VISUAL SCENE PERCEPTION IN THE HUMAN BRAIN: CONNECTIONS TO MEMORY, CATEGORIZATION, AND SOCIAL COGNITION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF COMPUTER SCIENCE
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Christopher Baldassano
February 2015

© 2015 by Christopher Anthony Baldassano. All Rights Reserved.
Re-distributed by Stanford University under license with the author.

This work is licensed under a Creative Commons Attribution-Noncommercial 3.0 United States License.
http://creativecommons.org/licenses/by-nc/3.0/us/

This dissertation is online at: http://purl.stanford.edu/hn881py5906

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Fei-Fei Li, Primary Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Percy Liang

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

Diane Beck

Approved for the Stanford University Committee on Graduate Studies.

Patricia J. Gumport, Vice Provost for Graduate Education

This signature page was generated electronically upon submission of this dissertation in electronic format. An original signed hard copy of the signature page is on file in University Archives.

Abstract

The human visual system faces a monumental data processing challenge: using about a pound of slow, inexact biological processors, it must analyze the barrage of constantly-shifting light patterns hitting the eye and quickly extract a stable, high-level model of the environment around us. Almost every piece of this process is mysterious: exactly what information is being gleaned from the visual signal, how this information is represented, and how this processing is implemented in neural circuits. Despite the superiority of silicon computers for most big-data processing, our emulations of the human visual system are still rudimentary, and can capture only basic information from visual images such as which objects are present. In this work, I describe a number of projects toward understanding higher-level processing of visual scenes. The first examines the neural basis of understanding human-object interactions, showing how an emergent property of a scene (created by the interaction of two scene parts) can activate representations in social cognition regions. The second investigates how scenes are categorized, arguing that one of the fundamental features encoded about a scene is the type of actions which could be performed in that environment. Finally, I present a large body of work on how scene processing interacts with long-term memory systems. These chapters describe several novel types of mathematical models for measuring connections between brain regions, and end with a new organizing proposal for scene perception regions.

To my parents, who launched me toward the heavens,
To my wife, who held my hand in the dark,
To my son, my shining star

Acknowledgments

Completing a PhD can easily turn into a demoralizing and isolating experience, as scientific research inevitably involves long strings of experimental failures, data analysis frustrations, and manuscript rejections. I count myself very lucky to have had such an extended network of colleagues, friends, and collaborators that kept me intellectually engaged and excited about research. These include too many people to name in the larger Vision Science and Machine Learning communities, both at Stanford and at other institutions through local and international conferences. Thank you for raising questions, giving feedback, and sharing your passion for pushing the boundaries of scientific knowledge.

Thank you to the members of Stanford Vision Lab and the University of Illinois at Urbana-Champaign Attention and Perception Lab, who helped me get started in a new field during my first year of graduate school and have offered me their help every day since. Conversations in the lab are never boring thanks to you.

Thank you to my co-authors, Marius Cătălin Iordan, Michelle R. Greene, and Andre Esteva, for being a constant sounding-board for both big ideas and minute tweaks to publications and presentations. I would like to especially thank Michelle for mentoring me over the past few years, bringing a wealth of knowledge into our office about the history and broad field of visual perception.

Thank you to my reading committee and defense committee members, Noah Goodman, Kalanit Grill-Spector, and Percy Liang, for being great role models both in your constructive scientific comments and your friendly, welcoming personalities.

Thank you to my advisors, Fei-Fei Li and Diane M. Beck, who taught me how to think like a scientist and have consistently challenged me to set a high bar for every

aspect of my experimentation and presentation. Nearly every word in this document has been influenced or edited by one or both of you, and the things I learned about experimental design, data analysis, paper writing, and scientific presentations from you are the most valuable skills I will take away from my PhD. Through all of the ups and downs of my grad school career - the frantic deadlines, the cautious optimism of positive results, real tragedies and imagined tragedies - you have contributed both passion and perspective.

Most importantly, thank you to my family. I would never had ended up at Stanford without my parents, who got me excited about science before I learned to read, and have made my education a priority over everything else in their lives. My wife Linda has been my devoted supporter, constant companion, and best friend, dropping everything to come with me to Stanford for my PhD. My two-year-old son Will doesn't quite know what I do all day, but constantly reminds me to be curious about the world and find joy in unexpected places.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Our eyes are our window on the world. Seeing is such a primary part of our sensory and conscious experience that we tend to describe most cognitive processes in terms of visual metaphors; to avoid being "in the dark" and "blind" to the truth, we want facts to "come into focus" so that we can "see" what's happening. The pattern of light entering our eyes is incredibly complicated, in constant flux due to motion in the world or eye motion, and spans a massive range of intensities. Processing this rich stream of information requires a large amount of real estate in the brain, with about 20% of the cortex dedicated almost exclusively to visual processing [303] and many other areas that are driven by visual input [108].

Given the complexity of the visual stimulus on the retina, a key question is to determine the correct level of abstraction for studying a given region of the brain. What properties of the image do particular neural circuits respond to? Early work, both in physiological experiments [139] and mathematical modeling [189], focused on the representation of oriented edges. For certain regions of the nervous system, including the retina and its primary cortical output at the back of the brain (V1), the edges present in an image predict a large portion of the neural response, abstracting away some of the details of the visual input and focusing on local, structured contrast differences. This representation, however, is still rather unsatisfying, since it has very little contact with semantically meaningful parts of the world; one would never describe the Mona Lisa in terms of a map of oriented edges.

A higher level of abstraction that currently a primary focus in neuroscience and computer vision is that of object recognition. Patterns of brain activity in certain regions, especially along the ventral stream extending forward from V1, show increasing abstraction from the particular edges present in an image and instead have responses predicted by the identities of the object(s) present in the image [249]. This level of analysis is an exciting and challenging one, since it connects visual input with semantic concepts, such as those used in human language. Building invariance to object pose, position, and lighting is a highly nontrivial task, though great progress has been made over the last several years in building computer models that can achieve near-human performance in some situations [247].

There are, however, even more complex descriptions of natural images, which go beyond listing objects and depend on larger structures of entire visual scenes. These include features like overall geometry, interactions between objects, or more abstract global properties such as aesthetic beauty or memorability of a scene. The brain has a number of regions that are related to these higher-order properties beyond object recognition, which show a larger response to full scenes than to isolated objects [170]. The field of scene perception is concerned with understanding these higher-order representations, discovering the neural mechanisms by which they are constructed, and describing their relationship with behaviors such as navigation, categorization, or memorization.

In this work I describe several projects looking at mental representations beyond object recognition. Chapter 2 investigates how adding an interaction between two parts of a scene (here a human and an object) changes neural activity patterns, producing an emergent representation that is more than the sum of its parts. Chapter 3 examines an even higher-level property, proposing that scene meaning is largely driven by the actions that it affords.

The remainder of the chapters examine neural responses that go beyond even the image itself. The processing in certain brain regions combines visual properties with past memories, such as contextual and navigational information, allowing for very long-term interactions between visual input over the course of a human lifetime. We characterize the properties of these regions largely in terms of their connections, using

brain imaging data from a variety of sources. Chapter 4 describes our novel approach for identifying connectivity differences between nearby brain regions, which we apply in Chapter 5 to discover specialized areas for visual and memory processing within a primary scene-processing region. We then extend this analysis to the whole brain, describing a new whole-brain connectivity clustering method in Chapter 6, and then use this approach in Chaper 7 to propose a large-scale framework for understanding how visual information is incorporated with past memories. The key results from each of these projects are summarized in Chapter 8.

Chapter 2

Human-object interactions are more than the sum of their parts

Understanding human-object interactions is critical for extracting meaning from everyday visual scenes, and requires integrating complex relationships between human pose and object identity into a new percept. To understand how the brain builds these representations, we compared conducted two fMRI experiments in which subjects viewed humans interacting with objects, non-interacting human-object pairs, and isolated humans and objects. A number of lateral visual regions are involved in processing human-object interactions, including the lateral occipital complex (LOC) and the extrastriate body area (EBA). The representations in these regions, however, are at least partially driven by object identity (for LOC) and/or human pose (for EBA), and not specifically the interaction between the two. However, a region anterior to EBA, in the posterior superior temporal sulcus (pSTS), represents interactions in a way that is not simply a linear combination of object and pose information, indicating that this region encodes human-object interactions as more than the sum of their parts. These results reveal the distributed networks underlying the representation of emergent visual concepts, such as the social perception of human-object interactions. This chapter is joint work with Diane M. Beck and Fei-Fei Li.

2.1 Introduction

Our visual experience consists not of a jumble of isolated objects but of coherent scenes, in which objects are arranged in meaningful relationships. Neuroscientists have long studied isolated object recognition, and we have at least a qualitative understanding of where and how the brain constructs invariant object representations [78]. A largely separate body of research has studied the perception of complex scene images containing diverse collections of objects, and has identified brain regions supporting the recognition of broad scene categories [302]. The connection between these two domains, however, has gone largely unstudied: how do objects come together to compose complex scenes with emergent semantic properties?

One scene category in which semantic meaning is critically driven by the relationship between scene components is that of human-object interactions. Understanding the differences between images of people riding horses, petting horses, leading horses, and feeding horses, for example, cannot be accomplished by simply recognizing the person and horse in isolation. Moreover, although observing human-object interactions is essential for both developmental learning about object manipulation [305] as well as everyday social cooperation, we know surprisingly little about how they are encoded in the brain. Information about object identity and the relative positions of body parts must be combined to produce a high-level percept of the human's actions and goals, requiring an integrated neural representation that is "more than the sum of its parts."

Human-object interactions can vary along two dimensions: the identity of the object, and the way in which the human is interacting with the object. We hypothesize, however, that extracting meaning from human-object interactions will require areas sensitive not just to object or pose, but also to higher-order emergent features of the interaction. Using multi-voxel pattern analysis (MVPA), we compared the representation of human-object interaction categories with linear combinations of responses evoked by isolated humans and objects. Although some multi-object scenes can be modeled by a linear pattern average of the responses to each object individually [13, 150, 182, 324], we find that human-object interactions break this linear assumption in

regions such as the posterior superior temporal sulcus (pSTS), evoking novel category representations distinct from pattern averages. In particular, this analysis revealed nonlinear representations across multiple components of the social cognition network [256].

We conclude that understanding human-object interactions involves distributed occipitotemporal networks, which support the creation of emergent representations in social cognition regions. These results demonstrate the critical impact of interactions between scene components on scene representation, providing a new bridge between isolated object perception and full scene recognition.

2.2 Materials and Methods

2.2.1 Stimuli

For Experiment 1, we created 128 person-riding-horse and 128 person-playing-guitar images by manually segmenting images from the Stanford 40 Actions database [318]. Each image was scaled to contain the same number of pixels, such that every image fit with a 450x450 square. We created 128 horse images (using images and masks from the Weizmann horse database [29]) and 128 guitar images (using images from the Caltech Guitar dataset, and manually segmenting them from the background [284]). We also created 128 person images using images and masks from INRIA Annotations for Graz-0 [193, 217] in additional to manually segmented people from the Stanford 40 Actions database. Each of the isolated images was scaled to contain half as many pixels as the interacting images. Half of the horses were horizontally mirrored (since all of the Weizmann horses face to the left) and the guitars were rotated so that the distribution of the neck angles exactly matched that of the person-playing-guitar images.

To create the non-interacting images, we overlaid an isolated person and isolated object, with the person and object chosen so as to avoid pairings that appeared to be interacting. The person and object images were each centered on a point drawn from a Gaussian distribution around the fixation point, with standard deviation set equal to

the standard deviation of objects and people relative to the image centers in the action images (0.62 degrees of visual angle). To make the images as qualitatively similar to the action images as possible, the person images were placed on top of (occluding) the horse images, but were placed behind the guitar images. The distribution of the relative sizes of the person and object was exactly matched to that of the action images, and the composite images were scaled to have the same number of pixels as the interacting images. The total number of stimuli in Experiment 1 was (3 isolated + 2 interacting + 2 non-interacting)*(128 images) = 896 images.

For Experiment 2, 40 images were collected from Google Images and Flickr for each of 4 action categories: pushing shopping carts, pulling luggage, using a computer, and using a typewriter. All of the 160 images were manually segmented to remove the person and object from the background, and scaled to have the same number of pixels such that every image fit within a 900x900 square. We manually separated the person and object, giving isolated object images, isolated human images, and human-object interaction images. Any overlap between the person and object was covered with a black rectangle, which was applied to all three versions of the image. All images were superimposed on a background containing 1/f noise in each color channel, in both their original orientation and mirrored left-to-right, for a total of (2 orientations)*(4 categories)*(3 conditions)*(40 images) = 960 stimuli.

2.2.2 Experimental Design

Each subject viewed blocks of images from different categories, with a 12s gap between blocks. Each block started with a 500ms fixation cross, and then 8 images each presented for 160ms with a 590ms blank inter-trial interval. Subjects were instructed to maintain fixation at the center of the screen, and perform a task using a button-box. In Experiment 1, subjects participated in 8 runs, each of which contained two blocks of each of the seven stimulus categories (isolated humans, guitars, and horses; non-interacting human-guitar and human-horse pairs; humans riding horses and humans playing guitars), for a total of 14 blocks (126 TRs) per run. Subjects performed a 1-back task, detecting consecutive repetitions of the same image, which occurred 0,

1, or 2 times per block. In the Experiment 2, subjects performed 14 runs, which were grouped by consecutive pairs into 7 pseudo-runs. Within each of the first 5 pseudo-runs, each run contained 8 blocks, one from every isolated (person/object) category, for a total of 79 TRs per run. The last 2 pseudo-runs each contained 20 blocks each (10 per run), with 5 blocks drawn from each interaction category, for a total of 97 TRs per run. Subjects performed a 1-back task, detecting consecutive images that were mirror images of each other, which occurred 0 or 1 times per block (with the same frequency for all categories and conditions). Regions of Interest

The locations of the category-selective ROIs for each subject's brain were obtained using standard localizer runs conducted in a separate fMRI experiment. Subjects performed 2 runs, each with 12 blocks drawn equally from six categories - child faces, adult faces, indoor scenes, outdoor scenes, objects (abstract sculptures with no semantic meaning), and scrambled objects - and an additional run with 12 blocks drawn from two categories (body parts and objects). Blocks were separated by 12 s fixation cross periods, and consisted of 12 image presentations, each of which consisted of a 900 ms image followed by a 100 ms fixation cross. Each image was presented exactly once, with the exception of two images during each block that were repeated twice in a row. Subjects were asked to maintain fixation at the center of the screen, and respond via button-press whenever an image was repeated. The ROIs were defined such that each subject had approximately the same total volume of clustered voxels: LOC, approx. 4800 mm^3 for Objects >Scrambled contrast in lateral occipital cortex; EBA, peak clusters of approx. 2900 mm^3 for Body Parts >Objects contrast in occipital cortex; parahippocampal place area, peak clusters of approx. 2900 mm^3 for Scenes >Objects contrast near the parahippocampal gyrus. The volume of each ROI in mm^3 was chosen conservatively, based on previous results [104].

We also defined a pSTS ROI for Experiment 2, based on the voxel showing the peak response in Experiment 1 (see Figure 2.4). This was defined in MNI space as all voxels within 10mm of the peak pSTS voxel, and then transformed into each subjects native space. Additionally, we defined retinotopic regions PHC1/2 using a group-level field map atlas [304].

2.2.3 Scanning parameters

For Experiment 1 and the ROI localizers, imaging data were acquired with a 3 Tesla G.E. Healthcare scanner. A gradient echo, echo-planar sequence was used to obtain functional images [volume repetition time (TR), 2 s; echo time (TE), 30 ms; flip angle, 80; matrix, 128x128 voxels; FOV, 20 cm; 29 oblique 3 mm slices with 1 mm gap; in-plane resolution, 1.56x1.56mm]. The first four volumes of each run were discarded, and the functional data were then motion-corrected and each voxel's mean value was scaled to equal 100, using the AFNI software package [69]. We collected a high-resolution (1x1x1mm voxels) structural scan (SPGR; TR, 5.9 ms; TE, 2.0 ms; flip angle, 11) in each scanning session. For computing whole-brain results at the group level, each subject's anatomy was registered by hand to the Talaraich coordinate system. Images were presented using a back-projection system (Optoma Corporation) operating at a resolution of 1024 x 768 pixels at 75 Hz, such that images covered approximately 14 degrees of visual angle.

For Experiment 2, imaging data were acquired with a different 3 Tesla G.E. Healthcare scanner. A gradient echo, echo-planar sequence was used to obtain functional images [volume repetition time (TR), 2 s; echo time (TE), 30 ms; flip angle, 77; matrix, 80x80 voxels; FOV, 23.2 cm; 42 oblique 2.9 mm slices; in-plane resolution, 2.9x2.9mm]. The first six volumes of each run were discarded, and the functional data were then motion-corrected and each voxel's mean value was scaled to equal 100, using the AFNI software package [69]. We collected a high-resolution (0.9x0.9x0.9mm voxels) structural scan (BRAVO; TR, 7.24 ms; TE, 2.78 ms; flip angle, 12) in each scanning session. For computing whole-brain results at the group level, each subject's anatomy was registered automatically to the Talaraich coordinate system. Images were presented using an LCD display (Resonance Technology) operating at a resolution of 640x480 at 240Hz, visible from a mirror within the head-coil, such that images covered approximately 12 degrees of visual angle.

2.2.4 Subjects

We collected data from 10 subjects (2 female, ages 22-28, including one of the authors) in Experiment 1, and 12 subjects (5 female, ages 20-32, including one of the authors, five subjects overlapping with first experiment). Subjects were in good health with no past history of psychiatric or neurological diseases, and with normal or corrected-to-normal vision. The experimental protocol was approved by the Institutional Review Board of Stanford University, and all subjects gave their written informed consent.

2.2.5 Mean Signal Analysis

In order to compare the mean signal response to noninteracting and interacting stimuli in each ROI in Experiment 1, we used a standard regression model. The stimulus regressors were modeled as step functions equal to 1 during a stimulus block and 0 elsewhere, convolved with the standard AFNI hemodynamic response function [69]. In addition, 30 nuisance regressors were added to the model: 3 for each of the 8 runs (constant offset, linear trend, quadratic trend), and 6 motion correction estimates (3 rotation and 3 translation). The estimated beta weights for the non-interacting and interacting regressors were then recorded in units of percent signal change.

2.2.6 ROI Decoding

For all MVPA decoding analyses in both Experiments, each fMRI timepoint was first assigned a stimulus label; all timepoints that occurred while a stimulus block was being presented (shifted by 6 seconds to account for hemodynamic lag) were assigned to the corresponding stimulus, while all other timepoints were labeled as inter-block timepoints. Classification was performed using linear support vector machines, using the MATLAB LIBSVM library [62]. In Experiment 1, we selected six runs for training, used one validation run to tune the soft-margin hyperparameter c, and tested on the remaining run (results are averaged over all possible choices of testing and validation runs). In Experiment 2, nine blocks of each stimulus category were selected for training, and the classifier was then tested on the remaining blocks, for fixed

c=0.1 (results are averaged over all choices of testing block). For cross-decoding, the classifier was also tested on all blocks corresponding to other stimulus conditions.

When applying this method to localizer ROIs LOC and EBA, we first excluded voxels that were not sensitive to visual stimulation, to improve decoding accuracy. All voxels were ranked based on the absolute value of their z-score for within-block timepoints versus inter-block timepoints. The top 40% of the voxels were used in decoding (the number of voxels retained was set to 40% of the group mean size for each region, so all subjects retained the same number of voxels in a given region), but our results are not sensitive to the number of voxels used (see Supplemental Figure A1). Note that this type of voxel selection does not introduce a circularity bias (as described in [298]) since (a) we are selecting only for visual sensitivity, not for between-condition effects, and (b) the selection is based only on training data.

In Experiment 1, two separate classifiers were trained: one to discriminate between non-interacting stimulus categories (humans with horses vs. humans with guitars) and one to discriminate between interacting stimulus categories (humans riding horses vs. humans playing guitars). In the first analysis, the performance of these classifiers was measured on the non-interacting and interacting testing timepoints, respectively. For the cross-decoding analysis, we created pattern-average testing timepoints, by averaging the mean response to humans in the testing run with all isolated object timepoints in the testing run. The non-interacting and interacting decoders were then applied to classify the category (human+horse vs. human+guitar) of these pattern-average timepoints.

In Experiment 2, three classifiers were trained: one to discriminate between isolated objects drawn from different action images, one to discriminate between isolated humans drawn from different action images, and one to discriminate between images of full human-object action images. This last classifier was also applied in a cross-decoding analysis, to decode isolated object timepoints, isolated human timepoints, and pattern-average timepoints (created by averaging the four timepoints corresponding to an isolated object category in a given run with the four timepoints corresponding to the isolated human from the same category in the same run, yielding a new set of four pattern-average timepoints).

2.2.7 MVPA Searchlight Analyses

We also ran these analyses in a whole-brain searchlight. Spheres with 7mm radius were centered on a grid with 8mm spacing. For each sphere, all voxels whose centers fell within its radius were used as a region of interest, and decoding analyses were performed as for the ROIs (without any voxel selection, and with soft-margin hyper-parameter set to the average of its value during the ROI experiments). Note that each sphere intersected with all 26 neighboring spheres, since the maximum distance between a sphere and its neighbors (38) is less than twice the radius (27). To produce a decoding accuracy map for each subject, the accuracy for each voxel was calculated as the mean accuracy of all searchlights that included that voxel. To determine which voxels showed significant differences between conditions, a Monte-Carlo permutation test was used. The analysis used on the real data was run 1000 times on data for which the timepoint labels were randomly shuffled between categories being used for training or testing. For example, when decoding riding-horse vs. playing-guitar, the labels of all riding-horse and playing-guitar timepoints were randomly shuffled. A threshold value was then fixed such that less than 5% of the sampled maps contained any above-threshold clusters larger than 100 voxels, and this same threshold was applied to the real data (see Supplemental Figure A2).

2.3 Results

2.3.1 Experiment 1

We constructed a stimulus set with three types of images (see Figure 2.1): isolated humans, guitars, and horses; non-interacting human-horse and human-guitar pairs, in which humans and objects were simply pasted together without an interaction; and interacting humans riding horses and humans playing guitars. These actions were chosen since both involve humans that are roughly vertical and centered, so that the non-interacting and interacting images had similar construction. As described in Experimental Procedures, the non-interacting images were constructed to match the statistics of the interacting images as closely as possible, so that the only difference

from the interacting images is that the human body is not correctly positioned to interact with the object.

Not surprisingly, given the subtle differences in the stimuli, a univariate analysis comparing interacting and non-interacting stimuli yielded no differences in occipitotemporal regions (LOC: t_9=-0.47, p=0.65; EBA: t_9=-0.78, p=0.46; FFA: t_9=-0.35, p=0.74; two-tailed t-test), and performing a whole-brain regression analysis contrasting interacting ¿ non-interacting failed to find any voxels meeting the threshold of FDR<0.05. Thus, we used MVPA decoding, which is more sensitive to fine-grained differences among stimuli, to find regions that showed a greater pattern difference between the interacting human-guitar and human-horse categories, compared to the non-interacting human-guitar and human-horse categories. Such a result would indicate that a region better distinguishes between the two human-object categories when an interaction is present, implying that this region contains specialized processing for human-object interactions. We found that the category (horse vs. guitar) could be decoded for both non-interacting and interacting stimuli in all three areas (Non-interacting: LOC: t_9=4.19, p <0.01; EBA: t_9=3.24, p <0.01; FFA: t_9=2.51, p=0.02. Interacting: LOC: t_9=3.50; EBA: t_9=5.41; FFA: t_9=3.47; all p <0.01; one-tailed t-test). LOC showed nearly identical decoding rates for both stimulus types (t_9=-0.13, p=0.90; two-tailed t-test), but EBA showed a consistent difference in the decoding rates for non-interacting and interacting stimuli, with significantly better category decoding for interacting stimuli (EBA: t_9=2.82, p=0.02; two-tailed t-test). These results are shown in Fig. 2.2 (solid bars, NN and II). A searchlight analysis for areas showing this same preference for interacting stimuli (Fig. 2.3) produced areas consistent with our ROI results; we found voxels in right EBA that gave better decoding for interacting stimuli. Additionally, this contrast revealed a more anterior patch of cortex around the right pSTS showing the same preference for interacting stimuli.

As discussed above, perceiving human-object interactions requires a representation which is more than the sum of its parts. As shown in previous work, some regions response to a pair of simultaneously-presented stimuli is simply the average of the

responses to the individual stimuli [13, 150, 182, 324]. If a region is sensitive to human-object interactions, however, we would expect the regions response to an interacting human and object to not be simply the sum of its parts, but to be qualitatively different from a simple average of human and object. We hypothesize that regions specifically sensitive to human-object interactions should have specialized (non-linear) representations for categories of interacting human-object pairs, but not for non-interacting categories.

We can find regions showing this behavior by using a cross-decoding approach. After training two classifiers, as before, to decode non-interacting human-horses vs. human-guitars, and to decode interacting human-horses vs. human-guitars, we can then attempt to use these classifiers to decode the pattern average of humans and horses vs. the pattern average of humans and guitars. If the features used to represent categories of human-object pairs are simply linear averages of the features of isolated humans and objects, then these classifiers trained on pairs should generalize well to the individual response averages. The ROI results in Fig. 2.2 show a compelling difference between cross-decoding in the non-interacting and interacting cases. When trained on non-interacting responses, classifiers for all three regions were able to decode the pattern-averaged stimuli above chance (NA bars; LOC: t_9=6.03, p <0.01; EBA: t_9=5.27, p <0.01; FFA: t_9=2.05, p=0.04; one-tailed t-test), with only a small drop in performance compared to decoding non-interacting stimuli (LOC: t_9=0.02, p=0.49; EBA: t_9=0.55, p=0.29; FFA: t_9=0.26, p=0.40; one-tailed t-test). This indicates that the features used to represent non-interacting stimulus categories can be effectively used to classify the average of the human and object patterns, demonstrating that none of these regions represent non-interacting human-object pairs in a specialized, non-linear way.

Cross-decoding results showed a different pattern, however, when the classifier was trained on interacting stimuli (IPA bars). In LOC, pattern-average responses could still be decoded above chance by the interacting-stimulus classifier (t_9=2.44, p=0.02; one-tailed t-test), and decoding performance showed only a small, nonsignificant drop compared to decoding interacting stimuli (t_9=0.91, n.s.; one-tailed t-test). In EBA, the classifiers trained on interacting stimuli showed a significant drop in performance

when used to decode pattern-averages (EBA: t_9=3.09, p <0.01), and were unable to decode the pattern-averaged stimuli above chance (EBA: t_9=1.49, n.s.). This drop was significantly larger than that in LOC (t_9=1.91, p=0.04; one-tailed t-test).

We can look outside our ROIs and search for all regions with this pattern of results for human-object interaction categories by performing a searchlight analysis, identifying searchlights with a greater nonlinearity (drop in performance when cross-decoding) for interacting stimuli than non-interacting stimuli (Fig. 2.4). In addition to EBA, this contrast reveals regions around the pSTS (peak voxel at MNI [54, -43, 12]) and temporoparietal junction (TPJ) in both hemispheres, right dorsal PCC, and the right angular gyrus in the inferior parietal lobule (IPL). As discussed below, these areas largely map onto the network of regions involved in social cognition and understanding action intent, consistent with interactions between the human and object being an important component of the semantic meaning of a social scene. These results indicate that the representation of human-object interaction categories in these body-related regions is not simply driven by a linear combination of isolated object identity and a person activity pattern.

2.3.2 Experiment 2

The results of Experiment 1 demonstrate that body-related regions do not represent person riding horse as a linear combination of person and horse, but it is possible that some of this effect is due to differences in pose; although pose is in a sense a configural property of the human, pose representations do not incorporate both human and object information into a single emergent unit. We tested this possibility using a new experiment, with a new set of stimuli (Figure 2.5). Subjects viewed four new action categories, but also viewed the objects and humans from these interaction images in isolation. This design ensured that objects and poses were exactly matched between the isolated and interacting images, so that a failure to generalize decoding from interacting to pattern averaged responses would necessarily indicate a nonlinearity in category representation of interaction.

We performed MVPA decoding using the same approach as in Experiment 1,

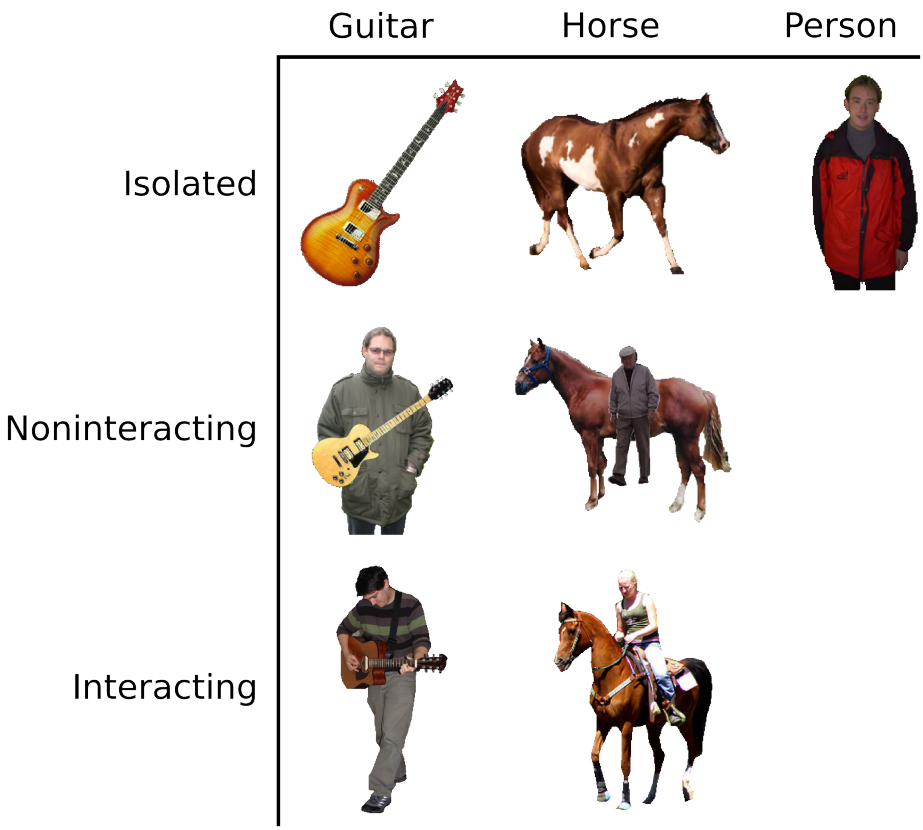


Figure 2.1: **Example stimuli from Experiment 1**. Subjects were shown 128 images in each of seven categories: isolated guitars, horses, and people; non-interacting human-guitar pairs and human-horse pairs; and interacting humans playing guitars and humans riding horses.

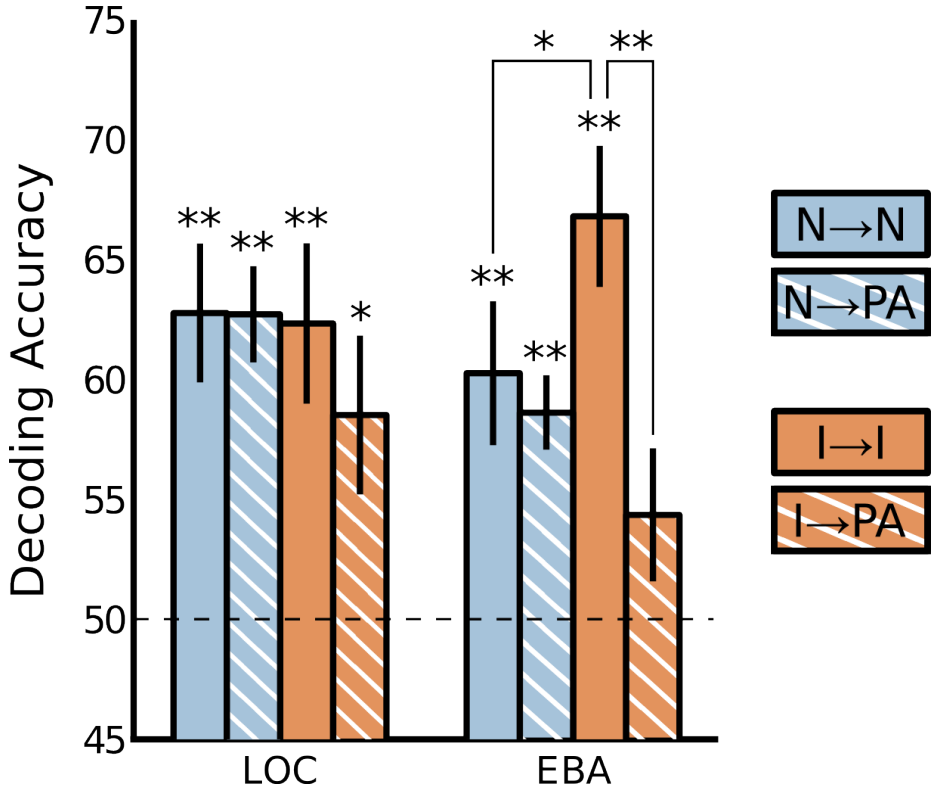


Figure 2.2: **MVPA decoding and cross-decoding for Experiment 1**. The stimulus category (person and horse vs. person and guitar) can be decoded in all three regions, whether an interaction is present (I) or not (N). However, EBA shows a significant increase in decoding accuracy for interacting stimuli (II) compared to non-interacting (NN), indicating that the image category is better represented in this region when an interaction is present. LOC, however, shows nearly identical decoding accuracies for the two conditions. Classifiers trained on responses to non-interacting stimuli in all three areas generalize well to pattern-averages of individual humans and objects (NPA), but the interacting classifier only generalizes to pattern-averaged responses in LOC (IPA). This indicates that EBA has a representation for human-object interaction categories which is not similar to the average of responses to isolated humans and objects. These results are consistent regardless of the number of voxels selected per region (see Figure S1). Error bars denote s.e.m., *p<0.05, **p<0.01.

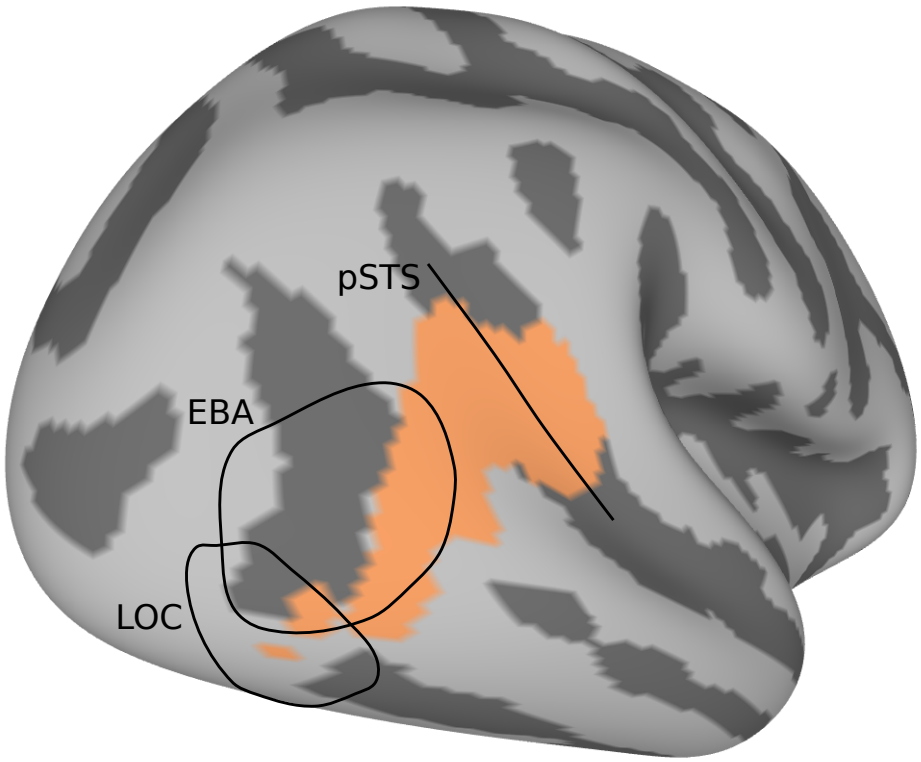


Figure 2.3: **MVPA decoding difference searchlight for Experiment 1**. Searching all of cortex for regions having higher decoding accuracy for interacting (II) than non-interacting (NN) stimuli yields a result consistent with the ROI-based analysis. Searchlights showing this preference for interacting stimuli consistently included voxels in the anterior EBA and posterior STS in the right hemisphere. $p<0.05$ cluster-level corrected.

Figure 2.4: **MVPA cross-decoding searchlight for Experiment 1**. Colored voxels are those showing a larger nonlinearity in the interacting condition (II minus IPA) compared to the nonlinearity in the non-interacting condition (NN minus NPA). In addition to EBA, this measure identifies regions around the posterior STS (peak effect marked with a dot) and TPJ in both hemispheres, the right dorsal PCC, and the right angular gyrus, $p<0.05$ cluster-corrected.

looking now at 4-way action classification for objects alone, people alone, and person-object interactions. As before, we measured whether the representation of human-object interactions was similar to the representation of its components using cross-decoding; we applied the classifier trained on full interactions to classify objects alone, people alone, and pattern averages of objects and people. In addition to the ROIs used in Experiment 1 (LOC and EBA), we also defined a pSTS ROI with a 10mm radius around the voxel that showed the strongest effect in Experiment 1 (Figure 2.4).

The decoding results are displayed in Figure 2.6. Both LOC and EBA show above-chance decoding for objects, poses, and interactions (Objects: LOC t_{11}=6.12, p <0.01; EBA t_{11}=2.09, p <0.05; Poses: LOC t_{11}=4.84, p <0.01; EBA t_{11}=2.30, p <0.05; Interactions: LOC t_{11}=4.32, p <0.01; EBA t_{11}=2.93, p <0.01; one-tailed t-test). When applying the interaction decoder to classify objects alone (such that the only information about stimulus category comes from objects, as in Experiment 1), we largely replicate our previous results LOC still shows above-chance decoding, while EBA does not (LOC t_{11}=2.48, p <0.05; EBA t_{11}=1.61, n.s.). We do see a significant performance drop in both ROIs (LOC t_{11}=2.54, p <0.05; EBA t_{11}=2.35, p <0.05), while the LOC cross-decoding drop did not reach significance in Experiment 1. A more substantial difference can be seen in the interaction to pose cross-decoding, in which the interaction decoder was used to classify isolated people posed for a particular action. Here LOC did not perform above chance, while EBA did (LOC t_{11}=0.52, n.s., EBA t_{11}=2.81, p <0.01). When trying to classify pattern averages of isolated objects and people (from a particular action class), both LOC and EBA perform above chance (LOC t_{11}=3.34, p <0.01; EBA t_{11}=3.22, p <0.01). The drop between interaction decoding and the mean of all cross-decoding conditions is significant in LOC, and marginally significant in EBA (LOC t_{11}=3.18, p <0.01; EBA t_{11}=1.73, p=0.056).

A different pattern of results was seen in the pSTS, just anterior to EBA. Here object and pose decoding was not significant (Object: t_{11}=-0.93, n.s.; Pose: t_{11}=1.57, n.s.), but full interactions could be decoded above chance (t_{11}=2.32, p <0.05). The interaction decoder did not generalize to isolated objects or poses, or the average of the two (Object: t_{11}=0.64, n.s.; Pose: t_{11}=0.45, n.s; Pattern Average: t_{11}=0.69, n.s),

and there was a significant overall drop between interaction decoding and the mean of all cross-decoding conditions (t_{11}=1.80, p <0.05). The classification accuracy on pattern averages is also significantly lower than the mean of LOC and EBA (t_{11}=1.93, p <0.05).

To further analyze the posterior-to-anterior decoding differences in lateral temporal cortex, we performed a searchlight analysis to measure both interaction classification and generalization to pattern averages. As shown in Figure 2.7, the results were largely consistent with the ROI analyses; classifying interactions was above chance in the majority of voxels within LOC and EBA, and each contained subregions (superior LOC and posterior EBA) where this classifier also generalized to decode pattern averages. In the anterior portion of EBA and pSTS, however, interaction cross-decoding fails on pattern averages; this posterior-anterior difference can be seen on an axial slice through lateral cortex (Figure 2.7b), showing that cross-decoding accuracy drops rapidly around pSTS while interaction decoding remains relatively high. Interestingly, the searchlight also revealed significant cross-decoding in the parahippocampal place area (PPA), restricted primarily to the retinotopic maps within this area (PHC1/2) [9] (interaction decoding was also above-chance in this region, but failed to meet the significance threshold).

2.4 Discussion

Using carefully constructed images of humans and objects, along with three different types of MVPA searchlight analysis, we identified regions in occipitotemporal cortex responsible for representing human pose and object identity, and for binding humans and objects together into a coherent interaction. Previous work has studied humans and objects in isolation (e.g. [81, 168]), but we have characterized for the first time how categories of pose and object identity are encoded in the context of human-object interactions. Decoding results in LOC revealed robust representations about action categories, which were at least partially driven by object identity information (Experiments 1 and 2) but seemingly unrelated to pose information (Experiment 2). EBA also showed consistent interaction decoding, but was not driven by object identity

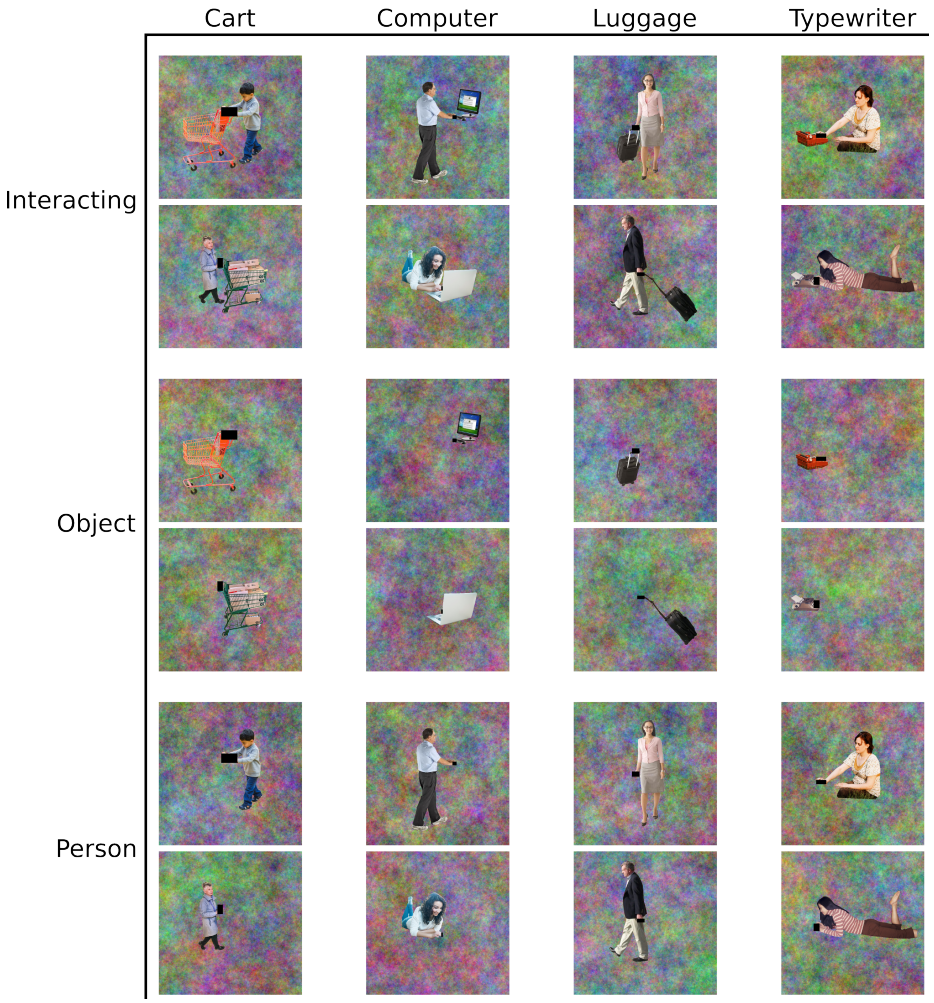


Figure 2.5: **Example stimuli from Experiment 2**. Subjects viewed images of human-object interactions from four different action categories (pushing carts, using computers, pulling luggage, and typing on typewriters), and also viewed the objects and people from these images in isolation.

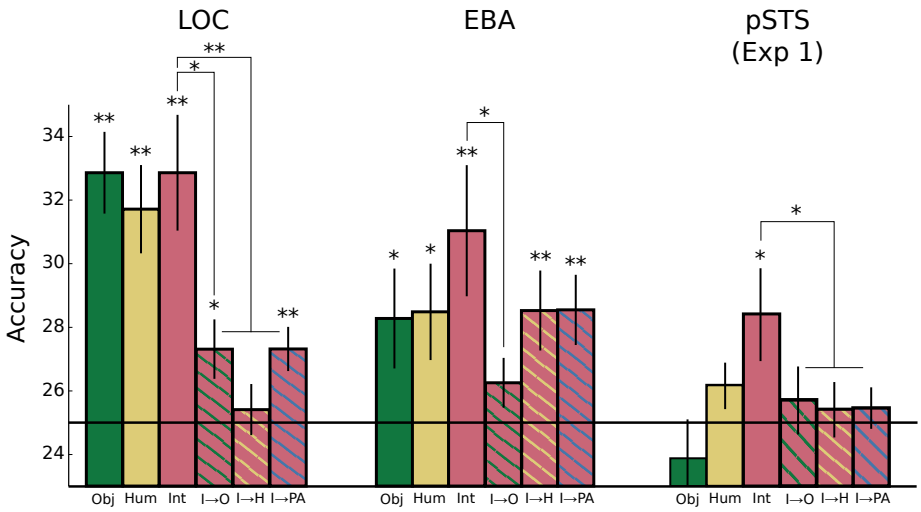


Figure 2.6: **MVPA decoding and cross-decoding for Experiment 2**. Both LOC and EBA show significant decoding of action category from isolated objects, isolated humans, or full actions. As in experiment 1, the classifier trained on full interactions performs above-chance on objects only in LOC, though the cross-decoding accuracy drop here is significant in both LOC and EBA. EBAs interaction classifier does, however, generalize well to human poses (while LOCs does not). Therefore both LOC and EBA classifiers show generalization to pattern averages, driven by object information in LOC and by pose information in EBA. The pSTS, on the other hand, localized based on results in Experiment 1, shows above-chance decoding only for human-object interactions, and does not generalize to pattern averages. Error bars denote s.e.m., *p<0.05, **p<0.01.

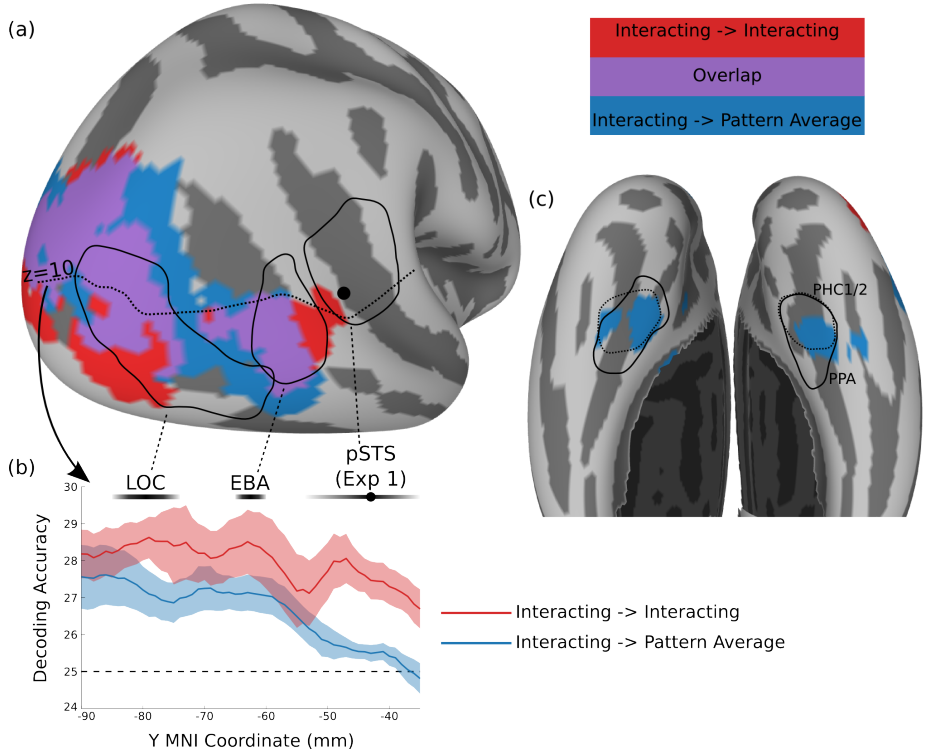


Figure 2.7: **MVPA cross-decoding searchlight for Experiment 2**. As in Figure 2.6, we identified voxels that could decode the action category of human-object interactions, and/or generalize this decoder to pattern averages. (a) A large swath of right lateral occipital and temporal regions (including LOC and EBA) can classify interaction timepoints, but in only some portions of LOC and EBA (superior LOC and posterior EBA) does this classifier generalize to pattern averages. (b) A z=10 slice of lateral cortex shows a clear difference between LOC/EBA and pSTS, with generalization to pattern averages much lower in pSTS. Error bars denote s.e.m. (c) We also found significant generalization to pattern averages within the retinotopic (PHC1/2) regions of PPA, indicating that this posterior subregion is somewhat insensitive to interactions.

information (Experiments 1 and 2) and showed a similarity to pattern-averaged responses only when pose was carefully controlled (Experiment 2). The most interesting decoding trends were observed in pSTS, which constructed representations of action categories that appear unrelated to object or pose information in isolation (Experiments 1 and 2). Overall, these results suggest that social cognition regions such as pSTS (possibly along with PCC, and IPL) represent human-object interaction categories using specialized features that are not present in the linear averages of human and object patterns, creating representations of human-object interactions that are more than the sum of their parts.

2.4.1 The role of EBA and pSTS

These results extend our current understanding of the role of EBA in action perception. It is well established that EBA represents body pose (reviewed in [81]). EBA, including the middle temporal gyrus (the most anterior portion of EBA, see [308]), has been implicated in action categorization through adaptation studies [148, 311], lesion studies [151] and a meta-analysis of object-related actions [56]. Exactly what type of information is represented in EBA has been more controversial, with proposals ranging from a cognitively unelaborated pose description [81] focused on "observable mechanics" [269] to an amodal hub for pairing gestures with semantic meaning [316]. The fact that noninteracting stimuli can be decoded above chance in Experiment 1 shows that EBA can discriminate based on object identity when the positioning of the human body is uninformative about the stimulus category, but the response to typical interactions appears to be primarily driven by body pose (Experiment 2). This fact that both object and pose information can be used by EBA raises the possibility that the representation in this region does represent more than simply body pose, though further work will be required to identify precisely how visual vs. semantic this representation is.

The pSTS (and adjacent TPJ) regions anterior to EBA have been associated with more abstract types of action perception, such as understanding unusual or deceptive human action [31, 117], recognizing whether an object is being grasped in a typical

way [322], and many other tasks involving perception of agency, theory of mind, and Gestalt integration [75, 129, 140, 223, 253, 254]. Interestingly, although pSTS shows little sensitivity to object identify or pose (Experiment 2), we found specialized representations for interacting stimuli here in both Experiments. Therefore pSTS appears to be much less related to visual features than EBA, and likely encodes more abstract semantic information about human actions and intentions. The PCC, another region identified in Experiment 2, may also be responsible for abstract action reasoning, as suggested by [269].

2.4.2 The neural basis of action recognition

There has been extensive prior work on the neural correlates of action perception, which is typically studied using video clips rather than controlled images (reviewed in [56, 74]). One controversy over the mechanism of action recognition is whether action recognition is carried out primarily in motor regions or in social reasoning areas. Under the simulation hypothesis, human actions are understood by mentally simulating the observed motor actions of the target and then inferring what the goals of the target must have been, a process presumed to be carried out in mirror neurons [35–37, 46, 47, 65, 240]. Under the teleological hypothesis, actions are understood by a more abstract social reasoning system, which does not depend on any mechanical "resonance" between the observer and target [31, 72, 126, 135]. Proponents of this view argue the activity seen in motor regions during action observation is involved in action prediction rather than action understanding [72, 177] and that the type of errors made by action observers is inconsistent with mirror simulation theories [255].

The social network proposed by [256] includes EBA, pSTS/TPJ, and PCC (in addition to medial frontal regions). Since our searchlight experiments show interaction effects almost exclusively in these regions (and show no effects in motor or premotor cortex), our results provide strong support for the view that action representations are built in social cognition regions, not in motor regions [309]. Additionally, our data reveal that social cognition regions process action stimuli even in the absence

of any social task, since our subjects were only performing one-back repetition detection. The only region outside the social network identified by our study is the right IPL, which has been previously linked with action perception but whose precise function is unclear. Some work has argued that this region contains mirror neurons due to its cross-adaptation properties [65] but the stimuli that activate this region do not activate macaque mirror neurons [135] and lesion studies suggest that IPL is involved in the spatial coding of object-related actions, but not actual semantic action understanding [151].

2.4.3 Comparison to object-object interaction studies

Previous work has attempted to link the perceptual grouping of interacting objects [112, 237, 241] with activity LOC, but the results have been controversial. Two studies have shown increased BOLD activity in LOC when objects are interacting [156] or positioned for interaction [242], while MVPA analyses have shown that the LOC response pattern for coherent scenes can be at least partially predicted as the average of responses to signature objects [183] and that the LOC response to pairs of action-oriented objects is similar to a linear combination of the two object responses [13]. One possibility is that a change in overall BOLD activity may not reflect a change in representation; for instance, BOLD differences could reflect the greater effort required to segment the objects when the objects overlap (as in [156]), or reduced competitive interaction among objects positioned to interact (as in [242]). Another proposal is that LOC is sensitive to lower-level nonaccidental visual relationships between objects or object parts, but not higher-level semantic interactions; notably, the observed difference in BOLD activity occurs regardless of whether the interaction is semantically meaningful, in both studies, suggesting that the interactions were based on low-level visual affordances of the objects (e.g. a wrench pointed toward a nut).

Our results suggest that both camps are correct. LOC did not show a decoding preference for interaction versus noninteracting categorization (Experiment 1) has interaction representations which are at least partially related to isolated object identity (Experiments 1 and 2), and does not incorporate human pose information

(Experiment 2), but does appear to encode some interaction information beyond object identity (Experiment 2), at least in the more inferior portion of LOC. Therefore it is possible that the representation in LOC is modulated in some way by interactions, while still being primarily driven by linear combinations of isolated object identity information.

Our finding of significant cross-decoding in PPA is surprising, in light of previous work showing that PPA generates scene representations that are not predictable from the constituent objects of the scene [183]. There are several possible ways of reconciling these results. One possibility is that PPAs global scene representation applies to full photographic images, but not two-element interactions, indicating that more complex stimuli are required to activate global processing in PPA (with more interactions, or more explicit 3D geometry). Alternatively, global representations may be generated only in the anterior portion of PPA, while the posterior PHC1/2 subregion of PPA accumulates local visual features in a way that is more similar to LOC [15].

2.4.4 Identifying configural processing

Our approach for identifying regions sensitive to a relationship between stimulus features is a general tool that could be used to investigate other types of configural processing. For example, placing walls and a floor together to form a 3D room likely evokes a novel representation in regions sensitive to scene layout and navigation. Our analysis suggests that these regions would exhibit a large cross-decoding penalty when training on rooms and testing on the average response to walls and floors. The regions responsible for processing relational or contextual interactions between objects [26] could be detected in a similar way, thereby avoiding the ambiguity of changes in mean BOLD activity. This approach for detecting configural representations gives neuroscientists a new way to locate the areas critical for creating our rich, complex experience of the visual world.

2.5 Acknowledgements

This work was funded by National Institutes of Health Grant 1 R01 EY019429 (to L.F.-F. and D.M.B.) and a National Science Foundation Graduate Research Fellowship under Grant No. DGE-0645962 (to C.B.). We thank the Richard M. Lucas Center for Imaging, the Center for Cognitive and Neurobiological Imaging, and the Stanford Vision lab for their comments and suggestions.

Chapter 3

Visual Scenes are Categorized by Function

How do we know that a kitchen is a kitchen by looking? Traditional models posit that scene categorization is achieved through recognizing features and objects, yet these models cannot account for mounting evidence that observers are relatively insensitive to the local image details. Although theoretical work has implicated scene function as a potential organizing principle, we have lacked the data to operationalize this idea. Using a large-scale experiment, we show that the activities afforded by a scene provide a fundamental categorization principle. Functions provided a good fit for human categorization patterns, outperforming alternative models based on objects or visual features. Moreover, nearly half of the explained variance was captured only by functions, implying that the predictive power of alternative models was due to their shared variance with the function-based model. These results challenge existing models of visual perception, providing immediately testable hypotheses for the functional organization of the visual system. This chapter is joint work with Michelle R. Greene, Andre Esteva, Diane M. Beck, and Fei-Fei Li, and an earlier version appeared as an arXiv preprint (1411.5340).

3.1 Introduction

"The question 'What makes things seem alike or different?' is one so fundamental to psychology that very few psychologists have been nave enough to ask it" [11].

Although more than half a century has passed since Attneave issued this challenge, we still have little understanding of how we categorize and conceptualize visual content. Traditionally, it has been assumed that scenes are categorized according to their component features and objects [25, 40, 190, 238, 270]. Mounting behavioral evidence, however, indicates that human observers have high sensitivity to the global meaning of an image [99, 114, 115, 226], and very little sensitivity to the local objects and features that are outside the focus of attention [236]. Consider the image of the kitchen in Figure 3.1. If scene categories are determined by objects, then we would expect the kitchen supply store (left) to be conceptually equivalent to the kitchen. Alternatively, if scenes are categorized from the similarity of spatial layout and surfaces [19, 216, 286], then observers might place the laundry room (center) into the same category as the kitchen. However, most of us share the intuition that the medieval kitchen (right) is in the same category, despite sharing few objects and features with the top image. Why is the image on the right a better category match to the modern kitchen than the other two?

Here we put forth the hypothesis that the conceptual structure of environments is driven primarily by the functions, or the actions that one could perform in the scene. We assert that representing a scene in terms of its high-level functions provides a better match to patterns human scene categorization than state-of-the-art models representing a scenes visual features or objects.

Figure 3.2 illustrates our approach. We constructed a large-scale scene category distance matrix by querying over 2,000 observers on over 63,000 images from 1055 scene categories (Figure 3.2A). We compared this human response pattern with an function-based similarity pattern created by asking hundreds of observers to indicate which of several hundred actions could take place in each scene (Figure 3.2B). We found a striking resemblance between function-based scene similarity and the human

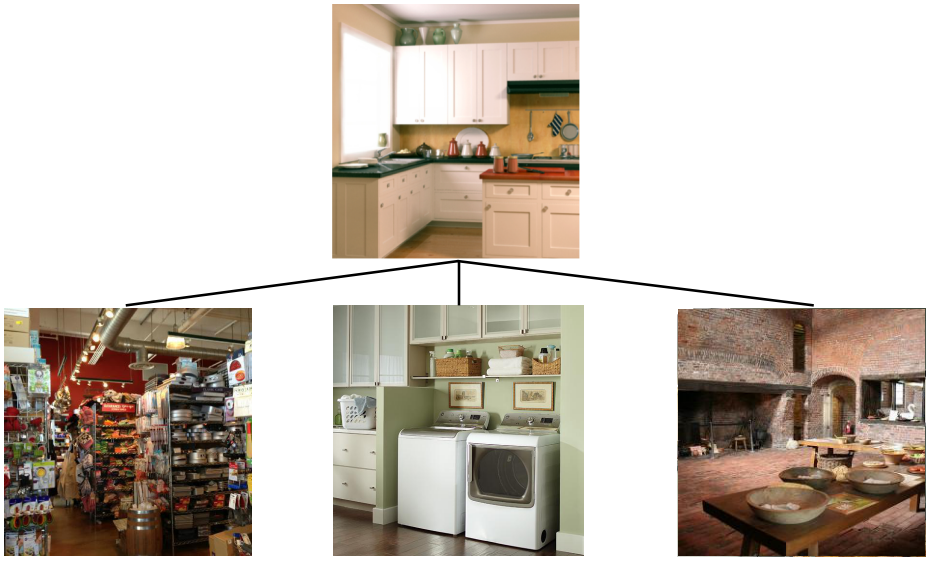


Figure 3.1: Which of the bottom images is in the same category as the kitchen image shown on top? Many influential models of visual perception would assume that scenes containing similar objects, such as the kitchen supply store (left), or similar layout, such as the laundry room (middle) would be placed into the same category by human observers. However, human observers tend to pick the medieval kitchen as the other category member despite having very different objects and features from the top kitchen.

similarity pattern. The function model not only explained more variance in the human category pattern than leading models of visual features and objects, but also contributed the most uniquely explained variance of any model, These results suggest that a scenes functions provide a fundamental coding scheme for human scene categorization.

3.2 Methods

3.2.1 Creating Human Scene Distance Matrix

Our aim was to amass a comprehensive collection of scene categories that have high human participant agreement about category membership. We started with 1,055 scene categories identified from the SUN and ImageNet databases [77, 315] and from literature review. These databases used the WordNet [196] hierarchy to identify potential scene concepts. We only included categories with at least 20 image exemplars, for a grand total of 63,988 images.

Human scene category distance was assessed using a large-scale online study using Amazons Mechanical Turk. Potential participants were recruited from a pool of trusted observers with at least 2,000 previously approved trials with at least 98% approval. Additionally, participants were required to pass a brief scene vocabulary test before participating.

We aimed to obtain at least 10 observations per pair of scene categories. In each trial, two images were presented side by side. Half of the image pairs came from the same putative scene category, while the other half were from two different categories that were randomly selected. Image exemplars were randomly selected within a category on each trial. Participants were instructed to indicate whether they would place the two images into the same category, and to type in the category name they would use for the left image (not analyzed, but used to assess understanding of the task). Workers were compensated $0.02 for each trial. We obtained 10 independent observations for each cell in the 1055 by 1055 scene matrix, for a total of over 5 million trials. Individual participants completed a median of 5 hits of this task (range: 1-36,497).

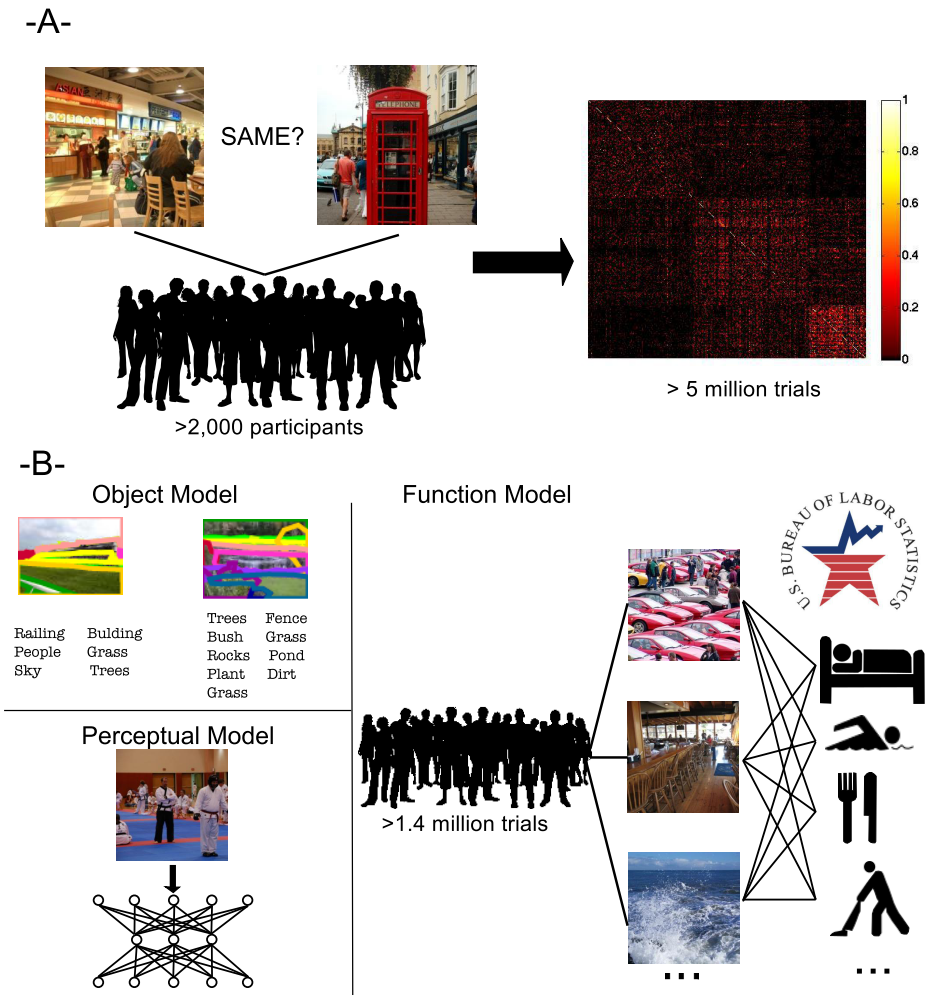


Figure 3.2: (A) We used a large-scale online experiment to generate a similarity matrix of scene categories. Over 2,000 individuals viewed more than 5 million trials in which participants viewed two images and indicated whether they would place the images into the same category. (B) Using the LabelMe tool [248] we examined the extent to which scene category similarity was related to scenes having similar objects. Our perceptual model used the output features of a state-of-the-art convolutional neural network [260] to examine the extent to which low-level visual features contribute to scene category. To generate the functional model, we took 227 actions from the American Time Use Survey. Using crowdsourcing, participants indicated which actions could be performed in which scene categories.

There was a median of 1,116 trials in each of the diagonal entries of the matrix, and a median of 11 trials in each cell of the off-diagonal entries.

From these data, we created a distance matrix in which each the distance between two scene categories was defined as the proportion of participants who indicated that the two categories were "different." From the 1,055 by 1,055 category distance matrix, we identified 311 categories with the strongest within-category cohesion (at least 70% of observers agreed that images were from the same category). Thus, the final dataset included 311 scene categories from 885,968 total trials, and from 2,296 individual workers.

3.2.2 Creating the Scene Function Space

In order to determine whether scene categories are governed by functional similarity, we needed a broad space of possible actions that could take place in our comprehensive set of scene categories. We gathered these actions from the lexicon of the American Time Use Survey (ATUS), a project sponsored by the US Bureau of Labor Statistics that uses U.S. census data to determine how people distribute their time across a number of activities. The lexicon used in this study was pilot tested over the course of three years [262], and therefore represents a complete set of goal-directed actions that people can engage in. The ATUS lexicon includes 428 specific activities organized into 17 major activity categories and 105 mid-level categories. The 227 actions included in our study included the most specific category levels with the following exceptions:

1. The superordinate category Caring for and Helping Non-household members was dropped as these actions would be visually identical to those in the Caring for and Helping Household members category.

2. In the ATUS lexicon, the superordinate-level category Work contained only two specific categories (primary and secondary jobs). Because different types of work can look very visually different, we expanded this category by adding 22 categories representing the major labor sectors from the Bureau of Labor Statistics.

3. The superordinate-level category Telephone calls was collapsed into one action because we reasoned that all telephone calls would look visually similar.

4. The superordinate-level category Traveling was similarly collapsed into one category because being in transit to go to school (for example) should be visually indistinguishable from being in transit to go to the doctor.

5. All instances of Security procedures have been unified under one category for similar reasons.

6. All instances of Waiting have been unified under one category.

7. All Not otherwise specified categories have been removed.

The final list of actions can be found in the Supplemental Materials.

3.2.3 Norming the Function Space

Using a separate large-scale online experiment, 484 workers indicated which of the 227 actions could take place in each of the 311 scene categories. Participants were screened using the same criterion described above. In each trial, a participant saw a randomly selected exemplar image of one scene category along with a random selection of 17 or 18 of the 227 actions. Each action was hyperlinked to its description in the ATUS lexicon. Participants were instructed to use check boxes to indicate which of the actions would typically be done in the type of scene shown.

Each individual participant did a median of 9 trials (range: 1-4,868). Each scene category action pair was rated by a median of 16 participants (range: 4-86), for a total of 1.4 million trials.

We created a 311-category by 227-matrix in which each cell represents the proportion of participants indicating that the action could take place in the scene category. Since scene categories varied widely in the number of actions they afford, we obtained a distance matrix by computing the cosine distance between categories. This measures the overlap between actions while being invariant to the absolute magnitude of the action vector.

3.2.4 Function Space MDS Analysis

To better understand the scene function space, we performed a classical multidimensional scaling (MDS) decomposition of the action distance matrix. This yielded an embedding of the scene categories such that inner products in this embedding space approximate the (double-centered) distances between scene categories, with the embedding dimensions ranked in order of importance [39]. In order to associate actions with each of these dimensions, we computed the correlation coefficient between each action (across scene categories) with the category coordinates for a given dimension.

3.2.5 Alternative Models

To put the performance of the function-based model in perspective, we compared it to eight alternative models. Five of the models represented visual features, and one model examined the objects that were present in the scenes. These models yielded scene category by feature matrices, and were converted to distance matrices using cosine distance. Additionally, two models measured distances directly, based either on the lexical distance between scene category names, or simply by whether scenes belonged to the same superordinate level category (indoor, urban or natural). We will detail each of the models below.

3.2.5.1 Perceptual Models

Convolutional Neural Network

We generated a perceptual feature vector using the publicly distributed OverFeat convolutional neural network (CNN) [260], which was trained on the ImageNet 2012 training set [77]. This 7-layer CNN takes an image of size 231x231 as input, and produces a vector of 4096 image features that are optimized for 1000-way object classification. This network achieves top-5 object recognition on ImageNet 2012 with approximately 16% error, meaning that the correct object is one of the models first five guesses in 84% of trials. Using the top layer of features, we averaged the features for all images in each scene category to create a 311-category by 4096-feature matrix.

Gist

We used the Gist descriptor features of [216]. This popular model for scene recognition provides a summary statistic representation of the dominant orientations and spatial frequencies at multiple scales coarsely localized on the image plane. We used spatial bins at 4 cycles per image and 8 orientations at each of 4 spatial scales for a total of 3,072 filter outputs per image. We averaged the gist descriptors for each image in each of the 311 categories to come up with a single 3,072-dimensional descriptor per category.

Color histograms

We represented color using LAB color space. For each image, we created a two-dimensional histogram of the a* and b* channels using 50 bins per channel. We then averaged these histograms over each exemplar in each category, such that each category was represented as a 2500 length vector representing the averaged colors for images in that category. The number of bins was chosen to be similar to those used in previous scene perception literature [215].

Tiny Images

Torralba and colleagues [286] demonstrated that human scene perception is robust to aggressive image downsampling, and that an image descriptor representing pixel values from such downsampled images could yield good results in scene classification. Here, we downsampled each image to 32 by 32 pixels (grayscale). We created our 311-category by 1024 feature matrix by averaging the downsampled exemplars of each category together.

Wavelets

We represented each image in this database as the output of a bank of multi-scale Gabor filters. This type of representation has been used to successfully model the representation in early visual areas [153]. Each image was converted to grayscale, down sampled to 128 by 128 pixels, and represented with a bank of Gabor filters at three spatial scales (3, 6 and 11 cycles per image with a luminance-only wavelet

that covers the entire image), four orientations (0, 45, 90 and 135 degrees) and two quadrature phases (0 and 90 degrees). An isotropic Gaussian mask was used for each wavelet, with its size relative to spatial frequency such that each wavelet has a spatial frequency bandwidth of 1 octave and an orientation bandwidth of 41 degrees. Wavelets were truncated to lie within the borders of the image. Thus, each image is represented by 3*3*2*4+6*6*2*4+11*11*2*4 = 1328 total Gabor wavelets. We created the feature matrix by averaging the Gabor weights over each exemplar in each category.

3.2.5.2 Object-based Model

In order to model the similarity of objects within scene categories, we employed the LabelMe tool [248] that allows users to outline and annotate each object in each image by hand. 7,710 scenes from our categories were already labeled in the SUN 2012 release [315], and we augmented this set by labeling an additional 223 images. There were a total of 3,563 unique objects in this set. Our feature matrix consisted of the proportion of scene images in each category containing a particular object. For example, if 10 out of 100 kitchen scenes contained a blender, the entry for kitchen-blender would be 0.10. In order to estimate how many labeled images we would need to robustly represent a scene category, we performed a bootstrap analysis in which we resampled the images in each category with replacement (giving the same number of images per category as in the original analysis), and then measured the variance in distance between categories. With the addition of our extra images, we ensured that all image categories either had at least 10 fully labeled images or had mean standard deviation in distance to all other categories of less than 0.05 (e.g. less than 5% of the maximal distance value of 1).

3.2.5.3 Semantic Models

We examined semantic similarity by examining the shortest path between category names in the WordNet tree using the Wordnet::Similarity implementation of [222]. The similarity matrix was normalized and converted into distance. We examined each

of the metrics of semantic relatedness implemented in Wordnet::Similarity and found that this path measure was the best correlated with human performance.

3.2.5.4 Superordinate-Category Model

As a baseline model, we examined how well a model that groups scenes only according to superordinate-level category would predict human scene similarity assessment. We assigned each of the 311 scene categories to one of three groups (natural outdoors, urban outdoors or indoor scenes). Then, each pair of scene categories in the same group was given a distance of 0 while pairs of categories in different groups were given a distance of 1.

3.2.6 Noise Ceiling

The variability of human categorization responses puts a limit on the maximum correlation expected by any of the tested models. In order to get an estimate of this maximum correlation, we used a bootstrap analysis in which we sampled with replacement observations from our scene categorization dataset to create two new datasets of the same size as our original dataset. We the correlated these two datasets to one another, and repeated this process 1000 times.

3.2.7 Hierarchical Regression Analysis

In order to understand the unique variance contributed by each of our feature spaces, we used hierarchical linear regression analysis, using each of the feature spaces both alone and in combination to predict the human similarity response pattern. In total, eight regression models were used: (1) all nine feature spaces used together; (2) the top 3 performing features together (functions, objects and the perceptual CNN); (3-5) each of the top three features alone; (6-8) each pair of the top three features. By comparing the r^2 values of a feature space used alone to the r^2 values of that space in conjunction with another feature space, we can infer the amount of variance that is independently explained by that feature space. In order to visualize this information in an Euler diagram, we used EulerAPE software [195].

3.3 Results

3.3.1 Human Scene Category Distance

To assess the conceptual structure of scene environments, we asked over 2,000 human observers to categorize images belonging to 311 scene categories in a large-scale online experiment. The resulting 311 by 311 category distance matrix is shown in Figure 3.3. In order to better visualize the category structure, we have ordered the scenes using the optimal leaf ordering for hierarchical clustering [21] allowing us to see what data-driven clusters emerge.

Several category clusters are visible. Some clusters appear to group several subordinate-level categories into a single entry-level concept, such as bamboo forest, woodland and rainforest being examples of forests. Other clusters seem to reflect broad classes of activities (such as sports) which are visually heterogeneous and cross other previously defined scene boundaries, such as indoor-outdoor [99, 132, 277, 287] or the size of the space [114, 216, 219]. Such activity-oriented clusters hint that the actions that one can perform in a scene (the scenes functions) could provide a fundamental grouping principle for scene category structure.

3.3.2 Function-based Similarity Best Correlates with Human Category Structure

For each of our nine feature spaces, we created a distance vector representing the distance between each pair of scene categories. We then correlated this distance vector with the human distance vector from the previously described experiment.

In order to quantify the performance of each of our models, we defined a noise ceiling based on the inter-observer reliability in the human scene distance matrix. This provides an estimate of the explainable variance in the scene categorization data, and thus provides an upper bound on the performance of any of our models. Using bootstrap sampling (see Methods), we found an inter-observer correlation of r=0.76. In other words, we cannot expect a correlation with any model to exceed this value.

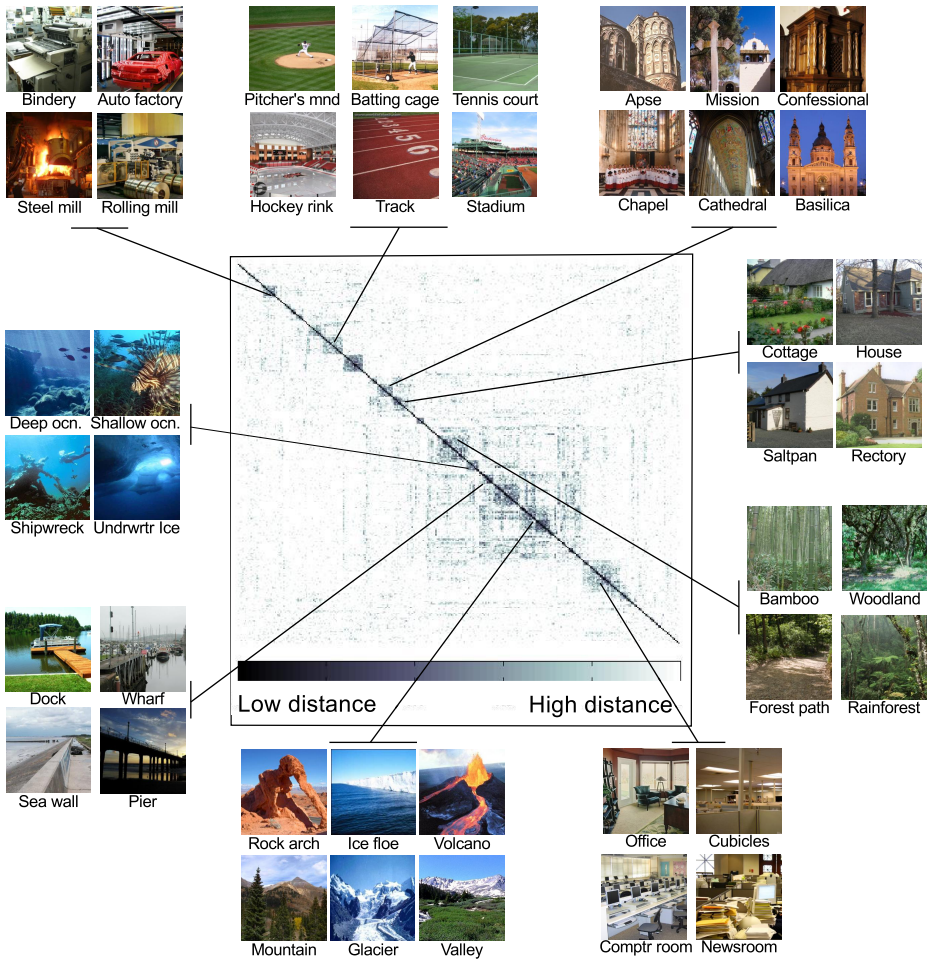


Figure 3.3: The human category distance matrix from our large-scale online experiment was found to be sparse. Over 2,000 individual observers categorized images in 311 scene categories. We visualized the structure of this data using optimal leaf ordering for hierarchical clustering, and show representative images from categories in each cluster.

Function-based similarity had the highest resemblance to the human similarity pattern (r=0.50). This represents about 2/3 of the maximum observable correlation obtained from the noise ceiling. As shown in Figure 3.4, this correlation is substantially higher than any of the alternative models we tested.

Of course, being able to perform similar actions often means manipulating similar objects, and scenes with similar objects are likely to share visual features. Therefore, we compared function-based categorization patterns to alternative models based on perceptual features, object-based similarity, and the semantic similarity of category names.

We tested five different models based on purely visual features. The most sophisticated used the top-level features of a state-of-the-art convolutional neural network model (CNN) [260] trained on the ImageNet database [77]. These features, computed by iteratively applying learned nonlinear filters to the image, have been shown to be a powerful image representation for a wide variety of visual tasks [234]. Category distances in CNN space produced a correlation with human distance of r=0.39. Simpler visual features, however, such as gist [216], color histograms [215], Tiny Images [286], and wavelets [153] had low correlations with human scene distance.

Category structure could also be predicted to some extent based on the similarity between the objects present in scene images (r=0.33, using human-labeled objects from the LabelMe database [248]), or the semantic distance between category names in the WordNet tree (r=0.27) [141, 196, 222]. Surprisingly, a model that merely groups scenes by superordinate-level categories (indoor, urban or natural environments) also had a substantial correlation (r=0.25) with human distance patterns.

Although each of these feature spaces had differing dimensionalities, this pattern of results also holds if the number of dimensions is equalized through dimensionality reduction (see Methods and Supplementary Figure B2).

3.3.3 Independent Contributions from Alternative Models

To what extent does function-based similarity uniquely explain the patterns of human scene categorization? Although function-based similarity was the best explanation

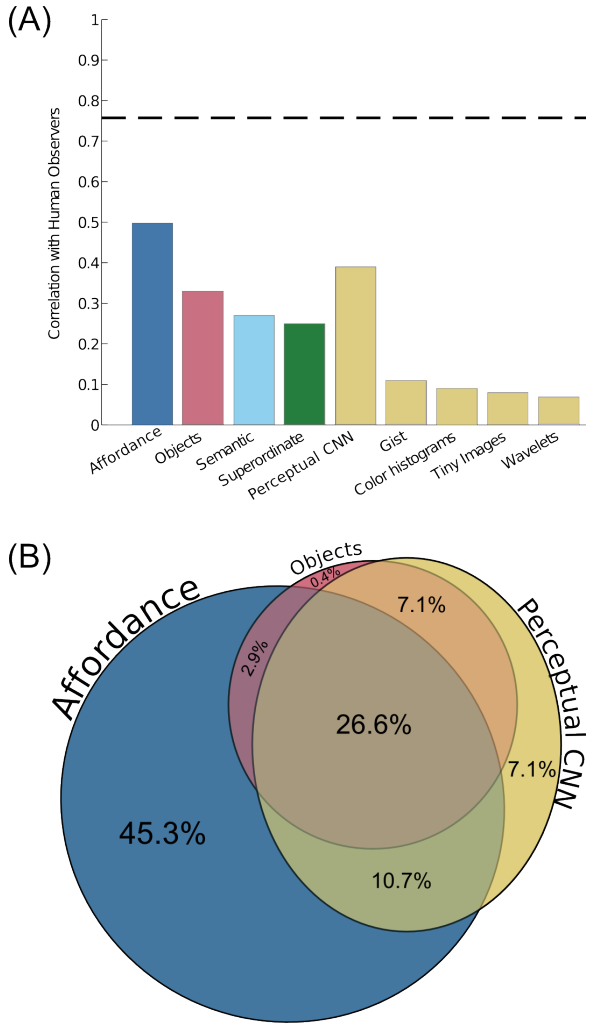


Figure 3.4: (A) Correlation of all models with human scene categorization pattern. Function-based similarity (dark blue, left) showed the highest resemblance to human behavior, achieving 2/3 of the maximum explainable similarity (black dotted line). Of the models based on visual features (yellow, right), only the model using the top-level features of the convolutional neural network (CNN) showed substantial resemblance to human data. Object-based similarity, semantic similarity and superordinate-level similarity all showed moderate correlations. (B) Euler diagram showing the distribution of explained variance for the three top-performing models. Function-based similarity independently explained 13.2% of the variance in the human similarity pattern (45% of total variance explained by all models). By contrast, perceptual similarity independently accounted for only 2% of the variance (7% of explained variance) and object-based similarity only accounted for 0.11% of the variance (0.4% of the explained variance).

of the human categorization pattern of the models we tested, perceptual and object-based similarities also had sizeable correlations with human behavior. To what extent do these models make the same predictions?

In order to assess the independent contributions made by each of the models, we used a hierarchical linear regression analysis in which each of the three top-performing models was used either separately or in combination to predict the human similarity pattern. By comparing the r^2 values from the individual models to the r^2 values for the combined model, we can assess the unique variance explained by each descriptor. A combined model with all nine features explained 29.8% of the variance in the human similarity pattern (r=0.55). This model is driven almost entirely by the top three feature spaces (functions, perceptual CNN, and object labels), which explained a combined 29.1% of the variance (r=0.54). Note that affordances explained 85.6% of this explained variance, indicating that the object and perceptual features only added a small amount of independent information (14.4% of the combined variance).

Although there was a sizable overlap between the portions of the variance explained by each of the models (see Figure 3.4B), nearly half of the total variance explained can be attributed only to functions (13.2% of total variance, or 45.3% of the explained variance), and was not shared by the other two models. In contrast, the independent variance explained by perceptual similarity and object-based similarity accounted for only 2% (7% of explained variance) and 0.11% (0.4% of explained variance) of the total variance respectively. Therefore, the contributions of perceptual and object-based similarities are largely shared with function-based similarity, further highlighting the utility of affordances for explaining human scene similarity patterns.

3.3.4 Examining Scene Function Space

In order to better understand the function space, we performed classical multi-dimensional scaling on the function distance matrix, allowing us to identify how patterns of functions contribute to the overall similarity pattern. We found that at least 10 MDS dimensions were necessary to explain 95% of the variance in the function distance matrix, suggesting that the efficacy of the function-based model

was driven by a number of distinct function dimensions. We examined the projection of categories onto the first three MDS dimensions. As shown in Figure 3.5, the first dimension appears to separate indoor locations that have a high potential for social interactions (such as socializing and attending meetings for personal interest) from outdoor spaces that afford more solitary activities, such as hiking and science work. The second dimension separates work from leisure. Later dimensions appear to separate environments related to transportation and industrial workspaces from restaurants, farming, and other food-related environments (see Supplementary Figure B1).

3.4 Discussion

We have shown that human scene categorization is better explained by the action possibilities, or functions, of a scene than by the scenes visual features or objects. Furthermore, function-based similarity explained far more independent variance than did alternative models, as these models were correlated with human category patterns only insofar as they were also correlated with the scenes functions. This suggests that a scenes functions contain essential information for categorization that is not captured by the scenes objects or low-level visual features.

The current results cannot be explained by the smaller dimensionality of the function-based features, as further analysis revealed that function-based features outperformed other similarity spaces using equivalent numbers of dimensions. Furthermore, this pattern was observed over a wide range of dimensions, suggesting that each function feature contained more information about scene categories than each perceptual or object-based feature.

The idea that the function of vision is for action has permeated the literature of visual perception, but it has been difficult to fully operationalize this idea for testing. Psychologists have long theorized that rapid and accurate environmental perception could be achieved by the explicit coding of an environments affordances, most notably in J.J. Gibsons influential theory of ecological perception [102]. This work is most often associated with the direct perception of affordances that reflect relatively simple

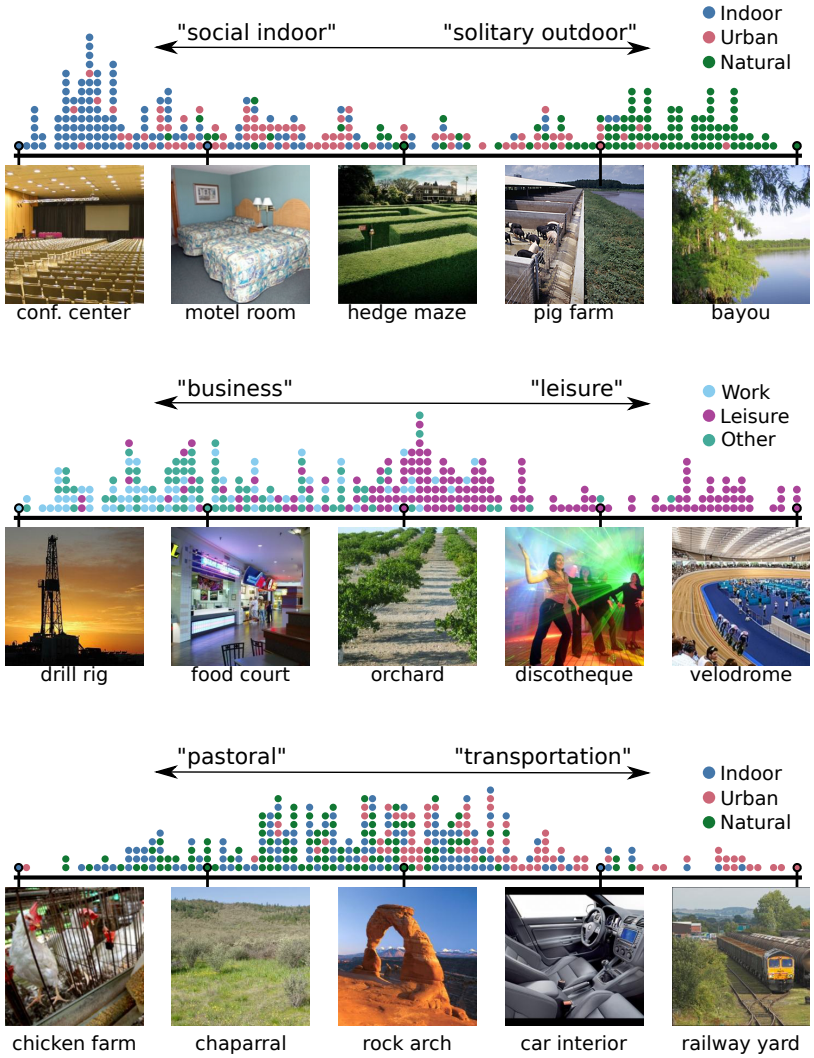


Figure 3.5: (Top): Distribution of superordinate-level scene categories along the first MDS dimension of the function distance matrix, which separates indoor scenes from natural scenes. Actions that were positively correlated with this component tend to be outdoor-related activities such as hiking while negatively correlated actions tend to reflect social activities such as eating and drinking. (Middle) The second dimension seems to distinguish environments for work from environments for leisure. Actions such as playing games are positively correlated while actions such as construction and extraction work are negatively correlated (Bottom). The third dimension distinguishes environments related to farming and food production (pastoral) from industrial scenes specifically related to transportation. Actions such as travel and vehicle repair are highly correlated with this dimension, while actions such as farming and food preparation are most negatively correlated.

motor patterns such as sitting or throwing. As the functions used in the current work often reflect higher-level, goal-directed actions, and because we are making no specific claims about the direct perception of these functions, we have opted not to use the term affordances here. Nonetheless, ideas from Gibsons ecological perception theory have inspired this work, and thus we consider our functions as conceptual extensions of Gibsons idea.

Previous small-scale studies have found that environmental functions such as navigability are reflected in patterns of human categorization [113, 114] and are perceived very rapidly from images [115].Our current results provide the first comprehensive, data-driven test of this hypothesis, using data from hundreds of scene categories and affordances. By leveraging the power of crowdsourcing, we were able to obtain both a large-scale similarity structure for visual scenes, but also normative ratings of functions for these scenes. Using hundreds of categories, thousands of observers and millions of observations, crowdsourcing allowed a scale of research previously unattainable. Previous research on scene function has also suffered from the lack of a comprehensive list of functions, relying instead on the free responses of human observers describing the actions that could be taken in scenes [114, 221]. By using an already comprehensive set of actions from the American Time Use Survey, we were able to see the full power of functions for predicting human categorization patterns.

Given the relatively large proportion of variance independently explained by function-based similarity, we are left with the question of why this model outperforms the more classic models. By examining patterns of variance in the function by category matrix, we found that functions can be used to separate scenes along previously defined dimensions of scene variance, such as superordinate-level category [147, 181, 287], and between work and leisure activities [83]. Although the variance explained by function-based similarity does not come directly from visual features or the scenes objects, human observers must be able to apprehend these functions from the image somehow. It is therefore a question open for future work to understand the extent to which human observers bring non-visual knowledge to bear on this problem.

Some recent work has examined large-scale neural selectivity based on semantic similarity [141] or object-based similarity [270], finding that both types of conceptual

structures can be found in the large-scale organization of human cortex. Our current work indeed shows sizeable correlations between these types of similarity structures and human behavioral similarity. However, we find that function-based similarity is a better predictor of behavior and may provide an even stronger grouping principle in the brain.

These results challenge many existing models of visual categorization that consider categories to be purely a function of shared visual features or objects. Just as the Aristotelian theory of concepts assumed that categories could be defined in terms of necessary and sufficient features, classical models of visual categorization have assumed that a scene category can be explained by necessary and sufficient objects [25, 270] or diagnostic visual features [235, 297]. However, just as the classical theory of concepts cannot account for important cognitive phenomena, the classical theory of scene categories cannot account for the fact that two scenes can share a category even when they do not share many features or objects. By contrast, the current results demonstrate that the possibility for action creates categories of environmental scenes. In other words, a kitchen is a kitchen because it is a space that affords cooking, not because it shares objects or other visual features with other kitchens.

3.5 Acknowledgements

Funding was provided by a National Science Foundation Graduate Research Fellowship under grant number DGE-0645962, National Institutes of Health Grant 1 R01 EY019429, National Institutes of Health F32 EY19815, and Office of Naval Research Multidisciplinary University Research Initiative grant number N000141410671.

Chapter 4

Spatially-regularized voxel-level connectivity

Discovering functional connectivity between and within brain regions is a key concern in neuroscience. Due to the noise inherent in fMRI data, it is challenging to characterize the properties of individual voxels, and current methods are unable to flexibly analyze voxel-level connectivity differences. We propose a new functional connectivity method which incorporates a spatial smoothness constraint using regularized optimization, enabling the discovery of voxel-level interactions between brain regions from the small datasets characteristic of fMRI experiments. We validate our method in two separate experiments, demonstrating that we can learn coherent connectivity maps that are consistent with known results. First, we examine the functional connectivity between early visual areas V1 and VP, confirming that this connectivity structure preserves retinotopic mapping. Then, we show that two category-selective regions in ventral cortex - the Parahippocampal Place Area (PPA) and the Fusiform Face Area (FFA) - exhibit an expected peripheral versus foveal bias in their connectivity with visual area hV4. These results show that our approach is powerful, widely applicable, and capable of uncovering complex connectivity patterns with only a small amount of input data.

We then present a method for identifying fine-grained functional connectivity between any two brain regions by simultaneously learning voxel-level connectivity maps

over both regions. We show how to formulate this problem as a constrained least-squares optimization, which can be solved using a trust region approach. Our method can automatically discover multiple correspondences between distinct voxel clusters in the two regions, even when these clusters have correlated timecourses. We validate our method by identifying a known division in the lateral occipital complex using only functional connectivity.

This chapter is joint work with Marius Cătălin Iordan, Diane M. Beck, and Fei-Fei Li, and portions have previously appeared in print in NeuroImage [18] and in the conference proceedings of the 2nd NIPS Workshop on Machine Learning and Interpretation in Neuroimaging [17].

4.1 Introduction

Functional Magnetic Resonance Imaging (fMRI) has been widely adopted by the neuroscience community primarily because it allows researchers to unobtrusively sample activity patterns from populations of neurons across the entire human brain, at a fine spatial scale (typically a few millimeters). However, many methods for identifying distributed functional networks underutilize the spatial resolution of fMRI, considering only the aggregate properties of groups of voxels. For example, when computing functional connectivity between brain regions, activity is often spatially averaged within each Region of Interest (ROI) and simple statistical relationships (e.g. correlation) between these mean timecourses are used as measures of connectivity between the regions (reviewed in [243]).

ROIs are generally defined by a contrast between two types of stimuli, constrained by rough anatomical location. However, there is no reason to assume that all voxels within an ROI have identical functional properties. Indeed, recent work has achieved some success in dividing existing ROIs into functional subregions. For example, lateral occipital complex (LOC, defined in [184]) has been shown to contain two functionally distinct subregions [118], and the extrastriate body area (EBA, defined in [80]) has been split into three separate limb-sensitive areas [308].

Recent work has begun to investigate intra-ROI structure using measures of functional connectivity. These methods have provided evidence of subdivisions within regions such as the thalamus [323], medial frontal cortex [157], the amygdala [244], anterior cingulate cortex [186], and the precuneus [187], and have been used to uncover the functional connectivity structure of early visual cortex [130].

However, these methods are unable to jointly model the functional connectivity properties of individual voxels for typical fMRI dataset sizes. Almost all current methods avoid simultaneously learning the connectivity properties for all voxels, by spatially downsampling to a small number of subregions [186, 244], only learning parameters for one voxel or subregion at a time [61, 66, 157, 323], or both [187]. Each of these approaches has some disadvantages. Downsampling requires prior knowledge of the anatomical subdivisions in a region [244] or of the relevant spatial scale of connectivity differences [186], making it ill-suited for exploratory studies. Learning voxel parameters separately can make comparisons between voxels difficult; for example, if two voxels are assigned different levels of connectivity with a seed region, there is generally no way to tell whether these two voxels predict different parts of the seed timecourse, or if one voxel is simply a noisy copy of the other. Jointly learning connectivity weights allows us to pinpoint those voxels that contribute unique information about the seed region, by simultaneously considering the timecourses of all voxels.

Support vector regression (SVR) can learn joint voxel-level connectivity maps, but requires a significant amount of data; for example, [130] uses more than 40 minutes of training data (1,600 timepoints) to learn connectivity structures in early visual areas. Scarcity of training data is a common obstacle for characterizing individual voxels in fMRI experiments. Typical fMRI datasets record activity from tens of thousands of voxels in the human brain, but with only about a thousand timepoints per voxel. Several methods have been successfully implemented to boost the number of recorded timepoints (e.g. rapidly scanning only a select portion of the brain, [30, 258]), but all fMRI studies must contend with a severe data shortage for individual subjects caused by this limitation. A recent survey of MVPA techniques [197] has demonstrated empirically that low-complexity models tend to perform better at decoding information

from patterns of activity than high-complexity models, which is theoretically plausible given the limited number of timepoints available for model training.

Therefore, there is still a need for a method that can estimate voxel-level connectivity structure with data set sizes more typical of fMRI experiments. For example, when investigating stimulus-category-dependent changes in connectivity patterns, the amount of data for each category can be on the order of only a hundred timepoints. To address this issue, we propose a spatially regularized method for examining connectivity differences within ROIs, which is specifically tailored to small training sets typical in the fMRI setting. Our regularization approach simply imposes the constraint that connectivity properties should vary smoothly across voxels, a highly plausible assumption given the nature of fMRI data. Much prior work has been dedicated to incorporating spatial regularization into MRI and fMRI analysis, with goals such as functional classification and regression [119, 212], classification of gray matter concentration maps [73], and inter-subject alignment [67]. However, none of these regularized models are specifically searching for evidence of voxel-level structure within an individual ROI.

In this paper, we present a spatially regularized method for uncovering connectivity differences within ROIs, and demonstrate that it is possible to discover consistent structures using only a small amount of training data. We validate our approach using two different experiments, for which the ground truth connectivity is already known. In the first experiment, we show that we can recover retinotopic connectivity patterns between early visual areas V1 and VP. In the second, we replicate the known eccentricity biases in the connectivity between visual area hV4 and both the Parahippocampal Place Area (PPA) and the Fusiform Face Area (FFA), without using a specialized experimental design.

4.2 Materials and Methods

4.2.1 Traditional Connectivity Analysis

The simplest way to characterize functional connectivity between two ROIs is to extract mean timecourses by spatially averaging over all the voxels in each ROI, then computing the Pearson product-moment correlation coefficient (r value) between the two mean timecourses. A high r^2 value indicates strong functional connectivity between the pair of ROIs.

We can reformulate this analysis as a linear regression problem in which we use voxel activation values from the first timecourse to predict the second timecourse. Specifically, we choose a slope a and an offset b minimizing

$$||(a \cdot \text{mean}_v(\boldsymbol{A^1}) + b \cdot \mathbb{1}) - \text{mean}_v(\boldsymbol{A^2})||_2^2 \tag{4.1}$$

where $\boldsymbol{A^1}$ and $\boldsymbol{A^2}$ are the (# voxels x # timepoints) data matrices from two ROIs, and mean_v denotes an average across voxels. The r^2 value is then equivalent to the fraction of variance explained (the increase in prediction accuracy from using a and b, as opposed to just predicting the mean of the second timecourse, [272]):

$$\begin{aligned} r^2 &= \text{Fraction of Variance Explained} \\ &= 1 - \frac{||(a \cdot \text{mean}_v(\boldsymbol{A^1}) + b \cdot \mathbb{1}) - \text{mean}_v(\boldsymbol{A^2})||_2^2}{||(\text{mean}_t(\text{mean}_v(\boldsymbol{A^2})) - \text{mean}_v(\boldsymbol{A^2})||_2^2} \end{aligned}$$

where mean_t denotes an average across time.

We can interpret $a \cdot \text{mean}_{v_1}(\boldsymbol{A^1})$ as a weighted sum, in which every voxel shares the same weight $c = a/(\#\text{ of voxels in } \mathrm{A}^1)$. This allows us to rewrite the traditional correlation method as an optimization problem in a more general form:

$$\begin{aligned} &\underset{\boldsymbol{a},c,b}{\text{minimize}} && ||(\boldsymbol{a}^T \cdot \boldsymbol{A^1} + b) - \text{mean}_v(\boldsymbol{A^2})||_2^2 \\ &\text{subject to} && \boldsymbol{a} = c \cdot \mathbb{1} \end{aligned} \tag{4.2}$$

where $\boldsymbol{a}$ is a vector with length equal to the number of voxels in A^1.

This is a convex optimization problem, and can be solved using a standard optimization package (all optimization problems in our paper are solved using `CVX`, a package for specifying and solving convex programs, [111]).

4.2.2 Regularized Connectivity Method

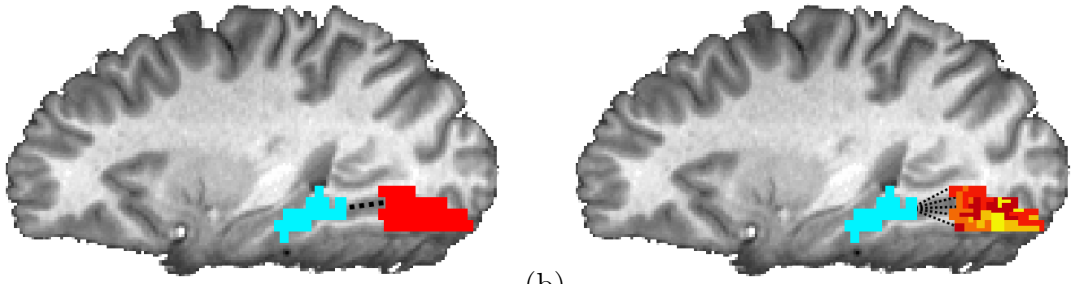
Although the basic connectivity method described in Section 4.2.1 provides valuable insight into the functional organization of the human brain, it lacks a principled way to take into account voxel-level spatial information present in the fMRI signal. However, simply removing the constraint that all voxels must have the same weight leads to severe overfitting on typical fMRI dataset, as will be demonstrated in sections 4.3.1 and 4.3.2. Rather than revealing interesting, generalizable connectivity patterns, the learned maps are driven mainly by noise in the training data and fail to replicate across runs. In order to obtain meaningful weight maps, we must place a constraint on the voxel weights which is less restrictive than that of the traditional method (all weights equal), but more restrictive than the unconstrained method (all weights independent).

One plausible assumption is that voxel connectivity properties are likely to be spatially correlated, with nearby voxels typically having more similar connectivity properties than spatially distant voxels. This reflects a common view of cortical organization, and is especially applicable to blood-oxygen-level dependent (BOLD) signals such as fMRI, since the hemodynamic response is spatially smooth.

To incorporate this assumption, we developed a new method of assessing functional connectivity patterns within ROIs (Fig. 4.1). We define an extension of the original optimization problem (Eq. 4.2), replacing the constraint that weights for all voxels must be equal with a spatial regularization term in the minimization objective:

$$\underset{\boldsymbol{a},b}{\text{minimize}} \quad ||(\boldsymbol{a}^T \cdot \boldsymbol{A^1} + b \cdot \mathbb{1}) - \text{mean}_v(\boldsymbol{A^2})||_2^2 + \lambda||\boldsymbol{D} \cdot \boldsymbol{a}||_2^2 \tag{4.3}$$

$\boldsymbol{D}$ is the *voxel connectivity matrix*, which we design to penalize the mean squared difference between the weight a_i of voxel i, and the weights of voxel i's neighbors. Each row of $\boldsymbol{D}$ represents a directed edge from a voxel i to an adjacent voxel j: all



(a) (b)

Figure 4.1: **Comparison of connectivity maps learned from traditional (a) and regularized (b) methods.** (a) In traditional functional connectivity analysis, connectivity with a seed region (blue) is assumed to be identical for all voxels in an ROI (red). (b) Our method can learn a map of weights in an ROI that describes the voxel-level connectivity between each voxel and the seed region. It is possible to learn these maps using a small amount of training data by imposing a spatial smoothness constraint.

entries in a row are zero, except for the jth element (equal to $1/\sqrt{d_i}$) and the kth element (equal to $-1/\sqrt{d_i}$), where d_i is the number of neighbors of voxel i. Thus the regularization term is $||\boldsymbol{D} \cdot \boldsymbol{a}||_2^2 = \sum_{i=1}^{N} \frac{1}{d_i} \sum_{j \in n_i} (a_i - a_j)^2$ where N is the number of voxels in $\boldsymbol{A^1}$ and n_i is the set of i's neighbors. The hyperparameter λ controls the strength of the regularization, trading off between an $\boldsymbol{a}$ that gives a good prediction of the seed timecourse $\boldsymbol{A^2}$ and an $\boldsymbol{a}$ that is spatially smooth. λ can take on any positive value, with $\lambda \to 0$ producing completely unregularized maps, and $\lambda \to \infty$ producing completely smooth (constant) maps.

In this paper, we define the voxel neighborhoods n_i to enforce smoothness along the cortical surface. After mapping an ROI onto a cortical flat map, we define the neighborhood of each voxel to be its k-nearest neighbors. This approach is suitable for ROIs that are known to have retinotopic structure on the cortical surface, such as early visual areas. Alternatively, a more general approach could simply define n_i to be all spatially adjacent voxels (touching voxel i at least on a corner in the three-dimensional representation of the particular subject's brain).

As in the traditional method, this optimization problem is convex and therefore has a global optimum that can be found efficiently. It is in fact possible to compute

the solution in closed form. If we assume that the timecourses have been normalized to have zero mean (such that the optimal $b = 0$ and can be disregarded), the minimization objective is

$$\begin{aligned}&(\boldsymbol{a}^T \cdot \boldsymbol{A^1} - \text{mean}_v(\boldsymbol{A^2}))(\boldsymbol{a}^T \cdot \boldsymbol{A^1} - \text{mean}_v(\boldsymbol{A^2}))^T + \lambda(\boldsymbol{D} \cdot \boldsymbol{a})(\boldsymbol{D} \cdot \boldsymbol{a})^T = \\ &\boldsymbol{a}^T(\boldsymbol{A^1}\boldsymbol{A^1}^T + \lambda \boldsymbol{D}\boldsymbol{D}^T)\boldsymbol{a} - \boldsymbol{a}\boldsymbol{A^1}\text{mean}_v(\boldsymbol{A^2})\end{aligned} \tag{4.4}$$

The quadratic term is positive definite for $\lambda > 0$, so this has the unique solution

$$\boldsymbol{a} = (\boldsymbol{A^1}\boldsymbol{A^1}^T + \lambda \boldsymbol{D}\boldsymbol{D}^T)^{-1}\boldsymbol{A^1}\text{mean}_v(\boldsymbol{A^2}) \tag{4.5}$$

4.2.3 Datasets

4.2.3.1 Human subjects

We tested our functional connectivity method on two separate datasets. Both experiments were approved by the Institutional Review Board of Stanford University, and all subjects gave their written informed consent. Subjects were in good health with no past history of psychiatric or neurological diseases, and had normal or corrected-to-normal vision. 13 subjects (1 female; age: 22-26 years; including one of the authors) participated in the first experiment, and 8 subjects (2 female; age: 23-26; including one of the authors) participated in the second experiment.

4.2.3.2 Scanning parameters

For both experiments, imaging data were acquired with a 3 Tesla G.E. Healthcare scanner. A gradient echo, echo-planar sequence was used to obtain functional images [volume repetition time (TR), 2 s; echo time (TE), 30 ms; flip angle, 80°; matrix, 128x128 voxels; FOV, 20 cm; 29 oblique 3 mm slices with 1 mm gap; in-plane resolution, 1.56x1.56mm]. The functional data were motion-corrected, each voxel's mean value was scaled to equal 100, and linear trends were removed from each run, using the AFNI software package [69]. No other preprocessing (e.g. spatial smoothing, slice timing correction, temporal smoothing) was applied. We collected a high-resolution

(1x1x1mm voxels) structural scan (SPGR; TR, 5.9 ms; TE, 2.0 ms; flip angle, 11°) in each scanning session. Images were presented using a back-projection system (Optoma Corporation) operating at a resolution of 1024 x 768 pixels at 75 Hz.

4.2.3.3 Visual stimuli and experimental design

For our first experiment, we collected early visual cortex responses from 13 subjects. We used a typical retinotopic mapping protocol, in which a checkerboard pattern undergoing contrast reversals at 5Hz moved through the visual field in discrete increments [257]. First, a wedge subtending an angle of 45 degrees from fixation was presented at 16 different polar angles for 2.4 seconds each. Next, an annulus subtending 3 degrees of visual angle was presented at 15 different radii for 2.4 seconds each. Each subject passively observed two runs of 6 cycles in each condition, yielding 512 timepoints per subject (see Fig. 4.2a).

Our second dataset consists of PPA, FFA, and hV4 responses from 8 subjects. We presented two types of stimuli, as shown in Fig. 4.2b: (1) boats and cars on a blank white background (isolated objects); and (2) boats and cars with a street or water scene background (objects in context). Images (450 x 450 pixels; subtending 24 x 24 degrees of visual angle) were presented 100 pixels (5 degrees) away from fixation in randomly determined directions. Subjects were informed that each image contained either a boat or a car, and were asked to indicate as quickly as possible whether the object was on the left half of the image or the right half of the image (using a button box). Subjects performed 4 runs, with 16 blocks per run (with a 14 s gap between blocks) and 9 images per block. The first 8 blocks of each run showed a boat or car placed in a photographic scene; for each block, the object could violate a semantic relationship (appearing in the wrong type of scene, e.g. a boat on a city street) and/or a geometric relationship (appearing in the wrong position in the scene, e.g. a car above a tree rather than on the street). Each presentation consisted of a 500 ms fixation cross, an image flashed for 100 ms, a 300 ms mask, and then a 1300 ms response period (blank gray screen). The last 8 blocks of each run showed a boat or car on a white background; these images were identical to those presented in the first eight blocks, with the backgrounds removed (and presented in a different random

order). Each presentation consisted of a 500 ms fixation cross, an image flashed for 350 ms, and then a 1300 ms response period (blank gray screen). The total number of timepoints for each of the 8 subjects was 1,224 (306 per run).

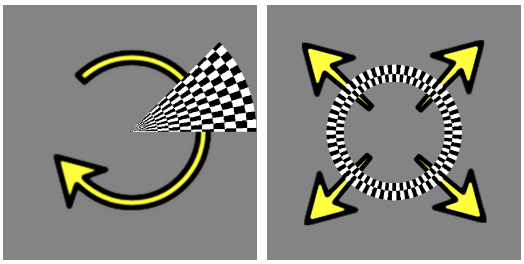
4.2.3.4 ROIs

In order to measure the eccentricity biases of PPA and FFA in the second experiment, we defined these regions using standard localizer runs conducted in a separate fMRI experiment. Subjects performed 2 runs, each with 12 blocks drawn equally from six categories: child faces, adult faces, indoor scenes, outdoor scenes, objects (abstract sculptures with no semantic meaning), and scrambled objects. Blocks were separated by 12 s fixation cross periods, and consisted of 12 image presentations, each of which consisted of a 900 ms image followed by a 100 ms fixation cross. Each image was presented exactly once, with the exception of two images during each block that were repeated twice in a row. Subjects were asked to maintain fixation at the center of the screen, and respond via button-press whenever an image was repeated. PPA was defined as the top 300 voxels near parahippocampal gyrus for the Scenes > Objects contrast, and FFA was defined as the top 100 voxels near fusiform gyrus for the Faces > Objects contrast. The volume of each ROI in mm^3 was chosen conservatively, based on previous results [104]. The locations of early visual areas V1, VP, and hV4 were delineated on a flattened cortical surface for each subject, using a horizontal meridian vs. vertical meridian general linear test from the retinotopic mapping data to give the boundaries between retinotopic maps.

4.3 Results

4.3.1 VP-V1 Connectivity

We know that voxels in early visual cortex exhibit strongly retinotopic population receptive fields [82]. Recent work has shown that the structure of functional connectivity between early visual areas preserves retinotopic organization. Specifically, the activity of a voxel in V3 is best predicted by voxels in V1 that correspond to the same



(a) The first dataset consists of responses to two flickering checkerboard patterns: a 45° wedge which rotates clockwise through the visual field, and an annulus subtending 3° of visual angle that expands outward from fixation.

(b) The second dataset consists of cars and boats, presented either in isolation or in a scene context.

Figure 4.2: **Stimuli used in our two datasets.**

retinotopic position in the visual field [130].

In this section, we validate our method by showing how it can be used to discover such connections between retinotopic areas of the early visual cortex. We apply our connectivity method to the early visual cortex dataset with V1 as area $\boldsymbol{A^1}$ and a single voxel in VP (ventral V3, or V3v) as area $\boldsymbol{A^2}$ (Eq. 4.3). For each voxel in VP, we obtain a separate connectivity map a of voxel weights in V1.

To quantitatively measure the precision of the learned V1 maps, we first assign a preferred angle and eccentricity to each voxel in V1 and VP. We use the t-statistics from a standard general linear model (GLM) to quantify the preference of each voxel to each wedge angle and each annulus radius [136]. Specifically, for each voxel v in the two areas, we take a weighted average of all stimulus angles, with weights proportional to that voxel's t-statistic for that angle θ_i (ignoring negative t-statistics):

$$\text{pref}_\theta(v) = \tan^{-1}\left(\frac{\sum\limits_{\{i|t_i^v>0\}} t_i^v \cdot \sin(\theta_i)}{\sum\limits_{\{i|t_i^v>0\}} t_i^v \cdot \cos(\theta_i)}\right)$$

where $\theta_i \in \{0, 22.5, 45, 67.5, ...337.5\}$ and t_i^v is the marginal t-statistic for angle i at voxel v.

Similarly, we compute the preferred eccentricity for each voxel v by taking a weighted average of the stimulus radii R_i:

$$\text{pref}_r(v) = \frac{\sum\limits_{\{i|t_i^v>0\}} t_i^v \cdot R_i}{\sum\limits_{\{i|t_i^v>0\}} t_i^v}$$

where $R_i \in \{0.73, 1.46, 2.92, 4.38, ...18.98, 19.71\}$ and t_i^v is the marginal t-statistic for radius i at voxel v.

Finally, we can estimate the position of the population receptive field for v by converting to cartesian coordinates:

$$\text{RF}(v) = \text{pref}_r(v) \cdot [\cos(\text{pref}_\theta(v)), \sin(\text{pref}_\theta(v))]$$

Given the population receptive field locations for each V1 and VP voxel, we can compare the receptive field RF(v) of each voxel v in VP with the receptive fields of the V1 voxels in v's connectivity map. If the V1 connectivity map for voxel v preserves retinotopic organization, then the V1 voxels with high positive weights should have the same retinotopic position as v. We therefore take a weighted average of the V1 receptive fields, in which the weight for each V1 voxel corresponds to its learned connectivity weight (negative weights are set to zero for this computation). This allows us to compare the receptive field of VP voxel v with that generated by the connected voxels in V1, as shown in Fig. 4.3. To ensure that the receptive field estimates are an independent measure of performance, we compute the receptive field positions using the first run of the wedge and annulus data, and learn connectivity maps using the second run.

Fig. 4.4 describes the results across all 13 subjects, with $\lambda = 10^3$ and $k = 10$. We observe a marked decrease in the magnitude of the receptive field differences between VP and V1 when adding regularization, with the median difference reduced by an average of 31% ($t(12) = 11.19, p << 0.01$, two-tailed paired t-test). With regularization, the V1 maps become much more precise, with the majority of the positive learned V1 weights falling in a retinotopic location similar to that of the VP voxel that generated them. This result demonstrates that our regularized method produces V1 maps that are not only spatially coherent, but also functionally correct. It also shows that our method can perform well even with very little data; we use only 256 timepoints to estimate connectivity maps over all ∼1000 V1 voxels. The performance of any connectivity method on this dataset will be limited by the uncertainty in our VP receptive field position estimates (introduced by the limited number of wedge and annulus positions used, and the small number of temporal samples); we can approximate this uncertainty by comparing the RF(v) calculated from a single run to the RF(v) calculated from both runs. This loose error bound is plotted in Fig. 4.4, indicating that our method makes significant progress toward the optimal result even with such a small number of training timepoints. Similar results for regularized maps are observed over a large range of λ and k values (see Supplementary Fig. C1).

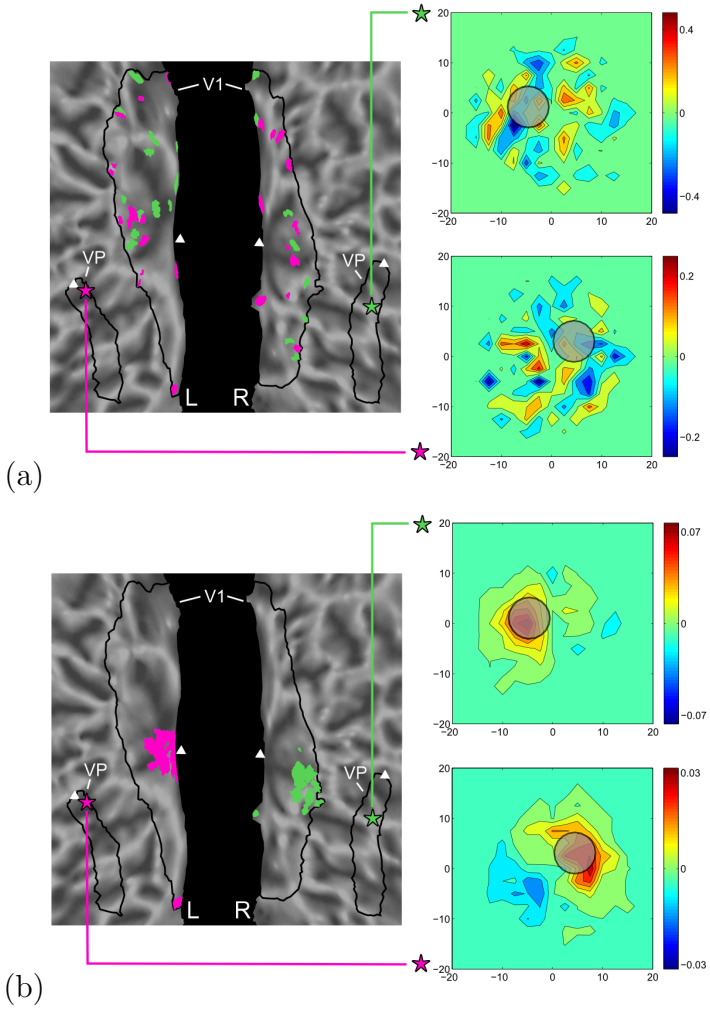


Figure 4.3: **Learned connectivity maps and receptive fields for 2 VP voxels, without regularization (a) and with regularization (b).** Two VP voxels are denoted by purple and green stars, and the top 30 voxels from the learned connectivity maps are shown in respective color in V1 (triangles indicate the location of the fovea). The inset plots compare the average receptive field of the connected V1 voxels (heatmap) with the actual population receptive field of each VP voxel (gray circle, radius given by the average uncertainty in our receptive field estimates). (a) The unregularized method produces maps with scattered weights, and the receptive fields of the connected V1 voxels are poor predictors of the VP receptive field. (b) The regularized connectivity method learns spatially coherent connectivity maps consistent with retinotopic organization, and the receptive fields of the connected V1 voxels are similar to that of the VP voxel.

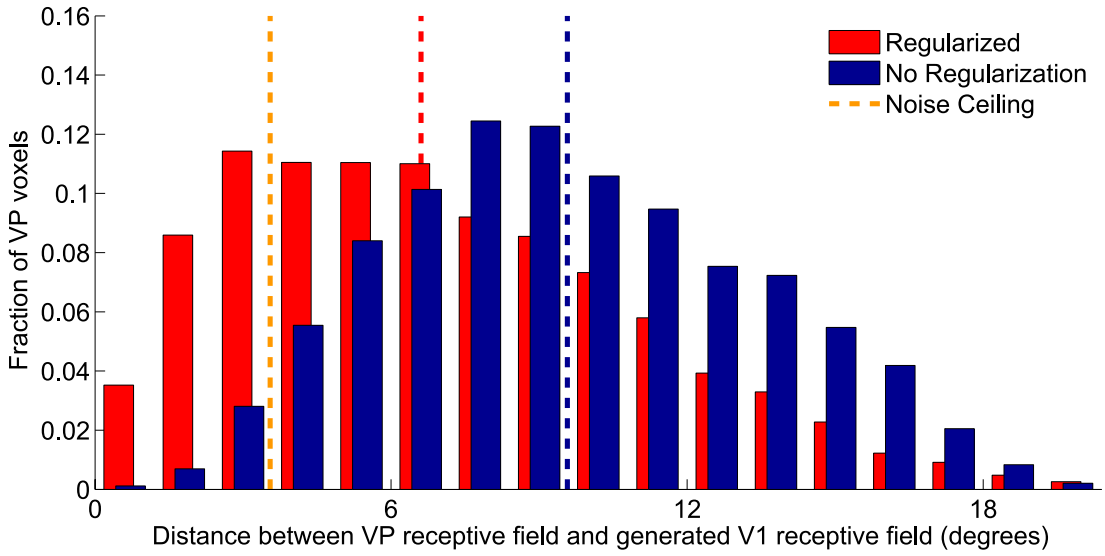


Figure 4.4: **Histogram comparing the precision of V1 maps generated from VP voxels.** The X-axis indicates the difference between the receptive field locations of VP voxels and the weighted average of the receptive fields in corresponding V1 connectivity maps. Since the actual functional connectivity between V1 and VP is known to preserve retinotopy, each VP voxel and its learned V1 connectivity map should have similar receptive field locations. The Y-Axis shows the fraction of VP voxels in each difference bin spanning 1.2 degrees of visual angle. Red bars (back) show results for regularized maps ($\lambda = 10^3, k = 10$), which demonstrate significantly smaller differences than blue bars (front), which show results for non-regularized maps ($\lambda = 0$). The dotted lines compare the median difference of both methods to a loose lower bound, based on the uncertainty in our receptive field estimates.

4.3.2 hV4-PPA/FFA Connectivity

Previous work has shown that there is a preferential response in PPA to peripherally-presented stimuli, and in FFA to foveally-presented stimuli; this effect has been measured both with discrete stimuli [175] and with traveling wave methods [103]. Experiments using diffusion tensor imaging (DTI) have provided evidence that this eccentricity bias is also present in the connectivity structure, with projections to early visual areas terminating at peripheral eccentricities for PPA and foveal eccentricities for FFA [158]. Our connectivity method provides a simple way of revealing such differential connectivity patterns, which does not require a specialized experimental design or a large amount of data. We chose to learn connectivity maps from PPA/FFA to area hV4 (as described in [299]), since it is the area in visual cortex most closely connected to ventral regions and is therefore most likely to show strong functional connectivity patterns.

We first examine the effect of varying λ on this dataset, and describe a principled approach for automatically selecting the regularization strength. λ controls the complexity of the learned connectivity patterns; as $\lambda \to \infty$, we can learn only constant-weight maps, while as $\lambda \to 0$, the weights are allowed to vary completely independently and maps can be arbitrarily complex.

We now use hV4 as area $\boldsymbol{A^1}$ and either PPA or FFA as area $\boldsymbol{A^2}$ ($k = 10$); the goal of our optimization is to find a map of weights for the hV4 voxels that allows for the best prediction of the mean PPA or FFA timecourse. For each subject, we train the model parameters on one run and then test on the other three runs (results are averaged across the choice of training run). The testing accuracies across a wide range of λ values (spaced logarithmically with step ratio of $10^{0.25}$) are shown in Fig. 4.5 (upper plot). At low values of λ, the connectivity maps are highly complex. These maps severely overfit to the training run, and fail to generalize to testing runs. At high values of λ, testing performance converges to essentially the same result as in the traditional connectivity method, in which all voxels have the same weight (unlike the traditional method, each hemisphere can have a different constant weight). However, the surprising characteristic of the testing accuracy curve is that it **does not increase monotonically** as λ increases. In every subject, the best testing

performance occurred at an intermediate value of λ, which shows that there exists a non-constant connectivity structure which is stable between runs; across subjects, testing performance was significantly increased over the traditional method ($\lambda = \infty$) for $10^{-0.25} < \lambda < 10^{6.75}$ for PPA and $10^{1.5} < \lambda < 10^{6}$ for FFA ($t(7) < -1.89, p < 0.05$, one-tailed paired t-test, uncorrected). This result shows that our method can carefully balance the trade-off between model complexity and data availability. Note that it is not possible to find generalizable connectivity maps using only pre-smoothing rather than spatial regularization (see Supplementary Fig. C2).

We obtain the best generalization performance around $\lambda = 10^{1}$, where we learn maps with a smoothness of approximately 9 mm FWHM (see Supplementary Fig. C3). As shown in the lower plot of Fig. 4.5, the connectivity maps in this regime have eccentricity biases in opposite directions for the two seed regions, with PPA biased toward peripheral eccentricities and FFA biased toward foveal eccentricities (correlation of learned weights with voxel eccentricities is significantly different for $10^{-1.25} < \lambda < 10^{3}, t(7) > 2.36, p < 0.05$, two-tailed paired t-test after z-transform, uncorrected).

Fig. 4.6 compares the eccentricity biases of the learned maps, with λ for each subject chosen to maximize generalization accuracy. Using all 306 timepoints from a run, the hV4 connectivity map with PPA is biased toward larger eccentricities, with an average correlation between eccentricity and connectivity weight of 0.21 ($t(7) = 2.83, p < 0.05$, one-tailed t-test after z-transform) while the hV4 connectivity map with FFA is biased toward smaller eccentricities, with an average correlation of -0.16 ($t(7) = -2.24, p < 0.05$, one-tailed t-test after z-transform) (PPA and FFA eccentricities significantly different, $t(7) = 4.19, p < 0.01$, two-tailed paired t-test after z-transform). We can obtain similar results using only the 148 "resting" timepoints in between stimulus blocks, in which subjects are simply fixating on a blank screen, suggesting that our method is sensitive to general functional connectivity rather than a stimulus mediated effect (PPA: $t(7) = 3.51, p < 0.01$, one-tailed t-test after z-transform; FFA: $t(7) = -2.39, p < 0.05$, one-tailed t-test after z-transform; Difference: $t(7) = 4.88, p < 0.01$, two-tailed paired t-test after z-transform).

To demonstrate that our method is more powerful than simpler approaches, the

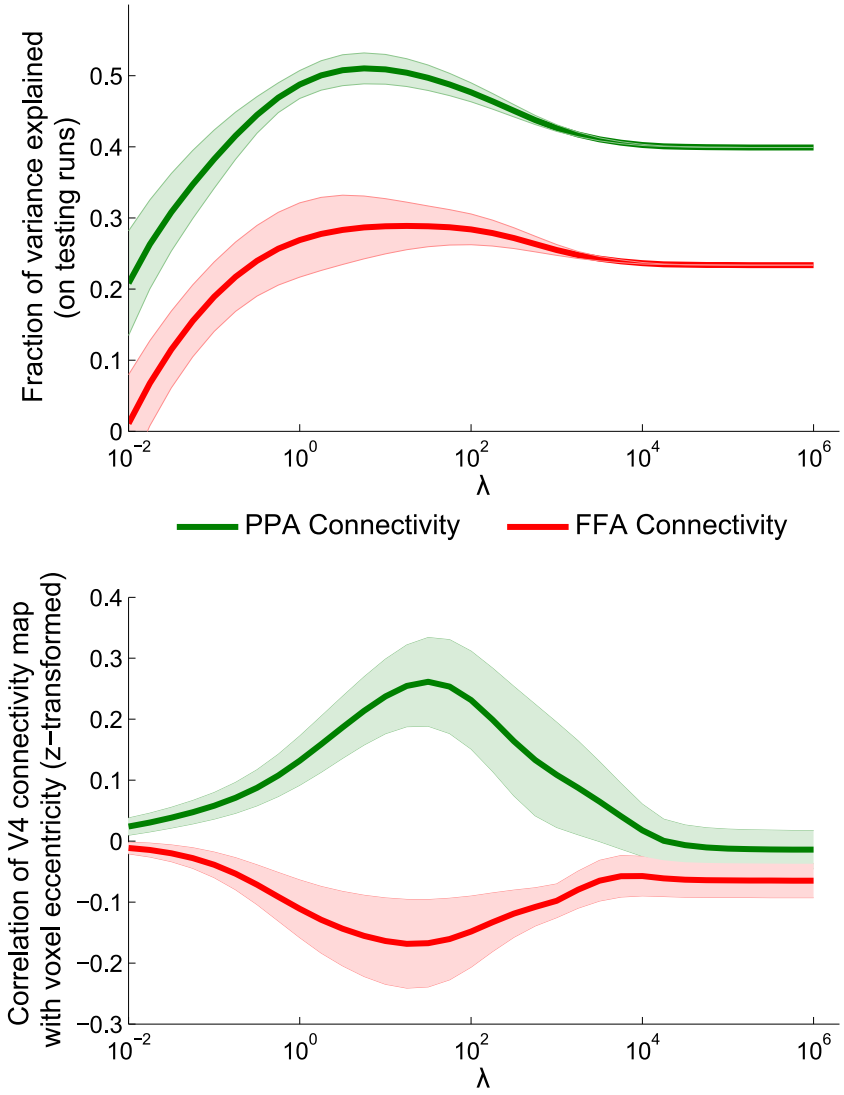


Figure 4.5: **Effects of changing λ on learned hV4 connectivity maps.** Connectivity maps over hV4 were learned with different regularization strengths λ, for seed regions PPA and FFA. An appropriate λ value can be chosen by maximizing the generalization performance of the learned maps, based on held-out testing runs (upper plot). At these values of λ, PPA and FFA show connectivity biases toward peripheral and central eccentricities, respectively (lower plot). Shaded regions indicate standard error across subjects (controlling for performance in the fully-regularized condition for the upper plot).

hV4 eccentricity biases for connectivity with PPA and FFA are computed in two additional ways: voxel-wise correlation (C), in which the weight of each hV4 voxel is set to the correlation between the timecourse of that voxel and PPA or FFA; and an unregularized version of our method (U) in which $\lambda = 0$. There are only two cases in which these methods give a significant result - the correlation method shows a foveal bias for FFA when using all TRs ($t(7) = -2.27, p < 0.05$, one-tailed t-test after z-transform) and the unregularized method shows a peripheral bias for PPA when using the resting TRs ($t(7) = 5.60, p < 0.01$, one-tailed t-test after z-transform). For both all TRs and the resting TRs, the difference between PPA and FFA eccentricity biases is significantly greater using our method than using the correlation method (all TRs: $t(7) = 3.63, p < 0.01$, resting TRs: $t(7) = 3.90, p < 0.01$, two-tailed paired t-test after z-transform) or using the unregularized method (all TRs: $t(7) = 4.20, p < 0.01$, resting TRs: $t(7) = 4.86, p < 0.01$, two-tailed paired t-test after z-transform). Our approach is therefore significantly more sensitive than either performing independent correlations between individual voxels and the seed region, or learning maps over all voxels without using spatial regularization.

A potential concern regarding functional connectivity measures is that they may be driven by local noise correlations, such that nearby voxels are good predictors of each other even if the underlying neural signals are unrelated. To ensure that our results are not being caused by relative positions of the ROIs, we ran a control analysis in which each hV4 voxel's connectivity weight was simply inversely proportional to its distance from the seed region. For bilateral ROIs, we set the weight of voxel $v = 1/(\text{dist from } v \text{ to left ROI}) + 1/(\text{dist from } v \text{ to right ROI})$. Since both PPA and FFA are closest to the anterior (peripheral) side of hV4, this model erroneously predicts that PPA and FFA should both show a peripheral eccentricity bias (PPA: $t(7) = 5.59, p < 0.01$; FFA: $t(7) = 3.03, p < 0.05$; two-tailed t-test after z-transform). Our results therefore cannot be explained simply by the physical arrangement of the ROIs.

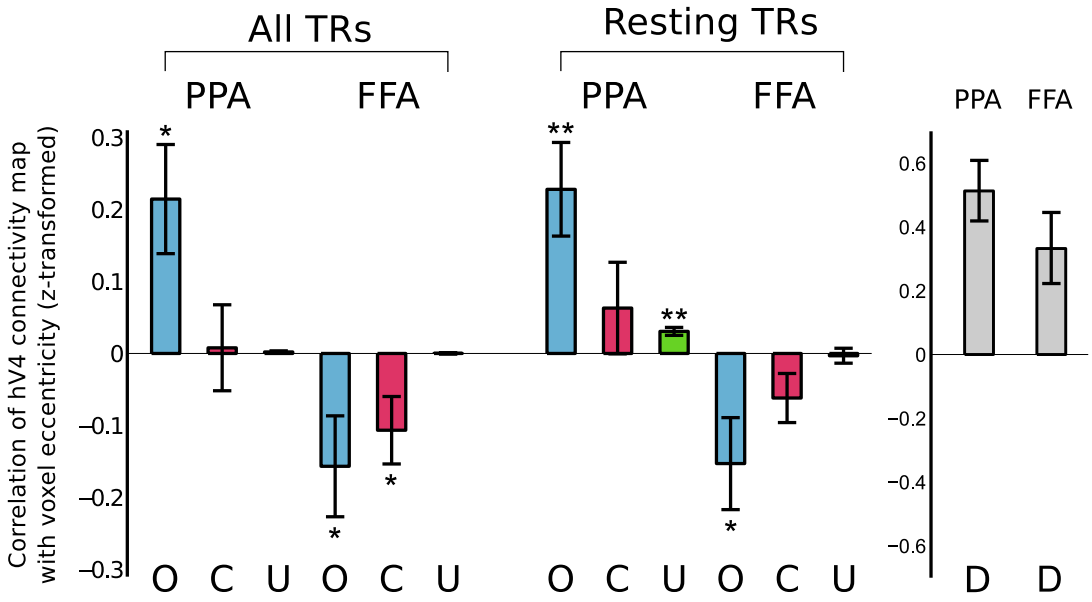


Figure 4.6: **hV4 eccentricity differences for optimal values of λ.** After choosing an optimal λ value for each subject based on generalization performance (see Fig. 5), we compute the eccentricity of hV4 connectivity maps for seed regions PPA and FFA, using our method (O), a voxel correlation method (C), and our method without regularization (U) (results averaged across four runs for each subject). Whether using all timepoints from a run (306 TRs) or using only those timepoints during which no stimulus was presented (approx. 148 TRs), our method finds that connectivity with PPA increases with increasing eccentricity, while the opposite is true for FFA. The correlation and unregularized controls are much less sensitive, showing significantly smaller differences between PPA and FFA eccentricity biases. Additionally, our results cannot be explained simply by local noise correlations; since both PPA and FFA are closer to the anterior (peripheral) side of hV4, such a model would predict similar peripheral eccentricity biases in PPA and FFA (D). Error bars indicate standard error, $*p < 0.05$, $**p < 0.01$.

4.4 Discussion

We have shown that our method can successfully extract known functional connectivity structures for two sets of regions. By adding spatial regularization to the traditional functional connectivity measure, our estimate of the connectivity between V1 and VP was made significantly more accurate, showing a clear retinotopic organization. We also demonstrated the expected eccentricity biases in the connectivity between V4 and PPA/FFA; unlike past experiments showing this effect [103, 175], this was accomplished without using a specialized experimental design, and could even be estimated from only resting-state data. The success of our method on these two different datasets demonstrates that this technique is likely to be applicable to a wide range of datasets and scientific questions. Note that we are able to learn these connectivity maps using only ~200 timepoints, in constrast to the ~2000 timepoints needed for complex models such as SVR [130]. Therefore, this method could be highly useful for detecting subtle variations in connectivity using small datasets. For example, it could plausibly be used to detect differences in connectivity across stimulus conditions, since only a small amount of data is required for learning.

Although these two experiments examined relatively simple characteristics of the learned weight maps (average retinotopic position or correlation with one of the spatial axes), our method should be applicable to any type of connectivity pattern, including multi-modal weight maps in which two separate sections of an ROI show high connectivity. Since the smoothness of the learned maps is controlled by a continuous parameter λ, our method is highly flexible and can learn arbitarily complex connectivity maps, given enough training data. For very large datasets, applying regularization will be less important, and the optimal value of λ (giving the best generalization accuracy) will decrease towards zero. Our method is therefore adaptive to the training set size, and will learn maps at finer and finer scales as the amount of training data increases.

Now that this method has been validated with known connectivity results, there are many opportunities to discover new connectivity patterns. One possible application would be to learn connectivity maps in frontal regions, where functional ROIs

are difficult to define. By locating the voxels in the frontal lobe that are connected to known ROIs in sensory regions, we may be able to identify how low-level sensory information converges in or is modulated by higher-level regions. Also, given any ROI, we can describe its connectivity with the entire rest of the cortex, by iteratively scanning a seed searchlight through all of cortex and learning a connectivity map over the ROI for each seed position. This will allow us to determine whether certain regions of cortex are connected to specific voxels in our ROI, as in "functional fingerprint" methods [157].

There are several ways that our method could be extended in future work. One current limitation is that weights can only be learned over one region at a time; that is, Eq. 4.3 is not symmetric with respect to $\boldsymbol{A^1}$ and $\boldsymbol{A^2}$. Simply replacing $\text{mean}_v(\boldsymbol{A^2})$ with a weighted average $\boldsymbol{a_2}^T \cdot \boldsymbol{A^2}$ will yield the degenerate solution $\boldsymbol{a} = \boldsymbol{a_2} = 0$, so (non-convex) constraints must be added to produce reasonable results. Another possible extension would be to learn weights simultaneously across multiple subjects. After first obtaining a voxel correspondence between subjects using a functional alignment technique (such as [127]), we could learn a global set of weights that is shared by all subjects. We could also allow the weights to vary between subjects, but introduce a new regularization term that encourages subjects to have similar weight maps.

4.5 Learning Maps over Both Regions

We can extend this method to identify voxel-level functional connectivity maps between any two regions, as shown in Fig. 4.7. This method is the first to learn voxel-level connectivity maps simultaneously in both regions, to automatically identify multiple functional correspondences between regions, and to utilize spatial regularization to prevent overfitting on small fMRI datasets. Our formulation makes no assumptions about the connectivity structure between regions, making it much more widely applicable than previous methods. We first review some related approaches in section 4.6, none of which can learn voxel-level maps over a pair of regions. We show in section 4.7 how rewriting the traditional correlational method as an optimization problem leads naturally to our approach, and we discuss how our formulation can be solved

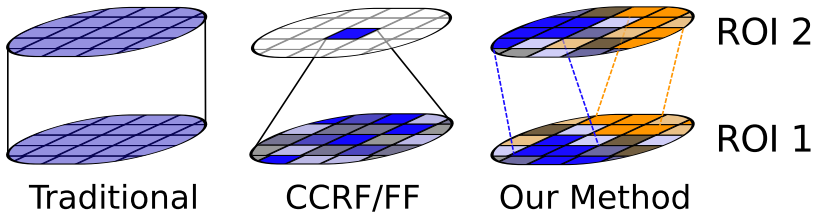


Figure 4.7: Functional connectivity methods. The standard measurement of functional connectivity between two regions averages together all voxels in each ROI, ignoring voxel-level connectivity differences. Recent CCRF/FF work produces a separate map over one region for each voxel in a seed region. Our method can learn connectivity structures over both ROIs simultaneously, and automatically identifies multiple connectivities between different sets of voxels.

efficiently using a trust region optimization method. In section 4.8 we validate our method on two pairs of ROIs, obtaining connectivity maps consistent with previous studies using only a small amount of training data. Finally, we conclude in section 4.9.

4.6 Related Work

Most fMRI studies measure functional connectivity between regions by simply computing the correlation between their mean timecourses [243], ignoring any connectivity differences at the subregion level. Methods that investigate *subregion* connectivity typically formulate the problem as learning 'cortico-cortical receptive fields" (CCRFs) [120, 130] or "functional fingerprints" (FFs) [157]. First, one of the two regions is chosen as a seed region. For each individual voxel (or cluster of voxels) in the seed region, these methods identify voxels in the second region that are most strongly functionally connected to the seed voxel. Seed voxels can then be grouped based on their connectivity signatures, by visual inspection [187], by a clustering method [157], or by an edge detection algorithm [66]. Such methods have discovered subregion connectivity patterns in a number of ROIs, including the thalamus [323], medial frontal cortex [157], and the precuneus [187]. The connectivity weights are generally computed using linear regression [187, 323] or correlation [157], but have also been

learned using support vector regression [130] and mutual information mapping [61]. All of these methods require treating the two regions asymmetrically, and cannot produce continuous-valued maps over both input regions; as shown in section 4.8.3, this reduces their sensitivity to fine-grained connectivity differences.

A small number of studies have computed connectivity maps by using timecourses from multiple seed clusters as predictors in a voxelwise linear regression, rather than learning maps independently for each seed cluster [186, 244]. However, these methods require downsampling the seed region to a small number of clusters based on prior anatomical knowledge (3 in [244], 16 in [186]) and manually comparing the connectivity maps for each seed cluster, making them difficult to apply to general ROI pairs. Canonical correlation analysis can be used to learn voxel-level maps over both regions [76], but has a number of limitations. Multiple correspondences between subregions can only be identified if their timecourses are not (positively) correlated (as shown in section 4.8.3, this assumption is not typically valid) and the number of voxels in each region must be smaller than the number of timepoints (limiting the datasets to which this method can be applied). Our method can learn voxel-level maps for regions of any size, and can identify distinct subregions even if their timecourses are correlated, making it widely applicable for investigating connectivity between any arbitrary pair of ROIs.

None of these previous functional connectivity methods apply any spatial regularization to the learned maps. Generally the voxel weights are set independently, which can require a large amount of training data to avoid overfitting. For example, in [130], even after training a support vector regression model on 44 minutes of data collection, a majority of the data points were still used as support vectors (indicating that the model is not yet saturated with data). Other groups have avoided this problem by constraining the learned maps to have a very simple shape, such as a Gaussian density [120]. Our model uses a spatial regularization term to compromise between these two extremes, allowing efficient estimation of smooth connectivity maps without making strong prior assumptions about the subregion shape.

4.7 Functional Connectivity as an Optimization Problem

4.7.1 Traditional Method

Functional connectivity between two ROIs is often measured by computing the Pearson product-moment correlation coefficient (r value) between the mean timecourses of the two ROIs [243]. Pairs of ROIs with a high r^2 value are then said to be strongly functionally connected. In order to generalize this method, we first recast it as an equivalent linear regression problem, in which we measure the similarity of the two mean timecourses, up to a scaling factor w:

$$\underset{w}{\text{minimize}} \; ||w \cdot \text{mean}_v(\boldsymbol{A^1}) - \text{mean}_v(\boldsymbol{A^2})||_2^2$$

where $\boldsymbol{A^1}$ and $\boldsymbol{A^2}$ are the (# voxels x # timepoints) data matrices from two ROIs, and $\text{mean}_v(\cdot)$ denotes an average across voxels. We assume that every voxel has been individually scaled to have zero mean across timepoints (a common fMRI preprocessing step) so the constant offset term in linear regression is not required. The r^2 value is then equivalent to the fraction of variance in $\text{mean}_v(\boldsymbol{A^2})$ explained by our predictor $w \cdot \text{mean}_v(\boldsymbol{A^1})$ [272].

We can therefore rewrite the traditional correlation method as an optimization problem in a more general form:

$$\begin{aligned} \underset{\boldsymbol{a^1},\boldsymbol{a^2},w}{\text{minimize}} \quad & ||\boldsymbol{a^1}^T\boldsymbol{A^1} - \boldsymbol{a^2}^T\boldsymbol{A^2}||_2^2 \\ \text{subject to} \quad & \boldsymbol{a^1} = \frac{w}{N_{A^1}} \cdot \mathbb{1}, \; \boldsymbol{a^2} = \frac{1}{N_{A^2}} \cdot \mathbb{1} \end{aligned} \tag{4.6}$$

where N_{A^k} is the number of voxels in ROI k, and $\boldsymbol{a^1}$ and $\boldsymbol{a^2}$ are vectors with lengths equal to the number of voxels in ROIs 1 and 2, respectively. We refer to $\boldsymbol{a^1}$ and $\boldsymbol{a^2}$ as *connectivity maps* over ROI 1 and ROI 2 (in this case, these maps take on a constant value for all voxels within an ROI). This is a convex optimization problem, and can be solved using a standard optimization package (all convex optimization problems in

our paper are solved using `CVX`, a package for specifying and solving convex programs [111]).

4.7.2 Voxel-Level Method

Although the traditional problem (4.6) can describe functional connectivity at the coarse scale of ROIs, it makes the simplistic assumption that all voxels within each region have the same functional connectivity properties. This prevents us from using the traditional method to explore connectivity differences at the voxel level, which are often of scientific interest [130, 157, 186, 187, 244, 323]. To learn voxel-level connectivity weights, we would like to relax the constraints on both $\boldsymbol{a^1}$ and $\boldsymbol{a^2}$ and allow the connectivity maps to be nonconstant. Note that a CCRF/FF method would relax only one of these constraints, learning a connectivity map over only one of the regions.

As will be shown in section 4.8, simply allowing each voxel to be chosen independently can lead to severe overfitting on the small datasets typical of fMRI experiments. It is possible to avoid overfitting by imposing a spatial regularization term that penalizes the average squared difference between every voxel i and its neighbors $n(i)$. This type of regularization encourages the maps to be spatially smooth, reflecting a common view of cortical organization, and has been applied in a variety of MRI and fMRI experiments [67, 73, 119, 212]. The neighborhoods $n(i)$ can be defined in a number of ways, with neighbors chosen based on physical distance between voxels or distance along the cortical surface. For the experiments in this paper, we choose the 10 voxels that are closest to i along the cortical surface (varying the number of neighbors from 5 to 15 has little effect). Adding these regularization terms to our objective function, we obtain:

$$\frac{1}{T}||\boldsymbol{a^1}^T\boldsymbol{A^1} - \boldsymbol{a^2}^T\boldsymbol{A^2}||_2^2 + \lambda\left[\sum_{i\in v_1}\sum_{j\in n(i)}\frac{1}{|n(i)|}(\boldsymbol{a_i^1} - \boldsymbol{a_j^1})^2 + \sum_{i\in v_2}\sum_{j\in n(i)}\frac{1}{|n(i)|}(\boldsymbol{a_i^2} - \boldsymbol{a_j^2})^2\right]$$

where T is the number of timepoints in our dataset, v_k is the set of all voxels in ROI k, and λ is a hyperparameter that controls the regularization strength. We can write

this objective compactly as

$$\left|\left|\begin{bmatrix} \frac{1}{\sqrt{T}}\boldsymbol{A^1}^T & -\frac{1}{\sqrt{T}}\boldsymbol{A^2}^T \\ \sqrt{\lambda}\boldsymbol{D_1} & 0 \\ 0 & \sqrt{\lambda}\boldsymbol{D_2} \end{bmatrix} \begin{bmatrix} \boldsymbol{a^1} \\ \boldsymbol{a^2} \end{bmatrix}\right|\right|_2^2$$

where $\boldsymbol{D_k}$ is a sparse connectivity matrix; each row represents an edge from a voxel i to a voxel j, with nonzero entries in column i ($1/\sqrt{|n(i)|}$) and column j ($-1/\sqrt{|n(i)|}$). Our objective therefore has the form $||\boldsymbol{X}_\lambda \cdot \boldsymbol{\beta}||_2^2$, where $\boldsymbol{\beta} = [\boldsymbol{a^1}\ \boldsymbol{a^2}]^T$ is the concatenation of the connectivity maps in both regions.

Since this is a homogeneous least-squares problem, it is clear that we must impose some constraint on the voxel weights $\boldsymbol{\beta}$ to avoid the degenerate solution $\boldsymbol{\beta} = 0$ (intuitively, the best timecourse prediction will always occur when the weight maps on both ROIs are identically zero, since this allows for perfect matching between the two regions). A standard method for choosing nonzero-weight solutions to homogeneous least-squares problems is to constrain the norm of $\boldsymbol{\beta}$ to be a constant. In addition, we impose the elementwise constraint $\boldsymbol{\beta} \succeq 0$; allowing negative connectivity weights makes the maps hard to interpret, since multiple solutions can be superimposed (with different signs) and inter-region connectivity can be confounded with intra-region connectivity (since weights in the same region can have opposite signs).

Our final optimization problem is therefore a constrained least-squares minimization:

$$\begin{aligned} \underset{\boldsymbol{\beta}}{\text{minimize}} \quad & ||\boldsymbol{X}_\lambda \cdot \boldsymbol{\beta}||_2^2 \\ \text{subject to} \quad & ||\boldsymbol{\beta}||_2 = 1,\ \boldsymbol{\beta} \succeq 0 \end{aligned} \tag{4.7}$$

4.7.3 Solving the Voxel-Level Optimization Problem

Due to the constraint $||\boldsymbol{\beta}||_2 = 1$, optimization problem (4.7) is not convex and may have multiple local minima. Although this makes the problem more complicated to analyze, the existence of multiple optima actually matches our intuition about functional connectivity structure; we know that for some pairs of regions (such as

in early visual cortex) there are multiple distinct connectivities between different subregions.

To find a locally optimal $\boldsymbol{\beta}$, we use a trust region approach [43, 60, 227]. This optimization method searches for local extrema by iteratively taking small steps in the parameter space. On each iteration, we create a convex approximation to the optimization problem by linearizing the norm constraint around the current set of parameters, and then find the optimal solution within a local trust region (see Algorithm 1). In our experiments, we use $\theta = 0.05$, $\Delta = \sqrt{\theta}$, and $\epsilon = \Delta/100$, but our results are not sensitive to this choice. We obtain multiple solutions by trying 20 initializations of $\boldsymbol{\beta_0}$ (below), each of which assigns all the connectivity weight to a single random voxel (in either region).

Algorithm 1: Iterative Trust Region Optimization

input : Initial $\boldsymbol{\beta_0}$, Trust region size Δ, Constraint tolerance θ, Convergence threshold ϵ

output: Locally optimal solution $\boldsymbol{\beta}$

while $||\mathbf{s}||_2 > \epsilon$ **do**

$$\begin{aligned} \underset{\boldsymbol{s}}{\text{minimize}} \quad & ||\boldsymbol{X}_\lambda \cdot (\boldsymbol{\beta_i} + \boldsymbol{s})||_2^2 \\ \text{subject to} \quad & |(||\boldsymbol{\beta_i}||_2^2 - 1) + 2{\boldsymbol{\beta_i}}^T \boldsymbol{s}| \leq \theta \\ & ||\boldsymbol{s}||_2 \leq \Delta,\ \boldsymbol{\beta_i} + \boldsymbol{s} \succeq 0 \end{aligned}$$

$\boldsymbol{\beta_{i+1}} \Leftarrow \boldsymbol{\beta_i} + \boldsymbol{s}$

4.7.4 Summary of Our Method

We have shown that the traditional correlational approach to functional connectivity can be rewritten as an equivalent optimization problem, in which all voxels in each region are constrained to take on a single fixed weight. Replacing this constraint with a spatial regularization term, we can learn connectivity weight maps by solving a constrained least-squares problem. A robust solution method for this problem is a trust region approach, which iteratively adjusts the connectivity maps in both regions

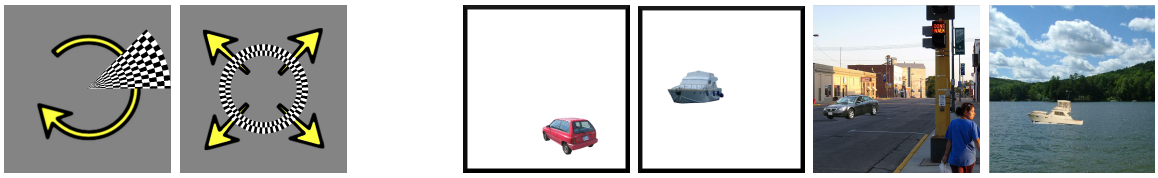
to find locally optimal solutions. In contrast to CCRF/FF methods, we learn maps in both regions simultaneously, and can automatically discover connectivities between different subregions.

4.8 Results

To demonstrate the versatility of our method, we show results for two separate experiments, each of which is performed on a separate dataset. As described in section 4.8.1, the first dataset consists of responses to moving checkerboard patterns, and the second dataset consists of responses to objects and scenes, with both datasets having a relatively small number of timepoints. In the first experiment (section 4.8.2), we apply our method to V1 and ventral V3 (VP), a pair of ROIs for which the ground-truth connectivity is known to be retinotopically organized. This allows us to quantitatively evaluate the quality of our learned maps. In the second experiment (section 4.8.3), we learn the connectivity pattern between the left and right halves of the lateral occipital complex (LOC), a region which is known to contain functional subdivisions along the anterior-posterior axis [118]. In both cases we demonstrate the importance of the spatial regularization term in our objective, by comparing the performance of our method without regularization ($\lambda = 0$) and with regularization. (When using regularization, we set $\lambda = 10^2$, but we obtain similar results for any λ within two orders of magnitude of this value.) In the second experiment, we also compare our results to those of the CCRF/FF method described in [157].

4.8.1 Experimental Design

For the V1-VP experiment, we use a retinotopic mapping dataset, illustrated in Fig. 4.8a. We collected early visual cortex responses from 13 subjects while a checkerboard pattern undergoing contrast reversals at 5Hz was moved through the visual field in discrete increments [257]. A wedge subtending an angle of 45 degrees from fixation was presented at 16 different polar angles for 2.4 seconds each, and an annulus



(a) (b)

Figure 4.8: Stimuli used in our two datasets. (a) In the first dataset, a flashing wedge pattern was presented at 16 different angles from fixation for two runs, and a flashing annulus was presented at 15 different eccentricities for two runs. (b) In the second dataset, images of boats and cars were presented both in isolation and in a scene context.

subtending 3 degrees of visual angle was presented at 15 different radii for 2.4 seconds each. Each subject passively observed two runs of 6 cycles in each condition, yielding 512 timepoints per subject. The first run of each condition is used to learn the preferred angle and eccentricity of each voxel: for each voxel, we perform a standard deconvolution to obtain a t-statistic for each of the 16 angle stimuli and each of the 15 eccentricity stimuli, and calculate a single preferred angle and eccentricity by separately averaging the directions and eccentricities weighted by their t-statistic. The borders of early visual regions were defined based on the horizontal and vertical meridian reversals. We then use the second run of each condition to train our model (256 timepoints total). Note that this is a very small number of timepoints compared to the number required by other approaches for learning connectivity maps (e.g. 1,760 timepoints are used in [130]).

For the LOC experiment, we use an objects-in-context dataset. We presented 10 subjects with two types of stimuli, as shown in Fig. 4.8b: (1) boats and cars on a blank white background (isolated objects); and (2) boats and cars with a street or water scene background (objects in context). Subjects performed 4 runs, viewing a total of 288 images of each type. We train our model on only a single run, consisting of 306 timepoints. In each subject, LOC was defined in an independent set of localizer scans as the top 500 voxels responding more to objects than scrambled images.

Imaging data were acquired with a 3 Tesla G.E. Healthcare scanner. A gradient echo, echo-planar sequence was used to obtain functional images (TR=2s,

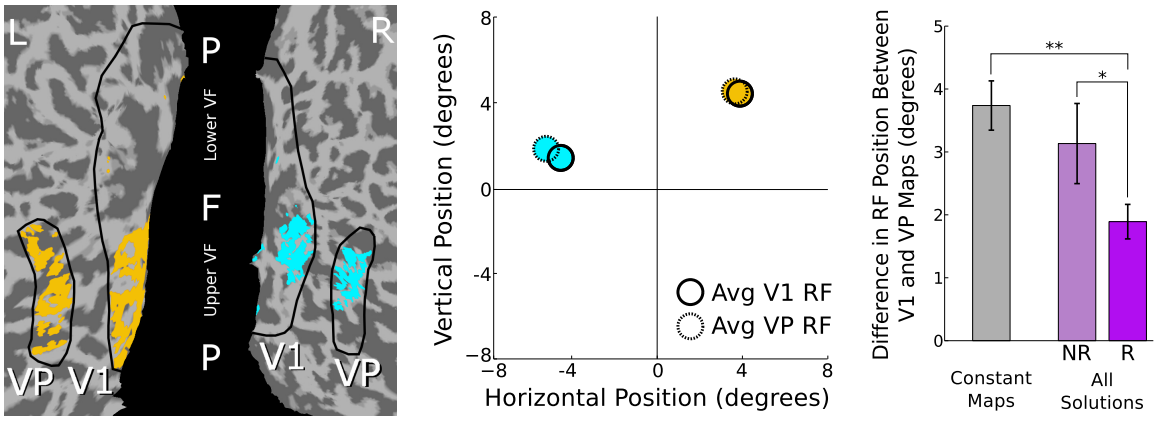
1.56x1.56x4 mm^3). The functional data were motion-corrected and each voxel's timecourse was z-scored to have zero mean and unit variance. We collected a high-resolution (1x1x1mm^3) SPGR structural scan in each scanning session. Stimuli were presented using a back-projection system (Optoma Corporation).

4.8.2 V1-VP Connectivity

We first apply our approach to the ROIs V1 and ventral V3 (VP), using the retinotopic mapping dataset. It is known that the functional connectivity between these regions is retinotopically organized [130], allowing us to objectively measure the quality of our learned maps. Since VP has receptive fields in the upper visual field, we expect that (at the coarsest level) VP should be functionally connected to the upper-visual-field portion of V1. Given a connectivity map over V1 and VP, we can compute the weighted average receptive field position in each area using the connectivity weights; if our maps preserve retinotopic correspondence, we should see a close match between the average receptive field positions in V1 and VP.

The results for a representative subject are shown in Fig. 4.9a. We obtain two locally optimal solutions to our optimization problem (4.7) in this subject, shown overlaid in two different colors. The resulting maps correctly identify that left and right VP are most strongly connected to left and right upper-field V1, respectively; this within-hemisphere connectivity was not included as a prior assumption, but was learned by our connectivity method. All subjects had between 2 and 4 solutions, with the left and right hemisphere correspondences appearing in every subject, and several subjects showing additional partitions based on eccentricity. Fig. 4.9b compares the average receptive fields in V1 and VP, for each of the two solutions in this subject. We find that the highly connected voxels correspond to similar positions in the visual field, confirming that we have identified a retinotopic correspondence between V1 and VP.

We then measure this match between the V1 and VP receptive field positions for all subjects. As a baseline, we measure the difference between the unweighted average receptive field positions in V1 versus VP; this corresponds to having constant weight



(a) (b) (c)

Figure 4.9: V1-VP connectivity results, for a representative subject (a-b) and all subjects (c). (a) We identify correspondences between voxels in V1 and VP, shown on a cortical flatmap (F: foveal region, P: peripheral regions, top 50 voxels from each solution shown in distinct colors). The two solutions in this subject identify the correspondence between subregions of VP and subregions of upper-visual-field V1 in the same hemisphere. (b) The average receptive field positions of the V1 and VP connectivity maps are very similar for each solution, indicating that these maps are consistent with retinotopic organization. (c) Learning maps without regularization (NR) yields only a small improvement over the baseline (lower is better), but our method significantly improves the match between average V1 and VP receptive fields when the spatial regularization term is included (R). * $p < 0.05$, ** $p < 0.01$, one-tailed paired t-test (n=13). (Best viewed in color)

maps in both regions. We compare this baseline to our method, without regularization (NR) and with spatial regularization (R). Fig. 4.9c shows that we obtain a significant reduction in error compared to our baseline ($t_{12} = 4.19, p < 0.01$, one-tailed paired t-test), but only when our spatial regularization term is included in the objective ($t_{12} = 1.88, p < 0.05$, one-tailed paired t-test). Without regularization, the maps severly overfit to our small dataset and give little improvement over the baseline measurement. Note that we have learned these maps using a very small number of timepoints, as described in section 4.8.1.

4.8.3 lLOC - rLOC Connectivity

We then learn the connectivity between left LOC (lLOC) and right LOC (rLOC), using the objects-in-context dataset. It has been previously shown that LOC consists of two functionally distinct subregions: a posterior-dorsal subdivision (LO), and an anterior-ventral subdivision (pFs) [118]. Since these two subregions have different functional response patterns, we expect distinct functional connections between the anterior side of lLOC and rLOC and/or between the posterior side of lLOC and rLOC.

For comparison purposes, we also apply a CCRF/FF correlation clustering method, which can only learn maps over one region at a time, from [157]. For each voxel in left LOC, the correlation of this voxel with all voxels in right LOC is computed. This method then applies k-means clustering (with a correlation distance measure) to partition the voxels in left LOC into two groups, based on the similarity of their right-LOC correlation maps. Finally, the process is repeated, using right LOC as the seed region and computing correlation maps over left LOC.

Fig. 4.10a shows the solutions for a representative subject. The top image shows the result of the CCRF/FF correlation clustering method, when using left LOC as the seed region. The two colors in left LOC denote the cluster to which each left LOC voxel was assigned, and the colors in right LOC show the top 40 voxels in the average correlation map for each cluster. We can see that there is no anterior-posterior spatial arrangement in either the cluster labels or the correlation maps (a similar result is obtained when using right LOC as the seed region). Applying our method without regularization (NR, middle image) also yields connectivity maps that highly overfit and have no consistent spatial structure. However, after adding the regularization term (R, bottom image), we obtain a clear posterior-anterior segregation of the two solutions in this subject. These two clusters partition left and right LOC into two functional units, corresponding to LO and pFs [118]. Note that our algorithm did not use any prior knowledge about the number of subregions or their spatial configuration. Across subjects, the number of solutions in the regularized case ranged from one to three.

To quantify the correspondence between the left and right connectivity maps, we

first separate all learned weights into five equally-spaced bins along the posterior-anterior axis. We then measure the correlation between these posterior-anterior connectivity profiles for the left and right hemisphere, resulting in Fig. 4.10b. Using correlation clustering, we obtain only a slight correlation between the posterior-anterior profiles (results averaged for the choice lLOC or rLOC as the seed region). This method is most likely failing for this connectivity task due to the small number of training timepoints, and the fine spatial scale at which we are attempting to discover these subregions (past applications of this method have been performed on data that is spatially downsampled to 5x5x5 mm^3 [157]). Note also that this clustering method cannot learn continuous-valued maps in both regions (since voxels in one of the regions are clustered into two discrete groups), seriously restricting its ability to represent fine-grained gradients of connectivity. Without regularization (NR), our method also fails to learn maps that have consistent anterior-posterior gradients in both hemispheres, again likely due to overfitting on the small input dataset. When adding regularization (R), however, our method produces maps which are highly correlated between hemispheres ($t_9 = 3.95, p < 0.01$, two-tailed t-test), giving a significant improvement over both the CCRF/FF method ($t_9 = 3.98, p < 0.01$, two-tailed paired t-test) and the unregularized method ($t_9 = 3.82, p < 0.01$, two-tailed paired t-test). Note that the representative timecourses for the anterior and posterior clusters are strongly positively correlated ($r = 0.83$ left, 0.81 right, averaged among subjects with exactly two clusters); the subtle distinction between these clusters therefore could not be identified by a CCA method [76]. Unlike [118], which used a specialized adaptation design, we are able to identify this anterior-posterior difference using only a single run from a dataset that was not tailored for this purpose.

4.8.4 Summary of Results

Using a single method (with a single setting of our hyperparameter λ), we are able to identify the true functional correspondences between two different pairs of ROIs. We have shown that we can successfully answer our original question, "which specific voxels in each of these two regions are most strongly connected?", without using

specialized datasets or a large number of training timepoints. By simultaneously learning weight maps over both regions and by including a spatial smoothness term, our method is much more sensitive to fine-grained connectivity differences than previous CCRF/FF methods.

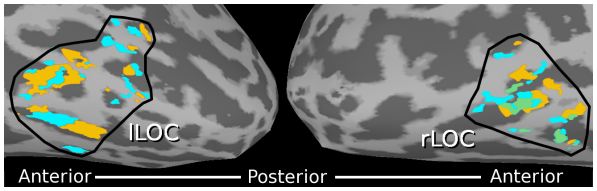
4.9 Conclusions

We have presented two new methods for discovering functional connectivity patterns between and within ROIs in the human brain. These methods are specifically tailored to the very small-size datasets typical of fMRI (addressing the known issue of data scarcity in this setting), and are capable of detecting subtle patterns at the voxel level. Our approach is fast, can operate efficiently with little input data, gives results consistent with prior work, and has proven to be a good candidate for investigating the structure of functional connectivity in the human brain.

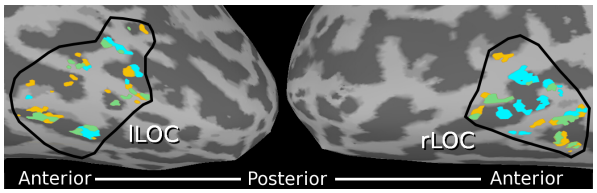
4.10 Acknowledgments

We thank Stephen Boyd for his suggestions and encouragement, and two anonymous reviewers for their detailed comments.

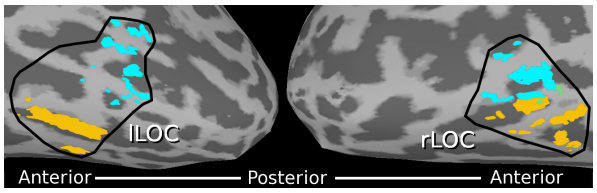
This work is funded by National Institutes of Health Grant 1 R01 EY019429 (to L.F.-F. and D.M.B.), a National Science Foundation Graduate Research Fellowship under Grant No. DGE-0645962 (to C.B.) and a William R. Hewlett Stanford Graduate Fellowship (to M.C.I.).



CCRF/FF Correlation Clustering

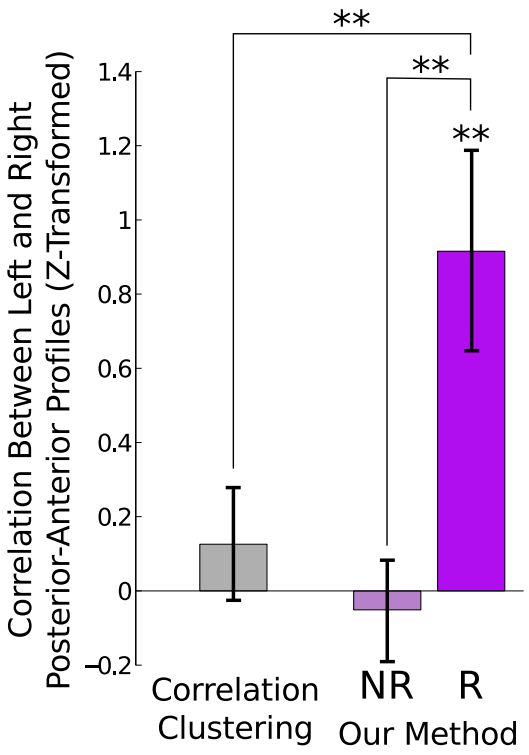


No Regularization (NR)



Regularized (R)

(a)



(b)

Figure 4.10: lLOC-rLOC results. (a) In this representative subject (top 40 voxels in each area shown in distinct colors), a CCRF/FF correlation clustering approach (top) fails to find anterior-posterior connectivity maps in left and right LOC, as does our method without the spatial regularization term (middle). Adding regularization (bottom) produces a separate posterior and anterior correspondence between hemispheres. (b) For each subject, we measure the correlation between the left and right hemisphere maps along the posterior-anterior dimension (larger is better). We see a strong correspondence between the left and right maps when using our proposed method with the spatial regularization term included (R), but not when the regularization term is removed (NR) or when we use the correlation clustering method. ** $p < 0.01$, two-tailed paired t-test (n=10). (Best viewed in color)

Chapter 5

Differential Connectivity Within the Parahippocampal Place Area

The Parahippocampal Place Area (PPA) has traditionally been considered a homogeneous region of interest, but recent evidence from both human studies and animal models has suggested that PPA may be composed of functionally distinct subunits. To investigate this hypothesis, we utilize a functional connectivity measure for fMRI that can estimate connectivity differences at the voxel level. Applying this method to whole-brain data from two experiments, we provide the first direct evidence that anterior and posterior PPA exhibit distinct connectivity patterns, with anterior PPA more strongly connected to regions in the default mode network (including the parieto-medial temporal pathway) and posterior PPA more strongly connected to occipital visual regions. We show that object sensitivity in PPA also has an anterior-posterior gradient, with stronger responses to abstract objects in posterior PPA. These findings cast doubt on the traditional view of PPA as a single coherent region, and suggest that PPA is composed of one subregion specialized for the processing of low-level visual features and object shape, and a separate subregion more involved in memory and scene context. This chapter is joint work with Diane M. Beck and Fei-Fei Li, and appeared previously in NeuroImage [15].

5.1 Introduction

Over the past two decades, functional magnetic resonance imaging (fMRI) has identified a number of category-selective regions involved in visual processing. Most of these regions have been defined based on differential activation to one category of stimuli over another, but this hypothesis-driven approach to mapping brain regions has significant drawbacks. Adjacent areas that have similar response profiles to the presented stimuli, but different functions, may be mistakenly conflated; for example, functionally distinct subregions have been identified in both object-sensitive lateral occipital complex (LOC) [118] and the extrastriate body area [308].

Another visual region that has been proposed as a candidate for subdivision is the Parahippocampal Place Area (PPA) [92]. This scene-sensitive area has been heavily implicated in visual scene perception, though the precise nature of the representation in this area has been controversial. Leading models have argued that PPA represents local scene geometry [93], spatial expanse [166, 218], space-defining objects [204], or contextual relationships [19]. All of these models have implicitly assumed that PPA is a homogeneous unit performing a single functional role, but this view has recently been called into question. In the last several years, a number of researchers have suggested that PPA could have multiple functional components. Differences in spatial frequency response [230], varying deficits resulting from PPA lesions [91], PPA's overlap with multiple visual field maps [9], and a clustering meta-analysis [261] all hint at the possibility that PPA may be comprised of at least two functionally distinct subunits along its posterior-anterior axis. However, studies explicitly searching for a distinction between posterior and anterior PPA have failed to identify major differences [51, 90].

Anatomical data from a proposed macaque homologue of PPA presents an interesting possibility for identifying subregions of human PPA. Although the definition of macque PPA is still a matter of ongoing research [209, 230, 261], a possible candidate spans cytoarchitectonically defined parahippocampal areas TH, TF, and TFO [167]. The most anterior area, TH, is primarily connected to retrosplenial cortex (RSC) [167, 274] and is also connected to the caudal inferior parietal lobule (cIPL) through

a parieto-medial temporal pathway [59, 167]. The more posterior TF is connected to a similar set of regions, but receives stronger input from ventral visual areas V4 and TEO [274]. The specific connectivity properties of the most posterior area (TFO) are not yet known, but it has been shown that TFO has a neuronal architecture highly similar to that of ventral visual regions [251]. In short, these macaque parahippocampal regions exhibit an anterior-posterior gradient, with the anterior side most related to RSC and cIPL and the posterior side most related to ventral visual areas.

Connectivity results in humans, using both diffusion tensor imaging (DTI) and fMRI, have shown that the parahippocampal region is connected to occipital visual cortex [154, 176, 246] as well as RSC and posterior parietal cortex [53, 149, 176, 246, 288], and PPA is known to combine both spatial and object identity information [121]. However, it is not known whether the posterior and anterior parts of the PPA connect differentially to these two networks. If human PPA corresponds to some or all of the macaque areas TH/TF/TFO, it should be possible to identify an anterior-posterior gradient in the functional connectivity properties of PPA. Such a finding would not only reinforce the proposed link between PPA and these macaque parahippocampal regions, but also demonstrate that PPA is actually composed of at least two regions operating on different types of visual information, shedding new light on the controversy over its functional properties.

To test whether voxels within PPA have differing connectivity properties, we apply our recent method for learning voxel-level connectivity maps [18]. Unlike standard functional connectivity measures that examine each voxel independently, our method considers all PPA voxels simultaneously to identify subtle differences in connectivity between voxels. After examining how several predefined ROIs connect to PPA, we perform a whole-brain searchlight analysis to identify the distinct cortical networks that connect preferentially to anterior or posterior PPA. We then demonstrate that these connectivity gradients are paired with gradients in functional selectivity, by evaluating the response to scenes and objects across PPA. Finally, we show that the connectivity gradients within PPA extend beyond PPA's borders, placing PPA in the context of ventral occipital and parahippocampal regions.

5.2 Materials and Methods

5.2.1 Regularized Connectivity Method

Investigating our hypothesis requires a method that characterizes functional connectivity patterns within a region of interest (ROI), at the voxel level. A number of studies have used fMRI functional connectivity measures to investigate structure within ROIs [61, 66, 157, 186, 187, 244, 323], but most previous approaches either do not measure connectivity at the voxel level (requiring spatial downsampling to a small number of subregions) and/or learn connectivity weights separately for each voxel (decreasing sensitivity and making comparisons between voxels more difficult). In our datasets, the PPA connectivity effects are too subtle to be detected by learning weights separately for each voxel (see Supplementary Fig. D1), and require the use of a method which can learn voxel-level connectivity maps that consider all voxels simultaneously. Support vector regression can learn these type of voxel-level connectivity maps [130], but does not utilize information about the spatial arrangement of the voxels and therefore requires a relatively large amount of data. To address this issue, we developed a method for examining connectivity differences within ROIs that is specifically tailored to small training sets typical in the fMRI setting. This method has been shown to recover voxel-level connectivity properties more accurately and efficiently than previous approaches [18].

The most common type of analysis for computing functional connectivity between two regions A^1 and A^2 measures how well the mean of all voxel timecourses in A^1 predicts the mean timecourse in A^2. We generalize this approach to identify voxel-level connectivity differences, by learning a *weighted* mean over the voxel timecourses in A^1 that best predicts the mean timecourse in A^2. The learned weights of the voxels in A^1 will then indicate the strength of the functional connection between each voxel and region A^2. Simply allowing each voxel weight to be learned independently leads to severe overfitting on typical fMRI datasets, but fMRI data naturally satisfies some regularity assumptions that can constrain our model. In particular, voxel connectivity properties are likely to be spatially correlated, with nearby voxels typically having more similar connectivity properties than spatially distant voxels. This reflects a

common view of cortical organization, and is especially applicable to blood-oxygen-level dependent (BOLD) signals such as fMRI, since the hemodynamic response is spatially smooth. To incorporate this assumption, we add a spatial regularization term to our model, which encourages each voxel in A^1 to have a connectivity weight similar to its spatially adjacent neighbors.

The learned connectivity maps are therefore a compromise between two objectives. Our first goal is to match the weighted average of the A^1 timecourses to the mean A^2 timecourse, by adjusting the weights. Our second goal is to make the weights spatially smooth, to prevent overfitting and allow our weights to generalize to independent data runs. The relative importance of this second goal is controlled by a hyperparameter λ, allowing us to trade off between having all weights be learned independently ($\lambda = 0$) and having all weights be identical ($\lambda = \infty$).

Mathematically, the connectivity weights are learned by minimizing the convex optimization objective

$$\underset{\boldsymbol{a},b}{\text{minimize}} \quad ||(\boldsymbol{a}^T \cdot \boldsymbol{A^1} + b) - \text{mean}_v(\boldsymbol{A^2})||_2^2 + \lambda ||\boldsymbol{D} \cdot \boldsymbol{a}||_2^2$$

where $\boldsymbol{a}$ is the connectivity weight map, b is a constant offset, $\boldsymbol{A^1}$ and $\boldsymbol{A^2}$ are the (# voxels x # timepoints) data matrices from two ROIs, and mean_v denotes an average across voxels. $\boldsymbol{D}$ is the voxel connectivity matrix, which we design to penalize the mean squared difference between the weight a_i of voxel i, and the weights of voxel i's neighbors: $||\boldsymbol{D} \cdot \boldsymbol{a}||_2^2 = \sum_{i=1}^{N} \frac{1}{|n_i|} \sum_{j \in n_i} (a_i - a_j)^2$ where N is the number of voxels in A^1 and n_i is the set of i's neighbors. The optimal $\boldsymbol{a}$ (for a given choice of λ) can be found efficiently by using a convex optimization package such as `CVX` [111]. For further details and validation experiments, see Baldassano et al. [18].

The following sections describe the collection of the datasets used to learn the connectivity weights $\boldsymbol{a}$. As will be shown in the Results, PPA's functional connectivity properties are not sensitive to the choice of experimental dataset; the specific details of the stimuli and tasks in these experiments are provided only for reference purposes.

5.2.2 Localizer and Object-in-Scene Experiments

5.2.2.1 Participants

10 subjects (3 female) with normal or corrected-to-normal vision participated in the object-in-scene and localizer fMRI experiment. The study protocol was approved by the Stanford University Institutional Review Board, and all subjects gave their written informed consent.

5.2.2.2 Scanning Parameters

Imaging data were acquired with a 3 Tesla G.E. Healthcare scanner. A gradient echo, echo-planar sequence was used to obtain functional images [volume repetition time (TR), 2 s; echo time (TE), 30 ms; flip angle, 80°; matrix, 128x128 voxels; FOV, 20 cm; 29 oblique 3 mm slices with 1 mm gap; in-plane resolution, 1.56x1.56mm]. The functional data was motion-corrected and each voxel's mean value was scaled to equal 100 (no spatial smoothing was applied). We collected a high-resolution (1x1x1mm voxels) structural scan (SPGR; TR, 5.9 ms; TE, 2.0 ms, flip angle, 11°) in each scanning session. The structural scan was used to calculate a transformation between each subject's brain and the Talairach atlas.

5.2.2.3 Localizer Stimuli and Procedure

For the localizer experiment, subjects performed 2 runs, each with 12 blocks drawn equally from six categories: child faces, adult faces, indoor scenes, outdoor scenes, objects (abstract sculptures with no semantic meaning), and scrambled objects (these stimuli have been used in previous studies such as [104]). Images (240 x 240 pixels; subtending 12.8 x 12.8 degrees of visual angle) were presented at fixation. Examples of scene and object stimuli are shown in Fig. 5.1a. Blocks were separated by 12s fixation cross periods, and consisted of 12 image presentations, each of which consisted of a 900 ms image followed by a 100 ms fixation cross. Each image was presented exactly once, with the exception of two images during each block that were repeated twice in a row. Subjects were asked to maintain fixation at the center of the screen, and

respond via button-press whenever an image was repeated. The total number of timepoints was 300 (150 per run).

5.2.2.4 Object-in-Scene Stimuli and Procedure

For the object-in-scene experiment, we presented two types of stimuli, as shown in Fig. 5.1b: (1) boats and cars on a blank white background (isolated objects); and (2) boats and cars with a street or water scene background (objects in context). Images (450 x 450 pixels; subtending 24 x 24 degrees of visual angle) were presented 100 pixels (5 degrees) away from fixation in randomly determined directions. Subjects were informed that each image contained either a boat or a car, and were asked to indicate as quickly as possible whether the object was on the left half of the image or the right half of the image (using a button box). Subjects performed 4 runs, with 16 blocks per run (with a 14 s gap between blocks) and 9 images per block. The first 8 blocks of each run showed a boat or car placed in a photographic scene; for each block, the object could violate a semantic relationship (appearing in the wrong type of scene, e.g. a boat on a city street) and/or a geometric relationship (appearing in the wrong position in the scene, e.g. a car above a tree rather than on the street). Each presentation consisted of a 500 ms fixation cross, an image flashed for 100 ms, a 300 ms mask, and then a 1300 ms response period (blank gray screen). The last 8 blocks of each run showed a boat or car on a white background; these images were identical to those presented in the first eight blocks, with the backgrounds removed (and presented in a different random order). Each presentation consisted of a 500 ms fixation cross, an image flashed for 350 ms, and then a 1300 ms response period (blank gray screen). The total number of timepoints was 1,224 (306 per run). Timepoints were classified as "resting" if they occured more than 4 seconds after the end of one stimulus block and less than 4 seconds after the start of the next stimulus block.

5.2.2.5 Functional Region of Interest Definition

Regressors for faces, scenes, objects, and scrambled objects in the localizer experiment were constructed by using the standard block hemodynamic model in `AFNI` [69].

LOC, PPA, RSC, and TOS were defined using the following contrasts: LOC, top 500 voxels for Objects > Scrambled near lateral occipital surface; PPA, top 300 voxels for Scenes > Objects near parahippocampal gyrus; RSC, top 200 voxels for Scenes > Objects near retrosplenial cortex; TOS, top 200 voxels for Scenes > Objects near the transverse occipital sulcus. The volume of each ROI in mm^3 was chosen conservatively, based on previous results [104]. Consistent with the meta-analysis by Nasr et al. [209], PPA in our subjects was found to be centered on the collateral sulcus adjacent to the parahippocampal gyrus.

5.2.3 Scene Category Experiment

5.2.3.1 Participants

8 subjects (4 female) with normal or corrected-to-normal vision participated in the scene category fMRI experiment (these subjects did not overlap with those in the object-in-scene experiment). The study protocol was approved by the University of Illinois Institutional Review Board, and all subjects gave their written informed consent.

5.2.3.2 Scanning Parameters

Functional imaging data were acquired with a 3 Tesla Siemens Trio scanner. A gradient echo, echo-planar sequence was used to obtain functional images [volume repetition time (TR), 1.75 s; echo time (TE), 30 ms; flip angle, 90°; matrix, 64x64 voxels; FOV, 19 cm; 29 oblique 3 mm slices with 0 mm gap; in-plane resolution, 3.0x3.0mm]. The functional data was motion-corrected and each voxel's mean value was scaled to equal 100 (no spatial smoothing was applied). We collected a high-resolution structural scan for each subject; 4 subjects were scanned in a 3 Tesla Siemens Trio scanner (MPRAGE; 1x1x1.2mm, TR, 1900 ms; TE, 2.25 ms, flip angle, 9°) and 4 subjects were scanned in a 3 Tesla Siemens Allegra (MPRAGE; 1.25x1.25x1.25mm, TR, 2000 ms; TE, 2.22 ms, flip angle, 8°). The structural scan was used to calculate a transformation between each subject's brain and the Talairach atlas.

5.2.3.3 Stimuli and Procedure

Images (800 x 600 pixels; subtending 24 x 18 degrees of visual angle) were presented in the center of the display using a back-projection system (Resonance Technologies) operating at a resolution of 800 x 600 pixels at 60 Hz. For each run, subjects were instructed to count the number of images belonging to a target category (beaches, cities, highways or mountains; see example stimuli in Fig. 5.1c). On average, there were 16 target images per run, ranging from 15-17 targets. Stimuli were presented in blocks of 8 images with a display time of 1.75 s for each image. Images within a block were primarily from the same natural scene category; however, in order to increase the difficulty of the counting task, one or two outgroup images from different scene categories (intrusions) occasionally appeared within a block. A fixation cross was presented throughout each block, and subjects were instructed to maintain fixation. There were 8 blocks in each run (2 blocks for each natural scene category), interleaved with 12 s fixation periods to allow for the hemodynamic response to return to baseline levels. A session contained 16 such runs, and the order of categories and intrusion images were counterbalanced and randomized across blocks. The total number of timepoints was 2,064 (129 per run). Timepoints were classified as "resting" if they occured more than 4 seconds after the end of one stimulus block and less than 4 seconds after the start of the next stimulus block.

5.2.3.4 Functional Region of Interest Definition

ROIs were defined using an independent localizer scan, consisting of blocks of face, object, scrambled object, landscape, and cityscape images. Each block consisted of 20 images presented for 450 ms each with a 330 ms interstimulus interval. Each of the five types of stimuli was presented four times during a run, with 12 s fixation periods after two or three blocks. Subjects completed two runs, performing a one-back task during the localizer by pressing a button every time an image was repeated. Regressors for faces, scenes, objects, and scrambled objects were constructed by using the standard block hemodynamic model in `AFNI` [69], and the following contrasts were used to define ROIs: LOC, Objects > Scrambled near lateral occipital surface; PPA, Scenes

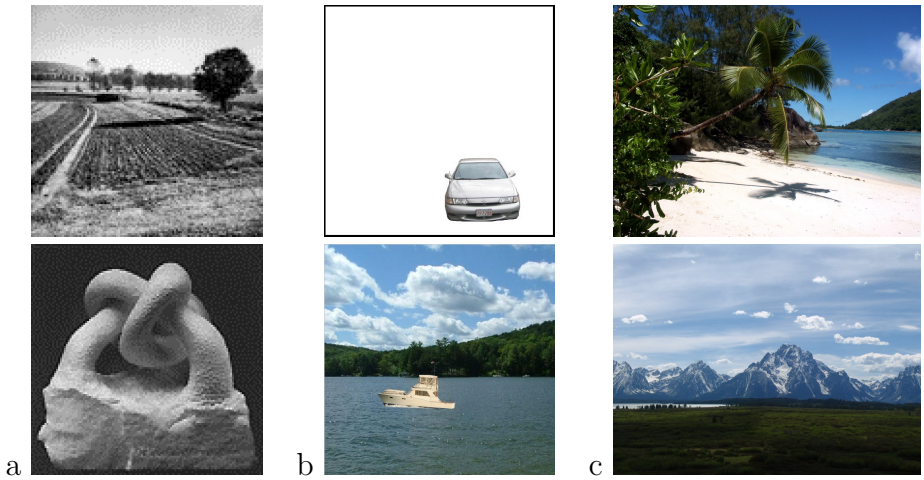


Figure 5.1: **Sample stimuli used in our experiments.** (a) Scene and object stimuli from the localizer experiment, which also included faces and scrambled objects. (b) Isolated object and object-in-scene stimuli from the object-in-scene experiment. (c) Beach and mountain stimuli from the scene category experiment, which also included cities and highways.

> Objects near parahippocampal gyrus; RSC, Scenes > Objects near retrosplenial cortex; TOS, Scenes > Objects near the transverse occipital sulcus. A threshold of $p < 2 \cdot 10^{-3}$ (uncorrected) was applied, and was tightened to break clusters if necessary.

5.2.4 Caudal IPL Definition

Caudal IPL is a region strongly connected to macaque parahippocampal cortex [167] for which we do not have a functional localizer. In order to evaluate the match between the macaque and human connectivity patterns, we sought to anatomically define a human region equivalent to cIPL. The two caudal-most areas of human IPL (defined using probabilistic cytoarchitectonic maps) are PGa and PGp, which are thought to correspond to the caudal-most sections of macaque IPL, PG and Opt [53]. Of these, PGp exhibits significantly stronger functional and structural connectivity with the parahippocampal gyrus [288], giving the best match with the proposed parieto-medial temporal pathway targeting parahippocampal areas from cIPL. We therefore define

cIPL in all subjects using the Eickhoff-Zilles PGp probabalistic cytoarchitectonic map [87] (based on [52, 54]). We thresholded the map at $p > 0.5$, and transformed the map into each subject's native space. Since cIPL slightly overlapped TOS in some subjects, any voxels shared between cIPL and TOS were excluded from both regions (no other ROIs included overlapping voxels).

5.2.5 PPA Connectivity Analysis: ROIs

We first learned PPA connectivity maps for four pre-defined seed regions: lateral occipital complex (LOC), transverse occipital sulcus (TOS, also referred to as the "occipital place area" in [79]), retrosplenial cortex (RSC), and caudal inferior parietal lobule (cIPL) by setting A^1 to be PPA and A^2 to be one of the four seed regions. To avoid functional connectivity idiosyncratic to a specific experiment or task, we used data from both the object-in-scene experiment and the scene category experiment (see above).

We first validated that our method could learn meaningful voxel-level connectivity maps which provide better generalization performance, compared to a connectivity map which is constant over left PPA and constant over right PPA. For each seed region and subject, we learned a connectivity map using one training run, and tuned the smoothness parameter λ to maximize the fraction of variance explained on a validation set consisting of all but one of the remaining runs. The classifier was then retrained on both the training run and validation set (using the selected λ value) and tested on the final held-out testing run. Results were averaged across all choices of training run, with a random testing run being chosen for each training run. These results were compared to those from ROI-level connectivity maps, in which all PPA voxels in each hemisphere were constrained to take on the same value (equivalent to $\lambda \to \infty$).

We then learned a weightmap over PPA for each subject and for each seed region using all experimental runs, with λ chosen such that the average fraction of variance explained, when training on one run and testing on the other runs, was maximized. We measured the correlation between the connectivity weights and the

anterior-posterior voxel coordinates, to obtain a simple measure of how the learned weights in PPA varied along the anterior-posterior axis. The correlation was computed separately for left and right PPA (except where specified, results below are collapsed across left and right PPA).

5.2.6 PPA Connectivity Analysis: Whole-Brain

To explore the connectivity patterns between PPA and the rest of the brain, we performed a whole-brain searchlight connectivity analysis in which our seed region was densely sampled throughout the entire cortex. We fixed A^1 to be PPA, and then placed a 3x3x3 voxel searchlight A^2 at each point on a lattice with 2 voxel spacing. For each searchlight, we used all experimental runs to learn a map of connectivity weights in PPA, and then measured the correlation between the learned weights and the anterior-posterior axis. We obtained an anterior PPA vs. posterior PPA preference for each brain voxel by averaging the correlation value of all searchlights which included that voxel. In order to speed up computation, we used a single value of $\lambda = 5.5$ for all subjects, equal to the average of the optimal λ values in the ROI experiment (in log space). Group-level statistics were computed by transforming each subject's results into Talairach space.

5.2.7 Scene- and Object-Sensitivity Analysis

After identifying connectivity differences among PPA voxels, we investigated whether these connectivity gradients corresponded to functional differences in stimulus selectivity. To measure the response properties of individual PPA voxels, we examined the statistics from the regressors in the localizer experiment. For each voxel, the t-statistics from the scene and object regressors were recorded, and each voxel was also given a binary label of "significantly activated" or "not significantly activated" based on whether its false discovery rate (FDR) for each category was less than or greater than 0.05. To detect a sensitivity gradient across PPA, the correlation between the anterior-posterior axis and the t-statistics was computed. For visualization purposes, each subject's PPA voxels were binned into 10 bins running anterior-posterior, and

the mean t-statistic and percentage of activated voxels was calcuated for each bin, to give a sensitivity profile.

5.2.8 LOC/TOS vs. RSC/cIPL Connectivity

After discovering that LOC/TOS and RSC/cIPL connect preferentially to different voxels in PPA (see Results), we sought to place these connectivity gradients in the context of the entire parahippocampal region. For each cortical voxel, we averaged the coefficients for the voxel's correlations with LOC and TOS, and compared it to the average of the coefficients for the voxel's correlations with RSC and cIPL. We transformed each subject's correlation maps into Talairach space, and identified voxels at the group level that showed a consistent different across subjects to LOC/TOS functional connectivity vs. RSC/cIPL functional connectivity. In addition to the parahippocampal region, we searched all of cortex for voxels with this connectivity pattern.

5.3 Results

Since we are interested in the intrinsic connectivity properties of PPA (rather than functional correlations idiosyncratic to a specific stimulus set), we localized PPA in two separate groups of subjects, each of which then performed a different experimental task with different stimuli. Although both experiments included scenes, in one case (identifying scene category) the scenes were directly relevant to the task, while in the other (locating a target object in scenes) scenes were not the primary focus. Given these datasets, do we see connectivity differences in anterior versus posterior PPA analogous to those in macaque parahippocampal cortex? Note that, since the connectivity patterns were similar in both datasets (see Supplementary Fig. D4), all connectivity results below are collapsed across both experiments.

5.3.1 PPA Connectivity Analysis: ROIs

We began our investigation of PPA's connectivity structure by learning PPA connectivity maps for four seed regions: two other scene-sensitive regions (TOS and RSC), an object-sensitive area in ventral occipital cortex (LOC), and a posterior parietal region known to exhibit parahippocampal connectivity (cIPL). We first confirmed that, for each individual subject, we could learn weight maps over PPA (describing its connectivity with each of these regions) that generalize well across runs. As shown in Fig. 5.2a, spatially smooth voxel-level connectivity maps in PPA predict activity in LOC, TOS, RSC, or cIPL better than a map which has only a single weight for left PPA and a single weight for right PPA (LOC: $t_{17} = 4.42$; TOS: $t_{17} = 4.63$; RSC: $t_{17} = 7.80$; cIPL: $t_{17} = 3.28$; all $p < 0.01$, two-tailed paired t-test). These results were computed by choosing λ to maximize the fraction of variance explained (on an independent validation set) but improvement over the traditional constant-weight connectivity held for a wide range of regularization strengths λ (see Supplementary Fig. D2). Although all regions showed at least some activity related to PPA's timecourse, a significantly smaller amount of the cIPL timecourse can be predicted by PPA (ROI-level: LOC>cIPL: $t_{17} = 7.31$; TOS>cIPL: $t_{17} = 10.58$; RSC>cIPL: $t_{17} = 10.23$; Voxel-level: LOC>cIPL: $t_{17} = 6.58$; TOS>cIPL: $t_{17} = 12.12$; RSC>cIPL: $t_{17} = 11.81$; all $p < 0.01$ two-tailed paired t-test), consistent with its proposed role as a general processing hub in parietal cortex with connections to many regions besides PPA [53, 55].

Since meaningful voxel-level weight maps can be learned for individual subjects, we can ask whether these weight maps show any anterior-posterior differences which are consistent across subjects. If PPA shows the same gradient of connectivity as TH/TF/TFO, we expect the posterior portion of PPA to be more strongly connected to occipital visual regions LOC and TOS, with the anterior portion of PPA more strongly connected to RSC and cIPL. As shown in Fig. 5.2b, this is precisely what we observed; LOC and TOS connectivity weights tend to increase moving anterior to posterior, while RSC and cIPL weights increase in the opposite direction (LOC: $t_{17} = 3.10, p < 0.01$; TOS: $t_{17} = 2.72, p = 0.01$; RSC: $t_{17} = -3.76, p < 0.01$; cIPL: $t_{17} = -3.24, p < 0.01$; two-tailed t-test after z-transform). These results are

collapsed across left and right PPA; both hemispheres showed similar connectivity patterns, though effects were somewhat stronger in left PPA, by an average of 0.13 ($t_{17} = 2.20, p = 0.042$; two-tailed t-test after z-transform). We did not observe significant differences along the inferior-superior axis or medial-lateral axis, except for preferential connectivity of cIPL to medial PPA (see Supplementary Fig. D3).

5.3.2 PPA Connectivity Analysis: Whole-Brain

Having established a consistent posterior-anterior gradient of connectivity between our regions of interest and PPA, we then performed a searchlight analysis to search for other brain regions with posterior-anterior PPA connectivity gradients; rather than using our fixed ROIs as seed regions, we swept a 3x3x3 voxel searchlight throughout the entire cortex. As in Fig. 5.2b, we learn a PPA connectivity map for each seed region and compute the correlation of this map with the anterior-posterior axis; those seed regions which induce a PPA weight map that is positively correlated with the anterior-posterior axis are preferentially connected to posterior PPA, while those inducing a negatively correlated weight map are preferentially connected to anterior PPA. The traditional (homogeneous) model of PPA predicts that consistent preferential connectivity should only occur for seed regions directly adjacent to posterior or anterior PPA (which will be correlated with the nearer part of PPA due to local noise). If PPA contains subregions similar to those in macaque, however, we would expect a number of regions throughout cortex to show preferential connectivity patterns which are both consistently non-zero and in opposite directions.

Our results are shown in Fig. 5.3. As predicted by our subregion hypothesis, seed regions in occipital visual areas (including LOC and TOS) showed preferential connectivity to posterior PPA, while RSC and cIPL showed preferential connectivity to anterior PPA. Note that these results cannot be explained by local noise correlations, since RSC and cIPL are physically closer to the posterior edge of PPA. We also observed connectivity to anterior PPA in ventral prefrontal cortex (PFC) (primarily on the medial surface) and on the lateral surface of the anterior temporal

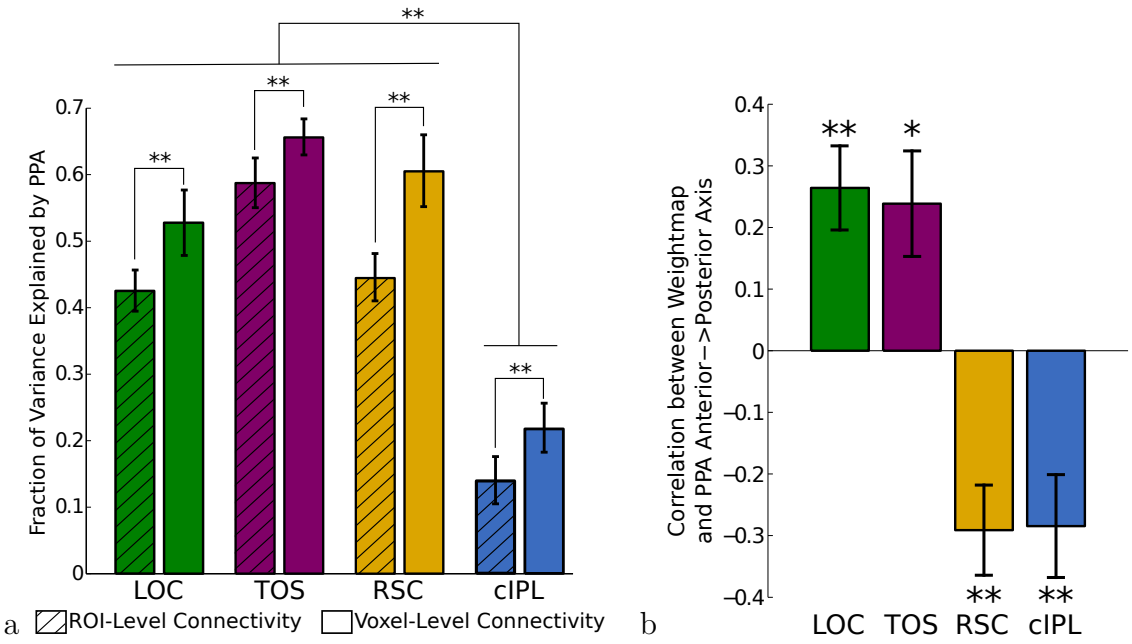


Figure 5.2: **A comparison of the learned PPA weightmaps and the overall connectivity strength, for our four ROIs.** (a) The timecourses of all four seed ROIs are better explained by a regularized voxel-level connectivity map in PPA, rather than a single connectivity weight for all of left and right PPA. Activity in LOC, TOS, and RSC is most closely related to PPA activity, while only a smaller amount of the cIPL timecourse is related to PPA activity. (b) To obtain a simple characterization of the learned maps, we compute the correlation between the connectivity weights and the anterior-posterior axis. This measure shows consistent differences between the four regions' connectivity maps. LOC and TOS are preferentially connected to posterior PPA (since their corresponding PPA weightmaps increase along the anterior to posterior axis) while RSC and cIPL are preferentially connected to anterior PPA. Error bars represent s.e.m. across subjects, * $p < 0.05$, ** $p < 0.01$.

lobe. Regions immediately anterior to PPA, including the hippocampus and anterior parahippocampal gyus, show preferential correlation with anterior PPA, but it is unclear if this effect is driven by intrinsic connectivity or local noise correlations. Coronal and axial slices are shown in Fig. 5.4, demonstrating that these connectivity patterns are bilaterally symmetric. This result can also be obtained by using only "resting" timepoints from between stimulus blocks or using a different value for λ, and is apparent for both the scene and object tasks (see Supplementary Fig. D4), suggesting that this connectivity pattern is intrinsic rather than task-specific. The fraction of variance explained for the searchlights is consistent with our ROI analysis, showing the strongest coupling between PPA and visual regions including LOC, TOS, and RSC (see Supplementary Fig. D5).

5.3.3 Scene- and Object-Sensitivity Analysis

Do these connectivity differences give rise to differences in functional response to stimulus categories? Although the functional roles of anterior and posterior PPA are likely complex, a simple functional anterior-posterior distinction can be seen in the scene and object responses during our localizer experiment. The selectivities of the PPA voxels to scenes and objects are shown in Fig. 5.5, binned based on position along the anterior-posterior axis. At the posterior side of PPA, the sensitivity to both scenes and objects is high, with nearly all voxels responding to scene stimuli and a majority of voxels responding to object stimuli. Moving posterior to anterior, scene selectivity decreases somewhat (average correlation between t-statistic and posterior-anterior axis of 0.25, $t_{10} = 3.00, p = 0.01$ two-tailed t-test after z-transform), although most voxels respond significantly to scene stimuli across all of PPA. Object sensitivity, however, substantially decreases (average correlation between t-statistic and posterior-anterior axis of 0.32, $t_{10} = 3.39, p < 0.01$ two-tailed t-test after z-transform), with a majority of voxels at the anterior edge showing no significant response to object stimuli.

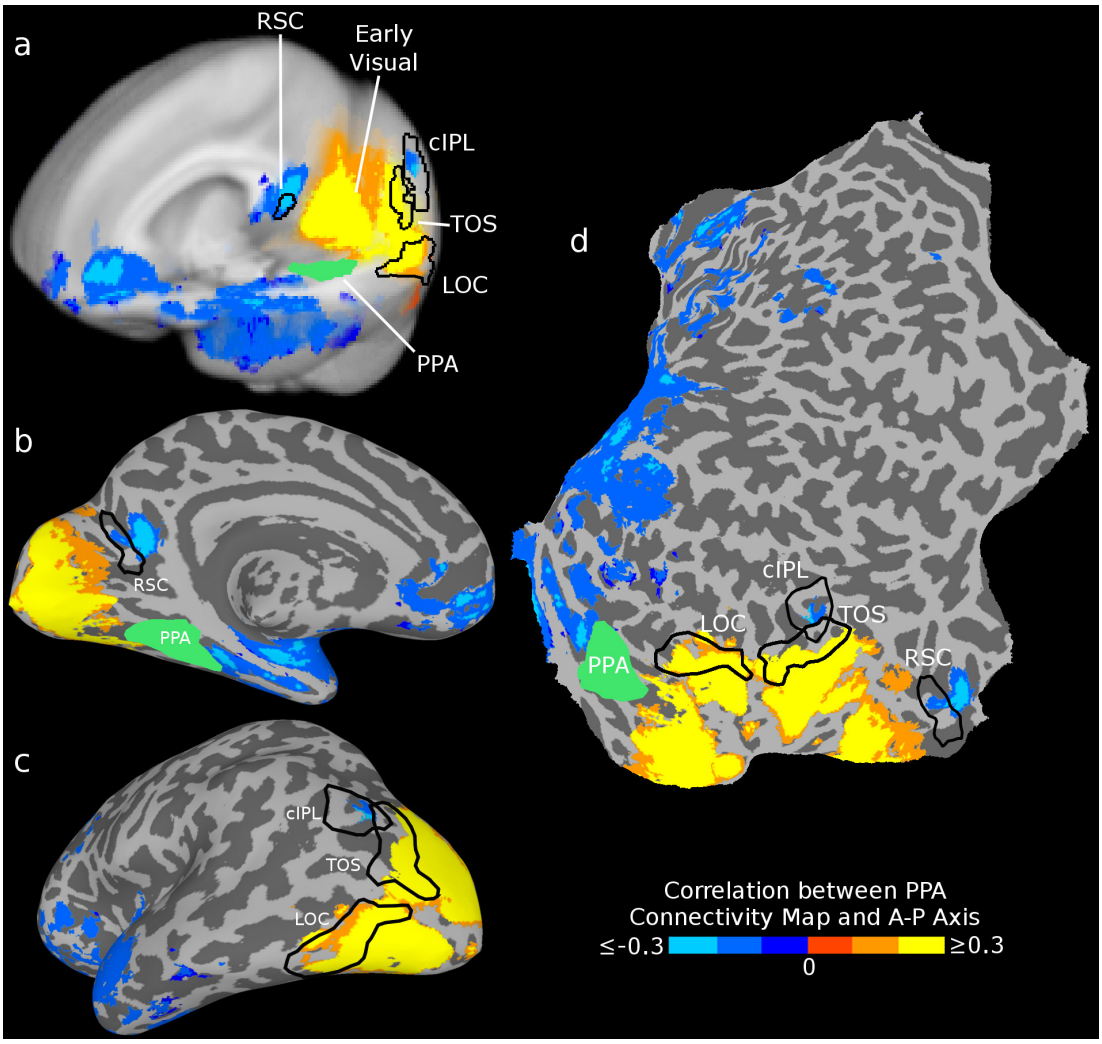


Figure 5.3: **Searchlight connectivity results.** (a) Rendering of the group connectivity bias map on the left hemisphere of the Talairach 452 brain. Colored voxels are those that showed highly significant (FDR <0.01, cluster size >300 mm^3) bias in anterior-posterior connectivity to PPA, computed as the correlation between the learned PPA connectivity map and the anterior-posterior axis. Bilateral areas RSC and cIPL, as well as ventral PFC and lateral anterior temporal regions, exhibited connectivity with anterior PPA (blue voxels), while occipital visual areas (including LOC and TOS) exhibited connectivity with posterior PPA (orange-yellow voxels). The borders of the group ROIs are shown for reference (outlining the location where at least 3 subjects' ROIs overlap). (b-d) The same connectivity map on an inflated surface and cortical flatmap .

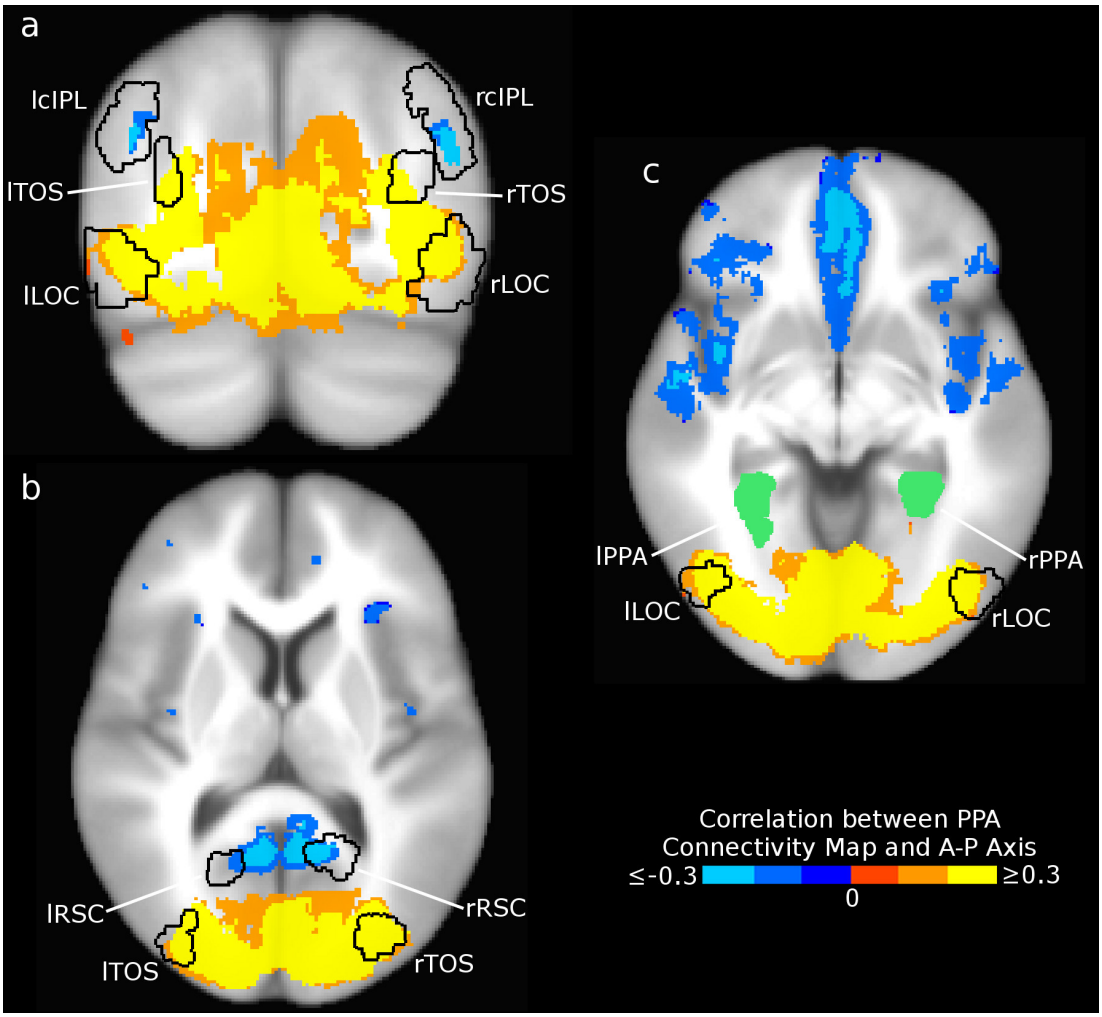


Figure 5.4: **Three slices of the group connectivity bias map.** Seed voxels for which the PPA connectivity map has a strong anterior-posterior gradient (FDR <0.01, cluster size >300 mm^3) are shown in blue (preferential connectivity to anterior PPA) and yellow (preferential connectivity to posterior PPA). (a) In this coronal slice (y=-73mm), we identify bilateral cIPL regions that show a different connectivity pattern from adjacent area TOS. (b) At z=10mm, we observe anterior PPA connectivity in RSC, as well as posterior PPA connectivity in TOS and early visual visual areas. (c) At z=-5mm, ventral occipital areas including LOC show connectivity to posterior PPA. Additionally, anterior PPA connectivity can be seen in the frontal and anterior temporal lobes.

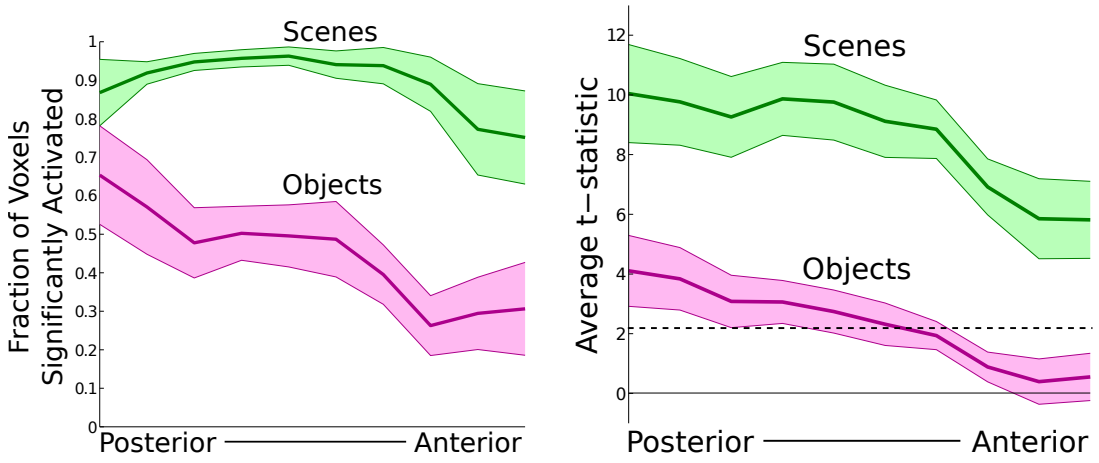


Figure 5.5: **Functional gradients across PPA.** The proportion of voxels responsive to scene and object stimuli, and the average t-statistic for the response to scene and object stimuli, were calculated in 10 bins along the anterior-posterior axis in each subject. The dotted line indicates the average t-statistic value corresponding to FDR=0.05 (across all subjects, for both stimulus categories). Scene sensitivity decreased from posterior to anterior PPA, but nearly all voxels across PPA responded significantly to scene stimuli. Object sensitivity substantially decreased from posterior to anterior PPA, with the majority of anterior PPA voxels failing to respond significantly to object stimuli. Error bars represent s.e.m. across subjects.

5.3.4 LOC/TOS vs. RSC/cIPL Connectivity

A number of studies have examined functional and connectivity gradients along the entire parahippocampal gyrus, which includes (in addition to PPA) a portion of parahippocampal cortex anterior to PPA, and the perirhinal cortex [5, 20, 176, 179, 271]. In order to examine how our gradients within PPA fit into the connectivity patterns of the broader medial temporal lobe, we searched for voxels which showed the same connectivity differences we observed within PPA. As shown in Fig. 5.6, the pattern of connectivity observed in anterior PPA (RSC and cIPL greater than LOC and TOS) extends anteriorly along the parahippocampal gyrus and into the hippocampus. The most anterior portion of the parahippocampal gyrus (around perirhinal cortex) did not show a connectivity pattern matching either anterior or posterior PPA, consistent with previous work on the connectivity properties of perirhinal cortex [176]. In general, the regions showing connectivity similar to anterior PPA overlap very well with the Default Mode Network [100] (see Supplementary Fig. D6).

5.4 Discussion

Our results demonstrate that human PPA exhibits a gradient in connectivity along the anterior-posterior axis analogous to the gradient in connectivity along macaque TH/TF/TFO. This connectivity gradient was also paired with a functional gradient of sensitivity to scene and abstract object stimuli. These results present a challenge to current models of PPA function which assume that PPA is functionally homogeneous, and demonstrate that anterior and posterior PPA connect differentially to two distinct cortical networks.

Note that, although our data suggest that PPA might contain identifiable subregions, these subregions should not be considered as completely independent modules. Both subregions activate selectively to scenes, and the parahippocampal region (at least in macaques) is densely self-connected [274], implying that these subregions cooperate to build a complete representation of a scene. Their distinct connectivity properties, however, do suggest that each may be involved in specific aspects of visual

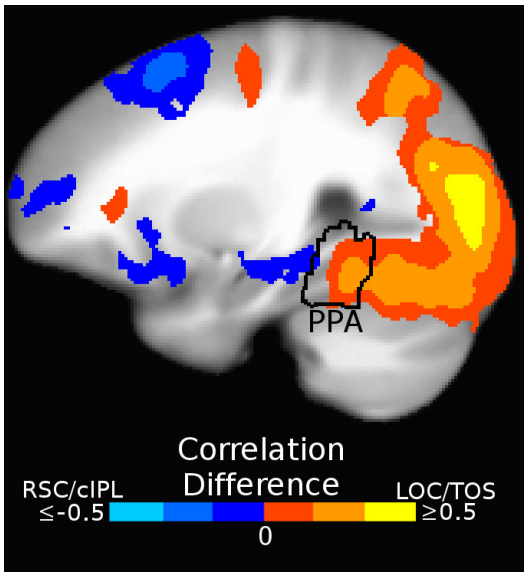


Figure 5.6: **Regions throughout cortex showing connectivity differences similar to anterior and posterior PPA.** In this sagittal slice (x=-26), colored voxels are those showing significantly (FDR <0.05, cluster size >1000 mm^3) different connectivity to LOC and TOS versus RSC and cIPL. The connectivity pattern in anterior PPA extends anteriorly along the parahippocampal gyrus and into the hippocampus. The connectivity patterns over the entire surface are shown in Supplementary Fig. D6.

and cognitive processing involved in the overarching goal of scene understanding. We discuss some possibilities for the functional roles of the subregions below.

5.4.1 Posterior PPA

Posterior PPA shows a stronger response to abstract objects than anterior PPA, and is more strongly connected to all of occipital visual cortex, including LOC and TOS. These regions have well-defined retinotopic maps [9, 209], and are associated with the perception of low-level visual features and object shape. Previous work has hinted that posterior PPA is more responsive to both simple visual textures and objects; Arcaro et al. [9] found that a posterior portion of PPA responded about four times as strongly to a flickering checkerboard stimulus compared to an anterior portion, and that the response to objects was greater than the response to scrambled images only in the posterior portion. In other words, posterior PPA may be more visually responsive than anterior PPA.

Posterior PPA may be specifically tuned to visual features in the high spatial frequency band; PPA has been shown to respond preferentially to higher spatial frequencies, and this effect tended to be strongest at the posterior end of PPA [230]. High-frequency edges could be the most important visual features for understanding the structure of a scene and navigating through it [230, 301]. Alternatively, this high-frequency preference could be related to the perception of large, landmark-like objects. A comparison of the Fourier spectra of 400 objects with either large or small real-world size (but matched visual size) found that larger objects tend to have more power at high spatial frequencies, especially along horizontal and vertical orientations; intuitively, larger objects are "boxier" while smaller objects are rounder [161]. A region of cortex selective for large objects has been shown to overlap with about half of PPA near the collateral sulcus, possibly corresponding to posterior PPA [162].

It is also possible that posterior PPA performs texture and ensemble processing, since these tasks tend to activate cortical regions around the collateral sulcus, overlapping with posterior PPA [49–51]). Although Cant and Xu [51] failed to find an

anterior-posterior difference within PPA for ensemble or texture processing by splitting PPA along the center of activation, a more sensitive voxel-level measure could potentially reveal such a gradient.

Our description of posterior PPA is in fact similar to the original proposal of Aguirre et al. for a "lingual landmark area" (LLA), slightly posterior to PPA, which was "specialized for the perception of visual stimuli with orienting value" [2] and carried out bottom-up perceptual analysis to recognize locations or landmarks [49, 89]. Although the LLA is no longer identified as an independent region from PPA in current studies, it is possible that the posterior portion of PPA corresponds to the properties of the proposed LLA, offering an explanation for why this region was localized more posteriorly than the full PPA.

5.4.2 Anterior PPA

Anterior PPA is specifically connected to RSC, cIPL, medial PFC, and the lateral surface of the anterior temporal lobe. In addition, this portion of PPA is less visually responsive to both scenes and objects, with notably low sensitivity to abstract object stimuli.

The set of regions connected more strongly to anterior PPA is strikingly similar to the Default Mode Network (DMN) [38, 100, 229], which is known to include the parahippocampal region. The only portions of the DMN that do not show differential connectivity to anterior PPA in our data are the rostral portion of the posterior cingulate cortex (PCC) and the superior frontal cortex. It is likely that PPA does not have direct connections to these regions; a DTI study showed that the medial temporal lobe directly connects only to RSC, rather than more rostral PCC [116], and a functional connectivity analysis showed that the hippocampal formation (including the parahippocampal region) is connected only to the more ventral portion of the prefrontal cortex, not the dorsal portion [38]. We confirmed that RSC and cIPL showed connectivity with the entire DMN (see Supplementary Fig. D6), suggesting that anterior PPA does have indirect connections to PCC and superior frontal cortex.

Although the DMN has been implicated in a large number of internally-focused

tasks, one of its key roles involves autobiographical memory [38]. Models of recognition memory have previously identified parahippocampal cortex as primarily encoding spatial context information [84], and data from Aminoff et al. [5] has suggested that an anterior portion of PPA may be involved in recall based on spatial context. Our results are consistent with anterior PPA playing a more central role in memory than posterior PPA, given anterior PPA's connectivity with the DMN.

PPA is known to represent global scene properties such as spatial expanse [166, 218] and to construct global scene representations which are not predictable from responses to signature objects [183]. Anterior PPA's connectivity to cIPL and RSC, along with its lower sensitivity to abstract objects, suggest that it may be more concerned with these types of spatial and non-object-based scene properties than posterior PPA. Future research contrasting global and object-based properties of scenes, however, would be necessary to test such a hypothesis.

The fact that anterior PPA had a lower sensitivity to our abstract object stimuli does not necessarily imply that this region does not use object information. Previous work has shown that PPA responds to objects that have spatial associations [5], are space-defining [204], and are navigationally-relevant [142]. These types of responses require spatial memory and cannot be based purely on visual features like object shape. If anterior PPA is involved in processing spatial context, then space-defining or navigationally-relevant objects could activate anterior PPA more strongly than our abstract objects, which were unfamiliar and provided no sense of context or orientation. Further experiments will be required to determine whether what type of object-related information is used in this region.

5.4.3 Homology with TH/TF/TFO

Given the close match between the connectivity gradients in macaque parahippocampal cortex and those in PPA, can we identify a precise correspondence between macaque regions TH/TF/TFO and our PPA subregions? Since the connectivity gradients extend anteriorly beyond PPA (which terminates in the most posterior part of the parahippocampal gyrus), a possible homology could identify posterior PPA with

TFO, anterior PPA with TF, and the anterior portion of parahippocampal cortex with TH. This labeling would be consistent with previous work showing that TFO is more visually responsive than TF and may have a coarse retinotopy [251], matching the properties of posterior PPA. This correspondence will only be definitively confirmed, however, if future electrophysiological measurements show that TH does not respond to scene stimuli (placing it anterior to anterior PPA) while TF and TFO do.

5.4.4 Implications for Future Work on PPA

Unraveling the functions of the PPA has proven to be a challenging problem, given the region's involvement in a variety of scene perception and navigation tasks [91]. Our results imply that a complete model of PPA's functional properties must account for the differences in connectivity and function between anterior and posterior PPA. Although the precise roles of PPA's subregions are yet to be determined, our results and previous work suggest that posterior PPA is concerned primarily with perception of low-level visual features and object shape, while anterior PPA is involved in memory and global contextual processing. Given the relatively small size of each of these subregions, voxel-level approaches (such as our connectivity method) as well as high-resolution fMRI imaging may be required to identify the representations evoked within the parts of PPA, and understand how these regions cooperate to build a coherent scene representation.

5.5 Conclusions

Our connectivity findings call into question the traditional view of PPA as a homogeneous region performing a single functional role, and provide a starting point for future experimental and modeling work investigating how different types of cortical networks interact for scene understanding and recognition. This discovery was made possible by our voxel-level functional connectivity approach, which may prove fruitful for uncovering subregions in other cortical systems.

5.6 Acknowledgments

We thank the Richard M. Lucas Center for Imaging, two anonymous reviewers, and Audrey Lustig (for the data from the Scene Category experiment). This work is funded by National Institutes of Health Grant 1 R01 EY019429 (to L.F.-F. and D.M.B.) and a National Science Foundation Graduate Research Fellowship under Grant No. DGE-0645962 (to C.B.).

Chapter 6

Parcellating connectivity in spatial maps

A common goal in biological sciences is to model a complex web of connections using a small number of interacting units. We present a general approach for dividing up elements in a spatial map based on their connectivity properties, allowing for the discovery of local regions underlying large-scale connectivity matrices. Our method is specifically designed to respect spatial layout and identify locally-connected clusters, corresponding to plausible coherent units such as strings of adjacent DNA base pairs, subregions of the brain, animal communities, or geographic ecosystems. Instead of using approximate greedy clustering, our nonparametric Bayesian model infers a precise parcellation using collapsed Gibbs sampling. We utilize an infinite clustering prior that intrinsically incorporates spatial constraints, allowing the model to search directly in the space of spatially-coherent parcellations. After showing results on synthetic datasets, we apply our method to both functional and structural connectivity data from the human brain. We find that our parcellation is substantially more effective than previous approaches at summarizing the brain's connectivity structure using a small number of clusters, produces better generalization to individual subject data, and reveals functional parcels related to known retinotopic maps in visual cortex. Additionally, we demonstrate the generality of our method by applying the same model to human migration data within the United States. This analysis reveals

that migration behavior is generally influenced by state borders, but also identifies regional communities which cut across state lines. Our parcellation approach has a wide range of potential applications in understanding the spatial structure of complex biological networks. This chapter is joint work with Diane M. Beck and Fei-Fei Li, and appeared previously in PeerJ [16].

6.1 Introduction

When studying biological systems at any scale, scientists are often interested not only in the properties of individual molecules, cells, or organisms, but also in the web of *connections* between these units. The rise of massive biological datasets has enabled us to measure these second-order interactions more accurately, in domains ranging from protein-protein interactions, to neural networks, to ecosystem food webs. We can often gain insight into the overall structure of a connectivity graph by grouping elements into clusters based on their connectivity properties. Many types of biological networks have been modeled in terms of interactions between a relatively small set of "modules" [22, 122], including protein-protein interactions [239], metabolic networks [232], bacterial co-occurrence [101], pollination networks [214], and food webs [165]. In fact, it has been proposed that modularity may be a necessary property for any network that must adapt and evolve over time, since it allows for reconfiguration [4, 122]. There are a large number of methods for clustering connectivity data, such as k-means [106, 157, 172], Gaussian mixture modeling [105], hierarchical clustering [68, 109, 205], normalized cut [133], infinite relational modeling [203], force-directed graph layout [71], weighted stochastic block modeling [3], and self-organized mapping [199, 312].

The vast majority of these methods, however, ignore the fact that biological networks almost always have some underlying spatial structure. As described by Legendre and Fortin: "In nature, living beings are distributed neither uniformly nor at random. Rather, they are aggregated in patches, or they form gradients or other kinds of spatial structures... the spatio-temporal structuring of the physical environment induces a similar organization of living beings and of biological processes, spatially as

well as temporally" [173]. In many biological datasets, we therefore wish to constrain possible clustering solutions to consist of *spatially-contiguous parcels*. For example, when dividing a DNA sequence into protein-coding genes, we should enforce that the genes are contiguous sequences of base pairs. Similarly, if we want to identify brain regions that could correspond to local cortical modules, we need each discovered cluster to be a spatially-contiguous region. Without spatial information, the discovered clusters may be difficult to interpret; for example, clustering functional brain connectivity data without spatial information yields spatially-distributed clusters that confound local modularity and long-distance interactions [172].

The problem is thus to a parcellate a spatial map into local, contiguous modules such that all elements in a module have the same connectivity properties (Fig. 6.1). In this paper we present the first general solution to this problem, using a generative probabilistic model to parcellate a spatial map into local regions with connectivity properties that are as uniform as possible. Scientific insights can be gained from both the clusterings themselves (which identify the local spatial sources of the interaction matrix) as well as the connections between the parcels, which summarize the original complex connectivity matrix. Our method yields better results than other approaches such as greedy clustering, and can help to determine the correct number of parcels in a data-driven way.

One of the most challenging spatial parcellation problems is in the domain of neuroscience. Modern human neuroimaging methods can estimate billions of connections between different locations in the brain, with complex spatial structures that are highly nonuniform in size and shape. Correctly identifying the detailed boundaries between brain regions is critical for understanding distributed neural processing, since even small inaccuracies in parcellation can yield major errors in estimating network structure [266].

Obtaining a brain parcellation with spatially coherent clusters has been difficult, since it is unclear how to extend standard clustering methods to include the constraint that only adjacent elements should be clustered together. Biasing the connectivity matrix to encourage local solutions can produce local parcels in some situations [64,

285], or distributed clusters can be split into their connected components after clustering [1], but these approximations will not necessarily find the best parcellation of the original connectivity matrix. It is also possible to add a Markov Random Field prior (such as the Ising model) onto a clustering model to encourage connected parcels [143, 250], but in practice this does not guarantee that clusters will be spatially connected [137].

Currently, finding spatially-connected parcels is often accomplished using agglomerative clustering [28, 131, 202, 282], which iteratively merges neighboring elements based on similarity in their connectivity maps. There are a number of disadvantages to this approach; most critically, the solution is only a greedy approximation (only a single pass over the data is made, and merged elements are never unmerged), which as will be shown below can lead to poor parcellations when there is a high level of noise. Edge detection methods [66, 110, 310] define cluster boundaries based on sharp changes in connectivity properties, which are also sensitive to localized patches of noisy data. Spectral approaches such as normalized cut [70] attempt to divide the spatial map into clusters by maximizing within-cluster similarity and between-cluster dissimilarity, but this approach has a strong bias to choose clusters that all have similar sizes [28]. It is also possible to incorporate a star-convexity prior into an MRF to efficiently identify connected parcels [137]. This approach, however, constrains clusters to be convex (in connectivity space); as will be shown below, our method finds structures in real datasets violating this assumption, such as nested regions in functional brain connectivity data. All of these methods require explicitly setting the specific number of desired clusters, and are optimizing a somewhat simpler objective function; they seek to maximize the similarity between the one-dimensional rows or columns of the connectivity matrix, while our method takes into account reordering of the both the rows and columns to make the between-parcel 2D connectivity matrix as simple as possible.

Our model is highly robust to noise, has no constraints on the potential sizes and shapes of brain regions, and makes many passes over the data to precisely identify region boundaries. We validate that our method outperforms previous approaches on synthetic datasets, and then show that we can more efficiently summarize both

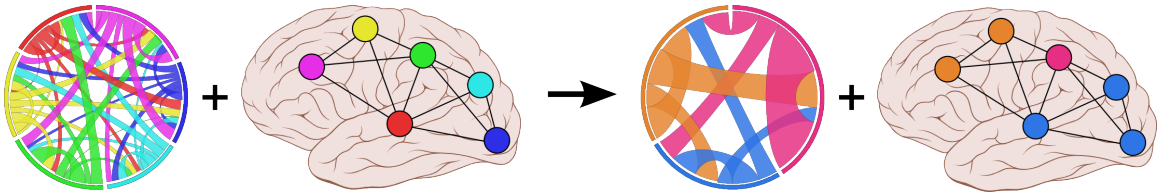


Figure 6.1: **Parcellating connectivity in spatial maps.** Given a set of elements arranged on a spatial map (such as points within the human cortex) as well as the connectivity between each pair of elements, our method finds the best parcellation of the spatial map into connected clusters of elements that all have similar connectivity properties. Brain image by Patrick J. Lynch, licensed under CC BY 2.5.

functional and structural brain connectivity data. Our parcellation of human cortex generalizes more effectively across subjects, and reveals new structure in the functional connectivity properties of visual cortex.

To demonstrate the wide applicability of our method, we apply the same model to find spatial patterns in human migration patterns within the United States. Despite the fact that this is an entirely different type of data at a different spatial scale, we are able to find new insights into how state borders shape migratory behavior. Our results on these diverse datasets suggest that our analysis could have a wide range of potential applications in understanding biological networks. It is also important to note that the "spatial adjacency" constraint of our method could also be used for other, nonspatial notions of adjacency; for example, clustering an organism's life into contiguous temporal segments based on its changing social interactions.

6.2 Materials and Methods

6.2.1 Probabilistic Model

Intuitively, we wish to find a parcellation $\mathbf{z}$ which identifies local regions, such that all elements in a region have the same connectivity "fingerprint." Specifically, for any two parcels m and n, all pairwise connectivities between an element in parcel m and an element in parcel n should have a similar value. Our method uses the full

distribution of all pairwise connectivities between two parcels, and finds a clustering for which this distribution is highly peaked. This makes our method much more robust than approaches which greedily merge similar clusters [28, 282] or define parcel edges where neighboring voxels differ [110, 281, 310]. The goal of identifying modules with similar connectivity properties is conceptually similar to weighted stochastic block models [3], but it is unclear how these models could be extended to incorporate the spatial-connectivity constraint.

We would like to learn the number of regions automatically from data, and additionally impose the requirement that all regions must be spatially-connected. We can accomplish both goals more efficiently in a single framework, by using an infinite clustering prior on our parcellation $\mathbf{z}$ which simultaneously constrains regions to be spatially coherent and does not limit the number of possible clusters. Specifically, since the mere existence of a element (even with unknown connectivity properties) changes the spatial connectivity and thus affects the most likely clustering, we must employ a nonparametric prior which is *not marginally invariant*. Other Bayesian nonparametric models allow for spatial dependencies between datapoints, but the only class of CRPs which is not marginally invariant is the distance-dependent Chinese Restaurant Process (dd-CRP) [27]. Instead of directly sampling a label for each element, the dd-CRP prior assigns each element i a link to a neighboring element c_i. The actual parcel labels $\mathbf{z}(\mathbf{c})$ are then defined implicitly as the undirected connected components of the link graph. Intuitively, this allows for changes in the labels of many elements when a single connection c_i is modified, since it may break apart or merge together two large connected sets of elements. Additionally, this construction allows the model to search freely in the space of parcel links $\mathbf{c}$, since every possible setting of the parcel links corresponds to a parcellation satisfying the spatial-coherence constraint.

Mathematically, our generative clustering model is:

$$\begin{aligned}
\mathbf{c} &\sim \text{dd-CRP}(\alpha, f) \\
A_{mn}, \sigma^2_{mn} &\sim \text{Normal-Inverse-}\chi^2(\mu_0, \kappa_0, \sigma^2_0, \nu_0) \\
D_{ij}|\mathbf{z}(\mathbf{c}) &\sim \text{Normal}(A_{\mathbf{z}(\mathbf{c})_i \mathbf{z}(\mathbf{c})_j}, \sigma^2_{\mathbf{z}(\mathbf{c})_i \mathbf{z}(\mathbf{c})_j})
\end{aligned}$$

For N elements and K parcels: $\mathbf{c}$ is a vector of length N which defines the cluster links for all elements (producing a region labeling vector $\mathbf{z}(\mathbf{c})$ of length N, taking values from 1 to K); α and f are the scalar hyperparameter and $N \times N$ distance function defining the dd-CRP; $\mathbf{A}$ and $\boldsymbol{\sigma^2}$ are the $K \times K$ connectivity strength and variance between regions; μ_0 and κ_0 are the scalar prior mean and precision for the connectivity strength; σ^2_0 and ν_0 are the scalar prior mean and precision for the connectivity variance; and $\mathbf{D}$ is the $N \times N$ observed connectivity between individual elements.

The probability of choosing a particular c_i in the dd-CRP is defined by a distance function f; we use $f_{ij} = 1$ if i and j are neighbors, and 0 otherwise, which guarantees that all clusters will be spatially connected. A hyperparameter α controls the probability that a voxel will choose to link to itself. Note that, due to our choice of distance function f, a random partition drawn from the dd-CRP can have many clusters even for $\alpha = 0$, since elements are only locally connected.

The connectivity strength A_{mn} and variance σ^2_{mn} between each pair of clusters m and n is given by a Normal-Inverse-χ^2 (NIχ^2) distribution, and the connectivity D_{ij} between every element i in one region and j in the other is sampled based on this strength and variance. The conjugacy of the Normal-Inverse-χ^2 and Normal distributions allows us to collapse over A_{mn} and σ^2_{mn} and sample only the clustering variables c_i. Empirically, we find that the only critical hyperparameter is the expected variance σ^2_0, with lower values encouraging parcels to be smaller (we set $\alpha = 10, \mu_0 = 0, \kappa_0 = 0.0001, \nu_0 = 1$ for all experiments).

To allow the comparison of hyperparameter values between problems with the same number of elements (e.g. the functional and structural datasets), we normalize the input matrix D to have zero mean and unit variance. We then initialize the

model using the Ward clustering (see below) with the most likely number of clusters under our model, and setting the links $\mathbf{c}$ to form a random spanning tree within each cluster.

In summary, we have introduced a novel connectivity clustering model which (a) uses the full distribution of connectivity properties to define the parcellation likelihood, and (b) employs an infinite clustering model which automatically chooses the number of parcels and enforces that parcels be spatially-connected.

6.2.2 Derivation of Gibbs Sampling Equations

To infer a maximum a posteriori (MAP) parcellation $\mathbf{z}$ based on the dd-CRP prior, we perform collapsed Gibbs sampling on the element links $\mathbf{c}$. A link c_i for element i is drawn from

$$\begin{aligned} p(c_i^{(new)}|\mathbf{c}_{-\mathbf{i}}, D) &\propto p(c_i^{(new)})p(D|\mathbf{z}(\mathbf{c}_{-\mathbf{i}} \cup c_i^{(new)})) = p(c_i^{(new)})p(D|\mathbf{z}^{(new)}) \\ &\propto \left\{ \begin{array}{ll} \alpha & \text{if } c_i^{(new)} = i \\ 1 & \text{else} \end{array} \right\} \prod_{k_1,k_2=1}^{|\mathbf{z}^{(new)}|} p(D_{z_{k_1}^{(new)}, z_{k_2}^{(new)}}) \end{aligned} \tag{6.1}$$

To compare the likelihood term for different choices of $c_i^{(new)}$, we first remove the current link c_i, giving the induced partition $\mathbf{z}(\mathbf{c}_{-\mathbf{i}})$ (which may split a region). If we resample c_i to a self-loop or to a neighbor j that does not join two regions, the likelihood term is based on the partition $\mathbf{z}(\mathbf{c}_{-\mathbf{i}}) = \mathbf{z}$. Alternatively, c_i can be resampled to a neighbor j such that two regions K' and K'' in $\mathbf{z}(\mathbf{c}_{-\mathbf{i}})$ are merged into one region K in $\mathbf{z}(\mathbf{c}_{-\mathbf{i}} \cup c_i^{(new)}) = \hat{\mathbf{z}}$. Numbering the regions so that $z_i \in \{1 \cdots (K-1), K', K''\}$ and $\hat{z}_i \in \{1 \cdots (K-1), K\}$ gives

$$\frac{p(D|\hat{\mathbf{z}})}{p(D|\mathbf{z})} = \frac{\prod_{k=1}^{K} p(D_{\hat{z}_k,\hat{z}_K}) \prod_{k=1}^{K-1} p(D_{\hat{z}_K,\hat{z}_k})}{\prod_{k=1}^{K'} p(D_{z_k,z_{K'}}) \prod_{k=1}^{K''} p(D_{z_k,z_{K''}}) \prod_{k=1}^{K-1} p(D_{z_{K'},z_k}) \prod_{k=1}^{K'} p(D_{z_{K''},z_k})} \tag{6.2}$$

Each term $p(D_{z_m,z_n})$ is a marginal likelihood of the NIχ^2 distribution, and can be

computed in closed form as shown in [206]:

$$p(D_{z_m,z_n}) = \frac{\Gamma(\nu_{mn}/2)}{\Gamma(\nu_0/2)} \left(\frac{\kappa_0}{\kappa_{mn}}\right)^{\frac{1}{2}} \frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_{mn}\sigma_{mn}^2)^{\nu_{mn}/2}} (\pi)^{-n/2}$$

$$L = |z_m||z_n| \qquad \kappa_{mn} = \kappa_0 + L \qquad \nu_{mn} = \nu_0 + L$$

$$\bar{d} = \frac{1}{L} \sum_{\substack{i \in z_m \\ j \in z_n}} D_{ij} \qquad s = \sum_{\substack{i \in z_m \\ j \in z_n}} (D_{ij} - \bar{d})^2 \qquad \mu_{mn} = \frac{\kappa_0\mu_0 + L\bar{d}}{\kappa_{mn}}$$

$$\sigma_{mn}^2 = \frac{1}{\nu_{mn}} (\nu_0\sigma_0^2 + s + \frac{L\kappa_0}{\kappa_0 + L}(\mu_0 - \bar{d})^2)$$

Intuitively, eq. 6.2 computes the probability of merging or splitting two regions at each step based on whether the connectivities between these regions' elements and the rest of the regions are better fit by one distribution or two.

In practice, the time-consuming portion of each sampling iteration is computing the sum of squared deviations s. This can be made more efficient by computing the s values for the merged $\hat{\mathbf{z}}$ in closed form. Given that the connectivities $D_{K'} = \{D_{iK'}\}_{i \in k}$ between parcel k and K' have sum of squares deviations $s_{K'}$ and mean $\bar{d}_{K'}$, and similarly for K'', then the sum of squares s_K for the connectivities between parcel k and the merged parcel K (merging K' and K'') is:

$$
\begin{aligned}
s_K &= \sum_{d \in D_{K'} \cup D_{K''}} (d - \bar{d})^2 \\
&= \left(\sum_{d \in D_{K'} \cup D_{K''}} d^2 \right) - (|D_{K'}| + |D_{K''}|) \cdot \left(\frac{|D_{K'}| \cdot \bar{d}_{K'} + |D_{K''}| \cdot \bar{d}_{K''}}{|D_{K'}| + |D_{K''}|} \right)^2 \\
&= \left(\sum_{d \in D_{K'} \cup D_{K''}} d^2 \right) - \frac{|D_{K'}|^2}{|D_{K'}| + |D_{K''}|} \bar{d}^2_{K'} - \frac{|D_{K''}|^2}{|D_{K'}| + |D_{K''}|} \bar{d}^2_{K''} - \\
&\quad 2 \frac{|D_{K'}||D_{K''}|}{|D_{K'}| + |D_{K''}|} \bar{d}_{K'} \bar{d}_{K''} \\
&= \left(\sum_{d \in D_{K'}} d^2 - |D_{K'}| \bar{d}^2_{K'} \right) + \left(\sum_{d \in D_{K''}} d^2 - |D_{K''}| \bar{d}^2_{K''} \right) + \\
&\quad \frac{|D_{K'}||D_{K''}|}{|D_{K'}| + |D_{K''}|} \left(\bar{d}^2_{K'} + \bar{d}^2_{K''} - 2 \bar{d}_{K'} \bar{d}_{K''} \right) \\
&= s_{K'} + s_{K''} + \frac{|D_{K'}||D_{K''}|}{|D_{K'}| + |D_{K''}|} (\bar{d}_{K'} - \bar{d}_{K''})^2
\end{aligned}
$$

6.2.3 Comparison Methods

In order to evaluate the performance of our model, we compared our results to those of four existing methods. All of them require computing a dissimilarity measure between the connectivity patterns of elements i and j. For a connectivity matrix D,

$$
W_{i,j} = \sqrt{\sum_{a \neq i,j} (D_{i,a} - D_{j,a})^2 + \sum_{a \neq i,j} (D_{a,i} - D_{a,j})^2} \tag{6.3}
$$

"Local similarity" computes the edge dissimilarity $W_{i,j}$ between each pair of neighboring elements, and then removes all edges above a given threshold. Here we set the threshold in order to obtain a desired number of clusters. This type of edge-finding approach has been used extensively for neuroimaging parcellation [66, 110, 310]. Additionally, this is equivalent to using a spectral clustering approach [281] if clustering in the embedding space is performing using single-linkage hierarchical clustering.

"Normalized cut" computes the edge similarity $S_{i,j} = 1/W_{i,j}$ between each pair of neighboring elements, then runs the normalized cut algorithm of [263]. This draws partitions between elements a and b when their edge similarity $S_{a,b}$ is low relative to their similarities with other neighbors. Although computing the globally optimal normalized cut is NP-complete, an approximate solution can be found quickly by solving a generalized eigenvalue problem. This approach has been specifically applied to neuroimaging data [70].

"Region growing" is based on the approach described in [28]. First, a set of seed points is selected which have high similarity to all their neighbors, since they are likely to be near the center of parcels. Seeds are then grown by iteratively adding neighboring elements with high similarity to the seed. Once every element has been assigned to a region, Ward clustering (see below) was used to cluster adjacent regions until the desired number of regions is reached.

"Ward clustering" requires computing $W_{i,j}$ between all pairs of elements (not just neighboring elements). Elements are each initialized as a separate cluster, and neighboring clusters are merged based on Ward's variance-minimizing linkage rule [307]. This approach has been previously applied to neuroimaging data [86, 282].

We also compared to random clusterings. Starting with each element in its own cluster, we iteratively picked a cluster uniformly at random and then merged it with a neighboring cluster (also picked uniformly at random from all neighbors). The process continued until the desired number of clusters remained.

6.2.4 Synthetic Data

To generate synthetic connectivity data, we created three different parcellation patterns on an $18x18$ grid (see Fig. 6.2), with the number of regions $K = 5, 6, 9$. Each element of the KxK connectivity matrix A was sampled from a standard normal distribution. For a given noise level σ, the connectivity value $D_{i,j}$ between element i in cluster $\mathbf{z_i}$ and element j in cluster $\mathbf{z_j}$ is sampled from a normal distribution with mean $A_{\mathbf{z_i},\mathbf{z_j}}$ and standard deviation σ. This data matrix was then input to our method with $\sigma_0^2 = 0.01$, which returned the MAP solution after 30 passes through the elements

(approximately 10,000 steps). Both our method and all comparison methods were run for 20 different synthetic datasets for each noise level σ and the results were averaged.

Parcellations were evaluated by calculating their normalized mutual information (NMI) with the ground truth labeling. We calculate NMI as in [273]. This measure ranges from 0 to 1, and does not require any explicit "matching" between parcels. For N total elements, if $\mathbf{z}$ assigns n_h elements to cluster h, $\mathbf{z_{gt}}$ assigns n_l^{gt} elements to cluster l, and $n_{h,l}$ elements are assigned to cluster h by $\mathbf{z}$ and cluster l by $\mathbf{z_{gt}}$, this is given by

$$\mathrm{NMI}(\mathbf{z}, \mathbf{z_{gt}}) = \frac{I(\mathbf{z}, \mathbf{z_{gt}})}{\sqrt{H(\mathbf{z})H(\mathbf{z_{gt}})}} = \frac{\sum_h \sum_l n_{h,l} \log(N n_{h,l}/(n_h n_l^{gt}))}{\sqrt{\left(\sum_h n_h \log(n_h/N)\right)\left(\sum_l n_l^{gt} \log(n_l^{gt}/N)\right)}} \quad (6.4)$$

6.2.5 Human Brain Functional Data

We utilized group-averaged resting-state functional MRI correlation data from 468 subjects, provided by the Human Connectome Project's 500 Subjects release [291]. Using a specialized Siemens 3T "Connectome Skyra" scanner, data was collected during four 15-minute runs, during which subjects fixated with their eyes open on a small cross-hair. A multiband sequence was used, allowing for acquisition of 2.0mm isotropic voxels at a rate of 720ms. Data for each subject was cleaned using motion regression and ICA+FIX denoising [252, 264] and then combined across subjects using an approximate group-PCA method yielding the strongest 4500 spatial eigenvectors [265]. The symmetric 59412 by 59412 functional connectivity matrix $D_{a,b}$ was computed as the correlation between the 4500-dimensional eigenmaps of voxels a and b. For each of $\sigma_0^2 = 2000, 3000, 4000, 5000$, we ran Gibbs Sampling for 10 passes (approximately 600,000 steps) to find the MAP solution. For comparison with individual subjects, we also computed functional connectivity matrices for the first 20 subjects with resting-state data in the 500 Subjects release.

The map of retinotopic regions in visual cortex was created by mapping the volume-based atlas from [304] onto the Human Connectome group-averaged surface.

6.2.6 Human Brain Structural Data

We obtained diffusion MRI data for 10 subjects from the Human Connectome Project's Q3 release [291]. This data was collected on the specialized Skyra described above, using a multi-shell acquisition over 6 runs. Probabilistic tractgraphy was performed using FSL [144], by estimating up to 3 crossing fibers with bedpostx (using gradient nonlinearities and a rician noise model) and then running probtrackx2 using the default parameters and distance correction. 2000 fibers were generated for each of the $1.7 \cdot 10^6$ white-matter voxels, yielding $3.4 \cdot 10^9$ total sampled tracks per subject (approximately 34 billion tracks in total). We assigned each of the endpoints to gray-matter voxels using the 32k/hemisphere Conte69 registered standard mesh distributed for each subject, discarding the small number of tracks that did not have both endpoints in gray matter (e.g. cerebellar or spinal cord tracks). Since we are using distance correction, the weight of a track is set equal to its length. In order to account for imprecise tracking near the gray matter border, the weight of a track whose two endpoints are closest to voxels a and b is spread evenly across the connection between a and b, the connections between a and b's neighbors, and the connections between a's neighbors and b. Since the gray-matter mesh has a correspondence between subjects, we can compute the group-average number of tracks between every pair of voxels. Finally, since connectivity strengths are known to have a lognormal distribution [188], we define the symmetric 59412 by 59412 structural connectivity matrix $D_{a,b}$ as the log group-averaged weight between voxels a and b. The hyperparameter σ_0^2 was set to 3000, and Gibbs Sampling was run for 10 passes (approximately 600,000 steps) to find the MAP solution.

6.2.7 Human Migration Data

We used the February 2014 release of the 2007-2011 county-to-county U.S. migration flows from the U.S. Census Bureau American Community Survey [290]. This dataset includes estimates of the number of annual movers from every county to every other county, as well as population estimates for each county. We restricted our analysis to the continential U.S. To reduce the influence of noisy measurements from small

counties, we preprocessed the dataset by iteratively merging the lowest-population county with its lowest-population neighbor (within the same state) until all regions contained at least 10000 residents. This process produced 2594 regions which we continue to refer to as "counties" for simplicity, though 306 cover multiple low-population counties. For visualization of counties and states, we utilized the KML Cartographic Boundary Files provided by the U.S. Census Bureau [289].

One major issue with analyzing this migration data is that counties have widely varying populations (even after the preprocessing above), making it difficult to compare the absolute number of movers between counties. We correct for this by normalizing the migration flows relative to chance flows driven purely by population. If we assume a chance distribution in which a random mover is found to be moving from county a to county b based purely on population, then the normalized flow matrix is

$$D_{a,b} = \frac{M_{a,b}}{\left(\sum_{i,j} M_{i,j}\right) \cdot \frac{P_a P_b}{\left(\sum_i P_i\right)^2}} \tag{6.5}$$

where $M_{i,j}$ is the absolute number of movers from county i to county j, and P_i is the population of county i. This migration connectivity matrix D is therefore a nonnegative, asymmetric matrix in which values less than 1 indicate below-chance migration, and values greater than 1 indicate above-chance migration. Setting $\sigma_0^2 = 10$, we ran Gibbs Sampling for 50 passes (approximately 130,000 steps) to find the MAP solution.

6.3 Results

6.3.1 Comparison on Synthetic Data

In order to understand the properties of our model and quantitatively compare it to alternatives on a dataset with a known ground truth, we performed several experiments with synthetic datasets. We compared against random parcellations (in which elements were randomly merged together) as well as four existing methods: local similarity, which simply thresholds the similarities between pairwise elements

(similar to [66, 110, 281, 310]); normalized cut [70] which finds parcels maximizing the within-cluster similarity and between-cluster difference; region growing [28], an agglomerative clustering method which selects stable points and iteratively merges similar elements; and Ward clustering [282], an agglomerative clustering method which iteratively merges elements to minimize the total variance. Since these methods cannot automatically discover the number of clusters, they (and the random clustering) are set to use the same number of clusters as inferred by our method. We varied the noise level of the synthetic connectivity matrix from low to high, and evaluated the learned clusters using the normalized mutual information with the ground truth, which ranges from 0 to 1 (with 1 indicating perfect recovery).

As shown in Fig. 6.2, our method identifies parcels that best match the ground truth, across all three datasets and all noise levels. The naive local similarity approach performs very poorly under even mild noise conditions, and becomes worse than chance for high noise levels (for which most parcellations consist of single noisy voxels). Normalized cut is competitive only when the ground-truth parcels are equally sized (matching results from [28]), and is near-chance in the other cases. Region growing is more consistent across datasets, but does not reach the performance of Ward clustering, which outperforms all methods other than ours. Our model correctly infers the number of clusters with moderate amounts of noise (using the same hyperparameters in all experiments), and finds near-perfect parcellations even at very high noise levels (see Fig. 6.2c).

6.3.2 Functional connectivity in the human brain

To investigate the spatial structure of functional connectivity in the human brain, we applied our model to data from the Human Connectome Project [291]. Combining data from 468 subjects, this symmetric 59412 by 59412 matrix gives the correlation between fMRI timecourses of every pair of vertices on the surface of the brain (at 2mm resolution) during a resting-state scan (in which subjects fixated on a blank screen). Using the anatomical surface models provided with the data, we defined vertices to be spatially adjacent if they were neighbors along the cortical surface.

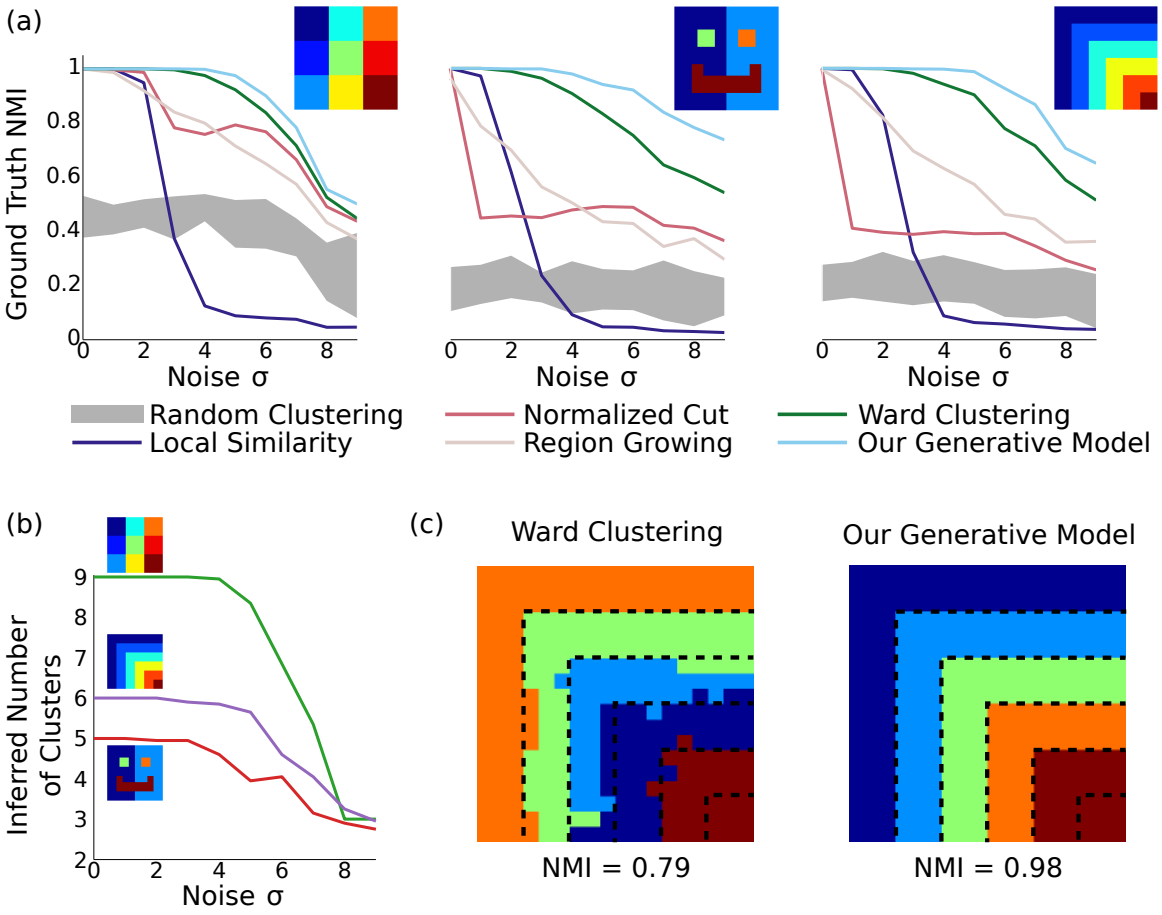


Figure 6.2: **Results on synthetic data.** (a) In three different synthetic datasets, our method is consistently better at recovering the ground-truth parcellation than alternative methods. This advantage is most pronounced when the parcels are arranged nonuniformly with unequal sizes, and the noise level is relatively high. Results are averaged across 20 random datasets for each noise level, and the gray region shows the standard deviation around random clusterings. (b) Our model can correctly infer the number of underlying clusters in the dataset for moderate levels of noise, and becomes more conserative about splitting elements into clusters as the noise level grows. (c) Example clusterings under the next-best clustering method and our model on the stripes dataset, for $\sigma = 6$. Although greedy clustering achieves a reasonable result, it is far noisier than the output of our method, which perfectly recovers the ground truth except for incorrectly merging the two smallest clusters.

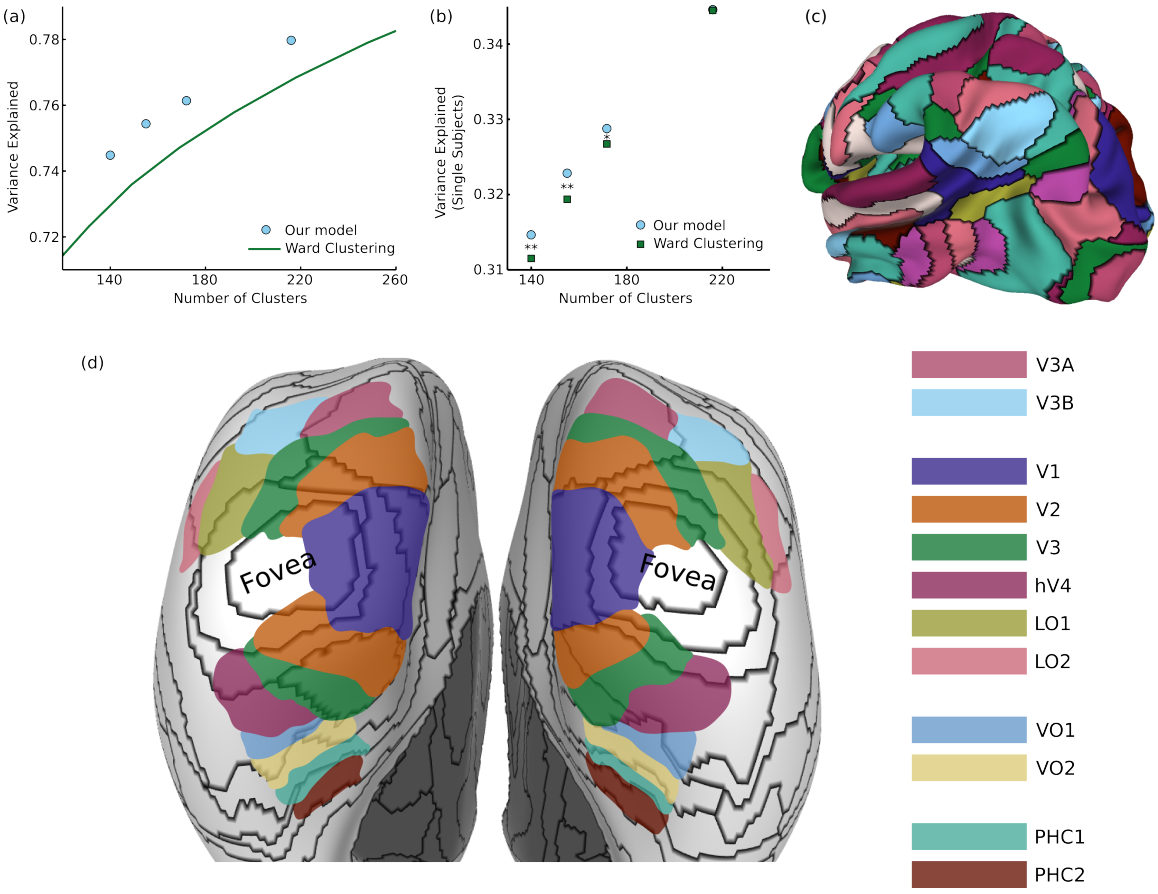


Figure 6.3: **Results on functional brain connectivity.** (a) Our model consistently provides a better fit to the data than greedy clustering, explaining the same amount of variance with 30 fewer clusters (different points were generated from different values of the hyperparameter σ_0^2). (b) When using our group-learned clustering to explain variance in 20 individual subjects, we consistently generalize better than the greedy clusters for cluster sizes less than 200 (* $p < 0.05$, ** $p < 0.01$). (c) A sample 172-cluster parcellation from our method. (d) Comparison between our parcels and retinotopic maps, showing a transition from eccentricity-based divisions to field map divisions.

Evaluating cortical parcellations is challenging since there is no clear ground truth for comparison, and different applications could require parcellations with different types of properties (e.g. optimizing for fitting individual subjects or for stability across subjects [282]). One simple measure of an effective clustering is the fraction of variance in the full 3.5 billion element matrix which is captured by the connectivity between parcels (consisting of only tens of thousands of connections). As shown in Fig. 6.3(a), our parcellation explains more variance for a given number of clusters than greedy Ward clustering; in order to achieve the same level of performance as our model, the simpler approach would require approximately 30 additional clusters. We can also measure how well this group-level parcellation (using data averaged from hundreds of subjects) fits the data from 20 individual subjects. Although the variance explained is substantially smaller for individual subjects, due both to higher noise levels and inter-subject connectivity differences, our model explains significantly more variance than Ward clustering with 140 clusters ($t_{19} = 2.97, p < 0.01$ one-tailed t-test), 155 clusters ($t_{19} = 3.67, p < 0.01$), or 172 clusters ($t_{19} = 1.77, p < 0.05$). The 220-cluster solutions from our model and Ward clustering generalize equally well, suggesting that our method's largest gains over greedy approximation occur in the more challenging regime of small numbers of clusters.

One part of the brain in which we do have prior knowledge about cortical organization is in visual cortex, which is segmented into well-known retinotopic field maps [304]. We can qualitatively examine the match between our 172-cluster parcellation (Fig. 6.3(c)) and these retinotopic maps on an inflated cortical surface, shown in Fig. 6.3(d). First, we observe a wide variety in the size and shape of the learned parcels, since the model places no explicit constraints on the clusters except that they must be spatially connected. We also see that we correctly infer very similar parcellations between hemispheres, despite the fact that bilateral symmetry is not enforced by the model. The earliest visual field maps (V1, V2, V3, hV4, LO1, LO2) all radiate out from a common representation of the fovea [34], and in this region, our model generates ring parcellations which divide the visual field based on distance from the fovea. The parcellation also draws a sharp border between peripheral V1 and V2. In the dorsal V3A/V3B cluster, V3A and V3B are divided into separate parcels. In

medial temporal regions, parcel borders show an approximate correspondence with known VO and PHC borders, with an especially close match along the PHC1-PHC2 border. Overall, we therefore see a transition from an eccentricity-based parcellation in the early visual cluster to a parcellation corresponding to known field maps in the later dorsal and ventral visual areas.

6.3.3 Structural connectivity in the human brain

Based on diffusion MRI data from the Human Connectome Project [291], we used probabilistic tractography [23] to generate estimates of the strength of the structural fiber connections between each pair of 2mm gray-matter voxels. Approximately 34 billion tracts were sampled across 10 subjects, yielding a symmetric 59412 by 59412 matrix in which about two-thirds of the elements are non-zero. Applying our method to this matrix parcellates the brain into groups of voxels that all had the same distribution of incident fibers. This problem is even more challenging than in the functional case, since this matrix is much less spatially smooth.

Fig. 6.4(a) shows a 190-region parcellation. Our clustering outperforms greedy clustering by an even larger margin than with the functional data, explaining as much variance as a greedy parcellation with 55 additional clusters. Fig. 6.4(b) also shows how the model fit evolves over many rounds of Gibbs sampling, when initialized with the greedy solution. Since our method can flexibly explore different numbers of clusters, it is able (unlike a greedy method) to perform complex splitting and merging operations on the parcels. Qualitatively evaluating our parcellation is even more challenging than in the previous functional experiment, but we find that our parcels match the endpoints of major known tracts. For example, Fig. 6.4(c) shows 35,000 probabilistically-sampled tracts intersecting with a parcel in the left lateral occipital sulcus, which (in addition to many short-range fibers) connects to the temporal lobe through the inferior longitudinal fasciculus, to the frontal lobe through the inferior fronto-occipital fasciculus, and to homologous regions in the right hemisphere through the corpus callosum [300]. Note that the full connectivity matrix was constructed from a million times as many tracks as shown in this figure, in order to estimate the

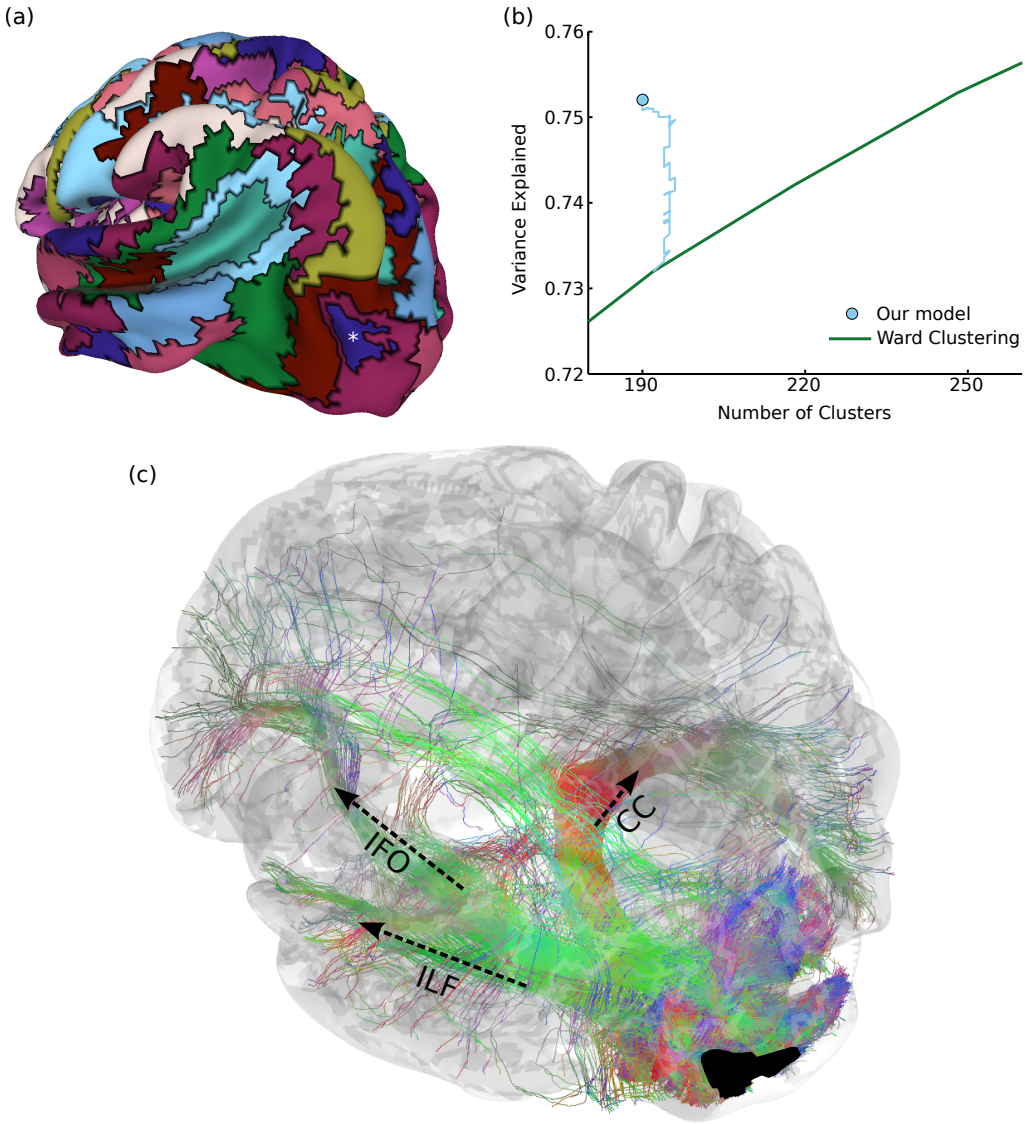


Figure 6.4: **Results on structural brain connectivity.** (a) A 190-cluster parcellation of the brain based on structural tractography patterns. (b) This parcellation fits the data substantially better than greedy clustering, which would require an additional 55 clusters to explain the same amount of variance. The blue path shows how our model fit improves over the course of Gibbs sampling when initialized with the greedy solution. (c) An example of 35,000 tracks (from one subject) connected to a parcel in the lateral occipital sulcus, marked with an asterisk in (a). These include portions of major fascicles such as the inferior longitudinal fasciculus (ILF), inferior fronto-occipital fasciculus (IFO), and corpus callosum (CC).

pairwise connectivity between every pair of gray-matter voxels.

6.3.4 Human migration in the United States

Given our successful results on neuroimaging data, we then applied our method to an entirely distinct dataset: internal migration within the United States. Using our probabilistic model, we sought to summarize the (asymmetric) matrix of migration between US counties as flows between a smaller number of contiguous regions. The model is essentially searching for a parcellation such that all counties within a parcel have similar (in- and out-) migration patterns. Note that this is a challenging dataset for clustering analyses since the county-level migration matrix is extremely noisy and sparse, with only 3.8% of flows having a nonzero value.

As shown in Fig. 6.5(a), we identify 83 regions defined by their migration properties. There are a number of interesting properties of this parcellation of the United States. Many clusters share borders with state borders, even though no information about the state membership of different counties was used during the parcellation. This alignment was substantially more prominent than when generating random 83-cluster parcellations, as shown in Fig. 6.5(b). As described in the Discussion, this is consistent with previous work showing behavioral differences caused by state borders, providing the first evidence that state membership also has an impact on intranational migration patterns. Greedy clustering performs very poorly on this sparse, noisy matrix, producing many clusters containing only one or a small number of counties, and has a lower NMI with state borders than even the random parcellations.

The 10 most populous clusters (Fig. 6.5(c)) cover 18 of the 20 largest cities in the US, with the two largest parcels covering the Northeast and the west coast. Some clusters roughly align with states or groups of states, while other divide states (e.g. the urban centers of east Texas) or cut across multiple states (e.g. the "urban midwest" cluster consisting of Columbus, Detroit, and Chicago). As shown in Fig. 6.5(d), our method succeeds in reordering the migration matrix to be composed of approximately piecewise constant blocks. In this case (and in many applications) the blocks along the main diagonal are most prominent, but this assortative structure is not enforced

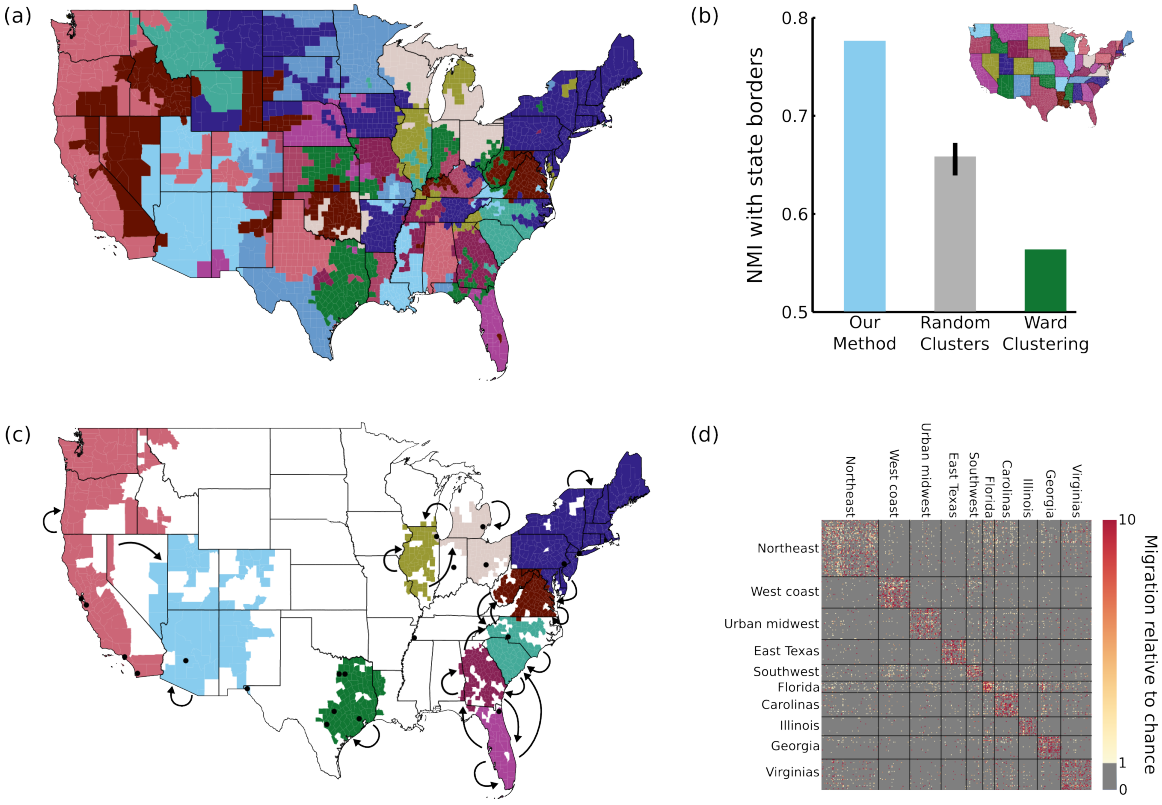


Figure 6.5: **Results on migration dataset.** (a) Our parcellation identified 83 contiguous regions within the continental US, such that migration between these regions summarizes the migration between all 2594 counties. (b) This parcellation was better aligned with state borders than an 83-cluster random parcellation (95% confidence interval shown) or an 83-cluster greedy Ward parcellation. (c) The top 10 clusters (by population) are shown, with arrows indicating above-chance flows between the clusters. The 20 most populous US cities are indicated with black dots for reference. (d) A portion of the migration matrix, showing the 1051 counties covered by the top 10 clusters.

by the model. Though largely symmetric, some flows do show large asymmetries. For example, the two most asymmetrical flows by absolute difference are between the urban midwest and Illinois (out of Illinois = 1.3, into Illinois = 2.0), and Florida and Georgia (out of Georgia = 1.3, into Georgia = 2.0).

6.4 Discussion

In this work we have introduced a new generative nonparametric model for parcellating a spatial map based on connectivity information. After showing that our model outperforms existing baselines on synthetic data, we applied it to three distinct real-world datasets: functional brain connectivity, structural brain connectivity, and US migration. In each case our method showed improvements over the current state-of-the-art, and was able to capture hidden spatial patterns in the connectivity data. The gap between our approach and past work varied with the difficulty of the parcellation problem; hierarchical clustering would require 17% more clusters for the relatively smooth functional connectivity data and 29% more clusters for the more challenging structural connectivity data, and fails completely for the most noisy migration dataset.

Finding a connectivity-based parcellation of the brain's cortical surface has been an important goal in recent neuroimaging research, for two primary reasons. First, the shapes and locations of connectivity-defined regions may help inform us about underlying modularity in cortex, providing a relatively hypothesis-free delineation of regions with distinct functional or structural properties. For example, connectivity clustering has been used to identify substructures in the posterior medial cortex [45], temporoparietal junction [192], medial frontal cortex [71, 145, 157, 160], occipital lobes [280], frontal pole [180, 200], lateral premotor cortex [285], lateral parietal cortex [191, 245], amygdala [64, 199], and insula [58]. Second, an accurate parcellation is necessary for performing higher-level analysis, such as analyzing distributed connectivity networks among parcels [7, 134, 228], using connectivity as a clinical biomarker [57], or pooling voxel features for classification [317]. Consistent with our results, previous work has found that greedy Ward clustering generally fits the datasets best

(in terms of variance explained) among these existing methods [282].

Our finding of eccentricity-based resting-state parcels in early visual areas is consistent with previous results showing a foveal vs. peripheral division of visual regions based on connectivity [172, 321]. Since our parcellation is much higher-resolution, we are able to observe nested clusters at multiple eccentricities. Our results are the first to suggest that higher-level retinotopic regions, especially PHC1 and PHC2, have borders that are related to changes in connectivity properties.

Parcellation based on structural tractography has generally been limited to specific regions of interest [71, 145, 160, 180, 191, 192, 200, 245, 280, 285], in part due to the computational difficulties of computing and analyzing a full voxel-by-voxel connectivity matrix. Our parcellation for this modality is somewhat preliminary; probabilistic tractography algorithms are still in their infancy, with recent work showing that they produce many tracts that are not well-supported by the underlying diffusion data [224] and are of questionable anatomical accuracy [283]. As diffusion imaging and tractography methods continue to improve, the input connectivity matrix to our method will become higher quality and allow for more precise parcellation.

There has been detailed scientific study of both inter- and intra-national migration patterns for over a century, beginning with the 1885 work of Ravenstein [233]. Even in this initial study (within the UK), it was clear that migration properties varied with spatial location; for example, rural areas showed large out-migration, while metropolitan areas showed greater in-migration, including long-distance migrants. The impact of state borders on migration behavior has not, to our knowledge, been specifically addressed, but there is a growing literature documenting differences in behaviors across state lines. Neighboring counties across state lines are less politically similar than those within a state, suggesting that a state border "creates a barrier to, or contains, political and economic institutions, policies, and possibly movement" [279]. State borders also play a role in isolating communities economically; this phenomenon gained a great of attention after Wolf's 2000 study [313], showing that trade was markedly lower between states than within states (controlling for distance using a gravity model). Our results demonstrate in a hypothesis-free way that migration behavior is influenced by state identities, since our method discovers a parcellation

related in many regions to state borders, without being given any information about the state membership of each county. Our results also show that state borders alone are not sufficient to capture the complexities of migration behavior, since other factors can override state identities to create other types of communities (such as in our "Urban midwest" parcel).

Since our algorithm is searching a much larger space of potential parcellations compared to previous methods, it does take longer to find the most likely clustering. There are a number of possible approaches for speeding up inference which could be explored in future work. One possibility is parallelize inference by performing Gibbs sampling on multiple elements simultaneously; although this would no longer be guaranteed to converge to the true posterior distribution, in practice this may not be an issue. Another option is to compute the Gibbs sampling probabilities only approximately [164], by using only a random subset of connectivities in a large matrix to approximate the likelihood of a proposed parcellation. It also may be possible to increase the performance of our algorithm even further by starting with many different initializations and selecting the solution with highest MAP probability.

6.5 Conclusions

In summary, we have proposed the first general-purpose probabilistic model to intrinsically incorporate spatial information in its clustering prior, allowing us to search directly in the space of contiguous parcellations using collapsed Gibbs sampling. Our approach is far more flexible and precise than previous work, with no constraints on the sizes and shapes of the learned parcels. This makes our model more resilient to noise in synthetic tests, and provides better fits to real-world data drawn from three different domains. This diverse set of results suggests that our model could be applied to a large set of biological network datasets to reveal fine-grained structure in spatial maps.

6.6 Acknowledgments

Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Thank you Henry Jung for porting our MATLAB code to python, to Mike Arcaro for providing the map of retinotopic regions in visual cortex, and to Michelle Greene for reviewing early versions of this draft.

Funding was provided by a National Science Foundation Graduate Research Fellowship under grant number DGE-0645962, and Office of Naval Research Multidisciplinary University Research Initiative grant number N000141410671.

Chapter 7

Two distinct scene processing networks connecting vision and memory

Research on visual scene understanding has identified a number of regions involved in processing natural scenes, but has lacked a unifying framework for understanding how these different regions are organized and interact. We propose a new organizational principle, in which scene processing relies on two distinct networks that split the classically defined Parahippocampal Place Area (PPA). The first network consists of the Transverse Occipital Sulcus (TOS, or the Occipital Place Area) and the posterior portion of the PPA (pPPA). These regions have a well-defined retinotopic organization and do not show strong memory or context effects, suggesting that this network primarily processes visual features from the current view of a scene. The second network consists of the caudal Inferior Parietal Lobule (cIPL), Retrosplenial Cortex (RSC), and the anterior portion of the PPA (aPPA). These regions are involved in a wide range of both visual and non-visual tasks involving episodic memory, navigation, and imagination, and connect information about a current scene view with a much broader temporal and spatial context. We provide evidence for this division from a diverse set of sources. Using a data-driven approach to parcellate resting-state fMRI data, we identify coherent functional regions corresponding to scene-processing

areas. We then show that a network clustering analysis separates these scene-related regions into two adjacent networks, which show sharp changes in connectivity properties. Additionally, we argue that the cIPL has been previously overlooked as a critical region for full scene understanding, based on a meta-analysis of previous functional studies as well as diffusion tractography results showing that cIPL is well-positioned to connect visual cortex with other cortical systems. This new framework for understanding the neural substrates of scene processing bridges results from many lines of research, and makes specific predictions about functional properties of these regions. This chapter is joint work with Andre Esteva, Diane M. Beck, and Fei-Fei Li.

7.1 Introduction

Natural scene perception has been shown to rely on a distributed set of cortical regions, including the parahippocampal place area (PPA) [92], retrosplenial cortex (RSC) [213], and the transverse occipital sulcus (TOS, aka the occipital place area, OPA) [125, 208]. More recent work has suggested that the picture is even more complicated, with PPA containing multiple subdivisions and the possible involvement of the parietal lobe [15]. Although there has been substantial progress in understanding the functional properties of each of these regions and the differences between them, the field has lacked a coherent overall framework for summarizing the overall architecture of the human scene processing system.

There is a long history of proposals for partitioning the visual system into separable components with different functions, such as spatial frequency channels [48], what versus where/how pathways [167, 198], or magnocellular, parcocellular, and koniocellular streams [152]. A division that is particularly relevant to natural scene perception is between the specific visual features present in the current glance of a scene, and the stable, high-level knowledge of where the place exists in the world, what has happened here in the past, and what possible actions we could take here in the future. For most cognitive and physical tasks we undertake in real-world places, the specific visual attributes we perceive are just a means to this end, of recalling and updating information about the physical environment; “the essential feature of

a landmark is not its design, but the place it holds in a city's memory" [207]. The connection between place and memory has been recognized for thousands of years, reflected in the ancient Greek method of loci that seeks to strengthen a memory by associating it with a physical location [319].

Some previous work has begun to point to this type of organizing principle among scene perception regions. Mapping functional connectivity differences between pairs of scene-sensitive regions has revealed some consistent distinctions, with some regions more connected to visual cortex and others to parietal and medial temporal regions [15, 210]. Contrasting activity evoked by perceptual categorization tasks compared to semantic retrieval tasks shows a similar division between visual and higher-level cortex [98]. These experiments, however, have all been targeted, hypothesis-driven comparisons between regions with similar functional properties. It is unclear whether these divisions are major organizing principles of the brains connectivity networks, or simply subtle differences within a single coherent scene-processing network.

To answer this question, we took a data-driven approach to identifying scene-sensitive regions and clustering cortical connectivity. After applying a state-of-the-art connectivity algorithm [16] to generate spatially-coherent parcels based on high-resolution resting-state connectivity, we associate these parcels with components of the scene-processing network using category localizers, retinotopic field maps, category decoding, and a meta-analysis of previous work. We then perform hierarchical clustering and multidimensional scaling to show that there is a prominent, bilaterally symmetric division of scene-related regions into two separate networks: one includes TOS and the posterior portion of PPA (retinotopic maps PHC1 and PHC2), while the other is composed of the RSC, anterior PPA (aPPA), and the caudal inferior parietal lobule (cIPL). We show that the least well-known of these regions, the cIPL, actually has unique structural connectivity properties which makes it well suited to link visual perception with processing throughout the rest of the cortex.

Based on these results, as well as a review of previous studies, we propose that scene processing is fundamentally divided into two collaborating but distinct networks, with one focused on the visual features of a scene image and the other related to contextual retrieval and navigation. Under this framework, scene perception is less

the function of a unified set of distributed neural machinery and more of "an ongoing dialogue between the material and symbolic aspects of the past and the continuously unfolding present" [14].

7.2 Materials and Methods

7.2.1 Imaging Data

The majority of the data used in this study was obtained from the Human Connectome Project (HCP), which provides detailed documentation on the experimental and acquisition parameters for these datasets [291]. We provide an overview of these datasets below.

Diffusion imaging data was used for the first 10 subjects from the January 2014 "Q3" HCP data release with complete data (subj ids 100408, 101915, 102816, 105216, 106016, 106319, 111009, 111514, 111716, 112819). Data were acquired using a multiband sequence at three different b-values (1000, 2000, 3000 s/mm2), with a total of 270 diffusion weighting directions and a resolution of 1.25mm isotropic.

The group-level functional connectivity data were derived from the 468-subject group-PCA eigenmaps, distributed with the June 2014 500 Subjects HCP data release. Resting-state fMRI data were acquired over four sessions (14 min, 33 seconds each) while subjects fixed on a bright cross-hair on a dark background, using a multiband sequence to achieve a TR of 720ms at 2.0mm isotropic resolution (59412 surface vertices). These timecourses were cleaned using FMRIB's ICA-based Xnoiseifier (FIX) [252], and then the top 4500 eigenvectors for each voxel were estimated across all subjects using Group-PCA [265].

For the first 20 subjects within the "500 Subjects" release with complete data (and non-overlapping with the Q3 subjects: subj ids 101006, 101107, 101309, 102008, 102311, 103111, 104820, 105014, 106521, 107321, 107422, 108121, 108323, 108525, 108828, 109123, 109325, 111413, 113922, 120515), we created individual subject

resting-state datasets by demeaning and concatenating their four resting-state sessions. We also obtained these subjects data from the HCP Working Memory experiment, in which they observed blocks of stimuli consisting of faces, places, tools, or body parts. We collapse across the two memory tasks being performed by participants (target-detection or 2-back detection).

To identify group-level scene localizers, we used data from a separate set of 24 subjects scanned at Stanford University (see below). Each subject viewed blocks of stimuli from six categories: child faces, adult faces, indoor scenes, outdoor scenes, objects (abstract sculptures with no semantic meaning), and scrambled objects. Functional data were acquired with an in-place resolution of 1.56mm, slice thickness of 3mm (with 1 mm gap), and a TR of 2s; a high-resolution (1mm isotropic) SPGR structural scan was also acquired to allow for transformation to MNI space. Full details of the localizer stimuli and acquisition parameters are given in our previous work [15].

7.2.2 Subjects

Scene localizer data was collected from 24 subjects (6 female, ages 22-32, including one of the authors). Subjects were in good health with no past history of psychiatric or neurological diseases, and with normal or corrected-to-normal vision. The experimental protocol was approved by the Institutional Review Board of Stanford University, and all subjects gave their written informed consent.

7.2.3 Resting-state Parcellation

We generated a voxel-level functional connectivity matrix by correlating the group-level eigenmaps for every pair of voxels and applying the arctangent function. We parcellated this 59412 by 59412 matrix into contiguous regions, using a generative probabilistic model [16]. This method finds a parcellation of the cortex such that the connectivity properties within each parcel are as uniform as possible, making multiple passes over the dataset to fine-tune the parcel borders. We set the scaling hyperparameter $\frac{2}{0} = 3000$ to produce a manageable number of parcels.

7.2.4 Scene localizers and retinotopic field maps

To identify PPA, RSC, and TOS, we deconvolved the localizer data from the 24 Stanford subjects using the standard block hemodynamic model in AFNI [69], with faces, scenes, objects, and scrambled objects as regressors. The Scenes ¿ Objects t-statistic was used to define PPA (top 300 voxels near the parahippocampal gyrus), RSC (top 200 voxels near retrosplenial cortex), and TOS (top 200 voxels near the transverse occipital sulcus). The ROI masks were then transformed to MNI space, summed across all subjects, and mapped to the closest vertices on the group cortical surface. The cluster denoting highest overlap between subjects was then manually annotated.

A volumetric group-level probabilistic atlas [304] was used to define retinotopic field maps, by mapping each field map to the closest vertices on the group-level surface.

7.2.5 Scene category decoding

For each cortical parcel (generated from resting-state connectivity as described above), we measured its sensitivity to scenes versus other visual categories through a category decoding analysis. We first used a hemodynamic model to associate timepoints within the 20 HCP working memory datasets with specific stimulus categories. We labeled timepoints as corresponding to bodies, faces, places, or tools by constructing a boxcar timecourse denoting when each stimulus category was being displayed, convolving these indicators with the standard SPM hemodynamic response function provided with AFNI [69], rescaling the maximum value to 1, then re-thresholding to a binary indicator. Effectively, this produced a shift of the stimulus blocks by 5.55s to account for hemodynamic delay. The fMRI timecourses were cleaned by regressing out movement (6 degree-of-freedom translation/rotation and derivatives) and constant, linear, and quadratic trends from each run, then normalizing each voxel to have unit variance. Voxel timecourses were then averaged within each parcel, yielding a vector of average parcel activities for each timepoint.

Linear support vector machines (SVMs) were trained separately for each subject

to discriminate scene timepoints from non-scene timepoints, and then tested on the other 19 subjects. We set the soft-margin hyperparameter c=1, but our results are not sensitive to this choice. Note that chance performance is 75%, since only 25% of the stimulus timepoints are scenes. Each subjects classifier assigned a weight to each parcel, indicating how strongly activity in this parcel predicted that a scene was being viewed. Parcels consistently assigned high positive weights were therefore most strongly associated with visual scene processing.

7.2.6 Meta-analysis

We sought to identify all fMRI studies involving scene memory, navigation, imagined experiences, or context memory that reported activation coordinates around the posterior parietal lobe. These coordinates were assumed to be in MNI space, unless identified as being in Talairach space, in which case we transformed the coordinates to MNI space [33]. Each coordinate was then mapped to the closest vertex on the group surface.

7.2.7 Parcel-to-parcel functional connectivity matrices

The 468-subject eigenmaps distributed by the HCP are approximately equal to performing a singular value decomposition on the concatenated timecourses of all 468 subjects, and then retaining the right singular values scaled by their eigenvalues [265]. This allows us to treat these eigenmaps as pseudo-timecourses, since dot products (and thus correlations) between eigenmaps approximate the dot products between the original voxel timecourses. Given a parcellation, we computed the group-level connectivity between a pair of regions by taking the mean over all eigenmaps in each region, then correlating these mean eigenmaps and applying the Fisher z-transform (hyperbolic arctangent). We computed subject-level connectivity in the same way, using the resting-state timecourse for each voxel rather than the eigenmap.

7.2.8 Network Clustering and Multidimensional scaling

The 172 by 172 parcel functional connectivity matrix was converted into a distance matrix by subtracting every entry from the maximum entry. Ward clustering (unconstrained by parcel position) was used to compute a hard clustering into 10 networks. Separately, classical multidimensional scaling was also applied to the distance matrix, and the first three dimensions were used to assign voxels RGB colors (with each color channel scaled to span the full range of 0 to 255 along each axis) and to plot parcels in a 3D space. We performed the same operation on each subject-level matrix as well, and then aligned each subjects 3D pointcloud to the group pointcloud using a procrustes transform.

7.2.9 Structural connectivity

Probabilistic tractgraphy was performed on each of the 10 HCP diffusion datasets using FSL [144], by estimating up to 3 crossing fibers with bedpostx (using gradient nonlinearities and a rician noise model) and then running probtrackx2 using the default parameters and distance correction. 2000 fibers were generated for each of the $1.7x10^6$ white-matter voxels, yielding $3.4x10^9$ total sampled tracks per subject (approximately 34 billion tracks in total). We assigned each of the endpoints to gray-matter voxels using the 32k/hemisphere Conte69 registered standard mesh distributed for each subject, discarding the small number of tracks that did not have both endpoints in gray matter (e.g. cerebellar or spinal cord tracks). Since we are using distance correction, the weight of a track is set equal to its length.

The distance-based connectivity profile of a voxel was obtained by summing all of the voxels connections within 1cm bins based on Euclidean distance from the voxel. The profile for a parcel was then computed as the average of all its voxel profiles (rather than the sum, which does not control for differing parcel areas). Connectivity profiles for cIPL parcels vs. other parcels were compared using a two-way repeated measures ANOVA, with cIPL vs. other as the first factor and distance bin as the second factor.

We computed the structural connectivity between a pair of parcels A and B as the

mean connectivity strength over all pairs of voxels with one voxel drawn from A and one drawn from B. Note that this also yields a measurement independent of parcel size.

7.3 Results

In order to reduce the complexity of the full 1.8-billion element whole-brain resting-state functional connectivity matrix, we first performed spatial parcellation using a generative modeling approach [16]. This parcellation consisted of 172 spatially-coherent regions across both hemispheres, each of which contains voxels with near-uniform connectivity properties. The connectivity matrix between these 172 parcels captures more than 76% of the variance in the original connectivity matrix, despite being dramatically smaller (by five orders of magnitude). Representing the connectivity matrix in this way allows us to identify locations where functional connectivity profiles change rapidly (the boundaries between parcels), and lets us examine functional and connectivity properties at the more manageable and meaningful parcel level rather than at the voxel level.

7.3.1 Identifying Scene-Sensitive Parcels

Our first goal was to identify parcels that were related to processing visual scenes, using several different approaches as shown in Figure 7.1. Mapping group-level retinotopic field maps to the surface shows that the parcels exhibit an eccentricity-based organization (dividing foveal and peripheral voxels) in early visual areas, but that parcel boundaries begin to align with field map boundaries in later dorsal and ventral regions, as we have previously reported [16]. This alignment is especially prominent in parahippocampal regions PHC1 and PHC2, which are divided into anterior and posterior parcels. In the left (right) hemisphere, 86% (87%) of PHC1 voxels fall into the posterior parcel and 97% (72%) of PHC2 voxels fall into the anterior parcel. We also overlaid group-level localizer data (from a separate group of subjects) for scene-sensitive regions TOS, RSC, and PPA. TOS and RSC fall largely within single parcels

(which we label the TOS and RSC parcels), while PPA runs perpendicular to parcel boundaries, extending through at least three separate parcels. The two posterior parcels correspond to PHC1 and PHC2 (which we collectively refer to as "posterior PPA", pPPA), and we label the most anterior parcel as "anterior PPA" (aPPA).

We can directly confirm that these parcels are scene-sensitive by applying our parcellation to task-fMRI data from the Human Connectome Project, and using the mean activity of each parcel as a feature for decoding scenes vs. other visual categories (faces, tools, bodies). These decoding accuracies were well above chance, even across subjects; a decoder trained on one subject could identify scene timepoints in other subjects with 85.1% accuracy (t_{19}=23.71, p<0.01; one-tailed t-test). Parcels that were consistently assigned positive weights for decoding scenes vs. other categories are identified in Figure 7.2. Scene-related parcels labeled from retinotopic maps and localizers exhibit high decoding weights (TOS: left t_{19}=3.95, p<0.01; right t_{19}=5.70, p<0.01; RSC: left t_{19}=4.95, p<0.01; right t_{19}=2.80, p<0.01; PHC1: left t_{19}=3.83, p<0.01; right t_{19}=1.06, n.s.; PHC2: left t_{19}=4.95, p<0.01; right t_{19}=5.66, p<0.01; aPPA: left t_{19}=1.73, p<0.05; right t_{19}=7.34, p<0.01; one-tailed t-test).

Interestingly, scene selectivity extends dorsally beyond TOS, into the caudal inferior parietal lobule (cIPL). Labeling the three parcels in this region cIPL1-3 (ordered posterior to anterior along the angular gyrus), both cIPL1 and cIPL2 consistently show discriminative weights for the (unfamiliar) localizer scenes (cIPL1: left t_{19}=9.61, p<0.01; right t_{19}=8.34, p<0.01; cIPL2: left t_{19}=3.87, p<0.01; right t_{19}=3.58, p<0.01) while cIPL3 does not (left t_{19}=-1.16, n.s; right t_{19}=1.48, n.s.). This result suggests that there may be scene-related activity anterior to typically-defined TOS, but does not provide clear evidence for a separate region with different functional properties. Scene localizers, however, are missing a critical component of real-world scene perception; since they typically include only unfamiliar scenes, they may fail to robustly activate memory and contextual networks engaged in processing familiar environments. A meta-analysis of previous studies shows that personally familiar places robustly activate cIPL, especially around cIPL2 and cIPL3 (Figure 7.3). This activation appears for a wide variety of tasks, including memory for visual scene images [10, 88, 94, 201, 278], learning navigational routes [32, 41], and even simply imagining

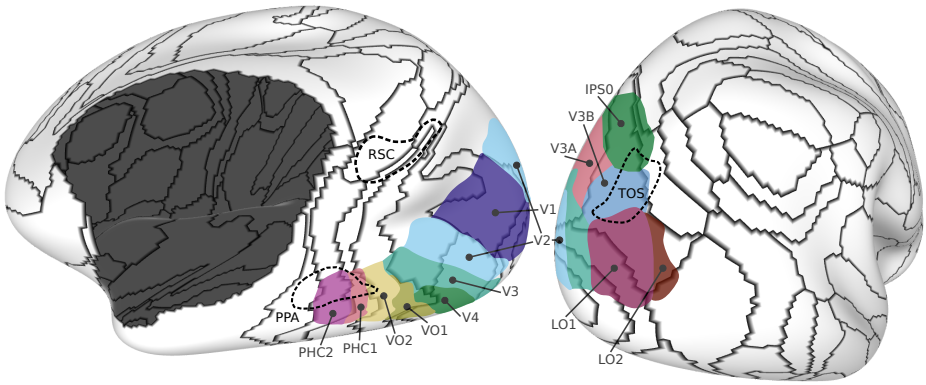


Figure 7.1: **Relationship between resting-state parcels, retinotipic maps, and scene localizers**. Group-level visual field maps and functional localizers are overlaid on parcels derived from resting-state connectivity patterns (black borders). RSC and TOS largely fall within a single parcel, with TOS corresponding roughly to V3B. Ventrally, PHC1 and PHC2 are well divided into two separate parcels, with PPA extending anteriorly into a parcel we denote aPPA.

past events or future events in familiar places [123, 275]. This same region can also be activated by recalling non-place stimuli (including words and objects), if the stimuli are associated with strong memory of the source context [146, 225, 294]. These studies, along with our previous work showing connectivity differences between TOS and cIPL [15], provide strong evidence that the caudal inferior parietal lobe is in fact a separate, important component of the scene-processing system.

7.3.2 Clustering Parcels into Networks

Having identified these eight (bilateral) parcels critical to scene perception, we clustered the whole-brain connectivity matrix to identify 10 functionally-connected networks. This data-driven analysis groups together parcels that all have high functional connectivity with one another, regardless of their spatial position.As shown in Figure 7.4, these networks are remarkably symmetric between hemispheres, and split scene perception regions into two separate categories. Posterior parcels - TOS, cIPL1, PHC1, and PHC2 - were clustered into visual network (dark blue) covering all of visual cortex outside of the early foveal cluster. Anterior parcels - cIPL2, cIPL3, RSC, and aPPA - were clustered into a separate parietal/medial-temporal network (pink),

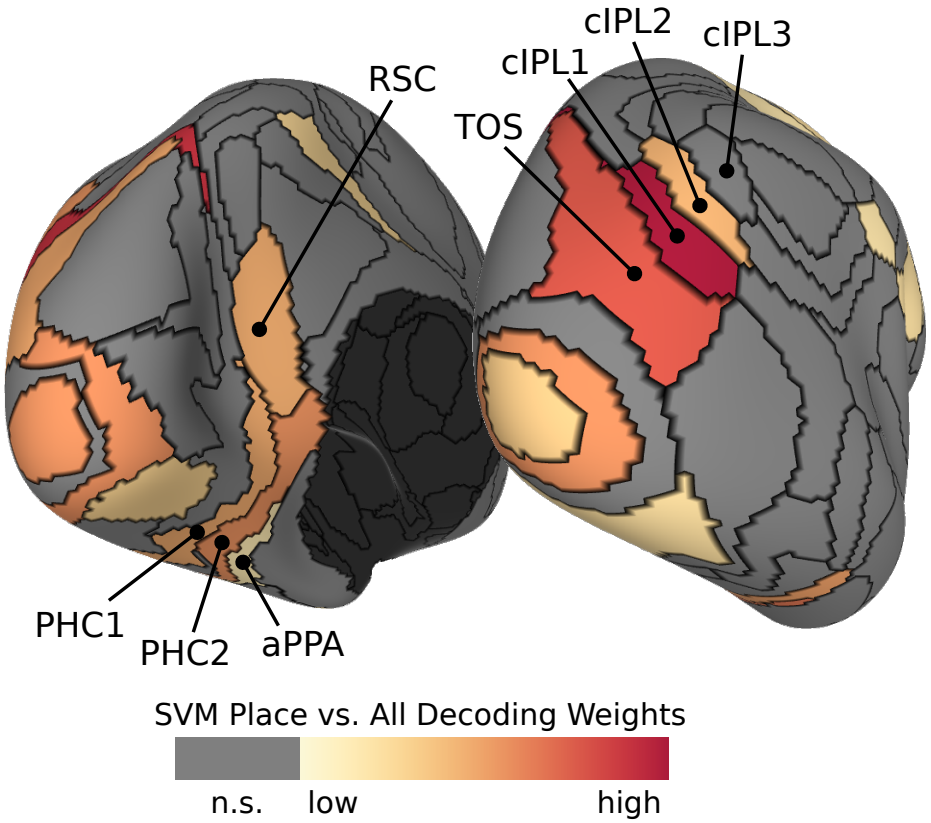


Figure 7.2: **Parcel scene decoding weights**. Linear SVMs were trained to classify unfamiliar scenes vs other images (faces, tools, bodies) based on mean activity in each resting-state parcel. Colored regions are those having significant positive weights across subjects ($p<0.05$). High activity in the parcels identified using field maps and scene localizers (Figure 1) predict that subjects are viewing scenes, and these positive weights extend from TOS partially onto the angular gyrus.

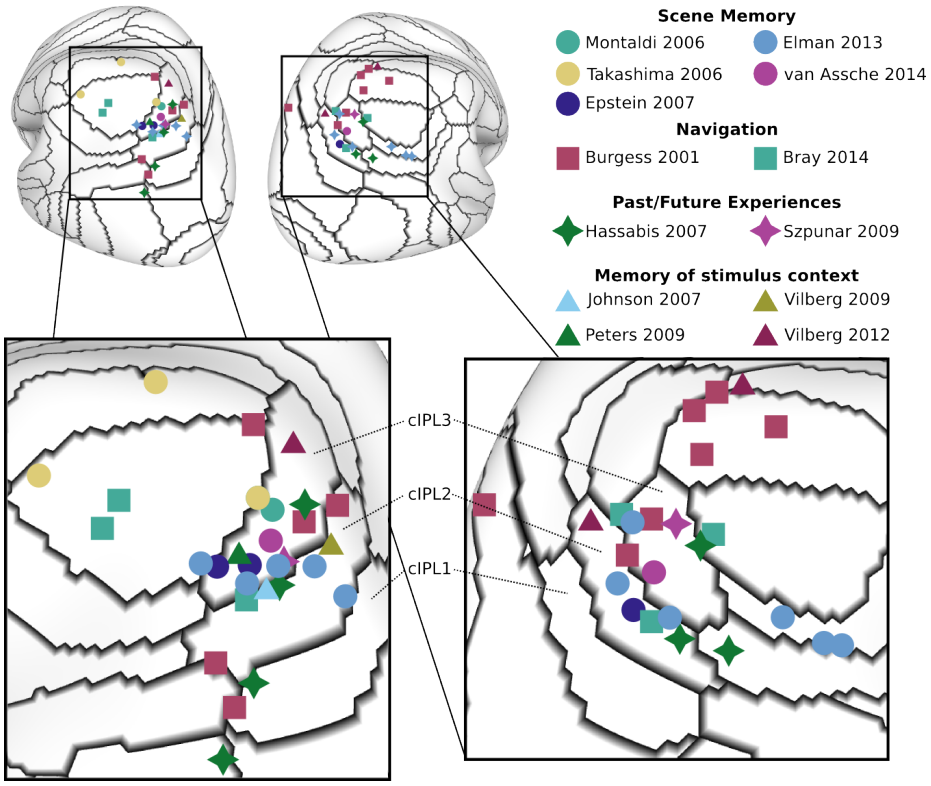


Figure 7.3: **Meta-analysis of cIPL involvement in place memory**. Although not typically identified as a scene-sensitive region, the posterior parietal lobe is consistently activated in studies involving familiar places. Perceiving images of familiar scenes, learning navigational routes, or imagining events in familiar places produces activation clustered around cIPL2-3. This same region also appears in memory studies of non-scene stimuli associated with a strong context.

which also included anterior temporal and medial frontal parcels. This corresponds to a portion of the known default mode regions, with other default mode regions being grouped into a separate network (green). The dividing line between the visual and context networks falls consistently near the edge of known retinotopic maps, suggesting a division between regions strongly tied to the current retinal input and those which are more driven by internally-driven processes and integrate information over longer time-scales. If the number of clusters is increased, divisions within these networks appear, first between TOS and pPPA, and then between RSC/cIPL and aPPA.

Rather than performing a hard clustering into distinct groups, we can use classical multidimensional scaling (MDS) to embed parcels into a three-dimensional space. Distances in this space approximate the functional connectivity strength between parcels, such that strongly-connected parcels are close together. Setting the RGB color of each parcel based on its position in this three-dimensional embedding space gives a soft clustering (Figure 7.5(a)). Moving along either the dorsal (TOS-cIPL) or ventral (PHC-aPPA) boundaries between scene regions produces rapid changes in functional connectivity properties, visualized in embedding space in Figure 7.5(b-c). In both cases, the most posterior regions (TOS and PHC1) show strong connectivity to other parcels in visual cortex, while the most anterior regions (cIPL3 and aPPA) are instead more related to default mode regions. To statistically evaluate this difference, we measure the connectivity between each scene-related parcel and a default-mode reference parcel on the opposite side of cortex (medial versus lateral), to avoid spurious connectivity due to local noise correlations. For the dorsal parcels, we measure connectivity to RSC, and for the ventral parcels, we measure connectivity to cIPL3. Along the dorsal boundary, we see significant increases in connectivity to RSC when moving from TOS to cIPL1 (Left: t_{19}=6.98, p<0.01; Right: t_{19}=6.35, p<0.01; two-tailed paired t-test), from cIPL1 to cIPL2 (Left: t_{19}=7.72, p<0.01; Right: t_{19}=6.16, p<0.01), and from cIPL2 to cIPL3 (Right: t_{19}=2.44, p<0.05). We observe a similar (though less dramatic) increase in connectivity to cIPL3 when moving from PHC1 to PHC2 (Left: t_{19}=4.21, p<0.01; Right: t_{19}=2.68, p<0.05) and PHC2 to aPPA (Right: t_{19}=3.03, p<0.01). These results (Figure 7.5(d-e)) indicate that the borders between

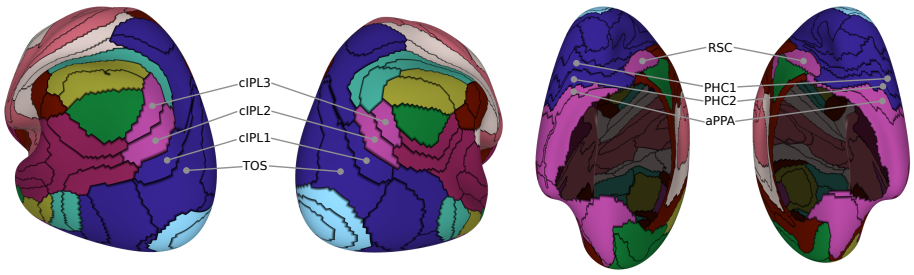


Figure 7.4: **Connectivity clustering of parcels**. Performing hierarchical clustering on the resting-state parcels based on their pairwise functional connectivity reveals that the scene processing network is split across two networks: a visual network (blue) which includes TOS and PHC1/2, and a parietal/medial-temporal network including cIPL, RSC, and aPPA. The visual network covers known retintopic field maps outside the early fovea, while the parietal/medial-temporal network corresponds to a portion of the default mode network.

the visual and context networks are not artifacts of the clustering procedure, but are in fact marked by rapid changes in connectivity properties.

Given the dramatic differences in functional connectivity properties among the scene parcels (especially cIPL, e.g. in Figure 7.5(d)), we examined whether these regions also differed in terms of structural connectivity, using diffusion imaging. We sampled 34 billion white matter seed locations across 10 subjects, and performed probabilistic tractography to identify the likely endpoints of the fiber tract passing through that seed. As shown in Figure 7.6, the cIPL parcels were qualitatively different from all other scene parcels, with both higher overall fiber incidence (per unit area) and a disproportionate number of long-range fibers (cIPL parcels vs. others, $F_{1,9}=191.24$, $p<0.01$; distance bin, $F_{19,171}=47.04$, $p<0.01$; interaction, $F_{19,171}=14.82$, $p<0.01$). These connections are widely distributed over posterior parietal, lateral and medial temporal, and prefrontal cortices, indicating the cIPL is structurally well-positioned to connect visual scene information with a wide variety of other cortical networks.

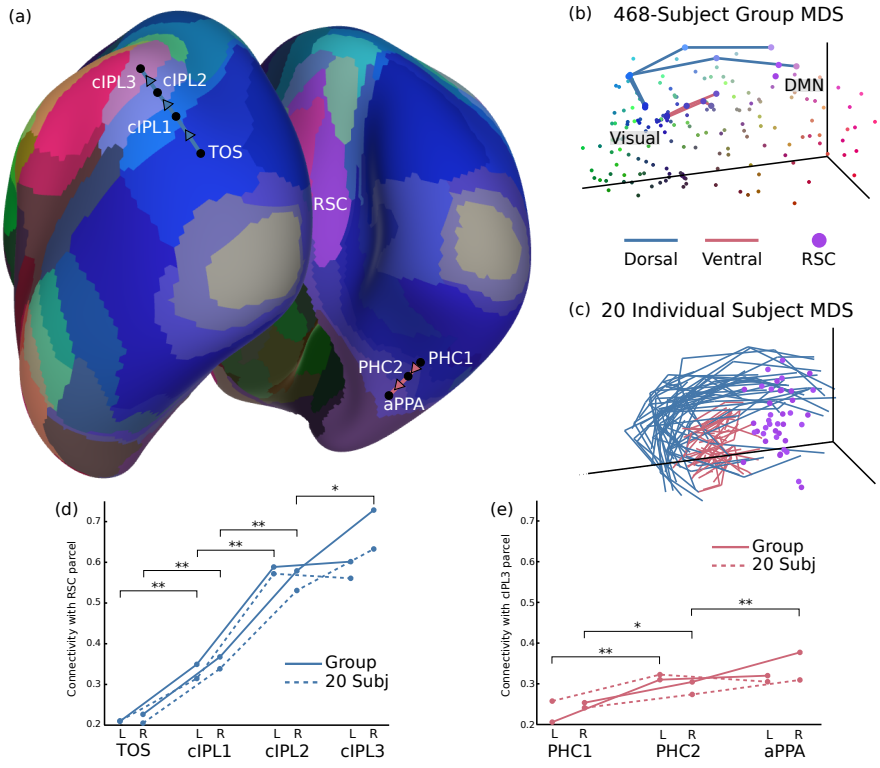


Figure 7.5: **Connectivity changes across the network border**. (a) Rather than performing a hard clustering assignment as in Figure 7.4, we can perform classical MDS on the parcel connectivity network and set regions RGB values based on their positions in a three-dimensional embedding space. This shows a similar result to hierarchical clustering, with abrupt connectivity changes across scene networks. (b) In MDS space, moving dorsally from TOS to cIPL3 produces the curves shown in blue, while moving ventrally from PHC1 to aPPA produces the curves shown in red. These curves move in parallel out of the retinotopic cluster toward the default mode cluster. (c) Plotting these curves for 20 individual subjects shows a similar pattern in each subject, with curves moving in parallel toward RSC (purple dots). (d) The connectivity between scene parcels and RSC increases dramatically as we move dorsally from TOS to cIPL3. (e) Connectivity with cIPL changes more subtly but significantly when moving ventrally from PHC1 to aPPA. *,** $p<0.05$, $p<0.01$

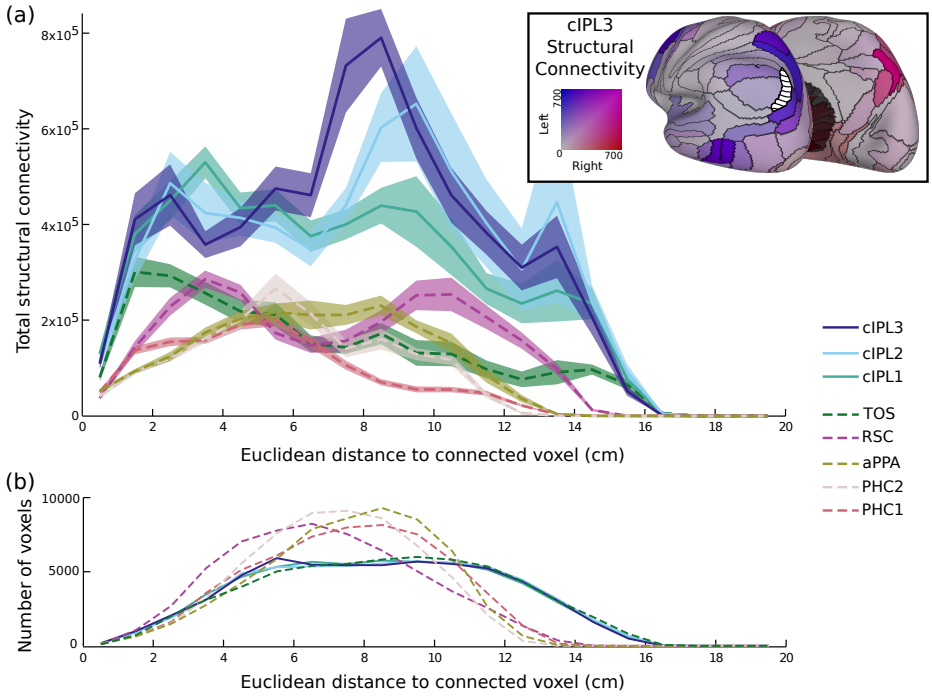


Figure 7.6: **Structural connectivity profiles of scene parcels**. (a) The connectivity between voxels in each parcel and the rest of the brain is plotted as a function of Euclidean distance (averaged between hemispheres, shaded regions show standard error of the mean). The cIPL parcels shows a distinct profile, both in overall connectivity strength and an emphasis on long-range connectivity. As shown in the inset, cIPL3 is structurally connected to a distributed set of cortical regions (primarily restricted to the same hemisphere). (b) The peak of cIPL connectivity around 10 cm is not driven by simple geometry, since the percentage of the cortex that is this distance away from cIPL is smaller than for other parcels such as RSC and those in PPA.

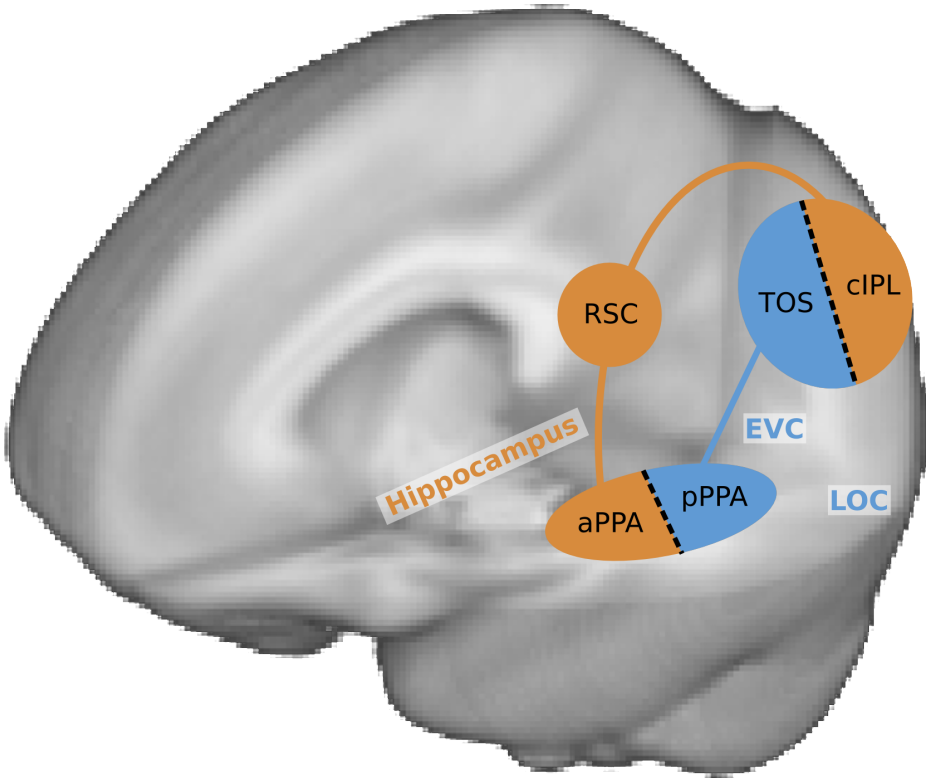


Figure 7.7: **Two-network model of scene perception**. Our results provide strong evidence for dividing scene-sensitive regions into two separate networks. TOS and posterior PPA (PHC1/2) process the current visual features of a scene (in concert with other visual areas, such early visual cortex and LOC), while cIPL, RSC, and anterior PPA perform higher-level context and navigation tasks (drawing on long-term memory structures such as the hippocampus).

7.4 Discussion

By combining a variety of data sources including function and structural connectivity data, task-fMRI, retinotopic maps, and a meta-analysis of previous results we have shown converging evidence for a functional division of scene-processing regions into two separate networks (summarized in Figure 7.7). The visual network covers retintopically-organized regions including TOS and posterior PPA (pPPA), while a separate memory-related network connects cIPL, RSC, and anterior PPA (aPPA). This division emerges from a purely data-driven network clustering, suggesting that this is a core organizing principle of the visual system. Our data also support a much more prominent role for cIPL in processing real-world familiar scenes, since it is well positioned both functionally and structurally to connect scene processing with the rest of the brain.

7.4.1 Subdivisions of the PPA

The division of the PPA into multiple anterior-posterior subregions with differing connectivity properties replicates our previous work (Baldassano et al., 2013) on an entirely different large-scale dataset, and shows that there is a strong connection between connectivity changes in PPA and the boundaries of retinotopic field maps. There is now a growing literature on anterior versus posterior PPA, including not only connectivity differences [210] but also the response to low-level [211] and high-level [178, 219] scene properties. Our results place this division into a larger context, and demonstrate that the connectivity differences within PPA are not just an isolated property of this region but a general organizing principle for scene-processing regions.

This subdivision may be the key to resolving a long-standing debate over the role of context effects in PPA. Some have proposed that PPA is primarily driven not by scenes per se but any stimuli with strong spatial contextual associations [5], and that these associations drive activity during even the early stages of perception [171]. Others have argued that PPA is only involved in visual spatial layout processing, and that context effects are mostly an artifact of later imagery [97]. We argue that both these descriptions may be correct, but for different portions of PPA, with pPPA

more related to concrete features of a visual scene and aPPA more related to general spatial context. In fact, the maps illustrated in these papers (Figure 4 in [5]; Figure 4 in [97]) suggest this type of anterior/posterior division.

7.4.2 The visual network

The visual network shows a close correspondence with the full set of retinotopic maps identified in previous studies [34, 138, 304], extending through the intraparietal sulcus (IPS) and laterally to hMT+. Our observation that TOS overlaps at the group level with retinotopic maps, primarily V3B, is consistent with prior measurements made in individual subjects [24, 209]. The only portion of cortex with known retinotopic maps that is not clustered in this network is the shared foveal representation of early visual areas, which segregates into its own cluster. One possible explanation is that our connectivity measures are based on eyes-open resting-state scans, during which a subjects fovea is being stimulated with a bright cross. This stimulation may be the dominant signal in this region, resulting in a suppression of the intrinsic fluctuations used to define resting-state networks.

TOS and posterior PPA have been shown to be responsive primarily to visual features of a stimulus, rather than higher-level attributes such as familiarity. Posterior PPA has a preferential response to high spatial frequencies [230], and both posterior PPA and TOS are activated by rectilinear shapes [211], even in non-scene images. Also, neither TOS nor posterior PPA show reliable familiarity effects ([94], but see further discussion below).

The functional distinction between pPPA and TOS is currently unclear. Previous work has speculated about the purpose of the apparent ventral and dorsal duplication of regions sensitive to large landmarks, proposing that it may be related to different output goals (e.g. action planning in TOS, object recognition in pPPA) [163], or to different input connections (e.g. lower visual field processing in TOS, upper visual field processing in pPPA) [168].

7.4.3 The context and navigation network

The network of parahippocampal, retrosplenial, and posterior parietal regions we identify has been emerged independently in many different fields of neuroimaging, outside of scene perception. Meta-analyses of internally-directed tasks such as theory of mind, autobiographical memory, and prospection have identified this as a core, re-occurring network [155, 268] (and component C10 of [320])). This network also appears in navigation [41, 267], recalling the study context of a stimulus [42, 128, 146, 225], recognition of personally familiar locations[10, 88], viewing objects with strong contextual associations [6], and thinking about past or imagined events in familiar contexts [123, 275, 276].

The broad set of tasks which recruit this network have been summarized in various ways, such as "scene construction" [124], "mnemonic scene construction" [7], or "relational processing" [85]. A review of memory studies referred to this network as the posterior medial (PM) memory system, and proposed that it is involved in any task requiring "situation models" relating entities, actions, and outcomes [231].

Sometimes this network appears as part of the larger default mode network, which includes other regions such as parts of medial prefrontal cortex. However, the functional and anatomical structure of the default mode network suggests that it not a single coherent structure, and that the parietal/medial-temporal portion is in fact a distinct subnetwork [7, 8, 321].

The specific functions of the individual components of this network have also been studied in a number of contexts. RSC appears to be most directly involved in orienting the viewer to the structure of the environment (both within and beyond the borders of the presented image) for the purpose of navigational planning; it encodes both absolute location and facing direction [96, 185, 293], integrates across views presented in a panoramic sequence [220], and shows strong familiarity effects [94, 95]. This is consistent with rodent neurophysiological studies, which have identified head direction cells in this region [63]. RSC is not sensitive to low-level rectilinear features in non-scene images such as objects or textures, though it does show some preference for rectilinear features in images of 3D scenes [211].

Anterior PPA has been less well-studied, since it was not recognized as a separate

region within the PPA until recently, but has been most strongly associated with coding the size of a scene [219]. Its representation of scene spaciousness draws on prior knowledge about the typical size of different scene categories, since it is affected by the presence of diagnostic objects [178].

The cIPL (also referred to as pIPL, PGp, or the angular gyrus) has been proposed as a "cross-modal hub" [8] that connects visual information with other sensory modalities as well as knowledge of the past. It is more intimately associated with visual cortex than most lateral parietal regions, since it has strong anatomical connections to higher-level visual regions in humans and macaques [53], and has a neurotransmitter receptor distribution similar to V3v and distinct from the rest of the IPL [55]. It is primarily involved in two related kinds of tasks. First, it supports contextual recall, showing both increases in mean activity [201, 294] as well as voxel-level activity patterns related to the specific context associated with an item [169]. Second, it performs temporal integration, sustaining activity under long delay periods [296], and accumulating both visual and auditory information over long time-scales [174]. Consistent with our structural connectivity results, its functional connections are distributed and flexible, coupling to the dorsal attention network during a spatial learning task [32] or to dorsolateral prefrontal and extrastriate visual cortex during successful recollection [159]. Based on these properties, it has been proposed [295] that this region implements the multi-modal episodic buffer proposed by [12].

Given cIPLs involvement in a diverse set of tasks, it has not traditionally been identified as a central part of the scene perception system. However, our results suggest a deep connection between cIPL and understanding real-world places, which (unlike typical localizer images) are associated with a wealth of memory, context, and navigational information. Our meta-analysis shows that cIPL is selectively responsive to familiar scenes (arguably the most common high-context stimuli in everyday life), but this property has largely gone unnoticed in the scene perception literature; for example, one of the studies in Figure 7.3 showing cIPL activation [94] described this location only as "near TOS." More importantly, our clustering analyses revealed that cIPL is tightly coupled (at rest) with RSC and aPPA, two regions that are widely recognized as performing scene-specific processing. Lesion studies support this view

that the posterior parietal lobe is primarily involved in scene-related functions (such as orienting to a previously learned map based on the current view), since these abilities can be selectively impacted without general memory deficits (reviewed in [167]).

7.4.4 Contrasting the two networks

Although our work is the first to propose the visual versus context networks as a general framework for scene perception, several previous studies have shown differential effects within these two networks. Contrasting the functional connectivity patterns of RSC vs. TOS or LOC [210] or anterior vs. posterior PPA [15] show a division between the two networks, consistent with our results. Contrasting scene-specific activity with general (image or word) memory retrieval showed an anterior vs. posterior distinction in PPA and cIPL/TOS, with only more anterior regions (aPPA and cIPL, along with RSC) responding to content-independent retrieval tasks [98, 146]. Our two-network division is also consistent with the dual intertwined rings model, which argues for a high-level division of cortex into a sensory ring and an association ring, the second of which is distributed but connected into a continuous ring through fiber tracts [194].

7.4.5 Open questions

The anterior/posterior pairing of aPPA/pPPA and cIPL/TOS raises the question of whether there is a similar anterior/posterior division in RSC. There is some evidence to suggest that this is the case: wide-field retinotopic mapping using natural scenes shows a partial retinotopic organization in RSC [138], and RSCs response to visual rectilinear features appears to be limited to the posterior portion [211]. However, we did not observe strong scene-selective responses in neighboring parcels near RSC (see Figure 7.2), a study of retinotopic coding in scene-selective regions failed to find any consistent topographic organization to RSC responses [306], and previous analyses of the functional properties of anterior versus posterior RSC have not found any significant differences [219].

Another interesting question is how spatial reference frames differ between and within the two networks. Given its retinotopic fieldmaps, the visual network presumably represents scene information relative to the current eye position; previous work has argued that this reference frame is truly retina-centered and not egocentric [107, 306]. The context network, however, likely transforms information between multiple reference frames. Models of spatial memory suggest that medial temporal lobe (possibly including aPPA) utilizes an allocentric representation, while the posterior parietal lobe (possibly including cIPL) is based on an egocentric reference frame, and that the two are connected via a transformation circuit in RSC that combines allocentric location and head direction [44, 292]. There is some recent evidence for this model in human neuroimaging: posterior parietal cortex codes the direction of attention in an egocentric reference frame (even for positions outside the field of view) [259], and RSC contains both position and head direction information (anchored to the local environment) [185]. This raises the possibility that another critical role of cIPL could be to transform retinotopic visual information into a stable egocentric scene over the course of multiple eye movements. The properties of aPPA, however, are much less clear; it seems unlikely that it would utilize an entirely different coordinate system than neighboring PHC1/2, and some aspects of the scene encoded in aPPA (such as overall scene size [219]) dont seem tied to any particular coordinate system.

7.4.6 Conclusion

Based on a review of previous literature, as well as novel comparisons of scene-related regions with data-driven clustering analyses, we have proposed a unifying framework for understanding the neural systems involved in processing both visual and non-visual properties of natural scenes. This new two-network classification system makes explicit the relationships between known scene-sensitive regions, re-emphasizes the importance of the functional subdivision within the PPA, and incorporates posterior parietal cortex as a primary component of the scene-understanding system. Our proposal, that much of the scene-processing network relates more to contextual and navigational information than to specific visual features, suggests that experiments

with unfamiliar natural scene images will give only a partial picture of the neural processes evoked in real-world places. Experiencing our visual environment requires a dynamic cooperation between distinct cortical systems, to extract information from the current view of a scene and then integrate it with our understanding of the world and determine our place in it.

7.5 Acknowledgements

Funding was provided by a National Science Foundation Graduate Research Fellowship under grant number DGE-0645962, and Office of Naval Research Multidisciplinary University Research Initiative grant number N000141410671. Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. We thank the Richard M. Lucas Center for Imaging, the Center for Cognitive and Neurobiological Imaging, and Michael Arcaro for helpful discussions.

Chapter 8

Conclusion

What makes a scene more than just a collection of objects? This work has provided several new answers to this question, and has also produced several novel methods for neuroimaging analysis that can be applied to many other questions about the human brain.

The emergent features present in a scene were first investigated in Chapter 2, in which subjects were shown objects, people, groups of noninteracting objects and people, and real human-object interactions. Decoding and cross-decoding analyses revealed that a collection of regions, especially the posterior superior temporal sulcus (pSTS), only represented the category of the stimuli drawn from real interactions and had category representations that were not predictable from the sum of individual object and human response patterns. This points to a neural mechanism underlying the perception of social features in multi-component scenes, which are not present in individual humans or objects.

Chapter 3 then tackled the question of emergent features in general scenes, full environments composed of many components. The results of this large-scale study found that the high-level meaning of a scene can be well-captured by the actions one could perform in that scene. In comparison to models based on individual scene components (objects or visual features), this high-level functionality description was a better predictor of which types of scenes participants thought were most similar. Scene functions are therefore a critical part of scene representation, and are not simply

inherited from the collection of objects in the image.

All of these experiments involved unfamiliar images with no associated memories or places for the participants. In real-world scene perception, almost every scene we experience has associated contextual and navigation information, which is not present in the image but must be retrieved from a representation in memory. Chapters 4 and 5 investigated how memory systems and visual perception systems intersect in the parahippocampal place area (PPA), the most prominent brain region underlying scene perception. Chapter 4 developed a set of tools for measuring fine-grained connectivity differences at the millimeter scale, which were then applied in Chapter 5 to show that the PPA consists of multiple subregions along the anterior-posterior axis, connected separately to visual and memory regions.

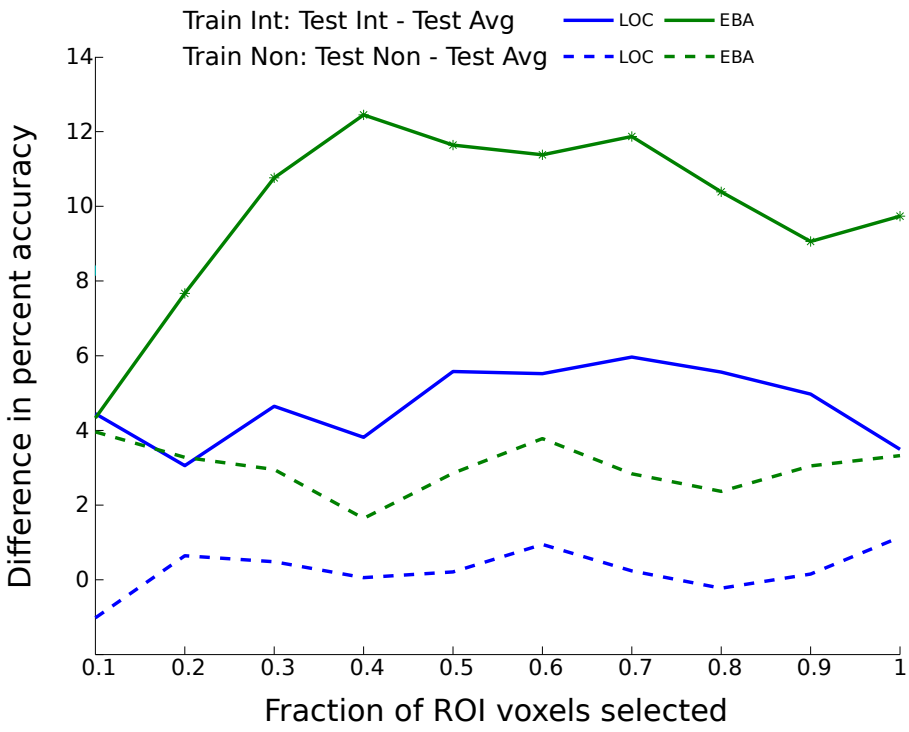
This visual vs. contextual division was extended into a general framework for understanding all scene-related processing in Chapters 6 and 7. Chapter 6 introduced a method for producing precise gray-matter clusters based on connectivity differences, and this parcellation was used in Chapter 7 to identify two distinct networks underlying scene perception: an occipital network responsible for processing the current visual input, and a parietal/medial-temporal network that connected visual information with long-term memories. This new organizing principle for scene processing shows another critical way in which scenes are not the sum of their parts: scenes evoke representations of a real location in time and space, and a large part of the brain's scene processing machinery is specialized for grounding a visual scene to a real-world place.

There are still many unanswered questions about high-order features of real-world scenes. How precisely are these features (such as functionality) implemented, at an algorithmic level and at a physical neural circuit level? How do these networks develop over a lifetime, from young children to older adults? How are these representations affected by task goals or demands? Are there individual differences in the way scenes or scene categories are encoded? What is the causal role of each component in the scene-processing system, and how can we help patients with damage to one or more of these regions? Answering these questions will require new computational and imaging techniques, larger-scale studies, and more effective translational research. This work

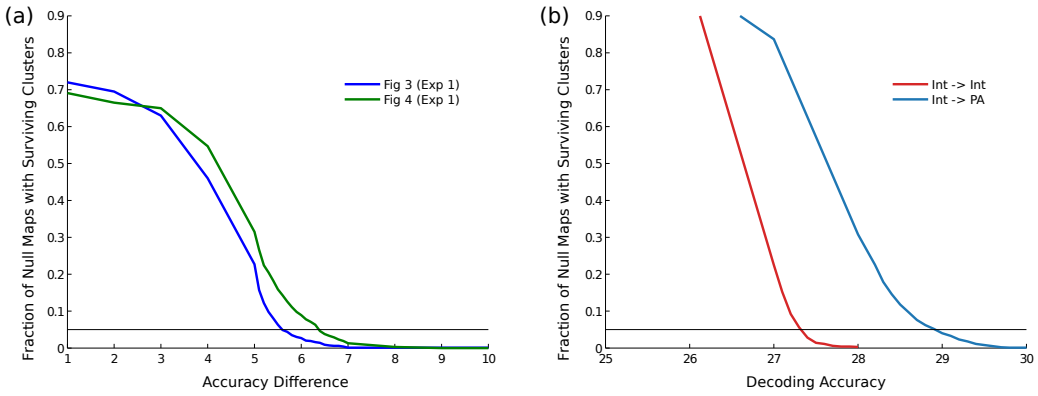
provides a starting point for diving deeper into the neural networks underlying the mysterious and critical mechanisms of visual scene perception.

Appendix A

Human-object interactions are more than the sum of their parts



Supplementary Figure A1: **Robustness of cross-decoding result to number of voxels selected**. The fraction of voxels selected for classifier training (based on overall visual responsiveness) did not have a major impact on the results reported in Figure 2. As long as at least 20% of the voxels in all areas were used in training, the same pattern of significant results can be shown. Starred points are those that are significantly greater than zero ($p<0.05$ one-tailed t-test).



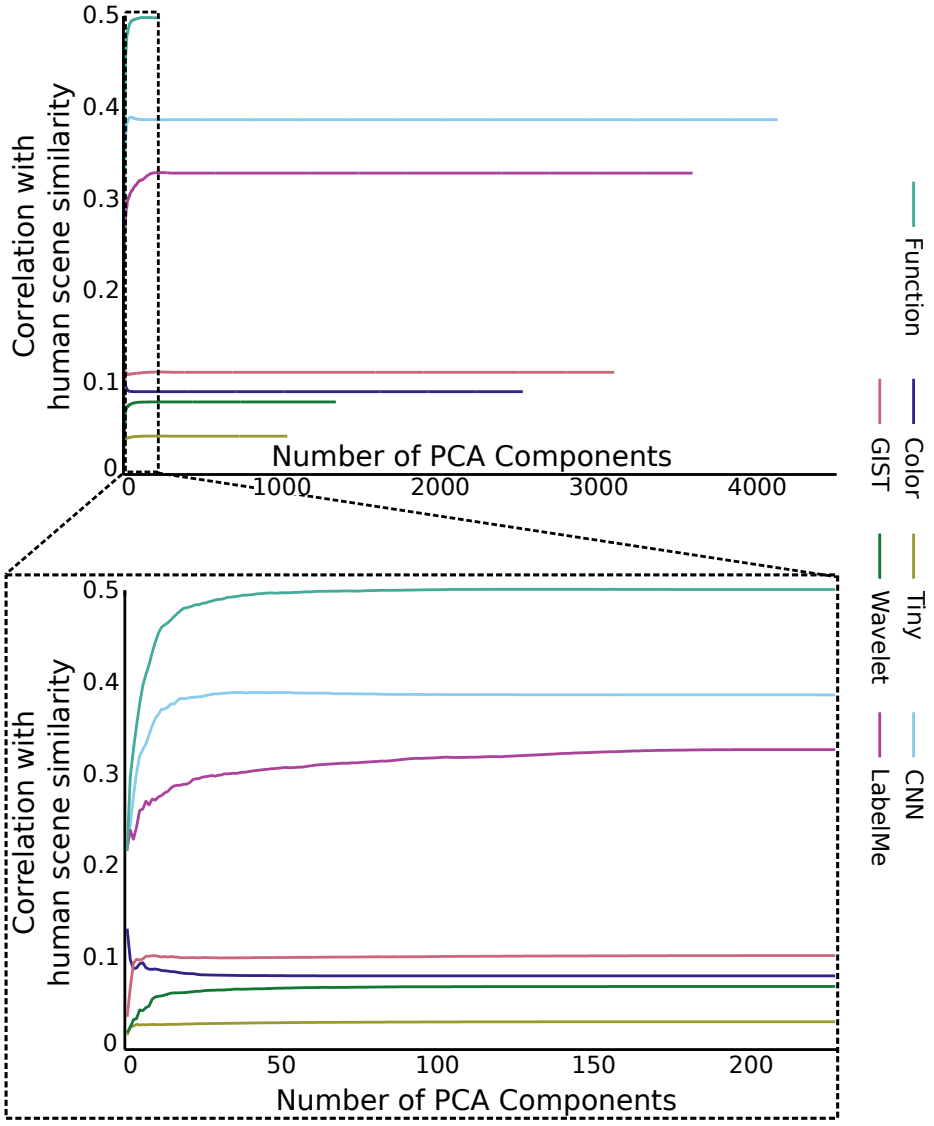
Supplementary Figure A2: **Determination of decoding significance threshold**. 1,000 null searchlight maps were generated for each of the searchlight analyses, by randomly permuting the stimulus labels for each classifier and then running the decoding searchlight. A threshold was chosen for each searchlight such that fewer than 5% of the null difference maps yielded false positive clusters larger than 100 voxels. (a) For experiment 1, in which we are measuring differences between classifier accuracies, we obtain thresholds of 5.6 and 6.4. (b) For experiment 2, in which we are measuring 4-way decoding accuracies, we obtain thresholds of 27.4 and 29.0.

Appendix B

Visual Scenes are Categorized by Function

Scene Scores		Action Correlations	
Most Positive	Most Negative	Most Positive	Most Negative
Dim 1: 45%			
bayou	conference center	Farming / Fishing and Forestry work	Arts / Design / Entertainment / Sports / Media work
swamp	music store	Travel	Socializing
chaparral	piano store	Hiking	Volunteer at event
forest	conference hall	Science work	Eating & drinking
waterfall	discotheque	Fishing	Sales work
waterfall (cascade)	video store	Camping	Attending school-related meetings & conferences
ice shelf	music studio	Rock climbing / caving	Attending meetings for personal interest
rainforest	theater (indoor round)	Watching fishing	Volunteer work: fundraising
sea cliff	movie theatre (indoor)	Hunting	Listening to music (not radio)
wheat field	bakery	Walking	Community and Social work
Dim 2: 20%			
wrestling ring	drill rig	Arts / Design / Entertainment / Sports / Media work	Income-generating services
bullpen	drugstore	Work-related sports	Architecture and Engineering work
velodrome (indoor)	auto factory	Playing games	Construction and Extraction work
batting cage (indoor)	call center	Playing sports with children	Sales work
batting cage (outdoor)	cubicle	Extracurricular club activities	Work-related eating/drinking
aquatic theater	control tower (indoor)	Hobbies	Job search activities
arena (basketball)	office cubicles	Watching weightlifting	Business and Financial Operations work
bullring	kitchenette	Playing basketball	Homework
stadium	pharmacy	Attending child's events	Food & drink preparation
track	chemical plant	Working out	Interior decoration & repair
Dim 3: 9%			
railway yard	indoor chicken farm	In transit / traveling	Farming / Fishing and Forestry work
tunnel (road outdoor)	pig farm	Transportation and Material Moving work	Food & drink preparation
arrival gate	vegetable garden	Architecture and Engineering work	Food Preparation and Serving work
access road	hayfield	Installation / Maintenance and Repair work	Food presentation
highway	dairy (indoor)	Travel	Eating & drinking
tunnel (rail outdoor)	corn field	Construction and Extraction work	Purchasing food (not groceries)
subway interior	pantry	Using vehicle maintenance & repair services	Volunteer work: food preparation
truss bridge	bakery	Security screening	Exercising & playing with animals
tollbooth	chicken yard	Vehicle repair & maintenance (self)	Kitchen & food clean-up
heliport	delicatessen	Attending museums	Using meal preparation services
Dim 4: 7%			
pump room	dining car	Architecture and Engineering work	In transit / traveling
nuclear power plant (indoor)	car interior	Construction and Extraction work	Eating & drinking
particle accelerator	bus interior	Production work	Travel
power plant (indoor)	airplane cabin	Installation / Maintenance and Repair work	Food Preparation and Serving work
water treatment plant (indoor)	limousine interior	Science work	Food presentation
bindery	restaurant	Computer and Mathematical work	Purchasing food (not groceries)
oil refinery	ice cream shop	Business and Financial Operations work	Food & drink preparation
machine shop	liquor store (indoor)	Using home repair & construction services	Transportation and Material Moving work
electrical substation	bus depot	Building and Grounds Cleaning and Maintenance work	Grocery shopping
rolling mill	subway interior	Income-generating services	Using meal preparation services

Supplementary Figure B1: **Principal components of action matrix**. MDS was performed on the scene by action matrix, yielding a coordinate for each scene along each MDS dimension, as well as a correlation between each action and each dimension.



Supplementary Figure B2: **Robustness to dimensionality reduction**. For each feature space, we reconstructed the feature matrix using a variable number of PCA components and then correlated the cosine distance in this feature space with the human scene distances. Although the number of features varies widely between spaces, all can be described in 100 dimensions, and the ordering of how well the features predict human responses is essentially the same regardless of the number of dimensions.

List of Affordances

- Personal care
  - Health related self-care
  - Sexual activity
  - Sleeping
  - Washing/dressing/grooming oneself
- Household activities
  - Appliance repair & maintenance (self)
  - Building & repairing furniture
  - Cleaning home exterior
  - Email
  - Exercising & playing with animals
  - Exterior home repair & decoration
  - Financial management
  - Food & drink preparation
  - Food presentation
  - Grocery shopping
  - Home heating / cooling
  - Home security
  - Home-schooling children
  - Household organization & planning
  - Interior decoration & repair
  - Interior home cleaning
  - Kitchen & food clean-up
  - Laundry
  - Lawn/garden & plant care

  - Mailing
  - Maintaining home pool/pond/hot tub
  - Non-veterinary pet care
  - Sewing & repairing textiles
  - Storing household items
  - Vehicle repair & maintenance (self)
- Caring for & helping household members
  - Arts & crafts with children
  - Attending childs events
  - Helping adult
  - Helping child with homework
  - Looking after adult
  - Looking after children
  - Obtaining medical care for adult
  - Obtaining medical care for child
  - Organizing & planning for adults
  - Organizing & planning for children
  - Physical care of adult
  - Physical care of children
  - Picking up / dropping off adult
  - Picking up / dropping off child
  - Playing sports with children
  - Playing with children (not sports)
  - Providing medical care to adult
  - Providing medical care to child
  - Reading with children

  - Talking with children
- Work & work-related activities
  - Architecture & engineering work
  - Arts / Design / Entertainment / Sports / Media work
  - Building and Grounds Cleaning and Maintenance work
  - Business and Financial Operations work
  - Community and social work
  - Computer and mathematical work
  - Construction and Extraction work
  - Education and library work
  - Farming / Fishing and Forestry work
  - Food Preparation and Serving work
  - Healthcare work
  - Income-generating hobbies & crafts
  - Income-generating performance
  - Income-generating rental property activity
  - Income-generating selling activities
  - Income-generating services
  - Installation / Maintenance and Repair work
  - Job interviewing
  - Job search activities
  - Legal work
  - Management/Executive work
  - Military work
  - Office and Administrative work
  - Personal Care and Service work

  - Production work
  - Protective services work
  - Sales work
  - Science work
  - Transportation and Material Moving work
  - Work-related eating/drinking
  - Work-related social activities
  - Work-related sports
- Education
  - Attending school-related meetings & conferences
  - Education-related administrative activities
  - Extracurricular club activities
  - Homework
  - School music activities
  - Student government
  - Taking class for degree or certification
  - Taking class for personal interest
- Consumer purchases
  - Comparison shopping
  - Purchasing food (not groceries)
  - Purchasing gasoline
  - Shopping (except food and gas)
- Professional & personal care services
  - Banking
  - Buying & selling real estate
  - Out-of-home medical services

  - Using clothing repair & cleaning services
  - Using legal services
  - Using meal preparation services
  - Using other financial services
  - Using personal care services
  - Using professional photography services
  - Using vehicle maintenance & repair services
  - Using veterinary services
- Household services
  - Using home repair & construction services
  - Using in-home medical services
  - Using interior home cleaning services
  - Using lawn & garden services
  - Using paid childcare services
  - Using pet services
- Government services & civic obligations
  - Civic obligations
  - Obtaining licenses & paying fees
  - Security screening
  - Using police & fire services
  - Using social services
  - Waiting
- Eating & drinking
  - Eating & drinking
- Socializing, relaxing & leisure
  - Arts & crafts

  - Attending meetings for personal interest
  - Attending movies
  - Attending museums
  - Attending or hosting parties
  - Attending the performing arts
  - Collecting as a hobby
  - Computer use (not games)
  - Dancing
  - Gambling
  - Hobbies
  - Listening to music (not radio)
  - Listening to radio
  - Playing games
  - Reading for personal interest
  - Relaxing
  - Socializing
  - Tobacco use
  - Watching television & movies
  - Writing for personal interest
- Sports, exercise & recreation
  - Biking
  - Boating
  - Bowling
  - Camping
  - Doing aerobics
  - Doing gymnastics

- Doing martial arts
- Fencing
- Fishing
- Golfing
- Hiking
- Hunting
- Participating in aquatic sports
- Participating in equestrian sports
- Participating in rodeo
- Playing baseball
- Playing basketball
- Playing billiards
- Playing football
- Playing hockey
- Playing racquet sports
- Playing rugby
- Playing soccer
- Playing softball
- Playing volleyball
- Rock climbing / caving
- Rollerblading / skateboarding
- Running
- Skiing / ice skating / snowboarding
- Using cardiovascular equipment
- Vehicle racing/touring
- Walking

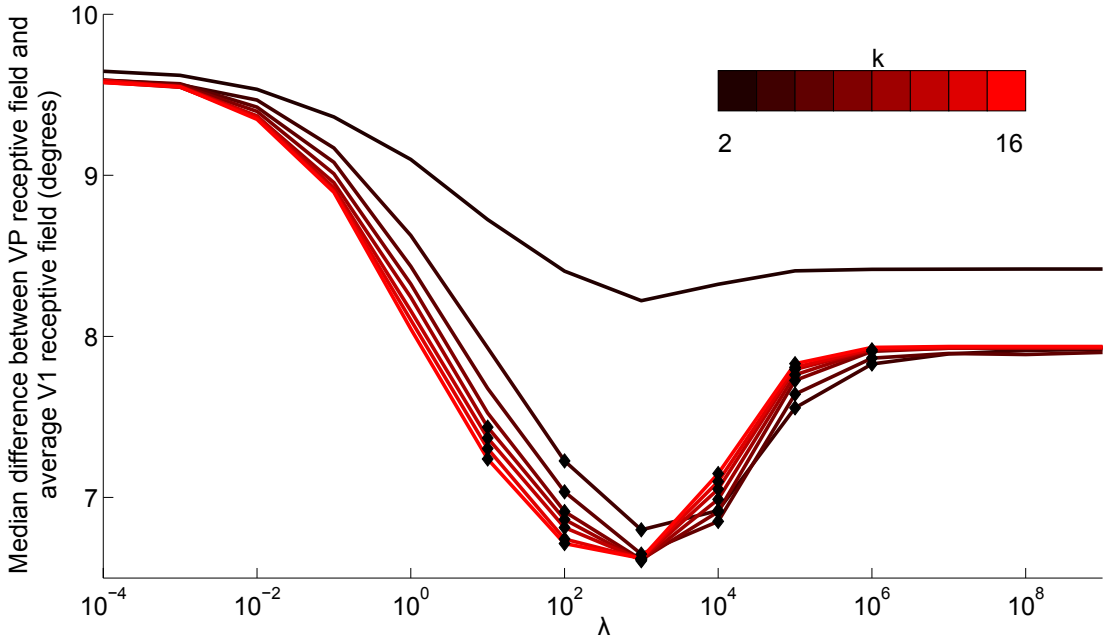
- Watching aerobics
- Watching aquatic sports
- Watching biking
- Watching billiards
- Watching boating
- Watching bowling
- Watching dance
- Watching equestrian sports
- Watching fencing
- Watching fishing
- Watching golf
- Watching gymnastics
- Watching hockey
- Watching live baseball
- Watching live basketball
- Watching live football
- Watching live soccer
- Watching live softball
- Watching live vehicle racing
- Watching martial arts
- Watching people walk
- Watching racquet sports
- Watching rock climbing / caving
- Watching rodeo
- Watching rollerblading / skateboarding
- Watching rugby

  - Watching running
  - Watching skiing / snowboarding
  - Watching volleyball
  - Watching weightlifting
  - Watching wrestling
  - Weightlifting
  - Working out
  - Wrestling
  - Yoga
- Religious & spiritual activities
  - Attending religious services
  - Religious education
  - Religious practices
- Volunteer activities
  - Volunteer at event
  - Volunteer work: attending meeting
  - Volunteer work: blood donation
  - Volunteer work: building
  - Volunteer work: clean up
  - Volunteer work: collecting goods
  - Volunteer work: computer use
  - Volunteer work: food preparation
  - Volunteer work: fundraising
  - Volunteer work: organizing
  - Volunteer work: performing
  - Volunteer work: providing care

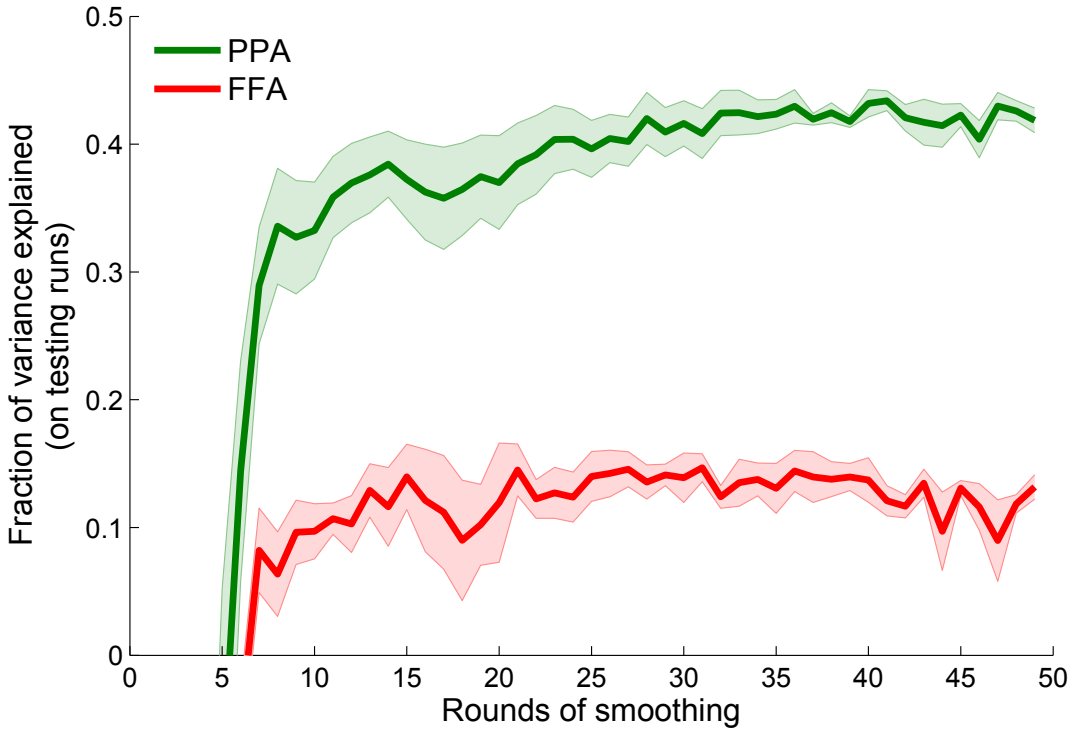
  - Volunteer work: public safety
  - Volunteer work: reading
  - Volunteer work: teaching
  - Volunteer work: telephone calls
  - Volunteer work: writing
- Telephone calls
  - Telephone calls
- Traveling
  - In transit / traveling
  - Travel

Appendix C

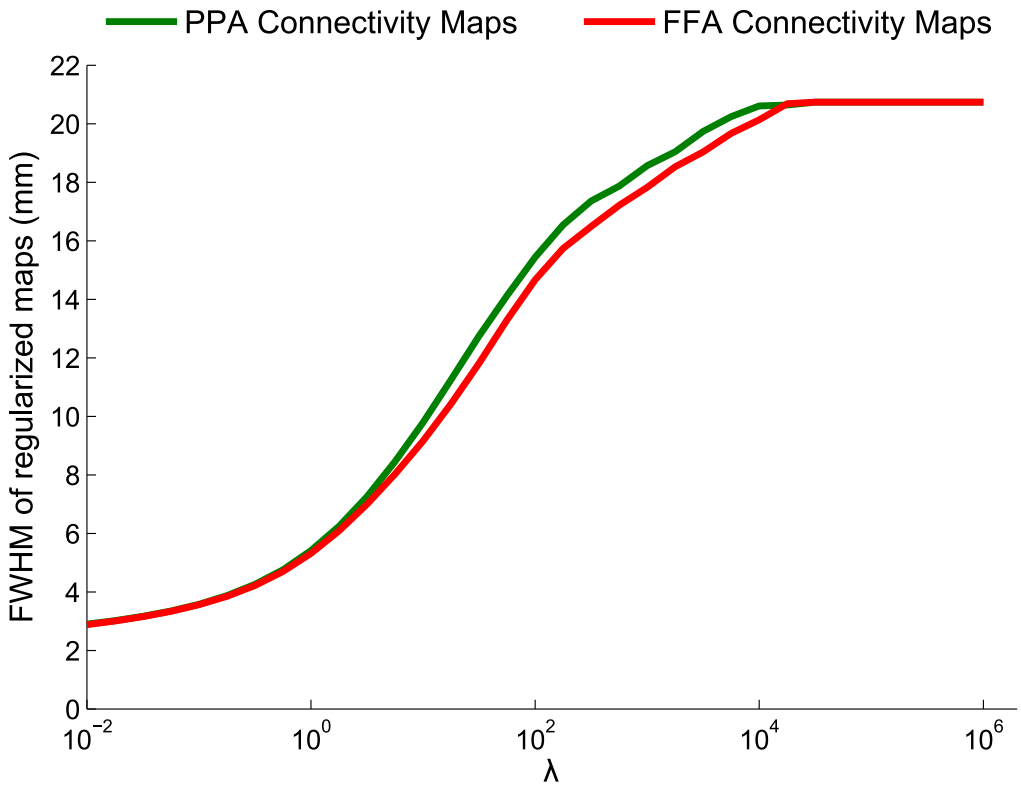
Spatially-regularized voxel-level connectivity



Supplementary Figure C1: **Effects of λ and k parameters on V1-VP connectivity**. We calculate the median difference between the VP receptive fields and the receptive fields generated by the V1 connectivity map for the VP voxels, averaged across subjects (smaller is better). Each curve corresponds to a k value between 2 and 16, and the x-axis corresponds to the λ value (log scale). Diamonds indicate (λ, k) combinations that give a significant reduction in error, compared with using a single weight for all voxels ($\lambda \to \infty$) ($t(12) > 1.78, p < 0.05$, one-tailed t-test). Improvement over the traditional approach is observed over a wide range of λ values (10^1 through 10^6), and for all $k > 2$.



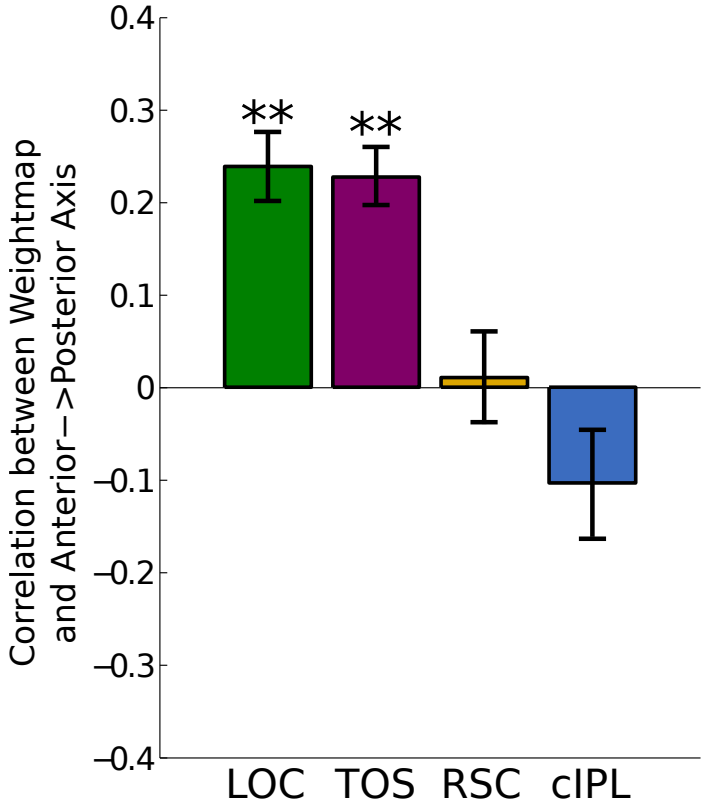
Supplementary Figure C2: **hV4 connectivity results using only pre-smoothing**. To demonstrate that spatial regularization is not equivalent to pre-smoothing, we smoothed the input data and then learned hV4 connectivity weights without regularization ($\lambda = 0$). This smoothing was performed by iteratively averaging the time-course of a voxel with those of its neighbors, for a given number of rounds ($k = 10$). The generalization performance of the learned hV4 maps on held-out testing data is plotted for seed regions PPA and FFA. In both cases, the generalization accuracy simply asymptotes as smoothing increases, and we are unable to identify non-constant maps that give better performance than constant maps. Our results with regularization (Fig. 5, top) are qualitatively different, since intermediate values of λ give a peak in prediction accuracy (achieving a performance level higher than any amount of pre-smoothing). The shaded region indicates standard error.



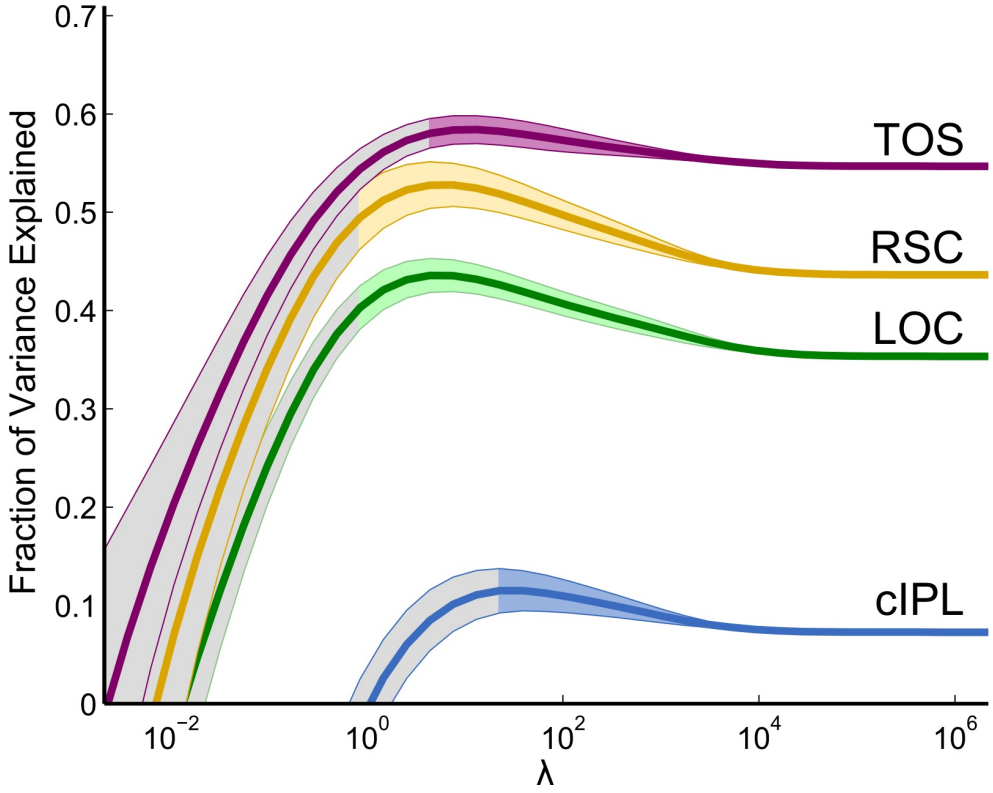
Supplementary Figure C3: **Smoothness of learned maps as a function of** λ. To quantify the relationship between the regularization strength λ and spatial smoothness, we compute the average FWHM (full width at half maximum) for the learned hV4 connectivity maps [314]. As $\lambda \to 0$, maps vary at the scale of individual voxels, while as $\lambda \to \infty$, maps are constant across the entire ROI.

Appendix D

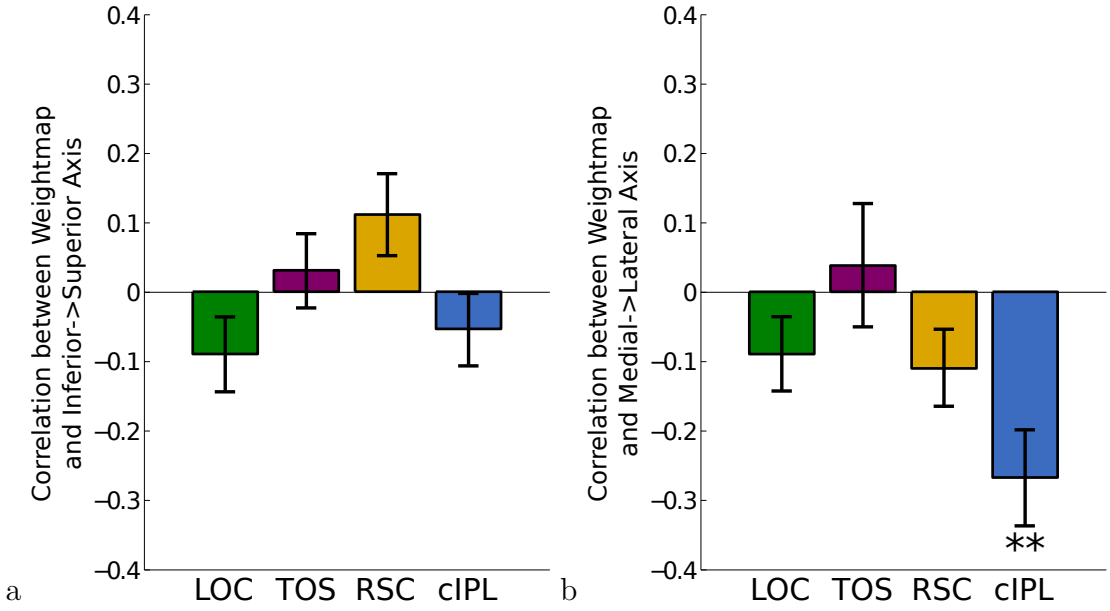
Differential Connectivity Within the Parahippocampal Place Area



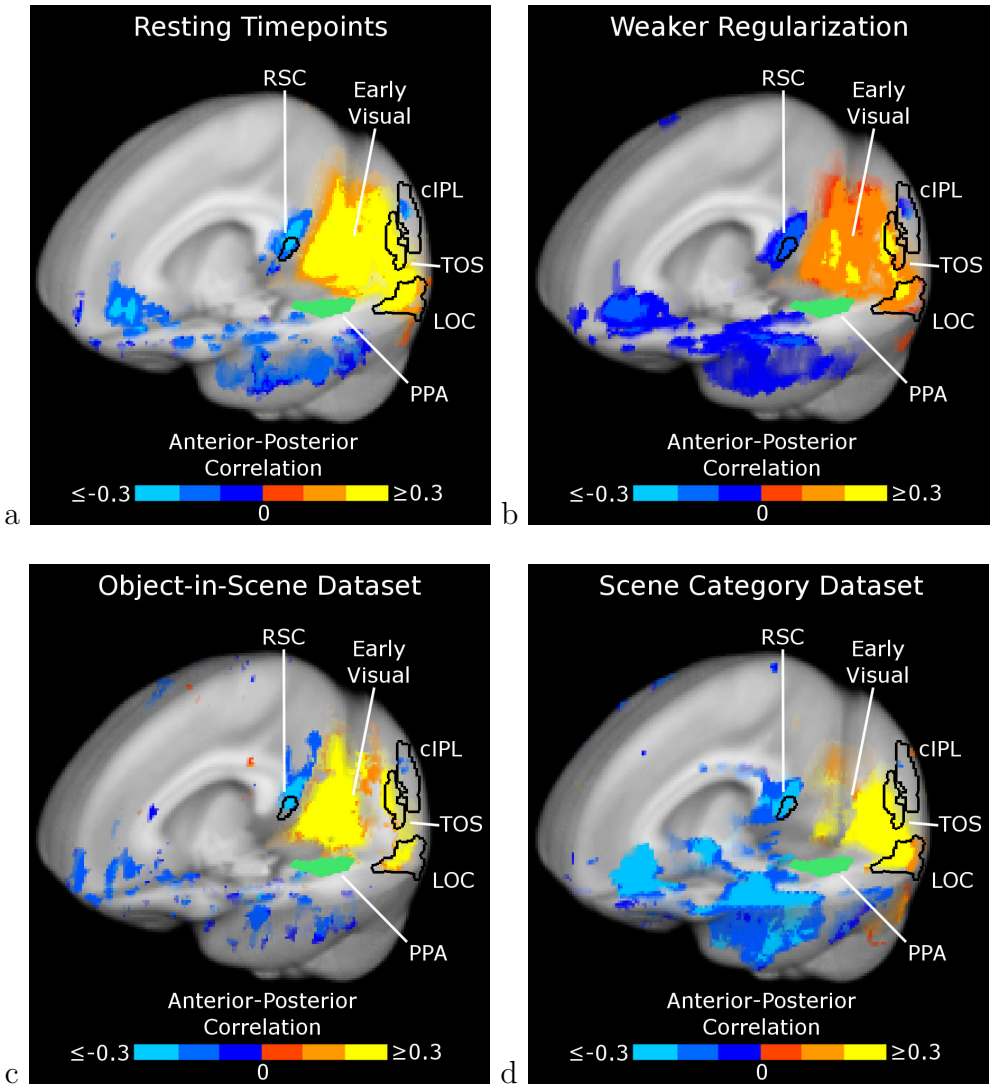
Supplementary Figure D1: **Weightmaps Learned by Voxelwise Correlation**. Rather than using our regularized connectivity approach, here the weight of each voxel for its connectivity with a seed ROI is simply set to the correlation between that voxel's timecourse and the seed ROI timecourse (constrast with main paper Fig. 2b). Although this approach can successfully detect that LOC and TOS are preferentially connected to posterior PPA, it fails to show significant effects for RSC and cIPL (LOC: $t_{17} = 6.02, p < 0.01$; TOS: $t_{17} = 7.03, p < 0.01$; RSC: $t_{17} = 0.22, p = 0.83$; cIPL: $t_{17} = -1.81, p = 0.09$; two-tailed t-test after z-transform). Error bars represent s.e.m. across subjects, ** $p < 0.01$.



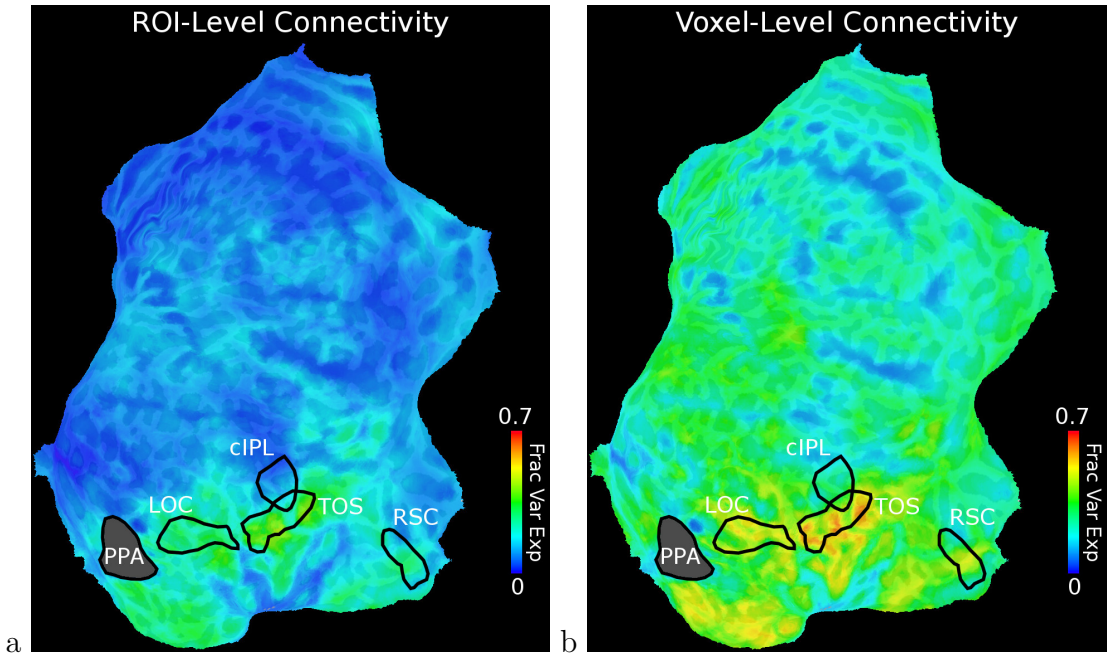
Supplementary Figure D2: **Predictive performance of the connectivity model with varying regularization strength**. After learning a map of connectivity weights over PPA for each seed region (LOC, TOS, RSC, and cIPL) using one run, we measured how well the weighted average of PPA timecourses predicted the mean seed timecourse on the held-out runs. The X-axis (log scale) indicates the strength of spatial regularization applied; at the left side of the graph voxel weights are estimated independently, while the right endpoint corresponds to the traditional connectivity model in which only constant weight maps are learned. Intermediate regularization values (colored) produce better significantly generalization accuracy than those at the endpoints of the graph. This improvement occurs for a wide range of regularization strengths λ (LOC: $10^{-0.07} < \lambda < 10^{6.58}$; TOS: $10^{0.64} < \lambda < 10^{5.63}$; RSC: $10^{-0.07} < \lambda < 10^{6.34}$; cIPL: $10^{1.36} < \lambda < 10^{6.10}$; $t_{17} > 1.74, p < 0.05$ one-tailed t-test, uncorrected). The error bars indicate the standard deviation across subjects (controlling for performance as $\lambda \to \infty$).



Supplementary Figure D3: **Weightmap Correlations along Other PPA Axes**. (a) The weightmaps for all areas show little correlation with the inferior to superior axis (LOC: $t_{17} = -1.71, p = 0.11$; TOS: $t_{17} = 0.63, p = 0.54$; RSC: $t_{17} = 1.87, p = 0.08$; cIPL: $t_{17} = -1.03, p = 0.32$; two-tailed t-test after z-transform). (b) Along the medial to lateral axis, cIPL is connected preferentially to the medial side of PPA, but other regions show no significant biases (LOC: $t_{17} = -1.73, p = 0.10$; TOS: $t_{17} = 0.55, p = 0.59$; RSC: $t_{17} = -1.95, p = 0.07$; cIPL: $t_{17} = -3.55, p < 0.01$; two-tailed t-test after z-transform). Error bars represent s.e.m. across subjects, ** $p < 0.01$.

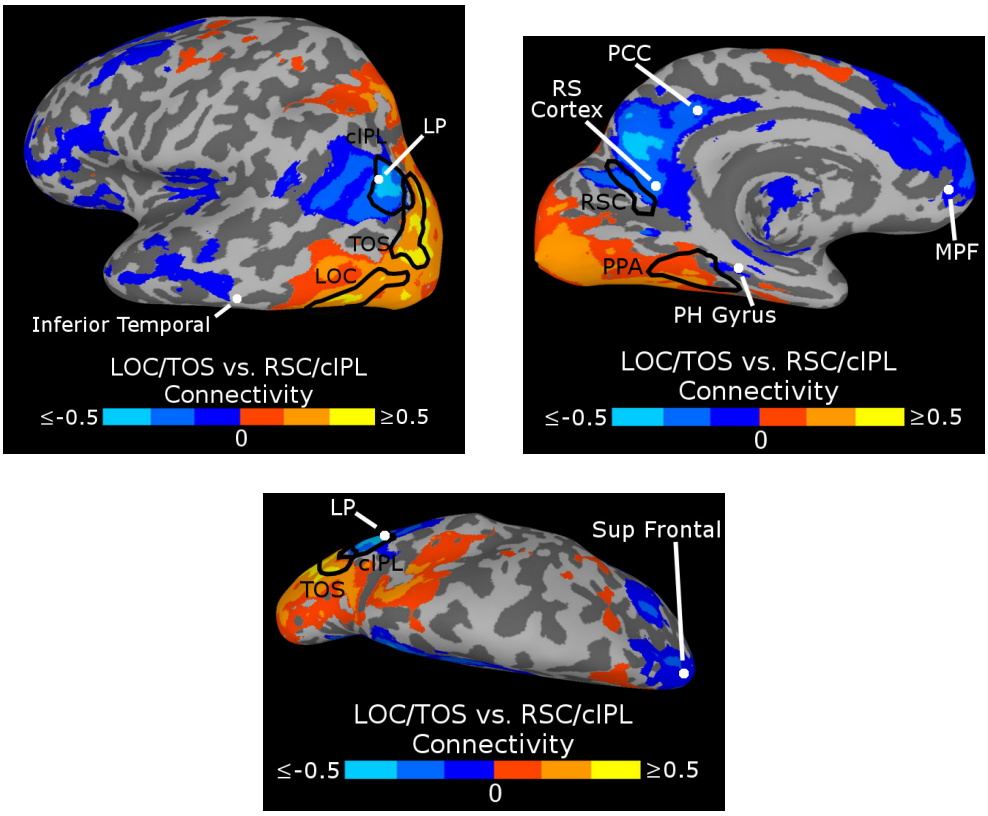


Supplementary Figure D4: **Robustness of Connectivity Result to Task and Regularization Parameter**. (a) Using only "resting" timepoints between stimulus blocks yields similar results as when using all timepoints (FDR < 0.01, cluster size $> 300\text{mm}^3$). (b) Rather than selecting an optimal regularization parameter using leave-one-run-in cross validation, we can optimize our regularization using leave-one-run-out cross validation, resulting in a smaller value of $\lambda = 0.54$. This does not change the overall pattern of connectivity. (c-d) Results for each set of subjects in the two experiments are similar to the whole-group results. These maps are thresholded at $p = 0.01$ (uncorrected) to show the trends in these smaller sample sizes.



a b

Supplementary Figure D5: **Fraction of Variance Explained in Searchlight Analysis**. The fraction of variance explained for each searchlight seed by PPA was calculated for both (a) the ROI-level method (using a spatially constant connectivity map over each PPA hemisphere, i.e. $\lambda \to \infty$) and (b) the voxel-level method. The fraction of variance explained by each voxel was computed as the average value of all searchlights including that voxel. Both methods show similar trends, with regions near LOC, TOS, and RSC having a large amount of shared variance with PPA, and other regions less related to PPA. The connectivity is substantially stronger overall for the voxel-level method, consistent with our results for the individual ROIs (main paper Fig. 2a).



Supplementary Figure D6: **LOC/TOS vs. RSC/cIPL Connectivity**. The data from Figure 6 is shown here across the entire inflated surface (FDR <0.05, cluster size >1000 mm^3). The Talairach coordinates of the cortical Default Mode Network (DMN) regions identified by [100] are indicated with white dots. Voxels showing the same connectivity pattern as anterior PPA (RSC/cIPL connectivity greater than LOC/TOS connectivity) overlap closely with the DMN regions, showing that our RSC and cIPL regions are key components of this network.

Bibliography

[1] A. Abraham, E. Dohmatob, B. Thirion, D. Samaras, and G. Varoquaux. "Extracting Brain Regions from Rest fMRI with Total-Variation Constrained Dictionary Learning". *MICCAI 2013*. 2013.

[2] G. K. Aguirre, E Zarahn, and M D'Esposito. "An area within human ventral cortex sensitive to building stimuli: evidence and implications." *Neuron* 21.2 (1998), pp. 373–383.

[3] C. Aicher, A. Z. Jacobs, and A. Clauset. "Learning latent block structure in weighted networks". *Journal of Complex Networks* (2014).

[4] U Alon. "Biological networks: the tinkerer as an engineer." *Science (New York, N.Y.)* 301.5641 (Sept. 2003), pp. 1866–7.

[5] E. Aminoff, N. Gronau, and M. Bar. "The parahippocampal cortex mediates spatial and nonspatial associations". *Cereb. Cortex* 17 (July 2007), pp. 1493–1503.

[6] E. Aminoff, D. L. Schacter, and M. Bar. "The cortical underpinnings of context-based memory distortion." *Journal of cognitive neuroscience* 20.12 (Dec. 2008), pp. 2226–37.

[7] J. R. Andrews-Hanna, J. S. Reidler, J. Sepulcre, R. Poulin, and R. L. Buckner. "Functional-anatomic fractionation of the brain's default network." *Neuron* 65.4 (Feb. 2010), pp. 550–62.

[8] J. R. Andrews-Hanna, J. Smallwood, and R. N. Spreng. "The default network and self-generated thought: component processes, dynamic control, and clinical relevance." *Annals of the New York Academy of Sciences* 1316 (May 2014), pp. 29–52.

[9] M. Arcaro, S. McMains, B. Singer, and S. Kastner. "Retinotopic organization of human ventral visual cortex". *J. Neurosci.* 29 (Aug. 2009), pp. 10638–10652.

[10] M. van Assche, V. Kebets, P. Vuilleumier, and F. Assal. "Functional Dissociations Within Posterior Parietal Cortex During Scene Integration and Viewpoint Changes". *Cereb Cortex* (Sept. 2014), bhu215.

[11] F. Attneave. "Dimensions of similarity". *Am J Psychol* 63.4 (Oct. 1950), pp. 516–556.

[12] A. Baddeley. "The episodic buffer: a new component of working memory?" *Trends in Cognitive Sciences* 4.11 (Nov. 2000), pp. 417–423.

[13] A. Baeck, J. Wagemans, and H. P. Op de Beeck. "The distributed representation of random and meaningful object pairs in human occipitotemporal cortex: The weighted average as a general rule". *NeuroImage* 70 (2013), pp. 37–47.

[14] K. Baker. "Identity, Memory and Place". *The Word Hoard* 1.1 (2012), Article 4.

[15] C. Baldassano, D. M. Beck, and L. Fei-Fei. "Differential connectivity within the Parahippocampal Place Area". *Neuroimage* 75 (July 2013), pp. 228–237.

[16] C. Baldassano, D. M. Beck, and L. Fei-Fei. "Parcellating connectivity in spatial maps". *PeerJ* 3.e784 (2015).

[17] C. Baldassano, M. C. Iordan, D. M. Beck, and L. Fei-Fei. "Discovering Voxel-Level Functional Connectivity Between Cortical Regions". *2nd NIPS Workshop on Machine Learning and Interpretation in Neuroimaging*. 2012.

[18] C. Baldassano, M. C. Iordan, D. M. Beck, and L. Fei-Fei. "Voxel-level functional connectivity using spatial regularization". *Neuroimage* 63.3 (July 2012), pp. 1099–1106.

[19] M. Bar. "Visual objects in context". *Nat. Rev. Neurosci.* 5 (Aug. 2004), pp. 617–629.

[20] M. Bar and E. Aminoff. "Cortical analysis of visual context." *Neuron* 38.2 (2003), pp. 347–58.

[21] Z. Bar-Joseph, D. K. Gifford, and T. S. Jaakkola. "Fast optimal leaf ordering for hierarchical clustering". *Bioinformatics* 17 Suppl 1 (2001), S22–29.

[22] A.-L. Barabási and Z. N. Oltvai. "Network biology: understanding the cell's functional organization." *Nature reviews. Genetics* 5.2 (Feb. 2004), pp. 101–13.

[23] T. E. Behrens, H. J. Berg, S. Jbabdi, M. F. Rushworth, and M. W. Woolrich. "Probabilistic diffusion tractography with multiple fibre orientations: What can we gain?" *Neuroimage* 34.1 (Jan. 2007), pp. 144–155.

[24] K. C. Bettencourt and Y. Xu. "The role of transverse occipital sulcus in scene perception and its relationship to object individuation in inferior intraparietal sulcus". *Journal of Cognitive Neuroscience* 25 (2013), pp. 1711–1722.

[25] I. Biederman. "Recognition-by-components: a theory of human image understanding". *Psychol Rev* 94.2 (Apr. 1987), pp. 115–147.

[26] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. "Scene perception: detecting and judging objects undergoing relational violations". *Cogn Psychol* 14.2 (Apr. 1982), pp. 143–177.

[27] D. M. Blei and P. I. Frazier. "Distance Dependent Chinese Restaurant Processes". *J. Mach. Learn. Res.* 12 (Nov. 2011), pp. 2461–2488.

[28] T. Blumensath, S. Jbabdi, M. F. Glasser, D. C. Van Essen, K. Ugurbil, T. E. Behrens, and S. M. Smith. "Spatially constrained hierarchical parcellation of the brain with resting-state fMRI". *Neuroimage* 76 (Aug. 2013), pp. 313–324.

[29] E. Borenstein and J. Malik. "Shape Guided Object Segmentation". *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on.* Vol. 1. June 2006, pp. 969–976.

[30] S. Bouvier and R. A. Epstein. "Early vs. late components of category selectivity in the parahippocampal place area: A rapid acquisition fMRI study". Presented at the Vision Sciences Society 11th Annual Meeting, Naples, FL, 2011.

[31] M. Brass, R. M. Schmitt, S. Spengler, and G. Gergely. "Investigating action understanding: inferential processes versus action simulation." *Current Biology* 17.24 (Dec. 2007), pp. 2117–21.

[32] S. Bray, A. E. Arnold, R. M. Levy, and G. Iaria. "Spatial and temporal functional connectivity changes between resting and attentive states". *Hum Brain Mapp* 36.2 (Feb. 2015), pp. 549–565.

[33] M. Brett. *The MNI brain and the Talairach atlas*. `http://imaging.mrc-cbu.cam.ac.uk/imaging/MniTalairach`. Accessed: 2015-01-27. 2002.

[34] A. Brewer and B. Barton. "Visual field map organization in human visual cortex". *Visual Cortex-Current Status and Perspectives*. Ed. by S. Molotchnikoff and J. Rouat. 2012. Chap. 2.

[35] G Buccino, F Binkofski, G. R. Fink, L Fadiga, L Fogassi, V Gallese, R. J. Seitz, K Zilles, G Rizzolatti, A. Parma, V. Volturno, and I Parma. "SHORT COMMUNICATION Action observation activates premotor and parietal areas in a somatotopic manner : an fMRI study". *Neuroscience* 13 (2001), pp. 400–404.

[36] G. Buccino, F. Binkofski, and L. Riggio. "The mirror neuron system and action recognition." *Brain and language* 89.2 (May 2004), pp. 370–6.

[37] G. Buccino, F. Lui, N. Canessa, I. Patteri, G. Lagravinese, F. Benuzzi, C. a. Porro, and G. Rizzolatti. "Neural circuits involved in the recognition of actions performed by nonconspecifics: an FMRI study." *Journal of cognitive neuroscience* 16.1 (2004), pp. 114–26.

[38] R. L. Buckner, J. R. Andrews-Hanna, and D. L. Schacter. "The brains default network: anatomy, function, and relevance to disease." *Annals Of The New*

York Academy Of Sciences 1124.1 (2008). Ed. by Sangeetha-ShyamEditors, pp. 1–38.

[39] A. Buja, D. F. Swayne, M. L. Littman, N. Dean, H. Hofmann, and L. Chen. "Data visualization with multidimensional scaling". *Journal of Computational and Graphical Statistics* (2008).

[40] H. H. Bulthoff, S. Y. Edelman, and M. J. Tarr. "How are three-dimensional objects represented in the brain?" *Cereb. Cortex* 5.3 (1995), pp. 247–260.

[41] N Burgess, E. a. Maguire, H. J. Spiers, and J O'Keefe. "A temporoparietal and prefrontal network for retrieving the spatial context of lifelike events." *NeuroImage* 14.2 (Aug. 2001), pp. 439–53.

[42] M. van Buuren, M. C. W. Kroes, I. C. Wagner, L. Genzel, R. G. M. Morris, and G. Fernandez. "Initial Investigation of the Effects of an Experimentally Learned Schema on Spatial Associative Memory in Humans". *Journal of Neuroscience* 34.50 (Dec. 2014), pp. 16662–16670.

[43] R. H. Byrd, R. B. Schnabel, and G. A. Shultz. "A Trust Region Algorithm for Nonlinearly Constrained Optimization". *SIAM Journal on Numerical Analysis* 24.5 (Oct. 1987), pp. 1152–1170.

[44] P. Byrne, S. Becker, and N. Burgess. "Remembering the past and imagining the future: a neural model of spatial memory and imagery." *Psychological review* 114.2 (Apr. 2007), pp. 340–75.

[45] D. Bzdok, A. Heeger, R. Langner, A. R. Laird, P. T. Fox, N. Palomero-Gallagher, B. A. Vogt, K. Zilles, and S. B. Eickhoff. "Subspecialization in the human posterior medial cortex". *Neuroimage* 106C (Nov. 2014), pp. 55–71.

[46] B Calvo-Merino, D. E. Glaser, J Grèzes, R. E. Passingham, and P Haggard. "Action observation and acquired motor skills: an FMRI study with expert dancers." *Cerebral cortex (New York, N.Y. : 1991)* 15.8 (Aug. 2005), pp. 1243–9.

[47] B. Calvo-Merino, J. Grèzes, D. E. Glaser, R. E. Passingham, and P. Haggard. "Seeing or doing? Influence of visual and motor familiarity in action observation." *Current Biology* 16.19 (Oct. 2006), pp. 1905–10.

[48] F. W. Campbell and J. G. Robson. "Application of Fourier analysis to the visibility of gratings." *The Journal of physiology* 197.3 (Aug. 1968), pp. 551–66.

[49] J. S. Cant and M. A. Goodale. "Attention to form or surface properties modulates different regions of human occipitotemporal cortex." *Cerebral Cortex* 17.3 (2007), pp. 713–731.

[50] J. S. Cant and M. A. Goodale. "Scratching beneath the surface: new insights into the functional properties of the lateral occipital area and parahippocampal place area." *Journal of Neuroscience* 31.22 (2011), pp. 8248–8258.

[51] J. S. Cant and Y. Xu. "Object ensemble processing in human anterior-medial ventral visual cortex." *Journal of Neuroscience* 32.22 (2012), pp. 7685–700.

[52] S. Caspers, S. B. Eickhoff, S. Geyer, F. Scheperjans, H. Mohlberg, K. Zilles, and K. Amunts. "The human inferior parietal lobule in stereotaxic space." *Brain structure function* 212.6 (2008), pp. 481–495.

[53] S. Caspers, S. B. Eickhoff, T. Rick, A. Von Kapri, T. Kuhlen, R. Huang, N. J. Shah, and K. Zilles. "Probabilistic fibre tract analysis of cytoarchitectonically defined human inferior parietal lobule areas reveals similarities to macaques." *NeuroImage* 58.2 (2011), pp. 362–380.

[54] S. Caspers, S. Geyer, A. Schleicher, H. Mohlberg, K. Amunts, and K. Zilles. "The human inferior parietal cortex: cytoarchitectonic parcellation and interindividual variability." *NeuroImage* 33.2 (2006), pp. 430–448.

[55] S. Caspers, A. Schleicher, M. Bacha-Trams, N. Palomero-Gallagher, K. Amunts, and K. Zilles. "Organization of the Human Inferior Parietal Lobule Based on Receptor Architectonics." *Cerebral Cortex* (2012), pp. 1–14.

[56] S. Caspers, K. Zilles, A. R. Laird, and S. B. Eickhoff. "ALE meta-analysis of action observation and imitation in the human brain." *NeuroImage* 50.3 (Apr. 2010), pp. 1148–67.

[57] F. X. Castellanos, A. Di Martino, R. C. Craddock, A. D. Mehta, and M. P. Milham. "Clinical applications of the functional connectome." *NeuroImage* 80 (Oct. 2013), pp. 527–40.

[58] F. Cauda, F. D'Agata, K. Sacco, S. Duca, G. Geminiani, and A. Vercelli. "Functional connectivity of the insula in the resting brain". *Neuroimage* 55.1 (Mar. 2011), pp. 8–23.

[59] C. Cavada and P. S. Goldman-Rakic. "Posterior parietal cortex in rhesus monkey: I. Parcellation of areas based on distinctive limbic and sensory corticocortical connections". *J. Comp. Neurol.* 287.4 (Sept. 1989), pp. 393–421.

[60] M. R. Celis, J. E. D. Jr., and R. A. Tapia. *A Trust Region Strategy for Equality Constrained Optimization*. Tech. rep. 84-1. Mathematical Sciences Department, Rice University, Sept. 1984.

[61] B. Chai, D. Walther, D. Beck, and L. Fei-Fei. "Exploring Functional Connectivity of the Human Brain using Multivariate Information Analysis". *Advances in Neural Information Processing Systems 22*. 2009.

[62] C.-C. Chang and C.-J. Lin. "LIBSVM: a library for support vector machines". *ACM Transactions on Intelligent Systems and Technology* 2.3 (2011). Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`, 27:1–27:27.

[63] L. L. Chen, L.-h. L. Edward, C. A. Barnes, and B. L. Mcnaughton. "Head-direction cells in the rat posterior cortex. I . anatomical distribution and behavioral modulation". *Experimental Brain Research* 101.1 (1994), pp. 24–34.

[64] H. Cheng and Y. Fan. "Semi-supervised clustering for parcellating brain regions based on resting state fMRI data". *SPIE Medical Imaging*. Ed. by S. Ourselin and M. A. Styner. International Society for Optics and Photonics, Mar. 2014, p. 903427.

[65] T. T.-J. Chong, R. Cunnington, M. a. Williams, N. Kanwisher, and J. B. Mattingley. "fMRI adaptation reveals mirror neurons in human inferior parietal cortex." *Current Biology* 18.20 (Oct. 2008), pp. 1576–80.

[66] A. L. Cohen, D. A. Fair, N. U. Dosenbach, F. M. Miezin, D. Dierker, D. C. Van Essen, B. L. Schlaggar, and S. E. Petersen. "Defining functional areas in individual human brains using resting functional connectivity MRI". *Neuroimage* 41 (May 2008), pp. 45–57.

[67] B. Conroy, B. Singer, J. Haxby, and P. Ramadge. "fMRI-Based Inter-Subject Cortical Alignment Using Functional Connectivity". *Advances in Neural Information Processing Systems 22*. 2009.

[68] D. Cordes, V. Haughton, J. D. Carew, K. Arfanakis, and K. Maravilla. "Hierarchical clustering to measure connectivity in fMRI resting-state data". *Magn Reson Imaging* 20.4 (May 2002), pp. 305–317.

[69] R. W. Cox. "AFNI: software for analysis and visualization of functional magnetic resonance neuroimages". *Comput. Biomed. Res.* 29.3 (June 1996), pp. 162–173.

[70] R. C. Craddock, G. A. James, P. E. Holtzheimer, X. P. Hu, and H. S. Mayberg. "A whole brain fMRI atlas generated via spatially constrained spectral clustering". *Hum Brain Mapp* 33.8 (Aug. 2012), pp. 1914–1928.

[71] A. Crippa, L. Cerliani, L. Nanetti, and J. B. T. M. Roerdink. "Heuristics for connectivity-based brain parcellation of SMA/pre-SMA through force-directed graph layout." *NeuroImage* 54.3 (Feb. 2011), pp. 2176–84.

[72] G. Csibra. "Action mirroring and action understanding: An alternative account". *Sensorimotor Foundations of Higher Cognition*. Ed. by P. Haggard, Y. Rosetti, and M. Kawato. Oxford: Oxford University Press, 2007.

[73] R. Cuignet, M. Chupin, H. Benali, and O. Colliot. "Spatial and anatomical regularization of SVM for brain image analysis". *Advances in Neural Information Processing Systems 23*. 2010.

[74] J. C. Culham and K. F. Valyear. "Human parietal cortex in action." *Current opinion in neurobiology* 16.2 (Apr. 2006), pp. 205–12.

[75] J. Decety and C. Lamm. "The role of the right temporoparietal junction in social interaction: how low-level computational processes contribute to meta-cognition." *The Neuroscientist : a review journal bringing neurobiology, neurology and psychiatry* 13.6 (Dec. 2007), pp. 580–93.

[76] F. Deleus and M. M. Van Hulle. "Functional connectivity analysis of fMRI data based on regularized multiset canonical correlation analysis". *J. Neurosci. Methods* 197.1 (Apr. 2011), pp. 143–157.

[77] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. "ImageNet: A Large-Scale Hierarchical Image Database". *CVPR09*. 2009.

[78] J. J. DiCarlo, D. Zoccolan, and N. C. Rust. "How does the brain solve visual object recognition?" *Neuron* 73.3 (Feb. 2012), pp. 415–434.

[79] D. D. Dilks, J. B. Julian, A. M. Paunov, and N. Kanwisher. "The occipital place area is causally and selectively involved in scene perception". *J. Neurosci.* 33.4 (Jan. 2013), pp. 1331–1336.

[80] P. E. Downing, Y. Jiang, M. Shuman, and N. Kanwisher. "A cortical area selective for visual processing of the human body". *Science* 293 (Sept. 2001), pp. 2470–2473.

[81] P. E. Downing and M. V. Peelen. "The role of occipitotemporal body-selective regions in person perception". *Cognitive Neuroscience* 2.3-4 (Sept. 2011), pp. 186–203.

[82] S. Dumoulin and B. A. Wandell. "Population receptive field estimates in human visual cortex". *Neuroimage* 39 (Jan. 2008), pp. 647–660.

[83] K. A. Ehinger, A. Torralba, and A. Oliva. "A taxonomy of visual scenes: Typicality ratings and hierarchical classification". *Journal of Vision* 10 (7 2010), pp. 1237–1237.

[84] H. Eichenbaum, A. P. Yonelinas, and C. Ranganath. "The medial temporal lobe and recognition memory". *Annu. Rev. Neurosci.* 30 (2007), pp. 123–152.

[85] H. Eichenbaum and N. J. Cohen. "Can We Reconcile the Declarative Memory and Spatial Navigation Views on Hippocampal Function?" *Neuron* 83.4 (Aug. 2014), pp. 764–770.

[86] S. B. Eickhoff, D. Bzdok, A. R. Laird, C. Roski, S. Caspers, K. Zilles, and P. T. Fox. "Co-activation patterns distinguish cortical modules, their connectivity and functional differentiation". *Neuroimage* 57.3 (Aug. 2011), pp. 938–949.

[87] S. B. Eickhoff, K. E. Stephan, H. Mohlberg, C. Grefkes, G. R. Fink, K. Amunts, and K. Zilles. "A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data." *NeuroImage* 25.4 (2005), pp. 1325–1335.

[88] J. A. Elman, B. I. Cohn-Sheehy, and A. P. Shimamura. "Dissociable parietal regions facilitate successful retrieval of recently learned and personally familiar information." *Neuropsychologia* 51.4 (Mar. 2013), pp. 573–83.

[89] R Epstein, A Harris, D Stanley, and N Kanwisher. "The parahippocampal place area: recognition, navigation, or encoding?" *Neuron* 23.1 (1999), pp. 115–125.

[90] R. A. Epstein and L. K. Morgan. "Neural responses to visual scenes reveals inconsistencies between fMRI adaptation and multivoxel pattern analysis". *Neuropsychologia* 50.4 (Mar. 2012), pp. 530–543.

[91] R. Epstein. "Parahippocampal and retrosplenial contributions to human spatial navigation". *Trends in Cognitive Sciences* 12.10 (2008), pp. 388 –396.

[92] R. Epstein and N. Kanwisher. "A cortical representation of the local visual environment". *Nature* 392 (Apr. 1998), pp. 598–601.

[93] R. Epstein, K. S. Graham, and P. E. Downing. "Viewpoint-Specific Scene Representations in Human Parahippocampal Cortex". *Neuron* 37.5 (2003), pp. 865 –876.

[94] R. A. Epstein, J. S. Higgins, K. Jablonski, and A. M. Feiler. "Visual scene processing in familiar and unfamiliar environments." *Journal of neurophysiology* 97.5 (May 2007), pp. 3670–83.

[95] R. A. Epstein, W. E. Parker, and A. M. Feiler. "Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 27.23 (2007), pp. 6141–6149.

[96] R. A. Epstein and L. K. Vass. "Neural systems for landmark-based wayfinding in humans". *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 369.1635 (2014).

[97] R. A. Epstein and E. J. Ward. "How reliable are visual context effects in the parahippocampal place area?" *Cerebral Cortex* 20.February (2010), pp. 294–303.

[98] S. L. Fairhall, S. Anzellotti, S. Ubaldi, and A. Caramazza. "Person- and place-selective neural substrates for entity-specific semantic access." *Cerebral cortex (New York, N.Y. : 1991)* 24.7 (July 2014), pp. 1687–96.

[99] L. Fei-Fei, A. Iyer, C. Koch, and P. Perona. "What do we perceive in a glance of a real-world scene?" *J Vis* 7.1 (2007), p. 10.

[100] M. D. Fox, A. Z. Snyder, J. L. Vincent, M. Corbetta, D. C. Van Essen, and M. E. Raichle. "The human brain is intrinsically organized into dynamic, anticorrelated functional networks". *Proc. Natl. Acad. Sci. U.S.A.* 102.27 (July 2005), pp. 9673–9678.

[101] S. Freilich, A. Kreimer, I. Meilijson, U. Gophna, R. Sharan, and E. Ruppin. "The large-scale organization of the bacterial network of ecological co-occurrence interactions." *Nucleic acids research* 38.12 (July 2010), pp. 3857–68.

[102] J. J. Gibson. *The Ecological Approach To Visual Perception*. New Ed. Psychology Press, Sept. 1986.

[103] E. Goesaert and H. P. Op de Beeck. "Continuous mapping of the cortical object vision pathway using traveling waves in object space". *Neuroimage* 49 (Feb. 2010), pp. 3248–3256.

[104] G. Golarai, D. G. Ghahremani, S. Whitfield-Gabrieli, A. Reiss, J. L. Eberhardt, J. D. Gabrieli, and K. Grill-Spector. "Differential development of high-level visual cortex correlates with category-specific recognition memory". *Nat. Neurosci.* 10 (Apr. 2007), pp. 512–522.

[105] P. Golland, Y. Golland, and R. Malach. "Detection of spatial activation patterns as unsupervised segmentation of fMRI data". *Med Image Comput Comput Assist Interv* 10.Pt 1 (2007), pp. 110–118.

[106] Y. Golland, P. Golland, S. Bentin, and R. Malach. "Data-driven clustering reveals a fundamental subdivision of the human cortex into two global systems". *Neuropsychologia* 46.2 (Jan. 2008), pp. 540–553.

[107] J. D. Golomb and N. Kanwisher. "Higher level visual cortex represents retinotopic, not spatiotopic, object location". *Cerebral Cortex* 22.December (2012), pp. 2794–2810.

[108] J. Gonzalez-Castillo, Z. S. Saad, D. A. Handwerker, S. J. Inati, N. Brenowitz, and P. A. Bandettini. "Whole-brain, time-locked activation with simple tasks revealed using massive averaging and model-free analysis". *Proc. Natl. Acad. Sci. U.S.A.* 109.14 (Apr. 2012), pp. 5487–5492.

[109] N. S. Gorbach, C. Schütte, C. Melzer, M. Goldau, O. Sujazow, J. Jitsev, T. Douglas, and M. Tittgemeyer. "Hierarchical information-based clustering for connectivity-based cortex parcellation." English. *Frontiers in neuroinformatics* 5 (Jan. 2011), p. 18.

[110] E. M. Gordon, T. O. Laumann, B. Adeyemo, J. F. Huckins, W. M. Kelley, and S. E. Petersen. "Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations". *Cereb. Cortex* (Oct. 2014).

[111] M. Grant and S. Boyd. *CVX: Matlab Software for Disciplined Convex Programming, version 1.21.* http://cvxr.com/cvx. Apr. 2011.

[112] C. Green and J. E. Hummel. "Familiar interacting object pairs are perceptually grouped." *Journal of experimental psychology. Human perception and performance* 32.5 (Oct. 2006), pp. 1107–19.

[113] M. R. Greene and A. Oliva. "High-level aftereffects to global scene properties". *J Exp Psychol Hum Percept Perform* 36.6 (Dec. 2010), pp. 1430–1442.

[114] M. R. Greene and A. Oliva. "Recognition of natural scenes from global properties: seeing the forest without representing the trees". *Cogn Psychol* 58.2 (Mar. 2009), pp. 137–176.

[115] M. R. Greene and A. Oliva. "The briefest of glances: the time course of natural scene understanding". *Psychol Sci* 20.4 (Apr. 2009), pp. 464–472.

[116] M. D. Greicius, K. Supekar, V. Menon, and R. F. Dougherty. "Resting-state functional connectivity reflects structural connectivity in the default mode network." *Cerebral Cortex* 19.1 (2009), pp. 72–8.

[117] J. Grèzes, C. Frith, and R. E. Passingham. "Brain mechanisms for inferring deceit in the actions of others." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 24.24 (June 2004), pp. 5500–5.

[118] K. Grill-Spector, T. Kushnir, S. Edelman, G. Avidan, Y. Itzchak, and R. Malach. "Differential Processing of Objects under Various Viewing Conditions in the Human Lateral Occipital Complex". *Neuron* 24.1 (1999), pp. 187 –203.

[119] L. Grosenick, B. Klingenberg, B. Knutson, and J. E. Taylor. "A family of interpretable multivariate models for regression and classification of whole-brain fMRI data". *ArXiv e-prints* (Oct. 2011). arXiv: `1110.4139 [stat.AP]`.

[120] K. V. Haak, J. Winawer, B. M. Harvey, S. O. Dumoulin, B. A. Wandell, and F. W. Cornelissen. "Cortico-cortical population receptive field modeling". *Perception 40 ECVP Abstract Supplement*. 2011, p. 49.

[121] A. Harel, D. J. Kravitz, and C. I. Baker. "Deconstructing Visual Scenes in Cortex: Gradients of Object and Spatial Layout Information". *Cereb Cortex* (Apr. 2012).

[122] L. Hartwell, J. Hopfield, S. Leibler, and A. Murray. "From molecular to modular cell biology". *Nature* 402.December (1999), pp. 47–52.

[123] D. Hassabis, D. Kumaran, and E. A. Maguire. "Using imagination to understand the neural basis of episodic memory." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 27.52 (Dec. 2007), pp. 14365–74.

[124] D. Hassabis and E. A. Maguire. "Deconstructing episodic memory with construction". *Trends in Cognitive Sciences* 11.7 (2007), pp. 299–306.

[125] U. Hasson, M. Harel, I. Levy, and R. Malach. "Large-scale mirror-symmetry organization of human occipito-temporal object areas". *Neuron* 37 (2003), pp. 1027–1041.

[126] M. Hauser and J. Wood. "Evolving the capacity to understand actions, intentions, and goals." *Annual review of psychology* 61 (Jan. 2010), pp. 303–24, C1.

[127] J. V. Haxby, J. S. Guntupalli, A. C. Connolly, Y. O. Halchenko, B. R. Conroy, M. I. Gobbini, M. Hanke, and P. J. Ramadge. "A common, high-dimensional model of the representational space in human ventral temporal cortex". *Neuron* 72 (Oct. 2011), pp. 404–416.

[128] H. R. Hayama, K. L. Vilberg, and M. D. Rugg. "Overlap between the neural correlates of cued recall and source memory: evidence for a generic recollection network?" *Journal of cognitive neuroscience* 24.5 (May 2012), pp. 1127–37.

[129] G. Hein and R. T. Knight. "Superior temporal sulcus–It's my area: or is it?" *Journal of cognitive neuroscience* 20.12 (Dec. 2008), pp. 2125–36.

[130] J. Heinzle, T. Kahnt, and J. Haynes. "Topographically specific functional connectivity between visual field maps in the human brain". *NeuroImage* 56.3 (2011), pp. 1426 –1436.

[131] R. Heller, D. Stanley, D. Yekutieli, N. Rubin, and Y. Benjamini. "Cluster-based analysis of FMRI data". *Neuroimage* 33.2 (Nov. 2006), pp. 599–608.

[132] J. M. Henderson, C. L. Larson, and D. C. Zhu. "Cortical activation to indoor versus outdoor scenes: an fMRI study". *Exp Brain Res* 179.1 (May 2007), pp. 75–84.

[133] M. van den Heuvel, R. Mandl, and H. Hulshoff Pol. "Normalized cut group clustering of resting-state FMRI data". *PLoS ONE* 3.4 (2008), e2001.

[134] M. P. van den Heuvel and O. Sporns. "Network hubs in the human brain." *Trends in cognitive sciences* 17.12 (Dec. 2013), pp. 683–96.

[135] G. Hickok. "Eight problems for the mirror neuron theory of action understanding in monkeys and humans". *J Cogn Neurosci* 21.7 (July 2009), pp. 1229–1243.

[136] A. Holmes, J. Poline, and K. Friston. "Characterizing brain images with the general linear model". *Human Brain Function*. Ed. by R. Frackowiak, K. Friston, C. Frith, R. Dolan, and J. Mazziotta. Academic Press USA, 1997, pp. 59–84.

[137] N. Honnorat, H. Eavani, T. D. Satterthwaite, R. E. Gur, R. C. Gur, and C. Davatzikos. "GraSP: Geodesic Graph-based Segmentation With Shape Priors for the Functional Parcellation of the Cortex". *Neuroimage* (Nov. 2014).

[138] R. S. Huang and M. I. Sereno. "Bottom-up Retinotopic Organization Supports Top-down Mental Imagery". *Open Neuroimag J* 7 (2013), pp. 58–67.

[139] D. H. Hubel and T. N. Wiesel. "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex". *The Journal of physiology* 160.1 (1962), pp. 106–154.

[140] E. Huberle and H.-O. Karnath. "The role of temporo-parietal junction (TPJ) in global Gestalt perception." *Brain structure & function* 217.3 (July 2012), pp. 735–46.

[141] A. G. Huth, S. Nishimoto, A. T. Vu, and J. L. Gallant. "A continuous semantic space describes the representation of thousands of object and action categories across the human brain". *Neuron* 76.6 (Dec. 2012), pp. 1210–1224.

[142] G. Janzen and M. Van Turennout. "Selective neural representation of objects relevant for navigation." *Nat Neurosci* 7.6 (2004), pp. 673–7.

[143] S. Jbabdi, M. W. Woolrich, and T. E. Behrens. "Multiple-subjects connectivity-based parcellation using hierarchical Dirichlet process mixture models". *Neuroimage* 44.2 (Jan. 2009), pp. 373–384.

[144] M. Jenkinson, C. F. Beckmann, T. E. Behrens, M. W. Woolrich, and S. M. Smith. "FSL". *Neuroimage* 62.2 (Aug. 2012), pp. 782–790.

[145] H Johansen-Berg, T. E. J. Behrens, M. D. Robson, I Drobnjak, M. F. S. Rushworth, J. M. Brady, S. M. Smith, D. J. Higham, and P. M. Matthews. "Changes in connectivity profiles define functionally distinct regions in human medial frontal cortex." *Proceedings of the National Academy of Sciences of the United States of America* 101.36 (Sept. 2004), pp. 13335–40.

[146] J. D. Johnson and M. D. Rugg. "Recollection and the reinstatement of encoding-related cortical activity." *Cerebral cortex (New York, N.Y. : 1991)* 17.11 (Nov. 2007), pp. 2507–15.

[147] O. R. Joubert, G. A. Rousselet, D. Fize, and M. Fabre-Thorpe. "Processing scene context: fast categorization and object interference". *Vision Res.* 47.26 (Dec. 2007), pp. 3286–3297.

[148] J. W. Kable and A. Chatterjee. "Specificity of action representations in the lateral occipitotemporal cortex." *Journal of cognitive neuroscience* 18.9 (Sept. 2006), pp. 1498–517.

[149] I. Kahn, J. R. Andrews-Hanna, J. L. Vincent, A. Z. Snyder, and R. L. Buckner. "Distinct cortical anatomy linked to subregions of the medial temporal lobe revealed by intrinsic functional connectivity." *Journal of Neurophysiology* 100.1 (2008), pp. 129–139.

[150] D. Kaiser, L. Strnad, K. N. Seidl, S. Kastner, and M. V. Peelen. "Whole person-evoked fMRI activity patterns in human fusiform gyrus are accurately modeled by a linear combination of face- and body-evoked activity patterns." *Journal of neurophysiology* 111.1 (Jan. 2014), pp. 82–90.

[151] S. Kalénine, L. J. Buxbaum, and H. B. Coslett. "Critical brain regions for action recognition: lesion symptom mapping in left hemisphere stroke." *Brain : a journal of neurology* 133.11 (Nov. 2010), pp. 3269–80.

[152] E. Kaplan. "The M, P, and K Pathways of the Primate Visual System". *The Visual Neuroscience Encyclopedia*. Ed. by L. Chalupa and J. Werner. Cambridge, MA: MIT Press, 2003, pp. 481–494.

[153] K. N. Kay, T. Naselaris, R. J. Prenger, and J. L. Gallant. "Identifying natural images from human brain activity". *Nature* 452.7185 (Mar. 2008), pp. 352–355.

[154] D.-S. Kim and M. Kim. "Combining Functional and Diffusion Tensor MRI". *Annals of the New York Academy of Sciences* 1064.1 (2005), pp. 1–15.

[155] H. Kim. "Dissociating the roles of the default-mode, dorsal, and ventral networks in episodic memory retrieval." *NeuroImage* 50.4 (May 2010), pp. 1648–57.

[156] J. G. Kim and I. Biederman. "Where do objects become scenes?" *Cereb. Cortex* 21.8 (Aug. 2011), pp. 1738–1746.

[157] J. H. Kim, J. M. Lee, H. J. Jo, S. H. Kim, J. H. Lee, S. T. Kim, S. W. Seo, R. W. Cox, D. L. Na, S. I. Kim, and Z. S. Saad. "Defining functional SMA and pre-SMA subregions in human MFC using resting state fMRI: functional connectivity-based parcellation method". *Neuroimage* 49 (Feb. 2010), pp. 2375–2386.

[158] M. Kim, M. Ducros, K. Ugurbil, and D.-S. Kim. "Topography of high-order human object areas measured with DTI and fMRI". *Proc. Intl. Soc. Mag. Reson. Med.* 13 (2005), p. 737.

[159] D. R. King, M. D. Chastelaine, R. L. Elward, T. H. Wang, and M. D. Rugg. "Recollection-Related Increases in Functional Connectivity Predict Individual Differences in Memory Accuracy". *The Journal of Neuroscience* 35.4 (2015), pp. 1763–1772.

[160] J. C. Klein, T. E. Behrens, M. D. Robson, C. E. Mackay, D. J. Higham, and H. Johansen-Berg. "Connectivity-based parcellation of human cortex using diffusion MRI: Establishing reproducibility, validity and observer independence in BA 44/45 and SMA/pre-SMA". *NeuroImage* 34.1 (2007), pp. 204–211.

[161] T. Konkle. "The role of real-world size in object representation". PhD thesis. Cambridge, MA: Massachusetts Institute of Technology, 2011.

[162] T. Konkle and A. Oliva. "A real-world size organization of object responses in occipitotemporal cortex". *Neuron* 74.6 (June 2012), pp. 1114–1124.

[163] T. Konkle and A. Caramazza. "Tripartite organization of the ventral stream by animacy and object size." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 33.25 (2013), pp. 10235–42.

[164] A. Korattikara, Y. Chen, and M. Welling. "Austerity in MCMC Land: Cutting the Metropolis-Hastings Budget". *Proceedings of the 31st International Conference on Machine Learning*. Apr. 2014. arXiv: 1304.5299.

[165] A. Krause, K. Frank, and D. Mason. "Compartments revealed in food-web structure". *Nature* 426.November (2003).

[166] D. J. Kravitz, C. S. Peng, and C. I. Baker. "Real-world scene representations in high-level visual cortex: it's the spaces more than the places". *J. Neurosci.* 31.20 (May 2011a), pp. 7322–7333.

[167] D. J. Kravitz, K. S. Saleem, C. I. Baker, and M. Mishkin. "A new neural framework for visuospatial processing". *Nature Reviews Neuroscience* 12.4 (2011), pp. 217–230.

[168] D. J. Kravitz, K. S. Saleem, C. I. Baker, L. G. Ungerleider, and M. Mishkin. "The ventral visual pathway: an expanded neural framework for the processing of object quality." *Trends in cognitive sciences* 17.1 (Jan. 2013), pp. 26–49.

[169] B. A. Kuhl and M. M. Chun. "Successful remembering elicits event-specific activity patterns in lateral parietal cortex." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 34.23 (June 2014), pp. 8051–60.

[170] K. Kveraga and M. Bar, eds. *Scene Vision: Making Sense of What We See*. The MIT Press, 2014.

[171] K. Kveraga, A. S. Ghuman, K. S. Kassam, E. a. Aminoff, M. S. Hämäläinen, M. Chaumon, and M. Bar. "Early onset of neural synchronization in the contextual associations network." *Proceedings of the National Academy of Sciences of the United States of America* 108 (2011), pp. 3389–3394.

[172] M. H. Lee, C. D. Hacker, A. Z. Snyder, M. Corbetta, D. Zhang, E. C. Leuthardt, and J. S. Shimony. "Clustering of resting state networks". *PLoS ONE* 7.7 (2012), e40370.

[173] P Legendre and M. Fortin. "Spatial pattern and ecological analysis". *Vegetatio* 80 (1989), pp. 107–138.

[174] Y. Lerner, C. J. Honey, L. J. Silbert, and U. Hasson. "Topographic mapping of a hierarchy of temporal receptive windows using a narrated story." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 31.8 (Feb. 2011), pp. 2906–15.

[175] I. Levy, U. Hasson, G. Avidan, T. Hendler, and R. Malach. "Center-periphery organization of human object areas". *Nat. Neurosci.* 4 (May 2001), pp. 533–539.

[176] L. A. Libby, A. D. Ekstrom, J. D. Ragland, and C Ranganath. "Differential Connectivity of Perirhinal and Parahippocampal Cortices within Human Hippocampal Subregions Revealed by High-Resolution Functional Imaging". *Journal of Neuroscience* 32.19 (2012), pp. 6550–6560.

[177] A. Lingnau, B. Gesierich, and A. Caramazza. "Asymmetric fMRI adaptation reveals no evidence for mirror neurons in humans." *Proceedings of the National Academy of Sciences of the United States of America* 106.24 (June 2009), pp. 9925–30.

[178] D. Linsley and S. P. Macevoy. "Encoding-Stage Crosstalk Between Object- and Spatial Property-Based Scene Processing Pathways." *Cerebral cortex (New York, N.Y. : 1991)* (Mar. 2014).

[179] L Litman, T Awipi, and L Davachi. "Category-specificity in the human medial temporal lobe cortex." *Hippocampus* 19.3 (2009), pp. 308–319.

[180] H. Liu, W. Qin, W. Li, L. Fan, J. Wang, T. Jiang, and C. Yu. "Connectivity-based parcellation of the human frontal pole with diffusion tensor imaging." *The Journal of neuroscience* 33.16 (Apr. 2013), pp. 6782–90.

[181] L. C. Loschky and A. M. Larson. "The natural/man-made distinction is made before basic-level distinctions in scene gist processing". *Visual Cognition* 18.4 (2010), pp. 513–536.

[182] S. P. Macevoy and R Epstein. "Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex." *Current Biology* 19.11 (June 2009), pp. 943–7.

[183] S. P. MacEvoy and R. A. Epstein. "Constructing scenes from objects in human occipitotemporal cortex". *Nature Neuroscience* 14.10 (2011), pp. 1323–1329.

[184] R. Malach, J. Reppas, R. Benson, K. Kwong, H. Jiang, W. Kennedy, P. Ledden, T. Brady, B. Rosen, and R. Tootell. "Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex". *Proc. Natl. Acad. Sci. U.S.A.* 92 (Aug. 1995), pp. 8135–8139.

[185] S. A. Marchette, L. K. Vass, J. Ryan, and R. A. Epstein. "Anchoring the neural compass: coding of local spatial reference frames in human medial parietal lobe." *Nature neuroscience* (Oct. 2014).

[186] D. S. Margulies, A. M. Kelly, L. Q. Uddin, B. B. Biswal, F. X. Castellanos, and M. P. Milham. "Mapping the functional connectivity of anterior cingulate cortex". *Neuroimage* 37 (Aug. 2007), pp. 579–588.

[187] D. S. Margulies, J. L. Vincent, C. Kelly, G. Lohmann, L. Q. Uddin, B. B. Biswal, A. Villringer, F. X. Castellanos, M. P. Milham, and M. Petrides. "Precuneus shares intrinsic functional architecture in humans and monkeys". *Proceedings of the National Academy of Sciences of the United States of America* 106.47 (2009), pp. 20069–20074.

[188] N. T. Markov, M. M. Ercsey-Ravasz, A. R. Ribeiro Gomes, C. Lamy, L. Magrou, J. Vezoli, P. Misery, A. Falchier, R. Quilodran, M. A. Gariel, J. Sallet, R. Gamanut, C. Huissoud, S. Clavagnier, P. Giroud, D. Sappey-Marinier, P. Barone, C. Dehay, Z. Toroczkai, K. Knoblauch, D. C. Van Essen, and H. Kennedy. "A weighted and directed interareal connectivity matrix for macaque cerebral cortex". *Cereb. Cortex* 24.1 (Jan. 2014), pp. 17–36.

[189] D. Marr and E. Hildreth. "Theory of Edge Detection". *Proceedings of the Royal Society of London B: Biological Sciences* 207.1167 (1980), pp. 187–217.

[190] D. Marr. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. New York, NY, USA: Henry Holt and Co., Inc., 1982.

[191] R. B. Mars, S. Jbabdi, J. Sallet, J. X. O'Reilly, P. L. Croxson, E. Olivier, M. P. Noonan, C. Bergmann, A. S. Mitchell, M. G. Baxter, T. E. J. Behrens, H. Johansen-Berg, V. Tomassini, K. L. Miller, and M. F. S. Rushworth. "Diffusion-weighted imaging tractography-based parcellation of the human parietal cortex and comparison with human and macaque resting-state functional connectivity." *The Journal of neuroscience* 31.11 (Mar. 2011), pp. 4087–100.

[192] R. B. Mars, J. Sallet, U. Schüffelgen, S. Jbabdi, I. Toni, and M. F. S. Rushworth. "Connectivity-based subdivisions of the human right "temporoparietal junction area": evidence for different areas participating in different cortical networks." *Cerebral cortex (New York, N.Y. : 1991)* 22.8 (Aug. 2012), pp. 1894–903.

[193] M. Marszatek and C. Schmid. "Accurate Object Localization with Shape Masks". *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*. June 2007, pp. 1–8.

[194] S. Mesmoudi, V. Perlbarg, D. Rudrauf, A. Messe, B. Pinsard, D. Hasboun, C. Cioli, G. Marrelec, R. Toro, H. Benali, and Y. Burnod. "Resting state networks' corticotopy: the dual intertwined rings architecture." *PloS one* 8.7 (Jan. 2013), e67444.

[195] L. Micallef and P. Rodgers. "eulerAPE: drawing area-proportional 3-Venn diagrams using ellipses". *PLoS ONE* 9.7 (2014), e101717.

[196] G. A. Miller. "WordNet: A Lexical Database for English". *Commun. ACM* 38.11 (Nov. 1995), pp. 39–41.

[197] M. Misaki, Y. Kim, P. Bandettini, and N. Kriegeskorte. "Comparison of multivariate classifiers and response normalizations for pattern-information fMRI". *NeuroImage* 53.1 (2010), pp. 103 –118.

[198] M. Mishkin, L. G. Ungerleider, and K. A. Macko. "Object vision and spatial vision: two cortical pathways". *Trends in Neurosciences* 6 (Jan. 1983), pp. 414–417.

[199] A. Mishra, B. P. Rogers, L. M. Chen, and J. C. Gore. "Functional connectivity-based parcellation of amygdala using self-organized mapping: a data driven approach". *Hum Brain Mapp* 35.4 (Apr. 2014), pp. 1247–1260.

[200] M. Moayedi, T. V. Salomons, K. A. Dunlop, J. Downar, and K. D. Davis. "Connectivity-based parcellation of the human frontal polar cortex". *Brain Struct Funct* (June 2014).

[201] D. Montaldi, T. J. Spencer, N. Roberts, and A. R. Mayes. "The neural system that mediates familiarity memory." *Hippocampus* 16.5 (Jan. 2006), pp. 504–20.

[202] D. Moreno-Dominguez, A. Anwander, and T. R. Knosche. "A hierarchical method for whole-brain connectivity-based parcellation". *Hum Brain Mapp* 35.10 (Oct. 2014), pp. 5000–5025.

[203] M. Morup, K. Madsen, A Dogonowski, H. Siebner, and L. K. Hansen. "Infinite Relational Modeling of Functional Connectivity in Resting State fMRI". *Advances in Neural Information Processing Systems 23*. Ed. by J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta. 2010, pp. 1750–1758.

[204] S. L. Mullally and E. A. Maguire. "A new role for the parahippocampal cortex in representing space". *J. Neurosci.* 31 (May 2011), pp. 7441–7449.

[205] J. A. Mumford, S. Horvath, M. C. Oldham, P. Langfelder, D. H. Geschwind, and R. A. Poldrack. "Detecting network modules in fMRI time series: a weighted network analysis approach". *Neuroimage* 52.4 (Oct. 2010), pp. 1465–1476.

[206] K. P. Murphy. *Conjugate bayesian analysis of the gaussian distribution*. Tech. rep. UBC, 2007.

[207] H. Muschamp. *The Secret History of 2 Columbus Circle*. Jan. 2006.

[208] K Nakamura, R Kawashima, N Sato, a Nakamura, M Sugiura, T Kato, K Hatano, K Ito, H Fukuda, T Schormann, and K Zilles. "Functional delineation of the human occipito-temporal areas related to face and scene processing. A PET study." *Brain : a journal of neurology* 123 (Pt 9 (2000), pp. 1903–1912.

[209] S. Nasr, N. Liu, K. J. Devaney, X. Yue, R. Rajimehr, L. G. Ungerleider, and R. B. Tootell. "Scene-selective cortical regions in human and nonhuman primates". *J. Neurosci.* 31.39 (Sept. 2011), pp. 13771–13785.

[210] S. Nasr, K. J. Devaney, and R. B. H. Tootell. "Spatial encoding and underlying circuitry in scene-selective cortex." *NeuroImage* 83 (Dec. 2013), pp. 892–900.

[211] S. Nasr, C. E. Echavarria, and R. B. H. Tootell. "Thinking outside the box: rectilinear shapes selectively activate scene-selective cortex." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 34.20 (May 2014), pp. 6721–35.

[212] B. Ng and R. Abugharbieh. "Generalized Sparse Regularization with Application to fMRI Brain Decoding". *IPMI*. Ed. by G. Székely and H. K. Hahn. Vol. 6801. Lecture Notes in Computer Science. Springer, 2011, pp. 612–623.

[213] K. M. O'Craven and N Kanwisher. "Mental imagery of faces and places activates corresponding stimulus-specific brain regions." *Journal of cognitive neuroscience* 12 (2000), pp. 1013–1023.

[214] J. M. Olesen, J. Bascompte, Y. L. Dupont, and P. Jordano. "The modularity of pollination networks." *Proceedings of the National Academy of Sciences of the United States of America* 104.50 (Dec. 2007), pp. 19891–6.

[215] A. Oliva and P. G. Schyns. "Diagnostic colors mediate scene recognition". *Cogn Psychol* 41.2 (Sept. 2000), pp. 176–210.

[216] A. Oliva and A. Torralba. "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope". *International Journal of Computer Vision* 42 (2001), pp. 145–175.

[217] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. "Generic object recognition with boosting". *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 28.3 (Mar. 2006), pp. 416–431.

[218] S. Park, T. F. Brady, M. R. Greene, and A. Oliva. "Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes". *J. Neurosci.* 31.4 (Jan. 2011), pp. 1333–1340.

[219] S. Park, T. Konkle, and A. Oliva. "Parametric Coding of the Size and Clutter of Natural Scenes in the Human Brain". *Cereb. Cortex* (Jan. 2014).

[220] S. Park and M. M. Chun. "Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in panoramic scene perception." *NeuroImage* 47.4 (Oct. 2009), pp. 1747–56.

[221] G. Patterson and J. Hays. "SUN attribute database: Discovering, annotating, and recognizing scene attributes". *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. June 2012, pp. 2751–2758.

[222] T. Pedersen, S. Patwardhan, and J. Michelizzi. "WordNet::Similarity: Measuring the Relatedness of Concepts". *Demonstration Papers at HLT-NAACL 2004*. HLT-NAACL–Demonstrations '04. Boston, Massachusetts: Association for Computational Linguistics, 2004, pp. 38–41.

[223] K. A. Pelphrey, J. P. Morris, and G. McCarthy. "Grasping the intentions of others: the perceived intentionality of an action influences activity in the superior temporal sulcus during social perception." *Journal of cognitive neuroscience* 16.10 (Dec. 2004), pp. 1706–16.

[224] F. Pestilli, J. D. Yeatman, A. Rokem, K. N. Kay, and B. A. Wandell. "Evaluation and statistical inference for human connectomes". *Nat. Methods* 11.10 (Oct. 2014), pp. 1058–1063.

[225] J. Peters, I. Daum, E. Gizewski, M. Forsting, and B. Suchan. "Associations evoked during memory encoding recruit the context-network." *Hippocampus* 19.2 (Feb. 2009), pp. 141–51.

[226] M. C. Potter. "Short-term conceptual memory for pictures". *J Exp Psychol Hum Learn* 2.5 (Sept. 1976), pp. 509–522.

[227] M. J. D. Powell and Y. Yuan. "A trust region algorithm for equality constrained optimization". *Mathematical Programming* 49 (1991), pp. 189–211.

[228] J. D. Power, B. L. Schlaggar, C. N. Lessov-Schlaggar, and S. E. Petersen. "Evidence for hubs in human functional brain networks." *Neuron* 79.4 (Aug. 2013), pp. 798–813.

[229] M. E. Raichle, A. M. MacLeod, A. Z. Snyder, W. J. Powers, D. A. Gusnard, and G. L. Shulman. "A default mode of brain function". *Proceedings of the National Academy of Sciences of the United States of America* 98.2 (2001), pp. 676–682.

[230] R. Rajimehr, K. J. Devaney, N. Y. Bilenko, J. C. Young, and R. B. Tootell. "The "parahippocampal place area" responds preferentially to high spatial frequencies in humans and monkeys". *PLoS Biol.* 9.4 (Apr. 2011), e1000608.

[231] C. Ranganath and M. Ritchey. "Two cortical systems for memory-guided behaviour." *Nature reviews. Neuroscience* 13.10 (Oct. 2012), pp. 713–26.

[232] E Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. "Hierarchical organization of modularity in metabolic networks." *Science (New York, N.Y.)* 297.5586 (Aug. 2002), pp. 1551–5.

[233] E. Ravenstein. "The Laws of Migration". *Journal of the Statistical Society of London* 48.2 (1885), pp. 167–235.

[234] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson. "CNN Features off-the-shelf: an Astounding Baseline for Recognition". *CoRR* abs/1403.6382 (2014).

[235] L. W. Renninger and J. Malik. "When is scene identification just texture recognition?" *Vision Research* 44.19 (Sept. 2004), pp. 2301–2311.

[236] R. A. Rensink. "Change detection". *Annu Rev Psychol* 53 (2002), pp. 245–277.

[237] M. J. Riddoch, G. W. Humphreys, S. Edwards, T. Baker, and K. Willson. "Seeing the action: neuropsychological evidence for action-based effects on object selection." *Nature neuroscience* 6.1 (Jan. 2003), pp. 82–9.

[238] M. Riesenhuber and T. Poggio. "Hierarchical models of object recognition in cortex". *Nat. Neurosci.* 2.11 (Nov. 1999), pp. 1019–1025.

[239] A. Rives and T. Galitski. "Modular organization of cellular networks". *Proceedings of the National Academy of Sciences* 100.3 (2003), pp. 1128–1133.

[240] G. Rizzolatti and L. Craighero. "The mirror-neuron system." *Annual review of neuroscience* 27 (Jan. 2004), pp. 169–92.

[241] K. L. Roberts and G. W. Humphreys. "Action relations facilitate the identification of briefly-presented objects." *Attention, perception & psychophysics* 73.2 (Feb. 2011), pp. 597–612.

[242] K. L. Roberts and G. W. Humphreys. "Action relationships concatenate representations of separate objects in the ventral visual system." *NeuroImage* 52.4 (Oct. 2010), pp. 1541–8.

[243] B. P. Rogers, V. L. Morgan, A. T. Newton, and J. C. Gore. "Assessing functional connectivity in the human brain by fMRI". *Magn Reson Imaging* 25 (Dec. 2007), pp. 1347–1357.

[244] A. K. Roy, Z. Shehzad, D. S. Margulies, A. M. Kelly, L. Q. Uddin, K. Gotimer, B. B. Biswal, F. X. Castellanos, and M. P. Milham. "Functional connectivity of the human amygdala using resting state fMRI". *Neuroimage* 45 (Apr. 2009), pp. 614–626.

[245] M. Ruschel, T. R. Knösche, A. D. Friederici, R. Turner, S. Geyer, and A. Anwander. "Connectivity Architecture and Subdivision of the Human Inferior Parietal Cortex Revealed by Diffusion MRI." *Cerebral cortex (New York, N.Y. : 1991)* 24.9 (Apr. 2013), pp. 2436–2448.

[246] M. F. Rushworth, T. E. Behrens, and H. Johansen-Berg. "Connection patterns distinguish 3 regions of human parietal cortex". *Cereb. Cortex* 16.10 (Oct. 2006), pp. 1418–1430.

[247] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". *CoRR* abs/1409.0575 (2014).

[248] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. *LabelMe: A Database and Web-Based Tool for Image Annotation*. 2008.

[249] N. C. Rust and J. J. DiCarlo. "Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT". *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience* 30.39 (Sept. 2010). PMID: 20881116, 12978–12995.

[250] S. Ryali, T. Chen, K. Supekar, and V. Menon. "A parcellation scheme based on von Mises-Fisher distributions and Markov random fields for segmenting brain regions using resting-state fMRI." *NeuroImage* 65 (Jan. 2013), pp. 83–96.

[251] K. S. Saleem, J. L. Price, and T. Hashikawa. "Cytoarchitectonic and chemoarchitectonic subdivisions of the perirhinal and parahippocampal cortices in macaque monkeys". *J. Comp. Neurol.* 500 (Feb. 2007), pp. 973–1006.

[252] G. Salimi-Khorshidi, G. Douaud, C. F. Beckmann, M. F. Glasser, L. Griffanti, and S. M. Smith. "Automatic denoising of functional MRI data: combining independent component analysis and hierarchical fusion of classifiers". *Neuroimage* 90 (Apr. 2014), pp. 449–468.

[253] R Saxe and N Kanwisher. "People thinking about thinking people: The role of the temporo-parietal junction in theory of mind". *NeuroImage* 19.4 (Aug. 2003), pp. 1835–1842.

[254] R Saxe, D.-K. Xiao, G Kovacs, D. I. Perrett, and N Kanwisher. "A region of right posterior superior temporal sulcus responds to observed intentional actions." *Neuropsychologia* 42.11 (Jan. 2004), pp. 1435–46.

[255] R. Saxe. "Against simulation: the argument from error." *Trends in cognitive sciences* 9.4 (Apr. 2005), pp. 174–9.

[256] R. Saxe. "Uniquely human social cognition." *Current opinion in neurobiology* 16.2 (Apr. 2006), pp. 235–9.

[257] R. Sayres and K. Grill-Spector. "Relating Retinotopic and Object-Selective Responses in Human Lateral Occipital Cortex". *Journal of Neurophysiology* 100.1 (2008), pp. 249–267.

[258] P. E. Scalf, P. E. Dux, and R. Marois. "Working memory encoding delays top-down attention to visual cortex". *J Cogn Neurosci* 23 (Sept. 2011), pp. 2593–2604.

[259] A. Schindler and A. Bartels. "Parietal cortex codes for egocentric space beyond the field of view." *Current Biology* 23.2 (Jan. 2013), pp. 177–82.

[260] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. "OverFeat: Integrated Recognition, Localization and Detection using Convolutional Networks". *International Conference on Learning Representations (ICLR 2014)*. CBLS, Apr. 2014.

[261] T. V. Sewards. "Neural structures and mechanisms involved in scene recognition: a review and interpretation". *Neuropsychologia* 49.3 (Feb. 2011), pp. 277–298.

[262] K. Shelley. "Developing the American Time Use Survey Activity Classification System". *Monthly Labor Review* 128.5 (2005), pp. 3–15.

[263] J. Shi and J. Malik. "Normalized cuts and image segmentation". *Pattern Analysis and Machine Intelligence, IEEE Transactions on* 22.8 (2000), pp. 888–905.

[264] S. M. Smith, C. F. Beckmann, J. Andersson, E. J. Auerbach, J. Bijsterbosch, G. Douaud, E. Duff, D. A. Feinberg, L. Griffanti, M. P. Harms, M. Kelly, T. Laumann, K. L. Miller, S. Moeller, S. Petersen, J. Power, G. Salimi-Khorshidi, A. Z. Snyder, A. T. Vu, M. W. Woolrich, J. Xu, E. Yacoub, K. Ugurbil, D. C. Van Essen, and M. F. Glasser. "Resting-state fMRI in the Human Connectome Project". *Neuroimage* 80 (Oct. 2013), pp. 144–168.

[265] S. M. Smith, A. Hyvarinen, G. Varoquaux, K. L. Miller, and C. F. Beckmann. "Group-PCA for very large fMRI datasets". *Neuroimage* 101 (Nov. 2014), pp. 738–749.

[266] S. M. Smith, K. L. Miller, G. Salimi-Khorshidi, M. Webster, C. F. Beckmann, T. E. Nichols, J. D. Ramsey, and M. W. Woolrich. "Network modelling methods for FMRI." *NeuroImage* 54.2 (Jan. 2011), pp. 875–91.

[267] H. J. Spiers and E. a. Maguire. "The neuroscience of remote spatial memory: a tale of two cities." *Neuroscience* 149.1 (Oct. 2007), pp. 7–27.

[268] R. Spreng, R. Mar, and A. Kim. "The common neural basis of autobiographical memory, prospection, navigation, theory of mind, and the default mode: a quantitative meta-analysis". *Journal of cognitive neuroscience* 7 (2009), pp. 489–510.

[269] R. P. Spunt, A. B. Satpute, and M. D. Lieberman. "Identifying the what, why, and how of an observed action: an fMRI study of mentalizing and mechanizing during action observation." *Journal of cognitive neuroscience* 23.1 (Jan. 2011), pp. 63–74.

[270] D. E. Stansbury, T. Naselaris, and J. L. Gallant. "Natural scene statistics account for the representation of scene categories in human visual cortex". *Neuron* 79.5 (Sept. 2013), pp. 1025–1034.

[271] B. P. Staresina, K. D. Duncan, and L. Davachi. "Perirhinal and parahippocampal cortices differentially contribute to later recollection of object- and scene-related event details." *Journal of Neuroscience* 31.24 (2011), pp. 8739–8747.

[272] D. Stockburger. *Introductory Statistics: Concepts, Models, and Applications*. 1996.

[273] A. Strehl and J. Ghosh. "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions". *Journal of Machine Learning Research* 3 (2002), pp. 583–617.

[274] W. A. Suzuki. "Comparative Analysis of the cortical afferents, intrinsic projections and interconnections of the parahippocampal region in monkeys and rats". *The Cognitive Neurosciences*. Ed. by M. S. Gazzaniga. The MIT Press, 2009, pp. 659–674.

[275] K. K. Szpunar, J. C. K. Chan, and K. B. McDermott. "Contextual processing in episodic future thought." *Cerebral cortex (New York, N.Y. : 1991)* 19.7 (July 2009), pp. 1539–48.

[276] K. K. Szpunar, P. L. St Jacques, C. A. Robbins, G. S. Wig, and D. L. Schacter. "Repetition-related reductions in neural activity reveal component processes of mental simulation." *Social cognitive and affective neuroscience* 9.5 (May 2014), pp. 712–22.

[277] M. Szummer and R. W. Picard. "Indoor-Outdoor Image Classification". *IEEE International Workshop on Content-Based Access of Image and Video Databases, CAIVD*. Bombay, India, Jan. 1998, 4251.

[278] A Takashima, K. M. Petersson, F Rutters, I Tendolkar, O Jensen, M. J. Zwarts, B. L. McNaughton, and G Fernández. "Declarative memory consolidation in humans: a prospective functional magnetic resonance imaging study." *Proceedings of the National Academy of Sciences of the United States of America* 103.3 (Jan. 2006), pp. 756–61.

[279] W. K. Tam Cho and E. P. Nicley. "Geographic Proximity Versus Institutions: Evaluating Borders as Real Political Boundaries". *American Politics Research* 36.6 (May 2008), pp. 803–823.

[280] M. Thiebaut de Schotten, M. Urbanski, R. Valabregue, D. J. Bayle, and E. Volle. "Subdivision of the occipital lobes: An anatomical and functional MRI connectivity study." *Cortex* 56 (July 2014), pp. 121–37.

[281] B. Thirion, G. Flandin, P. Pinel, A. Roche, P. Ciuciu, and J. B. Poline. "Dealing with the shortcomings of spatial normalization: multi-subject parcellation of fMRI datasets". *Hum Brain Mapp* 27.8 (Aug. 2006), pp. 678–693.

[282] B. Thirion, G. Varoquaux, E. Dohmatob, and J. B. Poline. "Which fMRI clustering gives good brain parcellations?" *Front Neurosci* 8 (2014), p. 167.

[283] C. Thomas, F. Q. Ye, M. O. Irfanoglu, P. Modi, K. S. Saleem, D. A. Leopold, and C. Pierpaoli. "Anatomical accuracy of brain connections derived from diffusion MRI tractography is inherently limited". *Proc. Natl. Acad. Sci. U.S.A.* (Nov. 2014).

[284] P. Tirilly, V. Claveau, and P. Gros. "Language Modeling for Bag-of-visual Words Image Categorization". *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*. CIVR '08. Niagara Falls, Canada: ACM, 2008, pp. 249–258.

[285] V. Tomassini, S. Jbabdi, J. C. Klein, T. E. J. Behrens, C. Pozzilli, P. M. Matthews, M. F. S. Rushworth, and H. Johansen-Berg. "Diffusion-weighted imaging tractography-based parcellation of the human lateral premotor cortex identifies dorsal and ventral subregions with anatomical and functional specializations." *The Journal of Neuroscience* 27.38 (Sept. 2007), pp. 10259–69.

[286] A. Torralba, R. Fergus, and W. T. Freeman. "80 million tiny images: a large data set for nonparametric object and scene recognition". *IEEE Trans Pattern Anal Mach Intell* 30.11 (Nov. 2008), pp. 1958–1970.

[287] B. Tversky and K. Hemenway. "Categories of environmental scenes". *Cognitive Psychology* 15 (1983), pp. 121–149.

[288] L. Q. Uddin, K. Supekar, H. Amin, E. Rykhlevskaia, D. A. Nguyen, M. D. Greicius, and V. Menon. "Dissociable connectivity within human angular gyrus and intraparietal sulcus: evidence from functional and structural connectivity." *Cerebral Cortex* 20.11 (2010), pp. 2636–2646.

[289] U.S. Census Bureau. *KML - Cartographic Boundary Files - Geography - U.S. Census Bureau*. `http://www.census.gov/geo/maps-data/data/tiger-kml.html`. Accessed: 2014-04-17.

[290] U.S. Census Bureau. *Migration/Geographic Mobility - County-to-County Migration Flows: 2007-2011 ACS - People and Households - U.S. Census Bureau*. `http://www.census.gov/hhes/migration/data/acs/county\_to\_county\_mig\_2007\_to\_2011.html`. Accessed: 2014-02-06.

[291] D. C. Van Essen, S. M. Smith, D. M. Barch, T. E. Behrens, E. Yacoub, K. Ugurbil, and W.-M. H. Consortium. "The WU-Minn Human Connectome Project: an overview". *Neuroimage* 80 (Oct. 2013), pp. 62–79.

[292] S. D. Vann, J. P. Aggleton, and E. A. Maguire. "What does the retrosplenial cortex do?" *Nature Reviews Neuroscience* 10.11 (Nov. 2009), pp. 792–802.

[293] L. K. Vass and R. A. Epstein. "Abstract representations of location and facing direction in the human brain." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 33.14 (Apr. 2013), pp. 6133–42.

[294] K. L. Vilberg and M. D. Rugg. "Functional significance of retrieval-related activity in lateral parietal cortex: Evidence from fMRI and ERPs." *Human brain mapping* 30.5 (May 2009), pp. 1490–501.

[295] K. L. Vilberg and M. D. Rugg. "Memory retrieval and the parietal cortex: a review of evidence from a dual-process perspective." *Neuropsychologia* 46.7 (Jan. 2008), pp. 1787–99.

[296] K. L. Vilberg and M. D. Rugg. "The neural correlates of recollection: transient versus sustained FMRI effects." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 32.45 (Nov. 2012), pp. 15679–87.

[297] J. Vogel and B. Schiele. "Semantic Modeling of Natural Scenes for Content-Based Image Retrieval". *Int. J. Comput. Vision* 72.2 (Apr. 2007), pp. 133–157.

[298] E Vul, C Harris, P Winkielman, and H Pashler. "Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition". *Perspectives on Psychological Science* 4.3 (2009), pp. 274–290.

[299] A. R. Wade, A. A. Brewer, J. W. Rieger, and B. A. Wandell. "Functional measurements of human ventral occipital cortex: retinotopy and colour". *Philos. Trans. R. Soc. Lond., B, Biol. Sci.* 357 (Aug. 2002), pp. 963–973.

[300] S. Wakana, H. Jiang, L. M. Nagae-Poetscher, P. C. M. van Zijl, and S. Mori. "Fiber tract-based atlas of human white matter anatomy." *Radiology* 230.1 (Jan. 2004), pp. 77–87.

[301] D. B. Walther, B. Chai, E. Caddigan, D. M. Beck, and L. Fei-Fei. "Simple line drawings suffice for functional MRI decoding of natural scene categories". *Proc. Nat. Acad. of Sci (PNAS)*. 2011.

[302] D. B. Walther, E. Caddigan, L. Fei-Fei, and D. M. Beck. "Natural scene categories revealed in distributed patterns of activity in the human brain." *The Journal of neuroscience : the official journal of the Society for Neuroscience* 29.34 (2009), pp. 10573–10581.

[303] B. Wandell, S. O. Dumoulin, and A. A. Brewer. "Visual cortex in humans". *Encyclopedia of neuroscience* 10 (2008), pp. 251–257.

[304] L. Wang, R. E. B. Mruczek, M. J. Arcaro, and S. Kastner. "Probabilistic Maps of Visual Topography in Human Cortex". *Cerebral Cortex* (2014).

[305] S. C. Want and P. L. Harris. "How do children ape? Applying concepts from the study of non-human primates to the developmental study of 'imitation' in children". *Developmental Science* 5.1 (2002), pp. 1–13.

[306] E. J. Ward, S. P. MacEvoy, and R. A. Epstein. "Eye-centered encoding of visual space in scene-selective regions." *Journal of vision* 10 (2010), p. 6.

[307] J. H. Ward. "Hierarchical Grouping to Optimize an Objective Function". *Journal of the American Statistical Association* 58.301 (1963), pp. 236–244.

[308] K. S. Weiner and K. Grill-Spector. "Not one extrastriate body area: using anatomical landmarks, hMT+, and visual field maps to parcellate limb-selective activations in human lateral occipitotemporal cortex". *Neuroimage* 56 (June 2011), pp. 2183–2199.

[309] T. Wheatley, S. C. Milleville, and A. Martin. "Understanding animate agents: distinct roles for the social network and mirror system." *Psychological science* 18.6 (June 2007), pp. 469–74.

[310] G. S. Wig, T. O. Laumann, and S. E. Petersen. "An approach for parcellating human cortical areas using resting-state correlations". *Neuroimage* 93 Pt 2 (June 2014), pp. 276–291.

[311] A. J. Wiggett and P. E. Downing. "Representation of action in occipito-temporal cortex." *Journal of cognitive neuroscience* 23.7 (July 2011), pp. 1765–80.

[312] J. L. Wiggins, S. J. Peltier, S. Ashinoff, S. J. Weng, M. Carrasco, R. C. Welsh, C. Lord, and C. S. Monk. "Using a self-organizing map algorithm to detect age-related changes in functional connectivity during rest in autism spectrum disorders". *Brain Res.* 1380 (Mar. 2011), pp. 187–197.

[313] H. Wolf. "Intranational home bias in trade". *Review of economics and statistics* 82.November (2000), pp. 555–563.

[314] K. J. Worsley, A. C. Evans, S. Marrett, and P. Neelin. "A three-dimensional statistical analysis for CBF activation studies in human brain". *J. Cereb. Blood Flow Metab.* 12.6 (Nov. 1992), pp. 900–918.

[315] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva. "SUN Database: Exploring a Large Collection of Scene Categories". *International Journal of Computer Vision* (2014).

[316] J. Xu, P. J. Gannon, K. Emmorey, J. F. Smith, and A. R. Braun. "Symbolic gestures and spoken language are processed by a common neural system." *Proceedings of the National Academy of Sciences of the United States of America* 106.49 (Dec. 2009), pp. 20664–9.

[317] R. Xu, Z. Zhen, and J. Liu. "Mapping informative clusters in a hierarchical [corrected] framework of FMRI multivariate analysis." *PloS one* 5.11 (Jan. 2010), e15065.

[318] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei. "Action Recognition by Learning Bases of Action Attributes and Parts". *International Conference on Computer Vision (ICCV)*. Barcelona, Spain, Nov. 2011.

[319] F. A. Yates. *The Art of Memory*. Chicago, IL: University of Chicago Press, 1966.

[320] B. T. T. Yeo, F. M. Krienen, S. B. Eickhoff, S. N. Yaakub, P. T. Fox, R. L. Buckner, C. L. Asplund, and M. W. L. Chee. "Functional Specialization and Flexibility in Human Association Cortex". *Cerebral Cortex* (Sept. 2014), bhu217.

[321] B. T. T. Yeo, F. M. Krienen, J. Sepulcre, M. R. Sabuncu, D. Lashkari, M. Hollinshead, J. L. Roffman, J. W. Smoller, L. Zöllei, J. R. Polimeni, B. Fischl, H. Liu, and R. L. Buckner. "The organization of the human cerebral cortex estimated by intrinsic functional connectivity." *Journal of neurophysiology* 106.3 (Sept. 2011), pp. 1125–65.

[322] E. Y. Yoon, G. W. Humphreys, S. Kumar, and P. Rotshtein. "The neural selection and integration of actions and objects: an fMRI study." en. *Journal of cognitive neuroscience* 24.11 (Nov. 2012), pp. 2268–79.

[323] D. Zhang, A. Z. Snyder, M. D. Fox, M. W. Sansbury, J. S. Shimony, and M. E. Raichle. "Intrinsic functional relations between human cerebral cortex and thalamus". *J. Neurophysiol.* 100 (Oct. 2008), pp. 1740–1748.

[324] D. Zoccolan, D. D. Cox, and J. J. DiCarlo. "Multiple object response normalization in monkey inferotemporal cortex". *J. Neurosci.* 25.36 (Sept. 2005), pp. 8150–8164.