Title: Evidence for a deep, distributed and dynamic semantic code in human ventral anterior temporal cortex

Authors: *Timothy T. Rogers[1], Christopher Cox[2], Qihong Lu[3], Akihiro Shimotake[4], Takayuki Kikuch[5], Takeharu Kunieda[5,6], Susumu Miyamoto[5], Ryosuke Takahashi[4], Akio Ikeda[7], Riki Matsumoto[8], *Matthew A. Lambon Ralph[9]

Affiliations:

[1] Dept. of Psychology, University of Wisconsin, Brogden Hall, Madison, WI 53706

[2] Dept. of Psychology, Louisiana State University, Audubon Hall, Baton Rouge, LA 70802

[3] Dept. of Psychology, Princeton University, South Dr, Princeton, NJ 08540

[4] Dept. of Neurology. Kyoto University Graduate School of Medicine, Kawaharacho, Shogoin,, Sakyo-ku, Kyoto, 606-8507, Japan

[5] Dept. of Neurosurgery, Kyoto University Graduate School of Medicine, Kyoto, Japan

[6] Dept. of Neurosurgery, Ehime University Graduate School of Medicine, Shizukawa Toon city, Ehime, 791-0295, Japan

[7] Dept. of Epilepsy, Movement Disorders and Physiology, Kyoto University Graduate School of Medicine, Kyoto, Japan

[8] Div. of Neurology, Kobe University Graduate School of Medicine, Kusunoki-cho, Chuo-ku, Kobe, 650-0017, Japan

[9] MRC Cognition and Brain Sciences Unit, 15 Chaucer Rd., Cambridge, UK, CB2-7EF

Correspondence to: Tim Rogers (ttrogers@wisc.edu; lead contact) or Matthew Lambon Ralph (Matt.Lambon-Ralph@mrc-cbu.cam.ac.uk)

ORCID: 0000-0001-6304-755X

Keywords: Semantic memory, ECoG, cognition, neural networks, temporal lobe, multivariate pattern classification, dynamic models, temporal generalization

Author Contributions

TTR conducted all simulations, contributed to ECoG analysis, and wrote the initial manuscript draft. CRC and QL contributed to ECoG analysis and manuscript revision. AS, TKi, TKu, SM, RT, AT and RM contributed to ECoG study design, and collected and prepared all ECoG data. MLR contributed to all aspects of the study design, analysis of model and ECoG data, and manuscript preparation.

Abstract

How does the human brain encode semantic information about objects? This paper reconciles two seemingly contradictory views. The first proposes that local neural populations independently encode semantic features; the second, that semantic representations arise as a dynamic distributed code that changes radically with stimulus processing. Combining simulations with a well-known neural network model of semantic memory, multivariate pattern classification, and human electrocorticography, we find that both views are partially correct: semantic information is distributed across ventral temporal cortex in a dynamic code that possesses stable feature-like elements in posterior regions but with elements that change rapidly and nonlinearly in anterior regions. This pattern is consistent with the view that anterior temporal lobes serve as a deep cross-modal "hub" in an interactive semantic network, and more generally suggests that tertiary association cortices may adopt dynamic distributed codes difficult to detect with common brain imaging methods.

Key words: semantic memory, ECoG, neural networks, MVPA, temporal generalization, dynamics, representation, temporal lobe

Introduction

Semantic memory supports the remarkable human ability to recognize new items and events, infer their unobserved properties, and comprehend and produce statements about them (1, 2). These abilities arise from neural activity propagating in a broadly distributed cortical network, with different components encoding different varieties of information (perceptual, motor, linguistic, etc (3–5). The ventral anterior temporal lobes (vATL) form a hub in this network that coordinates activation amongst the various surface representations(5, 6). In so doing the

vATL acquires distributed representations that allow the whole network to express conceptual similarity structure, supporting inductive generalization of acquired knowledge across conceptually-related items(7).

Within this framework, contemporary theories differ in their proposals about the semantic representation of visually-presented objects. Many researchers view such representations as arising from the propagation of activity upward through a feature hierarchy, beginning with simple perceptual features, moving through intermediate modality-specific features, and ending with abstract cross-modal features that express semantic category structure. This perspective has strongly influenced computational models of object recognition developed in visual neuroscience(8, 9) and associated brain imaging work(10, 11). Alternatively, semantic representations may arise from interactive settling processes within distributed and dynamic brain networks.(5, 12) On this view, semantic representations are activation patterns encoded jointly over many neural populations, with similar concepts evoking similar patterns. The information encoded by a given population depends on the states of other populations, so that the whole representation is best viewed as a point in a multidimensional state space.(2) Stimulus processing involves a transitioning of the system through the space, rather than the activation of increasingly abstract/complex feature detectors, as the whole system settles toward an interpretation of the input(13). Many models of semantic memory adopt this perspective(12, 14–17), which also aligns with the general view that the ventral visual processing stream is recurrent and interactive(18–20)

The two views carry critically different implications for understanding how neural systems represent semantic information. On featured-based views, local populations of neurons independently encode the presence of a particular feature, licensing a straightforward

interpretation of neural activity: when the population is active, the corresponding feature has been detected, inferred, or called to mind; when inactive, it hasn't. The temporal behavior of the population directly indicates the time-course with which the represented feature is available to influence processing, and the mean neural activity over time indicates the strength with which the feature was activated in a particular trial or task condition. These ideas motivate efforts to determine which cortical regions encode which features at which points in time by correlating/predicting the presence of a semantic feature (such as a category label or property) with/from local neural activity(3, 21). If, however, semantic representations are distributed, the local behavior of a single population may not be interpretable independent of the others;(22) and if semantic processing is also dynamic, the contribution of a local population to the distributed code may change in real time, potentially in highly nonlinear ways(23). These possibilities suggest that the treatment of neurophysiological signals as arising from independent feature detectors can mischaracterize how neural systems encode semantic information.

Figure 1 illustrates why. Suppose a participant names images of tools and mammals while the responses of two neural populations are measured with high temporal and spatial resolution. The top panels show hypothetical joint responses to each stimulus over time from onset to response (together with those of other untested living and manmade items), plotted as trajectories in a 2D space over time. The shaded panels show how a multivariate classifier partitions the space when trained to discriminate mammals from tools at one timepoint. The categories are always well differentiated, so the two populations jointly encode semantic

information at every timepoint, but because the trajectories are nonlinear each population's contribution to this structure changes over time—the category discrimination plane rotates.

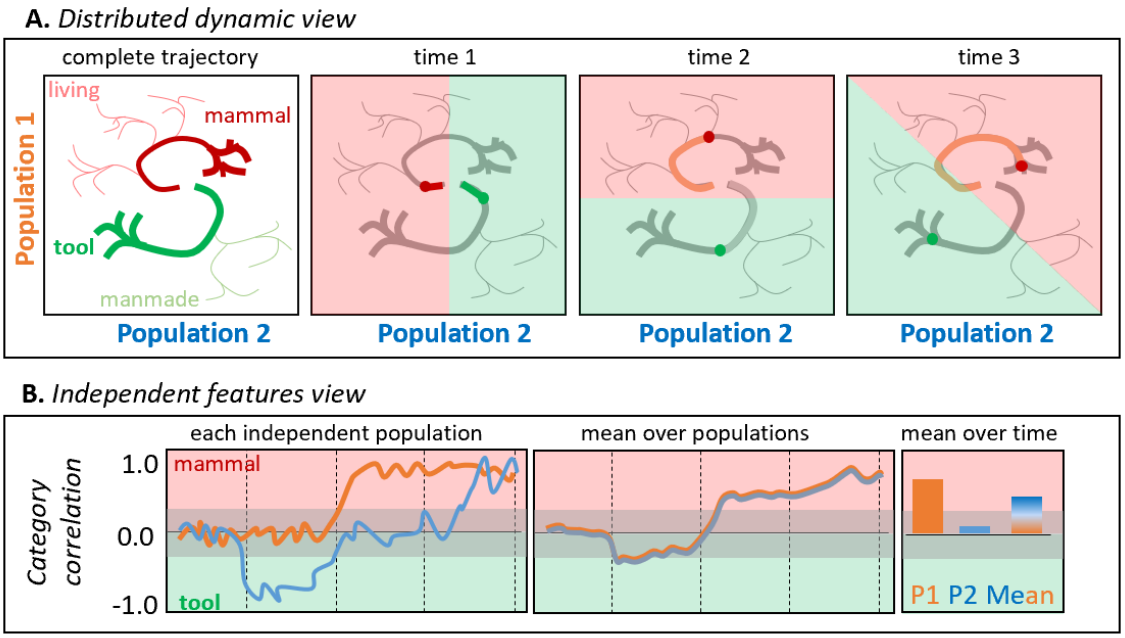

Figure 1. A. Hypothetical joint activations of two neural populations to living and manmade items (left), and the classification plane that would best discriminate tools from mammals at different timepoints. B. Independent correlations between each population's activity and a binary category label for the same trajectories plotted above, shown across time for each (left), averaged across the two populations (middle), and averaged over time for each population independently or for both populations (right).

The bottom panels show how each population's behavior would appear if its activity was analyzed independently via correlation with semantic information. The orange line shows how Population 1's activity (ie, the projection of the trajectories in panel A against the vertical axis only) correlates with the tool/mammal label at each timepoint—the correlation becomes positive about half-way through the time series The blue line shows how Population 2's activity(ie the Panel A trajectories projected onto the horizontal axis) correlates with the tool/mammal label: it quickly shows a negative correlation(tools more active than mammal), then no correlation, then a positive correlation (mammals more active than tools). From these independent analyses it might seem that (a) animals are detected only midway through the time-series, (b) tools are

111 detected earlier than animals and (c) there exist populations that "switch" from tool-detectors to
112 mammal-detectors. The bottom middle panel shows the mean response across the two
113 populations at each point in time, as might be observed if the neurophysiological measurement
114 lacks spatial resolution (e.g., EEG/MEG). The populations appear to detect animals late in the trial.
115 The right panel shows mean responses across time for each population (consistent with high
116 spatial and low temporal resolution; e.g. fMRI) and for both populations together (low spatial and
117 temporal resolution). The former suggests that population 2 plays no important role
118 distinguishing the categories, while the latter suggests the two populations together selectively
119 detect animals. These conclusions are incorrect from the distributed and dynamic perspective,
120 under which the two populations always jointly differentiate the categories but the contribution
121 of each changes over time (Figure 1A).

122 The interpretation of neurophysiological signals in the cortical semantic system thus
123 depends critically upon whether the neuro-semantic code is feature-based or distributed and
124 dynamic at whatever scale measurements are taken. Efforts to adjudicate the question face three
125 significant hurdles. First, spatial and/or temporal averaging can obscure signal if the code truly is
126 distributed and dynamic—discovery requires neural data with high temporal and spatial
127 resolution, ruling out the non-invasive methodologies that constitute the majority of work in this
128 area(24). Second, independent univariate analysis can mischaracterize information distributed
129 across multiple channels—discovery requires multivariate methods(25). Third, studies connecting
130 neurophysiologal measurements to computational models of visual recognition have primarily
131 focused on feed-forward models that do not exhibit dynamic processing(10, 26). While some
132 recent decoding evidence suggests visual processing in ventral temporal cortex is dynamic(20), no
133 prior work has assessed how representations evolve within distributed and dynamic semantic

134 models, and consequently it is unclear what this view predicts or how it differs from feature-based
135 approaches.

136 We therefore combined computational modeling, multivariate pattern classification, and
137 electrocorticography (ECoG) to assess whether semantic representations in human vATL are
138 distributed and dynamic. We focused on vATL due to its central role in semantic representation(5,
139 6) and because it is a common target for grid implantation in patients awaiting surgery to
140 remediate epilepsy. We first showed that the nonlinear dynamics highlighted in our thought
141 experiment arise in a well-studied neurocomputational model of semantic memory(7, 12), then
142 used temporal generalization of pattern classifiers(27) to establish the predicted "signature" of a
143 dynamically changing semantic code under this model. We next applied this technique to
144 intracranial local field potentials collected from the surface of the left vATL while participants
145 named line drawings of common items, and found the critical decoding signature—providing
146 strong evidence that, at the level of spatiotemporal resolution captured by ECoG, semantic
147 structure in vATL is expressed by a distributed code that changes dynamically with stimulus
148 processing. Finally we considered how the neural code exploited by classifiers changed over time
149 and across the ventral temporal surface, with results that reconcile the two divergent views of
150 semantic representation in the brain .

151 **Results**

152 *Simulation study***.** Simulations provide a formal basis for understanding the implications of the
153 distributed and dynamic view, because a computer model's architecture, behavior, learning and
154 testing patterns are fully known. We therefore adapted the well-known "hub and spokes" model

155 of semantic memory (5, 12, 28) to assess whether semantic representations change dynamically
156 with stimulus processing and how the temporal generalization method can uncover such a code.

157 The model was a fully continuous and recurrent neural network that learns cross-modal
158 associations between distributed visual and verbal representations via three reciprocally-
159 connected hidden layers (Figure 2A). Given positive input to a subset of visual or verbal units, it
160 learns to activate the corresponding item's unspecified visual and/or verbal attributes. Activity
161 propagates in both directions, from surface representations to hub and back, so that the trained
162 model settles to an attractor state representing any item specified by the input. We trained the
163 model to complete patterns representing ninety items from three model conceptual domains
164 (e.g., animals, objects, and plants), each organized into three categories containing ten items (see
165 Methods for details). Items from different categories in the same domain shared a few properties
166 while those in the same category shared many. We then presented the trained model with visual
167 input for each item and recorded the resulting activation patterns over hub units for each tick of
168 simulated time as the network settled.

169 To visualize how model internal representations changed during stimulus processing, we
170 computed a 3D multi-dimensional scaling (MDS) of activation patterns at all timepoints during
171 "perception" of each stimulus, then plotted the changing representation for each item as a line in
172 this space. The result in Figure 2B shows a systematic but nonlinear elaboration of the conceptual
173 structure latent in the stimulus patterns: the domains separate from one another early on, but
174 each item follows a curved trajectory over time. The curves are not an artifact of data-reduction—
175 Figure 2C shows these trajectories in the native activations of randomly-sampled unit pairs in one
176 network run, where they are clearly nonlinear (and see SI-1 for comparative results with linearly-
177 changing representations). Consequently, independent correlation analysis of each unit's

178 behavior produces mixed results (Figure 2D), with some behaving like tonic category detectors

179 (green squares), some like transient detectors (blue), some appearing to flip their category

180 preference (red) and others appearing not to code category information at all (gray).

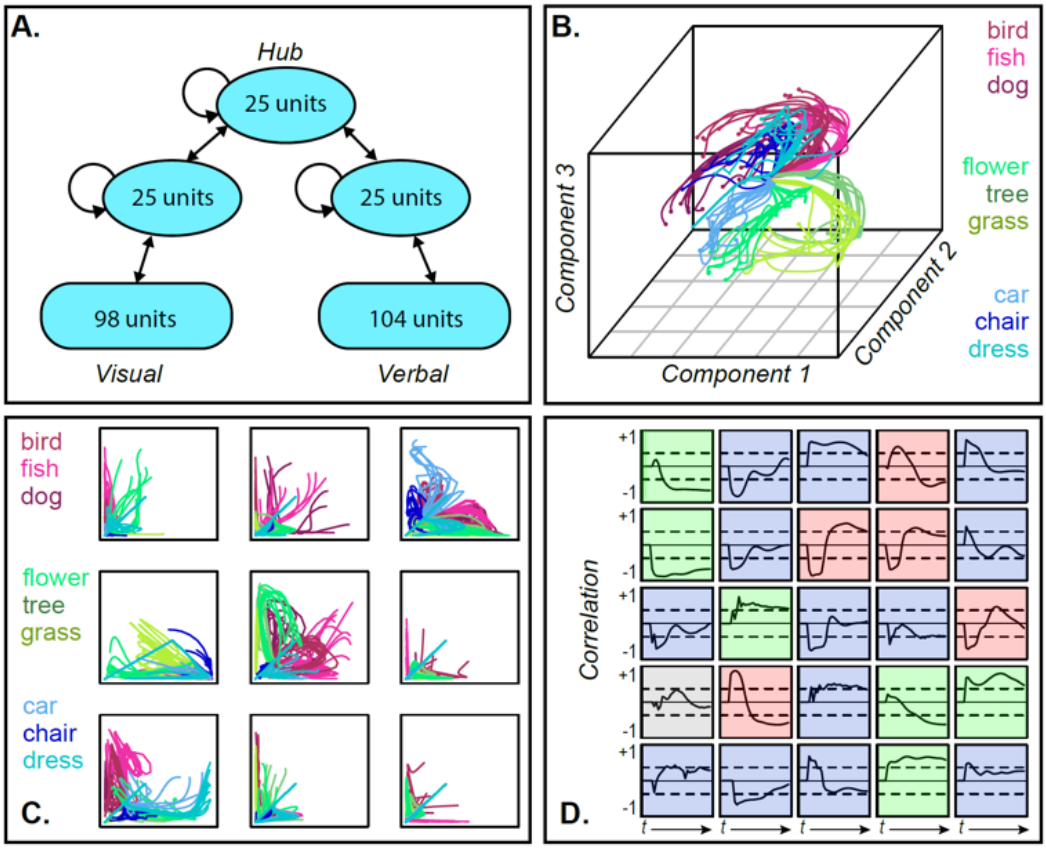

Figure 2. A. Model architecture. B. 3D MDS of hub activation patterns learned in one model run—each line shows the trajectory of a single item over time in the compressed space. C. The same trajectories shown in uncompressed unit activations for 9 randomly sampled unit pairs, horizontal and vertical axes each showing activation of one unit. D. Feature-based analysis of each hub unit in one network run. Each square shows one unit. Lines trace, across time, the correlation between unit activation and category labels across items category labels with dashed lines showing significance thresholds. Color indicates different patterns of responding (see text).

181

182 Thus the full distributed pattern across hub units elaborates conceptual structure from

183 early in processing, but the progression is nonlinear and only clearly discernable in a low-

184 dimensional embedding of the space. Such an embedding can be computed for the model

185 because we know which units are important and can apply the MDS to all and only those units.

186 The same approach cannot be applied to ECoG data for two reasons. First, one cannot know a-
187 priori which channels record signals relevant for semantic representation and thus cannot simply
188 compute a low-dimensional embedding of all data collected. Instead one must fit a statistical
189 model that will selectively weight signals useful for discerning semantic structure. Second,
190 whereas the model allows access to the entire network, a cortical surface sensor array only
191 sparsely samples the neural responses contributing to a semantic representation. The problem
192 thus requires a multivariate statistical model capable of revealing a dynamically changing neural
193 code when fitted to sparsely-sampled neural data.

194 We therefore used multivariate logistic classifiers to decode semantic category
195 information from hub activation patterns and assessed their behavior on simulated ECoG data.
196 For each simulated participant we selected a sparse random subsample (15%) of all hub units—
197 analogous to the sparse sampling of neural responses provided by a cortical sensor array—and
198 recorded their responses to each stimulus at every tick of time. We fitted a separate classifier at
199 each timepoint to distinguish two semantic domains from the activation patterns elicited over the
200 subsampled units. Figure 3A shows the cross-validated accuracy at each timepoint averaged over

many network runs and subsamples. The classifiers performed well above chance as soon as input activation reached the hub units and throughout the time window.

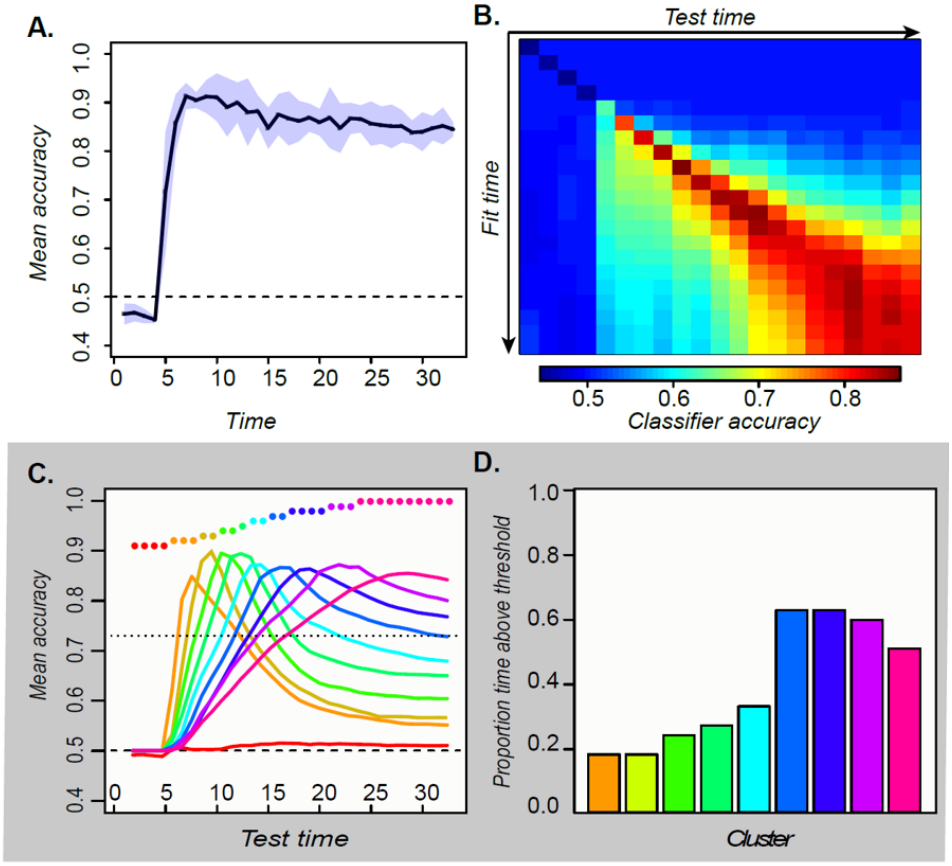

Figure 3. A. Mean and 95% confidence interval of the hold-out accuracy for classifiers trained at each tick of time in the model. B. Accuracy for each classifier (rows) tested at each point in time (columns). C. Mean accuracy for each cluster of classifiers at every point in time. Colored dots show the timepoints grouped together in each cluster. D. Proportion of the full time-window for which mean classifier accuracy in each cluster was reliably above chance.

To assess representational change over time we next adopted a temporal generalization approach(27), using the classifier fitted at one timepoint to decode the patterns observed at each other timepoint. Accuracy should remain high if the information a classifier exploits at the training time persists at other timepoints. The temporal generalization profile of each classifier thus

209 indicates how the underlying neural code persists or changes over time. Classifiers fitted to earlier
210 activation patterns generalized only to immediate temporal neighbors, while those fitted to later
211 patterns generalized over a wider window but failed at decoding earlier states (Figure 3B). To
212 better visualize these results, we clustered the rows of the matrix in Figure 3B and plotted the
213 mean accuracy of the classifiers in each cluster across time (Figure 3C). The results exhibit an
214 "overlapping waves" pattern: classifiers that work on early patterns quickly fail at later timepoints
215 where a different classifier succeeds. As time progresses, the clusters include more classifiers and
216 the breadth of time over which the classifiers perform well widens (Figure 3D).

217 This pattern reflects the nonlinear trajectories apparent in the sparsely-sampled
218 representational space as the network settles to an attractor state. When trajectories curve,
219 earlier classification planes fail later in time while later planes fail at earlier time-points. If
220 representations simply moved linearly from initial to final state, early classifiers would continue
221 to perform well throughout processing—a pattern observed in simulations with feature-based
222 models, models with distributed representations that evolve linearly, and recurrent but shallow
223 neural networks (see SI-1). In the deep network, the non-linear dynamic pattern was observed
224 only in the hub layer—in more superficial layers, the code remained stable (SI-2). The simulations
225 thus suggest that distributed and dynamic semantic representations can arise in deep layers of

226 interactive networks and will elicit a particular "signature" when multivariate pattern classifiers
227 are used to decode semantic structure from ECoG data. Specifically, such classifiers will show:

228 1. Constant decodability. Neural activity predicts stimulus category at every time point once
229 activation reaches the vATL.
230 2. Local temporal generalization. Classifiers generalize best to immediate temporal neighbors
231 and worst to distal timepoints.
232 3. Widening generalization window. The temporal window over which classifiers generalize
233 grows wider over time.
234 4. Change in code direction. Elevated neural activity can signify different semantic information
235 at different points in time.

236 These are the characteristics we looked for in the ECoG study.

237 *ECoG study.* The dataset included local field potentials (LFPs) collected at 1000Hz from 16-24
238 electrodes situated in the left ventral anterior temporal cortex of 8 patients awaiting surgery while
239 they named line-drawings of common animate and inanimate items matched for a range of
240 confounds (see Methods). We analyzed LFPs over the 1640ms following stimulus onset, a window
241 that allows us to assess decodeability both before and after the mean onset of naming (1190ms).
242 LFPs were decoded using a 50ms sliding-window, with separate classifiers fitted at each window
243 and the window advancing in 10ms increments. This yielded 160 classifiers per subject, each
244 decoding LFPs across all vATL electrodes in a 50ms window. Each classifier was then tested on all

160 time-windows. The classifiers were logistic regression models fitted with L1 regularization to encourage coefficients of 0 for many features (see Methods).

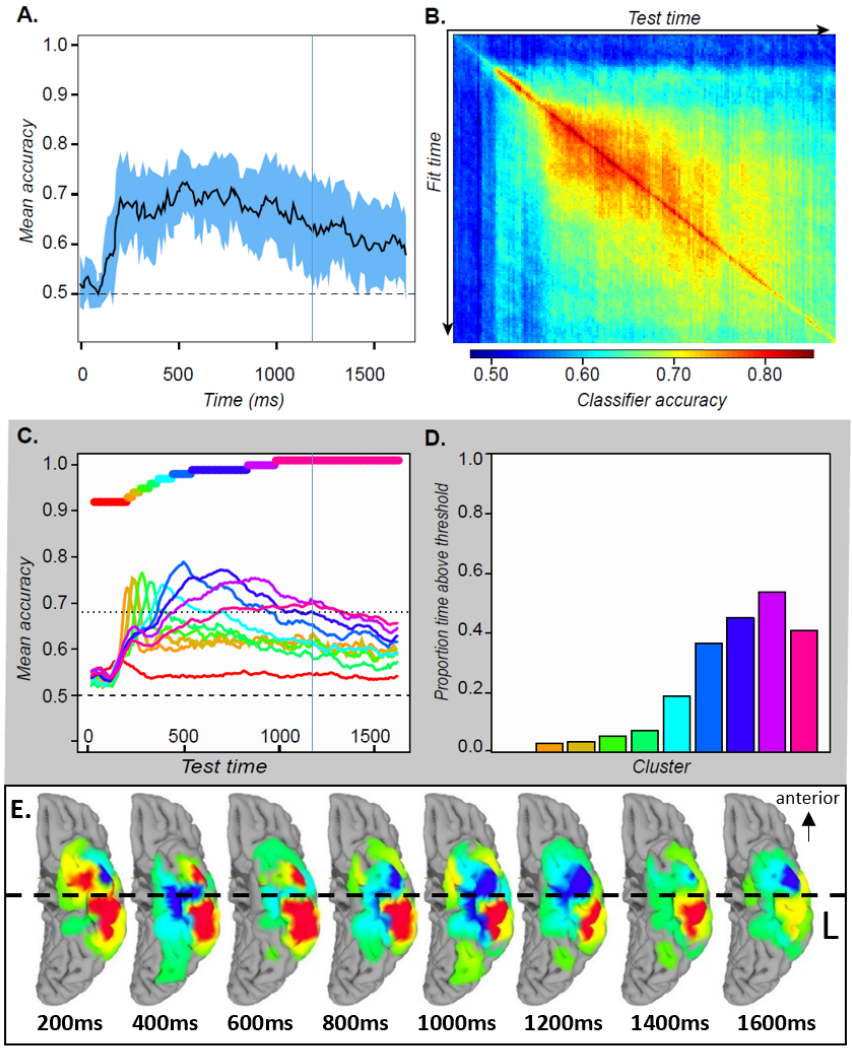

Figure 4. A. Mean and 95% confidence interval of the hold-out accuracy for classifiers trained at each 50ms time window of ECoG data. B. Mean accuracy across participants for each classifier (rows) tested at each timepoint (columns) in the ECoG data. C. Mean accuracy for each cluster of classifiers at every point in time. Colored bars show the timepoints grouped together in each cluster. D. Proportion of the full time-window for which mean classifier accuracy in each cluster was reliably above chance. E. Mean classifier coefficients across participants plotted on a cortical surface at regular intervals over the 1640ms window. Warm vs cool colors indicate positive versus negative mean coefficients, respectively. In A and C, vertical line indicates mean onset of naming.

Hold-out accuracy exceeded chance at about 200ms post stimulus onset, well before name initiation, and remained statistically reliable throughout the time window (Figure 4A) with

250 no obvious change at the mean time of naming onset. At 200ms classifiers generalized well to
251 timepoints near the training window but poorly to more distal timepoints, with the generalization
252 envelope widening as time progressed (4B). We again clustered the classifiers based on their
253 temporal accuracy profile, then plotted mean profiles for each cluster (4C). The result was an
254 "overlapping waves" pattern strikingly similar to the simulation: classifiers that performed well
255 early in processing quickly declined in accuracy, replaced by a different well-performing set. Over
256 time neighboring classifiers began to show similar temporal profiles, forming larger clusters that
257 performed above chance for a broader temporal window (4D).

258 Finally, we considered whether and how the neuro-semantic code changed over time. For
259 each time window we projected the classifier weights for all electrodes in all subjects to a cortical
260 surface model, then animated the results (see movie S1). Figure 4E shows snapshots every 200ms
261 post stimulus onset. In mid-posterior regions the code was spatially and temporally stable—
262 weights on the lateral aspect were positive while those on the medial aspect were negative,
263 consistent with feature-based views of representation. The anterior pattern differed, flipping
264 from mainly positive at 200ms to mainly negative by 800ms and fluctuating across time and space
265 throughout. In other words, the "meaning" of a positive deflection in the LFP—whether it signaled
266 animal or non-animal—stayed constant posteriorly but changed direction over time anteriorly,
267 consistent with the deep, distributed, dynamic view (see SI-3).

268 **Discussion**

269 We have combined computational modeling, multivariate pattern classification, and
270 human ECoG to better understand how ventral temporal cortex encodes semantic information
271 about visually-presented stimuli. The results suggest that, at the scale of intracranial LFPs, neural
272 signals throughout ventral temporal lobe jointly express semantic information from about 200ms

post stimulus-onset, via a distributed code that is stable and feature-like in middle/posterior regions but dynamic and nonlinearly-changing in anterior regions. In simulation we showed that similar phenomena arise in a deep, interactive neuro-semantic model, producing a characteristic decoding signature: classifiers perform well in the time-window when they were trained, but generalize over a narrow time envelope that widens as the system settles. This pattern was only observed in a model combining distributed representation, interactive processing, and a deep architecture (see SI). Remarkably similar phenomena were observed in ECoG data collected from the surface of ventral temporal lobe while participants named line-drawings, supporting the proposal that vATL encodes distributed semantic representations that change dynamically and nonlinearly with stimulus processing by virtue of this region's role as a deep hub within the cortical semantic network.

This proposal resolves a long-standing puzzle. Convergent methods have established the centrality of vATL for semantic memory, including studies of semantic impairment(29–31), lesion-symptom mapping(32), functional(33, 34) and structural37(35, 36) brain imaging, and transcranial magnetic stimulation(37, 38). Yet multivariate approaches to discovering neuro-semantic representations rarely identify the vATL, instead revealing semantic structure more posteriorly(39–41). One prominent study suggested that semantic representations may tile the entire cortex except for the vATL(4). Setting aside significant technical challenges of successful neuroimaging of this region36, almost all such studies have employed non-invasive imaging techniques that sacrifice either temporal or spatial resolution—a compromise that will destroy signal in vATL if semantic representations there are distributed and dynamic at the scale we have measured, but will preserve signal in posterior regions where the code is more stable. Thus the

295 widespread null result may arise precisely because semantic representations in vATL are
296 distributed and dynamic.

297 Consistent with this view, the current results in middle/posterior regions accord well with
298 evidence from classical functional brain imaging of visual semantics. Several fMRI studies have
299 reported category effects in the mid-posterior fusiform, with animate items eliciting greater
300 activation on the lateral bank and inanimate items eliciting greater activation on the medial
301 bank(42–44). In our analysis, these are regions where a stable, feature-like code arises, with
302 animate items signaled by positive LFPs on the lateral aspect and negative LFPs on the medial
303 aspect. Since the location and direction of the code are stable in these regions within and across
304 subjects, the signal remains detectable even with the spatial and temporal averaging that occurs
305 in univariate fMRI.

306 These conclusions rest partly on the analysis of classifier weights. When multiple channels
307 are affected by a common source of noise, multivariate decoders can place weights on
308 uninformative channels that "subtract out" the noise corrupting the signal on informative
309 channels(45). Might the fluctuating weights observed in vATL arise, not because the region
310 encodes semantic structure, but because it serves to "cancel out" effects of correlated noise on
311 the signal-carrying channels more posteriorly? Four lines of evidence suggest not. First, we fit
312 classifiers to only the anterior electrodes in each participant and observed reliable decoding (SI-
313 4). Second, an earlier study applied searchlight representational similarity analysis (RSA) to the
314 same data(46), and found that vATL was the only local region that reliably encodes semantic
315 similarity structure. Neither result could obtain if vATL simply subtracted out correlated noise
316 from more posterior areas. Third, the observed pattern of a stable code posteriorly and
317 fluctuating code anteriorly was predicted by the current simulations, using an model validated

318 against neuropsychological, anatomical, and brain-imaging results in prior work(5). Fourth, the
319 critical importance of ATL for semantic representation has been established by the broad range
320 of converging evidence cited previously.

321 Prior studies applying temporal generalization to MEG data in visual semantic tasks
322 uniformly report a very narrow and unchanging band of temporal generalization(47, 48), a pattern
323 consistent with the proposal that the neuro-semantic code changes rapidly over the course of
324 stimulus processing. Our results differ from the MEG pattern, and indeed from most other work
325 applying the temporal generalization approach(27), in showing a gradual widening of the
326 temporal generalization window. This phenomenon does not arise from the autocorrelational
327 structure of the data itself—the window of time over which an electrode reliably predicts its own
328 future state does not grow wider with stimulus processing (SI-5). Instead the widening must
329 reflect an increasingly stable semantic code. The simulation explains why the pattern arises in
330 anterior temporal cortex: hub representations in vATL change rapidly early on due to interactions
331 with modality-specific representations throughout cortex, but these changes slow as the full
332 activation pattern emerges across network components.

333 While ECoG provides simultaneously high spatial and temporal resolution compared to
334 non-invasive brain imaging, its measurements are nevertheless influenced by activity in hundreds
335 of neurons. For this reason we have emphasized that the anterior neuro-semantic code appears
336 distributed and dynamic at the scale of local neural population activity, as measured by LFPs. Our
337 conclusions do not shed light on the code employed by single neurons, or other fine-grained units
338 such as cortical columns, though the temporal-generalization approach we have deployed may
339 extend well to data collected at this scale.

Finally, why should a dynamic distributed code arise specifically within the vATL? The area is situated at the top of the ventral visual stream, but also connects directly to core language areas(49) and, via middle temporal gyrus, to parietal areas involved in object-directed action(12). It receives direct input from smell and taste cortices (50) , and is intimately connected with limbic structures involved in emotion, memory, and social cognition(51). Thus vATL anatomically forms the hub of a cross-modal network ideal for encoding associations among visual, linguistic, action, sensory, and social/motivational representations. Hub neurons interact with a wide variety of subsystems, each encoding a different kind of structure and content, potentially pushing the hub representations in different directions over time as activity propagates in the network. Other network components lying closer to the sensory or motor periphery connect mainly within individual modality-specific systems(35), so may be less impacted by such cross-modal interactions, as observed in the model. For this reason, the feature-based approach that has proven indispensable for characterizing neural representations in modality-specific cortices may be less suited to understanding the distributed and dynamic representations that arise in deeper and more broadly-connected (tertiary association) cortical regions. These include regions critical for human semantic knowledge, and potentially other higher-level cognitive functions.

Materials and Methods

Simulation study.

The model implements the “distributed-plus-hub” theory of semantic representation developed in prior work(5, 6, 12, 52, 53). It is a deep, fully continuous and recurrent neural network48 that learns associations among visual representations of objects, their names, and verbal descriptors, via a central cross-modal hub, with units and connectivity shown in Figure 2A of the main paper. Simulations were conducted using the open-source Light Efficient Network

Simulator updated for contemporary libraries (https://github.com/crcox/lens). Code for replicating the simulations and the data reported in this paper are available at https://github.com/ttrogers.

All units employed a continuous-time sigmoidal activation function with a time-constant of 0.25. Visual and Verbal units were given a fixed, untrainable bias of -3 that produced a low activation state without positive input. Hidden units had trainable biases. To simulate perception of an image, units encoding the item's visual representation were given direct positive stimulation of +6 so that, combined with the fixed bias, they received a net input of +3 (in addition to any inputs from other units in the model). The resulting changes in unit activations propagated through visual hidden, hub, and verbal hidden units to eventually alter activation states in the verbal units. Because the model was reciprocally connected, such downstream changes fed back to influence upstream states at each moment of simulated time, as the whole system settled to a stable state. To simulate verbal comprehension, the same process unfolded, but with positive input externally provided to verbal units. Units updated their activation states asynchronously in permuted order on each tick of time and were permitted to settle for 5 time intervals (a total of 20 updates) during training and 8 time intervals (32 updates) during testing.

Model environment. The model environment contained visual and verbal patterns for each of 90 simulated objects, conceived as belonging to 3 distinct domains (e.g. animals, objects, and plants). Each domain contained 10 items from each of 3 sub-categories—thus there were 30 "animals," 30 "objects" and 30 "plants." Visual patterns were constructed to represent each item by randomly flipping the bits of a binary category prototype vector in which items from the same domain shared a few properties and items from the same category shared many. The verbal patterns were constructed by giving each item a superordinate label true of all items within a

given domain (animal, object, plant), a basic-level label true of all items within a category (e.g. “bird”, “fish”, “flower”, etc), and a subordinate label unique to the item (e.g. “robin”, “salmon”, “daisy”, etc). These procedures, adopted from prior work(52), generated model input/target vectors that approximate the hierarchical relations among natural concepts in a simplified manner that permits clear understanding and control of the relevant structure.

Training. For each input, target patterns that fully specified the item’s visual and verbal characteristics were applied throughout the duration of stimulus processing. The model was trained with backpropagation to minimize squared error loss. Half of the training patterns involved generating verbal outputs from visual inputs, while the other half involved generating visual outputs from verbal inputs. The model was initialized with small random weights sampled from a uniform distribution ranging from -1 to 1, then trained for 30,000 epochs in full batch mode with a learning rate of 0.002 and without weight decay. For each pattern, the settling process was halted after 20 activation updates, or when all Visual and Verbal units were within 0.2 of their target values, whichever came first. For all reported simulations the model was trained 5 times with different random weight initializations. After training, all models generated correct output activations (ie, on the correct side of the unit midpoint) for more than 99% of output units across all training runs. Each model was analyzed independently, and the final results were then averaged across the five runs.

Testing. The picture-naming study was simulated by presenting visual input for each item, recording the resulting activations across the 25 hub units at each update as the model settled over 32 updates, and distorting these with uniform noise sampled from -0.005 to 0.005 to simulate measurement error. As the model settles over time it gradually activates the word unit corresponding to the item name, and in this sense simulates picture naming. Just as in the ECoG

409 data, we recorded unit activations across a fixed period of time, regardless of when the correct
410 name unit became active.

411 Note that, whereas the ECoG study employed items drawn from two general semantic
412 domains (living and nonliving), the model was trained on 3 domains. This provided a simple model
413 analog to the true state of affairs in which people know about more semantic kinds than just those
414 appearing in the ECoG stimulus set. To simulate the study, the model was presented with 60 items
415 selected equally from 2 of the 3 semantic domains—so as in the study, half the stimuli belonged
416 to one domain and half to another. To ensure results did not reflect idiosyncrasies of one domain,
417 we simulated the task with each pair of domains and averaged results across these.

418 *Analysis*. All analyses were conducted using R version 3.6. To visualize the trajectory of
419 hub representations through unit activation space, we computed a simultaneous 3-component
420 multidimensional scaling of the unit activation patterns for all 90 items at all 33 timepoints using
421 the native R function cmdscale. The resulting coordinates for a given item at each point in time
422 over the course of settling were plotted as lines in a 3D space using the scatterplot3d package in
423 R. Figure 2B shows the result for one network training run. Figure 2C shows the same trajectories
424 in the raw data (ie actual unit activation states rather than latent dimensions in a MDS) for
425 randomly-sampled pairs of hub units.

426 To simulate decoding of ECog data, we evaluated logistic classifiers in their ability to
427 discriminate superordinate semantic category from patterns of activity arising in the hub at each
428 timepoint. As explained in the main text, we assume that ECoG measures only a small proportion
429 of all the neural populations that encode semantic information. We therefore sub-sampled the
430 hub-unit activation patterns by selecting 3 units at random from the 25 hub units and using their
431 activations to provide input to the decoder. Classifiers were fitted using the glm function and the

binomial family in R. A separate decoder was fitted at each time-point, and unit activations were mean-centered independently at each time point prior to fitting. We assessed decoder accuracy at the time-point where it was fitted using leave-one-out cross-validation, and also assessed each decoder at every other time point by using it to predict the most likely stimulus category given the activation pattern at that time point and comparing the prediction to the true label. This process was repeated 10 times for each model with a different random sample of 3 hub units on each iteration. The reported results then show mean decoding accuracy averaged over the 5 independent network training runs, for decoders trained and tested at all 33 time points. The above procedure yielded the decoding accuracy matrix shown as a heat plot in Figure 3B.

Each row of this matrix shows the mean accuracy of decoders trained at a given timepoint, when those decoders are used to predict item domain at each possible timepoint. The diagonal shows hold-out accuracy for decoders at the same time point when they are trained, but off-diagonal elements show how the decoders fare for earlier (below diagonal) or later (above) timepoints. Decoders that perform similarly over time likely exploit similar information in the underlying representation, and so can be grouped together and their accuracy profiles averaged to provide a clearer sense of when the decoders are performing well. To this end, we clustered the rows of the decoding accuracy matrix by computing the pairwise cosine distance between these and subjecting the resulting similarities to a hierarchical clustering algorithm using the native hclust function in R with complete agglomeration. We cut the resulting tree to create 10 clusters, then averaged the corresponding rows of the decoding accuracy matrix to create a temporal decoding profile for each cluster (lines in Figure 3C). We selected 10 clusters because this was the highest number in which each cluster beyond the first yielded a mean classification

454 accuracy higher than the others at some point in time. Similar results were obtained for all cluster-
455 sizes examined, however.

456 Finally, to understand the time-window over which each cluster of decoders performs
457 reliably better than chance, we computed a significance threshold using a one-tailed binomial
458 probability distribution with Bonferroni correction. Each decoder discriminates two categories
459 from 60 items, with probability 0.5 of membership in either category. We therefore adopted a
460 significance threshold of 44 correct items out of 60, corresponding to a binomial probability of p
461 < 0.03 with Bonferroni correction for 330 tests (10 clusters at each of 33 time points). The barplot
462 in Figure 3D shows the proportion of the full time window during which each decoding cluster
463 showed accuracy above this threshold.

464 *ECoG study.*

465 *Participants.* Eight patients with intractable partial epilepsy (seven) or brain tumor (one)
466 originating in the left hemisphere participated in this study. These include all left-hemisphere
467 cases described in a previous study50, and we will use the same case numbers reported in that
468 work (specifically cases 1-5, 7, and 9-10). Background clinical information about each patient is
469 summarized in Table S1. Subdural electrode implantation was performed in the left hemisphere
470 for presurgical evaluation (mean 83 electrodes, range 56-107 electrodes/patient). 16-24
471 electrodes (mean 20 electrodes) covered the ventral ATL in each patient. The subdural electrodes
472 were constructed of platinum with an inter-electrode distance of 1 cm and recording diameter of
473 2.3 mm (ADTECH, WI). ECoG recording with subdural electrodes revealed that all epilepsy patients
474 had seizure onset zone outside the anterior fusiform region, except one patient for whom it was
475 not possible to localize the core seizure onset region. The study was approved by the ethics

476 committee of the Kyoto University Graduate School of Medicine (No. C533). Participants all gave

477 written information consent to participate in the study.

478

	Patient 1	Patient 2	Patient 3	Patient 4
Age, gender, handedness	22M R	29M R&L	17F R	38F R
WAIS-R (VIQ,PIQ,TIQ)	70, 78, 69	72, 78, 72	67, 76, 69	84,97,89
WMS-R (Verb, Vis , Gen, Attn, Del recall)	99, 64, 87, 91, 82	99, 92, 97, 87, 83	51, <50, <50, 81, 56	75,111,83,62,53
WAB	95.6	96	97.2	98.5
WADA test (Language)	Left	Bilateral	Left	Left
Age of seizure onset	16	10	12	29
Seizure type	non-specific aura → CPS, GTCS	aura (metamorphosia, ERS) → CPS	discomfort in throat → CPS	ERS → CPS
Ictal ECoG onset	aMTG	PHG	PHG	PHG
MRI	L basal frontal cortical dysplasia L anterior temporal arachnoid cyst	L posterior temporal cortical atrophy	L temporal tip arachnoid cyst	L hippocampal atrophy/sclerosis
Pathology	FCD type IA	FCD type IA Hippocampal sclerosis*	FCD type IB	Hippocampal sclerosis**

	Patient 5	Patient 7	Patient 9	Patient 10
Age, gender, handedness	55M R	41F R	51M R	38F R
WAIS-R (VIQ,PIQ,TIQ)	105,99,103	72, 83, 75	73, 97, 83	109, 115,112
WMS-R (Verb, Vis, Gen, Attn, Del recall)	71,117,84,109,72	83,111,89,94,82	80,101,85,1919,91	71,79,70,90,58
WAB	98	97.3	89.6	96.9
WADA test (Language)	Left	Right	Left	Left
Age of seizure onset	55	19	43	28
Seizure type	CPS (once)	aura (nausea,feeling pale) → CPS	CPS	non-specific aura → CPS
Ictal ECoG onset	none	PHG	mITG	SMG
MRI	Low-grade glioma L medial temporal lobe	L hippocampal atrophy/screrosis L parieto-occipital perinatal infarction	Left temporal cavernoma	L parietal opercurum tumor
Pathology	Diffuse astrocytoma	FCD IA Hippocampal sclerosis*	arteriovenous malformation	Oligoastrocytoma

Table 1. Patient characteristics. CPS: complex partial seizure; GTCS: generalized tonic clonic seizure; ECoG: electrocorticogram; ERS: epigastric rising sensation; a/pMTG: anterior/posterior part of the middle temporal gyrus; a/pMTG: anterior/posterior part of the middle temporal gyrus; FCD: focal cortical dysplasia; * dual pathology ** diagnosed by clinical findings

479 *Stimuli and Procedure*. One hundred line drawings (50 living and 50 nonliving items) were
480 obtained from previous norming studies(51,52). A complete list of all items can be found in (50).
481 Living and nonliving stimuli were matched on age of acquisition, visual complexity, familiarity and
482 word frequency, and had high name agreement. Independent-sample t-tests did not reveal any
483 significant differences between living and nonliving items for any of these variables.

484 Participants were presented with stimuli on a PC screen and asked to name each item as
485 quickly and accurately as possible. All stimuli were presented once in a random order in each
486 session and repeated over four sessions in the entire experiment. The responses of participants
487 were monitored by video recording. Each trial was time-locked to the picture onset using in-house
488 MATLAB scripts (version 2010a, Mathworks, Natick, MA). Stimuli were presented for 5 seconds
489 each and each session lasted 8 minutes 20 seconds. Participants' mean naming time was 1190ms.
490 Responses and eye fixation were monitored by video recording.

491 *Data preprocessing.* Data preprocessing was performed in MATLAB. Raw data were
492 recorded at sampling rate of 1000 Hz for six patients and at 2000Hz for two patients. The higher
493 sampling rates for the two patients were down-sampled to 1000Hz by averaging measurements
494 from each successive pair of time-points. The raw data from the target subdural electrodes for
495 the subsequent analysis were measured in reference to the electrode beneath the galea
496 aponeurotica in 4 patients (Patients 4,5,7 and 10) and to the scalp electrode on the mastoid
497 process contralateral to the side of electrode implantation in 4 patients (Patients 1-3 and 9).
498 Multivariate pattern-classification analyses were also conducted without such referencing and
499 yielded near-identical results. Baseline correction was performed by subtracting the mean pre-
500 stimulus baseline amplitude (200 ms before picture onset) from all data points in the epochs.
501 Trials with greater than +/-500 µV maximum amplitude were rejected as artifacts. Visual

502 inspection of all raw trials was conducted to reject any further trials contaminated by artifacts,
503 including canonical interictal epileptiform discharges. The mean waveform for each stimulus was
504 computed across repetitions. Data included, for each stimulus at each electrode, all
505 measurements beginning at stimulus onset and continuing for 1640ms. While this window
506 includes the onset of articulation toward the end, the critical results cannot reflect such motor
507 activity since all key phenomena are observe prior to mean time to initiate the utterance (1190
508 ms).

509 *Multivariate classification analysis.*The pre-processed data yielded, for each electrode in
510 each patient, a times-series of local field potentials sampled at 1000Hz over 1640ms for each of
511 100 stimuli. For each patient we trained classifiers to discriminate animal from non-animal images
512 given the LFPs evoked by each stimulus across all ventral-temporal electrodes in a 50ms time-
513 window. In a patient with 20 electrodes, one 50ms window contains 1000 measurements (50 LFPs
514 for each electrode x 20 electrodes). For each time window in every patient these measurements
515 were concatenated into feature vectors for each of the 100 stimuli, with the time windows
516 advancing along the time-series in 10ms steps. Thus the first window included 1-50ms post
517 stimulus onset, the next included 11-60ms, and so on. This procedure yielded feature vectors for
518 all 100 items in 160 time-windows for every subject.

519 The classifiers were logistic regression models fitted with L1 regularization53 using the
520 glmnet function in Matlab. L1-regularization applies an optimization penalty that scales with the
521 sum of the absolute value of the classifier coefficients and thereby encourages solutions in which
522 many features receive coefficients of 0. This approach is useful for a sliding-window analysis
523 because features receiving a 0 coefficient in the classifier have no impact on its performance when
524 it is assessed at other time points. So long as the information exploited by a classifier at time t is

present at a different time t +/- n, the classifier will continue to perform well, even if other features are in very different states. Thus classifiers trained with L1 regularization have the potential to show dynamic changes in the underlying code.

Classifier accuracy for a given time-window and subject was assessed using nested 10-fold cross-validation. In each outer fold, 10% of the data were held out, and the remaining 90% of the data were used with standard 9-fold cross-validation to search a range of values for the regularization parameter. When the best weight was selected, a model was fitted to all observations in the 90% of the training data and evaluated against the remaining 10% in the outer-loop hold-out set. This process was repeated 10 times with different final hold-outs, and classifier accuracy for each patient was taken as the mean hold-out accuracy across these folds. The means across patients are the data shown in Figure 4A and the diagonal of 4B in the main paper. A final classifier for the window was then fitted using all of the data and the best regularization parameter. This classifier was used to decode all other time-windows, yielding the off-diagonal accuracy values shown in Figure 4B.

The above procedures produced a pattern classifier for each of 160 50ms time-windows in every subject, with every classifier then tested at every time-window within each subject. Thus the classifier accuracy data were encoded in a 160x160-element decoding matrix in each subject. The matrices were averaged to create a single 160x160-element matrix indicating the mean decoding accuracy for each classifier at each point in time across subjects. This is the matrix shown in Figure 4B.

To better visualize how the code exploited by each classifier changes over time, we clustered the rows using the same agglomerative hierarchical approach described for the simulations. We considered solutions ranging from 4 to 15 clusters and plotted the mean

548 decoding accuracy over time across the classifiers within each cluster. All cluster sizes produced
549 the overlapping-waves pattern. In the main paper we show the 10-cluster solution as it is the
550 largest number in which each cluster after the first has a mean accuracy profile that is both
551 statistically reliable and higher than every other cluster at some point in time.

552 To assess the breadth of time over which a cluster showed reliable above-chance
553 classification accuracy, we again set Bonferroni-corrected significance thresholds using the
554 binomial distribution. Stimuli included 100 items, with a .5 probability of each item depicting an
555 animal. In the 1640ms measurement period there are 32 independent (ie non-overlapping) 50ms
556 time windows, and we assessed the mean classifier performance for each of 10 clusters at every
557 window. We therefore corrected for 320 multiple comparisons using a significance threshold of
558 68 correct ($p < 0.0001$ per comparison, $p < 0.03$ with correction).

559 *Surface plot visualization.* Magnetization-prepared rapid gradient-echo (MPRAGE)
560 volumetric scan was performed before and after implantation of subdural electrodes as a part of
561 presurgical evaluations. In the volumetric scan taken after implantation, the location of each
562 electrode was identified on the 2D slices using its signal void due to the property of platinum
563 alloy54. Electrodes were non-linearly co-registered to the patient MRI (MPRAGE) taken before
564 implantation, and then to MNI standard space (ICBM-152) using FNIRT
565 (www.fmrib.ox.ac.uk/fsl/fnirt/). The native coordinates of all the electrodes for all patients were
566 morphed into MNI space and resampled into 2 mm isotropic voxels55.

567 Projecting classifier coefficients to the surface. As described above, a separate logistic
568 classifier was fitted to each 50ms window in each subject. The classifier was specified as a set of
569 regression coefficients, with one coefficient for each timepoint at each electrode in the patient,
570 and many coefficients set to 0 due to L1-regularization. The sign of the classifier coefficient

571 indicates the “meaning” of a LFP deflection in a particular direction: a positive coefficient indicates
572 that animals are “signaled” by a positive deflection in the LFP, while negative coefficients indicate
573 that animals are signaled by a negative deflection. The magnitude of the coefficient indicates the
574 “importance” of the measurement, in the context of all other LFPs appearing in the classifier. The
575 distribution of coefficient directions and magnitudes across the cortex and over time thus
576 provides an indication of how the underlying neuro-semantic code changes over time.

577 For a single time window we computed, separately for each electrode in each participant,
578 the magnitudes (sum of absolute values) of the classifier weights across the 50 time points in the
579 window. The resulting data were exported from Matlab to NIFTI volumes using the NIFTI
580 toolbox(https://www.mathworks.com/matlabcentral/fileexchange/8797-tools-for-nifti-and-
581 analyze-image) and projected from all electrodes and subjects onto the common cortical surface
582 map using AFNI’s 3dVol2Surf relative to the smooth white matter and pial surfaces of the ICBM
583 152 surface reconstructions shared by the AFNI team and the NIH (https://afni.nimh.nih.gov/
584 pub/dist/tgz/suma_MNI152_2009.tgz). The space between corresponding nodes on the two
585 surfaces were spanned by a line segment sub-divided at 10 equally spaced points. The value
586 displayed on the surface is the average of the values where these 10 points intersect with the
587 functional volume along that line segment. Once mapped to the surface, the results were spatially
588 smoothed along the surface with an 8mm full-width half-max Gaussian kernel using the
589 SurfSmooth function in SUMA. We inclusively masked any surface point with a non-zero value in
590 this surface projection. A separate mask was generated for each time window.

591 To visualize how the representational code changes over time within the surface mask we
592 next carried out a similar procedure on the classifier coefficients themselves, without taking the
593 absolute values. At each electrode in every subject we summed the classifier coefficients over the

50ms time window, yielding a single positive or negative real-valued number at each electrode for each time window. These values were again projected onto a common brain surface and spatially smoothed with an 8mm FWHM Gaussian blur along the surface. In the resulting maps, any colored point indicates a cortical region that received a non-zero value in the weight magnitude mask, while the hue indicates the direction of the classifier coefficient in the area—that is, whether a positive deflection of the LFP for nearby electrodes indicated that the stimulus was an animal (warm colors), a non-animal (cool colors), or showed no systematic direction (green). A separate map of this kind was generated for each of 160 time windows. We animated the results to visualize how they change over time using the open-source ffmpeg software (https://ffmpeg.org/) with linear interpolation between successive frames. The animation is shown in Movie S1; snapshots of this visualization are shown in Figure 4E.

Acknowledgments

This work was partially supported by Medical Research Council Programme Grant MR/J004146/1 and by European Research Council grant GAP: 502670428 - BRAIN2MIND_NEUROCOMP.

Competing interests

The authors declare no competing interests.

612 **Supplementary Information**

613 **SI-1: Comparison of deep model results to control models.**

614 The main paper highlights four properties of the neural decoding results observed in both the
615 deep neural network model and in the human ECoG data: constant decodability, local temporal
616 generalization, a widening window of generalization, and change in neural code direction in the
617 ATL hub. We suggested that these properties arise because semantic structure is encoded as
618 distributed activation patterns that change in highly nonlinear ways due to their situation in the
619 deep cross-modal hub of a dynamic cortical network. This argument implies that the signature
620 pattern would not arise in models that adopt different kinds of representation and processing
621 mechanisms, nor in the shallower layers of the deep model. In this section we assess this
622 implication by comparing the main results with those observed in three alternative models of
623 semantic representation.

624 *Distributed versus feature-based representations.*

625 By distributed representation, we mean that many neural populations or units can jointly
626 contribute to representation of structure even if they do not each independently encode the
627 structure. Deep neural network models are capable of acquiring distributed representations of
628 this kind (22) and may be contrasted with models proposing that semantic representations are
629 comprised of elements that each independently detect a particular semantic feature, such as
630 membership in a particular conceptual domain or category. We therefore considered what the
631 decoding signature would look like in such a feature-based model. The 90 items were represented
632 with a vector in which two elements were dedicated to each conceptual domain (animal, plant,
633 object) and to each basic-level category (bird, fish, flower, etc; total of 9 categories). For instance,

a particular instance of flower would activate the two “plant” features and the two “flower” features; an instance of tree would activate the same two “plant” features and two “tree” features, etc. This yielded 24 elements total; to equate the number of features with the number of units in the deep network simulation, we added a 25th vector element that always adopted a low activation value.

We simulated the gradual activation of features over the course of processing by generating a 33-step time-series for each feature and each item presentation. All units began with an activation of 0, and features true of the stimulus would ramp up their activation according to a sigmoid function with a constant slope and a randomly-sampled offset term determining when in the stimulus presentation window the feature would begin to activate. This procedure yielded a dataset analogous to the evolution of internal representations in the deep network, but with feature-based semantic representations in which features activated with randomly-sampled time-courses.

Dynamic versus linear.

By dynamic processing, we mean that units can influence themselves via feedback from the other units to which they send connections. Reciprocally connected sets of units are coupled and so behave as a dynamic system in which states evolve together over time. Often the dynamics in such a system are non-linear, producing radical changes in the ways that neural states encode information. Thus the importance of dynamic processing in the deep neural network model can be assessed by contrasting the primary results with an alternative model that employs starting and ending representations identical to the deep model, but with intermediate states simply moving in a straight line from start to finish. We created such a model by recording the initial and

final representations arising in the deep neural network model, then creating a 33-step time-series for each stimulus representing a linear interpolation of the representation moving from initial to final state. For a given stimulus, each step of the time series was created as a proportional weighted average of the initial and final states, with the first step giving all the weight to the initial representation, subsequent steps gradually shifting more weight to the final representation, and the last step giving zero weight to the initial representation. This procedure yielded a dataset in which initial and final representations were distributed identically to the deep network, but the trajectories of the representations over time were linear interpolations between these.

Deep versus shallow.

By deep network, we mean a neural network that has multiple hidden layers interposing between Visual and Verbal representation units. Depth generally allows neural networks to discover and represent more complex statistical relations amongst connected units(54). It also allows for more complex temporal dynamics as mutual influences across distal network components take more processing time. We assessed the importance of network depth in two ways.

First we compared the behavior of the hub layer in the deep network to that of a shallow network employing just a single hidden layer containing 25 units reciprocally connected to Visual and Verbal units and parameterized identically to the hub units in the deep model. We trained the network for 30k epochs exactly as described for the deep network, using the same training and testing patterns and procedures. This procedure yielded a dataset in which internal representations were distributed and dynamic as in the deep model, but arose within a shallower network.

677 Second, we compared the behavior of the hub layer in the deep network to the patterns emerging
678 across intermediate (visual hidden and verbal hidden) and shallow (visual representation and
679 verbal representation) layers in the same model. For this comparison, we recorded the activation
680 time-series produced in response to each visual stimulus, for every unit in the model. We then
681 assessed the propensity for units in each layer to behave like individual feature detectors,
682 unresponsive units, or units that appear to "switch" their category preference over time, taking
683 the "switch" behavior as a marker of distributed and dynamic representation.

684 *Results*.

685 In the first analysis, we subjected each alternative model to the same analyses reported for the
686 primary model and assessed whether they also show the four signature properties identified in
687 the main paper.

688 *Constant decodeability*. All four models showed cross-validation accuracy reliably above chance
689 and consistently high across the time-window once input signals reached the representation
690 units—thus all models showed constant decodeability.

691 *Local temporal generalization*. Figure SI-1A shows a 3D MDS of the trajectories for all items
692 through the corresponding representation space in each model. For feature-based, linear, and
693 shallow models, the trajectories are strictly or nearly linear—only the deep, distributed and
694 dynamic model shows the nonlinearities discussed in the main paper. Consequently the models
695 show qualitatively different patterns of generalization over time: feature-based, linear, and
696 shallow models show a pattern in which all classifiers generalize poorly to earlier timepoints and
697 well to later timepoints. Thus local temporal generalization—in which classifiers do well only for

neighboring time points in both past and future—is only observed in the deep, distributed and dynamic model (SI-1B).

Widening generalization window. In contrast to the ECoG data and the deep, distributed and dynamic model, the alternative models all show a narrowing window of temporal generalization: models fitted early in processing show good performance over a wider window than those trained later. This follows from the linear progression of internal representations observed in these models: a classification plane that succeed for early representations must also succeed for later representations. Conversely, because representations are very well-separated toward the end of processing, classification planes that succeed later in time may not perform well for earlier representations, producing an symmetric temporal generalization profile.

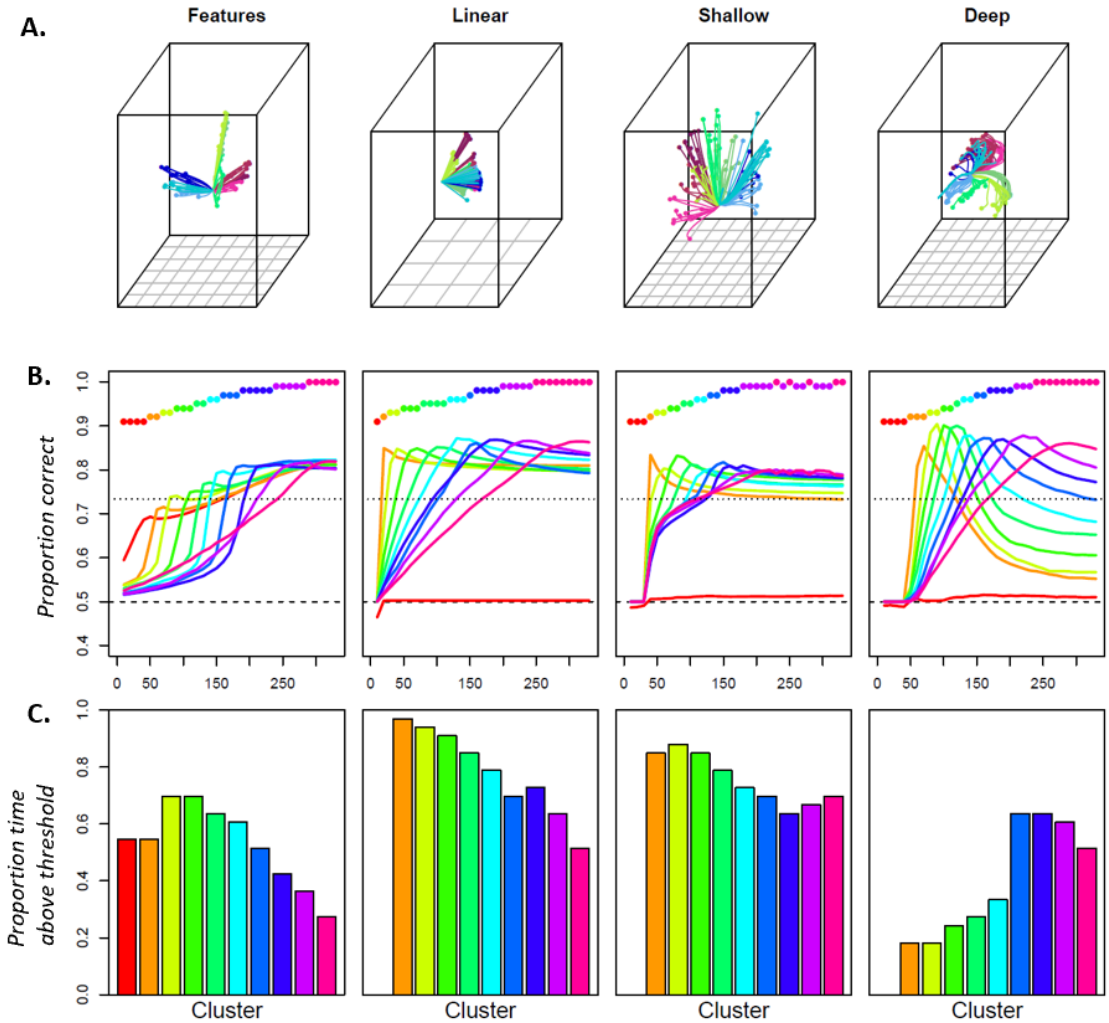

Figure SI-1. Comparison of simulation results for a feature-based model, a distributed linear model, a shallow recurrent network, and the deep, distributed and dynamic model. A. Multi-dimensional scaling showing the trajectory of each item through representation space under four different models. Only the deep model shows radically nonlinear change. B. Mean accuracy for clusters of classifiers under each model type. Only the deep model shows the overlapping-waves pattern. C. Proportion of time-window where classifiers in each cluster show reliably above-chance responding. Only the deep model shows a generalization window that widens over time.

Change in code direction. The deep, distributed and dynamic model acquired representations in which some single units, when analyzed independently, behaved like feature-detectors that change in direction over processing—with high activations initially predicting an animal stimulus, for instance, then later predicting a non-animal stimulus. A similar flipping of signal direction was

713 also observed in the more anterior parts of the ventral temporal lobe, via the changing sign of the

714 classifier coefficients identified in the ECoG data. We therefore considered whether a change in

715 code direction was observed for single units considered independently in the alternative models.

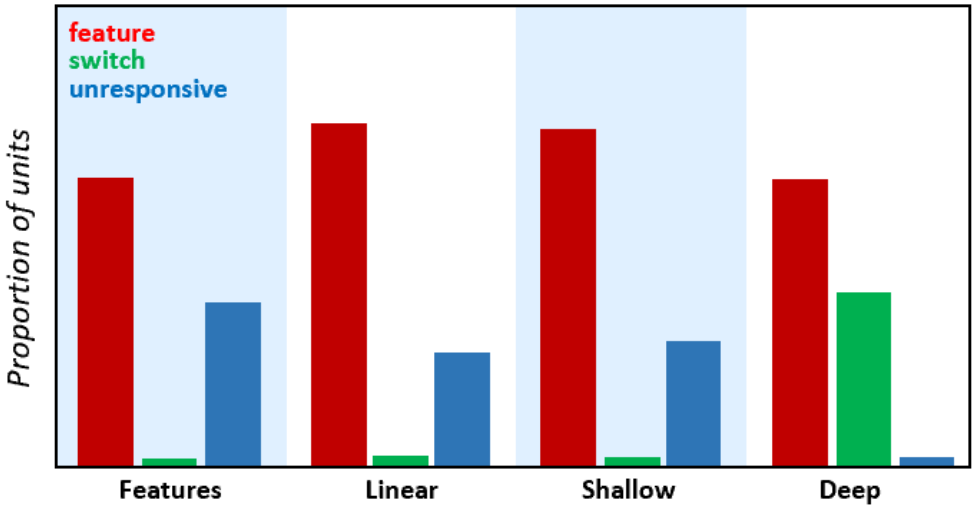

Figure SI-2. For each model type, the proportion of units that behave like feature-detectors (red), detectors that switch their category preference over time (green), and units that seem unresponsive to the semantic category (blue). Only the deep, distributed, dynamic model has units whose responses switch their category preference over time.

716

717 Specifically, we classified each unit in each simulation as (1) a feature-detector if its activity

718 correlated significantly with conceptual domain in only one direction over the time-course of

719 processing, (2) as a switch feature if it correlated significantly in both the positive and the negative

720 direction at different points in time, and as (3) non-responsive if it never correlated significantly

721 with conceptual domain. Across all five simulations and all three classification tasks, we computed

722 the proportion of units falling into each category for each model type. The results are shown in

723 Figure SI-2. For feature-based and linear models, the results are trivial, since the representations

724 are constrained to show only non-responses or feature-like behaviors, as was indeed observed.

725 That such a result is observed in these cases validates the analysis. More interestingly, the shallow

interactive model also learned unit responses that behaved either as feature-detectors or were non-responsive. Only the deep, distributed and dynamic model acquired units that appeared to switch their category preference.

SI-2: Stable features posteriorly with dynamic features anteriorly

Our analysis of ECoG data showed more consistent feature-like responding in more posterior ventral temporal regions, and more variability in code direction over time in vATL. We suggested that this difference between posterior and anterior regions arises because the vATL is deeper in the network and is connected cross-modally—that is, it resides in tertiary association cortex. In the model, the visual and verbal hidden layers are shallower and not directly cross-modal—they receive modality-specific inputs, and interact only with the hub layer. Thus our explanation of the ECoG data has a direct model analog: if it is correct, more superficial layers should be more likely to learn static, feature-like representations, while the hub should be more prone to acquire distributed dynamic representations. We tested this hypothesis by computing the number of units in each layer that behave as feature-detectors, switch-units, or unresponsive units, taking switch-units to indicate a distributed and dynamic code.

Results are shown in Figure SI-3. Units in both shallow and intermediate model layers behaved like consistent feature-detectors, or appeared unresponsive to the superordinate semantic category. The switching pattern diagnostic of nonlinear representational change only emerged in the deep hub layer. Note that the pattern cannot reflect overall distance in the network from the input activation, since the verbal hidden and representation units are further away from the visual inputs in the network than are the hub units but still show the same feature-like responding.

Instead the pattern must reflect the centrality of the hub units in computing interactive mappings between visual and verbal representations.

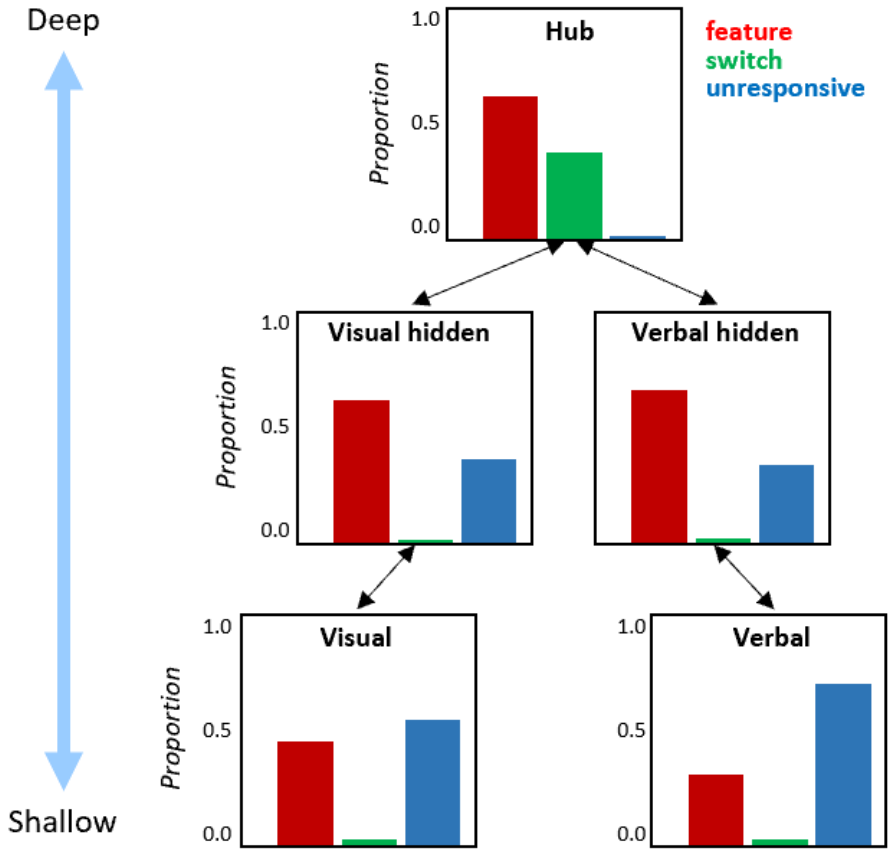

Figure SI-3. For each layer in the deep, distributed and dynamic model, the proportion of units that behave like feature-detectors (red), detectors that switch their category preference over time (green), and units that seem unresponsive to the semantic category (blue) when the model processes visual inputs. Only the hub layer of the network—the model analog to the ventral anterior temporal cortex—contained units whose responses switch their category preference over time.

Together these simulations establish that all three properties—distributed representation, dynamic processing, and network depth—conspire to yield the decoding signature observed in the ECoG data: local temporal generalization, a widening window of generalization, and neural populations whose code direction appears to change over time when considered independently.

SI-3: Independent correlation analysis of individual electrodes.

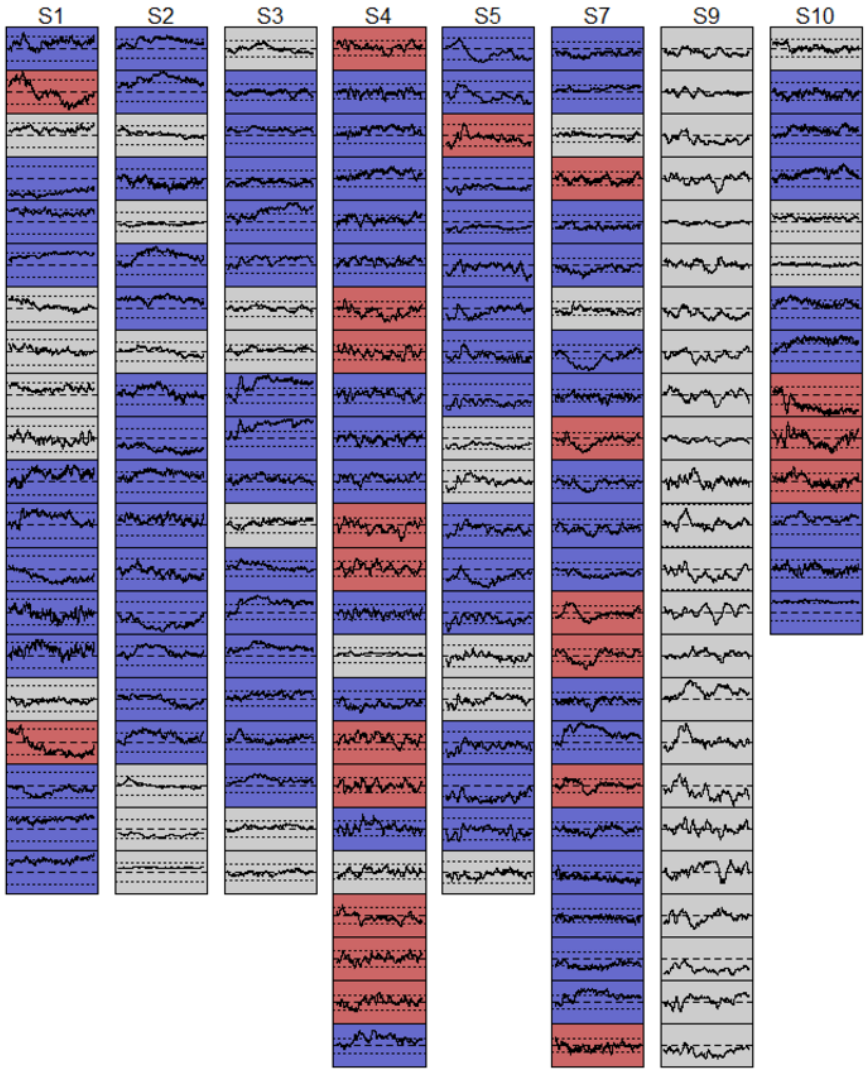

Figure SI-4. Each panel shows, for each electrode in each participant, the correlation over time between the measured LFP for each item and the category label. Dotted lines show statistical significance threholds for this correlations. Gray panels never exceed the threshold in either direction. Blue panels exceed it in one direction only. Red panels exceed it in both directions at different points in time.

The preceding simulations, together with those reported in the main paper, find model units that, considered independently, seem to "flip" their category preference over time. A similar phenomenon was observed in the anterior part of the ventral temporal lobe when visualizing

classifier weights plotted on the cortical surface: the direction and spatial organization of classifier coefficients fluctuated over time. The coefficients show how the LFPs measured at a particular location and timepoint contribute to the classifier's response when the responses of other regions are also taken into account, and so are useful for visualizing a distributed neural code. Because coefficients do not reflect the independent correlation of an electrode's activity with the category label, however, the model and ECoG analyses are not strictly analogous. For an apples-to-apples comparison, we conducted the same independent-correlation analysis of individual electrodes in each participant. At every timepoint following stimulus onset, for each electrode across all stimuli, we computed the correlation between the electrode's measured field potential and a binary category label.

Figure SI-4 shows these correlations plotted over time. Dotted lines indicate the correlation significance threshold of $p < 0.05$ for a two-tailed probability test on the correlation coefficient. Gray panels show electrodes that never exceeded the threshold; blue panels show electrodes that exceeded it in one direction only; red panels show electrodes that exceeded the threshold in different directions at different points in time. Two observations are warranted. First, electrodes in 5 of 8 patients show "switching" behavior analogous to that identified in the model, consistent with a distributed and dynamic code. Second, independent correlations never exceeded the significance threshold for one participant (patient 9)—but semantic category was nevertheless decodeable in this patient for some time windows, indicating that information can be present in a distributed code even when not discernable in individual channels.

781 **SI-4: Decoding anterior electrodes only.**

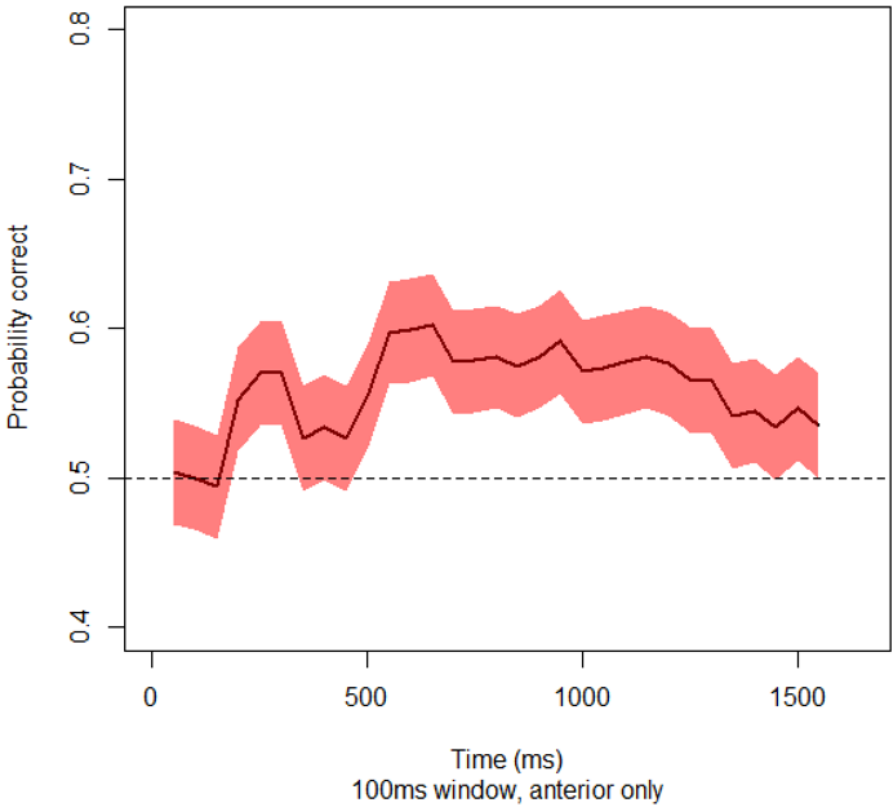

Figure SI-5. Accuracy for classifiers trained on anterior electrodes only. Curve shows expected probability of correct classification and 95% confidence intervals across participants at each window for classifiers trained only on vATL electrodes.

782

783 As noted in the main text, the interpretation of weights in a multivariate pattern classifier can be
784 problematic in the presence of correlated noise(45). When measurements from informative and
785 informative channels are corrupted with the same noise source, a classifier will place weights of
786 contrasting sign on both, with the weight on the uninformative channel serving to "cancel out"
787 the noise on the informative channel, boosting overall classifier performance. Thus an alternative
788 explanation of the ECoG results is that semantic information is encoded solely within the stable,
789 feature-like responses of mid/posterior regions of ventral temporal cortex, but correlated noise

790 across channels leads the classifier to place fluctuating non-zero weights on vATL areas. Under
791 this hypothesis, classifiers trained only on vATL electrodes should not show above-chance
792 classification. We therefore assessed whether semantic category (animate/inanimate) can be
793 decoded only from the more anterior electrodes.

794 We first computed a median split of electrodes based on their rostral/caudal location, discarding
795 the most posterior half of the electrodes (those more posterior than -17 in the anterior/posterior
796 dimension of MNI space). We then fit L1-regularized logistic classifiers to the anterior electrodes
797 only, using the same procedure described for the main analysis on a 100ms moving window,
798 computed every 50ms across the 1650ms post-stimulus response time. Classifiers were fit
799 separately at each window, for each participant, and were evaluated using 10-fold cross-
800 validation. From these data we computed, for each classifier, total number of items classified
801 correctly across the 10 holdouts. Across participants, we then computed expected probability of
802 correct classification and 95% confidence intervals (from the binomial distribution) on this
803 expectation at each window. The results are shown in Figure SI-5: from about 200ms post stimulus
804 onset, accuracy remained reliably above chance across most of the decoding window. Thus,
805 consistent with other arguments presented in the main paper, vATL regions encode semantic
806 structure even considered independent of the more posterior regions that adopt a more stable
807 and feature-like code.

808 SI-5: Autocorrelation analysis.

809 In both the simulation and ECoG data we observed that pattern classifiers generalize over a
810 temporal window that widens over time as a stimulus is processed. We suggested this widening
811 arises because the pattern of activation over hub units grows increasingly stable as the system

generates the correct activation pattern across the rest of the network. This section considers an alternative explanation of the ECoG result: perhaps the ECoG data itself is temporally structured so that it becomes self-predictive over a broader timescale as a stimulus processed. To test this possibility, we measured the temporal autocorrelation of each electrode at lags of 1-100ms, calculated over a 300ms sliding window, beginning with stimulus onset and moving the window forward in 100ms increments. If temporal autocorrelation broadens over time, we would expect significant correlation over a longer lag in windows selected from later in the time series.

Figure SI-5 shows, for each subject, the temporal autocorrelation across lagtime, averaged across all electrodes, for 300ms beginning at stimulus onset and sliding forward to the end of the time-series. Line colors indicate when in the time-series the 300ms window was selected. If electrode signals grow self-predictive over a broader timespan with processing, cooler colors would show a shallower slope and a later intersection with the 0 line on the y axis. Instead all lines are practically on top of one another, and showing that temporal autocorrelation within a 300ms window drops to 0 at a lag of about 60ms. Thus the "widening window" pattern observed in the ECoG temporal generalization data does not arise from shifts in the temporal autocorrelation of the data itself, and must reflect an increasingly stable distributed code.

References

1. J. L. McClelland, T. T. Rogers, The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience* **4** (2003).

2. T. T. Rogers, J. L. McClelland, *Semantic Cognition: A Parallel Distributed Processing Approach* (MIT Press, 2004).

3. A. Martin, GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic Bulletin & Review* **23**, 979–990 (2016).

4. A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, J. L. Gallant, Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).

5. K. Patterson, P. J. Nestor, T. T. Rogers, Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience* **8** (2007).

6. M. A. L. Ralph, E. Jefferies, K. Patterson, T. T. Rogers, The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience* **18**, 42–55 (2017).

7. T. T. Rogers, *et al.*, Structure and Deterioration of Semantic Memory: A Neuropsychological and Computational Investigation. *Psychological Review* **111** (2004).

8. L. Isik, E. M. Meyers, J. Z. Leibo, T. Poggio, The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology* **111**, 91–102 (2014).

9. T. Serre, A. Oliva, T. Poggio, A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences* **104**, 6424–6429 (2007).

10. R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports* **6**, 27755 (2016).

11. N. Kriegeskorte, Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science* **1**, 417–446 (2015).

12. L. Chen, M. A. Lambon Ralph, T. T. Rogers, A unified model of human semantic knowledge and its disorders. *Nature Human Behaviour* **1**, 0039 (2017).

13. T. T. Rogers, L. Watling, J. R. Hodges, K. Patterson, A basic-level disadvantage for speeded category verification in *Cognitive Neuroscience Society*, (2005), p. 211.

14. M. J. Farah, J. L. McClelland, A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General* **120**, 339–357 (1991).

15. G. Cree, K. McRae, C. McNorgan, An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science* **23**, 371–414 (1999).

16. M. W. Harm, M. Seidenberg, Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review* **111**, 662–720 (2004).

17. D. C. Plaut, T. Shallice, Perseverative and Semantic Influences on Visual Object Naming Errors in Optic Aphasia: A Connectionist Account. *JCN* **5**, 89–117 (1993).

18. D. J. Kravitz, K. S. Saleem, C. I. Baker, L. G. Ungerleider, M. Mishkin, The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences* **17**, 26–49 (2013).

19. E. Goddard, T. A. Carlson, N. Dermody, A. Woolgar, Representational dynamics of object recognition: Feedforward and feedback information flows. *NeuroImage* **128**, 385–397 (2016).

20. T. C. Kietzmann, *et al.*, Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences* **116**, 21854–21863 (2019).

21. F. Pulvermüller, How neurons make meaning: brain mechanisms for embodied and abstract-symbolic semantics. *Trends in Cognitive Sciences* **17**, 458–470 (2013).

22. C. R. Cox, M. S. Seidenberg, T. T. Rogers, Connecting functional brain imaging and Parallel Distributed Processing. *Language, Cognition and Neuroscience* **30**, 380–394 (2015).

23. V. Mante, D. Sussillo, K. v. Shenoy, W. T. Newsome, Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature* **503**, 78–84 (2013).

24. E. W. Contini, S. G. Wardle, T. A. Carlson, Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia* **105**, 165–176 (2017).

25. T. Grootswagers, S. G. Wardle, T. A. Carlson, Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience* **29**, 677–697 (2017).

26. A. Clarke, B. J. Devereux, B. Randall, L. K. Tyler, Predicting the Time Course of Individual Objects with MEG. *Cerebral Cortex* **25**, 3602–3612 (2015).

892 27. J.-R. King, S. Dehaene, Characterizing the dynamics of mental representations: the
893 temporal generalization method (2014) https:/doi.org/10.1016/j.tics.2014.01.002 (July 8,
894 2019).

895 28. T. T. Rogers, *et al.*, The structure and deterioration of semantic memory: a
896 computational and neuropsychological investigation. *Psychological Review* **111**, 205–235
897 (2004).

898 29. J. S. Snowden, P. J. Goulding, D. Neary, Semantic dementia: A form of circumscribed
899 temporal atrophy. *Behavioural Neurology* **2**, 167–182 (1989).

900 30. J. R. Hodges, N. Graham, K. Patterson, Charting the progression in semantic dementia:
901 Implications for the organisation of semantic memory. *Memory* **3**, 463–495 (1995).

902 31. M. A. Lambon Ralph, C. Lowe, T. T. Rogers, The neural basis of category-specific semantic
903 deficits for living things: Evidence from semantic dementia, HSVE and a neural network
904 model. *Brain* **130**, 1127–1137 (2007).

905 32. J. Acosta-Cabronero, *et al.*, Atrophy, hypometabolism and white matter abnormalities in
906 semantic dementia tell a coherent story. *Brain* **134**, 2025–2035 (2011).

907 33. M. Visser, E. Jefferies, M. A. Lambon Ralph, Semantic processing in the anterior temporal
908 lobes: A meta-analysis of the functional neuroimaging literature (2010) (July 31, 2013).

909 34. T. T. Rogers, *et al.*, Anterior temporal cortex and semantic memory: Reconciling findings
910 from neuropsychology and functional imaging. *Cognitive, Affective and Behavioral*
911 *Neuroscience* **6** (2006).

35. R. J. Binney, G. J. M. Parker, M. A. Lambon Ralph, Convergent Connectivity and Graded Specialization in the Rostral Human Temporal Lobe as Revealed by Diffusion-Weighted Imaging Probabilistic Tractography. *Journal of Cognitive Neuroscience* **24**, 1998–2014 (2012).

36. R. J. Binney, K. v Embleton, E. Jefferies, G. J. M. Parker, M. A. L. Ralph, 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. *Cerebral cortex (New York, N.Y. : 1991)* **20**, 2728–38 (2010).

37. G. Pobric, E. Jefferies, M. A. Lambon Ralph, Category-specific versus category-general semantic impairment induced by transcranial magnetic stimulation. *Current biology : CB* **20**, 964–8 (2010).

38. G. Pobric, E. Jefferies, M. A. Lambon Ralph, Anterior temporal lobes mediate semantic representation: mimicking semantic dementia by using rTMS in normal participants. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20137–41 (2007).

39. R. Bruffaerts, *et al.*, Similarity of fMRI activity patterns in left perirhinal cortex reflects semantic similarity between words. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **33**, 18597–607 (2013).

40. B. J. Devereux, A. Clarke, A. Marouchos, L. K. Tyler, Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *The Journal of neuroscience : the official journal of the Society for Neuroscience* **33**, 18906–16 (2013).

41. L. Sha, *et al.*, The Animacy Continuum in the Human Ventral Vision Pathway. *Journal of Cognitive Neuroscience* **27**, 665–678 (2015).

42. A. Martin, L. L. Chao, Semantic memory in the brain: Structure and processes. *Current Opinion in Neurobiology* **11**, 194–201 (2001).

43. S. Anzellotti, B. Z. Mahon, J. Schwarzbach, A. Caramazza, Differential activity for animals and manipulable objects in the anterior temporal lobes. *Journal of cognitive neuroscience* **23**, 2059–67 (2011).

44. B. Z. Mahon, S. Anzellotti, J. Schwarzbach, M. Zampini, A. Caramazza, Category-specific organization in the human brain does not require visual experience. *Neuron* **63**, 397–405 (2009).

45. S. Haufe, *et al.*, On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage* **87**, 96–110 (2014).

46. Y. Chen, *et al.*, The 'when' and 'where' of semantic coding in the anterior temporal lobe: Temporal representational similarity analysis of electrocorticogram data. *Cortex* **79**, 1–13 (2016).

47. T. Carlson, D. A. Tovar, A. Alink, N. Kriegeskorte, Representational dynamics of object vision: The first 1000 ms. *Journal of Vision* **13**, 1–1 (2013).

48. R. M. Cichy, D. Pantazis, A. Oliva, Resolving human object recognition in space and time. *Nature Neuroscience* **17**, 455–462 (2014).

49. A. C. Nobre, T. Allison, G. McCarthy, Word recognition in the human inferior temporal lobe. *Nature* **372**, 260–263 (1994).

50. P. Gloor, *The Temporal Lobe and Limbic System* (Oxford University Press, 1997).

51. M. Mesulam, From sensation to cognition. *Brain* **121**, 1013–1052 (1998).

52. T. T. Rogers, *et al.*, The structure and deterioration of semantic memory: A computational and neuropsychological investigation. *Psychological Review* **111**, 205–235 (2004).

53. L. Chen, M. A. Lambon Ralph, T. T. Rogers, A unified model of human semantic knowledge and its disorders. *Nature Human Behaviour* **1** (2017).

54. R. K. Srivastava, K. Greff, J. Schmidhuber, Training Very Deep Networks. 2377–2385 (2015).