**Introduction:**

With memory, storage, and processing power increasing at a dizzing pace and with new data processing frameworks including MapReduce and Spark, researchers in the field of software engineering are able to mine larger repositories of code data. Despite this advantage, developing mining programs using frameworks such as MapReduce tends to be far too low level and time-consuming for the needs of these researchers. To address this limitation, researchers at Iowa State University developed a domain specific language (DSL) called Boa that allows researchers to query large software repositories in a uniform way \cite(Boa\_Paper). In their DSL, the authors have information that pertains to projects, developers, and code.

In addition to this DSL, researchers at Iowa State developed a repository of GitHub data from September, 2015 that contains over 70,000 repositories and SurgeForge data from September, 2015. While this is a large amount of data, it is stored on Iowa State's servers. As such, Boa queries are compiled and run on external servers. Despite the drawback of not being local, there are advantages such as being able to offload computational power and being able to submit queries through a rich web interface, Eclipse plugin or Java API. Further, this data is included in the Mining Software Repositories (MSR) Challenge for 2016.

Due to these factors, Boa and it's affiliated repository dumps of GitHub and SurgeForge are an extremely attractive source of data for this project.