Turtles all the way down? Psychometric approaches to the reduction problem
Kievit, R.A.

Citation for published version (APA):
Chapter 6

Representational geometry: integrating cognition, computation, and the brain

Abstract
The cognitive concept of representation plays a key role in theories of brain information processing. However, linking neuronal activity to representational content and cognitive theory remains challenging. Recent studies have characterized the representational geometry of neural population codes by means of representational distance matrices, enabling researchers to compare representations across stages of processing and test cognitive and computational theories. Representational geometry provides a useful intermediate level of description capturing both the information represented in a neuronal population code and the format in which it is represented. We review recent insights gained with this approach in perception, memory, cognition, and action. Analyses of representational geometry can compare representations between models and the brain, and promise to explain brain computation as transformation of representational similarity structure.

This chapter is based on:

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (under review). The brain of the beholder – inferior temporal representations of visual objects are individually unique and predict perceived similarity.


* This author order reflects the editorial policy of Trends in Cognitive Sciences: The original pre-submission inquiry was ‘Kievit, R. A., & Kriegeskorte, N. Representational geometry: integrating cognition, computation, and the brain’
Introduction

The concept of representation is central to the cognitive and brain sciences. We interpret neuronal activity as serving the function of representing content, and of transforming representations of content, with the ultimate objective to produce successful behaviours. The content could be a visual image, a sound or odour, a semantic interpretation of sensory input, a proposition, a goal, a planned action, or a motor sequence. The representational interpretation (Dennett, 1988) provides a powerful explanatory framework that makes it easier to understand neuronal activity in the context of the overall function of the brain. Representation links cognition to brain activity and enables us to build functional theories of brain information processing (Forstmann, Wagenmakers, Eichele, Brown, & Serences, 2011). Neurophysiology has long interpreted the selectivity of neurons as serving to represent various kinds of sensory and higher-level information. The population of neurons within an area is thought to jointly represent the content in what is called a neuronal population code (Averbeck, Latham, & Pouget, 2006, see Appendix C for a glossary). It is the pattern of activity across the neurons that represents the content. The many possible combinations of activity states of the neurons provide a rich representational space. Motivated by this idea, recent analyses of neuronal recordings and functional imaging data have increasingly focused on patterns of activity across many neurons within a functional region (for an overview, see Kriegeskorte and Kreiman, 2012).

We can think of a region’s representation as a multidimensional space, where the dimensions correspond to the neurons and an activity pattern corresponds to a point (each neuron’s activity provides the coordinate value for one of the dimensions). A visually perceived object, for example, will correspond to a point in the representational space of a given visual area. The set of all possible objects (or pieces of mental content) corresponds to a vast set of points in the space. It is the geometry of the space of these points that defines the nature of the representation, in terms of its (dis)similarity compared to other representations to a wide variety of dimensions.

Mathematical and cognitive psychology have a long history of investigations of representational geometry on the basis of behavioural data (Edelman, 1998; Gärdenfors, 2004; Shepard, 1958; 1987; Torgerson, 1958; Tversky, 1977). However, the notion of representational geometry has only more recently been brought into the analysis of brain-activity data (Aguirre, 2007; Edelman, Grill-Spector, Kushner, & Malach, 1998; Kriegeskorte, Mur, & Bandettini, 2008; Laakso, & Cottrell, 2000; Op de Beeck, Wagemans, & Vogels, 2001)

In order to characterize the geometry of a representation, we can compare the brain-activity patterns representing a set of stimuli (or, more generally, experimental conditions) to each other. In this operationalization, the dissimilarity of two patterns corresponds to the distance between their points in the representational space. Having measured these distances, we can construct a matrix, the representational dissimilarity matrix (RDM), in which we can look up the representational distance (or dissimilarity) for each pair of stimuli (see Figure 6.1). Intuitively, the RDM tells us which distinctions between stimuli the population code honours, and which distinctions it disregards. Considering RDMs makes it very easy to compare different representations (e.g. different brain regions, a region to a computational model representation, or the same region in different individuals or species): just
compute the correlation between the RDMs (Box 1). Comparing activity patterns directly, by contrast, would require us to define the correspondence mapping between, say, the voxels of two regions, or between single neurons and the units of a computational network model, or between voxels of the same region in two individuals. Establishing this mapping can be difficult and generally requires a separate experimental data set (Haxby et al., 2011; Kay, Naselaris, Prenger, & Gallant, 2008; Mitchell et al., 2008; Naselaris, Kay, Nishimoto, & Gallant, 2011). Focusing on representational dissimilarity instead obviates the need for such a correspondence mapping.

A popular method for testing whether two classes of stimuli can be discriminated in the representation is pattern-classifier analysis (Carlson, Schrater, & He, 2003; Cox & Savoy, 2003; Haxby et al., 2001; Kamitani & Tong, 2005; Mur, Bandettini, & Kriegeskorte, 2009; Norman, Polyn, Detre, & Haxby, 2006; Haynes & Rees, 2006; Tong & Pratte, 2012). Classifier analysis, or decoding, typically focusses on binary distinctions and on revealing whether a region contains information about the class. In practice, the classifier is often linear and successful classification indicates some degree of linear discriminability between the classes. The analysis of representational geometry is complementary to classifier analysis, and goes beyond the question of discriminability of classes (and the presence of information). Two classes of stimuli can be discriminable for many different representational geometries (Figure 6.1, scenarios 2-9). However, the particular geometry matters for the computational function of the region. Beyond the question of what information is present (i.e. pieces of content distinctly represented), and what information is explicit (e.g. in the sense of being amenable to linear readout), a representation imposes a rich structure on the domain of its content. The detailed geometry of the representations of particular items can reflect their similarity, their categorical divisions, and their continuous variation along property dimensions of behavioural significance. Items that are clustered in a representation can easily be grouped together, and their differences abstracted from, when the code is read out by other brain regions. Representational geometry, thus, provides a basis for generalization and abstraction, important hallmarks of cognition (Tenenbaum & Griffiths, 2001). In fact, brain computation can be construed as the transformation of the representational geometry along stages of processing (DiCarlo & Cox, 2007).

Recent papers have reviewed results from pattern decoding of visual representations (Tong & Pratte, 2012), and pattern-information methods for testing computational models (Kriegeskorte, 2011). Here we give an overview of some of the insights from recent studies of representational geometry. The first section covers vision, the field that has been most active with this approach. The second section describes applications beyond vision, addressing other sensory modalities, memory, cognition, and action from this perspective. We then describe current challenges and future directions for studies of representational geometry.

Representational geometry in the visual system

The most rigorous account of a sensory representation is provided by a computational model that predicts neuronal responses to arbitrary stimuli. If we can accurately predict the responses of all neurons in an area, we have captured the computations up to that area. This method has been very successful for V1 and is being extended to higher-level cortical representations. If we could
predict neuronal responses throughout the brain, along with behavioural output, we might not need the abstraction of representational geometry, or indeed the concept of representation. However, even for V1, the degree to which we can predict responses is limited (e.g. Carandini et al., 2005). Predicting neural responses becomes more difficult as we move to higher-level regions. The space of computational mechanisms and model parameters becomes very complex, making it hard to consider all plausible models and to adjudicate between them with the limited amounts of data we can acquire. It is useful therefore to first seek a more abstract descriptive characterization of each area’s population code. Analyses of representational geometry have brought insights on all stages of visual representation, from early visual areas to the high-level object representations in the ventral stream.

For example, Hegdé and Van Essen (2007) found that considering the population representational geometry revealed clearer distinctions between early visual areas than single-cell selectivity measures. Neuronal population response patterns elicited by small symbols and visual patterns showed clustering according to complexity categories in V2 and V4, but not in V1 (Fig. 2a). Freeman and colleagues (2013) found a clear distinction between V1 and V2, when investigating the representation of natural textures. V2 neuronal representational distances better reflected perceived texture similarities (Ziemba, Freeman, Movshon, & Simoncelli, 2012). Human fMRI has suggested that the perceived similarity of natural textures, including metal, wood, glass, and fur, is best reflected in the representational geometry of higher ventral-stream regions (Hiramatsu, Goda, & Komatsu, 2011, Box 2, Fig. 1b). While V1 may represent local Fourier statistics, V2 and higher regions might compute higher-order statistics of V1 outputs, which are more predictive of the perceptual quality of a texture (see also Box 2, Fig. Ig, Groen, Ghebreab, Lamme, & Scholte, 2012a; 2012b).

The transformation of the representation from a low-level-feature representation to a representation that reflects perceptual qualities is a common theme among studies of representational geometry. In the domain of colour, a human fMRI study (Brouwer & Heeger, 2009) showed that the representational geometry in V4, but not V1-3, reflects perceptual colour space, although colour decoding was most accurate in V1 (Fig. 2b). This illustrates the need to reveal not only what information is present in a region, but its representational geometry, in order to understand the neural basis of perception. While textures and colours make the “stuff” of vision, a major function of the ventral stream is the visual recognition of “things” (Adelson, 2001). Ventral-stream representational geometry has been investigated using abstract parameterized shapes, which serve as a stepping stone toward real-world object images. Op de Beeck, Torfs and Wagemans (2008) found that the parameters of a simple 2D shape space were reflected in the representational geometry in monkey inferior temporal (IT) and in perceptual judgments of monkeys and humans (Fig. 2c). Human fMRI studies similarly support a representation of shape reflecting perception in lateral occipital complex (LOC) (op de Beeck et al., 2008), with the anterior LOC reflecting perceptual similarities most strongly (Hausofer, Livingstone, & Kanwisher, 2008) and lateral LOC tuned to smaller features (Drucker & Aguirre, 2009; Yamane, Carlson, Bowman, Wang, & Connor, 2008). More complex 3D shape parameterizations have also been used successfully to model IT single-neuron responses (Yamane et al., 2008) suggesting that IT represents 3D shape.
Psychometric approaches to the reduction problem

Figure 6.1. Representational geometries and their reflection in distance matrices. (a) Illustration of ten hypothetical brain regions’ representational geometries (numbered 1-10). Each dot corresponds to the representation of a particular piece of content (e.g., a visual object). The space in which the dots are placed is the space of representational patterns (illustrated as two dimensional, but high-dimensional in reality). The halo regions around the dots indicate the margin of error; dots with overlapping error halos are indistinguishable. The items fall into two categories (dark, light), or in the case of geometry 10, on a continuous manifold (graded colours). (1) No item is distinct from any other item. (2) Most items are distinctly represented, but the categories cannot be separated by any simple boundary. (3) Only the light items are distinctly represented, and they are separable from the dark items by a quadratic boundary. (4) Dark and light items are linearly separable, and arranged along parallel lines with pairs of dark and light dots matched up across the boundary. (5) The items form a single cluster, but the categories are linearly separable. (6) The items form two category clusters, which are linearly separable and within which all items are distinct. (7) Like previous, but the items in the dark category are indistinguishable. (8) Like previous, but only the category distinction is represented; items within each category are indistinguishable from each other. (9) The dark items are indistinguishable and located among the distinctly represented light items on a circle. (10) Items fall on two manifolds that closely follow each other, with pairs of items matched up across them. (b) The representational distance matrix for each of the ten geometries (in the corresponding panel location). Distances are colour-coded from blue (items indistinguishable) to red (items widely separated). Each matrix is indexed vertically (from the top down) and horizontally (from left to right) by the items as illustrated in lower left panel. Only geometry 10, lower right, has a different item set, and the upper left quadrant corresponds to the front manifold, the lower right quadrant to the back manifold. See Box 2 for actual brain representations exhibiting some of the geometrical features illustrated here.

Parameterized shapes afford good experimental control, but they lack naturalism and behavioural relevance. This has motivated the use of real-world photos, depicting faces, people, animals, objects, and scenes. Several studies using photos suggest that ventral-stream regions don’t merely represent objects in a continuous space of high-level visual features, but that they emphasize categorical boundaries and semantic dimensions of visual images in humans (Edelman et al., 1998; Haxby et al., 2001; Huth, Nishimoto, Vu, & Gallant,
Both the categorical divisions and the within-category representational geometry are strikingly similar between monkey IT and human ventral-temporal object-sensitive cortex (Kriegeskorte et al., 2008). Beyond the presence of category information (Haxby et al., 2001), several studies suggest that response patterns elicited by images of the same category form clusters in ventral-temporal response-pattern space (Edelman, 1998; Kiani et al., 2007; Kriegeskorte et al., 2008).

The major categorical divisions in neural responses are between animates and inanimates (see also Caramazza & Shelton, 1998; Naselaris, Stansbury, & Gallant, 2012) and between faces and bodies. Such clustering was not observed in either early visual representations or computational visual features of a range of complexities (Kriegeskorte et al., 2008), suggesting that the clusters do not simply reflect visual feature similarity. Instead, the ventral temporal code might be optimized to emphasize behaviourally relevant categorical divisions and semantic dimensions (see also Mitchell et al., 2008). The geometric centrality of an object in the representation has also been linked to the perception of typicality (Davis & Poldrack, 2013). Animates appear to form a representational cluster not only in IT, but also in the amygdala (Mormann et al., 2011). The representation appears to be sensitive to the dimension of animacy, even when comparing real faces and physically similar mannequin faces (Looser, Guntupalli, & Wheatley, 2012) strengthening the case for a semantic component to the code. The animate-inanimate division is associated with a large scale lateral-to-medial gradient in ventral temporal cortex, which has been demonstrated to be unaltered in congenitally blind individuals (Mahon, Anzellotti, Schwarzbach, Zampini, & Caramazza, 2009) suggesting that it does not require visual experience to develop. Finer categorical divisions within animates have also been observed in monkeys (Kiani et al., 2007) and humans (Connolly et al., 2012, Box 2, Fig. 1d).

Human ventral-temporal representational distances closely match human dissimilarity judgments (Adelson, 2001; Mur et al., 2013) which exhibit the same major categorical divisions and a similar within-category structure. However, human judgments have been found to transcend the ventral-temporal representation in that they additionally emphasize the division between human and nonhuman animals and the division between man-made and natural objects, which are not very pronounced in the ventral stream (Mur et al., 2013, Box 2, Fig. 1a).

Places (Epstein & Kanwisher, 1998) and faces (Kanwisher, McDermott, & Chun, 1997) are thought to have special behavioural relevance and specialized cortical regions dedicated to their analysis. Several studies have attempted to characterize the representational geometry of these regions. Walther, Caddigan, Fei-Fei, & Beck (2009, see also Walther, Beck, & Fei-Fei, 2012; Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011) found that natural scenes presented as either photos or line drawings could be decoded from V1, the parahippocampal place area (PPA), retrosplenial cortex (RSC), and LOC. However, only in the higher visual regions did pattern dissimilarities predict behavioural confusions. Kravitz, Peng, and Baker (2011) compared the representational geometry of scenes in early visual cortex and PPA, reporting that the early visual representation strongly reflected the distance of the perceived scene (near versus far), whereas PPA represented whether scenes were closed or open. A comparison of the representational geometries of LOC,
RSC, and PPA (Harel, Kravitz, & Baker, 2013) suggested that LOC represents the objects present in the scene, RSC the global spatial layout, and PPA combines information about both of these components of scenes. Morgan, MacEvoy, Aguirre, & Epstein (2011) found that geographic distances between landmarks of a college campus were reflected in hippocampal responses when subjects familiar with the location of the landmarks viewed them in the scanner. The hippocampus responded more strongly to changes spanning a greater physical distance. However, response-pattern dissimilarity was not significantly correlated with physical distance anywhere in the brain.

**A brief primer on representational similarity analysis**

Representational similarity analysis (RSA) is a pattern-information analysis that aims to compare representational geometries between brain regions, stimulus descriptions, conceptual and computational models, and behavioural reflections of similarity (Kriegeskorte, Mur, & Bandettini, 2008). It can be applied to functional imaging data (including fMRI, MEG and EEG) as well as neuronal-recording data. The three basic steps (Fig. 2) are as follows:

1. Choose a brain region and estimate the activity patterns. The region can be functionally or anatomically defined. The patterns can be estimated with standard methods used in univariate analyses. In fMRI, for example, a linear model with a hemodynamic response predictor for each stimulus might be used to estimate the response of each voxel to each stimulus. For neuronal recordings, a windowed spike count might be used. However, any other features of the responses, such as features reflecting temporal response structure or energy in different frequency bands, could equally be used to define the “representation” of each stimulus. (2) Estimate the representational dissimilarity matrix (RDM). The representation in a given brain region or computational model is characterized by the matrix of dissimilarities between the stimulus representations. A popular distance measure is the correlation distance (1 - Pearson correlation across voxels, neurons, features or model units). The correlation distance disregards the overall activation level (spatial mean), rendering the analysis complementary to analyses of overall activation. However, this is but one possible choice. One could also defend the position that perfectly anti-correlated measures are, although maximally dissimilar in terms of neural responses, also necessarily related in the sense that they are mutually predictive, so therefore statistically related. The choice of distance, or dissimilarity measure, is a matter of debate: Other distance measures like the Euclidean or Mahalanobis distance, or crossvalidated measures such as the discriminant t value or accuracy can also be used (e.g. see Kriegeskorte et al., 2008). (3) Compare RDMs from brains, behaviour, and models. The key step is to compare RDMs, in order to assess to what extent different representations are alike. We might want to know whether a brain representation (a) reflects stimulus properties, (b) reflects higher-level semantic properties, (c) can be accounted for by a computational model, (d) reflects representations in other brain regions, (e) is similar to a putative homologous representation in another species, (f) is reflected in behaviour, for example in similarity judgments, in stimulus confusions, or in reaction times in discrimination tasks. One useful way to compare RDMs is to compute the correlation between the corresponding dissimilarities. The rank correlation (Spearman) is often used for this purpose when a linear relationship between the dissimilarities cannot be assumed (e.g. when comparing fMRI-based RDMs to other RDMs). Statistical
inference is commonly performed by means of randomization testing (randomly permuting the stimulus labels to simulate the null distribution of the RDM correlation) and bootstrap techniques (to compare the relative performance of different models). Comparing two representations at the level of dissimilarities rather than at the level of the original patterns is a useful trick that obviates the need for defining the correspondence mapping between the representational units. Like classifier decoding, RSA is a pattern-information technique that is sensitive to information encoded combinatorially in fine-grained patterns of activity. However, rather than asking what information can be (linearly) read out from the representation, RSA attempts to characterize the representational geometry and compare it to various models. Like encoding models (also known as voxel, or population receptive field models in the fMRI literature, e.g. Dumoulin & Wandell, 2008; Kay, Naselaris, Prenger, & Gallant, 2008; Naselaris, Kay, Nishimoto, & Gallant, 2011). RSA captures the representation of a rich set of stimuli and aims to test computational models of brain information processing that generalize to novel stimuli. However, rather than comparing brains and models at the level of the activity patterns (requiring the fitting of weights that define the relationship between model units and voxels), RSA compares representations at the level of dissimilarity matrices.

**Figure 6.2.** Illustration of the steps of RSA for a simple design with six visual stimuli. (a) Stimuli (or, more generally, experimental conditions) are assumed to elicit brain representations of individual pieces of content (e.g. visual objects). Here each item’s representation is visualized as a set of voxels (an fMRI region of interest) that are active to different degrees (black-to-red colour scale). We compute the dissimilarity for each pair of stimuli, for example using 1 minus the correlation across voxels. (b) The representational dissimilarity matrix (RDM) assembles the dissimilarities for all pairs of stimuli (blue-to-red colour scale for small to large dissimilarities). The matrix can be used as a table to look up the dissimilarity between any two stimuli. The RDM is typically symmetric about a diagonal of zeros (white entries along the diagonal). RDMs can similarly be computed from stimulus descriptions (bottom left), from internal representations in computational models (bottom right), and from behaviour (top right). By correlating RDMs (black double arrows), we can then assess to what extent the brain representation reflects stimulus properties, can be accounted for by different computational models, and is reflected in behavior.
Studies of face-specific representations in humans have suggested that the fusiform face area (FFA, Kanwisher et al., 1997) emphasizes the distinction between faces and non-faces, whereas anterior temporal cortex discriminates individual faces (Fig. 2c, Anzellotti, Fairhall, & Caramazza, 2013; Kriegeskorte, Formisano, Sorger, & Goebel, 2007; Nestor, Plaut, & Behrmann, 2011). However, individual-level face information has also been reported in the FFA and other regions (Anzelotti et al., 2013, Goesaert & Op de Beeck, 2013; Nestor, Plaut, & Behrmann, 2011). The best demonstration of the transformation of face representational geometry across face regions comes from fMRI-targeted neuronal recordings in monkeys (Freiwald & Tsao, 2010, Box 2, Fig. Ic). The representation of identities was found to be view-specific in the middle face patches (ML and MF), partially view-tolerant in anterior face patch AL with mirror-symmetric views co-localized in the representational space, and almost view-invariant in the most anterior face patch AM. A mirror-symmetric representation of face views has also been reported in humans (Kietzmann, Swisher, König, & Tong, 2012).

Many higher visual areas, including LO, OFA, FFA, PPA, and dorsal regions showed a mirror-symmetric response to faces. For example, views of -60 and 60 degrees elicited similar response patterns, suggesting pooling of mirror-symmetric lower-level features as a computational step toward greater tolerance to view changes. Achieving tolerance to stimulus variations that are not relevant to a given task such as face identification is one of the central challenges of object recognition (DiCarlo, Zoccolan, & Rust, 2012). Although ventral-stream representations are not fully invariant (Kravitz, Kriegeskorte, & Baker, 2010) to the view, position, or scale of an object, they support linear readout with some robustness to these accidental properties (Hung, Kreiman, Poggio, & DiCarlo, 2005). In order to understand, how tolerance is achieved, Rust and DiCarlo (2010) compared the population representational geometries of V4 and IT on the basis of neuronal recordings in monkeys. Both regions discriminated individual images and their scrambled counterparts. IT exhibited reduced information about scrambled images, but increased generalization for intact images across position and context clutter. This suggests a transformation of representational geometry in which position and clutter is de-emphasized relative to the presence of complex feature conjunctions diagnostic of the differences between real-world objects.

A special case of clutter is the presence of additional objects in the image. Several studies have investigated ventral temporal response patterns to multiple simultaneously presented objects in monkeys (Zoccolan, Cox, & DiCarlo, 2005) and humans (MacEvoy, & Epstein, 2009; Reddy & Kanwisher, 2007; Reddy, Kanwisher, & VanRullen, 2009). Results suggest that the response pattern elicited by multiple objects is well predicted by average of the response patterns (Zoccolan et al. 2005). If one of the objects is attended, this object receives a greater weight in the average (Reddy et al., 2007). These results are broadly consistent with biased competition, or, more generally, the divisive normalization model (Carandini & Heeger, 2012) in which the summed neural population response to multiple inputs is held constant through recurrent suppression. As we observe people in the real world, we don’t just recognize them. We infer a host of socially important information. Facial expressions of emotion appear to be represented in posterior STS with the representational
geometry reflected in similarity judgments of the expressions (Said, Moore, Engell, Todorov, & Haxby, 2010).

The STS and medial prefrontal regions have been suggested to represent perceived emotions with invariance to the sensory source (dynamic faces, bodies, or voices; Peelen, Atkinson, & Vuilleumier, 2010). Another socially important feature of faces is gaze direction. The displacement of the dark iris and pupil provides a subtle, but socially important, indication of where someone is looking. Carlin, Calder, Kriegeskorte, Nili, and Rowe (2011) reported a representation of gaze direction in right anterior STS, which was tolerant to changes of both head view and physical image features (Box 2, Fig. 1e).

Visual representations are thought to be not merely perceptual, but also involved in mental imagery in the absence of visual input. A number of human fMRI studies have investigated the brain representations during visual imagery and their relationship to perceptual representation. Results support the idea that imagery and perception of the same visual content might be represented in the same cortical regions and in a similar representational format (Cichy, Heinzle, & Haynes, 2012; Lee, Kravitz, & Baker, 2012; Reddy et al., 2007; Stokes, Thompson, Cusack, & Duncan, 2009). Overall, studies of the geometry of visual representations have impressively documented the stage-wise transformation of the retinal image from low-level representations of local physical features to high-level holistic representations of objects, faces, and scenes that better reflect perceptual qualities, emphasizing behaviourally important categories and semantic dimensions and deemphasizing accidental and behaviourally irrelevant variation of the visual input. Further studies are needed to reveal the full computational mechanism giving rise to these transformations.

**Auditory perception**

Like vision, audition requires substantial tolerance to accidental variations of the signals to be recognized. The imposition of categorical boundaries on a fundamentally continuous space of stimuli is another shared feature between the two modalities. Giordano, McAdams, Zatorre, Kriegeskorte and Belin (2012) investigated the representation of natural environmental sounds and reported that the representational geometry in the planum temporale emphasized particular categorical distinctions more strongly than predicted by low-level feature models. The categories were living/non-living and human/non-human – highly behaviourally relevant divisions similar to those emphasized in ventral-visual representations (see also Staeren, Renvall, De Martino, Goebel & Formisano, 2009). The clearest examples of categorical representation might be expected in the domain of human language. Speech percepts are categorical not only at the level of conceptual and semantic content, but also at the phonetic level.
A well-known example is the categorical perception of phonemes. A recent study (Chang et al., 2010) used human intracranial electrode arrays to investigate the representational geometry of a continuum of artificial speech sounds ranging from 'ba' through 'da' to 'ga'. Response patterns in posterior superior temporal gyrus formed clear clusters corresponding to phonemes, despite the fact that the sound continuum was sampled in acoustically equal steps. Pattern dissimilarity emerged rapidly over time, peaking at the same time
(~110 ms) as the evoked potentials. A human fMRI study (Raizada, Tsao, Liu, & Kuhl, 2010) investigated the representations of the phonemes /la/ and /ra/ in native speakers of English and Japanese. The representational dissimilarity between /la/ and /ra/ phonemes in the right primary auditory cortex, but not the overall activation, predicted the extent to which speakers were able to discriminate between the two phonemes, between and even within the two groups. Beyond vision and audition, a close link between representational geometry and perception has also been observed for olfactory stimuli, where behavioural similarity ratings of smells correlated with neural pattern representational similarities in the posterior piriform cortex (Howard, Plailly, Grueschow, Haynes, & Gottfried, 2009). Across several sensory modalities, studies of representational geometry have demonstrated that brain representations emphasize behaviourally relevant categorical distinctions and predict perceptual similarities.

Memory
A growing number of studies is investigating representations of particular items during memory encoding and retrieval, and how the precision of representational reinstatement during encoding and retrieval predicts the success or failure of these memory operations. Polyn, Natu, Cohen, & Norman (2005) showed that the category-specific patterns arising during the encoding of faces, locations, and objects are reinstated in a subsequent free-recall phase. Activity patterns during free recall predicted the category about to be recalled several seconds in advance.

A more recent study by Xue and colleagues (2010) investigated the encoding process and found that more precise perceptual reinstatement of representations during encoding predicted better memory. Interestingly, the precision of perceptual reinstatement of representations has also been associated with conscious representation (Schurger, Pereira, Treisman, & Cohen, 2010). Subsequent studies suggested some refinements to this picture, in which activity patterns in the hippocampus and other medial temporal regions have distinct signatures that predict successful encoding (LaRocque et al., 2013) and retrieval (Ritchey, Wing, LaBar, & Cabeza, 2012; Staresina, Henson, Kriegeskorte, & Alink, 2012).

In the studies described so far, memory served as a storage facility, but the structure of the representational categories was never manipulated. A recent study (Visser, Scholte, & Kindt, 2011) associated particular exemplars from each category with a shock. This led to the emergence in the frontoparietal representation of a new categorical division between the shock-associated and the other images. Moreover, the patterns during fear learning predicted the long-term behavioural expression of fear memories (Visser, Scholte, Beemsterboer, & Kindt, 2013). In sum, studies of representational geometry are beginning to reveal how the perceptual representation of an individual piece of content affects its mnemonic encoding and how the reinstatement of the representation during recall enables successful recollection. We are also beginning to elucidate the specific roles of the hippocampus and other medial temporal regions and the plasticity of the representational space itself, including the formation of new behaviourally relevant categorical divisions.
Representational geometries in recent studies

Inspecting the representational dissimilarity matrices (RDMs) characterizing brain regions, various models, and behavioural data is an important exploratory process that can reveal interesting and unexpected representational structure. In Figure 6.4 we consider RDMs from a wide range of recent studies. Note that prior hypotheses and statistical inference, as reported in the original papers, are required to support any theoretical conclusions. (4a) The RDMs of human IT and explicit behavioural dissimilarity judgments reveal related, yet distinct structures. Both IT and judgments emphasize the animate/inanimate and the face/body divisions. However, the judgments additionally emphasize the human/nonhuman and the man-made/natural divisions (Mur et al., 2013). Note that the categorical structure is obvious only because the stimuli are ordered by category. In contrast to 2D arrangements by representational dissimilarity (Fig. 2), which do not depend on any choice of stimulus order, the appearance of an RDMs depends on the order chosen for display. (4b) The early visual representation of natural textures, including metal, ceramic, and glass, resembles a model based on low-level image features. The representation in the fusiform gyrus is distinct from the early visual representation and more consistent with human perceptual similarity ratings (Hiramatsu, et al. 2011) (4c) RDMs from neuronal recordings in macaque middle and anterior face patches illustrate the transformation of the representational geometry across stages of processing.

Neuronal population response patterns cluster by face view in the middle face patches (left RDM). The blocks of similar patterns correspond to faces of different identities in the same view. By contrast, the anterior face patch (right RDM) exhibits strong view-tolerance and selectivity for individual identities. Each identity elicited similar response patterns when presented in different views (subtle dark diagonal lines) (Freiwald & Tsao, 2010) (4d) Images of animals from six biological species are represented distinctly in early visual and lateral occipital cortex. The early visual RDM resembled that from a computational model of V1. The lateral occipital RDM resembled judgments of biological similarity, which reflected the categories (insects, birds, primates) (Connolly et al., 2012). (4e) An RDM model of gaze direction is contrasted with two competing models: a low-level model of physical stimulus features and a categorical model of gaze direction relative to the observer. Searchlight representational similarity analysis (not shown) revealed that anterior STS might represent gaze direction with tolerance to head view (Carlin et al., 2011) (4f) RDM of whole-brain activity patterns during pain perception and other mental states. The RDM reveals the similarity in terms of global brain activity of eight mental states, suggesting shared recruitment of specialized brain regions in pain, emotion, interoception, and reward (Cauda et al., 2012). For a comprehensive meta-analytical framework for analysing the discriminability of mental states from whole-brain activity patterns, see (Yarkoni, Poldrack, Nichols, Van Essen, & Wager, 2011). Note that the representational interpretation appears less natural for whole-brain patterns reflecting various kinds of task processing than for the localized representations in the other studies discussed here. However, similar caveats apply at both levels of analysis. (4g) The RDM for a set of natural images (based on spatial human EEG patterns 101 ms after stimulus onset) is substantially correlated with an RDM predicted by a model of spatially pooled image-contrast statistics, namely the parameters of a Weibull fit to the distribution of local contrast measurements (Groen et al., 2012a).
Action and motor control

We should be able to characterize representations along the entire perception-action cycle by their representational geometry. The primary motor representation is a classic example of a population code (Georgopoulos, Schwartz, & Kettner, 1986). We focus here on two recent studies that have explored the geometry of motor representations. Wiestler, McGonigle and Diedrichsen (2011) studied the representation of finger movements and sensations in the cerebellum, a structure thought to relate sensory and motor representations for smooth sensorimotor control. They compared cerebellar regions to cortical areas M1 and S1. Cerebellar and cortical representations both discriminated movements of different individual fingers. Both also discriminated sensations in different individual fingers. For a given finger, movement and sensation were associated with similar patterns in M1 and S1. In the cerebellum, however, the representations were not consistent between movement and sensation, with motor and sensory finger representations apparently randomly interdigitated.

Figure 6.4. Representational dissimilarity matrices in recent studies. (a) RDM from fMRI patterns in human IT and RDM based on human dissimilarity judgments. Reproduced, with permission, from Mur et al., 2013. (b) RDMs from fMRI responses to natural visual textures, from a model of low-level image statistics, and from human similarity ratings. Reproduced, with permission, from Hiramatsu et al. 2011. (c) RDMs from neuronal recordings for monkey middle face patches (middle lateral and middle fundus, 121 neurons) and anterior face patch (anterior medial, 158 neurons). The stimuli are faces of different identities and views. One per view labels a set of rows and columns for different identities in that view. Reproduced, with permission, from Freiwald and Tsao 2010. (d) RDMs from fMRI responses to images of animals from six species falling in three biological categories, and from a computational model of V1 and subject judgments of biological similarity. Reproduced, with permission, from Connolly et al. 2012. (e) RDMs for three different models of representation of faces and eye positions: (1) by gaze direction, (2) by head view (ignoring eyes), (3) by gaze categories relative to the observer (direct/eye contact, left, right). Reproduced, with permission, from Carlin et al. 2011. (f) Whole-brain activity pattern dissimilarities between different functional states. Global patterns were estimated meta-analytically. Reproduced, with permission, from Cauda et al. 2012. (g) RDMs based on spatial patterns of human event-related-potential amplitudes evoked by natural images 101 ms after stimulus onset (top) and a model based on parameters of a Weibull fit to the spatially pooled distribution of local contrast measurements (bottom). Reproduced, with permission, from Groen et al. 2012b.
This arrangement may enable the cerebellum to associate movements and their sensory consequences in flexible ways, a requirement of learning new motor tasks. In another study, Diedrichsen, Wiestler and Krakauer (2012) investigated co-localized motor representations of our two hands, which might serve to coordinate the hands during bimanual tasks. Representations of unimanual finger movements were represented mostly in contralateral M1 and S1, with a faint echo of a symmetrical representation in the ipsilateral areas. Such an arrangement might facilitate symmetrical bimanual movements. In premotor and parietal areas, unimanual movements also had an ipsilateral representation. However, it was not a symmetrical echo, but qualitatively different. Such a co-localized representation of both hands might serve to associate the movements of the two hands in flexible ways, so as to coordinate the hands during asymmetrical bimanual tasks.

**Current challenges for investigations of representational geometry**

**Testing many models: a simultaneously hypothesis- and data-driven approach**

Many studies have focused on one or two models of the representation. In the visual domain, this could be a particular computational model, such as a Gabor-filter model, a parameterized shape model, a semantic model, or a behavioural characterization of the representational geometry. The field is in an early phase where finding that a model explains significant variance (of neuronal or voxel responses, or response-pattern dissimilarities) is considered an advance. This is a low bar. The theoretical advance is not always substantial, because a great number of qualitatively different models may capture some component of the representation. For theoretical progress, we need statistical model comparisons along with estimates of the amount of non-noise variance left unexplained by each model. Ideally, we would like to cover the entire space of models that have not yet been strictly eliminated. Realistically, we may want to focus on a range of models that are qualitatively different in a single study. These should include models we have strong predictions for (so as to put current opinion to the test) as well as models whose relative performance we have no strong predictions for (so as to go beyond the state of the literature and advance theory). In fMRI studies, we can test our models in a variety of regions of interest, or continuously throughout the measured volume using a searchlight approach (Kriegeskorte, Goebel, & Bandettini, 2006). Such more exploratory analyses represent another major source of important information. We can combine exploration and confirmation in a single study by using multiple-testing correction and crossvalidation.

Testing a wide range of well-motivated models in multiple brain regions constitutes an approach that is simultaneously strongly hypothesis- and strongly data-driven (see also Kievit et al., 2011). This is the kind of approach we need to bridge the divide between the modelling and experimental literatures, and to richly constrain brain theory with empirical data. We must resist two temptations: (1) to shy away from disconfirmation of prevailing bias and (2) to restrict our analyses so as to be able to tell a better “story”. A story
that crumbles upon consideration of a broader view of the available evidence (i.e. broader analyses of our own data) is clearly not worth telling.

**Engaging temporal dynamics**
Representations are inherently dynamic, emerging over time as evidence from the sensory stream accumulates, or recurrent computations converge on an interpretation of the input. At the level of higher cognition, thoughts emerge dynamically through the interplay of perception, long-term memory associations, and the current contents of working memory. Representational dynamics can be investigated by analysing the representational geometry with a temporal sliding window. Recent studies have begun to move in this direction using sliding-window decoding techniques on neuronal recordings (Hung et al., 2005) and human MEG (Carlson, Hogendoorn, Kanai, Mesik, & Turret, 2011; Su, Fonteneau, Marslen-Wilson, & Kriegeskorte, 2012) and EEG data (Groen et al., 2012b)

**Improving characterizations of representational geometry**
We might seek improvements to the approach of comparing representations by the rank correlation between their distance matrices. This approach does not require the (often questionable) assumption of a linear relationship between the distances, and it deals gracefully with the fact that distances estimated from noisy pattern estimates are generally positively biased. However, it also discards potentially important geometrical information. A promising complementary approach to comparing representations based on noisy data is pattern component modelling, where the pattern variance is decomposed into components that correspond to experimental factors and noise (Diedrichsen, Ridgway, Friston, & Wiestler, 2011). The analysis of representational dissimilarity matrices is one important tool for understanding representational geometry. However, linear decoding remains useful as a straightforward test of linear separability (a particular feature of the representational geometry) and of the degree to which the linear decision boundary generalizes to different stimuli. Nonlinear decoders based on quadratic boundaries or radial basis functions (Poggio & Girosi, 1990) similarly reveal information that is available for immediate readout and might thus be considered “explicit” in the code, just like information available for linear readout. We will need a repertoire of specific models to test for a range of computationally relevant properties of the representational geometry.

**Considering different population codes and representational distance measures**
Current studies have largely defined the representations as spatial activity patterns (e.g. from fMRI or windowed spike counts), with the activity level in each voxel or neuron contributing one dimension to the space. However, population-coding theory has explored a much wider range of possible codes, including temporal codes. For example, the representational dissimilarity of two stimuli could be measured by comparing the temporal order in which neurons fire their first spike in response to the stimulus (VanRullen & Thorpe, 2001) or relative to the Gamma cycle (Fries, Nikolic, & Singer, 2007). Another approach would be to use spatiotemporal or time-frequency patterns to define the representational space. In addition, a range of distance measures based on neuronal population spike trains deserves to be explored (Houghton & Victor, 2010).
Individual differences
One of the challenges of understanding “the brain” (or “the mind”) is that it is not a unitary object. Everyone is different, and our brains change continuously. If we are to understand the mind or the brain, we must be interested in inter-individual variation and plasticity. Previous studies have focused largely on commonalities, both across subjects (e.g. Haxby et al., 2011; Shinkareva et al., 2012) and across species (e.g. Kriegeskorte et al., 2008), suggesting that the neural representation of visual stimuli is largely shared across people (Haxby et al., 2011; Raizada & Connolly, 2012; Shinkareva, Malave, Just, & Mitchell, 2012). However, this does not rule out the possibility of stable idiosyncrasies at lower levels of resolution. More recent work has focused on the individual differences in visual object representations, and to which extent the idiosyncrasies converge across behavioural and neural domains. Charest, Kievit, Schmitz, Deca and Kriegeskorte (under review) introduced a novel method called individual Representational Similarity Analysis (iRSA) to examine the extent to which there is a replicable, individually unique component to the representational geometry of each individual. They studied the neural responses of 20 people across two scanning sessions, separated by two weeks, as they perceived 72 different visual stimuli. The neural responses were measured in the early visual cortex, where visual information is initially processed, and in the inferior temporal cortex, commonly associated with higher level processing of visual categories. In addition to the measurement of the neural responses, subjects performed a behavioural test, rating the dissimilarity between each pair of stimuli (Kriegeskorte & Mur, 2012). Pattern analysis of the neural responses (see Figure 6.5) showed that the neural patterns were individually unique in both EVC and IT (that is the representational geometry, characterized by the dissimilarity of the activity patterns across all pairs of stimuli, was more similar within subjects than between subjects, when compared across scanning days). This suggests that individual idiosyncrasies in the neural representations of visual stimuli can be captured by iRSA. Secondly, the structure of the representational geometry of the neural responses predicted behavioural similarity judgments, but only in IT, not in EVC. This suggests that our subjective experience of visual objects, to the extent that it affects conscious behavioural judgments, is better reflected in higher visual areas. Finally, in addition to the individually unique detailed geometry, the neural representational dissimilarity pattern replicated previously found categorical clustering, with neural patterns in hIT showing clustering into meaningful categories at both more higher (animate/inanimate) and slightly lower (faces, places, objects and bodies) levels of abstraction. This shows that neural representations can have structured similarity across individuals, predictive of behavioural judgements, whilst retaining a unique component at lower levels predictive of individual differences in behavioural ratings. In the next section we show how this approach can be a useful general tool to examine theoretical positions discussed before, including identity theory and multiple realizability.
The framework of representational geometry is useful not just within traditional empirical cognitive neuroscience, but can be applied to theoretical discussions as discussed in previous chapters, such as the nature of identity theory. In chapters 2, 3 and 4, we examined the empirical corollaries of identity theory when relating structural brain properties to behavioural dimensions that differentiate between people (e.g. intelligence). However, the notion of identity theory becomes much more complex when studied in the context neural activity and behavioural dynamics over time. For instance, the size of the visual cortex can differ across individual differs by factor of 2.5 (e.g. Stensaas, Eddington, & Dobelle, 1974). This means that in a strict sense, it is impossible for two neural patterns to be completely identical, either between people (because they have different brains) or even within people (because every neural impulse affects the fine-grained details synaptic connectivity). Within the context of representational geometry, a more flexible and insightful interpretation of identity theory is possible that ignores irrelevant anatomical variations whilst retaining the conceptual weight of identity theory. We propose a geometrical perspective on identity theory for (behavioural and neural) representations as follows. Firstly, we propose that a psychological representation (e.g. the subjective experience of a picture of a face) is identical to a neural
representation if the position of that representation relative to other psychological representations is identical to the relative position of the concurrent neural representation in the neural domain. In other words, it is the fact that the structure of the neural representational geometry is identical to the structure of the behavioural pattern (across stimuli) that can be considered as evidence for identity theory. That is, if the relative position of a stimulus in some abstract space, compared to other stimuli (quantified by some similarity metric as discussed above), is the same regardless of whether we quantify it by neuroscientific or psychological metrics, we can consider them to display identity theory.

Secondly, we propose that an identity theory perspective on neural activity should allow for some degree of stochasticity. As the measurable variables, both behavioural and brain measurements, are inherently imperfect reflections of the underlying pattern (under the assumption that the underlying pattern is identical), this variability should be incorporated so as to rule out stronger forms of identity theory (type-token and type-type). Together, these two requirements lead to the following interpretation of an identity theory perspective on representational geometry: Two representational geometries, measured in two domains (e.g. behavioural judgments and neural responses, B and N) can be said to be the identical if the representational geometry we can use to describe them are as similar to each other as they are to themselves (e.g. by rank order correlation: r\text{BN} = r\text{N1N2}). The latter, self-similarity, (sometimes described as internal reliability) of neural geometry could be quantified by, for instance, a split-halves procedure. This perspective is represented in Figure 6.6.

An additional benefit of this approach is that multiple forms of identity theory can be naturally incorporated. As detailed in chapter 2 and 3, there are various forms of identity theory (e.g., Aydede, 2000). The most strict form, type-type identity theory, holds that every type of mental state (e.g. ‘being in pain’) corresponds to a type (i.e. a specific set of structurally identical group of neurons firing, such as ‘C-fibers’). Type-token identity theory holds that there is a specific type of neural activity associated with a mental state, but that that type differs between people. Token-token identity theory holds merely that whenever someone has a certain mental state, this corresponds to a certain neural state, but there need not be any consistency in this mapping.

The empirical results discussed above have shown how individual representations can be compared between brain and behaviour, and across individuals, and provide evidence to suggest that representations are similar across people (in terms of the overarching, broader categorical structure) yet also unique (the fine-grained details are idiosyncratic and replicable across sessions). We suggest that the results discussed above, including Kriegeskorte and colleagues (2008), Raizada & Connolly (2012) and Charest and colleagues provide tentative support for a hierarchical view of identity theory for visual representations. At the highest level of abstraction, such as the categorical distinction between animate and inanimate objects, we find evidence for a form of type-type identity theory: The representational clustering of these categories has been found in human behaviour (Charest et al. under review, Mur et al.,

---

26 Incidentally, this approach naturally accommodates multiple realizability (see Chapter 2 and 3): It is not the nature of the material substrate that counts, but the structure of the representational geometry.

27 To ensure that a trivial correlation of 0 does not count, one can set a criterion of minimal evidentiary strength for a non-zero correlation, e.g. a bayes factor > 10 (Wetzels & Wagenmakers, 2012).
Turtles all the way down?

Figure 6.6. Proposed representational model of identity theory. The underlying geometry is reflected, imperfectly, in both brain and behaviour. To assess identity at the latent level, the rank-order correlations of the RDM’s should be as high within a domain (reliability) as it is between two domains (for a given person).

Although this is a weaker form of identity theory than traditionally construed, as the exact neural firing is allowed to be different, it is nonetheless a useful conceptualization. The fact that evolutionarily adaptive visual categorizations such as animate/inanimate can be found consistently, across species and modalities, is in line with the relatively strong thesis of type identity theory. If we go down a level of abstraction (see dendrogram in Figure 6.6), examining the detailed representations of individual stimuli, we can consider the
results of Charest and colleagues (under review) to be partial evidence for type-token identity theory: Within people, there is a stable mapping of the geometry of neural responses to the geometry of behavioural judgments, but this pattern varies across individuals. At the lowest level, the repetition of individual stimuli, the data is most compatible with token-token identity theory: Even within a person with a stable representational structure, every instance of a stimulus is slightly different, as every repetition affects synaptic weights, however slightly (e.g. Garlick, 2002).

This hierarchical perspective captures both the informative regularities across individuals and even species for evolutionarily essential categories and the unique yet replicable within person idiosyncrasies. Of course these are initial empirical results, and not all patterns are clear cut. For instance, the within-subject similarity across sessions is considerable, but not perfect, nor is the between subject mapping always homogeneous: For instance, explicit similarity ratings can induce novel categories not seen when studying the neural patterns (Mur et al., 2012). Nonetheless, we think the study of representational geometry is one of the most promising frameworks to unify brain activity with behaviour, as it successfully finds the middle ground between abstracting away from the irrelevant details whilst retaining the fine-grained structure predictive of behaviour. This framework and these initial results should therefore be seen as the implementation of Heuristic Identity Theory (McCauley & Bechtel, 2001): By assuming there exists a mapping between brain activity and behavioural geometry, we can guide empirical research to ask important questions concerning the structure of our representations of the world. This empirical perspective on identity theory has been dubbed Heuristic identity Theory (Bechtel & McCauley, 2001). Future work (e.g. Kievit, Charest, Kriegeskorte, & Borsboom, in preparation) will further develop the theoretical and empirical predictions that follow from this framework.

**Conclusion**

The study of representational geometry has provided insight into such fundamental cognitive processes as memory (e.g. greater pattern similarity is associated with improved recall), conscious processing (consecutive consciously perceived stimuli are more similar than subliminally perceived stimuli) subjective similarity judgments (e.g. anterior LOC reflects perceptual similarity but posterior LOC reflects physical stimulus similarity), and related domains including vision, audition, categorical perception, emotion and motor control. Representational geometry can serve as a hub that enables us to connect experiment and theory at the crucial intermediate level of description, where computational network models meet massively multivariate data from electrophysiological population recordings and high-resolution functional imaging. Moreover, representational geometries can be compared between stages of processing, between model and brain representations, between individuals and species, and between brain and behaviour. This allows not only for empirical progress but for the empirical testability of hitherto metaphysical theories such as identity theory. Future studies will systematically test wider ranges of models, including computational models, better integrate population-coding theory, reveal representational dynamics, and elucidate representational plasticity and the individually unique component of representational geometries.