# Chunking Space and Time with Information Geometry

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Humans are exposed to a continuous stream of sensory data, yet understand the world in terms of discrete concepts. A large body of work has focused on chunking sensory data in time, i.e. finding event boundaries, typically identified by model prediction errors. Similarly, chucking sensory data in space is the problem at hand when building spatial maps for navigation. In this work, we argue that a single mechanism underlies both, which is building a hierarchical generative model of perception and action, where chunks at a higher level are formed by segments surpassing a certain information distance at the level below. We demonstrate how this can work in the case of robot navigation, and discuss how this could relate to human cognition in general.

## 1 Introduction

People observe the world around them through a continuous stream of multi-modal sensory data, yet they perceive and conceive in terms of discrete concepts. Event Segmentation Theory (EST) suggests that continuous sensory streams are segmented into discrete "events", which have a particular beginning and end [1], and their event boundaries are triggered by prediction errors [2]. This can be grounded in a predictive processing account of how the brain works [3–5]. In machine learning, this has inspired many to build predictive world models, and use these to detect event boundaries by either detecting spikes in prediction error [6, 7] or by inspecting changes of sparsely-gated latent space vectors [8]. Besides perceiving the world in discrete "events", this chunking has also been related to memory creation [9] and time perception [10], with a central role attributed to the hippocampus and medial prefrontal cortex in forming event memories [11].

A different (albeit similar) chunking is found in the hippocampal/entorhinal system in the context of spatial navigation. In particular, allocentric spatial representations in the hippocampus, so called "place cells", exhibit reliable responses when an organism visits a given discrete position in space [12]. These place cells appear to be tuned to the direction of motion, whether in 2 or 3 spatial dimensions [13], and seem to play an important role in planning, as forward sweeps of activation were found as organisms consider alternative routes, and the most robust sweeps predict the direction of subsequent locomotion [14, 15]. Similar to event segmentation, recent findings also show that place cell activity changes more abruptly in case of salient locations, i.e. locations that crossed cue boundaries [16].

In this paper, we propose that a single mechanism might underpin both phenomena. We suggest the brain, and in particular the hippocampus, is involved in learning a hierarchical, temporal predictive model [17] of multi-modal sensory inputs. In Section 2, we argue that the "chunk boundaries" of lower level states encoded in a higher level state are not solely determined by instantaneous peaks in Bayesian surprise or prediction error, but are rather a function(al) of the underlying trajectory surpassing a certain distance in information geometry. We then demonstrate in Section 3 how this can

Submitted to Information-Theoretic Principles in Cognitive Systems Workshop at the 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

be implemented in a 2-level hierarchical model targeted at robot navigation, and discuss in Section 4 how this could be extended further towards more general cognition.

## 2 Moving from continuous to discrete using information geometry

We consider an agent that acts in an environment by executing actions $a_t$ and observing sensory inputs $\{o_t^{(0)}, o_t^{(1)}, ...\}$ of different sensing modalities. Given a trajectory of observations of the $i$th modality $\tau = \{o_0^{(i)}, o_1^{(i)}, ..., o_t^{(i)}\}$, we want to assess the information contained in that sequence, as that will be our desired unit to "chunk". To that end, for each modality $i$, the agent builds a predictive probabilistic world model [18], consisting of:

$$\text{Transition:} \quad p(s_t^{(i)}|s_{t-1}^{(i)}, a_{t-1}), \quad \text{Likelihood:} \quad p(o_t^{(i)}|s_t^{(i)}), \quad \text{Posterior:} \quad q(s_t^{(i)}|s_{t-1}^{(i)}, a_{t-1}, o_t^{(i)}),$$

by optimizing the evidence lower bound, i.e. minimizing variational Free Energy [17]:

$$\mathcal{L} = D_{KL}[q(s_t^{(i)}|s_{t-1}^{(i)}, a_{t-1}, o_t^{(i)})\|p(s_t^{(i)}|s_{t-1}^{(i)}, a_{t-1})] - \mathbb{E}_{q(s_t^{(i)})}[\log p(o_t^{(i)}|s_t^{(i)})].$$

Hence, the agent effectively learns a statistical manifold for each modality, and can evaluate the information distance traveled along a trajectory using the Fisher information metric [19]. Note how the Fisher information metric is also related to the KL divergence. Effectively, if we evaluate $D_{KL}\big[x||x+\delta x\big]$ with $x$ a probability distribution and $x+\delta x$ a distribution close to $x$, we get that if $\delta x \to 0$ then $D_{KL}\big[x||x+\delta x\big] \to 1/2F(x)(\delta x)^2$. In other words, for infinitesimally small differences between distributions the KL divergence approaches the Fisher information metric [20]. This can be interpreted as integrating the agent's Bayesian surprise over infinitesimal timesteps to measure the information distance traveled. This means that instantaneous high Bayesian surprise peaks will still trigger event boundaries as in EST [2], but also gradual but slow increases in information distance will trigger discrete chunks in regular patterns as witnessed in place cells [12].

Based on this mechanism we can now instantiate a more coarse grained, higher level generative model, by determining an information threshold $\theta$ for each modality. As the agent interacts with the environment, we monitor the information distance traveled, and once it surpasses $\theta$ for any of the modalities, a node is formed for the higher level generative model, which we instantiate as a non-parametric graph structure. This is similar to the Subjective Timescale Model (STM) [7], but here we keep a graph structure for modeling possible state transitions, instead of fitting a recurrent neural model. In our case, the graph becomes the representation of a higher level generative model, with the different nodes being the high level, discrete states, and the (directed) edges between nodes representing the transition probabilities.

## 3 Example: Simultaneous Localization And Mapping (SLAM)

As a proof of principle we consider the case of a robot navigating an environment. In this case, the robot's actions are the commanded forward and angular velocity, and it has access to two sensing modalities: a first person view camera, and a proprioceptive odometry signal.

### 3.1 Low-level models

To learn a statistical manifold of camera views, we instantiate a Transition, Likelihood and Posterior model using deep neural networks, and train these on sequences of historical data of the robot driving around [21]. For the posterior model we use a convolutional neural network for processing the input pixels, after which we concatenate the previous state and action vectors and feed it to a multilayer perceptron. The likelihood model is composed of convolutional and upsampling layers, in order to turn a latent state vector back into pixels. The transition model consists of an LSTM [22].

For the priopriocepitive stream on the other hand, we use a pose continuous attractor network (CAN) [23] configured in a wrapped around three-dimensional grid which is topologically equivalent to a torus in SE(3). The dimensions of this grid represent the pose (i.e. the $x$ and $y$ position and rotation around the $z$-axis, i.e. the heading) of the robot. A cell in the grid is active if the robot is thought to be positioned closely to the corresponding pose, and cells are activated based on the proprioceptive sensory input, but also from top-down predictions (see later). At any given time, the

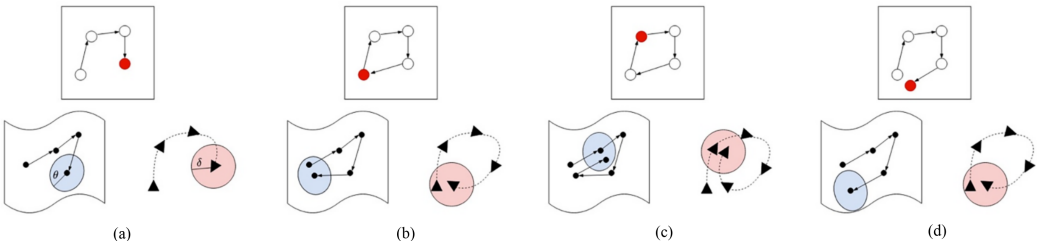


Figure 1: Different cases for the high-level map updating procedure. For each case we show the map (top), views (bottom left) and poses (bottom right) in their own respective spaces, and the current active node is always indicated in red. New map nodes are created when both (a) or one (d) modality surpasses a certain threshold on distance on the statistical manifold. Otherwise, one witnesses a loop closure event (b) or a shift in posterior belief at the higher level (c).

cell with the highest activity will be considered the best estimate of the current pose state [24]. The choice of a CAN to model the transition, likelihood and posterior in pose space, was made to better ground the model for spatial navigation, effectively performing path integration. Moreover, the pose CAN closely matches several features found in the entorhinal cortex, such as grid cells [25] and head direction cells [26].

### 3.2 High-level map

As mentioned earlier, as the agent surpasses the information threshold $\theta$, a new node is inserted in the graph which corresponds to the current view and pose state, and is linked to the previous graph node. This process results in an ever-growing map of the environment as the agent explores the world. However, in order to build a consistent map, there needs to be a principled way to determine whether a node is new or is already known to the agent. Therefore, we also monitor the distance to other, previously added nodes, and account for loop-closure events in case a matching node is found.

Figure 1 gives a visual overview of the various possible matching cases. If neither view nor pose state match with any of the previously stored view or pose states, a new node is created and inserted into the map (Fig. 1a). When the view and the pose both match, a loop-closure has occurred and the current node shifts to the stored node (Fig. 1b). If the currently observed node matches with a stored node further along the path, a relocation is required, and the estimate is shifted further along the path in the graph (Fig. 1c). At this point, we can also introduce a top-down prediction for the lower level, i.e. by injecting activity in the pose CAN at the grid cell that corresponds with this node. Finally, if the current pose estimate matches a stored node pose, but does not find the corresponding matching view state, a new node is inserted at the same location (Fig. 1d). This allows the agent to keep track of varying views of the same landmark throughout the map.

The result can be cast as a hierarchical generative model that is optimizing variational Free Energy [27]. Effectively, the resulting high-level graph yields a topological map of the environment, capturing the global structure of space. Figure 2 shows the resulting map when evaluating on trajectories of a robot driving through a warehouse like environment, with four distinct, but visually similar aisles.

### 3.3 Setting the threshold

Of course the threshold $\theta$ has a significant impact on the shape and content of the map. We empirically tuned the threshold parameter $\theta^\star$ to obtain a correct topological map in Figure 2b. A threshold much lower than this optimal value (Fig. 2c) will result in a mapping procedure with almost no loop-closure events. The map will contain every tiny perturbation in views and poses as a separate node and will be insufficient in countering odometry drift. Conversely, if the threshold is set much higher than $\theta^\star$, the mapping procedure will lump everything together in a small cluster of nodes (Fig. 2d). We hypothesize that this may potentially have a crucial source of individual differences in cognition, where a threshold ranging from low to high corresponds with a cognitive spectrum (and potential basis for differential diatheses) spanning from autism to schizophrenia [28, 29].

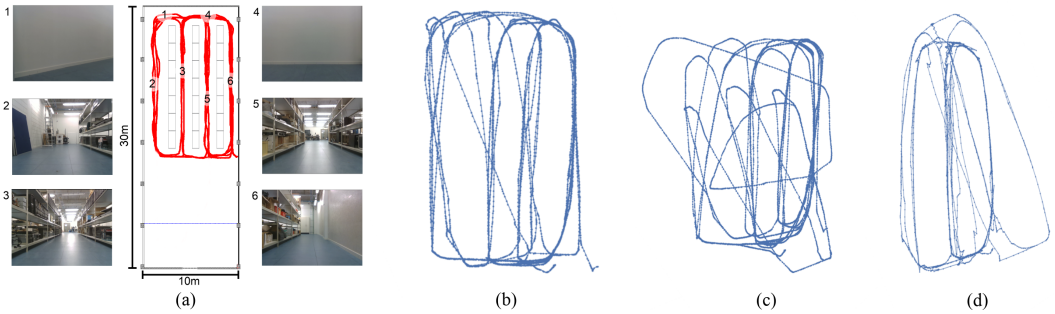


Figure 2: (a) A mobile robot drives around in a warehouse environment, with the ground truth path highlighted in red, and example camera views at distinct locations on the trajectory. (b) A topological map generated using an empirically determined optimal threshold $\theta^\star$, which recovers the ground truth layout. (c) When $\theta < \theta^\star$, many more distinct chunks are created, and loop closures are not detected. (d) When $\theta > \theta^\star$, visually similar aisles are mapped onto the same topological path.

## 4 G-SLAM: building general cognitive maps

Although we have focused on a SLAM example, we believe this concept can be extended to human cognition more generally. For instance, the Tolman-Eichenbaum machine (TEM) [30] similarly proposes a two-level hierarchical model, which generalized to both spatial navigation and relational memory tasks, but using an attractor network for memory retrieval [31] instead of our graph representation. In contrast to our approach, the TEM already assumed discrete inputs, and hence did not require any "chunking" mechanism. An interesting direction of future work might be to see to what extent the more flexible TEM could deal with chunked representations of a continuous lower level data stream.

More recently, Stoianov et al. [32] propose a three level hierarchical model, where each level models individual observations, trajectories and different maps respectively. Here, the first two layers are consistent with our hierarchical model, whereas the the third layer adds an extra context layer which allows to maintain beliefs over different environments or maps. However, also in this work, observations were one-hot encoded, and sequences were represented as a weighted sum of subsequent observations, effectively fixing the temporal depth of each sequence. It is interesting to note that our model also supports generative replay, and is consistent with the view of the hippocampus as a "sequence generator" [33]. In addition, our low level dynamics models can be used to discover novel edges in the map [34], which corresponds to the construction of never-experienced novel-path sequences found in hippocampal neural ensemble recordings [35].

We believe these kinds of architectures provide a general framework for understanding fundamental elements of minds and brains. The hippocampal/entorhinal system can be considered to be the top of the cortical heterarchy, while also providing foundations in terms of phylogenetic/ontogenic primacy. This core predictive-memory system allows for powerful structure learning, and in particular appears to have been selected (both evolutionarily and developmentally) for allowing large-scale dynamics to be remembered and orchestrated (for the sake of imaginative exploration and planning) along (generalized) spatiotemporal trajectories. Indeed, to localize something within a spatialized reference frame - which itself is impacted by the entities it maps/graphs - may be what it means to 'understand' and 'explain' something [36]. Therefore, we think that a Generalized SLAM (G-SLAM) architecture will be key to developing human-like cognition [37].

## References

[1] J. M. Zacks, N. K. Speer, K. M. Swallow, T. S. Braver, and J. R. Reynolds, "Event perception: a mind-brain perspective.," *Psychological bulletin*, vol. 133, no. 2, p. 273, 2007.

[2] C. A. Kurby and J. M. Zacks, "Segmentation in the perception and memory of events," *Trends in Cognitive Sciences*, vol. 12, pp. 72–79, Feb. 2008.

[3] K. Friston, "The free-energy principle: a unified brain theory?," *Nature Reviews Neuroscience*, vol. 11, pp. 127–138, Jan. 2010.

[4] J. Hohwy, *The Predictive Mind*. Oxford University Press, Nov. 2013.

[5] A. Clark, *Surfing Uncertainty*. Oxford University Press, Jan. 2016.

[6] J. Schmidhuber, "Learning complex, extended sequences using the principle of history compression," *Neural Computation*, vol. 4, pp. 234–242, March 1992.

[7] A. Zakharov, M. Crosby, and Z. Fountas, "Episodic memory for subjective-timescale models," in *ICML 2021 Workshop on Unsupervised Reinforcement Learning*, 2021.

[8] C. Gumbsch, M. V. Butz, and G. Martius, "Sparsely changing latent states for prediction and planning in partially observable domains," in *Advances in Neural Information Processing Systems* (A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), 2021.

[9] S. J. Gershman, A. Radulescu, K. A. Norman, and Y. Niv, "Statistical computations underlying the dynamics of memory updating," *PLOS Computational Biology*, vol. 10, pp. 1–13, 11 2014.

[10] Z. Fountas, A. Sylaidi, K. Nikiforou, A. K. Seth, M. Shanahan, and W. Roseboom, "A Predictive Processing Model of Episodic Memory and Time Perception," *Neural Computation*, vol. 34, pp. 1501–1544, 06 2022.

[11] W. Liu, Y. Shi, J. N. Cousins, N. Kohn, and G. Fernández, "Hippocampal-medial prefrontal event segmentation and integration contribute to episodic memory formation," *Cerebral Cortex*, vol. 32, pp. 949–969, Aug. 2021.

[12] J. O'Keefe and L. Nadel, *The Hippocampus as a Cognitive Map*. Oxford: Clarendon Press, 1978.

[13] A. Sarel, A. Finkelstein, L. Las, and N. Ulanovsky, "Vectorial representation of spatial goals in the hippocampus of bats," *Science*, vol. 355, pp. 176–180, Jan. 2017.

[14] A. Johnson and A. Redish, "Neural ensembles in ca3 transiently encode paths forward of the animal at a decision point," *The Journal of neuroscience : the official journal of the Society for Neuroscience*, vol. 27, pp. 12176–89, 12 2007.

[15] K. Kay, J. Chung, M. Sosa, J. Schor, M. Karlsson, M. Larkin, D. Liu, and L. Frank, "Constant sub-second cycling between representations of possible futures in the hippocampus," *Cell*, vol. 180, 01 2020.

[16] C.-H. Wang, J. D. Monaco, and J. J. Knierim, "Hippocampal place cells encode local surface-texture boundaries," *Current Biology*, vol. 30, no. 8, pp. 1397–1409.e7, 2020.

[17] K. J. Friston, R. Rosch, T. Parr, C. Price, and H. Bowman, "Deep temporal models and active inference," *Neuroscience &amp Biobehavioral Reviews*, vol. 77, pp. 388–402, June 2017.

[18] D. Ha and J. Schmidhuber, "Recurrent world models facilitate policy evolution," in *Advances in Neural Information Processing Systems* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds.), Curran Associates, Inc., 2018.

[19] S. I. R. Costa, S. A. Santos, and J. E. Strapasson, "Fisher information distance: A geometrical reading," *Discrete Applied Mathematics*, vol. 197, pp. 59–69, 2015.

[20] S. Kullback, *Information Theory and Statistics*. New York: Wiley, 1959. container-title: Information Theory and Statistics.

[21] O. Çatal, S. Wauthier, C. D. Boom, T. Verbelen, and B. Dhoedt, "Learning generative state space models for active inference," *Frontiers in Computational Neuroscience*, vol. 14, Nov. 2020.

[22] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[23] F. P. Battaglia and A. Treves, "Attractor neural networks storing multiple space representations: A model for hippocampal place fields," *Phys. Rev. E*, vol. 58, pp. 7738–7753, Dec 1998.

[24] M. Milford, G. Wyeth, and D. Prasser, "RatSLAM: a hippocampal model for simultaneous localization and mapping," in *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004*, IEEE, 2004.

[25] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex," *Nature*, vol. 436, p. 801–806, 2005.

[26] P. E. Sharp, H. T. Blair, and J. Cho, "The anatomical and computational basis of the rat head-direction cell signal," *Trends in Neurosciences*, vol. 24, no. 5, pp. 289–294, 2001.

[27] O. Çatal, T. Verbelen, T. Van de Maele, B. Dhoedt, and A. Safron, "Robot navigation as hierarchical active inference," *Neural Networks*, vol. 142, pp. 192–204, 2021.

[28] S. G. Byars, S. C. Stearns, and J. J. Boomsma, "Opposite risk patterns for autism and schizophrenia are associated with normal variation in birth size: phenotypic support for hypothesized diametric gene-dosage effects," *Proceedings of the Royal Society B: Biological Sciences*, vol. 281, p. 20140604, 11 2014.

[29] B. Crespi and N. Dinsdale, "Autism and psychosis as diametrical disorders of embodiment," *Evolution, Medicine, and Public Health*, vol. 2019, pp. 121–138, 1 2019.

[30] J. C. Whittington, T. H. Muller, S. Mark, G. Chen, C. Barry, N. Burgess, and T. E. Behrens, "The tolman-eichenbaum machine: Unifying space and relational memory through generalization in the hippocampal formation," *Cell*, vol. 183, no. 5, pp. 1249 – 1263.e23, 2020.

[31] J. Ba, G. E. Hinton, V. Mnih, J. Z. Leibo, and C. Ionescu, "Using fast weights to attend to the recent past," in *Advances in Neural Information Processing Systems* (D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds.), vol. 29, Curran Associates, Inc., 2016.

[32] I. Stoianov, D. Maisto, and G. Pezzulo, "The hippocampal formation as a hierarchical generative model supporting generative replay and continual learning," *Progress in Neurobiology*, vol. 217, p. 102329, Oct. 2022.

[33] G. Buzsáki and D. Tingley, "Space and time: The hippocampus as a sequence generator," *Trends in Cognitive Sciences*, vol. 22, pp. 853–869, Oct. 2018.

[34] D. de Tinguy, P. Mazzaglia, T. Verbelen, and B. Dhoedt, "Home run: Finding your way home by imagining trajectories," in *3rd International Workshop on Active Inference (IWAI), in conjunction with ECML PKDD*, 2022.

[35] A. S. Gupta, M. A. van der Meer, D. S. Touretzky, and A. D. Redish, "Hippocampal replay is not a simple function of experience," *Neuron*, vol. 65, pp. 695–705, Mar. 2010.

[36] G. Lakoff and M. Johnson, *Philosophy in the Flesh : The Embodied Mind and Its Challenge to Western Thought*. Basic Books, 12 1999.

[37] A. Safron, O. Çatal, and T. Verbelen, "Generalized simultaneous localization and mapping (g-slam) as unification framework for natural and artificial intelligences: towards reverse engineering the hippocampal/entorhinal system and principles of high-level cognition," *Frontiers in Systems Neuroscience*, 2022.