Mining Data Streams: Classification Methods

JERZY STEFANOWSKI
Inst. Informatyki PP
Wersja zmodyfikowana w 2020
dla przedmiotu PED

Outline of the presentation

1. Motivations and data streams requirements
2. Taxonomy of approaches
3. Classification in data streams
4. Concept drift in streams
5. Drift detection methods
6. Incremental learning, VFDT
7. Ensemble classifiers
8. Partially labeled streams
9. Approximate processing and basic operation
10. Research directions

Inspiracje

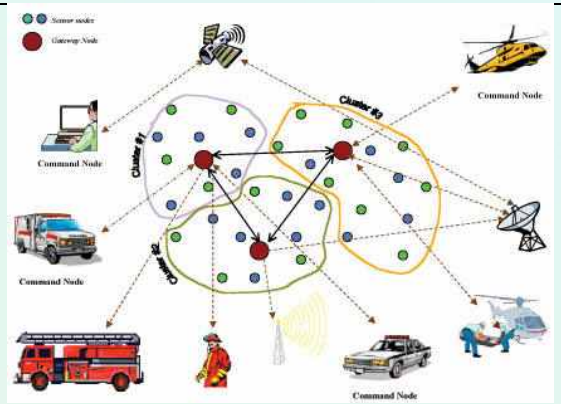
Some of slides are coming from lectures of:

- Mining High Speed Data Streams, talk by P. Domingos, G. Hulten, SIGKDD 2000.
- State of the art in data streams mining, talk by M.Gaber and J.Gama, ECML 2007.
- J.Han slides for a lecture on Mining Data Streams – available from Han's page on his book
- Myra Spiliopoulou, Frank Höppner, Mirko Böttcher - Knowledge Discovery from Evolving Data / tutorial at ECML 2008

Inne pomysły współpracownicy (D.Brzeziński, M.Deckert) + mój cykl wykładów
Ensemble Classifiers for Data Streams with Concept Drift → wykłady dla szkół doktoranckich

Powyższe → slajdy w języku angielskim

Processing new types of data



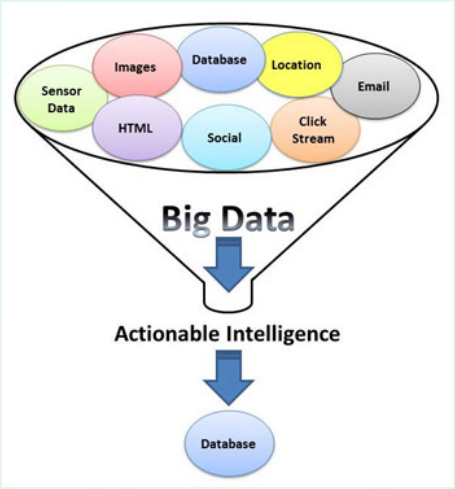
New data sources:

Internet, social networks, sensors, mobile devices, IoT, ..

„In many fields costs of data acquisition are now lower, much lower than costs of data analysis”

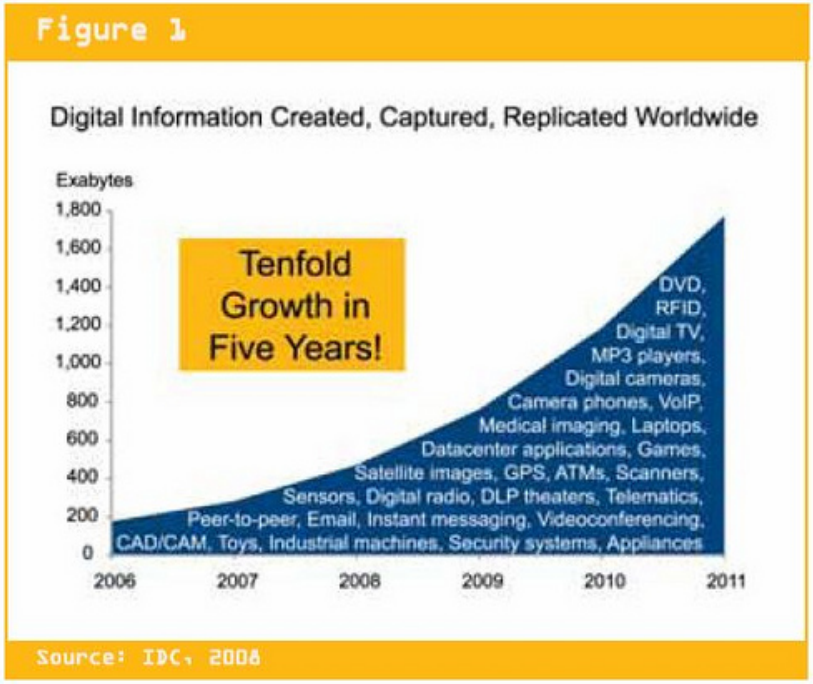
[Canadian Statistical Sci . Inst 2015]

Role of analysing Big Data (many V)



Challenges for processing new types of data

- Growing volumes of registered and stored data (exponential)
- Applications (machines) produce data (rapid rates, etc.)
- Difficulties with storing all produced data (massive volumes)
- Is this is a problem of big size only?
- Other characteristics of data
  - Static vs. dynamic
  - **Data streams**



Illustrative examples → machine generated data streams

Sensor networks

- J.Gama – an electrical grid
  – continuous measurements / telemetry
- Tasks: forecast electricity demands for sub-networks of consumer; discover profiles of users ; operational management, identifications of critical points [see. Chapter 1 J.Gama, Knowledge Discovery from Data Streams.]

Monitor quality of the production process → many measurement sensors in production machines

- M.Pechenizky, I.Zliobaite – changes over time (2-3 months)
- New supplied materials, production procedures, replacing broken components, new staff members, electricity new regulations,..

Data streams – also evolving

J. Gama – Sensor networks / Portugal

Electrical power Network: Sensors all around network monitor measurements of interest.

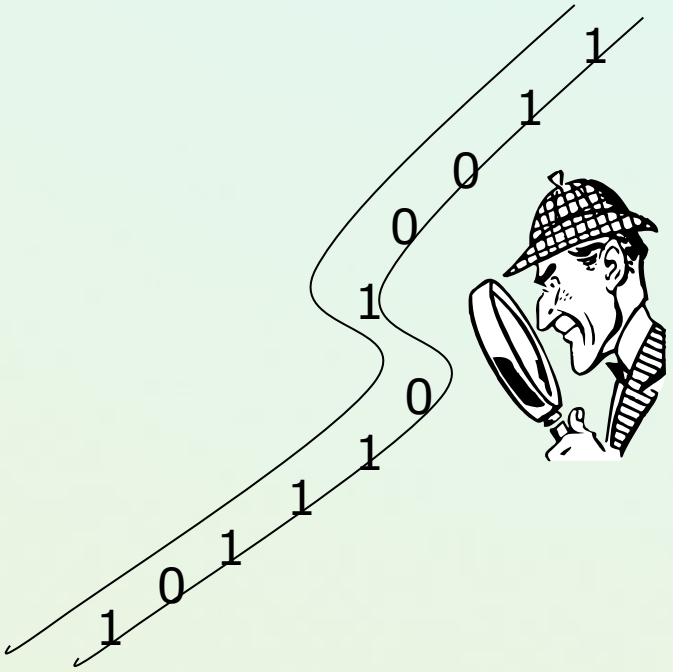
- Sensors produce continuous flow of data at high speed:
  - Send information at different time scales;
  - Act in adversary conditions: they are prone to noise, weather conditions, battery conditions, etc;
- Huge number of Sensors, variable along time
- Geographic distribution:
  - The topology of the network and the position of the sensors are known.

Sensor networks / electricity Portugal

- Cluster Analysis
  - Identification of Profiles: Urban, Rural, Industrial, etc.
- Predictive Analysis
  - Predict the value measured by each sensor for different time horizons.
  - Prediction of picks on the demand.
- Monitoring Evolution
  - Change Detection
    - Detect changes in the behaviour of sensors;
    - Detect Failures and Abnormal Activities;
  - Extreme Values, Anomaly and Outlier Detection
    - Identification of picks on the demand;
    - Identification of critical points in load evolution;

Applications

- Performance measurements in network monitoring and traffic management
- Call detail records in telecommunications
- Transactions in retail chains, ATM operations in banks
- Log records generated by Web Servers …
- Mobile devices, IoT
- Monitoring scientific objects
- Smart cities, traffic control
- Filtering information
- Recommenders systems
- Financial market analysis
- …



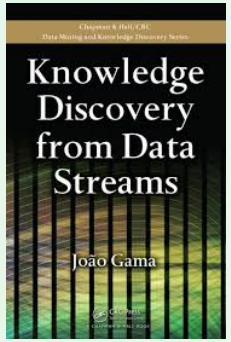
Data Streams - definition

- *"A **data stream** is a potentially unbounded, ordered sequence of data items, which arrive continuously at high-speeds"*

Springer Encyclopedia of Machine Leaning

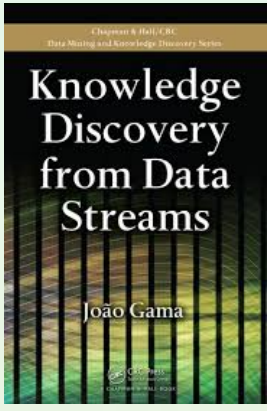
- *"It is impossible to control the order in which items arrive, nor is it feasible to locally **store** a stream in its entirety"*

- Other definitions see:



Data stream characteristic

- **Continuous flow** – the data elements arrive online one after another
  - Time intervals between element may vary
  - Each example can be processed only once (single scan)
  - The system has not control over the order of arriving elements
- **Huge volumes of data** (potentially unbounded in size)
- Data arrive at a rapid rate
  - With respect to the computational abilities of the processing system (time is costly)
- Data streams **may evolve** over time
  - Different types of concept drifts



Traditional vs. Stream Processing

Differences

	Traditional	Stream
No. of passes	Multiple	Single
Processing Time	Unlimited	Restricted
Memory Usage	Unlimited	Restricted
Type of Results	Accurate	Approximate
Distributed	Usually no	Yes

Data streams vs. time series

Both incremental and time dependent

However, there are strong differences

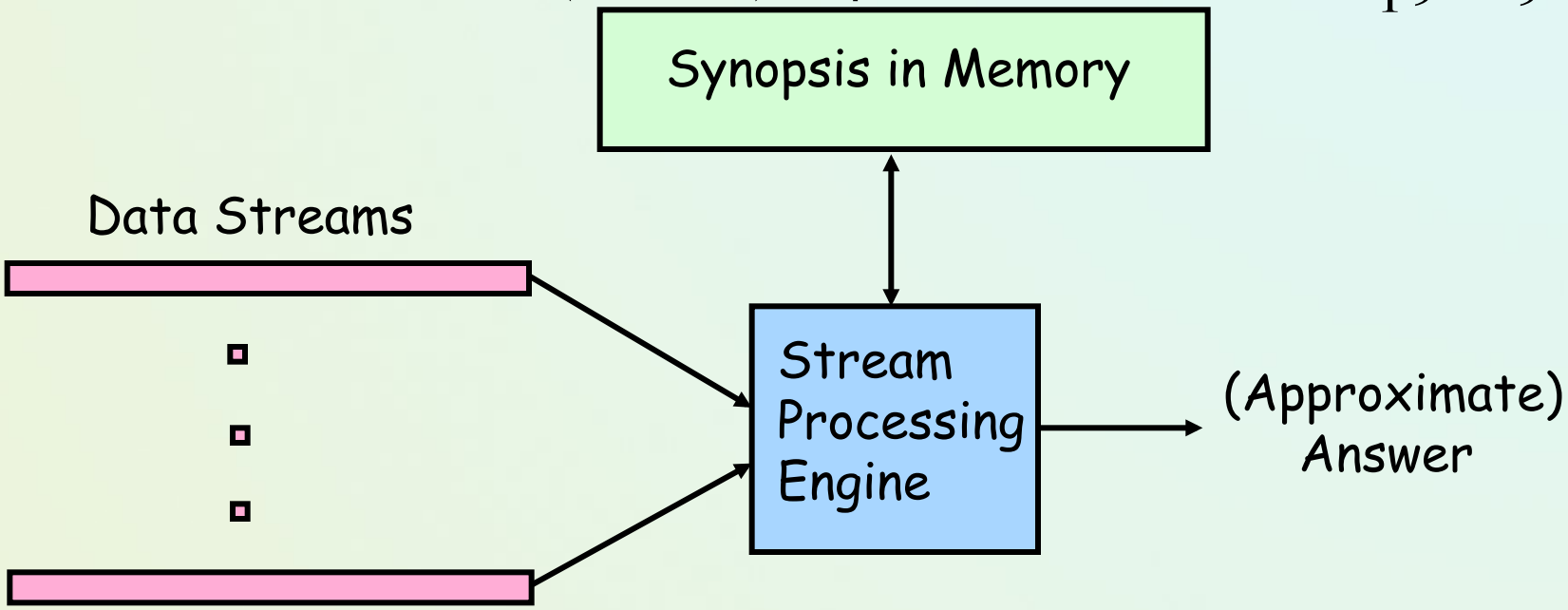
- Multi-dimensional attributes vs. focus on the main among them
- Different predictions
  - See the classification task / next lecture
- No typical auto-correlations and similar assumptions
- Other view of seasonal changes
- Non-stationary and concept-drifting characteristics
- Computational requirements
- And …

Stream Data Mining: Typical tasks

- Approximate queries and computations
- Frequent pattern mining
- Mining sequential patterns in data streams
- Classification in streams
- Stream clustering
- Mining outliers and unusual patterns in data streams
- Novelty detection
- Other: e.g. analysing text stream, sentiment and opinion modelling

Computation Model for Approximate Answers

- A data stream as a (massive) sequence of elements: $e_1, \ldots, e_n$



- Stream processing requirements
  - Single pass: Each record is examined at most once
  - Bounded storage: Limited Memory (M) for storing synopsis
  - Real-time: Per record processing time (to maintain synopsis) must be low

Data Stream – Querying Data

- Generally, algorithms compute approximate answers
  - Difficult to compute answers accurately with limited memory
- Approximate answers - Deterministic bounds
  - Algorithms only compute an approximate answer, but bounds on error
- Approximate answers - Probabilistic bounds
  - Algorithms compute an approximate answer with high probability
    - With probability at least $1-\delta$, the computed answer is within a factor ε of the actual answer

Illustrative Problems

Illustrative Problems:

- Count the number of distinct values in a stream;
- Count the number of 1's in a sliding window of a binary string;
- Count frequent items above a given support.

Approximate Answers

- Actual answer is within 5±1 with probability ≥ 0.9
- Trade off between sufficient accuracy of the answer and computational resources required to compute it
- Probabilistic tail inequalities
  - Chebyshev Inequality
  - Chernoff Bounds
  - Hoeffding Bound

Characterize the deviation between the true probability of some event and its frequency over m independent trials.

$$P(|\overline{X} - \mu| \geq \epsilon) \leq 2exp(-2m\epsilon^2/R^2),$$

where R is the range of the random variables.

Example: After seeing 100 examples of a random variable X, $x_i \in [0, 1]$, the sample mean is $\overline{x} = 0.6$;
The true mean is with confidence δ in $\overline{x} \pm \epsilon$, where

$$\epsilon = \frac{\sqrt{R^2 ln(1/\delta)}}{2n}$$

Stream Data Processing and Their Sampling

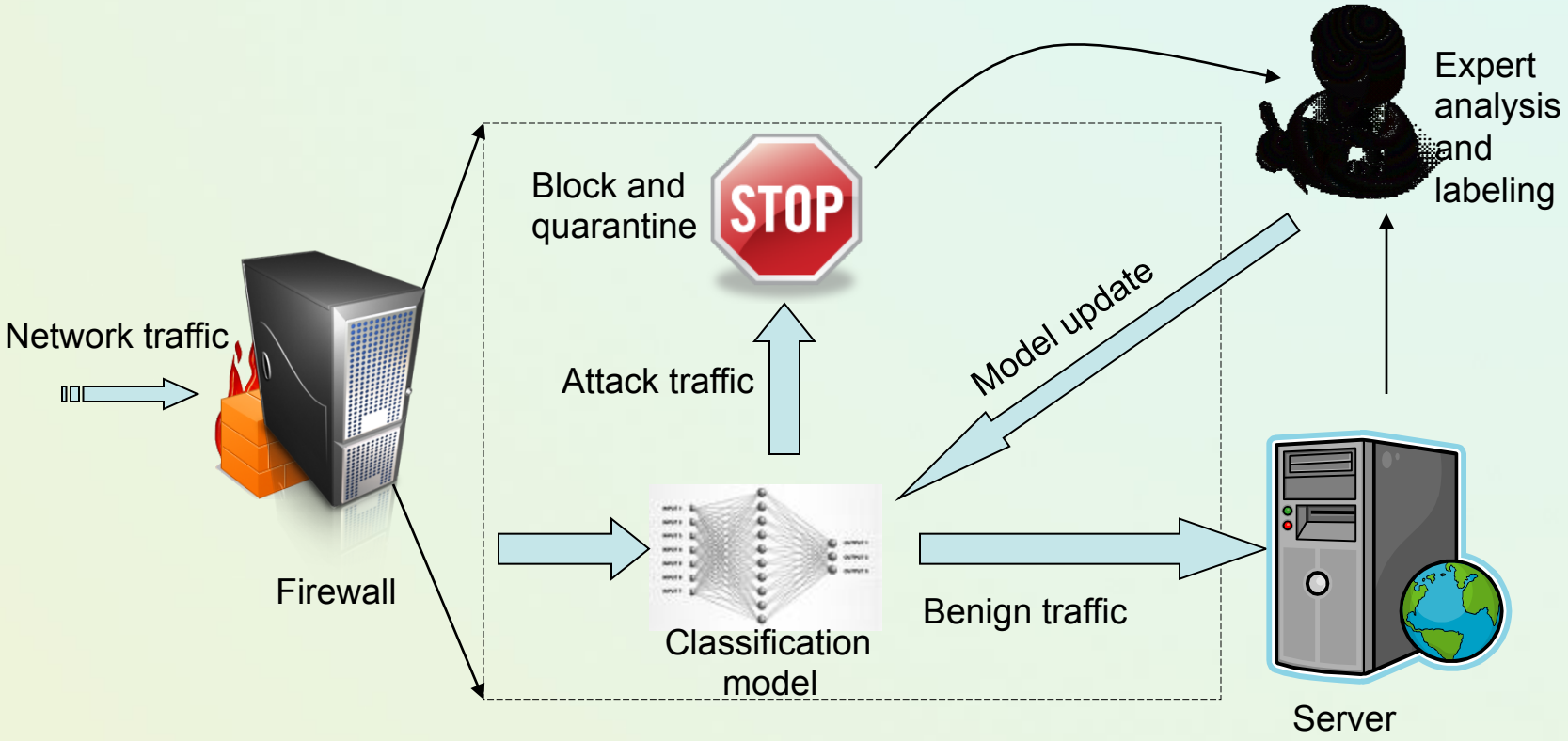
- Major challenges
  - Keep track of a large universe
- Methodology
  - Synopses (trade-off between accuracy and storage)
  - Use *synopsis data structure*, much smaller ($O(\log^k N)$ space) than their base data set (O(N) space)
  - Compute an *approximate answer* within a *small error range* (factor ε of the actual answer)
- Major methods
  - Random sampling
  - Histograms
  - Sliding windows
  - Sketches
  - Radomized algorithms

Stream Data Processing Methods (1)

- Random sampling (but without knowing the total length in advance)
  - *Reservoir sampling*: maintain a set of *s* candidates in the reservoir, which form a „true random sample" of the element seen so far in the stream. As the data stream flow, every new element has a certain probability (*s*/N) of replacing an old element in the reservoir.
- Sliding windows
  - Make decisions based only on ***recent*** *data* of sliding window size *w*
  - An element arriving at time *t* expires at time *t* + *w*
- Histograms
  - Approximate the frequency distribution of element values in a stream
  - Partition data into a set of contiguous buckets
  - Equal-width (equal value range for buckets) vs. V-optimal (minimizing frequency variance within each bucket)
- **Multi-resolution models**
  - Popular models: balanced binary trees, micro-clusters, and wavelets

Data Stream Classification

- Uses past labeled data to build classification model
- Predicts the labels of future instances using the model
- Are labels available for all future instances?



Do we have software support?

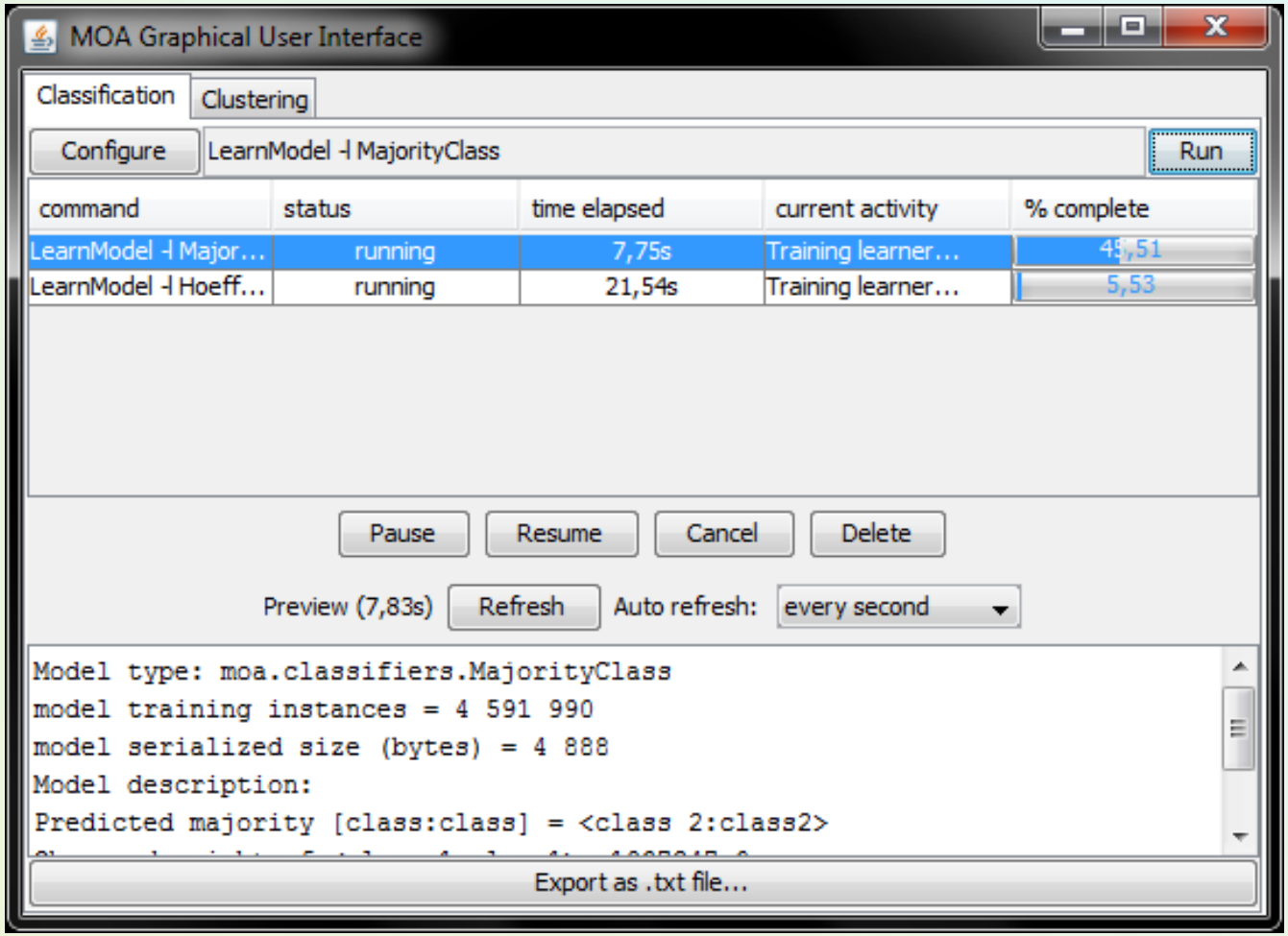
MOA framework/software

For traditional supervised learning WEKA, RapidMiner, R and other open-source DM libraries are popular

For streaming settings MOA: http://moa.cs.waikato.ac.nz/

- **Massive Online Analysis** software environment for implementing algorithms and running experiments for online learning from evolving data streams.;
- includes a collection of offline and online methods for online learning (boosting, bagging, Hoeffding Trees) with or without explicit change detection;
- tools for evaluation;
- bi-directional interaction with WEKA

MOA – an open source framework for massive data and data streams



See more at Waikato Univeristy web page

Let’s finish intro

Now we will move to the next part of the lecture devoted to classification of data streams, including concept drifts