Lecture 9

Scalable K-means

Haiping

http://www.dcs.shef.

COM6012 Scalable Machine Learning

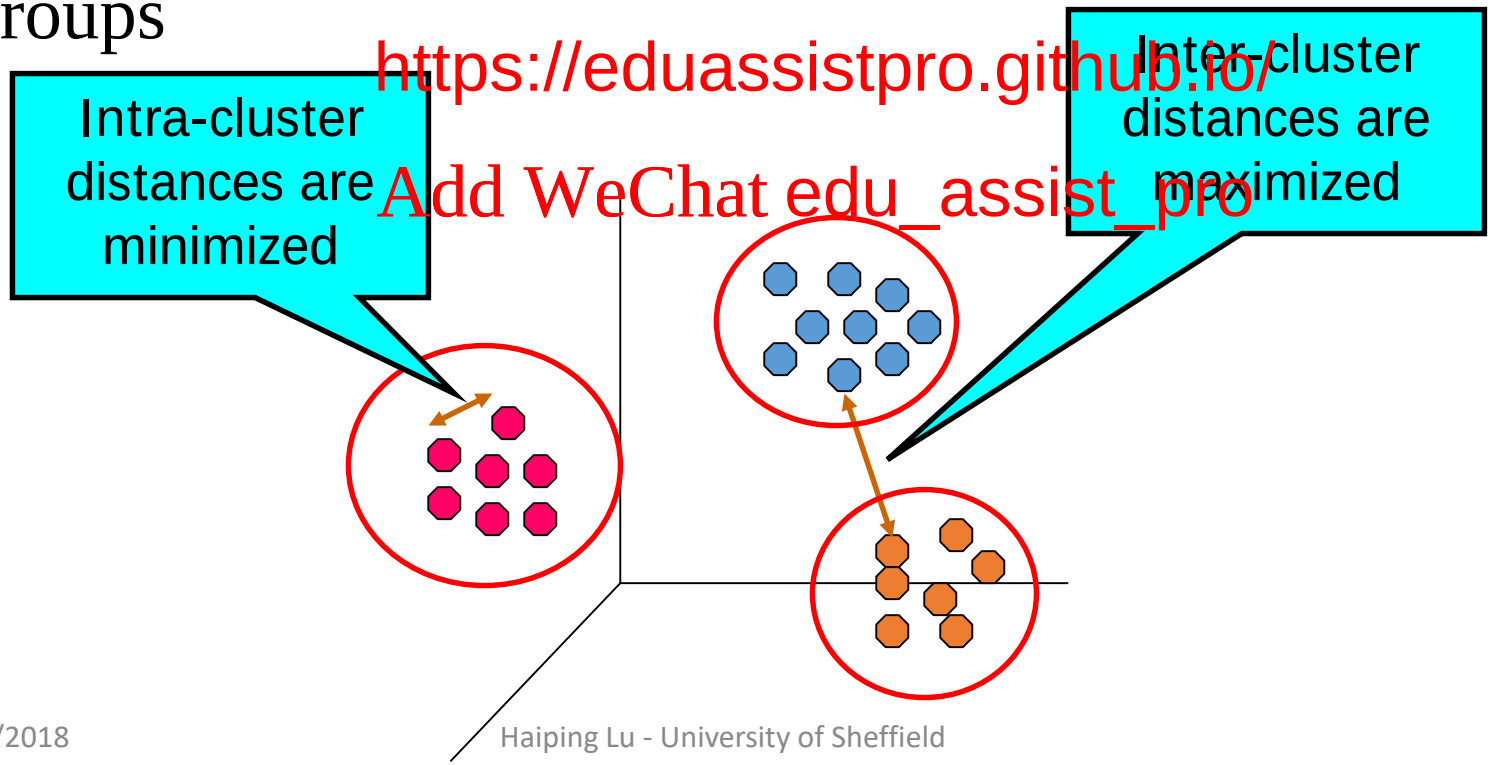
Spring 2018

Week 9 Contents

- **Introduction to Cluster Analysis**
- K-means Clu
- Scalable K-means
- Scalable K-means in Spark

What is Cluster Analysis?

- Finding groups of objects such that the objects in a group will be similar (or related) to one another and different from (or unrelated to) the objects in other groups



Cluster Analysis

- Divide data into (clusters) that are meaningful, useful, or both
- The study of t ically finding classes
- Clusters can help capture t tructure of the data
- A starting point to further analysis
- An important role in a wide variety of fields: psychology, biology, statistics, pattern recognition, information retrieval, machine learning and data mining, etc

Clustering for Understanding

- Classes, or conceptually meaningful groups of objects that share some similarities, play an important role in how people analyze the describe the world
- Human being g objects into groups (clustering) and assi icular objects to these groups (classification). E.g. children can quickly label the objects in a photograph as buildings, vehicles, people, animals, etc

Applications of Clustering

- Biology
  - Cluster analysis help create taxonomy of all living things: kingdom, phylum, class, order, family, etc
  - Cluster analys nnotate the function of genes / protein
- Information retrieval.
  - Clustering help group the search r al number of clusters, each of which captures a particular aspect of the query. E.g. a query of “movie” might return Web pages grouped into categories such as reviews, trailers, starts, and theaters
- Climate
  - Cluster analysis has been applied to find patterns in the atmospheric pressure of polar regions and areas of the ocean that have a significant impact on land climate

Applications of Clustering

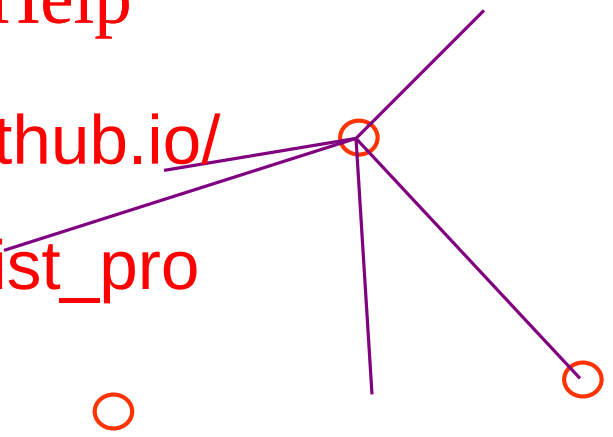
- Psychology and Medicine.
  - Identify different types of diseases (e.g. depression)
  - Detect patterns in the spatial or temporal distribution of a disease
  - Help group rns
- Business
  - Clustering analysis can be used to segment customers into a small number of groups for additional analysis and marketing activities
- **Anomaly/Outlier Detection (notebook/coursework data)**

Anomaly/Outlier Detection

- What are anomalies/outliers?
  - The set of data points that are considerably different than the remainder of the data
- Applications:
  - Credit card fraud detection:
  - Network intrusion detection
  - Ecosystem disturbances: typhoon, fire
  - Public health: SARS, bird flu, HxNx
  - Medicine: unusual symptoms/test results

Clustering-Based Anomaly/Outlier Detection

- Cluster the data into groups of different density
- Choose point as candidate
- Compute the distance between candidate points a non-candidate clusters.
- If candidate points are far from all other non-candidate points, they are outliers



About Cluster Analysis

- Cluster analysis groups data objects based only on information found in the data that describes the objects and their relationships
- The goal is th group be similar (or rel nd different from (or unrelated to) the object roups
- The greater the similarity (or homogeneity) within a group and the greater the difference between groups, the better or more distinct the clustering.

Week 9 Contents

- Introduction to Cluster Analysis
- **K-means Cl**
- Scalable K-means
- Scalable K-means in Spark

K-means Clustering

- A prototype-based, partitional clustering approach
- Each cluster is associated with a centroid (centre point)
- Each point is with the closest centroid
- Number of clusters, K, mus ied before clustering

K-means Clustering

- **Input:**
  - A set $X=\{x_1, x_2, \ldots, x_n\}$ of n data points
  - Number of clusters k
- For a set $C=\{$ "centres" define:

$$\varphi_X(C) = \sum_{x \in X} d(x,C)^2$$

where $d(x,C)$ = distance from x to closest centre in C

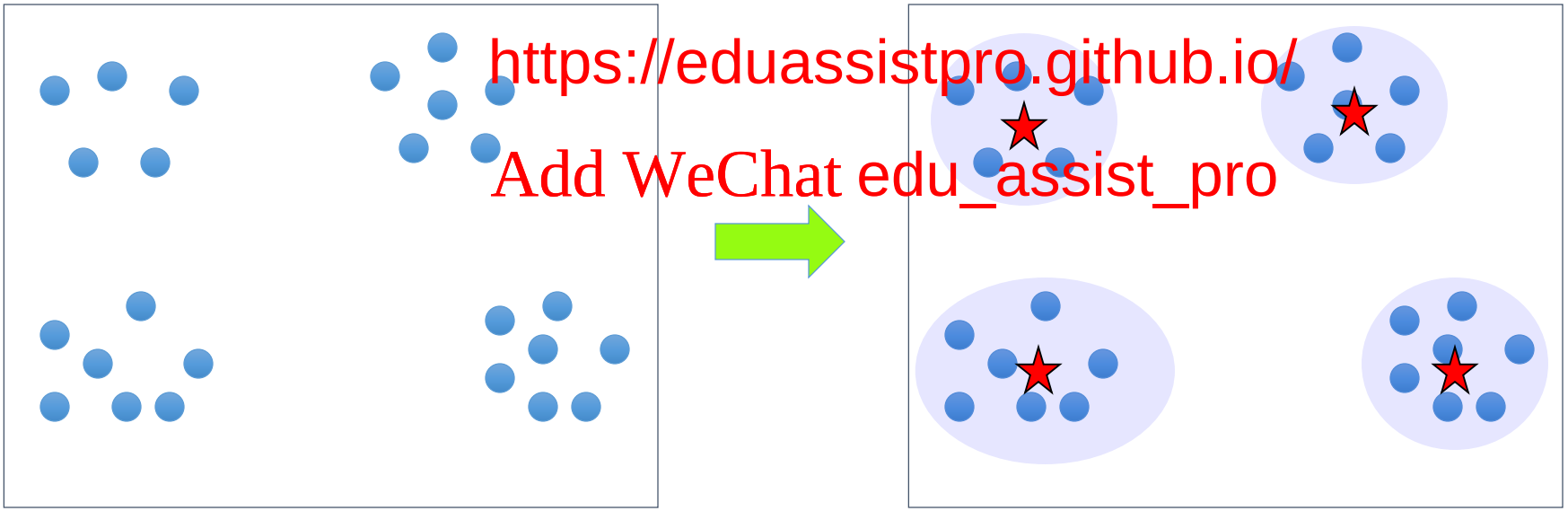
- **Goal:** To find a set C of centres that minimizes the objective function $\varphi_X(C)$

Determine the number of clusters

There are different approaches of determining K

- K can be arbitrarily set as any number
- K can be deter e need of further analysis
- K can be determined accord knowledge, or the knowledge obtained d visualisation
- Different K's can be initially set, and find the best K using some criteria

K-means Clustering: Example



K = 4

Lloyd Algorithm

- Start with k arbitrary centres $\{c_1, c_2, \ldots, c_k\}$ (typically chosen uniformly at random from data points)
- Performs an [illegible] convergence
- Main advantages: Simplicit[illegible]ity (iterations)

1: Select K points as the initial centroids.
2: **repeat**
3: Form K clusters by assigning all points to the closest centroid.
4: Recompute the centroid of each cluster.
5: **until** The centroids don't change

What's wrong with Lloyd Algorithm?

- Takes many iterations to converge
- Very sensitive to initialization
- Random initi two centres in the same clus
  - K-means gets stuck in a loc

Lloyd Algorithm: Initialization

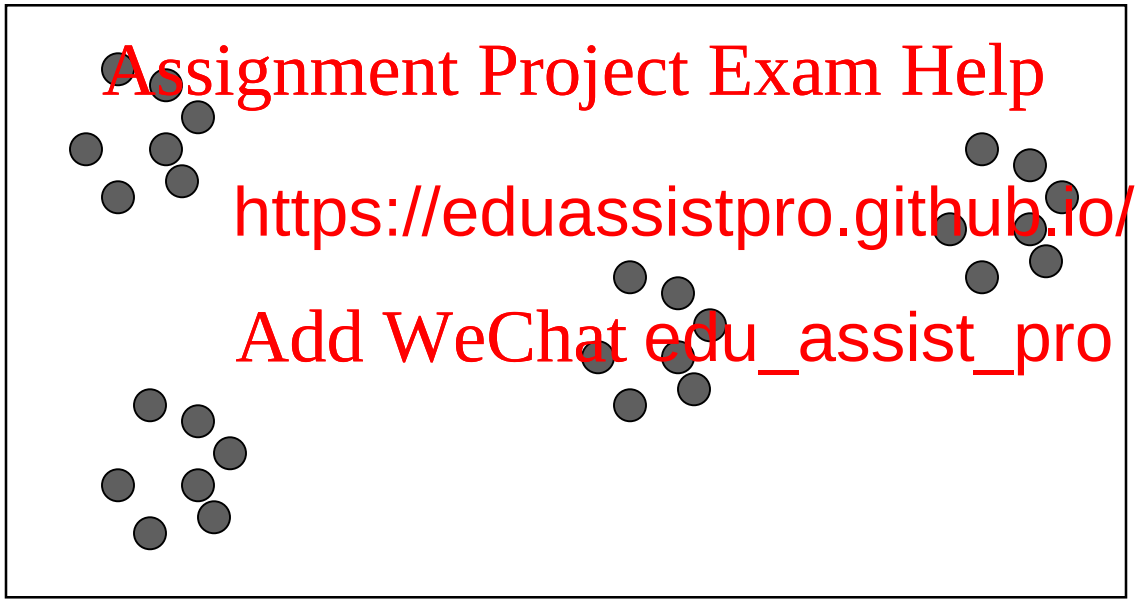


Figure credited to David Arthur

Lloyd Algorithm: Initialization

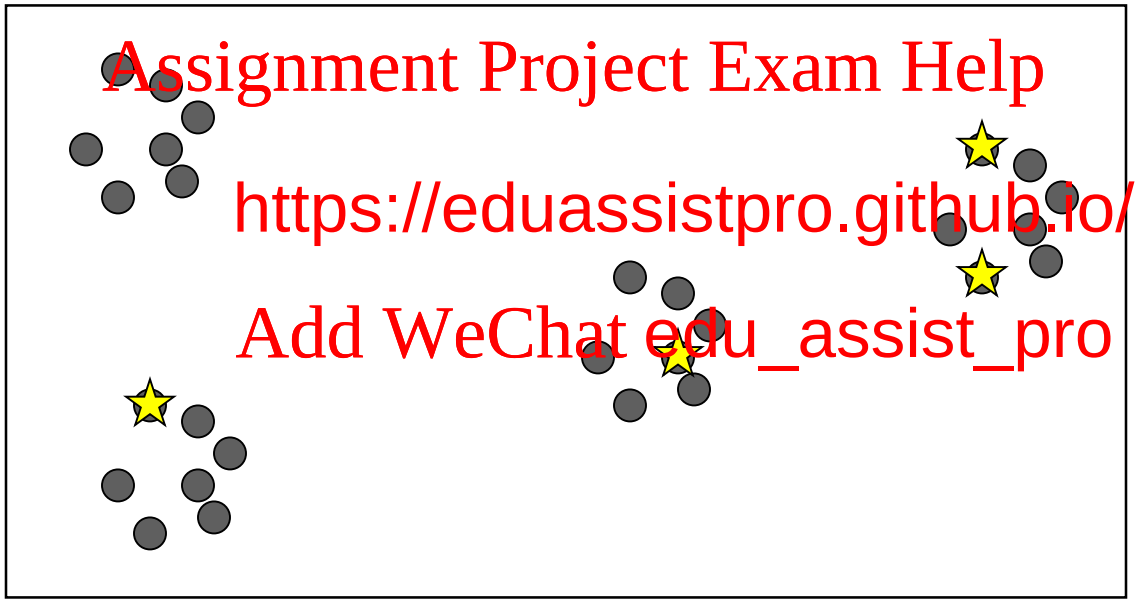


Figure credited to David Arthur

Lloyd Algorithm: Initialization

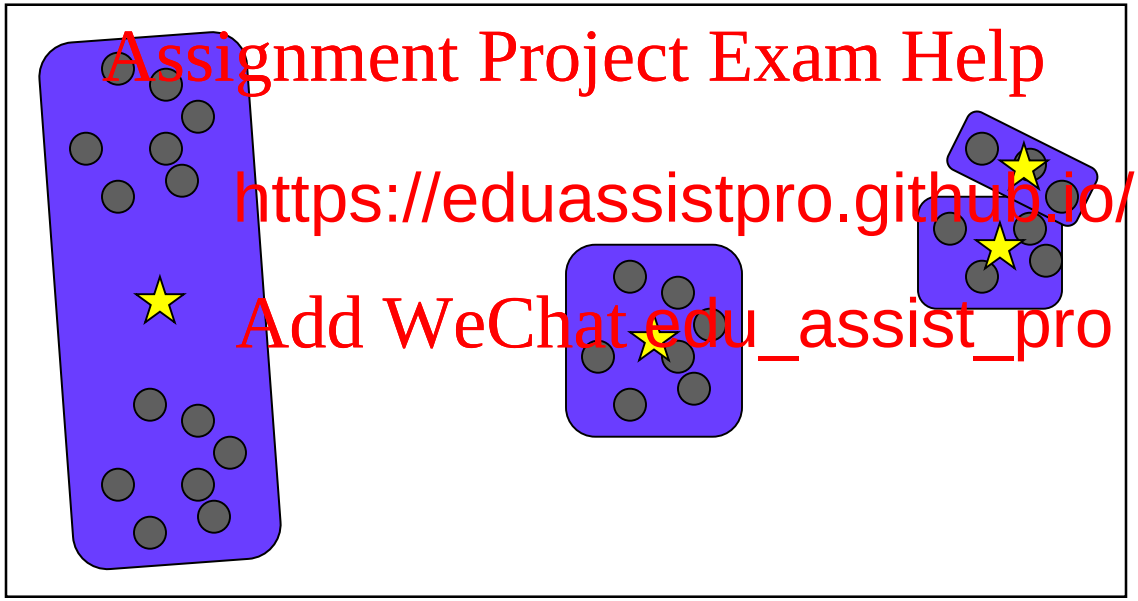


Figure credited to David Arthur

Lloyd Algorithm: Initialization

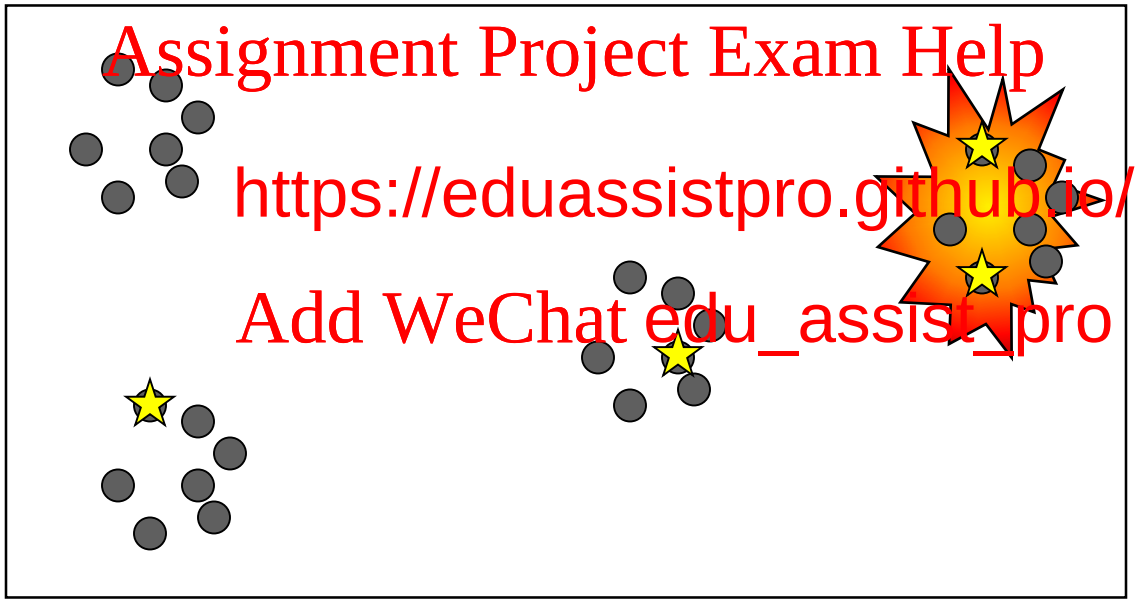


Figure credited to David Arthur

Week 9 Contents

- Introduction to Cluster Analysis
- K-means Clu
- **Scalable K-means**
- Scalable K-means in Spark

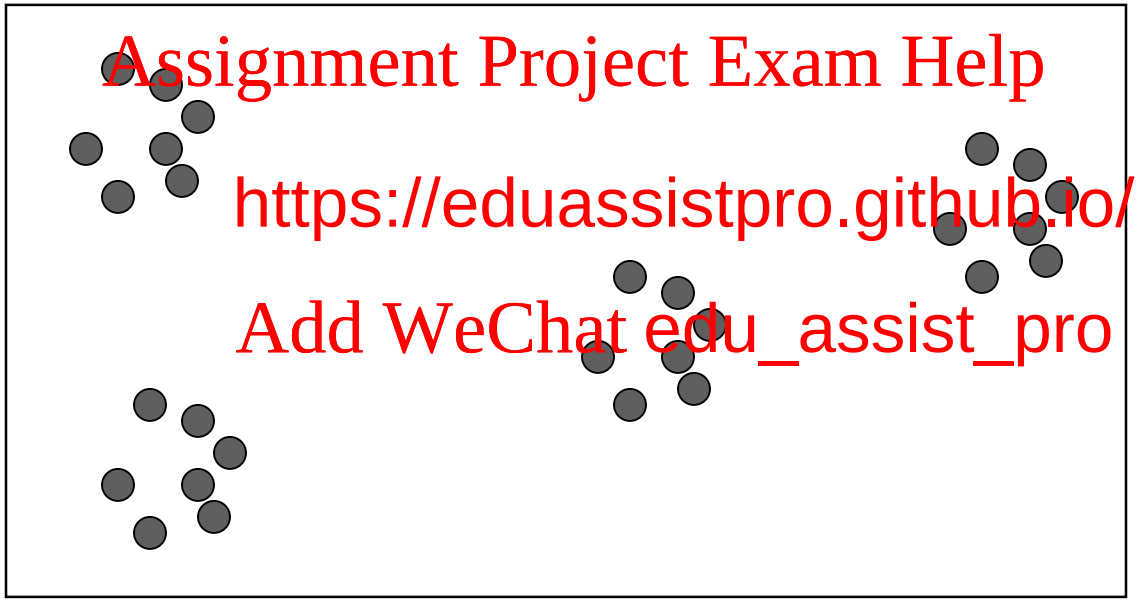
K-means++ [Arthur et al. ’07]

- Spreads out the centres
- Choose first centre c_1 uniformly at random from the data set
- Repeat for 2
  - Choose c_i to be equal to a da ampled from the distribution:

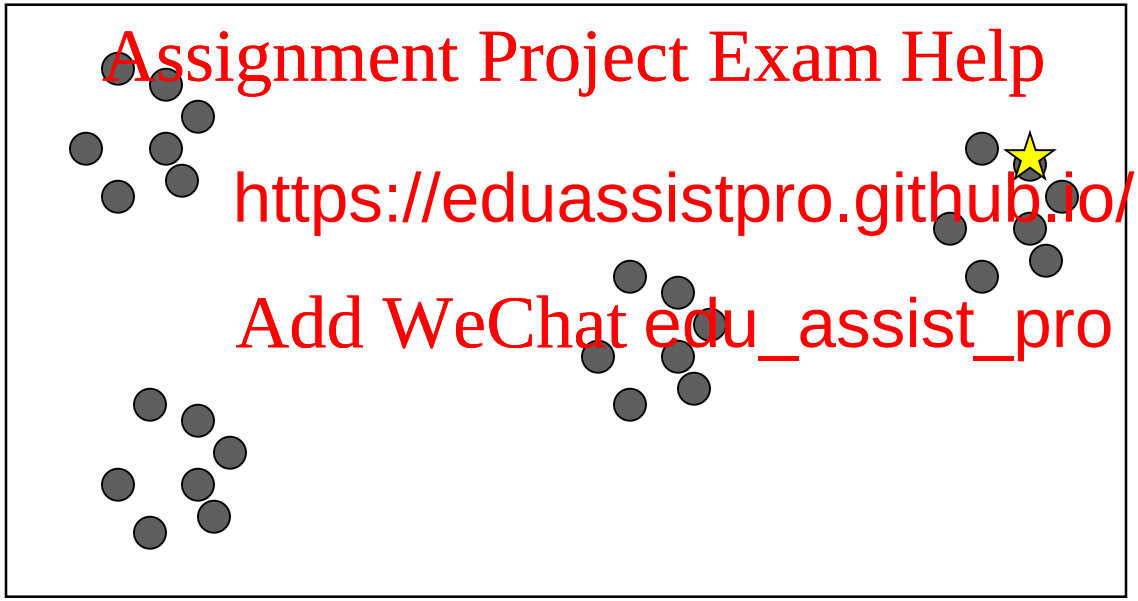
$$\frac{d(x_0, C)^2}{\varphi_X(C)} \propto d(x_0, C)^2$$

- **Theorem:** O(log k)-approximation to optimum, right after initialization

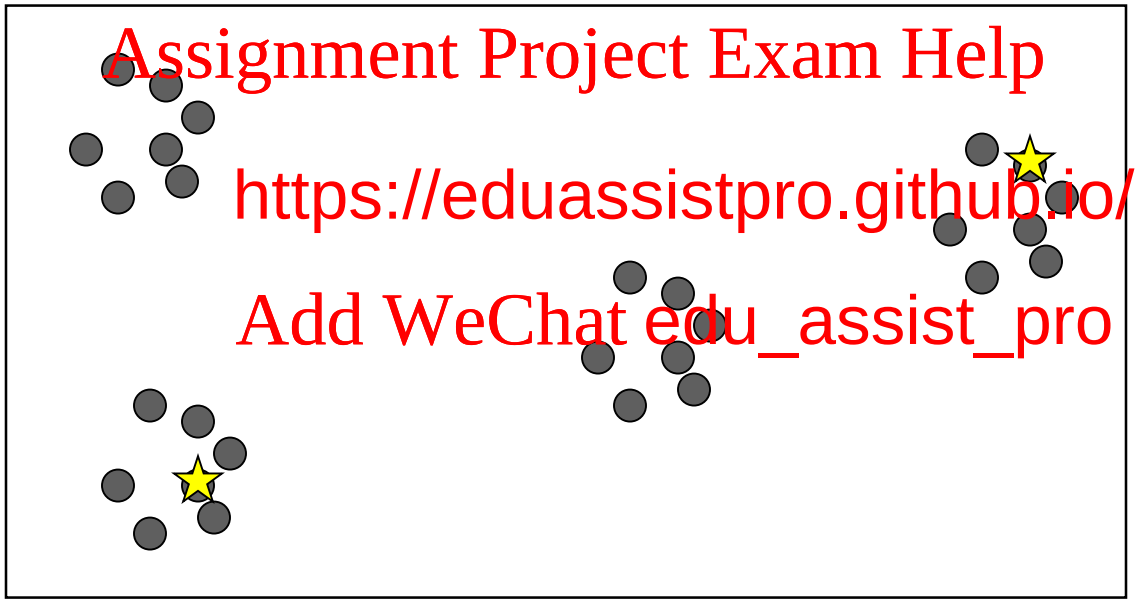
K-means++ Initialization



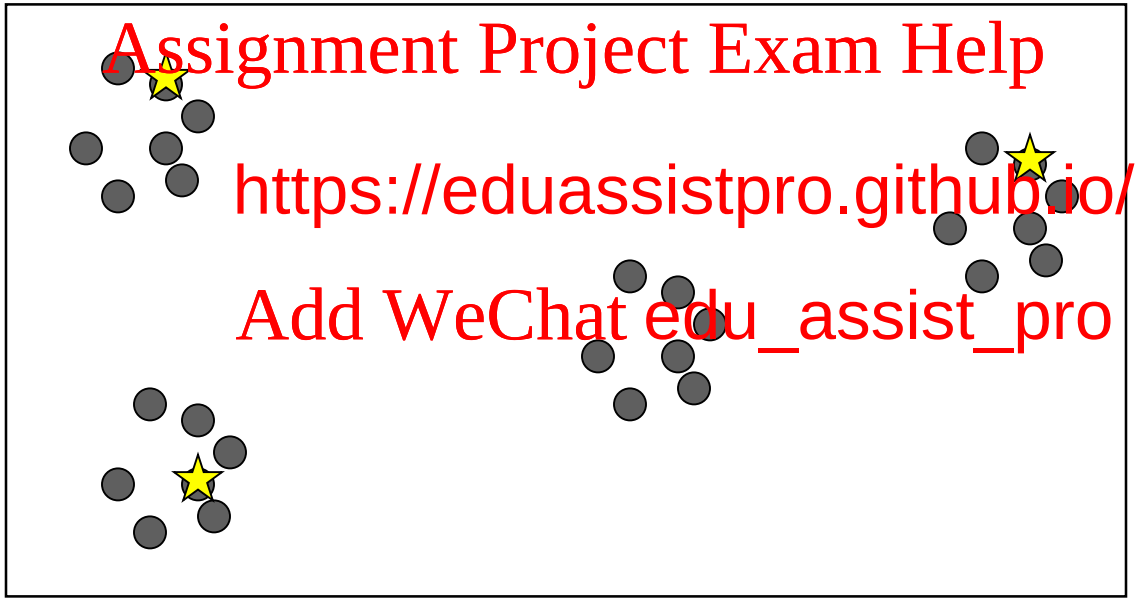
K-means++ Initialization



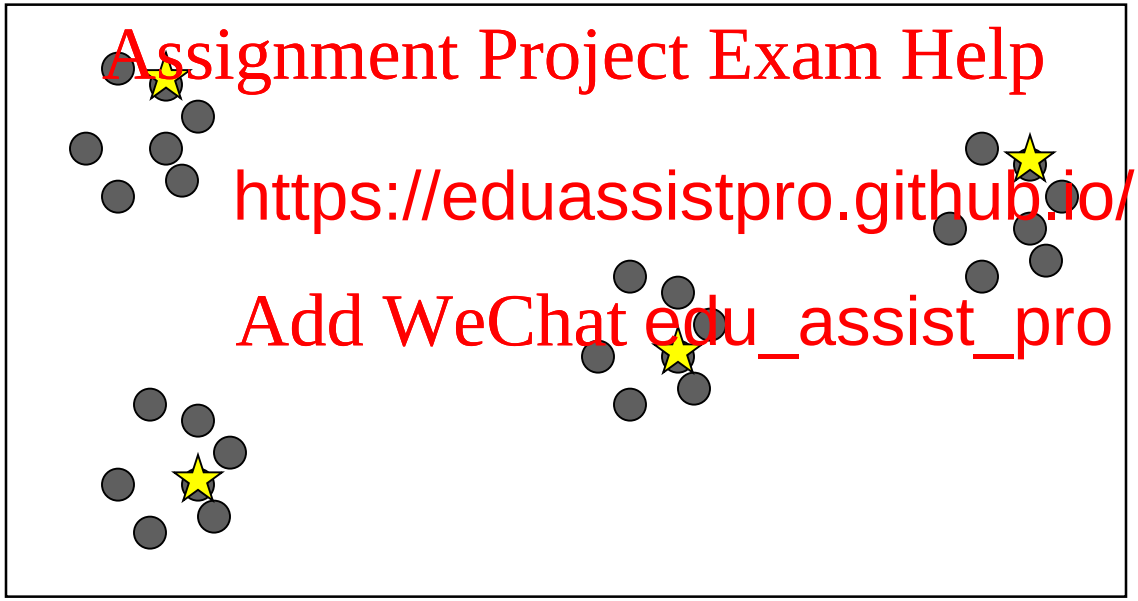
K-means++ Initialization



K-means++ Initialization



K-means++ Initialization



What's Wrong with K-means++?

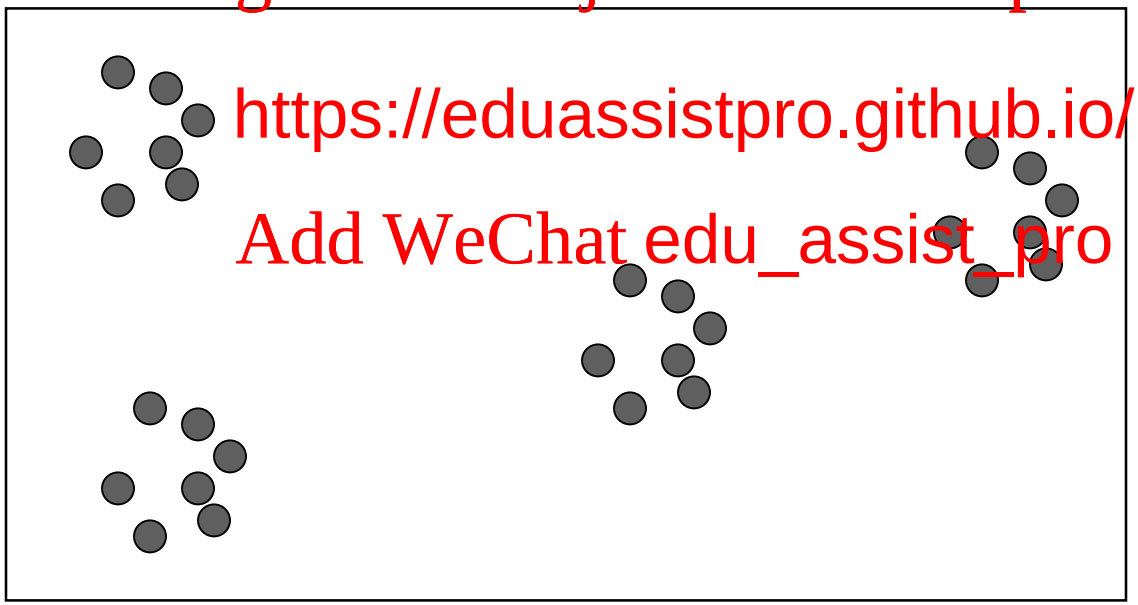
- Needs K passes over the data
- In large data applications, not only the data is massive, but [illegible] e (e.g., easily 1000).
- Does not scale!

Intuition for a Solution

- K-means++ samples one point per iteration and updates its distribution
- What if we **o** g each point independentl ity?
- Intuitively equivalent to up distribution much less frequently
  - Coarser sampling
- Turns out to be sufficient: K-means||

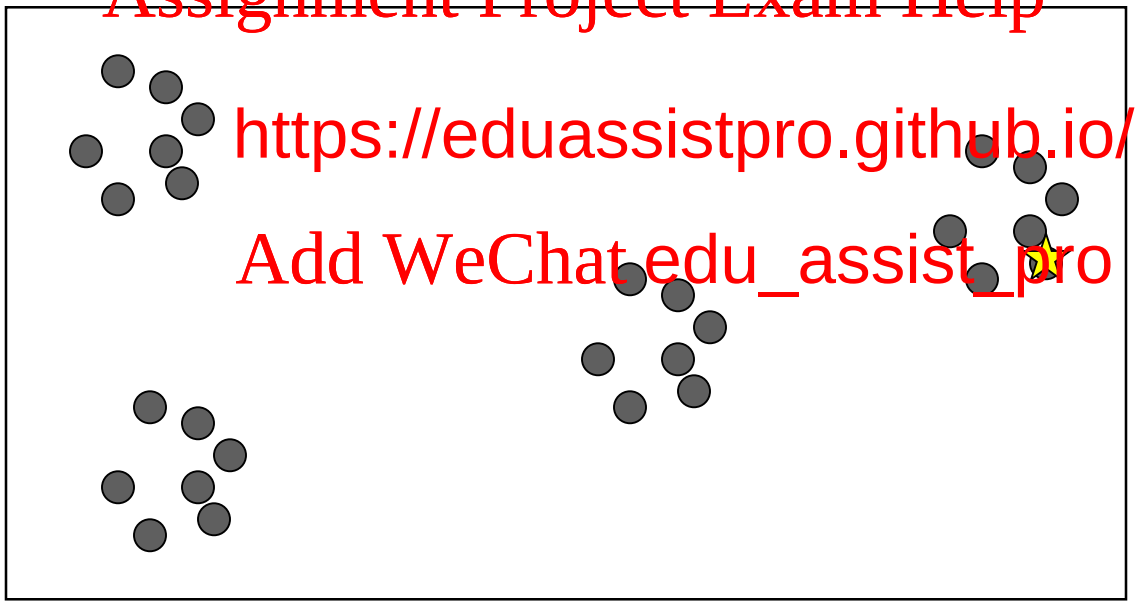
K-means|| Initialization [Bahmani et al. ’12]

K=4,
Oversampling factor L=3



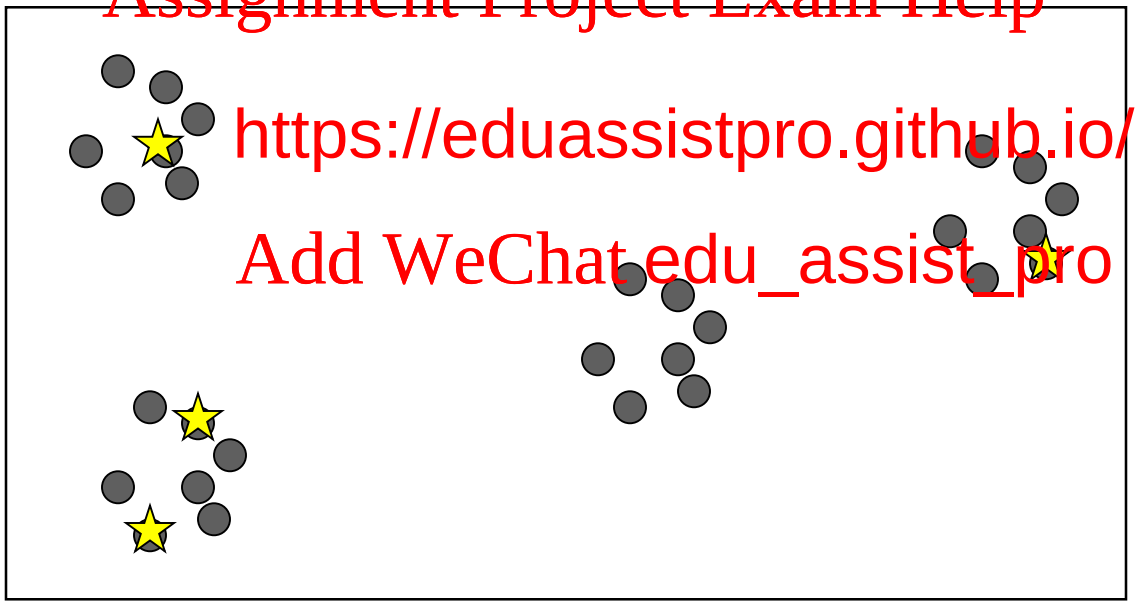
K-means|| Initialization

K=4,
Oversampling factor L=3



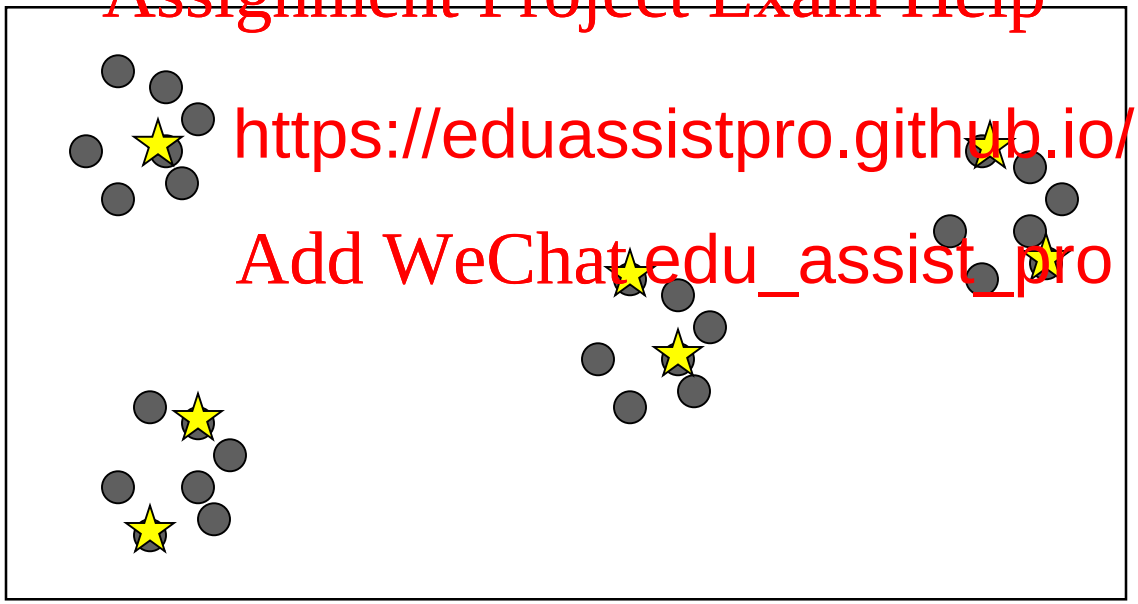
K-means|| Initialization

K=4,
Oversampling factor L=3



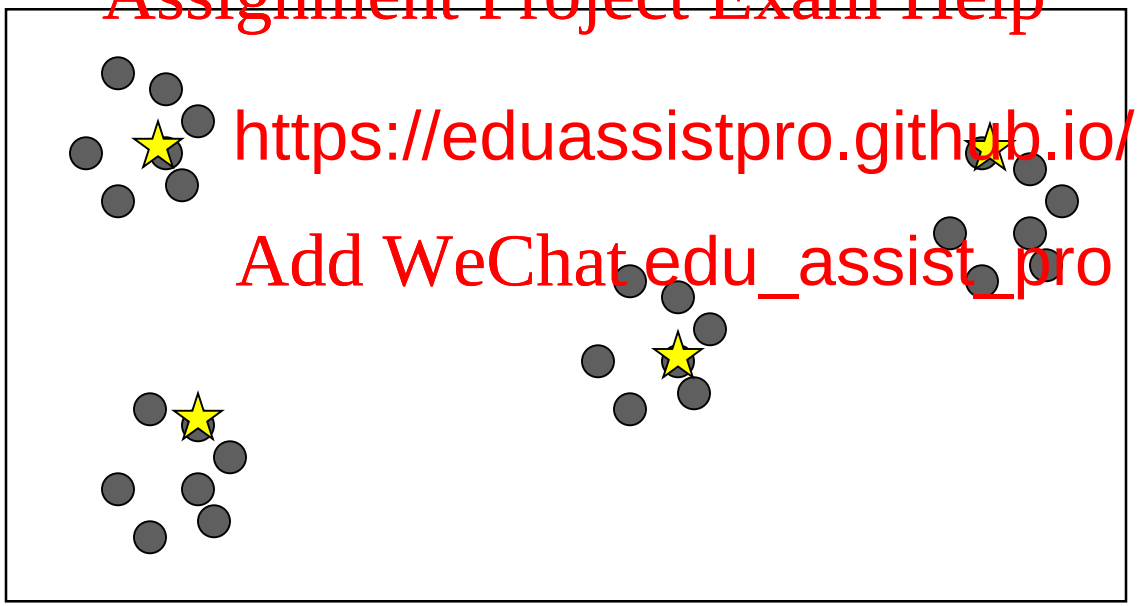
K-means|| Initialization

K=4,
Oversampling factor L=3



K-means|| Initialization

K=4,
Oversampling factor L=3



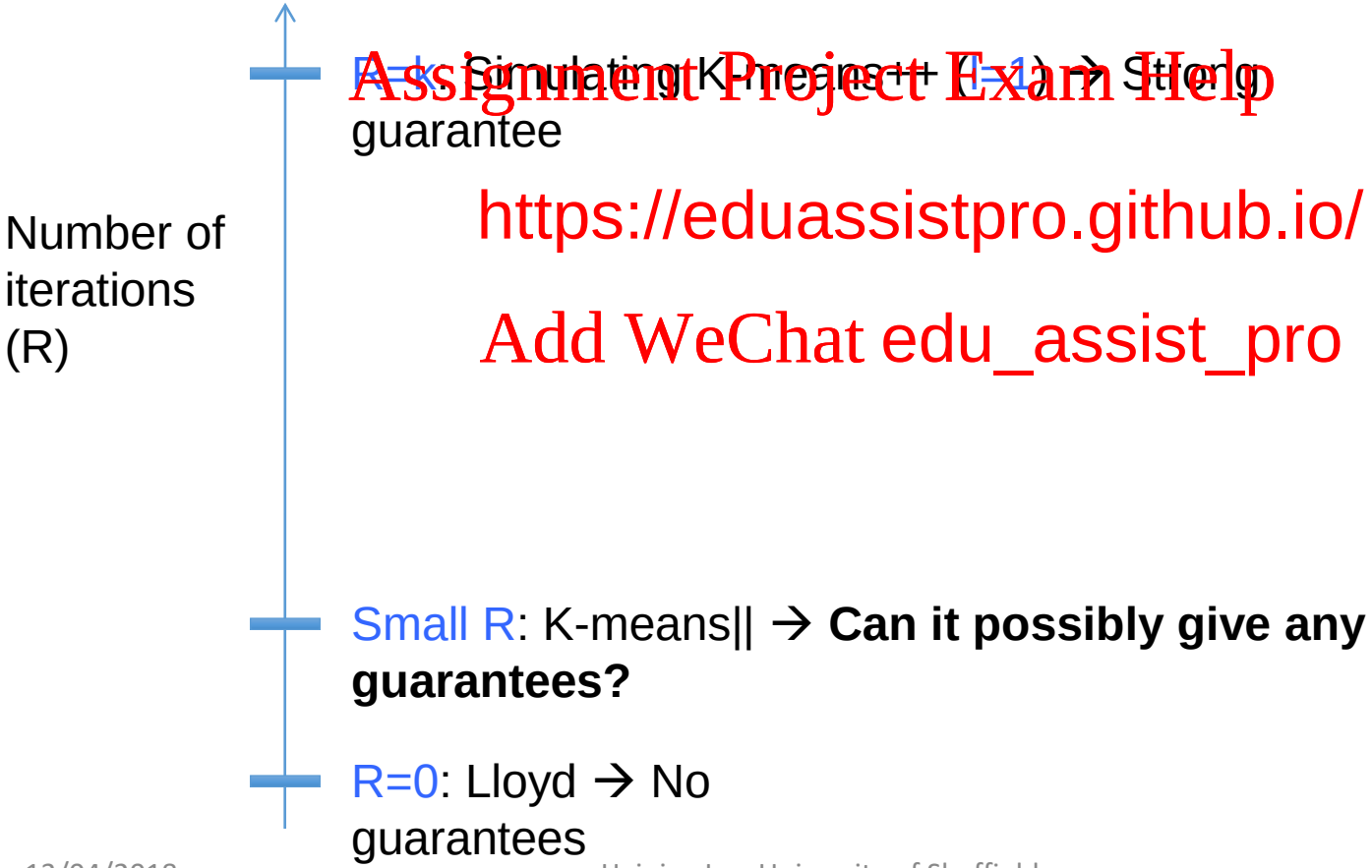
Cluster the intermediate centres

K-means|| [Bahmani et al. ’12]

- Choose L>1
- Initialize C to an arbitrary set of points
- For R iteratio
  - Sample eac ly with probability $p_x = Ld^2(x,C)/\phi_X(C)$
  - Add all the sampled points t
- Cluster the (weighted) points in C to find the final k centres

K-means||: Intuition

- An interpolation between Lloyd and K-means++



K-means||: Benefits

- Using K-means++ for clustering the intermediate centres, the overall approximati)
- K-means|| eans++ to get confused with noisy
- K-means|| reduces number of Lloyd iterations even more than K-means++

Week 9 Contents

- Introduction to Cluster Analysis
- K-means Clu
- Scalable K-means
- **Scalable K-means in Spark**

K-means in MLlib (notebook)

- Not scalable: Kmeans
- Scalable: Kmeans | | (default)
- Code: https://github.c...0/mllib/src/main/scala/org/apache/s...ans.scala
- Documentation: https://spark.apache.org/docs/2.1.0/api/scala/index.html#org.apache.spark.mllib.clustering.KMeans
- https://spark.apache.org/docs/2.1.0/mllib-clustering.html

K-means in Mllib (notebook)

- *k:* the number of desired clusters.
- *maxIterations:* the maximum number of iterations
- *initializationMo* m initialization or initialization vi
- *runs*: no effect since Spark 2.0.
- *initializationSteps*: determines t f steps in the k-means|| algorithm (default=2, advanced)
- *epsilon:* determines the distance threshold within which we consider k-means to have converged
- *initialModel:* manually set cluster centres for initialization

K-means in ML

- An Estimator
- Uses MLlib Kmeans (Kmeans | |)
- Code: https://github.com/apache/spark/blob/master/mllib/src/main/scala/org/apache/spark/mllib/clustering/KMeans.scala
- Documentation: https://spark.apache.org/docs/2.1.0/api/scala/index.html#org.apache.spark.ml.clustering.KMeans
- https://spark.apache.org/docs/2.1.0/ml-clustering.html

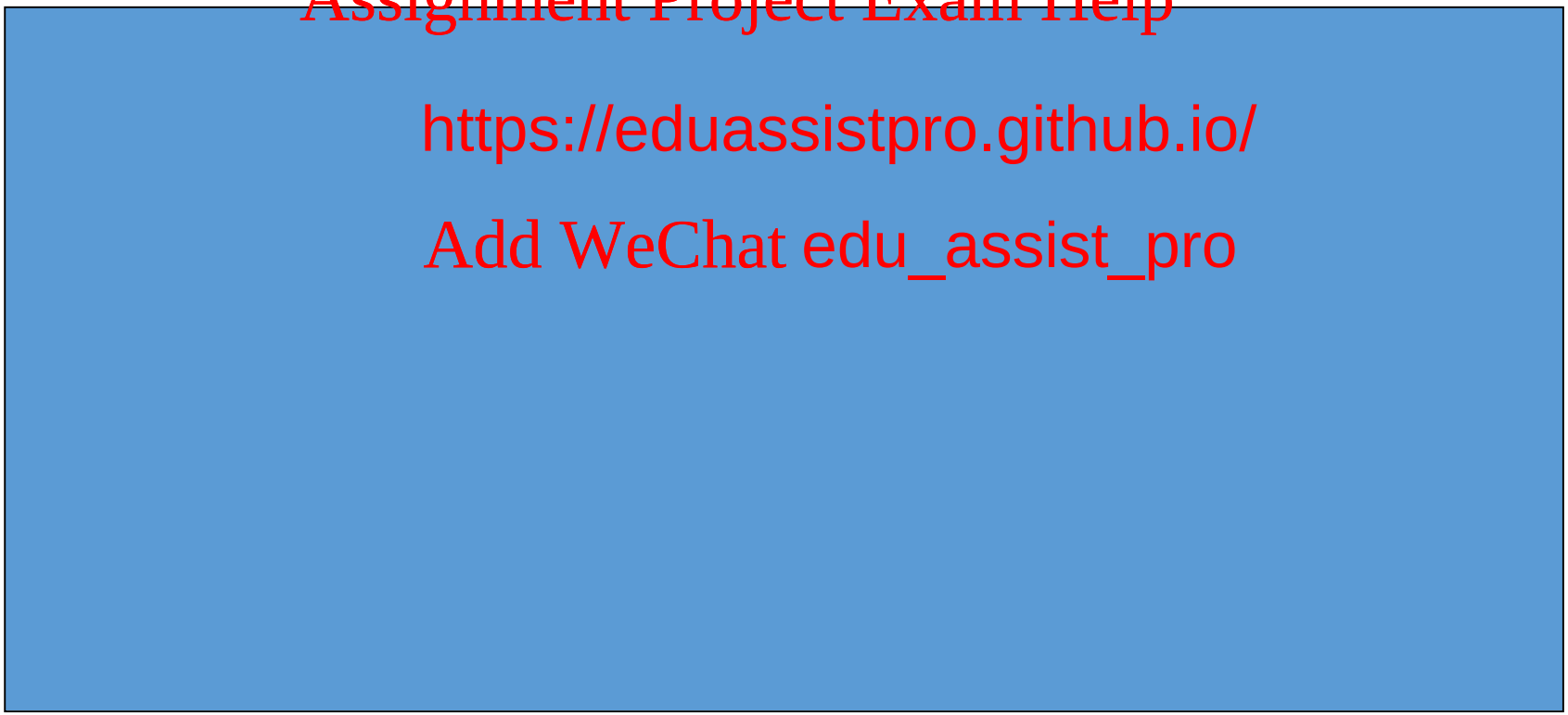
K-means in ML

- *k:* the number of desired clusters.
- *maxIter:* the maximum number of iterations
- *initMode:* speci ation or initialization via k-means|| (c
- *initSteps*: determines the numbe the k-means|| algorithm (default=2, advanced)
- *tol:* determines the distance threshold within which we consider k-means to have converged.
- *initialModel:* manually set cluster centres for initialization

Simplified from those for K-means in MLlib

Running Scalable K-means

- RDD should be cached for high performance (check warning when you run your program)



K-means++ in Spark

- Code: https://github.com/apache/spark/blob/master/mllib/src/main/scala/org/apache/spark/mllib/clustering/LocalKMeans.scala

Remark

- Acknowledgement
  - Some slides are adapted from the K-means|| slides by Bahman Bahmani, Stanford University, 2012
- References
  - Chapter on cluster pages): https://www-users.cs.umn.edu/ ering.pdf
  - K-means overview K_means%2B%2B
  - K-means ++ paper: http://ilpubs.stanf 78/1/2006-13.pdf
  - K-means || paper: http://dl.acm.org/cit =2180912.2180915
  - Spark ML: https://spark.apache.org/docs/2.1.0/api/scala/index.html#org.apache.spark.ml.clustering.KMeans
  - Spark MLLib: https://spark.apache.org/docs/2.1.0/api/scala/index.html#org.apache.spark.mllib.clustering.KMeans