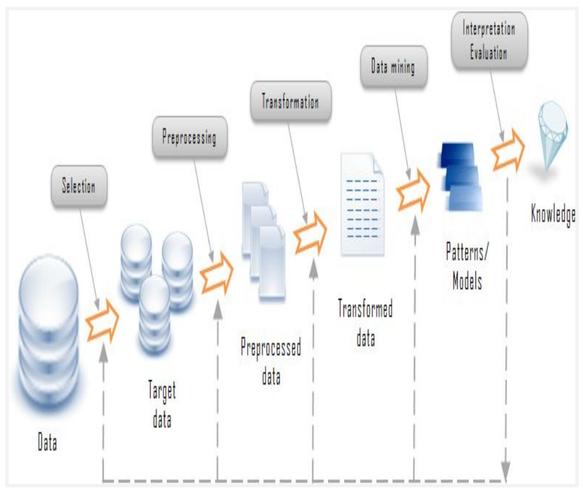
Machine Learning On Data Streams

Ankush Sharma
Kunal Arora

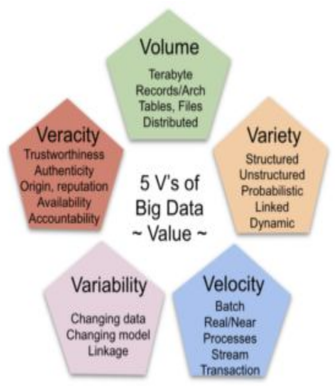
Goal

What is needed is a shift in mindset from mining databases to one pass mining algorithms that are able to mine on continuous, high volume, high velocity data streams as they arrive and take concept drift into account. In this report, we discuss the characteristics of such systems and the ongoing research and engineering effort that is currently in place.

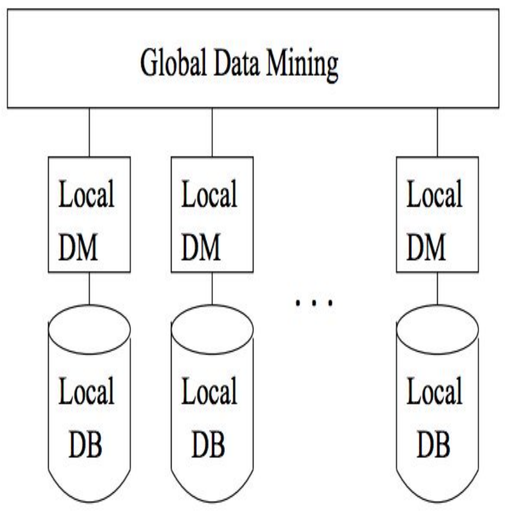
TRADITIONAL DATA MINING



The 5 V's of Big Data

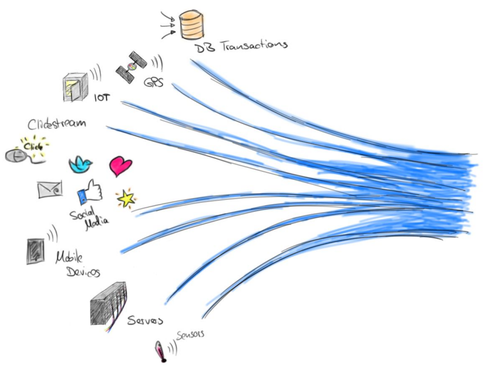


DISTRIBUTED DATA MINING



- Horizontal or Vertical partitioning
- Crosstalk is mighty expensive
- Consistency Problems
- Knowledge Integration step
- Data Privacy Issues

DATA STREAMS



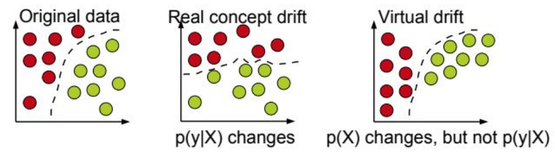
Challenges:

- **Volume** : Zero to possibly infinite
- **Velocity** : Arbitrary
- **Time and Querying Semantics** : Instantaneous and non blocking operations over time windows, ad hoc querying

CONCEPT DRIFT

- The statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen ways. This causes problems because the predictions become less accurate as time passes
- Everything is prone to drift, but more so with Data Streams

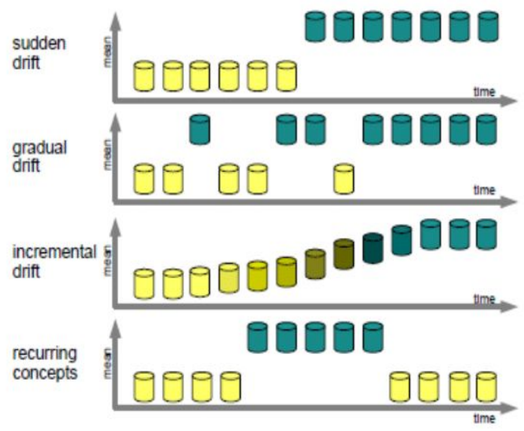
Formally: $\exists X: pt0(X,y) \neq pt1(X,y)$, where pt0 denotes the joint distribution at time t0 between the input vector X and the target variable y



Two different kinds of Concept Drifts have been identified :

1. *Real Drift*: This refers to a change in $p(y|X)$. Specifically, at time t and u, a real concept drift occurs when $pt(X,y) \neq pu(X,y)$. It changes class boundaries
2. *Virtual Drift*: A virtual drift occurs when $p(X)$ changes, without impacting $p(y|X)$

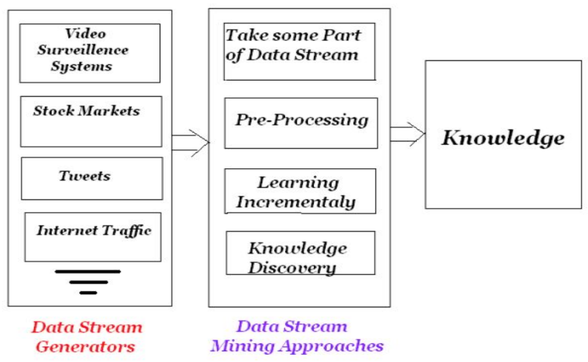
Data may change in unforeseen and unpredictable ways



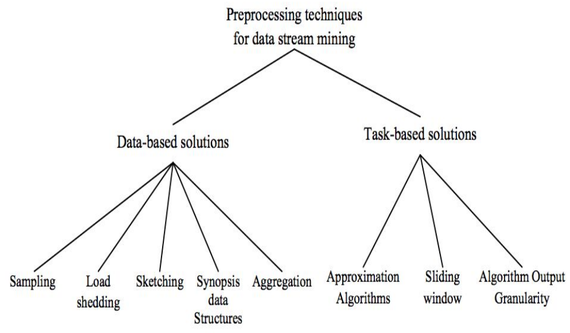
Different types of change in data have been identified :

1. *Sudden*: For example, the replacement of a sensor with another sensor that has a different calibration in a chemical plant
2. *Gradual*: A user gradually starts becoming interested in a different form of music
3. *Incremental*: A sensor becomes old and wears off
4. *Recurring*: Fashion trends

DATA STREAM MINING CHALLENGES



Data Stream Mining(DSM) is a process of mining continuous incoming real time streaming data with acceptable performance.



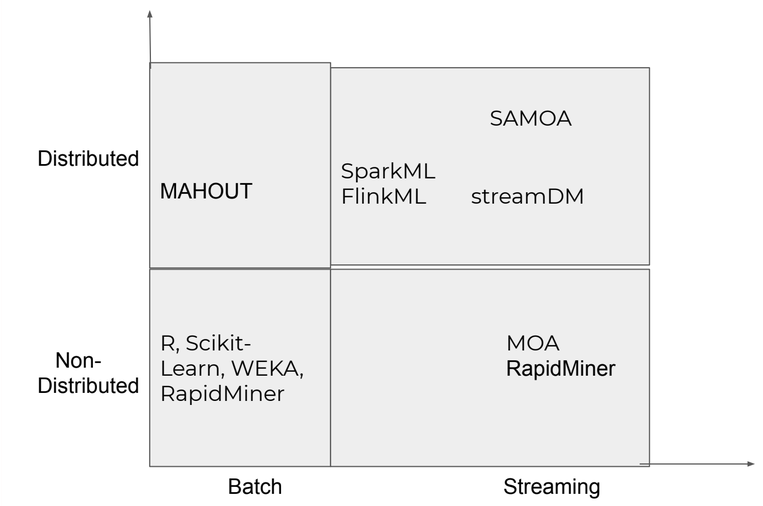
DSM algorithms can also do Pre Processing, but not all of them do.
We have identified 6 challenges that any DSM algorithm needs to address. A good DSM algorithm needs to confront all of them, though it is not necessary:

1. *Passes over data* : An ideal algorithm will have very few passes over data, ideally one
2. *No Random Access* : Since streaming data by definition is not stored, random access is not allowed
3. *Potentially infinite Volume* : The data coming into the system can be potentially infinite, and needs to be processed in limited time
4. *Optimized for Memory* : Since volume can be potentially infinite, the algorithm needs to be optimized for memory as not everything can be buffered into memory
5. *Time based operations*: The data in a stream can be multidimensional and low level, therefore query semantics for the model and the algorithm must allow fine grained access to the stream events on record level and time based operations (for example, windowing operations on a 5 minute window)
6. *Takes Drift into Account*

DATA STREAM MINING ALGORITHMS

Algorithm	Task	Advantages	Shortcomings
VFDT and CVFDT	Decision Trees	-Good processing speed -Low memory footprint -Does Pre Processing	-Does not resolve Concept Drift -Time consuming and costly learning
CDM	Decision tree and Bayes network	-Measures distance between events -Low memory footprint	-May or may not resolve Concept Drift -No PreProcessing
STREAM and LOCALSEARCH	K-Medians on streams	-Incremental Learning -Does Preprocessing	-Low accuracy -Low clustering quality on high speed
VFKM	Very fast K-means	-High Speed -Low Memory utilization -Does PreProcessing	-Multi-pass
CluStream	The concepts of a pyramidal time frame in conjunction with a micro-clustering approach. Has an online and offline component to perform learning	-Concept Drift Detection -Time and space efficiency -High accuracy -Does PreProcessing	-Offline clustering
D-Stream	The algorithm uses an online component which maps each input data record into a grid and an offline component.	- High quality and efficiency - concept drift detection in real-time data stream	-High Complexity
Approximate Frequent Counts	Frequent item sets	- Incremental update -Simplicity -Need less memory space -Single pass	-Approximate output with increasing error range possibility
FP Stream	Frequent Item Sets	-Incremental and dynamic update -Need less memory space	- High complexity

OPEN SOURCE SOFTWARE



CONCLUSION

We saw how Data mining has evolved from computation done on a centralized site, to distributed batch processing and now Data Stream Mining being a new field with research focused on concept drift adaptation and detection, memory efficient and fast algorithms. The field is still in its infancy, and the focus so far has been on Classification and Clustering algorithms, and tackling velocity and volume aspect of Big Data in streams. The tools are limited, SAMOA and streamDM being the only open source software to provide some Distributed Stream Mining algorithms.

REFERENCES

[1]https://www.researchgate.net/publication/324171539_Open_Source_Data_Mining_Programs_A_Case_Study_on_R
[2]http://www.essi.upc.edu/~aabello/publications/17.IST.pdf
[3]http://www.cs.put.poznan.pl/dbrzezinski/publications/DataStreamOpenChallenges.pdf
[4]https://flink.apache.org/img/blog/dynamic-tables/streams.png
[5]http://users.ics.aalto.fi/indre/surv.pdf
[6]https://www.researchgate.net/publication/270787580_A_Survey_on_Supervised_Classification_on_Data_Streams
[7]http://iieta.org/sites/default/files/Journals/RCES/04.1_06.pdf