Abstract:
In this talk, I will discuss new work from my lab which attempts to describe the “algorithms” that neural networks (NNs) implement implicitly, as a result of their training. I will focus specifically on NNs ability to encode abstract, compositional functions which consist of interpretable subroutines and which operate in a content-independent manner. I will discuss findings on models trained from scratch on language and vision tasks, as well as large language models (LLMs) in an in-context-learning setting.