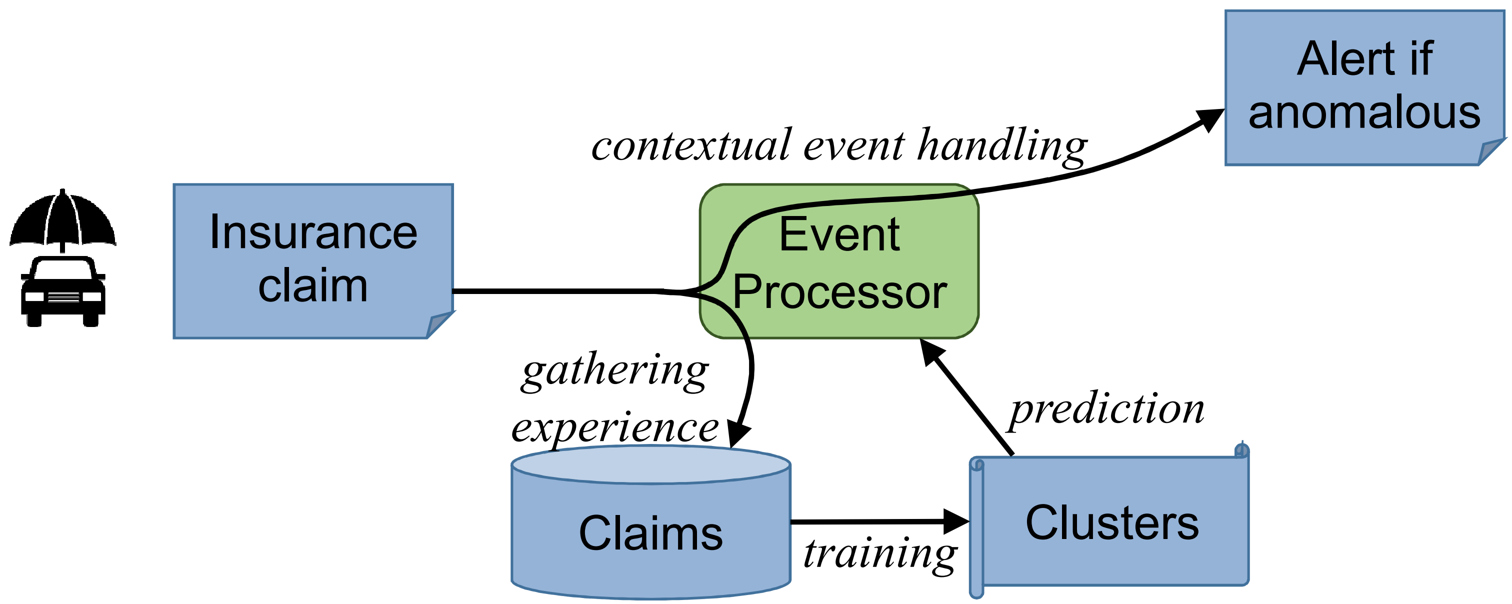
# AQuA: Adaptive Quality Analytics

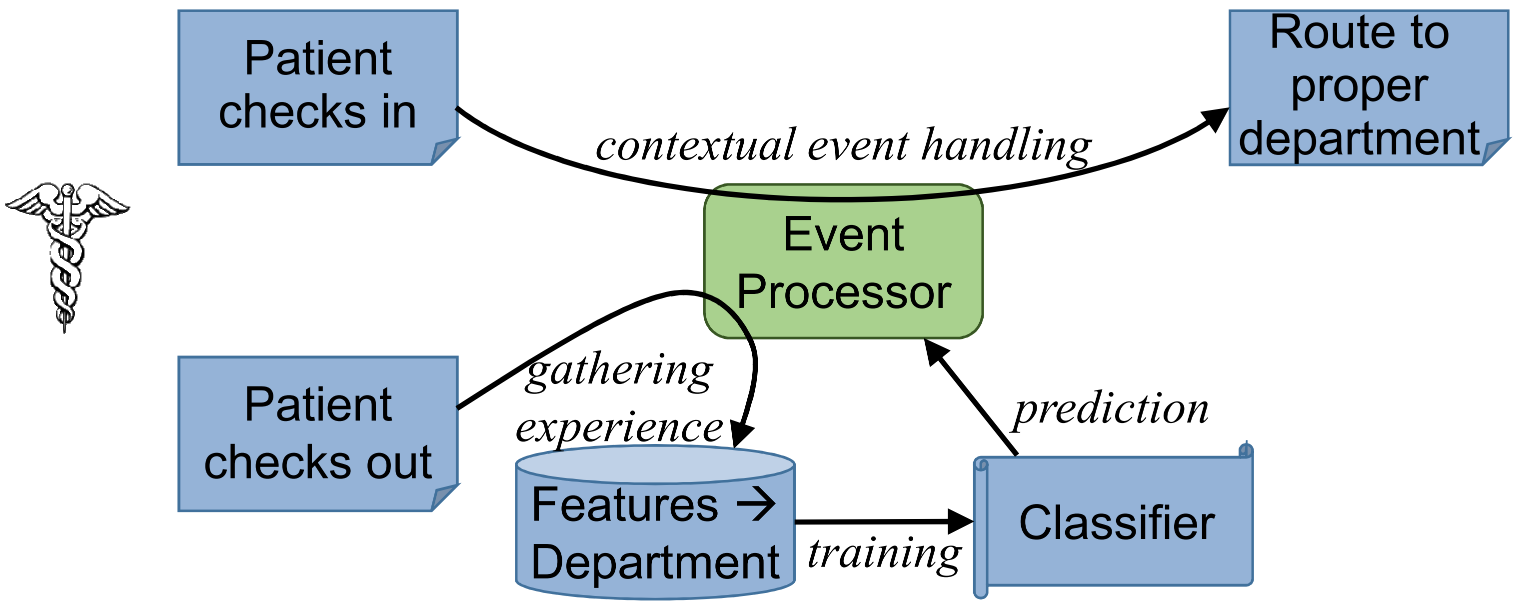
Conference on Distributed Event-Based Systems (DEBS), June 2016

Wei Zhang, Martin Hirzel, and David Grove

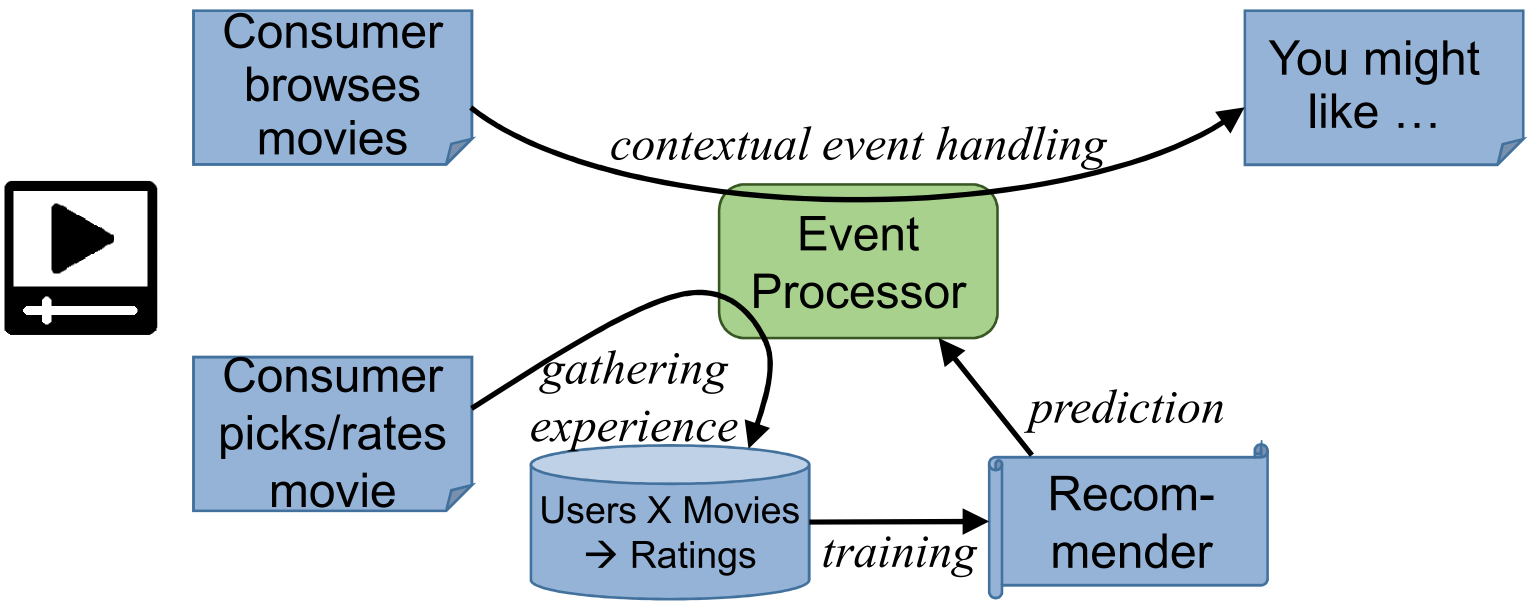
# Unsupervised Learning Scenario: Clustering



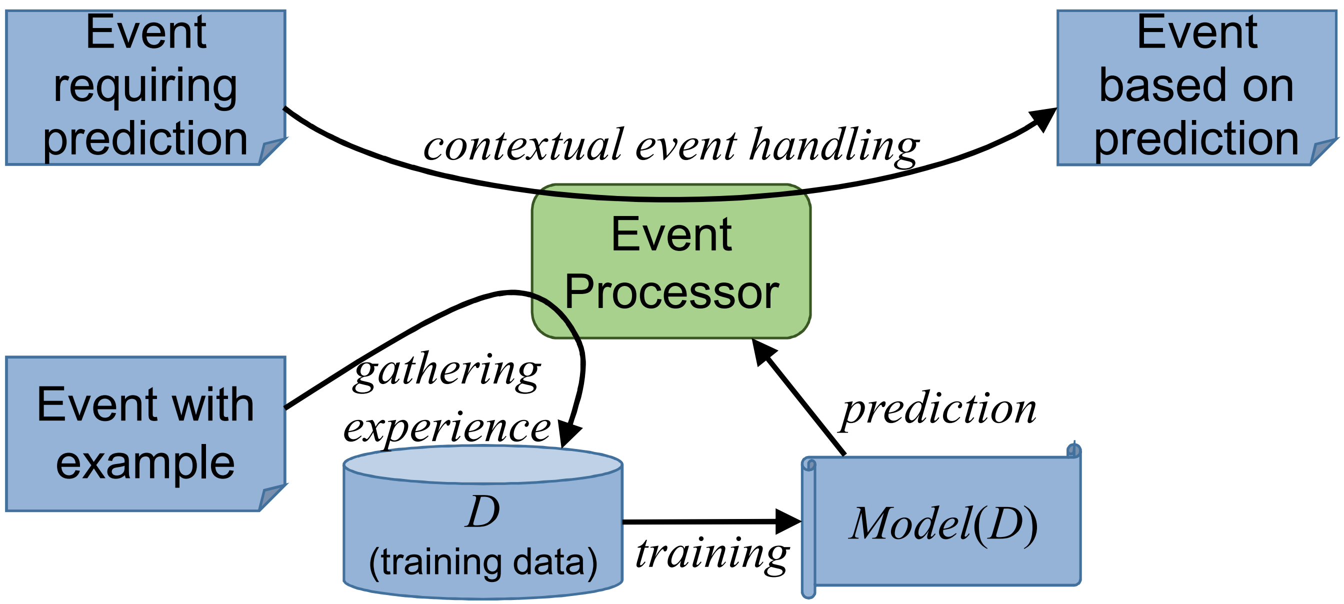
# Supervised Learning Scenario: Classification



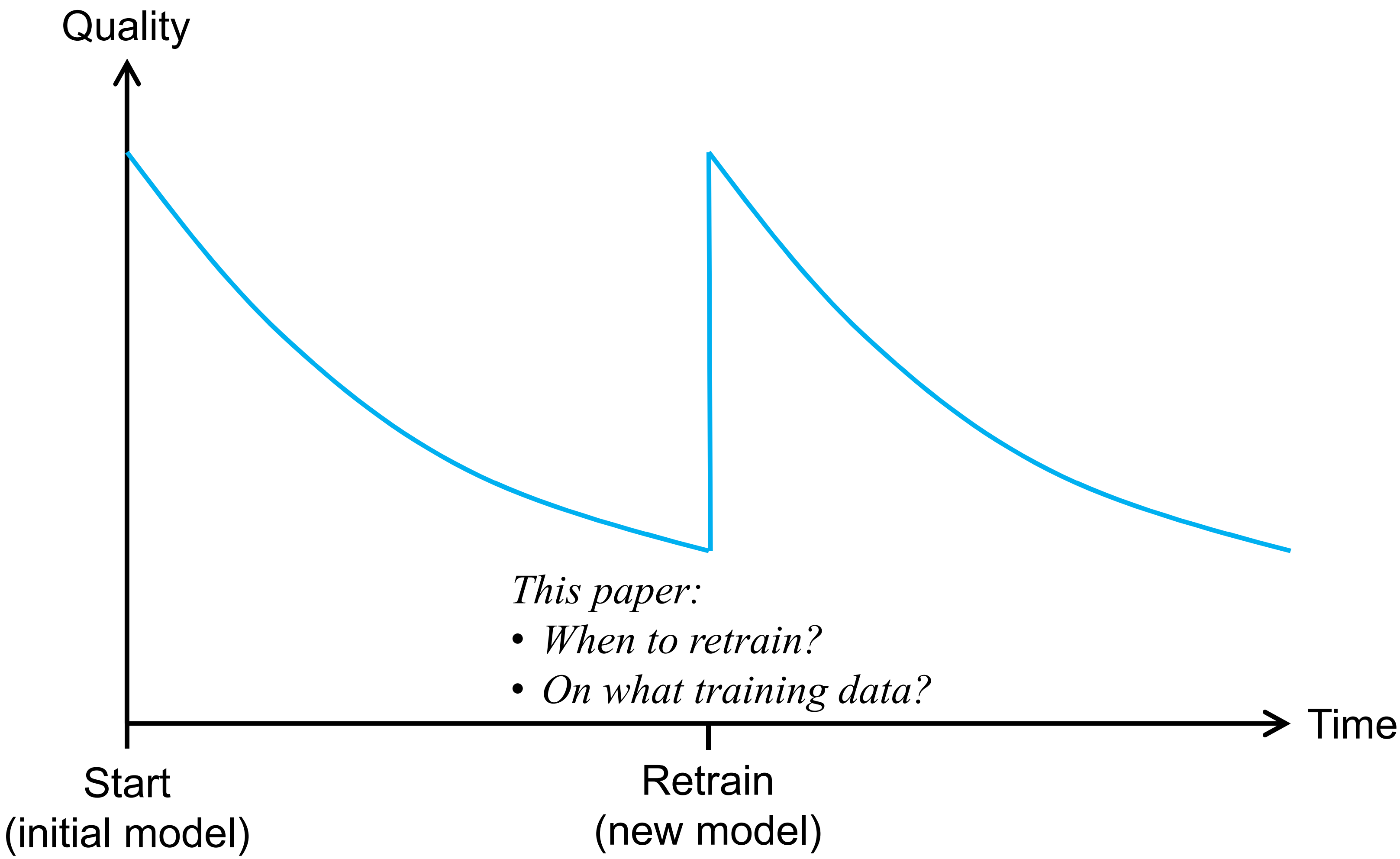
# Supervised Learning Scenario: Collaborative Filtering



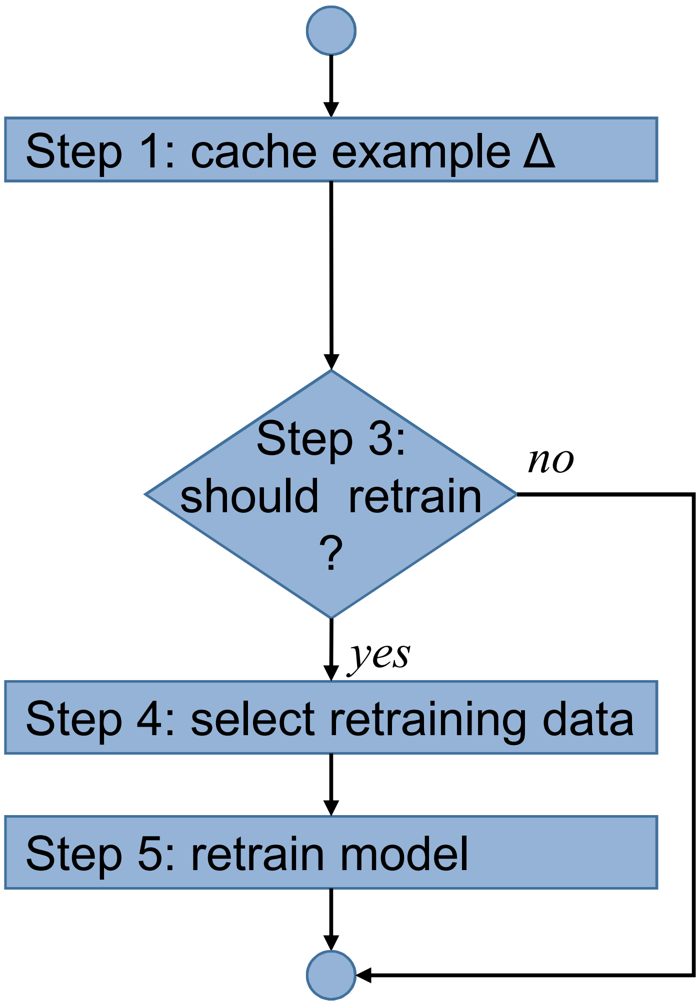
# META: Middleware for Events, Transactions, and Analytics [IBMRD'16]



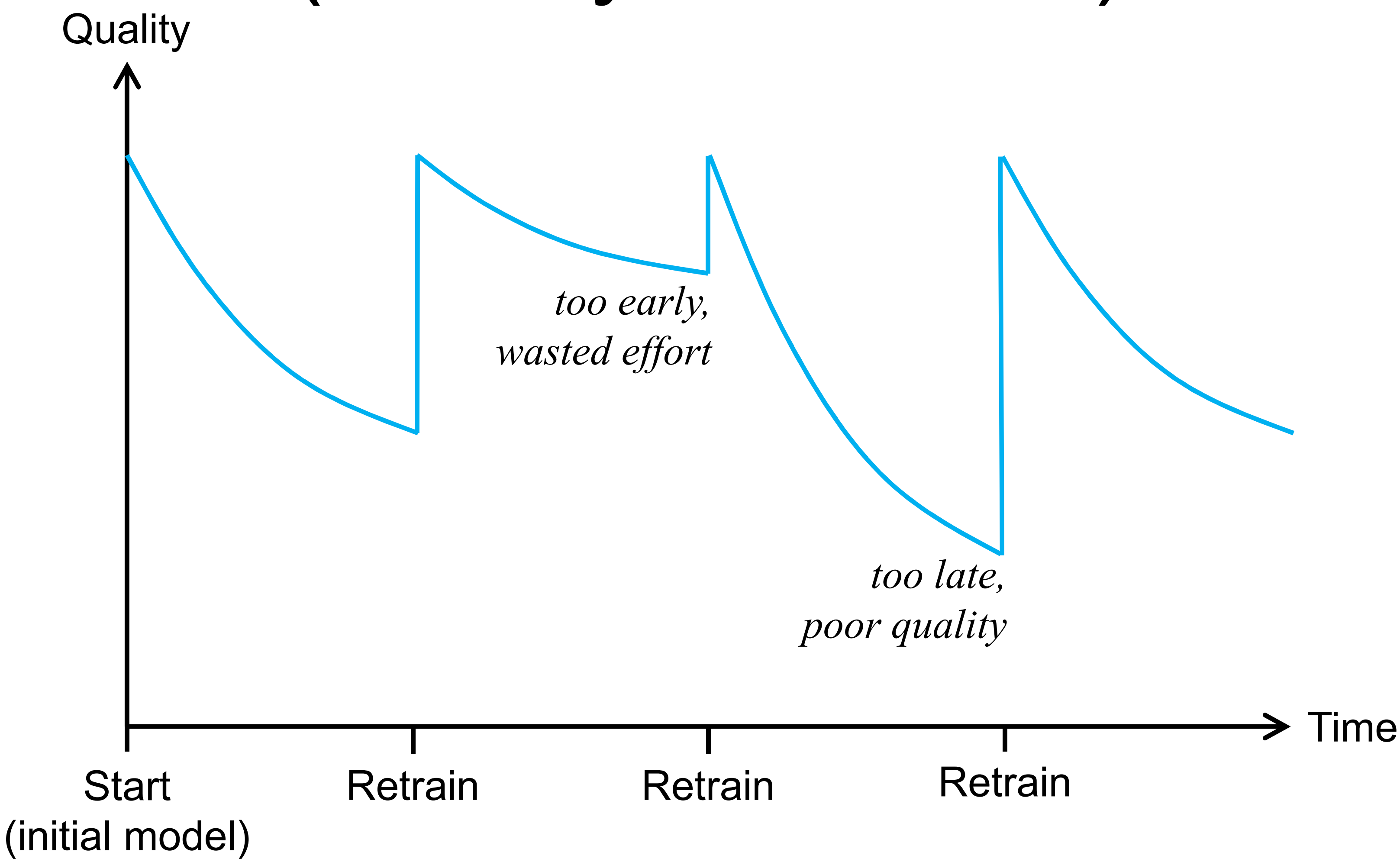
# Model Drift


Quality
Time
Start
(initial model)
Retrain
(new model)
*This paper:*
• *When to retrain?*
• *On what training data?*

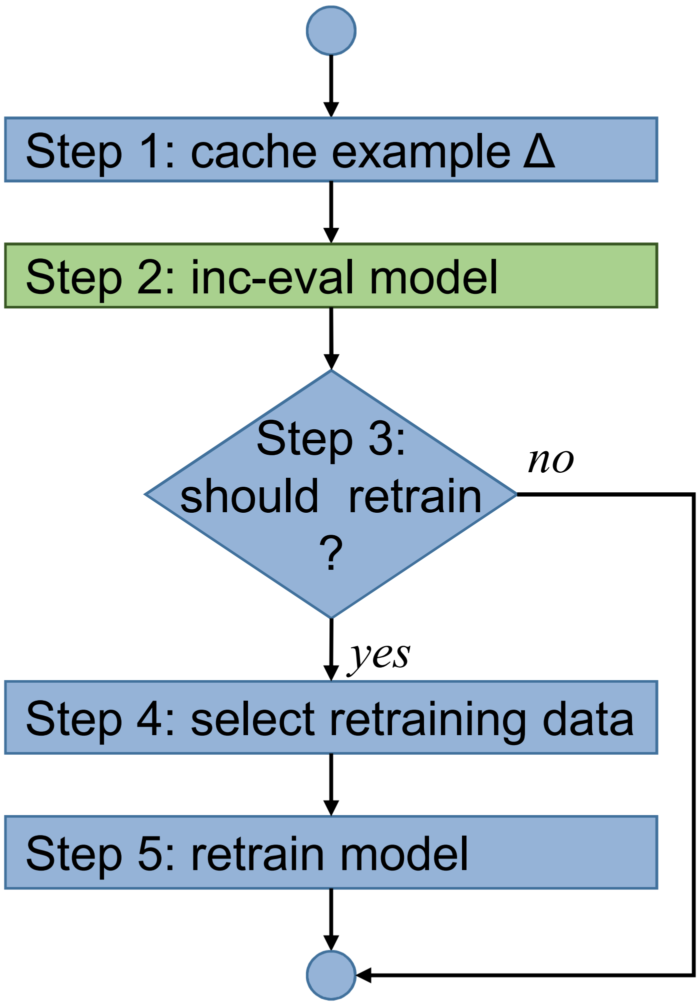
# Retraining Workflow (Attempt 1)



# ◆**Fix**ed-Size Retraining Strategy (Quality-Oblivious)



# Retraining Workflow (Attempt 2)



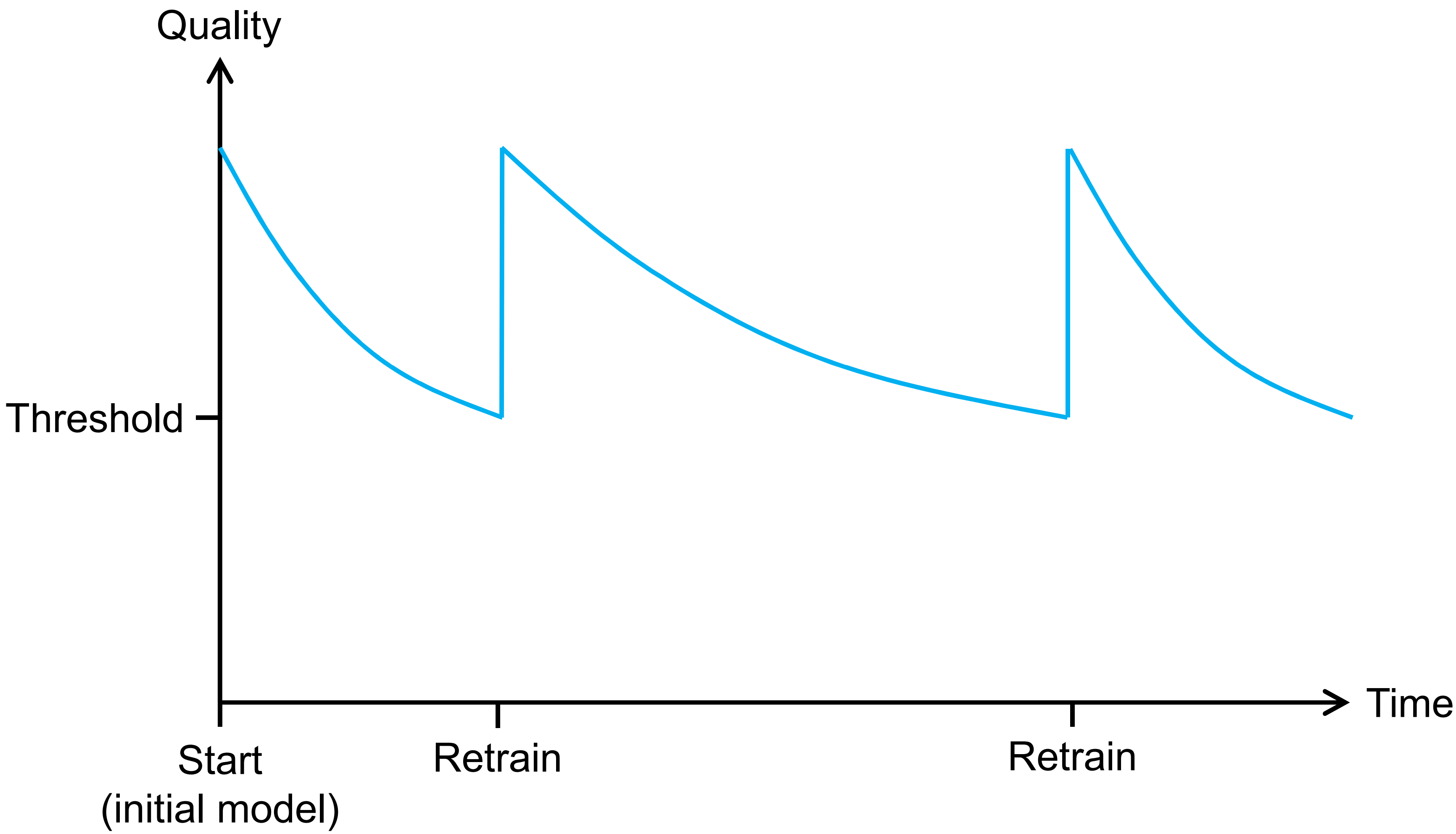

Training data $D$
Test data $T$
Evaluate error (inverse of quality):

$$evaluate(Model(D), T) = \sqrt{\frac{\sum_{t \in T} error(t)^2}{|T|}}$$

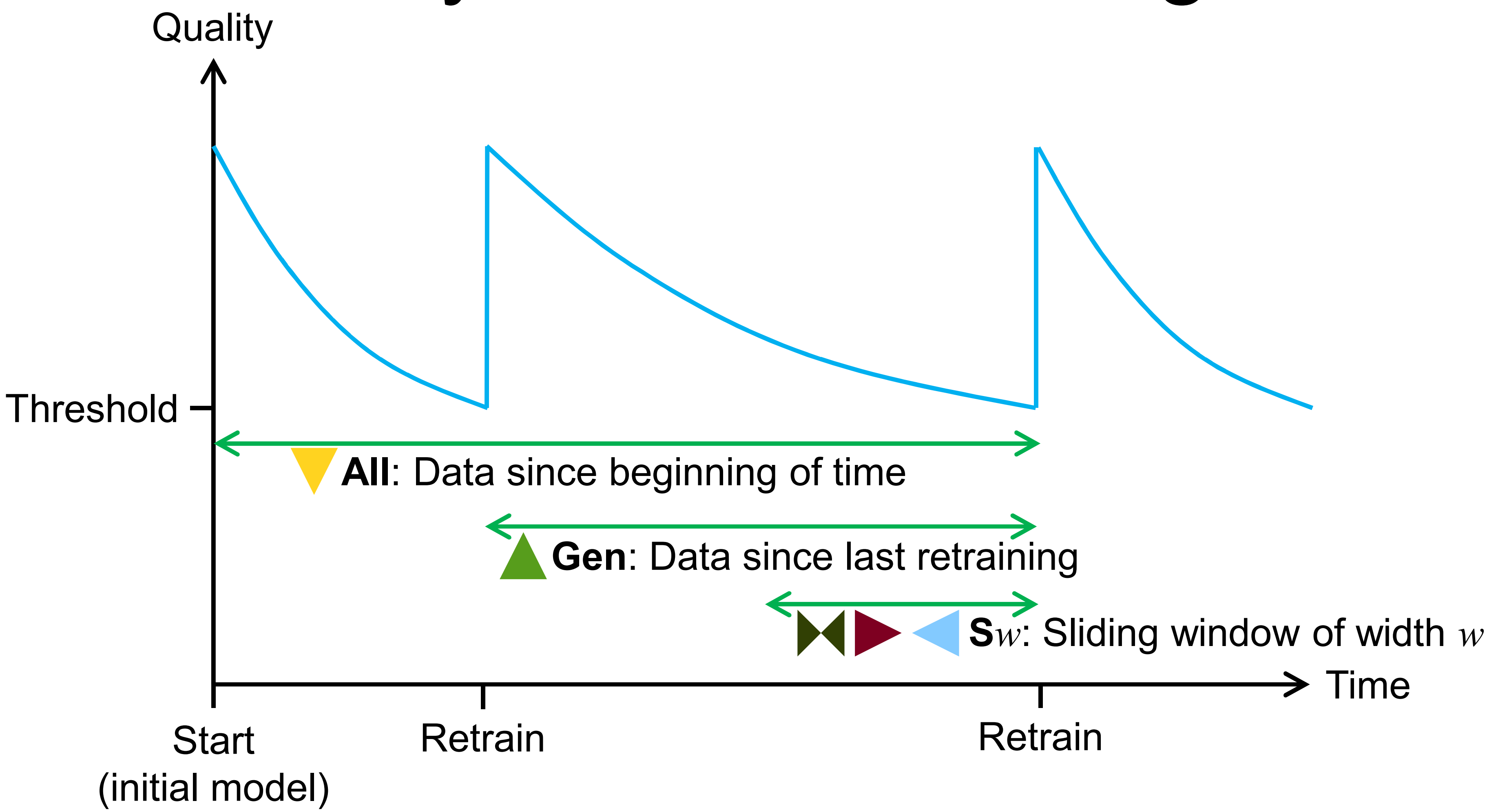
New test data $\Delta$
Incrementally update $sum$ and $count$

$$evaluate(Model(D), T \pm \Delta) = \sqrt{sum/count}$$

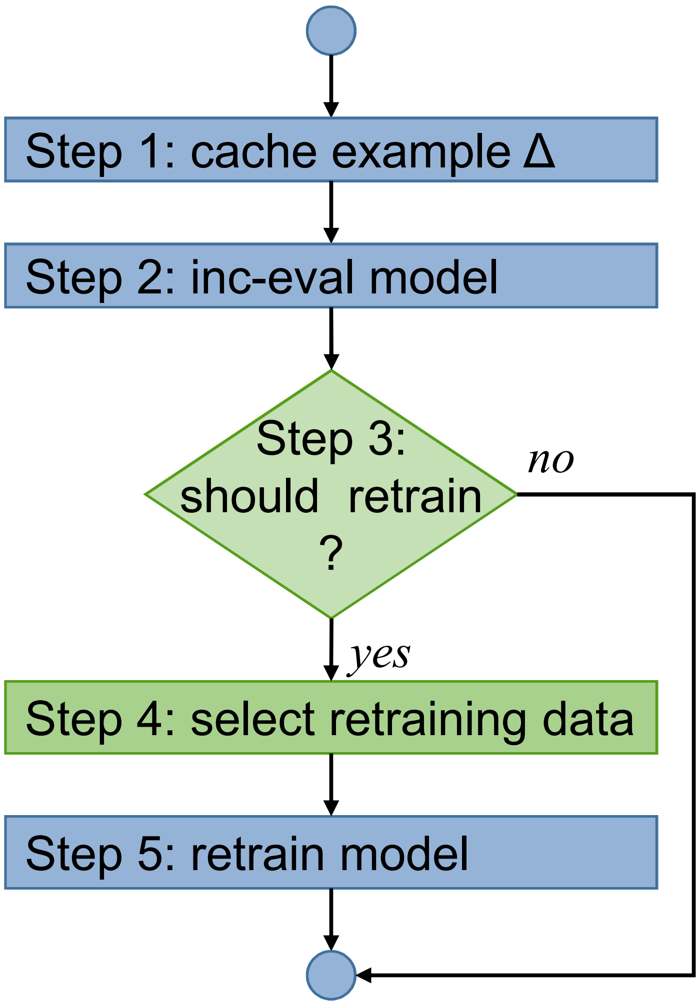
# Retrain When Quality falls Below Threshold



# Training Data Selection for Quality-Directed Strategies

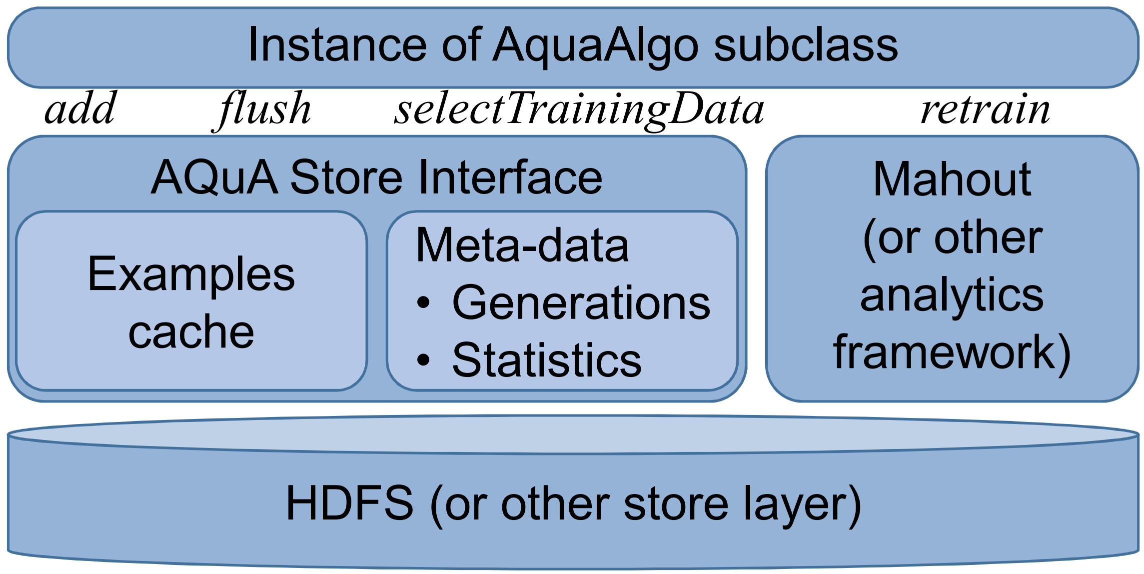


# Retraining Strategies

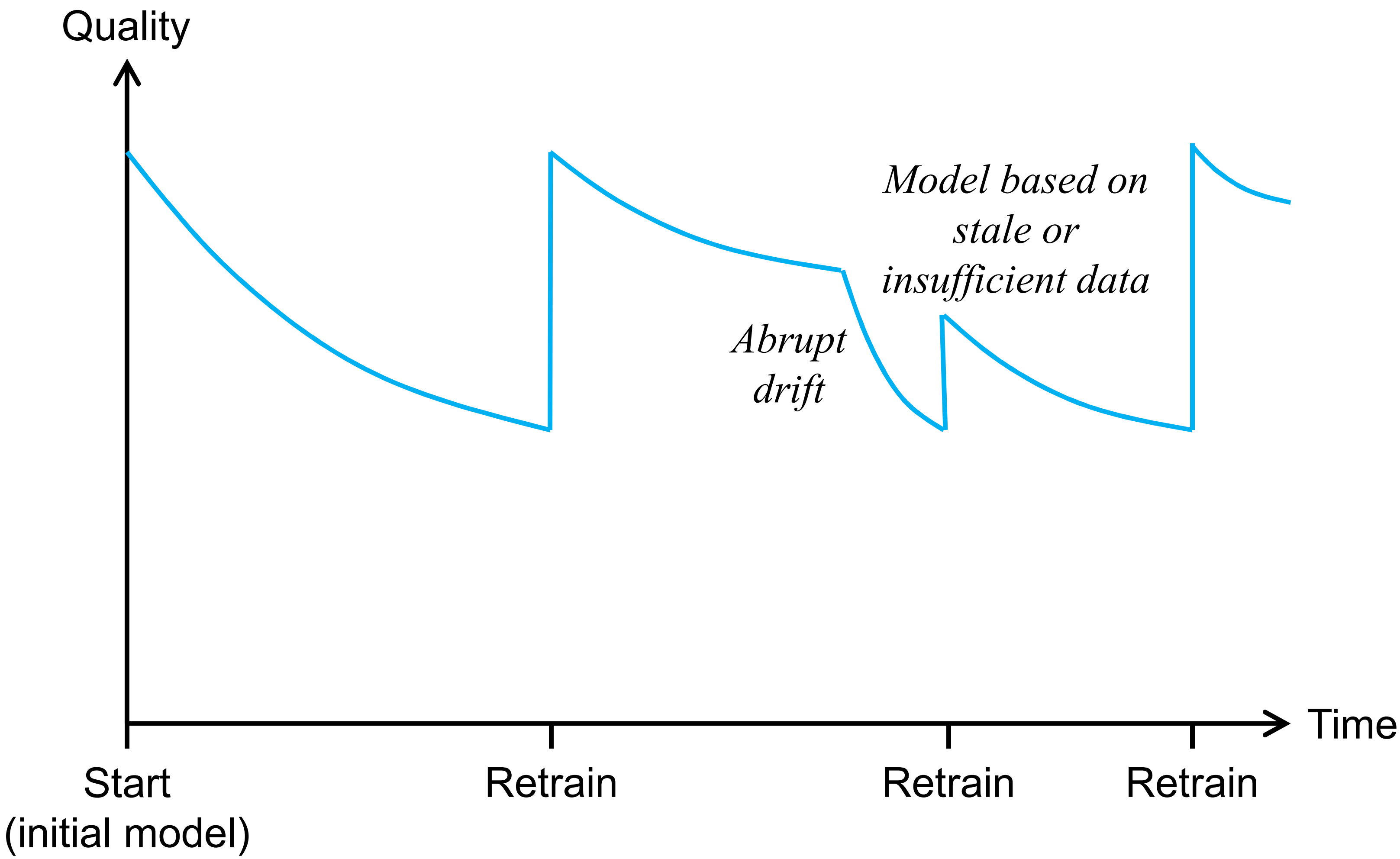



| Quality-oblivious strategies | Quality-directed strategies |
| --- | --- |
| **Never:** Never retrain model<br>**Fix**: Fixed-size retrain interval | **All**: Data since beginning of time<br>**Gen**: Data since last retraining<br>**S*w***: Sliding window of width *w* |

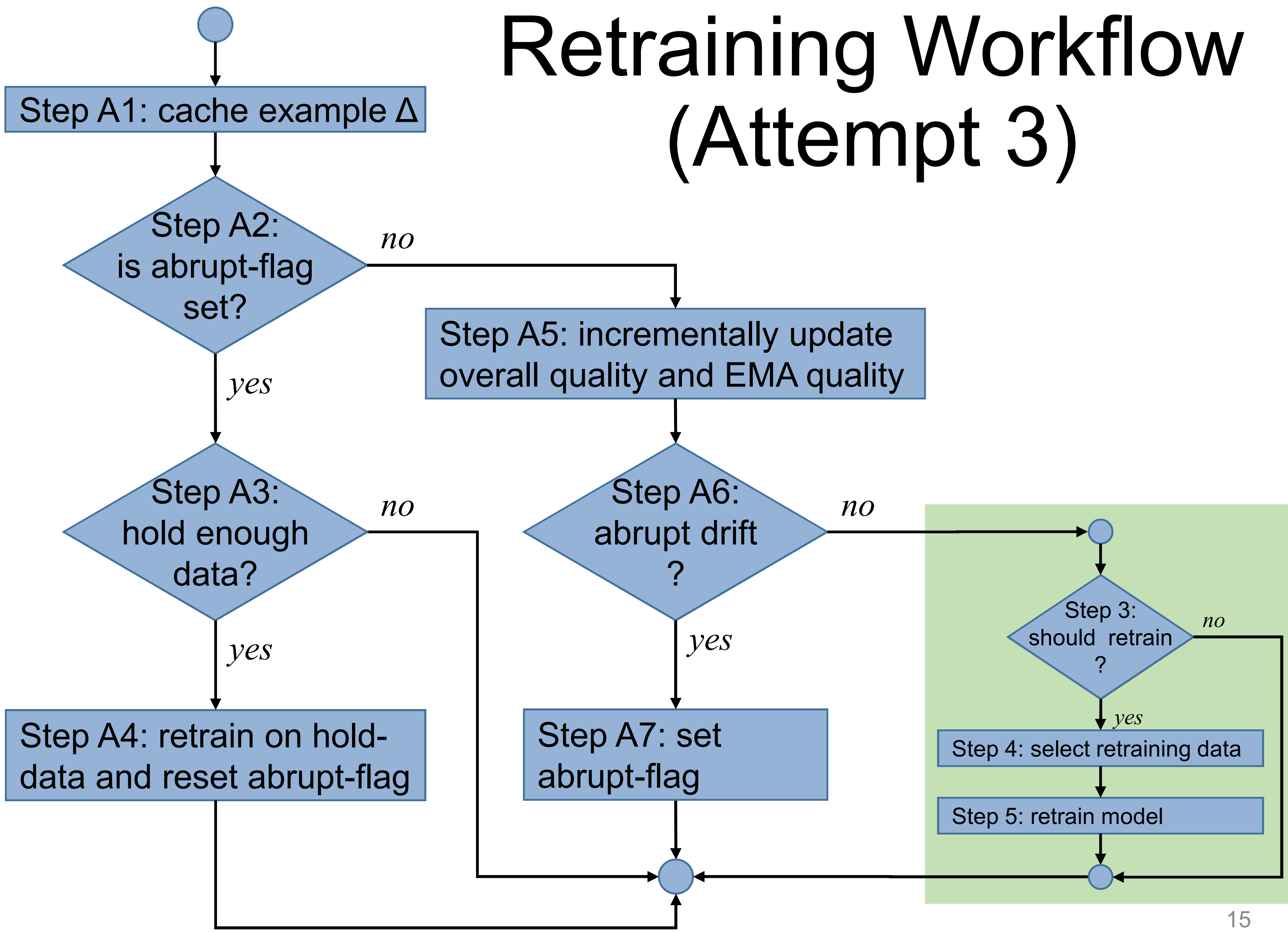
# AQuA Store Interface



# Gradual vs. Abrupt Model Drift

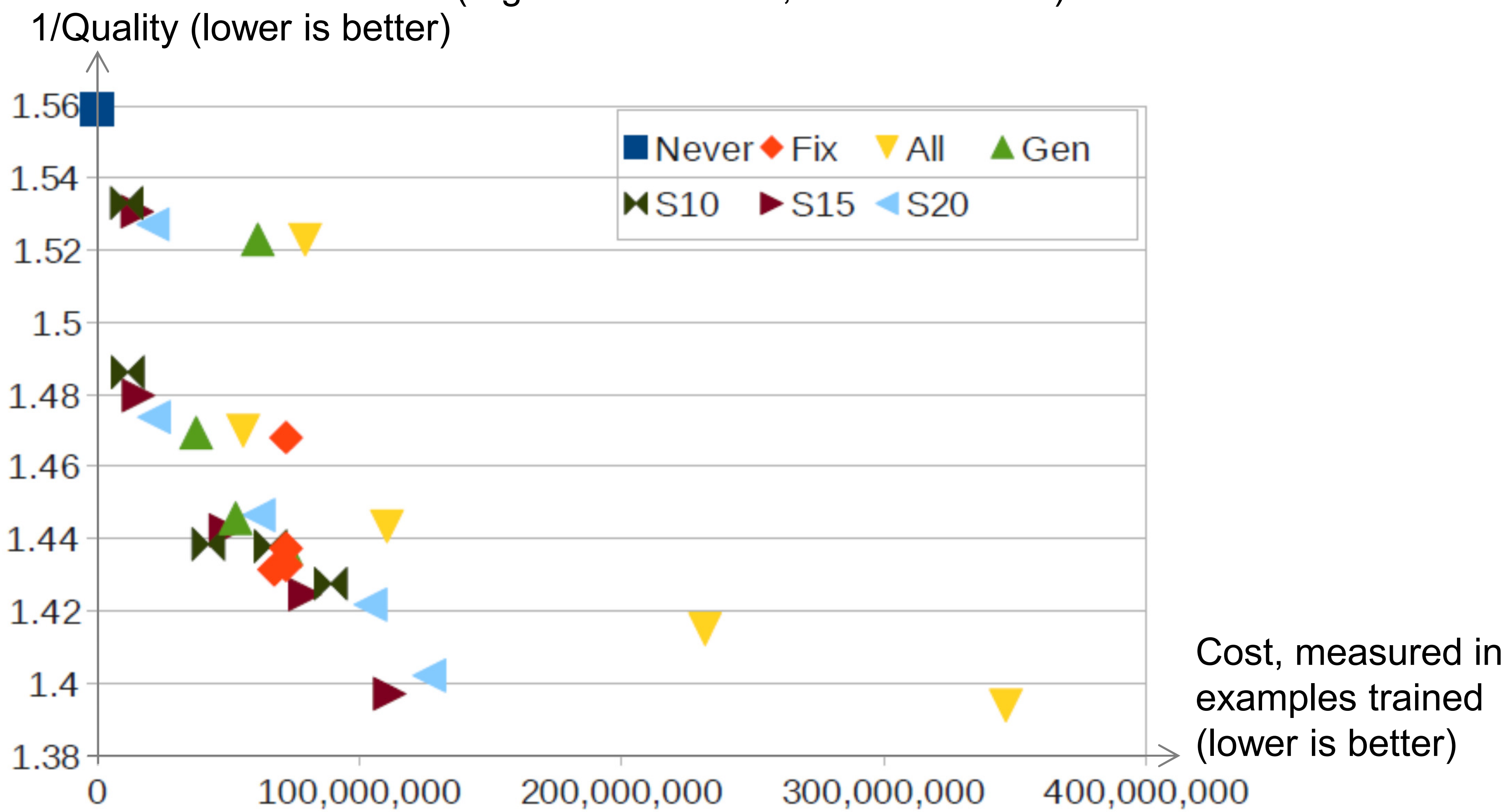


# Retraining Workflow (Attempt 3)



# Performance: Collaborative Filtering

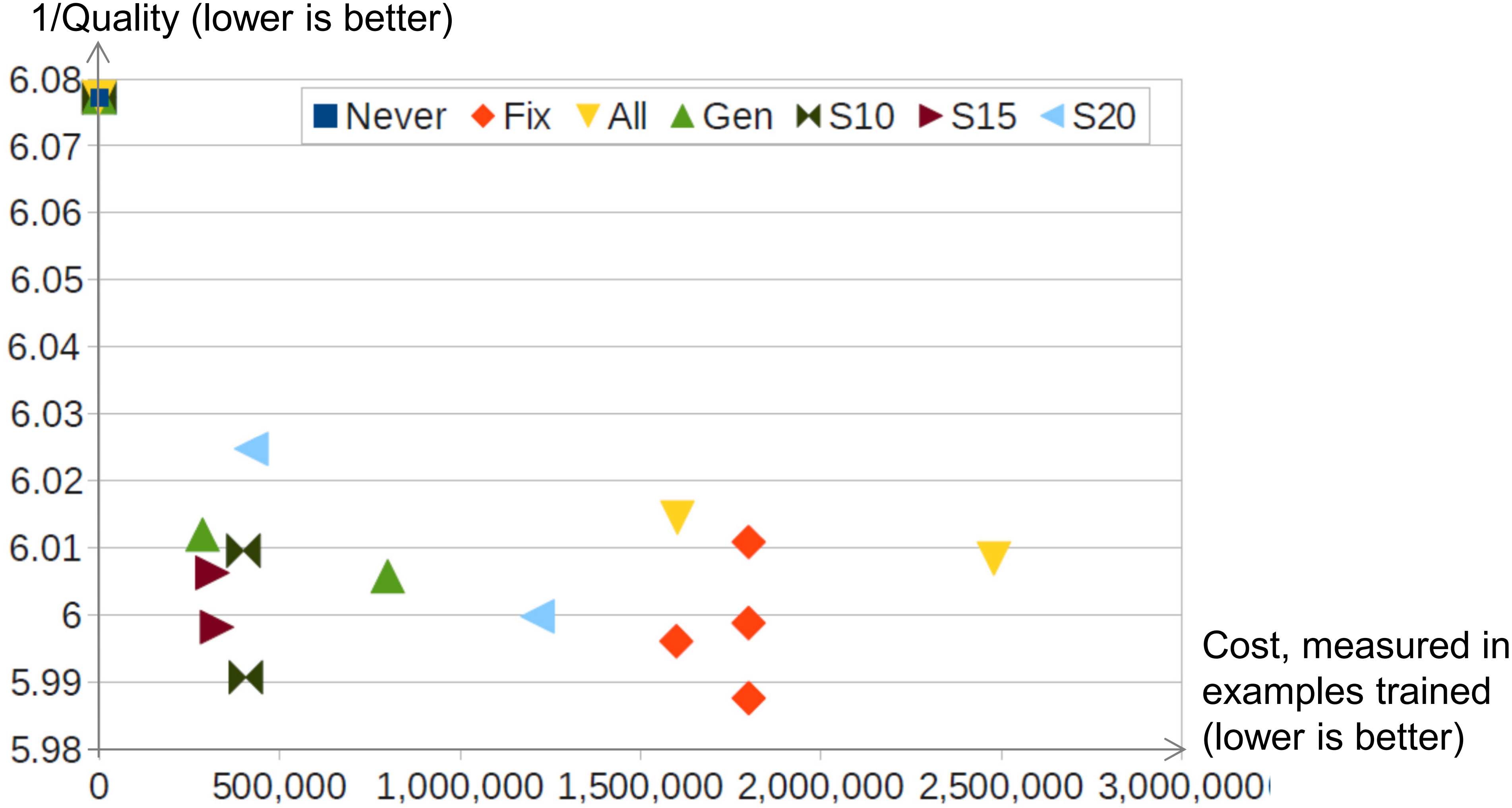
(Algorithm: ALSWR, dataset: Netflix)




S*w* (sliding window strategies) are on Pareto frontier

# Performance: Clustering

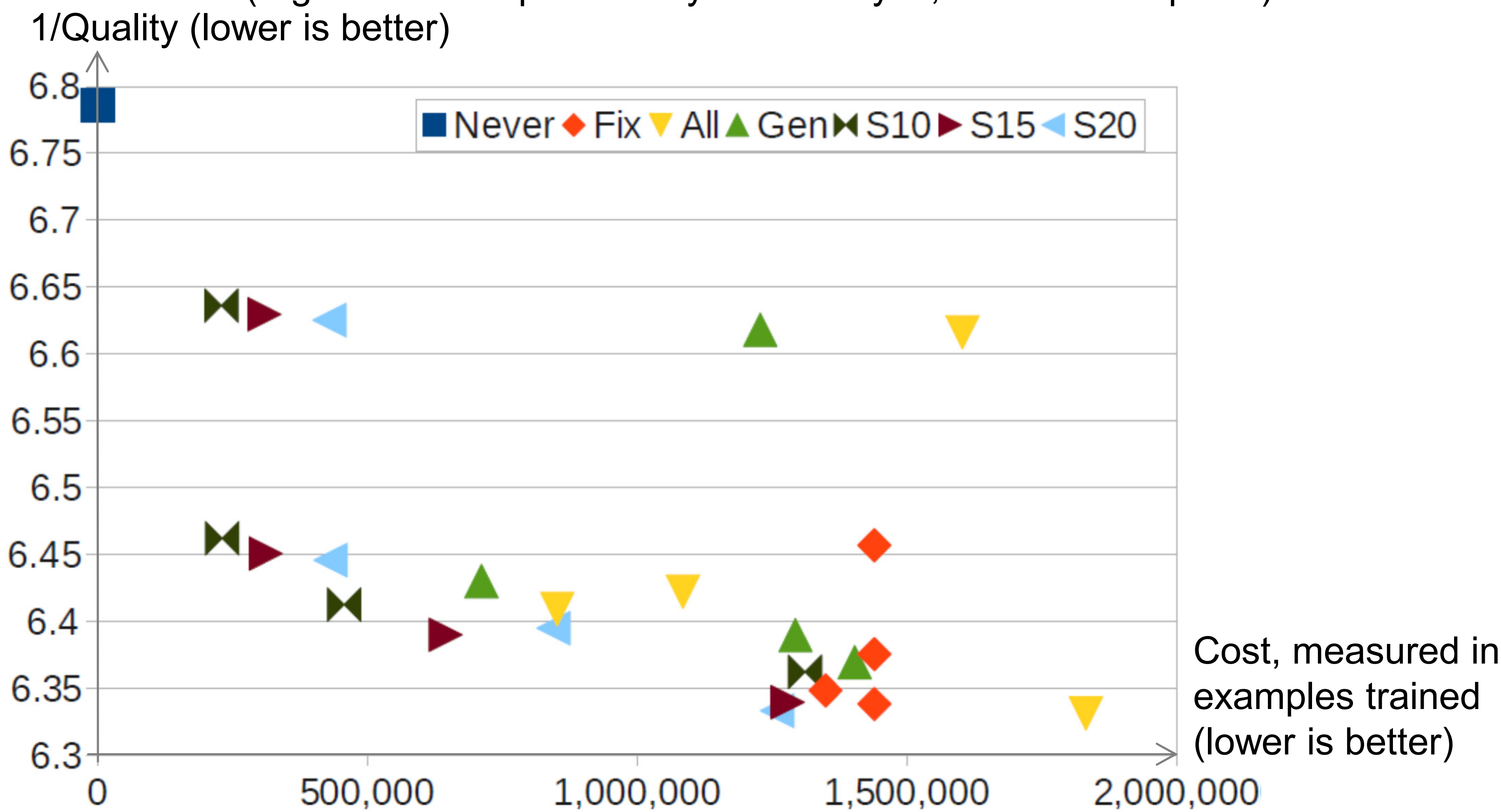
(Algorithm: KMeans, dataset: Wikipedia)




S$w$ (sliding window strategies) are on Pareto frontier

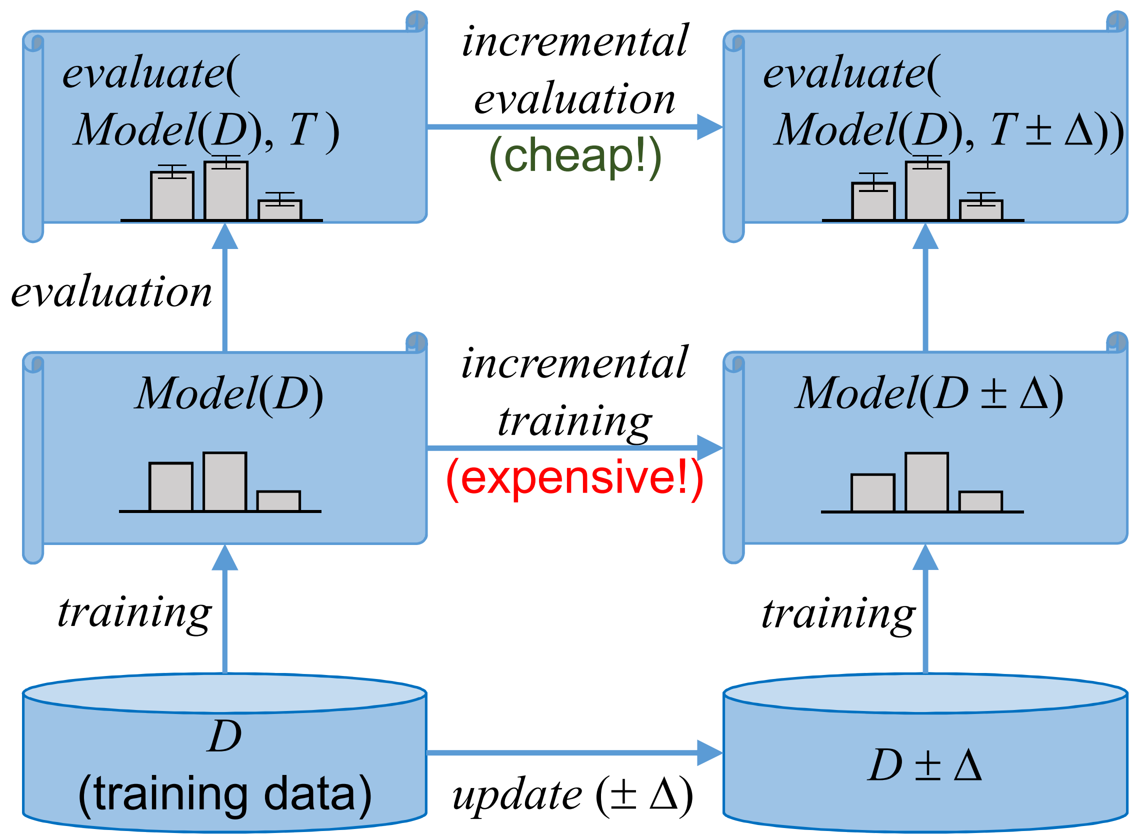
# Performance: Classification

(Algorithm: Complementary Naïve Bayes, dataset: Wikipedia)

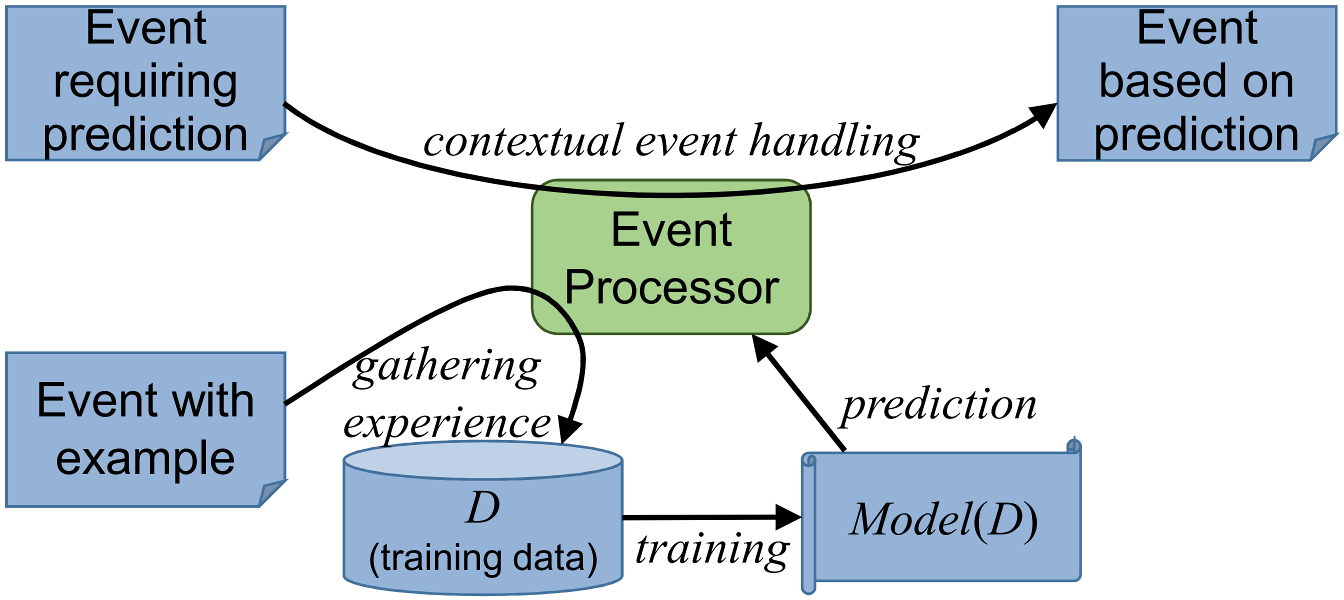



S*w* (sliding window strategies) are on Pareto frontier

# Related Work



# Conclusions




- Incremental evaluation of model quality
- Quality-directed retraining
- Strategies for gradual and abrupt model drift
- Sliding window strategies are on Pareto frontier