HHS Public Access
Author manuscript
Nat Neurosci. Author manuscript; available in PMC 2022 May 11.

Published in final edited form as:
Nat Neurosci. 2021 December ; 24(12): 1733–1744. doi:10.1038/s41593-021-00948-9.

A data-driven framework for mapping domains of human neurobiology

Elizabeth Beam, PhD[1,2,3], **Christopher Potts, PhD**[4], **Russell A. Poldrack, PhD**[1,2], **Amit Etkin, MD, PhD**[1,3,5,*]

[1]Wu Tsai Neurosciences Institute, Stanford University, Stanford, CA 94305, USA

[2]Department of Psychology, Stanford University, Stanford, CA 94305, USA

[3]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA

[4]Department of Linguistics, Stanford University, Stanford, CA 94305, USA

[5]Alto Neuroscience, Inc., Los Altos, CA 94022, USA

Abstract

Functional neuroimaging has been a mainstay of human neuroscience for the past 25 years. Interpretation of fMRI data has often occurred within knowledge frameworks crafted by experts, which have the potential to amplify biases that limit the replicability of findings. Here, we employ a computational approach to derive a data-driven framework for neurobiological domains that synthesizes the texts and data of nearly 20,000 human neuroimaging articles. Across multiple levels of domain specificity, the structure-function links within domains better replicate in held-out articles than those mapped from dominant frameworks in neuroscience and psychiatry. We further show that the data-driven framework partitions the literature into modular subfields, for which domains serve as generalizable prototypes of structure-function patterns in single articles. The approach to computational ontology we present here is the most comprehensive characterization of human brain circuits quantifiable with fMRI and may be extended to synthesize other scientific literatures.

Introduction

Scientific ontologies are popular tools for automated data synthesis,[1,2] yet relatively few attempts have been made to engineer ontologies in a data-driven manner.[3] Human

Users may view, print, copy, and download text and data-mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms

[*] **Correspondence and requests for materials should be addressed to:** Amit Etkin, Wu Tsai Neurosciences Institute, 290 Jane Stanford Way, Stanford, CA 94305, Phone: (650)725-5736, amitetkin@altoneuroscience.com.

Author Contributions

E.B. conceptualized the structure of the meta-analysis, collected the data, conducted the computational and statistical analyses, and composed the manuscript. C.P. provided guidance on the natural language processing and deep learning approaches, including their text descriptions. R.A.P. and A.E. advised the overall study design and provided input to the manuscript.

Competing Interests

A.E. holds equity for unrelated work in Akili Interactive, Mindstrong Health, and Alto Neuroscience, where he also receives salary. The remaining authors declare no competing interests.

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

neuroimaging seeks to build a framework for understanding how mental processes interrelate with patterns of brain activity. The flow of inquiry, however, has been largely unidirectional – taking mental constructs defined decades earlier in psychology as the premise for brain mapping efforts. As a consequence, neuroimaging studies have often served to reinforce the differences between concepts that originated in psychology, rather than to define new concepts grounded in brain function.[4] The lack of transparent, objective criteria for the structure of knowledge in human neuroscience will ultimately limit the replicability of findings and hinder progress.[5] Indeed, meta-analyses summarizing results across task paradigms have in many cases shown that brain circuits measured at the resolution of fMRI are surprisingly more similar than different between psychological constructs.[6,7,8] In turn, constructs accepted as "natural kinds" have been mapped to heterogeneous neural activation patterns across the literature.[9,10,11] Nonetheless, these meta-analyses have been limited to only addressing a subset of mental constructs or neural structures.

An empirical description of brain function is needed not only to guide the next generation of basic human neuroscience, but also as a biological foundation for classifying psychiatric disease. It is now well appreciated that the symptom-based mental illness diagnoses specified in the *Diagnostic and Statistical Manual* (DSM) are highly comorbid, biologically heterogeneous, and poorly predictive of treatment response.[12,13] The field of biological psychiatry has attempted to address these shortcomings by reframing mental illnesses as variations in the function of basic brain systems. This has yielded one of the most conceptually dominant expert-driven frameworks for mapping circuits, behaviors, and other units of analysis – called the Research Domain Criteria (RDoC) project.[12] At its present stage, the organizational principles of RDoC remain largely untested, the reproducibility of its circuit-function links is unknown, and the performance of this expert-driven framework relative to a fully data-driven one has not been investigated. To assess how well RDoC and the DSM explain structure-function relationships in the human brain, and to contrast these frameworks with a data-driven one, we conducted a comprehensive neuroimaging meta-analysis of 18,155 PET and fMRI articles.

By applying natural language processing (NLP) and machine learning techniques to the collective results of 25 years of human neuroimaging, we sought here to redefine mental constructs in relation to brain activation data, yielding an integrative data-driven framework. NLP is defined for our purposes as a collection of computational approaches for extracting terms from article texts and representing them quantitatively. Neurobiological domains were mapped from links between the terms for mental functions that occurred in article texts and from the brain coordinate data that studies reported. The resulting data-driven framework is evaluated by three ontological principles: reproducibility of circuit-function links, modularity of domains, and generalizability of domain-level information to single studies. These standards were applied further to assess meta-analytic mappings of the RDoC and DSM frameworks. The results demonstrate that the data-driven approach to ontology engineering summarizes scientific knowledge of human brain function in a manner that is more strongly predictive of results across the neuroimaging literature than the current leading frameworks in basic and clinical neuroscience.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Results

Mapping a data-driven framework for neurobiological domains

In engineering a data-driven framework for human neuroscience, we sought to identify the most coherent set of mental functions (from article texts) for describing the most distinct functional brain circuits (from coordinate data). The functions and circuits together comprise what we refer to as neurobiological “domains” (Fig. 1a). Broadly, our approach entailed clustering brain structures into circuits based on the similarity of their co-occurring mental functions, then mapping functions back onto the circuits to maximize the reproducibility of their interrelations. The first step applies a mutual information metric which gives highest weight to the most specific term-structure associations, while the latter applies a correlation metric which gives weight to terms that co-occur consistently and frequently with brain structures, rendering them representative of the semantic content of the domain.

The corpus used to generate the data-driven framework included 18,155 studies with coordinate data. We define a “corpus” (plural “corpora”) as a collection of article full texts. Corpora were assembled from the BrainMap[14] and Neurosynth[15] databases and by web scraping (Extended Data Fig. 1a). The predominant imaging modalities included fMRI (88% of articles) and PET (3.6%), as summarized along with subject demographics for a subset of studies in Supplementary Table 1. We next mined the neuroimaging data, which included 605,292 spatial coordinates in the human brain. Coordinates were processed as described in the Online Methods, then mapped probabilistically in a one-to-many fashion onto 118 gray matter structures in a neuroanatomical atlas spanning the human cerebrum[16] and cerebellum.[17] Each study was labeled with a given structure no more than once, resulting in a total of 625,714 neuroanatomical labels across studies. Data were randomized and split into a training set containing 70% of articles (*n* = 12,708) for generating the data-driven framework and fitting models, a validation set containing 20% of articles (*n* = 3,631) for optimizing model hyperparameters, and a test set containing 10% of articles (*n* = 1,816) for evaluating domain performance. Splitting the data *a priori* enabled us to avoid biased estimation of the reproducibility of circuit-function links in subsequent analyses. For modularity and generalizability analyses, articles were divided into a discovery set equivalent to the training set (*n* = 12,708) and a replication set combining validation and test sets (*n* = 5,447).

To define the set of possible mental functions, we compiled a broad and diverse lexicon of 1,683 words and phrases from public sources such as RDoC,[12] the BrainMap Taxonomy,[14] and the Cognitive Atlas[18] (Supplementary Table 2). Sources for the lexicon were required to be (a) comprehensive in their coverage of mental functions or illnesses, and (b) publicly available, facilitating accessibility and reproducibility of our analyses. The lexicon includes terms for mental constructs (e.g., “emotional memory”), processes (e.g., “retrieval”), percepts and stimuli (e.g., “face”), and task paradigms (e.g., “face identification task”). It omits neuroanatomical structures, which would otherwise be strongly related to the reported brain coordinate data. The mental function terms were extracted from the preprocessed full texts of articles in our corpus, which contained 128,170,267 total words and 4,831,488 occurrences of terms from our lexicon. Our NLP approach was validated by comparing

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

extracted mental functions to annotations manually assigned in BrainMap (Online Methods), demonstrating both qualitative and quantitative similarities (Supplementary Figs. 1-2).

Candidate domains of the data-driven framework were generated through an unsupervised learning approach that drew on insights from information theory (Fig. 1a). To map links between the 1,683 terms for mental functions and 118 brain structures in our neuroanatomical atlas, we computed their co-occurrences across the training set. Co-occurrence values were reweighted by pointwise mutual information (PMI), which captures the extent to which the joint probability of a term and a brain structure occurring in the same article deviates from the expected probability if the two occurred independently of one another. Because PMI is probability based, it results in high values for the most strongly associated functions and structures regardless of the raw number of times they were observed in the corpus. For instance, while "face identification task" is infrequent in article texts and few coordinates are mapped to the left amygdala, their co-occurrence has a high PMI value because they were both observed in the same small subset of articles. Circuits supporting distinctive sets of mental functions were then defined through *k*-means clustering of the brain structures by their PMI-weighted co-occurrences with mental function terms over a range of *k* values from 2 to 50.

Following this, we sought to identify the mental functions best representative of each circuit, which were selected in a manner that would reflect prevalence rates across the literature. We note that this was necessary because PMI gives high weight to connections that are specific and not necessarily common. While the pattern of structure-function connections is appropriate for grouping structures into circuits, a given connection may replicate only a few times in the literature. For the left amygdala, for instance, none of the top 25 terms with the strongest PMI weighting occurred in more than 0.2% of articles. We thus chose to instead assign the top 25 mental function terms to each circuit based on associations across the training set, computed as point-biserial correlations between binary term occurrences and the centroid of occurrences across structures in each circuit. For the circuit containing the left amygdala, the most strongly associated terms were "fear", "emotion", and "memory," which occurred in 10.82%, 18.12%, and 17.74% of articles, respectively.

Next, the number of terms per data-driven domain was determined through a supervised learning strategy. While up to 25 terms were initially assigned to a given circuit, fewer terms may suffice in representing its functional repertoire. To identify the set of terms and structures with the strongest predictive relationships, the number of mental function terms per circuit was determined by how well term occurrences predicted and were predicted by occurrences of structures. For each circuit, logistic regression classifiers were fit on the training set to predict structure occurrences from term occurrences (i.e., forward inference models), and to predict term occurrences from structure occurrences (i.e., reverse inference models), over a range of 5 to 25 mental function terms. The number of mental function terms for each circuit was selected to maximize area under the receiver operating characteristic curve (ROC-AUC) averaged between the forward and reverse inference models in the validation set. Each domain was then named by the mental function term with highest degree centrality of its term-term co-occurrences. The resultant mappings between term lists and brain circuits can be examined interactively over $k = 2$ to 50 domains at

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

http://neuro-knowledge.org, and subsets are printed in Fig. 2 and Extended Data Figs. 2-3. Similar results were obtained using neural networks in place of logistic regression (Extended Data Fig. 4, Supplementary Figs. 3-4).

Evaluating data-driven domains by organizational principles

The data-driven domains were validated by three principles for human brain organization: reproducibility, modularity, and generalizability (Figs. 1b, 1d, 1e; see Online Methods for details). Reproducibility has been called into question by the finding that some structures are so widely activated across tasks as to unreliably predict mental state.[15,19] To evaluate reproducibility (Fig. 1b), we computed the validation set ROC-AUC for logistic regression classifiers predicting structure occurrences combined within circuits (forward inference) and term occurrences combined within word lists (reverse inference). Second, modularity describes the extent to which domains are both internally homogeneous and distinctive from one another. Modularity is thought to constrain neural organization across scales ranging from cells studied in *Caenorhabditis elegans*[20] to distributed networks in humans.[21] However, because task-based neuroimaging studies are limited in the number of mental states they can reasonably induce, it remains unclear whether task-related human brain activity has multiscale modularity. We assessed modularity in the validation set by first assigning each article to the domain with the most similar functions and structures (Fig. 1c), then computing the ratio of mean Dice distance of articles between versus within domain partitions (Fig. 1d). Distance was computed over a vector including term and structure occurrences within each article. Finally, by the principle of generalizability, the circuit-function pattern for each domain is expected to be a representative prototype of the structures and functions reported in single articles. Previous meta-analyses have demonstrated that only a subset of psychological domains have generalizable representations in specialized brain regions.[22,23] We computed generalizability as the similarity of function and structure occurrences in each article to the prototypical function-structure pattern of the domain to which it was assigned, again within the validation set (Fig. 1e). Across all three measures, solutions at every k outperformed the null.

To understand how domain specificity varies with k, we visualized a hierarchy linking the k = 6, 8, and 22 levels which performed well by our evaluation metrics (Fig. 2). Specifically, these levels occurred at maxima of the geometric mean of reproducibility, modularity, and generalizability. Inter-level links were weighted by the Dice similarity of mental functions and brain structures. As k increases across the hierarchy, domains at lower k fractionate into subcircuits with more finely specified mental functions. For instance, memory begins at $k = 6$ as a medial temporal circuit with affective and memory components, then is subdivided at $k = 8$ into a memory domain with a hippocampal circuit and an emotion domain with an amygdalar circuit. The $k = 8$ memory domain is then parcellated at $k =$ 22 into domains for recall and encoding processes, and for semantic and episodic forms. While nesting relationships emerge across k, we acknowledge that this was not guaranteed by our approach, which generates independent solutions at each k level. To quantify the stability of functional mappings across k, we tracked the proportion of overlapping terms associated with each structure (Supplementary Fig. 5). This illustrates an expected dip in stability for the hippocampus and amygdala at $k = 8$, when these structures remap into

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

respective memory and emotion domains. By contrast, stability is maintained at a high level for language structures, which have consistent mappings across the literature and uphold a similar level of term specificity across the hierarchy.

The results up to this point illustrate that our approach captures varying levels of neurobiological specificity. For subsequent analyses, to facilitate between-framework comparisons, we selected a representative k level. We emphasize that the utility of a given clustering solution depends on context,[24] and as such, there is no "optimal" set of domains within the data-driven framework. The number of domains for follow-up analyses was chosen to balance interpretability with performance across our three evaluation metrics. The $k = 6$ solution was selected because it was not only the minimum value along the asymptote for mean ROC-AUC (Fig. 1b), but also the lowest of the top three k to maximize the geometric mean of reproducibility, modularity, and generalizability (Fig. 2a). As our intent here was to facilitate interpretation, we did not set out to statistically compare metrics across k.

Mapping the leading expert-determined knowledge frameworks

Expert-determined frameworks for basic brain function (i.e., RDoC) and psychiatric illness (i.e., the DSM) were mapped in a top-down fashion beginning with their terms for mental function and dysfunction. NLP was applied to translate the language of the frameworks into the language of the human neuroimaging literature, and the resulting term lists were subsequently mapped onto circuits of co-occurring brain structures (Fig. 3a). For RDoC, word embeddings were trained using GloVe[25] on a corpus of 29,828 general neuroimaging articles (Extended Data Fig. 1b, Supplementary Table 3, top). For the DSM, embeddings were trained with the same parameters on a corpus of 26,070 psychiatric neuroimaging articles (Extended Data Fig. 1c, Supplementary Table 3, bottom). Candidate synonyms were identified among terms for mental function or dysfunction (Supplementary Table 2) by the cosine similarity of their embeddings to the centroid of seed embeddings in each domain. This approach yielded synonyms of the RDoC domains with higher semantic similarity to seed terms than a previously published NLP approach[26] (Fig. 3b). Finally, we mapped brain circuits from each term list based on PMI-weighted co-occurrences with brain structures across the full corpus of articles with coordinates ($n = 18{,}155$ articles), restricting circuits to positive values with FDR < 0.01.

A comparison of the $k = 6$ data-driven domains with the expert-determined frameworks reveals notable differences in circuit-function mappings (Fig. 4). First, the data-driven framework offers novel combinations of emotional and cognitive terms in its domains for memory and cognition, which each relate to several domains in the RDoC framework. Likewise, the RDoC domain for social processes relates strongly to both memory and vision in the data-driven framework, indicating that further functional specification may be warranted in RDoC. While the reward domain of the data-driven framework is strongly related to the RDoC domain for positive valence at the FDR < 0.05 threshold, the reward circuitry is defined more specifically by frontomedial regions and the nucleus accumbens. Relations between the data-driven framework and the DSM are sparser, with four of the six data-driven domains exhibiting above-threshold similarity with seven of the nine DSM

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

domains. The memory domain in the data-driven framework is similar to the anxiety domain and the trauma and stressor domain (but not the bipolar domain or depressive domain with asymmetric fronto-subcortical circuitry). The developmental domain in the DSM maps onto two cortical domains of the data-driven framework, including the manipulation domain that also has high similarity with psychotic illness. To verify that inter-framework differences relied on the coordinate data of the data-driven domains, we repeated comparisons with a framework derived from terms alone (Extended Data Fig. 5). The term-based framework had higher similarity with both RDoC and the DSM, indicating that the coordinate data distinguish the data-driven framework to a meaningful degree.

Relating data-driven domains to leading knowledge frameworks

To assess the performance of the data-driven framework against that of the leading frameworks in neuroscience and psychiatry, we related the $k = 6$ solution to RDoC and the DSM across our measures of reproducibility, modularity, and generalizability. Each of these analyses was repeated with $k = 9$ data-driven domains to verify that differences from the nine DSM domains were not driven by domain number (Extended Data Fig. 6).

Establishing domains with reproducible circuit-function links is essential to advancing basic neuroscience and translating it into reliable neuropsychiatric biomarkers. We assessed the reproducibility of circuit-function links by the performance of logistic regression classifiers predicting the occurrence of mental function terms in article text from coordinate data mapped to the structures of our neuroanatomical atlas (Fig. 5). Test set ROC-AUC was higher across domains of the data-driven framework compared to RDoC and the DSM (Fig. 5h). Whereas all data-driven and RDoC domains achieved above-chance ROC-AUC (Figs. 5e-f), the psychotic domain of the DSM did not (Fig. 5g). These results suggest that orienting neurobiological and psychiatric frameworks to the circuits and term lists derived by a data-driven approach could improve the reproducibility of structure-function links. Further results supporting this conclusion were obtained when evaluating performance by F1 score (Supplementary Fig. 6), when using mental function terms to predict brain structures (Extended Data Fig. 7, Supplementary Fig. 7), and when substituting neural networks for logistic regression classifiers (Extended Data Figs. 8-9, Supplementary Figs. 8-9, Supplementary Tables 4-5).

We next evaluated frameworks by modularity, which captures how well domains partition the literature into homogenous and distinct subfields. Articles were partitioned into the most similar domain within the discovery and replication sets (Figs. 6a-f, Extended Data Fig. 10). Consistent with high comorbidity rates[27] and similar neural alterations[28] between affective disorders, there is visible overlap among the bipolar, depressive, and anxiety domains of the DSM (Figs. 6c, f). Modularity was then assessed by the ratio of mean Dice distance of articles between versus within subfields (Figs. 6g-i). The domain-level results exceeded chance for all domains in both splits across the three frameworks. Macro-averaging across domains in each framework, we find that modularity is higher for the data-driven framework compared to RDoC (Fig. 6j). Surprisingly, modularity of the DSM domains exceeded that of the RDoC domains. Whereas the partitioning results support the movement underway to ground psychiatric diagnoses in brain circuits for transdiagnostic mental constructs,[12,13,29]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the between-framework comparisons caution against the assumption that expert-determined domains of brain function will lead to improved ontological modularity.

The third principle we assessed was generalizability, or how well domain prototypes describe the circuit-function links reported in single articles. Similarity to the prototype exceeded chance for both discovery and replication sets across all frameworks tested (Figs. 6k-m), supporting the interpretation that they represent information which generalizes well within the subfields of human neuroscience. Yet, further gains in similarity to the prototype across domains were achieved by both the data-driven framework and RDoC relative to the DSM (Fig. 6n), highlighting the disconnect between current understanding of brain function and the way that mental disorders have historically been categorized. We anticipate that if mental disorders were redefined as disruptions in basic brain circuitry, their information content would better generalize within subfields of the human neuroscience literature.

Discussion

As human neuroscience expands beyond what any individual can reasonably interpret, the field demands computational approaches to ontology that have been developed in biomedical informatics.[2] We demonstrate here that it is feasible to synthesize the results of thousands of articles through text mining and machine learning techniques, organizing the knowledge of the human neuroimaging discipline into domains jointly defined by mental constructs and brain circuitry. Our procedure for modeling domains of brain function can be readily adapted to summarize other neuroscience subfields, such as direct electrophysiological recordings, cellular-level imaging, or tract tracing experiments. Ontologies developed across levels of analysis could ultimately be synthesized into a hierarchical framework for knowledge of brain function, facilitating advances in basic neuroscience and translational research.

Neural circuitry need not align with human intuition for the structure of mental phenomena, and indeed, the brain-based mapping of mental functions we present here yields several results that are unexpected from the psychology literature. Before the advent of neuroimaging, psychology had maintained a circumplex model for emotion as a distinct processing domain with underlying axes for valence and arousal.[30] RDoC reflects the influence of this model in its domain-level categorization of positive and negative affective valence. However, in the data-driven arrangement at the $k = 6$ level, affective processes are integrated into circuits for memory and reward, consistent with the theory of constructed emotion proposed more recently to account for neuroimaging findings.[31] Likewise, the data-driven framework combines stimulus and response processing stages, which have long been dissociated in psychology,[32] under a single manipulation domain. The frontoparietal circuit for this domain coincides with that found to underlie both attention and movement in human neuroimaging and primate electrophysiology.[33] Perhaps most surprising, the cognition domain spans a combination of terms for cognitive control, salience, and emotion which is unrecognizable among the RDoC domains. However counterintuitive, this set of processes has been consistently associated with activation in medial frontal regions across fMRI studies,[22] highlighting the need for a taxonomic structuring of mental constructs that is anchored in neurobiology.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

When boundaries between mental functions or disorders are drawn without respect to brain function, the resulting categories may be too narrow (i.e., over-specified) or too broad (i.e., under-specified) relative to their associated circuitry. This was a problem with the DSM criteria for psychiatric disorders, as evidenced by high comorbidity rates and internal heterogeneity, which ultimately led the National Institute of Mental Health to undertake the RDoC project.[12] Yet, RDoC domains are seen here to display both over- and under-specification. Mental functions in negative valence, positive valence, and arousal and regulation domains of RDoC have similarly high mutual information with the frontomedial cortex and amygdala, suggesting over-specification of functions relative to circuitry. By contrast, the $k = 6$ solution of the data-driven framework combines affective valence and arousal with a temporo-amygdalar circuit under its memory domain, while it retains a distinct reward domain with a frontomedial-accumbens circuit. Conversely, there is evidence that a domain like negative valence in RDoC is under-specified, encompassing constructs that recombine basic elements of memory, reward, and cognitive systems. At the $k = 6$ level of the data-driven framework, domains for memory, reward, and cognition were better predicted by their more narrowly defined circuits.

The comparative analyses undertaken here underscore the need for rigorous ontological standards in neuroscience and psychiatry, as well as for methods sufficiently flexible to hold such standards against ontologies built through varying approaches. Our assessments of reproducibility, modularity, and generalizability indicate that the data-driven framework meets criteria for predictive, discriminant, and convergent validity, respectively. RDoC likewise upheld these ontological standards, though gains in reproducibility and modularity were achieved with the data-driven organization of knowledge. The DSM, by contrast, failed to uniformly achieve predictive validity across its domains. In addition to testing frameworks in terms of ontological principles, standards for clinical utility should be established at the outset for biologically based frameworks intended to reclassify psychiatric diseases. Categories should not only minimize the comorbidities and internal heterogeneity observed for the symptom criteria,[12] but also exhibit predictive value for clinical outcomes and treatment decisions. In this vein, RDoC has been applied to generate dimensional ratings of domains in electronic medical records that were not only associated with genes relevant to psychopathology,[34] but also predictive of time to hospital readmission.[26] We expect that a data-driven classification system for psychiatric disease would facilitate targeting of neuromodulatory treatments to the patients and stimulation sites most likely to respond. Indeed, reclassification of major depression into brain-based subtypes has improved prediction of rTMS response.[35]

Neuroimaging meta-analysis has the advantage of accounting for functional similarities and differences on the whole-brain scale in human subjects, but it is crucial to note that the organization of the data-driven framework may in part reflect the spatial and temporal limitations of MRI and PET. For instance, while the data-driven framework defines a single brain circuit for affective valence in its memory domain, neuronal population recordings in the rodent amygdala reveal functional divisions for positive and negative affective valence.[36] This result does not necessarily contradict the finding that the poles of affective valence are more similar than they are different on the level of six domains, but rather, illustrates how more refined mappings can be achieved through intracranial recording techniques.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

A hierarchically organized ontology spanning levels of analysis is needed to reconcile the results of animal and human work. In addition to constraints on resolution, our meta-analytic approach is limited to spatial mappings, whereas brain function has also been shown to vary across temporal processing stages. The vision and language domains of the $k = 6$ solution collapse across processing streams which have been shown to first extract and subsequently integrate the features of perceptual inputs.[37,38,39] The neuroanatomical structures grouped together under any ontological domain should be seen not only to co-occur with semantically similar mental functions, but also to co-activate in temporal relation. This can be established through multimodal studies that combine techniques of relatively high spatial and temporal resolution, such as fMRI-electroencephalography. It can be accommodated ontologically through directed links between circuit components.

Another important limitation of our meta-analytic approach is the “resolution” of term and coordinate annotations, which were both assigned on the level of the article. This design decision enabled us to annotate a high volume of articles with mental functions and associated brain structures. By contrast, manually curated databases such as BrainMap contain fewer articles annotated more finely, at the level of the experiment. While our approach is susceptible to “blurring” of terms and coordinates co-reported within an article, we expect that with a large corpus, it will nonetheless detect relationships between terms and structures that are reported consistently together across studies. Our approach was developed with issues of domain resolution in mind and has the capacity to map finer-grained associations of function and circuitry by increasing the k of clustering in the initial step. Fig. 2 illustrates a hierarchy linking domains at the $k = 6$, 8, and 22 levels as a proof of this concept.

Beyond the technical constraints imposed by MRI and PET, human neuroimaging is held back by conceptual impasses that limit how results are translated into scientific knowledge. Several of these are reflected in the data-driven framework presented here. First, the data-driven framework is premised on the widespread assumption that mental functions are localizable to non-overlapping sets of neural structures. However, as observed with the evidence for the construct theory of emotion,[31] the circuitry that supports a phenomenologically constant mental function may vary between individuals, within an individual over time, and across environmental contexts. Second, from the reverse perspective, the data-driven framework assumes that every structure in the brain is responsible for a set of semantically related mental functions. Neuroimaging meta-analysis has revealed that a subset of regions such as the insula and anterior cingulate are activated across a wide range of task contexts.[15,19] These regions have been proposed to function atop information processing hierarchies as hubs that switch between and integrate across brain networks.[40,41] Intriguingly, the insula and anterior cingulate have been identified as sites of common vulnerability across psychotic and affective disorders,[42] suggesting that a deeper understanding of their processing roles would have widespread clinical implications. Going forward, human neuroscience will require a knowledge framework which can accommodate qualifications for various classes of structure-function mappings.

Finally, we note that both the human neuroimaging literature and the data-driven framework derived from it are subject to biases in research practice. Researchers may be more inclined

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

to seek and report evidence for mental functions which have been previously well described in psychology, a form of confirmation bias, and to treat these mental functions as stable constructs, a form of reification bias. Region-of-interest analyses may similarly confirm false assumptions about how structures are defined and mapped to function. We expect that each of these biases are evident in the primary data for meta-analysis, including both the terms for mental functions and the brain coordinate data. That said, these biases would likewise affect the reasoning processes of researchers tasked with crafting knowledge frameworks through individual and group-based deliberation. The advantage of our meta-analytic approach is that mental constructs and brain circuits are jointly defined, mitigating reification and confirmation biases by defining mental functions in direct relation to neurobiological data and, in turn, by drawing bounds around neuroanatomical circuits in direct relation to function. Indeed, redefining domains with respect to mental function terms alone was found to yield domains with higher similarity to the expert-determined frameworks (Extended Data Fig. 5). Finally, our approach to systematic meta-analysis reduces the selection bias that affects neuroscientists manually reviewing the massive volume of published findings.

An area of contention in the present-day design of neuroimaging studies is whether to apply existing knowledge frameworks in a top-down manner, or to perform bottom-up analyses that seek to “carve nature at its joints.” The former type of study has been incentivized throughout the history of the National Institutes of Health, which once required grants to define patient cohorts by DSM criteria and currently offers grants for research premised in the RDoC domains. The advantage of directing research with a single knowledge framework is that data can be collected and described consistently, facilitating interpretation across studies. However, the present work suggests that constraining research by expert-determined frameworks may introduce biases which limit progress. We thus encourage future research that leverages bottom-up approaches to derive and evaluate novel frameworks for knowledge in human neuroimaging. For example, an ontology of self-regulation was derived by combining results across tasks, surveys, and real-world outcomes.[43] Unconstrained analyses of primary neuroimaging data will be supported by the publication of multimodal task-based fMRI datasets like the Human Connectome Project, which has enabled the characterization of task-specific[44] and task-general brain circuits.[45]

The data-driven framework is presented here as the first step in what we expect will be an iterative process of knowledge engineering in neuroscience. In future work, it will be important to verify the contents and boundaries of the data-driven domains across levels and modes of neuroscientific analysis. Alongside advances in neurobiology, innovations in ontology mapping should focus on developing more refined categories of brain function, on annotating their relations, and on specifying their dependencies. Moreover, they should build on efforts to reduce quantifiable biases that are underway in basic science,[5] biomedical informatics,[46] and computational linguistics.[47] Neuroscience ontologies may even be leveraged for automated hypothesis generation and formal reasoning analyses if published in a structured form.[48] Going forward, we expect that computational approaches to ontology will pave the way for a new science of synthesis able to accelerate progress across a wide range of disciplines.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Methods

Neuroimaging Corpora

Three corpora of neuroimaging articles were curated from existing databases and by web scraping (Extended Data Fig. 1). All articles were manually inspected by E.B. to ensure they conducted neuroimaging of the human brain using MRI, PET, or other standard techniques. Metadata were gathered manually for a random subset of 250 articles in this corpus (Supplementary Table 1), demonstrating that functional neuroimaging was the predominant study modality (88% fMRI, 4.8% structural MRI, 3.6% PET). The samples had a mean of 43.35 participants ± 88.27 SD, comprised of 18.86 women ± 28.64 SD and 19.26 men ± 31.43 SD. Participants had a mean age of 30.25 years ± 13.35 SD. We note that analyses conducted on these corpora were not performed blind to framework (i.e., data-driven, RDoC, or DSM) or condition (e.g., discovery and replication split).

Neuroimaging Corpus with Coordinate Data—First, a corpus of brain activation coordinates and neuroimaging article full texts (*n* = 18,155; Extended Data Fig. 1a) was collected as the substrate for the data-driven framework and for coordinate-based analyses. Articles were included if they reported coordinates in the human brain in standard Montreal Neurological Institute (MNI) or Talairach space. Coordinates were gathered on the study level from BrainMap[14] (*n* = 3,346), then Neurosynth[15] (*n* = 12,676), and finally by deploying the Automated Coordinate Extractor (*n* = 2,133).[49] Of the studies with brain coordinate data, 13 were excluded because article full texts could not be obtained by automated or manual download.

General Neuroimaging Corpus—A comprehensive corpus of neuroimaging article full texts (*n* = 29,828; Extended Data Fig. 1b) served as the basis for our computational linguistics approach to selecting mental function terms for RDoC (Fig. 3a). Articles were retrieved in response to a PubMed query (Supplementary Table 3, top) and combined with the first corpus. The query returned 31,136 articles, of which 10,713 did not have full texts available by automated scraping methods.

Psychiatric Neuroimaging Corpus—A corpus of human neuroimaging articles enriched with studies of psychiatric illness (*n* = 26,070; Extended Data Fig. 1c) served as the basis for selecting mental function and dysfunction terms for the DSM. Articles were retrieved from PubMed through a query for DSM-5 disorders (Supplementary Table 3, bottom) and combined with those from the first corpus. The query returned 15,914 articles, of which 1,960 did not have full texts available by scraping.

Coordinate Processing

Brain activation coordinates were converted to MNI space if reported in Talairach using the Lancaster transform.[50] FSL atlasquery (version 5.0.10) was deployed within a Python (2.7.13) wrapper to map coordinates to 118 bilateral gray matter structures in probabilistic anatomical atlases of the human cerebrum[16] and cerebellum.[17] To ensure adequate coordinate data for each structure, our atlas combined substructures within the cerebellar lobules, crus, and vermis. Structures were retained if their probability of containing the

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

coordinate exceeded zero, resulting in one-to-many mappings from each coordinate to the brain. Differences in reporting practices were mitigated by binarizing the mappings for each study, meaning that a study could not report data in a given structure more than once.

Natural Language Processing

Text Preprocessing—Article full texts were extracted from PDF or HTML files downloaded through Stanford University subscription services, from PDF files available through the PMC Open Access Subset, or from XML files in the PMC Author Manuscript Collection. In order to tokenize articles, a lexicon was compiled from public sources (Supplementary Table 2). For RDoC and the data-driven framework, the lexicon was limited to terms describing mental processes and the paradigms used to study them. The lexicon for the DSM was extended to terms for psychopathology. Texts and the lexicon were preprocessed using NLTK (3.4.5). Preprocessing included case-folding, removal of stop words and punctuation, and lemmatization with WordNet (3.1). *N*-grams listed in the lexicon were combined with underscores in article texts.

Validation of Term Extraction—Our approach to extracting terms for mental functions from article full texts was validated by relating the extracted terms to the behavioral annotations manually assigned to articles in the BrainMap database. Both term and annotation occurrences were reweighted by term frequency-inverse document frequency (TF-IDF) so that associations would not reflect the length of texts or number of experiments. Pearson correlations were computed across the 3,346 articles in our corpus for which data had been obtained from BrainMap. The top three most strongly correlated terms are plotted for each subdomain in the BrainMap behavioral taxonomy (Supplementary Fig. 1). As an example, the extracted terms “movement,” “motor,” and “sensorimotor” were most strongly correlated with the execution subdomain of the action domain, meaning that they tended to occur in the full texts of articles which had been labeled manually with the execution label in BrainMap. For all terms shown, the correlations to subdomains were positive with FDR < 0.001.

Second, we offer quantitative evidence that our extracted terms are highly similar in meaning to manual annotations (Supplementary Fig. 2). To quantify semantic similarity, we computed one minus the cosine distance between the centroid of GloVe[25] embeddings for extracted mental function terms and for terms in the BrainMap annotations for behavioral domains. Because the cosine angle is not defined for all-zero vectors, articles were required to have at least one extracted term and one BrainMap annotation, leaving 2,842 for analysis. Centroids were computed as the TF-IDF weighted average of embeddings for terms that were either extracted or derived from BrainMap annotations. GloVe embeddings had been trained on the corpus of 29,828 neuroimaging articles, as described in the *Word Embeddings* section below. Permutation testing was performed by shuffling the annotation centroid over its embedding dimension ($n = 100$) for 1,000 iterations. Relative to scrambled embeddings, the semantic similarity between our extracted terms and the BrainMap annotations was higher than expected by chance (FDR < 0.001) for 89.97% of articles.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Word Embeddings—To identify synonyms of seed terms in existing frameworks as shown in Fig. 3a, word embeddings were trained with GloVe[25] on concatenated full texts in the general neuroimaging corpus (for RDoC) or the psychiatric neuroimaging corpus (for the DSM). The GloVe parameters were as follows: embedding dimension = 100, minimum word count = 5, window size = 15 words, iterations = 500.

Computational Ontology

Data-Driven Framework—The data-driven framework was built up from relationships between (1) activation coordinate data mapped to brain structures and (2) terms for mental functions in article full texts. To prevent over-estimation of classifier test set performance, the data-driven framework was generated in a training set that contained 70% of articles (n = 12,708). Co-occurrences were computed between brain structures (n_s = 118) and terms for mental functions (n_t = 1,683) across training set articles, with re-weighting by PMI to reduce the effects of structure and term frequency. PMI was computed as the logarithm of the observed co-occurrence value divided by the expected co-occurrence value. To avoid taking the logarithm of zero, terms and structures with no co-occurrences were assigned a value of zero. Circuits were then mapped by k-means clustering of brain structures by PMI-weighted links with mental function terms over a range of k values from 2 to 50. This range for the number of domains was deemed acceptable because the number we ultimately selected fell well within it and along a smooth curve for the selection criterion. Up to 25 terms for mental functions were assigned to each circuit by the point-biserial correlation (r_{pb}) of the centroid of occurrences of the circuit's structures with binary function term occurrences. This number of terms was chosen to yield term lists similar in length to those previously generated by an NLP approach to synonym detection for the RDoC framework.[26] We did not find sufficient reason to expand the range given that our analyses yielded a number of terms within it for the majority of domains in the data-driven, RDoC, and DSM frameworks. The number of circuits and terms for each circuit was determined based on the validation set performance of classifiers described below under *Classification Approaches: Data-Driven Mapping*. The resulting circuits and lists of their associated terms are what we refer to as "domains." Finally, each domain was named by the mental function in its list with highest degree centrality of term-term co-occurrences across training set articles.

To further validate our choice of k, several metrics were plotted across clustering solutions. First, the modularity (Fig. 1d) and generalizability (Fig. 1e) metrics described below were macro-averaged across domains at each k level. Whereas modularity tends to increase as k increases and domains reduce in size, generalizability decreases slightly. Our choice of k occurs at a nadir of modularity, and yet falls near the peak of generalizability. It is important to note that our approach yields domains at any level of k from 2 to 50 with reproducibility, modularity, and generalizability all higher than expected by chance. We next validated the stability of domain mappings by plotting the proportion of overlapping mental terms assigned to each brain structure between pairs of clustering solutions (Supplementary Fig. 5). The top five terms per structure were identified as those with the highest r_{pb} for the corresponding domain, and proportions were computed from the overlap of those terms at level k and $k + 1$. The majority of brain structures were assigned a stable set of mental function terms. With brain structures organized by their domain assignment at the $k = 6$

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

level, it is apparent that stability is particularly high across structures in domains known to be well localized, such as language.

RDoC and DSM Frameworks—RDoC and the DSM were modeled by translating their terms for mental functions and disorders into the language of the human neuroimaging literature (Fig. 3a). RDoC seed terms were preprocessed from RDoC behaviors, self-report items, and paradigms grouped by domain; DSM seed terms were preprocessed from disorders grouped by the headings of Section II in Edition 5 (which we refer to as "domains"). Candidate synonyms from the lexicon (Supplementary Table 2) were identified by cosine similarity to the centroid of seed terms in each domain over a range of list lengths from 5 to 25 terms. The list that maximized cosine similarity to the centroid of seed terms was selected for each domain. For RDoC, the word embeddings used to compute similarity were trained using GloVe[25] on a corpus of 29,828 article full texts (Extended Data Fig. 1b) that was curated by supplementing the original corpus with the results of a PubMed query for human neuroimaging studies (Supplementary Table 3, top). This yielded synonyms of the RDoC domains with higher semantic similarity to seed terms than a previously published NLP approach (Fig. 3b). For the DSM, embeddings were trained on a corpus of 26,070 psychiatric neuroimaging articles (Extended Data Fig. 1c) enriched with human neuroimaging studies of mental illness retrieved from PubMed (Supplementary Table 3, bottom). Further, overlap between DSM disorders was reduced by requiring seed terms to be unique to domains and candidate synonyms to have a cosine similarity ≥ 0.5. DSM domains presented here were additionally required to contain terms that appeared in at least 5% of articles in the corpus with coordinate data.

After identifying synonyms of the expert framework terms in the neuroimaging literature, co-occurrences were then mapped between brain structures and term lists. Co-occurrences were computed across the corpus of articles with coordinate data ($n = 18{,}155$) and weighted by positive PMI (PPMI). Brain circuits were defined for each domain by retaining structures for which the PPMI was greater than expected by chance (FDR < 0.01) determined by comparison to a null distribution generated over 10,000 iterations of shuffling term list occurrences across articles.

Framework Similarity—Domains of the data-driven framework were related to those of the RDoC and DSM frameworks by the Dice similarity of mental function terms and brain structures (Fig. 4). Dice similarity was computed from binarized vectors of the terms and structures contained in each domain.

A key distinction between the data-driven and expert-determined frameworks is that the former was jointly defined by mental functions and brain circuitry, while the latter were built on functions that were subsequently mapped to circuits. We would therefore expect that a data-driven framework derived from the mental functions in article texts would be more similar to the expert-determined frameworks. Indeed, this is shown to be the case in Extended Data Fig. 5 for term-based domains at the $k = 6$ level. Data-driven domains were first generated by clustering terms according to their PMI-weighted co-occurrences, then sorted by their r_{pb} with the centroid of cluster term occurrences. The number of terms per domain was selected (as for RDoC and the DSM) over a range of 5 to 25

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

according to the semantic similarity of their embeddings with the centroid of those for "seed" terms that resulted from the initial clustering. The embeddings had been trained in the general neuroimaging corpus (Supplementary Fig. 1b) and were the same as those used in generating the RDoC framework. Domain terms were then mapped to brain circuits by FDR-thresholded PPMI. As in Fig. 4, data-driven domains were related to RDoC and DSM domains by the Dice similarity of their mental functions and brain structures (Extended Data Fig. 5a). The similarity of links between frameworks is consistently lower for the data-driven framework defined jointly by terms and brain coordinate data (Extended Data Fig. 5b) relative to the framework defined by terms alone (Extended Data Fig. 5c). The differences averaged across domains between data-driven ontologies are significant for similarity to both RDoC (Extended Data Fig. 5d) and DSM (Extended Data Fig. 5e) domains. These results underscore that incorporating brain activation data at the first stage of our approach enhances the substantive differences of the resulting domains relative to RDoC and the DSM.

Classification Approaches

Logistic Regression—Classifiers were designed to predict the locations of brain activation coordinates reported by articles from the terms for mental functions occurring in article full texts (i.e., forward inference), and in turn, to predict term occurrences from coordinate locations (i.e., reverse inference). Articles with coordinate data were split into sets for training (70%; n = 12,708), validation (20%; n = 3,631) and test (20%; n = 1,816). Classifiers were fit using Scikit-learn (0.21.2) and stored with Pickleshare (0.7.5). All classifiers were fit in the training set with the following hyperparameters optimized over a grid search to maximize validation set ROC-AUC: regularization penalty (L1 or L2), regularization strength (0.001, 0.01, 0.1, 1, 10, 100, or 1,000), and intercept (included or not). The classifiers used in selecting the number of terms per data-driven domain (Fig. 1a, Step 3) were fit over 100 iterations, and those used in evaluating the number of circuits were fit over 500 iterations (Fig. 1a, Step 5). The test set was reserved for the final comparison of the data-driven framework with RDoC and the DSM, each of which scored article texts using a different list of terms for mental functions. The classifiers used in reproducibility analyses were optimized over the hyperparameter grid described above with training for 1,000 iterations.

Neural Networks—All neural network classifiers were fit in PyTorch (1.0.0) with the Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$). The architecture always included 8 fully connected (FC) layers. Batch normalization and a rectified linear unit (ReLU) activation function were applied to the first seven layers, with dropout during training. A sigmoid activation function was applied to the last layer for one-versus-all classification. The classifiers used in selecting the number of terms per data-driven domain were fit with learning rate = 0.001, weight decay = 0.001, neurons per layer = 100, dropout probability = 0.1, batch size = 1,024, epochs = 100. The classifiers used to select the number of data-driven domains (Extended Data Fig. 4a) were fit with the same hyperparameters over 500 epochs. Classifiers used in the comparison between the data-driven, RDoC, and DSM frameworks (Extended Data Figs. 8-9, Supplementary Figs. 8-9) were optimized through a randomized grid search over 100 combinations of the following hyperparameters: learning

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

rate = 10^{-5} to 1 on a log scale, weight decay = 10^{-5} to 1 on a log scale, number of neurons per layer = 25 to 150 by 25, and dropout probability = 0.1 to 0.9 by 0.1. Each classifier was trained for 500 epochs in this grid search, and the optimal hyperparameter combination was selected to maximize ROC-AUC evaluated on the validation set. Classifiers were subsequently trained for 1,000 epochs with the selected hyperparameters. The test set was reserved for the final comparison of the data-driven framework with RDoC and the DSM. Performance in the test set was highly similar overall to that for logistic regression classifiers performing both reverse (Supplementary Table 4) and forward inference (Supplementary Table 5), motivating us to proceed with the logistic regression architecture and its results in subsequent analyses.

Data-Driven Mapping—The data-driven framework was built through the computational meta-analyses outlined in Fig. 1a, which involved evaluating classifier performance at two stages. First, after mapping brain circuits and associating them with terms for mental functions, performance determined the number of terms per circuit. Specifically, classifiers were trained over a range of term list lengths (5 to 25 terms) and number of circuits (2 to 50). Forward inference classifiers predicted occurrences of coordinates mapped to structures in each circuit from occurrences of terms in its corresponding list; reverse inference classifiers predicted term occurrences from structure occurrences. For each circuit, the number of terms was chosen to maximize the mean ROC-AUC of forward and reverse inference classifiers applied to the validation set.

Second, after assigning terms to each circuit, domain reproducibility at each k was assessed by classifier performance. Classifiers were trained again over $k = 2$ to 50 circuits, this time using occurrences grouped by term list and circuit. Forward inference classifiers predicted occurrences of structures summed by circuit from occurrences of terms summed by list across the full set of domains; reverse inference classifiers predicted term list occurrences from circuit occurrences, again across the full set of domains. Input and output data were binarized by mean values across articles. Specifically, binary scores for the mental functions listed under each domain were computed by mean-thresholding term occurrences, then mean-thresholding the sum of terms within each domain list. Our approach to mean-thresholding was to count a term or structure occurrence as one if it occurred in a given article more frequently than in an average article, and to otherwise count it as zero.

Reproducibility—Test set performance was compared for classifiers that based their inputs or outputs on term lists for each domain of the data-driven framework, RDoC, and the DSM. Term lists were used to score article full texts through the following: (1) term occurrences were thresholded by their mean across articles, (2) binarized term occurrences were summed across terms in each list, (3) sums were thresholded by their mean across articles. Neural data for classification models included the binarized occurrences of coordinates within 118 brain structures. Reverse inference classifiers took brain structure occurrences as inputs to predict text scores as outputs (Fig. 5a); conversely, forward inference classifiers took text scores as inputs to predict brain structure occurrences as outputs (Extended Data Fig. 7a). Performance was assessed for each classifier based on overlap of bootstrap and null distributions for ROC-AUC computed in the test set of

Author Manuscript

articles. Bootstrap distributions were computed by resampling articles in the test set with replacement over 1,000 iterations. Null distributions were computed by shuffling true labels for term lists (Fig. 5) or brain structures (Extended Data Fig. 7) across articles, which we expect to be independent, again over 1,000 iterations.

Article Partitioning

The data-driven, RDoC, and DSM frameworks were applied to partition articles across the corpus into subfields of neuroscience. Articles were first split into a discovery set equivalent to the training set (*n* = 12,708) and a replication set that combined the validation and test sets (*n* = 5,447). Each domain was represented by a binary vector with one-valued cells for the terms and brain structures it entails. Likewise, each article was represented by a binary vector with one-valued cells for the terms that occurred in its full text and the brain structures that were mapped from its reported activation coordinates. Articles were assigned to the domain in each framework with the highest Dice similarity of its terms and structures.

Visualization

The Dice distance between articles in the discovery set (*n* = 12,708; Figs. 6a-c) and replication set (*n* = 5,447; Figs. 6d-f) was visualized with multidimensional scaling. Distances were computed between articles in the full corpus (*n* = 18,155) based on the binary vectors for the terms and structures they reported (Extended Data Figs. 10a-c). Metric multidimensional scaling was performed on the square distance matrix (18,155 x 18,155) with epsilon = 0.001 over 5,000 iterations. Subsets of the resulting multidimensional scaling matrix for articles in the discovery (12,708 x 2) and replication sets (5,447 x 2) were visualized with articles colored by their domain assignments in the data-driven framework (Figs. 6a, d), RDoC (Figs. 6b, e), and the DSM (Figs. 6c, f). For additional visualization by *t*-distributed stochastic neighbor embeddings (*t*-SNE), the 18,155 x 18,155 distance matrix was reduced by principal component analysis. The first 10 principal components (18,155 x 10) were taken as inputs to t-SNE, which was trained with perplexity = 25, early exaggeration = 15, learning rated = 500, and maximum iterations = 1,000 (Extended Data Figs. 10d-f).

Modularity—Within the discovery and replication sets, modularity of the partitions was assessed by the ratio of mean Dice distance of articles between versus within domain partitions (Figs. 6g-i). Dice distance was computed over the binary mental function term and brain structure occurrences in each article. Null distributions were computed by shuffling the distance matrix over 1,000 iterations. Bootstrap distributions were computed by resampling distances with replacement over 1,000 iterations (Fig. 6j).

Generalizability—Within the discovery and replication sets, generalizability of each domain to articles within its corresponding partition was assessed by Dice similarity of terms and structures (Figs. 6k-m). Null distributions were computed by shuffling terms and structures in each domain over 1,000 iterations. Bootstrap distributions were computed by resampling articles with replacement over 1,000 iterations (Fig. 6n).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Control Experiments

Inclusion of Coordinate Data—We expect that to some degree, brain coordinate data are correlated with certain mental function terms, potentially limiting the extent to which the coordinate data contribute meaningful information to domain organization. To verify that jointly defining domains by the neural data and function terms yielded results that could not be obtained from the function terms alone, we visualized the latter term-based solution. Specifically, we remapped a six-domain framework by clustering the mental function terms in our lexicon. These term groupings were subsequently mapped onto brain circuits, by contrast to the joint mapping approach presented in Fig. 1. The term-based framework was qualitatively distinct from the jointly defined framework, including domains for emotion and inference while omitting the reward and manipulation domains (Extended Data Fig. 5a). Importantly, its domain contents were more similar to RDoC ($p < 0.001$; Extended Data Fig. 5d) and the DSM ($p = 0.04$; Extended Data Fig. 5e) than those of the data-driven framework incorporating brain coordinate data. These results indicate that the brain coordinate data do indeed help to distinguish the content of data-driven framework from existing frameworks in neuroscience and psychiatry.

Number of Domains—Our evaluations of framework organization effectively controlled for domain number in comparisons with RDoC, which has six domains, as we had selected the $k = 6$ solution of the data-driven framework to present (Figs. 5-6). To control for domain number in comparisons with the DSM, which has nine domains, we repeated analyses using the $k = 9$ data-driven solution (Extended Data Fig. 6). As with the $k = 6$ domain solution, the $k = 9$ solution of the data-driven framework outperforms RDoC and the DSM on reproducibility indexed by both reverse inference (Extended Data Fig. 6a) and forward inference (Extended Data Fig. 6d), RDoC on modularity of domain partitions of the discovery and replication sets (Extended Data Fig. 6i), and the DSM on generalizability of the domain prototypes to discovery and replication set articles (Extended Data Fig. 6j). There were no significant differences between $k = 6$ and $k = 9$ data-driven solutions on the reproducibility or generalizability measures. Modularity was significantly increased for the $k = 9$ solution relative to $k = 6$, consistent with the positive trend with k shown in Fig. 1d. The increase in modularity with k is expected, as at a higher number of domains, articles within each domain's partition should be relatively more similar.

Statistics

Analyses were performed using Python (3.6.8) built-in modules and the following libraries: Matplotlib (3.1.1), Numpy (1.16.4), Pandas (0.24.2), Scikit-learn (0.21.2), Scipy (1.3.0), and Statsmodels (0.10.1). Unless stated otherwise, normality of distributions was not assumed. Statistical tests were nonparametric or derived from permutation testing or bootstrapping. Significance of framework evaluation metrics was determined by permutation tests estimating the probability of obtaining observed values by chance. Frameworks were compared by testing for a difference in means of bootstrap distributions. FDR estimates were computed by correcting p-values according to the Benjamini-Hochberg method.[51]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Data Availability

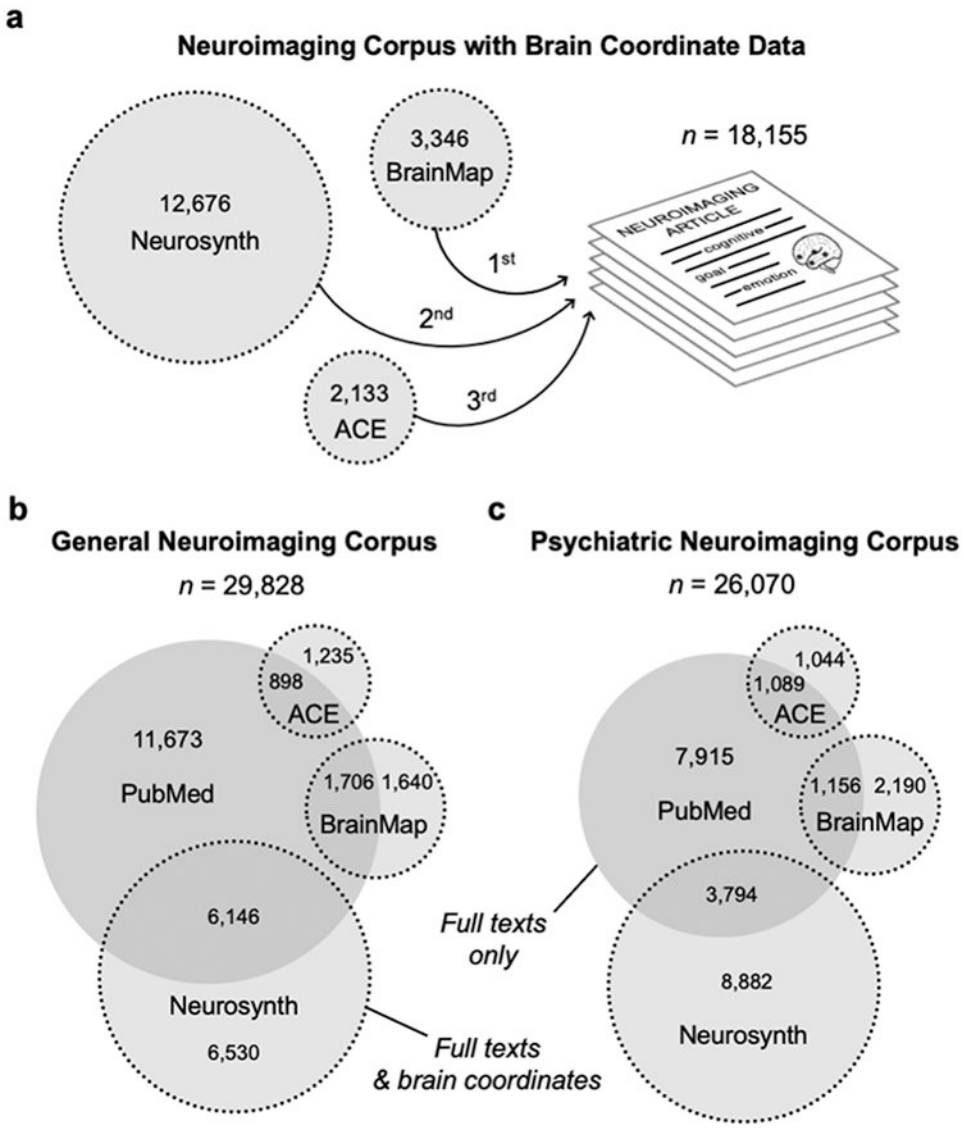
Data not subject to restrictions can be accessed at http://github.com/ehbeam/neuro-knowledge-engine. This repository contains data generated from the corpus of 18,155 human neuroimaging articles, including matrices of the terms and brain structures reported in each document, as well as GloVe[25] embeddings trained on the expanded neuroimaging and psychiatric corpora. Due to copyright restrictions, article PDF files and extracted texts have not been made publicly available. Article metadata including PMIDs are provided so that the corpus contents can be retrieved from PubMed. Subsets of the brain coordinate data were previously made available online by BrainMap (http://www.brainmap.org/software.html) and Neurosynth (http://github.com/neurosynth/neurosynth-data).

Code Availability

The code used to generate and assess knowledge frameworks is available at http://github.com/ehbeam/neuro-knowledge-engine. The code supporting the interactive viewer for the data-driven framework can be accessed at http://github.com/ehbeam/nke-viewer.

Author Manuscript

Extended Data



Extended data Fig. 1. Corpora of human neuroimaging articles

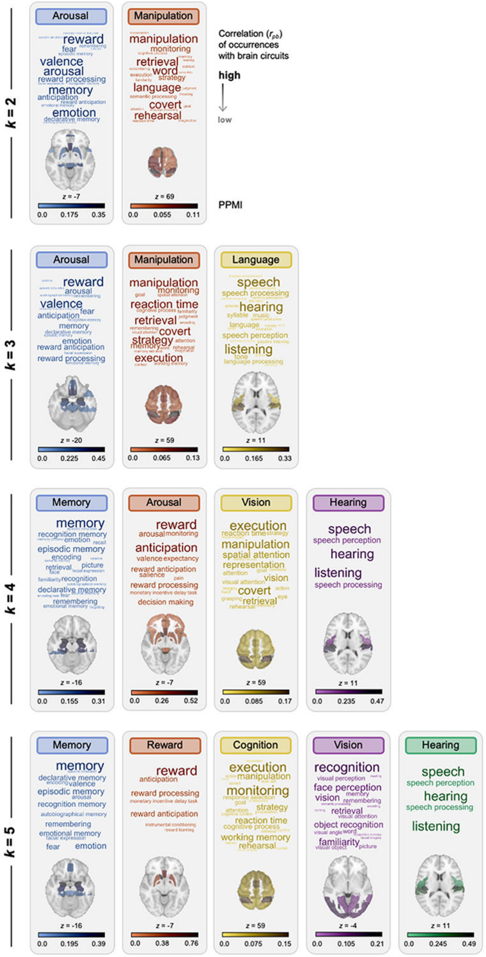
a, Articles reporting locations of activity in the human brain in standard MNI or Talairach space. Article metadata and coordinates were curated first from BrainMap (*n* = 3,346), then from Neurosynth (*n* = 12,676), then by deploying the Automated Coordinate Extractor (*n* = 2,133). **b,** A comprehensive corpus of human neuroimaging articles served as the basis for a computational linguistics approach to selecting mental function terms for the RDoC framework. Articles were retrieved in response to a PubMed query (Supplementary Table 3, top) and combined with those reporting coordinate data. **c,** A corpus of human neuroimaging articles enriched with studies addressing psychiatric illness served as the basis for selecting mental function and dysfunction terms for the DSM framework. As before, articles were retrieved through a PubMed query (Supplementary Table 3, bottom) and combined with those reporting coordinate data.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript



Extended data Fig. 2. Data-driven solutions for *k* = 2 to 5 using logistic regression classifiers to select function terms

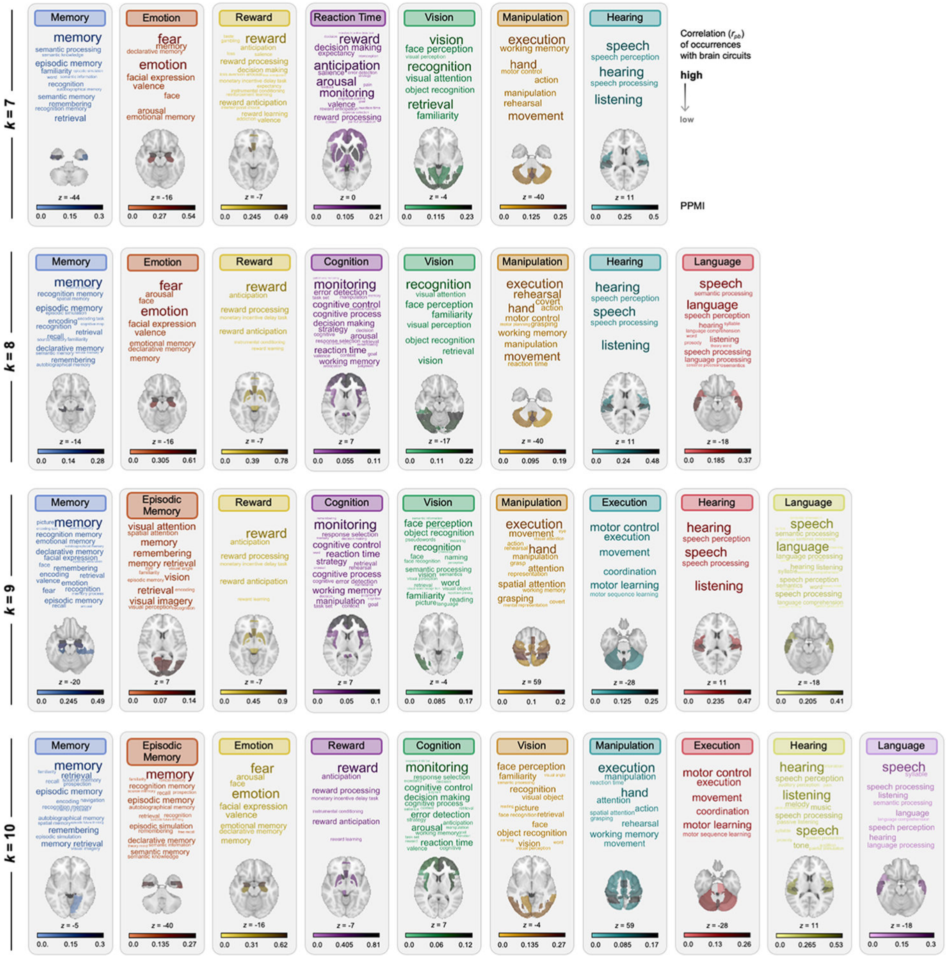
Domains generated at lower values of *k* than selected through the optimization procedure detailed in Fig. 1a.

Author Manuscript

Author Manuscript

Author Manuscript

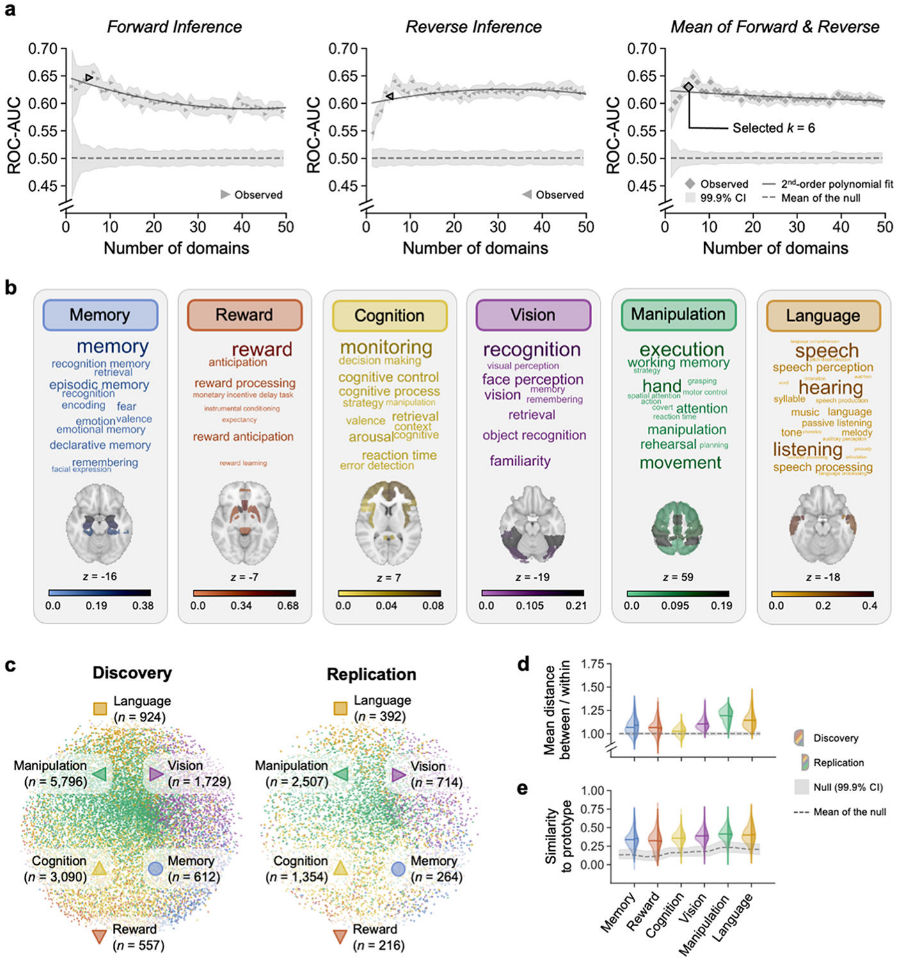
Author Manuscript



Extended data Fig. 3. Data-driven solutions for *k* = 7 to 10 using logistic regression classifiers to select function terms

Domains generated at higher values of *k* than selected through the optimization procedure detailed in Fig. 1a. Additional framework solutions up to *k* = 50 are visualized at http://neuro-knowledge.org.

Author Manuscript

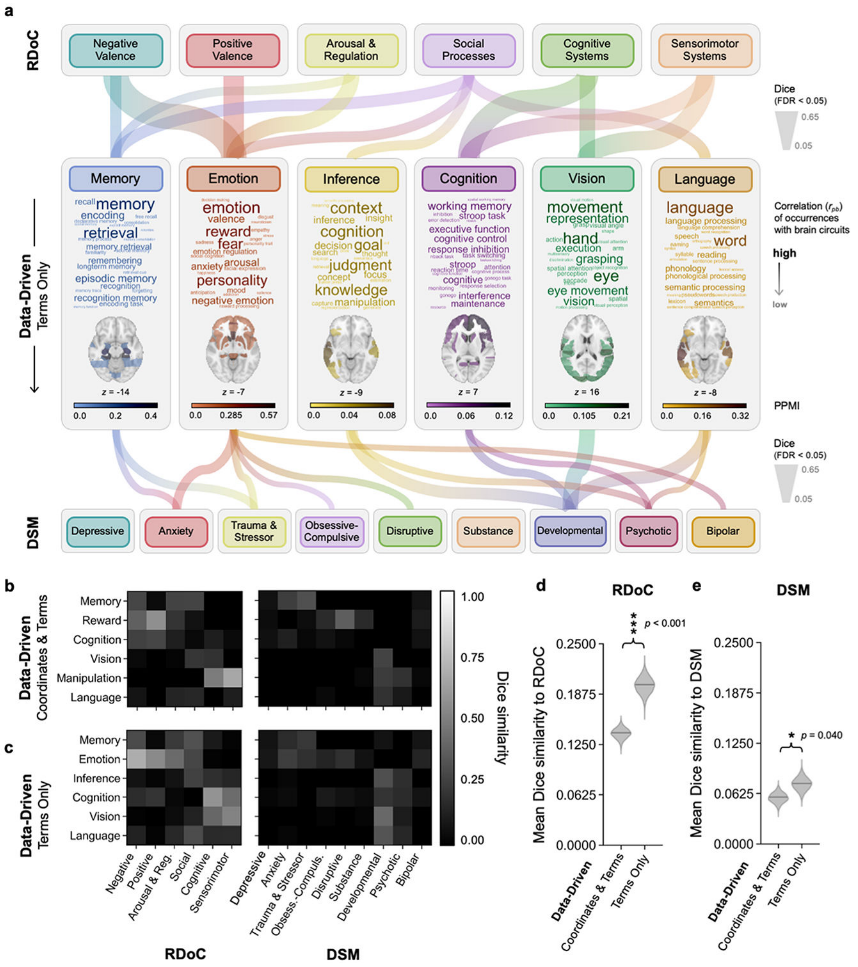


Extended data Fig. 4. Neural network approach to data-driven ontology generation

The procedure in Fig. 1a was repeated using neural network classifiers in place of logistic regression. Neural network classifiers were comprised of 8 fully connected layers and fit with learning rate = 0.001, weight decay = 0.001, neurons per layer = 100, dropout probability = 0.1 (last 3 layers), and batch size = 1,024. In Step 3, neural networks were trained over 100 epochs to predict term and structure occurrences within domains. In Step 4, neural networks were trained over 500 epochs to predict domain term list and circuit occurrences. **a,** Validation set ROC-AUC plotted for forward inference, reverse inference, and their average. **b,** Data-driven solution for *k* = 6. Term size is scaled to frequency in the corpus of 18,155 articles with activation coordinate data. The number of terms per domain was selected in Step 3 to maximize neural network performance in the validation set. Brain maps show structures included in each circuit as a result of clustering by PMI-weighted co-occurrences with function terms. **c,** Article partitioning based on maximal similarity to terms and structures in domain prototypes visualized by multidimensional scaling. **d,** Modularity was assessed by comparing the mean Dice distance of function and structure occurrences of articles between domains versus within domains. Observed values are colored by domain; null distributions in gray were computed by shuffling distance values across article partitions over 1,000 iterations. **e,** Generalizability was assessed by Dice similarity of each domain's

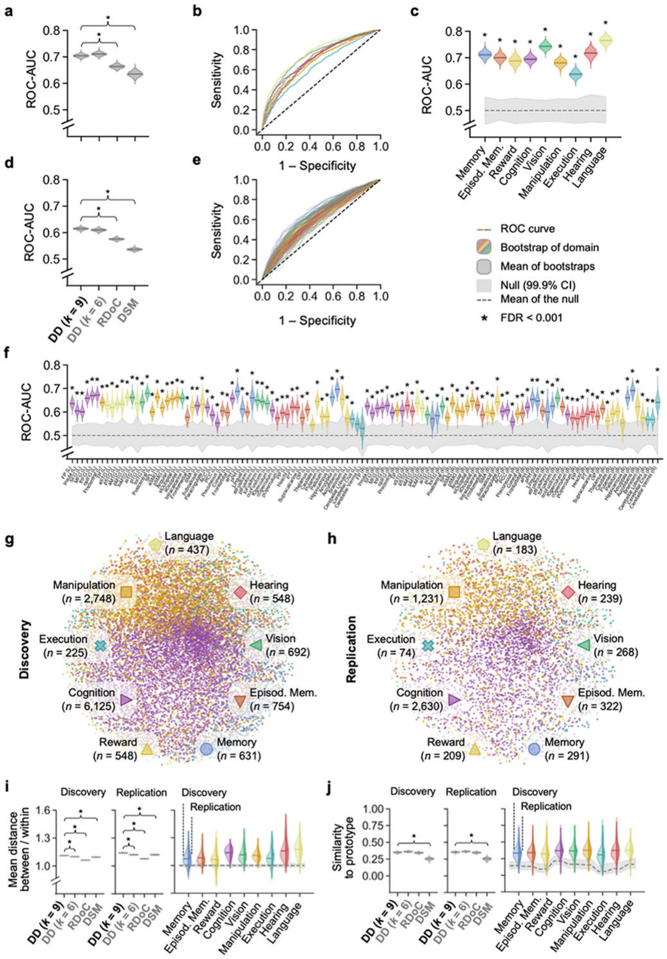
Author Manuscript Author Manuscript Author Manuscript Author Manuscript

"prototype" vector of function terms and brain structures with the occurrences of terms and structures in each article of the domain's partition. Observed values are colored by domain; null distributions in gray were computed by shuffling terms and structures in each prototype over 1,000 iterations.



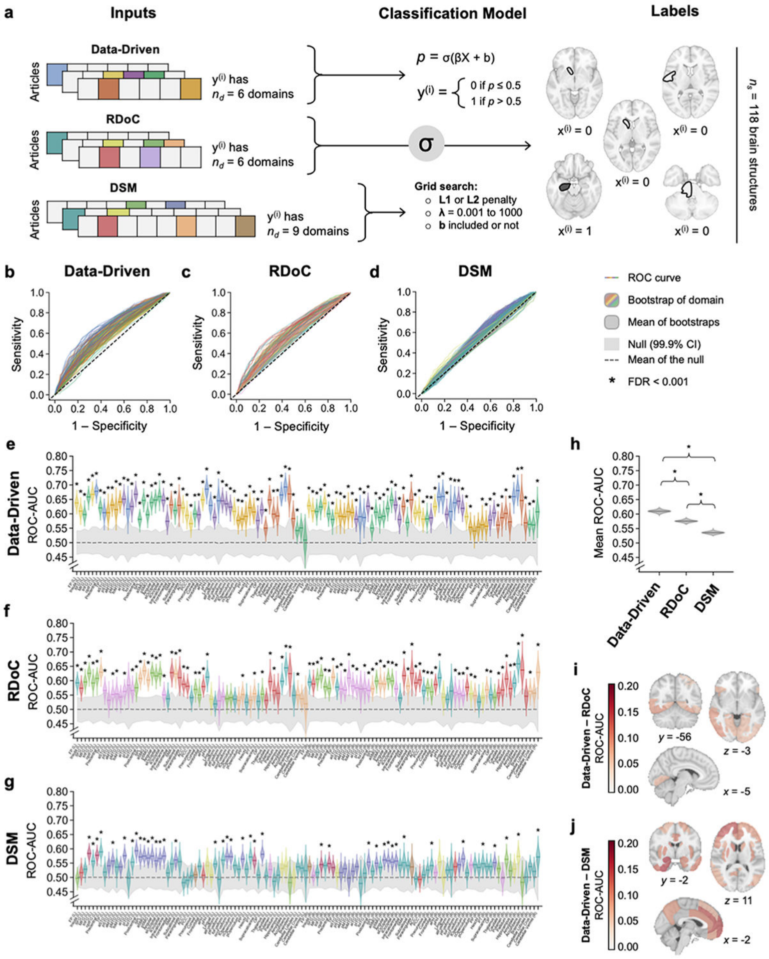
Extended data Fig. 5. Framework generated from mental function terms and subsequently mapped to brain circuits

To control for the contribution of brain coordinate data, the framework was rederived solely from mental function terms. Mental function terms were clustered by *k*-means according to PMI-weighted co-occurrences in the training set of 12,708 articles. The top 25 function terms were assigned to each domain by r_{pb} of binarized occurrences with the centroid of occurrences across "seed" terms from clustering. The number of terms per domain and name for each domain were determined as before. Circuits were mapped from PPMI of brain structures with the centroid of domain terms (FDR < 0.01). **a,** Domains are visualized for $k = 6$, the same dimensionality as the data-driven framework in the main text. Term size is scaled to r_{pb} with the centroid of seed terms. Term-based domains are linked to the RDoC and DSM domains illustrated in Fig. 4. Links between domains were computed across the corpus of 18,155 articles by Dice similarity of mental function terms and brain structures (FDR < 0.05 based on permutation testing over 10,000 iterations). The Dice similarity of links with RDoC and DSM frameworks across the corpus is shown for **b,** the data-driven framework based on brain coordinates and mental function terms (as in Fig. 4), and **c,** the framework based only on terms. Dice similarity with **d,** RDoC and **e,** the DSM was macro-averaged across domains, and a one-sided bootstrap test assessed the difference in means between the data-driven frameworks. The term-based framework was more similar to RDoC than the framework also based on coordinates (99.9% CI = [0.022, 0.098]), and to the DSM (95% CI = [0.001, 0.034]). Bootstrap distributions were computed by resampling function terms and brain structures over 10,000 iterations.



Extended data Fig. 6. Between-framework comparisons with *k* = 9 data-driven domains

To control for dimensionality in comparisons with the DSM, analyses were repeated with the *k* = 9 solution of the data-driven framework. Differences between framework pairs were assessed by two-sided bootstrap tests. With *k* = 9 in place of *k* = 6 domains, no differences in the data-driven, RDoC, and DSM rankings were observed. **a,** Reverse inference ROC-AUC in the test set was higher for the *k* = 9 data-driven framework than both RDoC (99.9% CI of the difference = [0.010, 0.066]) and the DSM (99.9% CI of the difference = [0.029, 0.102]). **b,** ROC curves of the *k* = 9 reverse inference classifiers. **c,** ROC-AUC of the *k* = 9 reverse inference classifiers. **d,** Forward inference ROC-AUC in the test set was higher for the *k* = 9 data-driven framework than both RDoC (99.9% CI of the difference = [0.027, 0.052]) and the DSM (99.9% CI of the difference = [0.061, 0.095]). **e,** ROC curves of the $k = 9$ forward inference classifiers. **f,** ROC-AUC of the $k = 9$ forward inference classifiers. Articles were partitioned into the the $k = 9$ data-driven domains within the **g,** discovery set ($n = 12{,}708$) and **h,** replication set ($n = 5{,}447$). **i,** Domain-averaged modularity (left panels) was higher for the $k = 9$ data-driven framework than for the $k = 6$ solution (99.9% CI of the difference = [0.007, 0.016] discovery, [0.009, 0.024] replication), RDoC (99.9% CI of the difference = [0.046, 0.053] discovery, [0.045, 0.059] replication), and the DSM (99.9% CI of the difference = [0.011, 0.022] discovery, [0.007, 0.026] replication). **j,** Domain-averaged generalizability (left panels) was higher for the $k = 9$ data-driven framework than for the DSM (99.9% CI of the difference = [0.043, 0.160] discovery, [0.039, 0.164] replication). Observed values for the $k = 9$ domains in the **i-j** right panels were compared against null distributions generated by shuffling over 1,000 iterations (* FDR < 0.001).



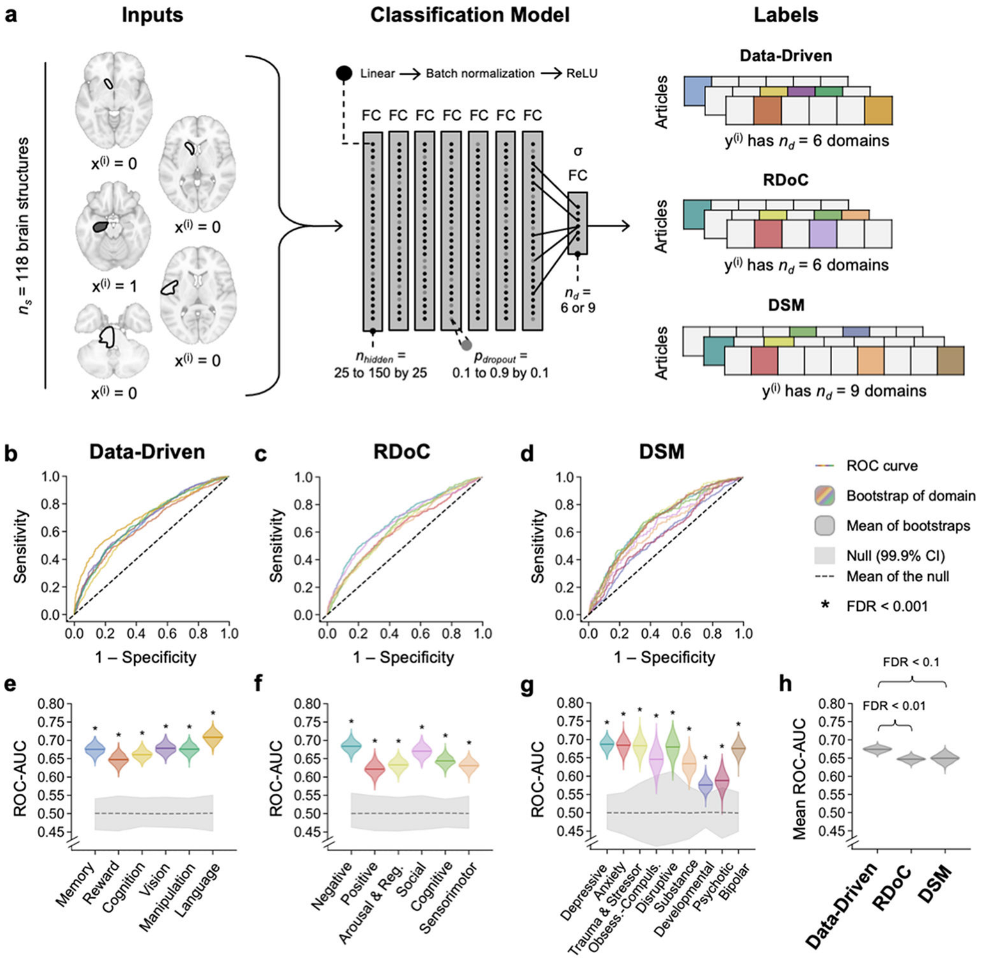
Extended data Fig. 7. Forward inference classification with logistic regression

a, Logistic regression classifiers were trained to predict whether coordinates were reported within brain structures based on the occurrences of mental function terms in full texts. Classifier features included term occurrences thresholded by mean frequency across the corpus, then the mean frequency of terms in each domain. Activation coordinate data were mapped to 118 structures in a whole-brain atlas. Training was performed in 70% of articles

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

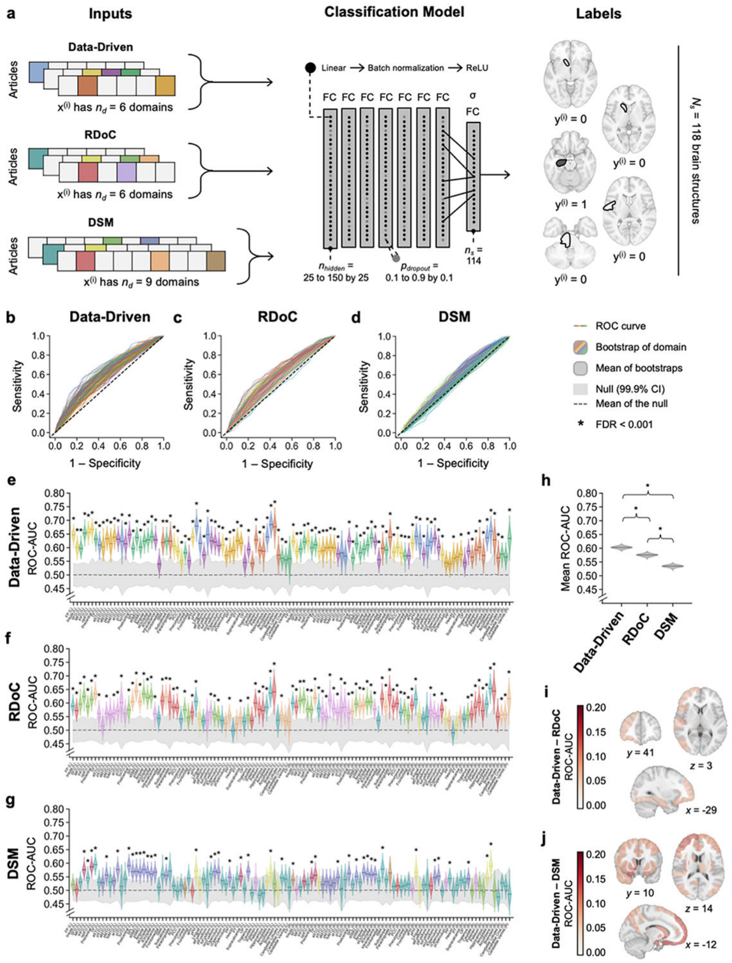
(*n* = 12,708), hyperparameters were tuned on a validation set containing 20% of articles (*n* = 3,631), then classifiers were evaluated in a test set containing 10% of articles (*n* = 1,816). Plots are colored by the domain to which structures were assigned in the data-driven framework, and by the domain with highest PPMI for the RDoC and DSM frameworks. Test set ROC curves are shown for **b,** the data-driven framework, **c,** RDoC, and **d,** the DSM. **e-g,** For each brain structure, the significance of the test set ROC-AUC was determined by a one-sided permutation test comparing the observed value to a null distribution, and the *p* value was FDR-corrected for multiple comparisons (* FDR < 0.001). Observed test set values are shown with solid lines. Null distributions (gray) were computed by shuffling true labels over 1,000 iterations. Bootstrap distributions (colored) were computed by resampling articles in the test set with replacement over 1,000 iterations. **h,** The difference in mean ROC-AUC was assessed for each framework pair by a two-sided bootstrap test. The data-driven framework had higher ROC-AUC than both RDoC (99.9% CI of the difference = [0.020, 0.049]) and the DSM (99.9% CI of the difference = [0.055, 0.091]). RDoC had higher ROC-AUC than the DSM (99.9% CI of the difference = [0.024, 0.058]). Solid lines denote means of the bootstrap distributions obtained by macro-averaging across brain structure classifiers. **i-j,** Difference in ROC-AUC between the data-driven and expert-determined frameworks. Maps were thresholded to show differences with FDR < 0.001 based on permutation testing.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



Extended data Fig. 8. Reverse inference classification with neural networks

Neural network classifiers were trained to perform reverse inference, using brain activation coordinates to predict occurrences of mental function terms grouped by domains shown in Extended Data Fig. 4. Classification models comprised 8 fully connected (FC) layers, all with ReLU activation functions except the output layer which was activated by a sigmoid. The optimal learning rate, weight decay, number of neurons per layer, and dropout probability were determined for each framework through a randomized grid search. ROC curves are shown for the test set performance of classifiers with mental function features defined by **b,** the data-driven framework, **c,** RDoC, and **d,** the DSM. **e-g,** For each domain, the significance of the test set ROC-AUC was determined by a one-sided permutation test comparing the observed value to a null distribution, and the *p* value was FDR-corrected for multiple comparisons (* FDR < 0.001). Observed values in the test set are shown with solid lines. Null distributions (gray) were computed by shuffling true labels for term list scores over 1,000 iterations; the 99.9% confidence interval is shaded, and distribution means are shown with dashed lines. Bootstrap distributions of ROC-AUC (colored) were computed by resampling articles in the test set with replacement over 1,000 iterations. **h,** Differences in bootstrap distribution means were assessed for each framework pair. While there were no differences between frameworks at the 99.9% confidence level, the data-driven framework had higher ROC-AUC than RDoC at the 99% confidence level (99% CI of the difference = [0.007, 0.050]), and higher ROC-AUC than the DSM at the 95% confidence level (95% CI of the difference = [0.0003, 0.049]). Solid lines denote bootstrap distribution means.

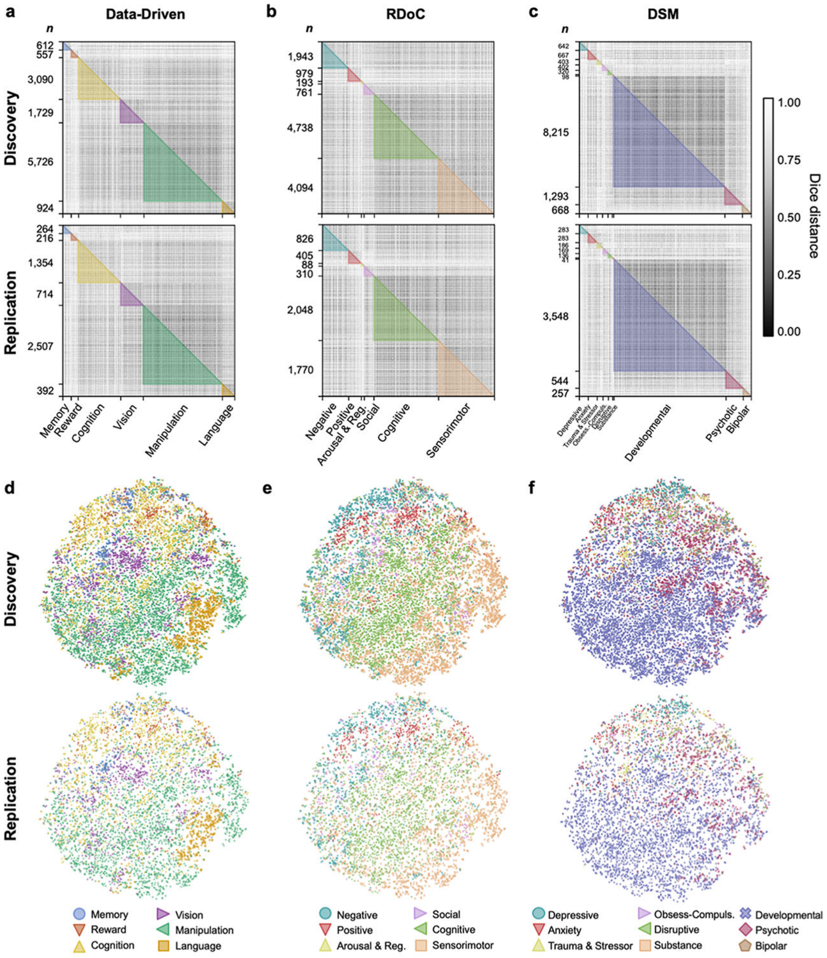


Extended data Fig. 9. Forward inference classification with neural networks

a, Neural network classifiers were trained to perform forward inference, using function term occurrences grouped by the domains in Extended Data Fig. 4 to predict brain activation coordinates. Forward inference classifiers were optimized over a grid search with the same hyperparameter values as reverse inference classifiers in Extended Data Fig. 8. ROC curves are shown for test set performance of classifiers with mental function features defined by **b,** the data-driven framework, **c,** RDoC, and **d,** the DSM. Plots are colored by the domain assignment for structures in the data-driven framework, and by the domain with the highest PPMI for the structure in RDoC and DSM frameworks. **e-g,** For each brain structure, the significance of the test set ROC-AUC was determined by a one-sided permutation test comparing the observed value to a null distribution, and the *p* value was FDR-corrected for multiple comparisons (* FDR < 0.001). Observed values in the test set are shown with solid lines. Null distributions (gray) were computed by shuffling true labels for term list scores over 1,000 iterations; the 99.9% confidence interval is shaded, and distribution means are shown with dashed lines. Bootstrap distributions of ROC-AUC (colored) were computed

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

by resampling articles in the test set with replacement over 1,000 iterations. **h,** Differences in bootstrap distribution means were assessed for each framework pair. The data-driven framework had higher ROC-AUC than both RDoC (99.9% CI of the difference = [0.014, 0.046]) and the DSM (99.9% CI of the difference = [0.049, 0.086]). RDoC had higher ROC-AUC than the DSM (99.9% CI of the difference = [0.018, 0.058]). Solid lines denote bootstrap distribution means. **i-j,** Difference in ROC-AUC scores between the data-driven and expert-determined frameworks. Maps were thresholded to show differences with FDR < 0.001 based on permutation testing.



Extended data Fig. 10. Additional visualizations of article partitioning by domains

Dice distance between articles is shown for binarized vectors of the mental function terms that occurred in the full text and the brain structures to which reported coordinate data were mapped. Articles were split into sets for discovery (*n* =12,708) and replication (*n* = 5,447), then matched to domains based on the Dice similarity of their term-structure vectors. Domain assignments are represented by the color coding scheme established in Fig. 4 for **a,** the data-driven framework, **b,** RDoC, and **c,** the DSM. Shaded areas represent the lower triangle of distances between articles within each domain partition. **d-f,** Dice distance between articles visualized with t-SNE. Distances were computed between the terms and structures of articles in the full corpus (*n* = 18,155), and dimensionality of

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

the 18,155 x 18,155 matrix was reduced by principal component analysis. The first 10 principal components (18,155 x 10) were taken as inputs to t-SNE (perplexity = 25, early exaggeration = 15, learning rate = 500, and maximum iterations = 1,000). Articles are visualized separately for the discovery and replication sets, with colors and shapes corresponding to domain assignments in each framework.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank D. McFarland and S. Baccus for discussing versions of the approaches and figures presented here. Research reported in this publication was supported by the National Institute of Mental Health of the National Institutes of Health under Award Number DP1 MH116506 to A.E. and Award Number F30 MH120956 to E.B. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

1. Bard JBL & Rhee SY Ontologies in biology: design, applications and future challenges. Nat. Rev. Genet 5, 213–222 (2004). [PubMed: 14970823]
2. Ashburner M et al. Gene Ontology: tool for the unification of biology. Nat. Genet 25, 25–29 (2000). [PubMed: 10802651]
3. Alterovitz G et al. Ontology engineering. Nat. Biotechnol 28, 128–130 (2010). [PubMed: 20139945]
4. Price CJ & Friston KJ Functional ontologies for cognition: the systematic definition of structure and function. Cogn. Neuropsychol 22, 262–275 (2005). [PubMed: 21038249]
5. Nuzzo R How scientists fool themselves – and how they can stop. Nature 526, 182–185 (2015). [PubMed: 26450039]
6. Lindquist KA, Satpute AB, Wager TD, Weber J & Barrett LF The brain basis of positive and negative affect: evidence from a meta-analysis of the human neuroimaging literature. Cereb. Cortex 26, 1910–1922 (2016). [PubMed: 25631056]
7. Liu X, Hairston J, Schrier M & Fan J Common and distinct networks underlying reward valence and processing stages: a meta-analysis of functional neuroimaging studies. Neurosci. Biobehav. Rev 35, 1219–1236 (2011). [PubMed: 21185861]
8. Wager TD, Jonides J & Reading S Neuroimaging studies of shifting attention: a meta-analysis. NeuroImage 22, 1679–1693 (2004). [PubMed: 15275924]
9. Siegel EH et al. Emotion fingerprints or emotion populations? A meta-analytic investigation of autonomic features of emotion categories. Psychol. Bull 144, 343–393 (2018). [PubMed: 29389177]
10. Redick TS & Lindsey DRB Complex span and n-back measures of working memory: a meta-analysis. Psychon. Bull. Review 20, 1102–1113 (2013).
11. Binder JR, Desai RH, Graves WW & Conant LL Where is the semantic system? A critical review and meta-analysis of 120 functional neuroimaging studies. Cereb. Cortex 19, 2767–2796 (2009). [PubMed: 19329570]
12. Insel T et al. Research Domain Criteria (RDoC): Toward a new classification framework for research on mental disorders. Am. J. Psychiatry 167, 748–751 (2010). [PubMed: 20595427]
13. Stephan KE et al. Charting the landscape of priority problems in psychiatry, part 1: classification and diagnosis. Lancet Psychiatry 3, 77–83 (2016). [PubMed: 26573970]
14. Fox PT & Lancaster JL Mapping context and content: the BrainMap model. Nat. Rev. Neurosci 3, 319–321 (2002). [PubMed: 11967563]
15. Yarkoni T, Poldrack RA, Nichols TE, Van Essen DC & Wager TD Large-scale automated synthesis of human functional neuroimaging data. Nat. Methods 8, 665–670 (2011). [PubMed: 21706013]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

16. Desikan RS et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. NeuroImage 31, 968–980 (2006). [PubMed: 16530430]

17. Diedrichsen J, Balster JH, Cussans E, & Ramnani N A probabilistic MR atlas of the human cerebellum. NeuroImage 46, 39–46 (2009). [PubMed: 19457380]

18. Poldrack RA et al. The Cognitive Atlas: toward a knowledge foundation for cognitive neuroscience. Front. Neuroinform 5, 1–11 (2011). [PubMed: 21472085]

19. Poldrack RA Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. Neuron 72, 692–697 (2011). PMCID: PMC3240863. [PubMed: 22153367]

20. Schröter M, Paulsen O & Bullmore ET Micro-connectomics: probing the organization of neuronal networks at the cellular scale. Nat. Rev. Neurosci 18, 131–146 (2017). [PubMed: 28148956]

21. Bullmore E & Sporns O Complex brain networks: graph theoretical analysis of structural and functional systems. Nat. Rev. Neurosci 10, 186–198 (2009). [PubMed: 19190637]

22. Kragel PA et al. Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. Nat. Neurosci 21, 283–289 (2018). [PubMed: 29292378]

23. Wang X et al. Representational similarity analysis reveals task-dependent semantic influence of the visual word form area. Sci. Rep 8, 1–10 (2018). [PubMed: 29311619]

24. von Luxburg U, Williamson RC & Guyon I Clustering: science or art? JMLR: Workshop Conf. Proc. 27, 65–79 (2012).

25. Pennington J, Socher R, & Manning C GloVe: global vectors for word representation. Proc. 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1532–1543 (2014).

26. McCoy TH et al. High throughput phenotyping for dimensional psychopathology in electronic health records. Biol. Psychiatry 83, 997–1004 (2018). [PubMed: 29496195]

27. Kessler RC et al. Prevalence, severity, and comorbidity of 12-month DSM-IV disorders in the National Comorbidity Survey Replication. Arch. Gen. Psychiatry 62, 617–627 (2005). [PubMed: 15939839]

28. Contractor AA et al. Latent profile analyses of posttraumatic stress disorder, depression and generalized anxiety disorder symptoms in trauma-exposed soldiers. J. Psychiatr. Res 68, 19–26 (2015). [PubMed: 26228395]

29. Williams LM Precision psychiatry: a neural circuit taxonomy for depression and anxiety. Lancet Psychiatry 3, 472–480 (2016). [PubMed: 27150382]

30. Russell JA A circumplex model of affect. J. Pers. Soc. Psychol 39, 1161–1178 (1980).

31. Barrett LF The theory of constructed emotion: an active inference account of interoception and categorization. Soc. Cogn. Affect. Neurosci 12, 1–23 (2017). [PubMed: 27798257]

32. Kornblum S, Hasbroucq T & Osman A Dimensional overlap: cognitive basis for stimulus-response compatibility – a model and taxonomy. Psychol. Rev 97, 253–270 (1990). [PubMed: 2186425]

33. Corbetta M Frontoparietal cortical networks for directing attention and the eye to visual locations: identical, independent, or overlapping neural systems? Proc. Natl. Acad. Sci. U. S. A 95, 831–838 (1998). [PubMed: 9448248]

34. McCoy TH et al. Genome-wide association study of dimensional psychopathology using electronic health records. Biol. Psychiatry 83, 1005–1011 (2018). [PubMed: 29496196]

35. Drysdale AT et al. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat. Med 23, 28–38 (2017). [PubMed: 27918562]

36. Janak PH & Tye KM From circuits to behaviour in the amygdala. Nature 517, 284–292 (2015). [PubMed: 25592533]

37. Cottaris NP & De Valois RL Temporal dynamics of chromatic tuning in macaque primary visual cortex. Nature 395, 896–900 (1998). [PubMed: 9804422]

38. Salmelin R, Hari R, Lounasmaa OV & Sams M Dynamics of brain activation during picture naming. Nature 368, 463–465 (1994). [PubMed: 8133893]

39. Gutschalk A, Patterson RD, Scherg M, Uppenkamp S & Rupp A Temporal dynamics of pitch in human auditory cortex. NeuroImage 22, 755–766 (2004). [PubMed: 15193604]

40. Menon V & Uddin LQ Saliency, switching, attention and control: a network model of insula function. Brain Struct. Func 214, 655–667 (2010).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

41. van den Heuvel MP & Sporns O Network hubs in the human brain. Trends Cogn. Sci 17, 683–696 (2013). [PubMed: 24231140]

42. McTeague LM et al. Identification of common neural circuit disruptions in cognitive control across psychiatric disorders. Am. J. Psychiatry 174, 676–685 (2017). [PubMed: 28320224]

43. Eisenberg IW et al. Uncovering the structure of self-regulation through data-driven ontology discovery. Nat. Commun 10, 1–13 (2019). [PubMed: 30602773]

44. Bolt T et al. Ontological dimensions of cognitive-neural mappings. Neuroinformatics 18, 451–463 (2020). [PubMed: 32067196]

45. Bertolero MA, Yeo BTT, Bassett DS & D'Esposito M A mechanistic model of connector hubs, modularity and cognition. Nat. Hum. Behav 2, 765–777 (2018). [PubMed: 30631825]

46. Ioannidis JPA, Fanelli D, Dunne DD & Goodman SN Meta-research: evaluation and improvement of research methods and practices. PLoS Biol. 13, e1002264 (2015). [PubMed: 26431313]

47. Bolukbasi T, Chang K-W, Zou J, Saligrama V & Kalai A Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Adv. Neural Inf. Proc. Syst 2016, 4349–4357 (2016).

48. Voytek JB & Voytek B Automated cognome construction and semi-automated hypothesis generation. J. Neurosci. Methods 208, 92–100 (2012). [PubMed: 22584238]

References (Methods)

49. Yarkoni T Automated Coordinate Extractor (ACE). (GitHub, 2015).

50. Lancaster JL et al. Bias between MNI and Talairach coordinates analyzed using the ICBM-152 brain template. Hum. Brain Mapp 28, 1194–1205 (2007). [PubMed: 17266101]

51. Benjamini Y & Hochberg Y Controlling the false discovery rate: a practical and powerful approach to multiple testing. J. R. Stat. Soc. Series B Stat. Methodol 57, 289–300 (1995).

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

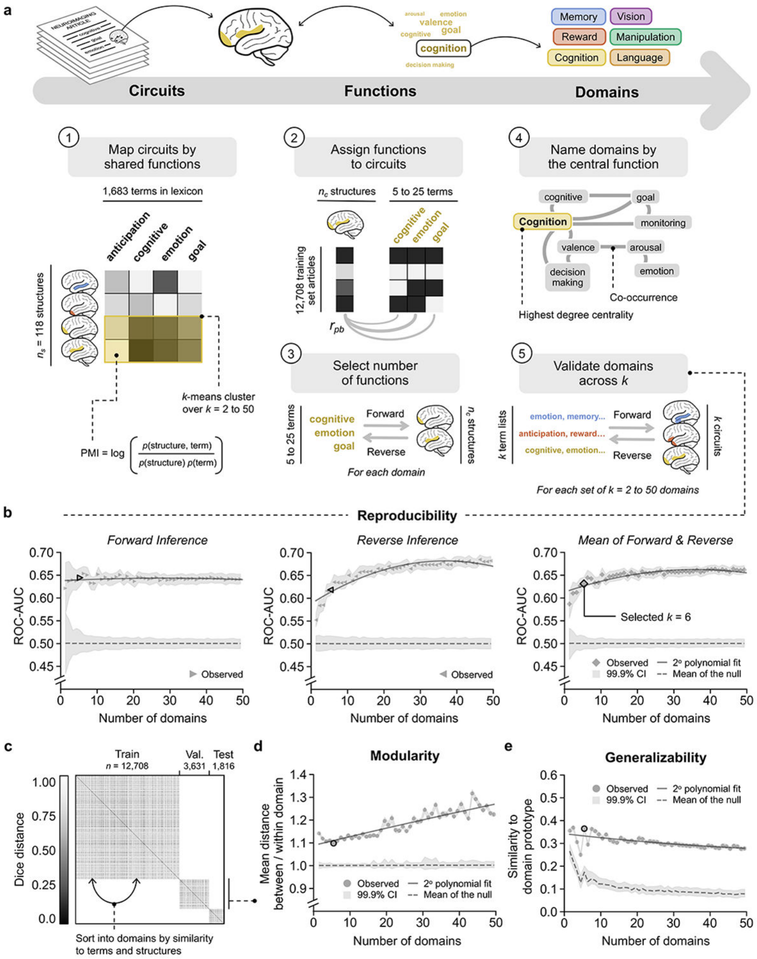


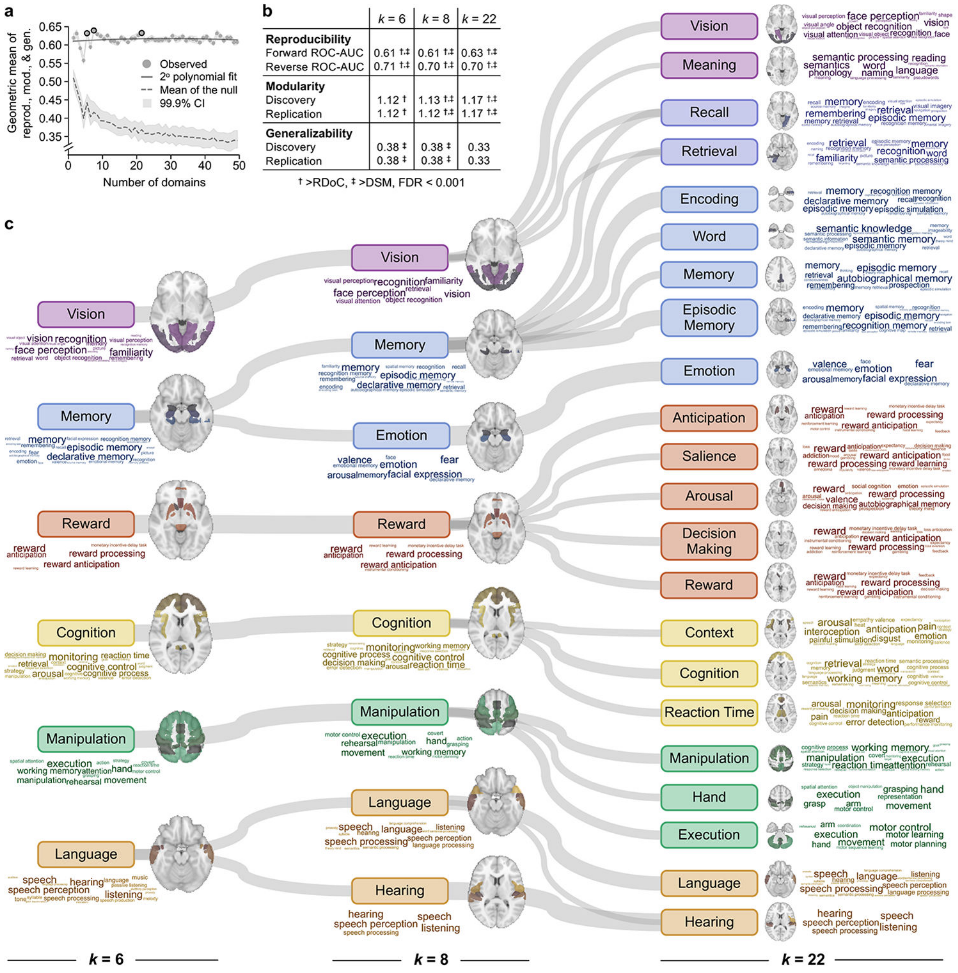
Fig. 1 |. Approach to data-driven ontology.

a, A framework for domains of brain function was generated in the training set. First, 118 brain structures were clustered by *k*-means according to co-occurrences with 1,683 terms for mental functions. The co-occurrence matrix was weighted by pointwise mutual information (PMI). Second, the top 25 mental function terms were assigned to each circuit based on the point-biserial correlation (r_{pb}) of binarized occurrences with the centroid of structure occurrences. Third, the number of terms was selected to maximize average ROC-AUC of logistic regression classifiers predicting structure occurrences from term occurrences (forward inference) and term occurrences from structure occurrences (reverse inference) over lists of 5 to 25 terms. Fourth, each domain was named by the mental function term with highest degree centrality of co-occurrences with other terms in the domain. Fifth, domains were validated by ontological principles such as reproducibility, the average ROC-AUC for forward and reverse inference classifiers. Forward classifiers predicted whether coordinates occurred within circuits from term occurrences summed across lists, and reverse classifiers predicted mean-thresholded occurrences of terms in the lists from coordinates within circuits. **b,** Validation set ROC-AUC across *k* for forward inference, reverse inference, and their average. **c,** To assess modularity and generalizability, articles were assigned to domains based on Dice similarity of mental function terms and brain coordinate data. **d,**

Modularity was computed as the mean ratio of Dice distances for articles between domain partitions versus within a domain partition. **e,** Generalizability was computed as the Dice similarity of each domain to its assigned articles. Markers are outlined in black for $k = 6$, which was selected for follow-up analyses. Performance at every k exceeded chance by reproducibility, modularity, and generalizability. Shaded areas around markers represent 99.9% confidence intervals computed by resampling validation set articles with replacement over 1,000 iterations. Dashed lines represent the mean of null distributions generated by shuffling true labels for validation set articles over 1,000 iterations, and the surrounding shaded areas are 99.9% confidence intervals.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Author Manuscript Author Manuscript Author Manuscript Author Manuscript



	k = 6	k = 8	k = 22
Reproducibility			
Forward ROC-AUC	0.61 †,‡	0.61 †,‡	0.63 †,‡
Reverse ROC-AUC	0.71 †,‡	0.70 †,‡	0.70 †,‡
Modularity			
Discovery	1.12 †	1.13 †,‡	1.17 †,‡
Replication	1.12 †	1.12 †,‡	1.17 †,‡
Generalizability			
Discovery	0.38 ‡	0.38 ‡	0.33
Replication	0.38 ‡	0.38 ‡	0.33

† >RDoC, ‡ >DSM, FDR < 0.001

Fig. 2 |. Top-performing solutions of the data-driven framework.

a, Solutions at the *k* = 6, 8, and 22 levels had the three highest values of a multi-objective function summarizing the reproducibility, modularity, and generalizability metrics. At a given *k*, each metric was computed by macro-averaging results across domains. In the case of reproducibility, ROC-AUC results were then averaged across forward and reverse inference classifiers. Finally, the geometric mean was taken across metrics to avoid over- and underweighting due to differences in scale. Shaded areas around markers represent 99.9% confidence intervals computed by resampling validation set articles with replacement over 1,000 iterations. Dashed lines represent the mean of null distributions generated by shuffling true labels for validation set articles over 1,000 iterations, and the surrounding shaded areas are 99.9% confidence intervals. **b,** For the selected solutions, values of each metric after macro-averaging across domains are shown. † Greater than RDoC, ‡ greater than the DSM, based on two-sided differences in bootstrap means tested for each framework pair (FDR < 0.001). **c,** Domains were linked across levels by the Dice similarity of their mental function terms and brain structures. Links were thresholded by Dice similarity > 0 and FDR < 0.001. The hierarchy can be viewed online at http://neuro-knowledge.org with the ability to magnify brain circuit maps and word clouds.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

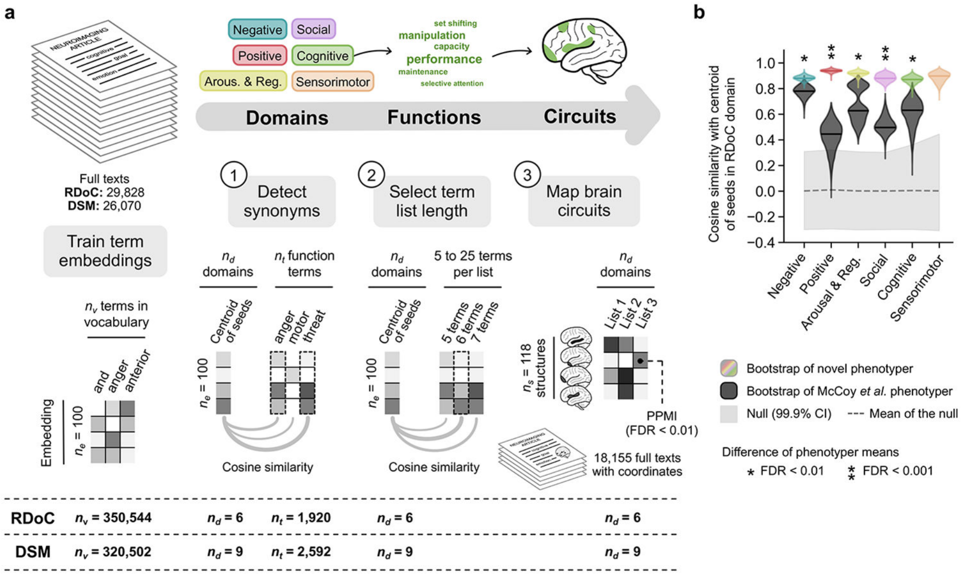


Fig. 3 |. Approach to mapping expert-determined frameworks for brain function (RDoC) and mental illness (DSM).

a, Seed terms from RDoC and the DSM were translated into the language of the human neuroimaging literature through a computational linguistics approach. Term embeddings of length 100 were trained using GloVe.[25] For RDoC, embeddings were trained on a general human neuroimaging corpus of 29,828 articles (Extended Data Fig. 1b). For the DSM, embeddings were trained on a psychiatric human neuroimaging corpus of 26,070 articles (Extended Data Fig. 1c). Candidate synonyms included mental functions in the case of RDoC and both mental functions and psychopathology in the case of the DSM (Supplementary Table 2). In the first step, synonyms were identified for each domain by the cosine similarity of their embeddings with the centroid of the domain's seed term embeddings. Second, the number of terms for each domain was selected to maximize cosine similarity. Third, the mental function term lists for each domain were mapped onto brain circuits based on positive pointwise mutual information (PPMI) of term and structure co-occurrences across the 18,155 articles with activation coordinate data (Extended Data Fig. 1a). Structures were included in a circuit if the FDR of the observed PPMI was <0.01, determined by comparison to a null distribution generated by shuffling term list features over 10,000 iterations. **b,** Semantic similarity to seed terms in the RDoC framework for our term lists generated using GloVe (colored) compared to a baseline from the literature (dark gray). The baseline model includes term lists generated by McCoy *et al.* through latent semantic analysis followed by filtering.[26] Bootstrap distributions for each domain were generated by resampling the 100-n embedding dimension with replacement over 10,000 iterations, then assessed for a one-sided difference in means (* FDR < 0.01, ** FDR < 0.001). Solid lines denote observed similarity values. No comparison is shown for sensorimotor systems, which was added to RDoC following the McCoy *et al.* publication. Null distributions were generated for the GloVe term lists by shuffling embeddings over 10,000 iterations. Gray dashed lines denote null distribution means.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

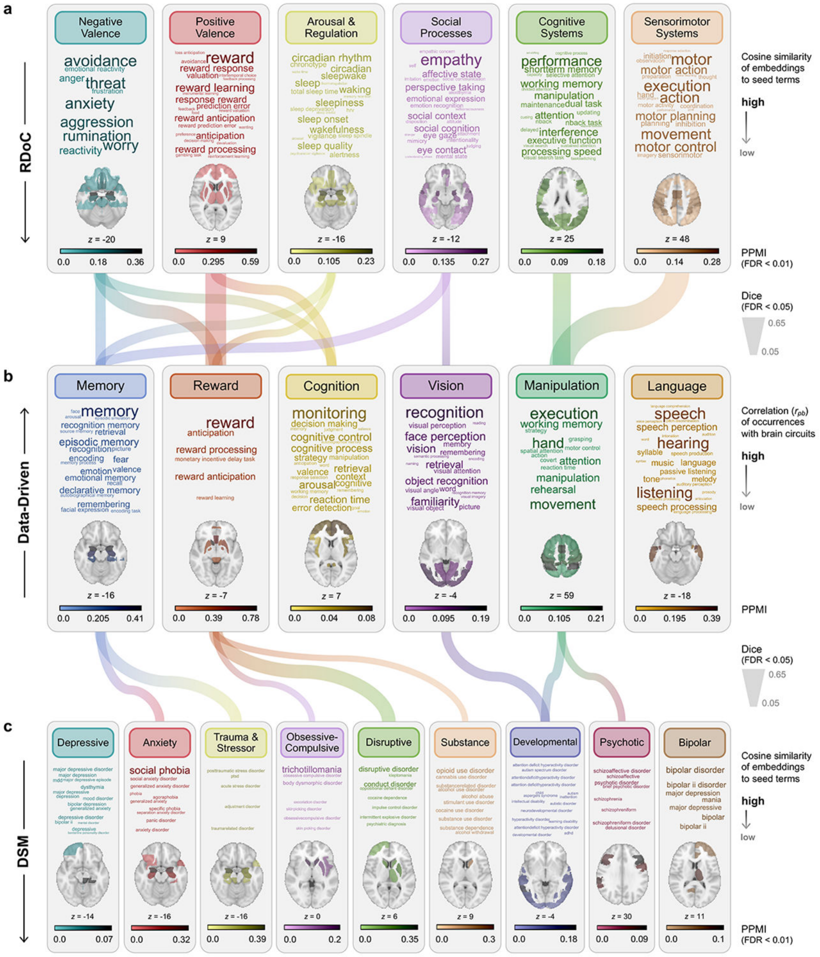


Fig. 4 |. Data-driven framework of brain functions related to expert-determined frameworks for brain function (RDoC) and mental illness (DSM).

Links are scaled to the Dice similarity of mental function terms and brain structures in each domain (FDR < 0.05 based on permutation testing over 10,000 iterations). Term size is scaled to frequency in the corpus of 18,155 articles with activation coordinate data. **a,** The RDoC framework was modeled in a top-down manner from terms to brain circuits (Fig. 3a). **b,** A data-driven framework for domains of brain function was engineered in a bottom-up manner beginning with circuits (Fig. 1a). **c,** The DSM framework for mental disorders was modeled in a top-down manner by a procedure analogous to that for RDoC (Fig. 3a).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

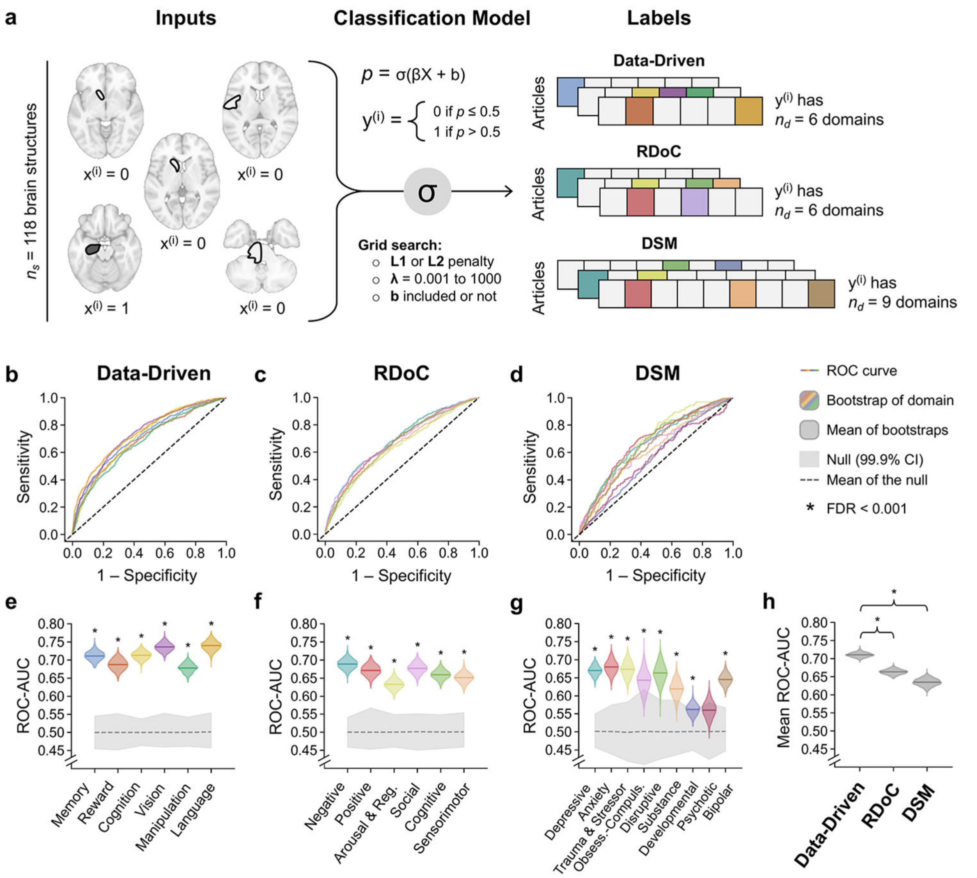


Fig. 5 |. Mental functions defined in a data-driven manner have more reproducible links with locations of brain activity.

a, Logistic regression classifiers were trained to predict the occurrence of terms for mental functions in neuroimaging article full texts from the brain activation coordinates that articles reported. Inputs (X) included activation coordinate data mapped to n_s = 118 brain structures in a whole-brain neuroanatomical atlas, where $x^{(i)}$ is a vector of length n_s for each article i. Labels ($y^{(i)}$) were term occurrences thresholded by mean frequency across the corpus, then the mean frequency of terms in each domain, of length n_d depending on the number of domains in the framework. Weights for the inputs (β) were fit by training classifiers using a sigmoid function (σ) over 1,000 iterations with hyperparameters selected through a grid search over the following values: penalty of L1 or L2; regularization strength (λ) of 0.001, 0.01, 0.1, 1, 10, 100 or 1,000; intercept (b) included or not. Training was performed in 70% of articles (n = 12,708), hyperparameters were tuned in 20% of articles (n = 3,631), and classifiers were evaluated in a test set containing 10% of articles (n = 1,816). ROC curves are shown for the test set performance of classifiers with mental function features defined by **b,** the data-driven framework, **c,** RDoC, and **d,** the DSM. For each domain in **e,** the data-driven framework, **f,** RDoC, and **g,** the DSM, the significance of the test set ROC-AUC was determined by a one-sided permutation test comparing the observed value to a null distribution, and the p value was FDR-corrected for multiple comparisons (* FDR < 0.001). Observed values in the test set are shown with solid lines. Null distributions (gray) were computed by shuffling true labels for term list scores over 1,000 iterations; the

Author Manuscript
Author Manuscript
Author Manuscript
Author Manuscript

99.9% confidence interval is shaded, and distribution means are shown with dashed lines. Bootstrap distributions of ROC-AUC (colored) were computed by resampling articles in the test set with replacement over 1,000 iterations. **h,** The difference in mean ROC-AUC was assessed for each framework pair by a two-sided bootstrap test. The data-driven framework had higher ROC-AUC than both RDoC (99.9% CI of the difference = [0.019, 0.073]) and the DSM (99.9% CI of the difference = [0.032, 0.111]). Solid lines denote means of the bootstrap distributions obtained by macro-averaging across domain classifiers.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

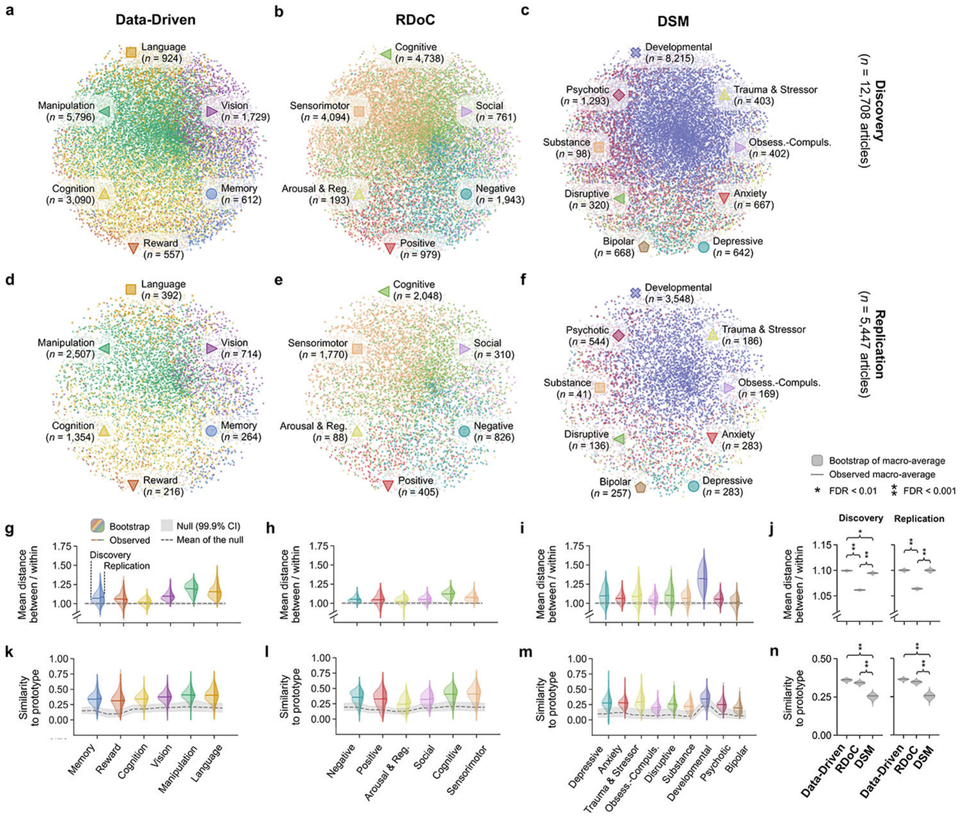


Fig. 6 |. The data-driven framework partitions the neuroimaging literature into modular subfields, for which domains are generalizable representations of brain circuits and mental functions.

Articles were matched to domains based on the Dice similarity of mental function terms and brain structures. Multidimensional scaling was applied to Dice distances between articles in the full corpus and visualized separately for the discovery set (for **a,** the data-driven, **b,** RDoC, and **c,** the DSM) and the replication set (for **d,** the data-driven, **e,** RDoC, and **f,** the DSM). Domain assignment is indicated by marker color and shape. Modularity of the article partitioning was assessed by comparing the mean Dice distance of function and structure occurrences of articles between versus within domains of **g,** the data-driven framework, **h,** RDoC, and **i,** the DSM. Observed values are colored by domain; null distributions in gray were computed by shuffling distance values across article partitions. **j,** Bootstrap distributions of modularity macro-averaged across the framework domains were assessed for differences in means between frameworks (* FDR < 0.001). Relative to RDoC, modularity was higher for both the data-driven framework (99.9% CI of the difference = [0.035, 0.042] discovery, [0.031, 0.044] replication) and the DSM (99.9% CI of the difference = [0.027, 0.039] discovery, [0.029, 0.045] replication). Generalizability was assessed by computing the Dice similarity of each domain's "prototype" vector of function terms and brain structures with the terms and structures occurring in each article of the domain's partition within **k,** the data-driven framework, **l,** RDoC, and **m,** the DSM. Observed values are colored by domain; null distributions in gray were computed by shuffling terms and structures in each prototype. **n,** Bootstrap distributions of generalizability macro-averaged across domains were assessed for two-sided differences in means between frameworks

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(* FDR < 0.001). Relative to the DSM, generalizability was higher for both the data-driven framework (99.9% CI of the difference = [0.055, 0.176] discovery, [0.051, 0.179] replication) and RDoC (99.9% CI of the difference = [0.043, 0.157] discovery, [0.012, 0.157] replication). All bootstrap and null distributions shown were generated over 1,000 iterations.