nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-021-01051-6

Geometric models reveal behavioural and neural signatures of transforming experiences into memories

Andrew C. Heusser[1,2,3], Paxton C. Fitzpatrick[1,3] and Jeremy R. Manning[1]✉

How do we preserve and distort our ongoing experiences when encoding them into episodic memories? The mental contexts in which we interpret experiences are often person-specific, even when the experiences themselves are shared. Here we develop a geometric framework for mathematically characterizing the subjective conceptual content of dynamic naturalistic experiences. We model experiences and memories as trajectories through word-embedding spaces whose coordinates reflect the universe of thoughts under consideration. Memory encoding can then be modelled as geometrically preserving or distorting the 'shape' of the original experience. We applied our approach to data collected as participants watched and verbally recounted a television episode while undergoing functional neuroimaging. Participants' recountings preserved coarse spatial properties (essential narrative elements) but not fine spatial scale (low-level) details of the episode's trajectory. We also identified networks of brain structures sensitive to these trajectory shapes.

What does it mean to remember something? In traditional episodic memory experiments (for example, list-learning or trial-based experiments[1,2]), remembering is often cast as a discrete, binary operation: each studied item may be separated from the rest of one's experience and labelled as having been either recalled or forgotten. More nuanced studies might incorporate self-reported confidence measures as a proxy for memory strength, or ask participants to discriminate between recollecting the (contextual) details of an experience and having a general feeling of familiarity[3]. Using well-controlled, trial-based experimental designs, the field has amassed a wealth of information regarding human episodic memory[4]. However, there are fundamental properties of the external world and our memories that trial-based experiments are not well suited to capture[5,6]. First, our experiences and memories are continuous, rather than discrete—isolating a naturalistic event from the context in which it occurs can substantially change its meaning. Second, whether or not the rememberer has precisely reproduced a specific set of words in describing a given experience is nearly orthogonal to how well they were actually able to remember it. In classic (for example, list-learning) memory studies, by contrast, the number or proportion of exact recalls is often considered to be a primary metric for assessing the quality of participants' memories. Third, one might remember the essence (or a general summary) of an experience but forget (or neglect to recount) particular low-level details. Capturing the essence of what happened is often a main goal of recounting an episodic memory to a listener, whereas the inclusion of specific low-level details is often less pertinent.

How might we formally characterize the essence of an experience, and whether it has been recovered by the rememberer? And how might we distinguish an experience's overarching essence from its low-level details? One approach is to start by considering some fundamental properties of the dynamics of our experiences. Each given moment of an experience tends to derive meaning from surrounding moments, as well as from longer-range temporal associations[7–9]. Therefore, the time course describing how an event unfolds is fundamental to its overall meaning. Further, this hierarchy formed by our subjective experiences at different timescales defines a context for each new moment[10,11] and has an important role in how we interpret that moment and remember it later[9,12]. Our memory systems can leverage these associations to form predictions that help guide our behaviours[13]. For example, as we navigate the world, the features of our subjective experiences tend to change gradually (for example, the room or situation we find ourselves in at any given moment is strongly temporally autocorrelated), allowing us to form stable estimates of our current situation and behave accordingly[14,15].

Occasionally, this gradual drift of our ongoing experience is punctuated by sudden changes or shifts (for example, when we walk through a doorway[16]). Previous research suggests that these sharp transitions (termed event boundaries) help to discretize our experiences (and their mental representations) into events[16–21]. 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 distils those experiences down to their essences. For example, previous work has shown that event boundaries can influence how we learn sequences of items[18,21], navigate[17] and remember and understand narratives[15,20]. This work also suggests a means of distinguishing the essence of an experience from its low-level details: the overall structure of events and event transitions reflects how the high-level experience unfolds (that is, its essence), while subtler event-level properties reflect its low-level details. Previous research has also implicated a network of brain regions (including the hippocampus and the medial prefrontal cortex) in having a critical role in transforming experiences into structured and consolidated memories[22].

Here we sought to examine how the temporal dynamics of a naturalistic experience were later reflected in participants' memories. We also sought to leverage the above conceptual insights into the distinctions between an experience's essence and its low-level details to build models that explicitly quantified these distinctions. We analysed an open dataset that comprised behavioural and functional

[1]Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA. [2]Akili Interactive Labs, Boston, MA, USA. [3]These authors contributed equally: Andrew C. Heusser, Paxton C. Fitzpatrick. ✉e-mail: Jeremy.R.Manning@Dartmouth.edu

magnetic resonance imaging (fMRI) data collected as participants viewed and then verbally recounted an episode of the BBC television show *Sherlock*[23]. We developed a computational framework for characterizing the temporal dynamics of the moment-by-moment content of the episode and of participants' verbal recalls. Our framework uses topic modelling[24] to characterize the thematic conceptual (semantic) content present in each moment of the episode and recalls by projecting each moment into a word-embedding space. We then use hidden Markov models (HMMs)[25,26] to discretize this evolving semantic content into events. In this way, we cast both naturalistic experiences and memories of those experiences as geometric trajectories through word-embedding space that describe how they evolve over time. Under this framework, successful remembering entails verbally traversing the content trajectory of the episode, thereby reproducing the shape (essence) of the original experience. Our framework captures the episode's essence in the sequence of geometric coordinates for its events, and its low-level details by examining its within-event geometric properties.

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

Results

To characterize the dynamic content of the *Sherlock* episode and participants' subsequent recountings, we used a topic model[24] to discover the episode's latent themes. Topic models take as inputs a vocabulary of words to consider and a collection of text documents, and return two output matrices. The first of these is a topics matrix whose rows are topics (or latent themes) and whose columns correspond to words in the vocabulary. The entries in the topics matrix reflect how each word in the vocabulary is weighted by each discovered topic. For example, a detective-themed topic might weight heavily on words such as '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 the mixture of discovered topics reflected in each document.

Chen et al. collected hand-annotated information about each of 1,000 (manually delineated) time segments spanning the roughly 50 min video used in their study[23]. Each annotation included a brief narrative description of what was happening, the location where the action took place, the names of any characters on the screen, and other similar details (for a full list of annotated features, see Methods). We took the union of all unique words (excluding stop words, such as 'and', 'or' and 'but') across all features from all annotations as the vocabulary for the topic model. We then concatenated the sets of words across all features contained in overlapping sliding windows of (up to) 50 annotations, and treated each window as a single document for the purpose of fitting the topic model. Next, we fit a topic model with (up to) $K = 100$ topics to this collection of documents. We found that 32 unique topics (with non-zero weights) were sufficient to describe the time-varying content of the episode (see Methods; Fig. 1 and Supplementary Fig. 2). We note that our approach is similar in some respects to dynamic topic models[27], in that we sought to characterize how the thematic content of the episode evolved over time. However, whereas dynamic topic models are designed to characterize how the properties of collections of documents change over time, our sliding-window approach enables us to examine the topic dynamics within a single document (or video). Specifically, our approach yielded (via the topic-proportions matrix) a single topic vector for each sliding window of annotations transformed by the topic model. We then stretched (interpolated) the resulting windows-by-topics matrix to match the time series of the 1,976 fMRI volumes collected as participants viewed the episode.

The 32 topics we found were heavily character focused (that is, the top-weighted word in each topic was nearly always a character) and could be roughly divided into themes centred around Sherlock Holmes (the titular character), John Watson (Sherlock's close confidant and assistant), supporting characters (for example, Inspector Lestrade, Sergeant Donovan or Sherlock's brother Mycroft), or the interactions between various groupings of these characters (Supplementary Fig. 2). This probably follows from the frequency with which these terms appeared in the episode annotations. Several of the identified 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 also sparse, in that only a small number of topics (typically one or two) tended to be active at any given timepoint (Fig. 2a). Further, the dynamics of the topic activations appeared to exhibit persistence (that is, given that a topic was active in one timepoint, it was likely to be active in the following timepoint) along with occasional rapid changes (that is, occasionally topic weights would change abruptly from one timepoint to the next). These two properties of the topic dynamics may be seen in the block-diagonal structure of the timepoint-by-timepoint correlation matrix (Fig. 2b) and reflect the gradual drift and sudden shifts fundamental to the temporal dynamics of many real-world experiences, as well as television episodes. Given this observation, we adapted an approach devised by Baldassano et al.[26], and used a HMM to identify the event boundaries where the topic activations changed rapidly (that is, the boundaries of the blocks in the temporal correlation matrix; event boundaries identified by the HMM are outlined in yellow in Fig. 2b). Part of our model-fitting procedure required selecting an appropriate number of events into which the topic trajectory should be segmented. To accomplish this, we used an optimization procedure that maximized the difference between the topic weights for timepoints within an event versus timepoints across multiple events (see Methods). We then created a stable summary of the content within each episode event by averaging the topic vectors across the timepoints spanned by each event (Fig. 2c).

Given that the time-varying content of the episode could be segmented cleanly into discrete events, we investigated whether participants' recalls of the episode also displayed a similar structure. We applied the same topic model (already trained on the episode annotations) to each participant's recalls. Analogously to the way in which we parsed the time-varying content of the episode, to obtain similar estimates for each participant's recall transcript, we treated each overlapping window of (up to) 10 sentences from their transcript as a document, and computed the most probable mix of topics reflected in each timepoint's sentences. This yielded, for each participant, a number-of-windows by number-of-topics topic-proportions matrix that characterized how the topics identified in the original episode were reflected in the participant's recalls. An important feature of our approach is that it allows us to compare participants' recalls to events from the original episode, despite the participants using widely varying language to describe the events, and their descriptions often diverging in content, quality and quantity from the episode annotations. This ability to match up conceptually related text that differs in specific vocabulary, detail and length is an important benefit of projecting the episode and recalls

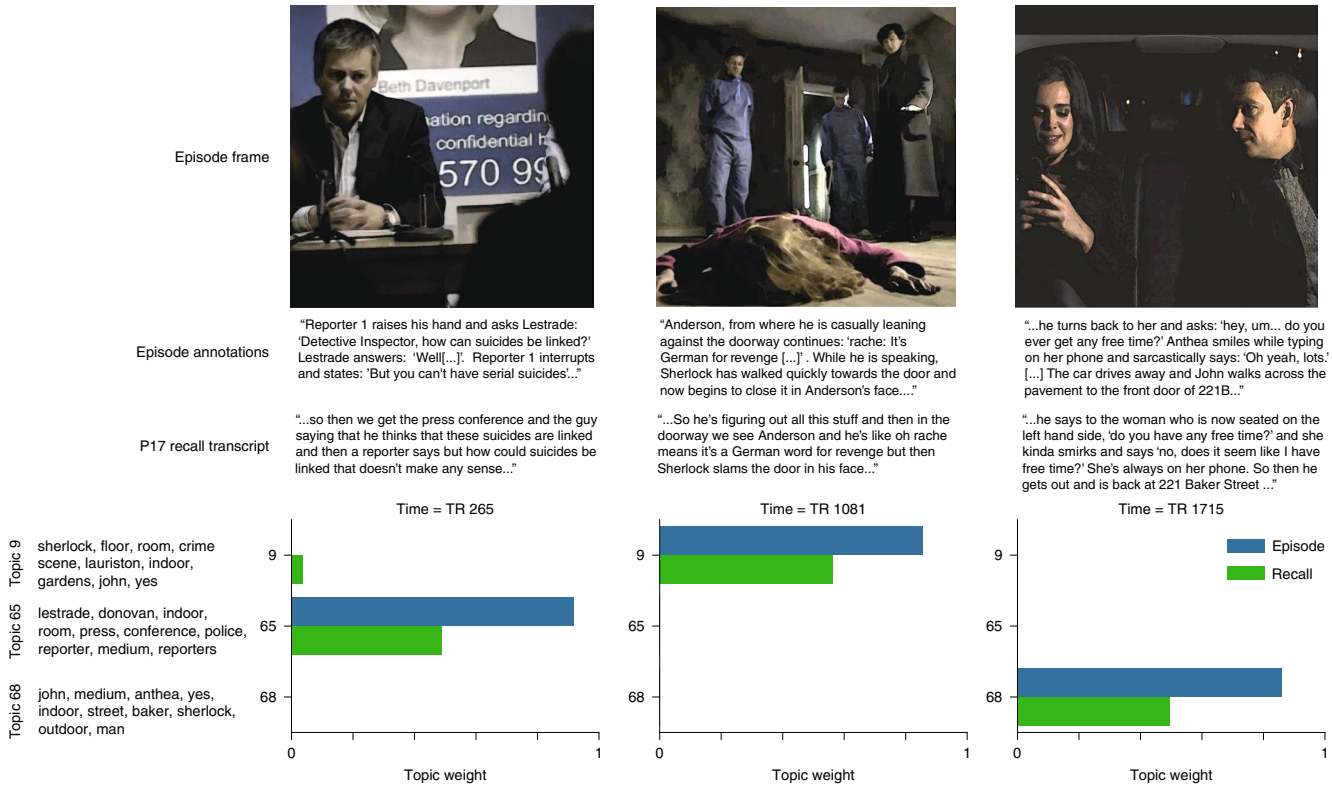


Fig. 1 | Topic weights in episode and recall content. We used detailed, hand-generated annotations describing each manually identified time segment from the episode to fit a topic model. Three example frames from the episode (first row) are displayed, along with their descriptions from the corresponding episode annotation (second row), an example participant's recall transcript (third row), and image repetition times (TR). We used the topic model (fit to the episode annotations) to estimate topic vectors for each moment of the episode and each sentence of participants' recalls. Example topic vectors are displayed in the bottom row (blue, episode annotations; green, example participant's recalls). Three topic dimensions are shown (the highest-weighted topics for each of the three example scenes), along with the ten highest-weighted words for each topic. Supplementary Fig. 2 provides a full list of the top 10 words from each of the discovered topics. Images are copyright of Hartswood Films Ltd.

into a shared topic space. An example topic-proportions matrix from one participant's recalls is shown in Fig. 2d.

Although the example participant's recall topic-proportions matrix shows some visual similarity to the episode topic-proportions matrix, the time-varying topic proportions for the example participant's recalls are not as sparse as those for the episode (compare Figs. 2a,d). Similarly, although there do appear to be periods of stability in the recall topic dynamics (that is, most topics are active or inactive over contiguous blocks of time), the changes in topic activations that define event boundaries appear less clearly delineated in participants' recalls than in the episode's annotations. To examine these patterns in detail, we computed the timepoint-by-timepoint correlation matrix for the example participant's recall topic-proportions matrix (Fig. 2e). As in the episode correlation matrix (Fig. 2b), the example participant's recall correlation matrix has a strong block-diagonal structure, indicating that their recalls are discretized into separated events. We used the same HMM-based optimization procedure that we had applied to the episode's topic-proportions matrix (Methods) to estimate an analogous set of event boundaries in the participant's recounting of the episode (outlined in yellow). We carried out this analysis on all 17 participants' recall topic-proportions matrices (Extended Data Fig. 2).

Two clear patterns emerged from this set of analyses. First, although every individual participant's recalls could be segmented into discrete events (that is, every individual participant's recall correlation matrix exhibited clear block-diagonal structure; Extended Data Fig. 2), each participant appeared to have a unique recall resolution, reflected in the sizes of those blocks. While some participants' recall topic proportions segmented into just a few events (for example, participant (P)4, P5 and P7), others' segmented into many shorter-duration events (for example, P12, P13 and P17). This suggests that different participants may be recalling the episode with different levels of detail—that is, some might recount only high-level essential plot details, whereas others might recount low-level details instead (or in addition). The second clear pattern present in every individual participant's recall correlation matrix was that, unlike in the episode correlation matrix, there were substantial off-diagonal correlations. One potential explanation for this finding is that the topic models, trained only on episode annotations, do not capture topic proportions in participants' held-out recalls as accurately. A second possibility is that, whereas each event in the original episode was (largely) separable from the others (Fig. 2b), in transforming those separable events into memory, participants appeared to be integrating across multiple events, blending elements of previously recalled and not-yet-recalled content into each newly recalled event[8,28,29] (Fig. 2e and Extended Data Fig. 2).

The above results demonstrate that topic models capture the dynamic conceptual content of the episode and participants' recalls of the episode. Further, the episode and recalls exhibit event boundaries that can be identified automatically using HMMs to segment the dynamic content. Next, we investigated whether some correspondence might be made between the specific content of the events the participants experienced while viewing the episode and the events they later recalled. We labelled each recall event as matching the episode event with the most similar (that is, most highly correlated) topic vector (Fig. 2g and Extended Data Fig. 3). This yielded

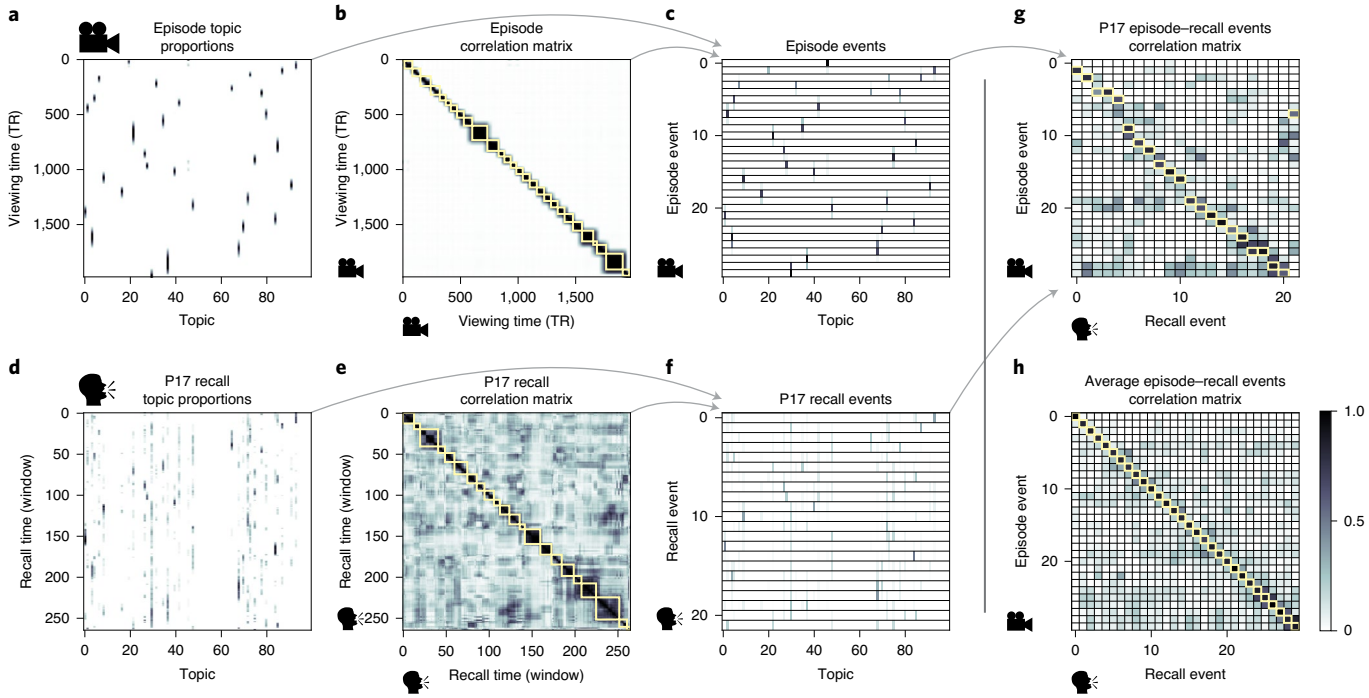


Fig. 2 | Modelling naturalistic stimuli and recalls. a–**h**, Darker colours indicate higher values; range: [0, 1]. **a**, Topic vectors ($K = 100$) for each of the 1,976 episode timepoints. **b**, Timepoint-by-timepoint correlation matrix of the topic vectors displayed in **a**. Event boundaries discovered by the HMM are denoted in yellow (30 events detected). **c**, Average topic vectors for each of the 30 episode events. **d**, Topic vectors for each of 265 sliding windows of sentences spoken by an example participant while recalling the episode. **e**, Timepoint-by-timepoint correlation matrix of the topic vectors displayed in **d**. Event boundaries detected by the HMM are denoted in yellow (22 events detected). Extended Data Fig. 2 shows similar plots for all participants. **f**, Average topic vectors for each of the 22 recall events from the example participant. **g**, Correlations between the topic vectors for every pair of episode events (**c**) and recall events (from the example participant in **f**). Extended Data Fig. 3 shows similar plots for all participants. **h**, Average correlations between each pair of episode events and recall events (across all 17 participants). To create the figure, each recalled event was assigned to the episode event with the most correlated topic vector (yellow boxes in **g**,**h**).

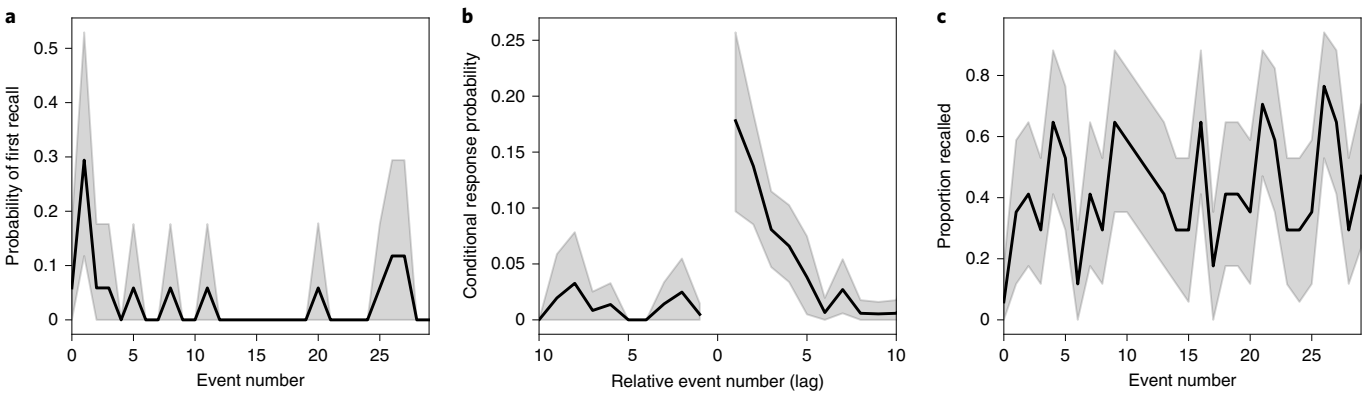


Fig. 3 | Naturalistic extensions of classic list-learning memory analyses. a, The probability of first recall as a function of the serial position of the event in the episode. **b**, The probability of recalling each event, conditioned on having most recently recalled the event that is lag events away in the episode. **c**, The proportion of participants who recalled each event as a function of the serial position of the events in the episode. Shaded regions denote the bootstrap-estimated 95% CI.

a sequence of presented events from the original episode, and a (potentially differently ordered) sequence of recalled events for each participant. Analogous to classic list-learning studies, we can then examine participants' recall sequences by asking which events they tended to recall first (probability of first recall[30–32]; Fig. 3a); how participants most often transitioned between recalls of the events as a function of the temporal distance between them (lag-conditional response probability[2]; Fig. 3b); and which events they were likely to remember overall (serial position recall analyses[1]; Fig. 3c). Some of the patterns we observed appeared to be similar to classic effects from the list-learning literature. For example, participants had a higher probability of initiating recall with early events (Fig. 3a) and a higher probability of transitioning to neighbouring events with an asymmetric forward bias (Fig. 3b). However, unlike typical observations in list-learning studies, we did not observe patterns comparable to the primacy or recency serial position effects (Fig. 3c). We

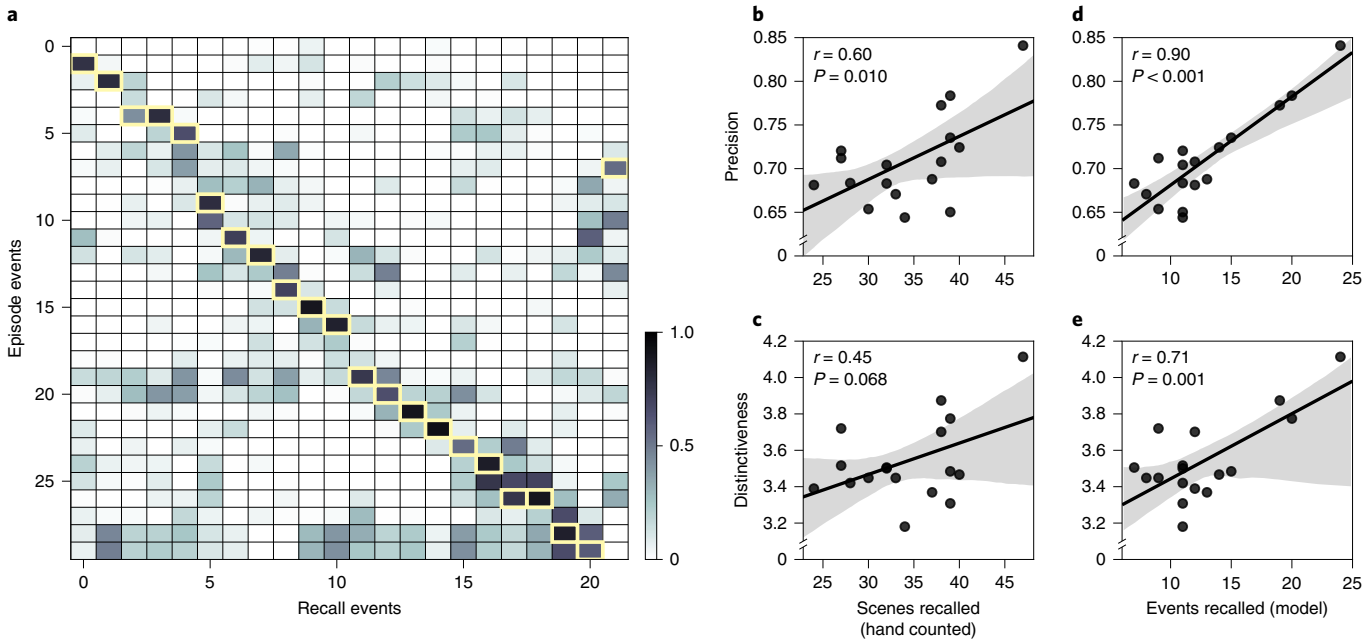


Fig. 4 | Novel content-based metrics of naturalistic memory: precision and distinctiveness. a, The episode-recall correlation matrix for an example participant (P17), chosen for their large number of recall events (for analogous figures for other participants, see Extended Data Fig. 2). The yellow boxes highlight the maximum correlation in each column. The example participant's overall precision score was computed as the average across the (Fisher *z*-transformed) correlation values in the yellow boxes. Their distinctiveness score was computed as the average (over recall events) of the *z*-scored (within-column) event precisions. **b**, The across-participants (Pearson's) correlation between precision and hand-counted number of recalled scenes. **c**, The correlation between distinctiveness and hand-counted number of recalled scenes. **d**, The correlation between precision and the number of recalled episode events, as determined by our model. **e**, The correlation between distinctiveness and the number of recalled episode events, as determined by our model.

hypothesized that participants might be leveraging meaningful narrative associations and references over long timescales throughout the episode.

Clustering scores are often used by memory researchers to characterize how people organize their memories of words on a studied list[33]. We defined analogous measures to characterize how participants organized their memories for episodic events (details in Methods). Temporal clustering refers to the extent to which participants group their recall responses according to encoding position. Overall, we found that sequentially viewed episode events tended to appear nearby in participants' recall-event sequences (clustering score 0.732 ± 0.033, mean $\pm$ s.e.m.). Participants with higher temporal clustering scores tended to exhibit better overall memory for the episode, according to both hand-counted numbers of recalled scenes from the episode reported by Chen et al.[23] (Pearson's $r(15) = 0.49$, $P = 0.046$, 95% confidence interval (CI) $= [0.25, 0.76]$) and the numbers of episode events that best matched at least one recall event (that is, model-estimated number of events recalled; Pearson's $r(15) = 0.59$, $P = 0.013$, 95% CI $= [0.31, 0.80]$). Semantic clustering measures the extent to which participants cluster their recall responses according to semantic similarity[34]. We found that participants tended to recall semantically similar episode events together (clustering score 0.650 ± 0.032), and that semantic clustering scores were also related to both hand-counted (Pearson's $r(15) = 0.65$, $P = 0.004$, 95% CI $= [0.31, 0.85]$) and model-estimated (Pearson's $r(15) = 0.58$, $P = 0.015$, 95% CI $= [0.10, 0.83]$) numbers of recalled events.

The above analyses illustrate how our framework for characterizing the dynamic conceptual content of naturalistic episodes enables us to carry out analyses that have traditionally been applied to much simpler list-learning paradigms. However, perhaps the most interesting aspects of memory for naturalistic episodes are those that have no list-learning analogues. The nuances of how one's memory for an event might capture some details, yet distort or neglect others, is central to how we use our memory systems in daily life. Yet, when researchers study memory in highly simplified paradigms, those nuances are not typically observable. We next developed two novel, continuous metrics, termed 'precision' and 'distinctiveness' aimed at characterizing distortions in the conceptual content of individual recall events, and the conceptual overlap between how people described different events.

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

Distinctiveness is intended to capture the specificity of recall. In other words, distinctiveness quantifies the extent to which a given recall event reflects the most similar episode event over and above other episode events. Intuitively, distinctiveness is like a normalized variant of our precision metric. Whereas precision measures only how much detail about an event was captured in someone's recall, distinctiveness penalizes details that also pertain to other episode events. We define the distinctiveness of an event's recall as its precision expressed in standard deviation units with respect to other episode events. Specifically, for a given recall event, we compute the correlation between its topic vector and that of each episode event. This yields a distribution of correlation coefficients (one per episode event). We subtract the mean and divide by the standard deviation of this distribution to obtain a *z*-score for each coefficient. The maximum value in this distribution (which, by definition, belongs

to the episode event that best matches the given recall event) is that recall event's distinctiveness score. In this way, recall events that match one episode event far better than all other episode events will receive a high distinctiveness score. By contrast, a recall event that matches all episode events roughly equally will receive a comparatively low distinctiveness score.

In addition to examining how precisely and distinctively participants recalled individual events, these metrics can also be used to summarize each participant's performance by averaging across a participant's event-wise precision or distinctiveness scores. This enables us to quantify how precisely a participant tended to recall subtle within-event details, as well as how specific (distinctive) those details were to individual events from the episode. Participants' average precision and distinctiveness scores were strongly correlated ($r(15) = 0.90$, $P < 0.001$, 95% CI = [0.66, 0.96]). This indicates that participants who tended to precisely recount low-level details of episode events also tended to do so in an event-specific way (for example, as opposed to detailing recurring themes that were present in most or all episode events; this behaviour would have resulted in high precision but low distinctiveness). We found that, across participants, higher precision scores were positively correlated with the numbers of both model-estimated events ($r(15) = 0.90$, $P < 0.001$, 95% CI = [0.54, 0.96]) and hand-annotated scenes ($r(15) = 0.60$, $P = 0.010$, 95% CI = [0.02, 0.83]) that participants recalled. Participants' average distinctiveness scores were also correlated with their numbers of model-estimated recalled events ($r(15) = 0.71$, $P = 0.001$, 95% CI = [−0.07, 0.90]) and marginally significantly correlated with their numbers of hand-annotated ($r(15) = 0.45$, $P = 0.068$, 95% CI = [−0.21, 0.79]).

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

Figure 5 also illustrates the differences between high and low distinctiveness scores. We extracted the set of sentences comprising the most distinctive (P9) and least distinctive (P6) recall events corresponding to the example episode event shown in Fig. 5c (event 21). We also extracted the annotations for all episode events whose content these participants' single recall events touched on. We assigned each episode event a unique colour (Fig. 5e), and coloured each recalled sentence (Fig. 5f) according to the episode events they best matched. Visual inspection of Fig. 5f reveals that the content of the most distinctive recall is tightly concentrated around event 21, whereas the least distinctive recall incorporates content from a much wider range of episode events.

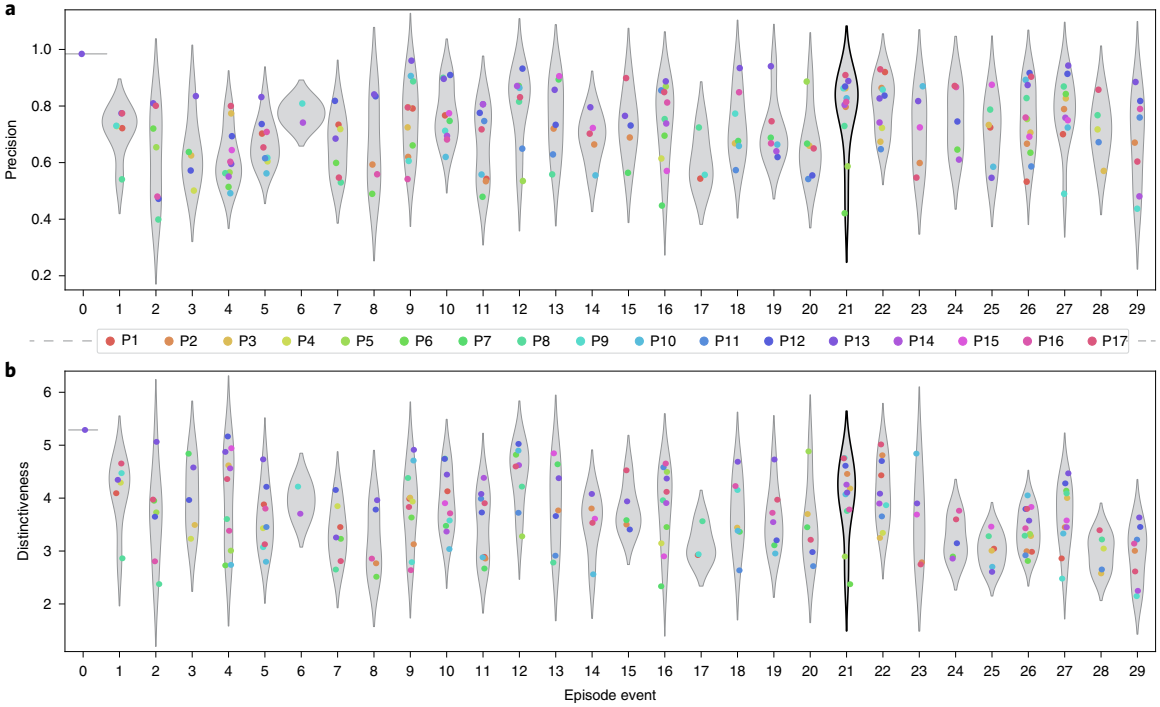
The preceding analyses sought to characterize how participants' recountings of individual episode events captured the low-level details of each event. Next, we sought to characterize how participants' recountings of the full episode captured its high-level essence—that is, the shape of the episode's trajectory through word-embedding (topic) space. To visualize the essence of the episode and each participant's recall trajectory[35], we projected the topic-proportions matrices for the episode and recalls onto a shared two-dimensional space using uniform manifold approximation and projection (UMAP)[36]. In this lower-dimensional space, each point represents a single episode or recall event, and the distances between the points reflect the distances between the events' associated topic vectors (Fig. 6). In other words, events that are nearer to each other in this space are more semantically similar, and those that are farther apart are are less so.

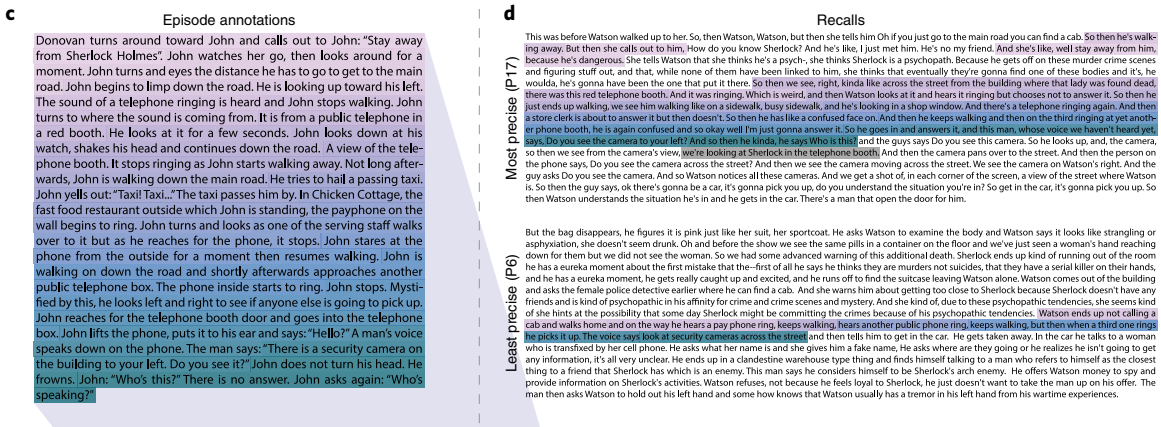
Visual inspection of the episode and recall topic trajectories reveals a striking pattern. First, the topic trajectory of the episode (which reflects its dynamic content; Fig. 6a) is captured nearly perfectly by the averaged topic trajectories of participants' recalls (Fig. 6b). To assess the consistency of these recall trajectories across participants, we asked: given that a participant's recall trajectory had entered a particular location in the reduced topic space, could the position of their next recalled event be predicted reliably? For each location in the reduced topic space, we computed the set of line segments connecting successively recalled events (across all participants) that intersected that location (see Methods and Extended Data Fig. 1). 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. 6b reflect significant peaks at $p < 0.05$, corrected). We observed that the locations traversed by nearly the entire episode trajectory exhibited such peaks. In other words, participants' recalls exhibited similar trajectories to each other that also matched the trajectory of the original episode (Fig. 6c). This is especially notable when considering the fact that the number of HMM-identified recall events (dots in Fig. 6c) varied considerably across people, and that every participant used different words to describe what they had remembered happening in the episode. Differences in the numbers of recall events appear in participants' trajectories as differences in the sampling resolution along the trajectory. We note that this framework also provides a means of disentangling classic proportion-recalled measures (that is, the proportion of episode events described in participants' recalls) from participants' abilities to recapitulate the episode's essence (that is, the similarity between the shapes of the original episode trajectory and that defined by each participant's recounting of the episode).

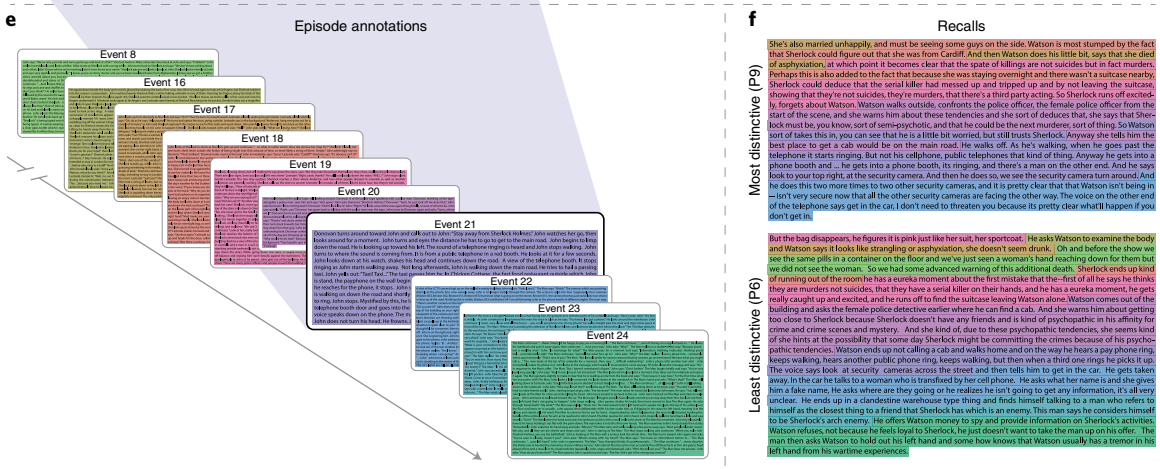
In addition to enabling us to visualize the episode's high-level essence, describing the episode as a geometric trajectory also

Fig. 5 | Precision reflects the completeness of recall, whereas distinctiveness reflects recall specificity. a, Recall precision by episode event. Grey violin plots display kernel density estimates for the distribution of recall precision scores for a single episode event. Coloured dots within each violin plot represent individual participants' recall precisions for the given event. **b**, Recall distinctiveness by episode event, analogous to **a**. **c**, The set of 'narrative details' episode annotations[23], comprising an example episode event (22) identified by the HMM. Each action or feature is highlighted in a different colour. **d**, Sentences comprising the most precise (P17) and least precise (P6) participants' recalls of episode event 21. Descriptions of specific actions or features reflecting those highlighted in **b** are highlighted in the corresponding colour. The text highlighted in grey denotes a (rare) false recall. The unhighlighted text denotes correctly recalled information about other episode events. **e**, The sets of 'narrative details' episode annotations[23] for scenes, comprising episode events described by the example participants in **f**. Each event's text is highlighted in a different colour. **f**, The sentences comprising the most distinctive (P9) and least distinctive (P6) participants' recalls of episode event 21. Sections of recall describing each episode event in **e** are highlighted with the corresponding colour.

enables us to drill down to individual words and quantify how each word relates to the memorability of each event. This provides another approach to examining participants' recall for low-level details beyond the precision and distinctiveness measures we defined above. The results displayed in Figs. 3c and 5a suggest that certain events were remembered better than others. Given this, we next asked whether the events that were generally remembered precisely or imprecisely tended to reflect particular content. Because our analysis framework projects the dynamic episode content and participants' recalls into a shared space, and because the dimensions of that space represent topics (which are, in turn, sets of weights over known words in the vocabulary), we are able to







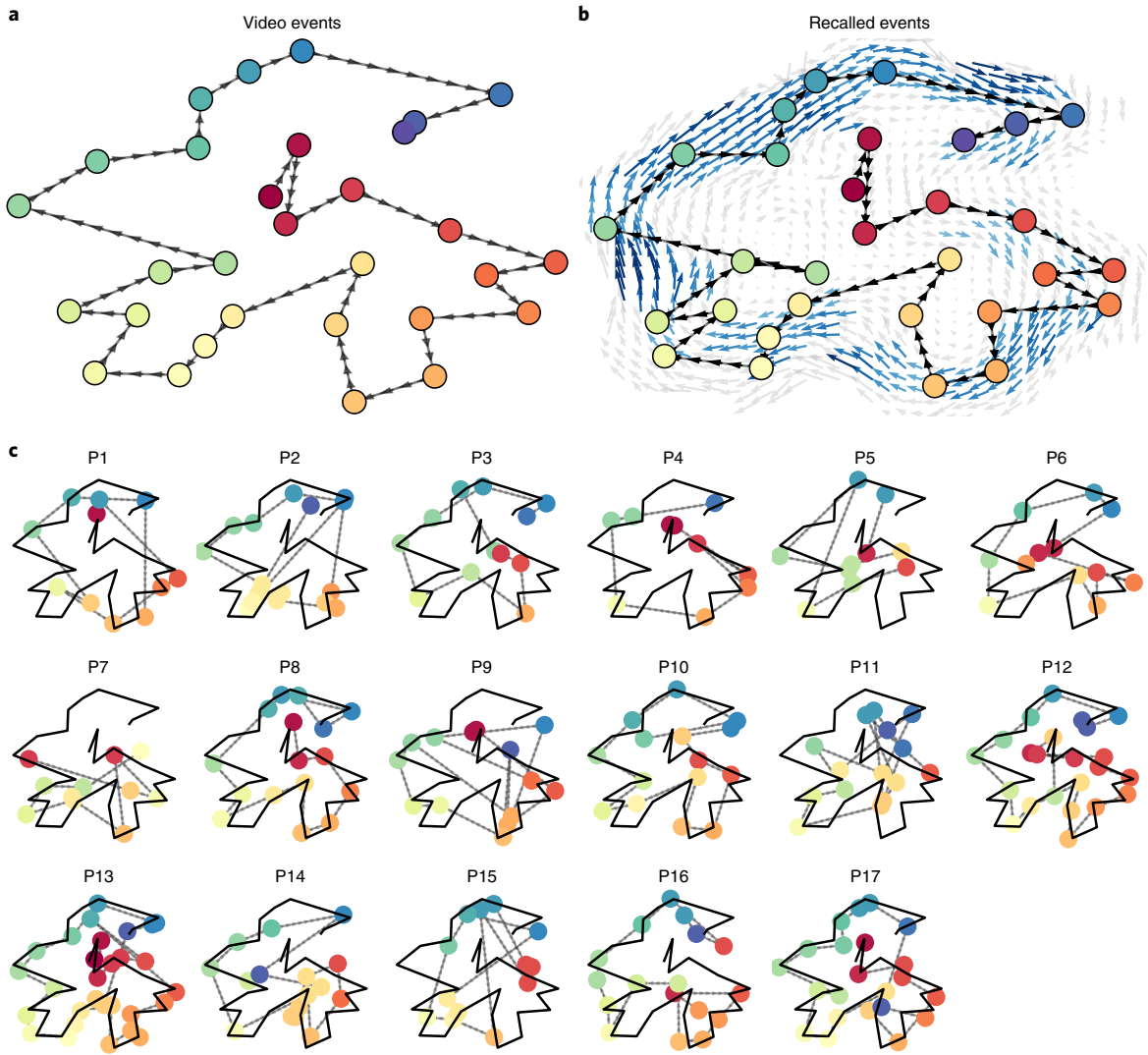


Fig. 6 | Trajectories through topic space capture the dynamic content of the episode and recalls. a–c, The topic-proportions matrices have been projected onto a shared two-dimensional space using UMAP. **a**, The two-dimensional topic trajectory taken by the episode of *Sherlock*. Each dot indicates an event identified using the HMM (see Methods). Dot colours 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 colouring as the trajectory in **a**. To compute the event positions, we matched each recalled event with an event from the original episode (see Results), and then we averaged the positions of all events with the same label. 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 as determined by a Rayleigh test ($P < 0.05$, corrected). Additional detail are provided in Methods and Extended Data Fig. 1. **c**, The recall topic trajectories (grey) taken by each individual participant (P1–P17). The episode's trajectory is shown in black for reference. Here, events (dots) are coloured by their matched episode event in **a**.

recover the weighted combination of words that make up any point (that is, topic vector) in this space. We first computed the average precision with which participants recalled each of the 30 episode events (Fig. 7a; note that this result is analogous to a serial position curve created from our precision metric). We then computed a weighted average of the topic vectors for each episode event, where the weights reflected how precisely each event was recalled. To visualize the result, we created a Wordle image (https://zenodo.org/record/1322068), in which words weighted more heavily by more precisely remembered topics appear in a larger font (Fig. 7b, green box). Across the full episode, content that weighted heavily on topics and words central to the major foci of the episode (for example, the names of the two main characters, 'Sherlock' and 'John', and the address of a major recurring location, '221B Baker Street') was best remembered. An analogous analysis revealed which themes were less precisely remembered. Here, in computing the weighted average over events' topic vectors, we weighted each event in inverse proportion to its average precision (Fig. 7b, red box). The least precisely remembered episode content reflected information that was extraneous to the episode's essence, such as the proper names of relatively minor characters (for example, 'Mike', 'Molly' and 'Lestrade') and locations (for example, 'St Bartholomew's Hospital').

A similar result emerged from assessing the topic vectors for individual episode and recall events (Fig. 7c). Here, for each of the three most and least precisely remembered episode events, we have constructed two Wordles: one from the topic vector for the original episode event (left) and a second from the average recall topic vector for that event (right). The three most precisely remembered events (circled in green) correspond to scenes integral to the central plotline: a mysterious figure spying on John in a phone booth;

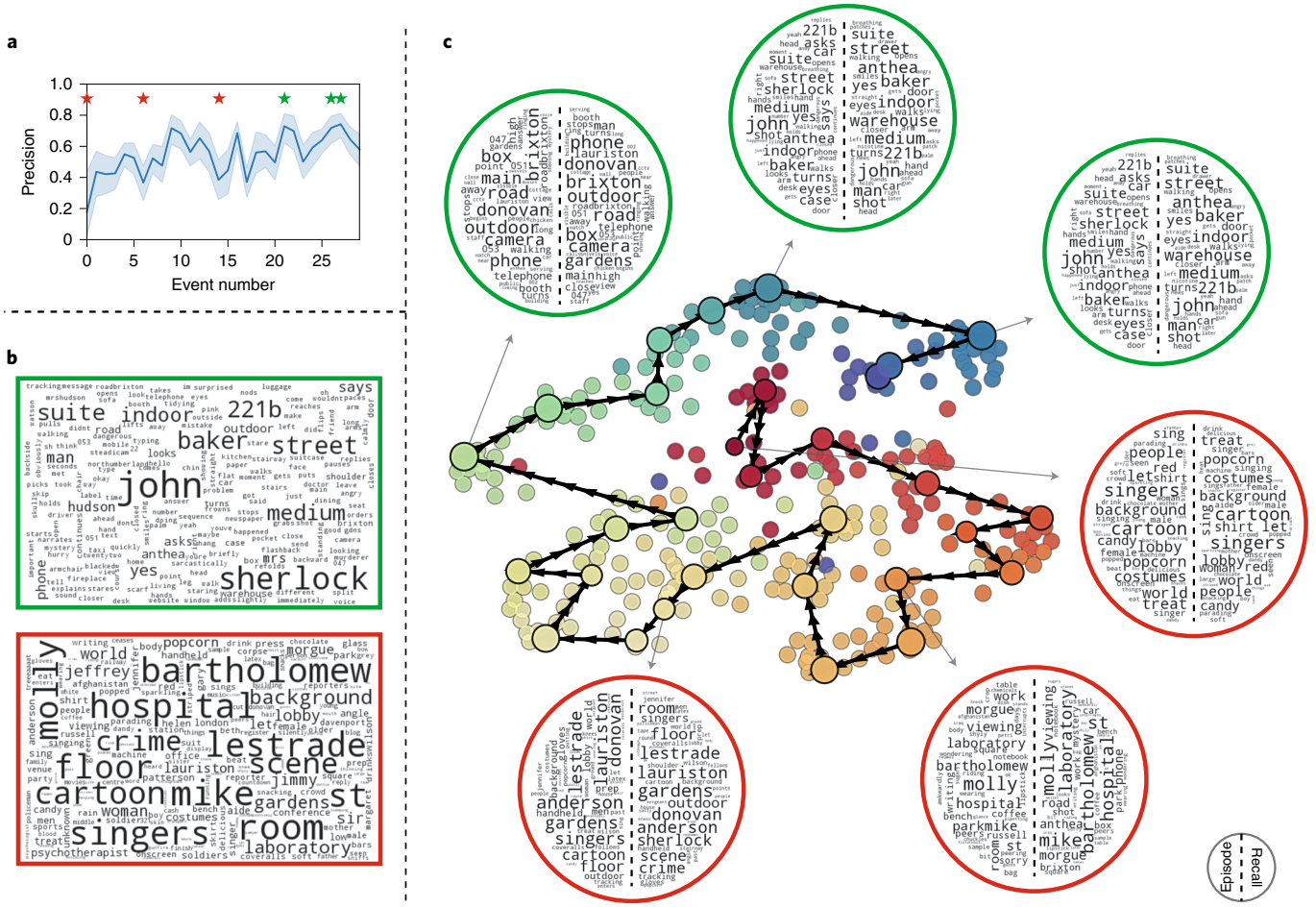


Fig. 7 | Language used in the most and least precisely remembered events. a, Average precision (episode event–recall event topic vector correlation) across participants for each episode event. Here we defined each episode event's precision for each participant as the correlation between its topic vector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

John meeting Sherlock at Baker St. to discuss the murders; and Sherlock laying a trap to catch the killer. Meanwhile, the least precisely remembered events (circled in red) reflect scenes that comprise minor plot points: a video of singing cartoon characters that participants viewed in an introductory clip before the main episode; John asking Molly about Sherlock's habit of over-analysing people; and Sherlock noticing evidence of Anderson's and Donovan's affair.

The results thus far inform us about which aspects of the dynamic content in the episode participants watched were preserved or altered in participants' memories. We next carried out a series of analyses aimed at understanding which brain structures might facilitate these preservations and transformations between the participants' shared experience of watching the episode and their subsequent memories of the episode. In the first analysis, we sought to identify brain structures that were sensitive to the dynamic unfolding of the episode's content, as characterized by its topic trajectory. We used a searchlight procedure to identify clusters of voxels whose activity patterns displayed a proximal temporal correlation structure (as participants watched the episode) matching that of the original episode's topic proportions (Fig. 8a; see Methods for additional details). In a second analysis, we sought to identify brain structures whose responses (during episode viewing) reflected how each participant would later structure their recounting of the episode. We used a searchlight procedure to identify clusters of voxels whose proximal temporal correlation matrices matched that of the topic-proportions matrix for each participant's recall transcript (Figs. 8b; see Methods for additional details). To ensure our searchlight procedure identified regions specifically sensitive to the temporal structure of the episode or recalls (that is, rather than those with a temporal autocorrelation length similar to that of the episode and recalls), we performed a phase shift-based permutation correction (see Methods). As shown in Fig. 8c, the episode-driven searchlight analysis revealed a distributed network of regions that may play a role in processing information relevant to the

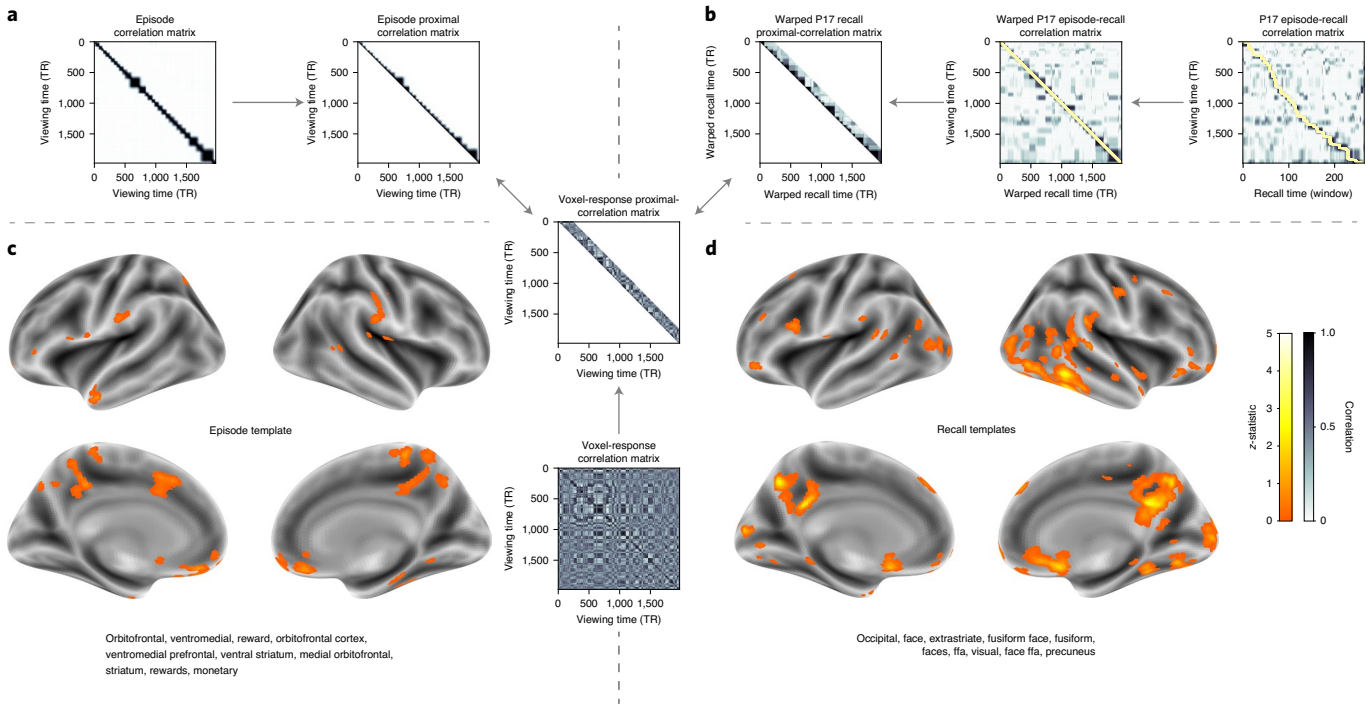


Fig. 8 | Brain structures that underlie the transformation of experience into memory. a, We isolated the proximal diagonals from the upper triangle of the episode correlation matrix and applied this same diagonal mask to the voxel response correlation matrix for each cube of voxels in the brain. We then searched for brain regions whose activation time series consistently exhibited a similar proximal correlational structure to the episode model, across participants. **b**, We used dynamic time warping[62] to align each participant's recall time series to the TR time series of the episode. We then computed the temporal correlation matrix of each participant's warped recalls. Next, we applied the same diagonal mask used in **a** to isolate the proximal temporal correlations and searched for brain regions whose activation time series for each participant consistently exhibited a similar proximal correlational structure to that participant's recalls. **c**, We identified a network of regions sensitive to the narrative structure of participants' ongoing experience. The map shown is thresholded at $P < 0.05$, corrected. The top ten Neurosynth terms displayed in the panel were computed using the unthresholded map. **d**, We also identified a network of regions sensitive to how individuals would later structure the episode's content in their recalls. The map shown is thresholded at $P < 0.05$, corrected. The top ten Neurosynth terms displayed in the panel were computed using the unthresholded map.

narrative structure of the episode. The recall-driven searchlight analysis revealed a second network of regions (Fig. 8d) that may facilitate a person-specific transformation of one's experience into memory. In identifying regions whose responses to ongoing experiences reflect how those experiences will be remembered later, this latter analysis extends classic subsequent memory effect analyses[37] to the domain of naturalistic experiences.

The searchlight analyses described above yielded two distributed networks of brain regions whose activity time courses tracked with the temporal structure of the episode (Fig. 8c) or participants' subsequent recalls (Fig. 8d). We next sought to gain greater insight into the structures and functional networks our results reflected. To accomplish this, we performed an additional, exploratory analysis using Neurosynth[38]. Given an arbitrary statistical map as input, Neurosynth performs a massive automated meta-analysis, returning a frequency-ranked list of terms used in neuroimaging papers that report similar statistical maps. We ran Neurosynth on the (unthresholded) permutation-corrected maps for the episode- and recall-driven searchlight analyses. The top ten terms with maximally similar meta-analysis images identified by Neurosynth are shown in Fig. 8.

Discussion

Explicitly modelling the dynamic content of a naturalistic stimulus and participants' memories enabled us to connect the present study of naturalistic recall with an extensive previous literature that has used list-learning paradigms to study memory[4], as in Fig. 3. We found some similarities between how participants in the present study recounted a television episode and how participants typically recall memorized random word lists. However, our broader claim is that word lists miss out on fundamental aspects of naturalistic memory that are more like the sort of memory we rely on in everyday life. For example, there are no random word-list analogues of character interactions, conceptual dependencies between temporally distant episode events, the sense of solving a mystery that pervades the *Sherlock* episode, or the myriad other features of the episode that convey deep meaning and capture interest. Nevertheless, each of these properties affects how people process and engage with the episode as they are watching it and how they remember it later. The overarching goal of the present study is to characterize how the rich dynamics of the episode affect the rich behavioural and neural dynamics of how people remember it.

Our work casts remembering as reproducing (behaviourally and neurally) the topic trajectory or shape of an experience, thereby drawing implicit analogies between mentally navigating through word-embedding spaces and physically navigating through spatial environments[39–41]. When we characterized memory for a television episode using this framework, we found that every participant's recounting of the episode recapitulated the low spatial frequency details of the shape of its trajectory through topic space (Fig. 6). We termed this narrative scaffolding the episode's essence. Where participants' behaviours varied most was in their tendencies to recount specific low-level details from each episode event. Geometrically, this appears as high spatial frequency distortions in participants' recall trajectories relative to the trajectory of the original episode (Fig. 7). We developed metrics to characterize the precision (recovery

of any and all event-level information) and distinctiveness (recovery of event-specific information). We also used word cloud visualizations to interpret the details of these event-level distortions.

The neural analyses we carried out (Fig. 8) also leveraged our geometric framework for characterizing the shapes of the episode and participants' recountings. We identified one network of regions whose responses tracked with temporal correlations in the conceptual content of the episode (as quantified by topic models applied to a set of annotations about the episode). This network included orbitofrontal cortex, ventromedial prefrontal cortex, and striatum, among others. As reviewed by Ranganath and Ritchey[13], several of these regions are members of the anterior temporal system, which has been implicated in assessing and processing the familiarity of ongoing experiences, emotions, social cognition and reward. A second network that we identified tracked with temporal correlations in the idiosyncratic conceptual content of participants' subsequent recountings of the episode. This network included occipital cortex, extrastriate cortex, fusiform gyrus, and the precuneus. Several of these regions are members of the posterior medial system[13], which has been implicated in matching incoming cues about the current situation to internally maintained situation models that specify the parameters and expectations inherent to the current situation[14,15]. Together, our results support the notion that these two (partially overlapping) networks work in coordination to make sense of our ongoing experiences, distort them in a way that links them with our prior knowledge and experiences, and encodes those distorted representations into memory for our later use. Our work also provides a potential framework for modelling and elucidating memory schemas—that is, cognitive abstractions that may be applied to multiple related experiences[42,43]. For example, the event-level geometric scaffolding of an experience (for example, Fig. 6a) might reflect its underlying schema, and experiences that share similar schemas might have similar shapes. This could also help explain how brain structures including the ventromedial prefrontal cortex[42] (Fig. 8) might acquire or apply schema knowledge across different experiences (that is, by learning patterns in the schema's shape).

Our general approach draws inspiration from previous work aimed at elucidating the neural and behavioural underpinnings of how we process dynamic naturalistic experiences and remember them later. Our approach to identifying neural responses to naturalistic stimuli (including experiences) entails building an explicit model of the stimulus dynamics and searching for brain regions whose responses are consistent with the model[44,45]. Building an explicit model of these dynamics also enables us to match up different people's recountings of a common shared experience, despite individual differences[46]. In previous work, a series of studies from Uri Hasson's group[7,23,26,47,48] has presented a clever alternative approach: rather than building an explicit stimulus model, these studies instead search for brain responses to the stimulus that are reliably similar across individuals. Inter-subject correlation and inter-subject functional connectivity analyses effectively treat other people's brain responses to the stimulus as a model of how its features change over time[49]. These purely brain-driven approaches are well suited to identifying which brain structures exhibit similar stimulus-driven responses across individuals. Further, because neural response dynamics are observed data (rather than model approximations), such approaches do not require a detailed understanding of which stimulus properties or features might be driving the observed responses. However, this also means that the specific stimulus features driving those responses are typically opaque to the researcher. Our approach is complementary. By explicitly modelling the stimulus dynamics, we are able to relate specific stimulus features to behavioural and neural dynamics. However, when our model fails to accurately capture the stimulus dynamics that are truly driving behavioural and neural responses, our approach necessarily yields an incomplete characterization of the neural basis of the processes we are studying.

Other recent work has used HMMs to discover latent event structure in neural responses to naturalistic stimuli[26]. By applying HMMs to our explicit models of stimulus and memory dynamics, we gain a more direct understanding of those state dynamics. For example, we found that although the events comprising each participant's recalls recapitulated the episode's essence, participants differed in the resolution of their recounting of low-level details. In turn, these individual behavioural differences were reflected in differences in neural activity dynamics as participants watched the television episode.

Our approach also draws inspiration from the growing field of word-embedding models. The topic models[24] we used to embed text from the episode annotations and participants' recall transcripts are just one of many models that have been studied in an extensive literature. The earliest approaches to word embedding, including latent semantic analysis[50], used word co-occurrence statistics (that is, how often pairs of words occur in the same documents contained in the corpus) to derive a unique feature vector for each word. The feature vectors are constructed so that words that co-occur more frequently have feature vectors that are closer (in Euclidean distance). Topic models are essentially an extension of those early models, in that they attempt to explicitly model the underlying causes of word co-occurrences by automatically identifying the set of themes or topics reflected across the documents in the corpus. More recent work on these types of semantic models, including word2vec[51], the Universal Sentence Encoder[52] and generative pre-trained transformers (for example, GPT-2[53] and GTP-3[54]) use deep neural networks to attempt to identify the deeper conceptual representations underlying each word. Despite the growing popularity of these sophisticated deep learning-based embedding models, we chose to prioritize interpretability of the embedding dimensions (for example, Fig. 7) over raw performance (for example, with respect to some predefined benchmark). Nevertheless, we note that our general framework is, in principle, robust to the specific choice of language model as well as other aspects of our computational pipeline. For example, the word-embedding model, time series segmentation model and the episode-recall matching function could each be customized to suit a particular question space or application. Indeed, for some questions, interpretability of the embeddings may not be a priority, and thus other text embedding approaches (including the deep learning-based models described above) may be preferable. Further work will be needed to explore the influence of particular models on our framework's predictions and performance.

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

Methods

Paradigm and data collection. Data were collected by Chen et al.[23]. In brief, participants ($n = 22$) viewed the first 48 min of 'A Study in Pink', the first episode of the BBC television show *Sherlock*, while fMRI volumes were collected (TR = 1,500 ms). Participants were pre-screened to ensure they had never seen any episode of the show before. The stimulus was divided into a 23 min (946 TR) and a 25 min (1030 TR) segment to mitigate technical issues related to the scanner. After finishing the clip, participants were instructed to "describe what they recalled of the [episode] in as much detail as they could, to try to recount events in the original order they were viewed in, and to speak for at least 10 min if possible, but that longer was better. They were told that completeness and detail were more important than temporal order, and that if at any point they realized they had missed something, to return to it. Participants were then allowed to speak for as long as they wished, and verbally indicated when they were finished (e.g., 'I'm done')"[23]. Five participants were dropped from the original dataset due to excessive head motion (2 participants), insufficient recall length (2 participants) or falling asleep during stimulus viewing (1 participant), resulting in a final sample size of $n = 17$. For additional details about the testing procedures and scanning parameters, see ref. [23]. The testing protocol was approved by Princeton University's Institutional Review Board.

After preprocessing the fMRI data and warping the images into a standard (3 mm^3 MNI) space, the voxel activations were given z-scores (within voxel) and spatially smoothed using a 6 mm (full width at half maximum) Gaussian kernel. The fMRI data were also cropped so that all episode-viewing data were aligned across participants. This included a constant 3 TR (4.5 s) shift to account for the lag in the haemodynamic response. All of these preprocessing steps followed Chen et al., where additional details may be found[23].

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

Statistics. All statistical tests performed in the behavioural analyses were two-sided. All statistical tests performed in the neural data analyses were two-sided, except for the permutation-based thresholding, which was one-sided. In this case, we were specifically interested in identifying voxels whose activation time series reflected the temporal structure of the episode and recall topic-proportions matrices to a greater extent than that of the phase-shifted matrices. The 95% confidence intervals we reported for each correlation were estimated by generating 10,000 bootstrap distributions of correlation coefficients by sampling (with replacement) from the observed data.

Modelling the dynamic content of the episode and recall transcripts. *Topic modelling.* The input to the topic model that we trained to characterize the dynamic content of the episode comprised 998 hand-generated annotations of short (mean 2.96s) time segments spanning the video clip (Chen et al. generated 1,000 annotations in total[23]; we removed two annotations referring to a break between the first and second scan sessions, during which no fMRI data were collected). We concatenated the text for all of the annotated features within each segment, creating a 'bag of words' describing its content, and performed some minor preprocessing (for example, stemming possessive nouns and removing punctuation). We then reorganized the text descriptions into overlapping sliding windows spanning (up to) 50 annotations each. In other words, we estimated the context for each annotated segment using the text descriptions of the preceding 25 annotations, the present annotations, and the following 24 annotations. To model the context for annotations near the beginning of the episode (that is, within 25 of the beginning or end), we created overlapping sliding windows that grew in size from one annotation to the full length. We also tapered the sliding-window lengths at the end of the episode, whereby time segments within fewer than 24 annotations of the end of the episode were assigned sliding windows that extended to the end of the episode. This procedure ensured that each annotation's content was represented in the text corpus an equal number of times.

We trained our model using these overlapping text samples with scikit-learn v.0.19.1[55], called from our high-dimensional visualization and text analysis software, HyperTools[35]. Specifically, we used the CountVectorizer class to transform the text from each window into a vector of word counts (using the union of all words across all annotations as the vocabulary, excluding English stop words); this yielded a number-of-windows by number-of-words word-count matrix. We then used the LatentDirichletAllocation class (topics = 100, method = 'batch') to fit a topic model[24] to the word-count matrix, yielding a number-of-windows (1,047) by number-of-topics (100) topic-proportions matrix. The topic-proportions matrix describes the gradually evolving mix of topics (latent themes) present in each annotated time segment of the episode. Next, we transformed the topic-proportions matrix to match the 1,976 fMRI volume acquisition times. We assigned each topic vector to the timepoint (in seconds) midway between the beginning of the first annotation and the end of the last annotation in its corresponding sliding text window. By doing so, we warped the linear temporal distance between consecutive topic vectors to align with the inconsistent temporal distance between consecutive annotations (whose durations varied greatly). We then rescaled these timepoints to 1.5 s TR units, and used linear interpolation to estimate a topic vector for each TR. This resulted in a number-of-TRs (1,976) by number-of-topics (100) matrix.

We created similar topic-proportions matrices using hand-annotated transcripts of each participant's verbal recall of the episode[23]. We tokenized the transcript into a list of sentences, and then reorganized the list into overlapping sliding windows spanning (up to) 10 sentences each, analogously to how we parsed the episode annotations. In turn, we transformed each window's sentences into a word-count vector (using the same vocabulary as for the episode model), then used the topic model already trained on the episode scenes to compute the most probable topic proportions for each sliding window. This yielded a number-of-windows (range 83–312) by number-of-topics (100) topic-proportions matrix for each participant. These reflected the dynamic content of each participant's recalls. For details on how we selected the episode and recall window lengths and number of topics, see Supplementary Information and Supplementary Fig. 1.

Segmenting topic-proportions matrices into discrete events using HMMs. We parsed the topic-proportions matrices of the episode and participants' recalls into discrete events using HMMs[25]. Given the topic-proportions matrix (describing the mix of topics at each timepoint) and a number of states K, an HMM recovers the set of state transitions that segments the time series into K discrete states. Following Baldassano et al.[26], we imposed an additional set of constraints on the discovered state transitions that ensured that each state was encountered exactly once (that is, never repeated). We used the BrainIAK toolbox (https://doi.org/10.5281/zenodo.59780) to implement this segmentation.

We used an optimization procedure to select the appropriate K for each topic-proportions matrix. Previous studies on narrative structure and processing have shown that we both perceive and internally represent the world around us at multiple, hierarchical timescales[7,23,26,43,56,57]. However, for the purposes of our framework, we sought to identify the single time series of event representations that was emphasized most heavily in the temporal structure of the episode and of each participant's recall. We quantified this as the set of K states that maximized the similarity between topic vectors for timepoints comprising each state, while minimizing the similarity between topic vectors for timepoints across different states. Specifically, we computed (for each matrix)

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

where a was the distribution of within-state topic vector correlations, and b was the distribution of across-state topic vector correlations. We computed the first Wasserstein distance (W_1, also known as Earth mover's distance)[58,59] between these distributions for a large range of possible K values (range [2, 50]), and selected the K that yielded the maximum value. Figure 2b displays the event boundaries returned for the episode, and Extended Data Fig. 2 displays the event boundaries returned for each participant's recalls. See Extended Data Fig. 4 for the optimization functions for the episode and recalls. After obtaining these event boundaries, we created stable estimates of the content represented in each event by averaging the topic vectors across timepoints between each pair of event boundaries. This yielded a number-of-events by number-of-topics matrix for the episode and recalls from each participant.

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

Perhaps the simplest and most widely used measure of memory performance is accuracy—that is, the proportion of studied (experienced) items (in this case, episode events) that the participant later remembered. Chen et al.[23] used this method to rate each participant's memory quality by computing the proportion of (50 manually identified) scenes mentioned in their recall. We found a strong across-participants correlation between these independent ratings and the proportion of 30 HMM-identified episode events matched to participants' recalls (Pearson's $r(15) = 0.71$, $P = 0.002$, 95% CI = [0.39, 0.88]). We further considered a number of more nuanced memory performance measures that are typically associated with list-learning studies. We also provide a software package, Quail, for carrying out these analyses[60].

Probability of first recall. Probability of first recall curves[30–32] reflect the probability that an item will be recalled first, as a function of its serial position

during encoding. To carry out this analysis, we initialized a number-of-participants (17) by number-of-episode-events (30) matrix of zeros. Then, for each participant, we found the index of the episode event that was recalled first (that is, the episode event whose topic vector was most strongly correlated with that of the first recall event) and filled in that index in the matrix with a 1. Finally, we averaged over the rows of the matrix, resulting in a 1 by 30 array representing the proportion of participants that recalled an event first, as a function of the order of the event's appearance in the episode (Fig. 3a).

Lag-conditional probability curve. The lag-conditional probability (lag-CRP) curve[2] reflects the probability of recalling a given item after the just-recalled item, as a function of their relative encoding positions (lag). In other words, a lag of 1 indicates that a recalled item was presented immediately after the previously recalled item, and a lag of −3 indicates that a recalled item came 3 items before the previously recalled item. For each recall transition (following the first recall), we computed the lag between the current recall event and the next recall event, normalizing by the total number of possible transitions. This yielded a number-of-participants (17) by number-of-lags (−29 to +29; 58 lags in total excluding lags of 0) matrix. We calculated the average over the rows of this matrix to obtain a group-averaged lag-CRP curve (Fig. 3b).

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

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

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

Averaging correlations. In all instances where we performed statistical tests involving precision or distinctiveness scores (Fig. 5), we used the Fisher z-transformation[61] to stabilize the variance across the distribution of correlation values before performing the test. Similarly, when averaging precision or distinctiveness scores, we used the z-transform of the scores to compute the mean, and inverse z-transformed the result.

Visualizing the episode and recall topic trajectories. We used the UMAP algorithm[36] to project the 100-dimensional topic space onto a two-dimensional space for visualization (Figs. 6 and 7). To ensure that all of the trajectories were projected onto the same lower-dimensional space, we computed the low-dimensional embedding on a stacked matrix created by vertically concatenating the events-by-topics topic-proportions matrices for the episode, the across-participants average recalls and all 17 individual participants' recalls. We then separated the rows of the result (a total number of events by two matrix) back into individual matrices for the episode topic trajectory, the across-participant average recall trajectory, and the trajectories for each individual participant's recalls (Fig. 6). This general approach for discovering a shared low-dimensional embedding for a collections of high-dimensional observations follows our previous work on manifold learning[35].

We optimized the manifold space for visualization on the basis of two criteria: first, that the 2D embedding of the episode trajectory should reflect its original 100-dimensional structure as faithfully as possible; and second, that the path traversed by the embedded episode trajectory should intersect itself a minimal number of times. The first criteria helps bolster the validity of visual intuitions about relationships between sections of episode content, based on their locations in the embedding space. The second criteria was motivated by the observed low off-diagonal values in the episode trajectory's temporal correlation matrix (suggesting that the same topic space coordinates should not be revisited; see Fig. 2a). For further details on how we created this low-dimensional embedding space, see Supplementary Information.

Estimating the consistency of flow through topic space across participants. In Fig. 6b, we present an analysis aimed at characterizing locations in topic space that different participants move through in a consistent way (via their recall topic trajectories; also see Extended Data Fig. 1). The two-dimensional topic space used in our visualizations (Fig. 6) comprised a 60×60 (arbitrary units) square. We tiled this space with a 50×50 grid of evenly spaced vertices, and defined a circular area centred on each vertex, whose radius was two times the distance between adjacent vertices (that is, 2.4 units). For each vertex, we examined the set of line segments formed by connecting each pair successively recalled events, across all participants, that passed through this circle. We computed the distribution of angles formed by those segments and the x-axis, and used a Rayleigh test to determine whether the distribution of angles was reliably peaked (that is, consistent across all transitions that passed through that local portion of topic space). To create Fig. 6b, we drew an arrow originating from each grid vertex, pointing in the direction of the average angle formed by the line segments that passed within 2.4 units. We set the arrow lengths to be inversely proportional to the P values of the Rayleigh tests at each vertex. Specifically, for each vertex we converted all of the angles of segments that passed within 2.4 units to unit vectors, and we set the arrow lengths at each vertex proportional to the length of the (circular) mean vector. We also indicated any significant results ($P<0.05$, corrected using the Benjamini–Hochberg procedure) by colouring the arrows in blue (darker blue denotes a lower P value, that is, a longer mean vector); all tests with $P\geq0.05$ are displayed in grey and given a lower opacity value.

Searchlight fMRI analyses. In Fig. 8, we present two analyses aimed at identifying brain regions whose responses (as participants viewed the episode) exhibited a particular temporal structure. We developed a searchlight analysis wherein we constructed a $5\times5\times5$ cube of voxels centred on each voxel in the brain[23], and for each of these cubes, computed the temporal correlation matrix of the voxel responses during episode viewing. Specifically, for each of the 1,976 volumes collected during episode viewing, we correlated the activity patterns in the given cube with the activity patterns (in the same cube) collected during every other timepoint. This yielded a $1{,}976\times1{,}976$ correlation matrix for each cube. Note: the scan of participant 5 ended 75 s early, and in the publicly released dataset for Chen et al.[23], their scan data was zero-padded to match the length of those of the other participants. For our searchlight analyses, we removed this padded data (that is, the last 50 TRs), resulting in a $1{,}925\times1{,}925$ correlation matrix for each cube in the brain of participant 5.

Next, we constructed a series of template matrices. The first template reflected the time course of the episode's topic-proportions matrix, and the others reflected the time course of each participant's recall topic-proportions matrix. To construct the episode template, we computed the correlations between the topic proportions estimated for every pair of TRs (before segmenting the topic-proportions matrices into discrete events; that is, the correlation matrix shown in Figs. 2b and 8a). We constructed similar temporal correlation matrices for each participant's recall topic-proportions matrix (Fig. 2d and Extended Data Fig. 2). However, to correct for length differences and potential non-linear transformations between viewing time and recall time, we first used dynamic time warping[62] to temporally align participants' recall topic-proportions matrices with the episode topic-proportions matrix. An example correlation matrix before and after warping is shown in Fig. 8b. This yielded a $1{,}976\times1{,}976$ correlation matrix for the episode template and for each participant's recall template.

The temporal structure of the episode's content (as described by our model) is captured in the block-diagonal structure of the episode's temporal correlation matrix (for example, Figs. 2b and 8a), with time periods of thematic stability represented as dark blocks of varying sizes. Inspecting the episode correlation matrix suggests that the episode's semantic content is highly temporally specific (that is, the correlations between topic vectors from distant timepoints are almost all near zero). By contrast, the activity patterns of individual (cubes of) voxels can encode relatively limited information on their own, and their activity frequently contributes to multiple separate functions[63–66]. By nature, these two attributes give rise to similarities in activity across large timescales that may not necessarily reflect a single task. To identify brain regions whose shifts in activity patterns mirrored shifts in the semantic content of the episode or recalls, we restricted the temporal correlations we considered to the timescale of semantic information captured by our model. Specifically, we isolated the upper triangle of the episode correlation matrix and created a proximal correlation mask that included only diagonals from the upper triangle of the episode correlation matrix up to the first diagonal that contained no positive correlations. Applying this mask to the full episode correlation matrix was equivalent to excluding diagonals beyond the corner of the

largest diagonal block. In other words, the timescale of temporal correlations we considered corresponded to the longest period of thematic stability in the episode, and by extension the longest period of thematic stability in participants' recalls and the longest period of stability we might expect to see in voxel activity arising from processing or encoding episode content. Figure 8 shows this proximal correlation mask applied to the temporal correlation matrices for the episode, an example participant's (warped) recall, and an example cube of voxels from our searchlight analyses.

To determine which (cubes of) voxel responses matched the episode template, we correlated the proximal diagonals from the upper triangle of the voxel correlation matrix for each cube with the proximal diagonals from episode template matrix[67]. This yielded, for each participant, a voxelwise map of correlation values. We then performed a one-sample *t*-test on the distribution of (Fisher *z*-transformed) correlations at each voxel, across participants. This resulted in a value for each voxel (cube), describing how reliably its time course followed that of the episode.

We further sought to ensure that our analysis identified regions where the activations' temporal structure specifically reflected that of the episode, rather than regions whose activity was simply autocorrelated at a timescale similar to the episode template's diagonal. To achieve this, we used a phase shift-based permutation procedure, whereby we circularly shifted the episode's topic-proportions matrix by a random number of timepoints (rows), computed the resulting null episode template, and re-ran the searchlight analysis, in full. (For each of the 100 permutations, the same random shift was used for all participants). We *z*-scored the observed (unshifted) result at each voxel against the distribution of permutation-derived null results, and estimated a *P* value by computing the proportion of shifted results that yielded larger values. To create the map in Fig. 8c, we thresholded out any voxels whose similarity to the unshifted episode's structure fell below the 95th percentile of the permutation-derived similarity results.

We used an analogous procedure to identify voxels whose responses reflected the recall templates. For each participant, we correlated the proximal diagonals from the upper triangle of the correlation matrix for each cube of voxels with the proximal diagonals from the upper triangle of their (time-warped) recall correlation matrix. As in the episode template analysis, this yielded a voxelwise map of correlation coefficients for each participant. However, whereas the episode analysis compared every participant's responses to the same template, here the recall templates were unique for each participant. As in the analysis described above, we *t*-scored the (Fisher *z*-transformed) voxelwise correlations, and used the same permutation procedure we developed for the episode responses to ensure specificity to the recall time series and assign significance values. To create the map in Fig. 8d we again thresholded out any voxels whose scores were below the 95th percentile of the permutation-derived null distribution.

Neurosynth decoding analyses. Neurosynth[38] parses a massive online database of over 14,000 neuroimaging studies and constructs meta-analysis images for over 13,000 psychology- and neuroscience-related terms, based on NIfTI images accompanying studies where those terms appear at a high frequency. Given a novel image (tagged with its value type; for example, *z*-, *t*-, *F*- or *P*-statistics), Neurosynth returns a list of terms whose meta-analysis images are most similar. Our permutation procedure yielded, for each of the two searchlight analyses, a voxelwise map of *z*-values. These maps describe the extent to which each voxel specifically reflected the temporal structure of the episode or individuals' recalls (that is, relative to the null distributions of phase-shifted values). We inputted the two statistical maps described above to Neurosynth to create a list of the ten most representative terms for each map.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The fMRI data we analysed are available online at https://dataspace.princeton.edu/jspui/handle/88435/dsp01nz8062179. The behavioural data are available at https://github.com/ContextLab/sherlock-topic-model-paper/tree/master/data/raw.

Code availability

All of our analysis code can be downloaded from https://github.com/ContextLab/sherlock-topic-model-paper.

Received: 14 September 2018; Accepted: 8 January 2021;
Published online: 11 February 2021

References

1. Murdock, B. B. The serial position effect of free recall. *J. Exp. Psychol.* **64**, 482–488 (1962).
2. Kahana, M. J. Associative retrieval processes in free recall. *Mem. Cogn.* **24**, 103–109 (1996).
3. Yonelinas, A. P. The nature of recollection and familiarity: a review of 30 years of research. *J. Mem. Lang.* **46**, 441–517 (2002).
4. Kahana, M. J. *Foundations of Human Memory* (Oxford Univ. Press, 2012).
5. Koriat, A. & Goldsmith, M. Memory in naturalistic and laboratory contexts: distinguishing accuracy-oriented and quantity-oriented approaches to memory assessment. *J. Exp. Psychol.* **123**, 297–315 (1994).
6. Huk, A., Bonnen, K. & He, B. J. Beyond trial-based paradigms: continuous behavior, ongoing neural activity, and naturalistic stimuli. *J. Neurosci.* **38**, 7551–7558 (2018).
7. Lerner, Y., Honey, C. J., Silbert, L. J. & Hasson, U. Topographic mapping of a hierarchy of temporal receptive windows using a narrated story. *J. Neurosci.* **31**, 2906–2915 (2011).
8. Manning, J. R. Episodic memory: mental time travel or a quantum 'memory wave' function? Preprint at *OSF* https://doi.org/10.31234/osf.io/6zjwb (2019).
9. Manning, J. R. in *Handbook of Human Memory* (eds Kahana, M. J. & Wagner, A. D.) (Oxford Univ. Press, in the press).
10. Howard, M. W. & Kahana, M. J. A distributed representation of temporal context. *J. Math. Psychol.* **46**, 269–299 (2002).
11. Howard, M. W. et al. A unified mathematical framework for coding time, space, and sequences in the medial temporal lobe. *J. Neurosci.* **34**, 4692–4707 (2014).
12. Manning, J. R., Norman, K. A. & Kahana, M. J. in *The Cognitive Neurosciences* 5th edition (ed. Gazzaniga, M.) 557–566 (MIT Press, 2015).
13. Ranganath, C. & Ritchey, M. Two cortical systems for memory-guided behavior. *Nat. Rev. Neurosci.* **13**, 713–726 (2012).
14. Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S. & Reynolds, J. R. Event perception: a mind–brain perspective. *Psychol. Bull.* **133**, 273–293 (2007).
15. Zwaan, R. A. & Radvansky, G. A. Situation models in language comprehension and memory. *Psychol. Bull.* **123**, 162 – 185 (1998).
16. Radvansky, G. A. & Zacks, J. M. Event boundaries in memory and cognition. *Curr. Opin. Behav. Sci.* **17**, 133–140 (2017).
17. Brunec, I. K., Moscovitch, M. M. & Barense, M. D. Boundaries shape cognitive representations of spaces and events. *Trends Cogn. Sci.* **22**, 637–650 (2018).
18. Heusser, A. C., Ezzyat, Y., Shiff, I. & Davachi, L. Perceptual boundaries cause mnemonic trade-offs between local boundary processing and across-trial associative binding. *J. Exp. Psychol. Learn. Mem. Cogn.* **44**, 1075–1090 (2018).
19. Clewett, D. & Davachi, L. The ebb and flow of experience determines the temporal structure of memory. *Curr. Opin. Behav. Sci.* **17**, 186–193 (2017).
20. Ezzyat, Y. & Davachi, L. What constitutes an episode in episodic memory? *Psychol. Sci.* **22**, 243–252 (2011).
21. DuBrow, S. & Davachi, L. The influence of contextual boundaries on memory for the sequential order of events. *J. Exp. Psychol.* **142**, 1277–1286 (2013).
22. Tompary, A. & Davachi, L. Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron* **96**, 228–241 (2017).
23. Chen, J. et al. Shared memories reveal shared structure in neural activity across individuals. *Nat. Neurosci.* **20**, 115 (2017).
24. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet allocation. *J. Mach. Learn. Res.* **3**, 993–1022 (2003).
25. Rabiner, L. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).
26. Baldassano, C. et al. Discovering event structure in continuous narrative perception and memory. *Neuron* **95**, 709–721 (2017).
27. Blei, D. M. & Lafferty, J. D. Dynamic topic models. In *Proc. 23rd International Conference on Machine Learning*, ICML '06 113–120 (ACM, 2006).
28. Manning, J. R., Polyn, S. M., Baltuch, G., Litt, B. & Kahana, M. J. Oscillatory patterns in temporal lobe reveal context reinstatement during memory search. *Proc. Natl Acad. Sci. USA* **108**, 12893–12897 (2011).
29. Howard, M. W., Viskontas, I. V., Shankar, K. H. & Fried, I. Ensembles of human MTL neurons 'jump back in time' in response to a repeated stimulus. *Hippocampus* **22**, 1833–1847 (2012).
30. Atkinson, R. C. & Shiffrin, R. M. Human memory: a proposed system and its control processes. *J. Learn. Motiv.* **2**, 89–105 (1968).
31. Postman, L. & Phillips, L. W. Short-term temporal changes in free recall. *Q. J. Exp. Psychol.* **17**, 132–138 (1965).
32. Welch, G. B. & Burnett, C. T. Is primacy a factor in association-formation. *Am. J. Psychol.* **35**, 396–401 (1924).
33. Polyn, S. M., Norman, K. A. & Kahana, M. J. A context maintenance and retrieval model of organizational processes in free recall. *Psychol. Rev.* **116**, 129–156 (2009).
34. Manning, J. R. & Kahana, M. J. Interpreting semantic clustering effects in free recall. *Memory* **20**, 511–517 (2012).
35. Heusser, A. C., Ziman, K., Owen, L. L. W. & Manning, J. R. HyperTools: a Python toolbox for gaining geometric insights into high-dimensional data. *J. Mach. Learn. Res.* **18**, 1–6 (2018).
36. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection for dimension reduction. Preprint at *arXiv* https://arxiv.org/abs/1802.03426 (2018).
37. Paller, K. A. & Wagner, A. D. Observing the transformation of experience into memory. *Trends Cogn. Sci.* **6**, 93–102 (2002).

38. Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C. & Wager, T. D. Large-scale automated synthesis of human functional neuroimaging data. *Nat. Methods* **8**, 665 (2011).
39. Bellmund, J. L. S., Gärdenfors, P., Moser, E. I. & Doeller, C. F. Navigating cognition: spatial codes for human thinking. *Science* **362**, eaat6766 (2018).
40. Bellmund, J. L. S. et al. Deforming the metric of cognitive maps distorts memory. *Nat. Hum. Behav.* **4**, 177–188 (2020).
41. Constantinescu, A. O., O'Reilly, J. X. & Behrens, T. E. J. Organizing conceptual knowledge in humans with a gridlike code. *Science* **352**, 1464–1468 (2016).
42. Gilboa, A. & Marlatte, H. Neurobiology of schemas and schema-mediated memory. *Trends Cogn. Sci.* **21**, 618–631 (2017).
43. Baldassano, C., Hasson, U. & Norman, K. A. Representation of real-world event schemas during narrative perception. *J. Neurosci.* **38**, 9689–9699 (2018).
44. Huth, A. G., Nisimoto, S., Vu, A. T. & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron* **76**, 1210–1224 (2012).
45. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).
46. Gagnepain, P. et al. Collective memory shapes the organization of individual memories in the medial prefrontal cortex. *Nat. Hum. Behav.* **4**, 189–200 (2020).
47. Simony, E., Honey, C. J., Chen, J. & Hasson, U. Dynamic reconfiguration of the default mode network during narrative comprehension. *Nat. Commun.* **7**, 1–13 (2016).
48. Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A. & Hasson, U. How we transmit memories to other brains: Constructing shared neural representations via communication. *Cereb. Cortex* **27**, 4988–5000 (2017).
49. Simony, E. & Chang, C. Analysis of stimulus-induced brain dynamics during naturalistic paradigms. *NeuroImage* **216**, 116461 (2020).
50. Landauer, T. K. & Dumais, S. T. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychol. Rev.* **104**, 211–240 (1997).
51. Mikolov, T., Chen, K., Corrado, G. & Dean, J. Efficient estimation of word representations in vector space. Preprint at *arXiv* https://arxiv.org/abs/1301.3781 (2013).
52. Cer, D. et al. Universal sentence encoder. Preprint at *arXiv* https://arxiv.org/abs/1803.11175 (2018).
53. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
54. Brown, T. B. et al. Language models are few-shot learners. Preprint at *arXiv* https://arxiv.org/abs/2005.14165 (2020).
55. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
56. Hasson, U., Yang, E., Vallines, I., Heeger, D. J. & Rubin, N. A hierarchy of temporal receptive windows in human cortex. *J. Neurosci.* **28**, 2539–2550 (2008).
57. Hasson, U., Chen, J. & Honey, C. J. Hierarchical process memory: memory as an integral component of information processing. *Trends Cogn. Sci.* **19**, 304–315 (2015).
58. Dobrushin, R. L. Prescribing a system of random variables by conditional distributions. *Theory Probab. Appl.* **15**, 458–486 (1970).
59. Ramdas, A., Trillos, N. & Cuturi, M. On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy* **19**, 47 (2017).
60. Heusser, A. C., Fitzpatrick, P. C., Field, C. E., Ziman, K. & Manning, J. R. Quail: a Python toolbox for analyzing and plotting free recall data. *J. Open Source Softw.* **2**, 424 (2017).
61. Fisher, R. A.*Statistical Methods for Research Workers* (Oliver and Boyd, 1925).
62. Berndt, D. J. & Clifford, J. Using dynamic time warping to find patterns in time series. In *AAAIWS '94: Proc. 3rd International Conference on Knowledge Discovery and Data Mining* 359–370 (1994).
63. Freedman, D., Riesenhuber, M., Poggio, T. & Miller, E. Categorical representation of visual stimuli in the primate prefrontal cortex. *Science* **291**, 312–316 (2001).
64. Sigman, M. & Dehaene, S. Brain mechanisms of serial and parallel processing during dual-task performance. *J. Neurosci.* **28**, 7585–7589 (2008).
65. Charron, S. & Koechlin, E. Divided representations of current goals in the human frontal lobes. *Science* **328**, 360–363 (2010).
66. Rishel, C. A., Huang, G. & Freedman, D. J. Independent category and spatial encoding in parietal cortex. *Neuron* **77**, 969–979 (2013).
67. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis—connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 1 – 28 (2008).

Acknowledgements

We thank L. Chang, J. Chen, C. Honey, C. Lee, L. Owen, E. Whitaker, X. Xu and K. Ziman for feedback and scientific discussions, and we thank J. Chen, Y. C. Leong, C. Honey, C. Yong, K. Norman and U. Hasson for sharing the data used in our study. Our work was supported in part by National Science Foundation Established Program to Stimulate Competitive Research Award number 1632738. The content is solely the responsibility of the authors and does not necessarily represent the official views of our supporting organizations. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

Conceptualization: A.C.H. and J.R.M.; methodology: A.C.H., P.C.F. and J.R.M.; software: A.C.H., P.C.F. and J.R.M.; analysis: A.C.H., P.C.F. and J.R.M.; writing, reviewing and editing: A.C.H., P.C.F. and J.R.M.; and supervision: J.R.M.

Competing interests

The authors declare no competing interests.

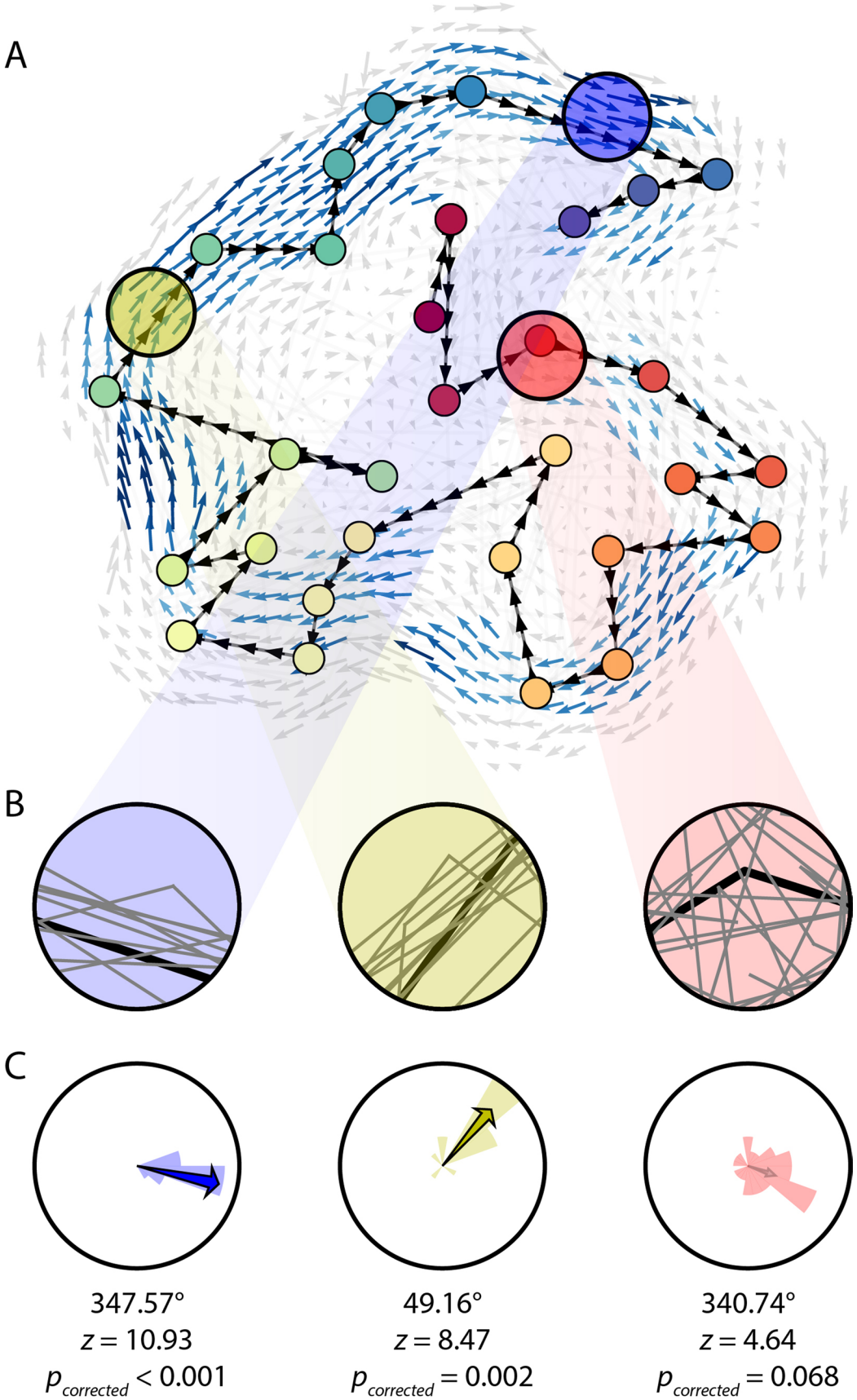
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41562-021-01051-6.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-021-01051-6.

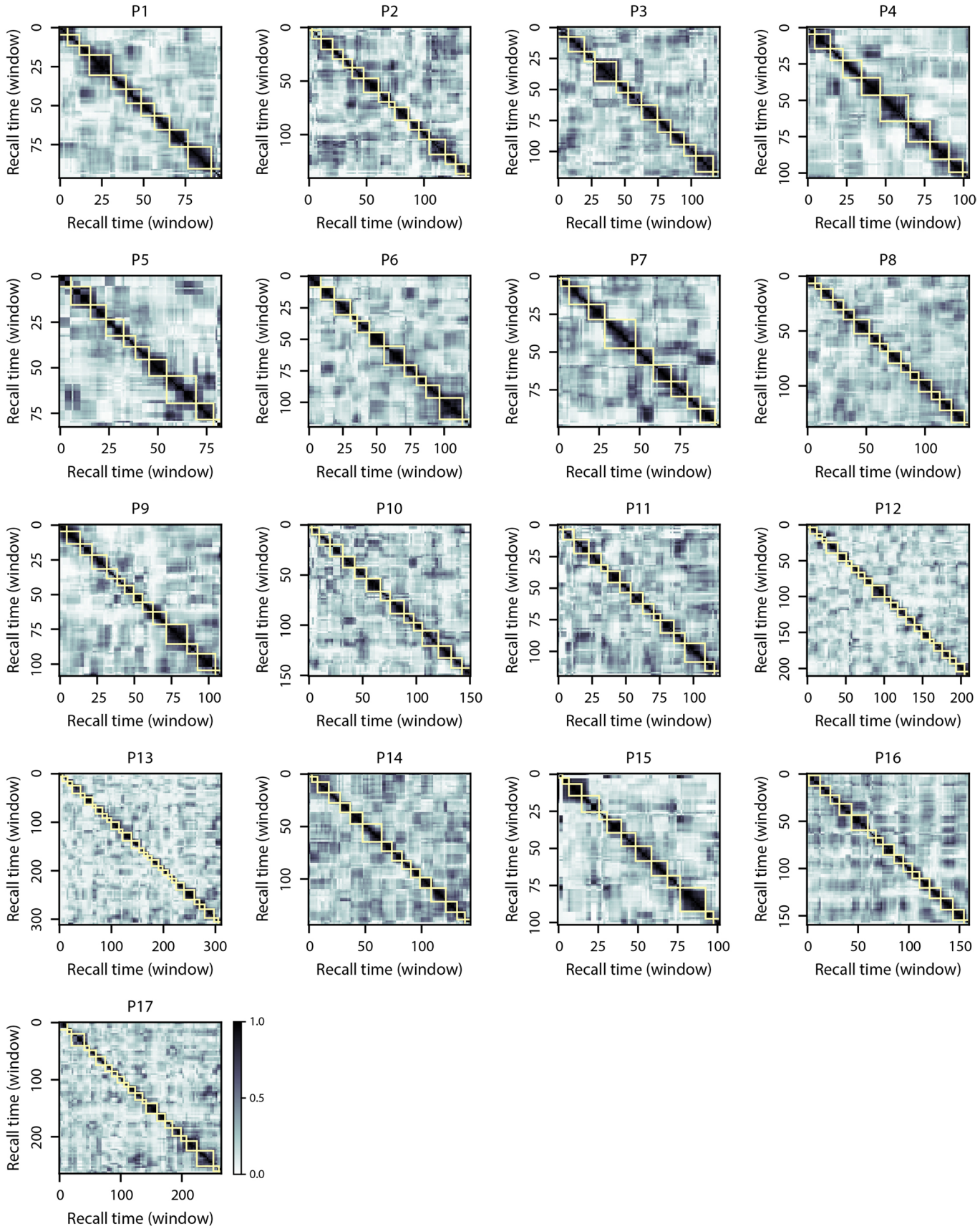
Correspondence and requests for materials should be addressed to J.R.M.

Peer review information Primary Handling Editor: Marike Schiffer.

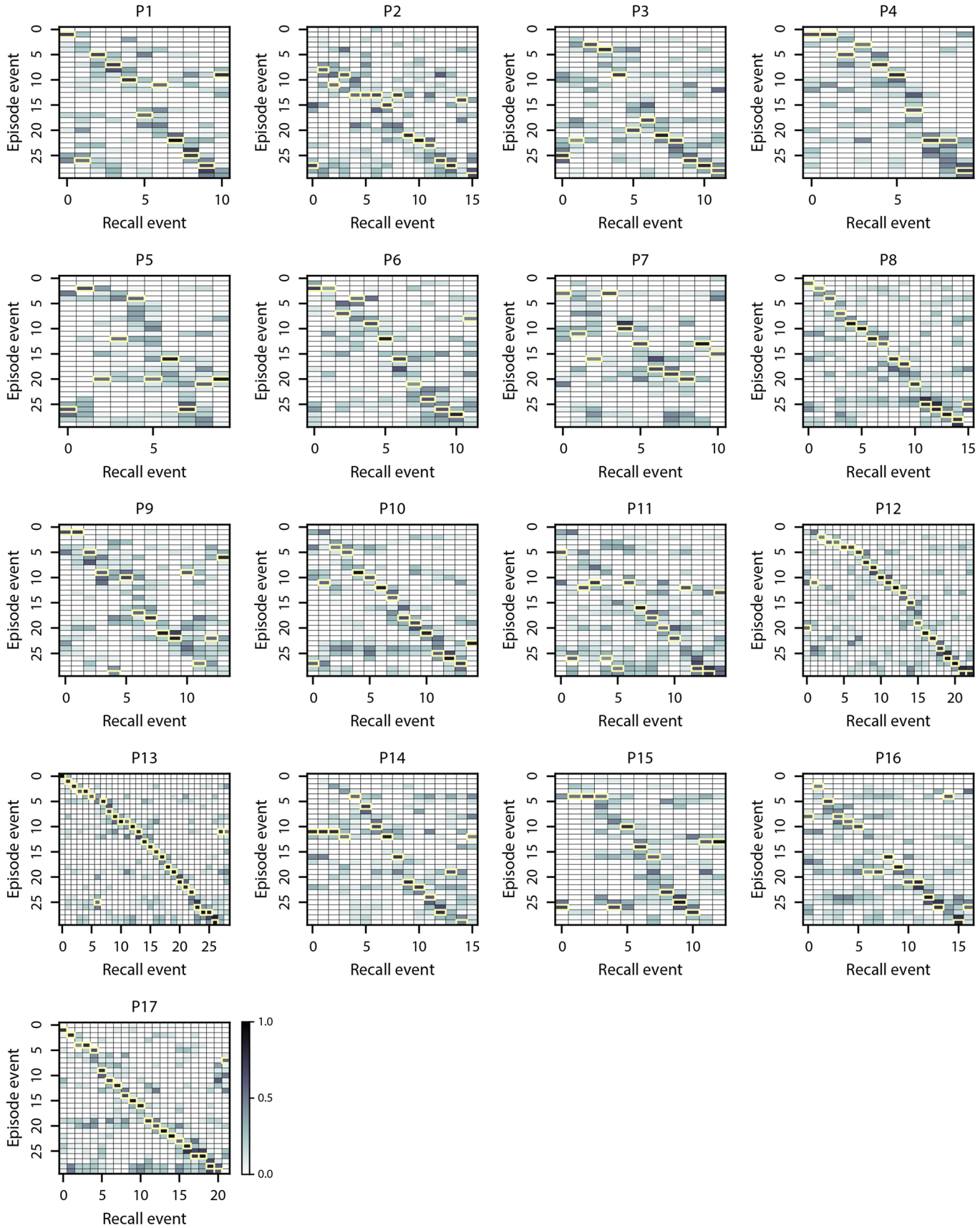


Extended Data Fig. 1 | See next page for caption.

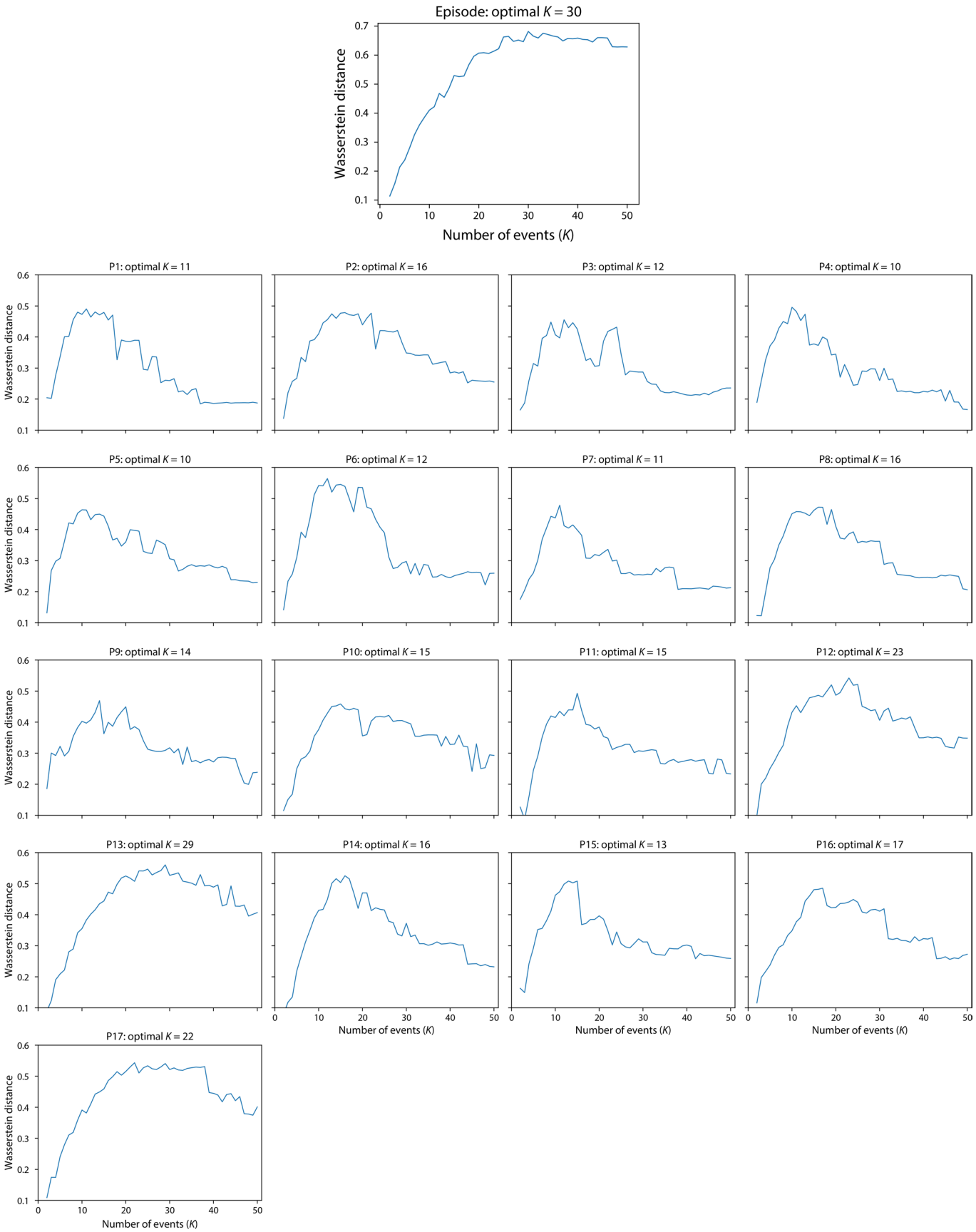
Extended Data Fig. 1 | Methods detail for recall trajectory analysis displayed in Figure 6B. A. This panel replicates Figure 6B, but with two additions. First, individual participants' recall trajectories are displayed (faintly) as light gray lines. Second, three locations on the trajectory have been highlighted (blue, yellow, and red circles). **B**. These zoomed-in views of the locations highlighted in Panel A show the average trajectory (black) and individual participants' trajectories (gray lines) that intersect the given region of topic space. **C**. For each circular region of topic space tiling the 2D embedding plane displayed in Panel A, we compute the distribution of angles formed between each participant's trajectory segment (that is, the point at which the trajectory enters and exists the region of topic space) and the *x*-axis. The distributions of angles for these three example regions are displayed in the colored rose plots. We use Rayleigh tests to assign an arrow direction, length, and color for that region of topic space. Arrows displayed in color are significant at the $p < 0.05$ level (corrected). The arrow directions are oriented according to the circular means of each distribution, and the arrow lengths are proportional to the lengths of those mean vectors. The example regions have been oriented from left to right in decreasing order of consistency across participants.



Extended Data Fig. 2 | Recall temporal correlation matrices and event segmentation fits. Each panel is in the same format as Figure 2E. The yellow boxes indicate HMM-identified event boundaries.



Extended Data Fig. 3 | Episode-recall event correlation matrices. Each panel is in the same format as Figure 2G. The yellow boxes mark the matched episode event for each recall event (that is, the maximum correlation in each column).



Extended Data Fig. 4 | Episode and recall topic proportions matrix *K*-optimization functions. We selected the optimal *K*-value for the episode and each recall topic proportions matrix using the formula described in *Methods*. This computation resulted in a curve for each matrix, describing the Wasserstein distance between the distributions of within-event and across-event topic vector correlations, as a function of *K*.

nature research

Corresponding author(s): Jeremy R. Manning

Last updated by author(s): Nov 26, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	PsychToolbox (http://psychtoolbox.org/) and MATLAB (https://www.mathworks.com/products/matlab.html) were used to present the video.
Data analysis	We used a number of open-source software in our analyses. All code was written in Python. All code used to analyze data and generate figures and text can be found here: https://github.com/ContextLab/sherlock-topic-model-paper. For topic modeling, we used our open-source library called Hypertools (https://hypertools.readthedocs.io/en/latest/), which utilizes scikit-learn (http://scikit-learn.org/stable/index.html). For the Hidden Markov Model, we used the brainIAK toolbox. For dimensionality reduction, we utilized HyperTools which calls UMAP (https://umap-learn.readthedocs.io/en/latest/). For brain-related analyses, we used BrainIAK and nilearn (http://nilearn.github.io/). For list learning analyses, we used our open-source software, Quail (https://cdl-quail.readthedocs.io/en/latest/). For plotting, we used matplotlib (https://matplotlib.org/), seaborn (https://seaborn.pydata.org/index.html) and word-cloud (https://github.com/amueller/word_cloud). Other packages used include pandas (https://pandas.pydata.org/), numpy (http://www.numpy.org/), scipy (https://www.scipy.org/), fastdtw (https://pypi.org/project/fastdtw/), pycircstat (https://github.com/circstat/pycircstat), statsmodels (https://www.statsmodels.org/stable/index.html).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All data and code to generate all figures can be found here: https://github.com/ContextLab/sherlock-topic-model-paper. We analyzed an open dataset originally published alongside Chen et al.'s "Shared memories reveal shared structure in neural activity across individuals" (Nature Neuroscience, vol. 20, p. 115, 2017).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	22 participants; Quoted from Chen et al, 2017: "No statistical methods were used to predetermine sample sizes, but our sample sizes are similar to those reported in previous publications"
Data exclusions	Quoted from Chen et al, 2017: "Twenty-two participants were recruited from the Princeton community (12 male, 10 female, ages 18–26, mean age 20.8)...Data from 5 of the 22 participants were discarded due to excessive head motion (greater than 1 voxel; 2 participants), because recall was shorter than 10 min (2 participants), or for falling asleep during the movie (1 participant). For one participant (#5 in Figs. 2c,f, and 3c) the movie scan ended 75 s early (that is, this participant was missing data for part of scene 49 and all of scene 50)."
Replication	The findings in this study were not replicated.
Randomization	All participants viewed, then verbally recalled, a common 48-minute video. Participants were not assigned to different groups/conditions, and the experimental design precluded randomization of trial order.
Blinding	Blinding was not necessary because there was no explicit experimental manipulation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology and archaeology
☐	☐ Animals and other organisms
☐	☒ Human research participants
☐	☐ Clinical data
☐	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☐ Flow cytometry
☐	☒ MRI-based neuroimaging

Antibodies

Antibodies used	*Describe all antibodies used in the study; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Validation	*Describe the validation of each primary antibody for the species and application, noting any validation statements on the manufacturer's website, relevant citations, antibody profiles in online databases, or data provided in the manuscript.*

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	*State the source of each cell line used.*
Authentication	*Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.*
Mycoplasma contamination	*Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.*
Commonly misidentified lines (See ICLAC register)	*Name any commonly misidentified cell lines used in the study and provide a rationale for their use.*

Palaeontology and Archaeology

Specimen provenance	*Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).*
Specimen deposition	*Indicate where the specimens have been deposited to permit free access by other researchers.*
Dating methods	*If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.*

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	*For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.*
Wild animals	*Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.*
Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Quoted from Chen et al, 2017: "Twenty-two participants were recruited from the Princeton community (12 male, 10 female, ages 18–26, mean age 20.8). All participants were right-handed native English speakers, reported normal or corrected-to-normal vision, and had not watched any episodes of Sherlock before the experiment."
Recruitment	See Chen et al, 2017 for specific recruitment information.
Ethics oversight	The testing protocol was approved by Princeton University's Institutional Review Board.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*

Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☐	☐	Public health
☐	☐	National security
☐	☐	Crops and/or livestock
☐	☐	Ecosystems
☐	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*

Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☐ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☐ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	*Describe the sample preparation, detailing the biological source of the cells and any tissue processing steps used.*
Instrument	*Identify the instrument used for data collection, specifying make and model number.*
Software	*Describe the software used to collect and analyze the flow cytometry data. For custom code that has been deposited into a community repository, provide accession details.*
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	*Describe the gating strategy used for all relevant experiments, specifying the preliminary FSC/SSC gates of the starting cell population, indicating where boundaries between "positive" and "negative" staining cell populations are defined.*

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	"Naturalistic" task: fMRI was recorded while participants watched and verbally recalled a video
Design specifications	The stimulus was divided into a 23 minute (946 TR) and a 25 min (1030 TR) segment to facilitate technical issues related to the scanner. After finishing the clip, participants were instructed to (quoting from Chen et al., 2017) “describe what they recalled of the movie in as much detail as they could, to try to recount events in the original order they were viewed in, and to speak for at least 10 minutes if possible but that longer was better. They were told that completeness and detail were more important than temporal order, and that if at any point they realized they had missed something, to return to it. Participants were then allowed to speak for as long as they wished, and verbally indicated when they were finished (e.g., ‘I’m done’).”
Behavioral performance measures	Verbal summaries of the video were collected.

Acquisition

Imaging type(s)	functional
Field strength	3
Sequence & imaging parameters	T2*-weighted echo-planar imaging (EPI) pulse sequence (TR 1,500 ms, TE 28 ms, flip angle 64, whole-brain coverage 27 slices of 4 mm thickness, in-plane resolution 3 × 3 mm2, FOV 192 × 192 mm2), with ascending interleaved acquisition.
Area of acquisition	whole brain

Diffusion MRI ☐ Used ☒ Not used

Preprocessing

Preprocessing software	Quoted from Chen et al., 2017: "Preprocessing was performed in FSL (http://fsl.fmrib.ox.ac.uk/fsl), including slice time correction, motion correction, linear detrending, high-pass filtering (140 s cutoff), and coregistration."
Normalization	Quoted from Chen et al., 2017: "..affine transformation of the functional volumes to a template brain (Montreal Neurological

Normalization	Institute (MNI) standard). Functional images were resampled to 3 mm isotropic voxels for all analyses. All calculations were performed in volume space."
Normalization template	MNI template brain (3mm isotropic)
Noise and artifact removal	FSL was used to remove artifacts related to motion and signal drift (linear detrending).
Volume censoring	N/A

Statistical modeling & inference

Model type and settings	We developed a searchlight analysis whereby we constructed a 5 mm3 cube centered on each voxel. For each of these cubes, we computed the temporal correlation matrix of the voxel responses during movie viewing. Specifically, for each of the 1976 volumes collected during movie viewing, we correlated the activity patterns in the given cube with the activity patterns (in the same cube) collected during every other timepoint. This yielded a 1976 by 1976 correlation matrix for each cube. Next, we constructed two sets of "template" matrices; one reflected the video's topic trajectory and the other reflected each participant's recall topic trajectory. To construct the video template, we computed the correlations between the topic proportions estimated for every pair of TRs (prior to segmenting the trajectory into discrete events). We constructed similar temporal correlation matrices for each participant's recall topic trajectory. However, to correct for length differences and potential non-linear transformations between viewing time and recall time, we first used dynamic time warping (Berndt and Clifford, 1994) to temporally align participants' recall topic trajectories with the video topic trajectory. This yielded a 1976 by 1976 correlation matrix for the video template and for each participant's recall template.
Effect(s) tested	To determine which (cubes of) voxel responses reliably matched the video template, we correlated the upper triangle of the voxel correlation matrix for each cube with the upper triangle of the video template matrix (Kriegeskorte et al., 2008). This yielded, for each participant, a single correlation value. We computed the average (Fisher z-transformed) correlation coefficient across participants.

Specify type of analysis: ☒ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)	voxel-wise p < .05 (corrected)
Correction	We used a permutation-based procedure to assess significance, whereby we re-computed the average correlations for each of 100 "null" video templates (constructed by circularly shifting the template by a random number of timepoints). (For each permutation, the same shift was used for all participants.) We then estimated a p-value by computing the proportion of shifted correlations that were larger than the observed (unshifted) correlation. To create the map in Figure 5A we thresholded out any voxels whose correlation values fell below the 95th percentile of the permutation-derived null distribution. We used a similar procedure to identify which voxels' responses reflected the recall templates. For each participant, we correlated the upper triangle of the correlation matrix for each cube of voxels with their (time warped) recall correlation matrix. As in the video template analysis this yielded a single correlation coefficient for each participant. However, whereas the video analysis compared every participant's responses to the same template, here the recall templates were unique for each participant. We computed the average z-transformed correlation coefficient across participants, and used the same permutation procedure we developed for the video responses to assess significant correlations. To create the map in Figure 5B we thresholded out any voxels whose correlation values fell below the 95th percentile of the permutation-derived null distribution.

Models & analysis

n/a	Involved in the study
☒	☐ Functional and/or effective connectivity
☒	☐ Graph analysis
☐	☒ Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis	We constructed two sets of "template" matrices; one reflected the video's topic trajectory and the other reflected each participant's recall topic trajectory. To construct the video template, we computed the correlations between the topic proportions estimated for every pair of TRs (prior to segmenting the trajectory into discrete events). We constructed similar temporal correlation matrices for each participant's recall topic trajectory. However, to correct for length differences and potential non-linear transformations between viewing time and recall time, we first used dynamic time warping (Berndt and Clifford, 1994) to temporally align participants' recall topic trajectories with the video topic trajectory. This yielded a 1976 by 1976 correlation matrix for the video template and for each participant's recall template. Our evaluation metric was the correlation (Pearson's r) between the upper triangle of the templates described about and the upper triangle of the searchlight cube (5mm) of voxel responses.