Behavioral/Cognitive

Finding Distributed Needles in Neural Haystacks

Christopher R. Cox[1] and Timothy T. Rogers[2]
[1]Department of Psychology, Louisiana State University, Baton Rouge, Louisiana 70803, and [2]Department of Psychology, University of Wisconsin, Madison, Wisconsin 53706

The human cortex encodes information in complex networks that can be anatomically dispersed and variable in their microstructure across individuals. Using simulations with neural network models, we show that contemporary statistical methods for functional brain imaging—including univariate contrast, searchlight multivariate pattern classification, and whole-brain decoding with L1 or L2 regularization—each have critical and complementary blind spots under these conditions. We then introduce the sparse-overlapping-sets (SOS) LASSO—a whole-brain multivariate approach that exploits structured sparsity to find network-distributed information—and show in simulation that it captures the advantages of other approaches while avoiding their limitations. When applied to fMRI data to find neural responses that discriminate visually presented faces from other visual stimuli, each method yields a different result, but existing approaches all support the canonical view that face perception engages localized areas in posterior occipital and temporal regions. In contrast, SOS LASSO uncovers a network spanning all four lobes of the brain. The result cannot reflect spurious selection of out-of-system areas because decoding accuracy remains exceedingly high even when canonical face and place systems are removed from the dataset. When used to discriminate visual scenes from other stimuli, the same approach reveals a localized signal consistent with other methods—illustrating that SOS LASSO can detect both widely distributed and localized representational structure. Thus, structured sparsity can provide an unbiased method for testing claims of functional localization. For faces and possibly other domains, such decoding may reveal representations more widely distributed than previously suspected.

Key words: **face representation; fMRI; multivariate pattern analysis; neural network models; structured sparsity**

Significance Statement

Brain systems represent information as patterns of activation over neural populations connected in networks that can be widely distributed anatomically, variable across individuals, and intermingled with other networks. We show that four widespread statistical approaches to functional brain imaging have critical blind spots in this scenario and use simulations with neural network models to illustrate why. We then introduce a new approach designed specifically to find radically distributed representations in neural networks. In simulation and in fMRI data collected in the well studied domain of face perception, the new approach discovers extensive signal missed by the other methods—suggesting that prior functional imaging work may have significantly underestimated the degree to which neurocognitive representations are distributed and variable across individuals.

Introduction

Cognitive neuroscience is embracing a network-based view: cognition arises from the propagation of activity over weighted connections among neural populations spanning multiple cortical areas (Sporns, 2011). Information inheres in the similarities among distributed activity patterns rather than the mean activation of local neural populations (Haxby et al., 2014). This raises a statistical challenge for functional neuroimaging, where various technologies yield thousands of measurements per second: cognitive structure must be encoded jointly over some subset of measurements, but the number of possible subsets is prohibitively large. How can the theorist find those that encode structure of interest? We propose a new answer motivated by neural network models (Rogers and McClelland, 2014) and show that it can lead to dramatically different conclusions about the neural bases of cognition even in well studied domains.

Received Apr. 14, 2020; revised Dec. 2, 2020; accepted Dec. 4, 2020.

Author contributions: C.R.C. and T.T.R. designed research; C.R.C. and T.T.R. performed research; C.R.C. and T.T.R. analyzed data; C.R.C. and T.T.R. wrote the paper.

This work was partially supported by Medical Research Council Program Grant MR/J004146/1 and by European Research Council Grant GAP: 670428 - BRAIN2MIND_NEUROCOMP. This research was performed using the computer resources and assistance of the University of Wisconsin (UW)-Madison Center for High Throughput Computing, which is supported by UW-Madison, the Advanced Computing Initiative, the Wisconsin Alumni Research Foundation, the Wisconsin Institutes for Discovery, and the National Science Foundation. We thank Nikhil Rao and Robert Nowak for developing the SOS Lasso loss function and its MATLAB implementation; Mark Seidenberg and Matthew A. Lambon Ralph for valuable discussions of this work in its development; and Brad Postle and Jarrod Lewis-Peacock for sharing functional imaging data with us.

The authors declare no competing financial interests.

Correspondence should be addressed to Christopher R. Cox at chriscox@lsu.edu.

https://doi.org/10.1523/JNEUROSCI.0904-20.2020

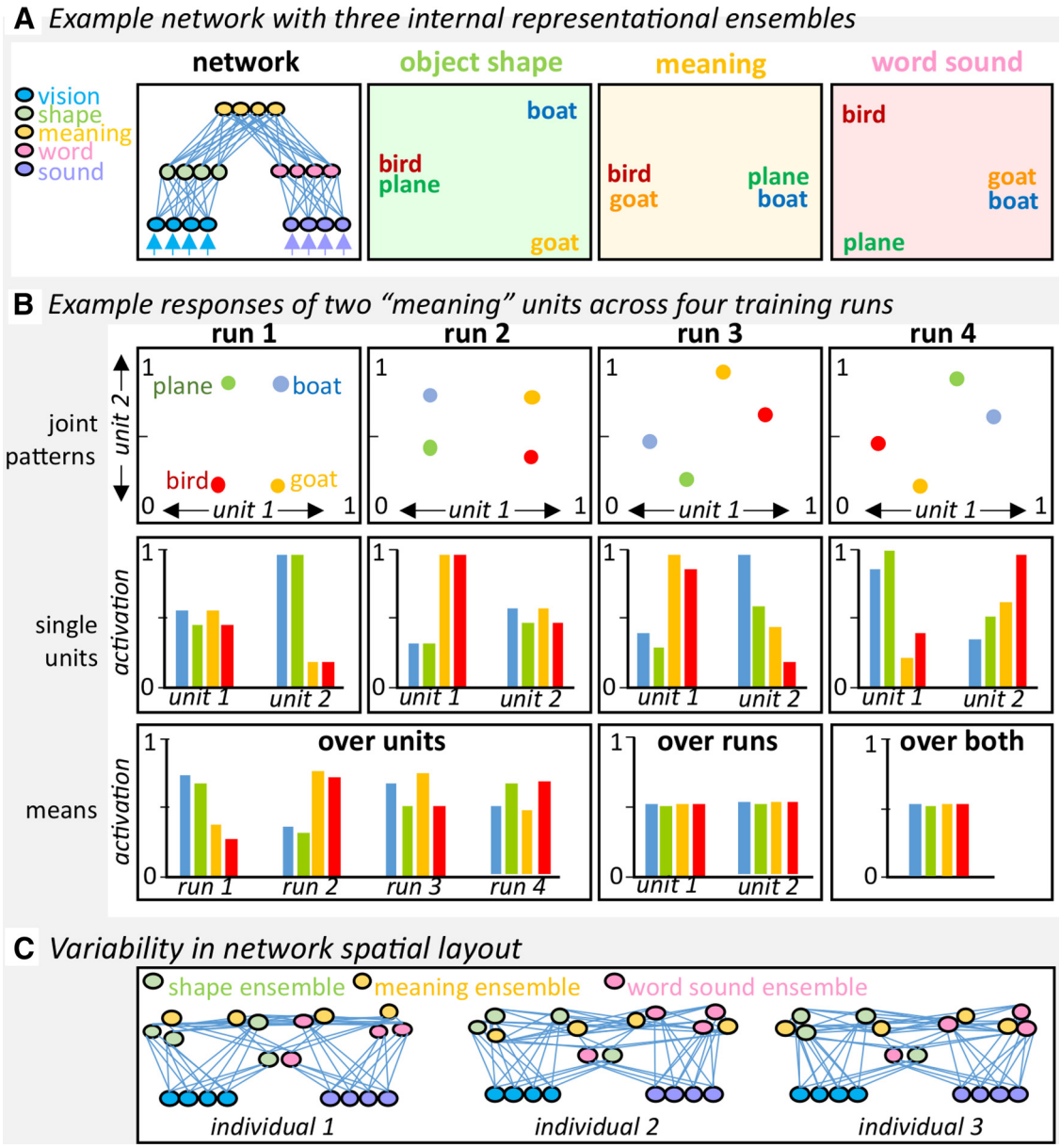

Figure 1. Challenges to functional brain imaging posed by network-based views of cognition. ***A***, Example network that maps visual and auditory inputs to ensembles expressing similarity in shape, meaning, or word sound. ***B***, Hypothetical contributions of two units to the representation of meaning in such a network across four different training runs. Jointly, the two units always encode the same distances among birds, goats, cars, and boats (top row). Considered independently each unit appears to show a different response pattern both within and across runs (middle row). Consequently, when the unit activations are spatially averaged within a run, or averaged across runs for each unit, their contributions to semantic structure are obscured (bottom row). ***C***, Variation in fine-grained connectivity. All three depicted networks have the same connectivity as that shown in ***A*** and so will function the same way despite having a more complex spatial layout. The same coarse topography is expected across individuals but the finer-grained spatial layout—exactly where the green, yellow, and pink units lie within a given spatial volume—may vary, challenging approaches that average within region or across individuals.

Neural network models propose that cognitive representations are patterns of activity distributed over neural populations or units (Rumelhart and McClelland, 1986). The patterns arise as units communicate their activity through weighted connections that determine the effect of a sending unit on a receiving unit. Pools of similarly connected units function as a representational ensemble that encodes a particular cognitive structure (e.g., phonological, semantic, visual). Network topography is initially specified, but learning shapes the connection weights that generate patterns over ensembles. Cognitive processing arises from the flow of activity through the network. Such models have proven useful for understanding the neural bases of healthy (McClelland et al., 2010; Rabovsky et al., 2018), disordered (Lambon Ralph et al., 2017), and developing (Saxe et al., 2019) cognition, but challenge functional imaging because they suggest that neural systems can encode information in ways that elude many statistical approaches (Fig. 1), as follows:

1. Unit activations may not be independently interpretable. If units in an ensemble jointly encode information of interest, independent analysis of each via univariate statistics can obscure critical signal.
2. Neighboring units need not encode the same information in the same way. Adjacent units in a distributed code may express different components of a represented structure or may express the same component through either increased

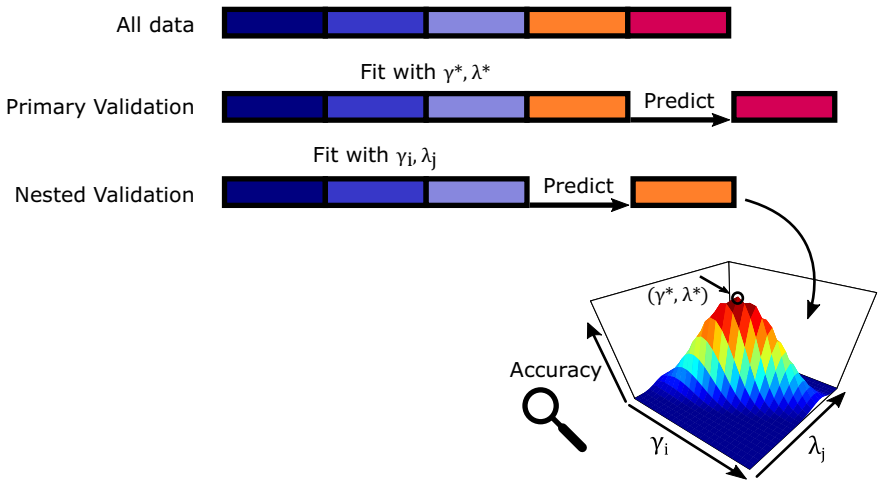

Figure 2. Cross-validation procedure. Multiple hyperparameter configurations are attempted in a nested cross-validation loop completely isolated from a primary validation set. The optimal configuration is used to fit a model to all but the primary validation set. The primary step is repeated holding out a different set of items each time, and each time $\gamma*$, $\lambda*$ is optimized through nested cross-validation.

or decreased activation. Thus, spatial averaging within subjects may destroy signal.

3. Single units can vary arbitrarily in their responses across individuals even when ensembles encode the same structure. Many different weight configurations can compute the same input/output mapping, so a particular unit in a given network can exhibit arbitrary patterns of responding across training runs in the same environment. In this case, response averaging across subjects at a given anatomical location will destroy a signal.
4. Representational ensembles need not be anatomically contiguous. Units with similar connectivity function as a representational ensemble—responding to the same inputs and contributing to the same outputs—even if situated in different cortical regions. Approaches that analyze different regions independently will miss information distributed in this way.
5. Fine connectivity varies across individuals. While coarse connectivity is innate, learning tunes individual weights, rendering precise unit-to-unit alignment across individuals impossible. Cross-subject averaging may therefore destroy information even with sophisticated alignment techniques.

To understand how contemporary methods address these challenges, Study 1 used a neural network model to generate simulated neuroimaging data. The model specifies which units carry information, allowing comparison of methods in their ability to discover different kinds of signal. Each has critical blind spots that suggest a new approach based on structured sparsity (Huang et al., 2011), developed in Study 2. Study 3 compares the different approaches when applied to functional magnetic resonance imaging (fMRI) data in a domain that has highlighted fundamental questions about neurocognitive representation: visual face processing. Each yields a different results, but established approaches all suggest that face processing arises within posterior temporal and occipital regions. The new approach uncovers a radically distributed network spanning all four lobes of the brain. The Discussion considers implications of these results for claims of functional localization from brain imaging data more broadly.

Study 1: assessing established approaches in simulation

The goal of Study 1 was to compare established statistical methods in their ability to discover representational signal encoded by units in a neural network under different assumptions about the nature of the neural code and the spatial (anatomic) layout of the units. To this end, we generated simulated imaging data from the simple auto-encoder network shown in Figure 2*A*. The model was trained 10 times under identical conditions except for weight initialization to reproduce 72 input patterns over the output units. Items were sampled equally from two categories (A and B) corresponding to some cognitive distinction of interest (e.g., faces vs places). Each trained model is analogous to an individual in an fMRI study, with the BOLD signal at a single voxel simulated as the activation of the unit perturbed by independently sampled noise.

Two subsets of units encode category information in different ways. Informative input/output (IIO) units adopt an independent code: each is weakly but reliably more active on average for items in one category (A or B) than the other. Informative hidden (IH) units connect informative input to output units, and thus learn a distributed code: after training, items from the same category always evoke similar patterns across units, but individual units vary arbitrarily in their independent correlations with category structure within and across networks. The model also includes arbitrary input-output and arbitrary hidden (AH) units that respond to stimuli but do not encode category information, and irrelevant units that take low random values.

We considered two anatomic layouts for the network (Fig. 2*A*), corresponding to two different assumptions about how units coding a distributed representation can be spatially organized. Both layouts situated the IO units of a given type (A, B, arbitrary) within a contiguous spatial region localized in the same way across model individuals, analogous to early perceptual areas known to have the same topographic organization across individuals. The localized layout also grouped hidden units by type (IH, AH, irrelevant) in this way (localized identically across model individuals) consistent with the common view that layers in a neural network model correspond approximately to contiguous cortical regions in the brain. The dispersed layout arranged hidden units in four anatomically distal "regions" containing a mix of IH, AH, and irrelevant units, with unit locations shuffled within region for each model subject. This condition represents the possibility suggested by neural network models that neural populations jointly contributing to a representation may be anatomically distal from one another, somewhat variable in their

spatial location across individuals, and intermingled with populations contributing to other representations. All models had the same connectivity—the layouts differed only in hypothesized spatial locations of the units.

We then applied the following four common statistical methods to find units that encode the A/B category structure: univariate contrast (Friston et al., 1994), searchlight multivariate pattern classification (MVPC; Kriegeskorte et al., 2006; Pereira et al., 2009), and whole-brain pattern classification (Lemm et al., 2011) regularized with either the L1 (Tibshirani, 1996) or the L2 (Hoerl and Kennard, 1970) norm. A good method should detect all informative units regardless of spatial layout and should indicate the code direction: more active for A versus B for IIO units and the heterogeneous code for IH units.

Methods

Simulations were conducted using the Light Efficient Network Simulator (Rohde, 1999) with minor revisions to support modern Tcl/Tk libraries (https://github.com/crcox/lens/releases/tag/v1.0) for software and networks files. The model was trained using backpropagation to minimize cross-entropy error with a learning rate of 0.1, a momentum ("Doug's") of 0.9, and weight decay of 0.001. The model achieved near-zero error with 1000 weight updates, each following a batch containing all 72 patterns. The model was fit 10 times with different initial weights sampled from a uniform distribution in [−1,1] to simulate 10 individuals in a brain imaging study. We computed the activation of every unit for each item in each model, then distorted this with i.i.d. (independent and identically distributed) Gaussian noise (mean = 0, SD = 1) to simulate the BOLD response to a stimulus at each voxel. The simulated activation patterns were extended with 28 irrelevant units (zero baseline activity and the same noise profile), which serve two important functions. Conceptually, they represent neural populations that are independent from the representations of interest and not involved with the task. Statistically, they allow us to evaluate the false alarm rate of our methods and construct tests of reliable unit selection.

Statistical analysis

Separate analyses were conducted for the localized and dispersed layouts. All methods used the same uncorrected but conservative criterion for significance ($\alpha = 0.002$).

Univariate contrast spatially smooths data, then identifies voxels whose mean activation across individuals at a given spatial location reliably differs for A versus B items (Friston et al., 1994). Smoothing used a boxcar average over a three unit window. Analysis then involved a two-tailed independent-samples t test (A vs B items) at each unit.

Searchlight MVPC seeks distributed representations by generating information maps across the cortex (Kriegeskorte et al., 2006; Pereira et al., 2009). At each voxel in each subject, a classifier is trained to discriminate items from different categories, based on the neural response evoked in neighboring voxels within a "searchlight" of fixed radius. Holdout accuracy is stored at the center voxel, and univariate tests across subjects then indicate which searchlights perform reliably above chance. Thus, information must be expressed within the searchlight radius and localized similarly across individuals. The model analysis leveraged the SearchMight toolbox (Pereira and Botvinick, 2011) for MATLAB and involved centering a searchlight of fixed radius on each unit. A support vector machine classifier was fit to unsmoothed unit activity within each searchlight with sixfold cross-validation. Unit groups were padded with noise units so that all searchlights contained the same number of units and a searchlight never encompassed units in different "regions." The mean cross-validation accuracy was stored at each searchlight center, and the mean accuracy over model subjects at each unit was compared against chance using a two-tailed t test. We assessed the performance of searchlights ranging in radius from 3 to 14 units and report the best performing searchlight (size 7) together with the smallest and largest.

Whole-brain MVPC with regularized regression

This approach fits and evaluates a single classifier per subject using all voxels and avoids overfitting by finding classifier coefficients β that jointly minimize prediction error and a regularization penalty (Lemm et al., 2011), as follows:

$$\arg\min_{\beta}\big((1-\lambda)f(\beta)+\lambda h(\beta)\big), \quad (1)$$

where $f(\beta)$ is the classifier prediction error, $h(\beta)$ is the regularization penalty, and $\lambda \in [0,1]$ is a free hyperparameter that scales the error versus regularization costs. The approach does not consider voxel location at all and so in principle can find nonlocal information. We used logistic loss for the error and considered two common regularization penalties: ridge regression, which increases with $\sqrt{\sum \beta^2}$, and LASSO, which increases with $\sum |\beta|$.

Analyses were conducted using the Whole-Brain Imaging with Sparse Correlations (WISC) workflow (https://github.com/crcox/WISC_MVPA/releases/tag/FirstMajor). The same procedure was used for all regularized regression analyses, so we will describe it in detail here.

Each analysis involves two rounds with different aims. In the performance round, the aim is to assess how accurately a decoding model, fit to a subset of data, can then classify held-out test items. In this evaluation, it is important that the data used to fit model parameters and hyperparameters are independent of those used to evaluate the fitted model. To this end, we adopted the nested cross-validation procedure shown in Figure 2. Data were divided into n partitions. One partition was held out as a final test set. The remaining partitions were used as a training set to find good model parameters in an inner cross-validation loop. In the inner loop, a series of models was fit to 80% of the training data using different values for λ, with each evaluated on a held-out 20% of the training data, and iterating over different holdout sets. This loop searched many hyperparameter values, seeking the one that yields the best mean accuracy across the inner-loop holdouts. A final model is then fit to all training data at the selected λ, and this model is evaluated for accuracy on the final completely held-out test set. This procedure is conducted for each of the n partitions, yielding n estimates of decoding accuracy on completely held-out items; the expected decoding accuracy is then taken as the mean of those n estimates.

While the performance round indicates the expected accuracy of a decoding model, it does not provide a clear picture of which features the model exploits, since the procedure fits n different decoding models (one for each partition), each specifying its own set of coefficients. We therefore conducted an importance mapping round with the aim of assessing which features (units or voxels) a decoding model selects as "important" for classification. In this round, we select the hyperparameter that yielded the

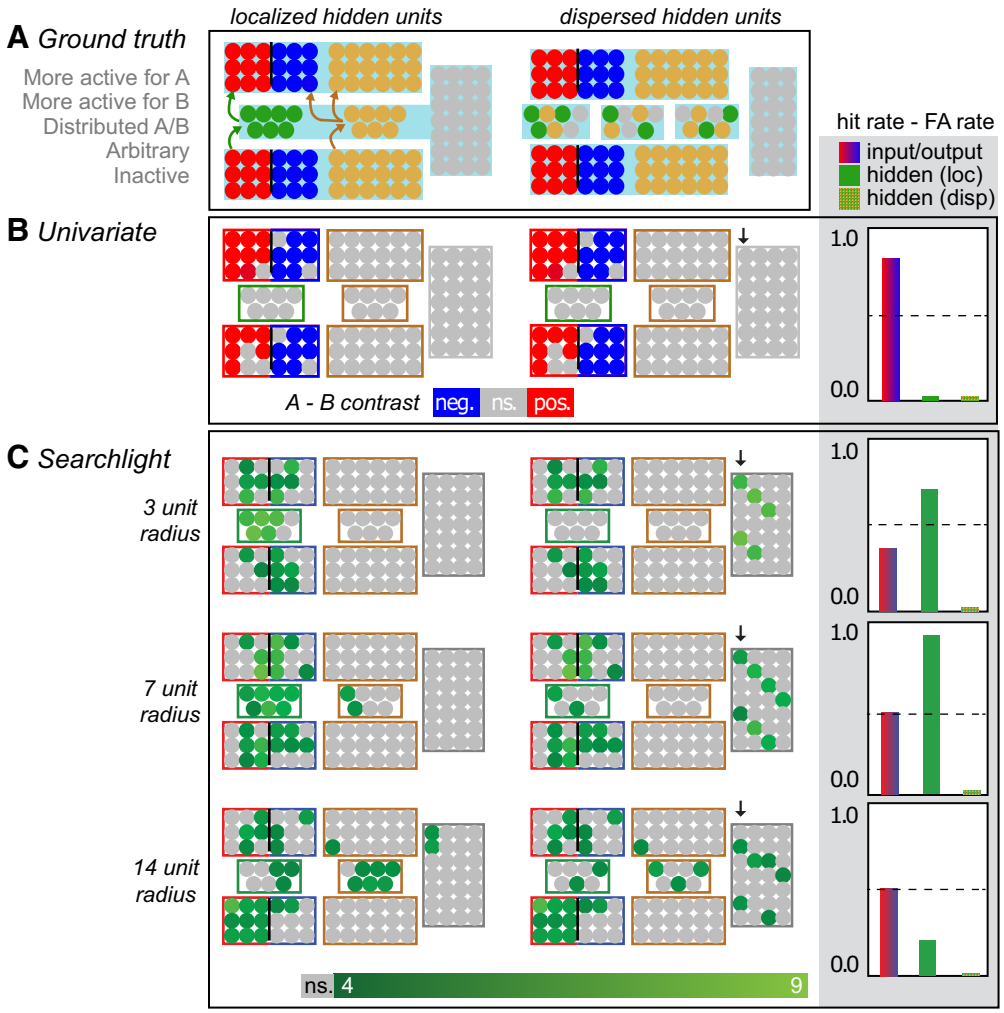

Figure 3. Simulation results for univariate and searchlight. ***A***, Model architecture. Arrows in the left panel indicate connectivity, which is the same in localized and dispersed layouts. Blue shading indicates anatomically contiguous units. ***B–D***, Results for univariate and searchlight. Colored circles indicate units reliably identified by the method. Blue/red indicates the direction of the effect for univariate contrast; searchlight results are shown in green because the method does not indicate direction of effect. Hidden units in localized and dispersed conditions have been laid out in the same way to facilitate comparison. For each plot, a successful method will identify all units in the three leftmost boxes and no units in the remaining boxes. For the dispersed layout, the arrow indicates the column of irrelevant units that were intermixed with arbitrary and systematic hidden units. Bar plots show how accurately each method discriminates signal carrying from arbitrary and irrelevant units for input/outputs (red/blue), localized (green bar), and dispersed (green/yellow bar) hidden units.

highest test accuracy in the performance round, then fit a decoding model to all data using the selected hyperparameter. This yields a single decoding model for each subject, which specifies one coefficient for each feature. This decoder was not evaluated for classification accuracy (since it was fit to all data) but was analyzed to assess which features reliably attract large or nonzero coefficients for a decoding model fit at the optimal hyperparameter.

For LASSO, it is straightforward to determine which features are selected: the regularizer pressures as many coefficients to 0 as possible, so any feature with a nonzero coefficient has been selected. To determine which features are selected in true data more often than expected from random data, we conducted a permutation analysis: the full procedure was run for each model subject on 1000 permutations, identical in all respects to the analysis above but with the category labels shuffled randomly each time. For each permutation, on every unit, we count how often the unit was selected across the 10 model subjects, providing a null distribution for probability of overlap across model subjects. We then count a unit as being reliably selected if the overlap across model subjects observed in the true data is higher than that observed in the permutation distribution with $p < 0.002$.

For ridge regression, all features always receive nonzero coefficients, so it can be unclear how to determine which are "selected" by the classifier. We adopted the common approach of treating units with large coefficients—specifically, those with an absolute value in the top quartile for each model—as having been selected by the classifier. To assess which units reliably attracted large coefficients across model runs, we tabulated, for each unit, the number of times its coefficient was in the top quartile by magnitude across model subjects, then computed the binomial probability of achieving this number or greater from 10 model runs given the base probability of landing in the top quartile by chance (0.25). We counted a unit as being reliably selected when this probability was <0.002.

The performance-round and importance-mapping round were conducted for L1 and L2 approaches using six data partitions.

Results

Simulation results are shown in Figures 3 and 4. We found the following. Univariate contrast (UC) uncovered the independent

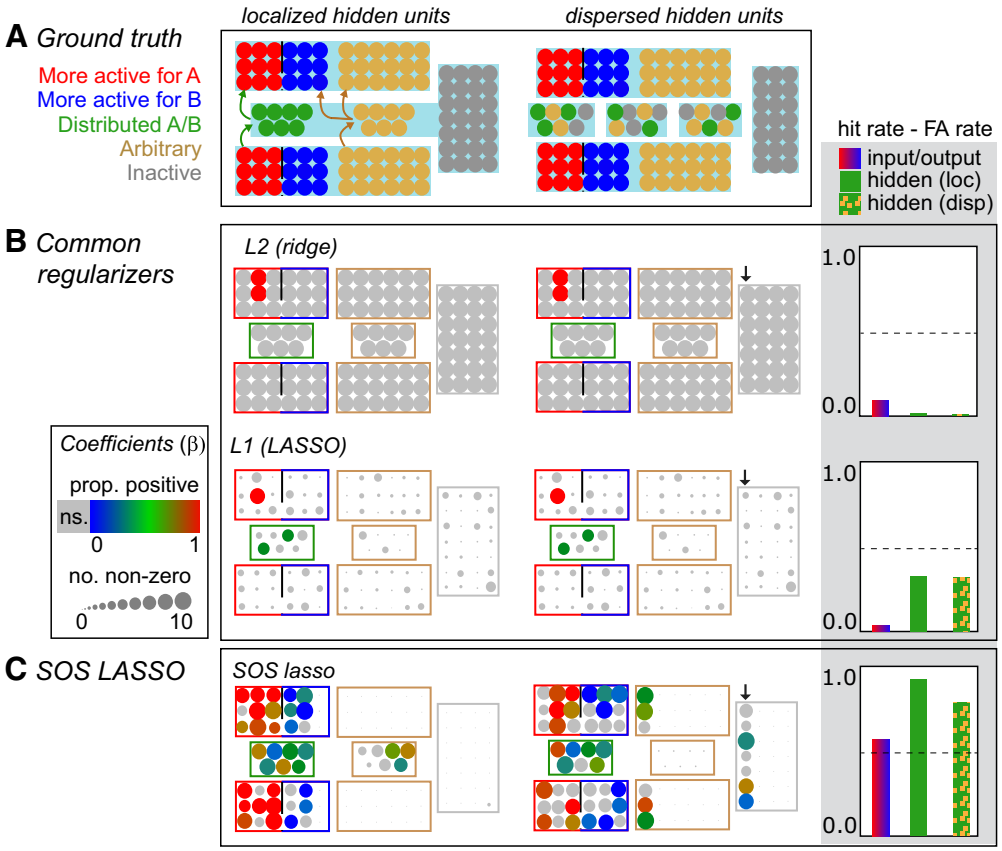

Figure 4. Simulation results for regularized regression. ***A***, Model architecture. Arrows in the left panel indicate connectivity, which is the same in localized and dispersed layouts. Blue shading indicates anatomically contiguous units. ***B***, ***C***, Whole-brain classification with L2 or L2 regularization (Study 1) and the SOS LASSO (Study 2). Colored circles indicate units identified by the method. Circle size indicates the number of times the unit received a nonzero coefficient across 10 model runs. Color indicates the proportion of coefficients that are positive. Hidden units in localized and dispersed cases have been laid out in the same way to facilitate comparison. For each plot, a successful method will always identify all units in the three leftmost boxes (large colored circles) and no units in the remaining boxes. For the dispersed layout, the arrow indicates the column of irrelevant units that were intermixed with arbitrary and systematic hidden units. Bar plots show how accurately each method discriminates signal-carrying from arbitrary and irrelevant units for input/outputs, localized (green bar), and dispersed (green/yellow bar) hidden units.

code and code direction in IIO units but missed the distributed code over IH units in both layouts. Searchlight MVPC, at the best performing radius, discovered half of the independent code in IIO units and almost all of the distributed code in the localized layout, but missed the distributed code in the dispersed layout and always obscured the code direction for individual units (since information maps only reveal classifier accuracy). Smaller searchlight sizes performed similarly, while larger sizes obscured the distributed code even in the localized layout. Whole-brain MVPC with regularized regression showed qualitatively different results depending on the regularization penalty (Fig. 4*B*). With the L2 norm (ridge regression), the classifier showed above-chance holdout accuracy (mean = 59.2%, $t_{(9)} = 5.36$, $p < 0.001$) but placed nonzero coefficients on all units, making it difficult to discriminate informative from arbitrary and irrelevant units. The strategy of selecting units with large coefficients did not identify signal-carrying units, which were no more likely to attract large weights in the classifier than expected by chance.

Regularizing with the L1 norm produced a classifier with equally good holdout accuracy (mean = 57.2%, $t_{(9)} = 2.34$, $p < 0.05$; Ridge vs LASSO paired, $t_{(9)} = 0.626$, n.s.) and a much sparser solution. Only three units (two IH and one IIO) reliably received nonzero coefficients, with no false alarms. Thus, conventional regularization either missed substantial signal (L1) or selected everything (L2).

Study 2: whole-brain MVPC with structured sparsity

We have suggested that structured sparsity (Jacob et al., 2009; Jenatton et al., 2011) can provide an avenue for preserving the strengths of established methods while avoiding their weaknesses. On this approach, a single classifier is fit to data from all subjects while a regularization penalty encourages desired sparsity patterns among the coefficients. In functional imaging, the solution should (1) clearly delineate selected and unselected voxels, (2) allow heterogeneous codes among neighboring units within and across individuals, (3) reveal code direction where this is consistent, (4) identify distal units that jointly express representational structure, (5) capitalize on shared location across subjects, where this exists, but also (6) tolerate individual variation in signal location.

In prior work, we defined the sparse-overlapping-sets (SOS) LASSO to meet these criteria (Rao et al., 2013, 2016). Voxels from all subjects are projected into a common reference space without interpolation or averaging. Sets are defined for grid points in the space, each encompassing all voxels within and across subjects that fall within a specified radius. Sets are analogous to searchlights in many respects; each includes all voxels within a spatially contiguous radius of a center point. As with searchlights, each voxel belongs to several sets, and sets overlap in the voxels they contain. The central difference lies in how this

information about spatial grouping is used in model fitting. Rather than fitting a different decoder to each searchlight/set independently, all sets instead contribute simultaneously to the regularization cost h in Equation 1 as follows:

$$h(S,\beta) = \sum_{S}\left((1-\gamma)\sum_{i}|\beta_{s,i}| + \gamma\sqrt{\sum_{i}\beta_{s,i}^{2}}\right), \qquad (2)$$

where S defines the grouping of voxels into sets, i indexes model coefficients within a set, and $\gamma \in [0,1]$ is a free parameter. The total cost is a sum over sets. The cost for each set is the proportional weighted sum of the LASSO sparsity penalty and a grouping penalty formulated as the root of the sum of squared coefficients. Because this root is taken over units within a set, the penalty is smaller when nonzero coefficients occupy the same set than when they occupy different sets (Rao et al., 2013). Thus, SOS LASSO (https://zenodo.org/record/3609239) encourages sparse solutions where selected voxels tend to occupy the same sets. The free hyperparameter γ controls the relative weighting of the sparsity versus grouping penalties within set. When $\gamma = 0$, the penalty reduces to LASSO. The optimization is convex and returns a unique solution for a given pair of hyperparameters (γ and λ). These are tuned via nested cross-validation on holdout error just as described for L1 and L2 regularization (Fig. 2).

SOS LASSO captures the spirit of searchlight analysis in seeking a whole-brain information map that identifies local regions where many signal-carrying voxels reside. The grouping penalty in the SOS LASSO cost function pressures the optimization to place nonzero coefficients in regions where information is localized similarly across participants. Yet, because SOS LASSO fits a single model to all data simultaneously, it can "see" all searchlights at once, and so can capitalize on information that may be distributed across multiple anatomically distal areas but not discernable within any single region (either because local correlations are weak or because information across regions must be combined for accurate decoding). Moreover, because the hyperparameter on the grouping penalty is tuned by data, the approach can down-weight or even ignore set membership if doing so leads to better prediction in the decoder. For instance, if signal-carrying voxels were not anatomically grouped within or across subjects, the optimization would select a value near 1 for the sparsity parameter (and near 0 for the grouping parameter), leading to a sparse solution that can be completely different in each participant.

SOS LASSO also has one property in common with L1 and L2 regularization that contrasts with searchlight MVPC: it returns a single whole-brain classifier for each individual subject. The holdout accuracy of this single classifier can be evaluated against chance to assess whether reliable signal has been found in each subject individually, without punishing correction for multiple comparisons and without assumption about homogeneity of signal location across subjects. While it is technically possible to conduct single-subject analyses using the searchlight procedure, this approach requires thousands of statistical tests in each participant to evaluate the classification accuracy of each searchlight. Thus, accuracy must be exceedingly high to survive correction for multiple comparisons. Consequently, the most common searchlight application involves testing the mean classification accuracy across subjects against chance at each voxel/location, which in turn relies on the assumption that important signal is localized similarly across individuals.

In these ways, SOS LASSO captures strengths while avoiding the limitations of other multivariate approaches. We also note that the approach is related to, but distinct from, other sparsity-based approaches that have been applied to neural decoding. The elastic net regularizes a whole-brain classifier with the weighted sum of the L1 and L2 norms (calculated across all model coefficients; Zou and Hastie, 2005). This is closely related to the regularization cost per set in SOS LASSO, but the set-wise evaluation of the cost has an important consequence: since the full cost is a sum over sets, the optimization seeks to have as many "empty sets" (all coefficients zero) as possible. Thus, SOS LASSO enforces a hard distinction between selected and unselected voxels (those with/without a nonzero coefficient). In contrast, elastic net, like ridge regression, often places small nonzero values on all features, and there exists no principled way of deciding which are important and which not. Our approach is also related to, and in fact generalizes, the overlapping group lasso, a common approach to multitask learning (Jacob et al., 2009) in which all features (e.g., voxels) in a selected group are constrained to have the same coefficient, and sparsity patterns are predefined. Our formulation allows different coefficients within set (so that solutions can vary across individual participants) and does not require a prespecified sparsity pattern. These relationships, together with mathematical analysis of the regularizer, are explained further in the studies by Rao et al. (2013, 2016).

Computational efficiency

It is worth noting that SOS LASSO demands considerably more computational resources than simple regularization with the L1 or L2 norm, for two reasons. First, it requires search over two independent hyperparameters, squaring the number of iterations that must be conducted during the inner loop of the nested cross-validation. Second, because the approach conducts a single optimization on data from all subjects simultaneously, then the full dataset must be transferred to each worker node for each model fit—that is, the approach cannot conduct separate analyses on each participant dataset in parallel. This exerts significant memory and data transfer demands. Yet, much of the workflow we have described is embarrassingly parallel—for instance, all steps of the inner loop cross-validation, including model fitting for each combination of hyperparameters on every fold, can be conducted in parallel on a high-throughput network such as Open Science Grid. Thus, while the analyses we report would require years of computing time on a single workstation, they typically complete overnight on the large HTCondor system deployed at the University of Wisconsin-Madison.

Statistical analyses

Decoding model data with the SOS LASSO was implemented using the same WISC workflow, cross-validation procedures, and permutation testing described for regularized regression in Study 1. To fit the SOS LASSO decoding models, each unit was assigned to one or more sets, with each set containing 14 consecutive units (analogous to searchlight radius 7 in this one-dimensional dataset), and with seven units overlapping between adjacent set pairs. The grouping and sparsity hyperparameters (γ and λ) were tuned using nested cross-validation with 10 data partitions. Decoding accuracy was evaluated in a performance round as described earlier, while feature selection was evaluated in an importance mapping round, also as described for Study 1. As with LASSO, the importance-mapping procedure returned a single decoder for each model "subject," with many coefficients pressured to zero. Thus, we again treated any unit in any model receiving a nonzero coefficient as having been selected by the method and used permutation testing to assess which units are

selected in true data more often than expected from random data.

Results

Applied to model data, SOS LASSO achieved the best holdout accuracy (68.3%), significantly better than the next best method (LASSO, paired $t_{(9)} = 4.869$, $p < 0.001$) while uncovering much of the independent code and almost all of the distributed code in both layouts (Fig. 4C). The sign of the classifier coefficients revealed the consistent code direction in IIO units and code heterogeneity in IH units. SOS LASSO alone discovered the anatomically dispersed distributed signal, while preserving the strengths of the other methods.

Study 3: finding distributed representation of faces in neural data

The simulations suggest that functional brain imaging studies may have missed important distributed signal even in well studied domains where multiple contemporary approaches have been applied. Study 3 assessed this possibility for visual face perception. We applied univariate contrast, searchlight MVPC, LASSO, and SOS LASSO to fMRI data collected in an unrelated study (Lewis-Peacock and Postle, 2008) where 10 participants evaluated 90 images depicting people, scenes, or objects (30 each). The study used a slow event-related design estimating the peak BOLD response to each image at each voxel without deconvolution. We applied each method to these data to find voxels whose activations discriminate face (people) from nonface (place and object) stimuli.

Methods

The fMRI data (Lewis-Peacock et al., 2018) were collected by a different group in an independent study that reports the full methodology and image acquisition details (Lewis-Peacock and Postle, 2008). In a slow event-related design, subjects viewed 30 images from each of three categories (celebrities, famous locations, objects) in permuted order while their brains were scanned with fMRI. Images were acquired with a gradient echo, echoplanar sequence with 2000 ms repetition time (TR) and 50 ms echo time to acquire data sensitive to the BOLD signal within a 64×64 matrix (30 axial slices coplanar with the T1 acquisition, $3.75 \times 3.75 \times 4$ mm). With button press on a 4 point Likert scale, they indicated their liking for the celebrity/location or familiarity with the object. Each trial consisted of a cue (2 s), stimulus (5 s), and judgment period (3 s) followed by an arithmetic task (16 s) to reduce interference between trials. Each stimulus appeared once. Functional data for each subject were masked to exclude noncortical voxels. The response to each stimulus at every voxel was taken as the BOLD signal recorded at the fifth TR following stimulus onset, without time-series deconvolution.

Univariate analysis

Functional images from the fifth TR following stimulus onset were projected to Talairach space using the T1 data with a combination of manual landmark identification (anterior and posterior commissures) and automated affine transformation obtained using 3dvolreg in AFNI (Analysis of Functional NeuroImages). The response to each stimulus was smoothed with 4 mm FWHM Gaussian kernel and downsampled to the original resolution. At each voxel, we computed the mean response for face and nonface stimuli for each subject. In a whole-brain analysis, we tested for a group-level difference between these means at each voxel with a two-tailed dependent-samples t test (clusterwise $\alpha = 0.05$). In a region of interest (ROI) analysis, we computed, for each subject, the difference in BOLD response to faces versus nonfaces averaged across voxels in the "right fusiform face area" (rFFA) mask from the study by Julian et al. (2012), then conducted a one-tailed t test (faces > nonfaces) against zero across subjects.

Searchlight MVPC

We used SearchMight (Pereira and Botvinick, 2011) with a linear support vector machine classifier to generate native-space information maps for each subject using both a standard 9 mm and a larger 12 mm searchlight. Holdout accuracy for each searchlight was measured as the average of the hit and correct rejection rates, a metric with range [0,1] and chance performance of 0.5, corresponding to 50% accurate categorization, regardless of the number of items in each category. Throughout, we refer to accuracy in percentage rather than proportional terms. Data for all subjects were projected to Talairach space, spatially smoothed, and downsampled to native resolution as in the univariate analysis. A one-tailed t test against 0 (true positives > false positives) was conducted on the mean of this metric across subjects (clusterwise $\alpha = 0.05$).

For whole-brain MVPC with L1 regularization, we evaluated holdout accuracy separately for each subject in a performance round using nested 10-fold cross-validation exactly as described for the simulation (Fig. 2). For each subject, the model was fit using all cortical voxels without any ROI or voxel preselection. This analysis allowed us to assess how accurately a classifier fit with L1 regularization could discriminate faces from nonfaces in each subject considered independently.

We then assessed whether decoding models across participants place nonzero coefficients in common regions using the same procedure used in the simulation. For each subject, we fit a single decoding model to all data, using the best performing hyperparameter discovered in initial assessment of decoding accuracy. This final model returned a single coefficient at each cortical voxel for every subject (with, of course, many coefficients set to 0). The classifier coefficients for each subject were projected to Talairach space with linear interpolation, smoothed with a 4 mm FWHM Gaussian kernel, and downsampled to the original resolution. We then counted how often each common-space voxel was selected across the 10 subjects. To statistically threshold this group map, we conducted a permutation test in which the identical procedure was followed, but with stimulus labels randomly permuted. From 1000 permutations, we estimated the base probability of selection under the null hypothesis for each voxel, then used this probability and the binomial distribution to threshold the group map of the true data at $p < 0.002$ uncorrected. For instance, if the probability of selection from permutations was 0.15, the voxel would exceed threshold if, in the true data, it was selected in $\geq$6 of the 10 participants (binomial probability, $\sim$0.0014).

SOS LASSO model fitting followed a similar procedure. To estimate decoding accuracy, we fit decoding models to all subject data simultaneously using the SOS LASSO optimization. Voxels were spatially aligned within Talairach coordinates without blurring or linear interpolation (i.e., we simply adjusted the spatial coordinates of each voxel based on the subject-to-common-space affine transform). Voxels within and across subjects were then assigned to multiple overlapping sets by tiling Talairach space with 18-mm-diameter cubes, overlapping by 9 mm along each axis. This diameter was chosen to

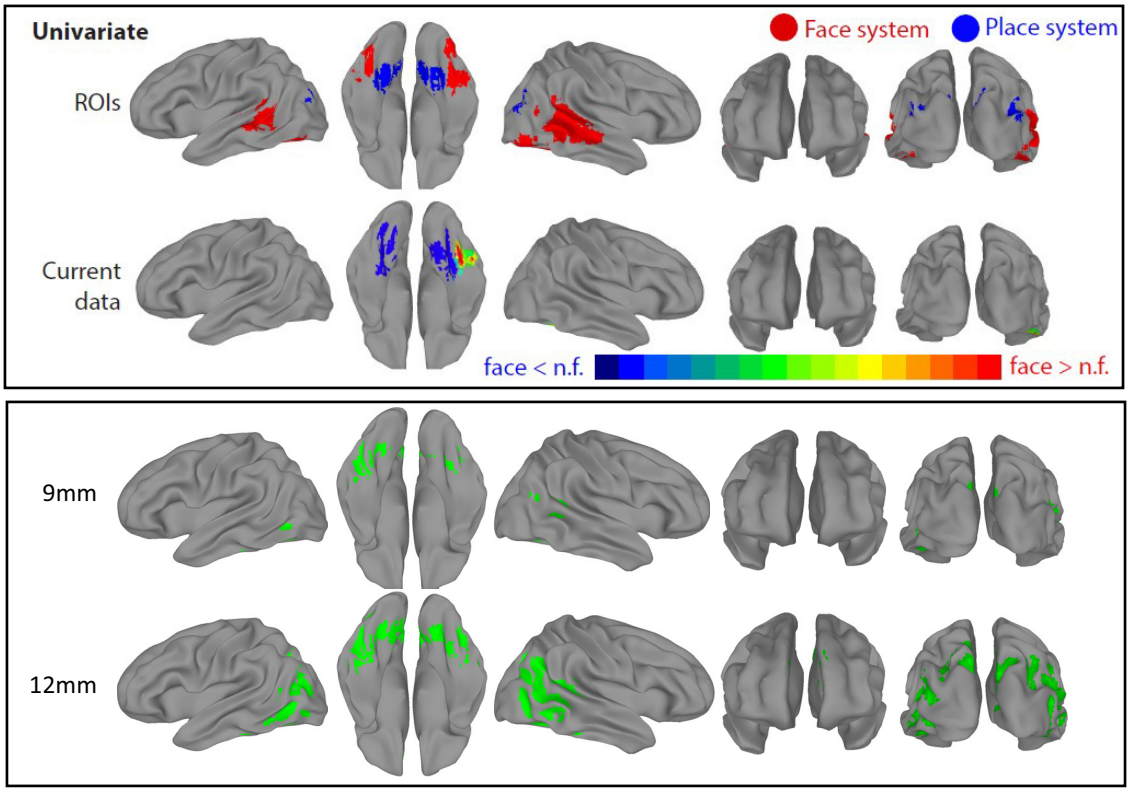

Figure 5. Functional imaging results with univariate and searchlight. Univariate: the canonical face and place systems (top) and results for the current data (bottom). Searchlight: Significant regions are shown in green since the code direction is undetermined in this approach. The top row shows results with a 9 mm searchlight radius, the bottom with 15 mm. n.f., Not face.

be comparable to the 9 mm radius commonly used in searchlight. We note, however, that SOS LASSO is less sensitive to the group size than searchlight because (1) groups overlap, (2) voxels can be selected from multiple groups simultaneously, and (3) solutions can be sparse within groups. We again applied nested 10-fold cross-validation to estimate decoding accuracy in each subject. To determine whether selected voxels accumulate in similar regions across subjects, we conducted an importance-mapping round as described earlier, yielding a single decoding coefficient for each voxel in every subject. To visualize the spatial/anatomic distribution of these coefficients across subjects, we flagged selected voxels with a binary mask (1 for voxels with a nonzero coefficient, 0 otherwise), then spatially smoothed the mask with a truncated 4 mm FWHM Gaussian kernel.

Because SOS LASSO fits all subjects simultaneously, selection of a voxel in one subject alters the probability of selection in a nearby region in other subjects. Thus, the binomial assumption of independence across subjects adopted to threshold the LASSO analysis is violated in the SOS LASSO case. We instead adopted a stricter criterion for significance based on permutation sampling. For each permutation round, the entire SOS LASSO pipeline was conducted using randomly shuffled category labels in the model fit. Selected voxels were subjected to a binary threshold and spatially smoothed in the identical way. For each voxel in the common space, we then counted how many subjects received a nonzero value in the smoothed map after applying the threshold. Across 1000 such analyses with randomly permuted labels and ~10,000 common-space voxels, no voxel was ever selected in more than seven subjects. We therefore masked the SOS LASSO maps to show voxels selected in eight or more subjects in the true data. Since all permutations are independent, this establishes a lower significance bound of $p < 0.001$ uncorrected. All follow-up analyses with SOS LASSO used the same procedure.

Results

Figure 5 shows canonical "face" and "place" systems from the study by Julian et al. (2012) together with results for the current data from univariate contrast and searchlight methods. The univariate approach revealed significantly less activation for faces around parahippocampus bilaterally ($p < 0.05$ cluster corrected), while the ROI analysis found greater activation for faces in the right rFFA ($p < 0.05$). Searchlight MVPC with a 9 mm radius identified areas near the FFA bilaterally ($p < 0.05$ cluster corrected) and in lateral occipitotemporal regions of both hemispheres, while a 12 mm radius identified similar areas with a broader anatomic spread ($p < 0.05$ cluster corrected). As noted for simulations, this approach was unable to reveal code direction.

Results for whole-brain decoding with sparse regularization are shown in the top panel of Figure 6. The decoding models fit with LASSO and SOS LASSO showed equally high holdout accuracy in the performance round (LASSO, 88.6%; SOS, 86.8%; $t_{(9)} = 1.23$, SE = 0.015, n.s. for within-subjects contrast). Analysis of final model coefficients for LASSO yielded a group result similar to searchlight: voxels selected in the group map more often than expected from permutations resided near FFA bilaterally, in right lateral occipitotemporal cortex, and small area in left occipitotemporal cortex ($p < 0.002$ uncorrected). In

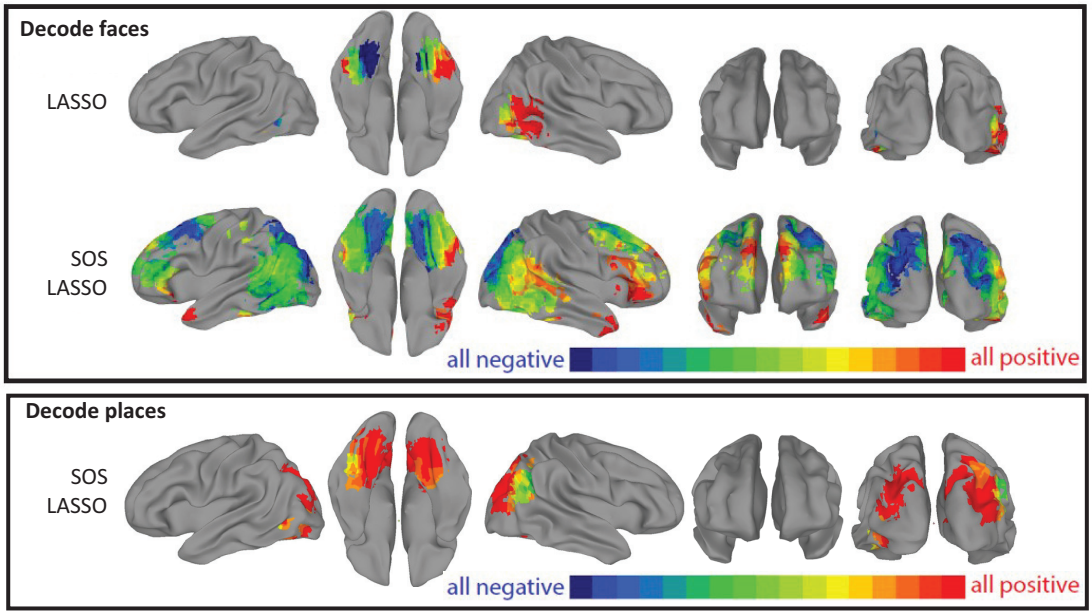

Figure 6. Decoding with sparse whole-brain pattern classification. Top, Regions reliably selected across participants by LASSO and SOS LASSO for classifiers discriminating faces from other stimuli. Bottom, Regions reliably selected by SOS LASSO for classifiers discriminating places from other stimuli. Hue indicates the proportion of nonzero classifier coefficients that were positive.

contrast with searchlight, the classifier coefficients indicated a consistent topographic code, with faces predicted by reduced activity around parahippocampus and elevated activity in lateral ventrotemporal regions of both hemispheres and in right occipitotemporal regions. Thus, these three approaches suggest similar but nonidentical conclusions about face representation in cortex, each with precedent in prior work: the UC result suggests that faces selectively activate the right FFA (Kanwisher et al., 1997); searchlight identifies localized bilateral signal around the FFA (Rivolta et al., 2014); and LASSO reveals a nonface-to-face gradient bilaterally in these regions plus a right-lateralized occipital face region (Martin, 2007).

Results from SOS LASSO differed strikingly. Colored areas in Figure 6 (top, bottom row) show regions consistently selected across participants more often than occurred in permutation testing. The hue depicts the proportion of participants receiving a positive coefficient in the corresponding region; positive coefficients signify that elevated BOLD response is associated with higher probability that the stimulus is a face. In addition to the canonical face and place systems, the results implicate regions spanning anterior temporal, frontal, and parietal cortex. Face stimuli were predicted by increased activity throughout the canonical face system, bilaterally in temporal poles and superior medial frontal areas, and in right inferior prefrontal cortex; and by decreased activity in parahippocampus, the dorsal visual stream, and left premotor cortex. Heterogeneous codes appeared in left occipito-temporo-parietal and lateral prefrontal cortices, and bilaterally in posterior fusiform.

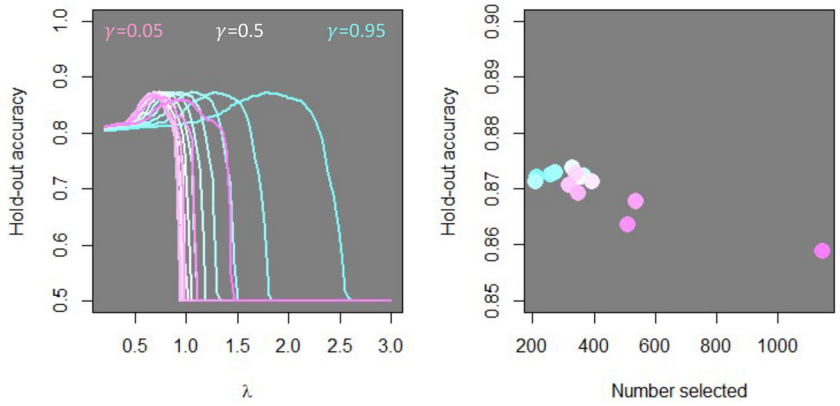

Figure 7. Decoding accuracy and sparsity for different hyperparameters. The left plot shows the mean classifier holdout accuracy during the inner loop of the performance round, for each of 15 different values of the grouping parameter γ (different lines) and across 100 values of the regularization parameter λ (*x*-axis). Pink curves have a low value on the grouping parameter and high sparsity; white curves equally weight grouping and sparsity; cyan curves strongly weight grouping with little weight on sparsity. Each curve peaks at a similar accuracy, indicating that there are many pairs of hyperparameters [γ, λ] that lead to similar decoding accuracy. The right panel shows the mean number of voxels selected per subject (*x*-axis) for the best performing model at each value of γ. Models that do comparably well all tend to select between 200 and 400 voxels per subject. Thus, the various hyperparameter pairs that produce the best decoding accuracy all tend to find solutions at a comparable level of sparsity.

How is the stark difference between SOS LASSO and other solutions to be interpreted? One possibility is that the approach exploits real signal within the canonical occipitotemporal face and place systems, but also places nonzero coefficients on voxels outside this system that do not carry actual signal—perhaps because they suppress correlated noise within the signal-carrying system (Henriksson et al., 2015) or for some other spurious reason. If that were so, the distributed-seeming signal revealed by SOS LASSO would be highly misleading. We therefore conducted two follow-up analyses to assess whether SOS LASSO can detect real signal outside of the canonical face and place systems.

In the first, we divided the data into the following two sets: a within-system set, defined as the canonical face and place system

ROIs dilated by 7 mm and then eroded by 5 mm with AFNI 3dmask_tool (effective dilation, ~3 mm in all directions while filling holes after back-projecting to 3 × 3 × 3.75 mm resolution in each subject with AFNI 3dfractionize); and an out-of-system set containing all remaining voxels. We used SOS LASSO to fit separate decoding models for each data subset. If the whole-brain result succeeds solely by decoding information contained within the canonical face and place systems (and selects out-of-system voxels for spurious reasons), then the within-system decoder should show better holdout accuracy than the out-of-system decoder. Instead, the reverse result was obtained: out-of-system classifiers showed significantly higher holdout accuracy (87%) than within-system classifiers (83%; two-tailed within-subjects test: $t_{(9)} = 2.4$, SE = 0.015, $p < 0.05$).

To address the possibility that good classification accuracy was still driven solely by signal within a temporo-occipital network (Haxby et al., 2000), we next used SOS LASSO to fit a decoding model using voxels in the parietal and frontal lobes only. Holdout classification accuracy remained high (85.52%). The analogous analysis with LASSO achieved 82.21% accuracy, clearly well above chance but reliably worse than SOS LASSO (two-tailed within-subjects test: $t_{(9)} = 2.4$, SE = 0.014, $p = 0.039$). Note that LASSO shows this good performance among a group of voxels that whole-brain LASSO did not initially select—clearly indicating that the strong emphasis on sparsity in LASSO can lead it to miss signal-carrying voxels.

Finally, if the widely distributed signal that SOS LASSO identifies arises because of correlated noise suppression or for other artifactual reasons, then a similarly distributed pattern should obtain regardless of the kind of stimulus being decoded. To assess whether this is so, we fit SOS LASSO decoding models using all cortical voxels to discriminate place from nonplace stimuli, applying procedures identical to those described earlier. The resulting models showed high test-set classification accuracy (79.3%), but from a very different distribution of selected voxels. As shown in the bottom panel of Figure 6, voxels discriminating place from nonplace stimuli were anatomically localized in a manner consistent with the canonical view of place representation in the brain (Epstein and Kanwisher, 1998; Epstein et al., 2003). The analysis demonstrates that SOS LASSO can yield anatomically localized solutions, and hence that the widely distributed result for faces is not artifactual. Together, these analyses suggest that the SOS maps differ from those produced by other methods, not because the method selects uninformative voxels, but because it detects reliable signal missed by other approaches.

Finally, we considered to what extent the SOS LASSO behavior depends on a particular choice of hyperparameters. Specifically, we examined the mean decoding accuracy for each pair of hyperparameter values evaluating during the performance round of model fitting, as well as the number of voxels selected for the best performing models selected at each level of the grouping parameter γ. These data are shown in Figure 7.

The left panel of Figure 7 shows that, for each level of the grouping parameter (the different lines), the decoding accuracy first rises as λ increases, then declines to chance (as the weight on the regularizer takes precedence over model accuracy). Each such curve peaks at a similar value, indicating that there are many pairs of parameters $[\gamma, \lambda]$ that produce comparable decoding accuracy. The right panel of Figure 7 shows, however, that most such pairs find solutions of comparable sparsity. The plot shows holdout accuracy for the best-performing models at each level of γ, plotted against the number of selected voxels per subject. All models that show decoding accuracy comparable to the best also select a comparable number of voxels (~300).

General discussion

Four contemporary approaches to functional image analysis have difficulties discovering distributed signal that are remedied by a new approach based on structured sparsity, the SOS LASSO. All yielded different results when applied to fMRI data collected while participants judged images of faces, places, or objects, but prior approaches supported the common view that face perception engages posterior temporal and occipital cortices. SOS LASSO uncovered a broader network encompassing anterior temporal, frontal, and parietal regions. The result does not solely reflect the selection of spurious voxels: the approach discriminated face from nonface stimuli with high accuracy even when fit only to voxels outside temporal and occipital cortices and yielded a more localized solution when discriminating place from nonplace stimuli. Nor does the result reflect idiosyncrasies of the stimuli or task since the same dataset yielded canonical results when analyzed using established methods. SOS LASSO was the only method capable of finding distributed signal in simulation, and the only one to reveal a radically distributed network for face perception in the brain.

Many aspects of the SOS LASSO result cohere with standard views of face/place perception and the broader literature. Positive coefficients picked out the canonical face system and social-cognitive areas including the temporal poles (Olson et al., 2007), right orbito-frontal cortex (Adolphs, 2002), and superior medial-frontal cortex (Amodio and Frith, 2006). Negative coefficients picked out areas that encode less socially critical information, including scenes (parahippocampal place area; Epstein and Kanwisher, 1998) and object-directed action (dorsal visual stream, left dorsal premotor area; Kalénine et al., 2010). Other regions received mixed coefficients, indicating that distributed patterns can represent stimulus category in a manner that varies within and across individuals. These results echo recent work suggesting that neural representations may be more broadly distributed than heretofore suspected, for face perception specifically (Hanson and Halchenko, 2008; Zhen et al., 2013; Nestor et al., 2016) and for conceptual structure more generally (Huth et al., 2016; Pereira et al., 2018).

The contrasting finding of distributed face versus more localized place representations also accords with recent literature. Whereas early observations of a highly localized face area in right posterior fusiform have given way to an elaborate network of face-relevant regions, the same is not true for place processing in the brain. A recent review from Epstein and Baker (2019) indicates that the place system involves just three anatomically proximal regions—medial ventrotemporal cortex, lateral occipital cortex, and medial occipital cortex—all of which are encompassed in the anatomically localized region identified by SOS LASSO when decoding places.

The contrasting results have two important implications for hypotheses about neural representation. First, even large (12 mm) searchlights missed the widely distributed signal, indicating that it does not reside within local cortical regions considered independently. Thus, anatomically distal regions can jointly encode multivariate representational structure. Second, SOS LASSO assigned consistently positive or negative coefficients in regions where UC yielded a null result. Whereas UC considers each voxel independently, classifier coefficients indicate how the activation of a voxel contributes to classification when combined

with those of other voxels. The contrast indicates that voxels can jointly contribute to representational structure in a directionally and locationally consistent manner even when they do not appear to independently correlate with that structure. In both cases, the joint contribution of multiple voxels or regions to representational structure may arise simply through the linear combination of correlations that are individually too weak to be detected by univariate methods, or because the contribution of one voxel or region to structure can be "masked" by the effects of others (e.g., in partial correlation). As a linear model that does not consider interactions, SOS LASSO cannot adjudicate these possibilities yet; but in either case it is clear that the joint consideration of multiple regions in parallel can uncover a broader structure than the analysis of each region independently.

In addition to core posterior temporal and occipital areas, maps of the extended face network often include anterior temporal cortex, inferior frontal cortex, and amygdala (Haxby et al., 2000; Ishai, 2008; Kanwisher and Barton, 2011; Avidan et al., 2014; Duchaine and Yovel, 2015). Our results with SOS LASSO suggest an even broader network in which face perception generates elevated activity in social-cognitive networks overlapping with the extended face network, reduced activity in networks relevant to navigation and object-directed action, and heterogeneous patterns in parietal and frontal cortices. Elements of the full pattern vary in the direction, independence, and localization of their code, so previous methods each provide only a partial view. Their agreement with canon arises because core and extended face systems comprise parts of the pattern discoverable via multiple methods—regions where information is encoded efficiently and independently within circumscribed cortical regions, often in a directionally consistent manner localized similarly across subjects.

Relationship to prior work

The relationship between representations acquired by neural network models and patterns of activation measured in the brain is currently an active area of research, especially in visual neuroscience (Kriegeskorte and Kievit, 2013; Cox and Dean, 2014; Khaligh-Razavi and Kriegeskorte, 2014; Clarke et al., 2015; Kriegeskorte, 2015; Cichy et al., 2016; Kheradpisheh et al., 2016; Marblestone et al., 2016). The most common approach compares the similarity structure arising in different layers of a neurocomputational model to those arising in different regions of the ventral visual processing stream, using correlation or other unsupervised methods. Such efforts have revealed important and remarkable similarities between the representations observed in real brains and those acquired by, for instance, deep convolutional image classifiers (Kriegeskorte, 2015). Like our work, these contributions illustrate how artificial neural network models provide a useful tool for bridging neural and computational levels of explanation for cognitive phenomena. Focusing on individual network layers and localized cortical regions, however, neglects the critical possibility suggested by network models that representation and processing can be distributed across anatomically distal regions. The current work illustrates not only how such a signal can be recovered in principle, but that radically distributed representations of this kind arise in human cortex.

The work also elaborates the evolving distributed view of visual object representations in the brain. Early perspectives emphasized a highly modular and localized view in which different brain regions were dedicated via evolutionary pressures to representing distinct categories of objects (Kanwisher et al., 1997; Caramazza and Shelton, 1998; Epstein and Kanwisher, 1998; Kanwisher, 2000; Downing et al., 2001). This began to change with the pioneering application by Haxby et al. (2000) of unsupervised multivariate methods to analysis of evoked patterns of activation in the ventral processing stream, which revealed a more distributed code localized within posterior ventral temporal cortex. As new multivariate methods have developed, many studies have argued for distributed visual object representations across occipital and posterior temporal regions (Haxby et al., 2001; Kriegeskorte et al., 2008b; Connolly et al., 2012), with some extending the face system up to anterior temporal cortex (Kriegeskorte et al., 2007; Nestor et al., 2008; Tsao et al., 2008; Rajimehr et al., 2009; Nestor et al., 2011; Collins and Olson, 2014). Our work continues the general trend to suggest that face representations are distributed even more broadly, across temporal, parietal, and frontal lobes in both hemispheres.

Zhen et al. (2013) reported a distributed network for face processing that spans all four lobes of the brain, making it, to the best of our knowledge, the one prior study that reports a pattern of results that resembles our own. Notably, they conducted a group constrained subject specific (GSS) univariate analysis (Fedorenko et al., 2010). While this study provides an important external validation of our results, it is worth noting a key methodological difference. Because GSS is univariate, it cannot detect information jointly encoded across voxels or regions; and because tests of significance are based solely on cross-group probabilities, it relies on common localization of signal across individual participants. As we have emphasized, SOS LASSO can exploit joint information across voxels, and returns an information map for each participant. Thus, while it is possible to inspect cross-subject consistency in location in a group map as we have done, the method does not *require* signal to be localized similarly across subjects. It therefore becomes an empirical question whether the signal-carrying voxels across subjects reside in similar anatomic locations. Perhaps because of these differences, we found the radically distributed face code with many fewer subjects (they included 42, 10 of whom were scanned seven times each) much more consistently (in 80% of participants vs a maximum of 25%), and in key regions not observed by Zhen et al. (2013) but known to be involved in face processing, such as the left temporal pole (Gorno-Tempini et al., 1998; Griffith et al., 2006).

We have focused on discriminating one experimental condition from another with multivariate classifiers. Alternatively, representational similarity analysis (RSA) seeks voxel sets that jointly express a target similarity structure (Kriegeskorte et al., 2008a), while "generative" approaches predict whole-brain images from externally derived stimulus features or experimental condition (Mitchell et al., 2008), each typically implemented in ways that limit their ability to detect network-distributed signal. Structured sparsity may likewise provide new insights for these kinds of problems (Oswal et al., 2016).

Other groups are also exploring the utility of structured sparsity for identifying distributed neuro-cognitive representations in neuroimaging data (Carroll et al., 2009; Baldassarre et al., 2012; Jenatton et al., 2012; Rish et al., 2012; Manning et al., 2014; Cohen et al., 2017). Understanding the relations among these approaches and SOS LASSO is a central goal for future research.

Data availability

The whole-brain imaging with Sparse Correlations (WISC) MVPA tools and code for specifying simulations can be found at

https://github.com/crcox/WISC_MVPA and https://github.com/crcox/SOSLassoSimulations, respectively.

References

Adolphs R (2002) Neural systems for recognizing emotion. Curr Opin Neurobiol 12:169–177.

Amodio DM, Frith CD (2006) Meeting of minds: the medial frontal cortex and social cognition. Nat Rev Neurosci 7:268–277.

Avidan G, Tanzer M, Hadj-Bouziane F, Liu N, Ungerleider LG, Behrmann M (2014) Selective dissociation between core and extended regions of the face processing network in congenital prosopagnosia. Cereb Cortex 24:1565–1578.

Baldassarre L, Mourao-Miranda J, Pontil M (2012) Structured sparsity models for brain decoding from fMRI data. In: Proceedings—2012 2nd International Workshop on Pattern Recognition in NeuroImaging, PRNI 2012, pp 5–8. Los Alamitos, CA: Conference Publishing Services, IEEE Computer Society.

Caramazza A, Shelton JR (1998) Domain-specific knowledge systems in the brain: the animate-inanimate distinction. J Cogn Neurosci 10:1–34.

Carroll MK, Cecchi GA, Rish I, Garg R, Rao AR (2009) Prediction and interpretation of distributed neural activity with sparse models. Neuroimage 44:112–122.

Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A (2016) Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep 6:27755.

Clarke A, Devereux BJ, Randall B, Tyler LK (2015) Predicting the time course of individual objects with MEG. Cereb Cortex 25:3602–3612.

Cohen JD, Daw N, Engelhardt B, Hasson U, Li K, Niv Y, Norman KA, Pillow J, Ramadge PJ, Turk-Browne NB, Willke TL (2017) Computational approaches to fMRI analysis. Nat Neurosci 20:304–313.

Collins JA, Olson IR (2014) Beyond the FFA: the role of the ventral anterior temporal lobes in face processing. Neuropsychologia 61:65–79.

Connolly AC, Guntupalli JS, Gors J, Hanke M, Halchenko YO, Wu Y-C, Abdi H, Haxby JV (2012) The representation of biological classes in the human brain. J Neurosci 32:2608–2618.

Cox DD, Dean T (2014) Neural networks and neuroscience-inspired computer vision. Curr Biol 24:R921–R929.

Downing PE, Jiang Y, Shuman M, Kanwisher N (2001) A cortical area selective for visual processing of the human body. Science 293:2470–2473.

Duchaine B, Yovel G (2015) A revised neural framework for face processing. Annu Rev Vis Sci 1:393–416.

Epstein R, Baker CI (2019) Scene perception in the human brain. Annu Rev Vis Sci 5:373–397.

Epstein R, Kanwisher N (1998) A cortical representation of the local visual environment. Nature 392:598–601.

Epstein R, Graham KS, Downing PE (2003) specific scene representations in human parahippocampal cortex. Neuron 37:865–876.

Fedorenko E, Hsieh P-J, Nieto-Castañón A, Whitfield-Gabrieli S, Kanwisher N (2010) New method for fMRI investigations of language: defining ROIs functionally in individual subjects. J Neurophysiol 104:1177–1194.

Friston KJ, Holmes AP, Worsley KJ, Poline JP, Frith CD, Frackowiak RSJ (1994) Statistical parametric maps in functional imaging: a general linear approach. Hum Brain Mapp 2:189–210.

Gorno-Tempini ML, Price CJ, Josephs O, Vandenberghe R, Cappa SF, Kapur N, Frackowiak RSJ (1998) The neural systems sustaining face and proper-name processing. Brain 121:2103–2118.

Griffith HR, Richardson E, Pyzalski RW, Bell B, Dow C, Hermann BP, Seidenberg M (2006) Memory for famous faces and the temporal pole: functional imaging findings in temporal lobe epilepsy. Epilepsy Behav 9:173–180.

Hanson SJ, Halchenko YO (2008) Brain reading using full brain support vector machines for object recognition: there is no "face" identification area. Neural Comput 20:486–503.

Haxby JV, Hoffman EA, Gobbini MI (2000) The distributed human neural system for face perception. Trends Cogn Sci 4:223–233.

Haxby JV, Gobbini MI, Furey ML, Ishai A, Schouten JL, Pietrini P (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293:2425–2430.

Haxby JV, Connolly AC, Guntupalli JS (2014) Decoding neural representational spaces using multivariate pattern analysis. Annu Rev Neurosci 37:435–456.

Henriksson L, Khaligh-Razavi S-M, Kay K, Kriegeskorte N (2015) Visual representations are dominated by intrinsic fluctuations correlated between areas. Neuroimage 114:275–286.

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12:55–67.

Huang J, Zhang T, Metaxas D (2011) Learning with structured sparsity. J Mach Learn Res 12:3371–3412.

Huth AG, de Heer WA, Griffiths TL, Theunissen FE, Gallant JL (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. Nature 532:453–458.

Ishai A (2008) Let's face it: it's a cortical network. Neuroimage 40:415–419.

Jacob L, Obozinski G, Vert J-P (2009) Group lasso with overlap and graph lasso. In: ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning, pp 433–440. New York: Association for Computing Machinery.

Jenatton R, Audibert J-Y, Bach F (2011) Structured variable selection with sparsity-inducing norms. J Mach Learn Res 12:2777–2824.

Jenatton R, Gramfort A, Michel V, Obozinski G, Eger E, Bach F, Thirion B (2012) Multiscale mining of fMRI data with hierarchical structured sparsity. SIAM J Imaging Sci 5:835–856.

Julian JB, Fedorenko E, Webster J, Kanwisher N (2012) An algorithmic method for functionally defining regions of interest in the ventral visual pathway. Neuroimage 60:2357–2364.

Kalénine S, Buxbaum LJ, Coslett HB, Kalénine S, Buxbaum LJ, Coslett HB (2010) Critical brain regions for action recognition: lesion symptom mapping in left hemisphere stroke. Brain 133:3269–3280.

Kanwisher N (2000) Domain specificity in face perception. Nat Neurosci 3:759–763.

Kanwisher N, Barton JJS (2011) The functional architecture of the face system: integrating evidence from fMRI and patient studies. Oxford: Oxford UP.

Kanwisher N, McDermott J, Chun MM (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. J Neurosci 17:4302–4311.

Khaligh-Razavi SM, Kriegeskorte N (2014) Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol 10:e1003915.

Kheradpisheh SR, Ghodrati M, Ganjtabesh M, Masquelier T (2016) Deep networks can resemble human feed-forward vision in invariant object recognition. Sci Rep 6:32672.

Kriegeskorte N (2015) Deep neural networks: a new framework for modeling biological vision and brain information processing. Annu Rev Vis Sci 1:417–446.

Kriegeskorte N, Kievit RA (2013) Representational geometry: integrating cognition, computation, and the brain. Trends Cogn Sci 17:401–412.

Kriegeskorte N, Goebel R, Bandettini PA (2006) Information-based functional brain mapping. Proc Natl Acad Sci U S A 103:3863–3868.

Kriegeskorte N, Formisano E, Sorger B, Goebel R (2007) Individual faces elicit distinct response patterns in human anterior temporal cortex. Proc Natl Acad Sci U S A 104:20600–20605.

Kriegeskorte N, Mur M, Bandettini PA (2008a) Representational similarity analysis - connecting the branches of systems neuroscience. Front Syst Neurosci 2:4.

Kriegeskorte N, Mur M, Ruff DA, Kiani R, Bodurka J, Esteky H, Tanaka K, Bandettini PA (2008b) Matching categorical object representations in inferior temporal cortex of man and monkey. Neuron 60:1126–1141.

Lambon Ralph MA, Jefferies E, Patterson K, Rogers TT (2017) The neural and computational bases of semantic cognition. Nat Rev Neurosci 18:42–55.

Lemm S, Blankertz B, Dickhaus T, Müller K-R (2011) Introduction to machine learning for brain imaging. Neuroimage 56:387–399.

Lewis-Peacock JA, Postle BR (2008) Temporary activation of long-term memory supports working memory. J Neurosci 28:8765–8771.

Lewis-Peacock JA, Postle BR, Cox CR, Rogers TT (2018) Long-term memory for famous faces, places, and common objects. OpenNeuro. 10.18112/openneuro.ds001497.v1.0.2.

Manning JR, Ranganath R, Norman KA, Blei DM (2014) Topographic factor analysis: a Bayesian model for inferring brain networks from neural data. PLoS One 9:e94914.

Marblestone AH, Wayne G, Kording KP (2016) Toward an integration of deep learning and neuroscience. Front Comput Neurosci 10:94.
Martin A (2007) The representation of object concepts in the brain. Annu Rev Psychol 58:25–45.
McClelland JL, Botvinick MM, Noelle DC, Plaut DC, Rogers TT, Seidenberg MS, Smith LB (2010) Letting structure emerge: connectionist and dynamical systems approaches to cognition. Trends Cogn Sci 14:348–356.
Mitchell TM, Shinkareva SV, Carlson A, Chang KM, Malave VL, Mason RA, Just MA (2008) Predicting human brain activity associated with the meanings of nouns. Science 320:1191–1195.
Nestor A, Vettel JM, Tarr MJ (2008) Task-specific codes for face recognition: how they shape the neural representation of features for detection and individuation. PLoS One 3:e3978.
Nestor A, Plaut DC, Behrmann M (2011) Unraveling the distributed neural code of facial identity through spatiotemporal pattern analysis. Proc Natl Acad Sci U S A 108:9998–10003.
Nestor A, Plaut DC, Behrmann M (2016) Feature-based face representations and image reconstruction from behavioral and neural data. Proc Natl Acad Sci U S A 113:416–421.
Olson IR, Plotzker A, Ezzyat Y (2007) The enigmatic temporal pole: a review of findings on social and emotional processing. Brain 130:1718–1731.
Oswal U, Cox CR, Lambon Ralph MA, Rogers TT, Nowak RD (2016) Representational similarity learning with application to brain networks. In: ICML'16 Proceedings of the 33rd International Conference on International Conference on Machine Learning, Vol 48 (Balcan MF, Weinberger KQ, eds), pp 1041–1049. New York: JMLR.org.
Pereira F, Botvinick MM (2011) Information mapping with pattern classifiers: a comparative study. Neuroimage 56:476–496.
Pereira F, Mitchell TM, Botvinick MM (2009) Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45 [1 Suppl]:S199–S209.
Pereira F, Lou B, Pritchett B, Ritter S, Gershman SJ, Kanwisher N, Botvinick M, Fedorenko E (2018) Toward a universal decoder of linguistic meaning from brain activation. Nat Commun 9:963.
Rabovsky M, Hansen SS, McClelland JL (2018) Modelling the N400 brain potential as change in a probabilistic representation of meaning. Nat Hum Behav 2:693–705.
Rajimehr R, Young JC, Tootell RB (2009) An anterior temporal face patch in human cortex, predicted by macaque maps. Proc Natl Acad Sci U S A 106:1995–2000.
Rao NS, Cox CR, Nowak RD, Rogers TT (2013) Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis. In: Advances in neural information processing systems, pp 2202–2210. La Jolla, CA: Neural Information Processing Systems Foundation.
Rao NS, Nowak RD, Cox CR, Rogers TT (2016) Classification with the sparse group lasso. IEEE Trans Signal Process 64:448–463.
Rish I, Cecchi GA, Heuton K, Baliki MN, Apkarian AV (2012) Sparse regression analysis of task-relevant information distribution in the brain. In: Proceedings Medical Imaging 2012: Image Processing (Haynor DR, Ourselin S, eds), p 8314. Bellingham, WA: SPIE.
Rivolta D, Woolgar A, Palermo R, Butko M, Schmalzl L, Williams MA (2014) Multi-voxel pattern analysis (MVPA) reveals abnormal fMRI activity in both the 'core' and 'extended' face network in congenital prosopagnosia. Front Hum Neurosci 8:925.
Rogers TT, McClelland JL (2014) Parallel Distributed Processing at 25: further explorations in the microstructure of cognition. Cogn Sci 38:1024–1077.
Rohde DLT (1999) LENS: The light efficient network simulator. Pittsburgh, PA: School of Computer Science, Carnegie Mellon University.
Rumelhart DE, McClelland JL (1986) Parallel distributed processing: explorations in the microstructure of cognition. Cambridge, MA: MIT.
Saxe AM, McClelland JL, Ganguli S (2019) A mathematical theory of semantic development in deep neural networks. Proc Natl Acad Sci U S A 116:11537–11546.
Sporns O (2011) Networks of the brain. Cambridge, MA: MIT.
Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Series B Stat Methodol 58:267–288.
Tsao DY, Moeller S, Freiwald WA (2008) Comparing face patch systems in macaques and humans. Proc Natl Acad Sci U S A 105:19514–19519.
Zhen Z, Fang H, Liu J (2013) The hierarchical brain network for face recognition. PLoS One 8:e59886.
Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodol 67:301–320.