Computational approaches to the neuroscience of social perception

Jeffrey A. Brooks

New York University

Ryan M. Stolier

Columbia University

Jonathan B. Freeman

New York University

Corresponding author:

Jeffrey A. Brooks or Jonathan B. Freeman, Ph.D.
Department of Psychology
New York University
6 Washington Place
New York, NY 10003
Email: jab1148@nyu.edu or jon.freeman@nyu.edu

Abstract

Across multiple domains of social perception - including social categorization, emotion perception, impression formation, and mentalizing - multivariate pattern analysis (MVPA) of fMRI data has permitted a more detailed understanding of how social information is processed and represented in the brain. As in other neuroimaging fields, the neuroscientific study of social perception initially relied on broad structure-function associations derived from univariate fMRI analysis to map neural regions involved in these processes. In this review, we trace the ways that social neuroscience studies using MVPA have built on these neuroanatomical associations to better characterize the computational relevance of different brain regions, and how MVPA allows explicit tests of the correspondence between psychological models and the neural representation of social information. We also describe current and future advances in methodological approaches to multivariate fMRI data and their theoretical value for the neuroscience of social perception.

Computational approaches to the neuroscience of social perception

Other people are tremendously complex, and humans must navigate their relationships and interactions with others under conditions of high uncertainty. Whether meeting a stranger, reading a description of someone, or trying to determine how a friend is feeling, we rely on a set of perceptions and inferences about the person to determine our behavior. Understanding how we form impressions about others has been a central focus in social psychology for decades, and more recently the topic has proven to be well-suited to methods from computational neuroscience, which can readily leverage the inherently high-dimensional nature of neuroimaging data alongside behavioral measures of social perception. In this article, we review recent advances in computational approaches to the neuroscience of social perception. We focus particularly on multivariate analyses of fMRI data, but computational analyses of behavioral data used in conjunction with fMRI, such as using fMRI and behavioral responses to estimate parameters of computational models, is an increasingly popular approach as well and is reviewed elsewhere (e.g., Cheong et al., 2017; Gonzalez & Chang, 2019; Hackel & Amodio, 2018).

As in other areas of social neuroscience, early fMRI studies on social perception generally focused on univariate activation-based analyses to associate relevant social cognitive processes with particular brain regions, which continues to be a valuable mainstay. This research described a number of regions important for social perception, such as the primacy of the fusiform gyrus (FG) in face processing (Haxby et al., 2000; 2002; Kanwisher et al., 1997); superior temporal sulcus (STS) in dynamic face and body perception (Grossman et al., 2000; Haxby et al., 2000; Said et al., 2010); and regions such as the medial prefrontal cortex (MPFC; Amodio & Frith, 2006; Mitchell, 2008) and temporo-parietal junction (TPJ; Saxe & Kanwisher, 2003) in thinking about others and representing their mental states ("mentalizing"). Multivariate pattern analysis (MVPA) was first introduced to neuroimaging research by Haxby and colleagues (2001), expanding the conceptual and methodological toolkit of neuroimaging researchers by providing a different way to conceptualize the patterns of activation that emerge in fMRI data. In the present review, we first briefly introduce MVPA techniques and discuss their theoretical basis as well as some related advantages and limitations. We then review in turn three domains of social perception research where these techniques have proven highly valuable: perceiving social group memberships, perceiving identity and associated traits and person knowledge, and perceiving others' emotional states.

MVPA Approaches

Univariate approaches primarily seek to relate the overall level of activation in a region to task conditions or experimental variables, which can provide neuroanatomical associations with those conditions or variables and their related psychological processes (although with exceptions, e.g., adaptation paradigms). However, neural regions assessed via fMRI do not only vary in their mean level of activation, but also in the spatial patterns of activation distributed across voxels. MVPA methods are sensitive to fine-grained differences in these spatial patterns of activation, whereas mass univariate testing treats each voxel individually, almost always ignoring the level of activation in contiguous voxels in statistical tests. This sensitivity enables MVPA to differentiate experimental conditions even in cases where mere differences in mean activation of a voxel cannot (Haxby et al., 2001; 2014).

A key assumption often made about MVPA and influencing how MVPA analyses are generally interpreted is that neural response patterns inherently contain information about an associated cognitive state (Davis et al., 2014; Haynes, 2015; Lewis-Peacock & Norman, 2014; Popov et al., 2018). Researchers can thus probe condition response patterns to see how they may differ between brain regions, revealing the involvement of different regions in processing information or representing states relevant for a given task. Thus, multivariate analyses often target regions already known to be involved in specific tasks to specify how that region is representing information throughout the task and what computational processes that region may support. Neuronal recordings in nonhuman primates have long shown that the aggregate activity of an assembly of neurons can provide a 'code' for various kinds of sensory and abstract cognitive information in the brain (i.e., a "population code"; Averbeck et al., 2006). Although fMRI voxels are far too large to be sensitive to individual neurons, neurons belonging to different neuronal assemblies (e.g., related to state 1 vs. state 2) may be distributed in different ways such that the precise assembly of neurons related to each state may vary across voxels. This could thereby give rise to distinct fMRI multi-voxel patterns associated with distinct states, despite a lack of sensitivity to individual neurons (Logothetis, 2008; Chaimow et al., 2011). In this way, MVPA provides a way to extend the population coding approach from systems neuroscience to the macro-scale populations measured with fMRI (Haynes, 2015; Haxby et al., 2001; 2014; Lewis-Peacock & Norman, 2014; but see the Conclusions section for a discussion of related limitations).

In its original application, Haxby and colleagues (2001) used MVPA to demonstrate that ventral temporal cortex shows spatially distributed response patterns to visual object categories that can be discriminated from the region's face response pattern using a classifier. Importantly, this included the fusiform face area (FFA), a region of fusiform gyrus (FG) that reliably shows higher mean activations for faces compared to other visual categories (Kanwisher, 1997). This approach demonstrated that regions which show selectivity in univariate signal for one category also can hold (perhaps even equivalent amounts of) information about other categories. Such a classification (or "decoding") analysis enables researchers to use neural response patterns to predict an associated cognitive state or stimulus condition (Lewis-Peacock & Norman, 2014). A classifier is typically trained on one set of the data and tested on at least one other held-out set. By training a classifier to discriminate between any given experimental factors (i.e., conditions or stimulus characteristics), testing the classifier on held-out data can reveal which regions are involved in representing those conditions or characteristics of the stimuli. Another common way to interpret the results of a classifier is that categorical boundaries between stimuli are "computationally relevant" in a given brain region if that region shows high classification accuracy for that category boundary. This interpretation assumes that if a brain region's patterns of activity discriminate between two stimulus categories, information about that category boundary is retained in that region's spatial response patterns because it is relevant for whatever computation that region is performing in that particular cognitive context.

In addition to being used to classify brain states by experimental conditions or stimulus categories, multi-voxel patterns associated with different conditions can be directly compared by measuring pairwise similarities in their response patterns. While sometimes quite illuminating in and of itself (showing, for example, that White individuals with a strong pro-White bias have more dissimilar neural response patterns for the Black and White race categories; Brosch et al., 2013), similarity between neural patterns can also be leveraged to test explicit theories about how the brain represents information. This technique, called Representational Similarity Analysis (RSA; Kriegeskorte et al., 2008), provides a shared framework for testing and comparing diverse models of neural computation (computational, theoretical, behavioral) by examining their second-order isomorphisms (i.e., similarities in the structure of their similarity spaces). This technique was developed by computational neuroscientists who have used it to assess how well the representational structure of the ventral-visual stream corresponds with

various computer vision models (Kriegeskorte et al., 2008; Khaligh-Razavi & Kriegeskorte, 2014; Jozwik et al., 2017).

Assuming that multi-voxel response patterns contain information about certain cognitive factors, RSA starts by computing how similar (or dissimilar) each condition or category is from each other in their neural response patterns throughout the brain. The resulting similarity space can be directly compared with any other second-order similarity space, most commonly those derived from computational models or behavioral task responses. Thus, while decoding and classification approaches can reveal which regions contain or process information about cognitive dimensions, RSA is able to directly test hypotheses about how that information is organized and represented, which can in turn address questions about which psychological or stimulus dimensions are computationally relevant in a given brain region (Kriegeskorte et al., 2008; Popal et al., 2019). Despite its explanatory power, RSA is highly computationally tractable and thus broadly approachable for researchers. RSA is especially relevant for social and affective neuroscience due to the large number of models proposed in the literature describing how social groups, emotion categories, and traits relate to one another along dimensions such as stereotype content, facial cues, and affective properties (e.g., Fiske et al., 2002; 2007; Oosterhof & Todorov, 2008; Russell, 1980). RSA allows researchers to specifically adjudicate between such models, testing competing and complementary explanations for how the brain represents and computes social information. We now turn our attention to specific areas of social perception research where MVPA and RSA have been leveraged to make important new insights.

Social Categories and Groups

An illustrative example of the contrast between univariate and multivariate fMRI analyses in the domain of social categorization comes from a set of papers that used both techniques to analyze the same dataset. In the experiment, participants were assigned to one of two arbitrarily defined mixed-race groups and then had to categorize faces from the two groups along either in-group vs. out-group or Black vs. White dimensions. The first paper (Van Bavel et al., 2011) reported an in-group selectivity effect in the FFA, such that a univariate in-group vs. out-group contrast showed greater mean activity in response to in-group vs. out-group members, despite the fact that both groups were mixed-race. However, a follow-up report showed that, despite the overall difference in activation in response to in-group members, the race of the faces

could still be discriminated by a multivariate pattern classifier (Ratner, Kaul, & Van Bavel, 2013). These results indicated that the race of the faces was still computationally relevant in the FFA, regardless of the difference in mean activation driven by the social context and current processing goals.

Many studies examining univariate social category responses in regions such as the FG/FFA interpreted greater activation for one category vs. another as enhanced processing of or greater attention to the target category in the contrast (Golby et al., 2001; Lieberman et al., 2005). Some have interpreted such results as providing a neural basis for long-standing out-group deficit effects in social psychology, such as better memory for in-group vs. out-group faces along a number of dimensions (Hugenberg et al., 2010; Meissner & Brigham, 2001). Multivariate decoding approaches simultaneously challenge and complement these results by demonstrating that even if a brain region does not show univariate selectivity for a given category, it still might represent and process information about that category.

Still, to date, relatively few studies have used multivariate decoding on social category responses. In one of the earliest applications of MVPA to social categorization, Kaul et al. (2011) found that face gender could be reliably decoded from an array of brain regions commonly associated with various levels of face processing. Classifier performance was highest in the medial orbitofrontal cortex (mOFC), FG, and inferior occipital gyrus (IOG). Contreras and colleagues (2013) showed that both face gender and race could be decoded from neural response patterns in a social categorization task, but after controlling for low-level differences in the stimuli from each category, the only region that showed accurate decoding was the FFA, further distinguishing the importance of this region in representing faces at the level of social categories. In a study that aimed to decode multi-voxel patterns associated with the broadest social group distinction possible (i.e., “us” vs. “them”), classification was used to show that the dorsal anterior cingulate cortex/middle cingulate cortex (dACC/MCC) and anterior insula (AI) contain high-level information about group boundaries ranging from arbitrarily-defined “minimal” groups to political groups (Cikara et al., 2017).

Reconciling and integrating findings from univariate and multivariate fMRI is an important ongoing task across all fields that use neuroimaging (Davis et al., 2014). RSA has recently proven quite useful for these purposes, since it can not only help reveal which brain regions represent social category-relevant information, but also the nature and organization of

those representations. In social psychology, social categorization has traditionally been considered an automatic and obligatory perceptual process that precedes any more "cognitive" social processes such as stereotyping (Allport, 1954). However, alternative approaches emphasize the idea that social-conceptual knowledge (i.e., stereotypical associations) and other top-down social cognitive processes can weigh in on face processing before a percept has solidified, thereby allowing implicit stereotypes to shape face perception (Freeman & Ambady, 2011; Freeman & Johnson, 2016). Recent work applying RSA has examined whether an individual's stereotypical associations are reflected in how the brain represents others' faces, an approach which can also reveal how "deeply" such top-down associations reach (i.e., regions that would suggest perceptual vs. post-perceptual processing of faces). Stolier and Freeman (2016) found that neural representations of faces' social categories (e.g., Black, Female, Happy) in the FG and orbitofrontal cortex (OFC) demonstrated a similarity structure that was predicted by a subject's own unique stereotype knowledge about those categories. That is, if someone held more similar stereotypes about the categories 'Male' and 'Anger', the multi-voxel patterns associated with those categories exhibited a greater similarity when subjects passively viewed faces belonging to those categories (Figure 1).

These findings suggest that the way face category representations are organized in regions important for face perception, such as the FG, is partially determined by stereotypes about those categories. Moreover, the fact that this similarity structure was also observed in the OFC suggests that domain-general perceptual processes associated with the OFC may be involved in driving the impact of stereotypes on face perception (Freeman & Johnson, 2016). In particular, the object recognition literature suggests that the OFC is recruited in perceptual categorization tasks when incoming visual input matches a pre-existing visual association or heuristic in memory (Bar, 2003; Bar et al., 2006; Summerfield & Egner, 2009). This effect is strengthened when the visual input is ambiguous or impoverished, suggesting that the OFC is involved in exerting visual predictions about category membership before those categorizations have fully solidified. One possibility is that the OFC is involved in a similar predictive capacity in perceiving social categories, supplying the FG with top-down visual predictions or expectations about social categories. In this case, the use of RSA permits inferences about the way social category representations are organized in visual processing regions as well as the high-level regions that may be involved in providing top-down social information to them.

It is sensible that visual face processing regions would partially depend on such rapid top-down input, as there are a number of challenges faced in social categorization, particularly when faces have atypical or ambiguous features. Any given facial feature can be more or less related to any set of social categories at once, and faces naturally vary on featural continua related to gender, race, and a host of other category dimensions. For example, a White face may not only have more or less White category-associated cues but may bear partial cues related to the Black category as well (Locke et al., 2005). Previous work suggests such multiple cues co-activate multiple category representations regardless of the ultimate categorization (Freeman et al, 2008; Freeman et al., 2010). MVPA has recently been used to investigate how the perceptual system is guided towards a final categorization despite features that may initially activate multiple social categories. Based on prior computational models of social categorization (Freeman & Ambady, 2011), Stolier and Freeman (2017) tested the possibility that initially the perceptual system co-activates any categories associated with features on a face, which then must compete and resolve over hundreds of milliseconds. They additionally examined whether cognitive monitoring processes may be recruited help resolve the competition, either to flag more attentional resources to be directed to the stimulus or perhaps to play an inhibitory role in the competition. For example, a feminine male face may initially elicit simultaneous partial activation of the categories Male and Female, then cognitive monitoring processes may help resolve the competition such that one category ('Male') wins out and the other category is cleared from processing ('Female'). A likely candidate for such processes would be the pre-supplementary motor area/dorsal anterior cingulate cortex (pre-SMA/dACC), a region central to cognitive monitoring and competition between decisions in tasks (Dosenbach et al., 2007).

In the study, subjects were presented with faces manipulated to vary in the typicality of their gender or race features, such that one category (e.g., Male or White) could have features more or less related to the alternate category (e.g., Female or Black). To measure perceivers' co-activation of multiple categories while viewing each face, participants performed a mouse-tracking task in the scanner, in which they made speeded face categorization decisions with a computer mouse while the trajectories of their mouse movements were recorded. The deviation of mouse trajectories toward unselected categories has been well-validated as a measure of multiple category co-activation during perception (Freeman, 2018; Freeman & Ambady, 2010). For instance, a mouse trajectory that deviated towards the 'Male' response option en route to a

categorization of 'Female' putatively reflects co-activation of the Male category despite a final categorization of 'Female'. To measure how brain regions involved in face perception held information pertaining to both competing categories during the mouse-tracking task, the FG multivariate response pattern to each individual trial's face (e.g., 'Male') for one category was compared to the mean response pattern to the alternate category (e.g., 'Female'). The results showed that, during trials in which mouse trajectories showed greater category co-activation (i.e., deviation toward the alternate category), multivariate response patterns in the FG for those trials showed a greater similarity to the mean response pattern for the alternate category. For instance, when subjects steered the mouse towards the 'Female' response option en route to 'Male' for a given face, the FG multi-voxel pattern in response to that face was more similar to the average multi-voxel response pattern for 'Female'. To explore how cognitive monitoring may assist perceivers in converging on their ultimate percept, the researchers also examined univariate activation in the pre-SMA/dACC. They found that, on trials where mouse-tracking showed more category co-activation (and overlapping neural response patterns), the pre-SMA/dACC became additionally engaged, suggesting a recruitment of conflict resolution processes to help the FG converge on a stable percept of a face. These findings suggest that other people's complex and sometimes ambiguous facial cues lead the FG to temporarily co-represent multiple categories, which through the help of the pre-SMA/dACC rapidly resolve over time to drive the stable categorization of other people.

Traits, identity, and person knowledge

Perhaps more than in any other domain of social neuroscience, enormous effort has been invested in determining which brain regions are most involved in thinking about other minds (i.e., mentalizing and theory of mind). This research has identified a set of brain regions that are reliably engaged by storing, retrieving, and updating knowledge about others, such as the medial prefrontal cortex (MPFC), temporo-parietal junction (TPJ), precuneus, and posterior cingulate cortex (PCC). The MPFC has been a particular focus in social neuroscience for its frequent associations with inferences about other minds, both in terms of fairly stable qualities like personality traits and more fleeting or situational mental states.

Our ability to recognize and tailor our behavior toward numerous personally familiar individuals would be computationally intractable without an extremely efficient coding scheme

in brain regions associated with thinking about others. Univariate fMRI approaches have shown that individual identities can activate associated person knowledge in regions such as the MPFC (Cloutier et al., 2011; Todorov et al., 2007), but that does not necessarily mean that the brain stores individual representations for each known identity. Recently, multivariate approaches have extended this work by demonstrating the complexity and flexibility of these identity representations. MVPA and RSA have been particularly well-suited to studying representations of personality and identity due to the multitude of theories in social psychology about the dimensions underlying our representations of others. In particular, RSA allows explicit tests of how well such dimensional theories of personality predict the brain's representational structure during social perception or mentalizing tasks (e.g., Tamir et al., 2016; Thornton & Mitchell, 2018). Due to the highly consistent set of regions associated with high-level social cognition in the univariate fMRI literature (i.e., MPFC, TPJ, PCC/precuneus; see Figure 2), MVPA investigations in this domain have had a well-defined set of a priori regions of interest to explore. Research using classification and RSA approaches has begun targeted assessments of the representational structure in these regions, revealing the granularity or discriminability of identity or personality representations, which category boundaries seem to be relevant for processing in these regions, and which aspects of social cognition each region is most computationally relevant for.

For example, research using a searchlight-based classification approach showed that multi-voxel response patterns in the MPFC, TPJ, and precuneus could discriminate between personally familiar and unfamiliar individuals, and further, that neural patterns in the MPFC could distinguish individual identities from both the familiar and unfamiliar conditions, potentially reflecting rapid encoding of coarse personality representations of the unfamiliar identities during the course of the experiment (Castello et al., 2017). Research introducing biographical information about novel targets found such information rapidly shapes multivoxel patterns in response to their faces in the bilateral FG, where face representations were grouped by the amount of biographical information participants learned about each individual (Verosky, Todorov, & Turke-Browne, 2013). Thornton and Mitchell (2017) similarly found that identity and person knowledge about personally familiar others are represented in discriminable patterns of neural activity in regions such as the MPFC, TPJ, and precuneus, but the task involved imagining these individuals in various contexts rather than looking at their faces. Subjects in this

experiment also made judgments about how accurate and vivid they felt their mental simulations were, and these judgments were predicted by how typical (for that target individual) the neural response pattern in the medial parietal cortex/precuneus was on the relevant trial. Moreover, the medial parietal cortex/precuneus was the only region that reliably encoded information about the situational context of the mental simulation in addition to information about the individual imagined in the simulation. While speculative, this possibly indicates the different computational roles of these regions, such that the MPFC stores and organizes person knowledge, but medial parietal regions are more involved in integrating person knowledge into social-cognitive judgments and contextual aspects of mental prospection.

But what are the nature of these representations of other minds, and how are they organized? While a comprehensive answer to this question remains elusive, recent work has used RSA to characterize the organization of person knowledge, indicating a number of complex yet efficient ways for organizing knowledge about numerous individuals along shared dimensions. Building off the extensive history of dimensional models in the person perception literature (e.g., Fiske et al., 2002; 2007; Oosterhof & Todorov, 2008), Tamir and colleagues (2016) tested how well a large set of hypothesized social-cognitive dimensions from the literature explained the representational space of multi-voxel patterns elicited when subjects thought abstractly about mental states such as “awe” and “self-consciousness”. Using principal components analysis, the researchers found that a smaller set of three dimensions explained a majority of the variance in mental state representations: rationality, social impact, and valence. Complementary work used RSA to test how well four specific models of person perception (including the big five factor model of personality traits and warmth-competence model of social cognition) predicted the neural representational space of response patterns during mentalizing (Thornton & Mitchell, 2018). The researchers found that all the four models significantly predicted the brain’s representational structure, but not as well as a synthetic model derived from a combination of all four candidate models. Thus, while there may be a low-dimensional space that largely explains how we represent information about others, ultimately our understanding of the nature of these representations is only emerging. Related work recently found neural representations of others hold information pertaining to the mental states most attributed to each individual, with those mental states weighted more strongly in a given target’s multi-voxel response patterns (Thornton et al., 2019a; see Figure 2).

One particularly informative approach for mental state and trait representation researchers has been to compare representations of self and other, and of many familiar others, within naturalistic groups of friends or acquaintances. Recently, Thornton and colleagues (2019) examined the relationship between a given individual's representation of their own self-knowledge with the neural representations of others' mental states. The researchers used RSA to show that the neural patterns associated with one's own self-concept are more distinct or discriminable than those associated with knowledge about others, as reflected in regions such as MPFC, TPJ, and dorsolateral prefrontal cortex (DLPFC). Other work employed a round-robin design in an fMRI study on a sample of close friends, finding that multi-voxel representations of a given individual's self-concept in the MPFC were correlated with their friend's MPFC representations of their personality (Chavez & Wagner, 2020). The strength of this relationship was associated with how similar trait ratings were between the target individual (rating their own traits) and the friend in question, suggesting that one factor driving "accuracy" in personality judgments is how well our mental models of an individual's personality match their own self-assessments.

Another fMRI study on a group of familiar individuals used a large sample drawn from a cohort of first-year MBA students (Parkinson et al., 2017), using RSA to show that representations of familiar others are organized along social network dimensions. Multiple social network attributes – including distance (number of intermediate paths between any given two people in the network) and eigenvector centrality (generally characterizing how well connected a given person is within their network) – were used to construct dissimilarity matrices reflecting pairwise relationships between each individual in the social network on these metrics. These network characteristics were found to predict the representational structure of neural response patterns in regions such as the MPFC, superior temporal cortex (STC), inferior parietal lobule (IPL), and precuneus/PCC when subjects viewed videos of their classmates introducing themselves in the scanner. This work suggests that representations of familiar others are organized in a manner partly determined by the overall structure of the social network in which an individual is embedded.

The brain's representational structure of identity and personality concepts is consistent with a number of different models of high-level social cognition that predict how different minds relate to one another. Together, these studies demonstrate that person representations (identity,

traits, personality) in the brain are highly structured, with high-level social dimensions such as personality similarity and social network position partly organizing neural response patterns. It is likely that a smaller set of underlying social cognitive dimensions explain most of the variance in pattern response in brain regions such as MPFC and TPJ, rather than representations in these regions being simultaneously constrained by several different psychological models of personality and behavior. However, research investigating any such low-dimensional representational space is only emerging.

Emotion Perception

Understanding how others are feeling is an important tool for navigating the social world safely and effectively, and understanding which emotion categories perceivers can successfully "recognize" in others has been a central focus of the literature on emotion perception. Early, seminal models of face perception (Bruce & Young, 1986) emphasized a processing dissociation between static and dynamic facial cues, and since facial emotion is often categorized based on dynamic facial movements, it has largely been treated separately in the literature from dimensions of social perception that have categorical boundaries defined by static cues (e.g., race and sex). The particular neuroanatomical dissociation is between the fusiform gyrus (FG), thought to be more important in processing configurations of static facial cues, and the superior temporal sulcus (STS), known to be involved in processing dynamic facial actions as well as socially relevant actions more broadly, such as body movements (Haxby et al., 2000; 2002). The neuroimaging literature on emotion perception has been further complicated by studies that aimed to isolate regions most involved in perception of specific discrete emotions, for example by showing participants facial expressions typically categorized as Angry, Disgusted, and Neutral, and computing univariate *Angry* > *Neutral* and *Disgusted* > *Neutral* contrasts to determine which regions are preferentially engaged by Angry and Disgusted facial expressions, respectively. Numerous studies associated different possible brain regions with different specific emotions, most famously strongly associating the amygdala with perceiving fear (Adolphs et al., 1995; Adolphs, 2008), although neuroimaging meta-analyses have been unable to find specific associations that are consistent across the literature (Lindquist et al., 2012).

In an analogous manner to the fMRI literature on social categorization, multivariate approaches have helped shift the neuroscientific study of emotion perception from a focus on

associating brain regions with emotion categories to a greater understanding of the relevant information different regions process. As with the study of traits and identity, this makes MVPA particularly beneficial to the study of emotion perception because of the degree of debate in the field surrounding different candidate models of emotion perception. The core of this debate concerns how to specify the relationship between emotion categories (e.g., Anger) and facial actions, with some assuming that specific facial expressions map directly onto corresponding emotion categories and others assuming a more context- and perceiver-dependent relationship (Barrett et al., 2019). Not only can diverse candidate models be directly tested against each other using RSA, classification and decoding approaches can demonstrate which aspects of emotion expressions determine boundaries between emotion categories and perceiver impressions, addressing theoretical predictions about the determinants of emotion percepts.

Multivariate classification approaches demonstrate that facial emotion categories can be decoded in regions such as early visual cortex (V1; Petro et al., 2013), the posterior superior temporal sulcus (pSTS; Said et al. 2010), and the fusiform gyrus (FG; Harry et al., 2013; Wegrzyn et al., 2015a). For static images of facial expressions, Wegrzyn et al. (2015a) compared classification performance in multiple brain regions, finding that classification performance was highest in the FG, indicating that this region contains and processes emotion category-relevant information. This is consistent with work showing that facial emotion categories exhibit category selectivity effects in a manner consistent with social category perceptions (Calder et al., 1996; de Gelder et al., 1997; Etcoff et al., 1992; Wegrzyn et al., 2015b). A recent study further challenged the common FG/STS distinction by showing that emotion category representations in response to dynamic facial movements could be decoded in the FG, and that emotion category representations in response to static facial expressions could be decoded in the STS (Liang et al., 2017). However, in studies characterizing the representation of fine-grained facial movements associated with emotion, the pSTS seems to be the most computationally relevant region (Deen & Saxe, 2019; Srinivasan et al., 2016).

As in the domain of perceiving social categories and groups, the idea that social cognitive processes and semantic representations of emotion categories may shape face processing has become important in understanding facial emotion perception. Classic categorical approaches such as basic emotion theory would predict a neural representational space in which different categories are primarily represented on the basis of their differences in facial cues. In a recent

study using RSA, we measured the neural representational space of emotion categories from a task in which subjects passively viewed emotional facial expressions, and also measured each subject's conceptual space of emotion categories (Brooks et al. 2019). We found that each individual's unique conceptual model of emotion categories was reflected in multi-voxel pattern structure in the FG when viewing faces belonging to those categories, even when acknowledging intrinsic visual differences in the emotion expressions themselves. Specifically, when an individual believed any given pair of emotions (e.g., Angry and Sad) were more conceptually similar, there was a corresponding similarity in the multi-voxel response patterns in the FG to faces belonging to those categories (see Figure 1c). Of relevance for comparing conceptually-shaped models of emotion perception with more categorical models, these results emerged even when controlling for three different models of similarity between categories in their displayed facial cues and low-level visual properties, strengthening the claim of conceptually-structured representations in the FG. These findings suggest that individual differences in conceptual understanding of what different emotions mean may impact the visual representation of facial expressions commonly associated with those categories.

Beyond using RSA to assess the correspondence between idiosyncratic conceptual structure and neural pattern structure in FG, RSA has also proven useful in adjudicating between different theoretical models of emotion in terms of how well they predict the brain's representational structure of emotion categories. In the domain of emotion perception, the dominant models have been the categorical "basic emotion" model and a dimensional "circumplex" model. Basic emotion theory posits a small set of psychologically distinct emotion categories thought to yield associated facial expressions that are universally recognized in a categorical fashion (Ekman, 1993; Ekman et al., 1969; Ekman & Friesen, 1971; 1976). The circumplex model emphasizes two underlying dimensions of valence (positive vs. negative) and arousal (high vs. low physiological activation or perceived intensity) that all emotion judgments map onto (Russell, 1980; 2003).

In a study using RSA to explicitly compare how well these models fit representations of 20 emotion categories inferred from reading stories about individuals, Skerry and Saxe (2015) found that neither model fit multi-voxel representations in the "theory of mind network" (including multiple sub-regions of MPFC as well as the rTPJ) as well as a higher-dimensional model generated from behavioral responses. This model contained 38 dimensions describing

various contextual and situational factors that constrain perceiver appraisals. While these brain regions, more often associated with mentalizing and theory of mind, have not been extensively studied in the context of emotion perception, it is of course true that emotion perception is fundamentally a case of mental state attribution, as with mentalizing. Indeed, an additional study (Skerry & Saxe, 2014) showed that the representational space of emotion categories in the MPFC was shared across situational inferences and facial emotion perception, indicating that these regions implicated in theory of mind may represent and compute high-level aspects of emotion perception across multiple modalities. Given that valence has also been found to partly organize the brain's mental state representations (Tamir et al., 2016), future work is needed to disentangle the representations and possibly shared underlying dimensions of emotion and mental state representations.

Conclusions

Across multiple domains of social perception, multivariate analyses of fMRI data have permitted a more fine-grained understanding of how social information is processed and represented in the brain, and an increased understanding of the computational relevance of specific brain regions in social-perceptual processes. Such an approach is equally informative from neuroscientific and psychological perspectives, better characterizing the computational characteristics of specific brain regions as well as allowing explicit tests of how well psychological models fit the brain's representational structure of social information. It is important to note that this review was relatively limited to classification and RSA approaches, but there are a number of other data-driven methods and advanced analysis techniques that have been similarly useful in moving the field past localization approaches (see Wagner et al., 2019 for a relevant review). Additional methodological tools are rapidly accumulating and gaining sophistication. For example, some recent work has expanded the use of multivariate pattern classifiers to include multivariate patterns of connectivity between brain regions, which can also be leveraged to make stronger causal claims about the relationships between distant brain regions in their representational spaces (Anzellotti et al., 2017). In the domain of social perception, this has already been used to decode wide-scale patterns of connectivity associated with facial emotion categories (Liang et al., 2018). Future work is needed to better understand the relationship between these various approaches.

While promising, these approaches are not without limitations. Notably, RSA is inherently correlational and does not permit any causal inferences about determinants of neural representational structure. Additionally, decoding approaches are sometimes particularly sensitive to idiosyncratic factors such as the specific sample, stimuli, and task context (Davis et al., 2014; Todd et al., 2013). For this reason, it is essential to test decoding models outside of the original sample, and recent applications have made progress incorporating this level of rigor. Large-scale openly shared fMRI datasets, and concurrent development of sophisticated tools to analyze them, also serve to mitigate these issues. It is also important to note that MVPA analyses are not immune to several issues faced by univariate fMRI. In particular, following a searchlight or whole-brain MVPA analysis that maps representational similarity or classification accuracy at each voxel in the brain, the resulting maps undergo voxelwise statistical tests, making MVPA vulnerable to the same statistical correction and reporting pitfalls observed in univariate fMRI research.

Another challenge of multivariate fMRI lies in interpreting MVPA results and using them to inform and develop psychological theory. What does it really mean to be able to discriminate patterns of fMRI signal along a cognitive dimension? Multi-voxel patterns associated with a particular category are often reported as the neural representation of that particular category, but this interpretation requires care. The precise relationship between neuronal activity and differences in spatial distribution of fMRI signal is unknown (but see Haynes, 2015 for a summary of existing knowledge about the relationship between multi-voxel patterns and the activity of neurons). However, empirical and theoretical efforts are underway to better characterize the nature of multi-voxel response patterns and their precise relationship with underlying neural activity as well as the cognitive dimensions they seem to encode (Davis et al., 2014; Grootswaters et al., 2018; Popov et al., 2018; Ritchie et al., 2019). The idiosyncrasies of sample and stimuli can also introduce issues of interpretation. Researchers should take caution when interpreting a classifier trained and tested on one sample as reflecting the "neural code" for that cognitive dimension. As noted above, the interpretability and theoretical utility of decoding models principally depends on their generalizability out of sample (Kriegeskorte & Douglas, 2019). Increased use of RSA and related methods may elucidate some of these open questions about the nature of representation in fMRI response patterns. In particular, while RSA does not permit causal inferences, its ability to test the dimensionality and informational content of neural

activation patterns in a targeted way affords stronger conclusions about how a given brain region encodes information.

While more work is needed to address these and other questions, the field stands to benefit from continued use of MVPA and related approaches. Techniques are constantly evolving and improving, building an exciting set of new avenues for computational neuroimaging research on social perception, such as MVPA in conjunction with naturalistic stimuli and tasks such as movies and real-time interactions (Wagner et al., 2019). Future work using RSA would benefit from choosing diverse candidate models to test at the level of neural representation. Of particular relevance to social perception are the astoundingly rapid advances in "deep" neural network models of computer vision, which approach or approximate human performance in a number of object recognition and categorization tasks (Goodfellow et al., 2016; Hassabis et al., 2017; Kriegeskorte & Golan, 2019). Comparing internal representations from advanced computer vision models to neural and behavioral category representations has already proven useful in the object recognition literature (Jozwik et al., 2017; Kriegeskorte & Golan, 2019; Storrs et al., 2017), and could benefit the neuroscientific study of visual social perception as well. These approaches, and other theoretical and methodological advancements, show promise in improving our understanding of how visual input from another person's face and body is transformed into a socially relevant category representation in the brain.

References

Amodio, D. M., & Frith, C. D. (2006). Meeting of minds: the medial frontal cortex and social cognition. *Nature Reviews Neuroscience*, *7*(4), 268–277.

Anzellotti, S., Caramazza, A., & Saxe, R. (2017). Multivariate pattern dependence. *PLOS Computational Biology*, *13*(11), e1005799.

Averbeck, B., Latham, P. & Pouget, A. (2006). Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, *7*, 358–366.

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object recognition. *Journal of Cognitive Neuroscience*, *15*(4), 600-609.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*(2), 449-454.

Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., & Pollak, S. D. (2019). Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, *20*, 1–68.

Brooks, J.A., Chikazoe, J., Sadato, N., & Freeman, J.B. (2019). The neural representation of facial emotion categories reflects conceptual structure. *Proceedings of the National Academy of Sciences*.

Bruce, V., & Young, A. W. (1986). A theoretical perspective for understanding face recognition. *British Journal of Psychology*, *77*, 305–327.

Calder, A.J., Young, A.W., Perrett, D.I., Etcoff, N.L., & Rowland, D. (1996). Categorical perception of morphed facial expressions. *Visual Cognition*, *3*, 81–118.

Chaimow, D., Yacoub, E., Ugurbil, K., & Shmuel, A. (2011). Modeling and analysis of mechanisms underlying fMRI-based decoding of information conveyed in cortical columns. *NeuroImage*, *56*(2), 627–642.

Chavez, R.S. & Wagner, D.D. (2020). The neural representation of self is recapitulated in the brains of friends: A round-robin fMRI study. *Journal of Personality and Social Psychology*, *118*(3), 407-416.

Cheong, J.H., Jolly, E., Sul, S., & Chang, L.J. (2017). Computational models in social neuroscience. In *Computational Models of Brain and Behavior*, Moustafa, A. (Ed.), Wiley-Blackwell.

Cikara, M., Van Bavel, J. J., Ingbretsen, Z., & Lau, T. (2017). Decoding "us" and "them:" Neural representations of generalized group concepts. *Journal of Experimental Psychology: General, 146*, 621-631.

Cloutier, J., Kelley, W. M., & Heatherton, T. F. (2011). The influence of perceptual and knowledge-based familiarity on the neural substrates of face perception. *Social Neuroscience*, *6*(1), 63–75.

Contreras, J. M., Banaji, M. R., & Mitchell, J. P. (2013). Multivoxel patterns in fusiform face area differentiate faces by sex and race. *PLoS ONE*, *8*(7), e69684.

Davis, T., LaRocque, K. F., Mumford, J. A., Norman, K. A., Wagner, A. D., & Poldrack, R. A. (2014). What do differences between multi-voxel and univariate analysis mean? How subject-, voxel-, and trial-level variance impact fMRI analysis. *NeuroImage*, *97*, 271-283.

Deen, B., & Saxe R. (2019). Parts-based representations of perceived face movements in the superior temporal sulcus. *Human Brain Mapping*.

De Gelder, B., Teunisse, J.-P., & Benson, P.J. (1997). Categorical perception of facial expressions: Categories and their internal structure. *Cognition and Emotion*, *11*, 1–23.

Dobs, K., Isik, L., Pantazis, D., & Kanwisher, N. (2019). How face perception unfolds over time. *Nature Communications*, *10*(1), 1258.

Dosenbach, N.U., Fair, D.A., Miezin, F.M., Cohen, A.L., Wenger, K.K., Dosenbach, R.A., Fox, M.D., Snyder, A.Z., Vincent, J.L., Raichle, M.E., Schlaggar, B.L., & Petersen, S.E. (2007). Distinct brain networks for adaptive and stable task control in humans. *Proceedings of the National Academy of Sciences*, *104*, 11073–11078.

Ekman, P. (1993). Facial expression of emotion. *American psychologist*, *48*, 384-392.

Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, *17*, 124–129.

Ekman, P., & Friesen, W. V. (1976). *Pictures of Facial Affect*. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Sorenson, E. R., & Friesen, W. V. (1969). Pan-cultural elements in facial displays of emotions. *Science*, *164*, 86–88.

Etcoff, N.L., & Magee, J.J. (1992). Categorical perception of facial expressions. *Cognition*, *44*, 227–240.

Fiske, S. T., Cuddy, A. J., & Glick, P. (2007). Universal dimensions of social cognition: Warmth and competence. *Trends in Cognitive Sciences*, *11*(2), 77-83.

Fiske, S. T., Cuddy, A. J., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, *82*(6), 878-902.

Freeman, J.B. (2018). Doing psychological science by hand. *Current Directions in Psychological Science*, *27*(5), 315-323.

Freeman, J.B. & Ambady, N. (2010). Mousetracker: Software for Studying Real-Time Mental Processing Using a Computer Mouse-Tracking Method. *Behavior Research Methods*, *42*(1) , 226-241.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*, 247-279.

Freeman, J.B. & Johnson, K.L. (2016). More than meets the eye: Split-second social perception. *Trends in Cognitive Sciences*, *20*, 362-374.

Golby, A.J., Gabrieli, J.D.E., Chiao, J.Y., & Eberhardt, J.L. (2001). Differential responses in the fusiform region to same-race and other-race faces. *Nature Neuroscience*, *4*, 845–850.

Gonzalez, B., & Chang, L. J. (2019, July 26). Computational models of mentalizing. https://doi.org/10.31234/osf.io/4tyd9.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.

Grootswaters, T., Cichy, R.M., & Carlson, T.A. (2018). Finding decodable information that can be read out in behaviour. *Neuroimage*, *179*, 252-262.

Hackel, L.M., & Amodio, D.M. (2018). Computational neuroscience approaches to social cognition. *Current Opinion in Psychology*, *24*, 92-97.

Harry, B., Williams, M.A., Davis, C., & Kim, J. (2013). Emotional expressions evoke a differential response in the fusiform face area. *Frontiers in Human Neuroscience*, *7*, 692.

Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-inspired artificial intelligence. *Neuron*, *95*, 245-258.

Haxby, J. V., Connolly, A. C., Guntupalli, J. S. (2014). Decoding neural representational spaces using multivari-ate pattern analysis. *Annual Review of Neuroscience*, *37*, 435–456.

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.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. A. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*, 223–233.

Haxby, J. V., Hoffman, E. A., & Gobbini, M. A. (2002). Human neural systems for face recognition and social communication. *Biological Psychiatry*, *51*, 59–67.

Haynes, J.-D. (2015). A primer on pattern-based approaches to fMRI: principles, pitfalls, and perspectives. *Neuron*, *87*, 257-270.

Hugenberg, K., Young, S. G., Bernstein, M. J., & Sacco, D. F. (2010). The categorization-individuation model: an integrative account of the other-race recognition deficit. *Psychological Review*, *117*(4), 1168–1187.

Józwik, K. M., Kriegeskorte, N., Storrs, K.R., & Mur, M. (2017). Deep convolutional neural networks outperform feature-based but not categorical models in explaining object similarity judgements. *Frontiers in Psychology*, *8*, 1726.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PloS Computational Biology*, *10*, e1003915.

Kriegeskorte, N., & Douglas, P.K. (2019). Interpreting encoding and decoding models. *Current Opinion in Neurobiology*, *55*, 167-179.

Kriegeskorte, N., & Golan, T. (2019). Neural network models and deep learning. *Current Biology*, 29(7), PR231-R236.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4.

Lewis-Peacock, J. A., & Norman, K. A. (2014). Multi-voxel pattern analysis of fMRI data. In M. S. Gazzaniga & G. R. Mangun (Eds.), *The Cognitive Neurosciences (5th ed.)*. MIT Press.

Liang, Y. , Liu, B. , Xu, J. , Zhang, G. , Li, X. , Wang, P. & Wang, B. (2017), Decoding facial expressions based on face-selective and motion-sensitive areas. *Human Brain Mapping*, *38*, 3113-3125.

Liang, Y., Liu, B., Li, X., & Wang, P. (2018). Multivariate pattern classification of facial expressions based on large-scale functional connectivity. *Frontiers in Human Neuroscience*, *12*, 94.

Lieberman, M.D., Hariri, A., Jarcho, J.M., Eisenberger, N.I., & Bookheimer, S.Y. (2005). An fMRI investigation of race-related amygdala activity in African-American and Caucasian-American individuals. *Nature Neuroscience*, *8*, 720–722.

Lindquist, K.A., Wager, T.D., Kober, H., Bliss-Moreau, E., & Barrett, L.F. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences, 35,* 121-143.

Locke, V., Macrae, C.N., Eaton, J.L. (2005). Is person categorization modulated by exemplar typicality? *Social Cognition*, *23*, 417–428.

Logothetis, N.K. (2008). What we can do and what we cannot do with fMRI. *Nature*, *453*, 869-878.

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: a meta-analytic review. *Psychology, Public Policy, and Law*, *7*, 3–35.

Mitchell, J. P. (2008). Contributions of functional neuroimaging to the study of social cognition. *Current Directions in Psychological Science*, *17*(2), 142–146.

Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*, 11087-11092.

Parkinson, C., Kleinbaum, A.M., & Wheatley, T. (2017). Spontaneous neural encoding of social network position. *Nature Human Behaviour*, *1*, 72.

Petro, L.S., Smith, F.W., Schyns, P.G., & Muckli, L. (2013). Decoding face categories in diagnostic subregions of primary visual cortex. *European Journal of Neuroscience*, *37*, 1130–1139.

Popal, H., Wang, Y., & Olson, I.R. (2019). A guide to representational similarity analysis for social neuroscience. *Social Cognitive and Affective Neuroscience*, *14*(11), 1243-1253.

Popov, V., Ostarek, M., & Tenison, C. (2018). Practices and pitfalls in inferring neural representations. *Neuroimage*, *174*, 340-351.

Ritchie, J.B., Kaplan, D.M., & Klein, C. (2019). Decoding the brain: neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*, *70*, 581-607.

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, *39*(6), 1161.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, *110*, 145–172.

Said, C.P., Moore, C.D., Engell, A.D., Todorov, A., & Haxby, J.V. (2010). Distributed representations of dynamic facial expressions in the superior temporal sulcus. *Journal of Vision*, *10*, 11.

Saxe, R., & Kanwisher N. (2003). People thinking about thinking people: The role of the temporo-parietal junction in "theory of mind". *NeuroImage*, *19*(4), 1835 - 1842.

Srinivasan, R., Golomb, J.D., & Martinez, A.M. (2016). A neural basis of facial action recognition in humans. *Journal of Neuroscience*, *36*(16), 4434-4442.

Stolier, R.M. & Freeman, J.B. (2016). Neural pattern similarity reveals the inherent intersection of social categories. *Nature Neuroscience*, *19*, 795-797.

Stolier, R. M. & Freeman, J. B. (2017). A neural mechanism of social categorization. *Journal of Neuroscience*, *37*(23), 5711-5721.

Storrs, K., Mehrer, J., Walter, A., & Kriegeskorte, N. (2017). Category-specialised neural networks best explain representations in category-selective visual areas. *Perception*, *46*, 1217–1218.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403- 409.

Tamir, D.I., Thornton, M.A., Contreras, J.M., & Mitchell, J.P. (2016). Neural evidence that three dimensions organize mental state representation: rationality, social impact, and valence. *Proceedings of the National Academy of Sciences*, *113*(1), 194-199.

Thornton, M. A., & Mitchell, J. P. (2017). Consistent neural activity patterns represent personally familiar people. *Journal of Cognitive Neuroscience, 29*(9), 1583-1594.

Thornton, M. A., & Mitchell, J. P. (2018). Theories of person perception predict patterns of neural activity during mentalizing. *Cerebral Cortex, 28*(10), 3505-3520.

Thornton, M. A., Weaverdyck, M. E., & Tamir, D. I. (2019a). The brain represents people as the mental states they habitually experience. *Nature Communications*, *10*, 2291.

Thornton, M. A., Weaverdyck, M. E., Mildner, J. N., & Tamir, D. I. (2019b). People represent their own mental states more distinctly than those of others. *Nature Communications*, *10*, 2117.

Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage*, *15*(77), 157-165.

Verosky, S. C., Todorov, A., & Turk-Browne, N. B. (2013). Representations of individuals in ventral temporal cortex defined by faces and biographies. *Neuropsychologia*, *51*(11), 2100-2108.

Visconti di Oleggio Castello, M., Halchenko, Y. O.,Guntupalli, J. S., Gors, J. D., Gobbini, M. I. (2017). The neural representation of personally familiar and unfamiliar faces in the distributed system for face perception. *Scientific Reports*, *7*(1), 12237.

Wagner, D. D., Chavez, R. S., & Broom, T. W. (2019). Decoding the neural representation of self and person knowledge with multivariate pattern analysis and data-driven approaches. *Wiley Interdisciplinary Reviews: Cognitive Science*, *10*(1), e1482.

M. Wegrzyn et al., Investigating the brain basis of facial expression perception using multi-voxel pattern analysis. Cortex 69, 131–140 (2015a).

Wegrzyn, M., Bruckhaus, I., & Kissler, J. (2015b). Categorical Perception of Fear and Anger Expressions in Whole, Masked and Composite Faces. *PloS one*, *10*(8), e0134790. doi:10.1371/journal.pone.0134790.

Figure Legends

Figure 1. MVPA shows that stereotypes and emotion concepts shape representations of other people's faces in the FG. Stolier and Freeman (2016) used multiple regression RSA on fMRI data from a task in which subjects viewed faces, showing that stereotypes partially structure how face's social categories are represented in regions important for face perception such as the FG. a) An example pair of corresponding dissimilarity matrices (DMs), depicting the corresponding representational structures of social categories in both stereotypes and subjective face perception. b) Results from a whole-brain searchlight analysis which performed multiple regression RSA at each searchlight sphere, measuring the correspondence between the subjective perceptual and neural DMs while controlling for three models of visual similarity. This analysis revealed that the right fusiform gyrus (rFG) and orbitofrontal cortex (OFC) represent social categories in a manner consistent with the influence of stereotypes on processing of faces' social categories. c) Similar results are shown from Brooks and colleagues (2019), which reported an fMRI study in which subjects passively viewed faces varying in emotion expression. The researchers also measured subjects' conceptual similarity between the emotion categories Anger, Disgust, Fear, Happiness, Sadness, and Surprise (corresponding to the facial expressions shown in the scanner). The correspondence between this idiosyncratic conceptual DM and the brain's representational structure (neural DM) was measured using multiple regression RSA in a whole-brain searchlight analysis, also controlling for three visual similarity models. This analysis revealed that the rFG represents facial emotion categories in a manner consistent with the influence of a perceiver's conceptual knowledge on processing of facial emotion. Figure adapted from Stolier and Freeman (2016) and Brooks and colleagues (2019).

Figure 2. MVPA sheds new light on how "social brain network" regions represent other people. The "social brain network", including regions such as the medial prefrontal cortex (MPFC), precuneus/posterior cingulate cortex (PCC), anterior temporal lobe (ATL), and temporoparietal junction (TPJ), has long been known to be involved in social cognition and person perception. MVPA and RSA have been important tools in recent progress made in understanding how this network of regions specifically represents and computes social information. In a study by Thornton and colleagues (2019a), RSA was used to show that these regions represent individual identities in a manner consistent with the sum of a person's mental state representations. Other studies have found that person and identity representations in these regions are structured by high-level social cognitive factors such as social network characteristics (Parkinson et al., 2017). In separate lines of work, MVPA and RSA have also proven helpful in disentangling the computational roles of these regions, e.g. suggesting that the MPFC is involved in representing information about individual people, while medial parietal regions such as the precuneus/PCC are more involved in representing information about the social context (Thornton & Mitchell, 2017). Figure adapted from Thornton and colleagues (2019a).

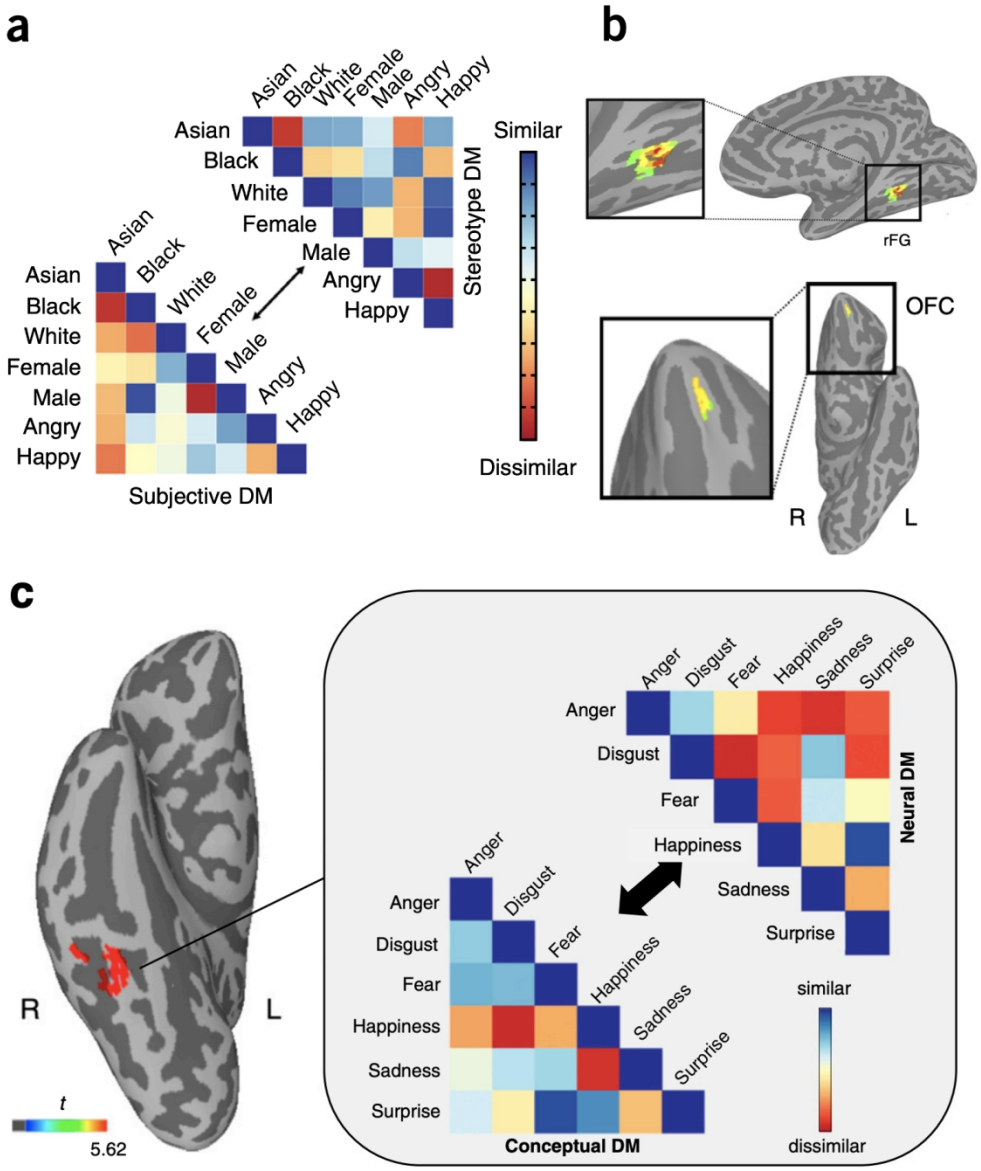

Figure 1. MVPA shows that stereotypes and emotion concepts shape representations of other people's faces in the FG. Stolier and Freeman (2016) used multiple regression RSA on fMRI data from a task in which subjects viewed faces, showing that stereotypes partially structure how face's social categories are represented in regions important for face perception such as the FG. a) An example pair of corresponding dissimilarity matrices (DMs), depicting the corresponding representational structures of social categories in both stereotypes and subjective face perception. b) Results from a whole-brain searchlight analysis which performed multiple regression RSA at each searchlight sphere, measuring the correspondence between the subjective perceptual and neural DMs while controlling for three models of visual similarity. This analysis revealed that the right fusiform gyrus (rFG) and orbitofrontal cortex (OFC) represent social categories in a manner consistent with the influence of stereotypes on processing of faces' social categories. c) Similar results are shown from Brooks and colleagues (2019), which reported an fMRI study in which subjects passively viewed faces varying in emotion expression. The researchers also measured subjects' conceptual similarity between the emotion categories Anger, Disgust, Fear, Happiness, Sadness, and Surprise (corresponding to the facial expressions shown in the scanner). The correspondence between this

idiosyncratic conceptual DM and the brain's representational structure (neural DM) was measured using multiple regression RSA in a whole-brain searchlight analysis, also controlling for three visual similarity models. This analysis revealed that the rFG represents facial emotion categories in a manner consistent with the influence of a perceiver's conceptual knowledge on processing of facial emotion. Figure adapted from Stolier and Freeman (2016) and Brooks and colleagues (2019).

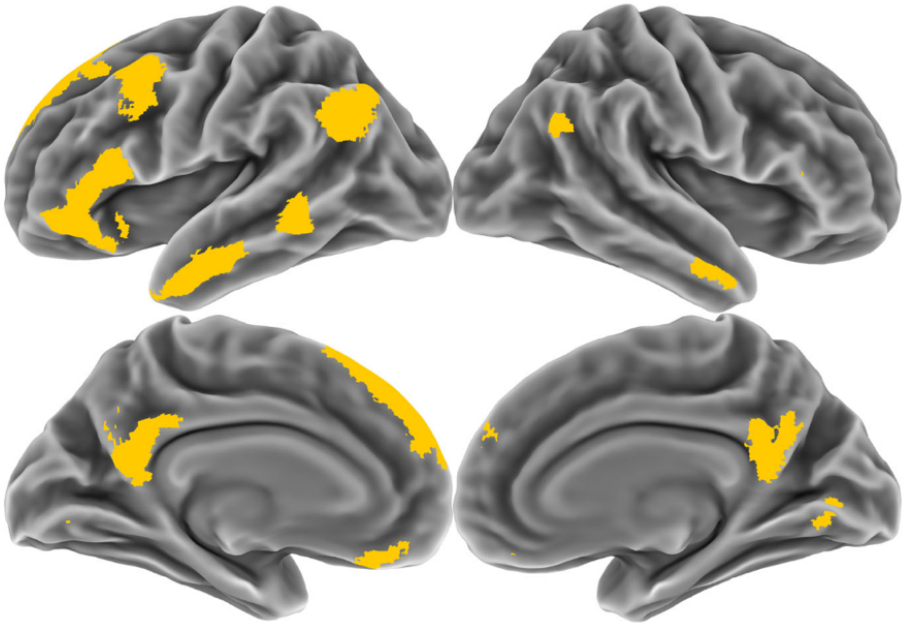

Figure 2. MVPA sheds new light on how "social brain network" regions represent other people. The "social brain network", including regions such as the medial prefrontal cortex (MPFC), precuneus/posterior cingulate cortex (PCC), anterior temporal lobe (ATL), and temporoparietal junction (TPJ), has long been known to be involved in social cognition and person perception. MVPA and RSA have been important tools in recent progress made in understanding how this network of regions specifically represents and computes social information. In a study by Thornton and colleagues (2019), RSA was used to show that these regions represent individual identities in a manner consistent with the sum of a person's mental state representations. Other studies have found that person and identity representations in these regions are structured by high-level social cognitive factors such as social network characteristics (Parkinson et al., 2017). In separate lines of work, MVPA and RSA have also proven helpful in disentangling the computational roles of these regions, e.g. suggesting that the MPFC is involved in representing information about individual people, while medial parietal regions such as the precuneus/PCC are more involved in representing information about the social context (Thornton & Mitchell, 2017). Figure adapted from Thornton and colleagues (2019).