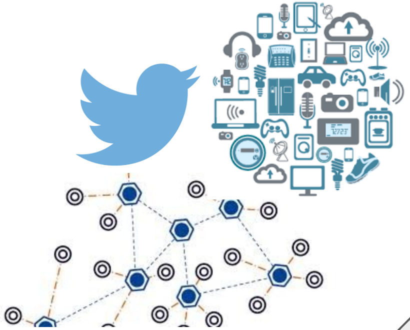
An Adaptive Framework for Multistream Classification

Swarup Chandra, Ahsanul Haque, Latifur Khan and Charu Aggarwal*
University of Texas at Dallas
*IBM Research

This material is based upon work supported by

Data Stream Classification



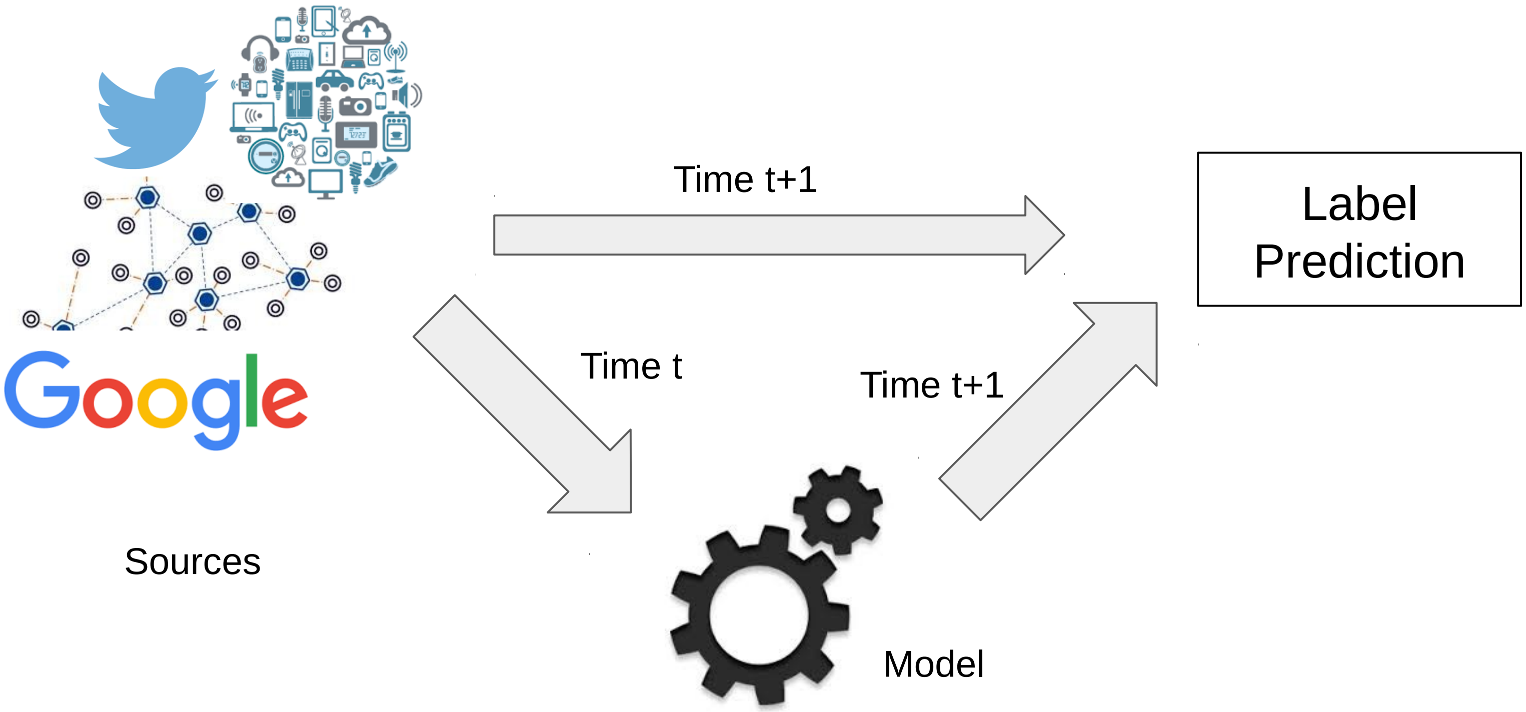
Time t

Sources

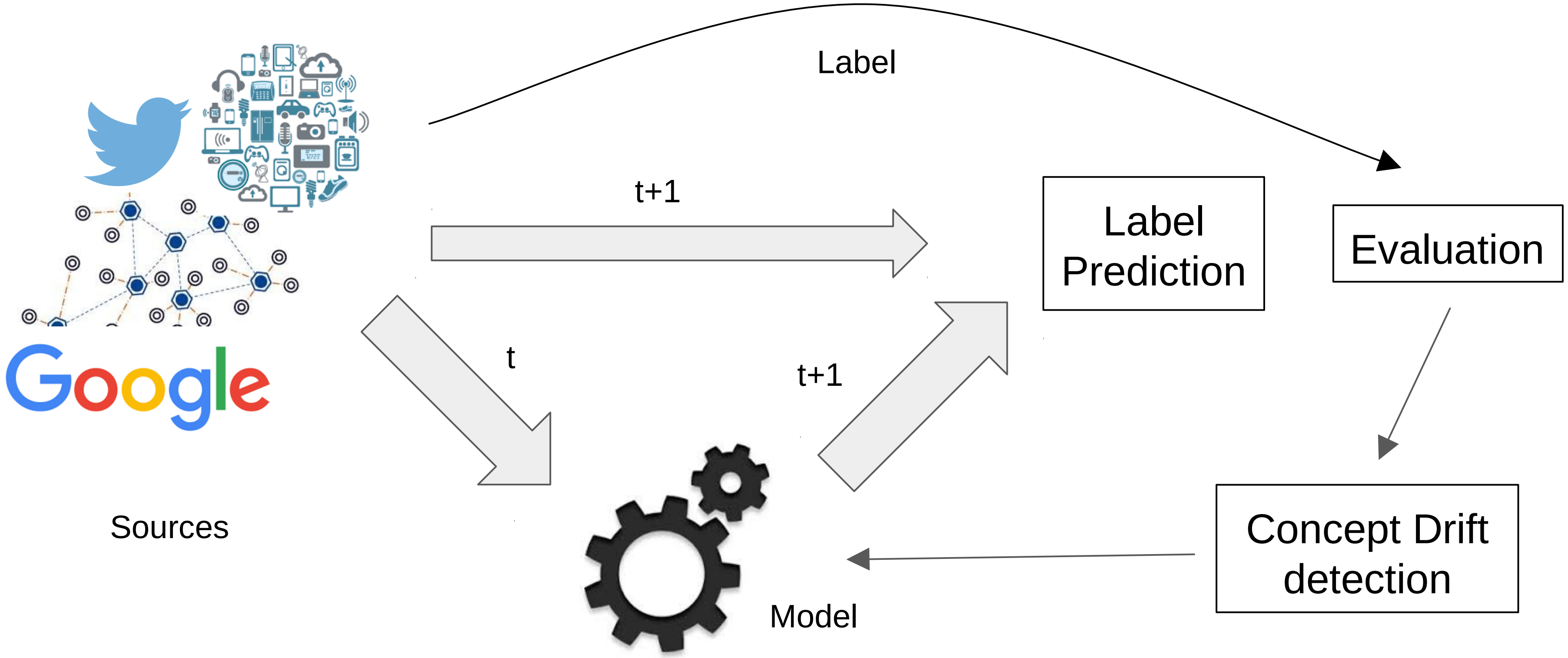
Model

- 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

Data Stream Classification

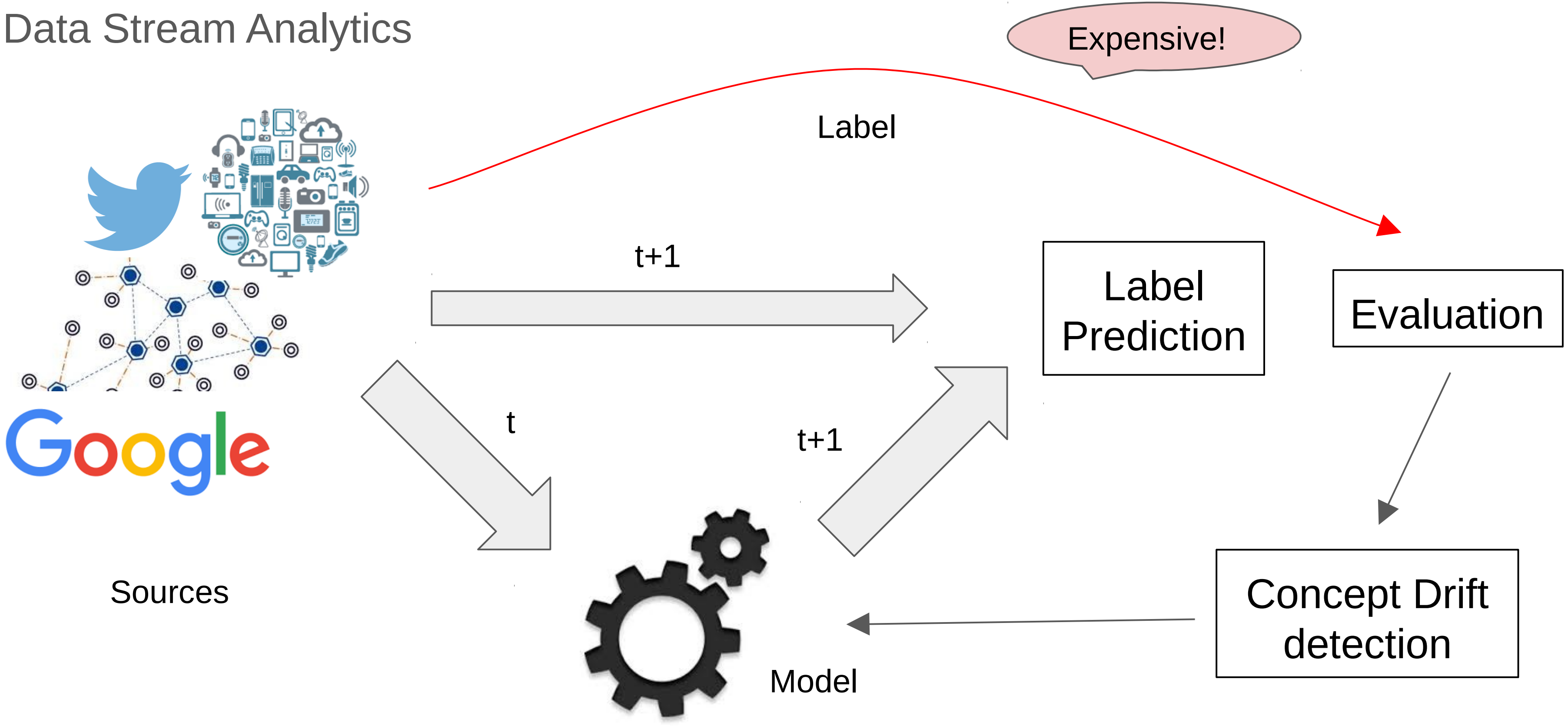


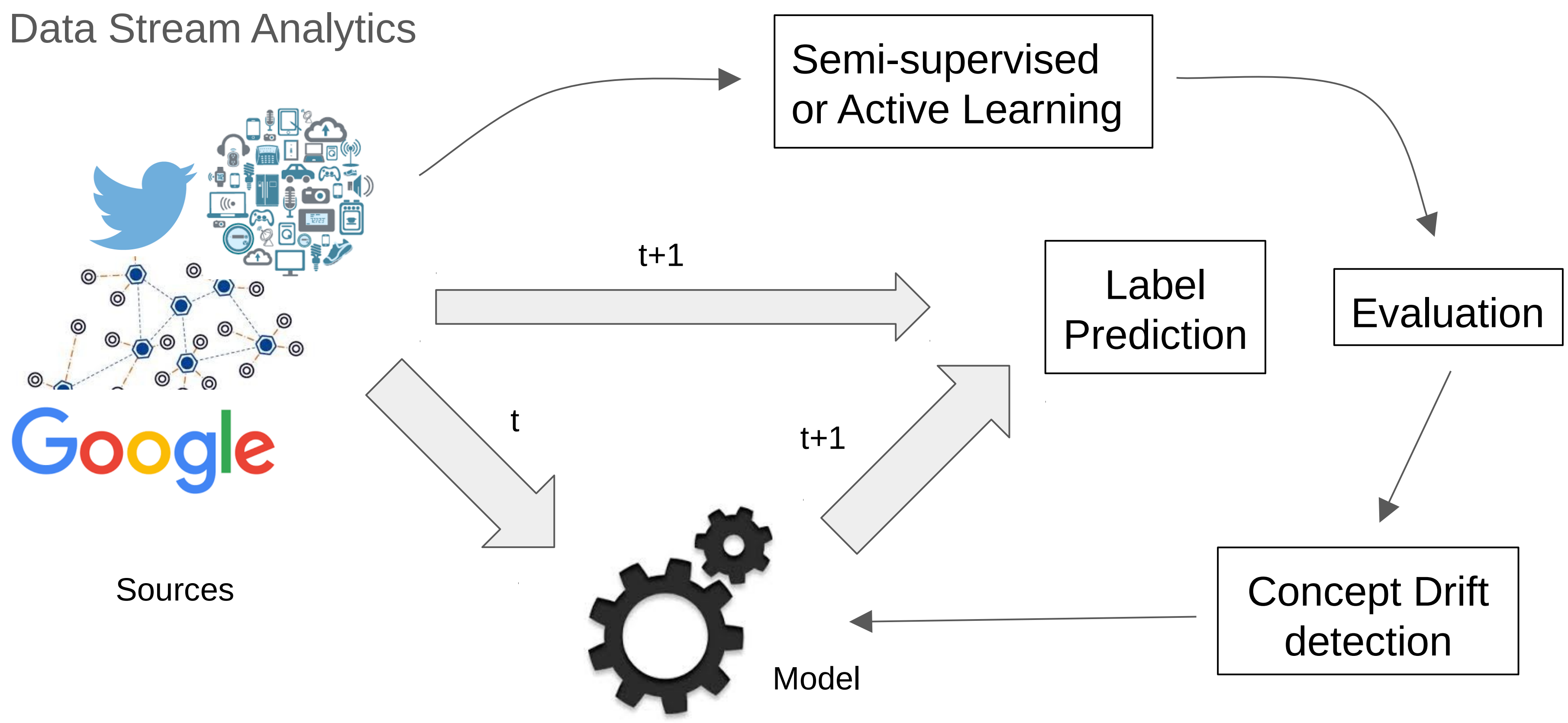
Data Stream Analytics



- 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.

Data Stream Analytics





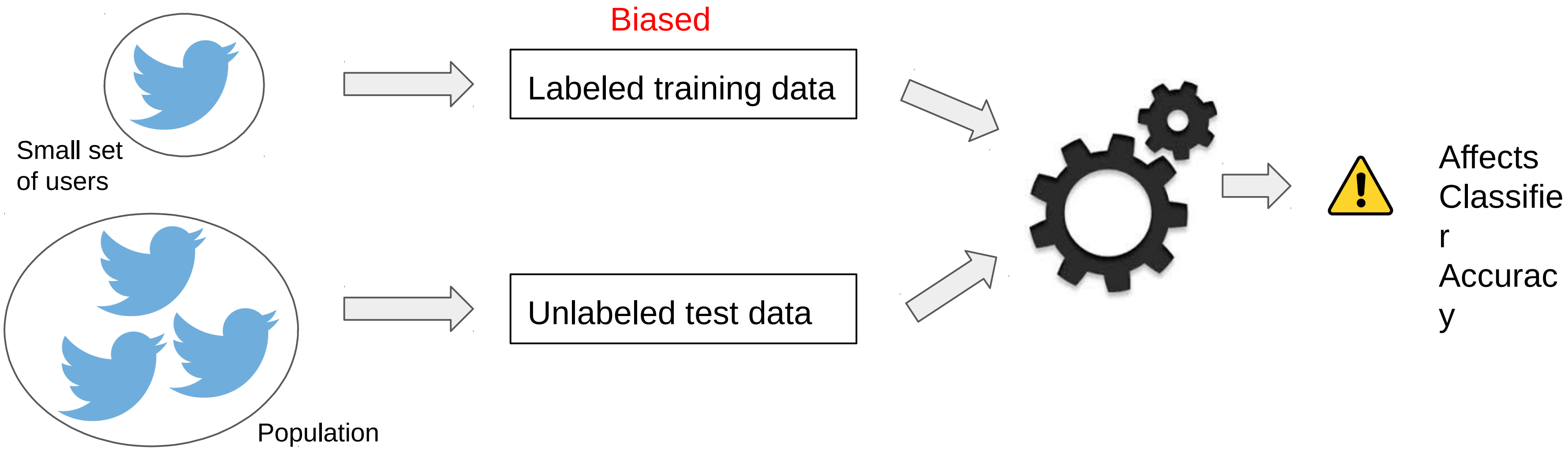
Data Stream Analytics
Semi-supervised or Active Learning
t+1
Label Prediction
Evaluation
t
t+1
Sources
Model
Concept Drift detection

Motivation

What if we do not find a good training set?

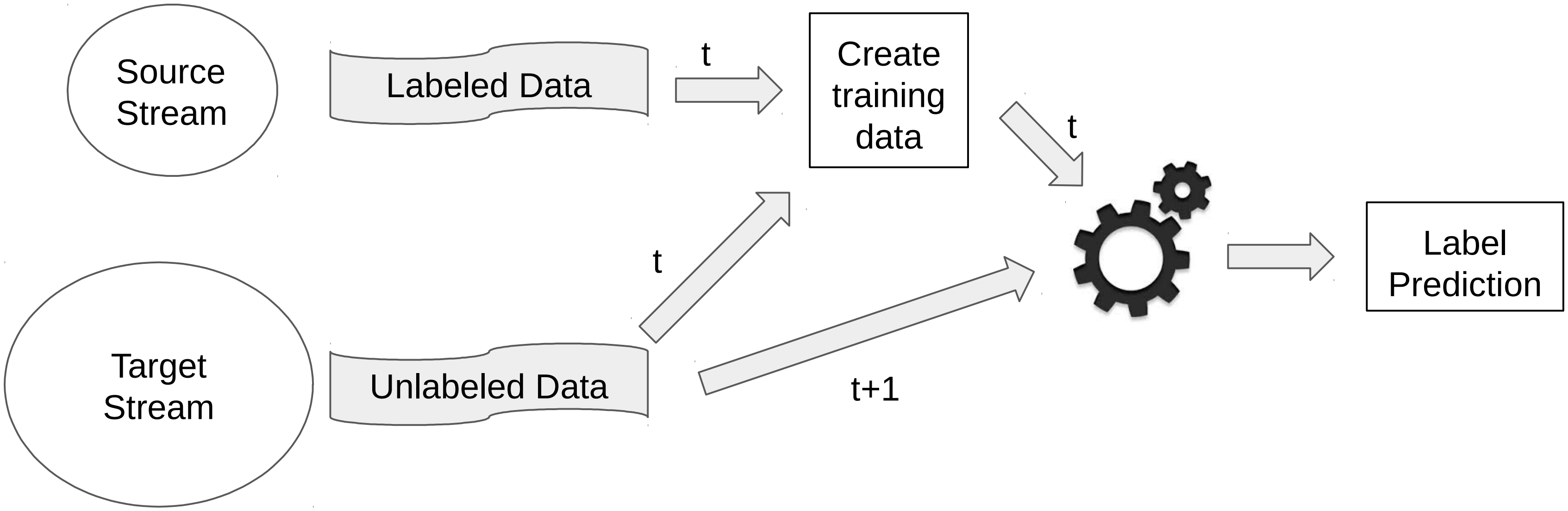
Biased training data selection mechanism.

<u>Example Scenario</u>



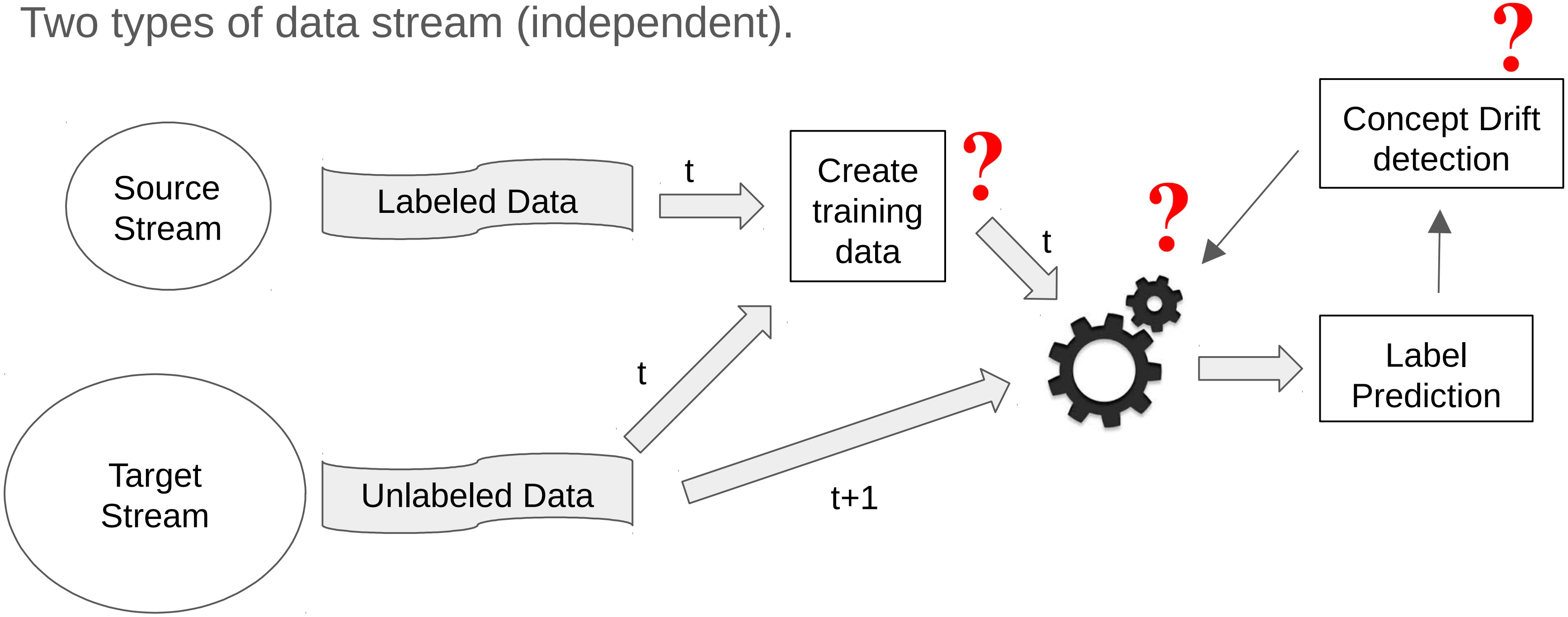
Problem (Multistream Classification)

Two types of data stream (independent).



Problem (Multistream Classification)

Two types of data stream (independent).



Potential Applications

- Domain Adaptation and Transfer Learning over data streams
  - Text Classification
  - Sensor-based location estimation
  - Collaborative filtering

Outline

Challenges

Solution Overview (MSC)

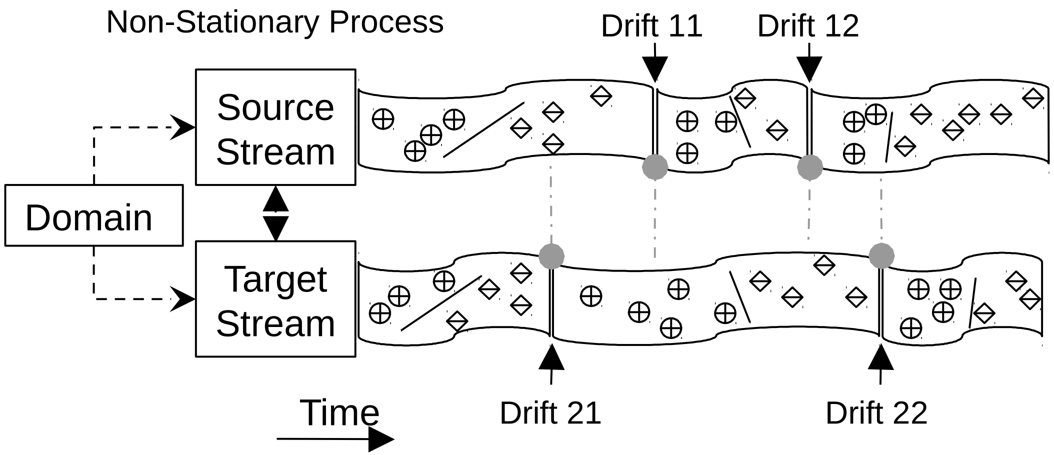
Framework Details

Empirical Evaluation

Conclusion

Challenges

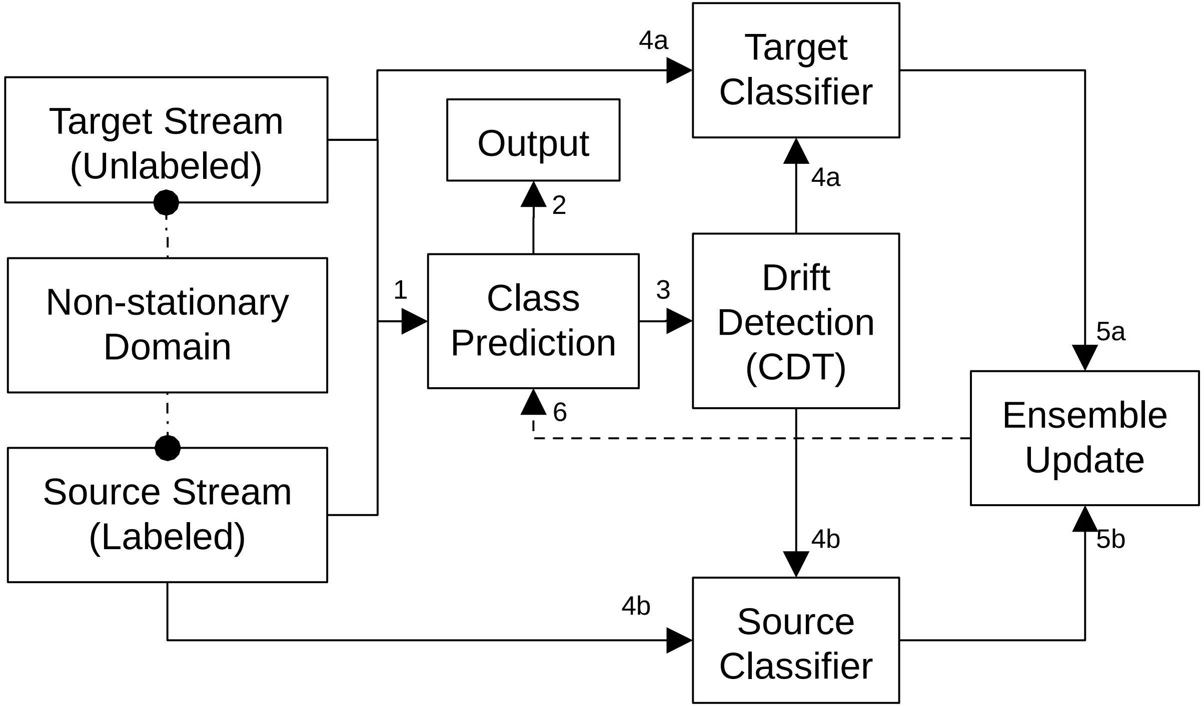
- Leveraging labeled and unlabeled data
  - bias-corrected training set.
- Asynchronous concept drift in source and target stream.
  - Drift detection
  - Drift correction



Challenges

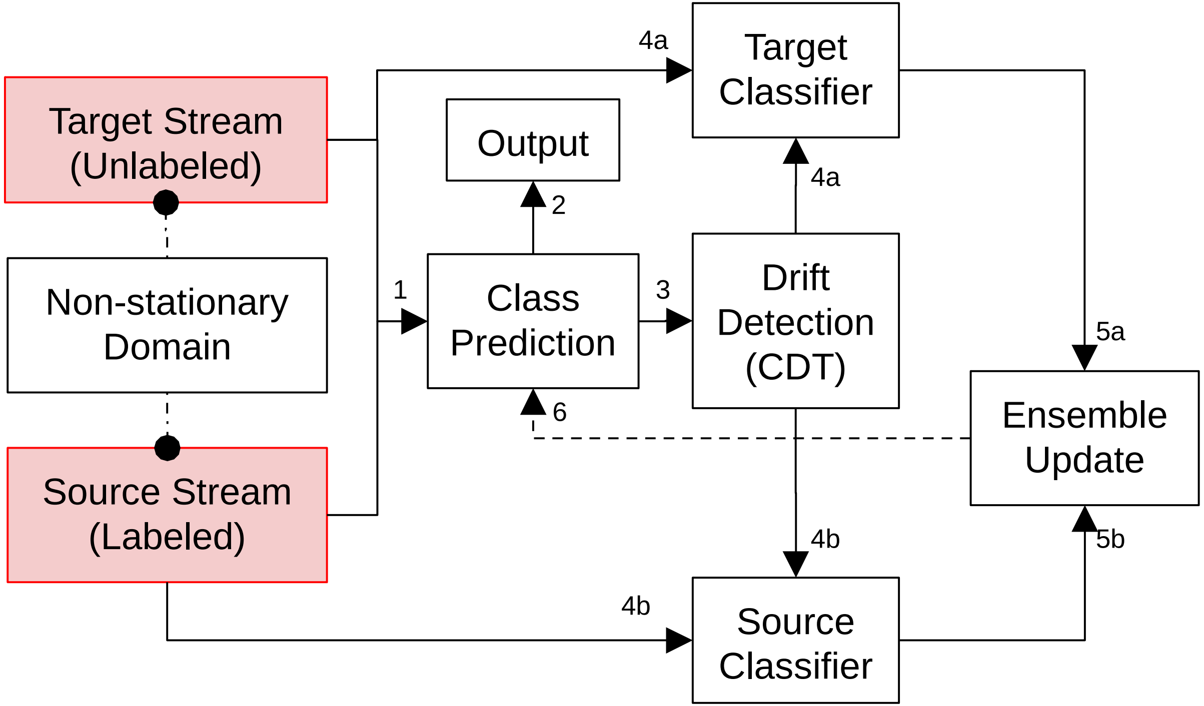
- Can the two streams be combined?
    - Data distributions are different.
    - Combination represent same distribution
    - Separate representation has advantages when multiple sources are present.

Solution Overview



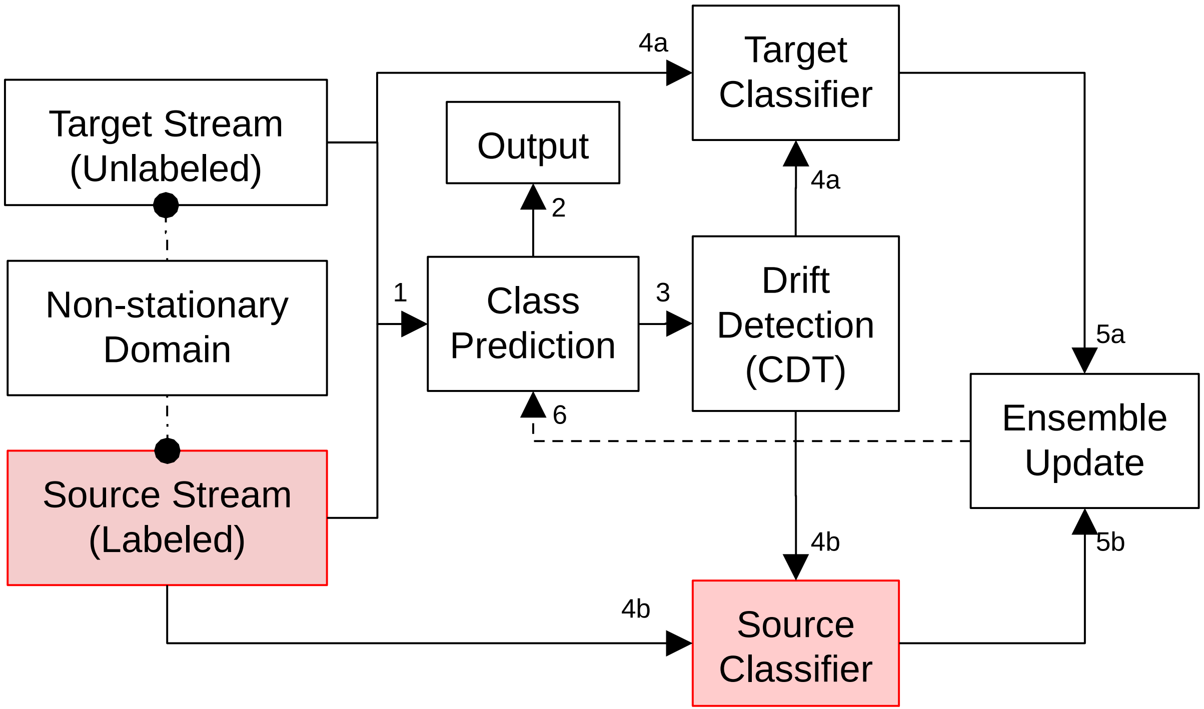
Design Overview

- Two data streams



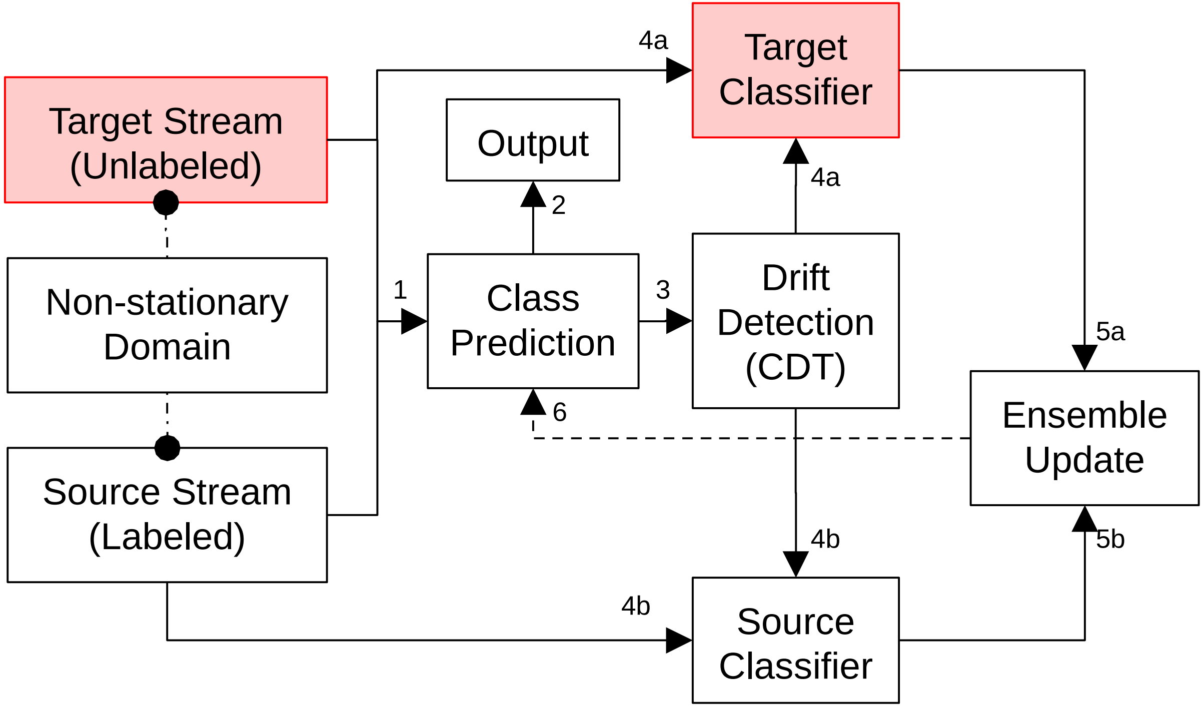
Design Overview

- Two data streams
- To address asynchronous concept drift.



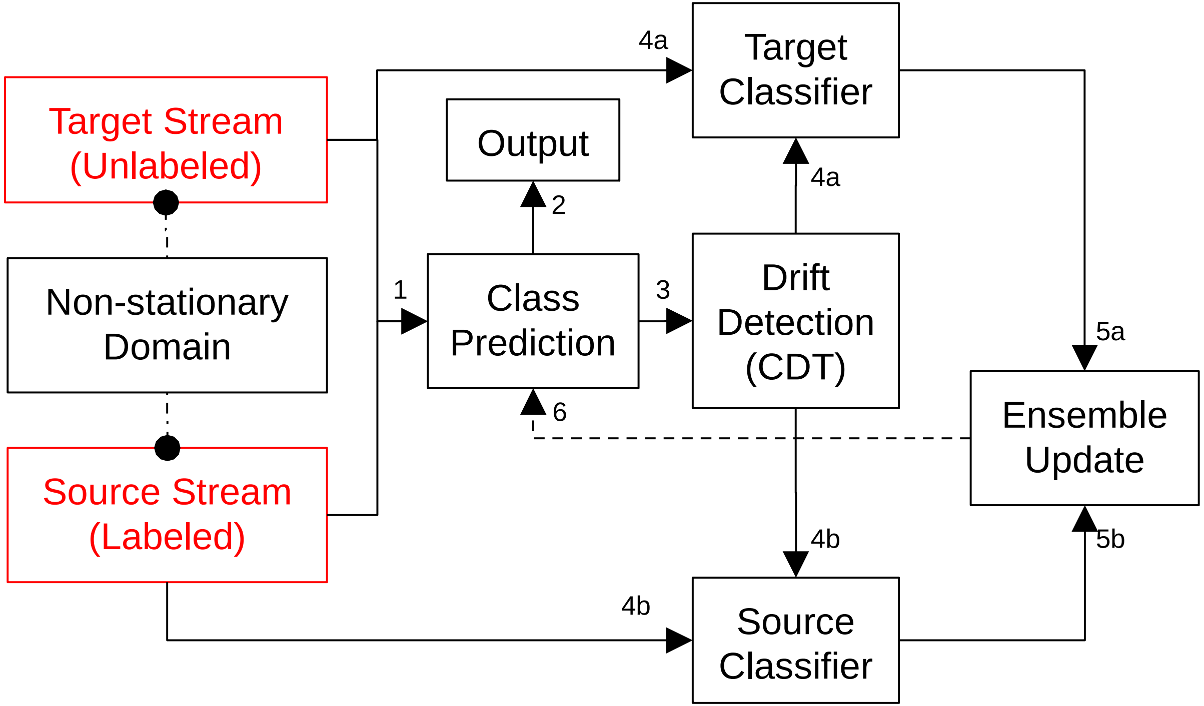
Design Overview

- Two data streams
- To address asynchronous concept drift.



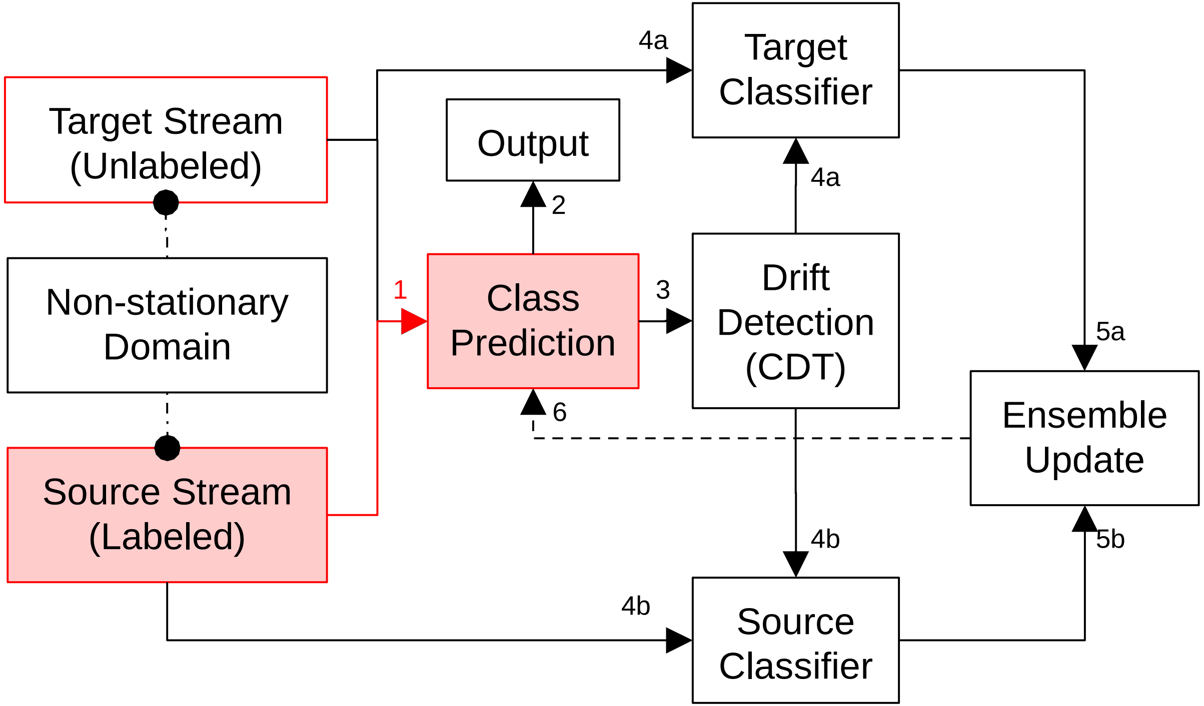
Solution Overview

- Data in source and target occur simultaneously.



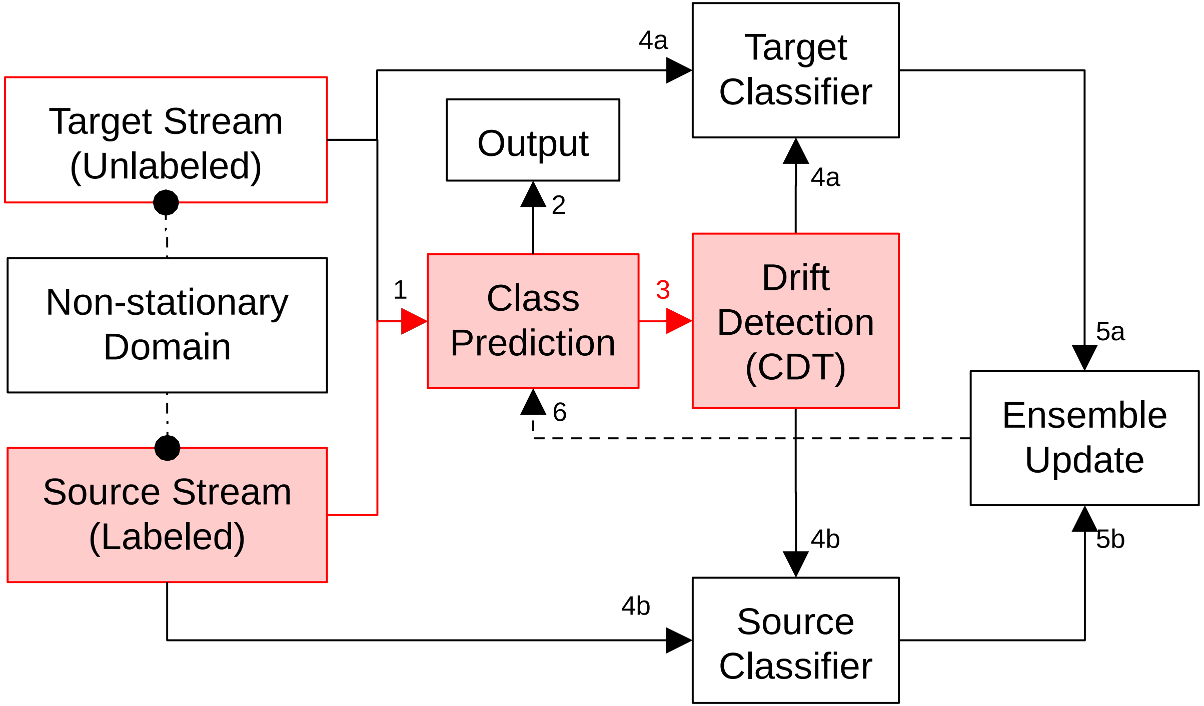
Solution Overview

- Data in source and target occur simultaneously.
- In the case of source data …



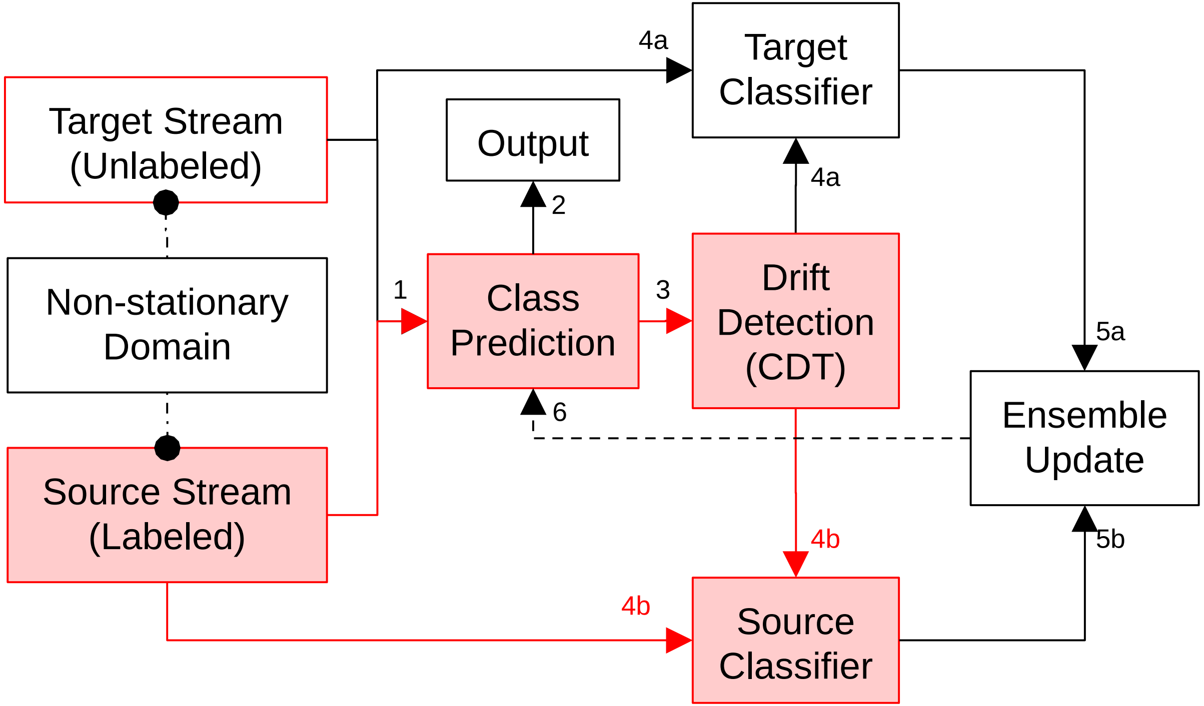
Solution Overview

- Data in source and target occur simultaneously.
- In the case of source data …



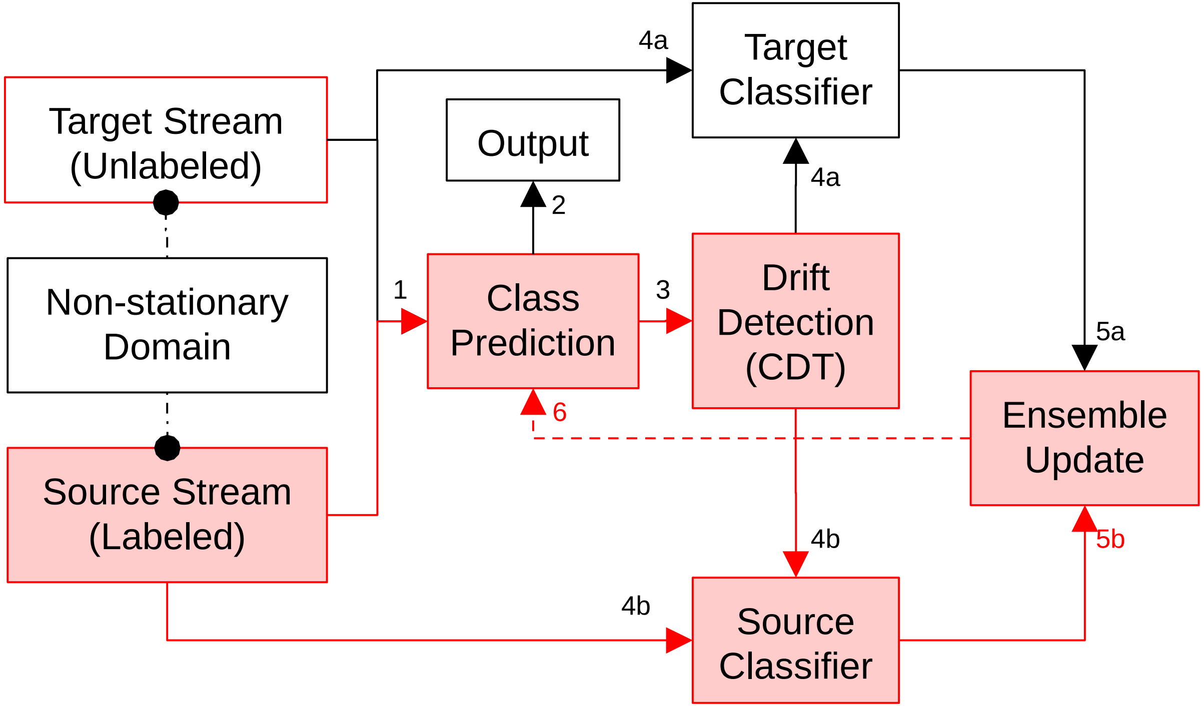
Solution Overview

- Data in source and target occur simultaneously.
- In the case of source data, drift detection output used to update source classifier.



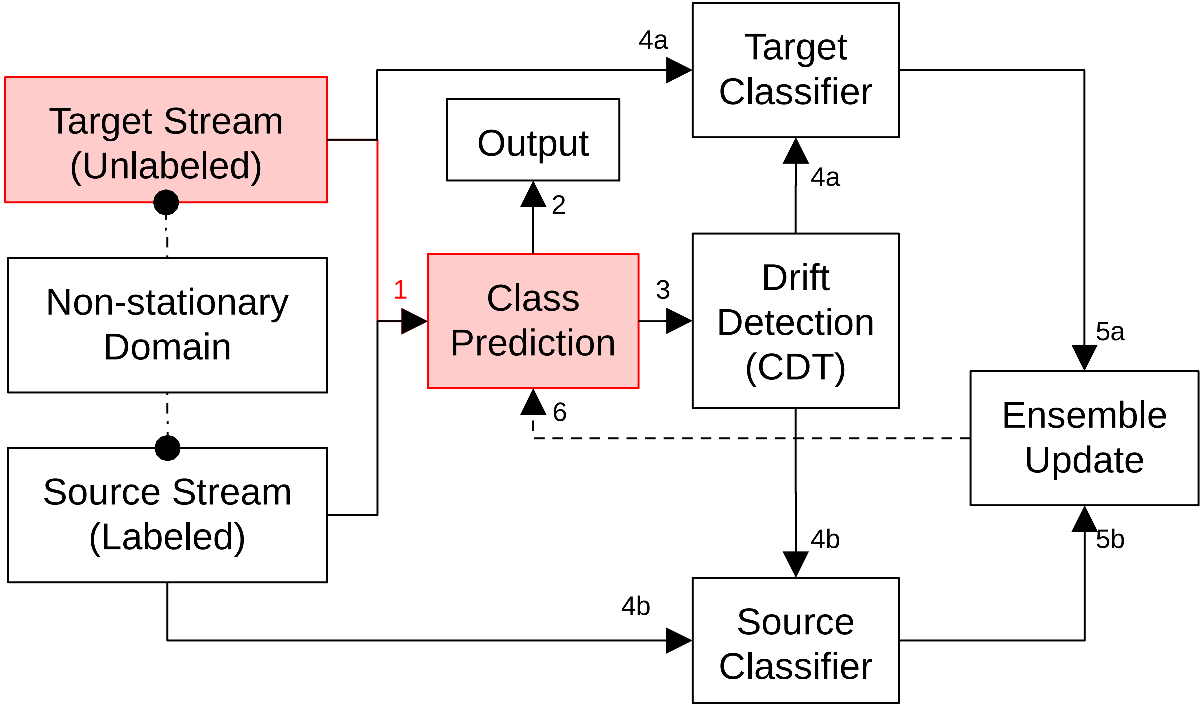
Solution Overview

- Data in source and target occur simultaneously.
- In the case of source data, drift detection output used to update source classifier.



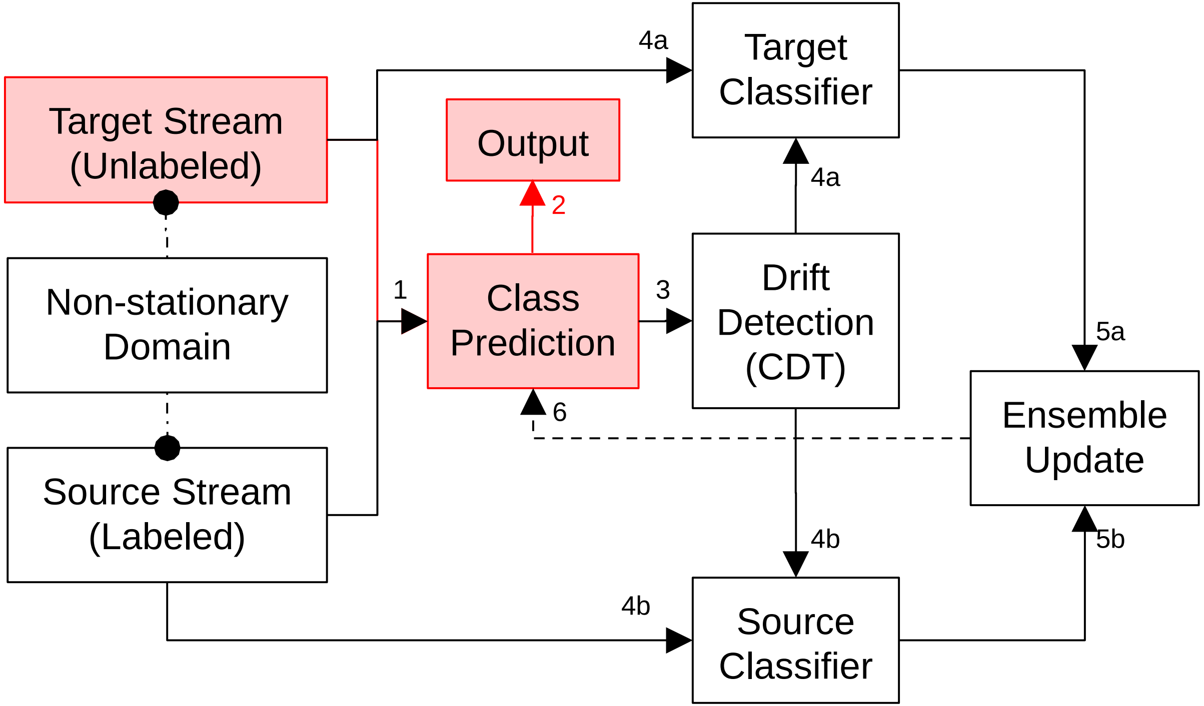
Solution Overview

- Data in source and target occur simultaneously.
- In the case of target data …



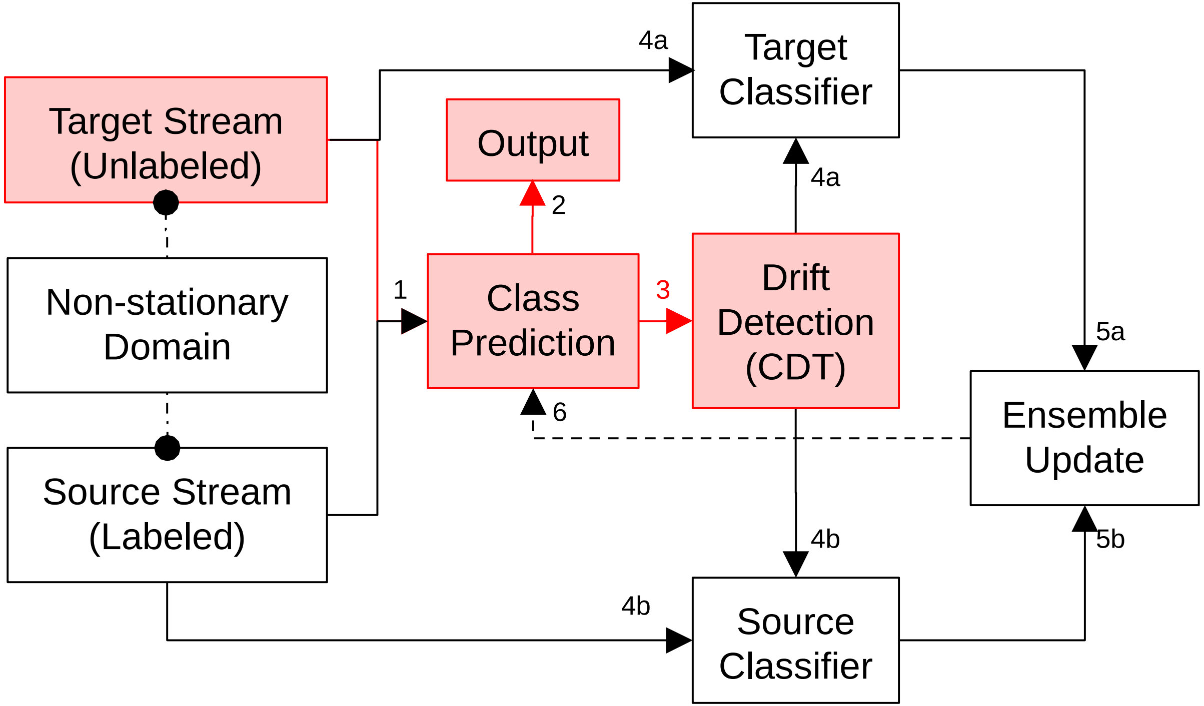
Solution Overview

- Data in source and target occur simultaneously.
- In the case of target data …



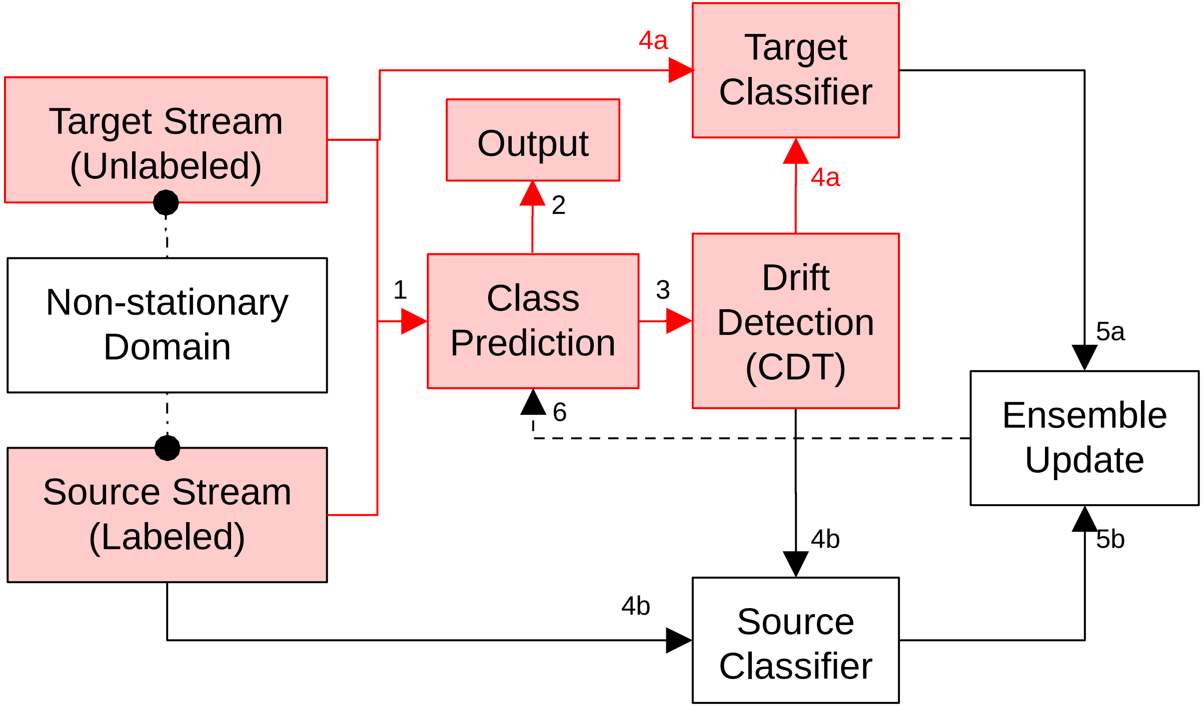
Solution Overview

- Data in source and target occur simultaneously.
- In the case of target data …



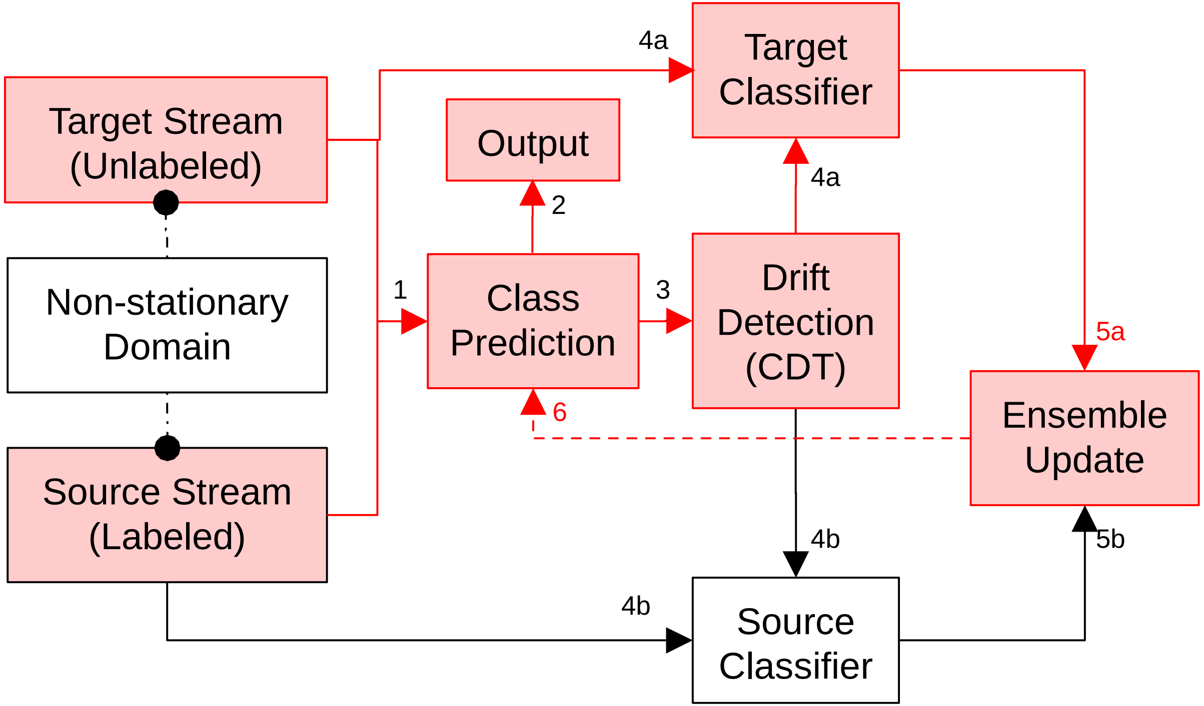
Solution Overview

- Data in source and target occur simultaneously.
- In the case of target data, drift detection output used to update target classifier.
- Target classifier corrects bias between source and target stream at time t.



Solution Overview

- Data in source and target occur simultaneously.
- In the case of target data, drift detection output used to update target classifier.
- Target classifier corrects bias between source and target stream at time t.



Classifier

- Source Classifier
  - Typical classifier using training data from source stream.
  - Predict labels of newly occuring source stream data.
- Target Classifier
  - Bias corrected source stream data for training.
  - Predict labels of newly occuring target stream data.

Target Classifier

- Training : Sampling bias correction via Kernel Mean Matching
  - Minimize mean discrepency between labeled source and unlabeled target distribution.

$$\boldsymbol{\beta}^{(t)*} \approx \underset{\boldsymbol{\beta}^{(t)}}{\text{minimize}} \frac{1}{2}\boldsymbol{\beta}^T \mathbf{K}\boldsymbol{\beta} - \boldsymbol{\kappa}^T \boldsymbol{\beta}$$

$$\text{subject to } \beta_i \in [0, B_{kmm}] \ \& \ \left| \frac{1}{n_{tr}} \sum_{i=1}^{n_{tr}} \beta_i - 1 \right| \leq \epsilon_{kmm}$$

Source data instance weight: $\beta(B_S) = \frac{P_T(B_T)}{P_S(B_S)}$

Matrices of kernel in RKHS: $\begin{cases} K_{ij} = k(x_{tr}^{(i)}, x_{tr}^{(j)}) \\ \kappa_i = \frac{n_{tr}}{n_{te}} \sum_{j=1}^{n_{te}} k(x_{tr}^{(i)}, x_{te}^{(j)}) \end{cases}$

B_S : Source window

B_T : Target window

Label Prediction

- Finite dynamic size window for incoming source and target data.
- Weighted hybrid ensemble
  - Fixed number of classifiers.
  - Contains both source and target classifiers.
  - Source classifier weight based on classifier error.. $w_S : \frac{1}{2}\ln\frac{1-\eta}{\eta}$
  - Target classifier weight based on classifier confidence on unlabeled target data.

Concept Drift Detection

- Source classifier error window
  - Contain binary values.
  - Follow Bernoulli distribution.
- Target classifier confidence window
  - Contain confidence value between 0 and 1.
  - Follow Beta distribution.

CUSUM-type change point detection to detect change point at element *q* of window *W*.

Sequential sub-window

$$\begin{cases} W_h^b = W[1 : q] \\ W_h^a = W[q+1 : n] \end{cases}$$

Likelihood ratio score at point q:

$$s(q,n) = \sum_{i=q+1}^{n} log\left(\frac{P\left(W_h[i] \mid \boldsymbol{\theta}_a\right)}{P\left(W_h[i] \mid \boldsymbol{\theta}_b\right)}\right)$$

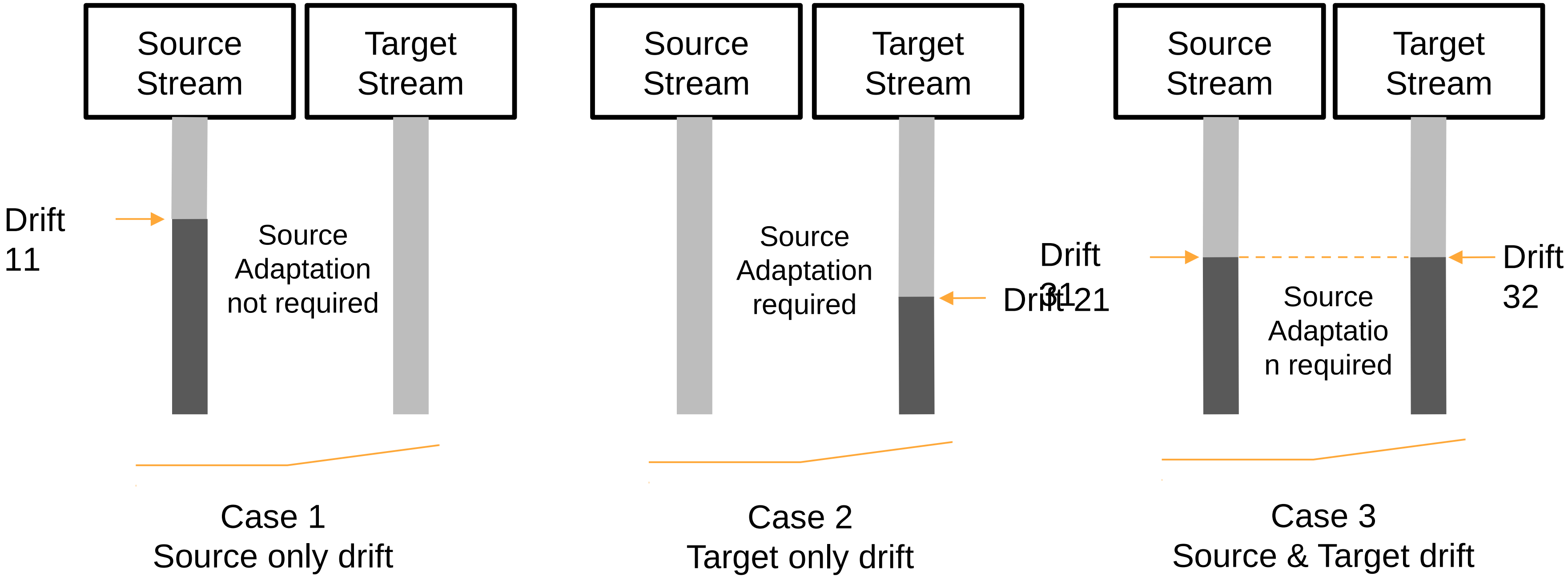
Change point is at *q* if:

$$\omega_n = \max_{\gamma \le q \le n-\gamma} s(q,n) > Threshold$$

Drift Adaptation

- Why not train both types of classifiers once a drift is detected on either stream?

- Sampling bias correction if target stream has a concept drift.



Empirical Evaluation

Real World

Synthetic

Dataset	# features	# classes	# instances
ForestCover	53	7	146,438
Sensor	5	58	150,000
SEA	3	3	58,000
SynEDC	40	20	98,816
SynRBF@002	70	7	98,000
SynRBF@003	70	7	98,686

Divide dataset into Source and Target Stream, with bias in source stream data selection according to: $e^{-|x-\bar{x}|^2}$

Empirical Evaluation

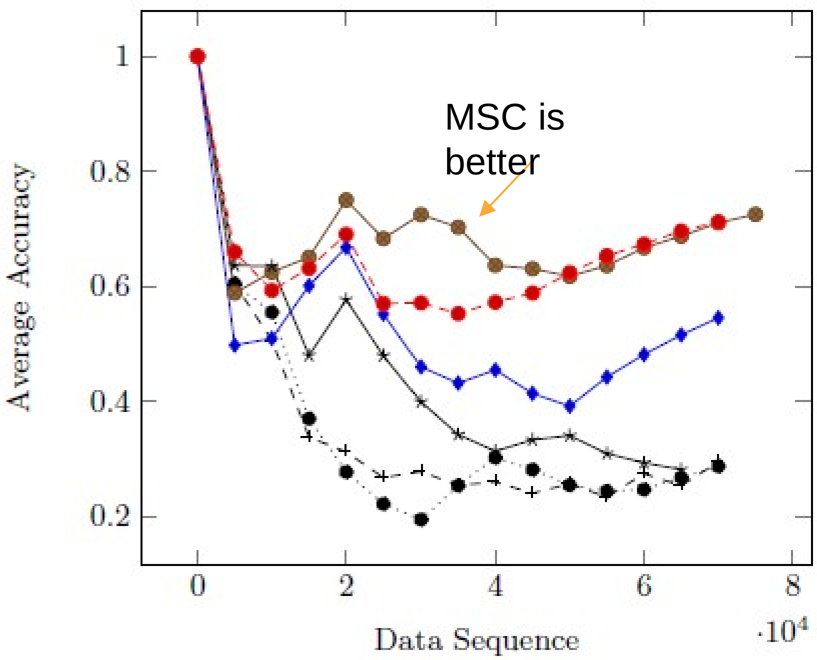
- SVM as base classifier
  - Source Classifier : Typical multiclass SVM.
  - Target Classifier : Weighted SVM
- Classifier confidence:
  - Distance of test data to hyperplane.

Empirical Evaluation

	Symbols	Description
Baseline	sKMM	Single target classifier without update.
	mKMM-5k	Single target classifier with update every 5k instances.
	srcMSC	CPD with source classifier only. No bias correction.
	trgMSC	CPD with target classifier only. No source drift adaptation.
Variants	MSC	Proposed method with hybrid ensemble.
	MSC2	Proposed method with separate source and target ensemble.

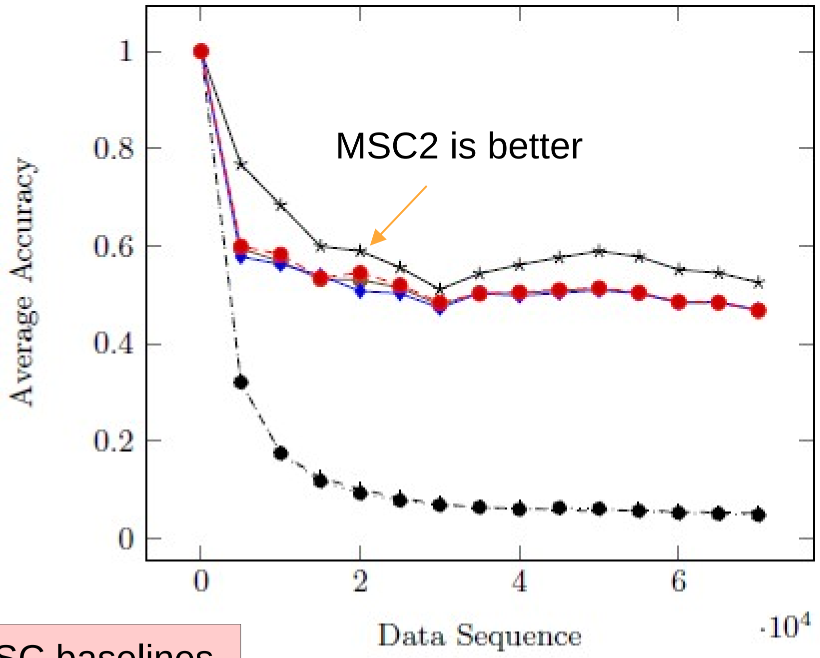
Results

sKMM; *mKMM-5k*; *MSC*; *MSC2*; *srcMSC*; *trgMSC*.



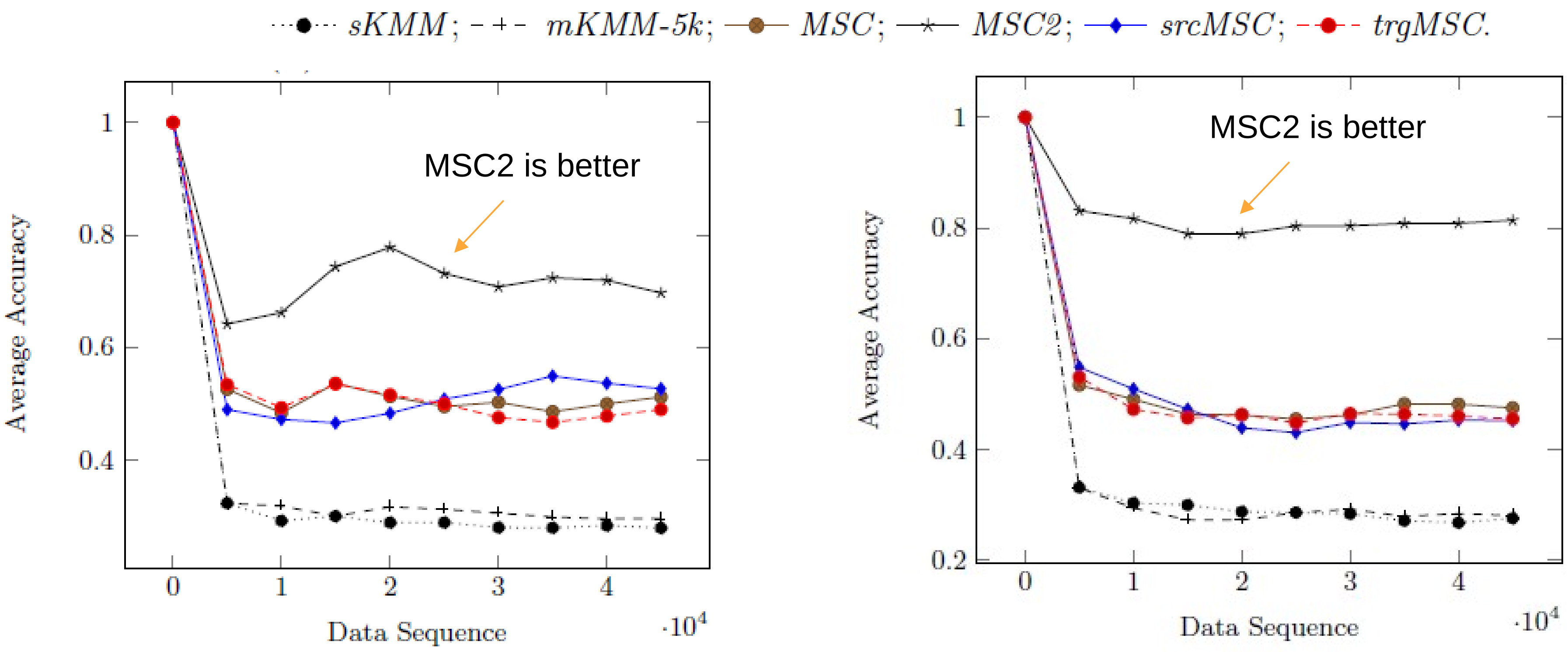
ForestCover Dataset

MSC baselines also good, but ..



Sensor Dataset

Results



SynRBF@002 Dataset

SynRBF@003 Dataset

Conclusion

- Introduce a new data stream mining setting with bias labeled data
- Propose a framework to address new challenges of concept drift in this setting.
- Empirical results achieve significantly better accuracy than baseline.

- Future work: Multi-source setting and Semi-supervised target stream classification.

Thank you

Q & A