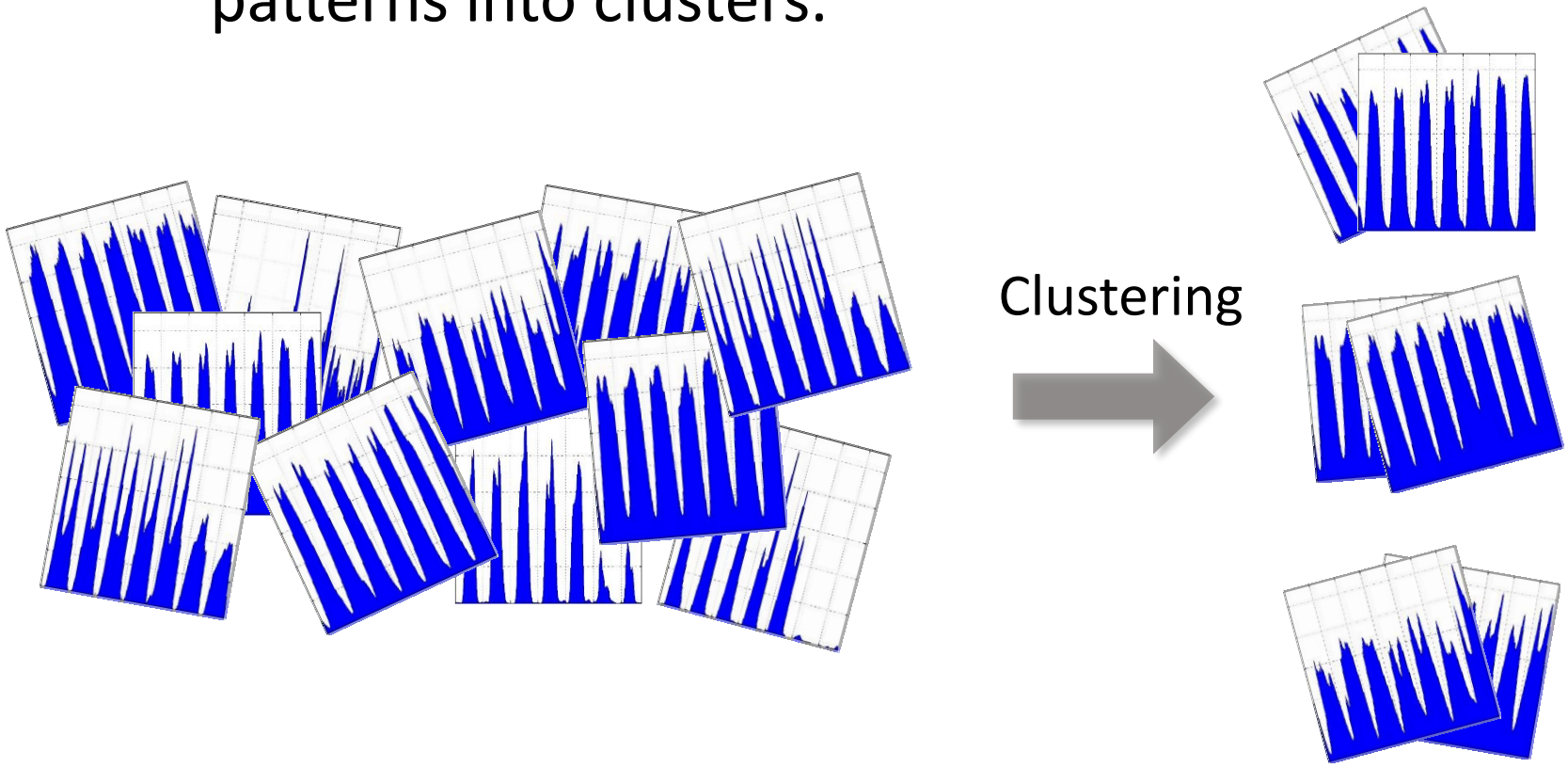
Homework 2: Clustering

Mobile Data Mining

Spring 2019

Goal

- Implement clustering algorithms on mobile big data
  - Group base stations (BSs) with similar traffic patterns into clusters.



Data

- Time distribution of BSs obtained in the experiments #2 of homework 1
  - Traffic volume of BSs in different time-bins (1 hour)
  - Each BS is described by a 744-sized vector (31days x 24 hour)

- Folding: Convert the traffic of one month to a week (7 days) by averaging
  - You can use traffic of 28 days in the folding (remove 3 days)

7x24 vector

- Normalization: Each vector is normalized by its z-score

$$\text{Normalized}(e_i) = \frac{e_i - \bar{E}}{\text{std}(E)}$$

traffic(第一个小时)-平均值/标准差

  - Where $\bar{E} = \frac{1}{n}\sum_{i=1}^{n} e_i$, $\text{std}(E) = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(e_i - \bar{E})}$

Experiments

- Implement one clustering algorithm introduced in the class on the BSs (regardless of language and platform)
  - Agglomerative hierarchical clustering
  - K-means
  - BFR
  - CURE

- Virtualization
  - Plot the traffic volume of the centroid or clustroid of each obtained cluster

Bonus

- Implement SVD to the vectors to reduce their dimensionality before clustering (5 points)

- Implement the clustering algorithm in the experiment by Spark (5 points)

Submission

- Submit this homework before April 28th. (Hard Deadline, please keep in mind)

- Submit as .zip file, including:
  - 1) A word document, Including:
    - Brief summary about the algorithms designed for data processing and clustering
    - All the results you obtained, presented in table or figure (using figure more, and show the results clearly and beautifully)
    - Interpretation/discussion for each result
    - Do not need to copy the code into this document
  - 2) Source code, Including:
    - Clustering algorithm
    - Other analysis code

Thank you!