University of Glasgow
School of Computing Science
ESSENCE
PERVASIVE & DISTRIBUTED INTELLIGENCE

Statistical Model Updates in Distributed Computing: An Optimal Stopping Theory Perspective

Ekaterina (Katie) Aleksandrova
2133352A@student.gla.ac.uk

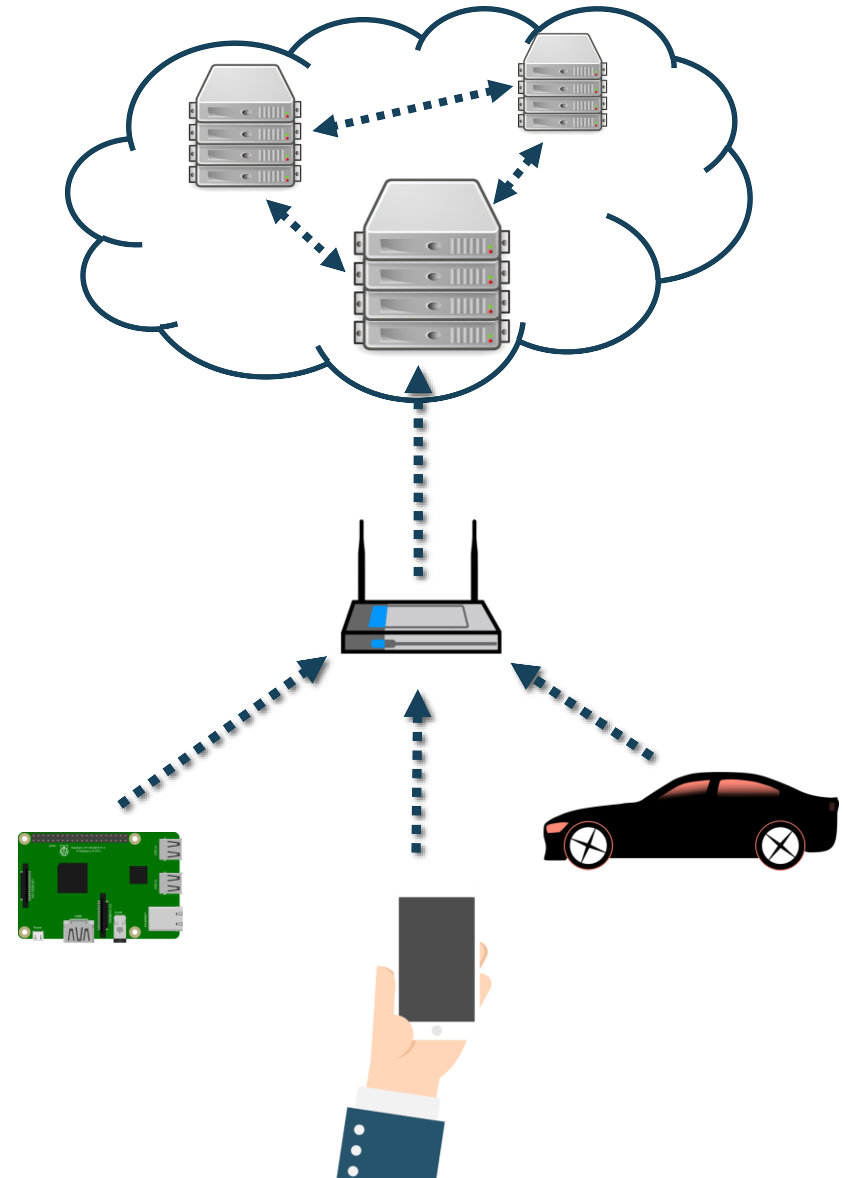
FRI 12th APRIL @ SoCS/SAWB 303

Supervision by: Dr Chris Anagnostopoulos

University of Glasgow
School of Computing Science

Problem statement

- **Internet of Things system:**
  - Edge sensors
  - Neighbourhood edge gateways
  - Data centres (the Cloud)
- **What we Are doing:**
  - Sense multivariate contextual data at the edge
  - Transfer the data to the Cloud for analysis
  - Have accurate and up-to-date knowledge in the Cloud
- **What we Don’t want:**
  - Computational overhead at the Data centres and sensors
  - Communication overhead
  - High network bandwidth

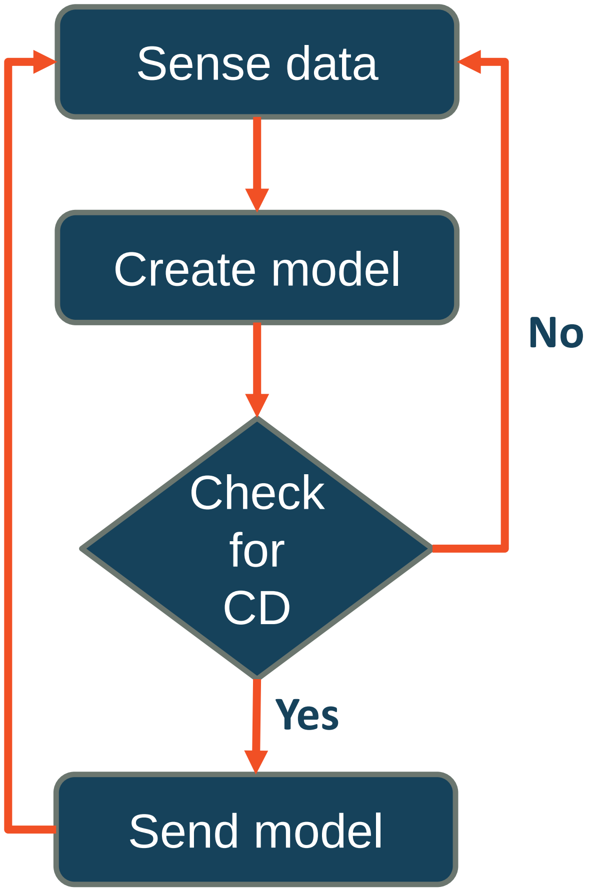


University of Glasgow
School of Computing Science

Problem statement (Cont'd)

- **What we Can do:**
  - **Gather** some of the sensed data in the sensor
  - **Create** a model from that data
  - **Communicate** the model
  - **Wait** until a Model Concept Drift (CD) has occurred
  - **Communicate** an updated model

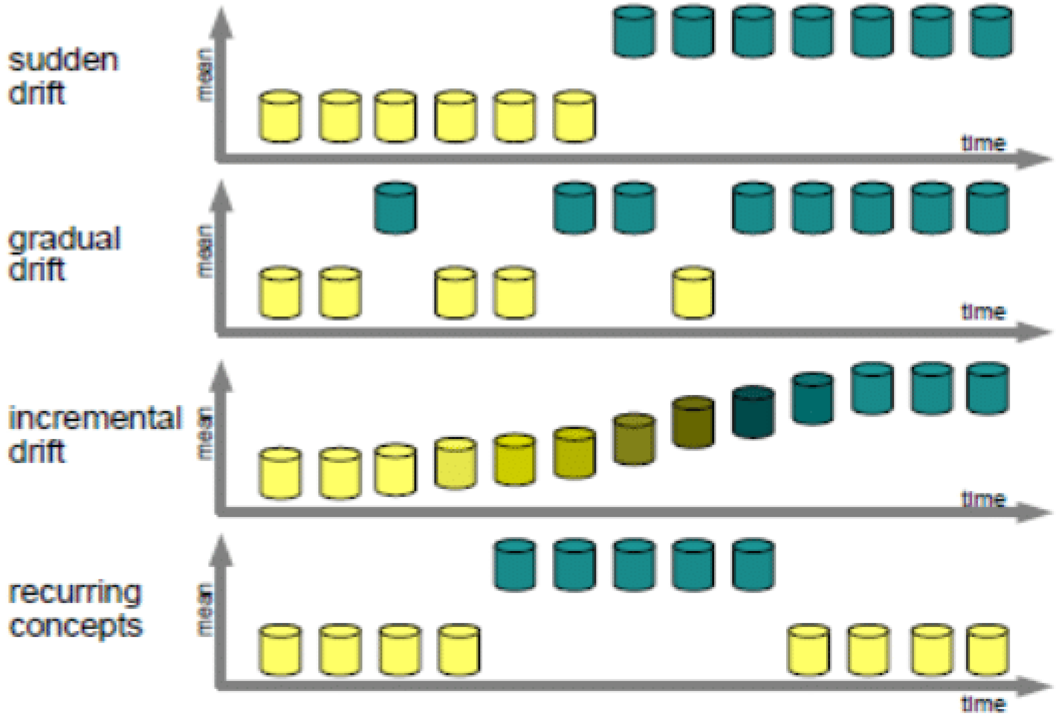
- **What we Will achieve:**
  - **Less communication** in the network
  - **Lower bandwidth** requirement
  - Data is delivered to the datacentre **partially analysed**
  - Data is **anonymised** by preserving the raw context at the sensor level



University of Glasgow
School of Computing Science

What is a Concept Drift?

- *Def.* A changing context which induces a change in the target concepts (Widmer & Kubat, 1996)



(Lemaire et. al., 2015)

University of Glasgow
School of Computing Science

Handling concept drift
Using Cumulative Sum (CuSum)[*]

- Using the Absolute Error Difference between current model and previously sent model on the most up-to-date subset of the data:
  - $\Delta e = |e - e'|$

- Make assumption on the **Good Distribution** and the **Bad distribution**

- Calculate the probability density functions of both:
  - P_{good} and P_{bad}

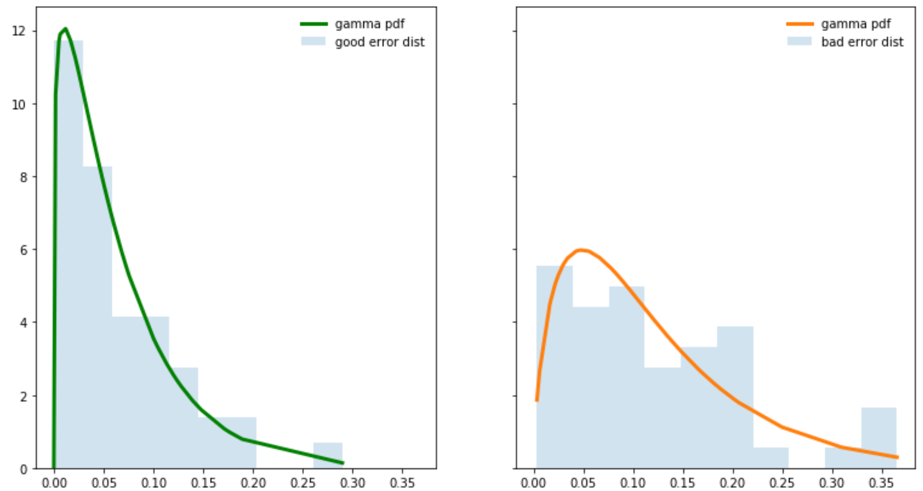


Fig. Good Distribution vs. Bad Distribution

[*] Invented by E. Page, Uni of Cambridge, 1954

University of Glasgow
School of Computing Science

Handling concept drift

Using Cumulative Sum (CuSum) (Cont'd)

- For each new Δe calculate the log-likelihood ratio:
  - $l_t = L_{\Delta e} = ln \frac{P_{\Delta e \mid bad}}{P_{\Delta e \mid good}}$
- Keep a record of all log-likelihood ratios sums:
  - $S[t] = \sum_{k=0}^{t} l_k$
- Decision value:
  - $g = S[t] - min_{0 \leq k \leq t-1}(S[k])$
- **Update criteria:** $g > h$, where $h \leftarrow threshold$

University of Glasgow
School of Computing Science

From CuSum to Optimal Stopping Theory

- **What does Optimal Stopping Theory deal with?**
  - How to estimate the best time to **stop** and gain the **highest reward** or suffer the least penalty?

- **Popular examples:**
  - The Secretary problem
  - The Blackjack Card game
  - The House Selling problem

- **Our problem:**
  - **Delay** sending an update as much as possible until a **change** in the distribution has occurred

University of Glasgow
School of Computing Science

From CuSum to OST (Cont'd)

- We use the cumulative sum principle on the absolute error rate, which is not allowed to exceed $\boldsymbol{\Theta}$
  - $Z_t = \Delta e_t = |e_t - e'_t|$
  - $S_t = \sum_{k=0}^{t} Z_k$

- Optimise t^* while maximising the reward function

$$V_t = \begin{cases} t, S_t \leq \Theta; & \textit{if we didn't update and the next } \Delta e \textit{ was low} \\ -B, S_t > \Theta; & \textit{if we didn't update and the next } \Delta e \textit{ was high} \end{cases}$$

- From here we can obtain the expected reward at time t

$$\Rightarrow \mathbb{E}[V_t] = t \cdot P(S_t \leq \Theta) + \left(-B \cdot P(S_t > \Theta)\right)$$
$$= (t + B) \cdot P(S_t \leq \Theta) - B$$

- Given the realisation of all random variables up to time t , let us have the **filtration**

$$\mathbb{F}_t = \{S_1, \ldots, S_t\} \cup \{Z_1, \ldots, Z_t\}$$

- From that we can express the conditional expectation of the future reward V_{t+1}

$$\mathbb{E}[V_{t+1} | \mathbb{F}_t] = (t + 1 + B) \cdot P(S_{t+1} \leq \Theta | \mathbb{F}_t) - B$$

University of Glasgow
School of Computing Science

From CuSum to OST (Cont'd)

- The equation $S_t = \sum_{k=0}^{t} Z_k$ can be used to express the sum at time $t+1$:

$$S_{t+1} = \sum_{k=0}^{t+1} Z_k = \sum_{k=0}^{t} Z_k + Z_{t+1} = S_t + Z_{t+1}$$

- This leads to finding that the **probability of the future sum** being less than Θ given our filtration equals the **cumulative distribution function** of Z being less than or equal to $\Theta - S_t$:

$$\begin{aligned} P(S_{t+1} \leq \Theta \mid \mathbb{F}_t) &= P(S_t + Z_{t+1} \leq \Theta \mid \mathbb{F}_t) \\ &= P(Z_{t+1} \leq \Theta - S_t \mid \mathbb{F}_t) \\ &= F_z(\Theta - S_t) \end{aligned}$$

- Now we can substitute for the conditional expectation of the future value and obtain:

$$\mathbb{E}[V_{t+1} \mid \mathbb{F}_t] = (t + 1 + B) \cdot F_z(\Theta - S_t) - B$$

- When the **currently obtained reward** is more than the **conditional expected future reward**, we want to send an updated model. That is when the following is satisfied:

$$V_t \geq \mathbb{E}[V_{t+1} \mid \mathbb{F}_t] \Leftrightarrow F_z(\Theta - S_t) \leq \frac{t+B}{t+1+B}$$

University of Glasgow
School of Computing Science

Other update policies and their rationale

- **Median-based policy**
  - **Update criteria:** $\Delta e_t > \alpha * median(\Delta e_1, \ldots, \Delta e_{t-1})$

- **Accuracy-based policy**
  - **Update criteria:** $\mathrm{e_t} > e_t'$

- **Random-based policy**
  - **Update criteria:** $generate(\{1|0\}, \mathrm{prob} = optimal\ policy)$

University of Glasgow
School of Computing Science

Experimenting using real life sensor data

- **GNFUV Unmanned Surface Vehicles Sensor Dataset**

  (Harth & Anagnostopoulos, 2018)
  - **collected data**: (humidity, temperature)
  - used with **Linear Regression** models

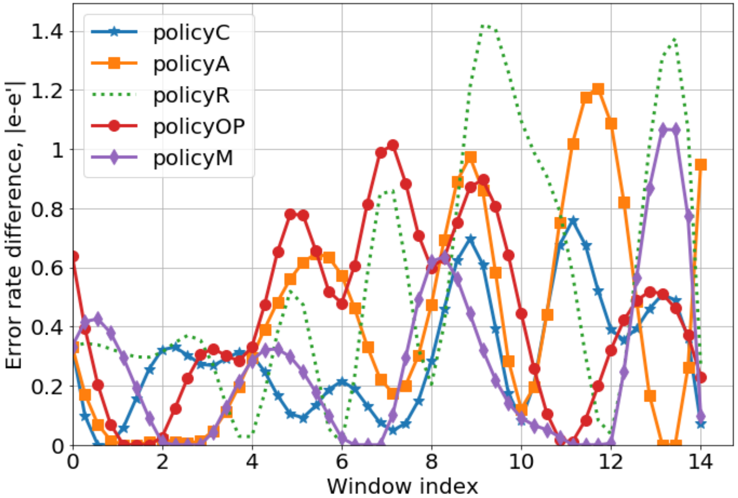
- **Gas Sensors for Home Activity Monitoring Dataset**

  (Huerta et. al., 2016)
  - **collected data**: (humidity, temperature, values from 8 metal-oxide sensors)
  - used with **Support Vector Regression** (with an RBF kernel) models
  - included artificial incremental concept drift in the data

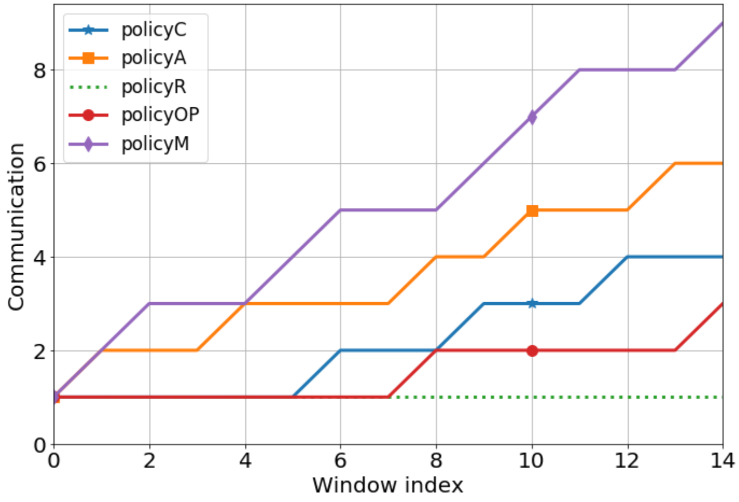
University of Glasgow
School of Computing Science

Our results for Linear Regression Models

- The <u>absolute error</u> for the optimal policy does not drastically deviate from the other policies



- The optimal policy saves on average **5 times** more <u>communication</u>



University of Glasgow
School of Computing Science

ANOVA test
for Linear Regression

- Looking for a statistical significant difference b/w policies:
  - waiting time

ANOVA p-value for waiting time		
sensor pi3	1.248e-30	<= 0.05
sensor pi4	7.893e-14	<= 0.05

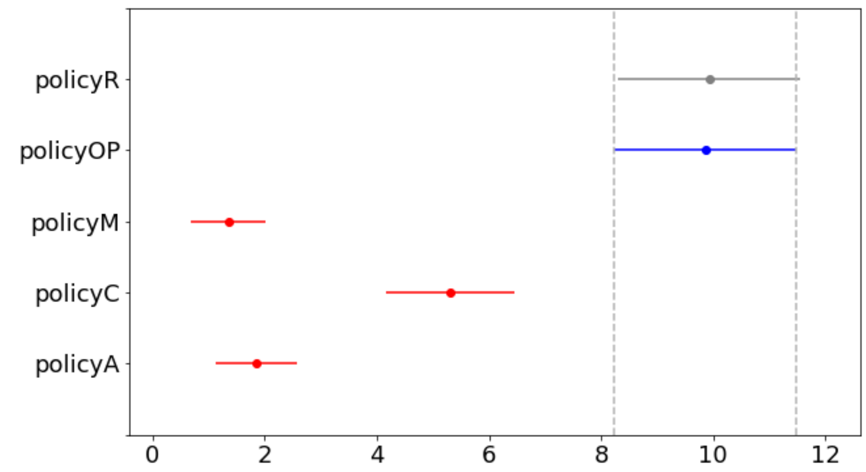
  - absolute error

ANOVA p-value for abs error		
sensor pi3	1.244e-13	<= 0.05
sensor pi4	2.723e-17	<= 0.05

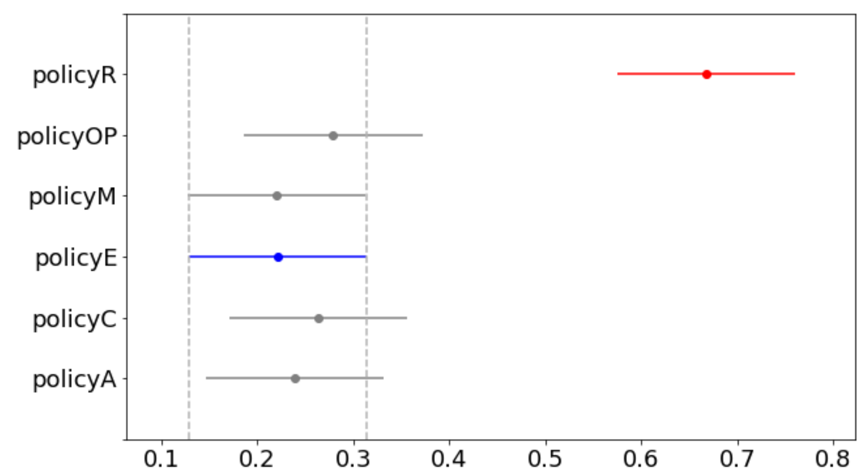
University of Glasgow
School of Computing Science

Tukey's HSD test for Linear Regression

The waiting time for the optimal policy has a higher mean
and
the difference is **statistically significant**.



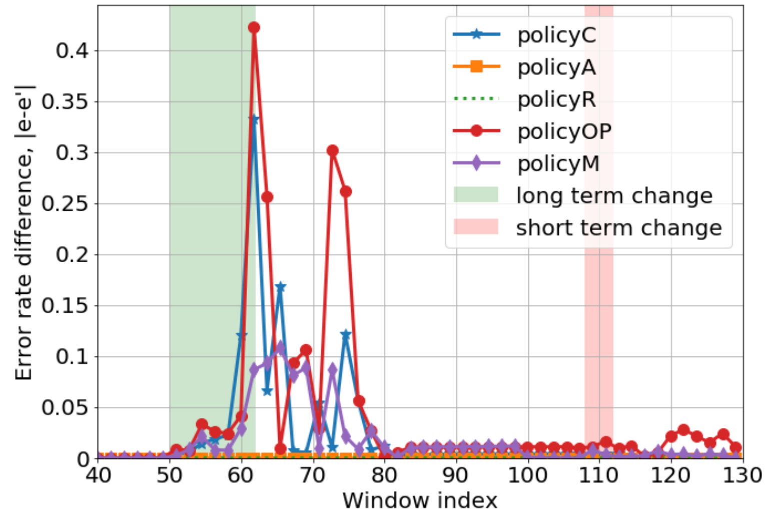
The difference in the absolute error between the optimal policy and when always having the most up-to-date model is **not statistically significant**.



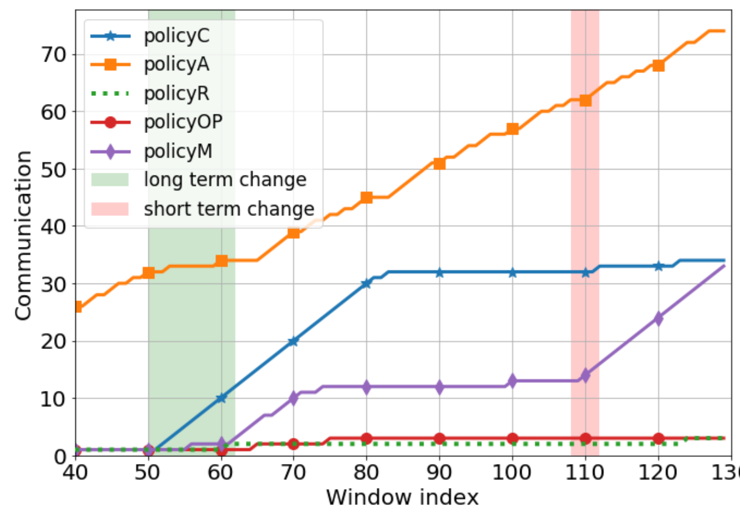
University of Glasgow
School of Computing Science

Our results
for Support Vector Regression Models

- The absolute error for the optimal policy and the cusum policy deviate the most from the accurate policy



- The optimal policy waits on average **30 times** longer than the other policies



University of Glasgow
School of Computing Science

ANOVA test
for Support Vector Regression Models

- Looking for a statistical significant difference b/w policies:
  - waiting time

ANOVA p-value for waiting time		
sensor R3	9.198e-18	<= 0.05
sensor R5	1.477e-32	<= 0.05

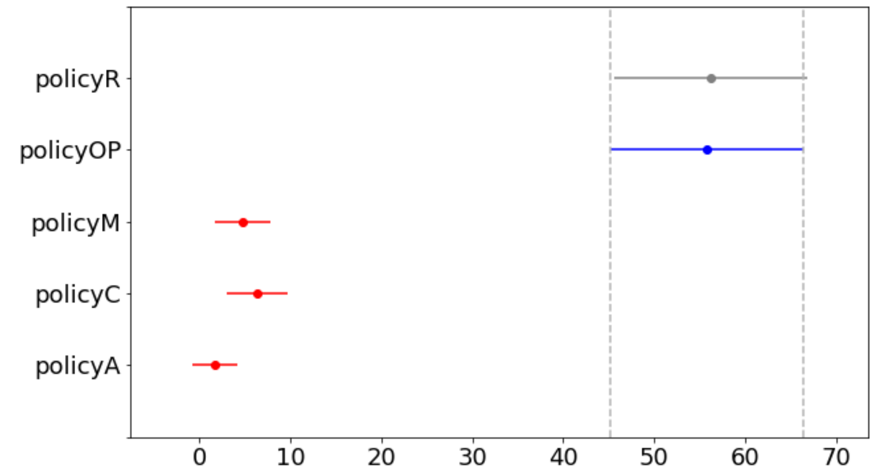
  - absolute error

ANOVA p-value for abs error		
sensor R3	6.003e-23	<= 0.05
sensor R5	1.062e-11	<= 0.05

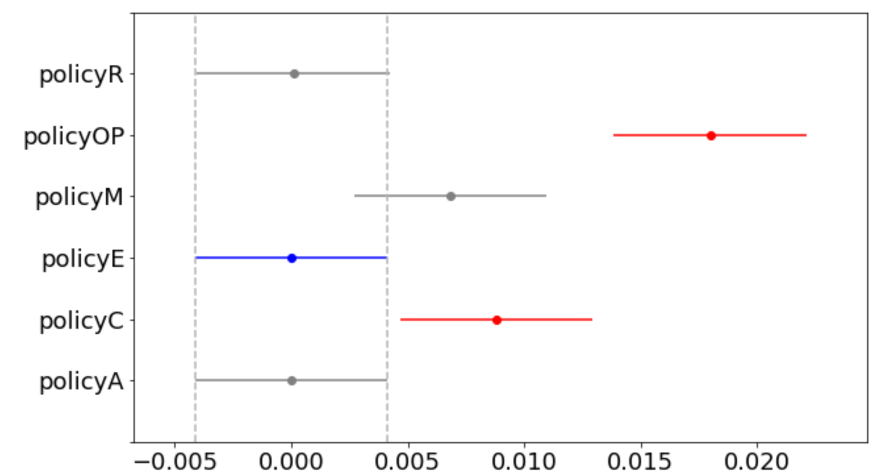
University of Glasgow
School of Computing Science

Tukey's HSD test for Support Vector Regression Models

The waiting time for the optimal policy has a higher mean and the difference is **statistically significant**.



The optimal policy has a higher absolute error rate and the difference is **statistically significant** from the rest of the policies.



University of Glasgow
School of Computing Science

Conclusion

Policy type	High quality prediction models	Lower quality prediction model
CuSum	high error and high communication	☺
Accuracy-based	☺	high communication
Optimal	high error	☺
Median-base	☺	high communication

University of Glasgow
School of Computing Science

ESSENCE
PERVASIVE & DISTRIBUTED INTELLIGENCE

Thank you!

Ekaterina (Katie) Aleksandrova