How is experience transformed into memory?

Andrew C. Heusser, Paxton C. Fitzpatrick, and Jeremy R. Manning
Department of Psychological and Brain Sciences
Dartmouth College, Hanover, NH 03755, USA
Corresponding author: jeremy.r.manning@dartmouth.edu

September 2, 2018

Abstract

How 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 they unfold over time. New insights emerge when we apply this approach to a naturalistic memory experiment which had participants view and recount a video. We found that the shapes of the trajectories formed by participants' recountings were all highly similar to that of the original video, but participants differed in the level of detail they remembered. We also identified a network of brain structures that are sensitive to the "shapes" of our ongoing experiences, and an overlapping network that is sensitive to how we will later remember those 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 the rest of one's experiences, and that item may be labeled as having been recalled versus forgotten. More nuanced studies might incorporate self-reported confidence measures as a proxy for memory strength, or ask participants to discriminate between "recollecting" the (contextual) details of an experience or having a general feeling of "familiarity" (Yonelinas, 2002). However, characterizing and evaluating memory in more realistic contexts (e.g., recounting a recent experience to a friend) is fundamentally different in at least three ways (also see Koriat and Goldsmith, 1994, for a review). First, real world recall is continuous, rather than discrete. Unlike in trial-based experiments, removing a (naturalistic) event from the context in which it occurs can substantially change its meaning. Second, the specific words used to describe an experience have 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, counting the number or proportion of precise recalls is often a primary metric of assessing the quality of participants' memories. Third, one might remember the *gist* or essence of an experience

but forget (or neglect to recount) particular details. Capturing the gist of what happened is typically the main "point" of recounting a memory to a listener whereas, depending on the circumstances, accurate recall of any specific detail may be irrelevant. There is no analog of the gist of an experience in most traditional memory experiments; rather we tend to assess participants' abilities to recover specific details (e.g., computing the proportion of specific stimuli they remember, which presentation positions the remembered stimuli came from, etc.).

How might one go about formally characterizing the gist of an experience, or whether that gist 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 (e.g., Lerner et al., 2011). Therefore the timecourse of 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 then 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 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). Although our experiences most often change gradually, they also occasionally change suddenly (e.g., when we walk through a doorway; Radvansky and Zacks, 2017). Prior research suggests that these sharp transitions (termed *event boundaries*) during an experience help to discretize our experiences into *events* (Radvansky and Zacks, 2017; Brunec et al., 2018; Heusser et al., 2018; 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– i.e., their gists. For example, prior work has shown that event boundaries can influence how we learn sequences of items (Heusser et al., 2018; DuBrow and Davachi, 2013), navigate (Brunec et al., 2018), and remember and understand narratives (Zwaan and Radvansky, 1998; Ezzyat and Davachi, 2011).

Here we sought to examine how the temporal dynamics of a "naturalistic" experience were reflected in participants' later memories of that experience. We analyzed an open dataset which comprised behavioral and functional Magnetic Resonance Imaging (fMRI) data collected as participants viewed and then verbally recalled 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 we use Hidden Markov Models (Rabiner, 1989; Baldassano et al., 2017) to discretize the evolving semantic content into events. In this way, we cast naturalistic experiences (and recalls of those experiences) as *topic trajectories* that describe how the experiences evolve over time. In other words, the episode's topic trajectory is a formalization of its gist. Under this framework, successful remembering entails verbally "traversing" the topic trajectory of the original episode, thereby reproducing the gist of the original episode. In addition, comparing the shapes of the topic trajectories of the original episode and of participants' retellings of the episode reveals which aspects of the episode were preserved (or lost) in the translation into memory. We also identified a network of brain structures whose responses (as participants watched the episode) reflected the gist of the episode, and a second network whose responses

reflected how participants would later recount the episode.

Results

To characterize the gists of the *Sherlock* episode participants watched and their subsequent recountings of the episode, we used a topic model (Blei et al., 2003) to discover the latent thematic content in the video. Topic models take as inputs a vocabulary of words to consider and a collection of text documents; they return as output two matrices. The first output 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 which mix topics is reflected in each document.

Chen et al. (2017) collected hand-annotated information about each of 1000 (manually identified) scenes spanning the roughly 45 minute video used in their experiment. This information included: a brief narrative description of what was happening; whether the scene took place indoors vs. outdoors; names of any characters on the screen; names of any characters who were in focus in the camera; names of characters who were speaking; the location where the scene took place; the camera angle (close up, medium, long, etc.); whether or not background music was present; 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 50-scene sliding windows, and treated each 50-scene sequence as a single "document" for the purposes 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 28 unique topics (with non-zero weights) were sufficient to describe the time-varying content of the movie (see *Methods*; Figs. 1, S1). 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 the how the properties of collections of documents change over time, our approach allows us to examine the topic dynamics within a single video. Specifically, our approach yielded (via the topic proportions matrix) a single *topic vector* for each timepoint of the episode (we set timepoints to match the acquisition times of the 1976 fMRI volumes collected as participants viewed the episode).

The topics we found were heavily character-focused (e.g., the top-weighted word in each topic was nearly always a character) and could be roughly divided into themes that were primarily Sherlock-focused; primarily John-focused (John is Sherlock's close confidant and assistant); or that involved Sherlock and John interacting (Fig. S1). Several of the topics were highly similar, which we hypothesized might allow us to distinguish between subtle 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) topics tended to be "active" in any given timepoint (Fig. 2A). Further, the dynamics of the topic activations appeared to exhibit *persistance* (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

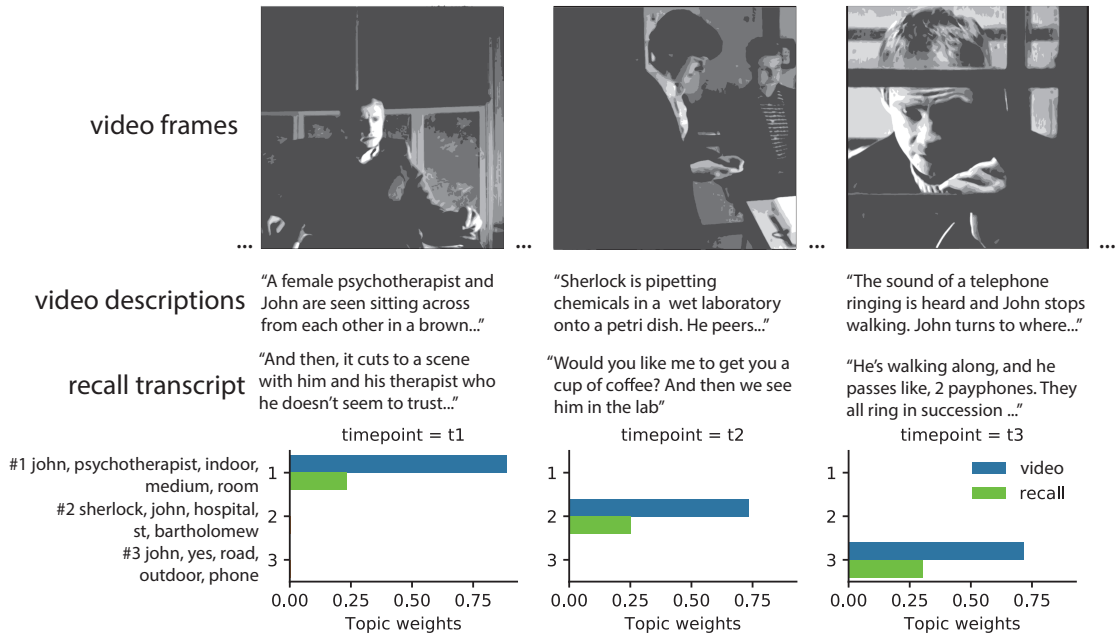

Figure 1: Methods overview. We used hand-annotated descriptions of each moment of video to fit a topic model. Three example video frames, and their associated descriptions are displayed (top two rows). Participants later recalled the video (in the third row, we show example recalls of the same three scenes from one participant). 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 (green: video annotations; blue: 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 5 highest-weighted words for each topic. Figure S1 provides a full list of the top 10 words from each of the discovered topics.

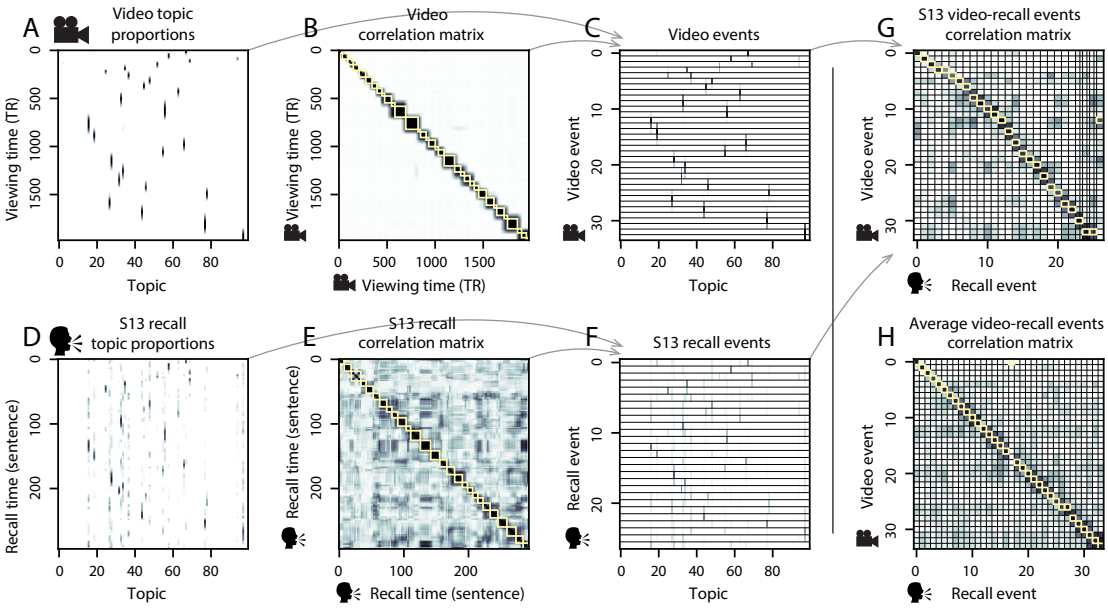

Figure 2: Modelling naturalistic stimuli and recalls. All panels: darker colors indicate greater values; range: [0, 1]. **A.** Topic vectors ($K = 100$) for each of the 1976 video timepoints. **B.** Timepoint-by-timepoint correlation matrix of the topic vectors displayed in Panel A. Event boundaries detected by the HMM are denoted in yellow (34 events detected). **C.** Average topic vectors for each of the 34 video events. **D.** Topic vectors for each of 294 sentences spoken by an example participant while recalling the video. **E.** Timepoint-by-timepoint correlation matrix of the topic vectors displayed in Panel D. Event boundaries detected by the HMM are denoted in yellow (27 events detected). **F.** Average topic vectors for each of the 27 recalled events from the example participant. **G.** Correlations between the topic vectors for every pair of video events (Panel C) and recalled events (from the example participant; Panel F). For similar plots for all participants see Figure S3. **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.

structure of the timepoint-by-timepoint correlation matrix (Fig. 2B). Following Baldassano et al. (2017), we 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). Part of our model fitting procedure required selecting an appropriate number of "events" to segment the timeseries into. We used an optimization procedure to identify the number of events that maximized within-event stability while also minimizing across-event correlations (see *Methods* for additional details). To create a stable "summary" of the video, we computed the average topic vector within each event (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 analyzed the time-varying content of the video, to obtain similar estimates for participants' recalls we treated each (overlapping) 10 sentence "window" of their transcript as a "document" and then computed the most probable mix of topics reflected in each

timepoint's sentences. This yielded, for each participant, a number-of-sentences 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 participant's recalls to events from the original video, despite that different participants may have used different language to describe the same event, and that those descriptions may not match the original annotations. This is a huge benefit of projecting the video and recalls into a shared "topic" space. An example topic proportions matrix from one participant's recalls is shown in Figure 2D.

Although the example participant's recall topic proportions matrix has some visual similarity to the video topic proportions matrix, the time-varying topic proportions for the example participant's recalls are not as sparse as for the video (e.g., compare Figs. 2A and D). Similarly, although there do appear to be periods of stability in the recall topic dynamics (e.g., most topics are active or inactive over contiguous blocks of time), the overall timecourses are not as cleanly delineated as the video topics are. To examine these patterns in detail, we computed the timepoint-by-timepoint correlation matrix for the example participant's recall topic proportions (Fig. 2E). As in the video correlation matrix (Fig. 2B), the example partipant'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. S2).

Two clear patterns emerged from this set of analyses. First, although every individual participant's recalls could be segmented into discrete events (e.g. every individual participant's recall correlation matrix exhibited clear block diagonal structure; Fig. S2), each participant appeared to have a unique *recall resolution*, reflected in the sizes of those blocks. For example, some participants' recall topic proportions segmented into just a few events (e.g. Participants 1, 4, and 15), while others' recalls segmented into many shorter duration events (e.g. Participants 12, 13, and 17). This suggests that different participants may be recalling the video with different levels of detail– e.g. some might touch on just the major plot points, whereas others might attempt to recall every minor scene. The second clear pattern present in every individual participant's recall correlation matrix is that, unlike in the video correlation matrix, there are substantial off-diagonal correlations in participant's recalls. Whereas each event in the original video (was largely) separable from the others (Fig. 2B), in transforming those separable events into memory participants appear to be integrating *across* different events, blending elements of previously recalled and not-yet-recalled events into each newly recalled event (Figs. 2D, S2; 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 using 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, S3). This yields a sequence of "presented" events from the original movie, and a sequence of (potentially differently ordered) "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 (e.g., probability of first recall; Fig. S5A; Welch and Burnett, 1924; Postman and Phillips, 1965; Atkinson and Shiffrin, 1968); how participants most often transition between recalls of the events as a function of the temporal distance between them (e.g., lag-conditional response probability; Fig. S5B; Kahana, 1996); and which events they were likely to remember overall (e.g., serial position recall analyses; Fig. S5C; Murdock, 1962). In list-learning studies, this set of three analyses may be used to gain a nearly complete view into the sequences of recalls participants made (e.g., Kahana, 2012). However, in naturalistic recall the analyses provide a wholly incomplete picture: they leave out any attempt to quantify participants' abilities to capture the *content* of what occurred in the video (i.e., their only experimental instruction!).

The dynamic content of the video and participants' recalls is quantified in the corresponding topic proportion matrices. However, it is difficult to gain deep insights into that content by directly examining the topic proportion matrices (e.g., Figs. 2A, D) or the corresponding correlation matrices (Figs. 2B, E, S2). To visualize the time-varying high-dimensional content in a more intuitive way (Heusser et al., 2018) we projected the topic proportion matrices onto a two-dimensional space using Uniform Manifold Approximation and Projection (UMAP; McInnes and Healy, 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. 3).

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. 3A) is captured nearly perfectly by the averaged topic trajectories of participants' recalls (Fig. 3B). 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 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. 3B). We observed that the locations traversed by nearly the entire the video trajectory exhibited such peaks. In other words, participants exhibited similar trajectories that also matched the trajectory of the original video (Fig. 3C). This is especially notable when considering the fact that the number of events participants recalled (dots in Fig. 3C) varied considerably across people, and that every participant used different words to describe what they had remembered happening in 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. S1), we can examine the topic trajectories to understand which specific content was remembered well (or poorly). For each video event, we can ask: for all participants who recalled that event, what was the average correlation (across participants) between the video event's topic vector and the 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. 4A). (We also examined how different comparisons between each video event's topic vector and the corresponding recall event topic vectors related to hand-annotated characterizations of memory performance; see *Supplemental Materials*). 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

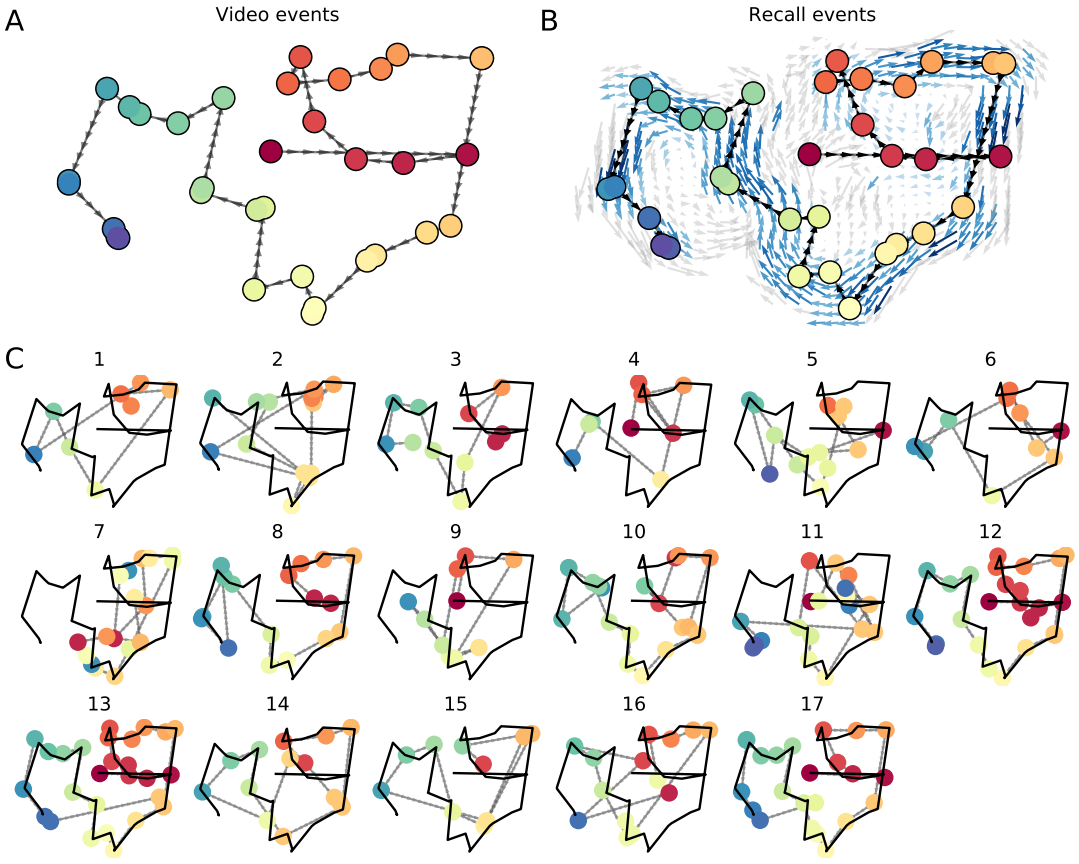

Figure 3: Trajectories through topic space capture the dynamic content of the video and recalls. All panels: the topic proportion matrices have been projected onto a shared two-dimensional space using UMAP. **A.** The two-dimensional topic trajectory taken by the episode of *Sherlock*. Each dot indicates 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 text), 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 taken by each individual participant (1–17). The video's trajectory is shown in black for reference. (Same format and coloring as Panel A.)

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. 4B, green box). Events that reflected topics weighting heavily on characters like "Sherlock" and "John" (i.e., the main characters) and locations like "221b Baker Street" (i.e., a major recurring location; the address of the flat that Sherlock and John share) 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. 4B, red box). This revealed that events with relatively minor characters such as "Mike," "Jeffrey," and "Molly," as well as less-integral plot locations (e.g., "hospital" and "office") were least well-remembered. This suggests that what is retained in memory are the major plot elements (i.e., the overall "gist" of what happened), whereas the more minor details are prone to pruning.

In addition to constructing overall summaries, assessing the video and recall topic vectors from individual recalls can provide further insights. Specifically, for any given event we can construct a wordle from that event's topic vector, and we can construct a similar wordle from the average topic vectors produced by all participants who recalled that event. We can then examine those wordles visually to gain an intuition for which aspects of the video event were recapitulated in participants' recalls of that event. Several example wordles are displayed in Figure 4C (wordles from the three best-remembered events are circled in green; wordles from the three worst-remembered events are circled in red). **[JRM NOTE: NEED SOME SORT OF "POINT" OR TAKE-AWAY FROM THIS ANALYSIS]**

The results thus far tell 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 (as the participants watched the video) whose temporal correlation matrix matched the temporal correlation matrix of the original video's topic proportion matrix (Fig. 2B). As shown in Figure 5A, the analysis revealed a network of regions including bilateral frontal and cingulate cortex, suggesting that these regions may play a role in maintaining information relevant 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. Specifically, 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. S2). As shown in Figure 5B, the analysis revealed a network of regions including the ventromedial prefrontal cortex, anterior cingulate, and right medial temporal lobe, suggesting that these regions may play a role in transforming each individual's experience into memory.

Discussion

Studying human memory is commonly distilled down to a process of matching specific moments of a past experience with specific mnemonic outcomes. In traditional trial-based free recall experiments, individual stimuli encountered during encoding are typically labeled as "remembered" or "forgotten" depending on whether the stimulus is recalled/recognized during a subsequent test.

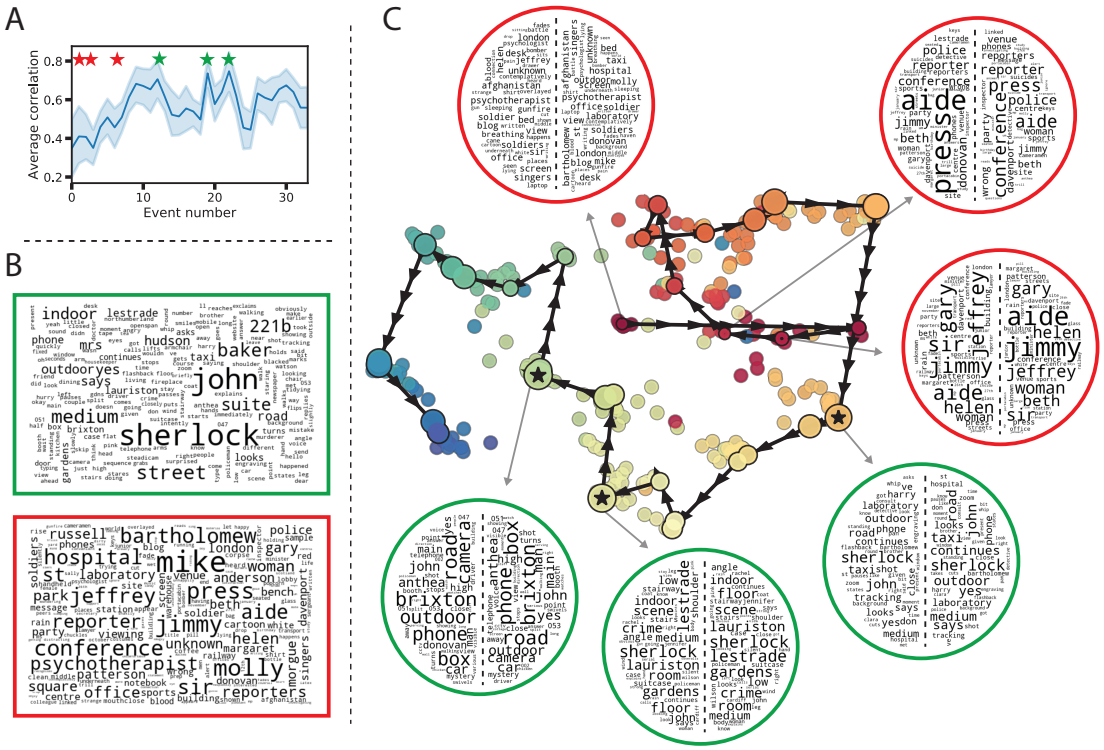

Figure 4: Transforming experience into memory. A. Average correlations (across participants) between the topic vectors from each video event and the closest-matching recall events. Error bars denote bootstrap-derived across-participant 95% confidence intervals. The stars denote the three best-remembered 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 3. The large dots denote video events and the smaller dots denote recalled events (same color scheme as Fig. 3A). 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.

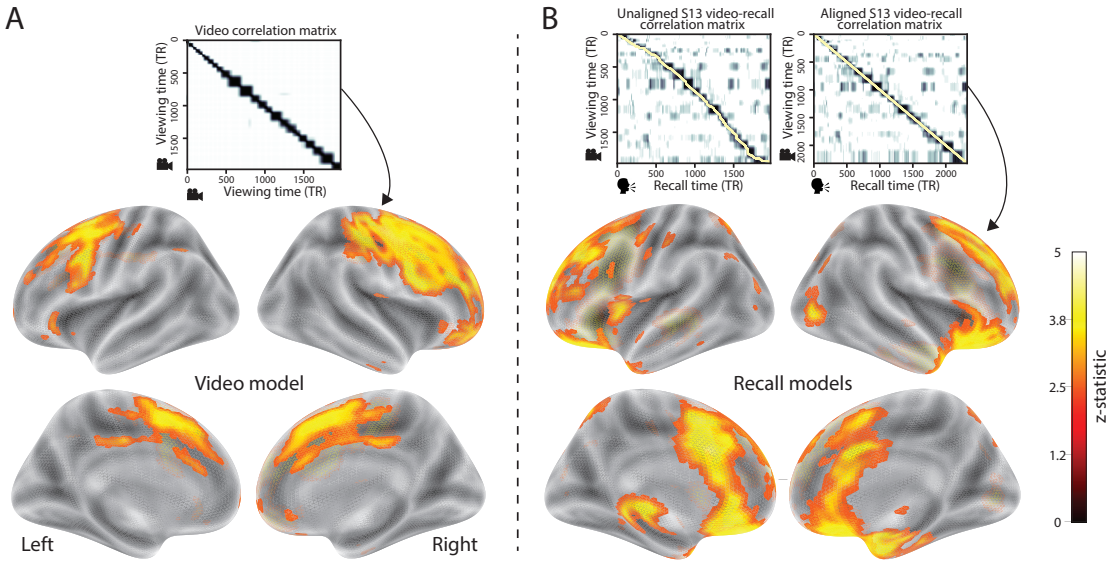

Figure 5: Brain structures that underlie the transformation of experience into memory. A. We searched for regions whose responses (as participants watched the video) matched the temporal correlation matrix of the video topic proportions. These regions are sensitive to the narrative structure of the video. **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.

While this approach has advanced our understanding of human memory immensely, it does not translate well to naturalistic experiences. For one thing, the contents of our recall at any given moment might be reflected in many prior experiences/moments. Furthermore, the particular words used to describe the experience will inevitably vary across people and even across repeated recollections within an individual. Thus, there is not a "one-to-one" mapping between naturalistic experiences and their mnemonic counterparts. For example, remembering the patterns and colors of each person's shirt in a crowd might be considered as excellent recall in a standard memory task setup. But if the rememberer failed to note that the people in the crowd were gathered for their surprise birthday party, then they would have missed the "point" of the experience.

Our topic modeling approach, whereby we consider the broad "theme" present in different moments of participants' experiences and their memories for those experiences, affords us the ability to flexibly and accurately characterize memory for naturalistic experiences. This approach allows us to quantify which moments from the past and the current recounted experience match in terms of their thematic content and critically, our ability to perform this matching does not require participants to use any specific overlapping words. Our work characterizes recollection of an experience by comparing the overall "shape" of a dynamic stimulus and a memory. We assess the quality of memory for the video participants viewed by measuring the match between the shapes of the video's trajectory and each participant's recall trajectory. By contrast, the number of recalls could be captured by the "sampling frequency" along that trajectory – but the number of recalls alone cannot tell us whether participants successfully recollected the meaning of the story by capturing the salient points of the narrative that define its main shape. In addition to providing a way to capture the shape of an experience, this method affords the ability to quantify the particular contents of memory. Whereas traditional approaches abstract over the content (e.g., percent correct), Our approach allows one to quantify which aspects of an experience are memorable and which fail to stick. This aspect of our approach opens the door for a much richer characterization of memory that considers not just how much information was recalled, but the particular details of that information as well.

Prior work on neural responses during naturalistic experiences and recall has largely focused on identifying brain regions whose responses are reliably similar across individuals (Lerner et al., 2011; Simony et al., 2016; Chen et al., 2017; Baldassano et al., 2017; Zadbood et al., 2017). This allows one to identify which regions might be processing or representing the stimulus, or retrieving details of the experience during recall. Our approach is fundamentally different. We ask: which brain regions during video viewing match the video structure and how individual participants will recall the video later (in terms of the temporal correlations of their recall topic trajectory)? This latter approach highlights regions such as the medial prefrontal cortex (mPFC) and medial temporal lobes (MTL) that may respond in idiosyncratic ways across individuals, but that nonetheless play an important role in encoding experience into memory. In other work, the mPFC has been suggested to play a role in the representation of a "schema", or a network of prior knowledge that is consistent with the current task (van Kesteren et al., 2012; Gilboa and Marlatte, 2017). One interpretation for mPFC's involvement in our work is that mPFC represents participant-specific relationships with the content of the video, and its representation during scene viewing is related to how a participant later describes the scene. MTL activity patterns also matched participant's recall, which is consistent with a large body of literature highlighting the role of the MTL in episodic memory encoding (Paller and Wagner, 2002; Davachi et al., 2003; Ranganath et al., 2004; Davachi, 2006).

More broadly, these findings have strong implications for how we assess memory in other

naturalistic contexts, such as the classroom or in a doctor's office. Whereas academic tests often measure students' performance using metrics such as the proportion of correctly answered questions, our work suggests that this approach might "miss the forrest for the trees". We view "true learning" as understanding key concepts (i.e., understanding central themes in the learned content and how they relate) rather than regurgitating the greatest number of facts. In addition to educational contexts, our approach may provide unique metrics that can be used to assist in the diagnosis of memory disorders, and other psychiatric disorders that influence memory. For example, while the quantity of information recalled could be roughly matched between a healthy and patient population, and other aspects of the memory (such as the shape, serial order, precision or distinctiveness) might be different. Thus, this work serves as a foundation for more nuanced approaches to memory assessment that consider the trajectory and specific contents of memory for a naturalistic experience.

An important question for future work concerns the factors that drive an individual to sample their recall trajectory finely or coarsely. For example, given a short recall interval, would participants intuitively gravitate towards coarser samplings that still outline the basic shape of the video's topic trajectory? Or if participants were told that their narrations would be played back to other participants (Zadbood et al., 2017), would that change the resolution or shape of their recalls? And over successive recountings of the same sequence of events or with more elapsed time between encoding and recall, how do the shapes of the trajectories change? For example, loss of detail would result in a "smoothing out" of the trajectories with each new retelling.

While we view this work as a major advance in characterizing and understanding human memory, as with any approach there are limitations. First, the approach relies on having a "good" model of the stimulus (e.g., one that describes the video in the same way as participants recall it), which is currently only achievable by a human hand-annotating each moment of the video. To increase the scalability of this approach, future work could explore automated methods for extracting meaning from videos (Haonan et al., 2016). Another potential limitation is that by its nature, the model extracts the "gist" of scenes from the video. This provides desirable flexibility (e.g., participants can use different combinations of words to describe the scenes), but this comes at the expense of capturing specific details. Thus, a future direction of this work will be to increase sensitivity to such details while maintaining flexibility.

To conclude, we'll revisit the question of what it means to remember something. Our view is that "successful remembering" is about recovering the "trajectory" of an experience, rather than the ability to recognize/recall any of its particular isolated details. Decades of research suggest that episodic memories are not veridical and context-free snapshots of the past, and so treating (and modelling) them as such is overly simplistic at best (and counterproductive at worst). While it's undeniable that these models have been useful in advancing our understanding of human memory, they are severely limited in their ability to explain memory for real life experiences. Real life experiences are highly structured in time, and so to have a complete understanding of the human memory system, our experiments and models of memory must not ignore this fact. Our work provides a theoretical advance in our understanding of what it means to remember as well as a novel methodological tool to study it.

Methods

Participants and Experimental Design

Participants (n = 17) viewed the first 50 minutes of "A Study in Pink", an episode of the BBC series, *Sherlock*. Immediately upon completion of the video, participants were instructed to (verbally) recount the events in the *Sherlock* episode in their original order and in as much detail as possible. During the entire experiment, participants were in an fMRI scanner. For comprehensive details of the experimental procedures, please refer to Chen et al. (2017).

Fitting the topic model to the video text and recall transcripts

A topic model was used to estimate the most likely mixture of topics for a given sample of text. First, the video was manually segmented into 1000 scenes, and a collection of descriptive features was manually transcribed. For each scene, we considered the following features: narrative details (a sentence or two describing what happened in that scene), whether the scene was indoor or outdoor, name of all the characters in the scene, name of the character in focus, name of the character speaking, location, camera angle, music presence, and text on the screen. We concatenated the text for all of these features within each segment, creating a "bag of words" describing each scene. We then transformed the text descriptions into overlapping windows of 50 scene segments. For example, the first text sample comprised the text from the first 50 segments, the 2nd comprised the text from n+1:n+51, and so on. We trained our model using these overlapping text samples with scikit-learn's (version 0.19.1) 'CountVectorizer' and 'LatentDirichletAllocation' classes (Pedregosa et al., 2011) implemented using our high-dimensional visualization/analysis software, Hypertools (Heusser et al., 2018). First, the text was transformed into a vector of word counts (after removing English stopwords). This gave a word count vector for each scene in the video. Then, the word count vectors were used to fit a topic model (topics=100, method='batch'). We transformed the text descriptions using the model resulting in a scenes (1000) by topics (100) matrix. The scene descriptions often spanned multiple timepoints (i.e., TRs). To account for this, we expanded the video model by copying the rows of the model for as many timepoints that the scene description spanned. After this expansion, the shape of the model was the length of the duration of the video (1976 TRs).

To create the recall models, for each participant we tokenized the recall transcript into a list of sentences and then mapped the list to overlapping windows of 10 sentences. We transformed the list of overlapping recall sentences using the model that was trained on the video text (as described in the paragraph above). The result of this was a sentences (range: 68-294) by topics (100) matrix for each participant that represented the most likely mixture of topics for a given chunk of sentences.

Choosing topic model parameters

There were 3 critical parameters related to fitting the topic model: 1) the number of topics, 2) the window size of text descriptions of the video used to fit the model, and 3) the window size of recall sentences used to transform the recall data. To chose these parameter values, we performed a grid search where the range of possible parameter values was 1, 5, 10, 25, 50, 100, 200, and 500. Our optimization objective was defined as the correlation between the hand annotated memory performance and the root mean squared distance between the video model and the recall model before any further processing (e.g., hidden Markov modeling, averaging within event, etc). While

many of the parameter combination elicited moderately high correlations, the optimal choice was 100 topics, 50 video segments and 10 recall sentences.

Extracting events using a hidden Markov model

The topic model timepoint-by-timepoint correlation matrices all exhibited a block-diagonal structure (with small off-diagonal values), suggesting that the models were comprised of a number of sequential 'states' (or events, see Fig. S2). To capture this structure, we fit the video and each recall model using a hidden Markov model (HMM). Given a number of states or events (k), the HMM recovers a set of labels that segments consecutive timepoints into k events (Rabiner, 1989; Baldassano et al., 2017). To implement this analysis, we used the BrainIAK toolbox (Baldassano et al., 2017; Capota et al., 2017).

Our metric for choosing the "best fitting" HMM was to choose the model with the k value that maximized the ratio of the average 'within-event' correlation values (i.e., the correlation values for blocks of consecutive timepoints the model identified as one event) and the average 'across-event' correlation (i.e., the rest of the correlation values). Additionally, we included a penalty parameter that was proportional to the smoothing of the model that preferred models with smaller k values. We chose k values separately for the video model and for each recall model. Then, using the best k values, We fit a separate HMM for the video and each recall model. Finally, we averaged over timepoints identified to be in the same event resulting in a events by topics matrix for the video model and each of the recall models.

Matching recall events to video events

To estimate which video event each recall event referred to, we correlated the video events model and each recall events model. This resulted in a video events (34) by recall events (8-27) correlation matrix (for each participant) which contains the similarity between each video event and each recall event (see Fig. S3). To find the most likely video event that a given recall event referred to, we computed the argmax over the columns of this matrix. The result was a list of video event indices for each participant. These indices are analogous to the values found in a "recall matrix" from a free recall list learning experiment, but represent the recall of particular events (instead of words, for example).

Visualizing the video and recall event models

To visualize the temporal structure of the video event model (34 events by 100 topics) and the recall event models (8-27 events by 100 topics), we used a technique called UMAP (McInnes and Healy, 2018) to reduce the "topic-space" from 100 dimensions down to 2 dimensions. UMAP is a nonlinear dimensionality reduction technique which models the manifold of the data with a fuzzy topological structure, and then searches for a (2D) projection of the data that has the closest equivalent fuzzy topological structure. We concatenated (vertically stacked) all event models (video, average recall, and individual recall), and then fit and transformed all of the models at once. This assured that the models were projected into the same space.

Vector field analysis

To quantify the flow of recall from event to event, we performed a vector field analysis. We tiled the 2D topic space (x, y: -6 to 6 by .25) with an evenly spaced grid. For each grid point, we drew a circle around the point (radius= 0.5). Then, we tested whether any line segments (formed by event recall transitions) passed through this area of the topic space. For example, say that a participant transitioned from recalling event 2 to event 3. These 2 recall events correspond to 2 points in topic space, and connecting them forms a line segment. We collected all line segments that passed through a given section of topic space (collapsing across participants). To plot the average direction of the line segments (i.e., the arrows for each grid point in Fig. 3B), we converted each of them to unit vectors and then averaged. For grid points where the direction was consistent (across all participants contributing to that point), the length of the arrow approaches 1, whereas if the direction was random the length of the arrow approaches 0. Lastly, we converted each unit vector to an angle (in radians) by taking the inverse tangent of the x, y components of the vector. To test whether the distribution of angles was significantly non-uniform (i.e., displayed a preferred direction across participants), we performed a Rayleigh test on the angles ($p < 0.001$, FDR-corrected at $p < 0.05$ using Benjamani-Hochberg procedure). Arrows where the Rayleigh test was significant are displayed in color (the darker the blue the more significant) while non-significant tests are displayed in gray with lower opacity.

fMRI analyses

Participants viewed and recalled the video stimulus inside an fMRI scanner. The video was split into two parts of approximately equal length (946 and 1030 TRs, $TR = 1.5 seconds$). All data were preprocessed and transformed to 3mm MNI space as described in (Chen et al., 2017). Data were z-scored across time at every voxel. 6mm smoothing was applied. Files are cropped so that all video-viewing data are aligned across participants, and all recall data are aligned to the scene timestamps below. The cropping includes a constant 3-TR (4.5 sec) shift to correct for hemodynamic lag.

Searchlight analysis

Our multivariate analyses were designed to capture brain regions whose timepoint-by-timepoint correlational structure mirrors the correlational structure of the video model as well as participant-specific recall topic models during video viewing. We conducted a searchlight analysis (5x5x5 voxel cube) where for each cube, we correlated the model timepoint-by-timepoint correlation matrix with the neural correlation matrix. To aggregate across participants, we Fisher's z-transformed the correlations and then averaged. To assess significance, we recomputed this group analysis 100 times, but randomly phase shifted the model by the same amount for each participant but different amounts for each permutation to build a null distribution of correlation values. Finally, we thresholded the group averaged correlation maps where the 'real' correlation value for a given voxel exceeded the 95th percentile of the null distribution. To correct for non-linearities between the viewing time and recall time, for each participant we used dynamic time warping to temporally align the matrices. The algorithm recovers a path of coordinates that would bring the video and recall model in maximal temporal alignment. We used this path to warp the fMRI data and the recall model into temporal alignment (separately for each participant).

Topic vector word clouds

We created word clouds to visualize the themes contained in the recall events. One component of the topic model comprises a words (2117) by topics (100) matrix (R), where the rows represent the weight of a given word in each topic. To find words that were maximally associated with a particular event vector, we computed the dot product between R and v, which gave a 1 by # of words vector where the values represent the "activation" of each of in the event. Activation is defined as the weight of a particular word in a particular event. Then, we created word clouds by extracting the top n words and plotting them where the size of the word is proportional to its activation in the event.

In the first analysis (Fig. 4A,B), we quantified the most and least remembered topics/words throughout the entire video by computing a weighted average over all recall events, where the weights were proportional to memory for each recall event. To measure memory for each event, for each participant we computed the correlation between the video event vector and the closest recall event vector. We then averaged these correlation values across participants. We then computed a weighted average of all video events using the correlation values as weights. Next, we computed the dot product between this weighted-average video event vector and the R matrix (described in the paragraph above) to get activations for each word. Finally, we plotted the top 200 words where the size of the word is proportional to its activation. To get the least remembered topics/words, we performed the same analysis but inverted memory weights.

In the second analysis (Fig. 4C), we created wordles for the top/bottom 3 remembered video events indexed by the average correlation values (Fig. 4A). To get the "activations" for words associated with the video events, we computed the dot product between the video event vector and the R matrix. The same procedure was used to get word activations for the recall events. We then plotted the top 200 words for the top/bottom 3 recalled events.

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.

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.

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.

Chen, J., Leong, Y. C., Norman, K. A., and Hasson, U. (2017). Shared experience, shared memory: a common structure for brain activity during naturalistic recall shared experience, shared memory: a common structure for brain activity during naturalistic recall. *Nature Neuroscience*, 20(1):115–125.

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.

DuBrow, S. and Davachi, L. (2013). The influence of contextual boundaries on memory for the sequential order of events. *Journal of Experimental Psychology: General*, page In press.

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

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

Haonan, Y., Jiang, W., Zhiheng, H., Yi, Y., and Wei, X. (2016). Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593.

Heusser, A. C., Ziman, K., Owen, L. L. W., and Manning, J. R. (2018). 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.

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

Kahana, M. J. (2012). *Foundations of Human Memory*. Oxford University Press, New York, NY.

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.

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

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. *Zenado*, 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.

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.

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.

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

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.

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

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.

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.