

<sup>1</sup> Geometric models reveal behavioral and neural  
<sup>2</sup> signatures of transforming naturalistic experiences into  
<sup>3</sup> episodic memories

<sup>4</sup> Andrew C. Heusser<sup>1, 2, †</sup>, Paxton C. Fitzpatrick<sup>1, †</sup>, and Jeremy R. Manning<sup>1, \*</sup>

<sup>1</sup>Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

<sup>2</sup>Akili Interactive

Boston, MA 02110

<sup>†</sup>Denotes equal contribution

<sup>\*</sup>Corresponding author: Jeremy.R.Manning@Dartmouth.edu

<sup>5</sup> September 3, 2020

## Abstract

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

## Introduction

What does it mean to *remember* something? In traditional episodic memory experiments (e.g., list-learning or trial-based experiments; Kahana, 1996; Murdock, 1962), remembering is often cast as a discrete binary operation: each studied item may be separated from the rest of one's experience and labeled as having been either recalled or forgotten. More nuanced studies might incorporate self-reported confidence measures as a proxy for memory strength, or ask participants to discriminate between recollecting the (contextual) details of an experience and having a general feeling of familiarity (Yonelinas, 2002). Using well-controlled, trial-based experimental designs, the field has amassed a wealth of information regarding human episodic memory (for review see Kahana, 2012). However, there are fundamental properties of the external world and our memories that trial-based experiments are not well suited to capture (for review, also see Huk et al., 2018; Koriat and Goldsmith, 1994). 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

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

40 How might we formally characterize the *essence* of an experience, and whether it has been  
41 recovered by the rememberer? And how might we distinguish an experience's overarching essence  
42 from its low-level details? One approach is to start by considering some fundamental properties  
43 of the dynamics of our experiences. Each given moment of an experience tends to derive meaning  
44 from surrounding moments, as well as from longer-range temporal associations (Lerner et al., 2011;  
45 Manning, 2019, 2020). Therefore, the timecourse describing how an event unfolds is fundamental  
46 to its overall meaning. Further, this hierarchy formed by our subjective experiences at different  
47 timescales defines a *context* for each new moment (e.g., Howard and Kahana, 2002; Howard  
48 et al., 2014), and plays an important role in how we interpret that moment and remember it  
49 later (for review see Manning, 2020; Manning et al., 2015). Our memory systems can leverage  
50 these associations to form predictions that help guide our behaviors (Ranganath and Ritchey,  
51 2012). For example, as we navigate the world, the features of our subjective experiences tend  
52 to change gradually (e.g., the room or situation we find ourselves in at any given moment is  
53 strongly temporally autocorrelated), allowing us to form stable estimates of our current situation  
54 and behave accordingly (Zacks et al., 2007; Zwaan and Radvansky, 1998).

55 Occasionally, this gradual drift of our ongoing experience is punctuated by sudden changes,  
56 or shifts (e.g., when we walk through a doorway; Radvansky and Zacks, 2017). Prior research  
57 suggests that these sharp transitions (termed *event boundaries*) help to discretize our experiences  
58 (and their mental representations) into *events* (Brunec et al., 2018; Clewett and Davachi, 2017;  
59 DuBrow and Davachi, 2013; Ezzyat and Davachi, 2011; Heusser et al., 2018a; Radvansky and  
60 Zacks, 2017). The interplay between the stable (within-event) and transient (across-event) temporal

61 dynamics of an experience also provides a potential framework for transforming experiences  
62 into memories that distills those experiences down to their essences. For example, prior work  
63 has shown that event boundaries can influence how we learn sequences of items (DuBrow and  
64 Davachi, 2013; Heusser et al., 2018a), navigate (Brunec et al., 2018), and remember and understand  
65 narratives (Ezzyat and Davachi, 2011; Zwaan and Radvansky, 1998). This work also suggests  
66 a means of distinguishing the essence of an experience from its low-level details. The overall  
67 structure of events and event transitions reflects how the high-level experience unfolds (i.e., its  
68 essence), while subtler event-level properties reflect low-level details. Prior research has also  
69 implicated a network of brain regions (including the hippocampus and the medial prefrontal  
70 cortex) in playing a critical role in transforming experiences into structured and consolidated  
71 memories (Tompry and Davachi, 2017).

72 Here, we sought to examine how the temporal dynamics of a naturalistic experience were later  
73 reflected in participants' memories. We also sought to leverage the above conceptual insights into  
74 the distinctions between an experience's essence and its low-level details to build models that  
75 explicitly quantified these distinctions. We analyzed an open dataset that comprised behavioral  
76 and functional Magnetic Resonance Imaging (fMRI) data collected as participants viewed and then  
77 verbally recounted an episode of the BBC television show *Sherlock* (Chen et al., 2017). We developed  
78 a computational framework for characterizing the temporal dynamics of the moment-by-moment  
79 content of the episode, and of participants' verbal recalls. Our framework uses topic modeling (Blei  
80 et al., 2003) to characterize the thematic conceptual (semantic) content present in each moment of the  
81 episode and recalls by projecting each moment into a word embedding space. We then use hidden  
82 Markov models (Baldassano et al., 2017; Rabiner, 1989) to discretize this evolving semantic content  
83 into events. In this way, we cast both naturalistic experiences and memories of those experiences  
84 as geometric *trajectories* through word embedding space that describe how they evolve over time.  
85 Under this framework, successful remembering entails verbally traversing the content trajectory  
86 of the episode, thereby reproducing the shape (essence) of the original experience. Our framework  
87 captures the episode's essence in the sequence of geometric coordinates for its events, and its  
88 low-level details by examining its within-event geometric properties.

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

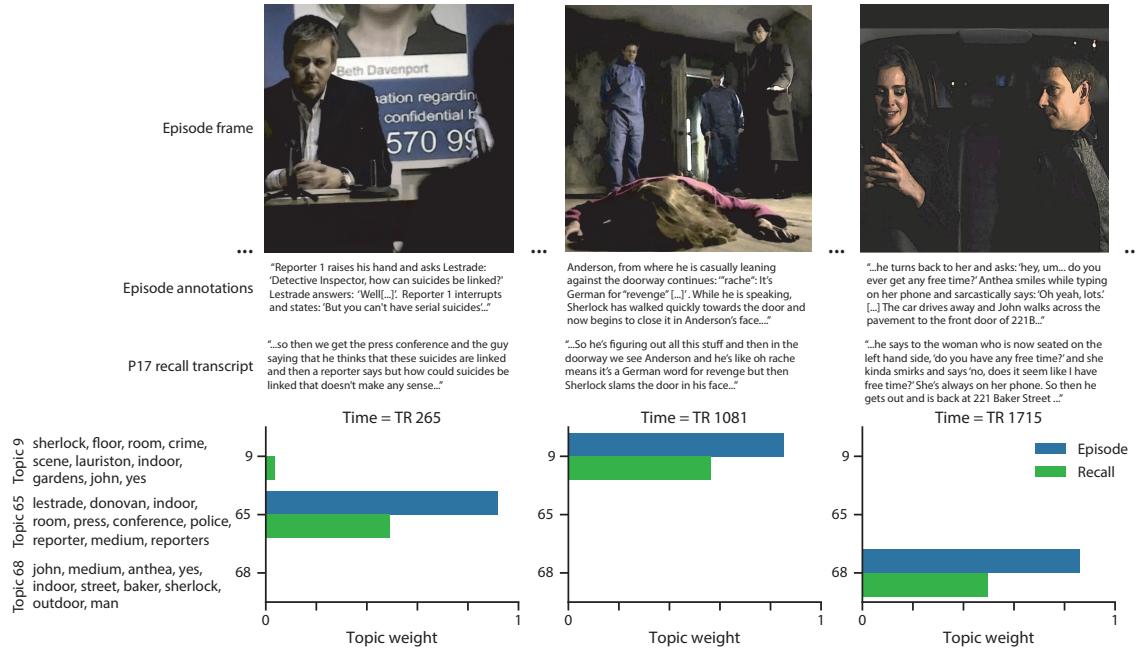
## 100 Results

101 To characterize the dynamic content of the *Sherlock* episode and participants' subsequent recounts  
102ings, we used a topic model (Blei et al., 2003) to discover the episode's latent themes. Topic models  
103 take as inputs a vocabulary of words to consider and a collection of text documents, and return  
104 two output matrices. The first of these is a *topics matrix* whose rows are *topics* (or latent themes)  
105 and whose columns correspond to words in the vocabulary. The entries in the topics matrix  
106 reflect how each word in the vocabulary is weighted by each discovered topic. For example, a  
107 detective-themed topic might weight heavily on words like "crime," and "search." The second  
108 output is a *topic proportions matrix*, with one row per document and one column per topic. The  
109 topic proportions matrix describes the mixture of discovered topics reflected in each document.

110 Chen et al. (2017) collected hand-annotated information about each of 1,000 (manually identi-  
111 fied) scenes spanning the roughly 50 minute video used in their study. This information included:  
112 a brief narrative description of what was happening, the location where the scene took place, the  
113 names of any characters on the screen, and other similar details (for a full list of annotated fea-  
114 tures, see *Methods*). We took from these annotations the union of all unique words (excluding stop

words, such as “and,” “or,” “but,” etc.) across all features and scenes as the vocabulary for the topic model. We then concatenated the sets of words across all features contained in overlapping sliding windows of (up to) 50 scenes, and treated each window as a single document for the purpose of fitting the topic model. Next, we fit a topic model with (up to)  $K = 100$  topics to this collection of documents. We found that 32 unique topics (with non-zero weights) were sufficient to describe the time-varying content of the episode (see *Methods*; Figs. 1, S2). We note that our approach is similar in some respects to Dynamic Topic Models (Blei and Lafferty, 2006), in that we sought to characterize how the thematic content of the episode evolved over time. However, whereas Dynamic Topic Models are designed to characterize how the properties of *collections* of documents change over time, our sliding window approach allows us to examine the topic dynamics within a single document (or video). Specifically, our approach yielded (via the topic proportions matrix) a single *topic vector* for each sliding window of annotations transformed by the topic model. We then stretched (interpolated) the resulting windows-by-topics matrix to match the time series of the 1,976 fMRI volumes collected as participants viewed the episode.

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



**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.

143 timepoint correlation matrix (Fig. 2B) and reflect the gradual drift and sudden shifts fundamental  
144 to the temporal dynamics of many real-world experiences, as well as television episodes. Given  
145 this observation, we adapted an approach devised by Baldassano et al. (2017), and used a hidden  
146 Markov model (HMM) to identify the *event boundaries* where the topic activations changed rapidly  
147 (i.e., the boundaries of the blocks in the temporal correlation matrix; event boundaries identified  
148 by the HMM are outlined in yellow in Fig. 2B). Part of our model fitting procedure required  
149 selecting an appropriate number of events into which the topic trajectory should be segmented.  
150 To accomplish this, we used an optimization procedure that maximized the difference between the  
151 topic weights for timepoints within an event versus timepoints across multiple events (see *Methods*  
152 for additional details). We then created a stable summary of the content within each episode event  
153 by averaging the topic vectors across the timepoints spanned by each event (Fig. 2C).

154 Given that the time-varying content of the episode could be segmented cleanly into discrete  
155 events, we wondered whether participants' recalls of the episode also displayed a similar structure.  
156 We applied the same topic model (already trained on the episode annotations) to each participant's  
157 recalls. Analogously to how we parsed the time-varying content of the episode, to obtain similar  
158 estimates for each participant's recall transcript, we treated each overlapping window of (up to)  
159 10 sentences from their transcript as a document, and computed the most probable mix of topics  
160 reflected in each timepoint's sentences. This yielded, for each participant, a number-of-windows  
161 by number-of-topics topic proportions matrix that characterized how the topics identified in the  
162 original episode were reflected in the participant's recalls. An important feature of our approach  
163 is that it allows us to compare participants' recalls to events from the original episode, despite  
164 that different participants used widely varying language to describe the events, and that those  
165 descriptions often diverged in content and quality from the episode annotations. This ability  
166 to match up conceptually related text that differs in specific vocabulary, detail, and length is an  
167 important benefit of projecting the episode and recalls into a shared topic space. An example topic  
168 proportions matrix from one participant's recalls is shown in Figure 2D.

169 Although the example participant's recall topic proportions matrix has some visual similarity  
170 to the episode topic proportions matrix, the time-varying topic proportions for the example par-



**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).

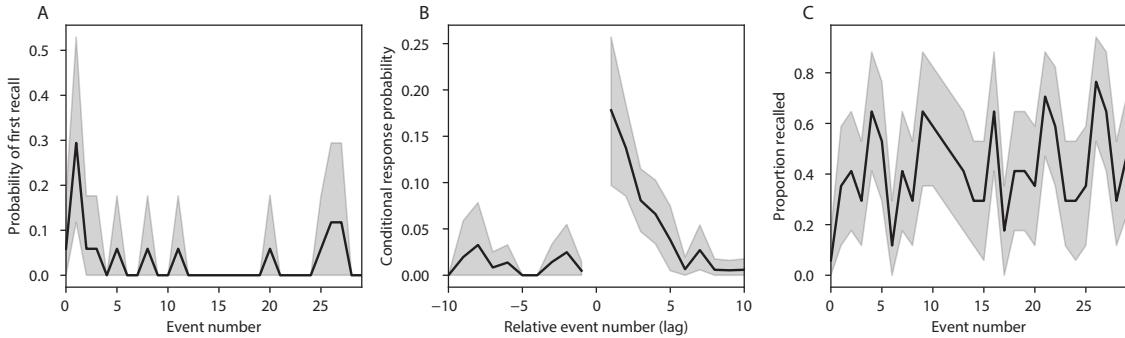
ticipant's recalls are not as sparse as those for the episode (compare Figs. 2A and D). Similarly, although there do appear to be periods of stability in the recall topic dynamics (i.e., most topics are active or inactive over contiguous blocks of time), the changes in topic activations that define event boundaries appear less clearly delineated in participants' recalls than in the episode's annotations. To examine these patterns in detail, we computed the timepoint-by-timepoint correlation matrix for the example participant's recall topic proportions matrix (Fig. 2E). As in the episode correlation matrix (Fig. 2B), the example participant's recall correlation matrix has a strong block diagonal structure, indicating that their recalls are discretized into separated events. We used the same HMM-based optimization procedure that we had applied to the episode's topic proportions matrix (see *Methods*) to estimate an analogous set of event boundaries in the participant's recounting of the episode (outlined in yellow). We carried out this analysis on all 17 participants' recall topic proportions matrices (Fig. S4).

Two clear patterns emerged from this set of analyses. First, although every individual participant's recalls could be segmented into discrete events (i.e., every individual participant's recall correlation matrix exhibited clear block diagonal structure; Fig. S4), each participant appeared to have a unique *recall resolution*, reflected in the sizes of those blocks. While some participants' recall topic proportions segmented into just a few events (e.g., Participants P4, P5, and P7), others' segmented into many shorter duration events (e.g., Participants P12, P13, and P17). This suggests that different participants may be recalling the episode with different levels of detail—i.e., some might recount only high-level essential plot details, whereas others might recount low-level details instead (or in addition). The second clear pattern present in every individual participant's recall correlation matrix was that, unlike in the episode correlation matrix, there were substantial off-diagonal correlations. Whereas each event in the original episode was (largely) separable from the others (Fig. 2B), in transforming those separable events into memory, participants appeared to be integrating across multiple events, blending elements of previously recalled and not-yet-recalled content into each newly recalled event (Figs. 2E, S4; also see Howard et al., 2012; Manning, 2019; Manning et al., 2011).

The above results demonstrate that topic models capture the dynamic conceptual content of

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

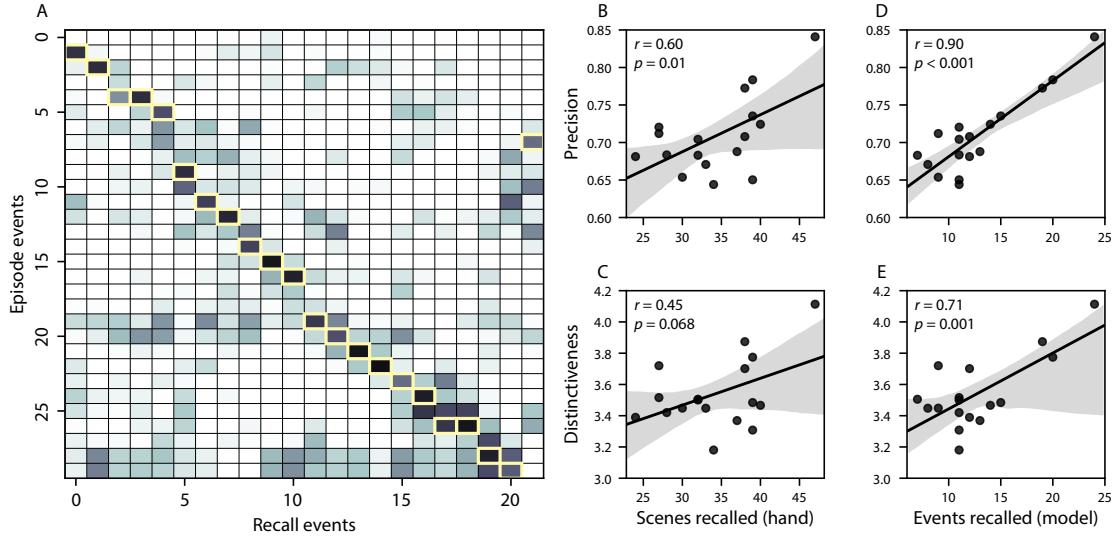
219 Clustering scores are often used by memory researchers to characterize how people organize  
220 their memories of words on a studied list (for review, see Polyn et al., 2009). We defined analogous  
221 measures to characterize how participants organized their memories for episodic events (see  
222 *Methods* for details). Temporal clustering refers to the extent to which participants group their recall  
223 responses according to encoding position. Overall, we found that sequentially viewed episode  
224 events tended to appear nearby in participants' recall event sequences (mean clustering score: 0.767,  
225 SEM: 0.029). Participants with higher temporal clustering scores tended to exhibit better overall  
226 memory for the episode, according to both Chen et al. (2017)'s hand-counted numbers of recalled



**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.

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



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

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

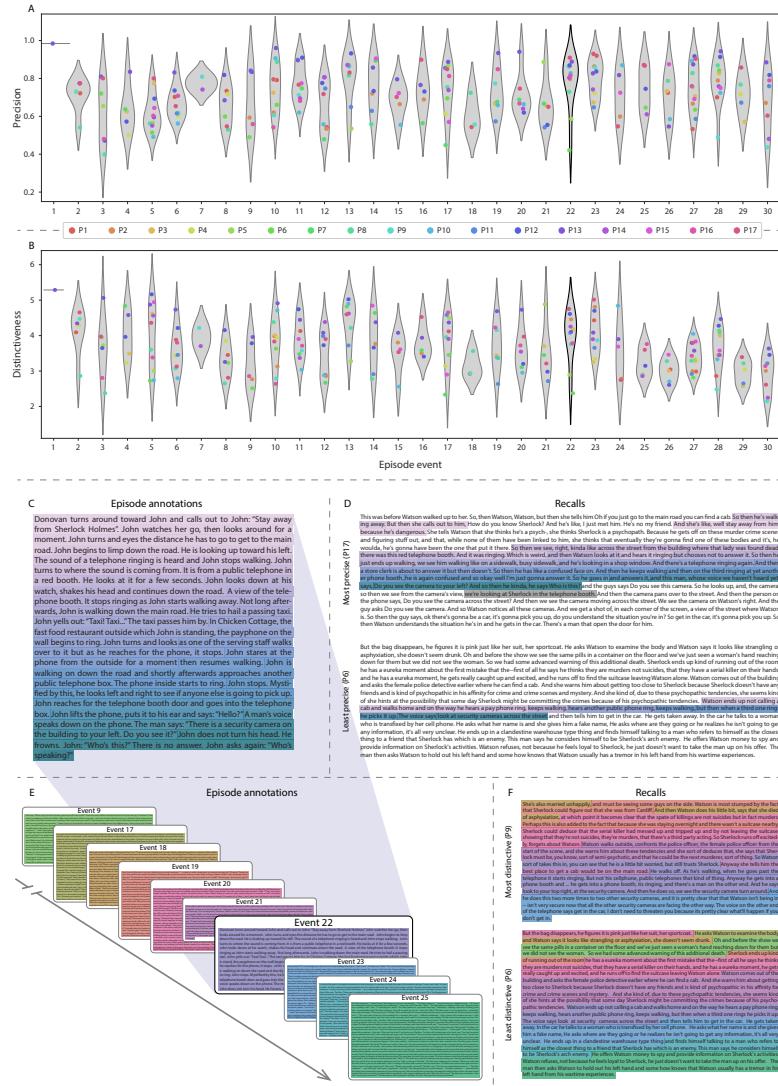
251     *Distinctiveness* is intended to capture the “specificity” of recall. In other words, distinctiveness  
252     quantifies the extent to which a given recalled event reflects the most similar episode event over  
253     and above its reflection of other episode events. Intuitively, distinctiveness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

266     In addition to examining how precisely and distinctively participants recalled individual events,  
267     one may also use these metrics to summarize each participant’s performance by averaging across  
268     a participant’s event-wise precision or distinctiveness scores. This enables us to quantify how pre-  
269     cisely a participant tended to recall subtle within-event details, as well as how specific (distinctive)  
270     those details were to individual events from the episode. Participants’ average precision and dis-  
271     tinctiveness scores were strongly correlated ( $r(15) = 0.90, p < 10^{-5}$ ). This indicates that participants

272 who tended to precisely recount low-level details of episode events also tended to do so in an  
273 event-specific way (e.g., as opposed to detailing recurring themes that were present in most or all  
274 episode events; this behavior would have resulted in high precision but low distinctiveness). We  
275 found that, across participants, higher precision scores were positively correlated with both the  
276 hand-annotated ( $r(15) = 0.60, p = 0.010$ ) and model-estimated ( $r(15) = 0.90, p < 0.001$ ) numbers of  
277 events that participants recalled. Participants' average distinctiveness scores were also correlated  
278 with both the hand-annotated ( $r(15) = 0.45, p = 0.068$ ) and model-estimated ( $r(15) = 0.71, p = 0.001$ )  
279 numbers of recalled events.

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

294 Figure 5 also illustrates the differences between high and low distinctiveness scores for the  
295 same event detailed in Figure 5B (i.e., event 22). Here, we have extracted the set of sentences  
296 comprising the most distinctive recall event (P9) and least distinctive recall event (P6) matched to  
297 the example episode event (Fig. 5F). We also extracted the annotations for the example episode  
298 event, as well as those from each other episode event whose content the example participants'  
299 single recall events described (Fig. 5E). We assigned each episode event a unique color (Panel E)

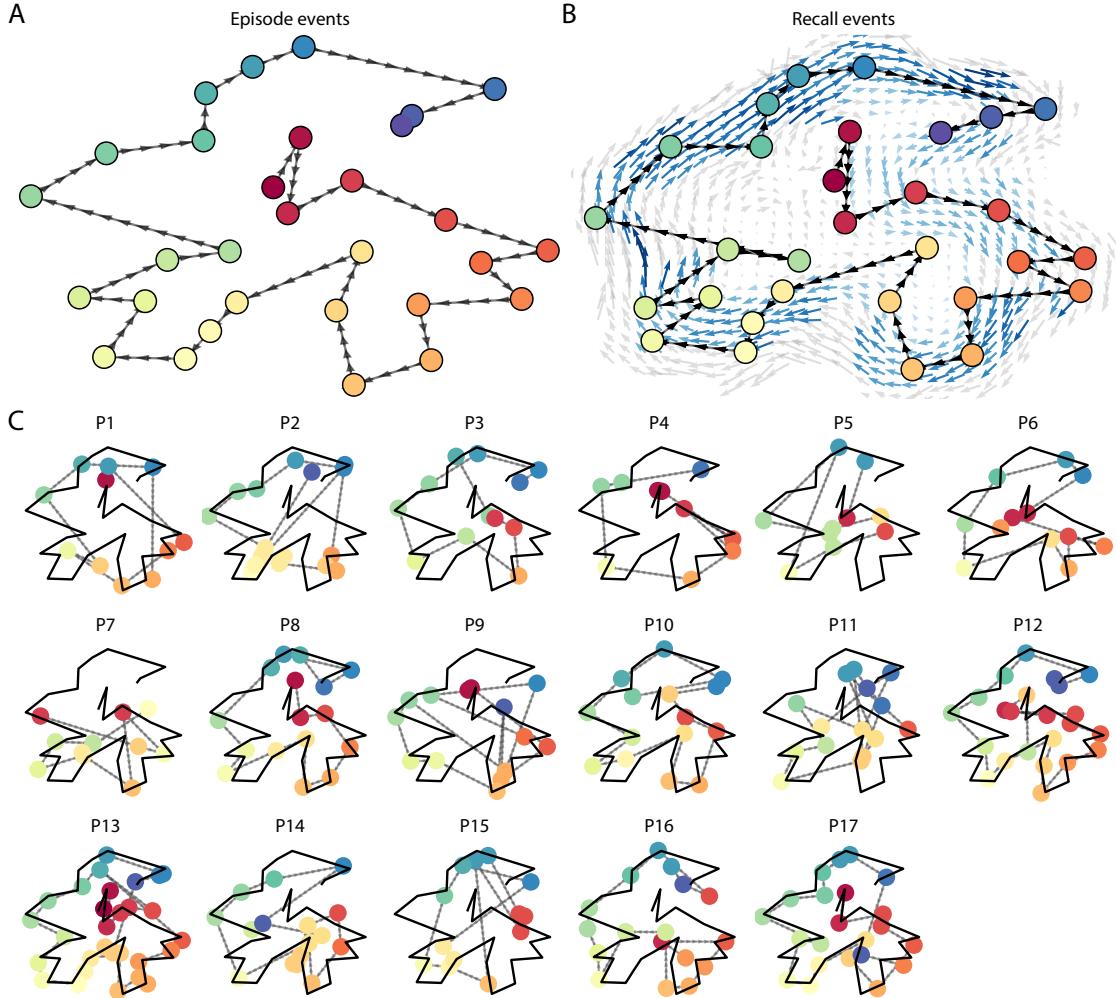


**Figure 5: Precision reflects the completeness of recall, whereas distinctiveness reflects recall specificity.** A. Recall precision by episode event. Grey violin plots display kernel density estimates for the distribution of recall precision scores for a single episode event. Colored dots within each violin plot represent individual participants' recall precision for the given event. Episode events are ordered along the x-axis by the average precision with which they were remembered. B. The set of "Narrative Details" episode annotations (generated by 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. D. Recall distinctiveness by episode event. Kernel density estimates for each episode event's distribution of recall distinctiveness scores, analogous to Panel A. E. 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 F. Each event's text is highlighted in a different color. F. The sentences comprising the most distinctive (P9) and least distinctive (P6) participants' recalls of episode event 22. Sections of recall describing each episode event in Panel E are highlighted with the corresponding color.

300 and colored each recalled phrase or sentence (Panel F) according to the episode events they best  
301 matched. Visual inspection of Panel F reveals that the most distinctive recall's content is tightly  
302 concentrated around event 22, whereas the least distinctive recall incorporates content from a much  
303 wider range of episode events.

304 The preceding analyses sought to characterize how participants' recounts of individual  
305 episode events captured the low-level details of each event. Next we sought to characterize how  
306 participants' recounts of the full episode captured its high-level essence— i.e., the shape of the  
307 episode's trajectory through word embedding (topic) space. To visualize the essence of the episode  
308 and each participant's recall trajectory (Heusser et al., 2018b), we projected the topic proportions  
309 matrices for the episode and recalls onto a shared two-dimensional space using Uniform Manifold  
310 Approximation and Projection (UMAP; McInnes et al., 2018). In this lower-dimensional space,  
311 each point represents a single episode or recall event, and the distances between the points reflect  
312 the distances between the events' associated topic vectors (Fig. 6). In other words, events that are  
313 nearer to each other in this space are more semantically similar, and those that are farther apart are  
314 less so.

315 Visual inspection of the episode and recall topic trajectories reveals a striking pattern. First, the  
316 topic trajectory of the episode (which reflects its dynamic content; Fig. 6A) is captured nearly per-  
317 fectly by the averaged topic trajectories of participants' recalls (Fig. 6B). To assess the consistency  
318 of these recall trajectories across participants, we asked: given that a participant's recall trajectory  
319 had entered a particular location in the reduced topic space, could the position of their *next* recalled  
320 event be predicted reliably? For each location in the reduced topic space, we computed the set of  
321 line segments connecting successively recalled events (across all participants) that intersected that  
322 location (see *Methods* for additional details). We then computed (for each location) the distribu-  
323 tion of angles formed by the lines defined by those line segments and a fixed reference line (the  
324 *x*-axis). Rayleigh tests revealed the set of locations in topic space at which these across-participant  
325 distributions exhibited reliable peaks (blue arrows in Fig. 6B reflect significant peaks at  $p < 0.05$ ,  
326 corrected). We observed that the locations traversed by nearly the entire episode trajectory exhib-  
327 ited such peaks. In other words, participants' recalls exhibited similar trajectories to each other



**Figure 6: Trajectories through topic space capture the dynamic content of the episode and recalls.** All panels: the topic proportion matrices have been projected onto a shared two-dimensional space using UMAP. **A.** The two-dimensional topic trajectory taken by the episode of *Sherlock*. Each dot indicates an event identified using the HMM (see *Methods*); the dot colors denote the order of the events (early events are in red; later events are in blue), and the connecting lines indicate the transitions between successive events. **B.** The average two-dimensional trajectory captured by participants' recall sequences, with the same format and coloring as the trajectory in Panel A. To compute the event positions, we matched each recalled event with an event from the original episode (see *Results*), and then we averaged the positions of all events with the same label. The arrows reflect the average transition direction through topic space taken by any participants whose trajectories crossed that part of topic space; blue denotes reliable agreement across participants via a Rayleigh test ( $p < 0.05$ , corrected). **C.** The recall topic trajectories (gray) taken by each individual participant (P1–P17). The episode's trajectory is shown in black for reference. Here, events (dots) are colored by their matched episode event (Panel A).

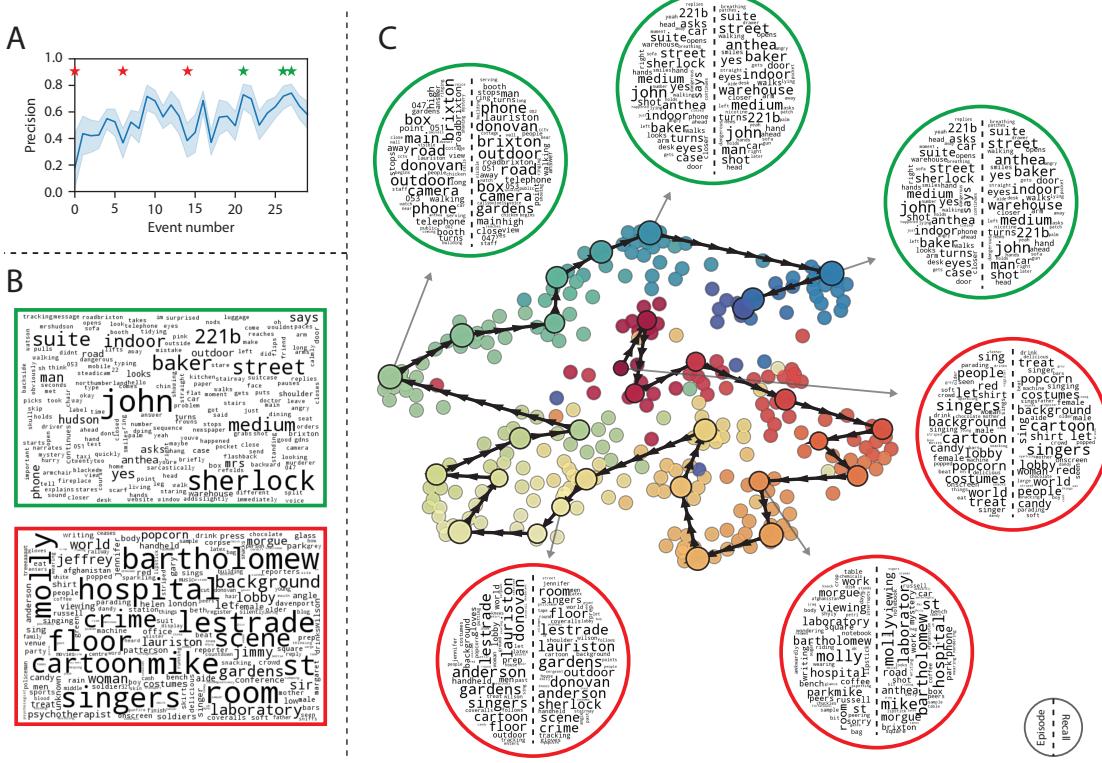
328 that also matched the trajectory of the original episode (Fig. 6C). This is especially notable when  
329 considering the fact that the numbers of events participants recalled (dots in Fig. 6C) varied con-  
330 siderably across people, and that every participant used different words to describe what they had  
331 remembered happening in the episode. Differences in the numbers of remembered events appear  
332 in participants' trajectories as differences in the sampling resolution along the trajectory. We note  
333 that this framework also provides a means of disentangling classic "proportion recalled" measures  
334 (i.e., the proportion of episode events described in participants' recalls) from participants' abilities  
335 to recapitulate the episode's essence (i.e., the similarity between the shapes of the original episode  
336 trajectory and that defined by each participant's recounting of the episode).

337 In addition to enabling us to visualize the episode's high-level essence, describing the episode  
338 as a geometric trajectory also enables us to drill down to individual words and quantify how each  
339 word relates to the memorability of each event. This provides another approach to examining  
340 participants' recall for low-level details beyond the precision and distinctiveness measures we  
341 defined above. The results displayed in Figures 3C and 5A suggest that certain events were  
342 remembered better than others. Given this, we next asked whether the events were generally  
343 remembered precisely or imprecisely tended to reflect particular content. Because our analysis  
344 framework projects the dynamic episode content and participants' recalls into a shared space, and  
345 because the dimensions of that space represent topics (which are, in turn, sets of weights over  
346 known words in the vocabulary), we are able to recover the weighted combination of words that  
347 make up any point (i.e., topic vector) in this space. We first computed the average precision with  
348 which participants recalled each of the 30 episode events (Fig. 7A; note that this result is analogous  
349 to a serial position curve created from our precision metric). We then computed a weighted average  
350 of the topic vectors for each episode event, where the weights reflected how precisely each event  
351 was recalled. To visualize the result, we created a "wordle" image (Mueller et al., 2018) where  
352 words weighted more heavily by more precisely-remembered topics appear in a larger font (Fig. 7B,  
353 green box). Across the full episode, content that weighted heavily on topics and words central to  
354 the major foci of the episode (e.g., the names of the two main characters, "Sherlock" and "John,"  
355 and the address of a major recurring location, "221B Baker Street") were best remembered. An

356 analogous analysis revealed which themes were less-precisely remembered. Here in computing  
357 the weighted average over events' topic vectors, we weighted each event in *inverse* proportion to  
358 its average precision (Fig. 7B, red box). The least precisely remembered episode content reflected  
359 information that was extraneous to the episode's essence, such as the proper names of relatively  
360 minor characters (e.g., "Mike," "Molly," and "Lestrade") and locations (e.g., "St. Bartholomew's  
361 Hospital").

362 A similar result emerged from assessing the topic vectors for individual episode and recall  
363 events (Fig. 7C). Here, for each of the three most and least precisely remembered episode events, we  
364 have constructed two wordles: one from the original episode event's topic vector (left) and a second  
365 from the average recall topic vector for that event (right). The three most precisely remembered  
366 events (circled in green) correspond to scenes integral to the central plot-line: a mysterious figure  
367 spying on John in a phone booth; John meeting Sherlock at Baker St. to discuss the murders; and  
368 Sherlock laying a trap to catch the killer. Meanwhile, the least precisely remembered events (circled  
369 in red) reflect scenes that comprise minor plot points: a video of singing cartoon characters that  
370 participants viewed in an introductory clip prior to the main episode; John asking Molly about  
371 Sherlock's habit of over-analyzing people; and Sherlock noticing evidence of Anderson's and  
372 Donovan's affair.

373 The results thus far inform us about which aspects of the dynamic content in the episode partic-  
374 ipants watched were preserved or altered in participants' memories. We next carried out a series of  
375 analyses aimed at understanding which brain structures might facilitate these preservations and  
376 transformations between the participants' shared experience of watching the episode and their  
377 subsequent memories of the episode. In the first analysis, we sought to identify brain structures  
378 that were sensitive to the dynamic unfolding of the episode's content, as characterized by its topic  
379 trajectory. We used a searchlight procedure to identify clusters of voxels whose activity patterns  
380 displayed a proximal temporal correlation structure (as participants watched the episode) match-  
381 ing that of the original episode's topic proportions (Fig. 8A; see *Methods* for additional details). In a  
382 second analysis, we sought to identify brain structures whose responses (during episode viewing)  
383 reflected how each participant would later structure their *recounting* of the episode. We used a



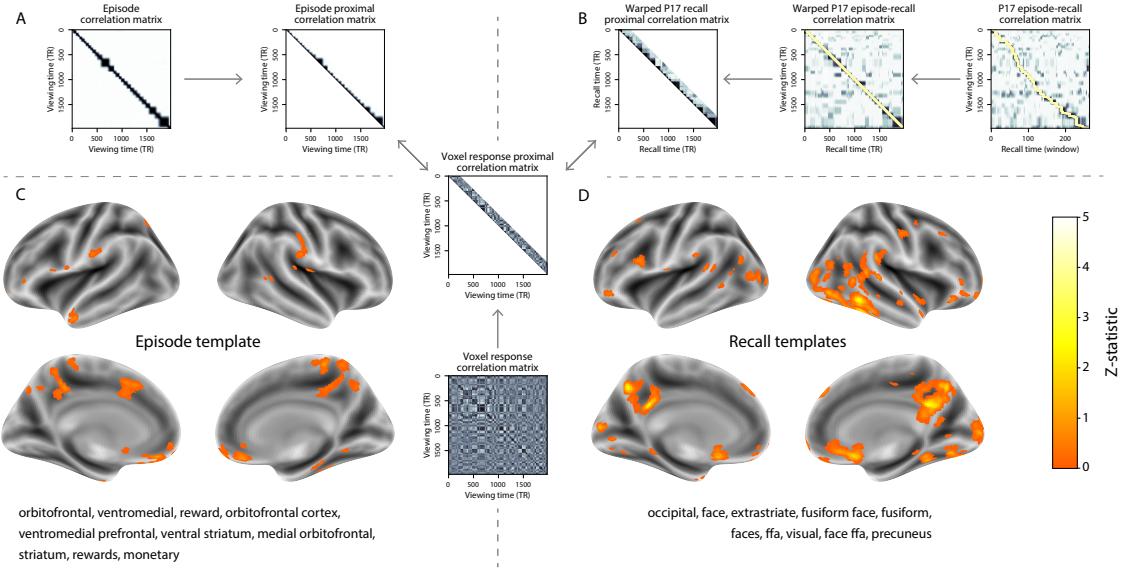
**Figure 7: Language used in the most and least precisely remembered events.** **A.** Average precision (episode event-recall event topic vector correlation) across participants for each episode event. Here we defined each episode event’s precision for each participant as the correlation between its topic vector and the most-correlated recall event’s topic vector from that participant. Error bars denote bootstrap-derived across-participant 95% confidence intervals. The stars denote the three most precisely remembered events (green) and least precisely remembered events (red). **B.** Wordles comprising the top 200 highest-weighted words reflected in the weighted-average topic vector across episode events. Green: episode events were weighted by their precision (Panel A). Red: episode events were weighted by the inverse of their precision. **C.** The set of all episode and recall events is projected onto the two-dimensional space derived in Figure 6. The dots outlined in black denote episode events (dot size is proportional to each event’s average precision). The dots without black outlines denote individual recall events from each participant. All dots are colored using the same scheme as Figure 6A. 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.

384 searchlight procedure to identify clusters of voxels whose proximal temporal correlation matrices  
385 matched that of the topic proportions matrix for each participant's recall transcript (Figs. 8B; see  
386 *Methods* for additional details). To ensure our searchlight procedure identified regions *specifically*  
387 sensitive to the temporal structure of the episode or recalls (i.e., rather than those with a tem-  
388 poral autocorrelation length similar to that of the episode and recalls), we performed a phase  
389 shift-based permutation correction (see *Methods* for additional details). As shown in Figure 8C,  
390 the episode-driven searchlight analysis revealed a distributed network of regions that may play  
391 a role in processing information relevant to the narrative structure of the episode. Similarly, the  
392 recall-driven searchlight analysis revealed a second network of regions (Fig. 8D) that may facil-  
393 itate a person-specific transformation of one's experience into memory. The top ten Neurosynth  
394 terms (Yarkoni et al., 2011) associated with each (unthresholded) map are displayed in each panel.  
395 In identifying regions whose responses to ongoing experiences reflect how those experiences will  
396 be remembered later, this latter analysis extends classic *subsequent memory effect analyses* (e.g., Paller  
397 and Wagner, 2002) to the domain of naturalistic experiences.

398 The searchlight analyses described above yielded two distributed networks of brain regions,  
399 whose activity timecourses tracked with the temporal structure of the episode (Fig. 8C) or par-  
400 ticipants' subsequent recalls (Fig. 8D). We next sought to gain greater insight into the structures  
401 and functional networks our results reflected. To accomplish this, we performed an additional,  
402 exploratory analysis using Neurosynth (Yarkoni et al., 2011). Given an arbitrary statistical map as  
403 input, Neurosynth performs a massive automated meta-analysis, returning a ranked list of terms  
404 reported in papers with similar significance maps. We ran Neurosynth on the significance maps  
405 for the episode- and recall-driven searchlight analyses. These maps, along with the 10 terms with  
406 maximally similar meta-analysis images identified by Neurosynth are shown in Figure 8.

## 407 Discussion

408 Explicitly modeling the dynamic content of a naturalistic stimulus and participants' memories  
409 enabled us to connect the present study of naturalistic recall with an extensive prior literature that



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

410 has used list-learning paradigms to study memory (for review see Kahana, 2012), as in Figure 3.  
411 We found some similarities between how participants in the present study recounted a television  
412 episode and how participants typically recall memorized random word lists. However, our broader  
413 claim is that word lists miss out on fundamental aspects of naturalistic memory more like the sort  
414 of memory we rely on in everyday life. For example, there are no random word list analogs of  
415 character interactions, conceptual dependencies between temporally distant episode events, the  
416 sense of solving a mystery that pervades the *Sherlock* episode, or the myriad other features of the  
417 episode that convey deep meaning and capture interest. Nevertheless, each of these properties  
418 affects how people process and engage with the episode as they are watching it, and how they  
419 remember it later. The overarching goal of the present study is to characterize how the rich  
420 dynamics of the episode affect the rich behavioral and neural dynamics of how people remember  
421 it.

422 Our work casts remembering as reproducing (behaviorally and neurally) the topic trajectory,  
423 or “shape,” of an experience. When we characterized memory for a television episode using this  
424 framework, we found that every participant’s recounting of the episode recapitulated the low  
425 spatial frequency details of the shape of its trajectory through topic space (Fig. 6). We termed  
426 this narrative scaffolding the episode’s *essence*. Where participants’ behaviors varied most was  
427 in their tendencies to recount specific low-level details from each episode event. Geometrically,  
428 this appears as high spatial frequency distortions in participants’ recall trajectories relative to the  
429 trajectory of the original episode (Fig. 7). We developed metrics to characterize the precision  
430 (recovery of any and all event-level information) and distinctiveness (recovery of event-specific  
431 information). We also used word cloud visualizations to interpret the details of these event-level  
432 distortions.

433 The neural analyses we carried out (Fig. 8) also leveraged our geometric framework for char-  
434 acterizing the shapes of the episode and participants’ recounts. We identified one network  
435 of regions whose responses tracked with temporal correlations in the conceptual content of the  
436 episode (as quantified by topic models applied to a set of annotations about the episode). This  
437 network included orbitofrontal cortex, ventromedial prefrontal cortex, striatum, among others. As

438 reviewed by Ranganath and Ritchey (2012), several of these regions are members of the *anterior*  
439 *temporal system*, which has been implicated in assessing and processing the familiarity of ongoing  
440 experiences, emotions, social cognition, and reward. A second network we identified tracked with  
441 temporal correlations in the idiosyncratic conceptual content of participants' subsequent recounts  
442 of the episode. This network included occipital cortex, extrastriate cortex, fusiform gyrus, and  
443 the precuneus. Several of these regions are members of the *posterior medial system* (Ranganath and  
444 Ritchey, 2012), which has been implicated in matching incoming cues about the current situation  
445 to internally maintained *situation models* that specify the parameters and expectations inherent to  
446 the current situation (also see Zacks et al., 2007; Zwaan and Radvansky, 1998). Taken together, our  
447 results support the notion that these two (partially overlapping) networks work in coordination  
448 to make sense of our ongoing experiences, distort them in a way that links them with our prior  
449 knowledge and experiences, and encodes those distorted representations into memory for our later  
450 use.

451 Our general approach draws inspiration from prior work aimed at elucidating the neural and  
452 behavioral underpinnings of how we process dynamic naturalistic experiences and remember them  
453 later. Our approach to identifying neural responses to naturalistic stimuli (including experiences)  
454 entails building an explicit model of the stimulus dynamics and searching for brain regions whose  
455 responses are consistent with the model (also see Huth et al., 2016, 2012). In prior work, a  
456 series of studies from Uri Hasson's group (Baldassano et al., 2017; Chen et al., 2017; Lerner et al.,  
457 2011; Simony et al., 2016; Zadbood et al., 2017) have developed a clever alternative approach:  
458 rather than building an explicit stimulus model, these studies instead search for brain responses  
459 (while experiencing the stimulus) that are reliably similar across individuals. So called *inter-*  
460 *subject correlation* (ISC) and *inter-subject functional connectivity* (ISFC) analyses effectively treat other  
461 people's brain responses to the stimulus as a "model" of how its features change over time (also  
462 see Simony and Chang, 2020). These purely brain-driven approaches are well-suited to identifying  
463 which brain structures exhibit similar stimulus-driven responses across individuals. Further,  
464 because neural response dynamics are observed data (rather than model approximations), such  
465 approaches do not require a detailed understanding of which stimulus properties or features might

466 be driving the observed responses. However, this also means that the specific stimulus features  
467 driving those responses are typically opaque to the researcher. Our approach is complementary.  
468 By explicitly modeling the stimulus dynamics, we are able to relate specific stimulus features to  
469 behavioral and neural dynamics. However, when our model fails to accurately capture the stimulus  
470 dynamics that are truly driving behavioral and neural responses, our approach necessarily yields  
471 an incomplete characterization of the neural basis of the processes we are studying.

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

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

494 Fig. 7) over raw performance (e.g., with respect to some predefined benchmark). Nevertheless, we  
495 note that our general framework is, in principle, robust to the specific choice of language model  
496 as well as other aspects of our computational pipeline. For example, the word embedding model,  
497 timeseries segmentation model, and the episode-recall matching function could each be customized  
498 to suit a particular question space or application. Indeed, for some questions, interpretability of  
499 the embeddings may not be a priority, and thus other text embedding approaches (including the  
500 deep learning-based models described above) may be preferable. Further work will be needed to  
501 explore the influence of particular models on our framework’s predictions and performance.

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

516 **Methods**

517 **Paradigm and data collection**

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

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

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

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

## 548 **Data and code availability**

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

## 551 **Statistics**

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

## 557 **Modeling the dynamic content of the episode and recall transcripts**

### 558 **Topic modeling**

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

567 for each scene using the text descriptions of the preceding 25 scenes, the present scene, and the  
568 following 24 scenes. To model the context for scenes near the beginning of the episode (i.e., within  
569 25 scenes of the beginning or end), we created overlapping sliding windows that grew in size  
570 from one scene to the full length. We also tapered the sliding window lengths at the end of the  
571 episode, whereby scenes within fewer than 24 scenes of the end of the episode were assigned  
572 sliding windows that extended to the end of the episode. This procedure ensured that each scene's  
573 content was represented in the text corpus an equal number of times.

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

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

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

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

603 We parsed the topic proportions matrices of the episode and participants' recalls into discrete  
604 events using hidden Markov Models (HMMs; Rabiner, 1989). Given the topic proportions matrix  
605 (describing the mix of topics at each timepoint) and a number of states,  $K$ , an HMM recovers the  
606 set of state transitions that segments the timeseries into  $K$  discrete states. Following Baldassano  
607 et al. (2017), we imposed an additional set of constraints on the discovered state transitions that  
608 ensured that each state was encountered exactly once (i.e., never repeated). We used the BrainIAK  
609 toolbox (Capota et al., 2017) to implement this segmentation.

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

$$617 \quad \underset{K}{\operatorname{argmax}} [W_1(a, b)],$$

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

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

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

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

645 **Probability of first recall (PFR).** PFR curves (Atkinson and Shiffrin, 1968; Postman and Phillips,  
646 1965; Welch and Burnett, 1924) reflect the probability that an item will be recalled first as a function  
647 of its serial position during encoding. To carry out this analysis, we initialized a number-of-  
648 participants (17) by number-of-episode-events (30) matrix of zeros. Then for each participant, we  
649 found the index of the episode event that was recalled first (i.e., the episode event whose topic  
650 vector was most strongly correlated with that of the first recall event) and filled in that index in  
651 the matrix with a 1. Finally, we averaged over the rows of the matrix, resulting in a 1 by 30 array  
652 representing the proportion of participants that recalled an event first, as a function of the order of  
653 the event's appearance in the episode (Fig. 3A).

654 **Lag conditional probability curve (lag-CRP).** The lag-CRP curve (Kahana, 1996) reflects the  
655 probability of recalling a given item after the just-recalled item, as a function of their relative  
656 encoding positions (or *lag*). In other words, a lag of 1 indicates that a recalled item was presented  
657 immediately after the previously recalled item, and a lag of -3 indicates that a recalled item came 3  
658 items before the previously recalled item. For each recall transition (following the first recall), we  
659 computed the lag between the current recall event and the next recall event, normalizing by the  
660 total number of possible transitions. This yielded a number-of-participants (17) by number-of-lags  
661 (-29 to +29; 58 lags total excluding lags of 0) matrix. We averaged over the rows of this matrix to  
662 obtain a group-averaged lag-CRP curve (Fig. 3B).

663 **Serial position curve (SPC).** SPCs (Murdock, 1962) reflect the proportion of participants that  
664 remember each item as a function of the items' serial positions during encoding. We initialized  
665 a number-of-participants (17) by number-of-episode-events (30) matrix of zeros. Then, for each  
666 recalled event, for each participant, we found the index of the episode event that the recalled  
667 event most closely matched (via the correlation between the events' topic vectors) and entered a  
668 1 into that position in the matrix. This resulted in a matrix whose entries indicated whether or  
669 not each event was recalled by each participant (depending on whether the corresponding entires  
670 were set to one or zero). Finally, we averaged over the rows of the matrix to yield a 1 by 30 array

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

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

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

### 691 **Averaging correlations**

692 In all instances where we performed statistical tests involving precision or distinctiveness scores  
693 (Fig. 5, we used the Fisher z-transformation (Fisher, 1925) to stabilize the variance across the  
694 distribution of correlation values prior to performing the test. Similarly, when averaging precision  
695 or distinctiveness scores, we z-transformed the scores prior to computing the mean, and inverse

696  $z$ -transformed the result.

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

698 We used the UMAP algorithm (McInnes et al., 2018) to project the 100-dimensional topic space onto  
699 a two-dimensional space for visualization (Figs. 6, 7). To ensure that all of the trajectories were  
700 projected onto the *same* lower dimensional space, we computed the low-dimensional embedding  
701 on a “stacked” matrix created by vertically concatenating the events-by-topics topic proportions  
702 matrices for the episode, across-participants average recall and all 17 individual participants’ re-  
703 calls. We then separated the rows of the result (a total-number-of-events by two matrix) back into  
704 individual matrices for the episode topic trajectory, across-participant average recall trajectory and  
705 the trajectories for each individual participant’s recalls (Fig. 6). This general approach for dis-  
706 covering a shared low-dimensional embedding for a collections of high-dimensional observations  
707 follows Heusser et al. (2018b).

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

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

718 In Figure 6B, we present an analysis aimed at characterizing locations in topic space that dif-  
719 ferent participants move through in a consistent way (via their recall topic trajectories). The  
720 two-dimensional topic space used in our visualizations (Fig. 6) comprised a  $60 \times 60$  (arbitrary  
721 units) square. We tiled this space with a  $50 \times 50$  grid of evenly spaced vertices, and defined a

722 circular area centered on each vertex whose radius was two times the distance between adjacent  
723 vertices (i.e., 2.4 units). For each vertex, we examined the set of line segments formed by connecting  
724 each pair successively recalled events, across all participants, that passed through this circle. We  
725 computed the distribution of angles formed by those segments and the  $x$ -axis, and used a Rayleigh  
726 test to determine whether the distribution of angles was reliably “peaked” (i.e., consistent across  
727 all transitions that passed through that local portion of topic space). To create Figure 6B we drew  
728 an arrow originating from each grid vertex, pointing in the direction of the average angle formed  
729 by the line segments that passed within 2.4 units. We set the arrow lengths to be inversely propor-  
730 tional to the  $p$ -values of the Rayleigh tests at each vertex. Specifically, for each vertex we converted  
731 all of the angles of segments that passed within 2.4 units to unit vectors, and we set the arrow  
732 lengths at each vertex proportional to the length of the (circular) mean vector. We also indicated  
733 any significant results ( $p < 0.05$ , corrected using the Benjamani-Hochberg procedure) by coloring  
734 the arrows in blue (darker blue denotes a lower  $p$ -value, i.e., a longer mean vector); all tests with  
735  $p \geq 0.05$  are displayed in gray and given a lower opacity value.

### 736 Searchlight fMRI analyses

737 In Figure 8, we present two analyses aimed at identifying brain regions whose responses (as partic-  
738 ipants viewed the episode) exhibited a particular temporal structure. We developed a searchlight  
739 analysis wherein we constructed a  $5 \times 5 \times 5$  cube of voxels (following Chen et al., 2017) centered  
740 on each voxel in the brain, and for each of these cubes, computed the temporal correlation matrix  
741 of the voxel responses during episode viewing. Specifically, for each of the 1976 volumes collected  
742 during episode viewing, we correlated the activity patterns in the given cube with the activity  
743 patterns (in the same cube) collected during every other timepoint. This yielded a  $1976 \times 1976$   
744 correlation matrix for each cube. Note: participant 5’s scan ended 75s early, and in Chen et al.,  
745 2017’s publicly released dataset, their scan data was zero-padded to match the length of the other  
746 participants’. For our searchlight analyses, we removed this padded data (i.e., the last 50 TRs),  
747 resulting in a  $1925 \times 1925$  correlation matrix for each cube in participant 5’s brain.

748 Next, we constructed a series of “template” matrices. The first template reflected the timecourse

749 of the episode’s topic trajectory, and the others reflected the timecourse of each participant’s recall  
750 trajectory. To construct the episode template, we computed the correlations between the topic  
751 proportions estimated for every pair of TRs (prior to segmenting the trajectory into discrete events;  
752 i.e., the correlation matrix shown in Figs. 2B and 8A). We constructed similar temporal correlation  
753 matrices for each participant’s recall topic trajectory (Figs. 2D, S4). However, to correct for length  
754 differences and potential non-linear transformations between viewing time and recall time, we  
755 first used dynamic time warping (Berndt and Clifford, 1994) to temporally align participants’  
756 recall topic trajectories with the episode topic trajectory. An example correlation matrix before and  
757 after warping is shown in Fig. 8B. This yielded a  $1976 \times 1976$  correlation matrix for the episode  
758 template and for each participant’s recall template.

759 The temporal structure of the episode’s content (as described by our model) is captured in the  
760 block-diagonal structure of the episode’s temporal correlation matrix (e.g., Figs. 2B, 8A), with time  
761 periods of thematic stability represented as dark blocks of varying sizes. Inspecting the episode  
762 correlation matrix suggests that the episode’s semantic content is highly temporally specific (i.e., the  
763 correlations between topic vectors from distant timepoints are almost all near zero). By contrast,  
764 the activity patterns of individual (cubes of) voxels can encode relatively limited information  
765 on their own, and their activity frequently contributes to multiple separate functions (Charron  
766 and Koechlin, 2010; Freedman et al., 2001; Rishel et al., 2013; Sigman and Dehaene, 2008). By  
767 nature, these two attributes give rise to similarities in activity across large timescales that may not  
768 necessarily reflect a single task. To enable a more sensitive analysis of brain regions whose shifts  
769 in activity patterns mirrored shifts in the semantic content of the episode or recalls, we restricted  
770 the temporal correlations we considered to the timescale of semantic information captured by our  
771 model. Specifically, we isolated the upper triangle of the episode correlation matrix and created a  
772 “proximal correlation mask” that included only diagonals from the upper triangle of the episode  
773 correlation matrix up to the first diagonal that contained no positive correlations. Applying this  
774 mask to the full episode correlation matrix was equivalent to excluding diagonals beyond the  
775 corner of the largest diagonal block. In other words, the timescale of temporal correlations we  
776 considered corresponded to the longest period of thematic stability in the episode, and by extension

777 the longest period of thematic stability in participants' recalls and the longest period of stability we  
778 might expect to see in voxel activity arising from processing or encoding episode content. Figure 8  
779 shows this proximal correlation mask applied to the temporal correlation matrices for the episode,  
780 an example participant's (warped) recall, and an example cube of voxels from our searchlight  
781 analyses.

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

789 We further sought to ensure that our analysis identified regions where the activations' temporal  
790 structure specifically reflected that of the episode, rather than regions whose activity was simply  
791 autocorrelated at a timescale similar to the episode template's diagonal. To achieve this, we used  
792 a phase shift-based permutation procedure, whereby we circularly shifted the episode's topic  
793 trajectory by a random number of timepoints, computed the resulting "null" episode template,  
794 and re-ran the searchlight analysis, in full. (For each of the 100 permutations, the same random shift  
795 was used for all participants). We *z*-scored the observed (unshifted) result at each voxel against  
796 the distribution of permutation-derived "null" results, and estimated a *p*-value by computing  
797 the proportion of shifted results that yielded larger values. To create the map in Figure 8C, we  
798 thresholded out any voxels whose similarity to the unshifted episode's structure fell below the 95<sup>th</sup>  
799 percentile of the permutation-derived similarity results.

800 We used an analogous procedure to identify which voxels' responses reflected the recall tem-  
801 plates. For each participant, we correlated the proximal diagonals from the upper triangle of the  
802 correlation matrix for each cube of voxels with the proximal diagonals from the upper triangle  
803 of their (time-warped) recall correlation matrix. As in the episode template analysis, this yielded  
804 a voxelwise map of correlation coefficients for each participant. However, whereas the episode

analysis compared every participant's responses to the same template, here the recall templates were unique for each participant. As in the analysis described above, we  $t$ -scored the (Fisher  $z$ -transformed) voxelwise correlations, and used the same permutation procedure we developed for the episode responses to ensure specificity to the recall timeseries and assign significance values. To create the map in Figure 8D we again thresholded out any voxels whose scores were below the 95<sup>th</sup> percentile of the permutation-derived null distribution.

## Neurosynth decoding analyses

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

## References

- Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Spence, K. W. and Spence, J. T., editors, *The psychology of learning and motivation*, volume 2, pages 89–105. Academic Press, New York.
- Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721.

- 828 Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas  
829 during narrative perception. *Journal of Neuroscience*, 38(45):9689–9699.
- 830 Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In  
831 *KDD workshop*, volume 10, pages 359–370.
- 832 Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International*  
833 *Conference on Machine Learning*, ICML ’06, pages 113–120, New York, NY, US. ACM.
- 834 Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*  
835 *Learning Research*, 3:993 – 1022.
- 836 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam,  
837 P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R.,  
838 Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S.,  
839 Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020).  
840 Language models are few-shot learners. *arXiv*, 2005.14165.
- 841 Brunec, I. K., Moscovitch, M. M., and Barense, M. D. (2018). Boundaries shape cognitive represen-  
842 tations of spaces and events. *Trends in Cognitive Sciences*, 22(7):637–650.
- 843 Capota, M., Turek, J., Chen, P.-H., Zhu, X., Manning, J. R., Sundaram, N., Keller, B., Wang, Y., and  
844 Shin, Y. S. (2017). Brain imaging analysis kit.
- 845 Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes,  
846 M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.  
847 *arXiv*, 1803.11175.
- 848 Charron, S. and Koechlin, E. (2010). Divided representations of current goals in the human frontal  
849 lobes. *Science*, 328(5976):360–363.
- 850 Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared  
851 memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*,  
852 20(1):115.

- 853 Clewett, D. and Davachi, L. (2017). The ebb and flow of experience determines the temporal  
854 structure of memory. *Curr Opin Behav Sci*, 17:186–193.
- 855 Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions.  
856 *Theory of Probability & Its Applications*, 15(3):458–486.
- 857 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the  
858 sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–1286.
- 859 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological  
860 Science*, 22(2):243–252.
- 861 Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd.
- 862 Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2001). Categorical representation of  
863 visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316.
- 864 Hasson, U., Chen, J., and Honey, C. J. (2015). Hierarchical process memory: memory as an integral  
865 component of information processing. *Trends in Cognitive Science*, 19(6):304–315.
- 866 Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal  
867 receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550.
- 868 Heusser, A. C., Ezzyat, Y., Shiff, I., and Davachi, L. (2018a). Perceptual boundaries cause mnemonic  
869 trade-offs between local boundary processing and across-trial associative binding. *Journal of  
870 Experimental Psychology Learning, Memory, and Cognition*, 44(7):1075–1090.
- 871 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail: a  
872 Python toolbox for analyzing and plotting free recall data. *The Journal of Open Source Software*,  
873 10.21105/joss.00424.
- 874 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018b). HyperTools: a Python  
875 toolbox for gaining geometric insights into high-dimensional data. *Journal of Machine Learning  
876 Research*, 18(152):1–6.

- Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46:269–299.
- Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., and H., E. (2014). A unified mathematical framework for coding time, space, and sequences in the medial temporal lobe. *Journal of Neuroscience*, 34(13):4692–4707.
- Howard, M. W., Viskontas, I. V., Shankar, K. H., and Fried, I. (2012). Ensembles of human MTL neurons “jump back in time” in response to a repeated stimulus. *Hippocampus*, 22:1833–1847.
- Huk, A., Bonnen, K., and He, B. J. (2018). Beyond trial-based paradigms: continuous behavior, ongoing neural activity, and naturalistic stimuli. *Journal of Neuroscience*, 10.1523/JNEUROSCI.1920-17.2018.
- Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532:453–458.
- Huth, A. G., Nisimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224.
- Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24:103–109.
- Kahana, M. J. (2012). *Foundations of Human Memory*. Oxford University Press, New York, NY.
- Koriat, A. and Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: distinguishing accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General*, 123(3):297–315.
- Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:1 – 28.
- Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato’s problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104:211–240.

- 902 Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy  
903 of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.
- 904 Manning, J. R. (2019). Episodic memory: mental time travel or a quantum ‘memory wave’ function?  
905 *PsyArXiv*, doi:10.31234/osf.io/6zjwb.
- 906 Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors, *Handbook*  
907 of Human Memory. Oxford University Press.
- 908 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory.  
909 In Gazzaniga, M., editor, *The Cognitive Neurosciences, Fifth edition*, pages 557–566. MIT Press.
- 910 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns  
911 in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National*  
912 *Academy of Sciences, USA*, 108(31):12893–12897.
- 913 McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and  
914 projection for dimension reduction. *arXiv*, 1802(03426).
- 915 Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations  
916 in vector space. *arXiv*, 1301.3781.
- 917 Mueller, A., Fillion-Robin, J.-C., Boidol, R., Tian, F., Nechifor, P., yoonsubKim, Peter, Rampin, R.,  
918 Corvellec, M., Medina, J., Dai, Y., Petrushev, B., Langner, K. M., Hong, Alessio, Ozsvald, I.,  
919 vkolmakov, Jones, T., Bailey, E., Rho, V., IgorAPM, Roy, D., May, C., foobuzz, Piyush, Seong,  
920 L. K., Goey, J. V., Smith, J. S., Gus, and Mai, F. (2018). WordCloud 1.5.0: a little word cloud  
921 generator in Python. *Zenodo*, <https://zenodo.org/record/1322068#.W4tPKZNKh24>.
- 922 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*,  
923 64:482–488.
- 924 Paller, K. A. and Wagner, A. D. (2002). Observing the transformation of experience into memory.  
925 *Trends in Cognitive Sciences*, 6(2):93–102.

- 926 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten-  
927 hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot,  
928 M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine*  
929 *Learning Research*, 12:2825–2830.
- 930 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model  
931 of organizational processes in free recall. *Psychological Review*, 116(1):129–156.
- 932 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal*  
933 *of Experimental Psychology*, 17:132–138.
- 934 Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech  
935 recognition. *Proceedings of the IEEE*, 77(2):257–286.
- 936 Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are  
937 unsupervised multitask learners. *OpenAI Blog*, 1(8).
- 938 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition. *Curr Opin*  
939 *Behav Sci*, 17:133–140.
- 940 Ramdas, A., Trillos, N., and Cuturi, M. (2017). On wasserstein two-sample testing and related  
941 families of nonparametric tests. *Entropy*, 19(2):47.
- 942 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature*  
943 *Reviews Neuroscience*, 13:713 – 726.
- 944 Rishel, C. A., Huang, G., and Freedman, D. J. (2013). Independent category and spatial encoding  
945 in parietal cortex. *Neuron*, 77(5):969–979.
- 946 Sigman, M. and Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during  
947 dual-task performance. *Journal of Neuroscience*, 28(30):7585–7589.
- 948 Simony, E. and Chang, C. (2020). Analysis of stimulus-induced brain dynamics during naturalistic  
949 paradigms. *NeuroImage*, 216:116461.

- 950 Simony, E., Honey, C. J., Chen, J., and Hasson, U. (2016). Uncovering stimulus-locked network  
951 dynamics during narrative comprehension. *Nature Communications*, 7(12141):1–13.
- 952 Tomary, A. and Davachi, L. (2017). Consolidation promotes the emergence of representational  
953 overlap in the hippocampus and medial prefrontal cortex. *Neuron*, 96(1):228–241.
- 954 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American Journal*  
955 *of Psychology*, 35:396–401.
- 956 Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale  
957 automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665.
- 958 Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research.  
959 *Journal of Memory and Language*, 46:441–517.
- 960 Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception:  
961 a mind-brain perspective. *Psychological Bulletin*, 133:273–293.
- 962 Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit  
963 memories to other brains: Constructing shared neural representations via communication. *Cereb*  
964 *Cortex*, 27(10):4988–5000.
- 965 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and  
966 memory. *Psychological Bulletin*, 123(2):162 – 185.

## 967 Supporting information

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

## 969 Acknowledgements

970 We thank Luke Chang, Janice Chen, Chris Honey, Lucy Owen, Emily Whitaker, and Kirsten Ziman  
971 for feedback and scientific discussions. We also thank Janice Chen, Yuan Chang Leong, Kenneth

972 Norman, and Uri Hasson for sharing the data used in our study. Our work was supported in part  
973 by NSF EPSCoR Award Number 1632738. The content is solely the responsibility of the authors  
974 and does not necessarily represent the official views of our supporting organizations.

975 **Author contributions**

976 Conceptualization: A.C.H. and J.R.M.; Methodology: A.C.H., P.C.F. and J.R.M.; Software: A.C.H.,  
977 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.,  
978 P.C.F. and J.R.M.; Supervision: J.R.M.

979 **Author information**

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