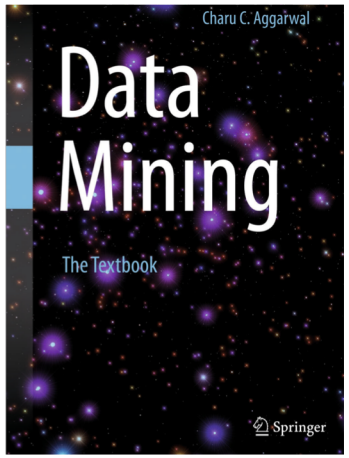


Chapter 12
Mining Data Streams

“You never step into the same stream twice.”
-Heraclitus

Karen Watts & Michael DiCicco

A data stream is a set of data the continuously accumulates rapidly over time.

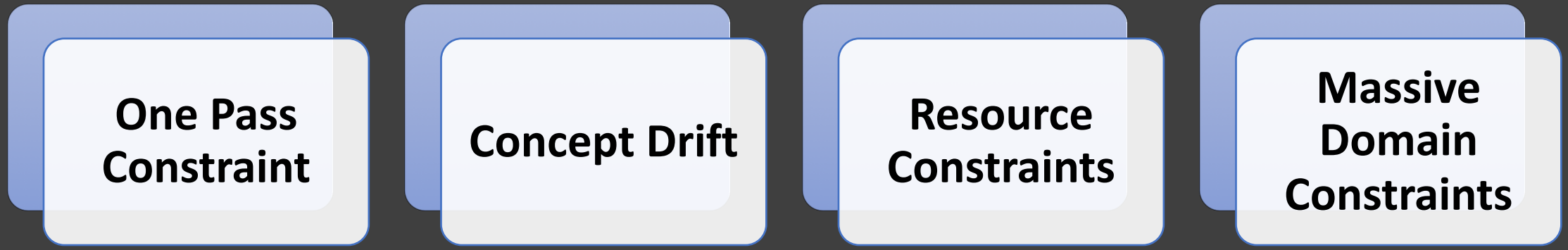
Impossible to store all the data.

What advantages do you have over big data frameworks?

4 Kinds of Streams

- Transaction Streams
- Web Click Streams
- Social Streams
- Network Streams

Design Constraints
That you have to be mindful of:



One Pass Constraint

Because of the volume and speed of data coming in, it’s assumed the data can only be processed once. Because of this, you can’t use most data mining algorithms as they are inherently iterative.

Concept Drift

Over time the relationships between the data may change.

Resource Constraints

You have very little control over the data stream.

You have very little control over the velocity of the data stream.

Because of these two things during peak periods it may be difficult to process the data in real time. To cope with this it may be necessary to drop data. This is called loadshedding.

Massive-Domain Constraints

When an attribute can have a large number of possible distinct values. Storing these can take massive amounts of space and simple things such as getting a count or the number of distinct entities can require specialized data structures.

Synopsis

Because the volume of the data.
The most common approach is to create a synopsis.
Then mine the synopsis.

Synopsis Structure Examples:

- Random Samples (Reservoir Sampling)
- Bloom Filter
- Sketches
- Distinct Element Counting Data Structures
- Clustering (Traditional Method of Data Mining)

Synopsis data structures are of two types:

Generic

Specific

Reservoir Sampling

It’s the most flexible method and it’s main advantage it can be used for any application.

Considered the method of choice for streaming scenarios.

Weaknesses

- Distinct Element Counting

The goal is to maintain a dynamically updated sample of points. Previously arrived data points that are not in the sample are lost.

With this you have to decide two things:

1

What rule should be used to be include an incoming data point in the sample?

2

What rule should be used to decide how to remove a data point from the sample?

Handling Concept Drift

Recent data considered more important.

Old data considered ”stale”.

Theoretical Bounds for Sampling

Random reservoir sampling provides samples but they may not meet some quality you need them to meet.

You may need to set bounds.

There are several ways you can do this.

The simplest of which is Markov Inequality.

Chebychev

Lower-Tail Chernoff Bound

Upper-Tail Chernoff Bound

Hoeffding Inequility

Table 12.1: Comparison of different methods used to bound tail probabilities

Result	**Scenario**	**Strength**
Chebychev	Any random variable	Weak
Markov	Nonnegative random variable	Weak
Hoeffding	Sum of independent bounded random variables	Strong
Chernoff	Sum of independent Bernoulli variables	Strong

Massive Domain

Streaming objects generally associated with pairs of identifiers.

The domain of possible values may be very large.

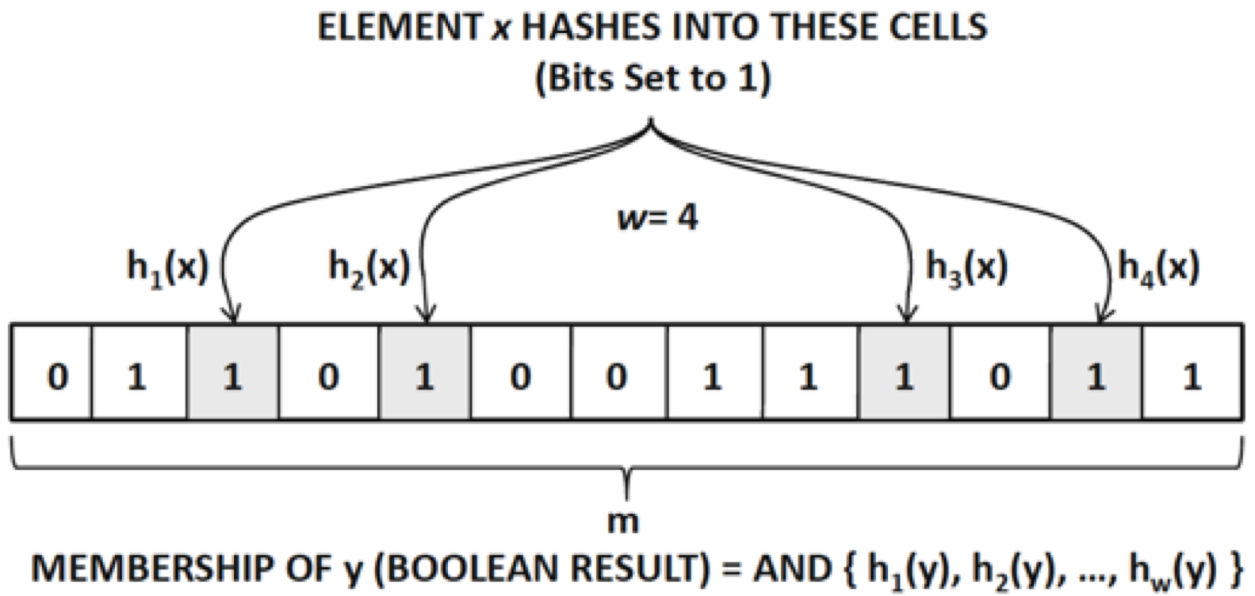
Synopsis Structures are optimized for different classes of queries.

Bloom Filter

– Optimized for set membership query.

Example: Given a particular element has it occurred in the data stream.

False positives are possible but false negatives are not possible.



Count-Min Sketch

Designed for count based summaries.

w * m 2-dimensional array in which stream elements are mapped to cells by hash functions

It can have collisions like the bloom filter, causing overestimates. Because of this, it can be helpful to set an upper bound on quality.

AMS Sketch (Alon-Matias-Szegedy)

Optimized for second moments. Random binary value from -1 to 1 is generated for each element using a hash function. They are assumed 4-wise independent.

The second-order moment can estimate the self-joined size of a data stream and can also be viewed as a variant of the Gini index.

The Flajolet-Martin Algorithm

Optimized for distinct element counting.

It uses a hash function to map an element to integer in the range [0,2^L-1]

Frequent Pattern Mining

Frequent items are also referred to as heavy hitters.

In a massive-domain scenario, just finding which items are frequent is difficult.

The other case is just a large set of numbers that fits in memory. In that case, your more likely to be interested in pattern frequency which could require multiple passes over the data set.

To achieve this there are two approaches:

Leveraging Synopsis Structures

Lossy Counting Algorithm

Leveraging Synopsis Structures

Synopsis structures are good because they allow a wider array of algorithms to be used, as well as being able to bias on the freshness of data.

Two forms:

Reservoir Sampling

Sketches

Lossy Counting Algorithm

Can be used for frequent item or itemset counting.

The frequencies are maintained in an array. If it get’s to many items in the space it will drop the less frequent items.

Clustering Data Streams

Especially good for data streams, because they provide a compact synopsis. Because of this stream clustering is often done before applying another application.

Stream Clustering Algorithms

1. STREAM based on k-medians.

Limitation of STREAM, it’s affect more by concept drift.

2. CluStream Algorithm

Applied as a two stage process. An online microclustering stage and offline macroclustering.

Pyramidal Time Frame

Table 12.2: An example [39] of snapshots stored for $\alpha = 2$ and $l = 2$

Order of Snapshots	Clock times (last five snapshots)
0	55 54 53 52 51
1	54 52 50 48 46
2	52 48 44 40 36
3	48 40 32 24 16
4	48 32 16
5	32

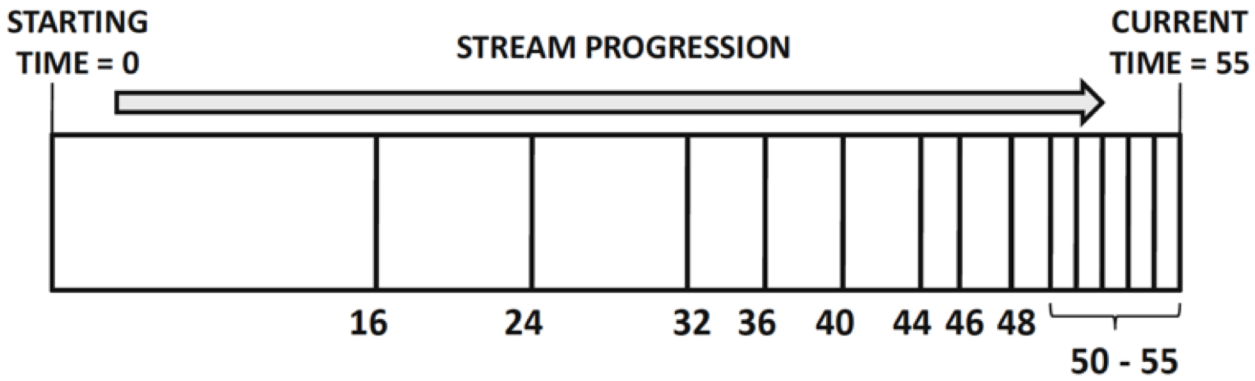


Figure 12.7: Recent snapshots are stored more frequently by pyramidal time frame

Massive-Domain Stream Clustering

CSketch was designed for clustering massive-domain data streams. Uses a count-min sketches to store the frequency of attribute values in each cluster.

Online k-means clustering is used on the sketches for each data point a dot product is computed with respect to each cluster. That product is divided by the total frequency of items in that cluster to avoid bias towards large clusters.

Streaming Outlier Detection

Two kinds of outliers for multidimensional data streams.

- Novelty - Detection of individual records.
- Aggregate – Change in trends.

Novelty Outliers|

Closely related to unsupervised novelty detection.

Distance Based Algorithms are useful within a time window.

Aggregate Change Points as Outliers

Sudden changes in trends often indicate anomalies in data.

One way of measuring how sudden these changes are is the concept of velocity density.

Streaming Classification

Especially challenging

Uses reservoir sample

Four choices:

1. VFDT
2. Supervised Microcluster (CluStream)
3. Ensemble (C4.5,RIPPER, naïve Bayesian)
4. Massive-Domain Streaming Classification

Summary

Streams present challenges related to high volume, concept drift, massive-domain nature, and resource constraints. As such, synopsis construction is the main problem when dealing with data streams.

As long as you have a quality synopsis, it can be used with stream mining algorithms.

The most common are reservoir-samples and sketches. Reservoir-samples should be used where ever possible.

Thank You!