Deep learning has become the gold standard of computer vision. The generation of high-quality labeled training data typically becomes the bottleneck of deep learning in areas such as natural science, ecology, and medical research, where domain expertise is required to correctly identify targets and thus crowdsourcing becomes non-viable. Yet it is in those areas deep learning has huge potential value. In the worst case, a single domain expert needs to sift through a large volume of unlabeled data to discover only a few positive examples.

In this talk, I will describe our ongoing work on Eureka, a system intended to improve the human expert's productivity in building a labeled training set. Eureka views a human's attention and time as the most precious resource throughout the system and helps to optimize the utilization of this critical resource. Eureka combines three techniques to achieve its goal: early discard, iterative discovery workflow, and edge computing. Experiments show that Eureka can reduce the amount of labeling effort by two orders of magnitude relative to a brute force approach.

Based on joint work with Shilpa George, Jan Harkes, Padmanabhan Pillai, Roberta Klatzky, and Mahadev Satyanarayanan.