# Economically-Efficient Data Stream Analysis

Roberto Oliveira Jr.
Advisor: Adriano Veloso
Co-advisor: Wagner Meira Jr.

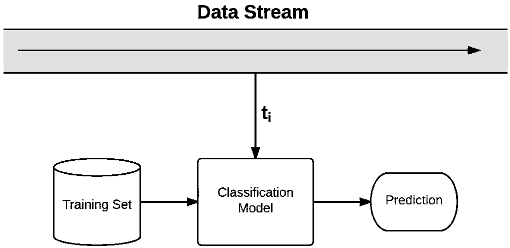
Computer Science Dept - UFMG - Brazil

# Data Stream

- Definition
  - Fast and possible unbounded sequence of data that arrives at time-varying.
- Motivation
  - It allows us to process huge volumes of data.
- Problem
  - Automatically extraction of relevant patterns and relations considering limited computational resources.

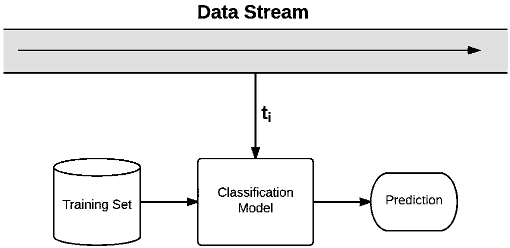
# Classification in Data Streams

- Classification models are applied to distinguish between pre-defined labels.



# Classification in Data Streams

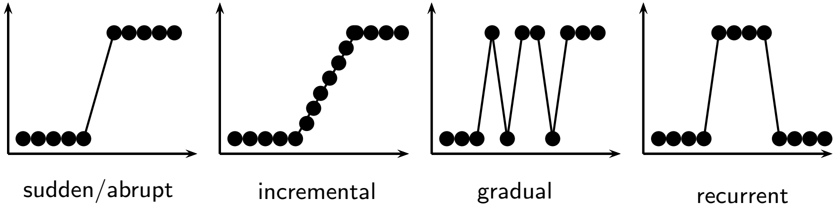
- Classification models are applied to distinguish between pre-defined labels.




- Data characteristics may change with time.

# Concept Drifts

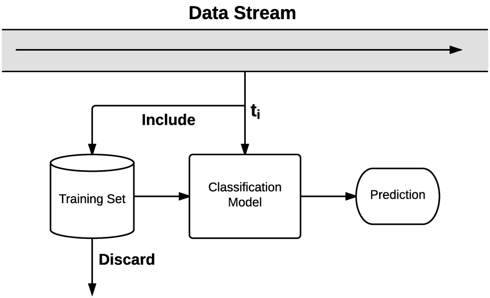
- Concept Drifts are unforeseen changes in nature of data over time.




- Data streams contains combination of such patterns.

# Classifying Data Streams

- Effective classification requires:
  - Updating the classification model as the stream evolves.
    - Limitation: memory, time and labeled examples.



# Research Question

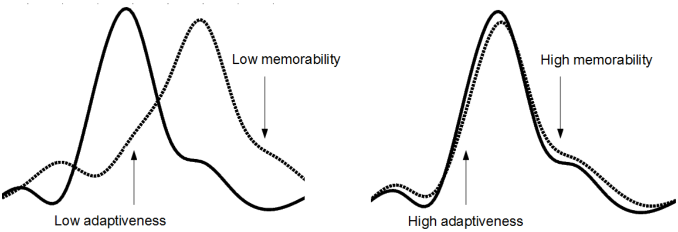
How to deal with concept drifts?

# Classification Model

- Classification models are composed by association rules.
  - $\{x \rightarrow y\}$, where $x \in X$ (input space) and $y \in Y$ (output space).
- Efficiently updated as the training set evolves.
- Models are built on-the-fly:
  - For a given $[x_i, *]$, rules $\{x \rightarrow y\}$ such that $x \in x_i$ are produced.
  - Prediction is performed from the combination of these rules.
- At each time step is produced a model $\mathcal{R}(x_i)$.

# Dealing with Drifts

- Two properties are necessary in order to produce classifiers that are robust to drifts:
  - Adaptiveness:
    - The ability to adapt itself to drifts.
  - Memorability:
    - The ability to recover itself from drifts.

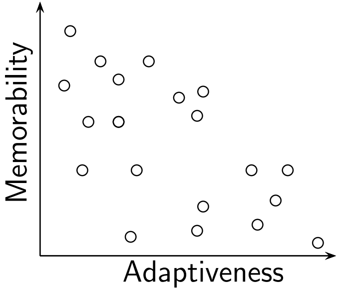


# Dealing with Drifts

- Improving both properties simultaneously may lead to a conflict-objective problem.
    - Improve adaptiveness may hurt memorability, and vice-versa.

# Economic Efficiency

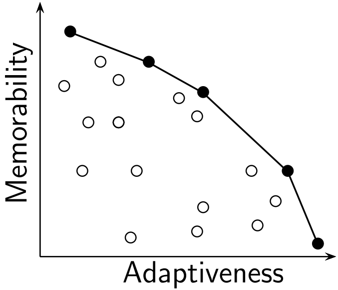
How to balance Adaptiveness and Memorability?



# Pareto Efficiency

## Pareto frontier: Dominant Points

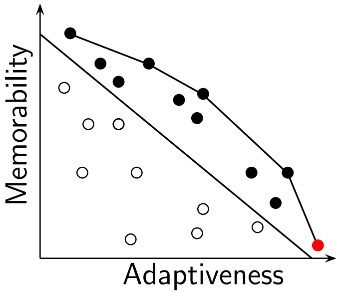
- $U_m(a) \geq U_m(b)$ and $U_a(a) \geq U_a(b)$
- $U_m(a) > U_m(b)$ or $U_a(a) > U_a(b)$



# Compensation — Kaldor-Hicks Principle

Region of compensation:

- Overall utility: $U(d_i) = U_m(d_i) + U_a(d_i)$
- Baseline point:
  $d^* = \{d_i \in \mathcal{P}_n | \forall d_j \in \mathcal{P}_n : U(d_i) \leq U(d_j)\}$

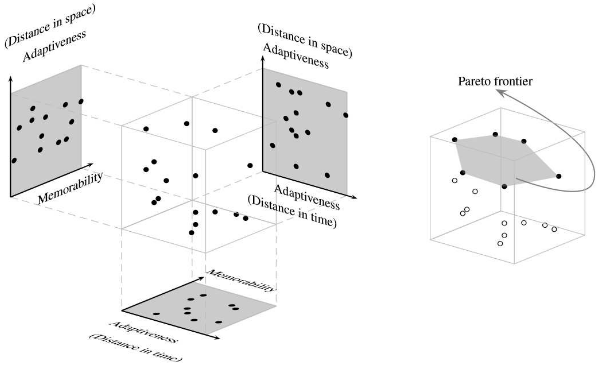


# Utility Measures

- Distance in space:
  - How similar training instance $t_j$ is to the newest instance $t_n$.
  - $U_s(t_j) = \frac{|\mathcal{R}(t_n) \cap \mathcal{R}(t_j)|}{|\mathcal{R}(t_n)|}$
- Distance in time:
  - How fresh is the training instance.
  - $U_t(t_j) = \frac{\gamma(t_j)}{\gamma(t_n)}$.
    - $\gamma(t_j)$ returns the time in which training instance $t_j$ arrived.
- Random permutation of training instances:
  - $U_r(t_j) = \frac{\alpha(t_j)}{|\mathcal{D}_n|}$
    - $\alpha(t_j)$ returns the position of $t_j$ in the shuffle.
    - $\mathcal{D}_n$ is the training set at time step $n$.

# Utility Space

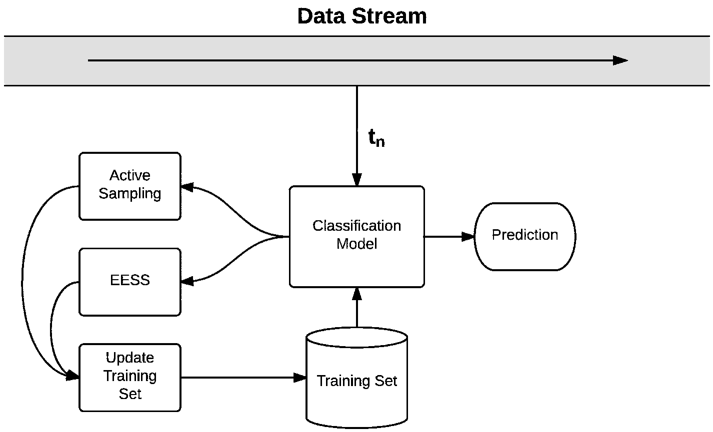
1. Place training instances in the utility space.
2. Select training instances in the Efficiency Region:
   - Pareto-Efficient Selective Sampling (PESS).
   - Kaldor-Hicks-Efficient Selective Sampling (KHSS).



# Reducing Labeling Efforts

- Random Active Learning
  - Naive strategy.
  - Simple to integrate.
  - Labeling Effort control: $\beta$.

# Economically-Efficient Selective Sampling



# Experimental Evaluation

## Setup

- Interleaved Test-Then-Train
- 1% of data provided as training seed;
- Massive Online Analysis (MOA) framework as evaluation environment;
- Baselines:

| Algorithm | Adaptiveness | Memorability |
|---|---|---|
| AC (KDD 2011) | Active Learning | Base Learner |
| HAT (JMLR 2011) | ADWIN | Decision Tree Model |
| ILAC (SIGIR 2011) | Data Projection | Incremental Training Set |

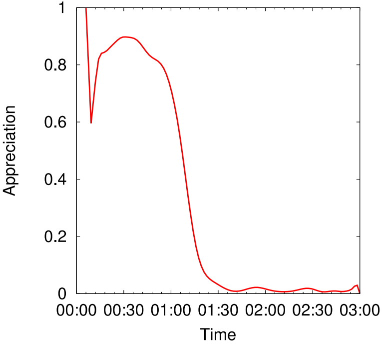
# Evaluation

- Measures used:
  - Mean Squared Error.
  - Labeling Effort: 10%; **25%**; 50%; 75% and 100%;
    - AC and EESS.
  - Training set size.
  - RAM-Hours.
- Datasets:

| Dataset | Concept Drift Pattern | | | |
|---|---|---|---|---|
| | Sudden | Incremental | Gradual | Recurrent |
| Presidential Elections | - | X | X | - |
| Person of the Year | - | X | X | - |
| **FIFA World Cup - EN** | X | - | - | - |
| **FIFA World Cup - PT** | X | - | - | - |
| **Cover Type** | X | - | X | X |
| Spam Filtering | X | - | X | X |
| **Poker Hand** | - | - | X | X |

# Evaluation

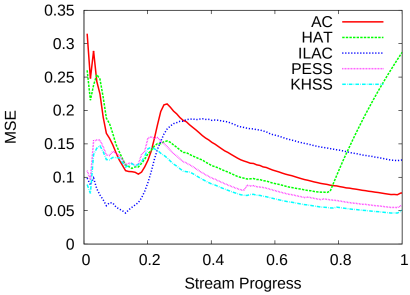
## FIFA World Cup - Portuguese

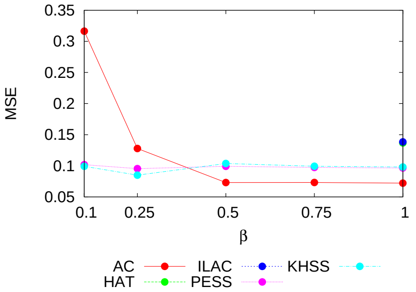



# Evaluation

## FIFA World Cup - Portuguese

### MSE and Labeling Efforts

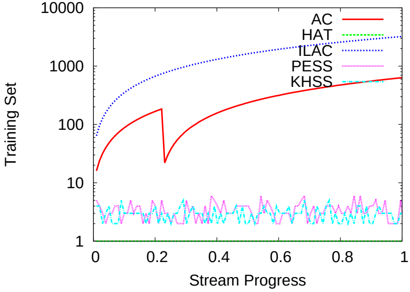


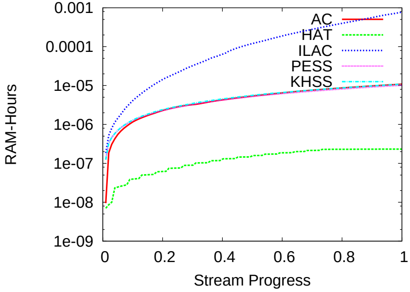



# Evaluation

## FIFA World Cup - Portuguese

### Training Size and RAM-Hours



# Conclusions

- Data analysis on streams.
  - Limited computing and training resources.
  - Concept drifts.

# Conclusions

- Data analysis on streams.
  - Limited computing and training resources.
  - Concept drifts.
- Efficiency and accuracy.
  - Incremental classifiers.
  - Adaptiveness and Memorability.
  - Pareto efficiency and compensation principle.
  - Simple-to-compute utility measures.
  - Ours algorithms were robust in different scenarios.

# Conclusions

- Data analysis on streams.
  - Limited computing and training resources.
  - Concept drifts.
- Efficiency and accuracy.
  - Incremental classifiers.
  - Adaptiveness and Memorability.
  - Pareto efficiency and compensation principle.
  - Simple-to-compute utility measures.
  - Ours algorithms were robust in different scenarios.
- Future work includes:
  - Other utility measures.
  - Employ our method to reduce Labeling Efforts.
  - Explore other classification models.

# Results


- Oliveira Jr., R., Veloso, A., Pereira, A., Meira Jr., W., Ferreira, R., and Parthasarathy, S. (2014). **Economically-efficient sentiment stream analysis**. In SIGIR. ACM.
- Veloso, A. (2012). **UFMG repository of machine learning**. https://code.google.com/p/machine-learning-dcc-ufmg/.
- Other results:
  - Buehrer, G., Oliveira Jr., R., Fuhry, D., and Parthasarathy, S. (2015). **Towards a parameter-free and parallel itemset mining algorithm in linearithmic time**. In ICDE. IEEE.

# Thank you!

robertolojr@dcc.ufmg.br