**Title:** Identifying PERILS of mining Github data and JIRA data

**Names:** Yiupang Chan

**Name of Faculty Mentor:** Davide Falessi

**Problem Context:** Github and JIRA projects have become important sources for software engineering researches. However, there are potential perils of mining data from them. My aim was to define and check a set of perils to help researchers measure the validity of the projects’ data. **Challenges:** API Rate Limits: My project exceeded the default rate limit of Github API after measuring one project. I have increased the rate by authenticating myself using personal access tokens on Github. Different names on JIRA and Github: Many projects use different names on JIRA and Github. I have resolved it by fetching the URLs of Github and JIRA from a JSON that contains metadata about all Apache projects. **Limitations of APIs:** For instance, GitPython does not provide users with detail about pull requests. I created a library that makes HTTP requests to Github to mine the data I needed. **Results:** I have successfully produced metrics of seven perils which show a low validity of data in both Github and JIRA projects. For example, some developers did not follow the rule of one commit per JIRA ticket. **Learned Topics:** I have learned three new APIs, GitPython, GitHub API, and JIRA API. Since the project involves data processing and parsing Strings, I have learned using regular expressions to speed up data fetching. Also, I gained more experience in object-oriented design. **Next Steps:** I will continue to work on this project as part of an Independent study CSC 400. The next steps are to measure perils for all the Apache projects and to develop a website application that allows other people to access our results. I am also writing a paper about the results of the projects which we plan to submit in a major internal peer-review conference such as the Mining Software Repository conference: https://conf.researchr.org/home/msr-2018.

*Figure: Class diagram of the Python script I have developed to measure perils.*