

¹ Memory for television episodes preserves event content
² while introducing new across-event similarities

³ Andrew C. Heusser^{1,2}, Paxton C. Fitzpatrick¹, and Jeremy R. Manning^{1,*}

¹Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

²Akili Interactive

Boston, MA 02110

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

⁴ February 4, 2020

⁵ **Abstract**

The ways our experiences unfold over time define unique *trajectories* through the relevant representational spaces. Within this geometric framework, one can compare the shape of the trajectory formed by an experience to that defined by our later remembering of that experience. We propose a framework for mapping naturalistic experiences onto geometric spaces that characterize how experiences are segmented into discrete events, and how the contents of event sequences evolve over time. We apply this approach to a naturalistic memory experiment which had participants view and recount a television episode. The content of participants' recounts of events from the original episode closely matched the original episode's content. However, the similarity patterns *across* events was much different in the original episode as compared with participants' recounts. We also identified a network of brain structures that are sensitive to the "shapes" of ongoing experiences, and an overlapping network that is sensitive (at the time of encoding) to how people later remembered those experiences in relation to other experiences.

18 In this way, modeling the content of richly structured experiences can reveal how (geometrically
19 and conceptually) those experiences are segmented into events and integrated into our memories
20 of other experiences.

21 **Introduction**

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

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

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

56 Occasionally, this gradual “drift” of our ongoing experience is punctuated by sudden changes,
57 or “shifts” (e.g., when we walk through a doorway; Radvansky and Zacks, 2017). Prior research
58 suggests that these sharp transitions (termed *event boundaries*) help to discretize our experiences
59 (and their mental representations) into *events* (Radvansky and Zacks, 2017; Brunec et al., 2018;
60 Heusser et al., 2018a; Clewett and Davachi, 2017; Ezzyat and Davachi, 2011; DuBrow and Davachi,
61 2013). The interplay between the stable (within-event) and transient (across-event) temporal
62 dynamics of an experience also provides a potential framework for transforming experiences into
63 memories that distill those experiences down to their essence. For example, prior work has shown
64 that event boundaries can influence how we learn sequences of items (Heusser et al., 2018a; DuBrow
65 and Davachi, 2013), navigate (Brunec et al., 2018), and remember and understand narratives (Zwaan
66 and Radvansky, 1998; Ezzyat and Davachi, 2011). Prior research has implicated the hippocampus
67 and the medial prefrontal cortex as playing a critical role in transforming experiences into structured
68 and consolidated memories (Tompry and Davachi, 2017).

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

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

90 Results

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

98 The second output is a *topic proportions matrix*, with one row per document and one column per
99 topic. The topic proportions matrix describes what mixture of discovered topics is reflected in each
100 document.

101 Chen et al. (2017) collected hand-annotated information about each of 1000 (manually identified)
102 time segments spanning the roughly 50 minute video used in their experiment. This information
103 included: a brief narrative description of what was happening, the location where the scene
104 took place, the names of any characters on the screen, and other similar details (for a full list of
105 annotated features, see *Methods*). We took from these annotations the union of all unique words
106 (excluding stop words, such as “and,” “or,” “but,” etc.) across all features and scenes as the
107 “vocabulary” for the topic model. We then concatenated the sets of words across all features
108 contained in overlapping, sliding windows of (up to) 50 scenes, and treated each window as a
109 single “document” for the purpose of fitting the topic model. Next, we fit a topic model with (up
110 to) $K = 100$ topics to this collection of documents. We found that 32 unique topics (with non-zero
111 weights) were sufficient to describe the time-varying content of the video (see *Methods*; Figs. 1, S2).
112 Note that our approach is similar in some respects to Dynamic Topic Models (Blei and Lafferty,
113 2006) in that we sought to characterize how the thematic content of the episode evolved over
114 time. However, whereas Dynamic Topic Models are designed to characterize how the properties
115 of *collections* of documents change over time, our sliding window approach allows us to examine
116 the topic dynamics within a single document (or video). Specifically, our approach yielded (via the
117 topic proportions matrix) a single *topic vector* for each sliding window of annotations transformed
118 by the topic model. We then stretched the resulting windows-by-topics matrix to match the time
119 series of the 1976 fMRI volumes collected as participants viewed the episode.

120 The 32 topics we found were heavily character-focused (i.e., the top-weighted word in each
121 topic was nearly always a character) and could be roughly divided into themes centered around
122 Sherlock Holmes (the titular character), John Watson (Sherlock’s close confidant and assistant),
123 supporting characters (e.g., Inspector Lestrade, Sergeant Donovan, or Sherlock’s brother Mycroft),
124 or the interactions between various pairs of these characters (see Fig. S2). Several of the identified
125 topics were highly similar, which we hypothesized might allow us to distinguish between subtle

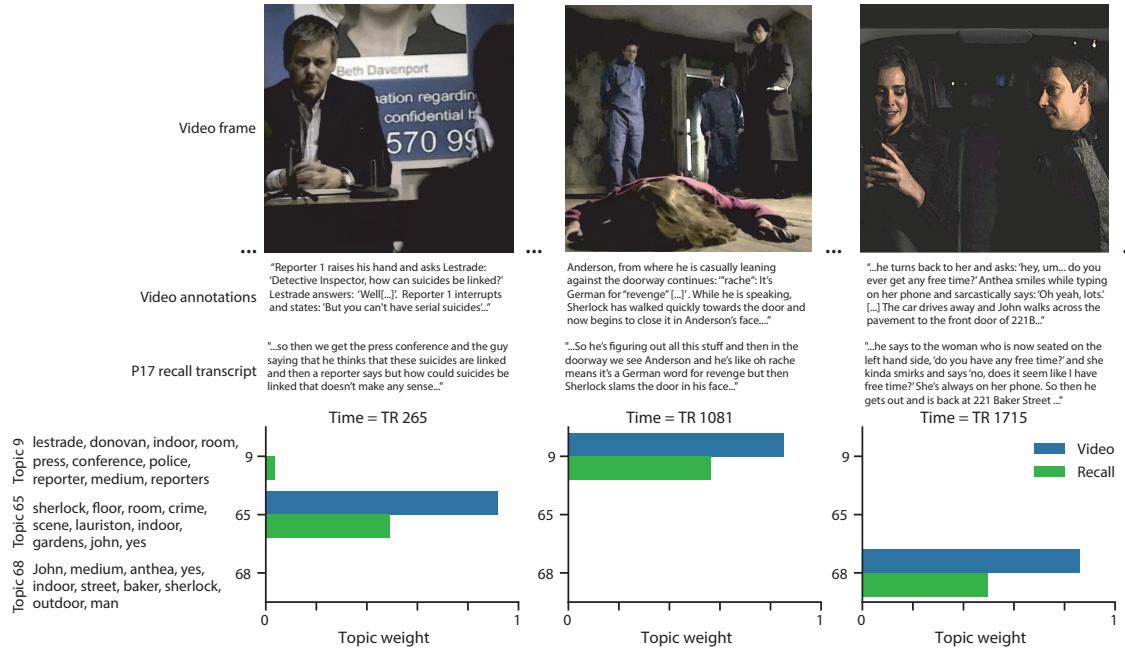


Figure 1: Methods overview. We used hand-annotated descriptions of each moment of video to fit a topic model. Three example video frames and their associated descriptions are displayed (top two rows). Participants later recalled the video (in the third row, we show example recalls of the same three scenes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

126 narrative differences if the distinctions between those overlapping topics were meaningful. The
127 topic vectors for each timepoint were *sparse*, in that only a small number (usually one or two) of
128 topics tended to be “active” in any given timepoint (Fig. 2A). Further, the dynamics of the topic
129 activations appeared to exhibit *persistence* (i.e., given that a topic was active in one timepoint, it was
130 likely to be active in the following timepoint) along with *occasional rapid changes* (i.e., occasionally
131 topics would appear to spring into or out of existence). These two properties of the topic dynamics
132 may be seen in the block diagonal structure of the timepoint-by-timepoint correlation matrix
133 (Fig. 2B) and reflect the gradual drift and sudden shifts fundamental to the temporal dynamics of
134 real-world experiences. Given this observation, we adapted an approach devised by Baldassano
135 et al. (2017), and used a Hidden Markov Model (HMM) to identify the *event boundaries* where the
136 topic activations changed rapidly (i.e., at the boundaries of the blocks in the correlation matrix;
137 event boundaries identified by the HMM are outlined in yellow in Fig. 2B). Part of our model fitting
138 procedure required selecting an appropriate number of “events” into which the topic trajectory
139 should be segmented. To accomplish this, we used an optimization procedure that maximized the
140 difference between the topic weights for timepoints within an event and across multiple events
141 (see *Methods* for additional details). We then created a stable “summary” of the content within
142 each video event by averaging the topic vectors across timepoints each event spanned (Fig. 2C).

143 Given that the time-varying content of the video could be segmented cleanly into discrete
144 events, we wondered whether participants’ recalls of the video also displayed a similar structure.
145 We applied the same topic model (already trained on the video annotations) to each participant’s
146 recalls. Analogous to how we parsed the time-varying content of the video, to obtain similar esti-
147 mates for each participant’s recall, we treated each overlapping “window” of (up to 10) sentences
148 from their transcript as a “document,” and computed the most probable mix of topics reflected in
149 each timepoint’s sentences. This yielded, for each participant, a number-of-windows by number-
150 of-topics topic proportions matrix that characterized how the topics identified in the original video
151 were reflected in the participant’s recalls. Note that an important feature of our approach is that it
152 allows us to compare participants’ recalls to events from the original video, despite different par-
153 ticipants using widely varying language to describe the same event, and that those descriptions



Figure 2: Modelling naturalistic stimuli and recalls. All panels: darker colors indicate greater values; range: [0, 1]. **A.** Topic vectors ($K = 100$) for each of the 1976 video timepoints. **B.** Timepoint-by-timepoint correlation matrix of the topic vectors displayed in Panel A. Event boundaries discovered by the HMM are denoted in yellow (30 events detected). **C.** Average topic vectors for each of the 30 video events. **D.** Topic vectors for each of 265 sliding windows of sentences spoken by an example participant while recalling the video. **E.** Timepoint-by-timepoint correlation matrix of the topic vectors displayed in Panel D. Event boundaries detected by the HMM are denoted in yellow (22 events detected). For similar plots for all participants see Figure S4. **F.** Average topic vectors for each of the 22 recalled events from the example participant. **G.** Correlations between the topic vectors for every pair of video events (Panel C) and recalled events (from the example participant; Panel F). For similar plots for all participants, see Figure S5. **H.** Average correlations between each pair of video events and recalled events (across all 17 participants). To create the figure, each recalled event was assigned to the video event with the most correlated topic vector (yellow boxes in panels G and H). The heat maps in each panel were created using Seaborn (Waskom et al., 2016).

¹⁵⁴ may not match the original annotations. This is a substantial benefit of projecting the video and
¹⁵⁵ recalls into a shared “topic” space. An example topic proportions matrix from one participant’s
¹⁵⁶ recalls is shown in Figure 2D.

¹⁵⁷ Although the example participant’s recall topic proportions matrix has some visual similarity to
¹⁵⁸ the video topic proportions matrix, the time-varying topic proportions for the example participant’s
¹⁵⁹ recalls are not as sparse as those for the video (compare Figs. 2A and D). Similarly, although there do
¹⁶⁰ appear to be periods of stability in the recall topic dynamics (i.e., most topics are active or inactive
¹⁶¹ over contiguous blocks of time), the individual topics’ overall timecourses are not as cleanly
¹⁶² delineated as the video topics’. To examine these patterns in detail, we computed the timepoint-
¹⁶³ by-timepoint correlation matrix for the example participant’s recall topic trajectory (Fig. 2E). As
¹⁶⁴ in the video correlation matrix (Fig. 2B), the example participant’s recall correlation matrix has a
¹⁶⁵ strong block diagonal structure, indicating that their recalls are discretized into separated events.
¹⁶⁶ As for the video correlation matrix, we can use an HMM, along with the aforementioned number-
¹⁶⁷ of-events optimization procedure (also see *Methods*) to determine how many events are reflected
¹⁶⁸ in the participant’s recalls and where specifically the event boundaries fall (outlined in yellow).
¹⁶⁹ We carried out a similar analysis on all 17 participants’ recall topic proportions matrices (Fig. S4).

¹⁷⁰ Two clear patterns emerged from this set of analyses. First, although every individual partic-
¹⁷¹ ipant’s recalls could be segmented into discrete events (i.e., every individual participant’s recall
¹⁷² correlation matrix exhibited clear block diagonal structure; Fig. S4), each participant appeared to
¹⁷³ have a unique *recall resolution*, reflected in the sizes of those blocks. While, some participants’ recall
¹⁷⁴ topic proportions segmented into just a few events (e.g., Participants P4, P5, and P7), others’ seg-
¹⁷⁵ mented into many shorter duration events (e.g., Participants P12, P13, and P17). This suggests that
¹⁷⁶ different participants may be recalling the video with different levels of detail— e.g., some might
¹⁷⁷ touch on just the major plot points, whereas others might attempt to recall every minor scene or ac-
¹⁷⁸ tion. The second clear pattern present in every individual participant’s recall correlation matrix is
¹⁷⁹ that, unlike in the video correlation matrix, there are substantial off-diagonal correlations. Whereas
¹⁸⁰ each event in the original video was (largely) separable from the others (Fig. 2B), in transforming
¹⁸¹ those separable events into memory, participants appear to be integrating across multiple events,

182 blending elements of previously recalled and not-yet-recalled content into each newly recalled
183 event (Figs. 2D, S4; also see Manning et al., 2011; Howard et al., 2012).

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

205 We can also apply two list-learning-native analyses that describe how participants group items
206 in their recall sequences: temporal clustering and semantic clustering (Polyn et al., 2009, see
207 *Methods* for details). Temporal clustering refers to the extent to which participants group their
208 recall responses according to encoding position. Overall, we found that sequentially viewed video
209 events were clustered heavily in participants' recall event sequences (mean: 0.767, SEM: 0.029),

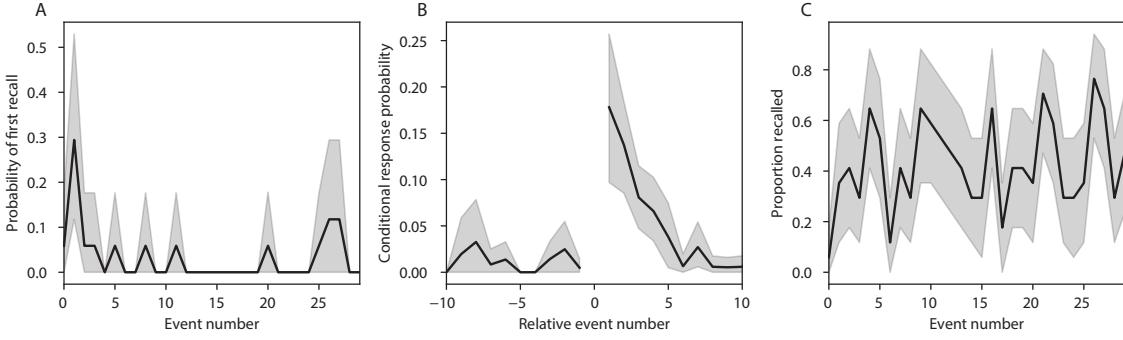


Figure 3: Naturalistic extensions of classic list-learning memory analyses. **A.** The probability of first recall as a function of the serial position of the event in the video. **B.** The probability of recalling each event, conditioned on having most recently recalled the event *lag* events away in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

and that participants with higher temporal clustering scores tended to perform better according to both Chen et al. (2017)'s hand-annotated memory scores (Pearson's $r(15) = 0.62$, $p = 0.008$) and our model's estimate (Pearson's $r(15) = 0.54$, $p = 0.024$). 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 video events together (mean: 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-derived (Pearson's $r(15) = 0.63$, $p = 0.007$) memory performance.

Statistical models of memory studies often treat recall success as binary (i.e., an item either was or was not recalled), or occasionally categorical (e.g., to distinguish familiarity from recollection; Yonelinas et al., 2002). Such approaches are tenable in classical list-learning or recognition memory paradigms, as the presented stimuli tend to be very simple (e.g., a sequence of individual words or items). However, the feature-rich content of a naturalistic experiences may later be described with many, highly variable levels of success. Our framework produces a content-based model of individual stimulus and recall events by projecting the dynamic content of the video and participants' recalls into a shared topic space. This allows for direct, quantitative comparison between all stimulus and recall events, as well as between the recall events themselves. Leveraging

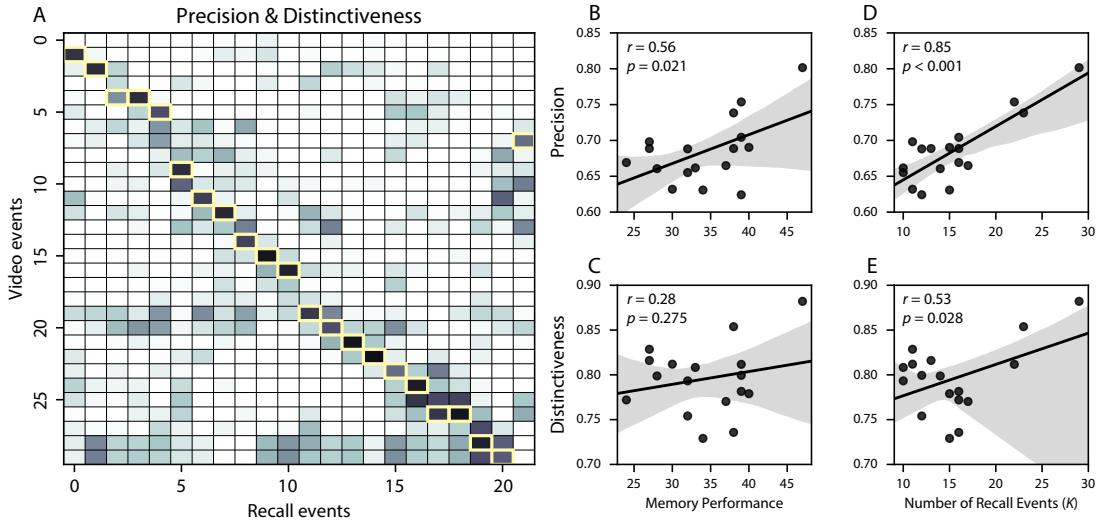


Figure 4: Novel content-based metrics of naturalistic memory: precision and distinctiveness. A. The video-recall correlation matrix for a representative participant (17). The yellow boxes highlight the maximum correlation in each column. The example participant's overall precision score was computed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

these content-based models of the stimulus/recall events, we developed two novel, *continuous* metrics for analyzing naturalistic memory: *precision* and *distinctiveness*. We define precision as the “completeness” of recall, or how fully the presented content was recapitulated in memory. Under our framework, we quantify this for a given recall event as the correlation between the topic proportions of the recall event and the maximally correlated video event (Fig. 4). A second novel metric we introduce here is *distinctiveness*, which we define as the “specificity” of recall, or how unique the description of a given section of content was, compared to descriptions for other sections of content. We quantify this for each recall event as 1 minus the average correlation between the given recall event and all other recall events not matched to the same video event. In addition to individual events, one may also use these metrics to describe each participant’s overall performance (i.e., by averaging across a participant’s event-wise precision or distinctiveness

238 scores). Participants whose recall events are more veridical descriptions of what happened in the
239 video event will presumably have higher precision scores. We find that, across participants,
240 a higher precision score is correlated to both hand-annotated memory performance (Pearson's
241 $r(15) = 0.56, p = 0.021$) and the number of recall events estimated by our model (Pearson's $r(15) =$
242 $0.85, p < 0.001$). We also hypothesized that participants who recounted events in a more distinctive
243 way would display better overall memory. We find that this distinctiveness score is related to
244 our model's estimated number of recalled events (Pearson's $r(15) = 0.53, p = 0.028$), and while
245 we do not find distinctiveness to be related to hand-annotated memory performance (Pearson's
246 $r(15) = 0.28, p = 0.275$), this is not entirely surprising given how the hand-annotated memory
247 scores were computed (see *Methods*).

248 Further intuition for the behaviors captured by these two metrics may be gained by directly
249 examining the content of the video and recalls our framework models. In Figure 5, we contrast
250 recalls for the same video event (event 22) from two participants: one with a high precision score
251 (P17), the other with a low precision score (P6). From the HMM-identified event boundaries,
252 we recovered the set of annotations describing the content of an example video event (Fig. 5B),
253 and divided them into different color-coded sections for each action or feature described. We
254 then similarly recovered the set of sentences comprising the corresponding recall event for each
255 of the two example participants. Because the recall sliding windows overlap heavily, and each
256 recall event spans multiple recall timepoints (i.e., windows), we have stripped any sentences from
257 the beginning and end that describe earlier or later video events for the sake of readability. In
258 other words, Fig. 5C shows a subset of the full recall event text, comprising sentences between
259 the first and last descriptions of content from the example video event. We then colored all words
260 describing actions and features coded in panel B by their corresponding color. Visual comparison
261 of the transcripts reveals that the most precise participant's recall both captures more of the video
262 event's content, and does so with far more detail.

263 Figure 6 similarly contrasts two example participants' recalls for a common video event (event
264 19) to illustrate the tangible differences between high and low distinctiveness scores. Here, we
265 have extracted the full set of sentences comprising the most distinctive recall event (P13) and least

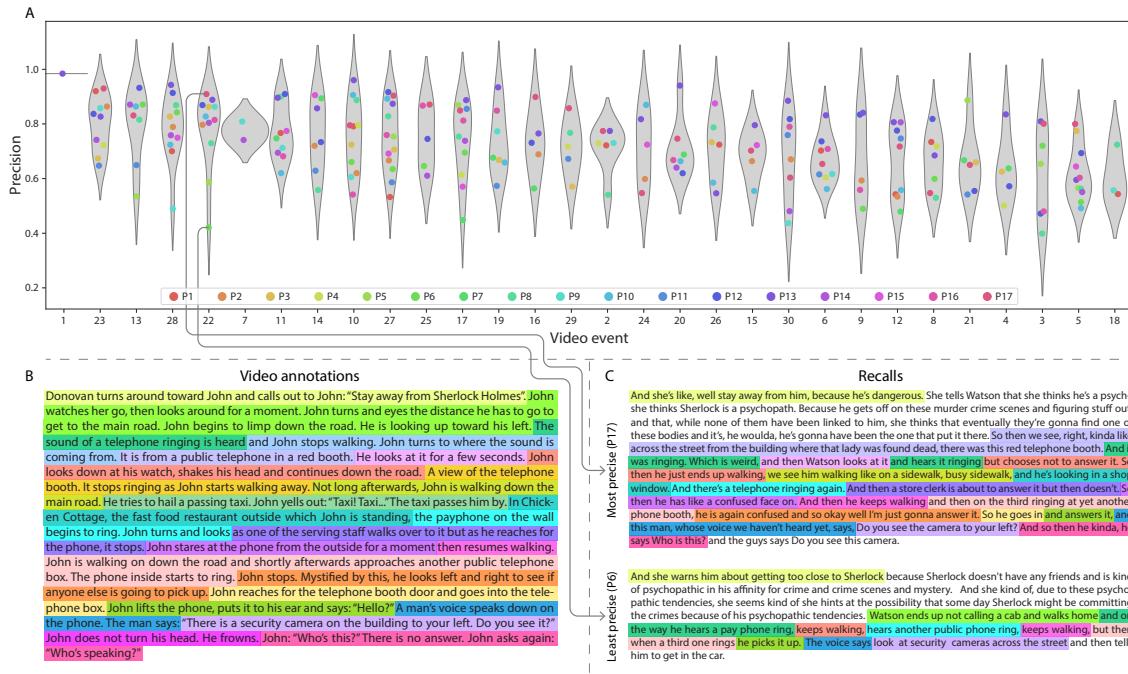


Figure 5: Precision metric reflects completeness of recall. **A.** Recall precision by video event. Grey violin plots display kernel density estimates for the distribution of recall precision scores for a single video event. Colored dots within each violin plot represent individual participants' recall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

266 distinctive recall event (P11) recall event matched to the example video event (Fig. 6C). We also
 267 extracted the annotations for the example video event, as well as those from each other video
 268 event whose content the example participants' single recall events described (Fig. 6B). We then
 269 shaded the annotation text for each video event with a different color, and shaded each word of
 270 the example participants' recall text by the color of the video event it describes. The majority of
 271 the most distinctive recall event text describes video event 19's content, with the first five and last
 272 one sentence describing the video events immediately preceding and succeeding the current one,
 273 respectively. Meanwhile, the least precise participant's recall for video event 19 blends the content
 274 from five separate video events, does not transition between them in order, and often combines

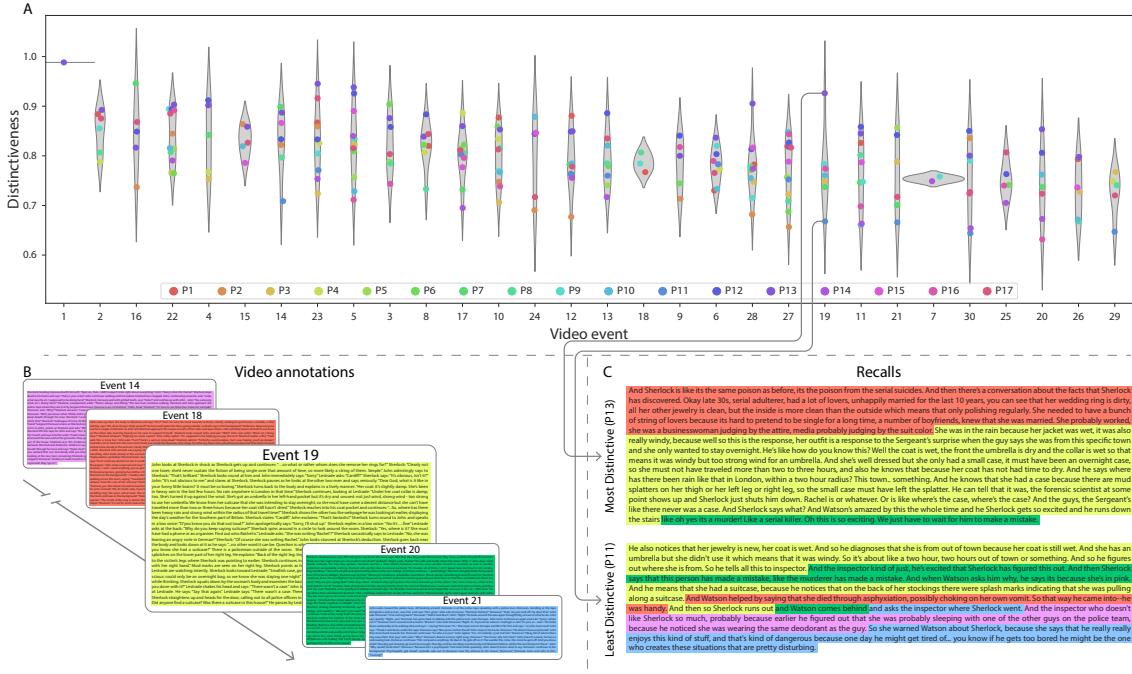


Figure 6: Distinctiveness metric reflects specificity of recall. A. Recall distinctiveness by video event. Kernel density estimates for each video event’s distribution of recall distinctiveness scores, analogous to Fig. 5A. B. The sets of “Narrative Details” video annotations (generated by Chen et al.,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

²⁷⁵ descriptions of two video events' content in the same sentence

The prior analyses leverage the correspondence between the 100-dimensional topic proportion matrices for the video and participants' recalls to characterize recall. However, it is difficult to gain deep insights into the content of (or relationships between) experiences and memories solely by examining these topic proportions (e.g., Figs. 2A, D) or the corresponding correlation matrices (Figs. 2B, E, S4). And while we can directly examine the original text underlying these topic vectors (e.g., Figs. 5, 6) to show how relationships between them reflect real-world behavior, this comparison becomes prohibitively cumbersome at larger timescales. To visualize the time-varying high-dimensional content in a more intuitive way (Heusser et al., 2018b) we projected the topic proportions matrices onto a two-dimensional space using Uniform Manifold Approximation and

285 Projection (UMAP; McInnes et al., 2018). In this lower-dimensional space, each point represents a
286 single video or recall event, and the distances between the points reflect the distances between the
287 events' associated topic vectors (Fig. 7). In other words, events that are nearer to each other in this
288 space are more semantically similar, and those that are farther apart are less so.

289 Visual inspection of the video and recall topic trajectories reveals a striking pattern. First,
290 the topic trajectory of the video (which reflects its dynamic content; Fig. 7A) is captured nearly
291 perfectly by the averaged topic trajectories of participants' recalls (Fig. 7B). To assess the consistency
292 of these recall trajectories across participants, we asked: given that a participant's recall trajectory
293 had entered a particular location in topic space, could the position of their *next* recalled event
294 be predicted reliably? For each location in topic space, we computed the set of line segments
295 connecting successively recalled events (across all participants) that intersected that location (see
296 *Methods* for additional details). We then computed (for each location) the distribution of angles
297 formed by the lines defined by those line segments and a fixed reference line (the *x*-axis). Rayleigh
298 tests revealed the set of locations in topic space at which these across-participant distributions
299 exhibited reliable peaks (blue arrows in Fig. 7B reflect significant peaks at $p < 0.05$, corrected). We
300 observed that the locations traversed by nearly the entire video trajectory exhibited such peaks.
301 In other words, participants exhibited similar trajectories that also matched the trajectory of the
302 original video (Fig. 7C). This is especially notable when considering the fact that the number of
303 events participants recalled (dots in Fig. 7C) varied considerably across people, and that every
304 participant used different words to describe what they had remembered happening in the video.
305 Differences in the numbers of remembered events appear in participants' trajectories as differences
306 in the sampling resolution along the trajectory. We note that this framework also provides a
307 means of disentangling classic "proportion recalled" measures (i.e., the proportion of video events
308 described in participants' recalls) from participants' abilities to recapitulate the overall unfolding
309 of the original video's content (i.e., the similarity between the shapes of the original video trajectory
310 and that defined by each participant's recounting of the video).

311 The results displayed in Figures 3C and 5A suggest that certain events were remembered better
312 than others. Given this, we next asked whether the events were generally remembered

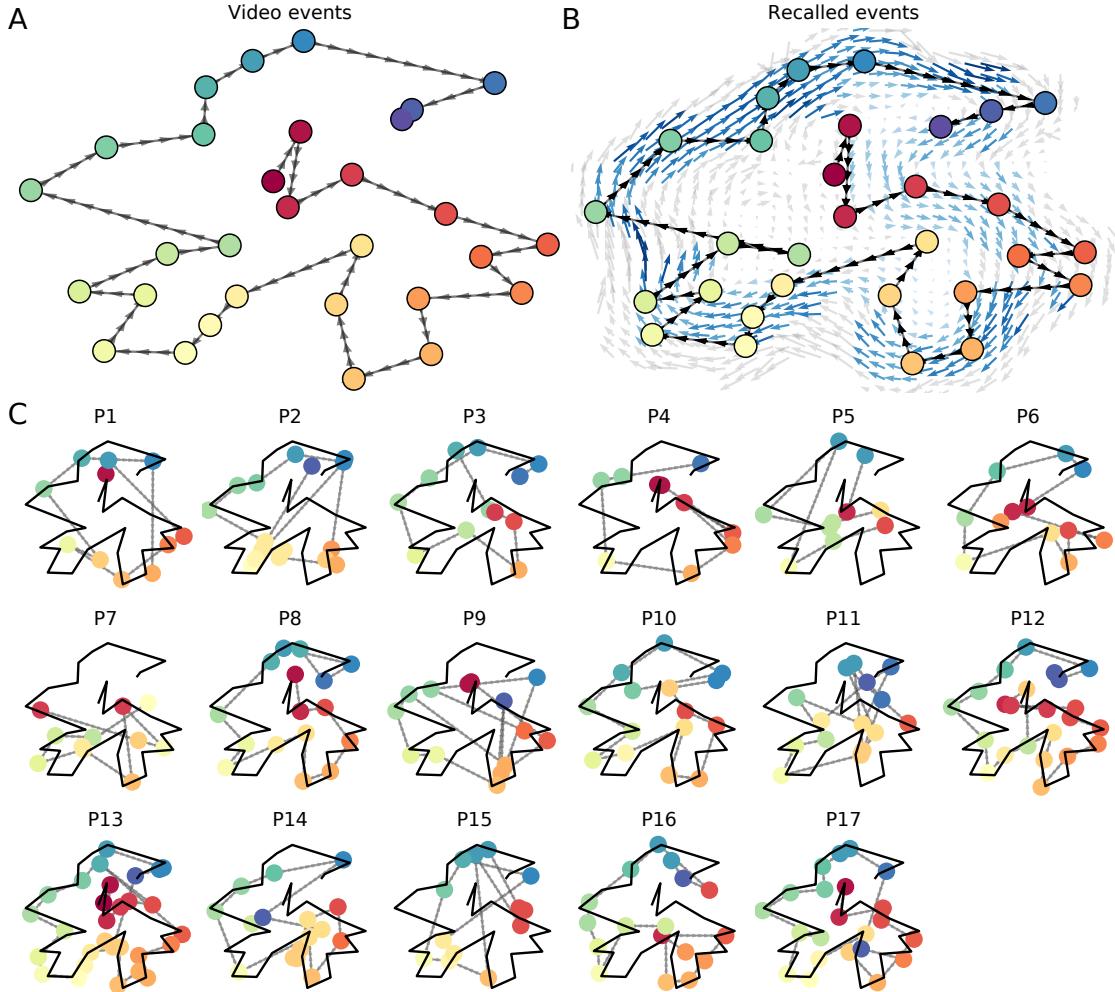


Figure 7: Trajectories through topic space capture the dynamic content of the video and recalls. All panels: the topic proportion matrices have been projected onto a shared two-dimensional space using UMAP. **A.** The two-dimensional topic trajectory taken by the episode of *Sherlock*. Each dot indicates an event identified using the HMM (see *Methods*); the dot colors denote the order of the events (early events are in red; later events are in blue), and the connecting lines indicate the transitions between successive events. **B.** The average two-dimensional trajectory captured by participants' recall sequences, with the same format and coloring as the trajectory in Panel A. To compute the event positions, we matched each recalled event with an event from the original video (see *Results*), and then we averaged the positions of all events with the same label. The arrows reflect the average transition direction through topic space taken by any participants whose trajectories crossed that part of topic space; blue denotes reliable agreement across participants via a Rayleigh test ($p < 0.05$, corrected). **C.** The recall topic trajectories (gray) taken by each individual participant (P1–P17). The video's trajectory is shown in black for reference. Here, events (dots) are colored by their matched video event (Panel A).

313 well or poorly tended to reflect particular content. Because our analysis framework projects the
314 dynamic video content and participants' recalls into a shared space, and because the dimensions
315 of that space represent topics (which are, in turn, sets of weights over words in the vocabulary), we
316 are able to recover the weighted combination of words that make up any point (i.e., topic vector) in
317 this space. We first computed the average precision with which participants recalled each of the 30
318 video events (Fig. 8A; note that this result is analogous to a serial position curve created from our
319 continuous recall quality metric). We then computed a weighted average of the topic vectors for
320 each video event, where the weights reflected how reliably each event was recalled. To visualize
321 the result, we created a "wordle" image (Mueller et al., 2018) where words weighted more heavily
322 by better-remembered topics appear in a larger font (Fig. 8B, green box). Across the full video,
323 content that reflected topics necessary to convey the central focus of the video (e.g., the names of the
324 two main characters, "Sherlock" and "John", and the address of a major recurring location, "221B
325 Baker Street") were best remembered. An analogous analysis revealed which themes were poorly
326 remembered. Here in computing the weighted average over events' topic vectors, we weighted
327 each event in *inverse* proportion to how well it was remembered (Fig. 8B, red box). The least well-
328 remembered video content reflected information not necessary to later convey a general summary
329 of the video, such as the proper names of relatively minor characters (e.g., "Mike," "Molly," and
330 "Lestrade") and locations (e.g., "St. Bartholomew's Hospital").

331 A similar result emerged from assessing the topic vectors for individual video and recall events
332 (Fig. 8C). Here, for each of the three best- and worst-remembered video events, we have constructed
333 two wordles: one from the original video event's topic vector (left) and a second from the average
334 recall topic vector for that event (right). The three best-remembered events (circled in green)
335 correspond to scenes important to the central plot-line: a mysterious figure spying on John in a
336 phone booth; John meeting Sherlock at Baker St. to discuss the murders; and Sherlock laying
337 a trap to catch the killer. Meanwhile, the three worst-remembered events (circled in red) reflect
338 scenes that are non-essential to summarizing the narrative's structure: the video of singing cartoon
339 characters participants viewed prior to the main episode; John asking Molly about Sherlock's habit
340 of over-analyzing people; and Sherlock noticing evidence of Anderson's and Donovan's affair.

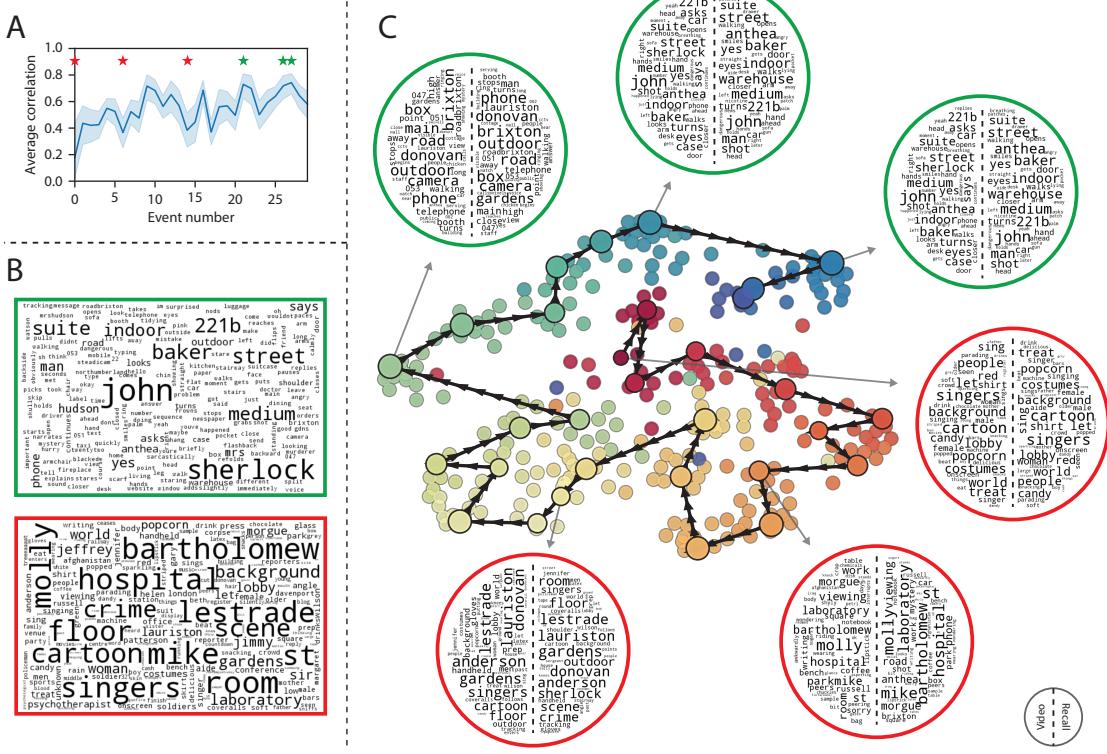


Figure 8: Transforming experience into memory. **A.** Average precision (video event-recall event topic vector correlation) across participants for each video event. Error bars denote bootstrap-derived across-participant 95% confidence intervals. The stars denote the three best-remembered events (green) and worst-remembered events (red). **B.** Wordles comprising the top 200 highest-weighted words reflected in the weighted-average topic vector across video events. Green: video events were weighted by how well the topic vectors derived from recalls of those events matched the video events' topic vectors (Panel A). Red: video events were weighted by the inverse of how well their topic vectors matched the recalled topic vectors. **C.** The set of all video and recall events is projected onto the two-dimensional space derived in Figure 7. The dots outlined in black denote video events (dot size reflects the average correlation between the video event's topic vector and the topic vectors from the closest matching recalled events from each participant; bigger dots denote stronger correlations). The dots without black outlines denote recalled events. All dots are colored using the same scheme as Figure 7A. Wordles for several example events are displayed (green: three best-remembered events; red: three worst-remembered events). Within each circular wordle, the left side displays words associated with the topic vector for the video event, and the right side displays words associated with the (average) recall event topic vector, across all recall events matched to the given video event.

341 The results thus far inform us about which aspects of the dynamic content in the episode
342 participants watched were preserved or altered in participants' memories of the episode. We next
343 carried out a series of analyses aimed at understanding which brain structures might facilitate
344 these preservations and alterations as the explicit content of the video was transformed into
345 participants' memories. In one analysis we sought to identify brain structures that were sensitive
346 to the dynamic unfolding of the video's content, as characterized by its topic trajectory. Specifically,
347 we used a searchlight procedure to identify clusters of voxels whose proximal temporal correlation
348 matrix of activity (as participants watched the video) matched the proximal temporal correlation
349 matrix of original video's topic proportions (Fig. 9A; see *Methods* for additional details). We
350 then performed a phase shift-based permutation correction to ensure our searchlight procedure
351 identified regions specifically sensitive to the temporal structure of the video (i.e., rather than those
352 with an autocorrelation length similar to the video; see *Methods*). We z-scored the searchlight results
353 at each voxel against the null distribution of (100) permuted searchlight results, thresholding at
354 $p = 0.05$. As shown in Figure 9C, the analysis revealed a distributed network of regions including
355 ????, suggesting that these regions may play a role in processing information relevant to the
356 narrative structure of the video. In a second analysis, we sought to identify which brain structures'
357 responses (while viewing the video) reflected how each participant would later structure their
358 *recall* of the video. We used an analogous searchlight procedure and permutation correction to
359 identify clusters of voxels whose proximal temporal correlation matrices reflected the proximal
360 temporal correlation matrix of the topic proportions for each individual's recall (Figs. 9B; see
361 *Methods* for additional details). As shown in Figure 9D, the analysis revealed a network of regions
362 including ????, suggesting that these regions may play a role in transforming each individual's
363 experience into memory. In identifying regions whose responses to ongoing experiences reflect
364 how those experiences will be remembered later, this latter analysis extends classic *subsequent*
365 *memory analyses* (e.g., Paller and Wagner, 2002) to domain of naturalistic stimuli.

366 The set of searchlight analyses described above yielded two distributed networks of brain
367 regions, whose activity timecourses mirrored to the temporal structure of the video (Fig. 9C) or
368 participants' eventual recalls (Fig. 9D). As an exploratory analysis, we sought to identify structures

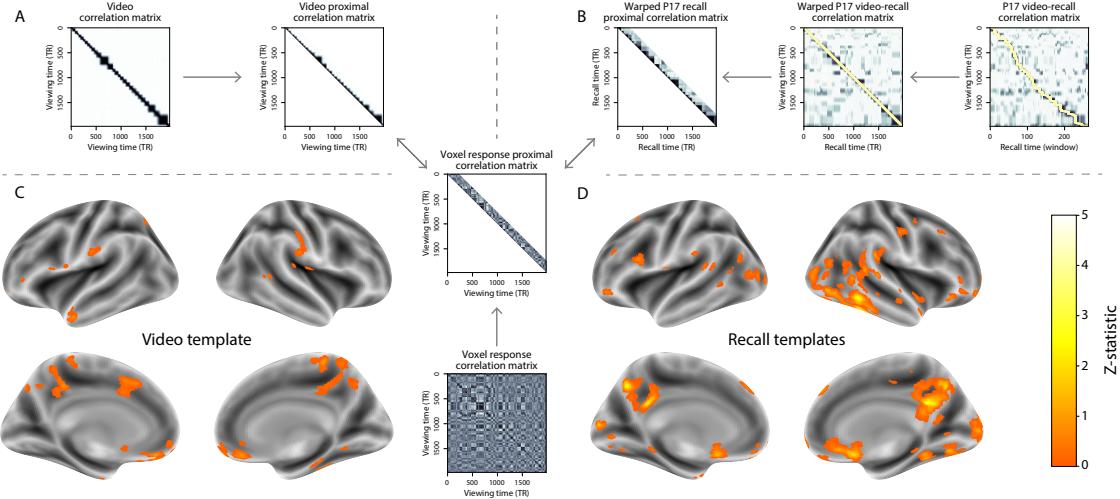


Figure 9: Brain structures that underlie the transformation of experience into memory. **A.** We isolated the proximal diagonals from the upper triangle of the video correlation matrix, and applied this same diagonal mask to the voxel response correlation matrix for each cube of voxels in the brain.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

369 comprised these networks and what functions are shared among them. To accomplish this in a
 370 blind, unbiased manner (i.e., without reverse inference via visual observation) we used Neurosynth
 371 (Yarkoni et al., 2011). Neurosynth parses a massive online database of over 14,000 neuroimaging
 372 studies and constructs meta-analysis images for over 13,000 psychology- and neuroscience-related
 373 terms based on NIfTI images accompanying studies where those terms appear at a high frequency.
 374 Given a novel image tagged with its value type (e.g., t -, F - or p -statistics), Neurosynth can then
 375 *decode* the list of terms whose meta-analysis images are most similar. We constructed brain maps of
 376 voxelwise p -values derived from our permutation correction (masking voxels with a corresponding
 377 permutation-derived z -score at or below zero). These maps describe the extent to which each
 378 voxel specifically reflected the temporal structure of the video or individuals' recalls (i.e., for

³⁷⁹ each voxel, the percentage of phase-shifted topic vector correlation matrices to which the voxel
³⁸⁰ activity correlation matrix was less similar). These *p*-value maps for the video- and recall-driven
³⁸¹ searchlight analyses are displayed in Figure 10, along with the top ten terms decoded from each
³⁸² using Neurosynth.

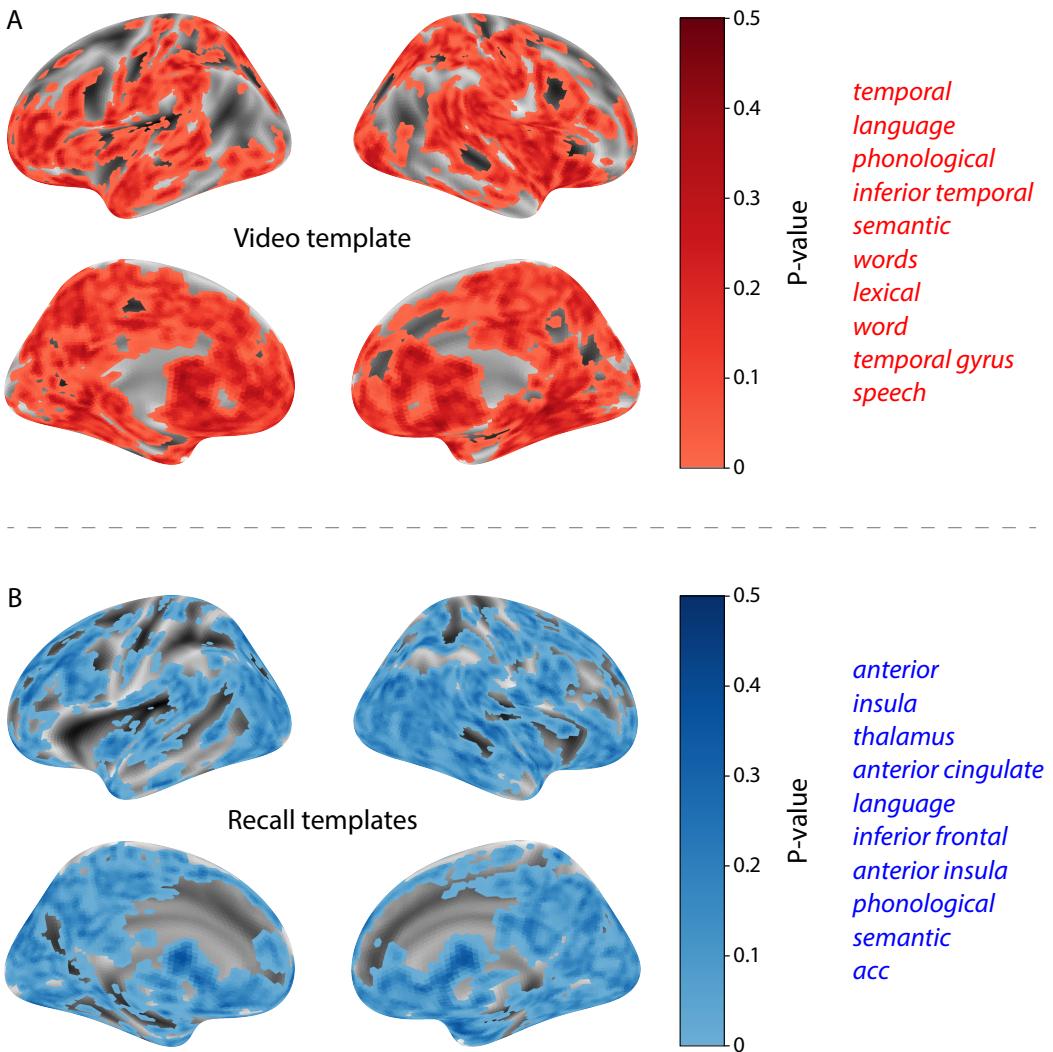


Figure 10: Decoding distributed statistical maps via Neurosynth meta-analyses. **A.** Video-searchlight p -map and top 10 decoded terms. We constructed a map of the permutation-derived p -values for the video-driven searchlight analysis (Fig. 9A, C) for all voxels with a positive permutation-derived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

383 **Discussion**

384 Our work casts remembering as reproducing (behaviorally and neurally) the topic trajectory, or
385 shape, of an experience. This view draws inspiration from prior work aimed at elucidating
386 the neural and behavioral underpinnings of how we process dynamic naturalistic experiences
387 and remember them later. One approach to identifying neural responses to naturalistic stimuli
388 (including experiences) entails building a model of the stimulus and searching for brain regions
389 whose responses are consistent with the model. In prior work, a series of studies from Uri
390 Hasson's group (Lerner et al., 2011; Simony et al., 2016; Chen et al., 2017; Baldassano et al., 2017;
391 Zadbood et al., 2017) have extended this approach with a clever twist: rather than building an
392 explicit stimulus model, these studies instead search for brain responses (while experiencing the
393 stimulus) that are reliably similar across individuals. So called *inter-subject correlation* (ISC) and
394 *inter-subject functional connectivity* (ISFC) analyses effectively treat other people's brain responses
395 to the stimulus as a "model" of how its features change over time. By contrast, in our present
396 work we used topic models and HMMs to construct an explicit stimulus model (i.e., the topic
397 trajectory of the video). When we searched for brain structures whose responses are consistent
398 with the video's topic trajectory, we identified a network of structures that overlapped strongly
399 with the "long temporal receptive window" network reported by the Hasson group (e.g., compare
400 our Fig. 9A with the map of long temporal receptive window voxels in Lerner et al., 2011). This
401 provides support for the notion that part of the long temporal receptive window network may be
402 maintaining an explicit model of the stimulus dynamics. When we performed a similar analysis
403 after swapping out the video's topic trajectory with the recall topic trajectories of each individual
404 participant, this allowed us to identify brain regions whose responses (as the participants viewed
405 the video) reflected how the video trajectory would be transformed in memory (as reflected by
406 the recall topic trajectories). The analysis revealed that the rMTL and vmPFC may play a role in
407 this person-specific transformation from experience into memory. The role of the MTL in episodic
408 memory encoding has been well-reported (e.g., Paller and Wagner, 2002; Davachi et al., 2003;
409 Ranganath et al., 2004; Davachi, 2006; Wiltgen and Silva, 2007; Diana et al., 2007; van Kesteren

et al., 2013). Prior work has also implicated the medial prefrontal cortex in representing “schema” knowledge (i.e., general knowledge about the format of an ongoing experience given prior similar experiences; van Kesteren et al., 2012, 2013; Schlichting and Preston, 2015; Gilboa and Marlatt, 2017; Spalding et al., 2018). Integrating across our study and this prior work, one interpretation is that the person-specific transformations mediated (or represented) by the rMTL and vmPFC may reflect schema knowledge being leveraged, formed, or updated, incorporating ongoing experience into previously acquired knowledge.

In extending classical free recall analyses to our naturalistic memory framework, we recovered two patterns of recall dynamics central to list-learning studies: a high probability of initiating recall with the first video event (Fig. 3A) and a strong bias toward transitioning from recalling a given event to recalling the event immediately following it (Fig. 3B). However, equally noteworthy are the typical free recall results not recovered in these analyses, as each highlights a fundamental difference between list-learning studies and naturalistic memory paradigms like the one employed in the present study. The most noticeable departure from hallmark free recall dynamics in these findings is the apparent lack of a serial position effect in Figure 3C, which instead shows greater and lesser recall probabilities for events distributed across the video stimulus. Stimuli in free recall experiments most often comprise lists of simple, common words, presented to participants in a random order. (In fact, numerous word pools have been developed based on these criteria; e.g., Friendly et al., 1982). These stimulus qualities enable two assumptions that are central to word list analyses, but frequently do not hold for real-world experiences. First, researchers conducting free recall studies may assume that the content at each presentation index is essentially equal, and does not bear qualities that would cause participants to remember it more or less successfully than others. Such is rarely the case with real-world experiences or experiments meant to approximate them, and the effects of both intrinsic and observer-dependent factors on stimulus memorability are well established (for review see Chun and Turk-Browne, 2007; Bylinskii et al., 2015; Tyng et al., 2017). Second, the random ordering of list items ensures that (across participants, on average) there is no relationship between the thematic similarity of individual stimuli and their presentation positions—in other words, two semantically related words are no more likely to be

438 presented next to each other than at opposite ends of the list. In most cases, the exact opposite
439 is true of real-world episodes. Our internal thoughts, our actions, and the physical state of the
440 world around us all tend to follow a direct, causal progression. As a result, each moment of our
441 experience tends to be inherently more similar to surrounding moments than to those in the distant
442 past or future. Memory literature has termed this strong temporal autocorrelation “context,” and
443 in various media that depict real-world events (e.g., movies and written stories), we recognize
444 it as a *narrative structure*. While a random word list (by definition) has no such structure, the
445 logical progression between ideas and actions in a naturalistic stimulus prompts the rememberer
446 to recount presented events in order, starting with the beginning. This tendency is reflected in our
447 findings’ second departure from typical free recall dynamics: a lack of increased probability of first
448 recall for end-of-sequence events (Fig. 3A).

449 Thus, analyses such as those in Figure 3 that address only the temporal dynamics of free re-
450 call paint an incomplete picture of memory for naturalistic episodes. While useful for studying
451 presentation order-dependent recall dynamics, they neglect to consider the stimuli’s content (or,
452 for example, that content’s potential interrelatedness). However, sensitivity to stimulus and recall
453 content introduces a new challenge: distinguishing between levels of recall quality for a stimulus
454 (i.e., an event) that is considered to have been “remembered.” When modeling memory experi-
455 ments, often times events (or items) and their later memories are treated as binary and independent
456 events (e.g., a given list item was simply either remembered or not remembered). Various models
457 of memory (e.g., Yonelinas, 2002) attempt to improve upon this by including confidence ratings,
458 rendering this binary judgement instead categorical. Our novel framework allows one to assess
459 memory performance in a more continuous way (*precision*), as well as analyze the correlational
460 structure of each encoding event to each memory event (*distinctiveness*). Further and importantly,
461 these two novel metrics we introduce here arise from comparisons of the actual content of the
462 experience/memories, which is not typically modeled. Leveraging this, we find that the successful
463 memory performance is related to 1) the precision with which the participant recounts each event
464 and 2) the distinctiveness of each recall event (relative to the other recalled events). The first finding
465 suggests that the information retained for *any individual event* may predict the overall amount of

466 information retained by the participant. The second finding suggests that the ability to distin-
467 guish between temporally or semantically similar content is also related to the quantity of content
468 recovered. Intriguingly, prior studies show that pattern separation, or the ability to discriminate
469 between similar experiences, is impaired in many cognitive disorders as well as natural aging
470 (Stark et al., 2010; Yassa et al., 2011; Yassa and Stark, 2011). Future work might explore whether
471 and how these metrics compare between cognitively impoverished groups and healthy controls.

472 While a large number of language models exist (e.g., WAS, LSA, word2vec, universal sentence
473 encoder; Steyvers et al., 2004; Landauer et al., 1998; Mikolov et al., 2013; Cer et al., 2018), here
474 we use latent dirichlet allocation (LDA)-based topic models for a few reasons. First, topic models
475 capture the *essence* of a text passage devoid of the specific set and order of words used. This
476 was an important feature of our model since different people may accurately recall a scene using
477 very different language. Second, words can mean different things in different contexts (e.g. “bat”
478 as the act of hitting a baseball, the object used for that action, or as a flying mammal). Topic
479 models are robust to this, allowing words to exist as part of multiple topics. Last, topic models
480 provide a straightforward means to recover the weights for the particular words comprising a topic,
481 enabling easy interpretation of an event’s contents (e.g. Fig. 8). Other models such as Google’s
482 universal sentence encoder offer a context-sensitive encoding of text passages, but the encoding
483 space is complex and non-linear, and thus recovering the original words used to fit the model is
484 not straightforward. However, it’s worth pointing out that our framework is divorced from the
485 particular choice of language model. Moreover, many of the aspects of our framework could be
486 swapped out for other choices. For example, the language model, the timeseries segmentation
487 model and the video-recall matching function could all be customized for the particular problem.
488 Indeed for some problems, recovery of the particular recall words may not be necessary, and thus
489 other text-modeling approaches (such as universal sentence encoder) may be preferable. Future
490 work will explore the influence of particular model choices on the framework’s accuracy.

491 Our work has broad implications for how we characterize and assess memory in real-world
492 settings, such as the classroom or physician’s office. For example, the most commonly used
493 classroom evaluation tools involve simply computing the proportion of correctly answered exam

494 questions. Our work indicates that this approach is only loosely related to what educators might
495 really want to measure: how well did the students understand the key ideas presented in the
496 course? Under this typical framework of assessment, the same exam score of 50% could be
497 ascribed to two very different students: one who attended the full course but struggled to learn
498 more than a broad overview of the material, and one who attended only half of the course but
499 understood the material perfectly. Instead, one could apply our computational framework to build
500 explicit content models of the course material and exam questions. This approach would provide
501 a more nuanced and specific view into which aspects of the material students had learned well
502 (or poorly). In clinical settings, memory measures that incorporate such explicit content models
503 might also provide more direct evaluations of patients' memories.

504 **Methods**

505 **Experimental design and data collection**

506 Data were collected by Chen et al. (2017). In brief, participants ($n = 22$) viewed the first 48 minutes
507 of "A Study in Pink", the first episode of the BBC television series *Sherlock*, while fMRI volumes
508 were collected (TR = 1500 ms). Participants were pre-screened to ensure they had never seen any
509 episode of the show before. The stimulus was divided into a 23 min (946 TR) and a 25 min (1030 TR)
510 segment to mitigate technical issues related to the scanner. After finishing the clip, participants
511 were instructed to (quoting from Chen et al., 2017) "describe what they recalled of the [episode]
512 in as much detail as they could, to try to recount events in the original order they were viewed
513 in, and to speak for at least 10 minutes if possible but that longer was better. They were told that
514 completeness and detail were more important than temporal order, and that if at any point they
515 realized they had missed something, to return to it. Participants were then allowed to speak for
516 as long as they wished, and verbally indicated when they were finished (e.g., 'I'm done')." Five
517 participants were dropped from the original dataset due to excessive head motion (2 participants),
518 insufficient recall length (2 participants), or falling asleep during stimulus viewing (1 participant),

519 resulting in a final sample size of $n = 17$. For additional details about the experimental procedure
520 and scanning parameters, see Chen et al. (2017). The experimental protocol was approved by
521 Princeton University's Institutional Review Board.

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

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

536 The video was also divided by an independent coder into 50 more broad “scenes” “following
537 major shifts in the narrative (e.g., location, topic, and/or time)” (Chen et al., 2017). The hand-
538 annotated memory scores for each participant we reference in our present study were generated
539 by considering a scene to have been recalled (in a binary fashion) “if the participant described any
540 part of the scene.”

541 **Data and code availability**

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

544 **Statistics**

545 All statistical tests performed in the behavioral analyses were two-sided. All statistical tests per-
546 formed in the neural data analyses were two-sided, except for the permutation-based thresholding,
547 which was one-sided. In this case, we were specifically interested in identifying voxels whose ac-
548 tivation time series reflected the temporal structure of the video and recall trajectories to a *greater*
549 extent than that of the phase-shifted trajectories.

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

551 **Topic modeling**

552 The input to the topic model we trained to characterize the dynamic content of the video comprised
553 998 hand-generated annotations of short (mean: 2.96s) scenes spanning the video clip (Chen et al.,
554 2017 generated 1000 annotations total; we removed two referring to the break between the first and
555 second scan sessions, during which no fMRI data was collected). The features annotated included:
556 narrative details (a sentence or two describing what happened in that scene); whether the scene
557 took place indoors or outdoors; names of any characters that appeared in the scene; name(s) of
558 characters in camera focus; name(s) of characters who were speaking in the scene; the location (in
559 the story) that the scene took place; camera angle (close up, medium, long, top, tracking, over the
560 shoulder, etc.); whether music was playing in the scene or not; and a transcription of any on-screen
561 text. We concatenated the text for all of these features within each segment, creating a “bag of
562 words” describing each scene. We then re-organized the text descriptions into overlapping sliding
563 windows of 50 scenes each. In other words, we created a “context” for each scene comprising the
564 text descriptions of the preceding 25 scenes, the present scene, and the following 24 scenes. To
565 model the “context” at the beginning and end of the video (i.e., within 25 scenes of the beginning or
566 end), we created overlapping sliding windows that grew in size from one scene to the full length,
567 then similarly tapered their length at the end. This bore the additional benefit of representing each
568 scene’s description in the text corpus an equal number of times.

569 We trained our model using these overlapping text samples with `scikit-learn` (version 0.19.1;

570 Pedregosa et al., 2011), called from our high-dimensional visualization and text analysis software,
571 `HyperTools` (Heusser et al., 2018b). Specifically, we used the `CountVectorizer` class to transform
572 the text from each window into a vector of word counts (using the union of all words across all
573 scenes as the “vocabulary,” excluding English stop words); this yielded a number-of-windows
574 by number-of-words *word count* matrix. We then used the `LatentDirichletAllocation` class
575 (`topics=100, method='batch'`) to fit a topic model (Blei et al., 2003) to the word count matrix,
576 yielding a number-of-windows (1047) by number-of-topics (100) *topic proportions* matrix. The
577 topic proportions matrix describes which mix of topics (latent themes) is present in and around
578 each scene. Next, we transformed the topic proportions matrix to match the 1976 fMRI volume
579 acquisition times. We assigned each topic vector to the timepoint midway between the beginning
580 of the first scene and the end of the last scene in its corresponding sliding text window. We
581 then transformed these timepoints to units of TRs and interpolated the dynamic topic proportions
582 matrix to obtain number-of-TRs (1976) by number-of-topics (100) matrix.

583 We created similar topic proportions matrices using hand-annotated transcripts of each partici-
584 pant’s recall of the video (annotated by Chen et al., 2017). We tokenized the transcript into a list of
585 sentences, and then re-organized the list into overlapping sliding windows spanning 10 sentences
586 each (and analogously tapered the lengths of the first and last 10 sliding windows). In turn, we
587 transformed each window’s sentences into a word count vector (using the same vocabulary as for
588 the video model). We then used the topic model already trained on the video scenes to compute
589 the most probable topic proportions for each sliding window. This yielded a number-of-windows
590 (range: 83–312) by number-of-topics (100) topic proportions matrix for each participant. These
591 reflected the dynamic content of each participant’s recalls. Note: for details on how we selected the
592 video and recall window lengths and number of topics, see *Supporting Information* and Figure S1.

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

594 We parsed the topic trajectories of the video and participants’ recalls into events using Hidden
595 Markov Models (Rabiner, 1989). Given the topic proportions matrix (describing the mix of topics
596 at each timepoint) and a number of states, K , an HMM recovers the set of state transitions that

597 segments the timeseries into K discrete states. Following Baldassano et al. (2017), we imposed an
598 additional set of constraints on the discovered state transitions that ensured that each state was
599 encountered exactly once (i.e., never repeated). We used the BrainIAK toolbox (Capota et al., 2017)
600 to implement this segmentation.

601 We used an optimization procedure to select the appropriate K for each topic proportions
602 matrix. Prior studies on narrative structure and processing have shown that we both perceive
603 and internally represent the world around us at multiple, hierarchical timescales (e.g., Hasson
604 et al., 2008; Lerner et al., 2011; Hasson et al., 2015; Chen et al., 2017; Baldassano et al., 2017, 2018).
605 However, for the purposes of our framework, we sought to identify the single timescale of event-
606 representations that is emphasized *most heavily* in the temporal structure of the video and each
607 participant's recalls. We quantified this as the set of K event boundaries that yielded the maximal
608 distinctiveness between the content (i.e., topics) within each event and that in all other events.
609 Specifically, we computed (for each matrix)

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

610 where a was the distribution of correlations between the topic vectors of timepoints within the
611 same state and b was the average correlation between the topic vectors of timepoints within
612 *different* states. For each possible K , we computed the first Wasserstein distance (W_1 ; also known as
613 "earth mover's distance"; Dobrushin, 1970; Ramdas et al., 2017) between these distributions, and
614 chose the K -value that yielded the greatest difference. Figure 2B displays the event boundaries
615 returned for the video, and Figure S4 displays the event boundaries returned for each participant's
616 recalls (See Fig. S6 for the optimization functions for the video and recalls). After obtaining these
617 event boundaries, we created stable estimates of each topic proportions matrix by averaging the
618 topic vectors within each event. This yielded a number-of-events by number-of-topics matrix for
619 the video and recalls from each participant.

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

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

627 Perhaps the simplest and most widely used measure of memory performance is *accuracy*—i.e.,
628 the proportion of studied (experienced) items (in this case, the 30 video events) that the participant
629 later remembered. Chen et al. (2017) developed a human rating system whereby the quality of
630 each participant’s memory was evaluated by an independent rater. We found a strong across-
631 participants correlation between these independant ratings and the overall number of events that
632 our HMM approach identified in participants’ recalls (Pearson’s $r(15) = 0.65, p = 0.004$).

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

636 **Probability of first recall (PFR).** PFR curves (Welch and Burnett, 1924; Postman and Phillips,
637 1965; Atkinson and Shiffrin, 1968) reflect the probability that an item will be recalled first as a
638 function of its serial position during encoding. To carry out this analysis, we initialized a number-
639 of-participants (17) by number-of-video-events (30) matrix of zeros. Then for each participant, we
640 found the index of the video event that was recalled first (i.e., the video event whose topic vector
641 was most strongly correlated with that of the first recall event) and filled in that index in the matrix
642 with a 1. Finally, we averaged over the rows of the matrix, resulting in a 1 by 30 array representing
643 the proportion of participants that recalled an event first, as a function of the order of the event’s
644 appearance in the video (Fig. 3A).

645 **Lag conditional probability curve (lag-CRP).** The lag-CRP curve (Kahana, 1996) reflects the
646 probability of recalling a given event after the just-recalled event, as a function of their relative
647 positions (or *lag*). In other words, a lag of 1 indicates that a recalled event came immediately after
648 the previously recalled event in the video, and a lag of -3 indicates that a recalled event came 3
649 events before the previously recalled event. For each recall transition (following the first recall),
650 we computed the lag between the current recall event and the next recall event, normalizing by
651 the total number of possible transitions. This yielded a number-of-participants (17) by number-
652 of-lags (-29 to +29; 61 lags total) matrix. We averaged over the rows of this matrix to obtain a
653 group-averaged lag-CRP curve (Fig. 3B).

654 **Serial position curve (SPC).** SPCs (Murdock, 1962) reflect the proportion of participants that
655 remember each item as a function of the items' serial position during encoding. We initialized
656 a number-of-participants (17) by number-of-video-events (30) matrix of zeros. Then, for each
657 recalled event, for each participant, we found the index of the video event that the recalled event
658 most closely matched (via the correlation between the events' topic vectors) and entered a 1 into
659 that position in the matrix (i.e., for the given participant and event). This resulted in a matrix
660 whose entries indicated whether or not each event was recalled by each participant (depending
661 on whether the corresponding entires were set to one or zero). Finally, we averaged over the rows
662 of the matrix to yield a 1 by 30 array representing the proportion of participants that recalled each
663 event as a function of the order of the event's appearance in the video (Fig. 3C).

664 **Temporal clustering scores.** Temporal clustering describes participants' tendency to organize
665 their recall sequences by the learned items' encoding positions. For instance, if a participant
666 recalled the video events in the exact order they occurred (or in exact reverse order), this would
667 yield a score of 1. If a participant recalled the events in random order, this would yield an expected
668 score of 0.5. For each recall event transition (and separately for each participant), we sorted
669 all not-yet-recalled events according to their absolute lag (i.e., distance away in the video). We
670 then computed the percentile rank of the next event the participant recalled. We averaged these

671 percentile ranks across all of the participant's recalls to obtain a single temporal clustering score
672 for the participant.

673 **Semantic clustering scores.** Semantic clustering describes participants' tendency to recall seman-
674 tically similar presented items together in their recall sequences. Here, we used the topic vectors
675 for each event as a proxy for its semantic content. Thus, the similarity between the semantic
676 content for two events can be computed by correlating their respective topic vectors. For each
677 recall event transition, we sorted all not-yet-recalled events according to how correlated the topic
678 vector of *the closest-matching video event* was to the topic vector of the closest-matching video event
679 to the just-recalled event. We then computed the percentile rank of the observed next recall. We
680 averaged these percentile ranks across all of the participant's recalls to obtain a single semantic
681 clustering score for the participant.

682 **Novel naturalistic memory metrics**

683 **Precision.** We tested whether participants who recalled more events were also more *precise* in
684 their recollections. For each participant, we computed the average correlation between the topic
685 vectors for each recall event and those of its closest-matching video event. This gave a single value
686 per participant representing the average precision across all recalled events. We then Fisher's z-
687 transformed these values and correlated them with both hand-annotated and model-derived (i.e.,
688 k or the number of events recovered by the HMM) memory performance.

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

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

695 where $\bar{c}(\text{event})$ is the average correlation between the given recalled event's topic vector and the
696 topic vectors from all video events *except* the best-matching video event. We then averaged these
697 distinctiveness scores across all of the events recalled by the given participant. As above, we used
698 Fisher's z -transformation before correlating these values with hand-annotated and model derived
699 memory performance scores across-subjects.

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

701 We used the UMAP algorithm (McInnes et al., 2018) to project the 100-dimensional topic space
702 onto a two-dimensional space for visualization (Figs. 7, 8). Importantly, to ensure that all of
703 the trajectories were projected onto the *same* lower dimensional space, we computed the low-
704 dimensional embedding on a “stacked” matrix created by vertically concatenating the events-
705 by-topics topic proportions matrices for the video, across-participants average recalls and all 17
706 individual participants’ recalls. We then divided the rows of the result (a total-number-of-events
707 by two matrix) back into separate matrices for the video topic trajectory and the trajectories for
708 each participant’s recalls (Fig. 7). This general approach for discovering a shared low-dimensional
709 embedding for a collections of high-dimensional observations follows Heusser et al. (2018b). Note:
710 for further details on how we created this low-dimensional embedding space, see *Supporting*
711 *Information*.

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

713 In Figure 7B, we present an analysis aimed at characterizing locations in topic space that dif-
714 ferent participants move through in a consistent way (via their recall topic trajectories). The
715 two-dimensional topic space used in our visualizations (Fig. 7) comprised a 60×60 (arbitrary
716 units) square. We tiled this space with a 50×50 grid of evenly spaced vertices, and defined a

717 circular area centered on each vertex whose radius was two times the distance between adjacent
718 vertices (i.e., 2.4 units). For each vertex, we examined the set of line segments formed by connecting
719 each pair successively recalled events, across all participants, that passed through this circle. We
720 computed the distribution of angles formed by those segments and the x -axis, and used a Rayleigh
721 test to determine whether the distribution of angles was reliably “peaked” (i.e., consistent across
722 all transitions that passed through that local portion of topic space). To create Figure 7B we drew
723 an arrow originating from each grid vertex, pointing in the direction of the average angle formed
724 by line segments that passed within its circular radius. We set the arrow lengths to be inversely
725 proportional to the p -values of the Rayleigh tests at each vertex. Specifically, for each vertex we
726 converted all of the angles of segments that passed within 2.4 units to unit vectors, and we set
727 the arrow lengths at each vertex proportional to the length of the (circular) mean vector. We also
728 indicated any significant results ($p < 0.05$, corrected using the Benjamani-Hochberg procedure) by
729 coloring the arrows in blue (darker blue denotes a lower p -value, i.e., a longer mean vector); all
730 tests with $p \geq 0.05$ are displayed in gray and given a lower opacity value.

731 **Searchlight fMRI analyses**

732 In Figure 9, we present two analyses aimed at identifying brain regions whose responses (as par-
733 ticipants viewed the video) exhibited a particular temporal structure. We developed a searchlight
734 analysis wherein we constructed a cube centered on each voxel (radius: 5 voxels) and for each
735 of these cubes, computed the temporal correlation matrix of the voxel responses during video
736 viewing. Specifically, for each of the 1976 volumes collected during video viewing, we correlated
737 the activity patterns in the given cube with the activity patterns (in the same cube) collected during
738 every other timepoint. This yielded a 1976 by 1976 correlation matrix for each cube.

739 Next, we constructed a series of “template” matrices: the first reflecting the timecourse of
740 video’s topic trajectory, and the others reflecting that of each participant’s recall topic trajectory.
741 To construct the video template, we computed the correlations between the topic proportions
742 estimated for every pair of TRs (prior to segmenting the trajectory into discrete events; i.e., the
743 correlation matrix shown in Figs. 2B and 9A). We constructed similar temporal correlation matrices

744 for each participant's recall topic trajectory (Figs. 2D, S4). However, to correct for length differences
745 and potential non-linear transformations between viewing time and recall time, we first used
746 dynamic time warping (Berndt and Clifford, 1994) to temporally align participants' recall topic
747 trajectories with the video topic trajectory. An example correlation matrix before and after warping
748 is shown in Fig. 9B. This yielded a 1976 by 1976 correlation matrix for the video template and for
749 each participant's recall template.

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

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

767 We used an analogous procedure to identify which voxels' responses reflected the recall tem-
768 plates. For each participant, we correlated the upper triangle of the correlation matrix for each cube
769 of voxels with their (time warped) recall correlation matrix. As in the video template analysis this
770 yielded a voxelwise map of correlation coefficients per participant. However, whereas the video
771 analysis compared every participant's responses to the same template, here the recall templates

772 were unique for each participant. As in the analysis described above, we t -scored the (Fisher
773 z -transformed) voxelwise correlations, and used the same permutation procedure we developed
774 for the video responses to ensure specificity to the recall timeseries and assign significance values.
775 To create the map in Figure 9B we again thresholded out any voxels whose correspondence values
776 fell below the 95th percentile of the permutation-derived null distribution.

777 References

- 778 Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control
779 processes. In Spence, K. W. and Spence, J. T., editors, *The psychology of learning and motivation*,
780 volume 2, pages 89–105. Academic Press, New York.
- 781 Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017).
782 Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–
783 721.
- 784 Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas
785 during narrative perception. *Journal of Neuroscience*, 38(45):9689–9699.
- 786 Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In
787 *KDD workshop*, volume 10, pages 359–370.
- 788 Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International
789 Conference on Machine Learning*, ICML ’06, pages 113–120, New York, NY, US. ACM.
- 790 Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine
791 Learning Research*, 3:993 – 1022.
- 792 Brunec, I. K., Moscovitch, M. M., and Barese, M. D. (2018). Boundaries shape cognitive represen-
793 tations of spaces and events. *Trends in Cognitive Sciences*, 22(7):637–650.
- 794 Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., and Oliva, A. (2015). Intrinsic and extrinsic
795 effects on image memorability. *Vision Research*, 116:165–178.

- 796 Capota, M., Turek, J., Chen, P.-H., Zhu, X., Manning, J. R., Sundaram, N., Keller, B., Wang, Y., and
797 Shin, Y. S. (2017). Brain imaging analysis kit.
- 798 Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes,
799 M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.
800 *arXiv*, 1803.11175.
- 801 Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared
802 memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*,
803 20(1):115.
- 804 Chun, M. and Turk-Browne, N. (2007). Interactions between attention and memory. *Current opinion*
805 in *neurobiology*, 17(2):177–184.
- 806 Clewett, D. and Davachi, L. (2017). The ebb and flow of experience determines the temporal
807 structure of memory. *Curr Opin Behav Sci*, 17:186–193.
- 808 Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in*
809 *Neurobiology*, 16(6):693—700.
- 810 Davachi, L., Mitchell, J. P., and Wagner, A. D. (2003). Multiple routes to memory: distinct medial
811 temporal lobe processes build item and source memories. *Proceedings of the National Academy of*
812 *Sciences, USA*, 100(4):2157 – 2162.
- 813 Diana, R. A., Yonelinas, A. P., and Ranganath, C. (2007). Imaging recollection and famil-
814 iarity in the medial temporal lobe: a three-component model. *Trends in Cognitive Sciences*,
815 doi:10.1016/j.tics.2007.08.001.
- 816 Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions.
817 *Theory of Probability & Its Applications*, 15(3):458–486.
- 818 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the
819 sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–1286.

- 820 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological*
821 *Science*, 22(2):243–252.
- 822 Friendly, M., Franklin, P. E., Hoffman, D., and Rubin, D. C. (1982). The Toronto Word Pool:
823 Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080
824 words. *Behavior Research Methods and Instrumentation*, 14:375–399.
- 825 Gilboa, A. and Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory.
826 *Trends Cogn Sci*, 21(8):618–631.
- 827 Hasson, U., Chen, J., and Honey, C. J. (2015). Hierarchical process memory: memory as an integral
828 component of information processing. *Trends in Cognitive Science*, 19(6):304–315.
- 829 Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal
830 receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550.
- 831 Heusser, A. C., Ezzyat, Y., Shiff, I., and Davachi, L. (2018a). Perceptual boundaries cause mnemonic
832 trade-offs between local boundary processing and across-trial associative binding. *Journal of*
833 *Experimental Psychology Learning, Memory, and Cognition*, 44(7):1075–1090.
- 834 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail: a
835 Python toolbox for analyzing and plotting free recall data. *The Journal of Open Source Software*,
836 10.21105/joss.00424.
- 837 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018b). HyperTools: a Python
838 toolbox for gaining geometric insights into high-dimensional data. *Journal of Machine Learning*
839 *Research*, 18(152):1–6.
- 840 Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal*
841 *of Mathematical Psychology*, 46:269–299.
- 842 Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., and H., E.
843 (2014). A unified mathematical framework for coding time, space, and sequences in the medial
844 temporal lobe. *Journal of Neuroscience*, 34(13):4692–4707.

- 845 Howard, M. W., Viskontas, I. V., Shankar, K. H., and Fried, I. (2012). Ensembles of human MTL
846 neurons “jump back in time” in response to a repeated stimulus. *Hippocampus*, 22:1833–1847.
- 847 Huk, A., Bonnen, K., and He, B. J. (2018). Beyond trial-based paradigms: continuous behavior, on-
848 going neural activity, and naturalistic stimuli. *Journal of Neuroscience*, 10.1523/JNEUROSCI.1920-
849 17.2018.
- 850 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24:103–109.
- 851 Koriat, A. and Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: distin-
852 guishing accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of*
853 *Experimental Psychology: General*, 123(3):297–315.
- 854 Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – con-
855 nnecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:1 – 28.
- 856 Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to latent semantic analysis.
857 *Discourse Processes*, 25:259–284.
- 858 Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy
859 of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.
- 860 Manning, J. R. (2019). Episodic memory: mental time travel or a quantum ‘memory wave’ function?
861 *PsyArXiv*, doi:10.31234/osf.io/6zjwb.
- 862 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory.
863 In Gazzaniga, M., editor, *The Cognitive Neurosciences, Fifth edition*, pages 557–566. MIT Press.
- 864 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns
865 in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National*
866 *Academy of Sciences, USA*, 108(31):12893–12897.
- 867 McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and
868 projection for dimension reduction. *arXiv*, 1802(03426).

- 869 Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations
870 in vector space. *arXiv*, 1301.3781.
- 871 Mueller, A., Fillion-Robin, J.-C., Boidol, R., Tian, F., Nechifor, P., yoonsubKim, Peter, Rampin, R.,
872 Corvellec, M., Medina, J., Dai, Y., Petrushev, B., Langner, K. M., Hong, Alessio, Ozsvald, I.,
873 vkolmakov, Jones, T., Bailey, E., Rho, V., IgorAPM, Roy, D., May, C., foobuzz, Piyush, Seong,
874 L. K., Goey, J. V., Smith, J. S., Gus, and Mai, F. (2018). WordCloud 1.5.0: a little word cloud
875 generator in Python. *Zenodo*, <https://zenodo.org/record/1322068#.W4tPKZNKh24>.
- 876 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*,
877 64:482–488.
- 878 Paller, K. A. and Wagner, A. D. (2002). Observing the transformation of experience into memory.
879 *Trends in Cognitive Sciences*, 6(2):93–102.
- 880 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten-
881 hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot,
882 M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine
883 Learning Research*, 12:2825–2830.
- 884 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model
885 of organizational processes in free recall. *Psychological Review*, 116(1):129–156.
- 886 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal
887 of Experimental Psychology*, 17:132–138.
- 888 Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech
889 recognition. *Proceedings of the IEEE*, 77(2):257–286.
- 890 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition. *Curr Opin
891 Behav Sci*, 17:133–140.
- 892 Ramdas, A., Trillos, N., and Cuturi, M. (2017). On wasserstein two-sample testing and related
893 families of nonparametric tests. *Entropy*, 19(2):47.

- 894 Ranganath, C., Cohen, M. X., Dam, C., and D'Esposito, M. (2004). Inferior temporal, prefrontal,
895 and hippocampal contributions to visual working memory maintenance and associative memory
896 retrieval. *Journal of Neuroscience*, 24(16):3917–3925.
- 897 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature
898 Reviews Neuroscience*, 13:713 – 726.
- 899 Schlichting, M. L. and Preston, A. R. (2015). Memory integration: neural mechanisms and impli-
900 cations for behavior. *Current Opinion in Behavioral Sciences*, 1:1–8.
- 901 Simony, E., Honey, C. J., Chen, J., and Hasson, U. (2016). Uncovering stimulus-locked network
902 dynamics during narrative comprehension. *Nature Communications*, 7(12141):1–13.
- 903 Spalding, K. N., Schlichting, M. L., Zeithamova, D., Preston, A. R., Tranel, D., Duff, M. C., and
904 Warren, D. E. (2018). Ventromedial prefrontal cortex is necessary for normal associative inference
905 and memory integration. *The Journal of Neuroscience*, 38(15):3767–3775.
- 906 Stark, S. M., Yassa, M. A., and Stark, C. E. L. (2010). Individual differences in spatial pattern
907 separation performance associated with healthy aging in humans. *Learning & Memory*, 17(6):284–
908 288.
- 909 Steyvers, M., Shiffrin, R. M., and Nelson, D. L. (2004). Word association spaces for predicting
910 semantic similarity effects in episodic memory. In Healy, A. F., editor, *Cognitive Psychology and
911 its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*. American
912 Psychological Association, Washington, DC.
- 913 Tompry, A. and Davachi, L. (2017). Consolidation promotes the emergence of representational
914 overlap in the hippocampus and medial prefrontal cortex. *Neuron*, 96(1):228–241.
- 915 Tyng, C. M., Amin, H. U., Saad, M. N. M., and S, M. A. (2017). The influences of emotion on
916 learning and memory. *Frontiers in psychology*, 8:1454.
- 917 van Kesteren, M. T. R., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., and Fernández, G.

- 918 (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent
919 encoding: from congruent to incongruent. *Neuropsychologia*, 51(12):2352–2359.
- 920 van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., and Henson, R. N. (2012). How schema and
921 novelty augment memory formation. *Trends Neurosci*, 35(4):211–9.
- 922 Waskom, M., Botvinnik, O., Okane, D., Hobson, P., David, Halchenko, Y., Lukauskas, S., Cole, J. B.,
923 Warmenhoven, J., de Ruiter, J., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E.,
924 Martin, M., Miles, A., Meyer, K., Augspurger, T., Yarkoni, T., Bachant, P., Williams, M., Evans,
925 C., Fitzgerald, C., Brian, Wehner, D., Hitz, G., Ziegler, E., Qalieh, A., and Lee, A. (2016). Seaborn:
926 v0.7.1.
- 927 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American Journal
928 of Psychology*, 35:396–401.
- 929 Wiltgen, B. J. and Silva, A. J. (2007). Memory for context becomes less specific with time. *Learning
930 & Memory*, 14(4):313–317.
- 931 Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale
932 automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665.
- 933 Yassa, M. A., Lacy, J. W., Stark, S. M., Albert, M. S., Gallagher, M., and Stark, C. E. L. (2011). Pattern
934 separation deficits associated with increased hippocampal ca3 and dentate gyrus activity in
935 nondemented older adults. *Hippocampus*, 21(9):968–979.
- 936 Yassa, M. A. and Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends In Neuro-
937 sciences*, 34(10):515–525.
- 938 Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research.
939 *Journal of Memory and Language*, 46:441–517.
- 940 Yonelinas, A. P., Kroll, N. E., Quamme, J. R., Lazzara, M. M., Sauvé, M. J., Widaman, K. F., and
941 Knight, R. T. (2002). Effects of extensive temporal lobe damage or mild hypoxia on recollection
942 and familiarity. *Nature Neuroscience*, 5(11):1236–41.

- 943 Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception:
944 a mind-brain perspective. *Psychological Bulletin*, 133:273–293.
- 945 Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit
946 memories to other brains: Constructing shared neural representations via communication. *Cereb*
947 *Cortex*, 27(10):4988–5000.
- 948 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and
949 memory. *Psychological Bulletin*, 123(2):162 – 185.

950 **Supporting information**

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

952 **Acknowledgements**

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

958 **Author contributions**

959 Conceptualization: A.C.H. and J.R.M.; Methodology: A.C.H., P.C.F. and J.R.M.; Software: A.C.H.,
960 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.,
961 P.C.F. and J.R.M.; Supervision: J.R.M.

⁹⁶² **Author information**

⁹⁶³ The authors declare no competing financial interests. Correspondence and requests for materials
⁹⁶⁴ should be addressed to J.R.M. (jeremy.r.manning@dartmouth.edu).