TESTING NEUROCOGNITIVE PREDICTIONS OF THE HUB-AND-SPOKE MODEL OF SEMANTIC MEMORY WITH NETWORK REPRESENTATIONAL SIMILARITY ANALYSIS

by

Christopher R. Cox

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Psychology)

at the

UNIVERSITY OF WISCONSIN–MADISON

2016

Date of final oral examination: 5 December, 2016

The dissertation is approved by the following members of the Final Oral Committee:
Timothy T. Rogers, Professor, Psychology
Mark S. Seidenberg, Professor, Psychology
Bradley Postle, Professor, Psychology
Robert Nowak, Professor, Electrical and Computer Engineering
Matthew A. Lambon Ralph, Professor, Psychological Science

Dedicated to my my wife and partner, my parents, and my brother.

Acknowledgments

I am filled with gratitude and humbled by the company I was able to keep and the experiences I was afforded at the University of Wisconsin-Madison. Foremost in my mind are Tim Rogers and Mark Seidenberg who have invested so much time, energy, and resources into my training. They helped me to set ambitious goals, and were patient with the inevitable failures, mistakes, disappointments, and setbacks that are both the debris of ambitious science and fertile soil of discovery. But they also never failed to point out the victories and forward progress. Their passion and excitement is an inspiration.

Mark's ability to see three steps ahead in a research program has granted me a beautiful parting gift of experimental ideas that I will only begin pursuing in my post-doctoral work… ideas posed circa 2011, when I was in my second year, and long before the means for addressing them—Network RSA—was developed.

Tim's penchant for taking big questions head on will always be evidence to me that science is not about publication counts, milking a paradigm to death, or taking a position and holding it. His spirit of "but this is just so interesting, how can you not want to know the answer?" can be distilled down to a motto that, if Tim never literally said it, he implicitly expresses in his attitude and demeanor: do good work that you value and are proud—*dying*—to tell your friends about.

I have a special thank you for Brad, who evaluated my and many of my first-year cohort's first graduate-level writing assignment for his "Memory and the

Brain" seminar. With a stroke of his red pen, Brad humbled and dumbstruck us. When our private horror and shame gave way to openness and consolation, it catalyzed the trust among my cohort, which would crystallize into friendships that I continue to share the best and worst times with, notably including Chris Potter, Callan Cooper, Brianna McMillan, and Libbie Brey. Brad's apparent severity quickly revealed itself to be the most genuine kind of caring, and his support and thoughtful guidance is made all the richer in light of his honesty.

This spirit of compassionate honesty is also a defining quality of the LCNL. Everyone knows that if you can present for Mark and Maryellen in room 617, you can be comfortable in front of any audience. This is the Wisconsin spirit, and I hope that I can bring that with me to whatever departments I am a part of in the future.

Beyond our department, I have had the great privilege to collaborate with Rob Nowak and his students, specifically Nikhil Rao and Urvashi Oswal. I entered those collaborations as a C student in calculus and a novice, self-taught programmer and was cultivated into someone with the technical expertise to complete this thesis. But Rob has also set an example for living a balanced life, and he and Tim have certainly demonstrated that some of the most important discussions can happen over a beer, like at HAMLET… but ideally on the terrace.

Looking forward to my post-doctoral life and beyond, I am optimistic for what I will be able to contribute to the field. This is largely due to the fantastic support I have received, and the investments people have always and continue to make in my forward progress. I have Matt Lambon Ralph to thank for my next crop of opportunities and for his outstanding support and insight during this dissertation process.

Circling back again on the department community, I grew so much and had a wonderful time in the process through my lab mates, past, present, and honorary.

In particular I would like to acknowledge Jon Willits, Jessica Montag, Chen Lang, April Murphy, Clint Jenson, and Pierce Edmiston. Thank you for your companionship and sharing you passion for science.

Finally, I would like to use this small platform to express my thanks and appreciation for Eileen Haebig. She has been a patient and steadfast companion, who tirelessly engaged with my obsessive deliberations over theory and deftly handled my anxiety over analyses and interpretations, while deliberately holding me to routine when the thesis threatened to be all-consuming. Each day, she motivated me to set goals, and appreciate incremental progress. And we never compromised on pancake Saturday's, or listening to the latest Serial release, or generally taking time each day just to be together. Her love and encouragement made a mark on every page of this document.

Contents

Contents **v**

1 Introduction **1**

1.1 How the ATL may contribute to semantic representation: Three hypotheses 4

1.2 Relation to prior work 12

1.3 Identifying sparse, distributed, whole brain patterns of functional activity 14

1.4 Cross-modal representations in the ATL 15

1.5 Summary 16

2 Distributed Representations **18**

2.1 Independence versus interdependence of representational elements 26

2.2 Graded versus discrete contributions to representation 30

2.3 Heterogeneity versus homogeneity of representation 33

2.4 Heterogeneity versus homogeneity of location 36

2.5 Conclusions 40

3 Comparing methods **41**

3.1 A brief overview of PDP models and their representational assumptions 46

3.2 Challenges for the discovery of PDP representational structure in the brain . . . 50
3.3 Model and Methods . . . 55
3.4 Simulation details . . . 61
3.5 Results . . . 63
3.6 Discussion . . . 86

4 Network Representational Similarity Analysis 94
4.1 Intuitions . . . 100
4.2 Optimization . . . 104

5 fMRI Experiments 108
5.1 Materials and methods . . . 112
5.2 Analysis methods . . . 118
5.3 Results . . . 124
5.4 Discussion . . . 146

6 ECoG Experiments 151
6.1 Materials and methods . . . 153
6.2 Results . . . 156
6.3 Discussion . . . 163

7 General Discussion 165
7.1 Estimating semantic structure . . . 170
7.2 Basis sets for meaning? . . . 172
7.3 Conclusion . . . 175

Bibliography 178

A Network RSA selection maps for individual subjects 199

List of Figures **202**

List of Tables **207**

Chapter 1

Introduction

The role of the anterior temporal lobe (ATL) in semantic cognition remains a controversial topic in cognitive neuroscience. It has been attributed with several specific kinds of semantic processing, such as social cognition (Simmons, Reddish, Bellgowan, & Martin, 2009), the representation of unique entities (Grabowski et al., 2001), and face recognition (Gainotti, 2007). On the other hand, the ATL is thought to play the critical "hub" role in hub-and-spoke theory of semantic cognition, as a part of the brain that participates in the representation of concepts of all kinds (Patterson & Lambon Ralph, 2016; Patterson, Nestor, & Rogers, 2007; Rogers et al., 2004).

To date, the ATL's hub status has been inferred from its anatomical connectivity (Binney, Parker, & Lambon Ralph, 2012; Morán, Mufson, & Mesulam, 1987), its multi-modal sensitivity (Abel et al., 2015; Shimotake et al., 2014), and its association with domain- and modality-general semantic impairments when damaged by disease (Lambon Ralph, Lowe, & Rogers, 2007; Mayberry, Sage, & Lambon Ralph, 2011; Rogers et al., 2004) or disrupted by trans-cranial magnetic stimulation (TMS; Pobric, Jefferies, and Lambon Ralph, 2007). L. Chen (2014, June 17) recently integrated these various sources of evidence to implement the hub-and-spoke hy-

pothesis as a computation model which succeeds at simulating a wide range of semantic behaviors, including the profiles of all known semantic disorders (Chen et al, unpublished). And although the functional neuroimaging literature used to be rather inconsistent with regard to whether the ATL played a functional role in semantic cognition (Visser, Jefferies, & Lambon Ralph, 2010), recent years have seen a rise in the number of studies showing increased activation of the anterior temporal lobes during semantic tasks. This is due to a variety of improvements in both experimental design and technical aspects of fMRI image acquisition (Binder et al., 2011; Binder, Swanson, Hammeke, & Sabsevitz, 2008; Rogers et al., 2006; Visser, Embleton, Jefferies, Parker, & Lambon Ralph, 2010; Visser, Jefferies, & Lambon Ralph, 2010).

While the general architecture of the model and in particular the location of the hub has now been well established, exactly how the ATL contributes to concept representation is unclear. The set of hypotheses that are consistent with the ATL being an integrative hub can be boiled down to three perspectives. Each perspective predicts that the ATL represents cross-modal structure in a different way, by encoding different kinds of information. These kinds of information are:

1. Points of peak convergence, associated with unique entities (A. R. Damasio, 1989a, 1989b; H. Damasio, Grabowski, Tranel, Hichwa, & Damasio, 1996; H. Damasio, Tranel, Grabowski, Adolphs, & Damasio, 2004; Tranel, Damasio, & Damasio, 1997).

2. Distributed representations containing all semantic content, including re-representations of unimodal and cross-modal structure that might be encoded elsewhere (L. Chen, Lambon Ralph, & Rogers, 2016; Lambon Ralph et al., 2007; Patterson et al., 2007; Rogers et al., 2004).

3. Distributed representations that encode the interactions among spokes, but

not the full semantic similarity structure structure. Unimodal structure is encoded in the spokes. The full semantic similarity structure is expressed in the joint activity over both the hub and the spokes, and not in any single region. This position is not articulated in the literature, but is hinted at by the "graded hub hypothesis" (Binney et al., 2012; Patterson & Lambon Ralph, 2016) which emphasizes the interactivity of the hub and the spokes (rather than feed-forward convergence from spoke to hub) and which more directly acknowledges the structural heterogeneity of the ATL than earlier expositions of the hub-and-spoke hypothesis.

Critically, each of these positions imply different patterns of neural activity over the ATL and the rest of cortex. If semantic knowledge is encoded as distributed representations in a hub-and-spoke network, this poses serious challenges from a neuroimaging perspective. Each concept would be encoded as some pattern of activity over the same neural regions as every other concept. Within these regions, the information would be encoded in fine-grained patterns of activity, rather than consistent activation or deactivation over gross anatomical areas. Furthermore, a distributed representation needs to be considered as a whole. This implies that a pattern which spans multiple anatomical regions needs to somehow be jointly identified, even if separated by expanses of irrelevant neural activity. Finally, the information content of distributed representations is expressed via their similarity to other distributed representations encoded over the same units. This implies that even if two people have encoded exactly the same content over exactly the same units, patterns of activity may still differ.

Until very recently, there did not exist a technique for analyzing neuroimages capable of identifying distributed representations that encode semantic similarity structure over potentially sparse patterns of neural activity that span multiple regions of the brain. It has therefore been difficult to adjudicate between these hy-

potheses with functional neuroimaging. In this dissertation, I will illustrate why contemporary approaches are ill-suited to whole brain multivariate analyses of distributed representations. I will then consider evidence that neural representations can be distributed in the ways just described (Chapter 2) and will articulate newly developed statistical methods better suited to discovering such representations (Chapters 3 and 4). Finally I will apply the methods to two datasets (fMRI in Chapter 5 and ECoG in Chapter 6) in order to adjudicate between the different hypotheses about how semantic similarity structure is encoded in neural activity, and the role of the ATL in that encoding.

Before proceeding, let us consider the three representational hypotheses alluded to above in more detail, with emphasis on the patterns of neural activity that would be associated with each in fMRI or ECoG datasets.

1.1 How the ATL may contribute to semantic representation: Three hypotheses

The convergence hypothesis

The convergence zone hypothesis of conceptual retrieval, developed by Damasio and colleagues (A. R. Damasio, 1989a, 1989b; H. Damasio et al., 2004; Tranel et al., 1997), predicts that conceptual retrieval involves time-locked reactivation of "fragmentary records", information encoded in modality-specific regions. Convergence regions contain "convergence zones", which are amodal nodes that bind together "fragments" of sensory information into coherent concepts. Convergence regions, roughly speaking, exist in a hierarchy where more anterior convergence regions contain the bindings relevant to more specific entities. The most anterior convergence regions, such as the ATL and inferior frontal lobe, bind together fragments

that can reconstruct unique entities (including spatial and temporal relationships), while posterior convergence regions bind together the features composing non-unique entities (A. R. Damasio, 1989a).

The convergence theory makes several clear predictions about the association between neural activity and concept representation. Ironically, the theory does not posit the existence of a convergence region that receives input from all modalities, and thus there is no notion of a pan-modal hub. Rather, the assumption is that time-locked reactivation of information fragments over disparate regions will be integrated by virtue of simultaneous activation, meaning that the convergence zone's primary role is to enforce the right patterns of activity in posterior cortical regions at the right time. Convergence zones do not re-represent information, and so by this account the ATL would not encode any similarity structure. What it does predict is that representational structure would be encoded in a widely distributed pattern over posterior cortical regions, and that there would be no representational structure/content within the hubs enabling semantic retrieval. [1]

I should emphasize that, for the purposes of the current work, I do not intend to pit convergence theory against the hub-and-spoke theory point-by-point. For instance, it is not important that the two theories are at odds with respect to the existence and essential role of a pan-modal hub. What is interesting is that the theories predict very different representation styles. Granting that the ATL is a high-level integration site relevant to conceptual knowledge of all kinds (which, again, is not Damasio's hypothesis), the information encoded in the ATL might be distributed representations or convergence zones. Just as distributed representation is a general means of information encoding not unique to the hub-and-spoke model, the

[1] One can imagine situations where a Damasio-esque representational scheme could result in structured ATL activation over the course of an experiment. If, for example, one was probed about family members and personal items in a way that evoked many unique entities on each trial (or was consistently ambiguous), the activated convergences zones could form a constellation of activity that might look reliably different in the person vs. item contrast. Note, however, that there would be no reason for unique person entities and unique item entities to be clustered in the ATL.

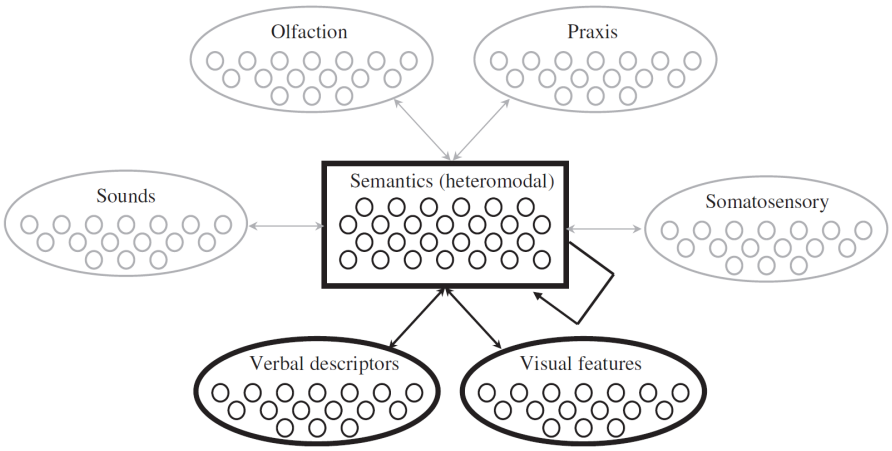

Figure 1.1: Schematic depiction of the hub-and-spoke model of semantic cognition. Reproduced from Lambon Ralph, Lowe, and Rogers (2007).

idea of "time-locked multiregional retroactivation" (A. R. Damasio, 1989b) can be applied in the context of neural architectures other than the specific one hypothesized by Damasio. In short, although devised in part to deal with the alleged absence of a pan-modal hub anywhere in the brain, the core representational hypothesis of the convergence theory is not fundamentally incompatible with such a hub existing.

The semantic hub hypothesis

A major advantage of the hub-and-spoke model is that it has been explicitly implemented as a neural network model that can be used to simulate semantic cognition under a wide range of interesting conditions (L. Chen, 2014, June 17; Pobric et al., 2007; Rogers et al., 2004). It also allows for an investigation of the underlying representational structure that gives rise to the model's behavior. Based on how the implemented model learns—through error correction—and the position of the semantic hub within the model (see Figure 1.1), the model predicts that the hub will come to encode a similarity space that represents all semantic structure. This means that the hub should participate in all semantic tasks regardless of stimu-

lus modality (e.g., reading a word, hearing a sound, viewing an image, feeling a texture, or smelling and odor). The structure expressed by the hub is abstracted away from specific attributes and properties of individual concepts—the hub encodes *similarity structure*, but does not encode specific object properties. That is to say, it is hypothesized that concepts that are similar will be associated with similar patterns of activation, and vice versa, but that a single pattern of activation, taken without the context of activation associated with other concepts, expresses nothing about the content of the concept. This similarity-based encoding scheme is strongly associated with distributed representation and is considered in detail in Chapters 2 and 3. The spokes contain modality specific structure, and is also where such properties would be encoded and retrieved from (Patterson et al., 2007; Rogers et al., 2004).

On this account, there is a degree of redundancy: each spoke represents modality specific structure, and the hub represents pan-modal structure that integrates over all modalities.

The semantic hub+spokes hypothesis

This perspective is not associated with an implemented model in the same sense that the semantic hub perspective is. This perspective posits that the ATL encodes *only* pan-modal structure, and does not reproduce any of the structure encoded in other cross-modal or unimodal association areas. The spokes and the hub encode independent components of semantic space. These disparate components might be integrated by virtue of synchronous activation, similar to the mechanism proposed under Damasio's convergence theory. Critically, and in stark contrast to the convergence theory, the hub+spokes perspective predicts that conceptual structure is encoded in cross-modal regions rather than only being present in unimodal fragments. Whereas the convergence theory requires time-locked reactivation of

sensory information, the hub+spokes perspective predicts simultaneous activation of information at several levels of abstraction. The full concept is encoded in a radically distributed code spanning modality specific and cross modal regions.

The representational predictions of this model are a little more complex. What does it mean for a region to encode *only* cross-modal structure? To develop the basic intuition, we can think about this hypothesis in terms of a linear model that includes interaction terms. Consider the following simple model to describe the relationship between random variables x_1 and x_2 and a 1-dimensional similarity structure, y:

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_{1\cdot2} x_{1\cdot2} \tag{1.1}$$

The similarity structure y can be perfectly predicted if you have several pieces of information—namely, access to the random variables x_1 and x_2, the weights β_1, β_2 and $\beta_{1\cdot2}$, and a function for combining—and the ability to work freely with that information. In the model above, the is no sense of space or division of labor. Consider instead the following:

$$y_a = \beta_1 x_1$$

$$y_b = \beta_2 x_2$$

$$y_c = \beta_{1\cdot2} x_{1\cdot2}$$

Rather than a single model that describes the *global* similarity space, there are three models, and each describes a *local* similarity space. Of course, by substitution:

$$y = y_a + y_b + y_c \tag{1.2}$$

That is, the local spaces can be combined to recover the global similarity structure. This may be made even more clear by visualizing a 2-dimensional example played out in a cartoon hub and spoke network. Imagine there are input features in the audio spoke, and two input features in the visual spoke, and the two spokes connect to the ATL. Each spoke can track the interaction between its two features, but cannot track interactions between features in the other spoke. That is what the ATL is for. For the sake of simplicity, let the ATL represent just the four 2-way interactions among spoke features. Finally, let y be a matrix of 2-dimensional coordinates, meaning that β will be a 10×2 matrix of weights (two for each feature and interaction of features). Assuming perfect knowledge of β, the local similarity spaces are shown in the top row of Figure 1.2. They each have access to different information, and so model only a portion of the structure. The plots show the points in similarity space associated with the β and x that the region has access to. The bottom row, in contrast, shows the global similarity structure. The panel labeled [true] is generated by considering all features and interactions at once. Compare that to the plot labeled "vis+aud+ATL", which is generated by simply summing the coordinates across the local representations displayed in the top row. The structures are the same. This merely allows us to visualize what was shown mathematically above: summing over representations or generating the global representations directly is equivalent.

This simple example also shows how "modality specific" information may play a critical role in pan-modal concept representations. Information received in the visual or audio spoke is unimodal, but because the ATL only codes the cross-modal aspects, they are very literally part of the pan-modal representation—the ATL representation cannot stand on its own, and the conceptual structure exists only in the aggregate local structures.

To be clear, the hub+spokes perspective is not wed to such simple models. The

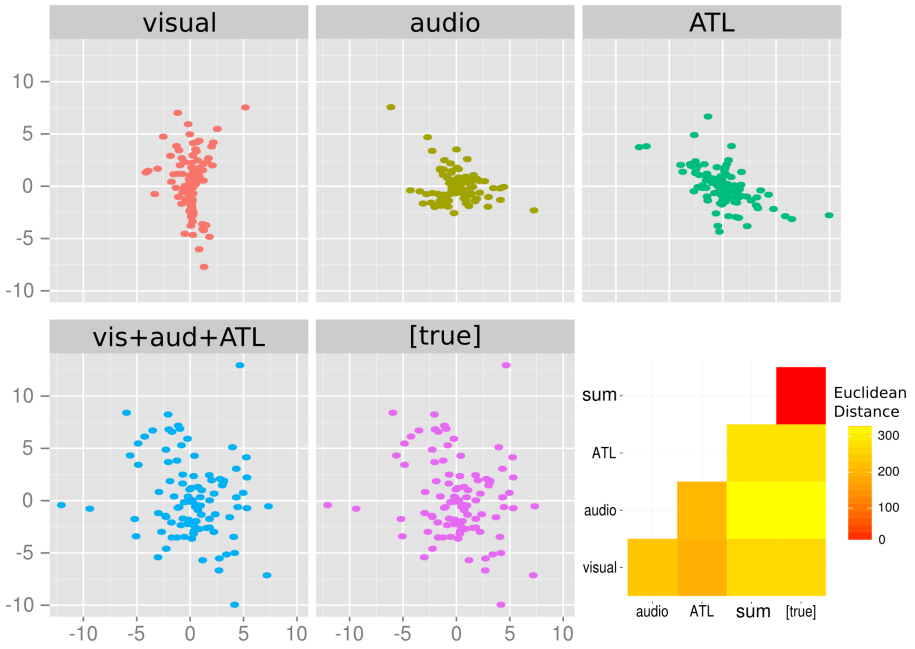

Figure 1.2: An example of how the ATL might encode only pan-modal structure, as under the *hub+spoke* perspective. The data points were generated by a simple linear model of 4 input features, 2 in each spoke (audio and visual) and their interactions (1 local interaction per spoke, and 4 global interactions reflected in the anterior temporal lobe). Given the appropriate weights, when taken all together, the set of features and their interactions express a *global* similarity structure. However, if the features are spread over the two spokes, no one region contains the full structure. Rather, each expresses *local* similarity structure. The insight is that aggregating the local "representations" across regions is mathematically equivalent to generating a single global representation directly (at least in the case of simple linear models).

demonstration above is meant only as a concrete example of how semantic representations may be supported over multiple regions, and to show the purely cross modal representations can exist, serve the proposed function, and are not so abstract as to not be grounded in very simple models.

Summary

These three hypotheses have many aspects in common, making them consistent with the general hub-and-spoke architecture. They all frame concept representation as a complex, experientially grounded process that involves some form of multisensory integration, and they all predict that modality-specific relationships

are encoded near perceptual and motor cortical areas, and the most abstracted relationships are encoded in anterior regions, including the ATL. In addition to being consistent with the hub-and-spoke model, these points are in line with virtually all neuroimaging work on semantics (Barsalou, 2008; Binder, Desai, Graves, & Conant, 2009; Kiefer & Pulvermüller, 2012; A. Martin, 2007; A. Martin & Chao, 2001; Patterson et al., 2007; Visser, Jefferies, & Lambon Ralph, 2010).

These baseline similarities neatly frame their differences, which pertain to how exactly this multisensory integration and abstraction is handled by the brain. Critically, the various hypotheses outlined above predict different patterns of neural activity associated with behaviors that require semantic cognition. In each case, the associated neural activity will take the form of patterns that have the features of distributed representations, including that the similarity among the patterns of activity corresponds to similarity among the mental states. However, the where exactly these patterns of activity can distinguish these hypotheses and thus provided insight into the underlying representations.

At a gross anatomical level, we have the following predictions associated with each of the three hypotheses:

1. *Convergence perspective*. The only neural patterns that correlate with conceptual structure will be in modality specific areas.

2. *Semantic hub perspective*. The neural patterns that correlate with conceptual structure will primarily exist in the ATLs. Neural patterns in modality specific areas may also correlate with semantic structure, but ideal images of the ATLs (with minimal distortion and sufficiently high spatial and temporal resolution) would provide a window into the entirety of the encoded conceptual structure.

3. *Semantic hub+spokes perspective*. The neural patterns that correlate with con-

> ceptual structure would be partially encoded by the ATL and partly encoded in modality specific areas. Studying the neural activity of the ATL in isolation, no matter how precisely, would only yield the pan-modal structural aspects that express the interactions among input from the spokes. Content encoded in unimodal areas is not re-represented in the ATL, and so the hub and spokes need to be included in the same mode in order to account for the relevant representational similarity structure encoded in the brain—indeed, the hub and spokes together support a global semantic space.

In the experimental work reported in Chapter 5, I will conduct a series of novel tests of the representational assumptions of the hub-and-spoke model, in an attempt to clarify the role of the ATL in domain general semantic cognition.

1.2 Relation to prior work

Although prior work has been technically limited to considering localized and/or regional structure in isolation of the rest of cortex, this dissertation does not mark the first attempt to characterize the representational structure expressed by that ATL or how semantic structure is expressed in the functional activity of the brain more generally. Devereux, Clarke, Marouchos, and Tyler (2013) had participants name the basic level category of specific images or names of category exemplars, sampled from six categories, and then performed a whole brain searchlight RSA considering the patterns of activity associated with words and pictures separately. They found that middle temporal gyrus and angular gyrus are areas that express semantic structure across both modalities, and not the anterior temporal lobe.

However, other studies have come to rather different findings. For example, Peelen and Caramazza (2012) presented garage and kitchen tools that are manipulated by either being squeezed or twisted (fully crossed design) in the context of

a one-back task, where participants monitored for either changes in *where* or *how* the tool is used. In contrast to Devereux et al. (2013), both searchlight RSA and a region of interest RSA indicated that activity patterns in ventral ATL carry information about how and where objects are typically used, and that this information was independent of the perceptual properties of the objects.

In another cross-modal study using words and images, Fairhall and Caramazza (2013) had participants rate how typical each stimulus was of its category (five categories total). They found an interesting dissociation between searchlight MVP classification and a region of interest RSA with respect to the ATL: 5-way classification was significantly above chance, but the RSA did not reveal that the semantic structure was significantly correlated with the target structure. This might be because classifiers involve modeling the neural activity, allowing the relative importance of each voxel to be scaled, whereas the RSA does not. If the similarity structure in that ATL is sparsely encoded, for instance, it may account for the dissociation. A cross modal study considering spoken and written words, however, seems to suggest that the left ventral and medial ATL expresses semantic similarity when words are written but not when spoken (Liuzzi et al., 2015). Studies that only used visual stimuli tend to identify posterior temporal regions (Connolly et al., 2012), but it is of course difficult to tell whether these results are reliant on visual or conceptual similarity (Bruffaerts, Dupont, et al., 2013; Jozwik, Kriegeskorte, & Mur, 2016).

The aforementioned research all measured neural responses with fMRI. A recent study of ECoG data collected while 10 patients were presented with line drawings and asked to name each one has revealed that a spatio-temporal searchlight RSA identifies semantic similarity structure in the patterns of local field potentials over electrodes implanted in the ventral ATL (8 out of 10 subjects had left hemisphere implants; Y. Chen et al., 2016).

In summary, there is a small but growing collection of functional evidence indicating that the ATL expresses semantic similarity structure. However, important questions remains. In particular, the distinction between audio and visual stimuli suggested by Liuzzi et al. (2015) bears further investigation, and whether these representations interact with local structure in modality specific spokes or are completely supported by the ATL has not been tested.

1.3 Identifying sparse, distributed, whole brain patterns of functional activity

Any analysis of complex high dimensional data, like that produced by neuroimaging, involves making assumptions about what constitutes meaningful signal. Multivoxel pattern analysis (MVPA; Haxby, Connolly, and Guntupalli, 2014; Haxby et al., 2001; Norman, Polyn, Detre, and Haxby, 2006) and Representational Similarity Analysis (RSA; Kriegeskorte, Goebel, and Bandettini, 2006; Kriegeskorte and Kievit, 2013) have greatly expanded the range of representational hypotheses that can be tested relative to the narrow sensitivity of univariate analyses. In addition to the tools and techniques that have taken firm hold in the neuroimaging community, there is active cross-disciplinary work with engineers producing novel methodologies that sit on the cutting edge of optimization theory and neuroscientific investigation. In Chapter 3, I will review several of the most prominent and influential of these techniques with a focus on their representational assumptions, which will lay the groundwork for introducing network RSA (Oswal, Cox, Lambon Ralph, Rogers, and Nowak, 2016) in Chapter 4, a novel analysis framework developed with the explicit goal of discovering signal with the characteristics of distributed representations.

Of course, it is not a given that artificial neural networks are worthy analogies to

real neural networks, a fact which may cast doubt on whether the brain utilizes distributed representation at all. If it does not, then this is two hefty marks against the hub-and-spoke hypothesis: it indicates invalid representational assumptions and eliminates a major argument for why the ATL appears inactive during semantic cognition when assessed by fMRI. This important consideration will be addressed in Chapter 2, where I will define distributed representations and make a case that brain does utilize them (see also C. R. Cox, Seidenberg, and Rogers, 2015).

1.4 Cross-modal representations in the ATL

If the ATL is a pan-modal hub, it must participate in semantic encoding regardless of the input modality. The semantic hub and hub+spoke hypotheses both predict that the semantic structure encoded by the ATL is referenced regardless of the input modality, and, given a unimodal stimulus like an image, is critical to determining that associated properties in other modalities (Patterson & Lambon Ralph, 2016; Patterson et al., 2007; Rogers et al., 2004). This critical hypothesis can be tested by presenting with stimuli that represent a common set of concepts from different modalities on different trials, for example presenting a picture of a labrador on some trials and a sound clip of a labrador barking on others. Network RSA can be used to estimate the network of voxels that express the semantic similarity structure for trials picture and sound trials independently. The test of this hypothesis could then be determined by a conjunction analysis. The fMRI dataset which is reported an analyzed in Chapter 5 was designed with this test in mind, and so I will be able to perform this test and potentially observe that the ATL encodes a cross-modal semantic space.

Although it has not previously been possible to assess the representational structure encoded in distributed networks in the brain, there does exist consid-

erable evidence that the ATL is involved in multi-modal representation, some of which was reviewed in Section 1.2. Two recent intracranial ECoG studies have detected local field potentials that that are similar whether recognizing a famous person by their voice or their face (Abel et al., 2015) and shown that direct stimulation at certain sites can disrupt both the comprehension of a spoken command and the ability to perform picture naming or reading tasks (Shimotake et al., 2014). Of course, that the ATL encodes cross modal structure is also evidenced by the pattern of impairment seen in patients with semantic dementia, whose impairments are not restricted by modality (Hodges, 1995; Hodges & Patterson, 2007; Hoffman, Jones, & Ralph, 2012; Nestor, Fryer, & Hodges, 2006; Snowden, Goulding, & Neary, 1989).

1.5 Summary

To date, ATL's hub status has been inferred from its anatomical connectivity (Binney et al., 2012; Morán et al., 1987), its multi-modal sensitivity (Shimotake et al., 2014), its association with domain- and modality-general semantic impairments when damaged by disease (Lambon Ralph et al., 2007; Mayberry et al., 2011; Rogers et al., 2004) or disrupted by transcranial magnetic stimulation (Ishibashi, Lambon Ralph, Saito, & Pobric, 2011; Jackson, Lambon Ralph, & Pobric, 2015; Pobric et al., 2007, 2010a, 2010b; Pobric, Lambon Ralph, & Zahn, 2016), and demonstrations of its functional involvement in semantic tasks performed by healthy participants relative to non-semantic tasks (Binney, Embleton, Jefferies, Parker, & Lambon Ralph, 2010; Visser, Embleton, et al., 2010; Visser, Jefferies, Embleton, & Ralph, 2012). Nevertheless, the contribution of the ATL to semantic representation remains in controversy because the representational predictions of the hub-and-spoke model have not been but to a fair test due to past methodological limitations.

In the proposed experimental work, the objective is to evaluate predictions the hub-and-spoke model make about how semantic representations are expressed in patterns of neural activity. In Chapter 2, I will review the functional neuroimaging literature to support the foundational assumption that the brain utilizes distributed representations. Then, in Chapter 3 I will consider the technical challenges associated with identifying representations of this kind, and demonstrate the need for a methodological innovation, which we call network RSA (Oswal et al., 2016). Chapter 4 will then be dedicated to introducing network RSA in detail to build intuitions about how it can be applied and what novel insights it can provide to cognitive neuroscientific research. With these foundations laid, Chapter 5 will contain a series of analyses peformed on a cross-modal fMRI dataset that has many qualities that make it ideal for testing the representational hypotheses associated with the hub-and-spoke model outlined in this chapter. Chapter 6 will follow up with an analysis of an ECoG dataset, including local field potentials collected from microelectrod arrays implanted in the ventral ATL. I will then conclude the dissertation in Chapter 7 with a general discussion in which I interpret my novel empirical work in broader context.

Chapter 2

Distributed Representations[1]

Distributed representation is one of the central tenets of the Parallel Distributed Processing (PDP) framework (Rumelhart, McClelland, & PDP Research Group, 1986). The basic notion is that entities such as words, concepts, objects, faces, places, and so on, are represented with patterns of activity over sets of neural processing units. An individual unit may participate in many different representations, while representations that express similar content will be coded with similar patterns over many units. The utility of such representations has been demonstrated in PDP models of many phenomena in many domains; a recent special issue of Cognitive Science, for instance, surveyed the impact of the PDP approach in the domains of learning, perception, language, memory, cognitive control, and consciousness (see Rogers and McClelland (2014), and accompanying articles). Together such models instantiate a theory of cognitive representation and processing that differs from traditional approaches involving rules (Pinker, 1991), "theories" (Gopnick & Wellman, 1994), and other symbolic representations (Tenenbaum, Griffiths, & Kemp, 2006). The models explain how representations of different types of knowledge develop, how such knowledge is structured and

[1]This chapter is based on C. R. Cox et al. (2015), adapted for clarity and to flow the the rest of the document. The full article situates distributed representations in a discussion about functional neuroimaging methods and high-level functional specificity in the brain.

organized, and how it is used in performing different tasks. Models using distributed representations have provided new accounts of important elements of intelligent behavior (e.g., generalization) and explain detailed aspects of behavior that other theories miss (e.g., the quasiregular character of language and other types of knowledge; Plaut, McClelland, Seidenberg, and Patterson, 1996; Seidenberg and Plaut, 2014). They have also been the foundation for influential accounts of many neuropsychological phemonena (Farah & McClelland, 1991; Harm & Seidenberg, 1999; Joanisse & Seidenberg, 1999; Lambon Ralph et al., 2007; Plaut, 2002; Plaut & Shallice, 1993).

Despite these successes, important questions remain about the epistemic status of distributed representations. The promise of the PDP approach is that the use of "neurally-inspired" constructs such as distributed representations would prepare the way for integrated theories of behavior and its brain bases. On the cognitive side, the relevance of distributed representations to understanding behavior is well-established, but the models obviously abstract away many complex properties of neural systems. On the neurobiological side, it is generally accepted that mental representations are instantiated as patterns of activity over large systems of individual neurons, which communicate through synaptic networks with structure at multiple spatial scales. It remains unclear, however, just how such networks represent entities such as words, concepts, objects, etc. (Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). The gulf between high-level cognitive and low-level biological understanding of representation thus raises questions about the extent to which the neural representations of cognitive entities are distributed in the PDP sense.

But why put so much emphasis on this aspect of PDP models? If this gulf were to be closed, what would be accomplished? Distributed representations are so important because they are an emergent property of neural network models.

Distributed representation is not *chosen* from among a set of alternative representational strategies when developing neural network simulations. The modeler can design the organization of units and links, choose the response functions for units and the learning rate on links, and even manipulate the error metric that supervised networks iteratively work to minimize. But PDP models will always develop distributed representations as they extract structure from the environment and learn to behave in desired ways. The content that they encode and the dynamics associated with their development, utilization, and degradation are inextricably bound. This means that these models are only relevant from a neuroscientific perspective if the brain also employs the basic mechanisms that would support the existence of PDP-like networks.

If PDP models can be judged to be neurally plausible, this is a boon to cognitive neuroscience. Questions that pertain to the computational qualities of billions of unmeasurable neurons in a human brain are virtually unsolvable without a suitable modeling environment that abstracts away confounding complexity while retaining essential core mechanisms. To explain how and why a PDP model performs the way it does will invariably include an analysis the distributed representations that the model has acquired, and the hub-and-spoke model is no exception. Consider the following apparently paradoxical qualities of the ATLs asserted by the hub-and-spoke model: the two ATLs jointly support a single, domain-general representational space for pan-modal conceptual knowledge, yet bilateral damage results in much worse semantic impairment than equivalent damage concentrated in a single hemisphere (Hermann, Seidenberg, Haltiner, & Wyler, 1995; Hermann et al., 1994; Lambon Ralph, Cipolotti, Manes, & Patterson, 2010; Patterson, 1995; Snowden et al., 1989; Warrington, 1975). An intuitive explanation might be that each hemisphere encodes its own similarity structure in parallel, and the structures encoded by each hemisphere are largely redundant. However, this is incon-

sistent with the notion of a single semantic system, and so would undermine a central tenant of the hub-and-spoke model. In a series of simulations, Schapiro, McClelland, Welbourne, Rogers, and Lambon Ralph (2013) explored how unilateral and bilateral lesions distort the internal distributed representations in a partial implementation of the hub-and-spoke model. Unilateral lesions reduced both the fidelity and magnitude of distributed activity in one hemisphere, allowing the other hemisphere to dominate. Critically, this pattern of results required that the two ATLs support a *single* representational space. If the ATLs were considered separately and then averaged, the difference between bilateral an unilateral damage vanishes; if considered separately and only the intact hemisphere is consulted, no deficits are observed which is at odds with the patient data. This is just one example where a serious challenge to a neurocognitive PDP model is addressed by appealing to the representational mechanics of PDP models. It means that the model is not vulnerable to this particular critique, but it places an increasing burden on the unconfirmed prediction that the brain utilizes distributed representation.

This chapter considers the status of distributed representations by examining their relevance at the level of analysis we will term "neurocognitive"—the level at which the processing units of neural network models arguably make closest contact with measurements of the neural activity underlying cognitive behaviors. Specifically, we consider whether measurements taken at the scale supported by fMRI and other contemporary functional brain imaging methods reveal neural representations that are distributed in the PDP sense. The grain at which these methods engage cognitive phenomena seems roughly similar: the units in PDP models are not neurons but capture, in simplified and abstract form, the aggregate behavior of many neurons. Likewise voxels in neuroimaging studies reflect, not the activity of individual neurons, but the aggregate behavior of many thousands of neurons. Many important phenomena have been explored using both

approaches. Our question, then, is whether neural representations of cognitive entities like words, objects, faces, and concepts are distributed at this level.

Before beginning, it is worth considering in more detail the rough correspondence suggested earlier between units in a PDP model and voxels in a brain imaging study. What motivates this analogy, beyond convenience? Brains are, of course, composed of neurons, and neural network models are sometimes described as assemblies of neuron-like processing units. Thus it might seem natural to think of a unit in a PDP model as roughly analogous to a single neuron. The analogy is tenuous, however. Whereas individual neurons exhibit all-or-nothing spiking behavior, units assume continuous activation states. Low level dynamics such as lateral inhibition, temporal coherence, and local extra-cellular conditions are glossed over in most connectionist models despite being critically important for understanding the behavior of individual neurons. The models also abstract away from morphological differences among neuron types, cytoarchitectonic details such as the organization of neurons into cortical columns, and other facts about brains. PDP units can instead be viewed as capturing, in a modest number of processing elements, the same informational states existing across vast numbers of heterogeneous spiking neurons in real nervous systems (Rogers & McClelland, 2014; Smolensky, 1986). The central assumption is that the representational content and cognitive functions expressed in the coordinated spiking behaviors of hundreds or thousands of neurons can be usefully approximated as a much smaller vector of continuous-valued activations, with individual units corresponding to single elements within the vector and summarizing the informational states of large populations of neurons.

Functional brain imaging adopts essentially the same central assumption. fMRI does not measure the activity of individual neurons but infers, via changes in blood oxygenation level at the scale of approximately $3mm^3$, the net synaptic input deliv-

ered to a population of thousands of individual spiking neurons (Arthurs & Boniface, 2002; Logothetis & Wandell, 2004). That is, each voxel provides approximate summary information about metabolic demands exerted by a large population of individual neurons. The effort to relate such measurements to cognitive representations and processes entails the assumption that there exists, in real brains, an important relationship between neural activity abstracted at this scale and the representations and processes that underlie cognition. Put differently, if functional brain imaging is to have any validity, it must be the case that the representational content and cognitive functions expressed in the coordinated spiking behaviors of hundreds of thousands of individual heterogeneous neurons can be usefully approximated with a smaller vector of continuous-valued activations. In this case, the elements of the vector are individual voxels and their values summarize a statistical relationship between the BOLD time-series at the voxel and other cognitive events, but the parallel to the central PDP assumption is clear. For this reason, in what follows we take the activation of a single unit to be a model analog of the mean activity in a population of neighboring neurons, similar to that estimated from changes in the BOLD response at a single voxel using fMRI. The central question is whether neural representations so measured have the properties that the PDP framework predicts.

Before turning to the literature to address this question, it is important to note that there are at least two profiles of functional activity that have been described as "distributed" which do not correspond to the kinds of representations that emerge in PDP-like systems. First, the word "distributed" is sometimes used in cases where univariate contrasts reveal reliable differences in BOLD response, not just in one cortical area, but in multiple anatomically well-separated areas. For instance, univariate fMRI studies of visual perception often show elevated BOLD responses for faces relative to other objects in parts of the occipital cortex, the pos-

terior fusiform, and the ventral ATL (Behrmann & Plaut, 2013). These regions are sometimes then described as forming a distributed network for face representation. Second, the word "distributed" sometimes refers to the case where a representation is encoded over multiple different regions, each encoding a different kind of information. For instance, theories of semantic representation often view the meaning of a word as being distributed over cortical regions that each individually encode a particular kind of sensory, motor, or linguistic information. Thus the color of an item is represented within a color area, shape is coded within a shape area, characteristic motion is encoded within a motion area, and so on (Fernandino et al., 2016; A. Martin & Chao, 2001). In this scenario, the meaning representation is distributed as a pattern of activation across many potentially widely dispersed cortical regions. This does not necessarily imply anything more than that concepts are supported by multiple cortical regions—it does not speak to the nature of the representations themselves. Indeed, A. Martin and Chao (2001) summarize a body of literature that exclusively employed univariate contrast analyses, which are not sensitive to distributed representations (see Chapter 3 and C. R. Cox et al., 2015 for a discussion of the assumptions made by univariate and multivariate analysis methods). Within the color area, for instance, the voxels are still viewed as always encoding color without contributing to the representation of other kinds of information; as encoding the color information independently, so that state of units outside the color area need not be taken into account; as being anatomically situated within the same contiguous region; and as being homogeneously located across individuals.

So what profiles of functional activity would provide support for distributed representations in the brain? Distributed representations that emerge through learning in PDP models have the following four characteristics:

1. *Interdependence of representational elements.* In PDP models, interesting repre-

sentational structure—phonological, morphological, conceptual, visual, etc.—is encoded in the patterns of activation evoked *across* whole ensembles of units, but may not be apparent in the individual activity of single units within the ensemble.

2. *Graded representation and functionality.* In PDP models, a given unit can participate in the representation of many different items. Within a distributed representation that robustly distinguishes, say, two different domains, a single unit may activate for subsets of items from both domains, or for all items in one domain and a few items in the other, or for only a subset of items within a single domain, and so on.

3. *Heterogeneity of representation within and across individuals.* In PDP models, the units that jointly encode a distributed representation—typically units within the same layer, which are connected in similar ways to other units in the network—can nevertheless exhibit very different responses to their inputs. Indeed, because the patterns of activation across units in a layer express representational structure suited to the task at hand, units within a layer must respond at least somewhat differently to different inputs.

4. *Heterogeneity of location within and across individuals.* Finally, even where different networks adopt the same representational code for various inputs, the localization of the code over units—the particular way that a given unit in a given layer responds to various stimuli—can vary arbitrarily.

In the remainder of the chapter, I will survey recent work in functional neuroimaging to assess the status of each of the four properties of representations predicted by PDP. The principal aim is to assess the face validity of the four properties, not to conclusively determine whether representations in some domain are distributed (in the PDP sense)—the current state of the evidence does not allow

this. Rather, the objective is to determine that neural representations are plausibly distributed in this sense. In each case I will first consider what evidence for distributed representation would look like, and then review studies that report relevant evidence.

2.1 Independence versus interdependence of representational elements

One point of contrast in the representational assumptions of classical univariate brain image analysis and PDP concerns the degree to which elements of a representation—units in a model or voxels in the brain—individually express important cognitive content. Standard brain imaging approaches assume that the important elements of representation can be discovered through univariate analyses, and hence that the elements contribute independently to representations. The fact that univariate methods often succeed in finding such elements indicates that clusters of voxels do indeed sometimes behave in ways amenable to discovery via univariate analysis. PDP, however, posits that information can sometimes exist in the pattern of activation across multiple units, without being reflected in the individual activations of the components. Is there any evidence supporting this hypothesis?

What might such evidence look like? As a start, consider that, if the univariate assumptions are always true, then the information encoded across all units in a representation will also be reflected in the individual elements of the representation. That is, there should be little or nothing gained in analyzing sets of voxels all together compared to analyzing individual voxels separately, since each element contributes to the representation independently. If the PDP assumptions are valid, however, there should be information contained in the patterns of activation across units that cannot be decoded from individual voxels taken separately. Thus,

if multivariate methods and univariate methods, applied to the same data-set in search of the same information, identify different voxel sets, this would suggest that the PDP assumption is valid.

Jimura and Poldrack (2012) conducted just such an analysis in a study of how the brain processes gain and loss in a gambling task. Many cortical regions were identified using a multivariate searchlight analysis (Kriegeskorte et al., 2006) that were not detected by the univariate method. If the searchlight method had simply identified a superset of the regions identified by the univariate method, the result might not be compelling—it may simply be that multivariate methods are "highly opportunistic" (Kriegeskorte et al., 2006, p.550), identifying regions with very weak signal just as might happen by relaxing the significance criterion in a univariate statistical test. What makes the result particularly interesting in the current context is that the voxel sets identified by the univariate analyses and were not simply a subset of those identified in the multivariate analyses overlapped very little. In addition to flagging regions that seemed irrelevant from the univariate analysis, the multivariate analysis did not flag several regions implicated by the univariate analysis.

Other work has demonstrated that the results of univariate and multivariate methods can actually doubly-dissociate. Riggall and Postle (2012) noted that regions in frontal and parietal cortex displayed sustained activation during the delay period of a working memory task in which participants were required to hold in mind the speed and direction of an array of moving dots. The authors trained a multivariate pattern classifier to determine which direction of motion was being held in memory, using the activations of voxels in these fronto-parietal regions as inputs. The classifier was unable to decode patterns at a level greater than chance, indicating that, despite the systematically elevated delay-period activity in these regions, the patterns did not encode the contents of working memory. Decoding

was possible, however, from classifiers trained on voxels in the occipital cortex, even though this area did not show elevated delay-period activity according to the univariate analysis.

Such result might initially seem counter-intuitive—surely an effect that can be detected by univariate methods must also be picked up by a multivariate approach. To see why this intuition is incorrect, consider the patterns shown in the schematic Figure 2.1. The three leftmost grids indicate a subset of voxels falling within a searchlight centered at the same anatomical location in three different individuals. The coloring of each grid cell indicates the degree to which each voxel's activations reliably predicts an experimental factor of interest: for instance, bright blue voxels might reliably predict that a stimulus item was from category A, while bright red squares reliably predict category B. Pale colors indicate voxels whose activity is only weakly correlated with the contrast of interest, while gray squares show uncorrelated voxels.

The top row exemplifies the case where the multivariate searchlight approach will identify signal missed by univariate analysis. Within each individual searchlight there are 2 or 3 voxels that reliably carry useful information about the stimulus class so that a trained classifier will successfully generalize to a hold-out cross-validation set. Such a classifier will therefore perform well in each individual subject, and the searchlight method will flag this searchlight location as encoding information relevant to the discrimination. Yet the particular way the information is encoded is highly variable across voxels in the searching for each individual, and the exact anatomical location of the signal-carrying voxels is highly variable across individuals. Blurring signal within subjects will thus destroy signal, and averaging signal at a given location across subjects will further eliminate signal. Thus the mean difference in BOLD response to category A versus category B items will be near zero for all voxels.

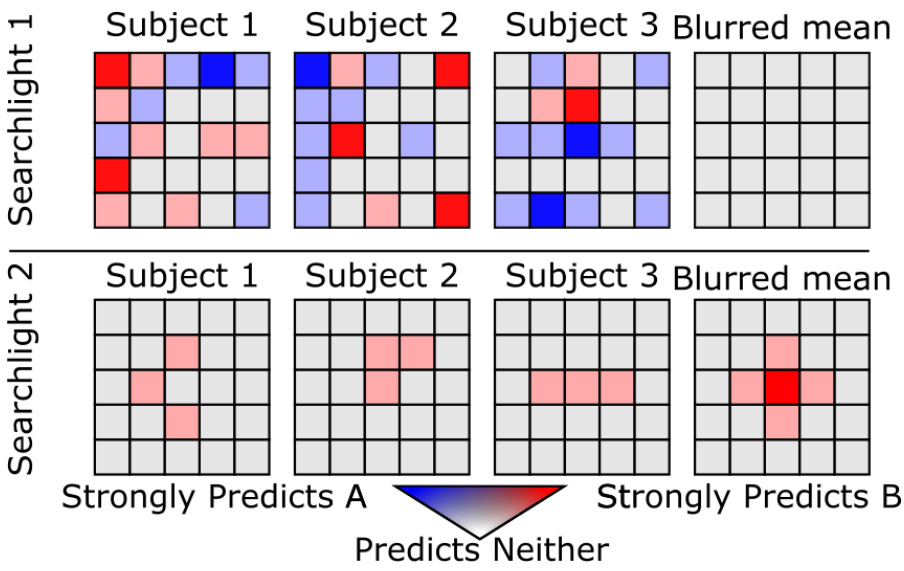

Figure 2.1: Each row corresponds to a searchlight that contains a set of 25 voxels. These voxels are the same across subjects, but different across the two searchlights. Searchlight 1 exemplifies a case where a searchlight MVPA will succeed but a univariate analysis, employing blurring within and averaging across subjects, will fail. Searchlight 2 exemplifies a case where a searchlight multivoxel pattern analysis (MVPA) is likely to fail but a univariate analysis will succeed.

The second searchlight shows the reverse case: here, most of the voxels in each subject are uncorrelated with the distinction between A and B, and only a small subset is weakly correlated with the distinction. A classifier trained on each subject individually has a high likelihood of failing a cross-validation assessment. If the classifier fails in many subjects, the searchlight centre will not be identified as reliably encoding information relevant to the distinction, meaning that this region will not be identified by a searchlight MVPA. The weakly covarying units, however, happen to encode the distinction of interest in the same manner, and to reside near one another within and across individuals. The univariate assumptions are met, so smoothing within and averaging across individuals reduces noise and allows detection of the voxel with univariate tests. Thus, the searchlight MVPA can fail to find signal in the very cases the univariate approach was designed to address—that is, when the signal is buried in noise within individuals, but is coded inde-

pendently in the same way and in the same location within and across individuals (see Chapter 3 for simulation examples of this case).

2.2 Graded versus discrete contributions to representation

The second point of contrast concerns the degree to which a given representational element can participate in many different representations. The classical view posits a discrete functional specialization, in which each element contributes only to a particular kind of representation—with, for instance, a given voxel activating only for faces (or a subset of faces), or for animals (or a subset of animals), and so on. Distributed representations, in contrast, are useful because they allow representational structure to be expressed as graded similarities across many representational elements. In such a scheme, any individual element will contribute, in graded fashion, to the representation of many different items or even to different representational domains. Thus a second important question for the literature is whether it contains evidence that individual voxels contribute in a graded fashion to different representations.

A clever indirect method for answering this question leverages neuronal adaptation (Grill-Spector, Henson, & Martin, 2006). Typically, the neural response to a stimulus will decrease over repeated presentations of the same item as active neurons deplete their resources with repeated firing. If representations are distributed so that two stimuli evoke overlapping patterns, the overlapping portions of the patterns would also be expected to adapt. Though the adaptation is happening at a scale much smaller than a functional voxel, if there is sufficient overlap across the representations the net effect will be to diminish the voxel's response relative to an appropriate control condition. The method naturally extends to any

domain where one is interested in testing whether neural representations overlap.

One particularly interesting domain to which fMRI adaptation analysis has been applied is lexical semantics evoked by word reading. Printed words are highly controlled stimuli, and orthography and phonology are both relatively uncorrelated with semantics, so it is possible to dissociate semantic from perceptual similarity (e.g., BIG and LARGE are semantically similar but formally dissimilar; HAIR and PAIR are semantically dissimilar but formally similar). Also, there is a deep psycholinguistic literature that has set a high bar for stimulus set composition; it is standard practice to control for word frequency and other potentially psycholinguistic dimensions, further isolating effects of interest.

With such stimulus sets, it is possible to use the adaptation procedure to assess the extent to which representations of different word meanings overlap. If such meanings are expressed as distributed patterns of activation, with similar meanings evoking similar and therefore overlapping patterns, the predictions for such a study are clear: The adaptation arising from successive presentations of semantically related words should be larger than that produced by successive unrelated words. That words from the same semantic category (e.g., two vehicles) result in more adaptation that words from different categories (e.g. a vehicle and an animal) is a widely replicated effect (Henson & Rugg, 2003; Rissman, Eliassen, & Blumstein, 2003; Wheatley, Weisberg, Beauchamp, & Martin, 2005). However, to address the PDP prediction that representations express graded similarity, more than two points are needed. With only two conditions, one cannot assess whether there is representational similarity among items from the same category. To our knowledge, graded similarity structure for the meanings of words referring to objects has not been explored using this method. It has, however, been explored in the domain of numbers and numeric magnitude. For example, (Piazza, Pinel, Bihan, & Dehaene, 2007) found that the degree of dis-habituation between a habituated

numeric quantity or numeral and a deviant stimulus was a function of the difference in magnitude. This suggests that there is graded similarity among number concepts, even when presented as Arabic numerals. This outcome would not be expected if the representations of meanings were discrete and non-overlapping.

Representational overlap can also be assessed by comparing the solutions found by two or more multivariate pattern classifiers within the same subject. Many such classifiers assign real-valued weights to each voxel that indicate the degree to which the voxel contributes to the relevant discrimination. When two or more classifiers are trained to perform different discriminations, the weights assigned by each classifier to each voxel can be compared. Voxels that receive large weights in both solutions can then be identified as important for both representational distinctions.

Studies of this kind are far less abundant, but do exist. One such study performed a three-way linear discriminant analysis of evoked brain activity measured by fMRI to distinguish trials in which subjects were presented with pictures of either faces, houses, or chairs (Carlson, Schrater, & He, 2003). After demonstrating above-chance pattern classification, the authors projected the solutions associated with each discrimination onto the brain, producing three maps of weights. The magnitude of each weight indicated how much the activation of a given voxel "pushes" the distributed representation away from the decision boundary, while the sign of the weight indicated to which side of the boundary the representation is being "pushed". The authors found that these solutions did overlap somewhat, meaning that some of the same voxels that were very indicative of "chairs" were also very indicative of "faces", and so on. Such results are particularly compelling given that the goal of linear discriminant analysis is to find the voxels that maximally discriminate between the three stimuli types. In principle this means that, so long as there are sufficient voxels responding uniquely to each category, other

voxels showing similar responses across two categories should be ignored—yet the analysis nevertheless identified voxels that appear to contribute simultaneously to two different object domains.

2.3 Heterogeneity versus homogeneity of representation

The third point of contrast concerns the degree to which elements of a distributed representation respond in the same way to objects of representation. By averaging neural responses across voxels in an individual, and again across individuals in group analysis, the standard approach appears to assume that, by and large, all elements respond to the objects of representation in the same way. For instance, it may seem reasonable to suppose that the voxels involved in coding perceptual representations of faces do so by showing consistently higher activation in response to visually-presented faces than to other objects. If this assumption is valid, and given the inherently noisy nature of the measurements in functional brain imaging, voxel-averaging is the correct thing to do: the noise at each voxel will cancel out across voxels, revealing the true underlying signal. The PDP view of representation, however, suggests the alternative possibility that the elements of a representation may respond to the objects of representation in quite different ways, both within individuals and across individuals. For instance, one face-relevant voxel might show elevated activation for one subset of faces and decreased activation of another subset; another voxel might show a different pattern of increased and decreased activation across various faces; and the ensemble together might express the degree to which different faces are perceptually similar. Since what matters is the similarity structure taken across elements, the responses of a single element within the representation can vary almost arbitrarily on this view, both in

an individual subject and across different subjects.

At first blush, there seems to be a substantial body of evidence in favor of representational homogeneity, both within and across subjects. After all, univariate methods that rely heavily on homogeneity have been applied effectively to fMRI data and yield replicable, consistent results, which would not be possible if neural responses to stimuli were purely heterogeneous and arbitrary across individuals. And, indeed, it has been demonstrated that cross-subject classification using multivariate classifiers is possible, albeit on coarse distinctions such as discriminating different tasks (Poldrack, Halchenko, & Hanson, 2009), sentences vs. pictures (Wang, Hutchinson, & Mitchell, 2004), or line drawings of tools vs. dwellings (Shinkareva et al., 2008).

Although these findings demonstrate that individual brains share important structure, they do not demonstrate representational homogeneity per se. To see this, consider the study of (Wang et al., 2004) and its later reanalysis (Rao, Cox, Nowak, & Rogers, 2013). The data were acquired while participants completed a cross modality match-to-sample task: one of the stimuli was a simple configuration of two symbols, and the other was a sentence which either did or did not correctly describe the configuration of symbols. Stimulus order was counterbalanced, and the goal was to determine, from the evoked BOLD response at a given time, whether the participant was reading a sentence of viewing an image. In the original study, classification across individuals was achieved by averaging voxel BOLD response within a small number of anatomically defined ROIs, and training a classifier using the averaged time series data from all but one participant. The solution was then used to classify each time-point in the functional data from the hold-out individual, and the results showed reliable above-chance performance. The analysis thus indicates a degree of consistency across individuals in the mean response to these different stimuli across coarse brain regions.

Still, the averaging at a broad grain ends up revealing little about the nature of the representations within and across individuals beyond this general consistency. Rao et al. (2013) looked for representational structure at a finer grain within and across subjects, using a whole-brain multivariate pattern classification method in which the responses of every individual voxel were provided as input, rather than the mean response averaged over pre-selected ROIs. To avoid over-fitting, the classifier employed a regularization penalty that preferred sparse solutions (i.e., most voxels receive weights of zero) in which selected voxels were located in roughly similar anatomical regions across participants.[2] In one sense the analysis replicated the original study: the majority of voxels that the classifier selected fell within the ROIs determined to be most informative by (Wang et al., 2004). The classifier solution also differed from that implied by the original analysis in important respects, however. Specifically, it did not identify some regions where all the weights were positive for all subjects (indicating, for instance, increased activation for sentences relative to pictures) and other regions where all the weights were negative (indicating the reverse). Instead, all regions identified included a mix of both positive and negative weights, consistent with the view that the representational code—whether high activation is observed for pictures or for words—can be heterogeneous even within a given circumscribed region, both within and between subjects.

Within this general mix, some regions showed a generally higher proportion of positive weights and others a generally higher proportion of negative weights, suggesting one explanation of the original result: when averaging across voxels within an ROI, the mean activity may carry signal because a majority of the underlying voxels code the information of interest in a particular way. But the analysis shows that such averaging can mask considerable underlying heterogeneity in the

[2]The SOS LASSO; see Rao et al., 2013 for a full explanation and demonstration of the technique.

representational code.

2.4 Heterogeneity versus homogeneity of location

The final point of contrast concerns the degree to which representations are localized homogeneously within and across individuals. The adoption of ROI averaging, cluster-thresholding, and spatial smoothing require the underlying assumption that the elements contributing to a given representation will be located near one another anatomically within subjects, whereas the anatomical alignment and averaging across participants require the additional assumption that localization will be largely consistent across individuals. The PDP view of representation, in contrast, suggests that the elements of a distributed representation may in fact vary substantially in their anatomical location both within and across individuals.

Several studies have now suggested that, in a variety of cognitive domains, neural representations are not confined to a small number of discrete and homogenous cortical regions but can be quite widely anatomically distributed. Recall that, in the study by Riggall and Postle (2012) discussed earlier, the authors were able to decode the direction of motion being held in working memory from activation patterns measured in occipital cortex. The same study further showed, however, that classification accuracy improved significantly when the logistic ridge-regression classifier was trained on data from the whole brain. The information maps generated from this analysis suggested the direction-of-motion signal was encoded in a very widely distributed cortical network, and not solely within a discrete region of visual cortex. Moreover, separate classifiers were trained and tested for each individual participant, so that the result did not arise from variability across subjects but illustrated heterogeneity of location within individual participants.

A similar result was obtained in a different domain in an interesting study by

Bulthé, Smedt, and de Beeck (2014). These authors applied multi-voxel searchlight, region of interest, and whole-brain classifiers to the same fMRI dataset, where the task was to decode numeric magnitude either from trials where Arabic numerals were presented (symbolic) or from trials where arrays of dots were presented (non-symbolic). This is a particularly interesting case, because prior univariate analyses implicated the intraparietal sulcus (IPS) as functionally specific for numerical magnitude, regardless of the stimulus modality (e.g., Dehaene and Cohen, 1997). The results indicated that both symbolic and non-symbolic magnitudes could be decoded from all lobes of the brain, and that whole brain decoding was on par with, if not better than, decoding from any individual lobe. The ROI analysis indicated that numeric magnitude could be decoded from nearly all ROIs during the non-symbolic trials (the visual word form area being the only exception), while only the IPS, fusiform, inferior occipital, left superior parietal, and the right superior frontal gyrus supported the decoding of magnitude during the symbolic trials. Finally, the searchlight analysis revealed that while non-symbolic magnitude could be decoded locally almost everywhere in the brain, symbolic magnitude could not be decoded anywhere from such local information. Thus in this case there appears to be information distributed across very widely situated voxel sets that cannot be extracted at more local scales, even by multivariate methods.

As a third example (Rish, Cecchi, Heuton, Baliki, & Apkarian, 2012) used elastic-net classifiers to predict judgements about the magnitude of a perceived stimulus in three quite different tasks, including magnitude judgments for visual object size, velocity of motion and pain intensity. In each task, an elastic net regression was run to select the 1000 most predictive voxels, a procedure that identified widely distributed sets of voxels that reliably predicted the magnitude of the pain. The authors then reran the analysis after excluding the 1000 voxels identified on the first run and found to their surprise that the predictive accuracy of the new so-

lution declined only negligibly relative to the original one. This process was repeated until performance reached floor. Remarkably, predictive accuracy in all three tasks declined very slowly. The authors interpreted this result as indicating that some kinds of information, such as stimulus magnitude, may be very broadly represented in the brain.[3]

Each example suggests that, at least in these particular cases, voxels that contribute to the discrimination of different cognitive states need not be situated near one another within a small set of cortical regions. What about localization across individuals? Is it possible that neural representations, even if they are widely dispersed anatomically within individuals, are nevertheless anatomically situated in similar ways across individuals?

The question can be very directly and elegantly addressed by leveraging a simple insight: if a representation is localized in the same way across a sample of subjects, the alignment of functional data should improve as the anatomical alignment improves. In turn, improving the functional alignment should increase the effect size in a univariate analysis. Tahmasebi et al. (2012) systematically varied how well participants' brains were anatomically aligned within a common space by applying a series of increasingly precise methods. He then assessed whether better anatomical alignment subsequently led to stronger effects in the analysis of functional data. In the experimental paradigm, subjects listened to sentences with ambiguous words ("His new post was in China"), matched unambiguous sentences ("The old tree was in danger"), signal-correlated noise (unintelligible noise matched to the intelligible sentences with respect to their length, spectral profile, and amplitude envelope), and silence in equal measures. Prior work had established where different univariate contrasts should produce reliable effects:

[3]Another possibility is that the tasks induce whole-brain metabolic changes that are correlated with stimulus magnitude but not involved in the cognitive representation of magnitude, an explanation that seems likely especially in the case of pain perception.

the contrast of sound to silence should activate the auditory thalamus, for instance, whereas the contrast of sentences to noise should activate primary auditory cortex, and the contrast of ambiguous to unambiguous sentences should activation the left posterior inferior temporal gyrus (ITG) and the left inferior frontal gyrus (IFG). With these predicted effects, the central question was whether improved anatomical alignment would increase the functional effect size in the relevant regions for each contrast of interest. The authors found that such an increase was indeed observed in the auditory thalamus, where auditory codes are presumably highly localized and consistent across subjects. A similar but weaker influence of alignment was also observed in primary auditory cortex, again consistent with the view that representations in this region should be relatively consistent across participants. This result was not obtained, however, for the contrast of ambiguous to unambiguous sentences. The size of the ambiguity effect was independent of the quality of the anatomical alignment, suggesting that the processes underlying ambiguity resolution are not anatomically localized in precisely the same way across subjects.

Other work has very directly assessed the degree to which the location of representational and processing structure varies across individuals. In one particularly compelling study, Feredoes, Tononi, and Postle (2007) considered a discrepancy in the neuroimaging literature related to working memory: group-level analyses tend to yield data consistent with the hypothesis that the PFC serves as a working memory buffer, evidenced by a delay-period sensitivity to memory load, whereas single-subject case study analyses tended to not show this effect. Instead, single-subject analyses implicated quite different regions in different people. These single-subject effects tended to be of greater magnitude than the group level effect in the PFC, leading to the hypothesis that working memory is supported by different regions in different people, with only weak involvement of the

PFC in any individual. Because the weak PFC activity is more consistently localized across individuals, however, this is the region that emerges in the group-level univariate analysis. An alternative hypothesis, and the one typically adopted in standard image analysis, is that the single subject effects are just noise. To adjudicate these interpretations, the authors applied TMS in each participant to either the PFC region identified in the group analysis or to an individual-specific location corresponding to the region of greatest activation during delay-period in the fMRI session. Larger effects were observed when TMS was applied to the individual hotspots than to the shared PFC region, suggesting that these regions—which were heterogeneous in location across individuals—nevertheless were playing a more important role in supporting the working memory task.

2.5 Conclusions

The preceding review supports the face validity of the representational claims staked by PDP. There is at least some evidence that, in at least some cognitive domains, neural representations measured at the scale of fMRI possess each of the four properties of distributed representations articulated earlier. What are the implications of these observations for our developing understanding of representation in the mind and brain? In particular, what can be concluded from the body of neuroimaging literature that has relied on univariate analyses, and how should future efforts proceed? These issues are given serious consideration in the General Discussion of C. R. Cox et al. (2015), the paper upon which this chapter is based.

Having now established the plausibility of distributed representations, I will now turn to addressing how they might be discovered in neuroimaging data.

Chapter 3

Comparing methods[1]

In the previous chapter, I reviewed a collection of primarily fMRI studies which, when surveyed as a whole, indicate that the brain may encode some kinds of information as distributed representations. The evidence had to be pieced together—no single study could serve as a fair test, because no analysis technique was well matched to modeling the kind of structure that exists among distributed representations when the elements of the representations are not localized. What would such a technique look like?

fMRI and other functional brain imaging technologies ubiquitous in human cognitive neuroscience typically yield vast amounts of noisy data. To discern interesting patterns and test particular hypotheses, the statistical models employed must adopt specific assumptions about the nature of the underlying signal. For many years, standard statistical approaches were built upon representational assumptions that essentially presupposed that neural representations are localized and not distributed (Kriegeskorte, 2008; S. L. Small & Nusbaum, 2004). These assumptions are considered in detail by C. R. Cox et al. (2015), but briefly stated they are:

[1]This chapter is based on an unpublished manuscript by Cox and Rogers, adapted for clarity and to flow the the rest of the document.

1. *Independence of representational elements.* The representational or processing significance of a given voxel's activation does not depend upon the states of other voxels. Each voxel encodes whatever it encodes, regardless of what other voxels are doing at the time.

2. *Discrete representation and functionality.* The brain is best thought of as an assembly of discrete regions, where each region contributes to one kind of representation or process and not to others.

3. *Homogeneity of representation within and across individuals.* The voxels contributing to a given representation respond to relevant items in essentially the same way, both within and across individuals.

4. *Homogeneity of location within and across individuals.* The voxels contributing to a given function or representation are localized similarly both within and across individuals.

Notice that these are exactly the reverse of the representational assumptions associated with distributed representations discussed in the previous chapter. This rather narrow perspective was famously contested and expanded by Haxby et al. (2001) work with fMRI. In particular, they conducted their analysis in such a way that assumptions 1–3 were not necessary, and thereby demonstrated that—at least for visual representations of faces, houses, and common objects in inferior temporal cortex—these assumptions are invalid and lead to an incorrect characterization of the functional organization of this part of the brain.

The analysis of Haxby et al. (2001) is the first example of multi-voxel pattern analysis (MVPA) in the cognitive neuroscience literature. Their analysis involved assessing the similarity among neural representations, where a neural representation was defined as the pattern of activity over a population of voxels. It leaned heavily on the assumption that the correlation between two neural representations

corresponds transparently to how similar the encoded content is. They were able to show that stimuli from the same category had more similar representations (i.e., their patterns of activity were more highly correlated) than stimuli from different categories, suggesting that the population of voxels encoded information that captured the similarity structure of the stimuli.

This work had such impact because of how the populations of voxels were selected. Haxby et al. (2001) selected specific parts of the inferior temporal cortex that responded maximally to particular categories, like faces or houses, and their MVPA revealed that each area contained distributed patterns that discriminated many kinds of stimuli from one another. That is, faces could be discriminated from other stimuli (including cats, chairs, and small man-made objects) with information encoded in voxels maximally active for houses, and vice versa. Discrimination among nearly all kinds of stimuli was successful based on patterns of activity in all sub-samples of inferior temporal cortex that they considered. These results were critically juxtaposed with the typical inference drawn from standard univariate analyses (that make the four assumptions above), which is: if a region of the brain is significantly more active in one condition relative to another, then it is disproportionately more important to representing one sort of information than another. The MVPA based on representational similarity provided the initial critical insight that mean regional activity is insufficient to describe the processes and content associated with a region of cortex.

In the 15 years since, cognitive neuroscience has embraced the multivariate nature of neural representation. The increasing prevalence of MVPA during this period raised an important distinction between *where* and *what* information is encoded in the brain. Most often, the questions are handled separately. Anatomical or functional regions of interest tend to be identified by univariate analyses, which is taken to address where the information is, and then MVPA can be applied to

some or all of the voxels within these regions. Thus, univariate analysis remains the standard for learning about where information is encoded, even as MVPA has become commonplace.

The notable exception is information-based functional brain mapping, often referred to as the searchlight technique (Kriegeskorte et al., 2006). The searchlight technique is an elegant marriage of *where* and *what*: the researcher first defines a volume (typically a sphere or cube) that would encompass a small selection of voxels. Then, this volume is swept through the fMRI dataset, centering on each voxel in turn. At each position, an MVPA is conducted based on the voxels in the volume, and the result is stored at the current position. The result is an "information map" (Kriegeskorte et al., 2006). The information map indicates locations in the brain where the local patterns of activity contain information that is statistically related to an experimental manipulation, where "local" is defined as "voxels which fall within the scope of the searchlight".

However, as referenced in Chapter 2, even the searchlight is not sensitive to all ways in which neural representations may be distributed over the cortex. Most obviously, searchlight analysis cannot detect representations that extend over a wider range of cortex than the scope of the searchlight. This might be addressed by increasing the size of the searchlight; however, doing so would diminish the resolution with which the technique localizes information. This might be problematic when investigating the representation of semantic knowledge, since there is considerable evidence that semantic representations might be, at least in part, represented across several sub-systems associated with sensory modalities (Barsalou, 2008; Binder et al., 2009; Binder & Desai, 2011; Desai, Binder, Conant, & Seidenberg, 2009; Desai, Conant, Binder, Park, & Seidenberg, 2013; Fernandino et al., 2016; Fernandino et al., 2013a, 2013b).

There are results in the literature that suggest this issue of scope ought to be

taken seriously. For example, standard univariate analyses have implicated the intraparietal sulcus (IPS) as functionally specific for numerical magnitude, regardless of the stimulus modality (Dehaene & Cohen, 1997). However, Bulthé et al. (2014) report being unable to decode symbolic magnitude from the IPS—or anywhere else in the brain using a searchlight approach. However, they were able to decode symbolic magnitude using a whole-brain MVPA, conducted using ridge regression. This suggests that the information is not localized to an area that can be decoded without also considering information from elsewhere.

The state of the art in MVPA has now advanced to include a host of techniques that allow researchers to conduct whole-brain multivariate analyses that do not require guidance from a prior univariate analysis or assume that representations are fundamentally local as the searchlight does. However, far from being a panacea, whole brain multivariate analyses bring with them important assumptions and caveats that often complicate their interpretation. Each approach to fMRI analysis provides a different perspective on the data that is colored by different simplifying assumptions that will be in line with some kinds of representational structure but not others. An appreciation for what an analysis is and is not sensitive to is essential for making valid inferences from experimental results and thus for meaningful progress in cognitive neuroscience.

The goal of this chapter is to introduce and motivate the application of whole brain multivariate analysis via regularized regression. This will be accomplished through a series of simulated analyses applied to fabricated data constructed so that they contain information represented in a wide variety of ways. The intuitions developed through the simulations will help situate the methodology introduce and applied in subsequent chapters among related methods, the problems they are and are not well suited to, and the particular challenges involved in doing whole brain multivariate analyses. Note that while these simulations involve multivoxel

pattern (MVP) classification, the intuitions they aim to engender will translate to representational similarity analysis (RSA) and, ultimately, network RSA (Oswal et al., 2016).

3.1 A brief overview of PDP models and their representational assumptions

We have at this point discussed distributed representation at length, both their potential neuropsychological relevance and their form. However, we have avoided a more formal consideration of the computational framework that gives rise to them. PDP models are composed of simple processing units that communicate via weighted synapse-like connections (Rogers & McClelland, 2014; Rumelhart et al., 1986). Each unit adopts an activation state, typically varying between 0 and 1, that can be viewed as analogous in some respects to the mean firing rate of a population of spiking neurons proportional to their maximal rate (Zipser & Andersen, 1988). Units transmit information about their current activation through weighted connections, which can be viewed as capturing the net effect of activity in one population of neurons on another. Weights are typically real-valued, with negative numbers indicating a net inhibitory effect and positive numbers indicating a net excitatory effect. Each unit computes a simple process: it adjusts its current activation state according to the input it receives from other units in the networks. If a given receiving unit receives inputs from a set of n sending units, then the input is usually computed as the inner product of the activation across all sending units and the values of the weights projecting from the sending units to the receiving units. The unit then converts the net input into a new activation state according to a specified transfer function (often a sigmoid function of the net input). All units are conceived as computing inputs and updating activation states in

parallel in continuous real time (hence "parallel" distributed processing), though on serial computers this parallel process is simulated by updating units in discrete steps in randomly permuted order.

Within a network, units are generally organized into layers, which govern the overall connectivity of the network: units within a layer tend to receive connections from, and direct connections toward, a similar set of units elsewhere in the network. Typically a subset of the units are specified to receive inputs directly from sensory systems (or other input systems outside the model), and to direct outputs toward motor systems (or other output systems outside the model). These unit subsets encode the input provided to the model and the outputs that simulate the model response. They are often referred to as visible units, because the theorist directly stipulates how different stimulus events and behaviors are represented with patterns of activation over the input and output units. Most models also include sets of units whose inputs and outputs are directed only to other units contained within the model—they do not receive external inputs from or direct outputs toward the model environment. For these hidden units, the theorist does not stipulate how different stimulus events or behaviors are to be coded with patterns of activation. Instead, the patterns of activation that arise across these units are determined solely by the values of the interconnecting weights.

The weights themselves are viewed as being shaped by learning and experience. Many different learning algorithms have been explored in this framework (see Hinton, 2014), but all share the general idea that the weights gradually change over time in order to optimize some objective function—for instance, minimizing the discrepancy between the outputs the model generates and the correct "target" outputs—as the network processes information from different stimulus events. Because the weights adapt to experience, and because the patterns of activation over hidden units depend upon the weight values, PDP models are therefore capable

of acquiring learned internal representations: the patterns of activation generated over hidden units by a given stimulus after the network has undergone learning in a model environment. One interesting aspect of PDP models, responsible for their utility in many different cognitive domains, concerns the nature of the internal representations they acquire after learning in a structured environment. Often the models can acquire internal representations that may seem counter-intuitive from other points of view, but that can be shown, through computer simulations, to support behaviors documented in the domain of interest. Figure 3.1 and its caption provide one example of a PDP model used to understand aspects of semantic memory.

With this overview of how PDP models work, we are ready to consider the challenges that the framework raises for the discovery of mental representations in functional brain imaging data. Many difficult issues arise, of course, in any effort to relate artificial neural networks to real neural networks. Because network models are functional abstractions of the neural processes they aim to uncover, they necessarily gloss the complexity, and many aspects of structure and behavior, known to be important in real nervous systems. In Chapter 2 (p. 2), I argued that PDP units can viewed as capturing, in a modest number of processing elements, the same informational states existing across vast numbers of heterogeneous spiking neurons in real nervous systems (Rogers & McClelland, 2014; Smolensky, 1986). The central assumption is that the representational content and cognitive functions expressed in the coordinated spiking behaviors of hundreds or thousands of neurons can be usefully approximated as a much smaller vector of continuous-valued activations. The effort to relate neural activity to cognitive events entails the assumption that important informational states over vast sets of neurons can be so abstracted.

Let us therefore adopt a fairly simplified stance on the relationship between network models and the brain networks we seek to discover in imaging data. Specifi-

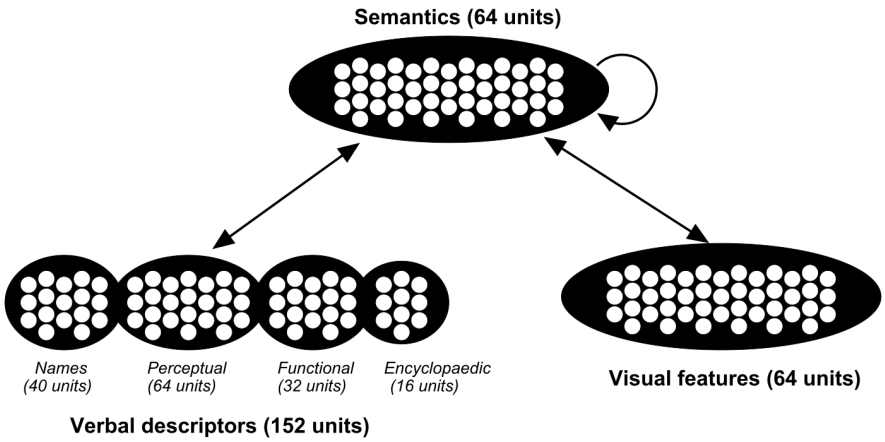

Figure 3.1: A PDP model used to understand semantic memory (from Rogers et al., 2004). Units in the *Visual* layer code visual features and units in the *Verbal* layer encode familiar words. The *Visual* and *Verbal* units can receive direct inputs from the environment, corresponding to direct perception of a visually-presented item or of a spoken statement. Units in both layers send connections to, and receive connections from, an intermediating hidden layer. To simulate a task such as object naming, visual features of the object are directly activated in the *Visual* layer and the activation propagates to other units via the weighted connections. If the weights are set to appropriate values, the model will ultimately activate the *Verbal* unit corresponding to the item name. Likewise name comprehension is simulated by directly activating the unit corresponding to the name and propagating activation throughout the network. With appropriate weights the visual features of the named item will activate, along with verbal units describing the item's properties. Appropriate weights are discovered through a predictive error-driven learning algorithm. Following learning, each input provokes a pattern of activation over hidden units that depends on the acquired weight configuration—a learned internal representation of the input. Though the particular pattern acquired for a given item varies across training runs, the representations always encode the same similarity structure among items in the environment, representing items that are conceptually related with similar patterns of activation.

cally, let us assume that the activation of a single unit in a network model is roughly analogous to mean neural activity in a population of hundreds of neighboring neurons within a small volume of cortex as estimated at a single voxel from BOLD activity in fMRI. Thus we will treat the pattern of activation generated by a given stimulus over units in a model network as analogous to the set of beta coefficients estimated over voxels from the BOLD response evoked by a given stimulus in a sparse event-related design.

3.2 Challenges for the discovery of PDP representational structure in the brain

Even with this relatively transparent view of the relation between model elements and measured physiological responses, PDP raises four difficult challenges for the discovery of representational structure in the brain.

The same content in the same location across individuals can be associated with different functional activity

For any given network, there are typically many different weight configurations that can generate appropriate outputs given the various inputs. The particular configuration that a network discovers with learning can depend on many things, including the initial random weight configuration, the ordering and distribution of the learning experiences sampled from the environment, and the effects of noise in the unit activations and/or weight changes. Thus a particular hidden unit in a given model can, across different training runs in the exact same environment, exhibit quite different patterns of activation in response to a given input. Yet the internal representations learned by a network are not arbitrary; the learning models are of interest because they reliably extract important similarity structure across the set of input and output patterns to which they are exposed. What varies is the particular way that individual units contribute to encoding the interesting structure across network runs. C. R. Cox et al. (2015) provide an example of this kind of variability in a simple model.

We can conceive of a single model training run as simulating the effects of learning and experience in a single individual person. The different weight configurations and internal codes that arise across model training runs thus indicate the kind of variability in representation that may exist across individuals under the

PDP view, even if the individuals show the same pattern of overt behavior in the domain and the same gross neural architecture. Specifically, the response generated by a given stimulus or process in a given patch of cortex may vary arbitrarily across individuals, even if the same representational structure is being encoded across the same cortical subregions. This possibility poses a challenge to imaging methods that focus on finding voxel clusters that reside in similar locations and respond in similar ways across individuals. If representations vary across individuals in the way that PDP models suggest, such methods will fail to discover them.

This consequence of distributed representations may pose greater problems for finding signal in some cortical regions than others. In peripheral regions (i.e., early sensory and motor cortices), it is clear that information is encoded in largely the same way, and with a largely similar neuroanatomical organization, across individuals. In association cortices, it may be that neural codes are less constrained are more strongly shaped by learning and experience, so that the way information is organized across cortex is more highly variable. PDP models provide a rough analog to this state of affairs, insofar as input and output units for a given model are stipulated to represent information in the same way in every model training run—that is, in every model "individual." The issues of variability in representation mainly apply to learned internal representations coded across hidden units.

Activation of individual units may not be interpretable independent of other units

A corollary of the preceding points is that the behavior of a given cortical unit may not be interpretable, or may have quite different interpretations across individuals, when analyzed independently from other units. This property of distributed representation is important because it suggests that univariate approaches to data

analysis—methods that assess the behavior of individual voxels or voxel clusters independently—can fail to uncover important components of neural representations. Wherever the interesting structure is embedded in activations across multiple cortical units, but is not reflected in individual units, such methods will yield null results (see C. R. Cox et al., 2015 for a concrete example).

The functional model architecture may not map transparently onto anatomical structure in the brain

A third issue concerns the relationship between the functional architecture of a computational model used to simulate performance on a task of interest and the actual anatomical structure of the corresponding neural network in the brain. As noted earlier, units in PDP models are organized into layers, with units in a given layer receiving connections from and directing connections toward the same subsets of units elsewhere in a network. The layer is a useful construct for understanding how a network functions, insofar as the units within a layer, by virtue of having similar connectivity to the broader network, "work together" to represent and process the same information. Distributed internal representations in PDP networks are typically viewed, therefore, as being encoded across units within a particular layer.

It may seem natural to view layers as model analogs of cortical regions, so that the gross architecture of a computational model maps transparently onto the anatomical structure of networks in the brain that carry out the modeled cognitive function. Though this analogy is reasonable, it is not the only possible way that the functional architecture of a computational model might relate to the neuroanatomical structure of a corresponding cortical network. In fact, the layers of a computational model do not, in principle, have any implications for how the corresponding cortical units might be anatomically situated in the brain. Units

that function together as a "layer" could be situated in multiple different cortical regions, or widely dispersed anatomically, or interdigitated with other units subserving different functions. The defining property of layers in a computational model is their pattern of connectivity in the gross architecture, and the same network connectivity can exist among many different spatial arrangements of units. In other words, the relationship between the functional architecture of a computational model—the grouping of units into layers as typically depicted in model figures, for instance—may not transparently reflect the topological arrangement of the corresponding cortical units in the brain.

This lack of transparency poses a problem for approaches to brain imaging that assume representations to be encoded over a volume of anatomically contiguous cortical units, including approaches that average signal over regions of interest, that spatially blur signal, or that restrict statistical analysis only to voxels within pre-specified areas. If cortical units that function together as a representational substrate do not happen to reside in a single contiguous cortical region, such methods may fail to discover important signal.

This is not to suggest that the PDP view predicts that anatomical structure is unimportant, or that shared structure across individuals is unexpected or meaningless. To the contrary, the connectivity of a given network strongly constrains its behavior. Thus the network architecture always constitutes an important aspect of the explanatory hypothesis a model is intended to exemplify. It is typically assumed that this architecture is largely shared across individuals, and that, however it is anatomically situated in the brain, there will be at least coarse similarities across individuals.

A network of interest will co-exist in the brain with many other networks serving other functions

Any given computational model is designed to aid understanding of a particular aspect of cognition, and typically includes only those elements that the theory stipulates to be important for the behavior of interest. Even if the model is a relatively faithful and accurate abstraction of a real cortical network in the brain, the physiological measurements generated by that network will be intermingled with measurements from a great many other cortical systems involved in other aspects of cognition unrelated to the task of interest. Odds are that the great majority of measurements taken will reflect metabolic activity unrelated to the representational structure we are searching for. Thus the effort to find distributed representations in brain imaging data raises concerns about needles and haystacks.

Summary

The representational assumptions of the PDP framework lead to rather bleak outlook. The behaviors of individual cortical units (i.e., voxels) may not independently covary with or otherwise reflect the objects of representation we are interested in finding in a systematic way across individuals. Mental representations may instead inhere in the patterns of activation evoked across whole sets of units that together function as a representational ensemble by virtue of their connectivity within the overall cortical network (like the layers of a neural network model). This is the core sense in which representations are distributed in the PDP view. The way that a particular unit contributes to different representations can be highly variable across individuals, even if the ensemble encodes the same representational structure across individuals. This means that the search for voxels exhibiting similar responses in similar anatomical locations across people will fail to reveal important

representational structure. Moreover, the units that operate together as a representational ensemble may not be anatomical neighbors, may vary in their location to some extent across individuals, and are certain to be buried within the mountain of measurements provided by functional imaging technologies across the whole brain. These possibilities raise a daunting challenge: representational structure can only be discerned by considering the whole pattern of activation over a representational ensemble; yet the units within such an ensemble may be anatomically dispersed and intermingled with a vast amount of irrelevant information. One cannot understand the representation without knowing which units together constitute an ensemble, but how is one to find the ensemble in the first place?

In what follows, we assess how well different approaches to fMRI data analysis meet this challenge by applying them to the discovery of representational structure in data generated by a simple neural network model as it processes different input patterns. We will see that all methods bring with them important biases in the kinds of representational structure they are capable of detecting, and that some methods are better-suited to finding the distributed structure that the PDP framework assumes. We will also see that patterns of results across methods can provide important information about the nature of the representations encoded in different parts of a network.

3.3 Model and Methods

The model we will employ for these analyses is illustrated in Figure 3.2A. It is an auto-encoder network: when presented with an experience in the form of a pattern of activity over its 36 input units, it learns to reproduce that same pattern over its 36 output units. Auto-encoder networks have been used as simple models of human memory, because once they have learned they are capable of both retrieving full

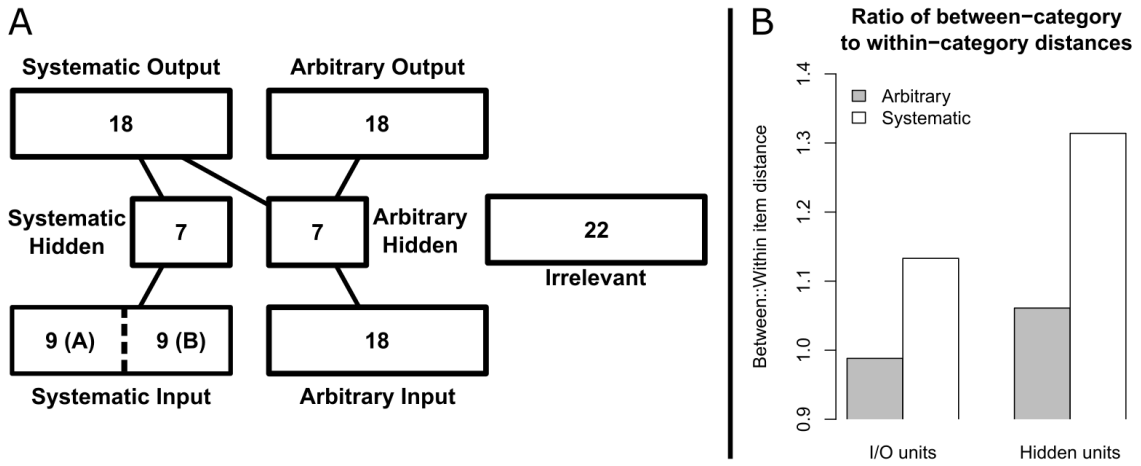

Figure 3.2: A) Architecture of the auto-encoder network used to generate the data for 10 model subjects used in subsequent simulations. The model has 36 input units (18 systematic), 14 hidden units (7 systematic), and 36 output units (18 systematic). The 22 irrelevant units are completely disconnected from the network, and stand for units that subserve an unrelated function but are anatomically adjacent to units of interest. B) Left: Ratio of between-domain to within-domain Euclidean distances for the representations coded over different sets of units in the network, averaged over the 10 model subjects. Distributed representations that encode the domain structure should have large distances for items from different domains and small distances for items from the same domain, and so should show a large ratio. While the systematic I/O units clearly code the domain structure to some degree, the systematic hidden units express the structure more strongly.

information from a partial cue and of generalizing prior learning to new items (McClelland & Rumelhart, 1985). In this case, however, we do not intend the model to embody a specific hypothesis about a particular real-world cognitive function. Instead, it is designed to make explicit the challenges noted in the introduction.

To this end, the patterns that the model processes are viewed as coming from two different domains, A and B, corresponding to some cognitive distinction of theoretical import. For instance, A and B might correspond to nouns versus verbs, or animals versus manmade objects, or faces versus non-faces, or any other binary distinction thought to be of potential relevance to behavior. Each individual item is represented with a unique pattern of activation over input units, and the network's task is simply to generate the same pattern over output units. In this sense, there is no explicit representation of the two classes A and B in the inputs, outputs, or

network task. The two domains are assumed, however, to be distinguishable from the distribution of input/output properties they possess. Specifically, one subset of input/output properties is marginally more likely to be active for items from domain A, while another subset is marginally more likely to be active for items in domain B. We will refer to these subsets together as systematic I/O units, because they each weakly co-vary with the representational distinction of interest. Each item also possesses several features coded by arbitrary I/O units whose activations do not systematically differ between domains.

After the model has learned, it is possible to "query" it by presenting an input pattern and generating patterns of activation throughout the rest of the network. As noted earlier, we take the activation at each unit in response to an input as a model analog of the neural response to a stimulus estimated from the BOLD signal at a single voxel in a single individual. Across different training runs, the model will always exhibit the same overt behavior (generating the correct pattern over output units), but arising from different configurations of weights, and hence from different internal representations. Variability in weight configurations and internal representations acquired across different training runs thus provides a model analog of individual variability in the neural representations acquired across the population. To simulate data generated by a functional brain imaging study with, say, 10 participants, we train the model 10 times with different random initial weight configurations. For each trained model, we record the pattern of activation generated over all model units by each input pattern (i.e., stimulus), taking these as model fMRI data. The question we then wish to ask, by applying different statistical methods to the analysis of this synthetic imaging data from a sample of trained models, is the following: which units in the network encode representations of the domains A and B, and how?

The network architecture is designed so that there are two possible answers to

this question. The first answer is that representations of A and B are directly encoded in the individual activations of the systematic I/O units. For all input and output units, the response of a given unit to a particular item is directly specified by the environment, so that these units will always respond to a given stimulus in the same way across model individuals. Each systematic I/O unit has a marginally different probability of being active depending upon the domain; in this sense the A units each independently encode a representation of the A domain and the B units encode a representation of the B domain. The relationship between domain and activation is, however, stipulated to be quite loose: for each domain, only a small number of the corresponding systematic units will be active for any given item—each unit participates in just a few patterns. Each item thus overlaps in their systematic properties with just a few other items in the domain, and the correlation between activation and domain is weak for any individual unit. We further stipulate that the A input and output units are anatomical neighbors, as are the B input and output units, and that this anatomical arrangement is exactly the same across individuals. Thus the systematic I/O units individually encode a weak distinction between A and B that is consistent across model individuals and is anatomically localized within input and output layers.

The second answer is that the representations of A and B domains are encoded in a distributed fashion over a subset of model hidden units. As shown in the Figure, the input units project to the output units by way of two separate hidden layers. The systematic hidden layer (SH) contains 7 hidden units that receive connections from the systematic input units and send connections to the systematic output units. The arbitrary hidden layer (AH) also contains 7 units that receive connections from the arbitrary inputs, and send connections to both the systematic and arbitrary outputs. The weights are shaped by learning, so every input generates a pattern of activation—a learned internal representation—over both the SH

and AH layers. The particular way that layers are connected, however, ensures that these internal representations will have specific representational properties. The SH layer connects systematic inputs to systematic inputs. Because items within a domain have a weak tendency to share systematic properties, the SH units can efficiently perform their mapping by representing the domain structure: items within a domain evoke similar patterns over units and items from different domains evoke quite different patterns. The AH layer receives inputs only from the arbitrary input units and directs outputs to all units. There is no tendency for items within a domain to share arbitrary features, so there is little pressure for these units to represent the domain structure. The AH layer thus acquires distributed internal representations that have little obvious structure. The weights in the arbitrary pathways effectively serve to "memorize" both the arbitrary features and the idiosyncratic differences among items in the same domain. In other words, the architecture produces a division of labor in which the SH layer learns distributed representations of the domain structure and the AH layer learns idiosyncratic differences among items. A good method, then, should identify SH units as important for representing the domains.

Indeed, the SH units arguably provide a better encoding of the domain structure than do the systematic I/O units. To illustrate this we first trained the model to saturation on 15 runs with different initial random weights, then analyzed, for each layer in the model, the Euclidean distance between the patterns of activation elicited by each pair of stimuli in the model. For each layer, we computed the mean distances for pairs within a domain and for pairs in different domains. We then took the ratio of between-domain to within-domain distances as a measure of how well the domain structure is expressed in each layer. A ratio of 1 indicates that between- and within-distances are about the same; a number greater than 1 indicates that between-domain distances are larger on average than within-domain

differences, indicating good differentiation of the domains. The results averaged across the 10 model subjects are shown in Figure 3.2B. For arbitrary units (both I/O and hidden), no domain structure is expressed: both ratios are near 1. For systematic I/O units, the ratio is clearly larger than 1, indicating reasonable encoding of the domain structure, but the ratio is much larger for the SH units, indicating that these distributed representations do a better job of systematically differentiating items from the two domains.

Finally, the model also includes 22 completely irrelevant units. These are units assumed to be anatomically near the SH and AH units but uninvolved in the task. These units always take a low activation value, and provide a simple model analog of the fact noted above that the units of interest may exist alongside other units that remain uninvolved in the task under investigation.

With this general understanding of the model behavior, let's consider how it makes explicit the four challenges for brain imaging noted earlier. First, although the SH units jointly encode the same representational structure across model subjects, the contribution of a given hidden unit to this structure varies arbitrarily across model subjects (Challenge 1). Second, the mean activations of SH units do not systematically differ for items in the A and B domains: to find the important structure, one must consider the pattern evoked over multiple units (Challenge 2). Third, the functional architecture of the model shown in Figure 3.2 can be anatomically arranged in many different ways (Challenge 3). To make this issue explicit, we consider two different topographic arrangements of the functional model. In the first, units within the same layer are always situated as anatomical neighbors, so that the representations encoded by the SH and the AH layers are anatomically localized. In the second arrangement, we assume that the SH units are spatially intermingled with the AH units, in a different way across model individuals, so that the representations they encode are anatomically dispersed. In the results we will

consider how well each method identifies the SH units as a function of whether they are localized or dispersed. Finally, the model captures the idea that the units of interest constitute only a small proportion of all the units measured (challenge 4). In the model itself, most of the units encode information irrelevant to the stimulus domain (the 43 arbitrary units plus 22 irrelevant units). The next largest set are the 36 systematic I/O units that encode the domain structure weakly but consistently across subjects. The units of greatest interest, the 7 SH units, constitute just 6% of all the units in the data.

Summary

Though very simple, this auto-encoder network captures each of the challenges noted in the introduction: it acquires distributed internal representations that express representational structure of interest; the way the structure is coded across units varies in different individual models; the structure cannot be discerned from the activations of single units but arises in patterns over multiple units; the relationship between the functional architecture and the underlying model topography can be opaque; and the units that encode the structure we wish to discover are buried in a large number of other measurements. The question we now address is how well different analysis methods fare at discovering representational structure across both systematic I/O units and the SH units, when they are applied to data generated from a sample of model training runs

3.4 Simulation details

The model shown in Figure 3.2 was trained on 72 items sampled from two domains, A and B. Each item activated exactly 2 systematic and 2 arbitrary input units, and across items each unit was active in exactly 8 items. Half of the system-

atic units were activated only by items from domain A, while the remaining half were activated only by items from domain B. Thus any pair of items in the same domain had a small probability of overlapping in some of their systematic properties, while items from different domains never overlapped in their systematic properties. Arbitrary units were equally likely to be active for items from domain A versus B.

The model was fit in LENS (Rohde, 1999) using back-propagation to minimize cross-entropy error. The weights were adjusted with a learning rate of 0.1, using momentum ("Doug's" momentum = 0.9) and subject to weight decay (decay constant = 0.001). The model was trained 10 times to asymptotic performance with very low error over 1000 epochs. Prior to each training run, the model was initialized with random weights sampled from a uniform distribution in the range [-1,1]. These 10 models were used to generate data for 10 model "subjects," based on the patterns of activity elicited by each input over the whole network. Each model was presented with the 72 input patterns in sequence, and the pattern of activation elicited over the 98 units in the network (including the 22 irrelevant units, which always had an activation of zero) was recorded. The dataset for each model subject thus consisted of a matrix with 72 rows corresponding to stimulus items and 98 columns corresponding to model voxels. Each matrix contained the "true" response pattern for each subject to each item. To simulate noise in the measurement of this activity, a random value sampled independently from a Gaussian distribution with a mean of zero and standard deviation of 1 was added to each cell of the matrix. We take the resulting values in each cell of a matrix to be a model analog of the estimated BOLD response to a single stimulus at a single voxel in a single subject in an fMRI study.

To apply different brain-imaging methods to the discovery of structure, it is necessary to further stipulate the anatomical locations of the different units in the

model. In all simulations, input units were situated all together, with domain-A units neighboring one another, domain-B units neighboring one another, and arbitrary units neighboring one another. Output units were organized the same way, though outputs were assumed to be anatomically distal to inputs. The anatomical arrangement of input and output units was assumed to be identical across model individuals. For hidden units, we considered two different anatomical organizations. For anatomically localized models, units within a layer (SH, AH, or irrelevant) were also assumed to be anatomical neighbors, localized in the same way across model individuals. In the anatomically dispersed condition, units from the three hidden layers were assumed to be randomly intermingled with one another anatomically, in a different manner across model individuals. In either case, units in the hidden layers (together with irrelevant units) were assumed to be anatomically distal from both the input and output layers. For each anatomical variant the activation patterns evoked across model units by different inputs, and the ways these patterns were distorted by measurement noise, were identical—all that differed was the assumption about the spatial locations of the units in each layer.

3.5 Results

With this understanding of the model, we are now ready to consider how different statistical methods for fMRI fare at discovering the model units that encode representations of the two domains, both in the case where the hidden units are anatomically localized and when they are anatomically dispersed. The methods we consider include the standard univariate contrast method and four forms of multivariate pattern classification (MVPC). Each method faces the challenges inherent in fMRI analysis—that of finding meaningful signal within a vast amount of quite noisy data. To address the challenge, each method adopts a different set of

assumptions about the nature of the underlying signal, and so brings with it biases in the kinds of results it yields. For each method, we will begin with a brief exposition of the basic logic and essential concepts and will explicitly note the underlying representational assumptions. We then report the implementational details and results of the analysis, with the aim of answering four questions:

1. Does the method identify the systematic I/O units, but not arbitrary units, as important for domain representation?

2. Does the method identify the systematic hidden units, but not arbitrary units, as important for the domain representation?

3. Do the results differ when hidden units are anatomically localized versus dispersed?

4. Does the method indicate differences in how the information of interest is coded across unit sets? Specifically, does it indicate that some units respond more to A items than to B items, others show the reverse pattern, and still others express the A/B distinction with a distributed code?

Univariate contrast analysis

The univariate contrast analysis is the standard method for interrogating fMRI data. Its goal is to identify regions of cortex that, across subjects, exhibit systematically different mean BOLD responses to two (or more) different kinds of cognitive events. Typically the BOLD signal is spatially smoothed, so that the raw response at each voxel is replaced with a weighted average of the responses from anatomically neighboring voxels. The smoothed time-series is then modeled independently at each voxel for each subject using a deconvolution procedure. This yields a beta coefficient for each experiment condition at each voxel indicating how

well the measured BOLD signal matches the response expected if the activation of neurons within the voxel varies systematically with the experiment condition. The beta coefficients for each subject are projected into a common anatomical reference space, and univariate statistical tests are computed at each voxel independently to assess whether the coefficients differ reliably in the two experimental conditions across subjects. Voxels that show significantly different responses across subjects are viewed as important for coding the representation of interest.

A major challenge for the approach lies in establishing a meaningful criterion of significance in the context of tens or even hundreds of thousands of individual statistical tests. To avoid both false-positives and punishing corrections for multiple comparisons, it is common to seek ways of reducing the number of tests performed. Several different methods have been employed, but all rely on the idea that the representations of interest can be localized to particular cortical regions, and that the responses of voxels within a functional region will be largely similar. With these assumptions, the number of tests can be reduced by (1) conducting regions-of-interest analyses, where the responses of voxels within a ROI are averaged and the test is performed on the result mean response, (2) applying cluster-thresholding, where tests are only performed on clusters of n anatomically contiguous voxels all showing a similar response across subjects, or (3) applying a topographic control of the false-discovery rate. The univariate contrast method thus favors the discovery of clusters of anatomically neighboring voxels located in similar regions across individuals and showing similar response profiles across experimental conditions. From this brief description we can see that the method relies on five assumptions about the nature of the neuro-cognitive representations.

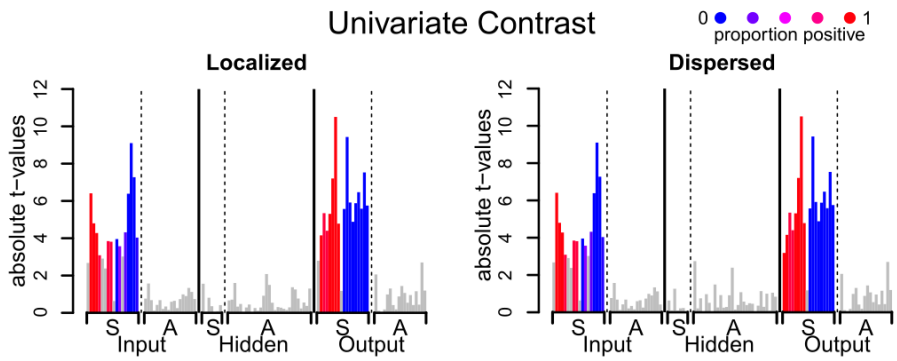

Figure 3.3: Results from the univariate analysis of simulated data. Bar height indicates the absolute value of the t statistic for the unit-wise contrast between conditions at the group level. Colored bars indicate units showing significant differences with p-values corrected to control the false discovery rate at $q < 0.05$. The red-blue scale indicates the direction of the contrast effect across model subjects, with red indicating units consistently showing greater activation for A items. S=systematic; A=arbitrary.

Implementation

The activity at each unit was modeled simultaneously for all subjects in a mixed effects model that treated subject as a random factor (G. Chen, Saad, Britton, Pine, & Cox, 2013; Friston, Stephan, Lund, Morcom, & Kiebel, 2005) using the lme4 package in R (Bates, 2007). Each model contains a single regressor, coding whether each item is an example of category A or B. The coefficients obtained from the mixed effects model were tested for significance using the Kenward-Roger approximation for the degrees of freedom (Kenward & Roger, 1997) and a standard F-test, numerator degrees of freedom = 1, denominator degrees of freedom = 9. The results are directly analogous to a repeated-measures ANOVA. The criterion for significance, alpha, is corrected to control the false discovery rate at q<0.05. The analysis was conducted for both the anatomically localized and the dispersed model. In both cases, the data were spatially smoothed, taking a weighted average over a three unit window, where the center unit was weighted about twice as much as the two flanking units.

Analysis

Figure 3.3 shows the results of applying the univariate method to the localized (left) and dispersed (right) models. In these plots, each bar corresponds to a single unit in the model. The bars are ordered according to their functional role in the network, as indicated by the X-axis labels. Colored bars indicate units showing statistically significant differences in mean activation across model individuals, while grey bars indicate units that did not show significant differences. Among the colored bars, red indicates units where activation was systematically higher for domain A across models, and blue indicates units where activation was systematically higher for domain B. Note that, in the anatomically dispersed plot (right), the units are shown in their standard functional location for ease of interpretation.

In both localized and dispersed cases, the univariate contrast method identifies systematic I/O units as important for representing the A/B distinction, and correctly indicates that different subsets of input units code this information differently (some responding more to A than B and others showing the opposite pattern). Note that these are the units for which the five univariate representational assumptions are all valid. In both localized and dispersed cases, however, the analysis completely misses the systematic hidden units, even though these jointly encode a cleaner representation of the A/B domain structure. The failure arises because, in both cases, the univariate assumptions are invalid. When hidden units are localized, assumptions 2 and 4 are violated: the way individual units encode information can vary across SH units in the same model individual, and across individuals at the same anatomical location. When the units are anatomically dispersed, assumption 3 is also invalid: the representation is coded in different anatomical locations across individuals. Because of these departures from the statistical assumptions, the mean activation of a unit at a given anatomical location across individuals does not differ reliably for SH units, even though these do reli-

ably encode the domain distinction in each individual.

Introduction of MVPA

The remaining methods we consider are all variants of multi-voxel pattern analysis (MVPA) that rely on pattern classification algorithms (Norman et al., 2006). Such approaches reverse the objective underlying univariate analyses: rather than using knowledge of the experimental design to explain variance in neural activity at individual voxels, MVPA uses the variance of neural activity across many voxels to make predictions about the experimental condition to which each trial, stimulus, or time point (henceforth, "example") belongs (Pereira, Mitchell, & Botvinick, 2009; Wang et al., 2004). To accomplish this, a classification algorithm is applied to a set of training data which include (a) the pattern of estimated activation evoked over a set of voxels for each of many examples and (b) a set of labels indicating the experimental condition or class associated with each pattern. For instance, in our model experiment, items in condition A might be labeled with a 0 while items from domain B are labeled with a 1. From the training data, the algorithm returns a pattern classifier—a statistical model that can be used to predict the label associated with any pattern of activation over voxels. Many different classification algorithms exist in the literature; as just one example, a logistic classifier will return a set of weights, one for each voxel, such that the estimated voxel activation, multiplied by its weight, summed over all voxels, and subject to a transformation function, yields a number that indicates the pattern label. In our example, a good logistic classifier should yield a number near 1 for all condition A items and near zero for all condition B items.

Even if there is no real signal at all in the data, it may be possible for a classifier to generate correct predictions for all items in the training set, especially when there are many predictors. Training set performance thus does not indicate

whether the classifier is exploiting real signal in the data. Instead, the classifier is typically assessed on a hold-out set: an additional set of examples and labels collected in the same experiment but excluded from the training data. The classifier learned from the training data is applied to patterns in the hold-out set, and for each pattern it generates a "guess" about the associated condition label. The classifier output is compared to the true label to get a measure of accuracy. If a model performs above chance at classifying the hold-out set, this indicates that it is likely exploiting real information in the data. To ensure that the results do not depend upon the particular items chosen for the training and hold-out sets, it is common to test a model using n-fold cross-validation. On each "fold" a subset of items is chosen for the hold-out set, and different hold-out sets are selected for different folds, such that, across folds, all items appear in exactly one hold-out set. Each hold-out set provides a measure of model classification accuracy, and this is usually averaged across folds to provide a single number indicating how accurately the trained model can classify hold-out patterns. We will refer to this number as the cross-validation accuracy of the classifier.

MVPA algorithms, like univariate analyses, are challenged by the abundance of data provided by fMRI, and so must adopt additional assumptions about the nature of the underlying signal. In any fMRI study (as in our model) there will always be more predictors (voxels) than things predicted (stimulus items or events), producing an over-fitting problem. In such cases, there exists no unique solution to the classification problem defined by the training set. Closed-form analyses are undefined, and other model-estimation procedures will produce a classifier that perfectly fits the training data without any guarantee of finding real signal. These problems can only be addressed by constraining the analysis based on an underlying hypothesis about how signal is truly encoded in the data. As with the univariate method, these constraints systematically affect the results. Each of the

remaining methods adopt different constraints to solve the over-fitting problem.

Searchlight pattern classification analysis

We begin with the well-known "searchlight" approach (Kriegeskorte et al., 2006), which was formulated specifically to address the challenge of finding distributed representations in brain imaging data. The method works as follows. Instead of training a classifier using all predictors at once, a separate classifier is trained for every individual voxel location in every individual subject. For each location, all voxels within a radius r of the center voxel are included as predictors in the classifier. This avoids the over-fitting problem by restricting the number of predictors included in any given classifier. The mean cross-validation accuracy for each classifier is stored in the searchlight center voxel, providing an information map for each subject. An univariate group-level analysis can be conducted on the information maps, similar in all respects to the analysis described in the previous section. Per the univariate assumptions, this means that each point in the accuracy map is considered independent of all others. However, within each searchlight, the effect of a given unit on the classification can differ depending upon the activations of other units in the searchlight, and these units can respond to various stimuli in quite different ways.

Thus the searchlight method relaxes assumptions about the consistency of the neural code within and across individuals, and about the independence of representational units, but retains assumptions about localization of information within and across individuals. What results are observed in the model with these differing representational assumptions?

Implementation

The searchlight analysis was conducted using the SearchMight toolbox (Pereira & Botvinick, 2011) for MATLAB (The Mathworks). The input units, hidden units, and output units were treated as three anatomically separated regions so that a searchlight never encompassed units in different regions. This was accomplished by inserting empty units between the layers, and providing a mask to SearchMight to omit those units during analysis while ensuring that no searchlight spans multiple regions. Within each searchlight, a Gaussian Naive Bayes (GNB) classifier was fit to distinguish between category A and B items. Although GNB classifiers are limited in some ways (Pereira & Botvinick, 2011), the concerns do not apply to this simple and idealized case where noise is truly identically and independently distributed (i.i.d.) with uniform variance. The amount of category information in each searchlight was estimated through 6-fold cross validation; the mean cross-validation accuracy was stored at each searchlight center; and the mean accuracy over model subjects was then computed for each unit and tested to see if it differed significantly from chance. The resulting map of p-values is FDR corrected, $q < 0.05$.

As previously, the analysis was performed on both the anatomically localized and the anatomically dispersed arrangements of units. The data were not smoothed prior to the searchlight analysis. The analysis was performed with various searchlight sizes, ranging from 3 to 28. A searchlight size of 7 or 9 should be roughly "optimal" given the size of the clusters of informative units in the localized data.

Analysis

Figure 3.4 shows the results of the searchlight analyses, for localized and dispersed model architectures, and for different searchlight sizes. The format is the same as in the preceding analysis, except the y-axis now indicates the mean classification

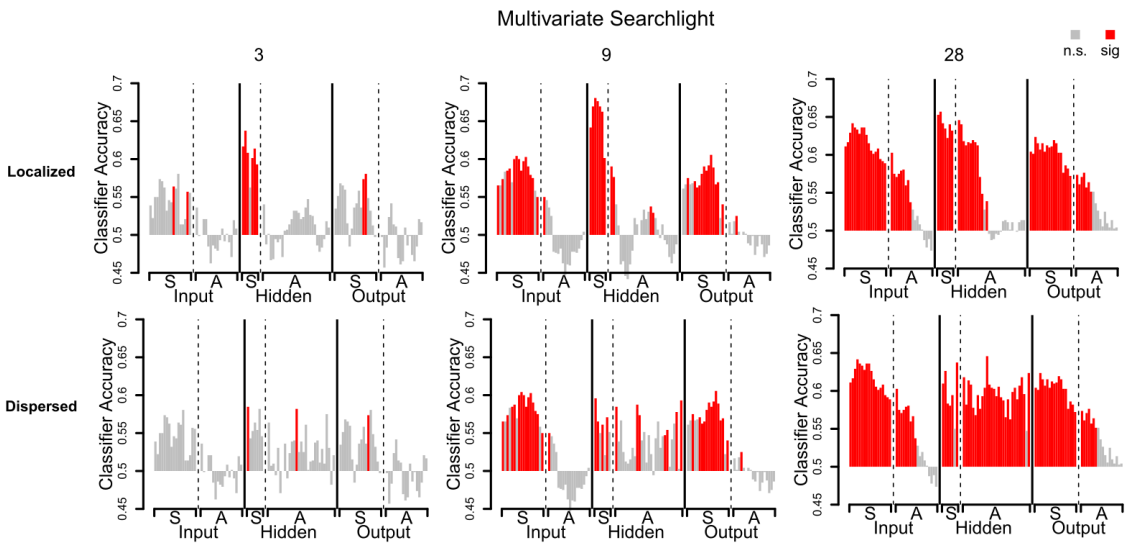

Figure 3.4: Result of the multivariate searchlight analysis of simulated data. Bar height indicates the mean classifier accuracy over subjects. Bars in red indicate searchlights where classification accuracy differed from chance over subjects, p-values corrected to control the false discovery rate at $q < 0.05$. Each column shows results from a different searchlight size indicated by the number on the far right. In the row labeled *Localized*, units were clustered by kind during the searchlight analysis; in the row labeled *Dispersed*, the systematic and arbitrary hidden units were shuffled together. S=systematic; A=arbitrary.

accuracy for searchlights centered on each unit, rather than a t-value at each unit. As before, colored bars indicate units that the method identifies as statistically significant—that is, units whose surrounding searchlights show classification accuracy reliably above chance across model individuals.

There are several points to note in these plots. First, when the hidden representations are anatomically localized, the method can do a quite good job of identifying both the systematic I/O and the systematic hidden units as important for the domain representations, though for both unit types the results vary substantially with the searchlight size. When the searchlight is small, the method reliably finds the SH units but misses most of the systematic I/O units. This happens because, as noted earlier, the SH units encode a clearer differentiation between domains, so that even if the searchlight does not encompass all 7 units there is sufficient information within it to classify stimuli above chance. The domain distinction is weaker

in the systematic-I/O units, so when only a small number of such units fall within the searchlight, there is insufficient information to classify correctly. With a larger searchlight (9 units), the method does very well at finding all relevant units. When it grows too large, however, it begins to incorrectly flag irrelevant units as being important for the representation (28 units). A very large searchlight, even when centered on an uninformative (arbitrary) unit, can have a broad enough span that it encompasses other informative units. In this case the classifier will perform well by virtue of the informative units appearing in the edge of the searchlight, but the above chance result will be "stored" in the searchlight center, making it appear as though there is useful information present at that location. Thus when the signal is anatomically localized, there is a tradeoff between searchlight size and discovery of representational structure, with searchlights that are too small missing weaker signal and those that are too large incorrectly flagging arbitrary signal.

In the model case, where we know a priori which are the signal-bearing units, it is easy to discern the optimal searchlight size, but it is less clear how this would be determined from real brain imaging data. One might initially expect the optimal searchlight to be identifiable from the accuracy of the resulting classifiers, but Figure 3.4 suggests that this is not the case: the very large searchlight, which flags many irrelevant units, shows almost as good classifier performance as the optimal size at the SH units and better performance at the systematic I/O units. If we did not already know which units were important for representation in the model, it would be difficult to know which searchlight size to choose, and hence which results to believe.

The second thing to note is that the searchlight analysis does a much poorer job overall of identifying the SH units when these are anatomically dispersed (right panels of Figure 3.4). The poor performance arises because the precise anatomical location of the signal-carrying units is assumed to vary across individuals in this

case. Within any individual model, a searchlight that includes a few of the informative units will show above-chance performance in classification, but the searchlight centers will differ across model individuals, especially when the searchlights are small. Thus the cross-subject statistical test at each location will yield a null result, leading to poor signal discovery. Larger searchlights will be more likely to contain the signal-carrying units, but also lead to poorer localization of the signal as already noted.

In sum, the method deals with the over-fitting problem by only including a small number of contiguous voxels in each classifier—an approach which assumes that useful representational structure can be localized within the searchlight radius, in the same locations across subjects. When these assumptions are met, the approach does a good job of discovering representational structure, even if the representational code (i.e., the way that individual units respond to particular stimuli) is highly variable within and across individuals. The limitations noted above arise when the assumptions are violated—when representational structure is anatomically distributed across multiple searchlights (as when searchlights are too small in the localized case), or in different ways across individuals (as in the dispersed model). Moreover, whether the assumptions are met depends, not only upon the anatomical distribution of the signal, but also upon the searchlight size, and it is not clear how the latter can be optimized for real brain imaging data.

Finally, it is worth noting that, in contrast to the univariate method, the searchlight approach does not provide information about how the contrast of interest in encoded in unit activations. Thus there is no way for the method to show, for instance, that there are some units systematically more active for A items than B items, others showing the reverse pattern, and still others that express the A/B distinction in a distributed code (the SH units).

Regularized logistic regression for whole-brain pattern classification

The limitations of the searchlight method arise from the relationship between a searchlight's field of view and the anatomical distribution of the underlying signal. A small searchlight provides better localization but is more likely to exclude signal-carrying units; a large searchlight is more likely to include the signal-carrying units but provides less information about where the signal really is.

An alternative approach that avoids this trade-off is to train a pattern classifier on the whole dataset simultaneously. While many classification algorithms exist, we focus here on variants of logistic regression because they are easily interpretable, powerful, and draw upon intuitions formed through experience with linear regression. A regression model is composed of a set of weights β_x, one for each predictor variable x plus an additional intercept term, tuned to make accurate predictions about a response variable y. In logistic multivariate pattern classification, the predictor variables are the voxels, and the response is a binary variable that codes class or condition label. For instance, in a contrast of conditions A and B, A events are labeled with $y = 1$ and B events are labeled with $y = 0$. To generate a prediction for a given item, the logistic regression model takes the weighted sum of the estimated response over voxels and passes it through a squashing function bounded at 0 and 1:

$$f(z) = \frac{e^z}{1 + e^z} \tag{3.1}$$

where $z = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_n X_{in} + \epsilon_i$—which is the model's linear response to a particular pattern of activity. Thus, $f(z)$ is a transformation of the weighted sum of predictor values expressing the probability that y=1 given the pattern of activity for the i^{th} item. Fitting a logistic regression model involves

finding coefficients that minimize the discrepancy between the true labels in $y \in 0, 1$ and the probabilities assigned by the model. This is typically measured by the logistic loss:

$$\arg\min_{\beta} \sum_{i=1}^{n} \log(1 + e^{-\bar{y}_i X_i \beta}) \tag{3.2}$$

where $\bar{y_i}$ is -1 when $y_i = 0$ and $+1$ when $y_i = 1$. This loss is minimized when the sign of the model's linear response $X_i\beta$ is positive for items labeled $y = 1$ and negative for items labeled $y = 0$.

As noted previously, the problem is that there are infinite possible solutions to the minimization when there are more predictors than items. One needs a way of deciding which among these is most likely to uncover the real signal. Regularized regression provides one way of doing this. Such approaches seek to jointly minimize the prediction error plus an additional cost, itself a function of the coefficients:

$$\arg\min_{\beta} \sum_{i=1}^{n} \log(1 + e^{-\bar{y}_i X_i \beta}) + \lambda h(\beta) \tag{3.3}$$

The additional penalty or *regularizer* represented by $h(\beta)$ prioritizes some model solutions over others, and in this way embodies a hypothesis about the nature of the true underlying signal. The constant λ is a free parameter that controls the degree to which the two terms (prediction error versus minimization of the regularizer) should be weighted in the joint optimization.

We here consider two varieties of regularized logistic regression recently employed in the fMRI literature: *LASSO* (Rish et al., 2012; Tibshirani, 1996) and *ridge regression* (Hoerl & Kennard, 1970/2000; Riggall & Postle, 2012). Though superficially similar, the two methods embody different implicit assumptions about the nature of the underlying signal and so yield quite different results. In LASSO, the

regularizer is the sum of the absolute values of the model coefficients:

$$h(\beta) = \sum_{j=1}^{m} |\beta_j| \tag{3.4}$$

For ridge regression, the penalty is the sum of their squared values:

$$h(\beta) = \sum_{j=1}^{m} \beta_j^2 \tag{3.5}$$

In both cases, the optimization is convex: for a given value of λ, there exists a unique set of coefficients that minimize the cost and that can be efficiently discovered by gradient descent. Yet the different penalties lead to quite different solutions. To understand why, it is useful to consider how they treat sets of predictors that covary together. Imagine four voxels whose responses across items are perfectly correlated, and suppose their activations are useful in predicting the condition label. In this scenario, there are many different ways of placing weights over the four voxels that will all have the same effect on the classifier output. For instance, placing a weight of 1 on each voxel will have exactly the same effect as placing a weight of 4 on one voxel and a weight of 0 on the other three. Because the voxel activations are perfectly correlated, and the classifier operates on a weighted sum over voxel activations, these different weight configurations have the same effect on the model output and hence on the prediction error. The regularization penalty, however, should prefer some weight configurations over others.

If the data really are perfectly correlated, the LASSO penalty won't be any help: the sum of the absolute value of the coefficients is the same for models that place a 1 on each unit versus models that put a 4 on one unit and zeroes on the rest. If we imagine, however, that all measurements are subject to some independent noise, the scenario is a bit different. In this case, one of the 4 units will, just by chance, covary slightly better with the category labels. In this case, the classifier can do a

slightly better job of minimizing the error term by loading up all of the weight on this single voxel. Thus the joint optimization will lead to a solution where just one (or perhaps a few) of the redundant voxels are selected.

Ridge regression behaves very differently. Here the penalty scales exponentially as weights increase on a single voxel, but only linearly as weights are added across voxels. Thus the penalty is minimized by placing small weights on many voxels. In the preceding example, placing a weight of 4 on one unit and 0 on the remaining three leads to a total penalty of 16 over the four units. Placing a weight of 1 on each unit, in contrast, leads to a penalty of 4. Ridge regression thus prefers solutions where small weights are "spread out" over redundant predictors. In fact, as the weight approaches zero, the ridge penalty becomes vanishingly small, so with a finite number of training examples, ridge regression will always place at least a tiny weight on every predictor. In real data, of course, voxel states are never perfectly correlated nor perfectly informative about the condition label, so the behaviors of the two approaches are less easy to intuit. In general, however, it is useful to think of LASSO as minimizing prediction error with the fewest possible predictors (i.e., as many zero coefficients as possible), while ridge regression can be viewed as "spreading" small weights over all predictors exhibiting any systematic relationship with the category labels, without care for the number of predictors.

All assumptions are relaxed relative to the univariate and searchlight methods. However, this does not mean that they are assumption-free. To the contrary, each approach entails additional assumptions about the *sparsity* and *redundancy* of the underlying signal:

- *Sparsity*: LASSO assumes the signal to be sparse, in that only a small proportion of voxels are involved in coding the information of interest. In this case, the best approach to finding true signal is to minimize prediction error using the smallest number of predictors possible. Ridge regression makes

no sparsity assumption.

- *Redundancy*: Ridge regression assumes that the signal is highly redundant, so that many voxels express essentially the same information. In this case the best approach to finding true signal is to minimize prediction error using distributions of weights that are as close to zero as possible, so that all informative predictors are included in the solution. LASSO assumes that the underlying signal is not highly redundant, so that different predictors carry different information.

With these assumptions, what signal do LASSO and ridge regression detect in the model?

Implementation

Logistic LASSO and ridge regression were conducted using GLMNET (Friedman, Hastie, & Tibshirani, 2010) in MATLAB (The Mathworks). Both methods have a free parameter λ that controls the importance of the regularization penalty relative to the prediction error, leading to greater sparsity in LASSO and more severe weight shrinkage in ridge regression. The analysis thus proceeded in two steps: one to estimate a useful λ for each subject, and a second to fit a model at the estimated λ and evaluate it on a hold-out set. The data for each model subject was first divided into 6 equal parts, each containing the same number of category A and B items. One part was set aside and the remaining 5 were passed to a function that conducted a 5-fold cross validation accuracy test at 100 values of λ. The function returns the λ producing the highest cross-validation accuracy, which is subsequently used to fit a model to all 5 parts of the data. The resulting model was then assessed on the original hold-out set (the 6th part). This procedure was carried out separately for all 10 model subjects, in both localized and anatomi-

cally dispersed model variants, for both LASSO and ridge regression. For each model subject, each method returns a vector of coefficients that indicates how the classifier interprets each unit's activation in generating a predicted class label. To understand which units contribute to the representation of interest and how these units encode information, the coefficients must be interpreted. A key difference between methods lies in the ease of interpretation. We will therefore consider results from the two methods separately, before contrasting them.

Analysis

For LASSO, interpretation of the classifier coefficients is straightforward: the method places zero weights on as many predictors as possible, so any unit receiving a non-zero weight can be viewed as having been "selected" by the classifier as important. If the classifier shows above-chance cross-validation accuracy, we can be certain that it has successfully identified signal-carrying units. In contrast to the preceding methods, there is no statistical test performed on a null hypothesis at each unit. Instead, any selected unit in a given model individual can be viewed as "significant" for the representation because, if it could be discarded without affecting classifier performance, LASSO would have done so. In this sense, for each subject, the method can be viewed as finding the smallest sufficient set for classification. The central questions then are (1) how well does the selected set pick out the signal-carrying voxels in I/O and hidden layers, (2) do the classifier weights indicate differences in how information is encoded for different voxel sets and (3) do the results differ for localized versus dispersed model variants?

Figure 3.5A shows how often LASSO selected each unit across the 10 model subjects for localized versus dispersed cases. To get a sense of which units were selected more often than expected by chance, we took the overall proportion of units selected across subjects as a base probability for conducting a binomial test

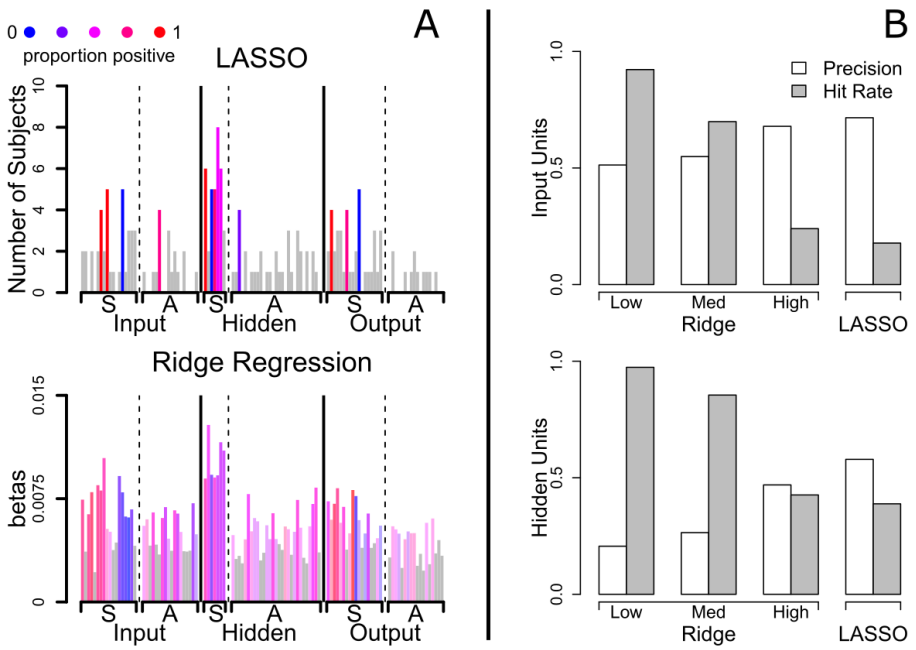

Figure 3.5: A) The results from the LASSO and ridge regression analyses. The blueness or redness of the bar conveys the frequency with which each unit was assigned a positive weight over subjects. Positive weights mean that activation at that unit will push the model towards labeling the current item as belonging to domain A. *LASSO*: Grey bars were selected less often than expected by chance given the overall rate of unit selection. *Ridge*: Bar saturation indicates which units would count as "selected" under three different policies, based on weight magnitude. Bright bars are in the top third of the distribution, pale bars are in the middle third, and gray bars are in the bottom third. S=systematic; A=arbitrary. B) Hit rate and precision for each regularization method, computed across the whole network (top), the I/O units only (middle), and the hidden units only (bottom)

at each unit.[2] Colored bars indicate units that were selected more frequently than expected if LASSO was choosing at random with this base rate, without correction for multiple comparisons. From this plot, the approach does a fairly good job of identifying the SH units, reliably tagging 5 of the 7 units (71%). The approach did less well discovering the systematic I/O units, reliably identifying only 6/36 (17%). This difference reflects the fidelity of the representations coded across different unit sets: as already noted, the 7 SH units encode the cleanest representation of the domain distinction, and so are more likely to be included in the smallest sufficient

[2]This binomial test for significance is not recommended for real analyses, but fits the need in this very simple context. This foreshadows the complexity of making statistical inferences based on weights obtained through regularized regression.

set for any individual. Also, note that the results are identical for localized versus dispersed cases. Since LASSO is conducted separately for each individual and is blind to anatomical structure, the results are literally identical regardless of how the units are spatially arranged.

This summary plot is misleading in one sense, however, since it applies an aggregate statistical test across model individuals to assess which units are reliably discovered. Such a test would not be possible with real data, since it would not be clear which voxels should be "lined up" across subjects to compute the binomial probabilities. The virtue of LASSO (and ridge regression) is that they are essentially single-subject analyses, and so are freed from assumptions about consistency in location and coding across individuals. What we really wish to know is how accurately the solution picks out the units of interest for each individual model. For every model individual, from the binary classification of selected versus unselected units, we can compute two numbers that jointly describe how well the solution identifies the important units. Specifically, we compute the hit rate, which is the proportion of actual signal-carrying units identified by the algorithm, and the precision, which indicates what proportion of the selected units are true signal-carrying voxels. Moreover, these figures can be tallied for just the I/O units, just the hidden units, and for the whole network, to provide an indication of how well the method singles out informative units in these different sets.

Figure 3.5B shows the mean of these figures across model individuals for LASSO (and other methods). The general pattern is clear: precision is relatively high, indicating that most of the units LASSO identifies are indeed signal-carrying units. LASSO is sub-optimal, however, in the hit rate: for the hidden layer, about half the important units are missed, while the great majority of signal-carrying units are missed in the I/O layer. Thus if LASSO selects a unit, one can have confidence that it does carry useful information, but one cannot have confidence that it has

discovered all the useful elements.

Finally, we can ask how well LASSO uncovers differences in the representational code. The red-blue spectrum of the colored bars in Figure 3.5A indicates the frequency with which each unit receives a positive weight across model subjects. Red and blue colors indicate that a unit's activation receives the same interpretation across model subjects, while shades in between indicate that the interpretation varies. Figure 3.5A shows that, where LASSO does identify systematic I/O units, it also reveals the correct code: all units are red or blue.[3] There are so few units identified, however, it is difficult to "see" the systematic layout of these responses. In the SH layer, LASSO correctly indicates that code can vary across individuals for some units, though it also appears to show consistent category-selective responses for some units. These differences arise because LASSO does not succeed in selecting all SH units in every model individual. Instead, each unit is identified in about half of the individuals. When the algorithm selects a unit in a small set of participants, all of whom happen to have acquired the same code, the selected unit appears to show a selective code.

The ridge regression classifier showed marginally better cross-validation accuracy (0.65 compared to 0.6 for LASSO), but with a very different distribution of weights. In fact, ridge regression placed a non-zero weight on every unit—effectively using the whole pattern of activation across all units in the network. Consequently it is difficult to know which predictors are playing an important role in the classifier behavior and which are not. Weight size (i.e., absolute value of a weight) provides one indicator of predictor importance, since the regularization penalty tries to keep weights as close to zero as possible. Any predictor receiving a weight that deviates strongly from zero must, therefore, be important for

[3] In general, interpreting the weights obtained by LASSO and similar methods is not necessarily as transparent as it is in these simple models (c.f. Haufe et al., 2014). See Section 3.6 (the discussion of this chapter) for more on this issue.

reducing prediction error. But this relationship is not perfectly transparent. Consider the case where a single unit carries important information for classifying one subset of items, while ten highly redundant units all carry information important for classifying another subset. In some sense all 11 units are equally informative for the classifier, but ridge regression will place a large weight on the singleton unit and many small weights across the ten redundant units. That is, the weight size under ridge is sensitive to both the informativeness of the unit activation and its redundancy with other units. Highly redundant units can receive quite small weights even if they carry useful information. For these reasons it is not clear, in the model and more so in real data, just how strong or weak a weight must be to "count" as having been selected by the classifier.

Figure 3.5A illustrates these points by showing the mean, over model subjects, of the absolute value of the classifier weight at each unit. It is clear that signal-carrying units receive somewhat stronger weights than the arbitrary units overall. It is also clear that the SH units receive stronger weights on average than do the systematic I/O units, reflecting their greater utility in reducing prediction error. Intuitively, one wants to draw a threshold below which units are classified as irrelevant, but it is not clear how the threshold is to be selected. The differences in weight magnitude are not large, and there is no a-priori basis for deciding how small a weight should be in order to conclude that it is not useful. Yet the conclusions one draws about where the signal is encoded can vary fairly dramatically depending upon this decision. The intensity of the shading in the ridge regression subplot in Figure 3.5A indicates which units would be "selected" under different thresholding policies. With a very strict policy (discarding 66% of the units), the representation would appear to reside mainly within the SH units. With a more lax policy (discarding 33% of the units) it would appear to be very broadly distributed over many units.

As with LASSO, the aggregate plot is somewhat misleading, since the different selection policies operate on mean coefficient values that could not be calculated in real data unless representations were localized identically across individuals. We therefore conducted the same analysis of hit rates and precision values across model individuals, adopting three different policies for discarding small weight values. In the lax policy, the 21 units (20%) with the weakest classifier weights were deemed unselected; in the moderate policy, half of the weights were discarded; and in the aggressive policy, only the 21 units with the strongest classifier weights were retained. For each policy we computed hit rates and precision, for I/O units, hidden units, and all units. The results are included in Figure 3.5B. When the policy is lax, the pattern is opposite to that observed in LASSO: relatively high hit rates but low precision for all unit subsets, indicating that the method has incorrectly selected many arbitrary units. When the policy is aggressive, the pattern is similar to LASSO: low hit rates but relatively high precision, especially for the SH units. Thus the accuracy with which the classifier weights pick out the signal-carrying units varies dramatically depending on the arbitrary selection of a weight threshold. In the model we can, in principle, discover an optimal thresholding policy—one that maximizes hit rate and precision—but only because we already know the ground truth. With real data, where the number of predictors is much larger, the representational structure likely to be much more complex, and with no knowledge of the ground truth, it is not clear how the set of weights discovered by ridge regression might be used to discover where the useful signal is coming from.

Finally, does ridge regression provide useful information about the different nature of the representational code at different units? As previously, the hue of the colored bars in Figure 3.5B indicate the frequency with which a unit receives a positive weight across model subjects. Both the independent coding of domain in I/O units and the variable nature of the code across subjects at the SH units

come across fairly clearly. Thus the method does a reasonable job of highlighting differences in the representational code across these units subsets. The chief problem with the approach is the difficulty it poses in understanding which units are contributing meaningfully to the classification.

3.6 Discussion

The PDP approach, which has made important theoretical contributions to cognitive science, adopts specific claims about the nature of mental representations. Despite this usefulness for understanding many aspects of cognition, representations of this kind have been difficult to empirically test in functional brain imaging studies. PDP suggests that mental representations are patterns of activation distributed over neural populations, with information encoded, not in the activity of individual cortical units taken independently, but by the complete pattern coded over a representational ensemble. It further suggests that the response of any given element in an ensemble may vary arbitrarily across individuals, even if the ensemble jointly codes the same structure across individuals; that elements of an ensemble need not all be anatomical neighbors within an individual; and that their location across individuals, while not completely random, may be subject to at least some variability. These assumptions about representation, paired with the observation that functional brain imaging technologies yield up vast amounts of quite noisy data, present serious challenges to discovery of the neural bases of mental representation.

We have now reviewed four different methods for analyzing brain imaging data, considered their underlying assumptions, and assessed their ability to discover distributed representations of the kind PDP assumes. From this analysis it is clear that very different results arise depending upon the statistical method em-

ployed or, in some cases, upon parameterization of the method in question. Each method succeeds best when the implicit assumptions it adopts are met in the data. Thus univariate contrast successfully identifies the systematic I/O units, which conform to the assumptions listed in the corresponding column of 1. The SH units, despite encoding cleaner domain representations, violate all of these assumptions and so are completely missed. Searchlight does well at detecting both systematic I/O and SH units, but only if the useful information is contained within the radius of a relatively small searchlight and is localized in the same way across individuals. LASSO performs well at discovering SH units, but in assuming no consistency across model individuals and no redundancy within individuals, becomes highly susceptible to noise and so misses many important units. Ridge regression, in assuming highly redundant signal without care for overall sparsity, spreads weights over all units, making the solution hard to decipher even if the classifier performs well.

However, saying the LASSO did well at recovering the SH units comes with a very important caveat. In any given subject (i.e., model run), only a subset of the SH units were discovered. Counting across subjects, the SH voxels stand out, but the degree to which those voxels stand out may be somewhat exaggerated by them being plotted as a neighbors in the figure. In reality, the voxels that LASSO discovers may not be neighbors, and certainly the number of voxels in an MRI dataset is orders of magnitude larger than the number of units in these simulations. This means there is more opportunity for false alarms, and the degree of overlaps across subjects may be diminished. The ruthless sparsity of LASSO, therefore, can result in solutions that are much less easily interpreted than those obtained through the analysis of these toy models.

Improving on LASSO

It is clear, then, that while regularized regression holds great promise for testing hypotheses about the representations that support different mental content and cognitive processes, neither is ideal. The limitations of LASSO and ridge regression have inspired a variety of methods, such as elastic net (which combines the ℓ_1 and ℓ_2 norms into a single model; Zou and Hastie, 2005), group LASSO (which allows the researcher to specify groups of features that are thought to be related a priori and exert some control over the sparse structure that is obtained; Simon, Friedman, Hastie, and Tibshirani, 2013; Yuan and Lin, 2007) and multi-task group LASSO methods such as sparse overlapping sets (SOS) LASSO (which was designed to encourage discovering sparse models that can leverage course spatial structure within and across fMRI datasets while retaining the benefits of LASSO; Rao et al., 2013), and the ordered weighted ℓ_1 (OWL) regularized regression(Figueiredo & Nowak, 2014; Figueiredo & Nowak, 2016; Zeng & Figueiredo, 2014). These techniques are different attempts to obtain more complete, interpretable models when there are highly correlated features. Each of the cited methods attempts to group features that are highly correlated or related, rather than assigning a weight to every feature (as in ridge regression) or seeking solutions that sample as few units as possible from sets of correlated features (as in LASSO). With group LASSO, features are assigned to groups a priori—grouping affects the optimization, but the optimization does not identify the groups. The researcher must approach the dataset with a hypothesis about which features should be grouped. This is not ideal in all cases. Elastic net and OWL, on the other hand, identify correlated sets of features as part of the optimization. While similar in this sense, OWL differs from elastic net in that it more explicitly groups variables, by assigning voxels deems to belong to the same group exactly the same weight (Figueiredo & Nowak, 2016). This means that OWL defines groups in transparently "human

readable" way, while elastic net allows correlated features to be assigned different weights, which obscures the group structure. In addition to producing more interpretable models, OWL has been shown to outperform elastic net in terms of how well it selects and groups features (Bondell & Reich, 2008; Zhong & Kwok, 2012)[4].

In these ways, OWL reflects the state of the art among regularized regression techniques. It will be reintroduced in the next chapter, as part of the optimization that drives network RSA, which will be critical to the experiments reported in Chapter 5.

Interpreting model weights in regularized regression

Interpreting sparse models of data like MRI datasets, characterized by tens of thousands of features that may be correlated with each other to various degrees and very few training examples, faces a complication we have not yet addressed. Intuitively, it seems that if including a voxel in the model reduces its prediction error, then the voxel must carry some information about the cognitive process under study. This is not necessarily the case, however. When voxels are correlated, they may be correlated in a way that is unrelated to the process of interest. Given two features with correlated noise, let one of the features participate in the process of interest and the other be completely unrelated to the process. A linear model can combine these two features to effectively cancel out the noise (Haufe et al., 2014). A model that is able to perform above chance on an out of sample prediction task has certainly discovered some voxels of interest, but it is likely that a subset of the voxels included in the model are "merely" being used to boost the signal of other voxels that are actually driving the model's performance. Because the noise applied to the simulated data above was i.i.d., and the overall number of voxels was

[4]Bondell and Reich (2008) and Zhong and Kwok (2012) performed their experiments using OSCAR, which is a special case of OWL.

small, correlated noise was not an issue, which made the models reported above easy to interpret; it will not be so in general.

One way to address this concern is to perform follow up analyses on subsets of voxels. Consider the following scenario: you have a group of participants make simple value judgments about faces and places. You then analyze these data using LASSO to learn a sparse model of each subjects data that can classify whether they are looking at a face or a house on any given trial. You identify voxels in the standard face and place regions, as well as many voxels in less typical places. Perhaps face and place judgments recruit a widely distributed neural network, or perhaps these other scattered voxels happen to have noise components that correlate with voxels in the standard areas. They are boosting the signal, but not contributing new face or place information. These two possibilities can be adjudicated by a follow up analysis, where the standard areas are "lesioned"—simply dropped from the fMRI dataset before refitting LASSO to the data. If this new model on the lesioned dataset can still perform above chance, and selects voxels in similar places as seen in the "intact" brain analysis, it suggests that neural regions beyond the standard areas really do support discrimination between judgments about faces or places.

It would be incorrect, however, to claim strongly that other areas of the brain are functionally independent of those that were "lesioned" from the dataset. To the extent that other areas of the brain are functionally connected with the "lesioned" areas, and given that when the data were collected the brain was healthy and intact, it is possible that these other areas express the representations that they do only because of the proper function of the areas we later omit from the models. While this certainly lacks an element of experimental control, we can at least say that these widespread areas, beyond the typically defined face-system, are adopting a state that distinguishes faces from other objects. This may because they are

central to face processing, or that they are ancillary to conceptual processing that is also relevant to the face stimuli in our task. But, nevertheless, an argument that these areas are functionally connected to the typically defined face system is interesting and suggests a wider base of relevant contributions to the processing of these stimuli than under the standard account.

Which is the best method to use?

This question is posed somewhat facetiously—the intention is not to advocate for one method to the exclusion of others. Rather, the point we wish to emphasize is that each statistical method is closely associated with an implicit hypothesis about what matters in neural signal. The question of what matters—what "makes" a pattern of activation over neurons a mental representation—is itself a central question, maybe the central question, of cognitive neuroscience. Each of the approaches we have considered begins with an implicit hypothesis about the answer to this question. The univariate contrast method begins with the hypothesis that consistency in location and neural response across individuals is what really matters to the discovery of representation. Searchlight begins with the hypothesis that what matters is the similarity of evoked responses over neural populations within a particular contiguous region of cortex. LASSO begins with the hypothesis that representations are sparse, are not highly redundant, and can be localized any which way. Ridge regression begins with the hypothesis that representations are highly redundant, but still can be localized any which way.

Like any hypothesis, each of these is potentially useful in guiding discovery, but is also subject to empirical assessment. From this point of view, brain imaging might best proceed, not by choosing the "best" method, but by comparing and contrasting the results obtained across different methods. With a good understanding of the underlying assumptions each method adopts, and consequently of

its blind spots, one may arrive at a better understanding of the neural code than any method individually provides. In the current work, for instance, the univariate contrast method excelled at identifying systematic I/O units but failed to find SH units. LASSO showed the reverse pattern. The juxtaposition of the results provides a fuller picture of how the network represents domain structure than does either method on its own. One could imagine conducting a univariate contrast analysis, masking out all voxels that show reliable effects, then conducting a whole-brain multivariate analysis on the remaining voxels to assess if there exists a distributed code. In general, it is not difficult to imagine how the different methods could be used in combination to better understand the neural basis of mental representations.

It is also worth noting, however, that the hypotheses underlying these different methods, like any hypothesis, must also play a coherent role within a broader theoretical account of the mechanisms that support the behavior of interest. In the case of the univariate contrast method, for instance, it is not generally sufficient simply to state that what matters is consistency in location and neural response across subjects. One wants some broader explanation as to why cross-subject consistency is an important indicator of the underlying representation. This is where we feel the connection to the PDP framework for cognition offers some real utility. The representational assumptions the approach adopts are not stipulated arbitrarily. They arise from a coherent set of principles stemming from neuroscience, computer science, and cognitive psychology that have proven useful for understanding a very broad range of behaviors in developing, healthy, and disordered populations. The hypothesis that mental representations are distributed in the particular way we have explored is, in this sense, deeply theoretically motivated. Statistical methods that begin with this hypothesis thus have a built-in theoretical justification that other methods may lack. What has been lacking, and what the new generation of

statistical methods—including the methods reviewed here and other approaches beginning to emerge in the literature—is a good understanding of how to look.

Chapter 4

Network Representational Similarity Analysis

In the previous chapters, I took a broad view on the challenges associated with testing the hypothesis that the brain utilizes distributed representations using readily available functional neuroimaging methods (primarity, fMRI), and made a case that testing this hypothesis is of great significance to the field. In Chapter 2, I reviewed a collection of fMRI literature that, taken as a whole, provide evidence that four defining qualities of distributed representation do appear to have a basis in the human brain. This review leaves one with an appreciation for the representational complexity of distributed representations, and with a desire for a more ideal means of identifying them. In Chapter 3, I compared several methods with respect to a simulated MRI dataset containing both localized, "nameable", functionally specified elements, and collections of elements that encode content via distributed representations. Here, we saw that distributed representations are fundamentally multivariate, and univariate techniques may overlook them even when the representations themselves are localized. Therefore, searchlight MVPA is a fine solution when the distributed representations are localized. When representations are dis-

tributed in space, and intermixed with other irrelevant representational elements, the searchlight method is less suited to the task. This led us to consider regularized regression, which eliminates the spatial constraint of the searchlight and may therefore be a more appropriate tool for studying distributed representations.

However, regularized regression is not a single method, but a whole class of methods. Regularization involves making additional assumptions about the structure of signal within the dataset, and models derived via regularized regression are in general much more complicated to interpret than standard regression. Some regularization techniques, such as applying the ℓ_2 norm as in ridge regression, make no attempt to yield an interpretable model: the objective is to make accurate predictions. This is not useful for neuroscientific investigation of where or how content is represented—it can only tell you if a certain prediction problem can be solved based on some linear combination of the available information. Others, such as applying the ℓ_1 norm as in LASSO, will identify a sparse subset of features which share as little variance as possible—the model pressured to not include redundant information because of how it is penalized. Further, the sparse model that is obtained can be difficult to interpret for the same reasons that weights returned by ridge regression are difficult to interpret. That all said, LASSO is more informative from a neuroscientific perspective than ridge regression. But it is clearly limited.

These limitations have inspired a variety of methods, such as elastic net (which combines the ℓ_1 and ℓ_2 norms into a single model; Zou and Hastie, 2005), group LASSO (which allows the researcher to specify groups of features that are thought to be related a priori and exert some control over the sparse structure that is obtained; Simon et al., 2013; Yuan and Lin, 2007) and multitask group LASSO methods such as sparse overlapping sets (SOS) LASSO (which was designed to encourage discovering sparse models that can leverage course spatial structure within

and across fMRI datasets while retaining the benefits of LASSO; Rao et al., 2013), and the ordered weighted ℓ_1 (OWL) regularized regression(Figueiredo & Nowak, 2014; Figueiredo & Nowak, 2016; Zeng & Figueiredo, 2014). These techniques are different attempts to obtain more complete, interpretable models when there are highly correlated features. Each of the cited methods attempts to group features that are highly correlated or related, rather than assigning a weight to every feature (as in ridge regression) or seeking solutions that sample as few units as possible from sets of correlated features (as in LASSO). With group LASSO, features are assigned to groups a priori—grouping affects the optimization, but the optimization does not identify the groups. The researcher must approach the dataset with a hypothesis about which features should be grouped. This is not ideal in all cases. Elastic net and OWL, on the other hand, identify correlated sets of features as part of the optimization. While similar in this sense, OWL differs from elastic net in that it more explicitly groups variables, by assigning voxels deems to belong to the same group exactly the same weight (Figueiredo & Nowak, 2016). This means that OWL defines groups in transparently "human readable" way, while elastic net allows correlated features to be assigned different weights, which obscures the group structure. In addition to producing more interpretable models, OWL has been shown to outperform elastic net in terms of how well it selects and groups features (Bondell & Reich, 2008; Zhong & Kwok, 2012)[1]. In these ways, OWL is the state of the art among regularized regression techniques, and it will provide the base for the analyses in this dissertation.

In the previous chapter, regularized regression was introduced in the context of classification. This was done to keep things simple and as closely comparable to the most prevalent analysis techniques in neuroscience, which are applied to identify sets of voxels that maximally discriminate between experimental condi-

[1]Bondell and Reich (2008) and Zhong and Kwok (2012) performed their experiments using OSCAR, which is a special case of OWL.

tions. However, not all hypotheses are best assessed by discrete, categorical designs. Most theories of concept representation, for example, hypothesize that the brain encodes information with greater fidelity than the level of category. This rich semantic content may be encoded via distributed representations (Chapter 2), or some other form of continuous feature space (e.g., Fernandino et al., 2016; Huth, de Heer, Griffiths, Theunissen, and Gallant, 2016; Huth, Nishimoto, Vu, and Gallant, 2012; Just, Cherkassky, Aryal, and Mitchell, 2010; Mitchell et al., 2008), but in either case performing classification (and, indeed, designing experiments with classification in mind by requiring participants to retrieve and work with conceptual information that are believed a priori to form clusters in "concept space") may lead to a somewhat distorted image of how and where concepts are represented in the brain. This is, in part, because a classifier does not need to model the structure of items within categories—it only needs to discriminate between them. By modeling more complex similarity structure among stimuli, the model will need to identify the components of the neural signal that carries this more nuanced information, which may cast distributed representations in stark relief relative to even whole brain multivariate classification.

Until very recently, the analysis that is suggested by this discussion had never been done. That is, regularized regression had only been applied to classify neural data, and never in the service of a representational similarity analysis. Representational similarity analysis (RSA) refers to a general methodology that involves comparing the similarity structure among patterns of neural activity (e.g., the structure expressed by the covariance of their activity over time or over stimuli) to a reference similarity structure (e.g., a similar covariance structure derived from another brain region or the same region in another (human or non-human) subject, or a structure based on an independent non-neural source like human similarity judgments or a computational model). Typically, these comparisons are direct, without

any sort of modeling to relate the two similarity structures. This involves nothing more than computing, for example, the correlation between the two structures.

While elegantly simple and flexible, the standard approach to RSA has several notable limitations. Consider how the representational similarity structure expressed by the brain is assessed: first, a set of voxels are selected, often either by hypothesis or as part of a searchlight analysis (3.5). Then, in a second independent step, the covariances among those voxels are computed directly. This means that the standard representational similarity analysis cannot be used to discover patterns; it involves analyzing the structure expressed by a predetermined selection of features.

Furthermore, standard RSA involves a strong assumption about the relation between neural activity and the represented structure: that the variance of every voxel in the predetermined selection contributes to the encoded similarity structure to the same extent. This is because the standard RSA procedure does not involve fitting any sort of model when relating the neural activity to the target structure. This puts any uninformative voxels which may exist in the set on equal footing with the most informative, adding noise to the estimate of the structure expressed over the set. The consequence is that selections that contain many uninformative voxels along with some voxels that actually contribute to the structure will be incorrectly overlooked. On the other hand, in regions where the ratio of signal to noise is favorable enough to express similarity structure that matches the target structure of interest, one does not know which voxels in the region actually express the structure. The structure is attributed to the whole region.

While not in the purview of standard RSA, there is great interest in discovering the network structure of the brain. DTI and related probabilistic tractography methods aim to map the major white matter tracts that connect various parts of the brain and allow information to be propagated and integrated among and be-

tween regions. In addition, functional connectivity analysis has begun to play an important role in identifying interactivity among brain regions in the context of specific tasks, clinical sub-populations, etc. Functional connectivity analysis (FCA) involves measuring the covariance among regions of the brain, and determining which of those regions are statistically related. However, functional connectivity analysis does not address what content is encoded in that covariance. In this sense, RSA and functional connectivity analysis are like two research disciplines that possess means to address the deep problems of the other, but have not found a common language to communicate effectively and so have historically talked past each other.

We recently proposed a new analysis technique that unites RSA and FCA into a single method, and which, like the unity of two formerly independent research disciplines, is ultimately more than the sum of its parts. We call this technique Network Representational Similarity Analysis (network RSA), and it addresses all three limitations of standard RSA by fitting a model of the voxel covariance in an fMRI dataset to a target similarity structure via sparse, multi-task regression (Oswal et al., 2016). Just as with the regularized regression approaches to multivoxel pattern analysis (MVPA) discussed in Chapter 3, this model of the data can generate predictions; unlike MVPA, the predictions are not attempts to categorize trials, but to properly situate the trial within a similarity space. The structure of the space expresses some psychologically relevant relationship among the trials, such as semantic or perceptual similarity.

Using network RSA, we will consider a simple cognitive task in a new light. When viewing an image and making a semantic judgment pertaining to the content of the images engages at least two processes: object perception and semantic evaluation. Visual object perception, a hierarchical process beginning with low level visual feature detection and ascending into more and more abstract abstrac-

tions of the input (Humphreys & Forde, 2001, 3; Logothetis & Sheinberg, 1996; Marr, 1982; Tanaka, Saito, Fukada, & Moriya, 1991, July), begins in the occipital lobe and extends into temporal and parietal areas. Perception relies heavily on knowledge, and the line between higher order vision and semantic processing continues to blur. That being said, while there is an agreed upon neural locus for several aspects of visual processing, the same cannot be said for semantic processing. Semantic processing of visual stimuli has been variously ascribed to the otherwise visual regions in posterior ventral temporal cortex (A. Martin & Chao, 2001), the inferior parietal lobe (Binder & Desai, 2011), and the anterior temporal lobe (Patterson et al., 2007). Surveys of the literature on semantic processing implicate the majority of cortex to some extent (Binder et al., 2009). In short, picking a semantic ROI a priori would be arbitrary, especially because most models of semantic processing predict that relevant semantic information may be represented in multiple regions and integrated in some way. This is inconsistent with a searchlight approach, which would not permit integration beyond the 6–8mm diameter of a standard searchlight.

4.1 Intuitions

Before continuing, it will be useful to establish some intuitions about how network RSA works and to make explicit the assumed relationship between psychological similarity structure (e.g., the perceptual or semantic similarity among a selection of images) and functional activity as measured by fMRI. For a full discussion of this method and its assumptions from a more technical perspective, see Oswal et al. (2016).

Previous applications of RSA in the neuroimaging literature are attempts at information mapping, in the sense of Kriegeskorte et al. (2006). The objective has

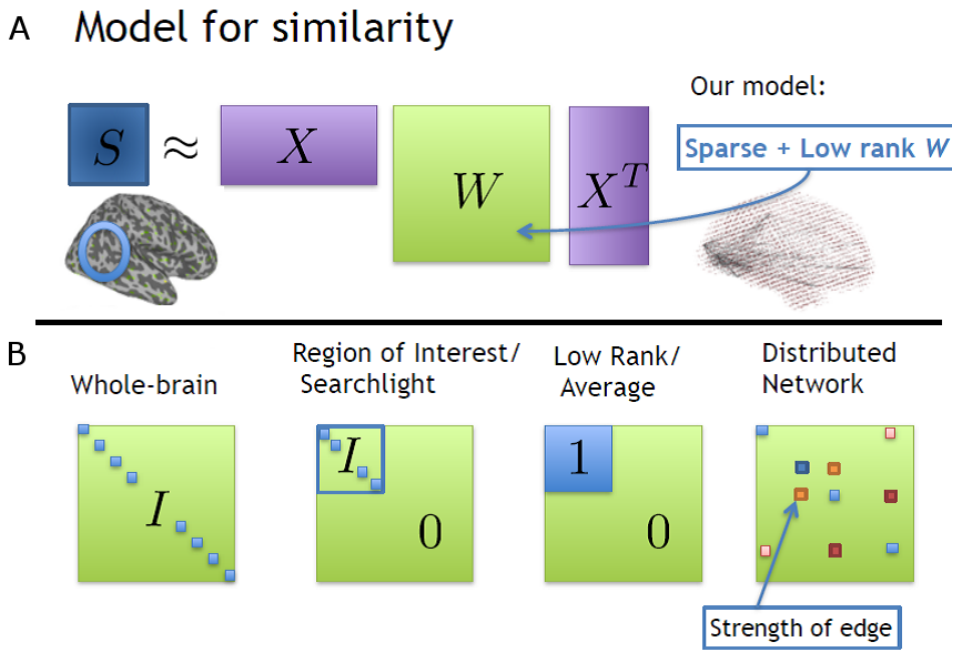

Figure 4.1: Overview of W matrix. See text for details.

been to identify brain regions in which voxel covariance directly expresses a target similarity structure, without modeling the variance or covariance in any way. As discussed in the introduction, this poses important challenges and limitations for a program of research interested in discovering the neural networks that encode structured knowledge about the world.

With network RSA, the objective is to model the covariance among voxels of X to fit the data to a target similarity structure. Symbolically, the assumed relationship between a target similarity structure, S, and neural activity, X, can be expressed as:

$$S \approx XWX^{T} \tag{4.1}$$

See figure 4.1a for a visual depiction of this equation. Here, the superscript T indicates that the associated matrix is transposed, S is a symmetric matrix expressing a measure of similarity between each pair of items, X is a matrix representation of

the fMRI data where rows correspond to items and columns correspond to voxels, and W is a matrix of weights that expresses the importance of each voxel (along the diagonal) and pair of voxels (off the diagonal) to expressing the structure in S. To be more explicit: the first value in the first row of W is a weight on voxel 1 which expresses how the activation of that voxel should be scaled before being linearly combined with the (weighted) activity of other voxels. This is the weight that would be applied to voxel 1 if all other voxels were to be held constant at zero. The remaining values in the first can be understood as interaction terms in a regression model. That is, they moderate the weight on voxel 1 by expressing how much this weight should be increased (or decreased, in the case of negative weights) for a unit change on each other voxel. Put another way, they express the influence of voxels 2 to n on voxel 1. This is consistent with conceiving of W as a representing graph structure, from which a network of voxels and their interconnections can be drawn.

Because it is so central to our method and the interpretation of our results, let us build intuitions about W by considering the four illustrative example graph structures depicted in figure 4.1b, which correspond to contrived configurations of weights in W. Let W have a row and column for every voxel in cortex, and let us consider each W structure in turn from left to right.

In the first example, W is fixed so that there are ones along the diagonal and zeros elsewhere (i.e., the identity matrix). In this case XWX^T would simplify to XX^T, which, if the columns of X are mean centered, yields the full covariance matrix of X. Conceptually, this W structure expresses that every voxel and pairwise combination of voxels in the brain contributes equally to expressing the target similarity structure. The second example W structure omits some of the voxels by setting rows and columns that correspond to the unwanted voxels (that is, columns in X) entirely to zero. Columns of X that correspond to all-zero rows (and columns) in

W are set to zero by XW. In this way, one can conceive of standard ROI and Searchlight based analyses in terms of W matrices with a subset of diagonal values set to 1 and all other values set to zero. Given such a W, XWX^T yields the covariance matrix among only the voxels with non-zero rows in W.

Next, we consider a less intuitive yet conceptually important structure, where off-diagonal values of W are non-zero. A block of ones in the W matrix, such as shown in the third example structure, expresses that all of the covariance, and the variance of individual voxels, are equally important. In fact, this particular structure of W expresses that the set of voxels that belong to a block *are effectively the same*. They ought to be treated as a single voxel. To cultivate this intuition, imagine that the set of voxels in the block are nodes in a network. W tells us the weights to assign the edges that connect these nodes, and in this case it is a block of ones: every node is connected to every other node with weight of 1. If any single node is stimulated, all nodes will adopt that same state. If multiple nodes were activated, the activation would sum. The network is most strongly active when all nodes are stimulated at the same time in the same way, less so when only a few units are active or when some units activate positively and others negatively. This is conceptually akin to aggregating voxel activity within ROIs.

The three example W structures we just considered are deliberately artificial, with fixed values in particular patterns. In reality, W is fit to the data and can take on whatever values are necessary, while respecting the constraints of symmetry (the influence of voxel i on voxel j must be the same as the influence of voxel j on voxel i) and of positive semi-definiteness (the Eigen decomposition of W must not contain negative eigenvalues). The fourth example depicts this, with an emphasis on the fact that W is sparse—most values in W are set to zero. The sparsity constraint will be explained in the next section, in which the optimization problem itself is described.

In summary, to conduct an network RSA is to solve for W in $S \approx XWX^T$, and the appropriate interpretation of W is that it contains the weighted edges of an undirected graph which represents the network structure among voxels.

4.2 Optimization

I will now consider in more detail how W is obtained for a given S and X. The fitted W should be sparse and low rank, which means that many voxels will receive zero row/column weights and be ignored (sparse) and that if one were to perform a principle components analysis on W, many of the eigenvalues would be zero (low rank). The full optimization problem can be written as:

$$\underset{W \in S_+}{\arg\min} \, ||S - XWX^T||_F^2 + \lambda_1 ||W||_1 + \lambda_2 ||W||_* \tag{4.2}$$

This essentially conveys three pieces of information. We want to find a W that will jointly minimize:

1. *Data fitting term*: The Frobenious norm of the difference between the target similarity matrix S and the predicted similarity matrix, given by XWX^T.

2. *Sparsity*: The ℓ_1 norm of W, which is the sum values in W. Technically, the ℓ_1 norm is the sum of the absolute values, but W cannot have negative values to begin with. The ℓ_1 norm encourages W to be sparse.

3. *Low rank structure*: The nuclear norm of W, which is the sum of the singular values of W. The nuclear norm encourages W to be low rank.

Note that λ_1 and λ_2 are simply non-negative scalar values that correspond to free parameters that allow the importance of the ℓ_1 and nuclear norms to be scaled up or down to fit the data.

Because the nuclear norm is computationally expensive, in practice the optimization routine is simplified by applying the "square root trick" (citation). Just as 4 can be expressed as 2×2, a matrix can often be expressed in terms of its square root. If $\sqrt{S} = Y$, then $S = YY^T$, and if $\sqrt{W} = B$ then $\sqrt{W} = BB^T$, and the original equation $S \approx XWX^T$ can be rewritten as $YY^T = XB(XB)^T$. Once the redundancy in the re-written form is eliminated, we are left with simply:

$$Y \approx XB \tag{4.3}$$

Again, W can be recovered by BB^T. The square root of a symmetric matrix is a basis set which can be linearly combined with itself to reproduce S. We are defining Y as:

$$T = U\Sigma \tag{4.4}$$

Here, $U\Sigma$ comes from the singular decomposition of S, $S = U\Sigma V$. U is a matrix where columns are eigenvectors and Σ is a matrix with the square roots of the eigenvalues on the diagonal. Because S is symmetric and positive semi-definite, these values are readily obtained and will ensure that $S = YY^T$.

Recall that the objective of the nuclear norm is to penalize based on the absolute sum of the eigenvalues, and that this penalty can be minimized by forcing eigenvalues to be zero. This is how we would enforce the constraint that W should be low rank if we were to work with $S \approx XWX^T$ directly. In the form $Y \approx XB$, we have the opportunity to directly set values in Σ to be zero, which will cause Y to have all-zero columns and thereby assert that the structure being modeled is low rank. Simply put, if we wanted to assert that the structure being modeled is rank 3, then we would set all but the first three eigenvalues to zero, and then only retain the first 3 columns of the resulting Y (columns 4 to n would be all zeros and contain

no information). Let us refer to this "truncated" Y as Y^r.

Now, rather than the optimization written above, we are going to pose $Y^r \approx XB^r$ as a multi-task learning problem.[2] Multi-task learning involves fitting several related models at once, and allowing the models to jointly constrain and guide one another as they converge towards a solution. Conceptually, a separate regression model will be fit to each column of Y^r. This will result in r weight vectors, where r is the number of columns in Y^r, amounting to r weights per voxel. What makes it a multi-task regression and not merely r independent regression models is that all models will be solved as part of the same optimization problem, which can now be stated as:

$$\arg\min_{B} \|Y^r - XB^r\|_F^2 + \lambda \sum_{i=1}^{p} w_i \|b_i\|_2 \tag{4.5}$$

Where Y^r is the truncated square root of the target similarity matrix S, X is a matrix representation of the fMRI data where columns correspond to voxels, B^r is a matrix of weights where rows correspond to voxels and there are r columns to match the r columns of Y, p is the number of voxels, and λ is a free parameter that can be adjusted to modulate the importance of the regularization term. The regularization term is a modification of the group LASSO penalty. For more information on this regularization penalty, please see Oswal et al. (2016) and Figueiredo and Nowak (2016). At a high level, this optimization equation conveys two pieces of information. We want to find a B^r that will jointly minimize:

1. The Frobenious norm of the difference between the truncated square root of the similarity structure S, represented by Y^r, and the predicted values for each component in Y^r, given by XB^r.

[2]The r superscript is intended to make explicit that the Y and B matrices have r columns, where r is defined by the rank truncation step explained in the previous paragraph.

2. The modified group LASSO regularization penalty, which forces XB^r to be "row sparse". That is, it forces many rows of B^r to be set to all zeros, but also imposes the constraint that if one element of a row is zero, all must be zero. To assign a zero to one element in a row requires setting the whole row to zero.

In fact, there is one other important aspect of this equation, and it has to do with $\sum_{i=1}^{p} w_i \|b_i\|_2$. This is the modification to group LASSO that makes it a variant of ordered weighted LASSO (OWL, Bondell and Reich, 2008; Figueiredo and Nowak, 2014; Figueiredo and Nowak, 2016). OWL is intended to address one of the key problems associated with LASSO, which is that if a set of voxels are all informative and are highly correlated with one another, LASSO will tend to select only one voxel from the set and set the rest to zero. This is problematic when a key objective is to test hypotheses about where in the brain information is being encoded and what that neural code is like. LASSO will tend to exaggerate sparsity in ways that can be misleading about the true nature of neural representations. OWL attempts to select whole sets of correlated voxels and provide a more complete picture of the neural representation.

In summary, the objective is to obtain a matrix of weights that maps between neural activity as measured by fMRI and a target similarity structure, which is expressed by Equation 1. The optimization that most obviously follows from this objective, expressed by Equation 2, is computationally intensive and motivated refactoring the problem as Equation 3, which will be optimized with all the same constraints by Equation 4. Once obtained, we can compute our estimate of W as $B^r(B^r)^T$.

Chapter 5

fMRI Experiments

How does the brain support semantic knowledge? The hub-and-spoke model predicts that concepts are supported by a distributed network of brain regions (the spokes) which each represent modality specific structure, and single hub region that all the spokes interact with. In this way, a single, domain general semantic system is enabled by pan-modal integration of our internal, external, and linguistic experiences. This all important semantic hub role is believed to be fulfilled by the bilateral anterior temporal lobes (ATL). However, how the ATL functionally performs this pan-modal integration is controversial: the hub-and-spoke model predicts that the semantic hub should be functionally involved with semantic processing of all kinds, but—as outlined in Chapter 1—there are three hypotheses about how the ATL contributes to semantic representation. Each hypothesis is associated with different representational structure being expressed among patterns of activity, both locally/regionally in the ATL and over cortex more generally. Briefly, those hypotheses about how semantic cognition might be supported by the ATL were:

1. *The convergence hypothesis*: The ATL contains high level convergence zones that describe how sensory features are distributed all over posterior, modal-

ity specific cortical regions.

2. *The semantic hub hypothesis*: The ATL encodes pan-modal representations that express all semantic similarity structure across all modalities.

3. *The hub+spoke hypothesis*: The ATL "hub" encodes the interactions among modality specific "spokes", so the hub and spokes jointy express the semantic similarity space.

All three hypotheses predict that representational similarity will be expressed among distributed patterns of neural activity associated with different stimuli, but differ in where that information is expressed. Respecting the same indexing as above, the predictions under each hypothesis are that representational similarity is encoded:

1. Over modality specific regions, each region expressing different dimensions of structure.

2. Over the bilateral anterior temporal lobes, which together express all dimensions of structure, *as well as over modality specific regions*. The hub taken by itself, or the spokes taken together but without the hub, are largely redundant from a brain decoding perspective. Of course, the hub is important from a neural processing perspective.

3. Over the bilateral anterior temporal lobes (hub) and modality specitic areas (spokes), where each spoke expresses different dimensions of structure, and the hub expresses dimensions of structure that capture the interactions among dimensions expressed in the spokes. Any given representation requires all areas, and from a brain decoding perspective a "full" model that includes the hub and all spokes should outperform any partial permutation of areas.

All three hypotheses predict distributed representations involving multiple anatomical regions. By this, I mean that all hypotheses predict that if one were able to simultaneously model information encoded across modality specific areas, and the model appropriately combined that information, that this model would have captured all dimensions that contribute to semantic structure. This ideal model would, in essence, be performing the role of the ATL.

The semantic hub hypothesis (2) differs from the others in that it predicts that the bilateral ATL support a locally consolidated high-dimensional semantic similarity space that has already combined, through relevant experience, the content encoded in the various spokes. This means that semantic representational similarity should be discovered in the ATL, without any additional information about activation states elsewhere in the brain. On the other hand, the hub+spoke hypothesis (3) predicts that the ATL does not encode a the full semantic space, but rather only encodes information about cross-modal interactions. This means that if the scope of analysis is limited to the ATL, it will not be possible to recover the semantic similarity structure, but considering the the ATL *and* modality specific regions together will. From an empirical modeling perspective, this predicts an interaction effect: modeling the hub and the spokes together should produce a more accurate model of the semantic similarity structure encoded in the brain than either the hub alone or the spokes alone. The convergence hypothesis (1) would be consistent with the ATL expressing no semantic similarity structure whatsoever, and so naturally including the ATL in a model's scope should make no difference.

This means that these three hypotheses may be adjudicated by comparing multivariate analyses with either locally restricted or whole-brain scope. Previous work with representational similarity analysis (RSA) has only performed analyses with locally restricted scope, due to limitations with the method. However, by employing network RSA, a technique introduced to the literature by Oswal et al.

(2016) and discussed at length in Chapter 4, we can perform a whole brain RSA and test this contrast. Also, because network RSA involves fitting a model that makes principled estimates of network structure associated with the selected voxels, the tools now exist that permit testing these hypotheses in much greater detail.

No matter which representational structure, the ATL's hub status means that it receives input from multiple modalities. Therefore it follows that the ATL should be functionally relevant regardless of the stimulus modality. That is, semantic processing of stimuli presented in audio or visual form should both involve the ATL. There is direct evidence, acquired via electrocorticography (ECoG) in patients preparing for neurosurgery, that neural sites within the ATL have tuning functions that respond to multiple modalities (Abel et al., 2015; Shimotake et al., 2014). However, to the extent that semantic representational similarity among items presented, say, visually can be detected in the ATL using local or whole brain RSA in an fMRI dataset, the hub-and-spoke model strongly predicts that a similar result would be obtained if the stimuli were presented in another modality.

In this first set of experiments these three representational hypotheses will be tested in a distortion corrected fMRI dataset, collected from participants who performed semantic judgments about a set of concepts, referenced by either a visual or audio stimulus. In combination with network RSA, this multimodal dataset collected with a protocol designed to correct MR imaging artifacts in the ATL and orbitofrontal regions has all the critical pieces in place to test, for the first time, whether semantic cognition is supported by distributed representations that span multiple brain regions.

5.1 Materials and methods

Participants

This dataset is comprised of 23 right handed participants (5 male). All participants were native English speakers, had normal or corrected-to-normal vision and none of them reported having any neurological disorder or dyslexia. The experiment was conducted at the Neuroscience and Aphasia Research Unit (NARU) at the University of Manchester, UK. Participants were paid 10 pounds for their participation. The experiment was approved by the Research Ethics Committee at University of Manchester.

Stimuli

In this experiment, 37 concepts were represented by both a black and white line drawing selected from the Snodgrass and Vanderwart (1980) set and a characteristic sound. Sounds were purchased purchased from Soundrangers[1] and edited using the software Audacity[2] to have equal duration (2s) with the same level of maximum amplitude. The set of line drawings was constrained two factors: 1) the need to chose items that could be matched with a characteristic sound available in this database, and 2) the need to select a set of line drawings where the low level visual similarity structure among items was uncorrelated with their estimated semantic similarity structure (see Models subsection, below).

Models

Representational similarity analysis (RSA) compares a target similarity structure selected by hypothesis to the similarity structure among patterns of functional ac-

[1]https://www.soundrangers.com/
[2]http://audacity.sourceforge.net/

tivity. In order to conduct RSA for the visual and semantic similarity structure among the 37 stimuli, we generated two matrices through independent means.

Visual similarity

The purpose of the RSA of visual structure is to serve as a control for low level similarity structure, and as a sanity check that our analyses localize this low level similarity structure to early visual cortex. To that end, we simply computed the minimum point distance to translate each black and white line drawing into every other black and white line drawing (the Chamfer method). The distance between each pair of items was arrived at through an iterative procedure where one of the images was scaled and translated relative to the other. The distance between the two images was taken to be the smallest value to emerge from these scaling and translation steps.

The resulting distance matrix was then converted to a similarity matrix via $S = e^{-cD}$, where c is a constant which scales how quickly similarity degrades towards zero with distance. The particular c we chose, 0.05, was selected because it produces a matrix S in which the rank similarity between each item perfectly corresponds to the rank dissimilarity expressed by D, and for which its eigenvalues were distributed similarly to the eigenvalues of D (indicating similar structural complexity).

Semantic similarity

Finally, for semantic dissimilarity, stimuli were scored using the binary features lists obtained in a recent feature norming study (Dilkina & Lambon Ralph, 2012). These are revised feature lists from Cree and McRae (2003), who originally asked undergraduates to list ten most salient features associated with each concept. Using these feature lists as a guide, Dilkina and Lambon Ralph (2012) effectively

cross-referenced the set of features produced by participants over all items with each item: rather than have participants list features themselves, they were provided with a word and a list of features, and their task was to indicate which of the features pertain to the concept referenced by the word. Relative to the feature listing task of Cree and McRae (2003), this feature *verification* task discovered richer structure, with more features being associated with each item. This is because there are many features of objects that do not spring to mind: you know that your desk chair is assembled with screws, but this is probably not as salient to you as whether it is comfortable (or not), whether it reclines, whether it has lumbar support, and so on, and being asked to list the features of your chair would reflect the availability of those facts. However, how your chair was assembled makes it similar to other things, in a way that may come out after explicitly cross-referencing features with concepts. The revised feature lists correlate well with other methods of getting semantic features, such as studying participants' drawings of different concepts (Rogers et al., 2004). The semantic similarity matrix is derived by computing the pairwise cosine similarity between these feature vectors.

Comparing structures

The visual and semantic structures are uncorrelated ($r = 0.06$, *n.s.*). Hierarchical cluster plots of each structure are displayed in Figure 5.1. The figures also contain examples of stimuli that are clustered close together in each structure, to give a sense of their visual similarity.

Experimental Procedure

The experiment consisted of four blocks, each of which lasted 8 minutes. In each block, all stimuli (both visual and auditory) were presented once. To avoid priming effects, trials were pseudorandomized with the constraint that visual and auditory

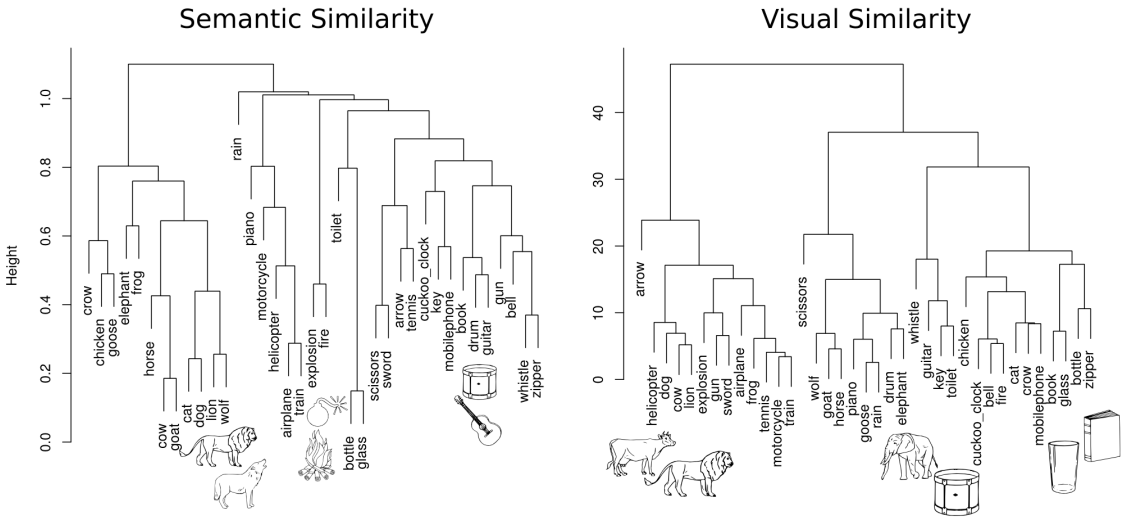

Figure 5.1: Hierarchical cluster analysis of the semantic (left) and visual (right) similarity structure among the 37 stimuli in the experiment.

versions of the same stimuli were separated by at least 10 trials. Pseudorandomization was achieved using Mix (van Casteren & Davis, 2006). Each trial lasted 4000ms. Visual trials started with a red fixation cross for 500ms, followed by the visual stimuli for 2000ms and a blank screen for 1500ms. Auditory trials started with a blue prompt "sound", followed by natural object sounds for 2000ms and a blank screen for 1500ms. Twenty-five null events (blank screen) were randomly inserted into each block to make ITIs variable to aid in deconvolution.

Participants were instructed to mentally perform a size judgment task: whether each item can be physically fit into a "wheelie bin" (outdoor trash can) in the real world (c.f. Horner & Henson, 2012). From time to time there would be a question mark at the end of a trial and participants needed to press buttons to indicate their size judgment regarding the previous item.

Image acquisition protocol

Functional MRI scanning was performed using a Philips 3T MR system with a head coil at the NIHR/Wellcome Trust Central Manchester Clinical Research Facility. Dual Echo sequence (Ajay D. Halai, Parkes, & Welbourne, 2015) was acquired using a $\mathtt{TR} = 2.8\mathtt{s}$, $\mathtt{TEshort} = 12\mathtt{ms}$, $\mathtt{TElong} = 25\mathtt{ms}$ and a flip angle of 85 degrees. Reconstructed images contained 31 slices covering the whole brain, with slice thickness $4\mathtt{mm}$, interslice distance $0\mathtt{mm}$, field-of-view $240\mathtt{mm}$ and in-plane resolution 80×80 voxels ($3 \times 3\mathtt{mm}$). Slices were tilted so that the front was up 30 degrees and the posterior end was down 30 degrees. A field map was acquired for each participant and used to perform distortion correction. Functional scans were preceded by a high-resolution structural T1-weighted MRI scan, acquired using a 3D MPRAGE sequence, field-of-view $240 \times 191\mathtt{mm}$, voxel size $1 \times 1 \times 1\mathtt{mm}$, matrix dimensions $256 \times 256 \times 256$, $\mathtt{TR} = 8.4\mathtt{ms}$, $\mathtt{TE} = 3.9\mathtt{ms}$.

Anatomical segmentation and surface construction

Cortical reconstruction and volumetric segmentation was performed with the Freesurfer image analysis suite, which is documented and freely available for download online (`http://surfer.nmr.mgh.harvard.edu/`). The technical details of these procedures are described in prior publications (Dale, Fischl, & Sereno, 1999; Fischl & Dale, 2000; Fischl, Liu, & Dale, 2001; Fischl et al., 2002; Fischl, Salat, et al., 2004; Fischl, Sereno, & Dale, 1999; Fischl, Sereno, Tootell, & Dale, 1999; Han et al., 2006; Jovicich et al., 2006; Reuter, Rosas, & Fischl, 2010; Reuter, Schmansky, Rosas, & Fischl, 2012; Segonne et al., 2004). Briefly, this processing includes removal of non-brain tissue using a hybrid watershed/surface deformation procedure (Segonne et al., 2004), automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures (Fischl et al., 2002; Fischl, Salat, et al., 2004) intensity normalization (Sled, Zijdenbos, & Evans, 1998), tessellation

of the gray matter white matter boundary, automated topology correction (Fischl et al., 2001; Segonne, Pacheco, & Fischl, 2007), and surface deformation following intensity gradients to optimally place the gray/white and gray/cerebrospinal fluid borders at the location where the greatest shift in intensity defines the transition to the other tissue class (Dale et al., 1999; Fischl & Dale, 2000). Once the cortical models are complete, a number of deformable procedures can be performed for further data processing and analysis including surface inflation (Fischl, Sereno, & Dale, 1999), parcellation of the cerebral cortex into units with respect to gyral and sulcal structure (Desikan et al., 2006; Fischl, van der Kouwe, et al., 2004). The maps are created using spatial intensity gradients across tissue classes and are therefore not simply reliant on absolute signal intensity. The maps produced are not restricted to the voxel resolution of the original data and thus are capable of detecting submillimeter differences between groups.

Cortical reconstruction and volumetric segmentation was conducted based on each subject's T1 anatomical scans and the methods cited above above. Cortical masks for each subject are defined as the space between the pial surface and white matter. Temporal lobe regions of interest were defined in reference to the Destrieux atlas (Destrieux, Fischl, Dale, & Halgren, 2010). Segmentations were visually inspected, and no corrections were deemed necessary.

Preprocessing

Functional images were corrected for slice timing, realigned and unwarped to remove any movement artifacts using tools and pipelines available in SPM 8 (Wellcome Department of Cognitive Neurology, London, UK). Images were coregistered to the T1-images using a mutual information coregistration procedure (Pluim, Maintz, & Viergever, 2003). The structural MRI was normalized to the `TT_N27` Talairach template using tools and pipelines available in AFNI (R. W. Cox, 1996).

Functional data remained in native space for all multivariate analyses, and solutions were warped into Talairach space using these transformations prior to group level statistics.

A general linear model (GLM) was used to model individual items' hemodynamic response using SPM 8. Each item was presented four times in each modality. The beta maps for these four repetitions were then averaged, resulting in one pattern per item in each modality.[3] These beta maps were then filtered to exclude voxels not belonging to the cortex. This was determined according to segmentations produced with Freesurfer. The trials associated with the visual and audio modalities were then separated. For each modality, voxels with estimated responses more than 5 standard deviations from the mean response across voxels were dropped.

The resulting censored whole brain beta maps for individual items in individual participant's native space were used in further analysis multivariate analyses.

5.2 Analysis methods

Univariate analysis

A univariate analysis was carried out by Yuanyuan Chen, PhD at the University of Manchester, and communicated to me through unpublished materials. These results will be reproduced in the following section to provide context relevant context. In these analyses, rather than considering individual stimuli, we consider 4 different conditions that a participant will experience over the course of the experiment: 1) audio trials, 2) visual trials, 3) size judgment "catch" questions, and 4) rest—points in time where nothing more than a fixation point is displayed and the participant is between trials. A general linear model is fit as describe above,

[3] Alternatively, the four repetitions of each item could have been modeled together so that that GLM would have simply produced one beta map per item per modality.

except that each subject's data are smoothed with an 8mm FHWM gaussian blur prior fitting the GLM, and rather than modeling each stimulus these 4 conditions are modeled. Four contrasts will be reported: 1–3) each trial type $>$ rest, and 4) visual vs. audio

Searchlight RSA

RSA analysis has been applied in previous papers (Connolly et al., 2012; Kriegeskorte, 2009; Kriegeskorte et al., 2008) and its pipeline has been described by Kriegeskorte (2008). Briefly, for a selection of voxels, the 3 dimensional pattern of activity for each item are formed into an item by voxel matrix. The correlation among a rows of this matrix will express the representational similarity structure expressed over that selection of voxels. The lower triangle of the similarity matrix for the region is then correlated with the lower triangle of target similarity matrix. The resulting correlation coefficient is a measure of how similar the two structures are.

Whole brain searchlight (Kriegeskorte et al., 2006) simply involves repeating this process at every voxel in the dataset. That is, for each voxel, select all voxels within a particular radius, convert the patterns of activity over items to a matrix as just described, compute the correlation among rows, and correlate with the target matrix. The resulting coefficient is then inserted at the center voxel of the searchlight, and move on to the next voxel. This can be applied to produce a map of correlation coefficients for each subjects, with can then be passed into a group univariate analysis to test which points in the brain reliably express the target similarity structure across subjects.

We performed this analysis using a spherical searchlight with 8mm radius, which is comparable to previous studies (Giordano, McAdams, Zatorre, Kriegeskorte, & Belin, 2013; Peelen & Caramazza, 2012). The Spearman correlation was used to assess the similarity between the target and searchlight similarity matrices. The

resulting correlation maps for each subject were interpolated to align with the `TT_N27` template brain, and smoothed with an 8mm FWHM gaussian kernel prior to group level analysis via a univariate t-test.

Network RSA

Derive target embeddings from model similarity matrices

Network RSA solves for a matrix of weights that will identify and linearly combine a set of voxels which jointly express a target similarity structure, here either visual or semantic similarity. However, as discussed in Chapter 4, the optimization does not model the full rank similarity matrix directly. Instead, the full model similarity matrix, S, is decomposed into its eigenvectors and values, and the r largest components are taken as a *low rank embedding* of S, which I will call Y to emphasize that this is what is being modeled and predicted by network RSA. I refer to Y as an embedding to express that the similarity structure has been "embedded" in a low dimensional metric space. A 2 dimensional embedding could rightly be plotted on a coordinate plane, and the distances among the points could be intuitively interpreted: points that are closer together are more similar.

The number of dimensions is chosen by setting a threshold for $||S - YY^T||_F / ||S||_F$ needs to be. With more dimensions, this error term will decrease, but not all variance in the similarity matrix is necessarily meaningful. When decomposing the visual model matrix, this threshold was set to 0.2, which results in a 3 dimensional embedding of visual structure. When decomposing the semantic model matrix, this threshold was set to 0.3, which results in a 8 dimensional embedding of semantic structure.

Voxel standardization

When model fitting, it can often help to standardize the variables in the data being modeled. In the case of regularized regression, this carries additional importance: voxels with more variance can be assigned a smaller weight. And since the regularization penalty will grow as the magnitude of the weights grows, this exerts influence on which voxels are selected. However, if the goal is to identify this most informative voxels rather than the "loudest" voxels, this is undesirable. Thus, every voxel is standardized by dividing by its standard deviation.

Cross validation

Model performance is assessed through cross validation. The 37 items are split into 9 groups (8 with 4 items, 1 with 5 items). When training the model, one of these groups is excluded from the training set. Once the model is fit to the truncated training set, it can be used to make predictions about the held out items. Error is assessed as $||Y - \hat{Y}||_F / ||Y||_F$, where Y is the r dimensional target embedding, with a coordinate for each item and $\hat{Y}$ is the models prediction of those values. This process is then repeated so that each of the 9 groups is held out one time. These 9 errors are then averaged to obtain the expected error for that model.

Likewise, in the analyses that follow, model solutions (i.e., the values attributed to each voxel) are also aggregated over cross validations: node strength estimates are averaged, and the number of times the voxel was assigned a weight over cross validations is taken as a metric of stability.

Parameter tuning

The network RSA procedure involves fitting two free parameters. Before cross validating to obtain a final error and a final solution map, the appropriate values for these must be determined. This parameter search is nested within each cross

validation noted above. This is necessary: estimating parameters for the whole dataset at once would provide some information about the portion of the data that will be held out during cross validation. By performing a separate parameter search within each cross validation loop, one can be sure that the model is not being "leaked" information about the test set.

The parameter space is searched via a manual grid search: At each point in the grid, cross validation is performed over the remaining item 8 groups and then averaged to obtain the expected value for that particular combination of parameters. Once the models at all points in the grid have completed, the point with the lowest cross validated error determines the parameters used to fit the models that will actually be inspected and analyzed further.

Group level significance testing

To determine whether a given voxel is statistically reliable at the group level, we have devised a permutation-based nonparametric test that obtains p-values with respect to a binomial distribution. A null distribution for the test statistic at each voxel is estimated by refitting a particular model many times, each time with the training examples permuted into a different, random order. Each time, a sparse set of voxels will be identified. At a given voxel, the distribution that emerges from these permutations will be left skewed with a large number of zeros—very non-normal, and also not well fit by a gamma function. However, one thing that can be safely said about the distribution is that the probability of the next random observation from this distribution being in the top half of the distribution is 0.5. The test statistic at a given voxel can be ranked against its empirical null distribution as determined by the permutation procedure, and if the value is ranked higher than half of the values in this distribution, then we can store a 1 at that voxel for that subject. Again, the probability of obtaining a 1 under the null hypothesis is 0.5. If

this procedure is applied for each subject, then for a given voxel one can simply count the number of subjects where it was assigned a one and perform a binomial test. The resulting probability can be used to determine whether a voxel value is statistically non-random at the group level.

In practice, because network RSA is applied to each subject's data before interpolating to a common coordinate space, the full procedure is as follows: 1) compute the test statistic for each voxel, 2) refit the model for 100 random permutations of the training set to form a null distribution for the test statistic at each voxel in the subject's native space, 3) interpolate all maps into the common space, to obtain a common space test statistic and associated common space null distribution, 4) rank the common space test statistic against the common space distribution, 5) if in the top half of the distribution, replace the test statistic with a 1, otherwise replace it with a 0, 6) repeat for all subjects, and 7) perform a binomial test at each voxel over subjects.

Each of these permuted models also yields a prediction error on an (also random) test set. The distribution of errors over permutations is normally distributed. Because the chance error rate for a model cannot be determined a priori, the mean and standard deviation of the error distribution can be used to standardize the error obtained by the "true" model. The standardized errors over subjects can be tested against a null hypothesis of zero with a simple t-test to determine whether performance is reliably better than chance in the sample.

High throughput computing with HTCondor

The network RSA procedure is computationally intensive. Parameter tuning, permutation ranking, and cross validation for each subject and modality and target structure involves fitting hundreds of thousands of models, and each model can take 20 or 30 minutes. The amount of computation time required to conduct this

analysis can be tallied in years. Obviously, this is grossly impractical to do on a workstation, or even a high performance super computer. Fortunately, because each model can be run without input from any other model, they can all be run as separate processes. Problems of this kind are very well suited to high throughout computing.

I, in collaboration with the Center for High Throughput Computing (CHTC) at the University of Wisconsin-Madison, have developed workflows for interfacing with distributed computing resources on campus and throughout the United States via the Open Science Grid (Pordes et al., 2007; Sfiligoi et al., 2009) using HTCondor (Thain, Tannenbaum, & Livny, 2005). The code necessary to reproduce these analyses are hosted at `https://github.com/crcox/WholeBrain_RSA` and `https://github.com/crcox/condortools`.

5.3 Results

Behavioral results

Participants performed the size judgment task in catch trials quite accurately, with mean accuracy of 91% (SD=13%) and mean response time of 1167 ms (SD=281 ms). This indicates that participants were paying attention and accessing the requisite semantic knowledge in order to perform accurately.

Univariate analysis

Before considering the searchlight RSA and network RSA solutions, let us consider the univariate solution (Figure 5.2). When contrasted with each other (bottom row), viewing stimuli in one or the other modality activated the associated primary motor cortex and little else. However, when considered relative to a passive baseline (rest), the audio trials are associated with increased BOLD in both hub-

and-spoke modelr temporal sulcus and inferior posterior temporal lobe, which is involved with processing visual form. This is consistent with prior work showing that that a stimulus presented in one modality can activate areas of the brain involved in processing other modalities of input, and that the very same stimulus can activate different modality specific cortical regions when the task demands change (see A. Martin, 2007 for review). Perhaps we see increased activity in visual areas on audio trials and not vice versa because the size judgment that participants were directed to make covertly is relies on form knowledge encoded in visual areas, and not on information encoded in the audio modality.

The audio, visual, and rest conditions are all passive in that they do not require any sort of overt response. The catch trials, which occur only periodically and do require an explicit behavior, show a much wider pattern of activation. Notable among the implicated areas are lateral inferior prefrontal cortex, bilaterally, which is thought to be a part of the semantic control system (Hoffman, Binney, & Ralph, 2015; A. Martin, 2007; A. Martin & Chao, 2001; Schnur et al., 2008). The inferior parietal lobe and middle temporal gyrus are also active on catch trials relative to baseline. The IPL is a often active in semantic tasks with an action component, particularly involving manipulation (Boronat et al., 2005). It is also associated with ideomotor apraxia (Buxbaum, 2001), a neurological disorder that affects tool use and other learned gestures. While the task does not explicitly demand knowledge about how to manipulate anything, when thinking about whether an object can fit in a bin one might consider how the object can be manipulated in order to make it fit. The middle temporal gyrus, up to and perhaps including the angular gyrus are also activated relative to rest. The angular gyrus in particular has been flagged as a major cross modal hub which might be important for integrating audio, visual, motion, and action information, making it semantically relevant (Binder et al., 2009; Binder & Desai, 2011).

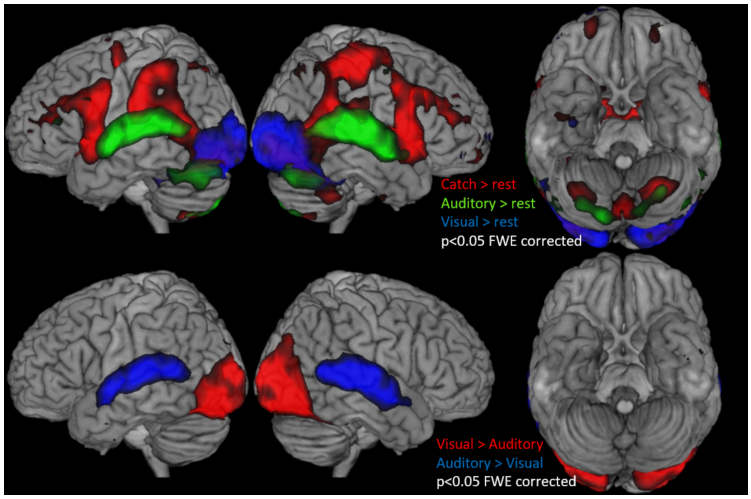

Figure 5.2: Univariate solutions obtained for visual, audio, and catch trials (where catch trials called for explicitly making a relative size judgment, and audio and visual trials were when stimuli were covertly processed). Maps for each trial type were thresholded at FDR corrected $p < 0.05$ and overlaid.

In all, the univariate results here capture the typical pattern of results seen in the literature rather well. The ATL is virtually silent.

Searchlight analysis

The preceding univariate analysis tests whether there are regions of the brain that show mean shifts in activity relative to a baseline, relative to 4 broad experimental conditions. It cannot test whether or how these regions express representational structure, and so does not speak directly to the central questions of this work. For that, we turn now to the multivariate representational similarity analyses, beginning with searchlight RSA.

Searchlight RSA can test whether there are *localized* patterns of activation that express similarity structure of theoretical interest, in *consistent* anatomical locations across participants. In the analyses to follow, trials where participants retrieved concepts based on their characteristic sounds—*audio trials*—or their black and white line drawing depictions—*visual trials*—will often be analyzed separately. This will allow us to ask whether the same cortical regions express semantic struc-

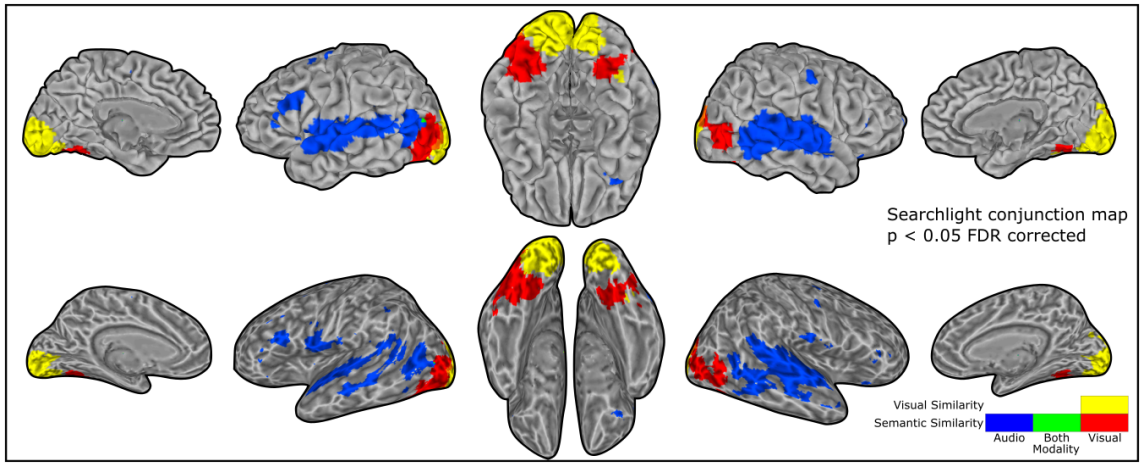

Figure 5.3: Searchlight conjunction $p < 0.05$ FDR corrected. 8mm searchlights, 8mm blur.

ture when concepts are retrieved, regardless of which modality the stimulus prompting retrieval was perceived through.

Conjunction analysis

The results of such a conjuction analysis are presented in Figure 5.3. This aggregates the results of three analyses, overlaid on top of one enough to reveal any overlap. Each analysis is exploring the fMRI dataset with a 8mm searchlight, considering the raw, unsmoothed beta coefficients associated with a GLM which statistically deconvolved the BOLD response using a canonical hemodynamic response function. This results in an information map for each subject and analysis, which has the Spearman correlation between the target and neural similarity structure stored at each voxel. These information maps are then warped to a common space and spatially smoothed with a 8mm FWHM gaussian kernel. These interpolated and smoothed maps were then used as the basis for a group-level univariate t-test. To compose Figure 5.3, each map was thresholded at false discovery rate (FDR) corrected $p < 0.05$ before being overlaid. One analysis compared the *visual* similarity structure to the correlation among patterns of activity on *visual* trials (shown in yellow); the other two compared the *semantic* similarity to the correlation among

patterns of activity on *audio* and *visual* trials (blue and red, respectively).

The first thing to note is that visual and semantic structure corresponds to clearly separate regions: visual structure correlates only with local structure expressed in early visual cortex, while semantic structure on visual trials is expressed in the posterior ventral temporal lobe (pvTL) and lateral occipital cortex (LOC), both bilaterally. This is consistent both with the univariate and a great deal of multivariate semantic studies in the literature.The pvTL and POC encode higher-order visual structure, and seem to play an important roll in the semantic representation of concrete objects, like animals and tools. However, the parts of the brain that express semantic structure on audio trials appear to be very different—the selections are non-overlapping. On audio trials, semantic structure is found along the hub-and-spoke modelr temporal gyrus (STG), and in the posterior middle temporal gyrus (MTG). The MTG is thought to particularly involved in representing concepts related to motion, both biological and mechanical, as well as being relevant to tool concepts. The STS is an important part of the language network.The left lateral prefrontal cortex (PFC) and right dorsolateral PFC also appear to reliably express semantic structure. The frontal lobe is generally thought to participate in cognitive control, including concept retrieval and selection among conceptual alternatives and irrelevant context (Hoffman et al., 2015). These areas were shown to be active during the catch trials in the univariate analysis, in keeping with their typically ascribed function. It is interesting to note that, in a more passive condition, that these areas are not reliably more active than baseline, in a univariate sense, but do seem to correlate with semantic structure.

The complete lack of overlap among localizable semantic similarity structure is interesting, as well as the absence of representational structure in the ATLs. When the statistical threshold is relaxed, the degree of overlap between the visual and audio semantic information maps remains low—however, a lower threshold reveals

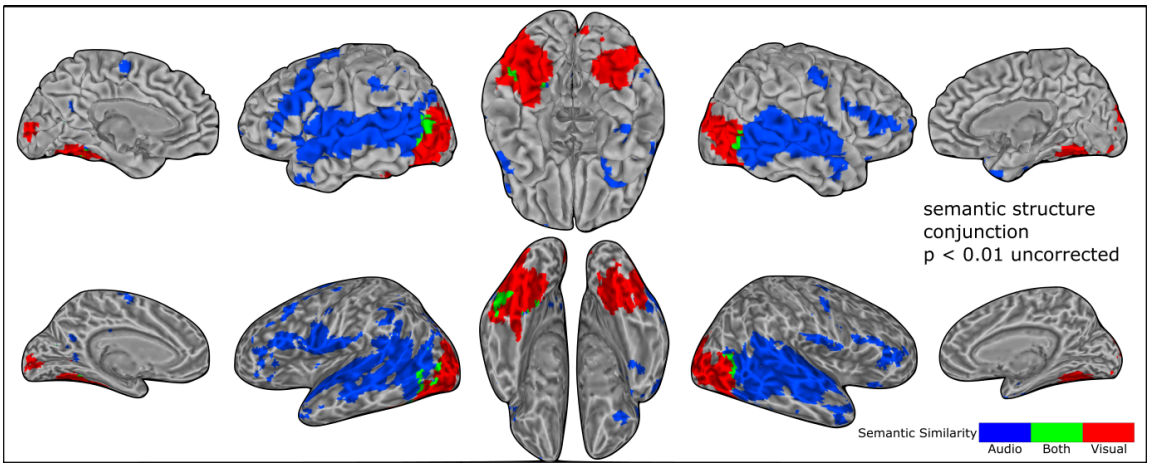

Figure 5.4: Searchlight conjunction plot, each component thresholded at uncorrected $p < 0.01$

that, on audio trials, the extent of areas that are more weakly correlated with the target semantic similarity structure expands considerably (Figure 5.4). The posterior extent comes to include the angular gyrus, and the anterior extent reaches down into lateral and even ventral ATL, both bilaterally. Even some overlap in pvTL seems to emerge. However, although these results would be very interesting if reflective of something true about the brain, the lack of statistical control coupled with the compounded distortion of location information (8mm searchlight solutions, interpolated into a common space and then further spatially smoothed) warns against drawing strong conclusions.

Averaging activity or structure

An alternative way to look for brain regions that express the cross-modal semantic similarity structure that is stable across stimulus modalities is to average the audio and visual datasets. Critically, this can be done in two ways. The first, most obvious approach is to average the beta coefficients for each item across modalities. This will boost signal in individual voxels that have similar response profiles over items. Alternatively, the representational similarity matrices constructed within each searchlight can be averaged together. This will enhance the fidelity of the

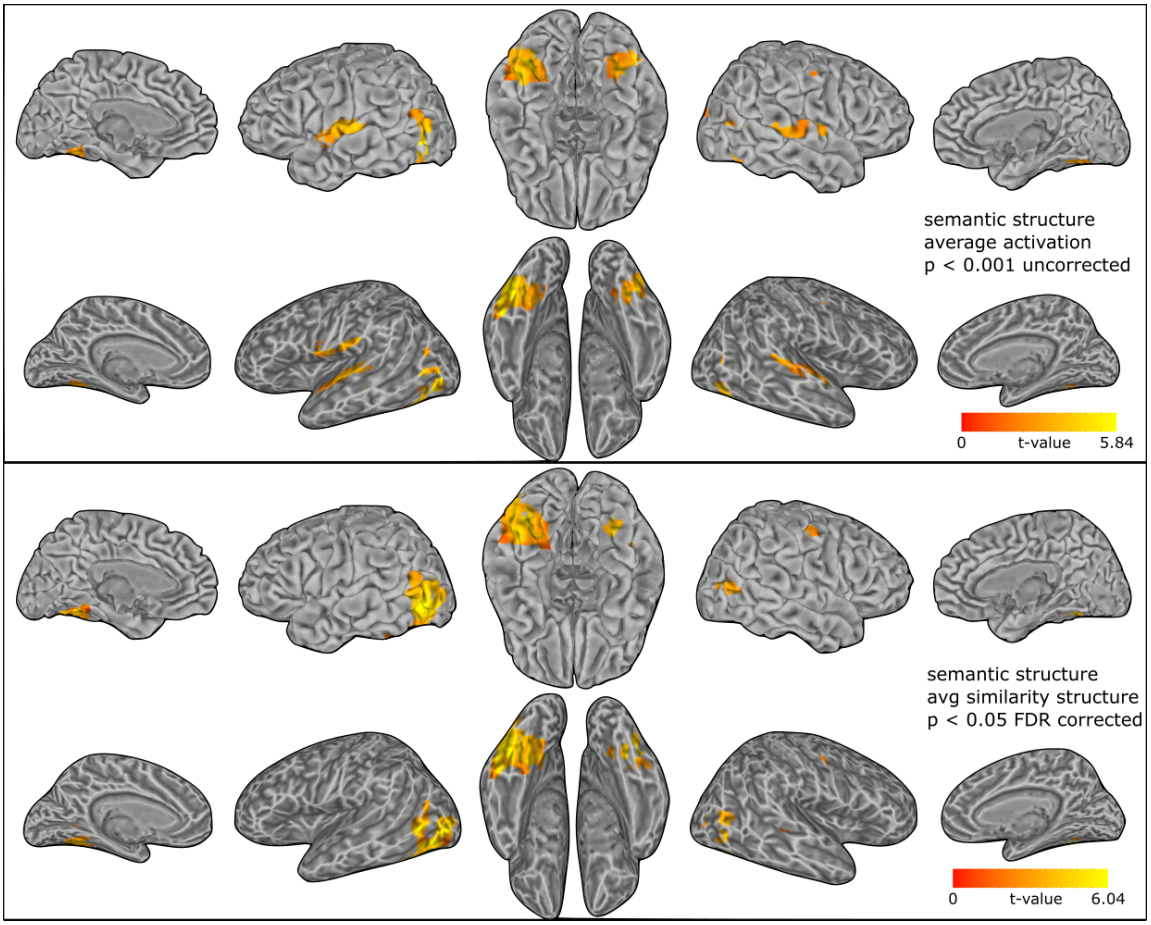

Figure 5.5: A follow up searchlight analysis, in which data were averaged across modalities in one of two ways before comparing the resulting similarity structure to the target semantic matrix. Searchlight analysis was peformed in parallel on data from each modality, each visiting the same voxel at the same time. The beta values in each searchlight were either averaged together directly, or a representational similarity matrix was generated for both sets of betas before averaging these resulting structures. The top row shows the result of doing the former, and the bottom row the latter. Note that the analysis of averaged betas resulting in a group level statistical map that could not be appropriately FDR corrected without excluding all voxels. While the averaged beta analysis echoes the univariate and searchlight conjunction analyses, the averaged structure analysis suggests that the similarity structure in the pvTL is more similar across modalities than in the STG. This may be because the size judgment is in part a visual, concrete-object oriented task.

similarity structure when they are similar across modalities, making them easier to detect. It is possible that these two approaches could lead to very different solutions.

In Figure 5.5, the results of this comparison are displayed. The top row shows the group level map associated with averaging over betas, and the bottom row for averaging over similarity structures in each searchlight. While averaging over betas seems to reiterate a diluted version of the familiar story, averaging over similarity structures suggests that the similarity structure in the pvTL is more similar across modalities than in the STS. However, averaging over local similarity structures does not change the story about the ATL, does not seem to locally express the target similarity structure.

Network RSA

So far, the results suggest that this task recruits many of the same regions typically identified in studies of semantic processing, and that local representational similarity structure appears to be restricted to areas that are relatively modality specific and not in the ATLs. However, the searchlight RSA makes very restrictive assumptions about how similarity structure is encoded: it must be localized, and it must be very transparent in the raw correlations among patterns of activity for each item. Network RSA, in contrast, allows for structure to be sparsely encoded, even across multiple regions, and for the relative importance of each voxel to be weighted, as is common practice in MVP classification analyses. Will this more flexible approach identify voxels that encode similarity structure in ways that we have previously been unable to measure?

The network RSA methodology is laid out in detail in the methods section, above. After obtaining appropriate values for the model's two free parameters, separately for each subject and cross validation fold within each analysis reported,

9 whole brain models were fit to each subject's visual and audio data (separately) and for each target structure (visual and semantic). Each of the 9 models were trained while withholding a different (mutually exclusive) set of trials to use as a cross validation set. Each model's prediction error was assessed with respect to these withheld trials, ultimately producing 9 error terms that are averaged together. This is the expected value for the error of network RSA when modeling under that particular combination of conditions (trial modality, target structure, parameter selections, and participant).

The following results will depict solutions maps in two ways that involve different summaries of each voxel's contribution to the overall solutions. Voxels will be described in terms of their *stability* or their *node strength*. A voxel's inclusion in a group-level solution map will determined by a *binomial test* conducted with respect to a non-parametric description of the voxel value with respect to an empirical null distribution obtained via permutation analysis. See the methods section for more detail on each point. Table 5.1 reports the average model performance over subjects for each whole brain model we fit, in standard deviation units. In these standard units, 0 is chance performance, and a good standardized error will be significantly negative. Chance performance and the standard deviation of the null distribution are determined based on the permuted models (see methods). Model error was determined to be normally distributed over permutations, which justifies scaling in this way. All models perform well above chance. Do keep in mind, however, that group level solution maps displayed in many of following analyses are have been thresholded based on correspondence across subjects, and do not necessarily reflect the models that generated this performance.

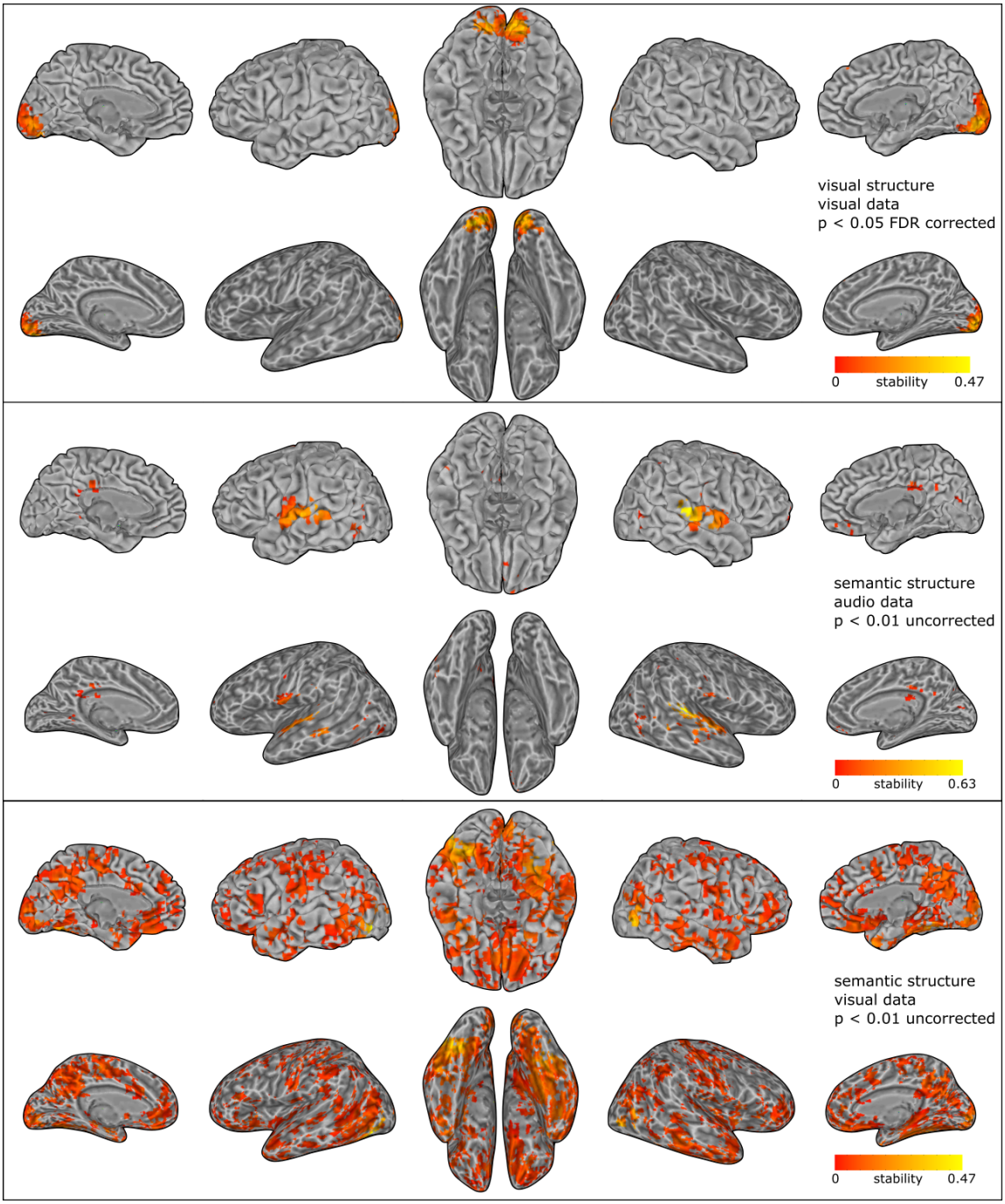

Figure 5.6: Network RSA solution maps displaying the mean stability of each over subjects. The maps are thresholded with respect to the binomial test based on empirically estimated null distributions, as described in the methods. Note that only the maps displaying the solutions for visual models of visual structure could be appropriately statistically thresholded at FDR corrected $p < 0.05$.

modality	model	mean	t(22)	SE	p-value
audio	semantic	-1.79	-7.34	0.24	<0.001
visual	semantic	-2.20	-10.30	0.21	<0.001
visual	visual	-3.06	-8.52	0.36	<0.001

Table 5.1: Whole brain network RSA error

Whole brain analysis

In Figure 5.6, stability maps are thresholded by binomial permutation test. Stability maps simply express which voxels are most reliably selected over variations in the training set, on average over subjects. In the top panel, we see that, again, even with whole brain scope, network RSA localized low level visual structure to early visual cortex when thresholded at FDR corrected $p < 0.05$. When modeling semantic similarity structure expressed over the whole brain in the audio trials, the largest concentration of stable voxels we located in the STG, similar to prior searchlight analyses. A few voxels from the lateral and ventral posterior temporal lobe are also identified. Also identified are bilateral posterior cingulate, which is active is a wide variety of contexts including spatial attention (Mesulam, 1990; D. Small et al., 2003) and visual imagery (Burgess, 2008; Hassabis, Kumaran, & Maguire, 2007; Johnson, Mitchell, Raye, D'Esposito, & Johnson, 2007), and ventral medial PFC which has been associated with motivation and reward processing (Bechara, Damasio, & Damasio, 2000; A. R. Damasio, 1994; Drevets et al., 1997; Mayberg et al., 1999; Phillips, Drevets, Rauch, & Lane, 2003).

When modeling semantic structure expressed over the visual trials, however, things look radically different. While the most stable voxels are to be found in vpTL and LOC, an extensive collection of other regions are also implicated, including the anterior temporal lobes bilaterally. Despite how extensive the map is, it does not identify structure in the STS, which is the primary site identified when considering the semantic models of audio data. Yet, in its relative sparseness, the semantic

model of *audio* data does extend into the middle and inferior temporal gyrii. This again could be taken to be consistent with the "wheelie bin" task being very visiospatial in addition to being semantic. Audio trials may involve activating high level audio structure in the STS as a part of activating the concept, and once the concept it retrieved, the visiospatial reasing involved in comparing the concept to a wheelie bin will require activating these posterior temporal, high level visual areas.

The plots in Figure 5.6 favor voxels that are very reliably selected, but the stability metric does not express about the role a voxel plays in the overall representation other than it is reliably useful to reducing error. Network RSA conveys a great deal more information about each voxel and its position in the distributed *network* supporting the representations that express the target structure. Each voxel selected by network RSA can be considered as a node in a network, that has edges that link it with other nodes. The W matrix (see Chapter 4) contains estimates of the importance of each node and how they relate to one another—in other words, the edge and vertex weights of a graph structure. One way to characterize the importance of a node is in terms of the sum of the magnitude of the edge weights that are connected with it. This statistic is referred to as *node strength*. Rather than describing each voxel in terms of how reliably it is selected for inclusion we can instead describe them in terms of their node strength.

While conceptually distinct, stability and node strength are bound to be somewhat correlated. Strong nodes are influential players in the representation, and so models will do well to include them. Alternatively, imagine that they are some uncorrelated in the case of some voxels: a voxel with very high node strength is only selected during a single cross validation fold. Since the voxel was only selected once, the every other time the voxel was given a zero value. After averaging, the expected value of the node strength will be small.

In the present analysis, plotting the node strength results in maps that are very

similar to those based on stability. They are shown in Figure 5.7. However, again, it may be possible to over-interpret these models. The semantic model of visual data is surprisingly inclusive, and none of the semantic models were associated with parametric maps that could be appropriately FDR corrected.[4]

A major hypothesis of the hub-and-spoke model is that the ATL is a semantic hub. This implies that units in the ATL should have larger node strengths than voxels in posterior areas. At first blush, Figure 5.7 may appear to disconfirm this prediction: if anything, it appears that node strengths are larger in the pvTL than in the ATL. However, this figure may be misleading. For instance, it may be the these posterior voxels are more densely sampled over subjects than the anterior voxels, and so interpolating, blurring, and averaging across subjects results in node-strengths that appear large at the group level, but in individual subjects are only moderately sized. Meanwhile, ATL voxels may be sparsely sampled, overlap less across subjects, and end up appearing small in the group map.

To inspect whether voxels in or near the ATL have larger node strength than voxels outside the ATL, each subject's individual node strength maps (before interpolation, blurring, or group level averaging) are projected onto the plane defined by the anterior–posterior (AP) and inferior–superior (IS) axes. Within each subject, a point in the ventral anterior temporal pole was determined, and the Euclidean distance from this point to every identified voxel on the AP-IS plane was computed. These distances were grouped into 10 equal bins defined by radial distance from the temporal pole point, and node strength was averaged by bin within each subject. Figure 5.8 plots mean node strength for each subject and radial bin as a function of distance from the temporal pole. Although crude, this visualization does not show the expected pattern with larger node strengths in the ATL. Instead, for models fit to visual data, the largest mean node strength is very distant from

[4]Because these models are by nature sparse, cluster thresholding is not necessarily a viable alternative.

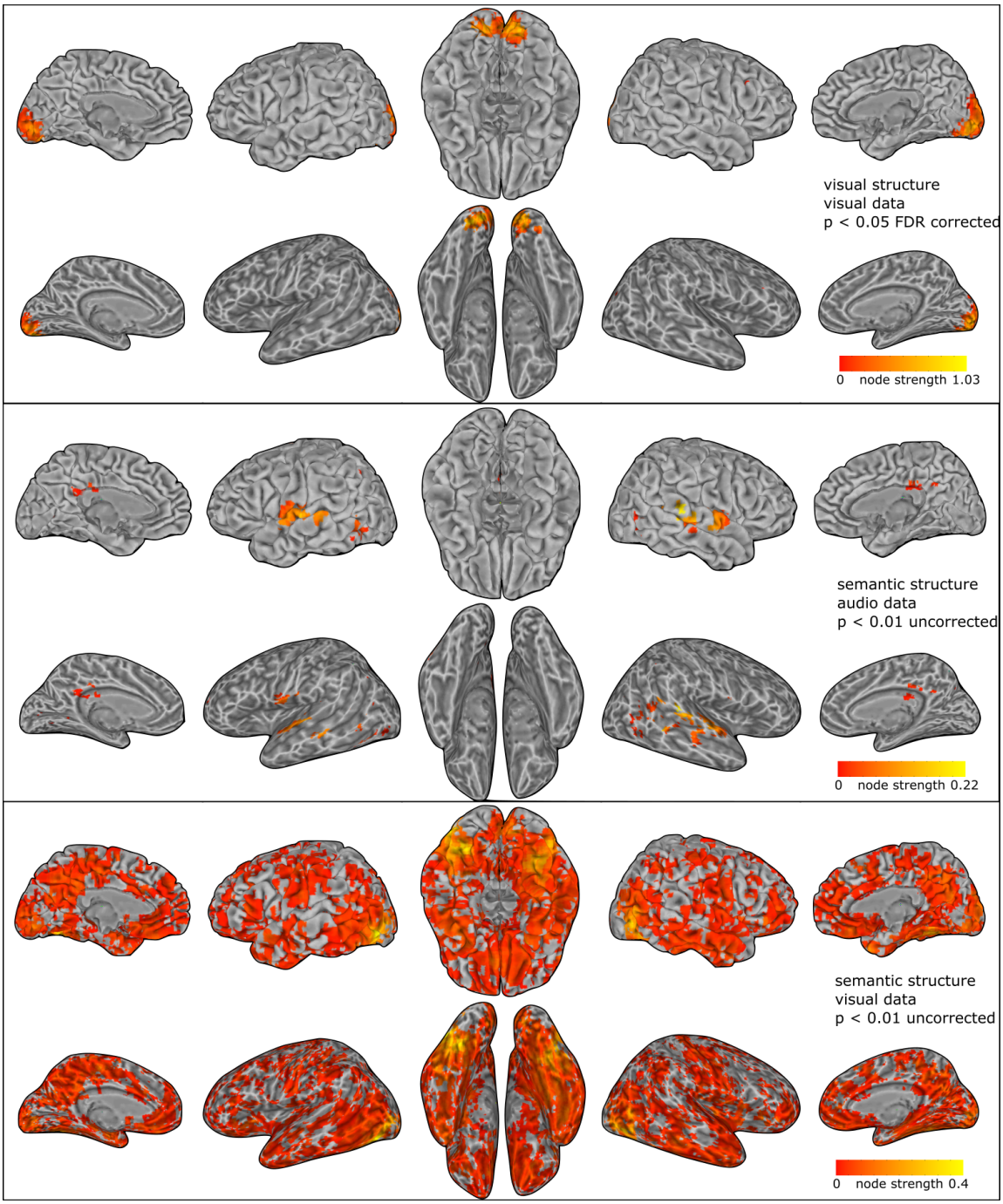

Figure 5.7: Network RSA solution maps displaying the mean nodestrength of each over subjects. The maps are thresholded as in Figure 5.6.

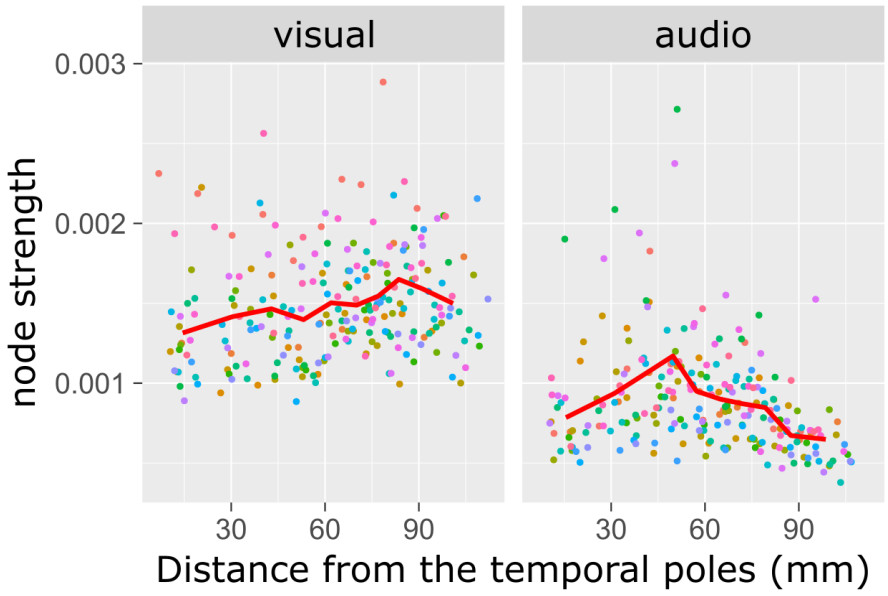

Figure 5.8: Node strength as a function of distance (mm) from temporal pole along the AP and IS axes. Each colored dot is the average over a bin of distances equal to 10% of maximum distance from the temporal pole within each subject; each color is a different subject, and position along the x-axis is the mean distance within each of the 10 bins. Peak nodestrength can be seen to roughly correspond to "spokes", rather than the "hub". The red line is tracks the mean over subjects at each bin.

the ATL, corresponding to the vpTL, and for models fit to audio data, the largest mean node strength is at a moderate distance from the temporal pole, corresponding to STG. In short, the individual subject data appears roughly consistent with the group maps with regard to the location of peak node strength.

However, the group maps presented so far have all relied on permutation distributions to provide an estimate of the null map against which statistical significance can be assessed. These maps resulting from the permuted models are surprisingly structured, however, for reasons that are not yet fully understood—either fragments of the target structure are surviving the permutation procedure, or there is a confound such that certain voxels will be selected regardless of the target structure, in data with certain degrees of covariance. The permutation maps, averaged by condition, are shown in Figure 5.9. Because there are still complications with the statistical procedure, I will present maps that are not referenced to the permu-

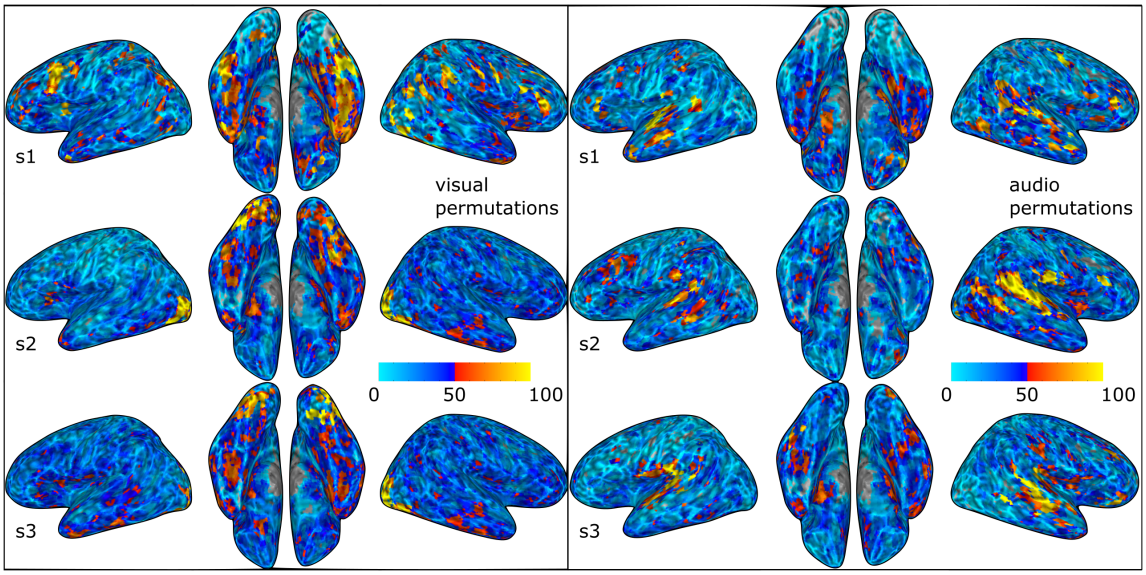

Figure 5.9: Overlap maps for 100 permutation models for subjects 1, 2, and 3 (representative sample). Voxels are not sampled at random by permutation test, and are clearly affected by the modality of the stimulus.

tation distribution at all. Figure 5.10 shows straight-forward subject overlap maps. The top row shows the overlap of the average model for each subject over cross validation folds, which were used in the group level statistical analyses reported previously. The bottom row shows the overlap considering just a single model for each subject, for a single model (that is, without averaging over hold sets). The color at each voxel tracks how many times that voxel was implicated over subjects after interpolating each subject's solutions map into the common space. No additional spatial blurring was applied. Regions showing the most overlap over subject are the same as those implicated in the statistical analyses. When considering the average model for each subject (which effectively counts all voxels selected on any cross validation fold with equal weight within each subject), it is not surprising the overlap is fairly widespread. In this case, the audio maps include ventral temporal lobe, including the ventral temporal lobe, but these selections are apparently not very common: the single-model overlap maps do not indicate that voxels are selected in the ventral temporal lobe in 6 or more subjects.

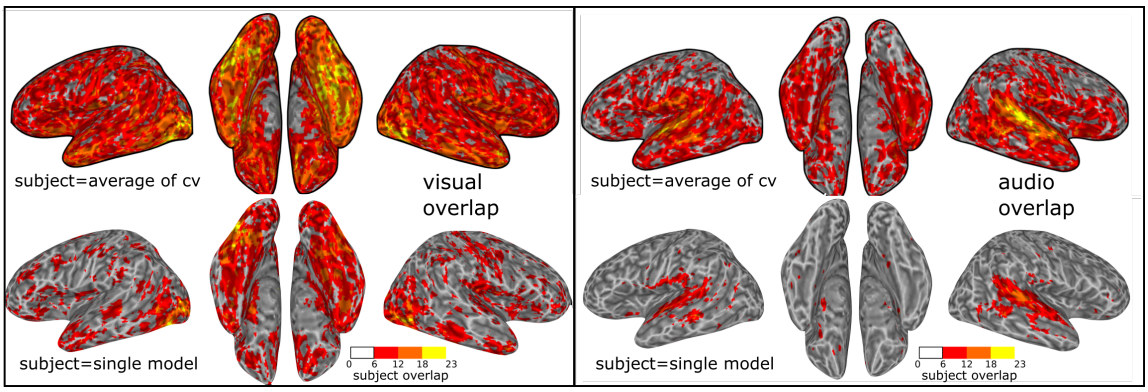

Figure 5.10: Network RSA overlap maps. Voxel color indicates the number of subjects for which that voxel was implicated after interpolating to the common space. The top row shows overlap across the average model for each subject; the bottom row shows overlap across a single model per subject. In these figures, no additional spatial smoothing is applied. For maps corresponding to the stability of voxels in individual subjects, see Appendix A.

In summary, this series of whole brain analyses has demonstrated that the network RSA procedure is capable of identifying both well localized and radically distributed representations in the brain. While the visual model and the semantic model of audio data resemble solutions obtained with other methods, the semantic model of visual data marks a major departure from the standard set of results.

ROI and lesion analysis

The *semantic hub* variant of the hub and spoke hypothesis predicts that conceptual knowledge is supported by pan-modal representations encoded in the bilateral ATLs, which if appropriately measured and be interpreted without information about activation states elsewhere in the brain. The *hub+spokes* variant, in contrast, predicts that the ATL extracts and encodes the interactions among the modality specific spokes, but does not locally re-represent modality specific structure that can be encoded by the spokes. This means that the hub and the spokes together jointly support the concept—they cannot be determined by ATL activity alone.

The preceding whole brain analyses were not restricted to just considering the

audio and visual spokes (in the STS and pvTL, respectively) and semantic hub (ATL). Although model performance was significantly better than chance, it is difficult to determine from these analysis, particularly at a group level, which combinations are regions expressed the structure that supported this performance.

In an attempt to more directly test hypotheses about the role of the ATL in supporting semantic representation, three regions of interest were defined, guided by anatomical segmentations for each subject automatically generated using the Freesurfer image analysis suite (see methods), prior literature, and the expert input of my dissertation committee. All are defined the same in both hemispheres. The definitions are mutually exclusive, meaning that the audio and visual spokes only extend anteriorly up to the point where the semantic hub begins.

1. *Audio spoke*: The hub-and-spoke modelr temporal gyrus, transverse temporal gyrus, and Heschel's gyrus.

2. *Visual spoke*: The fusiform gyrus, lingual gyrus, inferior temporal gyrus, and other structures of the posterior ventral temporal lobe.

3. *Hub:* The ATL is not a single monolithic structure (Binney et al., 2010; Binney et al., 2012; Sanjuán et al., 2015; Visser et al., 2012; Visser & Ralph, 2011), and so does not have clear anatomical demarcation. Jackson et al. (2015) focus their rTMS at three locations within $10mm$ of the time of the temporal pole. Likewise Nestor et al. (2006) defined the middle temporal gyrus into anterior, middle, and posterior by splitting it into three $10mm$ sections with the posterior commissure separating posterior from middle sections. On the other hand, the most common areas of hypometabolism in patients with semantic dementia extend, on average, back to around MNI $y = -21$ in the STS and $y = -39$ in the fusiform gyrus (Binney et al., 2010; Nestor et al., 2006). We have defined that ATL by a plane, defined in MNI coordinate space. The

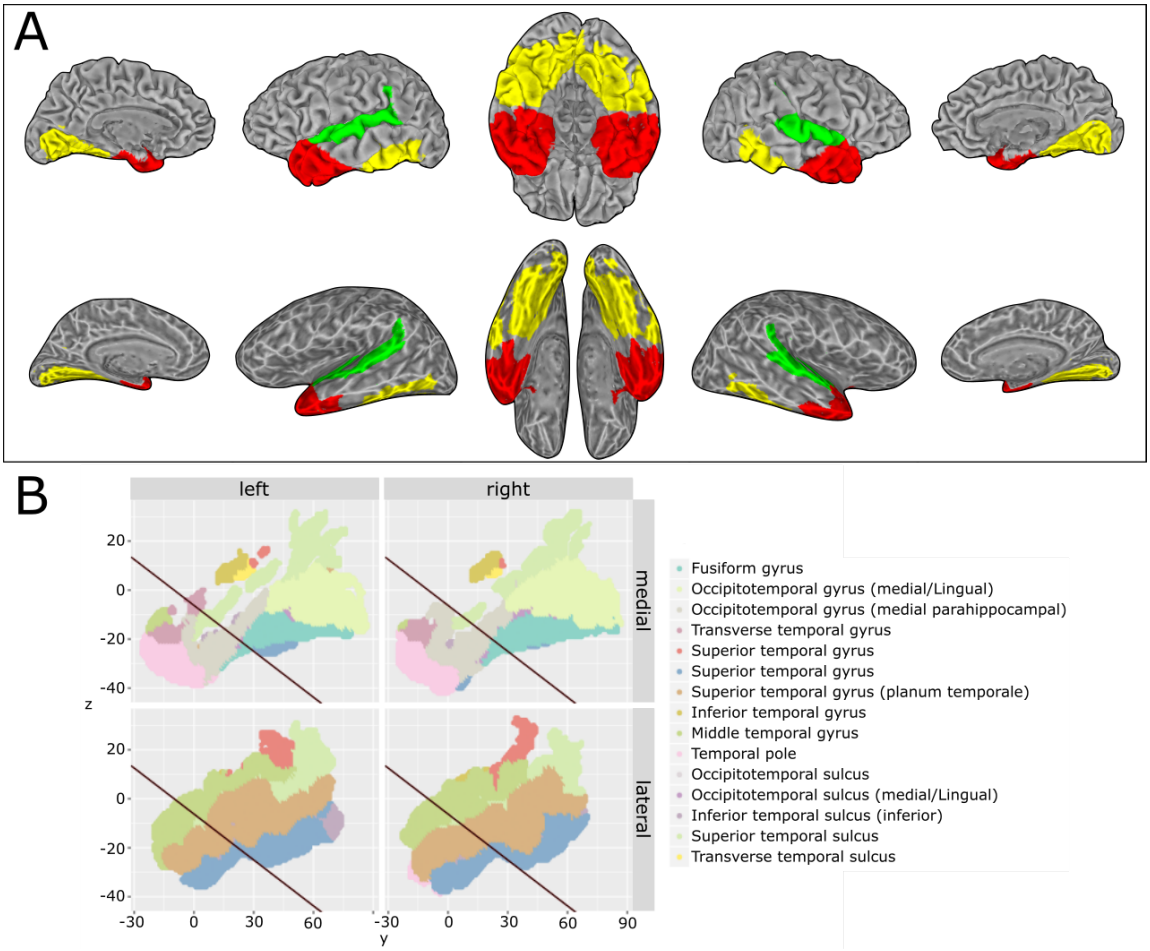

Figure 5.11: The definitions of the hub and spokes that will be used in the following "ROI" and "lesions analyses that follow. See the text for details on how they were defined.

plane intercepts the anterior posterior axis at $y = -10$ mm and has a slope of -1.6 mm. The plane is perpendicular to the anterior–posterior axis. See Figure 5.11B for a depiction of the plane relative to the y- and z-axes in MNI space, with lateral and medial projections of the Destrieux atlas (Destrieux et al., 2010) for reference.

Figure 5.11A presents all three ROIs on for a single subject. Using these ROIs, two sets of analyses were completed. The first uses the ROIs to select voxels in different combinations, and the second uses the ROIs to exclude voxels in different combinations. These are referred to informally as *ROI* and *lesion* analyses respectively, although those terms typically have other more rigorous meanings in the

literature. The claim here is not that omitting voxels is analogous to a brain lesion in terms of the claims that can be made about the functional organization of the brain. Rather than thinking of them as brain lesions, think of them as *model* lesions: the brain remains in tact, but network RSA is denied access to aspects of the functional patterns it expresses.

The ROI analyses include analyzing restricted versions of the dataset, comprised of:

1. The semantic hub (S) and both the audio (A) and visual (V) spokes.
2. The semantic hub and only the audio spoke.
3. The semantic hub and only the visual spoke.
4. The visual and audio spokes, excluding the hub.
5. Each ROI, individually.

Each of these restricted datasets were defined for visual and audio trials separately and modeled with network RSA to discover sets of voxels that encode semantic similarity structure. The left panel of Figure 5.12 shows the error associated with the various ROI analyses, as well as the error associated with the whole brain (WB) analysis for reference. Each subjects errors (colored dots) are standardized with respect to 100 permutations; red dots indicate the mean over subjects. Table 5.2 reports the error and t-statistics associated with each model.

There are several important points to make about this pattern of errors over different restricted datasets. First is that semantic structure can be decoded from the modality specific spoke that corresponds to the format of the input, but not vice versa. For instance, decoding from *just* the audio spoke is better than chance when stimuli were presented as characteristic sounds, but not when presented as images. To be concise, I will refer to the spoke that corresponds to the stimulus

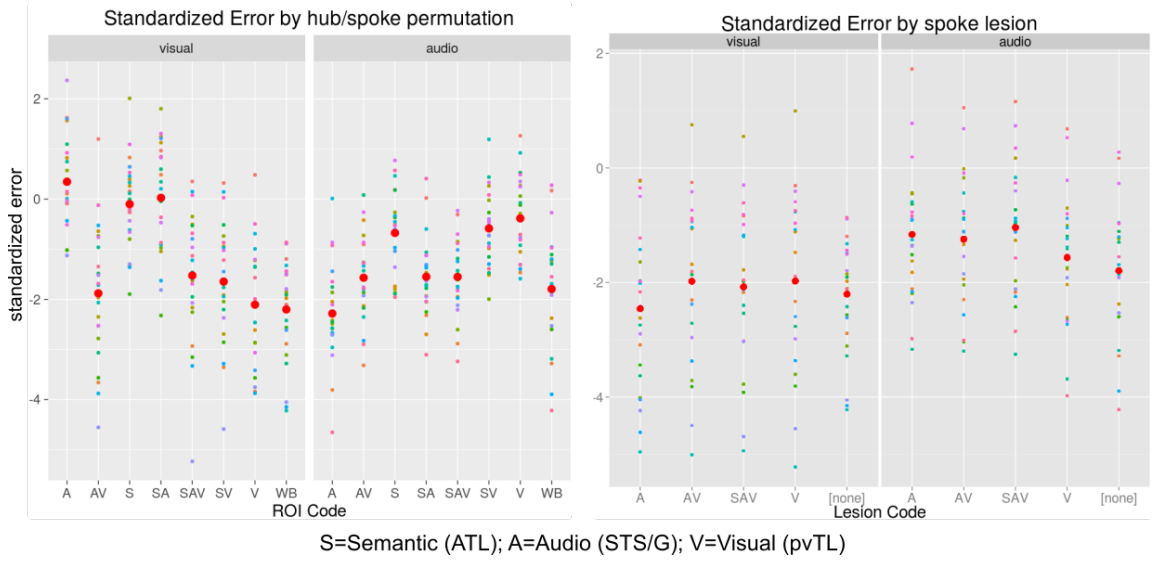

Figure 5.12: Error for Network RSA analyses.

modality	ROI code	mean	t(22)	SE	p-value
audio	A	-2.28	-11.24	0.20	<0.001
audio	AV	-1.57	-8.88	0.18	<0.001
audio	S	-0.68	-3.99	0.17	<0.001
audio	SA	-1.55	-9.42	0.16	<0.001
audio	SAV	-1.55	-9.70	0.16	<0.001
audio	SV	-0.58	-3.76	0.16	0.001
audio	V	-0.38	-2.28	0.17	n.s.
audio	WB	-1.79	-7.34	0.24	<0.001
visual	A	0.35	1.82	0.19	n.s.
visual	AV	-1.88	-6.65	0.28	<0.001
visual	S	-0.10	-0.54	0.19	n.s.
visual	SA	0.03	0.11	0.23	n.s.
visual	SAV	-1.52	-5.62	0.27	<0.001
visual	SV	-1.65	-6.53	0.25	<0.001
visual	V	-2.11	-8.42	0.25	<0.001
visual	WB	-2.20	-10.30	0.21	<0.001

Table 5.2: ROI error. Uncorrected $p < 0.001$ exceeds the Bonferroni criteria for 16 tests.

modality as the *congruent spoke*. Particularly remarkable is that the performance from just the congruent spoke is equivalent to a model fit to all voxels in cortex.

These ROI analyses also reveal a pattern of results that seems to contradict the whole brain solution maps studied in the previous section. Namely, although whole brain network RSA of audio trials did not reliably select voxels in the ATL, when restricted to only voxels in the ATL network RSA is able to identify sets of voxels in the region that encode the target semantic structure on average over subjects ($t(22) = -3.99$, $SE = 0.17$, $p < 0.001$). The opposite appears to be true when considering visual trials: although whole brain network RSA *did* tend to select voxels in the ATL, when trained exclusively on voxels from the ATL a model cannot be fit to the data that makes better than chance predictions, on average over subjects ($t(22) = -0.54$, $SE = 0.19$, *n.s.*). This raises additional important questions about our permutation and statistical procedures for determining voxel selections.

It also raises questions about how semantic representations are being supported by these regions and the rest of cortex. Particularly when considering the visual trials, the spoke region seems necessary to the semantic representation. On the other hand, the ATL seems to carry semantic structure on audio trials, but combining the hub and congruent spoke does not improve over the spoke alone.

To help test the necessity of each ROI, a series of "lesion" analyses were performed, where each group of voxels was omitted. This resulted in four additional restricted datasets, comprised of:

1. All of cortex minus the audio spoke (lA).
2. All of cortex minus the visual spoke (lV).
3. All of cortex minus the both spokes (lAV).
4. All of cortex minus the hub and both spokes (lSAV).

modality	lesion code	mean	t(22)	se	p-value
audio	lA	-1.16	-4.89	0.24	<0.001
audio	lAV	-1.24	-5.20	0.24	<0.001
audio	lSAV	-1.04	-4.33	0.24	<0.001
audio	lV	-1.56	-6.47	0.24	<0.001
audio	[none]	-1.79	-7.34	0.24	<0.001
visual	lA	-2.46	-8.29	0.30	<0.001
visual	lAV	-1.98	-6.61	0.30	<0.001
visual	lSAV	-2.07	-7.23	0.29	<0.001
visual	lV	-1.97	-6.34	0.31	<0.001
visual	[none]	-2.20	-10.30	0.21	<0.001

Table 5.3: Lesion error. Uncorrected $p < 0.001$ exceeds the Bonferroni criteria for 10 tests.

The model errors associated with each of these four restricted datasets are shown in the right panel of Figure 5.12, along with the error associated with the whole brain ([none], indicating "no lesion") model. Table 5.3 reports the error and t-statistics associated with each model. Model performance is above chance in all cases—even when removing the hub *and* both of the spokes from the data. This indicates that, while the congruent spoke is sufficient for decoding, it is not necessary. Semantic structure can be obtained from patterns of activity in regions beyond the temporal lobe.

5.4 Discussion

These experiments were approached with an eye towards testing a central prediction of the hub-and-spoke model, namely that the anterior temporal lobe is a pan-model hub that contributes to semantic representation regardless of stimulus modality. This prediction was decomposed into three nuanced hypotheses that are consistent with the overall architecture of the hub-and-spoke model, but make different predictions about how concepts are represented over the network and the specific contributions of the ATL. In particular, they variously predicted that (network) RSA would 1) not discover semantic structure in the ATL, 2) would discover

semantic structure entirely expressed by the ATL, or 3) discover semantic structure that was supported over multiple regions, and because the ATL only encodes interactions among spokes, the ATL would not be identified unless considered in combination with other areas.

Although the convergence hypothesis is technically consistent with the hub and spoke architecture, it runs counter to the computational principles that motivate the model. It is also unclear how such a model of semantic cognition would account for the patterns of impairment seen in, for example, semantic dementia. Critically, it predicts that the ATL should not express semantic similarity structure in its patterns of activation. However, at least when stimuli are presented aurally, the ROI analysis reported above provides evidence that such structure is encoded in the ATL, even when the region is considered in isolation. The hub+spokes hypothesis was tested by whole brain network RSA analyses and by contrasting performance on restricted datasets that included just the ATL or the ATL plus the STG/S (audio spoke) and vpTL (visual spoke). Specifically, the hub+spoke prediction is that semantic structure is expressed in the interaction of the hub and that spokes, so including both should lead to a superior model than when including just one or the other. This was not observed.

Taking these points together, the evidence so far seems to lean in favor of the semantic hub hypothesis. But this conclusion is not decisive. The evidence for semantic structure in the ATL is inconsistent, where it was observed the effect size was relatively small (standardized mean=-0.68, relative to -2.28 in the congruent spoke on audio trials), and the whole brain network RSA solutions maps resisted clear interpretation.

Although only weakly supported by the work reported above, prior multivariate work in fMRI has also identified semantic structure in the ATL (Bruffaerts, Dupont, et al., 2013; Clarke & Tyler, 2014; Fairhall & Caramazza, 2013; Liuzzi et

al., 2015; Peelen & Caramazza, 2012). And despite higher node strength not being correlated with distance from the temporal poles, the ATL has been shown to have extensive intrinsic connectivity temporal and parietal regions (Binney et al., 2012), and that these connections appear to become disordered in patients with semantic dementia (Guo et al., 2013).

It should also be emphasized that similar studies have revealed quite different results than the what we observed in the work reported above. Visser and Ralph (2011) conducted a cross modal study where participants were presented with a picture, a spoken name/label, or an environmental sound for each of a common set of concepts. While they found that the anterior superior temporal gyrus was only active in that auditory conditions, they found that all stimulus modalities elicited increased activity in the ventral anterior temporal lobe. Given the close experimental similarity, what can account for the different results? Visser and Ralph (2011) contrasted all semantic trials with an *active* non-semantic control condition. This may be a critical distinction—because passive "resting" states are likely filled with idle thought which relies on the semantic system, an active baseline may more clearly isolate the task-relevant semantic activity (Binder et al., 1999, January; Binder et al., 2011; Binney, Hoffman, & Ralph, 2016; Jackson, Hoffman, Pobric, & Ralph, 2016). While this is a clear concern for univariate contrast analyses looking to assess the conjunction between multiple semantic conditions, it may also be consequential to MVPA and RSA. Future work will explore this possibility.

In light of the rest of the literature, the results of this chapter are remarkable in at least two related senses. Unlike many others in the literature, we do not see that the ATL contributes significantly to semantic structure among visual stimuli. On the other hand, this effect *is* observed for the structure among non-word auditory stimuli. This has not been demonstrated before. Given that prior work has identified the region in visual and verbal tasks, demonstrating the semantic structure

is active in the ATL for audio stimuli might be taken as evidence for cross modal convergence. Our lack of a visual effect in the ROI is confounding in this regard. Whole brain network RSA maps may appear to implicate the ATL on visual studies, as does an inspection of the data in individual subjects (see Figure A.1 in Appendix A), but this has not been shown at a group level with appropriate statistical rigor.

Exciting work which relies on ECoG datasets have demonstrated that there are in fact cortical sites of multimodal convergence in the ATL (Abel et al., 2015; Shimotake et al., 2014). Even more recently, Y. Chen et al. (2016) reported that a spatio-temporal searchlight RSA has detected semantic structure in the patterns of activity over implanted grid electrodes in the same dataset studied by Shimotake et al. (2014).

This motivates an interesting hypothesis: perhaps the ATL represents semantic structure, but does so in a code that changes dynamically over time. If this were true, then what is visible to ECoG would be invisible to fMRI.

An important difference between the dataset I have reported above and others in the literature to which RSA has been applied is the category structure of the stimuli. In most studies the chosen stimuli are selected so that clusters of items neatly separate into categories, as necessary for contrast and classification analyses. The target similarity matrices employed are often binary, and often involve a relatively small number of distinctions. Regions of the brain that respond somewhat categorically, tending to activate more for one stimuli over the other, may correlate reasonable well with these similarity structures. In contrast, our stimuli were not chosen to be strongly categorical (see Figure 5.1). This may be holding the patterns of activity "to a higher standard", as it requires more subtle structure to be expressed, which are more easily drowned out by the inherent noisiness of fMRI.

It should be pointed out that the "ROI" and "lesion" analyses reported above do not test the full hub-and-spoke model, which of course would have included other modal regions and regions known to be associated with action and motion representations throughout the parietal lobe and middle temporal gyrus. The analysis was focused on the spokes associated with the modalities directly tested in this study, but stimuli were from a variety of categories including some that would be associated with motions, actions, and learned manipulation routines. These parts of the brain may be critical to the representation, and the semantic model of visual data touches on them—however, this map is so inclusive and resistant to appropriate statistical thresholding, that it is hard to draw strong conclusions.

Returning attention to the ATL, both possible methodological confounds (poor signal to noise and low temporal resolution of fMRI) can be better addressed by considering data collected by other means which are less subject to these concerns. In the next chapter, I will extend analyses performed by Shimotake et al. (2014) using the same ECoG dataset.

Chapter 6

ECoG Experiments

In the previous chapter, major predictions of the hub-and-spoke model were evaluated against an fMRI dataset that had many desirable qualities for a study aiming to discover patterns of neural activity that express cross-modal semantic representations. The data were distortion corrected (Ajay D. Halai et al., 2015; Ajay D Halai, Welbourne, Embleton, & Parkes, 2014; Visser, Embleton, et al., 2010) to enhance clarity in the anterior temporal lobes (ATL), and a common set of concepts were referenced with stimuli presented either visually (line drawing) or aurally (characteristic sound), and participants completed a semantic task involving a relative size judgment. While the functional activity evoked by trials of each kind independently expressed the target semantic structure among the relevant concepts, the areas that expressed this structure were not held in common as the stimuli modality changed.

Furthermore, evidence of the ATL's contribution to semantic representation was inconsistent. At the end of the previous chapter, I discussed several reasons why this may have been so. Some relate to experimental limitations: the stimulus set was small (37 items) and stimuli were sampled to emphasize the distinctiveness of their characteristic sounds. Others relate to limitations inherent to fMRI: the sig-

nal is noisy and temporally smoothed because each image acquired is separated by roughly 2 s. In this final set of experiments, I turn my attention to an ECoG dataset previously reported by Y. Chen et al. (2016) and Shimotake et al. (2014). Shimotake et al. (2014) provided direct evidence, through the ability to measure and stimulate the anterior temporal lobe directly using the same electrode grid, for cross modal integration in the region. Y. Chen et al. (2016) have demonstrated that a spatiotemporal searchlight reveals that patterns of activity over the electrode grid reliably express the target semantic similarity structure among visually presented stimuli (defined by feature norms very much like those used in the experiments reported in the previous chapter). Notably, they show that both living vs. nonliving classification performance and representational similarity expression begin to peak at around 200 ms after stimulus onset, *and remain level* for over a second.

At first this may seem to suggest that activations "come on and stay on". However, it is important to keep in mind that both classification performance and representational similarity can be supported by very different patterns of activity. In one sense, this was one of the critical concepts demonstrated in the simulations reported in Chapter 3. However, it could also be due to redundant coding, or perhaps each aspect in the dynamics emphasizes different dimensions of the semantic similarity structure and they are cycled through. A significant correlation between two similarity structures (i.e., the one expressed by the brain and the target structure determined by hypothesis) does not imply that they *match*. It only implies that they are more similar than would be expected by chance. This means it is possible that two different structures could express similarity that correlates with the target, yet convey different information about the stimuli. Indeed, this is precisely the hub+spoke hypothesis detailed in Chapter 1 and discussed at length in this thesis, except that the hub plus spoke prediction is that this happens over space.

The prior demonstration of Y. Chen et al. (2016) involved performing RSA with

target similarity matrices based on semantic features and superordinate semantic categories (living and non-living). Performance on these two models were similar for most time points. The correlation between the target similarity matrix and neural data peaked at around $r = 0.03$ in both cases, which might be achieved even if the neural signal only expressed relatively coarse categorical structure. In our unpublished own work with this data, it was not possible to train classifier models with LASSO to perform sub-category classification among the living things (mammals vs. non-mammals). In short, while the existing evidence is consistent with high-order semantic structure encoded in the ATL, it has not been demonstrated that the region actually encodes the fine-grained structure as predicted by the hub-and-spoke model,

In this final chapter of experiments, I test three hypotheses: 1) is semantic structure encoded in local field potentials evoked by naming visual stimuli and sampled from the ATL using ECoG, 2) does this structure contain meaningful sub-category structure, and 3) does this code have an important temporal component.

6.1 Materials and methods

These data have been reported elsewhere (Y. Chen et al., 2016; Shimotake et al., 2014). Please visit those publications for further detail about the patients and surgical procedures. Much of the study descriptions below are reproduced from Y. Chen et al. (2016).

Participants

Eight patients with intractable partial epilepsy (eight) or brain tumour (two, one associated with intractable partial epilepsy) participated in this study. Background clinical information about each patient is summarized in Table 1. Subdural elec-

trode implantation was performed in the left (seven) or right (one) hemisphere for presurgical evaluation (mean 83 electrodes, range 56–107 electrodes/patient). 6–30 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 (Morrison et al., 1997 and Snodgrass and Vanderwart, 1980). 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 (the patients' average naming time was 1190 msec) and each session lasted 8 minutes 20 seconds. Participants' 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. Trials with greater than ± 500 μV maximum amplitude were considered 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. Out of the full set of 10 subjects, 8 have been analyzed so far.

Data analysis

Both mutivariate classification with spatio-temporal LASSO of living and nonliving items and representational similarity analysis with spatio-temporal network RSA were performed.

LASSO analysis

LASSO was performed for all subjects over all electrodes simultaneously in a "moving window" procedure. A 50 ms time window was defined and, beginning with the window from 0–50 ms, a model was fit using LASSO. Then the window was slid forward in time by 10 ms, and the process repeated. This procedure was repeated 10 times, holding out a different set of 10 items each time to facilitate cross validation. This analysis confirmed the finding that there is useful signal over the 1000 ms time window following stimulus onset (Y. Chen et al., 2016). Subsequent analyses focus on this range.

Finally, once models were obtained for each time window, these models were applied to make predictions given the pattern of activity at each other time window. The mean error on the *training sets* for each combination is used to populate a full matrix of training and test windows.

Network RSA analysis

Network RSA was performed over all electrodes and over the time range from 200 ms to 1200 ms post stimulus onset. While during the LASSO classification analysis we modeled ever 1 ms time point, due to the large window size of the network RSA analysis data were reduced by average every consecutive 10 ms time window to a single data point for each electrode. The target semantic structure was determined by referencing the Leuven Concept Database (Deyne et al., 2008). These feature vectors were constructed in a similar way as Dilkina and Lambon Ralph (2012), which were the basis for the semantic features in the previous chaper. Sixty-three out of the 100 stimuli in this study had a corresponding entry in this database, meaning that only this smaller subset could be modeled with network RSA. The semantic similarity matrix was generated by computing the cosine similarity among the feature vectors for each of the 63 items (42 living, 21 nonliving). This cosine similarity was then embedded in a low dimensional space via the same procedure as in Chapter 5. An error threshold of 0.3 yielded a 5 dimensional embedding.

As with the fMRI analysis, an accurate interpretation of the error term requires standardizing by the mean and standard deviation of an empirical null error distribution obtained by fitting 100 models to randomly permuted versions of the target embedding.

6.2 Results

Network RSA analysis

The analysis has a simple logic: if the neural activity is conveying semantic structure via a time-varying code of some kind, averaging over large windows of time will destroy meaningful signal. On the other hand, a spatio-temporal network RSA

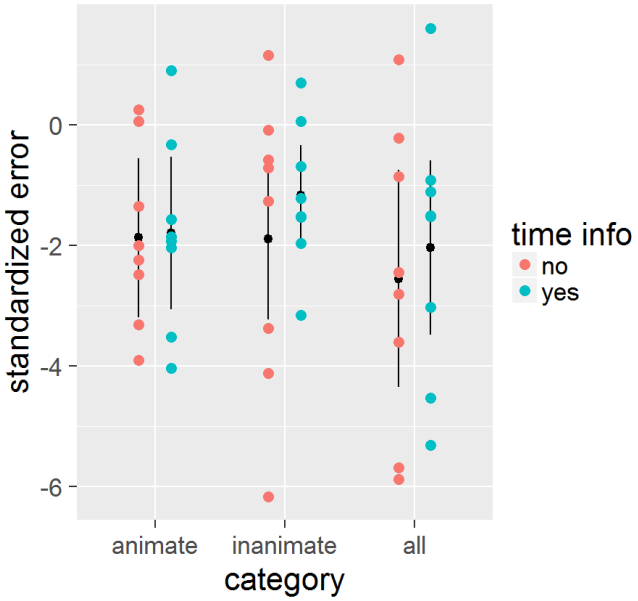

Figure 6.1: Error associated with two network RSA models fit to the raw ECoG local field potentials (LFP) over electrodes. In one case (time info=no), the time points from 200 to 1200 ms were averaged together for each electrode. This corresponds to the window of peak performance, both for Y. Chen et al. (2016) and in our own work involving superordinate classification (no shown). In the other (time info=yes), time was not averaged down to a single point. Instead, aside from moderate data reduction (consecutive 10 ms bins of LFP were averaged), all time points were entered into the model at once. This means that if different points in time express different elements of structure, they could be linearly combined to produce a single, more complete structure and allow a better fit to the target similarity structure. Error is shown for the whole dataset, and living and nonliving items separately (to show subcategory structure is learned). This manipulation of time information did not yield significantly different performance. Colored dots indicate subjects by time manipulation, black dots indicate the mean, and error bars correspond to 95% confidence intervals for the test of the mean against zero. Errors are standardized with respect to the permutation models.

analysis will be able to model the similarity structure and take advantage of representational structure that may exist at different points in time. The prediction the model ultimately makes will take all of these time points into account with an appropriate linear combination, rather than naively averaging. If the signal has meaningful structure at several points in time, the spatio-temporal network RSA will outperform a network RSA performed on the average of those time points.

These two analysis procedures were completed for each subject, with cross validation. The error for each subject was standardized relative to an empirical null

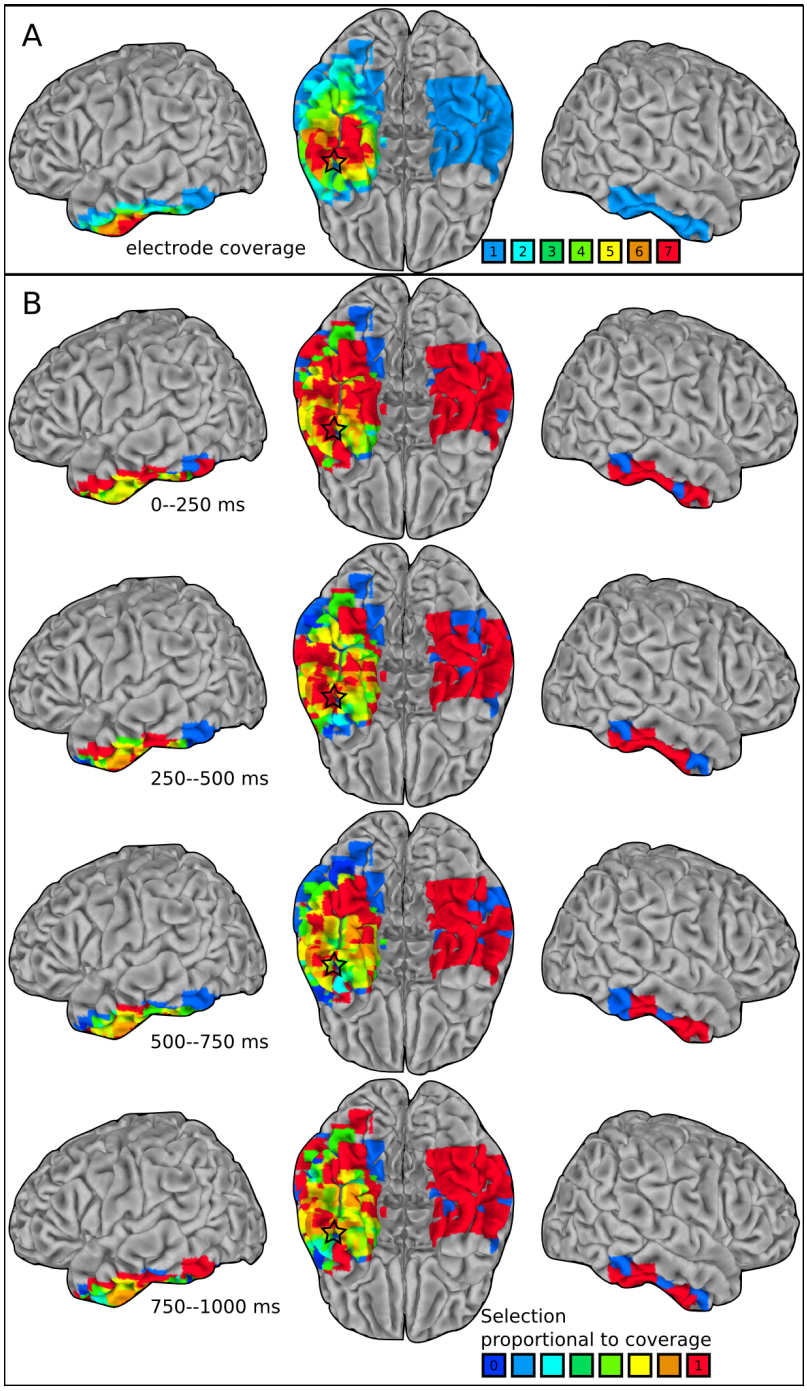

Figure 6.2: A) Electrode coverage over subjects, subject to a 4 mm FWHM Gaussian blur. One subject had electrodes implanted in the right hemisphere, so maximum overlap of coverage is 7 subjects in the right hemisphere. This blur is applied to make it comparable to the maps presented in B) which display the proportion of overlap at each electode for network RSA solutions obtained by modeling the target semantic structure in the LFPs collected with ECoG. To compose these maps, I first aggregated over cross validation models and consecutive 250 ms windows and applied a 4 mm FWHM Gaussian blur before assessing overlap across the 8 subjects. The count at each voxel was then divided by the number of times an electrode existed at that location, as shown in panel A).

distribution determined by the 100 permuted models (see methods). In both cases, we can ask whether 1) network RSA can capture meaningful semantic structure in the data, 2) if this structure exists within categories or only at the superordinate level, and 3) if these answers differ as a function of whether network RSA is allowed to freely combine information from different points in time or not.

The results are presented in Figure 6.1. Performance does not differ significantly as a consequence of whether or not time information is available. This indicates that the representations in the ATL that carry semantic structure may be fairly stable over time. However, network RSA does identify structure both within and across living and nonliving superordinate categories. This is consistent with Clarke and Tyler (2014), who found that the ventral medial ATL expressed semantic similarity structure that corresponded better with a rich semantic feature space as opposed to coarse category structure. It also enhances our interpretation of the original searchlight RSA analysis of these ECoG data presented by Y. Chen et al. (2016).

Although electrode coverage overlaps across patients most completely in the more anterior regions of the ventral temporal lobe, each patient has somewhat idiosyncratic coverage that, in some cases, extends into what is decidedly posterior ventral temporal lobe (see Figure 6.2A). It is therefore important to consider the network RSA solutions in more detail to determine that the models are reliant on LFPs sampled from the anterior areas. Figure 6.2B shows the voxel selection maps, for four equal time windows post stimulus onset (250 ms each). These maps are scaled by the number of subjects that have coverage at each point of the map (after applying a 4 mm blur to account for the fact that grid arrays are not perfectly aligned across subjects). Areas in red indicate that an electrode at roughly that position was selected in every patient for which there is available data. Of course, this metric is most sensitive where coverage overlaps the most over patients. For-

tunately, peak overlap occurs at the point in the anterior ventrolateral temporal lobe currently thought to be the "core" of the semantic hub (Binney et al., 2016; Jackson et al., 2016; Patterson & Lambon Ralph, 2016; Rice, Hoffman, & Lambon Ralph, 2015; Shimotake et al., 2014). A star has been overlaid to mark MNI (-39, 6, -39), the point chosen for the vATL seed by Jackson et al. (2016).

From these maps, we can see that the anterior aspects are important with a very high rate of consistency over time and subjects—electrodes in this area are virtually always selected. This indicates that even though some models had access to information in posterior ventral temporal lobe, the structure expressed by the vATL was important to accurately fitting the target semantic structure.

LASSO analysis

Another way to test whether signal is stable over time is to train a model at one time point and test at other time points. Because of the large number of models involved, I performed LASSO to classify all 100 items into living (50) and nonliving (50) categories. The full error matrices associated with training at each time point and testing each model at all other time points are shown in Figure 6.3. Each row is a trained model, and each column a timepoint where it was evaluated. Note that time each time window was 50 ms large, trained at ever 10 ms step through the data, so neighboring windows overlap by 40 ms.

While it is clearly possible for models trained at many time points to generalize over time, there are other time points that seem to generalize less well. This can be more clearly seen in the context of an analysis proposed and initially conducted by Tim Rogers (unpublished personal correspondence). Each row of the matrices shown in Figure 6.3 can be thought of as the "accuracy profile" of a model trained at a particular time window. Models with similar accuracy profiles must be tapping into similar structure with similar sets of weights (because they perform sim-

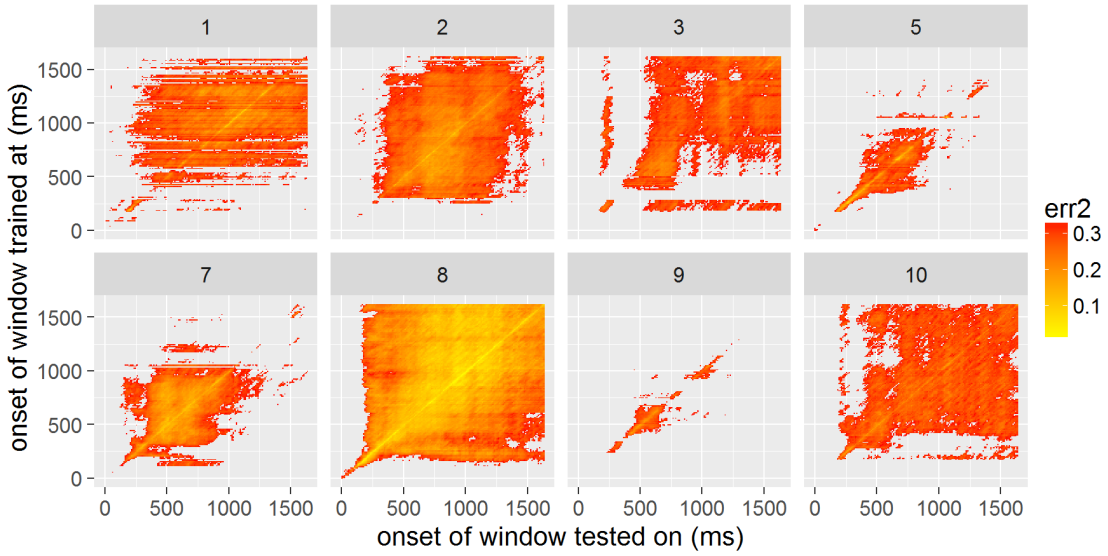

Figure 6.3: Binary classifiers fit with LASSO were trained on each time point, and then evaluated on every other time point. Maps are thresholded based on a binomial test of the error out of 90 items, where the probability of obtaining an error less than 45 under the null hypothesis is $p = 0.5$. Points are thresholded at uncorrected $p < 0.001$. Each panel shows the matrix of train- and test-time windows for one subject.

ilarly in the context of particular patterns over the LFPs). Thus, we can simplify Figure 6.3 by averaging models with similar response profiles, as determined by a cluster analysis. In the following, a simple k-means clustering was performed with k=5 (clustering accounts for greater than 85% of total variance in all subjects, and 88% on average over subjects), and so 5 mean temporal average profiles are displayed for each subject in Figure 6.4. Colored dots along the top of each figure indicate the points in time that were averaged to form each average accuracy profile of the same color.

Here, it is more clear to see that different clusters of models may perform better at different times, which indicates that the underlying code may be dynamically shifting. Consider patients 1, 5, and 8: each has a cluster of models that perform well (binomial $p < 0.001$) briefly early in the trial, before showing marked declines in performance over the rest of the time windows. At a more relaxed threshold, the same pattern is observed in patient 2. This is consistent with a rather stark

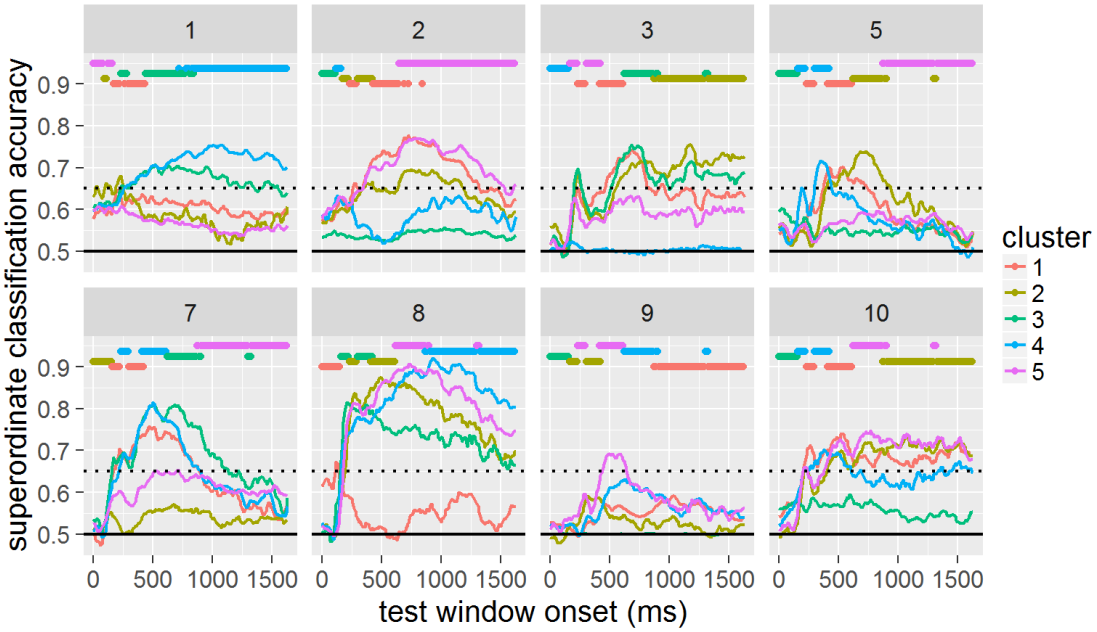

Figure 6.4: Accuracy profiles, based on a k-means clustering of the temporal error maps shown in Figure 6.3 (k=5). Each colored line is a average of several rows from the previous figure (exactly which rows are indicated by the corresponding colored dots along the top of each panel), with error inverted into accuracy. The dotted line indicates $p<0.001$ by binomial test as a fairly conservative reference for above-chance generalization.

change in the underlying activity. On the other hand, patients 1, 2, 5, 7, and 8 also show signs of more gradual, graded shifts over time (overlapping followed by gradual separation of performance for clusters 3 and 4 in patient 1, clusters 1 and 5 in patient 2, clusters 1 and 2 in patient 5, clusters 3 and 4 in patient 7, and the very intersting cascade of clusters 2, 3, 4, and 5 in patient 8). On the other hand, patients 3 and 10 show remarkable stability over time over several clusters.

However, there are typically one or two clusters that generalize well to many time points, which means at least some aspects of the code are relatively stable. Indeed, if this were not the case, the ability to discover structure after averaging over all time points as reported in Figure 6.1 would be rather difficult to comprehend. In short, there seem to be interesting dynamics that merit a great deal of further investigation, but there may be enough stability to support signal that is identifiable by fMRI.

6.3 Discussion

In this chapter, I considered three basic hypotheses about the semantic structure expressed in the ventral ATL. The first was that, contrary to what was seen in the fMRI analysis in Chapter 5, the ATL contains semantic structure that is active and relevant when processing visual stimuli in a semantic task (in this case naming). This was confirmed by the fact that network RSA can learn a model of the LFPs sampled from electrodes implanted in this region in 8 patients that performs significantly better than chance on predicting where out of sample stimuli should be positioned within the low rank embedding of the target similarity structure. The second hypothesis was that this it would be possible to recover within-category semantic structure, and this was confirmed by noting that within category predictions were also better than chance. The final hypothesis was that the ATL encodes semantic structure with a time-varying code that expresses different semantic dimensions at different points in time. This was generally not supported by the data. Averaging all time points within the window of peak decoding reported by Y. Chen et al. (2016) did not reduce or eliminate the ability of network RSA to model the semantic structure, indicating that the content being expressed is relatively stable over time.

The follow up analysis with LASSO took a different angle, considering if a model fit at one point in time will generalize to other points in time. Here we noted interesting dynamics around a reasonably stable core: the temporal accuracy profiles often indicated some time points (particularly early in the trial) that do not generalize well to other time points, which indicates that the underlying activity supported classification but with different patterns of activity than at other time points. On the other hand, models trained later in the trial do tend to generalize to many other time points, sometimes 500 ms into the the future and into the past. Taken together, these data do not robustly support the hypothesis that semantic

structure in the ATL was difficult to detect with fMRI due to it being expressed in a highly dynamic code.

It is important to note, however, that time series data contain a wealth of useful information which we have not considered here. It is typical, for example, to perform a spectral decomposition when analyzing high-temporal resolution data as a way of isolating different components in the time varying signal. These analyses conceive of fluctuations in the data as a mixture of sine waves of various frequencies. By decomposing a time series into its constituent frequency bands, it may become easier to isolate the aspects of the signal that carry information of interest. In other words, there may be be structure expressed in the time frequency domain that cannot be directly assessed in the raw LFP data.

There are many interesting open questions regarding the time course of semantic representation (Hauk, 2016; Jackson et al., 2015). Specifically in the context of the hub-and-spoke model, it may hold some of the keys to understanding why pan-modal semantic representations remain so elusive.

Chapter 7

General Discussion

In the preceding chapters, I framed important questions about the neural representation of semantic knowledge and specifically the role of the anterior temporal lobes (ATL) in supporting pan-modal concept representations, and then attempted to address these questions through experiments. Among the novel contributions of this work is the application of network RSA, a novel analysis technique for discovering representational similarity structure in sparse distributed networks of activity. Network RSA enables us to test hypotheses that involve representational structure encoded over multiple brain regions. This dissertation specifically focused on the representational predictions of the hub-and-spoke model, but virtually all contemporary hypotheses about the neural representation of concepts involve multiple brain regions working together. Network RSA would therefore be able to identify semantic structure, regardless of whether it was integrated in a hub region or if different parts of the structure are encoded throughout the brain.

The convergence hypothesis predicted that the ATL does not encode semantic structure, but that such structure would be expressed in a widely distributed network involving cortex in generally modality specific areas. The ATL contains convergence zones that facilitate reactivation of fragments of information encoded

over this vast network for specific unique entities, but does not encode similarity structure. Both the fMRI and ECoG datasets provide evidence that semantic similarity structure is expressed in the ATL, and so this critical prediction appears to be incorrect.

The semantic hub hypothesis, in contrast, predicted that the bilateral ATL support a domain general and cross modal semantic space, and so semantic structure should be present in the neural activity of the ATL during semantic tasks. However, it also predicts that the ATL functions by integrating and re-representing information expressed and coded, in isolated fragments, in other areas of the brain. Thus, the semantic hub hypothesis also predicts that a whole brain RSA would be able to identify patterns of activity expressing semantic similarity structure either over all the spokes taken together, or in the semantic hub in isolation. The hub may express this structure more concisely, which might make it more likely to be discovered by sparse logistic regression techniques (see Chapter 3), all else being equal. Of course, signal in the ATL may be less strong and more difficult to image, resulting in noisier response profiles in neuroimaging data, encouraging whole brain models that prefer the spokes. In short, a whole brain analysis of a brain adhering to all predictions of the semantic hub hypothesis might be expected to come out in various ways. And, indeed, it these whole brain maps ultimately do not provide compelling support for the hypothesis.

What *does* support the semantic hub hypothesis is the ability to recover semantic signal from the ATL in isolation, as demonstrated for audio stimuli in the fMRI experiment and for visual stimuli in the ECoG experiment. But, to merely demonstrate the semantic structure is available in ATL activation is not definitive evidence. The hub+spoke hypothesis does predict that *some* semantic structure will be present in the ATL on its own. On the hub+spoke account, ATL activation will be associated with cross modal interactions, and not express modality specific

structure. So, while the hub+spoke hypothesis predicts that different aspects of semantic structure are encoded across the hub and the spokes without redundancy, making the full structure available *only* if information is considered across all regions at once, is not incompatible with above chance performance of a model fit just to activity in the ATL. However, the fact that quite fine grained structured was available in the ECoG dataset, supporting accurate predictions about sub-category structure for both living and non-living things, is consistent with ATL representations being rather complete.

It is becoming clear that studying the representational structure that is present in the ATL will require devising a range of semantic target structures, expressing different components of the complete semantic structure. In the experiments reported in this dissertation, we considered semantic structure and low-level visual structure. However, considering whether the ATL represents higher order visual or audio structure, or whether it is better fit to a semantic structure with components that might be expressed by models of "spoke-level" representation removed than to a complete semantic structure. At present, it is difficult to discern whether even fine grained semantic structure can be conveyed in the absence of any perceptual structure that, under the hub+spoke account, would be represented beyond the ATL.

However, given the data we have available, the hub+spoke hypothesis predicts that semantic structure should be better expressed by the interaction of hub and spokes, rather than just in the hub or just in the spoke alone. This was not the case in either modality of the fMRI experiment: the hub and spoke together did not exceed the spoke on its own. Taken together, the evidence seems somewhat more consistent with the semantic hub hypothesis, in which the full semantic space is represented in the ATL.

The series of experiments reported here additionally emphasize the important

of the spokes to semantic representation (Patterson & Lambon Ralph, 2016). Although the importance of the ATL and the computational utility of a semantic hub have been thoroughly discussed, there is also considerable evidence that the spokes convey substantial semantic content in their own right. Retrieving concepts, even when only cued by the word, is associated with both brain activation in sensory or motor regions(Hauk, Shtyrov, & Pulvermüller, 2008; Pulvermüller, 2005; Simmons et al., 2007). Lesions near these same areas result in impairment relating to the aspects of knowledge that can be expressed in that modality on its own (Boulenger et al., 2008; Pulvermüller et al., 2010). Representational similarity analysis of the posterior ventral temporal lobe has demonstrated that some, but critically not all, semantic structure seems to be encoded in the reason (Mur et al., 2013). These all point towards important semantic contributions in a wide range of specialized regions, which then require a common hub to reveal cross modal structure.

While important progress has been made through the work reported in this dissertation, much remains to be done. Aside from developing tractable solutions to the statistical challenges facing whole brain network RSA, there are exciting open questions about how the brain supports semantic knowledge. One important aspect of semantic knowledge that is not often considered in cognitive neuroscience is that the meaning of a word, the interpretation of an event, and the function of an object can all be highly context dependent. Rather than a concept such as "dog" evoking a specific representation time after time, every encounter is different and our semantic system will very flexibly integrate the dog into appropriate contexts, and support constrained generalizations.

Representational similarity analysis has the potential to be a very sensitive method for assessing these shifts in representational structure. Given an appropriate set of stimuli, one would expect the items to become conceptually more similar or dis-

tinct as a function of changing context, and this should have corresponding shifts in the neural code. Peelen and Caramazza (2012) conducted an interesting experiment along these lines, in which they presented participants with images of tools that are either used in the kitchen or the garage by either squeezing or twisting them. Participants completed a one back task, where they were instructed to either monitor for repetitions of location or usage in different blocks. In other words, participants considered all the same stimuli, but in on block they cared *where* the items are used, and in another they cared *how* the items were used. This might be expected to change the underlying representations as expressed by patterns of activity over cortex. Their results do not bear out this prediction: their ROIs based on Brodmann areas 20 and 38 (the ventral temporal lobe and temporal poles, respectively) were sensitive to both kinds of information, regardless of whether the information was task relevant. Nevertheless, this line of questioning bears further research. For instance, their ROIs were very large, and representational structure was not modeled with a technique like network RSA, so there might be relevant differences in structure between the two conditions that were overwhelmed by noise and went undetected. Correlations between the target semantic similarity structure and the similarity structure expressed in activity within each ROI were low—$r < 0.1$—and the test only involved 12 items. Furthermore, the semantic similarity structure was very simply, encoding two dimensions in binary fashion (garage vs. kitchen and squeeze vs. twist). While it is exciting that they identified representational structure that expressed these dimensions, more nuanced questions about task effects on these representations are still wide open.

7.1 Estimating semantic structure

In order to ask these questions well, we need tools and methods for estimating semantic similarity structure in ways that capture the potentially high dimensionality if the space, are sensitive to context effects, and which can be targeted to specific populations in the service of representing individual differences. These are tall orders. Collecting feature norms, word associations, and category fluency responses are labor intensive on the part of the experimenter and depend on the participants to generatively map out the semantic space. In the case of feature norming, it is well understood that participants do not list all important features of cued concepts, even omitting ones that are quite obvious because, in that context, there is nothing driving attention to that aspect (Dilkina & Lambon Ralph, 2012; Hoffman & Lambon Ralph, 2013; Rogers et al., 2004). On the othe hand, corpus analytic techniques like latent semantic analysis (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997) are more data rich, and can assess the utilization of words across a huge range of contexts. However, such broad scope may gloss over semantic space may shift and change as a function of context.

A different angle would be to give participants a simple task that requires them to assess the *relative* similarity among pairs of items. Examples might include collecting similarity data using rating scales (Lee, Pincombe, & Welsh, 2005), multiple item arrangement (Kriegeskorte & Mur, 2012), sorting tasks (Rosenberg & Park Kim, 1975; Shaver, Schwartz, Kirson, & O'Connor, 1987) or forced-choice tasks (Jamieson, Jain, Fernandez, Glattard, & Nowak, 2015; Navarro & Lee, 2002). Carlson, Ritchie, Kriegeskorte, Durvasula, and Ma (2014) compare semantic spaces estimated from human judgments and multiple corpora and databases to one another and to representations in inferior temporal cortex. In particular, the utilization of a forced-choice task has been pursued fruitfully by our research group, and the problem of learning a similarity space adaptively is being studied (Jamieson

et al., 2015). The particular task involves comparing two stimuli to a point of reference, and judging which is more similar to the referent. The relevant dimensions of similarity can be specified by the researcher, imposing a particular context within which the judgments will be made, or left ambiguous. Each trial may involve a completely new set of stimuli, options and referent, which may emphasize completely different semantic dimensions.

There are several benefits to this approach. For one, while the task probes semantic structure from a variety of angles, each individual judgments is a fairly simply task. Participants do not need to consider all dimensions at once, as when listing features, or be clever in their word associations. As with other behavioral tasks, it can be targeted to specific individuals and populations, but relative to other behavior tasks there are no responses to code, spelling errors to correct, or feature labels to consolidate, making it an easier experiment to conduct.

A related method is the multiple item arrangement task introduced by Kriegeskorte and Mur (2012). In this task, participants arrange a large number of images or printed words on a screen, so that more similar items are closer together. Similarity structures estimated with this technique for individual participants has been shown to reflect individual differences that bear our in their patterns of neural activity (Charest, Kievit, Schmitz, Deca, & Kriegeskorte, 2014). Nevertheless, this technique may impose some limitations on participants: assessing many items at once and imposing a two-dimension work space may artificially restrict the dimensionality of the semantic space.

How best to estimate semantic structure is clearly an exciting topic of active research. Progress in this area will have important consequences for study of the neural basis of semantic knowledge and cognition.

7.2 Basis sets for meaning?

The preceding discussion about estimating semantic space considered some methods, like feature norming, that try to establish a list of properties that might be said to "define" a concept, and others that make no such attempt and simply aim to describe the similarity structure. The difference between these approaches is of deep theoretical significance, even if they both yield similarity structures that and be subjected to many of the same kinds of analysis.

In PDP models, distributed representations are composed of units that cannot in general be interpreted in isolation of the other units. This is distinctly different from neuroscience's roots in single cell recording. For example, the seminal work of Hubel and Wiesel (1959, 1968) characterized the receptive fields of neurons throughout striate cortex in cats and monkeys. They found that any given neuron responded to a particular, sometimes very narrow, set of visual stimuli. Different recording sites responded to different stimulus features, and different visual areas responded to more ot less complex combinations of features. Given a collection of "feature detectors" of this kind, complex visual representations can be expressed in the *population code* expressed by multiple feature detectors responding in unison (Singer & Gray, 1995)—all the information fragments are integrated into a whole. Population coding is an important feature of neural representation in all modalities (Averbeck, Latham, & Pouget, 2006; Pillow et al., 2008; Pouget, Dayan, & Zemel, 2000, November).

From a *decoding* perspective—fitting a model of neural activity that allows accurate prediction of a target structure—a population code and a PDP distributed representation can look similar. Both involve patterns of neural activity, where neighboring functional units may respond differently from one another, and the patterns of activity associated with a common stimulus may differ across people. The difference is that, unlike distribute representations, each element of a popula-

tion code is often considered to have a particular interpretation as a feature detector of some kind. On this perspective, one might aim to build an *encoding* model, which predicts how each functional unit will respond given the presence or absence of particular features. Encoding and decoding models are sometimes also referred to as forward and backward models. The difference is that forward models go from stimulus features to neural activity, while backward models go from neural activity to stimulus features. All MVPA and RSA analyses are decoding models. Standard univariate analyses of fMRI datasets are technically encoding models—the stimulus conditions and nuisance variables are used to predict the time series of each voxel. More sophisticated encoding models will involve more detailed stimulus information than a condition or category label.

Constructing an encoding model of visual stimuli or audio stimuli is relatively straight forward, in the sense that there is general consensus about the features that might be relevant to encoding content in each modality. For instance, visual stimuli might be encoded in terms of orientation, contour, luminance, and the like. Audio stimuli might be encoded in terms of energy in different frequency bands, amplitude, etc. The set of features that can be used to compose representations in a domain are called a *basis set*. Developing basis sets for visual and audio representations has strong neuroscientific motivation. However, this is not the case for all domains of interest. To highlight the most relevant example: is there such a thing as a basis set for semantic knowledge?

There are several examples of encoding models developed study the neural bases of semantic knowledge. The first model of this kind was presented by Mitchell et al. (2008). They defined a "basis set" of 25 verbs, and counted how many times each of a set larger set of nouns co-occurred with each of the verbs. So for each noun, there is a 25 element vector of numbers reflecting, in a sense, how related the noun is to each verb. These nouns were then presented to participants during

an fMRI study, which was pre-processed to result in a single beta map for each word. Then, the variance at each voxel was modeled with respect to the 25 "verb loadings" for a subset of the nouns. The model at each voxel can then be used to predict how the voxel will respond to nouns not included in the training set. Taken together, this results in a full brain map of predictions that can be compared to the true beta maps for each noun. They found that these predictions were more accurate than would be expected by chance (when comparisons are restricted to the 500 most stable voxels over stimulus repetitions). This might be taken to suggest that semantic knowledge might be at least partly encoded in terms of action.

Subsequent encoding models in the literature follow a similar methodology, differing primarily in the composition of the basis set. Just et al. (2010), for instance defined the basis set in terms of three principle factors: manipulation, shelter, and eating. Fernandino et al. (2016) normed each of 900 words with respect to their association with five physical attributes (sound, color, manipulation, visual motion, and shape) and treated these dimensions as a basis set. On the other hand, Huth et al. (2012) had participants watch hours of natural video, and coded every TR according to the presence or absence of 1,364 words and 341 categories, and modeled each voxel in terms or their response to these 1,705 entities. More recently, the same research group had participants listen to 2.5 hours of stories from *The Moth Radio Hour* and computed the co-occurance between every unique word in the stories and 985 common English words in (Huth et al., 2016). These studies are much less committed to a particular basis set of theoretical significance. While each research group brought a distinct theoretic perspective, they unanimously show that semantics are supported by a wide range of areas, and that individual units can have complex tuning functions responding to multiple basis dimensions.

Whether this approach to the study of the brain bases of semantics, marked by the effort to characterize the response profiles of individual functional units, is an

important point for theoretical discussion. From a purely PDP perspective, using encoding models to map out which functional units covary with a set of basis features may appear to be misguided. As has been rehearsed at multiple points in this thesis, PDP distributed representations are composed of units that cannot be interpreted in isolation, and the dimensions that are relevant for defining a PDP representational space are not known in advance. These points suggest that characterizing the response profiles of individual units in terms of a basis set is a lost cause. However, even if semantic knowledge is encoded by PDP distributed representations, encoding models can still be informative. The representational structure learned by PDP models can be embedded in low dimensional spaces, and these dimensions may correlate with theoretically relevant features. For example, the hub and spoke model would predict that representational structure in the visual spoke should be better described along visual dimensions like shape and color than representations in the audio spoke or the hub. If the first principle component of representational structure in the posterior ventral temporal lobe is shape, that means that an encoding model predicting activity in this regions with a basis set of shape features should be able to predict reasonably well. In other words, the success of encoding models does not necessarily imply that populations of neurons are tuned to particular basis features in an a priori sense, and may instead be picking up on learned dimensions among acquired distributed representations. Nevertheless, encoding models can give important insight into which dimensions are encoded in different brain regions.

7.3 Conclusion

This dissertation attempted to characterize the role of the ATL in supporting semantic representations, experimentally testing for the first time whether semantic

similarity structure is jointly encoded over multiple distant brain regions working together. This perspective was referred to as the hub+spoke hypothesis, to distinguish it from the semantic hub hypothesis which predicts that the ATL collects information from various sources and represents that content in a single, cross modal and domain general semantic space. When the hub and spoke model of semantic cognition was introduced to the literature as a neural network model, it was described quite explicitly in terms of that computational framework, which predicts a semantic space as described in the semantic hub hypothesis (Rogers et al., 2004). The most current descriptions of the theory, however, emphasize the importance of the hub and the spokes together in supporting semantic knowledge (Patterson & Lambon Ralph, 2016).

The results I have reported mark the first evidence that semantic structure of auditory sounds is encoded within ATL and that even fine-grained semantic structure for visual stimuli is likewise encoded in ATL (as evidenced by network RSA conducted on the ECoG dataset). It has also emphasized that critical aspects of semantic structure are reflected within modality-specific spokes, but only when the stimulus is from the corresponding modality. This has revealed a brain-wide picture of semantic representation that is consistent with the hub and the spokes both being relevant for encoding semantic structure, but not necessarily in the hub+spoke sense laid out in the introduction, which predicted that an interaction should have been observed between the hub and the spokes.

While this dissertation has not settled this representational question about the role of the ATL once and for all, it has provided additional evidence that semantic structure, even relatively fine grained structure, can be decoded from the ATL without considering the activity of the spokes. By demonstrating the presence of semantic structure in the ATL in the context of auditory stimuli in the fMRI experiments, this work also adds to the range of stimulus modalities that express

semantic similarity structure over this region.

Further work, and the refinement of the network RSA analysis procedure, will be required to further enhance our understanding of how, where, and when semantic structure is represented over cortex. The current evidence is consistent with an important role of the semantic hub which encodes a fully integrated cross modal and domain general semantic space.

Bibliography

Abel, T. J., Rhone, A. E., Nourski, K. V., Kawasaki, H., Oya, H., Griffiths, T. D., … Tranel, D. (2015). Direct physiologic evidence of a heteromodal convergence region for proper naming in human left anterior temporal lobe. *Journal of Neuroscience*, *35*(4), 1513–1520. doi:10.1523/jneurosci.3387-14.2015

Arthurs, O. J. & Boniface, S. (2002). How well do we understand the neural origins of the fMRI BOLD signal? *Trends in Neurosciences*, *25*(1), 27–31. doi:10.1016/s0166-2236(00)01995-0

Averbeck, B. B., Latham, P. E., & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, *7*(5), 358–366. doi:10.1038/nrn1888

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*(1), 617–645. doi:10.1146/annurev.psych.59.103006.093639

Bates, D. (2007). Linear mixed model implementation in lme4.

Bechara, A., Damasio, H., & Damasio, A. R. (2000). Emotion, decision making and the orbitofrontal cortex. *Cerebral cortex*, *10*(3), 295–307.

Behrmann, M. & Plaut, D. C. (2013). Distributed circuits, not circumscribed centers, mediate visual recognition. *Trends in Cognitive Sciences*, *17*(5), 210–219. doi:10.1016/j.tics.2013.03.007

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where is the semantic system? a critical review and meta-analysis of 120 Functional neuroimaging studies. *Cerebral Cortex*, *19*(12), 2767–2796. doi:10.1093/cercor/bhp055

Binder, J. R. & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. doi:10.1016/j.tics.2011.10.001

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Rao, S. M., & Cox, R. W. (1999, January). Conceptual processing during the conscious resting state. a functional MRI study. *Journal of Cognitive Neuroscience*, *11*(1), 80–95.

Binder, J. R., Gross, W. L., Allendorfer, J. B., Bonilha, L., Chapin, J., Edwards, J. C., … Weaver, K. E. (2011). Mapping anterior temporal lobe language areas with fMRI: a multicenter normative study. *NeuroImage*, *54*(2), 1465–1475. doi:10.1016/j.neuroimage.2010.09.048

Binder, J. R., Swanson, S. J., Hammeke, T. A., & Sabsevitz, D. S. (2008). A comparison of five fMRI protocols for mapping speech comprehension systems. *Epilepsia*, *49*(12), 1980–1997. doi:10.1111/j.1528-1167.2008.01683.x

Binney, R. J., Embleton, K. V., Jefferies, E., Parker, G. J. M., & Lambon Ralph, M. A. (2010). The ventral and inferolateral aspects of the anterior temporal lobe are crucial in semantic memory: evidence from a novel direct comparison of distortion-corrected fMRI, rTMS, and semantic dementia. *Cerebral Cortex*, *20*(11), 2728–2738. doi:10.1093/cercor/bhq019

Binney, R. J., Hoffman, P., & Ralph, M. A. L. (2016). Mapping the multiple graded contributions of the anterior temporal lobe representational hub to abstract and social concepts: evidence from distortion-corrected fMRI. *Cerebral Cortex*, *26*(11), 4227–4241. doi:10.1093/cercor/bhw260

Binney, R. J., Parker, G. J., & Lambon Ralph, M. A. (2012). Convergent connectivity and graded specialization in the rostral human temporal lobe as revealed by diffusion-weighted imaging probabilistic tractography. *Journal of cognitive neuroscience*, *24*(10), 1998–2014.

Bondell, H. D. & Reich, B. J. (2008). Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar. *Biometrics*, *64*(1), 115–123.

Boronat, C. B., Buxbaum, L. J., Coslett, H. B., Tang, K., Saffran, E. M., Kimberg, D. Y., & Detre, J. A. (2005). Distinctions between manipulation and function knowledge of objects: evidence from functional magnetic resonance imaging. *Cognitive Brain Research*, *23*(2-3), 361–373. doi:10.1016/j.cogbrainres.2004.11.001

Boulenger, V., Mechtouff, L., Thobois, S., Broussolle, E., Jeannerod, M., & Nazir, T. A. (2008). Word processing in parkinson's disease is impaired for action verbs but not for concrete nouns. *Neuropsychologia*, *46*(2), 743–756. doi:10.1016/j.neuropsychologia.2007.10.007

Bruffaerts, R., Dupont, P. [P.], Peeters, R., Deyne, S. D., Storms, G., & Vandenberghe, R. (2013). Similarity of fMRI activity patterns in left perirhinal cortex reflects semantic similarity between words. *Journal of Neuroscience*, *33*(47), 18597–18607. doi:10.1523/jneurosci.1548-13.2013

Bruffaerts, R., Dupont, P. [Patrick], Grauwe, S. D., Peeters, R., Deyne, S. D., Storms, G., & Vandenberghe, R. (2013). Right fusiform response patterns reflect visual object identity rather than semantic similarity. *NeuroImage*, *83*, 87–97. doi:10.1016/j.neuroimage.2013.05.128

Bulthé, J., Smedt, B. D., & de Beeck, H. O. (2014). Format-dependent representations of symbolic and non-symbolic numbers in the human cortex as revealed by multi-voxel pattern analyses. *NeuroImage*, *87*, 311–322. doi:10.1016/j.neuroimage.2013.10.049

Burgess, N. (2008). Spatial cognition and the brain. *Annals of the New York Academy of Sciences*, *1124*(1), 77–97. doi:10.1196/annals.1440.002

Buxbaum, L. J. (2001). Ideomotor apraxia: a call to action. *Neurocase*, *7*(6), 445–458. doi:10.1093/neucas/7.6.445

Carlson, T. A., Ritchie, J. B., Kriegeskorte, N., Durvasula, S., & Ma, J. (2014). Reaction time for object categorization is predicted by representational distance. *Journal of Cognitive Neuroscience*, *26*(1), 132–142. doi:10.1162/jocn_a_00476

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, *15*(5), 704–717. doi:10.1162/jocn.2003.15.5.704

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, *111*(40), 14565–14570. doi:10.1073/pnas.1402594111

Chen, G., Saad, Z. S., Britton, J. C., Pine, D. S., & Cox, R. W. (2013). Linear mixed-effects modeling approach to FMRI group analysis. *NeuroImage*, *73*, 176–190. doi:10.1016/j.neuroimage.2013.01.047

Chen, L. (2014, June 17). *White matter connectivity explains category-specific brain activation and impairment: a neurocomputational model of semantic cognition (unpublished doctoral dissertation)* (Doctoral dissertation, University of Wisconsin-Madison).

Chen, L., Lambon Ralph, M. A., & Rogers, T. T. (2016). Blank. *Nature*.

Chen, Y., Shimotake, A., Matsumoto, R., Kunieda, T., Kikuchi, T., Miyamoto, S., … Ralph, M. L. (2016). The 'when' and 'where' of semantic coding in the anterior temporal lobe: temporal representational similarity analysis of electrocorticogram data. *Cortex*, *79*, 1–13. doi:10.1016/j.cortex.2016.02.015

Clarke, A. & Tyler, L. K. (2014). Object-specific semantic coding in human perirhinal cortex. *Journal of Neuroscience*, *34*(14), 4766–4775. doi:10.1523/jneurosci.2828-13.2014

Connolly, A. C., Guntupalli, J. S., Gors, J., Hanke, M., Halchenko, Y. O., Wu, Y.-C., … Haxby, J. V. (2012). The representation of biological classes in the human brain. *Journal of Neuroscience*, *32*(8), 2608–2618. doi:10.1523/jneurosci.5547-11.2012

Cox, C. R., Seidenberg, M. S., & Rogers, T. T. (2015). Connecting functional brain imaging and parallel distributed processing. *Language, Cognition and Neuroscience*, *30*(4), 380–394. doi:10.1080/23273798.2014.994010

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, *29*(3), 162–173.

Cree, G. S. & McRae, K. (2003). Analyzing the factors underlying the structure and computation of the meaning of chipmunk, cherry, chisel, cheese, and cello (and many other such concrete nouns). *Journal of Experimental Psychology: General*, *132*(2), 163–201. doi:10.1037/0096-3445.132.2.163

Dale, A., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: i. segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194.

Damasio, A. R. (1989a). The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation*, *1*(1), 123–132. doi:10.1162/neco.1989.1.1.123

Damasio, A. R. (1989b). Time-locked multiregional retroactivation: a systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*(1-2), 25–62. doi:10.1016/0010-0277(89)90005-x

Damasio, A. R. (1994). Descartes' error: emotion, rationality and the human brain. *New York: Putnam*, *352*.

Damasio, H., Grabowski, T. J., Tranel, D., Hichwa, R. D., & Damasio, A. R. (1996). A neural basis for lexical retrieval. *Nature*, *380*(6574), 499–505. doi:10.1038/380499a0

Damasio, H., Tranel, D., Grabowski, T. J., Adolphs, R., & Damasio, A. R. (2004). Neural systems behind word and concept retrieval. *Cognition*, *92*(1-2), 179–229. doi:10.1016/j.cognition.2002.07.001

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391.

Dehaene, S. & Cohen, L. (1997). Cerebral pathways for calculation: double dissociation between rote verbal and quantitative knowledge of arithmetic. *Cortex*, *33*(2), 219–250. doi:10.1016/s0010-9452(08)70002-9

Desai, R. H., Binder, J. R., Conant, L. L., & Seidenberg, M. S. (2009). Activation of sensory-motor areas in sentence comprehension. *Cerebral Cortex*, *20*(2), 468–478. doi:10.1093/cercor/bhp115

Desai, R. H., Conant, L. L., Binder, J. R., Park, H., & Seidenberg, M. S. (2013). A piece of the action: modulation of sensory-motor regions by action idioms and metaphors. *NeuroImage*, *83*, 862–869. doi:10.1016/j.neuroimage.2013.07.044

Desikan, R. S., SÃ©gonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. doi:DOI:10.1016/j.neuroimage.2006.01.021

Destrieux, C., Fischl, B., Dale, A., & Halgren, E. (2010). Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, *53*(1), 1–15. doi:10.1016/j.neuroimage.2010.06.010

Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational similarity analysis reveals commonalities and differences in the semantic processing of words and objects. *Journal of Neuroscience*, *33*(48), 18906–18916. doi:10.1523/jneurosci.3809-13.2013

Deyne, S. D., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., & Storms, G. (2008). Exemplar by feature applicability matrices and other dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030–1048. doi:10.3758/brm.40.4.1030

Dilkina, K. & Lambon Ralph, M. A. (2012). Conceptual structure within and between modalities. *Frontiers in Human Neuroscience*, *6*. doi:10.3389/fnhum.2012.00333

Drevets, W. C., Price, J. L., Simpson, J. R., Todd, R. D., Reich, T., Vannier, M., & Raichle, M. E. (1997). Subgenual prefrontal cortex abnormalities in mood disorders.

Fairhall, S. L. & Caramazza, A. (2013). Brain regions that represent amodal conceptual knowledge. *Journal of Neuroscience*, *33*(25), 10552–10558. doi:10.1523/jneurosci.0051-13.2013

Farah, M. J. & McClelland, J. L. (1991). A computational model of semantic memory impairment: modality specificity and emergent category specificity. *Journal of Experimental Psychology: General*, *120*(4), 339–357. doi:10.1037/0096-3445.120.4.339

Feredoes, E., Tononi, G., & Postle, B. R. (2007). The neural bases of the short-term storage of verbal information are anatomically variable across individuals. *Journal of Neuroscience*, *27*(41), 11003–11008. doi:10.1523/jneurosci.1573-07.2007

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., … Seidenberg, M. S. (2016). Concept representation reflects multimodal abstraction: a framework for embodied semantics. *Cerebral Cortex*, *26*(5), 2018–2034. doi:10.1093/cercor/bhv020

Fernandino, L., Conant, L. L., Binder, J. R., Blindauer, K., Hiner, B., Spangler, K., & Desai, R. H. (2013a). Parkinson's disease disrupts both automatic and controlled processing of action verbs. *Brain and Language*, *127*(1), 65–74. doi:10.1016/j.bandl.2012.07.008

Fernandino, L., Conant, L. L., Binder, J. R., Blindauer, K., Hiner, B., Spangler, K., & Desai, R. H. (2013b). Where is the action? action sentence processing in parkinson's disease. *Neuropsychologia*, *51*(8), 1510–1517. doi:10.1016/j.neuropsychologia.2013.04.008

Figueiredo, M. A. T. & Nowak, R. D. [Robert D.]. (2014). Sparse estimation with strongly correlated variables using ordered weighted ℓ_1 regularization. *CoRR*, *abs/1409.4005*. Retrieved from http://arxiv.org/abs/1409.4005

Figueiredo, M. A. T. & Nowak, R. D. [Robert D]. (2016). Ordered weighted l1 regularized regression with strongly correlated covariates: theoretical aspects. In *Proceedings of the 19th international conference on artificial intelligence and statistics* (pp. 930–938).

Fischl, B. & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(20), 11050–11055. eprint: http://www.pnas.org/content/97/20/11050.full.pdf+html

Fischl, B., Liu, A., & Dale, A. M. (2001). Automated manifold surgery: constructing geometrically accurate and topologically correct models of the human cerebral cortex. *IEEE Medical Imaging*, *20*(1), 70–80.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., … Dale, A. M. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*, 341–355.

Fischl, B., Salat, D. H., van der Kouwe, A. J., Makris, N., SÃ©gonne, F., Quinn, B. T., & Dale, A. M. (2004). Sequence-independent segmentation of magnetic resonance images. *NeuroImage*, *23*(Supplement 1), S69–S84. Mathematics in Brain Imaging. doi:10.1016/j.neuroimage.2004.07.016

Fischl, B., Sereno, M. I., & Dale, A. (1999). Cortical surface-based analysis: ii: inflation, flattening, and a surface-based coordinate system. *NeuroImage*, *9*(2), 195–207. doi:10.1006/nimg.1998.0396

Fischl, B., Sereno, M. I., Tootell, R. B., & Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human Brain Mapping*, *8*(4), 272–284. doi:10.1002/(SICI)1097-0193(1999)8:4<272::AID-HBM10>3.0.CO;2-4

Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., SÃ©gonne, F., Salat, D. H., … Dale, A. M. (2004). Automatically Parcellating the Human Cerebral Cortex. *Cerebral Cortex*, *14*(1), 11–22. doi:10.1093/cercor/bhg087. eprint: http://cercor.oxfordjournals.org/content/14/1/11.full.pdf+html

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1). doi:10.18637/jss.v033.i01

Friston, K., Stephan, K., Lund, T., Morcom, A., & Kiebel, S. (2005). Mixed-effects and fMRI studies. *NeuroImage*, *24*(1), 244–252. doi:10.1016/j.neuroimage.2004.08.055

Gainotti, G. (2007). Face familiarity feelings, the right temporal lobe and the possible underlying neural mechanisms. *Brain Research Reviews*, *56*(1), 214–235. doi:10.1016/j.brainresrev.2007.07.009

Giordano, B. L., McAdams, S., Zatorre, R. J., Kriegeskorte, N., & Belin, P. (2013). Abstract encoding of auditory objects in cortical activity patterns. *Cerebral Cortex*, *23*(9), 2025–2037. doi:10.1093/cercor/bhs162

Gopnick, A. & Wellman, H. M. (1994). The theory theory. In L. A. Hirschfeld & S. A. Gelman (Eds.), *Mapping the mind: domain specificity in cognition and culture* (pp. 257–293). New York, NY: Cambridge University Press.

Grabowski, T. J., Damasio, H., Tranel, D., Ponto, L. L. B., Hichwa, R. D., & Damasio, A. R. (2001). A role for left temporal pole in the retrieval of words for unique entities. *Human Brain Mapping*, *13*(4), 199–212. doi:10.1002/hbm.1033.abs

Grill-Spector, K., Henson, R., & Martin, A. (2006). Repetition and the brain: neural models of stimulus-specific effects. *Trends in Cognitive Sciences*, *10*(1), 14–23. doi:10.1016/j.tics.2005.11.006

Guo, C. C., Gorno-Tempini, M. L., Gesierich, B., Henry, M., Trujillo, A., Shany-Ur, T., … Seeley, W. W. (2013). Anterior temporal lobe degeneration produces widespread network-driven dysfunction. *Brain*, *136*(10), 2979–2991. doi:10.1093/brain/awt222

Halai, A. D. [Ajay D.], Parkes, L. M., & Welbourne, S. R. (2015). Dual-echo fMRI can detect activations in inferior temporal lobe during intelligible speech comprehension. *NeuroImage*, *122*, 214–221. doi:10.1016/j.neuroimage.2015.05.067

Halai, A. D. [Ajay D], Welbourne, S. R., Embleton, K., & Parkes, L. M. (2014). A comparison of dual gradient-echo and spin-echo fMRI of the inferior temporal lobe. *Human Brain Mapping*, *35*(8), 4118–4128. doi:10.1002/hbm.22463

Han, X., Jovicich, J., Salat, D., van der Kouwe, A., Quinn, B., Czanner, S., … Fischl, B. (2006). Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, *32*(1), 180–194.

Harm, M. W. & Seidenberg, M. S. [Mark S.]. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review*, *106*(3), 491–528. doi:10.1037/0033-295X.106.3.491

Hassabis, D., Kumaran, D., & Maguire, E. A. (2007). Using imagination to understand the neural basis of episodic memory. *Journal of Neuroscience*, *27*(52), 14365–14374. doi:10.1523/JNEUROSCI.4549-07.2007. eprint: http://www.jneurosci.org/content/27/52/14365.full.pdf

Haufe, S., Meinecke, F., GÃ¶rgen, K., DÃ¤hne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. doi:10.1016/j.neuroimage.2013.10.067

Hauk, O. (2016). Only time will tell – why temporal information is essential for our neuroscientific understanding of semantics. *Psychonomic Bulletin & Review*, *23*(4), 1072–1079. doi:10.3758/s13423-015-0873-9

Hauk, O., Shtyrov, Y., & Pulvermüller, F. (2008). The time course of action and action-word comprehension in the human brain as revealed by neurophysiology. *Journal of Physiology-Paris*, *102*(1), 50–58.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annu. Rev. Neurosci. 37*(1), 435–456. doi:10.1146/annurev-neuro-062012-170325

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430. doi:10.1126/science.1063736. eprint: http://science.sciencemag.org/content/293/5539/2425.full.pdf

Henson, R. & Rugg, M. (2003). Neural response suppression, haemodynamic repetition effects, and behavioural priming. *Neuropsychologia*, *41*(3), 263–270. doi:10.1016/s0028-3932(02)00159-8

Hermann, B. P., Seidenberg, M., Haltiner, A., & Wyler, A. R. (1995). Relationship of age at onset, chronologic age, and adequacy of preoperative performance to verbal memory change after anterior temporal lobectomy. *Epilepsia*, *36*(2), 137–145. doi:10.1111/j.1528-1157.1995.tb00972.x

Hermann, B. P., Wyler, A. R., Somes, G., Dohan, F. C., Berry III, A. D., & Clemenet, L. (1994). Declarative memory following anterior temporal lobectomy in humans. *Behavioral Neuroscience*, *108*(1), 3–10. doi:10.1037/0735-7044.108.1.3

Hinton, G. (2014). Where do features come from? *Cognitive Science*, *38*(6), 1078–1101. doi:10.1111/cogs.12049

Hodges, J. R. (1995). Semantic dementia: progressive fluent aphasia with temporal lobe atrophy. *Neurocase*, *1*(1), 39g–54. doi:10.1093/neucas/1.1.39-g

Hodges, J. R. & Patterson, K. (2007). Semantic dementia: a unique clinicopathological syndrome. *The Lancet Neurology*, *6*(11), 1004–1014. doi:10.1016/s1474-4422(07)70266-1

Hoerl, A. E. & Kennard, R. W. (2000). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, *42*(1), 80–86. doi:10.1080/00401706.2000.10485983

Hoffman, P. A., Binney, R. J., & Ralph, M. A. L. (2015). Differing contributions of inferior prefrontal and anterior temporal cortex to concrete and abstract conceptual knowledge. *Cortex*, *63*, 250–266. doi:10.1016/j.cortex.2014.09.001

Hoffman, P. A., Jones, R. W., & Ralph, M. A. L. (2012). The degraded concept representation system in semantic dementia: damage to pan-modal hub, then visual spoke. *Brain*, *135*(12), 3770–3780. doi:10.1093/brain/aws282

Hoffman, P. A. & Lambon Ralph, M. A. (2013). Shapes, scents and sounds: quantifying the full multi-sensory basis of conceptual knowledge. *Neuropsychologia*, *51*(1), 14–25. doi:10.1016/j.neuropsychologia.2012.11.009

Horner, A. J. & Henson, R. N. (2012). Incongruent abstract stimulus–response bindings result in response interference: fMRI and EEG evidence from visual object classification priming. *Journal of Cognitive Neuroscience*, *24*(3), 760–773. doi:10.1162/jocn_a_00163

Hubel, D. H. & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, *148*(3), 574–591. doi:10.1113/jphysiol.1959.sp006308

Hubel, D. H. & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, *195*(1), 215–243. doi:10.1113/jphysiol.1968.sp008455

Humphreys, G. W. & Forde, E. M. (2001). Hierarchies, similarity, and interactivity in object recognition: "category-specific" neuropsychological deficits. *Behavioral and Brain Sciences*, (24), 453–476.

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. doi:10.1038/nature17637

Huth, A., Nishimoto, S., Vu, A., & Gallant, J. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224. doi:10.1016/j.neuron.2012.10.014

Ishibashi, R., Lambon Ralph, M. A., Saito, S., & Pobric, G. (2011). Different roles of lateral anterior temporal lobe and inferior parietal lobule in coding function and manipulation tool knowledge: evidence from an rTMS study. *Neuropsychologia*, *49*(5), 1128–1135. doi:10.1016/j.neuropsychologia.2011.01.004

Jackson, R. L., Hoffman, P., Pobric, G., & Ralph, M. A. L. (2016). The semantic network at work and rest: differential connectivity of anterior temporal lobe sub-

regions. *Journal of Neuroscience*, *36*(5), 1490–1501. doi:10.1523/jneurosci.2999-15.2016

Jackson, R. L., Lambon Ralph, M. A., & Pobric, G. (2015). The timing of anterior temporal lobe involvement in semantic processing. *Journal of Cognitive Neuroscience*, *27*(7), 1388–1396. doi:10.1162/jocn_a_00788

Jamieson, K. G., Jain, L., Fernandez, C., Glattard, N. J., & Nowak, R. (2015). Next: a system for real-world development, evaluation, and application of active learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28* (pp. 2656–2664). Curran Associates, Inc. Retrieved from http://papers.nips.cc/paper/5868-next-a-system-for-real-world-development-evaluation-and-application-of-active-learning.pdf

Jimura, K. & Poldrack, R. A. (2012). Analyses of regional-average activation and multivoxel pattern information tell complementary stories. *Neuropsychologia*, *50*(4), 544–552. doi:10.1016/j.neuropsychologia.2011.11.007

Joanisse, M. F. & Seidenberg, M. S. [M. S.]. (1999). Impairments in verb morphology after brain injury: a connectionist model. *Proceedings of the National Academy of Sciences*, *96*(13), 7592–7597. doi:10.1073/pnas.96.13.7592

Johnson, M. R., Mitchell, K. J., Raye, C. L., D'Esposito, M., & Johnson, M. K. (2007). A brief thought can modulate activity in extrastriate visual areas: top-down effects of refreshing just-seen visual stimuli. *NeuroImage*, *37*(1), 290–299. doi:http://dx.doi.org/10.1016/j.neuroimage.2007.05.017

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., … Dale, A. (2006). Reliability in multi-site structural mri studies: effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, *30*(2), 436–443. doi:DOI:10.1016/j.neuroimage.2005.09.046

Jozwik, K. M., Kriegeskorte, N., & Mur, M. (2016). Visual features as stepping stones toward semantics: explaining object similarity in IT and perception with non-negative least squares. *Neuropsychologia*, *83*, 201–226. doi:10.1016/j.neuropsychologia.2015.10.023

Just, M. A., Cherkassky, V. L., Aryal, S., & Mitchell, T. M. (2010). A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PLoS ONE*, *5*(1), e8622. doi:10.1371/journal.pone.0008622

Kenward, M. G. & Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics*, *53*(3), 983. doi:10.2307/2533558

Kiefer, M. & Pulvermüller, F. (2012). Conceptual representations in mind and brain: theoretical developments, current evidence and future directions. *Cortex*, *48*(7), 805–825. doi:10.1016/j.cortex.2011.04.006

Kriegeskorte, N. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*. doi:10.3389/neuro.06.004.2008

Kriegeskorte, N. (2009). Relating population-code representations between man, monkey, and computational models. *Front. Neurosci. 3*(3), 363–373. doi:10.3389/neuro.01.035.2009

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*(10), 3863–3868. doi:10.1073/pnas.0600244103

Kriegeskorte, N. & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. doi:10.1016/j.tics.2013.06.007

Kriegeskorte, N. & Mur, M. (2012). Inverse MDS: inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology*, *3*. doi:10.3389/fpsyg.2012.00245

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., … Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141. doi:10.1016/j.neuron.2008.10.043

Lambon Ralph, M. A., Cipolotti, L., Manes, F., & Patterson, K. (2010). Taking both sides: do unilateral anterior temporal lobe lesions disrupt semantic memory? *Brain*, *133*(11), 3243–3255. doi:10.1093/brain/awq264

Lambon Ralph, M. A., Lowe, C., & Rogers, T. T. (2007). Neural basis of category-specific semantic deficits for living things: evidence from semantic dementia, HSVE and a neural network model. *Brain*, *130*(4), 1127–1137. doi:10.1093/brain/awm025

Landauer, T. K. & Dumais, S. T. (1997). A solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240. doi:10.1037/0033-295x.104.2.211

Lee, M., Pincombe, B., & Welsh, M. (2005). An empirical evaluation of models of text document similarity. In *Proceedings of the cognitive science society*. Cognitive Science Society.

Liuzzi, A. G., Bruffaerts, R., Dupont, P., Adamczuk, K., Peeters, R., Deyne, S. D., … Vandenberghe, R. (2015). Left perirhinal cortex codes for similarity in meaning between written words: comparison with auditory word input. *Neuropsychologia*, *76*, 4–16. doi:10.1016/j.neuropsychologia.2015.03.016

Logothetis, N. K. & Sheinberg, D. L. (1996). Visual object recognition. *Annu. Rev. Neurosci. 19*(1), 577–621. doi:10.1146/annurev.ne.19.030196.003045

Logothetis, N. K. & Wandell, B. A. (2004). Interpreting the BOLD signal. *Annual Review of Physiology*, *66*(1), 735–769. doi:10.1146/annurev.physiol.66.082602.092845

Marr, D. (1982). *Vision: a computational investigation into the human representation and processing of visual information*. New York, NY.: Henry Holt and Co. Inc.

Martin, A. (2007). The representation of object concepts in the brain. *Annual Review of Psychology*, *58*(1), 25–45. doi:10.1146/annurev.psych.57.102904.190143

Martin, A. & Chao, L. L. (2001). Semantic memory and the brain: structure and processes. *Current Opinion in Neurobiology*, *11*(2), 194–201. doi:10.1016/s0959-4388(00)00196-3

Mayberg, H. S., Liotti, M., Brannan, S. K., McGinnis, S., Mahurin, R. K., Jerabek, P. A., …, Lancaster, J. L., et al. (1999). Reciprocal limbic-cortical function and negative mood: converging pet findings in depression and normal sadness. *American Journal of Psychiatry*.

Mayberry, E. J., Sage, K., & Lambon Ralph, M. A. (2011). At the edge of semantic space: the breakdown of coherent concepts in semantic dementia is constrained by typicality and severity but not modality. *Journal of Cognitive Neuroscience*, *23*(9), 2240–2251. doi:10.1162/jocn.2010.21582

McClelland, J. L. & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*(2), 159–188. doi:10.1037/0096-3445.114.2.159

Mesulam, M.-M. (1990). Large-scale neurocognitive networks and distributed processing for attention, language, and memory. *Annals of Neurology*, *28*(5), 597–613. doi:10.1002/ana.410280502

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science*, *320*(5880), 1191–1195. doi:10.1126/science.1152876

Morán, M. A., Mufson, E. J., & Mesulam, M.-M. (1987). Neural inputs into the temporopolar cortex of the rhesus monkey. *Journal of Comparative Neurology*, *256*(1), 88–103. doi:10.1002/cne.902560108

Mur, M., Meys, M., Bodurka, J., Goebel, R., Bandettini, P. A., & Kriegeskorte, N. (2013). Human object-similarity judgments reflect and transcend the primate-it object representation. *Frontiers in psychology*, *4*, 128.

Navarro, D. J. & Lee, M. D. (2002). Commonalities and distinctions in featural stimulus representations. In *Proceedings of the 24th annual conference of the cognitive science society* (Vol. 24, pp. 685–690).

Nestor, P. J., Fryer, T. D., & Hodges, J. R. (2006). Declarative memory impairments in alzheimer's disease and semantic dementia. *NeuroImage*, *30*(3), 1010–1020. doi:10.1016/j.neuroimage.2005.10.008

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. doi:10.1016/j.tics.2006.07.005

Oswal, U., Cox, C. R., Lambon Ralph, M. A., Rogers, T. T., & Nowak, R. (2016). Representational similarity learning with application to brain networks. In *Proceedings of the 33rd international conference on machine learning* (pp. 1041–1049).

Patterson, K. (1995). Deterioration of word meaning: implications for reading. *Neurocase*, *1*(2), 167e–172. doi:10.1093/neucas/1.2.167-e

Patterson, K. & Lambon Ralph, M. A. (2016). The hub-and-spoke hypothesis of semantic memory. In *Neurobiology of language* (pp. 765–775). Elsevier BV. doi:10.1016/b978-0-12-407794-2.00061-4

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? the representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976–987. doi:10.1038/nrn2277

Peelen, M. V. & Caramazza, A. (2012). Conceptual object representations in human anterior temporal cortex. *Journal of Neuroscience*, *32*(45), 15728–15736. doi:10.1523/jneurosci.1953-12.2012

Pereira, F. & Botvinick, M. (2011). Information mapping with pattern classifiers: a comparative study. *NeuroImage*, *56*(2), 476–496. doi:10.1016/j.neuroimage.2010.05.026

Pereira, F., Mitchell, T. M., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage*, *45*(1), 199–209. doi:10.1016/j.neuroimage.2008.11.007

Phillips, M. L., Drevets, W. C., Rauch, S. L., & Lane, R. (2003). Neurobiology of emotion perception ii: implications for major psychiatric disorders. *Biological psychiatry*, *54*(5), 515–528.

Piazza, M., Pinel, P., Bihan, D. L., & Dehaene, S. (2007). A magnitude code common to numerosities and number symbols in human intraparietal cortex. *Neuron*, *53*(2), 293–305. doi:10.1016/j.neuron.2006.11.022

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J., & Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, *454*(7207), 995–999. doi:10.1038/nature07140

Pinker, S. (1991). Rules of language. *Science*, *253*(5019), 530–535. doi:10.1126/science.1857983

Plaut, D. C. (2002). Graded modality-specific specialisation in semantics: a computational account of optic aphasia. *Cognitive Neuropsychology*, *19*(7), 603–639. doi:10.1080/02643290244000112

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56–115. doi:10.1037/0033-295x.103.1.56

Plaut, D. C. & Shallice, T. (1993). Deep dyslexia: a case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*(5), 377–500. doi:10.1080/02643299308253469

Pluim, J., Maintz, J., & Viergever, M. (2003). Mutual-information-based registration of medical images: a survey. *IEEE Transactions on Medical Imaging*, *22*(8), 986–1004. doi:10.1109/tmi.2003.815867

Pobric, G., Jefferies, E., & Lambon Ralph, M. A. (2007). Anterior temporal lobes mediate semantic representation: mimicking semantic dementia by using rtms in normal participants. *Proceedings of the National Academy of Sciences*, *104*(50), 20137–20141.

Pobric, G., Jefferies, E., & Lambon Ralph, M. A. (2010a). Amodal semantic representations depend on both anterior temporal lobes: evidence from repetitive transcranial magnetic stimulation. *Neuropsychologia*, *48*(5), 1336–1342. doi:10.1016/j.neuropsychologia.2009.12.036

Pobric, G., Jefferies, E., & Lambon Ralph, M. A. (2010b). Category-specific versus category-general semantic impairment induced by transcranial magnetic stimulation. *Current Biology*, *20*(10), 964–968. doi:10.1016/j.cub.2010.03.070

Pobric, G., Lambon Ralph, M. A., & Zahn, R. (2016). Hemispheric specialization within the superior anterior temporal cortex for social and nonsocial concepts. *Journal of Cognitive Neuroscience*, *28*(3), 351–360. doi:10.1162/jocn_a_00902

Poldrack, R. A., Halchenko, Y. O., & Hanson, S. J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science*, *20*(11), 1364–1372. doi:10.1111/j.1467-9280.2009.02460.x

Pordes, R., Petravick, D., Kramer, B., Olson, D., Livny, M., Roy, A., … Quick, R. (2007). The open science grid. *Journal of Physics: Conference Series*, *78*, 012057. doi:10.1088/1742-6596/78/1/012057

Pouget, A., Dayan, P., & Zemel, R. (2000, November). Information processing with population codes. *Nat Rev Neurosci*, *1*(2), 125–132. doi:10.1038/35039062

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(7), 576–582.

Pulvermüller, F., Cooper-Pye, E., Dine, C., Hauk, O., Nestor, P. J., & Patterson, K. (2010). The word processing deficit in semantic dementia: all categories are equal, but some categories are more equal than others. *Journal of Cognitive Neuroscience*, *22*(9), 2027–2041. doi:10.1162/jocn.2009.21339

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005). Invariant visual representation by single neurons in the human brain. *Nature*, *435*(7045), 1102–1107. doi:10.1038/nature03687

Rao, N. S., Cox, C. R., Nowak, R., & Rogers, T. T. (2013). Sparse overlapping sets lasso for multitask learning and its application to fmri analysis. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, & K. Weinberger (Eds.), *Advances in neural information processing systems 26* (pp. 2202–2210). Curran Associates, Inc.

Reuter, M., Rosas, H. D., & Fischl, B. (2010). Highly accurate inverse consistent registration: a robust approach. *NeuroImage*, *53*(4), 1181–1196. doi:10.1016/j.neuroimage.2010.07.020

Reuter, M., Schmansky, N. J., Rosas, H. D., & Fischl, B. (2012). Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, *61*(4), 1402–1418. doi:10.1016/j.neuroimage.2012.02.084

Rice, G. E., Hoffman, P., & Lambon Ralph, M. A. (2015). Graded specialization within and between the anterior temporal lobes. *Annals of the New York Academy of Sciences*, *1359*(1), 84–97. doi:10.1111/nyas.12951

Riggall, A. C. & Postle, B. R. (2012). The relationship between working memory storage and elevated activity as measured with functional magnetic resonance imaging. *Journal of Neuroscience*, *32*(38), 12990–12998. doi:10.1523/jneurosci.1892-12.2012

Rish, I., Cecchi, G. A., Heuton, K., Baliki, M. N., & Apkarian, A. V. (2012). Sparse regression analysis of task-relevant information distribution in the brain. In D. R. Haynor & S. Ourselin (Eds.), *Medical imaging 2012: image processing*. SPIE-Intl Soc Optical Eng. doi:10.1117/12.911318

Rissman, J., Eliassen, J. C., & Blumstein, S. E. (2003). An event-related fMRI investigation of implicit semantic priming. *Journal of Cognitive Neuroscience*, *15*(8), 1160–1175. doi:10.1162/089892903322598120

Rogers, T. T., Hocking, J., Noppeney, U., Mechelli, A., Gorno-Tempini, M. L., Patterson, K., & Price, C. J. (2006). Anterior temporal cortex and semantic memory: reconciling findings from neuropsychology and functional imaging. *Cognitive, Affective, & Behavioral Neuroscience*, *6*(3), 201–213. doi:10.3758/cabn.6.3.201

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). Structure and deterioration of semantic memory: a neuropsychological and computational investigation. *Psychological Review*, *111*(1), 205–235. doi:10.1037/0033-295x.111.1.205

Rogers, T. T. & McClelland, J. L. (2014). Parallel distributed processing at 25: further explorations in the microstructure of cognition. *Cognitive Science*, *38*(6), 1024–1077. doi:10.1111/cogs.12148

Rohde, D. L. T. (1999). *LENS: the light, efficient network simulator*. Carnegie Mellon University.

Rosenberg, S. & Park Kim, M. (1975). The method of sorting as a data-gathering procedure in multivariate research. *Multivariate Behavioral Research*, *10*(4), 489–502.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group. (1986). *Parallel distributed processing: explorations in the microstructure of cognition*. MIT Press.

Sanjuán, A., Hope, T. M., Jones, P̃., Prejawa, S., Oberhuber, M., Guerin, J., … Price, C. J. (2015). Dissociating the semantic function of two neighbouring subre-

gions in the left lateral anterior temporal lobe. *Neuropsychologia*, *76*, 153–162. doi:10.1016/j.neuropsychologia.2014.12.004

Schapiro, A. C., McClelland, J. L., Welbourne, S. R., Rogers, T. T., & Lambon Ralph, M. A. (2013). Why bilateral damage is worse than unilateral damage to the brain. *Journal of Cognitive Neuroscience*, *25*(12), 2107–2123. doi:10.1162/jocn_a_00441

Schnur, T. T., Schwartz, M. F., Kimberg, D. Y., Hirshorn, E., Coslett, H. B., & Thompson-Schill, S. L. (2008). Localizing interference during naming: convergent neuroimaging and neuropsychological evidence for the function of broca's area. *Proceedings of the National Academy of Sciences*, *106*(1), 322–327. doi:10.1073/pnas.0805874106

Segonne, F., Dale, A. M., Busa, E., Glessner, M., Salat, D., Hahn, H. K., & Fischl, B. (2004). A hybrid approach to the skull stripping problem in mri. *NeuroImage*, *22*(3), 1060–1075. doi:DOI:10.1016/j.neuroimage.2004.03.032

Segonne, F., Pacheco, J., & Fischl, B. (2007). Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Trans Med Imaging*, *26*, 518–529.

Seidenberg, M. S. [Mark S.] & Plaut, D. C. (2014). Quasiregularity and its discontents: the legacy of the past tense debate. *Cognitive Science*, *38*(6), 1190–1228. doi:10.1111/cogs.12147

Sfiligoi, I., Bradley, D. C., Holzman, B., Mhashilkar, P., Padhi, S., & Wurthwein, F. (2009). The pilot way to grid resources using glideinWMS. In *2009 WRI world congress on computer science and information engineering*. Institute of Electrical and Electronics Engineers (IEEE). doi:10.1109/csie.2009.950

Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (1987). Emotion knowledge: further exploration of a prototype approach. *Journal of Personality and Social Psychology*, *52*(6), 1061–1086. doi:10.1037/0022-3514.52.6.1061

Shimotake, A., Matsumoto, R., Ueno, T., Kunieda, T., Saito, S., Hoffman, P., … Lambon Ralph, M. A. (2014). Direct exploration of the role of the ventral anterior temporal lobe in semantic memory: cortical stimulation and local field potential evidence from subdural grid electrodes. *Cereb Cortex*, *25*(10), 3802–3817. doi:10.1093/cercor/bhu262

Shinkareva, S. V., Mason, R. A., Malave, V. L., Wang, W., Mitchell, T. M., & Just, M. A. (2008). Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS ONE*, *3*(1), 1394. doi:10.1371/journal.pone.0001394

Simmons, W. K., Ramjee, V., Beauchamp, M. S., McRae, K., Martin, A., & Barsalou, L. W. (2007). A common neural substrate for perceiving and knowing about color. *Neuropsychologia*, *45*(12), 2802–2810.

Simmons, W. K., Reddish, M., Bellgowan, P. S. F., & Martin, A. (2009). The selectivity and functional connectivity of the anterior temporal lobes. *Cerebral Cortex*, *20*(4), 813–825. doi:10.1093/cercor/bhp149

Simon, N., Friedman, J., Hastie, T., & Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*.

Singer, W. & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annu. Rev. Neurosci. 18*(1), 555–586. doi:10.1146/annurev.ne.18.030195.003011

Sled, J., Zijdenbos, A., & Evans, A. (1998). A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Trans Med Imaging*, *17*, 87–97.

Small, D., Gitelman, D., Gregory, M., Nobre, A., Parrish, T., & Mesulam, M.-M. (2003). The posterior cingulate and medial prefrontal cortex mediate the anticipatory allocation of spatial attention. *NeuroImage*, *18*(3), 633–641. doi:http://dx.doi.org/10.1016/S1053-8119(02)00012-5

Small, S. L. & Nusbaum, H. C. (2004). On the neurobiological investigation of language understanding in context. *Brain and Language*, *89*(2), 300–311. doi:10.1016/s0093-934x(03)00344-4

Smolensky, P. (1986). Neural and conceptual interpretation of parallel distributed processing models. In J. L. McClelland, D. E. Rumelhart, & C. PDP Research Group (Eds.), *Parallel distributed processing* (pp. 390–431). Cambridge, MA, USA: MIT Press. Retrieved from http://dl.acm.org/citation.cfm?id=104339.104348

Snodgrass, J. G. & Vanderwart, M. (1980). A standardized set of 260 pictures: norms for name agreement, image agreement, familiarity, and visual complexity. *Journal of Experimental Psychology: Human Learning & Memory*, *6*(2), 174–215. doi:10.1037/0278-7393.6.2.174

Snowden, J. S., Goulding, P., & Neary, D. (1989). Semantic dementia: a form of circumscribed cerebral atrophy. *Behavioural Neurology*.

Tahmasebi, A. M., Davis, M. H., Wild, C. J., Rodd, J. M., Hakyemez, H., Abolmaesumi, P., & Johnsrude, I. S. (2012). Is the link between anatomical structure

and function equally strong at all cognitive levels of processing? *Cerebral Cortex*, *22*(7), 1593–1603. doi:10.1093/cercor/bhr205

Tanaka, K., Saito, H., Fukada, Y., & Moriya, M. (1991, July). Coding visual images of objects in the inferotemporal cortex of the macaque monkey. *Journal of Neurophysiology*, *66*(1), 170–189.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318. doi:10.1016/j.tics.2006.05.009

Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: the condor experience. *Concurrency - Practice and Experience*, *17*(2-4), 323–356.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, *58*(1), 267–288. Retrieved from http://www.jstor.org/stable/2346178

Tranel, D., Damasio, H., & Damasio, A. R. (1997). A neural basis for the retrieval of conceptual knowledge. *Neuropsychologia*, *35*(10), 1319–1327. doi:10.1016/S0028-3932(97)00085-7

van Casteren, M. & Davis, M. H. (2006). Mix, a program for pseudorandomization. *Behavior Research Methods*, *38*(4), 584–589. doi:10.3758/bf03193889

Visser, M., Embleton, K., Jefferies, E., Parker, G., & Lambon Ralph, M. A. (2010). The inferior, anterior temporal lobes and semantic memory clarified: novel evidence from distortion-corrected fMRI. *Neuropsychologia*, *48*(6), 1689–1696. doi:10.1016/j.neuropsychologia.2010.02.016

Visser, M., Jefferies, E., & Lambon Ralph, M. A. (2010). Semantic processing in the anterior temporal lobes: a meta-analysis of the functional neuroimaging literature. *Journal of Cognitive Neuroscience*, *22*(6), 1083–1094. doi:10.1162/jocn.2009.21309

Visser, M., Jefferies, E., Embleton, K. V., & Ralph, M. A. L. (2012). Both the middle temporal gyrus and the ventral anterior temporal area are crucial for multimodal semantic processing: distortion-corrected fMRI evidence for a double gradient of information convergence in the temporal lobes. *Journal of Cognitive Neuroscience*, *24*(8), 1766–1778. doi:10.1162/jocn_a_00244

Visser, M. & Ralph, M. A. L. (2011). Differential contributions of bilateral ventral anterior temporal lobe and left anterior superior temporal gyrus to semantic processes. *Journal of Cognitive Neuroscience*, *23*(10), 3121–3131. doi:10.1162/jocn_a_00007

Wang, X., Hutchinson, R., & Mitchell, T. M. (2004). Training fmri classifiers to detect cognitive states across multiple human subjects. In S. Thrun, L. K. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems 16* (pp. 709–716). MIT Press. Retrieved from http : / / papers . nips . cc / paper / 2449 - training-fmri-classifiers-to-detect-cognitive-states-across-multiple-human-subjects.pdf

Warrington, E. K. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology*, *27*(4), 635–657. doi:10.1080/14640747508400525

Wheatley, T., Weisberg, J., Beauchamp, M. S., & Martin, A. (2005). Automatic priming of semantically related words reduces activity in the fusiform gyrus. *Journal of Cognitive Neuroscience*, *17*(12), 1871–1885. doi:10.1162/089892905775008689

Yuan, M. & Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *68*(1), 49–67. doi:10.1111/j.1467-9868.2005.00532.x

Zeng, X. & Figueiredo, M. A. T. (2014). The ordered weighted ℓ_1 norm: atomic formulation, dual norm, and projections. *CoRR*, *abs/1409.4271*. Retrieved from http://arxiv.org/abs/1409.4271

Zhong, L. W. & Kwok, J. T. (2012). Efficient sparse modeling with automatic feature grouping. *IEEE Transactions on Neural Networks and Learning Systems*, *23*(9), 1436–1447. doi:10.1109/tnnls.2012.2200262

Zipser, D. & Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*(6158), 679–684. doi:10.1038/331679a0

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, *67*, 301–320.

Appendix A

Network RSA selection maps for individual subjects

Network RSA is a regularized regression technique that obtains sparse models in individual subjects. In Chapters 5 and 6, these solutions are aggregated over subjects in an attempt to make more general claims about the contributions of various brain regions in in the sample and, ideally, in the population. In those chapters, we confronted some of the inherent challenges to this endeavor. Statistical thresholding presented a challenge because there is currently no good way to characterize the distribution from which the weights that comprise the models obtained with network RSA are sampled from, neither within nor across subjects. We attempted to address this by empirically simulating these distributions by performing analyses of permuted data. However, the sets of voxels selected by the permutation procedure were anything but random, and in fact were far more likely in precisely the areas one would expect true signal to be a priori, which saps a great deal of statistical power.

In light of the complications faced by attempting to aggregate over subjects in a principled way, I would like to present the solution maps for each individual subject each of the semantic models trained on visual and audio trials, respectively. In Figures A.1 and A.2, I show stability maps for each subject: voxels that are more yellow are selected more often over cross-validations (i.e., variations on the training set). These solutions are visualized on the common Talairach `TT_N27` template, so there is some degree of spatial smoothing due to interpolation, but no additional smoothing was applied.

In these figures, it is clear that group level aggregates may not be doing justice to the wide and idiosyncratic voxel selections observed in each subject.

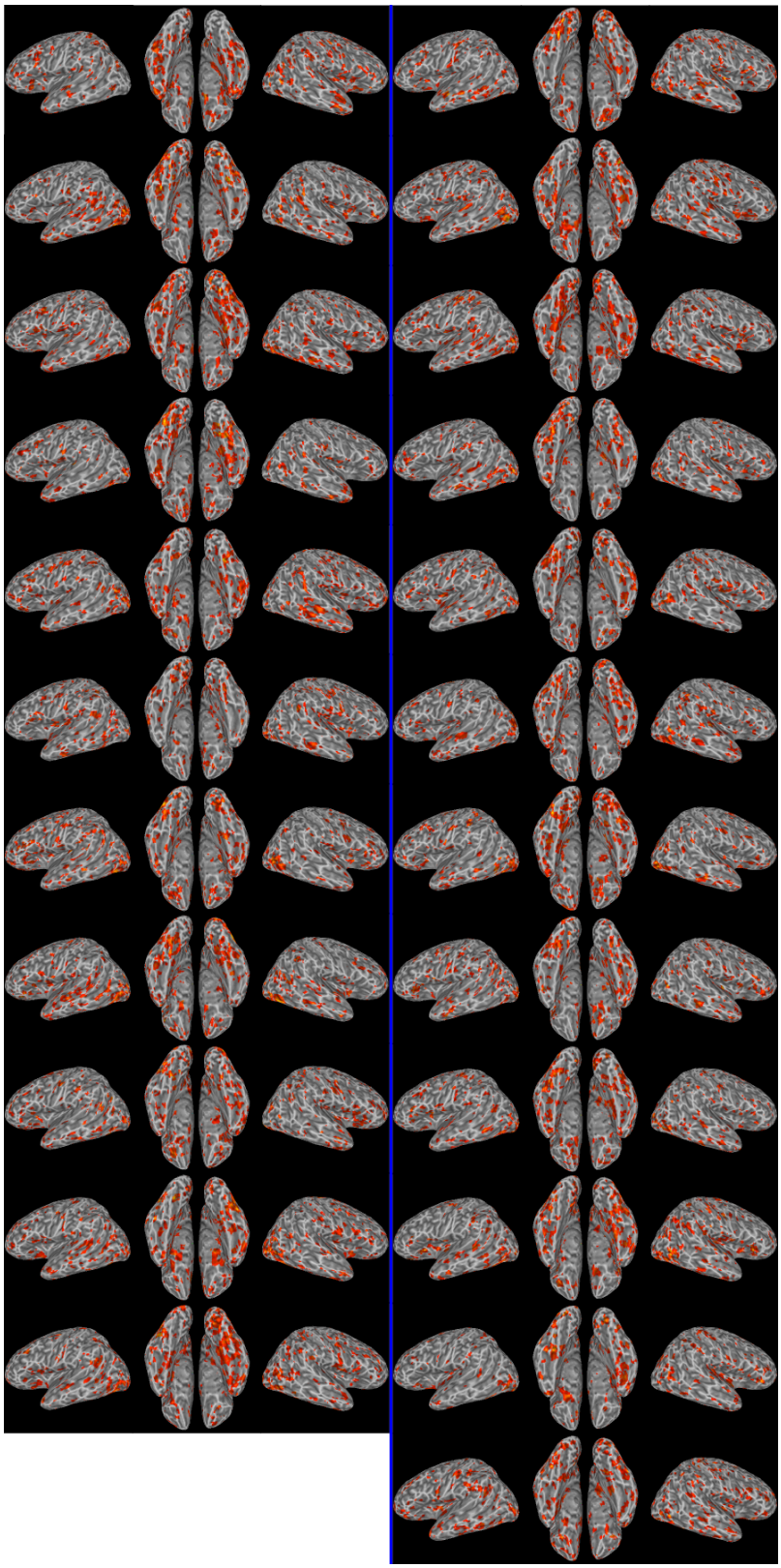

Figure A.1: Network RSA solution maps when modeling the target semantic structure described in Chapter 5 on visual trials, where stimuli were presented as line drawings.

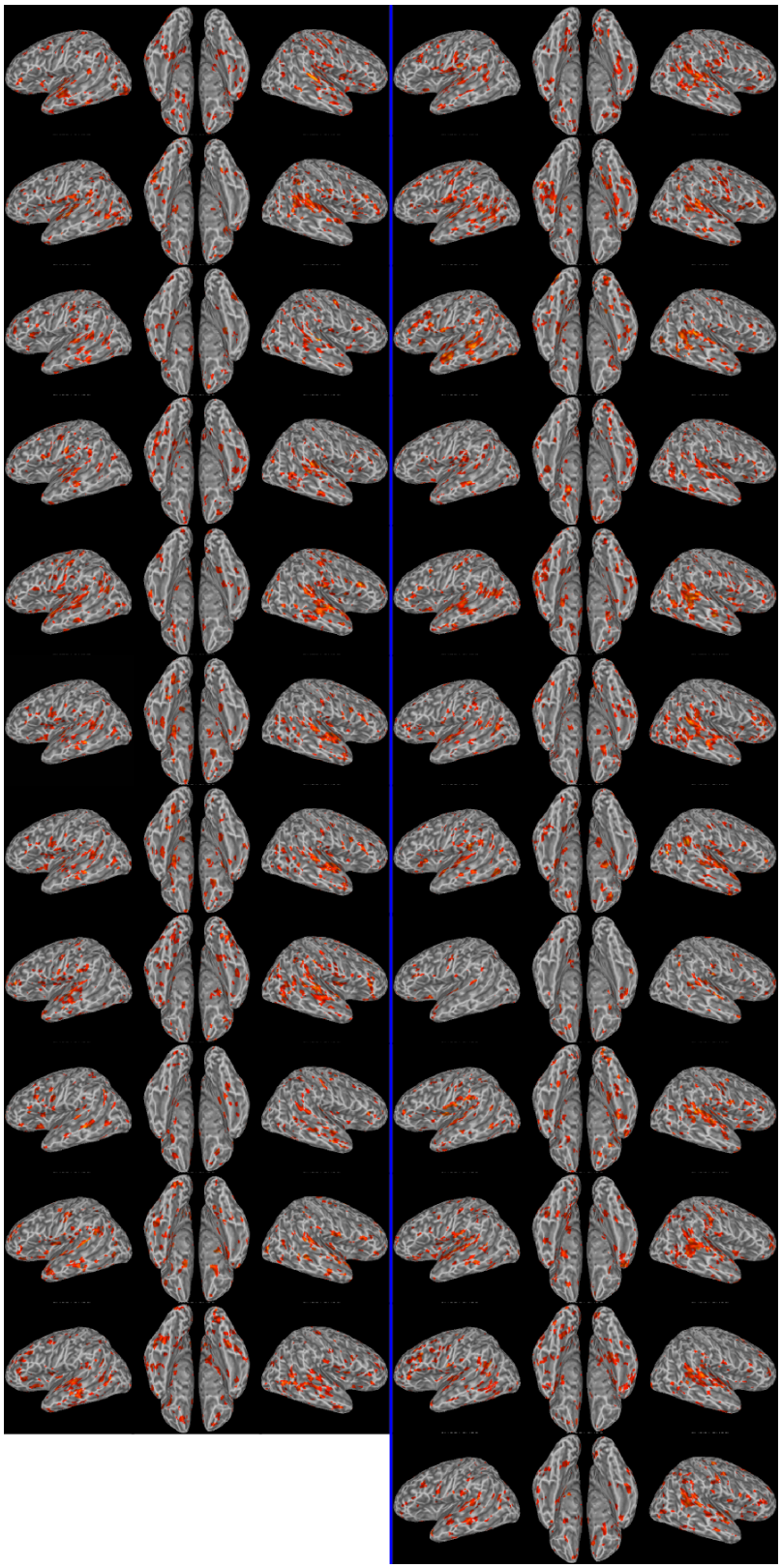

Figure A.2: Network RSA solution maps when modeling the target semantic structure described in Chapter 5 on audio trials, where stimuli were presented as characteristic sounds.

List of Figures

1.1 Schematic depiction of the hub-and-spoke model of semantic cognition. Reproduced from Lambon Ralph, Lowe, and Rogers (2007). 6

1.2 An example of how the ATL might encode only pan-modal structure, as under the *hub+spoke* perspective. The data points were generated by a simple linear model of 4 input features, 2 in each spoke (audio and visual) and their interactions (1 local interaction per spoke, and 4 global interactions reflected in the anterior temporal lobe). Given the appropriate weights, when taken all together, the set of features and their interactions express a *global* similarity structure. However, if the features are spread over the two spokes, no one region contains the full structure. Rather, each expresses *local* similarity structure. The insight is that aggregating the local "representations" across regions is mathematically equivalent to generating a single global representation directly (at least in the case of simple linear models). 10

2.1 Each row corresponds to a searchlight that contains a set of 25 voxels. These voxels are the same across subjects, but different across the two searchlights. Searchlight 1 exemplifies a case where a searchlight MVPA will succeed but a univariate analysis, employing blurring within and averaging across subjects, will fail. Searchlight 2 exemplifies a case where a searchlight multivoxel pattern analysis (MVPA) is likely to fail but a univariate analysis will succeed. 29

3.1 A PDP model used to understand semantic memory (from Rogers et al., 2004). Units in the *Visual* layer code visual features and units in the *Verbal* layer encode familiar words. The *Visual* and *Verbal* units can receive direct inputs from the environment, corresponding to direct perception of a visually-presented item or of a spoken statement. Units in both layers send connections to, and receive connections from, an intermediating hidden layer. To simulate a task such as object naming, visual features of the object are directly activated in the *Visual* layer and the activation propagates to other units via the weighted connections. If the weights are set to appropriate values, the model will ultimately activate the *Verbal* unit corresponding to the item name. Likewise name comprehension is simulated by directly activating the unit corresponding to the name and propagating activation throughout the network. With appropriate weights the visual features of the named item will activate, along with verbal units describing the item's properties. Appropriate weights are discovered through a predictive error-driven learning algorithm. Following learning, each input provokes a pattern of activation over hidden units that depends on the acquired weight configuration—a learned internal representation of the input. Though the particular pattern acquired for a given item varies across training runs, the representations always encode the same similarity structure among items in the environment, representing items that are conceptually related with similar patterns of activation. 49

3.2 A) Architecture of the auto-encoder network used to generate the data for 10 model subjects used in subsequent simulations. The model has 36 input units (18 systematic), 14 hidden units (7 systematic), and 36 output units (18 systematic). The 22 irrelevant units are completely disconnected from the network, and stand for units that subserve an unrelated function but are anatomically adjacent to units of interest. B) Left: Ratio of between-domain to within-domain Euclidean distances for the representations coded over different sets of units in the network, averaged over the 10 model subjects. Distributed representations that encode the domain structure should have large distances for items from different domains and small distances for items from the same domain, and so should show a large ratio. While the systematic I/O units clearly code the domain structure to some degree, the systematic hidden units express the structure more strongly. 56

3.3 Results from the univariate analysis of simulated data. Bar height indicates the absolute value of the t statistic for the unit-wise contrast between conditions at the group level. Colored bars indicate units showing significant differences with p-values corrected to control the false discovery rate at $q < 0.05$. The red-blue scale indicates the direction of the contrast effect across model subjects, with red indicating units consistently showing greater activation for A items. S=systematic; A=arbitrary. 66

3.4 Result of the multivariate searchlight analysis of simulated data. Bar height indicates the mean classifier accuracy over subjects. Bars in red indicate searchlights where classification accuracy differed from chance over subjects, p-values corrected to control the false discovery rate at $q < 0.05$. Each column shows results from a different searchlight size indicated by the number on the far right. In the row labeled *Localized*, units were clustered by kind during the searchlight analysis; in the row labeled *Dispersed*, the systematic and arbitrary hidden units were shuffled together. S=systematic; A=arbitrary. 72
3.5 A) The results from the LASSO and ridge regression analyses. The blueness or redness of the bar conveys the frequency with which each unit was assigned a positive weight over subjects. Positive weights mean that activation at that unit will push the model towards labeling the current item as belonging to domain A. *LASSO*: Grey bars were selected less often than expected by chance given the overall rate of unit selection. *Ridge*: Bar saturation indicates which units would count as "selected" under three different policies, based on weight magnitude. Bright bars are in the top third of the distribution, pale bars are in the middle third, and gray bars are in the bottom third. S=systematic; A=arbitrary. B) Hit rate and precision for each regularization method, computed across the whole network (top), the I/O units only (middle), and the hidden units only (bottom) 81

4.1 Overview of W matrix. See text for details. 101

5.1 Hierarchical cluster analysis of the semantic (left) and visual (right) similarity structure among the 37 stimuli in the experiment. 115
5.2 Univariate solutions obtained for visual, audio, and catch trials (where catch trials called for explicitly making a relative size judgment, and audio and visual trials were when stimuli were covertly processed). Maps for each trial type were thresholded at FDR corrected $p < 0.05$ and overlaid. 126
5.3 Searchlight conjunction $p < 0.05$ FDR corrected. 8mm searchlights, 8mm blur. 127
5.4 Searchlight conjunction plot, each component thresholded at uncorrected $p < 0.01$ 129

5.5 A follow up searchlight analysis, in which data were averaged across modalities in one of two ways before comparing the resulting similarity structure to the target semantic matrix. Searchlight analysis was performed in parallel on data from each modality, each visiting the same voxel at the same time. The beta values in each searchlight were either averaged together directly, or a representational similarity matrix was generated for both sets of betas before averaging these resulting structures. The top row shows the result of doing the former, and the bottom row the latter. Note that the analysis of averaged betas resulting in a group level statistical map that could not be appropriately FDR corrected without excluding all voxels. While the averaged beta analysis echoes the univariate and searchlight conjunction analyses, the averaged structure analysis suggests that the similarity structure in the pvTL is more similar across modalities than in the STG. This may be because the size judgment is in part a visual, concrete-object oriented task. 130
5.6 Network RSA solution maps displaying the mean stability of each over subjects. The maps are thresholded with respect to the binomial test based on empirically estimated null distributions, as described in the methods. Note that only the maps displaying the solutions for visual models of visual structure could be appropriately statistically thresholded at FDR corrected $p < 0.05$. 133
5.7 Network RSA solution maps displaying the mean nodestrength of each over subjects. The maps are thresholded as in Figure 5.6. 137
5.8 Node strength as a function of distance (mm) from temporal pole along the AP and IS axes. Each colored dot is the average over a bin of distances equal to 10% of maximum distance from the temporal pole within each subject; each color is a different subject, and position along the x-axis is the mean distance within each of the 10 bins. Peak nodestrength can be seen to roughly correspond to "spokes", rather than the "hub". The red line is tracks the mean over subjects at each bin. 138
5.9 Overlap maps for 100 permutation models for subjects 1, 2, and 3 (representative sample). Voxels are not sampled at random by permutation test, and are clearly affected by the modality of the stimulus. 139
5.10 Network RSA overlap maps. Voxel color indicates the number of subjects for which that voxel was implicated after interpolating to the common space. The top row shows overlap across the average model for each subject; the bottom row shows overlap across a single model per subject. In these figures, no additional spatial smoothing is applied. For maps corresponding to the stability of voxels in individual subjects, see Appendix A. 140
5.11 The definitions of the hub and spokes that will be used in the following "ROI" and "lesions analyses that follow. See the text for details on how they were defined. 142
5.12 Error for Network RSA analyses. 144

6.1 Error associated with two network RSA models fit to the raw ECoG local field potentials (LFP) over electrodes. In one case (time info=no), the time points from 200 to 1200 ms were averaged together for each electrode. This corresponds to the window of peak performance, both for Y. Chen et al. (2016) and in our own work involving superordinate classification (no shown). In the other (time info=yes), time was not averaged down to a single point. Instead, aside from moderate data reduction (consecutive 10 ms bins of LFP were averaged), all time points were entered into the model at once. This means that if different points in time express different elements of structure, they could be linearly combined to produce a single, more complete structure and allow a better fit to the target similarity structure. Error is shown for the whole dataset, and living and nonliving items separately (to show subcategory structure is learned). This manipulation of time information did not yield significantly different performance. Colored dots indicate subjects by time manipulation, black dots indicate the mean, and error bars correspond to 95% confidence intervals for the test of the mean against zero. Errors are standardized with respect to the permutation models. 157

6.2 A) Electrode coverage over subjects, subject to a 4 mm FWHM Gaussian blur. One subject had electrodes implanted in the right hemisphere, so maximum overlap of coverage is 7 subjects in the right hemisphere. This blur is applied to make it comparable to the maps presented in B) which display the proportion of overlap at each electode for network RSA solutions obtained by modeling the target semantic structure in the LFPs collected with ECoG. To compose these maps, I first aggregated over cross validation models and consecutive 250 ms windows and applied a 4 mm FWHM Gaussian blur before assessing overlap across the 8 subjects. The count at each voxel was then divided by the number of times an electrode existed at that location, as shown in panel A). 158

6.3 Binary classifiers fit with LASSO were trained on each time point, and then evaluated on every other time point. Maps are thresholded based on a binomial test of the error out of 90 items, where the probability of obtaining an error less than 45 under the null hypothesis is $p = 0.5$. Points are thresholded at uncorrected $p < 0.001$. Each panel shows the matrix of train- and test-time windows for one subject. 161

6.4 Accuracy profiles, based on a k-means clustering of the temporal error maps shown in Figure 6.3 (k=5). Each colored line is a average of several rows from the previous figure (exactly which rows are indicated by the corresponding colored dots along the top of each panel), with error inverted into accuracy. The dotted line indicates p<0.001 by binomial test as a fairly conservative reference for above-chance generalization. . 162

A.1 Network RSA solution maps when modeling the target semantic structure described in Chapter 5 on visual trials, where stimuli were presented as line drawings. 200

A.2 Network RSA solution maps when modeling the target semantic structure described in Chapter 5 on audio trials, where stimuli were presented as characteristic sounds. 201

List of Tables

5.1 Whole brain network RSA error 134
5.2 ROI error. Uncorrected $p < 0.001$ exceeds the Bonferroni criteria for 16 tests. 144
5.3 Lesion error. Uncorrected $p < 0.001$ exceeds the Bonferroni criteria for 10 tests. 146