Problem Statement

Data summarization is a complex and highly relevant task in the field of machine learning - finding a representative subset modeling a full dataset is a non-trivial task. One of the theoretical frameworks used for tackling this problem is that of submodularity. We say a function is submodular if it has the diminishing returns property (we get more gains from adding elements to a small data-set than to a large data-set). This idea has been discussed in several papers in the past years: Lin and Bilmes, 2011; 2012; Tschiatschek et al., 2014.

The scope of this project is to built upon the recent work of Eric Balkanski et al., 2016, who introduced a novel approach to data summarization based on two-stage submodular maximization. Even though the above paper shows promising results, its reliance on linear programming optimization techniques limits its functionality on really large datasets. The goal of my project is improving their result by developing a faster way of computing the solution set for the two stage submodular function described in their paper, making their solution scalable to really large problems.

Anticipated Challenges

Predicting how the development of a solution to a hard problem will proceed can never be taken as a certainty. For the present project, the hope is to center in on the linear programming approach described for solving the two stage submodular optimization problem and try to come up with a faster method for generating the same, or a similar enough result. If that proves successful, the next step would be to write an implementation of the algorithm and test its performance in practice.
However, if the above proves unsuccessful, the natural next step would be to analyze the initial components of the problem and try to come up with an alternate proof for the submodular maximization problem that permits a faster implementation. This would again have to be verified in practice on relevant datasets, but this approach would be harder than the previous one, requiring more in-depth knowledge of the research field.

**Project Components**

- Developing a proof
  - Analyzing the linear programming component of the base paper and trying to find a theoretical improvement
  - Analyzing the other aspects of the paper and trying to find improvements that can translate to a more scalable implementation of the algorithm
  - If the above fail, finding a novel approach to solve the problem that can translate into a faster overall algorithm
- Testing validity of theoretical results
  - Reading literature and comparing performance on examples present in other research papers with the method developed in this project
  - Verifying if the performance of the theoretical model holds in practice
- Testing performance on large scale datasets
  - Developing a robust implementation to be able to process large-scale datasets
  - Finding or building implementations of other results in the area to compare performance
**Deliverables**

1. Final Report, detailing the findings and obstacles encountered while developing said novel approach
2. Statistics showing how the approach provided in this project improves on the previous result, by showcasing the improved performance of the two stage submodular maximization problem on large data-sets
3. Source code that implements the above method