
Mining Stream and
Time-Series
Database

Outline

- **Define Time-Series Database**
  - Trend Analysis
  - Similarity Search
  - Data Reduction and Transformation Techniques
- **Define Mining Data Stream**
  - Clustering of Stream Date
  - Random Sampling
  - Data Reduction Methods
  - Data Stream Management Systems
  - Compression Technique

Time-Series Database

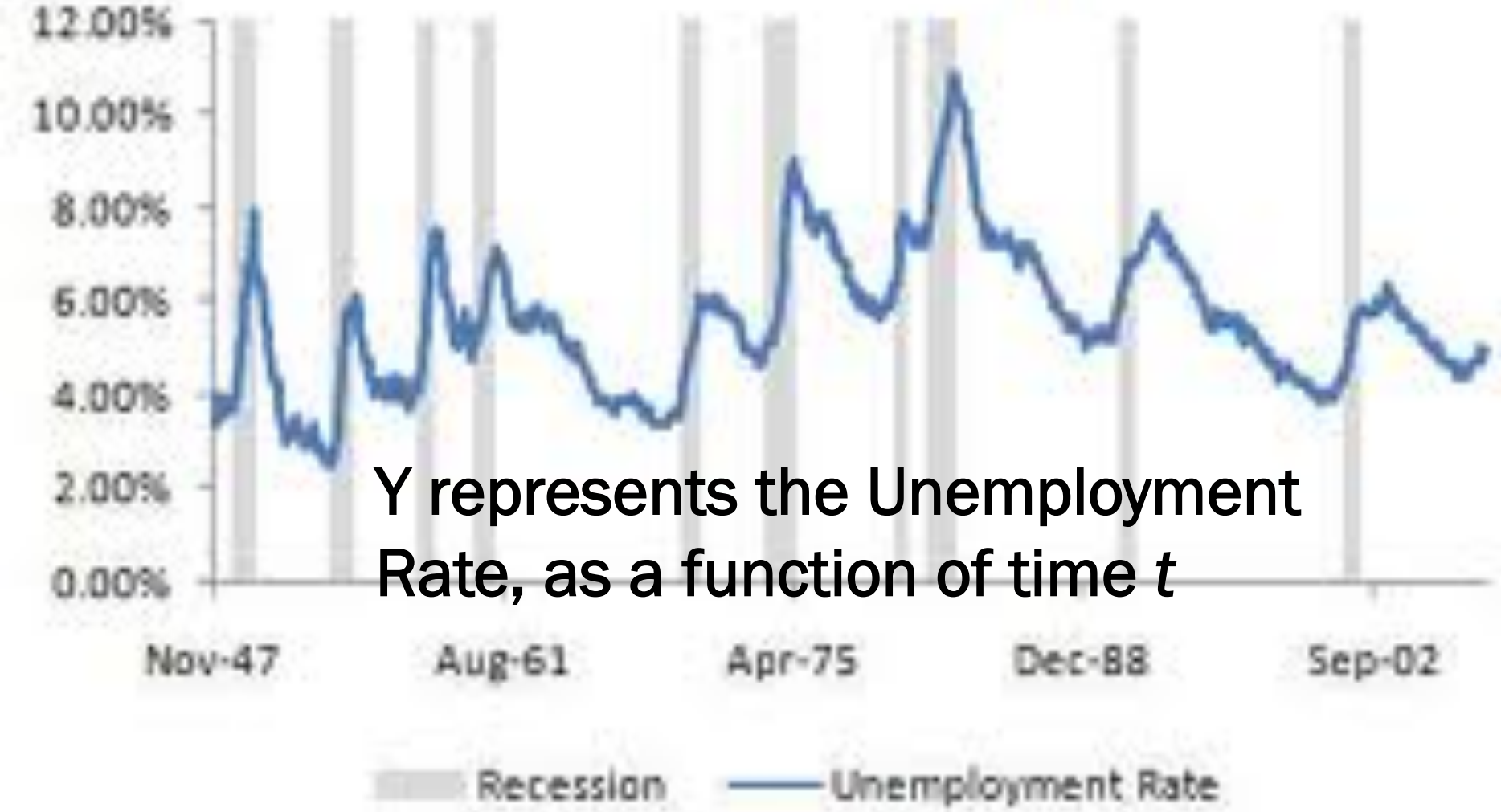
A time-series database consists of sequences of values or events obtained over repeated measurements of time.

- They are used in stock market analysis, sale forecasting, budgetary analysis, observation of natural phenomena, and etc

For example:

Y = F(t) can be illustrated as a time-series graph

U.S. Civilian Unemployment Rate
January 1948 to January 2008



Y represents the Unemployment Rate, as a function of time *t*

Trend Analysis

How can we study time-series data?

Two goals in time-series analysis:

1. **Modeling time**-series- to gain insight into underlying forces that generate the time-series

2. **Forecasting time-series**- to predict the future values of the time-series variables

Trend Analysis

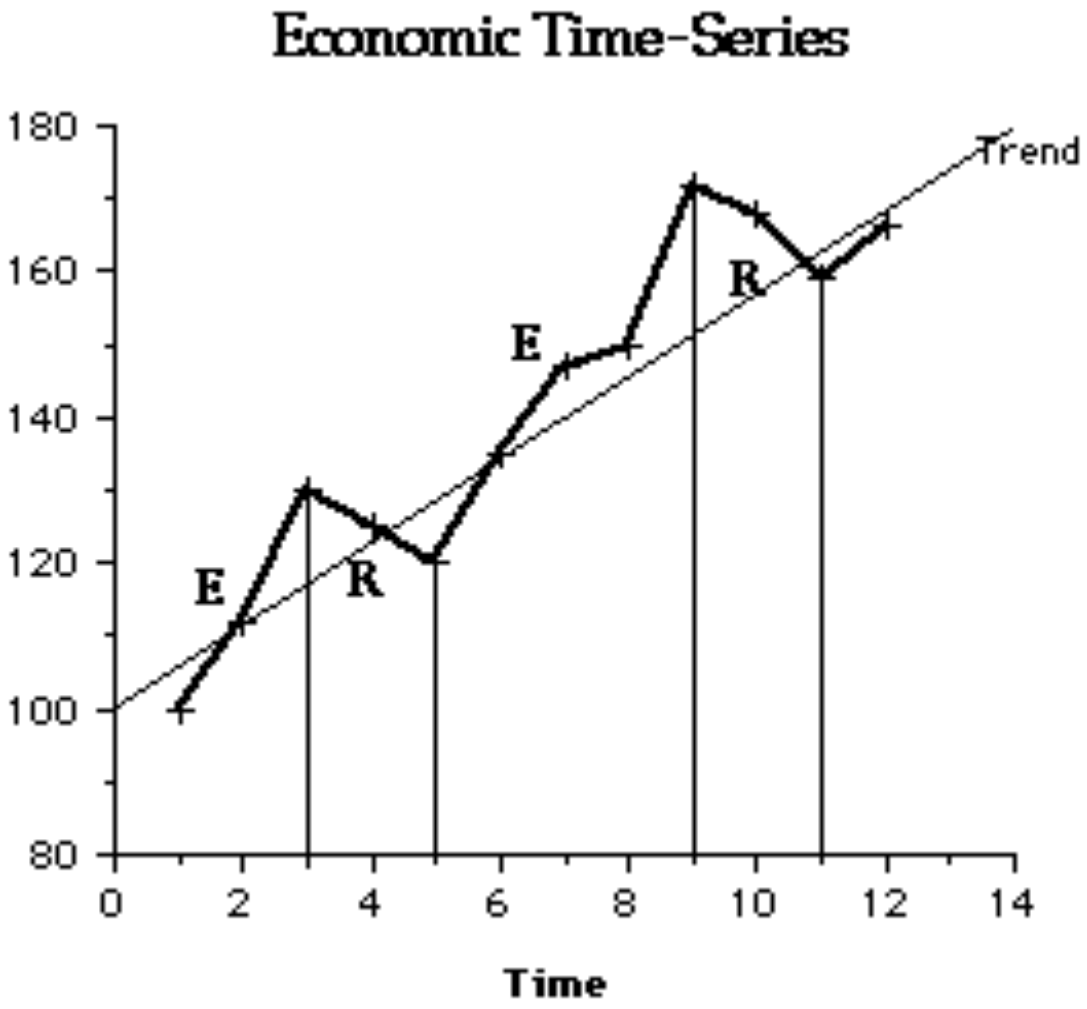
Trend analysis consist of four major components for characterizing time-series data:

1. Long-term movements
2. Cyclic movements
3. Seasonal movements

for examples:

- Sudden increase in sales of flowers before Valentine's Day

4. Random movements


Economic Time-Series
180
160
140
120
100
80
0
2
4
6
8
10
12
14
Time
Trend
E
R
E
R

Trend Analysis

Time-series modeling is also referred to as **decomposition** of time-series into these four basic movement.

- The time-series variable Y can be modeled as either the product of the *four(Y = T*C*S*I)* or their sum

Trend Analysis

How can we determine the trend of the data?

Moving Average of order n:

$$s_t = \frac{1}{n} \sum_{j=t}^{t+n-1} a_j.$$

- A moving average trends to reduce the amount of variation present in the data set.

Freehand

- An approximate curve line is drawn to fit a set of data based on the user's own judgment

Similarity Search

A **similarity search** finds data sequences that differ only slightly from given query sequence.

Two types of similarity searches:

1. Subsequence matching

2. Whole sequence matching

Similarity searches is useful for financial market analysis or medical diagnosis

Data Reduction and Transformation Techniques

Due to the tremendous size and high dimensionality of time-series, data reduction often serves as the first step. Leads to smaller **storage space** and **faster processing.**

Data Reduction and Transformation Techniques

Data Reduction Methods

1. Attribute subset selection
2. Dimensionality reduction
3. Numerosity reduction

Transformation Technique

Distance-preserving orthonormal transformations- are often used to transform the data from the time domain to the frequency domain.

Mining Data Stream

Stream data flow in and out of a computer system continuously and with varying update rates. They are temporally ordered, fast changing, massive, and potentially infinite.

For example:

Telecommunications data, satellite, data from electric power grids, and transaction data from retail industry

Clustering Stream Data

For effective clustering of stream data, serveral methodologies have been developed, as follows:

- Compute and Store summaries of past data
- Apply a divide-and-conquer strategy
- Perform microclustering and macroclustering analysis
- Divide stream clustering into on-line and off-line processes

Clustering Stream Data

Stream is a single-pass, constant factor approx. algorithm that was developed for the k-medians problem.

- Idea is to assign similar points to the same cluster, where these points are dissimilar from points in other clusters.

CluStream is an algorithm for the clustering of evolving data streams based on user-specified, online clustering queries.

Random Sampling

Stream data is gigantic in size that we generally cannot store the entire stream data set in main memory or even on disk.

Instead of dealing with the entire data stream, we sample the stream at periodic intervals.

"*To obtain an unbiased sampling of the data, we need to know the length of the stream in advance.*"

Random Sampling

What can we do if we do not know this length in advance?

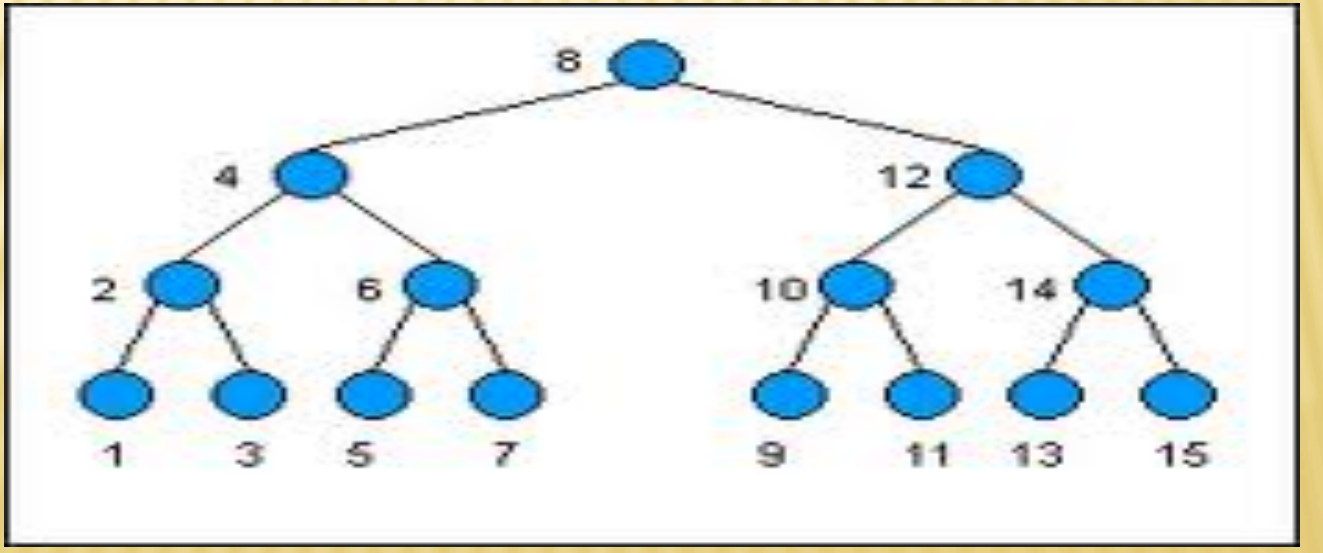
- Reservoir sampling

- Sliding Window
  - This element "expires" at time t + w, where w is the window "size" or length.

Data Reduction Methods

Allows a program to trade off between accuracy and storage, but also offer ability to understand a data stream at multiple levels of detail.

- Balanced Binary Tree



Data Reduction Methods

- **Wavelets**
  - Wavelets are a popular method for data stream compression

- **Sketches**
  - Provides probabilistic guarantees on the quality of the approx answer. Given *N* elements and a universe *U* of *v* values, such sketches can approx. ***F_0, F_1, and F_2 in O(log v + logN) space***

Data Reduction Methods

- **Randomized algorithm** is a probability distribution over a set of deterministic algorithms

- Randomized algorithms are often used to deal with massive, high dimensional data streams

- The use of randomization often leads to simpler and more efficient algorithms.

Management Systems

- In a DSMS, data streams arrive on-line and are continuous, temporally ordered, and potentially infinite.
- Once an element from a data stream has been processed, it is discarded or archived, and it cannot be easily retrieved unless it is explicitly stored in memory.

Management Systems

Stream data query includes three parts:

- End user
- Query processor
- Scratch space

Queries can be either:

- One-time query
- Continuous query

Compression Technique

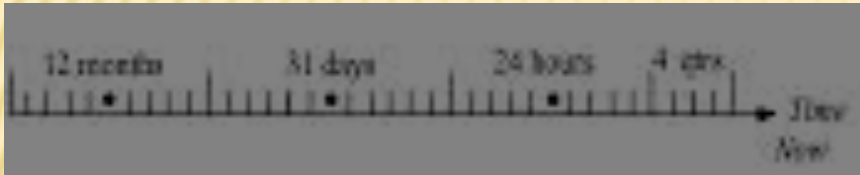
In stream data analysis, people are usually interested in recent changes at a fine scale but in long-term changes at a coarse scale.

Most recent time is registered at the finest granularity; the more distant time is registered at a coarser granularity. This time dimension model is called a **tilted time frame**.

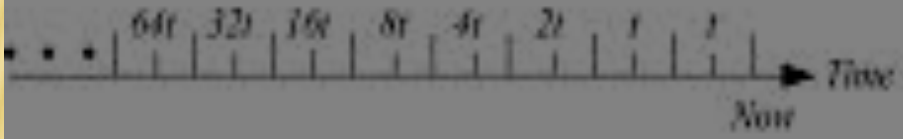
- This model ensures that the total amount of data to retain in memory or to be stored on disk is small

Compression Technique

Natural tilted time frame Model



a) A natural tilted time window model



b) A logarithmic tilted time window model

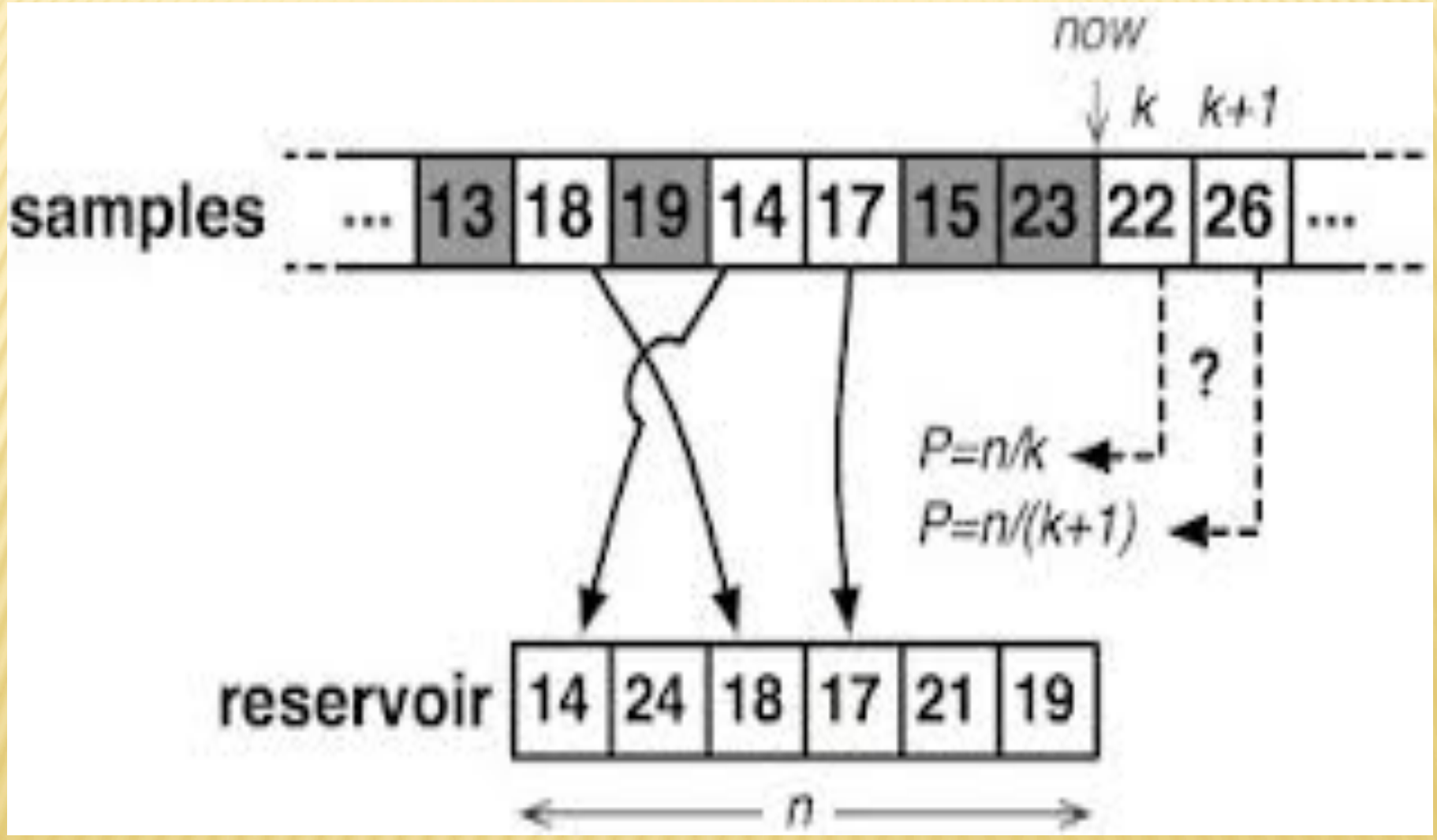
Frame no.	Snapshots (by clock time)
0	69 67 65
1	70 66 62
2	68 60 52
3	56 40 24
4	48 16
5	64 32

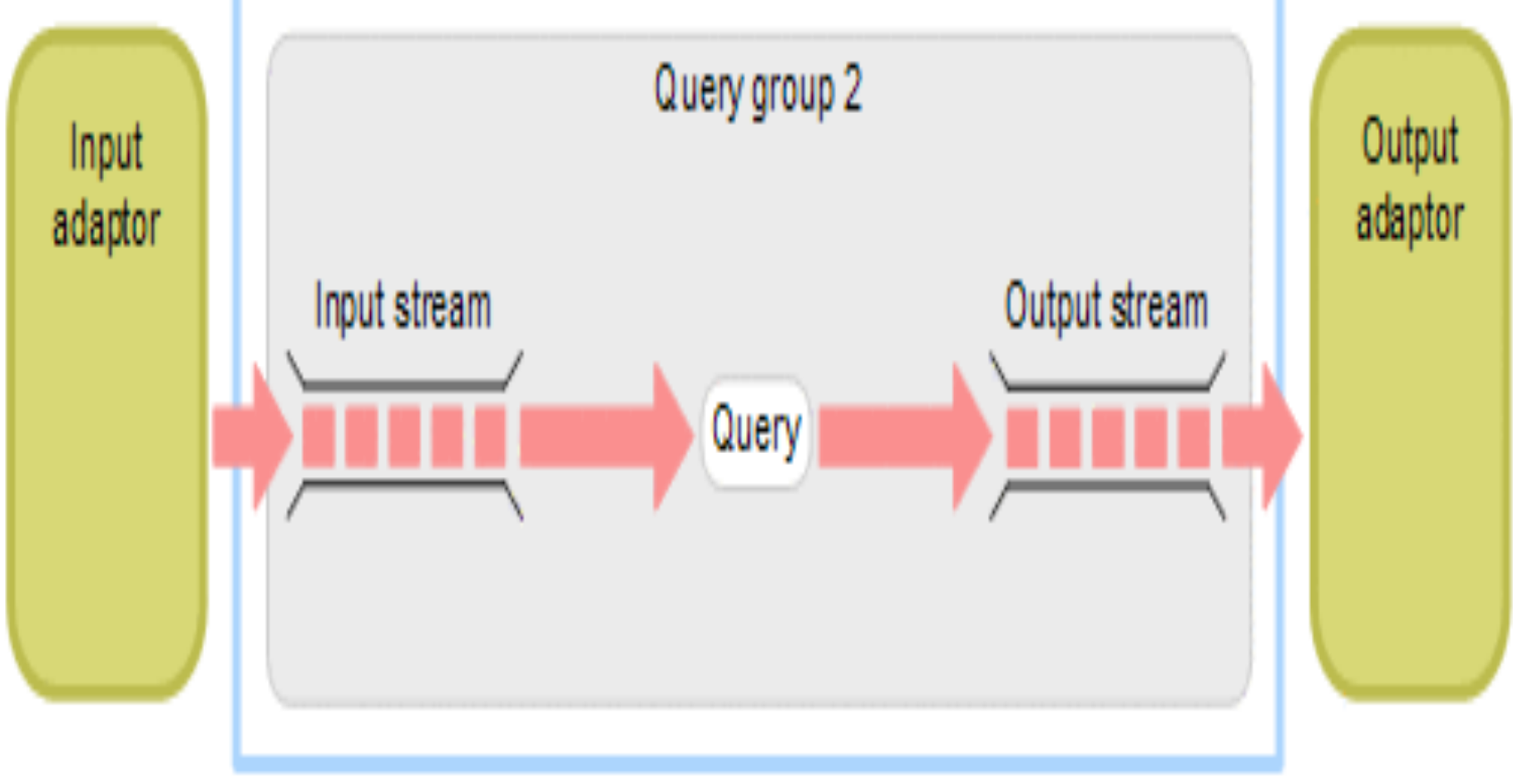
c) A progressive logarithmic tilted time window table

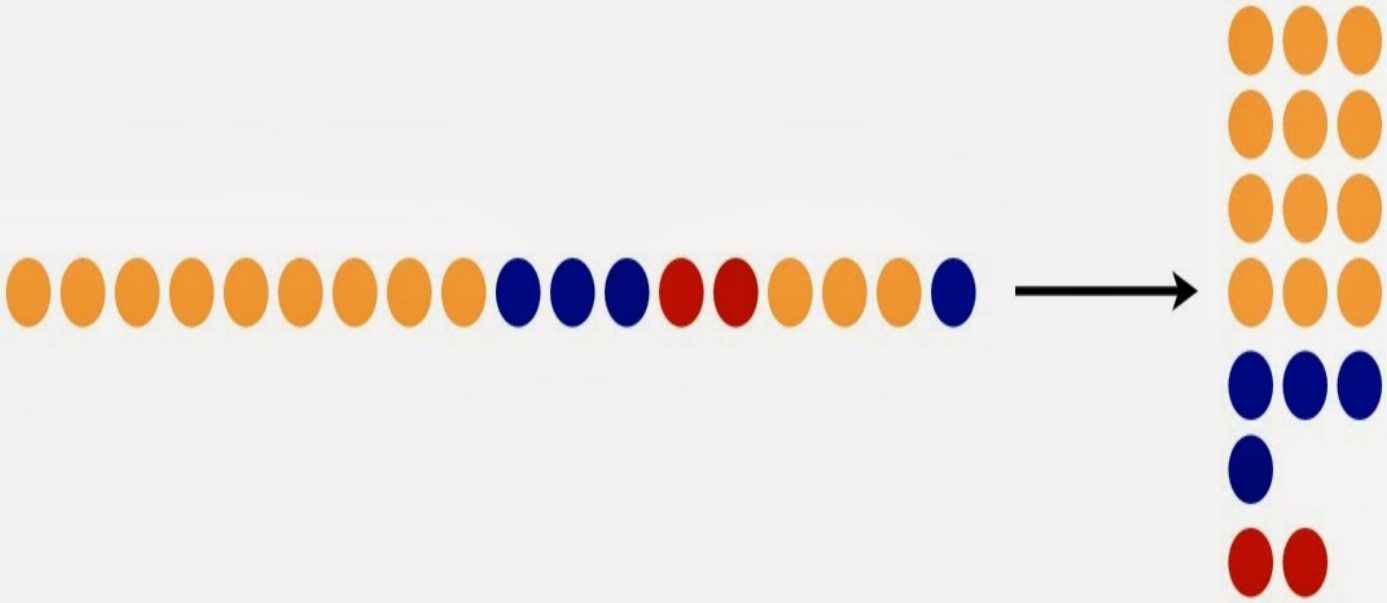
REFERENCE

- lionel-vinceslas.eurower.net
- www.tvmcalcs.com
- www.cise.ufl.edu
- www.codeproject.com
- www.cloudtm.eu
- wah.cse.cuhk.edu.hk
- www.uri.edu
- www.sciencedirect.com
- Koudas, N., & Srivastava, D. (2005, April). Data stream query processing. In*ICDE* (Vol. 5, p. 1145).
- Han. ((2006)). Mining Stream, Time-Series, and Sequence Data. In M. K. Jiawei, *Data mining: concepts and techniques* (pp. 467-534). Morgan kaufmann.

The End


now
k
k+1
samples
...
13
18
19
14
17
15
23
22
26
...
?
P=n/k
P=n/(k+1)
reservoir
14
24
18
17
21
19
n


Input adaptor
Query group 2
Input stream
Query
Output stream
Output adaptor

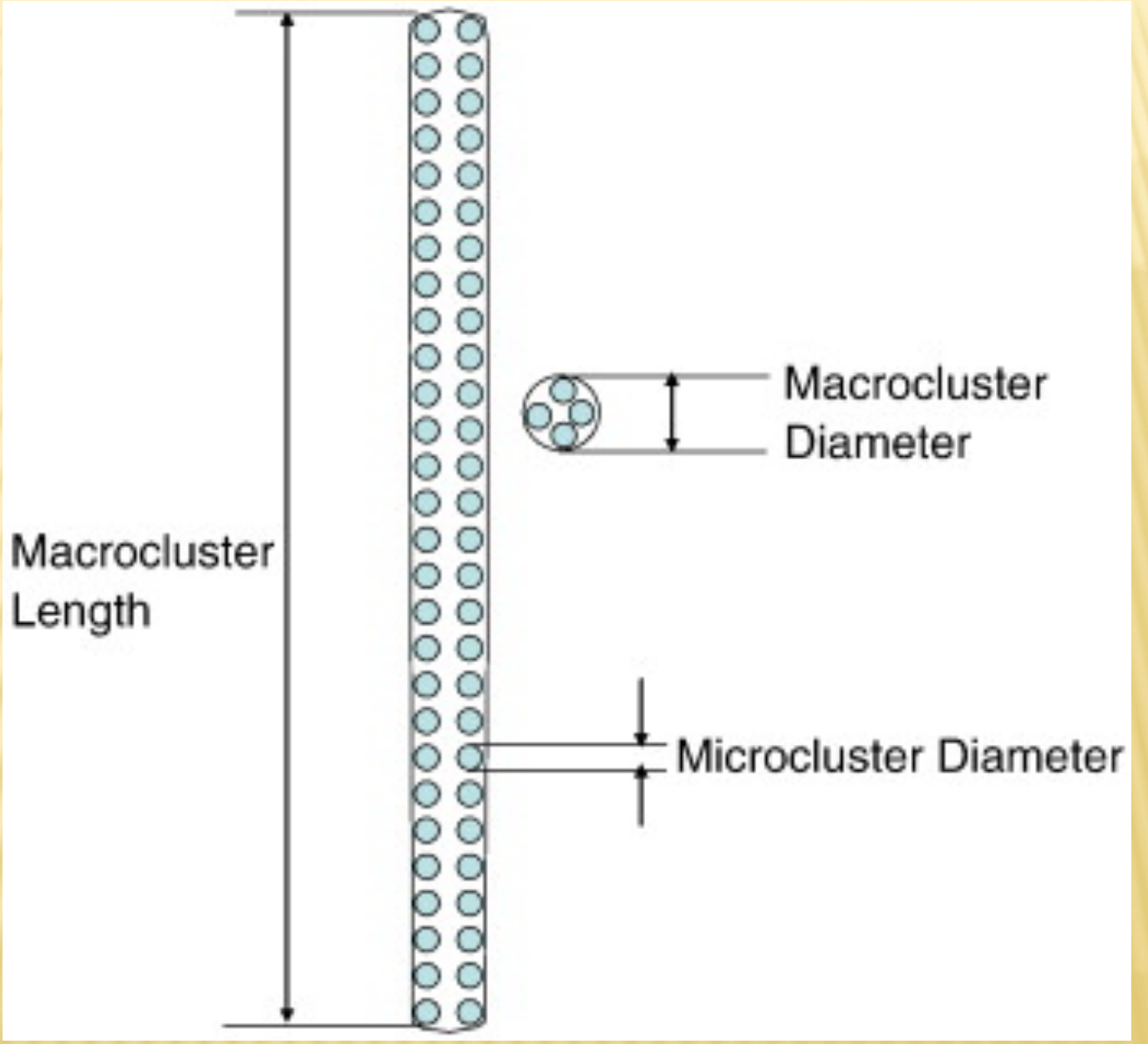


Build a summary
(start with "empty" summary)

Update with new information
(stream processing)

Merge summaries together
(allow for distributed processing)

Query : find out value of current summary
(tolerate approximate answers)


Macrocluster
Diameter
Macrocluster
Length
Microcluster Diameter