Introduction to Data Shift & Concept Drift

Mehdi Ataei – Vector Institute & University of Toronto

Ali Pesaranghader – CIBC Data Science & AI Research

March 26, 2021

VECTOR INSTITUTE

CIBC Banking that fits your life.

Outline

- **Chapter I: Introduction**
  - What is Data Shift?
  - Why Data Shift Happens?
  - Why Handle Data Shift?
    - What are the consequences of not properly addressing data shift?
  - How to approach data shift?
- **Chapter II: Data Shift**
  - Data Shift Types and Patterns
  - Data Shift Detection & Correction
  - Transfer and Active Learning
  - Evaluation and Discussion
  - Packages
- **Chapter III: Discussion and Q&A**

Introduction

VECTOR INSTITUTE

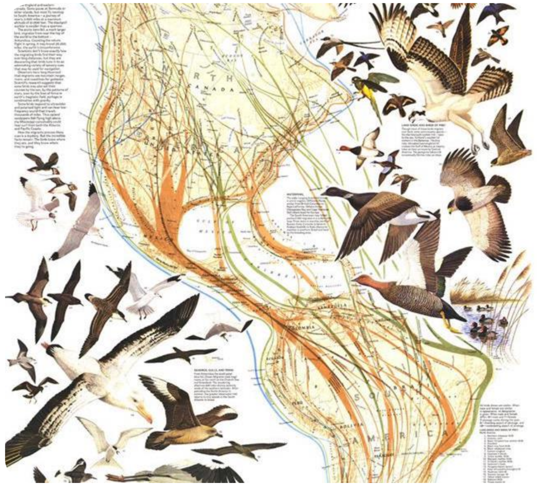
CIBC Banking that fits your life.

Definition
What is Data Shift?

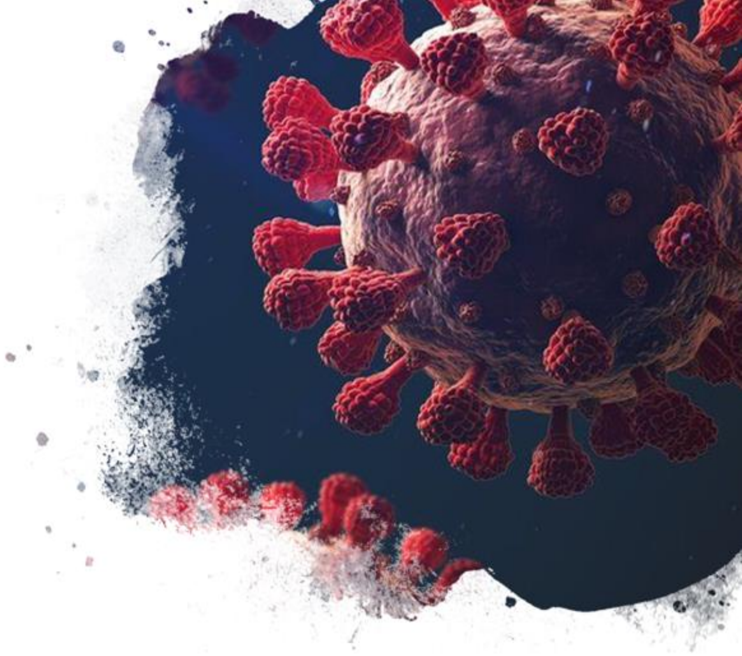
- In classic Machine learning, models are trained under the premise that the training and the real-world (i.e., both source and target) data are from the same distribution
- Such assumption may potentially result in predictive problems in dynamic industries and environments where the distribution of data changes over time
- The existence of such a difference between the dataset distributions is called as dataset shift in the machine learning community.
- In fact, most real-world applications should cope with some form of shift as the distribution of the data used to train a model differs from the distribution of the data that the model encounters after its deployment.

Why Data Shift Happens?

- **Reasons for experiences data shift could be:**
  - Political
  - Industrial (Solar & Clean Energy)
  - Financial (Shift from Fiat to Crypto…!)
  - Retail (supply and demand)
  - Pandemic, e.g., COVID 19 and SARS
  - War & Immigration
  - House Price (2007)
  - IT & Internet (e.g., Dot-com)
  - Security & Privacy (Cyber attack)
  - Environmental (Global warming, weather)
  - Natural (Bird Migrations)
  - Dynamic nature (Smart Houses)
  - Unexpected

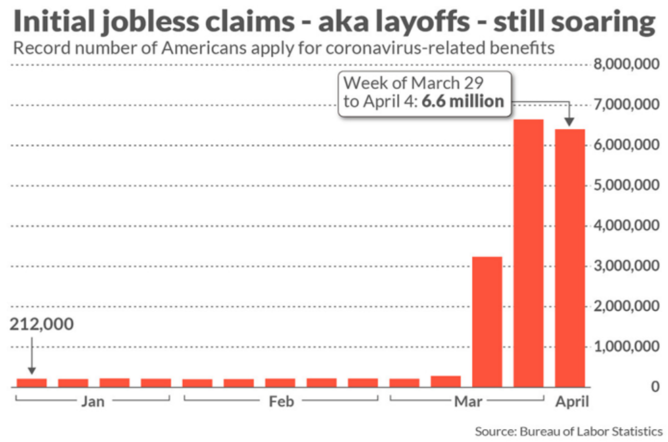


Motivation
Why Handle Data Shift?



- We train a model to predict "Jobless Claims" in the US
- In practice, the model correctly predicts 200k claims for the fourth week of March 2020
- COVID 19 hits and unemployment claims skyrocket
- The model now predicts 800k claims; it's higher than normal, but is it reliable? (In reality, it is over 6 millions…)

- After pandemic, the model parameters (e.g., layoffs, closures, mobility, data, etc.) may have dramatically changed
- The model trained on historical jobless claims data may underestimate the effect of the pandemic



Motivation
Why Handle Data Shift?

- **Pre-COVID:**
  - Trained a model to predict bankruptcy of an entity
  - Data:
    - 70% Non-bankrupted,
    - 20% Likely to Bankrupt,
    - 10% Will Bankrupt
  - The model had an accuracy of 80%
    (vs. 33% for a random classifier)

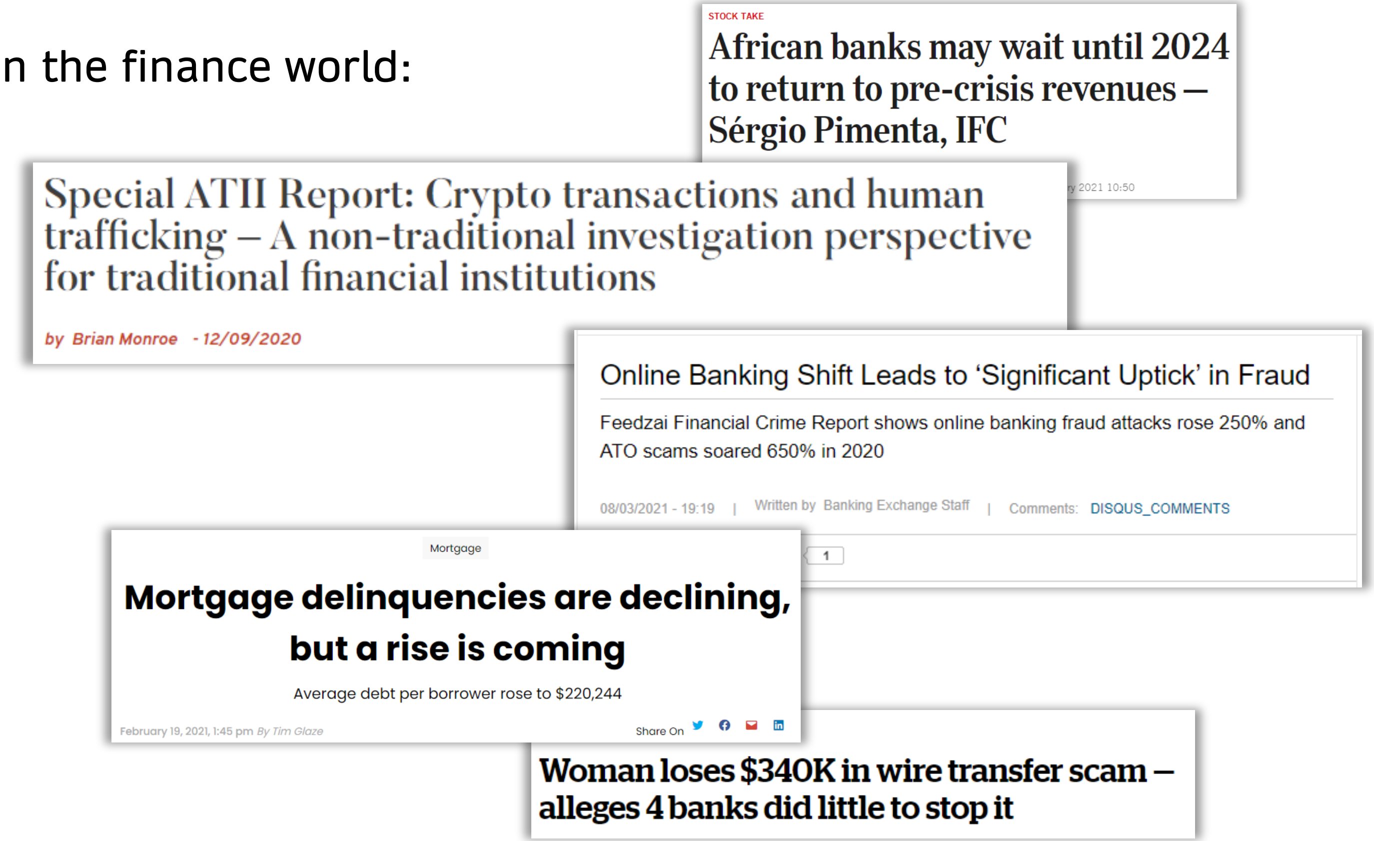
- **Post-COVID:**
  - After COVID, we observed:
    - 40% Non-bankrupted,
    - 60% bankrupted
    - Class distribution changed
    - The number of classes reduced
  - The model accuracy worsens although the problem is easier
    (a random classifier is now 50% accurate)

Motivation – Cont. Why Handle Data Shift?

- **Consequences of not handling data shift could be loss of:**
  - Lives
  - Clients
  - Resources
  - Funds
  - Time
  - Trust
  - Reputation

Motivation – Cont.
Why Handle Data Shift?

- Consequences of not handling data shift in the finance world:
  - **Financial crimes:**
    - Terrorist financing
    - Money laundering
    - Fraudulent transactions
    - Scamming
    - Slavery and human trafficking
  - **Client focused:**
    - Inadequate financial plans
    - Poor product recommendation
    - Mortgage delinquency
    - Credit and loan defaults
    - Attrition & losing clients
    - etc.


STOCK TAKE
African banks may wait until 2024 to return to pre-crisis revenues – Sérgio Pimenta, IFC

Special ATII Report: Crypto transactions and human trafficking – A non-traditional investigation perspective for traditional financial institutions
by Brian Monroe - 12/09/2020

Online Banking Shift Leads to 'Significant Uptick' in Fraud
Feedzai Financial Crime Report shows online banking fraud attacks rose 250% and ATO scams soared 650% in 2020
08/03/2021 - 19:19 | Written by Banking Exchange Staff | Comments: DISQUS_COMMENTS

Mortgage
Mortgage delinquencies are declining, but a rise is coming
Average debt per borrower rose to $220,244
February 19, 2021, 1:45 pm By Tim Glaze
Share On

Woman loses $340K in wire transfer scam – alleges 4 banks did little to stop it

How to approach data shift?

Reactive

- **React once something happened**
  - Transfer learning (reusing old models)
  - Adaptive learning (efficiently retraining)
  - Statistical correction
- **It is easier but more risky**

Proactive

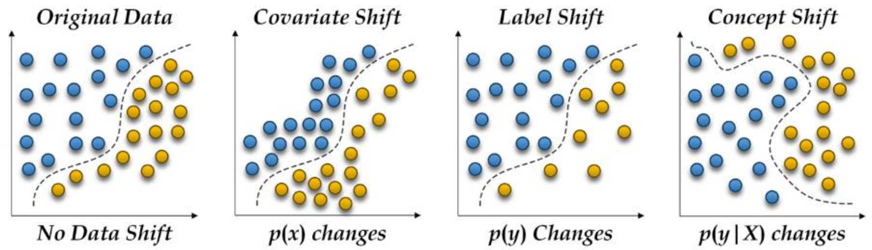
- **React before something happens**
  - Adding mitigation steps in ML pipelines
  - Using historical data to retrain the model (using COVID data for future pandemics)
  - Using synthetic datasets to estimate data shift e.g., adversarial training
  - It's not perfect. Must be an oracle.
- **It is harder but less risky**

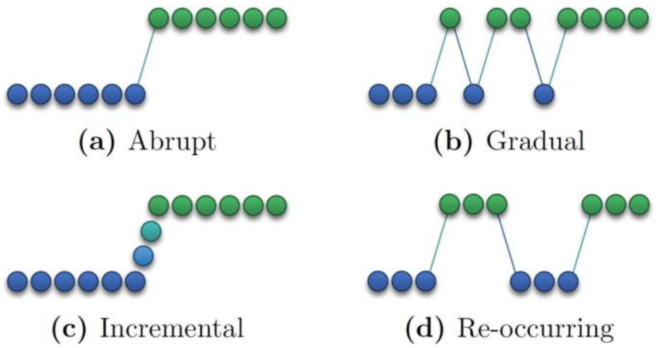
Data Shift

Data Shift

- **Common types of data shift:**
  - Covariate shift
  - Label shift
  - Concept shift

- **Data shift patterns:**
  - Abrupt
  - Gradual
  - Incremental
  - Re-occurring (or recurring)





Data Shift
Covariate Shift

Covariate Shift

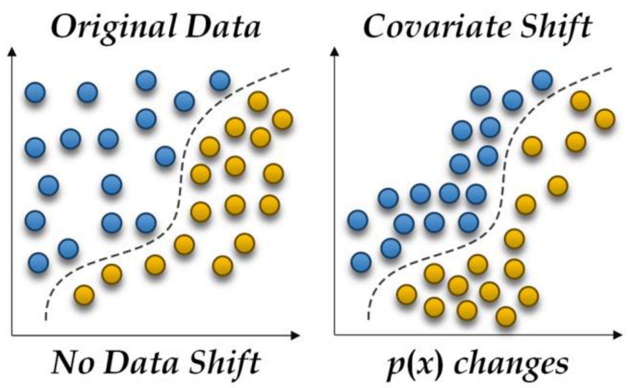
- Covariate shift happens when the conditional distribution $P_S(y|x)$ remains the same, i.e., that conditional distribution of the source and target domains are equal, but $P_S(x)$ changes. So, we have:

$$P_S(x)P_S(y|x) \neq P_T(x)P_T(y|x)$$

  where

$$P_S(y|x) = P_T(y|x)$$

- Covariate shift appears in data due to lack of randomness, inadequate sampling, biased sampling, and non-stationary environment.



Data Shift
Label Shift

Label Shift

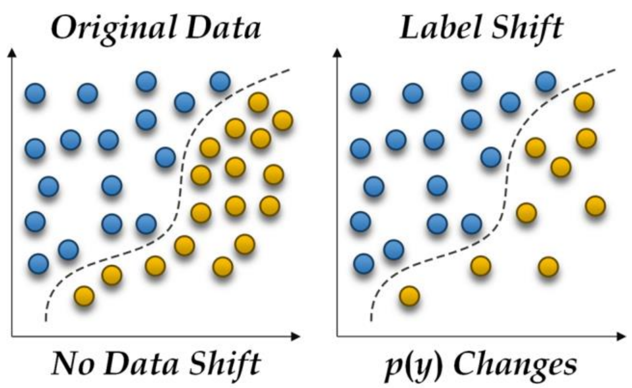
- Label shift is experienced when the conditional distribution $P_S(x|y)$ remains the same but $P_S(y)$ changes. So, we have:

$$P_S(y)P_S(x|y) \neq P_T(y)P_T(x|y)$$

where

$$P_S(x|y) = P_T(x|y)$$

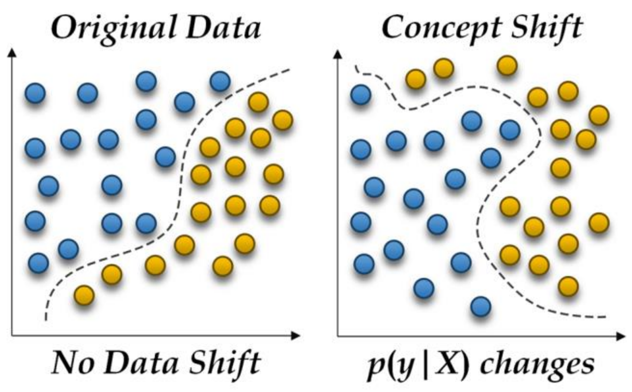
- Having $P_S(y) \neq P_T(y)$ implies that label shift happens when some concepts are undersampled or oversampled in the target domain compared to the source domain.



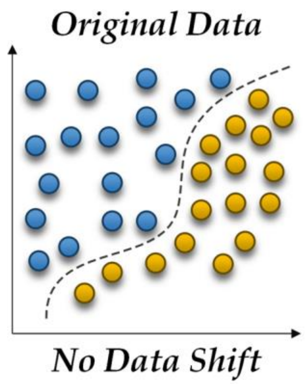
Data Shift
Concept Shift

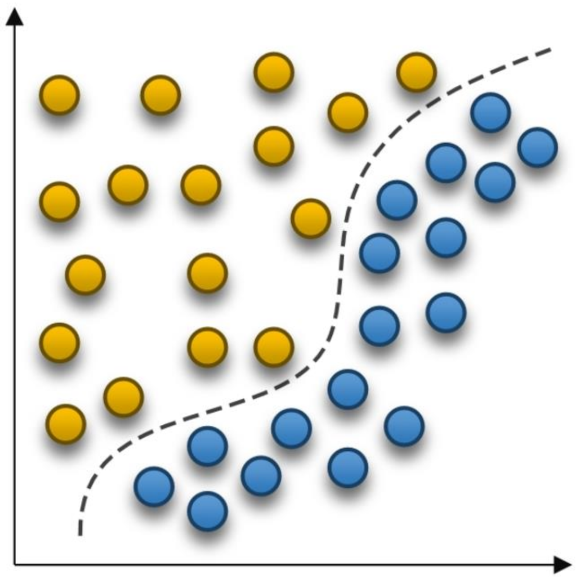
Concept Shift

- In the case of concept shift, $P_S(y)$ and $P_T(y)$ follow the same distribution but $P_S(y|x)$ differs from $P_T(y|x)$.
- To address concept shift, we adapt our model globally or locally.
- Global adaptation is training our model from scratch using the target data whereas local adaptation works for learning algorithms that can be refitted for some part of their decision regions; for example consider decision trees where we may update some branches to reflect the change in the real world.
- Concept shift detectors compare the performance of a learner against both the source and target data; and if there is a significant difference they alarm for a drift.



Quiz





What kind of shift is this?

Data Shift Detection

Covariate Shift
Domain Classifier

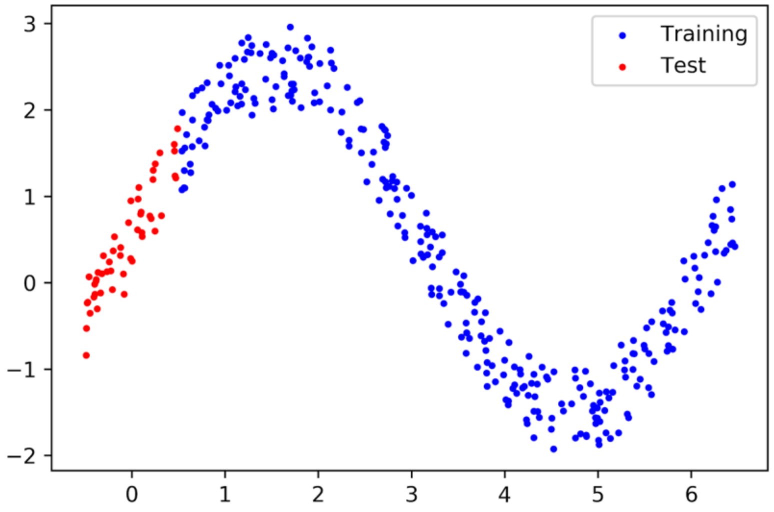
Covariate Shift Detection

Domain Classifier

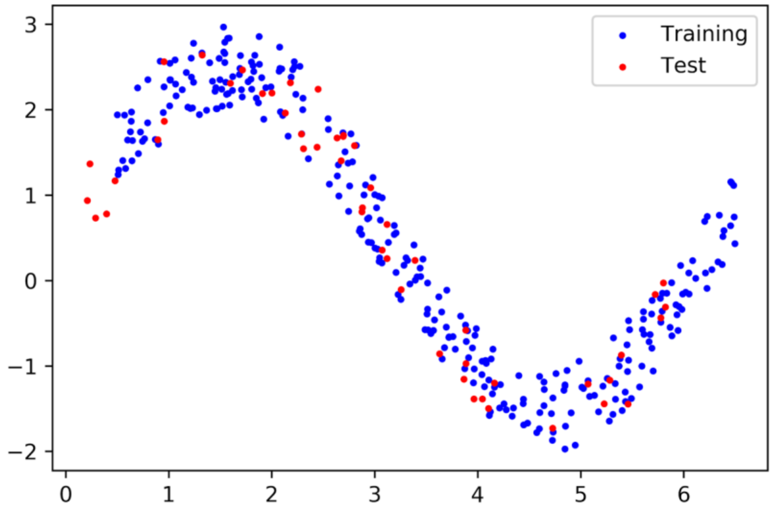
- Train a **domain classifier** to detect whether new data is from P_S or P_T
- That means we want to see if a data point is from source or target domain
- Domain classifier reduces dataset dimension to a **single dimension**, which specifically discriminate between source and target data
- The higher the error of the classifier $\longrightarrow$ the closer the distributions (i.e., unlikely to observe covariate shift)
- Applicable to high-dimensional data
- Can detect what feature(s) caused the shift using feature importance analysis
- Offline

Covariate Shift – Cont.
Domain Classifier

- ROC-AUC score can be used to check if the performance of the classifier is statistically better than random chance (i.e., ROC-AUC score of 0.5)
- ROC-AUC score larger than 0.8 can be considered major shift
- Bi-nominal testing can be used as well



ROC-AUC = 0.91



ROC-AUC = 0.52

Covariate Shift & Domain Classifier – Cont.
Important considerations

Considerations

- Requires training a classifier
- Requires access to large samples from $x_i \sim P_T$ and may perform poorly with small samples
- Choosing a classifier to distinguish between two distributions at high level is equivalent to picking a measure between distributions distances
- The choice of the classifier may yield very different results
- To improve the shift detection confidence, one may consider using multiple classifiers and aggregate their predictions in some manner

Label Shift
Black Box Shift Detection

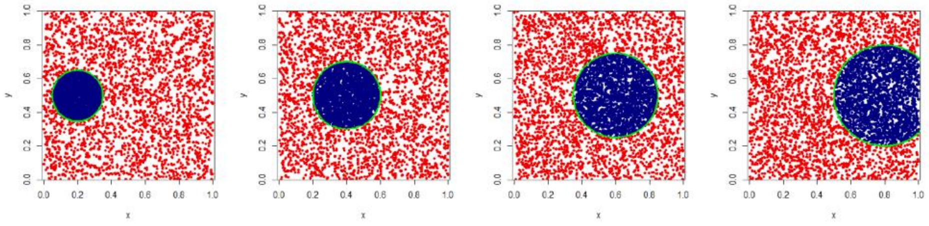
Label Shift Detection

- Detecting label shift is harder as we don't have access to labelled distribution
- Solution: Estimate it using a **pre-trained** classifier (Let's call it as *label classifier*)
- The label classifier must have an invertible confusion matrix
  - This condition is easily satisfied if the classifier is well-trained
- **Black Box Shift Detection**: Given a pre-trained label classifier $f(x)$ with invertible confusion matrix, detecting that the source distribution P_S is different from the target distribution P_T only requires detecting that $P_S(f(x)) \neq P_T(f(x))$

Concept Drift Detection

Concept Drift Detection

- **Idea**: Use probabilistic or statistical methods to bound the difference between $P_S(y|x)$ and $P_T(y|x)$ - a significant difference suggests concept shift.
- **In practice**: The *performance* of a learner is monitored; if it dropped significantly below a threshold, statistically bounded, system triggers for a drift
- Drift detection methods are categorized into three groups:
  - **Sequential Analysis based Methods**:
    Cumulative Sum (CUSUM),
    PageHinkley (PH)
  - **Statistical based Methods**:
    Drift Detection Method (DDM),
    Early Drift Detection Method (EDDM),
    Reactive Drift Detection Method (RDDM)
  - **Window based Methods**:
    Adaptive Windowing (ADWIN),
    SeqDrift detectors,
    Non-parametric Methods; e.g., HDDM, FHDDM, and MDDM



(a) Context 1 (b) Context 2 (c) Context 3 (d) Context 4

https://ruor.uottawa.ca/bitstream/10393/38190/1/Pesaranghader_Ali_2018_thesis.pdf

Concept Drift Detection – Cont.
FHDDM

The **FHDDM** algorithm slides a window with a size of n on the prediction results:

- It inserts a 1 into the window if the prediction result is *correct*, and 0, otherwise.

FHDDM updates two registers, while the predictions are processed:

- μ^t**:** the mean of elements in the window at time t.
- μ^m**:** the maximum mean observed so far.

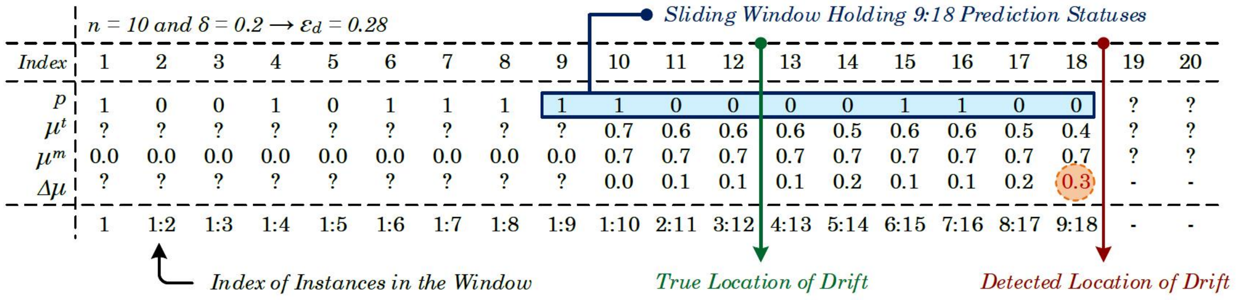
Considering the **PAC** learning model:

- μ^m should increase or remains steady as we process instances.
- ***Or*** the possibility of facing a concept drift increases if μ^m does not change and μ^t decreases over time

Eventually, a significant difference between μ^m and μ^t indicates the occurrence of a drift in the stream

In a streaming setting, assume μ^t is the mean of a sequence of n random entries, each in {0, 1}, at time t, and μ^m is the maximum mean observed so far.
Let $\Delta\mu = \mu^m - \mu^t \geq 0$ be the difference between the two mean. Then, given δ, i.e., the probability of error allowed, Hoeffding's inequality guarantees a drift has happened if $\Delta\mu \geq \varepsilon_d$, where:

$$\varepsilon_d = \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}$$



Data shift Correction

VECTOR INSTITUTE

Banking that fits your life.

Sample Re-weighting

Sample Re-weighting

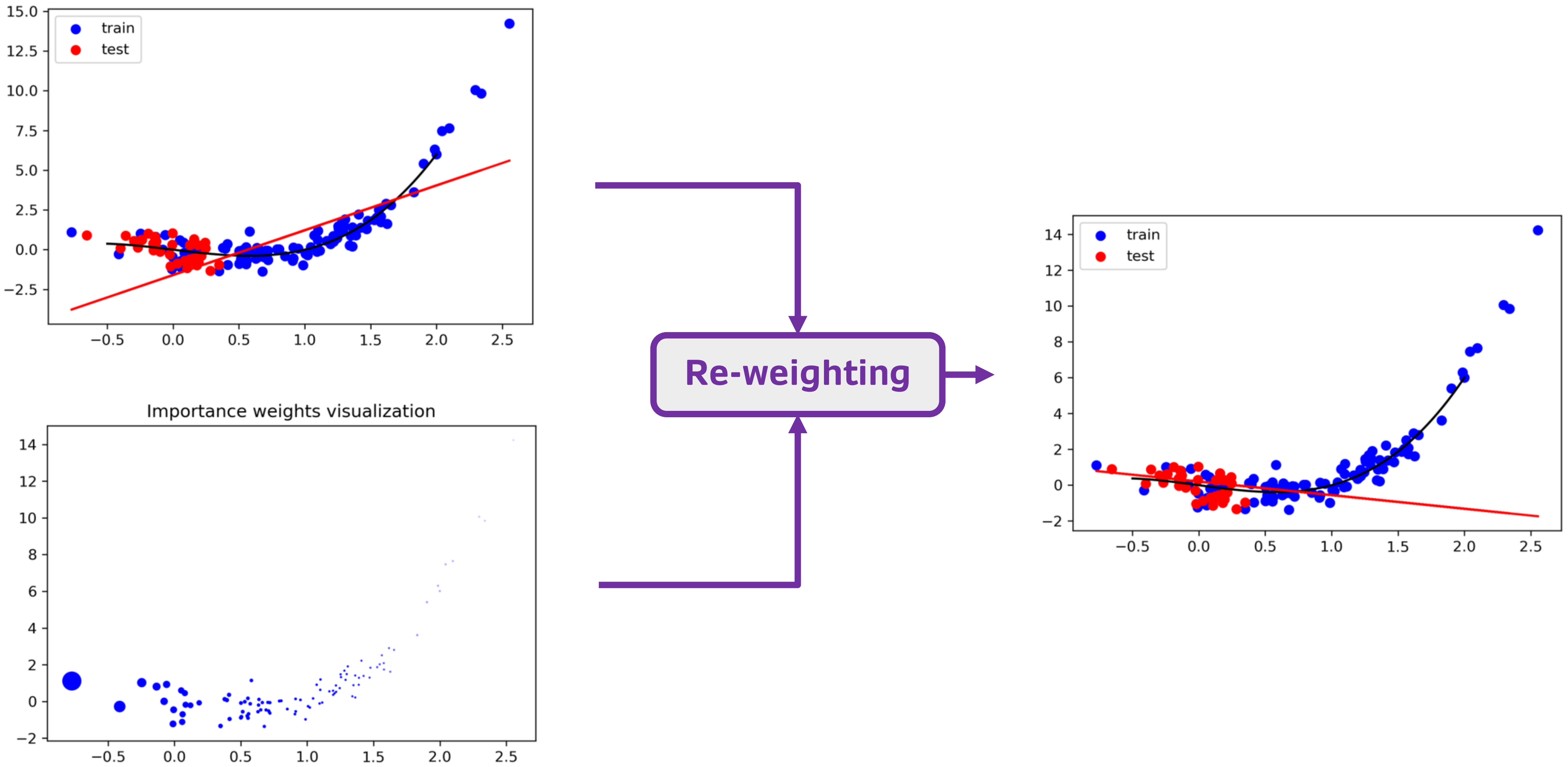
- Idea: Re-weight each data point by the ratios of the probabilities:

$$\beta_i \equiv \frac{P_T(x)}{P_S(x)}$$

- In order to calculate β_i, we need to estimate the distribution ratios
- One of the ways to estimate β is to train a classifier to distinguish between the training and test sets
- If the training and test data is drawn from the same distribution, the classifier would not be able to distinguish between them (equal likelihood that a sample is drawn from either one of the distributions)
- We hope that the classifier can find a useful re-weighting factor
- **NOTE**: The classifier may fail to detect a dataset shift (false negative)

$$\beta = \frac{p(z = -1|x)}{p(z = 1|x)}$$

Sample Re-weighting Example



Label Shift Correction

Label Shift Correction

- Idea: Re-weight each class using the ratio below:

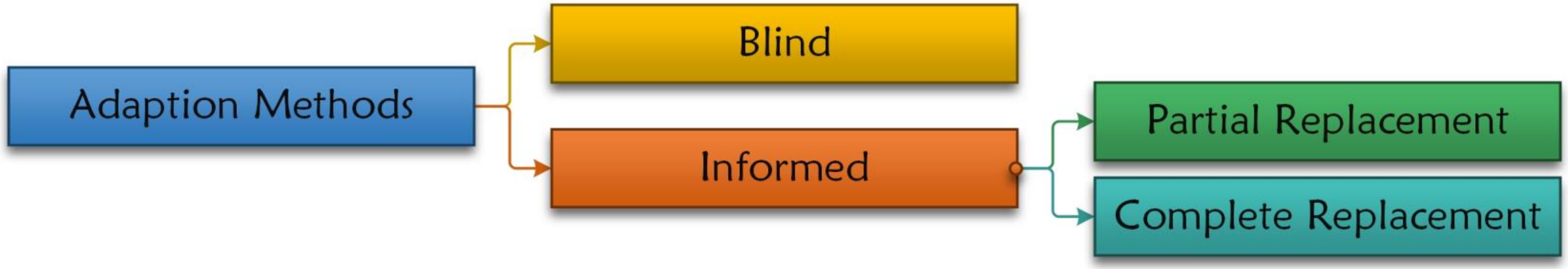
$$\beta_i \equiv \frac{P_T(y)}{P_S(y)}$$

- **Challenge**: *How to estimate $y_i \sim P_T(y)$?*
- To estimate target label distribution, we use *confusion matrix* $C_{k\times k}$ of a classifier that is trained on the source data
- Since we don't have access to the labels in the target data, we average model predictions on the test data to create $\mu(\hat{y})$ whose its ith element is the fraction of total prediction on the test set where the classifier predicted label i
- **Assuming that the confusion matrix is invertible** we can estiamte β by solving the following linear system:

$$C_{k\times k} \equiv \mu(\hat{y})$$

Concept Drift Correction

- Passive (Blind)
  - Update your model once a while without applying any shift detection
- Active (Informed)
  - Adapt your model once a shift detection triggers for a shift
  - Adaption or replacement could be globally or locally – depending on the learning algorithm



Advanced

Transfer Learning & Active Learning

VECTOR INSTITUTE

CIBC Banking that fits your life.

What if we could not correct a model?

- **Transfer learning - Reusing an existing model**
  - Same domain, different tasks
    - Target data has more/less classes (CIFAR-10 vs CIFAR-100)
    - Completely different tasks (trained for question-answering, used in sentiment analysis)
  - Different domains, same task
    - Training on grayscale images and testing on colored images

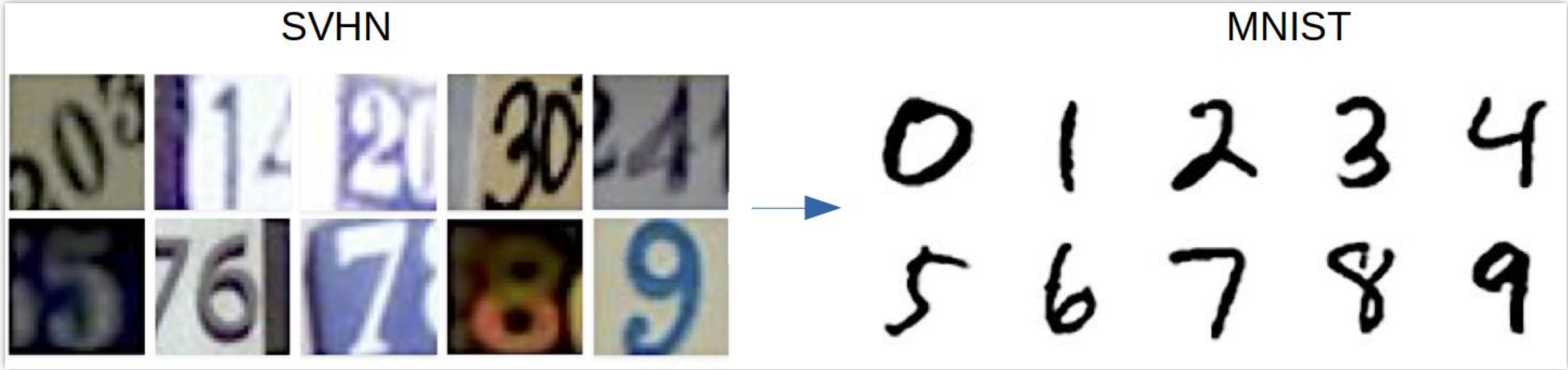
- **Active learning - Learning interactively with fewer training labels**
  - Not enough data from the target domain
  - Significant difference between the source and target distributions (no overlap)
  - We have the option to collect new samples: how can we do it more efficiently and effectively?

Transfer Learning

Transfer Learning

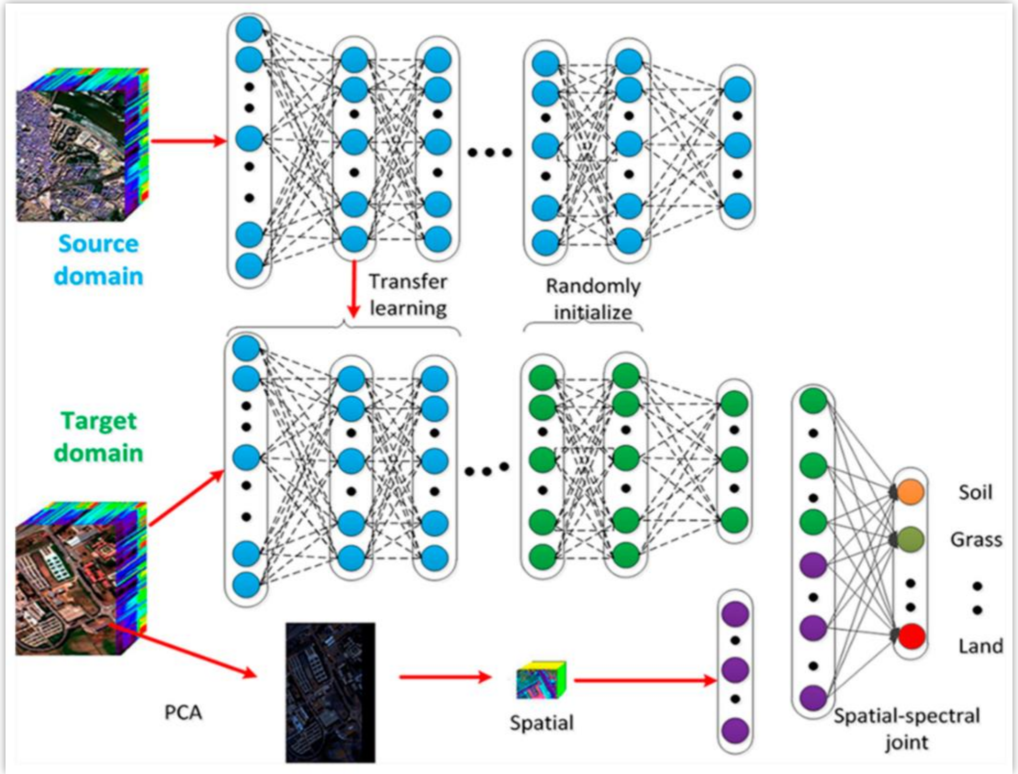
- Idea:
  - Layers in a neural network can be considered as feature representations
  - A common transfer paradigm is to **maintain the weights** in the earlier layers of network trained on some source task, and adapt only the weights in the last layer for a target task
  - Earlier layers may learn about abstract features such as edges and corners, which are common among different domains and tasks

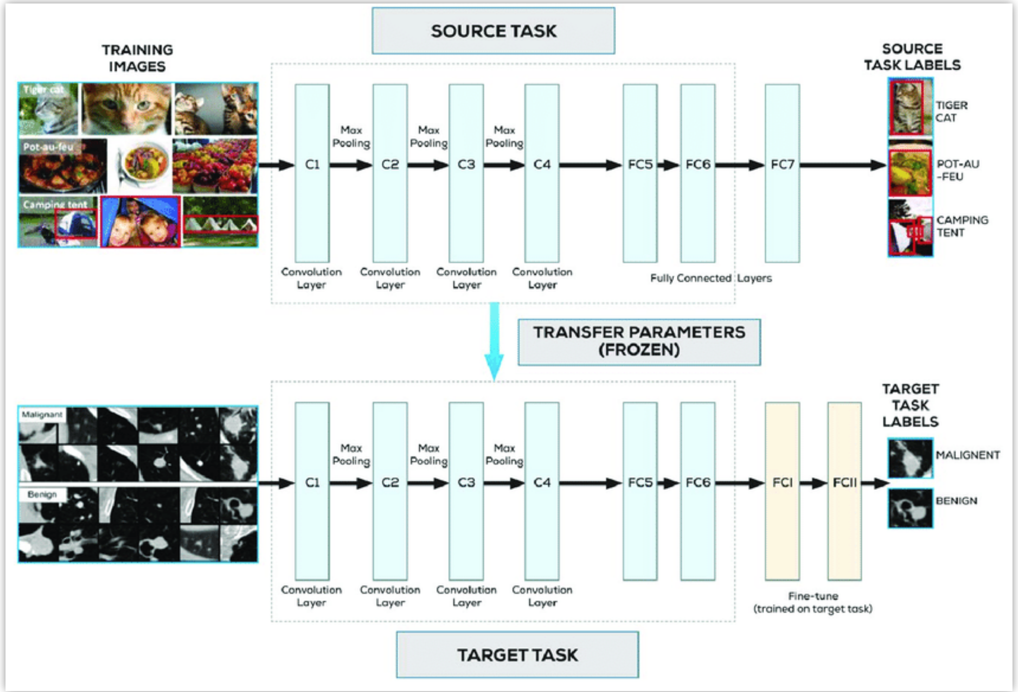
Transfer Learning – Cont.



Different domains, same task

Transfer Learning – Cont.





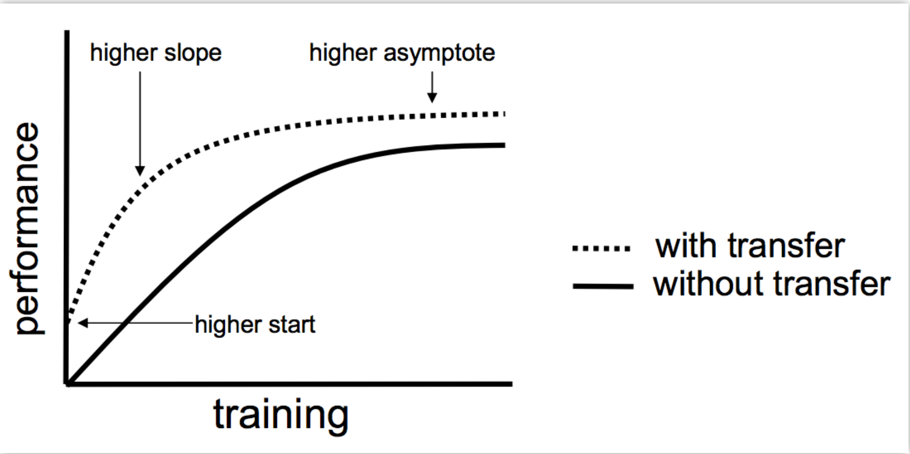
https://www.mdpi.com/2076-3417/9/7/1379/htm

https://pubmed.ncbi.nlm.nih.gov/30383900/

Same domain (colored images), different tasks

Transfer Learning – Cont.
Benefits

- Use the knowledge gained by a machine learning model from one task and apply it to a different task
- Possibility of reusing the same model
- The model training may have a higher start
- The training rate would improve
- The overall model performance may increase



Active Learning

- The model prioritize the labeling of new data such that training the model on the new data would have the maximum impact on model performance
- Interactively query the user to label new data points
- Can significantly reduce the number of new labeled data points required
- Query selection strategies
  - **Uncertainty Sampling**
    - Classification uncertainty: Being less confident about the model predictions - probabilities are not significantly different
    - Classification entropy: Uncertainty is proportional to the average number of guesses one must make to find the correct class

Active Learning – Cont.
Types

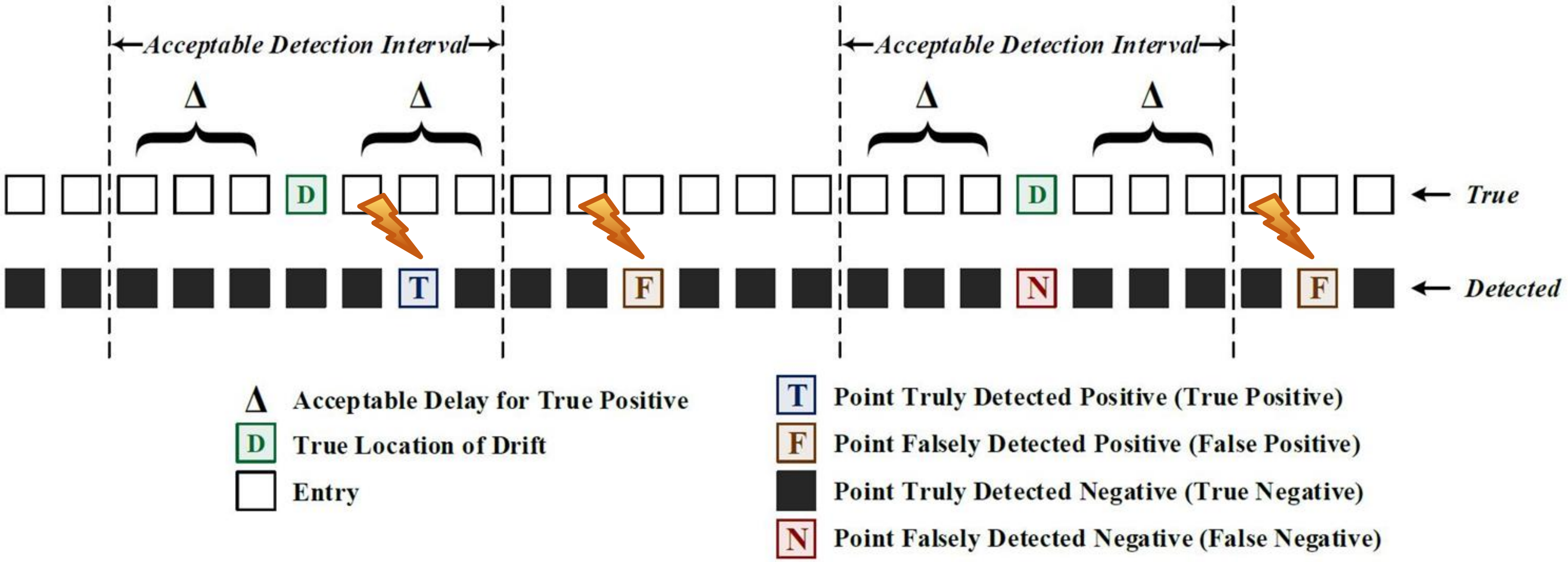
- **Pool-based active learning**
  - The model has access to a large pool of unlabeled data points
  - Query or rank the most informative samples
- **Stream-based active learning**
  - Stream of unlabeled samples
  - Decide to query the user for labeling of the streamed sample or not

Discussion

Evaluation Measures

- What shift detector is preferred?
  - ***Highest true positive***, ***lowest false positive*** and the ***lowest false negative***
  - The resources will be kept busy if the drift detector incorrectly alarms for concept drift repeatedly.
  - The error-rate of classification typically increases as does the false negative number.
- **The delay of detection:**
  - Shorter detection delay results in losing less data for learning, it means more instances from the new distribution can be used for learning.
- How about the model accuracy or loss:
  - It confirms whether using drift detection methods are beneficial or not!

Evaluation Measures



Discussion

- Always keep track of your data and your model performance
- If your model accuracy dropped (significantly), something is off…
  most likely due to some shift in data
- The significant level varies from my domain to another
- If you are not detecting a shift, there could still be a shift in your data
- Domain knowledge helps a lot
- Different types of data shift can co-occur
- Model repository for recurring concepts – Potentially for transfer learning
  - Some models trained in 2008 could be potentially used in 2020
- Track influencers

Packages

ALIBI DETECT

scikit-multiflow

TORNADO

moa

modAL

Q&A