# Clusterisation

## Flat

## Whats wrong with kmeans++ -

* Need k-passes on data, application with large data, k can be something like a 100 or even 1000.

## Intuition for a solution -

After it chooses the first center it has a distribution for the data points and it picks one and chooses one point from the joint distribution and picks a point and updates the distribution.

### So what Bahman (author) says is that instead of this we want to over sample the data -

* We want to pick each point independently so we do not want to choose points from a joint distribution we pick each point independently  
  This is pretty useful for distributed implementation because each point independently gets to decide if it should be in the set of centres or not.  
  This intuitively corresponds to updating your distribution much more infrequently which is a much more coarser sampling .So the way it works is that if you have a distribution you pick a lot of points from it and then update your distribution and then from the distribution you get after this step you again pick a lot of points.

The above algorithm is k means parallel.  
It may not be clear even from the algorithm that this is actually sufficient but they offered a proof for its soundness.

[**@zermelozf**](https://github.com/zermelozf) I would request you for any valuable inputs since you gave some time into reading and implementing this. Thanks a lot.

## Hierarchical