Encoding and Decoding Framework to Uncover the Algorithms of Cognition

King, J-R.[1-2], Gwilliams, L.[1,3], Holdgraf, C.[4], Sassenhagen, J.[5], Barachant, A.[6], Engemann, D.[7], Larson, E.[8], Gramfort, A[7].

1. Psychology Department, New York University, New York, USA; 2. Frankfurt Institute for Advanced Studies, Frankfurt, Germany; 3. NYUAD Institute, Abu Dhabi, UAE; 4. Berkeley Institute for Data Science, Helen Wills Neuroscience Institute, UC Berkeley, USA; 5. Department of Psychology, Goethe-University Frankfurt, Frankfurt, Germany; 6. CTRL-labs, New York, USA; 7. Parietal Team, INRIA, CEA, Université Paris-Saclay, Gif-sur-Yvette, France; 8. Institute for Learning and Brain Sciences, University of Washington, Seattle, WA;

Abstract

A central challenge to cognitive neuroscience consists in decomposing complex brain signals into an interpretable sequence of operations - an algorithm - which ultimately accounts for intelligent behaviors. Over the past decades, a variety of analytical tools have been developed to (i) isolate each algorithmic step and (ii) track their ordering from neuronal activity. In the present chapter, we briefly review the main methods to encode and decode temporally-resolved neural recordings, show how these approaches relate to one another, and summarize their main premises and challenges. Throughout, we illustrate with recent findings the increasing role of machine learning both as a method to extract convoluted patterns of neural activity, and as well as an operational framework to formalize cognitive processes. Overall, we discuss how modern analyses of neural time series help identify the algorithmic bases of cognition.

Introduction

An algorithm is a sequence of simple computations designed to solve a complex problem. Under this definition, a major goal of cognitive neuroscience therefore consists in uncovering the algorithms of the mind: i.e. identifying the nature and the order of computations implemented in the brain to adequately interact with the environment (Marr, 1982).

Over the years, this foundational endeavor has adopted a variety of methods, spanning from the decomposition of reaction times (Donders, 1969; Sternberg, 1998) to modern electrophysiology and neuroimaging paradigms. In the present chapter, we focus on two major pillars necessary to recover an interpretable sequence of operations from neuronal activity. First, we review how individual computations can be isolated by identifying and linking neural codes to mental representations. Second, we review how the analysis of dynamic neural responses can recover the order of these computations. Throughout, we discuss how the recent developments in machine learning not only offer complementary methods to analyze convoluted patterns of neural activity, but also help to formalize the computational foundations of cognition.

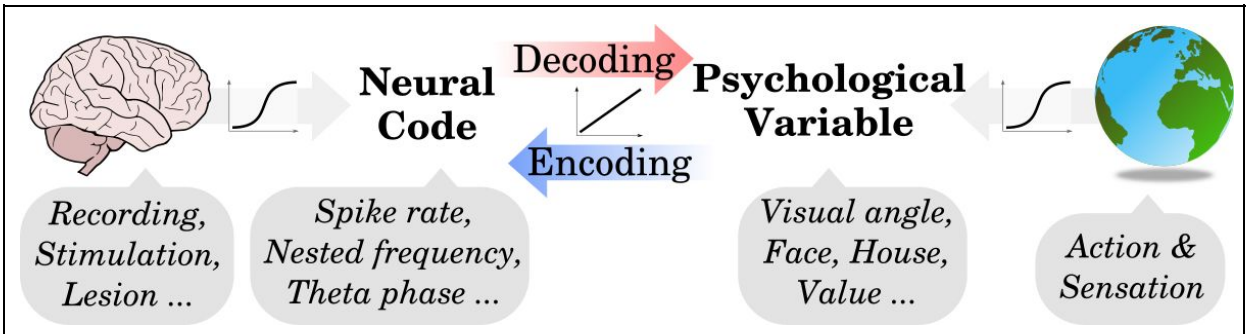


Figure 1. The representational paradigm consists of three challenges: (i) identifying the content of mental representations, (ii) determining how this information is encoded in the brain, and (iii) linking these two levels of description. These challenges increasingly benefit from statistical modeling and machine learning. In this view, decoding analyses predict experimental factors (e.g. the speed of a hand movement, the luminance of a flashed image etc) from specific brain activity features (e.g. spike rate, electric field, etc.), whereas encoding analyses predict the reverse. Currently, the relevant features of i) brain activity and ii) experimental factors are defined *a priori* by the experimenters and mapped onto one another with linear mapping. However, the capacity of machine learning to find non-linear structures in large datasets may ultimately help finding what the brain represents and how.

1. Neuronal activity: codes and contents.

1.1 The triple-quest of cognitive neuroscience.

Three challenges must be addressed to isolate the elementary computations underlying a given cognitive process (Fig. 1). First, we must identify what variables the brain generates and exploits to solve a particular task; in other words, what the psychologically relevant dimensions are. For example, speech has been formally described in terms of phonemes (e.g. /b/, /p/, /k/) but the existence of these psychological variables has been debated given their extensive overlap with low-level acoustic properties (e.g. (Hickok, 2014). In a recent study, Mesgarani et al. have shown that activity of the superior temporal gyrus in response to speech is better accounted for by phonetic features than by acoustic ones (Mesgarani, Cheung, Johnson, & Chang, 2014). This suggests that the brain computes phonetic variables to understand speech. More generally, the search for the relevant mental variables is ubiquitous in cognitive neuroscience. For example, studies have been able to characterize the neural bases of faces (Freiwald & Tsao, 2010; Haxby, 2006; Nancy Kanwisher, 2001), word strings (Dehaene & Cohen, 2007; Price, 2010), and semantics (Huth, de Heer, Griffiths, Theunissen, & Gallant, 2016), to name a few.

Second, we must identify *how* neurons read and communicate such informational content. For example, the fact that neurons usually do not discharge at precise moments has led some to claim that they transmit information through spike rate, rather than through the precise time at which they spike (Shadlen & Newsome, 1998). By contrast, the short duration of certain cognitive processes has led others to argue that spike timing may also carry additional information (Kistler & Gerstner, 2002). More generally, whether neurons and neural populations code information via their firing rates (Shadlen & Newsome, 1998), their oscillatory activity (Buzsaki, 2006; Fries, 2005; Singer & Gray, 1995), or even in the interaction between spikes and the phase of local field potentials (Bose & Recce, 2001; Lisman & Idiart, 1995) remains an outstanding research question.

Third, we must identify how these two levels of description—*what* is coded and *how* it is coded—relate with one another: i.e. we must find the patterns of neural activity (e.g. a spike or an oscillation) that are both sensitive and specific to putative variables (e.g. the position of a rat in a maze). In the context of correlational settings, such as neuroimaging and electrophysiological recordings, solving the code-content equation (Fig. 1.) is asymmetric. Either the code is assumed and multiple variables are comparatively tested, or *vice versa*. For example, one can assume a rate code and compare how the position of a rat in a maze predicts spike activity (O'Keefe & Dostrovsky, 1971). Reciprocally, one can assume that spatial locations are coded in the brain and compare how spike rates and the oscillations of the local field potentials predict this variable (Agarwal et al., 2014).

1.2 The statistical framework of encoding and decoding analyses.

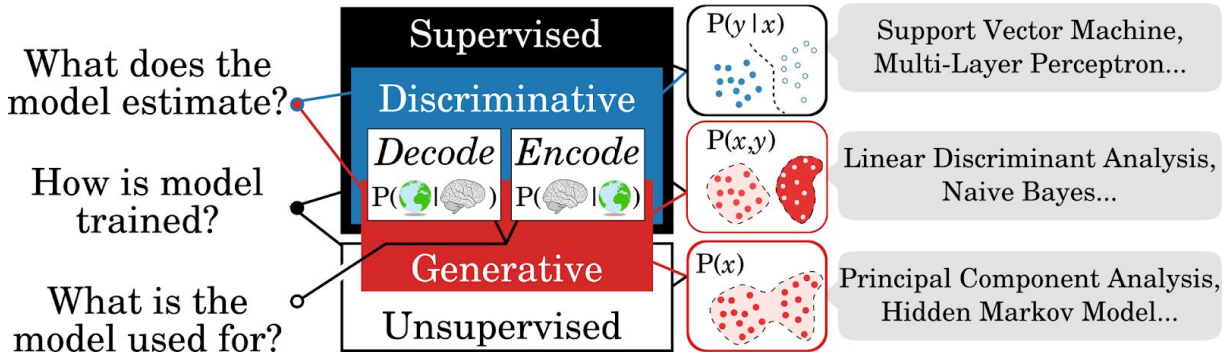
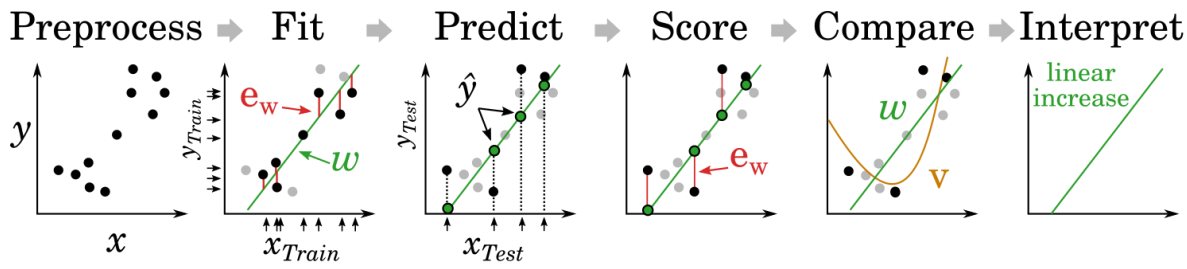


Figure 2. Left. The most common models used in cognitive neuroscience can be distinguished with three main dimensions of the statistical framework. Models that are trained to estimate a conditional probability P(y|x) between two sets of variables (e.g. x=psychological variables, y=neuronal recordings, or vice-versa) are referred to as 'discriminative'. By contrast, models that estimate the joint distribution P(x,y) are considered to be 'generative'. All generative models can thus be used then to derive P(y|x) for prediction. These models are 'supervised' in that they are trained to predict one variable (e.g. y) from another (e.g. x). By contrast, 'unsupervised' models estimate the distribution of a single (possibly multidimensional) variable (x). Finally, a trained model can ultimately be used for different purposes: e.g. decoding or encoding (see Fig. 1). **Right.** Examples of classical supervised and unsupervised models.

The asymmetry of the code-content mapping contributes to the distinction between encoding and decoding analyses. Specifically, encoding consists in predicting neuronal responses from internal (e.g. confidence) or environmental variables (e.g. the presence of an object): P(brain activity pattern | variables). Conversely, decoding consists in predicting variables from neuronal activity: P(variables | brain activity pattern) (Fig 1-2). Generally, encoding and decoding both depend on multivariate models whose objective is univariate, meaning that they fit several parameters to minimize a scalar that results from a loss function (Fig. 3) For example, fMRI studies routinely use encoding analyses by fitting a general linear model (GLM) to evaluate the extent to which multiple variables independently contribute to blood-oxygen-level dependent (BOLD) measurements. Such variables can be difficult to orthogonalize *a priori* (i) because of the slow temporal profile of the BOLD response or (ii) because the variables of interest can intrinsically covary (e.g. in natural images, the orientation of visual edges correlate with their spatial position: (Sigman, Cecchi, Gilbert, & Magnasco, 2001). Conversely, decoding analyses are predominantly used to maximally predict subjects' behavior or postdict their sensory stimulations. For example, brain-computer interfaces (BCI) studies typically examine several collinear patterns of brain activity in order to maximally predict

subjects' actions, intentions (Lebedev & Nicolelis, 2006) or mental state (Zander & Kothe, 2011).



Fitting Objective: Find parameters (w) to best predict targets (y) from input data (x).

$\max P(w \mid x, y)$, *where:* $x_i^T w = y_i$

Method: optimization

$P(w \mid x, y) \propto P(x, y \mid w)\, P(w)$

$\min_w Loss(y, x^T w) + Regularization(w)$

Name	Loss	Linear Model	Reg	Optimization
OLS:	l_2	$x_i^T w = \hat{y}_i$	*None*	$\min_w \sum_i (y_i - \hat{y}_i)^2$
Ridge:	l_2		l_2	$\min_w \sum_i (y_i - \hat{y}_i)^2 + \lambda \mid w \mid^2$
Logistic:	*log*		l_2	$\min_w \sum_i \log(1 + \exp(-y_i \hat{y}_i)) + \lambda \mid w \mid^2$
SVM:	*hinge*		l_2	$\min_w \sum_i \max(0, 1 - y_i \hat{y}_i) + \lambda \mid w \mid^2$

Figure 3. Top. Statistical modeling follows a standard multi-step pipeline which starts with 1) preprocessing (any transformation that can be defined independently of the data: e.g. filtering), followed by 2) model fitting on a subset of training data, 3) prediction and 4) scoring of independent and identically distributed held-out test data. The score (a.k.a "goodness-of-fit", "prediction error", here labelled as e_w) can be summarized with a variety of metrics (e.g. accuracy, AUC, R^2, cross-entropy, F-value etc.). Finally, one can subsequently perform model comparison to interpret results. **Middle.** The fitting stage can make use of a variety of models such as support vector machines (SVM) or ordinary least squares (OLS). These models can be formalized as maximum a posteriori (MAP), which balances the data likelihood (red) and the priors on the distribution of the model parameters (blue). These model parameters can be found with optimization, which consists of jointly minimizing the loss and regularization function. Note that for categorical models (e.g. logistic regression, SVM), the predicted values $\hat{y}$ are subsequently transformed into discrete categories. **Bottom.** Example of common linear models used in cognitive neuroscience, together with their corresponding loss, regularization and optimization functions.

A variety of multivariate linear analyses are routinely used in cognitive neuroscience, and range from linear discriminant analysis (LDA) and general linear model (GLM) to ridge and logistic regression and, more recently, to algorithms developed in the field of machine learning such as linear support vector machines (SVM). Despite their various denominations and historical origins, these analyses can be described within a common statistical framework (Fig.

3. Bottom). For example, they can be solved via the same convex optimization and identify the linear combination of features that maximally predict a brain response (e.g. encoding a spike) or a putative variable (e.g. decoding the presence of a face). In the context of electrophysiology and neuroimaging data, most of these analyses lead to similar results (Hastie, Tibshirani, & Friedman, 2009; Lebedev & Nicolelis, 2006; Gaël Varoquaux et al., 2017). However, distinct multivariate linear analyses assume distinct data distributions (e.g. LDA assumes normal distributions and equal variance-covariance matrices across classes, whereas logistic regression does not). Consequently, the choice of analysis depends on the problem (e.g. regression or classification), the amount of data, and its distribution (e.g. if the data are normally distributed, LDA can outperform logistic regression and *vice versa*). Interpreting the parameters of an analysis can be particularly challenging, because i) all parameters are simultaneously fitted, which makes the interpretation of individual parameters difficult, ii) some parameters may be related to the noise distribution, and iii) a given parameter need not actually impact the goodness of fit of a given model (Davis et al., 2014; Haufe et al., 2014; Hebart & Baker, 2018; Todd, Nystrom, & Cohen, 2013). For model interpretation, it is thus advised to supplement inspection of model parameters with an explicit model comparison evaluated on prediction error.

Encoding and decoding models are not always subject to comparable constraints, and can thus lead to different conclusions. In particular, decoding can pick up uncontrolled noise or signal structures in brain activity in a way that encoding cannot. For example, if an encoding model predicting the neural response to an image shows that its *luminosity* improves the prediction of brain activity, one can conclude that *luminosity* causally influences brain activity (provided that a number of assumptions are met, see (Weichwald et al., 2015). However, no causal conclusion may actually be drawn from an analogous decoding model: e.g. if including parietal neurons in a decoding model improves the decoding performance of image *luminosity*, parietal activity may not necessarily be caused by *luminosity*. Instead, parietal activity may simply reflect subjects' vigilance, which itself modulates the representation of *luminosity* in sensory regions; combining sensory and parietal regions may thus improve the decoding performance of *luminosity*. In this sensory-based paradigm, decoding can thus be less conclusive than encoding. However, this difference in conclusiveness comes with a benefit: because the decoding model can capture uncontrolled factors (e.g. vigilance), its predictive power may surpass the encoding model's (see Davis et al. (2014) for a related issue).

Overall, encoding and decoding models can therefore be profoundly distinct in their ability to detect and make use of uncontrolled factors and confounds. For a more detailed discussion on the causal inference and on the interpretation of encoding and decoding models, we refer the reader to (Haufe et al., 2014; Weichwald & Grosse-Wentrup, 2017; Weichwald et al., 2015).

1.3 Where do the linearity assumptions come from?

Most encoding and decoding analyses are based on linear modeling. This linear constraint is motivated by two theoretical principles: i) the general "linear superposition" principle, and ii) the neurocentric "linear readout" principle (Fig. 4).

Linear superposition. The linear superposition principle is a general assumption based on the notion that measurements derive from a weighted sum of underlying sources. For example, the electric potential measured by an electrode depends on the electric reference, the local field potential, as well as on the pre and postsynaptic activity of surrounding neurons. Following Maxwell's equations, under the quasi-static approximation, the electric fields of these sources linearly sum onto the electrode, and do not interact with one another. Similarly, the analysis of hemodynamic responses is often based on an analogous assumption: each voxel contains hundreds of thousands of neurons whose activity is summarized in a unique BOLD measurement. Under the linear superposition assumption, a measurement (from an electrode, from a voxel) linearly covaries with a variable only if one or a combination of sources (the underlying neural responses) linearly covaries with such variables. When multivariate measurements are available (e.g. spatially distributed electrodes), it is possible to separate the independent contribution of several sources, based on physical assumptions (as in magneto-encephalography (MEG) source reconstruction (Hämäläinen, Hari, Ilmoniemi, Knuutila, & Lounasmaa, 1993), or based on statistical assumptions (as in spike sorting: e.g. (Quiroga, Nadasdy, & Ben-Shaul, 2004). Note that the linear superposition assumption is generally applicable within a limited range. For example, the BOLD response is known to saturate above certain values, above which the linear superposition assumption fails to hold (Heeger & Ress, 2002).

Linear readout. The linear read-out principle is mainly relevant to decoding analyses. It builds upon the assumption that individual neurons can be approximated as a non-linear transformation (e.g. a spike) of a weighted sum of input (e.g. the sum of excitatory and inhibitory presynaptic potentials). Consequently, distributed patterns of simultaneous neuronal activity are thought to represent variables because any neuron connected to this distributed population can systematically covary with such variables (Hung, Kreiman, Poggio, & DiCarlo, 2005; Kamitani & Tong, 2005; King & Dehaene, 2014; Kriegeskorte & Kievit, 2013; Misaki, Kim, Bandettini, & Kriegeskorte, 2010). The linear readout principle clarifies the distinction between information and explicitly-represented features. For example, the retina may encode *information* about faces and letter strings, but would not explicitly *represent* these categories, in that faces and letter strings cannot be linearly separated from retinal activations. By contrast, the fusiform face regions (N. Kanwisher, McDermott, & Chun, 1997; Tsao, Freiwald, Tootell, & Livingstone, 2006) and the visual word form area (Dehaene & Cohen, 2011) have been shown to linearly map these two types of visual categories onto their patterns of neuronal activity.

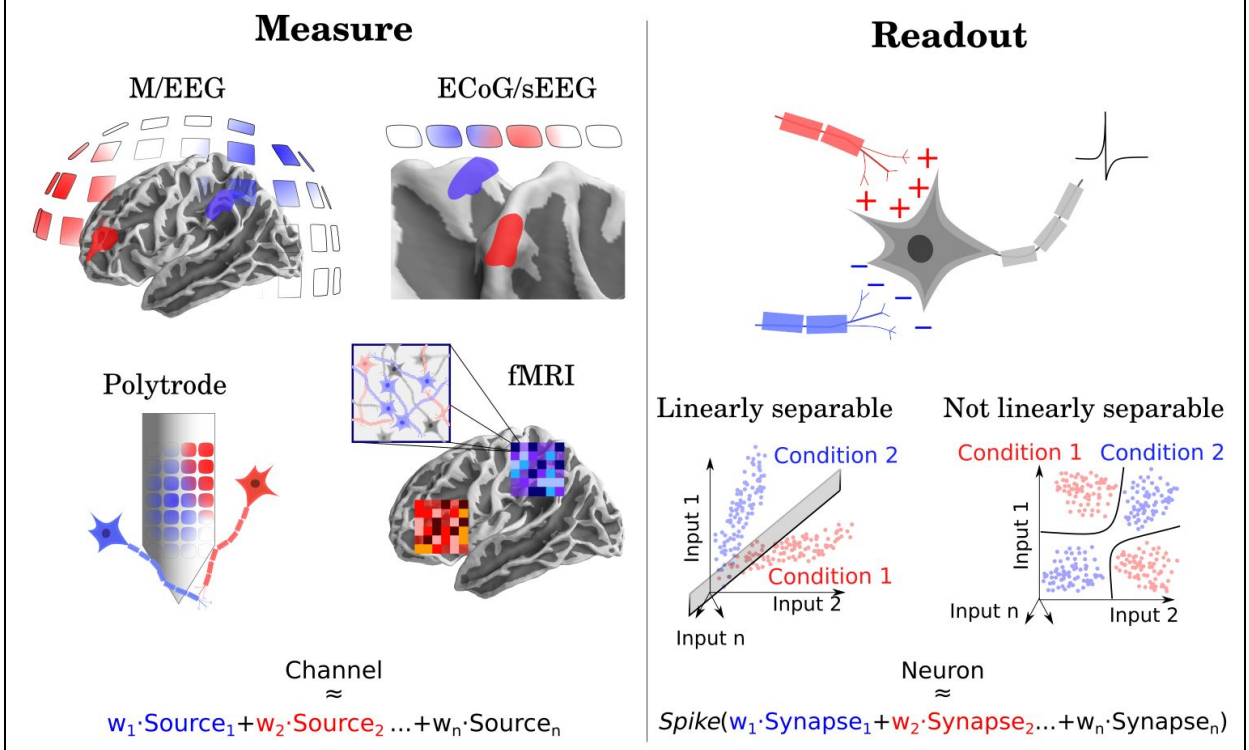


Figure 4. Left. The electro-magnetic fields produced by neurons linearly project onto polytrodes, electro-corticography (ECoG) or magneto-electroencephalography (M/EEG) sensors. A similar concept of linear mixture can be extended to fMRI: each voxel approximates a weighted sum of BOLD responses. Overall, measuring channels can thus be used to infer the underlying neural responses. **Right**. Neuronal computations can be approximated as non-linear functions (i.e. spike) applied onto weighted combinations of synaptic inputs. Statistically, this means that neurons are sensitive to (i.e. "read") linearly separable inputs.

It is important to highlight that encoding and decoding analyses are equally limited in their ability to determine whether a representation *de facto* constitutes information that the neural system *uses*. For example, one may find a linear relationship between a variable and (i) a spike, (ii) an increase in BOLD response, or (iii) an oscillation of a linear combination of EEG sensors, without that variable being effectively read and used by any neuron. Similarly to other correlational methods, encoding and decoding should thus be used in conjunction with comparative computational modeling and experimental manipulations in order to identify the causal or epiphenomenal nature of an identified pattern of brain activity.

1.4 Challenges of representational paradigm and the promises of Machine Learning.

Constraining the triple-quest of cognitive neuroscience (Fig. 1) to linear modeling leads to two main challenges. First, and as discussed elsewhere (Ritchie, Brendan Ritchie, Kaplan, & Klein, 2017), the linear readout assumption undermines the non-linear readout abilities of certain neurons (Brincat & Connor, 2004; Chichilnisky, 2001; Mineault, Khawaja, Butts, & Pack, 2012; Sahani & Linden, 2003; Van Steveninck & Bialek, 1988), cortical columns (Bastos et al., 2012) and large neural assemblies (Ritchie et al., 2017). The definition of an explicitly encoded variable is thus likely to change with our improved understanding of the neuronal codes.

Second, linear modeling implies a strong dependence on *a priori* human insight (Kording, Benjamin, Farhoodi, & Glaser, 2018). Specifically, linear models only fit the features explicitly provided by the experimenter. They are thus limited in their ability to identify unexpected patterns of neuronal activity, or unanticipated mental representations. For example, the discovery of grid cells — hippocampal neurons that fire when an animal is located at regularly-interspaced locations in an arena — resulted from human insights from visual data inspection. Indeed, Moser et al. had to view their electrophysiological recordings in a spatial representation before they could conjecture the grid coding scheme (Fyhn, Molden, Witter, Moser, & Moser, 2004; Moser, Kropff, & Moser, 2008). Only then did they implement a grid feature in a linear model to formally test and verify the robustness of this hypothesis (Hafting, Fyhn, Molden, Moser, & Moser, 2005). In other words, a linear model blindly fitting spiking activity to a two-dimensional spatial position variable would have missed the seminal discovery of grid-coding cells.

The rapid development of machine learning may partially roll back this epistemic dependence on human insights. For example, Benjamin and collaborators have recently investigated the ability of linear models to predict spiking activity in the macaque motor cortex given conventional variables of the arm movement, such as its instantaneous velocity and acceleration (Benjamin et al., 2017). The authors first show that linear encoding models can accurately predict the macaque's neural responses based on a weighted combination of these variables. However, they then demonstrate that linear models are outperformed by machine learning models that can efficiently capture non-linear relationships, such as random forests (Liaw, Wiener, & Others, 2002) and long short term memory neural networks (LSTMs, (Hochreiter & Schmidhuber, 1997). In other words, random forests and LSTMs can identify unsuspected features of the arm movements that are represented in the neural activity. More generally, this study illustrates how machine learning may supplement human insights and help to discover unanticipated representations.

Undoubtedly, applying machine-learning algorithms to cognitive neuroscientific data will lead to new challenges (Kording et al., 2018; Poldrack & Farah, 2015; Stevenson & Kording, 2011; Gael Varoquaux & Thirion, 2014). In particular, interpreting a multivariate model, and with greater reason, a non-linear one, can be particularly difficult. For example, in Benjamin et al.'s study discussed above, machine learning algorithms proved to be better at predicting the neural

activity associated with arm movements than linear models. However, this came at the cost of diminished interpretability: the exact nature of these unsuspected representations captured remain currently unclear. Note that interpretability is not a problem specific to non-linear models. For example, Huth and colleagues predicted the fMRI BOLD responses to spoken stories (Huth et al., 2016) from linear combinations of very large semantics vectors derived from latent semantic analysis of text corpora. The authors showed that this modeling approach was above chance level in a vast number of cortical regions, which thus strengthens the hypothesis of distributed representations of semantic features (Barsalou, 2017). However, to interpret such a model, one would need to investigate, for each voxel, the hundreds of coefficients associated with each semantic vector. To make things worse, these vectors are not directly interpretable. In fact, when the authors used an unsupervised linear model (principal component analysis) to summarize the main semantic dimensions that accounted for BOLD activity, they only managed to attribute a meaningful interpretation to a small subset of these principal components. Consequently, even linear modeling does not necessarily ensure a straightforward interpretation.

Overall, these two studies highlight how the interpretability of a neural representation, which has been essential for generating insights and novel hypotheses, runs a risk of becoming increasingly anecdotal as models are (justifiably) increasingly evaluated on the basis of their prediction accuracy.

2. From isolated computations to algorithms.

The above methods isolate the result of individual computations by linking putative variables with patterns of neural activity. However, to uncover the *algorithm* of a given cognitive ability, one must also identify the order in which these computations are performed. In this second section, we will first briefly review a variety of established sequences of neural activity and their algorithmic interpretations. We will then summarize the main methods that i) isolate specific neural sequences, ii) identify their selective input sequence, and iii) help interpret the computations associated with such neural dynamics.

2.1 Sequences of neural responses across spatial scales.

With the advances in temporally-resolved fMRI (Ekman, Kok, & de Lange, 2017) and the increasing ability to simultaneously record multiple neurons (Jun et al., 2017) and brain regions (Boto et al., 2018; Tybrandt et al., 2018), specific sequences of neural activity have been revealed across multiple spatial scales. At the network level for example, visual stimulations trigger a long cascade of neural responses from occipital to associative cortices (e.g. (Gramfort, Papadopoulo, Baillet, & Clerc, 2011; King, Pescetelli, & Dehaene, 2016), Fig. 5.C). This long sequence of brain responses has been successfully compared to the deep convolutional networks developed in artificial vision (Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017; Gwilliams & King, 2017; Kriegeskorte, 2015; Yamins et al., 2014). At the columnar level, neural activity has been shown to propagate from

and to the supra- and infragranular layers of the cortex via frequency-specific travelling waves (van Kerkoerle et al. (2014), Fig. 5.B), and has been argued to reflect a predictive coding algorithm (Bastos et al., 2012). Finally, at the cellular level, spatial positions (Girardeau & Zugaro, 2011; Jones & Wilson, 2005) are associated with specific sequences of spikes (Fig. 5.A) that may reflect learning and anticipatory simulation of spatial navigation (Girardeau & Zugaro, 2011).

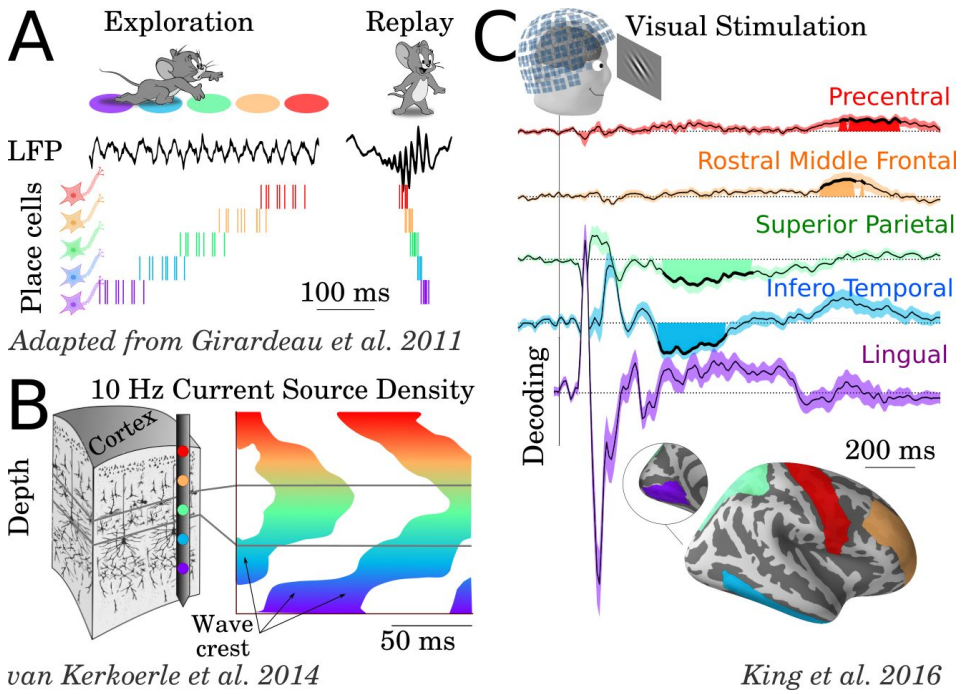


Figure 5. Sequences of neural responses across multiple spatial scales. Sequences of neural responses have been identified with a variety of methods and across various spatio-temporal scales. **A.** Hippocampal place cells spike when the rodent moves to specific spatial positions. When the rodent rests, the previous spike sequence is replayed in reverse order (adapted from (Girardeau & Zugaro, 2011). **B.** The 10-Hz oscillatory activity recorded in the macaque early visual cortex travels from superficial (red) and deep layers (purple) to the granular layer (green) (Adapted from (van Kerkoerle et al., 2014). **C.** Visual stimulation triggers a long-lasting cascade of macroscopic brain responses, starting from the early visual cortex and ultimately reaching the associative areas (King et al., 2016).

2.2 Isolating a sequence of processing stages from mixed neural activity.

Overall, these sequences of neural activations and their computational interpretations have been mainly established manually. However, a number of signal processing methods have recently been put forward to automatically detect and characterize specific spatio-temporal patterns of neural activity from high-dimensional recordings.

Analyzing neural recordings is often complicated by the fact that, as discussed above, they usually result from linear mixtures. For example, polytrodes, ECoG, and MEG all measure the fluctuations of the electro-magnetic fields of a linear combination of underlying neural electric sources (Fig 4. Left). To assess the spatial dynamics of such linear mixtures, it is thus common to use spatial filtering techniques. For example, one can decompose the evoked response to a visual stimulus by fitting a series of linear models, one at each time-sample locked to the stimulus onset, in order to track the rise and fall of a given variable (e.g. contrast: King et al., (2016) or the presence of a face: Cichy, Pantazis, & Oliva, (2014)) and test whether a model optimized to decode a variable at time *t* can generalize to decode at time *t'* (King & Dehaene, 2014; Meyers, Freedman, Kreiman, Miller, & Poggio, 2008; Stokes et al., 2013). If a model generalizes, this suggests that the brain activity pattern it isolated is present at both training and testing time samples. By systematically assessing how each model generalizes over time, i.e. by building a temporal generalization matrix, it is thus possible to determine whether the neural dynamics consists of a sequence of transient activations, a recurrent dynamic or simply a sustained activation directly from a linearly-mixed neural time series (Fig 6.A, (King & Dehaene, 2014).

The above methods focus on "evoked" spatial patterns - i.e. neural responses whose phases are consistent across trial repetitions. However, analogous spatial-filtering methods have been developed to specifically capture "induced" spatial patterns - i.e. linear mixtures of neural responses, such as oscillatory activity, whose dynamics are not phase-locked to an external event. These methods are generally based on the spatial covariance of electrophysiological recordings. For example, the common spatial pattern (CSP) method is a popular spatial filtering technique to identify the spatial pattern of neural activity that maximizes the discrimination between induced responses of two external events (e.g. left versus right hand, e.g. (Koles, Lazar, & Zhou, 1990). Similarly, the source power comodulation (SPoC) method extracts spatial patterns that are modulated by a continuous variable (e.g. hand position, e.g. (Dähne, Meinecke, et al., 2014). Recently, methods that directly use the spatial covariance as a feature proved to reach better results (Barachant, Bonnet, & Congedo, 2012; Barachant, Bonnet, Congedo, & Jutten, 2013; Farquhar, 2009).

These novel techniques have already proved valuable in establishing processing sequences. For example, a recent study (Heikel, Sassenhagen, & Fiebach, 2018) suggested that language processing depends on a sequence of distinct, but partially overlapping stages. Such a *cascading* architecture goes against the recurrent or the serial architecture argued by others (See (Friederici, 2011) for an overview). Similarly, visual selection has recently been shown to recruit parallel hierarchies of sensory processing converging into a serial executive processing stage (Marti & Dehaene, 2017; Marti, King, & Dehaene, 2015). Specifically, MEG recordings show that subjects can specifically attend to an image presented in a rapid stream, and represent its content in the associative cortices. However, temporal generalization showed that these late processing stages are systematically delayed when subjects are distracted by auditory stimuli, which suggests a strictly serial architecture. Overall, these studies thus illustrate how the automatic analyses of high-dimensional neural time series can help recover the underlying algorithmic organization of cognitive processes.

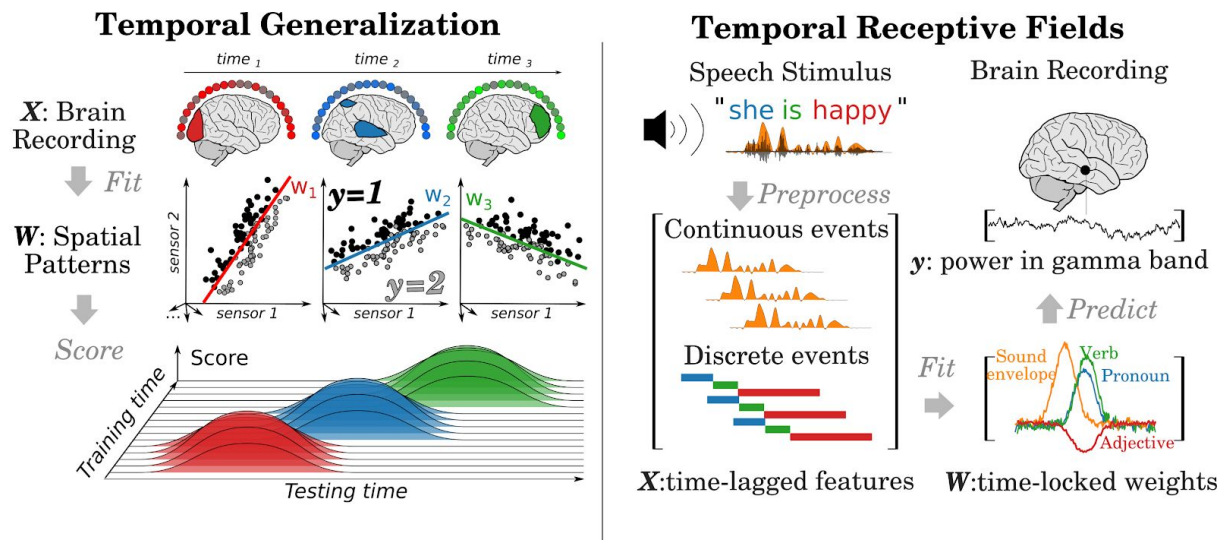


Figure 6. Decoding and encoding neural sequences. Left. Temporal generalization consists in fitting a spatial filter at each time sample locked to an event, and testing whether it can be used to decode the brain responses at all time samples. This analysis can be used to determine whether a sustained decoding score results from a stable pattern of neural responses, or whether it reflects a series of transient neural responses. Adapted from (King & Dehaene, 2014) **Right.** Temporal receptive field analyses involve first annotating the latent dimensions of that characterize the continuously-varying experimental variables - e.g. the spectral modulation of an acoustic waveform, the lexical categories of spoken words etc - (top left). These features are then transformed into a time-delay matrix whose linear modeling can be used to recover the temporal response profile of each feature (bottom right, adapted from (Holdgraf et al., 2017).

2.3 Identifying the sequence of inputs detected by a neuronal population.

A second major challenge consists in identifying the sequence of input detected by one or several neurons. This approach is particularly common in auditory and speech studies. In such cases, it is indeed common to model the neural dynamics with multiple "time-lags" of continuously-varying features (Almon, 1965; Ho & Kálmán, 1966), in order to capture the possibility that some neural responses are time-locked to a specific (combination of) events. For example, to model auditory and language processing, one can first (i) extract the continuously changing envelope of a recorded speech waveform and (ii) annotate the onsets of individual words (Fig 6.B, (de Heer, Huth, Griffiths, Gallant, & Theunissen, 2017; Holdgraf et al., 2017). A temporal "receptive field" model can finally be fitted to isolate the neural responses to these continuous fluctuations and discrete events respectively. Temporal receptive fields are similar to the predominant approach in fMRI analysis, where a design matrix is convolved with an impulse waveform which is either assumed (i.e. the canonical hemodynamic response function) or estimated from data, eventually improving fMRI encoding and decoding models (Pedregosa,

Eickenberg, Ciuciu, Thirion, & Gramfort, 2015). In the case of electrophysiology, the shapes of the impulse response functions are typically not known *a priori,* so the impulse response must be estimated from the data.

Temporal receptive fields and related methods can be used to (i) encode the average brain responses to categorical events (e.g. de Heer et al., 2017; Smith & Kutas, 2015) and spectro-temporal patterns of sensory input (Theunissen et al., 2001), as well as (ii) decode overlapping sequences of neural correlates of discrete (Rivet, Souloumiac, Attina, & Gibert, 2009; Theunissen et al., 2001) and continuously changing events (Dähne, Nikulin, et al., 2014). For example, DiLiberto et al. (2014) have shown that encoding categorical phonemic representations in EEG signals provides a superior model fit compared to acoustic features. Overall, these methods thus promise to help automatically find the sequence of input that maximally drive each brain response.

2.4 Toward directly mapping algorithms onto brain activity.

The last decade has been marked by an explosion of machine learning models that efficiently mimic basic cognitive operations. Most remarkably, computer vision models can now detect, locate, or describe objects in static natural images with accuracy that matches or even exceeds human performance (although see (Lake, Ullman, Tenenbaum, & Gershman, 2017). One of the most popular approaches, deep convolutional neural network, takes an image as input and sequentially applies a long series of non-linear transformations that are optimized to identify objects. Interestingly, this specific sequence of operations has been found to map onto both the spatial (Cichy et al., 2016; Eickenberg et al., 2017; Gwilliams & King, 2017; Kriegeskorte, 2015; Yamins et al., 2014) and the temporal organization (Cichy et al., 2016; Gwilliams & King, 2017; van de Nieuwenhuijzen et al., 2013) of the visual system in the mammalian brain. For example, the activity in the primary visual cortex specifically and linearly correlates with the activation in early layers of the artificial neural network, whereas the later responses of the inferior temporal cortex specifically and linearly correlate with the activation of the deepest layers. This result suggests that the sequence of computations applied by the human brain to solve a given task may be parsed and modeled with deep neural networks trained to solve the same task. This finding thus supports the notion that the visual system is organized as an extended hierarchy (Hubel & Wiesel, 1963; Riesenhuber & Poggio, 1999). More generally, it illustrates how sequences of brain activity patterns may be interpreted with performance-optimized algorithms: the algorithms that (i) can efficiently perform a task while (ii) doing so in a way that maps the spatio-temporal characteristics of brain activity, are likely to adequately model cognition.

Overall, the rapid development of machine learning provides a threefold promise to cognitive neuroscience. First, these tools support the automation, denoising, and summary of complex electrophysiological and neuroimaging time series (Jas, Engemann, Bekhti, Raimondo, & Gramfort, 2017). Second, these tools offer an operational ground to data-driven investigation: unanticipated patterns of data may be automatically identified from large datasets, without requiring the preface of human insight (Kording et al., 2018). Finally, machine learning and

cognitive neuroscience share the common goal of identifying the elementary components of knowledge acquisition and information processing. The interface between cognitive neuroscience and machine learning is thus mutually beneficial. On the one hand, machine learning can help define, identify, and formalize the computations of the brain. On the other hand, cognitive neuroscience can help provide insights and principled directions to shape the computational architecture of complex cognitive processes (Hassabis, Kumaran, Summerfield, & Botvinick, 2017; Lake et al., 2017).

References

Agarwal, G., Stevenson, I. H., Berényi, A., Mizuseki, K., Buzsáki, G., & Sommer, F. T. (2014). Spatially distributed local fields in the hippocampus encode rat position. *Science*, *344*(6184), 626–630.

Almon, S. (1965). The Distributed Lag Between Capital Appropriations and Expenditures. *Econometrica: Journal of the Econometric Society*, *33*(1), 178–196.

Barachant, A., Bonnet, S., & Congedo, M. (2012). Multiclass brain–computer interface classification by Riemannian geometry. *IEEE Transactions on*. Retrieved from http://ieeexplore.ieee.org/abstract/document/6046114/

Barachant, A., Bonnet, S., Congedo, M., & Jutten, C. (2013). Classification of covariance matrices using a Riemannian-based kernel for BCI applications. *Neurocomputing*, *112*, 172–178.

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

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

Benjamin, A. S., Fernandes, H. L., Tomlinson, T., Ramkumar, P., VerSteeg, C., Chowdhury, R., … Kording, K. P. (2017). Modern machine learning outperforms GLMs at predicting spikes. https://doi.org/10.1101/111450

Bose, A., & Recce, M. (2001). Phase precession and phase-locking of hippocampal pyramidal cells. *Hippocampus*, *11*(3), 204–215.

Boto, E., Holmes, N., Leggett, J., Roberts, G., Shah, V., Meyer, S. S., … Brookes, M. J. (2018). Moving magnetoencephalography towards real-world applications with a wearable system. *Nature*, *555*(7698), 657–661.

Brincat, S. L., & Connor, C. E. (2004). Underlying principles of visual shape selectivity in posterior inferotemporal cortex. *Nature Neuroscience*, *7*(8), 880–886.

Buzsaki, G. (2006). *Rhythms of the Brain*. Oxford University Press.

Chichilnisky, E. J. (2001). A simple white noise analysis of neuronal light responses. *Network*, *12*(2), 199–213.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755.

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462.

Dähne, S., Meinecke, F. C., Haufe, S., Höhne, J., Tangermann, M., Müller, K.-R., & Nikulin, V. V. (2014). SPoC: a novel framework for relating the amplitude of neuronal oscillations to behaviorally relevant parameters. *NeuroImage*, *86*, 111–122.

Dähne, S., Nikulin, V. V., Ramírez, D., Schreier, P. J., Müller, K.-R., & Haufe, S. (2014). Finding brain oscillations with power dependencies in neuroimaging data. *NeuroImage*, *96*, 334–348.

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.

Dehaene, S., & Cohen, L. (2007). Cultural recycling of cortical maps. *Neuron*, *56*(2), 384–398.

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, *15*(6), 254–262.
de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *37*(27), 6539–6557.
Donders, F. C. (1969). On the speed of mental processes. *Acta Psychologica*, *30*, 412–431.
Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, *152*, 184–194.
Ekman, M., Kok, P., & de Lange, F. P. (2017). Time-compressed preplay of anticipated events in human primary visual cortex. *Nature Communications*, *8*, 15276.
Farquhar, J. (2009). A linear feature space for simultaneous learning of spatio-spectral filters in BCI. *Neural Networks: The Official Journal of the International Neural Network Society*, *22*(9), 1278–1285.
Freiwald, W. A., & Tsao, D. Y. (2010). Functional compartmentalization and viewpoint generalization within the macaque face-processing system. *Science*, *330*(6005), 845–851.
Friederici, A. D. (2011). The brain basis of language processing: from structure to function. *Physiological Reviews*, *91*(4), 1357–1392.
Fries, P. (2005). A mechanism for cognitive dynamics: neuronal communication through neuronal coherence. *Trends in Cognitive Sciences*, *9*(10), 474–480.
Fyhn, M., Molden, S., Witter, M. P., Moser, E. I., & Moser, M.-B. (2004). Spatial representation in the entorhinal cortex. *Science*, *305*(5688), 1258–1264.
Girardeau, G., & Zugaro, M. (2011). Hippocampal ripples and memory consolidation. *Current Opinion in Neurobiology*, *21*(3), 452–459.
Gramfort, A., Papadopoulo, T., Baillet, S., & Clerc, M. (2011). Tracking cortical activity from M/EEG using graph cuts with spatiotemporal constraints. *NeuroImage*, *54*(3), 1930–1941.
Gwilliams, L., & King, J.-R. (2017). Performance-optimized neural network only partially account for the spatio-temporal organization of visual processing in the human brain. *BioRxiv*. https://doi.org/10.1101/221630
Hafting, T., Fyhn, M., Molden, S., Moser, M.-B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801–806.
Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., & Lounasmaa, O. V. (1993). Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, *65*(2), 413–497.
Hassabis, D., Kumaran, D., Summerfield, C., & Botvinick, M. (2017). Neuroscience-Inspired Artificial Intelligence. *Neuron*, *95*(2), 245–258.
Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised Learning. In T. Hastie, R. Tibshirani, & J. Friedman (Eds.), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (pp. 485–585). New York, NY: Springer New York.
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. (2006). Fine structure in representations of faces and objects. *Nature Neuroscience*, *9*(9), 1084–1086.
Hebart, M. N., & Baker, C. I. (2018). Deconstructing multivariate decoding for the study of brain function. *NeuroImage*, *180*(Pt A), 4–18.
Heeger, D. J., & Ress, D. (2002). What does fMRI tell us about neuronal activity? *Nature Reviews. Neuroscience*, *3*, 142.
Heikel, E., Sassenhagen, J., & Fiebach, C. J. (2018). Time-generalized multivariate analysis of EEG responses reveals a cascading architecture of semantic mismatch processing. *Brain and Language*, *184*, 43–53.
Hickok, G. (2014). The architecture of speech production and the role of the phoneme in speech processing. *Language and Cognitive Processes*, *29*(1), 2–20.
Ho, B. L., & Kálmán, R. E. (1966). Effective construction of linear state-variable models from input/output functions. *At-Automatisierungstechnik*, *14*(1-12), 545–548.
Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8),

1735–1780.
Holdgraf, C. R., Rieger, J. W., Micheli, C., Martin, S., Knight, R. T., & Theunissen, F. E. (2017). Encoding and Decoding Models in Cognitive Electrophysiology. *Frontiers in Systems Neuroscience*, *11*, 61.
Hubel, D. H., & Wiesel, T. N. (1963). Shape and arrangement of columns in cat's striate cortex. *The Journal of Physiology*, *165*, 559–568.
Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866.
Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.
Jas, M., Engemann, D. A., Bekhti, Y., Raimondo, F., & Gramfort, A. (2017). Autoreject: Automated artifact rejection for MEG and EEG data. *NeuroImage*, *159*, 417–429.
Jones, M. W., & Wilson, M. A. (2005). Theta Rhythms Coordinate Hippocampal–Prefrontal Interactions in a Spatial Memory Task. *PLoS Biology*, *3*(12), e402.
Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., … Harris, T. D. (2017). Fully integrated silicon probes for high-density recording of neural activity. *Nature*, *551*(7679), 232–236.
Kamitani, Y., & Tong, F. (2005). Decoding the visual and subjective contents of the human brain. *Nature Neuroscience*, *8*(5), 679–685.
Kanwisher, N. (2001). Faculty of 1000 evaluation for Distributed and overlapping representations of faces and objects in ventral temporal cortex. *F1000 - Post-Publication Peer Review of the Biomedical Literature*. https://doi.org/10.3410/f.1000496.16554
Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *17*(11), 4302–4311.
King, J.-R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: the temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210.
King, J.-R., Pescetelli, N., & Dehaene, S. (2016). Brain Mechanisms Underlying the Brief Maintenance of Seen and Unseen Sensory Information. *Neuron*, *92*(5), 1122–1134.
Kistler, W. M., & Gerstner, W. (2002). Stable propagation of activity pulses in populations of spiking neurons. *Neural Computation*, *14*(5), 987–997.
Koles, Z. J., Lazar, M. S., & Zhou, S. Z. (1990). Spatial patterns underlying population differences in the background EEG. *Brain Topography*, *2*(4), 275–284.
Kording, K. P., Benjamin, A., Farhoodi, R., & Glaser, J. I. (2018). The Roles of Machine Learning in Biomedical Science. In *Frontiers of Engineering: Reports on Leading-Edge Engineering from the 2017 Symposium*. National Academies Press.
Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*, 417–446.
Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412.
Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *The Behavioral and Brain Sciences*, *40*, e253.
Lebedev, M. A., & Nicolelis, M. A. L. (2006). Brain-machine interfaces: past, present and future. *Trends in Neurosciences*, *29*(9), 536–546.
Liaw, A., Wiener, M., & Others. (2002). Classification and regression by randomForest. *R News*, *2*(3), 18–22.
Lisman, J., & Idiart, M. (1995). Storage of 7 /- 2 short-term memories in oscillatory subcycles. *Science*, *267*(5203), 1512–1515.
Marr, D. (1982). Vision: A computational investigation into the human representation and processing of visual information. *WH San Francisco: Freeman and Company*. Retrieved from https://mitpress.mit.edu/books/978-0-262-29037-1
Marti, S., & Dehaene, S. (2017). Discrete and continuous mechanisms of temporal selection in rapid visual streams. *Nature Communications*, *8*(1), 1955.
Marti, S., King, J.-R., & Dehaene, S. (2015). Time-Resolved Decoding of Two Processing Chains during Dual-Task Interference. *Neuron*, *88*(6), 1297–1307.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in human superior temporal gyrus. *Science*, *343*(6174), 1006–1010.
Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K., & Poggio, T. (2008). Dynamic population coding of category information in inferior temporal and prefrontal cortex. *Journal of Neurophysiology*, *100*(3), 1407–1419.
Mineault, P. J., Khawaja, F. A., Butts, D. A., & Pack, C. C. (2012). Hierarchical processing of complex motion along the primate dorsal visual pathway. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(16), E972–E980.
Misaki, M., Kim, Y., Bandettini, P. A., & Kriegeskorte, N. (2010). Comparison of multivariate classifiers and response normalizations for pattern-information fMRI. *NeuroImage*, *53*(1), 103–118.
Moser, E. I., Kropff, E., & Moser, M.-B. (2008). Place cells, grid cells, and the brain's spatial representation system. *Annual Review of Neuroscience*, *31*, 69–89.
O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map. Preliminary evidence from unit activity in the freely-moving rat. *Brain Research*, *34*(1), 171–175.
Pedregosa, F., Eickenberg, M., Ciuciu, P., Thirion, B., & Gramfort, A. (2015). Data-driven HRF estimation for encoding and decoding models. *NeuroImage*, *104*, 209–220.
Poldrack, R. A., & Farah, M. J. (2015). Progress and challenges in probing the human brain. *Nature*, *526*(7573), 371–379.
Price, C. J. (2010). The anatomy of language: a review of 100 fMRI studies published in 2009. *Annals of the New York Academy of Sciences*, *1191*, 62–88.
Quiroga, R. Q., Nadasdy, Z., & Ben-Shaul, Y. (2004). Unsupervised spike detection and sorting with wavelets and superparamagnetic clustering. *Neural Computation*, *16*(8), 1661–1687.
Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025.
Ritchie, J. B., Brendan Ritchie, J., Kaplan, D., & Klein, C. (2017). Decoding The Brain: Neural Representation And The Limits Of Multivariate Pattern Analysis In Cognitive Neuroscience. https://doi.org/10.1101/127233
Rivet, B., Souloumiac, A., Attina, V., & Gibert, G. (2009). xDAWN algorithm to enhance evoked potentials: application to brain-computer interface. *IEEE Transactions on Bio-Medical Engineering*, *56*(8), 2035–2043.
Sahani, M., & Linden, J. F. (2003). How Linear are Auditory Cortical Responses? In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in Neural Information Processing Systems 15* (pp. 125–132). MIT Press.
Shadlen, M. N., & Newsome, W. T. (1998). The variable discharge of cortical neurons: implications for connectivity, computation, and information coding. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *18*(10), 3870–3896.
Sigman, M., Cecchi, G. A., Gilbert, C. D., & Magnasco, M. O. (2001). On a common circle: natural scenes and Gestalt rules. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(4), 1935–1940.
Singer, W., & Gray, C. M. (1995). Visual feature integration and the temporal correlation hypothesis. *Annual Review of Neuroscience*, *18*, 555–586.
Smith, N. J., & Kutas, M. (2015). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. *Psychophysiology*, *52*(2), 169–181.
Sternberg, S. (1998). Discovering mental processing stages: The method of additive factors. https://doi.org/ 1999-02657-014
Stevenson, I. H., & Kording, K. P. (2011). How advances in neural recording affect data analysis. *Nature Neuroscience*, *14*, 139.
Stokes, M. G., Kusunoki, M., Sigala, N., Nili, H., Gaffan, D., & Duncan, J. (2013). Dynamic coding for cognitive control in prefrontal cortex. *Neuron*, *78*(2), 364–375.
Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network*, *12*(3), 289–316.
Todd, M. T., Nystrom, L. E., & Cohen, J. D. (2013). Confounds in multivariate pattern analysis: Theory and rule representation case study. *NeuroImage*, *77*, 157–165.

Tsao, D. Y., Freiwald, W. A., Tootell, R. B. H., & Livingstone, M. S. (2006). A cortical region consisting entirely of face-selective cells. *Science*, *311*(5761), 670–674.

Tybrandt, K., Khodagholy, D., Dielacher, B., Stauffer, F., Renz, A. F., Buzsáki, G., & Vörös, J. (2018). High-Density Stretchable Electrode Grids for Chronic Neural Recording. *Advanced Materials* , *30*(15), e1706520.

van de Nieuwenhuijzen, M. E., Backus, A. R., Bahramisharif, A., Doeller, C. F., Jensen, O., & van Gerven, M. A. J. (2013). MEG-based decoding of the spatiotemporal dynamics of visual category perception. *NeuroImage*, *83*, 1063–1073.

van Kerkoerle, T., Self, M. W., Dagnino, B., Gariel-Mathis, M.-A., Poort, J., van der Togt, C., & Roelfsema, P. R. (2014). Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(40), 14332–14341.

Van Steveninck, R., & Bialek, W. (1988). Real-time performance of a movement-sensitive neuron in the blowfly visual system: coding and information transfer in short spike sequences. *Proc. R. Soc*. Retrieved from http://rspb.royalsocietypublishing.org/content/234/1277/379.short

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*(Pt B), 166–179.

Varoquaux, G., & Thirion, B. (2014). How machine learning is shaping cognitive neuroimaging. *GigaScience*, *3*, 28.

Weichwald, S., & Grosse-Wentrup, M. (2017). *The right tool for the right question --- beyond the encoding versus decoding dichotomy*. *arXiv [q-bio.NC]*. Retrieved from http://arxiv.org/abs/1704.08851

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.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624.

Zander, T. O., & Kothe, C. (2011). Towards passive brain-computer interfaces: applying brain-computer interface technology to human-machine systems in general. *Journal of Neural Engineering*, *8*(2), 025005.