# Analysing and Improving K-means Clustering Algorithms

**Project Proposal for CS 254, Design and Analysis of Algorithms**

29.03.2017

Mohit Mohta (150001018)

Kailas Sheregar (150001031)

B.Tech (CSE) 2nd Year, IIT Indore.

## Overview

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). So, K-means is an exclusive clustering algorithm i.e. if a data point belongs to one cluster, it can not belong to another at the same time. Here, we will partition n observations into k-clusters.

This project aims to analyse, implement and seek methods and techniques to improve the existing algorithm. This is an np-hard problem. A compilation of the most recent implementation can be found at :

https://github.com/mohtamohit/ClusteringAlgorithms/blob/master/K-means%20Clustering/Introduction.md

## Project Goals

1. To analyze and implement the popular K-means algorithm
2. To seek techniques to improve the existing algorithm

## Specifications

We plan to do the time and space complexity analysis of the popular K-means algorithm. As of yet, we plan to implement it in the Numpy or Dask framework of Python.

Implementation in Dask is better in sense because it's much more scalable and parallel computing can also be done easily with Dask, which helps us in running onto the full capacity of a machine and thus is much more scalable to the current techniques. But, Dask is a newly built framework when compared to NumPy and hence the community is not that huge. As a result, the tutorials for Dask are also limited. Thus, we might shift the implementation in Numpy if we get stuck.

We plan to get an insight into the world of Data Science through this project. We will analyze different techniques that might help us improve our algorithm and get better results in lesser time. We would also like to evaluate and document the cases in which the algorithm might fail.

## Future Goals

- We would like to dwell deep into data science later on, and solve some real life problem with the application of K-means, DPMM, Spectral Clustering or similar algorithms whose selection depend on the type of the dataset.
- We might even like to experiment with how the changes in the algorithm affect the optimality of the same by doing runtime analysis on different datasets.

## Courtsey

Wikipedia : https://en.wikipedia.org/wiki/K-means_clustering

Andrea Trevino's Blog : https://www.datascience.com/blog/introduction-to-k-means-clustering-algorithm-learn-data-science-tutorials