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
*ttrogers@wisc.edu

[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
**matthew.lambon-ralph@mrc-cbu.cam.ac.uk

Abstract: How does the human brain encode semantic information about objects? One view proposes that semantic representations arise from the propagation of information upward through a hierarchy of perceptual and conceptual features. We provide evidence supporting an alternative view: semantic representations arise as rapidly-changing neural patterns within a distributed dynamic system. In a novel application of pattern classification to simulated neural data, we identified the unique decoding "signature" of a dynamically-changing semantic code as predicted by a deep recurrent neural network model. Applying the technique to neural signals collected directly from human ventral temporal cortex, we observed the same signature—suggesting that the brain employs a distributed and dynamic semantic code possessing stable elements posteriorly and rapidly-changing elements anteriorly. The results challenge a feature-based view of semantic representation, resolve conflicting findings from past research, and provide a new framework for understanding the time-course of distributed representation in the brain.

Semantic memory supports human understanding of language and experience: our remarkable 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[6,7]. 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.[6,12] 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.[1] 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]. Models of semantic representation commonly adopt this perspective[14–18], which also aligns with the general view that the ventral visual processing stream is recurrent and interactive.[19,20]

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], or vice versa[5]. 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. These possibilities suggest that the feature-based research program can significantly 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 plot hypothetical joint responses to each stimulus over time from onset to response. The leftmost panel shows the full trajectories traced through the 2D space over time by various mammals (red) and tools (green). The remaining panels show how a multivariate classifier might partition 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. Because the trajectories are nonlinear, however, each population's contribution to this structure changes over time—the category discrimination plane rotates.

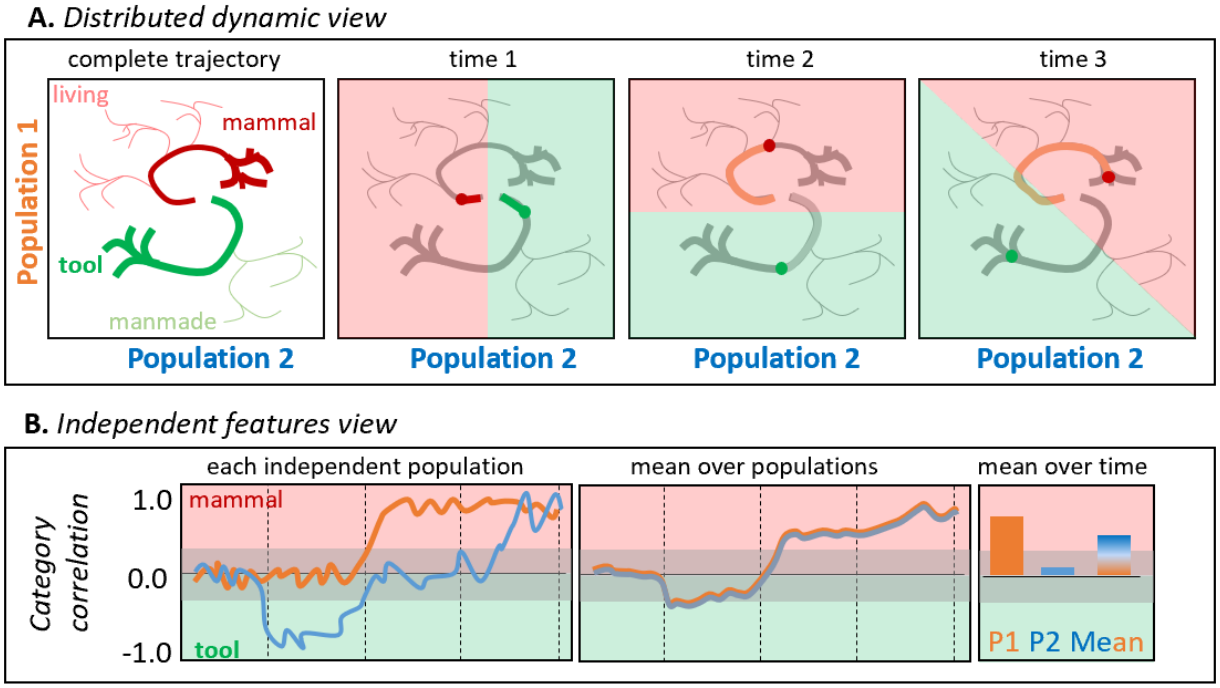

Figure 1. A. Hypothetical joint activations of two neural populations to living and manmade items (left), and the classification plane that would best discriminate tools from mammals at different timepoints. B. Independent correlations between each population's activity and a binary category label for 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 under a feature-based analysis. The lines show, for each population over time, the correlation between the population's response to various stimuli and a binary category label. Population 1 correlates positively with the category label beginning about half-way through the time series, while population 2 initially correlates negatively with the label, then shows no correlation, then correlates positively. It might seem that (a) animals are detected only midway through the time-series, (b) tools are detected earlier than animals and (c) there exist populations that "switch" from tool-detectors to mammal-detectors. The bottom middle panel shows the mean response across the two populations at each point in time, as might be observed if the neurophysiological measurement lacks spatial resolution (e.g., EEG/MEG). The populations appear to detect animals late in the trial. The right panel shows mean responses across time for each population (consistent with high spatial and low temporal resolution; e.g. fMRI) and for both populations together (low spatial and temporal resolution). The former suggests that population 2 plays no important role distinguishing the categories, while the latter suggests the two populations together selectively detect animals. These conclusions are incorrect from the distributed and dynamic perspective, under which the two populations always jointly differentiate the categories but the contribution of each changes over time (Figure 1A).

The interpretation of neurophysiological signals in the cortical semantic system thus depends critically upon whether the neuro-semantic code is feature-based or distributed and dynamic, but efforts to adjudicate the question face three significant hurdles. First, spatial and/or temporal averaging can obscure important signal if the code truly is distributed and dynamic—discovery requires neural data with high temporal and spatial resolution, ruling out the non-invasive methodologies that constitute the great majority of work in this area[23]. Second, independent univariate analysis can mischaracterize information distributed across multiple sites—discovery requires multivariate methods[24]. Third, recent studies connecting neurophysiologal measurements to computational models of recognition have focused exclusively on feed-forward models that do not exhibit dynamic processing.[10,25] No prior work has assessed how representations evolve within distributed and dynamic semantic models, and consequently it is unclear what this view predicts or how it differs from feature-based approaches.

We therefore combined computational modeling, multivariate pattern classification, and electrocorticography (ECoG) to assess whether semantic representations in human vATL are distributed and dynamic. We first showed that the nonlinear dynamics highlighted in our thought experiment arise in a well-studied neurocomputational model of semantic memory[26,27]. We next used temporal generalization of pattern classifiers[28] to establish the unique "signature" of a dynamically changing semantic code. We then applied this technique to ECoG data collected from the surface of human vATL while participants named line drawings of common items, and found the critical decoding signature—providing strong evidence that semantic structure is expressed in human vATL by a distributed code that changes dynamically with stimulus processing. The results challenge the view that semantic representations are encoded as activity patterns over independent feature detectors, and more generally offer a new framework for thinking about the time-course of representation in distributed neural systems.

Results

Simulation study. Simulations provide a formal basis for understanding the implications of the distributed and dynamic view, because a computer model's architecture, behavior, learning and testing patterns are fully known. Adapting prior models [16,26,27], we therefore assessed whether hub representations change dynamically with stimulus processing and how multivariate classification can uncover such a code.

The model was a fully continuous and recurrent neural network that learns cross-modal associations between distributed visual and verbal representations via three reciprocally-connected hidden layers (Figure 2A). Given positive input to a subset of visual or verbal units, it learns to activate the corresponding item's unspecified visual and/or verbal attributes. Activity propagates in both directions, from surface representations to hub and back, so that the trained model settles to an attractor state representing any item specified by the input. We trained the model with patterns representing ninety items from three model conceptual domains (e.g., animals, objects, and plants), each organized into three categories containing ten items. Items from different categories in the same domain shared a few properties while those in the same category shared many. We trained the model for 30k epochs, attaining a mean accuracy of 99% (see Methods). We then presented the model with visual input for each item and recorded the resulting activation patterns over hub units for each tick of simulated time as the network settled.

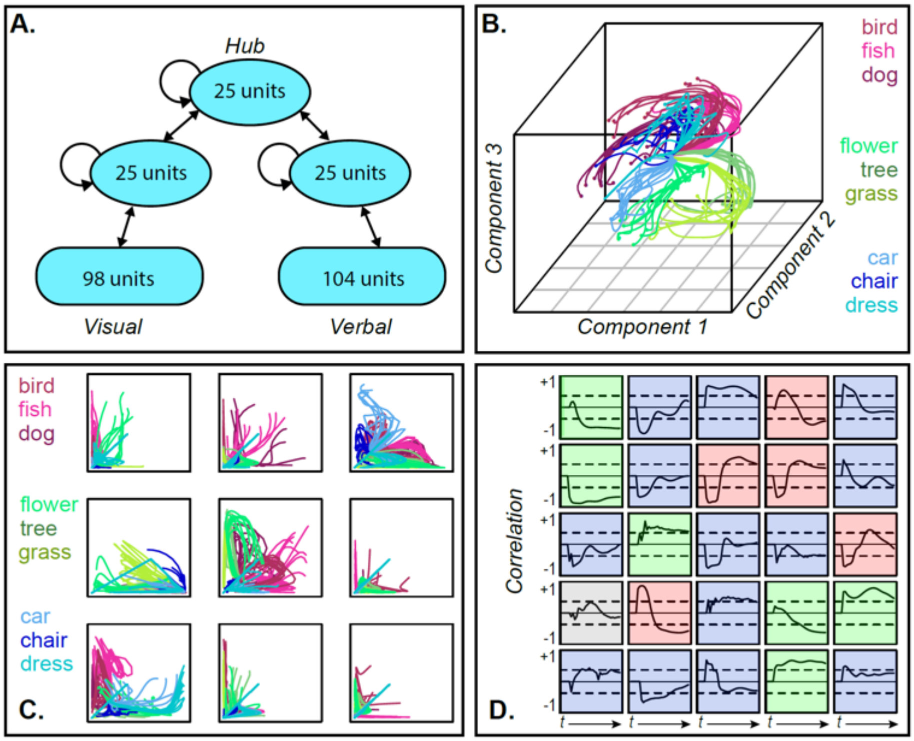

Figure 2. A. Model architecture. B. 3D MDS of hub activation patterns learned in one model run—each line shows the trajectory of a single item over time in the compressed space. C. The same trajectories shown in uncompressed unit activations for 9 randomly sampled unit pairs, horizontal and vertical 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).

To visualize how model internal representations changed during stimulus processing, we computed a 3D multi-dimensional scaling (MDS) of activation patterns at all timepoints during “perception” of each stimulus, then plotted the changing representation for each item as a line in this space. The result in Figure 2B shows a systematic but nonlinear elaboration of the conceptual structure latent in the stimulus patterns: the domains separate from one another early on, but each item follows a curved trajectory over time. Figure 2C shows these trajectories in the native activations of randomly-sampled unit pairs in one network run: they appear even more radically nonlinear. Consequently, independent analysis of each unit's behavior produces mixed results (Figure 2D), with some units behaving like tonic category detectors (green squares), some like transient detectors (blue), some appearing to flip their category preference (red) and others appearing not to code category information at all (gray).

Thus the full distributed pattern elaborates conceptual structure from early in processing, but the progression is nonlinear and only clearly discernable in a low-dimensional embedding of the space. Such an embedding can be computed for the model because we know which units are important and can apply the MDS to all and only those units. The same approach cannot be applied to ECoG data for two reasons. First, one cannot know *a-priori* which channels record signals relevant for semantic representation and thus cannot simply compute a low-dimensional embedding of all data collected. Instead one must fit a statistical model that will selectively weight signals useful for discerning semantic structure. Second, whereas the model allows access to the entire network, a cortical surface sensor array only sparsely samples the neural responses contributing to a semantic representation. The problem requires a multivariate statistical model capable of revealing a dynamically changing neural code when fitted to sparsely-sampled neural data.

We therefore used multivariate logistic classifiers to decode semantic category information from hub activation patterns and assessed their behavior on simulated ECoG data. For each simulated participant we selected a sparse random subsample (15%) of all hub units and recorded their responses to each stimulus at every tick of time. We fitted a separate classifier at each timepoint to distinguish two semantic domains from the activation patterns elicited over the 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.

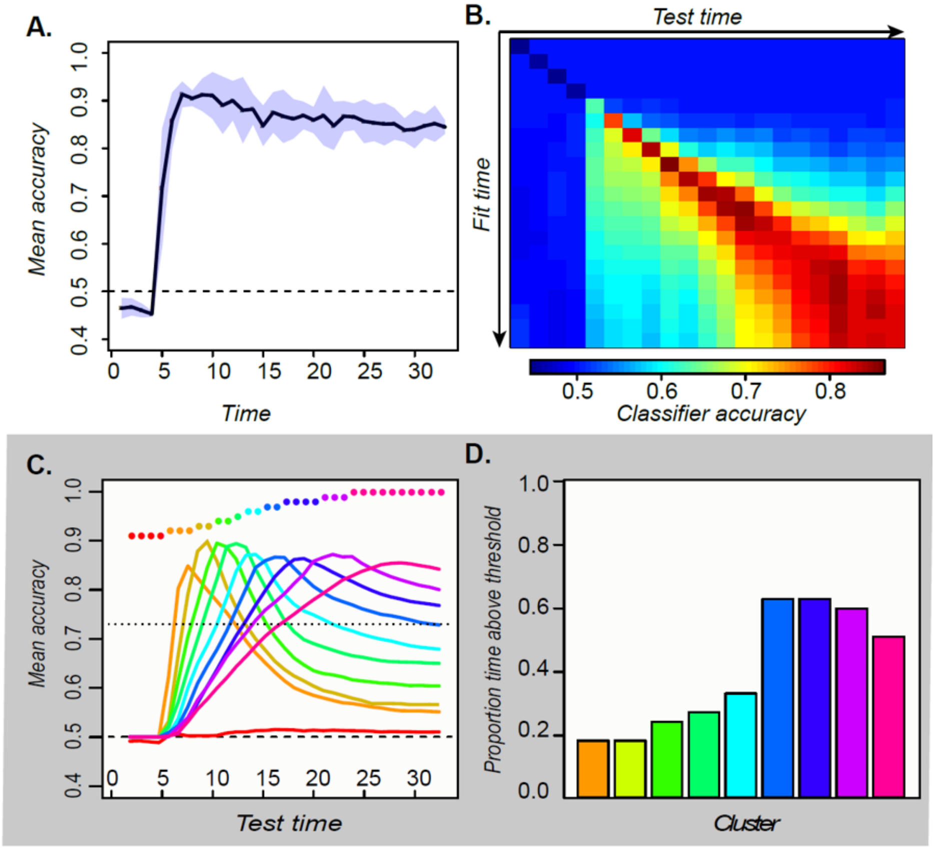

Figure 3. A. Mean and 95% confidence interval of the hold-out accuracy for classifiers trained at each tick of time in the model. B. Accuracy for each classifier (rows) tested at each point in time (columns). C. Mean accuracy for each cluster of classifiers at every point in time. Colored dots show the 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[28], 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 indicates how the underlying neural code persists or changes over time. Classifiers fitted to earlier activation patterns generalized only to immediate temporal neighbors, while those fitted to later patterns generalized over a wider window but failed at decoding earlier states (Figure

3B). To better visualize these results, we clustered the rows of the matrix in Figure 3B and plotted the mean accuracy of the classifiers in each cluster across time (Figure 3C). The results exhibit an “overlapping waves” pattern: classifiers that work on early patterns quickly fail at later timepoints where a different classifier succeeds. As time progresses, the clusters include more classifiers and the breadth of time over which the classifiers perform well widens (Figure 3D).

This pattern reflects the nonlinear trajectories apparent in the sparsely-sampled representational space. When trajectories curve, earlier classification planes fail later in time while later planes fail at earlier time-points (Figure 1A). If representations simply moved linearly from an initial to a final state, early classifiers would continue to perform well throughout processing—a pattern observed in further simulations with feature-based models, models with distributed representations that evolve linearly, and recurrent but shallow neural networks (see Supplementary Information). In the deep network, the non-linear dynamic pattern was observed only in the hub layer—in more superficial layers, the code remained stable (Supplementary Information). The simulations thus suggest that distributed and dynamic semantic representations can arise in deep layers of interactive networks and will elicit a particular “signature” when multivariate pattern classifiers are used to decode semantic structure from ECoG data. Specifically, such classifiers will show:

(1) *Constant decodability*. Neural activity predicts stimulus category at every time point once activation reaches the vATL.

(2) *Local temporal generalization.* Classifiers generalize best to immediate temporal neighbors and worst to distal timepoints.

(3) *Widening generalization window*. The temporal window over which classifiers generalize grows wider over time.

(4) *Change in code direction.* The independent correlation between neural activity at an individual site and semantic category can change direction over the course of a single trial.

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

ECoG study. The dataset included local field potentials (LFPs) collected at 1000Hz from 16-24 electrodes situated in the left ventral anterior temporal cortex of 8 patients awaiting surgery while they named line-drawings of common animate and inanimate items matched for a range of confounds (see Methods). We analyzed LFPs over the 1640ms following stimulus onset using a 50ms sliding-window approach in which separate classifiers were fitted for each window and the window advanced in 10ms increments. The approach yielded 160 classifiers per subject, each decoding the 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).

Hold-out accuracy exceeded chance at about 200ms post stimulus onset and remained statistically reliable throughout the time window (Figure 4A). By 200ms classifiers generalized well to timepoints near the training window but poorly to more distal timepoints, with the generalization envelope widening as time progressed (4B). We again clustered the classifiers based on their temporal accuracy profile, then plotted mean profiles for each cluster (4C). The result was an “overlapping waves” pattern strikingly similar to the simulation: classifiers that performed well early in processing quickly declined in accuracy, replaced by a different well-performing set. Over time neighboring classifiers began to show similar temporal profiles, forming larger clusters that performed above chance for a broader temporal window (4D).

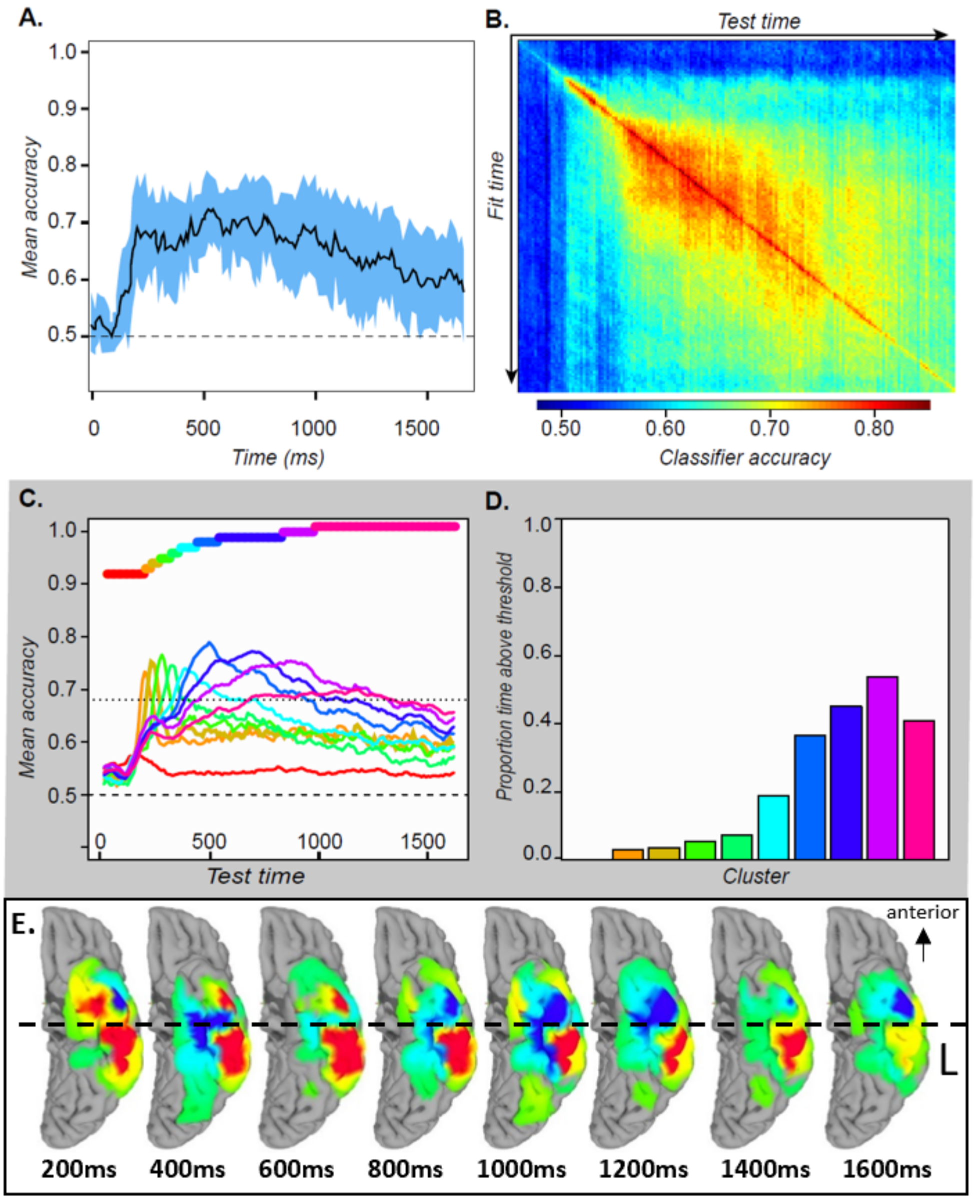

Figure 4. A. Mean and 95% confidence interval of the hold-out accuracy for classifiers trained at each 50ms time window of ECoG data. B. Mean accuracy across participants for each classifier (rows) tested at each timepoint (columns) in the ECoG data. C. Mean accuracy for each cluster of classifiers at 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.

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

Discussion.

We have presented evidence from computational modeling, multivariate pattern classification, and ECoG suggesting that semantic representations in human vATL are distributed and dynamic: neural populations throughout ventral temporal lobe jointly express semantic information from early in processing, but the distributed code changes over time, especially anteriorly. In simulation we showed that dynamic representational change arises in the deep layers of an interactive network, 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. These dynamics produced puzzling results when unit activations were analyzed independently, with some units behaving like tonic feature detectors, some like transient detectors, and some "flipping" the direction of their category preferences over time. 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 semantic representations in vATL are deep, distributed and dynamic.

Are the results also consistent with feature-based theories? They certainly rule out the simple view that stimulus perception drives tonic activation of feature detectors down the ventral visual stream. Were this the case, classifiers that perform well early on should show continued good performance later (see Supplementary Information). One could, perhaps, posit that feature detectors in different cortical regions become transiently active along different time-courses, with some engaging and disengaging early on and others only activating later. Under this view, each "wave" in Figure 3C could reflect the temporary activation of detectors in different cortical regions. If so the classifier weight maps should highlight different brain regions at different timepoints, but we found non-zero weights distributed across the entire field of view from the first moment that classification succeeds—a spatial distribution that changed hardly at all over time. What did change was the *direction* of the vATL semantic code: the "meaning" of a positive LFP flipped direction over time, contrasting with the stable positive-to-negative gradient observed more posteriorly. Several fMRI studies have reported a similar lateral-to-medial category-specific pattern in posterior fusiform [29,30], lending external validity to our analysis and suggesting in turn that the shifting vATL pattern is not artifactual.

The dynamic semantic code in vATL also resolves a long-standing puzzle. Convergent methods have established the centrality of this region for semantic memory, including studies of semantic impairment[31–33], lesion-symptom mapping[34], functional[35,36] and structural[37,38] brain imaging, and transcranial magnetic stimulation[39]. Yet multivariate approaches to discovering neuro-semantic representations almost never identify the vATL, instead revealing semantic structure more posteriorly[40–42]. One prominent study suggested that semantic representations may tile the entire cortex *except* for the vATL[5]. Setting aside significant technical challenges of successful neuroimaging of this region[36], 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 truly are distributed and dynamic, but will preserve signal in posterior regions where the code is more stable. Thus the widespread null result may arise precisely because semantic representations in vATL are distributed and dynamic.

Prior studies applying the temporal generalization method to MEG data in visual semantic tasks uniformly report a very narrow and unchanging band of temporal generalization[43,44]. This pattern is consistent with the general view that the neuro-semantic code changes rapidly over the course of stimulus processing[23]—a phenomenon difficult to reconcile with static feature-based models. Our results differ from the MEG pattern, and indeed from most other work applying the temporal generalization approach[28], in showing a gradual widening of the temporal generalization window. The simulation results explain why the pattern arises and why it is observed in ventral temporal cortex: hub representations in vATL change rapidly early on due to interactions with modality-specific representations throughout cortex, but these changes slow as the full activation pattern emerges across network components.

Why should a dynamic code arise in the vATL? The area is situated at the top of the ventral visual stream, but also connects directly to core language areas[45] and, via middle temporal gyrus, to parietal areas involved in object-directed action[26]. It receives direct input from smell and taste cortices[46], and is intimately connected with limbic structures involved in emotion, memory, and social cognition[47]. 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[38], 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 perception 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.

References

1. Rogers, T. T. & McClelland, J. L. *Semantic Cognition: A Parallel Distributed Processing Approach*. (MIT Press, 2004).
2. McClelland, J. L. & Rogers, T. T. The Parallel Distributed Processing approach to semantic cognition. *Nat. Rev. Neurosci.* **4**, 310–322 (2003).
3. Martin, A. 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 (2016).
4. Patterson, K., Nestor, P. J. & Rogers, T. T. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat. Rev. Neurosci.* **8**, (2007).
5. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
6. Patterson, K., Nestor, P. J. & Rogers, T. T. Where do you know what you know? The representation of semantic knowledge in the human brain. *Nat. Rev. Neurosci.* **8**, 976–987 (2007).
7. Ralph, M. A. L., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* **18**, 42–55 (2017).
8. Serre, T., Oliva, A. & Poggio, T. A feedforward architecture accounts for rapid categorization. *Proc. Natl. Acad. Sci.* **104**, 6424–6429 (2007).

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

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

11. Kriegeskorte, N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu. Rev. Vis. Sci.* **1**, 417–446 (2015).

12. Ralph, M. A. L., Jefferies, E., Patterson, K. & Rogers, T. T. The neural and computational bases of semantic cognition. *Nat. Rev. Neurosci.* **18**, (2016).

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

14. Farah, M. J. & McClelland, J. L. A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *J. Exp. Psychol. Gen.* **120**, 339–357 (1991).

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

16. Rogers, T. T. *et al.* The structure and deterioration of semantic memory: a computational and neuropsychological investigation. *Psychol. Rev.* **111**, 205–235 (2004).

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

18. Chen, L., Lambon Ralph, M. A. & Rogers, T. T. A unified model of human semantic knowledge and its disorders. *Nat. Hum. Behav.* **1**, (2017).

19. Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G. & Mishkin, M. The ventral

visual pathway: an expanded neural framework for the processing of object quality. *Trends Cogn. Sci.* **17**, 26–49 (2013).

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

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

22. Cox, C. R., Seidenberg, M. S. & Rogers, T. T. Connecting functional brain imaging and Parallel Distributed Processing. *Lang. Cogn. Neurosci.* **30**, (2015).

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

24. Grootswagers, T., Wardle, S. G. & Carlson, T. A. Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *J. Cogn. Neurosci.* **29**, 677–697 (2017).

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

26. Chen, L., Lambon Ralph, M. A. & Rogers, T. T. A unified model of human semantic knowledge and its disorders. *Nat. Hum. Behav.* **1**, (2017).

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

28. King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. (2014). doi:10.1016/j.tics.2014.01.002

29. Martin, A. & Chao, L. L. Semantic memory in the brain: Structure and processes. *Curr. Opin. Neurobiol.* **11**, 194–201 (2001).

30. Anzellotti, S., Mahon, B. Z., Schwarzbach, J. & Caramazza, A. Differential activity for animals and manipulable objects in the anterior temporal lobes. *J. Cogn. Neurosci.* **23**, 2059–67 (2011).

31. Snowden, J. S., Goulding, P. J. & Neary, D. Semantic dementia: A form of circumscribed temporal atrophy. *Behav. Neurol.* **2**, 167–182 (1989).

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

33. Lambon Ralph, M. A., Ehsan, S., Baker, G. A. & Rogers, T. T. Semantic memory is impaired in patients with unilateral anterior temporal lobe resection for temporal lobe epilepsy. *Brain* **135**, (2012).

34. Acosta-Cabronero, J. *et al.* Atrophy, hypometabolism and white matter abnormalities in semantic dementia tell a coherent story. *Brain* **134**, 2025–35 (2011).

35. Rogers, T. T. *et al.* The anterior temporal cortex and semantic memory: Reconciling findings from neuropsychology and functional imaging. *Cogn. Affect. Behav. Neurosci.* **6**, 201–213 (2006).

36. Visser, M., Jefferies, E. & Lambon Ralph, M. A. Semantic processing in the anterior temporal lobes: A meta-analysis of the functional neuroimaging literature. (2010).

37. Binney, R. J., Embleton, K. V, Jefferies, E., Parker, G. J. M. & Ralph, M. A. L. 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**, 2728–38 (2010).

38. Binney, R. J., Parker, G. J. M. & Lambon Ralph, M. A. 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 (2012).

39. Pobric, G., Jefferies, E. & Lambon Ralph, M. A. Anterior temporal lobes mediate semantic representation: mimicking semantic dementia by using rTMS in normal participants. *Proc. Natl. Acad. Sci. U. S. A.* **104**, 20137–41 (2007).

40. Bruffaerts, R. *et al.* Similarity of fMRI activity patterns in left perirhinal cortex reflects semantic similarity between words. *J. Neurosci.* **33**, 18597–607 (2013).

41. Devereux, B. J., Clarke, A., Marouchos, A. & Tyler, L. K. Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *J. Neurosci.* **33**, 18906–16 (2013).

42. Sha, L. *et al.* The Animacy Continuum in the Human Ventral Vision Pathway. *J. Cogn. Neurosci.* **27**, 665–678 (2015).

43. Carlson, T., Tovar, D. A., Alink, A. & Kriegeskorte, N. Representational dynamics of object vision: The first 1000 ms. *J. Vis.* **13**, 1–1 (2013).

44. Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and time. *Nat. Neurosci.* **17**, 455–462 (2014).

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

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

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

48. Rumelhart, D. E., McClelland, J. L. & the PDP Research Group. *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume I: Foundations & Volume II: Psychological and Biological Models*. (MIT Press, 1986).

49. Lambon Ralph, M. A., Lowe, C. & Rogers, T. T. Neural basis of category-specific

semantic deficits for living things: Evidence from semantic dementia, HSVE and a neural network model. *Brain* **130**, (2007).

50. Chen, Y. *et al.* The 'when' and 'where' of semantic coding in the anterior temporal lobe: Temporal representational similarity analysis of electrocorticogram data. *Cortex* (2016). doi:10.1016/j.cortex.2016.02.015
51. Barry, C., Morrison, C. M. & Ellis, A. W. Naming the Snodgrass and Vanderwart Pictures: Effects of Age of Acquisition, Frequency, and Name Agreement. *Q. J. Exp. Psychol. Sect. A* **50**, 560–585 (1997).
52. Snodgrass, J. G. & Vanderwart, M. A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. *J. Exp. Psychol. Learn. Mem. \& Cogn.* **6**, 174–215 (1980).
53. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288 (1996).
54. Matsumoto, R. *et al.* Subregions of human MT complex revealed by comparative MEG and direct electrocorticographic recordings. *Clin. Neurophysiol.* **115**, 2056–2065 (2004).
55. Matsumoto, R. *et al.* Left anterior temporal cortex actively engages in speech perception: A direct cortical stimulation study. *Neuropsychologia* **49**, 1350–1354 (2011).
56. Srivastava, R. K., Greff, K. & Schmidhuber, J. Training Very Deep Networks. 2377–2385 (2015).

Acknowledgements.

Funding: This work was partially supported by Medical Research Council Program grant MR/J004146/1 and by European Research Council grant GAP: 670428 - BRAIN2MIND_NEUROCOMP, and by the Grant-in-Aid for Scientific Research of the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) KAKENHI (Grant numbers:15H05874, 18H02709, 18K19514, and 19K17033). The Department of Epilepsy, Movement Disorders, and Physiology, Kyoto University Graduate School of Medicine is an endowment department, supported with grants by Eisai Co.Ltd, NIHON KOHDEN CORPORATION, Otsuka Pharmaceutical Co., and UCB Japan Co., Ltd.

Author contributions: Rogers conducted all simulations, contributed to ECoG data analyses, created figures, and wrote the initial draft. Cox and Lu contributed to the data analysis, figures, and paper revision. Shimotake and Matsumoto recorded the ECoG data during the naming task. Ikeda and Takahashi oversaw patient recruitment and neurological assessment. Kikuchi, Kunieda and Miyamoto arranged and managed the implantation the subdural electrodes. Lambon Ralph contributed to the design of the study and data collection, development of the model and analyses, and writing the manuscript.

Competing interests: Authors declare no competing interests.

Data and materials availability: All data is available via links in the main text or the Supplementary Information.

Supplementary Information:

Movie S1

Figures ED1-ED3

Table ED1

Methods

Model implementation details.

Model structure. The model implements the “distributed-plus-hub” theory of semantic representation developed in prior work to understand patterns of semantic impairment in patients with acquired neuropathology and patterns of functional activation observed in healthy participants during semantic task performance[6,7,16,18]. It is a deep, fully continuous and recurrent neural network[48] that learns associations among visual representations of objects, their names, and verbal descriptors, via a central cross-modal hub. All simulations were conducted using a variant of the open-source Light Efficient Network Simulator adapted to work with contemporary libraries and available at https://github.com/crcox/lens. Code for replicating the simulations and the data reported in this paper are available at https://github.com/ttrogers.

The model included a set of 98 *Visual* units conceived as encoding information about an object’s visual appearance, and a set of 102 *Verbal* units conceived as encoding object names and other verbal descriptors (“is furry”, “can fly”, etc). Visual perception was simulated by providing positive input to a subset of visual units, while perception of a word or phrase was simulated by providing positive input to one or more verbal units. The visual units were reciprocally connected to a *visual hidden layer*, while the verbal units connected reciprocally to a *verbal hidden* layer, both containing 25 units. The two hidden layers then connected reciprocally to a central “hub” layer, also containing 25 units. Units within each hidden layer were reciprocally connected to one another.

Processing visual and verbal inputs. 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 so that they adopted a low activation state in the absence of 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 then propagated through visual hidden, hub, and verbal hidden units to eventually alter activation states in the verbal units themselves. 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[16,18,49], 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 Verbals 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 model was tested by presenting visual input for each of the 90 items in its environment and recording the resulting activations in the 25 hub units at each update as the model settled over the course of 32 updates (producing 33 time-points including the initial state). These activations were then distorted with uniform noise sampled from -0.005 to 0.005 to simulate measurement error in ECog.

Analysis. All analyses were conducted using R version 3.6. To visualize the trajectory of hub representations through unit activation space, we computed a simultaneous 3-component multidimensional scaling of the unit activation patterns for all 90 items at all 33 timepoints. Pairwise Euclidean distances amongst all 2970 vectors were computed and subjected to a classical multi-dimensional scaling algorithm using the native R function cmdscale to extract three latent dimensions. The resulting coordinates for a given item at each point in time over the course of settling were plotted as lines in a 3D space using the scatterplot3d package in R. Figure 2B shows the result for one network training run. Figure 2C shows the same trajectories in the raw data (ie actual unit activation states rather than latent dimensions in a MDS) for randomly-sampled pairs of hub units.

To simulate decoding of ECog data, we evaluated the ability of pattern decoders (ie binary classifiers) to determine the correct superordinate category from patterns of activity arising in the hub for each item at each timepoint. As explained in the main text, we assume that ECoG measures only a small proportion of all the neural populations that encode semantic information. We therefore sub-sampled the hub-unit activation patterns by selecting 3 units at random from the 25 hub units and using their activations to provide input to the decoder. We fitted three decoders to discriminate, respectively, animals from objects, animals from plants, and plants from objects. The decoders were fitted with logistic regression 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 the classifier. We assessed decoder accuracy at the time-point where it was fitted using 90-fold 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. Thus for each trained neural network, decoding accuracy at a single timepoint reflected the mean accuracy across 30 decoders: animal vs object, animal vs plant, and object vs plant, each fitted to 10 independent sub-samples of 3 hub units. 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 accuracy higher than the others at some point in time. Similar results were obtained for all cluster-sizes examined, however.

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

<u>ECoG methods and materials</u>

Participants. Eight patients with intractable partial epilepsy (seven) or brain tumor (one) originating in the left hemisphere participated in this study. These include all left-hemisphere cases described in a previous study[50], and we will use the same case numbers reported in that work (specifically cases 1-5, 7, and 9-10). Background clinical information about each patient is summarized in Table S1. Subdural electrode implantation was performed in the left hemisphere for presurgical evaluation (mean 83 electrodes, range 56-107 electrodes/patient). 16-24 electrodes (mean 20 electrodes) covered the ventral ATL in each patient. The subdural electrodes were constructed of platinum with an inter-electrode distance of 1 cm and recording diameter of 2.3 mm (ADTECH, WI). ECoG recording with subdural electrodes revealed that all epilepsy patients had seizure onset zone outside the anterior fusiform region, except one patient for whom it was not possible to localize the core seizure onset region. The study was approved by the ethics committee of the Kyoto University Graduate School of Medicine (No. C533). Participants all gave written information consent to participate in the study.

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

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

Data preprocessing. Data preprocessing was performed in MATLAB. Raw data were recorded at sampling rate of 1000 Hz for six patients and at 2000Hz for two patients. The higher sampling rates for the two patients were down-sampled to 1000Hz by averaging measurements from each successive pair of time-points. The raw data from the target subdural electrodes for the subsequent analysis were measured in reference to the electrode beneath the galea aponeurotica in 4 patients (Patients 4,5,7 and 10) and to the scalp electrode on the mastoid process contralateral to the side of electrode implantation in 4 patients (Patients 1-3 and 9). Data included, for each stimulus at each electrode, all measurements beginning at stimulus onset and continuing for 1640ms. Baseline correction was performed by subtracting the mean pre-stimulus baseline amplitude (200 ms before picture onset) from all data points in the epochs. Trials with greater than +/-500 μV maximum amplitude were rejected as artifacts. Visual inspection of all raw trials was conducted to reject any further trials contaminated by artifacts, including canonical interictal epileptiform discharges. The mean waveform for each stimulus was computed across repetitions.

Multivariate classification analysis.

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

The classifiers were logistic regression models fitted with L1 regularization[53] using the glmnet function in Matlab. L1-regularization applies an optimization penalty that scales with the sum of the absolute value of the classifier coefficients and thereby encourages solutions in which many features receive coefficients of 0. This approach is useful for a sliding-window analysis because features receiving a 0 coefficient in the classifier have no impact on its performance when 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 decoding accuracy over time across the classifiers within each cluster. All cluster sizes produced the overlapping-waves pattern. In the main paper we show the 10-cluster solution as it is the largest number in which each cluster after the first has a mean accuracy profile that is both statistically reliable and higher than every other cluster at some point in time.

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

Visualizing solutions on surface plots.

Structural brain imaging and electrode localization. Magnetization-prepared rapid gradient-echo (MPRAGE) volumetric scan was performed before and after implantation of subdural electrodes as a part of presurgical evaluations. In the volumetric scan taken after implantation, the location of each electrode was identified on the 2D slices using its signal void due to the property of platinum alloy[54]. Electrodes were non-linearly co-registered to the patient MRI (MPRAGE) taken before implantation, and then to MNI standard space (ICBM-152) using FNIRT (www.fmrib.ox.ac.uk/fsl/fnirt/). The native coordinates of all the electrodes for all patients were morphed into MNI space and resampled into 2 mm isotropic voxels[55].

Projecting classifier coefficients to the surface. As described above, a separate logistic classifier was fitted to each 50ms window in each subject. The classifier was specified as a set of regression coefficients, with one coefficient for each timepoint at each electrode in the patient, and many coefficients set to 0 due to L1-regularization. The sign of the classifier coefficient indicates the "meaning" of a LFP deflection in a particular direction: a positive coefficient indicates that animals are "signaled" by a positive deflection in the LFP, while negative coefficients indicate that animals are signaled by a negative deflection. The magnitude of the coefficient indicates the "importance" of the measurement, in the context of all other LFPs appearing in the classifier. The distribution of coefficient directions and magnitudes across the cortex and over time thus provides an indication of how the underlying neuro-semantic code changes over time. We therefore analyzed the temporal and spatial distribution of mean coefficients across participants as follows.

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

To visualize how the representational code changes over time within the surface mask we next carried out a similar procedure on the classifier coefficients themselves, without taking the 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.

Supplementary Information

Comparison of deep model results to control models.

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

Distributed versus feature-based representations.

By *distributed representation*, we mean that many neural populations or *units* can jointly contribute to representation of structure even if they do not each independently encode the structure. Deep neural network models are capable of acquiring distributed representations of this kind [22] and may be contrasted with models proposing that semantic representations are comprised of elements that each independently detect a particular semantic feature, such as membership in a particular conceptual domain or category. We therefore considered what the decoding signature would look like in such a *feature-based* model. The 90 items were represented with a vector in which two elements were dedicated to each conceptual domain (animal, plant, 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[56]. 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.

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

Results.

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

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

Local temporal generalization. Figure ED-1A shows a 3D MDS of the trajectories for all items through the corresponding representation space in each model. For feature-based, linear, and shallow models, the trajectories are strictly or nearly linear—only the deep, distributed and dynamic model shows the nonlinearities discussed in the main paper. Consequently the models show qualitatively different patterns of generalization over time: feature-based, linear, and shallow models show a pattern in which all classifiers generalize poorly to earlier timepoints and 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 (ED-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.

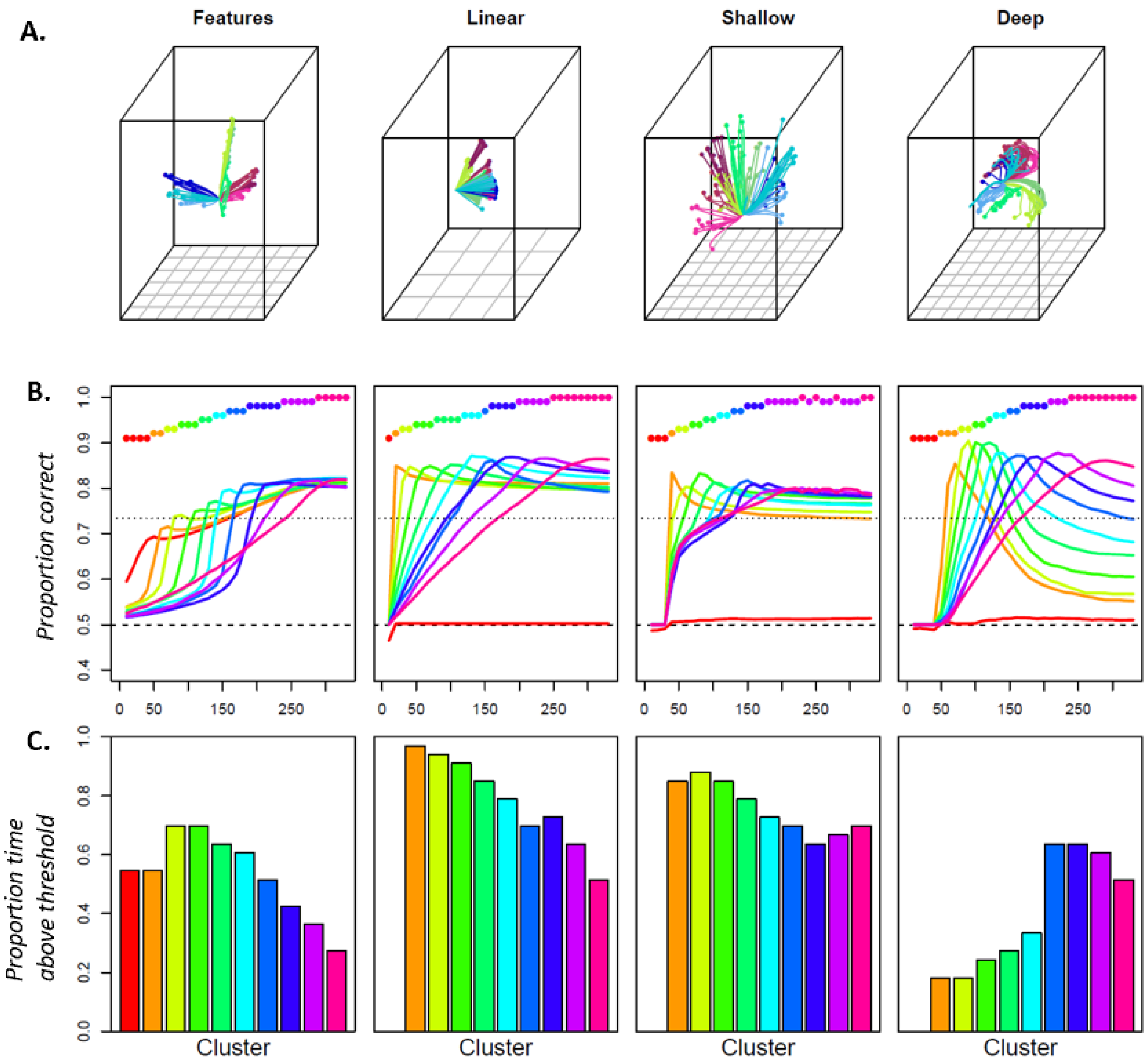

Figure ED-1. Comparison of simulation results for a feature-based model, a distributed linear model, a shallow recurrent network, and the deep, distributed and dynamic model. A. Multi-dimensional scaling showing the trajectory of each item through representation space under four different models. Only 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 also observed in the more anterior parts of the ventral temporal lobe, via the changing sign of the classifier coefficients identified in the ECoG data. We therefore considered whether a change in code direction was observed for single units considered independently in the alternative models.

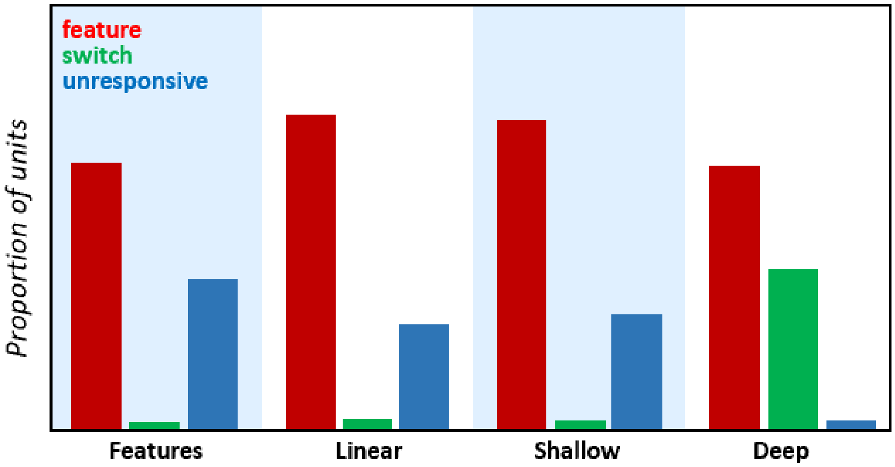

Figure ED-2. For each model type, the proportion of units that behave like feature-detectors (red), detectors that switch their category preference over time (green), and units that seem unresponsive to the semantic category (blue). Only the deep, distributed, dynamic model has units whose responses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

Specifically, we classified each unit in each simulation as (1) a *feature-detector* if its activity correlated significantly with conceptual domain in only one direction over the time-course of processing, (2) as a *switch feature* if it correlated significantly in both the positive and the negative direction at different points in time, and as (3) *non-responsive* if it never correlated significantly with conceptual domain. Across all five simulations and all three classification tasks, we computed the proportion of units falling into each category for each model type. The results are shown in Figure ED-2. For feature-based and linear models, the results are trivial, since the representations are constrained to show only non-responses or feature-like behaviors, as was indeed observed. 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.

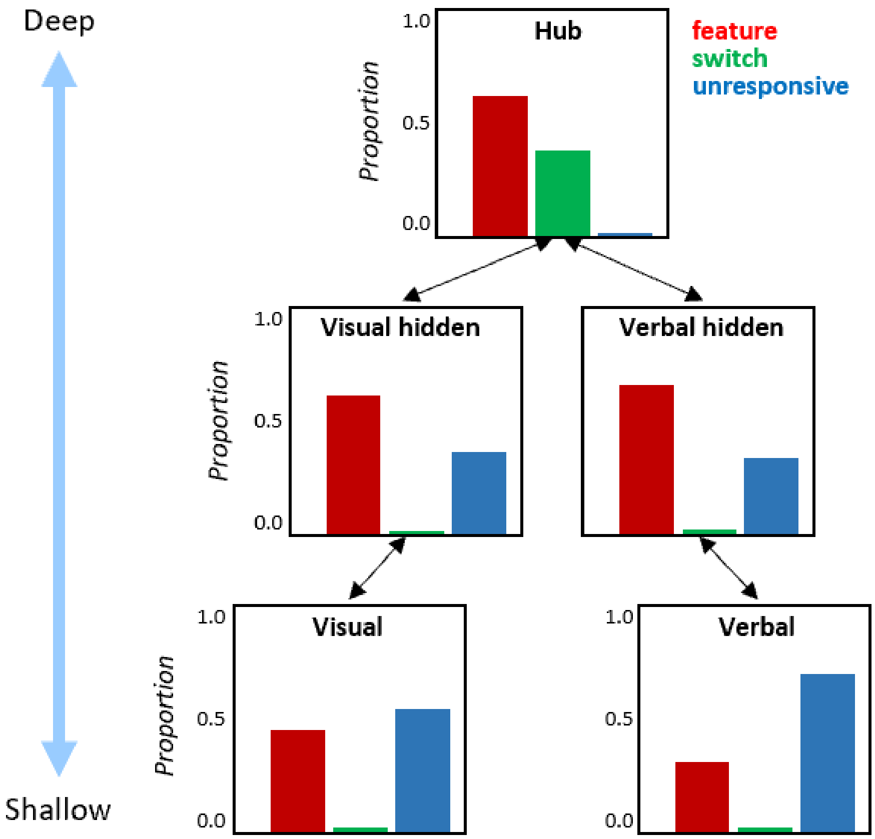

Figure ED-3. For each layer in the deep, distributed and dynamic model, the proportion of units that behave like feature-detectors (red), detectors that switch their category preference over time (green), and units that seem unresponsive to the semantic category (blue) when the model processes visual 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.

Finally, the 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 therefore performed the same analysis separately on each layer of the deep, distributed, and dynamic model, counting the proportion of units in each layer that behaved like feature-detectors, switch-features, or non-responsive units, as the network processed all 90 visual input patterns. Results are shown in Figure ED-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 diagnoistic 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. Instead the pattern must reflect the centrality of the hub units in computing interactive mappings between visual and verbal representations.

Summary.

These control 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.