

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

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

¹Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

²Akili Interactive

Boston, MA 02110

[†]Denotes equal contribution

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

5 August 28, 2020

6 **Abstract**

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

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

33 **Introduction**

34 What does it mean to *remember* something? In traditional episodic memory experiments (e.g.,
35 list-learning or trial-based experiments; Murdock, 1962; Kahana, 1996), remembering is often cast
36 as a discrete binary operation: each studied item may be separated from the rest of one's ex-
37 perience and labeled as having been either recalled or forgotten. More nuanced studies might
38 incorporate self-reported confidence measures as a proxy for memory strength, or ask participants
39 to discriminate between recollecting the (contextual) details of an experience or having a general
40 feeling of familiarity (Yonelinas, 2002). Using well-controlled, trial-based experimental designs,
41 the field has amassed a wealth of information regarding human episodic memory (for review see
42 Kahana, 2012). However, there are fundamental properties of the external world and our mem-

ories that trial-based experiments are not well suited to capture (for review, also see Koriat and Goldsmith, 1994; Huk et al., 2018). 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 able to remember it. In classic (e.g., list-learning) memory studies, by contrast, the number or proportion of *exact* recalls is often considered to be a primary metric for assessing the quality of participants' memories. Third, one might remember the essence (or a general summary) of an experience but forget (or neglect to recount) particular low-level details. Capturing the essence of what happened is often a main goal of recounting an episodic memory to a listener, whereas the inclusion of specific low-level details is often less pertinent.

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

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

71 suggests that these sharp transitions (termed *event boundaries*) help to discretize our experiences
72 (and their mental representations) into *events* (Radvansky and Zacks, 2017; Brunec et al., 2018;
73 Heusser et al., 2018a; Clewett and Davachi, 2017; Ezzyat and Davachi, 2011; DuBrow and Davachi,
74 2013). The interplay between the stable (within-event) and transient (across-event) temporal
75 dynamics of an experience also provides a potential framework for transforming experiences into
76 memories that distills those experiences down to their essences. For example, prior work has
77 shown that event boundaries can influence how we learn sequences of items (Heusser et al.,
78 2018a; DuBrow and Davachi, 2013), navigate (Brunec et al., 2018), and remember and understand
79 narratives (Zwaan and Radvansky, 1998; Ezzyat and Davachi, 2011). This work also suggests
80 a means of distinguishing the essence of an experience from its low-level details. The overall
81 structure of events and event transitions reflects how the high-level experience unfolds (i.e., its
82 essence), while subtler event-level properties reflect low-level details. Prior research has also
83 implicated a network of brain regions (including the hippocampus and the medial prefrontal
84 cortex) in playing a critical role in transforming experiences into structured and consolidated
85 memories (Tompry and Davachi, 2017).

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

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

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

113 Results

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

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

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

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

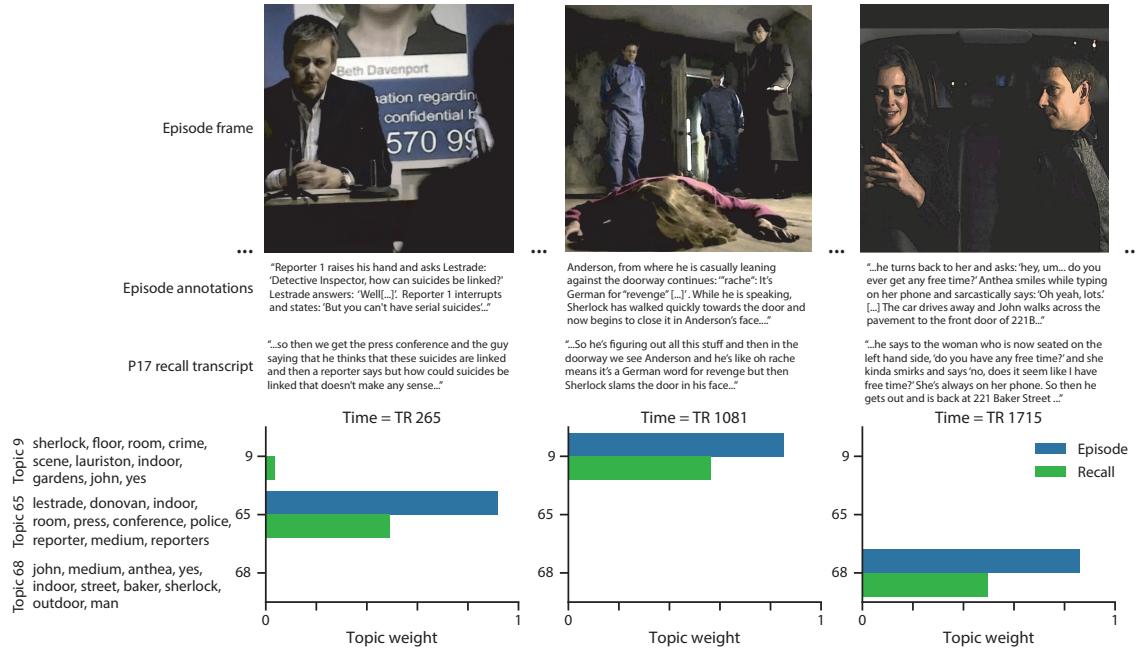


Figure 1: Topic weights in episode and recall content. We used hand-annotated descriptions of each manually identified scene from the episode to fit a topic model. Three example episode frames (first row) and their associated descriptions (second row) are displayed. The third row shows an example participant's 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.

timepoint, it was likely to be active in the following timepoint) along with *occasional rapid changes* (i.e., occasionally topic weights would change abruptly from one timepoint to the next). These two properties of the topic dynamics may be seen in the block diagonal structure of the timepoint-by-timepoint correlation matrix (Fig. 2B) and reflect the gradual drift and sudden shifts fundamental to the temporal dynamics of many real-world experiences, as well as television episodes. Given this observation, we adapted an approach devised by Baldassano et al. (2017), and used a hidden Markov model (HMM) to identify the *event boundaries* where the topic activations changed rapidly (i.e., the boundaries of the blocks in the temporal correlation matrix; event boundaries identified by the HMM are outlined in yellow in Fig. 2B). Part of our model fitting procedure required selecting an appropriate number of events into which the topic trajectory should be segmented. To accomplish this, we used an optimization procedure that maximized the difference between the topic weights for timepoints within an event versus timepoints across multiple events (see *Methods* for additional details). We then created a stable summary of the content within each episode event by averaging the topic vectors across the timepoints spanned by each event (Fig. 2C).

Given that the time-varying content of the episode could be segmented cleanly into discrete events, we wondered whether participants' recalls of the episode also displayed a similar structure. We applied the same topic model (already trained on the episode annotations) to each participant's 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 10) sentences from their transcript as a document, and computed the most probable mix of topics reflected in each timepoint's sentences. This yielded, for each participant, a number-of-windows by number-of-topics topic proportions matrix that characterized how the topics identified in the original episode were reflected in the participant's recalls. An important feature of our approach is that it allows us to compare participants' recalls to events from the original episode, despite that different participants used widely varying language to describe the events, and that those descriptions often diverged in content and quality from the episode annotations. This ability to match up conceptually related text that differs in specific vocabulary, detail, and length is an important benefit of projecting the episode and recalls into a shared topic space. An example topic



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).

¹⁸¹ proportions matrix from one participant’s recalls is shown in Figure 2D.

¹⁸² Although the example participant’s recall topic proportions matrix has some visual similarity
¹⁸³ to the episode topic proportions matrix, the time-varying topic proportions for the example par-
¹⁸⁴ ticipant’s recalls are not as sparse as those for the episode (compare Figs. 2A and D). Similarly,
¹⁸⁵ although there do appear to be periods of stability in the recall topic dynamics (i.e., most topics
¹⁸⁶ are active or inactive over contiguous blocks of time), the changes in topic activations that define
¹⁸⁷ event boundaries appear less clearly delineated in participants’ recalls than in the episode’s anno-
¹⁸⁸ tations. To examine these patterns in detail, we computed the timepoint-by-timepoint correlation
¹⁸⁹ matrix for the example participant’s recall trajectory (Fig. 2E). As in the episode correlation matrix
¹⁹⁰ (Fig. 2B), the example participant’s recall correlation matrix has a strong block diagonal structure,
¹⁹¹ indicating that their recalls are discretized into separated events. We used the same HMM-based
¹⁹² optimization procedure that we had applied to the episode’s topic proportions matrix (see *Meth-*
¹⁹³ *ods*) to estimate an analogous set of event boundaries in the participant’s recounting of the episode
¹⁹⁴ (outlined in yellow). We carried out this analysis on all 17 participants’ recall topic proportions
¹⁹⁵ matrices (Fig. S4).

¹⁹⁶ Two clear patterns emerged from this set of analyses. First, although every individual partic-
¹⁹⁷ ipant’s recalls could be segmented into discrete events (i.e., every individual participant’s recall
¹⁹⁸ correlation matrix exhibited clear block diagonal structure; Fig. S4), each participant appeared to
¹⁹⁹ have a unique *recall resolution*, reflected in the sizes of those blocks. While some participants’
²⁰⁰ recall topic proportions segmented into just a few events (e.g., Participants P4, P5, and P7), others’
²⁰¹ segmented into many shorter duration events (e.g., Participants P12, P13, and P17). This suggests
²⁰² that different participants may be recalling the episode with different levels of detail—i.e., some
²⁰³ might recount only high-level essential plot details, whereas others might recount low-level details
²⁰⁴ instead (or in addition). The second clear pattern present in every individual participant’s recall
²⁰⁵ correlation matrix was that, unlike in the episode correlation matrix, there were substantial off-
²⁰⁶ diagonal correlations. Whereas each event in the original episode was (largely) separable from the
²⁰⁷ others (Fig. 2B), in transforming those separable events into memory, participants appeared to be
²⁰⁸ integrating across multiple events, blending elements of previously recalled and not-yet-recalled

209 content into each newly recalled event (Figs. 2E, S4; also see Manning et al., 2011; Howard et al.,
210 2012; Manning, 2019).

211 The above results demonstrate that topic models capture the dynamic conceptual content of
212 the episode and participants' recalls of the episode. Further, the episode and recalls exhibit event
213 boundaries that can be identified automatically using HMMs to segment the dynamic content.
214 Next, we asked whether some correspondence might be made between the specific content of
215 the events the participants experienced in the episode, and the events they later recalled. We
216 labeled each recalled event as matching the episode event with the most similar (i.e., most highly
217 correlated) topic vector (Figs. 2G, S5). This yielded a sequence of "presented" events from the
218 original episode, and a (potentially differently ordered) sequence of "recalled" events for each
219 participant. Analogous to classic list-learning studies, we can then examine participants' recall
220 sequences by asking which events they tended to recall first (probability of first recall; Fig. 3A;
221 Atkinson and Shiffrin, 1968; Postman and Phillips, 1965; Welch and Burnett, 1924); how participants
222 most often transitioned between recalls of the events as a function of the temporal distance between
223 them (lag-conditional response probability; Fig. 3B; Kahana, 1996); and which events they were
224 likely to remember overall (serial position recall analyses; Fig. 3C; Murdock, 1962). Some of the
225 patterns we observed appeared to be similar to classic effects from the list-learning literature. For
226 example, participants had a higher probability of initiating recall with the first event in the sequence
227 (Fig. 3A) and a higher probability of transitioning to neighboring events with an asymmetric
228 forward bias (Fig. 3B). However, unlike what is typically observed in list-learning studies, we
229 did not observe patterns comparable to the primacy or recency serial position effects (Fig. 3C).
230 We hypothesized that participants might be leveraging the meaningful narrative associations and
231 references over long timescales throughout the episode.

232 Clustering scores are often used by memory researchers to characterize how people organize
233 their memories of words on a studied list (for review, see Polyn et al., 2009). We defined analogous
234 measures to characterize how participants organized their memories for episodic events (see
235 *Methods* for details). Temporal clustering refers to the extent to which participants group their recall
236 responses according to encoding position. Overall, we found that sequentially viewed episode

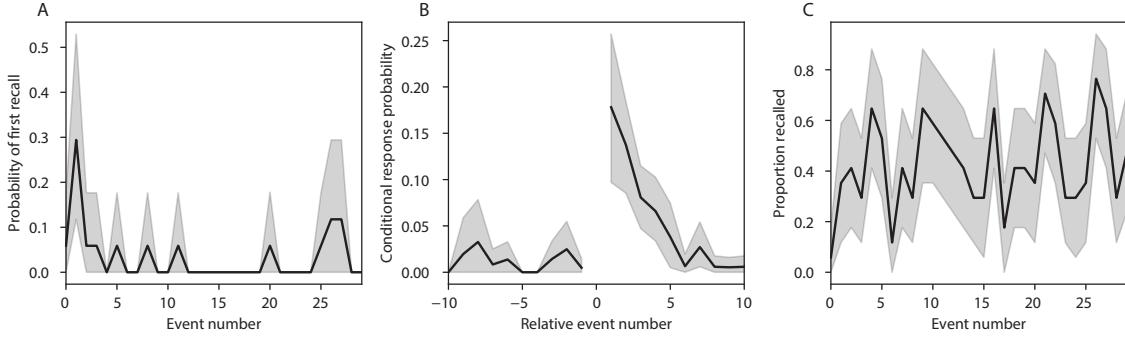


Figure 3: Naturalistic extensions of classic list-learning memory analyses. **A.** The probability of first recall as a function of the serial position of the event in the episode. **B.** The probability of recalling each event, conditioned on having most recently recalled the event *lag* events away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

events tended to appear nearby in participants' recall event sequences (mean clustering score: 0.767, SEM: 0.029). Participants with higher temporal clustering scores tended to exhibit better overall memory for the episode, according to both Chen et al. (2017)'s hand-counted numbers of recalled scenes from the episode (Pearson's $r(15) = 0.62, p = 0.008$) and the numbers of episode events that best-matched at least one recalled event (i.e., model-estimated number of recalled events; Pearson's $r(15) = 0.49, p = 0.0046$). Semantic clustering measures the extent to which participants cluster their recall responses according to semantic similarity. We found that participants tended to recall semantically similar episode events together (mean clustering score: 0.787, SEM: 0.018), and that semantic clustering score was also related to both hand-annotated (Pearson's $r(15) = 0.65, p = 0.004$) and model-estimated (Pearson's $r(15) = 0.61, p = 0.0092$) numbers of recalled events.

The above analyses illustrate how our framework for characterizing the dynamic conceptual content of naturalistic episodes enables us to carry out analyses that have traditionally been applied to much simpler list-learning paradigms. However, perhaps the most interesting aspects of memory for naturalistic episodes are those that have no list-learning analogs. The nuances of how one's memory for an event might capture some details, yet distort or neglect others, is central to how we use our memory systems in daily life. Yet when researchers study memory in highly simplified paradigms, those nuances are not typically observable. We next developed two novel

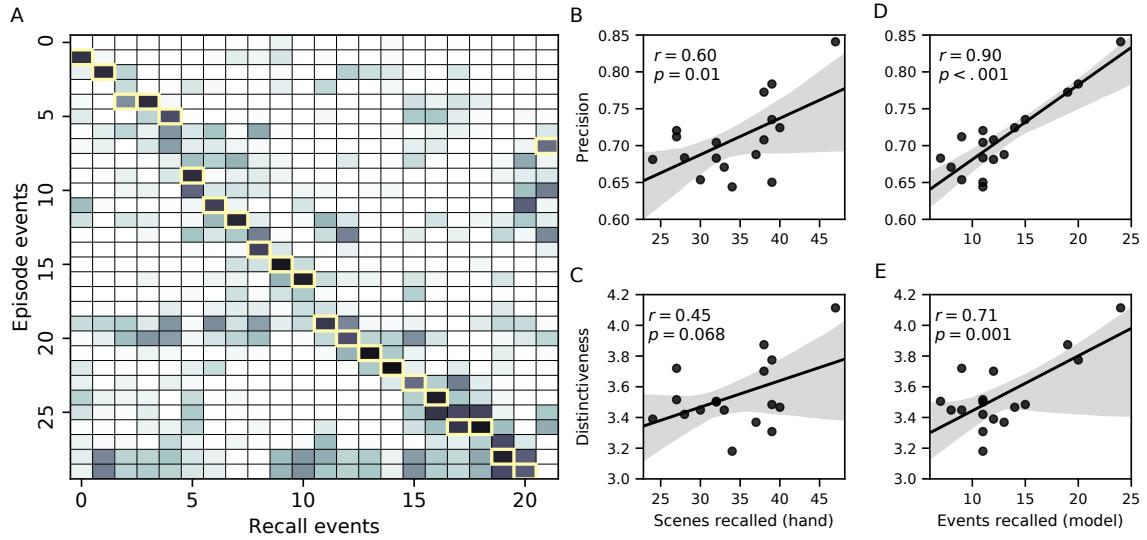


Figure 4: Novel content-based metrics of naturalistic memory: precision and distinctiveness. A. The episode-recall correlation matrix for a representative participant (17). The yellow boxes highlight the maximum correlation in each column. The example participant's overall precision score was computed as the average across correlation values in the yellow boxes. Their distinctiveness score was computed as the average (over recall events) of 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 episode events successfully recalled, as determined by our model. E. The correlation between distinctiveness and the number of episode events successfully recalled, as determined by our model.

254 continuous metrics, termed precision and distinctiveness, aimed at characterizing distortions in
255 the conceptual content of individual recalled events, and the conceptual overlap between how
256 people described different events.

257 *Precision* is intended to capture the “completeness” of recall, or how fully the presented content
258 was recapitulated in a participant’s recounting. We define a recall event’s precision as the maximum
259 correlation between the topic proportions of that recall event and any episode event (Fig. 4). In
260 other words, given that a recalled event best matches a particular episode event, more precisely
261 recalled events overlap more strongly with the conceptual content of the original episode event.
262 When a given event is assigned a blend of several topics, as is often the case (Fig. 2), a high precision
263 score requires recapitulating the relative topic proportions during recall.

264 *Distinctiveness* is intended to capture the “specificity” of recall. In other words, distinctiveness
265 quantifies the extent to which a given recalled event reflects the most similar episode event over
266 and above its reflection of other episode events. Intuitively, distinctiveness is like a normalized
267 variant of our precision metric. Whereas precision solely measures how much detail about an
268 episode was captured in someone’s recall, distinctiveness penalizes details that also pertain to
269 other episode events. We define the distinctiveness of an event’s recall as its precision expressed in
270 standard deviation units with respect to other episode events. Specifically, for a given recall event,
271 we compute the correlation between its topic vector and that of each episode event. This yields a
272 distribution of correlation coefficients (one per episode event). We subtract the mean and divide by
273 the standard deviation of this distribution to z -score the coefficients. The maximum value in this
274 distribution (which, by definition, belongs to the episode event that best matches the given recall
275 event) is that recall event’s distinctiveness score. In this way, recall events that match one episode
276 event far better than all other episode events will receive a high distinctiveness score. By contrast,
277 a recall event that matches all episode events roughly equally will receive a comparatively low
278 distinctiveness score.

279 In addition to examining how precisely and distinctively participants recalled individual events,
280 one may also use these metrics to summarize each participant’s performance by averaging across
281 a participant’s event-wise precision or distinctiveness scores. This enables us to quantify how pre-

cisely a participant tended to recall subtle within-event details, as well as how specific (distinctive) those details were to individual events from the episode. Participants' average precision and distinctiveness scores were strongly correlated ($r(15) = 0.90, p < 10^{-5}$). This indicates that participants who tended to precisely recount low-level details of episode events also tended to do so in an event-specific way (e.g., as opposed to detailing recurring themes that were present in most or all episode events; this behavior would have resulted in high precision but low distinctiveness). We found that, across participants, higher precision scores were positively correlated with both the hand-annotated ($r(15) = 0.60, p = 0.010$) and model-estimated ($r(15) = 0.90, p < 0.001$) numbers of events that participants recalled. Participants' average distinctiveness scores were also correlated with both the hand-annotated ($r(15) = 0.45, p = 0.068$) and model-estimated ($r(15) = 0.71, p = 0.001$) numbers of recalled events.

Examining individual recalls of the same episode event can provide insights into how the above precision and distinctiveness scores may be used to characterize similarities and differences in how different people describe the same shared experience. In Figure 5, we compare recalls for the same episode event (event 22) from two participants: one with a high precision score (P17), and the other with a low precision score (P6). From the HMM-identified event boundaries, we recovered the set of annotations describing the content of an example episode event (Fig. 5B), and divided them into different color-coded sections for each action or feature described. We used an analogous approach to identify the set of sentences comprising the corresponding recall events for each of the two example participants. Figure 5C shows excerpts of two participants' recall transcripts that comprised sentences between the first and last descriptions of content from the example episode event. We then colored all words describing actions and features in the transcripts shown in Panel C according to the color-coded annotations in Panel B. Visual comparison of these example recalls reveals that the more precise recall captures more of the episode event's content, and in greater detail.

Figure 6 illustrates the differences between high and low distinctiveness scores for the same event detailed in Figure 5 (i.e., event 22). Here, we have extracted the set of sentences comprising the most distinctive recall event (P9) and least distinctive recall event (P6) matched to the example

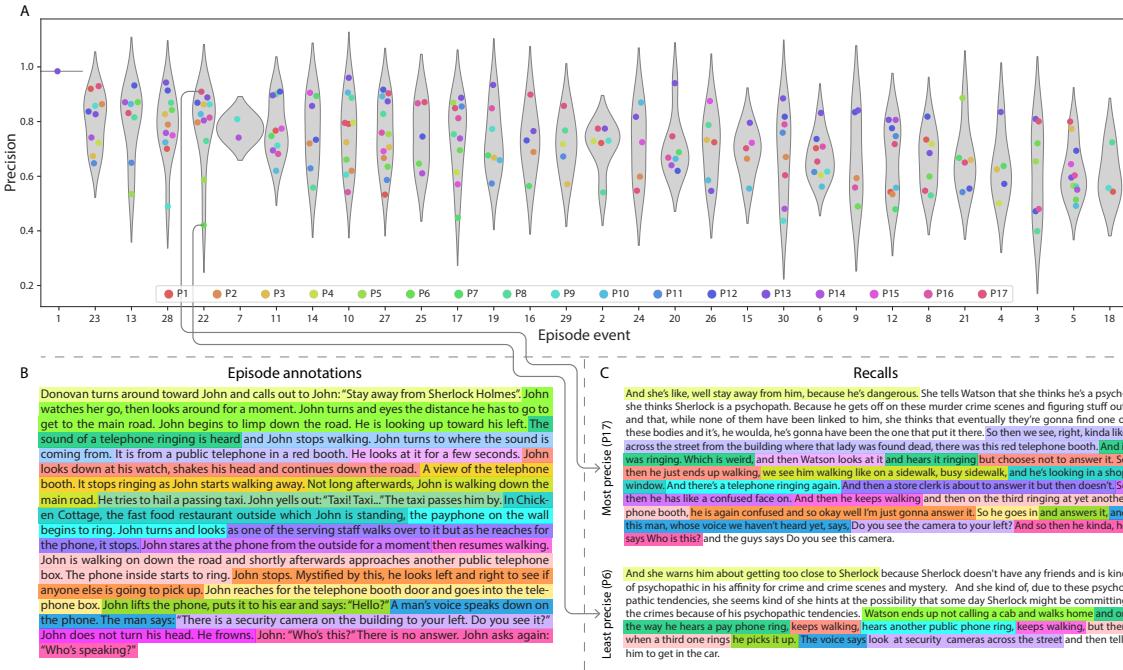


Figure 5: Precision metric reflects completeness of recall. **A.** Recall precision by episode event. Grey violin plots display kernel density estimates for the distribution of recall precision scores for a single episode event. Colored dots within each violin plot represent individual participants' 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 Chen et al., 2017) 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 B are highlighted in the corresponding color. The text highlighted in gray denotes a (rare) false recall. The unhighlighted text denotes correctly recalled information about other episode events.

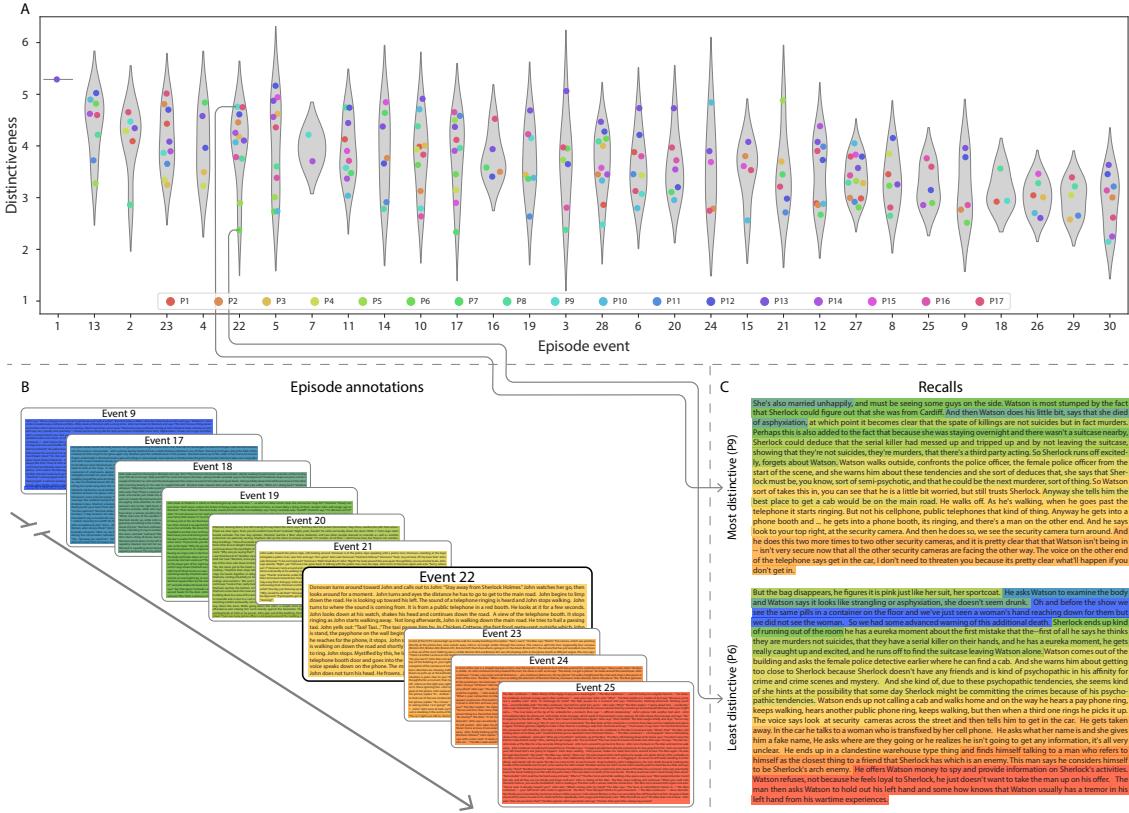


Figure 6: Distinctiveness metric reflects specificity of recall. **A.** Recall distinctiveness by episode event. Kernel density estimates for each episode event's distribution of recall distinctiveness scores, analogous to Fig. 5A. **B.** The sets of “Narrative Details” episode annotations (generated by Chen et al., 2017) for scenes comprising episode events described by the example participants in panel C. Each event's text is highlighted in a different color. **C.** The sentences comprising the most distinctive (P9) and least distinctive (P6) participants' recalls of episode event 22. Sections of recall describing each each episode event in Panel B are highlighted with the corresponding color.

310 episode event (Fig. 6C). We also extracted the annotations for the example episode event, as well
311 as those from each other episode event whose content the example participants' single recall
312 events described (Fig. 6B). We assigned each episode event a unique color (Panel B) and colored
313 each recalled phrase or sentence (Panel C) according to the episode events they best matched.
314 Visual inspection of Panel C reveals that the most distinctive recall's content is tightly concentrated
315 around event 22, whereas the least distinctive recall incorporates content from a much wider range
316 of episode events.

317 The preceding analyses sought to characterize how participants' recounts of individual
318 episode events captured the low-level details of each event. Next we sought to characterize how
319 participants' recounts of the full episode captured its high-level essence— i.e., the shape of the
320 episode's trajectory through topic space. To visualize the essence of the episode and each partici-
321 pant's recall trajectory (Heusser et al., 2018b), we projected the topic proportions matrices for the
322 episode and recalls onto a shared two-dimensional space using Uniform Manifold Approximation
323 and Projection (UMAP; McInnes et al., 2018). In this lower-dimensional space, each point repre-
324 sents a single episode or recall event, and the distances between the points reflect the distances
325 between the events' associated topic vectors (Fig. 7). In other words, events that are nearer to each
326 other in this space are more semantically similar, and those that are farther apart are less so.

327 Visual inspection of the episode and recall topic trajectories reveals a striking pattern. First, the
328 topic trajectory of the episode (which reflects its dynamic content; Fig. 7A) is captured nearly per-
329 fectly by the averaged topic trajectories of participants' recalls (Fig. 7B). To assess the consistency
330 of these recall trajectories across participants, we asked: given that a participant's recall trajectory
331 had entered a particular location in the reduced topic space, could the position of their *next* recalled
332 event be predicted reliably? For each location in the reduced topic space, we computed the set of
333 line segments connecting successively recalled events (across all participants) that intersected that
334 location (see *Methods* for additional details). We then computed (for each location) the distribu-
335 tion of angles formed by the lines defined by those line segments and a fixed reference line (the
336 *x*-axis). Rayleigh tests revealed the set of locations in topic space at which these across-participant
337 distributions exhibited reliable peaks (blue arrows in Fig. 7B reflect significant peaks at $p < 0.05$,

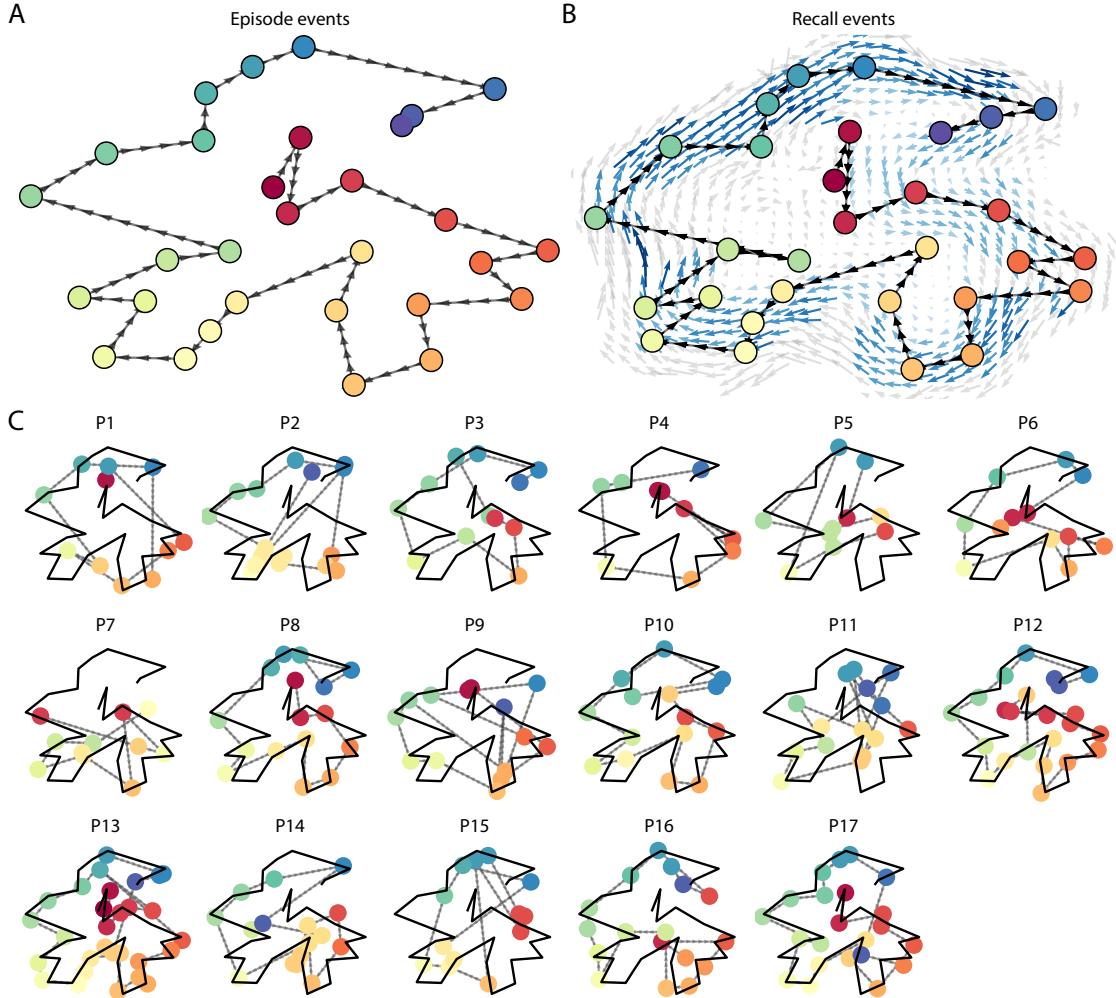


Figure 7: Trajectories through topic space capture the dynamic content of the episode and recalls. All panels: the topic proportion matrices have been projected onto a shared two-dimensional space using UMAP. **A.** The two-dimensional topic trajectory taken by the episode of *Sherlock*. Each dot indicates 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).

338 corrected). We observed that the locations traversed by nearly the entire episode trajectory exhib-
339 ited such peaks. In other words, participants' recalls exhibited similar trajectories to each other
340 that also matched the trajectory of the original episode (Fig. 7C). This is especially notable when
341 considering the fact that the numbers of events participants recalled (dots in Fig. 7C) varied con-
342 siderably across people, and that every participant used different words to describe what they had
343 remembered happening in the episode. Differences in the numbers of remembered events appear
344 in participants' trajectories as differences in the sampling resolution along the trajectory. We note
345 that this framework also provides a means of disentangling classic "proportion recalled" measures
346 (i.e., the proportion of episode events described in participants' recalls) from participants' abilities
347 to recapitulate the episode's essence (i.e., the similarity between the shapes of the original episode
348 trajectory and that defined by each participant's recounting of the episode).

349 In addition to enabling us to visualize the episode's high-level essence, describing the episode
350 as a geometric trajectory also enables us to drill down to individual words and quantify how each
351 word relates to the memorability of each event. This provides another approach to examining
352 participants' recall for low-level details beyond the precision and distinctiveness measures we
353 defined above. The results displayed in Figures 3C and 5A suggest that certain events were
354 remembered better than others. Given this, we next asked whether the events were generally
355 remembered precisely or imprecisely tended to reflect particular content. Because our analysis
356 framework projects the dynamic episode content and participants' recalls into a shared space, and
357 because the dimensions of that space represent topics (which are, in turn, sets of weights over
358 known words in the vocabulary), we are able to recover the weighted combination of words that
359 make up any point (i.e., topic vector) in this space. We first computed the average precision with
360 which participants recalled each of the 30 episode events (Fig. 8A; note that this result is analogous
361 to a serial position curve created from our precision metric). We then computed a weighted
362 average of the topic vectors for each episode event, where the weights reflected how precisely
363 each event was recalled. To visualize the result, we created a "wordle" image (Mueller et al., 2018)
364 where words weighted more heavily by more precisely-remembered topics appear in a larger font
365 (Fig. 8B, green box). Across the full episode, content that reflected topics necessary to convey the

366 central focus of the episode (e.g., the names of the two main characters, "Sherlock" and "John,"
367 and the address of a major recurring location, "221B Baker Street") were best remembered. An
368 analogous analysis revealed which themes were less-precisely remembered. Here in computing
369 the weighted average over events' topic vectors, we weighted each event in *inverse* proportion to
370 its average precision (Fig. 8B, red box). The least precisely remembered episode content reflected
371 information that was extraneous to the episode's essence, such as the proper names of relatively
372 minor characters (e.g., "Mike," "Molly," and "Lestrade") and locations (e.g., "St. Bartholomew's
373 Hospital").

374 A similar result emerged from assessing the topic vectors for individual episode and recall
375 events (Fig. 8C). Here, for each of the three most and least precisely remembered episode events, we
376 have constructed two wordles: one from the original episode event's topic vector (left) and a second
377 from the average recall topic vector for that event (right). The three most precisely remembered
378 events (circled in green) correspond to scenes integral to the central plot-line: a mysterious figure
379 spying on John in a phone booth; John meeting Sherlock at Baker St. to discuss the murders;
380 and Sherlock laying a trap to catch the killer. Meanwhile, the least precisely remembered events
381 (circled in red) reflect scenes that comprise minor plot points: a video of singing cartoon characters
382 that participants viewed in an introductory clip prior to the main episode; John asking Molly
383 about Sherlock's habit of over-analyzing people; and Sherlock noticing evidence of Anderson's
384 and Donovan's affair.

385 The results thus far inform us about which aspects of the dynamic content in the episode partic-
386 ipants watched were preserved or altered in participants' memories. We next carried out a series
387 of analyses aimed at understanding which brain structures might facilitate these preservations
388 and transformations between the external world and memory. In the first analysis, we sought to
389 identify brain structures that were sensitive to the dynamic unfolding of the episode's content,
390 as characterized by its topic trajectory. We used a searchlight procedure to identify clusters of
391 voxels whose activity patterns displayed a proximal temporal correlation structure (as participants
392 watched the episode) matching that of the original episode's topic proportions (Fig. 9A; see *Methods*
393 for additional details). In a second analysis, we sought to identify brain structures whose responses

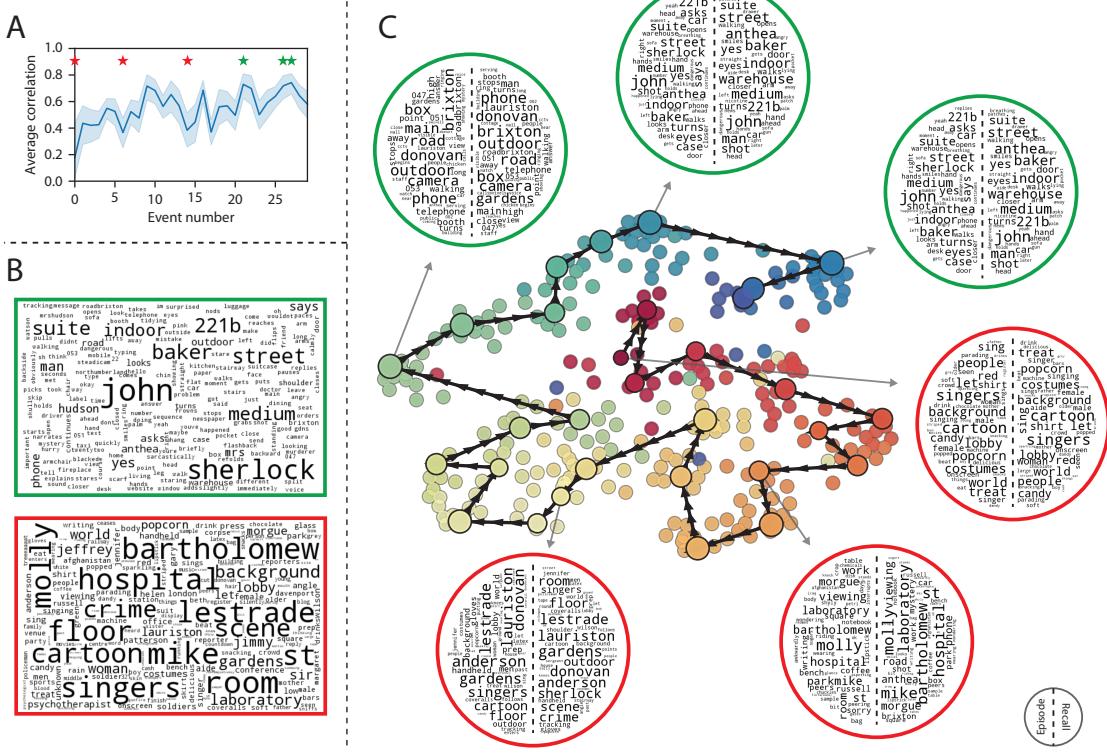


Figure 8: Language used in the most and least precisely remembered events. **A.** Average precision (episode event-recall event topic vector correlation) across participants for each episode event. Here we defined each episode event's precision for each participant as the correlation between its topic 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 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 7. The dots outlined in black denote episode events (dot size is proportional to each event's precision). The dots without black outlines denote recall events from each participant. All dots are colored using the same scheme as Figure 7A. 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.

394 (during episode viewing) reflected how each participant would later structure their *recounting* of
395 the episode. We used a searchlight procedure to identify clusters of voxels whose proximal tempo-
396 ral correlation matrices matched that of the topic proportions matrix for each participant's recall
397 transcript (Figs. 9B; see *Methods* for additional details). To ensure our searchlight procedure iden-
398 tified regions *specifically* sensitive to the temporal structure of the episode or recalls (i.e., rather
399 than those with a temporal autocorrelation length similar to that of the episode and recalls), we
400 performed a phase shift-based permutation correction (see *Methods* for additional details). As
401 shown in Figure 9C, the episode-driven searchlight analysis revealed a distributed network of
402 regions that may play a role in processing information relevant to the narrative structure of the
403 episode. Similarly, the recall-driven searchlight analysis revealed a second network of regions
404 (Fig. 9D) that may facilitate a person-specific transformation of one's experience into memory. The
405 top ten Neurosynth terms (Yarkoni et al., 2011) associated with each (unthresholded) map are dis-
406 played in each panel. In identifying regions whose responses to ongoing experiences reflect how
407 those experiences will be remembered later, this latter analysis extends classic *subsequent memory*
408 analyses (e.g., Paller and Wagner, 2002) to the domain of naturalistic experiences.

409 The searchlight analyses described above yielded two distributed networks of brain regions,
410 whose activity timecourses mirrored to the temporal structure of the episode (Fig. 9C) or partic-
411 ipants' eventual recalls (Fig. 9D). We next sought to gain greater insight into the structures and
412 functional networks our results reflected. To accomplish this, we performed an additional, ex-
413 ploratory analysis using Neurosynth (Yarkoni et al., 2011). Given an arbitrary statistical map as
414 input, Neurosynth performs a massive automated meta-analysis, returning a ranked list of terms
415 reported in papers with similar significance maps. We ran Neurosynth on the significance maps
416 for the episode- and recall-driven searchlight analyses. These maps, along with the 10 terms with
417 maximally similar meta-analysis images identified by Neurosynth are shown in Figure 9.

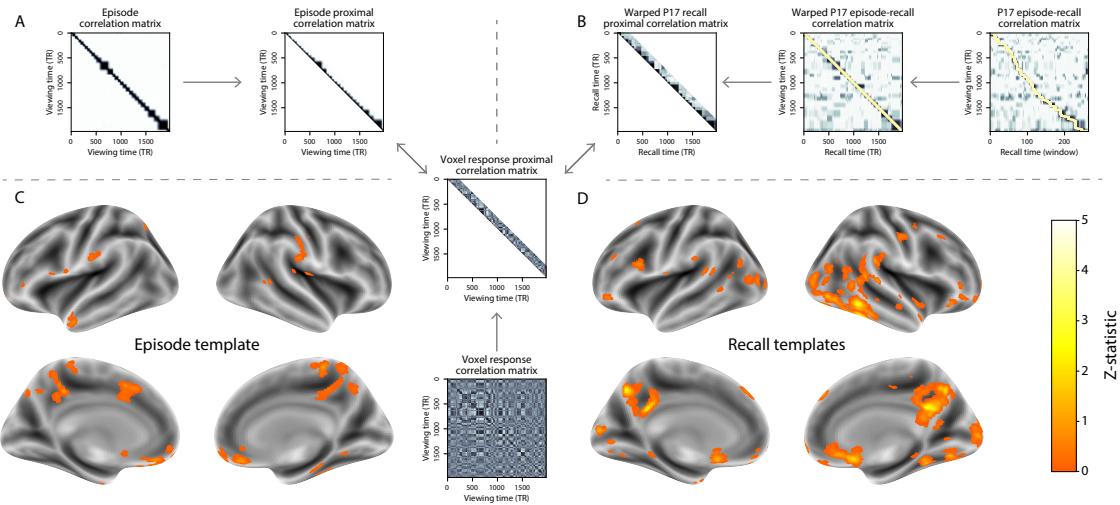


Figure 9: Brain structures that underlie the transformation of experience into memory. **A.** We isolated the proximal diagonals from the upper triangle of the episode correlation matrix, and applied this same diagonal mask to the voxel response correlation matrix for each cube of voxels in the brain. 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 (Berndt and Clifford, 1994) 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 or 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.

418 **Discussion**

419 Explicitly modeling the dynamic content of a naturalistic stimulus and participants' memories
420 enabled us to connect the present study of naturalistic recall with an extensive prior literature that
421 has used list-learning paradigms to study memory (for review see Kahana, 2012), as in Figure 3.
422 We found some similarities between how participants in the present study recounted a television
423 episode and how participants typically recall memorized random word lists. However, our broader
424 claim is that word lists miss out on fundamental aspects of naturalistic memory more like the sort
425 of memory we rely on in everyday life. For example, there are no random word list analogs
426 of character interactions, conceptual dependencies between temporally distant episode events,
427 the sense of solving a mystery that pervades the *Sherlock* episode, or the myriad other features
428 of the episode that convey deep meaning. Nevertheless, each of these properties affects how
429 people process the ongoing episode as they are watching it, and how they remember it later. The
430 overarching goal of the present study is to characterize how the rich dynamics of the episode affect
431 the rich behavioral and neural dynamics of how people remember it.

432 Our work casts remembering as reproducing (behaviorally and neurally) the topic trajectory,
433 or "shape," of an experience. When we characterized memory for a television episode using
434 this framework, we found that every participant's recounting of the episode recapitulated the
435 low spatial frequency details of the shape of its trajectory through topic space. We termed this
436 narrative scaffolding the episode's *essence*. Where participants' behaviors varied most was in
437 their tendencies to recount specific low-level details from each episode event. Geometrically,
438 this appears as high spatial frequency distortions in participants' recall trajectories relative to the
439 trajectory of the original episode. We developed metrics to characterize the precision (recovery of
440 any and all event-level information) and distinctiveness (recovery of event-specific information).
441 We also used word cloud visualizations to interpret the details of these event-level distortions.

442 The neural analyses we carried out (Fig. 9) also leveraged our geometric framework for char-
443 acterizing the shapes of the episode and participants' recounts. We identified one network
444 of regions whose responses tracked with temporal correlations in the conceptual content of the

445 episode (as quantified by topic models applied to a set of annotations about the episode). This
446 network included orbitofrontal cortex, ventromedial prefrontal cortex, striatum, among others. As
447 reviewed by Ranganath and Ritchey (2012), several of these regions are members of the *anterior*
448 *temporal system*, which has been implicated in assessing the familiarity of ongoing experiences,
449 emotional processing, social cognition, and rewards. A second network we identified tracked
450 with temporal correlations in the idiosyncratic conceptual content of participants' recounts of
451 the episode. This network included occipital cortex, extrastriate cortex, fusiform gyrus, and the
452 precuneus. Several of these regions are members of the *posterior medial system* (Ranganath and
453 Ritchey, 2012), which has been implicated in matching incoming cues about the current situation
454 to internally maintained *situation models* that specify the parameters and expectations inherent to
455 the current situation (also see Zacks et al., 2007; Zwaan and Radvansky, 1998). Taken together, our
456 results support the notion that these two (partially overlapping) networks work in coordination
457 to make sense of our ongoing experiences, distort them in a way that links them with our prior
458 knowledge and experiences, and encodes those distorted representations into memory for our later
459 use.

460 Our general approach draws inspiration from prior work aimed at elucidating the neural and
461 behavioral underpinnings of how we process dynamic naturalistic experiences and remember them
462 later. Our approach to identifying neural responses to naturalistic stimuli (including experiences)
463 entails building an explicit model of the stimulus and searching for brain regions whose responses
464 are consistent with the model (also see Huth et al., 2012, 2016). In prior work, a series of studies
465 from Uri Hasson's group (Lerner et al., 2011; Simony et al., 2016; Chen et al., 2017; Baldassano et al.,
466 2017; Zadbood et al., 2017) have developed a clever alternative approach: rather than building an
467 explicit stimulus model, these studies instead search for brain responses (while experiencing the
468 stimulus) that are reliably similar across individuals. So called *inter-subject correlation* (ISC) and
469 *inter-subject functional connectivity* (ISFC) analyses effectively treat other people's brain responses to
470 the stimulus as a "model" of how its features change over time (also see Simony and Chang, 2020).
471 These purely brain-driven approaches are well-suited to identifying which brain structures exhibit
472 similar stimulus-driven responses across individuals. Further, because neural response dynamics

473 are observed data (rather than model approximations), such approaches do not require a detailed
474 understanding of which stimulus properties or features might be driving the observed responses.
475 However, this also means that the specific stimulus features driving those responses are typically
476 opaque to the researcher. Our approach is complementary. By explicitly modeling the stimulus
477 dynamics, we are able to relate specific stimulus features to behavioral and neural dynamics.
478 However, when our model fails to accurately capture the stimulus dynamics that are truly driving
479 behavioral and neural responses, our approach necessarily yields an incomplete characterization
480 of the neural basis of the processes we are studying.

481 Other recent work has used HMMs to discover latent event structure in neural responses to
482 naturalistic stimuli (Baldassano et al., 2017). By applying HMMs to our explicit models of stimulus
483 and memory dynamics, we gain a more direct understanding of those state dynamics. For example,
484 we found that although the events comprising each participant’s recalls recapitulated the episode’s
485 essence, participants differed in the *resolution* of their recounting of low-level details. In turn,
486 these individual behavioral differences were reflected in differences in neural activity dynamics as
487 participants watched the television episode.

488 Our approach also draws inspiration from the growing field of word embedding models. The
489 topic models (Blei et al., 2003) we used to embed text from the episode annotations and participants’
490 recall transcripts are just one of many models that have been studied in an extensive literature.
491 The earliest approaches to word embedding, including latent semantic analysis (Landauer and
492 Dumais, 1997), used word co-occurrence statistics (i.e., how often pairs of words occur in the same
493 documents contained in the corpus) to derive a unique feature vector for each word. The feature
494 vectors are constructed so that words that co-occur more frequently have feature vectors that are
495 closer (in Euclidean distance). Topic models are essentially an extension of those early models, in
496 that they attempt to explicitly model the underlying causes of word co-occurrences by automatically
497 identifying the set of themes or topics reflected across the documents in the corpus. More recent
498 work on these types of semantic models, including word2vec (Mikolov et al., 2013), the Universal
499 Sentence Encoder (Cer et al., 2018), GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020) use
500 deep neural networks to attempt to identify the deeper conceptual representations underlying each

501 word. Despite the availability of these sophisticated deep learning-based embedding models, we
502 chose to prioritize interpretability of the embedding dimensions (e.g., Fig. 8) over raw performance,
503 e.g. by a predefined benchmark. Nevertheless, we note that our general framework is, in principle,
504 robust to the specific choice of language model as well as other aspects of our computational
505 pipeline. For example, the word embedding model, timeseries segmentation model, and the
506 episode-recall matching function could all be customized to suit a particular question space or
507 application. Indeed, for some questions, interpretability of the embeddings may not be a priority,
508 and thus other text embedding approaches (including the deep learning-based models described
509 above) may be preferable. Future work will be needed to explore the influence of particular models
510 on our framework’s predictions and performance.

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

524 **Methods**

525 **Experimental design and data collection**

526 Data were collected by Chen et al. (2017). In brief, participants ($n = 22$) viewed the first 48 minutes
527 of “A Study in Pink”, the first episode of the BBC television series *Sherlock*, while fMRI volumes
528 were collected (TR = 1500 ms). Participants were pre-screened to ensure they had never seen any
529 episode of the show before. The stimulus was divided into a 23 min (946 TR) and a 25 min (1030 TR)
530 segment to mitigate technical issues related to the scanner. After finishing the clip, participants
531 were instructed to (quoting from Chen et al., 2017) “describe what they recalled of the [episode]
532 in as much detail as they could, to try to recount events in the original order they were viewed
533 in, and to speak for at least 10 minutes if possible but that longer was better. They were told that
534 completeness and detail were more important than temporal order, and that if at any point they
535 realized they had missed something, to return to it. Participants were then allowed to speak for
536 as long as they wished, and verbally indicated when they were finished (e.g., ‘I’m done’).” Five
537 participants were dropped from the original dataset due to excessive head motion (2 participants),
538 insufficient recall length (2 participants), or falling asleep during stimulus viewing (1 participant),
539 resulting in a final sample size of $n = 17$. For additional details about the experimental procedure
540 and scanning parameters, see Chen et al. (2017). The experimental protocol was approved by
541 Princeton University’s Institutional Review Board.

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

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

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

556 **Data and code availability**

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

559 **Statistics**

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

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

566 **Topic modeling**

567 The input to the topic model we trained to characterize the dynamic content of the episode
568 comprised 998 hand-generated annotations of short (mean: 2.96s) scenes spanning the video
569 clip (Chen et al., 2017 generated 1000 annotations total; we removed two annotations referring to
570 a break between the first and second scan sessions, during which no fMRI data was collected).
571 We concatenated the text for all of the annotated features within each segment, creating a “bag of
572 words” describing each scene and performed some minor preprocessing (e.g., stemming possessive
573 nouns and removing punctuation). We then re-organized the text descriptions into overlapping
574 sliding windows spanning (up to) 50 scenes each. In other words, we estimated the “context”

575 for each scene using the text descriptions of the preceding 25 scenes, the present scene, and the
576 following 24 scenes. To model the context for scenes near the beginning of the episode (i.e., within
577 25 scenes of the beginning or end), we created overlapping sliding windows that grew in size
578 from one scene to the full length. We also tapered the sliding window lengths at the end of the
579 episode, whereby scenes within fewer than 24 scenes of the end of the episode were assigned
580 sliding windows that extended to the end of the episode. This procedure ensured that each scene's
581 content was represented in the text corpus an equal number of times.

582 We trained our model using these overlapping text samples with `scikit-learn` (version 0.19.1;
583 Pedregosa et al., 2011), called from our high-dimensional visualization and text analysis software,
584 `HyperTools` (Heusser et al., 2018b). Specifically, we used the `CountVectorizer` class to transform
585 the text from each window into a vector of word counts (using the union of all words across all
586 scenes as the “vocabulary,” excluding English stop words); this yielded a number-of-windows
587 by number-of-words *word count* matrix. We then used the `LatentDirichletAllocation` class
588 (`topics=100, method='batch'`) to fit a topic model (Blei et al., 2003) to the word count matrix,
589 yielding a number-of-windows (1047) by number-of-topics (100) *topic proportions* matrix. The
590 topic proportions matrix describes the gradually evolving mix of topics (latent themes) present in
591 each scene. Next, we transformed the topic proportions matrix to match the 1976 fMRI volume
592 acquisition times. We assigned each topic vector to the timepoint (in seconds) midway between the
593 beginning of the first scene and the end of the last scene in its corresponding sliding text window.
594 By doing so, we warped the linear temporal distance between consecutive topic vectors to align
595 with the inconsistent temporal distance between consecutive annotations (whose durations varied
596 greatly). We then rescaled these timepoints to 1.5s TR units, and used linear interpolation to
597 estimate a topic vector for each TR. This resulted in a number-of-TRs (1976) by number-of-topics
598 (100) matrix.

599 We created similar topic proportions matrices using hand-annotated transcripts of each partic-
600 ipant's verbal recall of the episode (annotated by Chen et al., 2017). We tokenized the transcript
601 into a list of sentences, and then re-organized the list into overlapping sliding windows spanning
602 (up to) 10 sentences each, analogously to how we parsed the episode annotations. In turn, we

603 transformed each window's sentences into a word count vector (using the same vocabulary as for
604 the episode model), and then we used the topic model already trained on the episode scenes to
605 compute the most probable topic proportions for each sliding window. This yielded a number-of-
606 windows (range: 83–312) by number-of-topics (100) topic proportions matrix for each participant.
607 These reflected the dynamic content of each participant's recalls. Note: for details on how we
608 selected the episode and recall window lengths and number of topics, see *Supporting Information*
609 and Figure S1.

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

611 We parsed the topic trajectories of the episode and participants' recalls into events using Hidden
612 Markov Models (HMMs; Rabiner, 1989). Given the topic proportions matrix (describing the mix
613 of topics at each timepoint) and a number of states, K , an HMM recovers the set of state transitions
614 that segments the timeseries into K discrete states. Following Baldassano et al. (2017), we imposed
615 an additional set of constraints on the discovered state transitions that ensured that each state was
616 encountered exactly once (i.e., never repeated). We used the BrainIAK toolbox (Capota et al., 2017)
617 to implement this segmentation.

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

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

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

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

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

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

653 **Probability of first recall (PFR).** PFR curves (Welch and Burnett, 1924; Postman and Phillips,
654 1965; Atkinson and Shiffrin, 1968) reflect the probability that an item will be recalled first as a
655 function of its serial position during encoding. To carry out this analysis, we initialized a number-
656 of-participants (17) by number-of-episode-events (30) matrix of zeros. Then for each participant,
657 we found the index of the episode event that was recalled first (i.e., the episode event whose topic
658 vector was most strongly correlated with that of the first recall event) and filled in that index in
659 the matrix with a 1. Finally, we averaged over the rows of the matrix, resulting in a 1 by 30 array
660 representing the proportion of participants that recalled an event first, as a function of the order of
661 the event's appearance in the episode (Fig. 3A).

662 **Lag conditional probability curve (lag-CRP).** The lag-CRP curve (Kahana, 1996) reflects the
663 probability of recalling a given item after the just-recalled item, as a function of their relative
664 encoding positions (or *lag*). In other words, a lag of 1 indicates that a recalled item was presented
665 immediately after the previously recalled item, and a lag of -3 indicates that a recalled item came 3
666 items before the previously recalled item. For each recall transition (following the first recall), we
667 computed the lag between the current recall event and the next recall event, normalizing by the
668 total number of possible transitions. This yielded a number-of-participants (17) by number-of-lags
669 (-29 to +29; 58 lags total excluding lags of 0) matrix. We averaged over the rows of this matrix to
670 obtain a group-averaged lag-CRP curve (Fig. 3B).

671 **Serial position curve (SPC).** SPCs (Murdock, 1962) reflect the proportion of participants that
672 remember each item as a function of the items' serial positions during encoding. We initialized
673 a number-of-participants (17) by number-of-episode-events (30) matrix of zeros. Then, for each
674 recalled event, for each participant, we found the index of the episode event that the recalled
675 event most closely matched (via the correlation between the events' topic vectors) and entered a
676 1 into that position in the matrix. This resulted in a matrix whose entries indicated whether or
677 not each event was recalled by each participant (depending on whether the corresponding entires
678 were set to one or zero). Finally, we averaged over the rows of the matrix to yield a 1 by 30 array

679 representing the proportion of participants that recalled each event as a function of the events'
680 order appearance in the episode (Fig. 3C).

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

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

699 **Novel naturalistic memory metrics**

700 **Precision.** We tested whether participants who recalled more events were also more *precise* in their
701 recollections. For each participant, we computed the average correlation between the topic vectors
702 for each recall event and those of its closest-matching episode event. This gave a single value per
703 participant representing the average precision across all recalled events. We then correlated these

704 values with both hand-annotated and model-derived (i.e., the number of unique episode events
705 matched by a participant’s recall events) memory performance.

706 **Distinctiveness.** We also considered the *distinctiveness* of each recalled event. That is, how unique
707 a participant’s description of a episode event was, versus their descriptions of other episode events.
708 We hypothesized that participants with high memory performance might describe each event in
709 a more distinctive way (relative to those with lower memory performance who might describe
710 events in a more general way). To test this hypothesis we define a distinctiveness score for each
711 recall event i as

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

712 where the average is taken over the correlation between the recall event i ’s topic vector and the
713 topic vectors from all other recall events from that participant. We averaged these distinctiveness
714 scores across all of the events recalled by the given participant to get the participant’s distinctiveness
715 score. We correlated these distinctiveness scores with hand-annotated and model-derived memory
716 performance scores across-subjects, as above.

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

722 Visualizing the episode and recall topic trajectories

723 We used the UMAP algorithm (McInnes et al., 2018) to project the 100-dimensional topic space onto
724 a two-dimensional space for visualization (Figs. 7, 8). To ensure that all of the trajectories were
725 projected onto the *same* lower dimensional space, we computed the low-dimensional embedding

726 on a “stacked” matrix created by vertically concatenating the events-by-topics topic proportions
727 matrices for the episode, across-participants average recall and all 17 individual participants’ re-
728 calls. We then separated the rows of the result (a total-number-of-events by two matrix) back into
729 individual matrices for the episode topic trajectory, across-participant average recall trajectory and
730 the trajectories for each individual participant’s recalls (Fig. 7). This general approach for dis-
731 covering a shared low-dimensional embedding for a collections of high-dimensional observations
732 follows Heusser et al. (2018b).

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

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

743 In Figure 7B, we present an analysis aimed at characterizing locations in topic space that dif-
744 ferent participants move through in a consistent way (via their recall topic trajectories). The
745 two-dimensional topic space used in our visualizations (Fig. 7) comprised a 60×60 (arbitrary
746 units) square. We tiled this space with a 50×50 grid of evenly spaced vertices, and defined a
747 circular area centered on each vertex whose radius was two times the distance between adjacent
748 vertices (i.e., 2.4 units). For each vertex, we examined the set of line segments formed by connecting
749 each pair successively recalled events, across all participants, that passed through this circle. We
750 computed the distribution of angles formed by those segments and the x -axis, and used a Rayleigh
751 test to determine whether the distribution of angles was reliably “peaked” (i.e., consistent across
752 all transitions that passed through that local portion of topic space). To create Figure 7B we drew

753 an arrow originating from each grid vertex, pointing in the direction of the average angle formed
754 by the line segments that passed within 2.4 units. We set the arrow lengths to be inversely propor-
755 tional to the p -values of the Rayleigh tests at each vertex. Specifically, for each vertex we converted
756 all of the angles of segments that passed within 2.4 units to unit vectors, and we set the arrow
757 lengths at each vertex proportional to the length of the (circular) mean vector. We also indicated
758 any significant results ($p < 0.05$, corrected using the Benjamani-Hochberg procedure) by coloring
759 the arrows in blue (darker blue denotes a lower p -value, i.e., a longer mean vector); all tests with
760 $p \geq 0.05$ are displayed in gray and given a lower opacity value.

761 **Searchlight fMRI analyses**

762 In Figure 9, we present two analyses aimed at identifying brain regions whose responses (as partic-
763 ipants viewed the episode) exhibited a particular temporal structure. We developed a searchlight
764 analysis wherein we constructed a $5 \times 5 \times 5$ cube of voxels (following Chen et al., 2017) centered
765 on each voxel in the brain, and for each of these cubes, computed the temporal correlation matrix
766 of the voxel responses during episode viewing. Specifically, for each of the 1976 volumes collected
767 during episode viewing, we correlated the activity patterns in the given cube with the activity
768 patterns (in the same cube) collected during every other timepoint. This yielded a 1976 by 1976
769 correlation matrix for each cube. Note: participant 5's scan ended 75s early, and in Chen et al.,
770 2017's publicly released dataset, their scan data was padded to match the length of the other partic-
771 ipants'. For our searchlight analyses, we removed this padded data (i.e., the last 50 TRs), resulting
772 in a 1925 by 1925 correlation matrix for each cube in participant 5's brain.

773 Next, we constructed a series of "template" matrices. The first template reflected the timecourse
774 of the episode's topic trajectory, and the others reflected the timecourse of each participant's recall
775 trajectory. To construct the episode template, we computed the correlations between the topic
776 proportions estimated for every pair of TRs (prior to segmenting the trajectory into discrete events;
777 i.e., the correlation matrix shown in Figs. 2B and 9A). We constructed similar temporal correlation
778 matrices for each participant's recall topic trajectory (Figs. 2D, S4). However, to correct for length
779 differences and potential non-linear transformations between viewing time and recall time, we

780 first used dynamic time warping (Berndt and Clifford, 1994) to temporally align participants'
781 recall topic trajectories with the episode topic trajectory. An example correlation matrix before and
782 after warping is shown in Fig. 9B. This yielded a 1976 by 1976 correlation matrix for the episode
783 template and for each participant's recall template.

784 The temporal structure of the episode's content (as described by our model) is captured in the
785 block-diagonal structure of the episode's temporal correlation matrix (e.g., Figs. 2B, 9A), with time
786 periods of thematic stability represented as dark blocks of varying sizes. Inspecting the episode
787 correlation matrix suggests that the episode's semantic content is highly temporally specific (i.e.,
788 the correlations between topic vectors from distant timepoints are almost all near zero). By contrast,
789 the activity patterns of individual (cubes of) voxels can encode relatively limited information on
790 their own, and their activity frequently contributes to multiple separate functions (Freedman
791 et al., 2001; Sigman and Dehaene, 2008; Charron and Koechlin, 2010; Rishel et al., 2013). By
792 nature, these two attributes give rise to similarities in activity across large timescales that may not
793 necessarily reflect a single task. To enable a more sensitive analysis of brain regions whose shifts
794 in activity patterns mirrored shifts in the semantic content of the episode or recalls, we restricted
795 the temporal correlations we considered to the timescale of semantic information captured by our
796 model. Specifically, we isolated the upper triangle of the episode correlation matrix and created a
797 "proximal correlation mask" that included only diagonals from the upper triangle of the episode
798 correlation matrix up to the first diagonal that contained no positive correlations. Applying this
799 mask to the full episode correlation matrix was analogous to excluding diagonals beyond the corner
800 of the largest diagonal block. In other words, the timescale of temporal correlations we considered
801 corresponded to the longest period of thematic stability in the episode, and by extension the longest
802 expected period of thematic stability in participants' recalls and the longest period of stability we
803 might expect to see in voxel activity arising from processing or encoding episode content. Figure 9
804 shows this proximal correlation mask applied to the temporal correlation matrices for the episode,
805 an example participant's (warped) recall, and an example cube of voxels from our searchlight
806 analyses.

807 To determine which (cubes of) voxel responses matched the episode template, we correlated

808 the proximal diagonals from the upper triangle of the voxel correlation matrix for each cube with
809 the proximal diagonals from episode template matrix (Kriegeskorte et al., 2008). This yielded, for
810 each participant, a voxelwise map of correlation values. We then performed a one-sample *t*-test
811 on the distribution of (Fisher *z*-transformed) correlations at each voxel, across participants. This
812 resulted in a value for each voxel (cube), describing how reliably its timecourse followed that of
813 the episode.

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

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

836 **Neurosynth decoding analyses**

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

848 **References**

- 849 Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control
850 processes. In Spence, K. W. and Spence, J. T., editors, *The psychology of learning and motivation*,
851 volume 2, pages 89–105. Academic Press, New York.
- 852 Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017).
853 Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–
854 721.
- 855 Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas
856 during narrative perception. *Journal of Neuroscience*, 38(45):9689–9699.
- 857 Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In
858 *KDD workshop*, volume 10, pages 359–370.

- 859 Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International*
860 *Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, US. ACM.
- 861 Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*
862 *Learning Research*, 3:993 – 1022.
- 863 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam,
864 P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R.,
865 Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S.,
866 Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020).
867 Language models are few-shot learners. *arXiv*, 2005.14165.
- 868 Brunec, I. K., Moscovitch, M. M., and Barense, M. D. (2018). Boundaries shape cognitive represen-
869 tations of spaces and events. *Trends in Cognitive Sciences*, 22(7):637–650.
- 870 Capota, M., Turek, J., Chen, P.-H., Zhu, X., Manning, J. R., Sundaram, N., Keller, B., Wang, Y., and
871 Shin, Y. S. (2017). Brain imaging analysis kit.
- 872 Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes,
873 M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.
874 *arXiv*, 1803.11175.
- 875 Charron, S. and Koechlin, E. (2010). Divided representations of current goals in the human frontal
876 lobes. *Science*, 328(5976):360–363.
- 877 Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared
878 memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*,
879 20(1):115.
- 880 Clewett, D. and Davachi, L. (2017). The ebb and flow of experience determines the temporal
881 structure of memory. *Curr Opin Behav Sci*, 17:186–193.
- 882 Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions.
883 *Theory of Probability & Its Applications*, 15(3):458–486.

- 884 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the
885 sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–1286.
- 886 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological
887 Science*, 22(2):243–252.
- 888 Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd.
- 889 Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2001). Categorical representation of
890 visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316.
- 891 Hasson, U., Chen, J., and Honey, C. J. (2015). Hierarchical process memory: memory as an integral
892 component of information processing. *Trends in Cognitive Science*, 19(6):304–315.
- 893 Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal
894 receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550.
- 895 Heusser, A. C., Ezzyat, Y., Shiff, I., and Davachi, L. (2018a). Perceptual boundaries cause mnemonic
896 trade-offs between local boundary processing and across-trial associative binding. *Journal of
897 Experimental Psychology Learning, Memory, and Cognition*, 44(7):1075–1090.
- 898 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail: a
899 Python toolbox for analyzing and plotting free recall data. *The Journal of Open Source Software*,
900 10.21105/joss.00424.
- 901 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018b). HyperTools: a Python
902 toolbox for gaining geometric insights into high-dimensional data. *Journal of Machine Learning
903 Research*, 18(152):1–6.
- 904 Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal
905 of Mathematical Psychology*, 46:269–299.
- 906 Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., and H., E.
907 (2014). A unified mathematical framework for coding time, space, and sequences in the medial
908 temporal lobe. *Journal of Neuroscience*, 34(13):4692–4707.

- 909 Howard, M. W., Viskontas, I. V., Shankar, K. H., and Fried, I. (2012). Ensembles of human MTL
910 neurons “jump back in time” in response to a repeated stimulus. *Hippocampus*, 22:1833–1847.
- 911 Huk, A., Bonnen, K., and He, B. J. (2018). Beyond trial-based paradigms: continuous behavior, on-
912 going neural activity, and naturalistic stimuli. *Journal of Neuroscience*, 10.1523/JNEUROSCI.1920-
913 17.2018.
- 914 Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural
915 speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532:453–458.
- 916 Huth, A. G., Nisimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes
917 the representation of thousands of object and action categories across the human brain. *Neuron*,
918 76(6):1210–1224.
- 919 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24:103–109.
- 920 Kahana, M. J. (2012). *Foundations of Human Memory*. Oxford University Press, New York, NY.
- 921 Koriat, A. and Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: distin-
922 guishing accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of*
923 *Experimental Psychology: General*, 123(3):297–315.
- 924 Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – con-
925 necting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:1 – 28.
- 926 Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic
927 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
928 104:211–240.
- 929 Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy
930 of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.
- 931 Manning, J. R. (2019). Episodic memory: mental time travel or a quantum ‘memory wave’ function?
932 *PsyArXiv*, doi:10.31234/osf.io/6zjwb.

- 933 Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors, *Handbook*
934 of Human Memory. Oxford University Press.
- 935 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory.
936 In Gazzaniga, M., editor, *The Cognitive Neurosciences, Fifth edition*, pages 557–566. MIT Press.
- 937 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns
938 in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National*
939 *Academy of Sciences, USA*, 108(31):12893–12897.
- 940 McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and
941 projection for dimension reduction. *arXiv*, 1802(03426).
- 942 Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations
943 in vector space. *arXiv*, 1301.3781.
- 944 Mueller, A., Fillion-Robin, J.-C., Boidol, R., Tian, F., Nechifor, P., yoonsubKim, Peter, Rampin, R.,
945 Corvellec, M., Medina, J., Dai, Y., Petrushev, B., Langner, K. M., Hong, Alessio, Ozsvald, I.,
946 vkolmakov, Jones, T., Bailey, E., Rho, V., IgorAPM, Roy, D., May, C., foobuzz, Piyush, Seong,
947 L. K., Goey, J. V., Smith, J. S., Gus, and Mai, F. (2018). WordCloud 1.5.0: a little word cloud
948 generator in Python. *Zenodo*, <https://zenodo.org/record/1322068#.W4tPKZNKh24>.
- 949 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*,
950 64:482–488.
- 951 Paller, K. A. and Wagner, A. D. (2002). Observing the transformation of experience into memory.
952 *Trends in Cognitive Sciences*, 6(2):93–102.
- 953 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten-
954 hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot,
955 M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine*
956 *Learning Research*, 12:2825–2830.

- 957 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model
958 of organizational processes in free recall. *Psychological Review*, 116(1):129–156.
- 959 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal*
960 *of Experimental Psychology*, 17:132–138.
- 961 Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech
962 recognition. *Proceedings of the IEEE*, 77(2):257–286.
- 963 Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are
964 unsupervised multitask learners. *OpenAI Blog*, 1(8).
- 965 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition. *Curr Opin*
966 *Behav Sci*, 17:133–140.
- 967 Ramdas, A., Trillos, N., and Cuturi, M. (2017). On wasserstein two-sample testing and related
968 families of nonparametric tests. *Entropy*, 19(2):47.
- 969 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature*
970 *Reviews Neuroscience*, 13:713 – 726.
- 971 Rishel, C. A., Huang, G., and Freedman, D. J. (2013). Independent category and spatial encoding
972 in parietal cortex. *Neuron*, 77(5):969–979.
- 973 Sigman, M. and Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during
974 dual-task performance. *Journal of Neuroscience*, 28(30):7585–7589.
- 975 Simony, E. and Chang, C. (2020). Analysis of stimulus-induced brain dynamics during naturalistic
976 paradigms. *NeuroImage*, 216:116461.
- 977 Simony, E., Honey, C. J., Chen, J., and Hasson, U. (2016). Uncovering stimulus-locked network
978 dynamics during narrative comprehension. *Nature Communications*, 7(12141):1–13.
- 979 Tompry, A. and Davachi, L. (2017). Consolidation promotes the emergence of representational
980 overlap in the hippocampus and medial prefrontal cortex. *Neuron*, 96(1):228–241.

- 981 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American Journal*
982 *of Psychology*, 35:396–401.
- 983 Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale
984 automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665.
- 985 Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research.
986 *Journal of Memory and Language*, 46:441–517.
- 987 Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception:
988 a mind-brain perspective. *Psychological Bulletin*, 133:273–293.
- 989 Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit
990 memories to other brains: Constructing shared neural representations via communication. *Cereb*
991 *Cortex*, 27(10):4988–5000.
- 992 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and
993 memory. *Psychological Bulletin*, 123(2):162 – 185.

994 Supporting information

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

996 Acknowledgements

997 We thank Luke Chang, Janice Chen, Chris Honey, Lucy Owen, Emily Whitaker, and Kirsten Ziman
998 for feedback and scientific discussions. We also thank Janice Chen, Yuan Chang Leong, Kenneth
999 Norman, and Uri Hasson for sharing the data used in our study. Our work was supported in part
1000 by NSF EPSCoR Award Number 1632738. The content is solely the responsibility of the authors
1001 and does not necessarily represent the official views of our supporting organizations.

1002 **Author contributions**

1003 Conceptualization: A.C.H. and J.R.M.; Methodology: A.C.H., P.C.F. and J.R.M.; Software: A.C.H.,
1004 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.,
1005 P.C.F. and J.R.M.; Supervision: J.R.M.

1006 **Author information**

1007 The authors declare no competing financial interests. Correspondence and requests for materials
1008 should be addressed to J.R.M. (jeremy.r.manning@dartmouth.edu).