OXFORD

Social Cognitive and Affective Neuroscience, 2021, 795–806

doi: https://doi.org/10.1093/scan/nsab010
Advance Access Publication Date: 27 January 2021
Original Manuscript

Special Issue: Computational Methods in Social Neuroscience

Multivariate spatial feature selection in fMRI

E. Jolly and L.J. Chang

Computational Social Affective Neuroscience Laboratory, Department of Psychological and Brain Science, Dartmouth College, Hanover, NH 03755, USA

Correspondence should be addressed to E. Jolly, Computational Social Affective Neuroscience Laboratory, Department of Psychological and Brain Science, Dartmouth College, 6207 Moore Hall, Hanover, NH 03755, USA. E-mail: eshin.jolly@dartmouth.edu.

Abstract

Multivariate neuroimaging analyses constitute a powerful class of techniques to identify psychological representations. However, not all psychological processes are represented at the same spatial scale throughout the brain. This heterogeneity is apparent when comparing hierarchically organized local representations of perceptual processes to flexible transmodal representations of more abstract cognitive processes such as social and affective operations. An open question is how the spatial scale of analytic approaches interacts with the spatial scale of the representations under investigation. In this article, we describe how multivariate analyses can be viewed as existing on a spatial spectrum, anchored by searchlights used to identify locally distributed patterns of information on one end, whole brain approach used to identify diffuse neural representations at the other and region-based approaches in between. We describe how these distinctions are an important and often overlooked analytic consideration and provide heuristics to compare these different techniques to choose based on the analyst's inferential goals.

Key words: multivariate; feature-selection; searchlight; biomarker; decoding; fMRI

Introduction

The past decade has witnessed an explosion in empirical studies employing advanced statistical methods to understand brain representations. Traditional univariate analyses of functional magnetic resonance imaging (fMRI) data have historically focused on differences in magnitudes of activation (Friston *et al.*, 1995), while more contemporary approaches have explored how spatial patterns of activity encode psychological information (multivariate pattern analysis; MVPA) (Haxby *et al.*, 2014) and how the temporal dynamics of neural responses are shared across individuals (intersubject correlation; ISC) (Nastase *et al.*, 2019). Unlike univariate techniques that independently model each voxel, these modern techniques often involve aggregating responses across multiple voxels during the modeling process (e.g. searchlights, regions of interest (ROIs) or whole brain). An underappreciated consideration when using these approaches is the spatial scale at which these analyses are performed. In this article, we will discuss how different psychological and cognitive processes may be reflected at different spatial scales and how this might impact choices in the analysis pipeline. We begin by exploring evidence for spatial-scale heterogeneity, then compare and contrast the most commonly employed techniques and conclude with practical considerations for choosing methods best suited for different research questions.

Spatial scale of representations in the brain

Many contemporary fMRI methods focused on mapping brain representations or modeling neural synchrony require selecting

Received: 22 May 2020; **Revised:** 16 September 2020; **Accepted:** 25 January 2021

specific spatial features to be used in an analysis (e.g. fMRI decoding, encoding, representational similarity analysis (RSA), ISC, intersubject RSA (Naselaris *et al.*, 2011; Diedrichsen and Kriegeskorte, 2017; van Baar *et al.*, 2019; Chen *et al.*, 2019; Nastase *et al.*, 2019; Finn *et al.*, 2020)). In this context, features refer to the specific information that is entered into a model (e.g. a group of voxels, the average activity in a cortical region or a neural distance matrix) and used to make inferences about a specific process, representation or psychological state. Numerous published papers have made general recommendations about setting up and interpreting analyses with different techniques (e.g. Haynes, 2015). However, these guides primarily make recommendations based on statistical considerations such as the interpretability of decoding accuracy (Etzel *et al.*, 2013), or highlight what contemporary techniques offer beyond simple univariate contrasts of brain activity (Kriegeskorte and Bandettini, 2007).

A key consideration often missing from these discussions is the spatial variability with which different kinds of neural and/or psychological information may be represented in the brain (Kragel *et al.*, 2018). For example, considerable evidence stemming from neuronal recordings, univariate fMRI studies, neuropsychological investigations, computational modeling and animal studies has demonstrated a reliable functional organizational scheme for sensory systems, with a particular focus on the visual system (Felleman and Van Essen, 1991; Grill-Spector and Malach, 2004; Hubel and Wiesel, 2004; Yamins *et al.*, 2014). This modular organizational structure has served as a scaffold for much contemporary research and has also importantly impacted the analytic approaches used to make scientific discoveries. The structure of the visual system affords researchers the ability to test specific predictions and build models at fine spatial scales. Some notable examples include direct recordings of populations in preselected cortical patches (Chang and Tsao, 2017), or using local patterns of neural activity to topographically map how representations change and transform as information moves through the visual system (Kriegeskorte *et al.*, 2006). It has also been a key driver of highly sophisticated contemporary work such as comparing features learned by layers of deep neural networks to neural representations in different stages of the ventral visual stream (Kriegeskorte, 2015; Cichy *et al.*, 2016; Yamins and DiCarlo, 2016). This scale of analysis comports well with consensus understanding of how perceptual systems are organized and is well-suited for examining the brain through the lens of functional compartments or locally distributed populations of activity (Haxby *et al.*, 2014; Kragel *et al.*, 2018).

In parallel, a large body of work has taken a more macroscopic view of brain organization by examining how diffusely distributed representations and networks subserve different cognitive functions by dynamically adapting to the task at hand (Kragel *et al.*, 2018). At this spatial scale, cortical areas can be seen as belonging to various subtypes such as primary sensorimotor, unimodal associative, transmodal associative, paralimbic and limbic (Mesulam, 1998). These subtypes demonstrate independent patterns of functional connectivity at rest (rsfMRI) and can be used to parcellate the brain into distinct networks (Power *et al.*, 2011; Yeo *et al.*, 2011; Glasser *et al.*, 2016; Schaefer *et al.*, 2016). Interestingly, several groups have demonstrated that subtypes of cortex vary markedly in the similarity between their structural and functional connectivity (Honey *et al.*, 2009). For example, functional connectivity most closely resembles anatomical connectivity and microstructural properties in sensory and unimodal regions, but this resemblance breaks down in transmodal areas such as the default mode network (DMN) (Paquola *et al.*, 2019; Vázquez-Rodríguez *et al.*, 2019). Further, the variability in functional connectivity patterns appears to be organized around functional gradients that range from unimodal primary sensory regions to transmodal associative regions (Margulies *et al.*, 2016). In other words, neural activity at rest is organized in a manner consistent with the geometric structure of the brain. Brain regions farther away from primary sensory areas are responsible for less externally focused computations and more abstract modes of cognition (e.g. associative, multimodal and internally directed). Transmodal regions often exhibit less hierarchical organization, denser interconnectivity, more top-down projections between cortical layers and less laminar differentiation, which are believed to facilitate more abstract and flexible responding to different kinds of information (Paquola *et al.*, 2019; Vázquez-Rodríguez *et al.*, 2019).

The contrast between these domains serves to highlight the breadth of spatial scales at which the brain represents and supports different psychological and cognitive functions. If tight, localized, hierarchical organization of primary sensory systems represent one end of this range, the other appears to be a more spatially diffuse, abstract and flexible organization of transmodal areas. In the field of social and affective neuroscience, there appears to be a network of brain regions, overlapping with the DMN, thought to reliably support socio-emotional processing (Lieberman, 2007; Adolphs, 2009). An open question, however, is whether the functional organization of these regions resembles primary sensory systems with circumscribed functional subdivisions, or a more general structure such that all regions support socio-emotional cognition by flexibly adapting their responsibilities to the particular task at hand.

There is some evidence that this social brain network may contain distinct cortical areas, patches and populations of neurons with highly circumscribed responsibilities functionally tuned to specific aspects of a socio-emotional experience, akin to functional specificity in primary sensory systems (Adolphs, 2009). Meta-analyses of the medial prefrontal cortex (mPFC), for example, posit the existence of distinct subdivisions for cognitive and emotional tasks (Amodio and Frith, 2006; De La Vega *et al.*, 2016) and a dorsal to ventral gradient which delineates representations about others or the self, respectively (Mitchell *et al.*, 2006; Wagner *et al.*, 2012; Sul *et al.*, 2015). The temporoparietal junction (TPJ) has been strongly associated with theory of mind and specifically reasoning about others' beliefs and intentions as distinct from their feelings and emotions (Saxe and Kanwisher, 2003; Peelen *et al.*, 2010; Young *et al.*, 2010; Carter *et al.*, 2012; Koster-Hale *et al.*, 2017), akin to the relationship between the fusiform gyrus and face processing (Kanwisher *et al.*, 1997). However, subdivisions within this area show different patterns of functional connectivity with the rest of the brain, suggesting distinct local representations despite cortical proximity (Mitchell, 2008; Mars *et al.*, 2012; Carter and Huettel, 2013). This work hints at a potentially fine-grained organizational structure within socio-emotional brain regions but has yet to be characterized to the same degree of functional and spatial granularity as primary sensory systems.

A different perspective proposes that socio-emotional representations might be more diffusely distributed because the phenomenological experiences themselves (e.g. feeling an emotion and inferring an intention) are by their very nature more abstract, consisting of the integration of numerous processes such as perception, memory, prediction, and interoception (Chang *et al.*, 2015; Barrett, 2017). Numerous studies support this account by demonstrating how regions within the DMN are critical for mental-state inference but also prospection, episodic memory, navigation, narrative comprehension, mind-wandering and high-level comprehension (Buckner and Carroll, 2007; Mason *et al.*, 2007; Spreng *et al.*, 2009; Simony *et al.*, 2016; Tamir *et al.*, 2016; Golchert *et al.*, 2017). A wide

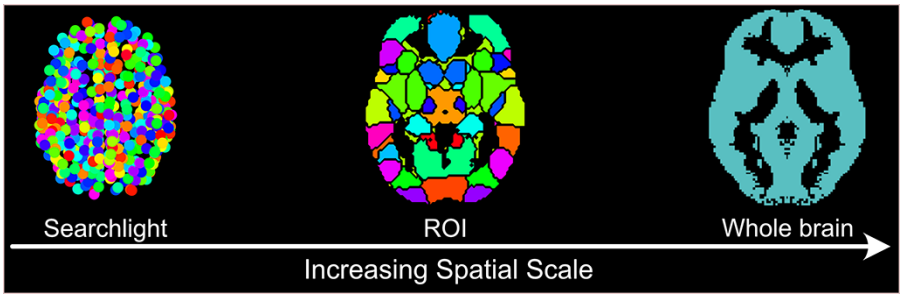

Fig. 1. Spatial scales of different analytic strategies. Most common analytic methods can be seen as lying on a spectrum of varying spatial scales. Searchlights (left) represent one endpoint of this spectrum as they are well suited for modeling information at small spatial scales such as fine-grained neural patterns in a local neighborhood around a voxel defined by a radius size. ROI (middle) approaches can be used to model larger spatial scales explicitly taking into account functional and anatomical divisions. Multiple ROIs can be combined together to model even larger spatial extents such as functional networks. Whole brain (right) approaches represent the other endpoint of this spectrum as they are well suited for modeling diffuse representations that extend beyond local neighborhoods, regions and networks.

range of brain regions, spanning multiple networks, including the default-mode, salience, and frontoparietal, appear to be involved in the representation of emotions (Kober *et al.*, 2008; Lindquist *et al.*, 2012; Chang *et al.*, 2015; Wager *et al.*, 2015; Kragel and LaBar, 2016). Further, even local neural patterns within specific areas such as the anterior TPJ demonstrate flexible responding as the same neural populations encode information about distances in space, time, as well as social ties (Parkinson *et al.*, 2014) or are broadly involved in establishing social context (Carter and Huettel, 2013). In this view, socio-emotional representations are entangled with other cognitive processes because they depend upon them. As such, neural representations appear to be correspondingly diffuse, recruiting distributed dynamic brain networks that can flexibly represent the highly abstract nature of social and emotional experiences.

What is the problem?

Given the heterogeneity of the spatial scale of different psychological processes, this immediately raises a question: how do the spatial scales of various analytic techniques interact with the representations they are measuring? For example, due to their inherently small spatial scale, searchlights are highly sensitive to identifying locally distributed patterns (Kriegeskorte *et al.*, 2008; Kriegeskorte and Diedrichsen, 2019), making them well suited to investigating representations that themselves are organized in a fine-grained manner (e.g. perceptual features). On the other hand, whole brain models, which jointly model functional responses across the entire brain, have been more successful than searchlights in identifying sensitive and specific predictive models of more abstract psychological processes such as pain (Wager *et al.*, 2013), negative affect (Chang *et al.*, 2015), guilt (Yu *et al.*, 2020), empathy (Krishnan *et al.*, 2016; López-Solà *et al.*, 2017) and identifying supramodal emotion categories (Kragel and LaBar, 2016). These examples raise the possibility that the efficient study of neural representations requires methods that coincide with the scale at which representations are organized. This problem is similar in nature to the choice of spatial smoothing kernel used in conventional fMRI analysis, whereby the optimal kernel size is dictated by the spatial extent of the hemodynamic response function as per the matched filter theorem (Friston, 2007). A large body of work has investigated how acquisition parameters like spatial resolution and pre-processing choices like smoothing affect the sensitivity of various analyses such as fMRI decoding (e.g. Gardumi *et al.*, 2016; Todd *et al.*, 2016; Yoo *et al.*, 2018). However, there have been far fewer studies investigating the optimal spatial scale (kernel size) of different multivariate analysis techniques (e.g. Stelzer *et al.*, 2014). This necessitates that researchers carefully consider the spatial scale of their analyses, rather than defaulting to particular pipelines. To aid in this process, we compare and contrast how common methodological conventions may interact with the spatial scale of neural representations.

Current conventions

Whether researchers are performing MVPA analyses to test information encoding or decoding, ISC analyses to measure neural synchrony, or connectivity analyses to examine networks, each technique implicitly or explicitly constrains the spatial scale at which statistics are computed. Should separate statistical models be built for different voxels, neighborhoods or regions of the brain (i.e. independent groups of voxels)? And if so, how should this be determined? Should predictions, weights and variability from these models be combined to make inferences? And if so, how? Because different answers to these questions ultimately test very different statistical models, spatial feature selection becomes a key decision that always adds additional assumptions or constraints to the hypotheses being tested and the conclusions being drawn. Fortunately, there are numerous options available to researchers that fall along a spectrum of fine grain to diffuse spatial scales[1] (Figure 1).

Searchlights. The popular searchlight approach (Kriegeskorte *et al.*, 2006; Kriegeskorte and Bandettini, 2007) lies at one end of the spectrum and can be viewed as the 'mass-multivariate' analogue to the 'mass-univariate' approach popular in conventional activation-based fMRI analyses (Friston *et al.*, 1995). Searchlight analyses only consider information contained in local, overlapping neighborhoods around each voxel defined by a radius, and ignore how information may be distributed across spatial scales outside of those local neighborhoods. In this way, searchlights may ignore relevant signals in more diffuse representations such as emotions and are consistently outperformed by whole brain or regional models in those situations (Kragel *et al.*, 2018; Chang *et al.*, 2015). When used for decoding analyses, searchlights are equivalent to feature subset-selection in the machine-learning literature, whereby subsets are determined

[1] Because all methods fundamentally operate on information contained in voxels, fine-grained in this context refers to information in voxel patterns comprised of small (often-contiguous) spatial neighborhoods, whereas diffuse refers to voxel patterns encompassing much larger contiguous or non-contiguous spatial extents.

by the coordinates of each voxel and the radius of each searchlight (Hastie *et al.*, 2009). Similar to their univariate counterpart, searchlights are agnostic to functional or anatomical subdivisions and typically require as many statistical computations as voxels in the brain. Though rarely directly contrasted, searchlights can be easily compared as they are most often computed with the same radius size and therefore different searchlights contain the same number of voxels.

Regions of interest. At a larger spatial scale, ROI approaches consist of groups of voxels determined by anatomical or functional divisions. There are broadly two types of ROI approaches: (i) contiguous and (ii) non-contiguous. Contiguous approaches consist of voxel groups that are spatially constrained to cover a continuous area of the brain, whereas non-contiguous approaches include both spatially contiguous but also spatially disjoint groups of voxels such as functional networks. Non-contiguous ROIs by their nature tend to encompass a larger spatial extent than contiguous regions. In both cases, spatial constraints are typically determined in two ways. One approach leverages functional responses, measured for example by using functional localizers from independent data (Saxe *et al.*, 2006) or by directly pruning voxels using techniques such as recursive feature elimination (De Martino *et al.*, 2008). The other approach relies on anatomical boundaries typically determined from brain atlases, rsfMRI connectivity network parcellations or meta-analyses (Yarkoni *et al.*, 2011; Chang *et al.*, 2013; De La Vega *et al.*, 2016; Eickhoff *et al.*, 2018; Shenton *et al.*, n.d.). The number of unique statistical computations estimated in the ROI approach is generally fewer than the searchlight approach and is determined based on the number of distinct regions selected. Unlike searchlights, ROI approaches can directly leverage known anatomical distinctions or functional response profiles as part of the spatial feature selection process. This flexibility enables them to capture a wide range of spatial scales, for example, modeling multiple distinct brain regions together or differentiating cortical sub-divisions across multiple models. More generally, ROI approaches are tests of focal hypotheses constrained to locations researchers often believe to be relevant a priori, such as social brain regions (Thornton and Mitchell, 2017). However, with this flexibility comes a trade-off in consistency across analyses. Comparisons across regions can become more complicated as ROIs typically don't contain the same number of voxels.

Whole brain models. Whole brain models reflect the largest spatial scale as they consider all voxels and their covariance during model estimation. In contrast to numerous small searchlights or ROIs, the whole brain approach can be viewed as a 'single searchlight/region' with a radius large enough to encompass all brain voxels. This approach can be used with unsupervised methods such as independent components analysis (Calhoun *et al.*, 2001; Beckmann *et al.*, 2005), or supervised methods such as decoding (Wager *et al.*, 2013; Chang *et al.*, 2015). Like searchlights, no anatomical information is explicitly used to determine the spatial scale of whole brain models. However, in decoding analyses, some more sophisticated algorithms can incorporate information about spatial smoothness or regional connectivity to find model estimates that better reflect the regional structure by forcing spatial constraints (Baldassarre *et al.*, 2012; Gramfort *et al.*, 2013; Grosenick *et al.*, 2013). Whole brain prediction analyses can provide a single model comprised of feature weights at each voxel that are simple to test in additional experimental contexts. Such generalization tests are highly valuable as they can provide valid reverse inference (Varoquaux and Poldrack, 2019) and also aid in identifying relative voxel importance (with caveats) (Haufe *et al.*, 2014; Kriegeskorte and Douglas, 2019). In addition, generalization tests can facilitate psychological construct validity, whereby model performance in different contexts can provide measurement information about the sensitivity and specificity of how a particular psychological construct is defined (e.g. different types of pain, memory and touch) (Kragel *et al.*, 2018). For this reason, these models have been particularly popular in translational and affective neuroscience, where whole brain decoders have been used as 'biomarkers' because they generalize well across populations and tasks even within a single subject (Wager *et al.*, 2013; Gabrieli *et al.*, 2015; Lindquist *et al.*, 2015; Krishnan *et al.*, 2016; Woo *et al.*, 2017; Kragel *et al.*, 2018).

Analytic considerations

There are several key factors that researchers might consider when choosing between different scales of spatial feature selection. We have organized these factors into three broad categories. The first concerns subjective choices such as the goals of a particular analysis and the types of inferences researchers hope to make. The second comprises practical considerations for reliable statistical estimation. The third concerns computational resource availability and the trade-offs between different approaches. A summary of these comparisons is listed in Table 1.

What is the goal?

A primary distinguishing factor between different analytic techniques is the type of inference researchers want to make. Broadly construed, modeling falls into two 'cultures,' (Breiman, 2001; Yarkoni and Westfall, 2016): inference emphasizes model interpretability and is evaluated using null-hypothesis-significance testing in a single context (e.g. a single dataset or task), while prediction emphasizes generalizability to new contexts and is evaluated based on out-of-sample model performance (Bzdok and Ioannidis, 2019). While this characterization cleanly distinguishes univariate magnitude-based analyses and multivariate predictive analyses, different multivariate analyses often conflate both goals in confusing ways (Hebart and Baker, 2018). For example, searchlight analysis was primarily conceived of as an information mapping technique and, when combined with cross-validated decoding, can approximate out-of-sample performance to make inferences about 'where information is represented' (Kriegeskorte *et al.*, 2006; Kriegeskorte and Bandettini, 2007). Decoding in the context of whole brain models has focused primarily on predictive performance and generalization to a variety of contexts such as developing brain–computer interfaces (Woo *et al.*, 2017; Hebart and Baker, 2018).

Reflecting these differences, results from searchlight analyses are typically reported as accuracy maps and inference is performed by comparing accuracy at each searchlight to empirical or permuted chance (Haynes, 2015) (Table 1 Conventional Inferences). However, the feasible conclusions that can be drawn from this approach only indicate whether at least one voxel in a local neighborhood is related to the outcome being predicted, not necessarily that every voxel in that neighborhood is reliably representing psychological information (Viswanathan *et al.*, 2012; Etzel *et al.*, 2013).[2] Feature weights within a searchlight

[2] This same criticism does not necessarily apply to searchlight-RSA or searchlight-ISC analyses.

Table 1. Comparison of different analytics strategies

	Searchlight	ROI	Whole brain
Spatial Scale	Fine-grained and fixed. Determined by searchlight radius which is typically the same for all searchlights.	Medium and flexible. Determined by how ROI was parcellated (e.g. functional responses, anatomy and network). Size reflects variable anatomy or functional response profiles.	Diffuse and fixed. Determined by sampling resolution of data (number of voxels).
Conventional inferences	Predictive performance of each searchlight (e.g. accuracy and correlation distance). Feature weights within searchlights typically not examined. Separate statistical models per individual and model performance aggregated at the group level.	Predictive performance for each ROI. Feature weights within ROIs highlight most informative voxels. Separate or common models across individuals.	Single predictive performance for model. Feature weights highlight most informative voxels. Separate or common models across individuals.
Estimation (decoding)	Independent models with overlapping features and some regularization (e.g. SVM). Anatomy is not part of estimation. $n > p;\ n \sim p;\ n < p$	Independent models with non-overlapping features and medium regularization (e.g. SVM and ridge). Anatomy can be used to define regions. $n < p;\ n \ll p$	Single model that uses global covariance across all features with high regularization and/or dimensionality reduction (e.g. LASSO-PCR[a]) Anatomy is not part of estimation but provide constraints.[b] $n \ll p$
Compute Cost (CPU-time)	High Large number of independent estimations required; more with permutation testing. Parallelization can reduce cost, but integrating results can be complicated	Medium Number of estimations depends on number of regions. Parallelization can reduce cost and integrating results is straightforward	Low Typically just one estimation and permutation regime performed. Parallelization is not trivial or not possible except for permutation testing or bootstrapping weights
Compute Cost (Memory)	Low/Medium memory Each searchlight has a small/medium memory footprint determined by radius and number of trials/conditions. Estimation rarely requires operating on all searchlight models simultaneously.	Medium memory Memory cost scales with the size of regions selected and number of trials/conditions/participants. Estimation rarely requires operating all ROI models simultaneously.	High memory Memory cost typically depends on total number of voxels (sampling resolution) and specific estimation routine (e.g. SVD). Estimation almost always requires operating on all voxels and observations simultaneously; exacerbated for between-subject models that require operating on many individual participants simultaneously
Compute cost (Storage + Ease of Sharing)	Low and simple if primarily working with performance only (e.g. accuracy maps, distance correlation) because each voxel is associated with a single value. High and complicated if intending to save feature weights because searchlights are overlapping. Data sharing typically consists of accuracy maps.	Low and simple because ROIs are most often non-overlapping and each voxel is associated with a single value (feature-weight or performance). Can represent performance and weight maps in a single standard format (array, NIfTI). Easy to apply to new datasets. Data sharing typically consists of accuracy maps, but feature weight maps are trivial to share as well.	Low and simple because just one model in which each voxel is associated with a single feature-weight. Can represent weightmaps in a single standard format. Data sharing typically consists of weight maps that are then applied to novel datasets

This table compares searchlight, ROI, and whole brain approaches in terms of their strengths and weaknesses along three categories: inferential goals, model estimation and computational resource demands. Legend: n: number of observations; p: number of features; ~: approximately equal; < or >: less or greater than; << or >>: much less or much greater than.

[a]The number of dimensions of predictive group models is typically limited by the number of participants in the dataset.

[b]See structured sparsity models (Baldassarre *et al.*, 2012; Gramfort *et al.*, 2013; Grosenick *et al.*, 2013).

are almost never examined nor used to make predictions on completely distinct datasets. This is due to the fact that searchlights are most often overlapping, leading each voxel to have a different feature weight depending upon the particular searchlight (local neighborhood) it belongs to. This makes it infeasible to perform traditional feature importance testing (e.g. bootstrapping/permutation testing) as there are numerous possible ways to integrate these different weights across

searchlights (e.g. see MIDAS (Varol *et al.*, 2018)). With increasing radius size, these issues make it nearly impossible to identify which voxels are most important for prediction, as accuracy scores are 'smeared' over spatial extents because searchlights are overlapping (Viswanathan *et al.*, 2012).[3] Searchlight analyses are also often computed on individual brains and performance metrics (e.g. accuracy) are aggregated at the group level to draw inferences (Stelzer *et al.*, 2014). This also means that the particular geometry of a representation (i.e. the spatial layout of feature weights within a local neighborhood) is likely to differ across individuals, greatly complicating what types of valid group inferences are possible. Unlike univariate activation analyses, rejecting the null-hypothesis of conventional parametric tests on accuracies (e.g. one-sample *t*-test) only suggests that some individuals demonstrate an effect not that the effect is typical in the population (Nichols *et al.*, 2005; Stelzer *et al.*, 2013; Allefeld *et al.*, 2016).

In contrast, whole brain analyses are often concerned with generalization to completely new datasets, which can be comprised of different individuals (Woo *et al.*, 2017). While predictive performance is essential in translational applications, the resulting feature weights at each voxel also provide some useful information as to the spatial layout of the representations e.g. 'neural signature' (Wager *et al.*, 2013). Feature importance (Table 1 Conventional Inferences) can be assessed by thresholding via resampling methods such as bootstrapping or permutation (Stelzer *et al.*, 2014; Chang *et al.*, 2015); however, the resulting thresholded maps must be interpreted with caution. Unlike univariate activation maps, reliable weight maps do not indicate that a voxel explicitly represents psychological information but that in concert with other voxels it can effectively predict an outcome (Haufe *et al.*, 2014). In other words, some voxels may indeed represent outcome-relevant information, but some may serve to denoise other voxels which share correlated noise (Kriegeskorte and Douglas, 2019).

ROI analyses are flexible enough to inherit the strengths and weaknesses of both searchlight and whole brain analyses depending on the details of an implementation. Separate models can be estimated for disjoint ROIs and aggregated to make predictions, similar to kernel learning in machine learning, where different kernels are used for different regions (Filippone *et al.*, 2012; Schrouff *et al.*, 2013). A single model encompassing multiple disjoint voxels can also be estimated to draw inferences about a network of regions or voxels that share similar functional response profiles, e.g. 'social-brain mask' (Thornton and Mitchell, 2017). Because ROI methods don't typically involve overlapping features like searchlights, accuracy maps do not suffer from spatial 'smearing,' and feature weights can be examined for relative voxel importance similar to whole brain models (Chang *et al.*, 2018). At the same time, performance metrics and generalization tests on separate datasets and contexts are feasible and straightforward, permitting inferences about the sensitivity and specificity of representations within single brain regions (Chang *et al.*, 2015; Krishnan *et al.*, 2016).

Thus, each end of the spectrum varies in its inferential goals. Searchlight decoding permits spatial inference based on isolated local neighborhoods tested in similar contexts while ignoring how that information is represented (ignoring feature weights) unless explicitly modeled with approaches like RSA. Because they are typically estimated separately across individuals, they do not identify shared or common representations, but rather whether any kind of task-relevant representations exist in the brain (Allefeld *et al.*, 2016). Whole brain models permit strong inferences about generalization, based on model performance, and diffuse inferences about the spatial location of representations based on feature weights. Most often in practice, whole brain models aim to learn a common representation that generalizes across individuals. Regional approaches land in-between these endpoints based on their particular implementation. All methods, however, can extend beyond simple decoding analyses to facilitate stronger inferences. Searchlight analyses can use cross-validated RSA or pattern-component modeling (PCM) with model comparison to test hypotheses about what stimulus features geometrically organize information within a neighborhood (Nastase *et al.*, 2017; Kriegeskorte and Diedrichsen, 2019). Different whole brain feature weight maps can be compared within the same context to determine representational specificity, share information and facilitate valid reverse inferences (Krishnan *et al.*, 2016; Varoquaux and Poldrack, 2019).

Model estimation

The most common multivariate[4] fMRI analyses are typically decoding models and RSA (Kriegeskorte *et al.*, 2006; Norman *et al.*, 2006). In decoding approaches, voxels are considered features, while time-points, trials, individuals or sessions serve as observations. Building a statistical model (e.g. a classifier, regression) requires estimating weights for features that can be combined to predict an outcome that generalizes over observations, such as properties of a task/stimulus (e.g. condition or category labels) or responses from individuals (e.g. behavior and emotional ratings).[5] Voxel-selection procedures are the primary determinant of inputs that a statistical model uses to predict an outcome. This means that successful statistical estimation is heavily affected by the ratio between the number of features (p) and number of observations (n) (Hastie *et al.*, 2009). When $n \geq p$, (more or equivalent observations than features) a model can be consistently[6] estimated without further constraints. However, situations where $n < p$ (fewer observations than features) yield a statistically underdetermined problem such that many unique combinations of features weights can yield the same predicted outcome. This issue is further exacerbated by the degree of independence between features. For example, spatial smoothing is a preprocessing step that can help boost signal-to-noise ratios but decreases spatial independence. Together these issues can lead to models that exhibit overfitting,[7] whereby feature

[3] Smearing, however, can happen in principal with searchlight-RSA and searchlight-ISC analyses.

[4] While encoding models can also be viewed as a kind of multivariate model, they are most often multivariate in stimulus feature space but univariate in brain space. In other words, high-dimensional models are primarily used to fit and predict a single voxel's responses rather than a local or global spatial pattern (Nishimoto *et al.*, 2011; Huth *et al.*, 2016).

[5] This delineation doesn't perfectly capture RSA analyses as models are typically distance matrices derived from stimulus or task features and outcomes are neural distance matrices based on responses to those stimulus or task features.

[6] Consistently here refers to a single solution (weights) that maps between features and outcomes conditional on some error/loss function (e.g. sum-of-squared errors/L_2 norm in linear regression).

[7] Underfitting is also possible, whereby feature weight fails to capture the true signal in a data, but occurs less often in fMRI analyses. This is because in most datasets, irrespective of spatial scale, researchers rarely have more observations being predicted (e.g. trials, conditions and individuals) than features used to make predictions (e.g. voxels), i.e. $n < p$ or $n << p$.

weights reflect both true signal but also idiosyncratic noise and generalize poorly to new data. To combat these issues, most estimation routines rely on some form of regularization, whereby constraints or penalties are used to limit the range of possible estimated feature weights. Common approaches include minimizing the squared (ridge and L_2 penalty) or absolute magnitude (lasso and L_1 penalty) (Hastie *et al.*, 2009) of feature weights. In many cases, these penalization techniques are similar to imposing differently shaped priors in Bayesian models (James *et al.*, 2013; Nunez-Elizalde *et al.*, 2019).

Since searchlights focus on local neighborhoods, their radius size, along with the details of an experimental task (e.g. number of conditions, trials, trials per condition, etc.), determine the ratio between features (voxels) and observations (trials, conditions) (Table 1 Estimation). Small neighborhoods comprise few features (e.g. ~28 voxels in a 6 mm radius searchlight collected at 2 mm voxel resolution volume) meaning approximately equivalent number of observations and features ($n \sim p$) or a smaller imbalance of more features than observations ($n < p$; e.g. 100 voxels to 80 observations (Nastase *et al.*, 2017)). This may facilitate algorithms that require less regularization as evidenced by the popular use of linear models (e.g. linear discriminant analysis and support vector machine (SVM)) that exhibit good performance using default or variance-scaled hyperparameters rather than optimal hyperparameters tuned via cross-validation (e.g. Norman *et al.*, 2006; Hanke *et al.*, 2009). However, radii are often arbitrarily chosen based on sizes in previous studies and can have large effects on this ratio and thus may require different statistical models and regularization strategies, e.g. cross-validated MANOVA (multivariate analysis of variance) (Allefeld and Haynes, 2014). In addition, multiple comparisons corrections are needed to adjust for the large number of estimated models (Etzel *et al.*, 2013).

Since whole brain models include all voxels and are often used to identify representations that generalize across individuals, the features greatly outnumber observations ($n \ll p$; e.g. 350k voxels to 182 individuals (Chang *et al.*, 2015)) often requiring stronger regularization (Kragel *et al.*, 2018) (Table 1 Estimation). For this reason, several studies use rigorous nested cross-validation along with independent hold-out sets to first tune regularization hyperparameters, then evaluate cross-validated predicted performance and, finally, test generalization performance on completely new individuals (Wager *et al.*, 2013; Chang *et al.*, 2015; López-Solà *et al.*, 2017; Kragel *et al.*, 2018). Another popular regularization approach is the LASSO-PCR (LASSO principal components regression), in which dimensionality reduction over all brain voxels is first performed using principal components analysis (PCA)[8] followed by a sparse regression model (LASSO) to estimate weights on each principal component that are later inverted back into voxel space (Wager *et al.*, 2011). This approach jointly considers large groups of voxels with similar responses as single features used for prediction and produces sparse weight maps where only a few such voxel groups contribute strongly to prediction.

As noted in the previous section, the flexibility of ROI approaches, and the particular implementation chosen, will largely dictate the properties of an estimation regime. However, using a particular implementation such as non-overlapping, but contiguous ROIs, it may be possible to balance the strengths and weakness of both searchlight and whole-brain approaches (e.g. smaller neighborhoods, necessitates less regularization, but with estimable feature importance maps that can be used for generalization testing).

[8] In practice, the maximum number of retainable components is limited to the number of observations, typically individuals, in the dataset.

Computational resources

The differences in inference and estimation routines between different techniques also impose different demands on computational resources (Table 1 Compute Cost). Broadly speaking, resources can be divided into three categories: (i) central processing units (CPU) time—the number of independent estimations required, the time required for each and the serial or parallelizability of the estimations; (ii) random access memory (RAM)—the 'temporary' working memory required to perform each estimation, typically determined by how and whether a particular algorithm needs to operate on all features and observations together, or can operate on them in a piecewise (batch) fashion and (iii) Storage—the hard disk space required to store the outputs of an estimation routine and the format of this storage that can determine ease of sharing models.

At the small spatial scale end of the spectrum, searchlights often demand high CPU costs, low to medium memory and, most often, low storage. This is because searchlight analyses require estimating as many models as there are voxels in a dataset. However, estimations can proceed in parallel and because features come from local neighborhoods with a small number of voxels, memory demands are typically low as well. Memory demands increase monotonically with increasing features and/or observations, i.e. larger radius or more task trials/conditions. If inferences are primarily made using accuracy maps, then storage is simple as a single value can be stored at each voxel location which can be easily shared. However, if researchers intend to store feature weights for each searchlight, storage becomes more complex due to large demands on disk-space and complicated indexing assigning feature weight vectors to each voxel location.

At the large spatial scale end of the spectrum, whole brain models often demand low CPU costs, high memory and low and simple storage. Because all voxels are used for estimation, only a single model needs be computed. However, because algorithms require operating on all voxels and observations simultaneously, they must hold and manipulate very large matrices (e.g. whole brain covariance matrix of 3k observations (100 participants with 30 trials each) by 200k voxels) in memory. Storage costs are low and straightforward as a model consists of a single scalar performance score and each voxel is only associated with a single feature weight, making whole brain models very easy to share and test on new datasets.

As with other analytic considerations, ROI approaches typically fall between searchlight and whole brain analyses with relatively medium CPU costs and memory but simple and low storage requirements. CPU costs can be minimized using parallelization like searchlight analyses. Memory demands scale with the size of each ROI as larger regions (e.g. non-contiguous DMN mask) require manipulating more features and observations together. Since ROI models are typically non-overlapping, they share storage demands similar to whole brain models as feature weights from different regions can be stored together in a single file along with binary masks to later extract the weights and apply them to new data. Accuracy maps derived from ROI models are similar to those estimated from searchlights, as only a single value needs to be associated with each voxel location.

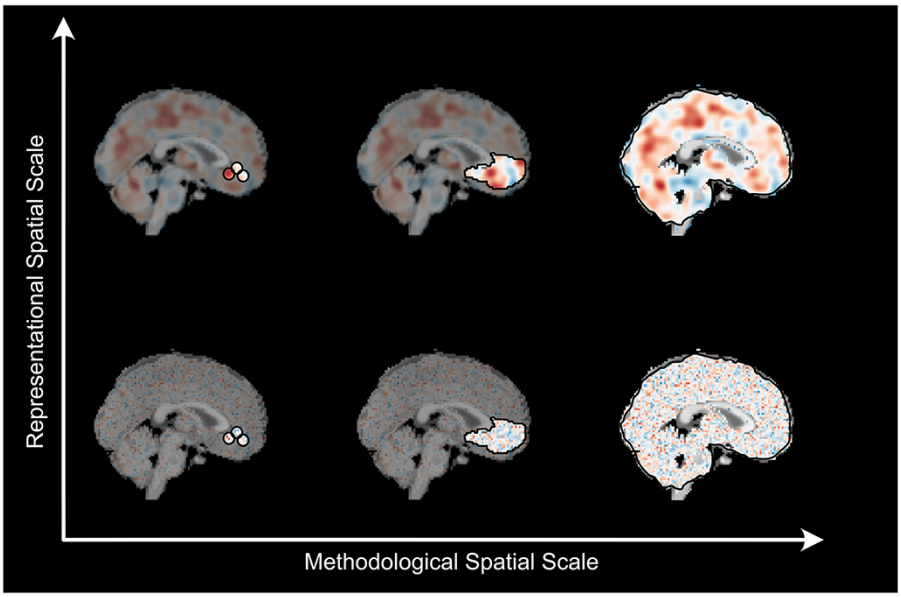

Fig. 2. Interactions between methodological and representational spatial scales. Depending on the type of phenomenon under inquiry some analytic techniques may be more or less optimal. Increasing spatial scale of analysis techniques are depicted on the *x*-axis with searchlights at the small (left) end and whole brain approaches on the large (right) end; these mirror the spectrum Figure 1. The *y*-axis depicts hypothetical endpoints of representational scales with fine-grained local patterns in the bottom row (e.g. perceptual processes) and more diffuse patterns in the top row (e.g. social and emotional processes). Fine scale methods like searchlights may fail to capture diffuse representations as local neighborhoods provide a distorted view of a diffuse representation (top-row; left). These same methods may be optimal for finer neural representation in which all relevant information is reflected in a local neighborhood (bottom-row; left). On the other hand, large-scale methods such as whole brain approaches may be unable to reliably identify informative voxels when representations are organized in local neighborhoods (bottom-row; right) and may be better suited to identifying diffuse representations with large spatial extents (top-row; right). ROI approaches (top/bottom-row; middle) offer a flexible compromise, inheriting both the strengths and weaknesses of searchlight and whole brain approaches depending on the particular ROI method employed. At the same time, the smallest spatial scale measurable by fMRI is likely limited by the BOLD point-spread-function at a particular magnetic field strength, e.g. 3–5 mm at 3T (Parkes *et al.*, 2005).

For all spatial scales, cross-validation or non-parametric inference using resampling methods such as bootstrapping and permutation testing, will dramatically increase CPU costs and can potentially increase memory or storage requirements. This is because resampling methods require re-estimating a completely new model for each cross-validation fold and bootstrapped/permuted iteration. In the case of cross-validation or permutation testing, only the performance of each iteration needs to be retained, keeping storage costs low. However, bootstrapping distributions of feature weights requires retaining each iteration in order to define upper and lower uncertainty bounds (e.g. confidence intervals), thereby increasing costs depending upon researchers' goals. For example, keeping feature weights in memory can reduce storage costs at the expense of increased RAM and decreased analytic flexibility down the line (e.g. loading and estimating a distribution). Saving feature weights to disk, on the other hand, increases storage costs by a factor of bootstrap iterations (each iteration produces a new set of feature weights of the same shape and size as the original model) but provides more analytic flexibility later on.

Conclusions and recommendations

In this article, we have highlighted literature demonstrating how neural representations can exist at multiple spatial scales across the brain. Representations related to perceptual processes are often localized to small neighborhoods with highly specific response properties and hierarchical organization. Representations related to more abstract modes of cognition like social and emotional processing have been observed at fine spatial scales but more often consist of diffuse spatial representations spanning multiple regions and networks. This representational heterogeneity can interact with the spatial scale of particular analytic techniques, ranging from fine-grain pattern sensitivity in local neighborhoods (searchlights), focal tests of specific regions and networks (ROI), to whole brain neural markers that generalize across experimental contexts.

While it may be tempting to iterate over many possible analyses and attempt to 'optimize' for the 'best' spatial scale, we caution researchers against framing the issue in this way given the lack of research specifically addressing this issue. For example, techniques like model comparison between searchlights and whole brain models are not trivial or even feasible to perform in most cases. Whole brain approaches estimate a single model, but other approaches estimate *N* models, where *N* is the number of ROIs or searchlights. Which of the *N* models should be used to compare to the whole brain model? Or should *N* models be combined into an ensemble? And if so how? One possible approach illustrated by Chang *et al.* (2015) (Supplementary Figure S4 Panel B) and Kragel *et al.* (2018) (Figure 2) compares the performance of whole brain models to the entire distribution of searchlight models but is unable to directly compare how different model weights capture the representation of emotions. Adding decision points to analysis pipelines without cross-validation multiplies analytic flexibility and will likely increase experiment level false-positive rates or facilitate '*p*-hacking' (Carp, 2012). Instead, we recommend researchers more carefully select their analytic approach using a combination of empirical goals, estimation techniques and computational

resources to determine what makes the most sense for the investigation at hand. At the same time, we believe the field may benefit from investigations directly examining the spatial scale of psychological phenomena thereby bringing greater clarity and more progress to this understudied issue.

Acknowledgements

The authors wish to thank Emma Templeton, Jin Cheong and Amanda Brandt for providing helpful feedback on earlier drafts of this manuscript.

Funding

This work was supported by the National Institute of Mental Health R01MH116026 and the National Science Foundation CAREER 1848370.

Conflict of interest

The authors report no conflicts of interest.

Open science statement

A preprint of this manuscript has been submitted to bioRxiv.

References

Adolphs, R. (2009). The social brain: neural basis of social knowledge. *Annual Review of Psychology*, **60**, 693–716.

Allefeld, C., Görgen, K., Haynes, J.-D. (2016). Valid population inference for information-based imaging: from the second-level t-test to prevalence inference. *NeuroImage*, **141**, 378–92.

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

Amodio, D.M., Frith, C.D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, **7**, 268–77.

van Baar, J.M., Chang, L.J., Sanfey, A.G. (2019). The computational and neural substrates of moral strategies in social decision-making. *Nature Communications*, **10**, 1483.

Baldassarre, L., Mourao-Miranda, J., Pontil, M. (2012). Structured sparsity models for brain decoding from fMRI data. In: *2012 Second International Workshop on Pattern Recognition in NeuroImaging*. 5–8.

Barrett, L.F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. *Social Cognitive and Affective Neuroscience*, **12**, 1833, London, UK: IEEE. https://doi.org/10.1109/PRNI.2012.31.

Beckmann, C.F., DeLuca, M., Devlin, J.T., et al. (2005). Investigations into resting-state connectivity using independent component analysis. *Philosophical Transactions of the Royal Society of London Series B, Biological Sciences*, **360**, 1001–13.

Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Statistical Science a Review Journal of the Institute of Mathematical Statistics*, **16**, 199–231.

Buckner, R.L., Carroll, D.C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, **11**, 49–57.

Bzdok, D., Ioannidis, J.P.A. (2019). Exploration, inference, and prediction in neuroscience and biomedicine. *Trends in Neurosciences*, **42**, 251–62.

Calhoun, V.D., Adali, T., Pearlson, G.D., et al. (2001). A method for making group inferences from functional MRI data using independent component analysis. *Human Brain Mapping*, **14**, 140–51.

Carp, J. (2012). On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. *Frontiers in Neuroscience*, **6**, 149.

Carter, R.M., Huettel, S.A. (2013). A nexus model of the temporal-parietal junction. *Trends in Cognitive Sciences*, **17**, 328–36.

Chang, L.J., Gianaros, P.J., Manuck, S.B., et al. (2015). A sensitive and specific neural signature for picture-induced negative affect. *PLoS Biology*, **13**, e1002180.

Chang, L.J., Jolly, E., Cheong, J.H., et al. (2018). Endogenous variation in ventromedial prefrontal cortex state dynamics during naturalistic viewing reflects affective experience. *bioRxiv*. https://doi.org/10.1101/487892.

Chang, L.J., Yarkoni, T., Khaw, M.W., et al. (2013). Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. *Cerebral Cortex*, **23**, 739–49.

Chang, L., Tsao, D.Y. (2017). The code for facial identity in the primate brain. *Cell*, **169**, 1013–28.e14.

Chen, P.-H.A., Jolly, E., Cheong, J.H., and Chang, L. J. (2020). Intersubject representational similarity analysis reveals individual variations in affective experience when watching erotic movies. *NeuroImage*, **216**, 116851.

Cichy, R.M., Khosla, A., Pantazis, D., et al. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, **6**, 27755.

De La Vega, A., Chang, L.J., Banich, M.T., and Yarkoni, MT. (2016). Large-scale meta-analysis of human medial frontal cortex reveals tripartite functional organization. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, **36**(24), 6553–6562.

De Martino, F., Valente, G., Staeren, N., et al. (2008). Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage*, **43**, 44–58.

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

Eickhoff, S.B., Yeo, B.T.T., Genon, S. (2018). Imaging-based parcellations of the human brain. *Nature Reviews Neuroscience*, **19**, 672–86.

Etzel, J.A., Zacks, J.M., Braver, T.S. (2013). Searchlight analysis: promise, pitfalls, and potential. *NeuroImage*, **78**, 261–9.

Felleman, D.J., Van Essen, D.C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, **1**, 1–47.

Filippone, M., Marquand, A.F., Blain, C.R.V., et al. (2012). Probabilistic prediction of neurological disorders with a statistical assessment of neuroimaging data modalities. *The Annals of Applied Statistics*, **6**, 1883–905.

Finn, E.S., Glerean, E., Khojandi, A.Y., et al. (2020). Idiosynchrony: from shared responses to individual differences during naturalistic neuroimaging. *NeuroImage*, **215**, 116828.

Friston, K. (2007). Statistical parametric mapping. *Statistical Parametric Mapping*, 10–31. https://www.sciencedirect.com/science/article/pii/B9780123725608500024.

Friston, K.J., Holmes, A.P., Poline, J.B., et al. (1995). Analysis of fMRI time-series revisited. *NeuroImage*, **2**, 45–53.

Gabrieli, J.D.E., Ghosh, S.S., Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. *Neuron*, **85**, 11–26.

Gardumi, A., Ivanov, D., Hausfeld, L., et al. (2016). The effect of spatial resolution on decoding accuracy in fMRI multivariate pattern analysis. *NeuroImage*, **132**, 32–42.

Glasser, M.F., Coalson, T.S., Robinson, E.C., et al. (2016). A multimodal parcellation of human cerebral cortex. *Nature*, **536**, 171–8.

Golchert, J., Smallwood, J., Jefferies, E., et al. (2017). Individual variation in intentionality in the mind-wandering state is reflected in the integration of the default-mode, fronto-parietal, and limbic networks. *NeuroImage*, **146**, 226–35.

Gramfort, A., Thirion, B., Varoquaux, G. (2013). Identifying predictive regions from fMRI with TV-L1 prior. In: *2013 International Workshop on Pattern Recognition in Neuroimaging*, Philadelphia, PA, USA. 17–20.

Grill-Spector, K., Malach, R. (2004). The human visual cortex. *Annual Review of Neuroscience*, **27**, 649–77.

Grosenick, L., Klingenberg, B., Katovich, K., et al. (2013). Interpretable whole-brain prediction analysis with GraphNet. *NeuroImage*, **72**, 304–21.

Hanke, M., Halchenko, Y.O., Sederberg, P.B., et al. (2009). PyMVPA: a python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, **7**, 37–53.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd edn, New York: Springer Science & Business Media.

Haufe, S., Meinecke, F., Görgen, K., et al. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, **87**, 96–110.

Haxby, J.V., Connolly, A.C., Guntupalli, J.S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, **37**, 435–56.

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

Hebart, M.N., Baker, C.I. (2018). Deconstructing multivariate decoding for the study of brain function. *NeuroImage*, **180**, 4–18.

Honey, C.J., Sporns, O., Cammoun, L., et al. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 2035–40.

Hubel, D.H., Wiesel, T.N. (2004). *Brain and Visual Perception: The Story of a 25-Year Collaboration*. Oxford: Oxford University Press.

Huth, A.G., de Heer, W.A., Griffiths, T.L., et al. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, **532**, 453–8.

James, G., Witten, D., Hastie, T., et al. (2013). *An Introduction to Statistical Learning: With Applications in R*. New York, NY: Springer.

Kanwisher, N., McDermott, J., Chun, M.M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience the Official Journal of the Society for Neuroscience*, **17**, 4302–11.

Kober, H., Barrett, L.F., Joseph, J., et al. (2008). Functional grouping and cortical-subcortical interactions in emotion: a meta-analysis of neuroimaging studies. *NeuroImage*, **42**, 998–1031.

Koster-Hale, J., Richardson, H., Velez, N., et al. (2017). Mentalizing regions represent distributed, continuous, and abstract dimensions of others' beliefs. *NeuroImage*, **161**, 9–18.

Kragel, P.A., Koban, L., Barrett, L.F., et al. (2018). Representation, pattern information, and brain signatures: from neurons to neuroimaging. *Neuron*, **99**, 257–73.

Kragel, P.A., LaBar, K.S. (2016). Decoding the nature of emotion in the brain. *Trends in Cognitive Sciences*, **20**, 444–55.

Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, **1**, 417–46.

Kriegeskorte, N., Bandettini, P. (2007). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, **38**, 649–62.

Kriegeskorte, N., Diedrichsen, J. (2019). Peeling the onion of brain representations. *Annual Review of Neuroscience*, **42**, 407–32.

Kriegeskorte, N., Douglas, P.K. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, **55**, 167–79.

Kriegeskorte, N., Goebel, R., Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 3863–8.

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

Krishnan, A., Woo, C.W., Chang, L.J., et al. (2016). Somatic and vicarious pain are represented by dissociable multivariate brain patterns. *eLife*, **5**, e15166.

Lieberman, M.D. (2007). Social cognitive neuroscience: a review of core processes. *Annual Review of Psychology*, **58**, 259–89.

Lindquist, K.A., Wager, T.D., Kober, H., et al. (2012). The brain basis of emotion: a meta-analytic review. *The Behavioral and Brain Sciences*, **35**, 121–43.

Lindquist, M.A., Krishnan, A., López-Solà, M., et al. (2015). Group-regularized individual prediction: theory and application to pain. *NeuroImage*, **145**, 274–287.

López-Solà, M., Koban, L., Krishnan, A., et al. (2017). When pain really matters: a vicarious-pain brain marker tracks empathy for pain in the romantic partner. *Neuropsychologia*, **145**, 1–8.

Margulies, D.S., Ghosh, S.S., Goulas, A., et al. (2016). Situating the default-mode network along a principal gradient of macroscale cortical organization. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 12574–9.

Mars, R.B., Sallet, J., Schüffelgen, U., et al. (2012). Connectivity-based subdivisions of the human right 'temporoparietal junction area': evidence for different areas participating in different cortical networks. *Cerebral Cortex*, **22**, 1894–903.

Mason, M.F., Norton, M.I., Van Horn, J.D., et al. (2007). Wandering minds: the default network and stimulus-independent thought. *Science*, **315**, 393–5.

Carter, R.M., Bowling, D.L., Reeck, C., and Huettel, S.A. (2012). A distinct role of the temporal-parietal junction in predicting socially guided decisions. *Science*, **337**(6090), 109–111.

Mesulam, M.M. (1998). From sensation to cognition. *Brain A Journal of Neurology*, **121**(Pt 6), 1013–52.

Mitchell, J.P. (2008). Activity in right temporo-parietal junction is not selective for theory-of-mind. *Cerebral Cortex*, **18**, 262–71.

Mitchell, J.P., Macrae, C.N., Banaji, M.R. (2006). Dissociable medial prefrontal contributions to judgments of similar and dissimilar others. *Neuron*, **50**, 655–63.

Naselaris, T., Kay, K.N., Nishimoto, S., et al. (2011). Encoding and decoding in fMRI. *NeuroImage*, **56**, 400–10.

Nastase, S.A., Connolly, A.C., Oosterhof, N.N., et al. (2017). Attention selectively reshapes the geometry of distributed semantic representation. *Cerebral Cortex*, **27**, 4277–91.

Nastase, S.A., Gazzola, V., Hasson, U., et al. (2019). Measuring shared responses across subjects using intersubject correlation. *Social Cognitive and Affective Neuroscience*, **14**, 667–685.

Nichols, T., Brett, M., Andersson, J., et al. (2005). Valid conjunction inference with the minimum statistic. *NeuroImage*, **25**, 653–60.

Nishimoto, S., Vu, A.T., Naselaris, T., et al. (2011). Reconstructing visual experiences from brain activity evoked by natural movies. *Current Biology*, **21**, 1641–6.

Norman, K.A., Polyn, S.M., Detre, G.J., et al. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, **10**, 424–30.

Nunez-Elizalde, A.O., Huth, A.G., Gallant, J.L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage*, **197**, 482–92.

Paquola, C., De Wael, R.V., Wagstyl, K., et al. (2019). Microstructural and functional gradients are increasingly dissociated in transmodal cortices. *PLoS Biology*, **17**, e3000284.

Parkes, L.M., Schwarzbach, J.V., Bouts, A.A., et al. (2005). Quantifying the spatial resolution of the gradient echo and spin echo BOLD response at 3 Tesla. *Magnetic Resonance in Medicine Official Journal of the Society of Magnetic Resonance in Medicine Society of Magnetic Resonance in Medicine*, **54**, 1465–72.

Parkinson, C., Liu, S., Wheatley, T. (2014). A common cortical metric for spatial, temporal, and social distance. *The Journal of Neuroscience the Official Journal of the Society for Neuroscience*, **34**, 1979–87.

Peelen, M.V., Atkinson, A.P., Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. *The Journal of Neuroscience the Official Journal of the Society for Neuroscience*, **30**, 10127–34.

Power, J.D., Cohen, A.L., Nelson, S.M., et al. (2011). Functional network organization of the human brain. *Neuron*, **72**, 665–78.

Saxe, R., Brett, M., Kanwisher, N. (2006). Divide and conquer: a defense of functional localizers. *NeuroImage*, **30**, 1088–96; discussion 1097–9.

Saxe, R., Kanwisher, N. (2003). People thinking about thinking people: the role of the temporo-parietal junction in 'theory of mind'. *NeuroImage*, **19**, 1835–42.

Schaefer, A., Kong, R., Gordon, E.M., et al. (2016). Cerebral cortex parcellation by fusion of local and global functional connectivity feature. In: *The International Society for Magnetic Resonance in Medicine Annual Meeting*, Singapore.

Schrouff, J., Rosa, M.J., Rondina, J.M., et al. (2013). PRoNTo: pattern recognition for neuroimaging toolbox. *Neuroinformatics*, **11**, 319–37.

Shenton, M.E., Kikinis, R., McCarley, W., et al. (n.d.) Harvard brain atlas: a teaching and visualization tool. In: Proceedings 1995 Biomedical Visualization, Atlanta, GA, USA.

Simony, E., Honey, C.J., Chen, J., et al. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension. *Nature Communications*, **7**, 12141.

Spreng, R.N., Mar, R.A., Kim, A.S.N. (2009). The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis. *Journal of Cognitive Neuroscience*, **21**, 489–510.

Stelzer, J., Buschmann, T., Lohmann, G., et al. (2014). Prioritizing spatial accuracy in high-resolution fMRI data using multivariate feature weight mapping. *Frontiers in Neuroscience*, **8**, 66.

Stelzer, J., Chen, Y., Turner, R. (2013). Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. *NeuroImage*, **65**, 69–82.

Sul, S., Tobler, P.N., Hein, G., et al. (2015). Spatial gradient in value representation along the medial prefrontal cortex reflects individual differences in prosociality. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 7851–6.

Tamir, D.I., Thornton, M.A., Contreras, J.M., et al. (2016). Neural evidence that three dimensions organize mental state representation: rationality, social impact, and valence. *Proceedings of the National Academy of Sciences of the United States of America*, **113**, 194–9.

Yeo, B.T., Krienen, F.M., Sepulcre, J., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology* **106**, 1125–1165.

Thornton, M.A., Mitchell, J.P. (2017). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex*, **28**, 3505–520.

Todd, N., Moeller, S., Auerbach, E.J., et al. (2016). Evaluation of 2D multiband EPI imaging for high-resolution, whole-brain, task-based fMRI studies at 3T: sensitivity and slice leakage artifacts. *NeuroImage*, **124**, 32–42.

Varol, E., Sotiras, A., Davatzikos, C. (2018). MIDAS: regionally linear multivariate discriminative statistical mapping. *NeuroImage*, **174**, 111–26.

Varoquaux, G., Poldrack, R.A. (2019). Predictive models avoid excessive reductionism in cognitive neuroimaging. *Current Opinion in Neurobiology*, **55**, 1–6.

Vázquez-Rodríguez, B., Suárez, L.E., Markello, R.D., et al. (2019). Gradients of structure–function tethering across neocortex. *Proceedings of the National Academy of Sciences of the United States of America*, **116**, 21219–27.

Viswanathan, S., Cieslak, M., Grafton, S.T. (2012). On the geometric structure of fMRI searchlight-based information maps. *arXiv [Q-bio.nc]*.

Wager, T.D., Atlas, L.Y., Leotti, L.A., et al. (2011). Predicting individual differences in placebo analgesia: contributions of brain activity during anticipation and pain experience. *The Journal of Neuroscience the Official Journal of the Society for Neuroscience*, **31**, 439–52.

Wager, T.D., Atlas, L.Y., Lindquist, M.A., et al. (2013). An fMRI-based neurologic signature of physical pain. *The New England Journal of Medicine*, **368**, 1388–97.

Wager, T.D., Kang, J., Johnson, T.D., et al. (2015). A Bayesian model of category-specific emotional brain responses. *PLoS Computational Biology*, **11**, e1004066.

Wagner, D.D., Haxby, J.V., Heatherton, T.F. (2012). The representation of self and person knowledge in the medial prefrontal cortex. *Wiley Interdisciplinary Reviews Cognitive Science*, **3**, 451–70.

Woo, C.-W., Chang, L.J., Lindquist, M.A., et al. (2017). Building better biomarkers: brain models in translational neuroimaging. *Nature Neuroscience*, **20**, 365–77.

Yamins, D.L.K., DiCarlo, J.J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, **19**, 356–65.

Yamins, D.L.K., Hong, H., Cadieu, C.F., et al. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, **111**, 8619–24.

Yarkoni, T., Poldrack, R.A., Nichols, T.E., et al. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, **8**, 665–70.

Yarkoni, T., Westfall, J. (2016). Choosing prediction over explanation in psychology: lessons from machine learning. *Figshare*, **12**(6), 1100–122.

Yoo, P.E., John, S.E., Farquharson, S., et al. (2018). 7T-fMRI: faster temporal resolution yields optimal BOLD sensitivity for functional network imaging specifically at high spatial resolution. *NeuroImage*, **164**, 214–29.

Young, L., Dodell-Feder, D., Saxe, R. (2010). What gets the attention of the temporo-parietal junction? An fMRI investigation of attention and theory of mind. *Neuropsychologia*, **48**, 2658–64.

Yu, H., Koban, L., Chang, L.J., et al. (2020). A generalizable multivariate brain pattern for interpersonal guilt. *Cerebral Cortex*, **30**, 3358–572.