

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 September 1, 2020

Abstract

The mental contexts in which we interpret experiences are often person-specific, even when the experiences themselves are shared. We developed a geometric framework for mathematically characterizing the subjective conceptual content of dynamic naturalistic experiences. We model experiences and memories as *trajectories* through word embedding spaces whose coordinates reflect the universe of thoughts under consideration. Memory encoding can then be modeled as geometrically preserving or distorting the *shape* of the original experience. We applied our approach to data collected as participants watched and verbally recounted a television episode while undergoing functional neuroimaging. Participants' recounts all preserved coarse spatial properties (essential narrative elements), but not fine spatial scale (low-level) details, of the episode's trajectory. We also identified networks of brain structures sensitive to these trajectory shapes. Our work provides insights into how our brains preserve and distort our ongoing experiences when we encode them into episodic memories.

Introduction

What does it mean to *remember* something? In traditional episodic memory experiments (e.g., list-learning or trial-based experiments; ??), remembering is often cast as a discrete binary operation: each studied item may be separated from the rest of one's experience and labeled as having been either recalled or forgotten. More nuanced studies might incorporate self-reported confidence measures as a proxy for memory strength, or ask participants to discriminate between recollecting the (contextual) details of an experience or having a general feeling of familiarity (?). Using well-controlled, trial-based experimental designs, the field has amassed a wealth of information regarding human episodic memory (for review see ?). However, there are fundamental properties of the external world and our memories that trial-based experiments are not well suited to capture (for review, also see ??). First, our experiences and memories are continuous, rather than discrete—isolating a naturalistic event from the context in which it occurs can substantially change its meaning. Second, whether or not the rememberer has precisely reproduced a specific set of words in describing a given experience is nearly orthogonal to how well they were actually

33 able to remember it. In classic (e.g., list-learning) memory studies, by contrast, the number or
34 proportion of *exact* recalls is often considered to be a primary metric for assessing the quality of
35 participants' memories. Third, one might remember the essence (or a general summary) of an
36 experience but forget (or neglect to recount) particular low-level details. Capturing the essence of
37 what happened is often a main goal of recounting an episodic memory to a listener, whereas the
38 inclusion of specific low-level details is often less pertinent.

39 How might we formally characterize the *essence* of an experience, and whether it has been
40 recovered by the rememberer? And how might we distinguish an experience's overarching essence
41 from its low-level details? One approach is to start by considering some fundamental properties
42 of the dynamics of our experiences. Each given moment of an experience tends to derive meaning
43 from surrounding moments, as well as from longer-range temporal associations (??). Therefore,
44 the timecourse describing how an event unfolds is fundamental to its overall meaning. Further,
45 this hierarchy formed by our subjective experiences at different timescales defines a *context* for
46 each new moment (e.g., ??), and plays an important role in how we interpret that moment and
47 remember it later (for review see ??). Our memory systems can leverage these associations to
48 form predictions that help guide our behaviors (?). For example, as we navigate the world, the
49 features of our subjective experiences tend to change gradually (e.g., the room or situation we
50 find ourselves in at any given moment is strongly temporally autocorrelated), allowing us to form
51 stable estimates of our current situation and behave accordingly (??).

52 Occasionally, this gradual drift of our ongoing experience is punctuated by sudden changes, or
53 shifts (e.g., when we walk through a doorway; ?). Prior research suggests that these sharp transi-
54 tions (termed *event boundaries*) help to discretize our experiences (and their mental representations)
55 into *events* (??????). The interplay between the stable (within-event) and transient (across-event)
56 temporal dynamics of an experience also provides a potential framework for transforming experi-
57 ences into memories that distills those experiences down to their essences. For example, prior work
58 has shown that event boundaries can influence how we learn sequences of items (??), navigate (?),
59 and remember and understand narratives (??). This work also suggests a means of distinguishing
60 the essence of an experience from its low-level details. The overall structure of events and event

61 transitions reflects how the high-level experience unfolds (i.e., its essence), while subtler event-
62 level properties reflect low-level details. Prior research has also implicated a network of brain
63 regions (including the hippocampus and the medial prefrontal cortex) in playing a critical role in
64 transforming experiences into structured and consolidated memories (?).

65 Here, we sought to examine how the temporal dynamics of a naturalistic experience were later
66 reflected in participants' memories. We also sought to leverage the above conceptual insights into
67 the distinctions between an experience's essence and its low-level details to build models that
68 explicitly quantified these distinctions. We analyzed an open dataset that comprised behavioral
69 and functional Magnetic Resonance Imaging (fMRI) data collected as participants viewed and
70 then verbally recounted an episode of the BBC television show *Sherlock* (?). We developed a
71 computational framework for characterizing the temporal dynamics of the moment-by-moment
72 content of the episode, and of participants' verbal recalls. Our framework uses topic modeling (?)
73 to characterize the thematic conceptual (semantic) content present in each moment of the episode
74 and recalls by projecting each moment into a word embedding space. We then use hidden Markov
75 models (??) to discretize this evolving semantic content into events. In this way, we cast both
76 naturalistic experiences and memories of those experiences as geometric *trajectories* through word
77 embedding space that describe how they evolve over time. Under this framework, successful
78 remembering entails verbally traversing the content trajectory of the episode, thereby reproducing
79 the shape (essence) of the original experience. Our framework captures the episode's essence in
80 the sequence of geometric coordinates for its events, and its low-level details by examining its
81 within-event geometric properties.

82 Comparing the overall shapes of the topic trajectories for the episode and participants' recalls
83 reveals which aspects of the episode's essence were preserved (or discarded) in the translation into
84 memory. We also develop two metrics for assessing participants' memories for low-level details:
85 (1) the *precision* with which a participant recounts details about each event, and (2) the *distinctiveness*
86 of each recalled event, relative to other events. We examine how these metrics relate to overall
87 memory performance as judged by third-party human annotators. We also compare and contrast
88 our general approach to studying memory for naturalistic experiences with standard metrics for

89 assessing performance on more traditional memory tasks, such as list-learning. Last, we leverage
90 our framework to identify networks of brain structures whose responses (as participants watched
91 the episode) reflected the temporal dynamics of the episode and/or how participants would later
92 recount it.

93 Results

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

103 ? collected hand-annotated information about each of 1,000 (manually identified) scenes span-
104 ning the roughly 50 minute video used in their experiment. This information included: a brief
105 narrative description of what was happening, the location where the scene took place, the names
106 of any characters on the screen, and other similar details (for a full list of annotated features, see
107 *Methods*). We took from these annotations the union of all unique words (excluding stop words,
108 such as "and," "or," "but," etc.) across all features and scenes as the vocabulary for the topic
109 model. We then concatenated the sets of words across all features contained in overlapping sliding
110 windows of (up to) 50 scenes, and treated each window as a single document for the purpose of
111 fitting the topic model. Next, we fit a topic model with (up to) $K = 100$ topics to this collection of
112 documents. We found that 32 unique topics (with non-zero weights) were sufficient to describe
113 the time-varying content of the episode (see *Methods*; Figs. ??, S2). We note that our approach is
114 similar in some respects to Dynamic Topic Models (?), in that we sought to characterize how the

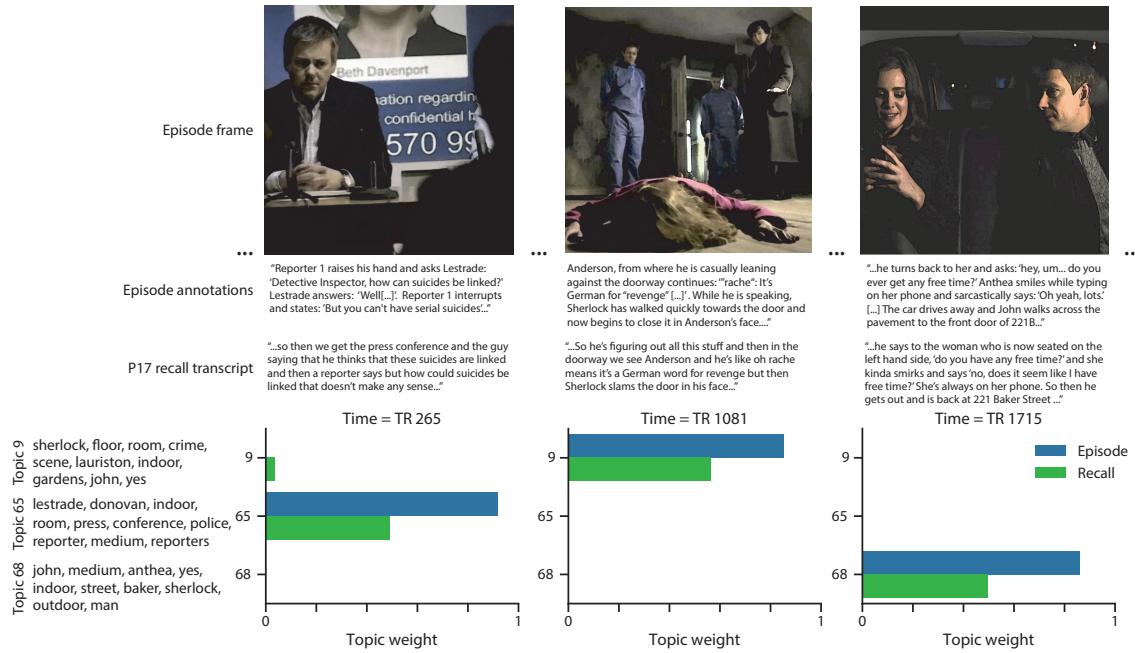


Figure 1: Topic weights in episode and recall content. We used hand-annotated descriptions of each manually identified scene from the episode to fit a topic model. Three example episode frames (first row) and their associated descriptions (second row) are displayed. The third row shows an example participant’s recounts of the same three scenes. We used the topic model (fit to the episode annotations) to estimate topic vectors for each moment of the episode and each sentence of participants’ recalls. Example topic vectors are displayed in the bottom row (blue: episode annotations; green: example participant’s recalls). Three topic dimensions are shown (the highest-weighted topics for each of the three example scenes, respectively), along with 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.

115 thematic content of the episode evolved over time. However, whereas Dynamic Topic Models
 116 are designed to characterize how the properties of *collections* of documents change over time, our
 117 sliding window approach allows us to examine the topic dynamics within a single document (or
 118 video). Specifically, our approach yielded (via the topic proportions matrix) a single *topic vector* for
 119 each sliding window of annotations transformed by the topic model. We then stretched (interpo-
 120 lated) the resulting windows-by-topics matrix to match the time series of the 1,976 fMRI volumes
 121 collected as participants viewed the episode.

122 The 32 topics we found were heavily character-focused (i.e., the top-weighted word in each
 123 topic was nearly always a character) and could be roughly divided into themes centered around

¹²⁴ Sherlock Holmes (the titular character), John Watson (Sherlock’s close confidant and assistant),
¹²⁵ supporting characters (e.g., Inspector Lestrade, Sergeant Donovan, or Sherlock’s brother Mycroft),
¹²⁶ or the interactions between various groupings of these characters (see Fig. S2). This likely follows
¹²⁷ from the frequency with which these terms appeared in the episode annotations. Several of the
¹²⁸ identified topics were highly similar, which we hypothesized might allow us to distinguish between
¹²⁹ subtle narrative differences if the distinctions between those overlapping topics were meaningful.
¹³⁰ The topic vectors for each timepoint were also *sparse*, in that only a small number (typically one or
¹³¹ two) of topics tended to be “active” in any given timepoint (see Fig. ??A). Further, the dynamics
¹³² of the topic activations appeared to exhibit *persistence* (i.e., given that a topic was active in one
¹³³ timepoint, it was likely to be active in the following timepoint) along with *occasional rapid changes*
¹³⁴ (i.e., occasionally topic weights would change abruptly from one timepoint to the next). These two
¹³⁵ properties of the topic dynamics may be seen in the block diagonal structure of the timepoint-by-
¹³⁶ timepoint correlation matrix (Fig. ??B) and reflect the gradual drift and sudden shifts fundamental
¹³⁷ to the temporal dynamics of many real-world experiences, as well as television episodes. Given
¹³⁸ this observation, we adapted an approach devised by ?, and used a hidden Markov model (HMM)
¹³⁹ to identify the *event boundaries* where the topic activations changed rapidly (i.e., the boundaries of
¹⁴⁰ the blocks in the temporal correlation matrix; event boundaries identified by the HMM are outlined
¹⁴¹ in yellow in Fig. ??B). Part of our model fitting procedure required selecting an appropriate number
¹⁴² of events into which the topic trajectory should be segmented. To accomplish this, we used an
¹⁴³ optimization procedure that maximized the difference between the topic weights for timepoints
¹⁴⁴ within an event versus timepoints across multiple events (see *Methods* for additional details). We
¹⁴⁵ then created a stable summary of the content within each episode event by averaging the topic
¹⁴⁶ vectors across the timepoints spanned by each event (Fig. ??C).

¹⁴⁷ Given that the time-varying content of the episode could be segmented cleanly into discrete
¹⁴⁸ events, we wondered whether participants’ recalls of the episode also displayed a similar structure.
¹⁴⁹ We applied the same topic model (already trained on the episode annotations) to each participant’s
¹⁵⁰ recalls. Analogously to how we parsed the time-varying content of the episode, to obtain similar
estimates for each participant’s recall transcript, we treated each overlapping window of (up to)

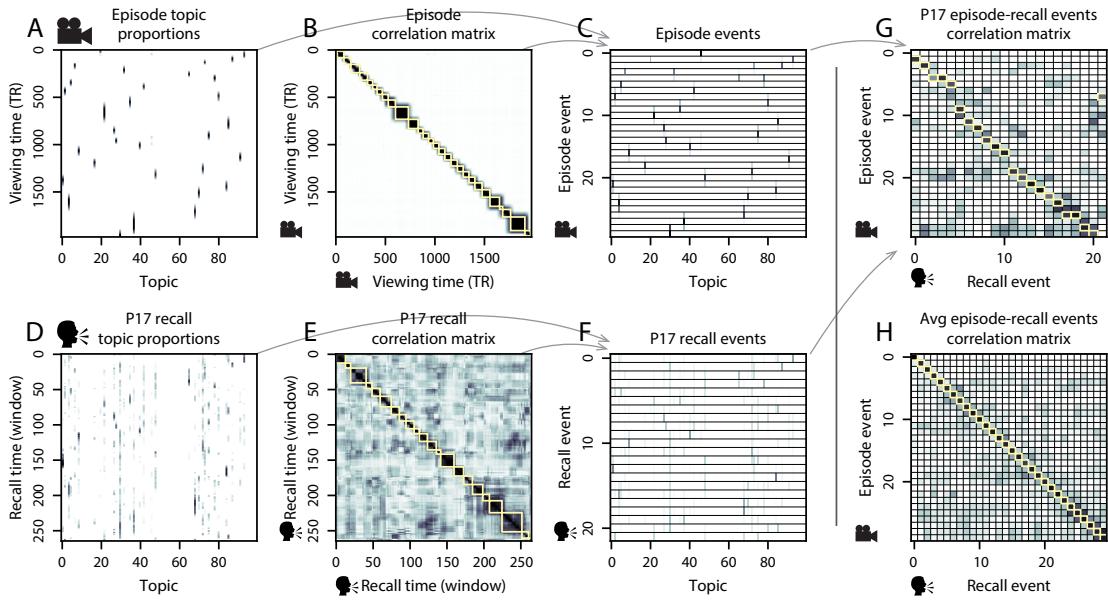


Figure 2: Modeling naturalistic stimuli and recalls. All panels: darker colors indicate greater values; range: [0, 1]. **A.** Topic vectors ($K = 100$) for each of the 1976 episode timepoints. **B.** Timepoint-by-timepoint correlation matrix of the topic vectors displayed in Panel A. Event boundaries discovered by the HMM are denoted in yellow (30 events detected). **C.** Average topic vectors for each of the 30 episode events. **D.** Topic vectors for each of 265 sliding windows of sentences spoken by an example participant while recalling the episode. **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 recall events from the example participant. **G.** Correlations between the topic vectors for every pair of episode events (Panel C) and recall events (from the example participant; Panel F). For similar plots for all participants, see Figure S5. **H.** Average correlations between each pair of episode events and recall events (across all 17 participants). To create the figure, each recalled event was assigned to the episode event with the most correlated topic vector (yellow boxes in panels G and H).

152 10 sentences from their transcript as a document, and computed the most probable mix of topics
153 reflected in each timepoint's sentences. This yielded, for each participant, a number-of-windows
154 by number-of-topics topic proportions matrix that characterized how the topics identified in the
155 original episode were reflected in the participant's recalls. An important feature of our approach
156 is that it allows us to compare participants' recalls to events from the original episode, despite
157 that different participants used widely varying language to describe the events, and that those
158 descriptions often diverged in content and quality from the episode annotations. This ability
159 to match up conceptually related text that differs in specific vocabulary, detail, and length is an
160 important benefit of projecting the episode and recalls into a shared topic space. An example topic
161 proportions matrix from one participant's recalls is shown in Figure ??D.

162 Although the example participant's recall topic proportions matrix has some visual similarity
163 to the episode topic proportions matrix, the time-varying topic proportions for the example par-
164 ticipant's recalls are not as sparse as those for the episode (compare Figs. ??A and D). Similarly,
165 although there do appear to be periods of stability in the recall topic dynamics (i.e., most topics
166 are active or inactive over contiguous blocks of time), the changes in topic activations that define
167 event boundaries appear less clearly delineated in participants' recalls than in the episode's anno-
168 tations. To examine these patterns in detail, we computed the timepoint-by-timepoint correlation
169 matrix for the example participant's recall topic proportions matrix (Fig. ??E). As in the episode
170 correlation matrix (Fig. ??B), the example participant's recall correlation matrix has a strong block
171 diagonal structure, indicating that their recalls are discretized into separated events. We used the
172 same HMM-based optimization procedure that we had applied to the episode's topic proportions
173 matrix (see *Methods*) to estimate an analogous set of event boundaries in the participant's recount-
174 ing of the episode (outlined in yellow). We carried out this analysis on all 17 participants' recall
175 topic proportions matrices (Fig. S4).

176 Two clear patterns emerged from this set of analyses. First, although every individual partic-
177 ipant's recalls could be segmented into discrete events (i.e., every individual participant's recall
178 correlation matrix exhibited clear block diagonal structure; Fig. S4), each participant appeared to
179 have a unique *recall resolution*, reflected in the sizes of those blocks. While some participants'

recall topic proportions segmented into just a few events (e.g., Participants P4, P5, and P7), others' segmented into many shorter duration events (e.g., Participants P12, P13, and P17). This suggests that different participants may be recalling the episode with different levels of detail—i.e., some might recount only high-level essential plot details, whereas others might recount low-level details instead (or in addition). The second clear pattern present in every individual participant's recall correlation matrix was that, unlike in the episode correlation matrix, there were substantial off-diagonal correlations. Whereas each event in the original episode was (largely) separable from the others (Fig. ??B), in transforming those separable events into memory, participants appeared to be integrating across multiple events, blending elements of previously recalled and not-yet-recalled content into each newly recalled event (Figs. ??E, S4; also see ???).

The above results demonstrate that topic models capture the dynamic conceptual content of the episode and participants' recalls of the episode. Further, the episode and recalls exhibit event boundaries that can be identified automatically using HMMs to segment the dynamic content. Next, we asked whether some correspondence might be made between the specific content of the events the participants experienced in the episode, and the events they later recalled. We labeled each recalled event as matching the episode event with the most similar (i.e., most highly correlated) topic vector (Figs. ??G, S5). This yielded a sequence of "presented" events from the original episode, and a (potentially differently ordered) sequence of "recalled" events for each participant. Analogous to classic list-learning studies, we can then examine participants' recall sequences by asking which events they tended to recall first (probability of first recall; Fig. ??A; ???); how participants most often transitioned between recalls of the events as a function of the temporal distance between them (lag-conditional response probability; Fig. ??B; ?); and which events they were likely to remember overall (serial position recall analyses; Fig. ??C; ?). Some of the patterns we observed appeared to be similar to classic effects from the list-learning literature. For example, participants had a higher probability of initiating recall with early events (Fig. ??A) and a higher probability of transitioning to neighboring events with an asymmetric forward bias (Fig. ??B). However, unlike what is typically observed in list-learning studies, we did not observe patterns comparable to the primacy or recency serial position effects (Fig. ??C). We hypothesized

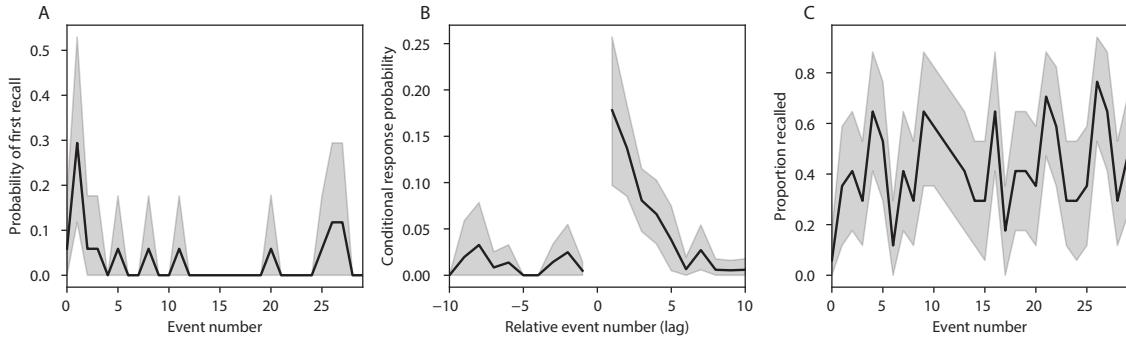


Figure 3: Naturalistic extensions of classic list-learning memory analyses. **A.** The probability of first recall as a function of the serial position of the event in the episode. **B.** The probability of recalling each event, conditioned on having most recently recalled the event *lag* events away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

208 that participants might be leveraging meaningful narrative associations and references over long
209 timescales throughout the episode.

210 Clustering scores are often used by memory researchers to characterize how people organize
211 their memories of words on a studied list (for review, see ?). We defined analogous measures
212 to characterize how participants organized their memories for episodic events (see *Methods* for
213 details). Temporal clustering refers to the extent to which participants group their recall responses
214 according to encoding position. Overall, we found that sequentially viewed episode events tended
215 to appear nearby in participants' recall event sequences (mean clustering score: 0.767, SEM: 0.029).
216 Participants with higher temporal clustering scores tended to exhibit better overall memory for
217 the episode, according to both ?'s hand-counted numbers of recalled scenes from the episode
218 (Pearson's $r(15) = 0.62$, $p = 0.008$) and the numbers of episode events that best-matched at least
219 one recalled event (i.e., model-estimated number of recalled events; Pearson's $r(15) = 0.49$, $p =$
220 0.0046). Semantic clustering measures the extent to which participants cluster their recall responses
221 according to semantic similarity. We found that participants tended to recall semantically similar
222 episode events together (mean clustering score: 0.787, SEM: 0.018), and that semantic clustering
223 score was also related to both hand-annotated (Pearson's $r(15) = 0.65$, $p = 0.004$) and model-
224 estimated (Pearson's $r(15) = 0.61$, $p = 0.0092$) numbers of recalled events.

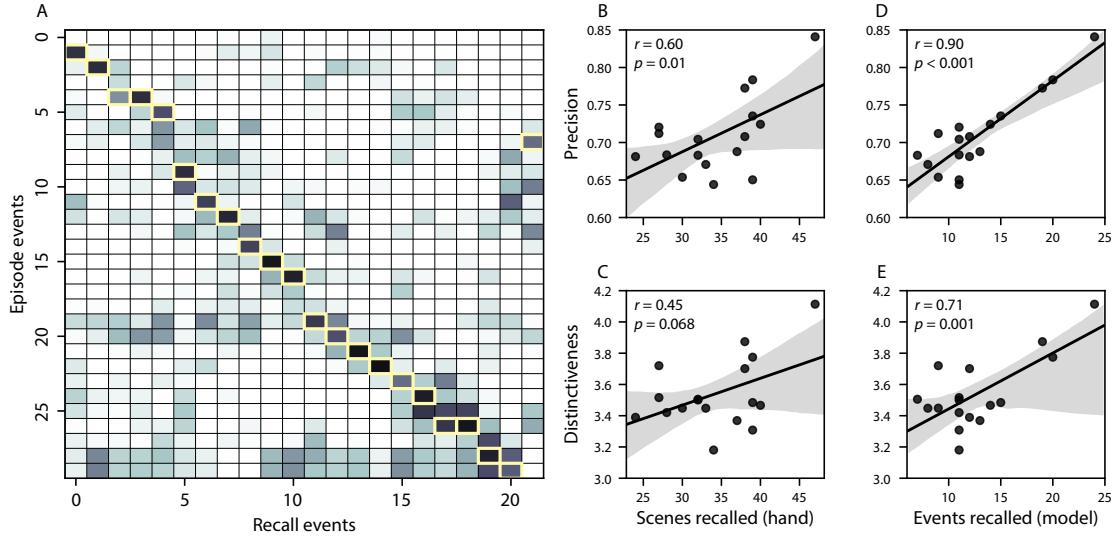


Figure 4: Novel content-based metrics of naturalistic memory: precision and distinctiveness. **A.** The episode-recall correlation matrix for a representative participant (P17). The yellow boxes highlight the maximum correlation in each column. The example participant’s overall precision score was computed as the average across correlation values in the yellow boxes. Their distinctiveness score was computed as the average (over recall events) of the z-scored (within column) event precisions. **B.** The (Pearson’s) correlation between precision and hand-counted number of recalled scenes. **C.** The correlation between distinctiveness and hand-counted number of recalled scenes. **D.** The correlation between precision and the number of recalled episode events, as determined by our model. **E.** The correlation between distinctiveness and the number of recalled episode events, as determined by our model.

225 The above analyses illustrate how our framework for characterizing the dynamic conceptual
226 content of naturalistic episodes enables us to carry out analyses that have traditionally been
227 applied to much simpler list-learning paradigms. However, perhaps the most interesting aspects
228 of memory for naturalistic episodes are those that have no list-learning analogs. The nuances of
229 how one's memory for an event might capture some details, yet distort or neglect others, is central
230 to how we use our memory systems in daily life. Yet when researchers study memory in highly
231 simplified paradigms, those nuances are not typically observable. We next developed two novel
232 continuous metrics, termed precision and distinctiveness, aimed at characterizing distortions in
233 the conceptual content of individual recalled events, and the conceptual overlap between how
234 people described different events.

235 *Precision* is intended to capture the “completeness” of recall, or how fully the presented content
236 was recapitulated in a participant’s recounting. We define a recall event’s precision as the maximum
237 correlation between the topic proportions of that recall event and any episode event (Fig. ??). In
238 other words, given that a recalled event best matches a particular episode event, more precisely
239 recalled events overlap more strongly with the conceptual content of the original episode event.
240 When a given event is assigned a blend of several topics, as is often the case (Fig. ??), a high
241 precision score requires recapitulating the relative topic proportions during recall.

242 *Distinctiveness* is intended to capture the “specificity” of recall. In other words, distinctiveness
243 quantifies the extent to which a given recalled event reflects the most similar episode event over
244 and above its reflection of other episode events. Intuitively, distinctiveness is like a normalized
245 variant of our precision metric. Whereas precision solely measures how much detail about an
246 episode was captured in someone’s recall, distinctiveness penalizes details that also pertain to
247 other episode events. We define the distinctiveness of an event’s recall as its precision expressed in
248 standard deviation units with respect to other episode events. Specifically, for a given recall event,
249 we compute the correlation between its topic vector and that of each episode event. This yields a
250 distribution of correlation coefficients (one per episode event). We subtract the mean and divide by
251 the standard deviation of this distribution to z-score the coefficients. The maximum value in this
252 distribution (which, by definition, belongs to the episode event that best matches the given recall

253 event) is that recall event's distinctiveness score. In this way, recall events that match one episode
254 event far better than all other episode events will receive a high distinctiveness score. By contrast,
255 a recall event that matches all episode events roughly equally will receive a comparatively low
256 distinctiveness score.

257 In addition to examining how precisely and distinctively participants recalled individual events,
258 one may also use these metrics to summarize each participant's performance by averaging across
259 a participant's event-wise precision or distinctiveness scores. This enables us to quantify how pre-
260 cisely a participant tended to recall subtle within-event details, as well as how specific (distinctive)
261 those details were to individual events from the episode. Participants' average precision and dis-
262 tinctiveness scores were strongly correlated ($r(15) = 0.90, p < 10^{-5}$). This indicates that participants
263 who tended to precisely recount low-level details of episode events also tended to do so in an
264 event-specific way (e.g., as opposed to detailing recurring themes that were present in most or all
265 episode events; this behavior would have resulted in high precision but low distinctiveness). We
266 found that, across participants, higher precision scores were positively correlated with both the
267 hand-annotated ($r(15) = 0.60, p = 0.010$) and model-estimated ($r(15) = 0.90, p < 0.001$) numbers of
268 events that participants recalled. Participants' average distinctiveness scores were also correlated
269 with both the hand-annotated ($r(15) = 0.45, p = 0.068$) and model-estimated ($r(15) = 0.71, p = 0.001$)
270 numbers of recalled events.

271 Examining individual recalls of the same episode event can provide insights into how the above
272 precision and distinctiveness scores may be used to characterize similarities and differences in how
273 different people describe the same shared experience. In Figure ??, we compare recalls for the same
274 episode event (event 22) from different participants: one with a high precision score (P17), and the
275 other with a low precision score (P6). From the HMM-identified event boundaries, we recovered
276 the set of annotations describing the content of an example episode event (Fig. ??B), and divided
277 them into different color-coded sections for each action or feature described. We used an analogous
278 approach to identify the set of sentences comprising the corresponding recall events for each of the
279 two example participants. Figure ??C shows excerpts of two participants' recall transcripts that
280 comprised sentences between the first and last descriptions of content from the example episode

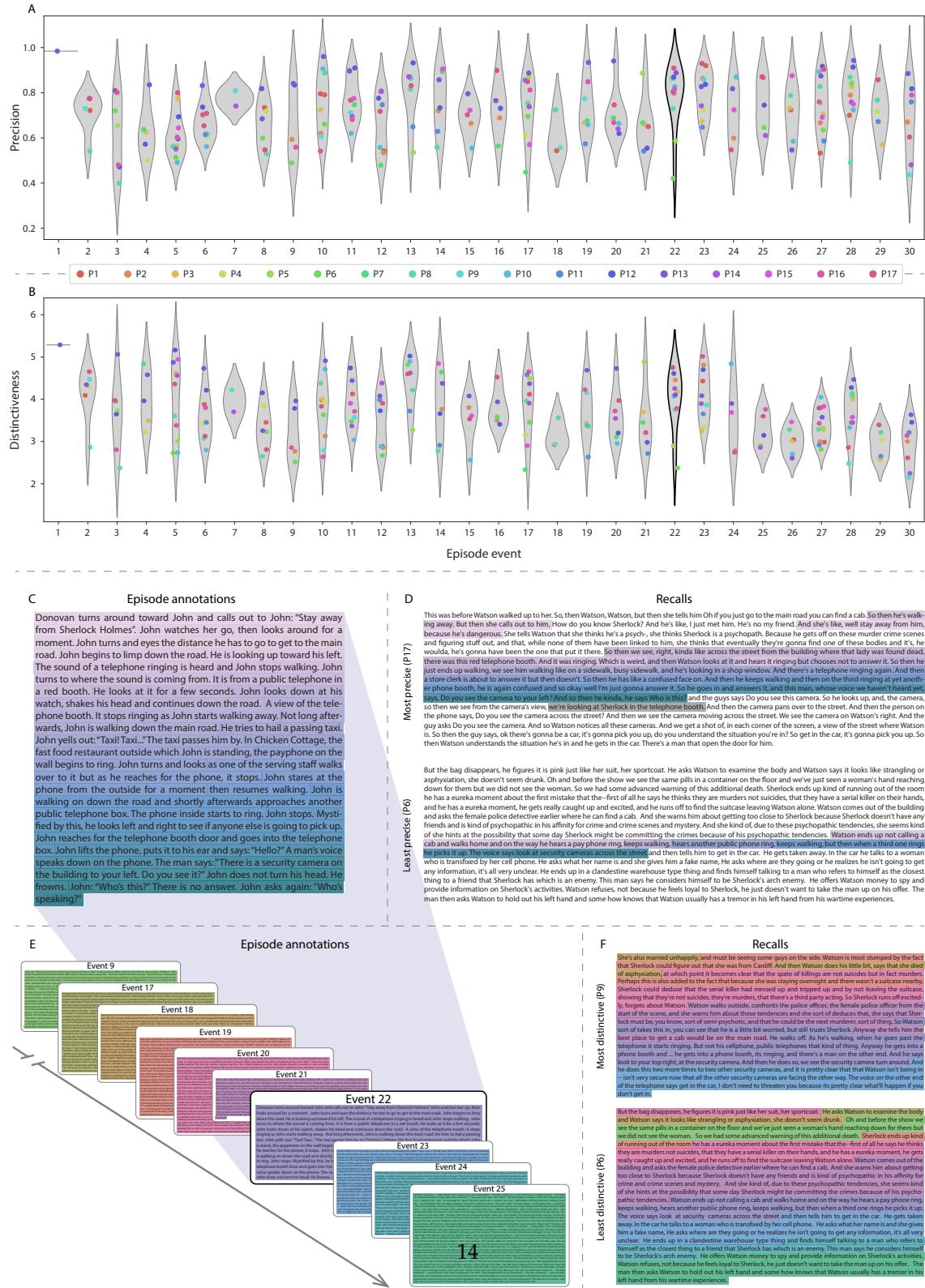


Figure 5: Precision reflects the completeness of recall, whereas distinctiveness reflects recall specificity. A. Recall precision by episode event. Grey violin plots display kernel density estimates for the distribution of recall precision scores for a single episode event. Colored dots within each violin plot represent individual participants' recall precision for the given event. Episode events are ordered along the x-axis by the average precision with which they were remembered. B. The set of “Narrative Details” episode annotations (generated by ?) for scenes comprising an example episode event (22) identified by the HMM. Each action or feature is highlighted in a different color. C. Excerpts from the most precise (P17) and least precise (P6) participants’ recalls of episode event 22. Descriptions of specific actions or features reflecting those highlighted in Panel

281 event. We then colored all words describing actions and features in the transcripts shown in Panel
282 C according to the color-coded annotations in Panel B. Visual comparison of these example recalls
283 reveals that the more precise recall captures more of the episode event's content, and in greater
284 detail.

285 Figure ?? also illustrates the differences between high and low distinctiveness scores for the
286 same event detailed in Figure ??B (i.e., event 22). Here, we have extracted the set of sentences
287 comprising the most distinctive recall event (P9) and least distinctive recall event (P6) matched to
288 the example episode event (Fig. ??F). We also extracted the annotations for the example episode
289 event, as well as those from each other episode event whose content the example participants'
290 single recall events described (Fig. ??E). We assigned each episode event a unique color (Panel E)
291 and colored each recalled phrase or sentence (Panel F) according to the episode events they best
292 matched. Visual inspection of Panel F reveals that the most distinctive recall's content is tightly
293 concentrated around event 22, whereas the least distinctive recall incorporates content from a much
294 wider range of episode events.

295 The preceding analyses sought to characterize how participants' recounts of individual
296 episode events captured the low-level details of each event. Next we sought to characterize how
297 participants' recounts of the full episode captured its high-level essence— i.e., the shape of the
298 episode's trajectory through word embedding (topic) space. To visualize the essence of the episode
299 and each participant's recall trajectory (?), we projected the topic proportions matrices for the
300 episode and recalls onto a shared two-dimensional space using Uniform Manifold Approximation
301 and Projection (UMAP; ?). In this lower-dimensional space, each point represents a single episode
302 or recall event, and the distances between the points reflect the distances between the events'
303 associated topic vectors (Fig. ??). In other words, events that are nearer to each other in this space
304 are more semantically similar, and those that are farther apart are less so.

305 Visual inspection of the episode and recall topic trajectories reveals a striking pattern. First, the
306 topic trajectory of the episode (which reflects its dynamic content; Fig. ??A) is captured nearly per-
307 fectly by the averaged topic trajectories of participants' recalls (Fig. ??B). To assess the consistency
308 of these recall trajectories across participants, we asked: given that a participant's recall trajectory

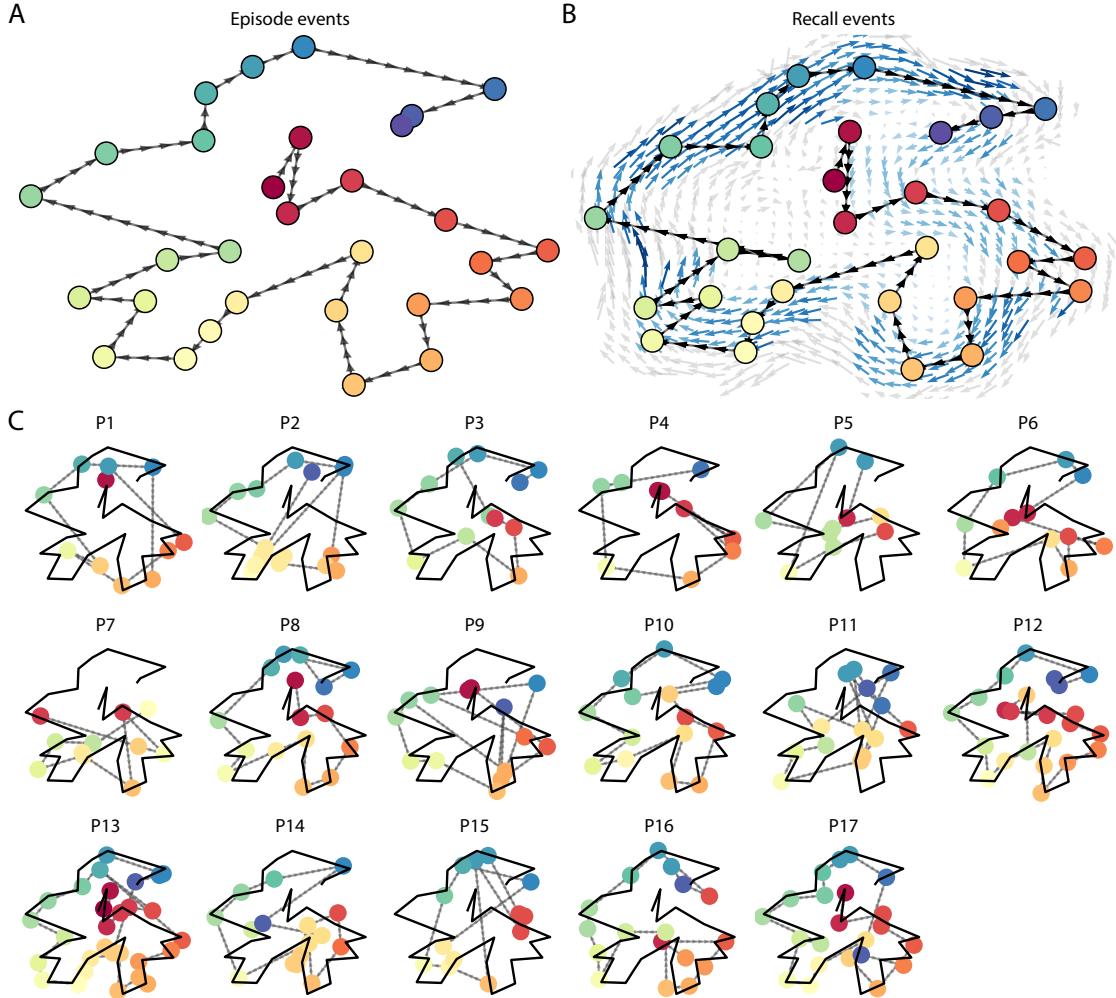


Figure 6: Trajectories through topic space capture the dynamic content of the episode and recalls. All panels: the topic proportion matrices have been projected onto a shared two-dimensional space using UMAP. **A.** The two-dimensional topic trajectory taken by the episode of *Sherlock*. Each dot indicates 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 episode (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 episode's trajectory is shown in black for reference. Here, events (dots) are colored by their matched episode event (Panel A).

had entered a particular location in the reduced topic space, could the position of their *next* recalled event be predicted reliably? For each location in the reduced topic space, we computed the set of line segments connecting successively recalled events (across all participants) that intersected that location (see *Methods* for additional details). We then computed (for each location) the distribution of angles formed by the lines defined by those line segments and a fixed reference line (the x -axis). Rayleigh tests revealed the set of locations in topic space at which these across-participant distributions exhibited reliable peaks (blue arrows in Fig. ??B reflect significant peaks at $p < 0.05$, corrected). We observed that the locations traversed by nearly the entire episode trajectory exhibited such peaks. In other words, participants' recalls exhibited similar trajectories to each other that also matched the trajectory of the original episode (Fig. ??C). This is especially notable when considering the fact that the numbers of events participants recalled (dots in Fig. ??C) varied considerably across people, and that every participant used different words to describe what they had remembered happening in the episode. Differences in the numbers of remembered events appear in participants' trajectories as differences in the sampling resolution along the trajectory. We note that this framework also provides a means of disentangling classic "proportion recalled" measures (i.e., the proportion of episode events described in participants' recalls) from participants' abilities to recapitulate the episode's essence (i.e., the similarity between the shapes of the original episode trajectory and that defined by each participant's recounting of the episode).

In addition to enabling us to visualize the episode's high-level essence, describing the episode as a geometric trajectory also enables us to drill down to individual words and quantify how each word relates to the memorability of each event. This provides another approach to examining participants' recall for low-level details beyond the precision and distinctiveness measures we defined above. The results displayed in Figures ??C and ??A suggest that certain events were remembered better than others. Given this, we next asked whether the events were generally remembered precisely or imprecisely tended to reflect particular content. Because our analysis framework projects the dynamic episode content and participants' recalls into a shared space, and because the dimensions of that space represent topics (which are, in turn, sets of weights over known words in the vocabulary), we are able to recover the weighted combination of words that

337 make up any point (i.e., topic vector) in this space. We first computed the average precision with
338 which participants recalled each of the 30 episode events (Fig. ??A; note that this result is analogous
339 to a serial position curve created from our precision metric). We then computed a weighted average
340 of the topic vectors for each episode event, where the weights reflected how precisely each event
341 was recalled. To visualize the result, we created a “wordle” image (?) where words weighted more
342 heavily by more precisely-remembered topics appear in a larger font (Fig. ??B, green box). Across
343 the full episode, content that weighted heavily on topics and words central to the major foci of
344 the episode (e.g., the names of the two main characters, “Sherlock” and “John,” and the address
345 of a major recurring location, “221B Baker Street”) were best remembered. An analogous analysis
346 revealed which themes were less-precisely remembered. Here in computing the weighted average
347 over events’ topic vectors, we weighted each event in *inverse* proportion to its average precision
348 (Fig. ??B, red box). The least precisely remembered episode content reflected information that was
349 extraneous to the episode’s essence, such as the proper names of relatively minor characters (e.g.,
350 “Mike,” “Molly,” and “Lestrade”) and locations (e.g., “St. Bartholomew’s Hospital”).

351 A similar result emerged from assessing the topic vectors for individual episode and recall
352 events (Fig. ??C). Here, for each of the three most and least precisely remembered episode events, we
353 have constructed two wordles: one from the original episode event’s topic vector (left) and a second
354 from the average recall topic vector for that event (right). The three most precisely remembered
355 events (circled in green) correspond to scenes integral to the central plot-line: a mysterious figure
356 spying on John in a phone booth; John meeting Sherlock at Baker St. to discuss the murders; and
357 Sherlock laying a trap to catch the killer. Meanwhile, the least precisely remembered events (circled
358 in red) reflect scenes that comprise minor plot points: a video of singing cartoon characters that
359 participants viewed in an introductory clip prior to the main episode; John asking Molly about
360 Sherlock’s habit of over-analyzing people; and Sherlock noticing evidence of Anderson’s and
361 Donovan’s affair.

362 The results thus far inform us about which aspects of the dynamic content in the episode partic-
363 ipants watched were preserved or altered in participants’ memories. We next carried out a series of
364 analyses aimed at understanding which brain structures might facilitate these preservations and

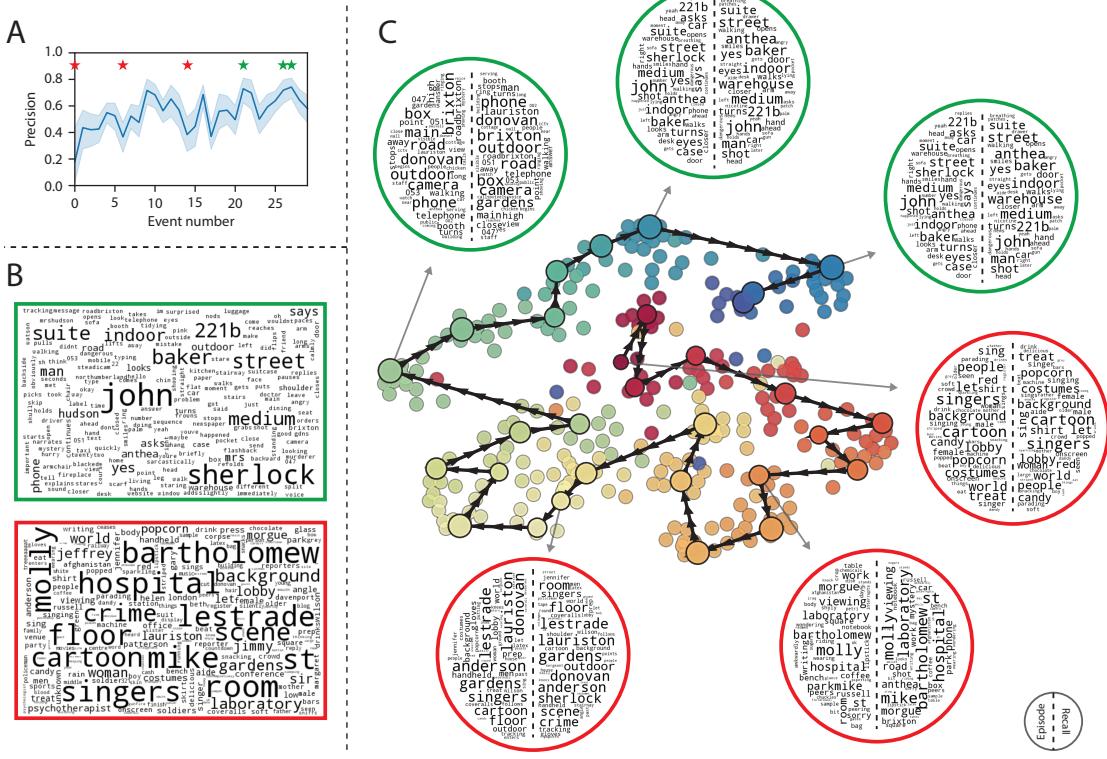


Figure 7: Language used in the most and least precisely remembered events. **A.** Average precision (episode event-recall event topic vector correlation) across participants for each episode event. Here we defined each episode event's precision for each participant as the correlation between its topic vector and the most-correlated recall event's topic vector from that participant. Error bars denote bootstrap-derived across-participant 95% confidence intervals. The stars denote the three most precisely remembered events (green) and least precisely remembered events (red). **B.** Wordles comprising the top 200 highest-weighted words reflected in the weighted-average topic vector across episode events. Green: episode events were weighted by their precision (Panel A). Red: episode events were weighted by the inverse of their precision. **C.** The set of all episode and recall events is projected onto the two-dimensional space derived in Figure ???. The dots outlined in black denote episode events (dot size is proportional to each event's average precision). The dots without black outlines denote individual recall events from each participant. All dots are colored using the same scheme as Figure ??A. Wordles for several example events are displayed (green: three most precisely remembered events; red: three least precisely remembered events). Within each circular wordle, the left side displays words associated with the topic vector for the episode event, and the right side displays words associated with the (average) recall event topic vector, across all recall events matched to the given episode event.

365 transformations between the participants' shared experience of watching the episode and their
366 subsequent memories of the episode. In the first analysis, we sought to identify brain structures
367 that were sensitive to the dynamic unfolding of the episode's content, as characterized by its topic
368 trajectory. We used a searchlight procedure to identify clusters of voxels whose activity patterns
369 displayed a proximal temporal correlation structure (as participants watched the episode) matching
370 that of the original episode's topic proportions (Fig. ??A; see *Methods* for additional details). In a
371 second analysis, we sought to identify brain structures whose responses (during episode viewing)
372 reflected how each participant would later structure their *recounting* of the episode. We used a
373 searchlight procedure to identify clusters of voxels whose proximal temporal correlation matrices
374 matched that of the topic proportions matrix for each participant's recall transcript (Figs. ??B; see
375 *Methods* for additional details). To ensure our searchlight procedure identified regions *specifically*
376 sensitive to the temporal structure of the episode or recalls (i.e., rather than those with a tem-
377 poral autocorrelation length similar to that of the episode and recalls), we performed a phase
378 shift-based permutation correction (see *Methods* for additional details). As shown in Figure ??C,
379 the episode-driven searchlight analysis revealed a distributed network of regions that may play
380 a role in processing information relevant to the narrative structure of the episode. Similarly, the
381 recall-driven searchlight analysis revealed a second network of regions (Fig. ??D) that may facil-
382 itate a person-specific transformation of one's experience into memory. The top ten Neurosynth
383 terms (?) associated with each (unthresholded) map are displayed in each panel. In identifying
384 regions whose responses to ongoing experiences reflect how those experiences will be remembered
385 later, this latter analysis extends classic *subsequent memory effect analyses* (e.g., ?) to the domain of
386 naturalistic experiences.

387 The searchlight analyses described above yielded two distributed networks of brain regions,
388 whose activity timecourses tracked with the temporal structure of the episode (Fig. ??C) or par-
389 ticipants' subsequent recalls (Fig. ??D). We next sought to gain greater insight into the structures
390 and functional networks our results reflected. To accomplish this, we performed an additional,
391 exploratory analysis using Neurosynth (?). Given an arbitrary statistical map as input, Neurosynth
392 performs a massive automated meta-analysis, returning a ranked list of terms reported in papers

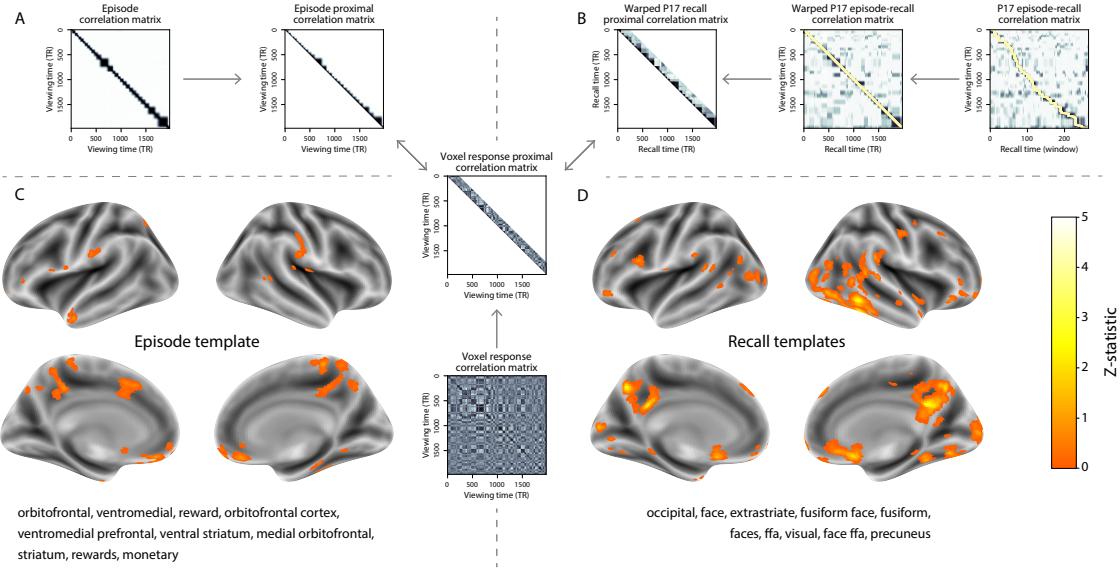


Figure 8: Brain structures that underlie the transformation of experience into memory. **A.** We isolated the proximal diagonals from the upper triangle of the episode correlation matrix, and applied this same diagonal mask to the voxel response correlation matrix for each cube of voxels in the brain. We then searched for brain regions whose activation timeseries consistently exhibited a similar proximal correlational structure to the episode model, across participants. **B.** We used dynamic time warping (?) to align each participant's recall timeseries to the TR timeseries of the episode. 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. The top ten Neurosynth terms displayed in the panel were computed using the unthresholded map. **D.** We also identified a network of regions sensitive to how individuals would later structure the episode's content in their recalls. The map shown is thresholded at $p < 0.05$, corrected. The top ten Neurosynth terms displayed in the panel were computed using the unthresholded map.

393 with similar significance maps. We ran Neurosynth on the significance maps for the episode- and
394 recall-driven searchlight analyses. These maps, along with the 10 terms with maximally similar
395 meta-analysis images identified by Neurosynth are shown in Figure ??.

396 Discussion

397 Explicitly modeling the dynamic content of a naturalistic stimulus and participants' memories
398 enabled us to connect the present study of naturalistic recall with an extensive prior literature that
399 has used list-learning paradigms to study memory (for review see ?), as in Figure ???. We found
400 some similarities between how participants in the present study recounted a television episode and
401 how participants typically recall memorized random word lists. However, our broader claim is that
402 word lists miss out on fundamental aspects of naturalistic memory more like the sort of memory
403 we rely on in everyday life. For example, there are no random word list analogs of character
404 interactions, conceptual dependencies between temporally distant episode events, the sense of
405 solving a mystery that pervades the *Sherlock* episode, or the myriad other features of the episode
406 that convey deep meaning and capture interest. Nevertheless, each of these properties affects how
407 people process and engage with the episode as they are watching it, and how they remember it
408 later. The overarching goal of the present study is to characterize how the rich dynamics of the
409 episode affect the rich behavioral and neural dynamics of how people remember it.

410 Our work casts remembering as reproducing (behaviorally and neurally) the topic trajectory,
411 or "shape," of an experience. When we characterized memory for a television episode using this
412 framework, we found that every participant's recounting of the episode recapitulated the low
413 spatial frequency details of the shape of its trajectory through topic space (Fig. ??). We termed
414 this narrative scaffolding the episode's *essence*. Where participants' behaviors varied most was
415 in their tendencies to recount specific low-level details from each episode event. Geometrically,
416 this appears as high spatial frequency distortions in participants' recall trajectories relative to the
417 trajectory of the original episode (Fig. ??). We developed metrics to characterize the precision
418 (recovery of any and all event-level information) and distinctiveness (recovery of event-specific

419 information). We also used word cloud visualizations to interpret the details of these event-level
420 distortions.

421 The neural analyses we carried out (Fig. ??) also leveraged our geometric framework for
422 characterizing the shapes of the episode and participants' recounts. We identified one network
423 of regions whose responses tracked with temporal correlations in the conceptual content of the
424 episode (as quantified by topic models applied to a set of annotations about the episode). This
425 network included orbitofrontal cortex, ventromedial prefrontal cortex, striatum, among others.
426 As reviewed by ?, several of these regions are members of the *anterior temporal system*, which
427 has been implicated in assessing and processing the familiarity of ongoing experiences, emotions,
428 social cognition, and reward. A second network we identified tracked with temporal correlations
429 in the idiosyncratic conceptual content of participants' subsequent recounts of the episode.
430 This network included occipital cortex, extrastriate cortex, fusiform gyrus, and the precuneus.
431 Several of these regions are members of the *posterior medial system* (?), which has been implicated
432 in matching incoming cues about the current situation to internally maintained *situation models*
433 that specify the parameters and expectations inherent to the current situation (also see ??). Taken
434 together, our results support the notion that these two (partially overlapping) networks work in
435 coordination to make sense of our ongoing experiences, distort them in a way that links them with
436 our prior knowledge and experiences, and encodes those distorted representations into memory
437 for our later use.

438 Our general approach draws inspiration from prior work aimed at elucidating the neural and
439 behavioral underpinnings of how we process dynamic naturalistic experiences and remember them
440 later. Our approach to identifying neural responses to naturalistic stimuli (including experiences)
441 entails building an explicit model of the stimulus dynamics and searching for brain regions whose
442 responses are consistent with the model (also see ??). In prior work, a series of studies from Uri
443 Hasson's group (?????) have developed a clever alternative approach: rather than building an
444 explicit stimulus model, these studies instead search for brain responses (while experiencing the
445 stimulus) that are reliably similar across individuals. So called *inter-subject correlation* (ISC) and
446 *inter-subject functional connectivity* (ISFC) analyses effectively treat other people's brain responses

447 to the stimulus as a “model” of how its features change over time (also see ?). These purely brain-
448 driven approaches are well-suited to identifying which brain structures exhibit similar stimulus-
449 driven responses across individuals. Further, because neural response dynamics are observed data
450 (rather than model approximations), such approaches do not require a detailed understanding of
451 which stimulus properties or features might be driving the observed responses. However, this
452 also means that the specific stimulus features driving those responses are typically opaque to the
453 researcher. Our approach is complementary. By explicitly modeling the stimulus dynamics, we
454 are able to relate specific stimulus features to behavioral and neural dynamics. However, when
455 our model fails to accurately capture the stimulus dynamics that are truly driving behavioral and
456 neural responses, our approach necessarily yields an incomplete characterization of the neural
457 basis of the processes we are studying.

458 Other recent work has used HMMs to discover latent event structure in neural responses to
459 naturalistic stimuli (?). By applying HMMs to our explicit models of stimulus and memory dy-
460 namics, we gain a more direct understanding of those state dynamics. For example, we found
461 that although the events comprising each participant’s recalls recapitulated the episode’s essence,
462 participants differed in the *resolution* of their recounting of low-level details. In turn, these individ-
463 ual behavioral differences were reflected in differences in neural activity dynamics as participants
464 watched the television episode.

465 Our approach also draws inspiration from the growing field of word embedding models. The
466 topic models (?) we used to embed text from the episode annotations and participants’ recall
467 transcripts are just one of many models that have been studied in an extensive literature. The
468 earliest approaches to word embedding, including latent semantic analysis (?), used word co-
469 occurrence statistics (i.e., how often pairs of words occur in the same documents contained in the
470 corpus) to derive a unique feature vector for each word. The feature vectors are constructed so that
471 words that co-occur more frequently have feature vectors that are closer (in Euclidean distance).
472 Topic models are essentially an extension of those early models, in that they attempt to explicitly
473 model the underlying causes of word co-occurrences by automatically identifying the set of themes
474 or topics reflected across the documents in the corpus. More recent work on these types of semantic

475 models, including word2vec (?), the Universal Sentence Encoder (?), GPT-2 (?), and GTP-3 (?) use
476 deep neural networks to attempt to identify the deeper conceptual representations underlying
477 each word. Despite the growing popularity of these sophisticated deep learning-based embedding
478 models, we chose to prioritize interpretability of the embedding dimensions (e.g., Fig. ??) over
479 raw performance (e.g., with respect to some predefined benchmark). Nevertheless, we note that
480 our general framework is, in principle, robust to the specific choice of language model as well as
481 other aspects of our computational pipeline. For example, the word embedding model, timeseries
482 segmentation model, and the episode-recall matching function could each be customized to suit
483 a particular question space or application. Indeed, for some questions, interpretability of the
484 embeddings may not be a priority, and thus other text embedding approaches (including the deep
485 learning-based models described above) may be preferable. Further work will be needed to explore
486 the influence of particular models on our framework’s predictions and performance.

487 Our work has broad implications for how we characterize and assess memory in real-world
488 settings, such as the classroom or physician’s office. For example, the most commonly used
489 classroom evaluation tools involve simply computing the proportion of correctly answered exam
490 questions. Our work indicates that this approach is only loosely related to what educators might
491 really want to measure: how well did the students understand the key ideas presented in the
492 course? Under this typical framework of assessment, the same exam score of 50% could be ascribed
493 to two very different students: one who attended to the full course but struggled to learn more than
494 a broad overview of the material, and one who attended to only half of the course but understood
495 the attended material perfectly. Instead, one could apply our computational framework to build
496 explicit dynamic content models of the course material and exam questions. This approach would
497 provide a more nuanced and specific view into which aspects of the material students had learned
498 well (or poorly). In clinical settings, memory measures that incorporate such explicit content
499 models might also provide more direct evaluations of patients’ memories, and of doctor-patient
500 interactions.

501 **Methods**

502 **Experimental design and data collection**

503 Data were collected by ?. In brief, participants ($n = 22$) viewed the first 48 minutes of “A Study
504 in Pink,” the first episode of the BBC television show *Sherlock*, while fMRI volumes were collected
505 (TR = 1500 ms). Participants were pre-screened to ensure they had never seen any episode of the
506 show before. The stimulus was divided into a 23 min (946 TR) and a 25 min (1030 TR) segment
507 to mitigate technical issues related to the scanner. After finishing the clip, participants were
508 instructed to (quoting from ?) “describe what they recalled of the [episode] in as much detail as
509 they could, to try to recount events in the original order they were viewed in, and to speak for at
510 least 10 minutes if possible but that longer was better. They were told that completeness and detail
511 were more important than temporal order, and that if at any point they realized they had missed
512 something, to return to it. Participants were then allowed to speak for as long as they wished, and
513 verbally indicated when they were finished (e.g., ‘I’m done’).” Five participants were dropped
514 from the original dataset due to excessive head motion (2 participants), insufficient recall length (2
515 participants), or falling asleep during stimulus viewing (1 participant), resulting in a final sample
516 size of $n = 17$. For additional details about the experimental procedure and scanning parameters,
517 see ?. The experimental protocol was approved by Princeton University’s Institutional Review
518 Board.

519 After preprocessing the fMRI data and warping the images into a standard (3 mm³ MNI) space,
520 the voxel activations were z-scored (within voxel) and spatially smoothed using a 6 mm (full width
521 at half maximum) Gaussian kernel. The fMRI data were also cropped so that all episode-viewing
522 data were aligned across participants. This included a constant 3 TR (4.5 s) shift to account for the
523 lag in the hemodynamic response. (All of these preprocessing steps followed ?, where additional
524 details may be found.)

525 The video stimulus was divided into 1,000 fine-grained “scenes” and annotated by an inde-
526 pendent coder. For each of these 1,000 scenes, the following information was recorded: a brief
527 narrative description of what was happening, the location where the scene took place, whether

528 that location was indoors or outdoors, the names of all characters on-screen, the name(s) of the
529 character(s) in focus in the shot, the name(s) of the character(s) currently speaking, the camera
530 angle of the shot, a transcription of any text appearing on-screen, and whether or not there was
531 music present in the background. Each scene was also tagged with its onset and offset time, in
532 both seconds and TRs.

533 **Data and code availability**

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

536 **Statistics**

537 All statistical tests performed in the behavioral analyses were two-sided. All statistical tests per-
538 formed in the neural data analyses were two-sided, except for the permutation-based thresholding,
539 which was one-sided. In this case, we were specifically interested in identifying voxels whose acti-
540 vation time series reflected the temporal structure of the episode and recall trajectories to a *greater*
541 extent than that of the phase-shifted trajectories.

542 **Modeling the dynamic content of the episode and recall transcripts**

543 **Topic modeling**

544 The input to the topic model we trained to characterize the dynamic content of the episode com-
545 prised 998 hand-generated annotations of short (mean: 2.96s) scenes spanning the video clip (?
546 generated 1000 annotations total; we removed two annotations referring to a break between the
547 first and second scan sessions, during which no fMRI data were collected). We concatenated the
548 text for all of the annotated features within each segment, creating a “bag of words” describing each
549 scene and performed some minor preprocessing (e.g., stemming possessive nouns and removing
550 punctuation). We then re-organized the text descriptions into overlapping sliding windows span-
ning (up to) 50 scenes each. In other words, we estimated the “context” for each scene using the text

descriptions of the preceding 25 scenes, the present scene, and the following 24 scenes. To model the context for scenes near the beginning of the episode (i.e., within 25 scenes of the beginning or end), we created overlapping sliding windows that grew in size from one scene to the full length. We also tapered the sliding window lengths at the end of the episode, whereby scenes within fewer than 24 scenes of the end of the episode were assigned sliding windows that extended to the end of the episode. This procedure ensured that each scene's content was represented in the text corpus an equal number of times.

We trained our model using these overlapping text samples with `scikit-learn` (version 0.19.1; ?), called from our high-dimensional visualization and text analysis software, `HyperTools` (?). Specifically, we used the `CountVectorizer` class to transform the text from each window into a vector of word counts (using the union of all words across all scenes as the “vocabulary,” excluding English stop words); this yielded a number-of-windows by number-of-words *word count* matrix. We then used the `LatentDirichletAllocation` class (`topics=100, method='batch'`) to fit a topic model (?) to the word count matrix, yielding a number-of-windows (1047) by number-of-topics (100) *topic proportions* matrix. The topic proportions matrix describes the gradually evolving mix of topics (latent themes) present in each scene. Next, we transformed the topic proportions matrix to match the 1976 fMRI volume acquisition times. We assigned each topic vector to the timepoint (in seconds) midway between the beginning of the first scene and the end of the last scene in its corresponding sliding text window. By doing so, we warped the linear temporal distance between consecutive topic vectors to align with the inconsistent temporal distance between consecutive annotations (whose durations varied greatly). We then rescaled these timepoints to 1.5s TR units, and used linear interpolation to estimate a topic vector for each TR. This resulted in a number-of-TRs (1976) by number-of-topics (100) matrix.

We created similar topic proportions matrices using hand-annotated transcripts of each participant's verbal recall of the episode (annotated by ?). We tokenized the transcript into a list of sentences, and then re-organized the list into overlapping sliding windows spanning (up to) 10 sentences each, analogously to how we parsed the episode annotations. In turn, we transformed each window's sentences into a word count vector (using the same vocabulary as for the episode

580 model), and then we used the topic model already trained on the episode scenes to compute the
581 most probable topic proportions for each sliding window. This yielded a number-of-windows
582 (range: 83–312) by number-of-topics (100) topic proportions matrix for each participant. These
583 reflected the dynamic content of each participant’s recalls. Note: for details on how we selected the
584 episode and recall window lengths and number of topics, see *Supporting Information* and Figure S1.

585 **Segmenting topic proportions matrices into discrete events using hidden Markov Models**

586 We parsed the topic proportions matrices of the episode and participants’ recalls into discrete events
587 using hidden Markov Models (HMMs; ?). Given the topic proportions matrix (describing the mix
588 of topics at each timepoint) and a number of states, K , an HMM recovers the set of state transitions
589 that segments the timeseries into K discrete states. Following ?, we imposed an additional set of
590 constraints on the discovered state transitions that ensured that each state was encountered exactly
591 once (i.e., never repeated). We used the BrainIAK toolbox (?) to implement this segmentation.

592 We used an optimization procedure to select the appropriate K for each topic proportions
593 matrix. Prior studies on narrative structure and processing have shown that we both perceive
594 and internally represent the world around us at multiple, hierarchical timescales (e.g., ??????).
595 However, for the purposes of our framework, we sought to identify the single timeseries of event-
596 representations that is emphasized *most heavily* in the temporal structure of the episode and of each
597 participant’s recall. We quantified this as the set of K states that maximized the similarity between
598 topic vectors for timepoints comprising each state, while minimizing the similarity between topic
599 vectors for timepoints across different states. Specifically, we computed (for each matrix)

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

600 where a was the distribution of within-state topic vector correlations, and b was the distribution of
601 across-state topic vector correlations . We computed the first Wasserstein distance (W_1 ; also known
602 as *Earth mover’s distance*; ??) between these distributions for a large range of possible K -values
603 (range [2, 50]), and selected the K that yielded the maximum value. Figure ??B displays the event

604 boundaries returned for the episode, and Figure S4 displays the event boundaries returned for
605 each participant's recalls. See Figure S6 for the optimization functions for the episode and recalls.
606 After obtaining these event boundaries, we created stable estimates of the content represented in
607 each event by averaging the topic vectors across timepoints between each pair of event boundaries.
608 This yielded a number-of-events by number-of-topics matrix for the episode and recalls from each
609 participant.

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

611 In traditional list-learning experiments, participants view a list of items (e.g., words) and then
612 recall the items later. Our episode-recall event matching approach affords us the ability to analyze
613 memory in a similar way. The episode and recall events can be treated analogously to studied and
614 recalled "items" in a list-learning study. We can then extend classic analyses of memory perfor-
615 mance and dynamics (originally designed for list-learning experiments) to the more naturalistic
616 episode recall task used in this study.

617 Perhaps the simplest and most widely used measure of memory performance is *accuracy*—i.e.,
618 the proportion of studied (experienced) items (in this case, episode events) that the participant
619 later remembered. ? used this method to rate each participant's memory quality by computing
620 the proportion of (50, manually identified) scenes mentioned in their recall. We found a strong
621 across-participants correlation between these independent ratings and the proportion of 30 HMM-
622 identified episode events matched to participants' recalls (Pearson's $r(15) = 0.71, p = 0.002$). We
623 further considered a number of more nuanced memory performance measures that are typically
624 associated with list-learning studies. We also provide a software package, Quail, for carrying out
625 these analyses (?).

626 **Probability of first recall (PFR).** PFR curves (???) reflect the probability that an item will be
627 recalled first as a function of its serial position during encoding. To carry out this analysis, we
628 initialized a number-of-participants (17) by number-of-episode-events (30) matrix of zeros. Then
629 for each participant, we found the index of the episode event that was recalled first (i.e., the episode

630 event whose topic vector was most strongly correlated with that of the first recall event) and filled
631 in that index in the matrix with a 1. Finally, we averaged over the rows of the matrix, resulting in a
632 1 by 30 array representing the proportion of participants that recalled an event first, as a function
633 of the order of the event’s appearance in the episode (Fig. ??A).

634 **Lag conditional probability curve (lag-CRP).** The lag-CRP curve (?) reflects the probability of
635 recalling a given item after the just-recalled item, as a function of their relative encoding positions
636 (or *lag*). In other words, a lag of 1 indicates that a recalled item was presented immediately after
637 the previously recalled item, and a lag of -3 indicates that a recalled item came 3 items before the
638 previously recalled item. For each recall transition (following the first recall), we computed the
639 lag between the current recall event and the next recall event, normalizing by the total number
640 of possible transitions. This yielded a number-of-participants (17) by number-of-lags (-29 to +29;
641 58 lags total excluding lags of 0) matrix. We averaged over the rows of this matrix to obtain a
642 group-averaged lag-CRP curve (Fig. ??B).

643 **Serial position curve (SPC).** SPCs (?) reflect the proportion of participants that remember each
644 item as a function of the items’ serial positions during encoding. We initialized a number-of-
645 participants (17) by number-of-episode-events (30) matrix of zeros. Then, for each recalled event,
646 for each participant, we found the index of the episode event that the recalled event most closely
647 matched (via the correlation between the events’ topic vectors) and entered a 1 into that position
648 in the matrix. This resulted in a matrix whose entries indicated whether or not each event was
649 recalled by each participant (depending on whether the corresponding entires were set to one or
650 zero). Finally, we averaged over the rows of the matrix to yield a 1 by 30 array representing the
651 proportion of participants that recalled each event as a function of the events’ order appearance in
652 the episode (Fig. ??C).

653 **Temporal clustering scores.** Temporal clustering describes a participant’s tendency to organize
654 their recall sequences by the learned items’ encoding positions. For instance, if a participant
655 recalled the episode events in the exact order they occurred (or in exact reverse order), this would

656 yield a score of 1. If a participant recalled the events in random order, this would yield an expected
657 score of 0.5. For each recall event transition (and separately for each participant), we sorted all
658 not-yet-recalled events according to their absolute lag (i.e., distance away in the episode). We
659 then computed the percentile rank of the next event the participant recalled. We averaged these
660 percentile ranks across all of the participant’s recalls to obtain a single temporal clustering score
661 for the participant.

662 **Semantic clustering scores.** Semantic clustering describes a participant’s tendency to recall se-
663 mantically similar presented items together in their recall sequences. Here, we used the topic
664 vectors for each event as a proxy for its semantic content. Thus, the similarity between the seman-
665 tic content for two events can be computed by correlating their respective topic vectors. For each
666 recall event transition, we sorted all not-yet-recalled events according to how correlated the topic
667 vector of *the closest-matching episode event* was to the topic vector of the closest-matching episode
668 event to the just-recalled event. We then computed the percentile rank of the observed next recall.
669 We averaged these percentile ranks across all of the participant’s recalls to obtain a single semantic
670 clustering score for the participant.

671 **Averaging correlations**

672 In all instances where we performed statistical tests involving precision or distinctiveness scores
673 (Fig. ??, we used the Fisher z-transformation (?) to stabilize the variance across the distribution of
674 correlation values prior to performing the test. Similarly, when averaging precision or distinctive-
675 ness scores, we z-transformed the scores prior to computing the mean, and inverse z-transformed
676 the result.

677 **Visualizing the episode and recall topic trajectories**

678 We used the UMAP algorithm (?) to project the 100-dimensional topic space onto a two-dimensional
679 space for visualization (Figs. ??, ??). To ensure that all of the trajectories were projected onto the *same*
680 lower dimensional space, we computed the low-dimensional embedding on a “stacked” matrix

681 created by vertically concatenating the events-by-topics topic proportions matrices for the episode,
682 across-participants average recall and all 17 individual participants' recalls. We then separated
683 the rows of the result (a total-number-of-events by two matrix) back into individual matrices for
684 the episode topic trajectory, across-participant average recall trajectory and the trajectories for
685 each individual participant's recalls (Fig. ??). This general approach for discovering a shared
686 low-dimensional embedding for a collections of high-dimensional observations follows ?.

687 We optimized the manifold space for visualization based on two criteria: First, that the 2D em-
688 bedding of the episode trajectory should reflect its original 100-dimensional structure as faithfully
689 as possible. Second, that the path traversed by the embedded episode trajectory should intersect
690 itself a minimal number of times. The first criteria helps bolster the validity of visual intuitions
691 about relationships between sections of episode content, based on their locations in the embedding
692 space. The second criteria was motivated by the observed low off-diagonal values in the episode
693 trajectory's temporal correlation matrix (suggesting that the same topic-space coordinates should
694 not be revisited; see Figure 2A in the main text). For further details on how we created this
695 low-dimensional embedding space, see *Supporting Information*.

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

697 In Figure ??B, we present an analysis aimed at characterizing locations in topic space that dif-
698 ferent participants move through in a consistent way (via their recall topic trajectories). The
699 two-dimensional topic space used in our visualizations (Fig. ??) comprised a 60×60 (arbitrary
700 units) square. We tiled this space with a 50×50 grid of evenly spaced vertices, and defined a
701 circular area centered on each vertex whose radius was two times the distance between adjacent
702 vertices (i.e., 2.4 units). For each vertex, we examined the set of line segments formed by connecting
703 each pair successively recalled events, across all participants, that passed through this circle. We
704 computed the distribution of angles formed by those segments and the x -axis, and used a Rayleigh
705 test to determine whether the distribution of angles was reliably "peaked" (i.e., consistent across
706 all transitions that passed through that local portion of topic space). To create Figure ??B we
707 drew an arrow originating from each grid vertex, pointing in the direction of the average angle

708 formed by the line segments that passed within 2.4 units. We set the arrow lengths to be inversely
709 proportional to the p -values of the Rayleigh tests at each vertex. Specifically, for each vertex we
710 converted all of the angles of segments that passed within 2.4 units to unit vectors, and we set
711 the arrow lengths at each vertex proportional to the length of the (circular) mean vector. We also
712 indicated any significant results ($p < 0.05$, corrected using the Benjamani-Hochberg procedure) by
713 coloring the arrows in blue (darker blue denotes a lower p -value, i.e., a longer mean vector); all
714 tests with $p \geq 0.05$ are displayed in gray and given a lower opacity value.

715 **Searchlight fMRI analyses**

716 In Figure ??, we present two analyses aimed at identifying brain regions whose responses (as partic-
717 ipants viewed the episode) exhibited a particular temporal structure. We developed a searchlight
718 analysis wherein we constructed a $5 \times 5 \times 5$ cube of voxels (following ?) centered on each voxel
719 in the brain, and for each of these cubes, computed the temporal correlation matrix of the voxel
720 responses during episode viewing. Specifically, for each of the 1976 volumes collected during
721 episode viewing, we correlated the activity patterns in the given cube with the activity patterns
722 (in the same cube) collected during every other timepoint. This yielded a 1976×1976 correlation
723 matrix for each cube. Note: participant 5's scan ended 75s early, and in ?'s publicly released
724 dataset, their scan data was zero-padded to match the length of the other participants'. For our
725 searchlight analyses, we removed this padded data (i.e., the last 50 TRs), resulting in a 1925×1925
726 correlation matrix for each cube in participant 5's brain.

727 Next, we constructed a series of "template" matrices. The first template reflected the timecourse
728 of the episode's topic trajectory, and the others reflected the timecourse of each participant's recall
729 trajectory. To construct the episode template, we computed the correlations between the topic
730 proportions estimated for every pair of TRs (prior to segmenting the trajectory into discrete events;
731 i.e., the correlation matrix shown in Figs. ??B and ??A). We constructed similar temporal correlation
732 matrices for each participant's recall topic trajectory (Figs. ??D, S4). However, to correct for length
733 differences and potential non-linear transformations between viewing time and recall time, we
734 first used dynamic time warping (?) to temporally align participants' recall topic trajectories with

735 the episode topic trajectory. An example correlation matrix before and after warping is shown
736 in Fig. ??B. This yielded a 1976×1976 correlation matrix for the episode template and for each
737 participant’s recall template.

738 The temporal structure of the episode’s content (as described by our model) is captured in the
739 block-diagonal structure of the episode’s temporal correlation matrix (e.g., Figs. ??B, ??A), with time
740 periods of thematic stability represented as dark blocks of varying sizes. Inspecting the episode
741 correlation matrix suggests that the episode’s semantic content is highly temporally specific (i.e.,
742 the correlations between topic vectors from distant timepoints are almost all near zero). By contrast,
743 the activity patterns of individual (cubes of) voxels can encode relatively limited information on
744 their own, and their activity frequently contributes to multiple separate functions (????). By
745 nature, these two attributes give rise to similarities in activity across large timescales that may not
746 necessarily reflect a single task. To enable a more sensitive analysis of brain regions whose shifts
747 in activity patterns mirrored shifts in the semantic content of the episode or recalls, we restricted
748 the temporal correlations we considered to the timescale of semantic information captured by our
749 model. Specifically, we isolated the upper triangle of the episode correlation matrix and created a
750 “proximal correlation mask” that included only diagonals from the upper triangle of the episode
751 correlation matrix up to the first diagonal that contained no positive correlations. Applying this
752 mask to the full episode correlation matrix was equivalent to excluding diagonals beyond the
753 corner of the largest diagonal block. In other words, the timescale of temporal correlations we
754 considered corresponded to the longest period of thematic stability in the episode, and by extension
755 the longest period of thematic stability in participants’ recalls and the longest period of stability
756 we might expect to see in voxel activity arising from processing or encoding episode content.
757 Figure ?? shows this proximal correlation mask applied to the temporal correlation matrices for
758 the episode, an example participant’s (warped) recall, and an example cube of voxels from our
759 searchlight analyses.

760 To determine which (cubes of) voxel responses matched the episode template, we correlated
761 the proximal diagonals from the upper triangle of the voxel correlation matrix for each cube with
762 the proximal diagonals from episode template matrix (?). This yielded, for each participant, a

763 voxelwise map of correlation values. We then performed a one-sample t -test on the distribution of
764 (Fisher z -transformed) correlations at each voxel, across participants. This resulted in a value for
765 each voxel (cube), describing how reliably its timecourse followed that of the episode.

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

777 We used an analogous procedure to identify which voxels' responses reflected the recall tem-
778 plates. For each participant, we correlated the proximal diagonals from the upper triangle of the
779 correlation matrix for each cube of voxels with the proximal diagonals from the upper triangle
780 of their (time-warped) recall correlation matrix. As in the episode template analysis, this yielded
781 a voxelwise map of correlation coefficients for each participant. However, whereas the episode
782 analysis compared every participant's responses to the same template, here the recall templates
783 were unique for each participant. As in the analysis described above, we t -scored the (Fisher z -
784 transformed) voxelwise correlations, and used the same permutation procedure we developed for
785 the episode responses to ensure specificity to the recall timeseries and assign significance values.
786 To create the map in Figure ??D we again thresholded out any voxels whose scores were below the
787 95th percentile of the permutation-derived null distribution.

788 **Neurosynth decoding analyses**

789 Neurosynth parses a massive online database of over 14,000 neuroimaging studies and constructs
790 meta-analysis images for over 13,000 psychology- and neuroscience-related terms, based on NIfTI
791 images accompanying studies where those terms appear at a high frequency. Given a novel image
792 (tagged with its value type; e.g., t -, F - or p -statistics), Neurosynth returns a list of terms whose
793 meta-analysis images are most similar. Our permutation procedure yielded, for each of the two
794 searchlight analyses, a voxelwise map of z -values. These maps describe the extent to which each
795 voxel *specifically* reflected the temporal structure of the episode or individuals' recalls (i.e., relative
796 to the null distributions of phase-shifted values). We inputted the two statistical maps described
797 above to Neurosynth to create a list of the 10 most representative terms for each map.

798 **References**

- 799 Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control
800 processes. In Spence, K. W. and Spence, J. T., editors, *The psychology of learning and motivation*,
801 volume 2, pages 89–105. Academic Press, New York.
- 802 Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017).
803 Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–
804 721.
- 805 Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas
806 during narrative perception. *Journal of Neuroscience*, 38(45):9689–9699.
- 807 Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In
808 *KDD workshop*, volume 10, pages 359–370.
- 809 Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International
810 Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, US. ACM.

- 811 Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*
812 *Learning Research*, 3:993 – 1022.
- 813 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam,
814 P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R.,
815 Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S.,
816 Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020).
817 Language models are few-shot learners. *arXiv*, 2005.14165.
- 818 Brunec, I. K., Moscovitch, M. M., and Barense, M. D. (2018). Boundaries shape cognitive represen-
819 tations of spaces and events. *Trends in Cognitive Sciences*, 22(7):637–650.
- 820 Capota, M., Turek, J., Chen, P.-H., Zhu, X., Manning, J. R., Sundaram, N., Keller, B., Wang, Y., and
821 Shin, Y. S. (2017). Brain imaging analysis kit.
- 822 Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes,
823 M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.
824 *arXiv*, 1803.11175.
- 825 Charron, S. and Koechlin, E. (2010). Divided representations of current goals in the human frontal
826 lobes. *Science*, 328(5976):360–363.
- 827 Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared
828 memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*,
829 20(1):115.
- 830 Clewett, D. and Davachi, L. (2017). The ebb and flow of experience determines the temporal
831 structure of memory. *Curr Opin Behav Sci*, 17:186–193.
- 832 Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions.
833 *Theory of Probability & Its Applications*, 15(3):458–486.
- 834 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the
835 sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–1286.

- 836 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological*
837 *Science*, 22(2):243–252.
- 838 Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd.
- 839 Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2001). Categorical representation of
840 visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316.
- 841 Hasson, U., Chen, J., and Honey, C. J. (2015). Hierarchical process memory: memory as an integral
842 component of information processing. *Trends in Cognitive Science*, 19(6):304–315.
- 843 Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal
844 receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550.
- 845 Heusser, A. C., Ezzyat, Y., Shiff, I., and Davachi, L. (2018a). Perceptual boundaries cause mnemonic
846 trade-offs between local boundary processing and across-trial associative binding. *Journal of*
847 *Experimental Psychology Learning, Memory, and Cognition*, 44(7):1075–1090.
- 848 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail: a
849 Python toolbox for analyzing and plotting free recall data. *The Journal of Open Source Software*,
850 10.21105/joss.00424.
- 851 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018b). HyperTools: a Python
852 toolbox for gaining geometric insights into high-dimensional data. *Journal of Machine Learning*
853 *Research*, 18(152):1–6.
- 854 Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal*
855 *of Mathematical Psychology*, 46:269–299.
- 856 Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., and H., E.
857 (2014). A unified mathematical framework for coding time, space, and sequences in the medial
858 temporal lobe. *Journal of Neuroscience*, 34(13):4692–4707.
- 859 Howard, M. W., Viskontas, I. V., Shankar, K. H., and Fried, I. (2012). Ensembles of human MTL
860 neurons “jump back in time” in response to a repeated stimulus. *Hippocampus*, 22:1833–1847.

- 861 Huk, A., Bonnen, K., and He, B. J. (2018). Beyond trial-based paradigms: continuous behavior, on-
862 going neural activity, and naturalistic stimuli. *Journal of Neuroscience*, 10.1523/JNEUROSCI.1920-
863 17.2018.
- 864 Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural
865 speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532:453–458.
- 866 Huth, A. G., Nisimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes
867 the representation of thousands of object and action categories across the human brain. *Neuron*,
868 76(6):1210–1224.
- 869 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24:103–109.
- 870 Kahana, M. J. (2012). *Foundations of Human Memory*. Oxford University Press, New York, NY.
- 871 Koriat, A. and Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: distin-
872 guishing accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of*
873 *Experimental Psychology: General*, 123(3):297–315.
- 874 Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – con-
875 necting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:1 – 28.
- 876 Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic
877 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
878 104:211–240.
- 879 Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy
880 of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.
- 881 Manning, J. R. (2019). Episodic memory: mental time travel or a quantum ‘memory wave’ function?
882 *PsyArXiv*, doi:10.31234/osf.io/6zjwb.
- 883 Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors, *Handbook*
884 *of Human Memory*. Oxford University Press.

- 885 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory.
- 886 In Gazzaniga, M., editor, *The Cognitive Neurosciences, Fifth edition*, pages 557–566. MIT Press.
- 887 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns
- 888 in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National*
- 889 *Academy of Sciences, USA*, 108(31):12893–12897.
- 890 McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and
- 891 projection for dimension reduction. *arXiv*, 1802(03426).
- 892 Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations
- 893 in vector space. *arXiv*, 1301.3781.
- 894 Mueller, A., Fillion-Robin, J.-C., Boidol, R., Tian, F., Nechifor, P., yoonsubKim, Peter, Rampin, R.,
- 895 Corvellec, M., Medina, J., Dai, Y., Petrushev, B., Langner, K. M., Hong, Alessio, Ozsváld, I.,
- 896 vkolmakov, Jones, T., Bailey, E., Rho, V., IgorAPM, Roy, D., May, C., foobuzz, Piyush, Seong,
- 897 L. K., Goey, J. V., Smith, J. S., Gus, and Mai, F. (2018). WordCloud 1.5.0: a little word cloud
- 898 generator in Python. *Zenodo*, <https://zenodo.org/record/1322068#.W4tPKZNKh24>.
- 899 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*,
- 900 64:482–488.
- 901 Paller, K. A. and Wagner, A. D. (2002). Observing the transformation of experience into memory.
- 902 *Trends in Cognitive Sciences*, 6(2):93–102.
- 903 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten-
- 904 hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot,
- 905 M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine*
- 906 *Learning Research*, 12:2825–2830.
- 907 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model
- 908 of organizational processes in free recall. *Psychological Review*, 116(1):129–156.

- 909 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal*
910 *of Experimental Psychology*, 17:132–138.
- 911 Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech
912 recognition. *Proceedings of the IEEE*, 77(2):257–286.
- 913 Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are
914 unsupervised multitask learners. *OpenAI Blog*, 1(8).
- 915 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition. *Curr Opin*
916 *Behav Sci*, 17:133–140.
- 917 Ramdas, A., Trillos, N., and Cuturi, M. (2017). On wasserstein two-sample testing and related
918 families of nonparametric tests. *Entropy*, 19(2):47.
- 919 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature*
920 *Reviews Neuroscience*, 13:713 – 726.
- 921 Rishel, C. A., Huang, G., and Freedman, D. J. (2013). Independent category and spatial encoding
922 in parietal cortex. *Neuron*, 77(5):969–979.
- 923 Sigman, M. and Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during
924 dual-task performance. *Journal of Neuroscience*, 28(30):7585–7589.
- 925 Simony, E. and Chang, C. (2020). Analysis of stimulus-induced brain dynamics during naturalistic
926 paradigms. *NeuroImage*, 216:116461.
- 927 Simony, E., Honey, C. J., Chen, J., and Hasson, U. (2016). Uncovering stimulus-locked network
928 dynamics during narrative comprehension. *Nature Communications*, 7(12141):1–13.
- 929 Tompany, A. and Davachi, L. (2017). Consolidation promotes the emergence of representational
930 overlap in the hippocampus and medial prefrontal cortex. *Neuron*, 96(1):228–241.
- 931 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American Journal*
932 *of Psychology*, 35:396–401.

- 933 Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale
934 automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665.
- 935 Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research.
936 *Journal of Memory and Language*, 46:441–517.
- 937 Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception:
938 a mind-brain perspective. *Psychological Bulletin*, 133:273–293.
- 939 Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit
940 memories to other brains: Constructing shared neural representations via communication. *Cereb*
941 *Cortex*, 27(10):4988–5000.
- 942 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and
943 memory. *Psychological Bulletin*, 123(2):162 – 185.

944 Supporting information

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

946 Acknowledgements

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

952 **Author contributions**

953 Conceptualization: A.C.H. and J.R.M.; Methodology: A.C.H., P.C.F. and J.R.M.; Software: A.C.H.,
954 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.,
955 P.C.F. and J.R.M.; Supervision: J.R.M.

956 **Author information**

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