Machine learning four NeuroImaging data analysis

Bertrand Thirion

Machine Learning four NeuroImaging data analysis

Bertrand Thirion

January 2024

Abstract

Machine learning has become an ubiquitous tool for neuroimaging data analysis over the last two decades. It has opened up the possibility of assessing relationships between brain characteristics - and most importantly, brain activity - with many different behavioural covariates, both in the field of cognitive neuroscience and for population studies. The use of machine learning requires some expertise, as there are some pitfalls to avoid, such as biased assessments due to some form of data leakage, or reliance on underpowered datasets leading to erroneous conclusions. The era of larger datasets is upon us, and the field of generative AI brings a new perspective to the field.

Keywords

Machine learning, supervised learning, decoding, encoding, functional Magnetic Resonance Imaging, interpretable machine learning, cross-validation, sample complexity, generative AI, high dimension.

Keypoints

- Machine learning has opened up new possibilities for the statistical analysis of neuroimaging, such as predictive modeling.
- Brain activity decoding can assess information content in brain maps, but it requires large sets of observations.
- Encoding models are the reference approach to map complex feature space to fMRI data, but are limited by the voxel-wise approach.
- Representational similarity analysis is an optimization-free alternative to encoding, but its statistical interpretation deserves some caution.
- Brain activity decoding can now be enhanced by generative AI technology to produce high-quality perceptual reconstructions.

1 Introduction

Over the last two decades, machine learning (ML) has become a key component of brain imaging data analysis. Its introduction originally came from the opportunity provided by the development of Machine learning, but was also motivated by several needs that emerged in the early 2000s. A prominent motivation is that while the in-sample statistics used at the time could provide a reliable inferential framework, they were unable to **make predictions** for a given observation, thus defeating the promise of individualised predictions, whether for diagnosis, prognosis or treatment response prediction. This need was pervasive across all brain imaging modalities. A second concern, related to the first, was the increasingly obvious **limitations of traditional null hypothesis testing** frameworks, which could produce frequentist statements that were not well suited to the discovery of novel insights into brain function and the accumulation of knowledge. A third motivation was the realisation that an important part of the information conveyed by brain activation images lies in the **patterns** they represent [18]. Here, a *pattern* should be understood as a configuration of activity across a given set of brain locations that is informative about a particular cognitive state or task. This immediately called for new methods to model and exploit these patterns in analytic tasks. The fourth motivation came a little later [26]: it was that richer models would be needed to correctly represent the content of complex stimuli and map it to brain activity. While traditional statistical inference only considers well-posed designs, where the number of observations exceeds the number of covariates, modelling **larger feature spaces** entailed high-dimensional models that can only be accommodated by relying on ML techniques.

A historical perspective We propose a perspective based on four key steps in the development of these machine learning approaches. Rather than dealing with a strict chronology, these should be seen as conceptual steps. We nevertheless provide indicative dates for the sake of concreteness.

- **Step 1: Pattern analysis and lightweight machine learning(2001-2008)** The whole approach started with the seminal contributions in [18] which mainly emphasized a novel pattern analysis approach to functional brain imaging. Soon, a supervised learning perspective [8] emerged, introducing well-posed frameworks and really exploiting the power of the ML techniques that were available at that time (mostly high-dimensional linear models, such as support vector machines). This type of framework quickly achieved important successes, such as the decoding of orientation from signal in the early visual cortex [20, 25] However, the pattern analysis received a further boost with the popularisation of Representational Similarity Analysis, [30, 31], a neuroimaging avatar of kernel-based data analysis [16], which soon became popular when used in conjunction with the so-called searchlight approach.
- **Step 2: Improving expressiveness and interpretability(2008-**

2018). A landmark paper [26] in 2008 introduced the use of high-dimensional stimulus representations that could be used in encoding models (see section 4). Fitting these models actually required the use of high-dimensional machine learning models. In parallel, researchers increasingly interpreted brain models [46], and realized that ML should be regularized to provide more meaningful results. These resulted in a series of technical developments, e.g. [17, 23, 7].

- **Step 3: Critique of standard ML frameworks and sample size increase (2017-)** Until then, neuroscience had worked in the realm of low-cost ML, with relatively simple models applied to sample-limited datasets, with risk of overfitting. The community began to realize the difficulty of evaluating models, or setting hyperparameters [43]. While some part of this was due to a poor use of machine learning tools, a prominent aspect was found to be the small sample size used in many neuroimaging works dealing with ML, leading to unreliable results [43]. Subsequently, some works started to perform large-scale analyses, in particular combining data from many datasets [32, 33].
- **Step 4: The age of deep phenotyping and generative modelling (2022-)** The last evolution in the field has come from intense developments in the fields of ML, now called *AI*, around the development of generative AI, namely the reproduction of content (images, video, speech, text), conditioned on some user inputs. On the neuroimaging side, the main trigger has been the development of large datasets such as BOLD5000[5] and the Natural Scenes dataset [2], which provide a large number of responses to visual stimuli in a few individuals. This has paved the way for a large number of works on reconstructing percepts from visual activity. See for example [38].

Outline In this chapter, we focus on supervised learning approaches, setting aside a part a large body of work on unsupervised learning also known as multivariate decompositions, which have proven to be useful for many tasks, but are typically one step in a larger pipeline. Those modeling approaches generally suffer from hard validation and model selection problems. This chapter is centered on functional Magnetic Resonance Imaging (fMRI), because the literature in this area is quite rich, and because it lends itself to the two perspectives outlines above: *i)* as an example of a medical imaging modality, it carries the perspective of cohort studies, individualised predictions, hence personalised medicine *ii)* as a temporally resolved modality informative about brain function, it carries information on cognition, as do electrophyiological modalities, and brain-computer interfaces.

The remainder of the chapter is organized as follows: in Section 2, we review the main motivations and principles underlying the use of machine learning techniques in brain imaging data analysis. Then in Section 3, we focus on supervised classification approaches, often referred to as *decoding* approaches in the context of functional brain mapping; in Section 4, we discuss pattern analysis and encoding methods that compare complex feature spaces with brain activity.

2 Application of machine learning to brain image analysis

The use of ML in neuroimaging usually corresponds to one of two main frameworks:

- **Intersubject settings**, where one tries to make a prediction about each individual, such as a diagnostic or prognostic task. This type of analysis has a clear medical relevance. The samples can generally be considered as independent (although genomic similarity can create some dependencies), but the classification problem typically suffers from a large variability between these samples. A special case is *fingerprinting*, where brain features, such as brain connectivity are used as markers of identity [13].
- **Intra-subject settings**, that discriminate between brain states and/or assess the global similarity between these brain states. This is used for cognitive brain mapping, and the result is an assessment of the regions whose activity is predictive of a particular cognitive state. The samples are not necessarily independent, because a dependency structure is inherited from the runs that produced the data. On the other hand, there is limited variability between the samples. This framework is illustrated in Figure 1.

The motivation for these machine learning frameworks is to infer the statistical significance of the association between brain image data and some associated target information, such as individual features or annotations of the presented stimuli. Classical tests consider univariate associations between the signal available at a given location and the target information; this is often called **mass-univariate** since a large number of tests are performed in parallel. However these models are limited in two ways: i) they do not take into account that different brain locations may be associated with different, and potentially complementary information of the target; ii) they suffer from multiple comparison problems, where the significance of the tests needs to be corrected for the number of tests performed, which compromises the sensitivity in the context of brain imaging.

In contrast, ML approaches rely on an integrated approach, where all the features in the data are used to jointly fit a given target. More specifically, a dataset consists of a number of individual observations (brain images), mathematically represented by a vector of values (e.g. voxel signals) called features. Each observation corresponds to a specific target value. On the basis of a certain sample of (observation, target) pairs, one wants to predict the target from the observation. Importantly, the inference now consists in ensuring that the prediction accuracy of the ML is higher than chance. This can be done by comparing the accuracy reached by the ML model with that of a *dummy* model that relies on a basic heuristic (random selection, or selection of the most frequent class). Prediction accuracy is measured in a way that is appropriate with the learning problem: for a classification problem (with categorical targets), it can be prediction accuracy, or the area under the receiver operating curve

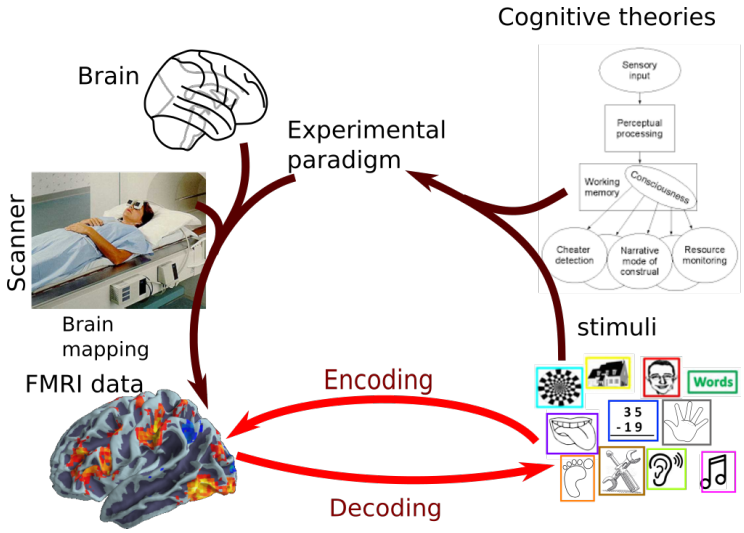

Figure 1: **Use of machine learning in functional neuroimaging data analysis**. Neuroimaging experiments are designed to probe brain activity in various contexts. At analysis time, machine learning approaches can be used to quantify the relationship between cognitive concepts that are probed with some classes of stimuli versus brain activity. This can be done using decoding (predicting stimulus characteristics from brain activity) or encoding (predicting brain activity from stimulus).

of the classification problem, or the area under the precision-recall curve —among many other metrics. When considering a regression problem, prediction accuracy is measured by the mean absolute or mean squared error on the prediction, or by the proportion of the variance of the target explained by the model (see e.g. `https://scikit-learn.org/stable/modules/model_evaluation.html` for an overview). For significance testing, a distribution of accuracy values should be considered, such as that obtained by considering multiple draws of a random dummy classifier.

This approach directly addresses the two limitations of classical inference raised above, although it should be emphasized that the inference performed is not equivalent to that of a mass-univariate model: whereas the latter is informative about the location of features associated with the target, the former is simply informative about the presence of a global association, representing a kind of global null hypothesis rejection.

Predictive validity A very important advantage of ML approaches is that they directly assess **predictive validity**: that is, the accuracy of the prediction made by the model on unseen data. This is because ML frameworks require the use of external validation: in fact, an ML model is typically a complex function that can combine many features to fit an output. It is well known that when complex models are allowed, it is always possible to perfectly fit a given target y based on high-dimensional data $\mathbf{x}$ at training time, even if there is no relationship between $\mathbf{x}$ and y in reality. For this reason, it is essential to take another independent dataset containing new instances of $\mathbf{x}$ and y to measure accuracy. In practice, *validation* procedures consist in dividing the available data into training and test data, so that a model is trained on the learning data and evaluated on independent test data to measure accuracy. However, particularly in the context of brain imaging, it is often the case that the total amount of data is limited, so that the data available for validation (typically 1/5 of the data) is even smaller, making the results highly dependent of this peculiar selection. For this reason, a cross-validation procedure must be considered instead: several splittings of training and testing are considered, and the results are averaged across the splittings, leading to a more reliable measure of performance. This is called *cross-validation*. The recommended cross-validation schemes are 5-fold or 10-fold, where k-fold means that the data are split into k subsamples of equal size, from which $k-1$ are used for training and 1 for testing [3, 43]. But in practice, more complex schemes are used, either to create more folds, or to take account of some latent grouping structure in the data or to stratify the cross-validation scheme with respect to the target distribution of some side information. The framework is often a bit more complex: as the learner relies on some *hyperparameters*, e.g. regularization parameters or dimension selection, which cannot be inferred upon at training time, another batch of observations has to be considered to measure the model accuracy for different values of hyperparameters. In this case, a nested cross-validation loop must be used; more simply, practitioners will rely on a tri-partitioning of the data [14]. A typical framework is illustrated in Fig. 2.

In summary, the main advantages of ML approaches to brain imaging is their reliance on predictive validity criteria, which leads to trustworthy

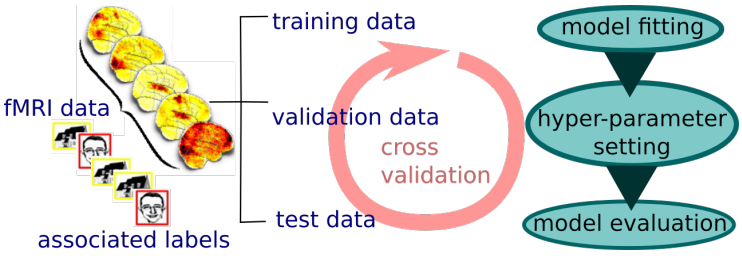

Figure 2: **Outline of the standard Machine learning analysis pipeline**.

results, without making untestable assumptions about the data [44].

Pros and cons of ML frameworks for neuroimaging It is useful to consider the advantages of ML frameworks and their limitations.

- Benefits
  - ML frameworks can **handle ill-posed settings**: ML learners can make a prediction, even if the number of samples n used for training is smaller than the number of dimensions p in the data. This is true even when relying on conceptually simple linear models: in such cases, using a shrinkage estimator such as Ridge or Lasso regression, makes the problem well-posed. The only cost is the introduction of an hyperparameter that controls the amount of shrinkage.
  - ML frameworks can **generalize** what was learned during training to data that is out-of-distribution, but with some potential latent relationships. For example, in [28], it was shown that a classifier trained to discriminate between leftward versus rightward saccades actually classify mental additions as rightward saccades, and mental subtractions as leftward saccades, but only when instantiated on some regions of interest in the brain volume.
- Limitations
  - Classification accuracy is a meaningful number (e.g. if the chance prediction accuracy is 50%, it is easy to interpret a 55% accuracy as the evidence of weak information, whereas a 95% classification accuracy indicates a strong effect), but the **statistical significance** of these numbers is difficult to establish. For example, accuracies from different folds are not statistically independent, which undermines the use of simple statistical tests of accuracy across folds.
  - Machine learning methods are **costly**. For a learning problem with n samples and p features, the computational complexity is at least $np\min(n, p)$. Other estimators can be much more expensive. This makes it difficult for practitioners to use such models. Fortunately, the development of ergonomic ML frame-

works in the years 2010 has made these approaches accessible to many researchers [1].

- ML methods are **data greedy** [43]: Since model training involves learning many parameters, a stable estimate can only be obtained when enough samples are available. It is increasingly recognized that most publications rely on too few samples to make reliable inferences.
- ML frameworks are more complex than traditional statistics, and have sometimes been misused by practitioners, leading to data leakage [45] and hence flawed performance evaluation. Even with the use of cross-validation, the reliance on too limited datasets to set hyperparameters can lead to also lead to overfitting [43]. Many contributions have led to non-reproducible results, e.g. [24].

In summary, machine learning frameworks have rapidly become popular as neuroscientists have realised their potential to address new questions. However, these frameworks require some technical insight and have sometimes been misused.

3 Decoding and individual prediction

The most emblematic and impactful ML tool for brain imaging is probably that of classification of high-dimensional data, often referred to as *decoding* in the context of cognitive neuroscience.

High-dimensional classification or regression In this section we focus on a key use case for ML for brain imaging: the *decoding* framework, where image data is used to make a prediction, either an individual assessment or some cognitive information related to stimulus processing in a given trial. This type of work relies on supervised classification or regression, which typically involves a high-dimensional learner that maps the information in the brain images, represented by a vector $\mathbf{x}$ of features, to the prediction of the target information y. Classification is used when y is a categorical variable, while regression is used when y represents a scalar quantity. A prominent problem is the optimal classifier f is typically sought in large family $\mathcal{F}$. Since it is possible to fit the training data almost perfectly with models in $\mathcal{F}$, there is overfitting, i.e. a large gap between the training and test errors. This gap can be reduced by using regularisation to penalise the complexity of the model learned within $\mathcal{F}$, and such penalisation is often beneficial for prediction accuracy. Another approach to improving prediction is not to consider the full set of features in $\mathbf{x}$, but to reduce it to an informative subset by using a *feature selection* approach. The most popular approaches are univariate feature selection, which selects the best features according to their association with y, and recursive feature elimination [9], but the latter is much more costly, and in general, not more powerful. Both approaches are heuristics, i.e. come without optimality guarantees, and require the tuning of additional hyperparameters, such as the number of features used. Hyperparameter and

feature selection must be done on the training set. This framework is well summarised in [21].

Intrestingly, this model has been easily reused in other areas of neuroscience, such as electrophysiology. There, the temporal resolution of the data provides additional flexibility; for example, whether a classifier trained on signal in one time window yields a prediction in another time window is indicative of the temporal structure of brain activity [27].

Interpretability and inference Making accurate predictions is not enough. Researchers also need to be able to make sense of the model. This requires that models are *interpretable*. Since the early days of the field, researchers have relied on linear high-dimensional models, that provide weights per feature, since the prediction model is nothing but a weighted sum of the inputs, e.g., $f(\mathbf{x}) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$, where $\mathbf{w} \in \mathbb{R}^p$ is a weight vector and b a bias parameter, for a binary classification problem. Then, for $j \in \{1 \cdots p\}$, $\mathbf{w}_j$ represents the importance of feature j. If the features are brain voxels, they form a map, the feature map. Inspection of weights has been done since [35] to interpret the learned model for classification.

However, several problems arose. The first one is that these weight maps were not easy to interpret, and sometimes lead to salt-and-pepper structure that does not make sense. The reason is simple: Learning $\mathbf{w}$ in the training phase is an ill-posed problem, since the number of features (brain voxels) is much larger than the number of observations. As a consequence, there are an infinite number of possible solutions that perform equally well. Considering a weight map means that one such solution is considered but this may not be the most meaningful one. The solution to this is to complement the learning problem (minimising of the training loss function $\ell(f)$, which measures the error rate on the training set) with a meaningful regularizer $R(f)$, which imposes some properties on the estimated f. In the case of a linear model, taking $R(f) = \|\mathbf{w}\|_2^2$ leads to Ridge regression $R(f) = \|\mathbf{w}\|_1$ to a Lasso problem with few non-zero coefficients in $\mathbf{w}$, but more complex regularization schemes penalize sharp variations of $\mathbf{w}$ when it is considered as an image [17]. Similarly, total variation penalisation has been used in this framework. In all cases, the estimation of f becomes much more complicated, and it involves additional hyperparameters.

Considering that much of this computation was overkill, alternative approaches have emerged, such as Fast Regularized Ensembles of Classifiers, which parcellate (group) voxels into small regions, average the signal within these regions, and then assign weights to the regions. To mitigate the effects of an arbitrary choice of parcellation, the procedure was repeated with different parcellation schemes that were equally good at representing the data. The resulting estimation procedure is more efficient and more stable than estimation schemes involving complex penalisation [23]. This is illustrated in Figure 3.

All this is helpful for interpretation, but one may still wonder how to make sense of the resulting weights, since they are not independent across features. The rigorous way to interpret these weight maps was first presented in [46]; a certain coefficient $\mathbf{w}_j$ representing the importance of feature j can be interpreted as a **conditional association test**: $\mathbf{w}_j \neq 0$

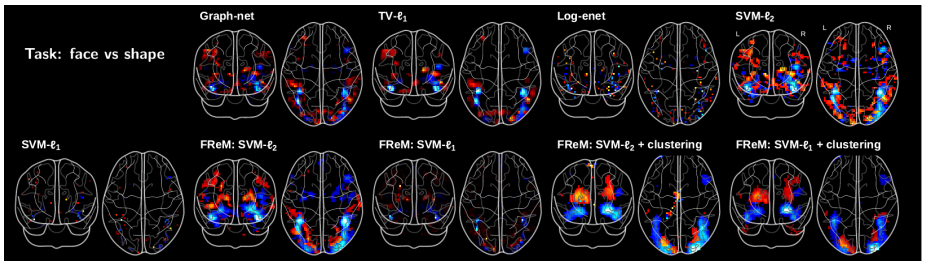

Figure 3: **Qualitative comparison of decoder weight maps**: Weight maps for different discriminative tasks on the HCP dataset, thus in an inter-subject setting. The maps are thresholded at the 99 percentile for visualization purposes. Note that this is an inter-subject setting, where the classification model has to generalize across individuals. The weight maps obtained with TV-L1 and FReM methods with clustering display a prediction driven by the functional areas of the visual mosaic, such as: primary visual areas, lateral occipital complex, the face and place specific regions in the Fusiform Gyrus. This figure was originally part of [23].

means that y depends on $\mathbf{x}_j$ conditionally on other features used, i.e. that it brings original information for prediction. This is in contrast to classical *marginal* importance scores, which consider the association between $\mathbf{x}_j$ and y, without taking other information into account. This explains in particular why this importance score is computationally difficult to obtain. However, it remained unclear at that point whether reliable statistical inference could be performed on these conditional importance coefficients. This was successfully addressed in [7], borrowing ideas from [23].

Overall, the difficulty of correctly interpreting the model coefficients has become increasingly apparent over the years, but rigorous interpretation has become feasible, although not at full brain resolution.

To conclude this section, let us briefly mention three extensions of the basic decoding framework: hyperalignment, large-scale decoding and generative decoding.

Hyperalignment It is often interesting, and sometimes necessary for reasons of sample size, to consider data from different individuals simultaneously when decoding brain activity. The problem then becomes one of intersubject differences, that reduce the ability of the classifier to generalise across individuals. Hyperalignment has been developed to align data from mutliple individuals, so that they can be pooled [19]. As there are many different flavors of hyperalignment, using different models of brain correspondence across individuals [42], a comparison of these techniques in the context of brain activity decoding has been presented in [4].

Large-scale analysis Acknowledging that the main weakness of decoding stems from two limitations: *i)* the lack of data, which leads to unreliable accuracy and coefficients *ii)* the discrimination of limited sets

of cognitive categories, which does not allow to build a consistent picture of the associations between brain regions and cognitive tasks, recent efforts have been paid to perform decoding at large scale, using public image repositories [32, 33], as long as they contain images with a range of values consistent with those of the task-related maps, and meaningful labels. This has highlighted the fact that one of the main problems is that consistently labelling images with cognitive concepts is a difficult task, and is likely to be the main limitation to large-scale decoding efforts [33]. Similarly, large-scale decoding has been performed by relying on the loci reported in publications, creating a novel type of predictive meta-analytic framework [10].

Generative decoding The impressive development of generative deep learning in recent years [15] has led to an unexpected development: since linguistic content or images can be conditionally generated, given some latent information, then decoding can simply be used to predict such latent representations. In such a setting, deep learning (often referred to as *AI* systems) generates *hallucinations*, i.e. images that could be held in mind, but those are constrained by the latent information conveyed by the brain data. The resulting reconstruction can reach impressive levels of accuracy [38]: seen images or heared speech can be reproduced with high fidelity using fMRI [41] or MEG data [12], but the whole field is moving very fast, suggesting than other cognitive content could be decoded in a near future. The utlimate limitation to such decoding efforts is the availability of large-scale neuroimaging data associated with stimuli, within subject, to get a sufficiently accurate decoding of brain activity towards latent space. The necessary data are currently available only in very few datasets [5, 2].

4 Encoding and Representational Similarity Analysis

Encoding models Since the seminal work of [26], it has become increasingly popular in the cognitive neuroscience community to build complex models of the stimuli and then to study how well these complex features together fit the brain activity. This framework was conceptualized in a subsequent paper [37]. The approach is to use a high-dimensional regression model to explain the activity in each voxel from a large set of features. Thus, it remains a mass-univariate approach, but with a high-dimensional feature space. As the problem is most constrained by the limited signal-to-noise ratio of each voxel time course, a simple and computationally efficient Ridge regression approach is typically sufficient to construct the encoding model. Its accuracy is measured by the prediction error of the model on a left-out run. If the feature space consists of several sub-spaces containing different classes of features, it may be better for fitting and prediction purposes to adapt the Ridge penalty to each sub-space. This is called Banded Ridge regression [11].

Representation similarity analysis An alternative framework has been proposed, namely representation similarity analysis (RSA) [31]. It consists in bypassing the optimization inherent to the fitting of brain data, but instead taking into account the similarities between the samples (each sample being associated with a stimulus event), measured by their mutual distance in high-dimensional feature spaces. These similarity values are then compared to the similarity in brain activity evoked by the response to the stimuli. This similarity is typically measured in a small brain region. For a given set of stimuli, we thus obtain a matrix of stimuli $\times$ stimuli similarities —typically the correlation matrix of brain activity. The core of the approach is then to assess how the feature space similarity and the brain activity similarity correspond. This statistical approach can be seen as a proxy for the mutual information between the stimuli in the high-dimensional feature space on the one hand, and brain activity on the other hand [30].

The advantage of this framework is that it is cost effective, since it does not rely on any optimisation procedure. It naturally provides a region-level mapping of the correspondence between brain activity and stimuli. The disadvantage of this approach is that it does not make it explicit which feature of the stimuli (within a given space) drives the correspondence. It has no predictive validity —whereas encoding models do. When used in a *searchlight* approach, i.e. the brain region is a small sphere centred on each brain location in turn, it becomes expensive again, and it is unclear how to declare statistical significance for the RSA values.

There have been few formal comparisons between encoding and RSA approaches so far, but there is evidence that encoding approaches may have more general validity [6].

Canonical Correlation Analysis and Partial Least Squares For population studies, RSA or encoding methods are rarely used. Yet, with the construction of richly phenotyped population cohorts, there has been some interest in exploring the relationships between brain activity and brain characteristics. To capture the essence of the association, there is a relatively broad consensus that brain traits should be compared with behavioral or genetic phenotypes, resulting in *many-to-many* association. While this can be done in the regression framework, e.g. using Random Forests or Reduced Rank Regression [39], a more popular solution so far has been, Canonical Correlation Analysis (CCA), which aims to build a linear combination of variables within each block that are maximally correlated with the corresponding variables in the other block [40]. The resulting correlation can be measured out-of-samples, thus avoiding concerns about statistical *circularity*.

However, this type of approach usually requires a prior reduction in the dimensionality of the data, in general done with a principal components analysis of the variables within blocks. This makes the computation quite efficient but blurs the identification of the correspondences measured between blocks of data. Another possibility is to use Partial Least Squares (PLS), an approach that maximises the covariance between blocks rather than the correlations, which makes the loadings within blocks somewhat correlated with the principal components in that block. Recently, the low

reproducibility of this type of association (whether CCA or PLS) when the sample-to-feature ratio is not very high, has called into question the reliability of this type of inference [22].

5 Conclusion: generative modeling in the AI toolbox

ML tools have become increasingly popular for neuroimaging data analysis. Decoding is certainly a major tool, that can be very powerful to issue individualized prediction or measure the presence of task-related signal in a given part of the brain. However, it can be hard to use properly, requires a lots of data, and needs some effort to yield valid interpretations. RSA and decoding, on the other hand, provide a local view of association that bypasses the interaction between brain regions that complexify the interpretation of decoding weights. However, they do not provide per-sample predictions.

The impressive development of artificial intelligence opens huge opportunities for the field of machine learning for brain imaging: being able to represent and generate visual or language content open very intriguing possibilities, such as decoding internal thoughts, visualize percepts and mental images, generate voice content [34], making an obvious link with Brain Computer interfaces (BCIs). It could be further fostered when used in association with inter-subject mapping tools, and by the development of AI models for other type of high-level content.

Another future direction is a nearing between ML-powered cognitive brain imaging and computational neuroscience in the framework of what is now known as cognitive computational neuroscience [29, 36] that relies extensively on the conceptual setting laid out by machine learning models.

References

[1] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8, 2014.

[2] Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, J Benjamin Hutchinson, Thomas Naselaris, and Kendrick Kay. A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.*, 25(1):116–126, January 2022.

[3] Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40, 2010.

[4] Thomas Bazeille, Elizabeth DuPre, Hugo Richard, Jean-Baptiste Poline, and Bertrand Thirion. An empirical evaluation of

functional alignment using inter-subject decoding. *Neuroimage*, 245(118683):118683, December 2021.

[5] Nadine Chang, John A Pyles, Austin Marcus, Abhinav Gupta, Michael J Tarr, and Elissa M Aminoff. BOLD5000, a public fMRI dataset while viewing 5000 visual images. *Sci. Data*, 6(1):49, May 2019.

[6] Luke Chen. An evaluation of representational similarity analysis for mode l selection and assessment in computational neuroscience. *bioRxiv*, 2023.

[7] Jérôme-Alexis Chevalier, Tuan-Binh Nguyen, Joseph Salmon, Gaël Varoquaux, and Bertrand Thirion. Decoding with confidence: Statistical control on decoder maps. *Neuroimage*, 234(117921):117921, July 2021.

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

[9] Federico De Martino, Giancarlo Valente, Noël Staeren, John Ashburner, Rainer Goebel, and Elia Formisano. Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *Neuroimage*, 43(1):44–58, October 2008.

[10] Jérôme Dockès, Russell A Poldrack, Romain Primet, Hande Gözükan, Tal Yarkoni, Fabian Suchanek, Bertrand Thirion, and Gaël Varoquaux. NeuroQuery, comprehensive meta-analysis of human brain mapping. *Elife*, 9, March 2020.

[11] Tom Dupré la Tour, Michael Eickenberg, Anwar O Nunez-Elizalde, and Jack L Gallant. Feature-space selection with banded ridge regression. *Neuroimage*, 264(119728):119728, December 2022.

[12] Alexandre Défossez, Charlotte Caucheteux, Jérémy Rapin, Ori Kabeli, and Jean-Rémi King. Decoding speech perception from non-invasive brain recordings. *Nature Machine Intelligence*, 5(10):1097–1107, October 2023.

[13] Emily S Finn, Xilin Shen, Dustin Scheinost, Monica D Rosenberg, Jessica Huang, Marvin M Chun, Xenophon Papademetris, and R Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat. Neurosci.*, 18(11):1664–1671, November 2015.

[14] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The Elements of Statistical Learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing̃ Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yõshua Bengio. Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014.

[16] Arthur Gretton, Kenji Fukumizu, Choon Teo, Le Song, Bernhard S̃chölkopf, and Alex Smola. A kernel statistical test of independence. In J. Platt, D. Koller, Y. Singer, and S. Roweis, editors, *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007.

[17] Logan Grosenick, Brad Klingenberg, Kiefer Katovich, Brian Knutson, and Jonathan E Taylor. Interpretable whole-brain prediction analysis with GraphNet. *Neuroimage*, 72:304–321, May 2013.

[18] James V. Haxby, Ida M. Gobbini, Maura L. Furey, et al. Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, 293:2425, 2001.

[19] James V Haxby, J Swaroop Guntupalli, Andrew C Connolly, Yaroslav O Halchenko, Bryan R Conroy, M Ida Gobbini, Michael Hanke, and Peter J Ramadge. A common, high-dimensional model of the representational space in human ventral temporal cortex. *Neuron*, 72(2):404–416, October 2011.

[20] John-Dylan Haynes and Geraint Rees. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat. Neurosci.*, 8(5):686–691, May 2005.

[21] John-Dylan Haynes and Geraint Rees. Decoding mental states from brain activity in humans. *Nat. Rev. Neurosci.*, 7:523, 2006.

[22] Markus Helmer, Shaun Warrington, Ali-Reza Mohammadi-Nej ad, Jie Lisa Ji, Amber Howell, Benjamin Rosand, Alan Anticevic, S̃tamatios N. Sotiropoulos, and John D. Murray. On stability of canonical correlation analysis and partial lea st squares with application to brain-behavior associations. *bioRxiv*, 2021.

[23] Andrés Hoyos-Idrobo, Gaël Varoquaux, Yannick Schwartz, and Bertrand Thirion. FReM - scalable and stable decoding with fast regularized ensemble of models. *Neuroimage*, 180(Pt A):160–172, October 2018.

[24] Marcel Adam Just, Lisa Pan, Vladimir L Cherkassky, Dana L McMakin, Christine Cha, Matthew K Nock, and David Brent. Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nat. Hum. Behav.*, 1:911–919, October 2017.

[25] Yukiyasu Kamitani and Frank Tong. Decoding the visual and subjective contents of the human brain. *Nat. Neurosci.*, 8(5):679–685, May 2005.

[26] Kendrick N Kay, Thomas Naselaris, Ryan J Prenger, and Jack L Gallant. Identifying natural images from human brain activity. *Nature*, 452(7185):352–355, March 2008.

[27] J-R King and S Dehaene. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.*, 18(4):203–210, April 2014.

[28] André Knops, Bertrand Thirion, Edward M Hubbard, Vincent Michel, and Stanislas Dehaene. Recruitment of an area involved in eye movements during mental arithmetic. *Science*, 324(5934):1583–1585, June 2009.

[29] Nikolaus Kriegeskorte and Pamela K Douglas. Cognitive computational neuroscience. *Nat. Neurosci.*, 21(9):1148–1160, September 2018.

[30] Nikolaus Kriegeskorte, Rainer Goebel, and Peter Bandettini. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, 103:3863, 2006.

[31] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.*, 2:4, November 2008.

[32] Arthur Mensch, Julien Mairal, Bertrand Thirion, and Gaël Varoquaux. Extracting representations of cognition across neuroimaging studies improves brain decoding. *PLoS Comput. Biol.*, 17(5):e1008795, May 2021.

[33] Romuald Menuet, Raphael Meudec, Jérôme Dockès, Gael Varoquaux, and Bertrand Thirion. Comprehensive decoding mental processes from web repositories of functional brain images. *Sci. Rep.*, 12(1):7050, April 2022.

[34] Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, Inga Zhuravleva, Adelyn Tu-Chan, Karunesh Ganguly, Gopala K Anumanchipalli, and Edward F Chang. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976):1037–1046, August 2023.

[35] Janaina Mourão-Miranda, Arun L W Bokde, Christine Born, Harald Hampel, and Martin Stetter. Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data. *Neuroimage*, 28(4):980–995, December 2005.

[36] Thomas Naselaris, Danielle S Bassett, Alyson K Fletcher, Konrad Kording, Nikolaus Kriegeskorte, Hendrikje Nienborg, Russell A Poldrack, Daphna Shohamy, and Kendrick Kay. Cognitive computational neuroscience: A new conference for an emerging discipline. *Trends Cogn. Sci.*, 22(5):365–367, May 2018.

[37] Thomas Naselaris, Kendrick N Kay, Shinji Nishimoto, and Jack L Gallant. Encoding and decoding in fMRI. *Neuroimage*, 56(2):400–410, May 2011.

[38] Furkan Ozcelik and Rufin VanRullen. Natural scene reconstruction from fMRI signals using generative latent diffusion. *Sci. Rep.*, 13(1):15666, September 2023.

[39] Mehdi Rahim, Bertrand Thirion, Danilo Bzdok, Irène Buvat, and Gaël Varoquaux. Joint prediction of multiple scores captures better individual traits from brain images. *Neuroimage*, 158:145–154, September 2017.

[40] Stephen M Smith, Thomas E Nichols, Diego Vidaurre, Anderson M Winkler, Timothy EJ Behrens, Matthew F Glasser, Kamil Ugurbil, Deanna M Barch, David C Van Essen, and Karla L Miller. A

positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature neuroscience*, 18(11):1565, 2015.

[41] Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G Huth. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nat. Neurosci.*, 26(5):858–866, May 2023.

[42] Alexis Thual, Huy Tran, Tatiana Zemskova, Nicolas Courty, Rémi Flamary, Stanislas Dehaene, and Bertrand Thirion. Aligning individual brains with Fused Unbalanced Gromov-Wasserstein. In *NeurIPS 2022 - Conference on Neural Information Processing Systems*, New-Orlean, France, November 2022.

[43] Gaël Varoquaux. Cross-validation failure: small sample sizes lead to large error bars. *NeuroImage*, June 2017.

[44] Gaël Varoquaux and Russell Poldrack. Predictive models avoid excessive reductionism in cognitive neuroimaging. *Current Opinion in Neurobiology*, 55, April 2019.

[45] Edward Vul, Christine Harris, Piotr Winkielman, and Harold Pashler. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspect. Psychol. Sci.*, 4(3):274–290, May 2009.

[46] Sebastian Weichwald, Timm Meyer, Ozan Özdenizci, Bernhard Schölkopf, Tonio Ball, and Moritz Grosse-Wentrup. Causal interpretation rules for encoding and decoding models in neuroimaging. *Neuroimage*, 110:48–59, April 2015.