Big Stream Data Analytics: Current & Future Trends

Latifur Khan
Professor, Department of Computer Science
The University of Texas at Dallas
www.utdallas.edu/~lkhan

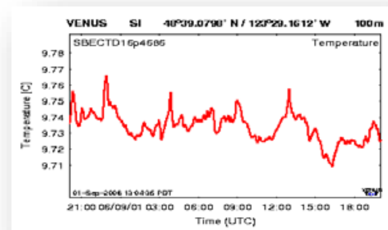
This material is based upon work supported by

Agenda

- Data Streams
- Challenges
- Shortcomings of Current Solutions
- Dynamic Chunk Management
- Limited Labeled Learning
- Experiments
- Applications
- Future Direction

Data Streams

- Data Stream:
  - is continuous flow of data.
  - very common in today's connected digital world.



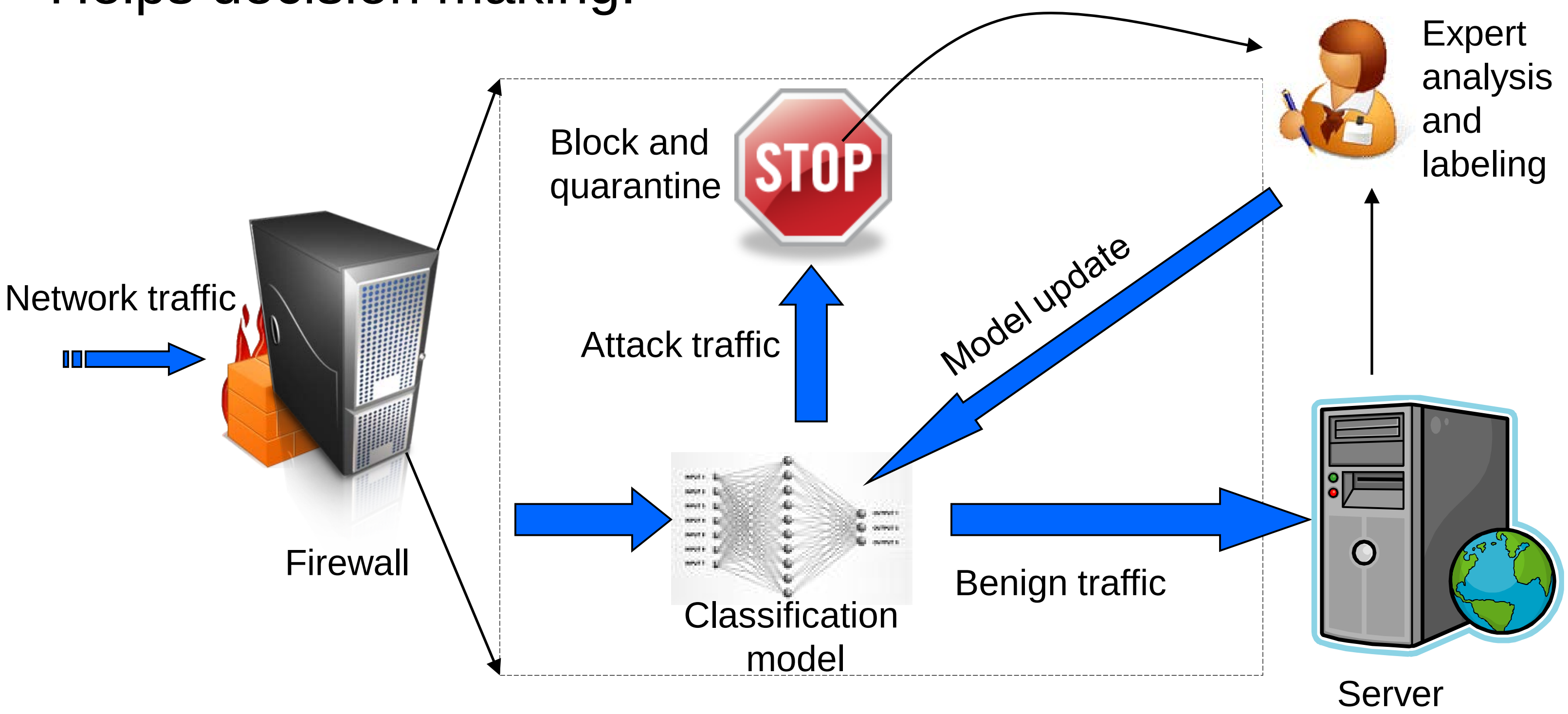
Sensor Data

Network Traffic

  - important source of knowledge that enables to take extremely important decisions in (near) real time.
- Hence, data stream mining is very important.

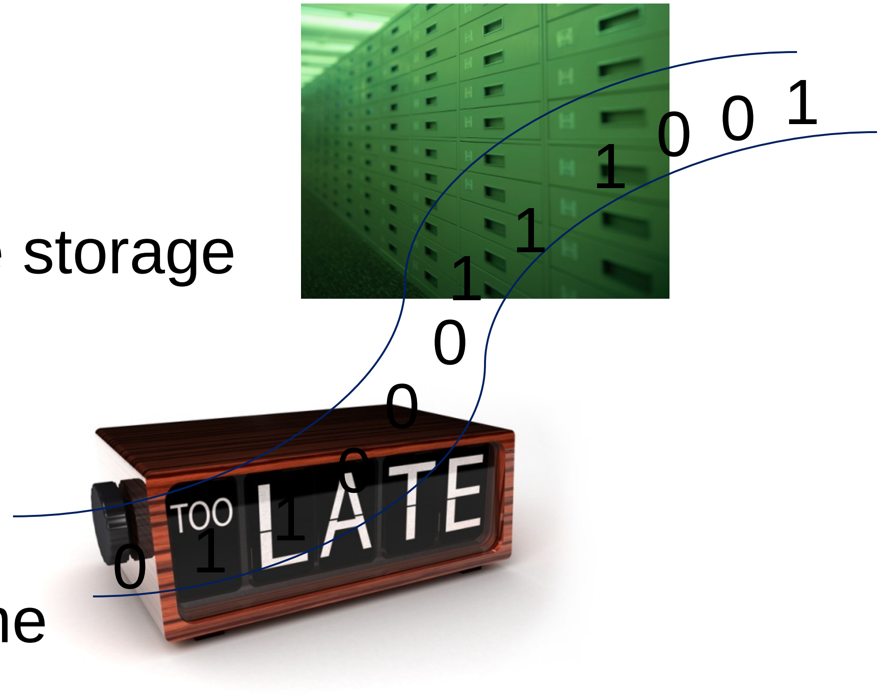
Data Stream Classification

- Uses past data to build classification model.
- Predicts the labels of future instances using the model.
- Helps decision making.

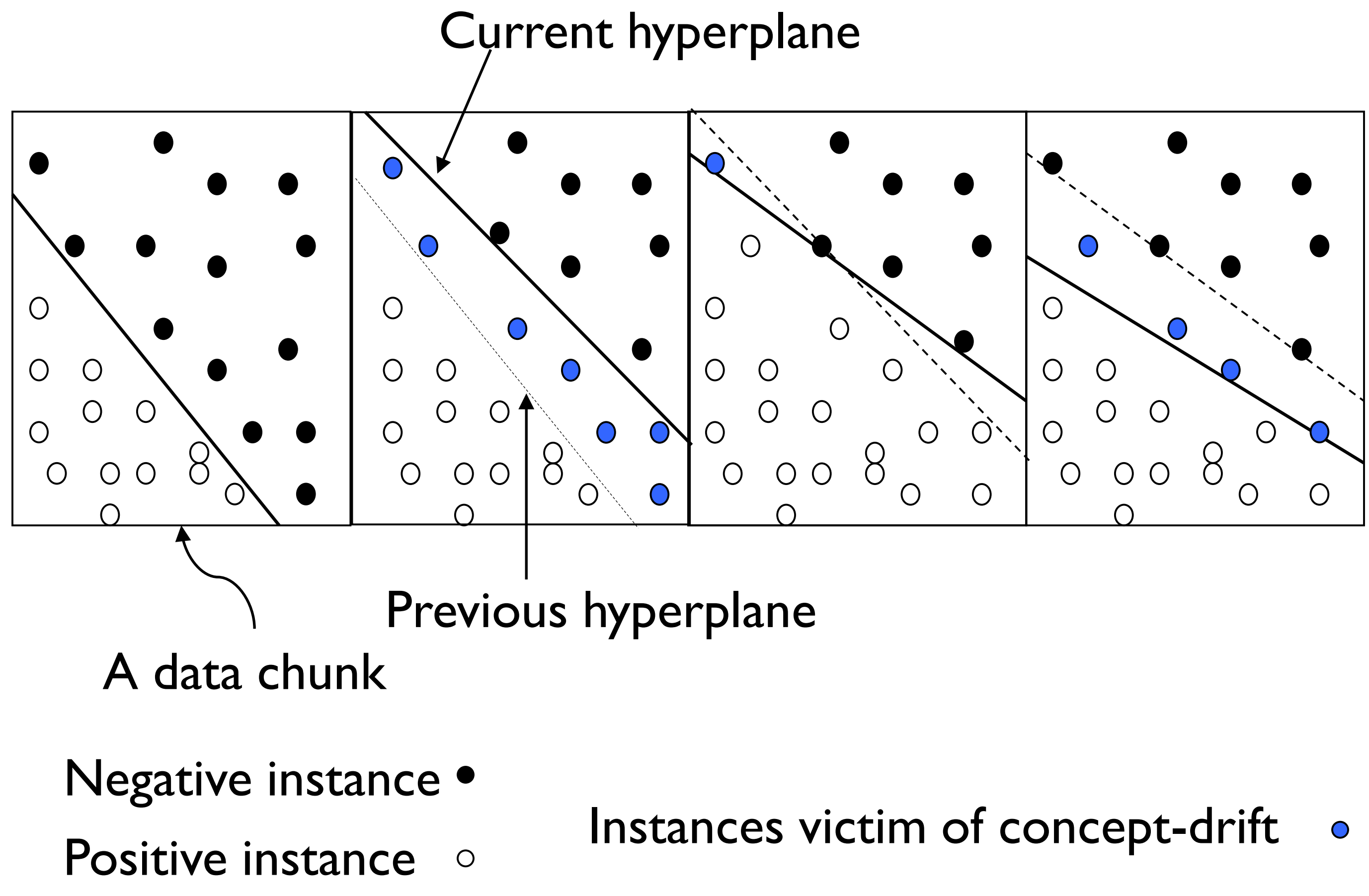


Challenge: Infinite Length

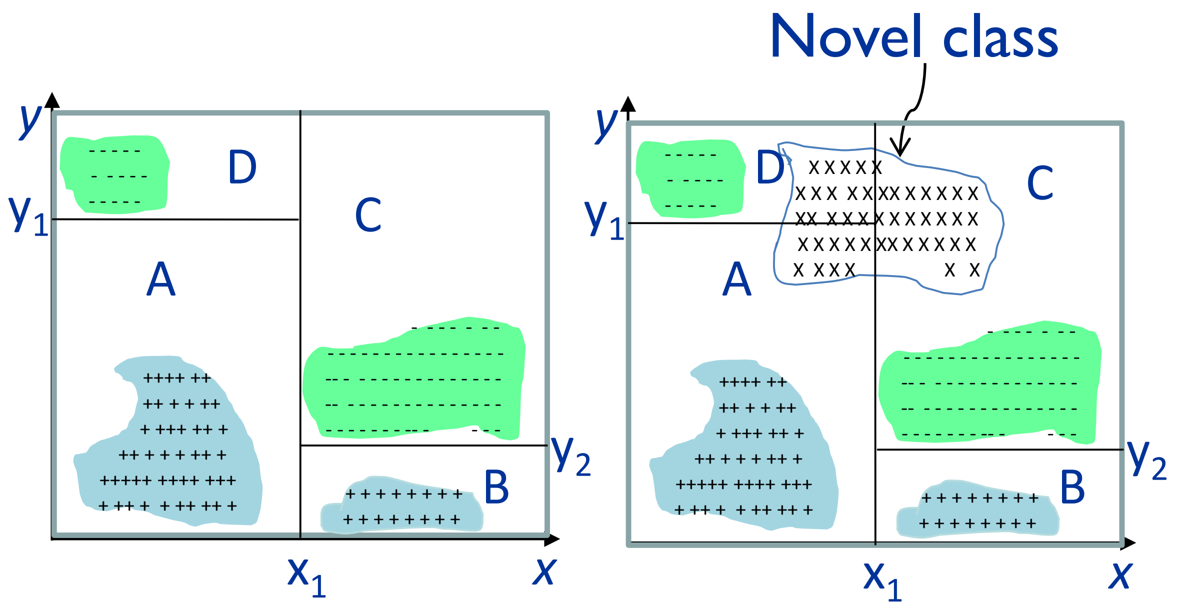
- Impractical to store and use all historical data
  - – requires infinite storage
  - – and running time



Challenge: Concept Drift



Challenge: Concept Evolution



Classification rules:

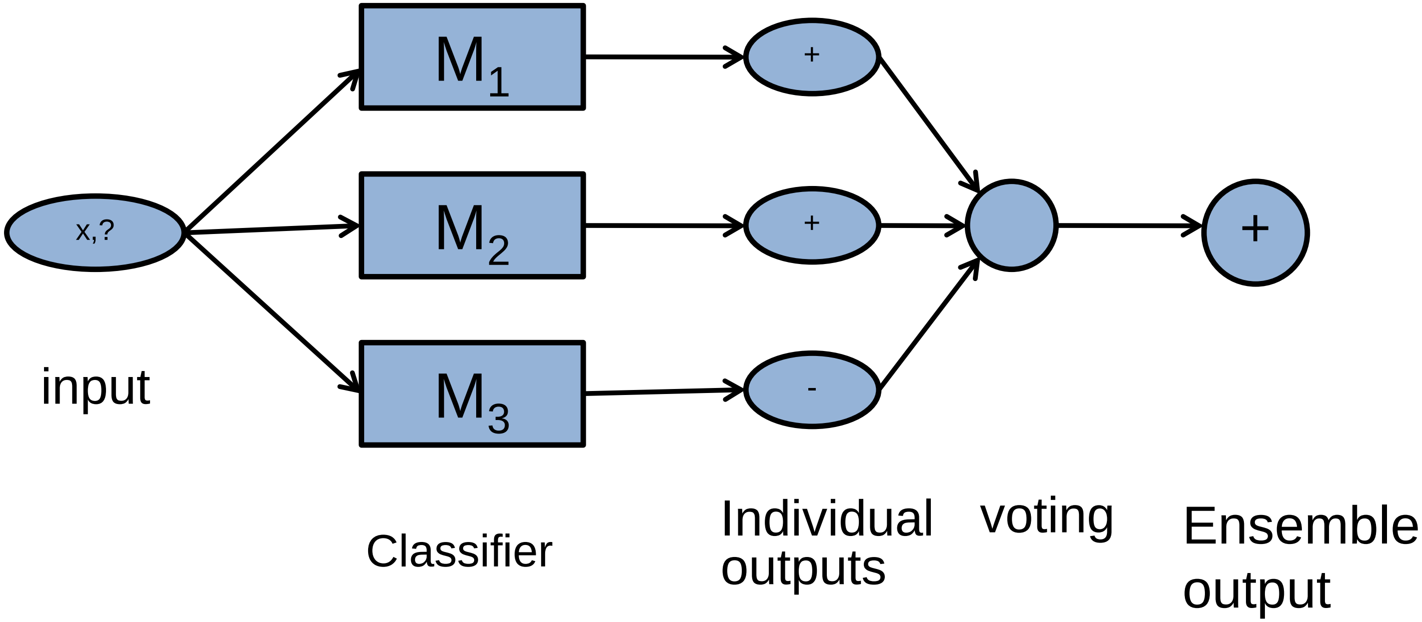
R1. if ($x > x_1$ and $y < y_2$) or ($x < x_1$ and $y < y_1$) then class = +

R2. if ($x > x_1$ and $y > y_2$) or ($x < x_1$ and $y > y_1$) then class = -

Existing classification models misclassify novel class instances

Existing Techniques: Ensemble based Approaches

Masud et al. [1][2]

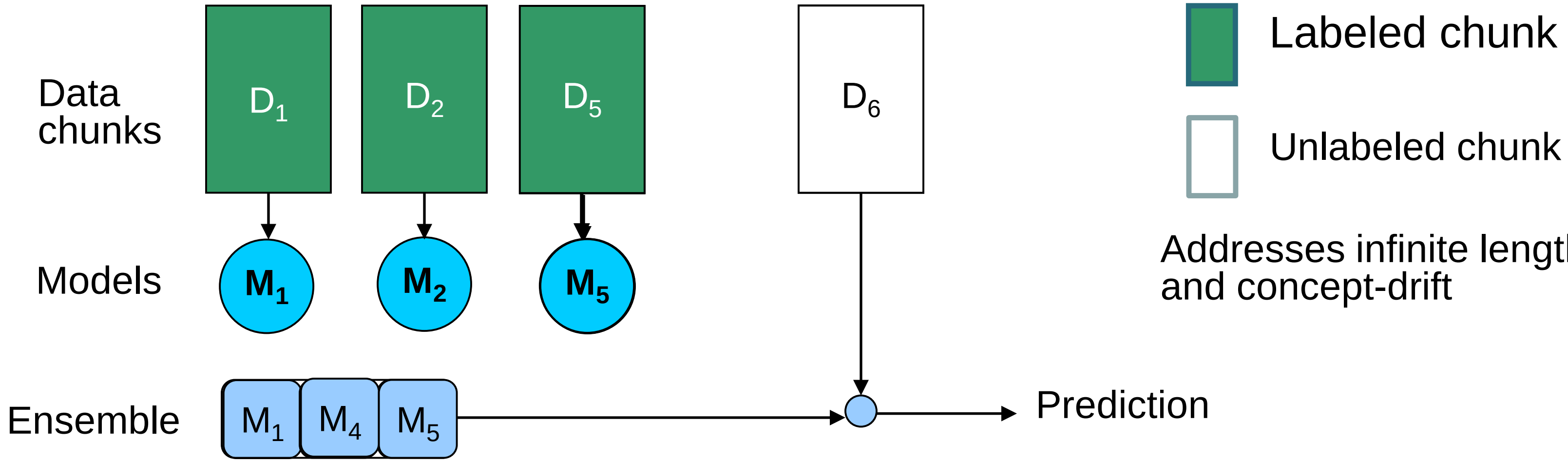


[1] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani M. Thuraisingham: A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data. ICDM 2008: 929-934
[2] Mohammad M. Masud, Clay Woolam, Jing Gao, Latifur Khan, Jiawei Han, Kevin W. Hamlen, Nikunj C. Oza: Facing the reality of data stream classification: coping with scarcity of labeled data. Knowl. Inf. Syst. 33(1): 213-244 (2011)

Existing Techniques: Ensemble Techniques

- Divide the data stream into equal sized chunks
  - Train a classifier from each data chunk
  - Keep the best *t* such classifier-ensemble
  - Example: for *t* = 3

Note: D_i may contain data points from different classes



Addresses infinite length and concept-drift

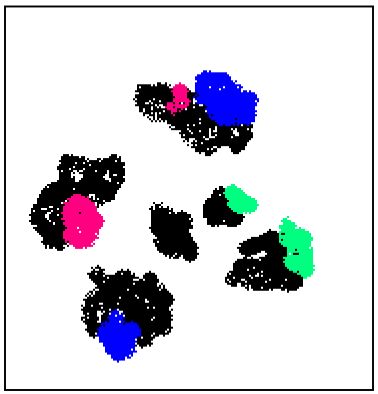
Novel Class Detection

Masud et al. [1][2], *Khateeb et al.* [3]

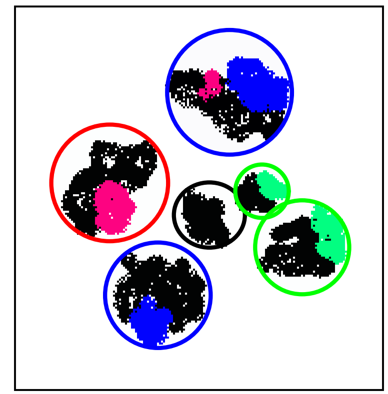
- Non parametric
  - – does not assume any underlying model of existing classes
- Steps:
  1. Creating and saving decision boundary during training
  2. Detecting and filtering outliers
  3. Measuring cohesion and separation among test and training instances

[1] Mohammad M. Masud, Qing Chen, Latifur Khan, Charu C. Aggarwal, Jing Gao, Jiawei Han, Ashok N. Srivastava, Nikunj C. Oza: Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams. IEEE Trans. Knowl. Data Eng. 25(7): 1484-1497 (2013)
[2] 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)
[3] Tahseen Al-Khateeb, Mohammad M. Masud, Latifur Khan, Charu C. Aggarwal, Jiawei Han, Bhavani M. Thuraisingham: Stream Classification with Recurring and Novel Class Detection Using Class-Based Ensemble. ICDM 2012: 31-40

Training with Semi-Supervised Clustering



Impurity based Clustering



Legend:

Black dots: unlabeled instances

Colored dots: labeled instances

Semi Supervised Clustering

Masud et al. [1][2]

- Objective function (dual minimization problem)

$$\mathcal{O}_{MCIKmeans} = \sum_{i=1}^{K} \Big(\sum_{\boldsymbol{x} \in \mathcal{X}_i} ||\boldsymbol{x} - \boldsymbol{u_i}||^2 + \sum_{\boldsymbol{x} \in \mathcal{L}_i} ||\boldsymbol{x} - \boldsymbol{u_i}||^2 * Imp_i \Big)$$

Intra-cluster dispersion ***Cluster impurity***

*Imp_i: = Aggregated dissimilarity count$_i$ * Entropy$_i$ = ADC$_i$ * Ent$_i$*

Aggregated dissimilarity count (*ADC*): $\mathcal{ADC}_i = \sum_{x \in \mathcal{L}_i} \mathcal{DC}_i(\boldsymbol{x}, y).$

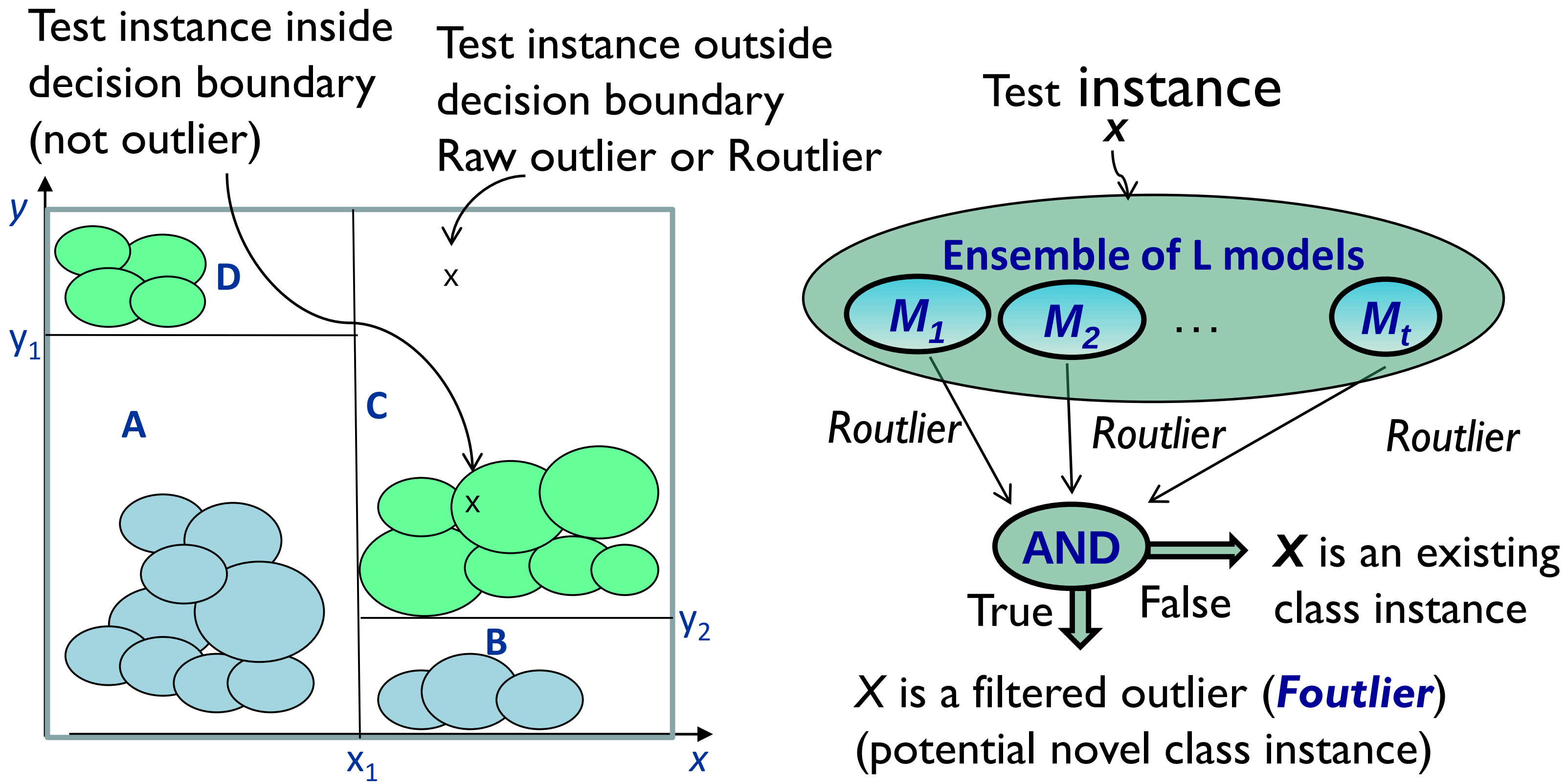
$$\mathcal{DC}_i(\boldsymbol{x}, y) = \begin{cases} 0 & \text{if } \boldsymbol{x} \text{ is unlabeled (i.e., } y = \phi) \\ |\mathcal{L}_i| - |\mathcal{L}_i(c)| & \text{if } \boldsymbol{x} \text{ is labeled and its label } y{=}c \end{cases}$$

Entropy (*Ent*): $Ent_i = \sum_{c=1}^{C} (-p_c^i * log(p_c^i))$

The minimization problem is solved using the Expectation-Maximization (E-M) framework

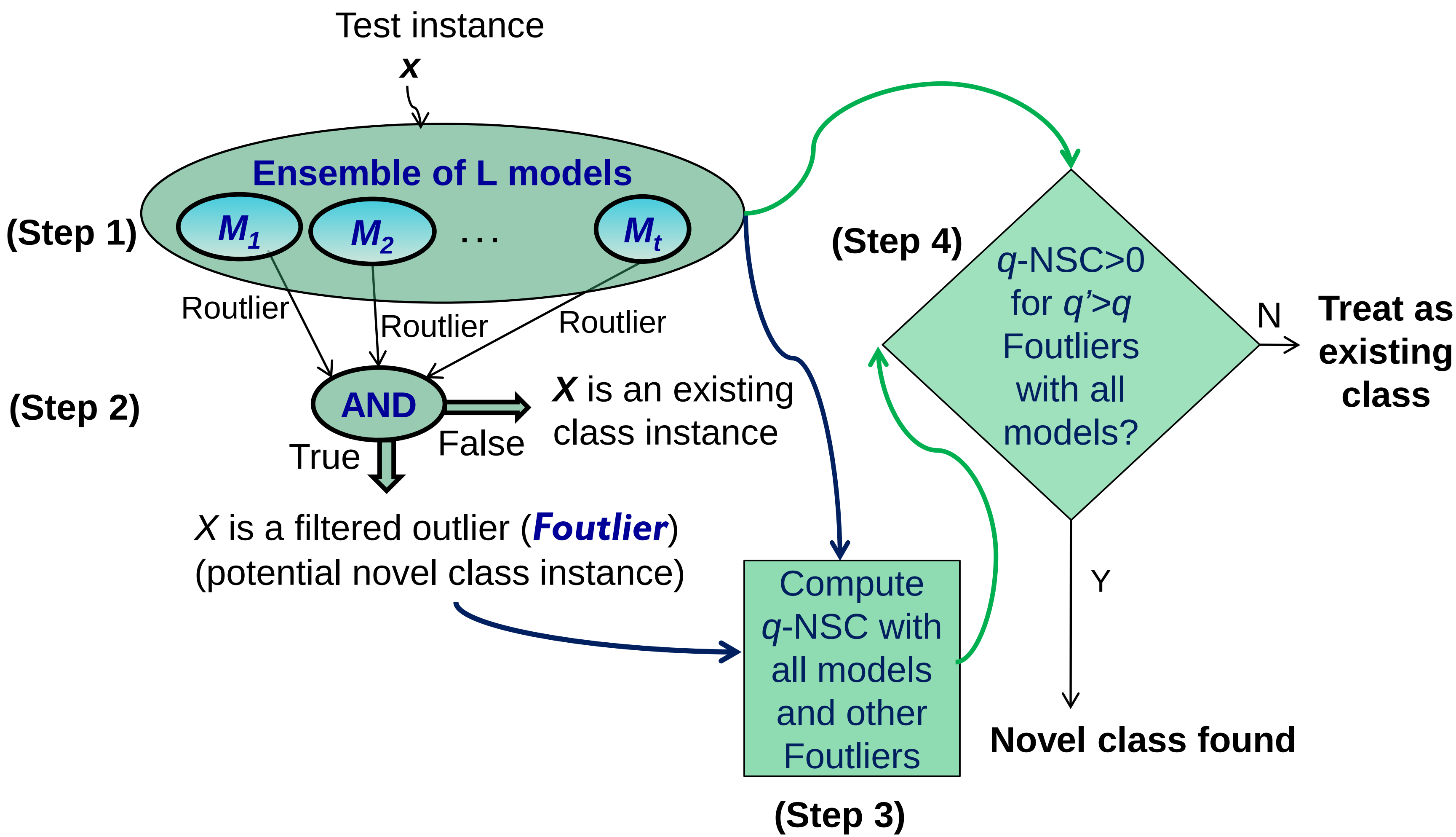
[1] Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani M. Thuraisingham: A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data. ICDM 2008: 929-934
[2] Mohammad M. Masud, Clay Woolam, Jing Gao, Latifur Khan, Jiawei Han, Kevin W. Hamlen, Nikunj C. Oza: Facing the reality of data stream classification: coping with scarcity of labeled data. Knowl. Inf. Syst. 33(1): 213-244 (2011)

Outlier Detection and Filtering

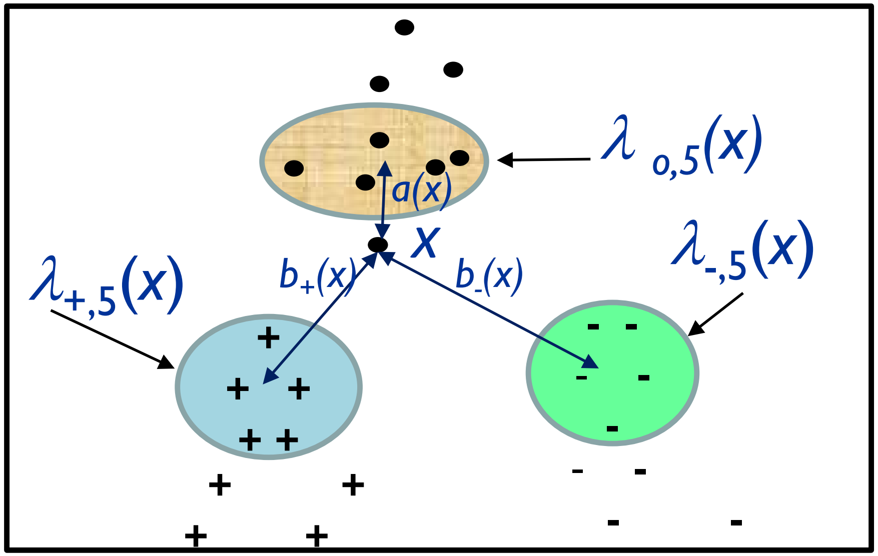


Foutliers may appear as a result of novel class, concept-drift, or noise. Therefore, they are filtered to reduce noise as much as possible.

Novel Class Detection



Computing Cohesion & Separation



- ➢λ_c(x) is the *set of nearest neighbors of x belonging to class c*
- ➢λ_o(x) is the *set of nearest Foutliers of x*

- *a(x)* = mean distance from an *Foutlier x* to the instances in $\lambda_{o,q}(x)$
- $b_{min}(x)$ = minimum among all $b_c(x)$ (e.g. $b_+(x)$ in figure)
- *q*-Neighborhood Silhouette Coefficient (*q*-NSC):

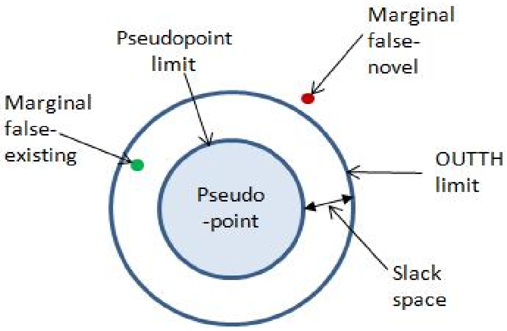
$$q\text{-}NSC(x) = \frac{(b_{min}(x) - a(x))}{\max(b_{min}(x), a(x))}$$

- If *q*-NSC(x) is positive, it means *x* is closer to *Foutliers* than any other class.

Detection of Concurrent Novel Classes

Masud et al. [1], *Faria et al.* [2]

- Challenges
  - High false positive (FP) (existing classes detected as novel) and false negative (FN) (missed novel classes) rates
  - Two or more novel classes arrive at a time

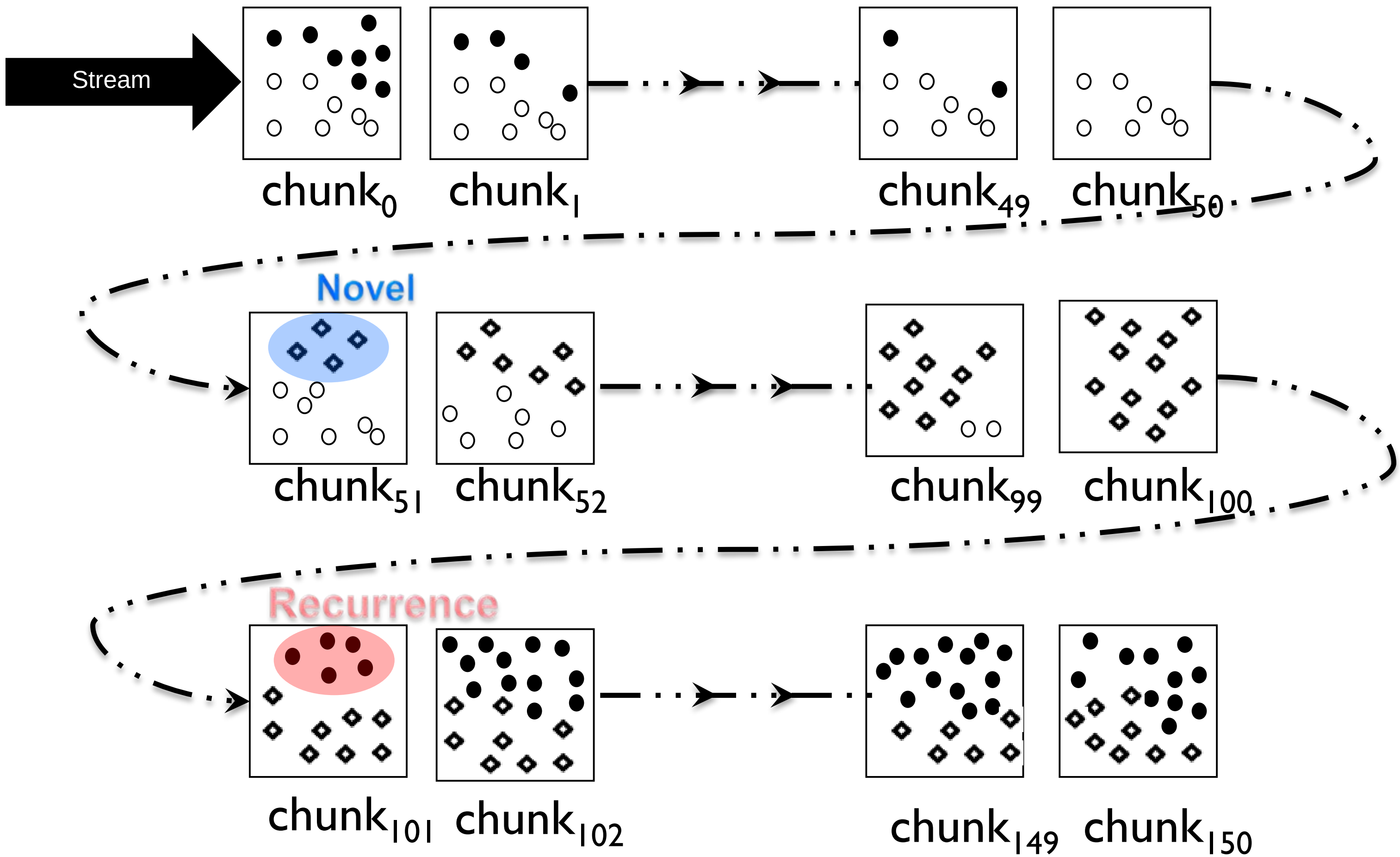


- Solutions
  - Dynamic decision boundary – based on previous mistakes
    - Inflate the decision boundary if high FP, deflate if high FN
  - Build statistical model to filter out noise data and concept drift from the outliers.
  - Multiple novel classes are detected by
    - Constructing a graph where outlier cluster is a vertex
    - Merging the vertices based on silhouette coefficient
    - Counting the number of connected components in the resultant (i.e., merged) graph

[1] Mohammad M. Masud, Qing Chen, Latifur Khan, Charu C. Aggarwal, Jing Gao, Jiawei Han, Bhavani M. Thuraisingham: Addressing Concept-Evolution in Concept-Drifting Data Streams. ICDM 2010: 929-934
[2] Elaine R. Faria, João Gama, André C. P. L. F. Carvalho: Novelty detection algorithm for data streams multi-class problems. SAC 2013: 795-800

Novel and Recurrence

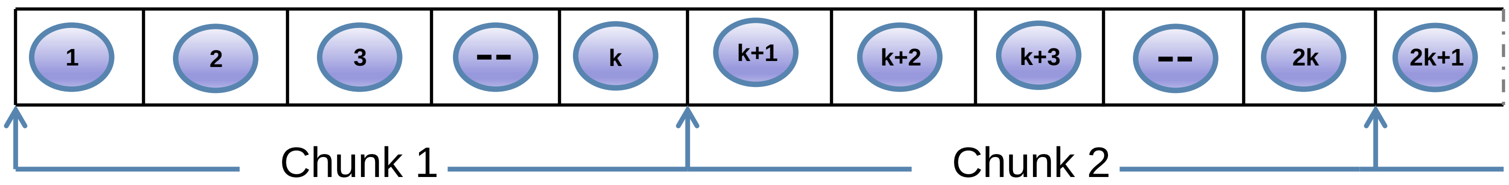
Khateeb et al. [1]



[1] Tahseen Al-Khateeb, Mohammad M. Masud, Latifur Khan, Charu C. Aggarwal, Jiawei Han, Bhavani M. Thuraisingham: Stream Classification with Recurring and Novel Class Detection Using Class-Based Ensemble. ICDM 2012: 31-40

Challenges: Fixed Chunk Size/ Decay Rate

Masud et al. [1], *Parker et al.* [2], *Aggarwal et al.* [3], *Klinkenberg*[4], *Cohen et al.* [5]

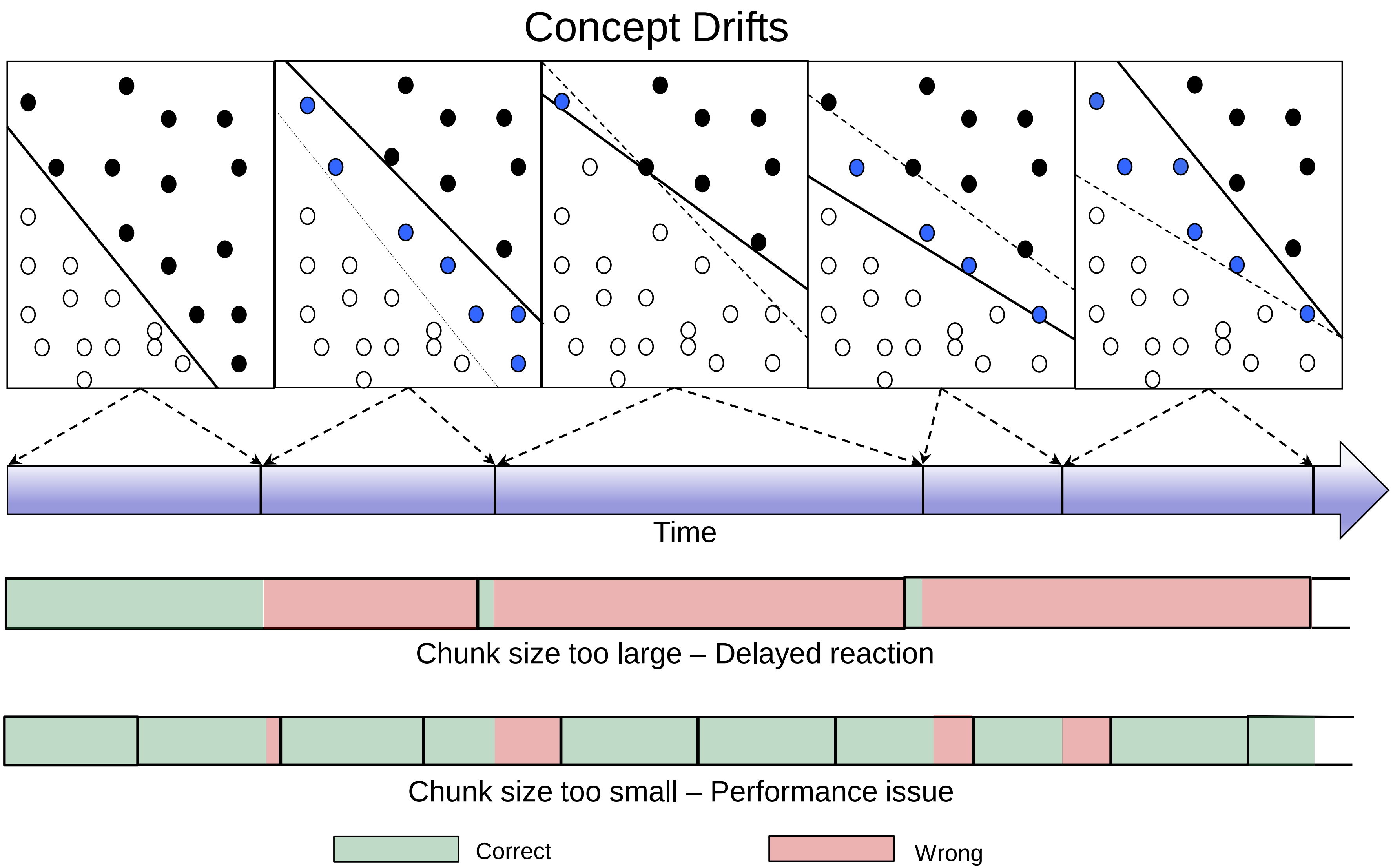


- Fixed chunk size
  - requires *a priori* knowledge about the time-scale of change.
  - delayed reaction if the chunk size is too large.
  - unnecessary frequent training during stable period if chunk size is too small.

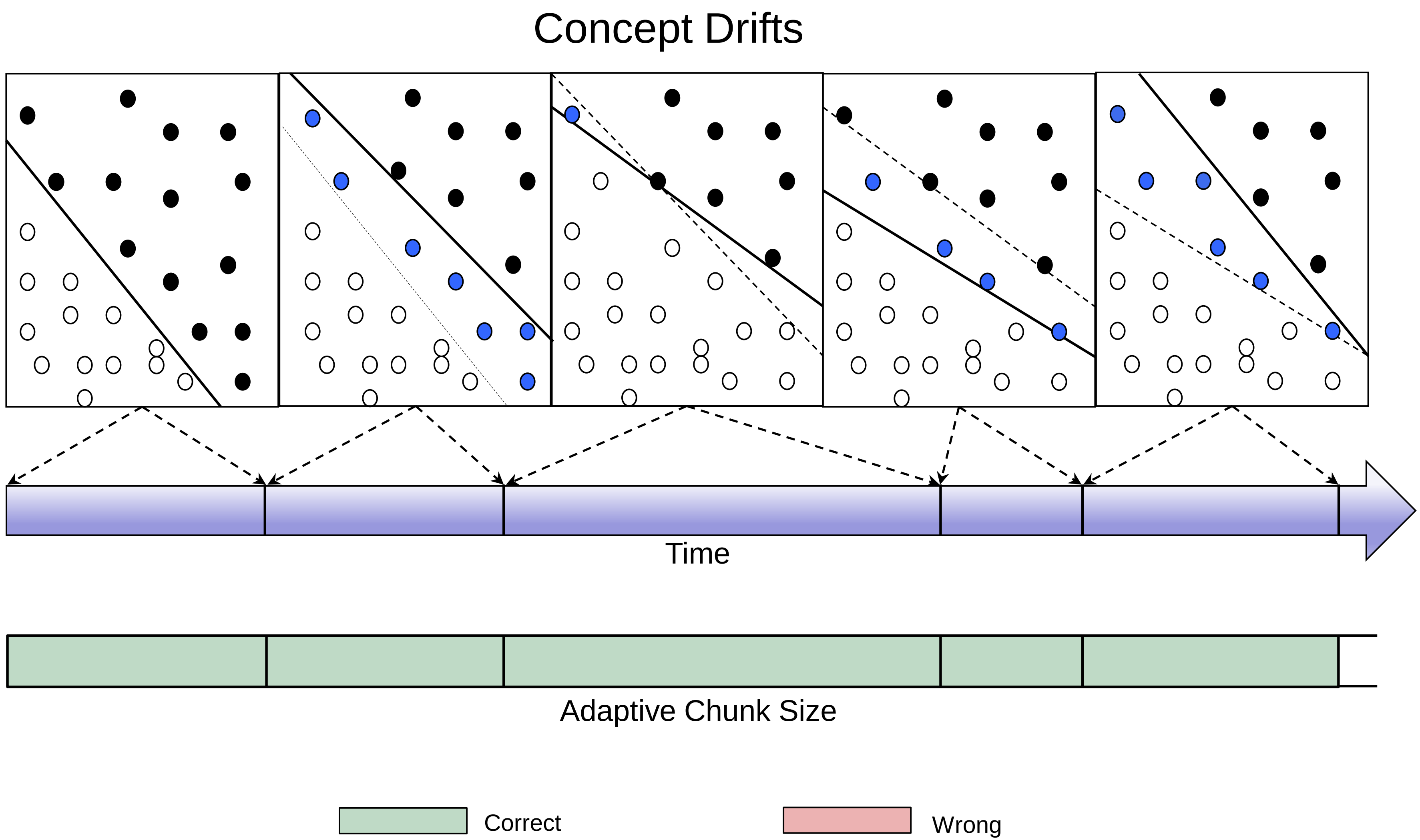
- Fixed decay rate
  - assigns weight to data instances based on their age.
  - decay constant must match the unknown rate of change.

[1] 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)
[2] Brandon Shane Parker, Latifur Khan: Detecting and Tracking Concept Class Drift and Emergence in Non-Stationary Fast Data Streams. AAAI 2015: 2908-2913
[3] Charu C. Aggarwal, Philip S. Yu: On Classification of High-Cardinality Data Streams. SDM 2010: 802-813
[4] Ralf Klinkenberg: Learning drifting concepts: Example selection vs. example weighting. Intell. Data Anal. 8(3): 281-300 (2004)
[5] Edith Cohen, Martin J. Strauss: Maintaining time-decaying stream aggregates. J. Algorithms 59(1): 19-36 (2006)

Challenges: Fixed Chunk Size

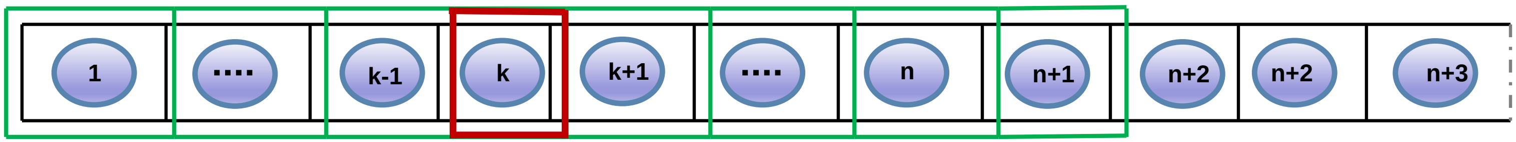


Solution: Adaptive Chunk Size



Adaptive Chunk - Sliding Window

Gamma et al. [1], *Bifet et al.* [2], *Harel et al.* [3]

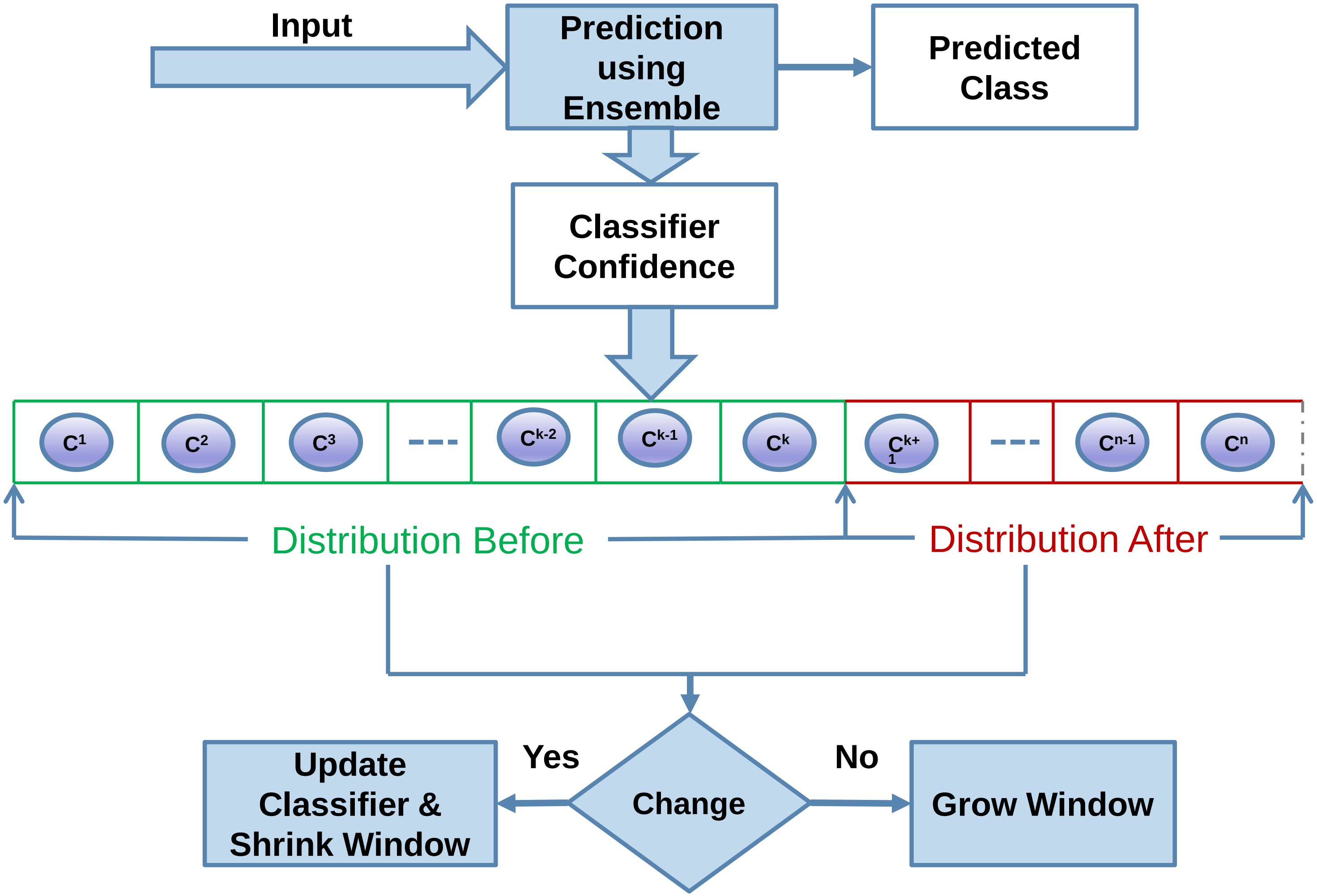


- Existing dynamic sliding window techniques
  - monitor error rate of the classifier.
  - Update classifier if starts to show bad performance.
  - **fully supervised**, which is not feasible in case of real-world data streams.

[1] João Gama, Gladys Castillo: Learning with Local Drift Detection. ADMA 2006: 42-55
[2] Albert Bifet, Ricard Gavaldà: Learning from Time-Changing Data with Adaptive Windowing. SDM 2007: 443-448
[3] Maayan Harel, Shie Mannor, Ran El-Yaniv, Koby Crammer: Concept Drift Detection Through Resampling. ICML 2014: 1009-1017

Adaptive Chunk - Unsupervised

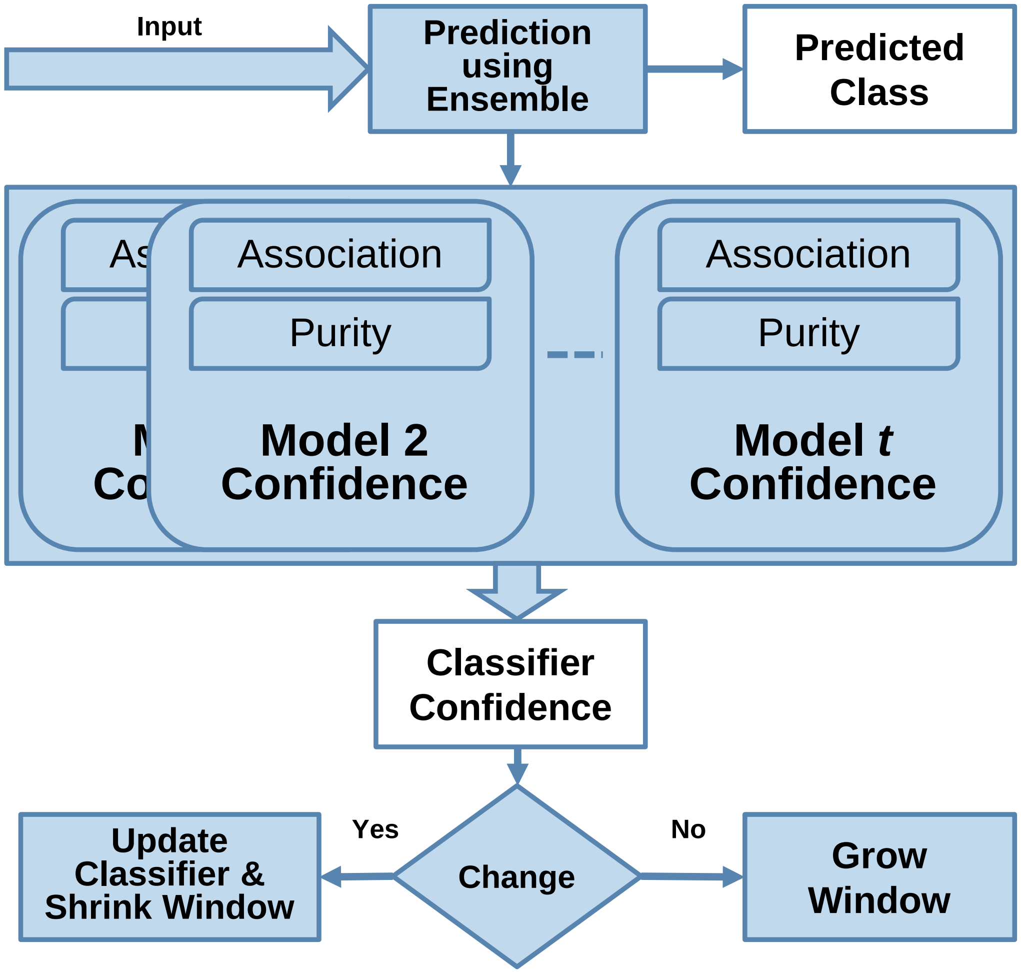
Haque et al. [1][2]



[1] Ahsanul Haque, Latifur Khan, Michael Baron, Bhavani M. Thuraisingham, Charu C. Aggarwal: Efficient handling of concept drift and concept evolution over Stream Data. ICDE 2016: 481-492.
[2] Ahsanul Haque, Latifur Khan, Michael Baron: SAND: Semi-Supervised Adaptive Novel Class Detection and Classification over Data Stream. AAAI 2016: 1652-1658.

Adaptive Chunk - Unsupervised

Haque et al. [1][2]



[1] Ahsanul Haque, Latifur Khan, Michael Baron, Bhavani M. Thuraisingham, Charu C. Aggarwal: Efficient handling of concept drift and concept evolution over Stream Data. ICDE 2016: 481-492

[2] Ahsanul Haque, Latifur Khan, Michael Baron: SAND: Semi-Supervised Adaptive Novel Class Detection and Classification over Data Stream. AAAI 2016: 1652-1658.

Confidence of a model

- For each testing instance x:
  - Confidence for i^{th} model, $c_i^x = \boldsymbol{h_i^x}.\boldsymbol{z_i}$
    - $\boldsymbol{h_i^x} = (a_i^x, p_i^x)$ is a vector of estimator values on test instance x.
    - $\boldsymbol{z_i}$ = *vector containing weights of the estimators for* i^{th} *model.*

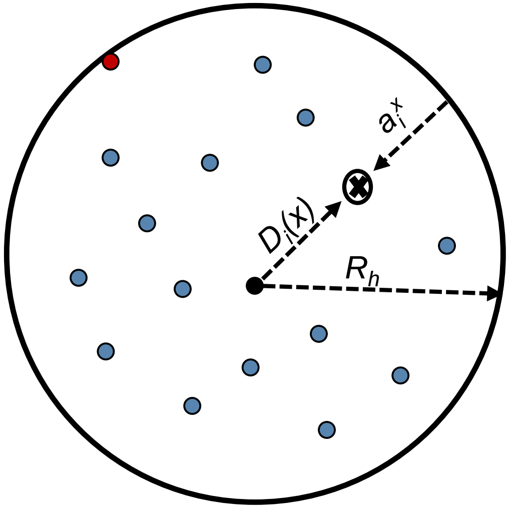
- To estimate confidence of the entire ensemble, we take the average confidence of the models towards the predicted class.

Confidence Estimators

- Let h be the closest cluster from data instance x in model M_i, confidence of M_i in classifying instance x is calculated based on the following estimators:

  - Association: $a_i^x = R_h - D_i(x)$, *where R_h is the radius of h and $D_i(x)$ is the distance of x from h.*

  - Purity: $p_i^x = N_m / N_s$, where N_s is the number of labeled instances in h, and N_m is the number of instances from the majority class in h.



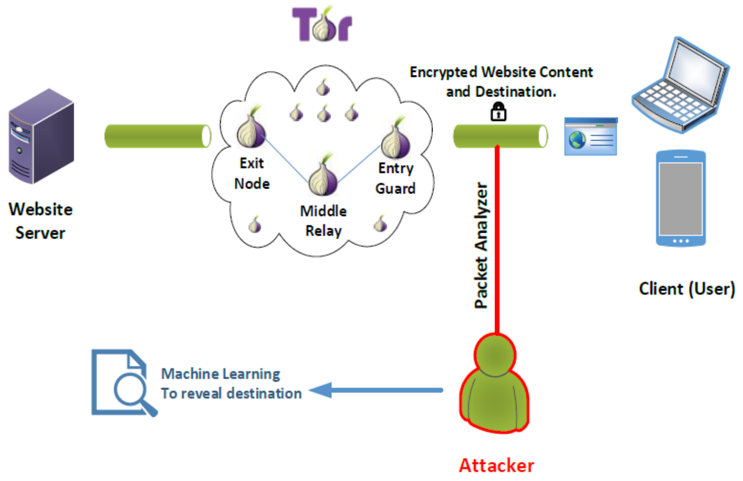
- N_s = 15, N_m = 14
- p_i^x = 14/15

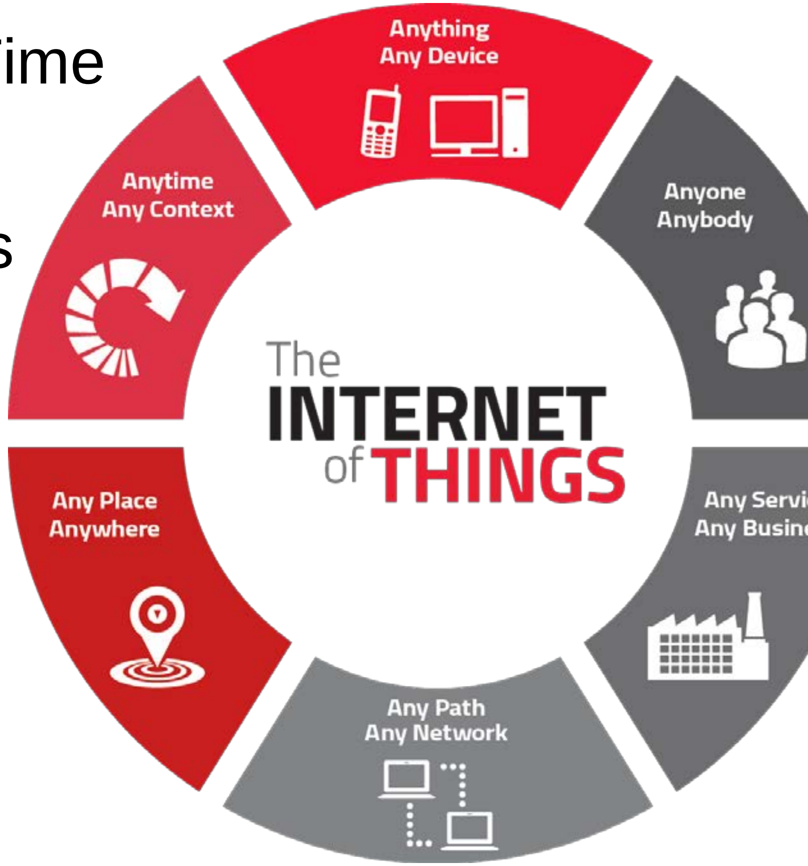
Big Stream Data: Current & Future

- Stream Mining*
  - IOT Big Stream Mining—Real Time
  - Security:

–Encrypted Stream Traffic Analysis

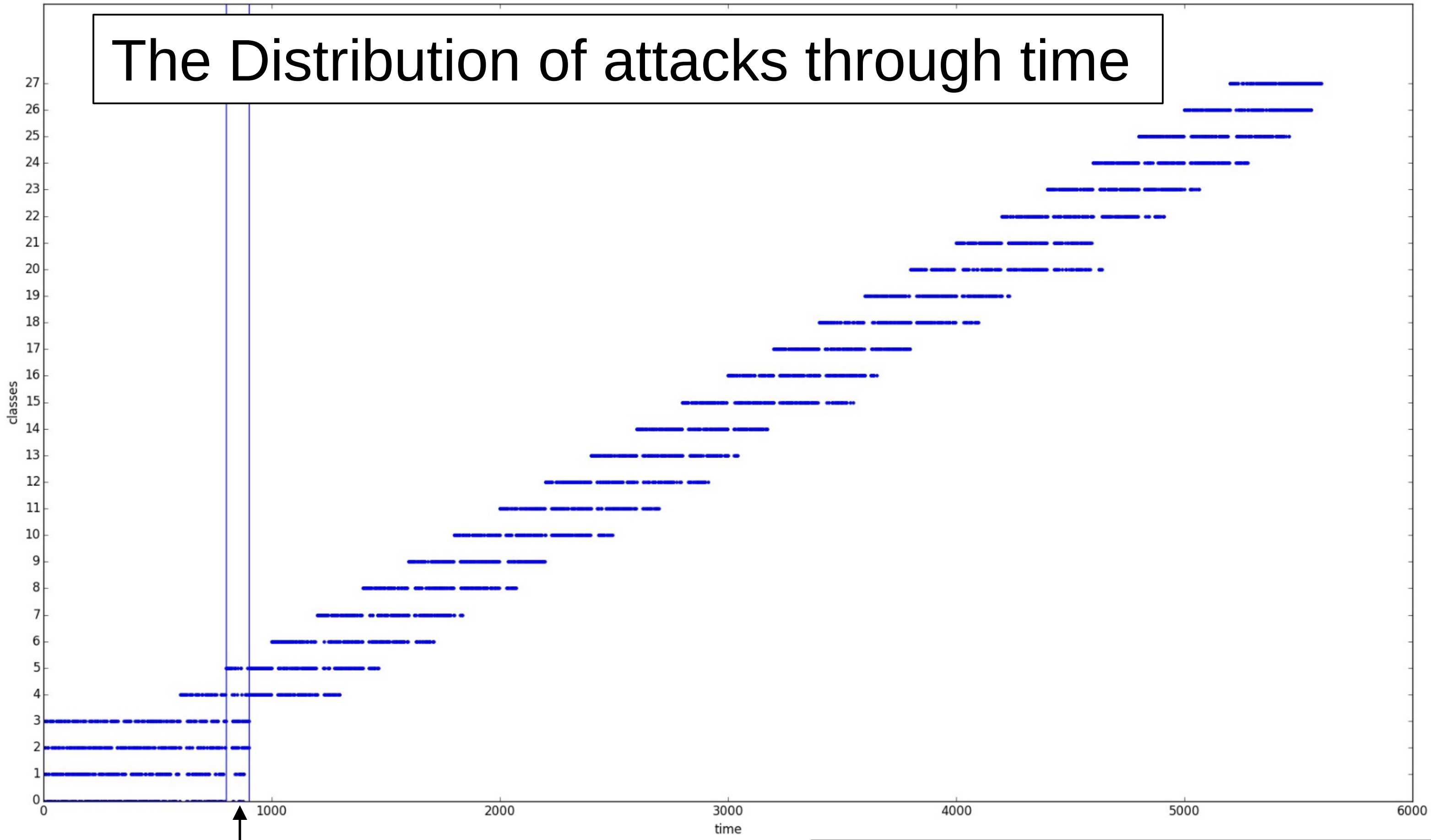
- Website Fingerprinting





*Parker, B., Khan, L.: Detecting and tracking concept class drift and emergence in non-stationary fast data streams. In Proc. Of Twenty-Ninth AAAI Conference on Artificial Intelligence. (Jan 2015).

Application (1): Detecting Zero-day attacks



Chunk contains 1 new attack and 5 existing classes.

- 28 classes
- Each class has 200 data points
- Chunk size = 100

Results Detecting Zero-day attacks

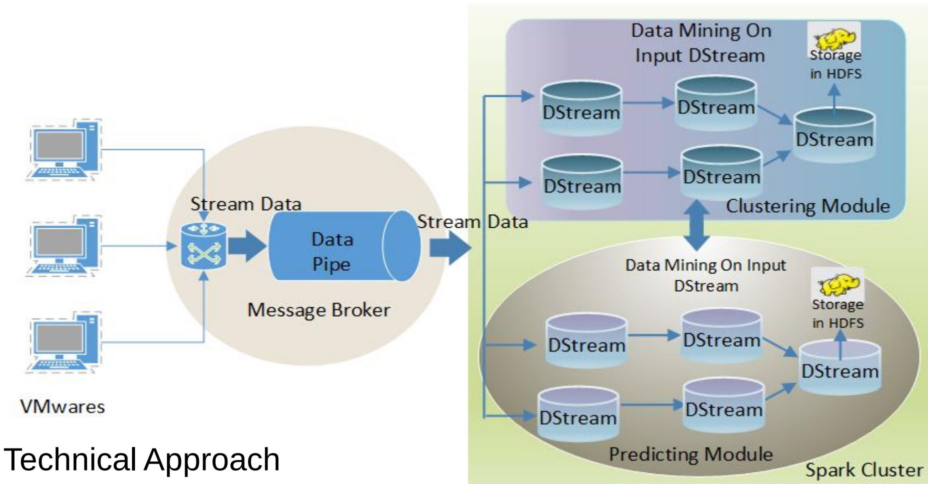
	FP%		FN%		Err%	
	Dxminer [1]	Dxminer+DAE features [2]	Dxminer	Dxminer + DAE features	Dxminer	Dxminer + DAE features
BiDi Packets:	26.988	**0.0**	24.869	**15.635**	42.037	**4.396**
N-grams SysCalls:	31.87	**19.33**	21.414	**4.761**	46.754	**17.66**

- **Dxminer[1] = novel class detection method**
- **DAE[2] = Denoising Autoencoders features**

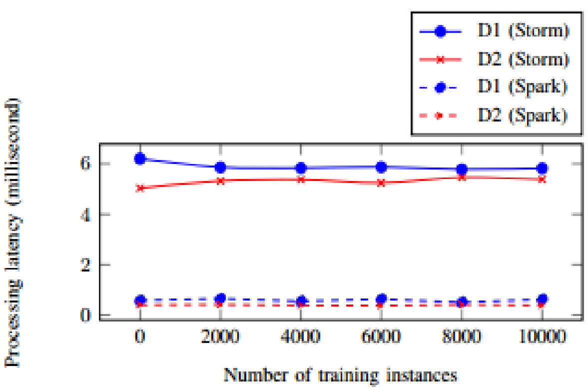
1. Tahseen Al-Khateeb et. al., **Recurring and Novel Class Detection Using Class-Based Ensemble for Evolving Data Stream.** *IEEE Trans. Knowl. Data Eng. 28(10):* 2752-2764 (2016)
2. Pascal Vincent et. al., **Stacked Denoising Autoencoders: Learning Useful Representations in a Deep Network with a Local Denoising Criterion.** *J. Mach. Learn. Res.* 11: 3371-3408 (2010)

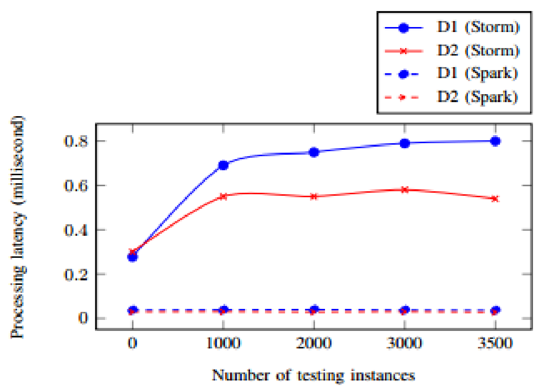
Spark-based Real-time Anomaly Detection: Framework (Application 2)

❑ Stream Data Mining Module



Technical Approach





❑ Experimental Result

- Dataset1 - Performance data of Spark jobs
- Dataset2 - Performance data for Yahoo Cloud Service Benchmark database operation.

Cluster Environment

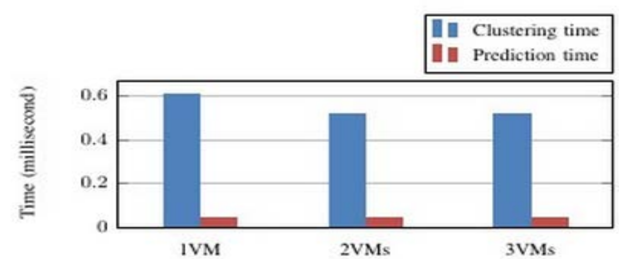
Component	Number of parallelism
Worker for emitting tuples	05
Worker for clustering	08
Worker for prediction	08

Training

Number of data point	Dataset 1	Dataset 2
Number of data points	10, 000	10, 000
Number of clusters	63	134

Testing

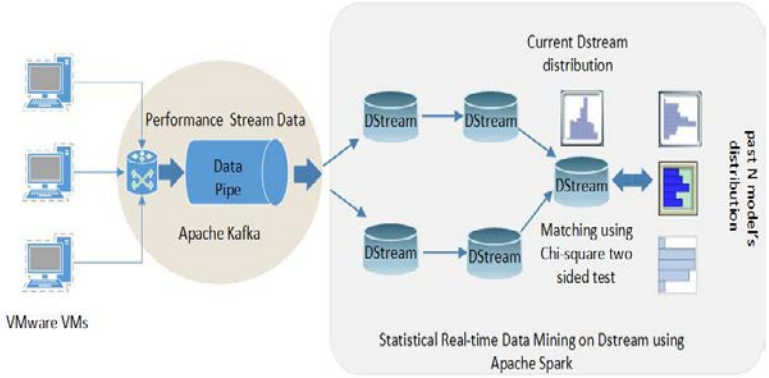
Dataset	TPR	FNR	TNR	FPR
D1	98.00%	2.00%	99.83%	0.17%
D2	99.20%	0.80%	99.06%	0.94%



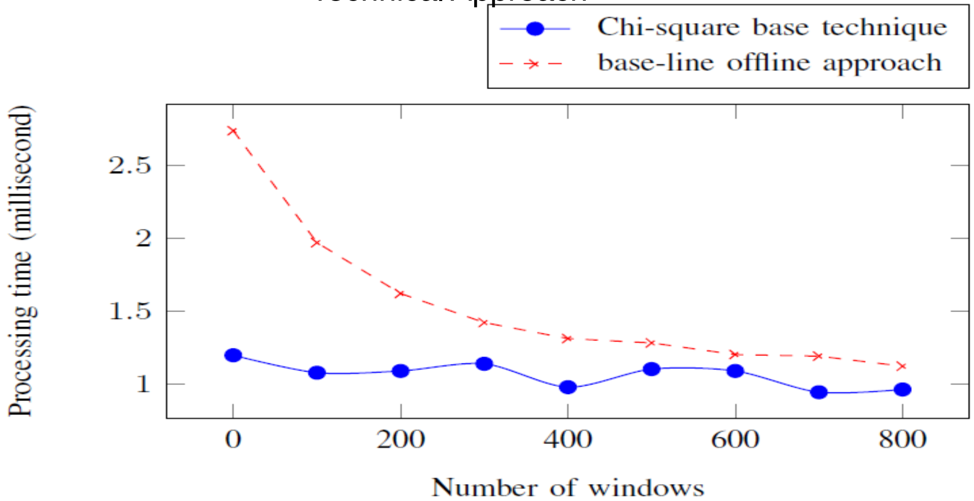
M. Solaimani, M. Iftekhar, L. Khan, B. Thuraisingham, J. Ingram, and S.E. Seker, “Online anomaly detection for multi-source VMware using a distributed streaming framework." Software: Practice and Experience (2016).

Statistical Technique for Online Anomaly Detection Using Spark: Framework

❑ Stream Data Mining Module



Technical Approach



❑ Experimental Result

Cluster Environment

Component	Number of parallelism
Worker for emitting tuples	05
Worker for statistical analysis	08

Statistical Model

Number of data point	Dataset 1	Dataset 2
Number of windows	800	800
Total Number of points	80, 000	80, 000

Testing

Method	TPR	FNR	TNR	FPR
Chi-square based Online model	90.00 %	10.00 %	98.80 %	1.2%
Base-line offline method	8.24%	91.76 %	99.16 %	0.84%

M. Solaimani, M. Iftekhar, L. Khan, and B. Thuraisingham, "Statistical Technique for Online Anomaly Detection Using Spark Over Heterogeneous Data from Multi-source VMware Performance Data," in proceedings of *the IEEE International Conference on Big Data 2014 (IEEE BigData 2014), Washington DC, USA.*

Application (3): Encrypted Traffic Fingerprinting

Al-Naami et al. [1][2]

- Traffic Fingerprinting (TFP) is a Traffic Analysis (TA) attack that threatens web/app navigation privacy.
- TFP allows attackers to learn information about a website/app accessed by the user, by recognizing patterns in traffic.
- Examples: Website Fingerprinting

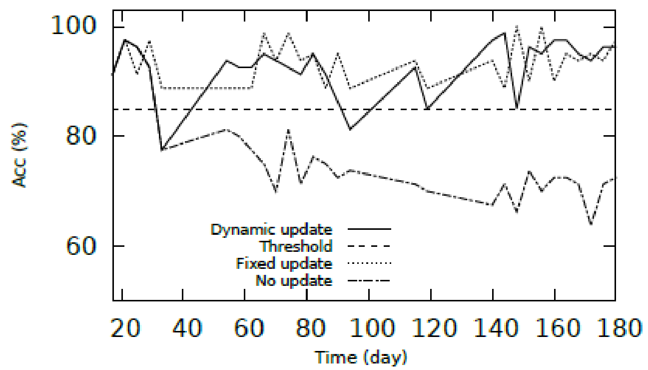
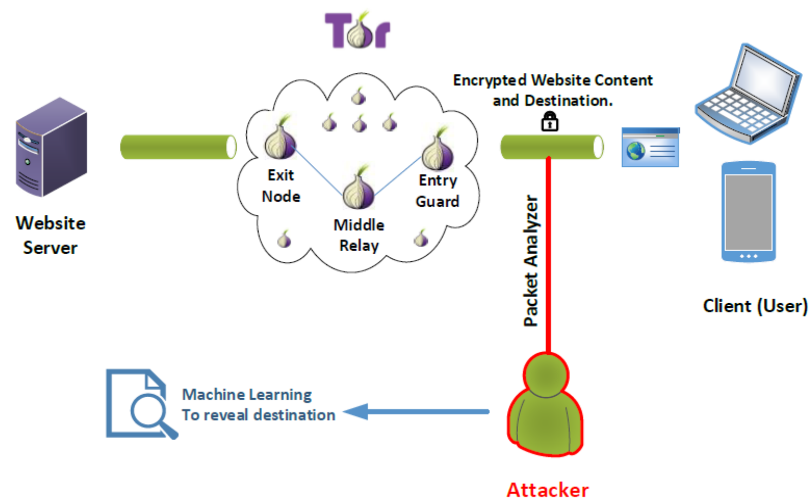


Figure 9: Adaptive Learning.



[1] K. Al-Naami, G. Ayoade, A. Siddiqui, N. Ruozzi, L. Khan and B. Thuraisingham, "P2V: Effective Website Fingerprinting Using Vector Space Representations," Computational Intelligence, 2015 IEEE Symposium Series on, Cape Town, 2015, pp. 59-66.
[2] K. Al-Naami, S. Chandra, A. Mustafa, L. Khan, Z. Lin, K. Hamlen, and B. Thuraisingham. 2016. Adaptive encrypted traffic fingerprinting with bi-directional dependence. In Proceedings of the 32nd Annual Conference on Computer Security Applications (ACSAC '16), Los Angeles, CA.

A Framework To Recommend New Political Actors With Role In Real-time (4)

- Dictionary (CAMEO) development requires
  - Human involvement
  - Not up-to-date
  - Higher Cost
  - Processing large number of articles
- Our Goal:
  - Reduce human effort and cost
  - Recommending news actor real-time
  - Update dictionary

CAMEO Dictionary

BARACK_OBAMA
+MR_OBAMA_
+MR._OBAMA_
+PRESIDENT_OBAMA_
+OBAMA
+PRESIDENT_BARACK_OBAMA
+US_PRESIDENT_BARACK_OBAMA
+AMERICAN_PRESIDENT_BARACK_OBAMA
……….
+OBAMA_ADMINISTRATION
[USAELI 780101-000101]
[USAGOV >090120]

A Framework To Recommend New

- Political with multiple alias names,
  - e.g., 'Barack Hussein Obama', 'Barack Obama', etc.

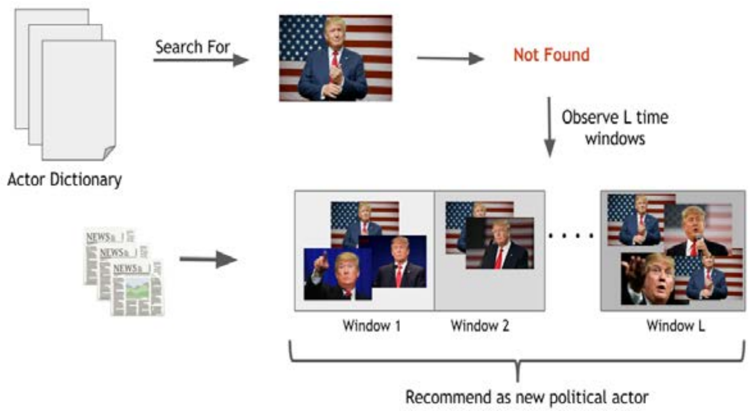
- Role of a political actor changes over time.
  - e.g., 'Shimon Peres' has multiple political roles in Israel

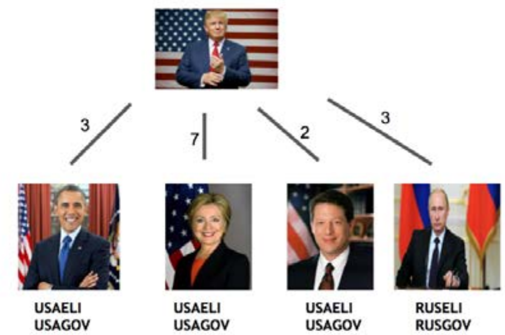
- Processing a large volume of news articles
  - demands scalable, distributed computing

A Framework To Recommend New Political Actors With Role In Real-time

- ❑ A real-time framework for recommendation
  - ➢ Possible new actors with their roles
  - ➢ Grouping actor aliases
- ❑ Frequency-based actor ranking algorithm
- ❑ A graph-based technique to recommend roles
  - ➢ A new actor
  - ➢ Existing actor whose role varies over time
  - ➢ Integrating external knowledge base (e.g., Wikipedia)
- ❑ Time window-based recommendation system.

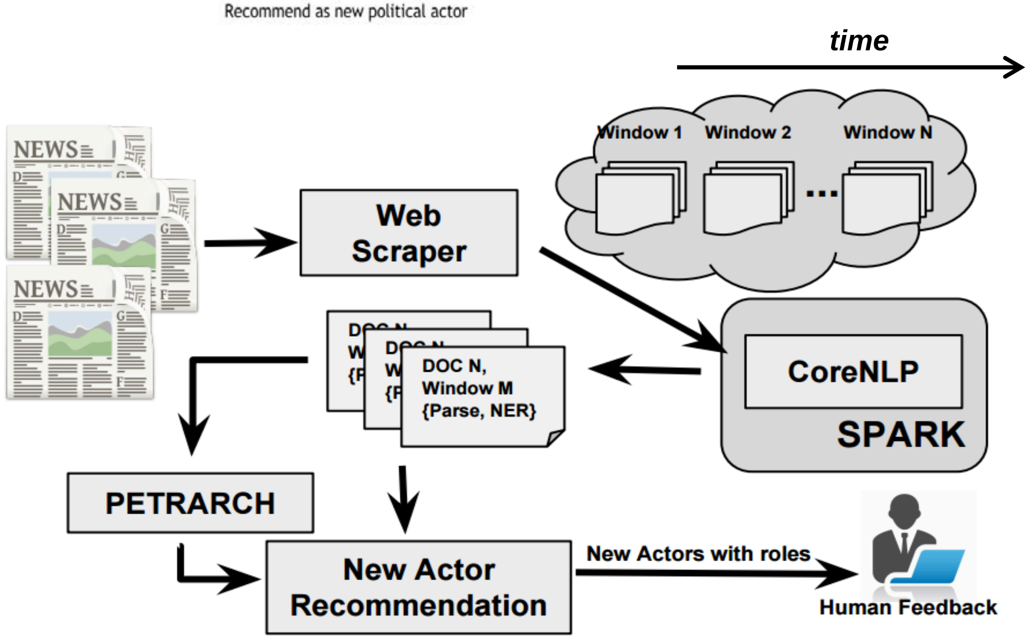
Real-time Political Actor Detection Over Textual Political Stream





Most probable role for Donald Trump will be

USAELI
USAGOV
RUSELI



Real-time new political actor recommendation framework.

Challenges

✓Same actor with multiple alias names

✓Identify novel actor along with roles

✓Existing political actor’s role changes over time

✓Processing high volume of news articles across the world

M. Solaimani, R. Gopalan, L. Khan, P. T. Brandt , and B. Thuraisingham, "Spark-based political event coding.” 2016 IEEE Second International Conference on Big Data Computing Service and Applications (BigDataService), pp. 14-23. IEEE, 2016, Oxford, UK.

Future Direction

- Adversarial active learning
  - Traditional algorithms are vulnerable to adversarial manipulation.
  - Instances should be selected carefully.
- Efficient online change detection
- Deep Learning Guided Stream Mining
- Multi-stream Analytics

References

- Ahsanul Haque, Latifur Khan, Michael Baron, Bhavani M. Thuraisingham, Charu C. Aggarwal: *Efficient handling of concept drift and concept evolution over Stream Data.* ICDE 2016: 481-492
- Ahsanul Haque, Latifur Khan, Michael Baron: *SAND: Semi-Supervised Adaptive Novel Class Detection and Classification over Data Stream*. AAAI 2016: 1652-1658
- Swarup Chandra, Ahsanul Haque, Latifur Khan, Charu C. Aggarwal: *An Adaptive Framework for Multistream Classification*. CIKM 2016: 1181-1190
- Khaled Al-Naami, Swarup Chandra, Ahmad M. Mustafa, Latifur Khan, Zhiqiang Lin, Kevin W. Hamlen, Bhavani M. Thuraisingham: *Adaptive encrypted traffic fingerprinting with bi-directional dependence*. ACSAC 2016: 177-188
- Parker, B., Khan, L.: *Detecting and tracking concept class drift and emergence in non-stationary fast data streams.* In: Twenty-Ninth AAAI Conference on Articial Intelligence. (Jan 2015).
- Tahseen Al-Khateeb, Mohammad M. Masud, Latifur Khan, Charu C. Aggarwal, Jiawei Han, Bhavani M. Thuraisingham: *Stream Classification with Recurring and Novel Class Detection Using Class-Based Ensemble*. ICDM 2012: 31-40
- Mohammad M. Masud, Qing Chen, Latifur Khan, Charu C. Aggarwal, Jing Gao, Jiawei Han, Ashok N. Srivastava, Nikunj C. Oza: *Classification and Adaptive Novel Class Detection of Feature-Evolving Data Streams.* IEEE Trans. Knowl. Data Eng. 25(7): 1484-1497 (2013)
- 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)
- Mohammad M. Masud, Jing Gao, Latifur Khan, Jiawei Han, Bhavani M. Thuraisingham: *A Practical Approach to Classify Evolving Data Streams: Training with Limited Amount of Labeled Data.* ICDM 2008: 929-934
- Mohammad M. Masud, Clay Woolam, Jing Gao, Latifur Khan, Jiawei Han, Kevin W. Hamlen, Nikunj C. Oza: *Facing the reality of data stream classification: coping with scarcity of labeled data.* Knowl. Inf. Syst. 33(1): 213-244 (2011)
- João Gama, Gladys Castillo: *Learning with Local Drift Detection.* ADMA 2006: 42-55
- Albert Bifet, Ricard Gavaldà: *Learning from Time-Changing Data with Adaptive Windowing.* SDM 2007: 443-448
- Maayan Harel, Shie Mannor, Ran El-Yaniv, Koby Crammer: *Concept Drift Detection Through Resampling.* ICML 2014: 1009-1017
- Charu C. Aggarwal, Philip S. Yu: *On Classification of High-Cardinality Data Streams*. SDM 2010: 802-813
- Wei Fan, Yi-an Huang, Haixun Wang, Philip S. Yu: *Active Mining of Data Streams*. SDM 2004: 457-461
- Xingquan Zhu, Peng Zhang, Xiaodong Lin, Yong Shi: *Active Learning from Data Streams*. ICDM 2007: 757-762
- Ralf Klinkenberg: *Learning drifting concepts: Example selection vs. example weighting.* Intell. Data Anal. 8(3): 281-300 (2004)
- Edith Cohen, Martin J. Strauss: *Maintaining time-decaying stream aggregates.* J. Algorithms 59(1): 19-36 (2006)