Neuron
Review

Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging

Philip A. Kragel,[1,2] Leonie Koban,[1] Lisa Feldman Barrett,[3,4,5] and Tor D. Wager[1,*]
[1]Department of Psychology and Neuroscience and the Institute of Cognitive Science, University of Colorado, Boulder, CO, USA
[2]Institute for Behavioral Genetics, University of Colorado, Boulder, CO, USA
[3]Department of Psychology, Northeastern University, Boston, MA, USA
[4]Department of Radiology, Athinoula A. Martinos Center for Biomedical Imaging, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
[5]Department of Psychiatry, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA
*Correspondence: tor.wager@colorado.edu
https://doi.org/10.1016/j.neuron.2018.06.009

Human neuroimaging research has transitioned from mapping local effects to developing predictive models of mental events that integrate information distributed across multiple brain systems. Here we review work demonstrating how multivariate predictive models have been utilized to provide quantitative, falsifiable predictions; establish mappings between brain and mind with larger effects than traditional approaches; and help explain how the brain represents mental constructs and processes. Although there is increasing progress toward the first two of these goals, models are only beginning to address the latter objective. By explicitly identifying gaps in knowledge, research programs can move deliberately and programmatically toward the goal of identifying brain representations underlying mental states and processes.

Introduction

In recent years, human neuroimaging research has undergone a paradigm shift from brain mapping to developing integrated, *multivariate brain models* of mental events (see Appendix for definitions of key terms). Traditional brain mapping approaches analyze brain-mind associations within isolated brain regions or voxels, or a series of them tested one at a time. A local brain response is treated as the outcome to be explained by statistical models, and effects in local regions are aggregated into maps. Brain models reverse this equation: sensory experiences, mental events, and behavior are the outcomes to be explained. A model specifies how to combine brain measurements to yield a prediction about the identity or intensity of a *mental process* (Figure 1). For example, a model might predict (or decode) the category of the objects a person is viewing (Haxby et al., 2001; Issa et al., 2013), which action a person is about to perform (Haynes et al., 2007; Soon et al., 2008), or the intensity of pain experience (Marquand et al., 2010; Wager et al., 2013). Thus, brain maps and models have fundamentally different goals: whereas maps describe local encoding of information, models attempt to specify the parts of a neural system and how their joint activity predicts mind and behavior.

Some models are simple, associating activity in a single brain region with an outcome. But increasingly, brain models are multivariate: they explain outcomes as patterns of brain activity and/or structure across large numbers of brain features, often distributed across anatomical regions and systems, or even types of measurements (fMRI activity, connectivity, structure, and/or neurochemistry). Multivariate models have been developed for diverse types of mental events and states—including object recognition (Haxby et al., 2011), speech content (Formisano et al., 2008), wakefulness (Tagliazucchi and Laufs, 2014), autonomic responses (Eisenbarth et al., 2016), memory (Harrison and Tong, 2009; Polyn et al., 2005), decision making (Hampton and O'Doherty, 2007; Kahnt et al., 2011), semantic concepts (Huth et al., 2016; Mitchell et al., 2008), cognitive tasks (Poldrack et al., 2009), attention (Esterman et al., 2009; Rosenberg et al., 2016), pain (Marquand et al., 2010; Wager et al., 2013), prosody (Ethofer et al., 2009), emotion (Kragel and LaBar, 2015; Saarimäki et al., 2016; Wager et al., 2015), empathy (Krishnan et al., 2016), and the content of dreams (Horikawa et al., 2013). They have also been applied to diverse neurological and mental disorders (for reviews, see Arbabshirani et al., 2017; Woo et al., 2017a).

In this review, we discuss the theoretical underpinnings that make multivariate brain models an informative and powerful approach, and provide a brief history of the expanding set of modeling tools and approaches in the field. We also explore the promise and challenges of a specific type of model, brain "signatures" or "neuromarkers", which identify brain patterns that predict mental and behavioral outcomes across individuals (Gabrieli et al., 2015). In the application areas listed above and beyond, the reversal of predictors and outcomes affords several advantages: (1) a better match to how mental and behavioral information is encoded in neurons; (2) larger effect sizes in brain-outcome associations than standard local region-based approaches; (3) quantitative predictions about outcomes that can be empirically falsified; (4) models with defined measurement properties that can, under certain use cases, be tested and validated across studies and labs; and (5) a path toward validating *mental constructs* and understanding how the brain carves the mind at its joints—i.e., which psychological distinctions are paralleled by strong neurological ones (Lenartowicz et al., 2010).

Finally, we discuss the difficult issues surrounding mental constructs and their validation, and how predictive brain models can help redefine the way we categorize and understand the mind. The progression from brain maps to models of mental states provides a strong foundation for empirical and theoretical

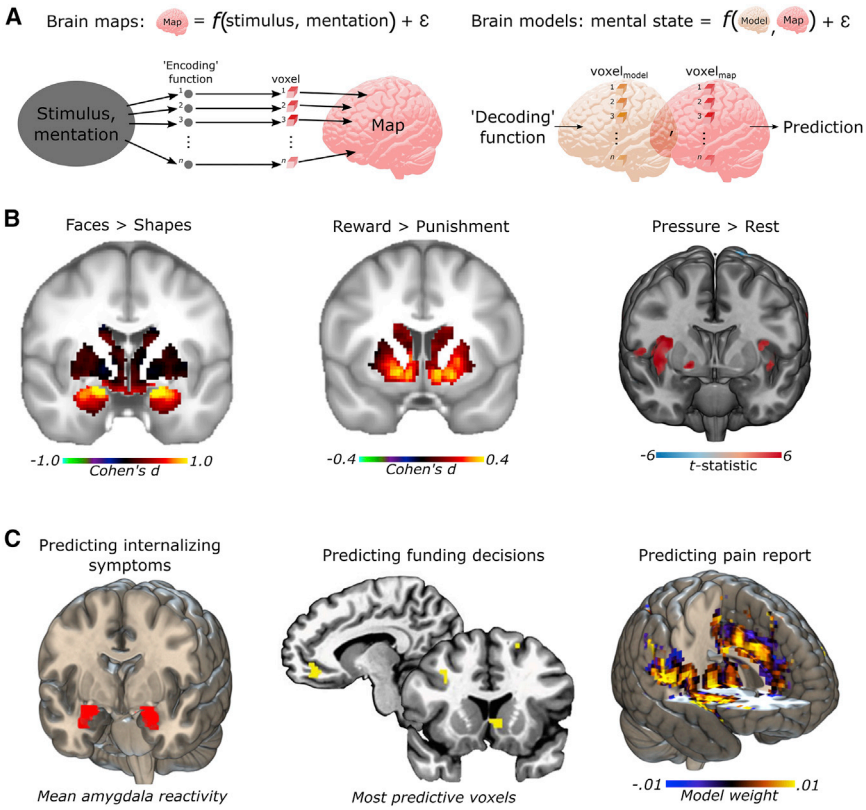

Figure 1. Brain Maps versus Brain Models

(A) The objective of conventional brain mapping is to identify which brain regions are *reliably* more active as a function of different kinds of stimulation or manipulations of *mental state* (in addition to error, E). The classical outcome of brain mapping study is a parametric map indicating the extent to which every brain measure (voxel) is associated with a given mental state. The objective of developing a multivariate brain model is to account for, and thus predict, an individual person's mental state or behavior (outcomes) based on their brain activity.

(B) Brain maps are displayed for comparisons of brain responses between emotional faces and shapes, reward and punishment (Barch et al., 2013), and painful pressure applied to the thumb and rest (study 5 from Kragel et al., 2018).

(C) Brain models can vary in complexity, ranging from the average activity of individual brain regions (e.g., a bilateral amygdala mask [left]; Swartz et al., 2015) to more complex patterns of brain activity optimized through statistical learning procedures (e.g., the top 1% of voxels that predict crowdfunding choices [center] [Genevsky et al., 2017] or the Neurological Pain Signature [right] [Wager et al., 2013]).

development. But it also raises some fundamental issues about how researchers define and evaluate mental constructs, and what it means to identify a *brain representation* that underlies them. As the science of multivariate brain models develops, the field must grapple with these questions. Scientists are already engaged in the demanding work of iteratively identifying potential mental constructs, developing neural measurement models for them, and validating, refining, and redefining those constructs based on empirical data. An explicit formalization of this process can identify gaps in current research and accelerate progress toward a fundamental goal of cognitive neuroscience and related fields: establishing mappings between mind and brain.

Theoretical Assumptions about Neural Representation

Brain mapping in neuroimaging emerged from a convention of thinking of mental processes as being modular and implemented in isolated, local brain regions. This view is grounded in long-standing assumptions in philosophy of mind (Lindquist and Barrett, 2012) and studies showing that lesions of distinct cortical areas produce deficits in speech production, language comprehension, perception, and action (reviewed in Banich, 2004; Brett et al., 2002). This work supported the notion that the brain can be thought of as a collection of functional modules—independent, separable processing units that access one another's inputs and outputs, but not intermediate processes (Fodor, 1985; Marr, 1977). Though challenged on theoretical grounds (e.g., Barrett, 2009a; Jonas and Kording, 2017; Sarter et al., 1996), this basic assumption was adopted in early neuroimaging studies, and it became popular to analyze brain-mind associations by analyzing each brain voxel independently (Brett et al., 2002; Logothetis, 2008).

In contrast, multivariate predictive models emerged from theories grounded in neural population coding and distributed representation. Neurophysiological studies have established that information about mind and behavior is encoded in the activity of intermixed populations of neurons. Many studies identify information encoded in single neurons—but often, activity in even the most stimulus- or task-predictive individual neurons contains too little information to accurately predict behavior. A literature on population coding demonstrates that behavior can often be more accurately predicted by joint activity across a population of cells (Averbeck et al., 2006; Pouget et al., 2000), including motor control (Georgopoulos et al., 1986), face perception and identification (Chang and Tsao, 2017; Young and Yamane, 1992), object recognition (Hung et al., 2005; Kiani et al., 2007), control of eye movements (Lee et al., 1988), odor perception (Iurilli and Datta, 2017), numerosity (Tudusciuc and Nieder, 2007), and more.

Population-coding studies have shown that most cells are not strongly selective for a single stimulus or action category, such as object type (Kiani et al., 2007) or saccade direction (Sparks et al., 1990), but rather respond to complex combinations of categories (Rigotti et al., 2013). Firing rates for non-preferred categories are stable and reproducible, and including them in predictive (i.e., "decoding") models results in stronger classification performance than models containing only strongly category-selective neurons (Kiani et al., 2007). Further, they provide strong classification performance after removing strongly responsive neurons from predictive models (Rigotti et al., 2013; Tudusciuc

and Nieder, 2007). Deactivation of neurons strongly responsive to one category (e.g., one saccade direction) does not eliminate that category of responses, but rather results in predictable shifts in behavior consistent with population coding (Lee et al., 1988). And in addition to variation in the mean activity in populations of cells, co-variation between neurons is also important (Ni et al., 2018). These findings indicate that when it comes to information coding, the whole is often greater than the sum of its parts.

Population codes have several adaptive benefits that may have driven their evolution, including robustness, noise filtering, and the ability to encode high-dimensional, nonlinear representations that can be used flexibly (Pouget et al., 2000). Distributed representation permits combinatorial coding (Osborne et al., 2008), providing the capacity to represent a great deal of information with limited neural "real estate." Neurons are elements that can be combined into a nearly infinite number of system states, exponentially increasing the network's coding capacity (Rolls, 2007). Such generative systems are pervasive. For example, 26 Latin letters are the basis for all the words in the English language. By contrast, the Chinese Zhonghua Zihai character dictionary includes over 85,000 logograms, each representing a single word or concept (Russell and Cohn, 2012).

These advantages have inspired artificial neural networks that capitalize on these principles (O'Reilly et al., 2012; Rumelhart et al., 1986). Neurons in these models encode features of input objects (e.g., images, text, etc.) in a highly distributed, "many-to-many" fashion. Each neuron represents many object features, and a representation of an object feature is distributed across many neurons, providing a rich way of representing similarities and associations across objects. Neurons in deep network layers encode complex combinations of features, which has proven critical for the improved predictive accuracy of deep learning models relative to other models (Krizhevsky et al., 2012; LeCun and Bengio, 1995; LeCun et al., 2015). Such models can also be used decode, and create, never-before-seen objects (Gregor et al., 2015; Miyawaki et al., 2008; Nguyen et al., 2016).

Multivariate modeling of how activity spanning many brain voxels jointly encodes behavioral outcomes in human neuroimaging is an extension of population-coding concepts in cellular neuroscience. Because human neuroimaging provides an indirect measure of neural activity that is more consistent with local field potentials and bulk calcium imaging than the activity of single neurons (Logothetis et al., 2001; Nir et al., 2007; Schulz et al., 2012), the activity in any individual voxel is not viewed as indicative of any specific computation or process, but as part of a distributed representation that can be dynamically transformed during cognition to accomplish different functional tasks. Rather than attempting to localize independent functional modules, as is done in conventional univariate approaches, multivariate methods characterize relationships between distributed patterns of activity and categories of mental events and behaviors (Haynes, 2015; Norman et al., 2006; Poldrack and Farah, 2015).

A Brief History of Multivariate Brain Models

Advances in Multi-voxel Pattern Analysis

Multivariate brain models are a diverse family of models, encompassing multiple goals and analytic approaches. One goal is to accurately predict outcomes (i.e., maximize variance explained by the model), which is useful for future prediction (prognosis). But there are other, complementary goals. Models can be designed to (1) generalize to new groups of people, mental states or behaviors, or testing contexts; (2) discriminate one category of mental events or behaviors from another; and/or (3) be more or less easily interpreted in the context of other neuroscientific data and validated against other findings. Models that are accurate, generalizable, and interpretable provide more than predictions; they provide explanations for the neural bases of mental events. Models also vary in the assumptions they make about how mental events are represented in the brain, with different goals and assumptions indicating different study designs and analytic methods.

Over the past two decades, the variation in the types of models employed has grown dramatically, as assumptions about how the brain represents mental events have changed and some goals originally thought to be unreachable now seem possible. Figure 2A shows a timeline of some of the most important developments, and corresponding choices about model goals and structure. We group these advances into multiple stages, each adding a set of techniques to the neuroscientist's toolbox.

Local Information Coding within Individuals. Early studies were grounded in the assumption that information is primarily encoded in local brain regions, in the activity of functional neuronal columns and ensembles with a fine spatial scale (i.e., ~1 mm or less, depending on the system; Duong et al., 2001; Fukuda et al., 2006) and whose precise topography varies across individuals (Issa et al., 2013). Modeling efforts thus focused on predicting mental states within individuals in spatially localized regions. The goal was not to develop a useful overall model of perception or behavior, but—as in traditional brain mapping—to understand local brain representation.

Using this approach, several seminal papers showed that brain activity in early visual cortex could be used to predict the orientation of line gratings a person was viewing (Kamitani and Tong, 2005); other work demonstrated that activity patterns identified in this way could be used as probes of working memory. For example, models developed to identify the perceived orientation of line gratings could be used to infer the contents of working memory in the absence of visual stimulation (Harrison and Tong, 2009). These studies and seminal work in other *domains* (e.g., Hampton and O'Doherty, 2007; Kay et al., 2008; Kuhl et al., 2011) have helped establish predictive analyses within local regions of interest as a way of understanding local representation of mental events. An extension, searchlight mapping (Kriegeskorte et al., 2006), involves multivariate prediction within local spherical "searchlights" across the brain to construct brain maps of *where* information about a mental/behavioral outcome is encoded, which has become a popular technique for mapping local brain information content (Haynes et al., 2007; Peelen et al., 2010; Rissman et al., 2010).

Although these studies illustrated a groundbreaking new approach, they are limited in some important ways. First, showing that a local fMRI model predicts an outcome above chance does not, in itself, license the use of the model as a "marker," or proxy for a brain representation. Using a brain model to infer the presence or strength of a mental event requires

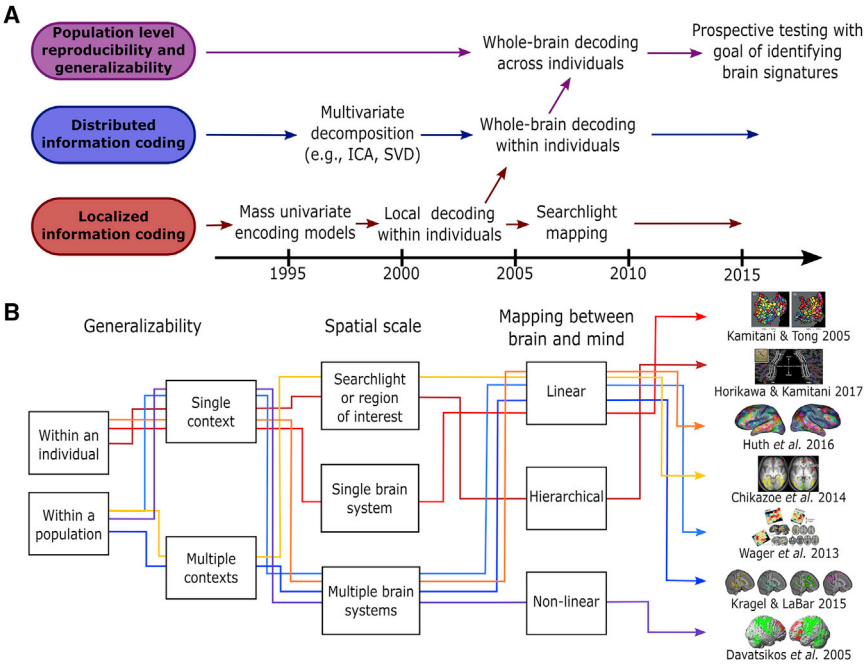

Figure 2. Advances in Multivariate Brain Modeling

(A) Timeline of methodological developments in predictive brain modeling. Advances in predicting behavioral and mental outcomes are influenced by ideas about local coding, distributed coding, and *generalizability*. These ideas fostered complementary tools for analyzing brain data.

(B) Decisions involved in developing multivariate models of brain activity. Three classes of decisions involve the intended generalizability of a model (should it work for a single individual, or a whole population?), the spatial scale of modeling (should activity within a local searchlight, a single brain system, or the whole brain be modeled?), and the complexity of relationships linking brain and mind (e.g., should a linear or quadratic function be used to map brain activity to model outcomes?). The rightmost column depicts multivariate brain models that are the result of different methodological decisions.

assuming—or, ideally, showing—that (1) the putative brain marker is causally related to the mental event, rather than confounding processes; (2) the brain marker is sufficient to capture the brain representation of the mental event and detect it with high *sensitivity*; and (3) the brain marker is specific to the mental event of interest (Davis et al., 2017; Woo et al., 2017a). The brain marker has high *positive predictive value* for the mental event—an indicator that activation of the brain marker implies that the mental event occurred—if and only if the latter two criteria are met. These criteria are particularly difficult to meet in single-subject, local decoding models.

Another limitation is that because single-subject decoding identifies a different model (e.g., different pattern of parameter estimates based on the observed fMRI activation) for each individual participant, it allows a great deal of flexibility in capturing artifacts and confounding processes not related to the mental process being studied (Davis and Poldrack, 2013; Gilron et al., 2017; Todd et al., 2013). Potential confounds (e.g., time-varying effects such as learning, habituation, and fatigue) are not typically modeled in local decoding studies, and avoiding systematic biases requires specialized experimental designs (e.g., within-person counterbalancing and stratification on confounds) that may be impractical in many cases. Individualized models also cannot be tested for accuracy, generalizability, or susceptibility to confounds in new studies, without bringing the same individuals back for re-testing. That is, it is possible to replicate the finding that premotor activity predicts future choices (Haynes et al., 2007; Soon et al., 2008), but not possible to test whether the precise models used for each individual represented intended actions or, rather, another correlated process unrelated to willed action.

Finally, local prediction relies heavily on the assumption that information is contained predominantly in fine-grained, local patterns. If this is not the case, then predictive models will perform poorly if restricted to focal brain regions. And the lower the decoding accuracy (or related measures of effect size; Box 1), the more likely it is that the brain measure is too noisy to serve as a proxy measure, the brain area targeted may play only a minor role in representing the mental process, or the association between brain and mental/behavioral outcome is artifactual. Unfortunately, it is not easy to know how accurate local searchlight models are in much of the published literature because post hoc effect sizes in searchlight maps are optimistically biased if significant regions are selected from among many regions tested (Reddan et al., 2017).

Thus, the development of brain measures as indicators for mental constructs like perception, working memory, pain, etc. is an important goal, but it requires inferences that are difficult to establish in single studies, let alone single participants. Several developments in the field address different aspects of these limitations, as we describe below.

From Local to Brain-wide Decoding. Rather than focusing on individual regions, other studies started from the assumption that information is encoded in distributed brain systems, and that characterizing complex behaviors may require models that capture patterns of activity across these systems. This assumption leads to models that make predictions based on joint patterns of activity and/or connectivity across many voxels (currently up to hundreds of thousands) spanning the brain.

This approach was slower to develop, in part due to the potential for overfitting when there are many more model parameters (e.g., voxels) than observations, producing models that do not generalize well (for more explanation in the neuroimaging context, see Pereira et al., 2009). However, machine learning techniques that regularize or reduce the complexity of predictive models with large numbers of features help to overcome this challenge and make whole-brain models viable (Gramfort et al., 2013; Grosenick et al., 2013; Michel et al., 2011). Studies emerged using brain-wide patterns to decode the contents of

Box 1. Effect Size as a Function of Spatial Scale

Measures of effect size (e.g., Cohen's *d*, Pearson's *r*, or Glass' Δ) indicate the strength of an observed relationship independent of sample size. The goal of predictive modeling is to develop models of brain activity that can detect specific mental states with large effect sizes.

Although many neuroimaging studies report effect sizes, most are not appropriately designed to estimate them in an unbiased manner (Reddan et al., 2017). Most studies report effects for a small subset of the most significant voxels from mass-univariate tests, which are not representative of future performance due to a voxel selection bias that makes them over-optimistic (Varma and Simon, 2006). This bias can be avoided by nested cross-validation and prospective, out-of-sample testing.

Multivariate brain models that accumulate information from many different neural sources increase sensitivity and effect sizes when the brain information of interest is distributed across regions. Recent studies examining brain representations of negative emotion (Chang et al., 2015) and vicarious pain (Krishnan et al., 2016) have shown that predictive models utilizing information spanning the entire brain perform better than models based on local patterns of activity within spherical "searchlights" (Kriegeskorte et al., 2006). These results suggest that complex emotional states are best characterized in terms of global brain states, as opposed to activity within localized neural substrates (Barrett, 2017).

Within the extremes of local searchlights and whole-brain models, predictive models can decode information at multiple different spatial scales. Figure 3 depicts how information is encoded across multiple brain systems, by plotting the performance of pain-predictive models developed using brain activity from multiple resting-state networks (i.e., uniform random sampling from the entire brain) versus models that sample from a single resting-state network and the most predictive searchlight from the entire brain. Models using features that span multiple networks ($\bar{r}$ = 0.544 for the whole-brain models) have larger effect sizes than those using features from a single network ($\bar{r}$ = 0.487 for the visual network models), indicating that information about pain experience is contained in patterns of activity that span multiple brain systems.

memory (Polyn et al., 2005) and semantic information (Mitchell et al., 2008), and differentiate among cognitive task types (Poldrack et al., 2009). More recent studies have shown that signals containing information about reward and punishment (Vickery et al., 2011), working memory (van Ast et al., 2016), semantics (Huth et al., 2012, 2016), pain (Brodersen et al., 2012; Marquand et al., 2010), sustained attention (Rosenberg et al., 2016), and other functions are not confined to single brain regions or systems, but are widely distributed across brain regions. A new direction is direct comparisons of models that operate at different spatial scales, permitting inferences about where and how broadly mental/behavioral information is encoded (Swisher et al., 2010). Early model comparison studies suggest that information about at least some classes of mental events is indeed distributed across regions and systems (e.g., Chang et al., 2015; Krishnan et al., 2016; Box 1).

Broadening the spatial scale of a model does not, in itself, address limitations inherent in single-subject models, including (1) susceptibility to confounds (especially diffuse neuromodulatory effects that are not specific to the target mental state), (2) poor interpretability of the model parameters (i.e., patterns within and across voxels), and (3) inability to test the generalizability and *specificity* of already-trained models across participants, contexts, and types of mental events. In addition, despite regularization and related modeling techniques, estimates of these parameters are generally noisier than standard univariate maps and their interpretation is more complex. For example, brain features important in the model may capture and control for sources of noise in the data, rather than being directly related to mental events (Haufe et al., 2014).

The interpretability of model parameters is particularly complicated when nonlinear mappings, such as those implemented by commonly used radial basis function support vector machines and deep neural nets, are used to predict mental states from brain activity. These approaches create mappings between parameters (generally, brain features) and outcomes that are complex and nonmonotonic (Kamitani and Tong, 2005; Norman et al., 2006). A classic example of this problem involves decoding object identity from retinal activity. A complex non-linear model can use retinal activity to predict the semantic category of the object one is viewing, even though individual neurons in the retina do not respond based on semantics. The representation of categories is encoded in the model, but not directly in the activity of retinal cells. A linear model will fail to predict semantic categories—an advantage in this case—because it relies only on information encoded in the system in a linear fashion. Despite this challenge facing nonlinear models, seminal studies using deep nets to model processing in the ventral visual pathway have shown a striking convergence with biological data (DiCarlo et al., 2012; Horikawa and Kamitani, 2017). More generally, issues related to interpretability can be partially addressed by assessing their relationship with univariate encoding weights (Haufe et al., 2014; Jimura and Poldrack, 2012; Woo et al., 2017b) and evaluating the reproducibility of model weights across individuals (Strother et al., 2002; Varoquaux et al., 2017).

From Modeling Individuals to Populations. Given the limitations raised above, researchers have increasingly focused on identifying models that generalize across individuals. Models that predict outcomes in a group of subjects are constrained to have identical model parameters and estimates across individuals, reducing idiosyncratic artifacts (Todd et al., 2013) and increasing interpretability. Additionally, model performance can be tested on out-of-sample individuals, yielding estimates of person-level performance as is the case for diagnostic tests used in medicine.

This approach assumes that there is useful information contained in patterns of brain activity that are consistent across

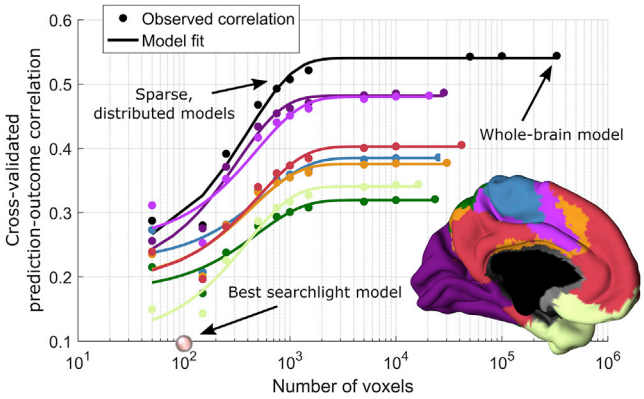

Figure 3. The Effect of Spatial Scale on Model Performance
The plot shows the average cross-validated performance (Pearson's r, averaged over 500 iterations) of models designed to predict pain reports following thermal stimulation using a 2-fold subject-independent cross-validation (data from study 2 of Wager et al., 2013; n = 33). The x axis denotes the number of voxels used in each model, which were sampled randomly from a uniform distribution spanning the entire brain (black) or individual resting-state networks (colored lines; inset render shows each network from a medial view). Solid curves display parametric fits of the form $A - Be^{-v/C}$, where v is the number of voxels, A is the performance of the whole-brain model, B is the performance of a single voxel, and C determines the rate of increase. Sampling voxels from the whole brain produces the most predictive models, compared to sampling within a single resting-state network or searchlight, although only ~1,000 randomly sampled voxels are needed to achieve this performance.

individuals—meso-scale and systems-level activity. In spite of initial assumptions to the contrary, cross-subject decoding proved effective in a number of different areas, including identification of attentional states (Mourão-Miranda et al., 2005), detecting the semantic category of perceived objects (Shinkareva et al., 2008), and diagnosis of dementia (Davatzikos et al., 2009), depression (Drysdale et al., 2017), chronic pain (Baliki et al., 2012; Mansour et al., 2016; Tétreault et al., 2016), and other clinical outcomes (for reviews, see Gabrieli et al., 2015; Woo et al., 2017a). These studies shaped the landscape of predictive models by establishing population-level models that generalize across participants.

One potential disadvantage is that population-level models are not always as predictive as individualized models (e.g., see direct comparisons in Clithero et al., 2011; Haxby et al., 2011; Lindquist et al., 2017; Shinkareva et al., 2008). One important limitation is inter-subject variability in structural and functional anatomy that reduces generalizability across subjects (Cox and Savoy, 2003). Statistical theory shows that the relative costs of between-person relative to within-person prediction depend on the ratio of between-person variance (individual differences) to within-person variance (individual measurement error; Lindquist et al., 2017). Larger individual differences and the ability to collect large amounts of data per person shift the advantage toward within-person models. However, as the amount of data that can be collected on one individual is often limited, as in standard brief "localizer" tasks, in some cases between-person models perform nearly as well as within-person models (Lindquist et al., 2017) or are actually *more* accurate (Chang et al., 2015). Several new models, including hyperalignment (Haxby et al., 2011) and other methods that align functional (rather than anatomical) regions across participants, can dramatically reduce inter-subject variability in functional anatomy, increasing the accuracy and specificity of population-level models.

Several other important limitations remain, including limitations in interpretability (Haufe et al., 2014), the potential for confounds, and questions about whether a given model generalizes to different contexts (i.e., predicts related outcomes in superficially different settings). However, these limitations can be partially overcome with considered choices of training and testing data and model structure, as we describe below.

Generalization across Contexts. A major challenge with all types of predictive models is ensuring that they reflect a particular target mental process (e.g., pain, attention, etc.). In some cases, the model may track correlated superficial variables, and in others, it may track the mental process in only one context. For example, is a brain classifier that predicts when individuals are viewing angry versus fearful faces picking up on the concepts of "anger" and "fear" in general, or rather some particular aspects of the faces and eye-movement patterns during viewing? An important direction in multivariate modeling is to explicitly train models that are robust to variations of the experimental context—e.g., angry versus neutral pictures, sounds, memories, etc. Systematically generalizing over experimental contexts makes models more likely to reflect a targeted mental category, rather than correlated sensorimotor and cognitive processes.

Several recent studies develop models that generalize across superficially different exemplars of a mental process. Work on modeling emotion categories (fear, anger, etc.) has trained population-level models to generalize across music and film clips (Kragel and LaBar, 2015), short movies and mental imagery (Saarimäki et al., 2016), and diverse emotion-induction methods (Wager et al., 2015). Other studies have predicted emotion and affective valence in a way that generalizes across visual and gustatory stimuli (Chikazoe et al., 2014); face, voice, and body cues (Peelen et al., 2010); and across direct perceptions and causal inferences about context (Skerry and Saxe, 2014).

As with other approaches, there are limitations here as well. Different manipulations may prove more or less effective at inducing targeted mental states; for example, video clips are frequently more robust elicitors of emotions than mental imagery or autobiographical recall (Westermann et al., 1996), and certain types of experience may be more difficult to manipulate with certain types of stimuli (Wager et al., 2015). This introduces a confound between category and intensity, which can be accounted for by (1) approximately matching stimuli from different categories on intensity and (2) controlling for intensity during model training. The same principle applies to other potential confounds, and while prior work has accounted for them in some cases, future modeling work should consider them carefully. In addition, it is often infeasible to manipulate more than a few variables in a single study. Combining contextual variation with population-level modeling can help by integrating data across multiple studies with more contextual heterogeneity in the combined dataset (e.g., Kragel et al., 2018). Finally, some mental constructs are hypothesized to vary as a function of context (e.g., Barrett, 2017; Barrett and

Satpute, 2017; Barsalou, 2008; Wilson-Mendenhall et al., 2011). For example, "anger" and "fear" are intrinsically linked to different action tendencies, and it may be impossible to separate emotions from these tendencies.

Brain Signatures: Toward Strong Inference in Relating Mind to Brain

Though multivariate brain models come in many varieties, a common goal is to predict mental events and thereby understand the brain representations underlying them. This can include (1) detecting whether a mental process has been engaged, (2) inferring the strength or intensity of engagement, (3) inferences about which mental categories are similar or distinct in their brain representation, (4) inferring how changes in context or treatments affect the engagement of a mental process and its brain representation, and more.

We argue that a particular class of models—*brain signatures*—is especially useful for this purpose. This class of models uses distributed information within and across brain systems to make population-level, between-subject predictions about the strength of engagement of a mental process, ideally across contexts, and to distinguish it from other categories of mental events (see Box 2 for guidelines on developing brain signatures). Such signatures are effectively brain biomarkers, or neuromarkers (Gabrieli et al., 2015). In our usage, the terms "signature" and "biomarker" are essentially interchangeable. A brain signature is not assumed to be unique to a particular mental process in the way that a handwritten signature is unique to a person. Its sensitivity, specificity, generalizability, and other measurement properties are empirical matters. Likewise, biomarkers in medicine may be more or less accurate, more or less specific to a particular disease, etc. Signatures with desirable properties should be carried forward and tested more extensively, and those with poor measurement properties discarded or redefined (Woo et al., 2017a). By extension, just because model development targets one type of mental event, we should not assume that the targeted event class is the best description of what the signature measures. For example, "pain signatures" trained to track pain may measure the engagement of attention or negative affect. Testing alternative psychological descriptions is also an empirical process, the heart of what we refer to as "construct identification" below.

In addition, a signature may not be a *complete* description of a mental process. It may be useful as an indicator without capturing relevant brain processes, just as a disease biomarker need not capture all aspects of disease physiology. Thus, there is ample room for multiple signatures for the same class of mental event.

The specific modeling choices involved in constructing brain signatures—distributed information and population-level modeling—allow these various types of empirical development and validation to occur across studies and laboratories, dramatically enhancing the ability to (1) falsify models by making strong predictions, (2) develop models with desirable measurement properties, (3) establish reproducibility across studies, (4) use pre-defined models as targets for interventions, and (5) identify the psychological constructs that are measured by brain signatures and develop new psychological ontologies.

Falsifying Models

Thinking of brain signatures as *measures* highlights one of their main advantages: they provide specific, quantitative predictions that can be tested and falsified. This enables a cycle of model formulation and rigorous testing that has been crucial for scientific advancement across fields (Platt, 1964). In physics, it has led to strong, theory-driven predictions that were empirically tested only many years later, like Einstein's prediction based on the general theory of relativity (Einstein, 1916) that starlight should bend around the sun as it reaches earth. This prediction was tested years later by Eddington (Dyson et al., 1920), validating the theory's predictive utility.

For example, a "pain signature" should respond whenever pain is strongly believed to be present, but not otherwise. If it does not respond to pain, and methodological errors can be ruled out, then the hypothesis that the signature reflects all types of pain can be ruled out, paving the way for new refinements. Alternatively, the signature might reflect only some types of pain or pain from some sources, leading to new hypotheses that the brain includes multiple distinct processes that can be labeled as painful. If the signature responds to events that are clearly not painful, like aversive images, bitter taste, or breathlessness, then the signature can be falsified as being unique to pain, and the understanding of what it measures can be refined.

Focus on Measurement Properties

Because much of the focus in both brain mapping and multivariate searchlight approaches has been on explaining local brain representations, relatively little attention has been paid to the measurement properties of brain signals as defined by signal detection theory (McNicol, 2005)—e.g., their sensitivity, specificity, positive predictive value for behavior, and generalizability. A second advantage of brain signatures is the fact that they have definable measurement properties that allow models to be empirically tested in subsequent research.

In addition, a closer match between multivariate patterns and underlying neural representation naturally leads to better measurement properties. This point is supported by studies of hyperacuity, the observation that multivariate models are sensitive to information coded at spatial resolutions finer than the resolution of neuroimaging data acquisition (Carlson, 2014; Kamitani and Tong, 2005; Wardle et al., 2017). It is also supported by observations of multiscale sensitivity, where distributed information within and across regions provides better prediction than single regions. These advantages have resulted in larger effect sizes for brain-wide multivariate compared to local multivariate and univariate models in direct comparisons in several studies (e.g., Chang et al., 2015; Krishnan et al., 2016) (Figure 3). Because the models that best match underlying neural representations are likely to fit best, model comparisons—univariate versus multivariate, local region versus distributed network—provide a way to use neuroimaging to probe the nature of underlying brain representations.

Reproducibility

The ability to reproduce results across studies and laboratories is a key part of scientific progress. Findings that cannot be replicated may simply be false positives, or the effects may be too context-sensitive for cumulative scientific progress and utility in practical applications. Questions about the reproducibility of scientific findings have become a major issue across fields

Box 2. Guidelines for the Development of Multivariate Brain Markers

Although the exact approach depends on the goal of a specific project, some general guidelines for experimental design—outlined below—apply to many multivariate model-development efforts.

Choice of outcome: outcomes can be either categorical or continuous, and can vary within-person, between-person, or both. One common approach is binary classification with two categories (e.g., two task conditions, or patients versus controls). These models can be useful, but they may often yield brain measures that are not specific. "Greedy" classifiers tend to reflect a range of processes that differ between the two conditions. For example, classifying pain versus no-pain conditions can capture activity related to arousal and attention as well as pain. Regression models trained on parametric variations across multiple conditions (e.g., pain intensity ratings) offer a partial solution, and allow for the estimation of dose-response relationships between brain responses and outcomes.

Choice of experimental conditions and level of analysis: models that capture variation of interest across multiple outcome levels (e.g., intensity of affect or cognitive performance) within-person can increase specificity and permit the evaluation of sensitivity across a dynamic range. This also reduces the potential for floor and ceiling effects. And importantly, training models on within-person outcomes reduces sources of noise that can make prediction of between-person differences intractable, including nuisance variation in both outcome and brain measures: for example, (1) scale rating usage (for self-reported assessments) and (2) vascularization, hematocrit, receptor density, arousal level, and other factors that produce inter-personal variation in brain responses.

Designing for specificity: in addition to target conditions of interest, it is also desirable to include a range of "foil" conditions that are similar to the outcome of interest (i.e., engage some overlapping processes). For example, when predicting taste aversion, foils could include other aversive conditions (pain, touch, pictures, and sounds) and other attention-demanding conditions. Including such foils during model training can increase specificity.

Designing for generalizability: including in training datasets multiple instances of the outcome of interest that vary superficially—e.g., different subjects, studies, populations, and task variants—can increase the generalizability of the model to a broader range of cases.

Avoiding bias in evaluating performance: cross-validation can be a valid and reasonable way to estimate the performance of a classifier in new samples. During cross-validation, one part of the data forms the training set (e.g., some participants, in models designed to generalize across participants) and the remaining data form the test set for performance evaluation. The procedure is repeated for several repetitions ("folds") until all samples have served as part of the test set. It is useful to stratify holdout sets on the outcome, keeping training and test sets balanced on outcome values, and balance outcomes on potential confounds if possible.

Cross-validation is useful, but it can provide over-optimistic estimates when (1) there is dependence among observations in training and test sets; (2) the intent is to generalize to populations that are not identical to the original one (e.g., different demographics and sample characteristics); and (3) fitting multiple models on the same dataset, e.g., to optimize parameters or feature selection. The latter is a form of researcher-induced bias that is often unreported in papers, making it difficult to assess whether reported results are optimistic or not.

We offer the following recommendations to mitigate dependence-related issues: prediction of time series data should control the influence of auto-correlation, for instance by using rolling-hvg block cross-validation (Racine, 2000) or leaving out entire runs. Ideally, prediction should be performed across participants if possible, depending on study goals (e.g., 5-fold cross-validation, leaving out all images collected from subjects in the test set). Dependence across participants can be related to cohort effects (time of day, experimenter, or time of year), family/twin structure, and other factors. If dependencies exist between participants (e.g., dyads or twin samples), those should be part of the same holdout sets. Dependence across participants can also be caused by mean-centering, *Z* scoring, or removing covariates estimated on the full sample; caution should be exercised when applying these types of transformations.

Permutation testing: permuting the order of outcome labels can be used to test for bias and statistical significance of the overall model. In the absence of bias, the permuted distribution should be centered on chance. For such tests, assessing root-mean-squared error is desirable, as correlations between predictions and outcomes can be negatively biased and accuracy is a coarse (imprecise) metric with small samples.

Independent holdout sets: to guard against over-optimism, performance should further be evaluated on a completely independent holdout test set, which should be tested only once to assess the final performance. All training, feature selection, and tweaking of machine-learning parameters should be done in one part of the data, and then tested only once in a completely novel, unseen dataset.

(Continued on next page)

Box 2. *Continued*

Prospective testing: given the early stages of brain patterns as biomarkers or diagnostic patterns, it will be crucial to keep testing and externally validating existing markers on new and independent datasets (in the sense of both convergent and discriminant validity) and to test the effects of different types of experimental manipulations and other factors. Prospective application requires creating a method (e.g., software) for applying the model to a new individual case and obtaining a model prediction, without reference to any normative sample or outcome information. We encourage the creation and sharing of such methods to encourage prospective testing.

(Munafò et al., 2017). But establishing reproducible findings has been particularly problematic in fields that involve massive numbers of tests, as in neuroimaging and genetics. It is also problematic when a large number of context variables may change how a process works, as in translational neuroscience (Begley and Ioannidis, 2015) and some areas of psychology (Doyen et al., 2012; Yong, 2012). Neuroimaging-based mapping of mind to brain lies at the intersection of these danger zones.

Reproducibility is limited in standard voxel-wise maps by a combination of noisy voxel-level measurements and the massive number of tests involved. The more stringent the multiple comparison threshold applied, the less likely that studies with the same true underlying neural activity will yield the same results—in effect, low statistical power can ensure that each study identifies a tiny, and often different, part of the true underlying pattern (e.g., Yarkoni, 2009). By contrast, signature-based approaches integrate brain information into a single optimized prediction and test predictions on new, independent individuals. This avoids the need for multiple comparisons and provides unbiased estimates of effect size (Reddan et al., 2017) when testing how experimental interventions affect pattern expression.

Pooling information over multiple brain regions can yield measures with much larger effect sizes. Whereas local effect sizes are limited (usually "moderate" effects around Cohen's d = 0.5; Poldrack et al., 2017), brain signatures often demonstrate large effect sizes when evaluated in independent studies. For example, a pain-predictive model called the Neurologic Pain Signature (NPS; Wager et al., 2013) yields effect sizes ranging from d = 1.2 to 3.50 for high versus low pain (Krishnan et al., 2016; Wager et al., 2013). In a recent analysis of n = 603 participants across 20 studies from different sites worldwide (Zunhammer et al., 2018), the NPS response is greater for pain versus rest in 95.4% of participants, with an effect size of Hedges' g = 2.30, 95% CI [1.92, 2.69] across studies. The Picture-Induced Negative Emotion Signature (PINES; Chang et al., 2015) differentiated emotionally negative from neutral images with an effect size of d = 4.69. The Vicarious Pain Signature (VPS; Krishnan et al., 2016) yielded effect sizes ranging from d = 1.63 to 1.75 for high versus low observed pain (Krishnan et al., 2016; López-Solà et al., 2017a).

These signatures have been evaluated in multiple studies, which have tested their properties in different ways. The NPS's responsiveness to pain has been replicated in 14 independent published study cohorts and one large-scale analysis (Zunhammer et al., 2018), which have begun to characterize its profile of sensitivity, specificity, and responses to drug and psychological interventions (e.g., Lindquist et al., 2017; Woo et al., 2017a). It responds to some interventions, including the opiate remifentanil (Wager et al., 2013; Zunhammer et al., 2018), serotonin reuptake inhibitor citalopram (Ma et al., 2016), and some conditioning paradigms that influence pain expectancies (Woo et al., 2017b). However, it is insensitive to others—including cognitive reappraisal (Woo et al., 2015), perceived control (Bräscher et al., 2016; Woo et al., 2017b), reward (Becker et al., 2017), and placebo (Zunhammer et al., 2018)—indicating that it tracks some of the neurophysiological processes that contribute to pain self-reports, but not others. The PINES response has been doubly dissociated from the NPS, indicating that it measures a distinct set of brain processes (Chang et al., 2015), and has been used as an outcome for emotion regulation (Gilead et al., 2016). In the latter study, taking the perspective of a "tough" individual reduces PINES responses to negative images. The double dissociation of NPS responses to somatic pain and VPS to vicarious pain has been replicated in two additional independent studies in Krishnan et al. (2016) and López-Solà et al. (2017a).

Novel Targets (and Measures) for Interventions

Given the better match to underlying processes and the improved measurement properties, multivariate brain models are promising targets for causal interventions that directly or indirectly influence neural activity (e.g., neurostimulation or neurofeedback, respectively). Population-level brain "signatures" are particularly useful in evaluating these properties because they enable cross-study testing and cumulative science. Brain stimulation targeting multivariate brain models has proven effective, particularly in shaping memory performance (Ezzyat et al., 2017, 2018; Rose et al., 2016). Often these techniques are aimed at changing activity in a single brain region, but their effects might be more widespread and instead alter representations that are distributed across multiple brain systems (Antal et al., 2008; Bestmann et al., 2004; Martin et al., 2013). In this case, multivariate markers can serve as outcome measures by measuring target processes altered by neurostimulation, and identifying brain mediators of effects on behavioral or clinical outcomes.

Construct Validation: Understanding Mental Events with Biologically Grounded Models

At the heart of the enterprise of mapping brain to mind is the definition of categories of mental events that should map onto brain processes. Painful heat, cold, and chemical stimuli all involve different peripheral receptors and populations of neurons; are the experiences they evoke all examples of a single type of "pain" that is represented similarly in the brain? Or is the category "pain" more like the category "furniture," a convenience of human thinking and language? Conversely, we use the term "pain" to describe sensations related to both bodily injury and,

Box 3. Why Mental Constructs Should Be Grounded in Biology

There is a long history in science of classification, induction (Mill, 1884), and searching for natural kinds (Venn, 2006). There are several tenets held by proponents of natural kind views that, although not necessarily valid (see, e.g., Searle, 1995, for criticism), shed light on the importance of biology in understanding mental constructs (adapted from Hacking, 1991):

Independence: it is a fact about nature that there are kinds of things; the differences among things are the work of nature, whereas the recognition of those things is the act of man

Definability: there are multiple ways to characterize natural kinds (e.g., they share a causal mechanism, or they have similar properties)

Utility: depending on the purpose, some classifications of objects are more useful than others

Uniqueness: there is a unique best taxonomy in terms of natural kinds that represents nature as it is, although it is not known to us

Thus, classification can be viewed as a way of organizing objects into useful groupings. Regardless of whether there is truly an optimal classification scheme, taxonomies can be compared to one another and evaluated based on their utility. Categories are often formulated in ways that prioritize human communication, rather than advancing understanding or utility in preventing disease or promoting health. This has historically been the case in medicine and psychology. In medicine, the classification of many cancers has traditionally been based on the tissue of origin and the degree of differentiation. While this classification has been useful, evolving research has shown that classification based on the presence or absence of specific genetic or molecular abnormalities may be more useful. The development of therapies tailored to these different subtypes has revolutionized cancer treatment.

Our understanding of the mind has not yet made such a transition. Constructs like "pain" and "emotion" have been defined based on symptoms reported by individuals—self-reports are clustered into categories such as "chest pain," "heartache," "depression," or "anxiety." "Memory" is one mental construct that has been heavily influenced by neuroscience research; types of "memory" are increasingly defined based on the biological systems that underlie them (Squire, 2004). However, we do not yet have biologically grounded ontologies for many other constructs.

sometimes, romantic rejection. Are their brain representations distinct, or should the category "pain" be extended to non-somatic events? The answers to these questions and others determine how we conceptualize the organization of mind and brain, and often have practical implications as well. Causing someone else somatic pain is grounds for legal prosecution; should causing emotional pain be considered equally harmful under the law?

All mental categories are ultimately psychological "constructs," conceptual categories organized into taxonomies of mental events (or alternatively as ontologies; see Poldrack, 2006; Poldrack and Yarkoni, 2016). Mental constructs have traditionally been "folk categories" (Barrett, 2017) defined based on similarities in phenomenology and linguistic usage rather than biological systems. Likewise, disease categories have historically been based on observable symptoms (stomach pain, shivering, etc.) rather than biological causes (bacterial, viral, etc.). Re-categorizing diseases based on their underlying biological pathology was the critical conceptual shift that separates modern allopathic medicine from previous systems for diagnosing and treating illness. Recent initiatives like the Research Domain Criterion (RDoC) framework attempt to accomplish a similar shift in how we think about mental illness (Insel et al., 2010; Insel, 2014).

Whether the outcomes are disease categories or other mental constructs, comparing brain models and their patterns of sensitivity and specificity can be used to validate existing mental constructs and even infer new ones, and thus use the brain to redefine how we think about the mind (Coltheart, 2013; Box 3). At present, studies aiming to develop multivariate brain models implicitly attempt to validate constructs, but do not systematically take advantage of construct validation theory. A paradigm shift toward explicitly evaluating brain models with a formal construct development approach may lead to a better understanding of both the brain and mind.

Construct Validation

Principled approaches to defining and validating constructs can be found in measurement theory (Cronbach and Meehl, 1955), but the strategies they use have rarely been applied to neuroscience (for a discussion, see Barrett, 2009b, 2011). A central tenet is the acknowledgment that constructs are not directly observable; rather, they are inferred from performance on multiple measures, called indicators (see, e.g., Strauss and Smith, 2009). For example, psychometric studies assume that constructs such as "general intelligence" or "math ability" cannot be directly observed, but that tests of math and reading can be used as indicators that reflect latent, underlying abilities. If different types of math tests with different material and presentation formats correlate together (convergent validity), it might be inferred that they all measure (load on) the same construct, and one can develop composite measures that track the latent construct better than any single indicator. If math tests are relatively uncorrelated with another coherent set of tests (discriminant validity)—say, of language performance—it might be inferred that the tests measure "math ability" instead of "general intelligence," willingness to follow instructions, etc. This approach uses the similarity structure across indicators to infer the nature of otherwise unobservable constructs.

Construct validation theory provides principled ways of evaluating multivariate brain models, and a path toward using brain models to infer which constructs have coherent neurophysiological mechanisms. Brain models provide putative measures (i.e.,

potential indicators) of latent constructs. Just as individual test items can be pooled together to measure a construct (e.g., subscales of clinical inventories; Henry and Crawford, 2005), brain activity across voxels and systems can be combined to create measures related to latent constructs.

For instance, if the same brain measure is activated by multiple distinct types of pain, but not by manipulations of other emotional or cognitive processes, it displays both convergent and discriminant validity as a measure of the construct "pain" (Kragel et al., 2018). Similarly, if a single brain measure is correlated with performance on multiple tasks requiring motor response inhibition (Wager et al., 2005), it shows some convergent validity as a measure of the construct "inhibition." However, it is not clear what the boundary conditions on the construct are. The brain pattern could be related narrowly to motor inhibition; to a broader construct of "inhibition" that encompasses actions, thoughts, percepts, and memories; or very broadly to "cognitive control." The pattern of convergent and discriminant evidence—manipulations the model is sensitive to and specific against, respectively—identifies the construct. In the inhibition example, such an analysis suggests new tasks that would be informative; for example, inhibition of memories and other cognitive control tasks that are not obviously related to motor inhibition. Recognizing that constructs are inferred makes it clear that we need explicit strategies for inferring what brain models actually measure, and conversely refining psychological taxonomies by identifying constructs with convergent and discriminant validity at the brain level.

Seen in this light, recent studies of generalizability across contexts and stimulus modalities begin to establish convergent validity for affective valence (Chikazoe et al., 2014; Kahnt et al., 2014) and other constructs. Some studies include manipulations of multiple contexts and multiple putative constructs, establishing convergent and discriminant validity in the same study (Kragel and LaBar, 2015; Saarimäki et al., 2016). "Mega-analysis" of person-level image data across studies can extend this process, allowing systematic sampling of multiple constructs each with multiple, distinct manipulations in a way that would be difficult in individual studies. For example, Kragel et al. (2018) analyzed participant-level contrast images from manipulations of pain, negative emotion, and cognitive control. Eighteen studies (with 270 participants) were selected to include three distinct methods for engaging each putative construct (e.g., noxious thermal, mechanical, and visceral stimulation for the construct "pain"), with two representative studies in each method. Modeling the similarity structure across constructs, methods, and studies provided convergent validity for a common representation of "pain" in the anterior midcingulate cortex and "negative emotion" in the ventromedial prefrontal cortex. The study also provided evidence that "cognitive control" may not map as clearly onto one underlying brain representation, but should rather be subdivided into more fine-grained subtypes.

Another way of validating constructs is external validity, which involves using a brain model to predict real-world outcomes (Knutson and Genevsky, 2018). For example, brain responses in the ventral striatum while viewing items predicts later purchasing decisions (Genevsky et al., 2017), ventromedial prefrontal responses predict long-term behaviors such as attempts to quit smoking (Chua et al., 2011), and amygdala activity predicts future anxiety (Swartz et al., 2015). More complex multivariate patterns predict the progression to chronic pain (Vachon-Presseau et al., 2016) and whether prefrontal brain stimulation will be an effective treatment for depression (Drysdale et al., 2017). In another series of studies, brain signatures for six different emotion categories developed in one study (Kragel and LaBar, 2015) were applied to resting-state data in an independent study and shown to correlate with individual differences in mood and personality traits (Kragel et al., 2016). Individuals with higher self-reports of depressive symptoms had greater expression of a "sadness" signature, whereas anxious individuals showed greater expression of a "fear" signature.

Other studies have validated constructs using interventions. Rose et al. (2016) used searchlight mapping to identify regions in which fMRI pattern activity related to an item in working memory (i.e., faces, words, or the direction of motion). They found that persistent activity drops to baseline over time, but transcranial magnetic stimulation of these regions preferentially reactivated memoranda-related patterns and enhanced subsequent memory. Interventions including brain stimulation, neurofeedback, and pharmacology can help validate brain measures by showing that they mediate intervention effects on behavior. Directly manipulating the brain also provides stronger inferences about causal effects of the brain system(s) measured.

Population-level models can bring these various kinds of validation together by allowing them to be tested across studies. The NPS has been tested on data from many laboratories around the world, allowing a provisional (and ongoing) identification of the construct it measures (Figure 4). It tracks pain evoked by noxious peripheral stimulation of multiple types, including thermal (Bräscher et al., 2016; Wager et al., 2013), mechanical (Krishnan et al., 2016), electrical (Krishnan et al., 2016; Ma et al., 2016), capsaicin-potentiated heat, laser, and visceral stimuli, demonstrating convergent validity for evoked pain (see Zunhammer et al., 2018, for a meta-analytic assessment). It does not respond to non-noxious warm stimuli (Wager et al., 2013), threat cues (Krishnan et al., 2016; Ma et al., 2016; Wager et al., 2013), social rejection-related stimuli (Wager et al., 2013), observed pain (Krishnan et al., 2016; López-Solà et al., 2017a), or aversive images (Chang et al., 2015), demonstrating discriminant validity against some kinds of related, non-somatic processes. It has external validity in predicting hypersensitivity in fibromyalgia patients (López-Solà et al., 2017b), though its generalizability to other forms of clinical pain remains unknown, and in showing responses to interventions thought to modulate pain, such as the opiate remifentanil (Wager et al., 2013; Zunhammer et al., 2018).

Challenges and Caveats

In addition to challenges facing the interpretation of model parameters, researchers are often tempted to go beyond the data and jump to broad conclusions about the biological significance of models. For instance, if a model predicts behavior, differentiates emotion categories, etc., it might be assumed to reflect brain systems that are (1) prewired or biologically determined (i.e., occur independently of learning or experience), (2) stable or invariant to context (new sample, individuals, test conditions, metabolic state of the body, etc.), and (3) superior to alternative explanations (that this classification scheme is the "right" or "best" classification). None of these conclusions are logical

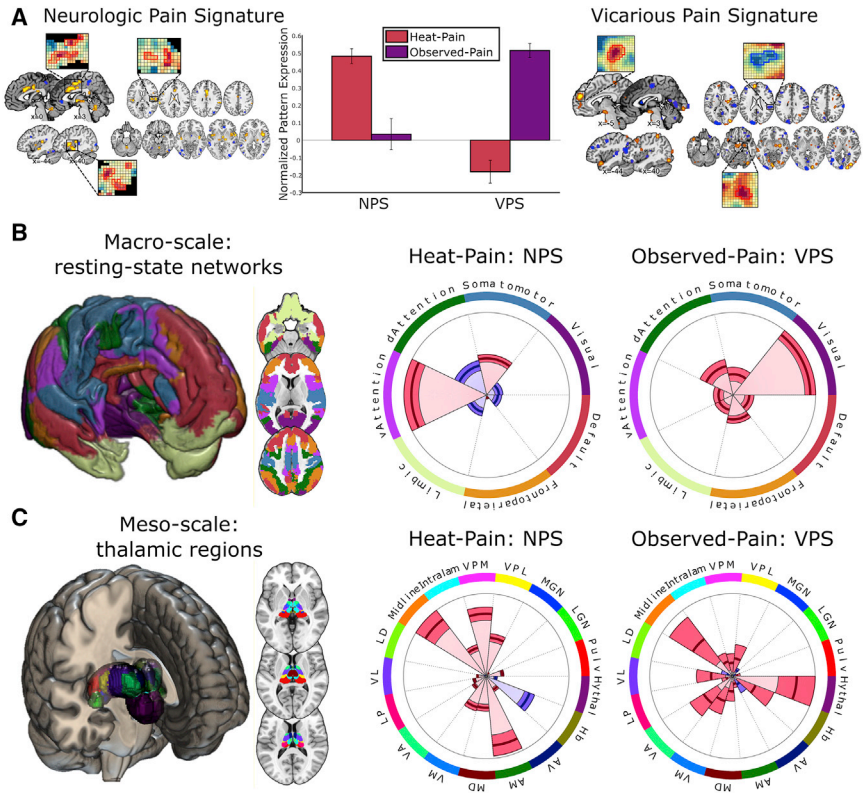

Figure 4. Examining Predictive Models at Multiple Spatial Scales

Brain-wide multivariate models can be understood by examining how pattern expression (i.e., model output for test data) varies across established brain networks and regions.

(A) Brain-wide expression of the Neurologic Pain Signature (NPS), a signature developed to predict physical pain intensity (Wager et al., 2013), and the Vicarious Pain Signature (VPS), a signature developed to predict observed pain intensity, in comparisons of high versus low levels of heat pain (red) and observed pain (purple) (data from study 1 of Krishnan et al., 2016). These two signatures are independently affected by these two manipulations. Adapted from Krishnan et al. (2016).

(B) Decomposing a distributed pattern into subsystems: expression of the NPS and VPS within seven resting-state networks (Yeo et al., 2011). Wedge plots of the same dataset depict normalized local pattern expression (using the signature weights in the local region), with red indicating positive values and blue negative values. The darker shaded area indicates the SEM across individuals. The NPS primarily has positive expression during pain in the "ventral attention" and "somatomotor" networks during heat pain, and negative expression in the "dorsal attention" and "limbic" networks. In contrast, the VPS has more evenly distributed expression across cortical networks, with a peak in the "visual network" during observed pain.

(C) Meso-scale organization: heat pain and observed pain also have distinct profiles of local pattern responses in the diencephalon, based on an anatomical delineation of thalamic nuclei and hypothalamus into 17 distinct regions (Krauth et al., 2010; Niemann et al., 2000). NPS expression during pain is positive in many thalamic nuclei and negative in the habenula, whereas VPS is expressed most reliably during observed pain in the hypothalamus. Error bars reflect SEM.

dAttention, dorsal attention; vAttention, ventral attention; Pulv, pulvinar; LGN, lateral geniculate nucleus; MGN, medial geniculate nucleus; VPL, ventral posterolateral nucleus; VPM, ventral posterior medial nucleus; Intralam, intralaminar nuclei; Midline, midline thalamic nuclei; LD, lateral dorsal nucleus; VL, ventral lateral nucleus; LP, lateral posterior nucleus; VA, ventral anterior nucleus; VM, ventral medial nucleus; MD, mediodorsal nucleus; AM, anteromedial nucleus; AV, anteroventral nucleus; Hb, habenula; Hythal, hypothalamus.

sequelae of the models or examples we discuss here. However, some are testable: new samples can be evaluated, at least in models that generalize across individuals, and context can be systematically varied. The innateness of a brain model can be indirectly inferred by evaluating it across development, or in populations with markedly different cultures and experiences. Such variations are at the core of construct development.

Toward Biologically Driven Construct Development

By explicitly identifying gaps in knowledge, research programs can move more deliberately and programmatically toward the goal of identifying brain representations for mental states and processes. This process is likely to be an iterative one: developing brain models that predict and explain mental constructs will require frequent revisions to both brain models and construct definitions. One goal is to maximize simple structure: to iteratively refine both psychological constructs and brain measures to approximate, as closely as possible, a 1:1 correspondence between them. Revisions to models will teach us about how the brain encodes mental states as we currently define them, and revisions to constructs will help us develop new, neuroscience-informed ideas about how the mind works.

Neuroimaging has contributed to a tension between psychology and neuroscience, as researchers have taken various positions on what studying the brain can tell us about the mind. One expression of this tension is a series of articles challenging whether neuroimaging has taught us anything about how the mind works (Berman et al., 2006; Coltheart, 2013; Henson, 2005; Mather et al., 2013; Poldrack, 2008; Poldrack and Wagner, 2004). While there are various empirical answers to this challenge, progress is often hard to track because it does not come in the form of proving or disproving a critical theory about the mind, but by shifts in the assumptions about how the mind works, which are often implicit, metaphorical, and embedded in our current understanding of physics, computation, and biology.

As an example, computational models of mind are grounded in the traditional concept of the five senses—a basic, implicit concept that few cognitive scientists question. But how many senses are there? Neuroscience has taught us that there is not simply one sense of "touch," but multiple types of somatic sensation that are mediated by distinct pathways and mechanisms (e.g., separate pathways exist for sensation of light touch, deep pressure, painful pressure, inflammation, and other somatic events). For many purposes, including specifying the computational processes involved and predicting outcomes,

these must not be lumped together—and constructs like "touch" have hampered progress toward understanding and clinical applications. In addition, some neuroscience findings challenge the assumption that our five exteroceptive senses are independent from one another; for example, auditory information is encoded in patterns of activity in primary visual cortex, and vice versa (Amedi et al., 2007; Luo et al., 2010).

The same can be said of "cognition," "emotion," "memory," "language," and other mental constructs—there is meaningful variation within each mental construct and meaningful similarity among constructs (Barrett, 2009a; Barrett and Satpute, 2013). Biological understanding allows us to discover meaningful categories, as well as new ideas about how assumed categories (like "working memory") can be affected by processes previously thought to be unrelated (like "inflammation"). Neuroscience can teach us much about the mind if we are open to using its insights to make novel inferences.

Conclusions

A new wave of multivariate predictive models is relating mind to brain in new, more powerful ways. In many domains of cognition, sensation, and affect, such models have highly reproducible relationships with mental phenomena and behavior. They can have very large effect sizes, and in some cases are sufficient to make accurate inferences about individual persons. They can generalize across individuals and testing contexts, providing quantitative models that can be falsified, and can have explanatory power for understanding the brain bases of mental events. This literature represents early steps in a process of construct validation, wherein mental constructs are validated against brain measures. In the future, neuroimaging research programs explicitly designed for construct validation will yield yet more generalizable and useful measures. Such models can serve as targets for interventions—psychological, pharmacological, and neurological. Ultimately, such models can also be used to refine psychological concepts to bring them closer in line with their biological underpinnings, and thus reinvent how we think about the mind.

Appendix: Definition of Key Terms

Brain representation: a latent variable that is inferred based on shared variance between measures of brain activity and outcomes of interest (e.g., observed behavior, self-report, or physiological responses)

Brain signature: a multivariate brain model that includes features that span multiple brain systems and is designed to make predictions on data from a population of individuals

Behavioral domain: a set of conceptually related observable behaviors that accomplish a common goal

Generalizability: the ability of a model to perform well when tested in different conditions, e.g., in different scanning sessions, experimental manipulations, individuals in the same study, or different studies

Mental construct: categories of mental phenomena that are inferred by the observation of multiple measurements and are not directly reducible to any single measure or indicator

Mental process: a sequence of operations (or events) thought to produce observable outcomes (e.g., behavior or brain activity)

Mental state: the status of ongoing mental processes at a given point in time

Multivariate brain model: an explicit mapping that transforms multivariate observations of brain activity (or connectivity) into an outcome of interest (either discrete or continuous).

Positive predictive value: the proportion of positive predictions (cases that a model assigns as being the positive class) that are true positives (cases that have positive ground truth labels); $PPV = TP/(TP+FP)$, where PPV = positive predictive value, TP = true positive, and FP = false positive

Reliability: the ability of a model to produce consistent results in the *same* conditions

Sensitivity: the proportion of true positives that are predicted to be positive, also known as the true positive rate; $TPR = TP/(TP+FN)$, where TPR = true positive rate, TP = true positive, and FN = false negative

Specificity: the proportion of true negatives that are predicted to be negative, also known as the true negative rate; $TNR = TN/(TN+FP)$, where TNR = true negative rate, TN = true negative, and FP = false positive

ACKNOWLEDGMENTS

We thank Jack Gallant for insightful discussions about multivariate prediction and for pointing out the example of decoding objects from retinal activity. This work was supported by the following sources of funding: NIH National Cancer Institute U01 CA193632, NIH National Institute of Mental Health R01 MH076136 and R01 MH116026, and NIH National Institute on Drug Abuse R01 DA035484 and T32 DA017637-14.

AUTHOR CONTRIBUTIONS

Conceptualization, P.A.K. and T.D.W.; Writing – Original Draft, P.A.K., L.K., L.F.B., and T.D.W.; Writing – Review & Editing, P.A.K., L.K., L.F.B., and T.D.W.; Visualization, P.A.K. and T.D.W.; Funding Acquisition, L.F.B. and T.D.W.; Supervision, T.D.W.

DECLARATION OF INTERESTS

T.D.W. has the following patents related to this work: (1) US 2016/0054409 fMRI-based Neurologic Signature of Physical Pain (PCT/US14/33538) and (2) US 2018/0055407 Neurophysiological signatures for fibromyalgia (CU4199B-PPA1).

REFERENCES

Amedi, A., Stern, W.M., Camprodon, J.A., Bermpohl, F., Merabet, L., Rotman, S., Hemond, C., Meijer, P., and Pascual-Leone, A. (2007). Shape conveyed by visual-to-auditory sensory substitution activates the lateral occipital complex. Nat. Neurosci. *10*, 687–689.

Antal, A., Brepohl, N., Poreisz, C., Boros, K., Csifcsak, G., and Paulus, W. (2008). Transcranial direct current stimulation over somatosensory cortex decreases experimentally induced acute pain perception. Clin. J. Pain *24*, 56–63.

Arbabshirani, M.R., Plis, S., Sui, J., and Calhoun, V.D. (2017). Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. Neuroimage *145* (Pt B), 137–165.

Averbeck, B.B., Latham, P.E., and Pouget, A. (2006). Neural correlations, population coding and computation. Nat. Rev. Neurosci. *7*, 358–366.

Baliki, M.N., Petre, B., Torbey, S., Herrmann, K.M., Huang, L., Schnitzer, T.J., Fields, H.L., and Apkarian, A.V. (2012). Corticostriatal functional connectivity predicts transition to chronic back pain. Nat. Neurosci. *15*, 1117–1119.

Banich, M.T. (2004). Cognitive Neuroscience and Neuropsychology (Houghton Mifflin College Division).

Barch, D.M., Burgess, G.C., Harms, M.P., Petersen, S.E., Schlaggar, B.L., Corbetta, M., Glasser, M.F., Curtiss, S., Dixit, S., Feldt, C., et al.; WU-Minn HCP Consortium (2013). Function in the human connectome: task-fMRI and individual differences in behavior. Neuroimage *80*, 169–189.

Barrett, L.F. (2009a). The future of psychology: connecting mind to brain. Perspect. Psychol. Sci. *4*, 326–339.

Barrett, L.F. (2009b). Understanding the mind by measuring the brain: lessons from measuring behavior (commentary on Vul et al., 2009). Perspect. Psychol. Sci. *4*, 314–318.

Barrett, L.F. (2011). Bridging token identity theory and supervenience theory through psychological construction. Psychol. Inq. *22*, 115–127.

Barrett, L.F. (2017). The theory of constructed emotion: an active inference account of interoception and categorization. Soc. Cogn. Affect. Neurosci. *12*, 1833.

Barrett, L.F., and Satpute, A.B. (2013). Large-scale brain networks in affective and social neuroscience: towards an integrative functional architecture of the brain. Curr. Opin. Neurobiol. *23*, 361–372.

Barrett, L.F., and Satpute, A.B. (2017). Historical pitfalls and new directions in the neuroscience of emotion. Neurosci. Lett. S0304-3940(17)30617-1. https://doi.org/10.1016/j.neulet.2017.07.045.

Barsalou, L.W. (2008). Grounded cognition. Annu. Rev. Psychol. *59*, 617–645.

Becker, S., Gandhi, W., Pomares, F., Wager, T.D., and Schweinhardt, P. (2017). Orbitofrontal cortex mediates pain inhibition by monetary reward. Soc. Cogn. Affect. Neurosci. *12*, 651–661.

Begley, C.G., and Ioannidis, J.P. (2015). Reproducibility in science: improving the standard for basic and preclinical research. Circ. Res. *116*, 116–126.

Berman, M.G., Jonides, J., and Nee, D.E. (2006). Studying mind and brain with fMRI. Soc. Cogn. Affect. Neurosci. *1*, 158–161.

Bestmann, S., Baudewig, J., Siebner, H.R., Rothwell, J.C., and Frahm, J. (2004). Functional MRI of the immediate impact of transcranial magnetic stimulation on cortical and subcortical motor circuits. Eur. J. Neurosci. *19*, 1950–1962.

Bräscher, A.K., Becker, S., Hoeppli, M.E., and Schweinhardt, P. (2016). Different brain circuitries mediating controllable and uncontrollable pain. J. Neurosci. *36*, 5013–5025.

Brett, M., Johnsrude, I.S., and Owen, A.M. (2002). The problem of functional localization in the human brain. Nat. Rev. Neurosci. *3*, 243–249.

Brodersen, K.H., Wiech, K., Lomakina, E.I., Lin, C.S., Buhmann, J.M., Bingel, U., Ploner, M., Stephan, K.E., and Tracey, I. (2012). Decoding the perception of pain from fMRI using multivariate pattern analysis. Neuroimage *63*, 1162–1170.

Carlson, T.A. (2014). Orientation decoding in human visual cortex: new insights from an unbiased perspective. J. Neurosci. *34*, 8373–8383.

Chang, L., and Tsao, D.Y. (2017). The code for facial identity in the primate brain. Cell *169*, 1013–1028.e14.

Chang, L.J., Gianaros, P.J., Manuck, S.B., Krishnan, A., and Wager, T.D. (2015). A sensitive and specific neural signature for picture-induced negative affect. PLoS Biol. *13*, e1002180.

Chikazoe, J., Lee, D.H., Kriegeskorte, N., and Anderson, A.K. (2014). Population coding of affect across stimuli, modalities and individuals. Nat. Neurosci. *17*, 1114–1122.

Chua, H.F., Ho, S.S., Jasinska, A.J., Polk, T.A., Welsh, R.C., Liberzon, I., and Strecher, V.J. (2011). Self-related neural response to tailored smoking-cessation messages predicts quitting. Nat. Neurosci. *14*, 426–427.

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

Coltheart, M. (2013). How can functional neuroimaging inform cognitive theories? Perspect. Psychol. Sci. *8*, 98–103.

Cox, D.D., and Savoy, R.L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": detecting and classifying distributed patterns of fMRI activity in human visual cortex. Neuroimage *19*, 261–270.

Cronbach, L.J., and Meehl, P.E. (1955). Construct validity in psychological tests. Psychol. Bull. *52*, 281–302.

Davatzikos, C., Xu, F., An, Y., Fan, Y., and Resnick, S.M. (2009). Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. Brain *132*, 2026–2035.

Davis, T., and Poldrack, R.A. (2013). Measuring neural representations with fMRI: practices and pitfalls. Ann. N Y Acad. Sci. *1296*, 108–134.

Davis, K.D., Flor, H., Greely, H.T., Iannetti, G.D., Mackey, S., Ploner, M., Pustilnik, A., Tracey, I., Treede, R.-D., and Wager, T.D. (2017). Brain imaging tests for chronic pain: medical, legal and ethical issues and recommendations. Nat. Rev. Neurol. *13*, 624–638.

DiCarlo, J.J., Zoccolan, D., and Rust, N.C. (2012). How does the brain solve visual object recognition? Neuron *73*, 415–434.

Doyen, S., Klein, O., Pichon, C.-L., and Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind? PLoS ONE *7*, e29081.

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R.N., Zebley, B., Oathes, D.J., Etkin, A., et al. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nat. Med. *23*, 28–38.

Duong, T.Q., Kim, D.S., Uğurbil, K., and Kim, S.G. (2001). Localized cerebral blood flow response at submillimeter columnar resolution. Proc. Natl. Acad. Sci. USA *98*, 10904–10909.

Dyson, F.W., Eddington, A.S., and Davidson, C. (1920). IX. A determination of the deflection of light by the sun's gravitational field, from observations made at the total eclipse of May 29, 1919. Philos. Trans. R. Soc. Lond. A. *220*, 291–333.

Einstein, A. (1916). Die grundlage der allgemeinen relativitätstheorie. Ann. Phys. *354*, 769–822.

Eisenbarth, H., Chang, L.J., and Wager, T.D. (2016). Multivariate brain prediction of heart rate and skin conductance responses to social threat. J. Neurosci. *36*, 11987–11998.

Esterman, M., Chiu, Y.C., Tamber-Rosenau, B.J., and Yantis, S. (2009). Decoding cognitive control in human parietal cortex. Proc. Natl. Acad. Sci. USA *106*, 17974–17979.

Ethofer, T., Van De Ville, D., Scherer, K., and Vuilleumier, P. (2009). Decoding of emotional information in voice-sensitive cortices. Curr. Biol. *19*, 1028–1033.

Ezzyat, Y., Kragel, J.E., Burke, J.F., Levy, D.F., Lyalenko, A., Wanda, P., O'Sullivan, L., Hurley, K.B., Busygin, S., Pedisich, I., et al. (2017). Direct brain stimulation modulates encoding states and memory performance in humans. Curr. Biol. *27*, 1251–1258.

Ezzyat, Y., Wanda, P.A., Levy, D.F., Kadel, A., Aka, A., Pedisich, I., Sperling, M.R., Sharan, A.D., Lega, B.C., Burks, A., et al. (2018). Closed-loop stimulation of temporal cortex rescues functional networks and improves memory. Nat. Commun. *9*, 365.

Fodor, J.A. (1985). Precis of the modularity of mind. Behav. Brain Sci. *8*, 1–5.

Formisano, E., De Martino, F., Bonte, M., and Goebel, R. (2008). "Who" is saying "what"? Brain-based decoding of human voice and speech. Science *322*, 970–973.

Fukuda, M., Moon, C.H., Wang, P., and Kim, S.G. (2006). Mapping iso-orientation columns by contrast agent-enhanced functional magnetic resonance imaging: reproducibility, specificity, and evaluation by optical imaging of intrinsic signal. J. Neurosci. *26*, 11821–11832.

Gabrieli, J.D.E., Ghosh, S.S., and Whitfield-Gabrieli, S. (2015). Prediction as a humanitarian and pragmatic contribution from human cognitive neuroscience. Neuron *85*, 11–26.

Genevsky, A., Yoon, C., and Knutson, B. (2017). When brain beats behavior: neuroforecasting crowdfunding outcomes. J. Neurosci. *37*, 8625–8634.

Georgopoulos, A.P., Schwartz, A.B., and Kettner, R.E. (1986). Neuronal population coding of movement direction. Science *233*, 1416–1419.

Gilead, M., Boccagno, C., Silverman, M., Hassin, R.R., Weber, J., and Ochsner, K.N. (2016). Self-regulation via neural simulation. Proc. Natl. Acad. Sci. USA *113*, 10037–10042.

Gilron, R., Rosenblatt, J., Koyejo, O., Poldrack, R.A., and Mukamel, R. (2017). What's in a pattern? Examining the type of signal multivariate analysis uncovers at the group level. Neuroimage *146*, 113–120.

Gramfort, A., Thirion, B., and Varoquaux, G. (2013). Identifying predictive regions from fMRI with TV-L1 prior. In 2013 International Workshop on Pattern Recognition in Neuroimaging (PRNI) (IEEE), pp. 17–20.

Gregor, K., Danihelka, I., Graves, A., Rezende, D.J., and Wierstra, D. (2015). DRAW: a recurrent neural network for image generation. arXiv, arXiv:150204623 https://arxiv.org/abs/1502.04623.

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., and Taylor, J.E. (2013). Interpretable whole-brain prediction analysis with GraphNet. Neuroimage *72*, 304–321.

Hacking, I. (1991). A tradition of natural kinds. Philos. Stud. *61*, 109–126.

Hampton, A.N., and O'Doherty, J.P. (2007). Decoding the neural substrates of reward-related decision making with functional MRI. Proc. Natl. Acad. Sci. USA *104*, 1377–1382.

Harrison, S.A., and Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. Nature *458*, 632–635.

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.D., Blankertz, B., and Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. Neuroimage *87*, 96–110.

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

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

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

Haynes, J.D., Sakai, K., Rees, G., Gilbert, S., Frith, C., and Passingham, R.E. (2007). Reading hidden intentions in the human brain. Curr. Biol. *17*, 323–328.

Henry, J.D., and Crawford, J.R. (2005). The short-form version of the Depression Anxiety Stress Scales (DASS-21): construct validity and normative data in a large non-clinical sample. Br. J. Clin. Psychol. *44*, 227–239.

Henson, R. (2005). What can functional neuroimaging tell the experimental psychologist? Q. J. Exp. Psychol. A *58*, 193–233.

Horikawa, T., and Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. Nat. Commun. *8*, 15037.

Horikawa, T., Tamaki, M., Miyawaki, Y., and Kamitani, Y. (2013). Neural decoding of visual imagery during sleep. Science *340*, 639–642.

Hung, C.P., Kreiman, G., Poggio, T., and DiCarlo, J.J. (2005). Fast readout of object identity from macaque inferior temporal cortex. Science *310*, 863–866.

Huth, A.G., Nishimoto, S., Vu, A.T., and Gallant, J.L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron *76*, 1210–1224.

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

Insel, T.R. (2014). The NIMH research domain criteria (RDoC) project: precision medicine for psychiatry. Am. J. Psychiatry *171*, 395–397.

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D.S., Quinn, K., Sanislow, C., and Wang, P. (2010). Research domain criteria (RDoC): toward a new classification framework for research on mental disorders. Am. J. Psychiatry *167*, 748–751.

Issa, E.B., Papanastassiou, A.M., and DiCarlo, J.J. (2013). Large-scale, high-resolution neurophysiological maps underlying FMRI of macaque temporal lobe. J. Neurosci. *33*, 15207–15219.

Iurilli, G., and Datta, S.R. (2017). Population coding in an innately relevant olfactory area. Neuron *93*, 1180–1197.e7.

Jimura, K., and Poldrack, R.A. (2012). Analyses of regional-average activation and multivoxel pattern information tell complementary stories. Neuropsychologia *50*, 544–552.

Jonas, E., and Kording, K.P. (2017). Could a neuroscientist understand a microprocessor? PLoS Comput. Biol. *13*, e1005268.

Kahnt, T., Heinzle, J., Park, S.Q., and Haynes, J.D. (2011). Decoding different roles for vmPFC and dlPFC in multi-attribute decision making. Neuroimage *56*, 709–715.

Kahnt, T., Park, S.Q., Haynes, J.D., and Tobler, P.N. (2014). Disentangling neural representations of value and salience in the human brain. Proc. Natl. Acad. Sci. USA *111*, 5000–5005.

Kamitani, Y., and Tong, F. (2005). Decoding the visual and subjective contents of the human brain. Nat. Neurosci. *8*, 679–685.

Kay, K.N., Naselaris, T., Prenger, R.J., and Gallant, J.L. (2008). Identifying natural images from human brain activity. Nature *452*, 352–355.

Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. J. Neurophysiol. *97*, 4296–4309.

Knutson, B., and Genevsky, A. (2018). Neuroforecasting aggregate choice. Curr. Dir. Psychol. Sci. *27*, 110–115.

Kragel, P.A., and LaBar, K.S. (2015). Multivariate neural biomarkers of emotional states are categorically distinct. Soc. Cogn. Affect. Neurosci. *10*, 1437–1448.

Kragel, P.A., Knodt, A.R., Hariri, A.R., and LaBar, K.S. (2016). Decoding spontaneous emotional states in the human brain. PLoS Biol. *14*, e2000106.

Kragel, P.A., Kano, M., Van Oudenhove, L., Ly, H.G., Dupont, P., Rubio, A., Delon-Martin, C., Bonaz, B.L., Manuck, S.B., Gianaros, P.J., et al. (2018). Generalizable representations of pain, cognitive control, and negative emotion in medial frontal cortex. Nat. Neurosci. *21*, 283–289.

Krauth, A., Blanc, R., Poveda, A., Jeanmonod, D., Morel, A., and Székely, G. (2010). A mean three-dimensional atlas of the human thalamus: generation from multiple histological data. Neuroimage *49*, 2053–2062.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. Proc. Natl. Acad. Sci. USA *103*, 3863–3868.

Krishnan, A., Woo, C.W., Chang, L.J., Ruzic, L., Gu, X., López-Solà, M., Jackson, P.L., Pujol, J., Fan, J., and Wager, T.D. (2016). Somatic and vicarious pain are represented by dissociable multivariate brain patterns. eLife *5*, e15166.

Krizhevsky, A., Sutskever, I., and Hinton, G.E. (2012). Imagenet classification with deep convolutional neural networks. In Advances in Neural Information Processing Systems, pp. 1097–1105.

Kuhl, B.A., Rissman, J., Chun, M.M., and Wagner, A.D. (2011). Fidelity of neural reactivation reveals competition between memories. Proc. Natl. Acad. Sci. USA *108*, 5903–5908.

LeCun, Y., and Bengio, Y. (1995). Convolutional networks for images, speech, and time series. In The Handbook of Brain Theory and Neural Networks, M.A. Arbib, ed. (MIT Press), pp. 255–258.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. Nature *521*, 436–444.

Lee, C., Rohrer, W.H., and Sparks, D.L. (1988). Population coding of saccadic eye movements by neurons in the superior colliculus. Nature *332*, 357–360.

Lenartowicz, A., Kalar, D.J., Congdon, E., and Poldrack, R.A. (2010). Towards an ontology of cognitive control. Top. Cogn. Sci. *2*, 678–692.

Lindquist, K.A., and Barrett, L.F. (2012). A functional architecture of the human brain: emerging insights from the science of emotion. Trends Cogn. Sci. *16*, 533–540.

Lindquist, M.A., Krishnan, A., López-Solà, M., Jepma, M., Woo, C.W., Koban, L., Roy, M., Atlas, L.Y., Schmidt, L., Chang, L.J., et al. (2017). Group-regularized individual prediction: theory and application to pain. Neuroimage *145* (Pt B), 274–287.

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

Logothetis, N.K., Pauls, J., Augath, M., Trinath, T., and Oeltermann, A. (2001). Neurophysiological investigation of the basis of the fMRI signal. Nature *412*, 150–157.

López-Solà, M., Koban, L., Krishnan, A., and Wager, T.D. (2017a). When pain really matters: A vicarious-pain brain marker tracks empathy for pain in the romantic partner. Neuropsychologia. S0028-3932(17)30265-8. Published online July 14, 2017. https://doi.org/10.1016/j.neuropsychologia.2017.07.012.

López-Solà, M., Woo, C.W., Pujol, J., Deus, J., Harrison, B.J., Monfort, J., and Wager, T.D. (2017b). Towards a neurophysiological signature for fibromyalgia. Pain *158*, 34–47.

Luo, H., Liu, Z., and Poeppel, D. (2010). Auditory cortex tracks both auditory and visual stimulus dynamics using low-frequency neuronal phase modulation. PLoS Biol. *8*, e1000445.

Ma, Y., Wang, C., Luo, S., Li, B., Wager, T.D., Zhang, W., Rao, Y., and Han, S. (2016). Serotonin transporter polymorphism alters citalopram effects on human pain responses to physical pain. Neuroimage *135*, 186–196.

Mansour, A., Baria, A.T., Tetreault, P., Vachon-Presseau, E., Chang, P.-C., Huang, L., Apkarian, A.V., and Baliki, M.N. (2016). Global disruption of degree rank order: a hallmark of chronic pain. Sci. Rep. *6*, 34853.

Marquand, A., Howard, M., Brammer, M., Chu, C., Coen, S., and Mourão-Miranda, J. (2010). Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. Neuroimage *49*, 2178–2189.

Marr, D. (1977). Artificial intelligence—a personal view. Artif. Intell. *9*, 37–48.

Martin, L., Borckardt, J.J., Reeves, S.T., Frohman, H., Beam, W., Nahas, Z., Johnson, K., Younger, J., Madan, A., Patterson, D., and George, M. (2013). A pilot functional MRI study of the effects of prefrontal rTMS on pain perception. Pain Med. *14*, 999–1009.

Mather, M., Cacioppo, J.T., and Kanwisher, N. (2013). How fMRI can inform cognitive theories. Perspect. Psychol. Sci. *8*, 108–113.

McNicol, D. (2005). A Primer of Signal Detection Theory (Psychology Press).

Michel, V., Gramfort, A., Varoquaux, G., Eger, E., and Thirion, B. (2011). Total variation regularization for fMRI-based prediction of behavior. IEEE Trans. Med. Imaging *30*, 1328–1340.

Mill, J.S. (1884). A System of Logic, Ratiocinative and Inductive: Being a Connected View of the Principles of Evidence and the Methods of Scientific Investigation*Volume 1* (Longmans, Green, and Company).

Mitchell, T.M., Shinkareva, S.V., Carlson, A., Chang, K.M., Malave, V.L., Mason, R.A., and Just, M.A. (2008). Predicting human brain activity associated with the meanings of nouns. Science *320*, 1191–1195.

Miyawaki, Y., Uchida, H., Yamashita, O., Sato, M.A., Morito, Y., Tanabe, H.C., Sadato, N., and Kamitani, Y. (2008). Visual image reconstruction from human brain activity using a combination of multiscale local image decoders. Neuron *60*, 915–929.

Mourão-Miranda, J., Bokde, A.L., Born, C., Hampel, H., and Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. Neuroimage *28*, 980–995.

Munafò, M.R., Nosek, B.A., Bishop, D.V.M., Button, K.S., Chambers, C.D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J.J., and Ioannidis, J.P.A. (2017). A manifesto for reproducible science. Nature Human Behaviour *1*, 1–9.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., and Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In Advances in Neural Information Processing Systems, pp. 3387–3395.

Ni, A.M., Ruff, D.A., Alberts, J.J., Symmonds, J., and Cohen, M.R. (2018). Learning and attention reveal a general relationship between population activity and behavior. Science *359*, 463–465.

Niemann, K., Mennicken, V.R., Jeanmonod, D., and Morel, A. (2000). The Morel stereotactic atlas of the human thalamus: atlas-to-MR registration of internally consistent canonical model. Neuroimage *12*, 601–616.

Nir, Y., Fisch, L., Mukamel, R., Gelbard-Sagiv, H., Arieli, A., Fried, I., and Malach, R. (2007). Coupling between neuronal firing rate, gamma LFP, and BOLD fMRI is related to interneuronal correlations. Curr. Biol. *17*, 1275–1285.

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

O'Reilly, R.C., Munakata, Y., Frank, M.J., Hazy, T.E., and Contributors. (2012). Computational Cognitive Neuroscience, First Edition (Wiki Book).

Osborne, L.C., Palmer, S.E., Lisberger, S.G., and Bialek, W. (2008). The neural basis for combinatorial coding in a cortical population response. J. Neurosci. *28*, 13522–13531.

Peelen, M.V., Atkinson, A.P., and Vuilleumier, P. (2010). Supramodal representations of perceived emotions in the human brain. J. Neurosci. *30*, 10127–10134.

Pereira, F., Mitchell, T., and Botvinick, M. (2009). Machine learning classifiers and fMRI: a tutorial overview. Neuroimage *45* (1, Suppl), S199–S209.

Platt, J.R. (1964). Strong Inference: Certain systematic methods of scientific thinking may produce much more rapid progress than others. Science *146*, 347–353.

Poldrack, R.A. (2006). Can cognitive processes be inferred from neuroimaging data? Trends Cogn. Sci. *10*, 59–63.

Poldrack, R.A. (2008). The role of fMRI in cognitive neuroscience: where do we stand? Curr. Opin. Neurobiol. *18*, 223–227.

Poldrack, R.A., and Farah, M.J. (2015). Progress and challenges in probing the human brain. Nature *526*, 371–379.

Poldrack, R.A., and Wagner, A.D. (2004). What can neuroimaging tell us about the mind? Insights from prefrontal cortex. Curr. Dir. Psychol. Sci. *13*, 177–181.

Poldrack, R.A., and Yarkoni, T. (2016). From brain maps to cognitive ontologies: informatics and the search for mental structure. Annu. Rev. Psychol. *67*, 587–612.

Poldrack, R.A., Halchenko, Y.O., and Hanson, S.J. (2009). Decoding the large-scale structure of brain function by classifying mental States across individuals. Psychol. Sci. *20*, 1364–1372.

Poldrack, R.A., Baker, C.I., Durnez, J., Gorgolewski, K.J., Matthews, P.M., Munafò, M.R., Nichols, T.E., Poline, J.B., Vul, E., and Yarkoni, T. (2017). Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat. Rev. Neurosci. *18*, 115–126.

Polyn, S.M., Natu, V.S., Cohen, J.D., and Norman, K.A. (2005). Category-specific cortical activity precedes retrieval during memory search. Science *310*, 1963–1966.

Pouget, A., Dayan, P., and Zemel, R. (2000). Information processing with population codes. Nat. Rev. Neurosci. *1*, 125–132.

Racine, J. (2000). Consistent cross-validatory model-selection for dependent data: hv-block cross-validation. J. Econom. *99*, 39–61.

Reddan, M.C., Lindquist, M.A., and Wager, T.D. (2017). Effect size estimation in neuroimaging. JAMA Psychiatry *74*, 207–208.

Rigotti, M., Barak, O., Warden, M.R., Wang, X.J., Daw, N.D., Miller, E.K., and Fusi, S. (2013). The importance of mixed selectivity in complex cognitive tasks. Nature *497*, 585–590.

Rissman, J., Greely, H.T., and Wagner, A.D. (2010). Detecting individual memories through the neural decoding of memory states and past experience. Proc. Natl. Acad. Sci. USA *107*, 9849–9854.

Rolls, E.T. (2007). The representation of information about faces in the temporal and frontal lobes. Neuropsychologia *45*, 124–143.

Rose, N.S., LaRocque, J.J., Riggall, A.C., Gosseries, O., Starrett, M.J., Meyering, E.E., and Postle, B.R. (2016). Reactivation of latent working memories with transcranial magnetic stimulation. Science *354*, 1136–1139.

Rosenberg, M.D., Finn, E.S., Scheinost, D., Papademetris, X., Shen, X., Constable, R.T., and Chun, M.M. (2016). A neuromarker of sustained attention from whole-brain functional connectivity. Nat. Neurosci. *19*, 165–171.

Rumelhart, D.E., Hinton, G.E., and McClelland, J.L. (1986). A general framework for parallel distributed processing. In Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1, D.E. Rumelhart and J.L. McClelland, eds. (MIT Press), pp. 45–76.

Russell, J., and Cohn, R. (2012). Zhonghua Zihai (Book on Demand).

Saarimäki, H., Gotsopoulos, A., Jääskeläinen, I.P., Lampinen, J., Vuilleumier, P., Hari, R., Sams, M., and Nummenmaa, L. (2016). Discrete neural signatures of basic emotions. Cereb. Cortex *26*, 2563–2573.

Sarter, M., Berntson, G.G., and Cacioppo, J.T. (1996). Brain imaging and cognitive neuroscience. Toward strong inference in attributing function to structure. Am. Psychol. *51*, 13–21.

Schulz, K., Sydekum, E., Krueppel, R., Engelbrecht, C.J., Schlegel, F., Schröter, A., Rudin, M., and Helmchen, F. (2012). Simultaneous BOLD fMRI and fiber-optic calcium recording in rat neocortex. Nat. Methods *9*, 597–602.

Searle, J.R. (1995). The Construction of Social Reality (Simon and Schuster).

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

Skerry, A.E., and Saxe, R. (2014). A common neural code for perceived and inferred emotion. J. Neurosci. *34*, 15997–16008.

Soon, C.S., Brass, M., Heinze, H.J., and Haynes, J.D. (2008). Unconscious determinants of free decisions in the human brain. Nat. Neurosci. *11*, 543–545.

Sparks, D.L., Lee, C., and Rohrer, W.H. (1990). Population coding of the direction, amplitude, and velocity of saccadic eye movements by neurons in the superior colliculus. Cold Spring Harb. Symp. Quant. Biol. *55*, 805–811.

Squire, L.R. (2004). Memory systems of the brain: a brief history and current perspective. Neurobiol. Learn. Mem. *82*, 171–177.

Strauss, M.E., and Smith, G.T. (2009). Construct validity: advances in theory and methodology. Annu. Rev. Clin. Psychol. *5*, 1–25.

Strother, S.C., Anderson, J., Hansen, L.K., Kjems, U., Kustra, R., Sidtis, J., Frutiger, S., Muley, S., LaConte, S., and Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: the NPAIRS data analysis framework. Neuroimage *15*, 747–771.

Swartz, J.R., Knodt, A.R., Radtke, S.R., and Hariri, A.R. (2015). A neural biomarker of psychological vulnerability to future life stress. Neuron *85*, 505–511.

Swisher, J.D., Gatenby, J.C., Gore, J.C., Wolfe, B.A., Moon, C.H., Kim, S.G., and Tong, F. (2010). Multiscale pattern analysis of orientation-selective activity in the primary visual cortex. J. Neurosci. *30*, 325–330.

Tagliazucchi, E., and Laufs, H. (2014). Decoding wakefulness levels from typical fMRI resting-state data reveals reliable drifts between wakefulness and sleep. Neuron *82*, 695–708.

Tétreault, P., Mansour, A., Vachon-Presseau, E., Schnitzer, T.J., Apkarian, A.V., and Baliki, M.N. (2016). Brain connectivity predicts placebo response across chronic pain clinical trials. PLoS Biol. *14*, e1002570.

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

Tudusciuc, O., and Nieder, A. (2007). Neuronal population coding of continuous and discrete quantity in the primate posterior parietal cortex. Proc. Natl. Acad. Sci. USA *104*, 14513–14518.

Vachon-Presseau, E., Tétreault, P., Petre, B., Huang, L., Berger, S.E., Torbey, S., Baria, A.T., Mansour, A.R., Hashmi, J.A., Griffith, J.W., et al. (2016). Corticolimbic anatomical characteristics predetermine risk for chronic pain. Brain *139*, 1958–1970.

van Ast, V.A., Spicer, J., Smith, E.E., Schmer-Galunder, S., Liberzon, I., Abelson, J.L., and Wager, T.D. (2016). Brain mechanisms of social threat effects on working memory. Cereb. Cortex *26*, 544–556.

Varma, S., and Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics *7*, 91.

Varoquaux, G., Raamana, P.R., Engemann, D.A., Hoyos-Idrobo, A., Schwartz, Y., and Thirion, B. (2017). Assessing and tuning brain decoders: Cross-validation, caveats, and guidelines. Neuroimage *145* (Pt B), 166–179.

Venn, J. (2006). The Logic of Chance (Courier Corporation).

Vickery, T.J., Chun, M.M., and Lee, D. (2011). Ubiquity and specificity of reinforcement signals throughout the human brain. Neuron *72*, 166–177.

Wager, T.D., Sylvester, C.-Y.C., Lacey, S.C., Nee, D.E., Franklin, M., and Jonides, J. (2005). Common and unique components of response inhibition revealed by fMRI. Neuroimage *27*, 323–340.

Wager, T.D., Atlas, L.Y., Lindquist, M.A., Roy, M., Woo, C.W., and Kross, E. (2013). An fMRI-based neurologic signature of physical pain. N. Engl. J. Med. *368*, 1388–1397.

Wager, T.D., Kang, J., Johnson, T.D., Nichols, T.E., Satpute, A.B., and Barrett, L.F. (2015). A Bayesian model of category-specific emotional brain responses. PLoS Comput. Biol. *11*, e1004066.

Wardle, S.G., Ritchie, J.B., Seymour, K., and Carlson, T.A. (2017). Edge-related activity is not necessary to explain orientation decoding in human visual cortex. J. Neurosci. *37*, 1187–1196.

Westermann, R., Stahl, G., and Hesse, F. (1996). Relative effectiveness and validity of mood induction procedures: analysis. Eur. J. Soc. Psychol. *26*, 557–580.

Wilson-Mendenhall, C.D., Barrett, L.F., Simmons, W.K., and Barsalou, L.W. (2011). Grounding emotion in situated conceptualization. Neuropsychologia *49*, 1105–1127.

Woo, C.W., Roy, M., Buhle, J.T., and Wager, T.D. (2015). Distinct brain systems mediate the effects of nociceptive input and self-regulation on pain. PLoS Biol. *13*, e1002036.

Woo, C.W., Chang, L.J., Lindquist, M.A., and Wager, T.D. (2017a). Building better biomarkers: brain models in translational neuroimaging. Nat. Neurosci. *20*, 365–377.

Woo, C.W., Schmidt, L., Krishnan, A., Jepma, M., Roy, M., Lindquist, M.A., Atlas, L.Y., and Wager, T.D. (2017b). Quantifying cerebral contributions to pain beyond nociception. Nat. Commun. *8*, 14211.

Yarkoni, T. (2009). Big correlations in little studies: inflated fMRI correlations reflect low statistical power-commentary on Vul et al. (2009). Perspect. Psychol. Sci. *4*, 294–298.

Yeo, B.T., Krienen, F.M., Sepulcre, J., Sabuncu, M.R., Lashkari, D., Hollinshead, M., Roffman, J.L., Smoller, J.W., Zöllei, L., Polimeni, J.R., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. J. Neurophysiol. *106*, 1125–1165.

Yong, E. (2012). Replication studies: Bad copy. Nature *485*, 298–300.

Young, M.P., and Yamane, S. (1992). Sparse population coding of faces in the inferotemporal cortex. Science *256*, 1327–1331.

Zunhammer, M., Bingel, U., Wager, T.D., and Consortium, P.I. (2018). Placebo effects on the Neurologic Pain Signature: a meta-analysis of individual participant functional magnetic resonance imaging data. JAMA Neurol. Published online July 30, 2018. https://doi.org/10.1001/jamaneurol.2018.2017.