Capturing the geometric structure of our experiences and how we remember them

Andrew C. Heusser & Jeremy R. Manning

August 16, 2018

1 Abstract

The human memory system is adept at cataloging the rich dynamics of ongoing experience. A defining feature of our everyday experiences is that they unfold in a structured and predictable manner, governed by the laws of the physical systems that produce them. However, this temporal structure is typically absent (or severly impoverished) in traditional trial-based memory experiments. As a consequence, our understanding of how the inherent temporal structure of our experiences influences memorability and memory organization is limited. In this study, we investigate how people verbally recount a video by characterizing and relating the thematic dynamics, or "trajectories," of the stimulus and participants' recalls. We found that despite large differences in the amount of information recalled across participants as well as the particular words used to describe the experience, the overall shape of the stimulus was recapitulated by most of the participants. These findings provide a window into which aspects of naturalistic experiences must be preserved, and which might be more flexible, in considering whether and how those experiences are remembered.

2 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 binary operation: either an item is recalled or it isn't. More nuanced studies might incorporate self-reported confidence measures as a proxy for memory strength, or ask participants to discriminate between "recollecting" an experience or a feeling of "familiarity" (Yonelinas, 2002). However, characterizing and evaluating memory in more realistic contexts (e.g., telling a story to a friend about a recent vacation) is much more nuanced. Real-world recall is continuous, rather than binary. The specific words used to describe an experience have very little bearing on whether the experience is considered to have been "remembered." Further, one might remember the gist of an experience but forget (or neglect to recount) particular details. Or different people who share an experience might recount the experience with a similar level of detail, but the specific details that were remembered might vary across people. Which aspects of those recollections should be considered fundamental, and which are extraneous to the main story?

Human memory studies typically assess memory for the "items" within an experience, such as the people, places and things encountered. However, real life experiences can also be described in terms of their temporal structure (i.e. how the contents relate/change over time). Naturalistic experiences are typically autocorrelated in space and time on short timescales: the contents of an experience are often similar from moment to moment and change gradually. For example, many of us spend our mornings/evenings at home and much of our days at work (e.g. an office). While the specific happenings within these contexts differ, they are often highly correlated (and thus predictable) over time. Furthermore, our experiences can also be characterized by longer timescale correlations. For example, consider returning to your office after taking a lunch break. The spatial contexts before and after lunch are highly similar, but separated in time. Thus, while evaluating the ability to recall specific contents of an experience is one way to assess memory, another important dimension to consider is the temporal structure of the remembered experience, and how the temporal structure of the memory relates to the original experience. However, these recurrent and gradually changing temporal dynamics are not typically present in traditional memory studies, but are likely critical if we wish to understand how our memory systems remember our life experiences.

Despite the continuous and (often) gradually changing nature of naturalistic experiences, prior research suggests that our memories for those same experiences may be more descrete and organized around particular "events" (**???**). Short timescale autocorrelations in the contents of an experience are thought associatively link information in memory, while more drastic changes in perceptual or semantic aspects of an experience (i.e. "event boundaries") may lead to discontinuous in memory (**???**). There is now compelling evidence that event boundaries can influence memory organization when learning sequences of items (**?**DuBrow and Davachi, 2013), during navigation (**?**), and narrative comprehension (Zwaan and Radvansky, 1998; Ezzyat and Davachi, 2011). However, whether people naturally recount their experiences in a way that is structured around events is not known.

In this study, we analyzed an open dataset in which participants view and then verbally recount an episode of the BBC series *Sherlock* (Chen et al., 2017). To capture the temporal structure present in naturalistic experiences, we developed a novel computational approach based on topic modeling (Blei et al., 2003) and hidden Markov models (Rabiner, 1989; Baldassano et al., 2017). We propose that the temporal structure of an experience gives it a unique geometric "shape" or "trajectory", defined by how the contents of the experience relate and change over time. We hypothesized that when recounting naturalistic experiences, the precise details that are recapitulated and language used may differ across people, but the trajectory of the stimulus is generally preserved.

2.1 Results

2.1.1 A model to capture the temporal and semantic structure of naturalistic stimuli

We fit a topic model (Blei et al., 2003) to hand-annotated text descriptions of scenes from the video. The text descriptions contained details of the scene such as the characters, location, and a short summary of the scene (see Fig.1 for example text, see 4 for analysis details). We then transformed the text descriptions using the (same) topic model, resulting in a scenes (1976) by topics (100) matrix, where each row of the matrix represents a probabalistic mixture of topics discovered in that scene (See Fig. 1 for example topic vectors). As depicted in Fig. 2a, the topic vectors are sparse and change slowly over time. Furthermore, there are clear transitions from one topic 'state' to the next, possibly indexing scene transitions in the stimulus. To get a better handle on this temporal structure, we computed a timepoints (1976) by timepoints (1976) correlation matrix of the

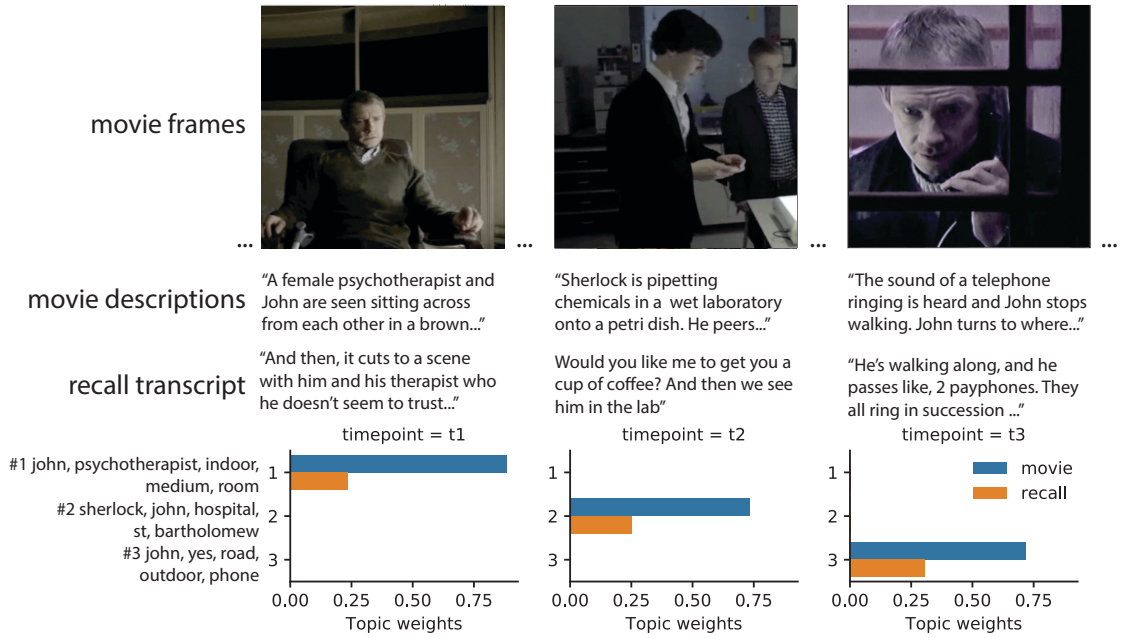

Figure 1: Schematic of the analysis approach. For each scene in the video, text descriptions were generated by hand. Three exemplary time points are displayed here. Below the video descriptions are text samples from an example participant's verbal recall transcript. We fit a topic model to the moment-by-moment video text descriptions and transformed participant's verbal recall transcripts using this same model. The bar charts display the resulting topic model weights for the video (in blue) and recall (in orange) for three example topic dimensions.

video model (Fig. 2c). This correlation matrix reveals that the model has a strong, block-diagonal structure. Another noteworthy characteristic is that there is very little correlation between blocks (i.e. the off-diagonal values are small), suggesting the representations of each block are unique and highly descriminable.

2.1.2 Modeling verbal recall

After watching the video, participants verbally recalled (in order) as much of the episode as they could. We used the same topic model (fit with the text descriptions of the video) to transform participants' verbal recall transcripts. The result was a sentences (range: 68 to 294, mean: 131.94) by topics (100) matrix for each participant, where each row represents a probabalistic mixture of topics for a given window of sentences during recall. An example participant's (#13) recall model is plotted in Fig. 2b. Like the video model, topic vectors were sparse and changed gradually. Note that the topics were derived solely from the text descriptions of the video, and so the topic models estimated for recall are dependent on the topics present in the video. This approach effectively projects the participants' verbal recall into the video's "topic space", allowing us to

directly compare them to the video model.

Next, we investigated the temporal structure of the recall matrices. For each participant, we computed a timepoint (68 to 294) by timepoints by timepoints (range: 68 to 294) correlation matrix from the recall models. An example participant's (#13) correlation matrix is plotted in Fig. 2d). Like the video model, each participants' recall correlation matrix exhibited a strong block-diagonal structure (Supp Fig. 1). Notably, this suggests that participants recounted the video in discriminable segments, potentially related to the recall of specific events (or scenes) from the video.

2.1.3 Segmenting the video and recall models into 'events'

As described above, a striking feature of the video and recall correlation matrices is a strong, block structure along the diagonal of the matrices (see Fig. 2c,d). We hypothesized that this structure might arise from transient stability in the contents of the video (i.e. "events"), as well as the subsequent language used to recount each event of the video. If true, we reasoned that it would be possible to identify the specific event a participant is describing by comparing the topic vector for a particular recall "event" with the topic vector for each video event. In other words, the video event that is most similar (i.e. correlated) to a particular recall event is the event that the participant is most likely describing.

To test this idea, first we segmented the video and recall models in k events (i.e. states) using a hidden Markov model (Baldassano et al., 2017). Our algorithm determined 34 events for the video model and a range of values (range: 8-27; mean: 15.41; SD: 5.6) for the recall models (see Methods for details on choosing an optimal k value). The events discovered for the model and a representative participant (#13) are highlighted as yellow rectangles outlining blocks along the diagonal of the correlation matrices (Fig. 2b,d).

Next, we created a video "event model by" averaging together neighboring topic vectors that the were classified to be in the same event, resulting in an events (34) by topics (100) matrix (Fig. 2e). We performed the same procedure for the recall matrices (see Fig. 2f for example). Then, we computed the correlation between video and recall event models, resulting in a video events (34) by recall events (8-27, depending on the participant) correlation matrix (Fig 2g). These matrices represent the similarity (correlation) between each video event and each recall event (for each participant). To determine which video event a particular recall event was most likely describing, we computed the index of the video event with the highest correlation to the recall event (i.e. the argmax). This is depicted in Fig. 2g as the cells highlighted with a yellow border. Interestingly, our algorithm suggests that the example participant recalled most of the video events in order.

Then, we computed a group-averaged recall event model and video-recall event correlation matrix. For each participant (and each recall event), we sorted the recall event vectors (across all participants) according to the video event with the highest correlation. We then averaged the recall event vectors within each group. This yielded an average recall event vector for all but one (of 34) video events, since no participant remembered one of the events according to our model. Lastly, we computed an average recall event (34) by video event (34) correlation matrix, and highlighted the cell with the highest correlation value with a yellow border (Fig. 2h). Notably, this matrix displayed high correlation values along the diagonal and low correlations in the off-diagonal cells. This suggests that on average, participants were able to recapitulate the events in the episode in a specific, highly descriminable way.

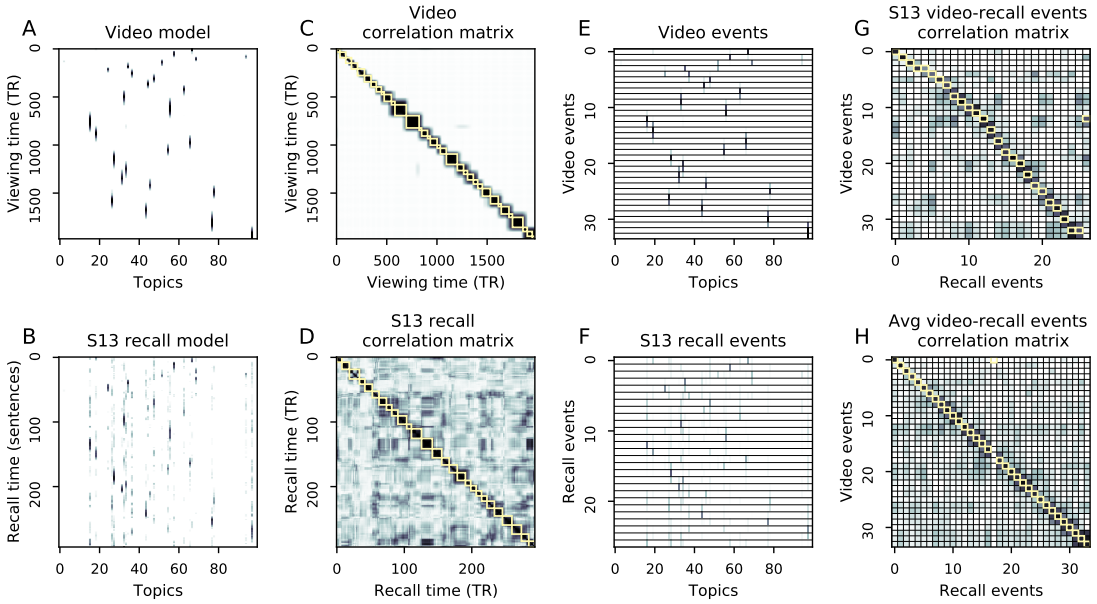

Figure 2: Modelling naturalistic stimuli and recall. A depiction of our analysis pipeline. For all plots, darker colors indicate greater values and the range of each plot is 0-1. A). A timepoints (1976) by topics (100) matrix representing the video stimulus. Each row represents the most likely mixture of topics for a given timepoint (i.e. topic weights). Each column represents a different topic. B). A timepoints (1976) by topics (100) matrix representing participant #13's recall. C). A viewing-time (1976) by viewing-time (1976) correlation matrix representing the correlation of each moment of the video model with every other moment of the video model. The white boxes represent 'events' recovered by a hidden Markov model. D). A recall-time (294 sentences) by recall-time (294) correlation matrix for participant #13. E). An events (34) by topics (100) matrix where each row represents the average topic vector for each event in the video model. F). An events (27) by topics (100) matrix where each row represents the average topic vector for each event in participant #13's recall model. G. A recall events (27) by video events (34) correlation matrix for participant #13. The cells with a yellow border identify the video event with the highest correlation to a given recall event. F. A group averaged recall events (34) by video events (34) correlation matrix. The cells with yellow borders are the video event with the highest correlation to a given average recall event.

2.1.4 Characterizing naturalistic memory with traditional list-learning analyses

Just like in a traditional "free-recall" list-learning experiment where participants view a list of words and then verbally recall them, our video-recall matching analysis approach affords us the ability to analyze memory in the same way. The recalled events can be treated as "items" analogous to words recalled in a list-learning study. In our first set of analyses, we sought to characterize memory performance/dynamics by extending classic analyses originally designed for list-learning experiments to more naturalistic settings.

First, we asked whether the estimated number of recall events (k) by participant was related to hand-annotated accuracy as published in Chen et al. (2017). We found a strong positive correlation where subjects with a greater number of recall events also had better overall memory performance

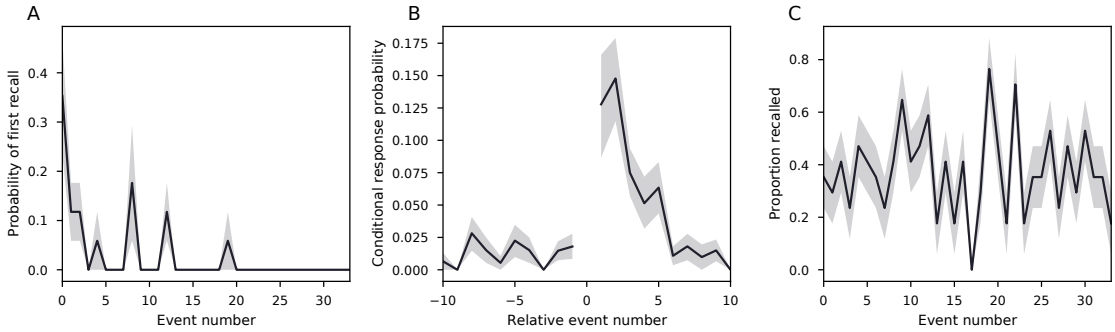

Figure 3: Naturalistic extensions of classic list-learning memory analyses. A). The probability of first recall as a function of the serial position of the event during encoding. B). A lag-conditional response probability curve. Given recall of event i, the probability that the next recalled item will be from serial position i +/- lag. C). Proportion of events recalled as a function of serial position. All error bars are the standard error of the mean derived from a bootstrap resampling procedure.

(Pearson's $r(16)=.67$, $p=.003$). Then, we considered how participants initiated the recall sequence (known in the literature as the 'probability of first recall' or 'PFR'). We found that participants tended to initiate their recall sequences with the first few events (Fig. 3a), which is qualitatively very similar to previously published list learning experiments (REF?). Next, we considered another well-studied memory measure in the list-learning literature, the lag conditional response probability curve (or lag-CRP) (Kahana, 1996). The result suggests a strong bias to transition sequentially events in the forward direction (Fig. 3b). Finally, we assessed memory performance for each event in the video as a function of its serial position during encoding (Fig. 3c). We found that there was substantial variability in memory for the different events (STAT).

We also considered two additional across-participant measures of recall that characterize memory organization: temporal clustering and semantic clustering. We found that participants who clustered in time also recalled a greater number of events (Pearson's $r(16)=.62$, $p=.007$). Next, we assessed semantic clustering. We found that the semantic clustering score was related to memory performance across participants (Pearson's $r(16)=.55$, $p=.02$). Thus, participants who organized their recalls with respect to the semantic information contained in the scene had better memory performance.

2.1.5 The shape of a naturalistic experience is preserved in recall

The analyses described in the previous sections are useful in that they allow us to quantify memory dynamics for naturalistic stimuli in a way that is comparable to a vast literature on trial-based list-learning experiments. Furthermore, we introduce novel measures (precision and distinctiveness) which provide additional details regarding the veracity of naturalistic memories. However, these methods fail to capture the rich temporal structure present in naturalistic stimuli and associated recall of those experiences. Thus, our next set of analyses test whether the "shape" of an experience is preserved in memory. We hypothesized that despite substantial individual variability in the precision and amount of content recalled across by participants, the overall shape of the episode would be recapitulated by the majority of participants.

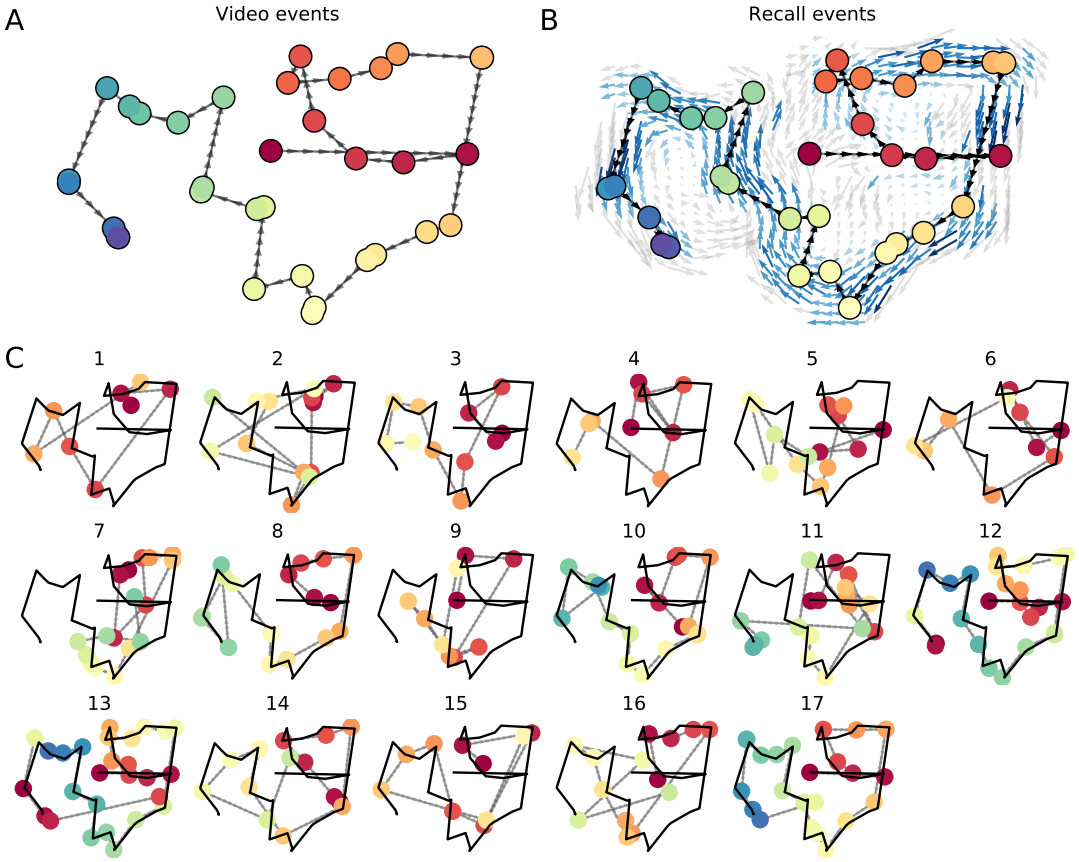

Figure 4: Video and recall trajectory plots. A). 2-dimensional embedding of the video events model. The arrows indicate the forward direction of the video events. B). 2-dimensional embedding of the average recall events model. The colors refer to the most similar video events. The directional lines connecting the points represent the true video event order. The arrows represent the average direction of all recall event transitions (i.e. a line segment connecting two consecutive recall event vectors) that intersected a box (width=.25) centered on the origin of the arrow.

To visualize the relationship between the video and recall event models, we embedded them into a 2D space (using the UMAP dimensionality reduction algorithm, for details see 4) where the points represent video/recall events and the distance between them represents their similarity in "topic space" (Fig. 4a). We observed that the two models have a very similar temporal evolution and geometric structure. To further quantify this correspondence and to characterize how participants navigated through the space during recall, we analyzed the angle between successive recall events. We created a grid of evenly spaced points (.25 units) in the 2D "topic space". For each point on the grid, we drew a circle (radius=.25) around the point and grouped together all recall transitions that intersected the circle (across subjects). To visualize the average angle, we converted each transition

angle to a unit vector and then averaged the vectors together. To assess consistency in the direction of the recall transition across participants, we performed a Rayleigh test ($p<.001$, corrected at $p<.05$ using permutation procedure described in Methods).Thus participants' recalls followed the same path as the video model (Fig. 4b), suggesting the shape of the stimulus was preserved despite idiosyncratic differences between participants in their recalls.

Notably, many of the participants' recall trajectories appear to recover the overall shape of the video stimulus. To quantify this similarity, we measured the average euclidean distance between each each segment of the video model (e.g. the line formed by successive video events) and the closest recall event for each participant. Inuitively, if the participant approximately remembered every video event, the measure would approach 0. The more missing recall events, the higher the value. We found that this shape measure was highly correlated to the number of events remembered across participants (Pearson's $r(16)=-.87$, $p<.001$). While these measures are highly correlated (as might be expected), there are cases where the measures dissociate. For example participant X and X recalled exactly the same number of events, but the shape measure is drastically different between them. Conversely, participant X and X had the same shape measure, but there was a large difference in the number of events recalled. Thus, measuring similarity in shape between the video and recall models provides unique information about participant's recall behavior that is not captured by classic measures of memory accuracy.

2.1.6 Measuring the quality of naturalistic memories

Representing the video and verbal recall as events in "topic space" also affords us the ability to characterize the quality of recall in a more fine-grained and nuanced way than what was previously possible. To quantify the similarity between the video model and individual recall models, we considered a number of novel metrics. First, we tested whether each participant's recall model matched the movie model in a general sense. To do this, for each participant we filtered the video model to only include the events that the participant remembered. Then, we computed the root mean squared difference (RMSD) between the video model and the recall model. As an example, if the participant remembered all the events in order (with perfect precision), the expected distance value would be 0. However, if they remembered a subset of events, events our of order or with low precision the expected distance would be greater than 0. To assess significance, we performed a permutation test where we circularly shifted the recall model (10000 times) and recomputed the RMSD. The recall model significantly matched the video model for nine of the participants (p¡.05; participants: 3-4, 8-13, 17 and the p-value for the rest of the participants was less than .25). Furthermore, the RMSD values were significantly correlated to hand annotated memory performance across participants (Pearson's $r(16)=-.57$, $p=.016$). Thus, a closer match between the video and recall event models was related to better recall performance.

Next, we tested whether participants who recalled more events were also more precise in their recollections. For each participant, we computed the correlation between each recall event and its matching video event (only for the events which they recalled). This resulted in a single number for each recalled event indexing how similar the recall event was to its matching movie event (i.e the "precision" of the recall). We then averaged the correlations within participant. In line with our prediction, there was a strong correlation between hand annotated memory performance and precision suggesting that participants who remembered more events also remembered them more veridically (Pearson's $r(16)=.74$, $p=.0006$). Next, we considered the distinctiveness of each recall event. That is, how uniquely a recall event matched a given video event compared to all 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 this end, we computed a 'distinctiveness' score for each participant (i.e. 1 - the correlation between a recall event and all non-matching video events). Then, we averaged this measure over recall events within participant. We found that participants with higher hand annotated memory performance also had higher distinctiveness scores (Pearson's $r(16)=.8$, $p=.0001$).

Lastly, we tested whether participants with better memory performance were also more likely to remember the events in order. For each participant, we computed the Spearman rank correlation between the order of events that the participant recalled and the actual order of events (filtering events that were actually recalled). We found that participants who recalled more events also recalled more of them in order (Pearson's $r(16)=.5$, $p=.04$). In summary, we found that better memory performance was associated with more precise, distinctive and ordered recalls.

2.1.7 Memorability by event and topics

A noteworthy advantage of representing naturalistic stimuli/memory using this approach is that the event models can be mapped back to the language used to fit the model(s). This allows for analysis and visualization of the language used when participant's recounted their experiences. In this next analysis, we 1) quantified the memorability of each video event and 2) used memorability to plot the top words for the most and least memorable scenes, and also aggregated across the entire video.

To measure video event memorability, we computed the correlation between each video event and the closest recall event and then averaged those values within participant (Fig. 5a). We used event memorability to create a weighted average over all video event vectors, where video events with better memory were weighted more heavily. Then, we extracted the top 200 words and created "wordles" representing the most memorable themes in the video. The result reveal that scenes containing "Sherlock" and "John" were highly memorable, and also scenes containing words such as "medium", "street" and "baker" (Fig. 5b). To find words related to the least memorable scenes, we inverted the memorability weights and again extracted the top 200 words. Scenes containing "Mike" and "Molly" were least remembered, and also words like "conference", "hospital" and "psychotherapist" (Fig. 5b).

Then, we took a closer look at the themes contained in the top/bottom 3 events, using the memorability weights to select the top/bottom events. We performed the same analysis as described above, but separately for the video and average recall event vectors representing each of the top/bottom 3 remembered events. In Figure 5c, the trajectory represents a low-dimensional embedding of the video model and the large colored dots represent the video events. The smaller colored dots represent the distribution of individual recall events across all participants, colored by the closest video event. We plotted wordles for the top 3 (circled in green) and bottom 3 (red) most/least remembered events, where the video is represented in the left half of the circle and the average recall is represented in the right half of the circle. Visual inspection reveals that the words contained in the memorability weighted worldles (Fig. 5b) overlap with the individual top/bottom event wordles (Fig. 5c). These analyses shed light on the contents of naturalistic stimuli and accompanying memories, and highlight the flexibility in transforming between human and machine readable representations of the data afforded by our modeling approach.

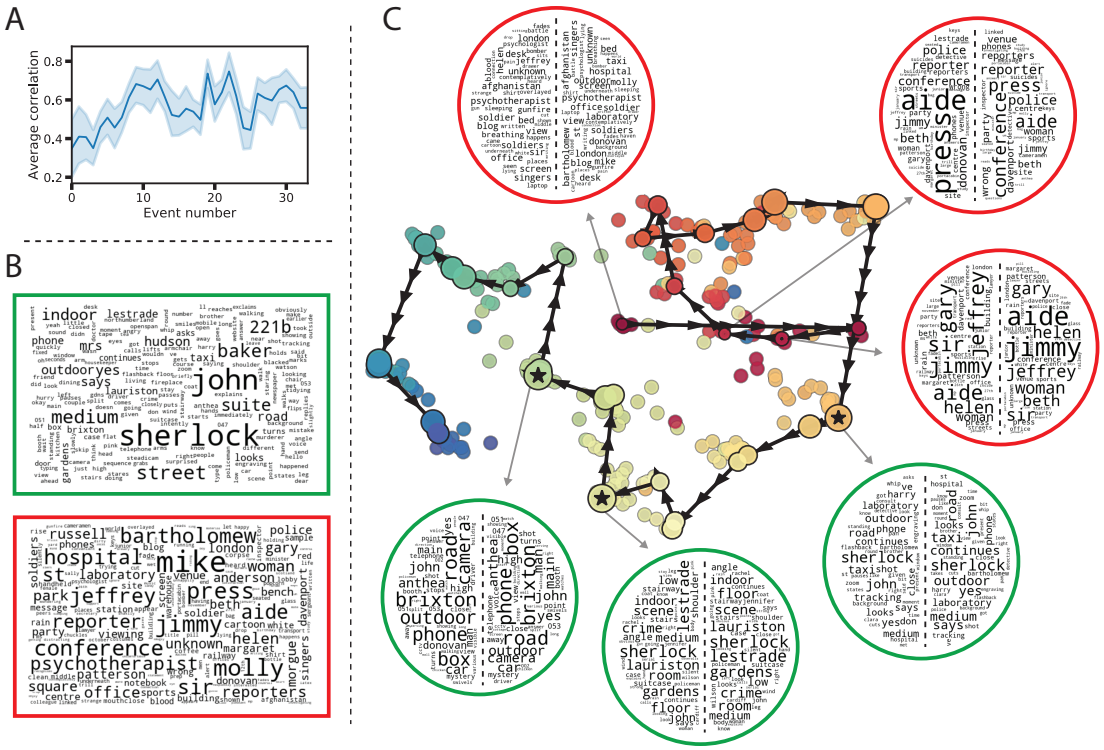

Figure 5: Analysis of topics by event memorability. A). Group-averaged correlation between each video event and closest recall event (per subject). Error bars represent 95% confidence intervals. B). Wordles (top 200 words) representing a weighted average of video event topic vectors weighted by event memorability (e.g. the correlation values in A). The top wordle (green) contains words from the most remembered events and the bottom (red) contains words from least memorable events. The word sizes are proportional to the word's "activation". C). Trajectory represents the video event model embedded in a 2D space using UMAP. The large colored dots are video events. The small colored dots are all individual recall events across all participants, where the colors refer to the closest video event. The circular wordles represent the top 3 most (green circles) and least (red) memorable video events. The left side of the wordle circles contain words associated with the video event vector and the right side contains words associated with the average recall event vector.

3 Discussion

Studying episodic 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 and item recognition 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. 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. 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.

Historically, human memory reseachers have focussed on measuring the quantity of information recovered by an individual (i.e. the proportion correct) under different experimental conditions. However, for real world experiences this is not a well-defined task. 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 work characterizes experience by comparing the overall "shape" of a dynamic stimulus and a memory. We assess the quality of memory for the movie participants viewed by measuring the match between the shapes of the movie'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.

More broadly, these findings have strong implications for how we assess memory in other naturalistic contexts, like the in the classroom or in a doctor's office. Whereas instructors 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". True learning is about understanding the key concepts (i.e. understanding key themes in the learned content and how they relate) rather than about 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 a memory disorder, 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, 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 movie's topic trajectory? Or if participants were told that their narrations would be played back to other participants (cite Hasson work), would that change the resolution or shape of their recalls? And over successive recountings of the same sequence of events, how do the shapes of the trajectories change? For example, loss of detail would result in a "smoothing our" of the trajectories with each new retelling. This could be quantified using our topic trajectory approach.

4 Methods

4.1 Participants and Experimental Design

Participants (n=17) viewed *A Study in Pink*, a 50 minute episode of the BBC series, *Sherlock*. Immediately upon completion of the video, participants were instructed to (verbally) recount the events in the video 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).

4.1.1 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: scene details (i.e. a sentence or two describing what happened in that scene, space (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 words 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 second comprised the text from n+1:n+51, and so on. We trained our model using these overlapping text samples using scikit-learn's (version 0.19.1) 'CountVectorizer' and 'LatentDirichletAllocation' classes. First, the text was transformed into a vector of word counts (default parameters). 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.

4.1.2 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 modelling, averaging within event, etc). While many of the parameter combination elicited moderately high correlations, the best choice was 100 topics, 50 video segments and 10 recall sentences.

4.1.3 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). 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 (REF markov paper). To implement this analysis, we used the Brainiak toolbox (Baldassano 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 each topic 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.

4.1.4 Matching recall events to video events

To figure out 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. 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).

4.1.5 Characterizing memory performance using traditional approaches

Overall Accuracy. To get an overall measure of the quantity of information recalled, we computed the proportion of sucessfully recalled events by counting the number of unique recall events identified by the HMM model and dividing by the total number of video events. We performed this analysis for each participant separately.

Probability of first recall (PFR). The (PFR) analysis represents the probability that an item will be recalled first as a function of its serial position during encoding. We initialized a # of participants (17) by # of video events (34) matrix. Then for each participant, we found the index of the video event that was recalled first 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 34 array representing the proportion of subject that recalled an event as a function of serial position during encoding.

Lag conditional probability curve (lag-CRP). The lag-CRP represents the probability that the next item recalled will be of lag i from the just recalled item. For each recall transition, we computed the lag between the current recall event and the next recall event, normalizing by the total number of possible transitions. This resulted in a # of participants (17) by lags (-33:+33) matrix. We averaged over the rows of this matrix to get a group-averaged lag-CRP.

Serial position curve (SPC). The SPC represents the proportion of participants that remember an item as a function of its serial position during encoding. We initialized a # of participants (17) by # of video events (34) matrix. Then, for each recall event (and each participant), we found the index of the video event that was recalled and filled it in with a 1. This resulted in a matrix where 1s indicate the successful recall of an event in serial position n and zeros indicate the lack of recall for that event. Lastly, we averaged over the rows of the matrix to get a 1 by 34 array representing the proportion of subjects that recalled an event as a function of its serial position.

Temporal clustering. Temporal clustering measures the extent to which participants group their recall responses according to encoding position. For instance, if the participant recalled each item in the presentation order, this would result in a score of 1. If the participant recalled randomly with respect to presentation order, this would result in a score of .5. For each event transition (and separately for each participant), we computed the rank similarity (euclidean distance) between the presentation position of the current and next recall events. The scores were then averaged within participant to get a single number representing the extent of temporal clustering exhibited by a given participant.

Semantic clustering. Similar to temporal clustering, semantic clustering measures the extent to which participants group their recall responses according to semantic similarity. Here, we are using the topic vectors for each event as a proxy for its semantic content. Thus, similarity between the semantic content for two events can be computed by correlating their respective topic vectors. For instance, if each consecutive recall was the next most similar event (in terms of its s), this would result in a score of 1. If the participant recalled randomly with respect to semantic similarity, this would result in a score of .5. For each event transition (and separately for each participant), we computed the rank similarity (correlation distance) between the current recall event and the next recall event. The scores were then averaged within participant to get a single number representing the extent of semantic clustering exhibited by a given participant.

4.1.6 Visualizing the video and recall event models

To better understand 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 to reduce the "topic-space" from 100 dimensions down to 2 dimensions (REF). 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.

4.1.7 Vector field analysis

To quantify (and visualize) 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 square around the point (height/width=.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 4b), 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 (REF). Arrows where the Rayleigh test was significant are displayed in color while non-significant tests are displayed in gray with lower opacity. Note that it was not computationally tractible to perform this analysis in the original 100 dimensional space because the number of grid points grows factorially with the number of topic dimensions.

4.1.8 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 (XX) 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. 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. 5a/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. 5c), we created wordles for the top/bottom 3 remembered video events indexed by the average correlation values (Fig. 5a). 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.

4.2 fMRI analyses

Seventeen participants watched the first 50 minutes of Episode 1 of BBC's Sherlock. The video was split into two parts of approximately equal length (946 and 1030 TRs). All data were preprocessed and transformed to 3mm MNI space as described in the paper. Data were zscored 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.

4.2.1 Searchlight analysis

Our multivariate analyses were designed to capture brain regions whose timepoint-by-timepoint correlational structure mirrors the correlational structure of the participant-specific recall topic models during video viewing. To correct for non-linearities between the viewing time and recall time, for each participant we used DTW 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). Then, 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 # of timepoints - 1) 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.

References

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

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.

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.

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

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

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

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.

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

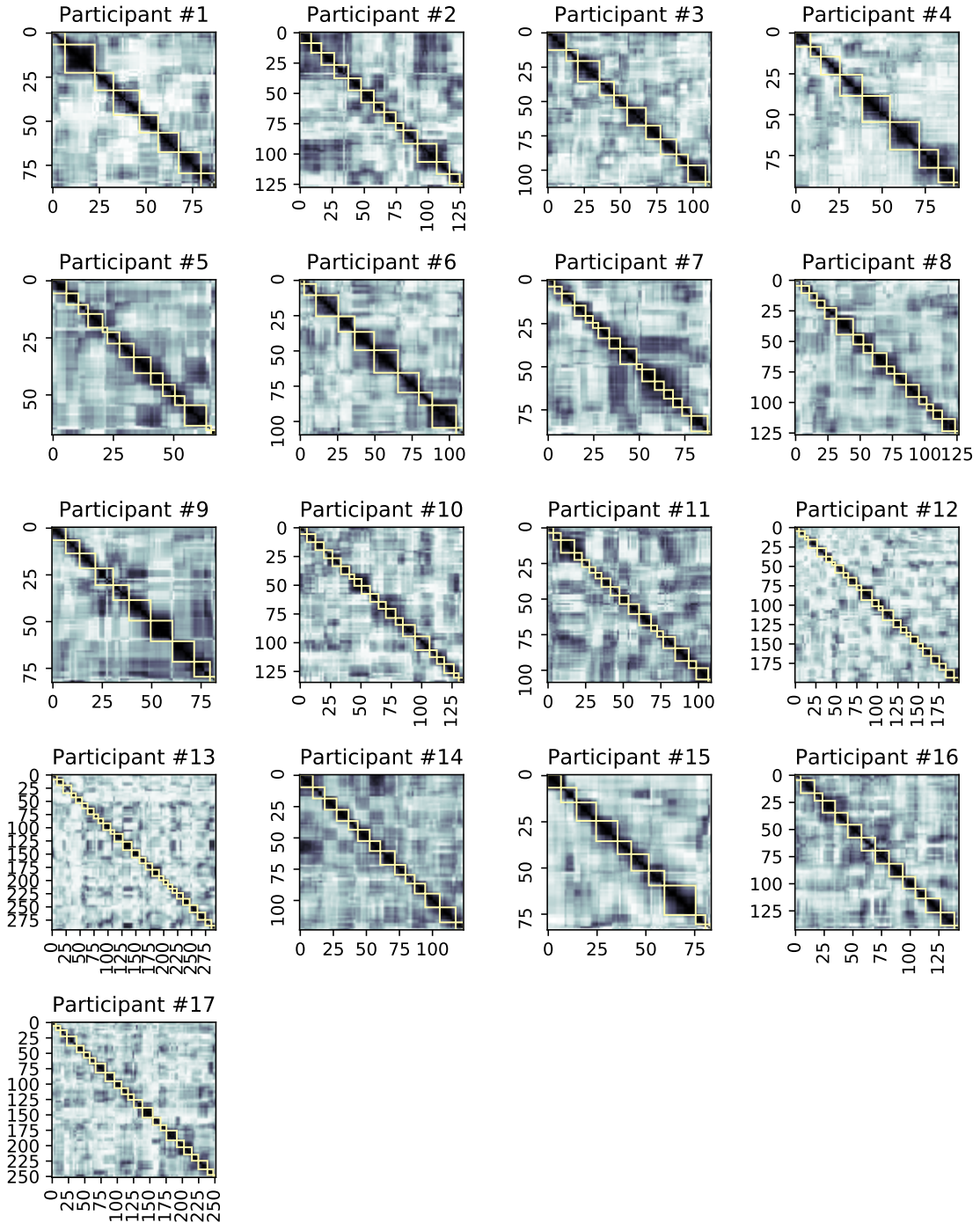

Figure 6: Recall model correlation matrices and event segmentation fits. Each participant's timepoint-by-timepoint recall correlation matrix. The yellow boxes represent "events" identified by a hidden Markov model.