II-105. The geometry of abstraction in hippocampus and pre-frontal cortex

Silvia Bernardi[1] SILVIA.BERNARDI@GMAIL.COM
Marcus K Benna[1] MKB2162@COLUMBIA.EDU
Mattia Rigotti[2] MR2666@COLUMBIA.EDU
Jerome Munuera[3] JEROME.MUNUERA@GMAIL.COM
Stefano Fusi[1] SF2237@COLUMBIA.EDU
C Daniel Salzman[1] CDS2005@COLUMBIA.EDU

[1]Columbia University
[2]IBM Research AI
[3]Columbia University, Centre National de la Recherche Scientifique (CNRS), Ecole Normale Superieure

Abstraction can be defined as a cognitive process that finds a common feature - an abstract variable, or concept - shared by a number of examples. Knowledge of an abstract variable enables generalization, which in turn allows one to apply inference to new examples based upon old ones. Neuronal ensembles could represent abstract variables by discarding all information about specific examples, but this allows for representation of only one variable. Here we show how to construct neural representations that encode multiple abstract variables simultaneously, and we characterize their geometry. Representations conforming to this geometry were observed in dorsolateral pre-frontal cortex, anterior cingulate cortex, and the hippocampus in monkeys performing a serial reversal-learning task. These neural representations allow for generalization, a signature of abstraction, and similar representations are observed in a simulated multi-layer neural network trained with back-propagation. These findings provide a novel framework for characterizing how different brain areas represent abstract variables, which is critical for flexible conceptual generalization and deductive reasoning.