Model-based fMRI Analysis of Memory

Bradley C. Love
University College London
The Alan Turing Institute

Abstract

Recent advances in Model-based fMRI approaches enable researchers to investigate hypotheses about the time course and latent structure in data that were previously inaccessible. Cognitive models, especially when validated on multiple datasets, allow for additional constraints to be marshalled when interpreting neuroimaging data. Models can be related to BOLD response in a variety of ways, such as constraining the cognitive model by neural data, interpreting the neural data in light of behavioural fit, or simultaneously accounting for both neural and behavioural data. Using cognitive models as a lens on fMRI data is complementary to popular multivariate decoding and representational similarity analysis approaches. Indeed, these approaches can realise greater theoretical significance when situated within a model-based approach.

Highlights

Cognitive models can formalise theories to make assumptions and predictions clearer.

Cognitive models offer additional constraints when interpreting the BOLD response.

Models can go from behaviour to BOLD, vice versa, or address both simultaneously.

Model-based approaches can incorporate pattern similarity and decoding approaches.

Correspondence: b.love@ucl.ac.uk
Word Count: 2365

Introduction

Memory by definition involves processes that extend over time and involve generalisation or similarity structure. Formal models offer a way to characterise these processes and better understand their brain basis. As I will review, there are a number of cases in which fMRI researchers could not have made an advance without a model-based analysis approach.

Models can play a number of constructive roles in psychology, neuroscience and science more broadly. One function is simply organising one's ideas and making assumptions clear. Each step needs to be detailed, which can reduce wiggle room relative to purely verbal theories. Whatever wiggle room is left (e.g., tuneable parameters) is made explicit.

As a consequence, what is predicted under different circumstances is made clear. Rather than debate what a theory predicts, a model can be simulated. For example, early work showing an advantage in processing category prototypes led researchers to believe that abstract prototypes were stored in memory, but subsequent work demonstrated such effects were compatible with exemplar models that store no abstractions in memory [1]. More recently, models have played a related role in the design and interpretation of fMRI studies of memory [2,3]. Models can play a constructive role in directing empirical investigations.

Science often progresses by evaluating competing theoretical accounts. Models afford the possibility of model comparison in which competing accounts can be pitted against one another and the model that performs best can be favoured. For example, Mack and colleagues [4] formally evaluated whether the representations in an exemplar or prototype model best matched the BOLD response and found the exemplar model was more consistent (also see [5]). Recent work evaluating whether the hippocampus learns to associate objects and words incrementally or in an all-or-none fashion used a related approach that favoured the all-or-none account [6].

Models can serve a powerful integrative role by linking seemingly disparate findings through common computational mechanisms. For example, a simple model of familiarity and recognition memory captured findings from both fMRI studies of visual categorisation and word list memory [7]. In my own work, the same clustering approach to understanding human learning has been applied to a number of fMRI studies [8–12], which helps to theoretically link studies. Recent work [13] has extended these same model mechanisms to offer an alternative explanation for place and grid cell responses in rodents and humans. This account makes novel predictions for how cell responses should change under different experimental conditions.

The aforementioned models can be considered cognitive models. These models are hypothesised to involve the same processes and representations as the human mind. Cognitive models reside at Marr's [14] algorithmic level and are well-placed to help explain how the brain implements higher-level computations [15]. Below, I will discuss the various ways that cognitive models can be related to fMRI data.

In addition to using cognitive models, neuroscientists also use formal models as purely data analysis tools. For example, the Generalised Linear Model (GLM) itself is a formal model that has assumptions and tuneable parameters that are fit to data. Of course, the GLM is not a model of how people process and represent information. Other examples of data analysis tools include Dynamic Causal Modelling [16], techniques to measure the intrinsic or functional dimensionality of fMRI data [17], and Multi-Voxel Pattern Analysis (MVPA).

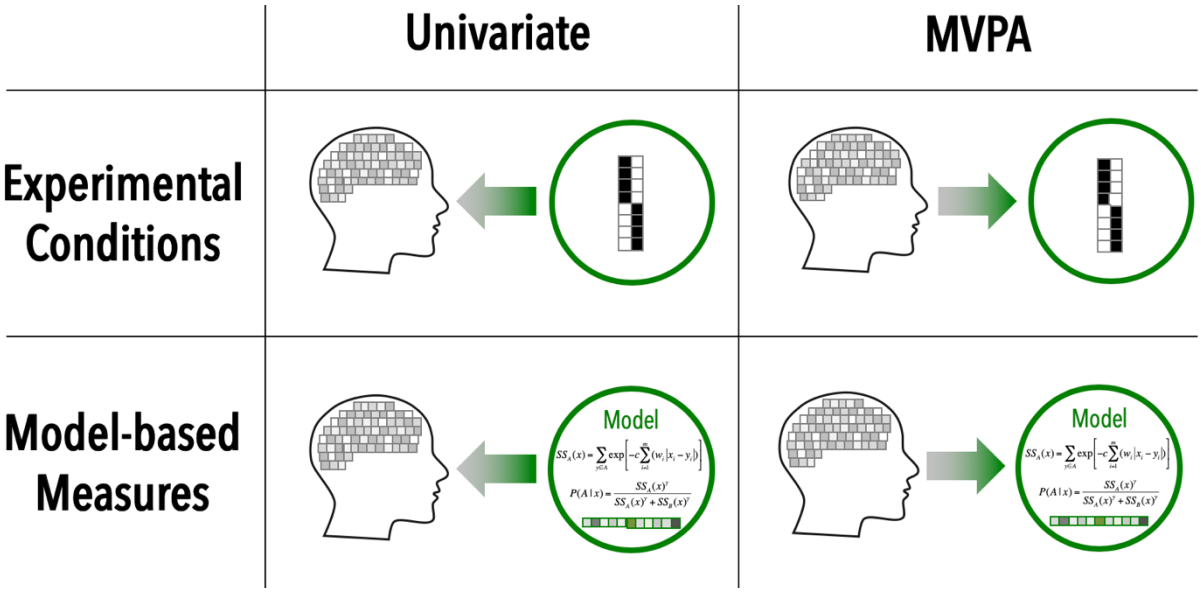

Figure 1: The top row illustrates approaches that are not model-based in that they don't leverage a cognitive model of the task. For example, in the top-left panel, a standard analysis might identify voxels that are more active for faces than for house stimuli, whereas in the top-right panel a decoder might try to classify whether the participant is viewing a house or a face stimulus on each trial. In the bottom row, a cognitive model is at the centre of the analysis. In the bottom-left panel, some measure from the cognitive model (which is usually fit to behavioural data), such as item familiarity, learning update, etc., is entered into the GLM. Such an analysis will identify voxels that show a similar activation profile to the model measure. In contrast, in the bottom-right quadrant, a classifier is applied to the brain to try to decode some internal measure from the cognitive model. In this case, models are favoured to the extent that their internal state is decodable [4].

MVPA decoding approaches apply a machine classifier to "mind read" from the BOLD response whether a participant, for example, is viewing a house or a face [18]. Although these are not psychological models, they can be used to make interesting behavioural predictions. For example, participants tend to have faster response times for stimuli that are further from the classifier's decision bound, which indicates the classifier is more confident about its decision [19]. Decoding approaches can also be used to determine when people are engaging in replay [20–23]. The line between what is a cognitive model and a data analysis tool can be blurred at times.

Linking cognitive models to the BOLD response

In a typical task fMRI analysis, experimental conditions are contrasted with one another. For example, one may contrast voxels that are more active for face than for house stimuli. The simplest model-based analysis replace the stimulus condition with some model measure (e.g., prediction error) that varies across trials [24]. By entering this regressor (e.g., prediction error) from the cognitive model into the GLM, one can evaluate which voxels co-vary with the cognitive construct. As shown in Figure 1, both the typical contrast approach and simple model-based analyses are univariate. Instead, standard MVPA analyses start from a collection of voxels (multivariate) and aim to predict some experimental condition,

such as whether the participant is viewing a house or a face. One innovation is to make the target of decoding a model measure, such as item familiarly according to a cognitive model [4]. The four quadrants shown in Figure 1 are not an exhaustive taxonomy of how to relate models to the BOLD response (for a more complete treatment, see [25,26]).

Perhaps because it's relatively straightforward, the univariate model-based approach is most common in the field. Typically, a model is fit to behavioural data and then used as a lens on the fMRI data. For example, an associative learning model was fit to behavioural data from a task where people formed impressions of various social groups through trial-by-trial feedback [27]. The fitted model provided a GLM trial-by-trial measure of valence or prejudice for each group, which tracked activity in the anterior temporal lobe in the model-based analysis. Model-based analysis was critical for capturing changes in memory *across* study trials.

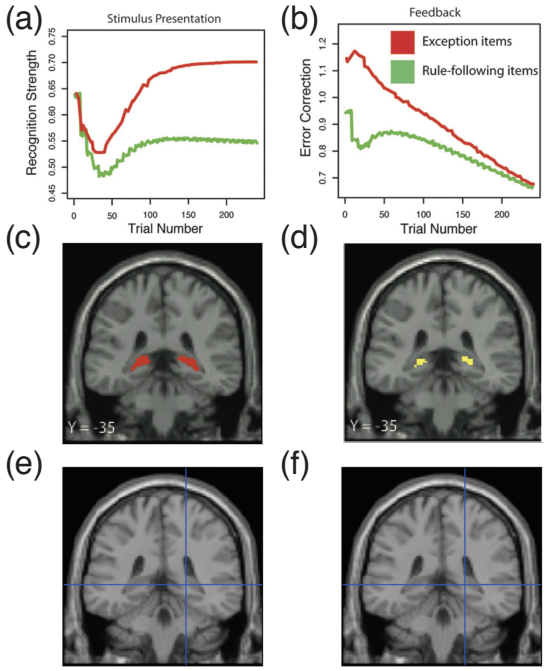

Figure 2: Panels a and b show model-based regressors for a measure of recognition strength (i.e., familiarity) and error correction (i.e., learning update). These model-based regressors track hippocampal activity at the stimulus presentation and feedback phases of trials, respectively [8]. In contrast, a standard contrast of exception>rule-following items (panels e and f) results in no statistically significant voxels, because this contrast does not track the time course of hippocampal activity.

In a category learning study [8], a model-based analysis with a clustering model of learning was critical to capturing two time courses, one across trials and one within. This study examined the hippocampus's role in acquiring categories in which most items followed a rule but some items (exceptions) did not. A clustering model [28] was fit to the behavioural data (i.e., the learning curves) and two model-based measures were entered into the GLM, one for recognition strength or familiarity and one for error correction or learning update. As shown in Figure 2, the hippocampus tracked the model's recognition measure at stimulus presentation and the error measure at feedback presentation. Interestingly, a standard analysis contrasting exception and rule-following items found no significant difference -- the cognitive model proved critical to capturing how hippocampal response changes over the course of study trials.

The same modelling approach can also be used to localise two simultaneous processes (by using two different model-based measures) within the same phase of a trial to draw distinctions between the function of anterior and posterior hippocampus [9]. Another way to scale up this basic univariate modelling approach is to adopt an encoder approach in which the fitted cognitive model provides a number of model-based regressors to enter into the GLM with the goal of explaining the most variance possible within brain regions of interest [29]. In the encoding approach, rather than trying to identify voxels that significantly regress on some specific model-based measure (e.g., prediction error), the goal is for multiple model measures to capture the most overall variance possible in the GLM.

Other model-based work [30,31] reverses the flow of information to incorporate brain measures directly into the operation of the model to better predict behaviour. For example, Kragel and colleagues [30] used a variant of the Context Maintenance and Retrieval (CMR) model of free recall [32] that took signals from the Medial Temporal Lobe (MTL) to determine whether contextual reactivation was successful at each potential recall event. The model that incorporated the BOLD input performed better than a baseline model in predicting behaviour.

Rather than link from model to brain or brain to model, joint modelling approaches [33,34] simultaneously model the mutual constraints between behavioural and brain measures through an intermediary cognitive model. This approach can deal with multiple brain measures (e.g., fMRI and EEG) and can make predictions about missing measures based on covariance with the observed measures.

There are number of other creative ways to link cognitive models to BOLD response. One way is to link a key event, as indexed by the cognitive model, to an operation in the brain. For example, a recent study finds that prediction errors during study are predictive of later replay events [20]. In other work, a Bayesian model determined the probability that an item would be remembered, which correlated with hippocampal activity during encoding [35].

Finally, a cognitive model's fitted parameters can be related to the BOLD response instead of a trial-by-trial measure from the model. During category learning, models [28,36] predict that goal-relevant aspects of the stimuli will receive greater weight or attention. A recent study found that the learned attentional weights from category learning models fit to behaviour were predictive of how well those stimulus aspects could be decoded from the BOLD response [37]. Related, in a study exploring vmPFC-hippocampal interactions during concept learning [12], the pattern of goal-directed representation compression in vmPFC paralleled the attention weights from a model fitted to behaviour.

Models can uncover useful latent states

Models can be useful in inferring latent states that can help explain behaviour and its brain basis. One example of a latent variable are the clusters in the aforementioned learning models [28,38] which detail how related items are stored together in memory. Models operationalise these hypothesised representational structures, which can be useful in analysing BOLD response.

Inferring latent state is more complex when researchers aim to characterise complex mental operations that unfold through time [39]. One popular approach is to use hidden Markov models (HMMs) to infer what operations people are currently undertaking and using this characterisation to interpret the BOLD response [40,41].

The importance of inferring latent state is also becoming appreciated in related fields, such as reinforcement learning [42]. Many of the same conceptual issues and brain systems are implicated in these tasks as in goal-directed concept learning. For example, strategic exploration relies on hippocampal-prefrontal cooperation [43] as is found during memory tasks [12].

Comparing model and brain representations

In addition to MVPA decoding, multivariate pattern analysis can be used to compare proposed (e.g., model) representations and voxel representations [44]. This pattern comparison analysis is popularly known as Representational Similarity Analysis (RSA, [45]). RSA correlates two similarity matrices, one from the cognitive model and one from the brain, to assess how well the two similarity spaces align. RSA can be used as confirmatory evidence that a model provides the correct representational account of a brain region or in an exploratory fashion such as in a whole-brain searchlight analysis. One application of RSA is to compare proposed memory representations acquired by models of concept learning to brain regions thought to implement those functions [4,46]. For example, RSA analyses found that hippocampal representations of objects are modulated by changes in the task goal [10].

For an RSA to be model-based, one of the similarity matrices should be generated by a cognitive model. RSA can involve the evaluation of several cognitive models. A variety of models can be considered and the model whose representations best align with the brain can be favoured [46]. However, not all RSAs are model-based and the dividing line can be blurry. For example, technically, finding that hippocampus CA1 codes distance to a goal [47] is not model-based (because distance is specified by the task), whereas coding distance to some model quantity, such as distance to a category prototype [48], is model-based (because the prototype is specified by the fitted cognitive model). For a model-based analysis to be useful, it should add something beyond a standard analysis. Ideally, a model-based analysis would improve both data fit and our understanding of the domain. For example, a model may largely code distance to goal, but diverge in informative ways under certain circumstances that could be empirically verified and in turn deepen our understanding of the domain.

Certainly, univariate analyses can be rigorous, interesting, and motivated but not model-based. The same is true in RSA. For example, a recent study [49] used similarity matrices designed to capture perceptual or conceptual similarity to hone in on the function of perirhinal cortex and other regions. This work is exciting and valuable, but because the similarity matrices were derived from human ratings rather than generated by a model of perceptual or conceptual processing, the analysis is not model-based.

Conclusions

Adopting a model-based approach to fMRI analysis can offer a number of advantages. In some cases, one can evaluate hypotheses that otherwise would not be possible with a standard analysis approach. Models, which are formalised as theories, offer the hope that results will be theoretically grounded. As related models are applied across data sets, models may promote a more systematic and cohesive science. Cognitive models are well positioned to integrate findings across levels of analysis [15].

I have reviewed a number of ways to relate cognitive models to the BOLD response. Possibilities include fitting models to be behaviour and incorporating derived trial-by-trial measures into the GLM, model decoding approaches [4], using BOLD response to drive the behavioural predictions of the model, joint modelling to simultaneously address brain and behavioural measures, and RSA comparisons of model representations and BOLD response. Which approach is suitable is largely a function of the study's design and the researcher's aims.

One key question to consider is why do model-based analyses work? Models are not magical, nor guaranteed to be helpful, so why are there so many cases in which model-based analyses succeed in pulling more from the data than would be possible through a standard analysis? The answer is that models have the ability to incorporate constraints that are outside the immediate study. In my own work, models are developed over years and honed while being applied to multiple behavioural and fMRI datasets. In this sense, the models have a reality and value outside their immediate application, which is critical because a model-based analyses is only as credible as the model used.

Conflict of interest

Nothing declared.

Acknowledgements

This work was supported by NIH Grant 1P01HD080679, Wellcome Trust Investigator Award WT106931MA, and Royal Society Wolfson Fellowship 183029 to B.C.L. Thanks to Franziska Broeker, Brett Roads, Maarten Speekenbrink for helpful comments.

Annotated References

*Berens et al. (2018) The authors compare two competing models of hippocampal-mediated learning and find evidence in favour of an all-or-none account.

*Mack et al. (2019) The authors explore vmPFC-hippocampal interactions and find that vmPFC establishes a goal-oriented compression code to support hippocampal learning.

*Momennejad et al. (2019) The authors show that replay events are predicted by error during previous study.

*Turner et al. (2019b) The authors review joint modelling approaches in which a cognitive model simultaneously addresses behavioural and neural measures.

*Braunlich & Love (2019) Individual differences in attention weights from learning models fitted to behavioural data predict how well stimulus information can be decoded from the brain.

References

1. Medin DL, Schaffer MM: **Context theory of classification learning**. *Psychol Rev* 1978, **85**:207–238.

2. Caplan JB, Madan CR: **Word Imageability Enhances Association-memory by Increasing Hippocampal Engagement**. *J Cogn Neurosci* 2016, **28**:1522–1538.

3. Nosofsky RM, Little DR, James TW: **Activation in the neural network responsible for categorization and recognition reflects parameter changes**. *Proc Natl Acad Sci U S A* 2012, **109**:333–338.

4. Mack ML, Preston AR, Love BC: **Decoding the Brain's Algorithm for Categorization from its Neural Implementation**. *Curr Biol* 2013, **23**:2023–2027.

5. Stillesjö S, Nyberg L, Wirebring LK: **Building Memory Representations for Exemplar-Based Judgment: A Role for Ventral Precuneus**. *Front Hum Neurosci* 2019, **13**:228.

6. Berens SC, Horst JS, Bird CM: **Cross-Situational Learning Is Supported by Propose-but-Verify Hypothesis Testing**. *Curr Biol* 2018, **28**:1132-1136.e5.

7. Davis T, Xue G, Love BC, Preston AR, Poldrack RA: **Global neural pattern similarity as a common basis for categorization and recognition memory**. *J Neurosci* 2014, **34**:7472–84.

8. Davis T, Love BC, Preston AR: **Learning the exception to the rule: model-based FMRI reveals specialized representations for surprising category members**. *Cereb Cortex* 2012, **22**:260–73.

9. Davis T, Love BC, Preston AR: **Striatal and hippocampal entropy and recognition signals in category learning: simultaneous processes revealed by model-based fMRI**. *J Exp Psychol Learn Mem Cogn* 2012, **38**:821–39.

10. Mack ML, Love BC, Preston AR: **Dynamic updating of hippocampal object representations reflects new conceptual knowledge**. *Proc Natl Acad Sci* 2016, **113**:13203–13208.

11. Inhoff MC, Libby LA, Noguchi T, Love BC, Ranganath C: **Dynamic integration of conceptual information during learning**. *PLOS ONE* 2018, **13**:e0207357.

12. Mack ML, Preston AR, Love BC: *Ventromedial prefrontal cortex compression during concept learning*. Nature Communications (in press); 2019.

13. Mok RM, Love BC: **A non-spatial account of place and grid cells based on clustering models of concept learning**. *Nat Commun* 2019, **10**:5685.

14. Marr D: *Vision*. W. H. Freeman; 1982.

15. Love BC: **The Algorithmic Level Is the Bridge Between Computation and Brain**. *Top Cogn Sci* 2015, **7**:230–242.

16. Friston KJ, Harrison L, Penny W: **Dynamic causal modelling**. *NeuroImage* 2003, **19**:1273–1302.

17. Ahlheim C, Love BC: **Estimating the functional dimensionality of neural representations**. *NeuroImage* 2018, **179**:51–62.

18. Cetron JS, Connolly AC, Diamond SG, May VV, Haxby JV, Kraemer DJM: **Decoding individual differences in STEM learning from functional MRI data**. *Nat Commun* 2019, **10**:2027.

19. Ritchie JB, de Beeck HO: **Using neural distance to predict reaction time for categorizing the animacy, shape, and abstract properties of objects**. *Sci Rep* 2019, **9**:13201.

20. Momennejad I, Otto AR, Daw ND, Norman KA: **Offline replay supports planning in human reinforcement learning**. *eLife* 2018, **7**:e32548.

21. Xue G: **The Neural Representations Underlying Human Episodic Memory**. *Trends Cogn Sci* 2018, **22**:544–561.

22. Shanahan LK, Gjorgieva E, Paller KA, Kahnt T, Gottfried JA: **Odor-evoked category reactivation in human ventromedial prefrontal cortex during sleep promotes memory consolidation**. *eLife* 2018, **7**:e39681.

23. Lee S-H, Kravitz DJ, Baker CI: **Differential Representations of Perceived and Retrieved Visual Information in Hippocampus and Cortex**. *Cereb Cortex* 2019, **29**:4452–4461.

24. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ: **Cortical substrates for exploratory decisions in humans**. *Nature* 2006, **441**:876–9.

25. Turner BM, Forstmann BU, Love BC, J.Palmeri T, Maanen LV: **Approaches to Analysis in Model-based Cognitive Neuroscience**. *J Math Psychol* in press,

26. Pratte MS, Tong F: **Integrating theoretical models with functional neuroimaging**. *J Math Psychol* 2017, **76**:80–93.

27. Spiers HJ, Love BC, Le Pelley ME, Gibb CE, Murphy RA: **Anterior Temporal Lobe Tracks the Formation of Prejudice**. *J Cogn Neurosci* 2017, **29**:530–544.

28. Love BC, Medin DL, Gureckis T: **SUSTAIN: A Network Model of Human Category Learning**. *Psychol Rev* 2004, **111**:309–332.

29. van Gerven MAJ: **A primer on encoding models in sensory neuroscience**. *J Math Psychol* 2017, **76**:172–183.

30. Kragel JE, Morton NW, Polyn SM: **Neural Activity in the Medial Temporal Lobe Reveals the Fidelity of Mental Time Travel**. *J Neurosci* 2015, **35**:2914–2926.

31. Palmeri TJ, Schall JD, Logan GD: *Neurocognitive Modeling of Perceptual Decision Making*. Oxford University Press; 2015.

32. Polyn SM, Norman KA, Kahana MJ: **A context maintenance and retrieval model of organizational processes in free recall.** *Psychol Rev* 2009, **116**:129–156.

33. Turner BM, Forstmann BU, Steyvers M: *Joint Models of Neural and Behavioral Data*. Springer International Publishing; 2019.

34. Turner BM, Palestro JJ, Miletić S, Forstmann BU: **Advances in techniques for imposing reciprocity in brain-behavior relations**. *Neurosci Biobehav Rev* 2019, **102**:327–336.

35. Gluth S, Sommer T, Rieskamp J, Büchel C: **Effective Connectivity between Hippocampus and Ventromedial Prefrontal Cortex Controls Preferential Choices from Memory**. *Neuron* 2015, **86**:1078–1090.

36. Nosofsky RM: **Attention, similairty, and the identification-categorization relationship**. *J Exp Psychol Gen* 1986, **115**:39–57.

37. Braunlich K, Love BC: **Occipitotemporal representations reflect individual differences in conceptual knowledge.** *J Exp Psychol Gen* 2019, **148**:1192–1203.

38. Anderson JR: **The adaptive nature of human categorization**. *Psychol Rev* 1991, **98**:409–429.

39. Wijeakumar S, Ambrose JP, Spencer JP, Curtu R: **Model-based functional neuroimaging using dynamic neural fields: An integrative cognitive neuroscience approach**. *J Math Psychol* 2017, **76**:212–235.

40. Tubridy S, Halpern D, Davachi L, Gureckis TM: *A neurocognitive model for predicting the fate of individual memories*. PsyArXiv; 2018.

41. Anderson JR, Borst JP, Fincham JM, Ghuman AS, Tenison C, Zhang Q: **The Common Time Course of Memory Processes Revealed**. *Psychol Sci* 2018, **29**:1463–1474.

42. Niv Y: **Learning task-state representations**. *Nat Neurosci* 2019, **22**:1544–1553.

43. Wang JX, Voss JL: **Brain Networks for Exploration Decisions Utilizing Distinct Modeled Information Types during Contextual Learning**. *Neuron* 2014, **82**:1171–1182.

44. Haxby JV: **Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex**. *Science* 2001, **293**:2425–2430.

45. Dimsdale-Zucker HR, Ranganath C: **Representational Similarity Analyses**. In *Handbook of Behavioral Neuroscience*. . Elsevier; 2018:509–525.

46. Ritchie JB, Op de Beeck H: **A Varying Role for Abstraction in Models of Category Learning Constructed from Neural Representations in Early Visual Cortex**. *J Cogn Neurosci* 2019, **31**:155–173.

47. Spiers HJ, Olafsdottir HF, Lever C: **Hippocampal CA1 activity correlated with the distance to the goal and navigation performance**. *Hippocampus* 2018, **28**:644–658.

48. Seger CA, Braunlich K, Wehe HS, Liu Z: **Generalization in Category Learning: The Roles of Representational and Decisional Uncertainty**. *J Neurosci* 2015, **35**:8802–8812.

49. Martin CB, Douglas D, Newsome RN, Man LL, Barense MD: **Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream**. *eLife* 2018, **7**:e31873.