

<sup>1</sup> A novel framework for linking dynamic experiences to  
<sup>2</sup> memories reveals event-like structure in naturalistic  
<sup>3</sup> episodic recall

<sup>4</sup> Andrew C. Heusser, Paxton C. Fitzpatrick, and Jeremy R. Manning

Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

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

<sup>5</sup> July 21, 2019

<sup>6</sup> **Abstract**

<sup>7</sup> Our life experiences unfold over time in highly complex manner, with the evolving presence  
<sup>8</sup> and absence of numerous intricate features describing our journey between each circumstance  
<sup>9</sup> or event we encounter. Here, we propose a framework for mapping dynamic naturalistic ex-  
<sup>10</sup> periences onto geometric spaces as *trajectories* that capture the temporal dynamics of real-world  
<sup>11</sup> content. Within this geometric framework, one may compare the shape of the trajectory formed  
<sup>12</sup> by an experience to that defined by one's later recollection to characterize our memories' re-  
<sup>13</sup> covery and distortion of the external world. Here, we apply this approach to a naturalistic  
<sup>14</sup> memory experiment in which participants viewed and verbally recounted a video, and find  
<sup>15</sup> that the video and subsequent recalls share both an experience-specific shape and a discernible  
<sup>16</sup> event-like structure. However, the level of *precision* with which individuals recounted various  
<sup>17</sup> events and the *distinctiveness* of recall for those events were varied and predictive of overall  
<sup>18</sup> memory performance. Finally, we identify a network of brain structures that is sensitive to the  
<sup>19</sup> "shapes" of our ongoing experiences, and an overlapping network sensitive to how we will later

remember them. These results highlight the rich event-like structure of the external world and our memories, and offer novel, content-sensitive alternatives to classical “proportion recalled” measures for assessing episodic memory

## Introduction

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

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

59 Although our experiences most often change gradually, they also occasionally change sud-  
60 denly (e.g., when we walk through a doorway; Radvansky and Zacks, 2017). Prior research  
61 suggests that these sharp transitions (termed *event boundaries*) during an experience help to dis-  
62 cretize our experiences (and their mental representations) into *events* (Radvansky and Zacks, 2017;  
63 Brunec et al., 2018; Heusser et al., 2018a; Clewett and Davachi, 2017; Ezzyat and Davachi, 2011;  
64 DuBrow and Davachi, 2013). The interplay between the stable (within event) and transient (across  
65 event) temporal dynamics of an experience also provides a potential framework for transforming  
66 experiences into memories that distill those experiences down to their essence. For example, prior  
67 work has shown that event boundaries can influence how we learn sequences of items (Heusser  
68 et al., 2018a; DuBrow and Davachi, 2013), navigate (Brunec et al., 2018), and remember and un-  
69 derstand narratives (Zwaan and Radvansky, 1998; Ezzyat and Davachi, 2011). Prior research has  
70 implicated the hippocampus and the medial prefrontal cortex as playing a critical role in trans-  
71 forming experiences into structured and consolidated memories (Tompry and Davachi, 2017).

72 Here we sought to examine how the temporal dynamics of a “naturalistic” experience were  
73 later reflected in participants’ memories. We analyzed an open dataset that comprised behavioral

74 and functional Magnetic Resonance Imaging (fMRI) data collected as participants viewed and then  
75 verbally recounted an episode of the BBC television series *Sherlock* (Chen et al., 2017). We developed  
76 a computational framework for characterizing the temporal dynamics of the moment-by-moment  
77 content of the episode and of participants' verbal recalls. Specifically, we use topic modeling (Blei  
78 et al., 2003) to characterize the thematic conceptual (semantic) content present in each moment of  
79 the episode and recalls, and Hidden Markov Models (Rabiner, 1989; Baldassano et al., 2017) to  
80 discretize the evolving semantic content into events. In this way, we cast naturalistic experiences  
81 (and recalls of those experiences) as *trajectories* that describe how the experiences evolve over  
82 time. Under this framework, successful remembering entails verbally "traversing" the content  
83 trajectory of the episode, thereby reproducing the shape (or essence) of the original experience.  
84 Comparing the shapes of the topic trajectories of the episode and of participants' retellings of the  
85 episode then reveals which aspects of the episode were preserved (or lost) in the translation into  
86 memory. We further examine whether 1) the *precision* with which a participant recounts each event  
87 and 2) the *distinctiveness* each recall event is (relative to the other recalled events) relates to their  
88 overall memory performance. Last, we identify networks of brain structures whose responses (as  
89 participants watched the episode) reflected the shape of the episode, and how participants would  
90 later recount the episode.

## 91 **Results**

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

100 column per topic. The topic proportions matrix describes which mixture of topics is reflected in  
101 each document.

102 Chen et al. (2017) collected hand-annotated information about each of 1000 (manually identified)  
103 scenes spanning the roughly 50 minute video used in their experiment. This information included:  
104 a brief narrative description of what was happening; whether the scene took place indoors vs.  
105 outdoors; names of any characters on the screen; names of any characters who were in focus in  
106 the camera shot; names of characters who were speaking; the location where the scene took place;  
107 the camera angle (close up, medium, long, etc.); whether or not background music was present;  
108 and other similar details (for a full list of annotated features see *Methods*). We took from these  
109 annotations the union of all unique words (excluding stop words, such as “and,” “or,” “but,” etc.)  
110 across all features and scenes as the “vocabulary” for the topic model. We then concatenated the  
111 sets of words across all features contained in overlapping, 50-scene sliding windows, and treated  
112 each 50-scene sequence as a single “document” for the purpose of fitting the topic model. Next,  
113 we fit a topic model with (up to)  $K = 100$  topics to this collection of documents. We found that 27  
114 unique topics (with non-zero weights) were sufficient to describe the time-varying content of the  
115 video (see *Methods*; Figs. 1, S2). Note that our approach is similar in some respects to Dynamic Topic  
116 Models (Blei and Lafferty, 2006), in that we sought to characterize how the thematic content of the  
117 episode evolved over time. However, whereas Dynamic Topic Models are designed to characterize  
118 how the properties of *collections* of documents change over time, our sliding window approach  
119 allows us to examine the topic dynamics within a single document (or video). Specifically, our  
120 approach yielded (via the topic proportions matrix) a single *topic vector* for each timepoint of the  
121 episode (we set timepoints to match the acquisition times of the 1976 fMRI volumes collected as  
122 participants viewed the episode).

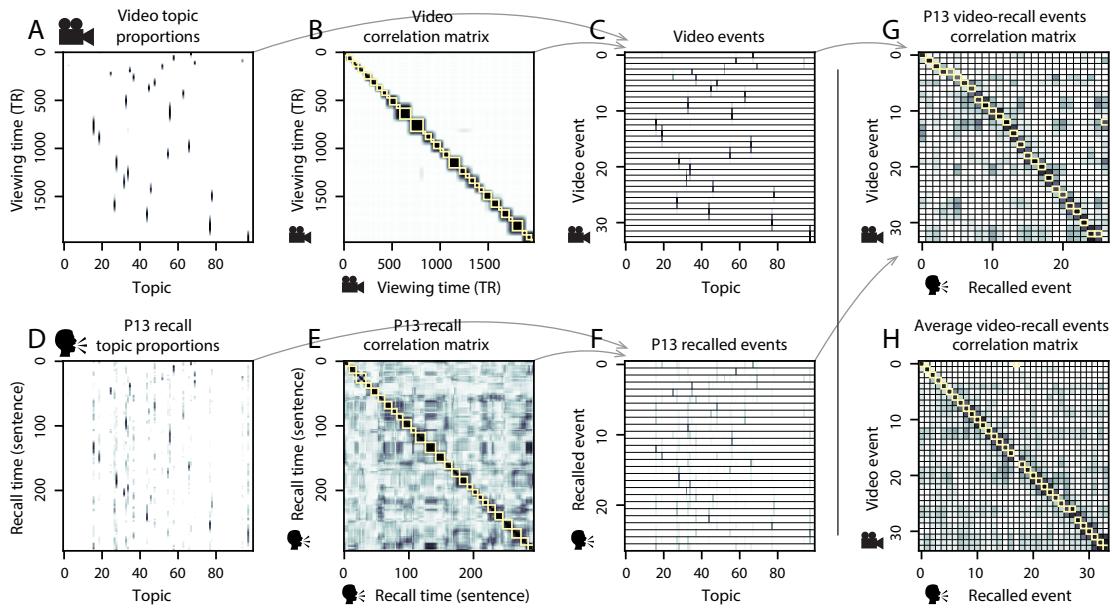
123 The topics we found were heavily character-focused (e.g., the top-weighted word in each topic  
124 was nearly always a character) and could be roughly divided into themes that were primarily  
125 Sherlock Holmes-focused (Sherlock is the titular character); primarily John Watson-focused (John  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



**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.

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

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



**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 (34 events detected). **C.** Average topic vectors for each of the 34 video events. **D.** Topic vectors for each of 294 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 (27 events detected). **F.** Average topic vectors for each of the 27 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).

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

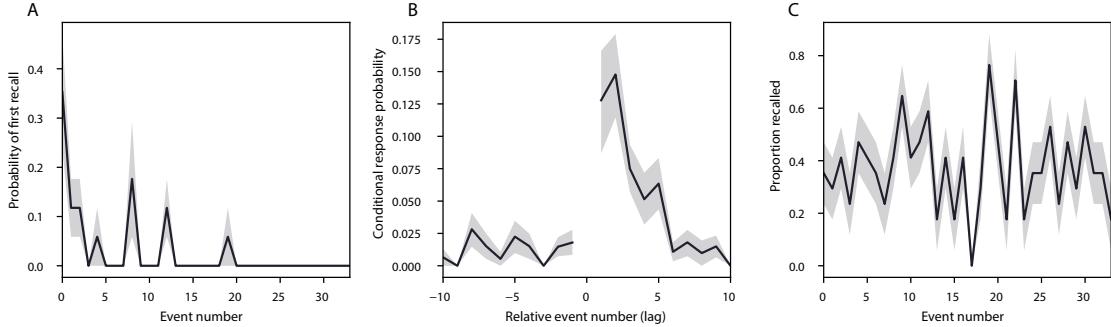
Although the example participant’s recall topic proportions matrix has some visual similarity to the video topic proportions matrix, the time-varying topic proportions for the example participant’s recalls are not as sparse as for the video (e.g., compare Figs. 2A and D). Similarly, although there do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

Two clear patterns emerged from this set of analyses. First, although every individual participant’s recalls could be segmented into discrete events (i.e., every individual participant’s recall correlation matrix exhibited clear block diagonal structure; Fig. S4), each participant appeared to have a unique *recall resolution*, reflected in the sizes of those blocks. For example, some participants’ recall topic proportions segmented into just a few events (e.g., Participants P1, P4, and P15), while others’ recalls segmented into many shorter duration events (e.g., Participants P12, P13, and P17). This suggests that different participants may be recalling the video with different levels of detail—e.g., some might touch on just the major plot points, whereas others might attempt to recall every minor scene. The second clear pattern present in every individual participant’s recall correlation matrix is that, unlike in the video correlation matrix, there are substantial off-diagonal correlations in participant’s recalls. Whereas each event in the original video (was largely) separable from the others (Fig. 2B), in transforming those separable events into memory participants appear to be

<sup>184</sup> integrating *across* different events, blending elements of previously recalled and not-yet-recalled  
<sup>185</sup> events into each newly recalled event (Figs. 2D, S4; also see Manning et al., 2011; Howard et al.,  
<sup>186</sup> 2012).

<sup>187</sup> The above results indicate that both the structure of the original video and participants' recalls  
<sup>188</sup> of the video exhibit event boundaries that can be identified automatically by characterizing the  
<sup>189</sup> dynamic content using a shared topic model and segmenting the content into events using HMMs.  
<sup>190</sup> Next we asked whether some correspondence might be made between the specific content of  
<sup>191</sup> the events the participants experienced in the video, and the events they later recalled. One  
<sup>192</sup> approach to linking the experienced (video) and recalled events is to label each recalled event as  
<sup>193</sup> matching the video event with the most similar (i.e., most highly correlated) topic vector (Figs. 2G,  
<sup>194</sup> S5). This yields a sequence of "presented" events from the original video, and a sequence of  
<sup>195</sup> (potentially differently ordered) "recalled" events for each participant. Analogous to classic list-  
<sup>196</sup> learning studies, we can then examine participants' recall sequences by asking which events  
<sup>197</sup> they tended to recall first (probability of first recall; Fig. 3A; Welch and Burnett, 1924; Postman  
<sup>198</sup> and Phillips, 1965; Atkinson and Shiffrin, 1968); how participants most often transition between  
<sup>199</sup> recalls of the events as a function of the temporal distance between them (lag-conditional response  
<sup>200</sup> probability; Fig. 3B; Kahana, 1996); and which events they were likely to remember overall (serial  
<sup>201</sup> position recall analyses; Fig. 3C; Murdock, 1962). Interestingly, for 2 of the analyses (probability  
<sup>202</sup> of first recall and lag-conditional response probability curves) we observe patterns comparable to  
<sup>203</sup> classic effects from the list-learning literature. Namely, a higher probability of initiating recall with  
<sup>204</sup> the first event in the sequence (Fig. 3A) and a higher probability of transitioning to neighboring  
<sup>205</sup> events with a forward asymmetric bias (Fig. 3C). In contrast, we do not observe a pattern comparable  
<sup>206</sup> to the serial position effect (Fig. 3C), but rather we see higher memory for specific events distributed  
<sup>207</sup> somewhat evenly throughout the video.

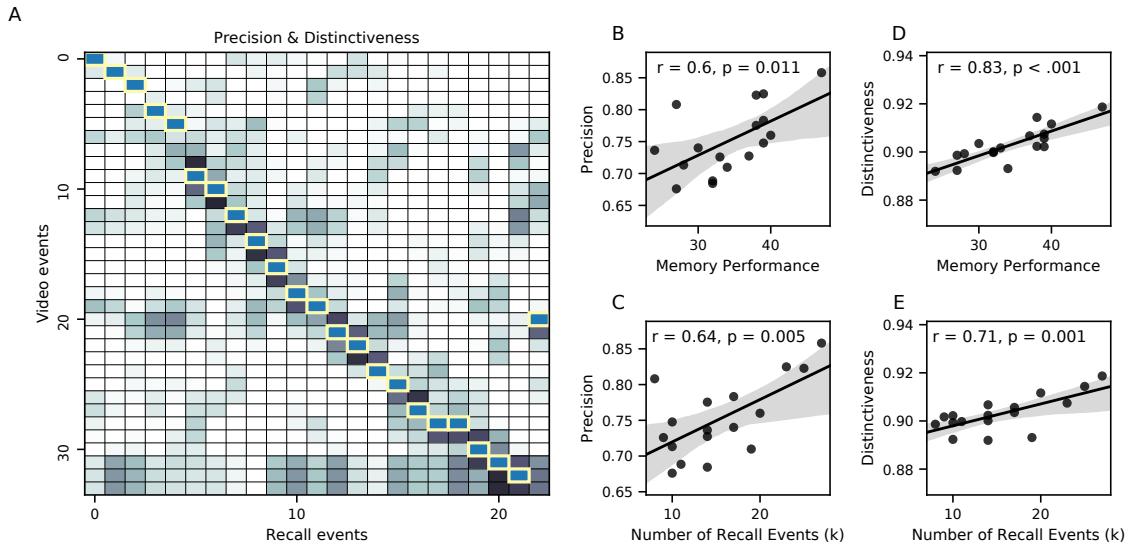
<sup>208</sup> Statistical models of memory studies often treat memory recalls as binary (e.g. the item was  
<sup>209</sup> recalled or not) and independent events. However, our framework produces a content-based model  
<sup>210</sup> of individual stimulus and recall events, allowing for direct quantitative comparison between all  
stimulus and recall events, as well as between the recall events themselves. Leveraging these



**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.

content-based models of the stimulus/recall events, we developed 2 novel metrics for quantifying naturalistic memory representations: precision and distinctiveness. We define precision as the average correlation between 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 (Pearson's  $r(15) = 0.6, p = 0.011$ ) as well as the number of recall events estimated by our model (Pearson's  $r(15) = 0.64, p = 0.005$ ). 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. Similar to precision, we find that the more distinct participants recalls are (on average), the more they remembered (hand-annotated memory: Pearson's  $r(15) = 0.83, p < 0.001$  and model derived memory: Pearson's  $r(15) = 0.71, p = 0.001$ ). 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.

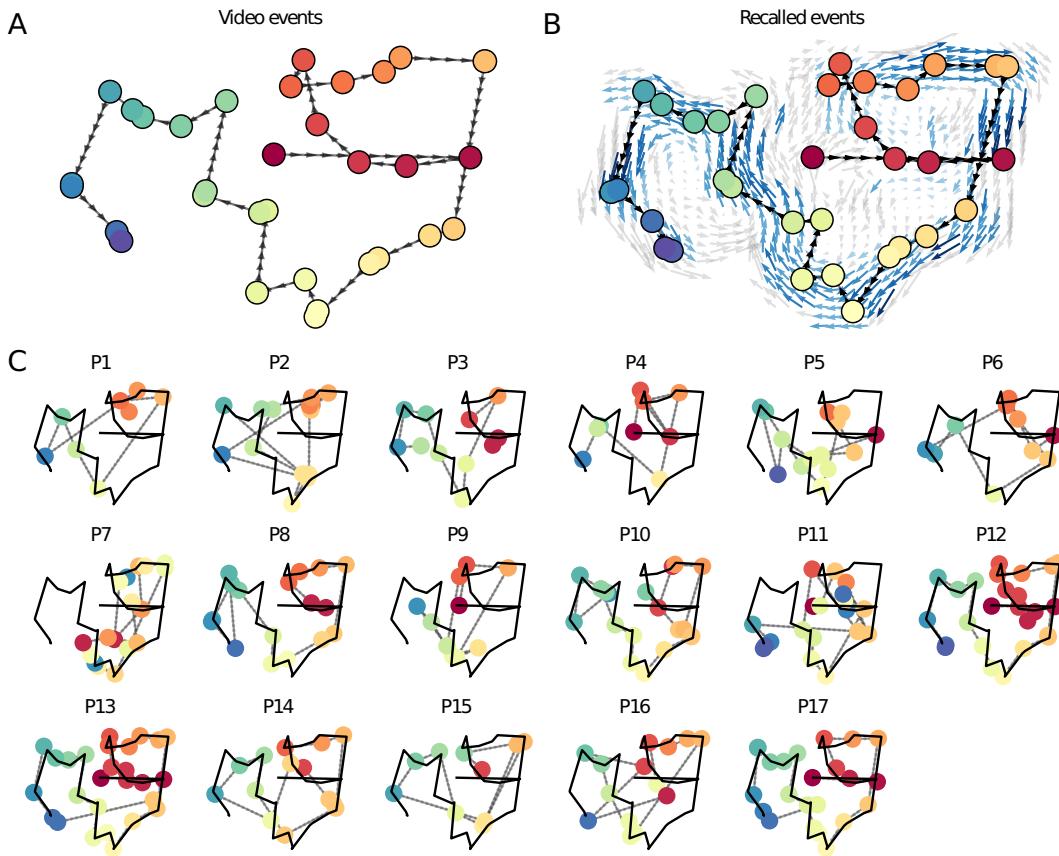
The prior analyses leverage the correspondence between the 100-dimensional topic proportion



**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 ( $k$ ).

229 matrices for the video and participants' recalls to characterize recall. However, it is difficult  
230 to gain deep insights into that content solely by examining the topic proportion matrices (e.g.,  
231 Figs. 2A, D) or the corresponding correlation matrices (Figs. 2B, E, S4). To visualize the time-  
232 varying high-dimensional content in a more intuitive way (Heusser et al., 2018b) we projected the  
233 topic proportions matrices onto a two-dimensional space using Uniform Manifold Approximation  
234 and Projection (UMAP; McInnes and Healy, 2018). In this lower-dimensional space, each point  
235 represents a single video or recall event, and the distances between the points reflect the distances  
236 between the events' associated topic vectors (Fig. 5). In other words, events that are near to each  
237 other in this space are more semantically similar.

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



**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.)

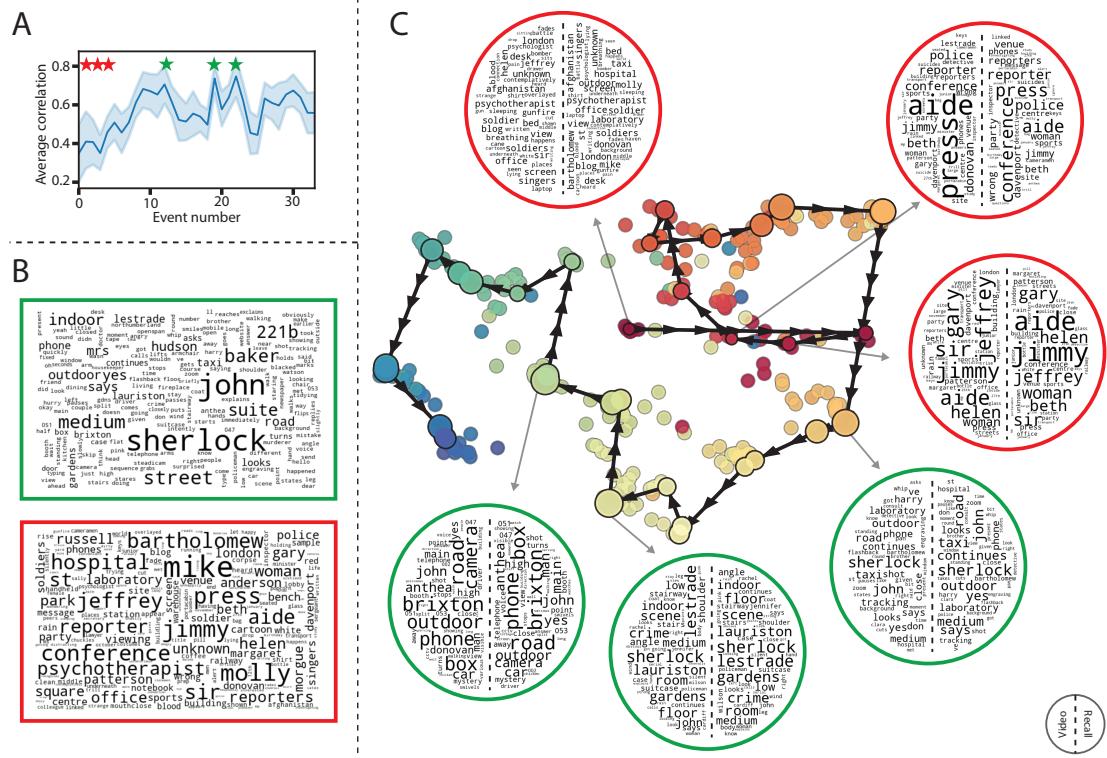
257 referenced in participants' recalls) from participants' abilities to recapitulate the full shape of the  
258 original video (i.e., the similarity in the shape of the original video trajectory and that defined by  
259 each participant's recounting of the video).

260 Because our analysis framework projects the dynamic video content and participants' recalls  
261 onto a shared topic space, and because the dimensions of that space are known (i.e., each topic  
262 dimension is a set of weights over words in the vocabulary; Fig. S2), we can examine the topic  
263 trajectories to understand which specific content was remembered well (or poorly). For each video  
264 event, we can ask: what was the average correlation (across participants) between the video event's  
265 topic vector and the closest matching recall event topic vectors from each participant? This yields a  
266 single correlation coefficient for each video event, describing how closely participants' recalls of the  
267 event tended to reliably capture its content (Fig. 6A). (We also examined how different comparisons  
268 between each video event's topic vector and the corresponding recall event topic vectors related  
269 to hand-annotated characterizations of memory performance; see *Supporting Information*). Given  
270 this summary of which events were recalled reliably (or not), we next asked whether the better-  
271 remembered or worse-remembered events tended to reflect particular topics. We computed a  
272 weighted average of the topic vectors for each video event, where the weights reflected how reliably  
273 each event was recalled. To visualize the result, we created a "wordle" image (Mueller et al., 2018)  
274 where words weighted more heavily by better-remembered topics appear in a larger font (Fig. 6B,  
275 green box). Events that reflected topics weighting heavily on characters like "Sherlock" and "John"  
276 (i.e., the main characters) and locations like "221b Baker Street" (i.e., a major recurring location and  
277 the address of the flat that Sherlock and John share) were best remembered. An analogous analysis  
278 revealed which themes were poorly remembered. Here in computing the weighted average over  
279 events' topic vectors we weighted each event in *inverse* proportion to how well it was remembered  
280 (Fig. 6B, red box). This revealed that events with relatively minor characters such as "Mike,"  
281 "Jeffrey," and "Molly," as well as less-integral plot locations (e.g., "hospital" and "office") were  
282 least well-remembered. This suggests that what is retained in memory are the major plot elements  
283 (i.e., the overall shape of what happened), whereas the more minor details are prone to pruning.

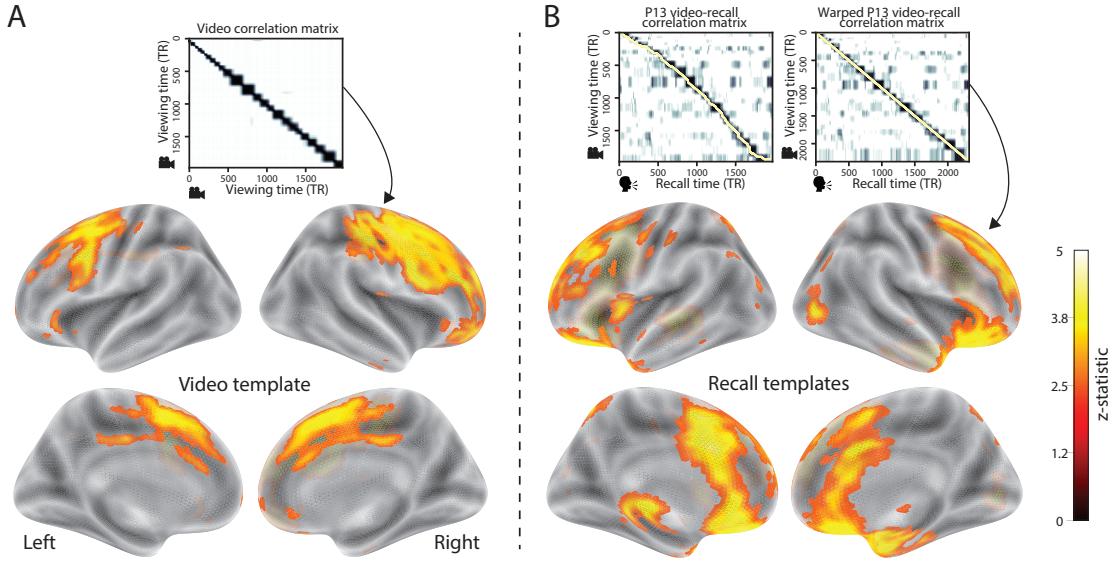
284 In addition to constructing overall summaries, assessing the video and recall topic vectors from

285 individual recalls can provide further insights. Specifically, for any given event we can construct  
286 two wordles: one from the original video event's topic vector, and a second from the average topic  
287 vectors produced by all participants' recalls of that event. We can then examine those wordles  
288 visually to gain an intuition for which aspects of the video event were recapitulated in participants'  
289 recalls of that event. Several example wordles are displayed in Figure 6C (wordles from the three  
290 best-remembered events are circled in green; wordles from the three worst-remembered events  
291 are circled in red). Using wordles to visually compare the topical content of each video event and  
292 the (average) corresponding recall event reveals the specific content from the specific events that  
293 is reliably retained in the transformation into memory (green events) or not (red events).

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



**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.



**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.

313 memory analyses (e.g., Paller and Wagner, 2002) to domain of naturalistic stimuli.

## 314 Discussion

315 Our work casts remembering as reproducing (behaviorally and neurally) the topic trajectory, or  
 316 shape, of the original experience. This view draws inspiration from prior work aimed at elucidating  
 317 the neural and behavioral underpinnings of how we process dynamic naturalistic experiences  
 318 and remember them later. One approach to identifying neural responses to naturalistic stimuli  
 319 (including experiences) entails building a model of the stimulus and searching for brain regions  
 320 whose responses are consistent with the model. In prior work, a series of studies from Uri  
 321 Hasson's group (Lerner et al., 2011; Simony et al., 2016; Chen et al., 2017; Baldassano et al., 2017;  
 322 Zadbood et al., 2017) have extended this approach with a clever twist. Rather than building an

323 explicit stimulus model, these studies instead search for brain responses (while experiencing the  
324 stimulus) that are reliably similar across individuals. So called *inter-subject correlation* (ISC) and  
325 *inter-subject functional connectivity* (ISFC) analyses effectively treat other people's brain responses  
326 to the stimulus as a "model" of how its features change over time. By contrast, in our present  
327 work we used topic models and HMMs to construct an explicit stimulus model (i.e., the topic  
328 trajectory of the video). When we searched for brain structures whose responses are consistent  
329 with the video's topic trajectory, we identified a network of structures that overlapped strongly  
330 with the "long temporal receptive window" network reported by the Hasson group (e.g., compare  
331 our Fig. 7A with the map of long temporal receptive window voxels in Lerner et al., 2011). This  
332 provides support for the notion that part of the long temporal receptive window network may be  
333 maintaining an explicit model of the stimulus dynamics. When we performed a similar analysis  
334 after swapping out the video's topic trajectory with the recall topic trajectories of each individual  
335 participant, this allowed us to identify brain regions whose responses (as the participants viewed  
336 the video) reflected how the video trajectory would be transformed in memory (as reflected by  
337 the recall topic trajectories). The analysis revealed that the rMTL and vmPFC may play a role in  
338 this person-specific transformation from experience into memory. The role of the MTL in episodic  
339 memory encoding has been well-reported (e.g., Paller and Wagner, 2002; Davachi et al., 2003;  
340 Ranganath et al., 2004; Davachi, 2006). Prior work has also implicated the medial prefrontal cortex  
341 in representing "schema" knowledge (i.e., general knowledge about the format of an ongoing  
342 experience given prior similar experiences; van Kesteren et al., 2012; Schlichting and Preston,  
343 2015; Gilboa and Marlatte, 2017; Spalding et al., 2018). Integrating across our study and this prior  
344 work, one interpretation is that the person-specific transformations mediated (or represented)  
345 by the rMTL and vmPFC may reflect schema knowledge being leveraged, formed, or updated,  
346 incorporating ongoing experience into previously acquired knowledge.

347 When modeling memory experiments, often times events (or items) and their later memories  
348 are treated as binary (or categorical in the case of confidence ratings) and independent events.  
349 Our novel framework allows one to assess memory performance in a more continuous way (e.g.  
350 precision), as well as analyze the correlational structure of each encoding event to each memory

351 event (e.g. distinctiveness). Further and importantly, it allows for consideration of the actual  
352 content of the experience/memories, which is not typically modeled. Leveraging this, using 2 novel  
353 memory metrics we find that the successful memory performance is related to 1) the *precision* with  
354 which the participant recounts each event and 2) how *distinctive* each recall event is (relative to the  
355 other recalled events). The first finding suggests to us that the accuracy of recall for *any individual*  
356 *event* may predict the overall amount of information recovered by the participant. The second  
357 finding suggests that remembering/describing events in a unique way (relative to other recalled  
358 events) is also related to the quantity of content recovered. Intriguingly, prior studies show that  
359 pattern separation, or the ability to discriminate between similar experiences, is impaired in many  
360 cognitive disorders as well as natural aging (Stark et al., 2010; Yassa et al., 2011; Yassa and Stark,  
361 2011). Future work might explore how/whether the novel metrics introduce here compare between  
362 cognitively impoverished groups and healthy controls.

363 While a large number of language models exist (e.g. WAS, LSA, word2vec, universal sentence  
364 encoder) (Landauer et al., 1998; Cer et al., 2018), here we use topic models for a few reasons. First,  
365 topic models capture the *essence* of a text passage devoid of the specific set and order of words  
366 used. This was an important feature of our model since different people may accurately recall  
367 a scene using very different language. Secondly, words can mean different things in different  
368 contexts (e.g. baseball bat vs. the animal bat), and topic models are robust to this since words can  
369 be part of multiple topics. Lastly, topic models provide a straight forward to recover the weights  
370 for the particular words comprising a topic, allowing for easy interpretation of an event's contents  
371 (e.g. Fig. 6). Other models such as Google's universal sentence encoder offer a context-sensitive  
372 encoding of text passages, but the encoding space is complex and non-linear and thus, recovering  
373 the original words used to fit the model is not straight forward. However, it's worth pointing out  
374 that our framework is divorced from the particular choice of language model. Moreover, many of  
375 the aspects of our framework could be swapped out for other choices. For example, the language  
376 model, the timeseries segmentation model and the video-recall matching function could all be  
377 customized for the particular problem. Indeed for some problems, recovery of the particular recall  
378 words may not be necessary, and thus other text-modeling approaches (such as universal sentence

379 encoder) may be preferable. Future work will explore the influence of particular model choices on  
380 the framework's accuracy.

381 Our work has broad implications for how we characterize and assess memory in real-world set-  
382 tings such as the classroom or physician's office. For example, the most commonly used classroom  
383 evaluation tools involve computing the proportion of correctly answered exam questions. Our  
384 work indicates that this approach is only loosely related to what educators might really want to  
385 measure: how well did the students understand the key ideas presented in the course? One could  
386 apply the computational framework we developed to construct topic trajectories for the video and  
387 participants' recalls to build explicit content models of the course material and exam questions.  
388 This approach would provide a more nuanced and specific view into which aspects of the material  
389 students had learned well (or poorly). In clinical settings, memory measures that incorporate such  
390 explicit content models might also provide more direct evaluations of patients' memories.

## 391 Methods

### 392 Experimental design and data collection

393 Data were collected by Chen et al. (2017). In brief, participants ( $n = 17$ ) viewed the first 48 minutes  
394 of "A Study in Pink", the first episode of the BBC television series *Sherlock*, while fMRI volumes  
395 were collected (TR = 1500 ms). The stimulus was divided into a 23 min (946 TR) and a 25 min  
396 (1030 TR) segment to mitigate technical issues related to the scanner. After finishing the clip,  
397 participants were instructed to (quoting from Chen et al., 2017) "describe what they recalled of the  
398 [episode] in as much detail as they could, to try to recount events in the original order they were  
399 viewed in, and to speak for at least 10 minutes if possible but that longer was better. They were told  
400 that completeness and detail were more important than temporal order, and that if at any point  
401 they realized they had missed something, to return to it. Participants were then allowed to speak  
402 for as long as they wished, and verbally indicated when they were finished (e.g., 'I'm done')."  
403 For additional details about the experimental procedure and scanning parameters see Chen et al.

404 (2017). The experimental protocol was approved by Princeton University's Institutional Review  
405 Board.

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

## 412 **Data and code availability**

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

## 415 **Statistics**

416 All statistical tests we performed were two-sided.

## 417 **Modeling the dynamic content of the video and recall transcripts**

### 418 **Topic modeling**

419 The input to the topic model we trained to characterize the dynamic content of the video comprised  
420 hand-generated annotations of each of 1000 scenes spanning the video clip (generated by Chen  
421 et al., 2017). The features included: narrative details (a sentence or two describing what happened  
422 in that scene); whether the scene took place indoors or outdoors; names of any characters that  
423 appeared in the scene; name(s) of characters in camera focus; name(s) of characters who were  
424 speaking in the scene; the location (in the story) that the scene took place; camera angle (close  
425 up, medium, long, top, tracking, over the shoulder, etc.); whether music was playing in the  
426 scene or not; and a transcription of any on-screen text. We concatenated the text for all of these

427 features within each segment, creating a “bag of words” describing each scene. We then re-  
428 organized the text descriptions into overlapping sliding windows spanning 50 scenes each. In  
429 other words, the first text sample comprised the combined text from the first 50 scenes (i.e., 1–50),  
430 the second comprised the text from scenes 2–51, and so on. We trained our model using these  
431 overlapping text samples with `scikit-learn` (version 0.19.1; Pedregosa et al., 2011), called from  
432 our high-dimensional visualization and text analysis software, `HyperTools` (Heusser et al., 2018b).  
433 Specifically, we use the `CountVectorizer` class to transform the text from each scene into a vector of  
434 word counts (using the union of all words across all scenes as the “vocabulary,” excluding English  
435 stop words); this yields a number-of-scenes by number-of-words *word count* matrix. We then  
436 use the `LatentDirichletAllocation` class (`topics=100, method='batch'`) to fit a topic model (Blei  
437 et al., 2003) to the word count matrix, yielding a number-of-scenes (1000) by number-of-topics  
438 (100) *topic proportions* matrix. The topic proportions matrix describes which mix of topics (latent  
439 themes) is present in each scene. Next, we transformed the topic proportions matrix to match the  
440 1976 fMRI volume acquisition times. For each fMRI volume, we took the topic proportions from  
441 whatever scene was displayed for most of that volume’s 1500 ms acquisition time. This yielded a  
442 new number-of-TRs (1976) by number-of-topics (100) topic proportions matrix.

443 We created similar topic proportions matrices using hand-annotated transcripts of each partici-  
444 pant’s recall of the video (annotated by Chen et al., 2017). We tokenized the transcript into a list of  
445 sentences, and then re-organized the list into overlapping sliding windows spanning 10 sentences  
446 each; in turn we transformed each window’s sentences into a word count vector (using the same  
447 vocabulary as for the video model). We then used the topic model already trained on the video  
448 scenes to compute the most probable topic proportions for each sliding window. This yielded a  
449 number-of-sentences (range: 68–294) by number-of-topics (100) topic proportions matrix, for each  
450 participant. These reflected the dynamic content of each participant’s recalls. Note: for details  
451 on how we selected the video and recall window lengths and number of topics, see *Supporting*  
452 *Information* and Figure S1.

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

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

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

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

463 where  $a$  was the average correlation between the topic vectors of timepoints within the same state;  
464  $b$  was the average correlation between the topic vectors of timepoints within *different* states; and  
465  $\alpha$  was a regularization parameter that we set to 5 times the window length (i.e., 250 scenes for  
466 the video topic trajectory and 50 sentences for the recall topic trajectories). Figure 2B displays the  
467 event boundaries returned for the video, and Figure S4 displays the event boundaries returned  
468 for each participant's recalls. After obtaining these event boundaries, we created stable estimates  
469 of each topic proportions matrix by averaging the topic vectors within each event. This yielded a  
470 number-of-events by number-of-topics matrix for the video and recalls from each participant.

471 We also evaluated a parameter-free procedure for choosing  $K$ , which finds the  $K$  value that  
472 maximizes the Wasserstein distance (a.k.a. "Earth mover's" distance) between the within and  
473 across event distributions of correlation values. This alternative procedure largely replicated the  
474 pattern of results found with the parameterized method described above, but recovered sub-  
475 stantially fewer events on average (Fig.S6). While both approaches seem to underestimate the  
476 number of video/recall events relative to the "true" number (as determined by human raters), the  
477 parameterized approach was closer to the true number.

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

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

485 Perhaps the simplest and most widely used measure of memory performance is *accuracy*— i.e.,  
486 the proportion of studied (experienced) items (in this case, the 34 video events) that the participant  
487 later remembered. Chen et al. (2017) developed a human rating system whereby the quality of  
488 each participant’s memory was evaluated by an independent rater. We found a strong across-  
489 participants correlation between these independant ratings and the overall number of events that  
490 our HMM approach identified in participants’ recalls (Pearson’s  $r(15) = 0.67, p = 0.003$ ).

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

494 **Probability of first recall (PFR).** PFR curves (Welch and Burnett, 1924; Postman and Phillips,  
495 1965; Atkinson and Shiffrin, 1968) reflect the probability that an item will be recalled first as a  
496 function of its serial position during encoding. To carry out this analysis, we initialized a number-  
497 of-participants (17) by number-of-video-events (34) matrix of zeros. Then for each participant, we  
498 found the index of the video event that was recalled first (i.e., the video event whose topic vector  
499 was most strongly correlated with that of the first recall event) and filled in that index in the matrix  
500 with a 1. Finally, we averaged over the rows of the matrix, resulting in a 1 by 34 array representing  
501 the proportion of participants that recalled an event first, as a function of the order of the event’s  
502 appearance in the video (Fig. 3A).

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

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

522 **Temporal clustering scores.** Temporal clustering refers to the extent to which participants group  
523 their recall responses according to encoding position (Polyn et al., 2009). For instance, if a par-  
524 ticipant recalled the video events in the exact order they occurred (or in exact reverse order), this  
525 would yield a score of 1. If a participant recalled the events in random order, this would yield  
526 an expected score of 0.5. For each recall event transition (and separately for each participant), we  
527 sorted all not-yet-recalled events according to their absolute lag (i.e., distance away in the video).  
528 We then computed the percentile rank of the next event the participant recalled. We averaged

529 these percentile ranks across all of the participant's recalls to obtain a single temporal clustering  
530 score for the participant (mean: 0.808, SEM: 0.022). Overall, we found that participants with higher  
531 temporal clustering scores also tended to recall more events (Pearson's  $r(15) = 0.62, p = 0.007$ ).

532 **Semantic clustering scores.** Semantic clustering measures the extent to which participants clus-  
533 tered their recall responses according to semantic similarity (Polyn et al., 2009). Here, we used the  
534 topic vectors for each event as a proxy for its semantic content. Thus, the similarity between the  
535 semantic content for two events can be computed by correlating their respective topic vectors. For  
536 each recall event transition, we sorted all not-yet-recalled events according to how correlated the  
537 topic vector of *the closest-matching video event* was to the topic vector of the closest-matching video  
538 event to the just-recalled event. We then computed the percentile rank of the observed next recall.  
539 We averaged these percentile ranks across all of the participant's recalls to obtain a single semantic  
540 clustering score for the participant (mean: 0.813, SEM: 0.022). We found that participants who  
541 exhibited stronger semantic clustering scores overall remembered more video events (Pearson's  
542  $r(15) = 0.55, p = 0.02$ ).

543 **Novel naturalistic memory metrics**

544 **Precision.** We tested whether participants who recalled more events were also more *precise* in their  
545 recollections. For each participant, we computed the correlation between the topic vectors for each  
546 recall event and that of its closest-matching video event (only for the events which they recalled).  
547 We Fisher's z-transformed the correlations, computed the average and then inverse Fisher's z-  
548 transformed the resulting value. This gave a single value per participant representing the average  
549 precision across all recalled events. We then correlated this value with hand-annotated as well as  
550 model derived (e.g.  $k$  or the number of events recovered by the HMM) memory performance.

551 **Distinctiveness.** We also considered the *distinctiveness* of each recalled event. That is, how  
552 uniquely a recalled event's topic vector matched a given video event topic vector, versus the  
553 topic vectors for the other video events. We hypothesized that participants with high memory

554 performance might describe each event in a more distinctive way (relative to those with lower  
555 memory performance who might describe events in a more general way). To test this hypothesis  
556 we define a distinctiveness score for each recalled event as

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

557 where  $\bar{c}(\text{event})$  is the average correlation between the given recalled event's topic vector and the  
558 topic vectors from all video events *except* the best-matching video event. We then averaged these  
559 distinctiveness scores across all of the events recalled by the given participant. As above, we used  
560 Fisher's z (transform and inverse-transform) before/after averaging correlation values. Finally,  
561 we correlated these values with hand-annotated and model derived memory performance scores  
562 across-subjects.

### 563 **Visualizing the video and recall topic trajectories**

564 We used the UMAP algorithm (McInnes and Healy, 2018) to project the 100-dimensional topic space  
565 onto a two-dimensional space for visualization (Figs. 5, 6). To ensure that all of the trajectories were  
566 projected onto the *same* lower dimensional space, we computed the low-dimensional embedding  
567 on a "stacked" matrix created by vertically concatenating the events-by-topics topic proportions  
568 matrices for the video and all 17 participants' recalls. We then divided the rows of the result (a  
569 total-number-of-events by two matrix) back into separate matrices for the video topic trajectory  
570 and the trajectories for each participant's recalls (Fig. 5). This general approach for discovering  
571 a shared low-dimensional embedding for a collections of high-dimensional observations follows  
572 Heusser et al. (2018b).

### 573 **Estimating the consistency of flow through topic space across participants**

574 In Figure 5B, we present an analysis aimed at characterizing locations in topic space that dif-  
575 ferent participants move through in a consistent way (via their recall topic trajectories). The  
576 two-dimensional topic space used in our visualizations (Fig. 5) ranged from -5 to 5 (arbitrary) units

577 in the  $x$  dimension and from -6.5 to 2 units in the  $y$  dimension. We divided this space into a grid  
578 of vertices spaced 0.25 units apart. For each vertex, we examined the set of line segments formed  
579 by connecting each pair successively recalled events, across all participants, that passed within 0.5  
580 units. We computed the distribution of angles formed by those segments and the  $x$ -axis, and used a  
581 Rayleigh test to determine whether the distribution of angles was reliably “peaked” (i.e., consistent  
582 across all transitions that passed through that local portion of topic space). To create Figure 5B we  
583 drew an arrow originating from each grid vertex, pointing in the direction of the average angle  
584 formed by line segments that passed within 0.5 units. We set the arrow lengths to be inversely  
585 proportional to the  $p$ -values of the Rayleigh tests at each vertex. Specifically, for each vertex we  
586 converted all of the angles of segments that passed within 0.5 units to unit vectors, and we set  
587 the arrow lengths at each vertex proportional to the length of the (circular) mean vector. We also  
588 indicated any significant results ( $p < 0.05$ , corrected using the Benjamani-Hochberg procedure) by  
589 coloring the arrows in blue (darker blue denotes a lower  $p$ -value, i.e., a longer mean vector); all  
590 tests with  $p \geq 0.05$  are displayed in gray and given a lower opacity value.

## 591 **Searchlight fMRI analyses**

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

599 Next, we constructed two sets of “template” matrices: one reflected the video’s topic trajectory  
600 and the other reflected each participant’s recall topic trajectory. To construct the video template, we  
601 computed the correlations between the topic proportions estimated for every pair of TRs (prior to  
602 segmenting the trajectory into discrete events; i.e., the correlation matrix shown in Figs. 2B and 7A).  
603 We constructed similar temporal correlation matrices for each participant’s recall topic trajectory

604 (Figs. 2D, S4). However, to correct for length differences and potential non-linear transformations  
605 between viewing time and recall time, we first used dynamic time warping (Berndt and Clifford,  
606 1994) to temporally align participants' recall topic trajectories with the video topic trajectory (an  
607 example correlation matrix before and after warping is shown in Fig. 7B). This yielded a 1976 by  
608 1976 correlation matrix for the video template and for each participant's recall template.

609 To determine which (cubes of) voxel responses reliably matched the video template, we cor-  
610 related the upper triangle of the voxel correlation matrix for each cube with the upper triangle  
611 of the video template matrix (Kriegeskorte et al., 2008). This yielded, for each participant, a  
612 single correlation value. We computed the average (Fisher  $z$ -transformed) correlation coefficient  
613 across participants. We used a permutation-based procedure to assess significance, whereby we  
614 re-computed the average correlations for each of 100 "null" video templates (constructed by circu-  
615 larly shifting the template by a random number of timepoints). (For each permutation, the same  
616 shift was used for all participants.) We then estimated a  $p$ -value by computing the proportion of  
617 shifted correlations that were larger than the observed (unshifted) correlation. To create the map  
618 in Figure 7A we thresholded out any voxels whose correlation values fell below the 95<sup>th</sup> percentile  
619 of the permutation-derived null distribution.

620 We used a similar procedure to identify which voxels' responses reflected the recall templates.  
621 For each participant, we correlated the upper triangle of the correlation matrix for each cube of  
622 voxels with their (time warped) recall correlation matrix. As in the video template analysis this  
623 yielded a single correlation coefficient for each participant. However, whereas the video analysis  
624 compared every participant's responses to the same template, here the recall templates were  
625 unique for each participant. We computed the average  $z$ -transformed correlation coefficient across  
626 participants, and used the same permutation procedure we developed for the video responses to  
627 assess significant correlations. To create the map in Figure 7B we thresholded out any voxels whose  
628 correlation values fell below the 95<sup>th</sup> percentile of the permutation-derived null distribution.

629 **References**

- 630 Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control  
631 processes. In Spence, K. W. and Spence, J. T., editors, *The psychology of learning and motivation*,  
632 volume 2, pages 89–105. Academic Press, New York.
- 633 Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017).  
634 Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–  
635 721.
- 636 Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In  
637 *KDD workshop*, volume 10, pages 359–370.
- 638 Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International  
639 Conference on Machine Learning*, ICML ’06, pages 113–120, New York, NY, US. ACM.
- 640 Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine  
641 Learning Research*, 3:993 – 1022.
- 642 Brunec, I. K., Moscovitch, M. M., and Barense, M. D. (2018). Boundaries shape cognitive represen-  
643 tations of spaces and events. *Trends in Cognitive Sciences*, 22(7):637–650.
- 644 Capota, M., Turek, J., Chen, P.-H., Zhu, X., Manning, J. R., Sundaram, N., Keller, B., Wang, Y., and  
645 Shin, Y. S. (2017). Brain imaging analysis kit.
- 646 Cer, D., Yang, Y., y Kong, S., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes,  
647 M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.  
648 *arXiv*, 1803.11175.
- 649 Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared  
650 memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*,  
651 20(1):115.
- 652 Clewett, D. and Davachi, L. (2017). The ebb and flow of experience determines the temporal  
653 structure of memory. *Curr Opin Behav Sci*, 17:186–193.

- 654 Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in*  
655 *Neurobiology*, 16(6):693—700.
- 656 Davachi, L., Mitchell, J. P., and Wagner, A. D. (2003). Multiple routes to memory: distinct medial  
657 temporal lobe processes build item and source memories. *Proceedings of the National Academy of*  
658 *Sciences, USA*, 100(4):2157 – 2162.
- 659 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the  
660 sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–1286.
- 661 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological*  
662 *Science*, 22(2):243–252.
- 663 Gilboa, A. and Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory.  
664 *Trends Cogn Sci*, 21(8):618–631.
- 665 Heusser, A. C., Ezzyat, Y., Shiff, I., and Davachi, L. (2018a). Perceptual boundaries cause mnemonic  
666 trade-offs between local boundary processing and across-trial associative binding. *Journal of*  
667 *Experimental Psychology Learning, Memory, and Cognition*, 44(7):1075–1090.
- 668 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail: a  
669 Python toolbox for analyzing and plotting free recall data. *The Journal of Open Source Software*,  
670 10.21105/joss.00424.
- 671 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018b). HyperTools: a Python  
672 toolbox for gaining geometric insights into high-dimensional data. *Journal of Machine Learning*  
673 *Research*, 18(152):1–6.
- 674 Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal*  
675 *of Mathematical Psychology*, 46:269–299.
- 676 Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., and H., E.  
677 (2014). A unified mathematical framework for coding time, space, and sequences in the medial  
678 temporal lobe. *Journal of Neuroscience*, 34(13):4692–4707.

- 679 Howard, M. W., Viskontas, I. V., Shankar, K. H., and Fried, I. (2012). Ensembles of human MTL  
680 neurons “jump back in time” in response to a repeated stimulus. *Hippocampus*, 22:1833–1847.
- 681 Huk, A., Bonnen, K., and He, B. J. (2018). Beyond trial-based paradigms: continuous behavior, on-  
682 going neural activity, and naturalistic stimuli. *Journal of Neuroscience*, 10.1523/JNEUROSCI.1920-  
683 17.2018.
- 684 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24:103–109.
- 685 Koriat, A. and Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: distin-  
686 guishing accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of*  
687 *Experimental Psychology: General*, 123(3):297–315.
- 688 Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – con-  
689 nnecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:1 – 28.
- 690 Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to latent semantic analysis.  
691 *Discourse Processes*, 25:259–284.
- 692 Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy  
693 of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.
- 694 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory.  
695 In Gazzaniga, M., editor, *The Cognitive Neurosciences, Fifth edition*, pages 557–566. MIT Press.
- 696 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns  
697 in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National*  
698 *Academy of Sciences, USA*, 108(31):12893–12897.
- 699 McInnes, L. and Healy, J. (2018). UMAP: Uniform manifold approximation and projection for  
700 dimension reduction. *arXiv*, 1802(03426).
- 701 Mueller, A., Fillion-Robin, J.-C., Boidol, R., Tian, F., Nechifor, P., yoonsubKim, Peter, Rampin, R.,  
702 Corvellec, M., Medina, J., Dai, Y., Petrushev, B., Langner, K. M., Hong, Alessio, Ozsvald, I.,

- 703 vkolmakov, Jones, T., Bailey, E., Rho, V., IgorAPM, Roy, D., May, C., foobuzz, Piyush, Seong,  
704 L. K., Goey, J. V., Smith, J. S., Gus, and Mai, F. (2018). WordCloud 1.5.0: a little word cloud  
705 generator in Python. *Zenodo*, <https://zenodo.org/record/1322068#.W4tPKZNKh24>.
- 706 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*,  
707 64:482–488.
- 708 Paller, K. A. and Wagner, A. D. (2002). Observing the transformation of experience into memory.  
709 *Trends in Cognitive Sciences*, 6(2):93–102.
- 710 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten-  
711 hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot,  
712 M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine  
713 Learning Research*, 12:2825–2830.
- 714 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model  
715 of organizational processes in free recall. *Psychological Review*, 116(1):129–156.
- 716 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal  
717 of Experimental Psychology*, 17:132–138.
- 718 Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech  
719 recognition. *Proceedings of the IEEE*, 77(2):257–286.
- 720 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition. *Curr Opin  
721 Behav Sci*, 17:133–140.
- 722 Ranganath, C., Cohen, M. X., Dam, C., and D’Esposito, M. (2004). Inferior temporal, prefrontal,  
723 and hippocampal contributions to visual working memory maintenance and associative memory  
724 retrieval. *Journal of Neuroscience*, 24(16):3917–3925.
- 725 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature  
726 Reviews Neuroscience*, 13:713 – 726.

- 727 Schlichting, M. L. and Preston, A. R. (2015). Memory integration: neural mechanisms and impli-  
728 cations for behavior. *Current Opinion in Behavioral Sciences*, 1:1–8.
- 729 Simony, E., Honey, C. J., Chen, J., and Hasson, U. (2016). Uncovering stimulus-locked network  
730 dynamics during narrative comprehension. *Nature Communications*, 7(12141):1–13.
- 731 Spalding, K. N., Schlichting, M. L., Zeithamova, D., Preston, A. R., Tranel, D., Duff, M. C., and  
732 Warren, D. E. (2018). Ventromedial prefrontal cortex is necessary for normal associative inference  
733 and memory integration. *The Journal of Neuroscience*, 38(15):3767–3775.
- 734 Stark, S. M., Yassa, M. A., and Stark, C. E. L. (2010). Individual differences in spatial pattern  
735 separation performance associated with healthy aging in humans. *Learning & Memory*, 17(6):284–  
736 288.
- 737 Tompry, A. and Davachi, L. (2017). Consolidation promotes the emergence of representational  
738 overlap in the hippocampus and medial prefrontal cortex. *Neuron*, 96(1):228–241.
- 739 van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., and Henson, R. N. (2012). How schema and  
740 novelty augment memory formation. *Trends Neurosci*, 35(4):211–9.
- 741 Waskom, M., Botvinnik, O., Okane, D., Hobson, P., David, Halchenko, Y., Lukauskas, S., Cole, J. B.,  
742 Warmenhoven, J., de Ruiter, J., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E.,  
743 Martin, M., Miles, A., Meyer, K., Augspurger, T., Yarkoni, T., Bachant, P., Williams, M., Evans,  
744 C., Fitzgerald, C., Brian, Wehner, D., Hitz, G., Ziegler, E., Qalieh, A., and Lee, A. (2016). Seaborn:  
745 v0.7.1.
- 746 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American Journal  
747 of Psychology*, 35:396–401.
- 748 Yassa, M. A., Lacy, J. W., Stark, S. M., Albert, M. S., Gallagher, M., and Stark, C. E. L. (2011). Pattern  
749 separation deficits associated with increased hippocampal ca3 and dentate gyrus activity in  
750 nondemented older adults. *Hippocampus*, 21(9):968–979.

- 751 Yassa, M. A. and Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends In Neuro-*  
752 *sciences*, 34(10):515–525.
- 753 Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research.  
754 *Journal of Memory and Language*, 46:441–517.
- 755 Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception:  
756 a mind-brain perspective. *Psychological Bulletin*, 133:273–293.
- 757 Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit  
758 memories to other brains: Constructing shared neural representations via communication. *Cereb*  
759 *Cortex*, 27(10):4988–5000.
- 760 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and  
761 memory. *Psychological Bulletin*, 123(2):162 – 185.

## 762 **Supporting information**

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

## 764 **Acknowledgements**

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

<sup>770</sup> **Author contributions**

<sup>771</sup> Conceptualization: A.C.H. and J.R.M.; Methodology: A.C.H. and J.R.M.; Software: A.C.H., P.C.F.  
<sup>772</sup> 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.  
<sup>773</sup> and J.R.M.; Supervision: J.R.M.

<sup>774</sup> **Author information**

<sup>775</sup> The authors declare no competing financial interests. Correspondence and requests for materials  
<sup>776</sup> should be addressed to J.R.M. (jeremy.r.manning@dartmouth.edu).