



# **BDA UNIT-2**

# DATA STREAM

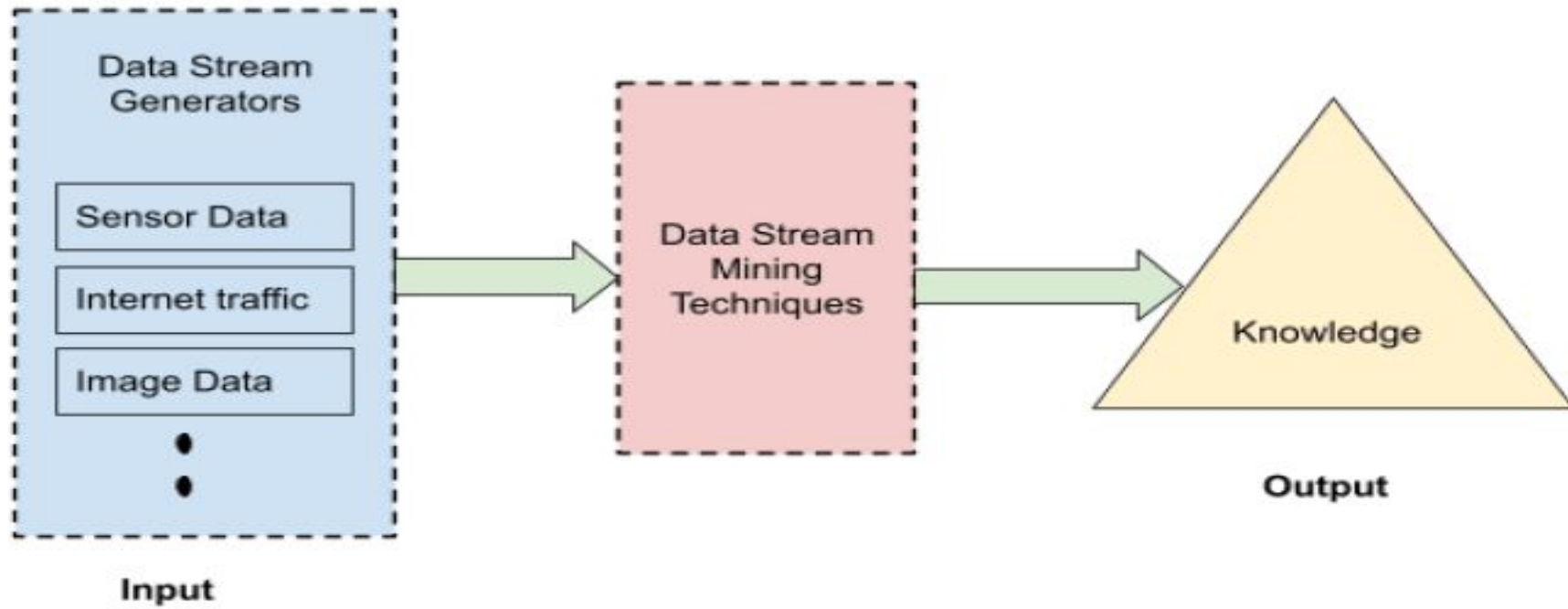
- Data Stream is a continuous, fast-changing, and ordered chain of data transmitted at a very high speed.
- It is an ordered sequence of information for a specific interval.
- The sender's data is transferred from the sender's side and immediately shows in data streaming at the receiver's side.
- Streaming does not mean downloading the data or storing the information on storage devices.

# Data Streams

- A **data stream** is a (potentially unbounded) sequence of tuples. Each tuple consist of a set of attributes, similar to a row in database table.
- **Transactional data streams**: log interactions between entities
  - Credit card: purchases by consumers from merchants
  - Telecommunications: phone calls by callers to dialed parties
  - Web: accesses by clients of resources at servers
- **Measurement data streams**: monitor evolution of entity states
  - Sensor networks: physical phenomena, road traffic
  - IP network: traffic at router interfaces
  - Earth climate: temperature, moisture at weather stations

# SOURCES OF DATA STREAM

- Internet traffic
- Sensors data
- Real-time ATM transaction
- Live event data
- Call records
- Satellite data
- Audio listening
- Watching videos
- Real-time surveillance systems
- Online transactions



### **1. Sensor Data –**

In navigation systems, sensor data is used. Imagine a temperature sensor floating about in the ocean, sending back to the base station a reading of the surface temperature each hour. The data generated by this sensor is a stream of real numbers. We have 3.5 terabytes arriving every day and we for sure need to think about what we can be kept continuing and what can only be archived.

### **2. Image Data –**

Satellites frequently send down-to-earth streams containing many terabytes of images per day. Surveillance cameras generate images with lower resolution than satellites, but there can be numerous of them, each producing a stream of images at a break of 1 second each.

### **3. Internet and Web Traffic –**

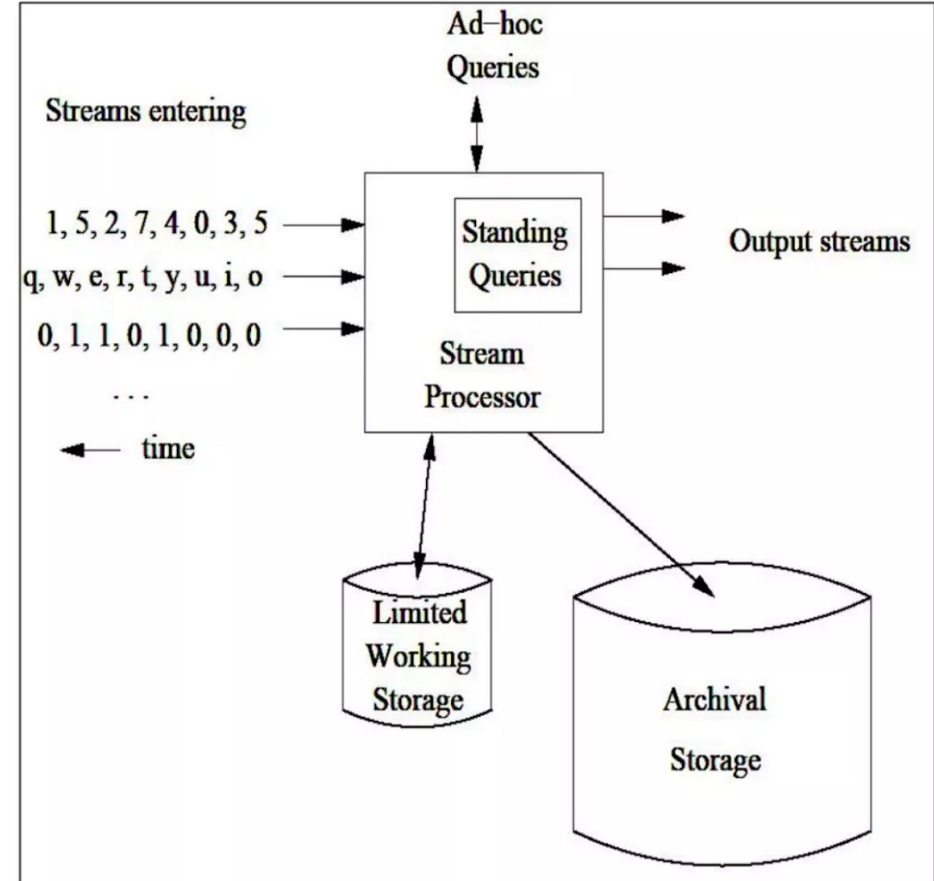
A bobbing node in the center of the internet receives streams of IP packets from many inputs and paths them to its outputs. Websites receive streams of heterogeneous types. For example, Google receives a hundred million search queries per day.

# DATA STREAMS IN DATA MINING

- Data Streams in Data Mining is extracting knowledge and valuable insights from a continuous stream of data using stream processing software.
- Data Streams in Data Mining can be considered a subset of general concepts of machine learning, knowledge extraction, and data mining.
- In Data Streams in Data Mining, data analysis of a large amount of data needs to be done in real-time.
- The structure of knowledge is extracted in data stream mining represented in the case of models and patterns of infinite streams of information.

# A data-stream-management system (DSMS)

- Streams may be archived in a large **archival store**, but we assume it is not possible to answer queries from the archival store.
- It could be examined only under special circumstances using time-consuming retrieval processes.
- There is also a **working store**, into which summaries or parts of streams may be placed, and which can be used for answering queries.
- The working store might be disk, or it might be main memory, depending on how fast we need to process queries.
- But either way, it is of sufficiently limited capacity that it cannot store all the data from all the streams.





- Any number of streams can enter the system.
- Each stream can provide elements at its own schedule; they need not have the same data rates or data types, and the time between elements of one stream need not be uniform.
- The fact that the rate of arrival of stream elements is not under the control of the system distinguishes stream processing from the processing of data that goes on within a database-management system.
- The latter system controls the rate at which data is read from the disk, and therefore never has to worry about data getting lost as it attempts to execute queries.

# Problems on Data Streams

- **Types of queries one wants on answer on a stream:**
  - **Sampling data from a stream**
    - Construct a random sample
  - **Queries over sliding windows**
    - Number of items of type  $x$  in the last  $k$  elements of the stream
- We will examine these two problems

- **Types of queries one wants on answer on a stream:**
  - **Filtering a data stream**
    - Select elements with property  $x$  from the stream
  - **Counting distinct elements**
    - Number of distinct elements in the last  $k$  elements of the stream
  - **Estimating moments**
    - Estimate avg./std. dev. of last  $k$  elements
  - **Finding frequent elements**

# Applications – (1)

- **Mining query streams**
  - Google wants to know what queries are more frequent today than yesterday
- **Mining click streams**
  - Yahoo wants to know which of its pages are getting an unusual number of hits in the past hour
- **Mining social network news feeds**
  - E.g., look for trending topics on Twitter, Facebook

## Applications – (2)

- **Sensor Networks**
  - Many sensors feeding into a central controller
- **Telephone call records**
  - Data feeds into customer bills as well as settlements between telephone companies
- **IP packets monitored at a switch**
  - Gather information for optimal routing
  - Detect denial-of-service attacks

# Applications of data stream processing

- ***Data stream processing***

- Process queries (compute statistics, activate alarms)
- Apply data mining algorithms

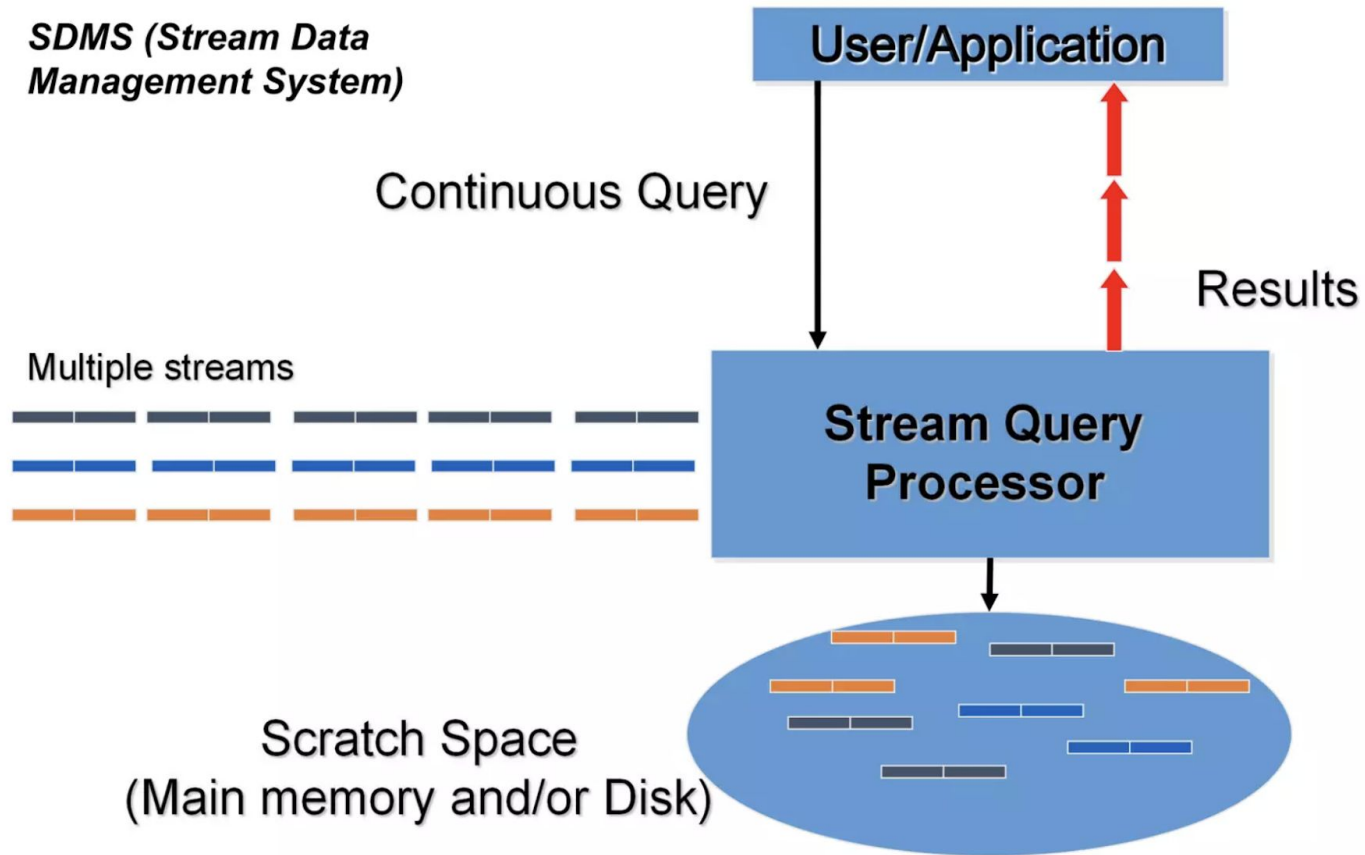
- ***Requirements***

- Real-time processing
- One-pass processing
- Bounded storage (no complete storage of streams)
- Possibly consider several streams

# CHARACTERISTICS OF DATA STREAM IN DATA MINING

- **Continuous Stream of Data:** The data stream is an infinite continuous stream resulting in big data. In data streaming, multiple data streams are passed simultaneously.
- **Time Sensitive:** Data Streams are time-sensitive, and elements of data streams carry timestamps with them. After a particular time, the data stream loses its significance and is relevant for a certain period.
- **Data Volatility:** No data is stored in data streaming as It is volatile. Once the data mining and analysis are done, information is summarized or discarded.
- **Concept Drifting:** Data Streams are very unpredictable. The data changes or evolves with time, as in this dynamic world, nothing is constant.

# Architecture: Stream Query Processing





# Data Stream Management Systems

	DBMS	DSMS
Data model	Permanent updatable relations	Streams and permanent updatable relations
Storage	Data is stored on disk	Permanent relations are stored on disk Streams are processed on the fly
Query	SQL language Creating structures Inserting/updating/deleting data Retrieving data (one-time query)	SQL-like query language Standard SQL on permanent relations Extended SQL on streams with windowing Continuous queries
Performance	Large volumes of data	Optimization of computer resources to deal with Several streams Several queries Ability to face variations in arrival rates without crash

# DBMS versus DSMS (Data Stream Management System)

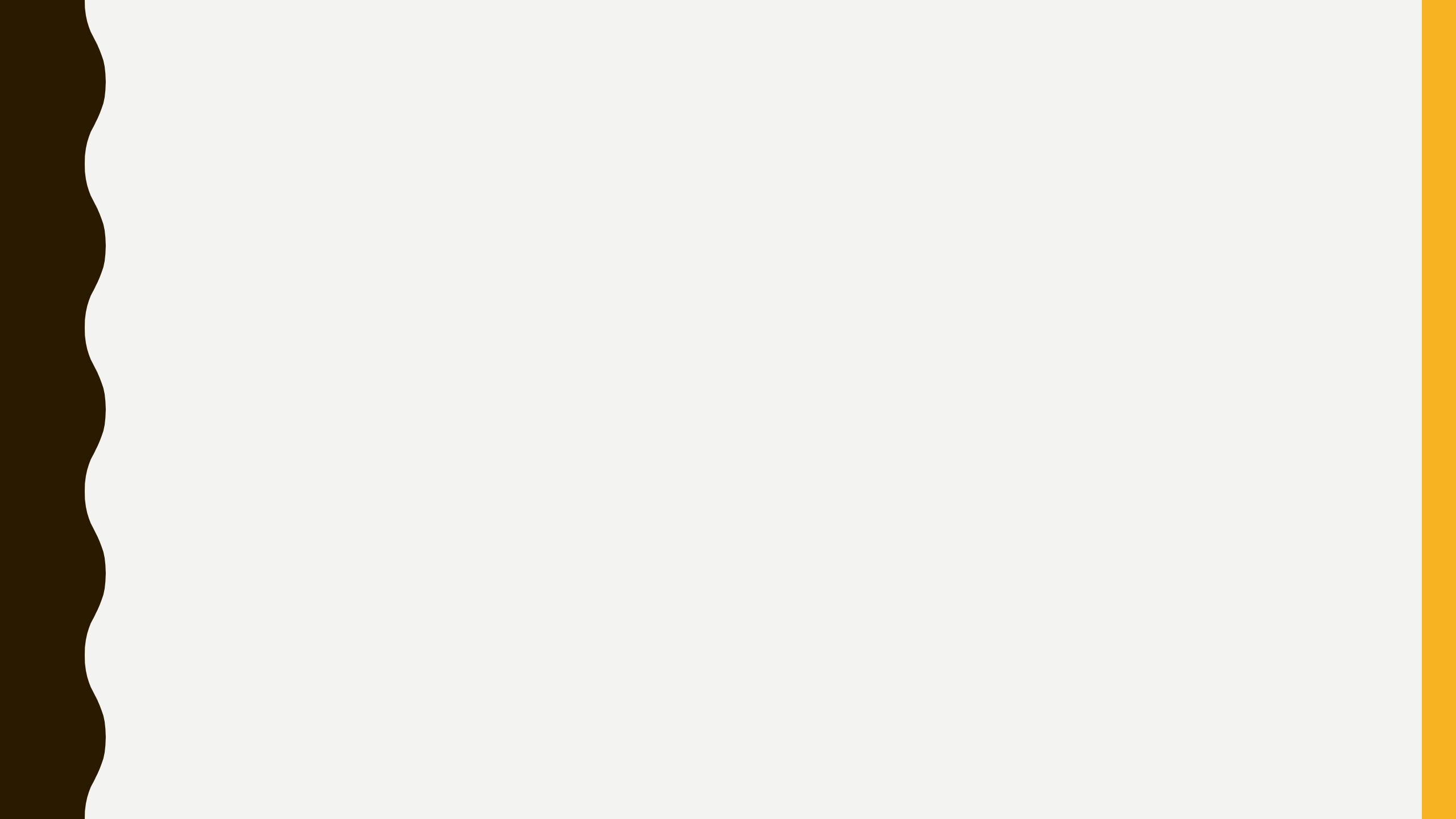
- Persistent relations
- One-time queries
- Random access
- “Unbounded” disk store
- Only current state matters
- No real-time services
- Relatively low update rate
- Data at any granularity
- Assume precise data
- Access plan determined by query processor, physical DB design
- Transient streams
- Continuous queries
- Sequential access
- Bounded main memory
- Historical data is important
- Real-time requirements
- Possibly multi-GB arrival rate
- Data at fine granularity
- Data stale/imprecise
- Unpredictable/variable data arrival and characteristics

# Challenges of Stream Data Processing

- **Multiple, continuous, rapid, time-varying, ordered** streams
- **Main memory** computations
- Queries are often **continuous**
  - Evaluated continuously as stream data arrives
  - Answer updated over time
- Queries are often **complex**
  - Beyond element-at-a-time processing
  - Beyond stream-at-a-time processing
  - Beyond relational queries (scientific, data mining, OLAP)
- **Multi-level/multi-dimensional** processing and data mining
  - Most stream data are at low-level or multi-dimensional in nature

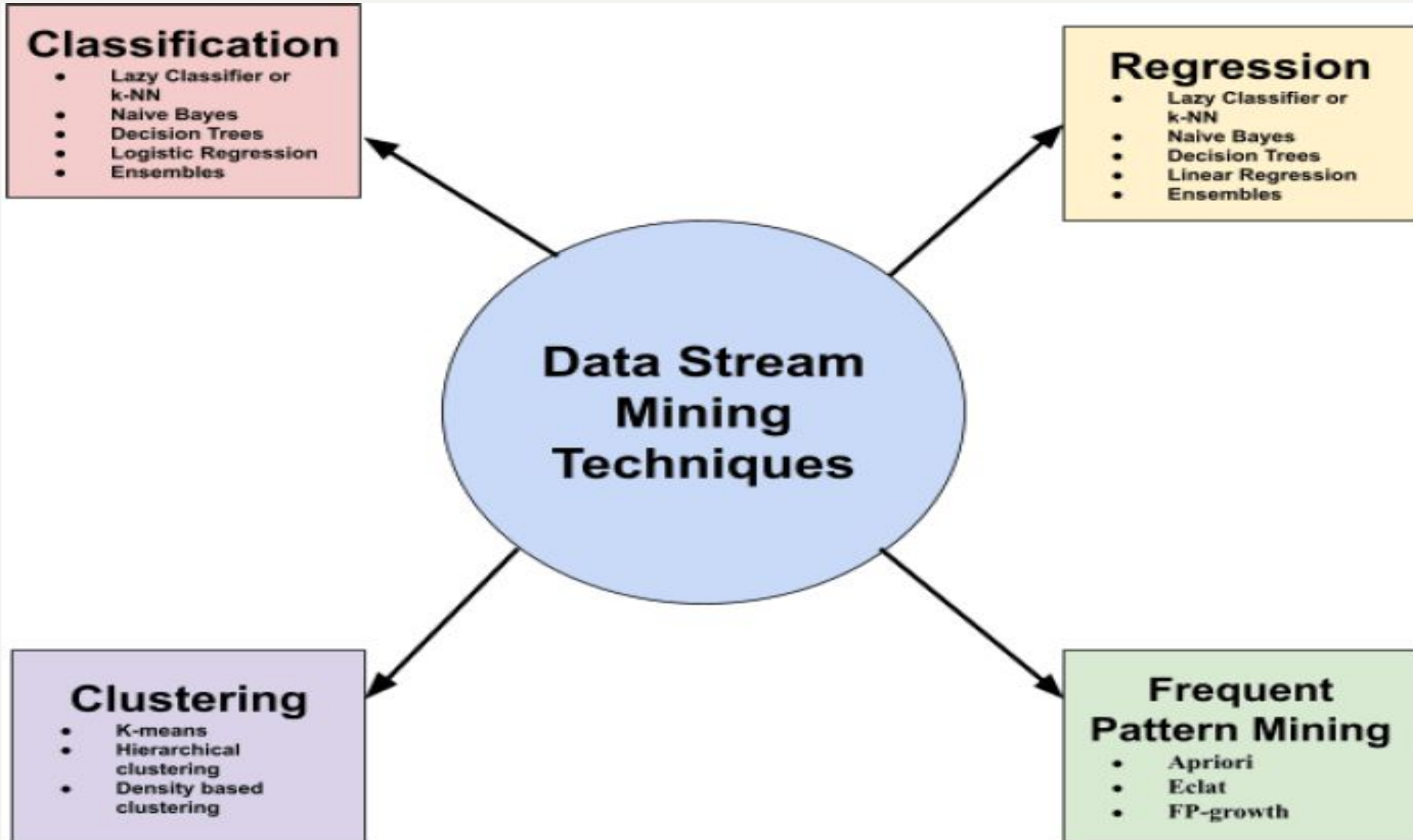
# Straming Computing Approaches

- Two approaches for handling such streams
  - Use a time window, and query the window as a static table
- When you can't store collected data, or to keep track of historical data
  - **Sampling**
  - **Filtering**
  - **Counting**



# DATA STREAMS IN DATA MINING TECHNIQUES

- Data Streams in Data Mining techniques are implemented to extract patterns and insights from a data stream.
- A vast range of algorithms is available for stream mining.
- There are four main algorithms used for Data Streams in Data Mining techniques.







# CLASSIFICATION

- Classification is a supervised learning technique. In classification, the classifier model is built based on the training data(or past data with output labels). This classifier model is then used to predict the label for unlabeled instances or items continuously arriving through the data stream. Prediction is made for the unknown/new items that the model never saw, and already known instances are used to train the model.
- Generally, a stream mining classifier is ready to do either one of the tasks at any moment:
- Receive an unlabeled item and predict it based on its current model.
- Receive labels for past known items and use them for training the model.



- 
- 
- **Lazy Classifier or k-Nearest Neighbour**
  - **Naive Bayes**
  - **Decision Trees**
  - **Logistic Regression**

# REGRESSION

- Regression is also a supervised learning technique used to predict real values of label attributes for the stream instances, not the discrete values like classification. However, the idea of regression is similar to classification either to predict the real-values label for the unknown items using the regressor model or train and adjust the model using the known data with the label.

- Regression Algorithms are also the same as classification algorithms. Below are the best-known regression algorithms for predicting the labels for data streams.
- Lazy Classifier or k-Nearest Neighbor
- Naive Bayes
- Decision Trees
- Linear Regression
- Ensembles

# CLUSTERING

- Clustering is an unsupervised learning technique. Clustering is functional when we have unlabeled instances, and we want to find homogeneous clusters in them based on the similarities of data items. Before the clustering process, the groups are not known. Clusters are formed with continuous data streams based on data and keep on adding items to the different groups.
-

- **K-means Clustering**

- The k-means clustering method is the most used and straightforward method for clustering. It starts by randomly selecting k centroids. After that, repeat two steps until the stopping criteria are met: first, assign each instance to the nearest centroid, and second, recompute the cluster centroids by taking the mean of all the items in that cluster.

- **Hierarchical Clustering**

- In hierarchical clustering, the hierarchy of clusters is created as dendrograms. For example, PERCH is a hierarchical algorithm used for clustering online data streams.

- **Density-based Clustering**

- DBSCAN is used for density-based clustering. It is based on the natural human clustering approach.

# FREQUENT PATTERN MINING

- Frequent pattern mining is an essential task in unsupervised learning. It is used to describe the data and find the association rules or discriminative features in data that will further help classification and clustering tasks. It is based on two rules.
  - **Frequent Item Set**- Collection of items occurring together frequently.
  - **Association Rules**- Indicator of the strong relationship between two items.
- 
- Apriori
  - Eclat
  - FP-growth

