**ORSP Requirements**

1. Summary of research findings

2. Number of students involved

3. Expected next steps or outcome of research

**Topics to dicsuss**

1. Discovery of DBSCAN and HDBSCAN algorithms

Mention teams that have developed these algorithms and their variants

2. Building clustering suite to run HDBSCAN with various parameters an the luteo viral dataset, an extremely large, high-dimensional dataset

3. Preparing clustering suite for general use, future datasets, particularly aquafex

4. Giving 3 talks (elaborate on the talks)

5. Sorin Lab Web Platform

6. Thesis preparations

7. Hadoop + Spark Cluster

8. hdbscan contributions

Discussions with Leland McInnes

**Highlights from ORSP Application Letter of Intent**

**Notes from Wiki**

**Draft**

On the third day of the Fall 2016 semester, I met with Dr. Eric Sorin for the first time to discuss his contributions to the field of *Computational Biophysics and Biochemistry* and possible graduate research projects. With my academic background and career interests in mind, he presented me with a fundamental problem that was halting progress in the data analysis portion his research. I accepted the challenge and immediately began digging deep into the vastly growing and widely applicable field of cluster analysis, a family of algorithmic techniques used to build relationships amongst unlabeled data points in a dataset. The lab was and is interested in the most powerful, efficient, and repeatable solution to this problem, due to the vast amounts of unlabeled, biological data that needs to undergo this form of analysis. Thus, I took it upon myself to find possible solutions, analyze and test those solutions, and build software for current and future lab projects. After receiving funding from ORSP, accomplishing this became the foremost goal of my summer.

I spent the first two weeks of summer learning about the basics of clustering, exploring the different approaches, and studying the most recent work of experts in the field today. Based on my initial studies and literature review, I learned that the following fundamental questions must be asked when taking on the goal of clustering a dataset of interest:

* What is a cluster?
* Which features should be used?
* Should the data be normalized?
* How do we define similarity?
* How many clusters are there?
* How should we perform clustering?
* Are the results valid?

Up until now, the lab had been using a popular, yet increasingly ineffective clustering algorithm called *K-Means*. The results produced by running this algorithm on our *luteo* dataset many times, for example, has at least a 40% deviation, where a 5% or less deviation is acceptable for noteworthy results.