Clustering Data Streams: A Second Look

Vasudha Bhatnagar
vbhatnagar@cs.du.ac.in
Deptt. Of Computer Science

Sharanjit Kaur
skaur@cs.du.ac.in
Acharaya Narendra Dev College

University of Delhi, India.

Acknowledgement: Dhriti Khanna, Tripti Gupta
Department of Computer Science, University of Delhi

So far….

- Google Scholar Search on Stream Clustering
- About 65,000 results
- Survey/Tutorial Papers

1. YH Lu: Mining Data Streams – A Survey, 2005
2. Alireza Rezaei Mahdiraji: Clustering data stream: A survey of algorithms, 2009

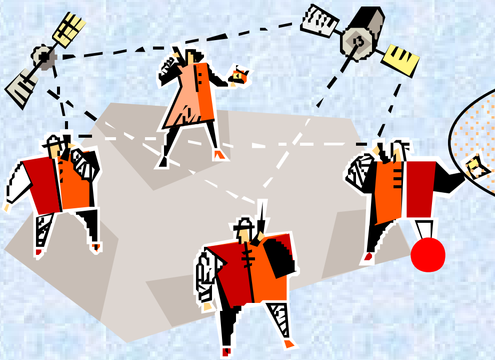
Based on papers…

i. [Barbara (2002)]: Requirements of Clustering Data Streams. SIGKDD Explorations 3(2):23-27.
ii. [Zhang et al. (1996)]: BIRCH: An Efficient Data Clustering Method for Very Large Databases. ACM SIGMOD : 103-110.
iii. [Callaghan et al. (2002)]: Streaming-Data Algorithms for High-Quality Clustering. ICDE: 685
iv. [Aggarwal et al. (2003)]: A Framework for Clustering Evolving Data Streams. VLDB: 81-92.
v. [Park et al. (2004)]: Statistical Grid-based Clustering over Data streams. ACM SIGMOD: 32-37.
vi. [Cao et al. (2006)] : Density-Based Clustering over an Evolving Data Stream with Noise. ICDM (SIAM): 326-337.
vii. [Orlowska et al. (2006)]: Can Exclusive Clustering on Streaming Data be Achieved? SIGKDD: 102-108.
viii. [Dang et al. (2009)]: Incremental and Adaptive Clustering Stream Data over Sliding Window. DEXA : 660-674.
ix. [Bhatnagar et al. (2009)]: A Parameterized Framework for Stream Clustering Algorithms. IJDWM (5):36-56.
x. [Aggarwal (2007)]: Data Streams: Models and Algorithms. Springer.
xi. [Gama (2009)]: Knowledge Discovery from Data Streams. Springer.

Agenda

- Introduction to data streams and clustering
- Contemporary stream clustering algorithms
- Influence of Synopsis
- Tailoring stream clustering algorithms

Streaming Data



Sensors networks

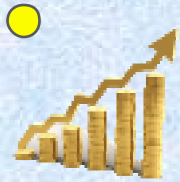
Data Stream:
A continuous inflow of data points, potentially unbounded

Web clicks

Patient monitoring

Communication networks

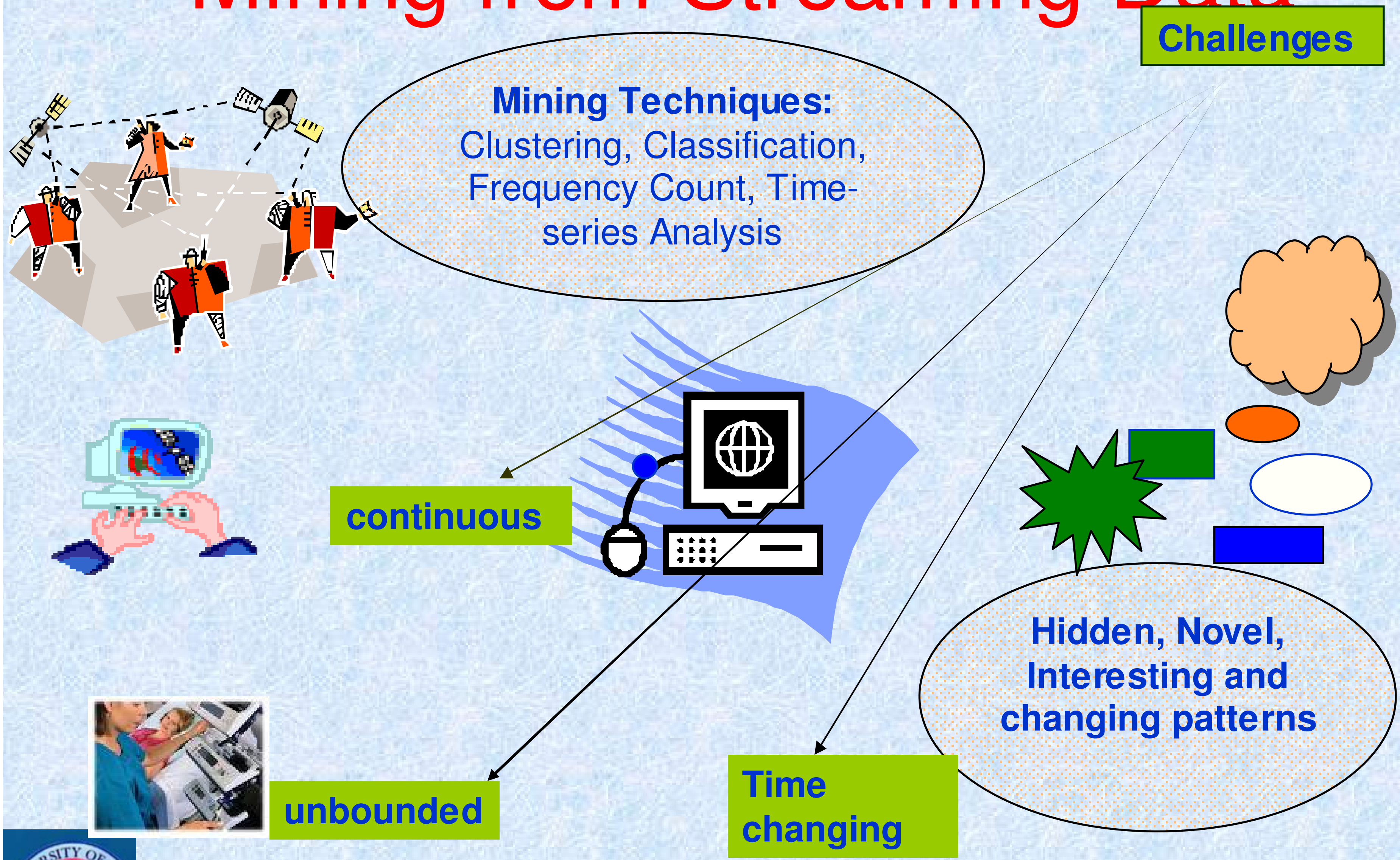
Stock ticks



What is Data Stream?

- On-line data with
  - Continuous flow
  - Potentially infinite
  - Time changing data characteristics

Mining from Streaming Data

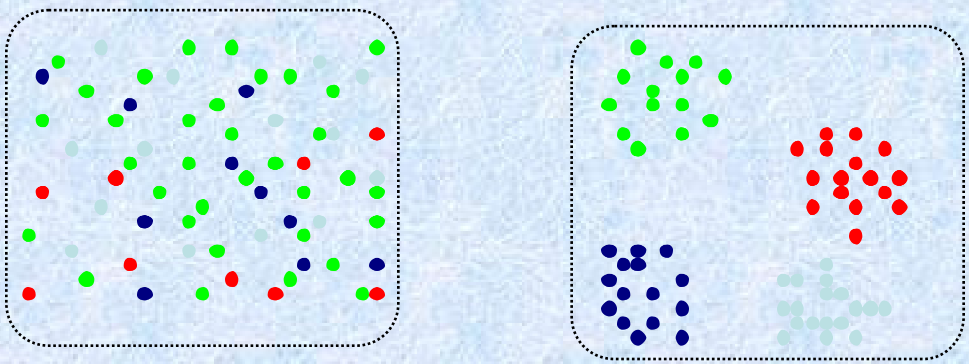


Challenges in Mining of Data Streams

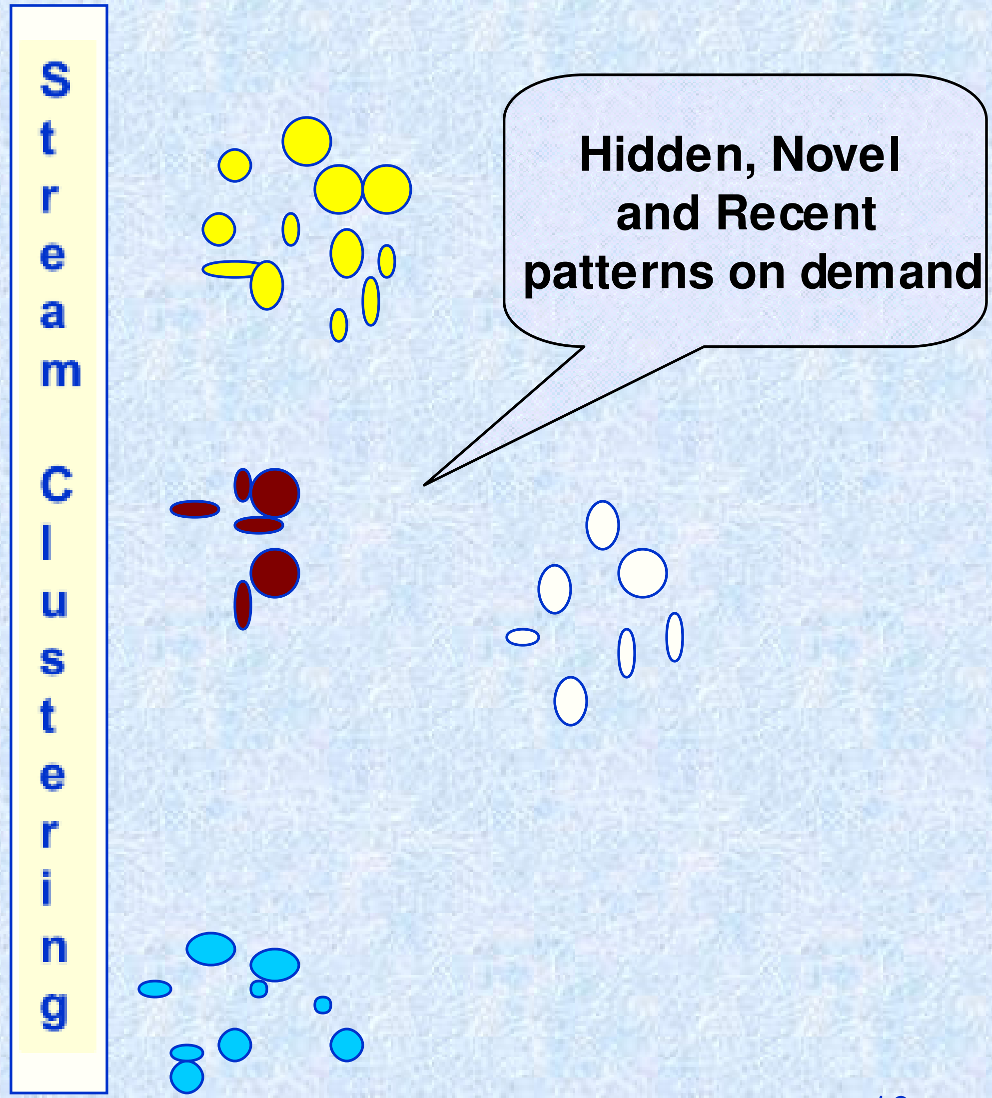
- Continuous inflow of data — Processing time
- Unbounded volume — Memory constraint & Single pass algorithm
- Evolution of data characteristics — Recent patterns

Clustering: Basic Idea

- Grouping a set of data objects into cluster
- Similar objects within the same cluster
- Dissimilar objects in different clusters
- No previous categorization known
- Descriptive Technique



Clustering of Data Stream



Requirements for Clustering Streams
[Barbara (2002)]

- Compactness of Synopsis
- Fast, incremental processing of new data points
- Clear and fast identification of outliers
- Insensitivity to order of incoming data points
- Capturing recency and data evolution

The overall goal is to get best possible clustering by making the best use of available resources.

Agenda

- Introduction to data streams and clustering
- Contemporary stream clustering algorithms
- Influence of Synopsis
- Tailoring stream clustering algorithms

PRECURSOR…

BIRCH (1996) : Balanced Iterative Reducing and Clustering using Hierarchies [Zhang et al. (1996)]

- Single scan
- Incremental algorithm
- Handles very large datasets
- First algorithm to detect outliers
- Opportunity for parallelism
- Introduces Cluster Feature

Clustering Feature (CF)

A triplet summarizing the information maintained for a cluster.

CF = <N, LS, SS>

N: d-dimensional data points in a cluster

LS: Linear sum of N data points

SS: Squared sum of N data points

Summary Representation of a cluster

Why Summary ???

unbounded

- Much less memory requirements compared to all data points in a cluster
- Sufficient for calculating measurements for a cluster

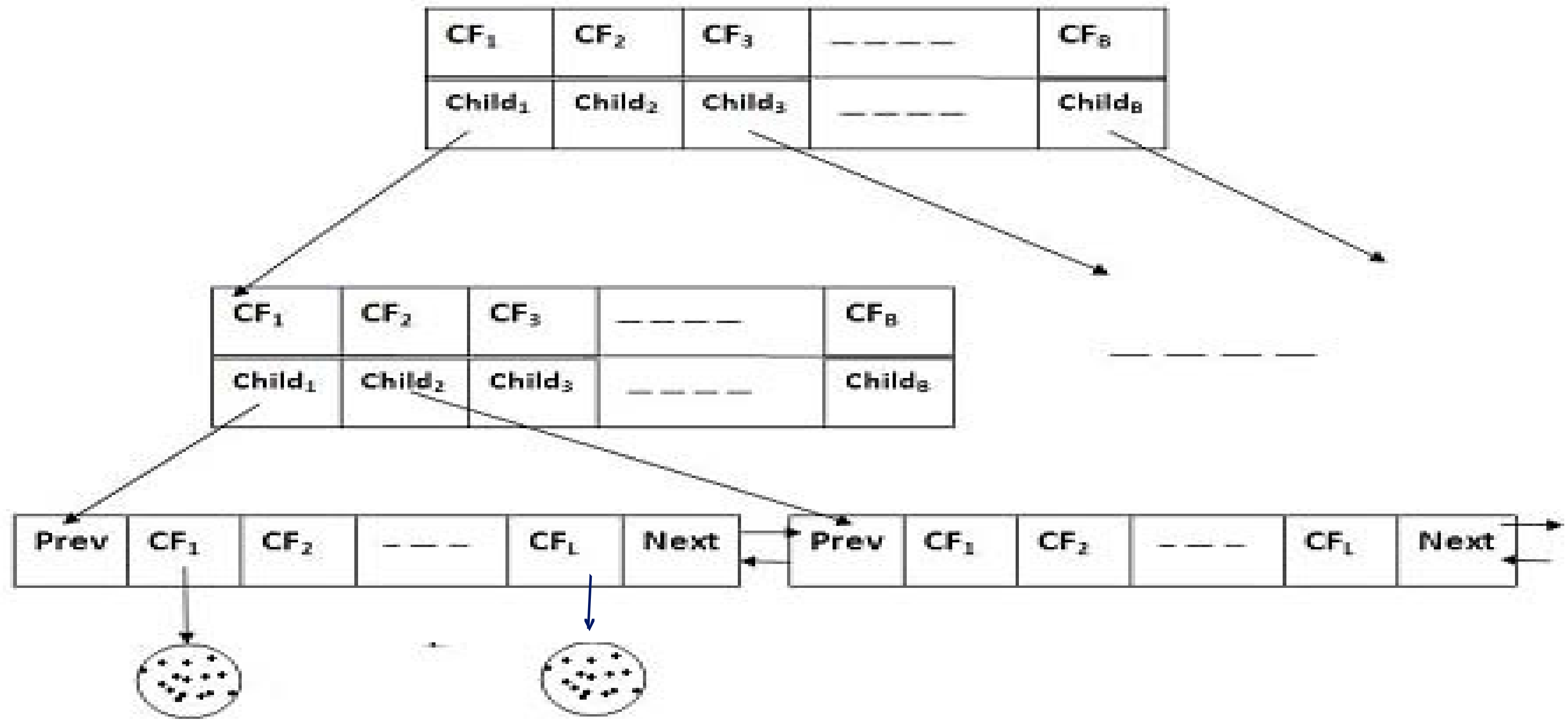
$$Centroid = \vec{LS}^{j} / N$$

$$Radius = \frac{\sqrt{N\sum_{j=1}^{d} SS^{j} - \sum_{j=1}^{d}(LS^{j})^{2}}}{N^{2}}$$

CF Tree: A compact representation of dataset

- A height-balanced tree with two parameters:
  - Branching Factor (B): Controls maximum entries in a non-leaf node
  - Radius/ Diameter Threshold (T): Controls the size of tree

CF-Tree



BIRCH Clustering Algorithm - Overview

Data

Phase1: Load into memory by building a CF tree

Initial CF tree

Phase2(Optional): Condense into desirable range by building a smaller CF tree

Smaller CF tree

Phase3: Global clustering

Good cluster

Phase4 (optional and off line): Clustering refinement

Better Cluster

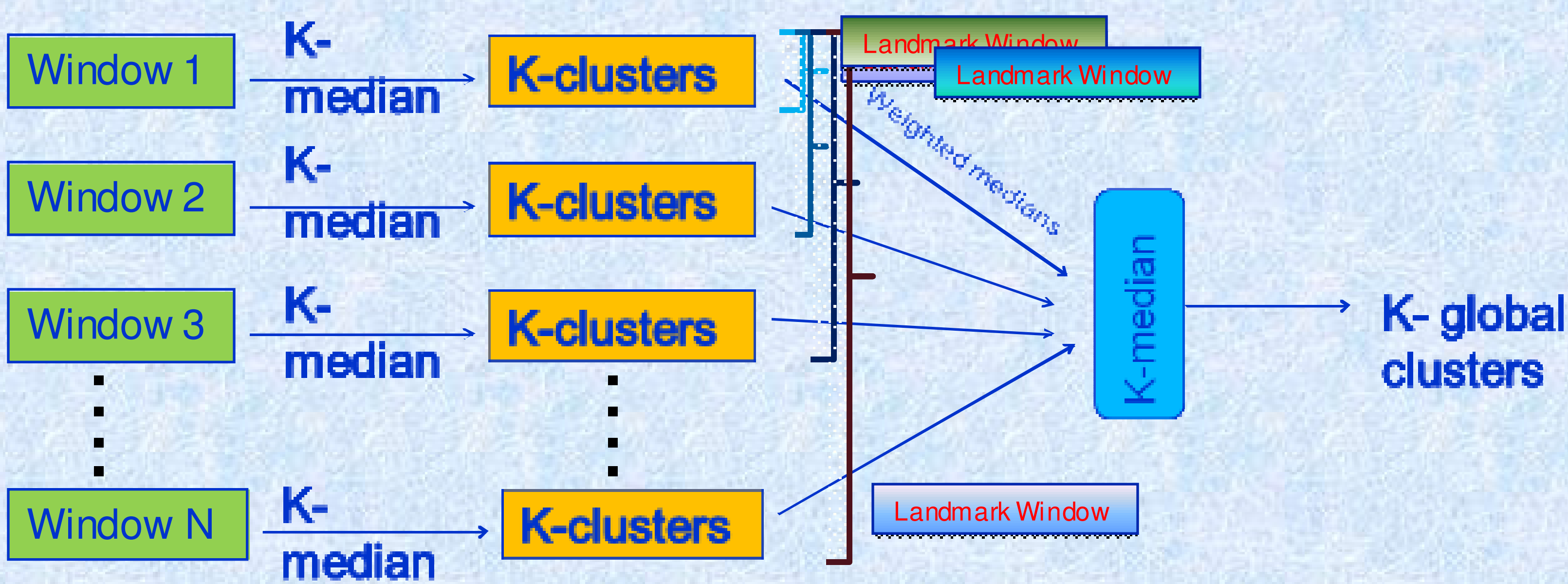
Why BIRCH is Unsuitable for Streams?

- High Per-point processing time
  - Identify appropriate leaf
  - Updating leaf statistics
  - Modifying the path to leaf
- Clustering results are sensitive to order of incoming data points
  - Points are inserted in a closest child node
- Not designed for capturing data evolution

STREAM
[Callaghan et al. (2002)]

- Uses divide and conquer strategy
- Clusters stream in fixed size windows
  - Small space algorithm
- Stores weighted medians for each window
  - Memory efficient
- Clusters medians after processing the current window
- Uses landmark window model

Clustering Process in STREAM



CluStream : A Pioneer Algorithm

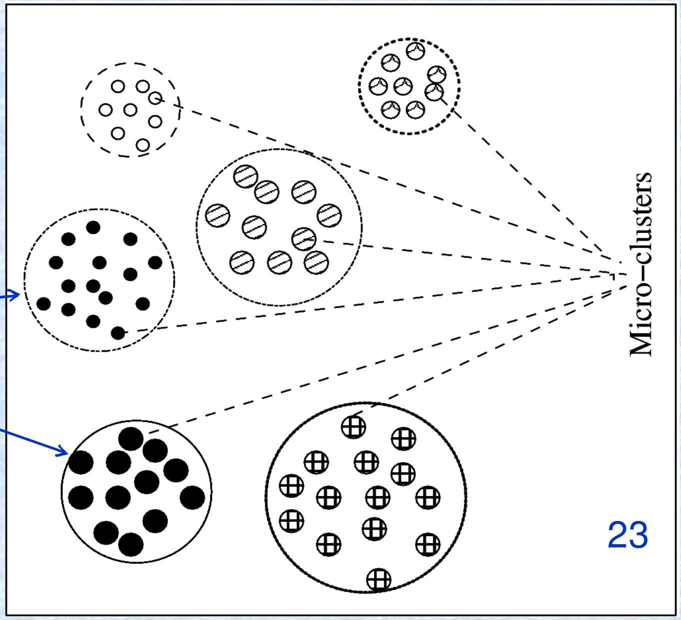
[Aggarwal et al. (2003)]

- Framework for clustering evolving stream
- Generates convex-shaped pre-specified number of clusters
- Reports clusters in user-defined time horizon
- Handles numeric data streams

Clustering process in Clustream

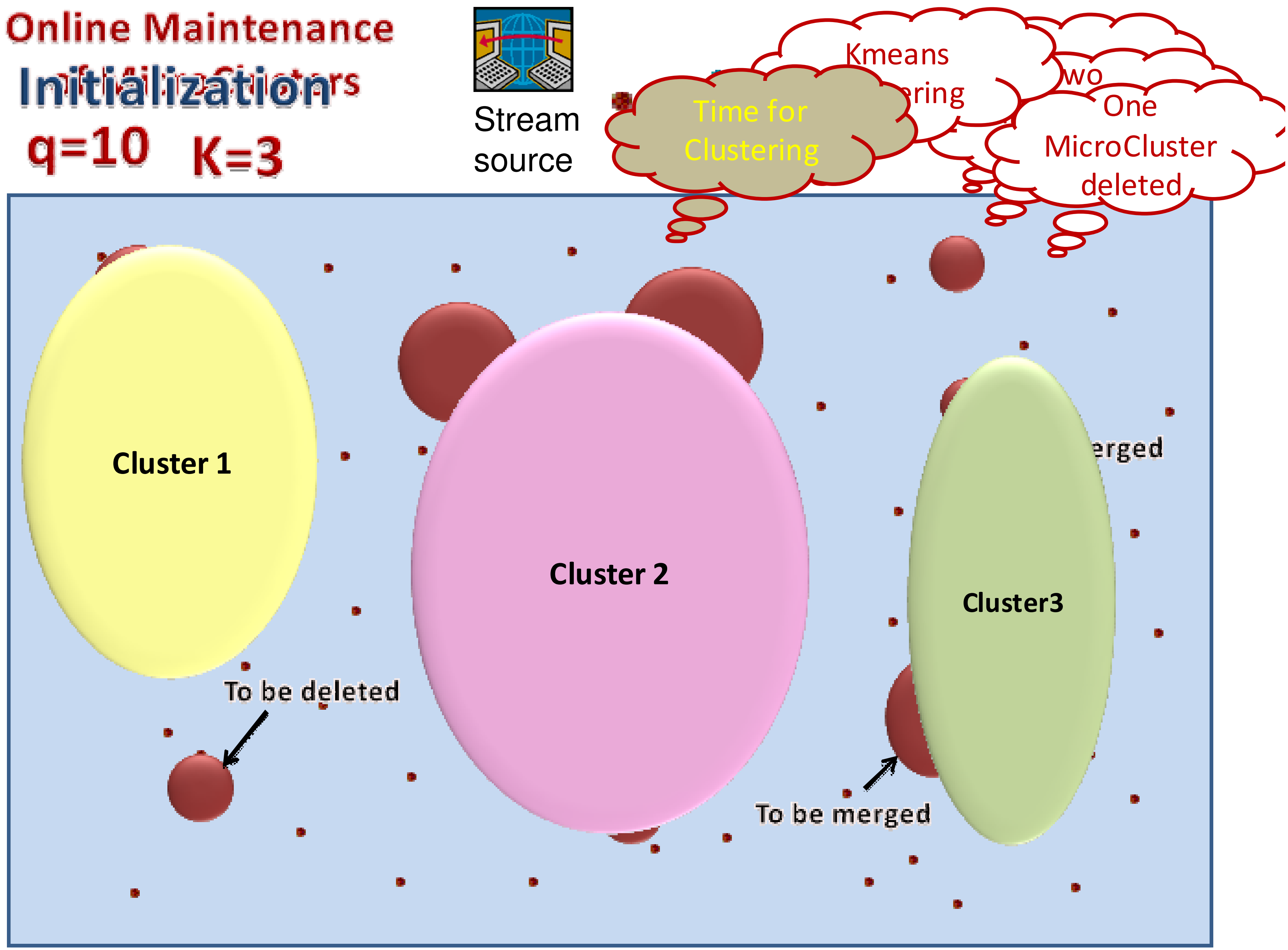
- Deploys micro-cluster based synopsis
- Micro-cluster
  - Represents set of points close to each other
  - Temporally extended Cluster Feature
- Macro-cluster
  - Inherent structures in data

Macro-cluster



Contd…

- Two underlying components
  - On-line
    - Incrementally updates synopsis
    - Stores snapshots of synopsis content
  - Off-line
    - Uses synopsis for generating clusters
    - On-demand clustering in user-specified time horizon


Online Maintenance
Initialization
q=10
K=3
Stream source
Kmeans
Time for Clustering
One MicroCluster deleted
Cluster 1
Cluster 2
Cluster3
To be deleted
To be merged

Limitations

- Number of clusters to be predefined
  - Infeasible in evolving data streams
- Convex-shaped clusters
  - Real clusters are arbitrarily shaped
- Overlapping clusters
  - Centroid of clusters changes with accumulation of points
- Inefficient for outlier handling
  - Appearance of outliers removes genuine clusters

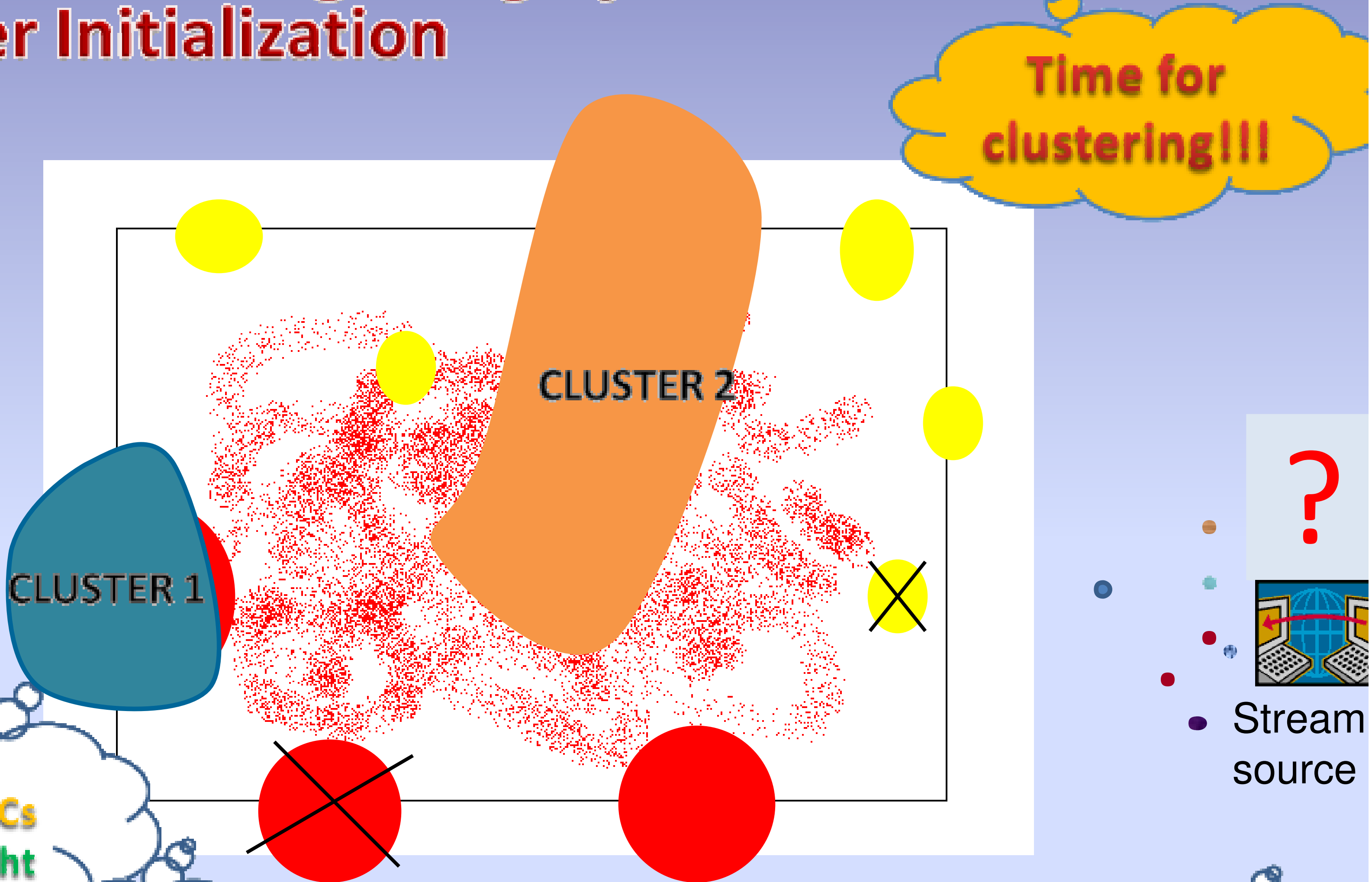
DenStream : A Density-based Algorithm
[Cao et al. (2006)]

- Reports arbitrarily-shaped clusters
- Uses damped window model to capture recency
- Segregates clusters from outliers and noise
- Capable of detecting clusters in noisy stream

Clustering Process in DenStream

- Synopsis consists of
  - Potential Micro clusters (PMC)
  - Outlier Micro clusters (OMC)
- Two components
  - On-line
- Synopsis updation and maintenance
- Periodically check status of PMCs and OMCs
  - Off-line
    - Generates clusters on user demand
    - Applies DBSCAN on stored PMCs

Clustering using DBSCAN
After Initialization



PMCs and OMCs differs in weight

Time for periodic status evaluation!!

Selecting Core microclusters

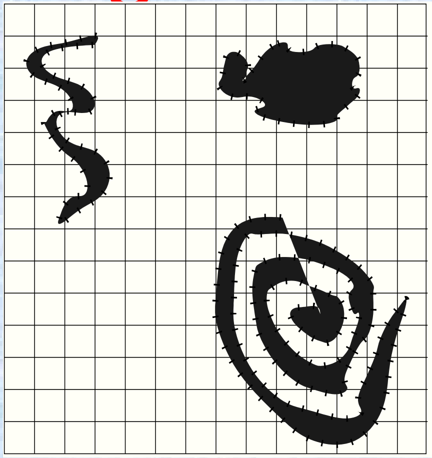
Applying DBSCAN

Limitations

- Overlapping clusters
  - Cluster feature maintenance
- Loss of spatial information
- Many user-defined parameters!!!
  - Density threshold, radius threshold, decaying factor etc..

Statistical Grid-based Clustering

[Park et al. (2004)]



- Detects arbitrarily-shaped clusters
  - Preserve spatial information
- Exclusive clustering
  - A point is member of exactly one cell
- Suitable for mixed attributes
- Summarizes data distribution

Clustering Process in Stats-grid

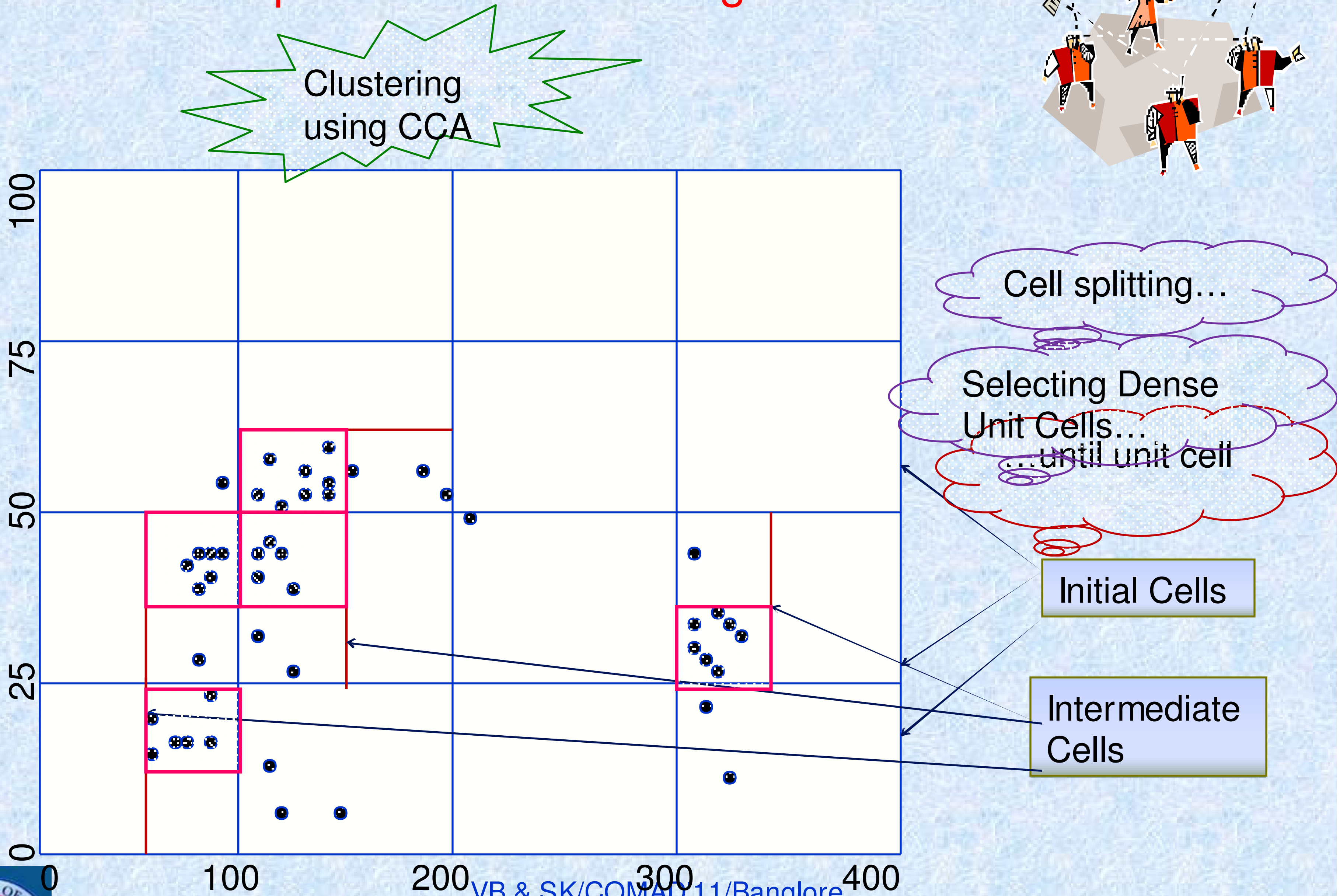
- On-line
  - Grid updation
  - Grid maintenance

Insertion and cell splitting

Pruning of sparse cells

- Off-line: Generates clusters using connected component analysis
  - Detects maximal connected regions by coalescing adjacent unit cells in the data space

Grid Updation and Clustering



To Summarize..

- Deploys grid as synopsis
- Initially, data space is partitioned into mutually exclusive equal-sized cells
- Dynamic cell partitioning to get unit cells
  - Smallest cell used in clustering

Limitations

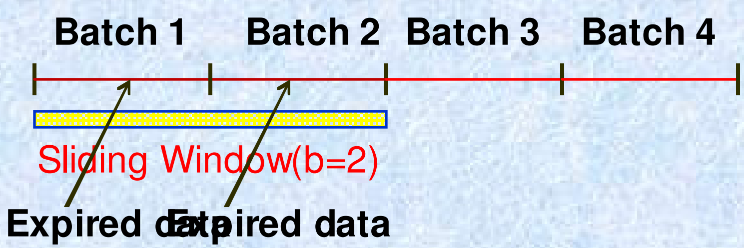
- Degraded performance for uniformly distributed data
  - Large number of cells
- Does not capture data evolution
  - Landmark window model

SWEM: A Statistical Approach for Stream Clustering [Dang et al. (2009)]

- Sliding Window with Expectation Maximization Technique
- Incremental and adaptive clustering
- Uses Expectation Maximization technique
  - L(□) = ln P(X| □)
  - Iterative and incremental approach
- Soft clustering

Data Processing in SWEM

- Processes data in a batch
- Focuses on the data in a pre-specified fixed size window (b batches)



Assumptions

- Streams consists of k mixture models
- Each model follows a multivariate normal distribution
- Distribution within one batch is always fixed

Clustering Process

- Initial Phase
  - Decides upon M micro-components
  - For each micro-component, following parameter set is maintained

$$\phi_h = \{\alpha_h, \mu_h, \Sigma_h\}$$

$$where\ \alpha_h : \text{weight},\ \mu_h : \text{mean},\ \Sigma_h : \text{covariance matrix}$$

  - Global k cluster models are fitted

Contd…

- Incremental Phase
  - Absorbs points of the batch and updates micro-components
  - Split and Merge micro-components
    - To discretely redistribute components across entire data space
- Expiring Phase
  - Removes impact of previous batch of data points

Limitation

- Generates pre-defined K clusters
- Assumption about underlying data distribution
  - Infeasible in case of evolving data streams
- Batch processing

Agenda

- Introduction to data streams and clustering
- Contemporary stream clustering algorithms
- Influence of Synopsis
- Tailoring stream clustering algorithms

Categorization of Stream Clustering Algorithms

- Distance-based
- Density-based
- Grid-based
- Statistical methods based

Microcluster as synopsis

Grid based synopsis

Distribution based synopsis

Commonality: Two components (on-line and off-line)
Synopsis for summarization of incoming data points

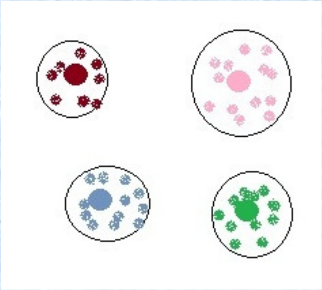
How to choose a suitable algorithm?

- Functional Characteristics
  - Shape of clusters
  - Sensitivity to order of data
  - Type of clustering (hard/soft)
  - Capturing data evolution
- Operational Characteristics
  - Per-point processing time
  - Initialization requirement
  - Memory requirement

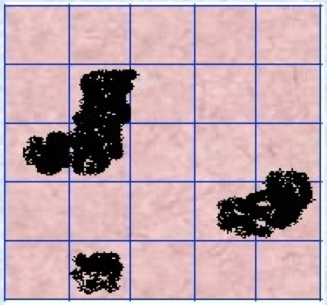
Synopsis Dependent

Alternatives for Synopsis

- Micro-clusters
  - Representatives set of points to which incoming points are absorbed using a distance metric



- Grid Structure
  - Divides multi-dimensional data space into a set of mutually exclusive cells



  - Incoming points are mapped according to their dimensional values in a cell

Synopsis Comparison

Characteristics	Micro-cluster	Grid Structure
Functional Characteristics		
Detection of inherent natural patterns	No	Yes
Sensitive to data ordering	Yes	No
Hard/ Exclusive clustering	No	Yes
Data evolution	Yes	Yes
Operational Characteristics		
Initialization required	Yes	No
Per-point processing time	Unpredictable, bounded	Constant
Memory requirement depends on	Distance threshold	Grid granularity

For Example…

- Telecom application
  - Streaming record consists of call duration, call type, call time, source and destination identities
  - No need to preserve spatial information
- Weather Monitoring
- Remote Sensing
  - Exclusive clustering desirable

Micro-cluster Based

Grid Based

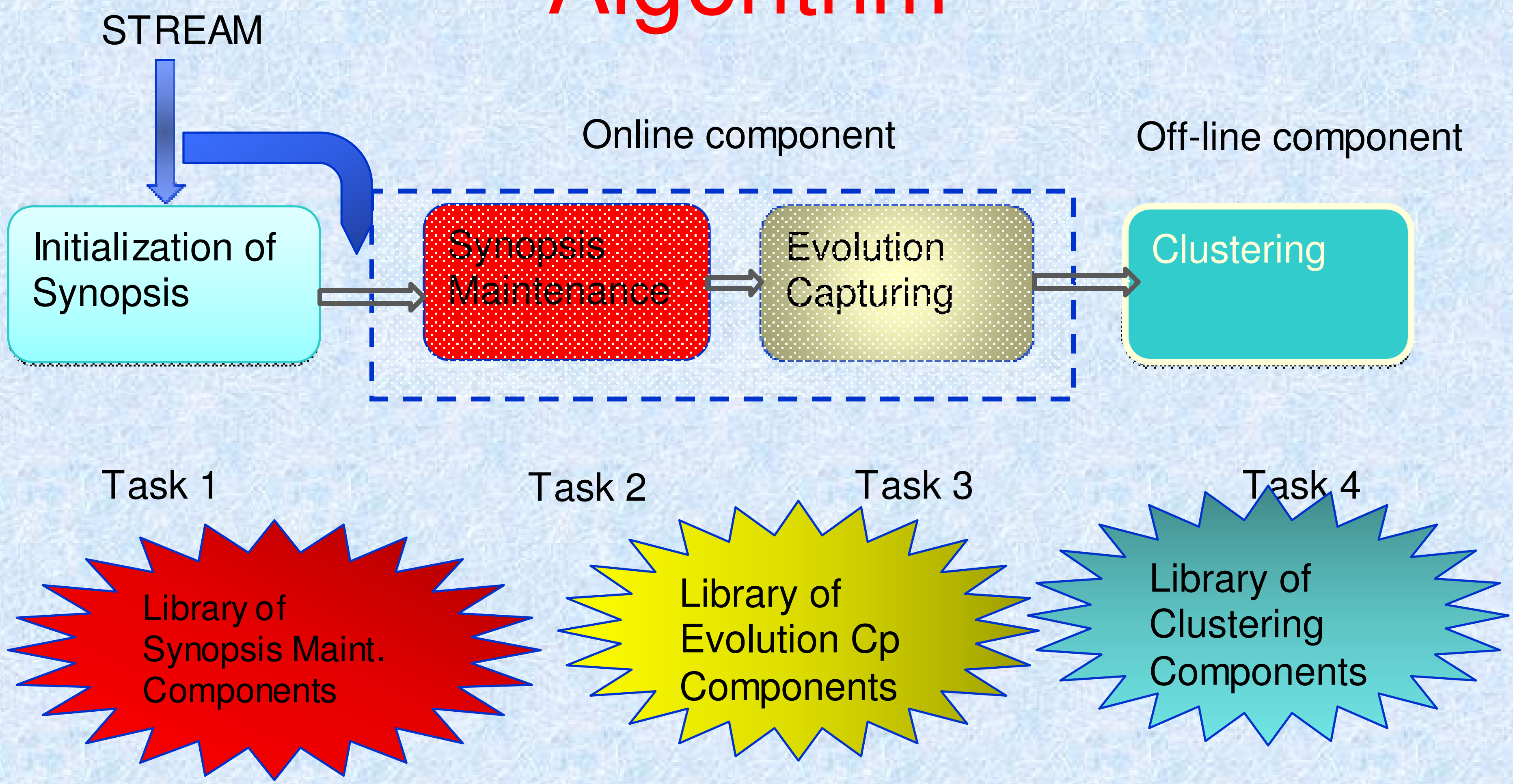
Agenda

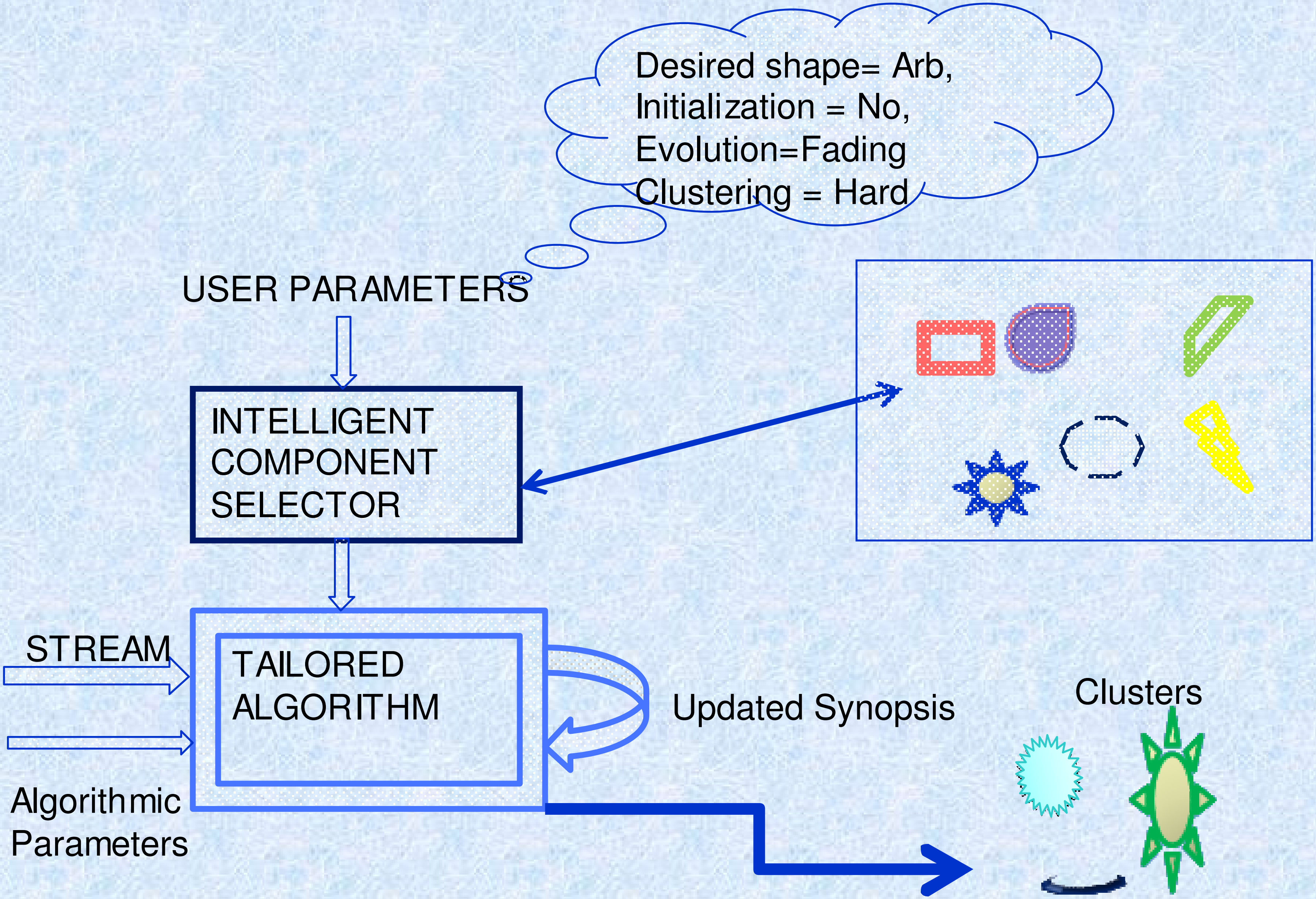
- Introduction to data streams and clustering
- Contemporary stream clustering algorithms
- Influence of Synopsis
- Tailoring stream clustering algorithms

A Parameterized Framework for Clustering Streams

- In general…
  - Ad-hoc approach for solving individual problems using KDD technology (Yang and Wu, 2006)
  - Need for a unified framework for integration of different tasks
- Specifically…
  - Assembling of stream clustering algorithms fulfilling user's application needs

Tasks in Stream Clustering Algorithm



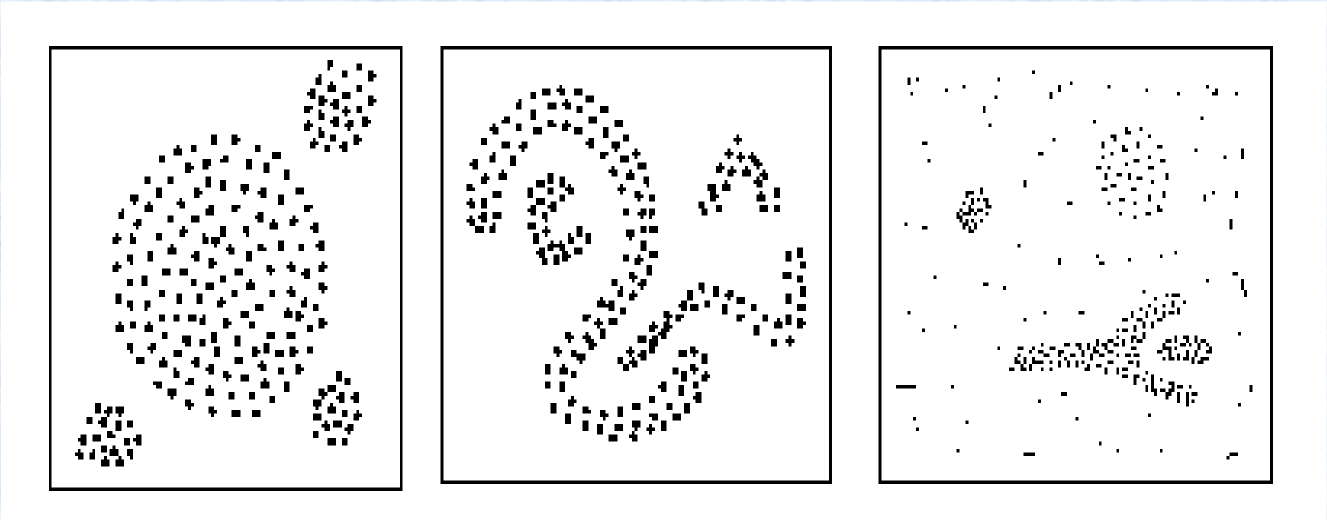

Desired shape= Arb,
Initialization = No,
Evolution=Fading
Clustering = Hard
USER PARAMETERS
INTELLIGENT COMPONENT SELECTOR
STREAM
TAILORED ALGORITHM
Updated Synopsis
Clusters
Algorithmic Parameters

Summary

- Four categories of stream clustering algorithms
- Influencing factor - synopsis
- Parameterized framework for tailoring stream clustering algorithms

Thank you

Density-based clustering



Each cluster has a considerable higher density of points than outside of the cluster