What makes different people's representations alike: neural similarity-space solves the problem of across-subject fMRI decoding

Rajeev D. S. Raizada[1] and Andrew C. Connolly[2]
[1]Department of Human Development, Cornell University, Ithaca NY 14853.
[2]Dept. of Psychological & Brain Sciences, Dartmouth College, Hanover NH 03755.

Abstract

A central goal in neuroscience is to interpret neural activation, and moreover to do so in a way that captures universal principles by generalising across individuals. Recent research in multivoxel pattern-based fMRI analysis has led to considerable success at decoding within individual subjects. However, the goal of being able to decode across subjects is still challenging: it has remained unclear what the common organising principles of neural representation might be which hold across individuals. Here we present a novel and highly accurate solution to this problem, which decodes across subjects between eight different stimulus conditions. The key to finding this solution was questioning the seemingly obvious idea that neural decoding should work directly upon neural activation patterns. On the contrary, we show that to decode across subjects it is necessary to abstract away from subject-specific patterns of neural activity, and instead to operate on the similarity-relations between those patterns: our new approach performs decoding purely within similarity-space. These results demonstrate a hitherto unknown common organising principle of neural representation, and also reveal a striking convergence between our empirical findings in fMRI and discussions in the philosophy of mind addressing the problem of conceptual similarity across neural diversity. Our results also have possible practical applications: previous approaches for interpreting neural activation in multiple individuals have needed conditions which specifically activate different parts of the brain, e.g. motor-imagery versus spatial-navigation. However, very few conditions meet that requirement, so such approaches allow discrimination between only two states, such as "Yes" and "No". In contrast, our new approach accurately distinguishes between eight different states, therefore opening up new possibilities for decoding fine-grained representations from multiple people's brains.

1 Introduction

In Cognitive Neuroscience, the goal is in general not to study the peculiarities of particular individuals' brains, but instead to find regularities which hold across individuals at the population level. An obstacle to that goal is the fact that different peoples' brains do not directly match up. They share the same gross anatomy, and they also share coarse-grained functional distinctions, e.g. between animate and inanimate object categories (Warrington & Shallice, 1984; Caramazza & Shelton, 1998; Martin, 2007) such as faces and houses (McCarthy et al., 1997; Kanwisher et al., 1997; Epstein & Kanwisher, 1998). However, at a finer-grain, there are diverse individual differences: for example, the size of V1 in different people can vary by a factor of more than two, and this size-variability has perceptual consequences (Duncan & Boynton, 2003). At the level of specific neural representations, pattern-recognition algorithms have been used to find the multivoxel neural "fingerprints" elicited

by given stimulus conditions (Haxby et al., 2001; Haynes & Rees, 2006; Norman et al., 2006; Pereira et al., 2009; Raizada & Kriegeskorte, 2010). However, just as the literal fingerprints on people's hands are idiosyncratic to individuals, the "neural fingerprints" of representations in their brains may also be subject-unique. Indeed, this has found to be the case. For example, Shinkareva, Mitchell and colleagues performed both within- and across-subject decoding, and found that "a critical diagnostic portion of the neural representation of the categories and exemplars is still idiosyncratic to individual participants" (Shinkareva et al., 2008).

Whatever commonality there might be between different people's neural representations, it must somehow abstract away from their subject-specific fine-grained neural patterns. Can we find a level of representation which is shared across individuals, and which is finer-grained than animate-vs.-inanimate, but which, unlike subject-specific neural fingerprints, succeeds in capturing across-subject commonalities? A shared level of representation that satisfies these conditions would be able to translate between different people's neural representational schemes. In other words, it would be able to perform across-subject neural decoding.

A potentially promising candidate level of representation is similarity-space, which is the set of pairwise relations between items defined by a similarity measure, and which has long served as a powerful tool in psychology for investigating cognitive processing (Shepard, 1962; Tversky, 1977; Medin et al., 1993; Edelman, 1998). In the neural domain, it has been used for visualising and comparing overall representational structure (Edelman et al., 1998; Hanson et al., 2004; O'Toole et al., 2007; Kriegeskorte et al., 2008; Connolly et al., 2011; Shinkareva et al., 2011). However, in seeking to decide whether or not different people's representational schemes are the same, we need to be able to do more than visualise the broad overall match between them. We need a translation dictionary enabling us to decode between the different sets of representations, i.e. to perform across-subject neural decoding. However, until now, no method of neural decoding using similarity-space has been available.

In all previous work on neural decoding, the inputs to the decoding algorithms have not been similarity-values, but instead have been neural activation values themselves (e.g. Haxby et al., 2001; Haynes & Rees, 2006; Norman et al., 2006; Pereira et al., 2009). However, at a fine-grain these neural activation patterns suffer from the subject-specific idiosyncrasies described above. Across-subject decoding of fine-grained neural representations has therefore remained a challenge.

It might seem almost too obvious to be worth stating that neural decoding should take as its input neural activation patterns. Here we argue that the seemingly tautological nature of that statement is deceptive. On the contrary, we argue here that to be effective across subjects the decoding should not take neural activation patterns as its input. Instead, its input must be the *similarity relations between those patterns*, rather than the neural patterns themselves. By operating on the similarity relations, the decoding can abstract away from the idiosyncratic and subject-specific nature of the neural activation. To support this claim, we present for the first time a method to perform neural decoding purely within similarity-space. We then demonstrate that this new method achieves highly accurate across-subject decoding.

2 Methods, data, and analysis approach

For the analyses in this paper, we used the classic Haxby et al. (2001) dataset of object-elicited activation in ventral temporal (VT) cortex, kindly made available online by Haxby and the developers of PyMVPA (http://dev.pymvpa.org/datadb/haxby2001.html). The VT-cortex masks in that study are included in the online dataset, and were manually traced from anatomical scans to consist of

the lingual, parahippocampal, fusiform, and inferior temporal gyri. The neural similarity-space for each subject was calculated simply as the spatial correlation between the various stimulus-conditions' activation patterns across VT-cortex. The stimulus categories spanned the animate-vs.-inanimate distinction, but they also included a lower level of multiple animate and inanimate subcategories. The animate stimuli were subdivided into cats and faces, and the inanimate stimuli were subdivided into bottles, chairs, houses, scissors, scrambled pictures, and shoes.

We first calculated the VT-cortex neural similarities between these eight stimulus-conditions for each of the six subjects. The similarity measure was the simplest possible: spatial correlation. Before this pattern-correlation step, the voxel time-courses were first normalised in intensity by being z-scored, i.e. by having their mean-values subtracted and being divided by their standard-deviations. In order to avoid normalising-out potentially informative stimulus-evoked signals, these means and standard deviations were calculated from the rest-condition TRs only. Such normalisation is standard for pattern-based fMRI analyses (Pereira et al., 2009) and indeed for machine-learning studies in general (Han & Kamber, 2006). It is particularly useful for correlation-based analyses, which would otherwise tend to be corrupted by outlier intensity values.

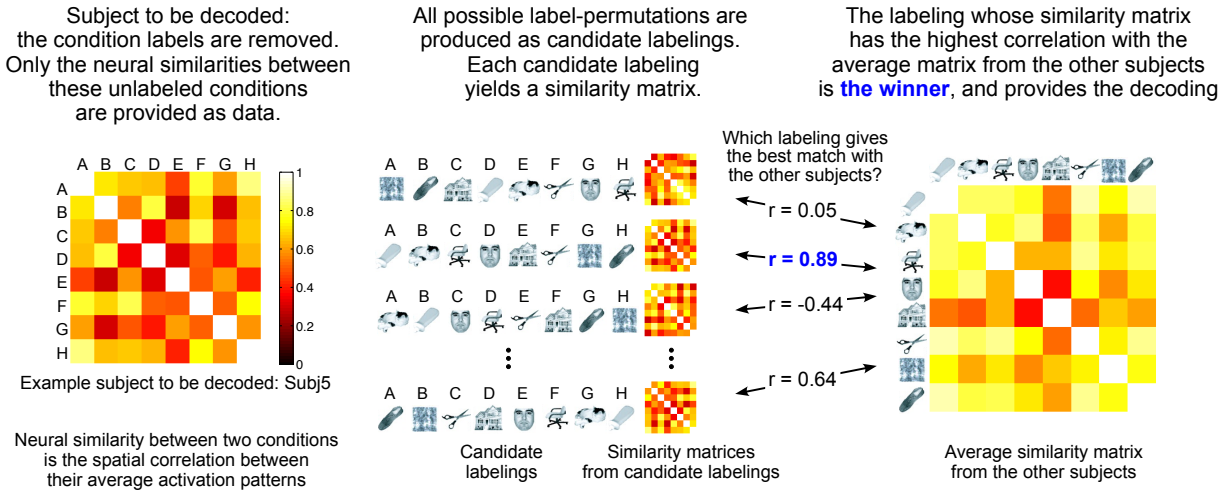

Figure 1: Our novel method of across-subject neural decoding: "Decoding by Matching Of Similarity-Spaces," or DEMOSS. The data entered into the model for each subject consists only of the values in their 8x8 similarity matrix, comprising 8x(8-1)/2 = 28 unique numbers. Only one permutation-matching computation is performed per subject, so there are no multiple comparisons. The illustrated similarity matrices are the actual data for the example subject shown.

In Figure 1 we present and explain our novel method, simple but highly effective, for performing neural decoding purely within similarity-space. Using our new method, which we call "Decoding by Matching Of Similarity-Spaces" or DEMOSS, we show here for the first time that similarity-space is indeed able to serve as a translation dictionary between different people's neural representations. We also show that people's shared representational structure goes beyond the animate-vs.-inanimate distinction, and extends to the finer-grained level of multiple animate and inanimate subcategories. Moreover, we demonstrate below that by operating purely within similarity-space, this across-subject decoding remains accurate even in the presence of high degree of neural diversity.

Our analysis code was written in Matlab, and the preprocessing and data-extraction were carried out in Python using scripts from PyMVPA (Hanke et al., 2009). To facilitate easy replication and verification of our results, all of the analysis-code is provided in the Supplementary Information.

3 Results

3.1 Visualisation of overall similarity structure leaves it unclear whether decoding can be achieved

As was remarked in the introduction, neural similarity-space has previously been used for visualising and comparing overall representational structure by combining it with multidimensional scaling (or MDS, Shepard, 1962). Examples of such studies include Edelman et al. (1998); Hanson et al. (2004); O'Toole et al. (2007); Kriegeskorte et al. (2008); Connolly et al. (2011) and Shinkareva et al. (2011). However, until now, no method for using similarity-space to perform neural decoding has been available. Given the existence of these visualisation studies, it is reasonable to ask whether visualisation alone is sufficient to judge whether neural decoding could be performed.

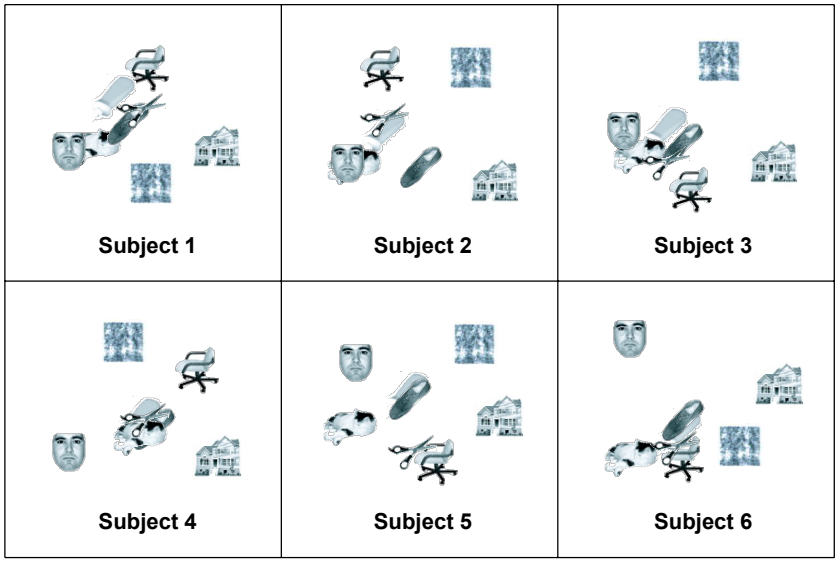

Figure 2: The neural similarity-spaces of each of the six subjects in the Haxby et al. (2001) data, visualised in 2-D using multidimensional scaling (Shepard, 1962). Although some broad commonalities are readily apparent, there are also major inter-subject differences. Such a visualisation therefore leaves it unclear whether similarity-space can serve as a translation dictionary between different people's neural representations. In order to directly test that, we need to see whether similarity-space enables across-subject decoding.

In Figure 2, we show 2-D multidimensional scaling (MDS) projections of each individual subject's neural similarity-space in the Haxby et al. (2001) dataset. Some broad commonalities are readily apparent: houses and scrambled-pictures always stand apart from the other stimuli, and bottles, shoes and scissors typically cluster together. But it is unclear whether these commonalities are sufficient to allow across-subject decoding. Categories that cluster together in some subjects are quite dissimilar in others (e.g. faces and cats are much more similar in Subjects 1, 2 and 3 than they are in Subjects 4, 5 and 6). Given this, one might expect that a similarity-based decoding would be able to distinguish between faces and cats in the first three subjects, but would confuse the two stimulus-categories in the remaining three. As we show below, this is not in fact the case: our similarity-based decoding did not confuse those two categories. The amount of variability across different people's category-clusterings means that the visualisation, on its own, does not tell us whether an attempt to decode the stimuli across subjects would succeed or fail.

3.2 Accurate across-subject decoding of fine-grained object categories in VT cortex

We used our new DEMOSS method, shown in Figure 1, to perform across-subject decoding of the Haxby data. With 8 categories per subject and 6 subjects, there were 48 decodings to perform in all. The method scored 91.7% correct (44 out of 48 categories correct). Software to replicate these analyses is provided in the Supplementary Information.

If the animate-vs.-inanimate distinction were the level at which different people's neural representational schemes are the same, then it would be predicted that across-subject decoding should succeed at that level, but fail at lower levels in the hierarchy. In contrast, if different people's neural representational schemes are the same not only at the animate-vs.-inanimate level but also at lower subdivisions of the hierarchy, then across-subject decoding would be predicted to succeed even at making distinctions between finer-grained subcategories, e.g. at distinguishing between different animate categories (faces vs. cats) and between different inanimate categories (e.g. bottles vs. shoes).

The latter prediction held true: the decoding was highly accurate at distinguishing between fine-grained animate and inanimate subcategories. Within the animate subdivision, decoding was 100% correct: a face was never confused with a cat. The more difficult decoding task was within the inanimate subdivision, in which some errors were made: five of the six subjects had all six of their inanimate categories perfectly decoded, and the remaining subject had two pairs of confusions: bottle-scissors and shoe-chair. However, decoding between inanimate subcategories was far above chance (32 out of 36 correct decodings, i.e. 88.9% correct). Chance-level performance is to get 1/8th of the decodings correct, i.e. 12.5%.

3.3 Decoding remains accurate, even across widespread neural diversity

The success of this across-subject decoding shows that neural similarity-space captures a representational scheme which is shared across individuals, even at the fine-grained level of multiple animate and inanimate subcategories. However, as was noted in the introduction, one of the main sources of difficulty for across-subject neural decoding is the fact that different people's brains do not directly match up. In the analysis above, that difficulty was not felt with its full force, because all of the neural signals were drawn from the same brain area: VT cortex.

As Figs. 3a and c show, the VT-cortex masks drawn by Haxby et al. on individual subjects' anatomical scans do not completely overlap when they are aligned to a common space. But they do mostly overlap. A stronger test would therefore be to use anatomically dispersed and highly variable sets of voxels in different individuals.

In order to carry that out, we devised a simple feature-selection scheme to find informative voxels within each individual subject. For each voxel, we calculated two measures to be used for selection. Comparing all of the visual-object stimuli together against the rest-blocks, we determined the t-statistic for the degree to which each voxel was active. Then, considering only the object-stimuli blocks, we calculated the F-statistic of the ratio of between-class variance to within-class variance. We then selected the voxels within each subject which scored not only in the top 5% of t-values but also in the top 5% of F-values, i.e. the voxels which were active and which differentiated between the various object stimuli. As before, the neural similarity-space for each subject was calculated simply as the spatial correlation between the various stimulus-conditions' activation patterns, but this time the patterns were the activations across the selected voxels, rather than across the VT-cortex region. (Matlab scripts used to perform this feature-selection, and to carry out the similarity-analyses on the selected voxels, are provided in the Supplementary Information). In order to compare the locations

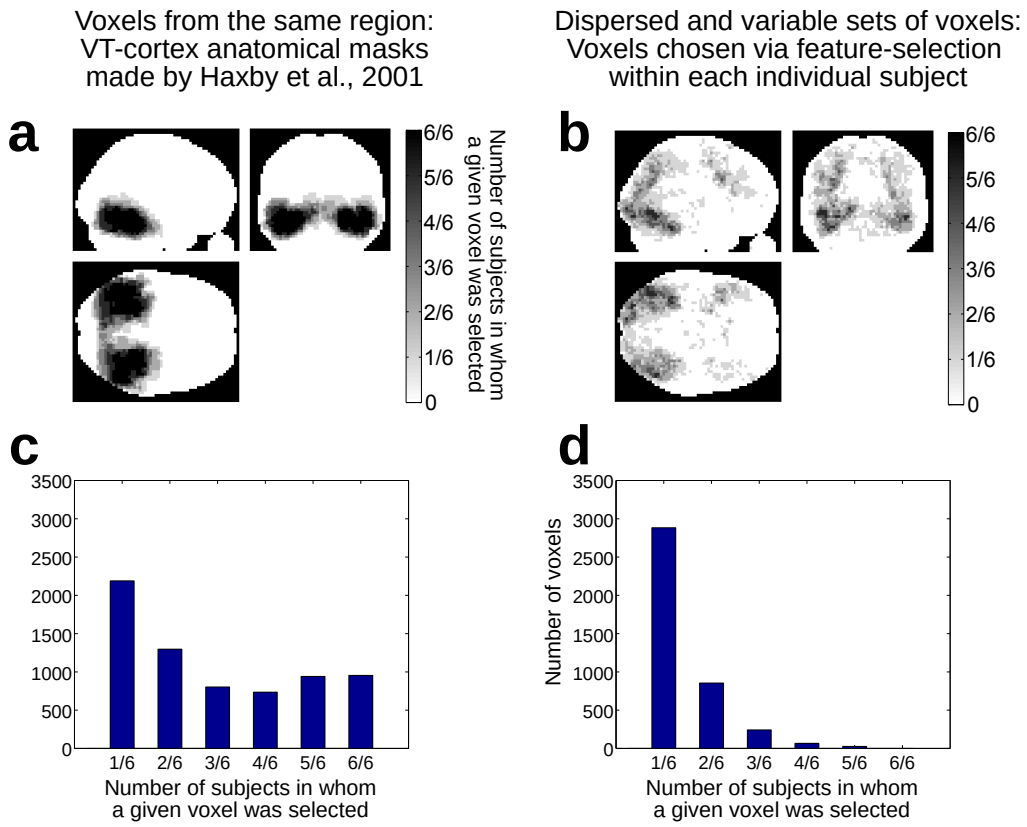

Figure 3: Maximum intensity projections and histograms, showing differing degrees of across-subject neural diversity. Panels **(a)** and **(c)**: when the voxels used for across-subject decoding were specified by the VT-cortex masks included in the Haxby dataset, the performance was 91.7% correct. However, there was relatively little neural diversity across different people's VT masks. Panels **(b)** and **(d)**: in a separate analysis, we used a simple feature-selection scheme to find different sets of informative voxels within each individual subject. The selected voxels were anatomically dispersed and highly variable across different individuals. Nonetheless, using these diverse sets of voxels, the across-subject decoding still achieved 87.5% correct. Chance-level performance is 12.5%.

of the selected voxels across different subjects, the brain volumes were all spatially normalised to the standard MNI152 template at 3x3x3mm resolution using SPM8, before feature-selection or similarity-analysis was carried out.

These feature-selected voxels showed a very high degree of diversity across subjects: the number of voxels selected within each subject ranged from 473 to 1346. In other words, the dimensionalities of people's neural-activation-spaces varied widely across different individuals. It is unclear even how to compare a 473-dimensional space with a 1346-dimensional space, let alone to try to decode between them. However, by calculating the spatial correlations between the stimulus-elicited activation patterns *within* each activation-space, the different subjects' activation-spaces, with their widely varying dimensionalities, all become transformed into eight-dimensional similarity-spaces defined by the eight stimulus categories. These similarity-spaces *can* be compared, and using our novel DEMOSS method presented in Figure 1, we can perform across-subject decoding between them.

Different people's selected voxels varied not only in their number, but also in their locations across the subjects' brains. As Figs. 3b and d show, the selected voxels were dispersed broadly throughout the brain, and their anatomical locations were highly variable across subjects. As would be expected, the greatest concentration of selected voxels was found in VT cortex; however, informative voxels were

found in many other regions, including parietal and frontal cortex. The selected voxel-locations in those areas were often shared by just one or two subjects, as can be seen from the light-gray regions in the maximum intensity projection in Fig. 3b. The histogram in Fig. 3d confirms that the majority of selected voxel-locations were specific to individual subjects, and that very few voxel-locations were shared by multiple individuals. As it happened, there was not a single voxel that was selected in all six subjects, not even in the heart of VT-cortex.

To what degree neural decoding be able to succeed, in the face of this very marked neural diversity? Using the similarity-spaces derived from these disparate sets of feature-selected voxels, the performance of across-subject decoding was 87.5%, only slightly lower than the 91.7% obtained when the voxels were specified by VT-cortex masks. Four of the six subjects were decoded perfectly. In one subject, there were four confusions: cat, chair, face and scissors. In the remaining subject, bottles and scissors were confused. As before, chance-level performance is 12.5%. Thus, similarity-space was able to capture the representational scheme shared across individuals, even though the neural populations used to match people's representations were extremely diverse.

3.4 Statistical significance tests for a permutation distribution

Given that there are eight stimulus categories, the number of possible category-labelings for each subject's decoding is equal to the factorial of eight, i.e. 40320. Only one of those more than forty thousand labelings gets all 8-out-of-8 labelings correct. It is therefore noteworthy that this solitary perfect 8-out-of-8 labeling often emerged as the decoding-output, in virtue of its having a higher level of across-subjects match in similarity-space than any of the other forty thousand labelings. That 8-out-of-8 perfect decoding was achieved for five of the six subjects when all of the neural information was drawn from cortical area VT, and for four of the six subjects when the neural information was drawn from anatomically dispersed and highly variable feature-selected sets of voxels. For the subjects whose decoding was not perfect, it was still significantly above chance. As the Supplementary Information and its accompanying Matlab code show, the chance-level is to get 1 out 8 correct, and in order to meet $p<0.05$ significance, 3 or more out of the 8 categories need to be decoded correctly. In our results, not only when decoding from VT-cortex but also when using the feature-selected voxels, no subject's decoding achieved fewer than 4 out of 8 correct. This suggests that similarity-space, even with its very simple construction and greatly reduced dimensionality, does indeed succeed in capturing a crucial aspect of what different people's representations have in common.

4 Discussion

The results above demonstrate for the first time, using real neural data, that similarity-space can provide a translation dictionary between different people's representational schemes. This across-subject decoding remains highly accurate even when the neural similarities are derived from widely diverse sets of voxels across different subjects.

4.1 Relation to previous fMRI studies

Previous across-subject decoding attempts have all involved feeding thousands of voxels and hundreds of time-points into classifier algorithms, so it has remained unclear which aspects of the complex neural signal have been the ones which different people shared. In contrast, our new across-subject

decoding method takes in an extremely reduced dataset as input: only the similarity-space of people's neural category representations. The success of its decoding is therefore driven entirely by across-subject commonalities in that abstract category-similarity structure. The structure of similarity-space serves as a translation dictionary between different people's representational schemes.

Moving from the nature of the input to questions of performance, our study decodes fine-grained category distinctions across subjects with high accuracy, whereas previous studies have either decoded coarse-grained changes in brain state, or with lower accuracy, or both. An example of decoding a large-scale change in brain state is distinguishing between the performance of different behavioural tasks, such as reading a sentence vs. looking at a picture (Wang et al., 2003), face-matching vs. location-matching (Mourao-Miranda et al., 2005) or between several different cognitive tasks (Poldrack et al., 2009). A different example of a coarse-grained distinction is between being rewarded with money vs. viewing an attractive face (Clithero et al., 2011). Shinkareva, Mitchell and colleagues (Shinkareva et al., 2008) went further, and were able to decode not only which general category of object a person was looking at (tool vs. dwelling), but also which of the five specific exemplars within each category they were looking at. However, their across-subject decoding, which operated directly on neural activation, worked for only eight of their twelve subjects, and achieved a considerably lower level of performance than our approach, which operates instead on neural similarities. Another interesting and important line of work in this area, still ongoing, is that of Guntupalli, Haxby and colleagues (Guntupalli & Haxby, 2009), who have proposed a high-dimensional mapping, called "hyper-alignment" of one person's voxel space onto another's.

Following its initial publication (Haxby et al., 2001), there have been a number of subsequent papers containing analyses of the Haxby 2001 data. These prior studies have performed within-subject voxel-wise sensitivity analyses (Hanson et al., 2004), have compared the performance of classifiers applied to the image stimuli themselves against that of classifiers applied to the neural data (O'Toole et al., 2005, 2007), have investigated ICA (Daubechies et al., 2009), and have explored the use of classifier ensembles (Kuncheva et al., 2010). Two of these Haxby-data studies (Hanson et al., 2004; O'Toole et al., 2007) used similarity-structure analyses to explore representational organisation. However, neither those studies, nor any previous investigations of neural similarity-space, have used similarity-space to perform across-subject decoding. This across-subject decoding is a key contribution of our new approach, along with what it tells us about people's shared hierarchy of object representations, and its ability to find the same representations across highly diverse neural populations. Our new approach also intersects with some longstanding conceptual debates in Cognitive Science, as we discuss in the following section.

4.2 Relation to longstanding conceptual debates in Cognitive Science

A striking aspect of the solution to across-subject decoding presented above is its parallelism to a proposal made more than 20 years ago by the neuro-philosopher Paul Churchland (Churchland, 1986). Churchland proposed, on purely theoretical grounds, that matched structure in people's neural similarity-spaces could explain how different brains can form the same mental representations. He referred to this as "the problem of conceptual similarity across neural diversity" (Churchland, 1998).

However, that proposal has faced opposition, most notably from Fodor & Lepore (1992, 1999), who argued that similarity-space theories cannot explain how different people could possess the same concept. Partly in response to such objections, researchers have presented computer simulations as evidence that similarity space could indeed, in principle, provide a solution (Laakso & Cottrell, 2000; Goldstone & Rogosky, 2002). However, no simulation can address the question of how *real* brains actually *do* solve the problem. Our results above, using real neural data, do precisely that.

The question of whether or not the philosophy of mind is a relevant part of Cognitive Science is far beyond the scope of this paper. We merely remark that when completely different lines of enquiry, originating respectively from conceptual and empirical concerns, end up converging on the same solution, it may be an indication that they are both being guided by something real. How robust and generalisable our proposed solution to across-subject decoding will turn out to be is, of course, an empirical question. To address that in future studies will require applying the analysis method to a broad variety of data sets, drawn from a wide range of task domains.

4.3 Why bother? Isn't it enough simply to decode at the single-subject level?

As with many questions in Cognitive Neuroscience, one answer to the question "why bother?" is simply "because it is so intrinsically interesting." That, we believe, is a valid reason in itself. However, there also exist some important practical reasons, whose examination serves to highlight some current limitations in the field of neural decoding, and how those limitations might be addressed.

What would it mean to truly understand people's neural representations in a robust and generalisable way? We would be able to record activation from people's brains, and then attribute meaning to that activation — not just for particular individuals whose personal neural codes we already had access to, but for all of those individuals, even people whom we had not tested before.

Perhaps the closest that the field has come to this ideal is the striking, but very limited, result of Monti et al. (2010), who attempted to extract meaning from the brains of several patients who were incapable of making behavioural responses, due to complete "locked in" paralysis after severe brain injury. To that end, the researchers needed to be able to pre-specify patterns of brain activation which, if present in any patient, could be unambiguously recognised and interpreted. Although it is impressive that this could be done at all, only two such types of activation could be found, thereby yielding only two possible meanings that the patients were able to express, namely "Yes" and "No". The two brain states were: motor cortex activation, elicited by mental motor imagery (imagining playing tennis), and parahippocampal activation, elicited by imagery of spatial navigation (imagining walking around a familiar building or town).

Why were only two distinguishable states achieved? The reason is that very few types of brain activation can be reliably interpreted as meaning just one thing. Almost all brain areas are highly multi-functional and multi-representational. For example, activation in parietal cortex might indicate the occurrence of many possible processes: attention, spatial planning, numerical processing, phonological awareness, and more. Activation in ventral temporal (VT) cortex can be triggered by the perception or imagery of thousands (perhaps millions) of different visual objects, all of which the brain somehow is able to represent within the same region of cortex. Even regions often thought of as having single functions have recently been shown to participate in multiple representations, such as the "Fusiform Face Area" (FFA), which responds strongly to faces but also to many other object categories (Hanson et al., 2004), and the amygdala, which was long believed to respond only to fear, but in fact responds not only to negatively valenced stimuli put to positive valenced ones as well (Morrison & Salzman, 2010). If the researchers had asked their locked-in patients to produce not two but four distinguishable neural states by adding face-imagery and fear-imagery to the list, then it is likely that the elicited patterns of neural activation would have become much harder to tell apart.

At this point, the following objection might come to mind: there exist many studies which have successfully performed multi-category neural decoding (e.g., Hanson & Schmidt, 2011), with several different object representations being distinguished from each other within the same brain area. Why could such an approach not be used to decode multiple different neural responses from the brains of

the several locked-in patients?

The answer to that question reveals a key weakness in current neural decoding approaches: the decoding does not generalise across subjects. In other words, the stimulus-evoked neural responses inside a given person's head are of very little use for decoding the neural responses from inside somebody else's head. As the quote from Shinkareva et al. (2008) above pointed out, the fine-scale "neural fingerprints" elicited by particular stimuli tend to be idiosyncratic and subject-specific.

The Monti et al. (2010) study, with its locked-in patients, highlights in a particularly clear-cut manner why constructing subject-specific neural-fingerprint lookup tables for decoding is not a viable option. For us to know which neural fingerprint corresponded to which mental-imagery condition, the subjects would need to be communicative and responsive, as they would need to be able to put themselves into each imagery condition when we requested them to do so. However, the patients are neither communicative nor responsive — they are locked-in! Thus, it is impossible to build up a codebook of subject-specific neural fingerprints. The only option is to use mental imagery tasks whose neural correlates can be unambiguously known in advance for all of the patients. That, in turn, requires using the very small subset of brain areas whose anatomical locations are constant across individuals, and whose activation unambiguously signals the occurrence of a specific mental imagery process. Very few brain areas meet that requirement, with the consequence that only two could be used: motor cortex and parahippocampal cortex, corresponding to just "Yes" and "No".

We hope that the above example clarifies why the ability to perform single-subject but non-generalisable neural decoding does not actually amount to achieving an understanding of the principles of neural representation. Instead, it consists of building a series of subject-specific neural-fingerprint lookup tables, one per person. That is a non-trivial task, and the fact that fMRI with its coarse resolution nonetheless allows such lookup tables to be successfully built is an interesting and important result. But it does not mean that one has succeeded in understanding the brain's representations. To approach that goal, we need to be able to carry out neural decoding that generalises across subjects.

4.4 Conclusion

It might seem obvious, at first sight, that neural decoding should take as its input neural activation patterns. Indeed, all previous neural decoding approaches have done exactly that; this paper is, as far as we are aware, the first to perform neural decoding using not the neural patterns themselves, but instead the similarities between those patterns. This, we wish to argue, is precisely why it is able to achieve accurate across-subject decoding. In order to capture the commonalities across subjects, it is necessary to abstract away from their idiosyncratic and subject-specific "neural fingerprints". Performing the decoding in similarity-space does exactly that.

The concept of similarity has been found to be a powerful tool in multiple domains of cognitive psychology (Shepard, 1962; Tversky, 1977; Medin et al., 1993; Edelman, 1998) and in studies of language and conceptual structure (Miller, 1995; Landauer & Dumais, 1997; Pedersen et al., 2004; Storms et al., 2010). The idea of classifying stimuli based on their similarities, rather than on the features of the stimuli themselves, has also attracted considerable attention in the machine-learning literature (Pekalska & Duin, 2005; Chen et al., 2009). In fMRI research, most investigations of neural similarity have been in the domain of visual object recognition (Edelman et al., 1998; Hanson et al., 2004; O'Toole et al., 2007; Kriegeskorte et al., 2008; Connolly et al., 2011; Shinkareva et al., 2011), but it has also been found to be important in memory (Xue et al., 2010) and olfaction (Howard et al., 2009). Indeed, in animal neurophysiology studies of olfaction, the concept of neural similarity is central (Cleland et al., 2002; Guerrieri et al., 2005; Haddad et al., 2008; Dupuy et al., 2010). These considerations suggest

that our new approach for decoding in similarity-space may have broad applicability, across multiple neural and behavioural domains.

Our across-subject neural decoding demonstrates the match between different people's representational schemes, by accurately translating between them. It achieves this by operating entirely within similarity-space. Whether drawing upon neural information from within a specific cortical area, or from disparate and diverse neural populations, this reveals the common organising principles of neural representation which make different people alike.

5 Acknowledgments

We would like to thank Jim Haxby for permission to use the dataset from his 2001 Science paper, and Daniel Ansari, Silvia Bunge, Shimon Edelman, Niko Kriegeskorte and Russ Poldrack for very helpful comments on earlier versions of the manuscript. Andrew Connolly was funded in part by NIMH NRSA grant 1F32MH085433-01A1. Correspondence should be addressed to raizada@cornell.edu

References

Caramazza, A. & Shelton, J. R. (1998). Domain-specific knowledge systems in the brain the animate-inanimate distinction. *J Cogn Neurosci*, *10*(1), 1–34.

Chen, Y., Garcia, E., Gupta, M., Rahimi, A., & Cazzanti, L. (2009). Similarity-based classification: Concepts and algorithms. *The Journal of Machine Learning Research*, *10*, 747–776.

Churchland, P. M. (1986). Some reductive strategies in cognitive neurobiology. *Mind*, *95*(379), 279–309.

Churchland, P. M. (1998). Conceptual similarity across sensory and neural diversity: The Fodor/Lepore challenge answered. *J Philosophy*, *95*(1), 5–32.

Cleland, T. A., Morse, A., Yue, E. L., & Linster, C. (2002). Behavioral models of odor similarity. *Behav Neurosci*, *116*(2), 222–31.

Clithero, J. A., Smith, D. V., Carter, R. M., & Huettel, S. A. (2011). Within- and cross-participant classifiers reveal different neural coding of information. *Neuroimage*, *56*(2), 699–708.

Connolly, A. C., Gobbini, M. I., & Haxby, J. V. (2011). Three virtues of similarity-based multivariate pattern analysis: An example from the human object vision pathway. In N. Kriegeskorte & G. Kreiman (Eds.), *Understanding visual population codes: Toward a common multivariate framework for cell recording and functional imaging*. Cambridge, MA: MIT Press.

Daubechies, I., Roussos, E., Takerkart, S., Benharrosh, M., Golden, C., D'Ardenne, K., Richter, W., Cohen, J. D., & Haxby, J. (2009). Independent component analysis for brain fMRI does not select for independence. *Proc Natl Acad Sci U S A*, *106*(26), 10415–22.

Duncan, R. O. & Boynton, G. M. (2003). Cortical magnification within human primary visual cortex correlates with acuity thresholds. *Neuron*, *38*(4), 659–71.

Dupuy, F., Josens, R., Giurfa, M., & Sandoz, J.-C. (2010). Calcium imaging in the ant camponotus fellah reveals a conserved odour-similarity space in insects and mammals. *BMC Neurosci*, *11*, 28.

Edelman, S. (1998). Representation is representation of similarities. *Behav Brain Sci*, *21*(4), 449–67; discussion 467–98.

Edelman, S., Grill-Spector, K., Kushnir, T., & Malach, R. (1998). Toward direct visualization of the internal shape representation space by fMRI. *Psychobiology*, *26*(4), 309–321.

Epstein, R. & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601.

Fodor, J. & Lepore, E. (1999). All at sea in semantic space: Churchland on meaning similarity. *J Philosophy*, *96*(8), 381–403.

Fodor, J. A. & Lepore, E. (1992). *Holism: a shopper's guide*. Oxford: Blackwell.

Goldstone, R. L. & Rogosky, B. J. (2002). Using relations within conceptual systems to translate across conceptual systems. *Cognition*, *84*(3), 295–320.

Guerrieri, F., Schubert, M., Sandoz, J.-C., & Giurfa, M. (2005). Perceptual and neural olfactory similarity in honeybees. *PLoS Biol*, *3*(4), e60.

Guntupalli, J. S. & Haxby, J. V. (2009). Inter-subject hyperalignment of neural representational space for objects. *Society for Neuroscience Abstracts*, *262.20*.

Haddad, R., Khan, R., Takahashi, Y. K., Mori, K., Harel, D., & Sobel, N. (2008). A metric for odorant comparison. *Nat Methods*, *5*(5), 425–9.

Han, J. & Kamber, M. (2006). *Data mining: concepts and techniques* (2nd ed ed.). Amsterdam: Elsevier.

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Olivetti, E., Fründ, I., Rieger, J. W., Herrmann, C. S., Haxby, J. V., Hanson, S. J., & Pollmann, S. (2009). PyMVPA: A unifying approach to the analysis of neuroscientific data. *Front Neuroinformatics*, *3*, 3.

Hanson, S. J., Matsuka, T., & Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *Neuroimage*, *23*(1), 156–66.

Hanson, S. J. & Schmidt, A. (2011). High-resolution imaging of the fusiform face area (FFA) using multivariate non-linear classifiers shows diagnosticity for non-face categories. *Neuroimage*, *54*(2), 1715–34.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430.

Haynes, J.-D. & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nat Rev Neurosci*, *7*(7), 523–34.

Howard, J. D., Plailly, J., Grueschow, M., Haynes, J.-D., & Gottfried, J. A. (2009). Odor quality coding and categorization in human posterior piriform cortex. *Nat Neurosci*, *12*(7), 932–8.

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *J Neurosci*, *17*(11), 4302–11.

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., Tanaka, K., & Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–41.

Kuncheva, L. I., Rodriguez, J. J., Plumpton, C. O., Linden, D. E. J., & Johnston, S. J. (2010). Random subspace ensembles for fMRI classification. *IEEE Trans Med Imaging*, *29*(2), 531–42.

Laakso, A. & Cottrell, G. (2000). Content and cluster analysis: assessing representational similarity in neural systems. *Philosophical Psychology*, *13*(1), 47–76.

Landauer, T. & Dumais, S. (1997). A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211.

Martin, A. (2007). The representation of object concepts in the brain. *Annu Rev Psychol*, *58*, 25–45.

McCarthy, G., Puce, A., Gore, J., & Allison, T. (1997). Face-specific processing in the human fusiform gyrus. *J Cogn Neurosci*, *9*(5), 605–610.

Medin, D. L., Goldstone, R. L., & Gentner, D. (1993). Respects for similarity. *Psychological Review*, *100*, 254–278.

Miller, G. (1995). WordNet: a lexical database for English. *Communications of the ACM*, *38*(11), 39–41.

Monti, M. M., Vanhaudenhuyse, A., Coleman, M. R., Boly, M., Pickard, J. D., Tshibanda, L., Owen, A. M., & Laureys, S. (2010). Willful modulation of brain activity in disorders of consciousness. *N Engl J Med*, *362*(7), 579–89.

Morrison, S. E. & Salzman, C. D. (2010). Re-valuing the amygdala. *Curr Opin Neurobiol*, *20*(2), 221–30.

Mourao-Miranda, J., Bokde, A. L. W., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data.

Neuroimage, *28*(4), 980–95.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci*, *10*(9), 424–430.

O'Toole, A. J., Jiang, F., Abdi, H., & Haxby, J. V. (2005). Partially distributed representations of objects and faces in ventral temporal cortex. *J Cogn Neurosci*, *17*(4), 580–90.

O'Toole, A. J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J. P., & Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *J Cogn Neurosci*, *19*(11), 1735–52.

Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004). Wordnet:: Similarity: measuring the relatedness of concepts. In *Demonstration Papers at HLT-NAACL 2004 on XX*, (pp. 38–41). Association for Computational Linguistics.

Pekalska, E. & Duin, R. P. W. (2005). *The dissimilarity representation for pattern recognition: foundations and applications*. Hackensack, N.J.: World Scientific.

Pereira, F., Mitchell, T., & Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. *Neuroimage*, *45*(1 Suppl), S199–209.

Poldrack, R. A., Halchenko, Y. O., & Hanson, S. J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychol Sci*, *20*(11), 1364–72.

Raizada, R. D. S. & Kriegeskorte, N. (2010). Pattern-information fMRI: new questions which it opens up, and challenges which face it. *International Journal of Imaging Systems and Technology*, *20*, 31–41.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, *27*(2), 125–140.

Shinkareva, S. V., Malave, V. L., Just, M. A., & Mitchell, T. M. (2011). Exploring commonalities across participants in the neural representation of objects. *Hum Brain Mapp*.

Shinkareva, S. V., Mason, R. A., Malave, V. L., Wang, W., Mitchell, T. M., & Just, M. A. (2008). Using fMRI brain activation to identify cognitive states associated with perception of tools and dwellings. *PLoS ONE*, *3*(1), e1394.

Storms, G., Navarro, D. J., & Lee, M. D. (2010). Introduction to the special issue on formal modeling of semantic concepts. *Acta Psychol (Amst)*, *133*(3), 213–5.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Wang, X., Hutchinson, R., & Mitchell, T. (2003). Training fMRI classifiers to detect cognitive states across multiple human subjects. In *Proceedings of the 2003 Conference on Neural Information Processing Systems*.

Warrington, E. K. & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*, 829–54.

Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. *Science*, *330*(6000), 97–101.