

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35

The temporal evolution of conceptual object representations revealed through models of behavior, semantics and deep neural networks

Bankson, B. B.^{1*}, Hebart, M. N.^{1*}, Groen, I. I. A.¹, & Baker, C. I.¹

¹Section on Learning and Plasticity, Laboratory of Brain and Cognition, National Institute of Mental Health, National Institutes of Health, Bethesda, MD 20892, USA.

* equal contribution

Correspondence should be addressed to:

Brett B. Bankson
Laboratory of Cognitive Neurodynamics
UPMC Presbyterian
Suite B-400
200 Lothrop Street
Pittsburgh, PA 15213
bbb17@pitt.edu

Conflict of Interest: The authors declare no competing financial interests.

Acknowledgements: This work was supported by the Intramural Research Program of the National Institute of Mental Health (ZIA-MH-002909) - National Institute of Mental Health Clinical Study Protocol 93-M-0170, NCT00001360, a Feodor-Lynen fellowship of the Humboldt Foundation to M.N.H., and a Rubicon Fellowship from the Netherlands Organization for Scientific Research to I.I.A.G.

36 **Abstract**

37

38 Visual object representations are commonly thought to emerge rapidly, yet it has remained
39 unclear to what extent early brain responses reflect purely low-level visual features of these
40 objects and how strongly those features contribute to later categorical or conceptual
41 representations. Here, we aimed to estimate a lower temporal bound for the emergence of
42 conceptual representations by defining two criteria that characterize such representations: 1)
43 conceptual object representations should generalize across different exemplars of the same
44 object, and 2) these representations should reflect high-level behavioral judgments. To test these
45 criteria, we compared magnetoencephalography (MEG) recordings between two groups of
46 participants ($n = 16$ per group) exposed to different exemplar images of the same object
47 concepts. Further, we disentangled low-level from high-level MEG responses by estimating the
48 unique and shared contribution of models of behavioral judgments, semantics, and different
49 layers of deep neural networks of visual object processing. We find that 1) both generalization
50 across exemplars as well as generalization of object-related signals across time increase after 150
51 ms, peaking around 230 ms; 2) behavioral judgments explain the most unique variance in the
52 response after 150 ms. Collectively, these results suggest a lower bound for the emergence of
53 conceptual object representations around 150 ms following stimulus onset.

54

55 **Introduction**

56

57 There is enormous variability in the visual appearance of objects, yet we can rapidly recognize
58 them without effort, even under difficult viewing conditions (DiCarlo & Cox, 2007; Potter et al.,
59 2013). Evidence from neurophysiological studies in human suggests the emergence of visual
60 object representations within the first 150 ms of visual processing (Thorpe et al., 1996; Carlson
61 et al., 2013, Cichy et al., 2014). For example, the specific identity of objects can be decoded from
62 the magnetoencephalography (MEG) signal with high accuracy around 100 ms (Cichy et al., 2014).
63 However, knowing when discriminative information about visual objects is available does not
64 inform us about the nature of those representations, in particular whether they primarily reflect
65 (low-level) visual features or (high-level) conceptual aspects of the objects (Clarke et al., 2015).
66 To address this issue, in this study we employed multivariate MEG decoding and model-based
67 representational similarity analysis (RSA) to elucidate the nature of object representations over
68 time.

69 Previous studies have demonstrated increasing category specificity (van de
70 Nieuwenhuijzen et al., 2013; Cichy et al., 2014), tolerance for position and size (Isik et al., 2014)
71 and semantic information (Clarke et al., 2013) over the first 200ms following stimulus onset,
72 suggesting some degree of abstraction from low-level visual features. However, identifying the
73 nature of object representations is an inherently difficult problem: low-level features may be
74 predictive of object identity, making it hard to disentangle the relative contribution of low and
75 high-level properties to measured brain signals (Groen et al., 2017). In this study, we addressed
76 this problem by combining tests for the generalization of object representations with methods
77 to separate the independent contributions of low- and high-level properties. We focused on two
78 specific criteria that would need to be fulfilled for a representation to be considered conceptual.

79 First, a conceptual representation should generalize beyond the specific exemplar presented, not
80 just variations of the same exemplar. Second, a conceptual representation should also reflect
81 high-level behavioral judgments about objects (Clarke & Tyler, 2015; Wardle et al., 2016). We
82 consider fulfillment of these two properties to provide a lower bound at which a representation
83 could be considered conceptual.

84 We collected MEG and behavioral data from 32 participants allowing us to probe the
85 temporal dynamics of conceptual object representations according to the two criteria above. To
86 test for generalization across specific exemplars, we assessed the reliability of object
87 representations across two independent sets of objects. Further, we assessed the relation of
88 those object representations to behavior by comparing participants' behavioral judgments with
89 the MEG response patterns using RSA. Importantly, to isolate the relative contributions of low-
90 level and conceptual properties to those MEG responses, we identified the variance uniquely
91 explained by behavioral judgments, isolating low-level representations using early layers of a
92 deep neural network, which have been shown to capture low- to mid-level responses in fMRI and
93 monkey ventral visual cortex (Cadieu et al., 2014; Cichy et al., 2016a; Eickenberg & Thirion, 2017;
94 Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014; Wen et al.,
95 2017). Finally, to achieve a more interpretable understanding of the contribution of behavior to
96 MEG responses, we identified the unique and shared variance explained in the MEG response by
97 behavior and two high-level conceptual models, one perceptual (upper layers in a deep neural
98 network) and one semantic (based on word co-occurrence statistics).

99

100 **Methods**

101

102 *Participants*

103 32 healthy participants (18 female, mean 25.8, range 19-47) with normal or corrected-to-normal
104 vision took part in this study. As a part of a pilot experiment used for purely illustrative purposes
105 (see Figure 4a), 8 participants (5 overlap) completed the same behavioral task with a different
106 set of stimuli. All participants gave written informed consent prior to participation in the study
107 as a part of the study protocol (93-M-0170, NCT00001360). The study was approved by the
108 Institutional Review Board of the National Institutes of Health and was conducted according to
109 the Declaration of Helsinki.

110

111 *Stimuli*

112 We created two independent sets of 84 object images each that were cropped and placed on a
113 grey background. Each stimulus set contained a unique exemplar for each of the 84 object
114 concepts, as shown in Figure 1a. We selected object concepts by using a combination of two
115 word databases, one of word frequency (Corpus of Contemporary American English, Davies,
116 2008) and the other of word concreteness (Brysbaert et al., 2014). First, based on our corpus we
117 selected the 5000 most frequent nouns in American English. From this set of words, we then
118 selected nouns with concreteness ratings $> 4/5$. Finally, for words that would be difficult or
119 impossible to distinguish when presented as an image (e.g. 'woman', 'mother', 'wife'), we used
120 only the most frequent entry. This selection left us with a set of 112 objects.

121 To evaluate whether those categories would be labeled consistently, we generated three
122 distinct images of each object concept and asked three individuals who were not involved in the
123 study to provide a verbal label for each of the three versions of the 112 objects. Images that were
124 not labeled correctly by all raters were discarded, leaving us with 84 object concepts. From the
125 three sets of object images, we then randomly sampled two per object concept. This generated
126 two sets of unique object exemplars for 84 object concepts, divided into Image Set 1 and Image
127 Set 2. The two sets of object stimuli are shown in Supplemental Figure S1.

128

129 *Procedure*

130

131 *MEG*

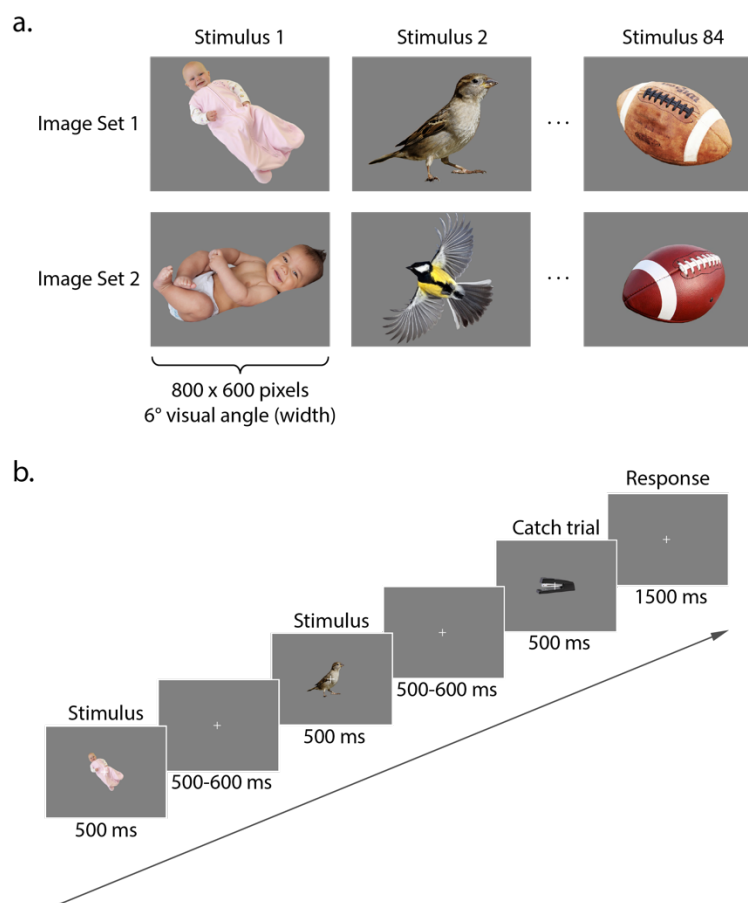
132 During MEG recordings, participants were seated upright in an electromagnetically shielded MEG
133 chamber. Stimuli were presented using the Psychophysics Toolbox (Brainard, 1997) in MATLAB
134 (version 2016a, Mathworks, Natick, MA). Visual stimulation was controlled by a Panasonic PT-
135 D3500U DLP projector with an ET-DLE400 lens, located outside of the chamber and projected
136 through a waveguide and series of mirrors onto a back-projection screen in front of the
137 participant. Participants were assigned to one of two groups and completed the experiment with
138 either Image Set 1 or Image Set 2. All stimuli were presented on a grey background with a white
139 fixation cross in the center (viewing distance: 70 cm, stimulus width: 6° of visual angle).
140 Participants completed an oddball detection task, pressing a button in response to catch trials
141 containing the oddball stimulus (desk stapler) that appeared pseudorandomly every 2-6 trials
142 (average 4, flat distribution). On each trial (Figure 1b), an object stimulus was presented at
143 fixation for 500 ms, followed by a variable fixation period (regular trials: pseudorandomly 500-
144 600 ms, catch trials: 1500 ms). In addition, participants were instructed to blink their eyes only
145 as they pressed the button of the MEG-compatible button box during catch trials, in order to
146 avoid any eye blink artifacts at other points of the experiment. Participants completed 18 runs
147 that were divided into 6 blocks of 3 runs each, with self-paced breaks between each block. Each
148 run lasted 240 s, resulting in a total experimental time of 72 min. In total, participants viewed
149 each of the 84 images 36 times over the course of the experiment.

150

151 *Behavior: Object arrangement task*

152 Within two days of completing the MEG session, participants took part in a follow-up behavioral
153 experiment to provide us with behavioral estimates of the representational similarity between
154 all possible object pairs. This was done using the object arrangement method (Goldstone 1994;
155 Kriegeskorte & Mur, 2012). In this method, participants arrange objects in a 2D “arena” based on
156 their subjective similarity, and the distance between the items is used to generate $(n \times n-1)/2$
157 pairwise distance estimates between object pairs. Participants were seated at a distance of
158 approximately 57 cm in front of a 30” monitor (resolution: 1440 × 900 pixels) and completed the
159 object arrangement task on the same 84 object images used in the MEG experiment. All items
160 were presented simultaneously but in random order and with equal distance around the circular
161 arena (image width: 1.5° of visual angle). Participants were instructed to use the computer mouse
162 and arrange the items according to their similarity at their own pace, taking ~20 minutes on
163 average to complete the task. In contrast to the original implementation of this method that used

164 additional trials with selective subsets of objects (Kriegeskorte et al., 2012), we only chose a
165 single arrangement, based on our experience with the multi-arrangement task exhibiting very
166 high correlations between results of the first and the last trial (unpublished data). We deliberately
167 did not provide participants with an explicit strategy or instructions on what object features to
168 focus, so as to not bias them to focus on any specific aspect of the stimuli. To facilitate the task,
169 when a participant clicked on a certain image around the arena, an enlarged version spanning
170 150×200 pixels ($6.75 \times 9^\circ$ of visual angle) was displayed in the top right of the computer screen.
171 After completion of the experiment, we extracted the pixel-wise distance between each pair of
172 items, yielding an 84×84 distance matrix for each participant. Note that the distance matrix
173 discards the absolute position of objects and only retains their relative location, which should
174 minimize bias related to the initial placement of objects.
175



176
177 **Figure 1.** Stimulus format and trial progression. **a.** Two unique object exemplars were selected for each of the 84
178 object concepts used in the study. **b.** Stimuli were presented on a grey background for 500 ms, followed by fixation
179 for 500-600 ms (catch trials: 1500 ms). All 84 stimuli from both image sets are shown in Supplemental Figure S1.
180

181 *MEG acquisition and preprocessing*

182 MEG data were recorded continuously at a sampling rate of 1200 Hz with a 275-channel CTF
183 whole-head MEG system (MEG International Services, Ltd., Coquitlam, BC, Canada). All analyses
184 were conducted in MATLAB (version 2016a, The Mathworks, Natick, MA). Preprocessing was
185 carried out using Brainstorm 3.4 (version 02/2016, Tadel et al., 2011) and custom-written code,

186 using similar preprocessing steps as previously published MEG decoding work (Cichy et al., 2014;
187 Grootswagers et al., 2016, Hebart et al., 2018). Recordings were available from 272 channels
188 (dead channels: MLF25, MRF43, MRO13). The whole-head array consists of radial first-order
189 gradiometer channels equipped with synthetic third-gradient balancing to remove background
190 noise online. At the beginning of the experiment and after every third experimental run,
191 participants' head position was localized based on fiducial coil placement at the nasion, left and
192 right preauricular points. Head position was recorded to provide the experimenter with feedback
193 about the head position to reposition the participant's head in the dewar if necessary. Data were
194 bandpass filtered between 0.1 and 300 Hz, and bandstop filtered at 60 Hz and harmonics. We
195 segmented the data into single trial bins, with each trial consisting of 100 ms baseline for
196 normalization purposes and 1000 ms post-stimulus activity, yielding a total of 1321 time samples
197 for each trial. Catch trials were discarded.

198 Three pre-analysis steps allowed us to increase SNR and reduce computational demand:
199 PCA dimensionality reduction, temporal smoothing on PCA components, and data
200 downsampling. Principal components analysis (PCA) was run to reduce the number of channels
201 into the set of most descriptive components. All data for an MEG channel across trials were
202 concatenated for PCA, and the components explaining the least variance were removed to speed-
203 up further processing, with a maximum removal of 50 % of the components (i.e. 136
204 components) or 1 % of the variance, whichever was reached first (Hebart et al., 2018). Since for
205 all participants the smallest 136 components explained less than 1 % of the variance, the data for
206 further analyses contained 136 components. Data across all time points were normalized
207 according to the baseline period of -100 to 0 ms relative to stimulus presentation. To do so, the
208 mean and standard deviation of the baseline period for each component were computed, and
209 the mean was subtracted from the data before dividing by the standard deviation. We then used
210 a Gaussian kernel of ± 15 ms half duration at half maximum (HDHM) to temporally smooth the
211 remaining components, and downsampled the components to 120 Hz (132 samples / trial).

212

213 *Multivariate decoding and temporal generalization analysis*

214

215 *Multivariate MEG decoding*

216 Our goal was to study the representational dynamics during visual object recognition and the
217 emergence of generalizable, conceptual object representations over time. To determine the
218 amount of object information contained in the MEG signal over time, we ran time-resolved
219 multivariate decoding of MEG data using a linear support vector machine classifier (SVM; Chang
220 & Lin, 2011). The analysis steps were chosen according to general recommendations
221 (Grootswagers et al., 2016) and a recent study from our lab (Hebart et al., 2018). Multivariate
222 analyses were conducted using functions from The Decoding Toolbox (Hebart et al., 2015) and
223 custom-written code. The following analysis steps were applied to all participants, regardless of
224 experimental group.

225 First, we created supertrials by averaging 6 trials of the same object concept drawn
226 randomly without replacement (Isik et al., 2014). For each time point, preprocessed MEG data
227 within each supertrial were arranged as P dimensional measurement vectors (corresponding to
228 the number of components from PCA preprocessing), yielding K pattern vectors for each time
229 point and object concept. For each pair of object concepts and each time point, we then trained

230 the classifier on $K-1$ pattern vectors and tested it on the pair of left-out pattern vectors, yielding
231 a decoding accuracy for each pair of object categories at each time point. Note that while leave-
232 one-out cross-validation can lead to some overfitting to the data at hand, when the purpose is to
233 demonstrate a statistical dependence in combination with classical statistics this is a valid
234 approach (Hebart & Baker, 2017). The assignment to training and testing sets and resulting
235 classification procedure was repeated 100 times for each pair of object concepts and each time
236 point, with a new random generation of supertrials in each iteration. The resulting decoding
237 accuracies were averaged across the 100 iterations and presented as an 84×84 matrix at every
238 time point, with rows and columns indexed according to object conditions, and with the diagonal
239 undefined. We used these matrices to evaluate average decoding accuracy at each time point by
240 computing the average of the lower triangular matrix.

241 Significance for the decoding analysis was assessed using a sign permutation test. A null
242 distribution of group means was generated by running the decoding procedure 1,000 times,
243 randomly generating a sign-permuted accuracy per participant and averaging those values. P -
244 values were determined as one minus the percentile of the original group mean in this null
245 distribution. Those p -values were corrected according to the false-discovery rate (FDR) and were
246 deemed significant if the corrected p -value did not exceed 0.05 (i.e. the test was one-sided).

247

248 *Temporal generalization of object representation*

249 While time-resolved multivariate decoding can reveal when specific mental representations are
250 present in patterns of neural activity, it cannot identify how said patterns at one time point relate
251 to other time points. We were interested in investigating the extent to which object-related
252 information is static or dynamic over time, which can give us an index of how rapidly neural
253 signals evolve. To investigate this, we conducted a cross-classification analysis over time, also
254 known as the temporal generalization method (King & Dehaene, 2014; Meyers et al., 2008). If a
255 classifier can successfully generalize from one time point to another, this shows that
256 representational content is highly similar between these two time points. Conversely, if the
257 classifier does not generalize, this shows that patterns of neural activity have evolved to an extent
258 that representational content is no longer similar.

259 To carry out this temporal generalization analysis, we used the same classification
260 approach described above; however, instead of only testing the classifier at the same time point
261 we also tested its performance at all other time points. We repeated the analysis with all time
262 points each serving as training data once for the classifier, and generated a 132×132 time-time
263 decoding matrix that shows the extent to which our classifier generalizes across time.

264

265 *Representational similarity analysis (RSA)*

266

267 RSA is a method to analyze and compare data patterns, for example brain activity patterns with
268 behavioral judgments or computational models (Kriegeskorte et al., 2008). Instead of comparing
269 these patterns directly, in RSA patterns are converted to representational similarity matrices
270 (RSMs), quantifying all pairwise similarities of all patterns. These RSMs can then be compared to
271 other RSMs based on other data.

272 In this study, we used RSA for two purposes. First, across participants we directly
273 compared the time courses of MEG RSMs evoked by the *same* exemplar with MEG RSMs evoked

274 by *different* exemplars. This allows an estimate of the generalizability of representations across
275 exemplars and thus the extent to which a representation reflects high-level versus low-level
276 properties, assuming that a generalized representation indicates a more high-level, conceptual
277 representation. Second, we used RSA to study the relationship between evoked MEG activity
278 patterns and computational, semantic, and behavioral models. In particular, we wanted to
279 identify time periods at which the MEG responses reflected predominantly behavioral
280 judgments, which we take as an index of high-level conceptual processing. To do this, we
281 quantified the unique and shared variance of each model RSM with RSMs based on MEG activity
282 patterns.

283

284 *Construction of MEG similarity matrices*

285 MEG RSMs were constructed as follows. For each time point, we averaged the preprocessed MEG
286 data for all 36 trials of each object concept, yielding 84 object concept MEG patterns. Then we
287 computed the similarity between all pairs of those 84 patterns across P principal components
288 using a Spearman correlation, yielding an 84×84 MEG RSM for each time point. We then
289 analyzed these RSMs further for the two purposes described above.

290

291 *Comparison of low-level image similarity between image sets*

292 To quantify the low-level similarity between image sets directly, we computed the pixelwise
293 similarity across both image sets, concatenating the three color channels of each images to a
294 vector and calculating the Spearman correlation between image vectors. This resulted in an $84 \times$
295 84 matrix, with the diagonal corresponding to the similarity within each object concept across
296 image sets (e.g. “baby” in Image Set 1 with “baby” in Image Set 2) and the off-diagonals
297 corresponding to the similarity across object concepts across object concepts (e.g. “baby” in
298 Image Set 1 with “woman” in Image Set 2). In addition, as a computational model of low-level
299 object processing we computed the GIST features (Oliva & Torralba, 2001) for each object image
300 using default model parameters. We then calculated the Spearman correlation between those
301 feature vectors in the same manner as described for pixelwise similarities.

302

303 *Generalization of MEG similarity patterns across exemplars*

304 To determine time periods that generalize between representations of object exemplars, we
305 compared the time courses of similarity of RSMs *within* each image set to the similarity *between*
306 image sets (see e.g. Guggenmos et al., 2018, for a similar methodological approach). To this end,
307 we split data between the groups for Image Set 1 and Image Set 2 and conducted within- and
308 between-group split-half correlation analyses with the RSMs for each participant. We chose a
309 repeated subsampling procedure within group to allow us to use the same analysis within and
310 between groups. The following analyses are described for one RSM at one time point, but were
311 repeated for all time points.

312 Within each group of participants ($n = 16$), we randomly assigned participants’ RSMs to
313 one of two arbitrary subsets of 8 participants and averaged participants’ RSMs within subsets.
314 Next, we calculated the Spearman rank correlation coefficient between the lower triangular part
315 of each 84×84 matrix, separately for every time point. We repeated this split-half analysis 1000
316 times with novel assignments of participants and averaged across repetitions, yielding a time
317 course of within-exemplar correlation. The same procedure was completed for the between-

318 group split-half analysis, but here the two subsets were each drawn from eight randomly selected
319 participants in each group, yielding a time-course of between-exemplar correlations.

320 To assess statistical significance, we conducted a randomization test. We repeated the
321 analysis above 1000 times (i.e. a total of 10^6 split-half analyses, for both within-exemplar and
322 between-exemplar comparisons). For each of those 1000 randomizations, we randomly
323 permuted the rows and columns of the matrices in one of the subgroups before calculating
324 Spearman's r . P -values were determined as one minus the percentile of the original split-half
325 analysis, and FDR-corrected to $p < 0.05$.

326

327 *Representational similarity matrices for computational models and behavior*

328 To identify and characterize the temporal evolution of the representational content of MEG
329 responses in relation to behavior judgments, we chose multiple behavioral and computational
330 models that we later compared to MEG data: a behavioral model based on the group mean
331 behavioral similarity, a semantic model to capture similarity at the semantic level, and two layers
332 of a deep neural network to capture different visual processing stages (low-to-mid level and high-
333 level, respectively). The purpose of including those models was to identify the contribution of
334 those processing stages to behavior, in order to gain a better understanding of the nature of the
335 behavioral judgments. For a first comparison, we characterized the pairwise similarity of these
336 models to assess their general similarity irrespective of MEG. We calculated Spearman's r for
337 each pair of models. Significance of correlations was tested using a randomization test: The rows
338 and columns of one model RSM were randomly permuted before computing the Spearman's r
339 between with the other model RSM. This procedure was repeated 1,000 times to generate a null
340 distribution of correlation coefficients, and results were deemed significant if they showed a
341 higher correlation coefficient than the distribution cut-off determined by a level of $p < 0.05$.

342

343 Behavior

344 We generated an RSM for behavioral judgments by extracting the 84×84 distance matrices from
345 each participant within a group and averaging them together. Next, we converted this distance
346 matrix to an RSM by subtracting the distances from 1. Note that subsequent analyses only use
347 the ranks of the entries of the distance matrices, which are simply inverted by this subtraction
348 procedure. This step yielded two group-level behavior RSMS corresponding to Image Set 1 and
349 Image Set 2.

350 Semantic model: Global Vectors for Word Representation (GloVe)

351 Global Vectors for Word Representations (GloVe) is an unsupervised algorithm that is trained on
352 corpus word co-occurrence statistics to yield vector representations for words in the corpus,
353 representing semantic relationships between words (Pennington et al., 2014). As a distributional
354 measure of the semantic relatedness of words based on their shared linguistic contexts, GloVe is
355 similar to other traditional co-occurrence models of word meaning but is particularly well-suited
356 to the analysis here because of the high-dimensional similarity structure that shows semantic
357 similarity between pairs of individual words, outperforming similar models in similarity tasks. As
358 such, the structure of GloVe provides a fine-grained metric to evaluate how the representational
359 space of MEG signals reflects semantic relationships as derived from shared lexical contexts. We
360 chose 50-dimensional word vectors pre-trained on a 6-billion token Wikipedia database,

361 extracted them for each object concept in the stimulus set and calculated Spearman's r between
362 each pair of vectors, generating an 84×84 RSM.

363 Visual model: Deep neural network VGG-F

364 We used the MatConvNet toolbox (Vedaldi & Lenc, 2015) to implement a pre-trained version of
365 the Visual Geometry Group-Fast deep neural network (VGG-F DNN) (Chatfield et al., 2014) that
366 was trained to perform the ImageNet ILSVRC 2012 object classification task. This network was
367 chosen based on its high classification performance, ease of implementation, and suitability for
368 our visual object concept stimuli. DNN representations for each image in both image sets were
369 extracted from both convolutional layers (1-5) and fully-connected layers (6-8) of the network.
370 We focused on representative examples of the convolutional and fully connected layers (3 and
371 7, respectively) to reflect low-to-midlevel vision and high-level vision, respectively. Within each
372 layer, we calculated Spearman's r between each of the object conditions that yielded an 84×84
373 RSM for both layers within each participant group. This yielded four distinct RSMs: DNN Layer 3
374 and Layer 7 for Image Set 1, and DNN Layer 3 and Layer 7 for Image Set 2.

375

376 *Representational similarity analysis: Model comparisons to MEG*

377 To directly compare each model to MEG activity patterns, we calculated Spearman's r between
378 the lower diagonals of the model variables and MEG RSMs at each time point within each group.
379 These group-specific correlations were averaged together to yield a time course showing the
380 level of correlation between the model and MEG responses. Upper and lower bounds for noise
381 ceilings were determined within each of the two groups of participants according to Nili et al.
382 (2014): The upper bound was estimated by calculating the correlation between each participant's
383 RSM and the mean group RSM *including* that participant, while the lower bound was estimated
384 by calculating the correlation between each participant's RSM and the mean group RSM
385 *excluding* that participant. The upper and lower bounds from each group were averaged together
386 to yield a mean noise ceiling across all participants. The statistical significance of this suite of
387 representational similarity analyses was determined using randomization tests as described
388 above, permuting the rows and columns of a given model RSM (behavior, GloVe, DNN Layer 3,
389 DNN Layer 7) and for each randomization computing correlation time courses with the original
390 MEG RSMs. Correlations were deemed significant if they exceeded a correlation cut-off
391 determined by a level of $p < 0.05$ (FDR-corrected).

392

393 *Establishing the unique and shared contributions of individual models*

394 To determine the unique and shared variance between models and MEG signals, we conducted
395 multiple linear regression analyses using the behavior RSM, DNN Layer 3 RSM, and DNN Layer 7
396 RSM as regressors to predict MEG RSMs from these variables (see See Groen et al., 2012;
397 Lescroart et al., 2015; Greene et al., 2016, 2018; Hebart et al., 2018 for similar approaches). Given
398 the complexities of describing the unique and shared variance partitions of more than three
399 regressors, we decided to exclude the GloVe model, which showed the weakest correlation with
400 MEG. Post-hoc analyses running different versions of the variance partitioning analysis (replacing
401 the behavior model by GloVe, and separately replacing the DNN Layer 7 model by GloVe)
402 demonstrated that the GloVe model overall explained less MEG variance than behavior
403 (Supplemental Figure S4), while explaining very little unique variance (Supplemental Figure S5).
404 By conducting a series of different multiple regressions with different combinations of model

405 variables, this approach allows us to determine not only the unique MEG variance explained by
406 each model RSM individually, but also the variance shared between any combination of model
407 RSMs. Before conducting variance partitioning analyses, we averaged the group-specific RSMs of
408 both image sets for behavior and DNN models, which yielded very similar results as compared to
409 calculating them separately and averaging results afterwards. We extracted the lower diagonal
410 from the mean MEG RSM at each time point as dependent variables, and assigned each of the
411 models as independent variables. In sum, 7 regression analyses were performed at each time
412 point that each included different combinations of models as regressors: 1) ‘full’ regression,
413 including all three models (DNN Layer 3, DNN Layer 7, behavior), (2-4) ‘combined-predictor’
414 regression, including all pairwise combinations of two models (DNN Layer 3 and behavior, DNN
415 Layer 7 and behavior, DNN Layer 3 and DNN Layer 7), and (5-7) ‘single-predictor regression’
416 including each model on its own. Subtracting the explained variance (R^2) values of these different
417 regression analyses yields portions of variance that are independently explained by each model,
418 the variance that each model shares with the other two models, and the variance shared by all
419 three.

420 For example, the unique variance explained by behavior (region c in the Venn diagram
421 depicted in Figure 6a) is computed as the difference in R^2 between the full regression model
422 (which includes all three regressors and therefore encompasses all regions described by the red,
423 green and blue circle, i.e. $a+b+c+ab+ac+bc+abc$) and a regression model including only DNN Layer
424 3 and 7 (encompassing all regions described by the green and blue circle, i.e. $a+b+ab+ac+bc+abc$).
425 Once the three regions of unique variance (a, b, and c) are obtained in this way, shared variances
426 can be computed. For example, the variance shared by behavior and DNN Layer 7 (region bc) is
427 computed by taking the R^2 resulting from including behavior and the third model (DNN Layer 3)
428 (corresponding to all regions covered by the red and blue circle, i.e. $a+ac+ab+abc+c+bc$) and
429 subtracting both the R^2 obtained when including DNN Layer 3 alone (blue circle, $a+ab+ac+abc$) as
430 well as the unique variance explained by behavior (region c). Finally, the variance shared by all
431 three models (region abc) is computed as the difference between the full regression R^2 and all
432 the sum of all unique variances ($a+b+c$) and shared variances between all combinations of two
433 models ($ab+ac+bc$). Statistical significance was determined using a randomization test as
434 described above, randomizing columns and rows of model matrices 1000 times and repeating
435 the original analysis. For a given iteration, the same randomization was used across all models to
436 fulfill the assumptions of the randomization test. Significance cutoffs for R^2 were set to $p < 0.05$
437 (FDR-corrected).

438
439

440 Results

441

442 Our aim in this study was to characterize the emergence of conceptual representations for visual
443 objects. We applied multivariate decoding and representational similarity analysis to MEG data
444 to examine (1) how object representations generalize across time and object exemplars, and (2)
445 to elucidate the unique and shared contributions of behavioral judgments to measured MEG
446 responses. The resulting temporal profiles inform us about stages of object processing from low-
447 level visual to conceptual representations.

448

449 *Time-resolved representation of object identity*

450 To characterize the time course by which neural signals in the human brain convey information
451 about object identity, we used time-resolved multivariate decoding, conducting pairwise
452 classification between MEG patterns in response to object stimuli (Figure 2a). Object identity
453 information rose rapidly in response to stimulus presentation, with decoding accuracy peaking
454 at 100 ms (mean accuracy: 91.1 %), followed by a slow decay of information that remained
455 significantly above chance after stimulus offset and for the duration of the trial time window
456 (1000 ms post stimulus onset). These results indicate that we were able to detect the temporal
457 unfolding of object-identity information encoded in MEG signals with high accuracy, establishing
458 a correspondence to previous research demonstrating that discriminable object representations
459 emerge well within 100 ms of visual recognition (Carlson et al., 2013; Cichy et al., 2014). Further,
460 these results lay an important foundation for the following analyses in which we delineate what
461 information specifically contributes to these discriminable object representations.

462

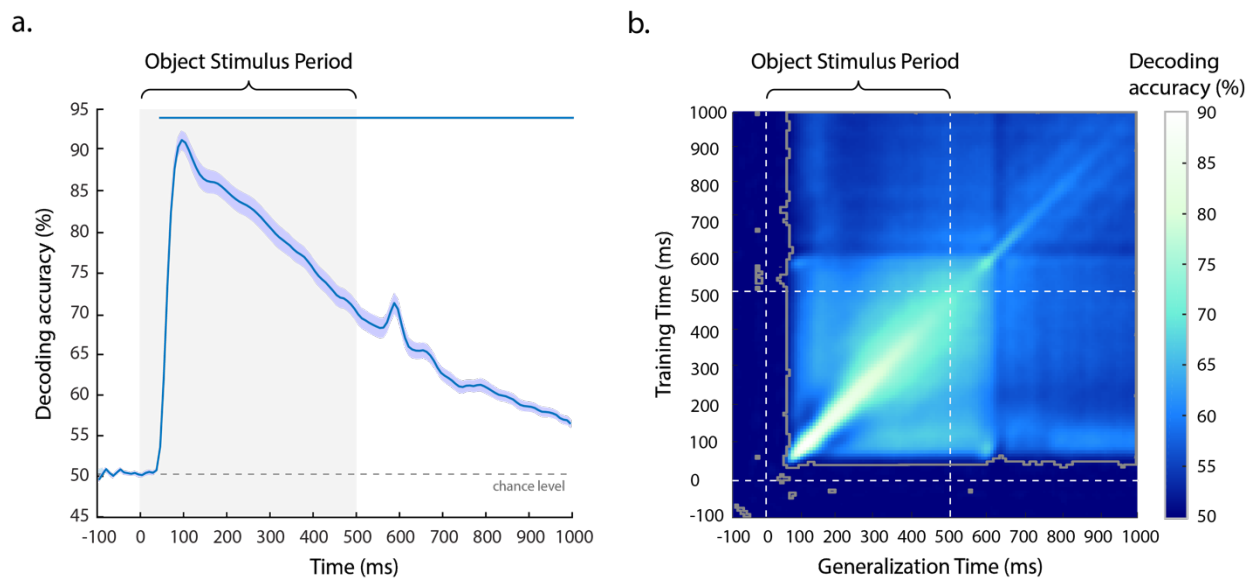
463 *Temporal generalization of object information*

464 While time-resolved multivariate decoding reveals the temporal evolution of discriminable
465 object representations, it does not inform about the dynamics and stability of those
466 representations across time. To identify the degree to which object representations generalize
467 across time, we ran a temporal generalization analysis by training a classifier on data at every
468 time point and testing it at all other time points. This yielded a temporal generalization matrix
469 (Figure 2b), with the diagonal representing training and testing at the same time points, mirroring
470 the results presented in Figure 2a. In a temporal generalization matrix, a dynamic representation
471 would be characterized by high accuracies around the diagonal and low accuracies everywhere
472 else, indicating little generalization across time. In contrast, a stable neural representation would
473 exhibit high decoding around the diagonal but also in the off-diagonal time points, demonstrating
474 a similar representation across time.

475 Our results exhibited significant generalization from ~70 ms onward, demonstrating a
476 shared representational format across the entire trial. While this result reveals a persistent
477 representation across time, the strength of generalization varies. Focusing on the first half of the
478 stimulus presentation period, the results revealed a period of increased temporal dynamics
479 between ~70-250 ms, indicated by the high decoding accuracy on the diagonal and lower
480 decoding accuracies away from the diagonal. This result suggests a relatively dynamic
481 representational format in this phase of visual processing. After ~250 ms, there was increased
482 generalization away from the diagonal, indicating a more persistent, shared representational
483 format during this later phase of visual processing. Interestingly, there was a generalization
484 period between time windows of ~70-100 ms and ~250-550 ms, suggesting an overlap of
485 representations between early visual and later conceptual processing. The markedly lower
486 information generalization between 150-250 ms and all other time points suggests the
487 information dynamics at these points are computationally dissimilar from other stages of
488 processing.

489 Taken together, these results reveal relatively weak but significant persistence of stable
490 object information throughout the entire trial. On top of this, the results reveal a general
491 broadening of information generalization after an early phase of visual processing. While the

492 results of this temporal generalization analysis do not reveal multiple distinct stages of
493 processing, this broadening suggests early dynamic neural activity followed by the emergence of
494 more stable object representations around 250 ms.
495



496 **Figure 2. a.** Time-resolved multivariate decoding of object identity across the trial. After onset of the object stimulus
497 (Object Stimulus Period), pairwise object decoding accuracy increased rapidly, followed by a slow decay towards
498 chance over the duration of the trial. Error bars reflect SEM across participants for each time point separately.
499 Significance is indicated by colored lines above the plot (non-parametric cluster-correction at $p < 0.05$). **b.** Temporal
500 generalization matrix for object identity. The y-axis depicts the classifier training time relative to stimulus onset, and
501 the x-axis classifier generalization time relative to stimulus onset. Dotted lines indicate stimulus onset and offset.
502 Areas bounded by a grey line contain significant temporal cross-decoding accuracy values ($p < 0.05$, FDR corrected).
503 See Supplemental Figure S2, for horizontal cross-sections across the temporal generalization plots.
504
505
506

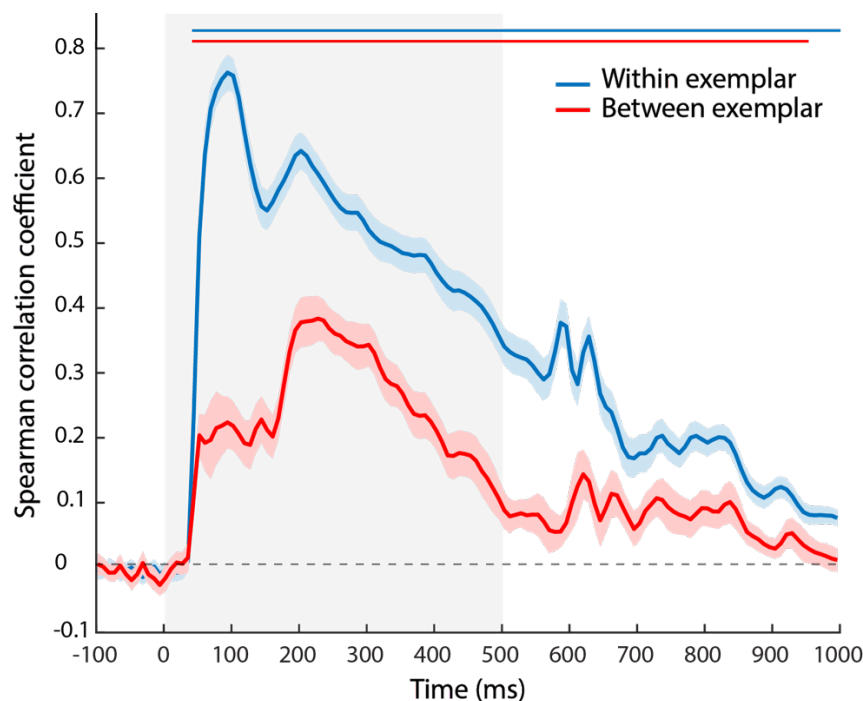
507 *Criterion 1 for conceptual object representation: Generalization between object exemplars*
508 Having established the time course of object identity-specific information, we investigated
509 when those brain responses reflect conceptual object representations. One prerequisite of a
510 conceptual object representation is a similar representational format between multiple
511 exemplars of the same object, since a conceptual representation is expected to generalize
512 beyond each individual exemplar. The data collected from Image Set 1 and 2 allow direct
513 comparison of representational similarity across exemplars for the same visual object concept
514 (Figure 3). We expected some low-level features to be shared across object exemplars, but that
515 this tendency would be reduced as compared to the same exemplar. Indeed, when comparing
516 the low-level similarity between exemplars of image sets, the similarity was higher within
517 object concept than between object concepts (mean pixelwise similarity within: $r = 0.15$,
518 between: $r = 0.04$; mean GIST similarity within: $r = 0.58$, between: $r = 0.41$), demonstrating
519 some preserved similarity between object exemplars. However, the overall similarity was
520 strongly reduced, and the maximal similarity across image sets was for the same object concept
521 in only 19% of the cases (based on the GIST similarity), demonstrating a strongly reduced low-
522 level similarity between image sets.

523 Having demonstrated the reduction in low-level similarity between image sets, we
524 measured this generalization of object concept-specific information by (1) calculating the
525 correlation of within-exemplar MEG RSMs for participants who were shown the same object
526 exemplar and (2) calculating the generalization of between-exemplar MEG RSMs for participants
527 who were shown different object exemplars. Then we compared the shape of these MEG
528 correlation time courses.

529 A comparison of within-exemplar and between-exemplar MEG RSM correlations revealed
530 a generally higher correlation within-exemplar than between-exemplar (mean difference across
531 time: Spearman's r : 0.18, $p < 0.001$, randomization test), indicating that differences between
532 exemplars persisted throughout most of the trial. Reliable structure for within-exemplar MEG
533 RSMs emerged rapidly, peaking at 93 ms (mean Spearman's r : 0.77). This was followed by a fast
534 drop in correlation, and then another rise beginning around 160 ms and peaking at 202 ms (mean
535 Spearman's r : 0.65), after which within-exemplar correlations decreased steadily for the
536 duration of the trial while remaining significantly above chance. The correlation of between-
537 exemplar MEG RSMs also initially increased rapidly, but then reached a plateau at a comparably
538 low level of correlation between ~ 70 and ~ 160 ms (mean Spearman's r : 0.21). Importantly,
539 between-exemplar reliability then increased again after ~ 160 ms, peaking at 227 ms (mean
540 Spearman's r : 0.39). Between-exemplar correlation then slowly decayed back to 0, but remained
541 significantly above chance until 960 ms after stimulus presentation.

542 These results reveal an important dissociation: While within-exemplar correlations
543 reached their maximum around 100 ms, between-exemplar generalization was maximal around
544 200 ms. Thus, this analysis reveals an early processing stage during which generalization is limited
545 by the variable visual features of each individual exemplar, and a later processing stage where
546 the increased generalization likely reflects the development of a more conceptual object
547 representation that is consistent across exemplars.

548



549

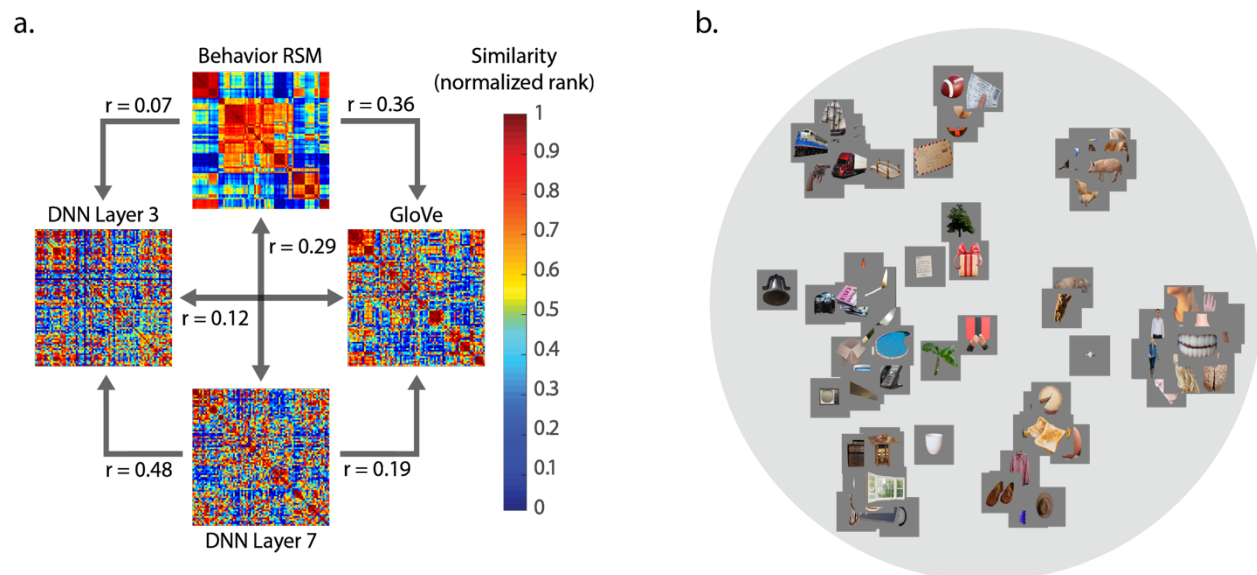
550 **Figure 3.** Within and between exemplar correlation of MEG RSMs. Within-exemplar correlation was generally higher
551 than between-exemplar correlation. Both within and between-exemplar correlations revealed an early peak (93 ms)
552 and a late peak (202 and 227 ms, respectively), with the early peak being higher than the late peak for within-
553 exemplar correlations, and the late peak being higher than the early peak for between-exemplar correlations. Error
554 bars reflect SEM. Significance is indicated by colored lines above the accuracy plot (non-parametric cluster-
555 correction at $p < 0.05$).

556

557 *Comparison of behavior and computational models of low-level and high-level processing*

558 To quantify how the RSMs derived from behavior (perceptual judgments, visualized in Figure 4b),
559 GloVe (lexical semantics), DNN Layer 3 (low/mid-level visual information), and DNN Layer 7 (high-
560 level visual information) relate to one another, we computed the correlation between each pair
561 of model RSMs (Figure 4a). For visualization purposes, we applied hierarchical clustering to
562 independent pilot data of the behavioral task to sort objects depicted in the model RSMs (Figure
563 4a). All model correlations were significant at a level of $p < 0.001$ (randomization test). An
564 estimate of the upper noise ceiling for possible model correlation values was calculated by the
565 correlation between behavior RSMs for the two groups of participants (Spearman's $r = 0.64$). The
566 greatest similarity to behavior was shown by the GloVe model. There was low similarity of
567 convolutional DNN Layer 3 with behavior and GloVe, but much greater similarity for fully-
568 connected DNN Layer 7. These results suggest an increase of semantic, behaviorally-related
569 information contained in the representational structure of the DNN Layer 7 as compared to Layer
570 3. Note that the lowest correlation observed was between DNN Layer 3 and behavior RSMs,
571 indicating that behavior was not strongly driven by low- to mid-level responses. As a post-hoc
572 analysis, we added the comparison of behavior to the GIST RSM, which was even lower
573 (Spearman's $r = 0.02$), highlighting the low explanatory power of low-level features in behavioral
574 judgments in the present study.

575



576

577

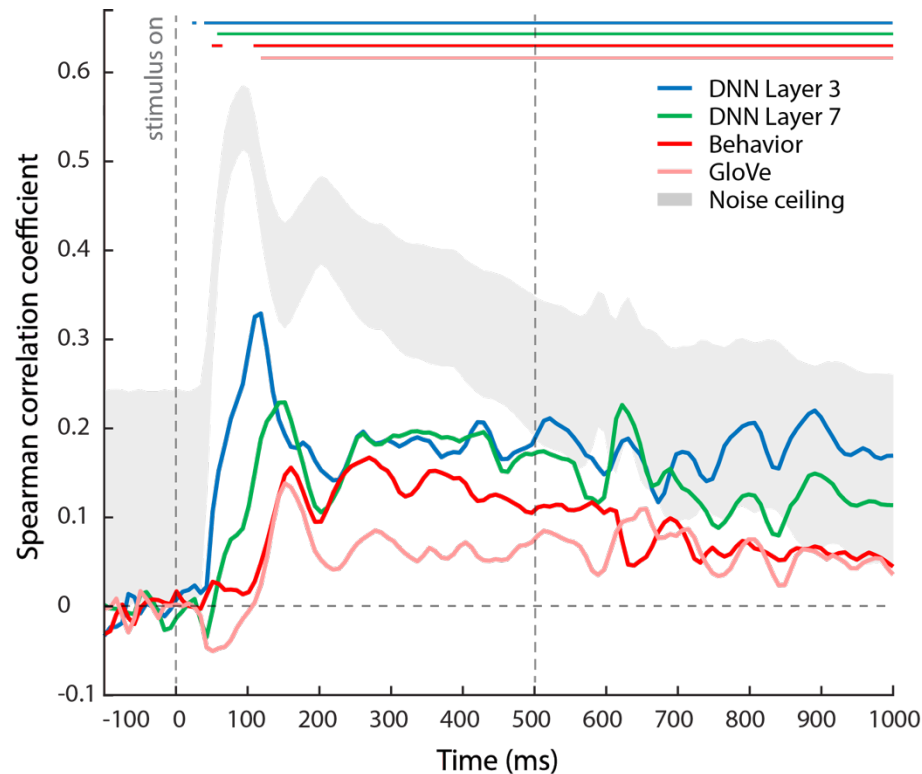
578 **Figure 4: a.** Explicit comparison of computational models and behavior using RSA. Models compared are group
579 average behavior, GloVe, DNN Layer 3 and DNN Layer 7. RSAs are plotted as ranks for higher visual contrast. Objects
580 are sorted based on clustering generated from independent pilot data. **b.** Group average inverse MDS plot
581 generated from behavioral arrangement task.

582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607

Criterion II for conceptual object representation: Behavioral and computational modeling of MEG data

To determine when there is a relationship between the MEG signal and high-level behavioral judgments, satisfying Criterion II, we first evaluated the time course of similarity between behavioral judgments and the MEG activity patterns (Figure 5). Further, to establish whether this relationship was uniquely explained by behavior, we additionally compared MEG to the computational models described above. Every model tested exhibited significant correlations with MEG activity patterns within the first 200 ms of visual processing. DNN Layer 3 showed peak correlation with MEG at 118 ms after stimulus onset (Spearman's $r = 0.33$), while DNN Layer 7 showed peak correlation with MEG at 151 ms (Spearman's $r = 0.23$). Further, the GloVe model was most strongly correlated with MEG at 151 ms (Spearman's $r = 0.13$), and behavior at 160 ms (Spearman's $r = 0.16$). Additional within-subject analyses, i.e. comparing each individual's behavioral RSM to their MEG RSM, revealed a very similar pattern of results but lower overall correlations (peak Spearman's $r = 0.06$) and no significant benefit of within-subject over between-subject analyses matched in size (all $p > 0.12$).

This sequence of peaks suggests an evolution from low-level visual to high-level conceptual representations, with the relationship to behavior peaking latest in time. However, given the significant correlations of all models with MEG throughout most of the trial and the presence of significant correlation between the models themselves (Figure 4a), it is unclear to what extent a given correlation was uniquely explained by one model, or whether this correlation could equally well be explained by other models. For example, the correlation of both DNN Layer 7 and behavior with MEG signals after 150 ms raises the question whether the behavioral correlations can be fully explained by the features represented in the DNN models.



608
609

610 **Figure 5.** Results of model-based representational similarity analysis with MEG data. Comparison includes models
611 based on DNN Layer 3, DNN Layer 7, GloVe and behavior. The results exhibit a progression of peaks from DNN Layer
612 3 to behavior, suggesting a temporal evolution of the underlying representation from more low-level to higher-
613 level/conceptual. Grey shaded area depicts the noise ceiling. Significant time points are indicated by a colored line
614 above the plots ($p < 0.05$, FDR-corrected permutation test).

615

616 *Variance Partitioning: Shared and unique model contributions*

617 To provide a deeper understanding of the unique contributions of different models to MEG
618 variance and how much explanatory power they share with behavioral judgments in explaining
619 MEG variance, we conducted a variance partitioning analysis in which we compared the results
620 of different multiple regression analyses applied to MEG RSMs (see Methods; Figure 6a). We first
621 considered the total percent of variance in the MEG RSMs explained when all three predictors
622 are combined in a single regression model ('full model') in comparison to the percent variance
623 explained by each model separately (Figure 6b). Since variance explained by each model
624 separately is identical to the square of the model correlation, the results of this analysis are very
625 similar to those of the previous section presented in Figure 5, with the only difference that these
626 results were collapsed across groups before conducting variance partitioning. Explained variance
627 of DNN Layer 3 peaked at 118 ms (R^2 : 11.0 %), DNN Layer 7 at 151 ms (R^2 : 7.0 %), and behavior
628 at 160 ms (R^2 : 4.8 %). Importantly, however, the dashed line indicates how these contributions
629 relate to the total variance accounted for by all three models combined. At its peak at 118 ms,
630 the full model explains 11.6 % of the variance, which is similar to the amount of variance
631 explained by DNN Layer 3 alone, suggesting that all variance captured at this time-point can be
632 attributed uniquely to DNN Layer 3, with limited additional contribution of DNN Layer 7 or
633 behavior. At later time points, however, the full model always substantially explains more

634 variance than the individual predictors, providing a first clue that some or all of these predictors
635 contribute unique (i.e., additive) variance.

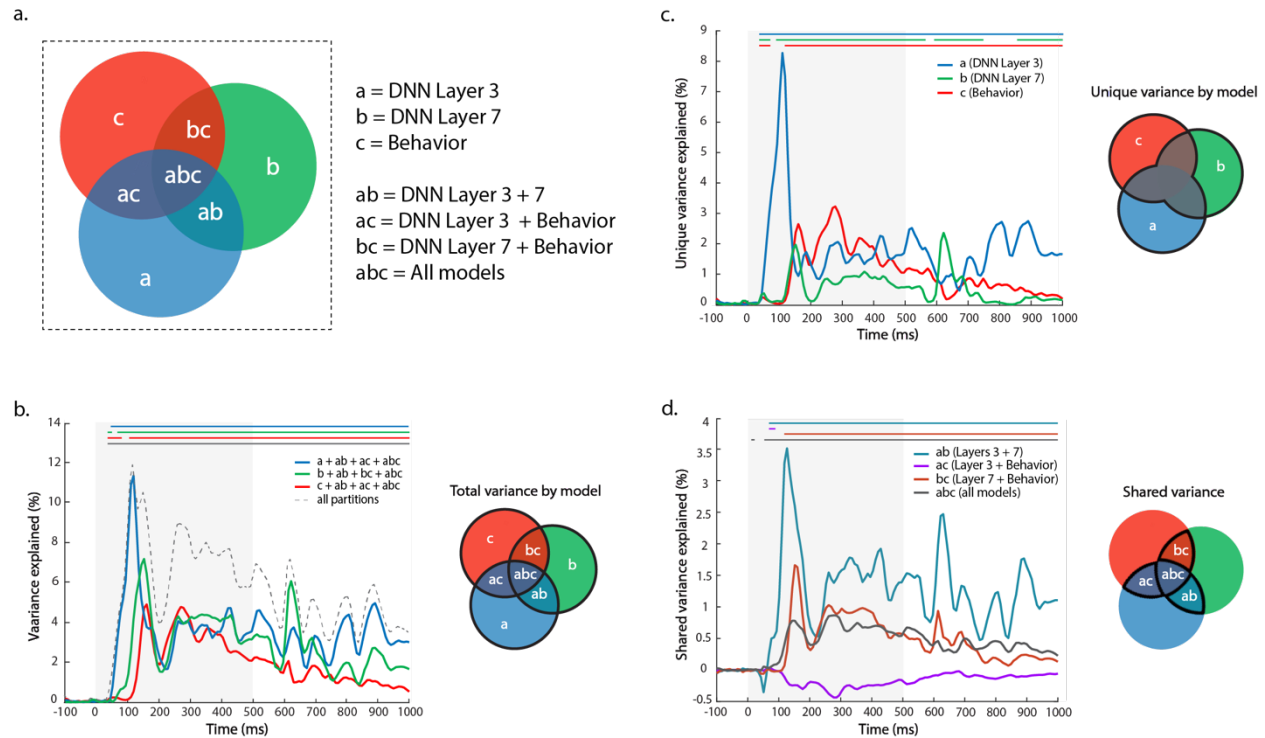
636 To directly quantify the unique and shared variance of each model, we compared the
637 regression outcomes with different model variables included (Figures 6c, 6d). The unique MEG
638 variance explained by DNN Layer 3 peaked very early in time, at 109 ms (R^2 : 8.3%). DNN Layer 7
639 peaked next at 151 ms (R^2 : 2.0 %), followed closely by behavior at 160 ms (R^2 : 2.6 %), with a
640 second peak at 277 ms (R^2 : 3.2 %). Importantly, DNN Layer 3 explained the most unique variance
641 until 143 ms, after which behavior predicted the most unique variance until ~400 ms. Thus, while
642 all three models (DNN Layer 3, DNN Layer 7 and behavior) captured some unique variance in
643 MEG activity throughout the trial, behavior dominated after around 150 ms.

644 Finally, to complete the picture, we partitioned the variance into shared contributions
645 from combinations of the different models. Both DNN Layers contributed the most shared
646 variance across all time points after stimulus onset, which is perhaps not surprising considering
647 that both layers are derived from the same computational model. This shared variance between
648 DNN Layer 3 and Layer 7 peaked at 126 ms (R^2 : 3.5 %). Interestingly, the shared variance between
649 behavior and DNN Layer 7 demonstrated a clear peak at 151 ms (R^2 : 1.7 %), suggesting that it is
650 around this time-point that DNN Layer 7 best captures neural information that is also reflected
651 in behavior. The shared variance between DNN Layer 3 and behavior was slightly negative, a
652 result that is not untypical for variance partitioning, indicative of small suppression effects
653 (Pedhazur, 1997) and suggesting that DNN Layer 3 does not capture information that is relevant
654 for behavioral judgments.

655 It is possible that DNN Layer 3 did not accurately capture the low-level responses. For this
656 reason, we ran additional variance partitioning analyses, replacing DNN Layer 3 with the GIST
657 model. The GIST RSM exhibited a strong correlation with the DNN Layer 3 RSM (Spearman's r :
658 0.65). As expected from this correlation, the variance partitioning results were qualitatively very
659 similar (Supplemental Figure S3), demonstrating that DNN Layer 3 likely captured relevant low-
660 level responses.

661 Collectively, the variance partitioning results indicate that behavioral judgments are
662 reflected in the MEG response above and beyond what is captured by the DNN, with behavioral
663 judgments explaining the most unique variance between 200 and 400 ms after stimulus onset.
664 Further, before 150 ms, DNN Layer 3 explains the most variance, suggesting that representations
665 prior to this point are unlikely to be conceptual in nature.

666



667
668
669
670
671
672
673
674
675

Figure 6. Time-resolved variance partitioning: Total, shared, and unique MEG variance explained by models: DNN Layer 3, DNN Layer 7, and behavior. **a.** Schematic of unique and shared variance components using a Venn diagram. **b.** Percent MEG variance explained by each model independently (colored lines), and total MEG variance explained at all time points (dotted line). **c.** Unique variance explained by each model. **d.** Shared variance between different model combinations. Significant time points are indicated by a colored line above the plots ($p < 0.05$, FDR-corrected permutation test).

676 Discussion

677

678 In this study, we investigated the temporal evolution of visual object representations. In
 679 particular we focused on determining a lower bound for the emergence of conceptual
 680 representations of objects. We proposed two criteria that would reflect conceptual
 681 representations: 1) generalization of representations between different exemplars of the same
 682 object, and 2) relationship to high-level behavioral judgments. We find qualitatively different
 683 processing of objects over time: Early responses (< 150 ms) were characterized by exemplar-level
 684 representations and similarity with computational visual models, whereas later responses (> 150
 685 ms) showed increasing generalization across exemplars and similarity with behavioral judgments,
 686 with greater stability of representations over time.

687 To evaluate generalization of representations reflecting conceptual processing, we
 688 compared the representational structure of MEG responses, both within exemplar and between
 689 sets of exemplars. This analysis revealed two interesting features. First, between-exemplar
 690 generalization was found to be consistently lower than within-exemplar generalization,
 691 demonstrating the persistence of exemplar-specific responses. This reduced between-exemplar
 692 generalization likely reflects the impact of low-level features varying between different

693 exemplars. The fact that this advantage is maintained throughout the trial, suggests some
694 persistence of low-level feature representation. This interpretation is supported by the temporal
695 generalization even for very early time points and the variance explained by DNN Layer 3 (which
696 likely corresponds to early to mid-level visual processing, Cichy et al., 2016a; Güçlü & van Gerven,
697 2015; Wen et al., 2017), throughout the trial. Second, both within and between-exemplar
698 generalization showed two distinct peaks, one early around 100 ms, and another late around 200
699 ms. However, their relative amplitude was reversed: While the early peak was stronger than the
700 second within-exemplar, this pattern was reversed between-exemplar. This striking increase in
701 generalization between exemplars that occurs for the later peak suggests the emergence of a
702 common representation across exemplars, a key marker for conceptual representations.
703 Together these results suggest that the earliest time point for the emergence of conceptual
704 representations is around 150 ms, but also suggest a prolonged representation of low-level visual
705 features.

706 To evaluate the relationship to high-level behavioral judgments, we compared models
707 derived from behavior, semantics (GloVe), and computational vision (DNN) with the MEG
708 response to objects. We found that all models show significant correlation with the MEG
709 response throughout most of the trial. The early DNN layer showed the strongest and earliest
710 correlation, while the GloVe model showed the weakest correlation. This result highlights the
711 importance of testing multiple models rather than relying on a significant effect for a single
712 model. Since the models themselves are correlated (Figure 4), this demonstrates that testing
713 multiple models is *also* not sufficient; it is important to determine the unique and shared variance
714 explained by the different models (Lescroart et al., 2015; Groen et al., 2012; Greene et al., 2016;
715 Hebart et al., 2018; Groen et al., 2018), motivating our variance partitioning analysis. Given the
716 complexities of describing the unique and shared variance partitions of more than three model
717 variables, we decided to exclude one of the four. Since the GloVe model showed the weakest
718 correlation with MEG and was mostly subsumed by the behavioral model, we focused on the
719 DNN and behavioral model variables.

720 The variance partitioning revealed several important features. Focusing on the unique
721 contribution of each model variable, it becomes clear that DNN Layer 3 dominates early MEG
722 responses peaking at 100 ms, whereas behavior explains the most variance after 150 ms, peaking
723 at 270 ms. This result fulfills our second criterion – relationship with high-level behavioral
724 judgments – converging with the results of both the temporal generalization analysis and the
725 representational generalization across exemplars in identifying the time period after around 150
726 ms as reflecting a lower bound for the emergence of conceptual representations. Focusing on
727 the shared contribution of model variables, the results largely reflect the correlations between
728 model variables (Figure 4), e.g. no shared variance between DNN Layer 3 and behavior, high
729 shared variance between Layers 3 and 7 of the DNN model. However, they provide important
730 information about the timing of the shared variances. In particular, the shared variance between
731 DNN Layers 3 and 7 persisted even late in time, again suggesting a sustained representation of
732 low-level visual information.

733 Our results are generally consistent with prior work investigating how visual processing
734 of objects evolves over time, showing the gradual emergence of high-level representations
735 (Contini et al., 2017). While early signals reflect low-level visual features (e.g. Groen et al., 2013;
736 Cichy et al., 2014; Coggan et al., 2016), later signals reflect perceptual similarity (Wardle et al.,

737 2016), some tolerance for changes in size and position (Isik et al., 2014), categorical processing
738 (Carlson et al., 2013; Cichy et al., 2014), and correlate with task performance and reaction times
739 (Van Rullen and Thorpe, 2001; Philiastides and Sajda, 2006; Martinovic et al., 2008; Ritchie et al.,
740 2015). Further, comparisons of deep neural networks with MEG have revealed a correspondence
741 of early layers with earlier MEG responses, likely reflecting initial stages of processing in early
742 visual cortex, while higher layers reflect later stages of processing in occipitotemporal cortex
743 (Cichy et al., 2016b; Seeliger et al., 2017). Our results significantly extend these results by
744 establishing a lower bound for the development of conceptual representations.

745 Other studies have also investigated high-level conceptual processing over time using
746 explicit semantic feature models (Clarke & Tyler, 2015) or behavioral judgments (Cichy et al.,
747 2017). For example, Clarke and colleagues showed semantic feature effects before 120 ms,
748 although including basic visual features based on the HMAX model revealed unique semantic
749 contributions to MEG signals only after ~200 ms (Clarke et al., 2013; Clarke et al., 2014). In
750 contrast to these studies, we used more recent deep convolutional neural networks which have
751 been shown to be more closely tied to neural and behavioral data (Khaligh-Razavi et al., 2016;
752 Jozwik et al., 2017; Cichy et al., 2016a). Further, we operationalized high-level conceptual
753 processing, using both a computational semantic model based on semantic co-occurrence
754 statistics (GloVe model), as well as behavioral judgments of object similarity that we take to more
755 broadly reflect conceptual processing. Indeed, our results suggest that MEG variance explained
756 by the GloVe model was comparably low and mostly covaried with behavioral judgments,
757 suggesting that conceptual representations extend beyond those relationships captured by the
758 GloVe model. Despite these differences, our results are generally consistent with the results of
759 Clarke and colleagues, but suggest a lower bound for conceptual processing around ~150 ms (see
760 also Cichy et al., 2017). Further, we show that the computational visual model and behavioral
761 judgments explain shared variance even prior to 150 ms. This shared variance indicates that the
762 neural activity captured by computational models is behaviorally relevant and argues against a
763 strong distinction between (low-level) visual features on the one hand, and high-level conceptual
764 processing on the other. At the same time, the presence of significant unique variance explained
765 by behavior after 150 ms suggests that not *all* aspects of conceptual object representations
766 reflected in MEG activity are explained by current generations of computational visual models.

767 While our study provides insight into the development of conceptual representations,
768 there are some important considerations. First, we used behavioral similarity judgments using
769 the multi-arrangement task (Kriegeskorte & Mur, 2012) to index conceptual processing.
770 However, this choice of method might constrain the ability to capture conceptual
771 representations. While the behavioral judgments explain more variance in the MEG signal than
772 the semantic GloVe model we tested, we do not know what aspects of conceptual processing are
773 reflected in those judgments. Further, it is unclear how sensitive those behavioral judgments are
774 to the context imposed by the stimuli and instructions. Second, we only employed two exemplars
775 per object concept to test generalization of representations, and this may not have contained
776 sufficient variability to fully disentangle low-level and high-level processing. In future work, it
777 would be useful to carry out similar analyses while presenting multiple image sets to each
778 participant, which would allow within-subject exploration of differences in temporal
779 generalization across exemplars. Finally, in the future alternative similarity metrics for MEG-
780 based RSA, such as the cross-validated Mahalanobis distance could be applied that may increase

781 the robustness over the current approach (Guggenmos et al., 2018). Future studies should
782 consider broader sets of stimuli, different behavioral tasks, and alternative computational
783 models that may better match the MEG signal.

784 In conclusion, by focusing on two criteria for conceptual object representations we
785 provide an estimate for a lower bound for the emergence of conceptual object representations
786 of around 150 ms. Prior to this time, our results demonstrate limited generalization across object
787 exemplars and time, and importantly little unique contributions of behavioral judgments to the
788 MEG response. The multifaceted nature of our findings here show that the combination of neural
789 data, behavior, and models are a viable method to probe the temporal dynamics of object
790 recognition and allow us to establish a novel profile of emergent conceptual representations in
791 time.

792 **References**

793

794 Brainard, D. H., 1997. The Psychophysics Toolbox. *Spat. Vis.* 10, 433-436.

795 Brysbaert, M., Wariner, A. B., & Kuperman, V., 2014. Concreteness ratings for 40 thousand
796 generally known English word lemmas. *Behav. Res. Methods* 46, 904-911.

797 doi:10.3758/s13428-013-0403-5.

798 Bullier, J., 2001. Integrated model of visual processing. *Brain Res. Rev.* 36, 96–107.

799 doi:10.1016/S0165-0173(01)00085-6.

800 Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., DiCarlo,
801 J. J., 2014. Deep neural networks rival the representation of primate IT cortex for core
802 visual object recognition. *PLoS Comp. Bio.* 10, e1003963.

803 doi:10.1371/journal.pcbi.1003963.

804 Carlson, T. A., Tovar, D., Alink, A., Kriegeskorte, N., 2013. Representational dynamics of object
805 vision: The first 1000 ms. *J. Vis.* 13, 1–19. doi:10.1167/13.10.1.doi.

806 Carlson, T. A., Hogendoorn, H., Kanai, R., Mesik, J., Turret, J., 2011. High temporal resolution
807 decoding of object position and category. *J. Vis.* 11, 1-17. doi:10.1167/11.10.9.

808 Chang, C. C., Lin, C. J., 2011. LIBSVM: a library for support vector machines. *ACM Trans. Intell.*
809 *Syst. Technol.*, 2-27. doi:10.1145/1961189.1961199.

810 Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A., 2014. Return of the devil in the details:
811 Delving deep into convolutional nets. *Brain Mach. Vis. Conf.*

812 Cichy, R. M., Khosla, A., Pantazis, D., Oliva, A., 2016a. Comparison of deep neural networks to
813 spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical
814 correspondence. *Sci. Rep.* 6, 27755. doi:10.1038/srep27755.

815 Cichy, R. M., Kriegeskorte, N., Jozwik, K. M., van den Bosch, J. J. F., Charest, I., 2017. Neural
816 dynamics of real-world object vision that guide behavior. *bioRxiv*. doi:

817 <http://dx.doi.org/10.1101/147298>.

818 Cichy, R. M., Pantazis, D., Oliva, A., 2014. Resolving human object recognition in space and time.
819 *Nat. Neurosci.* 17, 455-464. doi:10.1038/nn.3635.

820 Cichy, R.M., Pantazis, D., Oliva, A., 2016b. Similarity-based fusion of MEG and fMRI reveals
821 spatio-temporal dynamics in human cortex during visual object recognition. *Cereb. Cortex*
822 26, 3563-3579. doi:10.1093/cercor/bhw135.

823 Clarke, A., Devereux, B. J., Randall, B., Tyler, L. K., 2014. Predicting the time course of individual
824 objects with MEG. *Cereb. Cortex* 10, 3602-3612. doi:10.1093/cercor/bhu203.

825 Clarke, A., Taylor, K. I., Devereux, B., Randall, B., Tyler, L. K., 2013. From perception to
826 conception: how meaningful objects are processed over time. *Cereb. Cortex* 23, 187-197.

827 doi:10.1093/cercor/bhs002.

828 Clarke, A., Tyler, L. K., 2015. Understanding what we see: How we derive meaning from vision.
829 *Trends Cog. Sci.* 19, 677-687. doi:10.1016/j.tics.2015.08.008

830 Coggan, D.D., Baker, D.H., Andrews, T.J., 2016. The role of visual and semantic properties in the
831 emergence of category-specific patterns of neural response in the human brain. *eNeuro* 3,

832 e0158-16.2016 1-10.

- 833 Contini, E. W., Wardle, S. G., Carlson, T. A., 2017. Decoding the time-course of object
834 recognition in the human brain: From visual features to categorical decisions.
835 *Neuropsychologia* 105, 165-176. doi: 10.1016/j.neuropsychologia.2017.02.013.
- 836 Davies, M., 2008. The Corpus of Contemporary American English (COCA): 520 million words,
837 1990-present.
- 838 Eickenberg, M., Gramfort, A., Varoquaux, G., Thirion, B., 2017. Seeing it all: Convolutional
839 network layers map the function of the human visual system. *NeuroImage* 152, 184-194.
840 doi:<https://doi.org/10.1016/j.neuroimage.2016.10.001>.
- 841 Farah, M. J., McClelland, J. L., 1991. A computational model of semantic memory impairment:
842 Modality specificity and emergent category specificity. *J. Exp. Psychol. Gen.* 120, 339–357.
843 doi:10.1037/0096-3445.120.4.339.
- 844 Goldstone, R., 1994. An efficient method for obtaining similarity data. *Behav. Res. Methods*
845 *Instrum. Comput.* 26, 381-386. doi:10.3758/BF03204653.
- 846 Güçlü, U., van Gerven, M. A. J., 2015. Deep neural networks reveal a gradient in the complexity
847 of representations across the ventral stream. *J. Neurosci.* 35, 10005-10014.
848 doi:<https://doi.org/10.1523/JNEUROSCI.5023-14.2015>.
- 849 Greene, M. R., Baldassano, C., Esteva, A., Beck, D. M., Li, F. F., 2016. Visual scenes are
850 categorized by function. *J. Exp. Psych.* 145, 82-94. doi:10.1037/xge0000129.
- 851 Groen, I. I. A., Ghebreab, S., Lamme, V. A. F., Scholte, H. S., 2012. Spatially pooled contrast
852 responses predict neural and perceptual similarity of naturalistic image categories. *PLoS*
853 *Comput. Biol.* 8, e1002726. doi:10.1371/journal.pcbi.1002726.
- 854 Groen, I. I. A., Ghebreab, S., Prins, H., Lamme, V. A. F., Scholte, H. S., 2013. From image statistics
855 to scene gist: Evoked neural activity reveals transition from natural image structure to
856 scene category. *J. Neurosci.* 33, 18814-18824. doi:
857 <https://doi.org/10.1523/JNEUROSCI.3128-13.2013>.
- 858 Groen, I. I. A., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., Baker, C. I., 2018. Distinct
859 contributions of functional and deep neural network features to representational
860 similarity of scenes in human brain and behavior. *eLife* 7, e32962.
- 861 Groen, I. I. A., Silson, E. H., Baker, C. I., 2017. Contributions of low- and high-level properties to
862 neural processing of visual scenes in the human brain. *Phil. Trans. B.* 372.
863 doi:10.1098/rstb.2016.0102.
- 864 Grootswagers, T., Wardle, S. G., Carlson, T. A., 2016. Decoding dynamic brain patterns from
865 evoked responses: A tutorial on multivariate pattern analysis applied to time series
866 neuroimaging data. *J. Cogn. Neurosci.* 29, 677-697. doi:10.1162/jocn_a_01068.
- 867 Guggenmos, M., Sterzer, P., Cichy, R. M., 2018. Multivariate pattern analysis for MEG: A
868 comparison of dissimilarity measures. *Neuroimage*. doi:
869 10.1016/j.neuroimage.2018.02.044.
- 870 Hebart, M. N., Baker, C. I., 2017. Deconstructing multivariate decoding for the study of brain
871 function. *Neuroimage*. doi: <https://doi.org/10.1016/j.neuroimage.2017.08.005>.
- 872 Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I., Cichy, R., M., 2018. The representational
873 dynamics of task and object processing in humans. *eLife* 7, e32816.
- 874 Hebart M. N., Görgen K., Haynes J-D., 2015. The Decoding Toolbox (TDT): a versatile software
875 package for multivariate analyses of functional imaging data. *Front. Neuroinform.* 8:88.

- 876 Isik, L., Meyers, E. M., Leibo, J. Z., Poggio, T., 2014. The dynamics of invariant object recognition
877 in the human visual system. *J. Neurophysiol.* 111, 91-102. doi: 10.1152/jn.00394.2013.
- 878 Johnson, J. S., Olshausen, B. A., 2003. Timecourse of neural signatures of object recognition. *J.*
879 *Vis.* 3, 499–512. doi: 10:1167/3.7.4.
- 880 Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., Mur, M., 2017. Deep convolutional neural networks
881 outperform feature-based but not categorical models in explaining object similarity
882 judgments. *Front. Psychol.* 8, 1726. doi:<https://doi.org/10.3389/fpsyg.2017.01726>
- 883 Khaligh-Razavi S-M., Henriksson, L., Kay, K., Kriegeskorte, N., 2016. Fixed versus mixed RSA:
884 Explaining visual representations by fixed and mixed sets from shallow and deep
885 computational models. *J. Math. Psychol.* 76, 184-197.
886 doi:<https://doi.org/10.1016/j.jmp.2016.10.007>.
- 887 Khaligh-Razavi, S-M., Kriegeskorte, N., 2014. Deep supervised, but not unsupervised,
888 models may explain IT cortical representation. *PLoS Comp. Bio.* 10, e1003915. doi:
889 <https://doi.org/10.1371/journal.pcbi.1003915>.
- 890 King, J. R., Dehaene, S., 2014. Characterizing the dynamics of mental representations: the
891 temporal generalization method. *Trends Cog. Sci.* 18, 203-210.
892 doi:10.1016/j.tics.2014.01.002.
- 893 Kriegeskorte, N., Mur, M., 2012. Inverse MDS: Inferring dissimilarity structure from multiple
894 item arrangements. *Front. Psychol.* 3, 245. doi: 10.3389/fpsyg.2012.00245.
- 895 Lescroart, M. D., Stansbury, D. E., Gallant, J. L., 2015. Fourier power, subjective distance, and
896 object categories all provide plausible models of BOLD responses in scene-selective visual
897 areas. *Front. Comput. Neurosci.* 9. doi: 10.3389/fncom.2015.00135
- 898 Martinovic, J., Gruber, T., Müller, M. M., 2008. Coding of visual object features and feature
899 conjunctions in the human brain. *PLoS One* 3, e3781. doi:10.1371/journal.pone.0003781.
- 900 McRae, K., Seidenberg, M. S., De Sa, V. R., 1997. On the nature and scope of featural
901 representations of word meaning. *J. Exp. Psychol. Gen.* 126, 99–130.
902 doi:10.1080/10430719008404646.
- 903 Meyers, E. M., Freedman, D. J., Kreiman, G., Miller, E. K., Poggio, T., 2008. Dynamic population
904 coding of category information in inferior temporal and prefrontal cortex. *J. Neurophysiol.*
905 100, 1407-1419. doi:10.1152/jn.90248.2008.
- 906 Nili, H., Wingfield, C., Walther, A., Su, L., Marslen-Wilson, W., Kriegeskorte, N., 2014. A Toolbox
907 for Representational Similarity Analysis. *PLoS Comput. Biol.* 10, e1003553.
908 doi:10.1371/journal.pcbi.1003553.
- 909 Oliva, A., Torralba, A., 2001. Modeling the shape of the scene: A holistic representation of the
910 spatial envelope. *Int. J. Comput. Vis.* 42, 145–175.
- 911 Pedhazur, E. J., 1997. Multiple regression in behavioral research: Explanation and prediction, 3rd
912 Edition. Orlando, FL: Harcourt Brace.
- 913 Pennington, J., Socher, R., Manning, C. D., 2014. GloVe: Global vectors for word representation.
914 *Proc. of 2014 Conf. Empir. Methods Nat. Lang. Process.*
- 915 Piliastides, M. G., Sajda, P., 2006. Temporal characterization of the neural correlates of
916 perceptual decision making in the human brain. *Cereb. Cortex* 16, 509-518.
917 doi:10.1093/cercor/bhi130.

- 918 Potter, M. C., Wyble, B., Hagmann, C. E., McCourt, E. S., 2014. Detecting meaning in RSVP at 13
919 ms per picture. *Attent. Percept. Psychophys.* 76, 270-279. doi:[10.3758/s13414-013-0605-](https://doi.org/10.3758/s13414-013-0605-z)
920 [z](https://doi.org/10.3758/s13414-013-0605-z).
- 921 Ritchie, J. B., Kaplan, D., Klein, C., 2017. Decoding the brain: Neural representation and the
922 limits of multivariate pattern analysis in cognitive neuroscience. *BioRxiv*. doi:
923 [10.1101/127233](https://doi.org/10.1101/127233).
- 924 Ritchie, J. B., Tovar, D. A., Carlson, T. A., 2015. Emerging object representations in the visual
925 system predict reaction times for categorization. *PLoS Comp. Biol.* 11, e1004316.
926 doi:[10.1371/journal.pcbi.1004316](https://doi.org/10.1371/journal.pcbi.1004316).
- 927 Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., Boyes-Braem, P., 1976. Basic objects in
928 natural categories. *Cogn. Psychol.* 8, 382–439. doi:[10.1016/0010-0285\(76\)90013-X](https://doi.org/10.1016/0010-0285(76)90013-X).
- 929 Schendan, H. E., Maher, S. M., 2009. Object knowledge during entry-level categorization is
930 activated and modified by implicit memory after 200 ms. *NeuroImage* 44, 1423–1438.
931 doi:[10.1016/j.neuroimage.2008.09.061](https://doi.org/10.1016/j.neuroimage.2008.09.061).
- 932 Seeliger, K., Fritsche, M., Güçlü, U., Schoenmakers, S., Schoffelen, J., Bosch, S. E., Gerven, M. A.
933 J., 2017. Convolutional neural network-based encoding and decoding of visual object
934 recognition in space and time. *NeuroImage* 155. doi:
935 <https://doi.org/10.1016/j.neuroimage.2017.07.018>.
- 936 Tadel F., Baillet S., Mosher J. C., Pantazis D., Leahy R. M., 2011. Brainstorm: a user-friendly
937 application for MEG/EEG analysis. *Comput. Intell. Neurosci.* 8. doi:[10.1155/2011/879716](https://doi.org/10.1155/2011/879716).
- 938 Thorpe, S., Fize, D., Marlot, C., 1996. Speed of processing in the human visual system. *Nat.* 381,
939 520-522. doi:[10.1038/381520a0](https://doi.org/10.1038/381520a0).
- 940 Tyler, L. K., Moss, H. E., 2001. Towards a distributed account of conceptual knowledge. *Trends*
941 *Cogn. Sci.* 5, 244-262. doi:[10.1016/S1364-6613\(00\)01651-X](https://doi.org/10.1016/S1364-6613(00)01651-X).
- 942 VanRullen, R., 2007. The power of the feed-forward sweep. *Adv. Cogn. Psychol.* 3, 167–176.
943 doi:[10.2478/v10053-008-0022-3](https://doi.org/10.2478/v10053-008-0022-3).
- 944 VanRullen, R., Thorpe, S. J., 2001. The time course of visual processing: From early perception
945 to decision-making. *J. Cogn. Neurosci.* 13, 454-461. doi:[10.1162/08989290152001880](https://doi.org/10.1162/08989290152001880).
- 946 van de Nieuwenhuizen, M. E., Backus, A. R., Bahramisharif, A., Doeller, C. F., Jensen, O., van
947 Gerven, M. A. J., 2013. MEG-based decoding of the spatiotemporal dynamics of visual
948 category perception. *NeuroImage*. doi:<https://doi.org/10.1016/j.neuroimage.2013.07.075>
- 949 Vedaldi, A., Lenc, K., 2015. MatConvNet – convolutional neural networks for MATLAB. *ACM Int.*
950 *Conf. on Multimedia*.
- 951 Wardle, S. G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S-M., Carlson, T. A., 2016.
952 Perceptual similarity of visual patterns predicts dynamic neural activation patterns
953 measured with MEG. *NeuroImage*. doi:<https://doi.org/10.1016/j.neuroimage.2016.02.019>
- 954 Wen, H., Shi, J., Zhang, Y., Lu, K.-H., Cao, J., Liu, Z., 2017. Neural encoding and decoding with
955 deep learning for dynamic natural vision. *Cerebral Cortex*.
956 <https://doi.org/10.1093/cercor/bhx268>
- 957 Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., DiCarlo, J. J., 2014.
958 Performance-optimized hierarchical models predict neural responses in higher visual
959 cortex. *Proc. Natl. Acad. Sci. USA* 111, 8619:8624.