

1 How is experience transformed into memory? Geometric

2 models reveal behavioral and neural signatures of

3 transforming naturalistic experiences into episodic

4 memories

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

¹Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

²Akili Interactive Labs

Boston, MA 02110

†Denotes equal contribution

*Corresponding author: jeremy.r.manningJeremy.R.Manning@dartmouthDartmouth.edu

6 September 4, 2020

Abstract

The ways our experiences unfold over time define unique mental contexts in which we interpret experiences are often person-specific, even when the experiences themselves are shared. We developed a geometric framework for mathematically characterizing the subjective conceptual content of dynamic naturalistic experiences. We model experiences and memories as trajectories through 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 they unfold over time. We apply this approach to a naturalistic memory experiment which had participants view and recount a video. We found that the shapes of the trajectories formed by participants' recounts were all highly similar to that of the original video, but participants differed in the level of detail they remembered through word embedding spaces whose coordinates reflect the universe of thoughts under consideration. Memory encoding can then be modeled as geometrically preserving or distorting the shape of the original experience. We applied our approach to data collected as participants watched and verbally recounted a television episode while undergoing functional neuroimaging. Participants' recounts all preserved coarse spatial properties (essential narrative elements), but not fine spatial scale (low-level) details, of the episode's trajectory. We also identified a network of brain structures that are sensitive to the "shapes" of these trajectory shapes. Our work provides insights into how we preserve and distort our ongoing experiences, and an overlapping network that is sensitive to how we will later remember those experiences when we encode them into episodic memories.

Introduction

30 What does it mean to *remember* something? In traditional episodic memory experiments (e.g.,
31 list-learning or trial-based experiments; Kahana, 1996; Murdock, 1962), remembering is often
32 cast as a discrete ~~and~~^{and} binary operation: each studied item may be separated from the rest
33 of one's ~~experiences, and that item may be~~ experience and labeled as having been ~~recalled~~
34 ~~versus~~ either recalled or forgotten. More nuanced studies might incorporate self-reported con-

35 fidence measures as a proxy for memory strength, or ask participants to discriminate between
36 “recollecting” recollecting the (contextual) details of an experience or and having a general feeling
37 of “familiarity” (Yonelinas, 2002). familiarity (Yonelinas, 2002). Using well-controlled, trial-based
38 experimental designs, the field has amassed a wealth of information regarding human episodic
39 memory (for review see Kahana, 2012). However, characterizing and evaluating memory in more
40 realistic contexts (e.g., recounting a recent experience to a friend) is fundamentally different
41 in at least three ways (for review also see Huk et al., 2018; Koriat and Goldsmith, 1994) there are
42 fundamental properties of the external world and our memories that trial-based experiments are
43 not well suited to capture (for review, also see Huk et al., 2018; Koriat and Goldsmith, 1994). First,
44 real-world recall is our experiences and memories are continuous, rather than discrete. Unlike in
45 trial-based experiments, removing a (naturalistic) discrete—isolating a naturalistic event from the
46 context in which it occurs can substantially change its meaning. Second, the specific words used to
47 describe an experience have little bearing on whether the experience should be considered to have
48 been “remembered.” Asking whether the whether or not the rememberer has precisely reproduced
49 a specific set of words to describe in describing a given experience is nearly orthogonal to whether
50 how well they were actually able to remember it. In classic (e.g., list-learning) memory studies,
51 by contrast, counting the number or proportion of precise exact recalls is often considered to be a
52 primary metric of for assessing the quality of participants’ memories. Third, one might remember
53 the gist or essence essence (or a general summary) of an experience but forget (or neglect to recount)
54 particular low-level details. Capturing the gist essence of what happened is typically the main
55 “point” of recounting a often a main goal of recounting an episodic memory to a listenerwhereas,
56 depending on the circumstances, accurate recall of any specific detail may be irrelevant. There is
57 no analog of the gist of an experience in most traditional memory experiments; rather we tend to
58 assess participants’ abilities to recover specific details (e.g., computing the proportion of specific
59 stimuli they remember, which presentation positions the remembered stimuli came from, etc.).—
60 whereas the inclusion of specific low-level details is often less pertinent.

61 How might one go about formally characterizing the gist we formally characterize the essence
62 of an experience, or whether that gist and whether it has been recovered by the rememberer? Any

63 And how might we distinguish an experience's overarching essence from its low-level details?
64 One approach is to start by considering some fundamental properties of the dynamics of our
65 experiences. Each given moment of an experience ~~derives~~ tends to derive meaning from sur-
66 rounding moments, as well as from longer-range temporal associations (e.g., Lerner et al., 2011).
67 Therefore(Lerner et al., 2011; Manning, 2019, 2020). Therefore, the timecourse describing how an
68 event unfolds is fundamental to its overall meaning. Further, this hierarchy formed by our subjective
69 experiences at different timescales defines a *context* for each new moment (e.g., Howard et al., 2014; ?)
70 (e.g., Howard and Kahana, 2002; Howard et al., 2014), and plays an important role in how we inter-
71 pret that moment and remember it later (for review see Manning et al., 2015)(for review see Manning, 2020; Manning et
72 . Our memory systems can then leverage these associations to form predictions that help guide
73 our behaviors (Ranganath and Ritchey, 2012). For example, as we navigate the world, the features
74 of our subjective experiences tend to change gradually (e.g., the room or situation we are in find
75 ourselves in at any given moment is strongly temporally autocorrelated), allowing us to form
76 stable estimates of our current situation and behave accordingly (Zacks et al., 2007; Zwaan and
77 Radvansky, 1998). Although our experiences most often change gradually, they also occasionally
78 change suddenly (e.g., when we walk through a doorway; Radvansky and Zacks, 2017)

79 Occasionally, this gradual drift of our ongoing experience is punctuated by sudden changes, or
80 shifts (e.g., when we walk through a doorway; Radvansky and Zacks, 2017). Prior research sug-
81 gests that these sharp transitions (termed *event boundaries*) during an experience help to discretize
82 our experiences (and their mental representations) into events (Brunec et al., 2018; Clewett and
83 Davachi, 2017; DuBrow and Davachi, 2013; Ezzyat and Davachi, 2011; Heusser et al., 2018a; Rad-
84 vansky and Zacks, 2017). The interplay between the stable (within-eventwithin-event) and transient
85 (across-eventacross-event) temporal dynamics of an experience also provides a potential frame-
86 work for transforming experiences into memories that distill distills those experiences down to
87 their essence—i.e., their gists. essences. For example, prior work has shown that event boundaries
88 can influence how we learn sequences of items (DuBrow and Davachi, 2013; Heusser et al., 2018a),
89 navigate (Brunec et al., 2018), and remember and understand narratives (Ezzyat and Davachi, 2011;
90 Zwaan and Radvansky, 1998). This work also suggests a means of distinguishing the essence of an

91 experience from its low-level details: The overall structure of events and event transitions reflects
92 how the high-level experience unfolds (i.e., its essence), while subtler event-level properties reflect
93 its low-level details. Prior research has also implicated a network of brain regions (including the
94 hippocampus and the medial prefrontal cortex) in playing a critical role in transforming experiences
95 into structured and consolidated memories (Tompry and Davachi, 2017).

96 Here we sought to examine how the temporal dynamics of a “naturalistic” experience were
97 naturalistic experience were later reflected in participants’ later memories of that experience memories.
98 We also sought to leverage the above conceptual insights into the distinctions between an experience’s
99 essence and its low-level details to build models that explicitly quantified these distinctions. We
100 analyzed an open dataset that comprised behavioral and functional Magnetic Resonance Imaging
101 (fMRI) data collected as participants viewed and then verbally recalled recounted an episode of
102 the BBC television series show Sherlock (Chen et al., 2017). We developed a computational frame-
103 work for characterizing the temporal dynamics of the moment-by-moment content of the episode
104 and of participants’ verbal recalls. Specifically, we use Our framework uses topic modeling (Blei
105 et al., 2003) to characterize the thematic conceptual (semantic) content present in each moment
106 of the episode and recalls ,and we use Hidden Markov Models by projecting each moment into
107 a word embedding space. We then use hidden Markov models (Baldassano et al., 2017; Rabiner,
108 1989) to discretize the this evolving semantic content into events. In this way, we cast naturalistic
109 experiences (and recalls both naturalistic experiences and memories of those experiences)as as
110 geometric topic trajectories through word embedding space that describe how the experiences they
111 evolve over time. In other words, the episode’s topic trajectory is a formalization of its gist. Un-
112 der this framework, successful remembering entails verbally “traversing” the topic traversing
113 the content trajectory of the original episode, thereby reproducing the original shape (essence)
114 of the original experience. Our framework captures the episode’s gist. In addition, comparing
115 the essence in the sequence of geometric coordinates for its events, and its low-level details by
116 examining its within-event geometric properties.

117 Comparing the overall shapes of the topic trajectories of the original episode and of participants’
118 retellings of the episode for the episode and participants’ recalls reveals which aspects of the

episode's essence were preserved (or lost) in the translation into memory. We also identified a network of develop two metrics for assessing participants' memories for low-level details: (1) the precision with which a participant recounts details about each event, and (2) the distinctiveness of their recall for each event, relative to other events. We examine how these metrics relate to overall memory performance as judged by third-party human annotators. We also compare and contrast our general approach to studying memory for naturalistic experiences with standard metrics for assessing performance on more traditional memory tasks, such as list-learning. Last, we leverage our framework to identify networks of brain structures whose responses (as participants watched the episode) reflected the gist temporal dynamics of the episode, and a second network whose responses reflected and/or how participants would later recount the episode it.

Results

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

Chen et al. (2017) collected hand-annotated information about each of 1000 (manually identified) scenes 1,000 (manually delineated) time segments spanning the roughly 45–50 minute video used in their experiment. This information study. Each annotation included: a brief narrative description of what was happening; whether the scene took place indoors vs. outdoors; the location where

145 the action took place, the names of any characters on the screen; names of any characters who
146 were in focus in the camera shot; names of characters who were speaking; the location where the
147 scene took place; the camera angle (close up, medium, long, etc.); whether or not background
148 music was present; and other similar details (for a full list of annotated features, see *Methods*).
149 We took from these annotations the union of all unique words (excluding stop words, such as
150 "and," "or," "but," etc.) across all features and scenes as the "vocabulary" from all annotations
151 as the vocabulary for the topic model. We then concatenated the sets of words across all features
152 contained in overlapping 50-scene sliding windows sliding windows of (up to) 50 annotations,
153 and treated each 50-scene sequence window as a single "document" for the purposes document
154 for the purpose of fitting the topic model. Next, we fit a topic model with (up to) $K = 100$ topics
155 to this collection of documents. We found that 27–32 unique topics (with non-zero weights) were
156 sufficient to describe the time-varying content of the video episode (see *Methods*; Figs. 1, S2). Note
157 We note that our approach is similar in some respects to Dynamic Topic Models (Blei and Lafferty,
158 2006), in that we sought to characterize how the thematic content of the episode evolved over
159 time. However, whereas Dynamic Topic Models are designed to characterize how the properties
160 of collections of documents change over time, our sliding window approach allows us to examine
161 the topic dynamics within a single document (or video). Specifically, our approach yielded (via the
162 topic proportions matrix) a single *topic vector* for each timepoint of the episode (we set timepoints
163 sliding window of annotations transformed by the topic model. We then stretched (interpolated)
164 the resulting windows-by-topics matrix to match the acquisition times of the 1976 time series of
165 the 1,976 fMRI volumes collected as participants viewed the episode).

166 The 32 topics we found were heavily character-focused (e.g., the top-weighted word in
167 each topic was nearly always a character) and could be roughly divided into themes that were
168 primarily Sherlock Holmes focused (Sherlock is centered around Sherlock Holmes (the titular
169 character); primarily John Watson focused (John is, John Watson (Sherlock's close confidant and
170 assistant); or that involved Sherlock and John interacting supporting characters (e.g., Inspector
171 Lestrade, Sergeant Donovan, or Sherlock's brother Mycroft), or the interactions between various
172 groupings of these characters (Fig. S2). This likely follows from the frequency with which these

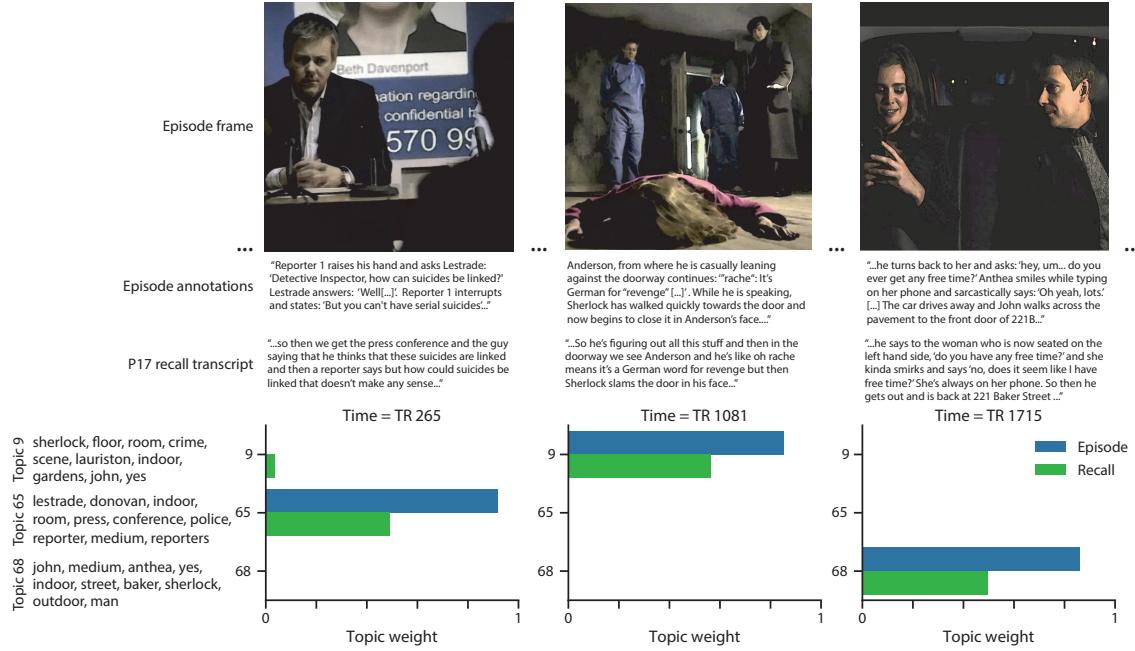


Figure 1: Methods overview. Topic weights in episode and recall content. We used hand-annotated descriptions of detailed, hand-generated annotations describing each moment of video manually identified time segment from the episode to fit a topic model. Three example video frames and their associated descriptions are displayed from the episode (top two rows first row). Participants later recalled are displayed, along with their descriptions from the video corresponding episode annotation (in the third second row, we show) and an example recalls of the same three scenes from participant P17's recall transcript (third row). We used the topic model (fit to the episode annotations) to estimate topic vectors for each moment of video—the episode and each sentence the of participants recalled' recalls. Example topic vectors are displayed in the bottom row (blue: video episode 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, along with the ten-10 highest-weighted words for each topic. Figure S2 provides a full list of the top 10 words from each of the discovered topics.

173 terms appeared in the episode annotations. Several of the identified topics were highly similar,
174 which we hypothesized might allow us to distinguish between subtle narrative differences (if
175 the distinctions between those overlapping topics were meaningful; ~~also see Fig. S3).~~ The
176 topic vectors for each timepoint were also sparse, in that only a small number (usually ~~of topics~~
177 (typically one or two) ~~of topics~~ tended to be “active” in any given timepoint (Fig. 2A). Further,
178 the dynamics of the topic activations appeared to exhibit ~~persistence~~persistence (i.e., given that a
179 topic was active in one timepoint, it was likely to be active in the following timepoint) along with
180 *occasional rapid changes* (i.e., occasionally ~~topics would appear to spring into or out of existence~~topic
181 weights would change abruptly from one timepoint to the next). These two properties of the topic
182 dynamics may be seen in the block diagonal structure of the timepoint-by-timepoint correlation
183 matrix (Fig. 2B). ~~Following Baldassano et al. (2017), we used a Hidden Markov Model and reflect~~
184 the gradual drift and sudden shifts fundamental to the temporal dynamics of many real-world
185 experiences, as well as television episodes. Given this observation, we adapted an approach
186 devised by Baldassano et al. (2017), and used a hidden Markov model (HMM) to identify the event
187 boundaries where the topic activations changed rapidly (i.e., at the boundaries of the blocks in the
188 temporal correlation matrix; event boundaries identified by the HMM are outlined in yellow in
189 Fig. 2B). Part of our model fitting procedure required selecting an appropriate number of “events
190 “~~to segment the timeseries into~~. We events into which the topic trajectory should be segmented.
191 To accomplish this, we used an optimization procedure ~~to identify the number of events that~~
192 ~~maximized within-event stability while also minimizing across-event correlations that maximized~~
193 ~~the difference between the topic weights for timepoints within an event versus timepoints across~~
194 ~~multiple events~~ (see *Methods* for additional details). To create a stable “summary” of the video,
195 ~~we computed the average topic vector within~~). We then created a stable summary of the content
196 ~~within each episode event by averaging the topic vectors across the timepoints spanned by~~ each
197 event (Fig. 2C).

198 Given that the time-varying content of the video episode could be segmented cleanly into
199 discrete events, we wondered whether participants’ recalls of the video episode also displayed
200 a similar structure. We applied the same topic model (already trained on the video episode

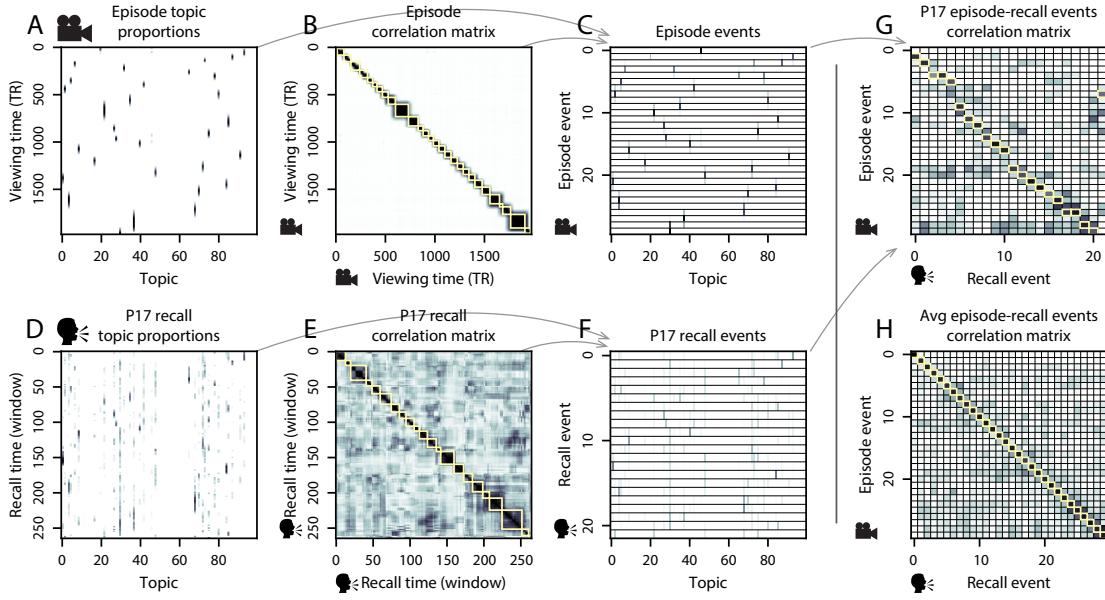


Figure 2: Modelling naturalistic stimuli and recalls. Modeling naturalistic stimuli and recalls. All panels: darker colors indicate greater values; range: [0, 1]. A. Topic vectors ($K = 100$) for each of the 1976 **video episode** timepoints. B. Timepoint-by-timepoint correlation matrix of the topic vectors displayed in Panel A. Event boundaries **detected** **discovered** by the HMM are denoted in yellow (3430 events detected). C. Average topic vectors for each of the 34 **video** 30 **episode** events. D. Topic vectors for each of 294 265 **sliding windows** **of** sentences spoken by an example participant while recalling the **video** **episode**. E. Timepoint-by-timepoint correlation matrix of the topic vectors displayed in Panel D. Event boundaries detected by the HMM are denoted in yellow (2722 events detected). **For similar plots for all participants, see Figure S4.** F. Average topic vectors for each of the 27 **recalled** 22 **recall** events from the example participant. G. Correlations between the topic vectors for every pair of **video** **episode** events (Panel C) and **recalled**-**recall** events (from the example participant; Panel F). **For similar plots for all participants, see Figure S5.** H. Average correlations between each pair of **video** **episode** events and **recalled**-**recall** events (across all 17 participants). To create the figure, each recalled event was assigned to the **video** **episode** 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).**

201 annotations) to each participant's recalls. Analogous to how we analyzed parsed the
202 time-varying content of the video episode, to obtain similar estimates for participants' recalls each
203 participant's recall transcript, we treated each (overlapping) overlapping window of (up to) 10
204 sentence "window" of sentences from their transcript as a "document" and then document,
205 and computed the most probable mix of topics reflected in each timepoint's sentences. This
206 yielded, for each participant, a number-of-sentences number-of-windows by number-of-topics
207 topic proportions matrix that characterized how the topics identified in the original video episode
208 were reflected in the participant's recalls. Note that an important feature of our approach
209 is that it allows us to compare participant's participants' recalls to events from the original
210 video episode, despite that different participants may have used different used widely varying
211 language to describe the same event events, and that those descriptions may not match the original
212 often diverged in content, quality, and quantity from the episode annotations. This is a huge ability
213 to match up conceptually related text that differs in specific vocabulary, detail, and length is an
214 important benefit of projecting the video episode and recalls into a shared "topic" topic space. An
215 example topic proportions matrix from one participant's recalls is shown in Figure 2D.

216 Although the example participant's recall topic proportions matrix has some visual similarity
217 to the video episode topic proportions matrix, the time-varying topic proportions for the example
218 participant's recalls are not as sparse as for the video (e.g., those for the episode) (compare Figs. 2A
219 and D). Similarly, although there do appear to be periods of stability in the recall topic dynamics
220 (e.g., i.e., most topics are active or inactive over contiguous blocks of time), the overall timecourses
221 are not as cleanly delineated as the video topics are changes in topic activations that define event
222 boundaries appear less clearly delineated in participants' recalls than in the episode's annotations.
223 To examine these patterns in detail, we computed the timepoint-by-timepoint correlation matrix
224 for the example participant's recall topic proportions matrix (Fig. 2E). As in the video episode
225 correlation matrix (Fig. 2B), the example participant's recall correlation matrix has a strong block
226 diagonal structure, indicating that their recalls are discretized into separated events. As for the
227 video correlation matrix, we can use an HMM, along with the aforementioned number-of-events
228 optimization procedure (also We used the same HMM-based optimization procedure that we had

229 applied to the episode's topic proportions matrix (see *Methods*) to determine how many events
230 are reflected estimate an analogous set of event boundaries in the participant's recalls and where
231 specifically the event boundaries fall recounting of the episode (outlined in yellow). We carried
232 out a similar this analysis on all 17 participants' recall topic proportions matrices (Fig. S4).

233 Two clear patterns emerged from this set of analyses. First, although every individual partic-
234 ipant's recalls could be segmented into discrete events (i.e., every individual participant's recall
235 correlation matrix exhibited clear block diagonal structure; Fig. S4), each participant appeared
236 to have a unique *recall resolution*, reflected in the sizes of those blocks. For example, While
237 some participants' recall topic proportions segmented into just a few events (e.g., Participants P1,
238 P4, and P15), while others' recalls P5, and P7), others' segmented into many shorter-duration
239 shorter-duration events (e.g., Participants P12, P13, and P17). This suggests that different partic-
240 ipants may be recalling the video episode with different levels of detail—e.g., detail—i.e., some might
241 touch on just the major plot points recount only high-level essential plot details, whereas others
242 might attempt to recall every minor scene recount low-level details instead (or in addition). The
243 second clear pattern present in every individual participant's recall correlation matrix is was that,
244 unlike in the video episode correlation matrix, there are were substantial off-diagonal correlations in
245 participant's recalls. Whereas each event in the original video (was episode was) (largely) sepa-
246 rable from the others (Fig. 2B), in transforming those separable events into memory participants
247 appear, participants appeared to be integrating across different across multiple events, blend-
248 ing elements of previously recalled and not-yet-recalled events content into each newly recalled
249 event (Figs. 2D, S4; also see Howard et al., 2012; Manning et al., 2011)(Figs. 2E, S4; also see Howard et al., 2012; Mannin
250 .

251 The above results indicate that both the structure of the original video demonstrate that topic
252 models capture the dynamic conceptual content of the episode and participants' recalls of the
253 video episode. Further, the episode and recalls exhibit event boundaries that can be identified
254 automatically by characterizing using HMMs to segment the dynamic content using a shared
255 topic model and segmenting the content into events using HMMs. Next, Next, we asked
256 whether some correspondence might be made between the specific content of the events the

257 participants experienced ~~in the video~~while viewing the episode, and the events they later re-
258 called. ~~One approach to linking the experienced (video) and recalled events is to label each~~
259 ~~We labeled each recall~~ event as matching the ~~video episode~~ event with the most similar
260 (i.e., most highly correlated) topic vector (Figs. 2G, S5). This ~~yields yielded~~ a sequence of “pre-
261 sented” events from the original ~~video episode~~, and a ~~sequence of~~ (potentially differently ordered)
262 ~~sequence of~~ “recalled” events for each participant. Analogous to classic list-learning studies,
263 we can then examine participants’ recall sequences by asking which events they tended to recall
264 first (~~probability of first recall; Fig. A; Atkinson and Shiffrin, 1968; Postman and Phillips, 1965; Welch and Burnett, 1924~~)
265 (~~probability of first recall; Fig. 3A; Atkinson and Shiffrin, 1968; Postman and Phillips, 1965; Welch and Burnett, 1924~~)
266 ; how participants most often ~~transition transitioned~~ between recalls of the events as a function of
267 the temporal distance between them (~~lag-conditional response probability; Fig. B; Kahana, 1996~~)
268 (~~lag-conditional response probability; Fig. 3B; Kahana, 1996~~); and which events they were likely to
269 remember overall (~~serial position recall analyses; Fig. C; Murdock, 1962~~). In (~~serial position recall analyses; Fig. 3C; Mu~~
270 . ~~Some of the patterns we observed appeared to be similar to classic effects from the list-learning~~
271 ~~studies, this set of three analyses may be used to gain a nearly complete view into the sequences~~
272 ~~of recalls participants made (e.g., Kahana, 2012)~~. Extending these analyses to apply to naturalistic
273 stimuli and recall (Heusser et al., 2017) highlights that, in naturalistic recall, these analyses provide
274 a wholly incomplete picture: they leave out any attempt to quantify participants’ abilities to
275 capture the ~~content of what occurred in the video their only experimental instruction!~~ literature.
276 For example, participants had a higher probability of initiating recall with early events (Fig. 3A)
277 and a higher probability of transitioning to neighboring events with an asymmetric forward bias
278 (Fig. 3B). However, unlike what is typically observed in list-learning studies, we did not observe
279 patterns comparable to the primacy or recency serial position effects (Fig. 3C). We hypothesized
280 that participants might be leveraging meaningful narrative associations and references over long
281 timescales throughout the episode.

282 The dynamic content of the video and participants’ recalls is quantified in the corresponding
283 topic proportion matrices. However, it is difficult to gain deep insights into that content solely by
284 examining the topic proportion matrices (Clustering scores are often used by memory researchers to

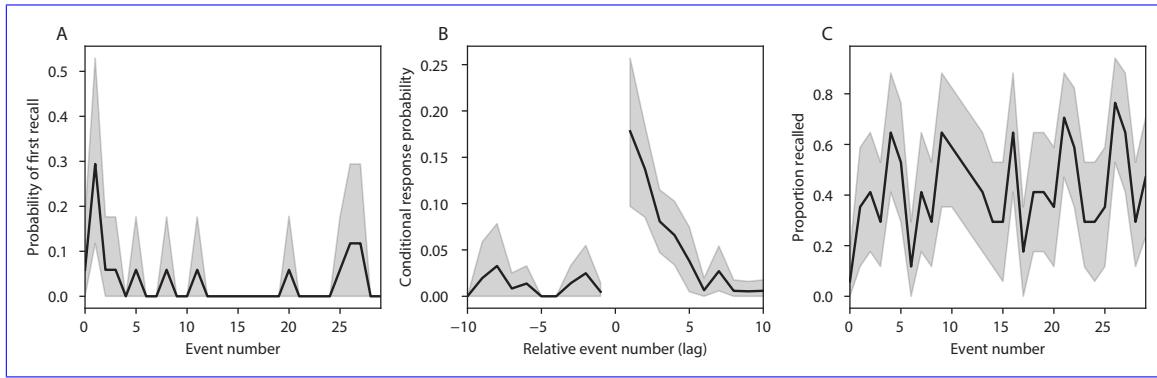


Figure 3: Naturalistic extensions of classic list-learning memory analyses. **A.** The probability of first recall as a function of the serial position of the event in the episode. **B.** The probability of recalling each event, conditioned on having most recently recalled the event *lag* events away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

285 characterize how people organize their memories of words on a studied list (for review, see Polyn et al., 2009)
 286 . We defined analogous measures to characterize how participants organized their memories
 287 for episodic events (see *Methods* for details). Temporal clustering refers to the extent to which
 288 participants group their recall responses according to encoding position. Overall, we found that
 289 sequentially viewed episode events tended to appear nearby in participants' recall event sequences
 290 (mean clustering score: 0.732, SEM: 0.033). Participants with higher temporal clustering scores
 291 tended to exhibit better overall memory for the episode, according to both Chen et al. (2017)'s
 292 hand-counted numbers of recalled scenes from the episode (Pearson's $r(15) = 0.49$, $p = 0.046$) and
 293 the numbers of episode events that best-matched at least one recall event (i.e., model-estimated
 294 number of events recalled; Pearson's $r(15) = 0.59$, $p = 0.013$). Semantic clustering measures the
 295 extent to which participants cluster their recall responses according to semantic similarity. We
 296 found that participants tended to recall semantically similar episode events together (mean clustering
 297 score: 0.650, SEM: 0.032), and that semantic clustering score was also related to both hand-counted
 298 (Pearson's $r(15) = 0.65$, $p = 0.005$) and model-estimated (Pearson's $r(15) = 0.58$, $p = 0.015$) numbers
 299 of recalled events.

300 The above analyses illustrate how our framework for characterizing the dynamic conceptual
 301 content of naturalistic episodes enables us to carry out analyses that have traditionally been

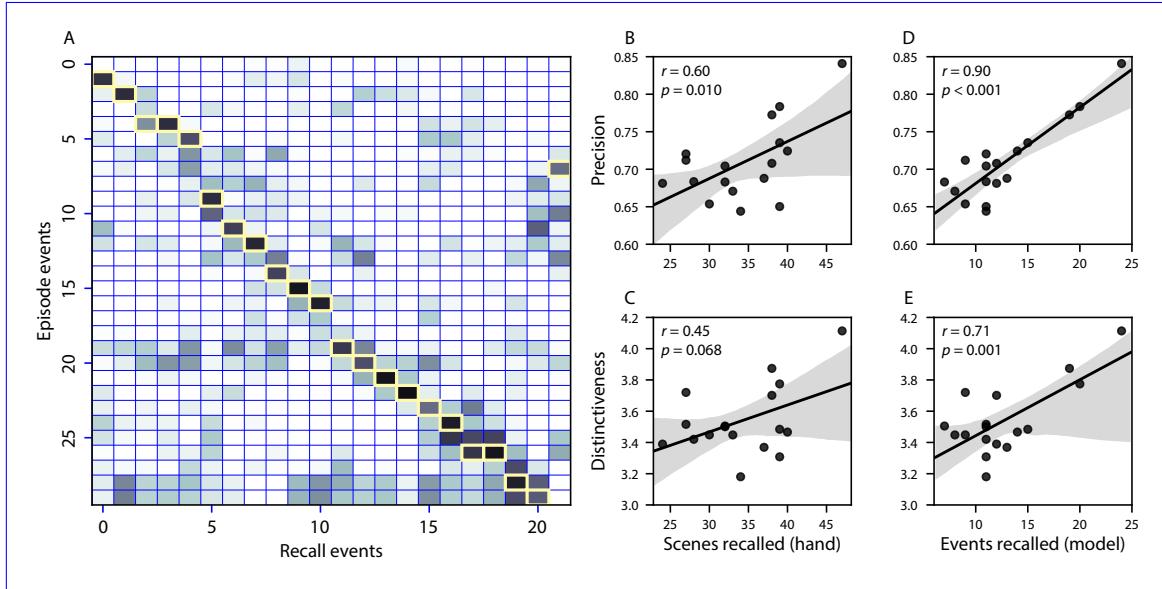


Figure 4: Novel content-based metrics of naturalistic memory: precision and distinctiveness. A. The episode-recall correlation matrix for a representative participant (P17). The yellow boxes highlight the maximum correlation in each column. The example participant's overall precision score was computed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

302 applied to much simpler list-learning paradigms. However, perhaps the most interesting aspects
303 of memory for naturalistic episodes are those that have no list-learning analogs. The nuances of
304 how one's memory for an event might capture some details, yet distort or neglect others, is central
305 to how we use our memory systems in daily life. Yet when researchers study memory in highly
306 simplified paradigms, those nuances are not typically observable. We next developed two novel,
307 continuous metrics, termed *precision* and *distinctiveness*, aimed at characterizing distortions in the
308 conceptual content of individual recall events, and the conceptual overlap between how people
309 described different events.

310 *Precision* is intended to capture the "completeness" of recall, or how fully the presented content
311 was recapitulated in a participant's recounting. We define a recall event's precision as the maximum
312 correlation between the topic proportions of that recall event and any episode event (Fig. 4). In
313 other words, given that a recall event best matches a particular episode event, more precisely
314 recalled events overlap more strongly with the conceptual content of the original episode event.
315 When a given event is assigned a blend of several topics, as is often the case (Fig. 2), a high precision
316 score requires recapitulating the relative topic proportions during recall.

317 *Distinctiveness* is intended to capture the "specificity" of recall. In other words, distinctiveness
318 quantifies the extent to which a given recall event reflects the most similar episode event over and
319 above other episode events. Intuitively, distinctiveness is like a normalized variant of our precision
320 metric. Whereas precision solely measures how much detail about an episode was captured in
321 someone's recall, distinctiveness penalizes details that also pertain to other episode events. We
322 define the distinctiveness of an event's recall as its precision expressed in standard deviation
323 units with respect to other episode events. Specifically, for a given recall event, we compute the
324 correlation between its topic vector and that of each episode event. This yields a distribution of
325 correlation coefficients (one per episode event). We subtract the mean and divide by the standard
326 deviation of this distribution to z-score the coefficients. The maximum value in this distribution
327 (which, by definition, belongs to the episode event that best matches the given recall event) is that
328 recall event's distinctiveness score. In this way, recall events that match one episode event far better
329 than all other episode events will receive a high distinctiveness score. By contrast, a recall event

330 that matches all episode events roughly equally will receive a comparatively low distinctiveness
331 score.

332 In addition to examining how precisely and distinctively participants recalled individual
333 events, one may also use these metrics to summarize each participant's performance by averaging
334 across a participant's event-wise precision or distinctiveness scores. This enables us to quantify
335 how precisely a participant tended to recall subtle within-event details, as well as how specific
336 (distinctive) those details were to individual events from the episode. Participants' average
337 precision and distinctiveness scores were strongly correlated ($r(15) = 0.90, p < 0.001$). This indicates
338 that participants who tended to precisely recount low-level details of episode events also tended
339 to do so in an event-specific way (e.g., Figs. 2A, D) or the corresponding correlation matrices
340 (Figs. 2B, E) as opposed to detailing recurring themes that were present in most or all episode
341 events; this behavior would have resulted in high precision but low distinctiveness). We found that,
342 across participants, higher precision scores were positively correlated with the numbers of both
343 hand-annotated scenes ($r(15) = 0.60, p = 0.010$) and model-estimated events ($r(15) = 0.90, p < 0.001$)
344 that participants recalled. Participants' average distinctiveness scores were also correlated with
345 both the hand-annotated ($r(15) = 0.45, p = 0.068$) and model-estimated ($r(15) = 0.71, p = 0.001$) numbers
346 of recalled events.

347 Examining individual recalls of the same episode event can provide insights into how the above
348 precision and distinctiveness scores may be used to characterize similarities and differences in how
349 different people describe the same shared experience. In Figure 5, we compare recalls for the same
350 episode event from the participants with the highest (P17) and lowest (P6) precision scores. From
351 the HMM-identified episode event boundaries, we recovered the set of annotations describing the
352 content of a single episode event (event 21; Fig. 5C), and divided them into different color-coded
353 sections for each action or feature described. Next, we used an analogous approach to identify
354 the set of sentences comprising the corresponding recall event from each of the two example
355 participants (Fig. 5D). We then colored all words describing actions and features in the transcripts
356 shown in Panel D according to the color-coded annotations in Panel C. Visual comparison of these
357 example recalls reveals that the more precise recall captures more of the episode event's content,

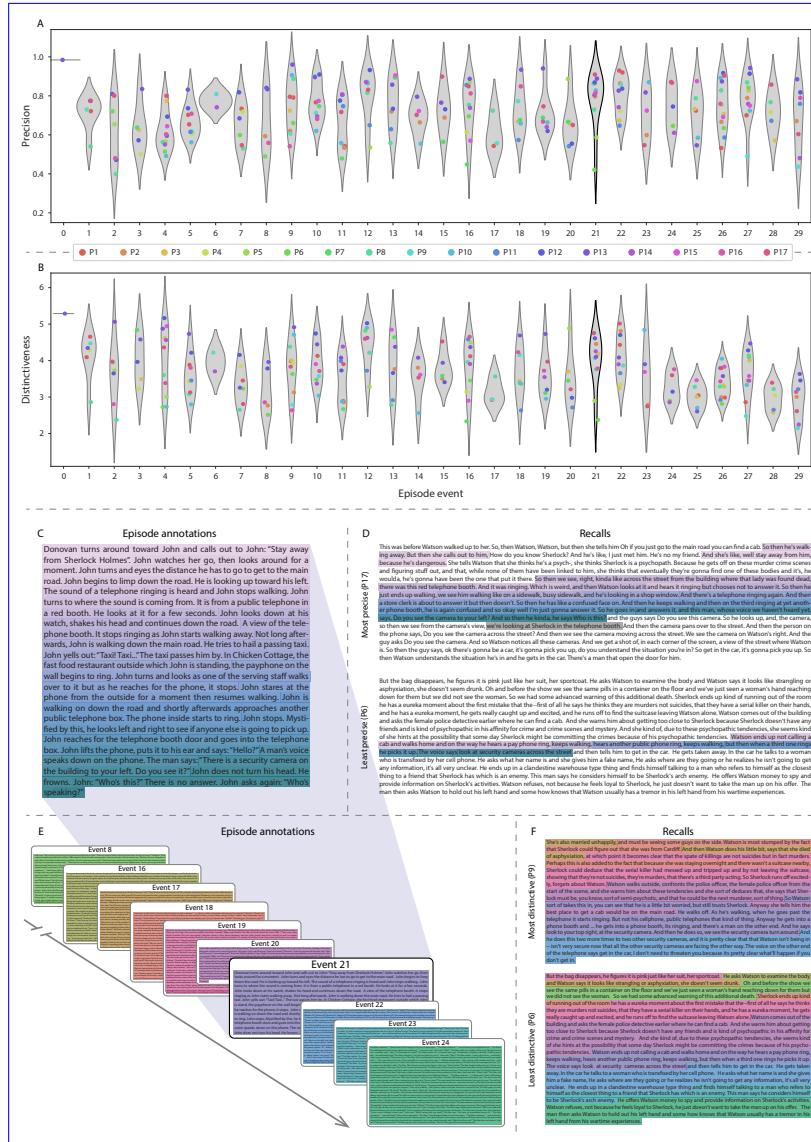


Figure 5: Precision reflects the completeness of recall, whereas distinctiveness reflects recall specificity. A. Recall precision by episode event. Grey violin plots display kernel density estimates for the distribution of recall precision scores for a single episode event. Colored dots within each violin plot represent individual participants' recall precision for the given event. B. Recall distinctiveness by episode event, analogous to Panel A. C. The set of "Narrative Details" episode annotations (generated by Chen et al., 2017) comprising an example episode event (22) identified by the HMM. Each action or feature is highlighted in a different color. D. Sentences comprising the most precise (P17) and least precise (P6) participants' recalls of episode event 21. Descriptions of specific actions or features reflecting those highlighted in Panel B are highlighted in the corresponding color. The text highlighted in gray denotes a (rare) false recall. The unhighlighted text denotes correctly recalled information about other episode events. E. The sets of "Narrative Details" episode annotations (generated by Chen et al., 2017) for scenes comprising episode events described by the example participants in Panel F. Each event's text is highlighted in a different color. F. The sentences comprising the most distinctive (P9) and least distinctive (P6) participants' recalls of episode event 21. Sections of recall describing each each episode event in Panel E are highlighted with the corresponding color.

358 and in greater detail.

359 Figure 5 also illustrates the differences between high and low distinctiveness scores. We
360 extracted the set of sentences comprising the most distinctive recall event (P9) and least distinctive
361 recall event (P6) corresponding to the example episode event shown in Panel C (event 21). We
362 also extracted the annotations for all episode events whose content these participants' single recall
363 events described. We assigned each episode event a unique color (Fig. 5E), and colored each
364 recalled sentence (Panel F) according to the episode events they best matched. Visual inspection
365 of Panel F reveals that the most distinctive recall's content is tightly concentrated around event
366 21, whereas the least distinctive recall incorporates content from a much wider range of episode
367 events.

368 The preceding analyses sought to characterize how participants' recounts of individual
369 episode events captured the low-level details of each event. Next, we sought to characterize how
370 participants' recounts of the full episode captured its high-level essence—i.e., the shape of the
371 episode's trajectory through word embedding (topic) space. To visualize the essence of the episode
372 and each participant's recall trajectory (Heusser et al., 2018b), S4. To visualize the time-varying
373 high-dimensional content in a more intuitive way (Heusser et al., 2018b) we projected the topic
374 proportions matrices onto a for the episode and recalls onto a shared two-dimensional space using
375 Uniform Manifold Approximation and Projection (UMAP; ?)(UMAP; McInnes et al., 2018). In
376 this lower-dimensional space, each point represents a single video-episode or recall event, and
377 the distances between the points reflect the distances between the events' associated topic vectors
378 (Fig. 6). In other words, events that are nearer to each other in this space are more semantically
379 similar, and those that are farther apart are less so.

380 Visual inspection of the video-episode and recall topic trajectories reveals a striking pattern.
381 First, the topic trajectory of the video-episode (which reflects its dynamic content; Fig. 6A) is cap-
382 tured nearly perfectly by the averaged topic trajectories of participants' recalls (Fig. 6B). To assess
383 the consistency of these recall trajectories across participants, we asked: given that a participant's
384 recall trajectory had entered a particular location in the reduced topic space, could the position of
385 their next recalled event be predicted reliably? For each location in the reduced topic space, we

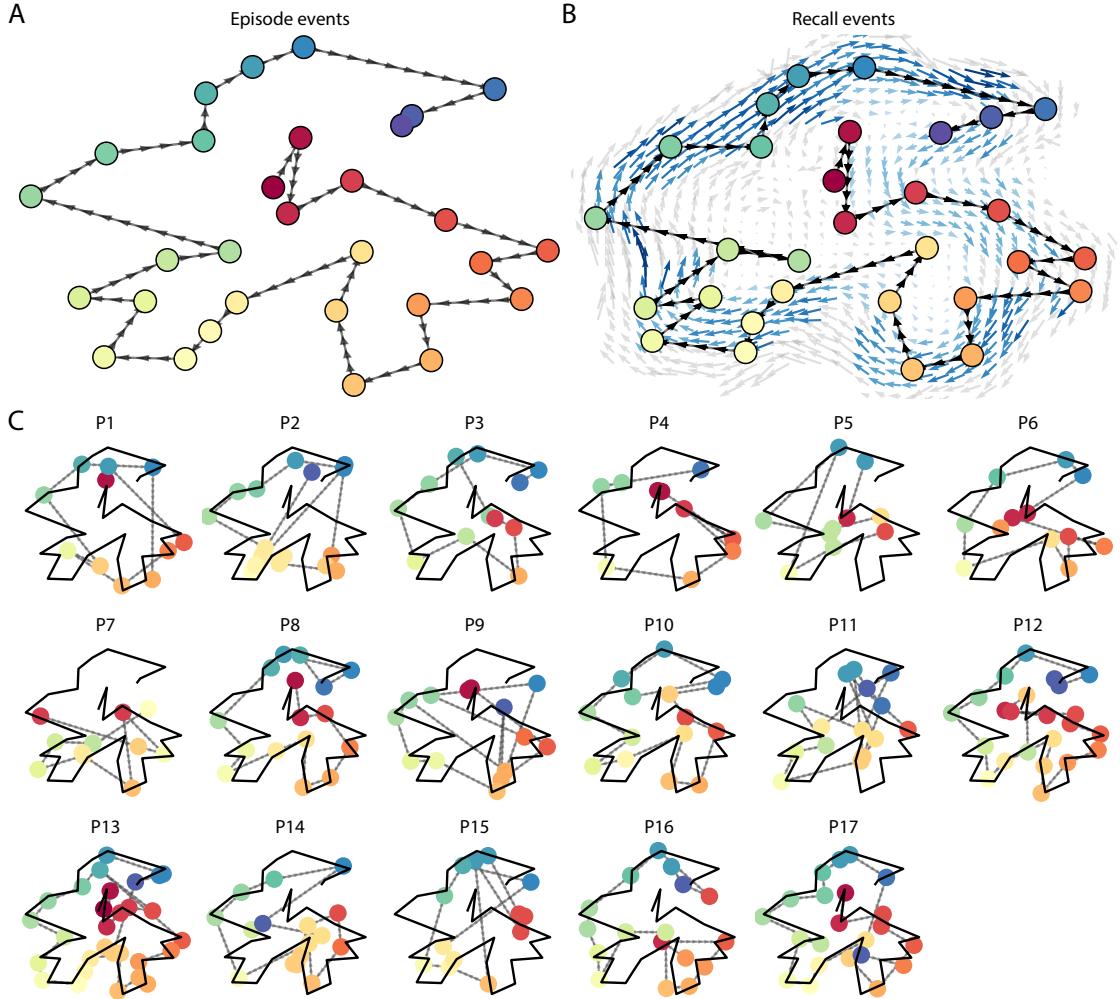


Figure 6: Trajectories through topic space capture the dynamic content of the video and recalls. All panels: the topic proportion matrices have been projected onto a shared two-dimensional space using UMAP. **A.** The two-dimensional topic trajectory taken by the episode of *Sherlock*. Each dot indicates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

386 computed the set of line segments connecting successively recalled events (across all participants)
387 that intersected that location (see *Methods*~~for additional details~~). We then computed (for each
388 location) the distribution of angles formed by the lines defined by those line segments and a fixed
389 reference line (the x -axis). Rayleigh tests revealed the set of locations in topic space at which these
390 across-participant distributions exhibited reliable peaks (blue arrows in Fig. 6B reflect significant
391 peaks at $p < 0.05$, corrected). We observed that the locations traversed by nearly the entire ~~video~~
392 ~~episode~~ trajectory exhibited such peaks. In other words, participants' ~~recalls~~ exhibited similar
393 trajectories ~~to each other~~ that also matched the trajectory of the original ~~video episode~~ (Fig. 6C).
394 This is especially notable when considering the fact that the number of ~~events participants recalled~~
395 ~~HMM-identified recall events~~ (dots in Fig. 6C) varied considerably across people, and that ev-
396 ery participant used different words to describe what they had remembered happening in the
397 ~~video episode~~. Differences in the numbers of ~~remembered recall~~ events appear in participants' tra-
398 jectories as differences in the sampling resolution along the trajectory. We note that this framework
399 also provides a means of ~~detangling disentangling~~ classic "proportion recalled" measures (i.e.,
400 the proportion of ~~video events referenced episode events described~~ in participants' recalls) from
401 participants' abilities to recapitulate the ~~full gist of the original video episode's essence~~ (i.e., the
402 similarity ~~in the shape between the shapes~~ of the original ~~video episode~~ trajectory and that defined
403 by each participant's recounting of the ~~video episode~~).

404 ~~In addition to enabling us to visualize the episode's high-level essence, describing the episode~~
405 ~~as a geometric trajectory also enables us to drill down to individual words and quantify how each~~
406 ~~word relates to the memorability of each event. This provides another approach to examining~~
407 ~~participants' recall for low-level details beyond the precision and distinctiveness measures we~~
408 ~~defined above. The results displayed in Figures 3C and 5A suggest that certain events were~~
409 ~~remembered better than others. Given this, we next asked asked whether the events that were~~
410 ~~generally remembered precisely or imprecisely tended to reflect particular content. Because our~~
411 ~~analysis framework projects the dynamic ~~video episode~~ content and participants' recalls onto a~~
412 ~~shared topic into a shared space, and because the dimensions of that space are known (i.e., each~~
413 ~~topic dimension is a set represent topics (which are, in turn, sets) of weights over known words in the~~

vocabulary; Fig. S2), we can examine the topic trajectories to understand which specific content was remembered well (or poorly). For each video event, we can ask: what was the average correlation (across participants) between the video event's topic vector and the closest matching recall event topic vectors from each participant? This yields a single correlation coefficient for each video event, describing how closely participants' recalls of the event tended to reliably capture its content are able to recover the weighted combination of words that make up any point (i.e., topic vector) in this space. We first computed the average precision with which participants recalled each of the 30 episode events (Fig. 7A). (We also examined how different comparisons between each video event's topic vector and the corresponding recall event topic vectors related to hand-annotated characterizations of memory performance; see *Supporting Information*). Given this summary of which events were recalled reliably (or not), we next asked whether the better-remembered or worse-remembered events tended to reflect particular topics. We ; note that this result is analogous to a serial position curve created from our precision metric). We then computed a weighted average of the topic vectors for each video episode event, where the weights reflected how reliably precisely each event was recalled. To visualize the result, we created a "wordle" image (Mueller et al., 2018) where words weighted more heavily by better-remembered more precisely-remembered topics appear in a larger font (Fig. 7B, green box). Events that reflected topics weighting heavily on characters like "Sherlock" and "John" (i.e., the main characters) and locations like "221b Baker Street" (i.e., a major recurring location Across the full episode, content that weighted heavily on topics and words central to the major foci of the episode (e.g., the names of the two main characters, "Sherlock" and "John," and the address of the flat that Sherlock and John share) were a major recurring location, "221B Baker Street") was best remembered. An analogous analysis revealed which themes were poorly less-precisely remembered. Here in computing the weighted average over events' topic vectors, we weighted each event in inverse proportion to how well it was remembered its average precision (Fig. 7B, red box). This revealed that events with The least precisely remembered episode content reflected information that was extraneous to the episode's essence, such as the proper names of relatively minor characters such as (e.g., "Mike," "JeffreyMolly," and "Molly" as well as less integral plot Lestrade") and locations (e.g., "hospital"

442 and “office”) were least well-remembered. This suggests that what is retained in memory are the
443 major plot elements (i.e., the overall “gist” of what happened), whereas the more minor details are
444 prone to pruning St. Bartholomew’s Hospital”).

445 In addition to constructing overall summaries, assessing the video and recall topic vectors from
446 individual recalls can provide further insights. Specifically, for any given event we can construct

447 A similar result emerged from assessing the topic vectors for individual episode and recall
448 events (Fig. 7C). Here, for each of the three most and least precisely remembered episode events,
449 we have constructed two wordles: one from the original ~~video~~-episode event’s topic vector
450 (~~left~~) and a second from the average ~~topic~~ vectors produced by all participants’ recalls of that
451 event. We can then examine those wordles visually to gain an intuition for which aspects of the
452 video event were recapitulated in participants’ recalls of that event. Several example wordles
453 are displayed in Figure 7C (wordles from the three best-remembered events are ~~recall~~ topic vector
454 for that event (right). The three most precisely remembered events (circled in green; ~~wordles~~
455 from the three worst-remembered events are circled in red). Using wordles to visually compare the
456 topical content of each video event and the (average) corresponding recall event reveals the specific
457 content from the specific events that is reliably retained in the transformation into memory (green
458 events) or not (red events) correspond to scenes integral to the central plot-line: a mysterious
459 figure spying on John in a phone booth; John meeting Sherlock at Baker St. to discuss the murders;
460 and Sherlock laying a trap to catch the killer. Meanwhile, the least precisely remembered events
461 (circled in red) reflect scenes that comprise minor plot points: a video of singing cartoon characters
462 that participants viewed in an introductory clip prior to the main episode; John asking Molly
463 about Sherlock’s habit of over-analyzing people; and Sherlock noticing evidence of Anderson’s
464 and Donovan’s affair.

465 **Transforming experience into memory.** A. Average correlations (across participants) between
466 the topic vectors from each video event and the closest matching recall events. Error bars
467 denote bootstrap-derived across-participant 95% confidence intervals. The stars denote the three
468 best-remembered events (green) and worst-remembered events (red). B. Wordles comprising the
469 top 200 highest-weighted words reflected in the weighted-average topic vector across video events.

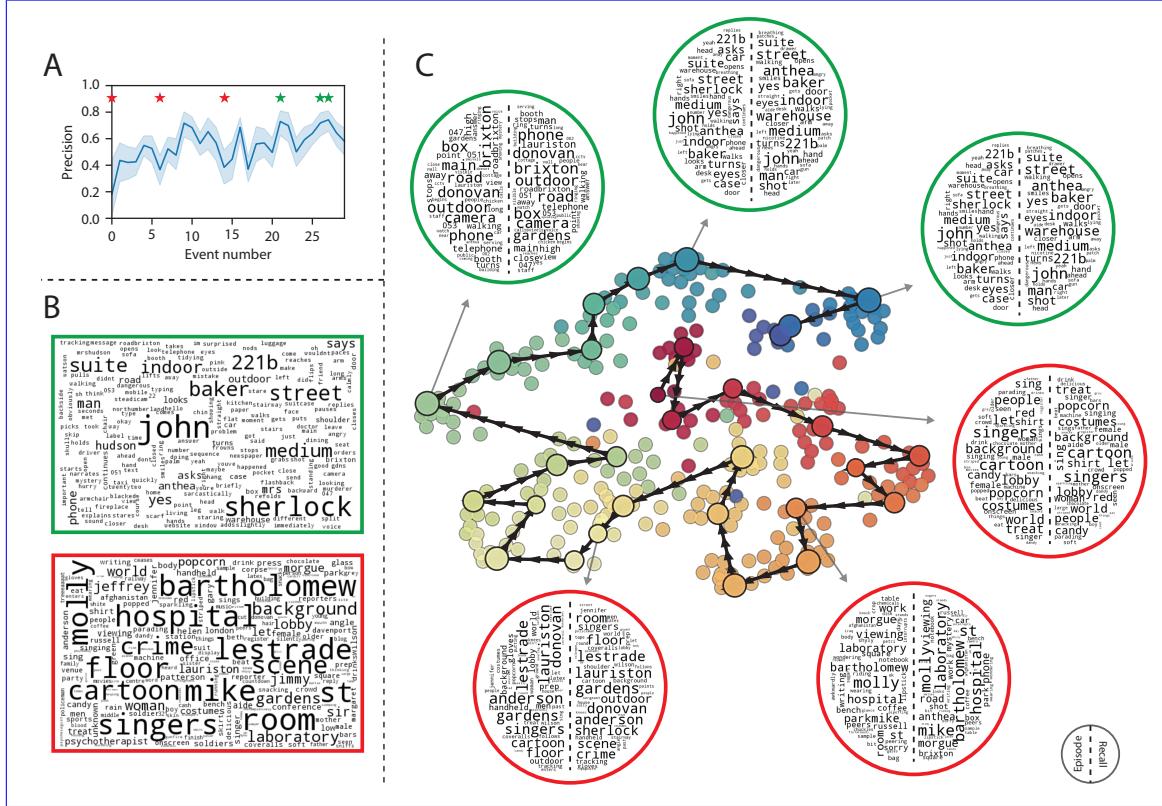


Figure 7: Language used in the most and least precisely remembered events. A. Average precision (episode event-recall event topic vector correlation) across participants for each episode event. Here we defined each episode event's precision for each participant as the correlation between its topic vector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

470 Green: video events were weighted by how well the topic vectors derived from recalls of those
471 events matched the video events' topic vectors (Panel A). Red: video events were weighted by the
472 inverse of how well their topic vectors matched the recalled topic vectors. C. The set of all video and
473 recall events is projected onto the two-dimensional space derived in Figure 6. The dots outlined
474 in black denote video events (dot size reflects the average correlation between the video event's
475 topic vector and the topic vectors from the closest matching recalled events from each participant;
476 bigger dots denote stronger correlations). The dots without black outlines denote recalled events.
477 All dots are colored using the same scheme as Figure 6A. Wordles for several example events are
478 displayed (green: three best-remembered events; red: three worst-remembered events). Within
479 each circular wordle, the left side displays words associated with the topic vector for the video
480 event, and the right side displays words associated with the (average) recall event topic vector,
481 across all recall events matched to the given video event.

482 The results thus far inform us about which aspects of the dynamic content in the episode participants watched were preserved or altered in participants' memories of the episode. We next carried out a series of analyses aimed at understanding which brain structures might implement these processes. In one analysis facilitate these preservations and transformations between the participants' shared experience of watching the episode and their subsequent memories of the episode. In the first analysis, we sought to identify which brain structures brain structures that were sensitive to the video's dynamic dynamic unfolding of the episode's content, as characterized by its topic trajectory. Specifically, we We used a searchlight procedure to identify the extent to which each cluster of voxels exhibited a timecourse (as the clusters of voxels whose activity patterns displayed a proximal temporal correlation structure (as participants watched the video) whose temporal correlation matrix matched the temporal correlation matrix of the original videoepisode's topic proportion matrix proportions (Fig. 2B). As shown in Figure 8A, the analysis revealed a network of regions including bilateral frontal cortex and cingulate cortex, suggesting that these regions may play a role in maintaining information relevant to the narrative structure of the video (see Methods for additional details). In a second analysis, we sought to identify which brain structures' responses (while viewing the video) brain

498 structures whose responses (during episode viewing) reflected how each participant would later
499 structure their *recall* the video *recounting* of the episode. We used *an analogous* a searchlight procedure
500 to identify clusters of voxels whose proximal temporal correlation matrices reflected the temporal
501 correlation matrix matched that of the topic proportions for each individual's recalls matrix for
502 each participant's recall transcript (Figs. 2D, S4) 8B; see *Methods* for additional details). To ensure
503 our searchlight procedure identified regions specifically sensitive to the temporal structure of the
504 episode or recalls (i.e., rather than those with a temporal autocorrelation length similar to that of the
505 episode and recalls), we performed a phase shift-based permutation correction (see *Methods*). As
506 shown in Figure 8B, the C, the episode-driven searchlight analysis revealed a distributed network of
507 regions including the ventromedial prefrontal cortex (vmPFC), anterior cingulate cortex, and right
508 medial temporal lobe (rMTL), suggesting that these regions that may play a role in transforming
509 each individual processing information relevant to the narrative structure of the episode. The
510 recall-driven searchlight analysis revealed a second network of regions (Fig. 8D) that may facilitate
511 a person-specific transformation of one's experience into memory. In identifying regions whose
512 responses to ongoing experiences reflect how those experiences will be remembered later, this
513 latter analysis extends classic *subsequent memory effect* analyses (e.g., Paller and Wagner, 2002) to the
514 domain of naturalistic stimuli experiences.

515 The searchlight analyses described above yielded two distributed networks of brain regions
516 whose activity timecourses tracked with the temporal structure of the episode (Fig. 8C) or participants'
517 subsequent recalls (Fig. 8D). We next sought to gain greater insight into the structures and
518 functional networks our results reflected. To accomplish this, we performed an additional,
519 exploratory analysis using Neurosynth (Yarkoni et al., 2011). Given an arbitrary statistical map as
520 input, Neurosynth performs a massive automated meta-analysis, returning a ranked list of terms
521 frequently used in neuroimaging papers that report similar statistical maps. We ran Neurosynth
522 on the (unthresholded) permutation-corrected maps for the episode- and recall-driven searchlight
523 analyses. The top ten terms with maximally similar meta-analysis images identified by Neurosynth
524 are shown in Figure 8.

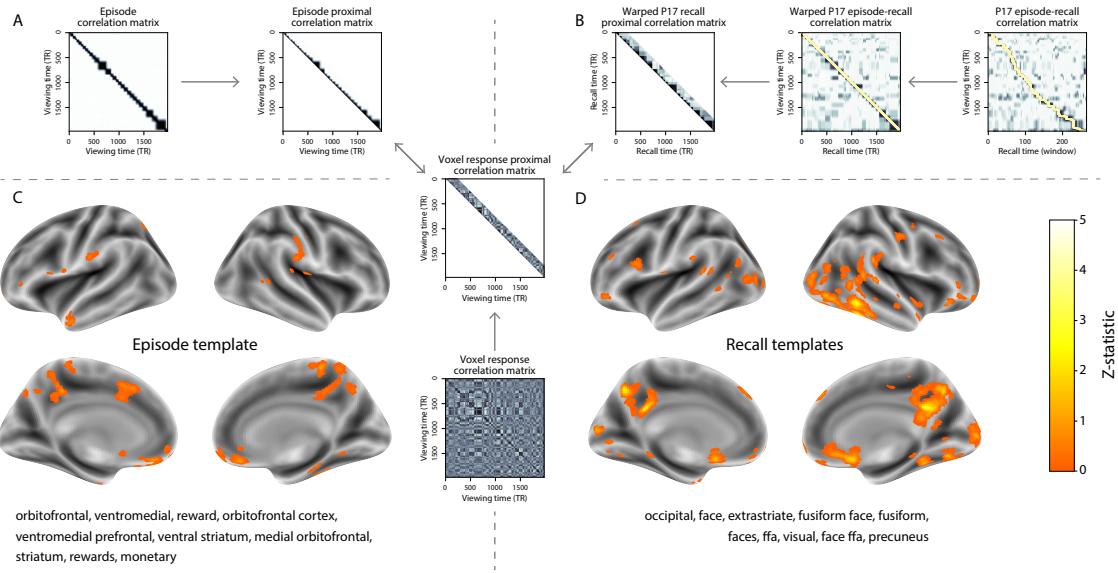


Figure 8: Brain structures that underlie the transformation of experience into memory. **A.** We searched for regions whose responses (as participants watched isolated the video) matched proximal diagonals from the temporal correlation matrix upper triangle of the video topic proportions. These regions are sensitive episode correlation matrix, and applied this same diagonal mask to the narrative structure voxel response correlation matrix for each cube of voxels in the videobrain. **B.** We then searched for brain regions whose responses (as activation timeseries consistently exhibited a similar proximal correlational structure to the episode model, across participants watched). **B.** We used dynamic time warping (Berndt and Clifford, 1994) to align each participant's recall timeseries to the video) matched the temporal correlation matrix TR timeseries of the topic proportions derived from episode. We then applied the same diagonal mask used in Panel A to isolate the proximal temporal correlations and searched for brain regions whose activation timeseries for an individual consistently exhibited a similar proximal correlational structure to each individual's later recall of video. **C.** We identified a network of regions are sensitive to how the narrative structure of the video participants' ongoing experience. The map shown is transformed into thresholded at $p < 0.05$, corrected. The top ten Neurosynth terms displayed in the panel were computed using the unthresholded map. **D.** We also identified a memory network of regions sensitive to how individuals would later structure the video episode's content in their recalls. Both panels: the maps are The map shown is thresholded at $p < 0.05$, corrected. The top ten Neurosynth terms displayed in the panel were computed using the unthresholded map.

525 **Discussion**

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

540 Our work casts remembering as reproducing (behaviorally and neurally) the topic trajectory, or
541 "gistshape," of the original experience. This view—an experience. When we characterized memory
542 for a television episode using this framework, we found that every participant's recounting of
543 the episode recapitulated the low spatial frequency details of the shape of its trajectory through
544 topic space (Fig. 6). We termed this narrative scaffolding the episode's *essence*. Where participants'
545 behaviors varied most was in their tendencies to recount specific low-level details from each
546 episode event. Geometrically, this appears as high spatial frequency distortions in participants'
547 recall trajectories relative to the trajectory of the original episode (Fig. 7). We developed metrics
548 to characterize the precision (recovery of any and all event-level information) and distinctiveness
549 (recovery of event-specific information). We also used word cloud visualizations to interpret the
550 details of these event-level distortions.

551 The neural analyses we carried out (Fig. 8) also leveraged our geometric framework for

552 characterizing the shapes of the episode and participants' recounts. We identified one network
553 of regions whose responses tracked with temporal correlations in the conceptual content of the
554 episode (as quantified by topic models applied to a set of annotations about the episode). This
555 network included orbitofrontal cortex, ventromedial prefrontal cortex, and striatum, among others.
556 As reviewed by Ranganath and Ritchey (2012), several of these regions are members of the *anterior*
557 *temporal system*, which has been implicated in assessing and processing the familiarity of ongoing
558 experiences, emotions, social cognition, and reward. A second network we identified tracked
559 with temporal correlations in the idiosyncratic conceptual content of participants' subsequent
560 recounts of the episode. This network included occipital cortex, extrastriate cortex, fusiform
561 gyrus, and the precuneus. Several of these regions are members of the *posterior medial system* (Ranganath and Ritchey, 201
562 , which has been implicated in matching incoming cues about the current situation to internally
563 maintained *situation models* that specify the parameters and expectations inherent to the current
564 situation (also see Zacks et al., 2007; Zwaan and Radvansky, 1998). Taken together, our results
565 support the notion that these two (partially overlapping) networks work in coordination to make
566 sense of our ongoing experiences, distort them in a way that links them with our prior knowledge
567 and experiences, and encodes those distorted representations into memory for our later use.

568 Our general approach draws inspiration from prior work aimed at elucidating the neural and
569 behavioral underpinnings of how we process dynamic naturalistic experiences and remember
570 them later. One-Our approach to identifying neural responses to naturalistic stimuli (includ-
571 ing experiences) entails building a-an explicit model of the stimulus dynamics and searching
572 for brain regions whose responses are consistent with the model (also see Huth et al., 2016, 2012)
573 . In prior work, a series of studies from Uri Hasson's group (Baldassano et al., 2017; Chen et al.,
574 2017; Lerner et al., 2011; Simony et al., 2016; Zadbood et al., 2017) have extended this approach
575 with a clever twist. Rather presented a clever alternative approach: rather than building an
576 explicit stimulus model, these studies instead search for brain responses (while experiencing
577 the stimulus) to the stimulus that are reliably similar across individuals. So called *inter-subject*
578 *correlation* (ISC) and *inter-subject functional connectivity* (ISFC) analyses effectively treat other peo-
579 ple's brain responses to the stimulus as a "model" of how its features change over time. By

580 contrast, in our present work we used topic models and HMMs to construct an explicit stimulus
581 model (i.e., the topic trajectory of the video). When we searched for brain structures whose
582 responses are consistent with the video's topic trajectory, we identified a network of structures that
583 overlapped strongly with the "long temporal receptive window" network reported by the Hasson
584 group (e.g., compare our Fig. 8A with the map of long temporal receptive window voxels in Lerner et al., 2011)
585 . This provides support for the notion that part of the long temporal receptive window network
586 may be maintaining an explicit model of the stimulus dynamics. When we performed a similar
587 analysis after swapping out the video's topic trajectory with the recall topic trajectories of each
588 individual participant, this allowed us to identify brain regions whose responses (as the participants
589 viewed the video) reflected how the video trajectory would be transformed in memory (as
590 reflected by the recall topic trajectories). The analysis revealed that the rMTL and vmPFC may
591 play a role in this person-specific transformation from experience into memory. The role of
592 (also see Simony and Chang, 2020). These purely brain-driven approaches are well suited to
593 identifying which brain structures exhibit similar stimulus-driven responses across individuals.
594 Further, because neural response dynamics are observed data (rather than model approximations),
595 such approaches do not require a detailed understanding of which stimulus properties or features
596 might be driving the observed responses. However, this also means that the MTL in episodic
597 memory encoding has been well-reported (e.g., Davachi, 2006; Davachi et al., 2003; Paller and Wagner, 2002; ?)
598 . Prior work has also implicated the medial prefrontal cortex in representing "schema" knowledge (i.e., general knowledge
599 . Integrating across our study and this prior work, one interpretation is that the person-specific
600 transformations mediated (or represented) by the rMTL and vmPFC may reflect schema knowledge
601 being leveraged, formed, or updated, incorporating ongoing experience into previously acquired
602 knowledge. Specific stimulus features driving those responses are typically opaque to the researcher.
603 Our approach is complementary. By explicitly modeling the stimulus dynamics, we are able to
604 relate specific stimulus features to behavioral and neural dynamics. However, when our model
605 fails to accurately capture the stimulus dynamics that are truly driving behavioral and neural
606 responses, our approach necessarily yields an incomplete characterization of the neural basis of
607 the processes we are studying.

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

615 Our approach also draws inspiration from the growing field of word embedding models.
616 The topic models (Blei et al., 2003) we used to embed text from the episode annotations and
617 participants’ recall transcripts are just one of many models that have been studied in an extensive
618 literature. The earliest approaches to word embedding, including latent semantic analysis (Landauer and Dumais, 1997)
619 , used word co-occurrence statistics (i.e., how often pairs of words occur in the same documents
620 contained in the corpus) to derive a unique feature vector for each word. The feature vectors
621 are constructed so that words that co-occur more frequently have feature vectors that are closer
622 (in Euclidean distance). Topic models are essentially an extension of those early models, in that
623 they attempt to explicitly model the underlying causes of word co-occurrences by automatically
624 identifying the set of themes or topics reflected across the documents in the corpus. More recent
625 work on these types of semantic models, including word2vec (Mikolov et al., 2013), the Universal
626 Sentence Encoder (Cer et al., 2018), GPT-2 (Radford et al., 2019), and GTP-3 (Brown et al., 2020)
627 use deep neural networks to attempt to identify the deeper conceptual representations underlying
628 each word. Despite the growing popularity of these sophisticated deep learning-based embedding
629 models, we chose to prioritize interpretability of the embedding dimensions (e.g., Fig. 7) over
630 raw performance (e.g., with respect to some predefined benchmark). Nevertheless, we note that
631 our general framework is, in principle, robust to the specific choice of language model as well as
632 other aspects of our computational pipeline. For example, the word embedding model, timeseries
633 segmentation model, and the episode-recall matching function could each be customized to suit
634 a particular question space or application. Indeed, for some questions, interpretability of the
635 embeddings may not be a priority, and thus other text embedding approaches (including the deep

636 learning-based models described above) may be preferable. Further work will be needed to explore
637 the influence of particular models on our framework's predictions and performance.

638 Our work has broad implications for how we characterize and assess memory in real-world
639 settings, such as the classroom or physician's office. For example, the most commonly used
640 classroom evaluation tools involve simply computing the proportion of correctly answered exam
641 questions. Our work indicates that this approach is only loosely related to what educators might
642 really want to measure: how well did the students understand the key ideas presented in the
643 course? One could apply the computational framework we developed to construct topic trajectories
644 for the video and participants' recalls to build explicit Under this typical framework of assessment,
645 the same exam score of 50% could be ascribed to two very different students: one who attended
646 to the full course but struggled to learn more than a broad overview of the material, and one who
647 attended to only half of the course but understood the attended material perfectly. Instead, one
648 could apply our computational framework to build explicit dynamic content models of the course
649 material and exam questions. This approach would provide a more nuanced and specific view into
650 which aspects of the material students had learned well (or poorly). In clinical settings, memory
651 measures that incorporate such explicit content models might also provide more direct evaluations
652 of patients' memories, and of doctor-patient interactions.

653 Methods

654 Experimental design Paradigm and data collection

655 Data were collected by Chen et al. (2017). In brief, participants ($n = 17$) viewed the first 48
656 minutes of "A Study in Pink," the first episode of the BBC television series show *Sherlock*, while
657 fMRI volumes were collected (TR = 1500 ms). Participants were pre-screened to ensure they had
658 never seen any episode of the show before. The stimulus was divided into a 23 min (946 TR) and
659 a 25 min (1030 TR) segment to mitigate technical issues related to the scanner. After finishing the
660 clip, participants were instructed to (quoting from Chen et al., 2017) "describe what they recalled

661 of the [episode] in as much detail as they could, to try to recount events in the original order they
662 were viewed in, and to speak for at least 10 minutes if possible but that longer was better. They
663 were told that completeness and detail were more important than temporal order, and that if at
664 any point they realized they had missed something, to return to it. Participants were then allowed
665 to speak for as long as they wished, and verbally indicated when they were finished (e.g., 'I'm
666 done')." Five participants were dropped from the original dataset due to excessive head motion (2
participants), insufficient recall length (2 participants), or falling asleep during stimulus viewing (1
participant), resulting in a final sample size of $n = 17$. For additional details about the experimental
procedure testing procedures and scanning parameters, see Chen et al. (2017). The experimental
testing protocol was approved by Princeton University's Institutional Review Board.

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

677 The video stimulus was divided into 1,000 fine-grained "time segments" and annotated by an
678 independent coder. For each of these 1,000 annotations, the following information was recorded:
679 a brief narrative description of what was happening, the location where the time segment took
680 place, whether that location was indoors or outdoors, the names of all characters on-screen, the
681 name(s) of the character(s) in focus in the shot, the name(s) of the character(s) currently speaking,
682 the camera angle of the shot, a transcription of any text appearing on-screen, and whether or not
683 there was music present in the background. Each time segment was also tagged with its onset and
684 offset time, in both seconds and TRs.

685 Data and code availability

686 The fMRI data we analyzed are available online here. The behavioral data and all of our analysis
687 code may be downloaded here.

688 **Statistics**

689 All statistical tests ~~we performed~~ performed in the behavioral analyses were two-sided. All
690 statistical tests performed in the neural data analyses were two-sided, except for the permutation-based
691 thresholding, which was one-sided. In this case, we were specifically interested in identifying
692 voxels whose activation time series reflected the temporal structure of the episode and recall topic
693 proportions matrices to a greater extent than that of the phase-shifted matrices.

694 **Modeling the dynamic content of the video episode and recall transcripts**

695 **Topic modeling**

696 The input to the topic model we trained to characterize the dynamic content of the ~~video comprised~~
697 ~~episode comprised~~ 998 hand-generated annotations of ~~each of 1000 scenes~~ short (mean: 2.96s) time
698 ~~segments~~ spanning the video clip (~~generated by Chen et al., 2017~~). The features included: narrative
699 details (~~a sentence or two describing what happened in that scene~~); whether the scene took place
700 indoors or outdoors; names of any characters that appeared in the scene; name(s) of characters in
701 camera focus; name(s) of characters who were speaking in the scene; the location (in the story) that
702 the scene took place; camera angle (close up, medium, long, top, tracking, over the shoulder,
703 etc.); whether music was playing in the scene or not; and a transcription of any on-screen text.
704 (~~Chen et al., 2017 generated 1000 annotations total; we removed two annotations referring to a~~
705 ~~break between the first and second scan sessions, during which no fMRI data were collected~~). We
706 concatenated the text for all of ~~these~~ the annotated features within each segment, creating a “bag of
707 words” describing ~~each scene~~ its content and performed some minor preprocessing (e.g., stemming
708 possessive nouns and removing punctuation). We then re-organized the text descriptions into
709 overlapping sliding windows spanning ~~(up to)~~ 50 ~~scenes~~ annotations each. In other words, ~~the~~
710 ~~first text sample comprised the combined text from the first 50 scenes~~ we estimated the “context”
711 ~~for each annotated segment using the text descriptions of the preceding 25 annotations, the present~~
712 ~~annotations, and the following 24 annotations~~. To model the context for annotations near the
713 beginning of the episode (i.e., 1–50), the second comprised the text from scenes 2–51, and so on.

714 within 25 of the beginning or end), we created overlapping sliding windows that grew in size from
715 one annotation to the full length. We also tapered the sliding window lengths at the end of the
716 episode, whereby time segments within fewer than 24 annotations of the end of the episode were
717 assigned sliding windows that extended to the end of the episode. This procedure ensured that
718 each annotation's content was represented in the text corpus an equal number of times.

719 We trained our model using these overlapping text samples with scikit-learn (version 0.19.1;
720 Pedregosa et al., 2011), called from our high-dimensional visualization and text analysis soft-
721 ware, HyperTools (Heusser et al., 2018b). Specifically, we ~~use~~used the CountVectorizer class
722 to transform the text from each ~~scene~~window into a vector of word counts (using the union
723 of all words across all ~~scenes~~annotations as the “vocabulary,” excluding English stop words);
724 this ~~yields a number-of-scenes~~yielded a number-of-windows by number-of-words *word count*
725 matrix. We then ~~use~~used the LatentDirichletAllocation class (topics=100, method='batch')
726 to fit a topic model (Blei et al., 2003) to the word count matrix, yielding a ~~number-of-scenes~~1000
727 ~~number-of-windows (1047)~~ by number-of-topics (100) *topic proportions* matrix. The topic pro-
728 portions matrix describes ~~which~~the gradually evolving mix of topics (latent themes) ~~is~~-present
729 in each ~~scene~~annotated time segment of the episode. Next, we transformed the topic proportions
730 matrix to match the 1976 fMRI volume acquisition times. ~~For each fMRI volume, we took the topic~~
731 ~~proportions from whatever scene was displayed for most of that volume's 1500 ms acquisition time.~~
732 ~~This yielded a new~~We assigned each topic vector to the timepoint (in seconds) midway between
733 ~~the beginning of the first annotation and the end of the last annotation in its corresponding sliding~~
734 ~~text window. By doing so, we warped the linear temporal distance between consecutive topic~~
735 ~~vectors to align with the inconsistent temporal distance between consecutive annotations (whose~~
736 ~~durations varied greatly). We then rescaled these timepoints to 1.5s TR units, and used linear~~
737 ~~interpolation to estimate a topic vector for each TR. This resulted in a~~ number-of-TRs (1976) by
738 number-of-topics (100) ~~topic proportions~~ matrix.

739 We created similar topic proportions matrices using hand-annotated transcripts of each partici-
740 pant's ~~verbal~~ recall of the ~~video (annotated by Chen et al., 2017)~~ episode ~~(annotated by Chen et al., 2017)~~
741 . We tokenized the transcript into a list of sentences, and then re-organized the list into overlapping

742 sliding windows spanning (up to) 10 sentences each; ~~in turn~~, analogously to how we parsed the
743 episode annotations. In turn, we transformed each window's sentences into a word count vector
744 (using the same vocabulary as for the video model). We episode model), then used the topic
745 model already trained on the video episode scenes to compute the most probable topic propor-
746 tions for each sliding window. This yielded a ~~number-of-sentences~~~~number-of-windows~~ (range:
747 68–29483–312) by number-of-topics (100) topic proportions matrix ~~r~~ for each participant. These
748 reflected the dynamic content of each participant's recalls. Note: for details on how we selected
749 the video episode and recall window lengths and number of topics, see *Supporting Information* and
750 Figure S1.

751 **Parsing Segmenting topic trajectories proportions matrices into discrete events using Hidden
752 hidden Markov Models**

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

761 We used an optimization procedure to select the appropriate K for each topic proportions matrix.
762 Prior studies on narrative structure and processing have shown that we both perceive and internally
763 represent the world around us at multiple, hierarchical timescales (e.g., Baldassano et al., 2017, 2018; Chen et al., 2017; H
764 . However, for the purposes of our framework, we sought to identify the single timeseries of
765 event-representations that is emphasized most heavily in the temporal structure of the episode
766 and of each participant's recall. We quantified this as the set of K states that maximized the
767 similarity between topic vectors for timepoints comprising each state, while minimizing the
768 similarity between topic vectors for timepoints across different states. Specifically, we computed

769 (for each matrix)

$$\operatorname{argmax}_K \left[\frac{a}{b} - \frac{K}{\alpha} W_1(a, b) \right],$$

770 where a was the average correlation between the topic vectors of timepoints within the same state;
771 distribution of within-state topic vector correlations, and b was the average correlation between
772 the topic vectors of timepoints within different states; and α was a regularization parameter that we
773 set to 5 times the window length (i.e., 250 scenes for the video topic trajectory and distribution of
774 across-state topic vector correlations). We computed the first Wasserstein distance (W_1 ; also known
775 as *Earth mover's distance*; Dobrushin, 1970; Ramdas et al., 2017) between these distributions for a
776 large range of possible K -values (range [2, 50 sentences for the recall topic trajectories]), and selected
777 the K that yielded the maximum value. Figure 2B displays the event boundaries returned for the
778 video episode, and Figure S4 displays the event boundaries returned for each participant's recalls.
779 See Figure S6 for the optimization functions for the episode and recalls. After obtaining these event
780 boundaries, we created stable estimates of each topic proportions matrix—the content represented
781 in each event by averaging the topic vectors within each event across timepoints between each pair
782 of event boundaries. This yielded a number-of-events by number-of-topics matrix for the video
783 episode and recalls from each participant.

784 **Visualizing the video and recall topic trajectories** We used the Naturalistic extensions of classic
785 list-learning analyses

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

792 Perhaps the simplest and most widely used measure of memory performance is *accuracy*—i.e.,
793 the proportion of studied (experienced) items (in this case, episode events) that the participant later

remembered. Chen et al. (2017) used this method to rate each participant's memory quality by computing the proportion of (50, manually identified) scenes mentioned in their recall. We found a strong across-participants correlation between these independent ratings and the proportion of 30 HMM-identified episode events matched to participants' recalls (Pearson's $r(15) = 0.71, p = 0.002$). We further considered a number of more nuanced memory performance measures that are typically associated with list-learning studies. We also provide a software package, Quail, for carrying out these analyses (Heusser et al., 2017).

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

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

819 **Serial position curve (SPC).** SPCs (Murdock, 1962) reflect the proportion of participants that
820 remember each item as a function of the item's serial position during encoding. We initialized
821 a number-of-participants (17) by number-of-episode-events (30) matrix of zeros. Then, for each
822 recalled event, for each participant, we found the index of the episode event that the recalled
823 event most closely matched (via the correlation between the events' topic vectors) and entered a
824 1 into that position in the matrix. This resulted in a matrix whose entries indicated whether or
825 not each event was recalled by each participant (depending on whether the corresponding entires
826 were set to one or zero). Finally, we averaged over the rows of the matrix to yield a 1 by 30 array
827 representing the proportion of participants that recalled each event as a function of the events'
828 order appearance in the episode (Fig. 3C).

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

838 **Semantic clustering scores.** Semantic clustering describes a participant's tendency to recall
839 semantically similar presented items together in their recall sequences. Here, we used the topic
840 vectors for each event as a proxy for its semantic content. Thus, the similarity between the semantic
841 content for two events can be computed by correlating their respective topic vectors. For each recall
842 event transition, we sorted all not-yet-recalled events according to how correlated the topic vector
843 of the closest-matching episode event was to the topic vector of the closest-matching episode event
844 to the just-recalled event. We then computed the percentile rank of the observed next recall. We

845 averaged these percentile ranks across all of the participant's recalls to obtain a single semantic
846 clustering score for the participant.

847 **Averaging correlations**

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

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

854 We used the UMAP algorithm (2) (McInnes et al., 2018) to project the 100-dimensional topic space
855 onto a two-dimensional space for visualization (Figs. 6, 7). To ensure that all of the trajectories were
856 projected onto the *same* lower dimensional space, we computed the low-dimensional embedding
857 on a "stacked" matrix created by vertically concatenating the events-by-topics topic proportions
858 matrices for the *video episode, across-participants average recall* and all 17 *individual* participants'
859 recalls. We then *divided-separated* the rows of the result (a total-number-of-events by two ma-
860 trix) back into *separate individual* matrices for the *video topic trajectory**episode topic trajectory*,
861 *across-participant average recall trajectory*, and the trajectories for each *individual* participant's
862 recalls (Fig. 6). This general approach for discovering a shared low-dimensional embedding for a
863 collections of high-dimensional observations follows Heusser et al. (2018b).

864 We optimized the manifold space for visualization based on two criteria: First, that the
865 2D embedding of the episode trajectory should reflect its original 100-dimensional structure as
866 faithfully as possible. Second, that the path traversed by the embedded episode trajectory should
867 intersect itself a minimal number of times. The first criteria helps bolster the validity of visual
868 intuitions about relationships between sections of episode content, based on their locations in the
869 embedding space. The second criteria was motivated by the observed low off-diagonal values in the
870 episode trajectory's temporal correlation matrix (suggesting that the same topic-space coordinates

871 should not be revisited; see Fig. 2A). For further details on how we created this low-dimensional
872 embedding space, see *Supporting Information*.

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

874 In Figure 6B, we present an analysis aimed at characterizing locations in topic space that dif-
875 ferent participants move through in a consistent way (via their recall topic trajectories). The
876 two-dimensional topic space used in our visualizations (Fig. 6) ranged from -5 to 5 (arbitrary)
877 units in the x dimension and from -6.5 to 2 units in the y dimension. We divided this space into a
878 grid of vertices spaced 0.25 units apart comprised a 60×60 (arbitrary units) square. We tiled this
879 space with a 50×50 grid of evenly spaced vertices, and defined a circular area centered on each
880 vertex whose radius was two times the distance between adjacent vertices (i.e., 2.4 units). For each
881 vertex, we examined the set of line segments formed by connecting each pair successively recalled
882 events, across all participants, that passed within 0.5 units through this circle. We computed the
883 distribution of angles formed by those segments and the x-axis, and used a Rayleigh test to deter-
884 mine whether the distribution of angles was reliably “peaked” (i.e., consistent across all transitions
885 that passed through that local portion of topic space). To create Figure 6B, we drew an arrow
886 originating from each grid vertex, pointing in the direction of the average angle formed by the line
887 segments that passed within 0.5–2.4 units. We set the arrow lengths to be inversely proportional
888 to the p-values of the Rayleigh tests at each vertex. Specifically, for each vertex we converted all
889 of the angles of segments that passed within 0.5–2.4 units to unit vectors, and we set the arrow
890 lengths at each vertex proportional to the length of the (circular) mean vector. We also indicated
891 any significant results ($p < 0.05$, corrected using the Benjamani-Hochberg procedure) by coloring
892 the arrows in blue (darker blue denotes a lower p-value, i.e., a longer mean vector); all tests with
893 $p \geq 0.05$ are displayed in gray and given a lower opacity value.

894 **Searchlight fMRI analyses**

895 In Figure 8, we present two analyses aimed at identifying brain structures regions whose responses
896 (as participants viewed the video) exhibited particular temporal correlations episode) exhibited

897 a particular temporal structure. We developed a searchlight analysis whereby wherein we con-
898 structed a cube $5 \times 5 \times 5$ cube of voxels (following Chen et al., 2017) centered on each voxel (radius:
899 5 voxels). For in the brain, and for each of these cubes, we computed the temporal correlation ma-
900 trix of the voxel responses during video episode viewing. Specifically, for each of the 1976 volumes
901 collected during video episode viewing, we correlated the activity patterns in the given cube with
902 the activity patterns (in the same cube) collected during every other timepoint. This yielded a 1976
903 by 1976 1976×1976 correlation matrix for each cube. Note: participant 5's scan ended 75s early,
904 and in Chen et al. (2017)'s publicly released dataset, their scan data was zero-padded to match the
905 length of the other participants'. For our searchlight analyses, we removed this padded data (i.e.,
906 the last 50 TRs), resulting in a 1925 \times 1925 correlation matrix for each cube in participant 5's brain.

907 Next, we constructed two sets a series of "template" matrices: one reflected the video. The
908 first template reflected the timecourse of the episode's topic trajectory and the other reflected
909 proportions matrix, and the others reflected the timecourse of each participant's recall topic
910 trajectory proportions matrix. To construct the video episode template, we computed the cor-
911 relations between the topic proportions estimated for every pair of TRs (prior to segmenting the
912 trajectory topic proportions matrices) into discrete events; i.e., the correlation matrix shown in
913 Figs. 2B and 8A). We constructed similar temporal correlation matrices for each participant's recall
914 topic trajectory proportions matrix (Figs. 2D, S4). However, to correct for length differences and
915 potential non-linear transformations between viewing time and recall time, we first used dynamic
916 time warping (Berndt and Clifford, 1994) to temporally align participants' recall topic trajectories
917 with the video topic trajectory (an proportions matrices with the episode topic proportions matrix).
918 An example correlation matrix before and after warping is shown in Fig. 8B). This yielded a 1976
919 by 1976 1976×1976 correlation matrix for the video episode template and for each participant's
920 recall template.

921 The temporal structure of the episode's content (as described by our model) is captured in
922 the block-diagonal structure of the episode's temporal correlation matrix (e.g., Figs. 2B, 8A),
923 with time periods of thematic stability represented as dark blocks of varying sizes. Inspecting
924 the episode correlation matrix suggests that the episode's semantic content is highly temporally

925 specific (i.e., the correlations between topic vectors from distant timepoints are almost all near
926 zero). By contrast, the activity patterns of individual (cubes of) voxels can encode relatively limited
927 information on their own, and their activity frequently contributes to multiple separate functions
928 (Charron and Koechlin, 2010; Freedman et al., 2001; Rishel et al., 2013; Sigman and Dehaene, 2008)
929 . By nature, these two attributes give rise to similarities in activity across large timescales that may
930 not necessarily reflect a single task. To enable a more sensitive analysis of brain regions whose shifts
931 in activity patterns mirrored shifts in the semantic content of the episode or recalls, we restricted
932 the temporal correlations we considered to the timescale of semantic information captured by our
933 model. Specifically, we isolated the upper triangle of the episode correlation matrix and created a
934 “proximal correlation mask” that included only diagonals from the upper triangle of the episode
935 correlation matrix up to the first diagonal that contained no positive correlations. Applying this
936 mask to the full episode correlation matrix was equivalent to excluding diagonals beyond the
937 corner of the largest diagonal block. In other words, the timescale of temporal correlations we
938 considered corresponded to the longest period of thematic stability in the episode, and by extension
939 the longest period of thematic stability in participants’ recalls and the longest period of stability we
940 might expect to see in voxel activity arising from processing or encoding episode content. Figure 8
941 shows this proximal correlation mask applied to the temporal correlation matrices for the episode,
942 an example participant’s (warped) recall, and an example cube of voxels from our searchlight
943 analyses.

944 To determine which (cubes of) voxel responses ~~reliably matched the video~~ matched the episode
945 template, we correlated the ~~proximal diagonals from the~~ upper triangle of the voxel correlation
946 matrix for each cube with the ~~upper triangle of the video~~ ~~proximal diagonals from episode~~ template
947 matrix (Kriegeskorte et al., 2008). This yielded, for each participant, a ~~single correlation value~~. We
948 ~~computed the average~~ voxelwise map of correlation values. We then performed a one-sample
949 *t*-test on the distribution of (Fisher *z*-transformed) ~~correlation coefficient~~ correlations at each voxel,
950 across participants. ~~We used a permutation-based procedure to assess significance~~ This resulted in
951 a value for each voxel (cube), describing how reliably its timecourse followed that of the episode.
952 We further sought to ensure that our analysis identified regions where the activations’ temporal

953 structure specifically reflected that of the episode, rather than regions whose activity was simply
954 autocorrelated at a timescale similar to the episode template's diagonal. To achieve this, we used
955 a phase shift-based permutation procedure, whereby we re-computed the average correlations for
956 each of 100 "null" video templates (constructed by circularly shifting the template circularly shifted
957 the episode's topic proportions matrix by a random number of timepoints) (rows), computed
958 the resulting "null" episode template, and re-ran the searchlight analysis, in full. (For each
959 permutation of the 100 permutations, the same random shift was used for all participants). We
960 then). We z-scored the observed (unshifted) result at each voxel against the distribution of
961 permutation-derived "null" results, and estimated a p -value by computing the proportion of shifted
962 correlations that were larger than the observed (unshifted) correlation results that yielded larger
963 values. To create the map in Figure 8A-C, we thresholded out any voxels whose correlation values
964 similarity to the unshifted episode's structure fell below the 95th percentile of the permutation-
965 derived null distribution similarity results.

966 We used a similar an analogous procedure to identify which voxels' responses reflected the
967 recall templates. For each participant, we correlated the proximal diagonals from the upper
968 triangle of the correlation matrix for each cube of voxels with their (time-warped the proximal
969 diagonals from the upper triangle of their (time-warped) recall correlation matrix. As in the video
970 template analysis episode template analysis, this yielded a single correlation coefficient voxelwise
971 map of correlation coefficients for each participant. However, whereas the video episode analysis
972 compared every participant's responses to the same template, here the recall templates were unique
973 for each participant. We computed the average As in the analysis described above, we t-scored
974 the (Fisher) z-transformed correlation coefficient across participants voxelwise correlations, and
975 used the same permutation procedure we developed for the video responses to assess significant
976 correlations episode responses to ensure specificity to the recall timeseries and assign significance
977 values. To create the map in Figure 8B-we D we again thresholded out any voxels whose correlation
978 values fell scores were below the 95th percentile of the permutation-derived null distribution.

979 **Neurosynth decoding analyses**

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

989 **References**

- 990 Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control
991 processes. In Spence, K. W. and Spence, J. T., editors, *The psychology of learning and motivation*,
992 volume 2, pages 89–105. Academic Press, New York.
- 993 Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017).
994 Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–
995 721.
- 996 Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas
997 during narrative perception. *Journal of Neuroscience*, 38(45):9689–9699.
- 998 Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In
999 *KDD workshop*, volume 10, pages 359–370.
- 1000 Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International*
1001 *Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, US. ACM.

- 1002 Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine*
1003 *Learning Research*, 3:993 – 1022.
- 1004 Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam,
1005 P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R.,
1006 Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S.,
1007 Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020).
1008 Language models are few-shot learners. *arXiv*, 2005.14165.
- 1009 Brunec, I. K., Moscovitch, M. M., and Barense, M. D. (2018). Boundaries shape cognitive represen-
1010 tations of spaces and events. *Trends in Cognitive Sciences*, 22(7):637–650.
- 1011 Capota, M., Turek, J., Chen, P.-H., Zhu, X., Manning, J. R., Sundaram, N., Keller, B., Wang, Y., and
1012 Shin, Y. S. (2017). Brain imaging analysis kit.
- 1013 Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes,
1014 M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder.
1015 *arXiv*, 1803.11175.
- 1016 Charron, S. and Koechlin, E. (2010). Divided representations of current goals in the human frontal
1017 lobes. *Science*, 328(5976):360–363.
- 1018 Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared
1019 memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*,
1020 20(1):115.
- 1021 Clewett, D. and Davachi, L. (2017). The ebb and flow of experience determines the temporal
1022 structure of memory. *Curr Opin Behav Sci*, 17:186–193.
- 1023 Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in*
1024 *Neurobiology*, 16(6):693—700.
- 1025 Davachi, L., Mitchell, J. P., and Wagner, A. D. (2003). Multiple routes to memory: distinct medial

- 1026 temporal lobe processes build item and source memories. *Proceedings of the National Academy of*
1027 *Sciences, USA*, 100(4):2157 – 2162.
- 1028 Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions.
1029 *Theory of Probability & Its Applications*, 15(3):458–486.
- 1030 DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the
1031 sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–1286.
- 1032 Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological*
1033 *Science*, 22(2):243–252.
- 1034 Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Oliver and Boyd.
- 1035 Freedman, D., Riesenhuber, M., Poggio, T., and Miller, E. (2001). Categorical representation of
1036 visual stimuli in the primate prefrontal cortex. *Science*, 291(5502):312–316.
- 1037 Gilboa, A. and Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory.
1038 *Trends Cogn Sci*, 21(8):618–631.
- 1039 Hasson, U., Chen, J., and Honey, C. J. (2015). Hierarchical process memory: memory as an integral
1040 component of information processing. *Trends in Cognitive Science*, 19(6):304–315.
- 1041 Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal
1042 receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550.
- 1043 Heusser, A. C., Ezzyat, Y., Shiff, I., and Davachi, L. (2018a). Perceptual boundaries cause mnemonic
1044 trade-offs between local boundary processing and across-trial associative binding. *Journal of*
1045 *Experimental Psychology Learning, Memory, and Cognition*, 44(7):1075–1090.
- 1046 Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail: a
1047 Python toolbox for analyzing and plotting free recall data. *The Journal of Open Source Software*,
1048 10.21105/joss.00424.

- 1049 Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018b). HyperTools: a Python
1050 toolbox for gaining geometric insights into high-dimensional data. *Journal of Machine Learning*
1051 *Research*, 18(152):1–6.
- 1052 Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal*
1053 *of Mathematical Psychology*, 46:269–299.
- 1054 Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., and H., E.
1055 (2014). A unified mathematical framework for coding time, space, and sequences in the medial
1056 temporal lobe. *Journal of Neuroscience*, 34(13):4692–4707.
- 1057 Howard, M. W., Viskontas, I. V., Shankar, K. H., and Fried, I. (2012). Ensembles of human MTL
1058 neurons “jump back in time” in response to a repeated stimulus. *Hippocampus*, 22:1833–1847.
- 1059 Huk, A., Bonnen, K., and He, B. J. (2018). Beyond trial-based paradigms: continuous behavior, on-
1060 going neural activity, and naturalistic stimuli. *Journal of Neuroscience*, 10.1523/JNEUROSCI.1920-
1061 17.2018.
- 1062 Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., and Gallant, J. L. (2016). Natural
1063 speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532:453–458.
- 1064 Huth, A. G., Nisimoto, S., Vu, A. T., and Gallant, J. L. (2012). A continuous semantic space describes
1065 the representation of thousands of object and action categories across the human brain. *Neuron*,
1066 76(6):1210–1224.
- 1067 Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24:103–109.
- 1068 Kahana, M. J. (2012). *Foundations of Human Memory*. Oxford University Press, New York, NY.
- 1069 Koriat, A. and Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: distin-
1070 guishing accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of*
1071 *Experimental Psychology: General*, 123(3):297–315.
- 1072 Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – con-
1073 necting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:1 – 28.

- 1074 Landauer, T. K. and Dumais, S. T. (1997). A solution to Plato's problem: the latent semantic
1075 analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*,
1076 104:211–240.
- 1077 Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy
1078 of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.
- 1079 Manning, J. R. (2019). Episodic memory: mental time travel or a quantum 'memory wave' function?
1080 *PsyArXiv*, doi:10.31234/osf.io/6zjwb.
- 1081 Manning, J. R. (2020). Context reinstatement. In Kahana, M. J. and Wagner, A. D., editors, *Handbook*
1082 *of Human Memory*. Oxford University Press.
- 1083 Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory.
1084 In Gazzaniga, M., editor, *The Cognitive Neurosciences, Fifth edition*, pages 557–566. MIT Press.
- 1085 Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns
1086 in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National*
1087 *Academy of Sciences, USA*, 108(31):12893–12897.
- 1088 McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and
1089 projection for dimension reduction. *arXiv*, 1802(03426).
- 1090 Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations
1091 in vector space. *arXiv*, 1301.3781.
- 1092 Mueller, A., Fillion-Robin, J.-C., Boidol, R., Tian, F., Nechifor, P., yoonsubKim, Peter, Rampin, R.,
1093 Corvellec, M., Medina, J., Dai, Y., Petrushev, B., Langner, K. M., Hong, Alessio, Ozsvald, I.,
1094 vkolmakov, Jones, T., Bailey, E., Rho, V., IgorAPM, Roy, D., May, C., foobuzz, Piyush, Seong,
1095 L. K., Goey, J. V., Smith, J. S., Gus, and Mai, F. (2018). WordCloud 1.5.0: a little word cloud
1096 generator in Python. *Zenodo*, <https://zenodo.org/record/1322068#.W4tPKZNKh24>.
- 1097 Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*,
1098 64:482–488.

- 1099 Paller, K. A. and Wagner, A. D. (2002). Observing the transformation of experience into memory.
1100 *Trends in Cognitive Sciences*, 6(2):93–102.
- 1101 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Pretten-
1102 hofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot,
1103 M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine
1104 Learning Research*, 12:2825–2830.
- 1105 Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model
1106 of organizational processes in free recall. *Psychological Review*, 116(1):129–156.
- 1107 Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal
1108 of Experimental Psychology*, 17:132–138.
- 1109 Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech
1110 recognition. *Proceedings of the IEEE*, 77(2):257–286.
- 1111 Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are
1112 unsupervised multitask learners. *OpenAI Blog*, 1(8).
- 1113 Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition. *Curr Opin
1114 Behav Sci*, 17:133–140.
- 1115 Ramdas, A., Trillos, N., and Cuturi, M. (2017). On wasserstein two-sample testing and related
1116 families of nonparametric tests. *Entropy*, 19(2):47.
- 1117 Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature
1118 Reviews Neuroscience*, 13:713 – 726.
- 1119 Rishel, C. A., Huang, G., and Freedman, D. J. (2013). Independent category and spatial encoding
1120 in parietal cortex. *Neuron*, 77(5):969–979.
- 1121 Schlichting, M. L. and Preston, A. R. (2015). Memory integration: neural mechanisms and impli-
1122 cations for behavior. *Current Opinion in Behavioral Sciences*, 1:1–8.

- 1123 Sigman, M. and Dehaene, S. (2008). Brain mechanisms of serial and parallel processing during
1124 dual-task performance. *Journal of Neuroscience*, 28(30):7585–7589.
- 1125 Simony, E. and Chang, C. (2020). Analysis of stimulus-induced brain dynamics during naturalistic
1126 paradigms. *NeuroImage*, 216:116461.
- 1127 Simony, E., Honey, C. J., Chen, J., and Hasson, U. (2016). Uncovering stimulus-locked network
1128 dynamics during narrative comprehension. *Nature Communications*, 7(12141):1–13.
- 1129 Spalding, K. N., Schlichting, M. L., Zeithamova, D., Preston, A. R., Tranel, D., Duff, M. C., and
1130 Warren, D. E. (2018). Ventromedial prefrontal cortex is necessary for normal associative inference
1131 and memory integration. *The Journal of Neuroscience*, 38(15):3767–3775.
- 1132 Tomrary, A. and Davachi, L. (2017). Consolidation promotes the emergence of representational
1133 overlap in the hippocampus and medial prefrontal cortex. *Neuron*, 96(1):228–241.
- 1134 van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., and Henson, R. N. (2012). How schema and
1135 novelty augment memory formation. *Trends Neurosci*, 35(4):211–9.
- 1136 Waskom, M., Botvinnik, O., Okane, D., Hobson, P., David, Halchenko, Y., Lukauskas, S., Cole, J. B.,
1137 Warmenhoven, J., de Ruiter, J., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E.,
1138 Martin, M., Miles, A., Meyer, K., Augspurger, T., Yarkoni, T., Bachant, P., Williams, M., Evans,
1139 C., Fitzgerald, C., Brian, Wehner, D., Hitz, G., Ziegler, E., Qalieh, A., and Lee, A. (2016). Seaborn:
1140 v0.7.1.
- 1141 Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American Journal
1142 of Psychology*, 35:396–401.
- 1143 Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., and Wager, T. D. (2011). Large-scale
1144 automated synthesis of human functional neuroimaging data. *Nature Methods*, 8(8):665.
- 1145 Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research.
1146 *Journal of Memory and Language*, 46:441–517.

- 1147 Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception:
1148 a mind-brain perspective. *Psychological Bulletin*, 133:273–293.
- 1149 Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit
1150 memories to other brains: Constructing shared neural representations via communication. *Cereb*
1151 *Cortex*, 27(10):4988–5000.
- 1152 Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and
1153 memory. *Psychological Bulletin*, 123(2):162 – 185.

1154 Supporting information

1155 Supporting information is available in the online version of the paper.

1156 Acknowledgements

1157 We thank Luke Chang, Janice Chen, Chris Honey, Lucy Owen, Emily Whitaker, and Kirsten Ziman
1158 for feedback and scientific discussions. We also thank Janice Chen, Yuan Chang Leong, Kenneth
1159 Norman, and Uri Hasson for sharing the data used in our study. Our work was supported in part
1160 by NSF EPSCoR Award Number 1632738. The content is solely the responsibility of the authors
1161 and does not necessarily represent the official views of our supporting organizations.

1162 Author contributions

1163 Conceptualization: A.C.H. and J.R.M.; Methodology: A.C.H., P.C.F. and J.R.M.; Software: A.C.H.,
1164 P.C.F. and J.R.M.; Analysis: A.C.H., P.C.F. and J.R.M.; Writing, Reviewing, and Editing: A.C.H.,
1165 P.C.F. and J.R.M.; Supervision: J.R.M.

1166 **Author information**

1167 The authors declare no competing financial interests. Correspondence and requests for materials
1168 should be addressed to J.R.M. (jeremy.r.manning@dartmouth.edu).