Painful issues in pain prediction

L. Hu[a,b,c], G.D. Iannetti[b]

[a] *Institute of Psychology, Chinese Academy of Sciences, Beijing, China;* [b] *Department of Neuroscience, Physiology and Pharmacology, University College London, UK;* [c] *Faculty of Psychology, Southwest University, Chongqing, China.*

*Correspondence: hulitju@gmail.com (L. Hu) or g.iannetti@ucl.ac.uk (G.D. Iannetti).

Keywords: Pain, Machine learning, Prediction, Functional magnetic resonance imaging (fMRI), Pain signature, Multivariate pattern analysis (MVPA).

Abstract: How perception of pain emerges from neural activity is largely unknown. Identifying a neural “pain signature” and deriving a way to predict perceived pain from brain activity would have enormous basic and clinical implications. Researchers are increasingly turning to functional brain imaging, often applying machine-learning algorithms to infer that pain perception occurred. Yet such sophisticated analyses are fraught with interpretive difficulties. Here we highlight some common and troublesome problems in the literature, and suggest methods to ensure researchers draw accurate conclusions from their results. Since functional brain imaging is increasingly finding practical applications with real-world consequences, it is critical to interpret brain scans accurately, as decisions based on neural data will only be as good as the science behind them.

Machine learning in pain research: objectives and protocols

Pain, as any other conscious sensation, is determined by a specific pattern of neural activity at the cortical level [1, 2]. To understand the perception of pain, many researchers use non-invasive functional neuroimaging techniques [3, 4], such as electroencephalography (EEG), magnetoencephalography (MEG), positron emission tomography (PET), and, especially, functional magnetic resonance imaging (fMRI). With these tools, researchers can now attempt to achieve the following key objectives. (1) Identify temporal and spatial patterns of neural activity that could serve as a cortical signature for human pain perception [5-8]. (2) Establish whether these patterns, or any other physiological measures of brain activity, can be used to reliably predict perceived pain [7, 9-14]. Achieving these objectives, which would have dramatic basic and clinical implications, is increasingly attempted through the application of sophisticated machine-learning algorithms to interpret functional brain imaging data [15-18]. However, correct interpretation requires proper protocol design and careful inferences. Here we point out some of the pitfalls of applying machine-learning techniques to functional brain imaging data related to pain perception, especially in light of recent divergent conclusions in the literature, and suggest possible remedies.

Machine learning is a scientific discipline exploiting algorithms that can learn from and make predictions on data [19-21]. When applied to functional brain

imaging data, machine learning has the potential (1) to identify response features that specifically *encode* a given experimental variable (e.g., the categories of visual objects [22]), and (2) to *decode* measured data to predict subjective percepts and intentions (e.g., the pain intensity reported by an individual [9]) (see Glossary and Box 1). Therefore, it is not surprising that machine learning has received immense interest in systems neuroscience, and it is now increasingly used in the field of human pain [7, 9-14, 23, 24].

While machine-learning techniques hold considerable promise for pain research, investigators must take special care to match machine-learning protocol design to the desired study objectives. Disregarding the tight relationship between protocol and objective can lead to inaccurate interpretation of results. In this article we explain how incorrect conclusions can result when deviating from a given machine-learning protocol's allowable objective. We first outline the two main objectives of machine learning in pain neuroscience. We then clarify some issues related to result interpretation, and finally provide guidelines to avoid unjustified claims.

Objective 1: Identifying a pain-specific neural signature

A main objective of machine learning is to identify a "neural signature" or "fingerprint", i.e., a neural correlate of fMRI activity that *uniquely* encodes a given experimental variable or perceptual experience [25, 26] (Box 1). This is

an extremely appealing objective in human pain neuroscience, given that the amplitude of the fMRI signal, when analysed with traditional mass-univariate analysis (i.e., general linear modeling, GLM [27, 28]), has failed to identify a unique signature for pain [29]. Indeed, transient painful stimuli elicit graded responses within a wide array of brain regions (which has been sometimes unfoundedly labelled as "pain matrix"), consistently including the primary and secondary somatosensory cortices (S1 and S2), the insula, and the anterior cingulate cortex (ACC) [30-33]. However, most of these areas are also activated by equally salient but never painful auditory, tactile, and visual stimuli [29, 34]. Given that these brain regions are also activated in situations where no pain is present, it is an incorrect reverse inference to conclude that this pattern of brain activation represents a pain signature [35-38].

Machine learning potentially offers a way forward, so long as the proper protocol is applied. Like traditional mass-univariate analysis, machine learning can exploit similar features of the functional neuroimaging response, such as spatial distribution and signal amplitude [39]. Yet, if machine learning simply exploits bulk differences in signal amplitude to successfully identify a given experimental variable (i.e., the perceived pain intensity), this does not reflect a *unique* pain signature, and the same problem of reverse inference applies to the interpretation of results [35]. Just like in mass-univariate analysis, it is valid

to interpret a given result as a "pain signature" *if and only if* the relationship between the brain response pattern and pain is unique for pain.

To overcome this issue, machine learning should be performed using a protocol that identifies the possible relationship between fine-grained spatial patterns of the brain response and pain (in this case machine learning is named multi-voxel pattern analysis, MVPA [40, 41]) *without making use of signal amplitude*. In addition, the specificity of a possible fine-grained spatial pattern should be verified against the brain responses elicited by non-painful but iso-salient stimuli, to rule out the possibility that the same spatial patterns could reflect equally salient stimuli of different sensory modalities. If these prerequisites are not satisfied, machine learning is no better than mass-univariate analysis, and the correct classification would be misinterpreted as a specific neural signature for pain.

Objective 2: Pain prediction from neural activity

When the objective is instead to *decode* a laboratory measure of brain activity to predict a subjective painful percept (see Box 1), machine learning can be performed using a protocol that exploits *all* signal components encoding the subjective percept (typically pain intensity, but also different qualities of pain). Therefore, both the amplitude and the spatial configuration of the signal can be preserved, as they both have the potential of encoding the reported pain

intensity. In particular, the amplitude information should be kept and exploited, given that this information is often, albeit not always, correlated well with subjective pain intensity [42-44]. Indeed, and rightly so, all studies using machine learning with the objective of *predicting* pain perception take advantage of the variability in signal amplitude [7, 9-14, 24]. It is important to note that, for the practical objective of predicting pain, the reverse inference issue highlighted in the previous section is less important (see Glossary). Indeed, even if some (or all) features of the signal exploited to predict pain are not pain specific, but a good prediction is achieved, this can still be useful. Of course, a practically important point is to estimate *how often* those features (despite not representing a unique pain signature) allow machine learning to predict pain. Indeed, most of the features that have been used to successfully predict pain (i.e., bulk signal changes in several brain regions [7, 9]) are likely to fail to predict pain in some contexts. For example, failure is likely when pain intensity is dissociated from stimulus saliency, given that it has been shown that these signal changes are also determined by iso-salient, but non-painful, sensory stimuli [29, 34, 44].

In the following sections we suggest some guidelines to improve the use of machine learning in interpreting fMRI data. We will describe in detail key aspects of the analytical steps needed to use machine learning in relation to the two objectives outlined above. The choices for each analytical step define

the potentially achievable objectives, as well as the physiological conclusions that can be inferred.

Signal normalization

As detailed earlier, the response amplitude of fMRI signal in regions of the so-called "pain matrix", although often correlated with the intensity of perceived pain, is largely not specific for pain, as non-painful stimuli can also elicit graded brain responses that correlate with intensity of perception [29]. Therefore, if successful machine learning relies on graded levels of response amplitude, the reverse inference that these features reflect a unique "pain signature" (see *Objective 1*) is unlikely to be correct. Implementing a strict normalization of fMRI signal amplitude is a possible strategy to minimize the contribution of graded levels of activation to successful machine learning (Fig. 1), and therefore increase the likelihood that the features exploited by machine learning represent a unique "pain signature" (see *Objective 1*). The amplitude of the brain activity at each time point can be normalized across a number of voxels, by subtracting from the signal of each voxel the mean signal across all voxels of a given region-of-interest (ROI) or the entire brain, and then dividing the result by the standard deviation of the signal from all voxels of the ROI (or the entire brain). As a result of this procedure, in each experimental condition the voxels constituting the ROI have a mean of zero and a standard deviation of 1.

This normalization strategy minimizes the contribution of non-pain-specific graded levels of activation, and should therefore be performed when aiming to identify a unique pain-specific spatial signature that cannot be disclosed using the mass-univariate analysis (*Objective 1*). In contrast, stimulus-evoked changes in signal amplitude can be preserved when aiming to predict subjective pain intensity (*Objective 2*), as perceived pain often correlates with signal amplitude, and therefore removing it usually entails a reduction of the accuracy of decoding. Exactly for this reason, studies aiming to predict pain avoid such a normalization step to maximize the predictive accuracy of the machine learning algorithm [9-12, 14]. An important note of caution is that successful pain predictions obtained when machine learning makes use of bulk signal amplitude likely exploit non-pain-specific neural responses [7, 9].

Within-subject vs between-subject prediction?

To achieve encoding objectives (i.e., identifying a pain signature), machine-learning analyses should be primarily performed within subjects, while to achieve decoding objectives (i.e., predicting pain), analyses should be primarily performed between subjects (Fig. 2). Indeed, to identify a *fine-grained* signature (using, for example, MVPA of fMRI signals), within-subject analyses will avoid the inevitable spatial blurring of responses caused by (1) the functional and anatomical differences between individuals

[45], and (2) the lack of optimal algorithms to co-register brains from different individuals [40, 41]. If performed at the between-subject level, any possible signature would be identified at least at the higher, mesoscopic scale of entire portions of brain regions. In contrast, machine learning for pain prediction is mostly performed between subjects, because, in practical applications, pain has to be predicted on new subjects, like a patient just after hospital admission, or a healthy participant in a drug trial [9, 24]. Machine learning for pain prediction can be also performed at within-subject level. Obviously, the usefulness of within-subject prediction is more limited, and the accuracy of such prediction is higher, as it is not affected by between-subject variability of the response features used to predict pain [9].

Use of prior knowledge when validating prediction performance

In basic and clinical applications of pain prediction, the quality or the intensity of subjective painful percepts is an unknown variable. Obviously, to predict unknown experimental variables, the use of prior knowledge about which variable each trial belongs to is only allowed when training the machine-learning model, but not when testing its prediction performance (Fig. 3) [19]. Therefore, the prediction performance of machine learning models should be validated strictly without using prior knowledge about those percepts. This important requirement is satisfied only when the prediction is performed at trial-by-trial level (called *Predicting* in Fig. 3) [10, 11]. However, in some studies

of pain prediction [7, 46], trials belonging to the same experimental condition (e.g., stimulus energy) were preliminarily averaged, and both training the prediction model and testing its performance were performed using averaged brain responses with increased signal-to-noise ratios. This strategy (called *Labeling* in Fig. 3) erroneously uses prior knowledge when testing the model's prediction performance, resulting in seemingly high accuracy of "pain prediction" (corresponding to extremely high sensitivity and specificity; e.g., Table 1 and Fig. 1 in [7]). The resulting "prediction" accuracy is not only artificially inflated, but also does not reflect the real prediction of an unknown pain level.

This is a crucial point. Indeed, the use of the prior knowledge in model testing artificially inflates the prediction accuracy, and therefore violates a fundamental rule when machine learning is used to predict a stimulus feature or a perceptual outcome (*Objective 2*) [19]. In contrast, when machine learning aims to identify a spatial signature that encodes a given experimental variable (*Objective 1*), it is acceptable to use prior knowledge about which experimental variable (e.g., reported subjective percept) each single trial belongs to when testing the model's prediction performance [40, 47]. Therefore, although incorrect for decoding objectives such as pain prediction, testing a model's prediction performance on trials averaged based on prior knowledge (as

previously done using stimulus energy [7]) makes sense for encoding objectives, such as identifying a new condition-specific spatial signature.

Conclusion and implications in the assessment of previous studies

Machine learning is extremely promising in pain research because it can identify response features that cannot be detected using mass-univariate analyses [23]. However, simply using machine-learning algorithms is not sufficient; the protocols must match the objectives to avoid erroneous conclusions. For example, given that machine learning can also exploit bulk differences in response amplitude, when these differences are not removed, a successful classification could simply rely on the same information identified by mass-univariate analyses [9]. This is acceptable if machine learning aims to predict pain (see *Objective 2*), but it represents a significant issue if machine learning aims to identify a *unique* signature for pain (see *Objective 1*).

Indeed, the validity issues of reverse inferences made from mass-univariate analyses of pain neuroimaging data [35, 36, 48] equally applies to the interpretation of the results obtained using machine learning. A given machine-learning result can be interpreted as reflecting a “pain signature” (Objective 1) *if and only* if the relationship between the brain response pattern and pain is unique for pain.

The conclusions we draw here warrant a more careful assessment of the interpretations of some recent machine-learning results in pain neuroscience [7, 46, 49]. Indeed, one particular study used a single, mixed machine-learning protocol: machine learning was performed on non-normalized fMRI data, at between-subject level, and making use of prior knowledge when estimating the prediction accuracy [7]. Using this approach the authors claimed to have achieved the two objectives of machine learning *together*. Indeed, they affirmed to have identified (1) a specific neurological pain signature ("NPS") relying on fine-grained spatial scales, which (2) can "reliably predict pain across different experiments" with extremely high accuracy.

However, the claim of having discovered a *unique* NPS that relies on fine-grained spatial scales is not entirely justified, as the employed machine-learning protocol violates the requirements needed to identify a *unique* brain signature of pain (see sections *Signal normalization* and *Within-subject vs between-subject prediction?*) [7, 46]. Furthermore, the seemingly impressive pain prediction accuracy was obtained by making use of prior knowledge when decoding the brain responses, a procedure that is incorrect when aiming to predict unknown experimental variables (see section *Use of prior knowledge when validating prediction performance*).

Such sweeping conclusions were only possible by incorrectly conflating encoding *(Objective* 1) vs. decoding *(Objective* 2) protocols, which must be applied separately to achieve those objectives (Box 1). Machine learning is a promising tool, but only by careful application one can take advantage of its full power to advance pain research (see Outstanding Questions). The stakes are high: functional brain imaging is increasingly finding practical applications with real-world consequences [49]. A neural "pain signature" could potentially serve as a biomarker for drug development, as evidence for pain perception in minimally conscious patients (or other patients that cannot report pain, such as infants [50]), or as an objective measure of pain to be used in legal cases. It is therefore critical to interpret brain scans accurately, as decisions based on neural data will only be as good as the science behind them.

Acknowledgements: LH is supported by the National Natural Science Foundation of China (31471082). GDI acknowledges the generous support of The Royal Society (for the experimental work that has provided the foundation of this article) and of The Wellcome Trust (COLL JLARAXR). We wish to thank all members of our research group for insightful comments on earlier versions of this manuscript.

Competing Financial Interests: The authors declare no competing financial interests.

References

1 Melzack, R. (1990) Phantom limbs and the concept of a neuromatrix. *Trends Neurosci* 13, 88-92

2 Tracey, I. (2011) Can neuroimaging studies identify pain endophenotypes in humans? *Nat Rev Neurol* 7, 173-181

3 Apkarian, A.V., *et al.* (2005) Human brain mechanisms of pain perception and regulation in health and disease. *Eur J Pain* 9, 463-484

4 Apkarian, A.V. (2015) *The Brain Adapting with Pain: Contribution of Neuroimaging Technology to Pain Mechanisms*. Wolters Kluwer Health

5 Schulz, E., *et al.* (2011) Neurophysiological coding of traits and states in the perception of pain. *Cereb Cortex* 21, 2408-2414

6 Gross, J., *et al.* (2007) Gamma oscillations in human primary somatosensory cortex reflect pain perception. *Plos Biology* 5, 1168-1173

7 Wager, T.D., *et al.* (2013) An fMRI-based neurologic signature of physical pain. *N Engl J Med* 368, 1388-1397

8 Kucyi, A. and Davis, K.D. (2015) The dynamic pain connectome. *Trends Neurosci* 38, 86-95

9 Huang, G., *et al.* (2013) A novel approach to predict subjective pain perception from single-trial laser-evoked potentials. *NeuroImage* 81, 283-293

10 Marquand, A., *et al.* (2010) Quantitative prediction of subjective pain intensity from whole-brain fMRI data using Gaussian processes. *NeuroImage* 49, 2178-2189

11 Brown, J.E., *et al.* (2011) Towards a physiology-based measure of pain: patterns of human brain activity distinguish painful from non-painful thermal stimulation. *PLoS ONE* 6, e24124

12 Prato, M.*, et al.* (2011) A regularization algorithm for decoding perceptual temporal profiles from fMRI data. *NeuroImage* 56, 258-267

13 Brodersen, K.H.*, et al.* (2012) Decoding the perception of pain from fMRI using multivariate pattern analysis. *NeuroImage* 63, 1162-1170

14 Cecchi, G.A.*, et al.* (2012) Predictive dynamics of human pain perception. *PLoS Comput Biol* 8, e1002719

15 Lindquist, M.A.*, et al.* (2015) Group-regularized individual prediction: theory and application to pain. *NeuroImage* DOI:10.1016/j.neuroimage.2015.10.074

16 Naselaris, T. and Kay, K.N. (2015) Resolving Ambiguities of MVPA Using Explicit Models of Representation. *Trends Cogn Sci* 19, 551-554

17 Haxby, J.V.*, et al.* (2014) Decoding neural representational spaces using multivariate pattern analysis. *Annu Rev Neurosci* 37, 435-456

18 Haxby, J.V. (2012) Multivariate pattern analysis of fMRI: the early beginnings. *NeuroImage* 62, 852-855

19 Bishop, C.M. (2006) *Pattern recognition and machine learning*. Springer

20 Pereira, F.*, et al.* (2009) Machine learning classifiers and fMRI: a tutorial overview. *NeuroImage* 45, S199-209

21 Wang, Z.*, et al.* (2007) Support vector machine learning-based fMRI data group analysis. *NeuroImage* 36, 1139-1151

22 Haxby, J.V.*, et al.* (2001) Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science* 293, 2425-2430

23 Rosa, M.J. and Seymour, B. (2014) Decoding the matrix: Benefits and limitations of applying machine learning algorithms to pain neuroimaging. *Pain* 155, 864-867

24 Schulz, E.*, et al.* (2012) Decoding an Individual's Sensitivity to Pain from the Multivariate Analysis of EEG Data. *Cereb Cortex* 22,1118-1123

25 Norman, K.A., *et al.* (2006) Beyond mind-reading: multi-voxel pattern analysis of fMRI data. *Trends Cogn Sci* 10, 424-430

26 Friston, K.J. (2009) Modalities, modes, and models in functional neuroimaging. *Science* 326, 399-403

27 Friston, K.J. (2007) *Statistical parametric mapping : the analysis of functional brain images*. Elsevier/Academic Press

28 Friston, K.J., *et al.* (1994) Statistical parametric maps in functional imaging: a general linear approach. *Hum Brain Mapp* 2, 189-210

29 Mouraux, A., *et al.* (2011) A multisensory investigation of the functional significance of the "pain matrix". *NeuroImage* 54, 2237-2249

30 Talbot, J.D., *et al.* (1991) Multiple representations of pain in human cerebral cortex. *Science* 251, 1355-1358

31 Bushnell, M.C., *et al.* (1999) Pain perception: is there a role for primary somatosensory cortex? *Proc Natl Acad Sci U S A* 96, 7705-7709

32 Vogt, B.A. (2005) Pain and emotion interactions in subregions of the cingulate gyrus. *Nat Rev Neurosci* 6, 533-544

33 Garcia-Larrea, L., *et al.* (2003) Brain generators of laser-evoked potentials: from dipoles to functional significance. *Neurophysiol Clin* 33, 279-292

34 Mouraux, A. and Iannetti, G.D. (2009) Nociceptive laser-evoked brain potentials do not reflect nociceptive-specific neural activity. *J Neurophysiol* 101, 3258-3269

35 Iannetti, G.D., *et al.* (2013) Beyond metaphor: contrasting mechanisms of social and physical pain. *Trends Cogn Sci* 17, 371-378

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

37 Legrain, V., *et al.* (2011) The pain matrix reloaded: a salience detection

system for the body. *Prog Neurobiol* 93, 111-124

38 Iannetti, G.D. and Mouraux, A. (2010) From the neuromatrix to the pain matrix (and back). *Exp Brain Res* 205, 1-12

39 Davis, T.*, et al.* (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

40 Haynes, J.D. and Rees, G. (2006) Decoding mental states from brain activity in humans. *Nat Rev Neurosci* 7, 523-534

41 Mur, M.*, et al.* (2009) Revealing representational content with pattern-information fMRI--an introductory guide. *Soc Cogn Affect Neurosci* 4, 101-109

42 Zhang, Z.G.*, et al.* (2012) Gamma-band oscillations in the primary somatosensory cortex--a direct and obligatory correlate of subjective pain intensity. *J Neurosci* 32, 7429-7438

43 Coghill, R.C.*, et al.* (1999) Pain intensity processing within the human brain: a bilateral, distributed mechanism. *J Neurophysiol* 82, 1934-1943

44 Iannetti, G.D.*, et al.* (2008) Determinants of laser-evoked EEG responses: pain perception or stimulus saliency? *J Neurophysiol* 100, 815-828

45 Fischl, B.*, et al.* (1999) High-resolution intersubject averaging and a coordinate system for the cortical surface. *Hum Brain Mapp* 8, 272-284

46 Woo, C.W.*, et al.* (2014) Separate neural representations for physical pain and social rejection. *Nat Commun* 5, 5380

47 Liang, M.*, et al.* (2013) Primary sensory cortices contain distinguishable spatial patterns of activity for each sense. *Nat Commun* 4, 1979

48 Poldrack, R.A. (2011) Inferring mental states from neuroimaging data: from reverse inference to large-scale decoding. *Neuron* 72, 692-697

49 Reardon, S. (2015) Neuroscience in court: The painful truth. *Nature* 518, 474-476

50 Goksan, S., *et al.* (2015) fMRI reveals neural activity overlap between adult and infant pain. *eLife* 4, e06356

Figure legends

Figure 1. Effects of signal normalization on spatial and amplitude differences in brain activation.

Normalization of fMRI signal is achieved by (1) subtracting from the signal of each voxel the mean signal across all voxels of a given ROI (or the entire brain), and (2) dividing the result by the standard deviation of the signal from all voxels of the ROI (or the entire brain). Before signal normalization (*top panel*), brain activity in different experimental conditions could differ in either signal amplitude (*left column*), spatial distribution (*middle column*), or both (*right column*). After signal normalization (*bottom panel*), brain activity mainly differs in its spatial distribution.

Figure 2. Comparison of within-subject and between-subject machine-learning protocols.

Left panel: Within-subject machine learning. The machine-learning model is trained on all trials except one (n-1), and tested on the remaining trial. The model is cross-validated using each trial as test trial once. Within-subject machine learning classifies the test trial into category A or B based on a model generated from the same subject. *Right panel*: Between-subject machine learning. The machine-learning model is trained on all trials of all subjects

except one (N-1), and tested on all trials of the remaining subject. Cross-validation is achieved by using each subject as test subject once. Between-subject machine learning classifies each single trial of the test subject into category A or B based on a model generated from the other subjects.

Figure 3. Predicting vs labeling: use of prior knowledge in machine learning.

At between-subject level, the machine-learning model is trained on all trials of all subjects except one (N-1; *top panel*), and tested on all trials of the remaining subject (*bottom panels*). Importantly, *predicting* the experimental variables A or B (*bottom left panel*) is achieved by classifying *each single trial* of the test subject into category A or B based on the trained model. Predicting does not exploit prior knowledge. In contrast, *labeling* is achieved by classifying two (or more) pre-defined groups (e.g., category A or B). Labeling uses prior knowledge about the experimental variable of interest, and typically results in higher accuracy than predicting (e.g., 100 vs. 72.5%). Such prior knowledge is obviously unavailable in most practical applications of machine learning for pain prediction.

Figure 1

Click here to download Figure Figure-1.tif

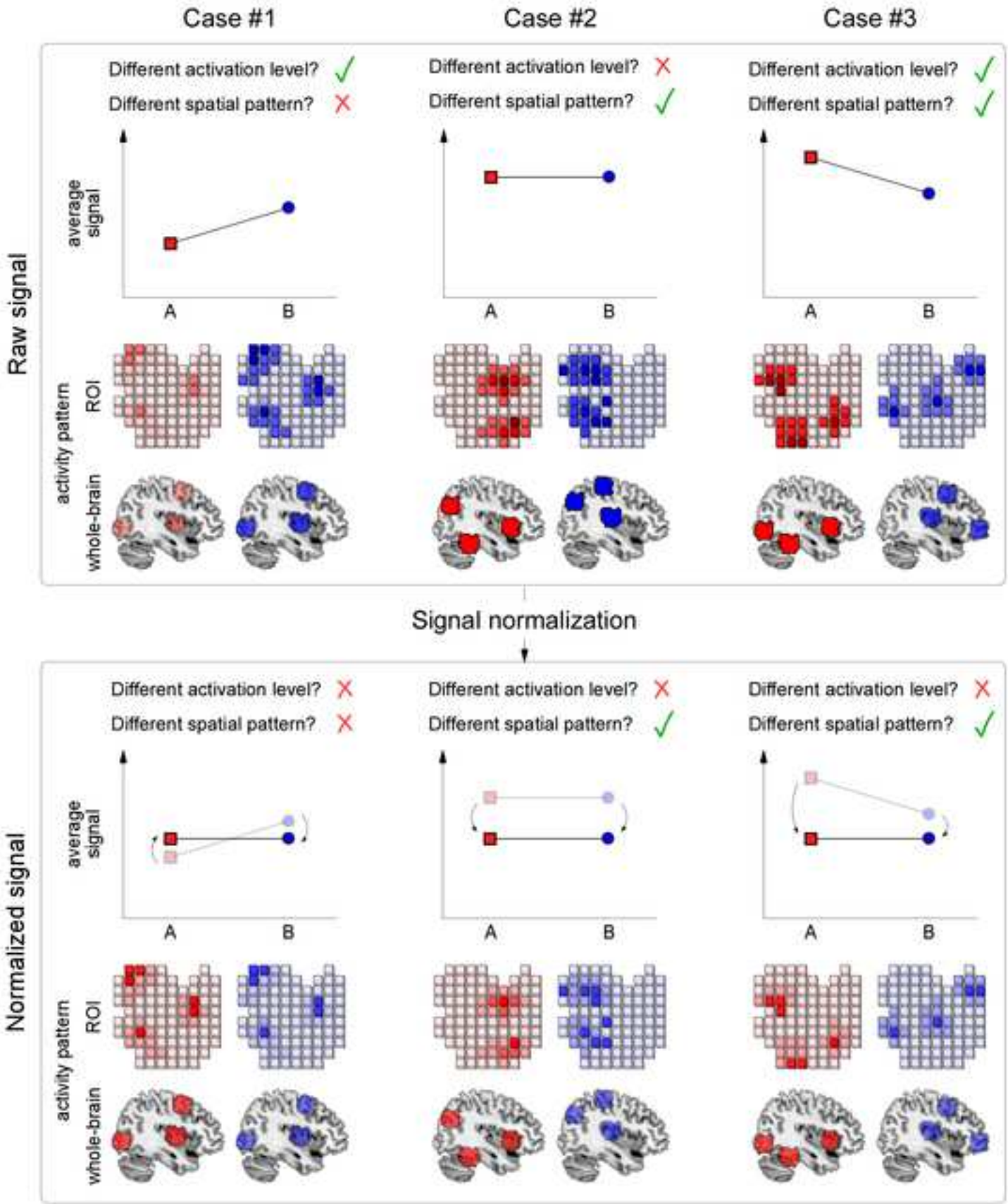


Figure 2

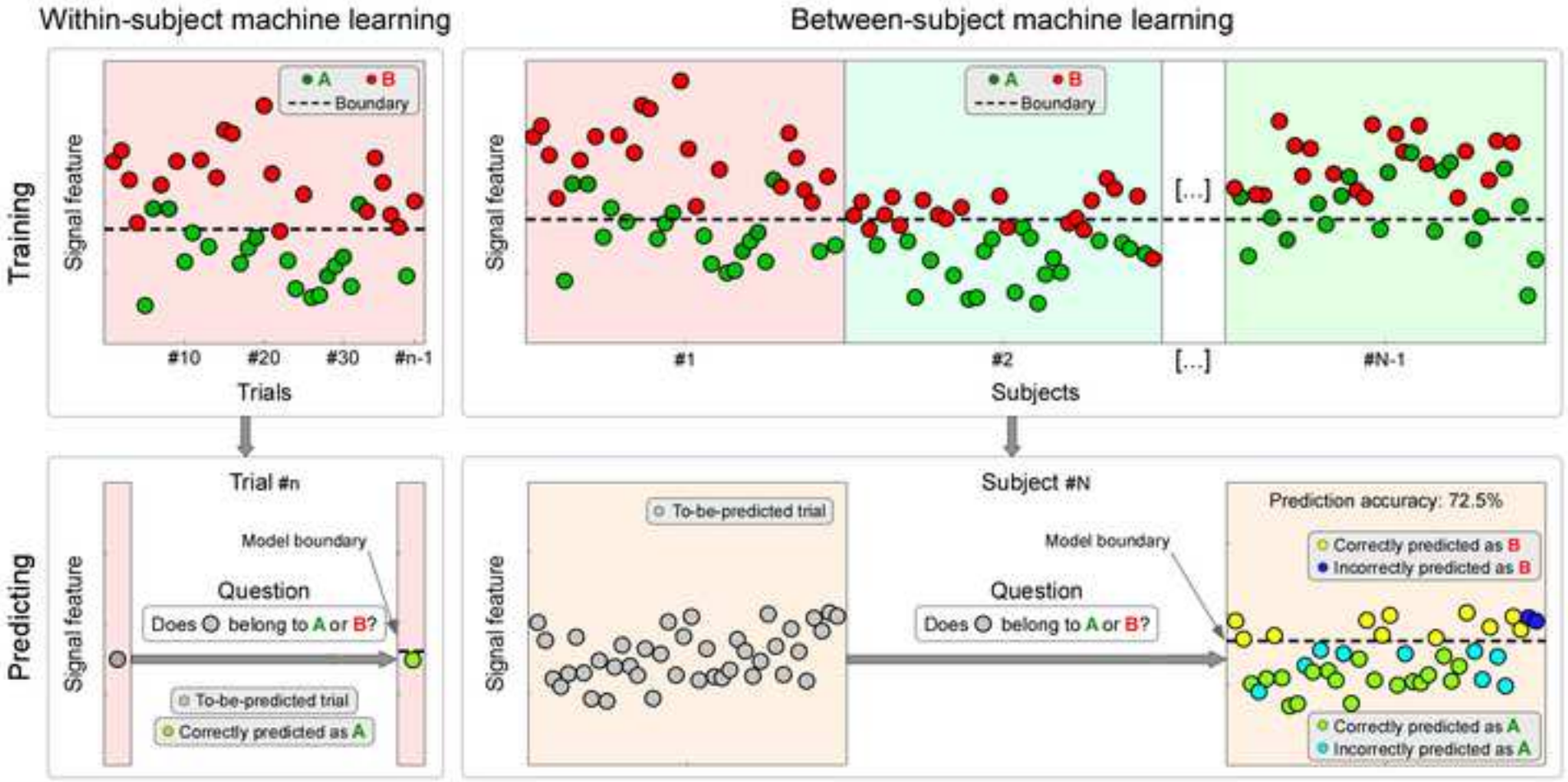
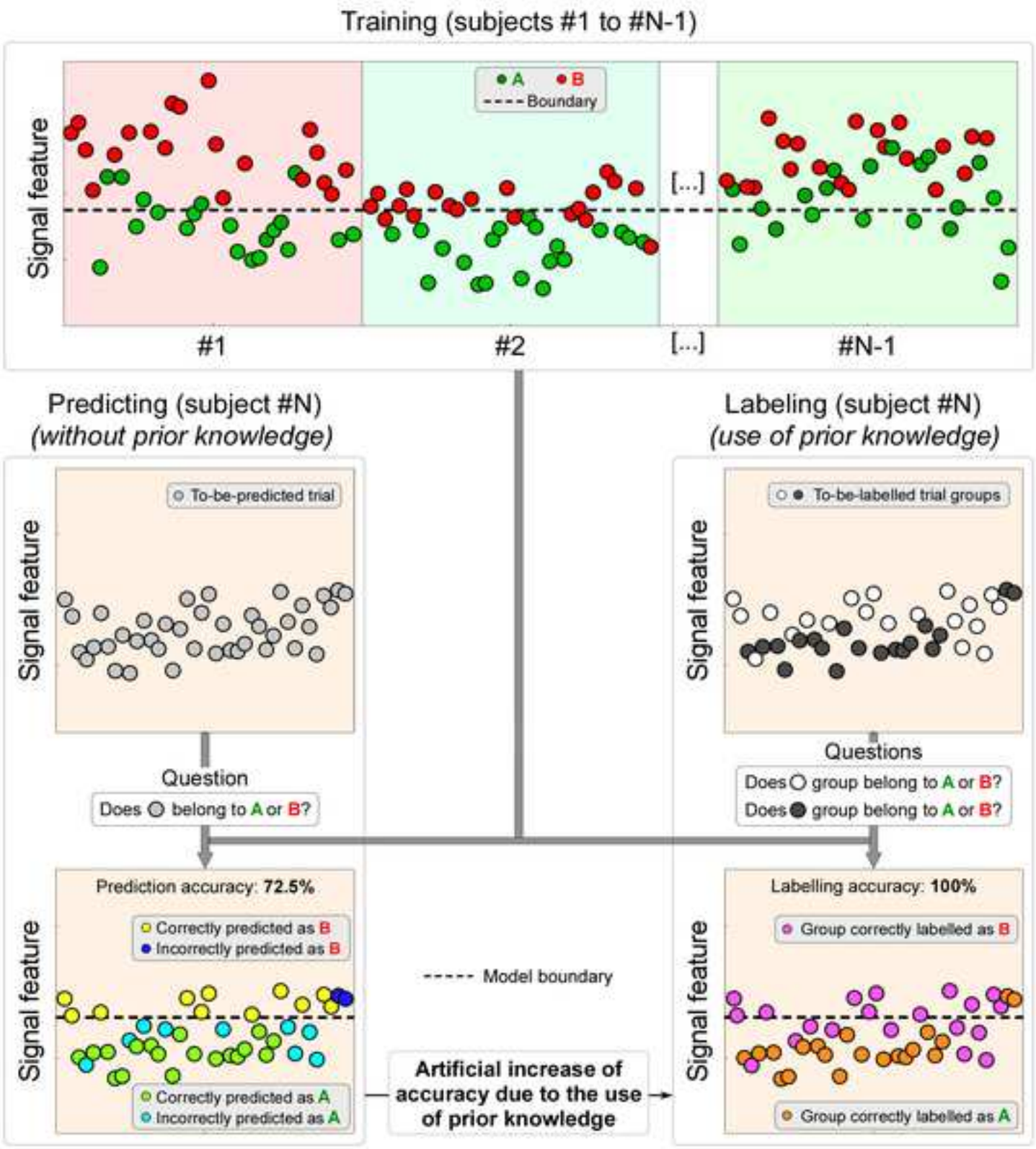


Figure 3

Click here to download Figure Figure-3.tif



Trends

- Predicting perceived pain from brain activity has enormous implications: “pain signatures” from brain imaging data are increasingly used as evidence for pain perception in minimally conscious patients or infants, or in legal settings.

- Sophisticated machine learning algorithms are increasingly applied to functional brain imaging data with two main objectives: (1) identifying a specific neural “pain signature” and (2) predicting perceived pain from brain activity.

- While machine learning approaches hold considerable promise for pain research, they are fraught with interpretive difficulties: disregarding the tight match between machine-learning protocol design and the desired study objectives could lead to incorrect interpretation of results.

Machine learning: an analysis approach that consists in using the ability of computers to learn from and make predictions on different kind of data. When applied to functional brain images, machine learning can be used to detect response patterns (e.g., intensity and spatial distribution of functional MRI signals) associated with a given experimental variable (e.g., the intensity of pain perception).

Machine learning prediction: once machine learning has identified a response pattern associated to an experimental variable, it can be used to *predict* that experimental variable on the basis of the detected response pattern.

Multi-voxel pattern analysis (MVPA): a kind of machine learning technique that identifies condition-specific spatial patterns of fMRI responses distributed across different voxels. These patterns of activity can be used to predict the occurrence of different experimental variables (e.g., different levels of subjective pain, or pain vs touch).

Neural signature: a feature of the brain response that is *uniquely* associated with a given experimental variable. To identify conclusively a neural signature it is crucial to ensure that its relationship with the experimental variable is exclusive, i.e., that other experimental variables do not produce the same pattern of brain response.

Pain prediction: the process of estimating *unknown* subjective intensity of pain perception using experimentally-measured functional brain imaging data. True pain prediction must not use prior knowledge about subjective reports of pain intensity when testing the prediction performance.

Prior knowledge: in the context of machine learning, prior knowledge refers to the information about the experimental variables that, although available, should not be used when testing the performance of the machine learning classifier in predicting an experimental variable. The incorporation of prior knowledge into the training is a necessary aspect of machine learning. In contrast, exploiting prior knowledge when testing the algorithm performance is incorrect, and results in an artificial inflation of performance (false positive results).

Reverse inference: in the context of human brain imaging reverse inference consists in inferring an experimental variable (e.g., pain perception) from a pattern of neural activity (e.g., the brain responses elicited by a nociceptive stimulus). The validity of a reverse inference drawn from neuroimaging depends on the exclusivity of the relationship between the experimental variable and the brain responses. For example, the validity of the inference that a person is experiencing pain because the pattern usually seen in response to nociceptive stimuli is observed, depends on whether the same pattern is also elicited by other stimuli that do not result in painful percepts.

Box 1. Encoding, decoding, and reverse inference.

In functional brain imaging *encoding* refers to the identification of a statistical dependency between experimental variables (e.g., pain perception) and measured brain responses. This encoding procedure is normally achieved using the traditional voxel-by-voxel mass-univariate analysis of fMRI timeseries (using, for example, general linear modeling: GLM, Fig. I).

In contrast, *decoding* consists in predicting the same experimental variables based on the measured brain responses. This decoding procedure is typically achieved using machine learning (e.g., multi-voxel pattern analysis, MVPA, Fig. I), which is based on certain features of the fMRI response (e.g., patterns of fMRI activity distributed over many voxels).

Reverse inferences are logically-flawed deductions based on affirming the consequent (e.g. if A determines B, when B is observed one infers that A has occurred). Reverse inferences are notoriously frequent in functional neuroimaging research, and typically consist in inferring a particular experimental variable (e.g., the perception of pain) from a given pattern of brain activation (e.g., the so-called “pain matrix”) [37-38]. Notably, reverse inferences have a probability of being correct, which depends on the exclusivity of the relationship between the experimental variable and the

recorded response (i.e., it depends on how many variables other than A determine B).

Even if decoding is the reverse prediction of experimental variables from the measured brain response, decoding is conceptually different from *reverse inference*: indeed, in most practical applications, decoding analysis does not require that the relationship between the experimental variable and the corresponding brain response is exclusive. For example, most currently available pain prediction algorithms rely on features of the brain response that are not tested for their necessity or sufficiency for the occurrence of pain perception.

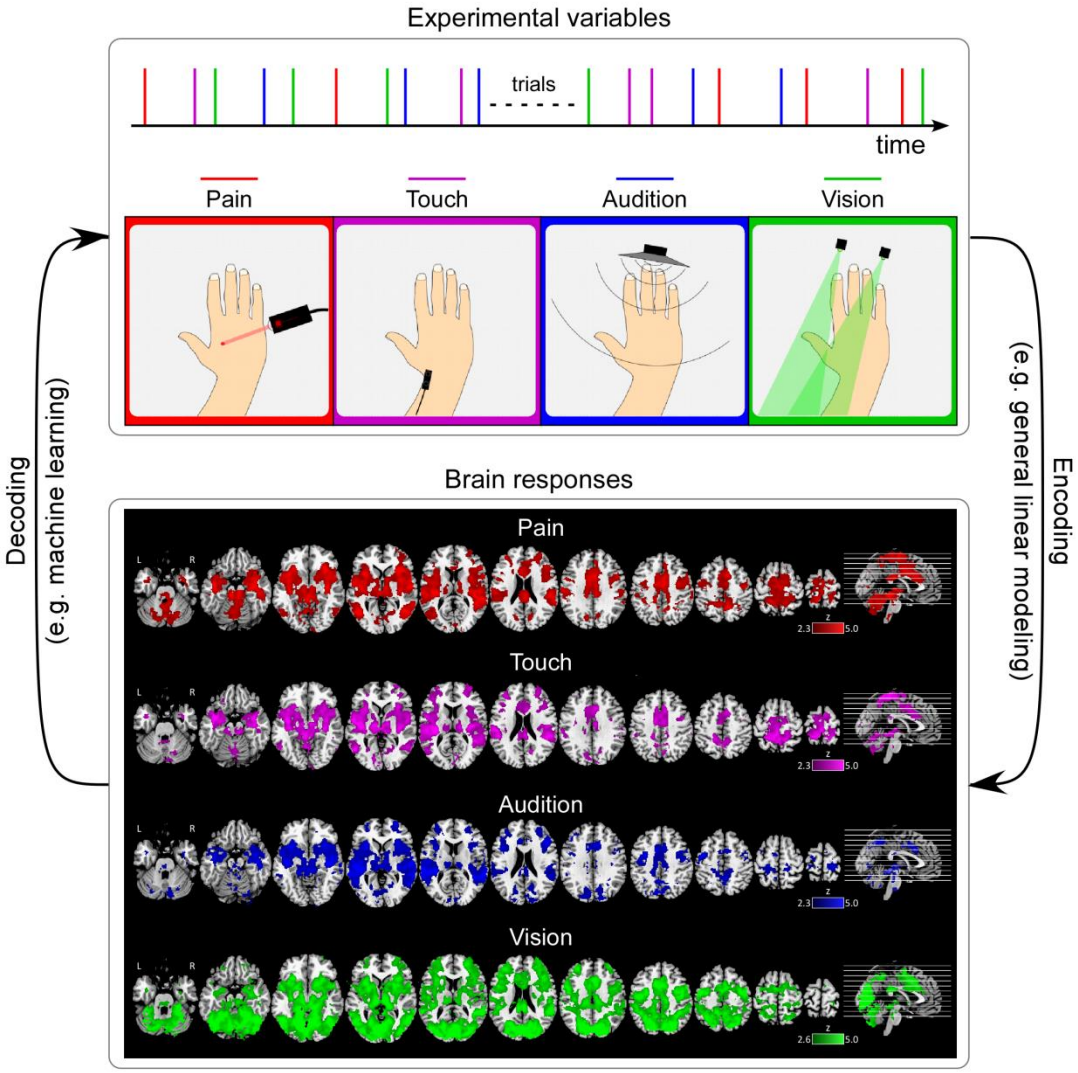


Fig. I. Relationship between *encoding* (identifying the statistical dependency between experimental variables and brain responses) and *decoding* (predicting unknown experimental variables from the brain responses). Bottom panel modified from [29].

Outstanding questions

- Do the functional neuroimaging features used to predict pain actually reflect neural activities that are causally related to the emergence of pain percepts? Or they reflect neural activities related to the consequences of painful percepts, but not directly involved in their emergence (e.g. attentional orienting, autonomic responses, motor preparation)?

- Which of these two kinds of neural activity (causally-specific for pain vs pain byproducts) is more likely to provide a reliable pain prediction?

- Will it be possible to use a machine learning classifier trained on functional neuroimaging data to predict perceived pain in real-life situations (e.g., when an individual is admitted to the hospital)?

- Should functional neuroimaging data be used as conclusive evidence of an experiential state of pain in medico-legal cases?

- Should the scientific community agree on some guidelines for avoiding the conflation of the objectives of pain prediction vs. the identification of pain signatures?