

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 December 11, 2019

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 Although our experiences most often change gradually, they also occasionally change sud-
57 denly (e.g., when we walk through a doorway; Radvansky and Zacks, 2017). Prior research
58 suggests that these sharp transitions (termed *event boundaries*) during an experience help to dis-
59 cretize our experiences (and their mental representations) into *events* (Radvansky and Zacks, 2017;
60 Brunec et al., 2018; Heusser et al., 2018a; Clewett and Davachi, 2017; Ezzyat and Davachi, 2011;
61 DuBrow and Davachi, 2013). The interplay between the stable (within event) and transient (across
62 event) temporal dynamics of an experience also provides a potential framework for transforming
63 experiences into memories that distill those experiences down to their essence. For example, prior
64 work has shown that event boundaries can influence how we learn sequences of items (Heusser
65 et al., 2018a; DuBrow and Davachi, 2013), navigate (Brunec et al., 2018), and remember and un-
66 derstand narratives (Zwaan and Radvansky, 1998; Ezzyat and Davachi, 2011). Prior research has
67 implicated the hippocampus and the medial prefrontal cortex as playing a critical role in trans-
68 forming experiences into structured 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 *trajectories* that describe how the experiences evolve over
79 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 the
82 episode then reveals which aspects of the episode were preserved (or lost) in the translation into
83 memory. We further examine whether 1) the *precision* with which a participant recounts each event
84 and 2) the *distinctiveness* each recall event is (relative to the other recalled events) relates to their
85 overall memory performance. Last, we identify networks of brain structures whose responses
86 (as participants watched the episode) reflected the temporal dynamics of the episode, and how
87 participants would later recount the episode.

88 Results

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

98 document.

99 Chen et al. (2017) collected hand-annotated information about each of 1000 (manually identified)
100 scenes spanning the roughly 50 minute video used in their experiment. This information included:
101 a brief narrative description of what was happening; whether the scene took place indoors or
102 outdoors; the names of any characters on the screen; the names of any characters who were in
103 focus in the camera shot; the names of characters who were speaking; the location where the scene
104 took place; the camera angle (close up, medium, long, etc.); whether or not background music was
105 present; and other similar details (for a full list of annotated features see *Methods*). We took from
106 these annotations the union of all unique words (excluding stop words, such as “and,” “or,” “but,”
107 etc.) across all features and scenes as the “vocabulary” for the topic model. We then concatenated
108 the sets of words across all features contained in overlapping, 50-scene sliding windows, and
109 treated each 50-scene sequence as a single “document” for the purpose of fitting the topic model.
110 Next, we fit a topic model with (up to) $K = 100$ topics to this collection of documents. We found that
111 32 unique topics (with non-zero weights) were sufficient to describe the time-varying content of the
112 video (see *Methods*; Figs. 1, S2). Note that our approach is similar in some respects to Dynamic Topic
113 Models (Blei and Lafferty, 2006) in that we sought to characterize how the thematic content of the
114 episode evolved over time. However, whereas Dynamic Topic Models are designed to characterize
115 how the properties of *collections* of documents change over time, our sliding window approach
116 allows us to examine the topic dynamics within a single document (or video). Specifically, our
117 approach yielded (via the topic proportions matrix) a single *topic vector* for each timepoint of the
118 episode (we set timepoints to match the acquisition times of the 1976 fMRI volumes collected as
119 participants viewed the episode).

120 The topics we found were heavily character-focused (e.g., the top-weighted word in each topic
121 was nearly always a character) and could be roughly divided into themes that were primarily
122 Sherlock Holmes-focused (Sherlock is the titular character), primarily John Watson-focused (John
123 is Sherlock’s close confidant and assistant), or focused on Sherlock and John interacting (Fig. S2).
124 Several of the topics were highly similar, which we hypothesized might allow us to distinguish
125 between subtle narrative differences (if the distinctions between those overlapping topics were



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 meaningful; also see Fig. S3). The topic vectors for each timepoint were *sparse*, in that only a small
127 number (usually one or two) of topics tended to be “active” in any given timepoint (Fig. 2A).
128 Further, the dynamics of the topic activations appeared to exhibit *persistance* (i.e., given that a
129 topic was active in one timepoint, it was likely to be active in the following timepoint) along with
130 *occasional rapid changes* (i.e., occasionally topics would appear to spring into or out of existence).
131 These two properties of the topic dynamics may be seen in the block diagonal structure of the
132 timepoint-by-timepoint correlation matrix (Fig. 2B) and reflect the gradual drift and sudden shifts
133 fundamental to the contextual dynamics of real-world experiences. Given this observation, we
134 adapted an approach devised by Baldassano et al. (2017), and used a Hidden Markov Model (HMM)
135 to identify the *event boundaries* where the topic activations changed rapidly (i.e., at the boundaries
136 of the blocks in the correlation matrix; event boundaries identified by the HMM are outlined in
137 yellow). Part of our model fitting procedure required selecting an appropriate number of “events”
138 to segment the timeseries into. We used an optimization procedure to identify the number of
139 events that maximized within-event stability while also minimizing across-event correlations (see
140 *Methods* for additional details). To create a stable “summary” of the video, we computed the
141 average topic vector within each event (Fig. 2C).

142 Given that the time-varying content of the video could be segmented cleanly into discrete
143 events, we wondered whether participants’ recalls of the video also displayed a similar structure.
144 We applied the same topic model (already trained on the video annotations) to each participant’s
145 recalls. Analogous to how we analyzed the time-varying content of the video, to obtain similar
146 estimates for participants’ recalls, we treated each (overlapping) 10-sentence “window” of their
147 transcript as a “document” and then computed the most probable mix of topics reflected in each
148 timepoint’s sentences. This yielded, for each participant, a number-of-windows by number-of-
149 topics topic proportions matrix that characterized how the topics identified in the original video
150 were reflected in the participant’s recalls. Note that an important feature of our approach is
151 that it allows us to compare participant’s recalls to events from the original video, despite that
152 different participants may have used different language to describe the same event, and that those
153 descriptions may not match the original annotations. This is a substantial benefit of projecting

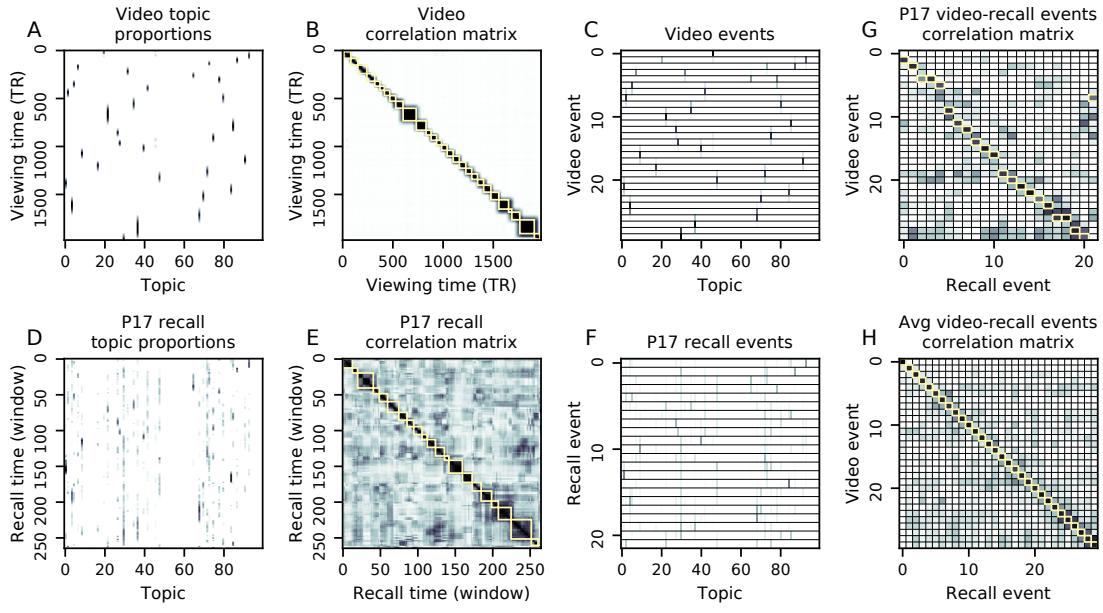


Figure 2: Modelling naturalistic stimuli and recalls. All panels: darker colors indicate greater values; range: [0, 1]. **A.** Topic vectors ($K = 100$) for each of the 1976 video timepoints. **B.** Timepoint-by-timepoint correlation matrix of the topic vectors displayed in Panel A. Event boundaries detected 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). **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 the video and recalls into a shared “topic” space. An example topic proportions matrix from one
¹⁵⁵ participant’s recalls is shown in Figure 2D.

¹⁵⁶ Although the example participant’s recall topic proportions matrix has some visual similarity to
¹⁵⁷ the video topic proportions matrix, the time-varying topic proportions for the example participant’s
¹⁵⁸ recalls are not as sparse as for the video (e.g., compare Figs. 2A and D). Similarly, although there
¹⁵⁹ do appear to be periods of stability in the recall topic dynamics (e.g., most topics are active or
¹⁶⁰ inactive over contiguous blocks of time), the overall timecourses are not as cleanly delineated as
¹⁶¹ the video topics are. To examine these patterns in detail, we computed the timepoint-by-timepoint
¹⁶² correlation matrix for the example participant’s recall topic proportions (Fig. 2E). As in the video
¹⁶³ correlation matrix (Fig. 2B), the example participant’s recall correlation matrix has a strong block
¹⁶⁴ diagonal structure, indicating that their recalls are discretized into separated events. As for the
¹⁶⁵ video correlation matrix, we can use an HMM, along with the aforementioned number-of-events
¹⁶⁶ optimization procedure (also see *Methods*) to determine how many events are reflected in the
¹⁶⁷ participant’s recalls and where specifically the event boundaries fall (outlined in yellow). We
¹⁶⁸ carried out a similar analysis on all 17 participants’ recall topic proportions matrices (Fig. S4).

¹⁶⁹ Two clear patterns emerged from this set of analyses. First, although every individual partic-
¹⁷⁰ ipant’s recalls could be segmented into discrete events (i.e., every individual participant’s recall
¹⁷¹ correlation matrix exhibited clear block diagonal structure; Fig. S4), each participant appeared to
¹⁷² have a unique *recall resolution*, reflected in the sizes of those blocks. For example, some participants’
¹⁷³ recall topic proportions segmented into just a few events (e.g., Participants P4, P5, and P7), while
¹⁷⁴ others’ recalls segmented into many shorter duration events (e.g., Participants P12, P13, and P17).
¹⁷⁵ This suggests that different participants may be recalling the video with different levels of detail-
¹⁷⁶ e.g., some might touch on just the major plot points, whereas others might attempt to recall every
¹⁷⁷ minor scene or action. The second clear pattern present in every individual participant’s recall
¹⁷⁸ correlation matrix is that, unlike in the video correlation matrix, there are substantial off-diagonal
¹⁷⁹ correlations. Whereas each event in the original video was (largely) separable from the others
¹⁸⁰ (Fig. 2B), in transforming those separable events into memory, participants appear to be integrat-
¹⁸¹ ing across multiple events, blending elements of previously recalled and not-yet-recalled events

182 into each newly recalled event (Figs. 2D, S4; also see Manning et al., 2011; Howard et al., 2012).

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

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

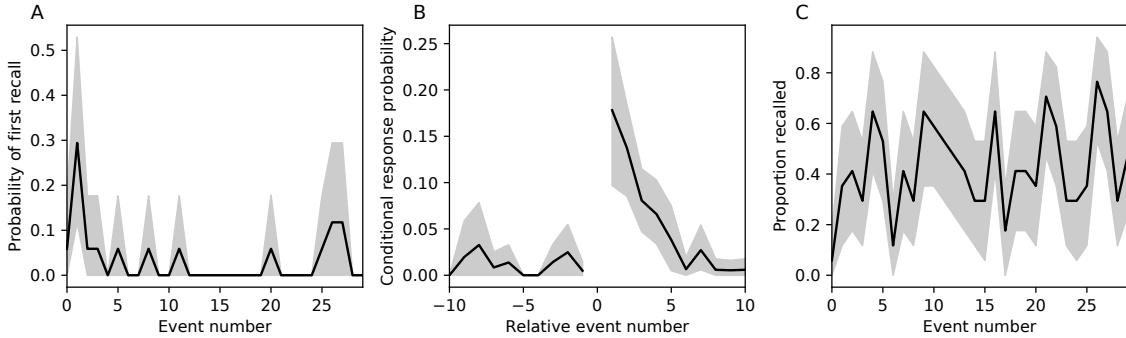


Figure 3: Naturalistic extensions of classic list-learning memory analyses. A. The probability of first recall as a function of the serial position of the event in the video. B. The probability of recalling each event, conditioned on having most recently recalled the event *lag* events away in the video.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

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 memory recalls as binary (e.g. the item was recalled or not) and independent events. However, our framework produces a content-based model of individual stimulus and recall events, allowing for direct quantitative comparison between all stimulus and recall events, as well as between the recall events themselves. Leveraging these content-based models of the stimulus/recall events, we developed two novel metrics for quantifying naturalistic memory representations: *precision* and *distinctiveness*. We define precision as the average correlation between the topic proportions of each recall event and the maximally correlated video event (Fig. 4). Participants whose recall events are more veridical descriptions of what happened in the video event will presumably have higher precision scores. We find that, across participants, a higher precision score is correlated to both hand-annotated memory performance

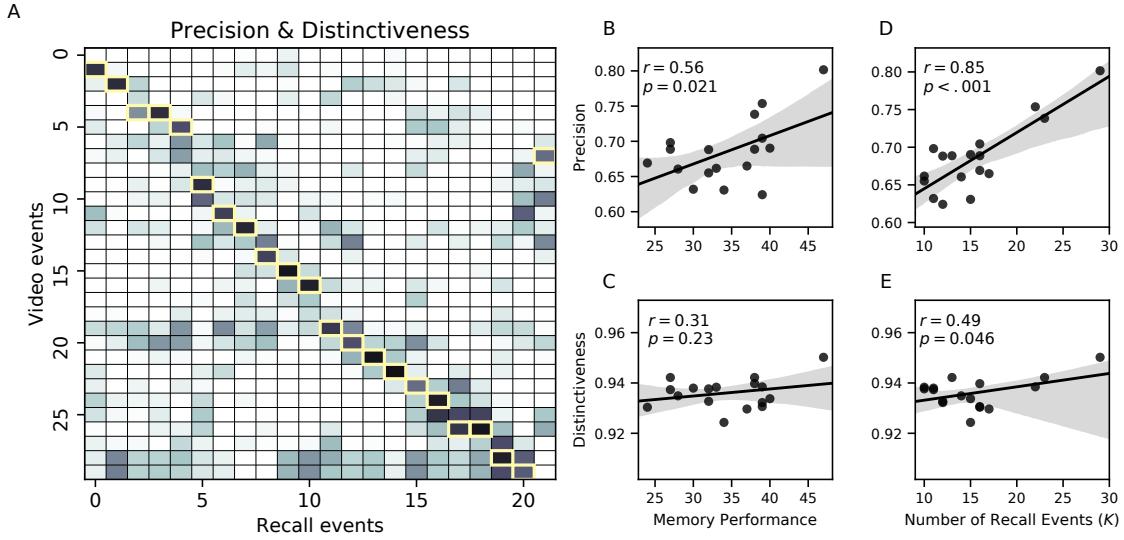


Figure 4: Novel content-based metrics of naturalistic memory: precision and distinctiveness. **A.** A video-recall correlation matrix for a representative participant (17). The yellow boxes highlight the maximum correlation in each column. Precision was computed as the average of the maximum correlation in each column. On the other hand, distinctiveness was defined as the average of everything except for the maximum correlation in each column. **B.** The (Pearson's) correlation between precision and hand-annotated memory performance. **C.** The correlation between precision and the number of events recovered by the model (k). **D.** The correlation between distinctiveness and hand-annotated memory performance. **E.** The correlation between distinctiveness and the number of events recovered by the model (k).

(Pearson's $r(15) = 0.56, p = 0.021$) and the number of recall events estimated by our model (Pearson's $r(15) = 0.85, p < 0.001$). A second novel metric we introduce here is distinctiveness, or how unique the recall description was to each video event. We define distinctiveness as 1 minus the average of all non-matching recall events from the video-recall correlation matrix. We hypothesized that participants who recounted events in a more distinctive way would display better overall memory. We find that this distinctiveness score is related to our model's estimated number of recalled events (Pearson's $r(15) = 0.49, p = 0.046$) but not to the analogous hand-annotated metric (Pearson's $r(15) = 0.31, p = 0.23$). In summary, using two novel metrics afforded by our approach, we find that participants whose recalls are both more precise and distinct remember more content.

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

238 to gain deep insights into that content solely by examining the topic proportion matrices (e.g.,
239 Figs. 2A, D) or the corresponding correlation matrices (Figs. 2B, E, S4). To visualize the time-
240 varying high-dimensional content in a more intuitive way (Heusser et al., 2018b) we projected the
241 topic proportions matrices onto a two-dimensional space using Uniform Manifold Approximation
242 and Projection (UMAP; McInnes and Healy, 2018). In this lower-dimensional space, each point
243 represents a single video or recall event, and the distances between the points reflect the distances
244 between the events' associated topic vectors (Fig. 5). In other words, events that are near to each
245 other in this space are more semantically similar.

246 Visual inspection of the video and recall topic trajectories reveals a striking pattern. First,
247 the topic trajectory of the video (which reflects its dynamic content; Fig. 5A) is captured nearly
248 perfectly by the averaged topic trajectories of participants' recalls (Fig. 5B). To assess the consistency
249 of these recall trajectories across participants, we asked: given that a participant's recall trajectory
250 had entered a particular location in topic space, could the position of their *next* recalled event
251 be predicted reliably? For each location in topic space, we computed the set of line segments
252 connecting successively recalled events (across all participants) that intersected that location (see
253 *Methods* for additional details). We then computed (for each location) the distribution of angles
254 formed by the lines defined by those line segments and a fixed reference line (the *x*-axis). Rayleigh
255 tests revealed the set of locations in topic space at which these across-participant distributions
256 exhibited reliable peaks (blue arrows in Fig. 5B reflect significant peaks at $p < 0.05$, corrected). We
257 observed that the locations traversed by nearly the entire video trajectory exhibited such peaks.
258 In other words, participants exhibited similar trajectories that also matched the trajectory of the
259 original video (Fig. 5C). This is especially notable when considering the fact that the number of
260 events participants recalled (dots in Fig. 5C) varied considerably across people, and that every
261 participant used different words to describe what they had remembered happening in the video.
262 Differences in the numbers of remembered events appear in participants' trajectories as differences
263 in the sampling resolution along the trajectory. We note that this framework also provides a
264 means of detangling classic "proportion recalled" measures (i.e., the proportion of video events
265 referenced in participants' recalls) from participants' abilities to recapitulate the full shape of the

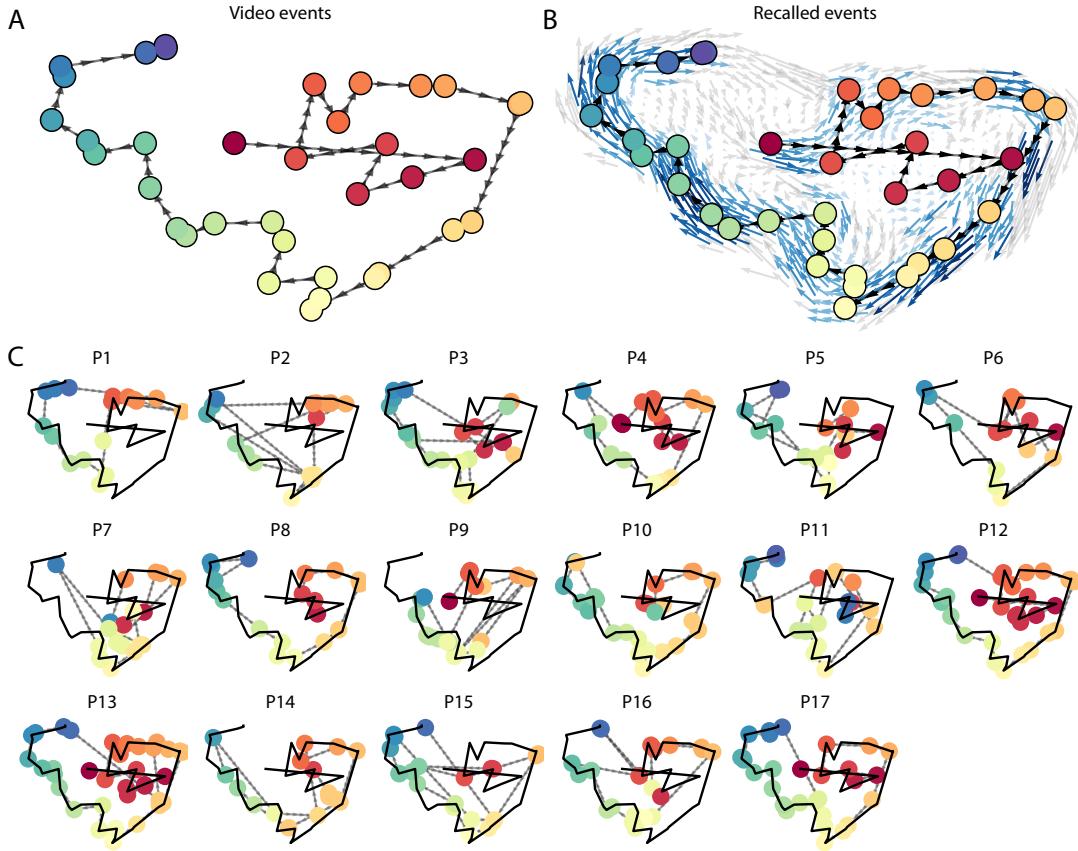


Figure 5: Trajectories through topic space capture the dynamic content of the video and recalls. All panels: the topic proportion matrices have been projected onto a shared two-dimensional space using UMAP. **A.** The two-dimensional topic trajectory taken by the episode of *Sherlock*. Each dot indicates 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. (Same format and coloring as Panel A.)

266 original video (i.e., the similarity in the shape of the original video trajectory and that defined by
267 each participant's recounting of the video).

268 Because our analysis framework projects the dynamic video content and participants' recalls
269 onto a shared topic space, and because the dimensions of that space are known (i.e., each topic
270 dimension is a set of weights over words in the vocabulary; Fig. S2), we can examine the topic
271 trajectories to understand which specific content was remembered well (or poorly). For each video
272 event, we can ask: what was the average correlation (across participants) between the video event's
273 topic vector and the closest matching recall event topic vectors from each participant? This yields a
274 single correlation coefficient for each video event, describing how closely participants' recalls of the
275 event tended to reliably capture its content (Fig. 6A). (We also examined how different comparisons
276 between each video event's topic vector and the corresponding recall event topic vectors related
277 to hand-annotated characterizations of memory performance; see *Supporting Information*). Given
278 this summary of which events were recalled reliably (or not), we next asked whether the better-
279 remembered or worse-remembered events tended to reflect particular topics. We computed a
280 weighted average of the topic vectors for each video event, where the weights reflected how
281 reliably each event was recalled. To visualize the result, we created a "wordle" image (Mueller
282 et al., 2018) where words weighted more heavily by better-remembered topics appear in a larger
283 font (Fig. 6B, green box). Across the full video, content that reflected topics necessary to convey
284 the central focus of the video (e.g., the names of the two main characters, "Sherlock" and "John",
285 and the address of a major recurring location, "221b Baker Street") were best remembered. An
286 analogous analysis revealed which themes were poorly remembered. Here in computing the
287 weighted average over events' topic vectors, we weighted each event in *inverse* proportion to how
288 well it was remembered (Fig. 6B, red box). The least well-remembered video content reflected
289 information not necessary to conveying the video's "gist," such as the names of relatively minor
290 characters (e.g., "Mike," "Jeffrey," "Molly," and "Jimmy") and locations (e.g., "St. Bartholomew's
291 Hospital").

292 A similar result emerged from assessing the topic vectors for individual video and recall events
293 (Fig. 6C). Here, for each of the three best- and worst-remembered video events, we have constructed

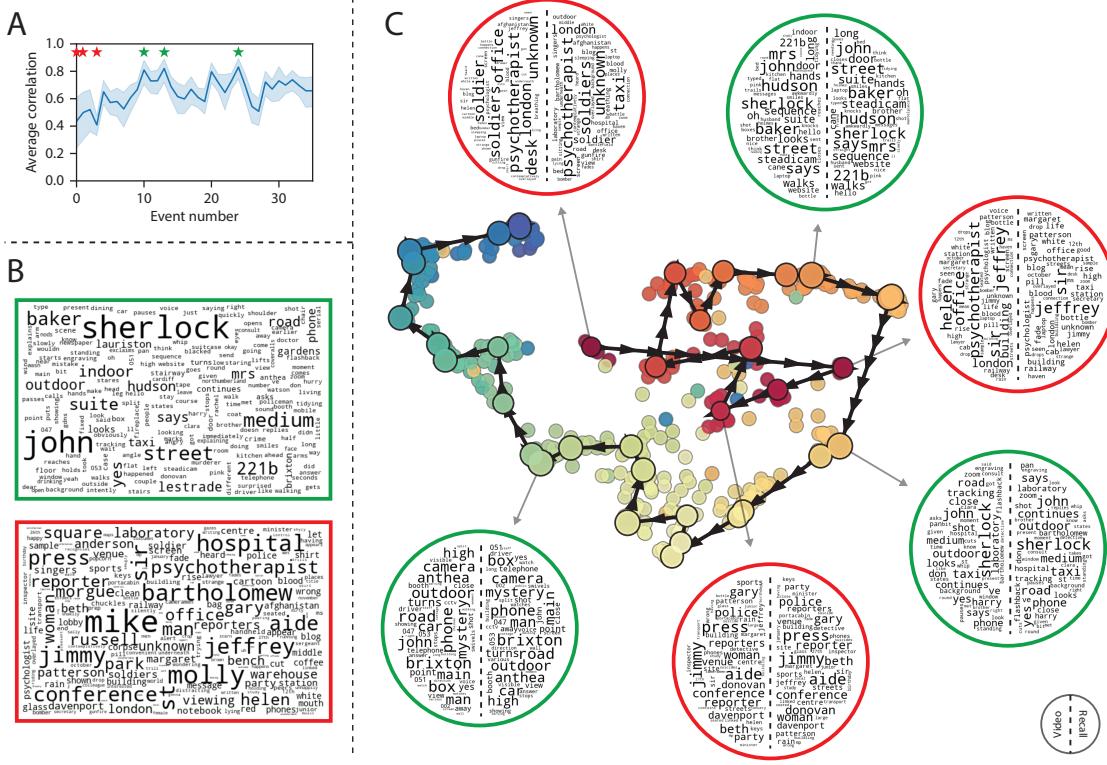


Figure 6: Transforming experience into memory. **A.** Average correlations (across participants) between the topic vectors from each video event and the closest-matching recall events. Error bars denote bootstrap-derived across-participant 95% confidence intervals. The stars denote the three best-remembered 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 5. 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 5A. 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.

294 two wordles: one from the original video event's topic vector (left) and a second from the average
295 recall topic vector for that event (right). The three best-remembered events (circled in green)
296 correspond to scenes important to the central plot-line (Sherlock and John meeting, Sherlock and
297 John chasing the killer, and the killer calling spying on John in a phone booth). Meanwhile, the three
298 worst-remembered events (circled in red) reflect various side-stories (John talking to his therapist
299 about the war, a soon-to-be-victim's affair with his assistant, and another soon-to-be-victim leaving
300 a party with his friend) that are not essentially to summarizing the video's narrative.

301 The results thus far inform us about which aspects of the dynamic content in the episode
302 participants watched were preserved or altered in participants' memories of the episode. We next
303 carried out a series of analyses aimed at understanding which brain structures might implement
304 these processes. In one analysis we sought to identify which brain structures were sensitive
305 to the video's dynamic content, as characterized by its topic trajectory. Specifically, we used a
306 searchlight procedure to identify the extent to which each cluster of voxels exhibited a timecourse
307 of activity (as the participants watched the video) whose temporal correlation matrix matched
308 the temporal correlation matrix of the original video's topic proportions (Fig. 2B). As shown
309 in Figure 7A, the analysis revealed a network of regions including bilateral frontal cortex and
310 cingulate cortex, suggesting that these regions may play a role in processing information relevant
311 to the narrative structure of the video. In a second analysis, we sought to identify which brain
312 structures' responses (while viewing the video) reflected how each participant would later *recall*
313 the video. We used an analogous searchlight procedure to identify clusters of voxels whose
314 temporal correlation matrices reflected the temporal correlation matrix of the topic proportions for
315 each individual's recalls (Figs. 2D, S4). As shown in Figure 7B, the analysis revealed a network of
316 regions including the ventromedial prefrontal cortex (vmPFC), anterior cingulate cortex (ACC), and
317 right medial temporal lobe (rMTL), suggesting that these regions may play a role in transforming
318 each individual's experience into memory. In identifying regions whose responses to ongoing
319 experiences reflect how those experiences will be remembered later, this latter analysis extends
320 classic *subsequent memory analyses* (e.g., Paller and Wagner, 2002) to domain of naturalistic stimuli.

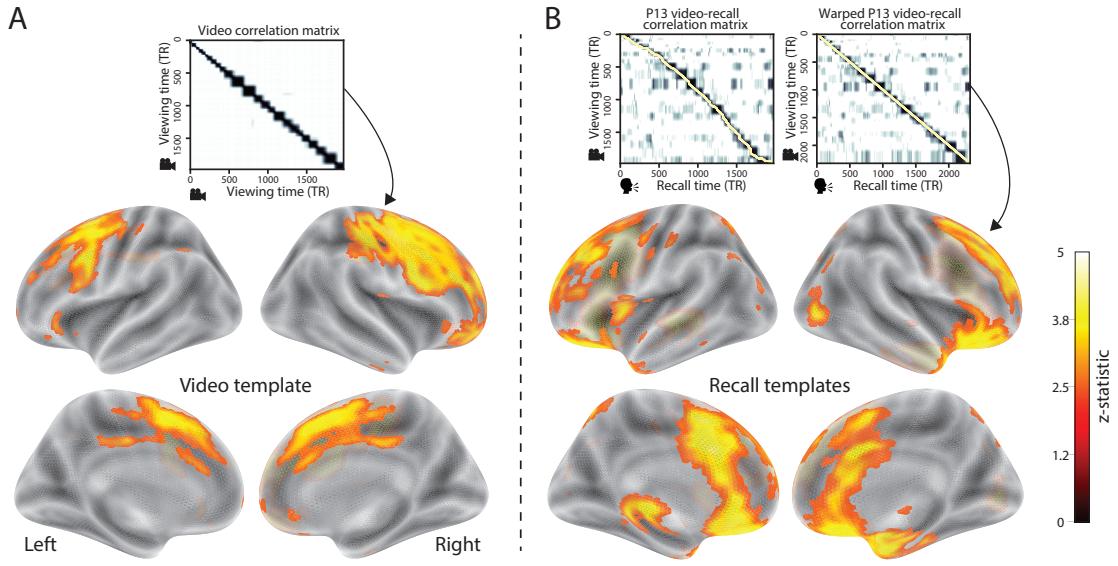


Figure 7: Brain structures that underlie the transformation of experience into memory. **A.** We searched for regions whose responses (as participants watched the video) matched the temporal correlation matrix of the video topic proportions. These regions are sensitive to the narrative structure of the video. **B.** We searched for regions whose responses (as participants watched the video) matched the temporal correlation matrix of the topic proportions derived from each individual's later recall of video. These regions are sensitive to how the narrative structure of the video is transformed into a memory of the video. Both panels: the maps are thresholded at $p < 0.05$, corrected.

321 **Discussion**

322 Our work casts remembering as reproducing (behaviorally and neurally) the topic trajectory, or
323 shape, of an experience. This view draws inspiration from prior work aimed at elucidating
324 the neural and behavioral underpinnings of how we process dynamic naturalistic experiences
325 and remember them later. One approach to identifying neural responses to naturalistic stimuli
326 (including experiences) entails building a model of the stimulus and searching for brain regions
327 whose responses are consistent with the model. In prior work, a series of studies from Uri
328 Hasson's group (Lerner et al., 2011; Simony et al., 2016; Chen et al., 2017; Baldassano et al., 2017;
329 Zadbood et al., 2017) have extended this approach with a clever twist: rather than building an
330 explicit stimulus model, these studies instead search for brain responses (while experiencing the
331 stimulus) that are reliably similar across individuals. So called *inter-subject correlation* (ISC) and
332 *inter-subject functional connectivity* (ISFC) analyses effectively treat other people's brain responses
333 to the stimulus as a "model" of how its features change over time. By contrast, in our present
334 work we used topic models and HMMs to construct an explicit stimulus model (i.e., the topic
335 trajectory of the video). When we searched for brain structures whose responses are consistent
336 with the video's topic trajectory, we identified a network of structures that overlapped strongly
337 with the "long temporal receptive window" network reported by the Hasson group (e.g., compare
338 our Fig. 7A with the map of long temporal receptive window voxels in Lerner et al., 2011). This
339 provides support for the notion that part of the long temporal receptive window network may be
340 maintaining an explicit model of the stimulus dynamics. When we performed a similar analysis
341 after swapping out the video's topic trajectory with the recall topic trajectories of each individual
342 participant, this allowed us to identify brain regions whose responses (as the participants viewed
343 the video) reflected how the video trajectory would be transformed in memory (as reflected by
344 the recall topic trajectories). The analysis revealed that the rMTL and vmPFC may play a role in
345 this person-specific transformation from experience into memory. The role of the MTL in episodic
346 memory encoding has been well-reported (e.g., Paller and Wagner, 2002; Davachi et al., 2003;
347 Ranganath et al., 2004; Davachi, 2006; Wiltgen and Silva, 2007; Diana et al., 2007; van Kesteren

348 et al., 2013). Prior work has also implicated the medial prefrontal cortex in representing “schema”
349 knowledge (i.e., general knowledge about the format of an ongoing experience given prior similar
350 experiences; van Kesteren et al., 2012, 2013; Schlichting and Preston, 2015; Gilboa and Marlatte,
351 2017; Spalding et al., 2018). Integrating across our study and this prior work, one interpretation is
352 that the person-specific transformations mediated (or represented) by the rMTL and vmPFC may
353 reflect schema knowledge being leveraged, formed, or updated, incorporating ongoing experience
354 into previously acquired knowledge.

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

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

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

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

410 While a large number of language models exist (e.g., WAS, LSA, word2vec, universal sentence
411 encoder; Steyvers et al., 2004; Landauer et al., 1998; Mikolov et al., 2013; Cer et al., 2018), here
412 we use latent dirichlet allocation (LDA)-based topic models for a few reasons. First, topic models
413 capture the *essence* of a text passage devoid of the specific set and order of words used. This
414 was an important feature of our model since different people may accurately recall a scene using
415 very different language. Second, words can mean different things in different contexts (e.g. “bat”
416 as the act of hitting a baseball, the object used for that action, or as a flying mammal). Topic
417 models are robust to this, allowing words to exist as part of multiple topics. Last, topic models
418 provide a straightforward means to recover the weights for the particular words comprising a topic,
419 enabling easy interpretation of an event’s contents (e.g. Fig. 6). Other models such as Google’s
420 universal sentence encoder offer a context-sensitive encoding of text passages, but the encoding
421 space is complex and non-linear, and thus recovering the original words used to fit the model is
422 not straightforward. However, it’s worth pointing out that our framework is divorced from the
423 particular choice of language model. Moreover, many of the aspects of our framework could be
424 swapped out for other choices. For example, the language model, the timeseries segmentation
425 model and the video-recall matching function could all be customized for the particular problem.
426 Indeed for some problems, recovery of the particular recall words may not be necessary, and thus
427 other text-modeling approaches (such as universal sentence encoder) may be preferable. Future
428 work will explore the influence of particular model choices on the framework’s accuracy.

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

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

442 Methods

443 Experimental design and data collection

444 Data were collected by Chen et al. (2017). In brief, participants ($n = 17$) viewed the first 48 minutes
445 of "A Study in Pink", the first episode of the BBC television series *Sherlock*, while fMRI volumes
446 were collected (TR = 1500 ms). The stimulus was divided into a 23 min (946 TR) and a 25 min
447 (1030 TR) segment to mitigate technical issues related to the scanner. After finishing the clip,
448 participants were instructed to (quoting from Chen et al., 2017) "describe what they recalled of the
449 [episode] in as much detail as they could, to try to recount events in the original order they were
450 viewed in, and to speak for at least 10 minutes if possible but that longer was better. They were told
451 that completeness and detail were more important than temporal order, and that if at any point
452 they realized they had missed something, to return to it. Participants were then allowed to speak
453 for as long as they wished, and verbally indicated when they were finished (e.g., 'I'm done')."
454 For additional details about the experimental procedure and scanning parameters, see Chen et al.
455 (2017). The experimental protocol was approved by Princeton University's Institutional Review
456 Board.

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

463 **Data and code availability**

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

466 **Statistics**

467 All statistical tests we performed were two-sided.

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

469 **Topic modeling**

470 The input to the topic model we trained to characterize the dynamic content of the video comprised
471 998 hand-generated annotations of short (mean: 2.96s) scenes spanning the video clip (Chen et al.,
472 2017 generated 1000 annotations total; we removed two referring to the break between the first and
473 second scan sessions, during which no fMRI data was collected). The features annotated included:
474 narrative details (a sentence or two describing what happened in that scene); whether the scene
475 took place indoors or outdoors; names of any characters that appeared in the scene; name(s) of
476 characters in camera focus; name(s) of characters who were speaking in the scene; the location (in
477 the story) that the scene took place; camera angle (close up, medium, long, top, tracking, over the
478 shoulder, etc.); whether music was playing in the scene or not; and a transcription of any on-screen
479 text. We concatenated the text for all of these features within each segment, creating a “bag of
480 words” describing each scene. We then re-organized the text descriptions into overlapping sliding

481 windows of 50 scenes each. In other words, we created a “context” for each scene comprising the
482 text descriptions of the preceding 25 scenes, the present scene, and the following 24 scenes. To
483 model the “context” at the beginning and end of the video (i.e., within 50 scenes of the beginning or
484 end), we created overlapping sliding windows that grew in size from one scene to the full length,
485 then similarly tapered their length at the end. This bore the additional benefit of representing each
486 scene’s description in the text corpus an equal number of times.

487 We trained our model using these overlapping text samples with `scikit-learn` (version 0.19.1;
488 Pedregosa et al., 2011), called from our high-dimensional visualization and text analysis software,
489 `HyperTools` (Heusser et al., 2018b). Specifically, we used the `CountVectorizer` class to transform
490 the text from each window into a vector of word counts (using the union of all words across all
491 scenes as the “vocabulary,” excluding English stop words); this yielded a number-of-windows
492 by number-of-words *word count* matrix. We then used the `LatentDirichletAllocation` class
493 (`topics=100, method='batch'`) to fit a topic model (Blei et al., 2003) to the word count matrix,
494 yielding a number-of-windows (1047) by number-of-topics (100) *topic proportions* matrix. The
495 topic proportions matrix describes which mix of topics (latent themes) is present in and around
496 each scene. Next, we transformed the topic proportions matrix to match the 1976 fMRI volume
497 acquisition times. We assigned each topic vector to the timepoint midway between the beginning
498 of the first scene and the end of the last scene in its corresponding sliding text window. We
499 then transformed these timepoints to units of TRs and interpolated the dynamic topic proportions
500 matrix to obtain number-of-TRs (1976) by number-of-topics (100) matrix.

501 We created similar topic proportions matrices using hand-annotated transcripts of each partici-
502 pant’s recall of the video (annotated by Chen et al., 2017). We tokenized the transcript into a list of
503 sentences, and then re-organized the list into overlapping sliding windows spanning 10 sentences
504 each (and analogously tapered the lengths of the first and last 10 sliding windows). In turn, we
505 transformed each window’s sentences into a word count vector (using the same vocabulary as for
506 the video model). We then used the topic model already trained on the video scenes to compute
507 the most probable topic proportions for each sliding window. This yielded a number-of-windows
508 (range: 83–312) by number-of-topics (100) topic proportions matrix for each participant. These

509 reflected the dynamic content of each participant's recalls. Note: for details on how we selected the
510 video and recall window lengths and number of topics, see *Supporting Information* and Figure S1.

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

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

519 We used an optimization procedure to select the appropriate K for each topic proportions
520 matrix. Specifically, we computed (for each matrix)

$$\operatorname{argmax}_K \left[POM\left(\frac{a}{b-J}\right) - \frac{K}{\alpha} \right],$$

521 where a was the average correlation between the topic vectors of timepoints within the same state;
522 b was the average correlation between the topic vectors of timepoints within *different* states; J was
523 a constant used to ensure a positive denominator, set equal to $\min_K \left[\frac{a}{b} \right]$; and α was a regularization
524 parameter that we set to 5 times the window length (i.e., 250 scenes for the video topic trajectory
525 and 50 sentences for the recall topic trajectories). Before subtracting the regularization term, we
526 scaled the ratio of correlations to be a proportion of the maximum ratio across all K 's. Figure 2B
527 displays the event boundaries returned for the video, and Figure S4 displays the event boundaries
528 returned for each participant's recalls (See Fig. S6 for the values of a and b for each K , Fig. S7 for
529 the optimization functions for the video and recalls). After obtaining these event boundaries, we
530 created stable estimates of each topic proportions matrix by averaging the topic vectors within
531 each event. This yielded a number-of-events by number-of-topics matrix for the video and recalls
532 from each participant.

533 We also evaluated a parameter-free procedure for choosing K , which finds the K value that
534 maximizes the Wasserstein distance (a.k.a. “Earth mover’s” distance) between the within and
535 across event distributions of correlation values. This alternative procedure largely replicated the
536 pattern of results found with the parameterized method described above, but recovered sub-
537 stantially fewer events on average (Fig S8). While both approaches seem to underestimate the
538 number of video/recall events relative to the “true” number (as determined by human raters), the
539 parameterized approach was closer to the true number.

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

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

547 Perhaps the simplest and most widely used measure of memory performance is *accuracy*—i.e.,
548 the proportion of studied (experienced) items (in this case, the 34 video events) that the participant
549 later remembered. Chen et al. (2017) developed a human rating system whereby the quality of
550 each participant’s memory was evaluated by an independent rater. We found a strong across-
551 participants correlation between these independant ratings and the overall number of events that
552 our HMM approach identified in participants’ recalls (Pearson’s $r(15) = 0.64, p = 0.006$).

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

556 **Probability of first recall (PFR).** PFR curves (Welch and Burnett, 1924; Postman and Phillips,
557 1965; Atkinson and Shiffrin, 1968) reflect the probability that an item will be recalled first as a
558 function of its serial position during encoding. To carry out this analysis, we initialized a number-

559 of-participants (17) by number-of-video-events (34) matrix of zeros. Then for each participant, we
560 found the index of the video event that was recalled first (i.e., the video event whose topic vector
561 was most strongly correlated with that of the first recall event) and filled in that index in the matrix
562 with a 1. Finally, we averaged over the rows of the matrix, resulting in a 1 by 34 array representing
563 the proportion of participants that recalled an event first, as a function of the order of the event's
564 appearance in the video (Fig. 3A).

565 **Lag conditional probability curve (lag-CRP).** The lag-CRP curve (Kahana, 1996) reflects the
566 probability of recalling a given event after the just-recalled event, as a function of their relative
567 positions (or *lag*). In other words, a lag of 1 indicates that a recalled event came immediately after
568 the previously recalled event in the video, and a lag of -3 indicates that a recalled event came 3
569 events before the previously recalled event. For each recall transition (following the first recall),
570 we computed the lag between the current recall event and the next recall event, normalizing by
571 the total number of possible transitions. This yielded a number-of-participants (17) by number-
572 of-lags (-33 to +33; 67 lags total) matrix. We averaged over the rows of this matrix to obtain a
573 group-averaged lag-CRP curve (Fig. 3B).

574 **Serial position curve (SPC).** SPCs (Murdock, 1962) reflect the proportion of participants that
575 remember each item as a function of the items' serial position during encoding. We initialized
576 a number-of-participants (17) by number-of-video-events (34) matrix of zeros. Then, for each
577 recalled event, for each participant, we found the index of the video event that the recalled event
578 most closely matched (via the correlation between the events' topic vectors) and entered a 1 into
579 that position in the matrix (i.e., for the given participant and event). This resulted in a matrix
580 whose entries indicated whether or not each event was recalled by each participant (depending
581 on whether the corresponding entires were set to one or zero). Finally, we averaged over the rows
582 of the matrix to yield a 1 by 34 array representing the proportion of participants that recalled each
583 event as a function of the order of the event's appearance in the video (Fig. 3C).

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

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

602 **Novel naturalistic memory metrics**

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

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

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

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

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

621 We used the UMAP algorithm (McInnes and Healy, 2018) to project the 100-dimensional topic space
622 onto a two-dimensional space for visualization (Figs. 5, 6). To ensure that all of the trajectories were
623 projected onto the *same* lower dimensional space, we computed the low-dimensional embedding
624 on a “stacked” matrix created by vertically concatenating the events-by-topics topic proportions
625 matrices for the video and all 17 participants’ recalls. We then divided the rows of the result (a
626 total-number-of-events by two matrix) back into separate matrices for the video topic trajectory
627 and the trajectories for each participant’s recalls (Fig. 5). This general approach for discovering
628 a shared low-dimensional embedding for a collections of high-dimensional observations follows
629 Heusser et al. (2018b).

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

631 In Figure 5B, we present an analysis aimed at characterizing locations in topic space that dif-
632 ferent participants move through in a consistent way (via their recall topic trajectories). The
633 two-dimensional topic space used in our visualizations (Fig. 5) comprised a 9x9 (arbitrary units)
634 square. We divided this space into a grid of vertices spaced 0.25 units apart. For each vertex, we
635 examined the set of line segments formed by connecting each pair successively recalled events,
636 across all participants, that passed within 0.5 units. We computed the distribution of angles formed
637 by those segments and the x -axis, and used a Rayleigh test to determine whether the distribution
638 of angles was reliably “peaked” (i.e., consistent across all transitions that passed through that local
639 portion of topic space). To create Figure 5B we drew an arrow originating from each grid vertex,
640 pointing in the direction of the average angle formed by line segments that passed within 0.5 units.
641 We set the arrow lengths to be inversely proportional to the p -values of the Rayleigh tests at each
642 vertex. Specifically, for each vertex we converted all of the angles of segments that passed within
643 0.5 units to unit vectors, and we set the arrow lengths at each vertex proportional to the length
644 of the (circular) mean vector. We also indicated any significant results ($p < 0.05$, corrected using
645 the Benjamani-Hochberg procedure) by coloring the arrows in blue (darker blue denotes a lower
646 p -value, i.e., a longer mean vector); all tests with $p \geq 0.05$ are displayed in gray and given a lower
647 opacity value.

648 **Searchlight fMRI analyses**

649 In Figure 7, we present two analyses aimed at identifying brain structures whose responses (as
650 participants viewed the video) exhibited particular temporal correlations. We developed a search-
651 light analysis whereby we constructed a cube centered on each voxel (radius: 5 voxels). For each
652 of these cubes, we computed the temporal correlation matrix of the voxel responses during video
653 viewing. Specifically, for each of the 1976 volumes collected during video viewing, we correlated
654 the activity patterns in the given cube with the activity patterns (in the same cube) collected during
655 every other timepoint. This yielded a 1976 by 1976 correlation matrix for each cube.

656 Next, we constructed two sets of “template” matrices: one reflecting the video’s topic trajectory
657 and the other reflecting each participant’s recall topic trajectory. To construct the video template, we
658 computed the correlations between the topic proportions estimated for every pair of TRs (prior to
659 segmenting the trajectory into discrete events; i.e., the correlation matrix shown in Figs. 2B and 7A).
660 We constructed similar temporal correlation matrices for each participant’s recall topic trajectory
661 (Figs. 2D, S4). However, to correct for length differences and potential non-linear transformations
662 between viewing time and recall time, we first used dynamic time warping (Berndt and Clifford,
663 1994) to temporally align participants’ recall topic trajectories with the video topic trajectory (an
664 example correlation matrix before and after warping is shown in Fig. 7B). This yielded a 1976 by
665 1976 correlation matrix for the video template and for each participant’s recall template.

666 To determine which (cubes of) voxel responses reliably matched the video template, we cor-
667 related the upper triangle of the voxel correlation matrix for each cube with the upper triangle
668 of the video template matrix (Kriegeskorte et al., 2008). This yielded, for each participant, a
669 single correlation value. We computed the average (Fisher z -transformed) correlation coefficient
670 across participants. We used a permutation-based procedure to assess significance, whereby we
671 re-computed the average correlations for each of 100 “null” video templates (constructed by circu-
672 larly shifting the template by a random number of timepoints). (For each permutation, the same
673 shift was used for all participants.) We then estimated a p -value by computing the proportion of
674 shifted correlations that were larger than the observed (unshifted) correlation. To create the map
675 in Figure 7A we thresholded out any voxels whose correlation values fell below the 95th percentile
676 of the permutation-derived null distribution.

677 We used a similar procedure to identify which voxels’ responses reflected the recall templates.
678 For each participant, we correlated the upper triangle of the correlation matrix for each cube of
679 voxels with their (time warped) recall correlation matrix. As in the video template analysis this
680 yielded a single correlation coefficient for each participant. However, whereas the video analysis
681 compared every participant’s responses to the same template, here the recall templates were
682 unique for each participant. We computed the average z -transformed correlation coefficient across
683 participants, and used the same permutation procedure we developed for the video responses to

684 assess significant correlations. To create the map in Figure 7B we thresholded out any voxels whose
685 correlation values fell below the 95th percentile of the permutation-derived null distribution.

686 References

- 687 Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control
688 processes. In Spence, K. W. and Spence, J. T., editors, *The psychology of learning and motivation*,
689 volume 2, pages 89–105. Academic Press, New York.
- 690 Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017).
691 Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–
692 721.
- 693 Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In
694 *KDD workshop*, volume 10, pages 359–370.
- 695 Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International
696 Conference on Machine Learning*, ICML ’06, pages 113–120, New York, NY, US. ACM.
- 697 Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine
698 Learning Research*, 3:993 – 1022.
- 699 Brunec, I. K., Moscovitch, M. M., and Barense, M. D. (2018). Boundaries shape cognitive represen-
700 tations of spaces and events. *Trends in Cognitive Sciences*, 22(7):637–650.
- 701 Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., and Oliva, A. (2015). Intrinsic and extrinsic
702 effects on image memorability. *Vision Research*, 116:165–178.
- 703 Capota, M., Turek, J., Chen, P.-H., Zhu, X., Manning, J. R., Sundaram, N., Keller, B., Wang, Y., and
704 Shin, Y. S. (2017). Brain imaging analysis kit.
- 705 Cer, D., Yang, Y., y Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes,
706 M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.
707 *arXiv*, 1803.11175.

- 708 Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared
709 memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*,
710 20(1):115.
- 711 Chun, M. and Turk-Browne, N. (2007). Interactions between attention and memory. *Current opinion*
712 in *neurobiology*, 17(2):177–184.
- 713 Clewett, D. and Davachi, L. (2017). The ebb and flow of experience determines the temporal
714 structure of memory. *Curr Opin Behav Sci*, 17:186–193.
- 715 Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in*
716 *Neurobiology*, 16(6):693—700.
- 717 Davachi, L., Mitchell, J. P., and Wagner, A. D. (2003). Multiple routes to memory: distinct medial
718 temporal lobe processes build item and source memories. *Proceedings of the National Academy of*
719 *Sciences, USA*, 100(4):2157 – 2162.
- 720 Diana, R. A., Yonelinas, A. P., and Ranganath, C. (2007). Imaging recollection and famil-
721 iarity in the medial temporal lobe: a three-component model. *Trends in Cognitive Sciences*,
722 doi:10.1016/j.tics.2007.08.001.
- 723 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the
724 sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–1286.
- 725 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological*
726 *Science*, 22(2):243–252.
- 727 Friendly, M., Franklin, P. E., Hoffman, D., and Rubin, D. C. (1982). The Toronto Word Pool:
728 Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080
729 words. *Behavior Research Methods and Instrumentation*, 14:375–399.
- 730 Gilboa, A. and Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory.
731 *Trends Cogn Sci*, 21(8):618–631.

- 732 Heusser, A. C., Ezzyat, Y., Shiff, I., and Davachi, L. (2018a). Perceptual boundaries cause mnemonic
733 trade-offs between local boundary processing and across-trial associative binding. *Journal of*
734 *Experimental Psychology Learning, Memory, and Cognition*, 44(7):1075–1090.
- 735 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail: a
736 Python toolbox for analyzing and plotting free recall data. *The Journal of Open Source Software*,
737 10.21105/joss.00424.
- 738 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018b). HyperTools: a Python
739 toolbox for gaining geometric insights into high-dimensional data. *Journal of Machine Learning*
740 *Research*, 18(152):1–6.
- 741 Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal*
742 *of Mathematical Psychology*, 46:269–299.
- 743 Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., and H., E.
744 (2014). A unified mathematical framework for coding time, space, and sequences in the medial
745 temporal lobe. *Journal of Neuroscience*, 34(13):4692–4707.
- 746 Howard, M. W., Viskontas, I. V., Shankar, K. H., and Fried, I. (2012). Ensembles of human MTL
747 neurons “jump back in time” in response to a repeated stimulus. *Hippocampus*, 22:1833–1847.
- 748 Huk, A., Bonnen, K., and He, B. J. (2018). Beyond trial-based paradigms: continuous behavior, on-
749 going neural activity, and naturalistic stimuli. *Journal of Neuroscience*, 10.1523/JNEUROSCI.1920-
750 17.2018.
- 751 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24:103–109.
- 752 Koriat, A. and Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: distin-
753 guishing accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of*
754 *Experimental Psychology: General*, 123(3):297–315.
- 755 Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – con-
756 necting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:1 – 28.

- 757 Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to latent semantic analysis.
758 *Discourse Processes*, 25:259–284.
- 759 Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy
760 of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.
- 761 Manning, J. R. (2019). Episodic memory: mental time travel or a quantum ‘memory wave’ function?
762 *PsyArXiv*, doi:10.31234/osf.io/6zjwb.
- 763 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory.
764 In Gazzaniga, M., editor, *The Cognitive Neurosciences, Fifth edition*, pages 557–566. MIT Press.
- 765 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns
766 in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National
767 Academy of Sciences, USA*, 108(31):12893–12897.
- 768 McInnes, L. and Healy, J. (2018). UMAP: Uniform manifold approximation and projection for
769 dimension reduction. *arXiv*, 1802(03426).
- 770 Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations
771 in vector space. *arXiv*, 1301.3781.
- 772 Mueller, A., Fillion-Robin, J.-C., Boidol, R., Tian, F., Nechifor, P., yoonsubKim, Peter, Rampin, R.,
773 Corvellec, M., Medina, J., Dai, Y., Petrushev, B., Langner, K. M., Hong, Alessio, Ozsvald, I.,
774 vkolmakov, Jones, T., Bailey, E., Rho, V., IgorAPM, Roy, D., May, C., foobuzz, Piyush, Seong,
775 L. K., Goey, J. V., Smith, J. S., Gus, and Mai, F. (2018). WordCloud 1.5.0: a little word cloud
776 generator in Python. *Zenodo*, <https://zenodo.org/record/1322068#.W4tPKZNKh24>.
- 777 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*,
778 64:482–488.
- 779 Paller, K. A. and Wagner, A. D. (2002). Observing the transformation of experience into memory.
780 *Trends in Cognitive Sciences*, 6(2):93–102.

- 781 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten-
782 hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot,
783 M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine*
784 *Learning Research*, 12:2825–2830.
- 785 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model
786 of organizational processes in free recall. *Psychological Review*, 116(1):129–156.
- 787 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal*
788 *of Experimental Psychology*, 17:132–138.
- 789 Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech
790 recognition. *Proceedings of the IEEE*, 77(2):257–286.
- 791 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition. *Curr Opin*
792 *Behav Sci*, 17:133–140.
- 793 Ranganath, C., Cohen, M. X., Dam, C., and D’Esposito, M. (2004). Inferior temporal, prefrontal,
794 and hippocampal contributions to visual working memory maintenance and associative memory
795 retrieval. *Journal of Neuroscience*, 24(16):3917–3925.
- 796 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature*
797 *Reviews Neuroscience*, 13:713 – 726.
- 798 Schlichting, M. L. and Preston, A. R. (2015). Memory integration: neural mechanisms and impli-
799 cations for behavior. *Current Opinion in Behavioral Sciences*, 1:1–8.
- 800 Simony, E., Honey, C. J., Chen, J., and Hasson, U. (2016). Uncovering stimulus-locked network
801 dynamics during narrative comprehension. *Nature Communications*, 7(12141):1–13.
- 802 Spalding, K. N., Schlichting, M. L., Zeithamova, D., Preston, A. R., Tranel, D., Duff, M. C., and
803 Warren, D. E. (2018). Ventromedial prefrontal cortex is necessary for normal associative inference
804 and memory integration. *The Journal of Neuroscience*, 38(15):3767–3775.

- 805 Stark, S. M., Yassa, M. A., and Stark, C. E. L. (2010). Individual differences in spatial pattern
806 separation performance associated with healthy aging in humans. *Learning & Memory*, 17(6):284–
807 288.
- 808 Steyvers, M., Shiffrin, R. M., and Nelson, D. L. (2004). Word association spaces for predicting
809 semantic similarity effects in episodic memory. In Healy, A. F., editor, *Cognitive Psychology and*
810 *its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*. American
811 Psychological Association, Washington, DC.
- 812 Tompry, A. and Davachi, L. (2017). Consolidation promotes the emergence of representational
813 overlap in the hippocampus and medial prefrontal cortex. *Neuron*, 96(1):228–241.
- 814 Tyng, C. M., Amin, H. U., Saad, M. N. M., and S, M. A. (2017). The influences of emotion on
815 learning and memory. *Frontiers in psychology*, 8:1454.
- 816 van Kesteren, M. T. R., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., and Fernández, G.
817 (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent
818 encoding: from congruent to incongruent. *Neuropsychologia*, 51(12):2352–2359.
- 819 van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., and Henson, R. N. (2012). How schema and
820 novelty augment memory formation. *Trends Neurosci*, 35(4):211–9.
- 821 Waskom, M., Botvinnik, O., Okane, D., Hobson, P., David, Halchenko, Y., Lukauskas, S., Cole, J. B.,
822 Warmenhoven, J., de Ruiter, J., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E.,
823 Martin, M., Miles, A., Meyer, K., Augspurger, T., Yarkoni, T., Bachant, P., Williams, M., Evans,
824 C., Fitzgerald, C., Brian, Wehner, D., Hitz, G., Ziegler, E., Qalieh, A., and Lee, A. (2016). Seaborn:
825 v0.7.1.
- 826 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American Journal*
827 *of Psychology*, 35:396–401.
- 828 Wiltgen, B. J. and Silva, A. J. (2007). Memory for context becomes less specific with time. *Learning*
829 *& Memory*, 14(4):313–317.

- 830 Yassa, M. A., Lacy, J. W., Stark, S. M., Albert, M. S., Gallagher, M., and Stark, C. E. L. (2011). Pattern
831 separation deficits associated with increased hippocampal ca3 and dentate gyrus activity in
832 nondemented older adults. *Hippocampus*, 21(9):968–979.
- 833 Yassa, M. A. and Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends In Neuro-*
834 *sciences*, 34(10):515–525.
- 835 Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research.
836 *Journal of Memory and Language*, 46:441–517.
- 837 Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception:
838 a mind-brain perspective. *Psychological Bulletin*, 133:273–293.
- 839 Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit
840 memories to other brains: Constructing shared neural representations via communication. *Cereb*
841 *Cortex*, 27(10):4988–5000.
- 842 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and
843 memory. *Psychological Bulletin*, 123(2):162 – 185.

844 Supporting information

- 845 Supporting information is available in the online version of the paper.

846 Acknowledgements

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

⁸⁵² **Author contributions**

⁸⁵³ Conceptualization: A.C.H. and J.R.M.; Methodology: A.C.H. and J.R.M.; Software: A.C.H., P.C.F.
⁸⁵⁴ and J.R.M.; Analysis: A.C.H., P.C.F. and J.R.M.; Writing, Reviewing, and Editing: A.C.H., P.C.F.
⁸⁵⁵ and J.R.M.; Supervision: J.R.M.

⁸⁵⁶ **Author information**

⁸⁵⁷ The authors declare no competing financial interests. Correspondence and requests for materials
⁸⁵⁸ should be addressed to J.R.M. (jeremy.r.manning@dartmouth.edu).