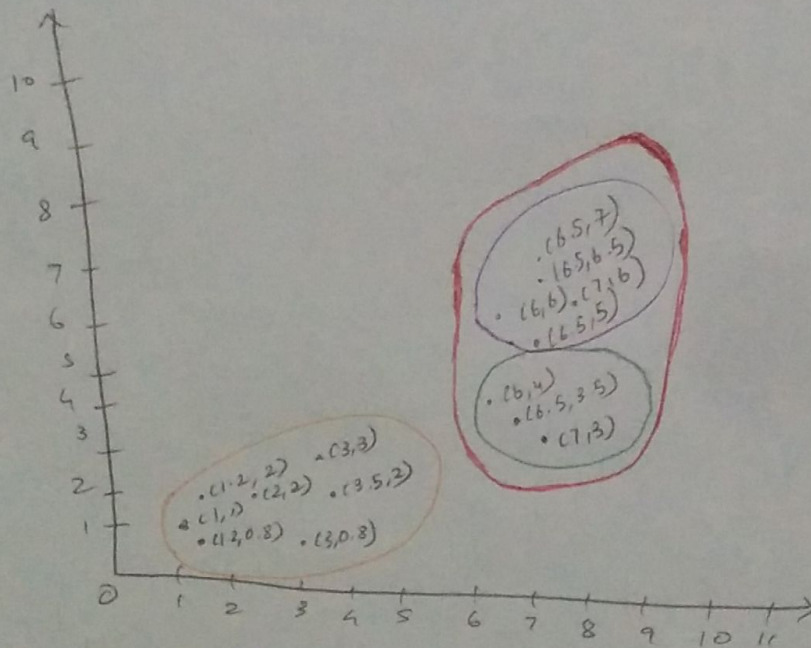## K-MEANS Implementation

→ ~~Basic K-Means simple~~
~~implementation~~ ✓

→ ~~Implementation in Map Reduce Hadoop~~ ✓

→ ~~Hadoop installation~~ ✓

→ shell script for different k-values

→ Python Visualization for optimal k value
(Elbow theorem)

→ ~~Test run/validate everything with different datasets~~

Sample Data

k = 2 or k = 3 ?
which is the optimum k

Basic Implementation

→ Read csv file ✓
→ Array list of points (input) ✓
→ choose random k points as center. ✓
→ Distance bet. each point + center to be
  calculated. (Euclidean distance) ✓
→ clusters with ((center, points) ✓
→ recompute new centers. ✓
→ iterate till convergence ✓

→ Extending Basic Implementation to Map Reduce

→ Map reduce installation.

→ Write Mapper class extending MapReduceBase

→ write reducer class

→ Test for smaller input

→ Understand example wordcount program.

Map Reduce framework (For k=3)

Distributed system
centroids file (with values)

Mappers

Input file chunk centroids

Input file chunk centroids

Input file chunk centroids

Input file chunk centroids

Input file chunk centroids

Distance calculator

(center, point)

(center, point)

(center, point)

(center pt)

(center point)

Reducers

center1, pt1, pt2

center2, pt1, pt2

center3, pt1, pt2

clustering

new centers

After iteration stops.

Output file.