ANNUAL REVIEWS

Annual Review of Neuroscience

Peeling the Onion of Brain Representations

Nikolaus Kriegeskorte[1] and Jörn Diedrichsen[2]

[1]Zuckerman Mind Brain Behavior Institute and Departments of Psychology, Neuroscience, and Electrical Engineering, Columbia University, New York, New York 10027, USA; email: n.kriegeskorte@columbia.edu

[2]Brain and Mind Institute and Departments of Computer Science and Statistical and Actuarial Sciences, Western University, London, Ontario N6A 3K7, Canada; email: jdiedric@uwo.ca

Annu. Rev. Neurosci. 2019. 42:407–32

The *Annual Review of Neuroscience* is online at neuro.annualreviews.org

https://doi.org/10.1146/annurev-neuro-080317-061906

Keywords

encoding, decoding, brain representations, neural code, pattern component model, representational similarity analysis

Abstract

The brain's function is to enable adaptive behavior in the world. To this end, the brain processes information about the world. The concept of representation links the information processed by the brain back to the world and enables us to understand what the brain does at a functional level. The appeal of making the connection between brain activity and what it represents has been irresistible to neuroscience, despite the fact that representational interpretations pose several challenges: We must define which aspects of brain activity matter, how the code works, and how it supports computations that contribute to adaptive behavior. It has been suggested that we might drop representational language altogether and seek to understand the brain, more simply, as a dynamical system. In this review, we argue that the concept of representation provides a useful link between dynamics and computational function and ask which aspects of brain activity should be analyzed to achieve a representational understanding. We peel the onion of brain representations in search of the layers (the aspects of brain activity) that matter to computation. The article provides an introduction to the motivation and mathematics of representational models, a critical discussion of their assumptions and limitations, and a preview of future directions in this area.

Contents

1. INTRODUCTION

1.1. The Representational Brain

Activity pattern: vector of activities for a single experimental condition (e.g., a perceptual stimulus) across measurement channels (e.g., neurons or voxels)

Representation: brain-activity pattern that serves the purpose of conveying information that specifies perceptions, thoughts, actions, or any other mental content in the context of the brain's overall function

Our brains give rise to a continuous stream of mental activities. We perceive things, have emotions, think thoughts, make decisions, and act. These mental activities are sustained by neurons activated in a dynamic weave of complex patterns. As neuroscientists, we try to understand the functional mechanism of this neural activity: how it enables us to interact with the world. We are therefore interested in how things in the world are reflected in the activity in our brains.

We interpret neural activity patterns as serving the function of conveying information about the world (Brentano 1874, Dennett 1987, Shea 2018). The content could be information about our environment, acquired through the senses, or any other mental content, such as thoughts, goals, plans, or actions. Beyond the mere presence of the information in the neural activity, a representational interpretation implies that the information is used by downstream neurons in a way that contributes to behavior (Millikan 1989, Kriegeskorte & Bandettini 2007, Shea 2018). We can test this hypothesis experimentally by manipulating the activity and studying the effects on behavior (Salzman et al. 1990, Afraz et al. 2006, Parvizi et al. 2012).

We could avoid representational interpretations altogether and approach the brain as a dynamical system (Bechtel 1998, Van Gelder 1998, Churchland et al. 2012, Shenoy et al. 2013). The dynamical systems perspective is fundamental (in that it captures what the brain does at the level of physical mechanism) and complete (in that it should be able to account for all aspects of brain function). However, the concepts of information and representation can help us understand the function of neuronal dynamics at a higher level of description (Dennett 1987, Shea 2018).

Consider the case of computers: They too can be understood as dynamical systems. However, interpreting the patterns of charges and currents as representations of data and instructions enables us to capture a computer's behavior more concisely in a high-level algorithmic description that reveals the dynamics in terms of the implemented functions. Like a computer, the brain is a dynamical system, and representational accounts can help us cope with its complexity.

1.2. Encoding and Decoding Models

Encoding models and decoding models (Paninski et al. 2007) attempt to capture, respectively, the causal process that gives rise to a representation and the causal process by which it might be read. Ideally, then, an encoding model should take stimuli as input and predict brain responses (Wu et al. 2006), and a decoding model should take brain responses as input and predict downstream brain or behavioral responses. Note, however, that a decoder, in this conceptualization, is not the inverse of an encoder.

In practice, decoders are often conceptualized as inverse encoders (Rieke et al. 1999, Cox & Savoy 2003, Hung et al. 2005, Kriegeskorte 2011, Tong & Pratte 2012, Carlson et al. 2013, Cichy et al. 2014, King & Dehaene 2014), mapping from brain responses back to the stimuli, rather than on to downstream brain or behavioral responses. The decoding model, then, serves not as a process model of brain computation, but rather as a tool of analysis that can help reveal what information is present in the code and in what format (Kriegeskorte & Douglas 2018b). To understand how the brain computes, we need to build process models of brain computation and test how well they can account for behavioral performance and brain activity (Kriegeskorte & Diedrichsen 2016, Kriegeskorte & Douglas 2018a).

Let us say that we have built a neural network model that can perform some cognitive task of interest. How do we assess whether it is a good model of how a brain performs the task? We need a way of comparing not just the architecture (the anatomy), but also the activity patterns (the physiology) between brain and model (Kriegeskorte 2015, Yamins & DiCarlo 2016, Kriegeskorte & Douglas 2018a). A detailed comparison of the internal representational spaces can be achieved using representational models (Diedrichsen & Kriegeskorte 2017).

We assume that the activity elicited by each stimulus (or, more generally, each experimental condition) in each measured response channel is estimated as a scalar. The scalar activity level could be defined as the firing rate (for neuronal recordings) or as the voxel response [for functional magnetic resonance imaging (fMRI)], averaged across repetitions of the same stimulus. We interpret each activity pattern across the neurons or voxels as representing the stimulus that elicited it. To test our neural network model, then, we must compare the representation in each of its layers to that in the corresponding cortical area while model and brain are processing the same stimuli. A good neural network model of brain information processing should recapitulate the representational transformations across stages of processing (Güçlü & van Gerven 2015). A model of the ventral visual stream, for example, should disentangle across successive stages the representational manifolds corresponding to different categories of object (DiCarlo & Cox 2007).

We consider in detail three types of representational model that can be used to test brain-computational models: (*a*) encoding models (Dumoulin & Wandell 2008, Kay et al. 2008, Mitchell et al. 2008, Naselaris et al. 2011, Naselaris & Kay 2015), (*b*) pattern component models (PCMs) (Diedrichsen et al. 2011, 2018), and (*c*) representational similarity analysis (RSA) (Kriegeskorte et al. 2008a, Kriegeskorte & Kievit 2013, Nili et al. 2014). An encoding model predicts each measured response channel as a linear combination of units of the brain-computational model. The other two types of representational model predict summary statistics of the representation, so as to simplify the inference. Each of them can enable us to adjudicate among task-performing

Encoding model: a model that predicts a brain-activity pattern from a description of the experimental condition in terms of a set of features

Decoding model: a model that predicts a feature from a brain-activity pattern

Representational model: a mathematical model that specifies a probability distribution over the space of activity profiles

Pattern component model (PCM): an approach to statistical inference on representational models that compares models to data in terms of the second moment of the activity profiles

Representational similarity analysis (RSA): an approach to statistical inference on representational models that compares models to data in terms of representational dissimilarity matrices

computational models (Kriegeskorte & Diedrichsen 2016, Diedrichsen & Kriegeskorte 2017) by comparing model representations to representations in cortical areas.

Researchers have developed many other methods for analyzing neural data, including dimensionality reduction methods (Cunningham & Yu 2014) and hybrid methods that attempt to explain activity data using a small number of components that relate to the stimuli (Kobak et al. 2016). Although the three types of representational model form the core of this article, we consider a broad range of representational analyses, from data driven (e.g., mapping selectivities across the cortical sheet) to hypothesis driven (e.g., testing for the presence of particular information with a decoder). Our goal is to clarify the relationships among all of these analyses by considering what information each discards.

1.3. The Onion of Brain Representations

We can think of the different aspects of brain representations as the layers of an onion (**Figure 1**). On the surface of the onion are the characteristics that first meet the eye when we look at the data. Functional imaging gives us spatial activity patterns across the cortex, with different techniques,

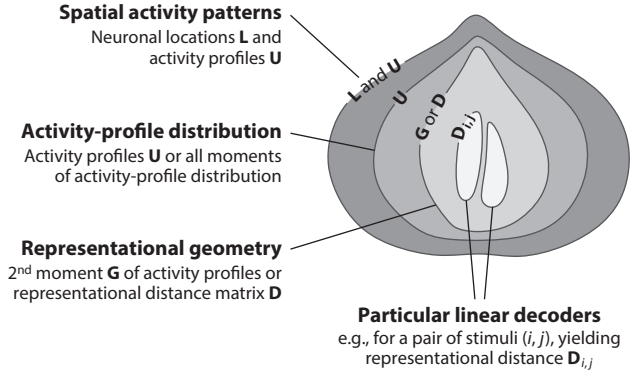


Figure 1

Peeling the onion. What information should researchers use to characterize brain representations? When we consider cortical maps of neuronal selectivity, we care about the activity profiles (tuning functions) **U** as well as the locations **L** of the neurons (entire onion). We can remove from our data the spatial coordinates specifying where in the map each neuron resides, thus peeling away the outer layer of the onion. This leaves us with the set of activity profiles. The distribution of the activity profiles specifies the prevalence in the neuronal population of each tuning function and, together with a noise model, determines both the content and format of the neuronal code, defining, for example, to what extent particular information is concentrated in a small number of neurons or distributed across the region. We can abstract from the activity profiles and consider only the representational geometry, which defines how separated the stimuli are in the multivariate response space. The representational geometry is fully and equivalently characterized by either the representational distance matrix **D** or the second moment of the activity profiles **G**. Different distributions of profiles can give rise to the same representational geometry (corresponding to rigid rotations and translations of the ensemble of response patterns in the space spanned by the response-channel activities). We can remove the information about the particular distribution of profiles and keep only **D** or **G**, the sufficient statistics of the geometry, thus peeling away the second layer. If the noise is isotropic and homoscedastic, and the noise distribution is known to us, then the representational geometry defines all encoded information, i.e., the mutual information between any stimulus feature (graded or categorical property of each stimulus) and the response pattern. In particular, the geometry defines how well a given feature or pairwise distinction between stimuli can be decoded. The geometry can be conceptualized as being composed of the linear dimensions of the multivariate response space (kernels of the onion), which are selectively explored in linear decoding analyses. By focusing on particular linear decoders, we peel the onion further, removing the other dimensions of the representational geometry.

such as fMRI and calcium imaging, enabling measurements at different spatial and temporal scales in humans and animals. Functional imaging reveals retinotopic maps in the visual cortex (Tootell et al. 1998), tonotopic maps in the auditory cortex (Formisano et al. 2003), and somatotopic maps in the sensory-motor cortices. The presence of an orderly spatial map for an encoded variable is sometimes taken as essential evidence that an area represents that variable.

Spatial organization matters to brain function. Although it is network topology, the connectivity graph, that defines the information flow (Felleman & Van Essen 1991), the costs of connection (in terms of axons, energy, and time lags) scale with the physical distance between two neurons. Network geometry, therefore, must constrain network topology to some extent (Chklovskii & Koulakov 2004). This motivates us to care about the spatial layout of functionally specialized components, both at the scale of the entire brain and at the scale of maps within cortical areas.

However, the precise layout of the maps at the columnar scale likely also reflects random developmental variation (Ejaz et al. 2015, Wilson & Bednar 2015, Diedrichsen 2019). Moreover, as illustrated by the example of *Caenorhabditis elegans* (Bargmann & Marder 2013), even knowing the locations of all the neurons and their connectome leaves mysterious the computational mechanism (Poeppel 2012). Might we be losing sight of the forest of computational function for the trees of individual neurons and their precise locations and tuning?

We can peel away the superficial layers of the onion, abstracting from the locations and tuning of individual neurons, to reveal the information that a population of neurons renders explicit for readout by downstream neurons. Deep inside the onion is the information accessible to linear decoders (Rieke et al. 1999, DiCarlo & Cox 2007). A linear decoder computes a weighted sum of the activities that it takes as its input. Because this is a biologically plausible computation for a single neuron, linear decoders let us probe what information a downstream neuron might extract from a neural population. A neural population that supports linear decoding of particular information is sometimes interpreted as an "explicit representation" of that information (deCharms & Zador 2000, Kriegeskorte 2011, DiCarlo et al. 2012, Hong et al. 2016)—with the tacit assumption that downstream neurons actually do use the information.

Linear decoders are widely used to analyze data from cell recording and functional imaging (Rieke et al. 1999; Haxby et al. 2001, 2014; Carlson et al. 2003; Cox & Savoy 2003; Hung et al. 2005; Kamitani & Tong 2005; Haynes & Rees 2006; Kriegeskorte et al. 2006; Norman et al. 2006; Mur et al. 2009; Pereira et al. 2009; Tong & Pratte 2012; Haynes 2015; Hebart & Baker 2017; Varoquaux et al. 2017; Kriegeskorte & Douglas 2018b). This multivariate approach contrasts with the univariate analyses of single sites of brain-activity measurement employed in both single-cell selectivity studies in animals and brain mapping studies in humans. Multivariate decoding lets us focus on the forest (the population code) and summarize the information conveyed by all of its trees (neurons) together.

Decoding strikes straight at the core of the onion, at the encoded information, stripping away intermediate layers that may deserve consideration. For example, a linear decoder does not tell us how the information is distributed across the population and what tuning single neurons exhibit. Might we be abstracting too aggressively when considering only the degree to which a particular variable can be linearly decoded?

To summarize the representational content, perhaps we should consider all possible dimensions of linear readout. As we discuss below, a sufficient set of linear projections of the multivariate response space defines the representational geometry (Kriegeskorte et al. 2008a). The representational geometry determines the total information content of the code and the format in which it is encoded, up to a translation and rotation of the response-pattern ensemble in the space spanned by the response channels. Enclosing this layer is the distribution of activity profiles (tuning functions), which additionally defines the set of axes that span the multivariate

Distance: a function mapping each pair of vectors in a space onto a nonnegative real number (includes metrics, and also the correlation distance, squared Euclidean distance, and Minkowski distances with $p < 1$)

Explicit representation: a representation of content in a format that enables it to be decoded in a single step by biological neurons

Representational geometry: the geometry of the ensemble of condition-related activity patterns in the multivariate response space, as defined by all pairwise representational dissimilarities

Activity profile: vector of activities for a single measurement channel (e.g., a neuron or voxel) across different experimental conditions (e.g., perceptual stimuli)

response space (Diedrichsen et al. 2018). The distribution of activity profiles defines to what extent a given variable is represented in a sparse or distributed way across the neurons (Simoncelli & Olshausen 2001, Olshausen & Field 2004).

***P*:** number of measurement channels (neurons, electrodes, voxels)

***K*:** number of experimental conditions

U: $K \times P$ matrix of activity profiles

Our goal in this review is to organize the different aspects of brain-activity measurements into a nested hierarchy: the onion. We argue that popular data-analysis methods in neuroscience correspond to successive peeling stages of the onion: from (*a*) spatial mapping of selectivities across the cortex, to (*b*) characterizing the distribution of neuronal tuning functions, to (*c*) investigating representational geometries, and finally to (*d*) decoding analyses. Rather than proposing that the onion be peeled to a particular layer, we aim to clarify the abstraction gained and the information lost with each step. First, we peel the onion layer by layer, from the outside toward the core. Then we proceed in the opposite direction. As we put the onion back together, we consider how adding layers back in can help us focus on portions of the encoded information that are more likely to be actually read out by neurons downstream.

2. PEELING THE ONION: FROM THE OUTSIDE IN

2.1. Entire Onion

Consider a wonderful data set: The activity of a large sample of P neurons within a cortical area has been measured in a large number K of experimental conditions. A scalar activity level has been estimated for each neuron and condition (e.g., the windowed spike count, averaged across trials of each condition). The activity estimates have been assembled in a large matrix **U** (K conditions by P neurons).

In the context of sensory systems, the experimental conditions will correspond to stimuli, and the activity measurements are typically referred to as responses. We use this terminology to convey a more concrete intuition, although the concepts that we describe are also applicable to cognitive and motor representations. The activity matrix **U**, then, is the stimulus–response matrix. For each neuron, **U** provides the activity profile in one of its columns (**Figure 2*a***). A neuron's activity profile enables us to characterize its tuning to different stimulus properties or its selectivity for different categories of stimuli.

In addition, we are given the locations matrix **L** of the neurons (P neurons by two or three coordinates defining the location of each neuron on a two-dimensional cortical flatmap or in three-dimensional brain space). **U** and **L** together define the entire onion (**Figure 1**). Using **L**, we could make a cortical map of selectivity, color coding neurons, for example, by the category of the experimental stimuli in which they are most active.

Assume, for the moment, that the neurons are affected by additive noise that is independent and identically distributed for each neuron and stimulus. We see below that this requirement can be relaxed (see the sidebar titled Dealing with Poisson and Correlated Noise). A neuron's activity profile then defines the accuracy with which we can discriminate any two stimuli from the neuron's activity. In conjunction with a noise model, **U** defines what information about the stimuli the code contains, and also how the information is encoded. **L** tells us how the neurons are laid out in the cortex.

2.2. Distribution of Activity Profiles

To focus on the information represented in the region, we might decide to disregard the locations **L** of the neurons. Removing **L** amounts to peeling off the outer layer of the onion. We are left with **U**, the set of activity profiles. We have lost none of the information in the code, only where the neurons were in the cortex (**Figure 3*a***).

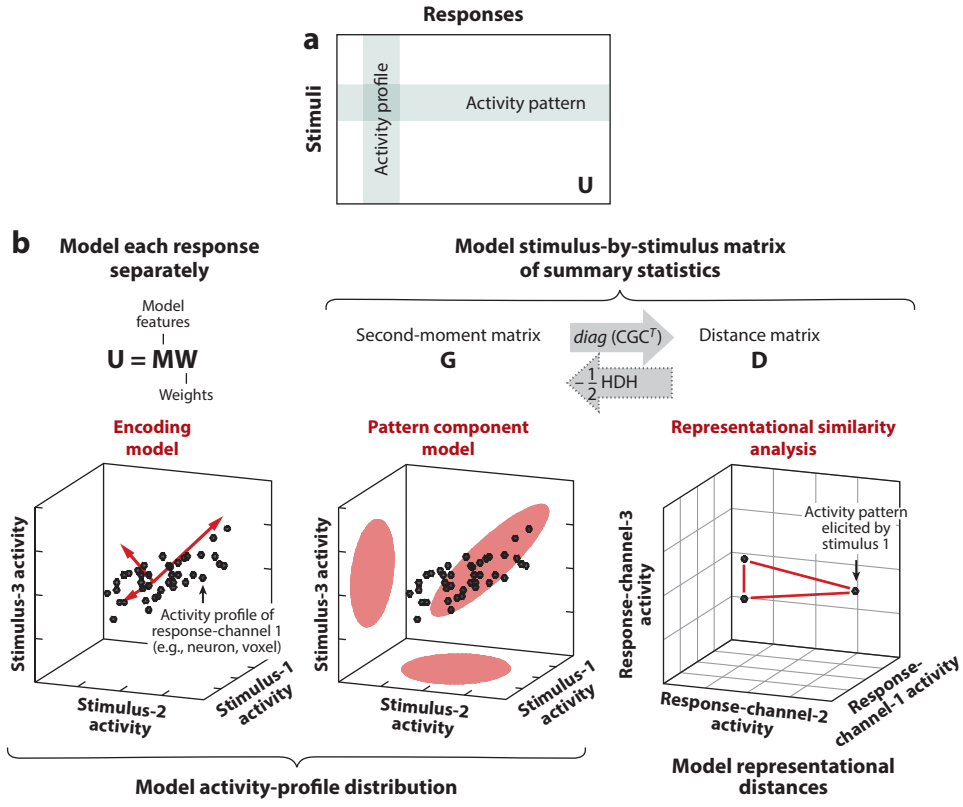


Figure 2

Three methods for specification and testing of representational models. (*a*) Representational models capture the relationship between stimuli (or, more generally, experimental conditions) and response channels (e.g., neurons or voxels). The stimulus–response matrix **U** contains the activity for each stimulus and response channel. (*b*) Representational models can be specified and tested using any of three methodological approaches: encoding models (*left*), pattern component models (PCMs) (*middle*), and representational similarity analysis (RSA) (*right*). An encoding model predicts each response as a linear combination of the activity profiles of a set of model features (*red arrows*), typically using a 0-mean Gaussian prior on the weights. A PCM predicts the distribution of activity profiles as a Gaussian distribution, characterized by the second moment **G** of the activity profiles. RSA predicts the geometry of the activity patterns, as characterized by the representational distance matrix **D**. Encoding models and PCMs, thus, target the distribution of activity profiles (columns of the matrix **U** in panel *a*; each axis in panel *b* represents the activity elicited by one stimulus), whereas RSA targets the geometry of the activity patterns (rows of the matrix **U** in panel *a*; each axis in panel *b* represents the activity elicited in one response channel). PCMs and RSA characterize a representation by a stimulus-by-stimulus matrix of summary statistics (**G** and **D**, respectively, which can be derived from each other), whereas encoding models characterize each response channel individually. When a 0-mean isotropic Gaussian weight prior is used for the encoding model (equivalent to ridge regression), all three methods test hypotheses captured by the second moment of the activity profiles **G** as a sufficient statistic (Diedrichsen & Kriegeskorte 2017).

The activity profiles capture the kind of information gathered by an electrophysiologist who records from one cell at a time with a tungsten electrode: She can characterize the tuning of many neurons, although she may not have precise information about where in the targeted cortical area the neurons were located. She might report the proportions of neurons exhibiting different types of tuning.

DEALING WITH POISSON AND CORRELATED NOISE

The representational geometry is equivalently defined by either the distance matrix or the second moment of the distribution of the activity profiles. These sufficient statistics define the decodability of any possible contrast of the experimental conditions (e.g., any stimulus category or continuous feature) when the noise on the activity is additive, independent and identically distributed across neurons (or measured response channels). These assumptions, however, are usually violated. In electrophysiological recordings, the trial-by-trial variability of neuronal responses is roughly proportional to their firing rate. Although the Fano factor (variance over mean) can deviate from 1, the variability of neuronal responses is often approximated as a Poisson process. Poisson variability can be handled by using a variance-stabilizing transform of the data: a nonlinear monotonic transformation that renders the variability of the firing independent of the firing rate. An approximate solution that is robust and efficient is using the square root of the instantaneous firing rates (Yu et al. 2009). Noise correlations (Abbott & Dayan 1999, Averbeck et al. 2006, Moreno-Bote et al. 2014) between recorded neurons or fMRI voxels can be dealt with following a similar logic. We assume that the noise is additive and multinormal, estimate the $P \times P$ error covariance matrix $\mathbf{\Sigma}$, and prewhiten the noise by transforming the patterns $\mathbf{U}$ into $\mathbf{U} \cdot \mathbf{\Sigma}^{-1/2}$. This transform renders the noise approximately independent and identically distributed across measurement channels (Walther et al. 2016). The Fisher linear discriminant includes this whitening transform. When all training and test patterns have been whitened, the Fisher linear discriminant $\mathbf{w}_{\text{Fisher}} = (\mathbf{r}_j - \mathbf{r}_i)^T \cdot \mathbf{\Sigma}^{-1}$ reduces to the difference $\mathbf{w}_{\text{Fisher}}^{\text{white}} = (\mathbf{r}_j^{\text{white}} - \mathbf{r}_i^{\text{white}})^T$ between the two activity patterns. Note that $\mathbf{w}_{\text{Fisher}} \neq \mathbf{w}_{\text{Fisher}}^{\text{white}}$. The former includes the whitening transform of the test data, whereas the latter assumes that the test data have been whitened.

We can view the set of activity profiles (the columns of $\mathbf{U}$) as a distribution. Consider the space of all possible activity profiles. Each axis corresponds to the activity elicited by a given stimulus. Each neuron is a point in this space, the coordinates of which specify the neuron's activity profile. The neurons, with their particular activity profiles, are scattered across this space (**Figure 2*b***, left and center).

The measured neurons are typically a sample from the neuronal population in the area that we are investigating. Our object of study might be a particular cortical area (say, area V1, MT, or M1) in a given species of animal. We can view the measured neurons as a sample from an idealized distribution associated with the cortical area under study. We may not need a comprehensive sample to characterize the representational space (Ganguli & Sompolinsky 2012, Gao & Ganguli 2015).

The activity-profile distribution $p(\mathbf{u})$ is a continuous probability density function over the space of activity profiles. It defines how well we can linearly discriminate any two stimuli given all the responses measured across neurons. Beyond pairs of stimuli, we can attempt to decode any feature (graded property of each stimulus) from the measured set of neurons. A feature corresponds to a direction in the space spanned by the stimuli. To the extent that the activity profiles of the measured neurons can be linearly combined to predict the feature, linear decoding will be successful. The noise and $\mathbf{U}$ together determine how well any feature can be decoded.

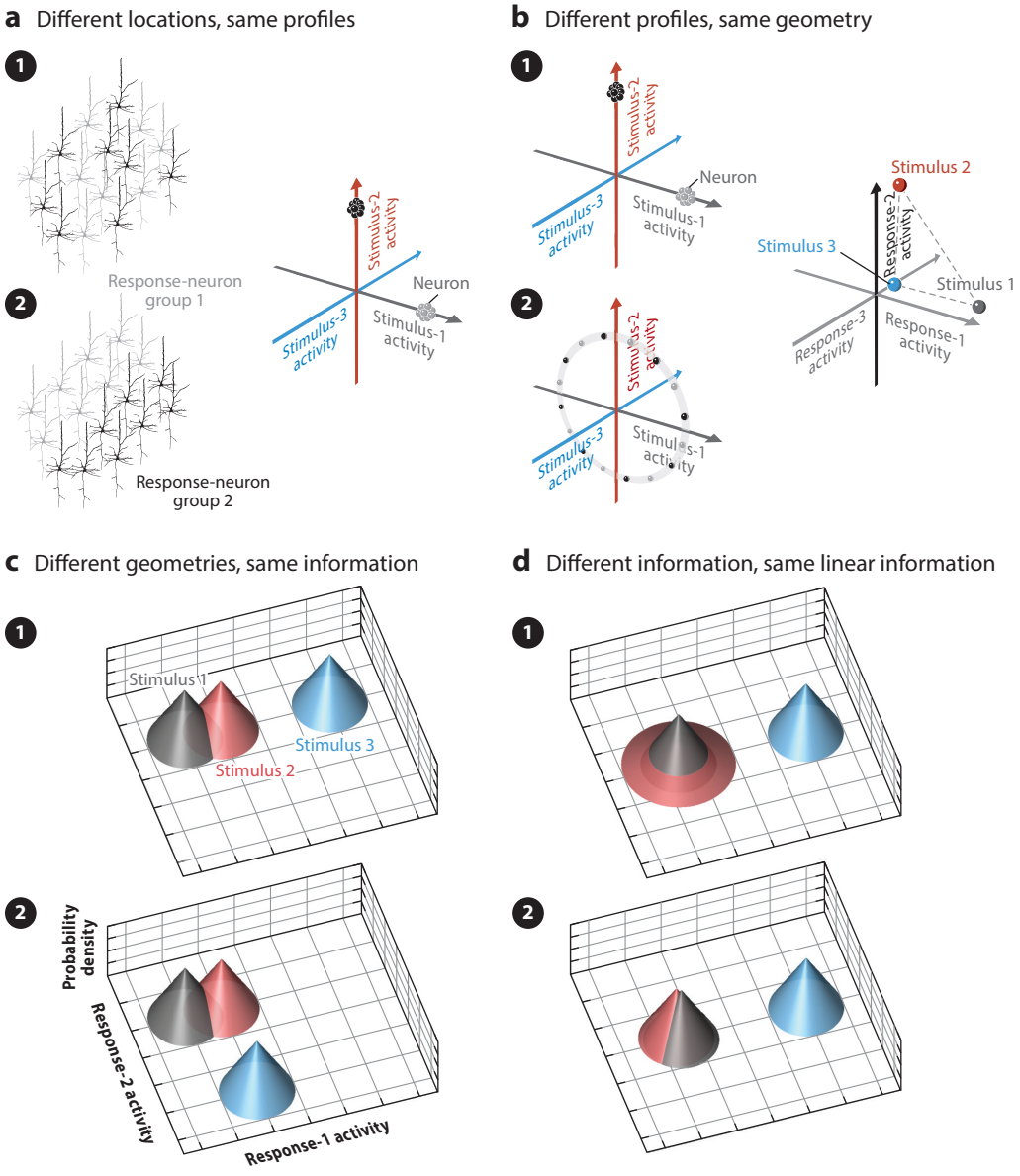
Feature: a scalar descriptor (e.g., a property or a category membership) for each experimental condition (e.g., each stimulus)

In conjunction with a noise model, such a probability density function $p(\mathbf{u})$ comprehensively captures the code, specifying both the information content and the representational format. We therefore define a representational model as a probability density function $p(\mathbf{u})$ over the space of activity profiles. Defining the activity profile as a random variable drawn from the activity-profile distribution $p(\mathbf{u})$ extends the classical framework of encoding models (Wu et al. 2006), which treats the responses as a fixed set given by the measurements.

One way to model the distribution of activity profiles is to specify a basis set of profiles (i.e., features) and assume that the measured activity profiles are linear combinations of this basis set. This particular type of representational model is called an encoding model (**Figure 2*b***, left) (Dumoulin

& Wandell 2008, Kay et al. 2008, Mitchell et al. 2008, Naselaris et al. 2011, Naselaris & Kay 2015, van Gerven 2017).

A basis set of profiles does not uniquely specify a distribution of profiles. However, most encoding models assume a Gaussian prior over the weights of the linear combination. Drawing the weights from a Gaussian prior, a basis set of profiles induces a Gaussian distribution of profiles $p(\mathbf{u})$. In practice, the Gaussian prior over the weights is often assumed to have a diagonal covariance matrix, which implies that the weights are uncorrelated (Kay et al. 2008; Huth et al. 2012, 2016). Each basis profile can capture variation among neurons along some oblique dimension of the space spanned by the stimulus-elicited activities (**Figure 2*b***), so the induced Gaussian



(*Caption appears on following page*)

Figure 3 (*Figure appears on preceding page*)

Understanding the layers: altering one without altering the next. (*a*) Changing neuronal locations does not alter the code or format. In this example, a population of neurons that fall into groups 1 (*gray*) and 2 (*black*) selectively respond to stimulus 1 and stimulus 2, respectively. If neurons were spatially rearranged in the brain (*i*, *ii*) but kept their activity profiles (tuning), then the code would carry the same information in the same format. (*b*) Example of a change of the distribution of activity profiles that does not alter the representational geometry. (*i*) Two groups of neurons, each preferring a different stimulus. (*ii*) Each neuron now prefers a different combination of stimuli 1 and 2, but the population spans the same subspace of activity profiles and equally weights all directions in this subspace. As a result, these very different distributions of activity profiles give rise to the same representational geometry. More generally, if the stimulus–response matrix is postmultiplied by an orthogonal matrix (rotating and/or reflecting the response-pattern ensemble rigidly in the multivariate response space), then the profiles are altered in a way that conserves the representational distances. (*c*) Example of a change to the representational geometry that does not alter the encoded information. The representational geometry determines the encoded information and aspects of the format of the code, such as linear decodability. In this case, by moving stimulus 3 (*blue*) in the multivariate response space, we have altered the geometry (*i*, *ii*). However, because the overlaps among the probability density functions do not change, this does not alter the encoded information. (*d*) Example of a change to the encoded information that does not alter the linearly decodable information. In both scenarios, stimulus 1 (*gray*) and stimulus 2 (*red*) elicit the same response pattern on average. (*i*) The code distinguishes the two stimuli to some extent because stimulus 2 has a wider noise distribution. This information is not amenable to linear decoding, but a radial-basis function readout could extract it. (*ii*) Stimuli 1 (*gray*) and 2 (*red*) have identical distributions (shifted slightly to make both surfaces visible), so no information about the distinction between stimuli 1 and 2 is encoded. The linearly decodable information is unaltered between subpanels *i* and *ii*. Note that, under isotropic, homoscedastic noise, the representational geometry determines both the encoded information and the linearly decodable information, and the former cannot be altered without altering the latter.

distribution of activity profiles can be nonisotropic and elongated along oblique directions, capturing, for example, that neurons driven by stimulus A might tend to also be driven by stimulus B.

If we assume the distribution of activity profiles to be Gaussian, then we can alternatively characterize it by its sufficient statistic, the second moment of the activity profiles **G** (*K* stimuli by *K* stimuli; see Equation 6 in Section 6). **G** is a matrix that captures, for each pair of stimuli A and B, to what extent neurons responsive to A are also responsive to B. A particular type of representational model, called a pattern component model (PCM) (**Figure 2*b***, middle) (Diedrichsen et al. 2011, 2018), uses **G** to characterize the distribution of activity profiles.

In brains, the distribution of activity profiles may seldom be Gaussian. For non-Gaussian profile distributions, **G** is not a sufficient statistic, and higher-order moments are required to fully define the distribution. An important area that needs to be further developed is the use of non-Gaussian models of the activity-profile distribution (Norman-Haignere et al. 2015). However, even if the distribution is not Gaussian, it can still be helpful to model it using linear combinations of basis activity profiles or to characterize it to a first approximation by its second moment **G**. We see below that, whether or not the activity-profile distribution is Gaussian, **G** (together with a noise model) defines the linearly decodable information.

2.3. Representational Geometry

Above, we considered the stimulus–response matrix **U** in terms of its columns, the activity profiles, which reflect neural tuning. We can alternatively consider **U** in terms of its rows, the activity patterns elicited by the stimuli (**Figure 2*a***). We can plot the activity patterns as points in the space spanned by the response channels (**Figure 2*b***, right). Each axis of this space corresponds to the activity of one of the neurons. Each point corresponds to the response pattern elicited by a stimulus.

Whether we plot the activity estimates in **U** in terms of the columns or the rows (**Figure 2*b***, left and right, respectively), we are plotting the same values and visualizing the same information, albeit from a different perspective. Plotting the activity profiles (**Figure 2*b***, left) reveals groups of neurons with similar tuning. Plotting the activity patterns (**Figure 2*b***, right) reveals groups of stimuli that are represented similarly. The latter perspective suggests that we should think of the

representation in terms of which stimuli it renders dissimilar, which it renders similar, and which it renders indiscriminable (Edelman 1998, Edelman et al. 1998).

If we assume additive Gaussian noise that is independent and identically distributed across neurons (isotropic) and stimuli (homoscedastic), then the Euclidean distance in the multivariate response space precisely defines the discriminability of a pair of stimuli in the representation. For two stimuli, the line passing through the two mean response patterns defines the best linear discriminant dimension. The dimensions orthogonal to this discriminant dimension carry no information discriminating the stimuli. Projecting the distributions onto the discriminant dimension yields two equal-variance univariate Gaussians, which capture all the information discriminating the stimuli. Their overlap is the error rate of the Bayes-optimal decoder and could be characterized by the sensitivity index d'. The accuracy of a linear decoder rises monotonically with the Euclidean distance between the two stimuli being discriminated. In practice, the noise is rarely additive, isotropic, and Gaussian. However, the conceptual points that we make in this section generalize to correlated multinormal and Poisson noise, which we can handle by transforming the responses. We discuss the issue of correlated noise in greater detail in the second half of the review, when putting the onion back together (see also the sidebar titled Dealing with Poisson and Correlated Noise).

Since representational distance reflects discriminability, we can characterize the representation by the matrix **D** (K stimuli by K stimuli; see Equation 7 in Section 6) of Euclidean distances between stimuli in the multivariate response space. The distance matrix defines the metric relationships between the stimuli in the representational space: the representational geometry. With all pairwise distances defined, none of the stimuli can move relative to the others.

The representational distance matrix is also known by its more general name, the representational dissimilarity matrix (RDM). The concept of dissimilarity includes Euclidean and Mahalanobis distances, along with other dissimilarity measures that do not conform to the mathematical definition of metric (which requires the triangle inequality to hold) or even to the looser definition of distance (which requires nonnegativity). Dissimilarity measures include the correlation distance (which can return 0 for nonidentical patterns; Equation 8 in Section 6), the squared Mahalanobis distance (which does not conform to the triangle inequality); and unbiased estimators (which can be negative), such as the crossnobis distance estimator, which we introduce below (Equation 3). RDMs provide the signatures used to compare representations between brains and models in RSA (**Figure 2*b***, right), another method for testing representational models (Kriegeskorte et al. 2008a,b; Kriegeskorte & Kievit 2013; Nili et al. 2014; Kriegeskorte & Diedrichsen 2016).

The representational geometry is defined by its sufficient statistic: the distance matrix **D**. The representational geometry determines not only how far any two stimuli are in the representation and how well they can be discriminated in the presence of noise, but also how well any feature (continuous or categorical stimulus property) can be linearly decoded from the representation. A linear decoder projects the representational geometry (the ensemble of patterns) onto an axis in multivariate response space (**Figure 2*b***, right). Say we set out to decode feature **f**, where **f** is a vector with a property value for each stimulus. The best linear decoder for **f** defines the axis projection onto which best matches **f**. The vector **w** defining this axis contains the weights with which we can linearly combine the activity profiles (**Figure 2*b***, left) to best recover the feature values.

The distance matrix **D** does not merely define how well any feature can be decoded linearly. It also defines how well any feature can be decoded nonlinearly (as long as the decoder can see the entire set of response channels and, like a linear decoder, is capable of translations and rotations in the space spanned by the response channels). In conjunction with an isotropic noise model, the distance matrix **D** defines the joint probability distribution $p(s, \mathbf{r}) = p(\mathbf{r}|s) \cdot p(s)$ of response

Representational dissimilarity: a measure of the dissimilarity between two brain-activity patterns interpreted as representations (includes metrics, distances, and dissimilarity estimators that are not either metrics or distances)

Metric: a function mapping each pair of vectors in a space onto a nonnegative real number, where 0 is returned if and only if the vectors are identical, the same number is returned if the vectors are swapped, and the numbers conform to the triangle inequality

patterns **r** and stimuli *s*, up to a rotation and translation of the ensemble of response patterns in the space spanned by the response channels (**Figure 2*b***, right). We assume in this case that $p(s)$ is a uniform prior over the stimuli. The total information that the representation contains about the stimuli is

$$I(s; \mathbf{r}) = \sum_s \int p(s, \mathbf{r}) \cdot log \frac{p(s, \mathbf{r})}{p(s)p(\mathbf{r})} \mathrm{d}\mathbf{r}. \quad 1.$$

When the noise distribution around each stimulus-related response pattern is isotropic, $I(s; \mathbf{r})$ is invariant to rigid rotations and translations of the ensemble of response patterns and depends only on the Euclidean distance matrix **D** and the noise. Similarly, for nonisotropic homoscedastic multinormal noise (Abbott & Dayan 1999, Averbeck et al. 2006, Moreno-Bote et al. 2014), $I(s; \mathbf{r})$ depends only on the Mahalanobis distance matrix.

If the representational geometry captures all of the information encoded in the representation, then have we really lost any information in going from the distribution of activity profiles to the geometry? The answer is yes. Although we have lost none of the encoded information about the stimuli, we have peeled away a layer of the onion that defines how the information is distributed over the neurons (**Figure 3*b***). For example, an equal representational distance could arise from large response differences in few neurons (localized, sparse) or from small response differences in many neurons (distributed). Similarly, two representational distinctions could be represented by disjoint or overlapping sets of neurons. This is illustrated in **Figure 3*b***.

Two representations of the same geometry could thus look very different to a single-cell electrophysiologist sampling cells with an electrode. To the extent that we sample a small number of neurons and analyze them separately, we will be more sensitive to distinctions that are strongly reflected in single-neuron responses. Focusing instead on the multivariate representational geometry makes us equally sensitive to information that is localized to a small subset of the region or distributed across the region. In fact, it makes us oblivious to the difference between them.

The information lost in going from distributions of activity profiles to representational geometry concerns rotations and translations of the ensemble of activity patterns in the space spanned by the response channels. We can think of the geometry as a rigid construction of points (stimuli) and edges (distances). The geometry remains unaltered as it is rigidly translated or rotated, but the activity-profile distribution changes in the process.

Like the second moment of the activity profiles **G**, the distance matrix **D** is a stimulus-by-stimulus matrix that summarizes a subset of the information defining the distribution of activity profiles. It turns out that there is a close relationship between these two matrices. We can think of **G** as a set of inner products among activity patterns, one for each pair of stimuli. The inner products can be interpreted as measuring the representational similarity for each pair of stimuli, whereas the distances measure the representational dissimilarity. This suggests that the two matrices might capture similar information.

The inner product of each pattern with itself reflects the squared norm of the pattern (i.e., the squared Euclidean distance from the origin). Each inner product for a pair of stimuli is proportional to the cosine of the angle spanned by the two activity patterns about the origin (and to the norms of the two patterns). By defining all pairwise angles and all distances from the origin, the second moment matrix **G**, in fact, fully defines the representational geometry. The distance matrix **D** can be computed from **G** (see Equation 9 in Section 6). If we add a baseline pattern before computing **D** or center the pattern ensemble in response space before computing **G**, then **G** can also be computed from **D** (see Equation 10 in Section 6). **G** and **D**, thus, capture equivalent information.

The equivalence of the second moment of the activity profiles **G** and the distance matrix **D** shows that the three methods for testing representational models—encoding models, PCMs, and RSA—are closely related. The former two focus on the activity profiles and thus could exploit the additional information that their distribution contains about the format of the code. In their typical implementations, however, they in fact predict Gaussian distributions of profiles the sufficient statistic of which is **G** (or, equivalently, **D**). All three, thus, compare representational models in terms of the representational geometries that they predict (Diedrichsen & Kriegeskorte 2017).

Measurement channel: a localized brain response for which scalar activity measurements have been performed, such as a voxel (fMRI hemodynamic response) or a neuron (extracellular electrode recording)

2.4. Decoding Analyses

Representational models make comprehensive predictions about the representational space. A decoding analysis merely asks whether a particular feature is accessible to a particular decoder. By focusing on particular features and particular decoders, we peel off more of the onion, ignoring other features that might be encoded, as well as other decoders that might be able to access more information. Representational models are generative models that predict representational spaces (at the level of the measured response channels in encoding models, the second moment of the activity profiles in PCMs, and the representational distances in RSA). Decoders are discriminative models that extract particular information. Because of their more limited focus, decoders, in general, provide weaker constraints for computational theory.

The most prominent type of decoder is the linear decoder, in which linear combinations of measurement channels (voxels or neurons) serve as discriminant functions. The motivation for testing for linear decodability derives from the notion that any information that can be linearly decoded is amenable to direct readout by downstream neurons that get input from the entire code. Above, we have seen that the representational geometry captures the content and format (up to a linear transform) of the representation. The distance matrix **D** tells us the discriminability of each pair of stimuli. If we want to know whether a given categorical division or feature **f** is linearly represented, we can inspect the quadratic form $\mathbf{f}^T\mathbf{G}\mathbf{f}$, which will reflect the accuracy of linear decoding within our sample of stimuli. Fitting a linear decoder makes the readout explicit, in terms of the required weighted combination, and enables us to estimate the accuracy that the decoder achieves on a test set, which could consist of responses to different stimuli.

With a view to understanding brain computation, researchers take the perspective of readout neurons. Linearly decodable information could be used by downstream neurons and is therefore sometimes described as explicit in the code (deCharms & Zador 2000, Kriegeskorte 2011, DiCarlo et al. 2012, Hong et al. 2016). Like the representational geometry, linear decoding focuses on the content of the code, but it peels off more of the onion, omitting information about any other features that may be present in the code, as well as information that would require a more sophisticated (e.g., nonlinear) decoder.

3. PUTTING IT BACK TOGETHER: FROM THE INSIDE OUT

As we peeled the onion, we explained what information is discarded with each layer and what information is kept. We emphasized the motivation for the peeling away of each layer. With our basic framework in place, we are in a position to take the opposite perspective below. We proceed in the reverse direction, putting the onion back together layer by layer. In the process, we emphasize the motivation for using more and more information about brain representations. We also move beyond the abstract definition of the layers and add more information on how analyses actually work, how they fall short in terms of their neuroscientific motivation, and what future directions might be desirable.

3.1. Linear Decoders

$\mathbf{w}_{\text{Fisher}}$: decoding weights for Fisher linear discriminant (P weights, one for each channel)

$\mathbf{r}_k$: activity pattern (across channels) in response to the kth condition (kth row of $\mathbf{U}$, transposed)

Linear decoding involves fitting a set of P weights (one for each neuron, site, or voxel) such that the weighted sum of the activities reflects the variable to be decoded. The decoded variable could be continuous or categorical. Computing the weighted sum of the responses to a stimulus is equivalent to projecting the response pattern elicited by the stimulus onto the dimension in the multivariate response space that is defined by the weight vector $\mathbf{w}$. A linear readout, then, provides a projection of the data, a perspective onto the geometry of the points corresponding to the response patterns. As we add more linear decoders, we capture more dimensions of the representational space. This insight suggests that we might want to consider all dimensions or, equivalently, the representational geometry. However, let us consider linear decoders in more detail first.

A linear decoder can be fitted to optimize different cost functions. For example, a linear support vector machine will maximize the margin of separation between two categories. A Fisher linear discriminant will maximize the ratio of between-category and within-category variance after projection onto the discriminant dimension. When the within-category distributions and noise are multinormal with equal covariance, the resulting one-dimensional projection retains all the information that the patterns contain about the category. The Fisher linear discriminant uses weights

$$\mathbf{w}_{\text{Fisher}} = (\mathbf{r}_j - \mathbf{r}_i)^T \Sigma^{-1}, \qquad 2.$$

where Σ is an estimate of the $P \times P$ within-class covariance matrix, and $\mathbf{r}_i$ and $\mathbf{r}_j$ and the activity patterns for the ith and jth stimulus (rows of $\mathbf{U}$, transposed), respectively. For linear decoding of brain activity, the Fisher linear discriminant often performs well (Mur et al. 2009, Misaki et al. 2010). Given the numbers of response channels and measurement time points typical in cell recordings and fMRI, the assumed multinormal noise model is about as rich a model of the dependencies between responses as is realistic to estimate.

When we fit a linear decoder, the P weights (one for each response channel) will be overfitted to some extent. The discriminant will therefore tend to separate the categories better in the data used for fitting (the training data) than in a new data set. In fact, when $P > K$ and there is noise in the data, some linear discriminant is sure to perfectly separate the training data as desired, even when all patterns are drawn from the same distribution, and the responses thus contain no information about the stimulus. To assess the actual degree of linear separability, we therefore need to test the discriminant on new data (the test data). This enables us to obtain an unbiased estimate of discriminability. Testing on a different data set provides a compelling empirical demonstration of decodability. Any assumptions that we have made (such as multinormal noise), if incorrect, will work against a significant result. The validity of frequentist inference of decodability, therefore, does not depend on the assumption of multinormality. This is an advantage of linear decoding over multivariate analysis of variance, which can also be used to test for pattern differences. In multivariate analysis of variance, the validity of inference depends on multinormality, and violations of this assumption might inflate the false-positive rate (Kriegeskorte et al. 2006, Kriegeskorte 2011, Allefeld & Haynes 2014).

When the target variable is categorical, decoding performance is often assessed in terms of accuracy (percentage of correctly classified test patterns). This requires the definition of a threshold on the discriminant dimension, enabling us to classify the test patterns by category. The thresholding slightly complicates the analysis. Moreover, the quantization involved in counting correct and incorrect classifications entails a loss of information, which can be substantial in practice when the number of test patterns is small. If we assume that the within-category distributions are multinormal (as we do above when choosing the Fisher linear discriminant), then the pattern distributions will be univariate normal distributions after projection onto the discriminant.

The univariate contrast on the discriminant then provides a more sensitive test statistic than the classification accuracy. The linear-discriminant contrast $c^2_{\mathbf{r}_i,\mathbf{r}_j} = \mathbf{w}_{\text{Fisher}} \cdot (\mathbf{r}_j^{\text{test}} - \mathbf{r}_i^{\text{test}})$ can be normalized by its standard error to obtain the linear discriminant t (LD-t) value, which can be converted into a p value, providing a frequentist test of the null hypothesis that the two patterns are identical (Kriegeskorte et al. 2007, Allefeld & Haynes 2014, Nili et al. 2014, Walther et al. 2016). The use of a continuous discriminability measure as the test statistic obviates the need for thresholding and prevents the loss of information associated with quantization when counting correct classifications. It therefore provides a more statistically efficient test of discriminability.

Researchers often train multiple linear decoders on the same data set to determine the degree to which different variables can be decoded. Each decoder provides the projection of the K stimulus-related response patterns onto a different dimension. We could push this approach to the extreme and fit a linear decoder for every dichotomy (categorical division into two subsets) of the stimulus set, or even for every feature (continuous property vector). However, the number of dimensions of the linear subspace containing all the patterns has the upper bound $\min(K-1, P)$. If every added linear decoder sampled a dimension of the linear subspace containing the stimuli that is independent of the previously sampled dimensions, then we would need at most $K-1$ linear decoders (for the frequent case of $K \leq P$) to fully define the representational geometry. We would have to store K values (the projections of the stimuli) for each of the $K-1$ decoder dimensions, and the characterization would depend on the particular decoders chosen.

3.2. Representational Geometry

An alternative way to fully define the representational geometry is to fit a linear discriminant for each pair of the K stimuli. There are K choose $2 = K \cdot (K-1)/2$ pairs. For each pair of stimuli, we store only the contrast (signed separation) of the projections of the two stimulus-related response patterns onto their linear discriminant. The contrast serves as an estimate of the representational distance between the two stimuli in the representation. With each pairwise discriminant, we reduce our uncertainty about the geometry of the ensemble. With the final pair, there is no more wiggle room for any point relative to any other point, and the geometry has been completely specified.

When we fit a separate Fisher linear discriminant for a stimulus pair, there is only one stimulus in each of the two classes to be discriminated. The within-class variability, then, just reflects the noise of the measurements. In practice, this often renders the assumption of equal-covariance multinormal within-class distributions (which the Fisher linear discriminant is based on) appropriate. The linear discriminant contrast

$$c^{(2)}_{\mathbf{r}_i,\mathbf{r}_j} = (\mathbf{r}_j - \mathbf{r}_i)^T \Sigma^{-1} (\mathbf{r}_j^{\text{test}} - \mathbf{r}_i^{\text{test}}) \qquad 3.$$

turns out to be a crossvalidated variant of the squared Mahalanobis distance

$$d^2_{\mathbf{r}_i,\mathbf{r}_j} = (\mathbf{r}_j - \mathbf{r}_i)^T \Sigma^{-1} (\mathbf{r}_j - \mathbf{r}_i). \qquad 4.$$

We therefore refer to $c^{(2)}_{\mathbf{r}_i,\mathbf{r}_j}$ as the crossnobis estimator (Nili et al. 2014, Diedrichsen et al. 2016, Kriegeskorte & Diedrichsen 2016, Walther et al. 2016). The crossnobis estimator provides an unbiased estimate of the true (i.e., noise-free) squared Mahalanobis distance between two response patterns. Note that the estimator can be negative (as is required for unbiasedness), and thus the parenthetical superscript does not indicate squaring, but instead emphasizes the relationship to the squared Mahalanobis distance. The signed square root of the crossnobis estimator $c_{\mathbf{r}_i,\mathbf{r}_j} = \text{sign}(c^{(2)}_{\mathbf{r}_i,\mathbf{r}_j}) \cdot |c^{(2)}_{\mathbf{r}_i,\mathbf{r}_j}|^{1/2}$ is an unbiased estimator of the Mahalanobis distance.

To understand the bias, consider the fact that a distance, by definition, is nonnegative. When a distance is estimated from measured data, the noise creates a positive bias. For example, when the

true distance is 0, the two noisy point estimates will still be separated by a positive distance. Computing distances from noisy data, therefore, can yield a distorted picture of the representational geometry and misleading inferences (Cai et al. 2016). One remedy is to restrict the analysis to the ranks of the distances (Kriegeskorte et al. 2008a,b; Nili et al. 2014). However, we might prefer to use a richer, ratio-scale characterization of the geometry.

The positive bias of distance functions used as distance estimators is the continuous equivalent of the overfitting bias that inflates training-set classifier accuracy. As for classifier accuracy, the bias can be removed by using an independent test set (or cross-validation, where results are averaged across different splits of the data into training and test sets). The crossnobis estimator uses this method to remove the bias. Unbiasedness entails that the estimator's expected value is 0 when the true distance is 0. As a result, an unbiased distance estimator must be able to return negative values and thus cannot itself be a distance or a metric. The crossnobis estimator combines a multinormal noise model (which captures spatial noise correlations) with cross-validation (which removes the bias) and provides continuous distance estimates (not compromised by quantization or saturation) with an interpretable 0 point. The crossnobis RDM is therefore an attractive way to estimate the representational geometry in practice.

Recall (from Section 2) that the Euclidean distance matrix [in conjunction with an isotropic, homoscedastic Gaussian noise model and a flat prior $p(s)$ over the stimuli] defines the joint distribution $p(s, \mathbf{r})$ up to a rotation and translation in the space spanned by the response channels and thus completely defines the mutual information $I(s; \mathbf{r})$ between stimulus and response. The Mahalanobis distance generalizes this relationship to Gaussian noise that is anisotropic (correlated between response channels) and homoscedastic (equal across stimuli). The Mahalanobis distance matrix completely defines the total encoded information $I(s; \mathbf{r})$, as well as the encoded information about any given particular stimulus feature. We can use the signed square root of the crossnobis estimator in practice, which provides an unbiased estimate of the Mahalanobis distance.

Beyond the encoded information, the representational geometry also characterizes the format of the code, up to an affine transformation (linear transformation and translation) of the pattern ensemble in the response space. This means that, given the true Mahalanobis distances, we know not only what information is encoded, but also how well any decoder can read out any feature if it has access to all neurons in the population and is capable of an affine transformation. This includes all reasonable nonlinear decoders.

The representational geometry thus contains two subsets of information: the encoded information and additional information about the format of the code. The encoded information might represent a good target for analysis because it captures all the information that the code might possibly provide to downstream computations. This suggests that we should peel off the format information contained in the representational geometry so as to arrive at the encoded information core of the onion (**Figure 4**).

The encoded information can be conceptualized as a function $I[\mathbf{f}, p(s, \mathbf{r})]$ of the stimulus feature of interest and the joint distribution of stimulus and response. The function is passed a feature $\mathbf{f}$ to be decoded (i.e., a property vector with an entry for each stimulus) as an input, and it returns the amount of information that the code contains about the feature. This definition strips away all format information and captures how well the code reflects each stimulus feature.

A complex network of neurons reading the code can implement an arbitrary nonlinear decoder and thus could theoretically access all encoded information. However, not all encoded information can be extracted directly by single neurons reading out the code. Readout neurons are limited to simple operations, such as linear decoding, and may not be able to access the entire population. The encoded information core of the onion thus invites us to peel further (**Figure 4**, right) in search of a biologically valid definition of explicit encoded information. We can think of

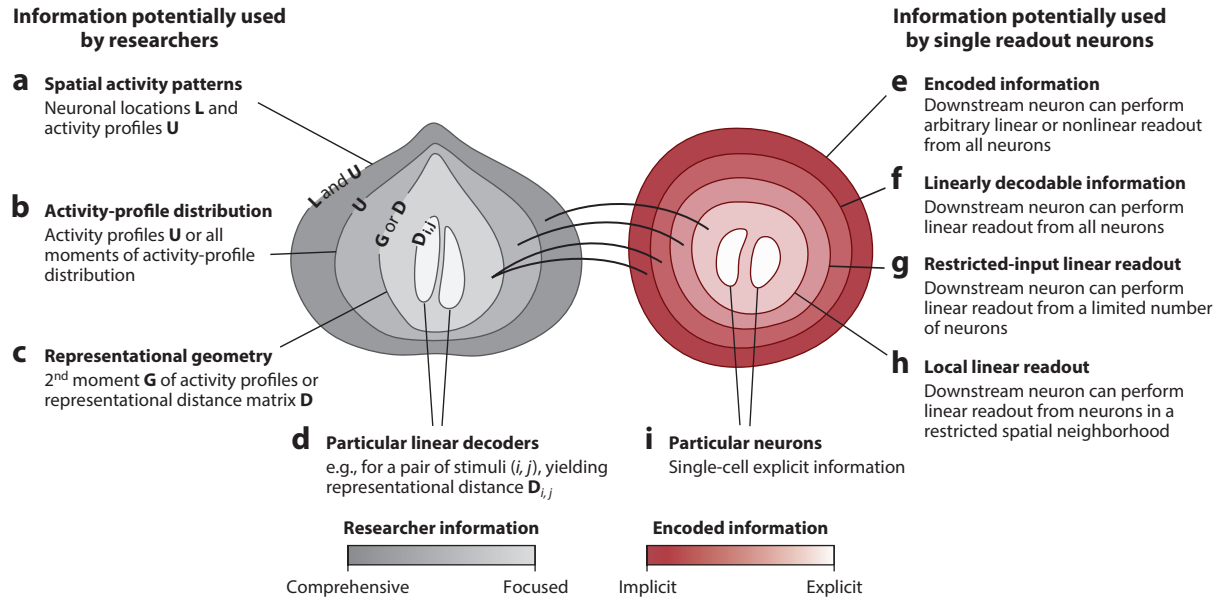


Figure 4

Peeling the encoded-information core of the onion. In search of the mental content represented by the code, a researcher might decide to peel off the outer layers of the onion (*a*,*b*). The representational geometry (*c*) captures the encoded information (assuming isotropic, homoscedastic noise), along with aspects of the format of the code (*d*). We can peel away the format information from the representational geometry to arrive at the encoded-information core of the onion (*e*). Readout neurons may be limited to simple operations, such as linear decoding, and may not be able to access the entire population. This suggests further peeling of the encoded-information core of the onion. However, to select subsets of the encoded information that are visible to particular biologically plausible decoders, we need to look back at the outer layers of the onion (at the geometry, the profiles, and the locations). To avoid confusion, we cut the encoded-information core transversely (*right*) and keep the outer layers close at hand (*left*). Peeling off the information not amenable to linear readout (*dark red*) requires knowledge of the representational geometry. It reveals the linearly decodable information (*f*) for each stimulus feature. A researcher may want to use more information about the code to further restrict what portion of the encoded information is considered. Using the distribution of activity profiles, a researcher can specify what information is available for readout if each readout neuron has access to a limited number of neurons from the code (*g*). Finally, using the spatial locations of the neurons, a researcher can specify what information is available for readout if the readout neuron can only see the code within a restricted spatial neighborhood (*h*). We can think of a kernel of the encoded-information core of the onion as the information gleaned by a linear decoder using only a single neuron (*i*). Note that, although linear decoding is a popular approach, it is not the only biologically plausible variant. For example, radial-basis-function decoding may be plausible and would give rise to an alternative decomposition of the encoded information.

the explicit information as information that is inferentially close, requiring only a single layer of readout neurons to be explicated.

As a first step, we peel off the portion of the encoded information that is not amenable to linear readout (the dark red layer in **Figure 4**). Selecting the linearly decodable subset requires knowledge of the representational geometry. The linearly decodable information is the union of all the information that can be directly linearly decoded. However, it does not contain the synergistic information of all linear decoders. If we included the synergistic information, then a set of linear decoders would always recover the entire encoded information because the decoders can always just copy the code. To capture the portion of the information that a given class of decoders can explicate, we need to exclude the synergistic information. The synergistic information is the implicit portion of the information in the decoder outputs and may, of course, be explicated by neurons further downstream.

Linear decoding of the entire population might be considered biologically implausible. For example, no single readout neuron is likely to receive input from the entire population of V1 neurons. We would therefore like to further peel the encoded information core, restricting what we consider explicit to subsets of the information that are accessible to more biologically plausible decoders. To define what information is accessible to decoders restricted to subsets of neurons, we need to bring back more of the outer layers of the onion (the profiles and the locations) because the representational geometry does not contain this information. So let us continue to put the onion back together.

3.3. Distribution of Activity Profiles

Although the representational geometry gives us much information about the format of the code, it fails to specify to what extent a variable is encoded in a localized or distributed fashion within our region of interest. A weak distributed selectivity can provide as much information as a strong localized selectivity. To a linear decoder that can access all neurons, these differences are irrelevant. The readout weight pattern can select localized signals or integrate weak distributed signals over the entire extent of the region, flipping signs as needed for decoding.

Consider a data set of V1 neuronal responses to Gabor stimuli. If someone snuck into the lab at night and replaced the stimulus–response matrix $\mathbf{U}$ with a randomly rotated matrix $\mathbf{U}' = \mathbf{UR}$, where $\mathbf{R}$ is a $P \times P$ orthonormal rotation matrix, then the electrophysiologist inspecting the tuning curves in the morning might no longer recognize the data as coming from V1. The receptive fields of the individual responses would no longer be localized in small regions of the visual field. Rather, every measured channel would respond somewhat to stimuli at every location. However, linear decoding analyses would show all the same results as on the previous day. RSA, PCMs, and encoding models using 0-mean Gaussian weight priors (i.e., ridge regression) would also all give the same results as on the previous day when used to evaluate representational models on the entire set of responses (i.e., in terms of overall prediction accuracy).[1] However, inspecting the weights of the encoding model for individual responses would reveal that sensitivities no longer appear localized to retinotopic locations.

The way that effects are distributed across the neuronal population likely matters to neuronal computation. For example, a downstream readout neuron might not have access to the entire code (e.g., when the code is spread out over a cortical area like V1). This motivates the use of further information in the distribution of activity profiles that is not contained in the representational geometry. We might restrict what we consider explicit information to what can be linearly read out from a limited number of neurons in the code. Focusing on this restricted-input linear-readout information (**Figure 4*g***) strips off another layer of the encoded information core of the onion.

Note that, to narrow our definition of the explicit portion of the encoded information, we use researcher information from layers outside the encoded information. Narrowing our focus to linearly decodable information (**Figure 4*f***) requires the use of the format information in the representational geometry (**Figure 4*c***). Further narrowing our definition of explicit information to restricted-input linear-readout information requires us to use information from the next enclosing layer: the distribution of activity profiles. There is one more layer to go to recover the entire onion.

[1] To understand why a ridge-regression encoding model will predict the data equally well after the rotation of the pattern ensemble in the space spanned by the response channels, consider **Figure 3*b*** and imagine that the encoding model contains two indicator predictor profiles, the first containing a one for stimulus 1 and the second containing a one for stimulus 2. Both scenarios (**Figure 3*b***, *i* and *ii*) can be explained equally well with the encoding model, and the optimal weights will yield identical L2 penalties.

3.4. Spatial Structure of Activity Patterns

Neurons can receive signals through long axonal projections from far-away locations in the opposite hemisphere. However, the development of the axonal tracts and the propagation of signals through them is costly. As a result, most connections are local, and the location of a neuron in three-dimensional brain space is thought to reflect its place in the topology of the network (Chklovskii & Koulakov 2004, Chen et al. 2006).

Imagine that we attempted to reengineer a brain with the locations of all neurons randomly shuffled while preserving the network topology and resulting dynamics. The volume of axonal tracts would likely vastly exceed the space available in the skull. The energy requirements would be much higher. Finally, it seems impossible that the dynamics could be preserved, since most signal latencies would be much longer (and some shorter). A neuron's function prominently depends on its connections in the network, not its location in the brain. However, its location in the brain determines which other neurons it can connect with cheaply at short latency.

These considerations motivate the use of the locations of the neurons **L** (outer layer of the onion) to further constrain what information is considered explicit in the code. At the simplest level, we could assume that a downstream neuron can only read out information from the code within a small radius of its own location. If we define the explicit information in the code as information that can be linearly read out from local clusters of neurons in the code, then we peel off another layer of the onion (**Figure 4*b***).

If we restrict the admissible input to a single neuron, then we recover the extreme definition of explicit coding. This definition was, in fact, widely used before larger numbers of neurons were routinely measured and analyzed with linear decoders. We can think of the single-neuron explicit information as the innermost bit of the onion—the union of a set of kernels, each of which contains the explicit information carried by a single neuron (**Figure 4*i***).

The assumptions that we made in peeling the encoded information core of the onion are not the only assumptions that make sense. For example, we could assume that downstream neurons read the code with radial basis functions, activating according to a nonlinear (e.g., Gaussian) function of the distance between a preferred pattern and the current input pattern. This would induce an alternative decomposition of the encoded information. Or we could assume a more biologically detailed model of the way downstream neurons read the code. Moreover, we could use connectomic information, rather than simply locations, to define what information is directly accessible. A different readout model of this type would lead us to a different core of encoded information that is to be considered explicit.

As we put the onion back together, access to successive outer layers enables us to progressively peel the encoded information core of the onion. We can use the representational geometry, the activity-profile distribution, and the locations of the neurons to define the explicit information in more biologically plausible ways.

4. WHAT LAYERS OF THE ONION SHOULD INFORM TESTS OF BRAIN-COMPUTATIONAL MODELS?

Representational models enable us to combine prior assumptions with brain-activity data to make inferences about brain representations. They support data-driven as well as theory-driven analyses. At the data-driven end of the spectrum, they can help us discover what features are prominently represented in different brain regions (e.g., Kriegeskorte et al. 2008b, Huth et al. 2012). At the theory-driven end of the spectrum, they enable us to test task-performing brain-computational models (Kriegeskorte & Douglas 2018a).

To test a model of brain information processing, we need to compare model and brain in terms of behavior and activity dynamics (Kriegeskorte 2015, Kriegeskorte & Diedrichsen 2016, Yamins

& DiCarlo 2016, Paninski & Cunningham 2017). This requires some measure of the goodness of fit to the data that the model achieves. Naively, we might map each unit of a spiking neural network model to a neuron in the brain and directly compare the spatiotemporal activity pattern during the same task, requiring precise prediction of each spike. However, such an approach of precise spatial and temporal correspondence is both unrealistic and undesirable. It might be a useful exercise to model a particular individual brain and the dynamics of a particular cognitive act. Given the idiosyncrasies of each animal's brain and the unique nature of each trial of some cognitive act, however, we are, in general, interested in a more abstract kind of functional correspondence. We should not expect our model's dynamics to be more precisely mapped in space and time to a given subject's brain dynamics than one subject's brain dynamics can be mapped to another's while both are performing the same cognitive task. We consider the spatial and temporal aspects of the mapping in turn.

Each individual primate brain is unique. A given cortical area to be modeled will have a different number of neurons in each individual, with each neuron having idiosyncratic functional properties. The approximate functional consistency reported in the literature resides at the level of cortical areas and populations of neurons. Spatial correspondence between individuals can sometimes be defined even within cortical areas, but breaks down at the level of cortical columns. Consider the primary visual cortex. The global retinotopic map could form the basis for defining a more precise spatial correspondence between individuals within the area. However, the organization of orientation columns and the specific neurons within them are not expected to have a spatially precise correspondence across individuals.

The temporal correspondence problem is similarly difficult. Each repeated trial of a cognitive task is unique, even in the same individual. Consider the act of recognizing a particular image. First, brain dynamics might be inherently stochastic, precluding precise reproduction of the measured response pattern. Furthermore, a brain's internal state at trial onset cannot be entirely controlled, as would be necessary to precisely repeat the trial with even a deterministic brain. Finally, the act of perceiving the image the first time permanently changes the brain and will specifically and measurably alter the way that the image is processed the second time (Grill-Spector et al. 2006). The same brain never sees a picture twice, as Heraclitus might assert today.

Despite the difficulties of defining precise spatial and temporal correspondence mappings, a tacit fundamental assumption of neuroscience is that different individual brains of the same species implement the same computational functions. This implies some appropriate level of abstraction at which functional correspondence is evident. Qualitative similarities among individuals abound, of course, at the levels of behavior, regional brain activation, and local neuronal tuning. However, to make functional correspondence a rigorous, mathematically defined concept, we will need to choose summary statistics of brain activity at the right level of abstraction. Our choice of statistics must trade off dynamic detail for interindividual generalizability to some extent, but we would like to retain rich signatures of the computational functions while discarding the idiosyncrasies of individual brains. Such summary statistics will be essential for testing and adjudicating among models of brain information processing with brain-activity measurements, enabling us to compare dynamics between models and brains. The onion of brain representations is a step in the direction of organizing some of the abstractions available.

For example, we could choose a level of spatial and temporal precision, say, a spatial unit like the cortical area and a temporal unit of 100 ms. We could then choose a degree of peelage for the onion, say, the representational geometry, as a basis for comparing model to brain. The brain dynamics within an experimental context comprising a finite set of representational states (stimuli) would then be characterized by a spatiotemporal field of RDMs $\mathbf{D}_{i,j}(l,t)$, where i, j indicate the representational states, l the location in the brain, and t the time after trial onset. Alternatively,

we could peel the onion further and focus comparisons on the encoded information or the explicit information. Or we could choose to peel the onion less and compare representations on the basis of the distribution of activity profiles.

We could also choose a different level of spatial or temporal precision. Greater precision of the spatial mapping is attractive because it gives us greater sensitivity to subtle differences between models. However, a fundamental question is whether a spatial correspondence at the desired precision even exists between individual animals. More practically, the mapping needs to be estimated from a separate data set, which can be costly. At the level of cortical areas, the correspondence mapping is relatively simple, realistic to estimate, and likely to generalize across subjects. At the level of small patches of cortex (e.g., cortical columns), the mapping is complex, is potentially unrealistic to estimate, and would likely have to be estimated separately for each subject. Similar trade-offs need to be considered in choosing the level of precision in the temporal domain.

M: $K \times Q$ matrix of model features (where Q is the number of features)

T: number of measurement points or trials

$\mathbf{y}_p$: measured activity time course for the pth channel (column vector of length T)

$\mathbf{u}_p$: activity profile (across conditions) of the pth channel (pth column of $\mathbf{U}$)

Z: $T \times K$ design matrix, indicating how measurements relate to experimental conditions

5. CONCLUSION

The onion of brain representations organizes the different aspects of brain-activity data into a nested hierarchy. As we peel it, we focus progressively on aspects that appear more directly related to the representational content. We hope to peel away layers that reflect developmental coincidences, random biological variation, and other epiphenomena without functional relevance, to arrive at the brain's representational core: the neural code used by the brain itself to mind and manipulate the world. Peer Gynt, the tragic hero of Henrik Ibsen's (1867, p. 218) eponymous play, peeled an onion in search of himself and came to conclude: "It's nothing but layers, smaller and smaller. Nature's a joker." We embrace the research program outlined in this review despite understanding its caveat: The layers might fall away to reveal no core of meaning. The brain is not a conventional computer. Nature, the joker, might leave us with nothing but a dynamical system.

6. APPENDIX: MATHEMATICAL DETAILS

6.1. Representational Models

Representational models define a probability distribution of activity profiles, $p(\mathbf{u}|\theta, M)$. These models often have some second-level parameters θ that determine the size or shape of the distribution. Given a set of brain observations $y_p = \mathbf{Z}\mathbf{u}_p + \epsilon$, we seek to evaluate the (marginal) likelihood of the data, given the model $\mathbf{M}$ and second-level parameters:

$$p(\mathbf{y}_p|\theta, \mathbf{M}) = \int p(\mathbf{y}_p|\mathbf{u}_p)\, p(\mathbf{u}_p|\theta, \mathbf{M})\, \mathrm{d}\mathbf{u}. \quad 5.$$

The integral is being taken over all possible activation profiles that may be the cause of our brain observations $\mathbf{y}_p$. In encoding approaches, this marginal likelihood is approximated using cross-validation; in pattern component modeling, the integral is evaluated directly. Both approaches assume (implicitly or explicitly) a Gaussian distribution of both signal ($\mathbf{u}$) and noise (ϵ).

6.2. Second-Moment Matrix

When evaluating the marginal likelihood, the critical statistic that fully defines the representational model is the second moment of the activity profiles,

$$\mathbf{G} = \sum_{p=1}^{P} \mathbf{u}_p \mathbf{u}_p^T / P = \mathbf{U}\mathbf{U}^T / P. \quad 6.$$

$\mathbf{w}_p$: encoding weights for channel p (Q weights, one for each model feature)

Assuming a mean of $\mathbf{0}$ across channels, $\mathbf{G}$ is the (co)variance matrix of the activity-profile distribution. In encoding models, each activity profile is expressed as a linear combination of model features: $\mathbf{u}_p = \mathbf{M}\mathbf{w}_p$, where $\mathbf{M}$ are the model features, and $\mathbf{w}_p$ are the feature weights. Furthermore, it is often assumed that the feature weights are identically and independently distributed across channels (ridge regression). In this case, the second-moment matrix becomes $\mathbf{G} = \mathbf{M}\mathbf{M}^T$.

The highest eigenvalues and associated eigenvectors of $\mathbf{G}$ are used to plot the representational space. An unbiased estimate of the second moment matrix, $\hat{\mathbf{G}}_{CV}$, can be obtained from the data using cross-validation (Diedrichsen et al. 2018).

6.3. Representational Dissimilarities

Representational dissimilarities measure how different the activity patterns for two conditions (i, j) are from each other. The Euclidean distance (or, by extension, the Mahalanobis distance) is commonly used and can be directly derived from the second moment matrix:

$$d_{i,j}^{\text{Euc}} = \sqrt{(\mathbf{r}_i - \mathbf{r}_j)^T (\mathbf{r}_i - \mathbf{r}_j)} = \sqrt{(\mathbf{G}_{i,i} + \mathbf{G}_{j,j} - 2\mathbf{G}_{i,j}) \cdot P}. \quad 7.$$

The Euclidean distance is sensitive to any difference between the activity patterns, including scaling of the intensity of the patterns. If a distances measure that is insensitive to this scaling is required, the cosine distance is a useful measure:

$$d_{i,j}^{\cos} = 1 - \frac{\mathbf{r}_i^T \mathbf{r}_j}{\sqrt{\mathbf{r}_i^T \mathbf{r}_i \cdot \mathbf{r}_j^T \mathbf{r}_j}} = 1 - \frac{\mathbf{G}_{i,j}}{\sqrt{\mathbf{G}_{i,i}\mathbf{G}_{j,j}}}. \quad 8.$$

The cosine distance becomes the correlation distance, when each pattern $\mathbf{r}_i$ is normalized by subtracting its mean from each element. If we apply these formulae for each pair (i, j) of conditions, the second-moment matrix can be transformed into an RDM $\mathbf{D}$. For the Euclidean distance, this transformation can be written more compactly as

$$\mathbf{d} = \text{diag}(\mathbf{C}\mathbf{G}\mathbf{C}^T), \quad 9.$$

where $\mathbf{C}$ is a matrix of contrasts (one for each pair of stimuli), and $\mathbf{d}$ is a vector containing all unique (lower triangular) entries of the squared Euclidean distance matrix $\mathbf{D}$ (which is symmetric about a diagonal of zeros). For the inverse transformation, we need to consider that the second moment matrix $\mathbf{G}$ retains a little bit more information than $\mathbf{D}$ in that it specifies the geometry with respect to the origin of the multivariate response space. We can rotate the geometry arbitrarily about the origin without changing $\mathbf{G}$. However, if we translate the geometry or rotate about a different point, then $\mathbf{G}$ changes, while $\mathbf{D}$ remains the same. All we have to do to make the two matrices entirely equivalent is to add the baseline condition (all-0 vector) to the set of stimuli captured by the distance matrix, yielding $\mathbf{D}'$ ($K + 1$ by $K + 1$). Alternatively, we can remove the baseline information from $\mathbf{G}$ by removing the mean activity pattern from each pattern (thus centering the geometry on the origin of multivariate response space). The second moment $\mathbf{G}'$ for the centered geometry can then be computed from $\mathbf{D}$ as follows:

$$\mathbf{G}' = -1/2\mathbf{H}\mathbf{D}\mathbf{H}, \quad 10.$$

where $\mathbf{H}$ is a centering matrix: $\mathbf{H} = \mathbf{I}_K - \mathbf{1}_K/K$. $\mathbf{1}_K$ is a square matrix of ones. Since $\mathbf{G}$ can be computed from $\mathbf{D}$ and vice versa, the two contain identical information (Diedrichsen & Kriegeskorte 2017).

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

The authors thank Peter Dayan, John Krakauer, John Morrison, Liam Paninski, and Xuexin Wei for helpful comments and discussions. J.D. was supported by an NSERC Discovery Grant (RGPIN-2016-04890) and the Canada First Research Excellence Fund (BransCAN).

LITERATURE CITED

Abbott LF, Dayan P. 1999. The effect of correlated variability on the accuracy of a population code. *Neural Comput.* 11:91–101

Afraz S-R, Kiani R, Esteky H. 2006. Microstimulation of inferotemporal cortex influences face categorization. *Nature* 442:692–95

Allefeld C, Haynes JD. 2014. Searchlight-based multi-voxel pattern analysis of fMRI by cross-validated MANOVA. *NeuroImage* 89:345–57

Averbeck BB, Latham PE, Pouget A. 2006. Neural correlations, population coding and computation. *Nat. Rev. Neurosci.* 7(5):358–66

Bargmann CI, Marder E. 2013. From the connectome to brain function. *Nat. Methods* 10(6):483–90

Bechtel W. 1998. Representations and cognitive explanations: assessing the dynamicist's challenge in cognitive science. *Cogn. Sci.* 22(3):295–318

Brentano F. 1874. *Psychology from an Empirical Standpoint*. Abingdon, UK: Routledge

Cai MB, Schuck NW, Pillow J, Niv Y. 2016. A Bayesian method for reducing bias in neural representational similarity analysis. In *Advances in Neural Information Processing Systems*, ed. DD Lee, M Sugiyama, UV Luxburg, I Guyon, R Garnett, pp. 4952–60. Cambridge, MA: MIT Press

Carlson T, Tovar DA, Alink A, Kriegeskorte N. 2013. Representational dynamics of object vision: the first 1000 ms. *J. Vis.* 13(10):1

Carlson TA, Schrater P, He SY. 2003. Patterns of activity in the categorical representation of objects. *J. Cogn. Neurosci.* 15:704–17

Chen BL, Hall DH, Chklovskii DB. 2006. Wiring optimization can relate neuronal structure and function. *PNAS* 103(12):4723–28

Chklovskii DB, Koulakov AA. 2004. Maps in the brain: What can we learn from them? *Annu. Rev. Neurosci.* 27:369–92

Churchland MM, Cunningham JP, Kaufman MT, Foster JD, Nuyujukian P, et al. 2012. Neural population dynamics during reaching. *Nature* 487:51–56

Cichy RM, Pantazis D, Oliva A. 2014. Resolving human object recognition in space and time. *Nat. Neurosci.* 17(3):455–62

Cox DD, Savoy RL. 2003. Functional magnetic resonance imaging (fMRI) brain reading: detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage* 19:261–70

Cunningham JP, Yu BM. 2014. Dimensionality reduction for large-scale neural recordings. *Nat. Neurosci.* 17(11):1500–9

deCharms RC, Zador A. 2000. Neural representation and the cortical code. *Annu. Rev. Neurosci.* 23:613–47

Dennett D. 1987. *The Intentional Stance*. Cambridge, MA: MIT Press

DiCarlo JJ, Cox D. 2007. Untangling invariant object recognition. *Trends Cogn. Sci.* 11:334–41

DiCarlo JJ, Zoccolan D, Rust NC. 2012. How does the brain solve visual object recognition? *Neuron* 73(3):415–34

Diedrichsen J. 2019. Representational models and the feature fallacy. In *The Cognitive Neurosciences*, ed. MS Gazzaniga, GR Mangun, D Poeppel. Cambridge, MA: MIT Press. 6th ed. In press

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

Diedrichsen J, Provost S, Hossein ZH. 2016. On the distribution of cross-validated Mahalanobis distances. arxiv:1607.01371 [stat.AP]

Diedrichsen J, Ridgway GR, Friston KJ, Wiestler T. 2011. Comparing the similarity and spatial structure of neural representations: a pattern-component model. *NeuroImage* 55(4):1665–78

Diedrichsen J, Yokoi A, Arbuckle SA. 2018. Pattern component modeling: a flexible approach for understanding the representational structure of brain activity patterns. *NeuroImage* 180:119–33

Dumoulin SO, Wandell BA. 2008. Population receptive field estimates in human visual cortex. *NeuroImage* 39(2):647–60

Edelman S. 1998. Representation is representation of similarities. *Behav. Brain Sci.* 21(4):449–67

Edelman S, Grill-Spector K, Kushnir T, Malach R. 1998. Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology* 26(4):309–21

Ejaz N, Hamada M, Diedrichsen J. 2015. Hand use predicts the structure of representations in sensorimotor cortex. *Nat. Neurosci.* 18:1034–40

Felleman DJ, Van Essen DC. 1991. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* 1(1):1–47

Formisano E, Kim DS, Di Salle F, van de Moortele PF, Ugurbil K, Goebel R. 2003. Mirror-symmetric tonotopic maps in human primary auditory cortex. *Neuron* 40(4):859–69

Ganguli S, Sompolinsky H. 2012. Compressed sensing, sparsity, and dimensionality in neuronal information processing and data analysis. *Annu. Rev. Neurosci.* 35:485–508

Gao P, Ganguli S. 2015. On simplicity and complexity in the brave new world of large-scale neuroscience. *Curr. Opin. Neurobiol.* 32:148–55

Grill-Spector K, Henson R, Martin A. 2006. Repetition and the brain: neural models of stimulus-specific effects. *Trends Cogn. Sci.* 10(1):14–23

Güçlü U, van Gerven MA. 2015. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* 35(27):10005–14

Haxby JV, Connolly AC, Guntupalli JS. 2014. Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* 37:435–56

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P. 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293(5539):2425–30

Haynes J-D. 2015. A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron* 87(2):257–70

Haynes J-D, Rees G. 2006. Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.* 7(7):523–34

Hebart MN, Baker CI. 2017. Deconstructing multivariate decoding for the study of brain function. *NeuroImage* 180:4–18

Hong H, Yamins DL, Majaj NJ, DiCarlo JJ. 2016. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* 19(4):613–22

Hung CP, Kreiman G, Poggio T, DiCarlo JJ. 2005. Fast readout of object identity from macaque inferior temporal cortex. *Science* 310(5749):863–66

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532:453–58

Huth AG, Nishimoto S, Vu AT, Gallant JL. 2012. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76:1210–24

Ibsen H. 1867. *Peer Gynt: A Dramatic Poem.* Copenhagen: Gyldendal

Kamitani Y, Tong F. 2005. Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.* 8(5):679–85

Kay KN, Naselaris T, Prenger RJ, Gallant JL. 2008. Identifying natural images from human brain activity. *Nature* 452(7185):352–55

King JR, Dehaene S. 2014. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* 18(4):203–10

Kobak D, Brendel W, Constantinidis C, Feierstein CE, Kepecs A, et al. 2016. Demixed principal component analysis of neural population data. *eLife* 5:e10989
Kriegeskorte N. 2011. Pattern-information analysis: from stimulus decoding to computational-model testing. *NeuroImage* 56(2):411–21
Kriegeskorte N. 2015. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* 1:417–46
Kriegeskorte N, Bandettini P. 2007. Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage* 38(4):649–62
Kriegeskorte N, Diedrichsen J. 2016. Inferring brain-computational mechanisms with models of activity measurements. *Philos. Trans. R. Soc. B* 371(1705):20160278
Kriegeskorte N, Douglas PK. 2018a. Cognitive computational neuroscience. *Nat. Neurosci.* 29:1148–60
Kriegeskorte N, Douglas PK. 2018b. Interpreting encoding and decoding models. arXiv:1812.00278 [q-bio.NC]
Kriegeskorte N, Formisano E, Sorger B, Goebel R. 2007. Individual faces elicit distinct response patterns in human anterior temporal cortex. *PNAS* 104(51):20600–5
Kriegeskorte N, Goebel R, Bandettini P. 2006. Information-based functional brain mapping. *PNAS* 103(10):3863–68
Kriegeskorte N, Kievit RA. 2013. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* 17:401–12
Kriegeskorte N, Mur M, Bandettini P. 2008a. Representational similarity analysis: connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4
Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, et al. 2008b. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60:1126–41
Millikan RG. 1989. Biosemantics. *J. Philos.* 86(6):281–97
Misaki M, Kim Y, Bandettini PA, Kriegeskorte N. 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage* 53(1):103–18
Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, et al. 2008. Predicting human brain activity associated with the meanings of nouns. *Science* 320:1191–95
Moreno-Bote R, Beck J, Kanitscheider I, Pitkow X, Latham P, Pouget A. 2014. Information-limiting correlations. *Nat. Neurosci.* 17(10):1410–17
Mur M, Bandettini PA, Kriegeskorte N. 2009. Revealing representational content with pattern-information fMRI: an introductory guide. *Soc. Cogn. Affect. Neurosci.* 4(1):101–9
Naselaris T, Kay KN. 2015. Resolving ambiguities of MVPA using explicit models of representation. *Trends Cogn. Sci.* 19(10):551–54
Naselaris T, Kay KN, Nishimoto S, Gallant JL. 2011. Encoding and decoding in fMRI. *NeuroImage* 56(2):400–10
Nili H, Wingfield C, Walther A, Su L, Marslen-Wilson W, Kriegeskorte N. 2014. A toolbox for representational similarity analysis. *PLOS Comput. Biol.* 10:e1003553
Norman KA, Polyn SM, Detre GJ, Haxby JV. 2006. Beyond mindreading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10:424–30
Norman-Haignere S, Kanwisher NG, McDermott JH. 2015. Distinct cortical pathways for music and speech revealed by hypothesis-free voxel decomposition. *Neuron* 88:1281–96
Olshausen B, Field D. 2004. Sparse coding of sensory inputs. *Curr. Opin. Neurobiol.* 14:481–87
Paninski L, Cunningham J. 2017. Neural data science: accelerating the experiment-analysis-theory cycle in large-scale neuroscience. bioRxiv 196949
Paninski L, Pillow J, Lewi J. 2007. Statistical models for neural encoding, decoding, and optimal stimulus design. *Prog. Brain Res.* 165:493–507
Parvizi J, Jacques C, Foster BL, Withoft N, Rangarajan V, et al. 2012. Electrical stimulation of human fusiform face-selective regions distorts face perception. *J. Neurosci.* 32:14915–20
Pereira F, Mitchell T, Botvinick M. 2009. Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage* 45(1 Suppl.):S199–209

Poeppel D. 2012. The maps problem and the mapping problem: two challenges for a cognitive neuroscience of speech and language. *Cogn. Neuropsychol.* 29(1–2):34–55
Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W. 1999. *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press
Salzman CD, Britten KH, Newsome WT. 1990. Cortical microstimulation influences perceptual judgements of motion direction. *Nature* 346(6280):174–77
Shea N. 2018. *Representation in Cognitive Science*. Oxford, UK: Oxford Univ. Press
Shenoy KV, Sahani M, Churchland MM. 2013. Cortical control of arm movements: a dynamical systems perspective. *Annu. Rev. Neurosci.* 36:337–59
Simoncelli EP, Olshausen BA. 2001. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* 24:1193–216
Tong F, Pratte MS. 2012. Decoding patterns of human brain activity. *Annu. Rev. Psychol.* 63:483–509
Tootell RB, Hadjikhani NK, Mendola JD, Marrett S, Dale AM. 1998. From retinotopy to recognition: fMRI in human visual cortex. *Trends Cogn. Sci.* 2(5):174–83
Van Gelder T. 1998. The dynamical hypothesis in cognitive science. *Behav. Brain Sci.* 21(5):615–28
van Gerven MA. 2017. A primer on encoding models in sensory neuroscience. *J. Math. Psychol.* 76:172–83
Varoquaux G, Raamana PR, Engemann DA, Hoyos-Idrobo A, Schwartz Y, Thirion B. 2017. Assessing and tuning brain decoders: cross-validation, caveats, and guidelines. *NeuroImage* 145:166–79
Walther A, Nili H, Ejaz N, Alink A, Kriegeskorte N, Diedrichsen J. 2016. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage* 137:188–200
Wilson SP, Bednar JA. 2015. What, if anything, are topological maps for? *Dev. Neurobiol.* 75(6):667–81
Wu MCK, David SV, Gallant JL. 2006. Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.* 29:477–505
Yamins DLK, DiCarlo JJ. 2016. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* 19:356–65
Yu BM, Cunningham JP, Santhanam G, Ryu SI, Shenoy KV, Sahani M. 2009. Gaussian-process factor analysis for low-dimensional single-trial analysis of neural population activity. *J. Neurophysiol.* 102(1):614–35