SFB 1315
Mechanisms and Disturbances in Memory Consolidation:
From synapses to systems

Tuesday

JUNE 30, 2020
6:00 pm CET

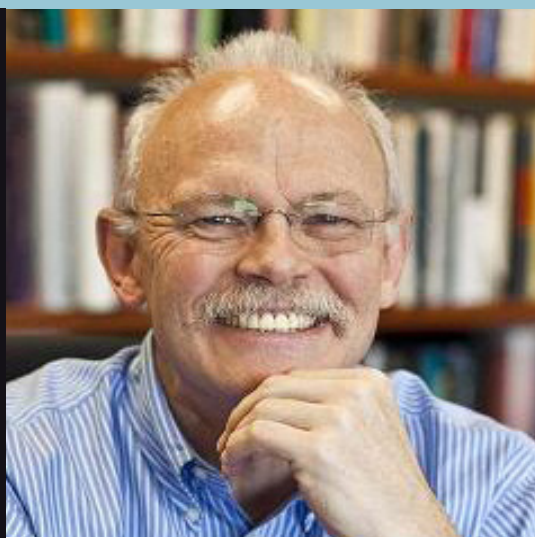
ZOOM ID: 7754910236
Register at:
SFB1315.ifb@hu-berlin.de

SFB 1315 LECTURE SERIES 2019-2020

INTEGRATING NEW KNOWLEDGE WITHOUT CATASTROPHIC INTERFERENCE: COMPUTATIONAL AND THEORETICAL INVESTIGATIONS IN A HIERARCHICALLY STRUCTURED ENVIRONMENT

JAMES L McCLELLAND

Lucie Stern Professor in the Social Sciences
Director, Center for Mind, Brain and Computation
Department of Psychology
Stanford University
Stanford, CA



Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation

SFB 1315
Mechanisms and Disturbances in Memory Consolidation:
From synapses to systems

Tuesday

JUNE 30, 2020
6:00 pm CET

ZOOM ID: 7754910236
Register at:
SFB1315.ifb@hu-berlin.de

INTEGRATING NEW KNOWLEDGE WITHOUT CATASTROPHIC INTERFERENCE: COMPUTATIONAL AND THEORETICAL INVESTIGATIONS IN A HIERARCHICALLY STRUCTURED ENVIRONMENT

According to complementary learning systems theory, integrating new memories into a multi-layer neural network without interfering with what is already known depends on interleaving presentation of the new memories with ongoing presentations of items previously learned.

I use deep linear neural networks in hierarchically structured environments previously analyzed by Saxe, McClelland, and Ganguli (SMG) to gain new insights into this process. For the environment I will consider in this talk, its content can be described by the singular value decomposition (SVD) of the environment's input-output covariance matrix, in which each successive dimension corresponds to categorical split in the hierarchical environment. Prior work showed that deep linear networks are sufficient to learn the content of the environment, and they do so in a stage-line way, with each dimension strength rising from near-zero to its maximum strength after a delay inversely proportional to the strength of the dimension, as previously demonstrated by Saxe et al.

Several observations are then accessible when we consider learning a new item previously not encountered in the micro-environment. (1) The item can be examined in terms of its projection onto the existing structure, and whether it adds a new categorical split. (2) To the extent the item projects onto existing structure, including it in the training corpus leads to the rapid adjustment of the representation of the categories involved, and effectively no adjustment occurs to categories onto which the new item does not project at all. (3) Learning a new split is slow, and its learning dynamics show the same delayed rise to maximum that depends on the dimension's strength. These observations then motivate the development of a similarity-weighted interleaved learning scheme in which only items similar to the to-be-learned new item need be presented to avoid catastrophic interference.

McClelland, J. L., McNaughton, B. L., & Lampinen, A. K. (2020). Integration of New Information in memory: New insights from a complementary learning systems perspective. *Philos Trans R Soc B.* 375: 20190637.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *PNAS.* 116(23), 11537-11546.

Funded by
DFG Deutsche Forschungsgemeinschaft
German Research Foundation