

1 Geometric models reveal behavioral and neural
2 signatures of how naturalistic experiences are
3 transformed into episodic memories

4 Andrew C. Heusser^{1, 2, †}, Paxton C. Fitzpatrick^{1, †}, and Jeremy R. Manning^{1, *}

¹Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

²Akili Interactive

Boston, MA 02110

[†]Denotes equal contribution

^{*}Corresponding author: Jeremy.R.Manning@Dartmouth.edu

5 August 20, 2020

6 **Abstract**

7 Our ongoing subjective experience reflects external sensory information from each moment,
8 along with additional information from our past that we carry with us into that moment. The
9 blend of memories, knowledge, emotions, goals, and other internal perceptual and mental states
10 that color our subjective experience provides a *context* for interpreting new information and
11 conceptually linking what is happening now with our prior experiences. Because this contextual
12 information is often person-specific, the subjective experience that each person encodes into their
13 memory is often idiosyncratic, even for shared experiences and sensory perspectives. We sought
14 to study which aspects of a shared naturalistic experience were preserved or distorted, and how
15 those distortions compared across individuals. To this end, we developed a geometric frame-

16 work for mathematically characterizing the subjective conceptual content of dynamic naturalistic
17 experiences. We model experiences as *trajectories* through word embedding spaces whose coor-
18 dinates reflect the universe of thoughts under consideration. We also demonstrate how *memories*
19 may also be modeled as trajectories through the same spaces. According to this view, encod-
20 ing an experience into memory entails geometrically distorting or transforming the *shape* of the
21 original experience’s trajectory. This translates qualitative, neuropsychological questions about
22 how we remember naturalistic experiences into quantitative, geometric questions about the spatial
23 configurations of trajectory shapes. We applied our framework to data collected as participants
24 watched and verbally recounted a television episode while undergoing functional neuroimaging.
25 We found that the trajectories of participants’ recounts of the episode nearly all captured
26 the coarse spatial properties of the original episode’s trajectory (i.e., the essential plot points),
27 but participants differed in their memory for fine details. We also identified a network of brain
28 structures that were sensitive to the shape of the episode’s trajectory through word embedding
29 space, and an overlapping network that predicted, at the time of encoding, how people would
30 distort (transform) the episode’s trajectory when they recounted the episode later. Our work
31 provides insights into how our brains distort and transform our ongoing experiences when we
32 encode them into episodic memories.

33 **Introduction**

34 What does it mean to *remember* something? In traditional episodic memory experiments (e.g.,
35 list-learning or trial-based experiments; Murdock, 1962; Kahana, 1996), remembering is often cast
36 as a discrete and binary operation: each studied item may be separated from the rest of one’s
37 experience and singularly labeled as having been recalled or forgotten. More nuanced studies
38 might incorporate self-reported confidence measures as a proxy for memory strength, or ask
39 participants to discriminate between “recollecting” the (contextual) details of an experience or
40 having a general feeling of “familiarity” (Yonelinas, 2002). Using well-controlled, trial-based
41 experimental designs, the field has amassed a wealth of information regarding human episodic
42 memory. However, there are fundamental properties of the external world and our memories that

43 trial-based experiments are not well suited to capture (for review, also see Koriat and Goldsmith,
44 1994; Huk et al., 2018). First, our experiences and memories are continuous, rather than discrete—
45 isolating a (naturalistic) event from the context in which it occurs can substantially change its
46 meaning. Second, whether or not the rememberer has precisely reproduced a specific set of words
47 in describing a given experience is nearly orthogonal to how well they were actually able to
48 remember it. In classic (e.g., list-learning) memory studies, by contrast, the number or proportion
49 of *exact* recalls is often considered to be a primary metric for assessing the quality of participants'
50 memories. Third, one might remember the *essence* (or a general summary) of an experience but
51 forget (or neglect to recount) particular details. Capturing the essence of what happened is often
52 a main goal of recounting an episodic memory to a listener, whereas the inclusion of specific,
53 low-level details is often less pertinent.

54 How might we formally characterize the *essence* of an experience, and whether it has been
55 recovered by the rememberer? And how might we distinguish an experience's overarching essence
56 from its low-level details? One approach is to start by considering some fundamental properties
57 of the dynamics of our experiences. Each given moment of an experience tends to derive meaning
58 from surrounding moments, as well as from longer-range temporal associations (Lerner et al., 2011;
59 Manning, 2019, 2020). Therefore, the timecourse describing how an event unfolds is fundamental
60 to its overall meaning. Further, this hierarchy formed by our subjective experiences at different
61 timescales defines a *context* for each new moment (e.g., Howard and Kahana, 2002; Howard
62 et al., 2014), and plays an important role in how we interpret that moment and remember it
63 later (for review see Manning et al., 2015; Manning, 2020). Our memory systems can leverage
64 these associations to form predictions that help guide our behaviors (Ranganath and Ritchey,
65 2012). For example, as we navigate the world, the features of our subjective experiences tend
66 to change gradually (e.g., the room or situation we find ourselves in at any given moment is
67 strongly temporally autocorrelated), allowing us to form stable estimates of our current situation
68 and behave accordingly (Zacks et al., 2007; Zwaan and Radvansky, 1998).

69 Occasionally, this gradual “drift” of our ongoing experience is punctuated by sudden changes,
70 or “shifts” (e.g., when we walk through a doorway; Radvansky and Zacks, 2017). Prior research

suggests that these sharp transitions (termed *event boundaries*) help to discretize our experiences (and their mental representations) into *events* (Radvansky and Zacks, 2017; Brunec et al., 2018; Heusser et al., 2018a; Clewett and Davachi, 2017; Ezzyat and Davachi, 2011; DuBrow and Davachi, 2013). The interplay between the stable (within-event) and transient (across-event) temporal dynamics of an experience also provides a potential framework for transforming experiences into memories that distills those experiences down to their essence. For example, prior work has shown that event boundaries can influence how we learn sequences of items (Heusser et al., 2018a; DuBrow and Davachi, 2013), navigate (Brunec et al., 2018), and remember and understand narratives (Zwaan and Radvansky, 1998; Ezzyat and Davachi, 2011). This work also suggests a means of distinguishing the essence of an experience from its low-level details. The overall structure of events and event transitions reflects how the high-level experience unfolds (i.e., its essence), while subtler event-level properties reflect low-level details. Prior research has also implicated a network of brain regions (including the hippocampus and the medial prefrontal cortex) in playing a critical role in transforming experiences into structured and consolidated memories (Tompry and Davachi, 2017).

Here, we sought to examine how the temporal dynamics of a “naturalistic” experience were later reflected in participants’ memories. We also sought to leverage the above conceptual insights into the distinctions between an experience’s essence and low-level details to build models that explicitly quantified these distinctions. We analyzed an open dataset that comprised behavioral and functional Magnetic Resonance Imaging (fMRI) data collected as participants viewed and then verbally recounted an episode of the BBC television series *Sherlock* (Chen et al., 2017). We developed a computational framework for characterizing the temporal dynamics of the moment-by-moment content of the episode, and of participants’ verbal recalls. Specifically, we use topic modeling (Blei et al., 2003) to characterize the thematic conceptual (semantic) content present in each moment of the episode and recalls, and hidden Markov models (Rabiner, 1989; Baldassano et al., 2017) to discretize this evolving semantic content into events. In this way, we cast both naturalistic experiences and memories of those experiences as geometric *trajectories* that describe how they evolve over time. Under this framework, successful remembering entails verbally “traversing” the content trajectory

99 of the episode, thereby reproducing the shape (essence) of the original experience. Our framework
100 captures the episode’s essence in the sequence of geometric coordinates for its events, and its
101 low-level details by examining its within-event geometric properties.

102 Comparing the overall shapes of the topic trajectories for the episode and participants’ recalls
103 reveals which aspects of the episode’s essence were preserved (or discarded) in the translation into
104 memory. We also develop two metrics for assessing participants’ memories for low-level details:
105 (1) the *precision* with which a participant recounts details about each event, and (2) the *distinctiveness*
106 of each recall event, relative to other recalled events. We examine how these metrics relate to overall
107 memory performance as judged by third-party human annotators. We also compare and contrast
108 our general approach to studying memory for naturalistic experiences with standard metrics for
109 assessing performance on more traditional memory tasks, such as list-learning. Last, we leverage
110 our framework to identify networks of brain structures whose responses (as participants watched
111 the episode) reflected the temporal dynamics of either the episode or how participants would later
112 recount it.

113 Results

114 To characterize the dynamic content of the *Sherlock* episode and participants’ subsequent recounts
115 we used a topic model (Blei et al., 2003) to discover the episode’s latent themes. Topic models
116 take as inputs a vocabulary of words to consider and a collection of text documents, and return
117 two output matrices. The first of these is a *topics matrix* whose rows are *topics* (or latent themes)
118 and whose columns correspond to words in the vocabulary. The entries in the topics matrix
119 reflect how each word in the vocabulary is weighted by each discovered topic. For example, a
120 detective-themed topic might weight heavily on words like “crime,” and “search.” The second
121 output is a *topic proportions matrix*, with one row per document and one column per topic. The
122 topic proportions matrix describes the mixture of discovered topics reflected in each document.

123 Chen et al. (2017) collected hand-annotated information about each of 1,000 (manually iden-
124 tified) scenes spanning the roughly 50 minute video used in their experiment. This information

125 included: a brief narrative description of what was happening, the location where the scene took
126 place, the names of any characters on the screen, and other similar details (for a full list of annotated
127 features, see *Methods*). We took from these annotations the union of all unique words (excluding
128 stop words, such as “and,” “or,” “but,” etc.) across all features and scenes as the “vocabulary” for
129 the topic model. We then concatenated the sets of words across all features contained in overlap-
130 ping sliding windows of (up to) 50 scenes, and treated each window as a single “document” for
131 the purpose of fitting the topic model. Next, we fit a topic model with (up to) $K = 100$ topics to this
132 collection of documents. We found that 32 unique topics (with non-zero weights) were sufficient
133 to describe the time-varying content of the video (see *Methods*; Figs. 1, S2). Note that our approach
134 is similar in some respects to Dynamic Topic Models (Blei and Lafferty, 2006), in that we sought
135 to characterize how the thematic content of the episode evolved over time. However, whereas
136 Dynamic Topic Models are designed to characterize how the properties of *collections* of documents
137 change over time, our sliding window approach allows us to examine the topic dynamics within
138 a single document (or video). Specifically, our approach yielded (via the topic proportions matrix)
139 a single *topic vector* for each sliding window of annotations transformed by the topic model. We
140 then stretched (interpolated) the resulting windows-by-topics matrix to match the time series of
141 the 1,976 fMRI volumes collected as participants viewed the episode.

142 The 32 topics we found were heavily character-focused (i.e., the top-weighted word in each topic
143 was nearly always a character) and could be roughly divided into themes centered around Sherlock
144 Holmes (the titular character), John Watson (Sherlock’s close confidant and assistant), supporting
145 characters (e.g., Inspector Lestrade, Sergeant Donovan, or Sherlock’s brother Mycroft), or the
146 interactions between various groupings of these characters (see Fig. S2). Several of the identified
147 topics were highly similar, which we hypothesized might allow us to distinguish between subtle
148 narrative differences if the distinctions between those overlapping topics were meaningful. The
149 topic vectors for each timepoint were also *sparse*, in that only a small number (typically one or
150 two) of topics tended to be “active” in any given timepoint (see Fig. 2A). Further, the dynamics
151 of the topic activations appeared to exhibit *persistence* (i.e., given that a topic was active in one
152 timepoint, it was likely to be active in the following timepoint) along with *occasional rapid changes*

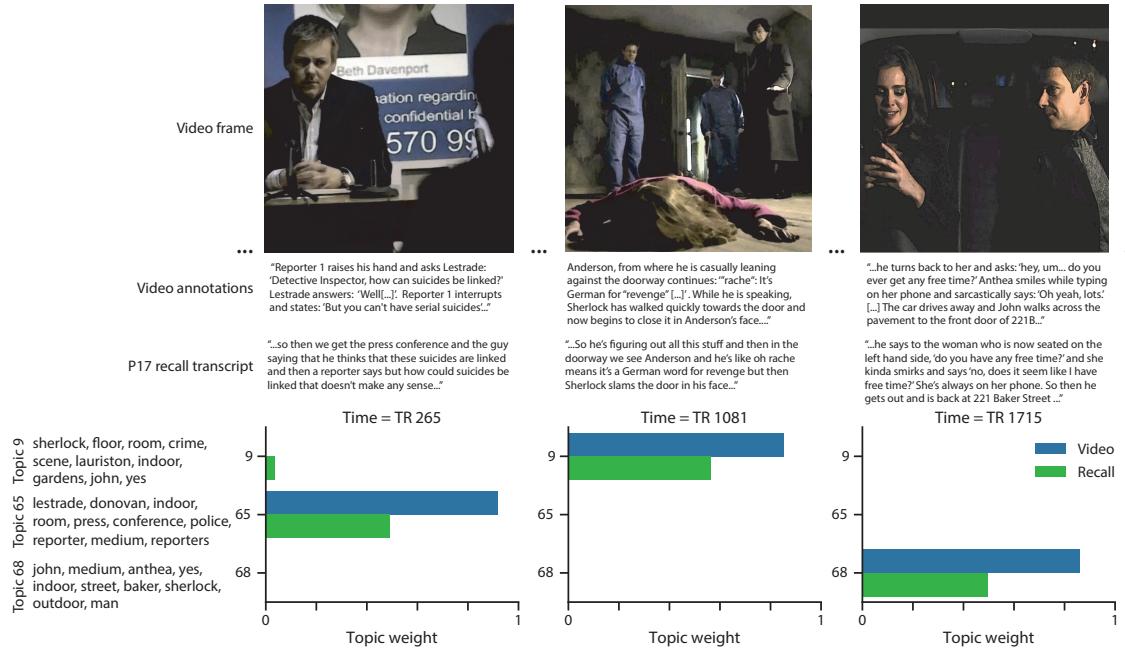


Figure 1: Methods overview. We used hand-annotated descriptions of each moment of video to fit a topic model. Three example video frames and their associated descriptions are displayed (top two rows). Participants later recalled the video (in the third row, we show example recalls of the same three scenes from participant 17). 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 10 highest-weighted words for each topic. Figure S2 provides a full list of the top 10 words from each of the discovered topics.

153 (i.e., occasionally topics would appear to spring into or out of existence). These two properties
154 of the topic dynamics may be seen in the block diagonal structure of the timepoint-by-timepoint
155 correlation matrix (Fig. 2B) and reflect the gradual drift and sudden shifts fundamental to the
156 temporal dynamics of real-world experiences. Given this observation, we adapted an approach
157 devised by Baldassano et al. (2017), and used a hidden Markov model (HMM) to identify the *event*
158 *boundaries* where the topic activations changed rapidly (i.e., the boundaries of the blocks in the
159 temporal correlation matrix; event boundaries identified by the HMM are outlined in yellow in
160 Fig. 2B). Part of our model fitting procedure required selecting an appropriate number of “events”
161 into which the topic trajectory should be segmented. To accomplish this, we used an optimization
162 procedure that maximized the difference between the topic weights for timepoints within an event
163 versus timepoints across multiple events (see *Methods* for additional details). We then created a
164 stable “summary” of the content within each video event by averaging the topic vectors across the
165 timepoints spanned by each event (Fig. 2C).

166 Given that the time-varying content of the video could be segmented cleanly into discrete
167 events, we wondered whether participants’ recalls of the video also displayed a similar structure.
168 We applied the same topic model (already trained on the video annotations) to each participant’s
169 recalls. Analogously to how we parsed the time-varying content of the video, to obtain similar
170 estimates for each participant’s recall, we treated each overlapping window of (up to 10) sentences
171 from their transcript as a “document,” and computed the most probable mix of topics reflected in
172 each timepoint’s sentences. This yielded, for each participant, a number-of-windows by number-
173 of-topics topic proportions matrix that characterized how the topics identified in the original video
174 were reflected in the participant’s recalls. Note that an important feature of our approach is that
175 it allows us to compare participants’ recalls to events from the original video, despite different
176 participants using widely varying language to describe the events, and that those descriptions
177 often diverged in content and quality from the video annotations. This is a substantial benefit of
178 projecting the video and recalls into a shared “topic” space. An example topic proportions matrix
179 from one participant’s recalls is shown in Figure 2D.

180 Although the example participant’s recall topic proportions matrix has some visual similarity to

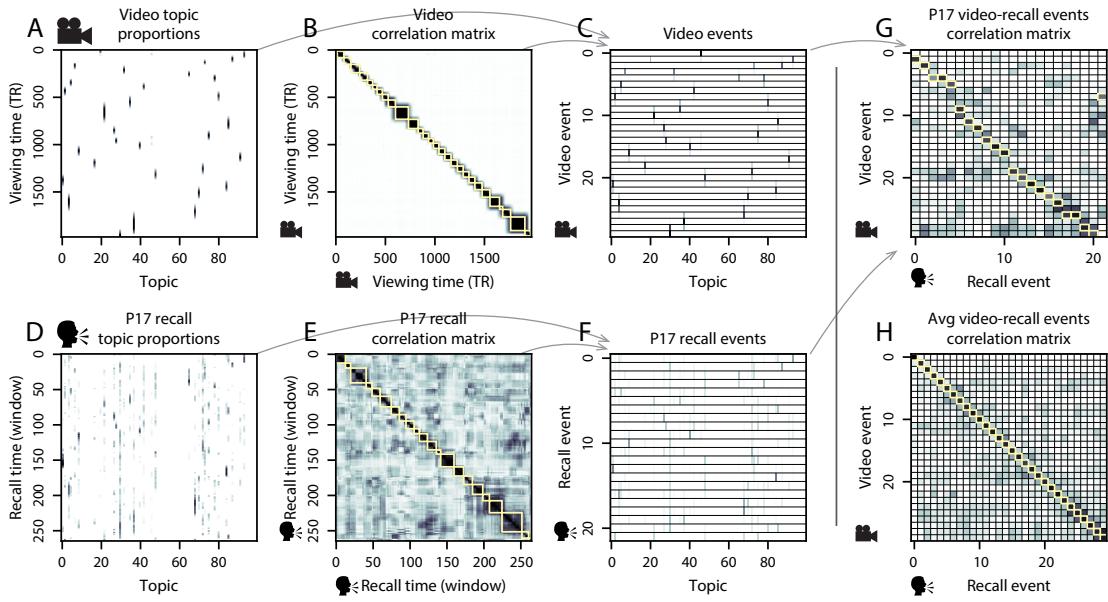


Figure 2: Modelling naturalistic stimuli and recalls. All panels: darker colors indicate greater values; range: [0, 1]. **A.** Topic vectors ($K = 100$) for each of the 1976 video timepoints. **B.** Timepoint-by-timepoint correlation matrix of the topic vectors displayed in Panel A. Event boundaries discovered by the HMM are denoted in yellow (30 events detected). **C.** Average topic vectors for each of the 30 video events. **D.** Topic vectors for each of 265 sliding windows of sentences spoken by an example participant while recalling 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 (22 events detected). For similar plots for all participants, see Figure S4. **F.** Average topic vectors for each of the 22 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).

181 the video topic proportions matrix, the time-varying topic proportions for the example participant's
182 recalls are not as sparse as those for the video (compare Figs. 2A and D). Similarly, although
183 there do appear to be periods of stability in the recall topic dynamics (i.e., most topics are active
184 or inactive over contiguous blocks of time), the changes in topic activations that define event
185 boundaries appear less clearly delineated in participants' recalls than in the episode's annotations.
186 To examine these patterns in detail, we computed the timepoint-by-timepoint correlation matrix
187 for the example participant's recall trajectory (Fig. 2E). As in the video correlation matrix (Fig. 2B),
188 the example participant's recall correlation matrix has a strong block diagonal structure, indicating
189 that their recalls are discretized into separated events. As for the video correlation matrix, we
190 leveraged an HMM-based optimization procedure (see *Methods*) to determine how many events
191 are reflected in the participant's recalls and where specifically the event boundaries fall (outlined
192 in yellow). We carried out a similar analysis on all 17 participants' recall topic proportions matrices
193 (Fig. S4).

194 Two clear patterns emerged from this set of analyses. First, although every individual partic-
195 ipant's recalls could be segmented into discrete events (i.e., every individual participant's recall
196 correlation matrix exhibited clear block diagonal structure; Fig. S4), each participant appeared to
197 have a unique *recall resolution*, reflected in the sizes of those blocks. While some participants' recall
198 topic proportions segmented into just a few events (e.g., Participants P4, P5, and P7), others' seg-
199 mented into many shorter duration events (e.g., Participants P12, P13, and P17). This suggests that
200 different participants may be recalling the video with different levels of detail—i.e., some might
201 touch on just the major plot points, whereas others might attempt to recall every minor scene or
202 action. The second clear pattern present in every individual participant's recall correlation matrix
203 was that, unlike in the video correlation matrix, there were substantial off-diagonal correlations.
204 Whereas each event in the original video was (largely) separable from the others (Fig. 2B), in
205 transforming those separable events into memory, participants appeared to be integrating across
206 multiple events, blending elements of previously recalled and not-yet-recalled content into each
207 newly recalled event (Figs. 2E, S4; also see Manning et al., 2011; Howard et al., 2012; Manning,
208 2019).

209 The above results indicate that both the structure of the original video and participants' recalls
210 of the video exhibit event boundaries that can be identified automatically by characterizing the
211 dynamic content using a shared topic model and segmenting the content into events via HMMs.
212 Next, we asked whether some correspondence might be made between the specific content of the
213 events the participants experienced in the video, and the events they later recalled. One approach
214 to linking the experienced (video) and recalled events is to label each recalled event as matching
215 the video event with the most similar (i.e., most highly correlated) topic vector (Figs. 2G, S5). This
216 yields a sequence of "presented" events from the original video, and a (potentially differently
217 ordered) sequence of "recalled" events for each participant. Analogous to classic list-learning
218 studies, we can then examine participants' recall sequences by asking which events they tended
219 to recall first (probability of first recall; Fig. 3A; Atkinson and Shiffrin, 1968; Postman and Phillips,
220 1965; Welch and Burnett, 1924); how participants most often transition between recalls of the
221 events as a function of the temporal distance between them (lag-conditional response probability;
222 Fig. 3B; Kahana, 1996); and which events they were likely to remember overall (serial position
223 recall analyses; Fig. 3C; Murdock, 1962). Interestingly, for two of these analyses (probability of
224 first recall and lag-conditional response probability curves) we observed patterns comparable to
225 classic effects from list-learning literature: namely, a higher probability of initiating recall with the
226 first event in the sequence (Fig. 3A) and a higher probability of transitioning to neighboring events
227 with an asymmetric forward bias (Fig. 3B). In contrast, we did not observe a pattern comparable
228 to the serial position effect (Fig. 3C), but rather greater memory for specific events distributed
229 approximately evenly throughout the video.

230 We can also apply two list-learning-native analyses that describe how participants group items
231 in their recall sequences: temporal clustering and semantic clustering (Polyn et al., 2009, see
232 *Methods* for details). Temporal clustering refers to the extent to which participants group their
233 recall responses according to encoding position. Overall, we found that sequentially viewed video
234 events were clustered heavily in participants' recall event sequences (mean clustering score: 0.767,
235 SEM: 0.029), and that participants with higher temporal clustering scores tended to perform better
236 according to both Chen et al. (2017)'s hand-annotated memory scores (Pearson's $r(15) = 0.62$, $p =$

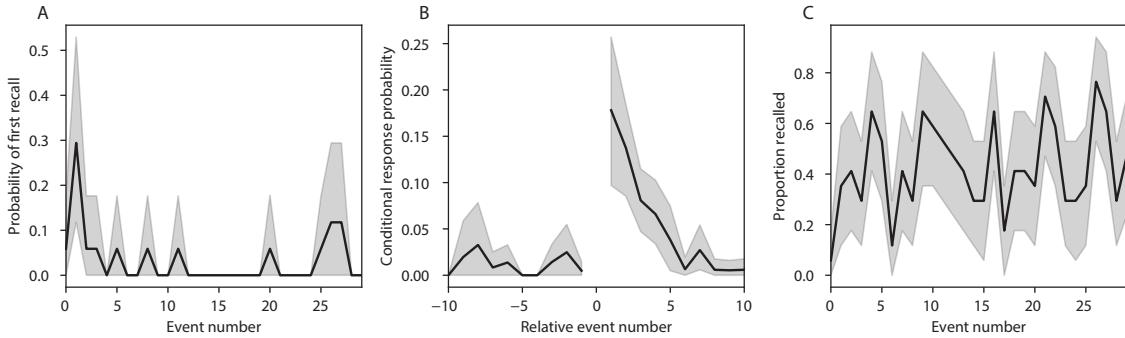


Figure 3: Naturalistic extensions of classic list-learning memory analyses. **A.** The probability of first recall as a function of the serial position of the event in the video. **B.** The probability of recalling each event, conditioned on having most recently recalled the event *lag* events away in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

237 0.008) and our model's estimate (Pearson's $r(15) = 0.54, p = 0.024$). Semantic clustering measures
 238 the extent to which participants cluster their recall responses according to semantic similarity.
 239 We found that participants tended to recall semantically similar video events together (mean
 240 clustering score: 0.787, SEM: 0.018), and that semantic clustering score was also related to both
 241 hand-annotated (Pearson's $r(15) = 0.65, p = 0.004$) and model-derived (Pearson's $r(15) = 0.63, p =$
 242 0.007) memory performance.

243 Statistical models of memory studies often treat recall success as binary (in other words, an
 244 item either was or was not recalled), or occasionally categorical (e.g., to distinguish familiarity
 245 from recollection; Yonelinas et al., 2002). Such approaches are tenable in classical list-learning or
 246 recognition memory paradigms, as the presented stimuli tend to be very simple (e.g., a sequence
 247 of individual words or items). However, memory for naturalistic experiences is much more
 248 nuanced. For example, certain aspects of an experience might be correctly remembered at varying
 249 levels of detail, or distorted, or forgotten entirely. Further, each remembering is itself a richly
 250 structured phenomenon. Our framework produces a content-based model of individual video
 251 and recall events by projecting the dynamic content of the video and participants' recalls into a
 252 shared topic space. This allows for direct, quantitative comparisons between all stimulus and recall
 253 events, as well as between the recall events themselves. Leveraging these content-based models of

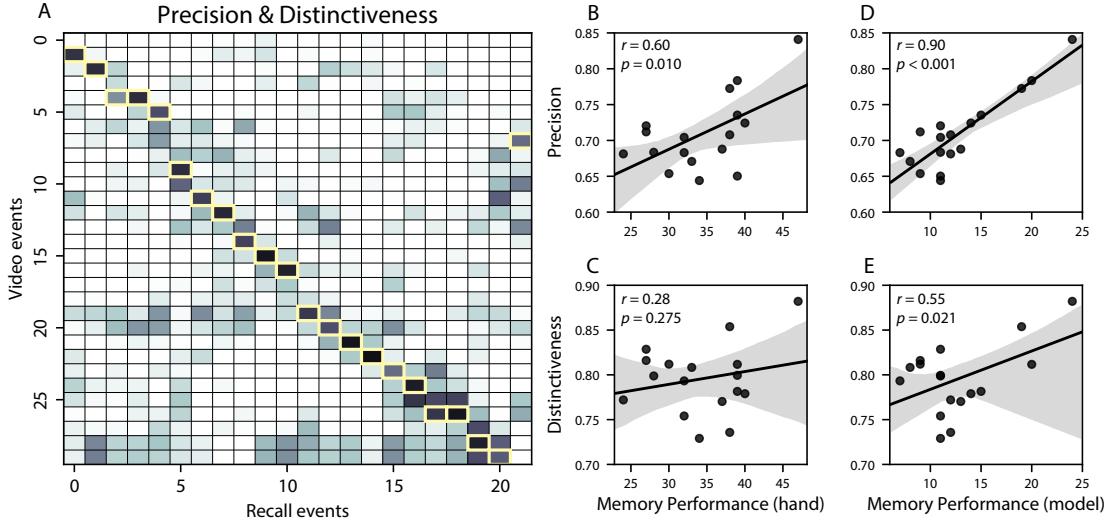


Figure 4: Novel content-based metrics of naturalistic memory: precision and distinctiveness. **A.** The video-recall correlation matrix for a representative participant (17). The yellow boxes highlight the maximum correlation in each column. The example participant's overall precision score was computed as the average across correlation values in the yellow boxes. Their distinctiveness score was computed as the average (over recall events) of 1 minus the average correlation between each recall event and all other recall events that do not display a box in the same row. **B.** The (Pearson's) correlation between precision and hand-annotated memory performance. **C.** The correlation between distinctiveness and hand-annotated memory performance. **D.** The correlation between precision and the number of video events successfully recalled, as determined by our model. **E.** The correlation between distinctiveness and the number of video events successfully recalled, as determined by our model.

the stimulus/recall events, we developed two novel, *continuous* metrics for analyzing naturalistic memory: *precision* and *distinctiveness*. Precision is intended to capture the “completeness” of recall, or how fully the presented content was recapitulated in memory. We define a recall event’s precision as the maximum correlation between the topic proportions of that recall event and any video event (Fig. 4). A second novel metric we introduce here is *distinctiveness*, which is intended to capture the “specificity” of recall. In other words, distinctiveness quantifies the extent to which a given recalled event reflects the most similar presented event more so than it does other presented events. To compute a recall event’s distinctiveness, we first identify the video event to which its topic vector is most strongly correlated. We then define distinctiveness as one minus the average correlation between the given recall event and all *other* video events.

264 In addition to individual events, one may also use these metrics to describe each participant's
265 overall performance by averaging across a participant's event-wise precision or distinctiveness
266 scores. Participants whose recall events are more veridical descriptions of what happened in
267 the video event will presumably have higher precision scores. We find that, across participants,
268 higher precision scores are positively correlated with both hand-annotated memory performance
269 (as collected by Chen et al., 2017; Pearson's $r(15) = 0.60, p = 0.010$) and the number of video events
270 successfully remembered, as determined by our model (Pearson's $r(15) = 0.90, p < 0.001$). We also
271 hypothesized that participants who recounted events in a more distinctive way would display
272 better overall memory. We find that participants' distinctiveness scores were positively correlated
273 with our model's estimated number of recall events (Pearson's $r(15) = 0.55, p = 0.021$). However,
274 we found no evidence that distinctiveness scores were correlated with hand-annotated memory
275 performance (Pearson's $r(15) = 0.28, p = 0.275$). We elaborate on this potential discrepancy in the
276 *Discussion* section.

277 Further intuition for the behaviors captured by these two metrics may be gained by directly
278 examining the content of the video and recalls our framework models. In Figure 5, we contrast
279 recalls for the same video event (event 22) from two participants: one with a high precision score
280 (P17), the other with a low precision score (P6). From the HMM-identified event boundaries,
281 we recovered the set of annotations describing the content of an example video event (Fig. 5B),
282 and divided them into different color-coded sections for each action or feature described. We
283 then similarly recovered the set of sentences comprising the corresponding recall event for each
284 of the two example participants. Because the recall sliding windows overlap heavily, and each
285 recall event spans multiple recall timepoints (i.e., windows), we have stripped any sentences from
286 the beginning and end that describe earlier or later video events for the sake of readability. In
287 other words, Fig. 5C shows a subset of the full recall event text, comprising sentences between
288 the first and last descriptions of content from the example video event. We then colored all words
289 describing actions and features coded in panel B by their corresponding color. Visual comparison
290 of these example transcripts reveals that the more precise recall captures more of the video event's
content, and with more detail.

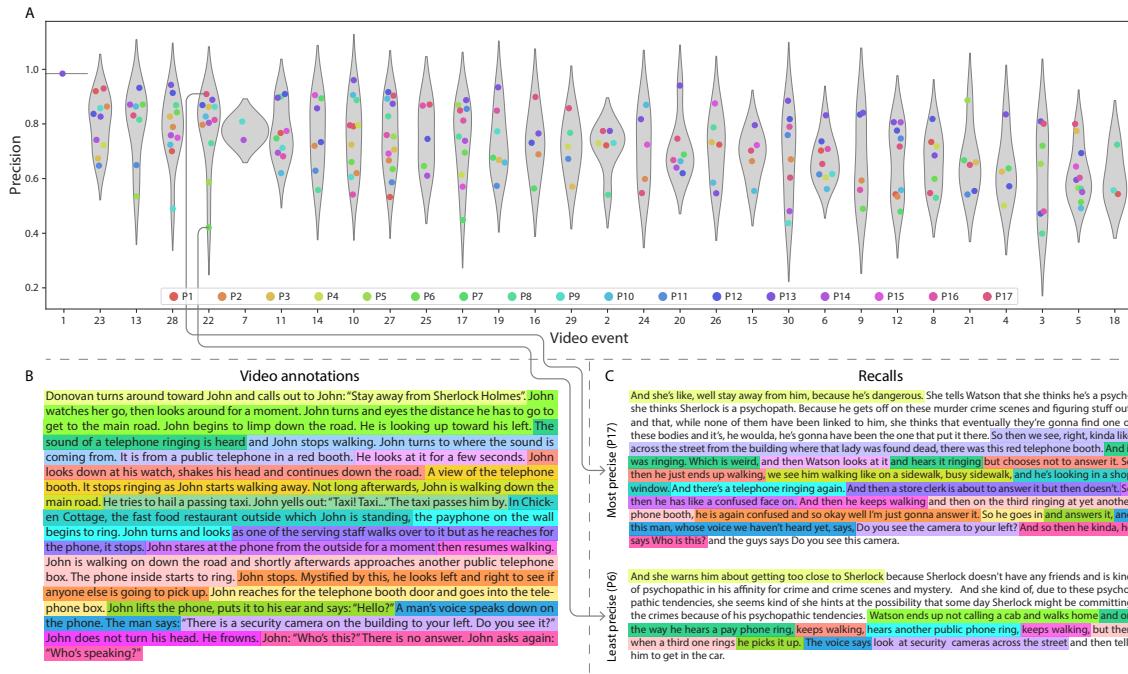


Figure 5: Precision metric reflects completeness of recall. **A.** Recall precision by video event. Grey violin plots display kernel density estimates for the distribution of recall precision scores for a single video event. Colored dots within each violin plot represent individual participants' recall precision for the given event. Video events are ordered along the *x*-axis by the average precision with which they were remembered. **B.** The set of "Narrative Details" video annotations (generated by Chen et al., 2017) for scenes comprising an example video event (22) identified by the HMM. Each action or feature is highlighted in a different color. **C.** A subset of the sentences comprising the most precise (P17) and least precise (P6) participants' recalls of video event 22. Descriptions of specific actions or features reflecting those highlighted in panel B are highlighted in the corresponding color.

292 Figure 6 similarly contrasts two example participants' recalls for a common video event (event
 293 19) to illustrate the tangible differences between high and low distinctiveness scores. Here, we
 294 have extracted the full set of sentences comprising the most distinctive recall event (P13) and least
 295 distinctive recall event (P11) matched to the example video event (Fig. 6C). We also extracted the
 296 annotations for the example video event, as well as those from each other video event whose content
 297 the example participants' single recall events described (Fig. 6B). We then shaded the annotation
 298 text for each video event with a different color, and shaded each word of the example participants'
 299 recall text by the color of the video event it describes. The majority of the most distinctive recall
 300 event text describes video event 19's content, with the first five and last one sentence describing

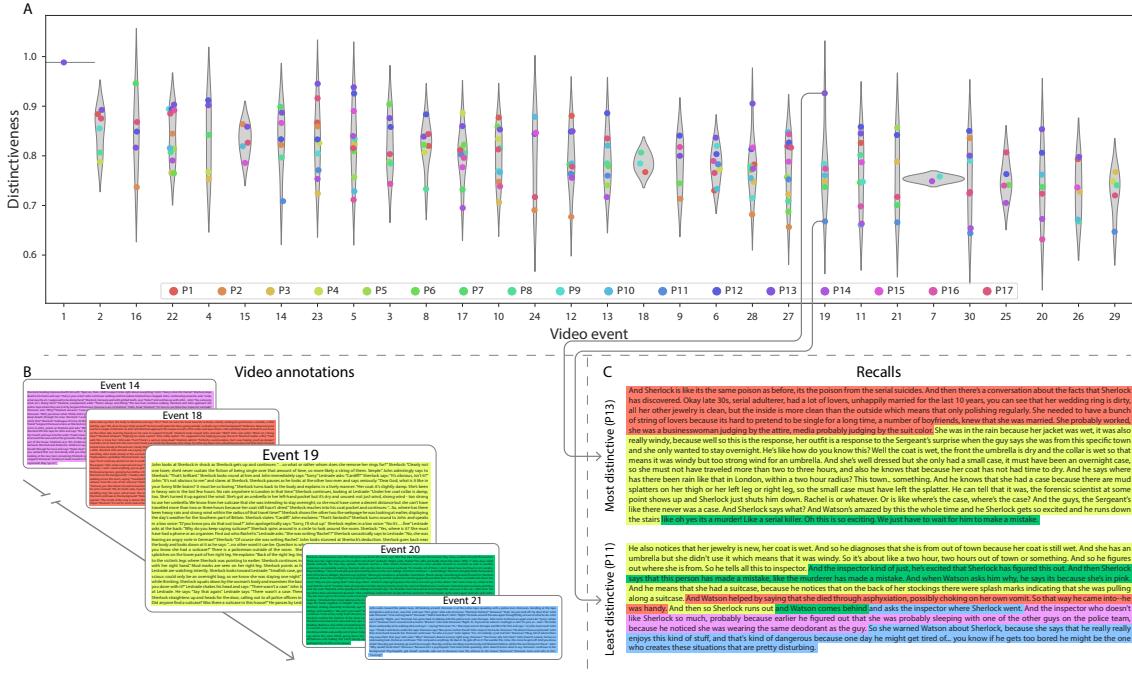


Figure 6: Distinctiveness metric reflects specificity of recall. A. Recall distinctiveness by video event. Kernel density estimates for each video event’s distribution of recall distinctiveness scores, analogous to Fig. 5A. B. The sets of “Narrative Details” video annotations (generated by Chen et al., 2017) for scenes comprising video events described by the example participants in panel C. Each event’s text is highlighted in a different color. C. The sentences comprising the most distinctive (P13) and least distinctive (P11) participants’ recalls of video event 19. Sections of recall describing each video event in panel B are highlighted with the corresponding color.

301 the video events immediately preceding and succeeding the current one, respectively. In contrast,
302 the least distinctive recall of video event 19 blends the content from five separate video events,
303 does not transition between them in order, and often combines descriptions of two video events'
304 content in the same sentence.

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 the content of (or relationships between) experiences and memories solely by examining these topic proportions (e.g., Figs. 2A, D) or the corresponding correlation matrices (Figs. 2B, E, S4). And while we can directly examine the original text underlying these topic vectors (e.g., Figs. 5, 6) to show how relationships between them reflect real-world behavior, this

311 comparison becomes prohibitively cumbersome at larger timescales. To visualize the time-varying
312 high-dimensional content in a more intuitive way (Heusser et al., 2018b), we projected the topic
313 proportions matrices onto a two-dimensional space using Uniform Manifold Approximation and
314 Projection (UMAP; McInnes et al., 2018). In this lower-dimensional space, each point represents a
315 single video or recall event, and the distances between the points reflect the distances between the
316 events' associated topic vectors (Fig. 7). In other words, events that are nearer to each other in this
317 space are more semantically similar, and those that are farther apart are less so.

318 Visual inspection of the video and recall topic trajectories reveals a striking pattern. First, the
319 topic trajectory of the video (which reflects its dynamic content; Fig. 7A) is captured nearly perfectly
320 by the averaged topic trajectories of participants' recalls (Fig. 7B). To assess the consistency of these
321 recall trajectories across participants, we asked: given that a participant's recall trajectory had
322 entered a particular location in the reduced topic space, could the position of their *next* recalled
323 event be predicted reliably? For each location in the the reduced topic space, we computed the set of
324 line segments connecting successively recalled events (across all participants) that intersected that
325 location (see *Methods* for additional details). We then computed (for each location) the distribution
326 of angles formed by the lines defined by those line segments and a fixed reference line (the *x*-
327 axis). Rayleigh tests revealed the set of locations in topic space at which these across-participant
328 distributions exhibited reliable peaks (blue arrows in Fig. 7B reflect significant peaks at $p < 0.05$,
329 corrected). We observed that the locations traversed by nearly the entire video trajectory exhibited
330 such peaks. In other words, participants exhibited similar trajectories that also matched the
331 trajectory of the original video (Fig. 7C). This is especially notable when considering the fact that
332 the number of events participants recalled (dots in Fig. 7C) varied considerably across people, and
333 that every participant used different words to describe what they had remembered happening in
334 the video. Differences in the numbers of remembered events appear in participants' trajectories
335 as differences in the sampling resolution along the trajectory. We note that this framework also
336 provides a means of disentangling classic "proportion recalled" measures (i.e., the proportion
337 of video events described in participants' recalls) from participants' abilities to recapitulate the
338 overall unfolding of the original video's content (i.e., the similarity between the shapes of the

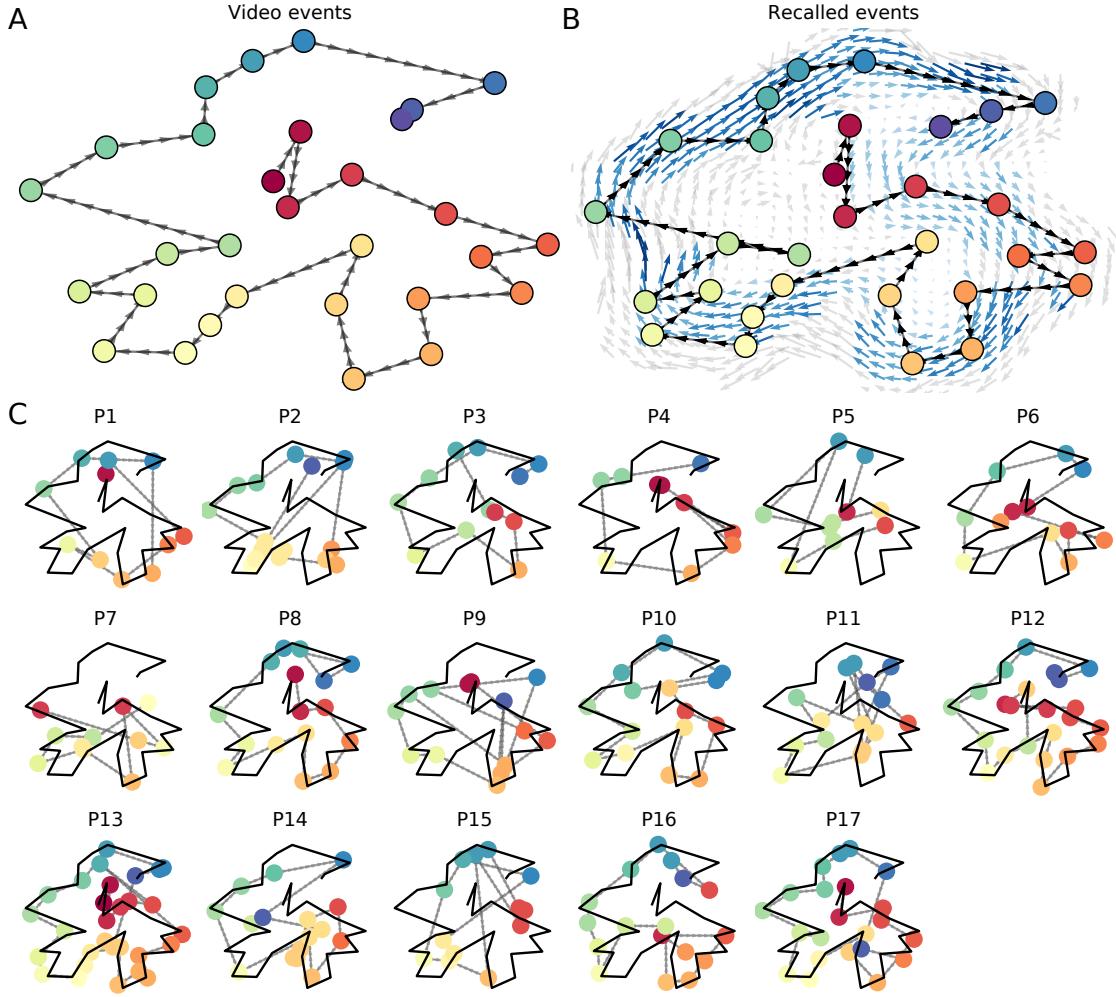


Figure 7: Trajectories through topic space capture the dynamic content of the video and recalls. All panels: the topic proportion matrices have been projected onto a shared two-dimensional space using UMAP. **A.** The two-dimensional topic trajectory taken by the episode of *Sherlock*. Each dot indicates 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. Here, events (dots) are colored by their matched video event (Panel A).

339 original video trajectory and that defined by each participant’s recounting of the video).

340 In addition to the more “holistic” measure of memory described in the previous section, our
341 framework also affords the ability to drill down to individual words and quantify how each word
342 relates to the memorability of each event. The results displayed in Figures 3C and 5A suggest that
343 certain events were remembered better than others. Given this, we next asked whether the
344 events were generally remembered well or poorly tended to reflect particular content. Because
345 our analysis framework projects the dynamic video content and participants’ recalls into a shared
346 space, and because the dimensions of that space represent topics (which are, in turn, sets of
347 weights over known words in the vocabulary), we are able to recover the weighted combination
348 of words that make up any point (i.e., topic vector) in this space. We first computed the average
349 precision with which participants recalled each of the 30 video events (Fig. 8A; note that this result
350 is analogous to a serial position curve created from our continuous recall quality metric). We
351 then computed a weighted average of the topic vectors for each video event, where the weights
352 reflected how reliably each event was recalled. To visualize the result, we created a “wordle”
353 image (Mueller et al., 2018) where words weighted more heavily by better-remembered topics
354 appear in a larger font (Fig. 8B, green box). Across the full video, content that reflected topics
355 necessary to convey the central focus of the video (e.g., the names of the two main characters,
356 “Sherlock” and “John,” and the address of a major recurring location, “221B Baker Street”) were
357 best remembered. An analogous analysis revealed which themes were poorly remembered. Here
358 in computing the weighted average over events’ topic vectors, we weighted each event in *inverse*
359 proportion to how well it was remembered (Fig. 8B, red box). The least well-remembered video
360 content reflected information not necessary to later convey a general summary of the video, such
361 as the proper names of relatively minor characters (e.g., “Mike,” “Molly,” and “Lestrade”) and
362 locations (e.g., “St. Bartholomew’s Hospital”).

363 A similar result emerged from assessing the topic vectors for individual video and recall events
364 (Fig. 8C). Here, for each of the three best- and worst-remembered video events, we have constructed
365 two wordles: one from the original video event’s topic vector (left) and a second from the average
366 recall topic vector for that event (right). The three best-remembered events (circled in green)

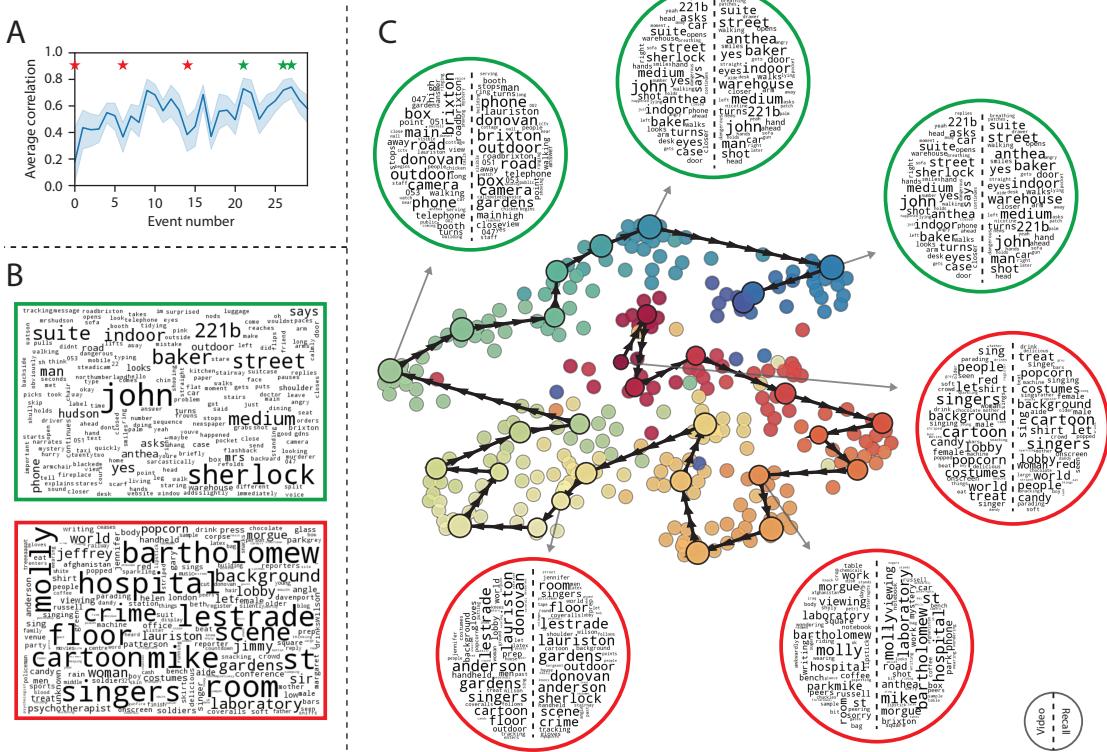


Figure 8: Language used in the most and least memorable events. **A.** Average precision (video event-recall event topic vector correlation) across participants for each video event. Error bars denote bootstrap-derived across-participant 95% confidence intervals. The stars denote the three best-remembered 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 7. 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 7A. 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.

367 correspond to scenes integral to the central plot-line: a mysterious figure spying on John in a
368 phone booth; John meeting Sherlock at Baker St. to discuss the murders; and Sherlock laying
369 a trap to catch the killer. Meanwhile, the three worst-remembered events (circled in red) reflect
370 scenes that are non-essential to summarizing the narrative's structure: the video of singing cartoon
371 characters participants viewed in an introductory clip prior to the main episode; John asking Molly
372 about Sherlock's habit of over-analyzing people; and Sherlock noticing evidence of Anderson's
373 and Donovan's affair.

374 The results thus far inform us about which aspects of the dynamic content in the episode partic-
375 ipants watched were preserved or altered in participants' memories. We next carried out a series
376 of analyses aimed at understanding which brain structures might facilitate these preservations
377 and transformations between the external world and memory. In the first analysis, we sought
378 to identify brain structures that were sensitive to the dynamic unfolding of the video's content,
379 as characterized by its topic trajectory. We used a searchlight procedure to identify clusters of
380 voxels whose activity patterns displayed a proximal temporal correlation structure (as participants
381 watched the video) matching that of the original video's topic proportions (Fig. 9A; see *Methods* for
382 additional details). In a second analysis, we sought to identify brain structures whose responses
383 (during video viewing) reflected how each participant would later structure their recounting of the
384 video. We used an analogous searchlight procedure to identify clusters of voxels whose proximal
385 temporal correlation matrices matched that of the topic proportions for each individual's recall
386 (Figs. 9B; see *Methods* for additional details). To ensure our searchlight procedure identified re-
387 gions *specifically* sensitive to the temporal structure of the video or recalls (i.e., rather than those
388 with a temporal autocorrelation length similar to that of the video/recalls), we performed a phase
389 shift-based permutation correction (see *Methods* for additional details). As shown in Figure 9C, the
390 video-driven searchlight analysis revealed a distributed network of regions that may play a role in
391 processing information relevant to the narrative structure of the video. Similarly, the recall-driven
392 searchlight analysis revealed a second network of regions (Fig. 9D) that may facilitate a person-
393 specific transformation of one's experience into memory. In identifying regions whose responses
394 to ongoing experiences reflect how those experiences will be remembered later, this latter analysis

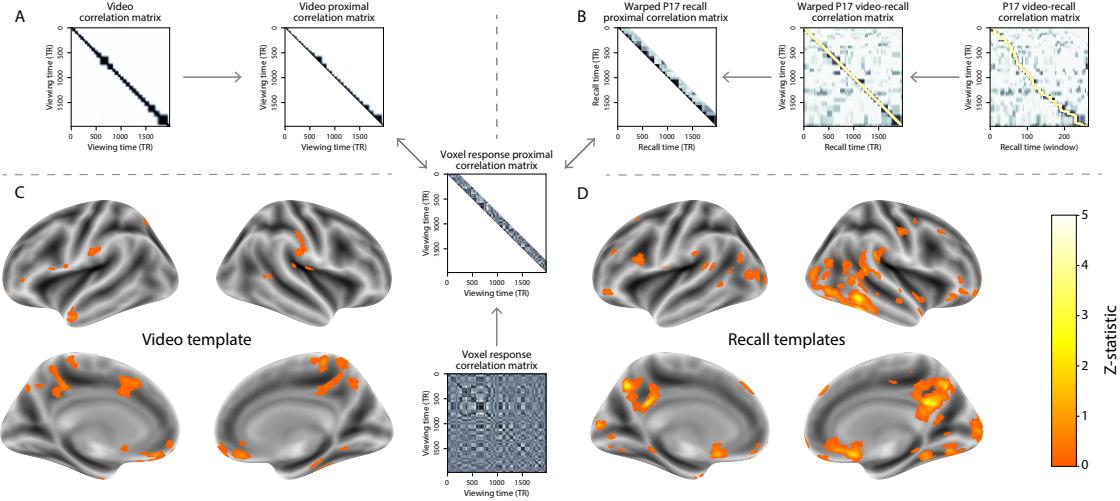


Figure 9: Brain structures that underlie the transformation of experience into memory. **A.** We isolated the proximal diagonals from the upper triangle of the video correlation matrix, and applied this same diagonal mask to the voxel response correlation matrix for each cube of voxels in the brain. We then searched for brain regions whose activation timeseries consistently exhibited a similar proximal correlational structure to the video model, across participants. **B.** We used dynamic time warping (Berndt and Clifford, 1994) to align each participant's recall timeseries to the TR timeseries of the video. We then applied the same diagonal mask used in Panel A to isolate the proximal temporal correlations and searched for brain regions whose activation timeseries for an individual consistently exhibited a similar proximal correlational structure to each individual's recall. **C.** We identified a network of regions sensitive to the narrative structure of participants' ongoing experience. The map shown is thresholded at $p < 0.05$, corrected. **D.** We also identified a network of regions sensitive to how individuals would later structure the video's content in their recalls. The map shown is thresholded at $p < 0.05$, corrected.

395 extends classic *subsequent memory analyses* (e.g., Paller and Wagner, 2002) to domain of naturalistic
 396 stimuli.

397 The searchlight analyses described above yielded two distributed networks of brain regions,
 398 whose activity timecourses mirrored to the temporal structure of the video (Fig. 9C) or participants'
 399 eventual recalls (Fig. 9D). We next sought to gain greater insight into the structures and functional
 400 networks our results reflected. To accomplish this, we performed an additional, exploratory
 401 analysis using Neurosynth (Yarkoni et al., 2011). Given an arbitrary statistical map as input,
 402 Neurosynth performs a massive automated meta-analysis, returning a ranked list of terms reported
 403 in papers with similar significance maps. We ran Neurosynth on the significance maps for the video-
 404 and recall-driven searchlight analyses. These maps, along with the 10 terms with maximally similar

⁴⁰⁵ meta-analysis images identified by Neurosynth are shown in Figure 10.

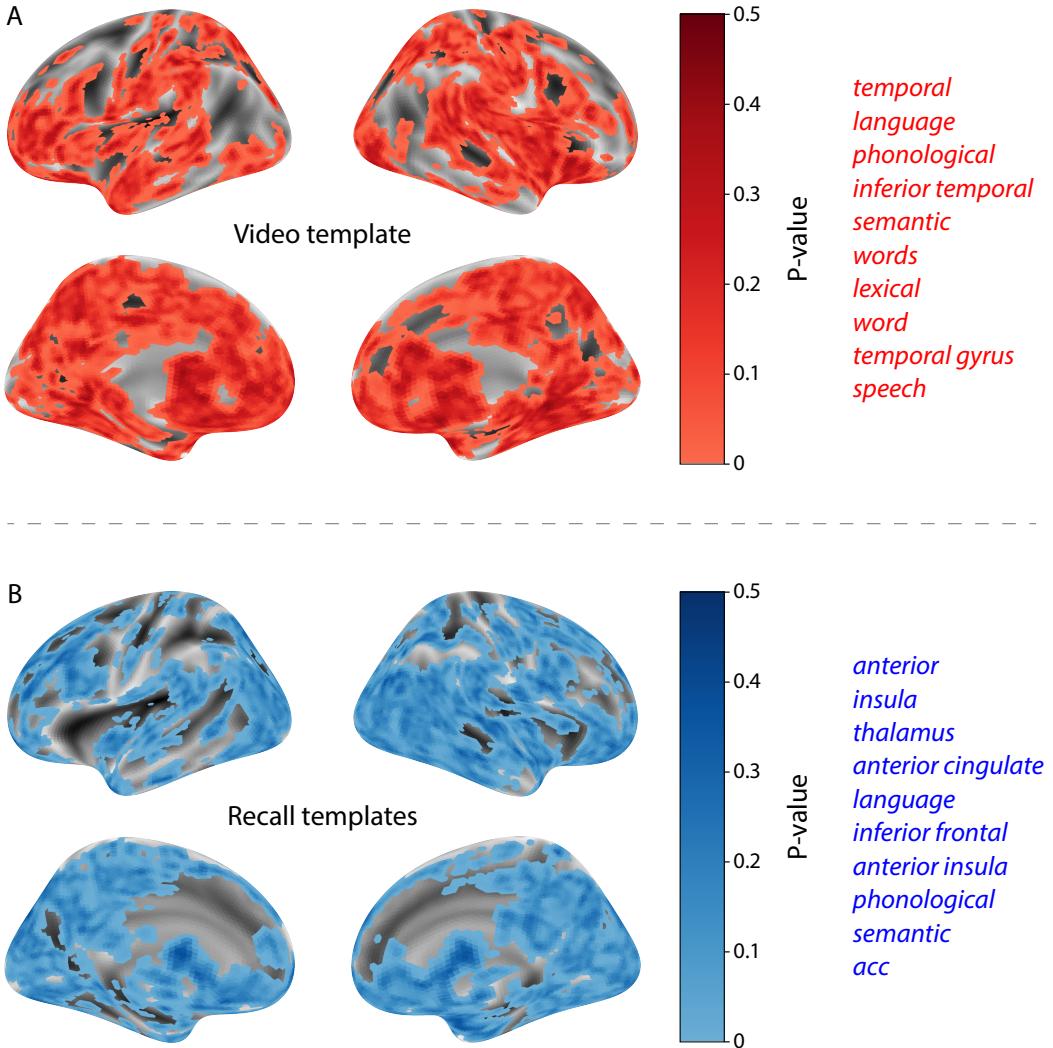


Figure 10: Decoding distributed statistical maps via Neurosynth meta-analyses. **A.** Video-searchlight significance and top 10 decoded terms. We constructed a map of the permutation-derived p -values for the video-driven searchlight analysis (Fig. 9A, C) at each voxel with a positive permutation-derived z -score. The top 10 terms decoded from this significance map are shown in red. **B.** Recall-searchlight significance and top 10 decoded terms. We constructed a map of the permutation-derived p -values for the recall-driven searchlight analysis (Fig. 9A, C) at each voxel with a positive permutation-derived z -score. The top 10 terms decoded from this significance map are shown in blue.

406 **Discussion**

407 Our work casts remembering as reproducing (behaviorally and neurally) the topic trajectory, or
408 shape, of an experience. This view draws inspiration from prior work aimed at elucidating
409 the neural and behavioral underpinnings of how we process dynamic naturalistic experiences
410 and remember them later. One approach to identifying neural responses to naturalistic stimuli
411 (including experiences) entails building a model of the stimulus and searching for brain regions
412 whose responses are consistent with the model. In prior work, a series of studies from Uri Hasson's
413 group (Lerner et al., 2011; Simony et al., 2016; Chen et al., 2017; Baldassano et al., 2017; Zadbood
414 et al., 2017) have extended this approach with a clever twist: rather than building an explicit
415 stimulus model, these studies instead search for brain responses (while experiencing the stimulus)
416 that are reliably similar across individuals. So called *inter-subject correlation* (ISC) and *inter-subject*
417 *functional connectivity* (ISFC) analyses effectively treat other people's brain responses to the stimulus
418 as a "model" of how its features change over time. By contrast, in our present work, we use topic
419 models to construct an explicit content model directly from the stimulus (i.e., the topic trajectory
420 of the video). Projecting each participant's recall into a space shared by both the stimulus and
421 other participants then allows us to compare recalls both directly to the stimulus and to each other.
422 Similarly, prior work introducing the use of HMMs to discover latent event structure in naturalistic
423 stimuli and recall (Baldassano et al., 2017) used between-subjects cross-validation to identify event
424 boundaries shared across participants, and between stimulus and recall. Our framework allows
425 us to break from the restriction of a common, shared event-timeseries and identify the unique
426 *resolution* of each participant's recall event structure, and how that may differ from the video and
427 that of other participants.

428 Word embedding models are a rapidly growing area of machine learning research. Early ap-
429 proaches including latent semantic analysis (Landauer and Dumais, 1997) use word co-occurrence
430 statistics (i.e., how often pairs of words occur in the same documents contained in the corpus) to
431 derive a unique feature vector for each word. The feature vectors are constructed so that words
432 that co-occur more frequently have feature vectors that are closer (in Euclidean distance). Related

433 approaches, such as latent dirichlet allocation (Blei et al., 2003) attempt to explicitly model the
434 underlying causes of word co-occurrences by automatically identifying the set of themes or topics
435 reflected across the documents in the corpus. More recent work on these types of semantic mod-
436 els, including word2vec (Mikolov et al., 2013), the Universal Sentence Encoder (Cer et al., 2018),
437 GPT-2 (Radford et al., 2019), and GTP-3 (Brown et al., 2020) use deep neural networks to attempt
438 to identify the deeper conceptual representations underlying each word. Despite the growing
439 popularity of more sophisticated deep learning-based embedding models, here we leverage latent
440 dirichlet allocation (i.e., topic modelling) to embed video and recall text. This decision was mo-
441 tivated by several factors. First, topic models capture the *essence* of a text passage devoid of the
442 specific set and order of words used. This was an important feature of our model since different
443 people may accurately recall a scene using very different language. Second, words can mean dif-
444 ferent things in different contexts (e.g. “bat” may be the act of hitting a baseball, the object used for
445 that action, or as a flying mammal). Topic models are robust to this, allowing words to exist as part
446 of multiple topics. Last, topic models provide a straightforward means of recovering the weights
447 for the particular words comprising a topic, enabling straightforward interpretation of an event’s
448 contents (e.g. Fig. 8). Other models such as the Universal Sentence Encoder, GPT-2, and GPT-3
449 offer context-sensitive encoding of text passages, but the encoding space is complex and non-linear,
450 and thus recovering the original words used to fit the model is not straightforward. However, it is
451 worth pointing out that our general framework is divorced from the particular choice of language
452 model. Moreover, many of the aspects of our framework could be swapped out for other choices.
453 For example, the language model, the timeseries segmentation model and the video-recall match-
454 ing function could all be customized to suit a particular question space or application. Indeed for
455 some questions, recovery of the particular words used to describe a memory may not be necessary,
456 and thus other text-modeling approaches (including the deep learning-based embedding models
457 described above) may be preferable. Future work will explore the influence of particular model
458 choices on the framework’s efficacy.

459 In extending classical free recall analyses to our naturalistic memory framework, we recovered
460 two patterns of recall dynamics central to list-learning studies: a heightened probability of initiating

461 recall with the first presented “item” (in our case, video events; Fig. 3A) and a strong bias toward
462 transitioning from recalling a given event to recalling the one immediately following it (Fig. 3B).
463 However, equally noteworthy are the typical free recall results *not* recovered in these analyses,
464 as each highlights a fundamental difference between the list-learning paradigm and naturalistic
465 memory paradigms like the one employed in the present study. The most noticeable departure
466 from hallmark free recall dynamics in these findings is the apparent lack of a serial position effect in
467 Figure 3C, which instead shows greater and lesser recall probabilities for events distributed across
468 the video. Stimuli in free recall experiments most often comprise lists of simple, common words,
469 presented to participants in a random order. (In fact, numerous word pools have been developed
470 based on these criteria; e.g., Friendly et al., 1982). These stimulus qualities enable two assumptions
471 that are central to word list analyses, but frequently do not hold for real-world experiences. First,
472 researchers conducting list-learning studies may assume that the content at each presentation index
473 is essentially equal, and does not possess attributes that would render it, on average, more or less
474 memorable than others. Such is rarely the case with real-world experiences or experiments meant
475 to approximate them, and the effects of both intrinsic and observer-dependent factors on stimulus
476 memorability are well established (for review see Chun and Turk-Browne, 2007; Bylinskii et al.,
477 2015; Tyng et al., 2017). Second, the random ordering of list items ensures that (across participants,
478 on average) there is no relationship between the thematic similarity of individual stimuli and their
479 presentation positions—in other words, two successively presented items are no more likely to be
480 highly semantically similar than they are to be highly dissimilar. In most cases, the exact opposite
481 is true of real-world episodes. Our internal thoughts, our actions, and the physical state of the
482 world around us all tend to follow a direct (often causal) progression. As a result, each moment
483 of our experience tends to be inherently more similar to surrounding moments than to those in
484 the distant past or future. Memory literature has termed this strong temporal autocorrelation
485 “context,” and in various media that depict real-world events (e.g., movies or written stories),
486 we recognize it as a *narrative structure*. While a random word list (by definition) has no such
487 structure, the logical progression between ideas and actions in a naturalistic stimulus prompts the
488 rememberer to recount presented events in order, starting with the beginning. This tendency is

489 reflected in our findings' second departure from typical free recall dynamics: a lack of increased
490 probability of first recall for end-of-sequence events (Fig. 3A).

491 Because they disregard presentation order-dependent variability in the stimulus content, anal-
492 yses such as those in Figure 3 enable a more sensitive analysis of presentation order-dependent
493 temporal dynamics in free recall. Yet by the same token, they paint a wholly incomplete picture of
494 memory for naturalistic episodes. In an attempt to address this shortcoming, we have developed a
495 framework in the present study that characterizes the explicit semantic content of the stimulus and
496 subsequent recalls. However, sensitivity to stimulus and recall content introduces a new challenge:
497 distinguishing between levels of recall quality for a stimulus (e.g., an event) that is considered to
498 have been "remembered." When modeling memory in an experimental setting, recall quality for
499 individual events is often cast as binary (e.g., a given list item was simply either remembered or
500 not remembered). Various models of memory (e.g., Yonelinas, 2002) attempt to improve upon this
501 by including confidence ratings, rendering this binary judgement instead categorical. To better
502 evaluate naturalistic memory quality, we introduce a continuous metric (*precision*), which reflects
503 the level of completeness of a participant's recall for a feature-rich experience. Additionally, recall
504 quality for a single event is typically assessed independently from that for all other events (e.g., it
505 is difficult to "compare" a participant's binary recall success for list item 1 to that of list item 10).
506 The second novel metric we introduce (*distinctiveness*) is based on analyzing of the correlational
507 structure of an individual's full set of recall events, and reflects the specificity of their memory for
508 a single experienced event. We find that both of these metrics relate to the overall number of video
509 events participants successfully recalled, and that our precision metric additionally relates to Chen
510 et al. (2017)'s hand-annotated memory memory scores.

511 We did not find evidence that participants' average recall distinctiveness was related to their
512 hand-annotated memory scores computed by Chen et al. (2017). One possible explanation is that,
513 in hand-scoring each participant's verbal recall for each of 50 (manually-delimited) scenes, "[a]
514 scene was counted as recalled if the participant described any part of the scene" (Chen et al.,
515 2017). In other words, both an extensive description of a scene's content and a brief mention of
516 some subset of its content were (binarily) considered equally successful recalls. By contrast, we

517 identify the event structure in participants' recalls in an unsupervised manner, independent of
518 the video event-timeseries, prior to mapping between video and recall content. Our HMM-based
519 event-segmentation produces boundaries between timepoints where the topic proportions shift in
520 a substantial way, and because a small handful of words is unlikely to contribute significantly to
521 the topic proportions for any sliding window, such brief scene descriptions will most often not
522 result in a sufficiently large shift in the resulting topic proportions for the HMM to identify an
523 event boundary. Instead, they will be grouped with a neighboring event, consequently lowering
524 that event's distinctiveness score and by extension, the participant's overall distinctiveness score.
525 This is in essence the qualitative difference between distinctive and indistinctive recall, and reflects
526 the comparison shown in Figure 6C. Intriguingly, prior studies show that pattern separation, or the
527 ability to cleanly discriminate between similar experiences, is impaired in many cognitive disorders
528 as well as natural aging (Stark et al., 2010; Yassa et al., 2011; Yassa and Stark, 2011). Future work
529 might explore whether and how these metrics compare between cognitively impoverished groups
530 and healthy controls.

531 In the analyses outlined in Figure 9, we identified two networks of brain regions whose re-
532 sponses during video viewing were consistent with the temporal structure of the video and recall
533 topic trajectories, respectively. The network identified by the video trajectory analysis included the
534 ventromedial prefrontal cortex, left anterior temporal lobe, superior parietal and dorsal anterior
535 cingulate cortex. The network from the video-recall trajectory analysis also included the ventro-
536 medial prefrontal and superior parietal cortices, in addition to the posterior medial cortex (PMC)
537 and the inferior temporal regions. Notably, Chen et al. (2017) also observed the PMC in a number
538 of analyses including one that searched for regions whose activity patterns during encoding were
539 reinstated during free recall. The PMC has been consistently identified in studies involving stimuli
540 with meaningfully structured events (Cohn-Sheely and Ranganath, 2017). Further, the PMC is
541 part of the "posterior medial" system, a network of brain regions thought to represent situation
542 models (Zacks et al., 2007) in support of memory, spatial navigation and social cognition (Ran-
543 ganath and Ritchey, 2012). Given that we constructed our video-recall searchlight model to capture
544 temporal structure in the episode's semantic content (and how one's later recall aligns with that

545 structure), we speculate that the PMC may play a role in constructing mnemonic events from
546 meaningfully structured experiences.

547 Decoding the associated significance maps with Neurosynth revealed two intriguing results.
548 First, the top 10 terms returned for the video-driven searchlight significance map were centered
549 around themes of language and semantic meaning (Fig. 10A). In other words, the voxels identified
550 as more reflective of the video content's temporal structure (i.e., voxels with lower permutation
551 correction-derived p -values), as defined by our model, were most likely to be reported as active in
552 studies focused on the the neural underpinnings of semantic processing. This finding is interesting,
553 as our model specifically captures the temporal structure of the video's *semantic* content (e.g., as
554 opposed to that of the visual, auditory, or affective content). This suggests that the network of
555 structures displayed in Figure 9C may play a roll in processing the evolving semantic content of
556 ongoing experiences.

557 Our second searchlight analysis identified a partially overlapping network of regions (Fig. 9D)
558 whose patterns of activity as participants viewed the video reflected the idiosyncratic structure of
559 each individual's later recalls. The associated significance map yielded a set of Neurosynth terms
560 that primarily reflected names of specific structural regions (such as "thalamus," "anterior insula,"
561 "anterior cingulate" and "inferior frontal"; Fig. 10B). Interestingly, these regions share membership
562 in a common, large-scale functional network (termed the "salience network") involved in detect-
563 ing and processing affective cues. In particular, the latter three regions have been implicated in
564 functions relevant to assigning personal meaning to an experience, including: ascribing subjective
565 value to raw, sensory input (Medford and Critchley, 2010); modulating semantic and phonological
566 processing in response to personally salient stimuli (Kelly et al., 2007); and directing and reallo-
567 cating attention and working memory resources towards the most relevant stimuli (Menon and
568 Uddin, 2010). This suggests that the network of structures displayed in Figure 9D may be play a roll
569 in transforming and restructuring ongoing experiences through the lens of one's prior experience
570 and subjective emotions as they are encoded in memory.

571 Our work has broad implications for how we characterize and assess memory in real-world
572 settings, such as the classroom or physician's office. For example, the most commonly used

573 classroom evaluation tools involve simply computing the proportion of correctly answered exam
574 questions. Our work indicates that this approach is only loosely related to what educators might
575 really want to measure: how well did the students understand the key ideas presented in the
576 course? Under this typical framework of assessment, the same exam score of 50% could be
577 ascribed to two very different students: one who attended the full course but struggled to learn
578 more than a broad overview of the material, and one who attended only half of the course but
579 understood the material perfectly. Instead, one could apply our computational framework to build
580 explicit content models of the course material and exam questions. This approach would provide
581 a more nuanced and specific view into which aspects of the material students had learned well
582 (or poorly). In clinical settings, memory measures that incorporate such explicit content models
583 might also provide more direct evaluations of patients' memories.

584 Methods

585 Experimental design and data collection

586 Data were collected by Chen et al. (2017). In brief, participants ($n = 22$) viewed the first 48 minutes
587 of "A Study in Pink", the first episode of the BBC television series *Sherlock*, while fMRI volumes
588 were collected (TR = 1500 ms). Participants were pre-screened to ensure they had never seen any
589 episode of the show before. The stimulus was divided into a 23 min (946 TR) and a 25 min (1030 TR)
590 segment to mitigate technical issues related to the scanner. After finishing the clip, participants
591 were instructed to (quoting from Chen et al., 2017) "describe what they recalled of the [episode]
592 in as much detail as they could, to try to recount events in the original order they were viewed
593 in, and to speak for at least 10 minutes if possible but that longer was better. They were told that
594 completeness and detail were more important than temporal order, and that if at any point they
595 realized they had missed something, to return to it. Participants were then allowed to speak for
596 as long as they wished, and verbally indicated when they were finished (e.g., 'I'm done')." Five
597 participants were dropped from the original dataset due to excessive head motion (2 participants),

598 insufficient recall length (2 participants), or falling asleep during stimulus viewing (1 participant),
599 resulting in a final sample size of $n = 17$. For additional details about the experimental procedure
600 and scanning parameters, see Chen et al. (2017). The experimental protocol was approved by
601 Princeton University's Institutional Review Board.

602 After preprocessing the fMRI data and warping the images into a standard (3 mm³ MNI) space,
603 the voxel activations were z-scored (within voxel) and spatially smoothed using a 6 mm (full width
604 at half maximum) Gaussian kernel. The fMRI data were also cropped so that all video-viewing
605 data were aligned across participants. This included a constant 3 TR (4.5 s) shift to account for the
606 lag in the hemodynamic response. (All of these preprocessing steps followed Chen et al., 2017,
607 where additional details may be found.)

608 The video stimulus was divided into 1,000 fine-grained “scenes” and annotated by an inde-
609 pendent coder. For each of these 1,000 scenes, the following information was recorded: a brief
610 narrative description of what was happening, the location where the scene took place, whether
611 that location was indoors or outdoors, the names of all characters on-screen, the name(s) of the
612 character(s) in focus in the shot, the name(s) of the character(s) currently speaking, the camera
613 angle of the shot, a transcription of any text appearing on-screen, and whether or not there was
614 music present in the background. Each scene was also tagged with its onset and offset time, in
615 both seconds and TRs.

616 **Data and code availability**

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

619 **Statistics**

620 All statistical tests performed in the behavioral analyses were two-sided. All statistical tests per-
621 formed in the neural data analyses were two-sided, except for the permutation-based thresholding,
622 which was one-sided. In this case, we were specifically interested in identifying voxels whose ac-

623 tivation time series reflected the temporal structure of the video and recall trajectories to a *greater*
624 extent than that of the phase-shifted trajectories.

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

626 **Topic modeling**

627 The input to the topic model we trained to characterize the dynamic content of the video comprised
628 998 hand-generated annotations of short (mean: 2.96s) scenes spanning the video clip (Chen
629 et al., 2017 generated 1000 annotations total; we removed two annotations referring to a break
630 between the first and second scan sessions, during which no fMRI data was collected). We
631 concatenated the text for all of the annotated features within each segment, creating a “bag of
632 words” describing each scene and performed some minor preprocessing (e.g., stemming possessive
633 nouns and removing punctuation). We then re-organized the text descriptions into overlapping
634 sliding windows spanning (up to) 50 scenes each. In other words, we estimated the “context”
635 for each scene using the text descriptions of the preceding 25 scenes, the present scene, and the
636 following 24 scenes. To model the context for scenes near the beginning of the video (i.e., within
637 25 scenes of the beginning or end), we created overlapping sliding windows that grew in size
638 from one scene to the full length. We also tapered the sliding window lengths at the end of the
639 video, whereby scenes within fewer than 24 scenes of the end of the video were assigned sliding
640 windows that extended to the end of the video. This procedure ensured that each scene’s content
641 was represented in the text corpus an equal number of times.

642 We trained our model using these overlapping text samples with `scikit-learn` (version 0.19.1;
643 Pedregosa et al., 2011), called from our high-dimensional visualization and text analysis software,
644 `HyperTools` (Heusser et al., 2018b). Specifically, we used the `CountVectorizer` class to transform
645 the text from each window into a vector of word counts (using the union of all words across all
646 scenes as the “vocabulary,” excluding English stop words); this yielded a number-of-windows
647 by number-of-words *word count* matrix. We then used the `LatentDirichletAllocation` class
648 (`topics=100, method='batch'`) to fit a topic model (Blei et al., 2003) to the word count matrix,

649 yielding a number-of-windows (1047) by number-of-topics (100) *topic proportions* matrix. The
650 topic proportions matrix describes the gradually evolving mix of topics (latent themes) present in
651 each scene. Next, we transformed the topic proportions matrix to match the 1976 fMRI volume
652 acquisition times. We assigned each topic vector to the timepoint (in seconds) midway between the
653 beginning of the first scene and the end of the last scene in its corresponding sliding text window.
654 By doing so, we warped the linear temporal distance between consecutive topic vectors to align
655 with the inconsistent temporal distance between consecutive annotations (whose durations varied
656 greatly). We then rescaled these timepoints to 1.5s TR units, and used linear interpolation to
657 estimate a topic vector for each TR. This resulted in a number-of-TRs (1976) by number-of-topics
658 (100) matrix.

659 We created similar topic proportions matrices using hand-annotated transcripts of each partici-
660 pant's verbal recall of the video (annotated by Chen et al., 2017). We tokenized the transcript into a
661 list of sentences, and then re-organized the list into overlapping sliding windows spanning (up to)
662 10 sentences each, analogously to how we parsed the video annotations. In turn, we transformed
663 each window's sentences into a word count vector (using the same vocabulary as for the video
664 model), and then we used the topic model already trained on the video scenes to compute the most
665 probable topic proportions for each sliding window. This yielded a number-of-windows (range:
666 83–312) by number-of-topics (100) topic proportions matrix for each participant. These reflected
667 the dynamic content of each participant's recalls. Note: for details on how we selected the video
668 and recall window lengths and number of topics, see *Supporting Information* and Figure S1.

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

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

676 to implement this segmentation.

677 We used an optimization procedure to select the appropriate K for each topic proportions
678 matrix. Prior studies on narrative structure and processing have shown that we both perceive
679 and internally represent the world around us at multiple, hierarchical timescales (e.g., Hasson
680 et al., 2008; Lerner et al., 2011; Hasson et al., 2015; Chen et al., 2017; Baldassano et al., 2017, 2018).
681 However, for the purposes of our framework, we sought to identify the single timeseries of event-
682 representations that is emphasized *most heavily* in the temporal structure of the video and of each
683 participant's recall. We quantified this as the set of K states that maximized the similarity between
684 topic vectors for timepoints comprising each state, while minimizing the similarity between topic
685 vectors for timepoints across different states. Specifically, we computed (for each matrix)

$$\operatorname{argmax}_K [W_1(a, b)],$$

686 where a was the distribution of within-state topic vector correlations, and b was the distribution of
687 across-state topic vector correlations . We computed the first Wasserstein distance (W_1 ; also known
688 as *Earth mover's distance*; Dobrushin, 1970; Ramdas et al., 2017) between these distributions for a
689 large range of possible K -values (range [2, 50]), and selected the K that yielded the maximum value.
690 Figure 2B displays the event boundaries returned for the video, and Figure S4 displays the event
691 boundaries returned for each participant's recalls. See Figure S6 for the optimization functions
692 for the video and recalls. After obtaining these event boundaries, we created stable estimates of
693 the content represented in each event by averaging the topic vectors across timepoints between
694 each pair of event boundaries. This yielded a number-of-events by number-of-topics matrix for
695 the video and recalls from each participant.

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

697 In traditional list-learning experiments, participants view a list of items (e.g., words) and then recall
698 the items later. Our video-recall event matching approach affords us the ability to analyze memory
699 in a similar way. The video and recall events can be treated analogously to studied and recalled

700 “items” in a list-learning study. We can then extend classic analyses of memory performance and
701 dynamics (originally designed for list-learning experiments) to the more naturalistic video recall
702 task used in this study.

703 Perhaps the simplest and most widely used measure of memory performance is *accuracy*—i.e.,
704 the proportion of studied (experienced) items (in this case, video events) that the participant later
705 remembered. Chen et al. (2017) used this method to rate each participant’s memory quality by
706 computing the proportion of (50, manually identified) scenes mentioned in their recall. We found a
707 strong across-participants correlation between these independent ratings and the proportion of 30
708 HMM-identified video events matched to participants’ recalls (Pearson’s $r(15) = 0.71, p = 0.002$).
709 We further considered a number of more nuanced memory performance measures that are typically
710 associated with list-learning studies. We also provide a software package, Quail, for carrying out
711 these analyses (Heusser et al., 2017).

712 **Probability of first recall (PFR).** PFR curves (Welch and Burnett, 1924; Postman and Phillips,
713 1965; Atkinson and Shiffrin, 1968) reflect the probability that an item will be recalled first as a
714 function of its serial position during encoding. To carry out this analysis, we initialized a number-
715 of-participants (17) by number-of-video-events (30) matrix of zeros. Then for each participant, we
716 found the index of the video event that was recalled first (i.e., the video event whose topic vector
717 was most strongly correlated with that of the first recall event) and filled in that index in the matrix
718 with a 1. Finally, we averaged over the rows of the matrix, resulting in a 1 by 30 array representing
719 the proportion of participants that recalled an event first, as a function of the order of the event’s
720 appearance in the video (Fig. 3A).

721 **Lag conditional probability curve (lag-CRP).** The lag-CRP curve (Kahana, 1996) reflects the
722 probability of recalling a given item after the just-recalled item, as a function of their relative
723 encoding positions (or *lag*). In other words, a lag of 1 indicates that a recalled item was presented
724 immediately after the previously recalled item, and a lag of -3 indicates that a recalled item came 3
725 items before the previously recalled item. For each recall transition (following the first recall), we

726 computed the lag between the current recall event and the next recall event, normalizing by the
727 total number of possible transitions. This yielded a number-of-participants (17) by number-of-lags
728 (-29 to +29; 58 lags total excluding lags of 0) matrix. We averaged over the rows of this matrix to
729 obtain a group-averaged lag-CRP curve (Fig. 3B).

730 **Serial position curve (SPC).** SPCs (Murdock, 1962) reflect the proportion of participants that
731 remember each item as a function of the items' serial positions during encoding. We initialized
732 a number-of-participants (17) by number-of-video-events (30) matrix of zeros. Then, for each
733 recalled event, for each participant, we found the index of the video event that the recalled event
734 most closely matched (via the correlation between the events' topic vectors) and entered a 1 into
735 that position in the matrix. This resulted in a matrix whose entries indicated whether or not each
736 event was recalled by each participant (depending on whether the corresponding entires were
737 set to one or zero). Finally, we averaged over the rows of the matrix to yield a 1 by 30 array
738 representing the proportion of participants that recalled each event as a function of the events'
739 order appearance in the video (Fig. 3C).

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

749 **Semantic clustering scores.** Semantic clustering describes a participant's tendency to recall se-
750 mantically similar presented items together in their recall sequences. Here, we used the topic
751 vectors for each event as a proxy for its semantic content. Thus, the similarity between the seman-

tic content for two events can be computed by correlating their respective topic vectors. For each recall event transition, we sorted all not-yet-recalled events according to how correlated the topic vector of the closest-matching video event was to the topic vector of the closest-matching video event to the just-recalled event. We then computed the percentile rank of the observed next recall. We averaged these percentile ranks across all of the participant's recalls to obtain a single semantic clustering score for the participant.

758 Novel naturalistic memory metrics

759 **Precision.** We tested whether participants who recalled more events were also more *precise* in
760 their recollections. For each participant, we computed the average correlation between the topic
761 vectors for each recall event and those of its closest-matching video event. This gave a single value
762 per participant representing the average precision across all recalled events. We then correlated
763 these values with both hand-annotated and model-derived (i.e., the number of unique video events
764 matched by a participant's recall events) memory performance.

765 **Distinctiveness.** We also considered the *distinctiveness* of each recalled event. That is, how unique
766 a participant's description of a video event was, versus their descriptions of other video events.
767 We hypothesized that participants with high memory performance might describe each event in
768 a more distinctive way (relative to those with lower memory performance who might describe
769 events in a more general way). To test this hypothesis we define a distinctiveness score for each
770 recall event i as

$$d(i) = 1 - \frac{1}{N-1} \sum_{j=i} \text{corr}(\text{event}_i, \text{event}_j)$$

771 where the average is taken over the correlation between the recall event i 's topic vector and the
772 topic vectors from all other recall events from that participant. We averaged these distinctiveness
773 scores across all of the events recalled by the given participant to get the participant's distinctiveness
774 score. We correlated these distinctiveness scores with hand-annotated and model-derived memory

775 performance scores across-subjects, as above.

776 **Averaging correlations** In all instances where we performed statistical tests involving precision
777 or distinctiveness scores, we used the Fisher z -transformation (Fisher, 1925) to stabilize the variance
778 across the distribution of correlation values prior to performing the test. Similarly, when averaging
779 precision or distinctiveness scores, we z -transformed the scores prior to computing the mean, and
780 inverse z -transformed the result.

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

782 We used the UMAP algorithm (McInnes et al., 2018) to project the 100-dimensional topic space onto
783 a two-dimensional space for visualization (Figs. 7, 8). To ensure that all of the trajectories were
784 projected onto the *same* lower dimensional space, we computed the low-dimensional embedding
785 on a “stacked” matrix created by vertically concatenating the events-by-topics topic proportions
786 matrices for the video, across-participants average recall and all 17 individual participants’ recalls.
787 We then separated the rows of the result (a total-number-of-events by two matrix) back into
788 individual matrices for the video topic trajectory, across-participant average recall trajectory and the
789 trajectories for each individual participant’s recalls (Fig. 7). This general approach for discovering
790 a shared low-dimensional embedding for a collections of high-dimensional observations follows
791 Heusser et al. (2018b).

792 We optimized the manifold space for visualization based on two criteria: First, that the 2D
793 embedding of the video trajectory should reflect its original 100-dimensional structure as faithfully
794 as possible. Second, that the path traversed by the embedded video trajectory should intersect
795 itself a minimal number of times. The first criteria helps bolster the validity of visual intuitions
796 about relationships between sections of video content, based on their locations in the embedding
797 space. The second criteria was motivated by the observed low off-diagonal values in the video
798 trajectory’s temporal correlation matrix (suggesting that the same topic-space coordinates should
799 not be revisited; see Figure 2A in the main text). For further details on how we created this
800 low-dimensional embedding space, see *Supporting Information*.

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

802 In Figure 7B, we present an analysis aimed at characterizing locations in topic space that dif-
803 ferent participants move through in a consistent way (via their recall topic trajectories). The
804 two-dimensional topic space used in our visualizations (Fig. 7) comprised a 60 x 60 (arbitrary
805 units) square. We tiled this space with a 50 x 50 grid of evenly spaced vertices, and defined a
806 circular area centered on each vertex whose radius was two times the distance between adjacent
807 vertices (i.e., 2.4 units). For each vertex, we examined the set of line segments formed by connecting
808 each pair successively recalled events, across all participants, that passed through this circle. We
809 computed the distribution of angles formed by those segments and the x -axis, and used a Rayleigh
810 test to determine whether the distribution of angles was reliably “peaked” (i.e., consistent across
811 all transitions that passed through that local portion of topic space). To create Figure 7B we drew
812 an arrow originating from each grid vertex, pointing in the direction of the average angle formed
813 by the line segments that passed within 2.4 units. We set the arrow lengths to be inversely propor-
814 tional to the p -values of the Rayleigh tests at each vertex. Specifically, for each vertex we converted
815 all of the angles of segments that passed within 2.4 units to unit vectors, and we set the arrow
816 lengths at each vertex proportional to the length of the (circular) mean vector. We also indicated
817 any significant results ($p < 0.05$, corrected using the Benjamani-Hochberg procedure) by coloring
818 the arrows in blue (darker blue denotes a lower p -value, i.e., a longer mean vector); all tests with
819 $p \geq 0.05$ are displayed in gray and given a lower opacity value.

820 **Searchlight fMRI analyses**

821 In Figure 9, we present two analyses aimed at identifying brain regions whose responses (as par-
822 ticipants viewed the video) exhibited a particular temporal structure. We developed a searchlight
823 analysis wherein we constructed a 5 x 5 x 5 cube of voxels (following Chen et al., 2017) centered on
824 each voxel in the brain, and for each of these cubes, computed the temporal correlation matrix of
825 the voxel responses during video viewing. Specifically, for each of the 1976 volumes collected dur-
826 ing video viewing, we correlated the activity patterns in the given cube with the activity patterns

827 (in the same cube) collected during every other timepoint. This yielded a 1976 by 1976 correlation
828 matrix for each cube. Note: participant 5's scan ended 75s early, and in Chen et al., 2017's publicly
829 released dataset, their scan data was padded to match the length of the other participants'. For
830 our searchlight analyses, we removed this padded data (i.e., the last 50 TRs), resulting in a 1925 by
831 1925 correlation matrix for each cube in participant 5's brain.

832 Next, we constructed a series of "template" matrices. The first template reflected the timecourse
833 of the video's topic trajectory, and the others reflected the timecourse of each participant's recall
834 trajectory. To construct the video template, we computed the correlations between the topic
835 proportions estimated for every pair of TRs (prior to segmenting the trajectory into discrete events;
836 i.e., the correlation matrix shown in Figs. 2B and 9A). We constructed similar temporal correlation
837 matrices for each participant's recall topic trajectory (Figs. 2D, S4). However, to correct for length
838 differences and potential non-linear transformations between viewing time and recall time, we
839 first used dynamic time warping (Berndt and Clifford, 1994) to temporally align participants'
840 recall topic trajectories with the video topic trajectory. An example correlation matrix before and
841 after warping is shown in Fig. 9B. This yielded a 1976 by 1976 correlation matrix for the video
842 template and for each participant's recall template.

843 The temporal structure of the video's content (as described by our model) is captured in the
844 block-diagonal structure of the video's temporal correlation matrix (e.g., Figs. 2B, 9A), with time
845 periods of thematic stability represented as dark blocks of varying sizes. Inspecting the video
846 correlation matrix suggests that the video's semantic content is highly temporally specific (i.e., the
847 correlations between topic vectors from distant timepoints are almost all near zero). By contrast,
848 the activity patterns of individual (cubes of) voxels can encode relatively limited information
849 on their own, and their activity frequently contributes to multiple separate functions (Freedman
850 et al., 2001; Sigman and Dehaene, 2008; Charron and Koechlin, 2010; Rishel et al., 2013). By
851 nature, these two attributes give rise to similarities in activity across large timescales that may not
852 necessarily reflect a single task. To enable a more sensitive analysis of brain regions whose shifts
853 in activity patterns mirrored shifts in the semantic content of the video or recalls, we restricted
854 the temporal correlations we considered to the timescale of semantic information captured by our

model. Specifically, we isolated the upper triangle of the video correlation matrix and created a “proximal correlation mask” that included only diagonals from the upper triangle of the video correlation matrix up to the first diagonal that contained no positive correlations. Applying this mask to the full video correlation matrix was analogous to excluding diagonals beyond the corner of the largest diagonal block. In other words, the timescale of temporal correlations we considered corresponded to the longest period of thematic stability in the video, and by extension the longest expected period of thematic stability in participants’ recalls and the longest period of stability we might expect to see in voxel activity arising from processing or encoding video content. Figure 9 shows this proximal correlation mask applied to the temporal correlation matrices for the video, an example participant’s (warped) recall, and an example cube of voxels from our searchlight analyses.

To determine which (cubes of) voxel responses matched the video template, we correlated the proximal diagonals from the upper triangle of the voxel correlation matrix for each cube with the proximal diagonals from video template matrix (Kriegeskorte et al., 2008). This yielded, for each participant, a voxelwise map of correlation values. We then performed a one-sample *t*-test on the distribution of (Fisher z-transformed) correlations at each voxel, across participants. This resulted in a value for each voxel (cube), describing how reliably its timecourse followed that of the video.

We further sought to ensure that our analysis identified regions where the activations’ temporal structure specifically reflected that of the video, rather than regions whose activity was simply autocorrelated at a width similar to the video template’s diagonal. To achieve this, we used a phase shift-based permutation procedure, whereby we circularly shifted the video’s topic trajectory by a random number of timepoints, computed the resulting “null” video template, and re-ran the searchlight analysis, in full. (For each of the 100 permutations, the same random shift was used for all participants). We z-scored the observed (unshifted) result at each voxel against the distribution of permutation-derived “null” results, and estimated a *p*-value by computing the proportion of shifted results that yielded larger values. To create the map in Figure 9C, we thresholded out any voxels whose similarity to the unshifted video’s structure fell below the 95th percentile of the permutation-derived similarity results.

883 We used an analogous procedure to identify which voxels' responses reflected the recall tem-
884 plates. For each participant, we correlated the proximal diagonals from the upper triangle of the
885 correlation matrix for each cube of voxels with the proximal diagonals from the upper triangle
886 of their (time-warped) recall correlation matrix. As in the video template analysis, this yielded a
887 voxelwise map of correlation coefficients per participant. However, whereas the video analysis
888 compared every participant's responses to the same template, here the recall templates were unique
889 for each participant. As in the analysis described above, we *t*-scored the (Fisher *z*-transformed)
890 voxelwise correlations, and used the same permutation procedure we developed for the video
891 responses to ensure specificity to the recall timeseries and assign significance values. To create the
892 map in Figure 9D we again thresholded out any voxels whose scores were below the 95th percentile
893 of the permutation-derived null distribution.

894 **Neurosynth decoding analyses**

895 Neurosynth parses a massive online database of over 14,000 neuroimaging studies and constructs
896 meta-analysis images for over 13,000 psychology- and neuroscience-related terms, based on NIfTI
897 images accompanying studies where those terms appear at a high frequency. Given a novel image
898 (tagged with its value type; e.g., *t*-, *F*- or *p*-statistics), Neurosynth returns a list of terms whose
899 meta-analysis images are most similar. Our permutation procedure yielded, for each of the two
900 searchlight analyses, a voxelwise map of significance (*p*-statistic) values. These maps describe the
901 extent to which each voxel *specifically* reflected the temporal structure of the video or individuals'
902 recalls (i.e., for each voxel, the proportion of phase-shifted topic vector correlation matrices less
903 similar to the voxel activity correlation matrix than the unshifted video's correlation matrix). We
904 inputted the two statistical maps described above to Neurosynth to create a list of the 10 most
905 representative terms for each map.

906 **References**

- 907 Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control
908 processes. In Spence, K. W. and Spence, J. T., editors, *The psychology of learning and motivation*,
909 volume 2, pages 89–105. Academic Press, New York.
- 910 Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017).
911 Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–
912 721.
- 913 Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas
914 during narrative perception. *Journal of Neuroscience*, 38(45):9689–9699.
- 915 Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In
916 *KDD workshop*, volume 10, pages 359–370.
- 917 Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International
918 Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, US. ACM.
- 919 Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine
920 Learning Research*, 3:993 – 1022.
- 921 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam,
922 P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R.,
923 Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S.,
924 Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020).
925 Language models are few-shot learners. *arXiv*, 2005.14165.
- 926 Brunec, I. K., Moscovitch, M. M., and Barene, M. D. (2018). Boundaries shape cognitive represen-
927 tations of spaces and events. *Trends in Cognitive Sciences*, 22(7):637–650.
- 928 Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., and Oliva, A. (2015). Intrinsic and extrinsic
929 effects on image memorability. *Vision Research*, 116:165–178.

- 930 Capota, M., Turek, J., Chen, P.-H., Zhu, X., Manning, J. R., Sundaram, N., Keller, B., Wang, Y., and
931 Shin, Y. S. (2017). Brain imaging analysis kit.
- 932 Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes,
933 M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.
934 *arXiv*, 1803.11175.
- 935 Charron, S. and Koechlin, E. (2010). Divided representations of current goals in the human frontal
936 lobes. *Science*, 328(5976):360–363.
- 937 Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared
938 memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*,
939 20(1):115.
- 940 Chun, M. and Turk-Browne, N. (2007). Interactions between attention and memory. *Current opinion
941 in neurobiology*, 17(2):177–184.
- 942 Clewett, D. and Davachi, L. (2017). The ebb and flow of experience determines the temporal
943 structure of memory. *Curr Opin Behav Sci*, 17:186–193.
- 944 Cohn-Sheely, B. I. and Ranganath, C. (2017). Time regained: how the human brain constructs
945 memory for time. *Current Opinion in Behavioral Sciences*, 17:169–177.
- 946 Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions.
947 *Theory of Probability & Its Applications*, 15(3):458–486.
- 948 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the
949 sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–1286.
- 950 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological
951 Science*, 22(2):243–252.
- 952 Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd.

- 953 Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2001). Categorical representation of
954 visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316.
- 955 Friendly, M., Franklin, P. E., Hoffman, D., and Rubin, D. C. (1982). The Toronto Word Pool:
956 Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080
957 words. *Behavior Research Methods and Instrumentation*, 14:375–399.
- 958 Hasson, U., Chen, J., and Honey, C. J. (2015). Hierarchical process memory: memory as an integral
959 component of information processing. *Trends in Cognitive Science*, 19(6):304–315.
- 960 Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal
961 receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550.
- 962 Heusser, A. C., Ezzyat, Y., Shiff, I., and Davachi, L. (2018a). Perceptual boundaries cause mnemonic
963 trade-offs between local boundary processing and across-trial associative binding. *Journal of*
964 *Experimental Psychology Learning, Memory, and Cognition*, 44(7):1075–1090.
- 965 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail: a
966 Python toolbox for analyzing and plotting free recall data. *The Journal of Open Source Software*,
967 10.21105/joss.00424.
- 968 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018b). HyperTools: a Python
969 toolbox for gaining geometric insights into high-dimensional data. *Journal of Machine Learning*
970 *Research*, 18(152):1–6.
- 971 Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal*
972 *of Mathematical Psychology*, 46:269–299.
- 973 Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., and H., E.
974 (2014). A unified mathematical framework for coding time, space, and sequences in the medial
975 temporal lobe. *Journal of Neuroscience*, 34(13):4692–4707.
- 976 Howard, M. W., Viskontas, I. V., Shankar, K. H., and Fried, I. (2012). Ensembles of human MTL
977 neurons “jump back in time” in response to a repeated stimulus. *Hippocampus*, 22:1833–1847.

- 978 Huk, A., Bonnen, K., and He, B. J. (2018). Beyond trial-based paradigms: continuous behavior, on-
979 going neural activity, and naturalistic stimuli. *Journal of Neuroscience*, 10.1523/JNEUROSCI.1920-
980 17.2018.
- 981 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24:103–109.
- 982 Kelly, S., Lloyd, D., Nurmikko, T., and Roberts, N. (2007). Retrieving autobiographical memories
983 of painful events activates the anterior cingulate cortex and inferior frontal gyrus. *The Journal of
984 Pain*, 8(4):307–314.
- 985 Kriat, A. and Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: distin-
986 guishing accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of
987 Experimental Psychology: General*, 123(3):297–315.
- 988 Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – con-
989 necting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:1 – 28.
- 990 Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic
991 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
992 104:211–240.
- 993 Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy
994 of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.
- 995 Manning, J. R. (2019). Episodic memory: mental time travel or a quantum ‘memory wave’ function?
996 *PsyArXiv*, doi:10.31234/osf.io/6zjwb.
- 997 Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors, *Handbook
998 of Human Memory*. Oxford University Press.
- 999 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory.
1000 In Gazzaniga, M., editor, *The Cognitive Neurosciences, Fifth edition*, pages 557–566. MIT Press.

- 1001 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns
1002 in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National*
1003 *Academy of Sciences, USA*, 108(31):12893–12897.
- 1004 McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and
1005 projection for dimension reduction. *arXiv*, 1802(03426).
- 1006 Medford, N. and Critchley, H. D. (2010). Conjoint activity of anterior insular and anterior cingulate
1007 cortex: awareness and response. *Brain Structure and Function*, 214(5-6):535–549.
- 1008 Menon, V. and Uddin, L. Q. (2010). Saliency, switching, attention and control: a network model of
1009 insula function. *Brain Structure and Function*, 214(5-6):655–667.
- 1010 Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations
1011 in vector space. *arXiv*, 1301.3781.
- 1012 Mueller, A., Fillion-Robin, J.-C., Boidol, R., Tian, F., Nechifor, P., yoonsubKim, Peter, Rampin, R.,
1013 Corvellec, M., Medina, J., Dai, Y., Petrushev, B., Langner, K. M., Hong, Alessio, Ozsvald, I.,
1014 vkolmakov, Jones, T., Bailey, E., Rho, V., IgorAPM, Roy, D., May, C., foobuzz, Piyush, Seong,
1015 L. K., Goey, J. V., Smith, J. S., Gus, and Mai, F. (2018). WordCloud 1.5.0: a little word cloud
1016 generator in Python. *Zenodo*, <https://zenodo.org/record/1322068#.W4tPKZNKh24>.
- 1017 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*,
1018 64:482–488.
- 1019 Paller, K. A. and Wagner, A. D. (2002). Observing the transformation of experience into memory.
1020 *Trends in Cognitive Sciences*, 6(2):93–102.
- 1021 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten-
1022 hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot,
1023 M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine*
1024 *Learning Research*, 12:2825–2830.

- 1025 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model
1026 of organizational processes in free recall. *Psychological Review*, 116(1):129–156.
- 1027 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal*
1028 of *Experimental Psychology*, 17:132–138.
- 1029 Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech
1030 recognition. *Proceedings of the IEEE*, 77(2):257–286.
- 1031 Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are
1032 unsupervised multitask learners. *OpenAI Blog*, 1(8).
- 1033 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition. *Curr Opin*
1034 *Behav Sci*, 17:133–140.
- 1035 Ramdas, A., Trillos, N., and Cuturi, M. (2017). On wasserstein two-sample testing and related
1036 families of nonparametric tests. *Entropy*, 19(2):47.
- 1037 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature*
1038 *Reviews Neuroscience*, 13:713 – 726.
- 1039 Rishel, C. A., Huang, G., and Freedman, D. J. (2013). Independent category and spatial encoding
1040 in parietal cortex. *Neuron*, 77(5):969–979.
- 1041 Sigman, M. and Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during
1042 dual-task performance. *Journal of Neuroscience*, 28(30):7585–7589.
- 1043 Simony, E., Honey, C. J., Chen, J., and Hasson, U. (2016). Uncovering stimulus-locked network
1044 dynamics during narrative comprehension. *Nature Communications*, 7(12141):1–13.
- 1045 Stark, S. M., Yassa, M. A., and Stark, C. E. L. (2010). Individual differences in spatial pattern
1046 separation performance associated with healthy aging in humans. *Learning & Memory*, 17(6):284–
1047 288.

- 1048 Tomrary, A. and Davachi, L. (2017). Consolidation promotes the emergence of representational
1049 overlap in the hippocampus and medial prefrontal cortex. *Neuron*, 96(1):228–241.
- 1050 Tyng, C. M., Amin, H. U., Saad, M. N. M., and S, M. A. (2017). The influences of emotion on
1051 learning and memory. *Frontiers in psychology*, 8:1454.
- 1052 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American Journal
1053 of Psychology*, 35:396–401.
- 1054 Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale
1055 automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665.
- 1056 Yassa, M. A., Lacy, J. W., Stark, S. M., Albert, M. S., Gallagher, M., and Stark, C. E. L. (2011). Pattern
1057 separation deficits associated with increased hippocampal ca3 and dentate gyrus activity in
1058 nondemented older adults. *Hippocampus*, 21(9):968–979.
- 1059 Yassa, M. A. and Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends In Neuro-
1060 sciences*, 34(10):515–525.
- 1061 Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research.
1062 *Journal of Memory and Language*, 46:441–517.
- 1063 Yonelinas, A. P., Kroll, N. E., Quamme, J. R., Lazzara, M. M., Sauvé, M. J., Widaman, K. F., and
1064 Knight, R. T. (2002). Effects of extensive temporal lobe damage or mild hypoxia on recollection
1065 and familiarity. *Nature Neuroscience*, 5(11):1236–41.
- 1066 Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception:
1067 a mind-brain perspective. *Psychological Bulletin*, 133:273–293.
- 1068 Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit
1069 memories to other brains: Constructing shared neural representations via communication. *Cereb
1070 Cortex*, 27(10):4988–5000.
- 1071 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and
1072 memory. *Psychological Bulletin*, 123(2):162 – 185.

1073 **Supporting information**

1074 Supporting information is available in the online version of the paper.

1075 **Acknowledgements**

1076 We thank Luke Chang, Janice Chen, Chris Honey, Lucy Owen, Emily Whitaker, and Kirsten Ziman
1077 for feedback and scientific discussions. We also thank Janice Chen, Yuan Chang Leong, Kenneth
1078 Norman, and Uri Hasson for sharing the data used in our study. Our work was supported in part
1079 by NSF EPSCoR Award Number 1632738. The content is solely the responsibility of the authors
1080 and does not necessarily represent the official views of our supporting organizations.

1081 **Author contributions**

1082 Conceptualization: A.C.H. and J.R.M.; Methodology: A.C.H., P.C.F. and J.R.M.; Software: A.C.H.,
1083 P.C.F. and J.R.M.; Analysis: A.C.H., P.C.F. and J.R.M.; Writing, Reviewing, and Editing: A.C.H.,
1084 P.C.F. and J.R.M.; Supervision: J.R.M.

1085 **Author information**

1086 The authors declare no competing financial interests. Correspondence and requests for materials
1087 should be addressed to J.R.M. (jeremy.r.manning@dartmouth.edu).