frontiers in Human Neuroscience

PERSPECTIVE
published: 08 July 2021
doi: 10.3389/fnhum.2021.682661

# Overfitting the Literature to One Set of Stimuli and Data

*Tijl Grootswagers*[1,2,3]* *and Amanda K. Robinson*[3]

[1] The MARCS Institute for Brain, Behaviour and Development, Sydney, NSW, Australia, [2] School of Psychology, Western Sydney University, Sydney, NSW, Australia, [3] School of Psychology, University of Sydney, Sydney, NSW, Australia

A large number of papers in Computational Cognitive Neuroscience are developing and testing novel analysis methods using one specific neuroimaging dataset and problematic experimental stimuli. Publication bias and confirmatory exploration will result in overfitting to the limited available data. We highlight the problems with this specific dataset and argue for the need to collect more good quality open neuroimaging data using a variety of experimental stimuli, in order to test the generalisability of current published results, and allow for more robust results in future work.

**Keywords: MEG, fMRI, EEG, vision, objects**

OPEN ACCESS

***Edited by:***
*Adolfo M. García,*
*Universidad de San Andrés, Argentina*

***Reviewed by:***
*Sara Invitto,*
*University of Salento, Italy*
*Sebastian Moguilner,*
*Comisión Nacional de Energía Atómica (CNEA), Argentina*
*Eugenia Fatima Hesse Rizzi,*
*Universidad de San Andrés, Argentina*

****Correspondence:***
*Tijl Grootswagers*
*t.grootswagers@westernsydney.edu.au*

***Specialty section:***
*This article was submitted to Cognitive Neuroscience, a section of the journal Frontiers in Human Neuroscience*

***Received:*** *18 March 2021*
***Accepted:*** *16 June 2021*
***Published:*** *08 July 2021*

***Citation:***
*Grootswagers T and Robinson AK (2021) Overfitting the Literature to One Set of Stimuli and Data. Front. Hum. Neurosci. 15:682661. doi: 10.3389/fnhum.2021.682661*

## BACKGROUND

How many ways are there to look at one set of data? In a highly influential paper (Kriegeskorte et al., 2008b) compared human fMRI responses to electrophysiological recordings from Monkey inferior temporal cortex (obtained from Kiani et al., 2007), revealing a striking similarity in the representations of objects between species (Kriegeskorte et al., 2008b). In addition, this study introduced the Representational Similarity Analysis (RSA) framework (Kriegeskorte et al., 2008a) to compare information representations between fMRI and electrophysiological recordings. This framework is now a widely used method for comparing information across modalities (Kriegeskorte and Kievit, 2013). The RSA framework was later used in a landmark study that used MEG and fMRI to track object representations in space and time (Cichy et al., 2014). For ease of comparing their results, this study used the same stimulus set (**Figure 1**) as Kriegeskorte et al. (2008b). So far so good, but these stimuli and corresponding fMRI and MEG data have now formed the basis for over 35 publications (estimated by going through a *Scopus* list of citations to these papers; **Figure 2**). These studies have yielded important information about how new analysis methods can be used to give insight into the visual system. Yet, it is undeniable that overuse of the same stimuli and data will eventually lead to a bias in the literature. A major factor to consider when designing a study is how generalisable the findings will be. Any one study is characterised by details (and limitations) of the experimental design, data collection, and analyses. Analysing the same sets of data in different ways or using the same stimulus sets will lead to over-representation and over-generalisation of experiment-specific trends. The intention of this commentary is not to undermine or refute any of these studies, but rather to point out that the field needs to diversify.

## A PROBLEMATIC STIMULUS SET

The stimulus set described above consists of 92 segmented visual objects of animals, people, places, and things (**Figure 1**). At first glance, there is nothing wrong with the set itself. To study object representations, we need stimuli, but controlling for all possible covariates in a set of stimuli is challenging, thus no set is perfect. In fact, the 92 objects were a considerable advance over previous work that had stronger limits imposed by the experimental designs. However, there are still issues with this set, as highlighted in **Figure 1**. Some stimuli were reported to be ambiguous (Kriegeskorte et al., 2008b). Some do not, on close inspection, belong in the manually specified categories, for example an image of hair is classified as "human body." There are also clear categorical differences between simple image features that covary with the imposed category structure. For example, many animate stimuli contain faces, which on average are visually more similar than stimuli within the inanimate category (**Figure 1**). These reliable visual similarities likely lead to large-scale pattern differences in neuroimaging data (cf. Vanrullen, 2011), which could account for the strong animate-inanimate distinction that is often observed in studies that used the 92-object stimuli (e.g., Kriegeskorte et al., 2008b; Carlson et al., 2013; Cichy et al., 2014; Kaneshiro et al., 2015; Grootswagers et al., 2018), and is less prominent in studies that used stimulus sets that controlled for systematic visual differences (e.g., Rice et al., 2014; Bracci and Op de Beeck, 2016; Proklova et al., 2016, 2019; Long et al., 2018; Grootswagers et al., 2019).

These limitations would not constitute a huge problem on their own. They could be addressed in follow-up work that replicates results using a different stimulus set, or a set that specifically controls for the issues above. Indeed, several studies have used variations or entirely different sets to highlight contrasting and complementary findings (For a recent review, see Wardle and Baker, 2020). However, the problem arises because a large amount of published work has used the exact same stimulus set. This leads to a wide-spread issue of generalisability. In addition, the current academic landscape encourages only publishing positive results (Ioannidis et al., 2014), which could mean that we are getting a skewed picture of published results that are specific to the 92-object stimulus set.

## OVERFITTING TO ONE NEUROIMAGING DATASET

Overusing the same stimuli is certainly an issue, but it is arguably more worrying that so many studies also use the exact same data (**Figure 2**). The MEG and fMRI responses to the 92-object stimulus set were made publicly available (Cichy et al., 2016, 2014). This is a gold standard open science practice, and the dataset has certainly been useful for the field: since its release, a large proportion of experimental papers have used this exact dataset to develop new analyses or models, decide on optimal analysis pipelines, or assess the similarity of the data to other modalities. An unavoidable result, however, is that these new and interesting developments are possibly specific to one particular dataset (note, some have shown that their conclusions were supported by multiple datasets). Looking at the same dataset from many different angles over and over again will lead to several findings that are dataset specific. In other words, there is a risk of overfitting to one set of data. Eventually, this will leave us with a body of literature that does not replicate or generalise to new data, which is a waste of time, money, and other resources.

We strongly point out that this commentary is not an argument against making data public. On the contrary, if data sharing practices were more common, we may not have had this problem. The current issue lies in the fact that data reuse is very common, but there are very few similar open data sets in this literature. They may exist but are either hard to find, or in a difficult-to-use format. Open data is incredibly useful to validate existing analyses and test new ideas and methods without using immense resources to collect expensive neuroimaging data (resources that do not exist for many researchers). A good balance needs to be struck between collecting new data and reusing existing open data. If resources allow, existing open data could serve as pilot data for new neuroimaging experiments. The resulting output would contribute to greater experimental diversity, yield a new open dataset that can be used in the future and ultimately allow for better generalisability of conclusions. We also point out that reusing a dataset for the sake of benchmarking models is a different endeavour from the uses of the problematic dataset we highlight here. For comparing model performance, it is indeed important to use the same dataset as others in the literature. For example, the *imagenet* and *mnist* benchmarks used in computer vision research. Previous work has used the 92-object dataset to demonstrate promising new analyses, so it could be considered a benchmark dataset. However, it is not well-suited for this purpose, as new analyses might appear too good to be true considering the limitations of the stimuli. Therefore, even for benchmarking purposes we urge to not rely (solely) on the 92-object set.

## THE WAY FORWARD

A large number of papers in Computational Cognitive Neuroscience have used the same dataset and stimulus set, which raises questions about the generalisability of their influential and exciting results. This problem is ongoing, with many of the papers in **Figure 2** published in the last 5 years, and several forthcoming works currently on preprint servers. Yet, there are promising signs on the horizon.

First, not all work in the field has relied on this problematic dataset. Many studies have collected new stimuli and data, re-used different stimuli and datasets, or generalised their results to multiple datasets. Second, efforts to develop large-scale, systematically selected stimulus databases are a huge step forward, such as THINGS (Hebart et al., 2019), or *ecoset* (Mehrer et al., 2021). These large sets will hopefully will soon be accompanied by high-quality open neuroimaging datasets. Third, data sharing has also become easier through several (free) hosting platforms (e.g., *figshare*, *osf*, *openneuro*), and it is increasingly

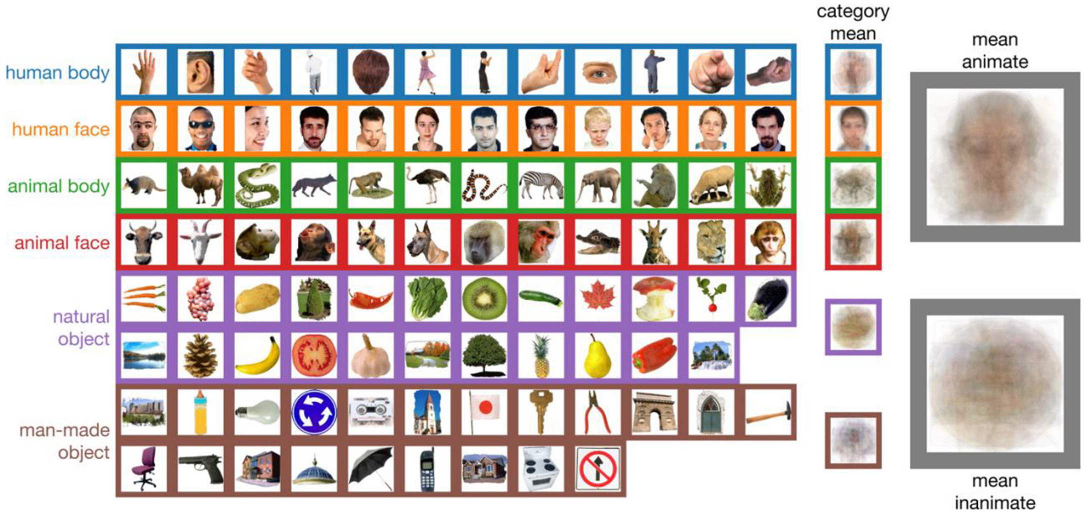

**FIGURE 1 |** The 92-object stimulus set used in over 35 computational cognitive neuroscience papers. Some of the objects in the "animate" category are not animate (e.g., hair; row 1 column 5, or a cut-out wolf figurine; row 3 column 4). Some inanimate images are not objects, but rather scenes (e.g., row 6/7 column 1). This stimulus set is often used to highlight a strong animate/inanimate dichotomy in human brain responses, but the categories have consistent visual differences (the rightmost two columns show the means of all images in a category).

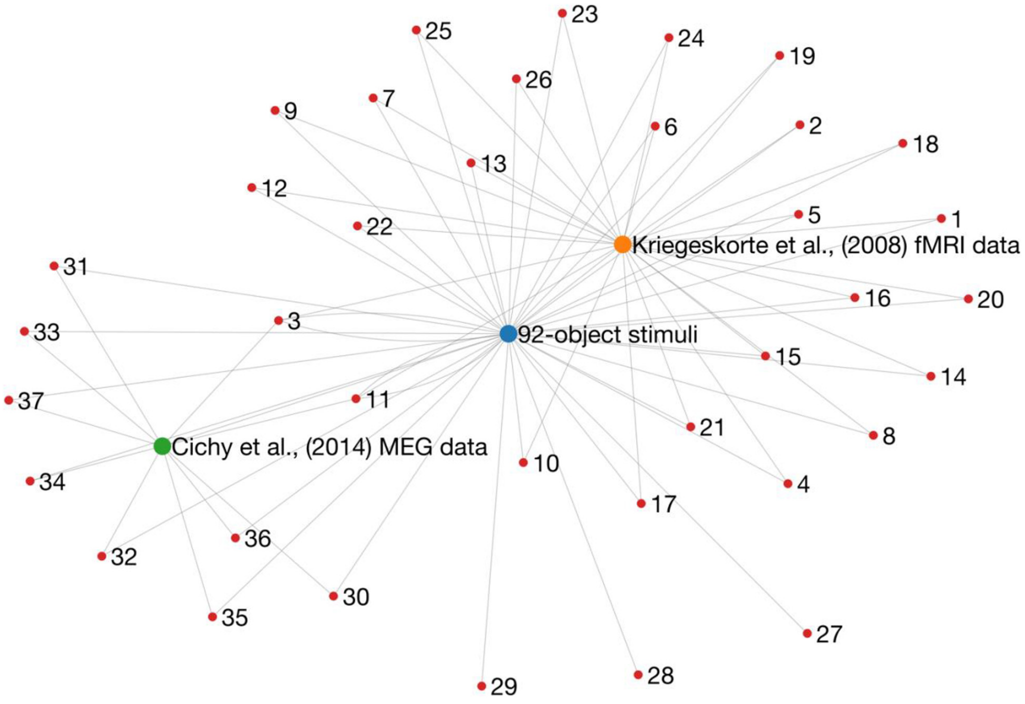

**FIGURE 2 |** A graph representation of research outputs (nodes) that re-used (edges) the same dataset or stimuli from two influential papers (larger nodes). Importantly, this is not intended to question or refute the findings of any of these studies, but rather to point to a potential issue of generalisability in the literature. These data were obtained by going through *Scopus* citation lists, and therefore it is likely that the graph is not complete.

more common to make data available upon publication. Finally, data formatting standards have been established, such as the brain imaging data structure (BIDS) (Gorgolewski et al., 2016; Niso et al., 2018; Holdgraf et al., 2019; Pernet et al., 2019), which makes it easier to re-use data.

In conclusion, we need to strike a delicate balance between taking advantage of existing resources and being aware of the limitations that come with re-using existing data and stimulus sets. While open data will allow the field to keep exploring new ideas without spending huge amounts of public funds or devoting many hours to operating neuroimaging equipment, we equally often need to consider collecting new data to test the reliability of these ideas and improve the body of research as a whole.

## DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/supplementary material, further inquiries can be directed to the corresponding author/s.

## AUTHOR CONTRIBUTIONS

TG created the figures and first draft. TG and AR wrote the manuscript. Both authors contributed to the article and approved the submitted version.

## FUNDING

AR was supported by ARC DE200101159.

## REFERENCES

Bracci, S., and Op de Beeck, H. P. (2016). Dissociations and associations between shape and category representations in the two visual pathways. *J. Neurosci.* 36, 432–444. doi: 10.1523/JNEUROSCI.2314-15.2016

Carlson, T. A., Tovar, D. A., Alink, A., and Kriegeskorte, N. (2013). Representational dynamics of object vision: the first 1000 ms. *J. Vis.* 13:1. doi: 10.1167/13.10.1

Cichy, R. M., Pantazis, D., and Oliva, A. (2014). Resolving human object recognition in space and time. *Nat. Neurosci.* 17, 455–462. doi: 10.1038/nn.3635

Cichy, R. M., Pantazis, D., and Oliva, A. (2016). Similarity-Based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cereb. Cortex* 26, 3563–3579. doi: 10.1093/cercor/bhw135

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., et al. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci. Data* 3:160044. doi: 10.1038/sdata.2016.44

Grootswagers, T., Cichy, R. M., and Carlson, T. A. (2018). Finding decodable information that can be read out in behaviour. *NeuroImage* 179, 252–262. doi: 10.1016/j.neuroimage.2018.06.022

Grootswagers, T., Robinson, A. K., Shatek, S. M., and Carlson, T. A. (2019). Untangling featural and conceptual object representations. *NeuroImage* 202:116083. doi: 10.1016/j.neuroimage.2019.116083

Hebart, M. N., Dickter, A. H., Kidder, A., Kwok, W. Y., Corriveau, A., Wicklin, C. V., et al. (2019). THINGS: a database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS One* 14:e0223792. doi: 10.1371/journal.pone.0223792

Holdgraf, C., Appelhoff, S., Bickel, S., Bouchard, K., D'Ambrosio, S., David, O., et al. (2019). iEEG-BIDS, extending the Brain Imaging Data Structure specification to human intracranial electrophysiology. *Sci. Data* 6:102. doi: 10.1038/s41597-019-0105-7

Ioannidis, J. P. A., Munafò, M. R., Fusar-Poli, P., Nosek, B. A., and David, S. P. (2014). Publication and other reporting biases in cognitive sciences: detection, prevalence, and prevention. *Trends Cogn. Sci.* 18, 235–241. doi: 10.1016/j.tics.2014.02.010

Kaneshiro, B., Guimaraes, M. P., Kim, H.-S., Norcia, A. M., and Suppes, P. (2015). A representational similarity analysis of the dynamics of object processing using single-Trial EEG classification. *PLoS One* 10:e0135697. doi: 10.1371/journal.pone.0135697

Kiani, R., Esteky, H., Mirpour, K., and Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *J. Neurophysiol.* 97, 4296–4309. doi: 10.1152/jn.00024.2007

Kriegeskorte, N., and Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn. Sci.* 17, 401–412. doi: 10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., and Bandettini, P. A. (2008a). Representational similarity analysis - connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* 2:4. doi: 10.3389/neuro.06.004.2008

Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., et al. (2008b). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* 60, 1126–1141. doi: 10.1016/j.neuron.2008.10.043

Long, B., Yu, C.-P., and Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proc. Natl. Acad. Sci. U.S.A.* 115, E9015–E9024. doi: 10.1073/pnas.1719616115

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., and Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proc. Natl. Acad. Sci. U.S.A.* 118:e2011417118. doi: 10.1073/pnas.2011417118

Niso, G., Gorgolewski, K. J., Bock, E., Brooks, T. L., Flandin, G., Gramfort, A., et al. (2018). MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. *Sci. Data* 5:180110. doi: 10.1038/sdata.2018.110

Pernet, C. R., Appelhoff, S., Gorgolewski, K. J., Flandin, G., Phillips, C., Delorme, A., et al. (2019). EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Sci. Data* 6:103. doi: 10.1038/s41597-019-0104-8

Proklova, D., Kaiser, D., and Peelen, M. V. (2016). Disentangling representations of object shape and object category in human visual cortex: the animate–inanimate distinction. *J. Cogn. Neurosci.* 28, 680–692. doi: 10.1162/jocn_a_00924

Proklova, D., Kaiser, D., and Peelen, M. V. (2019). MEG sensor patterns reflect perceptual but not categorical similarity of animate and inanimate objects. *NeuroImage* 193, 167–177. doi: 10.1016/j.neuroimage.2019.03.028

Rice, G. E., Watson, D. M., Hartley, T., and Andrews, T. J. (2014). Low-Level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. *J. Neurosci.* 34, 8837–8844. doi: 10.1523/JNEUROSCI.5265-13.2014

Vanrullen, R. (2011). Four common conceptual fallacies in mapping the time course of recognition. *Percept. Sci.* 2:365. doi: 10.3389/fpsyg.2011.00365

Wardle, S. G., and Baker, C. (2020). Recent advances in understanding object recognition in the human brain: deep neural networks, temporal dynamics, and context. *F1000Res.* 9:590. doi: 10.12688/f1000research.22296.1

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.