Data Mining the Brain to Decode the Mind

Daniel A. Weiskopf

Abstract: In recent years, neuroscience has begun to transform itself into a "big data" enterprise with the importation of computational and statistical techniques from machine learning and informatics. In addition to their translational applications such as brain-computer interfaces and early diagnosis of neuropathology, these tools promise to advance new solutions to longstanding theoretical quandaries. Here I critically assess whether these promises will pay off, focusing on the application of multivariate pattern analysis (MVPA) to the problem of reverse inference. I argue that MVPA does not inherently provide a new answer to classical worries about reverse inference, and that the method faces pervasive interpretive problems of its own. Further, the epistemic setting of MVPA and other decoding methods contributes to a potentially worrisome shift towards prediction and away from explanation in fundamental neuroscience.

1. Neuroscience and the data revolution

From genetics to astronomy and climatology, the sciences now routinely deal with extraordinarily large quantitative datasets and deploy computational techniques to manage and extract information from them. Neuroscience is no exception to this trend. The quantity and kinds of neural data available have shifted radically in the last two decades (Van Horn & Toga, 2014), a transition striking enough to prompt declarations that "massive data is the new reality in neuroscience and medicine" (Bzdok & Yeo, 2017, p. 560). With this shift has come a transformation in the analytic tools used to share and process this data, as well as a new wave of optimism about the ability of these methods to overcome long-standing theoretical challenges.

The data revolution has several different fronts. Here I will focus on the impact that machine learning (ML) techniques have had on theory and practice in neuroscience. Machine learning allows us to efficiently partition complex datasets, make inferences, conduct searches, and extract hidden patterns.[1] In the following discussion I sketch one way that machine learning has transformed neuroscientific practice, namely through the application of data analytic tools to imaging studies. One such application is in the use of multivariate pattern analysis (MVPA) to uncover neural structure. MVPA-based methods have proliferated since their introduction in studies of visual processing (Haxby, 2001). However, it is by no means clear how best to interpret the outputs of the increasingly complicated machine learning algorithms that lie at the heart of these methods.

My aim in this chapter is twofold. First, I argue that MVPA does not provide a new solution to the longstanding problem of reverse inference, a claim that has been advanced by Guillermo del Pinal and Marco Nathan in several papers (Del Pinal & Nathan, 2017; Nathan & Del Pinal, 2017), and that also comports with the interpretation of MVPA presupposed by many prominent studies. If MVPA enabled us to break new ground in overcoming the challenges of reverse inference, this would be a powerful argument in favor of multivariate studies over traditional mass-univariate approaches. I present three interpretive challenges that cast doubt on the claim that MVPA can singlehandedly resolve the reserve inference debate. These challenges center on these techniques' sensitivity and globality, the instability and interpretive opacity of their results, and their agnosticism with respect to causal structure.

Second, I want to sound a cautionary note about these new tools and statistical techniques. Such technologies are not ideologically neutral. They come with certain preferred uses, as well as a specific deployment of rhetoric which they carry over from their original

[1] Much recent work using machine learning in neuroscience has centered on deep convolutional neural networks (DCNNs). Classifiers such as the ones discussed here are sometimes used to assign labels to the layers of DCNNs, so the two are not entirely unrelated. Nevertheless, DCNNs are substantially different in their structure and uses from the kinds of models I focus on, so I omit further discussion of them.

computational contexts to their neuroscientific applications. With respect to machine learning, the key term often involved is *prediction*. Indeed, some neuroscientists have explicitly begun to couch their epistemic aims in terms not of explanation or understanding, but of greater predictive accuracy. While this may not yet be the prevalent view among practitioners, I suggest that in light of the ease with which machine learning tools can be turned to purely predictive ends we should be cautious about interpreting models that are based on them, and conscious about the subtle effects they may be having on the studies we design and the epistemic aims that we adopt. Prediction and explanation need not inherently be in conflict with one another, and neuroscience should develop multifaceted modeling practices that integrate these goals rather than favoring one over the other.

2. Two forms of reverse inference: Functional and predictive

The ultimate aim of cognitive neuroscience and neuropsychology is to construct interfield theories bridging brain and mind. Ideally, such bridges would comprise an explanatory implementation theory that would make it comprehensible *how* and *why* specific patterns of brain activity realize the cognitive processes that they do. Explaining the neural basis of cognition requires an account of how low-level neural processes give rise to specific cognitive functions, spelled out in terms of their causal capacities and organization. Most acknowledge that this is at present a utopian prospect. A more modest goal would be to make neuroscientific data evidentially relevant to determining the structure of cognition. Debate has raged, however, over how optimistic we should be even about this goal, with skeptics such as Max Coltheart (2006, 2013) doubting whether neural evidence could *ever* be sufficient to distinguish between competing psychological models.

Building interfield theories and bringing neural evidence to bear in psychology requires a reliable inferential framework capable of crossing ontological, epistemic, and methodological boundaries. Here I focus on one facet of this framework, namely *reverse inference*.[2] Reverse inferences move from the fact that neural process N occurs to the conclusion that (with some probability) cognitive process C is engaged (Poldrack, 2006). Reverse inferences hold when N's activation provides sufficient evidence for C, to the exclusion of other cognitive processes that might be taking place. For instance, suppose (1) that activity in Broca's area makes it probable that processing of sequentially structured information is taking place, (2) that this processing is unlikely to be taking place in the absence of this underlying activity, (3) that no other regional neural activity is evidence for this form of sequence processing, and (4) that activity in this area is not strong evidence for the engagement of any *other* type of cognitive process. Knowing these facts, we can use such activation to conclude that a novel experimental task that activates Broca's area involves sequential processing, which may help to decide between two different psychological models of how it is performed.

The present impasse over reverse inference centers on how to respond to the comprehensive failure of functional localization for many cognitive processes of interest. Localization is the claim that particular cognitive processes are realized by neuroanatomically circumscribed brain regions that are relatively small and functionally dedicated. It has increasingly become clear that many, perhaps most, brain regions seem to participate to some degree in several different cognitive processes (Anderson, 2014; Rathkopf, 2013; Weiskopf, 2016). So from the fact that a pattern N occurs there is some probability that *at least one* of the

[2] Its other face, *forward inference*, involves moving in the opposite direction, viz. from the engagement of a cognitive process to the fact that a specific neural process is occurring (Henson, 2006). For discussion of forward inferences in the context of dissociation studies rather than imaging contexts, see Davies (2010).

processes C_1, C_2, …, C_n is being engaged. From this probability distribution we can't conclude that any one of them, to the exclusion of all others, is localized in N. Given that some regions are involved in a wildly heterogeneous-seeming array of activities across many domains, it is hard to conclude that realizing any one of them is that region's determinate and unique function.

A number of strategies have been proposed to deal with the problem of reverse inference under conditions of functional heterogeneity (Burnston, 2016a; Glymour & Hanson, 2016; Hutzler, 2014; Klein, 2012; Machery, 2014; McCaffrey, 2015; Roskies, 2009). Rather than survey all of these here, I will consider a recent proposal to solve the problem by applying machine learning techniques to imaging data.

First, though, we should distinguish two purposes for which one may seek out reverse inferences. Call these *functional* reverse inference and *predictive* reverse inference. A functional reverse inference involves two claims: that activity in N indicates engagement of C, *and* that this relationship holds because the function of N is to realize C.[3] Functional RI is distinguished by the fact that it incorporates a justification for why the inferential relationship is reliable. It is not merely an accidental-but-reliable co-occurrence: the neural process or region that is active is one that has a certain assigned cognitive function. Having such a function imposes highly specific requirements on the causal organization of the underlying region, namely that it be capable of underwriting the pattern of effects that characterize the target cognitive process. This constraint in turn secures an *explanatory* connection between what is happening in N and C's engagement.

[3] Alternately, the second claim can be formulated in mechanistic terms: the neural mechanism involved in N has the function of realizing or implementing cognitive process C. I won't make any assumptions here about whether all realizing structures for cognitive processes are mechanistic.

Seeking out functional RIs such as these is essential to fleshing out the sort of interfield theory sketched earlier.

Predictive RI, by contrast, merely says that activity in N indicates a certain probability of engagement of C. Its focus is on finding reliable indicators of cognition, no matter what the function of those markers is within the mind/brain system. These might be thought of as "cognitive biomarkers". In medicine, a biomarker is any detectable biological signature that is correlated either with the presence or progression of a disorder. Examples include hemoglobin A1C levels for diabetes or the BRCA1 gene for breast cancer. Biomarkers are sometimes linked directly with the underlying causal factors that drive a disorder, but often may reflect effects or other secondary processes that are restricted to their clinical or prognostic utility.

By analogy, neural activity in a region understood as a cognitive biomarker can serve predictive RI perfectly well, despite not being apt for functional RI. The reason is that this activity can be exploited to predict cognitive processing even when it is not what *realizes* that processing. For instance, the "ground truth" might be that N_1 realizes C, but N_1 might also reliably co-occur with N_2—either because N_1 and N_2 are directly causally related (e.g., N_1 causes N_2), or because they are common effects of a distinct cause. Here both N_1 and N_2 would be equally suited for predictive RI, but not equally good for functional RI. The grounds for predicting cognitive processing differ from those that explain it.

Functional and predictive RI are distinguished in terms of the purposes or goals that lie behind them. This is not to deny that they may work together in many contexts. There is no contradiction between gathering information about brain-mind correlations for the purpose of finding realizers and seeking such correlations for the aim of finding strongly predictive neural signatures. Nevertheless, they can also be pursued exclusively, and prescribe different programs of experimental interventions, interpretation of evidence, and statistical analysis. A neural signature of deception, for instance, might be highly predictive and legally probative without tracking the neural implementation of the intent to deceive. Theorists have not always been explicit on which conception of reverse inference is at issue, although most of the debate over bringing neuroscientific evidence to bear on cognitive theories has tacitly assumed a functional conception of RI. Carefully observing this distinction becomes especially important with the recent turn to machine learning methods, because the rhetoric of decoding, and the striking success of ML classifiers on prediction tasks, has begun to drive some neuroscientists towards abandoning explanation in favor of prediction. It is not an accident that the rise of decoding methods in neuroscience has coincided with the more general adoption of predictive machine learning tools in science, medicine, industry, and marketing (see, e.g., Agrawal, Gans, & Goldfarb, 2018).

Proponents of this "predictive turn" argue that it injects much needed rigor into neuroscience and psychology. They correctly point out that these fields have disappointing track records of real-world prediction. The traditional significance tests they frequently use are hard to interpret in predictive terms, and merely fitting statistical models to existing datasets often leaves us unable generate any useful forecasts. These shortcomings have also been obscured to some degree by the focus of recent philosophy of science on questions concerning explanation, to the exclusion of prediction.[4]

The extent to which the predictive turn is becoming more prominent in neuroscience at large is hard to measure given the size and diversity of the field. Nevertheless, passages such as

[4] There are some notable exceptions to this. For instance, Douglas (2009) argues that despite the philosophical neglect of prediction, it remains central to defining the scientific enterprise, and Northcott (2017) points out that in many domains such as political polling, prediction is often a more desirable epistemic trait than understanding.

the following, drawn from position papers by major participants in the debate, represent a few straws in the wind:

> "Perhaps the biggest benefits of a prediction oriented within psychology are likely to be realized when psychologists start asking research questions that are naturally amenable to predictive analysis. Doing so requires setting aside, at least some of the time, deeply ingrained preoccupations with identifying the underlying causal mechanisms that are mostly likely to have given rise to some data." (Yarkoni & Westfall, 2017, p. 18)

> "Isolating components of mental processing leads to studying them only via oppositions, and this reductionism prevents the building of broad theories of the mind. We believe that predictive modeling provides new tools to tackle this formidable task" (Varoquaux & Poldrack, 2019, p. 1)

> "the main goal of the prediction enterprise is to put the built model, with already estimated model parameters, to the test against some independent data… she [the investigator] is not necessarily worrying about how the model works or whether its fitted parameters carry biological insight" (Bzdok & Ioannidis, 2019, p. 3)

The thrust of these passages is clear: prediction should be given *at least* equal (if not greater) epistemic weight as explanation in modeling cognitive and neural phenomena.

Of course, these are merely three papers that stake out their high-level methodological claims relatively quickly. For another indicator of prediction’s rise, consider the rapidly growing field of neuroforecasting, the explicit goal of which is to find neural signals that predict individual, group, or society-wide behaviors, attitudes, and trends (Berkman & Falk, 2013). In some representative studies, activity in medial prefrontal regions of individual smokers exposed to antismoking public health messages has been said to predict the population-level success of those campaigns (Falk, Berkman, & Lieberman, 2012), and nucleus accumbens activation has been singled out as a predictor of aggregate success of crowdfunded projects on the Internet (Genevsky, Yoon, & Knutson, 2017). Often these neural predictors outperform behaviors or expressed attitudes, which makes them especially attractive targets for marketing purposes.

To the extent that there is a move towards predictively oriented studies taking place, this may in part be an effect of the new tools that neuroscientists have at their disposal. The predictive turn is a concomitant of the adoption of techniques from machine learning. Since these tools have a natural epistemic habitat in data science tasks where computationally efficient prediction is the goal, they tend to carry aspects of this habitat with them when they take root in new domains.

3. Decoding the mind with multivariate pattern analysis

Much of the excitement surrounding the use of machine learning in neuroscience is that it offers the possibility of decoding brain activity, a process that its advocates often colorfully refer to as “reading” the mind off of the brain (Norman, Polyn, Detre, & Haxby, 2006; Poldrack, 2018).[5] In a typical decoding experiment, participants perform a set of tasks during a data collection phase. In principle any sort of data can serve as input to a decoding process (EEG, MEG, direct electrode recordings, etc.), but I will focus on functional MRI studies. Participants are scanned while performing tasks that are typically selected for their differences in the information and the processes that they draw on.[6]

The data from these tasks consists of a vector of numbers measuring the change in the BOLD signal at each voxel at each time step of the scanning sequence. In a procedure known as cross-classification validation, each input sequence is labeled according to the task or stimulus condition that it was gathered in (with labels just being binary features), and the data is separated into two piles: a training set and a test set. Typically, data from a certain number of subjects is reserved for testing. The labeled training sequences are then fed into a supervised machine learning classifier until it reaches criterion performance. Testing is then carried out on the remaining reserved data. This process is iterated across different training subsets, and the classifier's overall performance is reported as the average of its performance on each run.[7]

There are many possible classifiers to use in MVPA studies. To streamline discussion, I will focus on a single commonly used example, namely support vector machines (SVM). SVMs

[5] The mindreading rhetoric is handled cagily in the literature. For instance, despite his book's title, Poldrack hedges on the aptness of the "reading" metaphor, referring to it as "audacious" at one point (p. 2). Others have been less cautious: Haynes et al. (2007) explicitly refer to "reading intentions" out from brain activity, and in a review essay Haynes (2012) remarks that thanks to "combining fMRI with pattern recognition" it "has been possible to read increasingly detailed contents of a person's thoughts" (p. 30). He later comments that in practice this form of mindreading will likely be most useful with respect to broad categories of mental states such as the intent to deceive. Finally, Tong & Pratte (2012) helpfully distinguish between "brain reading" and "mind reading", where the former refers to predicting overt or observable behaviors from brain activity, while the latter refers to predicting subjective cognitive states. They regard MVPA methods as having contributed to progress in both (pp. 485-6).

[6] Many studies also use naturalistic tasks (e.g., movie watching) that engage more widespread cognitive processes. For more details on experimental design, see Tong & Pratte (2012), Haxby, Connolly & Guntupalli (2014), and Haynes (2015).

[7] There is reason to think that these prevalent leave-*k*-out training regimes aren't adequately variance-minimizing, however; see Varoquaux et al. (2017), who recommend leaving out 10-20% of the data and using repeated random splits. Because of the relative youth of these paradigms, best experimental practices are still stabilizing.

efficiently learn to assign each voxel a weight according to how well its activity can help to predict the target category. In linear SVMs, each voxel is assigned a positive or negative weight according to its contribution to correct labeling. The SVM's goal is to draw an optimal hyperplane in voxel (feature) space partitioning the space of possible activity patterns into regions corresponding to each label. There are usually many linear partitions available, but optimality means that the hyperplane maximizes the margin from itself to the nearest members of each category. Data sets that cannot be linearly partitioned in their raw form can be transformed using kernel methods into spaces where such partitioning is possible.[8] Once an SVM learns to achieve an optimal degree of separation with the training set, its weights are frozen and its performance is judged by averaging over repeated folds of out-of-sample transfer (i.e., how well it classifies members of the unseen test set).

Decoding, then, is defined as a classifier's performing adequately well at inferring from neural data to a category label standing for something extra-neural, e.g., a perceptual stimulus, a behavioral response, a task condition, or a cognitive process.[9] This decoding paradigm can be illustrated by Kamitani & Tong's (2005) landmark study of visual attention. Participants were initially scanned while viewing gratings oriented at either 45^{o} or 135^{o}, and the resulting images were used to train a classifier on voxels selected from regions V1-V4. They were then shown a grating that superimposed both of the previous ones and asked to direct their attention selectively

[8] Most neuroimaging studies use the standard linear kernel. Higher-order relationships among voxels are considered only in nonlinear classifiers, including so-called "deep" neural networks. Since almost everyone considers these too powerful and unconstrained for use with imaging data, I continue to omit them here.

[9] *Encoding*, by contrast, involves the reverse operation: training classifiers to predict measurements of neural activation given an experimental task, condition, or stimulus input. Note that the encoding/decoding distinction has to do with the *direction of inference* relative to available neural data. In either direction, it is couched in terms of the measured information made available. Further inferences are required to move from this data to conclusions about content or actual neural ground truths. The encoding/decoding distinction also shouldn't be confused with direction of *causality*. Both decoding and encoding are predictive modeling techniques that can be applied to experimental setups in which neural activity is either the cause or effect of the state being predicted.

to one or another of the orientations. The data from the second phase was fed into the classifier trained on the first phase, which was able to discriminate between the two attention conditions with nearly 80% accuracy. They concluded that information about a participant's attentional state can be decoded from activity in visual cortex.

As Varoquaux & Poldrack (2019) emphasize, classifiers' "validity is established by successful predictions from new data, and not by isolating significant differences across observations" (p. 2). In this sense the statistical regime that underlies MVPA is fundamentally different from that of mass univariate analysis. It focuses not primarily on detection of univariate statistical differences in activation patterns, but on extracting predictive information—in any form whatsoever—from distributed neural activity (Hebart & Baker, 2018). The epistemic regime of prediction is therefore entwined with MVPA at a fundamental level.

Machine learning applied to neural data has proven fruitful across many practical domains. Examples include classifying patients into neuropsychiatric groups on the basis of resting scans, diagnosis of neuropathological conditions by biomarkers rather than symptoms, creating brain-computer interfaces and other neuroprosthetics, extracting the contents of ongoing visual perception, and detection of consciousness in unresponsive patients. The success of these clinical and translational applications is more than enough to justify the interest in solving more fundamental theoretical problems using the same analytic toolkit.

4. Decoding as a solution to reverse inference

In experimental setups where what is being decoded is the occurrence of a cognitive process (rather than, say, the presence of a disorder), decoding can be interpreted as the use of classifiers to perform reverse inference tasks. It is a very short step from (1) MVPA reveals that information about mental states can be extracted from measured brain activity to (2) MVPA can be used to infer the occurrence of mental states on the basis of measured brain activity.[10] In several papers, Guillermo del Pinal and Marco Nathan have taken this step. They argue that MVPA provides a new solution to the problem of reverse inference (Del Pinal & Nathan, 2017; Nathan & Del Pinal, 2017). They call this pattern-based reverse inference, by contrast with classical location-based reverse inference.

Their central argument for preferring MVPA to location-based approaches rests on the fact that classifier-based studies satisfy what they call the linking condition (Del Pinal & Nathan, 2017, p. 129). Suppose we want to know whether a task-evoked pattern of neural activity N engages cognitive processes C_1 or C_2. To do so requires *independent evidence* that N is positively linked with, say, C_1 (rather than C_2). In traditional univariate analysis this evidence is precisely what is missing, thanks to the multifunctionality of regions across studies (see Section 2). However, MVPA involves training classifiers on data gathered within phases of the same experiment, rather than making comparisons across experiments. It therefore circumvents the problem by directly comparing activation patterns, where the reliability with which these patterns are distinguishable is determined within the experiment (pp. 135-6). Moreover, MVPA does this without importing any problematic assumptions either about the localization of cognitive processes in brain regions, or about the previously established cognitive functions of those regions.

[10] A closely related inference concerning the decoding of representational content from MVPA classification studies has been challenged by Ritchie, Kaplan, & Klein (2019). See especially pp. 11-13 for a detailed unpacking of the premises that these inferences rely on.

From these points we can extract the following methodological prescription concerning the utility of *decoding for cognitive difference*:

> (DCD): If a decoder can be trained to distinguish neural patterns elicited by two tasks, then the tasks involve different cognitive processes.

DCD relies on the principle that any differences in cognitive processing will be reflected in their underlying neural realization, so no two processes can have (within an individual performing a specific task) the same realization.

Appeal to the DCD principle is implicit in Del Pinal and Nathan's arguments. They propose that multivariate imaging analysis can "overcome the challenge of determining the reliability of bridge laws and, as a result, promise to be a more useful technique for discriminating among competing cognitive-level hypotheses" (Nathan & Del Pinal, 2017, p. 5). Suppose that we begin with a classifier trained to decode cognitive processes C_1 and C_2 from distinct equivalence classes of neural patterns. Then we have the leverage needed to decide whether an arbitrary novel task taps that one or the other of these processes by seeing how that classifier performs on data collected from imaging that task (pp. 5-7). Successful decoding here is presented as sufficiently strong evidence to license functional reverse inferences.

In a related vein, Ritchie, Kaplan, & Klein (2019) articulate a principle they call the "decoder's dictum" that they argue persuasively drives the interpretation of many MVPA studies. According to the dictum, "If information can be decoded from patterns of neural activity, then this provides strong evidence about what information those patterns represent" (p. 2). DCD as presented here can be viewed as complementary to the decoder's dictum: the latter focuses on the decodability of information, while the former concerns the use of decoding to discover cognitive processes. Information and processing are tightly related but nevertheless distinct. Cognitive processes may differ in the informational or representational content that they manipulate, but they may also make distinct uses of the same body of information (if, for instance, the goal of the information processing is different in each case). Ritchie, Kaplan, & Klein's arguments against the decoder's dictum thus dovetail with the ones presented here against the DCD principle. Each attempts to separate and target one strand in the familiarly entwined notion of "information processing."

DCD can also be seen as tacitly driving the interpretation of a number of imaging studies. Varoquaux & Thirion (2014), for instance, propose that decoding provides a "principled methodological framework for reverse inferences" (p. 4), where the latter are understood in the functional sense. Moreover, DCD-like principles aren't confined to the pages of theoretical papers. Consider studies of visual perception such as Haynes & Rees (2005), in which participants simultaneously viewed two stimuli designed to induce binocular rivalry while indicating via button-pressing which of the two they were experiencing at a particular moment. A pattern classifier was trained on activity in 50 voxels of V1 and used to predict the timing with which one or the other visual stimulus became conscious, achieving an 80% success rate. In a separate condition, a classifier trained to distinguish presentations of monocular non-rivalrous stimuli could predict binocular switching similarly well. Haynes & Rees conclude that "[their] data could be taken to represent a simple form of 'mind reading,' in which brain responses were sufficient to predict dynamic changes in conscious perception in the absence of any behavioral clues" (p. 1302). That is, they interpret this study's methods as licensing an inference from accurate machine classification of neural patterns to changes in people's perceptual states.

Similar inferences crop up in studies of pain perception. In one widely cited study, Wager et al. (2013) subjected participants to thermal stimuli varying from warm to painful. These stimuli were both classified and rated according to intensity on a 100-point scale. A sparse pattern classifier (see Section 4.2 below) was trained on a map of anatomical regions preselected for their known involvement in pain processing, and this classifier was tested on scans of neural activity during the stimulation period. The classifier was used to generate predictions of how the stimulus was experienced, and to predict its intensity.[11] It was able to discriminate painful from nonpainful conditions with 93% specificity and sensitivity, and to predict pain intensity well (although warmth intensity was less successfully captured). These results, among others, lead them to conclude that the regions of interest (ROIs) driving classifier performance constitute a "neurologic signature" (p. 1396) or biomarker of subjective pain experience. This again is consistent with DCD, since biomarker regions (as determined by classifier weight assignments) are singled out for their role in predicting participants' experiential reports, which are assumed to reflect their phenomenal state. The logic of this study is representative of that presented in a recent survey and critique of the pain prediction literature by Hu & Iannetti (2016).[12]

Finally, moving from experiential states to cognitive ones, DCD also drives studies aimed at predicting intentions to act. Soon, He, Bode, & Haynes (2013) trained classifiers to find regions that are predictive of conscious decisions to carry out abstract actions (in this case, adding or subtracting single digit numbers). Participants viewed a sequence of slides containing a

[11] These predictions were calculated in terms of a "signature response", here defined as the dot product of the trained classifier weights and the activation map for each temperature within participants (see p. 1391 and the Supplementary Materials). Signature response was used in two ways: to directly predict rated intensity of a stimulus, and with an imposed threshold to predict pain/no pain.

[12] This review also distinguishes between two objectives in decoding: discovering a pain-specific *neural signature* and discovering a reliable pain *predictor*. This approximately corresponds to the distinction drawn here between functional and predictive RI. As the authors note, these two goals prescribe distinct experimental and statistical logics and should be more cleanly separated in practice.

matrix of numbers plus a single letter cue, and were free to choose at any time to either add or subtract the numbers. After indicating readiness and carrying out the arithmetic operation, they reported the result along with which letter was present when they became aware of their decision. Classifiers were trained on scans from the 8-18 seconds preceding their awareness, with the aim of distinguishing between the operations that later they carried out. At 4 seconds prior to awareness of the intention, two regions were able to successfully decode (with 59% accuracy) which type of mental arithmetic the participants carried out. This decoding success was interpreted as evidence for the presence of an unconscious intention to execute a mental action. In their discussion section, they say: "Our results show that regions of medial frontopolar cortex and posterior cingulate/precuneus encode freely chosen abstract intentions before the decisions have been consciously made" (p. 6219). An additional explicit invocation of a DCD-like principle occurs in their methods section, where they note that "[g]ood classification implied that the local cluster of voxels spatially encoded information about the participant's specific current intention" (p. 6221).

These examples suggest that DCD-style inferences of the kind recommended by Del Pinal and Nathan are employed across a number of domains in contemporary imaging studies. Nevertheless, I argue we should reject the claim that decodability of differences between tasks is generally sufficient to reveal cognitive differences. Classifiers are powerful tools, but they often achieve their results for reasons that are opaque or flat out in conflict with the wider epistemic purposes that drive the debate over reverse inference. In the following sections I survey three problems that plague the interpretation of decoding results. The picture that emerges is one on which even when they can attain a high degree of predictive success, we may not be able to confidently infer from this fact to either ground truths about neural functioning or to facts about cognitive processing.

4.1. The problem of sensitivity and globality

Two core traits for which classifiers are touted are their high degree of sensitivity to variations in neural activity and their globality, meaning that in making predictions they inherently take into account spatially distributed voxel patterns. Del Pinal and Nathan specifically cite globality as a virtue when they note that MVPA does not rely on assumptions about localization of cognitive functions in the brain. They remark that “classifiers can employ multi-voxel patterns, which are distributed across traditional brain regions of interest. Hence, the use of [pattern-based reverse inference] is compatible with the possibility that the sources from which to decode cognitive processes are widely distributed patterns” (Nathan & Del Pinal, 2017, p. 7). And this sensitivity to distributed or global patterns in turn means that MVPA methods can be used to detect cognitive processes whose realization spans several multifunctional local regions. This emphasis on the ability of MVPA to track global patterns of interest is often couched in terms of evidence for a highly distributed neural code, with task-relevant information being encoded by subtle activation differences within and across regions (Kragel, Koban, Barrett, & Wager, 2018).[13]

[13] However, despite the fact that it remains common to see successful applications of MVPA described in terms of distributed neural representations, it has been shown that we cannot infer from the dimensionality of the measurements to that of the underlying neural code itself. Linear classifiers will use any number of voxel features that they are trained on, but this does not establish that the brain itself encodes this information in this way (Davis et al., 2014). For a real-world example, single electrode studies can recover information about face identity in macaque visual cortex, but this information cannot be decoded with MVPA, plausibly because of weak clustering of similarly-responding neurons (Dubois, de Berker, & Tsao, 2015).

From this perspective the globality of classifiers is a virtue, since it meshes appropriately with the structure of the underlying neural realizers. Both sensitivity and globality, however, can lead to scenarios in which labeled patterns are distinguished with high accuracy without this necessarily being a sign that different cognitive processes are engaged. In short, classifiers can be *oversensitive* relative to our interest in reverse inference.

To see this, consider that classifiers may succeed for reasons that do not seem related to the functions of the underlying regions or the task being carried out. For example, regions of motor cortex frequently show distinctive activity across task contexts, due to the demands of the specific responses each task requires. A classifier might assign these some predictive value, without their being relevant to the "core" cognitive processes of interest (Jimura & Poldrack, 2012, p. 550). Indeed, in one often-cited study dozens of cortical regions could support successful classification between 30 and 50% of the time (Poldrack, Halchenko, & Hanson, 2009). Cases like these show that a neural pattern can be useful for distinguishing the engagement of two processes without being the realizer of either.

One response to this problem is to be more selective about the regions that are used to train and test classifiers. If motor processes are not thought to be functionally relevant, voxels in motor cortex should be stripped out by deleting them from the input vectors prior to classifier training. But while *a priori* selection of ROIs can remove regions that are believed to be irrelevant to the cognitive processing that we are interested in, this solution doesn't generalize. Sometimes the information that classifiers exploit is present in regions that we *are* interested in, and so it can't be successfully stripped out. For example, regions of primary visual cortex contribute to discrimination of high level visual features despite the fact that we don't have strong reasons to think that they actually compute using the information that can be decoded from them (Cox & Savoy, 2003). Decoders' sensitivity to available information *within* ROIs can easily outstrip the ground truths about whether and how that information is causally used (de-Wit, Alexander, Ekroll, & Wagemans, 2016; Ritchie et al., 2019).

This interpretive problem arises even in the most methodologically sophisticated of studies. Searchlight analysis is a widely used exploratory technique that avoids presupposing anything about specific assignments of functions to local regions (Etzel, Zacks, & Braver, 2013; Kriegeskorte & Bandettini, 2007a, 2007b; Kriegeskorte, Goebel, & Bandettini, 2006). Briefly, it involves dividing the brain (or some region thereof) up into three-dimensional volumes each of which centers on a voxel. A new classifier is then trained on the signals within each such volume to see whether its activity patterns can discriminate among conditions of interest. The metaphorical "searchlight" can be visualized as the iteration of an MVPA detector systematically through these relatively small brain regions. In principle this gives an unbiased procedure for sorting brain volumes by how predictive their activity patterns are. The output of searchlight analysis is typically a map of those regions that can decode the target condition with greater than chance accuracy.

As an illustration, consider Vickery, Chun, & Lee's (2011) study of reward processing. In one of their experiments, they asked participants to play a penny-matching game against a computer, in which the players won on average 48% of the time. They then examined trials on which players won to see whether there were regions from which signs of reward or reinforcement (operationalized as wins vs. losses) could be decoded using both an ROI-based and a searchlight analysis. In the former, they found reinforcement signals decodable in 37 of 43 prechosen bilateral regions, while in the latter ~30% of all voxels elicited significant decoding. In the authors' words, "[v]irtually every major cortical and subcortical division contained a significant cluster in one or both hemispheres" (p. 169). Neural signals linked with reinforcement, then, are far from localized. They can be decoded from an extremely widespread set of brain regions. Moreover, these globally distributed patterns are also more sensitive than paired univariate analyses: only between 9 and 7 of the 43 ROIs were significant when a standard general linear model is applied to the same data.

However, the ability to sensitively decode winning trials from globally distributed patterns does not inherently support the claim that these regions realize or have the function of tracking wins. There may be some very general cognitive process labeled "reinforcement" that is involved in these regions' activity—although whether it is precisely the same process in each case or not would require much more precise specification. But there are many forms that this involvement may take. Detecting wins may modulate other processes carried out within those regions without those regions being in any sense *for* detecting wins. Vickery, Chun, & Lee themselves are cautious on this point, saying that "the functional neuroanatomy exists for positive and negative outcomes to directly influence neural processing throughout nearly the entire brain" (p. 175). A region's processing being influenced by the valence of an outcome does not require that the region has the function of processing that valence, nor that there be any single cognitive process that those regions share. It is compatible with any form of influence strong enough to make the region a good predictor.

It is certainly defensible for some translational purposes to focus just on decoding success. Perhaps engineering brain-computer interfaces or clinical diagnosis are examples. However, doing so involves privileging predictive RI over functional RI. This carries the risk that our models are ignoring potentially explanatory ground truths. Insofar as a model is insensitive to such truths, we should not treat it as directly illuminating cognitive processing.

4.2. The problem of tradeoffs and interpretability

A second problem facing MVPA methods is that even when classifiers can distinguish between task states, increased prediction accuracy *per se* does not guarantee other epistemically desirable properties. Here the problem lies in the fact that what is decoded depends in part on the specific modeling choices made by experimenters. Because classifier performance turns on model selection and tuning of parameters, it embodies certain familiar trade-offs. In particular there is a tension between the *stability* of the weights and the performance of the classifier (Baldassarre, Pontil, & Mourão-Miranda, 2017; Rasmussen, Hansen, Madsen, Churchill, & Strother, 2012; Varoquaux et al., 2017). Stability is a measure of how reliably the same weight pattern will be reproduced by different classifiers, or by different runs of the same classifier. Machine learning research has increasingly focused on the quantifying these tradeoffs, and one consistent result that emerges from these studies is that if we choose parameter assignments that maximize the predictive success of a classifier, we are necessarily sacrificing other potentially important properties.

A typical linear classifier like SVM has a soft margin parameter that determines how much misclassifications are counted against a weight assignment.[14] Sparse classifiers include various regularization terms, which impose parsimony constraints (degree of fit to the data, contiguity, smoothness, etc.) on the resulting weights. These classifiers are used to select only some of the possible input features to drive the weight vector, but a great deal turns on exactly how these parameters are tuned. In one study, Rasmussen et al. (2012) found that as the

[14] The choice of kernel is also significant, but many neuroimaging applications use a linear kernel, so I ignore this complication here.

regularization parameter is varied, predictive accuracy decreases (from ~72% to 50% correct) while pattern reproducibility as measured by Pearson's correlation increases (from 0.0 to 0.5). More accurate prediction, in other words, is purchased at the cost of high variability in the spatial weight map. This implies that credit assigned to one region could be revoked if the same classifier were retrained without alteration.

The tradeoff for a model's high degree of success, then, is a lack of reliable informativeness about what regions are most responsible for that success. This has obvious consequences for the interpretation of classifier performance: we may know that a certain region is predictive without having generalizable insight into why this is the case. These types of tradeoffs apply even within the domain of sparse classifiers, which attempt to group weights into relatively few internally homogeneous or structurally adjacent clusters. In a comparison across six sparse models trained on fMRI datasets, systematic accuracy-stability tradeoffs arise for each one (Baldassarre et al., 2017). A typical sparse classifier such as LASSO can achieve high accuracy (85%) at a corrected overlap score of just under 0.6, while a higher overlap score (around 0.7) returns much worse accuracy (~65%).

If predictive accuracy is all that we care about, it is clear which parameter tuning we should prefer. But in practice, modelers often prefer sparse solutions. What sparseness costs in predictive accuracy it purportedly gains in making models more interpretable and biologically plausible. A non-sparse model can assign decoding importance to a scattered, buckshot-like distribution of regions that lacks any neurophysiological sense. Even sparse models are not interpretively transparent, though. While the best-performing sparse classifiers converged in assigning the same five regions the highest weight (although not in the same order), they still varied widely in how many regions they included overall (from 10 to 106 total). Human-legible interpretation remains challenging with dozens of small, anatomically insignificant regions participating.

The situation with respect to tradeoffs among classifier performance, stability, and interpretability is strongly akin to what Gelman and Loken (2014) famously refer to as the "garden of forking paths" in statistical analysis. The number of available off-the-shelf classifiers plus the number of tunable parameters for each gives rise to potentially quite distinct assignments to each of these three valuable properties. The choice of any particular model-parameter pairing in imaging studies can be epistemically consequential, and can even shape whether a result is considered significant. But such choices are often undermotivated. We should be cautious about interpreting results where the choice of data analysis methods is largely unconstrained except by custom and experimenter preference, and where these choices can make a difference to the outcome of the analysis.

Classifier interpretability is further complicated by the fact that weights may be assigned to voxels that are not the origin of the underlying neural signal, and that low (or even negative) weights may be assigned to voxels where the signal is located.[15] To take one example of this phenomenon, suppose that we have BOLD measurements from two regions, and that the ground truth is that one of these regions contains information that can discriminate moderately well between two labeled conditions, while the other contains no such information. Nevertheless, the weight vector to achieve optimal discrimination can (under the right circumstances) be one that

[15] A related warning is that positive weights on a voxel can reflect *decreases* in its activation, since if these decreases are reliable they may convey information about certain stimulus conditions.

assigns double the weight to the latter region than to the former—despite the fact that the latter region is by hypothesis one that is informationally empty (Haufe et al., 2014).[16]

This idealized case illustrates two broader points about MVPA: first, it runs together signal and noise, treating both as potential information; and second, it assigns weights based not on individual voxel importance but on how well overall classification performance is affected. Classifiers are holistic and opportunistic (see also Ritchie et al., 2019, p. 14). Two voxels that contain no genuine information about which condition obtains can nevertheless be used for discrimination if they have different noise variances in each condition (Hebart & Baker, 2018). So weight assignments are at least sometimes performed on grounds other than the causal-explanatory significance of voxel activity, further complicating interpretability.

Practitioners may object that typical studies look mainly at overall classifier accuracy as their measure of interest. So it may seem unclear why these issues about their precise internal structure may matter. With respect to the reverse inference debate, however, the concern is that we cannot easily analyze classifier successes in terms of the underlying neural ground truths. One can't, for example. conclude from the fact that a successful classifier assigns a certain weight to a voxel that the voxel's activity contains a signal whose function is realizing the cognitive operation that is being decoded. Weights are assigned for the purpose of maximizing overall success using any available cues. As Haufe et al. note: "*A widespread misconception about multivariate classifier weights is that (the brain regions corresponding to) measurement*

[16] This artificial example has been criticized by Schrouff & Mourão-Miranda (2018), who argue that it holds only for low signal-to-noise ratio cases. However, given that it is often unclear what the SNR is for particular ROIs, it is fair to say we cannot across the board rule out the presence of "false positive" voxel weights. Moreover, the type of noise matters. As Haufe et al. point out, it is sometimes possible to correct for the presence of Gaussian noise to recover underlying signal, but this doesn't hold for noise induced by scanner drift, head motion, and periodic noise (P. K. Douglas & Anderson, 2017), all of which are present in imaging data.

channels with large weights are strongly related to the experimental condition" (Haufe et al., 2014, p. 97). If this assumption doesn't hold in general, the undeniable success of classifiers may end up being causally opaque.

Even so, one may wonder why issues such as the interpretability of models should matter from a perspective such as that of DCD, where the express goal of decoding is simply to find evidence that decides between two possible cognitive models. Given del Pinal and Nathan's emphasis on the fact that MVPA does not depend on any specific localizationist assignment of functions to regions, prioritizing sparseness at all might seem beside the point. DCD as a criterion of reverse inference cares only about predictive success, not other epistemic traits of models. Once we no longer seek to map cognitive functions onto regions in a way that respects their underlying causal organization, there is no added evidential value in the mere fact that a weight map is sparsely interpretable, let alone stable.

For these purposes, decoding that is based on an unstable weight map or one that is hard to interpret may indeed be adequate. A more traditional concern for functional RI might lead us to have a different set of goals in mind, however, including the desire to *explain* how neural patterns realize cognitive processes. For these goals, interpretability and plausibility matter. Focusing attention on a sparse subset of regions is best understood as motivated by a search for neural structures that play the appropriate causal and explanatory roles. As we will see, though, even this goal often proves elusive.

4.3. The problem of causality

Suppose a classifier achieves what we regard as a good balance of accuracy and production of a reproducible and plausibly interpretable weight map. It is tempting to infer from such success to claims about causality and processing. Kriegeskorte and Douglas (2019), for instance, propose that classifiers can perform double duty as causal models: “if a decoder is used to predict behavioral responses, for example judgments of categorization or continuous stimulus variables… then the decoder can be interpreted as a model (at a high level of abstraction) of the brain computations generating the behavioral responses from the encoding of the stimuli in the decoded brain region” (p. 171).

However, decoders do not give us enough evidence to conclude that the predictively weighted regions *cause* behavioral effects. There are several reasons for this. One is that decoders have no inherent causal directionality built into them. Procedures to find the best boundary to enable pattern-to-label associations are agnostic on whether there are any causal relations between the two. This is easy to see once we step outside the domain of neural data, since classifiers are frequently used on datasets that have no such causal relations among features and labels. SVMs can be used to parse handwritten ZIP codes on envelopes, or for image analysis and facial recognition. Success in these contexts implies nothing about causal structure in the target materials. Even within neuroscience, it is common to train classifiers on multimodal data sets (combining imaging, MEG/EEG, and other physiological or clinical biomarkers) that do not have a clear joint causal interpretation of their features (Meng et al., 2017; Woo, Chang, Lindquist, & Wager, 2017).

Moreover, good predictors in machine learning tasks do not always overlap well with good targets of intervention (Athey, 2017). Consider a non-neural example. Marketing firms use machine learning tools to discover “high churn” customers—those that are likely to stop using a company's products or services. But the population of customers who respond well to interventions such as marketing appeals only overlaps by 50% with those who are in the predictively isolated high-churn group. So we can often know that churn will take place in a certain population without being able to use that information to intervene causally on it. The same holds for many neurodiagnostic classifications. A classifier might use anatomical features such as hippocampal volume or the presence of amyloid plaques to diagnose Alzheimer's disease, but neither of these is a *cause* of the disorder. Drugs targeting amyloid, in particular, have regularly failed to produce clinical improvements in DAT patients. Decoders in this case are not reliably tracking causes. This is a recurring problem across fields using similar classifier-based analyses, such as genome-wide association studies: significant variables are often not predictive, and vice-versa (Lo, Chernoff, Zheng, & Lo, 2015).

Indeed, decoders can just as easily operate in an *anti*-causal direction (Weichwald et al., 2015). Consider two experimental designs, one in which a stimulus is presented followed by BOLD imaging, and another in which BOLD imaging occurs prior to production of a behavioral or cognitive response. In a stimulus-first design, decoding the category of the stimulus operates *against* the direction of causation in the experiment. Clearly in this design we can't treat decoders as causal models. But we cannot do so even in cases where the direction of decoding and the direction of causation are consistent. The reason is that decoding weights are not designed to be measures of causal contribution. As noted above, some factors may result in weights being assigned to features that are not causes of a phenomenon, such as incidentally correlated noise within voxels. Some actual causes may even receive low or zero weights simply because they are not most useful for decoding purposes *in the context* of the other voxel weight assignments.

Neither can we generally regard classifiers as processing models. Decoding is typically presented as a way of extracting the information present in regional activation patterns. But within cognitive modeling, there is an important distinction between information and processes, as evinced by the fact that such models posit representation-process pairs that work together to execute cognitive functions. This is a fundamental commitment of cognitive modeling within the broadly Marrian tradition (Barsalou, 2017). Algorithmic cognitive models describe how representations are constructed, stored, and transformed in carrying out a cognitive operation. Task performance is a product of the joint operation of both factors (along with architectural facts such as resource constraints). The neural decodability of a distinction between two task conditions does not tell us whether this stems from representational differences, processing differences, demand characteristics and resource usage, or some combination of these. From the point of view of causal modeling, this simply amounts to conflating potentially separate contributions. The sort of information that decoding provides, then, does not inherently tell us about causal-explanatory structure, particularly as it relates to cognitive processing.

This point can be illustrated by studies that explicitly attempt to use classifiers to derive causal structure by using their performance as predictive of later events such as behavior. Consider a nicely designed set of studies by Grootswagers, Cichy, & Carlson (2018). They asked participants to view images and make binary categorization decisions about them, e.g., judging whether a banana was animate or inanimate. In the first analysis phase they used a SVM-based searchlight procedure to generate a map of regions whose activation predicted correct category decisions. The crucial step lies in the second analysis phase, which involved running another searchlight analysis in which a new SVM classifier was trained for each region and the distance of each presented exemplar from the classifier's decision hypersurface was computed. The logic behind this analysis stems from signal detection theory, which holds that an option's distance to a decision boundary is determined by the evidence; i.e., stronger evidence places items farther away in space. This, in turn, generates the prediction that items located close to the decision boundary should be ones for which there is relatively little evidence, or evidence of ambiguous quality. For items such as these, choice is more difficult. Finally, there is the assumption that choice difficulty is reflected linearly in decision time. The second map created depicts the averaged degree of negative correlation between distance to the classifier's decision hyperplane and RT.

Their key finding is that the two maps overlap somewhat, but not entirely: while animacy can be successfully decoded throughout the ventral visual processing stream, RT is predicted by only a subset of those regions, predominantly ones located in anterior ventral temporal cortex. This suggests that mere decodability does not imply that the information present in regional patterns is formatted correctly to be "read out" in behavior (see also Williams, Dang, & Kanwisher, 2007). To bridge this gap successfully requires analysis that systematically links the properties of classifiers with behavioral variables. Specifically, it depends on interpreting the formal structure of classifiers in terms of established computational theories (Ritchie & Carlson, 2016). Here the relevant theory is a distance-to-bound model of choice transposed to the neural domain. Applying this model turns essentially on giving explanatory significance to the distances defined by classifier hypersurfaces, since these are treated as both reflecting the processing of evidence and as affecting behavioral responses.

While the approach taken by Grootswagers et al. is promising, some caveats remain. Most importantly for present purposes, the success of this computational framework only drives home further the need to take seriously the model choice considerations raised in Sect. 4.2, particularly since different classifier structures may give rise to different RT predictions. If classifiers are to be treated as causal models of the computational processes that mediate behavior, they need to be both stable and interpretable. This, again, is just to emphasize that they need to be chosen with an eye towards facilitating functional RI.

5. Decoding as data exploration

The problems surveyed here converge on the following conclusions. In terms of our original distinction, classifiers can be extraordinarily useful tools for *predictive* reverse inference. For *functional* reverse inference—the discovery not only of neural activity that is indicative of cognitive processing, but also of a prospective implementation theory for that processing—their utility is significantly less clear. The reason is that they are driven, in an unknown proportion of cases, by factors besides the ground truth concerning what patterns of neural activation are causally and explanatorily responsible for the cognitive processing we are investigating. Disentangling genuinely explanatory factors from the rest is difficult given that classifiers inherently conflate them. Decoding, in short, allows reverse inference of an often opaque kind that does not suit all of our investigative ends equally well.[17]

In fact, MVPA itself is demonstrably not a panacea for the ills of localization-based reverse inference, since the same problem of multiple functional assignments can arise just as readily within it as in univariate analysis. To see this, consider a widely discussed study by Knops, Thiriel, Hubbard, Michel, and Dehaene (2009). In the first phase of their study they

[17] To be clear, the preceding arguments are obviously not meant as blanket condemnations of the use of MVPA and machine learning in neuroscience. The issue concerns only whether the successful use of ML-based decoding methods is sufficient for making reverse inferences.

scanned participants during a random left/right eye movement task and trained a classifier on a group of six pre-selected cortical ROIs. This classifier could decode direction of motion with ~70% accuracy across all participants. They then had the same participants perform a simple arithmetic task: either add or subtract two displayed numbers and choose the closest correct answer (out of seven choices). The classifier trained on activation patterns from the bilateral posterior superior parietal lobule (PSPL) was then applied, without alteration, to activation patterns in that region from the arithmetic task. The classifier succeeded ~55% of the time with addition mapped onto rightward eye motion and subtraction onto leftward motion (breakdown by condition was 61% correct for addition and 49% for subtraction).

Knops et al. concluded from the fact that the same classifier achieved predictive success on both datasets that that the PSPL is involved in computations underlying both L/R eye motion and addition/subtraction. But now we face exactly the problem of multifunctional regions again. The information present in PSPL might indicate rightward eye motion *or* (some unknown cognitive component of) mental addition—or, for that matter, some more abstract but unknown operation that is implicated in both of them. We are not appreciably closer to understanding what "the" function of PSLP is, except to say that it contains information that can contribute to this pattern of discriminative success across tasks. The classifier transfer paradigm, then, is not *in itself* an advance in understanding the cognitive processing that goes on in particular brain regions.

I should stress that this conclusion is one that del Pinal and Nathan might not object to. At one point they seem to reject the search for an explanatory implementation theory of the sort that functional RI is concerned with, arguing that rather than focusing on "how cognitive algorithms are neurally implemented", reverse inference should address only the question of “which cognitive processes are more or less likely to be engaged in certain tasks whose nature is under dispute” (Nathan & Del Pinal, 2017, p. 9). This approach is quite consistent with the predictive turn, although they don’t couch their claim in those terms. As Bzdok and Ionnadis note, “predictive approaches put less emphasis on mechanistic insight into the biological underpinnings of the coherent behavioral phenotype” (2019, p. 3). Shifting focus away from discovering the cognitive function of brain regions (or even distributed brain networks) is of a piece with this move from explanatory understanding towards successful prediction.

Supposing, however, that neuroscientists wish to retain explanation as an epistemic goal, how can we reconceive the role of experimental practices such as MVPA, given that decoding models are not *themselves* explanatory? The rhetoric surrounding prediction depicts it as competing with explanation, or at least on the opposite side of a continuum from it (Bzdok & Yeo, 2017). I suggest, to the contrary, that we view decoding models not as competing for epistemic real estate with causal-explanatory ones, but as cooperating as part of a modeling pipeline. Decoding results constitute both *heuristic input* to explanatory models as well as *constraints* on them.

Decoding is a useful heuristic insofar as it suggests a menu of possible sites for further investigation and intervention. Regions whose activity can be decoded to distinguish between presented visual objects can also be scrutinized for whether that activity predicts behavioral outcomes. There is no guarantee that it will, as shown by the Grootswagers et al. study discussed above, but this needs to be investigated using methods that are geared towards generating and testing potential causal explanations, not just predictive adequacy. Such regions can also be probed using other methods to uncover their operations. For example, regions that support decoding of information might exhibit repetition suppression for that same information (though see Ward, Chun, & Kuhl, 2013 for some doubts about this).

In a case where we know that decodable information can be correlated with behavioral or cognitive outcomes, we have the following constraint: for any region from which information can be read out, any account of that region's function should explain how the region can contribute towards producing the behavior in question. That is, decoding results help to establish and clarify the explanandum phenomena that characterize the regions targeted by explanatory modeling. Something like this provides a useful way to understand Representational Similarity Analysis (RSA), a procedure of correlating the geometry of the space of stimuli (such as pictures of artifacts or faces) with that of the activation space of a collection of voxels (Kriegeskorte, 2011; Kriegeskorte & Kievit, 2013). RSA returns a numerical measure (using, e.g., rank correlation) of the extent to which stimuli that are close together in visual similarity space remain so in activation space. As its name suggests, RSA is sometimes described as characterizing what a region represents. But as normally practiced it does not offer hypotheses about algorithmic-level vehicles or processes. Rather, it articulates abstract structures of correspondence between regions and stimuli or behaviors. These correspondences, often discovered through decoding studies, can be regarded as tentative functional assignments. In this way they become part of the phenomena to be fed into the explanatory modeling pipeline.

In short, the place of decoding models is not in competition with explanatory modeling, but prior to and in concert with it. This aligns with the guiding and constraining role that data modeling has often been assigned within a mechanistic framework (Bechtel & Abrahamsen, 2005; but see Burnston, 2016b for an alternate interpretation). Imaging data is noisy and complex. Machine learning tools provide one way of extracting useful patterns from this data, which can help to stabilize new phenomena and discover new explanatory targets. This dovetails nicely with one of the original functions for which linear classifiers were developed, namely the partitioning of large datasets according to the varieties of hidden information that they contain. As tools for simplifying and exploring neuroscientific datasets, they can contribute to explanatory modeling without displacing it.

Finally, with respect to the question of reverse inference, the mere existence of decoding differences between task conditions does not establish differences in the underlying cognitive processes. What it does, however, is provide a set of phenomena to be investigated further; specifically, it suggests a plausible hypothesis about the structured information that is robustly detectable (in the brain at large, in an ROI, or within a cluster of searchlights) and that can be connected with measurable outcomes. If decoding results are stable across a wide range of models, parameter settings, and training regimes, and if they are systematically connectable with cognitive or behavioral outcomes, then the most predictive interpretable regions these converging models pick out are plausible targets for explanatory modeling. MVPA achieves the role of potential evidence for or against cognitive hypotheses by playing a supporting (though not individually sufficient) role in this sort of data modeling pipeline.[18]

6. Conclusion

[18] This point is similar to Kriegeskorte & Douglas's (2019) warning against committing the *single-model-significance* fallacy: that is, assuming that because a model explains some significant variance that it *thereby* captures facts about processing or causal structure. To reach such conclusions we need to integrate information from many models operating over a wide range of training data and parameter settings. This many-model integration process is what I have referred to here as a modeling pipeline. This notion is also discussed at length by Wright (2018), who emphasizes that in practice multiple analyses of data make distinct contributions to the characterization of phenomena in neuroimaging.

I’ve argued that MVPA’s ability to make predictive inferences from activation patterns does not offer us a transparent interpretive window onto the ground truths that drive this success. This form of predictive modeling is useful not because it can serve as a replacement for explanatory modeling, but because, seen in the proper perspective, it is an essential complement to it. Techniques from data science have their natural home in the analysis and modeling of data, even when deployed within neuroscience. To the extent that neuroscience continues to import and adapt machine learning tools, with their associated epistemic focus on prediction over explanation, there may be strong temptations to focus on the success of these tools without inquiring into the underlying causal-explanatory facts that enable them to succeed or fail. This temptation is understandable, given their striking translational successes, but I’ve argued that giving in to it would be a mistake. We should welcome the return of prediction as an important scientific desideratum without granting it dominance over our epistemic regime.

Acknowledgments

Thanks to Nikolaus Kriegeskorte, Marco Nathan, J. Brendan Ritchie, and Jessey Wright for their thoughtful comments on a previous draft of this paper, which was presented as part of the Neural Mechanisms webinar in June, 2019. I’m grateful to Fabrizio Calzavarini and Marco Viola for generously inviting me to participate in this series.

References

Agrawal, A., Gans, J., & Goldfarb, A. (2018). *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston, MA: Harvard Business School Publishing.

Anderson, M. L. (2014). *After Phrenology*. Cambridge, MA: MIT Press.

Athey, S. (2017). Beyond prediction: Using big data for policy problems. *Science*, *355*, 483–485.

Baldassarre, L., Pontil, M., & Mourão-Miranda, J. (2017). Sparsity Is Better with Stability: Combining Accuracy and Stability for Model Selection in Brain Decoding. *Frontiers in Neuroscience, 11*. https://doi.org/10.3389/fnins.2017.00062

Barsalou, L. W. (2017). What does semantic tiling of the cortex tell us about semantics? *Neuropsychologia*, *105*, 18–38.

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A mechanist alternative. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *36*, 421–441.

Berkman, E. T., & Falk, E. B. (2013). Beyond Brain Mapping: Using Neural Measures to Predict Real-World Outcomes. *Current Directions in Psychological Science*, *22*, 45–50.

Burnston, D. C. (2016a). A contextualist approach to functional localization in the brain. *Biology & Philosophy*, *31*, 527–550.

Burnston, D. C. (2016b). Data graphs and mechanistic explanation. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *57,* 1–12.

Bzdok, D., & Ioannidis, J. P. A. (2019). Exploration, Inference, and Prediction in Neuroscience and Biomedicine. *Trends in Neurosciences, 42*, 251–262.

Bzdok, D., & Yeo, B. T. T. (2017). Inference in the age of big data: Future perspectives on neuroscience. *NeuroImage*, *155,* 549–564.

Coltheart, M. (2006). What has functional neuroimaging told us about the mind (so far)? *Cortex*, *42*, 323–331.

Coltheart, M. (2013). How Can Functional Neuroimaging Inform Cognitive Theories? *Perspectives on Psychological Science*, *8*, 98–103.

Cox, D. D., & 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.

Davies, M. (2010). Double Dissociation: Understanding its Role in Cognitive Neuropsychology. *Mind & Language*, *25*, 500–540.

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

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic Bulletin & Review*, *23*, 1415–1428.

Del Pinal, G., & Nathan, M. J. (2017). Two Kinds of Reverse Inference in Cognitive Neuroscience. In J. Leefman & E. Hildt (Eds.), *The Human Sciences after the Decade of the Brain* (pp. 121–139). Elsevier.

Douglas, H. E. (2009). Reintroducing Prediction to Explanation. *Philosophy of Science*, *76*, 444–463.

Douglas, P. K., & Anderson, A. (2017). Interpreting fMRI Decoding Weights: Additional Considerations. *31st Conference on Neural Information Processing Systems (NIPS 2017)*, 1–7.

Dubois, J., de Berker, A. O., & Tsao, D. Y. (2015). Single-Unit Recordings in the Macaque Face Patch System Reveal Limitations of fMRI MVPA. *Journal of Neuroscience*, *35*, 2791–2802.

Etzel, J. A., Zacks, J. M., & Braver, T. S. (2013). Searchlight analysis: Promise, pitfalls, and potential. *NeuroImage*, *78*, 261–269.

Falk, E. B., Berkman, E. T., & Lieberman, M. D. (2012). From Neural Responses to Population Behavior: Neural Focus Group Predicts Population-Level Media Effects. *Psychological Science*, *23*, 439–445.

Gelman, A., & Loken, E. (2014). The Statistical Crisis in Science. *American Scientist*, *102*, 460–465.

Genevsky, A., Yoon, C., & Knutson, B. (2017). When Brain Beats Behavior: Neuroforecasting Crowdfunding Outcomes. *The Journal of Neuroscience*, *37*, 8625–8634.

Glymour, C., & Hanson, C. (2016). Reverse Inference in Neuropsychology. *The British Journal for the Philosophy of Science*, *67*, 1139–1153.

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

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

Haxby, J. V. (2001). Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science*, *293*, 2425–2430.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding Neural Representational Spaces Using Multivariate Pattern Analysis. *Annual Review of Neuroscience*, *37*, 435–456.

Haynes, J.-D. (2012). Brain reading. In S. Richmond, G. Rees, & S. Edwards (Eds.), *I know what you're thinking: Brain imaging and mental privacy* (pp. 29–40). Oxford: Oxford University Press.

Haynes, J.-D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, *87*, 257–270.

Haynes, J.-D., & Rees, G. (2005). Predicting the Stream of Consciousness from Activity in Human Visual Cortex. *Current Biology, 15*, 1301–1307.

Haynes, J.-D., Sakai, K., Rees, G., Gilbert, S., Frith, C., & Passingham, R. E. (2007). Reading Hidden Intentions in the Human Brain. *Current Biology*, *17*, 323–328.

Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *NeuroImage, 180*, 4–18.

Henson, R. (2006). Forward inference using functional neuroimaging: Dissociations versus associations. *Trends in Cognitive Sciences*, *10*, 64–69.

Hu, L., & Iannetti, G. D. (2016). Painful Issues in Pain Prediction. *Trends in Neurosciences*, *39*, 212–220.

Hutzler, F. (2014). Reverse inference is not a fallacy per se: Cognitive processes can be inferred from functional imaging data. *NeuroImage*, *84*, 1061–1069.

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

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

Klein, C. (2012). Cognitive Ontology and Region- versus Network-Oriented Analyses. *Philosophy of Science*, *79*, 952–960.

Knops, A., Thirion, B., Hubbard, E. M., Michel, V., & Dehaene, S. (2009). Recruitment of an Area Involved in Eye Movements During Mental Arithmetic. *Science*, *324*, 1583–1585.

Kragel, P. A., Koban, L., Barrett, L. F., & Wager, T. D. (2018). Representation, Pattern Information, and Brain Signatures: From Neurons to Neuroimaging. *Neuron*, *99*, 257–273.

Kriegeskorte, N. (2011). Pattern-information analysis: From stimulus decoding to computational-model testing. *NeuroImage*, *56*, 411–421.

Kriegeskorte, N., & Bandettini, P. (2007a). Analyzing for information, not activation, to exploit high-resolution fMRI. *NeuroImage*, *38*, 649–662.

Kriegeskorte, N., & Bandettini, P. (2007b). Combining the tools: Activation- and information-based fMRI analysis. *NeuroImage*, *38*, 666–668.

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

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, *103*, 3863–3868.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*, 401–412.

Lo, A., Chernoff, H., Zheng, T., & Lo, S.-H. (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences*, *112*, 13892–13897.

Machery, E. (2014). In Defense of Reverse Inference. *The British Journal for the Philosophy of Science*, *65*, 251–267.

McCaffrey, J. B. (2015). The Brain's Heterogeneous Functional Landscape. *Philosophy of Science*, *82*, 1010–1022.

Meng, X., Jiang, R., Lin, D., Bustillo, J., Jones, T., Chen, J., … Calhoun, V. D. (2017). Predicting individualized clinical measures by a generalized prediction framework and multimodal fusion of MRI data. *NeuroImage*, *145*, 218–229.

Nathan, M. J., & Del Pinal, G. (2017). The Future of Cognitive Neuroscience? Reverse Inference in Focus. *Philosophy Compass*, *12*, 1–11.

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

Northcott, R. (2017). When are Purely Predictive Models Best? *Disputatio*, *9*, 631–656.

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

Poldrack, R. A. (2018). *The New Mind Readers: What Neuroimaging Can and Cannot Reveal about Our Thoughts*. Princeton, NJ: Princeton University Press.

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

Rasmussen, P. M., Hansen, L. K., Madsen, K. H., Churchill, N. W., & Strother, S. C. (2012). Model sparsity and brain pattern interpretation of classification models in neuroimaging. *Pattern Recognition, 45*, 2085–2100.

Rathkopf, C. A. (2013). Localization and Intrinsic Function. *Philosophy of Science, 80,* 1–21.

Ritchie, J. B., & Carlson, T. A. (2016). Neural Decoding and “Inner” Psychophysics: A Distance-to-Bound Approach for Linking Mind, Brain, and Behavior. *Frontiers in Neuroscience, 10*. https://doi.org/10.3389/fnins.2016.00190

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the Brain: Neural Representation and the Limits of Multivariate Pattern Analysis in Cognitive Neuroscience. *British Journal for the Philosophy of Science*, *70,* 581–607.

Roskies, A. (2009). Brain-Mind and Structure-Function Relationships: A Methodological Response to Coltheart. *Philosophy of Science*, *76*, 1–14.

Schrouff, J., & Mourao-Miranda, J. (2018). Interpreting weight maps in terms of cognitive or clinical neuroscience: Nonsense? *2018 International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 1–4. Singapore: IEEE.

Soon, C. S., He, A. H., Bode, S., & Haynes, J.-D. (2013). Predicting free choices for abstract intentions. *Proceedings of the National Academy of Sciences*, *110,* 6217–6222.

Tong, F., & Pratte, M. S. (2012). Decoding Patterns of Human Brain Activity. *Annual Review of Psychology*, *63*, 483–509.

Van Horn, J. D., & Toga, A. W. (2014). Human neuroimaging as a “Big Data” science. *Brain Imaging and Behavior*, *8,* 323–331.

Varoquaux, G., & Poldrack, R. A. (2019). Predictive models avoid excessive reductionism in cognitive neuroimaging. *Current Opinion in Neurobiology*, *55*, 1–6.

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

Varoquaux, G., & Thirion, B. (2014). How machine learning is shaping cognitive neuroimaging. *GigaScience*, *3*, 1–7.

Vickery, T. J., Chun, M. M., & Lee, D. (2011). Ubiquity and Specificity of Reinforcement Signals throughout the Human Brain. *Neuron*, *72*, 166–177.

Wager, T. D., Atlas, L. Y., Lindquist, M. A., Roy, M., Woo, C.-W., & Kross, E. (2013). An fMRI-Based Neurologic Signature of Physical Pain. *New England Journal of Medicine*, *368*, 1388–1397.

Ward, E. J., Chun, M. M., & Kuhl, B. A. (2013). Repetition Suppression and Multi-Voxel Pattern Similarity Differentially Track Implicit and Explicit Visual Memory. *Journal of Neuroscience*, *33*, 14749–14757.

Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., & Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. *NeuroImage*, *110*, 48–59.

Weiskopf, D. A. (2016). Integrative modeling and the role of neural constraints. *Philosophy of Science*, *83*, 674–685.

Williams, M. A., Dang, S., & Kanwisher, N. G. (2007). Only some spatial patterns of fMRI response are read out in task performance. *Nature Neuroscience*, *10*, 685–686.

Woo, C.-W., Chang, L. J., Lindquist, M. A., & Wager, T. D. (2017). Building better biomarkers: Brain models in translational neuroimaging. *Nature Neuroscience*, *20*, 365–377.

Wright, J. (2018). The Analysis of Data and the Evidential Scope of Neuroimaging Results. *British Journal of the Philosophy of Science*, *69*, 1179–1203.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science, 12*, 1100–1122.