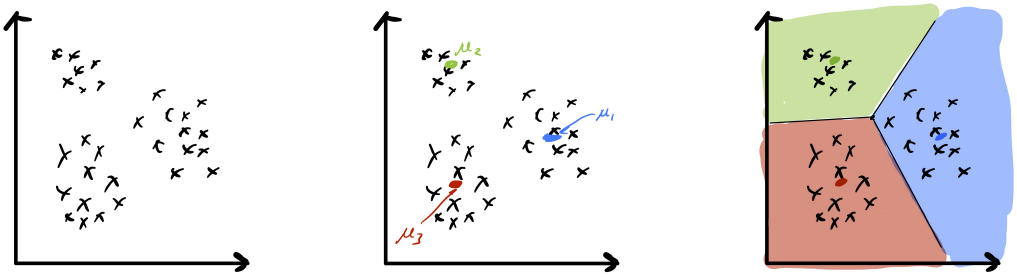
K-means

Prof. Luca Formaggia, Matteo Caldana

September 22, 2023



1 Introduction

K-means [1] is a clustering algorithm that tries to partition a set of points into K sets (clusters) such that the points in each cluster tend to be near each other.

2 Objectives

Implement the k-means clustering algorithm (also called Lloyd's algorithm) in parallel with MPI and the filtering algorithm that uses kd-trees to speed up each k-means step [2]. Your code should be able to read the data from a CSV file and process it in parallel. Choose a large dataset to test your algorithm. Run strong and weak scalability test and report the results.

3 Ideas for an exam project

K-means has emerged as powerful tools for uncovering patterns, communities, and structures within complex networks (*graphs*). For instance, you could apply k-means to perform mesh agglomeration of a complex geometry, such as the human brain.

References

[1] https://en.wikipedia.org/wiki/K-means_clustering
[2] Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. IEEE transactions on pattern analysis and machine intelligence, 24(7), 881-892. (https://www.cs.umd.edu/ mount/Projects/KMeans/pami02.pdf)