NeuroImage xxx (2017) 1–13

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

# Ghosts in machine learning for cognitive neuroscience: Moving from data to theory

Thomas Carlson[a,b,*,1], Erin Goddard[b,c,1], David M. Kaplan[b,d,e,1], Colin Klein[b,f,1], J. Brendan Ritchie[g,1]

[a] School of Psychology, The University of Sydney, Sydney, NSW, 2006, Australia
[b] ARC Centre of Excellence in Cognition and its Disorders, Macquarie University, Sydney, NSW, 2109, Australia
[c] McGill Vision Research Group, McGill University, Montreal, QC, Canada
[d] Department of Cognitive Science, Macquarie University, Sydney, NSW, 2109, Australia
[e] Perception in Action Research Centre, Macquarie University, Sydney, NSW, 2109, Australia
[f] Department of Philosophy, Macquarie University, Sydney, NSW 2109, Australia
[g] Laboratory of Biological Psychology, KU Leuven, 3000 Leuven, Flemish Brabant, Belgium

ARTICLE INFO

*Keywords:*
Multivariate pattern analysis
Brain decoding
Exploratory methods
fMRI
Magnetoencephalography

ABSTRACT

The application of machine learning methods to neuroimaging data has fundamentally altered the field of cognitive neuroscience. Future progress in understanding brain function using these methods will require addressing a number of key methodological and interpretive challenges. Because these challenges often remain unseen and metaphorically "haunt" our efforts to use these methods to understand the brain, we refer to them as "ghosts". In this paper, we describe three such ghosts, situate them within a more general framework from philosophy of science, and then describe steps to address them. The first ghost arises from difficulties in determining what information machine learning classifiers use for decoding. The second ghost arises from the interplay of experimental design and the structure of information in the brain – that is, our methods embody implicit assumptions about information processing in the brain, and it is often difficult to determine if those assumptions are satisfied. The third ghost emerges from our limited ability to distinguish information that is merely decodable from the brain from information that is represented and used by the brain. Each of the three ghosts place limits on the interpretability of decoding research in cognitive neuroscience. There are no easy solutions, but facing these issues squarely will provide a clearer path to understanding the nature of representation and computation in the human brain.

## 1. Introduction: data, pattern, theory

Textbooks present scientific confirmation as a matter of fitting theory to data. Savvy philosophers and scientists have long known better. High-level theories do not make direct predictions about data. To borrow a framework from philosophy of science, scientific inference is not a one-step process from data to theory but a *two-step* process from data to *phenomenon* to theory (Bogen and Woodward, 1988; Suppes, 1962). For example, the Standard Model in physics is not tested directly against the voluminous data from particle colliders. Instead, that collider data is processed to give evidence for some stable, replicable phenomenon – $Z^0$ decay, for example – and then the Standard Model is checked to see if it can account for that phenomenon. Similarly, plate tectonics did not explain magnetometer readings but rather *the spreading of the mid-Atlantic ridge*. General relativity did not explain a series of telescopic observations but the *precession of Mercury*.

So too with various types of data in cognitive neuroscience. What one typically aims to explain is not raw data itself (e.g., changes in BOLD signal), or even a particular set of results from a single experiment. Rather, the goal is arguably to uncover and explain stable and replicable patterns of activation in response to a stimulus or task. It is of only mild interest that inferior temporal (IT) cortex was activated in this or that experiment. It is, however, of great importance that IT cortex is reliably activated by a wide variety of object recognition tasks.

Many early critiques of neuroimaging focused on these two inferential steps as they applied to univariate analyses of brain activation. Insofar as

* Corresponding author. 318 Griffith-Taylor Bldg., School of Psychology, University of Sydney, Sydney, Australia, 2006.
*E-mail address:* thomas.carlson@sydney.edu.au (T. Carlson).

[1] All authors contributed equally to this work.

http://dx.doi.org/10.1016/j.neuroimage.2017.08.019
Received 17 February 2017; Received in revised form 17 July 2017; Accepted 4 August 2017
Available online xxxx

simple univariate analyses seemed problematic, it was precisely because of weak links in the inference from data to replicable phenomenon (Klein, 2010; Logothetis et al., 2001; Nair, 2005; Poldrack, 2006). At the same time as the weaknesses in univariate analyses were becoming apparent, developments in machine learning techniques were changing the world of science, technology, medicine, and industry (Jordan and Mitchell, 2015). Perhaps unsurprisingly, machine learning methods have also found their way into cognitive neuroscience, most prominently under the banner of multivariate pattern analysis (MVPA) or "brain decoding". Some uses of machine learning in neuroscience directly address practical problems. For example, machine learning methods can be used to decipher patterns in neural data for clinical diagnosis and rehabilitation purposes including brain-machine interfaces (Hatsopoulos and Donoghue, 2009). Such uses are judged solely by their utility, and are otherwise unconstrained in the data and methods they use. We mention these to put them aside. Our focus will be on the application of decoding methods in the pursuit of basic knowledge about brain function.

Machine learning methods have become popular in part because they do not require many of the problematic auxiliary assumptions that plague univariate analyses. Specifically, MVPA arguably does not require strong commitments about the viability of reverse inference (Poldrack, 2006). Nor does MVPA assume a simple relationship between brain activity and the BOLD response (Logothetis et al., 2001), or the specifics of process decomposition (Sternberg, 2011). Further, MVPA allows researchers to deal with extremely large datasets utilising a wide range of techniques including structural MRI, DTI, fMRI, EEG, and MEG. The combination of large datasets and comparatively fewer assumptions gives machine learning methods an air of objectivity: rather than relying on old assumptions about cognitive architecture, we might simply let the brain tell us which categories provide the best fit (Anderson, 2014).

Yet machine learning does not directly connect theory and data any more than univariate analyses. The primary outcome from machine learning analyses is not (we suggest) a direct test of theory but rather evidence concerning stable patterns of brain activity – phenomena, in the above parlance. Such patterns are typically characterised in terms of a neural population's representational space: that is, how activity in the population activity relates both to the world and to other neural representations. The phenomena thus uncovered are what provide a basis for our tests of theories about cognition and brain function.

Machine learning brings with it its own set of problems. Precisely because it offers up simple patterns, it can be easy to read too much into data – to see phenomena that are not really there. This article outlines three of these metaphorical "ghosts" in machine learning techniques, as applied in cognitive neuroscience. The first involves the source of MVPA data itself, and the need to achieve greater specificity about the information we are measuring in the brain. The second involves the move from data to phenomenon, in particular when using dimensionality reduction techniques to go from complex datasets to simple patterns. The third and final challenge comes in moving from phenomenon to theory, and the difference between measuring information in the brain and inferring how the brain might actually use this information. Each of the three ghosts place limits on the interpretability of decoding research in cognitive neuroscience. Although there are no easy solutions, awareness of these issues will provide a clearer path to understanding the nature of representation and computation in the human brain.

Most will be familiar with some of these challenges, and some will be familiar with all of them. Many researchers have expressed related concerns about the interpretation of MVPA decoding results in cognitive neuroscience, as well as offering similar recommendations that this issue must be handled with care (e.g., Davis and Poldrack, 2014; de-Wit et al., 2016; Dubois et al., 2015; Guest and Love, 2017; Haynes, 2015; Poldrack and Farah, 2015; Ritchie et al., in press). One of our goals in this paper is to show that these problems can be fit into a common framework that connects them to ones faced previously by other, more well-established scientific disciplines. This is not an exercise in pessimism, however. We think that by clarifying the different steps of scientific inference and identifying the points at which problems often arise, we can arrive at useful constraints on the design and interpretation of machine learning studies.

Finally, in highlighting several field-specific challenges facing decoding research, we do not mean to imply that other interpretive and inferential issues associated with neuroimaging in general are somehow irrelevant. Importantly, the inferences licensed by decoding methods – like all neuroimaging methods – are limited by the fact that they are inherently correlational (Poldrack, 2011). Consequently, demonstrating significant decoding in a given brain region during task performance cannot by itself establish that it plays a causal role in that performance. Interventions, which include transcranial magnetic stimulation, reversible inactivation, lesions, and optogenetics, provide essential causal information that complements the evidence supplied by decoding studies (Pearl, 1995; Spirtes et al., 2000; Woodward, 2003). Related general critiques of decoding research based on their reliance on reverse inference (e.g., Poldrack, 2006, 2008) may also be germane, but fall outside the scope of this article to address. Importantly, we are squarely focused on internal steps that decoding researchers can take to overcome the field-specific interpretative and inferential challenges described above – without depending on help from other methods.

## 2. The ghost of source ambiguity

In science, data is the foundation upon which we discover phenomena and test theories. The same is true in cognitive neuroscience. But what exactly is the nature of the data we rely on in decoding research? Although there is consensus that machine learning methods measure information in the brain, it is quite common for there to be uncertainty about the underlying source of this information. The first ghost arises from the gap between our ability to measure information and our capacity to determine the underlying neural source. The former enables us to tell whether, and perhaps even how much, decodable information is present about the stimulus or task condition in a brain representation. Yet only the latter – identifying the neural source of this information – permits the data to act as a foundation for interpretation and brings us closer to the aim of understanding neural representations and processes.

Ascertaining the true neural source of decodable information, however, is extremely difficult because the mere presence of decodable information is ambiguous between potential sources (Bartels et al., 2008; Naselaris and Kay, 2015; Op de Beeck, 2010). To illustrate this, consider a hypothetical scenario from another branch of science. Suppose a simple linear classifier such as Gaussian Naïve Bayes (GNB) is successfully trained to predict whether a hurricane will form based on data from a large array of meteorological sensors. At this stage, we would have learned that information about hurricanes is present in the multivariate data collected from the sensors. Although this result would be useful for all kinds of practical purposes, we would not have appreciably deepened our understanding of hurricanes. At a minimum, if the classifier is to help us understand hurricanes, we would have to determine what information in the sensor data is driving the classification. To do this, one might inspect the classifier weights. Perhaps one would then find that a combination of dew point and humidity drove the classification. Only now would we begin to understand the relationship between these meteorological variables and hurricanes, and thereby add to our knowledge of hurricanes. Moreover, having identified these variables as important factors for hurricanes puts us in the position to study how these factors interact with other variables (e.g. wind speed, atmospheric pressure, etc.), potentially deepening our knowledge of hurricanes still further. The lesson here is that not all data is equal; even useful and predictive data can fail to give us the sort of information we need for advancing understanding.

### 2.1. *Case study: source ambiguity in orientation decoding*

The most rigorous investigation of the link between decodable