Deep learning has enabled valuable computer vision applications in many domains. It can be used to assist domain experts, such as scientists, military, and medical doctors, to detect phenomena of interest quickly. Unfortunately, the manual effort to collect large training sets of domain-specific targets remains a deterrent to deep learning in these domains. Crowd-sourcing is not a viable solution, because the crowds do not have the professional knowledge to accurately label examples. Interesting objects are usually scarce. As a result, a single expert may need to go through millions of unlabeled examples to find a few positive instances of a target.

My thesis is that the above challenges can be addressed by a discard-based system that supports rapid creation of more accurate filters through the use of just-in-time machine learning. I will validate my thesis by building a system called Eureka. Eureka allows an expert to search a large volume of unlabeled data using filters that discard obviously negative data. To further improve the efficacy of early discard, I propose an iterative workflow where the expert labels a small number of candidates returned from Eureka, uses them to create more accurate filters immediately, and restarts the search with better filters, and repeats. The expert’s productivity can thus be improved iteratively throughout the process.

The goal of Eureka is to enable human-efficient discovery of training examples. To this end, computational efficiency, programming abstractions, and query interfaces are all important. My research will first address how to run Eureka efficiently in different computing landscapes, such as edge computing, cloud computing, and smart storage devices. I will identify various performance bottlenecks in different software and hardware architectures, and develop optimizations to alleviate them. Next, I will show Eureka is efficient across different problem domains, such as object detection in images and activity recognition in videos. I will describe how Eureka provides a unified framework to search over different data and query types with unique characteristics.

Thesis Committee:
Mahadev Satyanarayanan, Chair
Martial Hebert
Roberta Klatzky
Padmanabhan Pillai, Intel Labs

Thesis Summary: http://www.cs.cmu.edu/~ziqiangf/proposal.pdf