Data Stream Mining

Bastian Mohrmann

University of Paderborn

Motivation

Motivation

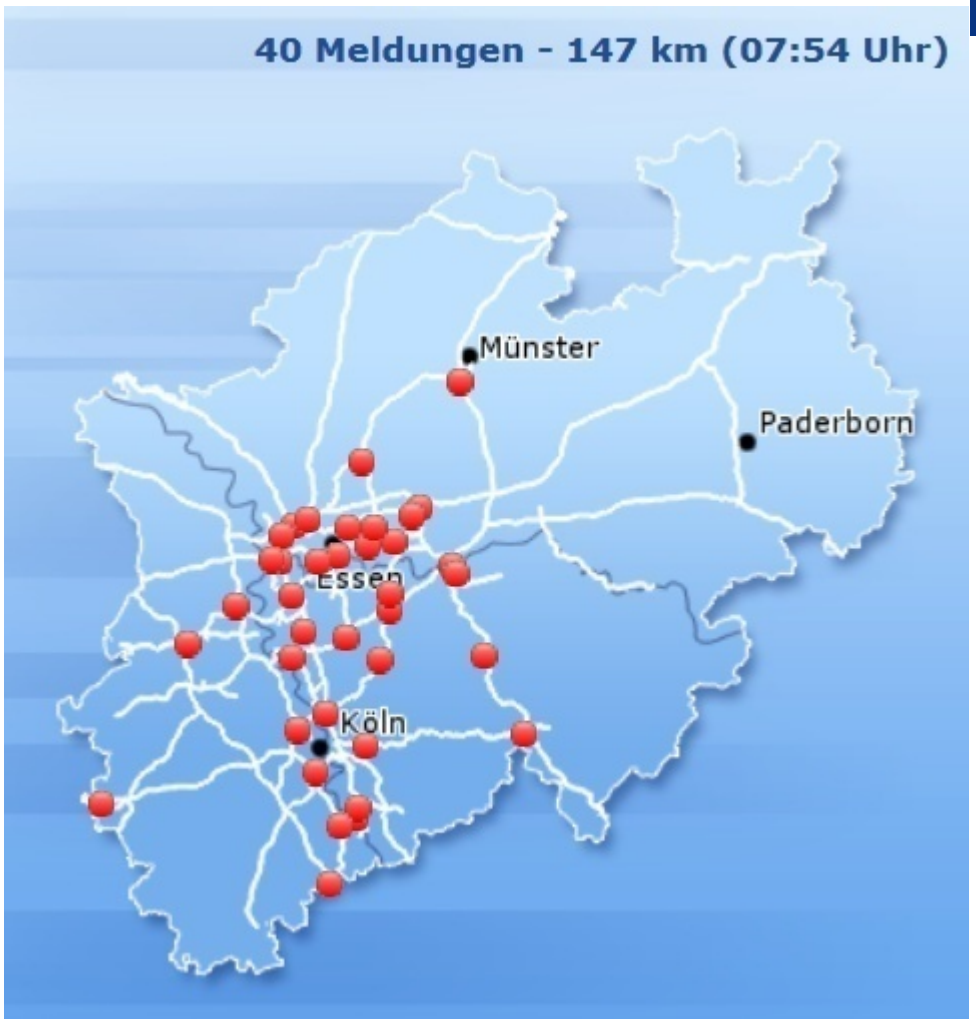
- 1. Time lost (320.000 hours)
- 2. Money lost (288 million litre fuel)

Motivation

- 1. Time lost (320.000 hours)
- 2. Money lost (288 million litre fuel)

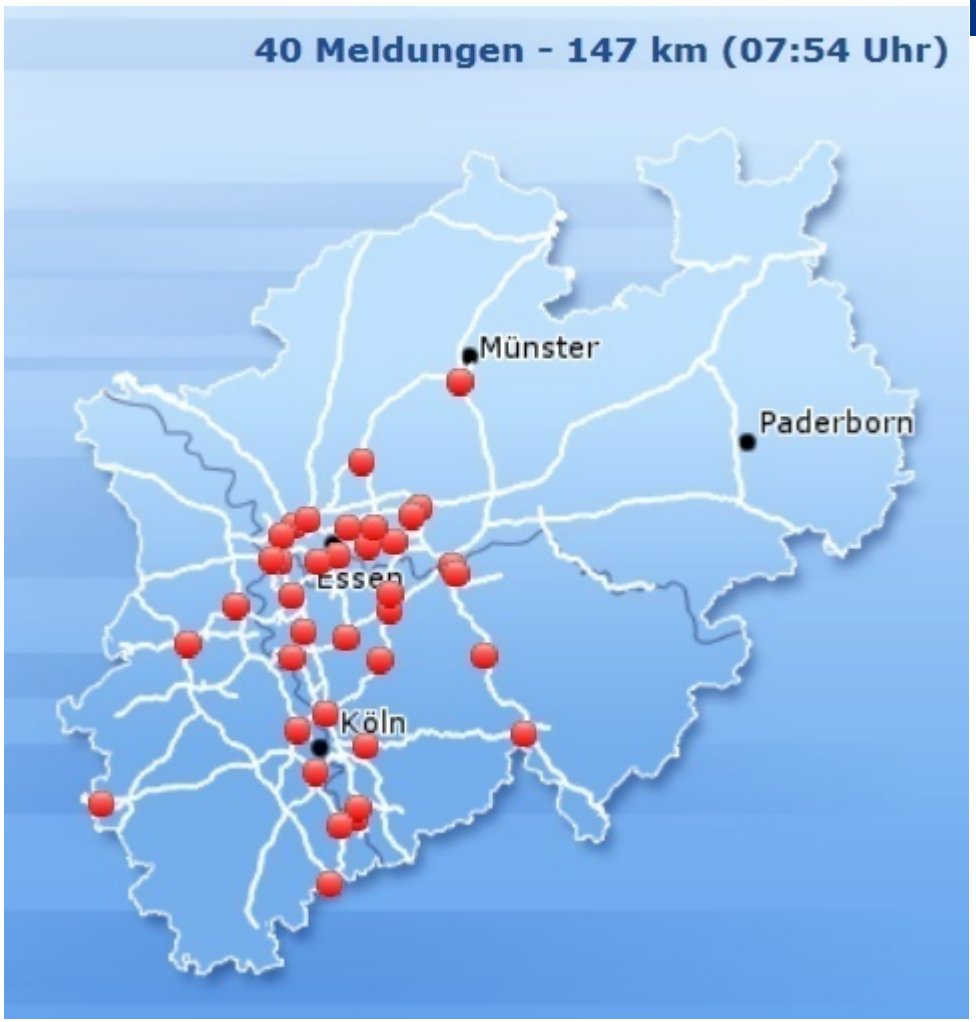
Optimization potential

Motivation – Example



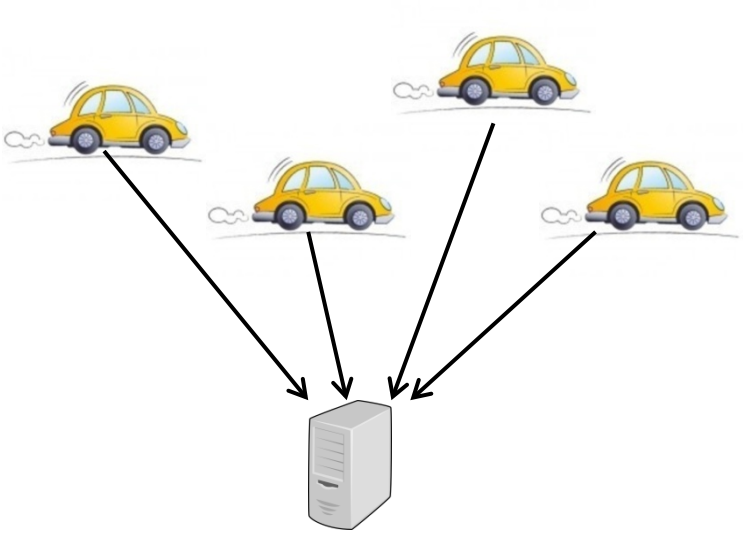
Motivation – Example

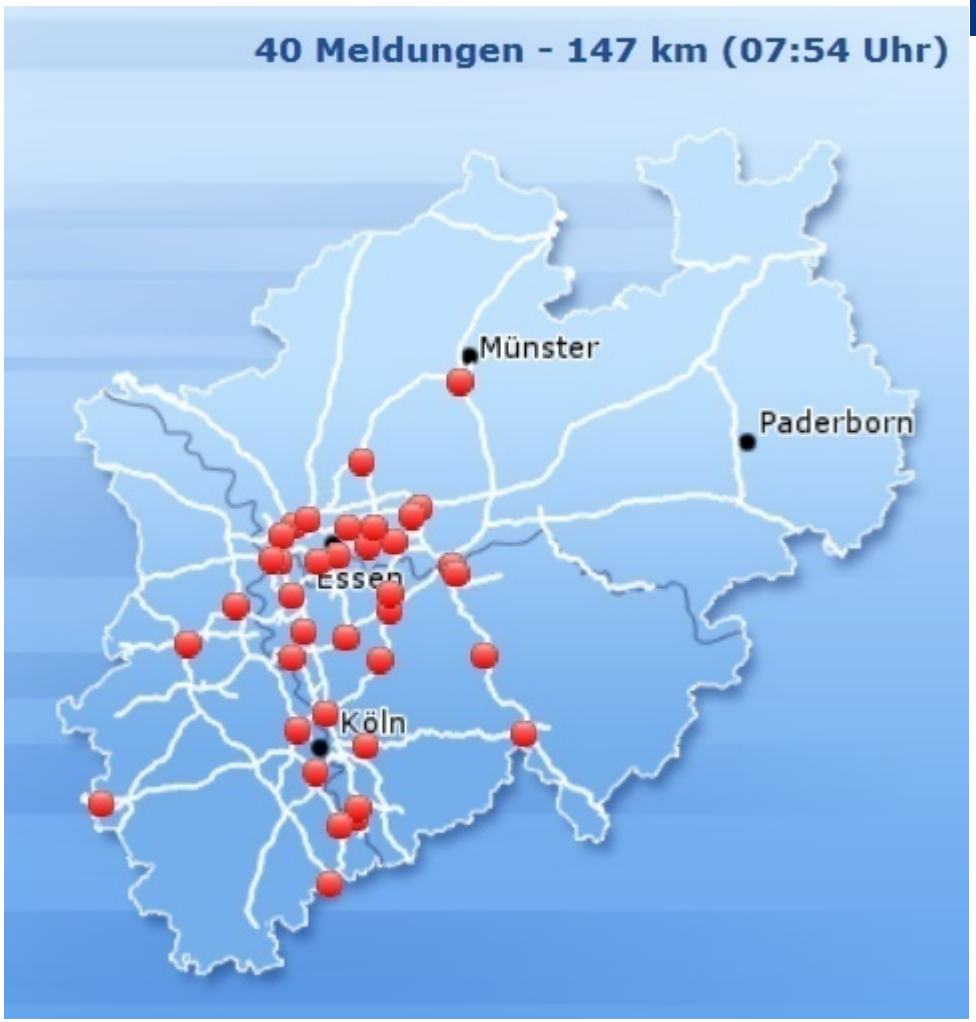
- Prediction



Motivation – Example

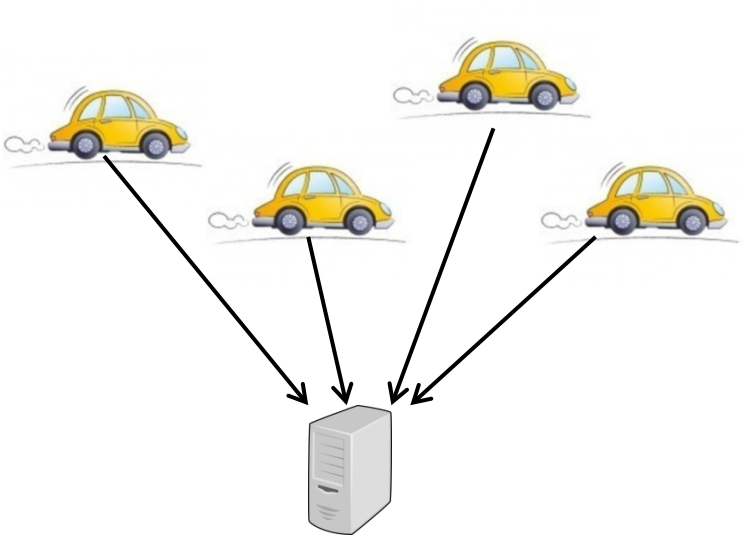
- Prediction

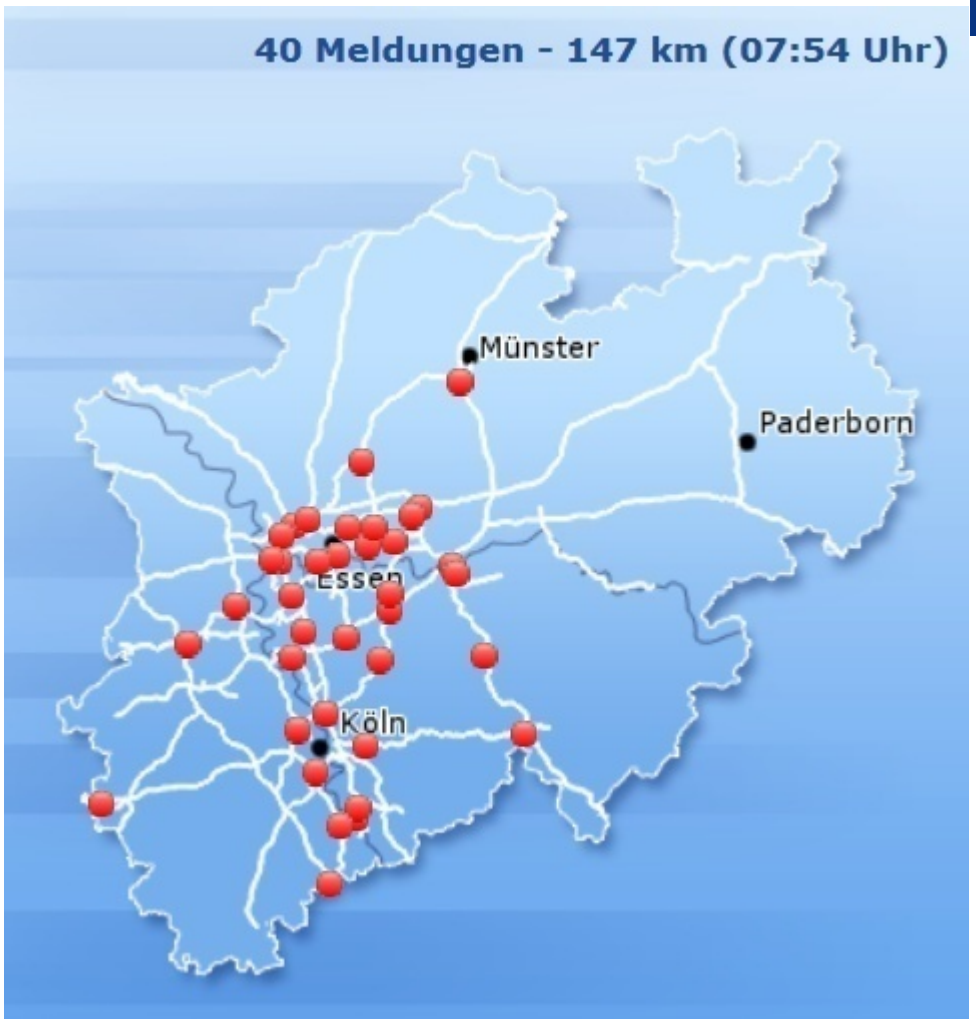




Motivation – Example

- Prediction

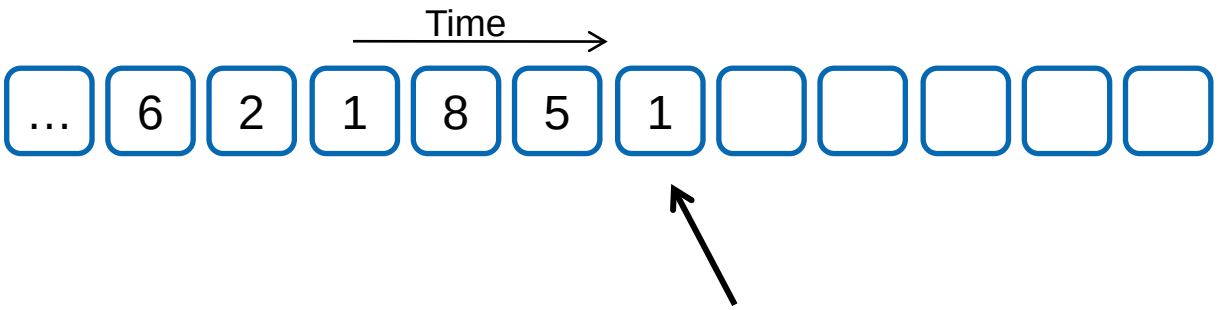




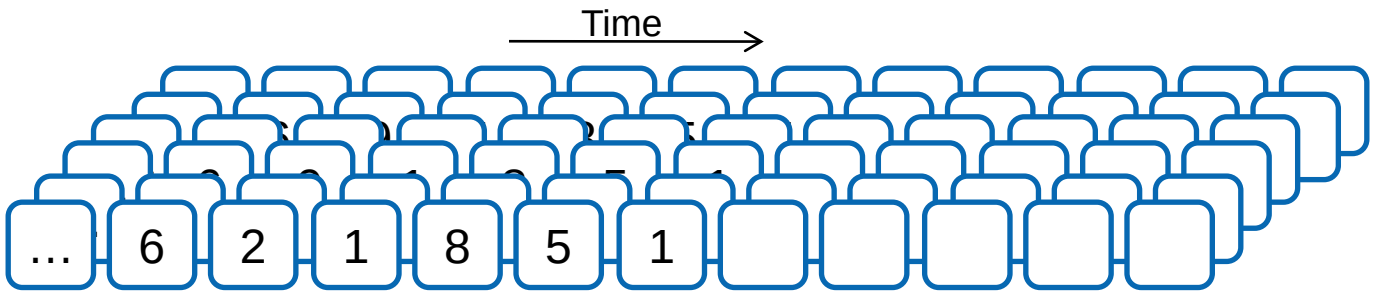
- Data Streams hold valuable information:

Hidden context

What are Data Streams?

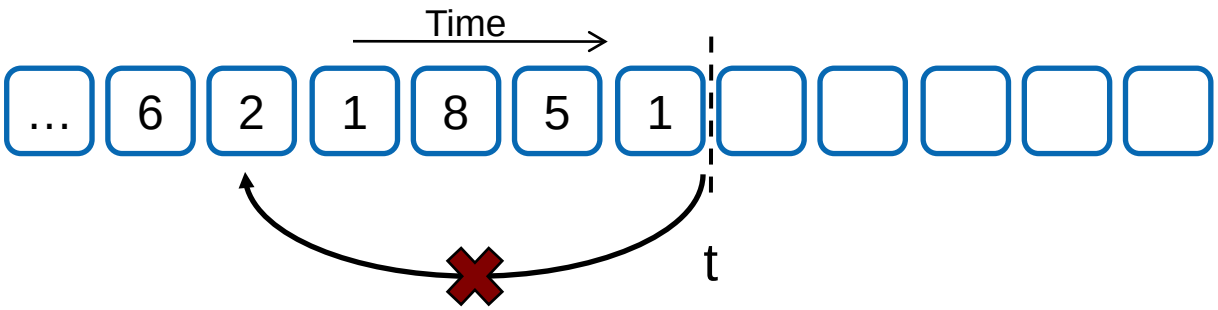


Data Streams - Characteristics



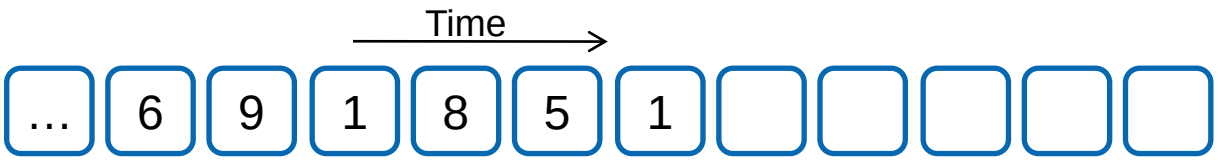
- 1. Big volume
  - Not stored in main memory, no random access
  - Storage of data in a **compressed** way

Data Streams - Characteristics



- 1. Big volume
  - Not stored in main memory, no random access
  - Storage of data in a **compressed** way
- 2. Arriving continously and unbounded
  - Accessing data in sequential order
  - Processing in one ***pass***

Data Streams - Characteristics



- 1. Big volume
  - Not stored in main memory, no random access
  - Storage of data in a **compressed** way
- 2. Arriving continously and u… …ded
  - Accessing data in sequential orde…
  - Processing in one ***pass***

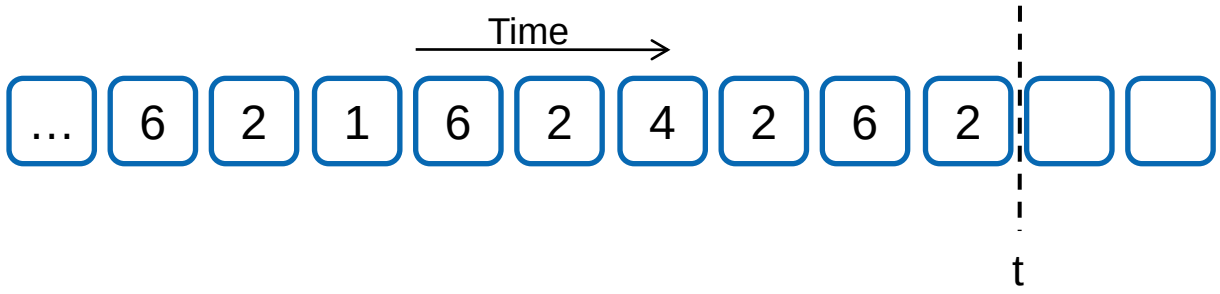
Methods/Algorithms for **Processing** and **Storing** Data Streams efficiently

Outline

1. Frequent Pattern Mining
2. Data Stream Clustering
3. Synopsis Construction
4. Summary

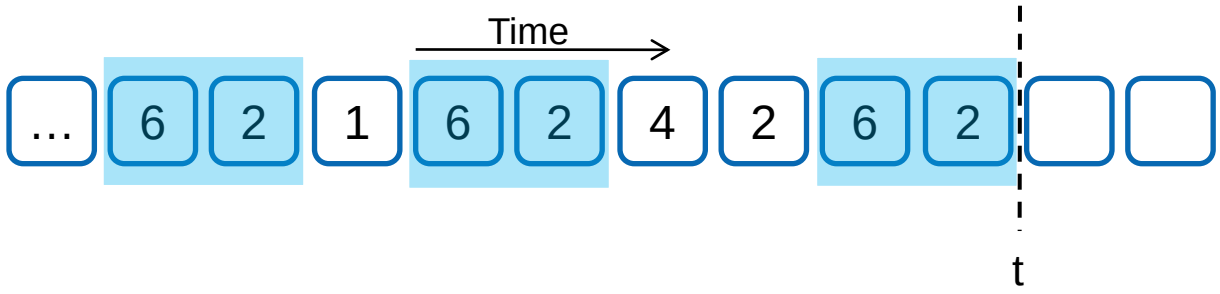
Frequent Pattern Mining

- Goal: Frequently occurring patterns/subsequences



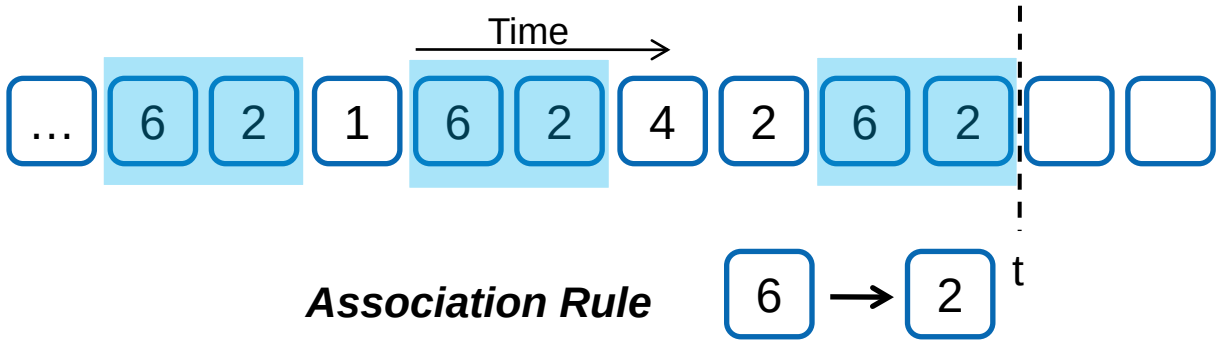
Frequent Pattern Mining

- Goal: Frequently occurring patterns/subsequences



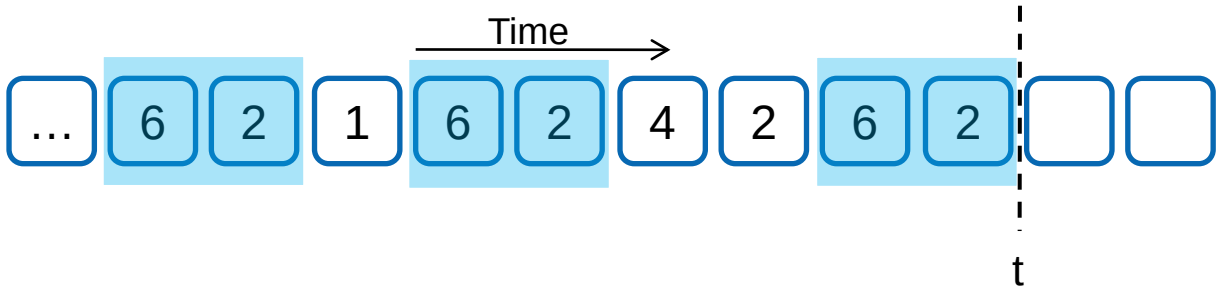
Frequent Pattern Mining

- Goal: Frequently occurring patterns/subsequences



Frequent Pattern Mining - Examples

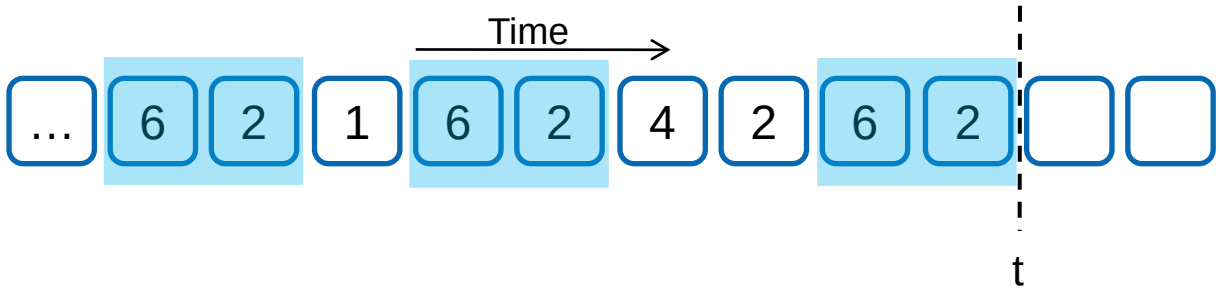
- Goal: Frequently occurring patterns/subsequences



- Web usage mining -> page association

Frequent Pattern Mining - Challenges

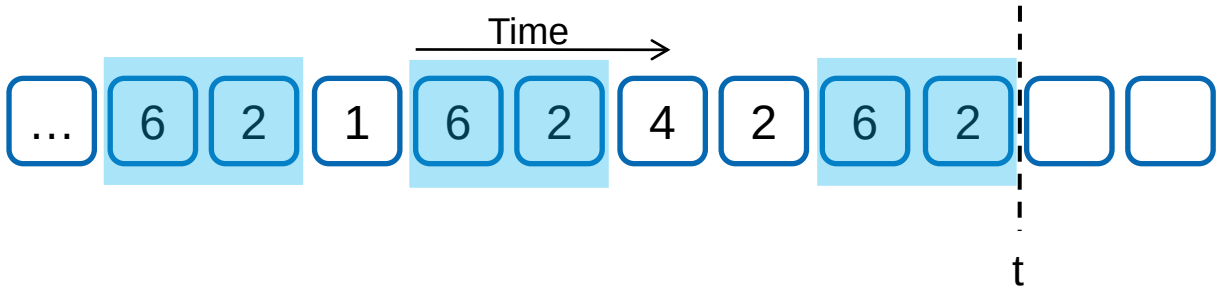
- Goal: Frequently occurring patterns/subsequences



- Mining patterns over the entire data stream is too expensive

Frequent Pattern Mining - Challenges

- Goal: Frequently occurring patterns/subsequences

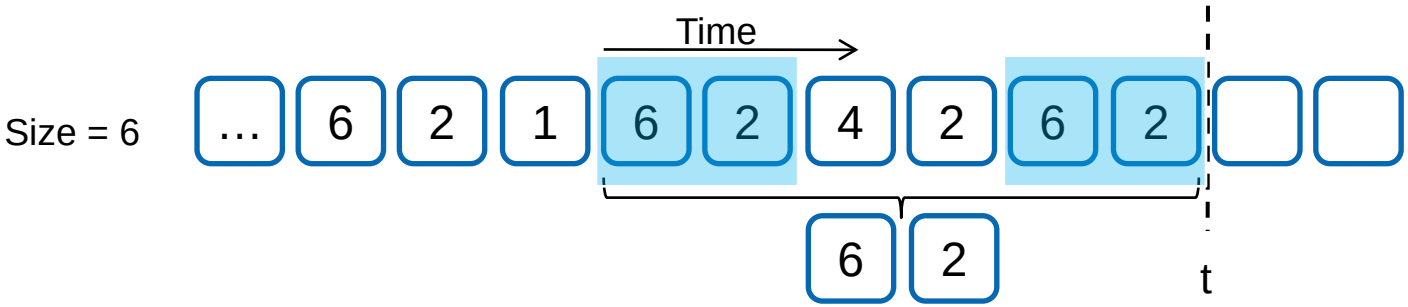


- Mining patterns over the entire data stream is too expensive

window model

Frequent Pattern Mining – Window model

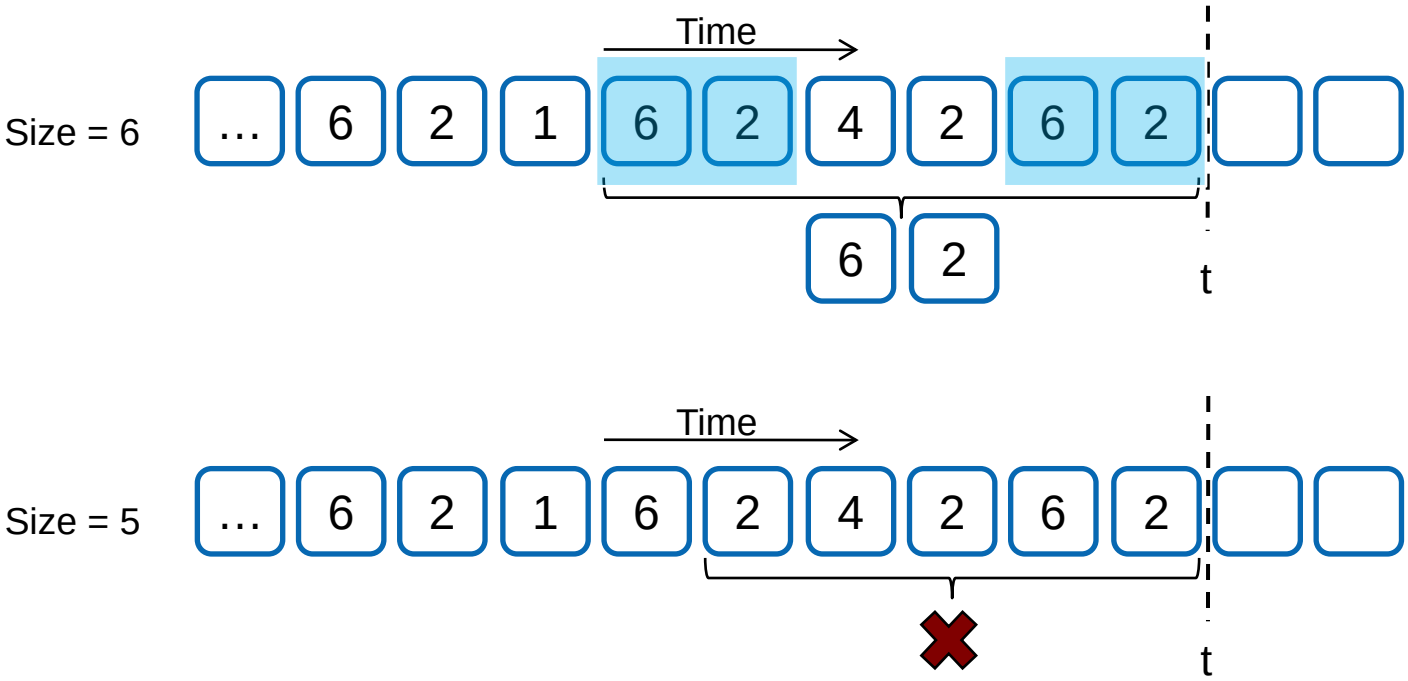
- Sliding Window



- Can be held in main memory

Frequent Pattern Mining – Window model

- Sliding Window



Resources vs. Accuracy

Frequent Pattern Mining - Summary

- Finds frequently occuring subsequences
- Window model
  - Damped window, Landmark window…
- Patterns as fundamentals for further processing

Outline

1. Frequent Pattern Mining
2. Data Stream Clustering
3. Synopsis Construction
4. Summary

Data Stream Clustering

- Goal: Uncover Structure of Data

  Identify **Evolution** of Data -> React to evolution

Data Stream Clustering - Examples

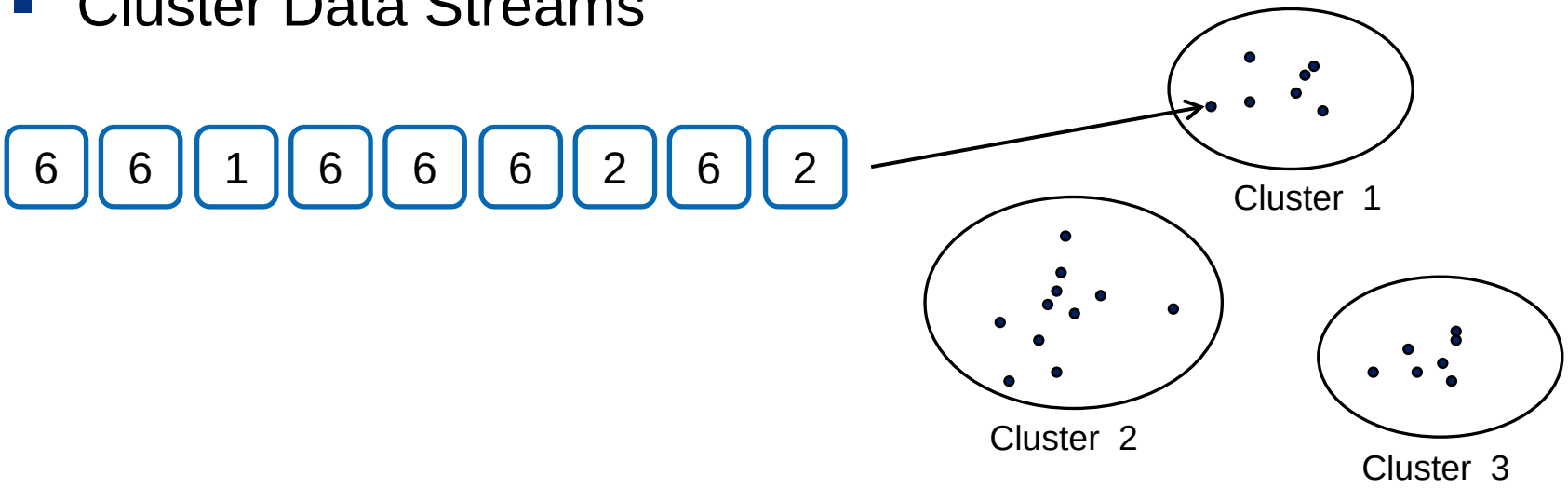
- Goal: Uncover Structure of Data

  Identify **Evolution** of Data -> React to evolution

- Tracking network data to study changes/evolution in the traffic and adapt routing
- Tracking stock market data to identify anomalies

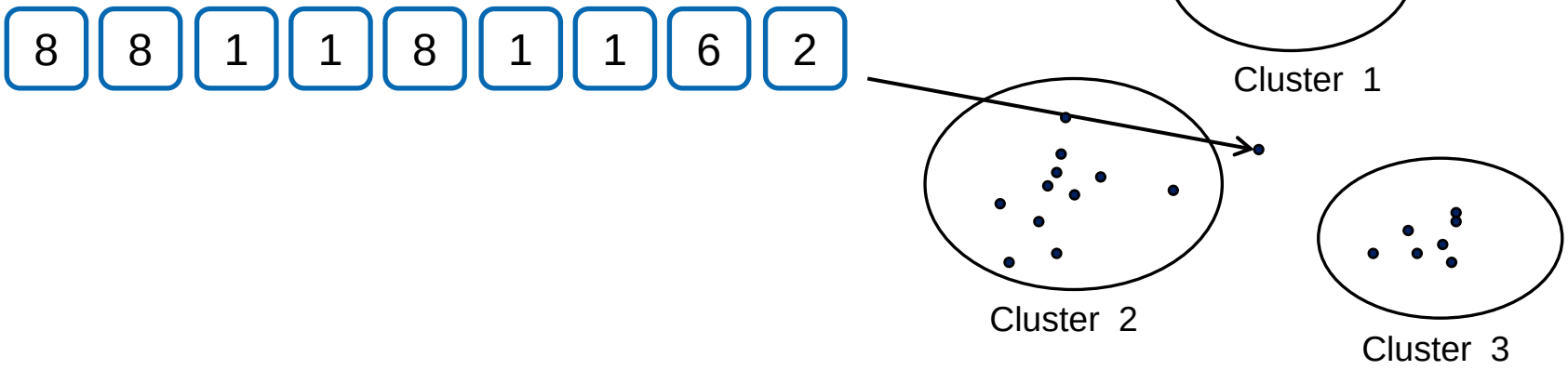
Data Stream Clustering – How?

- Cluster Data Streams



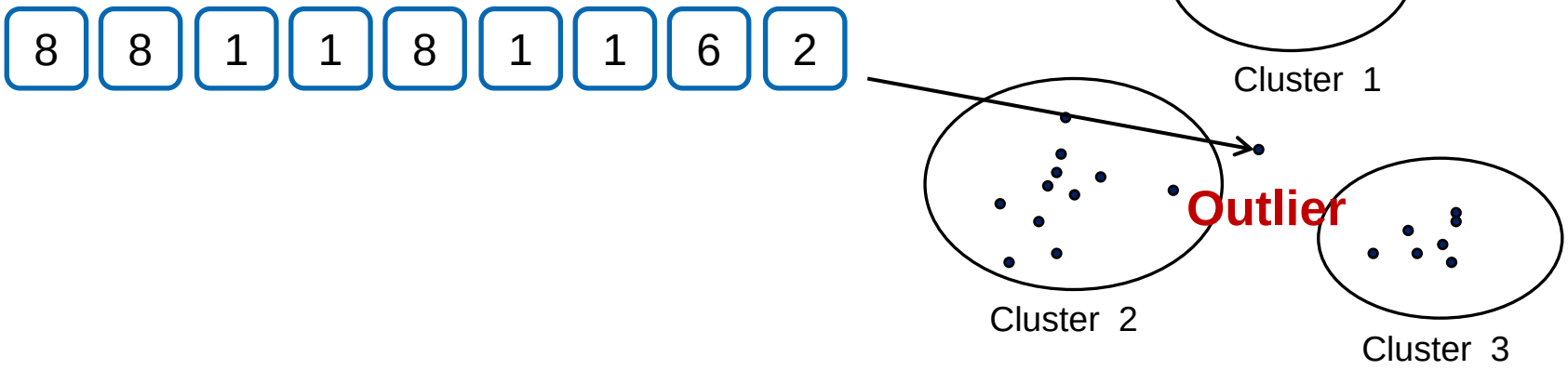
Data Stream Clustering – How?

- Cluster Data Streams



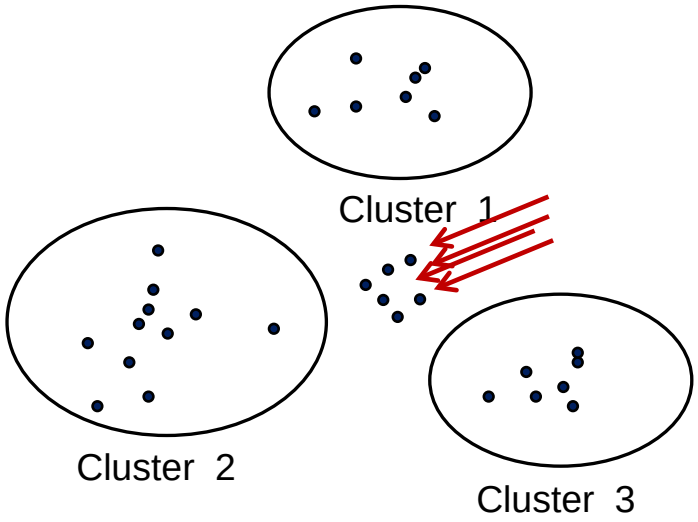
Data Stream Clustering – How?

- Cluster Data Streams



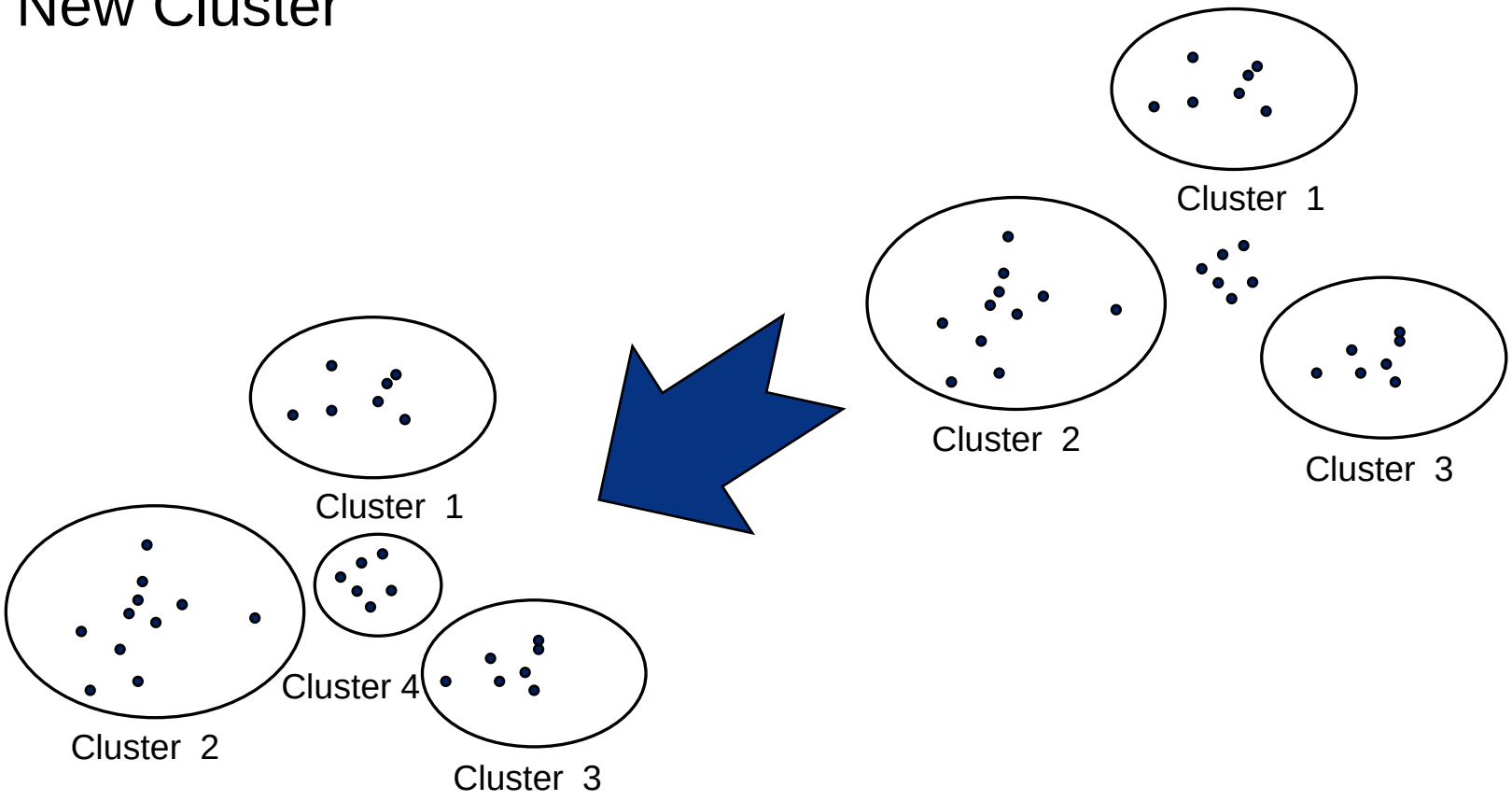
Data Stream Clustering – How?

- Multiple Outliers



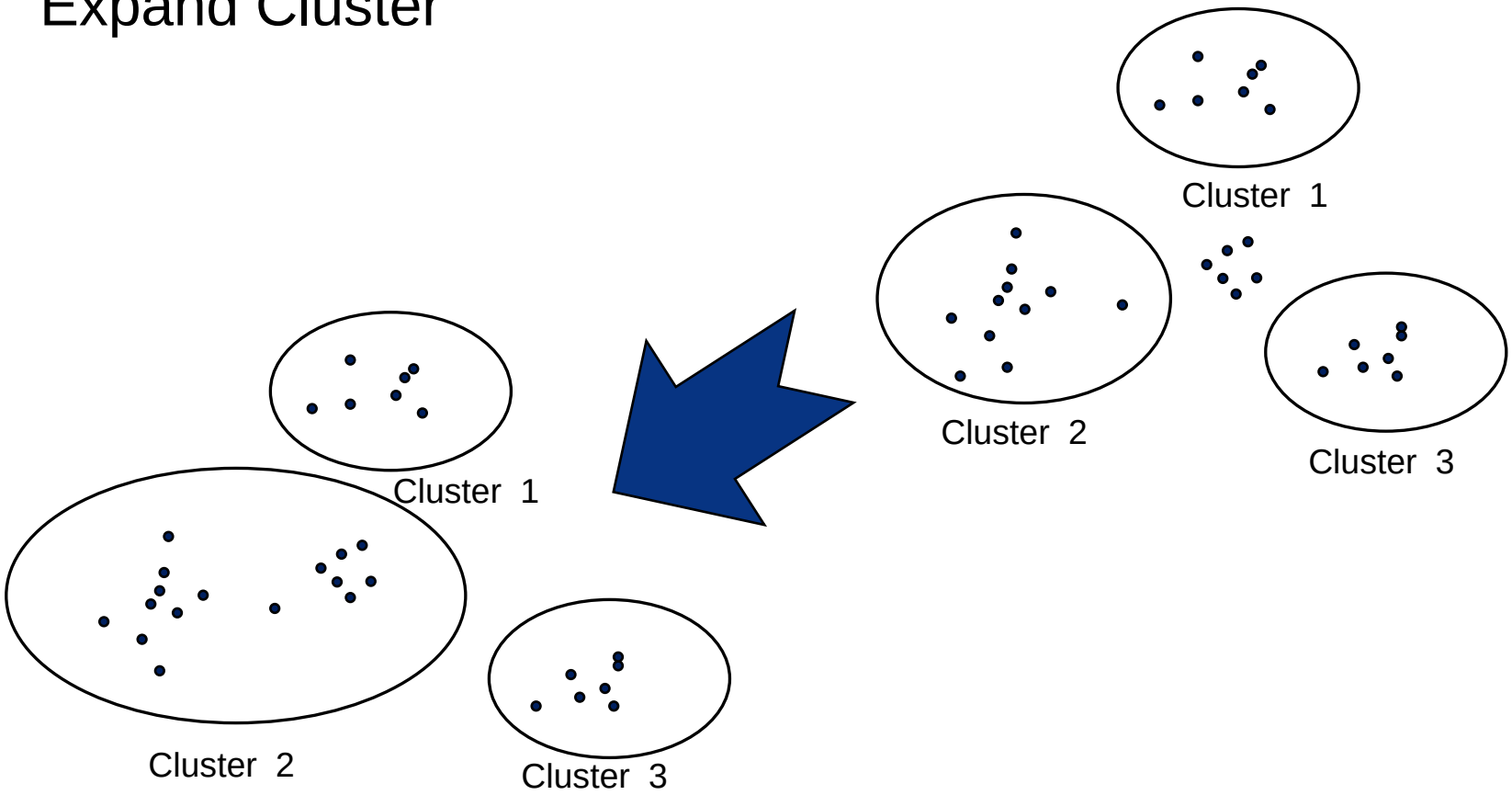
Data Stream Clustering – How?

- New Cluster



Data Stream Clustering – How?

- Expand Cluster



Data Stream Clustering - Summary

- Detect Outliers -> Data Evolution
  - Expand existing cluster or create new cluster
- Open issue: Cluster in memory?

Data Stream Clustering - Summary

- Detect Outliers -> Data Evolution
  - Expand existing cluster or create new cluster

- Open issue: Cluster in memory?

Synopsis Construction

Outline

1. Frequent Pattern Mining
2. Data Stream Clustering
3. Synopsis Construction
4. Summary

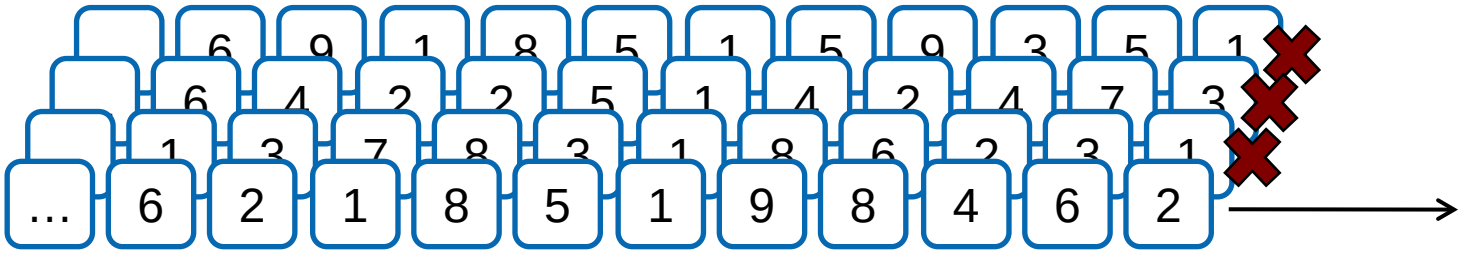
Synopsis Construction - Definition

- Goal: Compress data stream
  Do not lose information

- 1. Sampling
- 2. Histograms

Synopsis Construction - Sampling

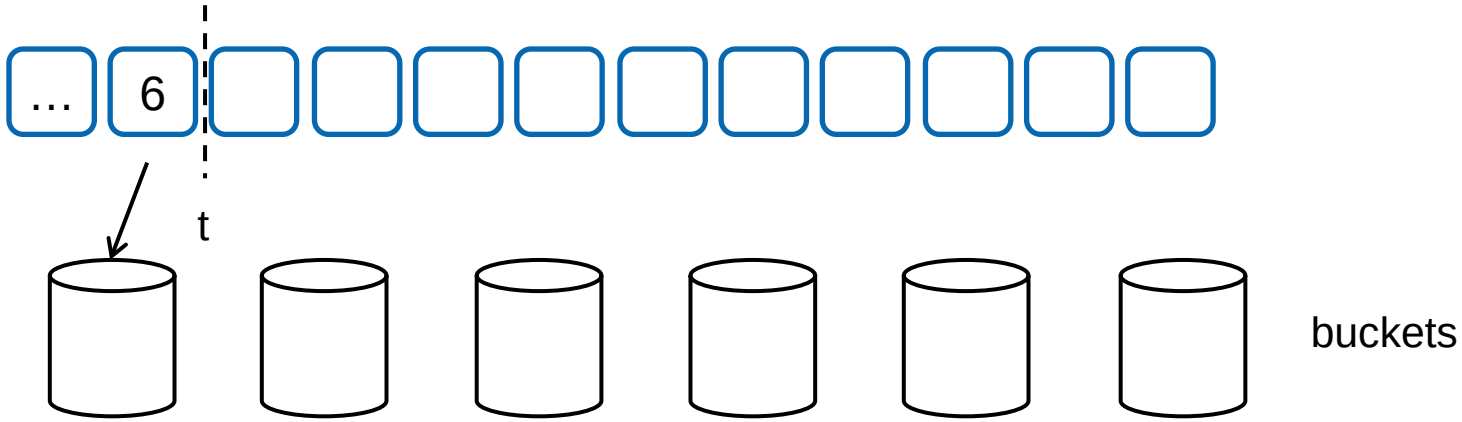
- Sampling



- processed data is equally distributed like whole data
- Broad Applicability
- Not good for identifying rare events!

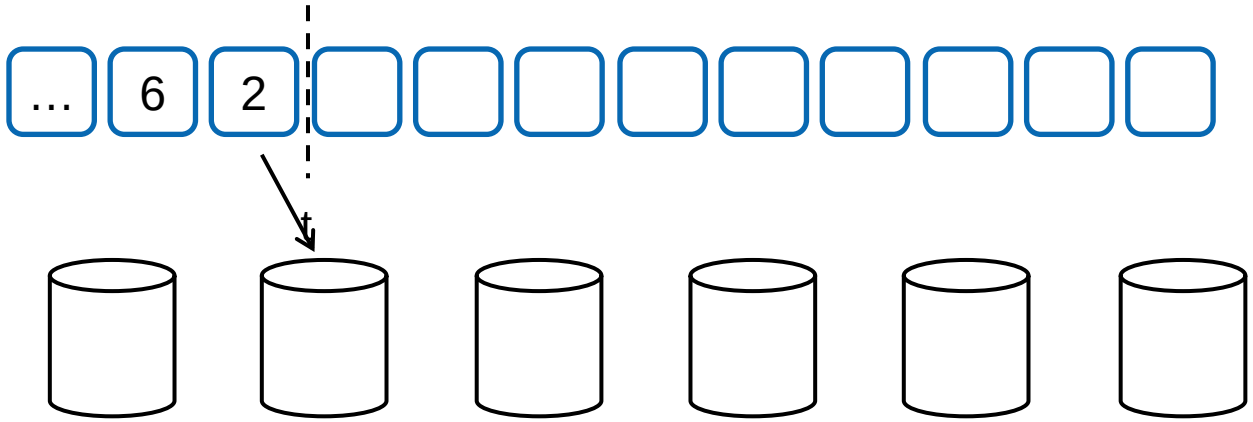
Synopsis Construction – Histogram

- Histograms



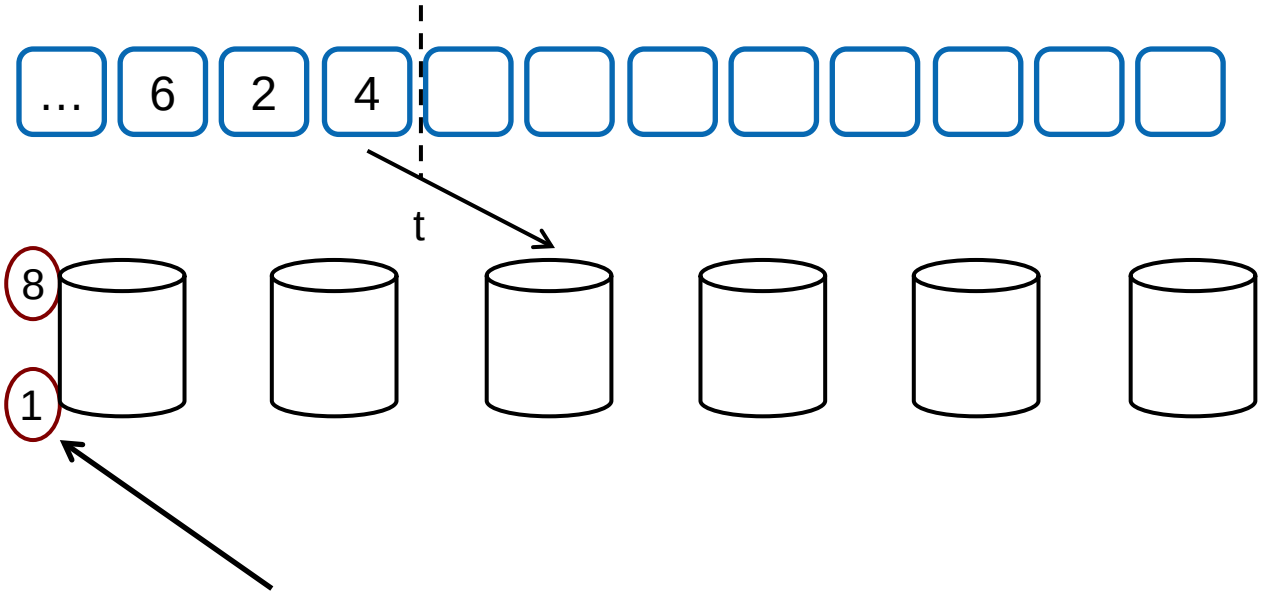
Synopsis Construction – Histogram

- Histograms



Synopsis Construction – Histogram

- Histograms



- Only boundaries are saved for each bucket

- Heavy Hitters, Anomalies still considered

Synopsis Construction - Summary

- Used for compressing data
- Synopsis structures are often used heuristically
- Different Concrete Synopsis Construction -> depends on the application field

Outline

1. Frequent Pattern Mining
2. Data Stream Clustering
3. Synopsis Construction
4. Summary

Summary

- Challenge is processing frequently arriving big amounts of data
- Different algorithms can be combined
- Concrete algorithm highly depend on application field and hidden context

References

- http://www.focus.de/auto/ratgeber/unterwegs/co2/spritvergeudung_aid_227544.html
- http://www.wdr.de/verkehrslage/

Backup Slides