Finding motifs in mental representations of faces, places, and objects

Y. Ivette Colón (ycolon@wisc.edu)

Timothy T. Rogers (ttrogers@wisc.edu)

Abstract

Cognitive science has long relied on the assumption that mental representations are universal across healthy adults, treating individual differences as random noise. We challenge this assumption by proposing that representations instead conform to a limited set of organizational motifs—systematic patterns shared across subgroups of individuals. Using triadic comparisons and embedding analysis, we examine how people conceptually organize a set of DallE-generated faces, places, and objects that systematically vary in five attributes of interest per domain. We show that individuals cluster into distinct "representational motifs" when organizing faces, places, and objects. Logistic regression analyses show that these motifs differ in the relative use of image attributes to form mental representations. Our findings demonstrate that variability in conceptual organization is not merely noise, but rather reflects meaningful patterns of shared representational frameworks that emerge naturally.

Keywords: conceptual organization; mental representation; universality; dalle; triadic comparison; triplet embedding

Introduction

In cognitive science, efforts to estimate the structure of mental representations often rely on a convenient fiction: that the structures of interest are more-or-less the same within a given population of healthy adults, and that individual differences reflect random variation around the common underlying structure. This idea, sometimes called the *assumption of universality* (Caramazza, 1984), is convenient for two reasons. First, it licenses the parametric statistics that underlie much empirical work: if any individual's behavior is a noisy reflection of a universal underlying structure, then aggregating behaviors across individuals is an effective way to cancel the noise and reveal the structure. Second, it is not clear what the alternative might be. If different (healthy, adult) members of a common population do *not* all possess largely shared mental structures with random variation—if in fact different individuals or subgroups organize representations in a given domain in systematically different ways—how then can we figure out what various organizations exist in which individuals or subgroups, short of treating everyone as an individual case study?

Despite its convenience, however, it has long been clear the assumption of universality is indeed a fiction. Pioneering studies of expertise (Atran et al., 2004; Chi et al., 1981; Tanaka & Gauthier, 1997) and conceptual development (Carey, 1992) showed decades ago that mental representations restructure themselves as they develop, so that people can discern quite different relations among the same set of stimulus items. Studies of phonological and visual perception find that the very sounds we hear in speech (Kuhl, 1991) or differences we perceive in color (Winawer et al., 2007) change depending on the language we speak. Studies of reading (Seidenberg & McClelland, 1989) or mathematical reasoning (Alibali & GoldinMeadow, 1993) reveal varied strategies for each, with implications for the way elements of each domain are organized. In these and many other cases, people organize their thoughts and perceptions about a given domain in quite different ways, depending upon a variety of factors.

Here, we propose an alternative to the assumption of universality that preserves some of its convenience while still allowing systematic variation across individuals. Specifically, we hypothesize that mental representations in a given domain hew to a limited set of possible *motifs*. Although individuals may vary, they do not vary randomly, but rather along a somewhat constrained set of representational possibilities. If this were the case, then empirical and statistical methodologies could be designed to map out the space of possible representations across a population, as well as the specific instantiation arising within any given individual.

The main goal of this paper is to test whether mental representations in three example domains accord with this hypothesis. We first consider how representational motifs might be discovered among members of a (relatively, ostensibly) homogeneous population, using a data-driven approach. By data-driven, we mean that we need not divide the population into hypothetical subgroups, such as experts vs non-experts, or well-educated vs poorly-educated, or some division of that nature. Instead, the idea is to collect behavioral data sufficient to estimate how an individual person organizes their mental representations within a given concept domain, and then measure how similar the various structures are to one another. The central question is whether conceptual organization varies randomly from person to person, or whether they fall into a small repertoire of representational motifs.

Our approach relies on *triadic comparisons* to estimate embeddings of stimuli that reveal the underlying similarities that a person discerns among them. In each trial, participants view a *reference* and two *option* items, and must decide which option is more similar to the target. From many such judgments one can embed the various items in a low-dimensional metric space, such that items often chosen as "more similar" to one

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

another are nearby and those rarely chosen are far apart. This approach is commonly known as ordinal multidimensional scaling, and differs from classical multidimensional scaling (MDS) in the nature of responses required to reach an embedding solution: ordinal embeddings do not require rank ordering of all pairwise comparisons between a set of items, and can instead be reliably derived from samples of triadic comparisons (Vankadara et al., 2023). The ordinal embedding generated for an individual provides a kind of snapshot of the similarity relations the person uses to guide behavior, and we can compare how these snapshots relate to one another across individuals. In this work, we not only employ ordinal embedding as a means of estimating individuals' conceptual similarity spaces, but we can also *interpret* the relative contributions of stimulus attributes to such similarity spaces and uncover groups of individuals who think similarly in systematic and interpretable ways.

We apply this approach to evaluate how mental representations vary in three commonly-studied visual knowledge domains: faces, places and objects. We selected these domains because they encompass highly familiar and arguably universal kinds, sometimes thought to be supported by dedicated systems in the brain (Kanwisher, 2001). That is, if any variety of mental representation should accord with the assumption of universality, it should be faces, places, and objects; and conversely, if mental representations vary along limited motifs in these domains, they likely do for many others.

Finding groups that "think alike"

Below, we establish a pipeline for uncovering groups of people that respond similarly to a shared set of stimuli. To do so, we capture individual conceptual organization for a controlled yet naturalistic set of images through triadic comparison tasks. The data generated from this task yield multidimensional embeddings that reflect meaningful conceptual organization along manipulated stimulus attributes. By clustering the similarity of embeddings between participants, we show that these potentially unconstrained, idiosyncratic mental representations cluster into "kinds of thinkers" with similar mental motifs. Logistic regression models predicting feature attributes from cluster embeddings reveal the specific mix of attributes that different groups prioritize in their conceptual organization. We examine the validity of these embeddings and clustering solutions by predicting trial-level decision-making from one's own embedding and those in and out of their cluster.

Generating stimuli

We first generated a set of stimuli across the three domains of faces, places, and objects that systemically varied along five attributes. For each possible combination of five attributes within a domain, we created two example images, resulting in 32 images per domain ($2^5 = 32$). To establish whether there were clear, dissociable dimensions of semantic dissimilarity within a population, we chose to manipulate attributes along binarized features (see Figure 1A).

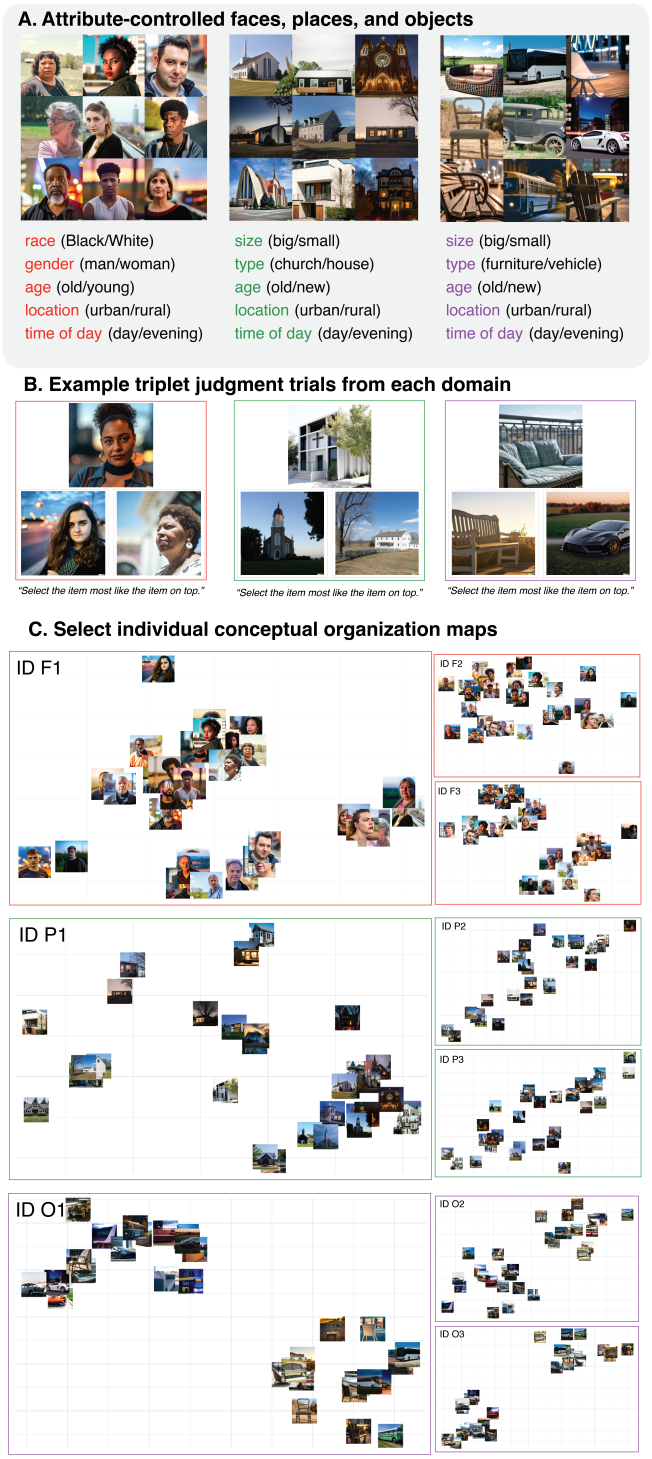

Figure 1: A: Example stimuli for each task domain and the binary attributes used to generate attribute-controlled stimuli. B: Example trials from each triadic comparison task. C: Randomly-selected example conceptual maps from different identities.

Each domain varies along the dimension of age (e.g. older/younger people, old/modern buildings, old/modern objects) and the environmental context of each exemplar: location (urban/rural) and time of day (mid-day/evening). Place stimuli varied along categorical membership of the type of building (house/church), and the size of the building (big/small). Objects too varied in size (big/small) and categorical membership of the type of object (vehicle/furniture).

For our face stimuli, we additionally manipulate exemplars along the dimensions of gender (men/women) and race (Black/White). We acknowledge that these binarized social categories are extremely limited and do not represent the full spectrum of gender, race, nor the presentation thereof. Although categorizing people into broad "types" is problematic in its own right, we want to acknowledge the real, intersectional differences in the experiences of between Black and white men and women, which are influenced by how others organize social attributes.

Capturing triadic comparisons

For each domain, we conducted procedurally identical but separate triadic comparison tasks. Tasks were custom-coded using JsPsych and deployed online via Prolific. Each task took approximately 30 minutes to complete, and participants were compensated at a rate of $12 USD per hour.

In each trial of each task, participants saw three images: one "reference" item and two "choice" items beneath the reference item (see Figure 1B). Participants were instructed to "Choose the item most like the item on top" and were allowed to use any criteria in making their decisions. For each task domain, we used all 32 images to construct triadic comparison trials ("triads").

Each participant saw 610 total trials: 300 randomly-generated trials; 10 vigilance trials (where one of the choice items was identical to the reference item, used as attention checks); and 300 validation trials. The validation trials were constructed such that, for each person, 200 unique trials were randomly sampled from a larger validation pool of 1,000 randomly-generated triads shared across all participants, and were then further sampled from the 200 selected triads, resulting in 300 trials that have some proportion of trials shared across individuals, and some proportion repeated within an individual. Random, vigilance, and validation trials were interspersed throughout each study, and participants had the opportunity to take a break halfway through the experiment. Following the triadic comparison trials, participants completed a short demographic survey collecting information on their race and ethnicity, age, and gender.

As part of the study, participants received feedback about their performance if they responded to triadic comparison trials in less than 1 second, or if they responded incorrectly to a vigilance trial (not choosing the item identical to the reference item). We excluded participants who had an average triadic comparison reaction time (RT) that was faster than the log of the mean RT across participants, and/or got less than 80% of vigilance trials correct.

Table 1: Demographics

Task	*n*	Gender (W/M/O%)	Race/Ethnicity (W/B/H/A/O%)	Age (*M*, Range)
Faces	41	64/36/0	85/9/6/3/0	33.2, 18–68
Places	39	51/44/5	85/8/0/18/5	35.0, 18–71
Objects	47	51/47/2	75/9/17/9/6	30.0, 18–72

Note: W=Women, M=Men, O=Other/Not disclosed;
W=White, B=Black/African American, H=Hispanic/Latino,
A=Asian, O=Other (incl. Pacific Islander, Am. Indian/Alaska Native)

Computing embeddings

To assess conceptual organization at the individual level, we first computed separate ordinal embeddings for each participant. For each participant, we input all of their non-vigilance, and non-repeated validation trial responses into a modified version of an ordinal embedding algorithm available in the Salmon Python package (Sievert et al., 2023). This embedding algorithm requires specification of several key parameters, including the number of output dimensions, the proportion of training and testing examples for learning embeddings, the measure to we wish to optimize, and the number of epochs to train and test the embedding model.

As our stimuli in each domain vary along five key dimensions, we chose to generate 5-dimensional embeddings, whereby each of the 32 stimulus items in a domain is represented by a five-dimensional coordinate. If all five dimensions of stimulus variation are represented in participants' responses, 5-dimensional embeddings should be sufficient to capture such variation (assuming that there are no unintentional dimensions of variation discernible in our stimuli). Based on pilot data for this experiment, embedding responses in five dimensions (vs. two, three, four, and six) best minimized prediction error in embedding models across participants. This is not to say that embeddings need be 5-dimensional to encompass individual similarity spaces– some participants may only be using a subset of stimulus attributes to base their decisions on. As will be described, our analyses accommodate such a possibility, and allow us to identify which attributes significantly contribute to five-dimensional similarity spaces.

The embedding model relies on a set of training trials to learn from and generate an embedding space, and a set of held-out trials with which to test its predicted embedding space. Here, we used 80% of each person's trials as the training set, and the remaining 20% of their trials as the test set. Training and testing sets were chosen randomly for each participant.

Our modifications to this algorithm primarily concern the optimization performance criterion, where rather than rewarding the *accuracy* of estimates for predicting unseen test triads at each epoch, we reward estimates that *minimize the error* for unseen test triads. Our rationale for this modification is that, when using accuracy to test triadic comparisons, we are essentially using a binary measure— either the embedding correctly predicts which two items are most similar, or it does not. This is a coarse-grained measure of the "goodness" of an embedding space, relative to the rather fine-grained con-

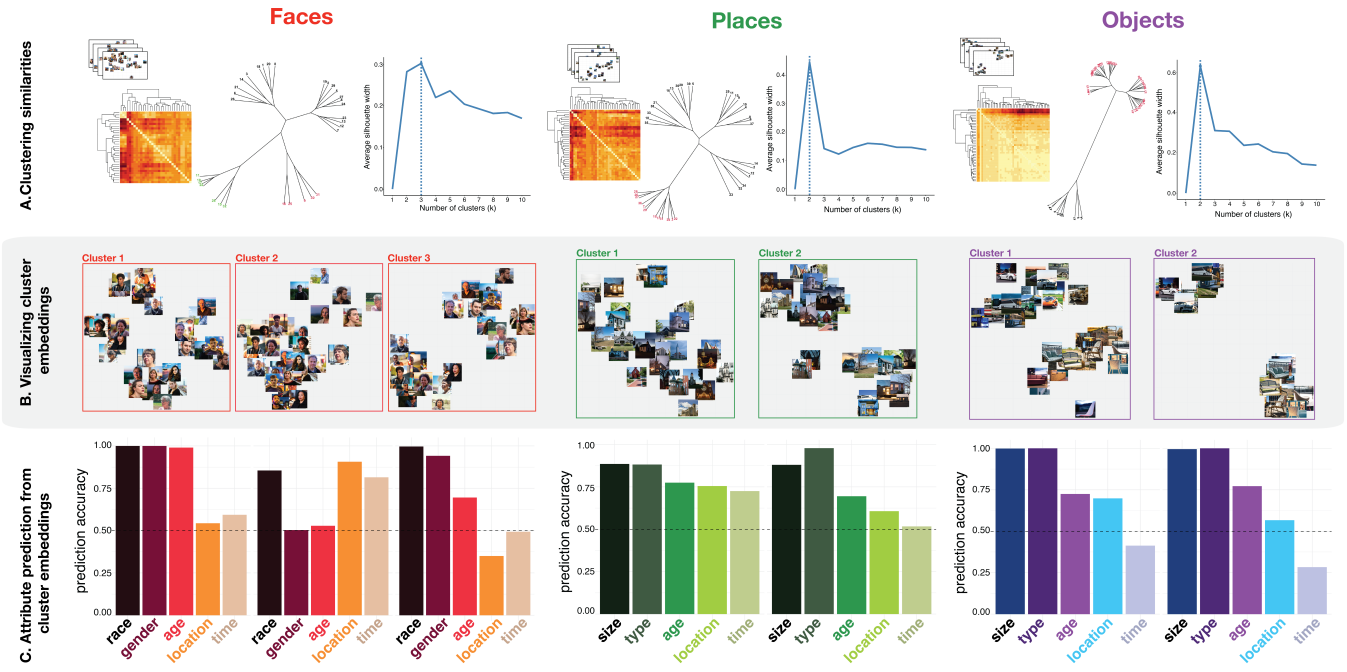

Figure 2: A: Clustering similarities by first taking all participants embeddings, Procrustes aligning them and capturing their similarity, then hierarchically clustering these similarities. Silhouette scores help determine the number of clusters. B: Visualizations of cluster-level embeddings with the actual images as points. C: Prediction of image attributes for each cluster's embeddings using logistic regression to predict binary category membership (chance = 50%).

ceptual spaces we develop. By contrast, prediction error metrics allow us to get a sense of just *how* wrong a prediction is by reflecting distance error for each triad member prediction, which in turn allows us to land on embeddings that better reflect participants' nuanced similarity judgments.

If the test loss error does not improve for 10,000 consecutive epochs, we stop training and use the last error-minimizing embedding as the "best fit" embedding. We chose 50,000 as the maximum number of training epochs for generating embeddings. This training scheme allows the optimization process to fully converge on a stable solution while avoiding over-training, as the algorithm may need many iterations to find an optimal arrangement that satisfies the complex network of ordinal relationships but may not need 50,000 epochs to find such an embedding.

Measuring embedding similarity

Distances in the 5-dimensional space generated by the ordinal embedding are based on the *relative* similarity of items, where items closer in space are considered more conceptually similar than items further apart, but the axes of this space are not inherently meaningful. As such, we require a measure of similarity between embeddings that takes this relativity into account. Procrustes analysis allows us to compute the difference between the 5-dimensional "shapes" formed by individual participants' embeddings. Procrustes analysis calculates the similarity between two embeddings by finding the optimal shape transformations to align the two sets (e.g. scaling, rotating, reflecting), then measuring the distance (error) between corresponding points. In the present study, we use the sum of squared errors between embeddings as a measure of their similarity, and compute this error measurement for each pair of participant embeddings within a domain. The result is a matrix of embedding similarities, where rows and columns are individual participants and each entry is the similarity of embeddings between pairs of participants. We use this matrix of participant similarities in further analyses.

Clustering individual embeddings

To do so, we perform an iterative process of hierarchically clustering participants based on their embedding similarity at different numbers of clusters (k), then calculate the silhouette score at each level of k to find the optimal number of participant clusters for each domain (see Figure 2A). Once we establish an optimal k, we visualize these clusters and their relative similarity as phylogenetic tree, and inspect these clustering solutions. Finding the optimal number of clusters in which data lies can be more of an art than a science— sometimes visual inspection of clustering data and quantitative measures of optimal clustering (e.g. silhouette scores or mutual information) can offer different solutions. For the purposes of this work, we use the silhouette statistic as the basis for the number of participant groups in our data.

Clustering analyses reveal three distinct clusters or groups of respondents for face stimuli, two groups for place stimuli, and two for object stimuli. We next establish what makes

these groups different from one another by examining the information recoverable from the embeddings generated for each cluster group.

Predicting attributes from cluster embeddings

Having established unique groups of respondents directly from the triadic comparison data for each domain, we next compute cluster-level ordinal embeddings. For each cluster group, we used all non-vigilance and non-duplicate validation trials from group members to compute a cluster-level embedding. The embedding computation procedure for group-level embeddings was identical to the computation of the individual-level embeddings (i.e. five-dimensional output, 80%/20% training/testing split, minimizing test loss error, and 50,000 epoch maximum). This procedure results in seven cluster-level embeddings: three for each group in the face task, two for the place task, and two for the object task.

How can we examine what information is contained within these embeddings? The two-dimensional visualizations of group-level cluster embeddings in Figure 2B can reveal the salient guiding dimensions within each groups' conceptual organization. For example, Cluster 1 in the face task shows distinct quadrants of images along dimensions of race and gender; Cluster 2 in the place task shows clear delineation between churches and homes; and Cluster 1 in the object task shows categorical separation of vehicles and furniture, and within each category reflects separation by size.

Two-dimensional visualizations, however, do not reveal the full picture of conceptual organization. Instead, we leverage the binary nature of our stimuli attributes for use in classification. We used the five-dimensional embedding coordinates for each stimulus as predictors in separate logistic regression models predicting each of the binary stimulus attributes. Binary attribute prediction allowed us to quantify the presence of an attribute within embedding— the greater the prediction accuracy for a given attribute from an embedding, the more meaningful that attribute is for conceptual organization.

For faces, separate models predicted age (older adult/younger adult), race (Black/white), and gender (men/women); place models predicted age (old/modern), type (church/house), and size (small/large); object models predicted age (old/modern), type (furniture/vehicle), and size (small/large); for all domains, separate models predicted location (urban/rural) and time of day (midday/evening). Each logistic regression model used cross-validation, whereby it was trained on embeddings for 28 of the 32 stimulus items in each domain and tested on the remaining 4 items. This process was repeated 100 times with random training and testing samples to generate distributions of prediction performance. Figure 2C depicts the mean prediction performance for each attribute across all cluster groups and domains.

What attributes do groups use?

As shown in Figure 2C, cluster groups seem to rely on different stimulus attributes for conceptual organization.

For faces, Cluster 1 embeddings show perfect or near-perfect prediction accuracy for race (100%), gender (100%), and age (99%), and near-chance (chance = 50%) performance for location (54.33%) and time of day (59.33%), suggesting that this group relies on all manipulated *person-focused* attributes for organizing people, but not on environmental context. By contrast, location and time of day are both well predicted by Cluster 2 embeddings (90.67% and 81.50%, respectively) along with race (85.50%), but not age (52.83%) or gender (50.33%) which are both relatively flexible social attributes. Finally, Cluster 3's conceptual organization shows varied prediction capacity by attribute, with race (99.66%), gender (94.16%), and age (69.50%) predicted above chance, time of day around chance (49.33%), and location well-below chance (35%).

Place clusters only split into two groups: Cluster 1 shows above chance, but not exceptional, performance on all attributes, prioritizing place size (88.50%) and type (88.17%)) over age (77.50%), time (72.50%), and location (60.67%). Cluster 2 embeddings can predict type near perfectly (97.83%), followed by size (88%), age (69.50%), location (60.67%) and time at near-chance accuracy (51.67%).

Object clusters were also split in two, however the distinction between the two clusters based on attribute prediction is less pronounced: Cluster 1 showed embeddings predicted object type and size near perfectly (100% and 99.83%, respectively) and both age and location above chance (72.50% and 69.83% respectively), and time of day below chance (41.33%). Similarly, Cluster 2 embeddings predicted object type and size near perfectly (100% and 99.50%, respectively) and age above chance (77.17%), followed by location (56.67%), and time of day well-below chance (28.33%).

Embeddings reflect meaningful individual and group-level differences

How can we be sure that our individualized embeddings reflect meaningful idiosyncrasies in conceptual organization and subsequent decision-making? Uncovering embedding similarity does not necessarily reveal meaningful recovery of conceptual and behavioral motifs. In other words, are these embeddings (and their clusters) "real"?

To evaluate whether participants' embeddings could predict choice behavior, we implemented a held-out validation framework based on the triadic judgments, predicting responses to unseen triads from embeddings. We first defined a distance-based prediction function that operates over each participant's 5-dimensional embedding. For a given trial, we calculated the Euclidean distance in the embedding between the reference item and each of the choice items. Our model predicts that for any given triad, participants would choose the item closer to the reference item. To test how well embeddings predicted actual triadic judgments, we computed prediction accuracy for each participant using three different sources of embeddings: 1) self (participant's own embedding), 2) cluster mates (embeddings from other participants assigned to the same cluster), and 3) non-cluster mates.

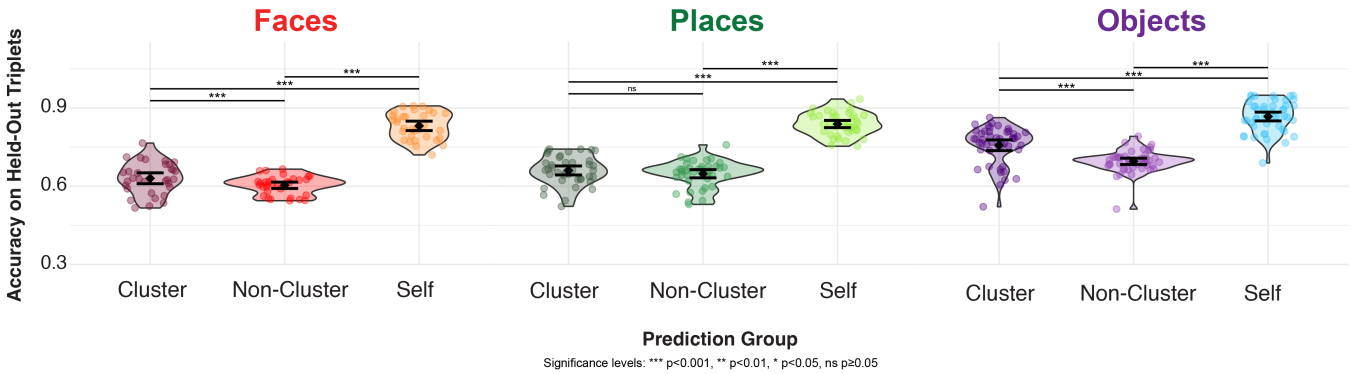

Figure 3: Accuracy for predicted responses to held-out triadic judgment trials using one's own embedding (self), the embeddings of those in the same cluster, and embeddings from those in different clusters.

Each participant's trials were randomly split into an 80%/20% training/testing sets. For each test trial, the model used an embedding (from self, cluster mates, or non-cluster mates) to generate a prediction. Cluster-based predictions were computed by averaging accuracy across all relevant participants. Prediction accuracy scores were aggregated by participant and group (self, cluster mates, non-cluster mates), and compared using paired-sample t-tests.

Prediction analyses show that domain clusters vary in how well they reflect triadic judgment behavior. For all three domains, one's own embedding best predicts trial responses (p 2.2e-16). For faces and objects, those in one's own cluster better predicts their responses than those out of their cluster, suggesting that for at least faces and objects, embedding cluster groups and their associated representational motifs correspond to actual differences in behavior (faces p = 4.085e-05; objects p = 6.501e-08; places p = 0.21).

Discussion

In this work, we use an approach that captures high-dimensional similarity judgments without relying on predefined categories or linguistic labels, and identify distinct groups of individuals who share common organizational frameworks for either faces, places, and objects. Our findings show that mental representations need not be treated as either universal across individuals or entirely idiosyncratic. Instead, we find evidence for a middle ground where individuals cluster into a limited set of organizational patterns.

Each cluster represents a coherent organizational strategy shared by multiple individuals. Particularly telling is the fact that we found these motifs in a domain (faces) often thought to rely on dedicated neural systems. If the assumption of universality were to hold anywhere, it should be for the face domain. The fact that we instead find systematic variation, even here, suggests that similar organizational motifs likely exist across other domains of human knowledge.

The varying number of clusters across domains— three for faces vs. two for objects vs. one for places— offers insight into how domain content influences representational diversity. Social stimuli showed the greatest variety in organizational motifs. Face task clusters showed marked differences in their prioritization of social attributes: while one group organized faces primarily by person-related characteristics (race, gender, age), another incorporated environmental context, and a third showed a mixed pattern of attribute weighting. This increased diversity in social representation makes theoretical sense: our engagement with others, and face stimuli at all, is deeply influenced by personal experience, cultural context, social identity, and clinical diagnoses, leading to more varied ways of organizing social information (Freeman et al., 2020).

The relative consistency in place and object organization suggests that some domains may show greater universality than others. However, crucially, even the more constrained object domain showed systematic variation, rather than random individual differences around a universal structure. The relative uniformity of place-related organization strategies may also reflect more shared experiences with this domain, or possibly more constrained ways of organizing non-social concepts based on functional or physical properties. Another possibility is that the increased similarity of place stimuli categories (houses and churches share many features) lead to less predictable trial-level responses. Further investigation is necessary to parse why behaviorally distinct groups are not recoverable from place embedding organization.

Conclusion

The findings presented here demonstrate that the assumption of universality, while convenient, fails to capture the systematic ways in which mental representations vary across individuals. At the same time, we show that such variation is not random but falls into coherent patterns— representational motifs— that can be discovered through controlled empirical work. This finding supports a larger trend in cognitive science: one that preserves the tractability of studying shared mental structures while acknowledging and investigating systematic individual differences in how people organize their knowledge of the world.

Acknowledgments

This work was supported by an NSF grant to TTR (NSF 21-517).

References

Alibali, M. W., & GoldinMeadow, S. (1993). Gesture-speech mismatch and mechanisms of learning: What the hands reveal about a child s state of mind. *Cognitive psychology*, *25*(4), 468–523.

Atran, S., Medin, D., & Ross, N. (2004). Evolution and devolution of knowledge: A tale of two biologies. *Journal of the Royal Anthropological Institute*, *10*(2), 395–420.

Caramazza, A. (1984). The logic of neuropsychological research and the problem of patient classification in aphasia. *Brain and language*, *21*(1), 9–20.

Carey, S. (1992). The origin and evolution of everyday concepts. *Cognitive models of science*, *15*, 89–128.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. *Cognitive science*, *5*(2), 121–152.

Freeman, J. B., Stolier, R. M., & Brooks, J. A. (2020). Dynamic interactive theory as a domain-general account of social perception. In *Advances in experimental social psychology* (pp. 237–287, Vol. 61). Elsevier.

Kanwisher, N. (2001). Faces and places: Of central (and peripheral) interest. *Nature neuroscience*, *4*(5), 455–456.

Kuhl, P. K. (1991). Human adults and human infants show a "perceptual magnet effect" for the prototypes of speech categories, monkeys do not. *Perception & psychophysics*, *50*(2), 93–107.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, *96*(4), 523.

Sievert, S., Nowak, R., & Rogers, T. T. (2023). Efficiently learning relative similarity embeddings with crowdsourcing. *Journal of open source software*, *8*(84).

Tanaka, J. W., & Gauthier, I. (1997). Expertise in object and face recognition. *Psychology of learning and motivation*, *36*, 83–125.

Vankadara, L. C., Lohaus, M., Haghiri, S., Wahab, F. U., & Von Luxburg, U. (2023). Insights into ordinal embedding algorithms: A systematic evaluation. *Journal of Machine Learning Research*, *24*(191), 1–83.

Winawer, J., Witthoft, N., Frank, M. C., Wu, L., Wade, A. R., & Boroditsky, L. (2007). Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, *104*(19), 7780–7785.