# Atlases of cognition with large-scale brain mapping

Gaël Varoquaux, Yannick Schwartz, Russell Poldrack, Baptiste Gauthier, Danilo Bzdok, Jean-Baptiste Poline, Bertrand Thirion

# Atlases of cognition with large-scale brain mapping

Gaël Varoquaux[1,2,3,¤,☯] Yannick Schwartz[1,2,3,☯] Russell A. Poldrack[4], Baptiste Gauthier[5,2], Danilo Bzdok[1,2,6,7], Jean-Baptiste Poline[8], Bertrand Thirion[1,2,3,*]

**1** Parietal, Inria, Saclay, France
**2** Neurospin, CEA, Saclay, France
**3** Université Paris-Saclay
**4** Stanford University, Stanford, CA 94305
**5** Cognitive Neuroimaging Unit, INSERM, Gif sur Yvette, France
**6** JARA-BRAIN, Jülich-Aachen Research Alliance, Germany
**7** Department of Psychiatry, Psychotherapy and Psychosomatics, RWTH Aachen University, 52072 Aachen, Germany
**8** McGill University, Montreal, Canada

☯These authors contributed equally to this work.
¤Current Address: Neurospin, CEA Saclay, 91191 Gif sur Yvette, France
* bertrand.thirion@inria.fr

## Abstract

To map the neural substrate of mental function, cognitive neuroimaging relies on controlled psychological manipulations that engage brain systems associated with specific cognitive processes. In order to build comprehensive atlases of cognitive function in the brain, it must assemble maps for many different cognitive processes, which often evoke overlapping patterns of activation. Such data aggregation faces contrasting goals: on the one hand finding correspondences across vastly different cognitive experiments, while on the other hand precisely describing the function of any given brain region. Here we introduce a new analysis framework that tackles these difficulties and thereby enables the generation of brain atlases for cognitive function. The approach leverages ontologies of cognitive concepts and multi-label brain decoding to map the neural substrate of these concepts. We demonstrate the approach by building an atlas of functional brain organization based on 30 diverse functional neuroimaging studies, totaling 196 different experimental conditions. Unlike conventional brain mapping, this functional atlas supports robust *reverse inference*: predicting the mental processes from brain activity in the regions delineated by the atlas. To establish that this reverse inference is indeed governed by the corresponding concepts, and not idiosyncrasies of experimental designs, we show that it can accurately decode the cognitive concepts recruited in new tasks. These results demonstrate that aggregating independent task-fMRI studies can provide a more precise global atlas of selective associations between brain and cognition.

## Author summary

Cognitive neuroscience uses neuroimaging to identify brain systems engaged in specific cognitive tasks. However, linking unequivocally brain systems with cognitive functions is difficult: each task probes only a small number of facets of cognition, while brain systems are often engaged in many tasks. We develop a new approach to generate a functional atlas of cognition, demonstrating brain systems selectively associated with specific cognitive functions. This approach relies upon an ontology that defines specific cognitive functions and the relations between them, along with an analysis scheme tailored to this ontology. Using a database of thirty neuroimaging studies, we show that this approach provides a

highly-specific atlas of mental functions, and that it can decode the mental processes engaged in new tasks.

## Introduction

A major challenge to reaching a global understanding of the functional organization of the human brain is that each neuroimaging experiment only probes a small number of cognitive processes. Cognitive neuroscience is faced with a profusion of findings relating specific psychological functions to brain activity. These are like a collection of anecdotes that the field must assemble into a comprehensive description of the neural basis of mental functions, akin to "playing twenty questions with nature" [1]. However, maps from individual studies are not easily assembled into a functional atlas. On the one hand, the brain recruits similar neural territories to solve very different cognitive problems. For instance, the intra-parietal sulcus is often studied in the context of spatial attention; however, it is also activated in response to mathematical processing [2], cognitive control [3], and social cognition and language processing [4]. On the other hand, aggregating brain responses across studies to refine descriptions of the function of brain regions faces two challenges: First, experiments are often quite disparate and each one is crafted to single out a specific psychological mechanism, often suppressing other mechanisms. Second, standard brain-mapping analyses enable conclusions on responses to tasks or stimuli, and not on the function of given brain regions.

Cognitive subtraction, via the opposition of carefully-crafted stimuli or tasks, is used to isolate differential responses to a cognitive effect. However, scaling this approach to many studies and cognitive effects leads to neural activity maps with little functional specificity, hard to assemble in an atlas of cognitive function. Indeed, any particular task recruits many mental processes; while it may sometimes be possible to cancel out all but one process across tasks (e.g. through the use of conjunction analysis [5]), it is not feasible to do this on a large scale. Furthermore, it can be difficult to eliminate all possible confounds between tasks and mental processes. An additional challenge to the selectivity of this approach is that, with sufficient statistical power, nearly all regions in the brain will respond in a statistically significant way to an experimental manipulation [6].

The standard approach to the analysis of functional brain images maps the response of brain regions to a known psychological manipulation [7]. However, this is most often not the question that we actually wish to answer. Rather, we want to understand the mapping between brain regions/networks and psychological functions (i.e. "what function does the fronto-parietal network implement?"). If we understood these mappings, then in theory we could predict the mental state of an individual based solely on patterns of activation; this is often referred to as *reverse inference* [8], because it reverses the usual pattern of inference from mental state to brain activation. Whereas informal reverse inference (e.g. based on a selective review of the literature) can be highly biased, it is increasingly common to use meta-analytic tools such as Neurosynth [9] to perform formal reverse inference analyses (also know as *decoding*). However, these inferences remain challenging to interpret due to the trade-off between breadth and specificity that is necessary to create a sufficiently large database (e.g. see discussion in [10,11]).

The optimal basis for brain decoding would be a large database of task fMRI datasets spanning a broad range of mental functions. Previous work has demonstrated that it is possible to decode the task being performed by an individual, in a way that generalizes across individuals [1], but this does not provide insight into the specific cognitive functions being engaged, which is necessary if we wish to infer mental functions associated with novel tasks. The goal of decoding cognitive functions rather than tasks requires that the data are annotated using an ontology of cognitive functions [13–15], which can then become the target for decoding. Some recent work has used a similar approach in restricted domains, such as pain [16], and was able to isolate brain networks selective to physical pain. Extending this success to the entire scope of cognition requires modeling a broad range of experiments with sufficient annotations to serve as the basis for decoding.

To date, the construction of human functional brain atlases has primarily relied upon the combination of resting-state fMRI and coordinate-based meta-analyses. This approach is attractive because of the widespread availability of resting-state fMRI data (from which brain functional networks can be inferred through statistical approaches [17]), and the ability to link function to structure through the use of annotated coordinate-based data (such as those in the BrainMap [18] and Neurosynth [9] databases). This approach has identified a set of large-scale networks that are consistently related to specific sets of cognitive functions [19,20], and provides

decompositions of specific regions [21, 22]. However, resting-state analysis is limited in the set of functional states that it can identify [23], and meta-analytic databases are limited in the specificity of their annotation of task data, as well as in the quality of the data, given that it is reconstructed merely from activation coordinates reported in published papers.

A comprehensive functional brain atlas should link brain structures and cognitive functions in both forward and reverse inferences [7]. To build such a bilateral mapping, we introduce the concept of "ontology-based decoding,", in which the targets of decoding are specific cognitive features annotated according to an ontology. This idea was already present in [1, 9, 24]; here we show how an ontology enables scaling it to many cognitive features, to increase breadth. In the present case, we use the Cognitive Paradigm Ontology (CogPO) [15], that provides a common vocabulary of concepts related to psychological tasks and their relationships (see Modeling brain response to cognitive-ontology concepts). Forward inference then relies on ontology-defined contrasts across experiments, while reverse inference is performed using an ontology-informed decoder to leverage this specific set of oppositions (see Fig. 1 and methodological details). We apply these forward and reverse inferences to the individual activation maps of a large task-fMRI database: 30 studies, 837 subjects, 196 experimental conditions, and almost 7000 activation maps (see Modeling brain response to cognitive-ontology concepts). We use studies from different laboratories, that cover various cognitive domains such as language, vision, decision making, and arithmetics. We start from the raw data to produce statistical brain maps, as this enables homogeneous preprocessing and thorough quality control. The results of this approach demonstrate that it is possible to decode specific cognitive functions from brain activity, even if the subject is performing a task not included in the database.

# Materials and methods

## An ontology to describe cognitive neuroimaging studies

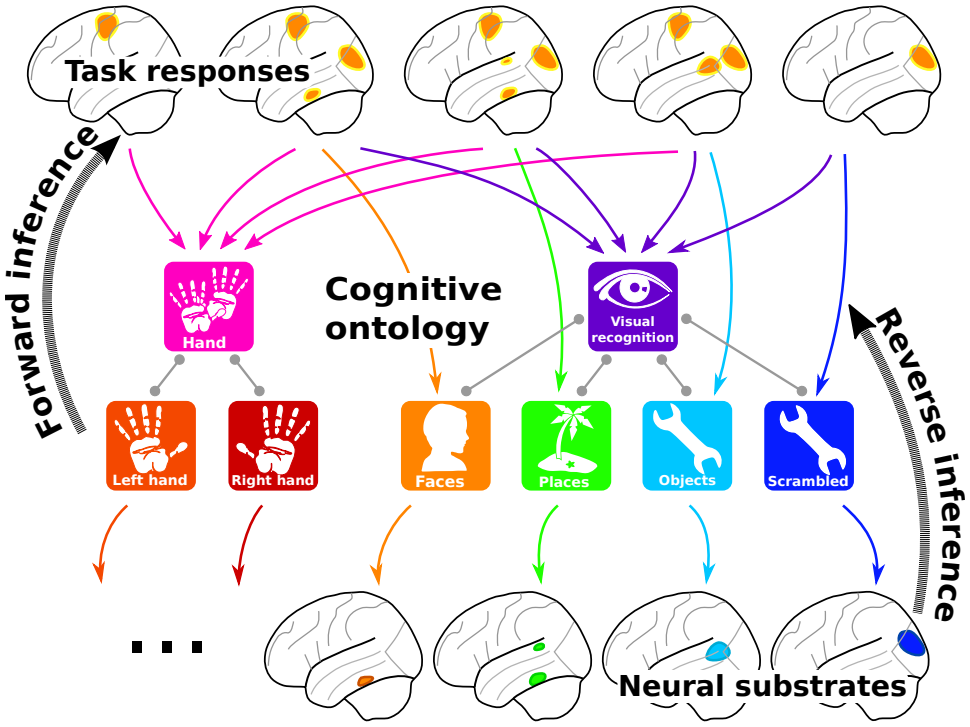
The main challenge to accumulate task fMRI is to account for the disparity in experimental paradigms. One solution is the use of cognitive ontologies that define terms describing the cognitive tasks at hand and enable to relate them. The choice of the ontology must meet two opposite goals: have a good coverage of the cognitive space, and document overlap between studies. In practice, each cognitive term describing mental processes must be expressed in several studies of our database to ensure the generalizability of our inference.

**Terms** The cognitive ontologies currently being developed in the neuroimaging community follow two directions. The Cognitive Paradigm Ontology (CogPO) [15], which is derived from the BrainMap taxonomy [18], concentrates on the description of the experimental conditions that characterize an experimental paradigm. A taxonomy is a special case of ontology in which links between concepts are captured in categories: high-level concepts from categories that encompass lower-level concepts. In CogPO, experimental tasks are described via different categories that represent the stimuli, the expected responses, and the instructions given to the subjects, *e.g.,* "stimulus modality", "explicit stimulus", "explicit response". The CogPO terms are rather broad, but enable to find common task descriptors regardless of the original intent of the study. More tailored towards cognitive processes, the Cognitive Atlas [14] lists a large number of cognitive tasks and concepts, and increasingly links them together. We decide to mainly use terms from CogPO, and extend it where our database can benefit from more precise or high-level descriptions. Not all terms of CogPO in our database are present over multiple studies, and thus we only use a subset of CogPO. Similarly, with the limited number of studies in our database, there is only little overlap in high-level cognition. We added only the "language" label from the Cognitive Atlas.

It should be noted that the ontology does not have a full hierarchical structure, as *stimulus modality*, *explicit stimulus* and *explicit response* convey different level of information. Further work with growing databases will however need to add more and more terms. Finding a consistent structure underlying all these terms is a hard task.

**Categories** Functional MRI experiments are carefully designed to balance conditions of interest with control conditions to cancel out effects related to the stimulation. As we do not want to ignore the designs, but rather leverage them in the

**Fig 1. Brain mapping with a cognitive ontology.** Our approach characterizes the task conditions that correspond to each brain image with terms from a cognitive ontology. *Forward inference* maps differences between brain responses for a given term and its neighbors in the ontology, i.e. closely related psychological notions. *Reverse inference* is achieved by predicting the terms associated with the task from brain activity. The figure depicts the analysis of visual object perception tasks with motor response. A forward inference captures brain responses in motor, primary visual and high-level visual areas. Reverse inference captures which regions or *neural substrate* are predictive of different terms, discarding common response to different tasks, here in the primary visual cortex.




context of a large-scale inference, we introduce an additional category level for our terms, that groups together terms –or conditions– that are typically contrasted in individual studies. These new categories strongly relate to the paradigm classes from BrainMap and the tasks from the Cognitive Atlas. The categories we choose are relevant to our database, and reflect the contrasts found in the studies. They nonetheless could be modified or extended further to test other hypotheses. This hierarchy of terms enables to co-analyze heterogeneous studies. Table 4 references the categories and associated terms used in this paper.

## Forward inference

Standard forward inference in functional neuroimaging uses the GLM (general linear model), which models brain responses as linear combinations of multiple effects. We use a *one-hot-encoding* of the concepts, *i.e.* we represent their presence in the tasks by a binary design matrix. We test for response induced by each concept with a second-level analysis using cross-studies contrasts.

To disentangle various experimental factors, brain mapping uses contrasts. Individual studies are crafted to isolate cognitive processes with control conditions, e.g. a face-recognition study would rely on a "face versus place" or a "face versus scrambled picture" contrast. To separate cognitive factors without a strong prior on control conditions, the alternative is to contrast a term against all related terms, e.g., "face versus place and scrambled picture".

We use the categories of our ontology to define such contrasts in a systematic way for the wide array of cognitive concepts touched in our database. This approach yields groups of terms within the task categories, as described in Table 1:

the task categories are used to define the conditions and their controls. Inside each group, we perform a GLM analysis with all the "one versus others" contrasts. We denote these *ontology contrasts*. Compared to a standard group analysis, the benefit of this GLM is that the control conditions for each effect studied span a much wider range of stimuli than typical studies.

## Reverse inference

For reverse inference, we rely on large-scale decoding [1]. Prior work [1, 24] tackles this question using a multi-class predictive model, the targets of the classification being separate cognitive labels. Our formulation is different as our goal is to predict the presence or absence of a term, effectively inverting the inference of our forward model based on one-hot-encoding. This implies that each image is associated with more than a single label, which corresponds to multi-label classification in a decoding setting.

**A hierarchical decoder** Linear models are widely used for decoding as they give good prediction and their parameters form brain maps. However, in a multi-label setting, they give rise to a profusion of separate one-versus-all (OvA) problems and cannot exploit the shared information between each label. We use a method based on stacked regressions [25]: two layers of linear models (logistic regressions) discriminating different cognitive terms. The first layer is tuned to specific oppositions between terms related in the ontology, while the second is tuned to predict which specific term is most relevant. This peculiar classifier architecture is tailored to the ontology that defines the structure of the targeted cognitive information. In the future, more complex cognitive ontologies may entail further refinements of the classifier.

**First layer** First, we stack the decisions of the OvA classifiers, that capture specific activation patterns across all tasks. This allows to relate cognitive processes across independent cognitive disciplines. Second, we build one-versus-one (OvO) classifiers by opposing terms that belong to the same task category (see Supplementary Table 4). This enables to generalize the notion of contrasts and subtraction-logic that is implicit to the majority of fMRI experiments. Finally, we build classifiers predicting the actual task categories from Supplementary Table 4. It enables to build a hierarchical decoding framework, that combines the decisions of simpler problems, namely classifying the task categories, and more subtle, within-category problems: the OvO classifiers. There may be better choices of classifiers, but the final predictor weights them, and therefore mitigates the introduction of unnecessary or sub-optimal classifiers. We list in Supplementary Table 2 all the classifiers that we use in the first level to learn the feature space capturing the ontology.

**Second layer** In a second layer, we learn the terms on the reduced representation with an OvA scheme, which also uses $\ell_1$-penalized logistic regression. The final output of this method is one linear classifier per term, that can be recovered by the linear combination of the coefficients of the base classifiers, with the coefficients of the final classifiers. The resulting ontology-informed decoder combines fined-grain information captured by opposing matching conditions in the first level with more universal decisions in the second level that outputs the presence or absence of a term. This combination is itself a linear classifier per label, and thus yields discriminant brain patterns for each term. Fig. 2 summarizes this decoding procedure and Section 4 gives more specific details.

Such a two-step classification is important because binary classifiers opposing one term to another exhibit undesirable properties in rich output settings: for instance a binary classifier that would detect occurrences of *right hand* task would typically classify all *left hand* task occurrence as *right hand*, given that the negative class for this problem typically involved mostly non-manual tasks. Leveraging a two- instead of one-layer classification architecture creates the possibility to capture more subtle effects, a trick systematically used in recent deep learning models.

**Cross validation** To evaluate the procedure, we perform the classification in randomized leave-3-study-out cross validation scheme. Cross-study prediction ensures that the representation of the cognitive labels generalizes across paradigms. We run 100 iterations of the cross validation to get a good estimate of the classifier's performance.

| CogPO Categories | Task Categories | Terms | contrasts |
|---|---|---|---|
| Stimulus modality | - | visual<br>auditory | visual - auditory<br>auditory - visual |
| Explicit stimulus | Sounds | human voice<br>sound | human voice - sound<br>sound - human voice |
| | Retinotopy | vertical checker-board<br>horizontal checkerboard | vertical checkerboard - horizontal checker-board<br>horizontal checkerboard - vertical checker-board |
| | Object recognition | faces<br>places<br>objects<br>scramble | faces - $\frac{1}{3}$(places + object + scramble)<br>places - $\frac{1}{3}$(faces + object + scramble)<br>object - $\frac{1}{3}$(faces + places + scramble)<br>scramble - $\frac{1}{3}$(faces + places + object) |
| | Symbol recognition | words<br>digits | words - digits<br>digits - words |
| Response modality | Motor - hands | left hand<br>right hand | left hand - right hand<br>right hand - left hand |
| | Motor - feet | left foot<br>right foot | left foot - right foot<br>right foot - left foot |
| | Arithmetics | saccades | saccades |
| Instructions | Arithmetics | calculation | calculation |
| Cognitive Atlas term | No category | language | language |

**Table 1. Contrasts used to characterize tasks effects in our database.** We used CogPO categories for task-related description, and add necessary terms from Cognitive Atlas to describe higher-level cognitive aspect. Here we report only terms that were present in more than one study –aside from the "left foot", which maps in the analysis as maps in "feet" task category, but not "right foot". The task categories group terms typically used as conditions and their controls to test a hypothesis. The *stimulus modality* category stands for CogPO and task categories. Some terms do not belong to any task category and are referred as such. The *arithmetics* task category spans across the *response modality* and *instructions* CogPO categories.

# Results

## An atlas of areas linked to function

Using a database of 30 studies, we demonstrate that our approach captures a rich mapping of the brain, identifying networks with a specific link to cognitive concepts. Prediction of cognitive components in new paradigms validates this claim.

## Linking brain networks and cognitive concepts

We combine forward and reverse inference to construct a one-to-one mapping between brain structures and cognitive concepts. Forward inference across studies requires adapting brain mapping analysis to leverage the ontology. Mapping the brain response to the presence of a concept in tasks selects unspecific regions, as it captures other related effects, *e.g.* selecting the primary visual cortex for any visual task (Fig. 3). To obtain a more focal mapping, we remove these effects by opposing the concept of interest to related concepts in the ontology. Reverse inference narrows down to regions specific to the term. However, as we use a multivariate procedure, some of its variables may model sources of noise [26]. For instance, when using visual n-back tasks with a motor response to map the visual system, the motor response creates confounding signals. A multivariate procedure could use signal from regions that capture these confounds to subtract them from vision-specific activity, leading to better prediction. As such regions are not directly related to the task, they are well filtered with a standard GLM (General Linear Model) used in forward inference. For this reason, our final maps combine statistics from forward and reverse inference: functional regions are composed of voxels that are both recruited by the cognitive process of interest *and* predictive of this process; see Section Consensus between forward and reverse inference for statistical arguments and [27] for more fundamental motivations regarding causal inference. Fig. 3a–d shows how the neural-activity patterns for the "places" label progressively narrow on the PPA with the different approaches. Thus we link each cognitive concept to a set of focal regions, resulting in a brain-wide functional atlas.

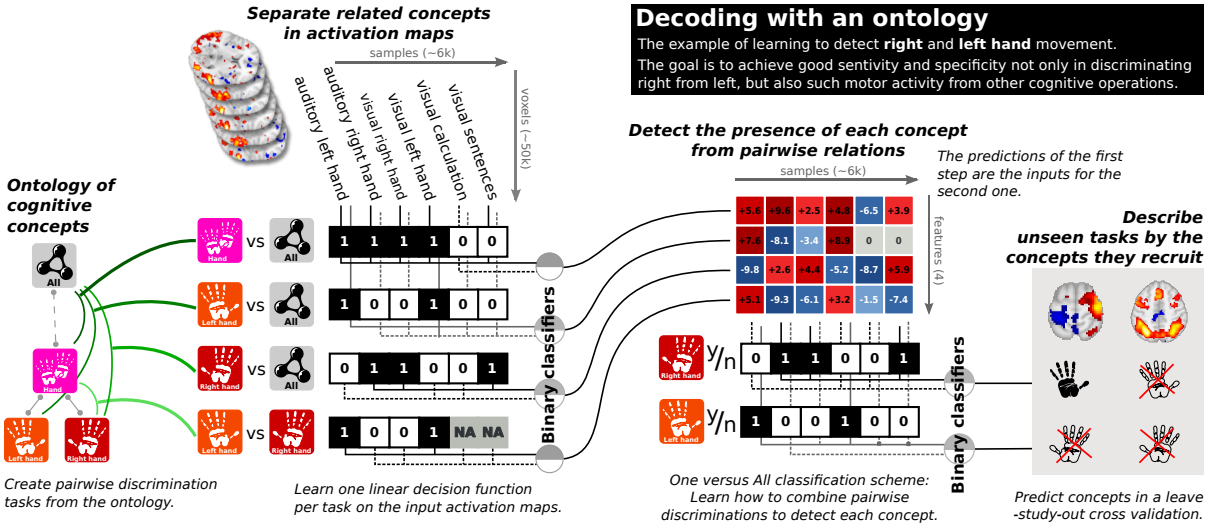



**Fig 2. Ontology informed decoding** The hierarchical decoding procedure reduces the dimensionality by stacking the decision functions of several simple binary classifiers, which mimic study-level contrasts by opposing each term to matching ones. A second level of one-versus-all (OvA) classifiers predicts the presence of terms using the output of the first level. The first layer may be seen as capturing whether a given brain activity map looks more like face or place recognition, objects or scrambled images, visual or motor stimuli. The second layer combines this information to conclude on what cognitive terms best describe the given activity. Final linear classifiers may be recovered by combining the coefficients of the first and second level classifiers.

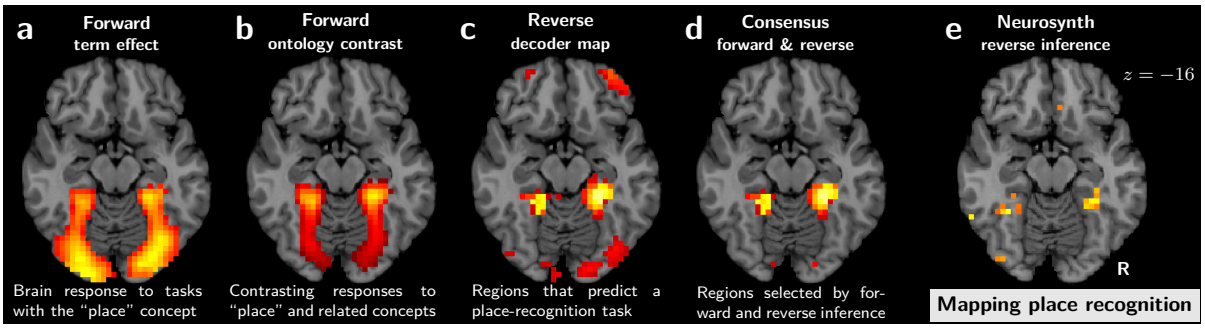
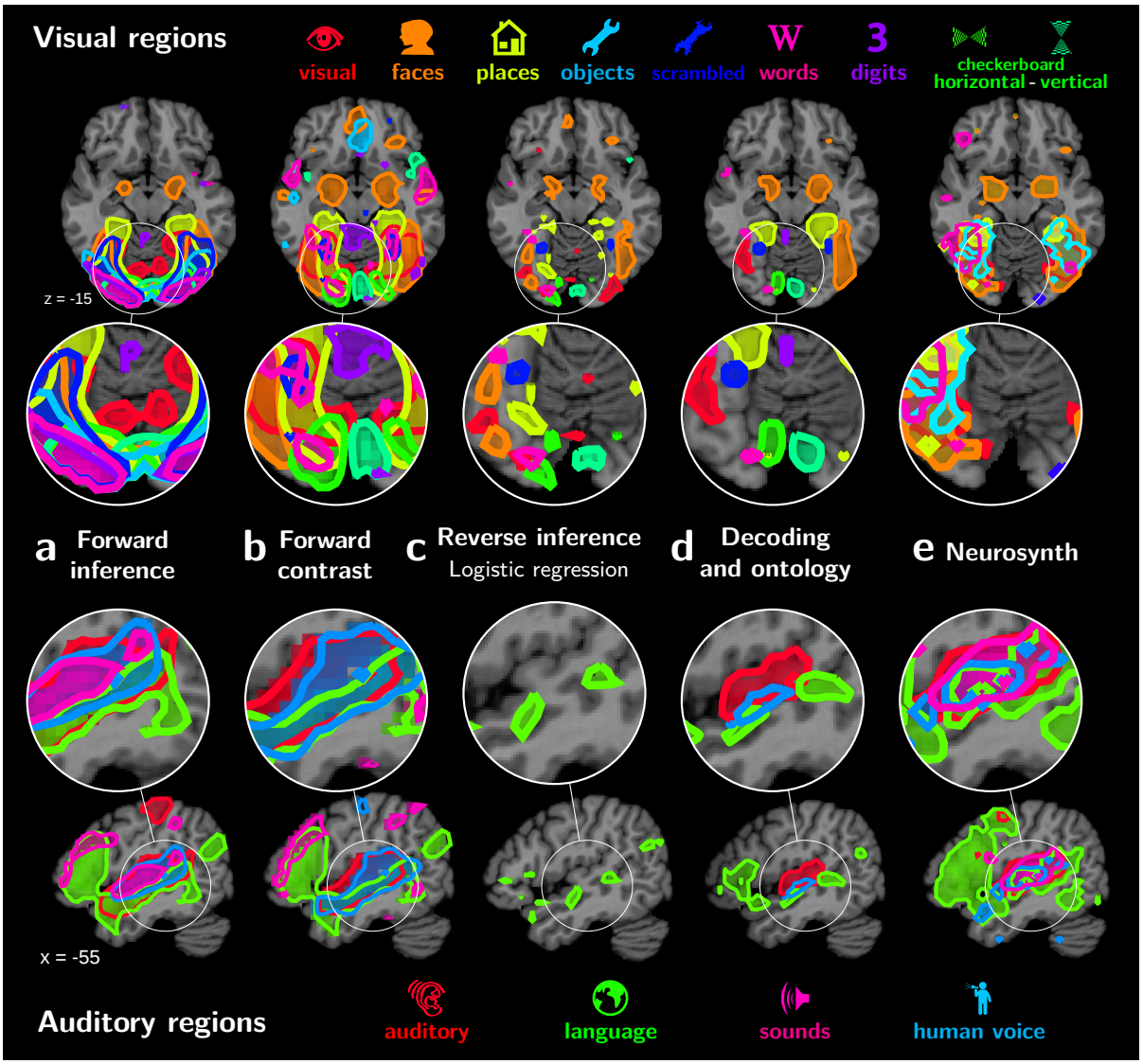



**Fig 3. Maps for the different inference types**. Left (**a**–**d**): maps of the different inferences on our database for the "place" concept. The consensus between reverse inference and forward inference based on contrasts defined from the ontology singles out the "parahippocampal place area" (PPA) for the "place" concept. Right (**d**): the NeuroSynth reverse-inference map for this concept. Reverse inference with Neurosynth also narrows well on the PPA, but is more noisy.

**Fig 4. Different functional atlases** – Regions outlined using different functional mapping approaches, from left to right: a. forward term mapping; b. forward inference with ontology contrasts (standard analysis); c. reverse inference with logistic regression; d. NeuroSynth reverse inference; and e. our approach, mapping with decoding and an ontology. The top part shows visual regions, and the lower one auditory regions in the left hemisphere. Forward term mapping outlines overlapping regions, as brain responses capture side effects such as the stimulus modality: for visual and auditory regions every cognitive term is represented in the corresponding primary cortex. Forward mapping using contrasts removes the overlap in primary regions, but a large overlap persists in mid-level regions, as control conditions are not well matched across studies. Standard reverse inference, specific to a term, creates overly sparse regions though with little overlap. Reverse inference with Neurosynth also displays large overlap in mid-level regions. Finally, ontology-based decoding maps recover known functional areas the visual and auditory cortices.



## Atlases with various mapping approaches

To build functional atlases, it is important to clearly identify the regions associated with different cognitive concepts. Fig. 3e shows that reverse-inference meta-analysis with Neurosynth also associates the PPA with the "place" term, but the region is not as well delineated as with our approach. Fig. 4 shows functional atlases of auditory and visual regions extracted with various mapping strategies. The relative position and overlap of the various maps is clearly visible. Forward-inference mapping of the effect of each term versus baseline on our database gives regions that strongly overlap (Fig. 4a). Indeed, the maps are not functionally specific and are dominated by low-level visual mechanisms in the occipital cortex and language in the temporal cortex. Using contrasts helps decreasing this overlap (Fig. 4b), and hence reveals some of the functional segregation of the visual system. However, as the stimuli are not perfectly balanced across experiments, contrasts also capture unspecific regions, such as responses in the lateral occipital cortex (LOC) for faces or places. Reverse inference with a logistic-regression decoder gives well separated regions, albeit small and scattered (Fig. 4c). The ontology-informed approach identifies well-separated regions that are consistent with current knowledge of brain functional organization (Fig. 4d). Finally, meta analysis with NeuroSynth separates maps related to the various terms better than forward analysis on our database of studies (Fig. 4e). Yet some overlap remains, for instance in the LOC for maps related to visual concepts. In addition, the outline of regions is ragged, as the corresponding maps are noisy (Fig. 3e), probably because they are reconstructed from peak coordinates. Note that overlaps across term-specific topographies are ultimately expected to remain, especially in associative cortices. In the following, we first discuss quantitative validation of the reverse-inference atlases, and then study in detail the atlas obtained with the ontology-informed approach.

## Decoding cognition validates the atlas

Upon qualitative inspection, the regions extracted by our mapping approach provide a good functional segmentation of the brain. For an objective test of this atlas, we quantify how well these regions support reverse inference. For this, we use the ontology-informed decoder to predict cognitive concepts describing tasks in new paradigms and measure the quality of the prediction. This approach was tested using a cross-validation scheme in which 3 studies were held out of each training fold for subsequent testing. Fig. 5 shows the corresponding scores: ontology-informed decoding accurately predicts cognitive concepts in unseen tasks. It predicts these concepts better than other commonly used decoders (logistic regression and naive Bayes, see also Evaluating prediction accuracy: cross-validation) and NeuroSynth decoding based on meta-analysis. This confirms that the corresponding atlas captures areas specialized in cognitive functions better than conventional approaches.

Very general labels such a "visual" are found in most studies, and therefore easy to predict. However, higher-level or more specialized cognitive concepts such as viewing digits or moving the left foot are seldom present (see Modeling brain response to cognitive-ontology concepts). For these rare labels, the fraction of prediction errors is not a useful measure. Indeed, simply assigning them to zero images would lead to a small fraction of errors. For this reason, Fig. 5 reports the area under the receiver operating characteristic (ROC) curve. This is a standard metric that summarizes both false positives and false negatives and is not biased for rare labels. This analysis showed that even for relatively rare concepts, successful decoding was possible.

## Regions in our functional atlas

Our approach links different cognitive terms to functionally-specialized brain regions:

### Visual regions

(Fig. 6a) Visual object recognition is linked to the ventral stream of specialized regions: primary visual areas associated with vertical and horizontal checkerboards in a basic but accurate retinotopic mapping; regions in the LOC linked to objects and scrambled objects; the Fusiform Face Area (FFA) and parahippocampal place area (PPA) associated respectively with "faces" and "places" terms; the region called visual word form area (VWFA) [28] linked to word recognition. Interestingly, both

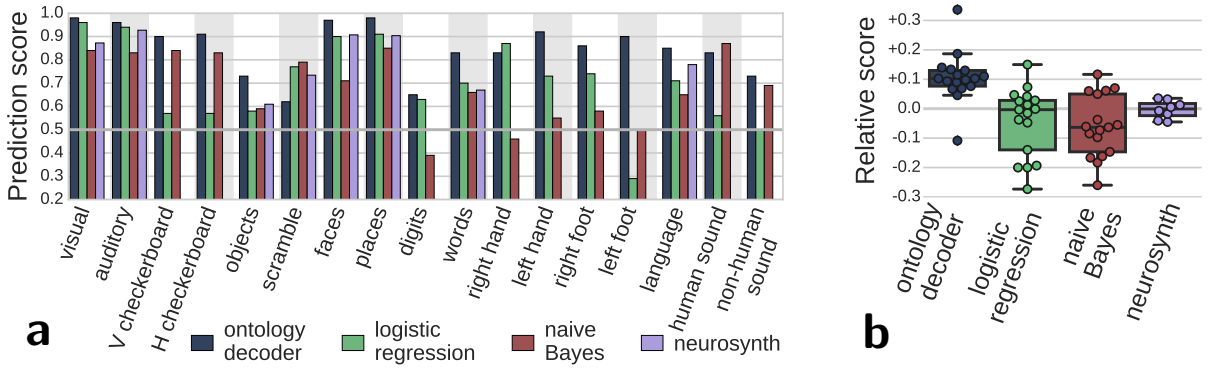



**Fig 5. Prediction scores for different methods**: area under the ROC curve (1 is perfect prediction, while chance is at 0.5); **a** score for each term; **b** score relative to the average per term for each decoding approach. As the terms in NeuroSynth do not fully overlap with the terms used in our database, not every term has a prediction score with NeuroSynth. The ontology-informed decoder is almost always able to assign the right cognitive concepts to an unknown task and clearly out-perform standards decoders: logistic regression and naive Bayes classifier trained on our database. It also outperforms the NeuroSynth decoding based on meta-analysis.

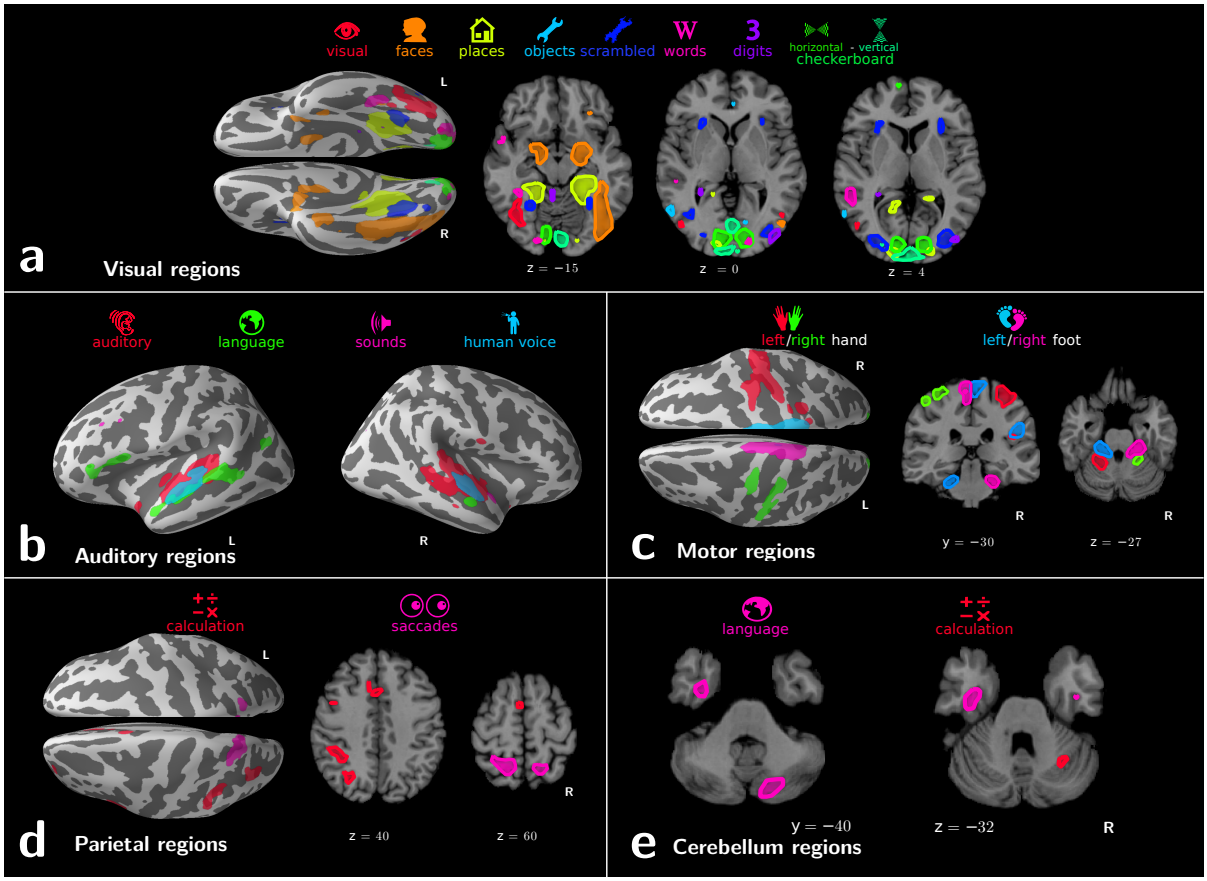



**Fig 6. Functional atlases with decoding in an ontology** – Regions linked to the various cognitive terms by our mapping approach. They are displayed in 5 different panels depending on their location in the brain: a. visual regions; b. auditory regions ; c. motor regions ; d. parietal regions ; e.cerebellum regions.

amygdalas also appear related to faces, which could be due to emotional effects of face processing not modeled in the ontology. Digit-viewing does not outline meaningful regions. Corresponding decoding scores are poor (Fig. 5): our database is not suited to cross-study mapping of digit viewing. This example confirms that decoding scores can serve as Occam's razor, validating or falsifying functional regions.

### Auditory regions

(Fig. 6b) Four cognitive terms are represented in the temporal lobe: "auditory", "sounds", "language", and "human voice". These correspond to increasingly specific concepts in our ontology, and map increasingly focal regions: The "auditory" label denotes the stimulus modality, a fairly general concept, and is linked to the entire auditory cortex. The more precise "sounds" label is associated with Heschl's gyrus. The "language" label highlights a prototypical left-lateralized language network: anterior and posterior superior temporal sulcus (STS), temporal lobe, supramarginal gyrus, and Broca's area. The "human voice" label reveals regions in the upper bank of the STS that were previous identified as voice-selective regions by contrasting human voices with closely-matched scrambled voices control conditions [29]. That the mapping singles out such regions from the data is an impressive feat given that only one study in our database [30] features both human voices and non-voice auditory conditions.

### Motor regions

(Fig. 6c) Motor labels reveal the lateralized hands and feet representations in the primary motor cortex, as well as in the cerebellum.

### Parietal regions

(Fig. 6d) Saccadic eye movements and mental arithmetic are known to recruit almost overlapping parietal areas [2], which are difficult to separate with standard analysis. In the IPS (intra-parietal sulcus), we find bilateral regions for saccades but calculation appears left lateralized, consistent with previous reports [31]. Cross-study analysis of activation maps is important to study such nearly-colocalized functions from different cognitive domains. Indeed, meta-analysis based on coordinates suffers a loss of spatial resolution (Fig. 3e).

### Cerebellar regions

While the cerebellum is involved in a variety of mental processes, there are very few systematic mapping results. Previous work [32] studied the somatotopic organization of the cerebellum visible on Fig. 6c, with an inverted laterality of functional areas with respect to cortical somatotopy. Other higher-level cognitive functions are represented in the cerebellum with the same inversion. Notably our analysis links the "language" term to a right-lateralized cerebellum region in Crus II (Fig. 6e), consistent with language studies [33]. Finally, the "calculation" term is also represented in the right cerebellar cortex, in the superior medial section of the lobule VI. This location has been linked to working memory [34]. It appears here linked to calculation, consistent with the fact that mental arithmetic has a strong working-memory component [35], and our cognitive ontology does not explicitly model working memory.

## Discussion

The inference framework introduced here represents a new approach to developing functional atlases of the human brain. It formally characterizes representations for various cognitive processes that evoke overlapping brain responses, and makes it possible to pool many task-fMRI experiments probing different cognitive domains. Existing meta-analysis approaches face the risk of being unspecific, as demonstrated by our standard analysis results on our database (Fig. 3 and 4). Databases of coordinates, such as NeuroSynth, can more easily accumulate data on many different cognitive concepts and support formal reverse inference. This data accumulation is promising, but existing reverse-inference approaches do not suffice to fully remove the overlap in functional regions (Fig. 6). Our approach gives more differentiated maps for cognitive concepts by analyzing them in a way that leverages the cognitive ontology. They are also sharper, presumably because they are derived from images rather than coordinates. In a multi-modal framework [17], these maps could be combined with resting-state and anatomical data to provide cognitive resolution to brain

parcellations. Note that our framework is meant to be used at the population level and does not address individual brain mapping or decoding.

## Reverse inference mapping

Our analysis framework overcomes the loss in specificity typical of data aggregation. As a result, it enables analyzing jointly more cognitive processes. These richer models can map qualitatively different information. Analyzing more diverse databases of brain functional images can bring together two central brain-mapping questions: *where* is a given cognitive process implemented, and *what* cognitive processes are represented by a given brain structure. Answers to the "what" question have traditionally been provided by invasive studies or neurological lesion reports. Indeed, in a given fMRI study, brain activity results from the task. Concluding on what processes are implied by the observed activity risks merely capturing this task. Decoding across studies can answer this question, by demonstrating the ability to perform accurate inference from brain activity to cognitive function [36].

Reverse-inference maps are essential to functional brain mapping. A key insight comes from the analysis in NeuroSynth [9]: some brain structures are activated in many tasks. Hence, a standard analysis –forward inference– showing such a structure as activated does not provide much information about what function is being engaged. Reverse inference puts the observed brain activity in a wider context by characterizing the behavior that it implies. The analysis performed in NeuroSynth accounts for the multiple tasks that activate a given structure, performing a Bayesian inversion with the so-called *Naive Bayes* model; however, it does not account for other activation foci in the brain that characterize the function. Put differently, our approach departs from the model used by NeuroSynth for reverse inference by what it conditions upon: NeuroSynth's model asserts functional specialization *conditional to* other terms, while we condition on other brain locations when predicting concept occurrence. This difference should be kept in mind when interpreting differences between the two types of approaches. The Inferior Temporal Gyrus (ITG), for instance, is more active in object-recognition tasks than in other paradigms. However, observing activity in the ITG does not help deciding whether the subject is recognizing faces or other types of objects: the information is in the Fusiform gyrus. An important difference between reverse-inference maps with a Naive Bayes –as in Neurosynth– and using a linear model –as in our approach– is that the Naive Bayes maps do no capture dependencies across voxels. On the opposite, linear models map how brain activity in a voxel relates to behavior *conditionally* on other voxels. Technically, this is the reason why Neurosynth reverse-inference maps related to object recognition overlap in the IT cortex (Fig. 3e) while maps produced by our approach separate the representations of the various terms in the ventral mosaic (Fig. 3d).

Another, more subtle, benefit of the two-layer model over more classical multi-label approaches is that it combines the decisions of classifiers based on subsets of the data, such as the OvO classifiers, which helps learning relevant local discriminative information.

In sum, our mapping approach provides a different type of brain maps: They quantify how much observing activity in a given brain location, as opposed to other brain locations, informs on whether the subject was engaged in a cognitive operation.

## Generalizing beyond single studies

Brain functional atlases are hard to falsify: is a functional atlas specific to the experimental paradigms employed to build it, or is it more generally characteristic of human brain organization? The success of statistically-grounded reverse inference, which generalizes to new paradigms from unseen studies, suggests that there must be some degree of generality in the present atlas. In demonstrating this generalization, the present work goes beyond previous work that had shown generalization to new subjects under known task conditions [1], but not to unknown protocols. However, it is worth noting that here too we found that it was easier to predict on held-out subjects (from one of the training studies) than on held-out studies (see Evaluating prediction accuracy: cross-validation), consistent with a substantial effect of the specific task (see section 2). Despite this, our ontology-enabled approach was able to successfully predict cognitive processes for new tasks. Interestingly, it opens the possibility to perform prospective decoding analyses on novel data, hence makes it easier to grasp the added information of incoming data.

To enable this generalization across paradigms,

we characterize each task by the multiple cognitive concepts that it recruits, that are specified in the ontology. Departing from the subtractions often used in brain mapping, our framework relies on quantifying full descriptions of the tasks. In the context of decoding, this approach leads to *multi-label prediction*, predicting multiple terms for an activation map, as opposed to *multi-class prediction*, used in prior works [1, 16], that assigns each new map to a single class. The use of the multi-label approach combined with an ontology capturing the relationships between terms provides a principled way of modeling the multiple components of cognition and thus avoids the need for hand-crafted oppositions that are customarily used in subtraction studies. Defining good ontologies is yet another challenge for the community, but it is not unlikely that brain imaging will become part of that process [36, 37]. Providing a methodological approach founded on an explicit hierarchy of cognitive concepts would allow to test for different cognitive ontologies, and, provided with a comparison metric, select the best ontology according to the available data. Although the present analysis is limited to a relatively small set of cognitive functions, such an approach will be essential as the field attempts to scale such analyses to the breadth of human cognition.

## Conclusion

To build brain functional atlases that map many cognitive processes, we have found that reverse inference and an ontology relating these processes were key ingredients. Indeed, because of the experimental devices used in cognitive neuroimaging, some regions –e.g. attentional or sensory regions– tend to be overly represented in forward inferences. An ontology encodes the related cognitive processes that must be studied together to best establish forward or reverse inferences.

Using a relatively small number of independent task fMRI datasets, our brain-mapping approach reconciles the conundrum of multiple cognitive processes/labels mapping to often overlapping brain regions in activation studies. More data will enable even more fine-grained process-region mappings. In particular higher-level cognitive processes elude the present work, limited by the amount and the diversity of the studies in our database. Indeed, high-level terms form very rare classes in the datasets employed here (see Modeling brain response to cognitive-ontology concepts). With increased data sharing in the neuroimaging community [38], there is a growing opportunity to perform this kind of analysis on a much larger scale, ultimately providing a comprehensive atlas of neurocognitive organization. A major challenge to such analyses is the need for detailed task annotation; whereas annotation of task features such as the response effector is relatively straightforward, annotation of complex cognitive processes (e.g., whether a task involves attentional selection or working memory maintenance) is challenging and often contentious. The utility of the ontology in the present work suggests that this effort is worthwhile, and that the increased utilization of ontologies in cognitive neuroscience may be an essential component to solving the problem of how cognitive function is organized in the brain.

## References


1. Newell A. You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In: Visual information processing; 1973.
2. Knops A, Thirion B, Hubbard EM, Michel V, Dehaene S. Recruitment of an area involved in eye movements during mental arithmetic. Science. 2009;324:1583.
3. Dosenbach NU, Fair DA, Miezin FM, Cohen AL, Wenger KK, Dosenbach RA, et al. Distinct brain networks for adaptive and stable task control in humans. P Natl Acad Sci Usa. 2007;104:11073–11078.
4. Bzdok D, Hartwigsen G, Reid A, Laird AR, Fox PT, Eickhoff SB. Left inferior parietal lobe engagement in social cognition and language. Neurosci & Biobehav Rev. 2016;.
5. Price CJ, Friston KJ. Cognitive conjunction: a new approach to brain activation experiments. Neuroimage. 1997;5:261.
6. Thyreau B, Schwartz Y, Thirion B, Frouin V, Loth E, Vollstädt-Klein S, et al. Very large fMRI study using the IMAGEN database: Sensitivity–specificity and population effect modeling in relation to the underlying anatomy. NeuroImage. 2012;61:295.
7. Henson R. Forward inference using functional neuroimaging: Dissociations versus associations. Trends Cogn Sci. 2006;10:64–69.
8. Poldrack R. Can cognitive processes be inferred from neuroimaging data? Trends Cogn Sci. 2006;10:59.
9. Yarkoni T, Poldrack R, Nichols T, Essen DV, Wager T. Large-scale automated synthesis of human functional neuroimaging data. Nat Methods. 2011;8:665.
10. Wager TD, Atlas LY, Botvinick MM, Chang LJ, Coghill RC, Davis KD, et al. Pain in the ACC? Proc Natl Acad Sci USA. 2016;113:E2474–E2475.
11. Lieberman MD, Burns SM, Torre JB, Eisenberger NI. Reply to Wager et al.: Pain and the dACC: The importance of hit rate-adjusted effects and posterior probabilities with fair priors. Proc Natl Acad Sci USA. 2016; p. 201603186.
12. Poldrack RA, Halchenko YO, Hanson SJ. Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychol Sci. 2009;20:1364.
13. Price CJ, Friston KJ. Functional ontologies for cognition: The systematic definition of structure and function. Cognitive Neuropsychology. 2005;22:262.
14. Poldrack RA, Kittur A, Kalar D, Miller E, Seppa C, Gil Y, et al. The cognitive atlas: toward a knowledge foundation for cognitive neuroscience. Front neuroinform. 2011;5:17.
15. Turner J, Laird A. The cognitive paradigm ontology: design and application. Neuroinformatics. 2012;10:57.
16. Wager TD, Atlas LY, Lindquist MA, Roy M, Woo CW, Kross E. An fMRI-based neurologic signature of physical pain. N Engl J Med. 2013;368:1388.
17. Glasser M, Coalson T, Robinson E, Hacker C, Harwell J, Yacoub E, et al. A Multi-modal parcellation of human cerebral cortex. Nature. 2016;536:171.
18. Laird A, Lancaster J, Fox P. Brainmap. Neuroinformatics. 2005;3:65.
19. Smith SM, Fox PT, Miller KL, Glahn DC, Fox PM, Mackay CE, et al. Correspondence of the brain's functional architecture during activation and rest. Proc Natl Acad Sci. 2009;106:13040.
20. Laird AR, Fox PM, Eickhoff SB, Turner JA, Ray KL, McKay DR, et al. Behavioral interpretations of intrinsic connectivity networks. J cog neurosci. 2011;23:4022.
21. Chang LJ, Yarkoni T, Khaw MW, Sanfey AG. Decoding the role of the insula in human cognition: functional parcellation and large-scale reverse inference. Cereb Cortex. 2012; p. bhs065.
22. Bzdok D, Heeger A, Langner R, Laird AR, Fox PT, Palomero-Gallagher N, et al. Subspecialization in the human posterior medial cortex. Neuroimage. 2015;106:55.

23. Cole MW, Bassett DS, Power JD, Braver TS, Petersen SE. Intrinsic and task-evoked network architectures of the human brain. Neuron. 2014;83:238.

24. Poldrack RA, Barch D, Mitchell J, Wager T, Wagner A, Devlin J, et al. Towards open sharing of task-based fMRI data: The OpenfMRI project. Front Neuroinform. 2013;7:12.

25. Breiman L. Stacked regressions. Machine learning. 1996;24:49.

26. Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage. 2014;87:96–110.

27. Sebastian Weichwald, Timm Meyer, Ozan Özdenizci, Bernhard Schölkopf, Tonio Ball, and Moritz Grosse-Wentrup. Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, 110:48–59, 2015.

28. Cohen L, Dehaene S. Specialization within the ventral stream: the case for the visual word form area. NeuroImage. 2004;22:466.

29. Belin P, Zatorre RJ, Lafaille P, Ahad P, Pike B. Voice-selective areas in human auditory cortex. Nature. 2000;403:309.

30. Pinel P, Dehaene S. Genetic and environmental contributions to brain activation during calculation. NeuroImage. 2013;81:306.

31. Andres M, Seron X, Olivier E. Hemispheric lateralization of number comparison. Cognitive Brain Research. 2005;25:283.

32. Grodd W, Hulsmann E, Lotze M, Wildgruber D, Erb M. Sensorimotor mapping of the human cerebellum: fMRI evidence of somatotopic organization. Hum brain mapp. 2001;13:55.

33. Stoodley CJ, Schmahmann JD. Functional topography in the human cerebellum: a meta-analysis of neuroimaging studies. Neuroimage. 2009;44:489.

34. Durisko C, Fiez JA. Functional activation in the cerebellum during working memory and simple speech tasks. Cortex. 2010;46:896.

35. Zago L, Pesenti M, Mellet E, Crivello F, Mazoyer B, Tzourio-Mazoyer N. Neural correlates of simple and complex mental calculation. Neuroimage. 2001;13:314.

36. Poldrack RA, Yarkoni T. From brain maps to cognitive ontologies: Informatics and the search for mental structure. Annual review of psychology. 2016;67:587.

37. Danilo Bzdok, Thomas B T Yeo. Inference in the age of big data: Future perspectives on neuroscience. *NeuroImage*, 155:549-564, 2017.

38. Poldrack RA, Gorgolewski KJ. Making big data open: data sharing in neuroimaging. Nat neurosci. 2014;17:1510.

# Supporting information

## 1 Distribution of terms in our database

Our database is comprised of data from 30 studies, assembled from various sources. We have uploaded the subject-level maps resulting from our first-level analysis on NeuroVault. These form the inputs to our approach. We list in supplementary Table 1 references to these datasets as well as the NeuroVault URLs to download the maps.

The terms with which we label tasks cover very different concepts of cognitive science, ranging from very broad and general such as "visual", used to denote visual stimuli, to much more specific, such as "calculation". As a result, their distribution across study is very inhomogeneous. Broad terms are present in much more tasks and studies than specific terms. We observe a power-law like behavior in the distribution of terms (see Supplementary Fig. 1).

Supplementary Fig. 2 shows the overlap between term presence and studies. It outlines the difficulty of analysis and prediction across studies: given two terms to compare, there are often few studies with both of these terms. Note that the "left foot" is present in only one study, and we report results under the "left foot" label for what is detected as "not right foot" in the "feet category".

# References


1. Barch DM, Burgess GC, Harms MP, Petersen SE, Schlaggar BL, Corbetta M, et al. Function in the human connectome: task-fMRI and individual differences in behavior. Neuroimage. 2013;80:169–189.

2. Amalric M, Dehaene S. Origins of the brain networks for advanced mathematics in expert mathematicians. Proceedings of the National Academy of Sciences. 2016;113(18):4909–4917.

3. Cauvet E. Traitement des Structures Syntaxiques dans le langage et dans la musique. Paris 6; 2012.

4. Hara N, Cauvet E, Devauchelle A, DEHAENE S, PALLIER C, et al. Neural correlates of constituent structure in Language and Music. Neuroimage. 2009;47:S143.

5. Devauchelle AD, Oppenheim C, Rizzi L, Dehaene S, Pallier C. Sentence syntax and content in the human temporal lobe: an fMRI adaptation study in auditory and visual modalities. Journal of Cognitive Neuroscience. 2009;21(5):1000–1012.

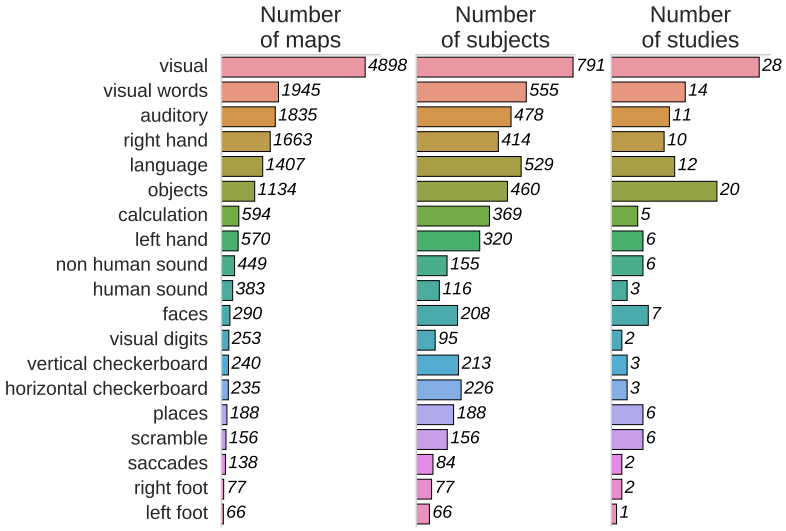
6. Schonberg T, Fox C, Mumford J, Congdon C, Trepel C, Poldrack R. Decreasing ventromedial prefrontal cortex activity during sequential risk-taking: an fMRI investigation of the balloon analog risk task. Frontiers in Neuroscience. 2012;6.

7. Aron A, Gluck M, Poldrack R. Long-term test–retest reliability of functional MRI in a classification learning task. Neuroimage. 2006;29:1000.

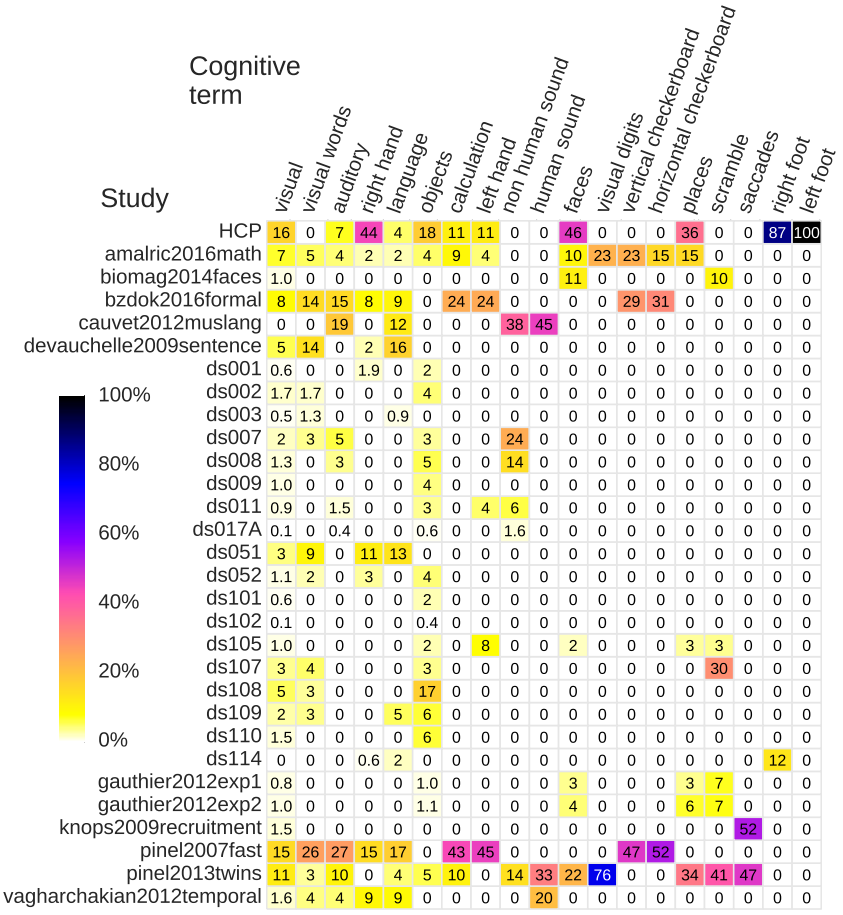
8. Xue G, Poldrack RA. Rhyme judgment;. https://openfmri.org/dataset/ds000003.

9. Xue G, Poldrack R. The neural substrates of visual perceptual learning of words: implications for the visual word form area hypothesis. J Cognitive Neurosci. 2007;19:1643.

10. Xue G, Aron A, Poldrack R. Common neural substrates for inhibition of spoken and manual responses. Cerebral Cortex. 2008;18:1923.




**Supplementary Fig. 1. Term distribution**: the number of times a term appears in our database, per map, subject, or study.




**Supplementary Fig. 2. Terms distribution in studies**: percentage of term occurrence in each study.

**Supplementary Table 1.**
**Studies in the database** The subject-level maps output by our preprocessing and first-level analysis have been uploaded to NeuroVault, for reproducibility of the analysis. These maps form the input of our analytic scheme. All the subject-level statistical maps that we computed as part of our preprocessing are available on http://neurovault.org/collections/1952

| Internal name | Ref |
|---|---|
| HCP release 1 | [1] |
| amalric2016math | [2] |
| cauvet2012muslang | [3, 4] |
| devauchelle2009sentence | [5] |
| ds001 | [6] |
| ds002 | [7] |
| ds003 | [8, 9] |
| ds007 | [10] |
| ds008 | [11] |
| ds009 | [12] |
| ds011 | [13] |
| ds017a | [14] |
| ds051 | [15, 16] |
| ds052 | [17] |
| ds101 | [18] |
| ds102 | [19] |
| ds105 | [20] |
| ds107 | [21] |
| ds108 | [22] |
| ds109 | [23] |
| ds110 | [24] |
| ds114 | [25] |
| gauthier2012exp1 | [26] |
| gauthier2012exp2 | [26] |
| biomag2014faces | [27] |
| knops2009recruitment | [28] |
| pinel2007fast | [29] |
| pinel2013twins | [30] |
| bzdok2016formal | [31] |
| vagharchakian2012temporal | [32] |


11. Aron AR, Behrens TE, Smith S, Frank MJ, Poldrack RA. Triangulating a cognitive control network using diffusion-weighted magnetic resonance imaging (MRI) and functional MRI. The Journal of Neuroscience. 2007;27:3743–3752.

12. Cohen JR. The development and generality of self -control. UCLA; 2009.

13. Foerde K, Knowlton B, Poldrack R. Modulation of competing memory systems by distraction. Proc Natl Acad Sci. 2006;103:11778.

14. Rizk-Jackson A, Aron AR, Poldrack RA. Classification learning and stop-signal (1 year test-retest);. https://openfmri.org/dataset/ds000017.

15. Alvarez RP, Poldrack RA. Cross-language repetition priming;. https://openfmri.org/dataset/ds000051.

16. Alvarez RP, Jasdzewski G, Poldrack RA. Building memories in two languages: An fMRI study of episodic encoding in bilinguals. In: Society for Neuroscience Abstracts; 2002.

17. Poldrack R, Clark J, Pare-Blagoev E, Shohamy D, Creso Moyano J, Myers C, et al. Interactive memory systems in the human brain. Nature. 2001;414:546.

18. Kelly A, Milham M. Cross-language repetition priming;. https://openfmri.org/dataset/ds000101.

19. Kelly A, Uddin LQ, Biswal BB, Castellanos F, Milham M. Competition between functional brain networks mediates behavioral variability. Neuroimage. 2008;39:527.

20. Haxby J, Gobbini I, Furey M, Ishai A, Schouten J, Pietrini P. Distributed and overlapping representations of faces and

objects in ventral temporal cortex. Science. 2001;293:2425.

21. Duncan K, Pattamadilok C, Knierim I, Devlin J. Consistency and variability in functional localisers. Neuroimage. 2009;46:1018.

22. Wager TD, Davidson ML, Hughes BL, Lindquist MA, Ochsner KN. Prefrontal-subcortical pathways mediating successful emotion regulation. Neuron. 2008;59:1037.

23. Moran JM, Jolly E, Mitchell JP. Social-cognitive deficits in normal aging. The Journal of Neuroscience. 2012;32:5553–5561.

24. Uncapher MR, Hutchinson JB, Wagner AD. Dissociable effects of top-down and bottom-up attention during episodic encoding. The Journal of Neuroscience. 2011;31:12613–12628.

25. Gorgolewski KJ, Storkey A, Bastin ME, Whittle IR, Wardlaw JM, Pernet CR. A test-retest fMRI dataset for motor, language and spatial attention functions. GigaScience. 2013;2:1.

26. Gauthier B, Eger E, Hesselmann G, Giraud AL, Kleinschmidt A. Temporal tuning properties along the human ventral visual stream. The Journal of Neuroscience. 2012;32:14433–14441.

27. Wakeman DG, Henson RN. A multi-subject, multi-modal human neuroimaging dataset. Scientific data. 2015;2.

28. Knops A, Thirion B, Hubbard EM, Michel V, Dehaene S. Recruitment of an area involved in eye movements during mental arithmetic. Science. 2009;324:1583.

29. Pinel P, Thirion B, Meriaux S, Jobert A, Serres J, Bihan DL, et al. Fast reproducible identification and large-scale databasing of individual functional cognitive networks. BMC neuroscience. 2007;8:91.

30. Pinel P, Dehaene S. Genetic and environmental contributions to brain activation during calculation. NeuroImage. 2013;81:306.

31. Bzdok D, Varoquaux G, Grisel O, Eickenberg M, Poupon C, Thirion B. Formal models of the network co-occurrence underlying mental operations. PLoS Comput Biol. 2016;12:e1004994.

32. Vagharchakian L, Dehaene-Lambertz G, Pallier C, Dehaene S. A temporal bottleneck in the language comprehension network. The Journal of Neuroscience. 2012;32:9089–9102.

33. Salimi-Khorshidi G, Smith SM, Keltner JR, Wager TD, et al. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. Neuroimage. 2009;45:810.

34. Varoquaux G, Gramfort A, Thirion B. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. ICML. 2012;.

35. Dietterich TG. Ensemble methods in machine learning. In: Multiple classifier systems. Springer; 2000. p. 1–15.

36. Hoyos-Idrobo A, Schwartz Y, Varoquaux G, Thirion B. Improving sparse recovery on structured images with bagged clustering. In: Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop on. IEEE; 2015. p. 73–76.

37. Schwartz Y, Thirion B, Varoquaux G. Mapping paradigm ontologies to and

from the brain. Advances in Neural Information Processing Systems. 2013; p. 1673–1681.

38. Breiman L. Stacked regressions. Machine learning. 1996;24:49.

39. Wellcome Department of Cognitive Neurology. SPM8; 2008. http://www.fil.ion.ucl.ac.uk/spm.

40. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825.

41. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Frontiers in Neuroinformatics. 2014;8:14.

42. Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage. 2014;87:96–110.

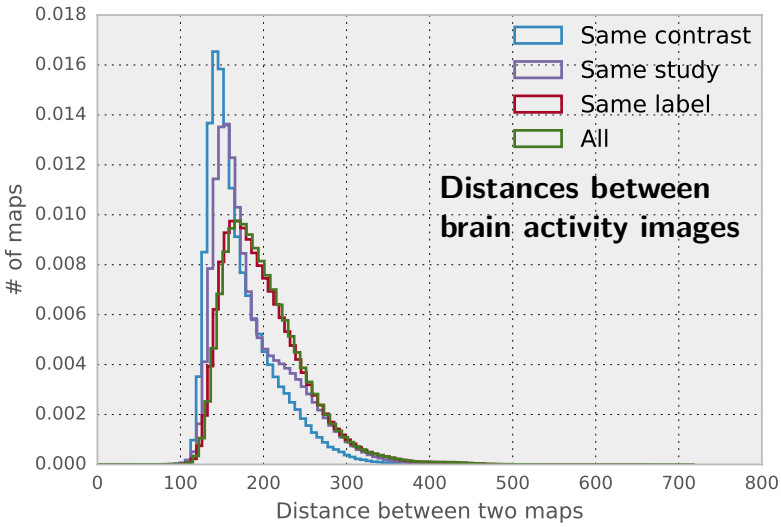
43. Poldrack RA, Halchenko YO, Hanson SJ. Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychol Sci. 2009;20:1364.

44. Poldrack RA, Barch D, Mitchell J, Wager T, Wagner A, Devlin J, et al. Towards open sharing of task-based fMRI data: The OpenfMRI project. Front Neuroinform. 2013;7:12.

# 2 Similarities of activations across the database

To understand how the variance is distributed in our database, we compute the pairwise distances between any two brain activity images in the database. In Supplementary Fig. 3 we show histograms of distance for all image pairs in the database, then images sharing a cognitive label, images taken from the same study, or corresponding to the same exact experimental contrast. We can see that while images drawn from the same contract or the same study tend to be close (small distance), distances between images sharing a cognitive label are distributed no differently than distances between images drawn randomly in the database.

These distributions show that, despite our uniform preprocessing, the inter-study variance is larger that the variance across labels of cognitive concepts. Indeed, images drawn from the same study tend to be closer than images sharing a cognitive label. Such lack of similarity is evidence that the idiosyncrasies of the experiment –imaging details but also implementation details of the paradigm such as specific choice of stimuli– can explain more variance than the cognitive concepts we are interested in capturing for large-scale functional mapping of the brain.




**Supplementary Fig. 3. Histograms of the distances between brain activity images**: pairwise distances across all the images of our 30-study database: comparing all images, images sharing a cognitive label, in the same study, or in the same exact contrast.

# 3 Forward analysis: ontology-based design across studies

## 3.1 Modeling brain response to cognitive-ontology concepts

In a standard GLM framework, we use a design matrix capturing the effect on brain activity of the presence of a term in the task description, followed by a set of contrasts to isolate contribution of the term of interest opposed to related terms in the ontology.

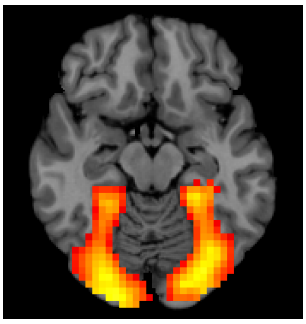
**Term effect** We assign a set of terms to each image, forming a *one-hot-encoding* of the database, *i.e.* representing the occurrence of terms by a binary design matrix. We follow the standard fMRI analysis framework and perform a General Linear Model (GLM). This gives the correlation of each separate voxel with the terms within a set of images, and enables to test for their significance. Using the GLM formulation:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon,$$

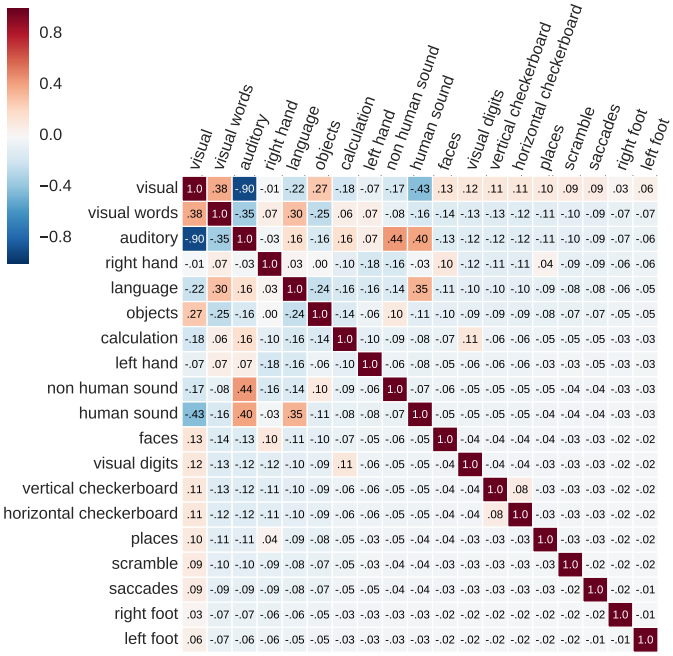
$\boldsymbol{y}$ corresponds to the activation maps, $\boldsymbol{X}$ to the design matrix modeling the presence of terms, and $\boldsymbol{\beta}$ to the term effects. The input activation maps are subject-level condition versus baseline maps. Supplementary Fig. 4 shows the effect map for the *places* term. We will use this term in the following to illustrate the differences between the types of inference.

Correlations in the terms induce correlations in the design matrix: effects of terms that appear always together in tasks cannot be teased out. Supplementary Fig. 5 shows this correlation matrix for our database. We can see that the "visual" and "auditory" terms are very anticorrelated (their correlation is -.9). Indeed, our tasks are exclusively either visual or auditory, aside from the `ds114` study in which there is no explicit stimuli. For this reason, we remove the regressor "auditory". The auditory map can be defined as the negated map for the visual term. Other terms suffer from strong correlation, in particular the "voice" and "auditory" terms, as most auditory stimuli are voices. However, some tasks involved non-voice auditory stimuli, such as the *muslang* study (see Supplementary Fig. 2). Using contrasts, as detailed below, can then separate the terms corresponding to multiple different types of auditory stimuli.

**Term contrasts** A GLM estimates responses for each voxel with respect to a



**Supplementary Fig. 4. Term effect for the "place" term.**
The "place" term denotes visual place recognition tasks. As such a task involves viewing images, it recruits also the low-level and mid-level visual areas.

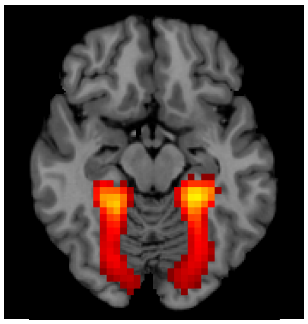



**Supplementary Fig. 5. Terms correlations.**
Correlation matrix between terms across images.

combination of terms. This entails that maps corresponding to the individual term effects show a certain degree of specificity: the effect of that term is *conditional* to the other terms. However, there is shared variance between the terms. To better isolate cognitive processes, a standard analysis in individual studies relies on contrasts in the GLM, e.g., a "face versus place" and a "face versus scrambled picture" contrast for a face recognition study. To disentangle the experimental factors without a too strong a priori on the control conditions, the alternative is to contrast a $\beta$ map against all others, e.g., "face versus place and scrambled picture". To define such contrasts in a systematic way for the wide array of cognitive concepts touched in our database, we use the categories of our ontology. We form groups of terms within the task categories described in Supplementary Table 4: these are used to define the conditions and their controls. Inside each group, we perform a GLM analysis with all the "one versus all" contrasts. We denote these *ontology contrasts*. Note that we do not perform a 3rd level analysis salimi2009 in the sense that the term effects are estimated directly from the subject-level maps, jointly across all studies.

**Other regression approaches** As outlined by one reviewer another potential approach to drawing relationships between cognitive concepts and brain activity is to rely on Partial Least Squares of Canonical

**Supplementary Fig. 6. Ontology contrasts for the "place" term.**
We contrast the "place" with other visual recognition tasks as defined in Supplementary Table 4: recognizing faces, objects, and scrambled images.
The contrast is efficient at suppressing low-level visual areas, but does not completely remove mid-level visual areas. Indeed, mid-level features are probably not balanced across studies, as some objects with no background, some full pictures of objects, and some cropped pictures.



Correlation Analysis methods – or more precisely, their predictive variants, namely reduced rank regression. These methods typically find combinations of terms that are highly correlated with combination of regional activities. However, they tend to combine many terms to form their prediction, creating latent factors –"loadings"– distributed across labels. In the present work we prefer to rely on term-specific mappings that avoid the additional difficulty of studying the cognitive loadings of the obtained components. The combination across terms is then done explicitly through contrasts and discriminative models.

## References

1. Salimi-Khorshidi G, Smith SM, Keltner JR, Wager TD, et al. Meta-analysis of neuroimaging data: a comparison of image-based and coordinate-based pooling of studies. Neuroimage. 2009;45:810.

# 4 Reverse inference: Decoding with cognitive ontologies

**Linear decoders with good map recovery** We want to build a linear model to be able to map the predictive features onto the brain. Feature recovery is the ability to recover stable and meaningful predictive features from our model. Three issues usually get in the way in fMRI multivariate analyses: the high dimensionality of the data, the local correlation of the features –voxels–, and model selection. [1] show that it is possible to come around the dimensionality and correlation problems by using sparse regression models with randomization techniques and feature clustering. This actually amounts to building an ensemble of sparse linear classifiers [2], on a set of randomized parcellations generated by a Ward agglomerative clustering algorithm combined with a resampling method. Note that parcel-averages of the signals are used in the next steps. We add a cross validation procedure in the training of our ensemble in order to select the model. For each random parcellation, we keep the best model. Ensemble classifiers typically either use a voting or an averaging strategy for the final prediction. We choose the latter to keep a linear model, in line with our brain mapping goals. We also perform a non-conservative

univariate screening of the features, and keep 30% of the features. This step is primarily due to computational concerns. On the specifics of our model, we choose to use an $\ell_1$-logistic regression, and 5K parcels for the clustering. We run 5-fold cross validation for model averaging. For each fold, following [3, 4], we set the $\ell_1$ regularization parameter to minimize the error on the left-out data. We average the resulting models.

**Class imbalance and rare classes problem** The class imbalance problem is inherent to our data since mental processes are not uniformly investigated in the literature, and even less so in our database. This is a common problem for meta-analyses, known as the literature bias. There are several ways to account for class imbalance such as using resampling methods or decomposition strategies to project the classes samples into a balanced space. We choose to use a resampling method akin to bagging (Bootstrap AGGregatING), in which each classifier is given a balanced sub-sample of the whole dataset. This results in an ensemble of classifiers that retains a good coverage of the majority class but suffers less from the imbalanced class distributions.

**Hierarchical decoding: using the ontology for an intermediate feature space** The previous paragraphs describe the necessary steps to build a classifier for a single label, *i.e.* a single term, but we are in a multi-label classification setting. The usual approach to solve this kind of problem in machine learning is to train one binary classifier per label in a One versus All (OvA) scheme. The approach has successfully been used in our initial contribution [5], but in our opinion suffers from two main limitations in this context. First an OvA classification models each label separately, and by doing so misses potentially useful connections between the labels that could improve their individual prediction. Second, it ignores the experimental design of the studies from which the images are drawn: an OvA approach uses blindly all the data to learn a label, regardless of whether the images are from a study designed to expose this kind of mental process.

We introduce a new model to alleviate these shortcomings, that relies on **stacked regressions** [6]. A stacked regression model is an ensemble method that uses the linear combinations of different classifiers to improve the final prediction. The general idea of this model is to generate different predictors on the same data. The predictors can be generated through resampling methods, or merely use different underlying models (*e.g.* to combine a collection of linear and non-linear models). We *stack* the decision functions from this *first-level* collection of classifiers, and use them to train a final, *second-level*, predictor that forms a linear combination of the base models. This model has the advantage of building a linear classifier if we avoid introducing non-linearities in the ensemble classifiers. Another interesting property is that it enables to use classifiers that do multi-class prediction, *ie* choose *one* label, to perform multi-label classification, *ie* predict the presence or not of multiple labels. It does so by combining their predictions. Finally, the first level may be seen as a supervised dimensionality reduction method, as we condense the original space to a number of dimensions equal to the number of base classifiers in the ensemble. Note that as all classifiers combined are linear, the resulting complete model is also a linear model, which means that its weights form brain maps.

**Software aspects** Standard preprocessing was performed with SPM [7]. The ontology-informed decoder as well as the other decoding experiments were implemented

| Category classifier | Terms classifier |
|---|---|
| Modality terms vs all | visual vs auditory<br>visual vs all<br>auditory vs all |
| Sounds terms vs all | human voice vs sound<br>human voice vs all<br>sound vs all |
| Retinotopy terms vs all | vertical vs horizontal checkerboard<br>horizontal checkerboard vs all<br>vertical checkerboard vs all |
| Object recognition terms vs all | faces vs places & objects & scramble<br>places vs (faces & objects & scramble)<br>objects vs (faces & places & scramble)<br>scramble vs (faces & places & objects)<br>faces vs all<br>places vs all<br>objects vs all<br>scramble vs all |
| Symbol recognition terms vs all | words vs digits<br>words vs all<br>digits vs all |
| Hands vs all | left vs right hand<br>left hand vs all<br>right hand vs all |
| Feet vs all | left vs right foot<br>left foot vs all<br>right foot vs all |
| Saccades & calculation vs all | saccades vs calculation<br>saccades vs all<br>calculation vs all |
| No category classifier | language vs all |

**Supplementary Table 2. First-level classifiers used:** We train three types of classifier to learn the hierarchy of terms: category classifiers (with a OvA approach), and terms classifiers (both with OvA and OvO approaches). The classifiers' decision functions span an intermediate feature space tailored to our ontology, upon which we perform a standard OvA approach to predict our labels.

using classifiers from scikit-learn [8] with the nilearn toolbox [9] for data preparation steps.

## References


1. Varoquaux G, Gramfort A, Thirion B. Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering. ICML. 2012;.
2. Dietterich TG. Ensemble methods in machine learning. In: Multiple classifier systems. Springer; 2000. p. 1–15.
3. Hoyos-Idrobo A, Schwartz Y, Varoquaux G, Thirion B. Improving sparse recovery on structured images with bagged clustering. In: Pattern Recognition in NeuroImaging (PRNI), 2015 International Workshop on. IEEE; 2015. p. 73–76.
4. Hoyos-Idrobo A, Varoquaux G, Schwartz Y, Thirion B. FReM - Scalable and stable decoding with fast regularized ensemble of models. Neuroimage; 2017 S1053-8119(17)30818-2.
5. Schwartz Y, Thirion B, Varoquaux G. Mapping paradigm ontologies to and from the brain. Advances in Neural Information Processing Systems. 2013; p. 1673–1681.
6. Breiman L. Stacked regressions. Machine learning. 1996;24:49.
7. Wellcome Department of Cognitive Neurology. SPM8; 2008. http://www.fil.ion.ucl.ac.uk/spm.
8. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825.
9. Abraham A, Pedregosa F, Eickenberg M, Gervais P, Mueller A, Kossaifi J, et al. Machine learning for neuroimaging with scikit-learn. Frontiers in Neuroinformatics. 2014;8:14.

# 5 Consensus between forward and reverse inference

By taking into account several cognitive concepts at the same time, reverse inference maps are more specific than the ones from forward inference, but may also capture irrelevant noise. Indeed, regions that are not marginally[1] linked to the concept, e.g. noise regions, can be included because, conditioning on them removes noise [1]. These regions are not linked to the concept of interest in a forward inference, even with a low threshold. We thus want to use forward inference to remove them from reverse inference, capturing the consensus between the two approaches, as in Supplementary Fig. 7.

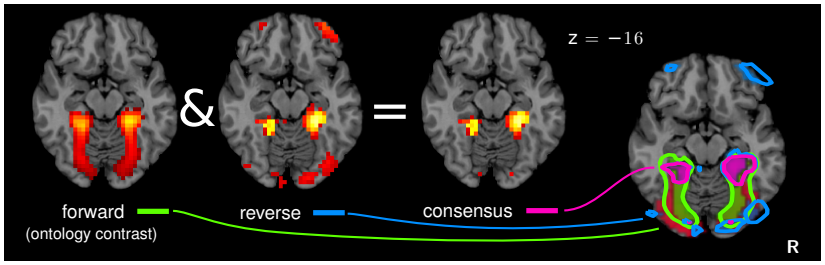
However, using both inferences in conjunction is not straightforward, as they do not perform the same statistical tests and do not have the same statistical power. As we are only interested in the common patterns between both approaches, we use a noise independent procedure to delineate those patterns. Specifically, we compute z-scores for the classifier coefficients by dividing the raw coefficients by their standard error (obtained by cross-validation). The scores' distributions are displayed on the right of Supplementary Fig. 8, and shows the difficulty to find a scale at which to threshold forward and reverse maps to find the common patterns. For this reason, we normalize independently the forward and reverse maps. The left of Supplementary Fig. 8 shows the z-scores' distributions after normalization. From this figure, a fair choice of threshold that yields common patterns lies between $z = 1.5$ and $z = 2$. We mask out the reverse inference maps with those from forward inference using a threshold of 1.5 on the normalized statistic.

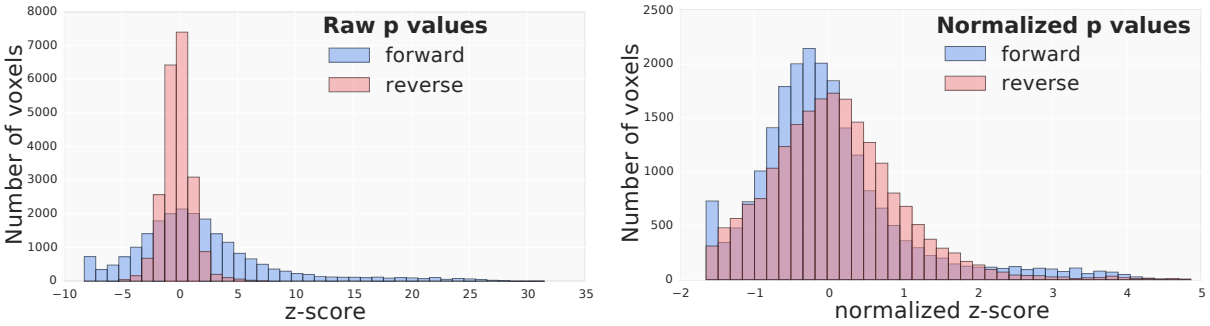
[1]Marginally in the statistical sense: marginal dependence between two variates as opposed to independence conditionally on others.

# References

1. Haufe S, Meinecke F, Görgen K, Dähne S, Haynes JD, Blankertz B, et al. On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage. 2014;87:96–110.

**Supplementary Fig. 7. Maps for consensus between forward and reverse.** Left: maps for the different inferences on the "place" concept. Right: the overlaid inferences for this concept. The consensus singles out the PPA for the "place" concept.




**Supplementary Fig. 8.** Distributions of the z-scores for forward and reverse inference. for the maps related to the *place* concept. **Right**: raw p-values. **Left**: after normalization.

# 6 Evaluating prediction accuracy: cross-validation

## 6.1 Model evaluation procedures

**Cross-validation scheme** To evaluate prediction accuracy, we use a randomized leave-3-study out cross-validation scheme. Using cross-study prediction ensures that the representation of the cognitive labels generalizes across paradigms. Failure to do so might result in over fitting the data, and learning studies idiosyncrasies. This is the first time this type of cross validation is used, as previous multi-study decoding experiments [1,2] relied on a leave-subject out cross-validation. Given the distribution of labels in the database (Supplementary Fig. 2), a left-out study only represents a fraction of labels. To measure the prediction error better, we leave out 3 studies in the test set. However each fold enables to test only a subset of the terms. We complete 100 iterations of the cross validation to get a good estimate of the classifiers performance even for the minority classes.

We give results for both for our leave-3-study out cross-validation and for a simpler cross-validation scheme in which left out subjects are drawn randomly from the database

**Error metric: AUC for ROC** We report the area under the curve (AUC) for the ROC (receiver operator characteristic). This metric summarizes the fraction of misses and false detections on the labels when varying the bias on the decision of the

classifier: biasing to a large number of predicted labels to minimize the misses, or conversely being conservative and risking false detections. It is a standard metric used in machine learning to evaluate performance for unbalanced problems. Indeed, for rare classes the compromise between misses and false detections is difficult to capture by reporting only the number of errors. This number is not affected by class imbalance, and it can thus be compared across our various labels.

**Other classifiers** We also provide classification scores for other common decoders not relying on the ontology: a logistic regression and a naive Bayes classifier. The logistic regression is a linear model, very close to the much used linear SVM (that gives similar results and maps, as its mathematical formulation is not very different). The naive Bayes classifier models voxels as independent, and thus leads to univariate estimation of the weights (though multivariate prediction from them).

## 6.2 Prediction accuracy results

Supplementary Fig. 9 summarizes results for prediction accuracy. Supplementary Table 3 gives details for each class.

We find that imposing the ontology structure is beneficial when predicting to new studies, but not when predicting to new subjects from the same studies. This comes from the fact that when predicting in a given study, there always exists in the training dataset an activation map close to that of the new subject. Thus forcing to focus on the difference between labels is counter productive. The classifier equivalent to our ontology aware classifier, but without the ontology, ie a simple logistic regression, performs best.

The Naive Bayes classifier has an overall performance below that of linear models (Supplementary Table 3), suggesting that due to its univariate nature, it cannot capture distributed patterns of activity to predict cognitive labels. In other words, different concepts leads to overlapping brain activity patterns. Estimating them in a linear model, that captures the dependence between voxels (leading to partialing out the activation of other voxels for each voxel), is beneficial for prediction.

# References

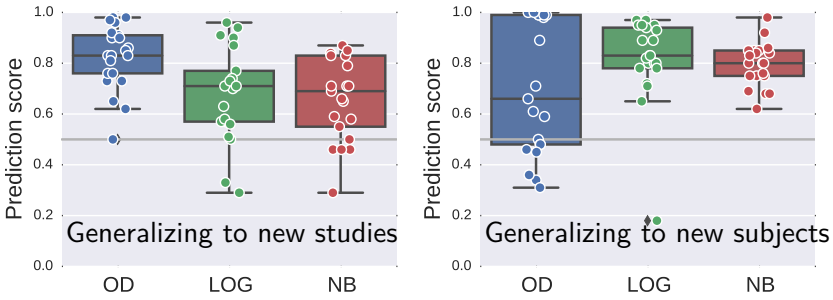
1. Poldrack RA, Halchenko YO, Hanson SJ. Decoding the large-scale structure of brain function by classifying mental states across individuals. Psychol Sci. 2009;20:1364.

2. Poldrack RA, Barch D, Mitchell J, Wager T, Wagner A, Devlin J, et al. Towards open sharing of task-based fMRI data: The OpenfMRI project. Front Neuroinform. 2013;7:12.

| | | visual | auditory | digits | words | V checkerboard | H checkerboard | objects | scramble | faces | places |
|---|---|---|---|---|---|---|---|---|---|---|---|
| New studies | OD | **0.98** | **0.96** | **0.65** | **0.83** | **0.90** | **0.91** | **0.73** | 0.62 | **0.97** | **0.98** |
| | LOG | 0.96 | 0.94 | 0.63 | 0.70 | 0.57 | 0.57 | 0.58 | 0.77 | 0.90 | 0.91 |
| | NB | 0.84 | 0.83 | 0.39 | 0.66 | 0.84 | 0.83 | 0.59 | **0.79** | 0.71 | 0.85 |
| | NS | 0.87 | 0.92 | – | 0.67 | – | – | 0.61 | 0.73 | 0.91 | 0.90 |
| New subjects | OD | 0.31 | **0.99** | 0.61 | 0.46 | **0.99** | **0.99** | 0.66 | 0.45 | **1.00** | 0.36 |
| | LOG | **0.97** | 0.97 | **0.80** | **0.83** | 0.80 | 0.80 | **0.72** | **0.82** | 0.95 | **0.95** |
| | NB | 0.85 | 0.85 | 0.75 | 0.68 | 0.85 | 0.84 | 0.62 | 0.80 | 0.75 | 0.85 |

| | | right hand | left hand | right foot | left foot | language | voice | sound | classification | congruent | loss | gain |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| New studies | OD | 0.83 | **0.92** | **0.86** | **0.90** | **0.85** | 0.83 | **0.73** | **0.76** | 0.50 | **0.81** | **0.76** |
| | LOG | **0.87** | 0.73 | 0.74 | 0.29 | 0.71 | 0.56 | 0.50 | 0.51 | 0.33 | 0.73 | 0.71 |
| | NB | 0.46 | 0.55 | 0.58 | 0.50 | 0.65 | **0.87** | 0.69 | 0.71 | **0.71** | 0.46 | 0.46 |
| | NS | – | – | – | – | 0.78 | – | – | – | – | – | – |
| New subjects | OD | **0.98** | 0.50 | **1.00** | **1.00** | 0.48 | 0.59 | 0.34 | **0.99** | 0.48 | **0.89** | 0.71 |
| | LOG | 0.94 | **0.89** | 0.89 | 0.95 | **0.83** | **0.93** | 0.65 | 0.71 | 0.18 | 0.78 | 0.78 |
| | NB | 0.69 | 0.83 | 0.92 | 0.98 | 0.68 | 0.84 | **0.75** | 0.76 | **0.86** | 0.80 | **0.80** |

**Supplementary Table 3. Prediction scores for different methods**: AUC (area under the curve) of the ROC curve. OD: ontology decoding, LOG: logistic regression, NB: Naive Bayes, NS: NeuroSynth. The OD (ontology decoding) method performs very well (chance is at .5), including when predicting to new studies. Leave-subject-out cross-validation scheme tend to display a higher prediction score than with a leave-study-out cross-validation. This higher prediction accuracy corroborates the observation that activations in the same study are more similar than activations related to the same cognitive term (fig 3).




**Supplementary Fig. 9. Prediction scores for different methods**: AUC (area under the curve) of the ROC. OD: ontology decoding, LOG: logistic, NB: Naive Bayes. Left: leave-study-out cross-validation, Right: leave-subject-out cross-validation.

| CogPO Categories | Task Categories | Terms | % studies | # subjects |
|---|---|---|---|---|
| Stimulus modality | - | visual | 93% | 791 |
| | | auditory | 37% | 478 |
| Explicit stimulus | Sounds | human voice | 23% | 422 |
| | | sound | 20% | 156 |
| | Retinotopy | vertical checkerboard | 10% | 215 |
| | | horizontal checkerboard | 10% | 228 |
| | Object recognition | faces | 23% | 209 |
| | | places | 20% | 188 |
| | | objects | 67% | 460 |
| | | scramble | 20% | 157 |
| | Symbol recognition | words | 47% | 555 |
| | | digits | 7% | 95 |
| Response modality | Motor - hands | left hand | 20% | 321 |
| | | right hand | 33% | 415 |
| | Motor - feet | left foot | 3% | 66 |
| | | right foot | 7% | 77 |
| | Arithmetics | saccades | 7% | 84 |
| Instructions | Arithmetics | calculation | 17% | 369 |
| Cognitive Atlas term | No category | language | 40% | 509 |

**Supplementary Table 4. Terms and categories we use to characterize tasks associated with images in our database.** We used CogPO categories for task-related description, and add necessary terms from Cognitive Atlas to describe higher-level cognitive aspect. Here we report only terms that were present in more than one study –aside from the "left foot", which maps in the analysis as maps in "feet" task category, but not "right foot". The task categories group terms typically used as conditions and their controls to test a hypothesis. The *stimulus modality* category stands for CogPO and task categories. Some terms do not belong to any task category and are referred as such. The *arithmetics* task category spans across the *response modality* and *instructions* CogPO categories.