Building stream summaries

Looking for a new approach of exploratory analysis of data streams under time and memory constraints

Valentina Zelaya Mendizábal
PhD candidate in Applied Mathematics
Paris Panthéon-Sorbonne University | Orange Labs Research

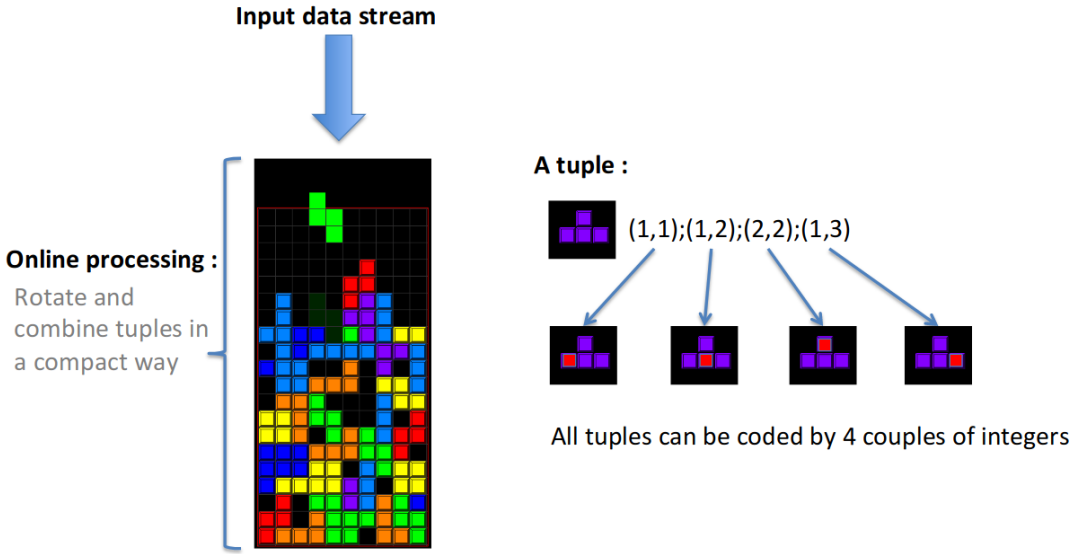
July 9th 2019

MOTIVATION

Data streams

Fraud detection, health monitoring, smart cities...

- Tuples (data units) arrive continuously and at high speed
- Data can be accessed only once
- Distributions are constantly changing



MOTIVATION

Summaries *Compact representations of past tuples that should allow to query or analyze the whole history of the data*

- Estimate the distribution with a controlled memory space and precision trade-off
- Making as few **assumptions** on the nature of the data as possible (*exploratory analysis*)
- Taking into account **time** and **memory** use constraints defined beforehand

Specific summaries

Answer a specific question

Flajolet-Martin Sketch : approximates the number of unique objects in a stream;

Count-Min Sketch : enumerates the number of elements with a paricular value, or within an interval of values.

Generic summaries

Answer a range of questions

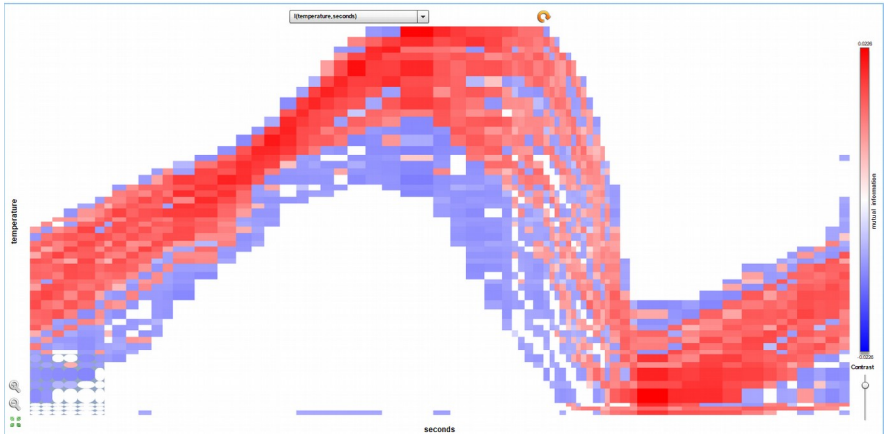
CluStream : micro-clusters over time and '*snapshots*';

StreamSamp : successive windowing and sampling.

OUR APPROACH (SO FAR)

Adaptative grid models

- Piece-wise constant estimators : Time x Data
- Variable width grid cells
- Automatic best model selection with the *Minimum Description Length* principle



Temperature evolution of 65 sensors in one day (Intel Lab Sensors dataset)

MDL

- Information theory for statistical inference

$$cost(Model) = -\log(p(Model|Data)) \propto -\log(p(Model) \times p(D|M))$$

Natural compromise between the precision and the robustness of the model

ROADMAP

- Extend the grid construction to the stream setting

  → As data arrives instead of batch mode

- Beware of change in distribution as it can affect the quality of the density estimation

  → Drift detection mechanism for streams

- Keep merging the produced summaries so we have the most informative yet compact representation

- **Quality of a general summary ?**

  → Compare performance of models that learned from the raw data and from the summaries ?

HPC IN ALL OF THIS

- The optimal MDL criterion is expensive to compute

  → Already parallelized by Orange team

- Streams are high volume and high dimensional by nature

  → process stream chunks

  → process attributes separately by distributing them over the processors available

- Optimise the memory use for grid construction and summary storage

  → Automatic ressource allocation for streams ?