Self-attention, an architectural motif designed to model long-range interactions in sequential data, has driven numerous recent breakthroughs in natural language processing and beyond. In this talk, we will focus on studying the inductive bias of self-attention blocks by rigorously establishing which functions and long-range dependencies they statistically represent. Our main result shows that bounded-norm Transformer layers can represent sparse functions of the input sequence, with sample complexity scaling only logarithmically with the context length. Furthermore, we propose new experimental protocols to support this analysis, built around the large body of work on provably learning sparse Boolean functions.

Based on joint work with Benjamin L. Edelman, Sham Kakade and Cyril Zhang.