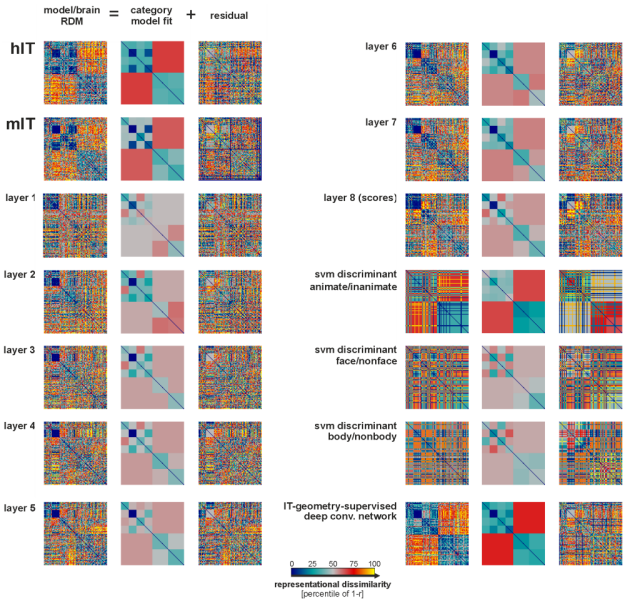

Figure 3: Category-cluster analysis of a supervised DCNN model reveals that a similar representational structure to human and monkey IT neuronal population responses emerges across the various layers of the model. The final weighted combination of layers (bottom right) shows a similar clustering structure to that of the hIT and mIT (reproduced from Khaligh-Razavi and Kriegeskorte (2014)).

levels of processing in the DCNN begin to approximate the same patterns. Each subsequent layer represents and processes higher-level properties of the input stimuli with greater tolerance for noise and nuisance variation than the layers preceding it. Comparing brain and model RDMs establishes a kind of second-order isomorphism between model and brain region representations that mirrors the category-cluster patterns constructed by a well-defined conceptual space. This suggests that both our computational model and its target system represent the same real patterns at later stages of processing. This allows these systems to efficiently identify new conditions and generalize successful performance to these conditions. If this interpretation is right, then DCNNs turn out to be predictive of neural activity in the primate visual system precisely because they learn to represent and exploit the same kinds of high-level patterns.

However, the modelers' choice of similarity measure is a crucial decision for RSA. Which measure of similarity one chooses has implications for how the representational space is reconstructed from the underlying data. While it has been commonplace to use correlation distance because it is scale invariant, there are open questions as to whether this is a principled reason for choosing it as a similarity measure (Walther et al., 2016; Bobadilla-Suarez et al., 2020). Results from Walther et al. (2016) and Bobadilla-Suarez et al. (2020) both

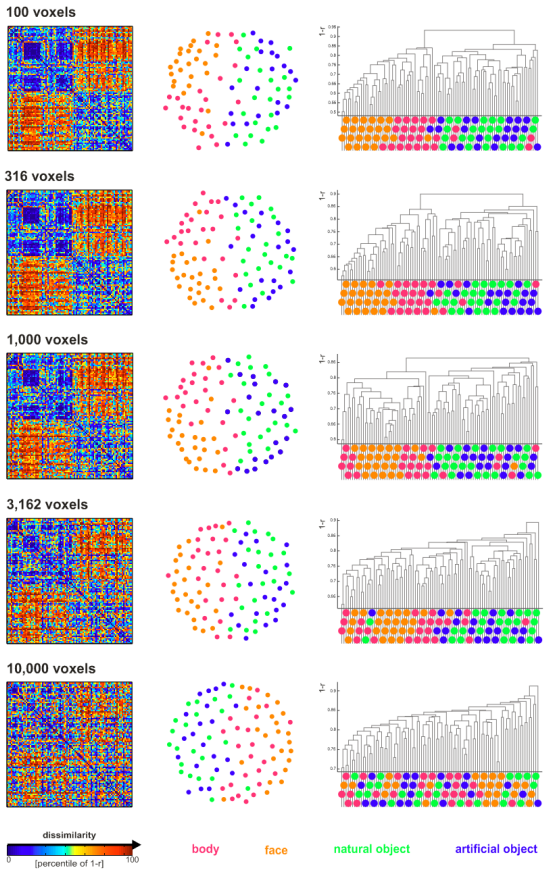

Figure 4: RDMs (left), MDS arrangements (middle), and dendrogram trees (right) were computed on the same region while varying the number of voxels selected for inclusion. The similarity structure and categorical clustering of human IT varies based on the number of voxels (reproduced from Kriegeskorte et al. (2008b) supplemental data).

seem to suggest that other measures of similarity better capture the decoding processes used by the brain for certain kinds of stimuli. Moreover, modelers typically opt for some form of noise normalization to better capture the dimensions of variation in the representational space. This typically precedes some kind of feature selection and subsequent dimensionality reduction. Dimensionality reduction can prove helpful for navigating intractably complex activation spaces and representing only the dimensions that are functionally relevant to a representational space with lower intrinsic dimensionality. However, these data-driven methods are not guaranteed to generate a hypothesis-neutral transition from data to phenomena (Goddard et al., 2018; Carlson et al., 2018). How the modeler chooses to carve up the data effects the structure of the representational space.

Figure 4 illustrates this effect. Kriegeskorte et al. (2008b) used data-driven methods to

identify the responsiveness of voxels in the IT cortex to visual stimuli. Sets of voxels were then selected for inclusion in the similarity analysis according to their responsiveness. This marks an implicit decision to treat voxels that are highly responsive to stimuli as functionally relevant to the representational space and to treat less responsive voxels as noise. The number of voxels selected for inclusion impacts the structure and discernibility of patterns in the representational space. As the number of selected voxels increases the categorical structure found in the RDM becomes less distinct. At 10,000 voxels the organization between faces and bodies appears almost completely obliterated. Of course, we have good, theory-driven reasons to suspect that structure is there. Stripping away irrelevant features reveals that structure. However, selecting for certain sets of voxels over others, even when using data-driven methods, runs the risk of biasing results towards a favored hypothesis.

To deal with these worries, we should emphasize an important difference between encoding and decoding models. Encoding model approaches like RSA differ from decoding approaches to MVPA in that a model constitutes a testable hypothesis about the representational structure of a sensory region of interest in the brain. Kriegeskorte et al. (2008a) emphasize the fact that RSA is hypothesis-driven when we test whether a predefined model fits the representational space of a region of interest. This kind of hypothesis testing can be cashed out in terms of an exploratory search for causal patterns in the underlying neural data. The focal pattern of interest constrains the hypothesis space. Which patterns are useful for drawing out representational explanations will depend on a host of goals and interests of particular researchers. Because the underlying neural data is exceptionally complex and noisy, testing for and representing salient, causal patterns in the data often requires various abstractions and idealizations. These might include opting to select for specific features, dimensionality reduction, normalization, or choosing a particular similarity measure. Modelers must make choices like these to structure and represent the state space in such a way that emphasizes the presence of real patterns to facilitate understanding. Idealizations of this kind are often inseparable from the actual practices involved in studying neural representations, even when those methods purport to be data-driven and thus "objective." If we understand encoding approaches as an exploratory search for real patterns in neural data that allow us to make testable predictions about the structure of representational space, then the rampant idealization found in RSA becomes innocuous. After all, there are many ways to represent a focal pattern that a phenomenon embodies. A real pattern is present in the data so long as there is a way of representing the information carried by the pattern that is more efficient than a bit map of the data.

This view of encoding models as a search for real patterns in neural data aligns with the suggestion that hypothesis-laden methods in neuroscience should be properly conceived of as exploratory analyses (Goddard et al., 2018; Carlson et al., 2018). According to Goddard et al. (2018), an exploratory analysis "gives you evidence that, for the stimuli, parameters, and contexts that were examined, there is a feature space that can capture the variation in those parameters" (Goddard et al., 2018, p. 61). It is a real virtue of encoding model approaches

that they use hypothesis-driven methods to carve up an activation space to capture the dimensions of variation associated with different experimental conditions. Ritchie et al. (2019) suggest that we can make more secure inferences about neural representations when information encoded by neural activity can be used to predict observer behavior based on a psychological model (Ritchie et al., 2019, p. 599-601). The idea is that if a state space implements a psychological similarity space, then we can apply psychological hypotheses directly to that state space. RSA can be understood as reconstructing a representational space according to a particular functional decomposition of neural activity. That functional decomposition consists of a set of theoretically and pragmatically motivated idealizations and abstractions that structure the state space into a more well-defined representational space. We can then use a task-optimized DNN as an encoding model that generates testable predictions about the structure of that representational space. These predictions are linked to the target representational space because the behavior of both model and target system are optimized to solve the same task. By testing model predictions against the reconstructed neural data, we can get confirmatory evidence that certain categorical features are actually represented in that space.

My claim then is that we are justified in using connectionist models as experimental tools in virtue of the real patterns of representational similarity identified by RSA. We can make hypotheses that groups of stimuli should be represented as more or less similar to each other on the basis of some shared semantic feature. We can then use an encoding model, like a DCNN, to generate testable predictions about how these patterns are to be represented in a sensory region of the brain. These predictions can then be verified by comparing the structure visualized in different RDMs. This approach has a significant advantage over traditional cluster analysis. The cluster approach attempts to explicitly identify regions or manifolds within a space as the vehicles of content in virtue of clustering of neural activity. It then groups representations into types on the basis of their falling within the bounds of an identified cluster. But, as we have seen, drawing well-defined borders around activity in a state space such that the identified regions capture any coherent notion of semantic content becomes intractable in even moderately high-dimensional spaces. RDMs represent an idealized reconstruction of the geometry of state space by computing the similarity between ordered pairs of individual activation points. Rather than attempting to identify semantic categories with discrete, well defined regions in a high-dimensional state space, RDMs visualize the similarity between individual representations and model how real patterns arise from the underlying neural data.

By identifying category clusters represented in the RDM as markers of content we abstract away from specific architectures and physical implementation. We can instead view an RDM as an idealized model of a network's representational space. We can use this model to make hypotheses and test predictions about representational information and discover confirmatory evidence about the content of neural representations. As seen above, a virtue of this proposal is that RDMs not only enable comparisons between different models, but they

can also be used to compare representations obtained from multiple modalities. This enables us to make meaningful comparisons between representations produced by neural networks and regions of interest in the brain. This provides a promising experimental paradigm for using connectionist models to investigate the representational mechanisms in the brain.

The further upshot of this view is that we use RSA to acquire *evidence* that a pattern of neural activity has a particular representational content. I have laid out the case that understanding representational content in terms of real patterns is fruitful for hypothesis-driven, experimental approaches to cognitive neuroscience. Researchers can use connectionist models to make testable predictions about real patterns in representational geometry of sensory regions in the brain. This emphasis on exploratory searches for real patterns means that the relationship between ML and cognitive neuroscience turns out to be highly experimental rather than purely conceptual. Moreover, connectionist models lend themselves to experimental methods that are hypothesis-driven and that make use of idealizations and abstractions that are both theoretically and pragmatically motivated. Against the backdrop of this complex web of ancillary assumptions, researches can use connectionist models to acquire confirmatory or dis-confirmatory evidence that a set of neural data embodies a particular real pattern.

5 Conclusion

I have just suggested a framework for generalizing the principles of the cluster approach to state-of-the-art deep learning models. The upshot of this is that it motivates a rigorous, empirical approach to studying localized aspects of cognition with connectionist models. If we want to understand the various mechanisms underlying cognition, we should begin by isolating regions of the brain which instantiate those mechanisms. We can extend the principles of the above framework to test different models for similarity to the region of interest in the brain. By testing models with varying designs and degrees of neurophysiological inspiration we can begin to uncover the features of the underlying mechanism which are essential for reproducing the cognitive phenomenon of interest. Through such an iterative process we approach a genuinely explanatory mechanistic model of that cognitive phenomenon.

References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1973:420–434.

Andersen, H. K. (2017). Patterns, Information, and Causation. *The Journal of Philosophy*, 114(11):592–622.

Bennett, K. P., Fayyad, U., and Geiger, D. (1999). Density-based indexing for approximate nearest-neighbor queries. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99*, pages 233–243, New York, New York, USA. ACM Press.

Bernardi, S., Benna, M. K., Rigotti, M., Munuera, J., Fusi, S., and Salzman, C. D. (2020). The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, 183(4):954–967.e21.

Beyer, K., Goldstein, J., Ramakrishnan, R., and Shaft, U. (1999). When is "Nearest Neighbor" Meaningful?

Bobadilla-Suarez, S., Ahlheim, C., Mehrotra, A., Panos, A., and Love, B. C. (2020). Measures of Neural Similarity. *Computational Brain & Behavior*, 3(4):369–383.

Botvinick, M., Barrett, D. G. T., Battaglia, P., de Freitas, N., Kumaran, D., Leibo, J. Z., Lillicrap, T., Modayil, J., Mohamed, S., Rabinowitz, N. C., Rezende, D. J., Santoro, A., Schaul, T., Summerfield, C., Wayne, G., Weber, T., Wierstra, D., Legg, S., and Hassabis, D. (2017). Building Machines that Learn and Think for Themselves: Commentary on Lake et al., Behavioral and Brain Sciences, 2017.

Buckner, C. (2018). Empiricism without magic: transformational abstraction in deep convolutional neural networks. *Synthese*, 195(12):5339–5372.

Buckner, C. (2019). Deep learning: A philosophical introduction. *Philosophy Compass*, 14(10):1–19.

Buckner, C. and Garson, J. (2019). Connectionism.

Cao, R. and Yamins, D. (2021). Making sense of mechanism : How neural network models can explain brain function.

Carlson, T., Goddard, E., Kaplan, D. M., Klein, C., and Ritchie, J. B. (2018). Ghosts in machine learning for cognitive neuroscience: Moving from data to theory. *NeuroImage*, 180(July 2017):88–100.

Cartwright, N. (1994). *Nature's Capacities and Their Measurement*. OUP Catalogue. Oxford University Press.

Churchland, P. (1998). Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered. *The Journal of Philosophy*, 95(1):5–32.

Clark, A. (1989). *Microcognition: Philosophy, Cognitive Science, and Parallel Distributed Processing*. MIT Press, Cambridge, Ma.

Dennett, D. (1991). Real Patterns. 88(1):27–51.

DiCarlo, J. J. and Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, 11(8):333–341.

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3):415–434.

Diedrichsen, J. and Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4):e1005508.

Domingos, P. (2012). A Few Useful Things to Know About Machine Learning. *Communications of the ACM*, 55(10).

Dretske, F. (1988). *Explaining behavior: Reasons in a world of causes.* The MIT Press, Cambridge, MA, US.

Dretske, F. (1989). Reasons and Causes. *Philosophical Perspectives*, 3(1989):1.

Durrant, R. J. and Kabán, A. (2009). When is 'nearest neighbour' meaningful: A converse theorem and implications. *Journal of Complexity*, 25(4):385–397.

Elgin, C. Z. (2004). True Enough. *Philosophical Issues*, 14(1):113–131.

Eliasmith, C. (2013). *How to Build a Brain*, volume 4. Oxford University Press.

Elliott-Graves, A. and Weisberg, M. (2014). Idealization. 3:176–185.

Fodor, J. (2000). *The mind doesn't work that way : the scope and limits of computational psychology.* MIT Press, Cambridge, Ma.

Fodor, J. and Lepore, E. (1999). All at Sea in Semantic Space: Churchland on Meaning Similarity. *The Journal of Philosophy*, 96(8):381–403.

Fodor, J. A. (1990). *A Theory of Content and Other Essays.* MIT Press.

Fodor, J. A. and Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, 28:3–71.

Ganguli, S. and Sompolinsky, H. (2012). Compressed Sensing, Sparsity, and Dimensionality in Neuronal Information Processing and Data Analysis. *Annual review of neuroscience*, 35(1):463–483.

Gardenfors, P. (2000). *Conceptual Spaces : the geometry of thought*. MIT Press, Cambridge, Mass, first mit edition.

Gärdenfors, P. (2004). Conceptual spaces as a framework for knowledge representation. *Mind and Matter*, 2(2):9–27.

Gauker, C. (2011). *Words and Images: An Essay on the Origin of Ideas*. Oxford University Press.

Gładziejewski, P. (2016). Predictive coding and representationalism. *Synthese*, 193:559–582.

Goddard, E., Klein, C., Solomon, S. G., Hogendoorn, H., and Carlson, T. A. (2018). Interpreting the dimensions of neural feature representations revealed by dimensionality reduction. *NeuroImage*, 180(May 2017):41–67.

Gracia, A., González, S., Robles, V., and Menasalvas, E. (2014). A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality. *Information Sciences*, 270:1–27.

Grootswagers, T., Robinson, A. K., and Carlson, T. A. (2019). The representational dynamics of visual objects in rapid serial visual processing streams. *NeuroImage*, 188(October 2018):668–679.

Hassabis, D., Kumaran, D., Summerfield, C., and Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2):245–258.

Haxby, J. V. (2012). Multivariate pattern analysis of fMRI: The early beginnings. *NeuroImage*, 62(2):852–855.

Horgan, T. and Tienson, J. (1994). A nonclassical framework for cognitive science. *Synthese*, 101(3):305–345.

Horgan, T. and Tienson, J. (1995). Connectionism and the Commitments of Folk Psychology. *Philosophical Perspectives*, 9(8):127.

Horgan, T. and Tienson, J. (1996). *Connectionism and the philosophy of psychology*. MIT Press, Cambridge, Mass.

Horgan, T. and Tienson, J. (1997). Précis of connectionism and the philosophy of psychology. *Philosophical Psychology*, 10(3):337–356.

Jia, Y., Wang, H., Shao, S., Long, H., Zhou, Y., and Wang, X. (2019). On geometric structure of activation spaces in neural networks. *arXiv*.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Tunyasuvunakool, K., Ronneberger, O., Bates, R., Žídek, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Potapenko, A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Stanislav, N., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Steinegger, M., Pacholska, M., Silver, D., Vinyals, O., Senior, A. W., Kavukcuoglu, K., Kohli, P., and Hassabis, D. (2020). High Accuracy Protein Structure Prediction Using Deep Learning. In *Fourteenth Critical Assessment of Techniques for Protein Structure Prediction (Abstract Book)*.

Kaplan, D. M. and Craver, C. F. (2011). The explanatory force of dynamical and mathematical models in neuroscience: A mechanistic perspective. *Philosophy of Science*, 78(4):601–627.

Khaligh-Razavi, S. M. and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11).

Kriegeskorte, N. and Diedrichsen, J. (2019). Peeling the Onion of Brain Representations. *Annual Review of Neuroscience*, 42:407–432.

Kriegeskorte, N. and Douglas, P. K. (2019). Interpreting encoding and decoding models Encoding and decoding: concepts with caveats HHS Public Access. *Curr Opin Neurobiol*, 55:167–179.

Kriegeskorte, N. and Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2(NOV):1–28.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., and Bandettini, P. A. (2008b). Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron*, 60(6):1126–1141.

Krizhevsky, B. A., Sutskever, I., and Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems*, volume 25, pages 1097–1105. Curran Associates, Inc.

Laakso, A. and Cottrell, G. (2000). Content and cluster analysis: Assessing representational similarity in neural systems. *Philosophical Psychology*, 13(1):47–76.

Lecun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Levinthal, C. (1969). How to Fold Graciously. In DeBrunner, J. T. P. and Munck, E., editors, *Mossbauer Spectroscopy in Biological Systems: Proceedings of a meeting held at Allerton House, Monticello, Illinois.*, pages 22–24.

Lin, B., Mur, M., Kietzmann, T., and Kriegeskorte, N. (2019). Visualizing Representational Dynamics with Multidimensional Scaling Alignment. In *2019 Conference on Cognitive Computational Neuroscience*, Brentwood, Tennessee, USA. Cognitive Computational Neuroscience.

Machamer, P., Darden, L., and Craver, C. F. (2000). Thinking about mechanisms. *Philosophy of Science*, 67(1):1–25.

Mante, V., Sussillo, D., Shenoy, K. V., and Newsome, W. T. (2013). Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84.

Marcus, G. (2018). Deep learning: A critical appraisal. *arXiv*, pages 1–27.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., and Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, 11(1):1–12.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*, pages 1–12.

Millikan, R. G. (1984). *Language, Thought, and Other Biological Categories*. MIT Press.

Montúfar, G., Pascanu, R., Cho, K., and Bengio, Y. (2014). On the number of linear regions of deep neural networks. *Advances in Neural Information Processing Systems*, 4(January):2924–2932.

Naselaris, T. and Kay, K. N. (2015). Resolving Ambiguities of MVPA Using Explicit Models of Representation. *Trends in Cognitive Sciences*, 19(10):551–554.

Naselaris, T., Kay, K. N., Nishimoto, S., and Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, 56(2):400–410.

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., and Yamins, D. L. (2018). Task-driven convolutional recurrent models of the visual system. *Advances in Neural Information Processing Systems*, 2018-Decem(NeurIPS):5290–5301.

Nick, W., Shelton, J., Bullock, G., Esterline, A., and Asamene, K. (2015). Comparing dimensionality reduction techniques. *Conference Proceedings - IEEE SOUTHEASTCON*, 2015-June(June):0–1.

Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., and Kriegeskorte, N. (2014). A Toolbox for Representational Similarity Analysis. *PLoS Computational Biology*, 10(4).

Pestov, V. (2000). On the geometry of similarity search: dimensionality curse and concentration of measure. *Information Processing Letters*, 73:47–51.

Potochnik, A. (2017). *Idealization and the Aims of Science*. Chicago: University of Chicago Press.

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

Ritchie, J. B., Kaplan, D. M., and Klein, C. (2019). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *British Journal for the Philosophy of Science*, 70(2):581–607.

Rohwer, Y. and Rice, C. (2013). Hypothetical pattern idealization and explanatory models. *Philosophy of Science*, 80(3):334–355.

Rumelhart, D. E., McClelland, J. L., and PDP Research Group, C., editors (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1: Foundations*. MIT Press, Cambridge, MA, USA.

Rupert, R. D. (1998). On the Relationship between Naturalistic Semantics and Individuation Criteria for Terms in a Language of Thought. *Synthese*, 117(1):95–131.

Rupert, R. D. (2001). Coining Terms in the Language of Thought: Innateness, Emergence, and the Lot of Cummins's Argument against the Causal Theory of Mental Content. *The Journal of Philosophy*, 98(10):499.

Shea, N. (2007). Content and its vehicles in connectionist systems. *Mind and Language*, 22(3):246–269.

Shea, N. (2018). *Representation in Cognitive Science*.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., and Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587):484–489.

Slater, M. H. (2015). Natural kindness. *British Journal for the Philosophy of Science*, 66(2):375–411.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences*, 11(1):1–74.

Stinson, C. (2018). Explanation and Connectionist Models. *The Routledge handbook of the computational mind*, pages 1–17.

Stinson, C. (2020). From Implausible Artificial Neurons to Idealized Cognitive Models: Rebooting Philosophy of Artificial Intelligence. *Philosophy of Science*, (2019):1–38.

Tiffany, E. (1999). Semantics San Diego Style. *The Journal of Philosophy*, 96(8):416.

van der Maaten, L. and Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605.

van Gelder, T. (1995). What Might Cognition Be If Not Computation? *Journal of Philosophy*, 92(7):345–381.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., and Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137:188–200.

Wang, L., Hu, L., Gu, J., Wu, Y., Hu, Z., He, K., and Hopcroft, J. (2018). Towards Understanding Learning Representations: To What Extent Do Different Neural Networks Learn the Same Representation. *Advances in Neural Information Processing Systems*.

Weisberg, M. (2007). Three Kinds of Idealization. *Journal of Philosophy*, 104(12):639–659.

Woodward, J. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press.

Yamins, D. L. and DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, 19(3):356–365.

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

Zimek, A., Schubert, E., and Kriegel, H.-P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, 5(5):363–387.