CellPress
OPEN ACCESS

Trends in
Cognitive Sciences

Feature Review

Decoding semantic representations in mind and brain

Saskia L. Frisby,[1,*] Ajay D. Halai,[1] Christopher R. Cox,[2] Matthew A. Lambon Ralph,[1] and Timothy T. Rogers[3,*]

A key goal for cognitive neuroscience is to understand the neurocognitive systems that support semantic memory. Recent multivariate analyses of neuroimaging data have contributed greatly to this effort, but the rapid development of these novel approaches has made it difficult to track the diversity of findings and to understand how and why they sometimes lead to contradictory conclusions. We address this challenge by reviewing cognitive theories of semantic representation and their neural instantiation. We then consider contemporary approaches to neural decoding and assess which types of representation each can possibly detect. The analysis suggests why the results are heterogeneous and identifies crucial links between cognitive theory, data collection, and analysis that can help to better connect neuroimaging to mechanistic theories of semantic cognition.

Highlights

State-of-the-art brain imaging studies have recently produced a variety of sometimes contradictory conclusions about the neural systems that support human semantic memory.

Multivariate techniques deployed in this work adopt implicit or explicit assumptions that limit the types of signal they can detect, and thus the types of hypotheses they can test.

We lay out the space of possible cognitive and neural representations and then critically review contemporary methods to determine which analyses can test which hypotheses.

The results account for the heterogeneity of recent findings and identify an important empirical and methodological gap that makes it difficult to connect the imaging literature to neurocomputational models of semantic processing.

The neurocognitive quest for semantic representations

Cognitive science has long sought to understand the mechanisms underlying human semantic memory – the storehouse of knowledge that supports our ability to comprehend and produce language, recognize and classify objects, and understand everyday events. Recently, cross-fertilization of cognition, neuroscience, and machine learning has generated a plethora of new analysis methods to aid the discovery of neural systems that encode semantic information [1–5]. Although this renaissance has produced a remarkable array of new findings, the evolution of different approaches across research groups makes it difficult to track them all, understand their respective strengths and limitations, and compare results across studies. Consequently, the literature contains sometimes startlingly different conclusions about the nature, structure, and organization of semantic representations in the mind and brain, and the field has little recourse for understanding why the differences arise or how they might be reconciled.

We address this challenge by reviewing hypotheses about how semantic information may be encoded computationally and neurally, then critically evaluating the types of representational structure that contemporary multivariate methods can possibly discover in functional neuroimaging data. Crucially, each method encapsulates assumptions about how neural systems encode mental structure that then constrain the types of neural coding it can, and cannot, detect. Hypothesis, data collection, and analysis are therefore linked in ways that sometimes go unremarked and may explain the heterogeneity of findings in the literature. Through exposition of these points, we present an overview of the current empirical landscape with the aim of both organizing current thinking about semantic representations in mind and brain, and of providing a more general field guide to contemporary multivariate methods for brain imaging.

What might semantic representations be like computationally?

Semantic representations serve at least two crucial cognitive functions. First, they express conceptual similarity structure – knowledge that items can be similar in kind even if they are distinct

[1]Medical Research Council (MRC) Cognition and Brain Sciences Unit, Chaucer Road, Cambridge CB2 7EF, UK
[2]Department of Psychology, Louisiana State University, Baton Rouge, LA 70803, USA
[3]Department of Psychology, University of Wisconsin–Madison, 1202 West Johnson Street, Madison, WI 53706, USA

*Correspondence:
saskia.frisby@mrc-cbu.cam.ac.uk (S.L. Frisby) and ttrogers@wisc.edu (T.T. Rogers).

in appearance (e.g., hummingbird and ostrich), verbal labels (e.g., dog and wolf), or the action plans that engage them (e.g., glue and tape). Children as young as 9 months of age detect such relationships and use them to guide reaches even when they contravene perceptual similarity [6–8]. Adults can reliably judge relatedness in kind and sort items into conceptual groups on this basis [9–11], and both children and adults use conceptual similarity as a primary basis for generalizing names and other properties [12–14]. Second, semantic representations support knowledge retrieval or inference – attributing to an item or event properties that are not directly observed or stated. For instance, when observing a picture of a parrot in a textbook, the student may infer that the item can fly even though the image is static; reading about a trip to the restaurant, she may infer that the diner had to pay even if this is not mentioned; observing the new neighbor's pet, a toddler may call it 'doggie' even if it is an unfamiliar breed, and so on. Semantic representations thus can be defined as the cognitive and neural states that express conceptual structure and support semantic retrieval/inference. Hypotheses about the cognitive mechanisms that support these functions reside within a fairly constrained space of possibilities (Figure 1).

Considering conceptual structure, most approaches adopt one of three positions. The first proposes that semantic memory contains many discrete and independent **category** (see Glossary) representations, each corresponding roughly to a basic-level natural language concept such as *tree* or *boat* [15,16] (Figure 1, top) and possibly to more general (*plant*, *vehicle*) or specific (*elm*, *yacht*) classes [17,18]. On this view, verbal comprehension involves discerning the category to which a word refers [19] whereas comprehension of visual and other sensory inputs involves correctly classifying a perceived item [4,18,20,21]. Category-based theories explain conceptual structure by proposing that conceptually similar items activate the same category representation – for instance, parrots, hummingbirds, and robins are viewed as being conceptually related because they all activate the mental category *bird*.

The second view proposes that semantic representations are composed of local **features**, each independently indicating the presence/absence of a property such as *is red*, *can fly*, or *has eyes* (Figure 1, middle row). Each perceived item or word activates associated features, indicating properties that are likely to be true of the item [22–26]. Conceptual similarity structure arises from property overlap: hummingbirds and ostriches are understood to be similar in kind because they possess many common properties (wings, feathers, etc.), but are also known to be non-identical because they possess individuating properties as well [27,28].

Category-based approaches are often distinguished from feature-based views because of the special role that category representations play in determining conceptual similarity and supporting inference. For instance, prototype theories [29], 'entry-level' [18,30] and spreading-activation views [31], rational approaches [15], and some neurally inspired models of object categorization [32] all propose that access to semantic information depends upon first matching a stimulus (image, word, sound, etc.) to a semantic category. Successful categorization then provides direct access to semantic information or initiates a 'search' of the semantic system, allowing retrieval of other properties. On such views, semantic categories constitute more than merely an additional feature that is attributed to a perceived item.

Nevertheless, under both approaches semantic representations can also be viewed as vectors in a high-dimensional representation space. For categorical theories, dimensions encode membership of distinct and mutually exclusive categories, and the representation of an item is a multinomial probability distribution indicating the probability that a stimulus belongs to each class. For instance, observing an item with wings, feathers, and a beak would generate a high probability density on the *bird* axis and a low density on axes corresponding to *fish*, *car*, *boat*, etc. because

Glossary

Category: (of a representation) composed of discrete, independent units that each correspond to a concept (such as *boat*, *vehicle*, or *yacht*).
Conjoint: (of a representation) consisting of units that express different semantic information depending on the states of other units.
Consistent: (of a representation) associated with the same direction of change in activation across individuals – for example, homologous voxels in different individuals become more active when representing *cat*.
Contiguous: (of a representation) composed of units residing in the same brain region.
Decoding: predicting the stimulus (or sometimes the properties of the stimulus, or of the task) experienced by a participant using patterns of activity across multiple neural units.
Dispersed: (of a representation) composed of units residing in different brain regions.
Electrocorticography (ECoG): a method of measuring brain activity via intracranial electrodes placed on the cortical surface.
Electroencephalography (EEG): a method of measuring brain activity via electrodes placed on the scalp.
Encoding model: a model that predicts the activity of a single neural unit using multiple independently interpretable features of the stimulus. Multiple encoding models are used to predict activity across multiple neural units.
Feature-based: (of a representation) composed of multiple independently interpretable features (such as *is red* or *can fly*).
Functional magnetic resonance imaging (fMRI): a method of measuring brain activity by detecting changes in blood flow.
Grounded: (of a representation) requiring the generation of modality-specific surface representations to produce retrieval/inference.
Heterogeneous: (of a representation) consisting of units that adopt different activation states when representing a concept.
Homogeneous: (of a representation) consisting of units that all adopt the same activation state when representing a concept.
Inconsistent: (of a representation) associated with different directions of

the probability that the item is a bird is high and the probability of it belonging to other categories is low. For feature-based theories, dimensions encode various directly interpretable properties, and the representation of an item indicates, independently on each dimension, the binomial probability that the item possesses the corresponding property. On this view, *cardinal* is a vector with high values on dimensions such as *is red* and *can fly*, but low values on dimensions such as *has scales* and *can swim*. Moreover, some such features may directly indicate the semantic category label of an item (e.g., 'bird', 'fish'), although, in contrast to category-based theories, such labels have no special function beyond that of other features. In both cases, conceptual structure reflects the similarity of different points in the **vector space**.

The third proposal likewise views semantic representations as points in a high-dimensional vector space, but without assigning any directly interpretable meaning to the corresponding dimensions (Figure 1, bottom). Perception of a stimulus or word evokes an activation pattern across an ensemble of representation units, corresponding to a point in the space where the proximity between points expresses conceptual similarity [33–35]. Unlike feature- and category-based approaches, however, one cannot discern what information is encoded in the representation by looking at the activation of each element taken independently. Instead, what matters is the similarity of a given vector to those elicited by other items, taken across all units in the ensemble. On this view, *cardinal* is a vector with high values on some dimensions and low values on others. Examining each dimension reveals no information about the properties of the cardinal, but information can be gleaned from the fact that *cardinal* is located very close to *goldfinch*, reasonably close to *ostrich*, and far from *canoe* (Box 1).

Considering retrieval/inference, most approaches adopt one of two proposals, both compatible with the perspectives on conceptual structure outlined above. First, semantic information may be **self-contained** within the representation such that activation brings retrieval/inference along with it (Figure 1; left column). For categorical models, the category representation might encapsulate knowledge of properties essential to or characteristic of category members, as in classical, prototype, and rational models [36–38]. In feature-based models, because each element of the representation vector corresponds to an explicit property, the system need only 'read off' the vector elements active above some threshold to attribute the corresponding properties to the perceived/named item. Such a view is captured by semantic feature-based neural network models [22–24], spreading-activation models [31,39,40], and distributional semantic models that constrain representations to have interpretable dimensions (such as topic models and non-negative sparse embeddings; Box 1) [41,42]. For vector space models, although the dimensions of the representation space are not independently interpretable, retrieval/inference can still be self-contained by proposing that these functions rely on similarity and/or direction within the representation space [34]. For instance, the system may infer that the cardinal can fly and breathe because the vectors for the words 'fly' and 'breathe' are both near to the vector for 'cardinal' and are situated along a direction in the space that separates behavioral 'can' properties from other property types (such as parts, names, colors, etc.). Such a perspective is captured by distributional semantic models that are not constrained to yield interpretable dimensions (e.g., latent semantic analysis [33], holistic analog to language [43], word2vec [34], and language neural networks [44]) (Box 1).

Self-contained approaches face a significant hurdle, however: retrieving the content of a representation requires a labeling scheme, without which it would be impossible to know which semantic content 'goes with' which representation vectors (sometimes called the symbol grounding problem [45]). The second approach to retrieval/inference (Figure 1, right column) addresses this problem by proposing that semantic content is **grounded** in perception, action, and language systems

change in activation across individuals – for example, homologous voxels in multiple individuals behave differently when representing *cat*, some becoming more active and others becoming less active.
Independent: (of a representation) consisting of units that express the presence or absence of the same semantic information irrespective of the states of other units.
Labeled data: a dataset specifying both input and output values for fitting an encoding or decoding model.
Magnetoencephalography (MEG): a method of measuring brain activity by measuring magnetic fields generated by neural activity.
Multivariate pattern classification (MVPC): the categorization of stimuli based on the neural patterns they evoke (a form of decoding).
Region of interest (ROI): a subset of neural units, chosen in a hypothesis-guided way, upon which an analysis is conducted.
Regularization: a method of avoiding overfitting by finding classifier weights that jointly minimize classification error and an additional loss which is a function of the classifier weights.
Representational similarity analysis (RSA): a method of investigating representational structure by comparing the similarity structure recorded to that hypothesized.
Self-contained: (of a representation) encapsulating semantic information within itself such that mere activation of the representation brings about retrieval/inference.
Surface representation: a sensory representation of a stimulus that is modality-specific – for example, color (specific to the visual modality) or a paddling action (specific to the motor modality).
Transcranial magnetic stimulation (TMS): the use of magnetic fields to temporarily and reversibly disrupt brain function.
Vector space: (of a representation) composed of a pattern across representational units, the meanings of which cannot be independently interpreted.

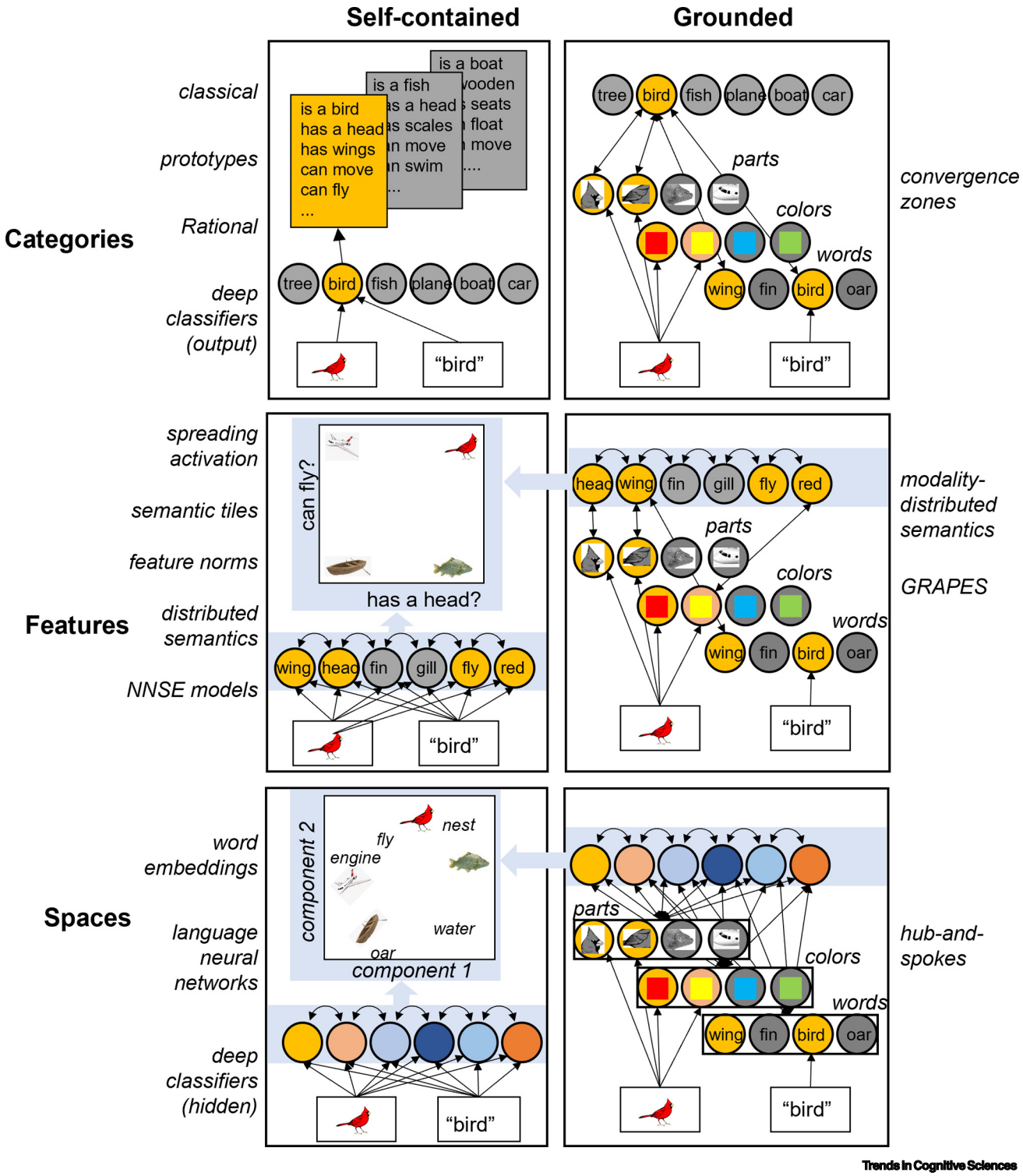


Figure 1. Computational hypotheses about semantic representation. There are three ways in which conceptual structure could be encoded. First, information may be encoded in discrete, independent category representations (top row). On this view, sensory inputs recruit discrete and independent category representations

(Figure legend continued at the bottom of the next page.)

Box 1. Ways of estimating semantic structure

Category-based theories propose that distinct representations encode information about different semantic categories. Some have argued that different brain regions are specialized to represent categories that are important for survival over evolution, such as faces, tools, animals, foods, body parts, and shelter [73,90,107,108], but the general question of which categories are stored in memory and why remains controversial [109,110].

Feature-based theories cast semantic representations as vectors that denote the properties of a given item, such as *is red*, *can fly*, or *has blood inside* for the concept *cardinal*. Three methods have been used to construct such vectors.

(i) Semantic norming studies ask participants to list the properties that are true of a given concept. Properties generated and/or verified by many participants are compiled in a matrix with rows corresponding to the tested concepts and columns corresponding to the various properties generated by the participants across all study concepts [28,111] (J. Tanaka and L. Szechter, unpublished data).

(ii) Brain-inspired feature vectors identify semantic properties that, from univariate brain imaging, selectively engage different cortical areas. Participants then rate the strength of association between a given concept and each such property. The procedure produces many fewer features than norming studies, but still captures rich conceptual structure [26,52].

(iii) Non-negative sparse word embeddings (NNSE) estimate feature vectors from text corpora by exploiting the tendency for words with similar meanings to occur in similar contexts. Standard techniques {e.g., latent semantic analysis (LSA) [33,113,114] and word2vec [34]} generate embeddings with uninterpretable dimensions, but, when embeddings are constrained to be both sparse (zeros on most dimensions) and non-negative (only positive values on the rest), the resulting elements are more interpretable and each word can be viewed as a semantic feature vector [115].

Vector spaces cast semantic representations as points in a high-dimensional space where pairwise distances capture conceptual relatedness, but with uninterpretable dimensions. Two methods are used to compute such spaces.

(i) Unconstrained word embeddings adopt the same corpus-based approach as non-negative sparse embeddings without sparsity or positivity constraints. The resulting spaces express comparable structure to NNSE using fewer dimensions, but the dimensions are not typically independently interpretable.

(ii) Deep neural networks trained on natural language and/or large image datasets learn vector space representations for photographs, words, or larger units of language. Deep image classifiers represent color photographs with activation vectors across many serial processing layers [116,117]; sentence-processing networks represent words, phrases, or whole passages of text as activation vectors over internal units {e.g., bidirectional encoder representations from transformers (BERT) [44] and generative pretrained transformer 3 (GPT3) [118]}.

that directly encode **surface representations** of the environment: shapes, colors, parts, movements, affordances, words, and so on [46–48]. On this view, the activation of a categorical, feature-based, or vector space representation does not in itself cause information retrieval/inference. Instead, retrieval/inference arises when these structure-encoding representations activate modality-specific representations that are identical or intimately related to those that directly mediate perception and action. Thus the categorical/featural/vector space representation of *canoe* is meaningful only in virtue of its ability to generate mental images of what a canoe looks like (including shape, color, parts, etc.), motor actions associated with canoes (e.g., paddling), words used to describe canoes ('boat', 'light', 'floats'), and so on.

which either encapsulate semantic information within themselves [15,20,36,105,106] (top left) or connect and bind modality-specific surface representations encoding characteristics of category members [49,50] (top right). Second, semantic information may be distributed across independent and interpretable semantic feature representations, with featural overlap indicating conceptual similarity (middle). Features may independently and intrinsically encode the presence of stipulated semantic features within a concept [22–24,75] (middle left) or gain meaning via connection to surface representations that directly encode such information [2,25,51,52] (middle right). Third, semantic information may be encoded by a continuous distributed representation space that expresses conceptual similarities among items even though its dimensions are not independently interpretable (bottom). Semantic information may be self-contained by the distances encoded in such a space [33,34,41,44] (bottom left) or grounded via mappings from the space to modality-specific surface representations of specific properties [9,53,54] (bottom right). Black arrows illustrate how information may flow through the network given the stimuli shown. Text on either side indicates well-known perspectives in the literature that characterize each view. For feature-based and vector space representations, representational spaces are schematized on a blue background. Blue arrows point to the type of representational similarity structure encoded by the corresponding layers – note that both self-contained and grounded approaches can encode the same representational space. Abbreviations: GRAPES, grounding representations in action, perception, and emotion systems; NNSE, non-negative sparse embeddings.

On a grounded category-based approach, a discrete category representation connects the surface representations encoding characteristics of category members, and binds these together so that they are understood as all inhering in the same concept. For example, *bird* connects surface representations of the visual appearance of feathers, the motion of flight, the word 'bird', and so on; the 'convergence zone' hypothesis provides an example of this view [49,50]. Under grounded feature-based approaches, the featural dimensions that encode the semantic representation are 'labeled' by virtue of their direct/preferential connectivity to surface representations that directly encode the corresponding content – for instance, a semantic dimension encoding the color of an object may be directly connected to color-perception areas; a dimension encoding its associated action may be connected to motion-perception areas; and so on. Several proposals motivated by functional imaging data align with this view, including the GRAPES (grounding representations in action, perception, and emotion systems) framework [51] and the neurally inspired 'experiential features' view [26,52]. Finally, grounded vector-space models suggest that the representational ensemble that encodes conceptual similarity structure connects reciprocally to a variety of different surface representations such that the generation of an activity pattern across the ensemble activates surface representations that encode the specific, embodied properties associated with the corresponding item – a view consistent with the hub-and-spokes model of semantic representation. [9,53–55]

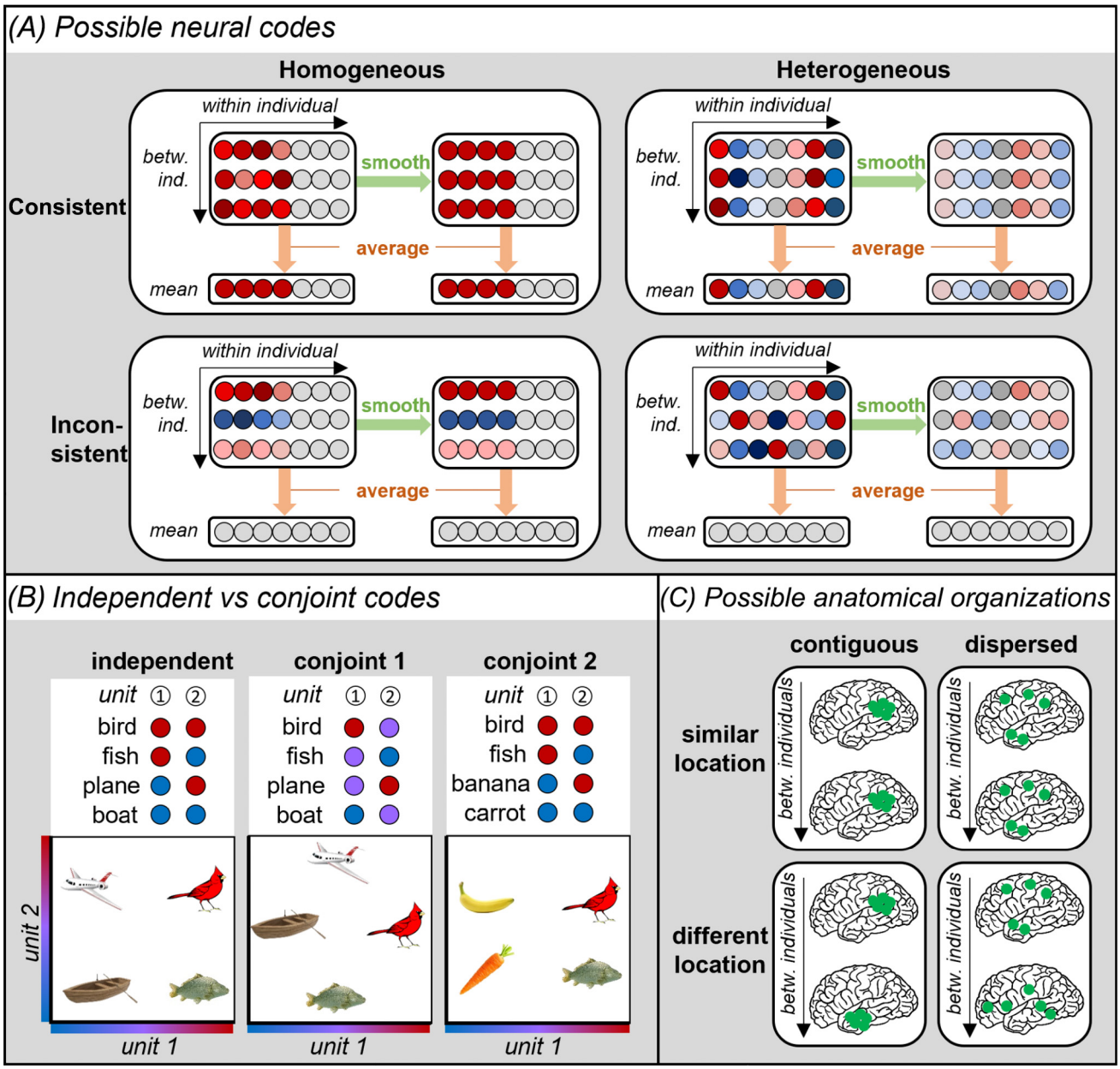
In sum, considering how semantic representations might serve their defining functions – expressing conceptual structure and supporting semantic retrieval/inference – delineates a well-constrained space of hypotheses in which cognitive theories of semantic representation can be situated. The different views, and examples of theories aligning with each, are shown in Figure 1. Each cognitive hypothesis has implications for how neural data are best collected and analyzed; for instance, adjudicating grounded versus self-contained theories may require participants to semantically process stimuli in different modalities. The next section considers how these views constrain the search for neural systems that encode semantic information.

How might semantic representations be organized in the brain?

Next, we consider how these different computational schemes might be implemented in neural systems in ways that can be measured by functional brain imaging. All such technologies can be viewed as summarizing the responses of many different neural populations to a cognitive event. Different technologies such as **functional magnetic resonance imaging (fMRI)**, **electroencephalography (EEG)**, **magnetoencephalography (MEG)**, and **electrocorticography (ECoG)** yield summary estimates at different spatial and temporal granularities (e.g., voxels, EEG sources, and electrodes). We will use the term 'unit' to refer to the summary estimate provided by a given technology over its characteristic window of space and time. Therefore, regardless of imaging modality, the neural response to a stimulus is characterized as a pattern of activation across many units over a particular window of time. Discovering the neural underpinnings of semantic representations then requires close consideration of (i) how the representational elements proposed by a cognitive theory are encoded in unit activation patterns within and across individuals, (ii) how the representational work might be divided among units participating in a representation, and (iii) how signal-carrying units might be anatomically organized within and across individuals.

Variation of the neural code

Within an individual, the neuro-semantic code – how changes in unit activity express semantic information – can be either **homogeneous** or **heterogeneous** (Figure 2A). In a homogeneous code, signal-carrying units all adopt the same activation when the represented information is present – for instance, all voxels representing *cat* become more active when a cat is semantically



Trends in Cognitive Sciences

Figure 2. Hypotheses about the neuro-semantic code. (A) Within individuals a representation may adopt a homogeneous code (all involved units adopt the same activation change – i.e., all become more active or all become less active) or a heterogeneous code (the units involved adopt different changes to activation – i.e., some become more active than others, and/or some become more active and some less active). Across individuals the code may be consistent (the same magnitude and direction of change in all individuals) or inconsistent (different magnitudes and/or directions of change in different individuals). Spatial smoothing and cross-subject averaging can either help or hinder discovery depending on the code. (B) In the independent code shown, unit 1 activation indicates whether the item is animate, while unit 2 independently encodes whether it can fly. In the first conjoint code, the two units express the same similarity relations among the four items, but considered independently, neither unit clearly expresses either dimension. For instance, *fish* and *plane* both moderately activate unit 1, whereas *bird* and *boat* moderately activate unit 2. In the second conjoint example, unit 2 activation is difficult to interpret considered independently, but discriminates birds from fish when unit 1 is active, and fruits from vegetables when unit 1 is inactive. In both conjoint examples, understanding the neural code requires joint consideration of both units. (C) Anatomically, the units in a representation may be localized to a contiguous region or dispersed across multiple distal areas, and the units may occupy either the same or different locations across individuals. The two brains within each white box denote two different individuals. Abbreviation: Betw. individuals, between individuals.

processed. In a heterogeneous code, different units express the same information differently – some voxels representing *cat* may be greatly activated when a cat is present, some greatly suppressed, and some only moderately active, etc. Approaches that average unit activations within participants [e.g., via spatial smoothing or **region of interest (ROI)** averaging] favor the discovery of homogeneous over heterogeneous codes.

Across individuals, the neural code may be **consistent** – a given piece of information is always expressed with the same activity change in homologous units (e.g., *cat* always being signaled by the same activation pattern across aligned voxels of different individuals) – or **inconsistent** (*cat* being signaled by different activation patterns across aligned voxels of different individuals; Figure 2A). Methods that aggregate or summarize unit activation across individuals – for instance, fitting a single model to decode all participants, computing the mean blood oxygen level-dependent (BOLD) response at each voxel before applying a **decoding** model, or averaging predictions of **encoding models** across participants before passing the result to further analysis – favor the discovery of consistent over inconsistent codes. Likewise, methods that align voxels across individuals on the basis of their having similar activation patterns across stimuli (e.g., hyper-alignment) [56] implicitly assume a consistent code.

Independent and conjoint codes

Categorical and feature-based approaches both suggest that each unit **independently** encodes a piece of semantic information: its activity expresses the presence or absence of that information (such as category membership or a semantic feature) regardless of the states of other units. For the example shown in the left panel of Figure 2B, unit 1 encodes whether the stimulus is living or non-living independently of unit 2, whereas unit 2 encodes whether the stimulus can fly independently of unit 1. For any stimulus, it is possible to determine whether the item is alive solely by inspecting the state of unit 1, without needing to consider the activation of other units.

By contrast, vector space hypotheses suggest that units **conjointly** encode a representational space, and that semantic information is expressed in the activity pattern considered across multiple units such that single-unit activation may not be interpretable without consideration of other units in the ensemble. Figure 2B shows two examples. In the middle panel, one cannot determine whether a stimulus is living or whether it can fly solely by inspecting the activation of unit 1 (because *fish* and *plane* elicit equal activation) or unit 2 (because *boat* and *cardinal* elicit equal activation). Considering the joint activation of both units clearly separates living and non-living things along one diagonal, and flying from non-flying things along the other. In the right panel, unit 1 clearly encodes whether a stimulus is a plant or animal, but the behavior of unit 2 considered independently might appear to be arbitrary (activating for *banana* and *cardinal*, but not for *carrot* or *fish*). Joint consideration of both units makes the interpretation of unit 2 clear: if unit 1 is active, it differentiates birds from fish; if inactive, it differentiates fruit from vegetables.

Variation of anatomical location

Within an individual, units representing a given semantic element may be anatomically **contiguous** (situated within the same brain region) or **dispersed** (residing in multiple separate regions; Figure 2C). Methods that analyze different areas separately (e.g., analysis of different ROIs) favor the discovery of contiguous over dispersed representations. Finally, irrespective of whether units are contiguous or dispersed within an individual, signal-carrying units may be anatomically localized in the same or different areas across individuals. Averaging data across anatomically aligned brains (e.g., in searchlight analyses) favors the discovery of similarly over differently localized representations, whereas techniques that align on the basis of similar responses to stimuli rather than anatomical location (e.g., hyper-alignment) relax the localization assumption.

Together these factors delineate 24 different possibilities for the organization of the neuro-semantic code within and across individuals (Table 1). These are not mutually exclusive – different aspects of a representation, or representations in different conceptual domains, may be organized according to different principles. Understanding which principles best explain which aspects of representation thus requires methods capable of finding each variety of signal.

Assumptions implicit in analytic approaches

We next consider how different analytic approaches in functional brain imaging might favor the evaluation of some hypotheses over others. Such studies aim to find the units whose measured responses to stimuli encode the representational elements specified by the cognitive theory.

Table 1. Twenty-four hypotheses about the nature and anatomical organization of the neuro-semantic code[a]

Code	Within subject		Across subjects		Single voxel	Spatial blurring	ROI/SL	Average before model fitting	Average after model fitting
Type	Code[b]	Location	Code	Location	*n* = 46	*n* = 40	*n* = 63	*n* = 45	*n* = 64
Independent	Homo	Contiguous	Consistent	Same	**100**	**100**	**100**	**100**	**100**
Independent	Homo	Contiguous	Consistent	Different	**100**	**100**	**100**	**100**	*10*
Independent	Homo	Contiguous	Inconsistent	Same	**100**	**100**	**100**	*62*	**62**
Independent	Homo	Contiguous	Inconsistent	Different	**100**	**100**	**100**	*62*	*9*
Independent	Homo	Dispersed	Consistent	Same	**100**	**100**	*42*	**42**	**42**
Independent	Homo	Dispersed	Consistent	Different	**100**	**100**	*42*	**42**	*9*
Independent	Homo	Dispersed	Inconsistent	Same	**100**	**100**	*42*	*23*	**23**
Independent	Homo	Dispersed	Inconsistent	Different	**100**	**100**	*42*	*23*	*8*
Independent	Hetero	Contiguous	Consistent	Same	**100**	*60*	**60**	**60**	**60**
Independent	Hetero	Contiguous	Consistent	Different	**100**	*60*	**60**	**60**	*9*
Independent	Hetero	Contiguous	Inconsistent	Same	**100**	*60*	**60**	*36*	**36**
Independent	Hetero	Contiguous	Inconsistent	Different	**100**	*60*	**60**	*36*	*8*
Independent	Hetero	Dispersed	Consistent	Same	**100**	*60*	*30*	**30**	**30**
Independent	Hetero	Dispersed	Consistent	Different	**100**	*60*	*30*	**30**	*8*
Independent	Hetero	Dispersed	Inconsistent	Same	**100**	*60*	*30*	*17*	**17**
Independent	Hetero	Dispersed	Inconsistent	Different	**100**	*60*	*30*	*17*	*7*
Conjoint	Hetero	Contiguous	Consistent	Same	*46*	*23*	**23**	**23**	**23**
Conjoint	Hetero	Contiguous	Consistent	Different	*46*	*23*	**23**	**23**	*2*
Conjoint	Hetero	Contiguous	Inconsistent	Same	*46*	*23*	**23**	*15*	**15**
Conjoint	Hetero	Contiguous	Inconsistent	Different	*46*	*23*	**23**	*15*	*2*
Conjoint	Hetero	Dispersed	Consistent	Same	*46*	*23*	*3*	**3**	**3**
Conjoint	Hetero	Dispersed	Consistent	Different	*46*	*23*	*3*	**3**	*1*
Conjoint	Hetero	Dispersed	Inconsistent	Same	*46*	*23*	*3*	*3*	**3**
Conjoint	Hetero	Dispersed	Inconsistent	Different	*46*	*23*	*3*	*3*	*1*

[a]Each row indicates one hypothesis and the first five columns show corresponding combinations of key factors discussed in the text (code type, within-subject homogeneity and localization, and between-subject consistency and localization). The remaining columns summarize a review of 100 papers using multivariate methods to uncover neuro-semantic representations. Each column represents a common analysis step that entails an implicit assumption about the neural code, including independent analysis of single voxels (assuming an independent code), spatial blurring of BOLD (assuming a homogeneous code), independent consideration of different areas via ROI or searchlight (assuming contiguous localization within area), averaging the neural signal across subjects before model fitting (assuming a consistent code), and averaging of model fit data across subjects (assuming similar localization). The *n* indicates how many papers adopted the corresponding step. Emphasis shows hypotheses where the associated step will benefit (bold font) or hinder (italic) discovery. The numbers indicate how many reports are capable of detecting each possible neural code considering the analysis decisions taken at each step from left to right. The final column indicates the number of reports that adopt choices capable of finding each possible code.
[b]Abbreviations: Hetero, heterogeneous; Homo, homogeneous.

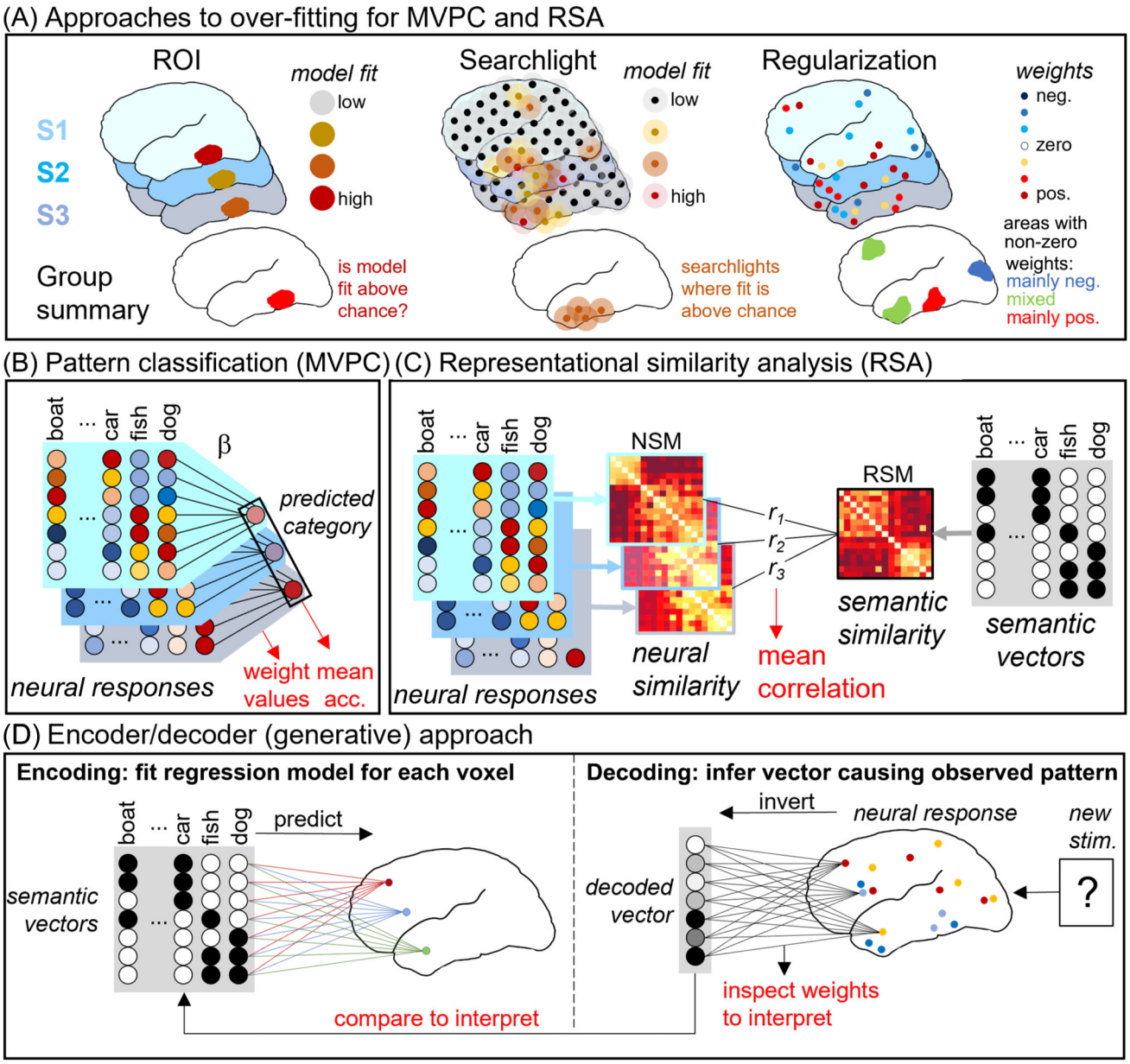
Because all imaging methods yield thousands of noisy measurements for each stimulus in each participant, statistical models that seek informative units must be constrained in some way. Multivariate methods vary in their approach to this problem and thus in their ability to detect different types of representations. We consider three broad approaches and their variants (Figure 3) with an eye to highlighting their respective strengths and limitations. Box 2 additionally considers crucial but commonly overlooked issues for collecting the data that feed these different approaches.

Multivariate pattern classification (MVPC) fits models (Gaussian naive Bayes, support vector machines, logistic/multinomial regression, etc.) to categorize stimuli from the neural activity they evoke [57,58]. During a training phase, the model receives **labeled data** consisting of the neural responses across units to each of many stimuli (e.g., various images of objects) and, for each item, a label indicating the stimulus category. Training involves fitting classifier weights to output the correct label for each item in the training set. The trained model is then evaluated by assessing whether it outputs the correct category label when given neural responses for test stimuli that are not present in the training set. Where a fitted model reliably classifies held-out items, input units are interpreted as encoding information about the target categories. The approach is transparently consistent with category-based semantic representations but will also yield positive results for both feature-based and vector space representations provided that the target categories are separable in the corresponding neural activation patterns (i.e., it is possible to fit a flat hyperplane that reliably divides the target categories in the high-dimensional representation space). Because the output of a classifier depends on activation patterns across multiple units, MVPC can detect both independent and conjoint codes. Classifiers assign unique weights to each unit, and the approach can therefore detect both homogeneous and heterogeneous codes. Because separate classifiers are typically fitted for each participant, the method can potentially find inconsistent and variably localized representations as well.

A key challenge for MVPC concerns over-fitting. With more predictors (neural measurements) than datapoints (stimuli), model fitting is underdetermined without additional constraint – even with random data, an infinite set of coefficients will perfectly predict the category membership of training items [35]. MVPC variants differ in the constraints they impose to handle this issue; this has important implications for signal discovery (Figure 3A).

One method is to reduce the number of neural features provided as the input to the model by applying an explicit anatomical constraint. For instance, ROI-based approaches look only at the units contained in a predefined ROI – discovery therefore requires that the representation is anatomically contiguous and localized similarly across individuals, and also that a sufficient amount of the representation falls within the preselected region to drive classifier accuracy above chance. ROI selection also crucially determines how neural evidence can relate to the space of cognitive hypotheses. For instance, ROIs falling outside modality-specific areas cannot offer evidence relevant to testing grounded theories of representation, whereas those falling solely within a given modality-specific region cannot evaluate self-contained hypotheses.

Relatedly, searchlight approaches fit a separate classifier at each spatial location in each participant (e.g., each voxel, source, or electrode), including as predictors all units within a prespecified anatomical radius ('searchlight') [58,59]. Thus, different brain regions are analyzed separately. Typically cross-participant univariate statistics at each location assess where in the brain the classifier hold-out accuracy is reliably better than chance; this approach therefore requires that the representation is localized similarly across individuals. If this criterion is met, the searchlight can reveal anatomically dispersed codes, but only if each searchlight independently contains sufficient information to drive classifier accuracy above chance. If accurate classification depends



Trends in Cognitive Sciences

Figure 3. Approaches to neural decoding. (A) Different solutions to the over-fitting problem faced by multivariate pattern classification (MVPC) and representational similarity analysis (RSA) approaches. Region of interest (ROI) approaches look only at a prespecified area in each participant and evaluate whether the mean model fit (i.e., hold-out error or correlation) across participants differs reliably from chance. Searchlight methods independently evaluate model fit at many 'searchlights' throughout the brain in each participant, then find areas where searchlights produce above-chance fits reliably across participants. Regularization fits a single model in each participant using all neural features, but constrains the model to minimize prediction error jointly with an additional cost that prevents over-fitting (discussed in the main text). Non-zero coefficients in the decoding model of a subject indicate neural units that carry signal; these can be distributed across the brain and can be different for each participant. Group maps indicate areas where non-zero coefficients accumulate more than expected by chance across individuals. (B) Multivariate pattern classification fits a model to predict a stimulus category label from the neural pattern it evokes across selected neural units. Mean hold-out accuracy across participants indicates whether the selected units carry category information and classifier weights can indicate whether category membership is signaled by increased or decreased neural activation. (C) RSA computes similarity in the neural responses generated across selected units by various stimuli, and then correlates this with a target semantic similarity matrix. Mean correlation across subjects indicates whether the selected neural units encode semantic structure. (D) Generative approaches use regression to

(Figure legend continued at the bottom of the next page.)

Box 2. Implications for data acquisition

Hypotheses about the cognitive and neural systems supporting semantic cognition have crucial implications, not only for how neural data are analyzed, but also for how data are collected.

Stimulus selection

Each modality of stimulus has advantages and disadvantages. Words are easily presented in the scanner, allow all concept types to be probed, and have a perceptual/orthographic structure that is unconfounded with semantic structure. However, decoding is less successful with words than with picture stimuli generally [82] and written words generate a strongly asymmetric (left hemisphere) distribution of activation that contrasts with the bilateral pattern found for pictures and spoken words [119].

Task selection

Tasks used to elicit semantic activation vary across studies in ways that are known to strongly impact the engagement of underlying neural systems, including their overall difficulty [120], the specificity with which an item must be identified for good performance [121], reliance on strongly versus weakly encoded information [122], aspects of knowledge the task foregrounds [25,123], and the degree to which the task can be performed via alternative, non-semantic processing routes [124].

Temporal and spatial resolution

Neuroimaging methods vary in spatial and temporal resolution, limitations that may or may not affect discovery depending on the nature of the underlying code. For instance, the lag in BOLD means that successive stimuli blend into one another in fast event-related designs, which can hinder discovery if the neural code is heterogeneous. Slow event-related methods avoid temporal blending [125] but cannot be used for richer tasks such as connected speech or movie-viewing. EEG and MEG offer higher temporal resolution and thus avoid stimulus-to-stimulus blending, but at the cost of spatial blending that can compromise discovery if the neural code is heterogeneous or anatomically dispersed. ECoG offers temporal and spatial precision, but only a minority of regions are ever probed because the sensors are placed for clinical need and only in patients who need neurosurgical intervention.

Image acquisition

The possibility that semantic representations are anatomically dispersed must be tested with whole-brain imaging, thus posing a challenge for fMRI acquisition where the signal-to-noise ratio varies substantially across the brain [126]. Standard sequences yield especially poor signal in orbitofrontal and ventral anterior temporal regions that are thought to be crucial for semantic cognition [127]. Strategies for improving the signal, including distortion-corrected spin-echo [127,128] and multi-echo protocols [129,130], have been available for several years but have only rarely been applied in semantic studies [131]. Indeed, many studies have restricted the field of view to exclude ventral anterior temporal lobe (ATL) completely [132].

on joint consideration of units that fall in separate searchlights, the code will be missed. In this sense, the searchlight may fail to find dispersed, conjoint codes [5,60].

Note that, in principle, classifier accuracy for searchlights and ROIs could be analyzed separately in each individual, relaxing the assumption of similar localization across participants. We are not aware of such an approach being applied to semantic decoding and we therefore focus on the more usual method of using cross-subject univariate statistics to create group-level information maps for these approaches.

A second approach chooses classifier inputs based on a summary univariate statistic that is computed independently for each unit (such as an *F*-statistic that contrasts unit activation for different category members [3], or a correlation-based metric that assesses the stability of the response of a voxel across stimuli [61]). This avoids the anatomical assumptions of ROI and searchlight

fit models that predict the response of each neural unit to various stimuli. After fitting, the regression weights can be inspected to determine the information that each unit encodes, and novel brain responses can be 'decoded' by finding the semantic vector most likely to have generated the observed neural pattern and then comparing this to known semantic vectors. Abbreviations: acc., accuracy; Neg., negative; NSM, neural similarity matrix; Pos., positive; RSM, representational similarity matrix; S1–S3, brains from three different subjects; stim., stimulus.

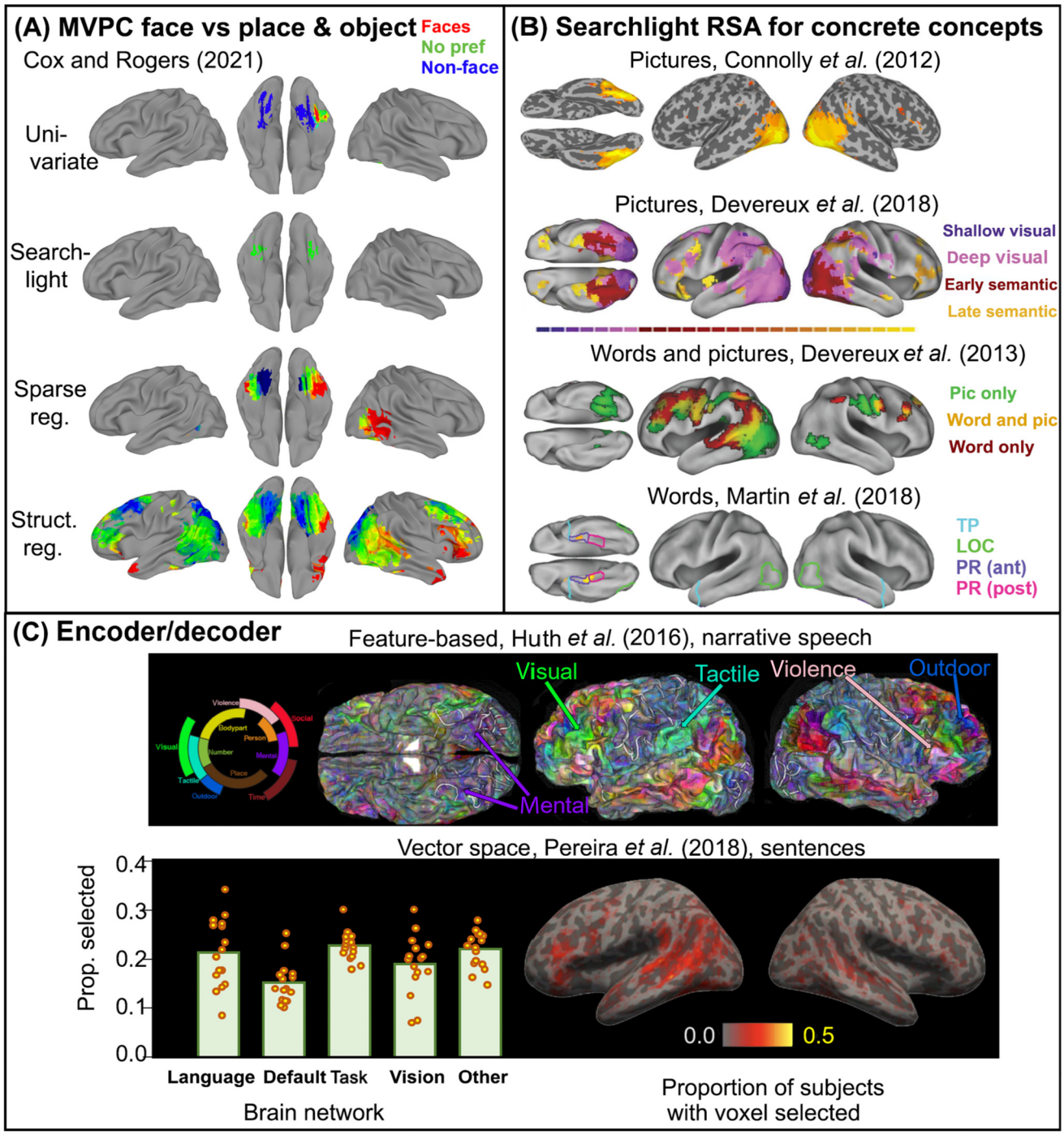
approaches but lacks a principled rationale for setting a cut-off threshold and may fail to discover conjoint representations because each included unit must independently survive the preselection criterion.

A third strategy employs model **regularization**: all units in the cortex provide input to the classifier, which avoids over-fitting by jointly minimizing classification error and an additional loss that is itself a function of the classifier weights [5]. Common losses include the sum of the squared coefficients (L2-norm, also known as 'ridge' regression [62]), the sum of their absolute values {L1-norm, also known as 'LASSO' (least absolute shrinkage and selection operator) [63]}, or a weighted average of these (also known as 'elastic net' [64]). The approach makes no assumption about the anatomical location of signal-carrying units within or across participants, can detect conjoint representations (because it does not require independent preselection of classifier units), and offers a principled way to guide parameterization via nested cross-validation of prediction error [5].

Crucially, however, different regularizers impose different constraints on model fitting, leading to wildly different solutions [5]. Regularization with the L1 norm zeros out as many predictors as possible while still maximizing predictive accuracy, and typically 'selects' (i.e., places non-zero coefficients on) a very small proportion of units. By contrast, the L2 norm spreads similar weights across correlated units and places non-zero weights on all units. The choice of regularizer thus implements an assumption about the likely nature of the true signal: that signal-carrying units are sparse and uncorrelated (L1) or that they are dense and highly redundant (L2). An alternative approach designs loss functions that explicitly incorporate prior knowledge about the likely neural and cognitive structure. For instance, the sparse overlapping sets (SOS) LASSO penalty encourages patterns of 'structured sparsity' where selected units reside in roughly similar locations across participants, promoting loose anatomical clustering that still permits some variation in signal location across participants [65,66].

These differences can yield radically different views of the neuro-semantic code when applied to the same data. In Figure 4A, neural representations of face stimuli appear to be increasingly widely distributed and heterogeneous as analytic methods progressively relax tacit assumptions about the independence, heterogeneity, and localization of the neural code. Standard univariate contrast (assuming a consistently localized, independent, and homogeneous code) replicates the classic finding of a right-lateralized posterior fusiform area that is more active for faces. Searchlight (assuming a similarly localized and contiguous but potentially conjoint and heterogeneous code) suggests a bilateral representation localized to posterior ventral temporal cortex. Whole-brain MVPC regularized with the L1 norm (assuming a sparse code that can be dispersed, heterogeneous, and differently localized) shows a bilateral face-to-nonface gradient in posterior ventral temporal cortex and a face-selective region in right lateral occipital cortex. Regularization with the SOS LASSO (allowing dispersed, heterogeneous, and differently localized codes, but preferring solutions with roughly similar anatomical distributions) suggests a much more broadly distributed code encompassing anterior temporal, parietal, and prefrontal regions in both hemispheres [5].

Representational similarity analysis (RSA) searches for sets of units whose responses express semantic similarities among stimuli [58,59,67]. The analysis first computes a target representational similarity matrix (RSM; sometimes defined in terms of dissimilarity where it is called a target representational dissimilarity matrix) that expresses semantic relatedness for all pairs of stimuli (Box 1). It then estimates a neural similarity matrix (NSM; sometimes called a neural representational dissimilarity matrix) that encodes pairwise similarities in stimulus-evoked neural activity across a set of units. The correlation between RSM and NSM indicates whether the selected units encode the target structure (Figure 3C).



Trends in Cognitive Sciences

Figure 4. Example results from various decoding methods applied to fMRI data. (A) Four different multivariate pattern classification (MVPC) approaches applied to the same dataset. Participants made pleasantness judgments in response to images of faces, places, or objects, and each analysis sought voxel sets that differentiate face from non-face stimuli. Approaches that assume consistently localized signals (univariate and searchlight) suggest that representations are localized to posterior ventrotemporal cortex, whole-brain decoding with sparse regularization suggests a somewhat more distributed representation, whereas decoding with structured sparsity suggests a widely distributed representation [5]. (B) Searchlight representational similarity analysis (RSA) decoding of semantic structure from pictures, words, or both.

(Figure legend continued at the bottom of the next page.)

Similarly to MVPC, RSA can detect categorical, feature-based, and vector space representations provided that the NSM and semantic RSM correlate positively. Because neural similarities are computed across multiple units, the technique can detect conjoint or independent codes and heterogeneous or homogeneous codes. A central challenge concerns how neural units are selected and evaluated for significance. Most studies employ either a prespecified ROI or a searchlight technique. The correlation between RSM and NSM is computed for each ROI or searchlight individually in each participant and, if these are reliably positive across individuals, the ROI/searchlight is interpreted as encoding semantic structure. As with MVPC, information maps could be analyzed separately in each individual, but RSA as typically practiced requires that (i) representations are localized similarity across individuals, (i) information is not conjointly encoded across different searchlights or ROIs, and (iii) individual searchlights contain sufficient information to drive correlations with the target matrix reliably above chance.

RSA views even small correlations as meaningful provided that they are reliably positive across participants. Because semantic structure covaries with many confounding factors, the results can be difficult to interpret. For instance, early studies using visual stimuli suggested that posterior temporo-occipital areas encode semantic structure [68], but a recent comparative analysis found that these areas more strongly encode high-order visual structure and semantic structure was better encoded in more anterior ventro-temporal regions (Figure 4B, top) [69]. Studies that do not control for visual similarity suggest that semantic structure for both words and pictures is encoded within a left perisylvian network [70], but when stimuli orthogonally vary semantic and visual similarity, semantic structure for words appears to be localized to the medial-ventral anterior temporal lobe [71] (Figure 4B, bottom). Thus, very different patterns are obtained depending upon the target RSMs, the selection of stimuli, and the input modality (Box 2).

Finally, encoder/decoder (also known as generative) approaches use regression to fit a separate encoding model for each unit, predicting its response to a stimulus from the semantic features of the item [72–74]. Successful prediction indicates that the corresponding unit independently encodes semantic information. A whole-brain response can be estimated by passing a stimulus feature vector forward through each encoder, yielding a predicted activation at every unit [72]. Alternatively, the whole-brain response generated by a new, unknown item can be decoded by inverting the encoding models to find the semantic vector most likely to have generated the observed neural response, and then interpreting the resulting vector [1,74] (Figure 3D). Because separate models are fitted for each voxel and participant, generative approaches make no assumption about code homogeneity, cross-participant consistency, or anatomical organization within or across individuals. However, they do face two non-trivial challenges.

First, generative approaches can fail to predict the independent activity of a unit that forms part of a conjoint code. To see this, consider the second conjoint example in Figure 2B right, where two units both contribute to a semantic representation. If unit 1 is active, unit 2 differentiates fish from

Results vary remarkably depending on several factors, including the representational similarity matrices (RSMs) considered (semantic similarity alone [68] produces different results from comparing semantic versus visual similarity; top two images [69]) and experimental control of stimulus properties (semantic structure for words appears to be encoded in perisylvian regions when visual structure is uncontrolled [70], but in ventral anterior temporal lobe (ATL) when controlled [71]). (C) Generative approaches for decoding semantic representations of narrative speech/sentences. When predictor vectors have semantically interpretable dimensions, and encoder weights are used to interpret the meaning of a voxel's activation, the results seem to show a mosaic of localized semantic features across cortex within each subject, but callouts show areas where the proposed semantic content is at odds with traditional understanding of function (top; images generated from online visualization tool at https://gallantlab.org/huth2016/). Approaches that invert encoding models to decode whole-brain states (bottom) can recover sentence meanings with good accuracy, but the nature of the underlying code is difficult to discern because the approach selects thousands of voxels widely distributed across cortex in each participant (right), with approximately equal proportions residing in various pre-defined brain networks [1] (left). In both cases verbal semantic representations appear to be widely distributed across cortex and highly variable across individuals. For references see [1,5,68–71,75]. Abbreviations: ant, anterior; LOC, lateral occipital complex; Pic, picture; post, posterior; pref, preference; PR, perirhinal cortex; Prop., proportion; reg., regularization; TP, temporal pole.

birds; if inactive, unit 2 instead differentiates fruits from vegetables. The 'meaning' of unit 2 is clear when unit 1 is taken into consideration, but might appear arbitrary when considered independently. An encoder model might struggle to predict the independent behavior of unit 2 from semantic features such as *can move*, *has feathers*, *is sweet*, etc., and thus might suggest that it is not involved in semantic representation.

The second challenge concerns interpretation. One strategy fits the encoders using semantic vectors whose elements are each individually interpretable (such as a semantic feature vector; Box 1), and then inspects the encoder weights for each unit to understand what content it encodes [2,75,76]. For instance, if the activation of a voxel is reliably predicted by semantic features such as *can move*, *can grow*, and *has eyes*, these features will receive non-zero weights in the regression model for that voxel, which might then be interpreted as encoding animacy. The goal is to understand each unit as independently encoding a subset of semantic features, thereby yielding an interpretable semantic feature map of cortex that is consistent with feature-based cognitive models. Because there are many potential semantic features, however, the encoder fit must be regularized using techniques such as those described earlier for MVPC (commonly L2 norm, e.g., [16], although other approaches are also popular, e.g., [77]). As we have seen, different regularizers can produce dramatically different configurations of weights, and the interpretation of encoder weights therefore hinges crucially upon the choice of the regularizer. Perhaps for this reason, approaches adopting this strategy have yielded puzzling findings – suggesting a mosaic-like organization of local semantic features across many cortical areas that is difficult to reconcile with the wealth of cognitive and clinical neuroscience information about the functions of these regions [75] (Figure 4C, top).

An alternative strategy eschews the effort to identify a 'meaning' for individual units and instead decodes the full activation pattern evoked across cortical units by inverting the encoder models to find the semantic vector that is most likely to have generated the whole-brain response. The recovered vector is interpreted by comparing its similarity to vectors corresponding to known words or sentences [1,74]. For instance, if the decoded vector is near to the known vectors for *grow*, *move*, *eat*, *eyes*, *legs*, *fur*, it will be interpreted as encoding a meaning such as *animal*. Because no effort is made to interpret each dimension, this method is consistent with vector space approaches, but can also detect category or feature-based representations. One recent study showed remarkably good decoding of sentence-level meaning using this approach [1] – but the implications of the study for understanding neural organization of semantics remain unclear because the results identified thousands of voxels scattered across the cortex in each individual, with approximately equal involvement of many different brain networks and no voxels selected in more than half of the participants (Figure 4C, bottom).

It is worth noting that each general approach encompasses several variants – for instance, in the particular classification model adopted by MVPC [58] and the specific similarity metric used by RSA [78,79]. Although a full characterization of each is beyond the scope of this review, it seems likely that such variation further contributes to the heterogeneity of the findings reported in the literature.

Analytic implications of grounded versus self-contained theories

The issues described above arise regardless of whether neuro-semantic representations are grounded or self-contained, but this important distinction in cognitive theories carries two additional implications for the design, analysis, and interpretation of multivariate imaging studies. First, primary and secondary perceptual and motor cortices conform to localization assumptions that are central to particular analytic choices – specifically, such areas are both contiguous and

localized similarly across individuals. Grounded approaches suggest that such areas can encode semantic information about stimuli, and studies designed specifically to assess whether semantic structure arises within a given modality [80,81] therefore have good motivation to employ ROI or searchlight-based feature selection. The anatomical organization of tertiary and association cortices is less well understood and may be more likely to vary across individuals, therefore studies seeking semantic structure outside the earlier modality-specific regions are better served by the adoption of approaches that loosen localization, homogeneity, and consistency assumptions. Assessment of self-contained hypotheses will depend crucially on such methods because they propose that semantic representations encode information in a modality-independent manner.

Second, adjudication of grounded versus self-contained hypotheses requires studies that probe semantic information through different stimulus modalities. Self-contained views hold that the same system of semantic representation is engaged regardless of whether the stimulus is a word, picture, image, sound, etc. Such a view cannot be disconfirmed by evidence that, for instance, semantic information is decodable from visual areas when a visual stimulus appears because such a result might also arise if the structure of purely perceptual visual representations is confounded with semantic structure (e.g., Figure 4B). Evaluating the proposal instead requires searching for neural systems from which semantic information can be decoded across multiple different stimulus modalities. Currently, the literature contains relatively few such studies, and these have yielded mixed findings [70,82–84] (further details are given in the supplemental information online).

Toward best practices

To understand how the preceding issues may have shaped current thinking about semantic representation in mind and brain, we reviewed 100 papers applying multivariate techniques to the discovery of neuro-semantic representations in fMRI data (supplemental information). For each, we considered five analytic decisions, each reflecting a latent assumption about the neural code, and we evaluated which of the 24 representational possibilities the study was capable of detecting as each choice was made. The results are summarized in Table 1. All methods were capable of detecting neural representations that adopt an independent, homogeneous, and anatomically contiguous code that, across individuals, is consistent and similarly localized – the type of representation sought by univariate analysis. Fewer could detect other types of representational structure, and very few were capable of finding representations that are dispersed in the brain, localized differently across participants, and/or encode semantic information conjointly across units rather than independently. In this sense, methodological choices made during data analysis determine which types of neural signal can and cannot be detected – the analytic decisions effectively filter the empirical record.

A central question thus concerns how the field might best proceed given the complexity and heterogeneity of contemporary methods and the filtering that inevitably results. No analytic approach is assumption-free, and we doubt that the universal adoption of any single method will resolve the issues we have identified. Instead, we believe the field would be well served by adopting some best practices in the way that studies are designed and results are communicated.

Articulating explicit hypotheses about the neural code

In laying out the motivation and design of a study, it is helpful for researchers to explicitly state their working hypothesis about the nature and structure of the neuro-semantic code – what form the cognitive representation is hypothesized to take, how its neural instantiation is reflected in the measurements taken, and how it is expected to vary within and across individuals. The cognitive and neural possibilities developed in this review provide a frame of reference for such statements,

which are important because they allow the reader to understand why a given analysis method was chosen and how the observed results relate to the working hypothesis.

Explicit consideration of alternative hypotheses

When designing/motivating an analysis and when drawing conclusions from the results, it is helpful for researchers to consider other possible ways that the target information might be encoded in neural activity, beyond the working hypothesis. Before data collection, such habits can prompt new design or analysis ideas that allow adjudication of a richer variety of hypotheses. When drawing conclusions, explicit consideration of alternative possibilities and whether/how the current data can possibly disconfirm them can help the community to better understand seemingly heterogeneous patterns of results.

Connection to neurocognitive computational models

One way to make working assumptions about representation explicit is to connect the experimental design and analysis plan to a neuro-computational model of the behavior [4,5,60,85–88]. Figure 5 shows three recent examples. This connection serves several purposes. First, it provides a bridge between functional imaging results and explicit hypotheses about the mechanisms supporting the behavior of interest, rendering the neural data a supporting part of a broader set of ideas about how the system works. Second, such models can offer new hypotheses about the nature of the neural code that might not otherwise occur to the theorist. Third, neurocomputational models can be used to better understand the strengths and weaknesses of different analytic approaches: the theorist can probe model analogs of neural signals and evaluate whether a given technique is capable of discovering information of the type captured by the model. Fourth, models allow exploration of alternative possibilities – the strengths and limitations of a given approach can be illuminated by comparing and contrasting its results when applied to models that embody different assumptions about the neural signal.

Simplified open data

Multivariate imaging studies pose unique challenges for the open data movement. The path from raw data to published result is often complex, software- or system-dependent, contains default parameterizations that may go unexplained, and involves many intermediate data products between raw measurements and summary results that can be exceedingly large and difficult to document. Any single workflow can require extensive effort for outside scientists to fully understand and, because new approaches arrive with daunting frequency, it is difficult to know which bespoke pathways are worth mastering. Nevertheless, each method we have described makes use, at some level, of common data elements that are easy to understand and not too large to document and share. These include (i) the matrix that encodes, for each subject, the estimated response of each neural unit (voxel, electrode, source, etc.) to each stimulus, (ii) the coordinates of the units in a standard reference frame [e.g., Montreal Neurological Institute (MNI) coordinates of voxels, time and location information for ECoG, etc.], and (iii) meta-information about the stimuli (e.g., category labels used for decoding, semantic feature vectors used in an encoding model, the similarity matrix used for RSA, etc.). Sharing only these elements in standardized form would provide minimally sufficient information for scientists to apply a variety of different techniques to a dataset, thus promoting better understanding of how results vary with the method of analysis.

Convergence with other forms of evidence

Functional imaging alone will not resolve the quest for neuro-semantic representations. A fuller understanding will require relating multivariate imaging results to other diverse sources of evidence in cognitive neuroscience, including (i) the rich neuropsychology literature documenting patterns of verbal and nonverbal semantic impairment and their underlying neuropathology

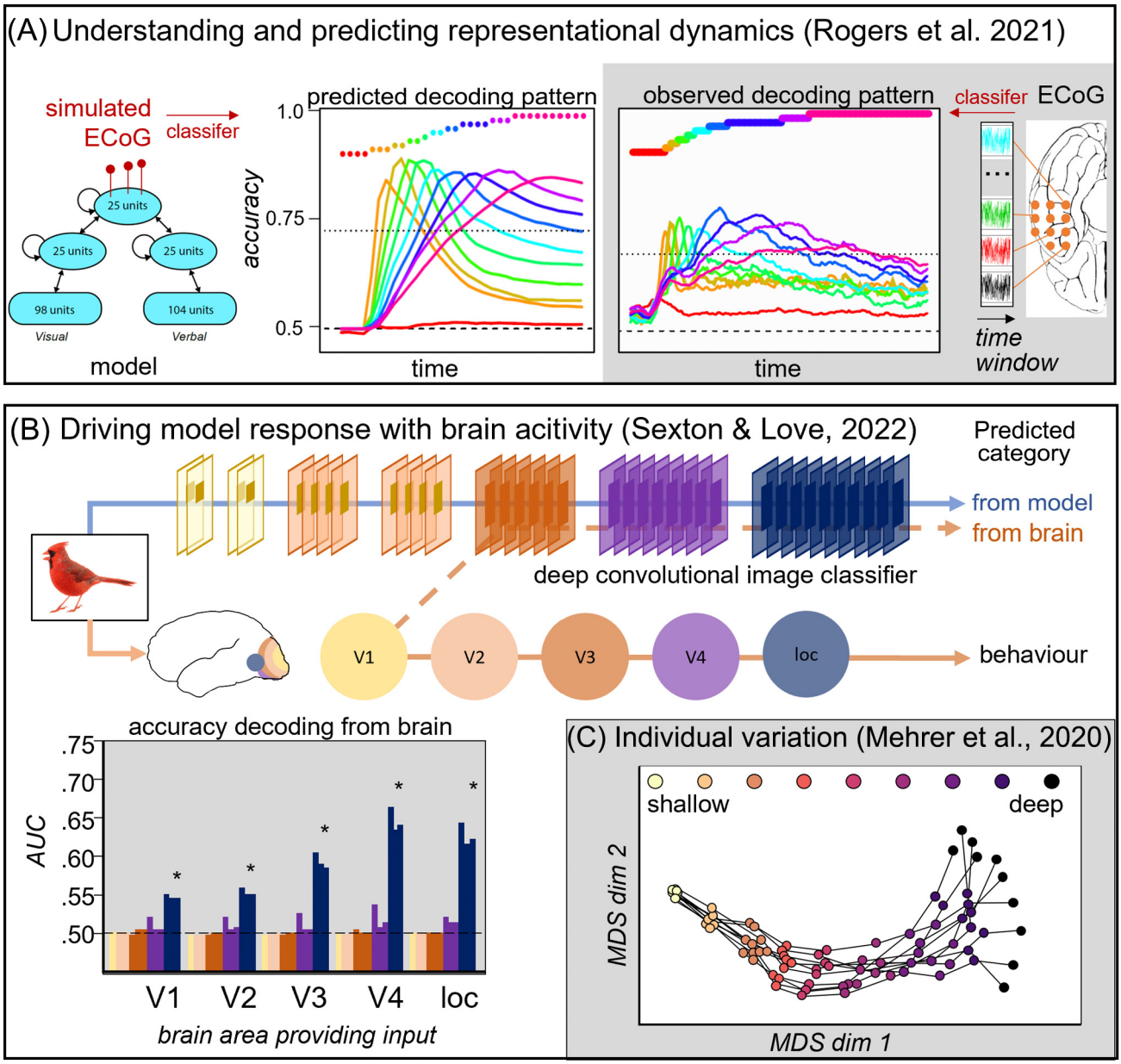


Figure 5. Recent examples of computational models informing neural decoding. (A) In recurrent models the activation patterns that encode semantic information change over the course of stimulus processing. In simulated electrocorticography (ECoG, left), classifiers fit to different temporal windows (colored dots) decode well within the same and neighboring time-windows, but poorly for more distal time-windows (colored lines). A similar pattern arises when the same approach is used to decode ECoG from human anterior temporal cortex while participants name pictures, suggesting rapid nonlinear change in the neuro-semantic code [133]. (B) Deep convolutional neural networks (DCNNs) may provide a useful framework for understanding visual object semantics [134,135]. A recent study assessed whether a trained DCNN could classify images when activations at a given model layer were replaced by neural responses (measured by fMRI) of different visual areas [136]. Neural patterns from each area were successfully decoded, but only when they were input to the deeper model layers (barplot) – suggesting that the richer semantic structure encoded in such layers is reflected throughout the ventral visual stream. (C) Other work uses similar models to evaluate individual differences across parts of the vision-to-semantics system [137]. In the plot shown the authors trained several models, measured similarity in the representational geometry acquired in each layer across models, and embedded these in two dimensions. The proximity of colored circles indicates the similarity of the representational structure acquired by the corresponding layers. Lines connect layers in the same model. Shallower model layers (light colors) always learned relatively similar structure, whereas deeper layers – those most likely to express abstract semantic structure – learned more variable structure, suggesting that neural codes may differ more across individuals in the regions that are most likely to encode semantic structure. For references see [133,136,137]. Abbreviations: AUC, area under the curve; dim, dimension; LOC, lateral occipital complex; MDS, multidimensional scaling; V1–V4, visual cortex areas 1–4.

Box 3. The importance of converging evidence

The heterogeneity of imaging findings may be resolved by considering how conclusions from various studies relate to converging evidence from other methods. Some examples are given below.

Neuropsychology

Several varieties of brain damage cause semantic impairment and distinct deficits are observed depending on the neuropathology. Close consideration of these can illuminate brain imaging results. For instance, cross-modal semantic impairment can arise both from bilateral damage to the anterior temporal lobes (ATLs) [93,138] and from left frontoparietal or posterior-lateral temporal stroke [92,139], but whereas ATL damage erodes conceptual structure, frontoparietal/posterior-lateral temporal damage instead disrupts the ability to shape semantic processing to the task context [54]. Thus, results implicating frontoparietal/posterior lateral temporal areas in semantics might best be interpreted by considering the demands on semantic control, whereas studies seeking conceptual structure in the brain should employ methods that are capable of resolving ATL signal.

Neural disruption

If imaging results suggest that a brain region selectively represents/processes a particular type of semantic information, transient disruption of the area via **transcranial magnetic stimulation (TMS)** should selectively affect retrieval of the target information. For instance, TMS applied to left or right ATL slows semantic judgments equally for animates and inanimates, but does not affect number judgments, supporting the view that bilateral ATLs encode semantic information across domains [95]. Such studies will be especially important for testing the implications of multivariate imaging studies indicative of highly unorthodox semantic functions for various cortical areas [75].

Neural connectivity

The neural response of a given area can reflect its broader connectivity, with implications for understanding its function. For instance, medial posterior fusiform cortex responds more to artifact than animal names – a pattern observed both in sighted and congenitally blind individuals [140,141]. One interpretation suggests that different brain areas natively specialize to represent distinct semantic categories [142]. However, the area of interest is functionally [98] and structurally [100] connected to dorsal areas that aid in object-directed actions, suggesting that the seeming category effect may instead arise from more effective interactions between this visual area and parts of the action system [98,143].

Neurocognitive development

Developmental trajectories can likewise aid the understanding of mature activation patterns. For instance, the right posterior fusiform responds strongly to face images in most literate adults, perhaps suggesting an innately dedicated system for face representation [144,145]. However, face perception engages the fusiform bilaterally in pre-literate children [146], and the left hemifield/right hemisphere advantage for face recognition emerges late in development as a child learns to read [147]. Such data suggest that the mature pattern reflects, not innate specialization for a visual category, but experienced-based tuning of visual perception [101].

Outstanding questions

Which cognitive hypotheses best describe semantic representations? The multivariate methods considered in this review do not indicate whether the underlying representation is categorical, feature-based, or a vector space, or is self-contained versus grounded. MVPC can produce a positive result even if neural representations are vector spaces rather than categories, and RSA can generate a positive result even if neural representations are categories and not vector spaces. How then can brain imaging adjudicate between these views?

When different brain areas all encode semantic structure, what data can determine whether they support the same or different functions? Semantic structure has been observed across multiple brain areas, but disruption caused by brain damage or transcranial magnetic stimulation (TMS) can produce qualitatively different patterns of impairment – suggesting that these regions serve different functions in semantic cognition.

Can imaging data resolve which aspects of a target representational structure are, or are not, encoded within a neural system? Many studies report above-chance decoding that is nevertheless relatively weak (e.g., RSA correlations as small as $r = 0.03$, binary classification accuracy of 0.55, etc.). Such effects might arise because neural data are noisy, because the neural system encodes weak confounds with the target structure, or because it encodes only part of the target structure.

Can a combination of approaches overcome the individual limitations of each method? Each technique has strengths and limitations; perhaps the fullest picture of semantics in the brain will arise from a combination of approaches that will allow the community to evaluate the full space of representational possibilities outlined in this review.

[89–94], (ii) methods for disrupting neural processing in healthy participants, which can provide crucial evidence about causality [95–97], (iii) structural and functional brain connectivity [98–100], (iv) patterns of behavior and functional activation arising over typical and atypical development [101,102], and (v) results of behavioral studies arising in cognitive science [30,103,104]. Box 3 considers how these sources of evidence can aid the interpretation of imaging data. Of course, not every paper can comprehensively review a large and complex literature – but in drawing conclusions it can be helpful for authors to explicitly consider where these cohere with results from other methodologies, where they contradict such results, and where the relevant experiments have not yet been conducted.

Concluding remarks

Our review illustrates that methodological choices in multivariate neuroimaging analysis selectively filter data to promote discovery of some types of neuro-semantic codes over others. These considerations compel a re-evaluation of the literature. Over three decades many neuroimaging studies have reported cortical areas that locally encode a particular type of semantic information in a systematic way across individuals. The preponderance and replicability of such

findings suggest that some elements of neuro-semantic representation must indeed be independent, contiguous, and localized similarly across individuals. However, because this is precisely the one form of neuro-semantic code that, among many possibilities, is most robust to methodological choices, the ubiquity of such findings does not signify that these are the only, or even the most important, elements of semantic representation. On the contrary, neurocomputational models of healthy and disordered semantic cognition typically acquire internal representations that are conjoint rather than independent, are distributed across units that may be anatomically dispersed, are heterogeneous in code, and are potentially localized differently across individuals [5,60,85]. These latter forms of semantic representation are the least likely to be revealed by most current analytical methods. The few studies capable of finding such structure often reveal a more widely distributed, heterogeneous, and variable semantic code than other studies suggest [1,5,75]. Thus there exists an important lacuna in the empirical landscape that must be filled if we are to develop a mechanistic understanding of semantic cognition in the brain. We hope that this article provides a first step toward an organizing framework that can bring the current heterogeneity of findings under a common explanatory umbrella (see Outstanding questions).

Acknowledgments

This work was supported by an MRC Career Development Award (MR/V031481/1) to A.D.H., by a grant from the Rosetrees Trust (A1699) to A.D.H. and M.A.L.R., and by an Advanced European Research Council (ERC) award (GAP 670428-30 BRAIN2MIND_NEUROCOMP), MRC programme grant (MR/R023883/1), and intramural funding (MC_UU_00005/18) to M.A.L.R.

Declaration of interests

The authors declare no conflicts of interest.

Supplemental information

Supplemental information associated with this article can be found online at https://doi.org/10.1016/j.tics.2022.12.006.

References

1. Pereira, F. *et al.* (2018) Toward a universal decoder of linguistic meaning from brain activation. *Nat. Commun.* 9, 1–13
2. Popham, S.F. *et al.* (2021) Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nat. Neurosci.* 24, 1628–1636
3. Visconti di Oleggio Castello, M. Visconti *et al.* (2021) Shared neural codes for visual and semantic information about familiar faces in a common representational space. *Proc. Natl. Acad. Sci. U. S. A.* 118, e2110474118
4. Kriegeskorte, N. (2015) Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annu. Rev. Vis. Sci.* 1, 417–446
5. Cox, C.R. and Rogers, T.T. (2021) Finding distributed needles in neural haystacks. *J. Neurosci.* 41, 1019–1032
6. Mandler, J.M. (2006) *The Foundations of Mind: Origins of Conceptual Thought* (1st edn), Oxford University Press
7. Pauen, S. (2002) Evidence for knowledge-based category discrimination in infancy. *Child Dev.* 73, 1016–1033
8. Pauen, S. (2002) The global-to-basic shift in infants' categorical thinking: first evidence from a longitudinal study. *Int. J. Behav. Dev.* 26, 492–499
9. Rogers, T.T. *et al.* (2004) The structure and deterioration of semantic memory: a computational and neuropsychological investigation. *Psychol. Rev.* 111, 205–235
10. Lopez, A. *et al.* (1997) The tree of life: universal and cultural features of folkbiological taxonomies and inductions. *Cogn. Psychol.* 32, 251–295
11. Hodges, J.R. *et al.* (1995) Charting the progression in semantic dementia: implications for the organisation of semantic memory. *Memory* 3, 463–495
12. Waxman, S.R. and Markow, D.B. (1995) Words as invitations to form categories: evidence from 12- to 13-month-old infants. *Cogn. Psychol.* 29, 257–302
13. Booth, A.E. and Waxman, S.R. (2008) Taking stock as theories of word learning take shape. *Dev. Sci.* 11, 185–194
14. Lin, E.L. and Murphy, G.L. (2001) Thematic relations in adults' concepts. *J. Exp. Psychol. Gen.* 130, 3–28
15. Anderson, J.R. (1991) The adaptive nature of human categorization. *Psychol. Rev.* 98, 409–426
16. Rosch, E. *et al.* (1976) Basic objects in natural categories. *Cogn. Psychol.* 8, 382–439
17. Collins, A.M. and Quillian, M.R. (1969) Retrieval time from semantic memory. *J. Verbal Learn. Verbal Behav.* 8, 240–247
18. Jolicoeur, P. *et al.* (1984) Pictures and names: making the connection. *Cogn. Psychol.* 19, 31–53
19. Xu, F. and Tenenbaum, J.B. (2007) Word learning as Bayesian inference. *Psychol. Rev.* 114, 245–272
20. Serre, T. *et al.* (2007) A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci.* 104, 6424–6429
21. Humphreys, G.W. and Forde, E.M. (2001) Hierarchies, similarity, and interactivity in object-recognition: on the multiplicity of 'category-specific' deficits in neuropsychological populations. *Behav. Brain Sci.* 24, 453–509
22. Farah, M.J. and McClelland, J.L. (1991) A computational model of semantic memory impairment: modality-specificity and emergent category-specificity. *J. Exp. Psychol. Gen.* 120, 339–357
23. Cree, G. *et al.* (1999) An attractor model of lexical conceptual processing: simulating semantic priming. *Cogn. Sci.* 23, 371–414

24. Tyler, L. *et al.* (2000) Conceptual structure and the structure of concepts: a distributed account of category-specific deficits. *Brain Lang.* 75, 195–231
25. Martin, A. (2007) The representation of object concepts in the brain. *Annu. Rev. Psychol.* 58, 25–45
26. Anderson, A.J. *et al.* (2019) An integrated neural decoder of linguistic and experiential meaning. *J. Neurosci.* 39, 8969–8987
27. McRae, K. *et al.* (1997) On the nature and scope of featural representations of word meaning. *J. Exp. Psychol. Gen.* 126, 99–130
28. Ruts, W. *et al.* (2004) Dutch norm data for 13 semantic categories and 338 exemplars. *Behav. Res. Methods Instrum. Comput.* 36, 506–515
29. Mervis, C.B. and Rosch, E. (1981) Categorization of natural objects. *Annu. Rev. Psychol.* 32, 89–115
30. Mack, M. and Palmeri, T. (2011) The timing of visual object categorization. *Front. Psychol.* 2, 165
31. Collins, A.M. and Loftus, E.F. (1975) A spreading-activation theory of semantic processing. *Psychol. Rev.* 82, 407–428
32. Riesenhuber, M. and Poggio, T. (1999) Hierarchical models of object recognition in cortex. *Nat. Neurosci.* 2, 1019–1025
33. Landauer, T.K. and Dumais, S.T. (1997) A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* 104, 211–240
34. Mikolov, T. *et al.* (2013) Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems* (Vol. 2), pp. 3111–3119, Curran Associates Inc.
35. Pereira, F. *et al.* (2016) A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cogn. Neuropsychol.* 33, 175–190
36. Katz, J. (1972) *Semantic Theory*, Addison-Wesley Educational Publishers
37. Rosch, E. (1978) Principles of categorization. In *Cognition and Categorization* (Lloyd, B. and Rosch, E., eds), pp. 27–48, Lawrence Erlbaum Associates
38. Hampton, J.A. (2015) Categories, prototype and exemplars. In *The Routledge Handbook of Semantics* (Riemer, N., ed.), pp. 141–157, Routledge
39. Rotaru, A.S. *et al.* (2018) Modeling the structure and dynamics of semantic processing. *Cogn. Sci.* 42, 2890–2917
40. Kumar, A.A. *et al.* (2022) A critical review of network-based and distributional approaches to semantic memory structure and processes. *Top. Cogn. Sci.* 14, 54–77
41. Griffiths, T.L. *et al.* (2007) Topics in semantic representation. *Psychol. Rev.* 114, 211–244
42. Derby, S. *et al.* (2018) Using sparse semantic embeddings learned from multimodal text and image data to model human conceptual knowledge. *ArXiv* Published online September 7, 2018. https://doi.org/10.48550/arXiv.1809.02534
43. Burgess, C. and Lund, K. (1997) Modelling parsing constraints with high-dimensional context space. *Lang. Cogn. Process.* 12, 177–210
44. Devlin, J. *et al.* (2019) BERT: pre-training of deep bidirectional transformers for language understanding. *ArXiv* Published online May 24, 2019. https://doi.org/10.48550/arXiv.1810.04805
45. Barsalou, L.W. (2008) Grounded cognition. *Anuual Rev. Psychol.* 59, 617–645
46. Barsalou, L.W. (2003) Situated simulation in the human conceptual system. *Lang. Cogn. Process.* 18, 513–562
47. Glenberg, A.M. and Robertson, D.A. (2000) Symbol grounding and meaning: a comparison of high-dimensional and embodied theories of meaning. *J. Mem. Lang.* 43, 379–401
48. Glenberg, A.M. (2010) Embodiment as a unifying perspective for psychology. *Wiley Interdiscip. Rev. Cogn. Sci.* 1, 586–596
49. Damasio, A.R. (1989) The brain binds entities and events by multiregional activation from convergence zones. *Neural Comput.* 1, 123–132
50. Damasio, H. *et al.* (2004) Neural systems behind word and concept retrieval. *Cognition* 92, 179–229
51. Martin, A. (2016) GRAPES – grounding representations in action, perception, and emotion systems: how object properties and categories are represented in the human brain. *Psychon. Bull. Rev.* 23, 979–990
52. Fernandino, L. *et al.* (2022) Decoding the information structure underlying the neural representation of concepts. *Proc. Natl. Acad. Sci.* 119, e2108091119
53. Patterson, K. *et al.* (2007) Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat. Rev. Neurosci.* 8, 976–987
54. Lambon Ralph, M.A. *et al.* (2017) The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* 18, 42–55
55. Rogers, T.T. and McClelland, J.L. (2004) *Semantic Cognition: A Parallel Distributed Processing Approach*, MIT Press
56. Guntupalli, J.S. *et al.* (2016) A model of representational spaces in human cortex. *Cereb. Cortex* 26, 2919–2934
57. Pereira, F. and Botvinick, M. (2011) Information mapping with pattern classifiers: a comparative study. *NeuroImage* 56, 476–496
58. Norman, K.A. *et al.* (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn. Sci.* 10, 424–430
59. Kriegeskorte, N. *et al.* (2006) Information-based functional brain mapping. *Proc. Natl. Acad. Sci.* 103, 3863–3868
60. Cox, C.R. *et al.* (2015) Connecting functional brain imaging and parallel distributed processing. *Lang. Cogn. Neurosci.* 30, 380–394
61. Vargas, R. and Just, M.A. (2020) Neural representations of abstract concepts: identifying underlying neurosemantic dimensions. *Cereb. Cortex* 30, 2157–2166
62. Hoerl, A.E. and Kennard, R.W. (2000) Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 42, 80–86
63. Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B Methodol.* 58, 267–288
64. Jia, J. and Yu, B. (2008) On model selection consistency of the elastic net when p >> n. *Stat. Sin.* 20, 595–611
65. Rao, N. *et al.* (2013) Sparse overlapping sets lasso for multitask learning and its application to fMRI analysis. *Adv. Neural Inf. Proces. Syst.* 26, 2202–2210
66. Rao, N. *et al.* (2016) Classification with the sparse group lasso. *IEEE Trans. Signal Process.* 64, 448–463
67. Pereira, F. *et al.* (2009) Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage* 45, S199–S209
68. Connolly, A.C. *et al.* (2012) The representation of biological classes in the human brain. *J. Neurosci.* 32, 2608–2618
69. Devereux, B.J. *et al.* (2018) Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Sci. Rep.* 8, 1–12
70. Devereux, B.J. *et al.* (2013) Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *J. Neurosci.* 33, 18906–18916
71. Martin, C.B. *et al.* (2018) Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *eLife* 7, e31873
72. Mitchell, T.M. *et al.* (2008) Predicting human brain activity associated with the meanings of nouns. *Science* 320, 1191–1195
73. Just, M.A. *et al.* (2010) A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS One* 5, e8622
74. Pereira, F. *et al.* (2011) Generating text from functional brain images. *Front. Hum. Neurosci.* 5, 72
75. Huth, A.G. *et al.* (2016) Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* 532, 453–458
76. Huth, A.G. *et al.* (2012) A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* 76, 1210–1224
77. Nunez-Elizalde, A.O. *et al.* (2019) Voxelwise encoding models with non-spherical multivariate normal priors. *NeuroImage* 197, 482–492
78. Haxby, J.V. *et al.* (2014) Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci.* 37, 435–456
79. Diedrichsen, J. and Kriegeskorte, N. (2017) Representational models: a common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Comput. Biol.* 13, e1005508
80. Clarke, A. and Tyler, L.K. (2014) Object-specific semantic coding in human perirhinal cortex. *J. Neurosci.* 34, 4766–4775

81. Carota, F. *et al.* (2021) Category-specific representational patterns in left inferior frontal and temporal cortex reflect similarities and differences in the sensorimotor and distributional properties of concepts. *BioRxiv* Published online September 3, 2021. https://doi.org/10.1101/2021.09.03.458378
82. Shinkareva, S.V. *et al.* (2011) Commonality of neural representations of words and pictures. *NeuroImage* 54, 2418–2425
83. Simanova, I. *et al.* (2014) Modality-independent decoding of semantic information from the human brain. *Cereb. Cortex* 24, 426–434
84. Handjaras, G. *et al.* (2016) How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *NeuroImage* 135, 232–242
85. Rogers, T.T. (2020) Neural networks as a critical level of description for cognitive neuroscience. *Curr. Opin. Behav. Sci.* 32, 167–173
86. Yuste, R. (2015) From the neuron doctrine to neural networks. *Nat. Rev. Neurosci.* 16, 487–497
87. Yang, G.R. *et al.* (2019) Task representations in neural networks trained to perform many cognitive tasks. *Nat. Neurosci.* 22, 297–306
88. Richards, B.A. *et al.* (2019) A deep learning framework for neuroscience. *Nat. Neurosci.* 22, 1761–1770
89. Patterson, K. and Hodges, J. (2000) Semantic dementia: one window on the structure and organisation of semantic memory. In *Handbook of Neuropsychology Vol. 2: Memory and Its Disorders* (Cermak, J., ed.), pp. 313–333, Elsevier Science
90. Caramazza, A. and Mahon, B.Z. (2003) The organization of conceptual knowledge: the evidence from category-specific semantic deficits. *Trends Cogn. Sci.* 7, 354–361
91. Mesulam, M.M. *et al.* (2013) Words and objects at the tip of the left temporal lobe in primary progressive aphasia. *Brain J. Neurol.* 136, 601–618
92. Jefferies, E. and Lambon Ralph, M.A. (2006) Semantic impairment in stroke aphasia versus semantic dementia: a case-series comparison. *Brain* 129, 2132–2147
93. Acosta-Cabronero, J. *et al.* (2011) Atrophy, hypometabolism and white matter abnormalities in semantic dementia tell a coherent story. *Brain* 134, 2025–2035
94. Chen, L. and Rogers, T.T. (2014) Revisiting domain-general accounts of category specificity in mind and brain. *Wiley Interdiscip. Rev. Cogn. Sci.* 5, 327–344
95. Pobric, G. *et al.* (2010) Category-specific versus category-general semantic impairment induced by transcranial magnetic stimulation. *Curr. Biol.* 20, 964–968
96. Pobric, G. *et al.* (2007) Anterior temporal lobes mediate semantic representation: mimicking semantic dementia by using rTMS in normal participants. *Proc. Natl. Acad. Sci. U. S. A.* 104, 20137–20141
97. Lambon Ralph, M.A. *et al.* (2009) Conceptual knowledge is underpinned by the temporal pole bilaterally: convergent evidence from rTMS. *Cereb. Cortex* 19, 832–838
98. Mahon, B.Z. *et al.* (2007) Action-related properties shape object representations in the ventral stream. *Neuron* 55, 507–520
99. Binney, R.J. *et al.* (2012) Convergent connectivity and graded specialization in the rostral human temporal lobe as revealed by diffusion-weighted imaging probabilistic tractography. *J. Cogn. Neurosci.* 24, 1998–2014
100. Chen, L. *et al.* (2017) A unified model of human semantic knowledge and its disorders. *Nat. Hum. Behav.* 1, 1–10
101. Plaut, D.C. and Behrmann, M. (2011) Complementary neural representations for faces and words: a computational exploration. *Cogn. Neuropsychol.* 28, 251–275
102. Behrmann, M. and Plaut, D.C. (2012) Bilateral hemispheric processing of words and faces: evidence from word impairments in prosopagnosia and face impairments in pure alexia. *Cereb. Cortex* 24, 1102–1118
103. Van Rullen, R. and Thorpe, S.J. (2001) Is it a bird? Is it a plane? Ultra-rapid visual categorization of natural and artifactual objects. *Perception* 30, 655–668
104. Rogers, T.T. and Patterson, K. (2007) Object categorization: reversals and explanations of the basic-level advantage. *J. Exp. Psychol. Gen.* 136, 451
105. Rosch, E. (1975) Cognitive representations of semantic categories. *J. Exp. Psychol. Gen.* 104, 192–233
106. Armstrong, S.L. *et al.* (1983) What some concepts might not be. *Cognition* 13, 263–308
107. Caramazza, A. and Shelton, J.R. (1998) Domain-specific knowledge systems in the brain: the animate–inanimate distinction. *J. Cogn. Neurosci.* 10, 1–34
108. Kanwisher, N. (2010) Functional specificity in the human brain: a window into the functional architecture of the mind. *Proc. Natl. Acad. Sci. U. S. A.* 107, 11163–11170
109. Murphy, G. (2002) *The Big Book of Concepts*, MIT Press
110. Murphy, G. and Medin, D.L. (1985) The role of theories in conceptual coherence. *Psychol. Rev.* 92, 289–316
111. McRae, K. *et al.* (2005) Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods Instrum. Comput.* 37, 547–559
113. Landauer, T.K. (1998) Learning and representing verbal meaning: the latent semantic analysis theory. *Curr. Dir. Psychol. Sci.* 7, 161–164
114. Landauer, T.K. *et al.* (1998) An introduction to latent semantic analysis. *Discourse Process.* 25, 259–284
115. Panigrahi, A. *et al.* (2019) Word2Sense: sparse interpretable word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5692–5705, ACL
116. Krizhevsky, A. *et al.* (2017) ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 84–90
117. Simonyan, K. and Zisserman, A. (2015) Very deep convolutional networks for large-scale image recognition. *ArXiv* Published online April 10, 2015. https://doi.org/10.48550/arXiv.1409.1556
118. Floridi, L. and Chiriatti, M. (2020) GPT-3: its nature, scope, limits, and consequences. *Mind. Mach.* 30, 681–694
119. Liuzzi, A.G. *et al.* (2015) Left perirhinal cortex codes for similarity in meaning between written words: comparison with auditory word input. *Neuropsychologia* 76, 4–16
120. Sabsevitz, D.S. *et al.* (2005) Modulation of the semantic system by word imageability. *NeuroImage* 27, 188–200
121. Rogers, T.T. *et al.* (2006) Anterior temporal cortex and semantic memory: reconciling findings from neuropsychology and functional imaging. *Cogn. Affect. Behav. Neurosci.* 6, 201–213
122. Noonan, K.A. *et al.* (2013) Going beyond inferior prefrontal involvement in semantic control: evidence for the additional contribution of dorsal angular gyrus and posterior middle temporal cortex. *J. Cogn. Neurosci.* 25, 1824–1850
123. Chiou, R. *et al.* (2018) Controlled semantic cognition relies upon dynamic and flexible interactions between the executive 'semantic control' and hub-and-spoke 'semantic representation' systems. *Cortex* 103, 100–116
124. Graves, W.W. *et al.* (2010) Neural systems for reading aloud: a multiparametric approach. *Cereb. Cortex* 20, 1799–1815
125. Lewis-Peacock, J.A. and Postle, B.R. (2008) Temporary activation of long-term memory supports working memory. *J. Neurosci.* 28, 8765–8771
126. Liu, T.T. (2016) Noise contributions to the fMRI signal: an overview. *NeuroImage* 143, 141–151
127. Embleton, K.V. *et al.* (2010) Distortion correction for diffusion-weighted MRI tractography and fMRI in the temporal lobes. *Hum. Brain Mapp.* 31, 1570–1587
128. Binney, R.J. *et al.* (2010) The ventral and inferolateral aspects of the anterior temporal lobe are crucial in semantic memory: evidence from a novel direct comparison of distortion-corrected fMRI, rTMS, and semantic dementia. *Cereb. Cortex* 20, 2728–2738
129. Halai, A.D. *et al.* (2014) A comparison of dual gradient-echo and spin-echo fMRI of the inferior temporal lobe. *Hum. Brain Mapp.* 35, 4118–4128
130. Kundu, P. *et al.* (2017) Multi-echo fMRI: a review of applications in fMRI denoising and analysis of BOLD signals. *NeuroImage* 154, 59–80
131. Asyraff, A. *et al.* (2021) Stimulus-independent neural coding of event semantics: evidence from cross-sentence fMRI decoding. *NeuroImage* 236, 118073

132. Visser, M. *et al.* (2010) Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *J. Cogn. Neurosci.* 22, 1083–1094
133. Rogers, T.T. *et al.* (2021) Evidence for a deep, distributed and dynamic code for animacy in human ventral anterior temporal cortex. *eLife* 10, e66276
134. Kriegeskorte, N. *et al.* (2008) Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141
135. Cadieu, C.F. *et al.* (2014) Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Comput. Biol.* 10, e1003963
136. Sexton, N.J. and Love, B.C. (2022) Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Sci. Adv.* 8, eabm2219
137. Mehrer, J. *et al.* (2020) Individual differences among deep neural network models. *Nat. Commun.* 11, 1–12
138. Adlam, A.-L.R. *et al.* (2006) Semantic dementia and fluent primary progressive aphasia: two sides of the same coin? *Brain* 129, 3066–3080
139. Rogers, T.T. *et al.* (2015) Disorders of representation and control in semantic cognition: effects of familiarity, typicality, and specificity. *Neuropsychologia* 76, 220–239
140. Mahon, B.Z. *et al.* (2009) Category-specific organization in the human brain does not require visual experience. *Neuron* 63, 397–405
141. Mahon, B.Z. *et al.* (2010) The representation of tools in left parietal cortex is independent of visual experience. *Psychol. Sci.* 21, 764–771
142. Bedny, M. and Saxe, R. (2012) Insights into the origins of knowledge from the cognitive neuroscience of blindness. *Cogn. Neuropsychol.* 29, 56–84
143. Chen, L. and Rogers, T.T. (2015) A model of emergent category-specific activation in the posterior fusiform gyrus of sighted and congenitally blind populations. *J. Cogn. Neurosci.* 27, 1981–1999
144. Kanwisher, N. (2000) Domain specificity in face perception. *Nat. Neurosci.* 3, 759–763
145. Kanwisher, N. *et al.* (1997) The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J. Neurosci.* 17, 4302–4311
146. Behrmann, M. *et al.* (2016) Neural mechanisms of face perception, their emergence over development, and their breakdown. *WIREs Cogn. Sci.* 7, 247–263
147. Dundas, E.M. *et al.* (2013) The joint development of hemispheric lateralization for words and faces. *J. Exp. Psychol. Gen.* 142, 348–358