Mining Data Streams

Online Prediction of Data Instance Labels

Presenters: Brandon S. Parker (PhD Student)
Ahsanul Haque (PhD Student)

Supervising Professor: Dr. Latifur Khan

Big Data Management and Analytics Lab

Agenda


Applications


Problem Statement


Challenges


Approaches

Data Streams

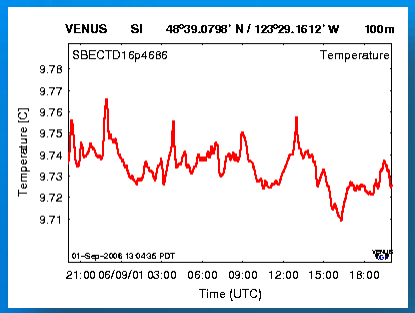
Data Streams:

- are continuous, effectively infinite, flows of data
- are increasingly common in today's connected and data driven world
- may come from disparate sources combined into a single larger stream
- evolve over time

Micro-blogs

News Feeds

Network Traffic



Sensor Data

Call center records

Use Case: Categorization of Textual Media

- Social media, blogs/micro-blogs, and aggregated news feeds.

- Addressable Problems:
  - Author attribution,
  - Sentiment categorization,
  - Syndromic surveillance
    - Computational Epidemiology (CDC)
    - Emergency Response (FEMA)
    - Natural/Weather phenomena (NOAA, USGS)

- Illustrative data sets:
  - Twitter
  - RSS feeds

Use Case: Network Monitoring

- Network protection:
  - insider threat detection
  - bandwidth allocation/ resource management
  - Worm/virus/malware propagation
  - trending analysis
- Illustrative data sets:
  - KDD Cup ’99
    - Salvatore J. Stolfo, Wei Fan, Wenke Lee, Andreas Prodromidis, and Philip K. Chan. *Cost-based Modeling and Evaluation for Data Mining With Application to Fraud and Intrusion Detection: Results from the JAM Project*.

Use Case: Sensor Data Monitoring

- Systems need to discern the global or entity states from a collection of sensor feeds in near real-time
  - Patient health monitoring
  - Environmental monitoring
  - Industrial monitoring

- Illustrative data set:
  - PAMAP2 Physical Activity Monitoring Data Set
    - A. Reiss and D. Stricker. Introducing a New Benchmarked Dataset for Activity Monitoring. The 16th IEEE International Symposium on Wearable Computers (ISWC), 2012.

Problem Statement

How do we assign accurately predicted labels to instances in a continuous, non-stationary and evolving data stream?

Generally Recognized Challenges

- Data set is effectively infinite, so:
  - the algorithm has only a single opportunity to use each data instance (i.e. one-pass),
  - must limit the memory utilization (i.e. cannot grow indefinitely),
  - cannot pre-normalize or pre-inspect the data as a whole
- The algorithm must limit the time complexity of the training *and* prediction.
- The algorithm should not *unnecessarily* reduce the feature space.
- The algorithm should be able to predict a label in near real-time.
- The algorithm should handle evolving data, including:
  - Concept Drift: changes in the feature values
  - Feature evolution: addition of new features, removal of old features, and changes in feature usage
  - Novel class appearances: completely new concept appear in the stream

Challenges:
Data Drift and Evolution

Concept Drift

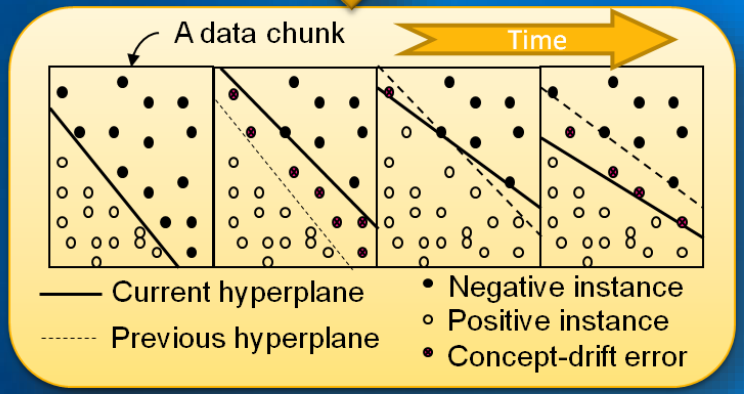
- Class boundaries change over time
- Per-concept feature subspace may change

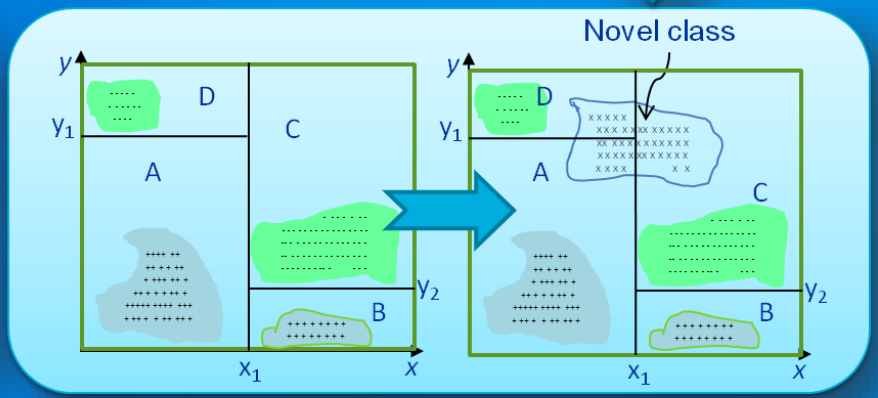
Feature Evolution

- New features appear
- Feature type/distribution changes

Concept Evolution

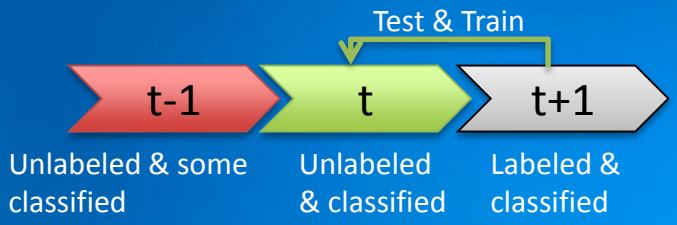
- Novel classes
- Recurrent Novel Classes

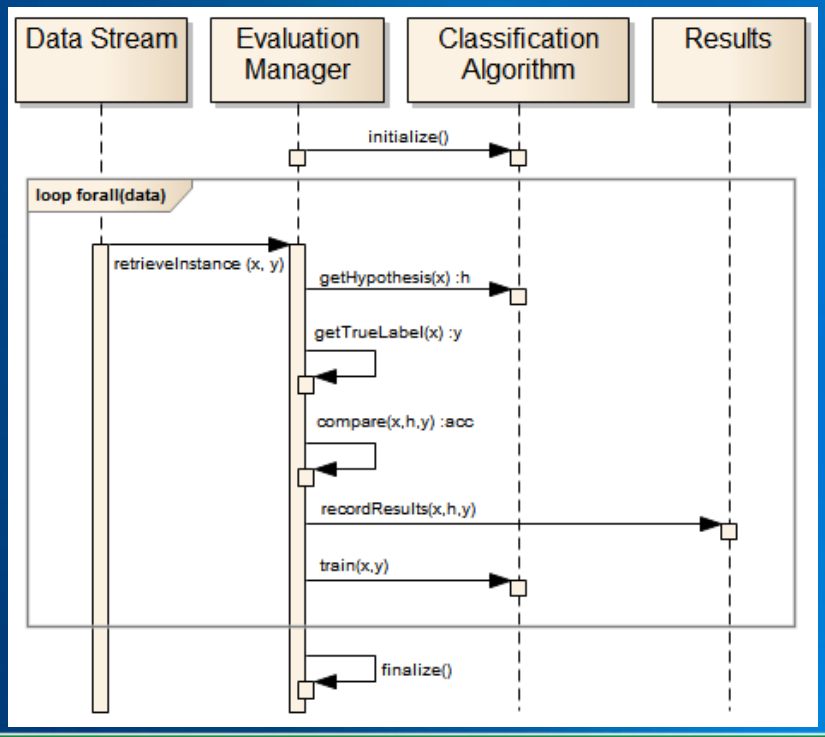


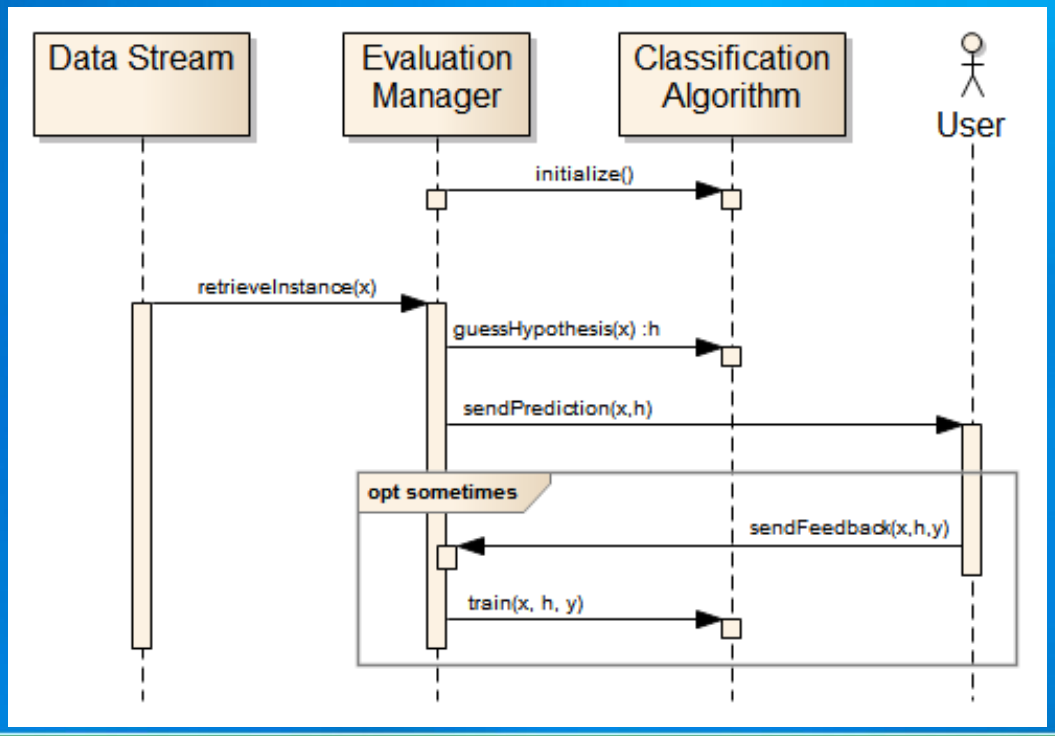


Challenges:
Required Training Data

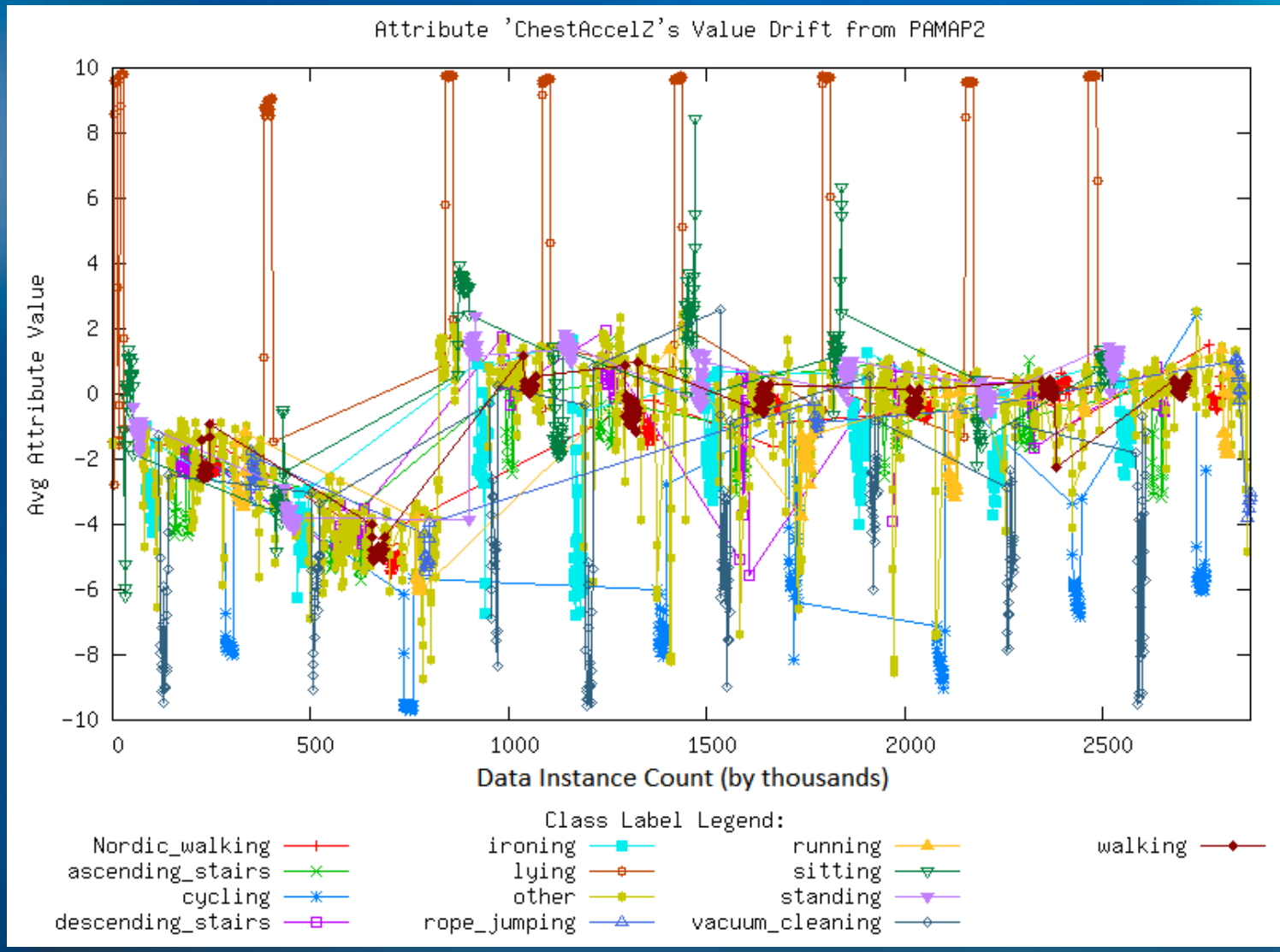
Current state-of-the-art algorithms use a fully-supervised methodology, but in real data sets, only a *fraction* of the data is actually labeled, if any.



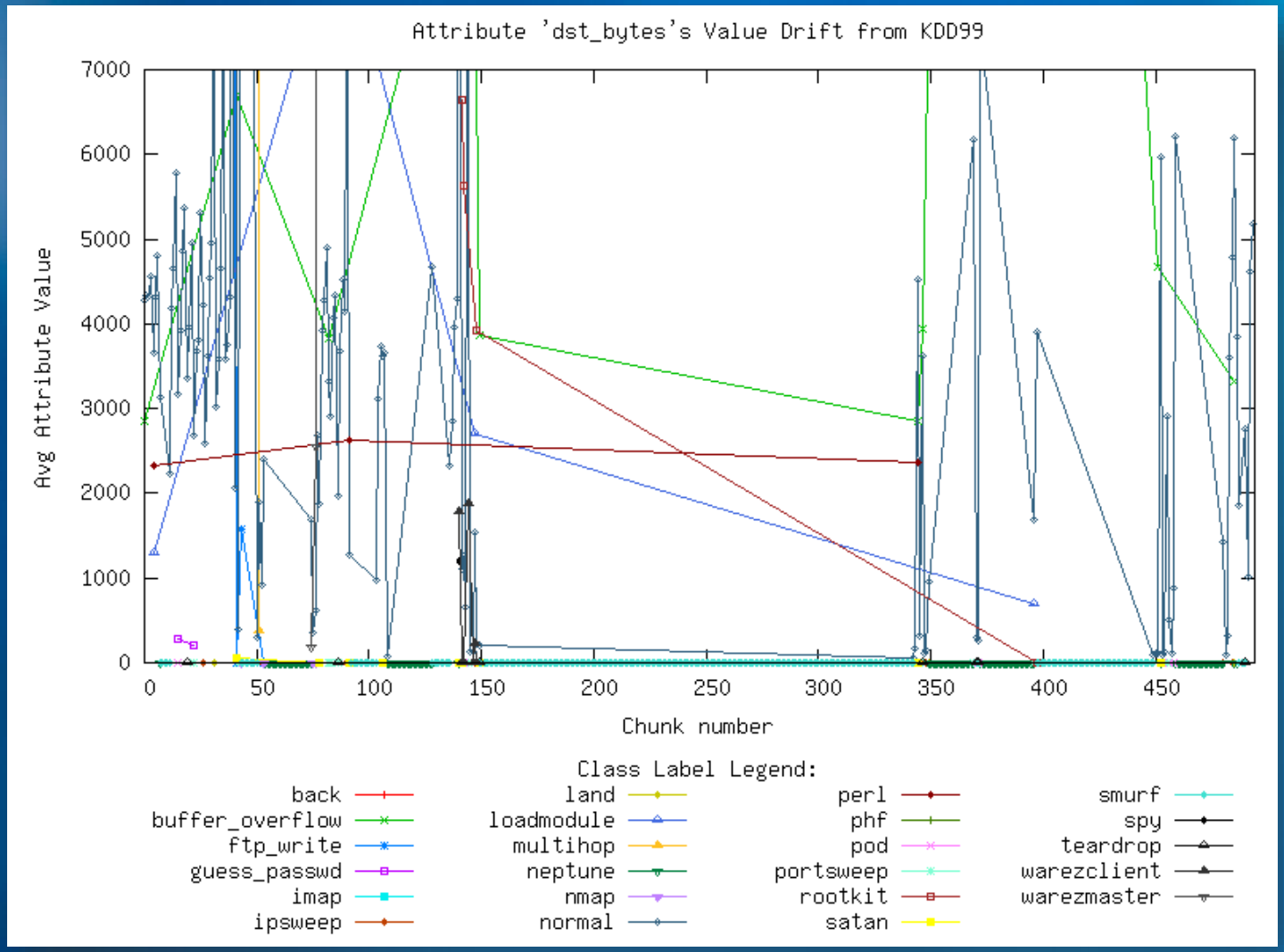




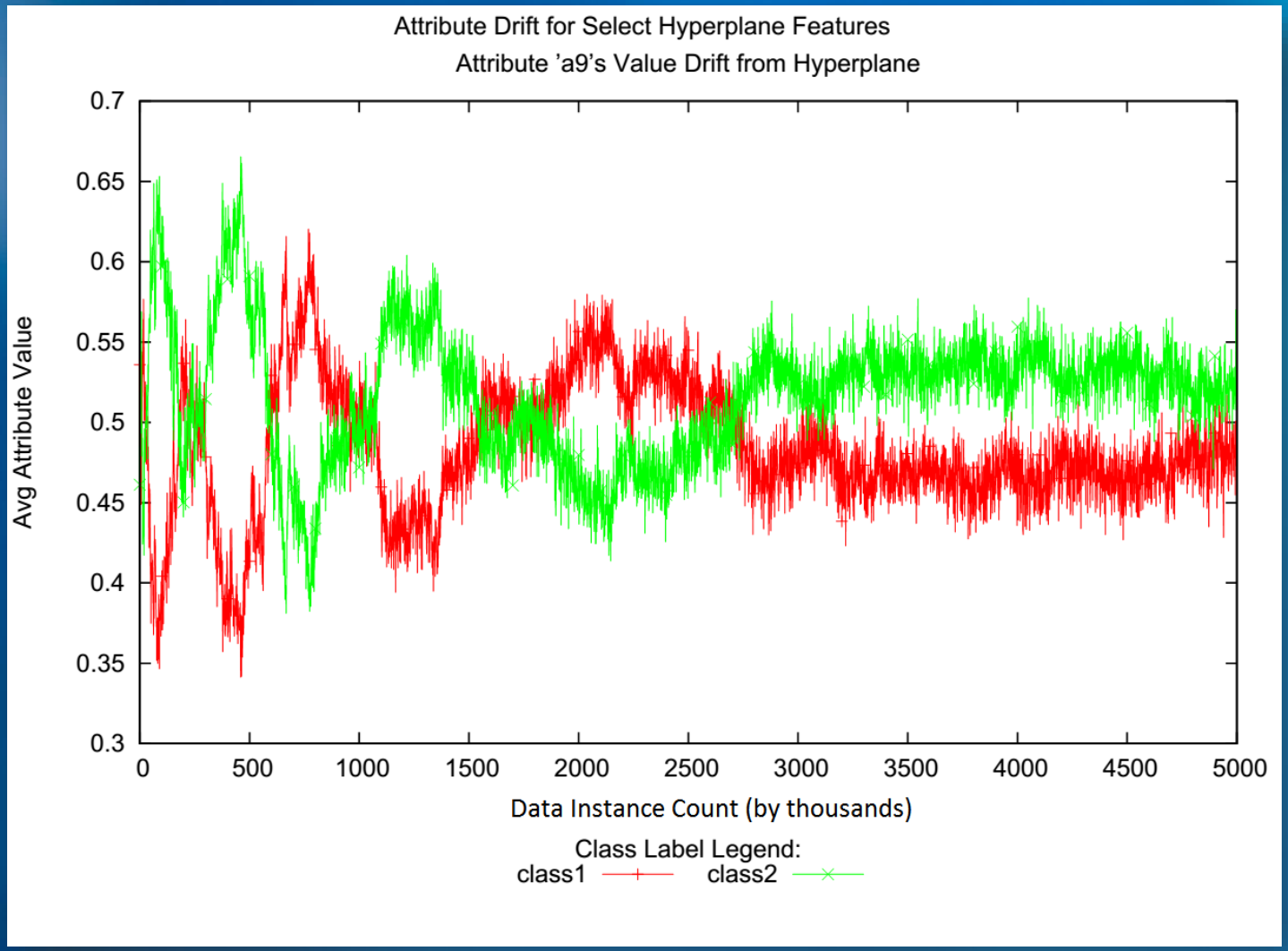
Challenges:
Lack of Test Harness



Challenges:
Lack of Test Harness



Challenges:
Lack of Test Harness



Challenges:
Conjectures of Data Streams

Conjecture #1:

A data stream requiring automated label classification will have ground truth for ***at most*** a minority of the data tuples present in the stream.

Conjecture #2:

A continuous data stream consists of more data than a static data set.

Conjecture #3:

An evolving continuous data stream consists of continuous fluctuations in observed data distributions.

Approach Comparison

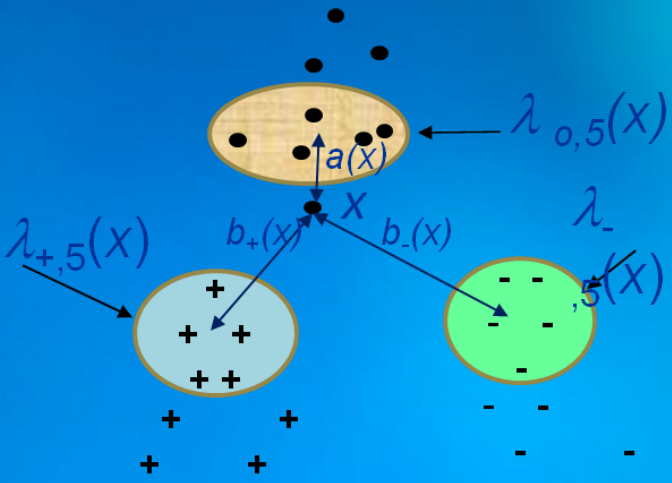
	SluiceBox	DXMiner	IncPreDeCon	DenStream[1]	FRAHST	Naive Bayes[1]	AHOT[1,2]
Labels Instances	●	●	○	○	○	●	●
Concept Subspace Tracking	●	○	●	○	◐	○	◐
Fully *Online* Process	●	●	●	○	●	●	●
One-pass	●	●	●	●	●	●	●
Multi-type Attributes	●	○	○	○	○	●	●
Not Limited to Gaussian	●	◐	●	●	◐	○	○
Concept Drift Tracking	●	●	●	●	●	●	●
Feature Evolution	●	◐	◐	○	○	◐	◐
Novel Class Detection (NCD)	●	●	◐	◐	◐	○	○
Arbitrary Shape NCD	●	◐	●	●	◐	○	○
Subspace Embedded NCD	●	○	●	○	●	○	○
Outlier Insensitivity	●	●	●	●	●	○	●

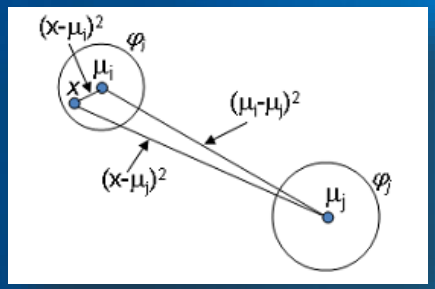
Legend: ●= full support, ◐= partial support ○= no support. [1] As implemented in MOA, [2] AdaHoeffdingOptionTree

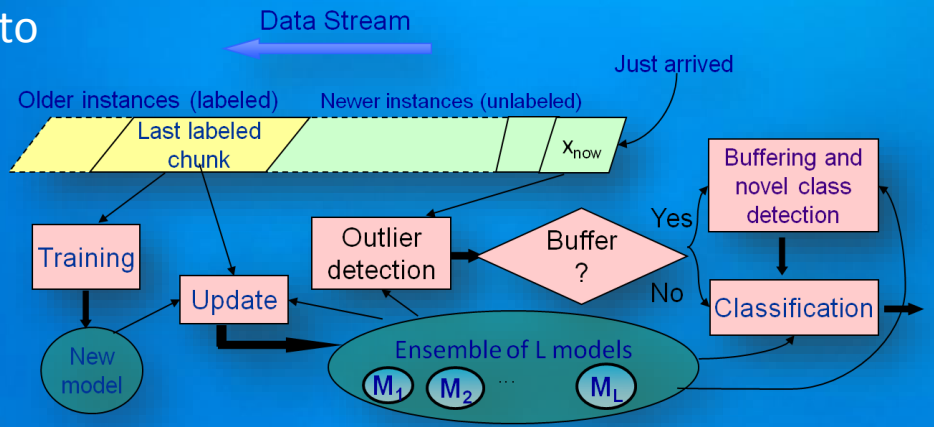
In addition, no other current approach addresses semi-supervised learning in the dynamic streaming context.

Approach: DXMiner

- Uses a chunk-based approach
- Creates hyper-sphere clusters
- Uses majority voting of per-chunk classifiers
- Uses a unified cohesion/ separation metric to discover novel classes among outliers



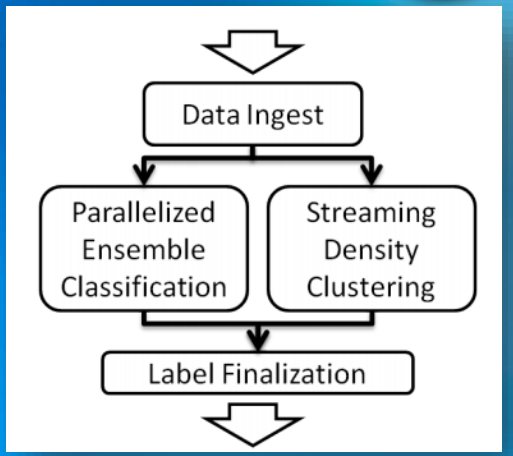


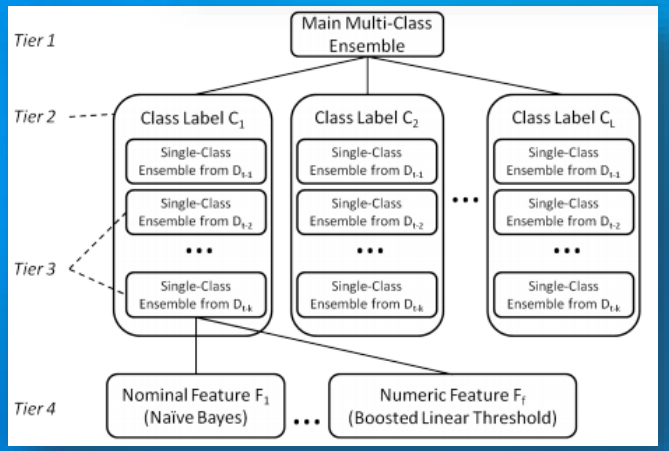


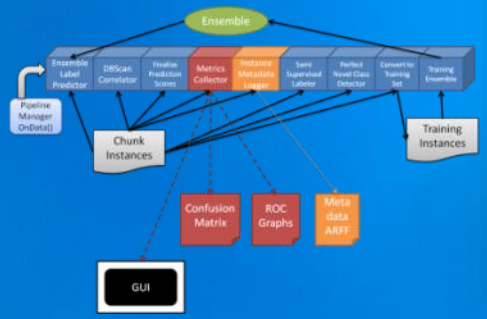
Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani M. Thuraisingham: *Classification and Novel Class Detection in Concept-Drifting Data Streams under Time Constraints*. IEEE Trans. Knowl. Data Eng. 23(6): 859-874 (2011)

Approach: SluiceBox V1.0

- Benefits:
  - Detects Novel Classes,
  - Tracks concept drift,
  - Handles feature evolution
  - Uses targeted distance and classifier algorithms per data type
  - Uses Density-based clustering for Novel Class Detection and data correlation
  - Enables semi-supervised learning
  - Both Ensemble and Clustering easily parallelized
    - ◊ QtConcurrent MapReduce on multi-Core systems
    - ◊ Multi-node MapReduce via Hadoop
    - ◊ GPU massive vector parallelism
- Weaknesses:
  - Potentially slower without parallelism





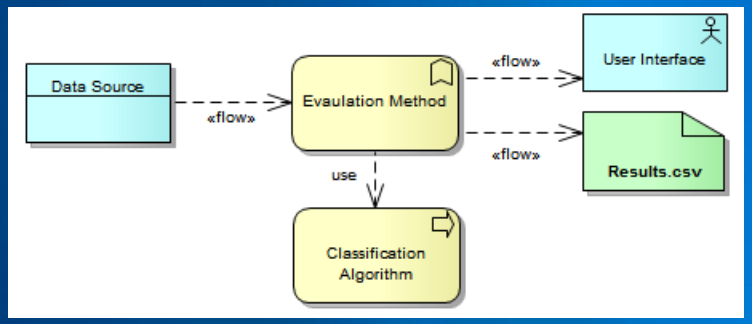


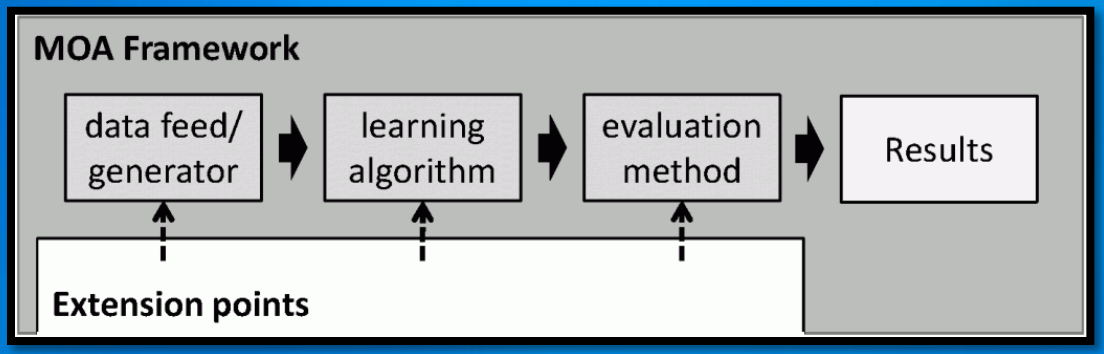
[1] B. Parker, A. Mustafa, and L. Khan, "Novel class detection and feature via a tiered ensemble approach for stream mining," in Proceedings of the 2012 IEEE 24th International Conference on Tools with Artificial Intelligence, ser. ICTAI '12. IEEE Computer Society, 2012, pp. 1171– 1178

[2] A. Haque, B. Parker, and L. Khan, "Labeling instances in evolving data streams with MapReduce." 2013 IEEE International Congress on Big Data. Santa Clara, CA: IEEE, 2013.

Approach: MOA

- Benefits:
  - Available algorithms for stream classification, including handling of concept drift
  - Available algorithms for stream generation
  - Available algorithms for stream clustering
  - Available methods for result testing
- Weaknesses:
  - Not horizontally scalable alone (see SOMOA)
  - No current methods for novel class detection nor feature evolution
  - Currently only provides fully supervised methods



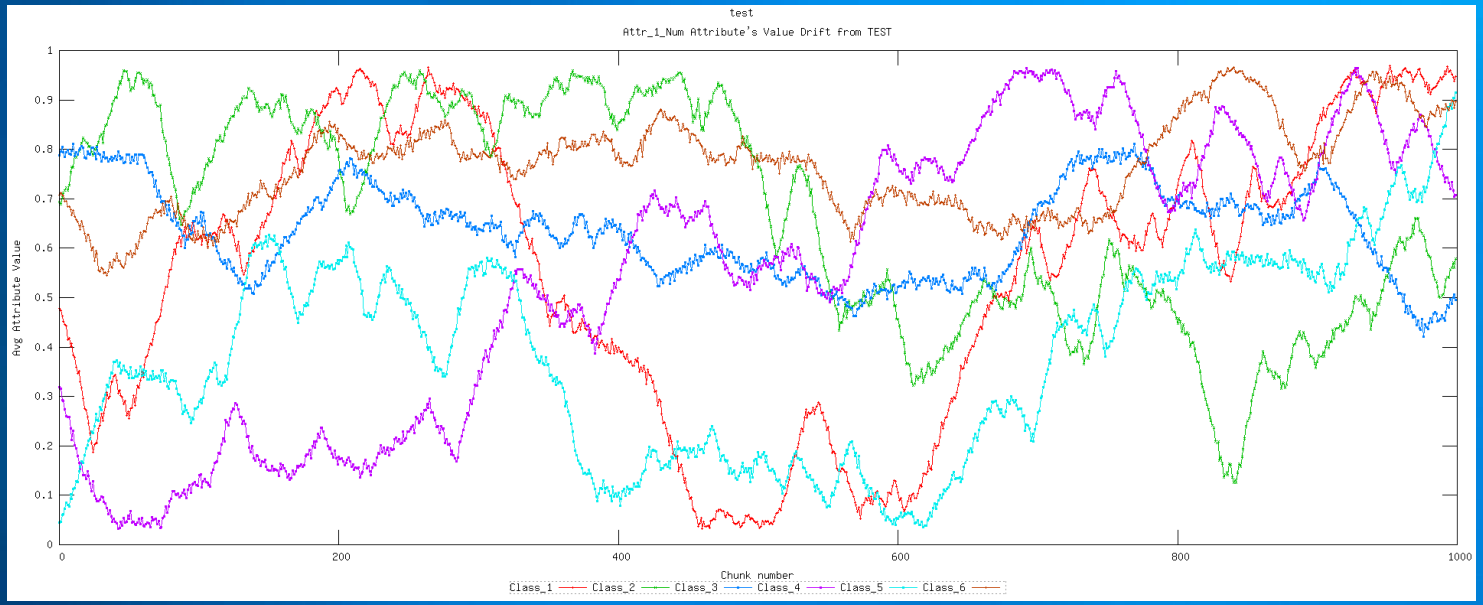


P. Kranen, H. Kremer, T. Jansen, T. Seidl, A. Bifet, G. Holmes, and B. Pfahringer, "Clustering performance on evolving data streams: Assessing algorithms and evaluation measures within moa," in Data Mining Workshops (ICDMW), 2010 IEEE International Conference on, 2010, pp. 1400–1403

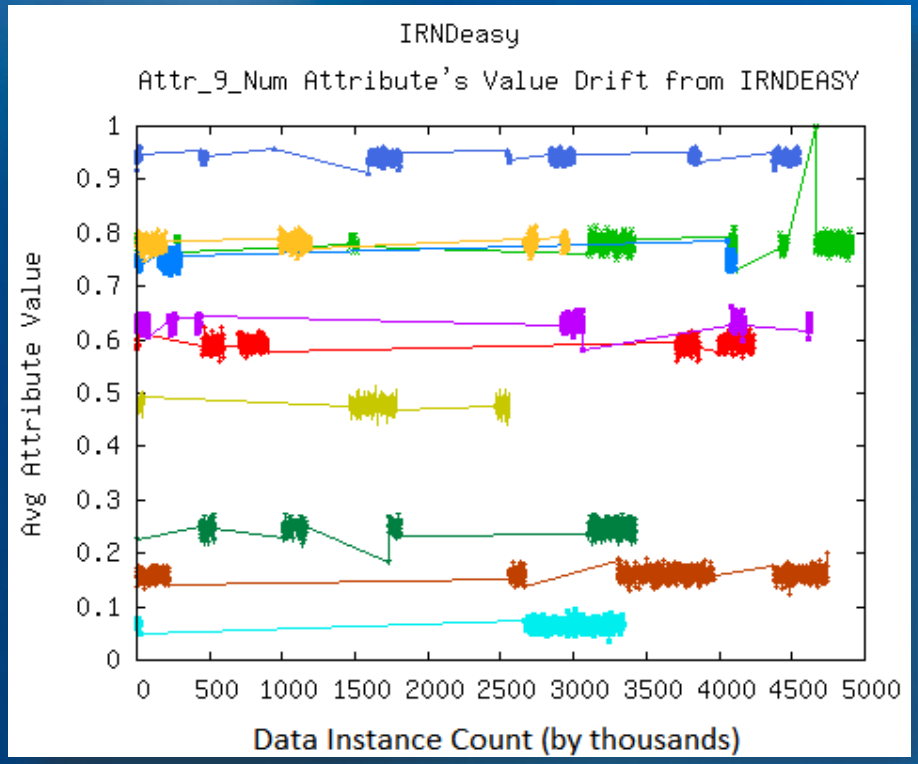
Approach: IRND Harness

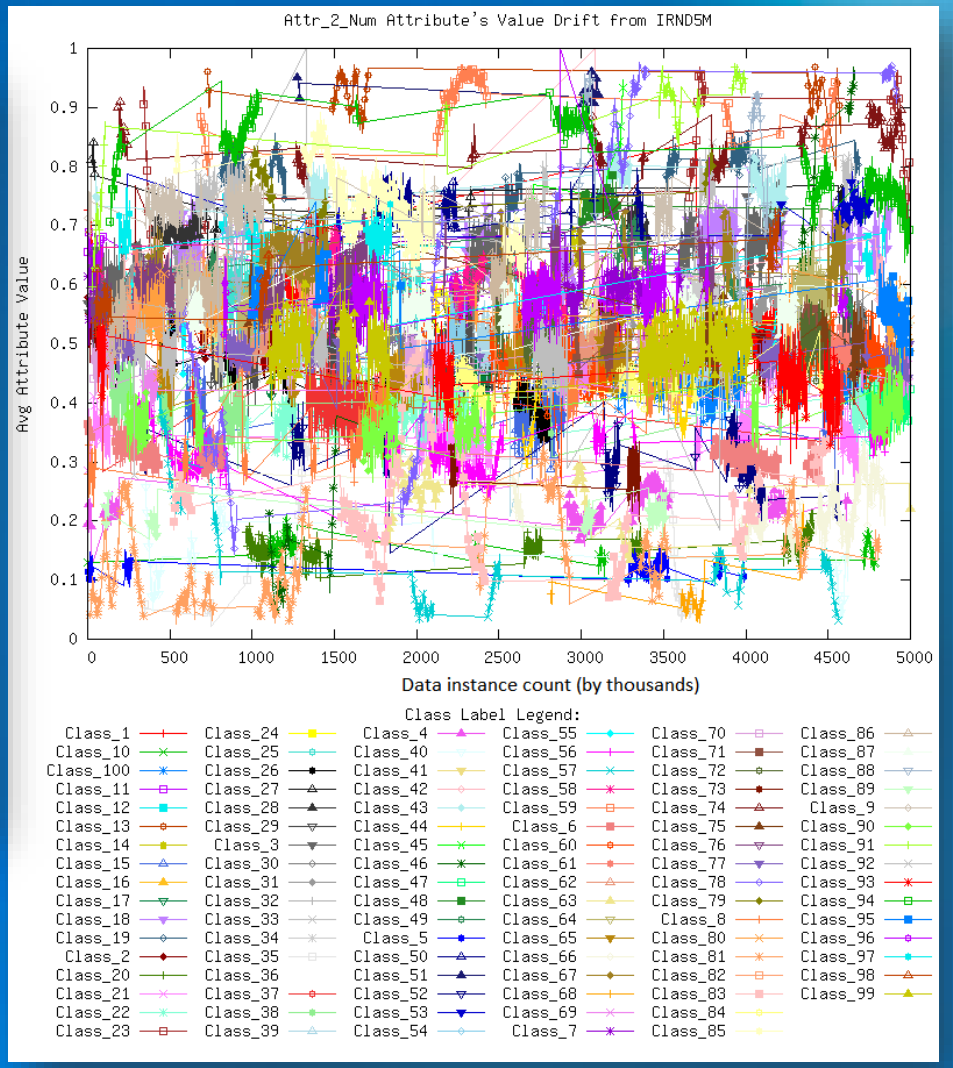
Induced Random Non-Stationary Data (IRND) Generator

- Large number of distinct concept definitions
- large number of numeric and/or nominal features
- multiple centroids per concept
- non-Gaussian feature value distributions
- Induced noise for feature value (variance) and label (labeling error)
- Concept evolution via limiting number of active rotating concepts
- Feature evolution via limiting number of active rotating attributes *per concept*
- Concept drift via tunable attribute value velocity thresholds and velocity shift probabilities



Approach: IRND Harness





Approach: SluiceBox V2.0

M^3 Algorithm (Modal Mixture Model)

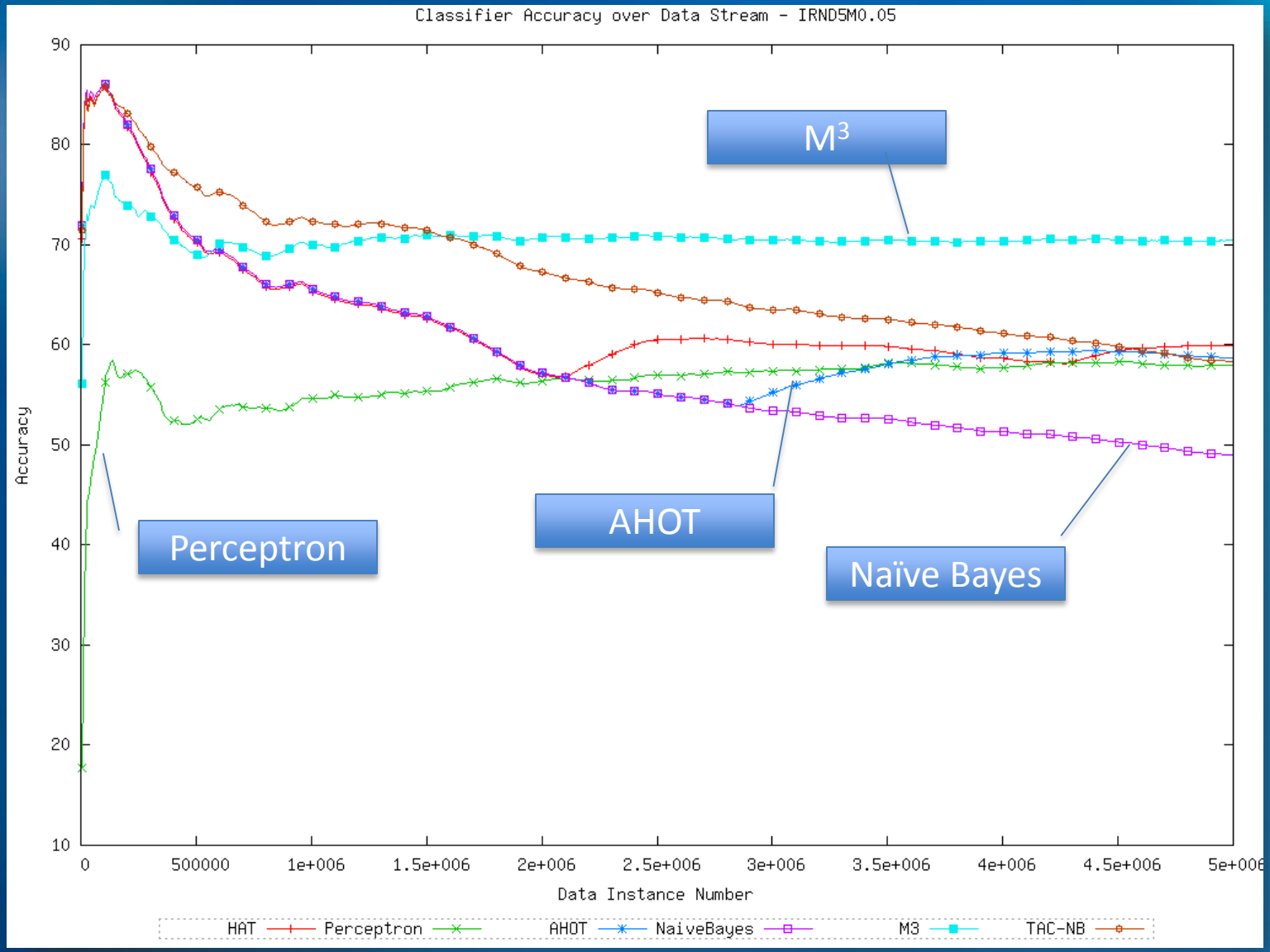
- Ensemble Method,
- Weighting based on Reinforcement Learning,
- Uses online base learners/classifiers
- Developed within the MOA framework
- Contributions to MOA Framework:
  - Reinforcement Learning Ensemble
  - IRND test harness
  - Novel Class Detection tasks
  - Additional test-case classifiers

Algorithm 3 M^3.Train

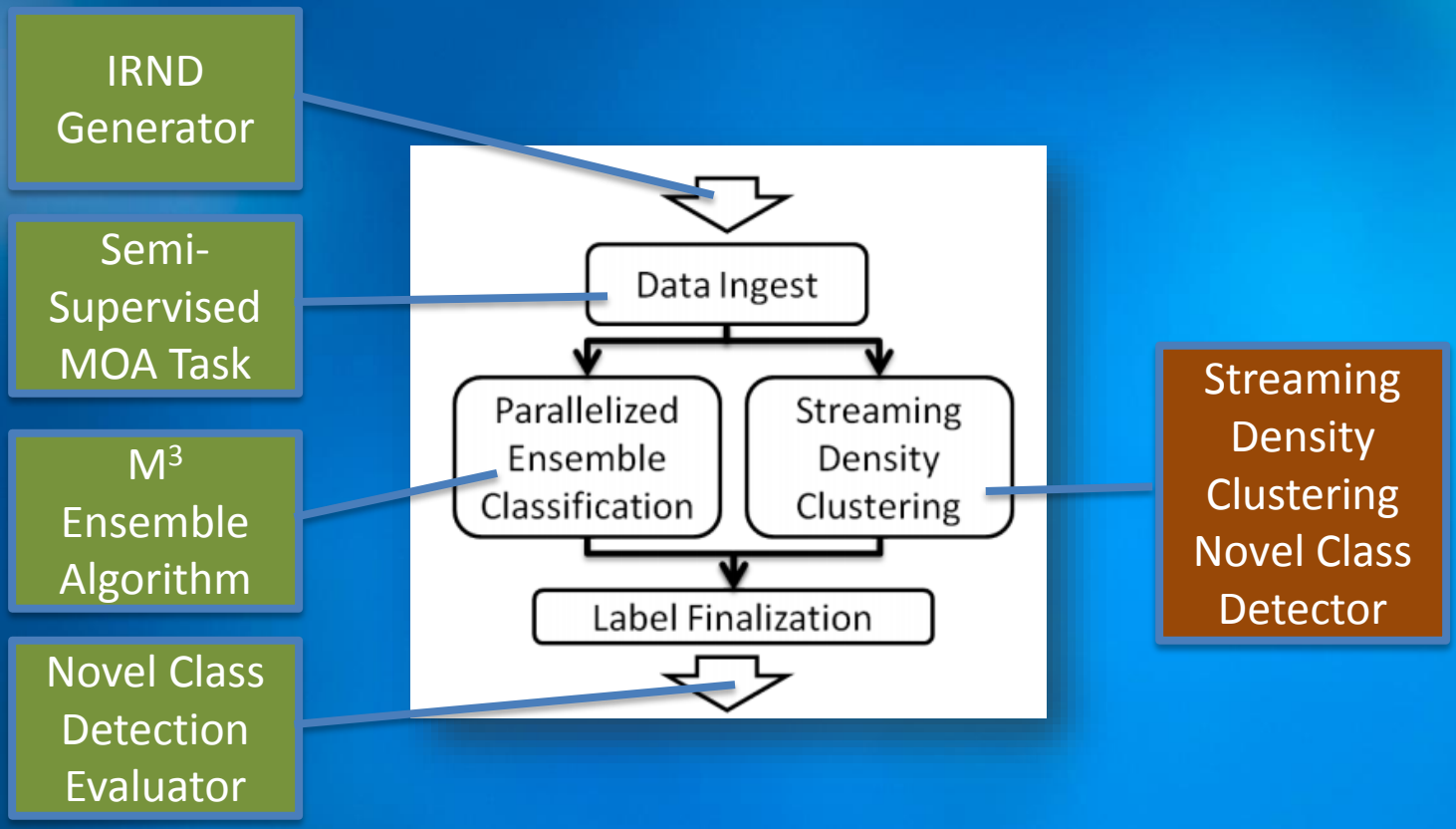
Input: data x_i, ensemble Ξ, learningFactor α, weights ϖ
for all $c \in \Xi$ **do**
  h := c.predict(x_i)
  $\varpi_c = (1-\alpha)(x.y_i = h) + \alpha \times \varpi_c$
  if $\varpi_c < threshold$ **then**
    c.reset()
    $\varpi_c := 1.0$
  end if
  c.train(x_i)
end for

	HAT		AHOT		TAC-NB		M3		DXM	
Data Set	% Acc	Time	% Acc	Time	% Acc	Time	% Acc	Time	% Acc	Time
PAMAP2	92.83	00:02:02	90.08	00:10:22	86.64	00:03:28	83.91	00:14:15	51.70	01:13:52
KDD99	95.86	00:00:19	97.11	00:00:41	96.74	00:00:19	94.04	00:01:12	88.10	00:12:14
Hyperplane	86.79	00:02:16	79.92	00:26:30	68.50	00:00:34	91.40	00:01:05	99.99	00:14:04
SEA	89.50	00:03:34	89.53	00:02:46	88.22	00:00:14	88.26	00:00:34	89.21	03:15:07
Netlogo	100.00	00:00:52	100.00	00:00:42	100.00	00:00:46	100.00	00:00:54	31.70	05:00:00
IRNDeasy	99.31	00:06:10	99.34	00:14:06	99.65	00:07:36	99.17	00:26:04	66.04	00:53:39
IRND5M	87.34	00:06:04	76.50	00:32:26	69.58	00:07:12	77.99	00:20:48	63.27	00:49:10
IRND5M 50%	83.67	00:04:21	74.48	00:28:31	69.39	00:07:27	77.02	00:14:24	-	-
IRND5M 25%	78.69	00:04:04	72.64	00:23:14	69.12	00:07:21	75.21	00:11:20	-	-
IRND5M 5%	62.63	00:05:25	61.51	00:08:54	67.08	00:07:09	70.65	00:09:06	-	-
Average	87.66		84.11		81.50		85.77		70.00	

Approach: SluiceBox V2.0



Approach: SluiceBox V2.0

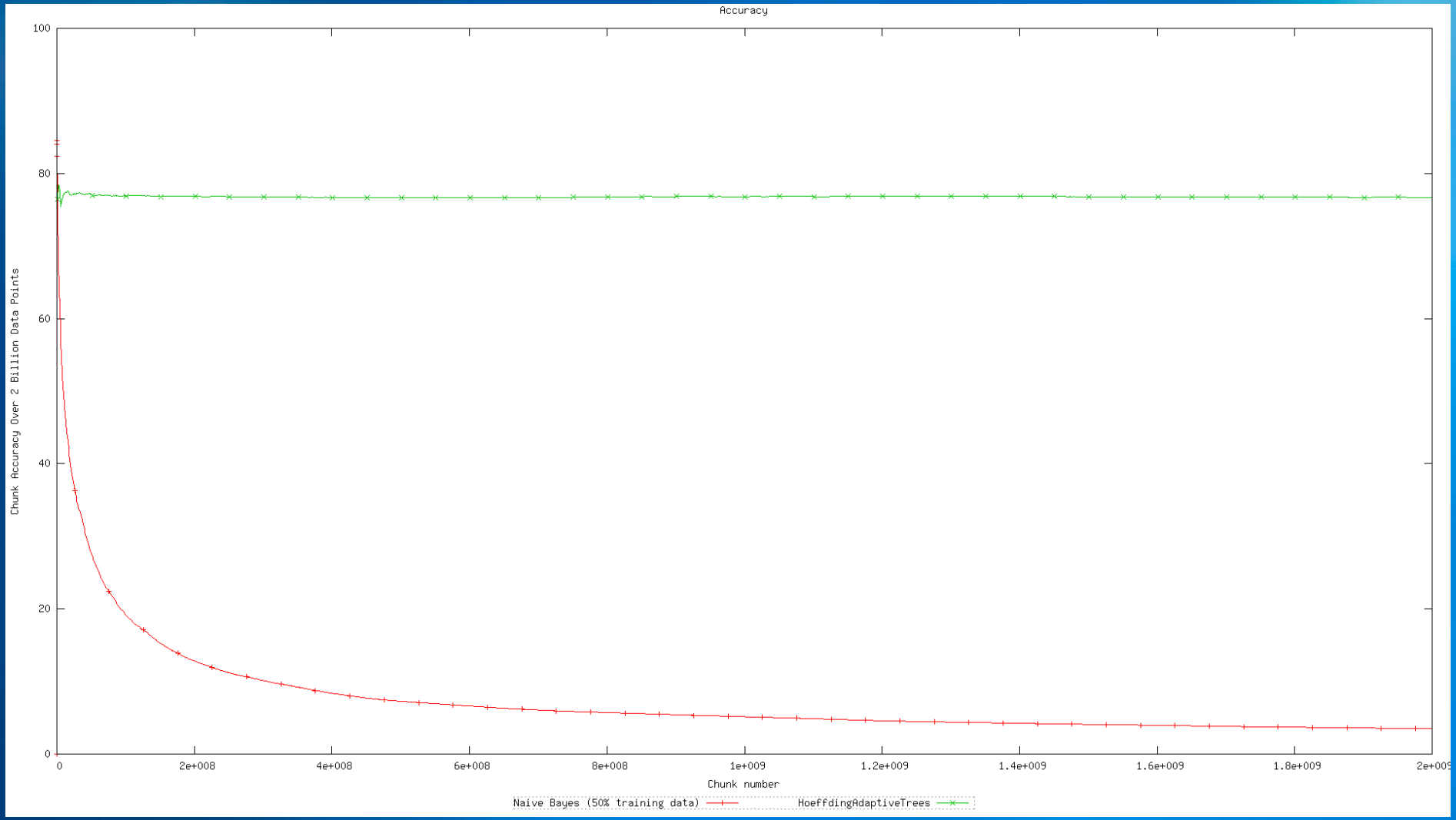


Developed in accordance with the moa framework

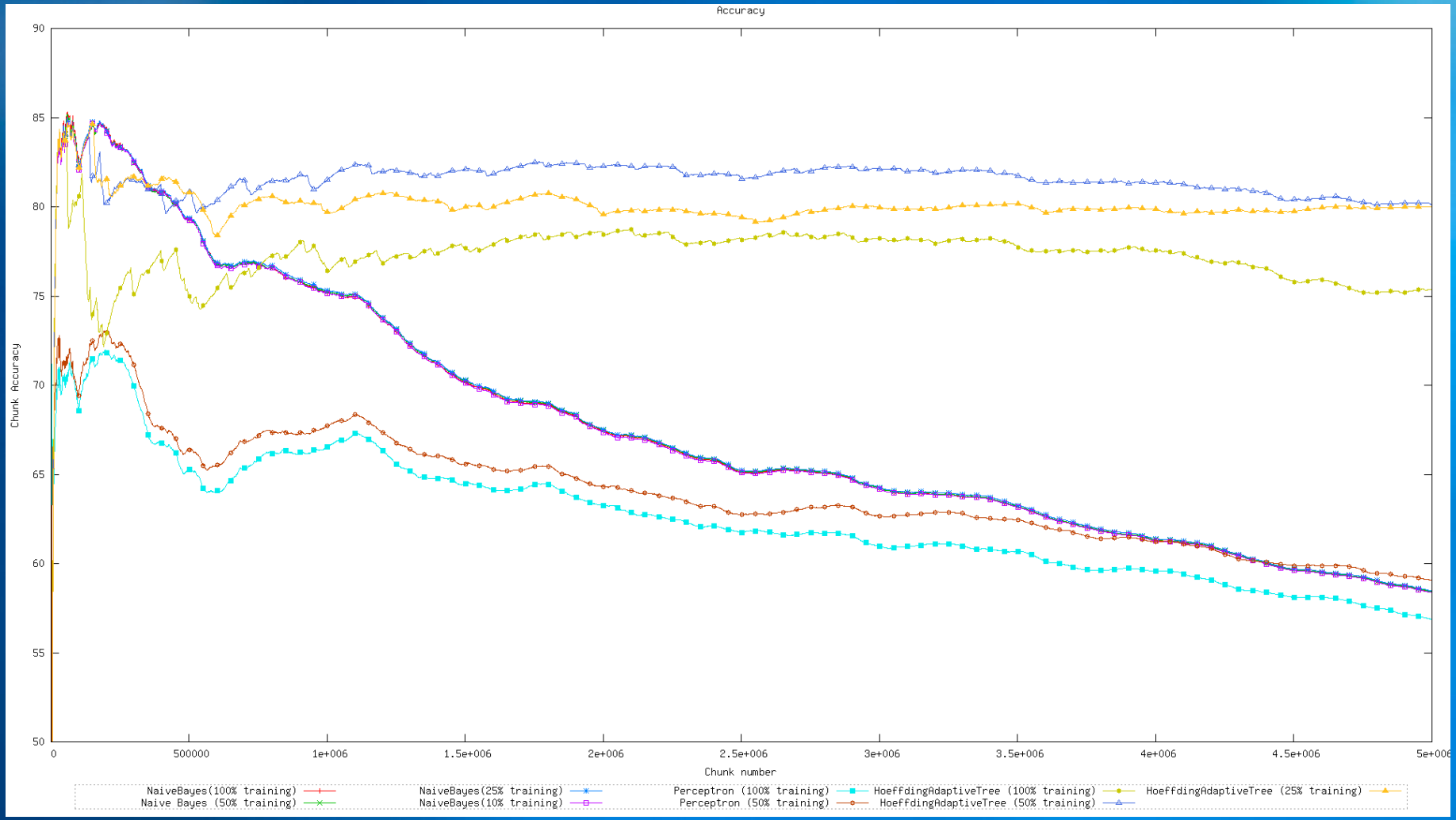
Questions?

Backup Slides

Accuracy Curve for 2 billion records



Accuracy Curve for Reduced Training



SluiceBox V1.7 Workflow