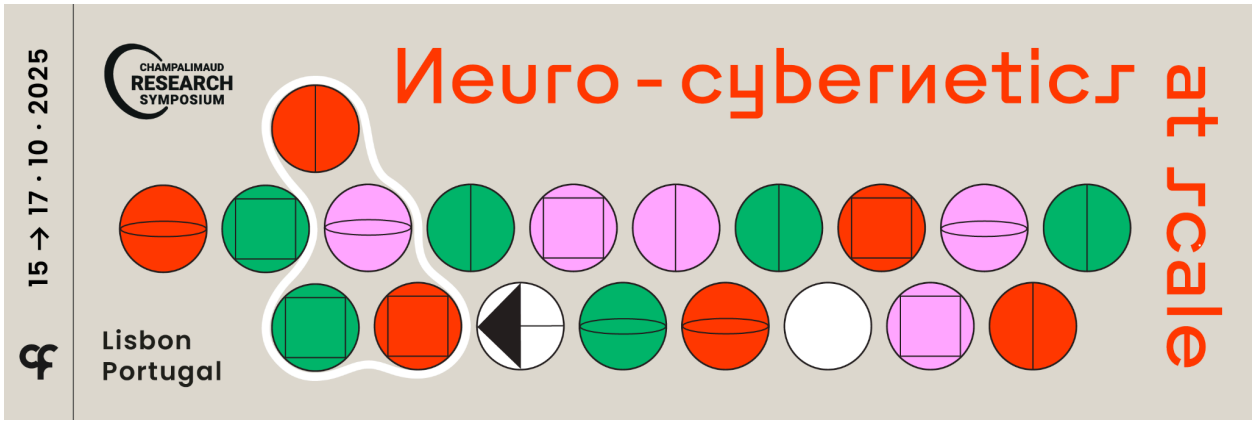

Dynamical Archetype Analysis: Autonomous Computation

Abel Sagodi 1, Il Memming Park 1
1. Champalimaud Research, Lisbon, PT

The study of neural computation aims to understand the function of a neural system as an information processing machine. Neural systems are undoubtedly complex, necessitating principled and automated tools to abstract away details to organize and incrementally build intuition. We argue that systems with the same effective behavior should be abstracted by their ideal representative, i.e., archetype, defined by its asymptotic dynamical structure. We propose a library of archetypical computations and a new measure of dissimilarity that allows us to group systems based on their effective behavior by explicitly considering both deformations that break topological conjugacy as well as diffeomorphisms that preserve it. The proposed dissimilarity can be estimated from observed trajectories. Numerical experiments demonstrate our method's ability to overcome previously reported fragility of existing (dis)similarity measures for approximate continuous attractors and high-dimensional recurrent neural networks. For example, our model can reconstruct deformed and perturbed ring attractors, while the methods Dynamical Similarity Analysis [1] and Smooth Prototype Equivalences [2] struggle to identify expected similarities and differences. Our experiments focus on working memory systems – the backbone of most temporal computation, but our theoretical approach naturally extends to general mechanistic interpretation of recurrent dynamics in both biological and artificial neural systems. We argue that abstract dynamical archetypes, rather than detailed dynamical systems, offer a more useful vocabulary for describing neural computation.

interpretability, dynamical dissimilarity, dynamical archetypes, neural computation, working memory, computation through dynamics