This thesis is focused on building knowledge bases (KBs) for scientific domains. Specifically, we create structured representations of technical-domain information using unsupervised or semi-supervised learning methods. This work is inspired by recent advances in knowledge base construction based on Web text. However, in the technical domains we consider here, we have grounded data about the objects named by text entities. For example, in the software domain, we can consider the implementation of classes in a code repository, and we can observe the way they are being used in programs. In the biomedical realm, biological ontologies define interactions and relations between entities in this domain, and there is experimental information on domain entities such as proteins and genes. The additional resources available in technical domains, present an opportunity for learning, not only how entities are discussed in text, but also what are their real-world properties.

The main contribution of this thesis is in addressing challenges from the following research areas, in the context of learning about a technical domain: (1) Knowledge representation: How should knowledge about a technical domain be represented and used? (2) Grounding: How can existing resources of technical domains be used in learning? (3) Applications: What applications can benefit from the use of structured knowledge bases dedicated to scientific data?

We explore grounded learning and knowledge base construction for the biomedical and software domains. We first discuss approaches for improving applications based on a grounded statistical model. Next, we construct a deeper representation of domain entities by building a grounded ontology, and through an adaption of an ontology-driven KB learner to scientific input. Finally, we present a topic model framework for knowledge base construction, which jointly optimizes the KB schema and learned facts, and we show that this framework produces high precision KBs in our two domains of interest. We discuss unsupervised and semi-supervised extensions to our model that allow to incorporate grounded data and pre-existing knowledge into the KB learning process.

Thesis Committee:
William Cohen, Chair
Tom Mitchell
Roni Rosenfeld
Alon Halevy, Google Research