

1 A content-sensitive framework for modeling episodic
2 memory reveals event-like structure in naturalistic
3 experience, recall, and neural processing

4 Andrew C. Heusser, Paxton C. Fitzpatrick, and Jeremy R. Manning

5 Department of Psychological and Brain Sciences

6 Dartmouth College, Hanover, NH 03755, USA

7 Corresponding author: jeremy.r.manning@dartmouth.edu

8 August 29, 2019

9 **Abstract**

10 Our life experiences unfold over time in highly complex manner, with the evolving presence
11 of numerous intricate features describing our journey between each event we encounter. Here,
12 we propose a framework for mapping dynamic naturalistic episodes onto geometric spaces as
13 *experience trajectories* that capture the temporal dynamics of real-world content. Within this
14 geometric framework, one may compare the shape of the trajectory formed by an experience to
15 that defined by one's later recollection to characterize our memories' recovery and distortion of
16 the external world. Here, we apply this approach to a naturalistic memory experiment in which
17 participants viewed and verbally recounted a video, and find that the video and subsequent
18 recalls share both an experience-specific shape and a discernible event-like structure. Despite
19 this apparent similarity, we find that the level of *precision* with which individuals recounted
various events and the *distinctiveness* of recall for those events were varied and predictive of
overall memory performance. Finally, we identify a network of brain structures that is sensitive
to the "shapes" of our ongoing experiences, and an overlapping network sensitive to how we will

20 later remember them. Our framework and findings underscore the rich event-like structure of
21 the external world and our memories, and offer novel, content-sensitive measures for assessing
22 episodic memory

23 **Introduction**

24 What does it mean to *remember* something? In traditional episodic memory experiments (e.g.,
25 list-learning or trial-based experiments; Murdock, 1962; Kahana, 1996), remembering is often cast
26 as a discrete and binary operation: each studied item may be separated from all others, and la-
beled as having been recalled or forgotten. More nuanced studies might incorporate self-reported
27 confidence measures as a proxy for memory strength, or ask participants to discriminate between
28 “recollecting” the (contextual) details of an experience or having a general feeling of “familiarity”
29 (Yonelinas, 2002). Using well-controlled, trial-based experimental designs, the field has amassed
30 a wealth of valuable information regarding human episodic memory. However, there are funda-
31 mental properties of the external world and our memories that trial-based experiments are not well
32 suited to capture (for review also see Koriat and Goldsmith, 1994; Huk et al., 2018). First, our expe-
33 riences and memories are continuous, rather than discrete—removing a (naturalistic) event from
34 the context in which it occurs can substantially change its meaning. Second, the specific language
35 used to describe an experience has little bearing on whether the experience should be considered to
36 have been “remembered.” Asking whether the rememberer has precisely reproduced a specific set
37 of words to describe a given experience is nearly orthogonal to whether they were actually able to
38 remember it. In classic (e.g., list-learning) memory studies, by contrast, the number or proportion
39 of precise recalls is often a primary metric for assessing the quality of participants’ memories.
40 Third, one might remember the *essence* (or a general summary) of an experience but forget (or
41 neglect to recount) particular details. Capturing the essence of what happened is typically the
42 main “point” of recounting a memory to a listener, while the addition of highly specific details
43 may add comparatively little to successful conveyance of an experience.

45 How might one go about formally characterizing the “essence” of an experience, or whether

46 it has been recovered by the rememberer? Any given moment of an experience derives meaning
47 from surrounding moments, as well as from longer-range temporal associations (Lerner et al.,
48 2011; Manning, 2019). Therefore, the timecourse describing how an event unfolds is fundamental
49 to its overall meaning. Further, this hierarchy formed by our subjective experiences at different
50 timescales defines a *context* for each new moment (e.g., Howard and Kahana, 2002; Howard et al.,
51 2014), and plays an important role in how we interpret that moment and remember it later (for
52 review see Manning et al., 2015). Our memory systems can leverage these associations to form
53 predictions that help guide our behaviors (Ranganath and Ritchey, 2012). For example, as we
54 navigate the world, the features of our subjective experiences tend to change gradually (e.g., the
55 room or situation we are in at any given moment is strongly temporally autocorrelated), allowing
56 us to form stable estimates of our current situation and behave accordingly (Zacks et al., 2007;
57 Zwaan and Radvansky, 1998).

58 Although our experiences most often change gradually, they also occasionally change sud-
59 denly (e.g., when we walk through a doorway; Radvansky and Zacks, 2017). Prior research
60 suggests that these sharp transitions (termed *event boundaries*) during an experience help to dis-
61 cretize our experiences (and their mental representations) into *events* (Radvansky and Zacks, 2017;
62 Brunec et al., 2018; Heusser et al., 2018a; Clewett and Davachi, 2017; Ezzyat and Davachi, 2011;
63 DuBrow and Davachi, 2013). The interplay between the stable (within event) and transient (across
64 event) temporal dynamics of an experience also provides a potential framework for transforming
65 experiences into memories that distill those experiences down to their essence. For example, prior
66 work has shown that event boundaries can influence how we learn sequences of items (Heusser
67 et al., 2018a; DuBrow and Davachi, 2013), navigate (Brunec et al., 2018), and remember and un-
68 derstand narratives (Zwaan and Radvansky, 1998; Ezzyat and Davachi, 2011). Prior research has
69 implicated the hippocampus and the medial prefrontal cortex as playing a critical role in trans-
70 forming experiences into structured and consolidated memories (Tompry and Davachi, 2017).

71 Here we sought to examine how the temporal dynamics of a “naturalistic” experience were
72 later reflected in participants’ memories. We analyzed an open dataset that comprised behavioral
73 and functional Magnetic Resonance Imaging (fMRI) data collected as participants viewed and then

74 verbally recounted an episode of the BBC television series *Sherlock* (Chen et al., 2017). We developed
75 a computational framework for characterizing the temporal dynamics of the moment-by-moment
76 content of the episode and of participants' verbal recalls. Specifically, we use topic modeling (Blei
77 et al., 2003) to characterize the thematic conceptual (semantic) content present in each moment of
78 the episode and recalls, and Hidden Markov Models (Rabiner, 1989; Baldassano et al., 2017) to
79 discretize this evolving semantic content into events. In this way, we cast naturalistic experiences
80 (and recalls of those experiences) as *trajectories* that describe how the experiences evolve over
81 time. Under this framework, successful remembering entails verbally "traversing" the content
82 trajectory of the episode, thereby reproducing the shape (or essence) of the original experience.
83 Comparing the shapes of the topic trajectories of the episode and of participants' retellings of the
84 episode then reveals which aspects of the episode were preserved (or lost) in the translation into
85 memory. We further examine whether 1) the *precision* with which a participant recounts each event
86 and 2) the *distinctiveness* each recall event is (relative to the other recalled events) relates to their
87 overall memory performance. Last, we identify networks of brain structures whose responses
88 (as participants watched the episode) reflected the temporal dynamics of the episode, and how
89 participants would later recount the episode.

90 Results

91 To characterize the shape of the *Sherlock* episode and participants' subsequent recounts of its
92 unfolding, we used a topic model (Blei et al., 2003) to discover the latent themes in the episode's
93 dynamic content. Topic models take as inputs a vocabulary of words to consider and a collection
94 of text documents, and return two output matrices. The first of these is a *topics matrix* whose rows
95 are topics (latent themes) and whose columns correspond to words in the vocabulary. The entries
96 of the topics matrix define how each word in the vocabulary is weighted by each discovered topic.
97 For example, a detective-themed topic might weight heavily on words like "crime," and "search."
98 The second output is a *topic proportions matrix*, with one row per document and one column per
99 topic. The topic proportions matrix describes what mixture of discovered topics is reflected in each

100 document.

101 Chen et al. (2017) collected hand-annotated information about each of 1000 (manually identified)
102 scenes spanning the roughly 50 minute video used in their experiment. This information included:
103 a brief narrative description of what was happening; whether the scene took place indoors or
104 outdoors; the names of any characters on the screen; the names of any characters who were in
105 focus in the camera shot; the names of characters who were speaking; the location where the scene
106 took place; the camera angle (close up, medium, long, etc.); whether or not background music was
107 present; and other similar details (for a full list of annotated features see *Methods*). We took from
108 these annotations the union of all unique words (excluding stop words, such as “and,” “or,” “but,”
109 etc.) across all features and scenes as the “vocabulary” for the topic model. We then concatenated
110 the sets of words across all features contained in overlapping, 50-scene sliding windows, and
111 treated each 50-scene sequence as a single “document” for the purpose of fitting the topic model.
112 Next, we fit a topic model with (up to) $K = 100$ topics to this collection of documents. We found that
113 27 unique topics (with non-zero weights) were sufficient to describe the time-varying content of the
114 video (see *Methods*; Figs. 1, S2). Note that our approach is similar in some respects to Dynamic Topic
115 Models (Blei and Lafferty, 2006) in that we sought to characterize how the thematic content of the
116 episode evolved over time. However, whereas Dynamic Topic Models are designed to characterize
117 how the properties of *collections* of documents change over time, our sliding window approach
118 allows us to examine the topic dynamics within a single document (or video). Specifically, our
119 approach yielded (via the topic proportions matrix) a single *topic vector* for each timepoint of the
120 episode (we set timepoints to match the acquisition times of the 1976 fMRI volumes collected as
121 participants viewed the episode).

122 The topics we found were heavily character-focused (e.g., the top-weighted word in each topic
123 was nearly always a character) and could be roughly divided into themes that were primarily
124 Sherlock Holmes-focused (Sherlock is the titular character), primarily John Watson-focused (John
125 is Sherlock’s close confidant and assistant), or focused on Sherlock and John interacting (Fig. S2).
126 Several of the topics were highly similar, which we hypothesized might allow us to distinguish
127 between subtle narrative differences (if the distinctions between those overlapping topics were



Figure 1: Methods overview. We used hand-annotated descriptions of each moment of video to fit a topic model. Three example video frames and their associated descriptions are displayed (top two rows). Participants later recalled the video (in the third row, we show example recalls of the same three scenes from participant 13). We used the topic model (fit to the annotations) to estimate topic vectors for each moment of video and each sentence the participants recalled. Example topic vectors are displayed in the bottom row (blue: video annotations; green: example participant’s recalls). Three topic dimensions are shown (the highest-weighted topics for each of the three example scenes, respectively). We also show the ten highest-weighted words for each topic. Figure S2 provides a full list of the top 10 words from each of the discovered topics.

128 meaningful; also see Fig. S3). The topic vectors for each timepoint were *sparse*, in that only a small
129 number (usually one or two) of topics tended to be “active” in any given timepoint (Fig. 2A).
130 Further, the dynamics of the topic activations appeared to exhibit *persistance* (i.e., given that a
131 topic was active in one timepoint, it was likely to be active in the following timepoint) along with
132 *occasional rapid changes* (i.e., occasionally topics would appear to spring into or out of existence).
133 These two properties of the topic dynamics may be seen in the block diagonal structure of the
134 timepoint-by-timepoint correlation matrix (Fig. 2B) and reflect the gradual drift and sudden shifts
135 fundamental to the contextual dynamics of real-world experiences. Given this observation, we
136 adapted an approach devised by Baldassano et al. (2017), and used a Hidden Markov Model (HMM)
137 to identify the *event boundaries* where the topic activations changed rapidly (i.e., at the boundaries
138 of the blocks in the correlation matrix; event boundaries identified by the HMM are outlined in
139 yellow). Part of our model fitting procedure required selecting an appropriate number of “events”
140 to segment the timeseries into. We used an optimization procedure to identify the number of
141 events that maximized within-event stability while also minimizing across-event correlations (see
142 *Methods* for additional details). To create a stable “summary” of the video, we computed the
143 average topic vector within each event (Fig. 2C).

144 Given that the time-varying content of the video could be segmented cleanly into discrete
145 events, we wondered whether participants’ recalls of the video also displayed a similar structure.
146 We applied the same topic model (already trained on the video annotations) to each participant’s
147 recalls. Analogous to how we analyzed the time-varying content of the video, to obtain similar
148 estimates for participants’ recalls, we treated each (overlapping) 10-sentence “window” of their
149 transcript as a “document” and then computed the most probable mix of topics reflected in each
150 timepoint’s sentences. This yielded, for each participant, a number-of-sentences by number-of-
151 topics topic proportions matrix that characterized how the topics identified in the original video
152 were reflected in the participant’s recalls. Note that an important feature of our approach is
153 that it allows us to compare participant’s recalls to events from the original video, despite that
154 different participants may have used different language to describe the same event, and that those
155 descriptions may not match the original annotations. This is a substantial benefit of projecting

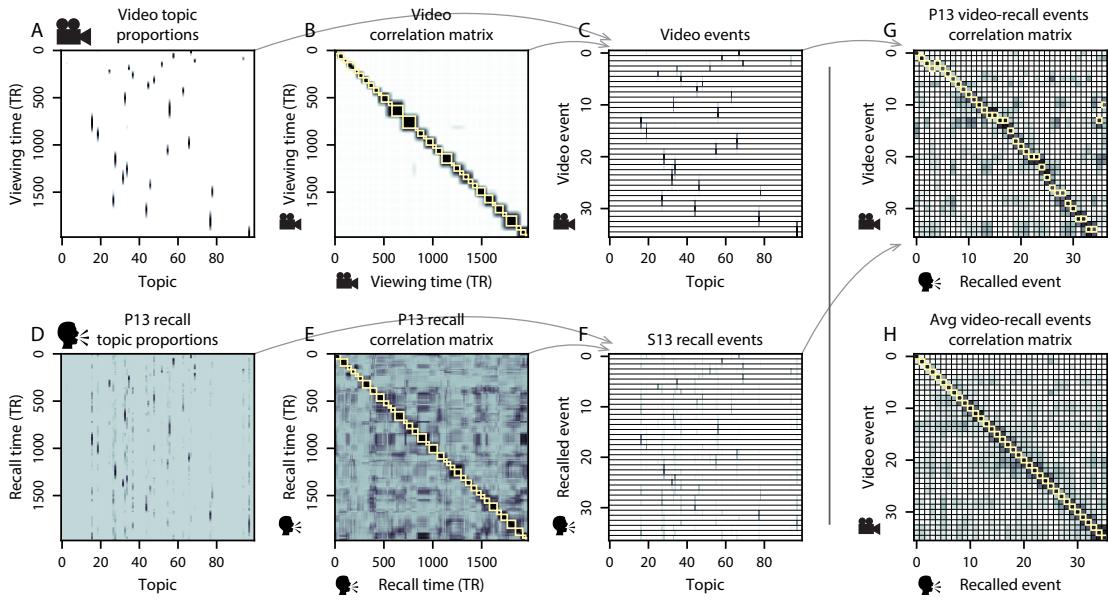


Figure 2: Modelling naturalistic stimuli and recalls. All panels: darker colors indicate greater values; range: [0, 1]. **A.** Topic vectors ($K = 100$) for each of the 1976 video timepoints. **B.** Timepoint-by-timepoint correlation matrix of the topic vectors displayed in Panel A. Event boundaries detected by the HMM are denoted in yellow (36 events detected). **C.** Average topic vectors for each of the 36 video events. **D.** Topic vectors for each of 294 sentences spoken by an example participant while recalling the video, resampled to match the timeseries of the video. **E.** Timepoint-by-timepoint correlation matrix of the topic vectors displayed in Panel D. Event boundaries detected by the HMM are denoted in yellow (37 events detected). **F.** Average topic vectors for each of the 37 recalled events from the example participant. **G.** Correlations between the topic vectors for every pair of video events (Panel C) and recalled events (from the example participant; Panel F). For similar plots for all participants see Figure S5. **H.** Average correlations between each pair of video events and recalled events (across all 17 participants). To create the figure, each recalled event was assigned to the video event with the most correlated topic vector (yellow boxes in panels G and H). The heat maps in each panel were created using Seaborn (Waskom et al., 2016).

156 the video and recalls into a shared “topic” space. An example topic proportions matrix from one
157 participant’s recalls is shown in Figure 2D.

158 Although the example participant’s recall topic proportions matrix has some visual similarity to
159 the video topic proportions matrix, the time-varying topic proportions for the example participant’s
160 recalls are not as sparse as for the video (e.g., compare Figs. 2A and D). Similarly, although there
161 do appear to be periods of stability in the recall topic dynamics (e.g., most topics are active or
162 inactive over contiguous blocks of time), the overall timecourses are not as cleanly delineated as
163 the video topics are. To examine these patterns in detail, we computed the timepoint-by-timepoint
164 correlation matrix for the example participant’s recall topic proportions (Fig. 2E). As in the video
165 correlation matrix (Fig. 2B), the example participant’s recall correlation matrix has a strong block
166 diagonal structure, indicating that their recalls are discretized into separated events. As for the
167 video correlation matrix, we can use an HMM, along with the aforementioned number-of-events
168 optimization procedure (also see *Methods*) to determine how many events are reflected in the
169 participant’s recalls and where specifically the event boundaries fall (outlined in yellow). We
170 carried out a similar analysis on all 17 participants’ recall topic proportions matrices (Fig. S4).

171 Two clear patterns emerged from this set of analyses. First, although every individual partic-
172 ipant’s recalls could be segmented into discrete events (i.e., every individual participant’s recall
173 correlation matrix exhibited clear block diagonal structure; Fig. S4), each participant appeared to
174 have a unique *recall resolution*, reflected in the sizes of those blocks. For example, some participants’
175 recall topic proportions segmented into just a few events (e.g., Participants P2, P6, and P14), while
176 others’ recalls segmented into many shorter duration events (e.g., Participants P12, P13, and P17).
177 This suggests that different participants may be recalling the video with different levels of detail-
178 e.g., some might touch on just the major plot points, whereas others might attempt to recall every
179 minor scene or action. The second clear pattern present in every individual participant’s recall
180 correlation matrix is that, unlike in the video correlation matrix, there are substantial off-diagonal
181 correlations. Whereas each event in the original video was (largely) separable from the others
182 (Fig. 2B), in transforming those separable events into memory, participants appear to be integrat-
183 ing across multiple events, blending elements of previously recalled and not-yet-recalled events

184 into each newly recalled event (Figs. 2D, S4; also see Manning et al., 2011; Howard et al., 2012).

185 The above results indicate that both the structure of the original video and participants' recalls
186 of the video exhibit event boundaries that can be identified automatically by characterizing the
187 dynamic content using a shared topic model and segmenting the content into events using HMMs.
188 Next, we asked whether some correspondence might be made between the specific content of the
189 events the participants experienced in the video, and the events they later recalled. One approach
190 to linking the experienced (video) and recalled events is to label each recalled event as matching
191 the video event with the most similar (i.e., most highly correlated) topic vector (Figs. 2G, S5). This
192 yields a sequence of "presented" events from the original video, and a (potentially differently
193 ordered) sequence of "recalled" events for each participant. Analogous to classic list-learning
194 studies, we can then examine participants' recall sequences by asking which events they tended
195 to recall first (probability of first recall; Fig. 3A; Welch and Burnett, 1924; Postman and Phillips,
196 1965; Atkinson and Shiffrin, 1968); how participants most often transition between recalls of the
197 events as a function of the temporal distance between them (lag-conditional response probability;
198 Fig. 3B; Kahana, 1996); and which events they were likely to remember overall (serial position
199 recall analyses; Fig. 3C; Murdock, 1962). Interestingly, for two of these analyses (probability of first
200 recall and lag-conditional response probability curves) we observe patterns comparable to classic
201 effects from the list-learning literature: namely, a higher probability of initiating recall with the
202 first event in the sequence (Fig. 3A) and a higher probability of transitioning to neighboring events
203 with an asymmetric forward bias (Fig. 3C). In contrast, we do not observe a pattern comparable to
204 the serial position effect (Fig. 3C), but rather we see higher memory for specific events distributed
205 somewhat evenly throughout the video.

206 We can also apply two list-learning-native analyses that describe how participants group items
207 in their recall sequences: temporal clustering and semantic clustering (Polyn et al., 2009, see
208 *Methods* for details). Temporal clustering refers to the extent to which participants group their
209 recall responses according to encoding position. Semantic clustering measures the extent to which
210 participants cluster their recall responses according to semantic similarity. Overall, we found that
participants heavily clustered video events in their recalls by both temporal proximity (mean:

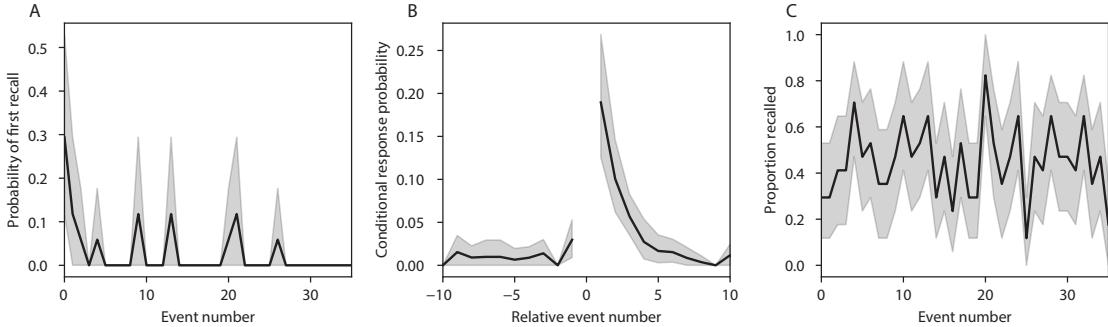


Figure 3: Naturalistic extensions of classic list-learning memory analyses. **A.** The probability of first recall as a function of the serial position of the event in the video. **B.** The probability of recalling each event, conditioned on having most recently recalled the event *lag* events away in the video. **C.** The proportion of participants who recalled each event, as a function of the serial position of the events in the video. All panels: error bars denote bootstrap-estimated standard error of the mean.

212 0.860, SEM: 0.016) and semantic content (mean: 0.888, SEM: 0.014).

213 Statistical models of memory studies often treat memory recalls as binary (e.g. the item was re-
 214 called or not) and independent events. However, our framework produces a content-based model
 215 of individual stimulus and recall events, allowing for direct quantitative comparison between all
 216 stimulus and recall events, as well as between the recall events themselves. Leveraging these
 217 content-based models of the stimulus/recall events, we developed two novel metrics for quanti-
 218 fying naturalistic memory representations: *precision* and *distinctiveness*. We define precision as the
 219 average correlation between the topic proportions of each recall event and the maximally corre-
 220 lated video event (Fig. 4). Participants whose recall events are more veridical descriptions of what
 221 happened in the video event will presumably have higher precision scores. We find that, across
 222 participants, a higher precision score is correlated to both hand-annotated memory performance
 223 (Pearson's $r(15) = 0.55, p = 0.021$) and the number of recall events estimated by our model (Pear-
 224 son's $r(15) = 0.66, p = 0.004$). A second novel metric we introduce here is distinctiveness, or how
 225 unique the recall description was to each video event. We define distinctiveness as 1 minus the av-
 226 erage of all non-matching recall events from the video-recall correlation matrix. We hypothesized
 227 that participants who recounted events in a more distinctive way would display better overall
 228 memory. Similarly to precision, we find that the more distinct participants recalls are (on average),

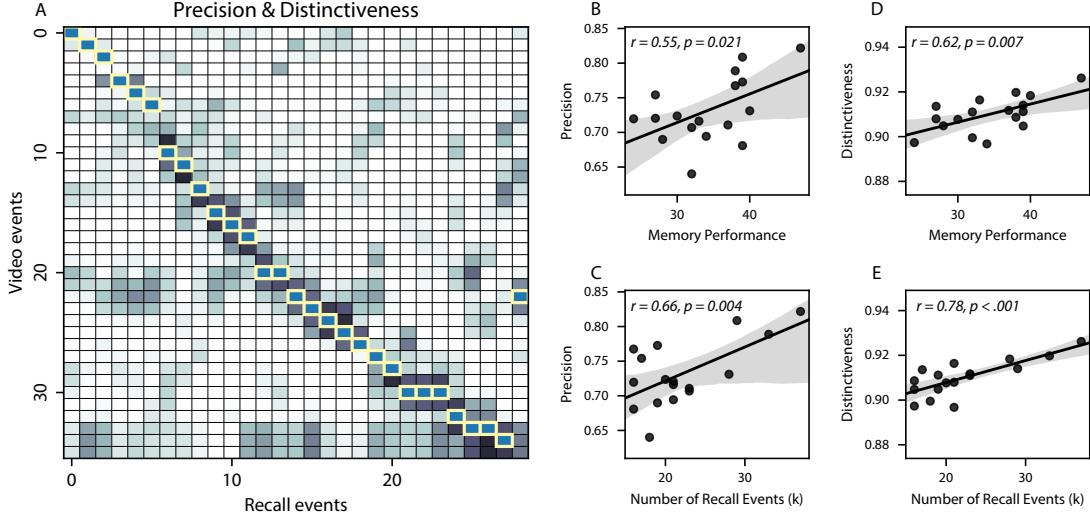


Figure 4: Novel content-based metrics of naturalistic memory: precision and distinctiveness. **A.** A video-recall correlation matrix for a representative participant (17). The yellow boxes highlight the maximum correlation in each column. Precision was computed as the average of the maximum correlation in each column. On the other hand, distinctiveness was defined as the average of everything except for the maximum correlation in each column. **B.** The (Pearson's) correlation between precision and hand-annotated memory performance. **C.** The correlation between precision and the number of events recovered by the model (k). **D.** The correlation between distinctiveness and hand-annotated memory performance. **E.** The correlation between distinctiveness and the number of events recovered by the model (k).

the more they remembered (hand-annotated memory: Pearson's $r(15) = 0.62, p = 0.007$; number of events: Pearson's $r(15) = 0.78, p < 0.001$). In summary, using two novel metrics afforded by our approach, we find that participants whose recalls are both more precise and distinct remember more content.

The prior analyses leverage the correspondence between the 100-dimensional topic proportion matrices for the video and participants' recalls to characterize recall. However, it is difficult to gain deep insights into that content solely by examining the topic proportion matrices (e.g., Figs. 2A, D) or the corresponding correlation matrices (Figs. 2B, E, S4). To visualize the time-varying high-dimensional content in a more intuitive way (Heusser et al., 2018b) we projected the topic proportions matrices onto a two-dimensional space using Uniform Manifold Approximation and Projection (UMAP; McInnes and Healy, 2018). In this lower-dimensional space, each point

240 represents a single video or recall event, and the distances between the points reflect the distances
241 between the events' associated topic vectors (Fig. 5). In other words, events that are near to each
242 other in this space are more semantically similar.

243 Visual inspection of the video and recall topic trajectories reveals a striking pattern. First,
244 the topic trajectory of the video (which reflects its dynamic content; Fig. 5A) is captured nearly
245 perfectly by the averaged topic trajectories of participants' recalls (Fig. 5B). To assess the consistency
246 of these recall trajectories across participants, we asked: given that a participant's recall trajectory
247 had entered a particular location in topic space, could the position of their *next* recalled event
248 be predicted reliably? For each location in topic space, we computed the set of line segments
249 connecting successively recalled events (across all participants) that intersected that location (see
250 *Methods* for additional details). We then computed (for each location) the distribution of angles
251 formed by the lines defined by those line segments and a fixed reference line (the *x*-axis). Rayleigh
252 tests revealed the set of locations in topic space at which these across-participant distributions
253 exhibited reliable peaks (blue arrows in Fig. 5B reflect significant peaks at $p < 0.05$, corrected). We
254 observed that the locations traversed by nearly the entire video trajectory exhibited such peaks.
255 In other words, participants exhibited similar trajectories that also matched the trajectory of the
256 original video (Fig. 5C). This is especially notable when considering the fact that the number of
257 events participants recalled (dots in Fig. 5C) varied considerably across people, and that every
258 participant used different words to describe what they had remembered happening in the video.
259 Differences in the numbers of remembered events appear in participants' trajectories as differences
260 in the sampling resolution along the trajectory. We note that this framework also provides a
261 means of detangling classic "proportion recalled" measures (i.e., the proportion of video events
262 referenced in participants' recalls) from participants' abilities to recapitulate the full shape of the
263 original video (i.e., the similarity in the shape of the original video trajectory and that defined by
264 each participant's recounting of the video).

265 Because our analysis framework projects the dynamic video content and participants' recalls
266 onto a shared topic space, and because the dimensions of that space are known (i.e., each topic
267 dimension is a set of weights over words in the vocabulary; Fig. S2), we can examine the topic

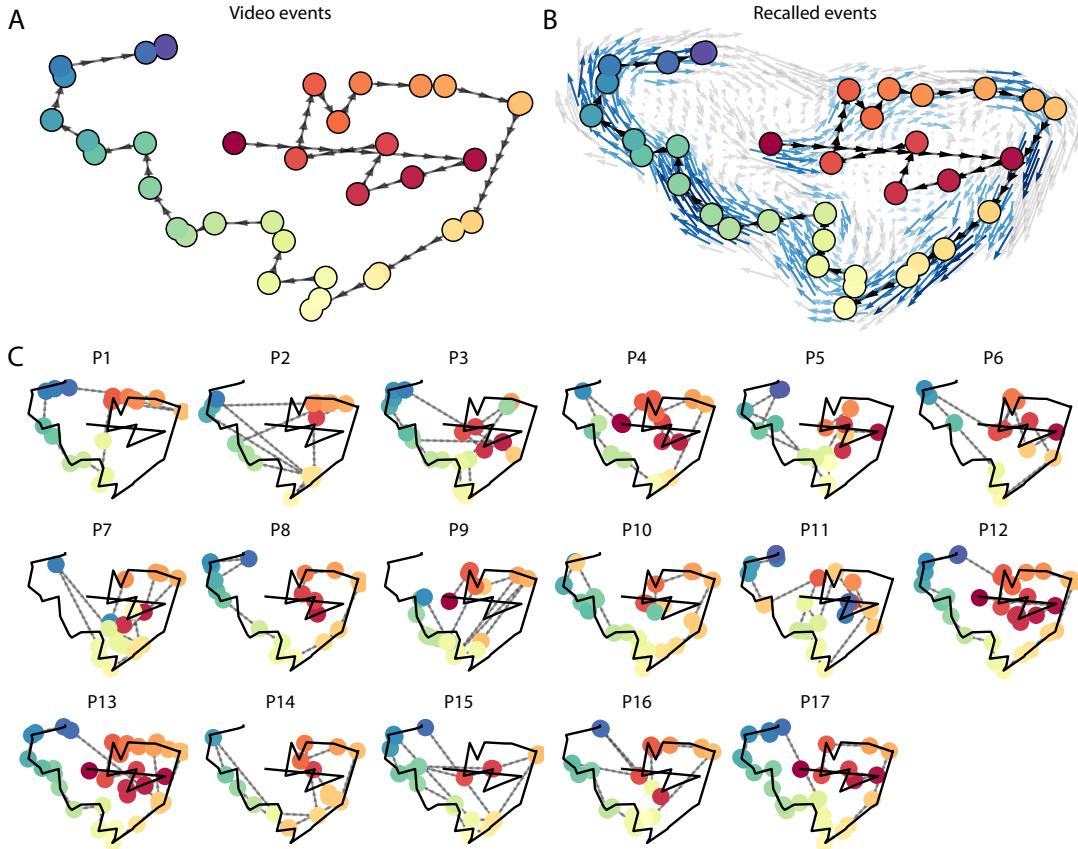


Figure 5: Trajectories through topic space capture the dynamic content of the video and recalls. All panels: the topic proportion matrices have been projected onto a shared two-dimensional space using UMAP. **A.** The two-dimensional topic trajectory taken by the episode of *Sherlock*. Each dot indicates an event identified using the HMM (see *Methods*); the dot colors denote the order of the events (early events are in red; later events are in blue), and the connecting lines indicate the transitions between successive events. **B.** The average two-dimensional trajectory captured by participants' recall sequences, with the same format and coloring as the trajectory in Panel A. To compute the event positions, we matched each recalled event with an event from the original video (see *Results*), and then we averaged the positions of all events with the same label. The arrows reflect the average transition direction through topic space taken by any participants whose trajectories crossed that part of topic space; blue denotes reliable agreement across participants via a Rayleigh test ($p < 0.05$, corrected). **C.** The recall topic trajectories (gray) taken by each individual participant (P1–P17). The video's trajectory is shown in black for reference. (Same format and coloring as Panel A.)

268 trajectories to understand which specific content was remembered well (or poorly). For each video
269 event, we can ask: what was the average correlation (across participants) between the video event's
270 topic vector and the closest matching recall event topic vectors from each participant? This yields a
271 single correlation coefficient for each video event, describing how closely participants' recalls of the
272 event tended to reliably capture its content (Fig. 6A). (We also examined how different comparisons
273 between each video event's topic vector and the corresponding recall event topic vectors related
274 to hand-annotated characterizations of memory performance; see *Supporting Information*). Given
275 this summary of which events were recalled reliably (or not), we next asked whether the better-
276 remembered or worse-remembered events tended to reflect particular topics. We computed a
277 weighted average of the topic vectors for each video event, where the weights reflected how reliably
278 each event was recalled. To visualize the result, we created a “wordle” image (Mueller et al., 2018)
279 where words weighted more heavily by better-remembered topics appear in a larger font (Fig. 6B,
280 green box). Events that reflected topics weighting heavily on characters like “Sherlock” and “John”
281 (the main characters) and locations like “221b Baker Street” (a major recurring location and the
282 address of the flat that Sherlock and John share) were best remembered. An analogous analysis
283 revealed which themes were poorly remembered. Here in computing the weighted average over
284 events' topic vectors, we weighted each event in *inverse* proportion to how well it was remembered
285 (Fig. 6B, red box). This revealed that events with relatively minor characters such as “Mike,”
286 “Jeffrey,” and “Molly,” as well as less-integral plot locations (e.g., “hospital” and “office”) were
287 least well-remembered. This suggests that what is retained in memory are the major plot elements
288 (i.e., the overall shape of what happened), whereas the more minor details are prone to pruning.

289 In addition to constructing overall summaries, assessing the video and recall topic vectors
290 from individual events can provide further insights. Specifically, for any given event we can
291 construct two wordles: one from the original video event's topic vector, and a second from the
292 average topic vectors produced by all participants' recalls of that event. We can then examine those
293 wordles visually to gain an intuition for which aspects of the video event were recapitulated in
294 participants' recalls. Several example wordles are displayed in Figure 6C (wordles from the three
295 best-remembered events are circled in green; wordles from the three worst-remembered events

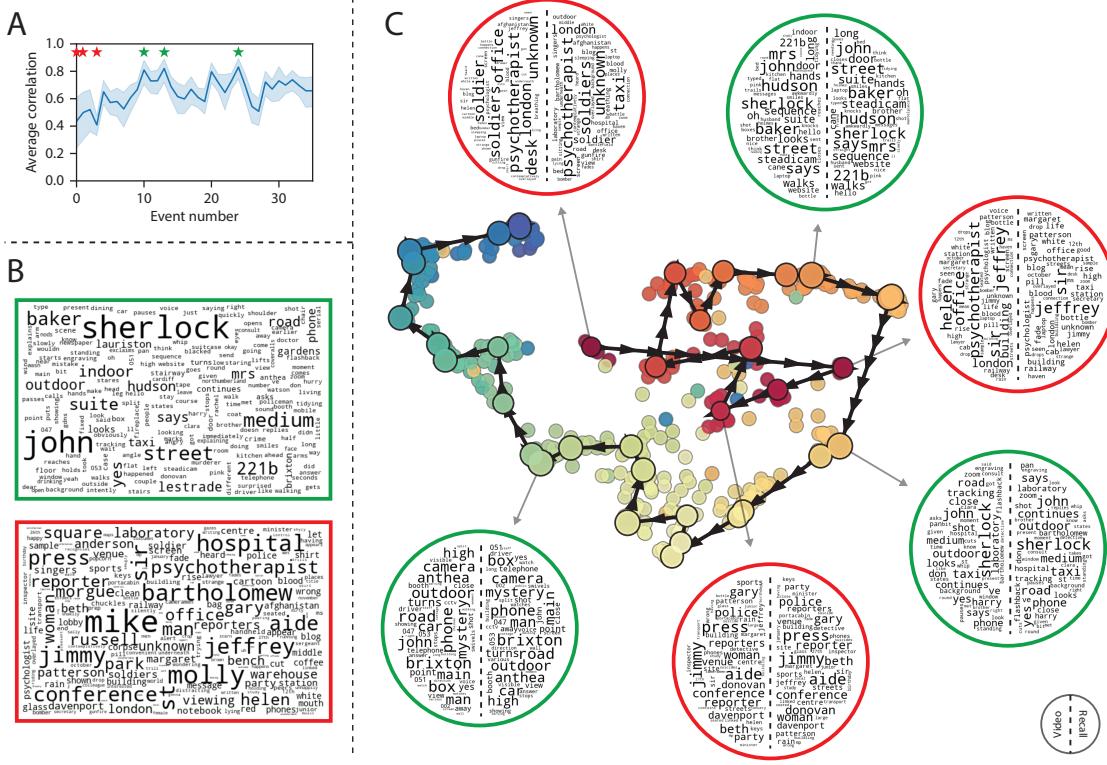


Figure 6: Transforming experience into memory. **A.** Average correlations (across participants) between the topic vectors from each video event and the closest-matching recall events. Error bars denote bootstrap-derived across-participant 95% confidence intervals. The stars denote the three best-remembered events (green) and worst-remembered events (red). **B.** Wordles comprising the top 200 highest-weighted words reflected in the weighted-average topic vector across video events. Green: video events were weighted by how well the topic vectors derived from recalls of those events matched the video events' topic vectors (Panel A). Red: video events were weighted by the inverse of how well their topic vectors matched the recalled topic vectors. **C.** The set of all video and recall events is projected onto the two-dimensional space derived in Figure 5. The dots outlined in black denote video events (dot size reflects the average correlation between the video event's topic vector and the topic vectors from the closest matching recalled events from each participant; bigger dots denote stronger correlations). The dots without black outlines denote recalled events. All dots are colored using the same scheme as Figure 5A. Wordles for several example events are displayed (green: three best-remembered events; red: three worst-remembered events). Within each circular wordle, the left side displays words associated with the topic vector for the video event, and the right side displays words associated with the (average) recall event topic vector, across all recall events matched to the given video event.

296 are circled in red). Using wordles to visually compare the topical content of each video event and
297 the (average) corresponding recall event reveals the specific content from the specific events that
298 is reliably retained in the transformation into memory (green events) or not (red events).

299 The results thus far inform us about which aspects of the dynamic content in the episode
300 participants watched were preserved or altered in participants' memories of the episode. We next
301 carried out a series of analyses aimed at understanding which brain structures might implement
302 these processes. In one analysis we sought to identify which brain structures were sensitive
303 to the video's dynamic content, as characterized by its topic trajectory. Specifically, we used a
304 searchlight procedure to identify the extent to which each cluster of voxels exhibited a timecourse
305 of activity (as the participants watched the video) whose temporal correlation matrix matched
306 the temporal correlation matrix of the original video's topic proportions (Fig. 2B). As shown
307 in Figure 7A, the analysis revealed a network of regions including bilateral frontal cortex and
308 cingulate cortex, suggesting that these regions may play a role in processing information relevant
309 to the narrative structure of the video. In a second analysis, we sought to identify which brain
310 structures' responses (while viewing the video) reflected how each participant would later *recall*
311 the video. We used an analogous searchlight procedure to identify clusters of voxels whose
312 temporal correlation matrices reflected the temporal correlation matrix of the topic proportions for
313 each individual's recalls (Figs. 2D, S4). As shown in Figure 7B, the analysis revealed a network of
314 regions including the ventromedial prefrontal cortex (vmPFC), anterior cingulate cortex (ACC), and
315 right medial temporal lobe (rMTL), suggesting that these regions may play a role in transforming
316 each individual's experience into memory. In identifying regions whose responses to ongoing
317 experiences reflect how those experiences will be remembered later, this latter analysis extends
318 classic *subsequent memory analyses* (e.g., Paller and Wagner, 2002) to domain of naturalistic stimuli.

319 Discussion

320 Our work casts remembering as reproducing (behaviorally and neurally) the topic trajectory, or
321 shape, of an experience. This view draws inspiration from prior work aimed at elucidating

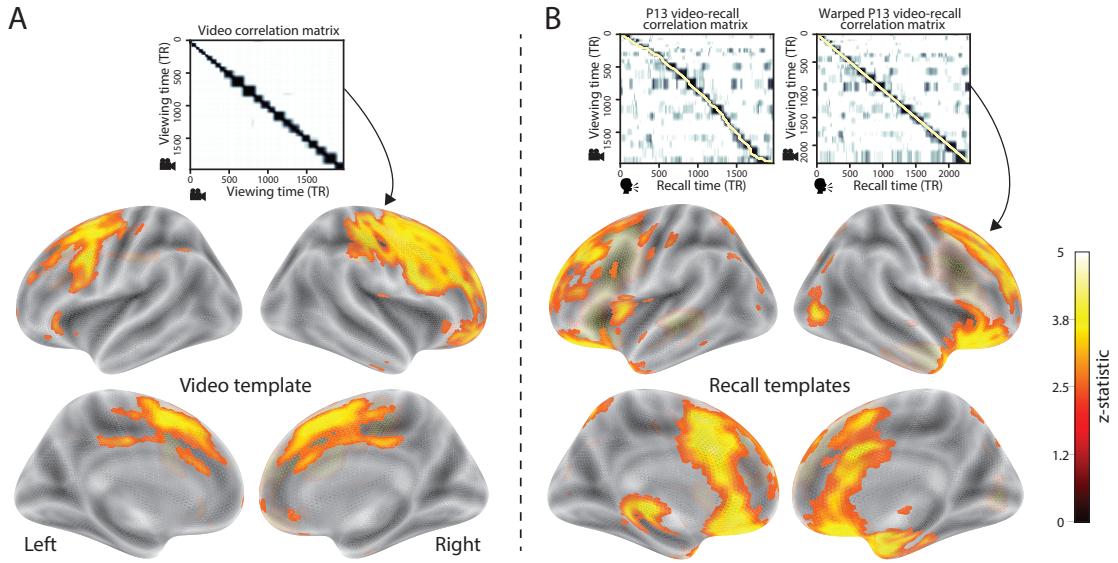


Figure 7: Brain structures that underlie the transformation of experience into memory. **A.** We searched for regions whose responses (as participants watched the video) matched the temporal correlation matrix of the video topic proportions. These regions are sensitive to the narrative structure of the video. **B.** We searched for regions whose responses (as participants watched the video) matched the temporal correlation matrix of the topic proportions derived from each individual's later recall of video. These regions are sensitive to how the narrative structure of the video is transformed into a memory of the video. Both panels: the maps are thresholded at $p < 0.05$, corrected.

the neural and behavioral underpinnings of how we process dynamic naturalistic experiences and remember them later. One approach to identifying neural responses to naturalistic stimuli (including experiences) entails building a model of the stimulus and searching for brain regions whose responses are consistent with the model. In prior work, a series of studies from Uri Hasson’s group (Lerner et al., 2011; Simony et al., 2016; Chen et al., 2017; Baldassano et al., 2017; Zadbood et al., 2017) have extended this approach with a clever twist: rather than building an explicit stimulus model, these studies instead search for brain responses (while experiencing the stimulus) that are reliably similar across individuals. So called *inter-subject correlation* (ISC) and *inter-subject functional connectivity* (ISFC) analyses effectively treat other people’s brain responses to the stimulus as a “model” of how its features change over time. By contrast, in our present work we used topic models and HMMs to construct an explicit stimulus model (i.e., the topic trajectory of the video). When we searched for brain structures whose responses are consistent with the video’s topic trajectory, we identified a network of structures that overlapped strongly with the “long temporal receptive window” network reported by the Hasson group (e.g., compare our Fig. 7A with the map of long temporal receptive window voxels in Lerner et al., 2011). This provides support for the notion that part of the long temporal receptive window network may be maintaining an explicit model of the stimulus dynamics. When we performed a similar analysis after swapping out the video’s topic trajectory with the recall topic trajectories of each individual participant, this allowed us to identify brain regions whose responses (as the participants viewed the video) reflected how the video trajectory would be transformed in memory (as reflected by the recall topic trajectories). The analysis revealed that the rMTL and vmPFC may play a role in this person-specific transformation from experience into memory. The role of the MTL in episodic memory encoding has been well-reported (e.g., Paller and Wagner, 2002; Davachi et al., 2003; Ranganath et al., 2004; Davachi, 2006; Wiltgen and Silva, 2007; Diana et al., 2007; van Kesteren et al., 2013). Prior work has also implicated the medial prefrontal cortex in representing “schema” knowledge (i.e., general knowledge about the format of an ongoing experience given prior similar experiences; van Kesteren et al., 2012, 2013; Schlichting and Preston, 2015; Gilboa and Marlatte, 2017; Spalding et al., 2018). Integrating across our study and this prior work, one interpretation is

350 that the person-specific transformations mediated (or represented) by the rMTL and vmPFC may
351 reflect schema knowledge being leveraged, formed, or updated, incorporating ongoing experience
352 into previously acquired knowledge.

353 In extending classical free recall analyses to our naturalistic memory framework, we recovered
354 two patterns of recall dynamics central to list-learning studies: a high probability of initiating
355 recall with the first video event (Fig. 3A) and a strong bias toward transitioning from recalling a
356 given event to recalling the event immediately following it (Fig. 3B). However, equally noteworthy
357 are the typical free recall results not recovered in these analyses, as each highlights a fundamental
358 difference between list-learning studies and naturalistic memory paradigms like the one employed
359 in the present study. The most noticeable departure from hallmark free recall dynamics in these
360 findings is the apparent lack of a serial position effect in Figure 3C, which instead shows greater
361 and lesser recall probabilities for events distributed across the video stimulus. Stimuli in free recall
362 experiments most often comprise lists of simple, common words, presented to participants in a
363 random order. (In fact, numerous word pools have been developed based on these criteria; e.g.,
364 Friendly et al., 1982). These stimulus qualities enable two assumptions that are central to word
365 list analyses, but frequently do not hold for real-world experiences. First, researchers conducting
366 free recall studies may assume that the content at each presentation index is essentially equal, and
367 does not bear qualities that would cause participants to remember it more or less successfully than
368 others. Such is rarely the case with real-world experiences or experiments meant to approximate
369 them, and the effects of both intrinsic and observer-dependent factors on stimulus memorability
370 are well established (for review see Chun and Turk-Browne, 2007; Bylinskii et al., 2015; Tyng
371 et al., 2017). Second, the random ordering of list items ensures that (across participants, on
372 average) there is no relationship between the thematic similarity of individual stimuli and their
373 presentation positions—in other words, two semantically related words are no more likely to be
374 presented next to each other than at opposite ends of the list. In most cases, the exact opposite
375 is true of real-world episodes. Our internal thoughts, our actions, and the physical state of the
376 world around us all tend to follow a direct, causal progression. As a result, each moment of our
377 experience tends to be inherently more similar to surrounding moments than to those in the distant

378 past or future. Memory literature has termed this strong temporal autocorrelation “context,” and
379 in various media that depict real-world events (e.g., movies and written stories), we recognize
380 it as a *narrative structure*. While a random word list (by definition) has no such structure, the
381 logical progression between ideas and actions in a naturalistic stimulus prompts the rememberer
382 to recount presented events in order, starting with the beginning. This tendency is reflected in our
383 findings’ second departure from typical free recall dynamics: a lack of increased probability of first
384 recall for end-of-sequence events (Fig. 3A).

385 Thus, analyses such as those in Figure 3 that address only the temporal dynamics of free re-
386 call paint an incomplete picture of memory for naturalistic episodes. While useful for studying
387 presentation order-dependent recall dynamics, they neglect to consider the stimulus’s content (or,
388 for example, that content’s potential interrelatedness). However, sensitivity to stimulus and recall
389 content introduces a new challenge: distinguishing between levels of recall quality for a stimulus
390 (i.e., an event) that is considered to have been “remembered.” When modeling memory experi-
391 ments, often times events (or items) and their later memories are treated as binary and independent
392 events (e.g., a given list item was simply either remembered or not remembered). Various models
393 of memory (e.g., Yonelinas, 2002) attempt to improve upon this by including confidence ratings,
394 rendering this binary judgement instead categorical. Our novel framework allows one to assess
395 memory performance in a more continuous way (*precision*), as well as analyze the correlational
396 structure of each encoding event to each memory event (*distinctiveness*). Further and importantly,
397 these two novel metrics we introduce here arise from comparisons of the actual content of the
398 experience/memories, which is not typically modeled. Leveraging this, we find that the successful
399 memory performance is related to 1) the precision with which the participant recounts each event
400 and 2) the distinctiveness of each recall event (relative to the other recalled events). The first finding
401 suggests that the information retained for *any individual event* may predict the overall amount of
402 information retained by the participant. The second finding suggests that the ability to distin-
403 guish between temporally or semantically similar content is also related to the quantity of content
404 recovered. Intriguingly, prior studies show that pattern separation, or the ability to discriminate
405 between similar experiences, is impaired in many cognitive disorders as well as natural aging

406 (Stark et al., 2010; Yassa et al., 2011; Yassa and Stark, 2011). Future work might explore whether
407 and how these metrics compare between cognitively impoverished groups and healthy controls.

408 While a large number of language models exist (e.g., WAS, LSA, word2vec, universal sentence
409 encoder; Steyvers et al., 2004; Landauer et al., 1998; Mikolov et al., 2013; Cer et al., 2018), here
410 we use latent dirichlet allocation (LDA)-based topic models for a few reasons. First, topic models
411 capture the *essence* of a text passage devoid of the specific set and order of words used. This
412 was an important feature of our model since different people may accurately recall a scene using
413 very different language. Second, words can mean different things in different contexts (e.g. “bat”
414 as the act of hitting a baseball, the object used for that action, or as a flying mammal). Topic
415 models are robust to this, allowing words to exist as part of multiple topics. Last, topic models
416 provide a straightforward means to recover the weights for the particular words comprising a topic,
417 enabling easy interpretation of an event’s contents (e.g. Fig. 6). Other models such as Google’s
418 universal sentence encoder offer a context-sensitive encoding of text passages, but the encoding
419 space is complex and non-linear, and thus recovering the original words used to fit the model is
420 not straightforward. However, it’s worth pointing out that our framework is divorced from the
421 particular choice of language model. Moreover, many of the aspects of our framework could be
422 swapped out for other choices. For example, the language model, the timeseries segmentation
423 model and the video-recall matching function could all be customized for the particular problem.
424 Indeed for some problems, recovery of the particular recall words may not be necessary, and thus
425 other text-modeling approaches (such as universal sentence encoder) may be preferable. Future
426 work will explore the influence of particular model choices on the framework’s accuracy.

427 Our work has broad implications for how we characterize and assess memory in real-world
428 settings, such as the classroom or physician’s office. For example, the most commonly used
429 classroom evaluation tools involve simply computing the proportion of correctly answered exam
430 questions. Our work indicates that this approach is only loosely related to what educators might
431 really want to measure: how well did the students understand the key ideas presented in the
432 course? Under this typical framework of assessment, the same exam score of 50% could be
433 ascribed to two very different students: one who attended the full course but struggled to learn

434 more than a broad overview of the material, and one who attended only half of the course but
435 understood the material perfectly. Instead, one could apply our computational framework to build
436 explicit content models of the course material and exam questions. This approach would provide
437 a more nuanced and specific view into which aspects of the material students had learned well
438 (or poorly). In clinical settings, memory measures that incorporate such explicit content models
439 might also provide more direct evaluations of patients' memories.

440 **Methods**

441 **Experimental design and data collection**

442 Data were collected by Chen et al. (2017). In brief, participants ($n = 17$) viewed the first 48 minutes
443 of "A Study in Pink", the first episode of the BBC television series *Sherlock*, while fMRI volumes
444 were collected (TR = 1500 ms). The stimulus was divided into a 23 min (946 TR) and a 25 min
445 (1030 TR) segment to mitigate technical issues related to the scanner. After finishing the clip,
446 participants were instructed to (quoting from Chen et al., 2017) "describe what they recalled of the
447 [episode] in as much detail as they could, to try to recount events in the original order they were
448 viewed in, and to speak for at least 10 minutes if possible but that longer was better. They were told
449 that completeness and detail were more important than temporal order, and that if at any point
450 they realized they had missed something, to return to it. Participants were then allowed to speak
451 for as long as they wished, and verbally indicated when they were finished (e.g., 'I'm done')."
452 For additional details about the experimental procedure and scanning parameters, see Chen et al.
453 (2017). The experimental protocol was approved by Princeton University's Institutional Review
454 Board.

455 After preprocessing the fMRI data and warping the images into a standard (3 mm³ MNI) space,
456 the voxel activations were z-scored (within voxel) and spatially smoothed using a 6 mm (full width
457 at half maximum) Gaussian kernel. The fMRI data were also cropped so that all video-viewing
458 data were aligned across participants. This included a constant 3 TR (4.5 s) shift to account for the

459 lag in the hemodynamic response. (All of these preprocessing steps followed Chen et al., 2017,
460 where additional details may be found.)

461 **Data and code availability**

462 The fMRI data we analyzed are available online [here](#). The behavioral data and all of our analysis
463 code may be downloaded [here](#).

464 **Statistics**

465 All statistical tests we performed were two-sided.

466 **Modeling the dynamic content of the video and recall transcripts**

467 **Topic modeling**

468 The input to the topic model we trained to characterize the dynamic content of the video comprised
469 hand-generated annotations of each of 1000 scenes spanning the video clip (generated by Chen
470 et al., 2017). The features annotated included: narrative details (a sentence or two describing
471 what happened in that scene); whether the scene took place indoors or outdoors; names of any
472 characters that appeared in the scene; name(s) of characters in camera focus; name(s) of characters
473 who were speaking in the scene; the location (in the story) that the scene took place; camera angle
474 (close up, medium, long, top, tracking, over the shoulder, etc.); whether music was playing in
475 the scene or not; and a transcription of any on-screen text. We concatenated the text for all of
476 these features within each segment, creating a “bag of words” describing each scene. We then
477 re-organized the text descriptions into overlapping sliding windows spanning 50 scenes each.
478 In other words, the first text sample comprised the combined text from the first 50 scenes (i.e.,
479 1–50), the second comprised the text from scenes 2–51, and so on. We trained our model using
480 these overlapping text samples with `scikit-learn` (version 0.19.1; Pedregosa et al., 2011), called
481 from our high-dimensional visualization and text analysis software, `HyperTools` (Heusser et al.,
482 2018b). Specifically, we used the `CountVectorizer` class to transform the text from each scene

483 into a vector of word counts (using the union of all words across all scenes as the “vocabulary,”
484 excluding English stop words); this yielded a number-of-scenes by number-of-words *word count*
485 matrix. We then used the `LatentDirichletAllocation` class (`topics=100, method='batch'`) to fit
486 a topic model (Blei et al., 2003) to the word count matrix, yielding a number-of-scenes (1000) by
487 number-of-topics (100) *topic proportions* matrix. The topic proportions matrix describes which mix
488 of topics (latent themes) is present in each scene. Next, we transformed the topic proportions
489 matrix to match the 1976 fMRI volume acquisition times. For each fMRI volume, we took the topic
490 proportions from whatever scene was displayed for most of that volume’s 1500 ms acquisition time.
491 This yielded a new number-of-TRs (1976) by number-of-topics (100) topic proportions matrix.

492 We created similar topic proportions matrices using hand-annotated transcripts of each par-
493 ticipant’s recall of the video (annotated by Chen et al., 2017). We tokenized the transcript into
494 a list of sentences, and then re-organized the list into overlapping sliding windows spanning 10
495 sentences each; in turn we transformed each window’s sentences into a word count vector (using
496 the same vocabulary as for the video model). We then used the topic model already trained on
497 the video scenes to compute the most probable topic proportions for each sliding window. This
498 yielded a number-of-sentences (range: 68–294) by number-of-topics (100) topic proportions matrix,
499 for each participant. These reflected the dynamic content of each participant’s recalls. Finally, we
500 resampled each recall model to match the timecourse of the video model. Note: for details on how
501 we selected the video and recall window lengths and number of topics, see *Supporting Information*
502 and Figure S1.

503 **Parsing topic trajectories into events using Hidden Markov Models**

504 We parsed the topic trajectories of the video and participants’ recalls into events using Hidden
505 Markov Models (Rabiner, 1989). Given the topic proportions matrix (describing the mix of topics
506 at each timepoint) and a number of states, K , an HMM recovers the set of state transitions that
507 segments the timeseries into K discrete states. Following Baldassano et al. (2017), we imposed an
508 additional set of constraints on the discovered state transitions that ensured that each state was
509 encountered exactly once (i.e., never repeated). We used the BrainIAK toolbox (Capota et al., 2017)

510 to implement this segmentation.

511 We used an optimization procedure to select the appropriate K for each topic proportions
512 matrix. Specifically, we computed (for each matrix)

$$\operatorname{argmax}_K \left[\frac{a}{b} - \frac{K}{\alpha} \right],$$

513 where a was the average correlation between the topic vectors of timepoints within the same state;
514 b was the average correlation between the topic vectors of timepoints within *different* states; and
515 α was a regularization parameter that we set to 5 times the window length (i.e., 250 scenes for
516 the video topic trajectory and 50 sentences for the recall topic trajectories). Figure 2B displays the
517 event boundaries returned for the video, and Figure S4 displays the event boundaries returned
518 for each participant's recalls. After obtaining these event boundaries, we created stable estimates
519 of each topic proportions matrix by averaging the topic vectors within each event. This yielded a
520 number-of-events by number-of-topics matrix for the video and recalls from each participant.

521 We also evaluated a parameter-free procedure for choosing K , which finds the K value that
522 maximizes the Wasserstein distance (a.k.a. “Earth mover’s” distance) between the within and
523 across event distributions of correlation values. This alternative procedure largely replicated the
524 pattern of results found with the parameterized method described above, but recovered sub-
525 stantially fewer events on average (Fig.S6). While both approaches seem to underestimate the
526 number of video/recall events relative to the “true” number (as determined by human raters), the
527 parameterized approach was closer to the true number.

528 **Naturalistic extensions of classic list-learning analyses**

529 In traditional list-learning experiments, participants view a list of items (e.g., words) and then recall
530 the items later. Our video-recall event matching approach affords us the ability to analyze memory
531 in a similar way. The video and recall events can be treated analogously to studied and recalled
532 “items” in a list-learning study. We can then extend classic analyses of memory performance and
533 dynamics (originally designed for list-learning experiments) to the more naturalistic video recall

534 task used in this study.

535 Perhaps the simplest and most widely used measure of memory performance is *accuracy*—i.e.,
536 the proportion of studied (experienced) items (in this case, the 34 video events) that the participant
537 later remembered. Chen et al. (2017) developed a human rating system whereby the quality of
538 each participant’s memory was evaluated by an independent rater. We found a strong across-
539 participants correlation between these independent ratings and the overall number of events that
540 our HMM approach identified in participants’ recalls (Pearson’s $r(15) = 0.67, p = 0.003$).

541 As described below, we next considered a number of memory performance measures that are
542 typically associated with list-learning studies. We also provide a software package, Quail, for
543 carrying out these analyses (Heusser et al., 2017).

544 **Probability of first recall (PFR).** PFR curves (Welch and Burnett, 1924; Postman and Phillips,
545 1965; Atkinson and Shiffrin, 1968) reflect the probability that an item will be recalled first as a
546 function of its serial position during encoding. To carry out this analysis, we initialized a number-
547 of-participants (17) by number-of-video-events (34) matrix of zeros. Then for each participant, we
548 found the index of the video event that was recalled first (i.e., the video event whose topic vector
549 was most strongly correlated with that of the first recall event) and filled in that index in the matrix
550 with a 1. Finally, we averaged over the rows of the matrix, resulting in a 1 by 34 array representing
551 the proportion of participants that recalled an event first, as a function of the order of the event’s
552 appearance in the video (Fig. 3A).

553 **Lag conditional probability curve (lag-CRP).** The lag-CRP curve (Kahana, 1996) reflects the
554 probability of recalling a given event after the just-recalled event, as a function of their relative
555 positions (or *lag*). In other words, a lag of 1 indicates that a recalled event came immediately after
556 the previously recalled event in the video, and a lag of -3 indicates that a recalled event came 3
557 events before the previously recalled event. For each recall transition (following the first recall),
558 we computed the lag between the current recall event and the next recall event, normalizing by
559 the total number of possible transitions. This yielded a number-of-participants (17) by number-

560 of-lags (-33 to +33; 67 lags total) matrix. We averaged over the rows of this matrix to obtain a
561 group-averaged lag-CRP curve (Fig. 3B).

562 **Serial position curve (SPC).** SPCs (Murdock, 1962) reflect the proportion of participants that
563 remember each item as a function of the items' serial position during encoding. We initialized
564 a number-of-participants (17) by number-of-video-events (34) matrix of zeros. Then, for each
565 recalled event, for each participant, we found the index of the video event that the recalled event
566 most closely matched (via the correlation between the events' topic vectors) and entered a 1 into
567 that position in the matrix (i.e., for the given participant and event). This resulted in a matrix
568 whose entries indicated whether or not each event was recalled by each participant (depending
569 on whether the corresponding entires were set to one or zero). Finally, we averaged over the rows
570 of the matrix to yield a 1 by 34 array representing the proportion of participants that recalled each
571 event as a function of the order of the event's appearance in the video (Fig. 3C).

572 **Temporal clustering scores.** Temporal clustering describes participants' tendency to organize
573 their recall sequences by the learned items' encoding positions. For instance, if a participant
574 recalled the video events in the exact order they occurred (or in exact reverse order), this would
575 yield a score of 1. If a participant recalled the events in random order, this would yield an expected
576 score of 0.5. For each recall event transition (and separately for each participant), we sorted
577 all not-yet-recalled events according to their absolute lag (i.e., distance away in the video). We
578 then computed the percentile rank of the next event the participant recalled. We averaged these
579 percentile ranks across all of the participant's recalls to obtain a single temporal clustering score
580 for the participant.

581 **Semantic clustering scores.** Semantic clustering describes participants' tendency to recall seman-
582 tically similar presented items together in their recall sequences. Here, we used the topic vectors
583 for each event as a proxy for its semantic content. Thus, the similarity between the semantic
584 content for two events can be computed by correlating their respective topic vectors. For each
585 recall event transition, we sorted all not-yet-recalled events according to how correlated the topic

586 vector of the closest-matching video event was to the topic vector of the closest-matching video event
587 to the just-recalled event. We then computed the percentile rank of the observed next recall. We
588 averaged these percentile ranks across all of the participant's recalls to obtain a single semantic
589 clustering score for the participant.

590 **Novel naturalistic memory metrics**

591 **Precision.** We tested whether participants who recalled more events were also more *precise* in their
592 recollections. For each participant, we computed the correlation between the topic vectors for each
593 recall event and that of its closest-matching video event (only for the events which they recalled).
594 We Fisher's z -transformed the correlations, computed the average and then inverse Fisher's z -
595 transformed the resulting value. This gave a single value per participant representing the average
596 precision across all recalled events. We then correlated this value with hand-annotated as well as
597 model derived (e.g. k or the number of events recovered by the HMM) memory performance.

598 **Distinctiveness.** We also considered the *distinctiveness* of each recalled event. That is, how
599 uniquely a recalled event's topic vector matched a given video event topic vector, versus the
600 topic vectors for the other video events. We hypothesized that participants with high memory
601 performance might describe each event in a more distinctive way (relative to those with lower
602 memory performance who might describe events in a more general way). To test this hypothesis
603 we define a distinctiveness score for each recalled event as

$$d(\text{event}) = 1 - \bar{c}(\text{event}),$$

604 where $\bar{c}(\text{event})$ is the average correlation between the given recalled event's topic vector and the
605 topic vectors from all video events *except* the best-matching video event. We then averaged these
606 distinctiveness scores across all of the events recalled by the given participant. As above, we used
607 Fisher's z -transform (and inverse-transform) before/after averaging correlation values. Finally,
608 we correlated these values with hand-annotated and model derived memory performance scores

609 across-subjects.

610 **Visualizing the video and recall topic trajectories**

611 We used the UMAP algorithm (McInnes and Healy, 2018) to project the 100-dimensional topic space
612 onto a two-dimensional space for visualization (Figs. 5, 6). To ensure that all of the trajectories were
613 projected onto the *same* lower dimensional space, we computed the low-dimensional embedding
614 on a “stacked” matrix created by vertically concatenating the events-by-topics topic proportions
615 matrices for the video and all 17 participants’ recalls. We then divided the rows of the result (a
616 total-number-of-events by two matrix) back into separate matrices for the video topic trajectory
617 and the trajectories for each participant’s recalls (Fig. 5). This general approach for discovering
618 a shared low-dimensional embedding for a collections of high-dimensional observations follows
619 Heusser et al. (2018b).

620 **Estimating the consistency of flow through topic space across participants**

621 In Figure 5B, we present an analysis aimed at characterizing locations in topic space that dif-
622 ferent participants move through in a consistent way (via their recall topic trajectories). The
623 two-dimensional topic space used in our visualizations (Fig. 5) ranged from -5 to 5 (arbitrary) units
624 in the x dimension and from -6.5 to 2 units in the y dimension. We divided this space into a grid
625 of vertices spaced 0.25 units apart. For each vertex, we examined the set of line segments formed
626 by connecting each pair successively recalled events, across all participants, that passed within 0.5
627 units. We computed the distribution of angles formed by those segments and the x -axis, and used a
628 Rayleigh test to determine whether the distribution of angles was reliably “peaked” (i.e., consistent
629 across all transitions that passed through that local portion of topic space). To create Figure 5B we
630 drew an arrow originating from each grid vertex, pointing in the direction of the average angle
631 formed by line segments that passed within 0.5 units. We set the arrow lengths to be inversely
632 proportional to the p -values of the Rayleigh tests at each vertex. Specifically, for each vertex we
633 converted all of the angles of segments that passed within 0.5 units to unit vectors, and we set
634 the arrow lengths at each vertex proportional to the length of the (circular) mean vector. We also

635 indicated any significant results ($p < 0.05$, corrected using the Benjamani-Hochberg procedure) by
636 coloring the arrows in blue (darker blue denotes a lower p -value, i.e., a longer mean vector); all
637 tests with $p \geq 0.05$ are displayed in gray and given a lower opacity value.

638 **Searchlight fMRI analyses**

639 In Figure 7, we present two analyses aimed at identifying brain structures whose responses (as
640 participants viewed the video) exhibited particular temporal correlations. We developed a search-
641 light analysis whereby we constructed a cube centered on each voxel (radius: 5 voxels). For each
642 of these cubes, we computed the temporal correlation matrix of the voxel responses during video
643 viewing. Specifically, for each of the 1976 volumes collected during video viewing, we correlated
644 the activity patterns in the given cube with the activity patterns (in the same cube) collected during
645 every other timepoint. This yielded a 1976 by 1976 correlation matrix for each cube.

646 Next, we constructed two sets of “template” matrices: one reflecting the video’s topic trajectory
647 and the other reflecting each participant’s recall topic trajectory. To construct the video template, we
648 computed the correlations between the topic proportions estimated for every pair of TRs (prior to
649 segmenting the trajectory into discrete events; i.e., the correlation matrix shown in Figs. 2B and 7A).
650 We constructed similar temporal correlation matrices for each participant’s recall topic trajectory
651 (Figs. 2D, S4). However, to correct for length differences and potential non-linear transformations
652 between viewing time and recall time, we first used dynamic time warping (Berndt and Clifford,
653 1994) to temporally align participants’ recall topic trajectories with the video topic trajectory (an
654 example correlation matrix before and after warping is shown in Fig. 7B). This yielded a 1976 by
655 1976 correlation matrix for the video template and for each participant’s recall template.

656 To determine which (cubes of) voxel responses reliably matched the video template, we cor-
657 related the upper triangle of the voxel correlation matrix for each cube with the upper triangle
658 of the video template matrix (Kriegeskorte et al., 2008). This yielded, for each participant, a
659 single correlation value. We computed the average (Fisher z -transformed) correlation coefficient
660 across participants. We used a permutation-based procedure to assess significance, whereby we
661 re-computed the average correlations for each of 100 “null” video templates (constructed by circu-

larly shifting the template by a random number of timepoints). (For each permutation, the same shift was used for all participants.) We then estimated a p -value by computing the proportion of shifted correlations that were larger than the observed (unshifted) correlation. To create the map in Figure 7A we thresholded out any voxels whose correlation values fell below the 95th percentile of the permutation-derived null distribution.

We used a similar procedure to identify which voxels' responses reflected the recall templates. For each participant, we correlated the upper triangle of the correlation matrix for each cube of voxels with their (time warped) recall correlation matrix. As in the video template analysis this yielded a single correlation coefficient for each participant. However, whereas the video analysis compared every participant's responses to the same template, here the recall templates were unique for each participant. We computed the average z-transformed correlation coefficient across participants, and used the same permutation procedure we developed for the video responses to assess significant correlations. To create the map in Figure 7B we thresholded out any voxels whose correlation values fell below the 95th percentile of the permutation-derived null distribution.

References

- Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Spence, K. W. and Spence, J. T., editors, *The psychology of learning and motivation*, volume 2, pages 89–105. Academic Press, New York.
- Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721.
- Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370.
- Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, US. ACM.

- 687 Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*
688 *Learning Research*, 3:993 – 1022.
- 689 Brunec, I. K., Moscovitch, M. M., and Barense, M. D. (2018). Boundaries shape cognitive represen-
690 tations of spaces and events. *Trends in Cognitive Sciences*, 22(7):637–650.
- 691 Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., and Oliva, A. (2015). Intrinsic and extrinsic
692 effects on image memorability. *Vision Research*, 116:165–178.
- 693 Capota, M., Turek, J., Chen, P.-H., Zhu, X., Manning, J. R., Sundaram, N., Keller, B., Wang, Y., and
694 Shin, Y. S. (2017). Brain imaging analysis kit.
- 695 Cer, D., Yang, Y., y Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes,
696 M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.
697 *arXiv*, 1803.11175.
- 698 Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared
699 memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*,
700 20(1):115.
- 701 Chun, M. and Turk-Browne, N. (2007). Interactions between attention and memory. *Current opinion*
702 *in neurobiology*, 17(2):177–184.
- 703 Clewett, D. and Davachi, L. (2017). The ebb and flow of experience determines the temporal
704 structure of memory. *Curr Opin Behav Sci*, 17:186–193.
- 705 Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in*
706 *Neurobiology*, 16(6):693—700.
- 707 Davachi, L., Mitchell, J. P., and Wagner, A. D. (2003). Multiple routes to memory: distinct medial
708 temporal lobe processes build item and source memories. *Proceedings of the National Academy of*
709 *Sciences, USA*, 100(4):2157 – 2162.

- 710 Diana, R. A., Yonelinas, A. P., and Ranganath, C. (2007). Imaging recollection and famil-
711 iarity in the medial temporal lobe: a three-component model. *Trends in Cognitive Sciences*,
712 doi:10.1016/j.tics.2007.08.001.
- 713 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the
714 sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–1286.
- 715 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological
716 Science*, 22(2):243–252.
- 717 Friendly, M., Franklin, P. E., Hoffman, D., and Rubin, D. C. (1982). The Toronto Word Pool:
718 Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080
719 words. *Behavior Research Methods and Instrumentation*, 14:375–399.
- 720 Gilboa, A. and Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory.
721 *Trends Cogn Sci*, 21(8):618–631.
- 722 Heusser, A. C., Ezzyat, Y., Shiff, I., and Davachi, L. (2018a). Perceptual boundaries cause mnemonic
723 trade-offs between local boundary processing and across-trial associative binding. *Journal of
724 Experimental Psychology Learning, Memory, and Cognition*, 44(7):1075–1090.
- 725 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail: a
726 Python toolbox for analyzing and plotting free recall data. *The Journal of Open Source Software*,
727 10.21105/joss.00424.
- 728 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018b). HyperTools: a Python
729 toolbox for gaining geometric insights into high-dimensional data. *Journal of Machine Learning
730 Research*, 18(152):1–6.
- 731 Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal
732 of Mathematical Psychology*, 46:269–299.
- 733 Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., and H., E.

- 734 (2014). A unified mathematical framework for coding time, space, and sequences in the medial
735 temporal lobe. *Journal of Neuroscience*, 34(13):4692–4707.
- 736 Howard, M. W., Viskontas, I. V., Shankar, K. H., and Fried, I. (2012). Ensembles of human MTL
737 neurons “jump back in time” in response to a repeated stimulus. *Hippocampus*, 22:1833–1847.
- 738 Huk, A., Bonnen, K., and He, B. J. (2018). Beyond trial-based paradigms: continuous behavior, on-
739 going neural activity, and naturalistic stimuli. *Journal of Neuroscience*, 10.1523/JNEUROSCI.1920-
740 17.2018.
- 741 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24:103–109.
- 742 Koriat, A. and Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: distin-
743 guishing accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of*
744 *Experimental Psychology: General*, 123(3):297–315.
- 745 Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – con-
746 nnecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:1 – 28.
- 747 Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to latent semantic analysis.
748 *Discourse Processes*, 25:259–284.
- 749 Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy
750 of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.
- 751 Manning, J. R. (2019). Episodic memory: mental time travel or a quantum ‘memory wave’ function?
752 *PsyArXiv*, doi:10.31234/osf.io/6zjwb.
- 753 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory.
754 In Gazzaniga, M., editor, *The Cognitive Neurosciences, Fifth edition*, pages 557–566. MIT Press.
- 755 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns
756 in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National
757 Academy of Sciences, USA*, 108(31):12893–12897.

- 758 McInnes, L. and Healy, J. (2018). UMAP: Uniform manifold approximation and projection for
759 dimension reduction. *arXiv*, 1802(03426).
- 760 Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations
761 in vector space. *arXiv*, 1301.3781.
- 762 Mueller, A., Fillion-Robin, J.-C., Boidol, R., Tian, F., Nechifor, P., yoonsubKim, Peter, Rampin, R.,
763 Corvellec, M., Medina, J., Dai, Y., Petrushev, B., Langner, K. M., Hong, Alessio, Ozsvald, I.,
764 vkolmakov, Jones, T., Bailey, E., Rho, V., IgorAPM, Roy, D., May, C., foobuzz, Piyush, Seong,
765 L. K., Goey, J. V., Smith, J. S., Gus, and Mai, F. (2018). WordCloud 1.5.0: a little word cloud
766 generator in Python. *Zenodo*, <https://zenodo.org/record/1322068#.W4tPKZNKh24>.
- 767 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*,
768 64:482–488.
- 769 Paller, K. A. and Wagner, A. D. (2002). Observing the transformation of experience into memory.
770 *Trends in Cognitive Sciences*, 6(2):93–102.
- 771 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten-
772 hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot,
773 M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine
774 Learning Research*, 12:2825–2830.
- 775 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model
776 of organizational processes in free recall. *Psychological Review*, 116(1):129–156.
- 777 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal
778 of Experimental Psychology*, 17:132–138.
- 779 Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech
780 recognition. *Proceedings of the IEEE*, 77(2):257–286.
- 781 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition. *Curr Opin
782 Behav Sci*, 17:133–140.

- 783 Ranganath, C., Cohen, M. X., Dam, C., and D'Esposito, M. (2004). Inferior temporal, prefrontal,
784 and hippocampal contributions to visual working memory maintenance and associative memory
785 retrieval. *Journal of Neuroscience*, 24(16):3917–3925.
- 786 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature
787 Reviews Neuroscience*, 13:713 – 726.
- 788 Schlichting, M. L. and Preston, A. R. (2015). Memory integration: neural mechanisms and impli-
789 cations for behavior. *Current Opinion in Behavioral Sciences*, 1:1–8.
- 790 Simony, E., Honey, C. J., Chen, J., and Hasson, U. (2016). Uncovering stimulus-locked network
791 dynamics during narrative comprehension. *Nature Communications*, 7(12141):1–13.
- 792 Spalding, K. N., Schlichting, M. L., Zeithamova, D., Preston, A. R., Tranel, D., Duff, M. C., and
793 Warren, D. E. (2018). Ventromedial prefrontal cortex is necessary for normal associative inference
794 and memory integration. *The Journal of Neuroscience*, 38(15):3767–3775.
- 795 Stark, S. M., Yassa, M. A., and Stark, C. E. L. (2010). Individual differences in spatial pattern
796 separation performance associated with healthy aging in humans. *Learning & Memory*, 17(6):284–
797 288.
- 798 Steyvers, M., Shiffrin, R. M., and Nelson, D. L. (2004). Word association spaces for predicting
799 semantic similarity effects in episodic memory. In Healy, A. F., editor, *Cognitive Psychology and
800 its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*. American
801 Psychological Association, Washington, DC.
- 802 Tompry, A. and Davachi, L. (2017). Consolidation promotes the emergence of representational
803 overlap in the hippocampus and medial prefrontal cortex. *Neuron*, 96(1):228–241.
- 804 Tyng, C. M., Amin, H. U., Saad, M. N. M., and S, M. A. (2017). The influences of emotion on
805 learning and memory. *Frontiers in psychology*, 8:1454.
- 806 van Kesteren, M. T. R., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., and Fernández, G.

- 807 (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent
808 encoding: from congruent to incongruent. *Neuropsychologia*, 51(12):2352–2359.
- 809 van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., and Henson, R. N. (2012). How schema and
810 novelty augment memory formation. *Trends Neurosci*, 35(4):211–9.
- 811 Waskom, M., Botvinnik, O., Okane, D., Hobson, P., David, Halchenko, Y., Lukauskas, S., Cole, J. B.,
812 Warmenhoven, J., de Ruiter, J., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E.,
813 Martin, M., Miles, A., Meyer, K., Augspurger, T., Yarkoni, T., Bachant, P., Williams, M., Evans,
814 C., Fitzgerald, C., Brian, Wehner, D., Hitz, G., Ziegler, E., Qalieh, A., and Lee, A. (2016). Seaborn:
815 v0.7.1.
- 816 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American Journal
817 of Psychology*, 35:396–401.
- 818 Wiltgen, B. J. and Silva, A. J. (2007). Memory for context becomes less specific with time. *Learning
819 & Memory*, 14(4):313–317.
- 820 Yassa, M. A., Lacy, J. W., Stark, S. M., Albert, M. S., Gallagher, M., and Stark, C. E. L. (2011). Pattern
821 separation deficits associated with increased hippocampal ca3 and dentate gyrus activity in
822 nondemented older adults. *Hippocampus*, 21(9):968–979.
- 823 Yassa, M. A. and Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends In Neuro-
824 sciences*, 34(10):515–525.
- 825 Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research.
826 *Journal of Memory and Language*, 46:441–517.
- 827 Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception:
828 a mind-brain perspective. *Psychological Bulletin*, 133:273–293.
- 829 Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit
830 memories to other brains: Constructing shared neural representations via communication. *Cereb
831 Cortex*, 27(10):4988–5000.

⁸³² Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and
⁸³³ memory. *Psychological Bulletin*, 123(2):162 – 185.

⁸³⁴ **Supporting information**

⁸³⁵ Supporting information is available in the online version of the paper.

⁸³⁶ **Acknowledgements**

⁸³⁷ We thank Luke Chang, Janice Chen, Chris Honey, Lucy Owen, Emily Whitaker, and Kirsten Ziman
⁸³⁸ for feedback and scientific discussions. We also thank Janice Chen, Yuan Chang Leong, Kenneth
⁸³⁹ Norman, and Uri Hasson for sharing the data used in our study. Our work was supported in part
⁸⁴⁰ by NSF EPSCoR Award Number 1632738. The content is solely the responsibility of the authors
⁸⁴¹ and does not necessarily represent the official views of our supporting organizations.

⁸⁴² **Author contributions**

⁸⁴³ Conceptualization: A.C.H. and J.R.M.; Methodology: A.C.H. and J.R.M.; Software: A.C.H., P.C.F.
⁸⁴⁴ and J.R.M.; Analysis: A.C.H., P.C.F. and J.R.M.; Writing, Reviewing, and Editing: A.C.H., P.C.F.
⁸⁴⁵ and J.R.M.; Supervision: J.R.M.

⁸⁴⁶ **Author information**

⁸⁴⁷ The authors declare no competing financial interests. Correspondence and requests for materials
⁸⁴⁸ should be addressed to J.R.M. (jeremy.r.manning@dartmouth.edu).