Introduction to Data Science

Course 094219

Lab 3:

K-Means

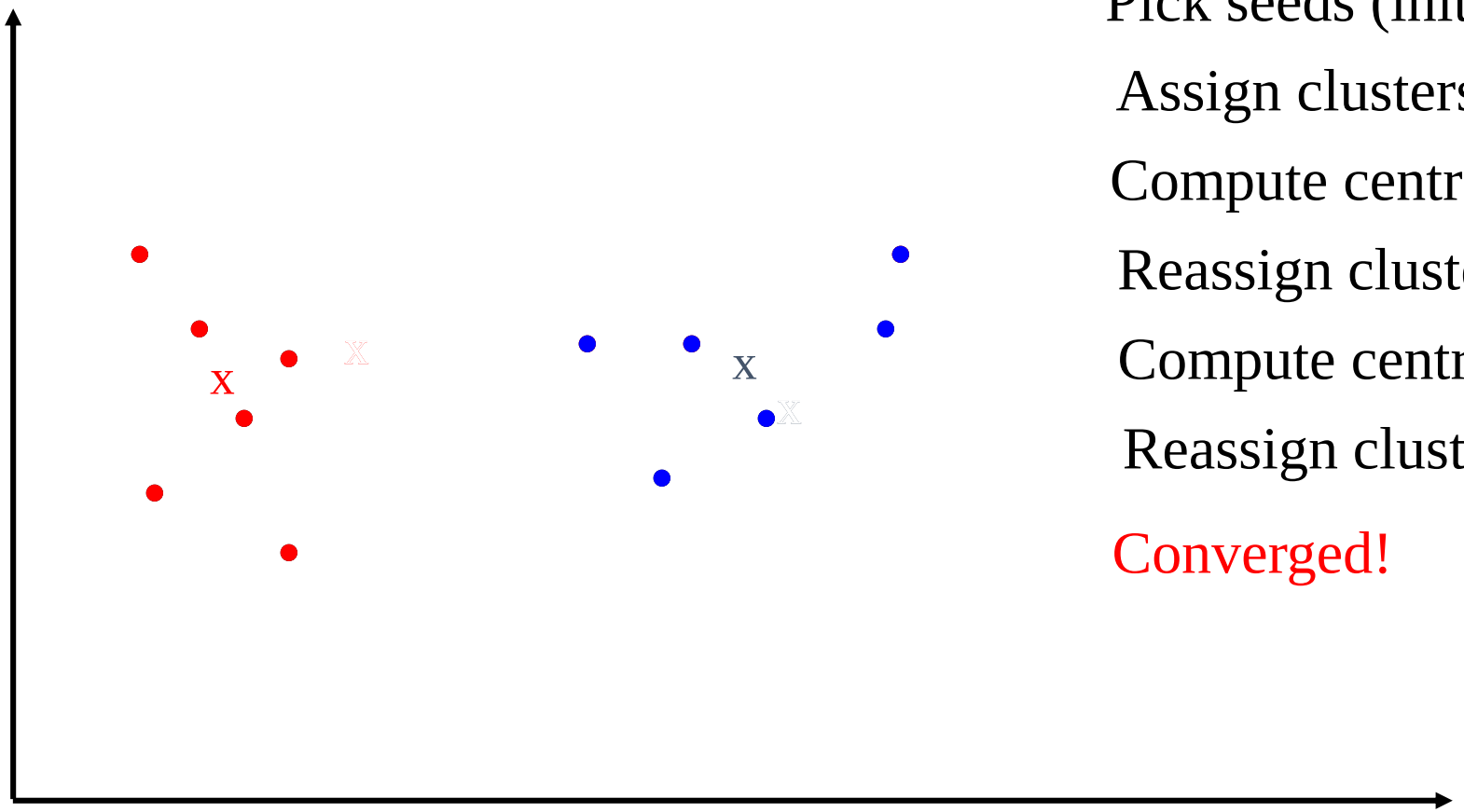
Spring 2017

K-Means

- Partitional clustering method
- Items are represented as points in space (we points/items terminology interchangeably)
- Divide items to K clusters iteratively, until we find a partition that doesn't change
  - Each cluster is associated with a centroid (the arithmetic mean of clusters' items)
  - Each item is assigned to the cluster with the closest centroid
  - After the assignment the centroids are not necessarily correct, thus updated
  - Next we again assign the items to the updated centroids and so on

```
Select K points as the initial centroids.
repeat
    Form K clusters by assigning all points to the closest centroid.
    Recompute the centroid of each cluster.
until The centroids don't change
```

2D Example



Pick seeds (initial centroids)

Assign clusters

Compute centroids

Reassign clusters

Compute centroids

Reassign clusters

Converged!

The dataset and the code

- The code and the data can be found at:

/mnt/share/students/LAB3

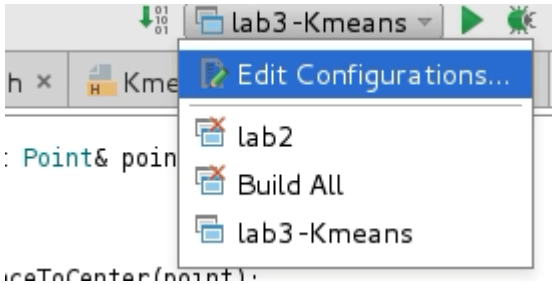
- **Copy everything to your local folder and unzip the code**:

tar –xvzf **lab3_Kmeans.tgz**

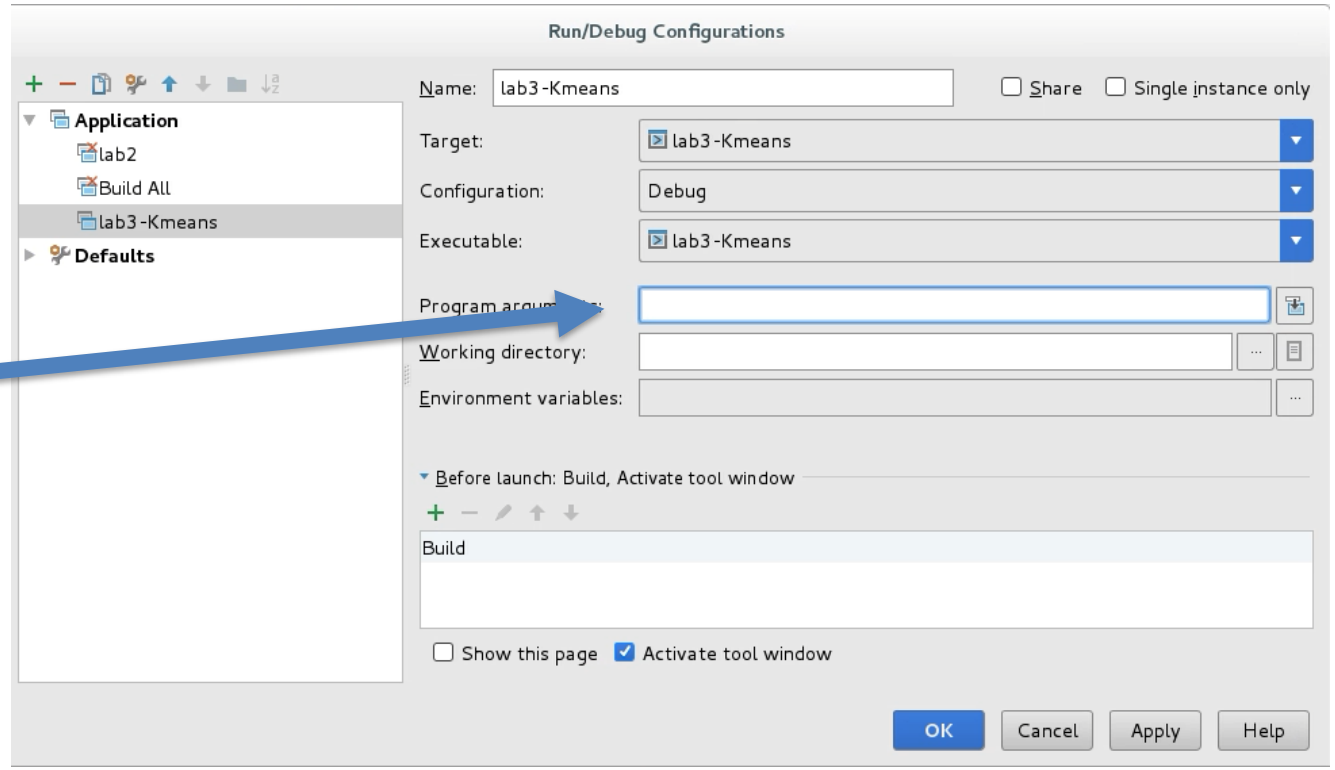
- The file **in1.txt** is a very simple, small input to the algorithm, you can use it to check it is working (after you add the missing functions), but this doesn't replace exhaustive testing
- The file **color_dataset_ready.txt** contains colors' names and their RGB values. The dataset was created by showing a person a color and asking him to name it. If we cluster the data based on RGB values each cluster will contain names of similar colors
- The clusters can be used to disambiguate the common color name based on its RGB value or to find synonyms for different color names

Assignment – Let's start

- What are arguments for main and how do we set them in CLion?



- Here you can edit the arguments for your program (see next slide)



Assignment – Arguments

- The arguments for the program are:
  - K : number of clusters.
  - max_iterations : set limit for the number of iterations.
  - has_name : does the data file contains names for each point.
  - fileName : Path for your file.
- The main class of the project is the KMeans class
- Try to identify which methods in the class KMeans are responsible for which steps in the algorithm
- The header files of the classes Cluster and Point are fully provided to you, try to see which implementations are missing and what is their purpose

Assignment – Complete Class Point

- Implement the missing methods for the class Point.
- For example, the class Point has Euclidean distance method

$$dist(x, y) = ||x - y|| = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

- Try to implement it yourself.

Assignment – Complete Class Cluster

- Implement the missing methods for the class Cluster.

- For example, getDistanceToPrototype calculates the distance from the input point to the centroid of the cluster.

Assignment - Home

- This week there is no assignment to do at home ☺

- You will use your K-Means code in your homework.