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

Andrew C. Heusser[1, 2], Paxton C. Fitzpatrick[1], and Jeremy R. Manning[1, *]

[1]Department of Psychological and Brain Sciences

Dartmouth College, Hanover, NH 03755, USA

[2]Akili Interactive

Boston, MA 02110

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

February 4, 2020

Abstract

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

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

Introduction

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

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

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

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

Results

To characterize the "essence" of the *Sherlock* episode and participants' subsequent recountings of its unfolding, we used a topic model (Blei et al., 2003) to discover the latent themes in the episode's dynamic content. Topic models take as inputs a vocabulary of words to consider and a collection of text documents, and return two output matrices. The first of these is a *topics matrix* whose rows are topics (latent themes) and whose columns correspond to words in the vocabulary. The entries of the topics matrix define 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 what mixture of discovered topics is reflected in each document.

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

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

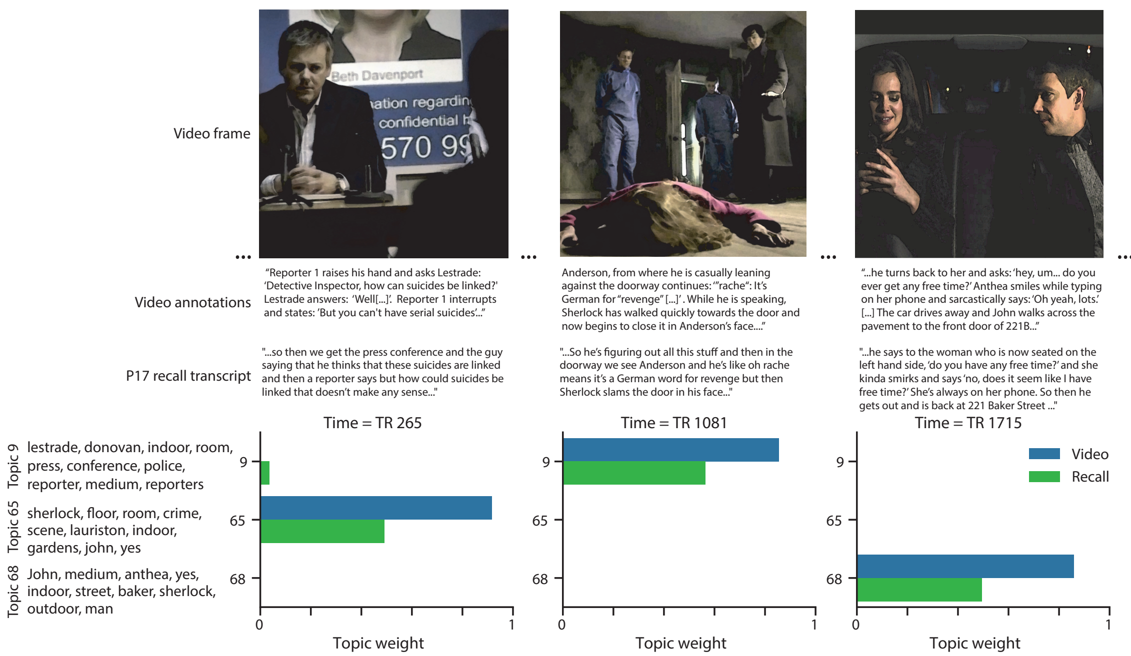

Figure 1: Methods overview. We used hand-annotated descriptions of each moment of video to fit a topic model. Three example video frames and their associated descriptions are displayed (top two rows). Participants later recalled the video (in the third row, we show example recalls of the same three scenes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

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

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

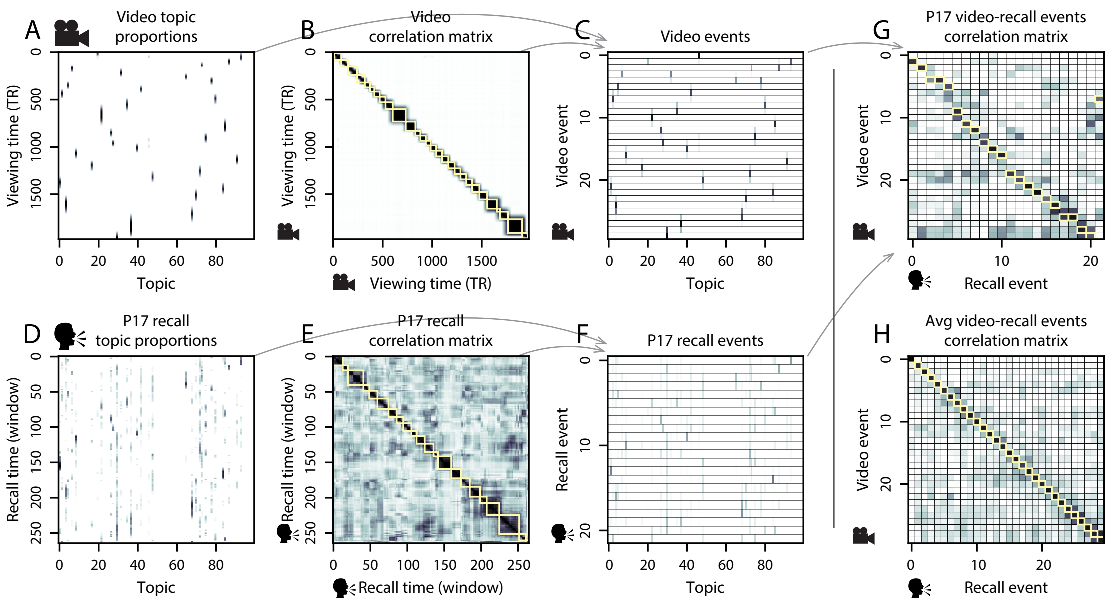

Figure 2: Modelling naturalistic stimuli and recalls. All panels: darker colors indicate greater values; range: $[0, 1]$. **A.** Topic vectors ($K = 100$) for each of the 1976 video timepoints. **B.** Timepoint-by-timepoint correlation matrix of the topic vectors displayed in Panel A. Event boundaries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

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

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

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

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

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

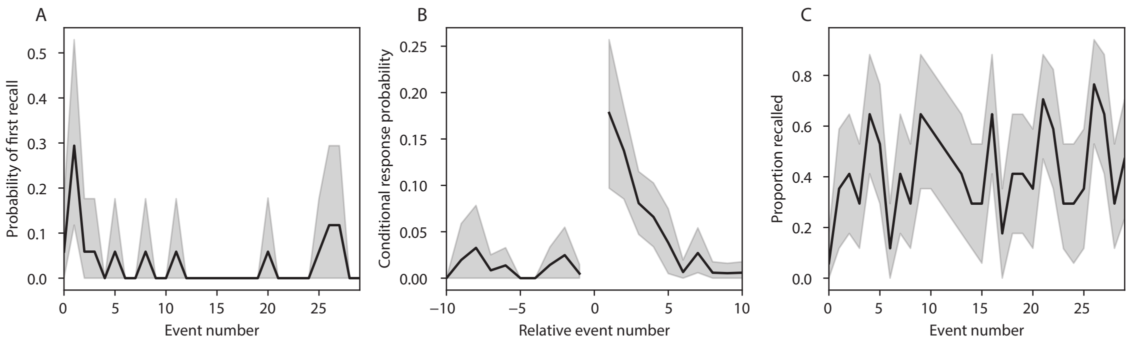

Figure 3: Naturalistic extensions of classic list-learning memory analyses. A. The probability of first recall as a function of the serial position of the event in the video. **B**. The probability of recalling each event, conditioned on having most recently recalled the event *lag* events away in the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

and that participants with higher temporal clustering scores tended to perform better according to both Chen et al. (2017)'s hand-annotated memory scores (Pearson's $r(15) = 0.62$, $p = 0.008$) and our model's estimate (Pearson's $r(15) = 0.54$, $p = 0.024$). Semantic clustering measures the extent to which participants cluster their recall responses according to semantic similarity. We found that participants tended to recall semantically similar video events together (mean: 0.787, SEM: 0.018), and that semantic clustering score was also related to both hand-annotated (Pearson's $r(15) = 0.65$, $p = 0.004$) and model-derived (Pearson's $r(15) = 0.63$, $p = 0.007$) memory performance.

Statistical models of memory studies often treat recall success as binary (i.e., an item either was or was not recalled), or occasionally categorical (e.g., to distinguish familiarity from recollection; Yonelinas et al., 2002). Such approaches are tenable in classical list-learning or recognition memory paradigms, as the presented stimuli tend to be very simple (e.g., a sequence of individual words or items). However, the feature-rich content of a naturalistic experiences may later be described with many, highly variable levels of success. Our framework produces a content-based model of individual stimulus and recall events, allowing for direct quantitative comparison between all stimulus and recall events, as well as between the recall events themselves. Leveraging these content-based models of the stimulus/recall events, we developed two novel, *continuous* metrics for quantifying naturalistic memory representations: *precision* and *distinctiveness*. We define precision as the "completeness" of recall, or how fully the presented content was recapitulated in memory. Under our framework, we quantify this for a given recall event as the correlation between the topic proportions of the recall event and the maximally correlated video event (Fig. 4). A second novel metric we introduce here is *distinctiveness*, which we define as the "specificity" of recall, or how unique the description of a given section of content was, compared to descriptions for other sections of content. We quantify this for each recall event as 1 minus the average correlation between the given recall event and all other recall events not matched to the same video event. In addition to individual events, one may also use these metrics to describe each participant's overall performance (i.e., by averaging across a participant's event-wise precision or distinctiveness scores). Participants whose recall events are more veridical descriptions of what happened in the video event will presumably have higher precision scores. We find that, across participants, a higher precision score is correlated to both hand-annotated memory performance (Pearson's $r(15) = 0.56, p = 0.021$) and the number of recall events estimated by our model (Pearson's $r(15) = 0.85, p < 0.001$). We also hypothesized that participants who recounted events in a more distinctive way would display better overall memory. We find that this distinctiveness score is related to our model's estimated number of recalled events (Pearson's $r(15) = 0.53, p = 0.028$), and while we do not find distinctiveness to be related to hand-annotated memory performance (Pearson's $r(15) = 0.28, p = 0.275$), this is not entirely surprising given how the hand-annotated memory scores were computed (see *Methods*).

Further intuition for the behaviors captured by these two metrics may be gained by directly examining the content of the video and recalls our framework models. In Figure 5, we contrast recalls for the same video event from two participants: one with a high precision score (P17), the other with a low precision score (P6). From the HMM-identified event boundaries, we recovered the set of annotations describing the content of an example video event (event 22; Fig. 5B), and divided them into different color-coded sections for each action or feature described. We then similarly recovered the set of sentences comprising the corresponding recall event for each of the two example participants. Because the recall sliding windows overlap heavily, and each recall

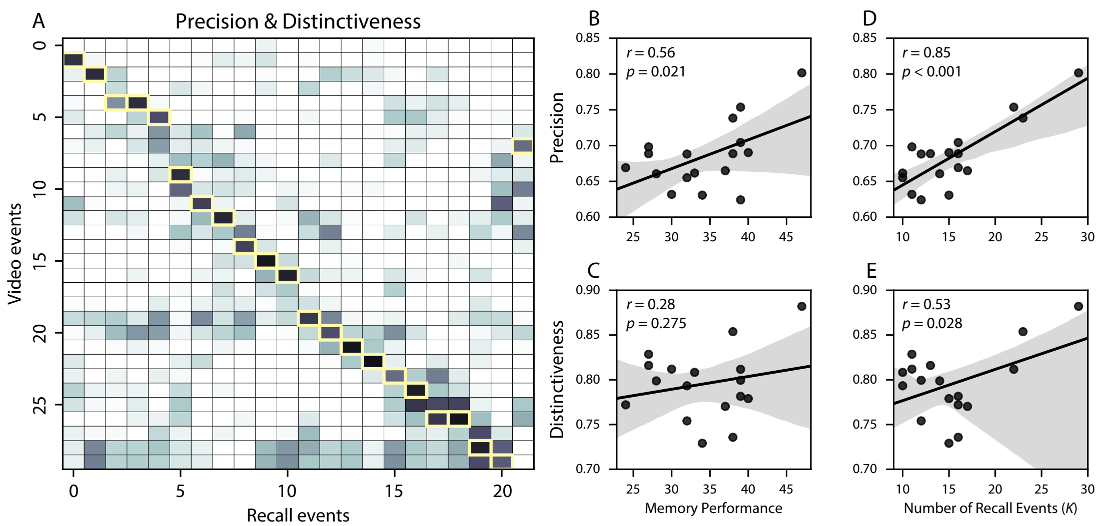

Figure 4: Novel content-based metrics of naturalistic memory: precision and distinctiveness. A. The video-recall correlation matrix for a representative participant (17). The yellow boxes highlight the maximum correlation in each column. The example participant's overall precision score was computed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

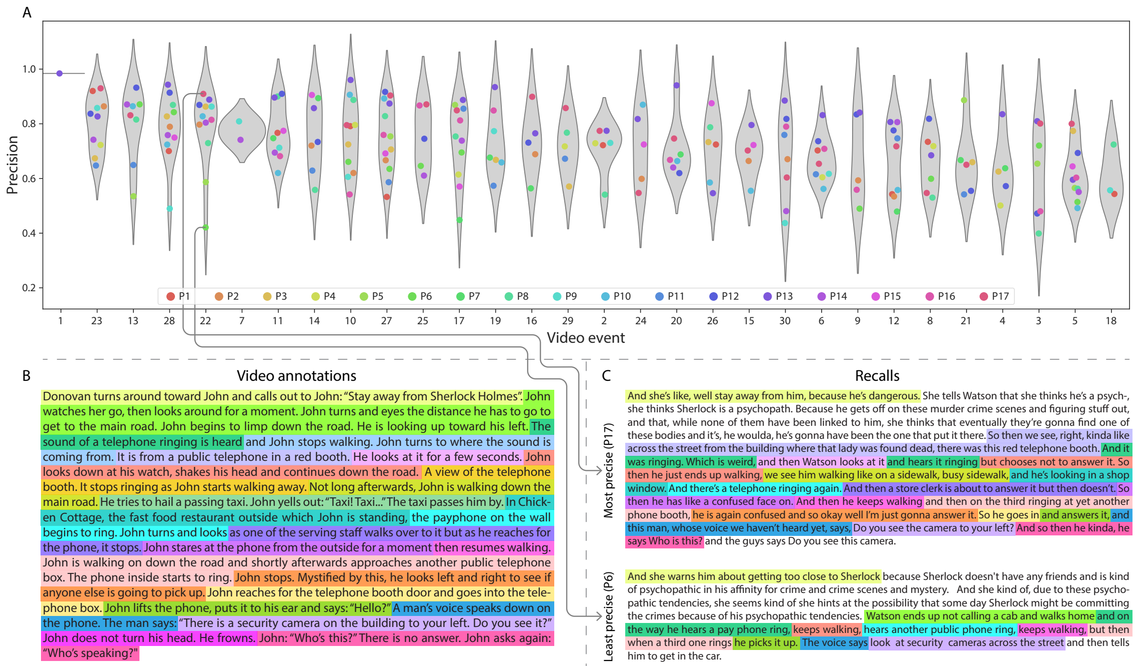

Figure 5: Precision metric reflects completeness of recall. A. Recall precision distributions over participants, for each video event. Grey violin plots display kernel density estimates for the distribution of recall precision scores for a single video event. Colored dots within the violin plots represent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

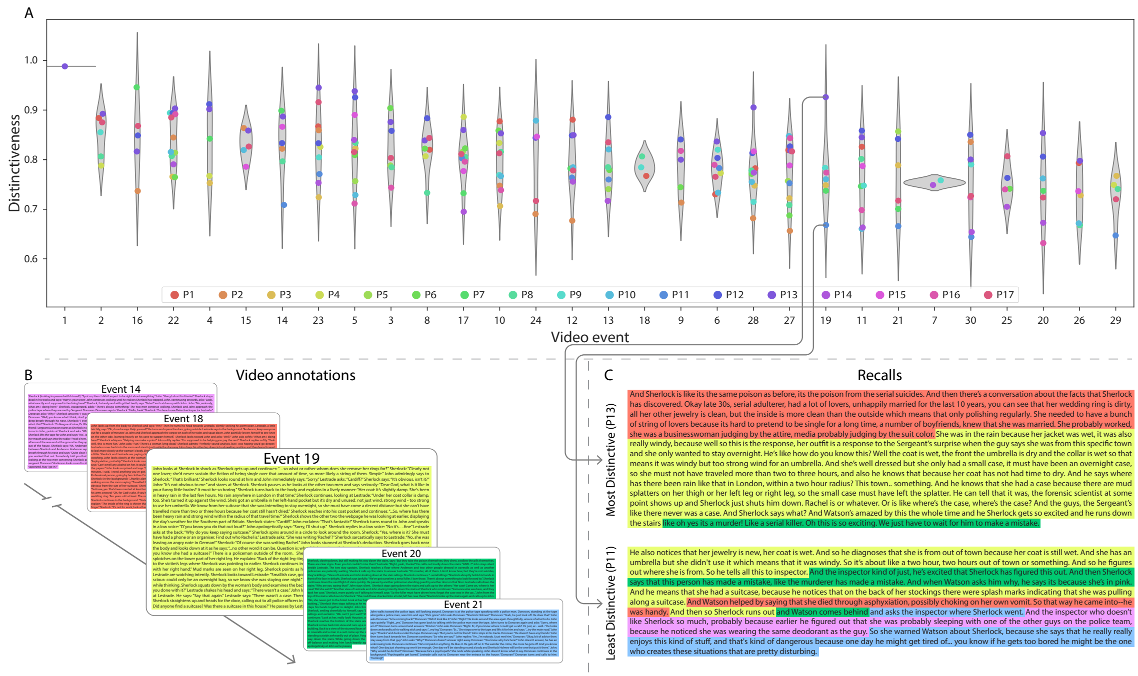

Figure 6: Distinctiveness metric reflects specificity of recall. A. Recall distinctiveness distributions over participants, for each video event. Kernel density estimates for each video event's distribution of recall distinctiveness scores, analogous to Fig. 5A. **B.** The set of "Narrative Details" video annotations (generated by Chen et al., 2017) for scenes comprising an example video event (19) identified by the HMM. **C.** The sentences comprising the most distinctive (P13) and least distinctive (P11) participants' recalls of video event 19. Descriptions of recalled content specific to video event 19 (as opposed to that combining surrounding events) are highlighted.

event spans multiple recall timepoints (i.e., windows), we have stripped any sentences from the beginning and end that describe earlier or later video events for the sake of readability. In other words, Fig. 5C shows a subset of the full recall event text, comprising sentences between the first and last descriptions of content from the example video event. We then colored all words describing actions and features coded in panel B by their corresponding color. Visual comparison of the transcripts reveals that the most precise participant's recall captures a greater amount of the video event's content, and with far more detail.

Figure 6 similarly contrasts two participants' recall for an example video event (event 19) to illustrate the tangible differences between high and low distinctiveness scores. Here, we have extracted the full text from the most

The prior analyses leverage the correspondence between the 100-dimensional topic proportion matrices for the video and participants' recalls to characterize recall. However, it is difficult to gain deep insights into that content solely by examining the topic proportion matrices (e.g., Figs. 2A, D) or the corresponding correlation matrices (Figs. 2B, E, S4). To visualize the time-varying high-dimensional content in a more intuitive way (Heusser et al., 2018b) we projected the topic proportions matrices onto a two-dimensional space using Uniform Manifold Approximation and Projection (UMAP; McInnes et al., 2018). In this lower-dimensional space, each point represents a single video or recall event, and the distances between the points reflect the distances between the events' associated topic vectors (Fig. 7). In other words, events that are near to each other in this space are more semantically similar.

Visual inspection of the video and recall topic trajectories reveals a striking pattern. First, the topic trajectory of the video (which reflects its dynamic content; Fig. 7A) is captured nearly perfectly by the averaged topic trajectories of participants' recalls (Fig. 7B). To assess the consistency of these recall trajectories across participants, we asked: given that a participant's recall trajectory had entered a particular location in topic space, could the position of their *next* recalled event be predicted reliably? For each location in topic space, we computed the set of line segments connecting successively recalled events (across all participants) that intersected that location (see *Methods* for additional details). We then computed (for each location) the distribution of angles

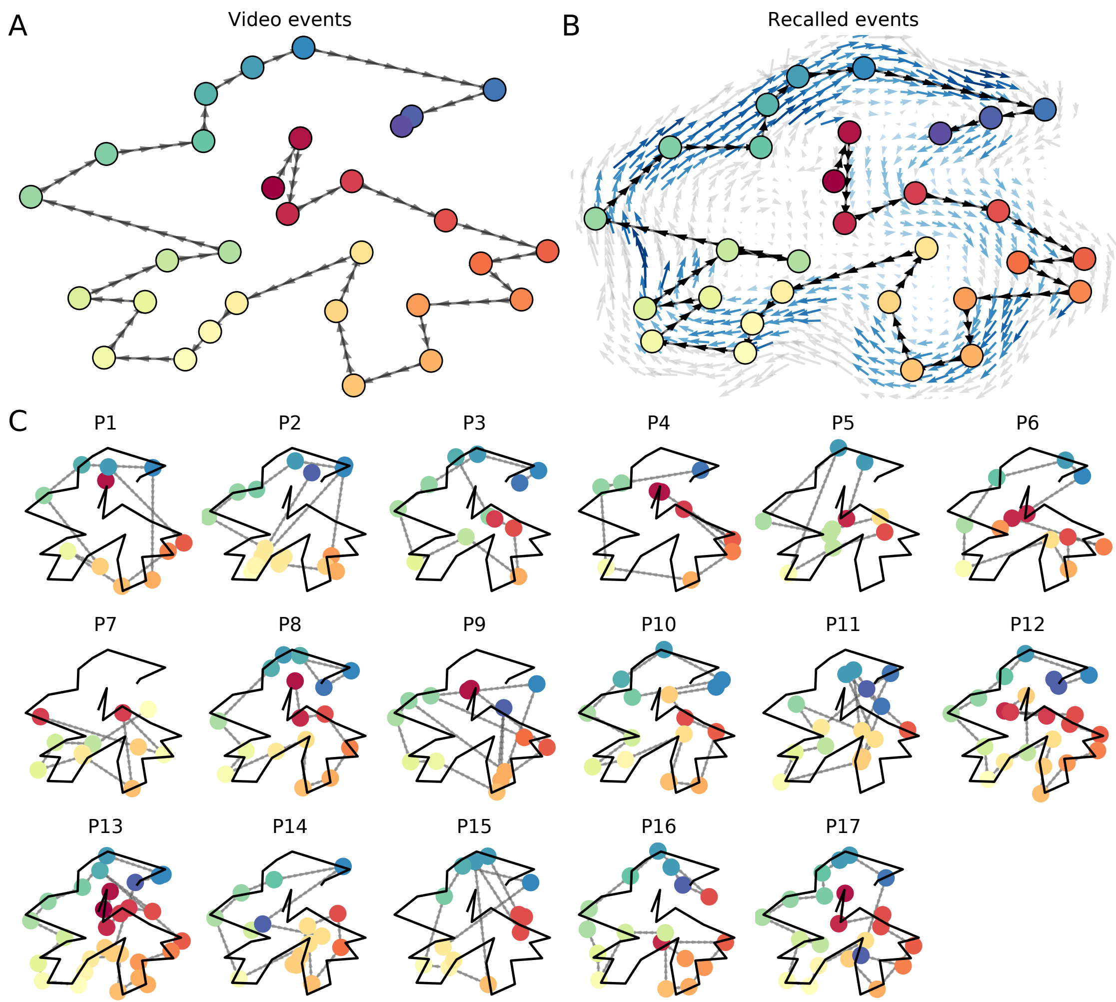

Figure 7: Trajectories through topic space capture the dynamic content of the video and recalls. All panels: the topic proportion matrices have been projected onto a shared two-dimensional space using UMAP. **A.** The two-dimensional topic trajectory taken by the episode of *Sherlock*. Each dot indicates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

formed by the lines defined by those line segments and a fixed reference line (the x-axis). Rayleigh tests revealed the set of locations in topic space at which these across-participant distributions exhibited reliable peaks (blue arrows in Fig. 7B reflect significant peaks at $p < 0.05$, corrected). We observed that the locations traversed by nearly the entire video trajectory exhibited such peaks. In other words, participants exhibited similar trajectories that also matched the trajectory of the original video (Fig. 7C). This is especially notable when considering the fact that the number of events participants recalled (dots in Fig. 7C) varied considerably across people, and that every participant used different words to describe what they had remembered happening in the video. Differences in the numbers of remembered events appear in participants' trajectories as differences in the sampling resolution along the trajectory. We note that this framework also provides a means of detangling classic "proportion recalled" measures (i.e., the proportion of video events referenced in participants' recalls) from participants' abilities to recapitulate the full shape of the original video (i.e., the similarity in the shape of the original video trajectory and that defined by each participant's recounting of the video).

Because our analysis framework projects the dynamic video content and participants' recalls onto a shared topic space, and because the dimensions of that space are known (i.e., each topic dimension is a set of weights over 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 (Fig. 8A). 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 computed a weighted average of the topic vectors for each video event, where the weights reflected how reliably 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 topics appear in a larger font (Fig. 8B, green box). Across the full video, content that reflected topics necessary to convey the central focus of the video (e.g., the names of the two main characters, "Sherlock" and "John", and the address of a major recurring location, "221B Baker Street") were best remembered. An analogous analysis revealed which themes were poorly remembered. Here in computing the weighted average over events' topic vectors, we weighted each event in *inverse* proportion to how well it was remembered (Fig. 8B, red box). The least well-remembered video content reflected information not necessary to conveying the video's "gist," such as the proper names of relatively minor characters (e.g., "Mike," "Molly," and "Lestrade") and locations (e.g., "St. Bartholomew's Hospital"), as well as the brief, animated clip participants viewed at the beginning of each of the two scan session (involving "singing" "cartoon" characters).

A similar result emerged from assessing the topic vectors for individual video and recall events (Fig. 8C). Here, for each of the three best- and worst-remembered video events, we have constructed two wordles: one from the original video event's topic vector (left) and a second from the average recall topic vector for that event (right). The three best-remembered events (circled in green) correspond to scenes important to the central plot-line: a mysterious figure spying on John in a phone booth; John and Sherlock discussing the murders in their apartment; and Sherlock laying a trap to catch the murderer. Meanwhile, the three worst-remembered events (circled in red) reflect scenes that are non-essential to summarizing the narrative's structure: the two appearances of singing cartoon characters; Molly watching as Sherlock beats a corpse in the morgue; and Sherlock noticing evidence of Anderson's and Donovan's affair.

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 we sought to identify which brain structures were sensitive to the video's dynamic content, as characterized by its topic trajectory. Specifically, we used a searchlight procedure to identify the extent to which each cluster of voxels exhibited a timecourse of activity (as the participants watched the video) whose temporal correlation matrix matched the temporal correlation matrix of the original video's topic proportions (Fig. 2B). As shown in Figure 10A, the analysis revealed a network of regions including bilateral frontal cortex and cingulate cortex, suggesting that these regions may play a role in processing information relevant

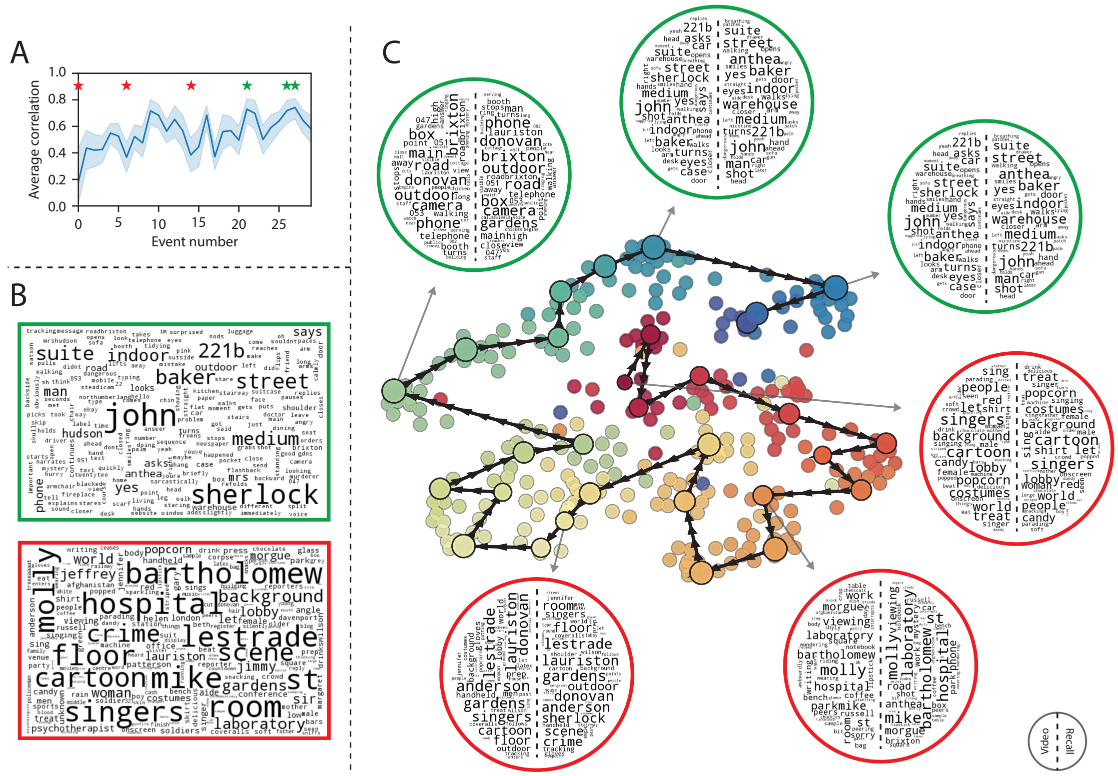

Figure 8: Transforming experience into memory. A. Average correlations (across participants) between the topic vectors from each video event and the closest-matching recall events. Error bars denote bootstrap-derived across-participant 95% confidence intervals. The stars denote the three best-remembered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

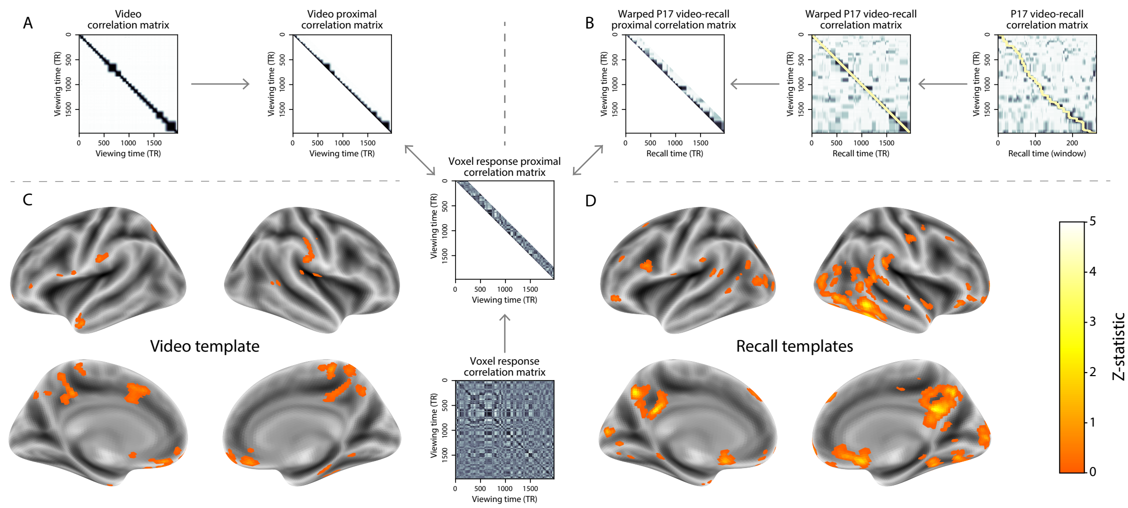

Figure 9: Brain structures that underlie the transformation of experience into memory. A. We searched for regions whose responses (as participants watched the video) matched the temporal correlation matrix of the video topic proportions. These regions are sensitive to the narrative structure of the video. **B.** We searched for regions whose responses (as participants watched the video) matched the temporal correlation matrix of the topic proportions derived from each individual's later recall of video. These regions are sensitive to how the narrative structure of the video is transformed into a memory of the video. Both panels: the maps are thresholded at $p < 0.05$, corrected.

to the narrative structure of the video. In a second analysis, we sought to identify which brain structures' responses (while viewing the video) reflected how each participant would later *recall* the video. We used an analogous searchlight procedure to identify clusters of voxels whose temporal correlation matrices reflected the temporal correlation matrix of the topic proportions for each individual's recalls (Figs. 2D, S4). As shown in Figure 10B, the analysis revealed a network of regions including the ventromedial prefrontal cortex (vmPFC), anterior cingulate cortex (ACC), and right medial temporal lobe (rMTL), suggesting that these regions may play a role in transforming each individual's experience into memory. In identifying regions whose responses to ongoing experiences reflect how those experiences will be remembered later, this latter analysis extends classic *subsequent memory analyses* (e.g., Paller and Wagner, 2002) to domain of naturalistic stimuli.

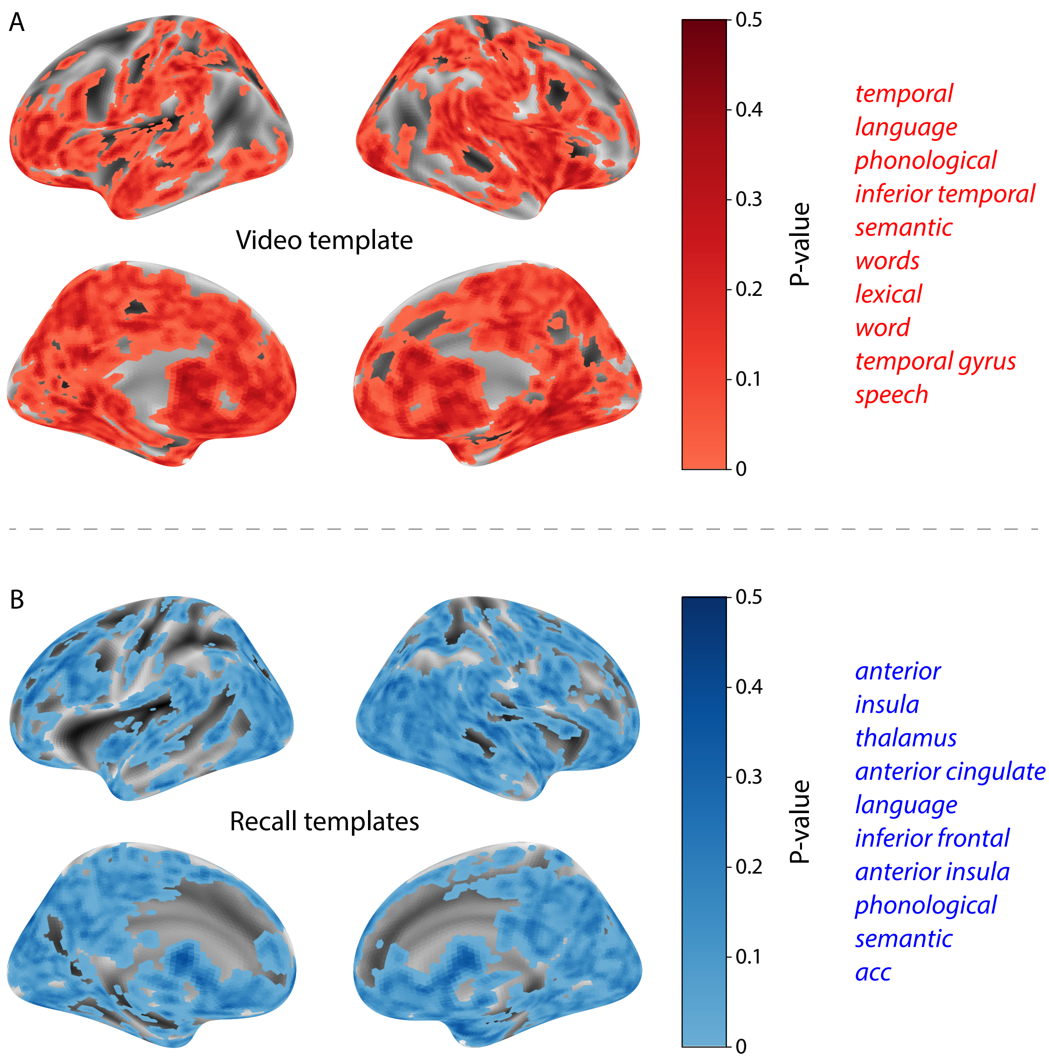

Figure 10: INSERT CAPTION HERE

Discussion

Our work casts remembering as reproducing (behaviorally and neurally) the topic trajectory, or shape, of an experience. This view draws inspiration from prior work aimed at elucidating the neural and behavioral underpinnings of how we process dynamic naturalistic experiences and remember them later. One approach to identifying neural responses to naturalistic stimuli (including experiences) entails building a model of the stimulus and searching for brain regions whose responses are consistent with the model. In prior work, a series of studies from Uri Hasson's group (Lerner et al., 2011; Simony et al., 2016; Chen et al., 2017; Baldassano et al., 2017; Zadbood et al., 2017) have extended this approach with a clever twist: rather than building an explicit stimulus model, these studies instead search for brain responses (while experiencing the stimulus) that are reliably similar across individuals. So called *inter-subject correlation* (ISC) and *inter-subject functional connectivity* (ISFC) analyses effectively treat other people's brain responses to the stimulus as a "model" of how its features change over time. By contrast, in our present work we used topic models and HMMs to construct an explicit stimulus model (i.e., the topic trajectory of the video). When we searched for brain structures whose responses are consistent with the video's topic trajectory, we identified a network of structures that overlapped strongly with the "long temporal receptive window" network reported by the Hasson group (e.g., compare our Fig. 10A with the map of long temporal receptive window voxels in Lerner et al., 2011). This provides support for the notion that part of the long temporal receptive window network may be maintaining an explicit model of the stimulus dynamics. When we performed a similar analysis after swapping out the video's topic trajectory with the recall topic trajectories of each individual participant, this allowed us to identify brain regions whose responses (as the participants viewed the video) reflected how the video trajectory would be transformed in memory (as reflected by the recall topic trajectories). The analysis revealed that the rMTL and vmPFC may play a role in this person-specific transformation from experience into memory. The role of the MTL in episodic memory encoding has been well-reported (e.g., Paller and Wagner, 2002; Davachi et al., 2003; Ranganath et al., 2004; Davachi, 2006; Wiltgen and Silva, 2007; Diana et al., 2007; van Kesteren et al., 2013). Prior work has also implicated the medial prefrontal cortex in representing "schema" knowledge (i.e., general knowledge about the format of an ongoing experience given prior similar experiences; van Kesteren et al., 2012, 2013; Schlichting and Preston, 2015; Gilboa and Marlatte, 2017; Spalding et al., 2018). Integrating across our study and this prior work, one interpretation is that the person-specific transformations mediated (or represented) by the rMTL and vmPFC may reflect schema knowledge being leveraged, formed, or updated, incorporating ongoing experience into previously acquired knowledge.

In extending classical free recall analyses to our naturalistic memory framework, we recovered two patterns of recall dynamics central to list-learning studies: a high probability of initiating recall with the first video event (Fig. 3A) and a strong bias toward transitioning from recalling a given event to recalling the event immediately following it (Fig. 3B). However, equally noteworthy are the typical free recall results not recovered in these analyses, as each highlights a fundamental difference between list-learning studies and naturalistic memory paradigms like the one employed in the present study. The most noticeable departure from hallmark free recall dynamics in these findings is the apparent lack of a serial position effect in Figure 3C, which instead shows greater and lesser recall probabilities for events distributed across the video stimulus. Stimuli in free recall experiments most often comprise lists of simple, common words, presented to participants in a random order. (In fact, numerous word pools have been developed based on these criteria; e.g., Friendly et al., 1982). These stimulus qualities enable two assumptions that are central to word list analyses, but frequently do not hold for real-world experiences. First, researchers conducting free recall studies may assume that the content at each presentation index is essentially equal, and does not bear qualities that would cause participants to remember it more or less successfully than others. Such is rarely the case with real-world experiences or experiments meant to approximate them, and the effects of both intrinsic and observer-dependent factors on stimulus memorability are well established (for review see Chun and Turk-Browne, 2007; Bylinskii et al., 2015; Tyng et al., 2017). Second, the random ordering of list items ensures that (across participants, on average) there is no relationship between the thematic similarity of individual stimuli and their presentation positions—in other words, two semantically related words are no more likely to be presented next to each other than at opposite ends of the list. In most cases, the exact opposite is true of real-world episodes. Our internal thoughts, our actions, and the physical state of the world around us all tend to follow a direct, causal progression. As a result, each moment of our experience tends to be inherently more similar to surrounding moments than to those in the distant past or future. Memory literature has termed this strong temporal autocorrelation "context," and in various media that depict real-world events (e.g., movies and written stories), we recognize it as a *narrative structure*. While a random word list (by definition) has no such structure, the logical progression between ideas and actions in a naturalistic stimulus prompts the rememberer to recount presented events in order, starting with the beginning. This tendency is reflected in our findings' second departure from typical free recall dynamics: a lack of increased probability of first recall for end-of-sequence events (Fig. 3A).

Thus, analyses such as those in Figure 3 that address only the temporal dynamics of free recall paint an incomplete picture of memory for naturalistic episodes. While useful for studying presentation order-dependent recall dynamics, they neglect to consider the stimuli's content (or, for example, that content's potential interrelatedness). However, sensitivity to stimulus and recall content introduces a new challenge: distinguishing between levels of recall quality for a stimulus (i.e., an event) that is considered to have been "remembered." When modeling memory experiments, often times events (or items) and their later memories are treated as binary and independent events (e.g., a given list item was simply either remembered or not remembered). Various models of memory (e.g., Yonelinas, 2002) attempt to improve upon this by including confidence ratings, rendering this binary judgement instead categorical. Our novel framework allows one to assess memory performance in a more continuous way (*precision*), as well as analyze the correlational structure of each encoding event to each memory event (*distinctiveness*). Further and importantly, these two novel metrics we introduce here arise from comparisons of the actual content of the experience/memories, which is not typically modeled. Leveraging this, we find that the successful memory performance is related to 1) the precision with which the participant recounts each event and 2) the distinctiveness of each recall event (relative to the other recalled events). The first finding suggests that the information retained for *any individual event* may predict the overall amount of information retained by the participant. The second finding suggests that the ability to distinguish between temporally or semantically similar content is also related to the quantity of content recovered. Intriguingly, prior studies show that pattern separation, or the ability to discriminate between similar experiences, is impaired in many cognitive disorders as well as natural aging (Stark et al., 2010; Yassa et al., 2011; Yassa and Stark, 2011). Future work might explore whether and how these metrics compare between cognitively impoverished groups and healthy controls.

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

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

Methods

Experimental design and data collection

Data were collected by Chen et al. (2017). In brief, participants (n = 22) viewed the first 48 minutes of "A Study in Pink", the first episode of the BBC television series *Sherlock*, while fMRI volumes were collected (TR = 1500 ms). Participants were pre-screened to ensure they had never seen any episode of the show before. The stimulus was divided into a 23 min (946 TR) and a 25 min (1030 TR) segment to mitigate technical issues related to the scanner. After finishing the clip, participants were instructed to (quoting from Chen et al., 2017) "describe what they recalled of the [episode] in as much detail as they could, to try to recount events in the original order they were viewed in, and to speak for at least 10 minutes if possible but that longer was better. They were told that completeness and detail were more important than temporal order, and that if at any point they realized they had missed something, to return to it. Participants were then allowed to speak for as long as they wished, and verbally indicated when they were finished (e.g., 'I'm 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 and scanning parameters, see Chen et al. (2017). The experimental protocol was approved by Princeton University's Institutional Review Board.

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

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

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

Data and code availability

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

Statistics

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

Modeling the dynamic content of the video and recall transcripts

Topic modeling

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

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

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

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

Parsing topic trajectories into events using Hidden Markov Models

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

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

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

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

Naturalistic extensions of classic list-learning analyses

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

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

As described below, we next considered a number of 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 (Welch and Burnett, 1924; Postman and Phillips, 1965; Atkinson and Shiffrin, 1968) 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-video-events (30) matrix of zeros. Then for each participant, we found the index of the video event that was recalled first (i.e., the video 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 video (Fig. 3A).

Lag conditional probability curve (lag-CRP). The lag-CRP curve (Kahana, 1996) reflects the probability of recalling a given event after the just-recalled event, as a function of their relative positions (or *lag*). In other words, a lag of 1 indicates that a recalled event came immediately after the previously recalled event in the video, and a lag of -3 indicates that a recalled event came 3 events before the previously recalled event. 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; 61 lags total) matrix. We averaged over the rows of this matrix to obtain a group-averaged lag-CRP curve (Fig. 3B).

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

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

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

Novel naturalistic memory metrics

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

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

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

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

Visualizing the video and recall topic trajectories

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

Estimating the consistency of flow through topic space across participants

In Figure 7B, we present an analysis aimed at characterizing locations in topic space that different participants move through in a consistent way (via their recall topic trajectories). The two-dimensional topic space used in our visualizations (Fig. 7) comprised a 60 x 60 (arbitrary units) square. We tiled this space with a 50 x 50 grid of evenly spaced vertices, and defined a circular area centered on each vertex whose radius was two times the distance between adjacent vertices (i.e., 2.4 units). For each vertex, we examined the set of line segments formed by connecting each pair successively recalled events, across all participants, that passed through this circle. We computed the distribution of angles formed by those segments and the x-axis, and used a Rayleigh test to determine whether the distribution of angles was reliably "peaked" (i.e., consistent across all transitions that passed through that local portion of topic space). To create Figure 7B we drew an arrow originating from each grid vertex, pointing in the direction of the average angle formed by line segments that passed within its circular radius. We set the arrow lengths to be inversely proportional to the p-values of the Rayleigh tests at each vertex. Specifically, for each vertex we converted all of the angles of segments that passed within 2.4 units to unit vectors, and we set the arrow lengths at each vertex proportional to the length of the (circular) mean vector. We also indicated any significant results ($p < 0.05$, corrected using the Benjamani-Hochberg procedure) by coloring the arrows in blue (darker blue denotes a lower p-value, i.e., a longer mean vector); all tests with $p \geq 0.05$ are displayed in gray and given a lower opacity value.

Searchlight fMRI analyses

In Figure 10, we present two analyses aimed at identifying brain regions whose responses (as participants viewed the video) exhibited a particular temporal structure. We developed a searchlight analysis wherein we constructed a cube centered on each voxel (radius: 5 voxels) and for each of these cubes, computed the temporal correlation matrix of the voxel responses during video viewing. Specifically, for each of the 1976 volumes collected during video viewing, we correlated the activity patterns in the given cube with the activity patterns (in the same cube) collected during every other timepoint. This yielded a 1976 by 1976 correlation matrix for each cube.

Next, we constructed a series of "template" matrices: the first reflecting the timecourse of video's topic trajectory, and the others reflecting that of each participant's recall topic trajectory. To construct the video template, we computed the correlations between the topic proportions estimated for every pair of TRs (prior to segmenting the trajectory into discrete events; i.e., the correlation matrix shown in Figs. 2B and 10A). We constructed similar temporal correlation matrices for each participant's recall topic trajectory (Figs. 2D, S4). However, to correct for length differences and potential non-linear transformations between viewing time and recall time, we first used dynamic time warping (Berndt and Clifford, 1994) to temporally align participants' recall topic trajectories with the video topic trajectory. An example correlation matrix before and after warping is shown in Fig. 10B. This yielded a 1976 by 1976 correlation matrix for the video template and for each participant's recall template.

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

We further sought to ensure that our analysis identified regions where the activations' temporal structure specifically reflected that of the video, rather than regions whose activity was simply autocorrelated at a width similar to the video template's diagonal. To achieve this, we used a phase shift-based permutation procedure, wherein we circularly shifted the video's topic trajectory by a random number of timepoints, computed the resulting "null" video template, and re-ran the searchlight analysis, in full. (For each of the 100 permutations, the same random shift was used for all participants). We z-scored the observed (unshifted) result at each voxel against the distribution of permutation-derived "null" results, and estimated a p-value by computing the proportion of shifted results that yielded larger values. To create the map in Figure 10A, we thresholded out any voxels whose similarity to the unshifted video's structure fell below the 95^{th} percentile of the permutation-derived similarity results.

We used an analogous procedure to identify which voxels' responses reflected the recall templates. For each participant, we correlated the upper triangle of the correlation matrix for each cube of voxels with their (time warped) recall correlation matrix. As in the video template analysis this yielded a voxelwise map of correlation coefficients per participant. However, whereas the video analysis compared every participant's responses to the same template, here the recall templates were unique for each participant. As in the analysis described above, we t-scored the (Fisher z-transformed) voxelwise correlations, and used the same permutation procedure we developed for the video responses to ensure specificity to the recall timeseries and assign significance values. To create the map in Figure 10B we again thresholded out any voxels whose correspondence values fell below the 95$^{\text{th}}$ percentile of the permutation-derived null distribution.

References

Atkinson, R. C. and Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Spence, K. W. and Spence, J. T., editors, *The psychology of learning and motivation*, volume 2, pages 89–105. Academic Press, New York.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., and Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, 95(3):709–721.

Baldassano, C., Hasson, U., and Norman, K. A. (2018). Representation of real-world event schemas during narrative perception. *Journal of Neuroscience*, 38(45):9689–9699.

Berndt, D. J. and Clifford, J. (1994). Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370.

Blei, D. M. and Lafferty, J. D. (2006). Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 113–120, New York, NY, US. ACM.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993 – 1022.

Brunec, I. K., Moscovitch, M. M., and Barense, M. D. (2018). Boundaries shape cognitive representations of spaces and events. *Trends in Cognitive Sciences*, 22(7):637–650.

Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A., and Oliva, A. (2015). Intrinsic and extrinsic effects on image memorability. *Vision Research*, 116:165–178.

Capota, M., Turek, J., Chen, P.-H., Zhu, X., Manning, J. R., Sundaram, N., Keller, B., Wang, Y., and Shin, Y. S. (2017). Brain imaging analysis kit.

Cer, D., Yang, Y., Kong, S. Y., Hua, N., Limtiaco, N., John, R. S., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Sung, Y.-H., Strope, B., and Kurzweil, R. (2018). Universal sentence encoder. *arXiv*, 1803.11175.

Chen, J., Leong, Y. C., Honey, C. J., Yong, C. H., Norman, K. A., and Hasson, U. (2017). Shared memories reveal shared structure in neural activity across individuals. *Nature Neuroscience*, 20(1):115.

Chun, M. and Turk-Browne, N. (2007). Interactions between attention and memory. *Current opinion in neurobiology*, 17(2):177–184.

Clewett, D. and Davachi, L. (2017). The ebb and flow of experience determines the temporal structure of memory. *Curr Opin Behav Sci*, 17:186–193.

Davachi, L. (2006). Item, context and relational episodic encoding in humans. *Current Opinion in Neurobiology*, 16(6):693—700.

Davachi, L., Mitchell, J. P., and Wagner, A. D. (2003). Multiple routes to memory: distinct medial temporal lobe processes build item and source memories. *Proceedings of the National Academy of Sciences, USA*, 100(4):2157 – 2162.

Diana, R. A., Yonelinas, A. P., and Ranganath, C. (2007). Imaging recollection and familiarity in the medial temporal lobe: a three-component model. *Trends in Cognitive Sciences*, doi:10.1016/j.tics.2007.08.001.

Dobrushin, R. L. (1970). Prescribing a system of random variables by conditional distributions. *Theory of Probability & Its Applications*, 15(3):458–486.

DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the sequential order of events. *Journal of Experimental Psychology: General*, 142(4):1277–1286.

Ezzyat, Y. and Davachi, L. (2011). What constitutes an episode in episodic memory? *Psychological Science*, 22(2):243–252.

Friendly, M., Franklin, P. E., Hoffman, D., and Rubin, D. C. (1982). The Toronto Word Pool: Norms for imagery, concreteness, orthographic variables, and grammatical usage for 1,080 words. *Behavior Research Methods and Instrumentation*, 14:375–399.

Gilboa, A. and Marlatte, H. (2017). Neurobiology of schemas and schema-mediated memory. *Trends Cogn Sci*, 21(8):618–631.

Hasson, U., Chen, J., and Honey, C. J. (2015). Hierarchical process memory: memory as an integral component of information processing. *Trends in Cognitive Science*, 19(6):304–315.

Hasson, U., Yang, E., Vallines, I., Heeger, D. J., and Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience*, 28(10):2539–2550.

Heusser, A. C., Ezzyat, Y., Shiff, I., and Davachi, L. (2018a). Perceptual boundaries cause mnemonic trade-offs between local boundary processing and across-trial associative binding. *Journal of Experimental Psychology Learning, Memory, and Cognition*, 44(7):1075–1090.

Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K., and Manning, J. R. (2017). Quail: a Python toolbox for analyzing and plotting free recall data. *The Journal of Open Source Software*, 10.21105/joss.00424.

Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018b). HyperTools: a Python toolbox for gaining geometric insights into high-dimensional data. *Journal of Machine Learning Research*, 18(152):1–6.

Howard, M. W. and Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46:269–299.

Howard, M. W., MacDonald, C. J., Tiganj, Z., Shankar, K. H., Du, Q., Hasselmo, M. E., and H., E. (2014). A unified mathematical framework for coding time, space, and sequences in the medial temporal lobe. *Journal of Neuroscience*, 34(13):4692–4707.

Howard, M. W., Viskontas, I. V., Shankar, K. H., and Fried, I. (2012). Ensembles of human MTL neurons "jump back in time" in response to a repeated stimulus. *Hippocampus*, 22:1833–1847.

Huk, A., Bonnen, K., and He, B. J. (2018). Beyond trial-based paradigms: continuous behavior, ongoing neural activity, and naturalistic stimuli. *Journal of Neuroscience*, 10.1523/JNEUROSCI.1920-17.2018.

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition*, 24:103–109.

Koriat, A. and Goldsmith, M. (1994). Memory in naturalistic and laboratory contexts: distinguishing accuracy-oriented and quantity-oriented approaches to memory assessment. *Journal of Experimental Psychology: General*, 123(3):297–315.

Kriegeskorte, N., Mur, M., and Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:1 – 28.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Lerner, Y., Honey, C. J., Silbert, L. J., and Hasson, U. (2011). Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *Journal of Neuroscience*, 31(8):2906–2915.

Manning, J. R. (2019). Episodic memory: mental time travel or a quantum 'memory wave' function? *PsyArXiv*, doi:10.31234/osf.io/6zjwb.

Manning, J. R., Norman, K. A., and Kahana, M. J. (2015). The role of context in episodic memory. In Gazzaniga, M., editor, *The Cognitive Neurosciences, Fifth edition*, pages 557–566. MIT Press.

Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B., and Kahana, M. J. (2011). Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proceedings of the National Academy of Sciences, USA*, 108(31):12893–12897.

McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 1802(03426).

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*, 1301.3781.

Mueller, A., Fillion-Robin, J.-C., Boidol, R., Tian, F., Nechifor, P., yoonsubKim, Peter, Rampin, R., Corvellec, M., Medina, J., Dai, Y., Petrushev, B., Langner, K. M., Hong, Alessio, Ozsvald, I., vkolmakov, Jones, T., Bailey, E., Rho, V., IgorAPM, Roy, D., May, C., foobuzz, Piyush, Seong, L. K., Goey, J. V., Smith, J. S., Gus, and Mai, F. (2018). WordCloud 1.5.0: a little word cloud generator in Python. *Zenodo*, https://zenodo.org/record/1322068#.W4tPKZNKh24.

Murdock, B. B. (1962). The serial position effect of free recall. *Journal of Experimental Psychology*, 64:482–488.

Paller, K. A. and Wagner, A. D. (2002). Observing the transformation of experience into memory. *Trends in Cognitive Sciences*, 6(2):93–102.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Polyn, S. M., Norman, K. A., and Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review*, 116(1):129–156.

Postman, L. and Phillips, L. W. (1965). Short-term temporal changes in free recall. *Quarterly Journal of Experimental Psychology*, 17:132–138.

Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286.

Radvansky, G. A. and Zacks, J. M. (2017). Event boundaries in memory and cognition. *Curr Opin Behav Sci*, 17:133–140.

Ramdas, A., Trillos, N., and Cuturi, M. (2017). On wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47.

Ranganath, C., Cohen, M. X., Dam, C., and D'Esposito, M. (2004). Inferior temporal, prefrontal, and hippocampal contributions to visual working memory maintenance and associative memory retrieval. *Journal of Neuroscience*, 24(16):3917–3925.

Ranganath, C. and Ritchey, M. (2012). Two cortical systems for memory-guided behavior. *Nature Reviews Neuroscience*, 13:713 – 726.

Schlichting, M. L. and Preston, A. R. (2015). Memory integration: neural mechanisms and implications for behavior. *Current Opinion in Behavioral Sciences*, 1:1–8.

Simony, E., Honey, C. J., Chen, J., and Hasson, U. (2016). Uncovering stimulus-locked network dynamics during narrative comprehension. *Nature Communications*, 7(12141):1–13.

Spalding, K. N., Schlichting, M. L., Zeithamova, D., Preston, A. R., Tranel, D., Duff, M. C., and Warren, D. E. (2018). Ventromedial prefrontal cortex is necessary for normal associative inference and memory integration. *The Journal of Neuroscience*, 38(15):3767–3775.

Stark, S. M., Yassa, M. A., and Stark, C. E. L. (2010). Individual differences in spatial pattern separation performance associated with healthy aging in humans. *Learning & Memory*, 17(6):284–288.

Steyvers, M., Shiffrin, R. M., and Nelson, D. L. (2004). Word association spaces for predicting semantic similarity effects in episodic memory. In Healy, A. F., editor, *Cognitive Psychology and its Applications: Festschrift in Honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*. American Psychological Association, Washington, DC.

Tompary, A. and Davachi, L. (2017). Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron*, 96(1):228–241.

Tyng, C. M., Amin, H. U., Saad, M. N. M., and S, M. A. (2017). The influences of emotion on learning and memory. *Frontiers in psychology*, 8:1454.

van Kesteren, M. T. R., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., and Fernández, G.

(2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: from congruent to incongruent. *Neuropsychologia*, 51(12):2352–2359.

van Kesteren, M. T. R., Ruiter, D. J., Fernández, G., and Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends Neurosci*, 35(4):211–9.

Waskom, M., Botvinnik, O., Okane, D., Hobson, P., David, Halchenko, Y., Lukauskas, S., Cole, J. B., Warmenhoven, J., de Ruiter, J., Hoyer, S., Vanderplas, J., Villalba, S., Kunter, G., Quintero, E., Martin, M., Miles, A., Meyer, K., Augspurger, T., Yarkoni, T., Bachant, P., Williams, M., Evans, C., Fitzgerald, C., Brian, Wehner, D., Hitz, G., Ziegler, E., Qalieh, A., and Lee, A. (2016). Seaborn: v0.7.1.

Welch, G. B. and Burnett, C. T. (1924). Is primacy a factor in association-formation. *American Journal of Psychology*, 35:396–401.

Wiltgen, B. J. and Silva, A. J. (2007). Memory for context becomes less specific with time. *Learning & Memory*, 14(4):313–317.

Yassa, M. A., Lacy, J. W., Stark, S. M., Albert, M. S., Gallagher, M., and Stark, C. E. L. (2011). Pattern separation deficits associated with increased hippocampal ca3 and dentate gyrus activity in nondemented older adults. *Hippocampus*, 21(9):968–979.

Yassa, M. A. and Stark, C. E. L. (2011). Pattern separation in the hippocampus. *Trends In Neurosciences*, 34(10):515–525.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, 46:441–517.

Yonelinas, A. P., Kroll, N. E., Quamme, J. R., Lazzara, M. M., Sauvé, M. J., Widaman, K. F., and Knight, R. T. (2002). Effects of extensive temporal lobe damage or mild hypoxia on recollection and familiarity. *Nature Neuroscience*, 5(11):1236–41.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., and Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological Bulletin*, 133:273–293.

Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). How we transmit memories to other brains: Constructing shared neural representations via communication. *Cereb Cortex*, 27(10):4988–5000.

Zwaan, R. A. and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123(2):162 – 185.

Supporting information

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

Acknowledgements

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

Author contributions

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

Author information

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