Neuropsychologia 82 (2016) 134–141

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

A meta-analysis of fMRI decoding: Quantifying influences on human visual population codes

Marc N. Coutanche [a,*], Sarah H. Solomon [b], Sharon L. Thompson-Schill [b]

[a] Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA
[b] Department of Psychology, University of Pennsylvania, Philadelphia, PA, USA

ARTICLE INFO

Article history:
Received 14 September 2015
Received in revised form
13 January 2016
Accepted 15 January 2016
Available online 19 January 2016

Keywords:
Decoding
MVPA
Patterns
Meta-analysis
Vision
Objects

ABSTRACT

Information in the human visual system is encoded in the activity of distributed populations of neurons, which in turn is reflected in functional magnetic resonance imaging (fMRI) data. Over the last fifteen years, activity patterns underlying a variety of perceptual features and objects have been decoded from the brains of participants in fMRI scans. Through a novel multi-study meta-analysis, we have analyzed and modeled relations between decoding strength in the visual ventral stream, and stimulus and methodological variables that differ across studies. We report findings that suggest: (i) several organizational principles of the ventral stream, including a gradient of pattern granulation and an increasing abstraction of neural representations as one proceeds anteriorly; (ii) how methodological choices affect decoding strength. The data also show that studies with stronger decoding performance tend to be reported in higher-impact journals, by authors with a higher h-index. As well as revealing principles of regional processing, our results and approach can help investigators select from the thousands of design and analysis options in an empirical manner, to optimize future studies of fMRI decoding.

1. Introduction

Activity patterns in the human brain encode information that is processed during perception and cognition (Haxby et al., 2001; Tong and Pratte, 2012). These distributed multi-voxel patterns—recorded with functional magnetic resonance imaging (fMRI)—can be "decoded" to reveal the current processing target of a region's population of neurons. The most typical decoding approach uses machine learning techniques, which can be trained to successfully identify perceived images or cognitive states (O'Toole et al., 2007). This "multi-voxel pattern analysis" (MVPA) has given new insights into how information is organized in the brain (Tong and Pratte, 2012; Coutanche, 2013) and has relevance for clinical issues, such as the ability to track symptom severity (Coutanche et al., 2011).

Each individual fMRI study of visual population codes is affected by the particular properties of the employed stimuli, the study's design and the analysis approaches selected by the investigator. The role of methodological and analysis decisions are particularly important when using advanced analytical tools, where thousands of potential design and analysis combinations exist (Carp, 2012). Several investigations have empirically examined the impact of particular methodological options, including the number of fMRI acquisition runs (Coutanche and Thompson-Schill, 2012), type of classifier (Misaki et al., 2010) and trial sequences (Mumford et al., 2014). Others have examined how activity patterns in visual regions are modulated by certain stimulus properties, such as color (Parkes et al., 2009) and visual context (Troiani et al., 2014).

Within one study, it is possible to empirically examine the effect of stimulus or methodological variables that are explicitly manipulated. In contrast, through compiling and analyzing the outcomes of a large collection of studies, a cross-study meta-analysis has the potential to identify and quantify influences that are not evident from a single study alone. Previously, fMRI meta-analyses have been used to identify brain locations of consistent univariate activation (e.g., Bartra et al., 2013). In areas of behavioral science in which brain locations are not relevant, meta-analyses are used to statistically model how the size of effects relate to predictors (e.g., Hallion and Ruscio, 2011).

In this study, we perform the first meta-analysis of fMRI decoding strength. The presence of information in patterns of brain activity is frequently measured through the performance of machine learning classifiers (O'Toole et al., 2007) or through correlation values (Haxby et al., 2001). Here, we draw on decoding accuracy and pattern correlation values to ask how the strength of information extracted from activity patterns varies with a host of stimuli and methodological differences, across over one hundred investigations of perceptual decoding.

[*] Corresponding author.
E-mail address: marc.coutanche@pitt.edu (M.N. Coutanche).

http://dx.doi.org/10.1016/j.neuropsychologia.2016.01.018
0028-3932/© 2016 Elsevier Ltd. All rights reserved.

Combining studies is a key (and necessary) approach in meta-analyses. Studies will typically have many differences, such as the employed stimuli, fine methodological details, tasks, as well as variations that most investigators experience in their own studies, such as level of participant engagement and natural individual differences. Nonetheless, pooling studies has proven effective in the psychology literature, even with (or perhaps because of) variation in how their dependent variables (or "outcome measures") are produced. For instance, meta-analyses have collapsed across outcome measures from distinct tasks believed to tap one cognitive function (e.g., "visuospatial skill" from WAIS-R Block Design test, Copy subtest of the Rey-Osterrieth Complex Figure test and the Clock drawing task; Bäckman et al. (2005)), across distinct sensory modalities (by collapsing reading span and listening span to assess "verbal processing and storage"; Daneman and Merikle (1996)), and even across differing assessments of more subjective phenomena such as creativity (Baas et al., 2008). In this investigation, we will also be pooling different outcome measures. Here, we will pool outcome measures from studies that decode different conditions and stimuli, in order to understand common influences on decoding strength. Although different stimuli can vary significantly, our hope is that by collapsing across different comparisons - analogous to how outcomes from visual and auditory processing have been combined in prior meta-analyses - we can shed light on general organizational principles and influences on the ventral stream.

In this study, we employ a meta-analysis approach to examine issues such as pattern-granularity from V1 to later visual regions, and the degree of abstraction-from-exemplars as the ventral stream moves to anterior regions. We also identify the stimulus and methodological properties that are associated with stronger or weaker decoding across different visual regions.

2. Materials and methods

2.1. Study selection

We analyzed empirical investigations of visual decoding in the human ventral stream. We identified candidate studies in two ways to maximize our chance of collecting all relevant results. We first searched for peer-reviewed publications using search terms: "fMRI" and at least one of "MVPA", "decoding", "pattern analysis", "multivoxel", "multi-voxel", or "classification" in online databases Scopus, PubMed, Web of Science, and Google Scholar. The search in Google Scholar was restricted to articles' titles due to the large number of irrelevant studies that is otherwise returned. Second, we compiled all publications citing a seminal study by Haxby et al. (2001). Combining the publications identified through both methods resulted in 3920 unique entries. We restricted this list to empirical work published in peer-reviewed journals; eliminating conference proceedings, dissertations, book chapters, review articles, and non peer-reviewed letters (Fig. 1).

From the resulting peer-reviewed empirical papers, we focused our investigation on fMRI decoding studies of perception in the ventral stream, based on the following characteristics:.

(1) Research subjects. We included papers that recruited healthy adult human participants: eliminating papers studying non-human animals, and those exclusively investigating patients or children. If a clinical or developmental paper also examined healthy adults (e.g., as a comparison group), we included the data from just those healthy adult subjects.

(2) Research design. We focused our analyses on fMRI data collected from occipital and temporal cortex. We did not include classifications based on whole-brain data, which can be influenced by non-occipital/temporal regions. The extracted results were collected during visual presentations, including different visual categories (e.g., objects, faces, etc.) and basic visual properties (e.g., line orientation). We did not include analyses of words or instances where behavioral responses (rather than the identity of the visual stimulus) were being predicted.

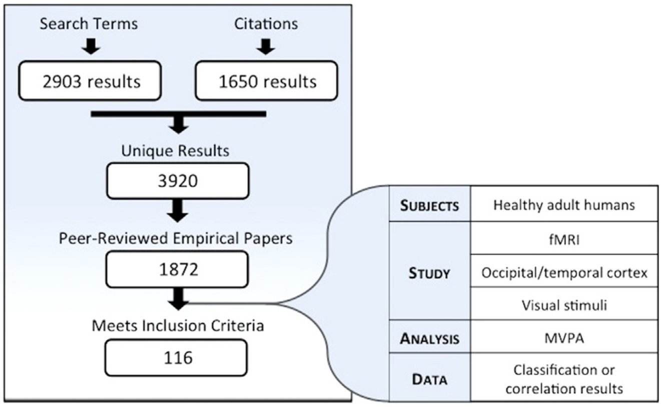


Fig. 1. A flow-chart of the paper identification and selection process. See also Supplementary Table 1 for the complete list of included papers.

(3) Format of reported data. In order to combine decoding results across diverse analysis approaches, it was necessary to focus on papers reporting results in terms of: (i) classification accuracy; (ii) correlation values (a standard approach to comparing activity patterns for different conditions). This removed a subset of papers reporting results in alternative formats, such as d-prime, and papers that did not directly report correlation/classification values (e.g., reporting only *t*-values or difference-scores without the underlying values).

(4) The included results were conducted on 3T scanners, removing any possible confounding influence of magnet strength. There were too few examples of non-3T scans in relevant papers to investigate this as a meta-analysis predictor.

2.2. Meta-analysis variables

For each included paper, we extracted the study's methodological decisions and stimulus properties (Table 1).

To examine how the effects of these variables on decoding strength might vary along the ventral stream, we separately collected results from data in V1–V5, fusiform face area (FFA), lateral occipital complex (LOC), parahippocampal place area (PPA), fusiform gyrus, anterior temporal lobe (ATL) and ventral temporal (VT) cortex as a whole. As the area of "ventral temporal" cortex has no agreed-upon definition, for present purposes we included studies applying an approximate anatomical definition, or defining visually responsive voxels in this area of temporal cortex. Due to a low numbers of data-points for ATL, FG and V5, we could not analyze these regions and removed five papers that only analyzed these areas.

2.3. Dependent variables

The most common formats of results from MVPA studies are classification performance and correlation values. When not clearly included in studies' text or tables, we extracted relevant classification and correlation values from graphs or from color-coded matrices with scales. We extracted data in this fashion, rather than contacting authors, to avoid a potential human-induced selection bias in which values were, and were not, included in analyses. Classification accuracies are only informative in relation

Table 1
The methodological and stimulus variables extracted from included studies, with a description of how each variable was coded.

Extracted variable	Coding scheme
Voxel size	Voxel size at recording
Block vs. event design	Binary variable of whether the study was blocked or featured events
Number of runs	Number of runs used in the included analysis
Stimulus time	Calculated duration of the data for each class (e.g., sixty seconds if sixty 1-second stimuli were presented in a classified condition)
Spatial smoothing (mm)	Amount of spatial smoothing applied (in studies that reported more than one smoothing level, the best performing results were included)
Degree of trial averaging	Percentage of a condition's available data that was averaged to create one time-point for the decoding analysis (e.g., if the classified data-points are block-averages, with 10 blocks, this would be 10%)
Percentage of training data	Percentage of data used for training or as input in a correlation analysis
Classification across vs. within runs	Binary variable of whether trials were classified across runs or from the same run (e.g., leave-one-run-out would be across-runs)
SVM vs. correlation classifier	When a classifier is used, this binary variable compares the two most popular variants
Voxel count	Mean number of voxels used in analyses
Task	The tasks employed were coded into five categories based on depth of stimulus processing: 1=fixation change; 2=passive viewing; 3=monitoring stimulus change (e.g., size); 4=working memory (e.g., 1-back); 5=semantic meaning (e.g., naming)
Photograph vs. rendered	Binary variable of whether presented images were photographs or rendered (computer-generated stimuli and line-drawings)
Color vs. grayscale	Binary variable of whether color was present in the stimuli
Cluttered vs. isolated	Binary variable of the context of presented stimuli: "cluttered" includes instances of additional items on-screen or images with a background
Number of exemplars per class	Number of unique items (exemplars) in each class

to chance performance. To combine results across studies (where chance performance differs), we created a common decoding metric:

$$\text{decoding strength} = \frac{\text{accuracy} - \text{chance}}{1 - \text{chance}}$$

.

This formula gives the degree to which a classification accuracy value surpasses chance (numerator), while being bound by the potential range of classification values (denominator). This scaling by the denominator is necessary to account for differing ranges of available performance (e.g., when chance is 50%, 50% points are available to reach 100%, whereas when chance is 20%, 80% points are available). This approach normalizes accuracy values by the size of the interval of possible values. A score of zero then reflects no information (classification performance=chance) and a score of 1.0 reflects the maximum available performance. We then *z*-scored the resulting full set of values.

Although this metric has a number of advantages, one possible concern is whether differences in accuracy-to-chance ratios are sufficiently incorporated (e.g., 50% accuracy is 2 times chance for 4 classes, but 4 times chance for 8 classes). A Spearman's rank correlation of the metric's values with the corresponding accuracy-to-chance ratios extracted from the same data showed that both are extremely closely related ($p=1.59\times10^{-28}$), giving confidence that information contained in the accuracy-to-chance ratio is being tracked by the above metric. The regression results generated from each metric were also strongly related ($p<0.0001$).

To analyze correlation values from studies, we first ensured that *r*-values were Fisher-transformed to *z*-values, and then created "discrimination" values ($1-z$), so that higher values indicate greater dissimilarity between patterns (i.e., they are more discriminable). The full set of discrimination values were then *z*-scored, to give a mean of zero. We removed one paper with an extremely high value on the resulting metric (due to a unique methodological choice), resulting in 110 papers within our final corpus of studies.

To quantitatively examine how the extracted variables in Table 1 relate to decoding strength, we used linear regressions to predict classifier performance and pattern discrimination for each region. We could then examine coefficient weights to understand the relations between the variables and decoding strength. All continuous predicting variables were centered on zero and—in order to prevent unreliable conclusions—we have indicated when certain ROI-and-predictor combinations lacked sufficient data-points to reliably draw conclusions. To be included, binary variables were required to have a minimum of five studies in each binary option, and continuous variables were required to have at least five contributing investigations.

3. Results

We examined how decoding visual multi-voxel patterns in the ventral stream is influenced by stimulus properties and methodological decisions. To do this, we modeled classification performance and pattern correlations from over one-hundred studies in linear regressions. The resulting coefficients reflect the relationship between each variable and the ability to extract information from visual regions through MVPA within the examined studies. Positive coefficients reflect positive relations between predictors and decoding, while negative coefficients indicate inverse relations.

We first examined two principles of organization of the ventral stream. The first concerns pattern granularity. A current question among investigators of the early visual system regards the granulation of information-carrying patterns within early visual cortex (Kamitani and Sawahata, 2010; Op de Beeck, 2010; Freeman et al., 2011). One approach taken by individual studies to examine this issue has been applying varying levels of spatial smoothing and measuring the effect on decoding (Op de Beeck, 2010), although this approach has also been criticized (Kamitani and Sawahata, 2010). With a meta-analysis approach, we can instead examine imaging acquisition parameters that are varied *across* studies. One such parameter is the spatial resolution (i.e., voxel size) of the acquired functional data. Multi-voxel decoding should be optimized (all else being equal) when the voxel size of acquired data matches the spatial resolution (i.e., granularity) of a region's information-containing patterns. We hypothesized that if V1 holds a more fine-grained map of information than later visual regions, employing larger voxels should not benefit decoding in V1, but may benefit decoding in post-V1 regions (through greater signal-to-noise at the scale of these patterns). The results of our regression analyses supported this: using larger voxels improved decoding in V2 ($B=0.01$, $p=0.049$), unlike V1 ($B=0.002$, $p=0.451$). The decoding boost was at least 2.5 times greater in post-V1

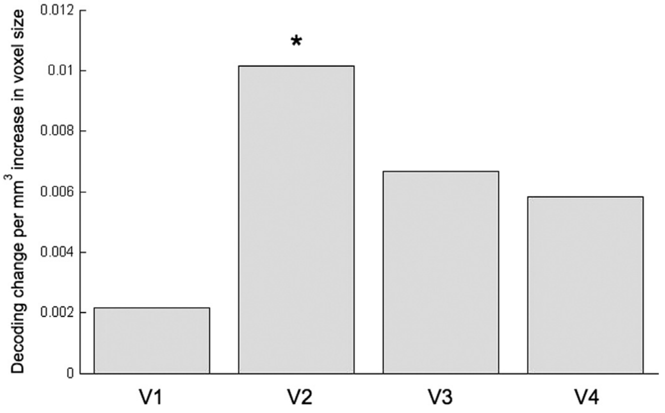


Fig. 2. Modeled influence of voxel size on decoding strength. Predicted changes in decoding strength are shown for every 1 mm^3 added to the size of acquired voxels. Positive values indicate greater decoding strength for each visual region.

regions than in V1 (Fig. 2). These findings are consistent with later visual regions holding coarser multi-voxel codes, while V1 relies on fine-grained patterns. The voxel-size benefit was most apparent in V2. Although we can only speculate at this point, this may be due to the same organizational principle that leads V2 to have the strongest level of response-adaptation (which is also at the voxel level) of the early visual cortical areas (Sapountzis et al., 2010).

Naturally, at a certain point, increasing the voxel size is expected to impair performance for any region. The upper end of the voxel sizes in these studies were small by fMRI standards (with almost half at 15.6 mm^3 or smaller; equivalent to 2.5 mm isotropic), but we also examined whether employing larger voxels begins to reduce performance, by testing if a quadratic term (e.g., an inverted-U) would better fit the data. Within the range of voxel sizes in these studies, a quadratic term did not improve the model for the visual regions (all $p > 0.77$ in chi-square comparisons of model fits), although we stress that this result may not apply for voxel sizes beyond this range.

We next used the multi-study data to examine the robustness of visual population codes to different exemplars of categories. The start of the ventral stream (V1) is retinotopically organized. With this kind of organization, any change in visual stimulation will significantly change activity. Accordingly, early visual cortex decoding should be weaker when a trained classifier is tested on data collected from viewing visually different stimuli (e.g., training to distinguish chairs from tables and being tested on a new visually distinct chair). If a key principle of the ventral stream is for information to become more categorical and abstracted from basic visual features (from posterior to anterior regions; Coutanche and Thompson-Schill (2015)), then classifiers in more anterior regions should be affected less by visually different stimuli within a category. We tested this prediction by modeling decoding strength based on the number of exemplars that were included in each class. This analysis revealed a strikingly linear relation between the posterior-to-anterior progression of the ventral stream, and the robustness of trained classifiers to data from new exemplars (r=0.94; $p < 0.001$; Fig. 3). In other words, patterns become more generalizable as the ventral stream progresses.

Analyzing the full set of variables allows us to examine both common and region-specific relations with decoding strength. The regression coefficients (i.e., relation strengths) for classification are shown in Fig. 4. A coefficient's value reflects a relation within our set of studies, so that a positive coefficient indicates a positive association between decoding strength and the variable's values employed within studies of the region.

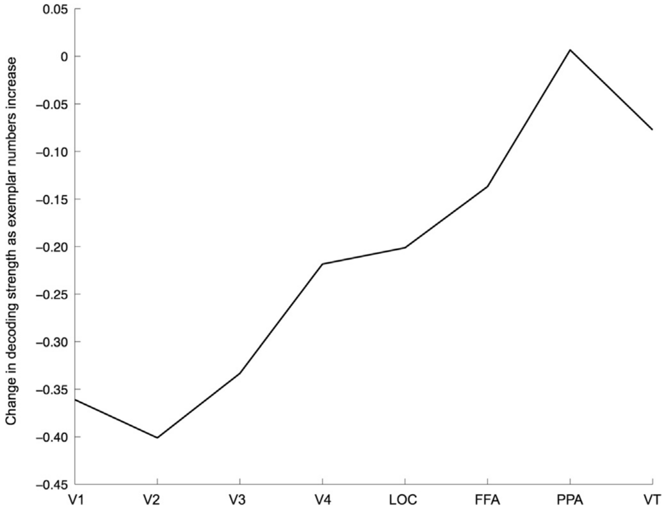


Fig. 3. Change in decoding strength with increasing exemplar counts in each class. The *y*-axis shows the regression coefficient from a model predicting classification performance based on the number of exemplars within classified categories. Positive values indicate a positive relation between classification performance and class exemplars (i.e., exemplar robustness). Negative values reflect inverse relations.

Particular results are placed in the context of the literature and interpreted in the Discussion (Section 4), however particular patterns of results give confidence in our method. For example, regions with a retinotopic organization are expected to show improved classification with increasing voxel counts, as including more voxels provides access to more of the visual field (unlike, for example, placing a small sphere at the calcarine sulcus). On the other hand, less retinotopic regions should not show this benefit, and may even show reduced performance due to overfitting from greater numbers of features. We observe exactly this (Fig. 4): all early visual regions showed improved decoding with more voxels, in addition to LOC (which is known to contain retinotopic information; Larsson and Heeger (2006)). Equally, we note that the proportion of data used for training positively predicts classification performance across early visual regions, as would be predicted for models trained with more data (Coutanche and Thompson-Schill, 2012; O'Toole et al., 2007). In order to further view which factors play important roles (and in which direction), we have collapsed across early and late visual regions to display the size and direction of each relationship in Fig. 5. Similarly, Table 2 presents some specific stimulus and methodological choices that we observed as associated with strong decoding. We wish to emphasize that the best choice for a variable in any particular study is affected by a number of factors (including other variables in Table 2), so this table should be considered a guide to aid study design, rather than a prescription. Relatedly, we can only speak to the range of values used in the investigated studies: even stronger decoding might be observed for as-yet untested options (e.g., new behavioral tasks or voxel sizes).

For pattern correlation discriminability values (Fig. 6), voxel counts were again important for V1 and the LOC region, unlike a later area, which supports this finding with an alternative metric of pattern strength and different data. Greater spatial smoothing was associated with weaker discriminability in the FFA. Additionally, "deeper" stimulus processing in the scanner task (e.g., memory retrieval rather than visual fixation) positively predicted discriminability in the ventral temporal region.

Having collated methodological and stimulus parameters, we took the opportunity to examine patterns of co-occurrence between different stimulus and methodological variables in MVPA studies. By treating these as features, we cross-correlated papers

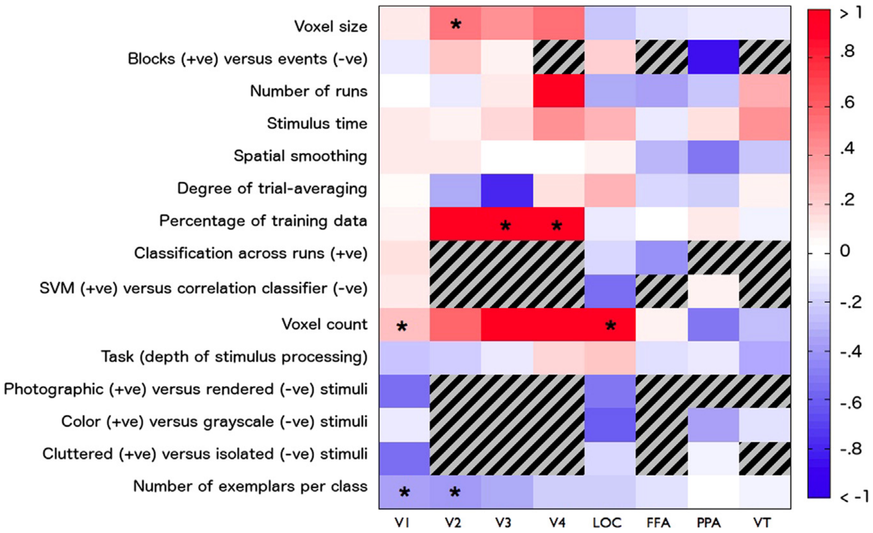


Fig. 4. Influences on classification performance. Colors reflect standardized coefficients from a regression predicting each region's decoding strength. Red indicates that higher values of each variable predict greater decoding. Blue indicates that lower values predict greater decoding. White reflects an absence of a relation. For binary variables, a value of 1 was assigned to the first option listed in the *y*-axis. Predictors with too few data-points for an ROI are striped-out. Asterisks indicate relations that generalize beyond the current sample of papers ($p < 0.05$). (Readers of the print version are referred to the web version for color coding.)

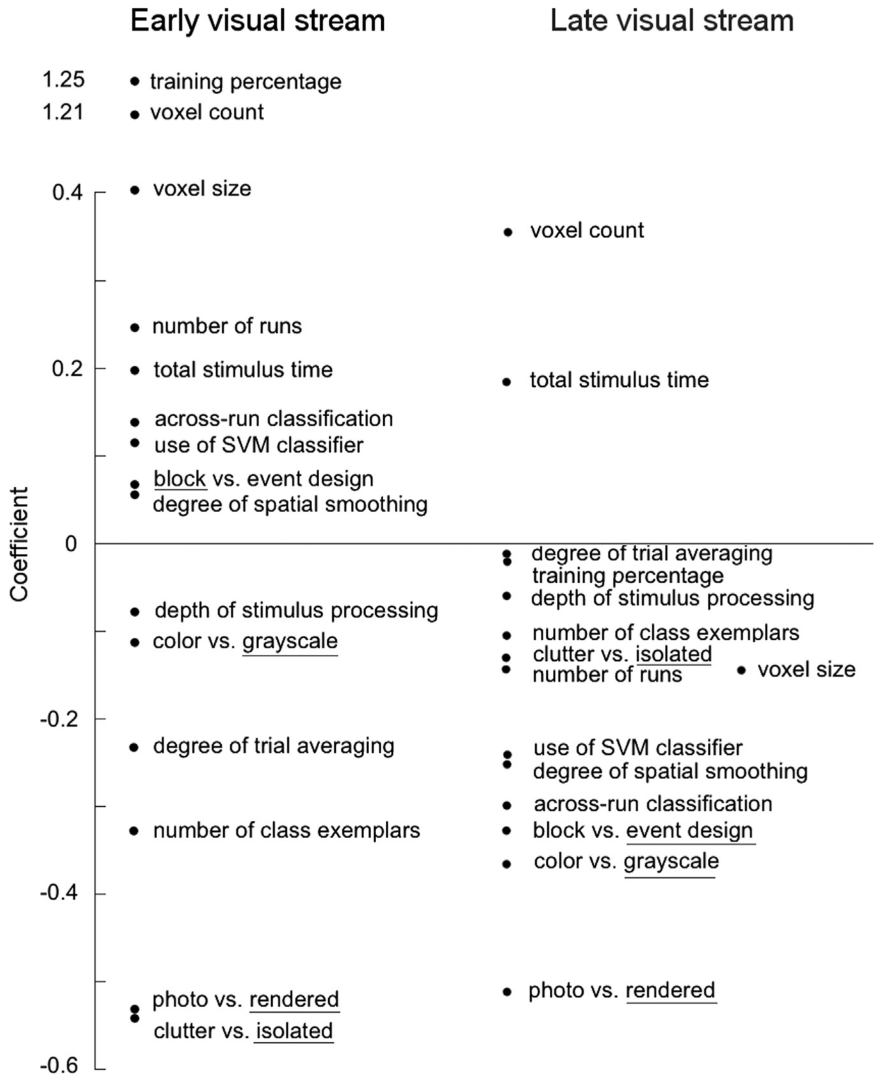


Fig. 5. Variables ordered by the strength and direction of their relationship with decoding performance. Variables at the top and bottom have the greatest influence (in a positive and negative direction respectively). For ease of viewing, V1, V2, V3 and V4 have been collapsed, as have LOC, FFA, PPA and VT regions.

Table 2
Values of each variable that are associated with maximal decoding performance in the analyzed studies. V1, V2, V3 and V4 have been collapsed into early visual cortex (EVC), as have LOC, FFA, PPA and VT regions (VT). When the strongest performance is associated with the lowest or highest values, “min” or “max” is listed respectively. If two separate values show strong performance, both are given. The “task depth of processing” lists which of the five task categories is associated with the strongest decoding (Methods). Task 2 is passive viewing. Task 3 is monitoring for a stimulus change (e.g., color). The dash indicates that the optimum value was unclear, due to variability across the collapsed regions (for spatial smoothing) or because both binary options give performance similar enough that selecting one might be arbitrary.

	EVC	**VT**
Voxel size	6.5 and 33.5 mm^3	11.4 mm^3
Blocks vs. events	–	–
Run count	20	17
Total stimulus time	1141 s	max
Spatial smoothing (full width half-maximum)	–	–
Degree of trial averaging	38%	37% and max
Training data percentage	69% and max	53%
Across vs. within run classification	–	within
SVM vs. correlation classifier	SVM	correlation classifier
Voxel count	max	1706
Task depth of processing	2/5	3/5
Photographic vs. rendered	rendered	rendered
Color vs. grayscale	grayscale	grayscale
Cluttered vs. isolated	isolated	isolated
Number of exemplars per class	min	min and 204

based on their unique set of stimulus / methodological properties (Fig. 7A). The results showed that investigations of early visual cortex are methodologically similar to each other (mean correlation=0.82, s.d.=0.30), while ventral temporal studies are relatively more diverse (mean correlation=0.72, s.d.=0.27; Wilcoxon rank sum test of difference: z=9.51, $p<0.001$; Fig. 7B). We confirmed that this is not an artifact of early visual cortex investigations originating from a smaller proportion of unique labs (based on whether papers share a last author; χ^2=0.01, $p>0.9$).

As a final exploration of variables that might predict decoding strength, we asked if meta-data about the studies could predict decoding strength (collapsing across regions). Specifically, we hypothesized that studies with stronger results might be conducted by authors with more experience in selecting fMRI parameters and analyses. A common metric for an author's publication experience is the h-index, which combines an author's productivity with the influence of their published work. We modeled decoding performance based on the h-index of the first author, in addition to the last author for comparison purposes. The h-index of the studies' first author was highly predictive of decoding performance (R^2=0.19, p=0.003): first authors with more influential work publish studies with higher levels of decoding. In contrast, the last author's h-index did not predict decoding (R^2=0.007, p=0.56). We note, however, that we cannot identify a direction from this relationship. Greater decoding performance in a study could equally lead to the paper becoming cited more frequently, thus increasing the first author's h-index. Our second hypothesis for the paper meta-data was that studies with stronger decoding are more likely to be published in more prestigious journals, either due to self-selection by the submitting authors, or requirements of these journals to publish striking results. The publishing journal's Impact Factor (reflecting the citation rate of articles) was also predictive of decoding strength (r^2=0.06, p=0.02): Journals with larger impact factors publish papers that report stronger decoding.

4. Discussion

In this investigation, we extracted and analyzed results from 110 MVPA studies of visual population codes through a cross-study meta-analysis. We modeled pattern decoding strength based on the stimulus and methodological properties of each study to discover relations between these factors and decoding. To probe the organizational principles of the ventral stream, we first analyzed voxel size, finding that multi-voxel pattern codes are coarser (less granulated) in regions downstream from V1. In a second analysis,

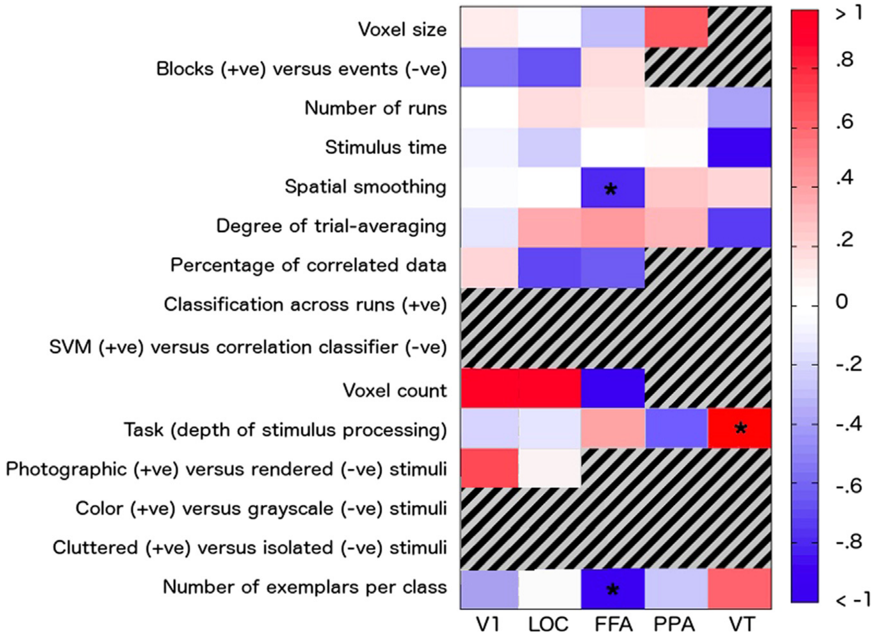
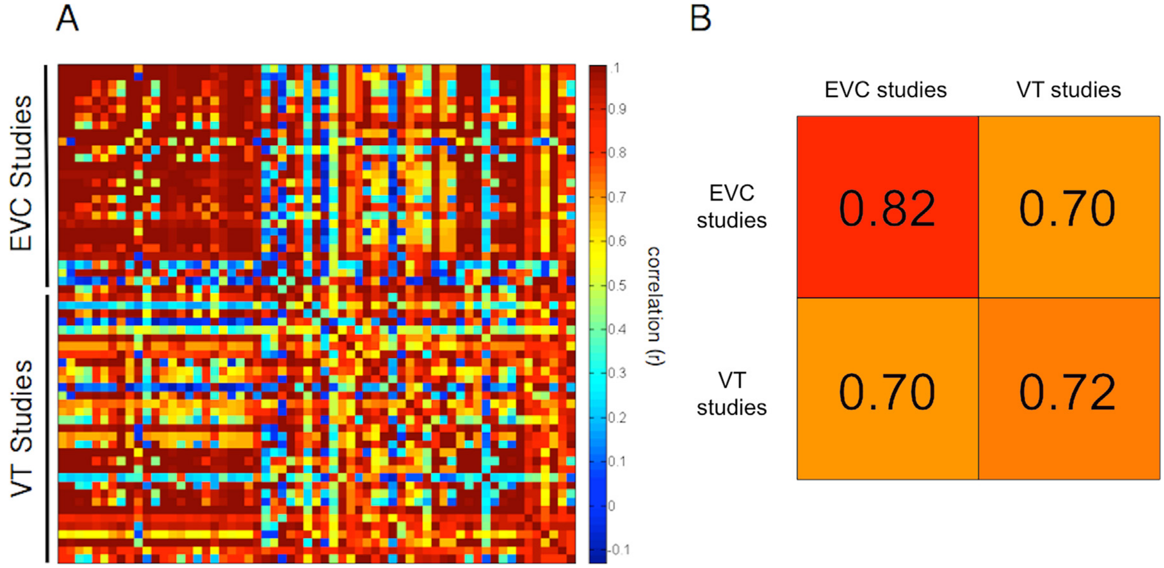


Fig. 6. Influences on pattern correlations. Colors reflect standardized coefficients from a regression predicting each region's pattern discrimination based on correlations (i.e., $1-r$). Red indicates that higher values predict greater pattern discriminability. Blue indicates that lower values predict greater discriminability. White reflects an absence of a relation. For binary variables, a value of 1 was assigned to the first option listed in the y-axis. Predictors that have too few data-points, and that are irrelevant for correlations (e.g., classifier type) are striped-out. Asterisks indicate relations that generalize beyond the current sample of papers ($p<0.05$). (Readers of the print version are referred to the web version for color coding.)



	EVC studies	VT studies
EVC studies	0.82	0.70
VT studies	0.70	0.72

Fig. 7. The methodological space of the study set. A: The similarity matrix shows pairwise Pearson correlations between investigations based on their methodological decisions, grouped by investigated region. B: Mean correlation values extracted from the quadrants of panel A.

we found that activity patterns become increasingly generalizable/robust to different exemplars as the ventral stream progresses anteriorly. This finding is consistent with a general organizational principle of greater abstraction from visual features to more abstract concepts (Coutanche and Thompson-Schill, 2015). We also identified stimulus and methodological factors that co-occur with stronger decoding (discussed further below). Finally, we found that studies with stronger decoding are more likely to be first-authored by scientists with "larger" publication records, and published in more impactful journals.

The VT organizational principles we report—greater granulation and abstraction over exemplars—concern two distinct, but combinable, properties. Combining these principles leads to a prediction that pattern granularity should be related to the categorical level of its representation. The results of a recent study are consistent with this: spatial smoothing has been used to argue that patterns can contain multiple scales of organization at differing levels of granularity, where coarser patterns are more closely linked to the representations of broader (less specific) categories (Brants et al., 2011).

For the methodological variables that show a relation with decoding, the importance of the amount of training data has been previously described (e.g., O'Toole et al., 2007), and acts as a useful sense-check of the meta-analysis method. The number of included voxels was found to be a particularly strong predictor of decoding performance in V1 and LOC. V1 is organized retinotopically, so that including more voxels is likely to better ensure that the entire visual field is available to a classifier. The LOC—commonly associated with shape and object processing—is composed of multiple sub-regions that differ in the character of their encoded information (Drucker and Aguirre, 2009). This can explain why voxel count is particularly important for this region, and LOC investigators may wish to take special care to include all its sub-regions if they wish to maximize the region's decoding potential. At first glance, our finding that greater decoding is associated with both voxel-count and size might appear contradictory. Although these variables are linked within one study, however, they can have different relationships *across* studies. In this instance, higher voxel counts may benefit decoding for the reasons discussed above, while increasing voxel size can give similar benefits to those observed from a small amount of spatial smoothing (Op de Beeck et al., 2008a, 2008b).

For influences on pattern correlations, greater spatial smoothing was particularly associated with decreasing FFA discriminability. Interestingly, this particular region has been the focus of a number of studies with conflicting results. Some MVPA investigations of the FFA have failed to decode facial identity (e.g., Kriegeskorte et al., 2007), while others have succeeded (e.g., Axelrod and Yovel, 2015). The identified importance of spatial smoothing here may provide a clue to the reason behind this variability. Future investigations that obtain successful decoding may wish to examine how spatial smoothing (or voxel size) impacts facial identity discrimination. Our second FFA-related finding is that the region's discriminability is particularly vulnerable to adding exemplars, which suggests that the number and visual differences of discriminated facial identities might also play a key role in the mixed findings in the literature. The meta-analysis results for correlations also suggest that pattern discriminability in the overall VT region improves as depth of stimulus processing increases. This is in line with recent findings that ventral temporal cortex encodes semantic information as well as basic visual information (e.g., Carlson et al., 2014). A task that triggers neuronal populations underlying meaning may contribute to this improved discriminability.

Finally, our finding that first-author researchers with greater *h*-indices have studies with stronger decoding may reflect an informal qualitative "meta-analysis" that fMRI researchers employ as we draw on our own experiences and expertize to select stimulus and methodological properties. More experienced investigators will often have a greater awareness of the optimal choices in a study's design and analysis stages. Alternatively, the relationship may reflect that larger stimuli differences were examined by (now, more senior) investigators when MVPA was first developed, compared to subtler stimuli differences being investigated more recently by (still junior) investigators. A further alternative is that papers with greater decoding performance will be cited more often, which in turn would increase an author's *h*-index. This would, however, need to account for the presence of a relationship with the studies' first, but not last, author. The second meta-study predictor of decoding strength—the journal's Impact Factor—might similarly reflect a tendency for high impact journals to publish studies that decode basic differences for the first time, but not subtler follow-up comparisons.

In our meta-analysis approach, we have used metrics of decoding strength from particular regions in order to examine regional specificity. In contrast, typical GLM meta-analyses examine regional specificity by spatially plotting the foci of significant GLM results on a standardized brain's voxels (e.g., Bartra et al., 2013). The machine learning classifiers often employed in MVPA themselves usually have a set of voxel weights as part of their training. It is, however, worth noting that it would not be appropriate to combine voxel maps of classifier weights in the same way. As others have described (Haufe et al., 2014), although the weights of "forward models" (such as the GLM) are interpretable, weight values from "backward models" (such as multivariate classifiers) do not always reflect a voxel's contribution to a brain state. Thus, without appropriate conversions, classifier weight maps should not be combined. One area that our general approach might be extendable to in the future is "encoding" analyses of fMRI data (Naselaris and Kay, 2015), although a common or convertible dependent variable will be needed across investigations.

The approach we have employed has limitations that are important to acknowledge and consider. First, one of the method's greatest strengths is also a weakness: Generalizing across the results of decoding different stimuli and conditions allows us to examine general principles, but also smooths over differences across comparisons. For example, the categorical levels of different comparisons may influence the particular spatial scale being tapped, among other properties. More narrowly defined sets of classification studies can be examined in the future provided that sufficient numbers of studies are available.

We hope that the relations we have identified between methodological choices and decoding strength can be referenced by investigators as they seek to make design and analysis choices from the thousands of combinations available (Carp, 2012). More generally, this empirical approach has the potential to support the field's efforts toward replicability and standardized pipelines of analysis, as well as helping to understand the design and analysis approaches that are associated with optimal fMRI decoding.

Acknowledgments

The work in this paper was supported by a grant from the National Institutes of Health (R01EY021717). The first author was supported by a National Research Service Award from the National Eye Institute of the National Institutes of Health (F32EY024851).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at http://dx.doi.org/10.1016/j.neuropsychologia.2016.01.018.

References

Axelrod, V., Yovel, G., 2015. Successful decoding of famous faces in the fusiform face area. PLOS One 10, e0117126.

Baas, M., De Dreu, C.K.W., Nijstad, B.A., 2008. A meta-analysis of 25 years of mood-creativity research: hedonic tone, activation, or regulatory focus? Psychol. Bull. 134 (6), 779–806.

Bäckman, L., Jones, S., Berger, A.-K., Laukka, E.J., Small, B.J., 2005. Cognitive impairment in preclinical Alzheimer's disease: a meta-analysis. Neuropsychology 19 (4), 520–531.

Bartra, O., McGuire, J.T., Kable, J.W., 2013. The valuation system: a coordinate-based meta-analysis of BOLD fMRI experiments examining neural correlates of subjective value. NeuroImage 76, 412–427.

Brants, M., Baeck, A., Wagemans, J., Op de Beeck, H.P., 2011. Multiple scales of organization for object selectivity in ventral visual cortex. NeuroImage 56, 1372–1381.

Carlson, T.A., Simmons, R.A., Kriegeskorte, N., Slevc, L.R., 2014. The emergence of semantic meaning in the ventral temporal pathway. J. Cogn. Neurosci. 26, 120–131.

Carp, J., 2012. On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. Front. Neurosci. 6, 149.

Coutanche, M.N., 2013. Distinguishing multi-voxel patterns and mean activation: why, how, and what does it tell us? Cogn. Affect. Behav. Neurosci. 13, 667–673.

Coutanche, M.N., Thompson-Schill, S.L., 2015. Creating concepts from converging features in human cortex. Cereb. Cortex 25 (9), 2584–2593.

Coutanche, M.N., Thompson-Schill, S.L., 2012. The advantage of brief fMRI acquisition runs for multi-voxel pattern detection across runs. NeuroImage 61, 1113–1119.

Coutanche, M.N., Thompson-Schill, S.L., Schultz, R.T., 2011. Multi-voxel pattern analysis of fMRI data predicts clinical symptom severity. NeuroImage 57, 113–123.

Daneman, M., Merikle, P.M., 1996. Working memory and language comprehension: a meta-analysis. Psychon. Bulletin & Rev. 3 (4), 422–433.

Drucker, D.M., Aguirre, G.K., 2009. Different spatial scales of shape similarity representation in lateral and ventral LOC. Cereb. Cortex 19, 2269–2280.

Freeman, J., Brouwer, G.J., Heeger, D.J., Merriam, E.P., 2011. Orientation decoding depends on maps, not columns. J. Neurosci. 31, 4792–4804.

Hallion, L.S., Ruscio, A.M., 2011. A meta-analysis of the effect of cognitive bias modification on anxiety and depression. Psychol. Bull. 137, 940–958.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., et al., 2014. On the interpretation of weight vectors of linear models in multivariate neuroimaging. NeuroImage 87, 96–110.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

Kamitani, Y., Sawahata, Y., 2010. Spatial smoothing hurts localization but not information: pitfalls for brain mappers. NeuroImage 49, 1949–1952.

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

Larsson, J., Heeger, D.J., 2006. Two retinotopic visual areas in human lateral occipital cortex. J. Neurosci. 26 (51), 13128–13142.

Misaki, M., Kim, Y., Bandettini, P.A., Kriegeskorte, N., 2010. Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. NeuroImage 53, 103–118.

Mumford, J.A., Davis, T., Poldrack, R.A., 2014. The impact of study design on pattern estimation for single-trial multivariate pattern analysis. NeuroImage 103, 130–138.

Naselaris, T., Kay, K.N., 2015. Resolving ambiguities of MVPA using explicit models of representation. Trends Cogn. Sci. 19 (10), 551–554.

Op de Beeck, H.P., 2010. Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? NeuroImage 49, 1943–1948.

Op de Beeck, H.P., Deutsch, J.A., Vanduffel, W., Kanwisher, N.G., DiCarlo, J.J., 2008a. A stable topography of selectivity for unfamiliar shape classes in monkey inferior temporal cortex. Cereb. Cortex 18 (7), 1676–1694.

Op de Beeck, H.P., Torfs, K., Wagemans, J., 2008b. Perceived shape similarity among unfamiliar objects and the organization of the human object vision pathway. J. Neurosci. 28 (40), 10111–10123.

O'Toole, A.J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J.P., Parent, M.A., 2007. Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. J. Cogn. Neurosci. 19, 1735–1752.

Parkes, L.M., Marsman, J.-B.C., Oxley, D.C., Goulermas, J.Y., Wuerger, S.M., 2009. Multivoxel fMRI analysis of color tuning in human primary visual cortex. J. Vis. 9 (1), 1–13.

Sapountzis, P., Schluppeck, D., Bowtell, R., Peirce, J.W., 2010. A comparison of fMRI adaptation and multivariate pattern classification analysis in visual cortex. NeuroImage 49 (2), 1632–1640.

Tong, F., Pratte, M.S., 2012. Decoding patterns of human brain activity. Annu. Rev. Psychol. 63, 483–509.

Troiani, V., Stigliani, A., Smith, M.E., Epstein, R.A., 2014. Multiple object properties drive scene-selective regions. Cereb. Cortex 24, 883–897.