Online Machine Learning Techniques for Predicting Operator Performance

MASTER'S THESIS BY
AHMET ANIL PALA

Thesis Advisor: MAX HEIMEL
Thesis Supervisor: Prof. Dr. VOLKER MARKL

September 3rd, 2015

Fachgebiet Datenbanksysteme und Informationsmanagement
Technische Universität Berlin

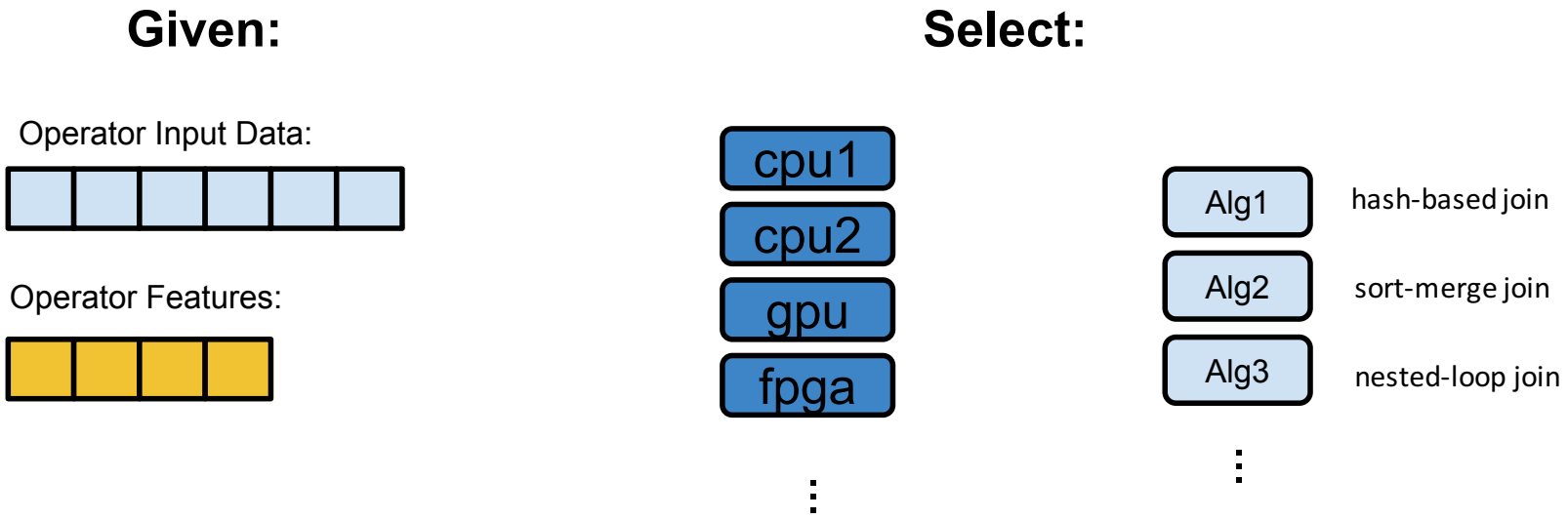
http://www.dima.tu-berlin.de/

Agenda

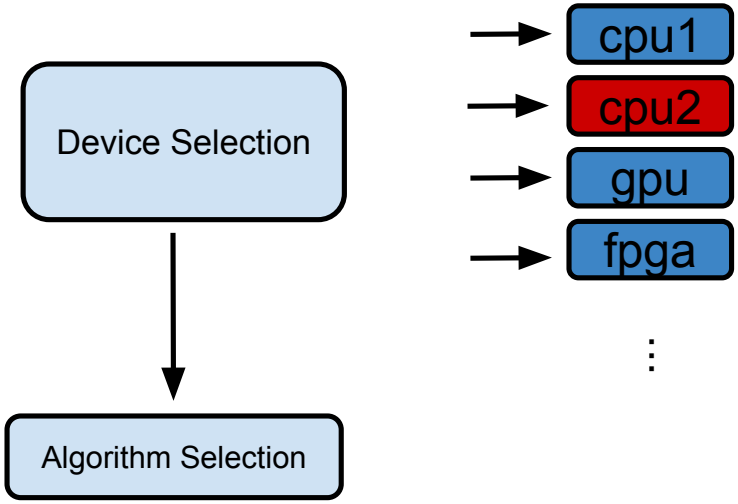
- Motivation
- Problem
- Existing Methods
- Approach
- Evaluation
- Summary

Tuning a hardware-oblivious database

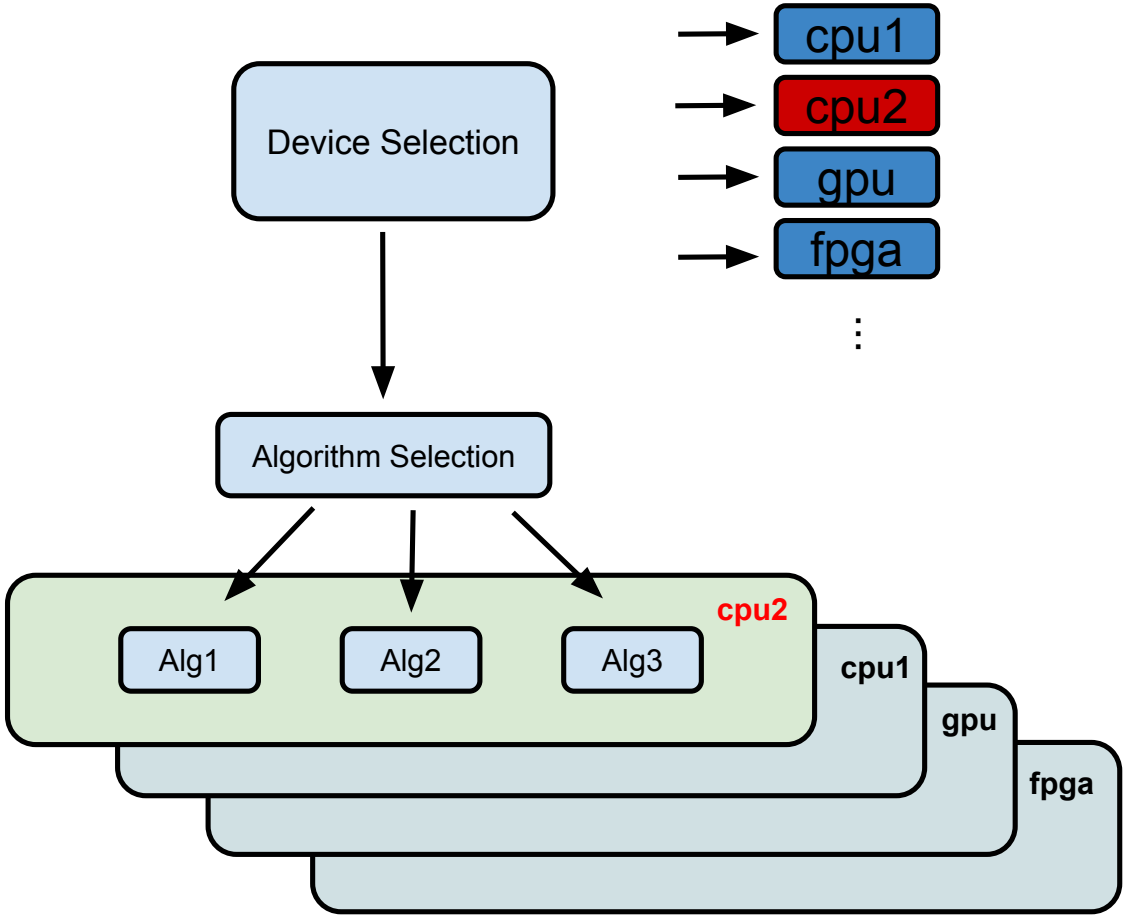
- Tuning a hardware-oblivious database involves two decision-making tasks
  - Device Selection
  - Algorithm Selection



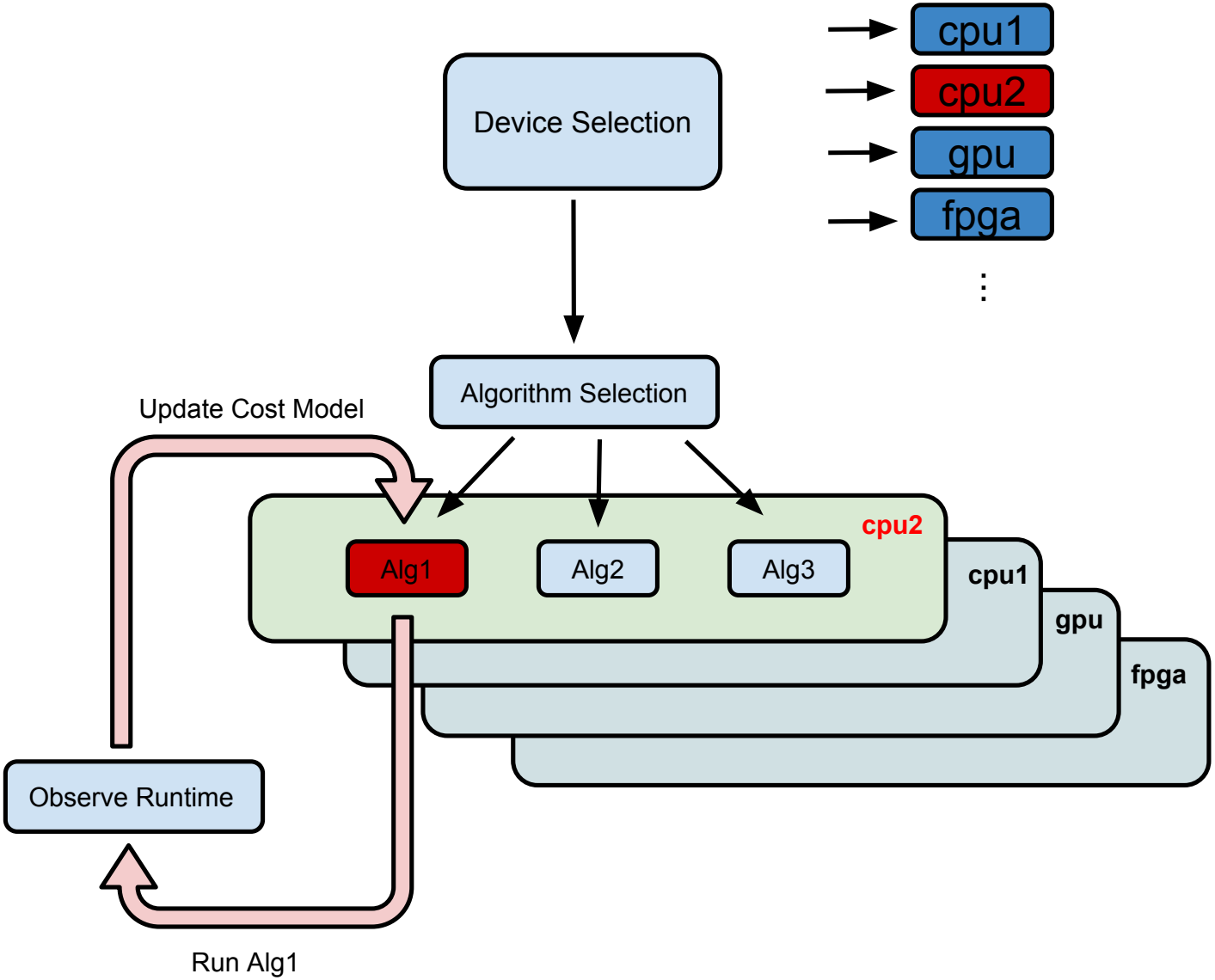
Device Selection



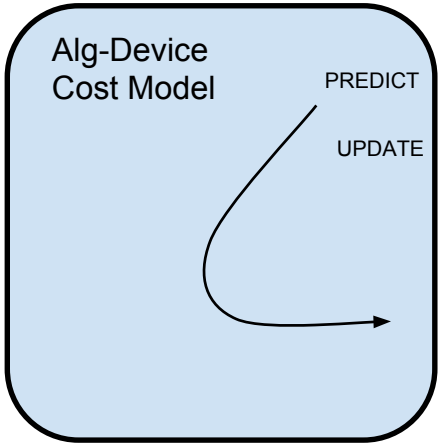
Algorithm Selection

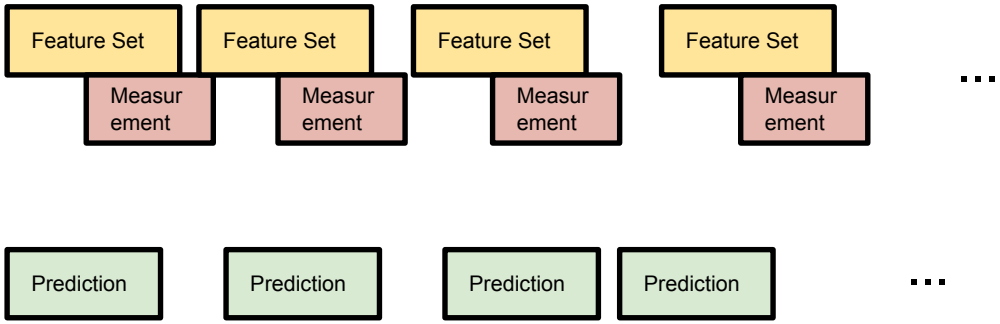


Algorithm Selection

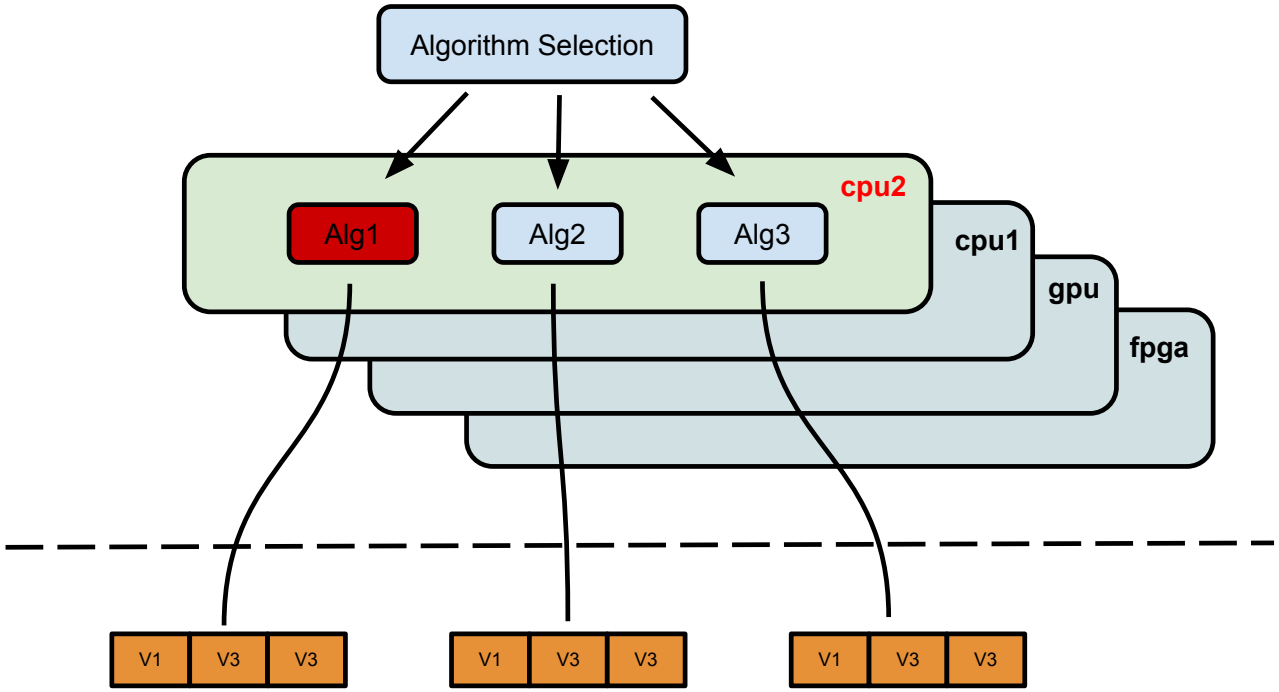


Learning Scenario





Concept Drifts

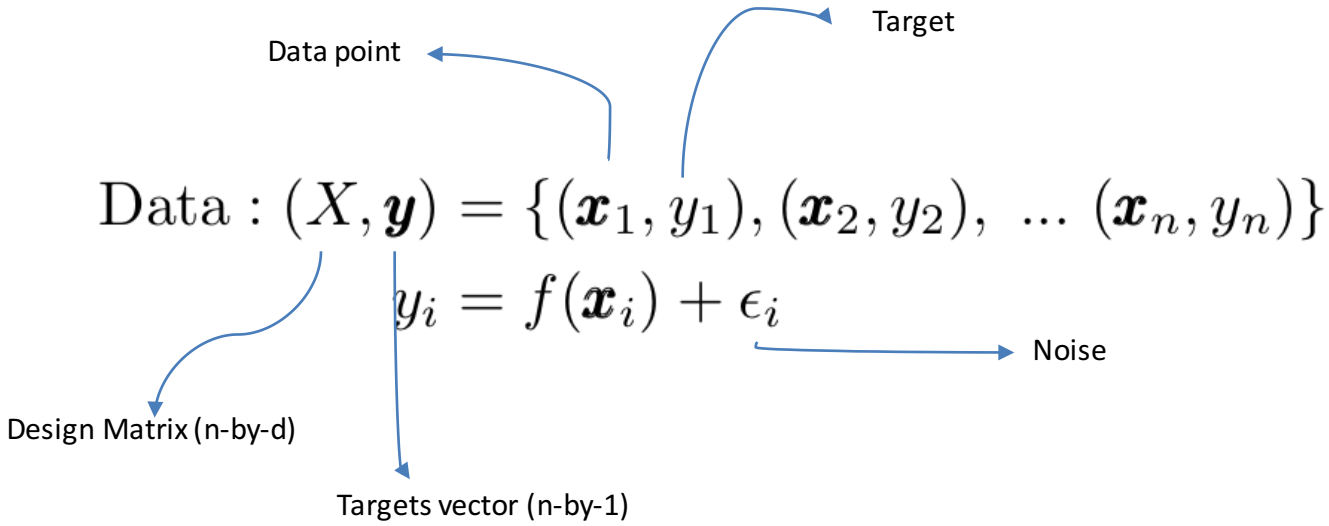


Variant Library for Alg1:

Variant Library for Alg2:

Variant Library for Alg3:

Online Regression Problem



Find $y_{n+1} | (X, \boldsymbol{y})$

Learning Model Requirements

- High Accuracy
- Robust to Abrupt Concept Changes
- Robust to Measurement Noise
- Provide Estimation Bounds
- Constant Space
- Efficiency (≤ 1ms/item)

Learning Models

- Parametric models
  - MAP-method
  - MLE-method
- Non-parametric models
  - Gaussian Process Regression
  - Kernel Regression

Parametric models

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \boldsymbol{x} = w_1 x_1 + w_2 x_2 + w_3 x_3 + \ldots$$

$$f(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x}) = w_1 x_1^2 + w_2 x_1 x_2 + w_3 x_1 x_3 + w_4 x_2 x_3 + \ldots$$

Aim is to find $\boldsymbol{w}$!

MLE/MAP methods

$$\boldsymbol{w}_{MLE} = (XX^\top)^{-1}X\boldsymbol{y} \qquad \boldsymbol{w}_{MAP} = (XX^\top + \sigma_y^2\Sigma_w^{-2})^{-1}X\boldsymbol{y}$$

regularization term

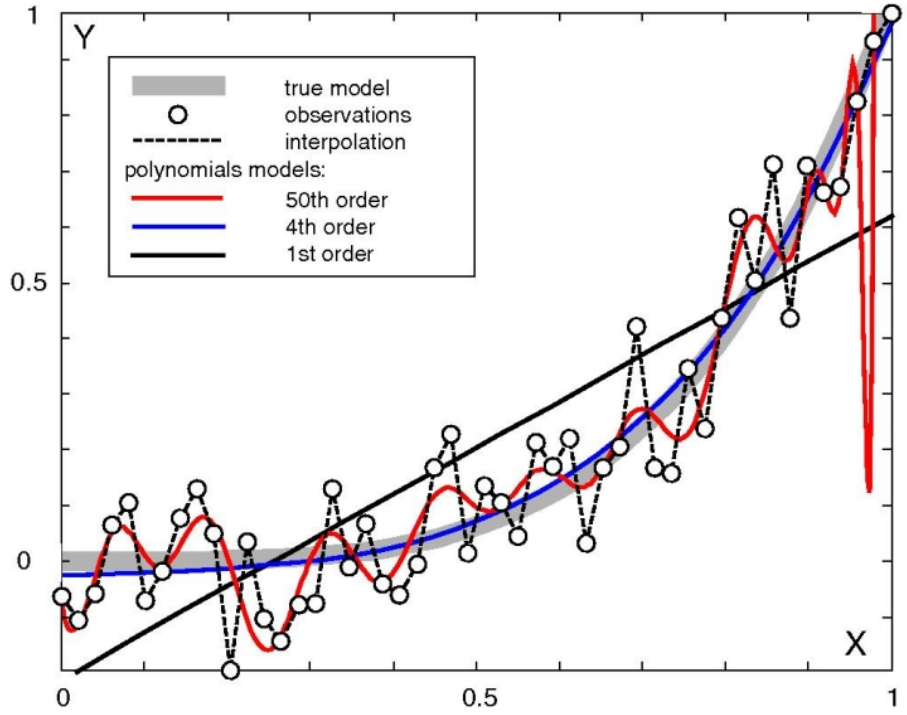
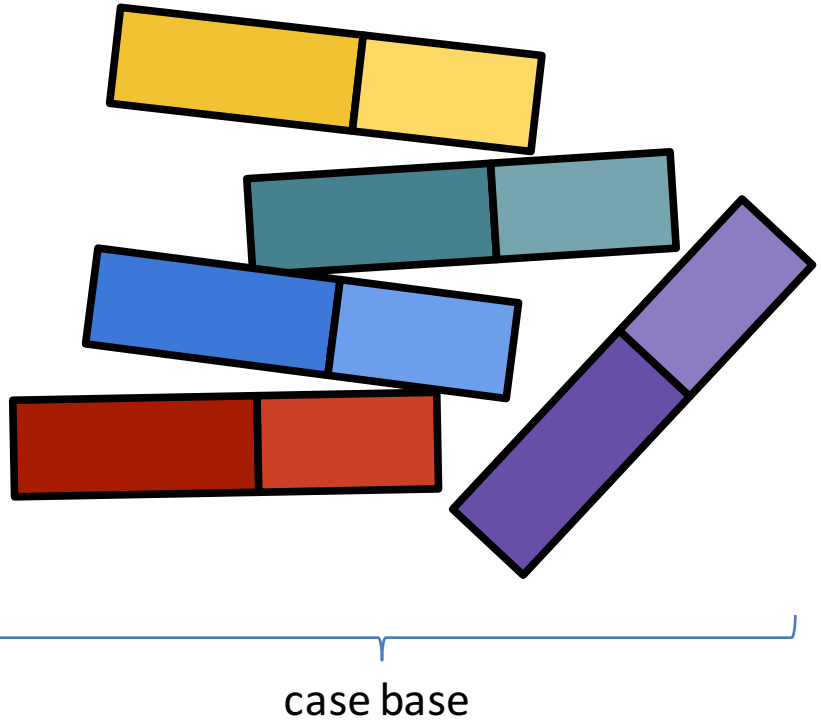
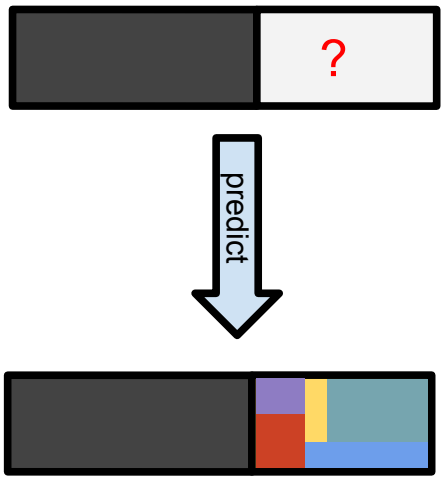


Image taken from http://pvermees.andropov.org/noble/disc/discPaper/node19.html

Non-parametric Models

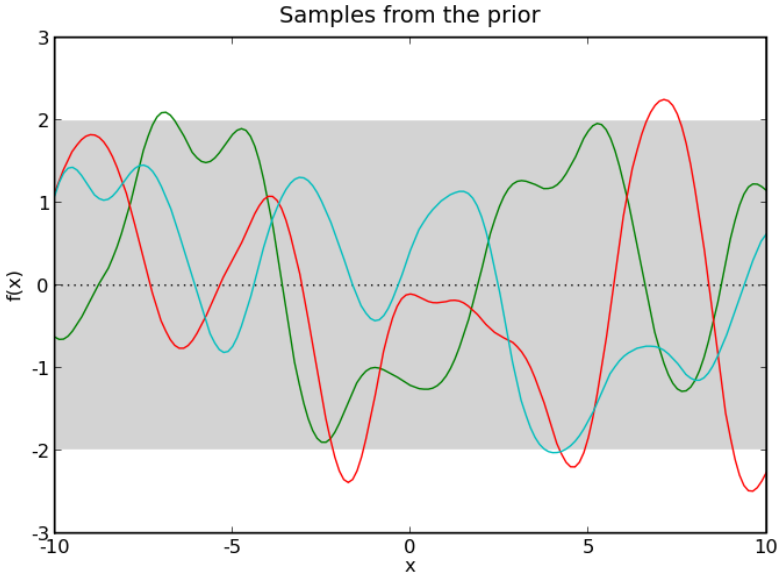
- Data-driven
- Accumulative



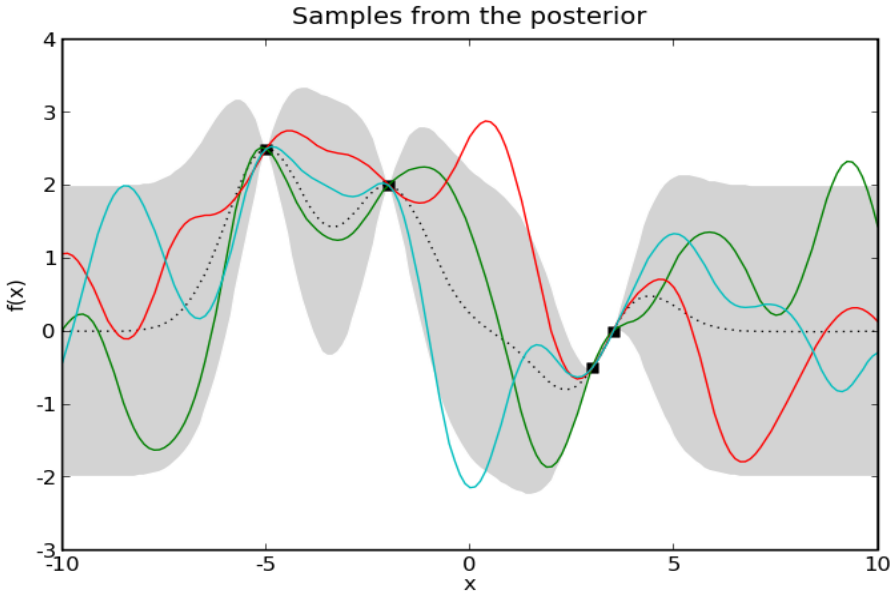


Gaussian Process Regression

- All targets as a single point in Gaussian distribution
- Considering the functions that can map data points to the observed targets in the function space



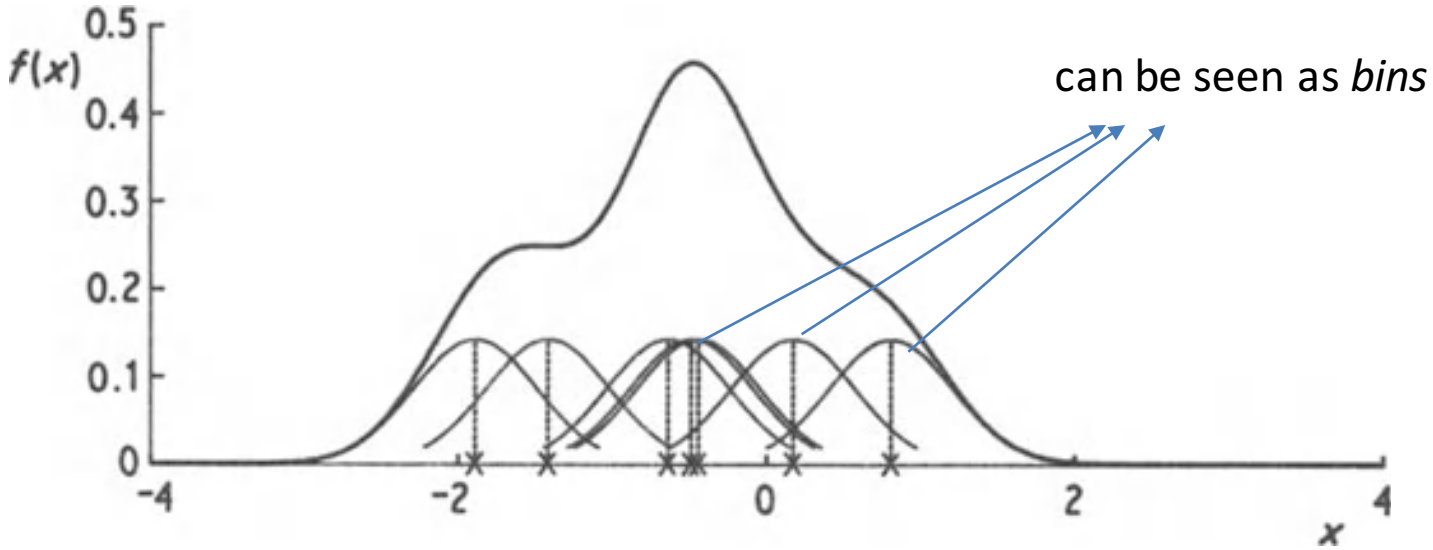
$$f(\mathbf{x}) \sim \mathcal{GP}\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big)$$



Images taken from http://pythonhosted.org/infpy/gps.html

Kernel Regression

- Based on kernel-density estimation
- Generalization of k-NN method

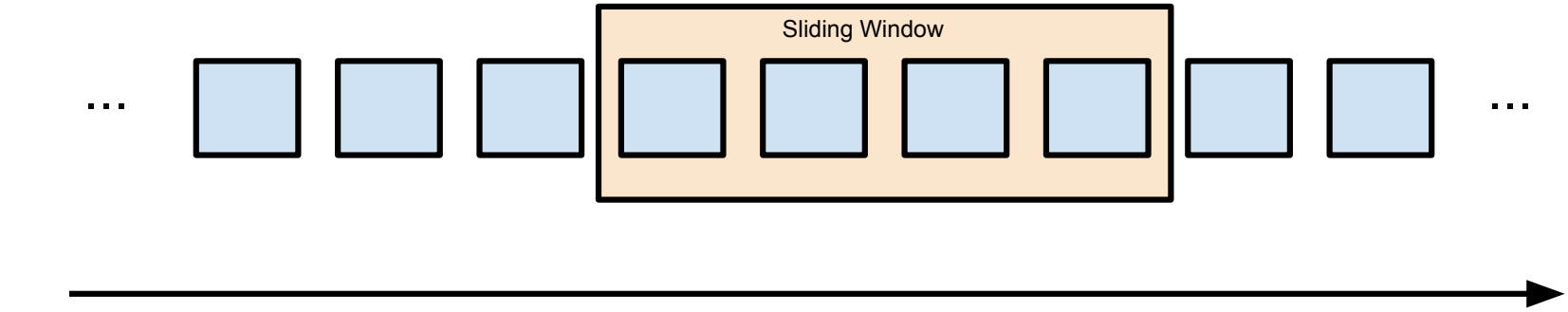


Graph taken from Silverman, Bernard W. *Density estimation for statistics and data analysis*.

Approach

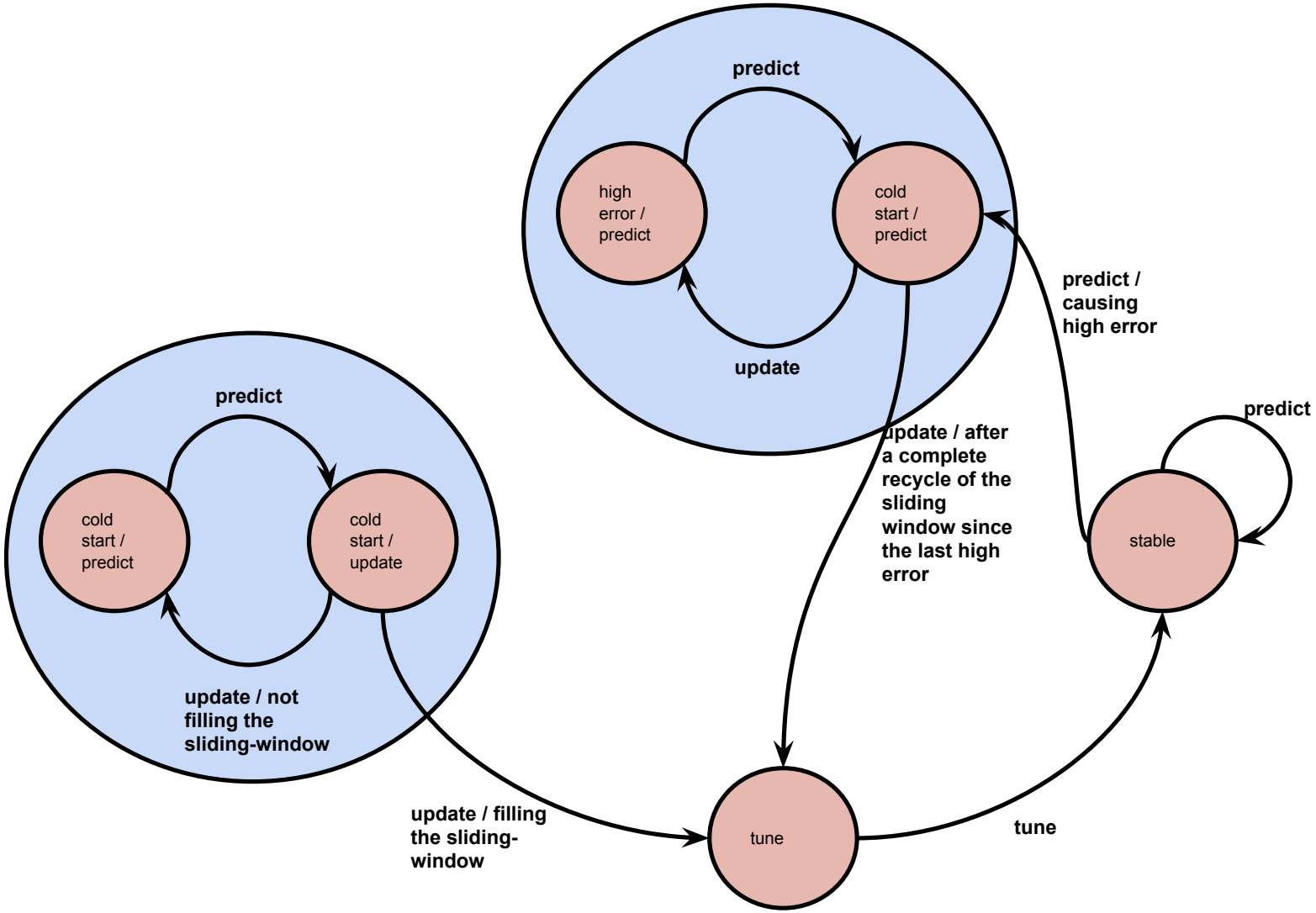
- Incremental update mechanism
- Removal of obsolete examples (downdate)

Sliding-windowed design

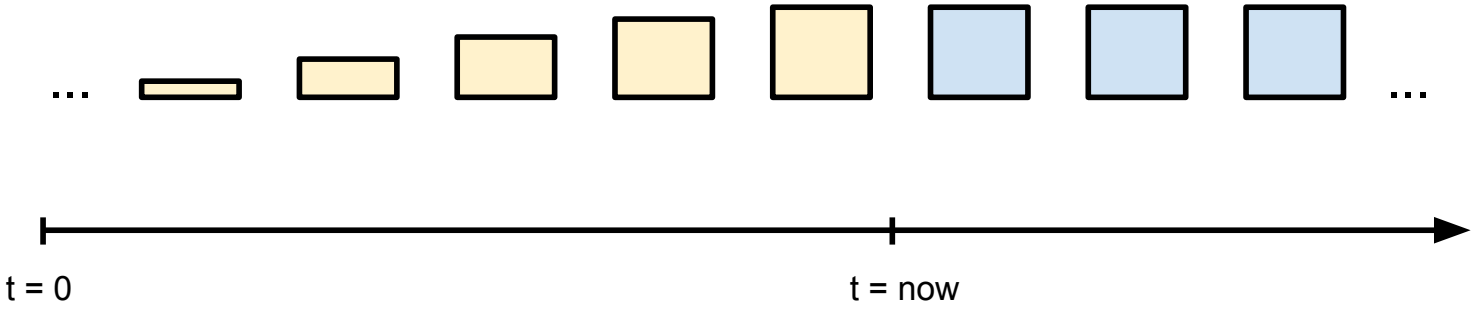


t = 0

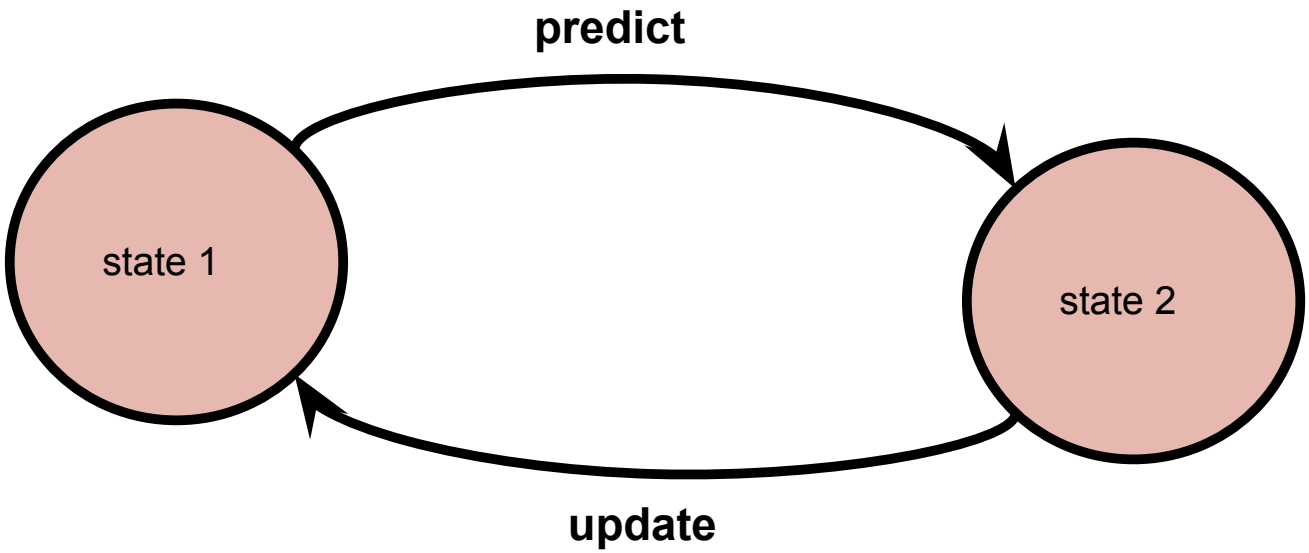
Sliding-windowed design – state transition diagram



Forgetting-based design



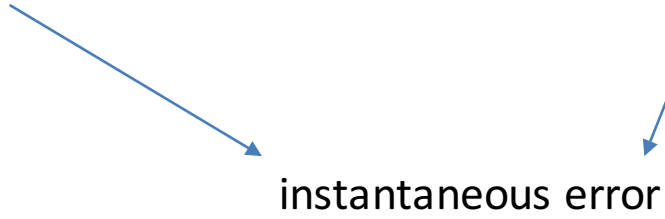
Forgetting-based design – state transition diagram



Implementation of 'predict' for MLE/MAP learners

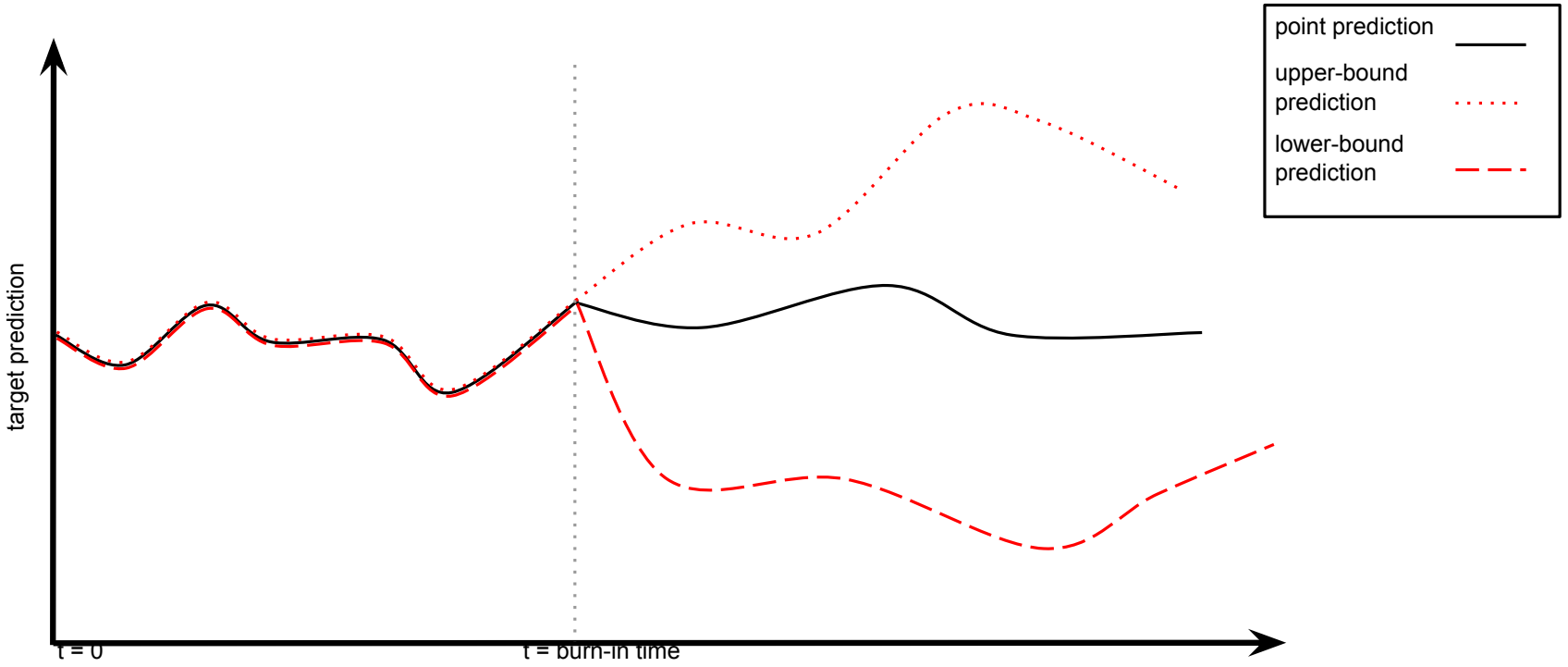
$$\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i$$

$$y_i \in [\boldsymbol{w}^\top \boldsymbol{x}_i \pm z_{1-\frac{\alpha}{2}} \sqrt{\underbrace{s^2}\boldsymbol{x}_i^\top (XX^\top)^{-1}\boldsymbol{x}_i + \underbrace{s^2}]}$$



3-learner ensemble for estimating prediction bounds for MLE/MAP learners

$$\hat{y}_i = \boldsymbol{w}^\top \boldsymbol{x}_i$$



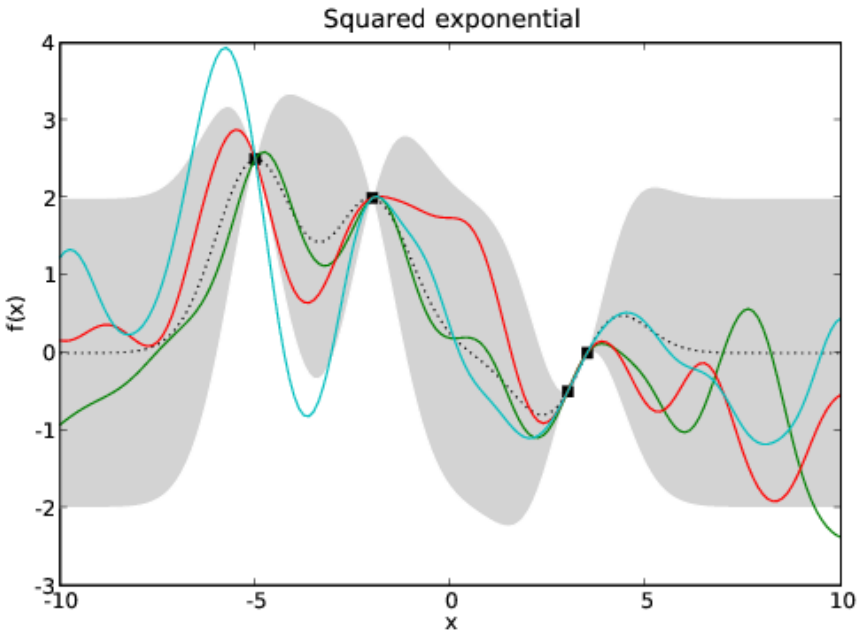
Implementation of 'tune' for GP Regression

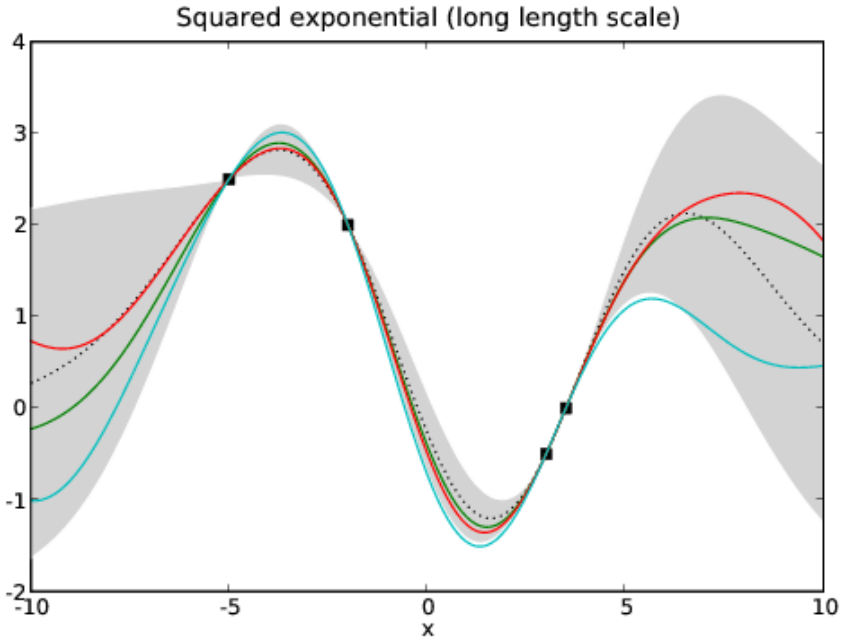
$$k(\boldsymbol{x}_p, \boldsymbol{x}_q) = \sigma_w^2 exp(-\frac{1}{2}(\boldsymbol{x}_p - \boldsymbol{x}_p)^\top D^{-2}(\boldsymbol{x}_p - \boldsymbol{x}_p) + \sigma_y^2 \delta_{pq}$$

signal variance

characteristic lengthscale matrix

noise variance

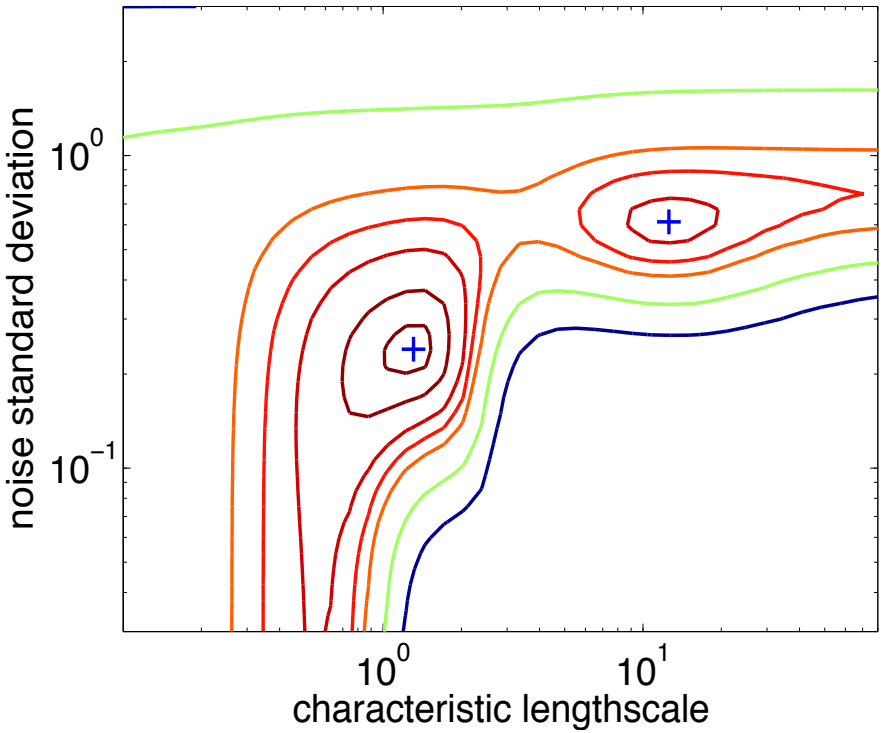




Images taken from http://pythonhosted.org/infpy/gps.html

Implementation of 'tune' for GP Regression

$$log(p(\boldsymbol{y}|X,\theta)) = \underbrace{-\frac{n}{2}log(2\pi)}_{\text{constant}} \underbrace{-\frac{1}{2}log(|K(X,X)|)}_{\text{complexity penalty}} \underbrace{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{M})^\top K(X,X)(\boldsymbol{y}-\boldsymbol{M})}_{\text{data-fit}}$$



Graph taken from Williams, Christopher KI, and Carl Edward Rasmussen. "Gaussian processes for machine learning." *the MIT Press* 2.3 (2006): 4.

Implementation of 'tune' for Kernel Regression

$$\hat{f}(\boldsymbol{x}) = \frac{1}{n|H|} \sum_{i=1}^{n} k((\boldsymbol{x}_i - \boldsymbol{x})H^{-1})$$

bandwidth matrix

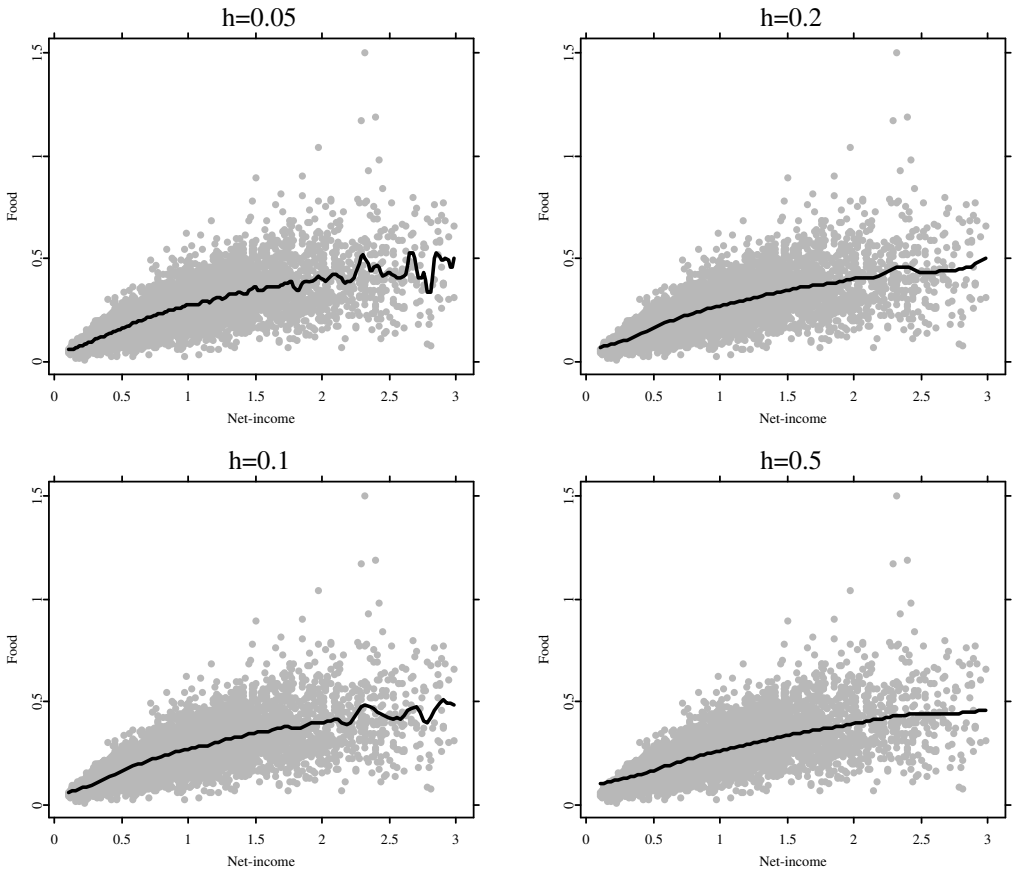


Image taken from http://pvermees.andropov.org/noble/disc/discPaper/node19.html

Evaluation

- 52 Online learners & 12 Batch learners
- Rigorous testing on synthetic streams

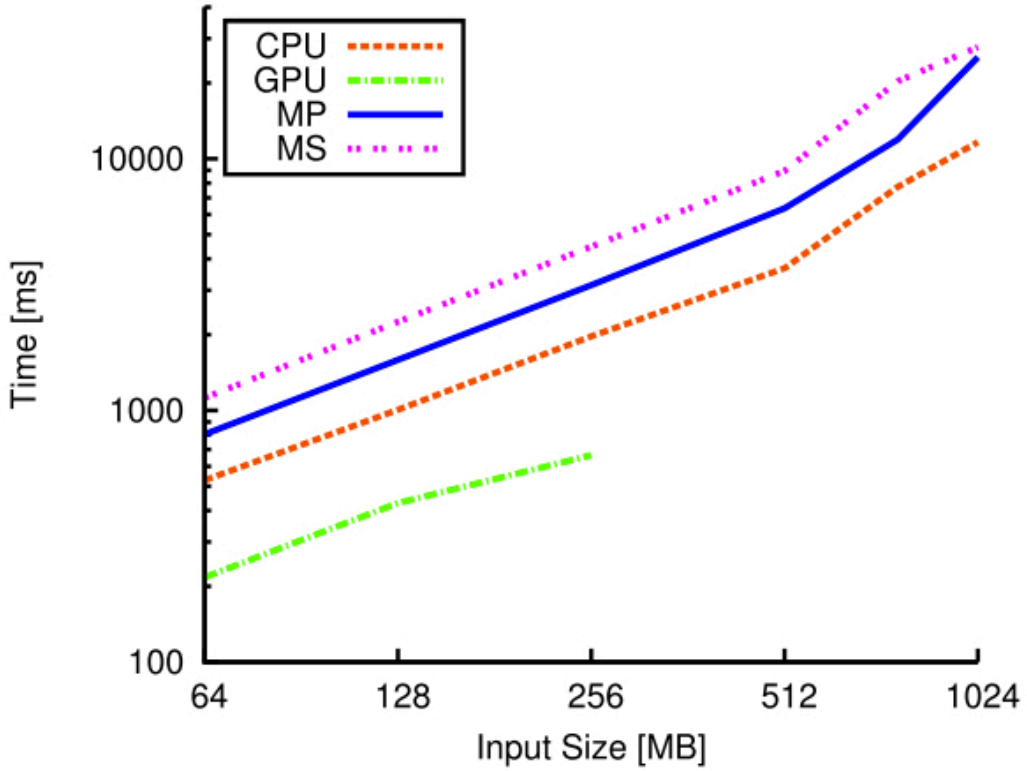


Image taken from Heimel, Max, et al. "Hardware-oblivious parallelism for in-memory column-stores." *Proceedings of the VLDB Endowment* 6.9 (2013): 709-720.

Evaluation

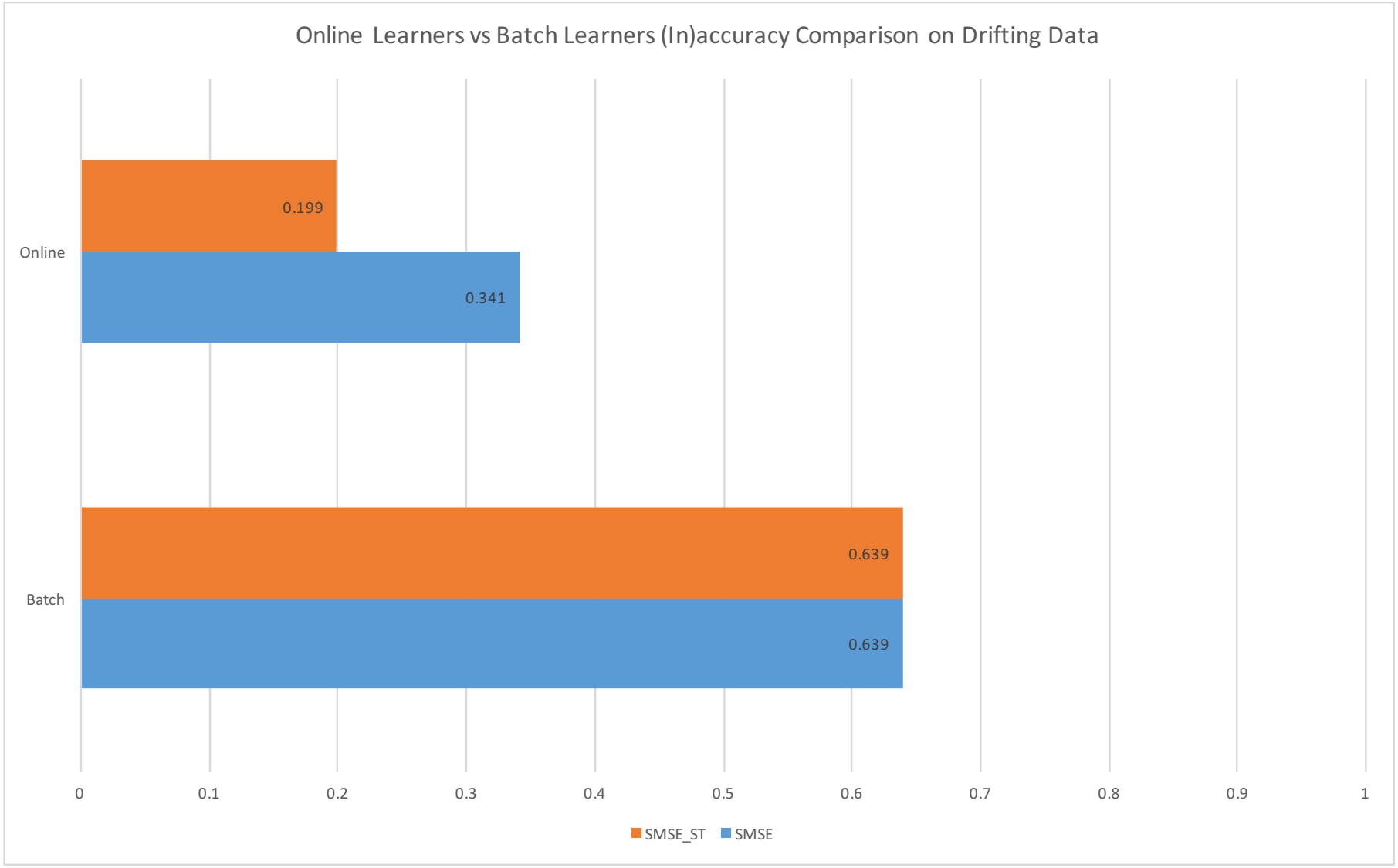
- 52 Online learners & 12 Batch learners
- Rigorous testing on synthetic streams
- Evaluation dimensions: accuracy, prediction bounds quality, time efficiency

Accuracy	Prediction Bounds	Time Efficiency
RMSE	ICR	ATPI (ms/item)
RMSE_ST	AIW	Max. Data Rate (item/ms)
SMSE	SAIW	
SMSE_ST		

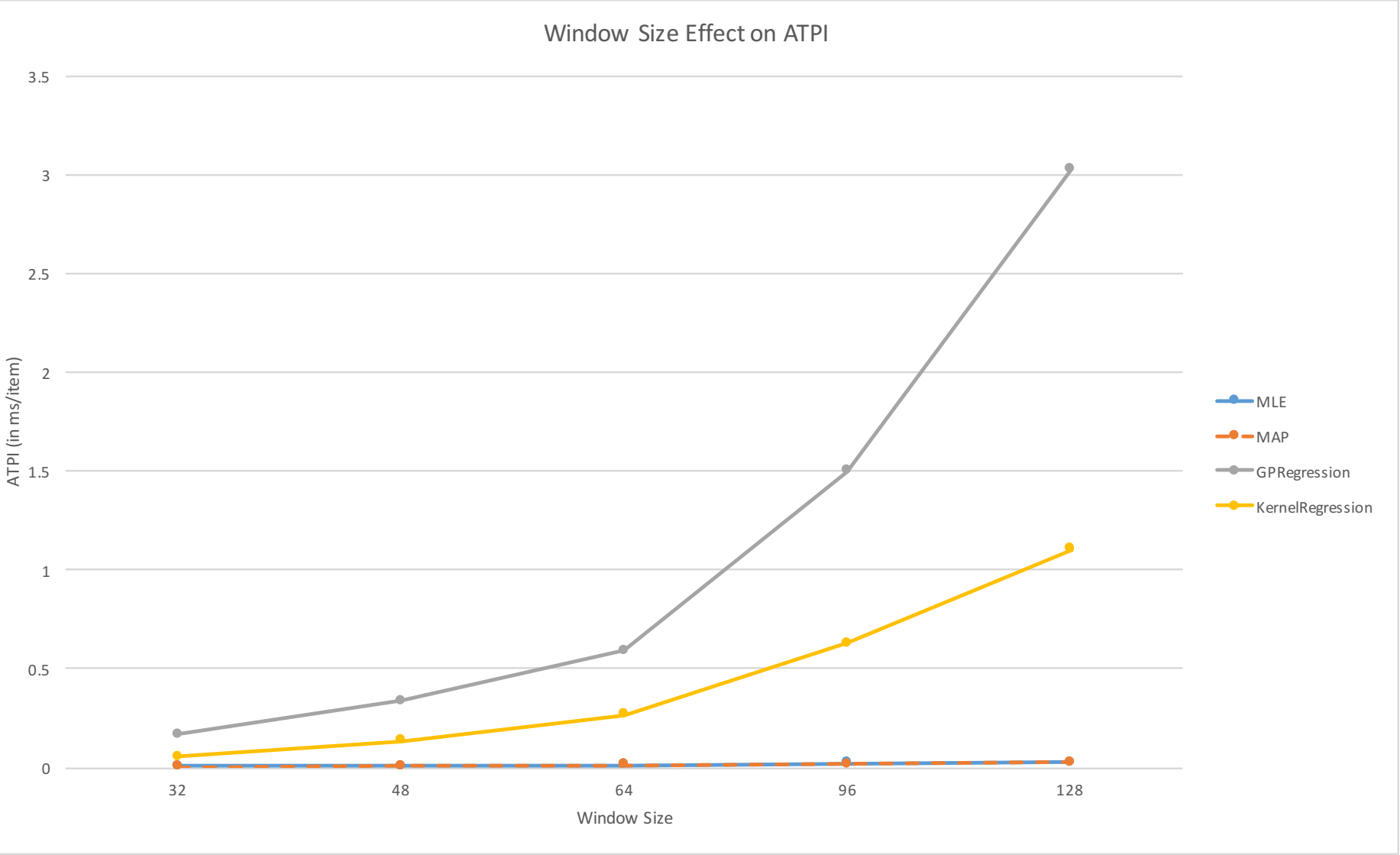
Evaluation

- 52 Online learners & 12 Batch learners
- Rigorous testing on synthetic streams
- Evaluation dimensions: accuracy, prediction bounds quality, time efficiency
- Prequential method and windowed prequential method

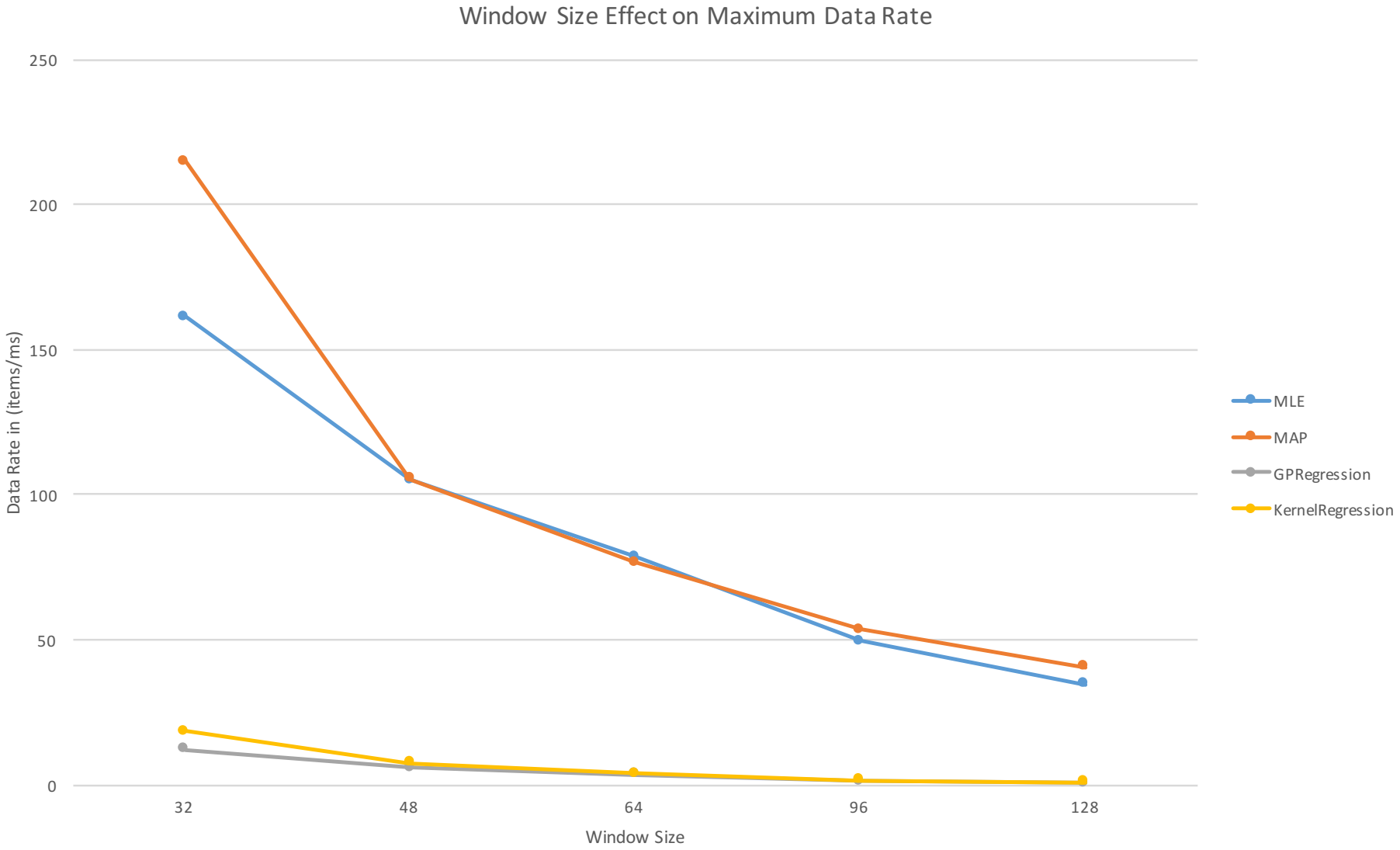
Online Learners vs Batch Learners



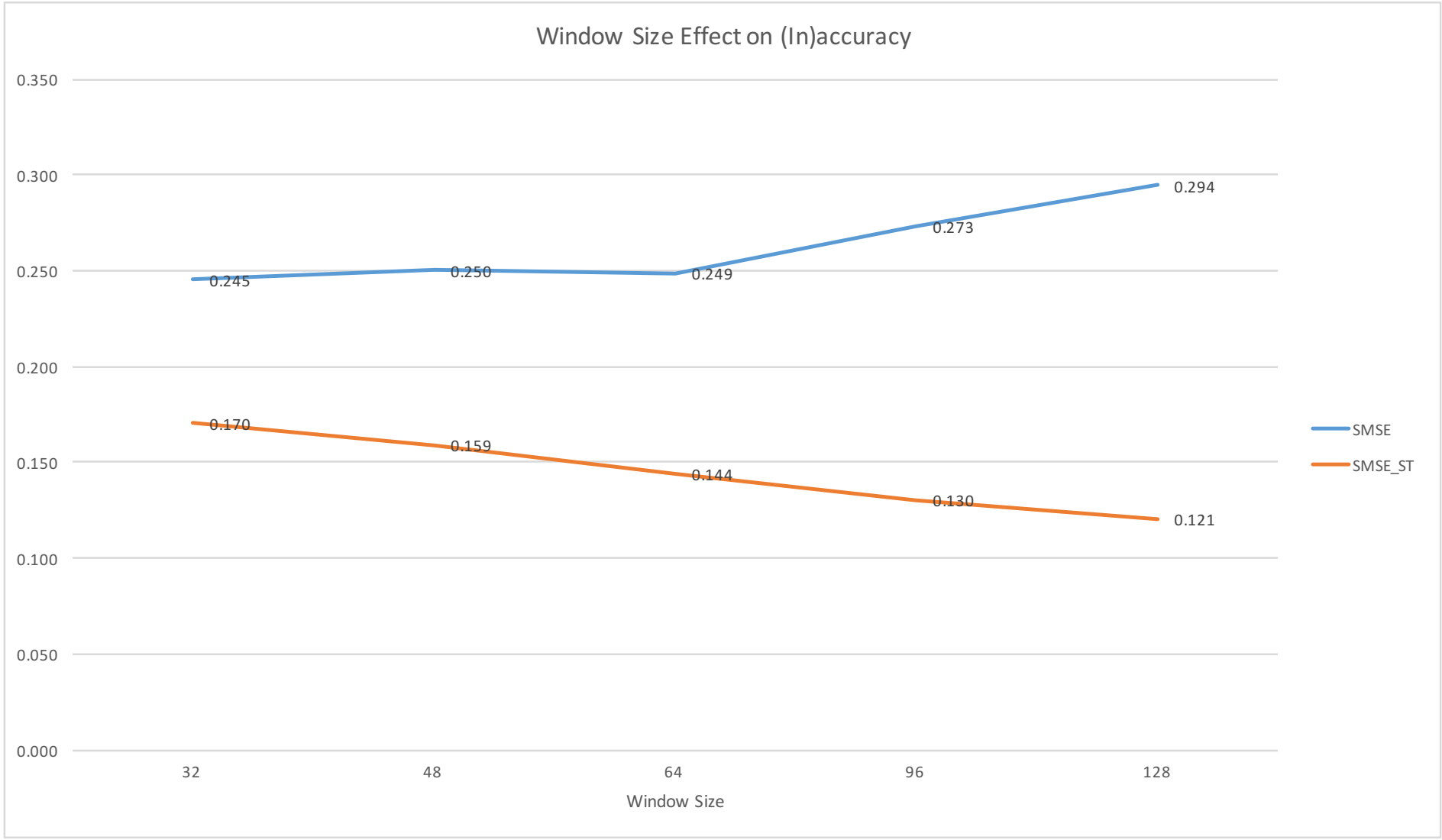
Window size on efficiency



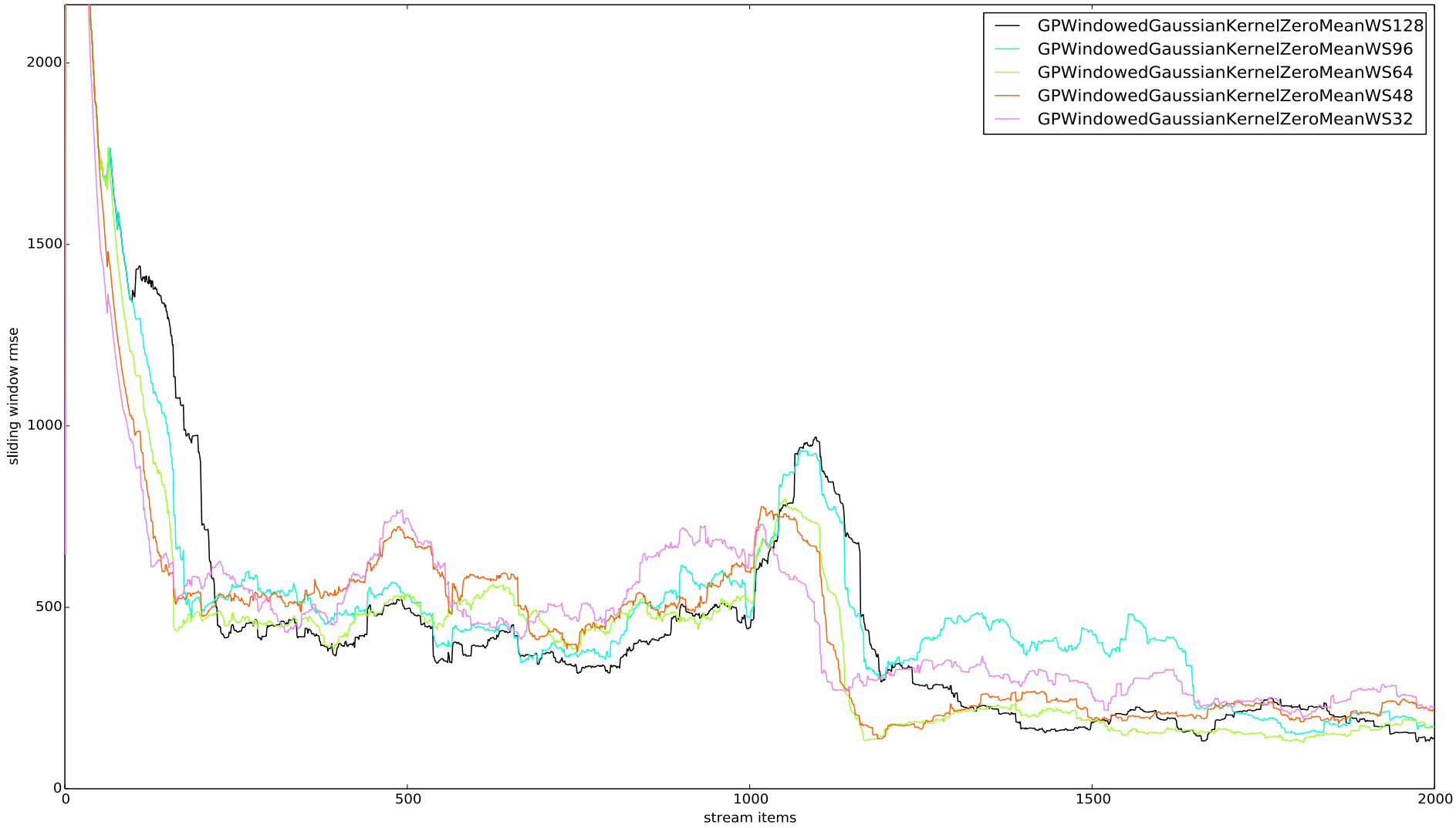
Window size on efficiency



Window size on accuracy

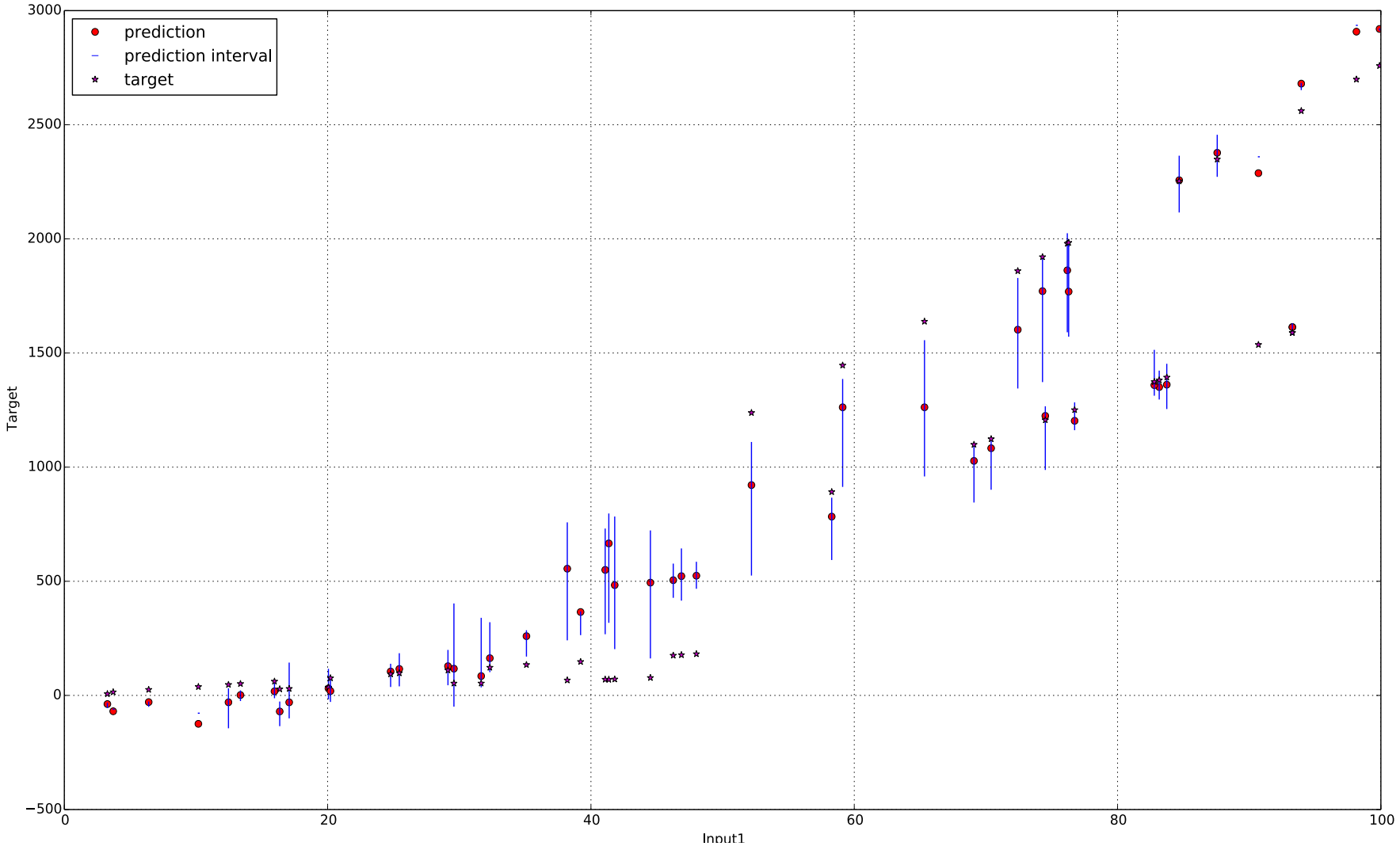


Window size on accuracy



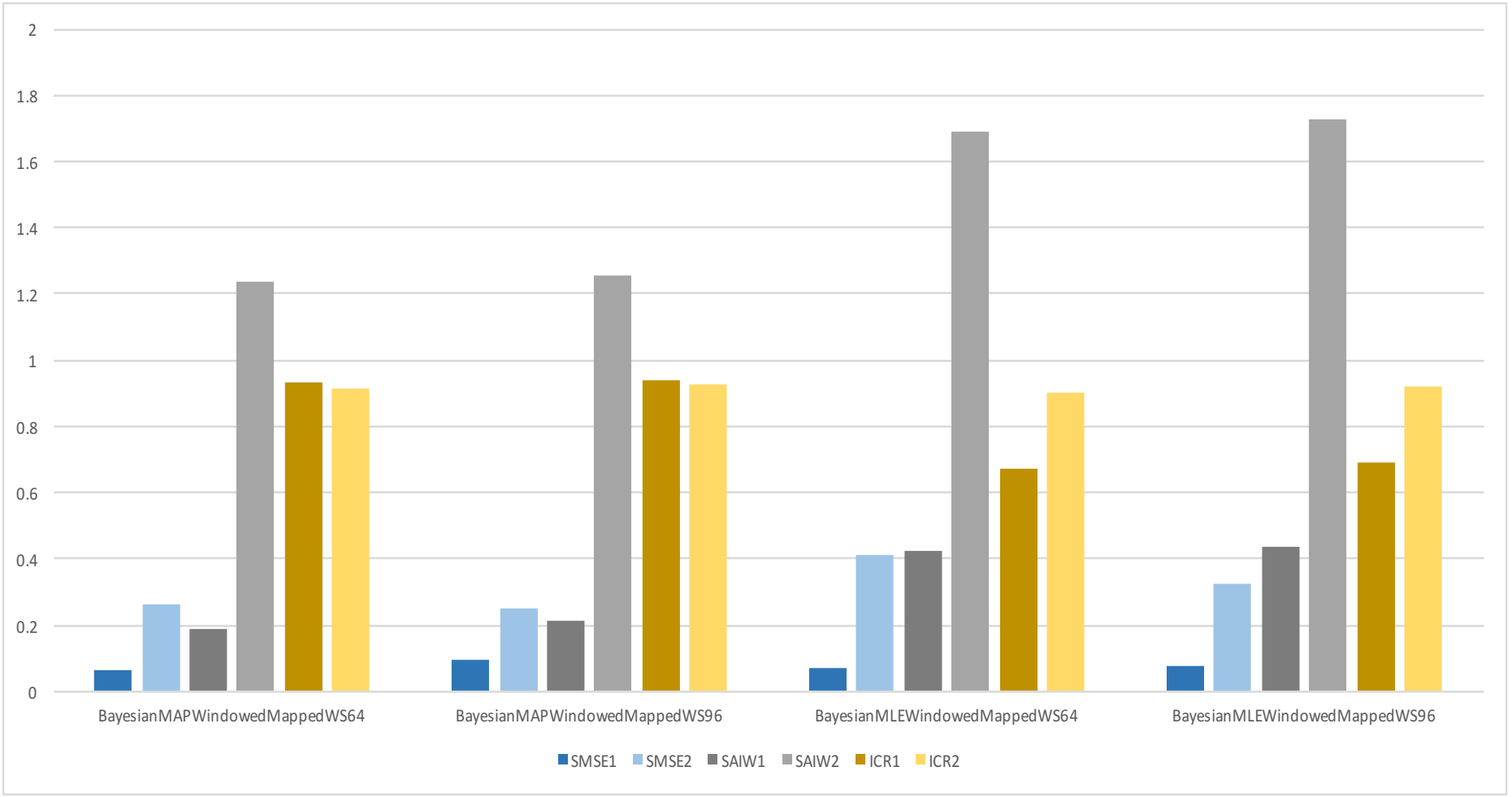
GPWindowedGaussianKernelZeroMean variants with different window sizes on **SYNTH_D_CD_2000_4_10_1_24**

3-Ensemble Prediction Bounds Estimation Method Performance

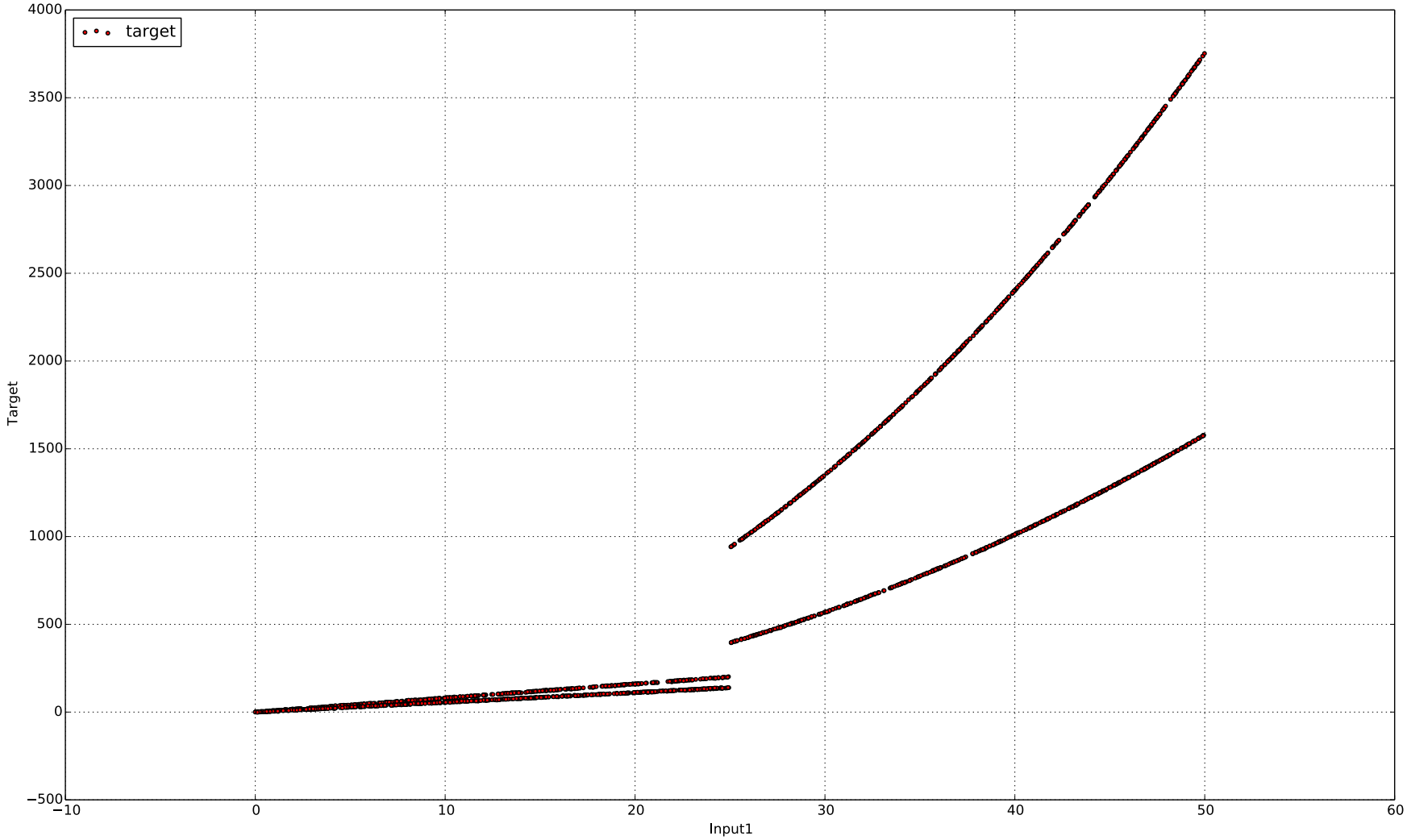


50 sample predictions of **BayesianMAPForgettingMapped_FF0.05** on **SYNTH_D_CD_2000_1_100_1_12**

MAP/MLE on streams with continuous and non-discontinuous data

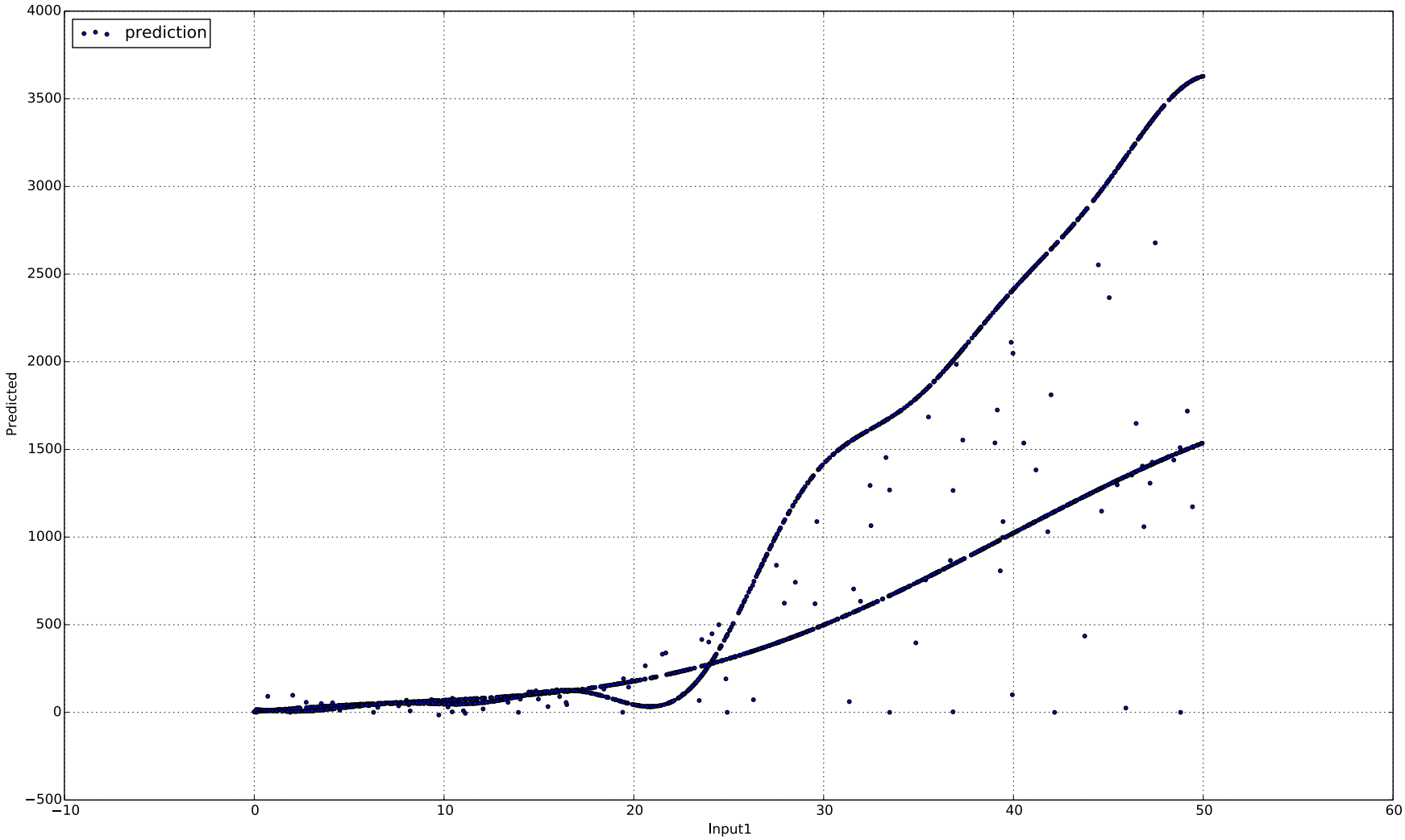


Targets vs Predictions by the 5 selected online learners



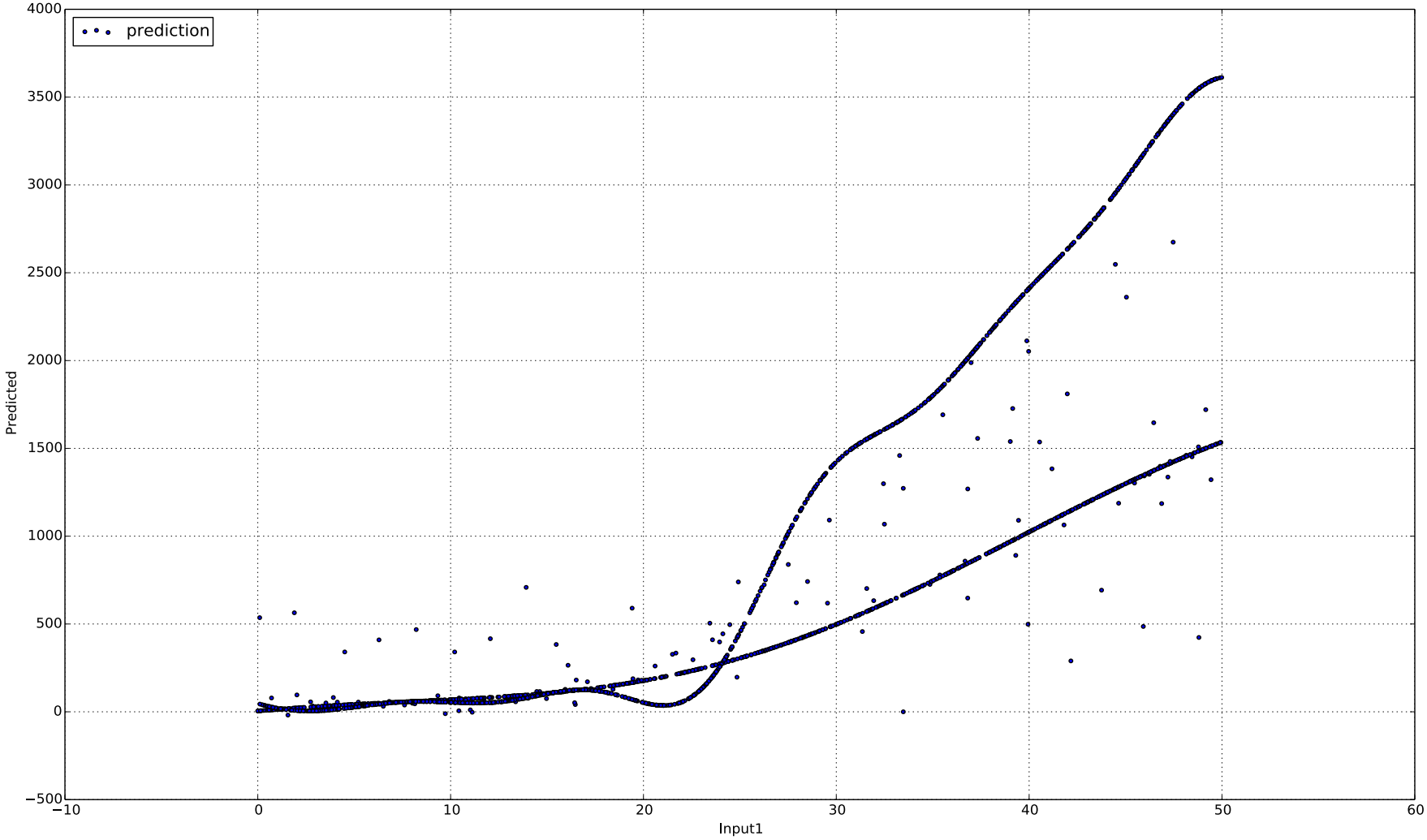
SYNTH_D_CD_2000_1_50_1_13

Targets vs Predictions by the 5 selected online learners



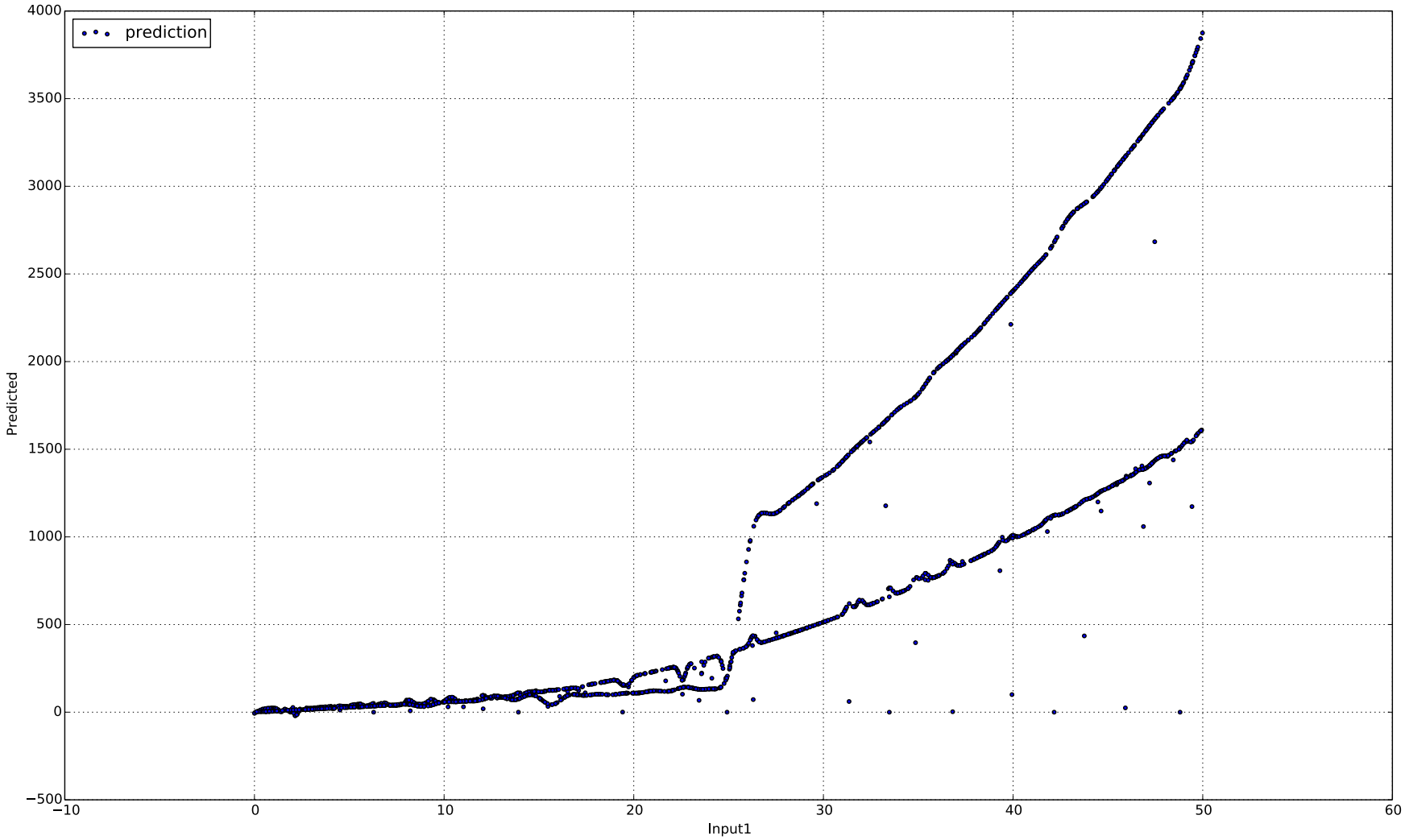
GPRegressionZeroMean_WS64 on **SYNTH_D_CD_2000_1_50_1_13**

Targets vs Predictions by the 5 selected online learners



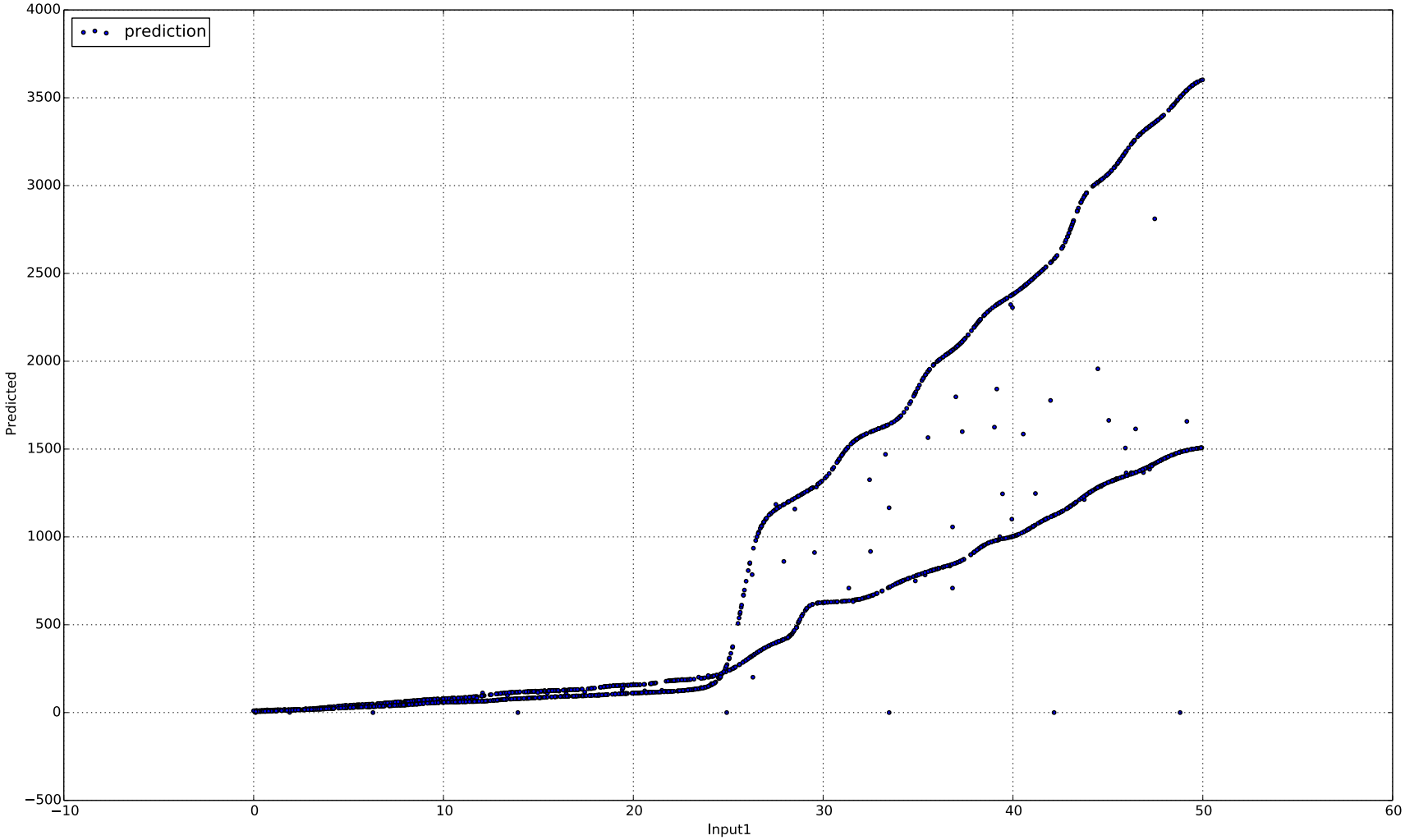
GPRegressionAvgMean_WS64 on **SYNTH_D_CD_2000_1_50_1_13**

Targets vs Predictions by the 5 selected online learners



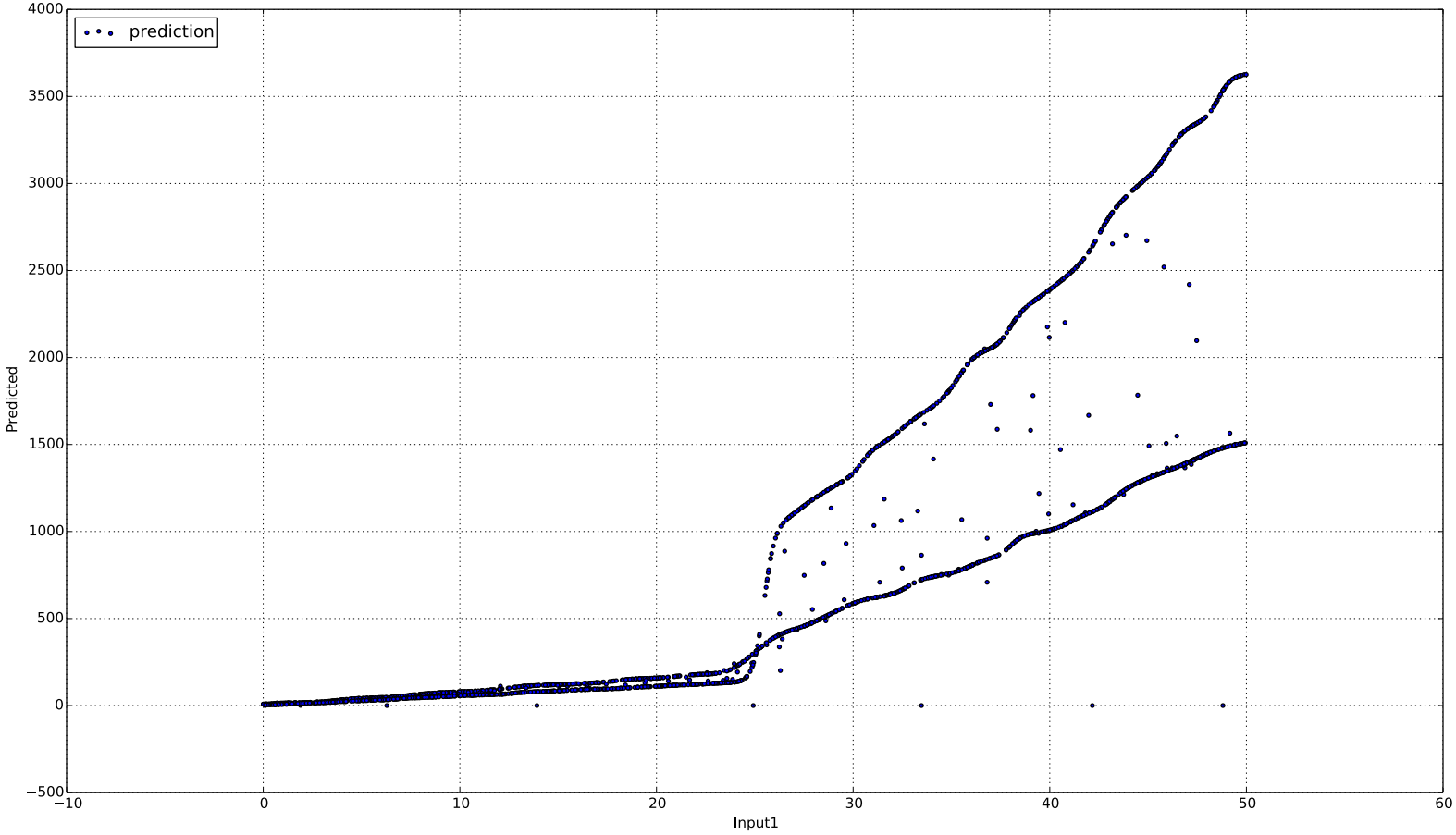
GPRegressionOLSMean_WS64 on **SYNTH_D_CD_2000_1_50_1_13**

Targets vs Predictions by the 5 selected online learners



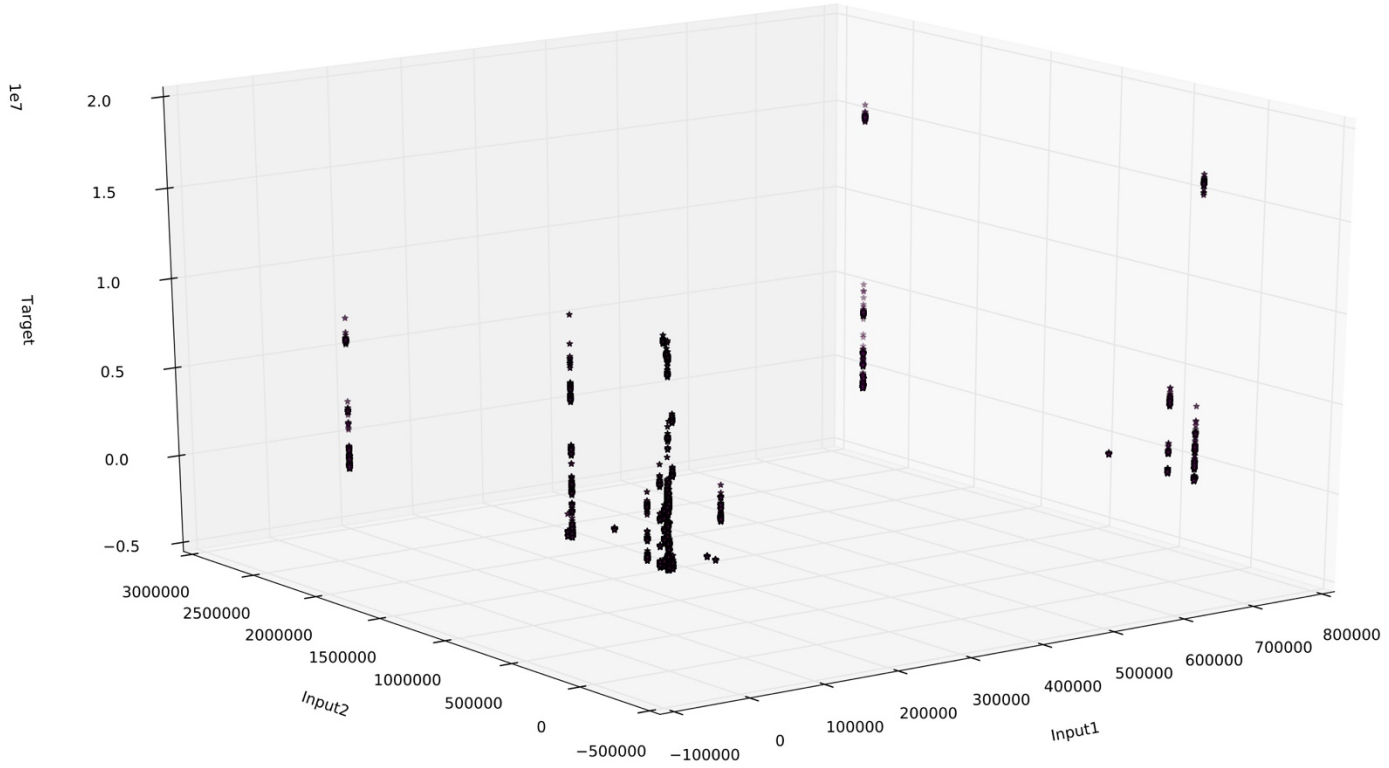
KernelRegression_WS64 on **SYNTH_D_CD_2000_1_50_1_13**

Targets vs Predictions by the 5 selected online learners



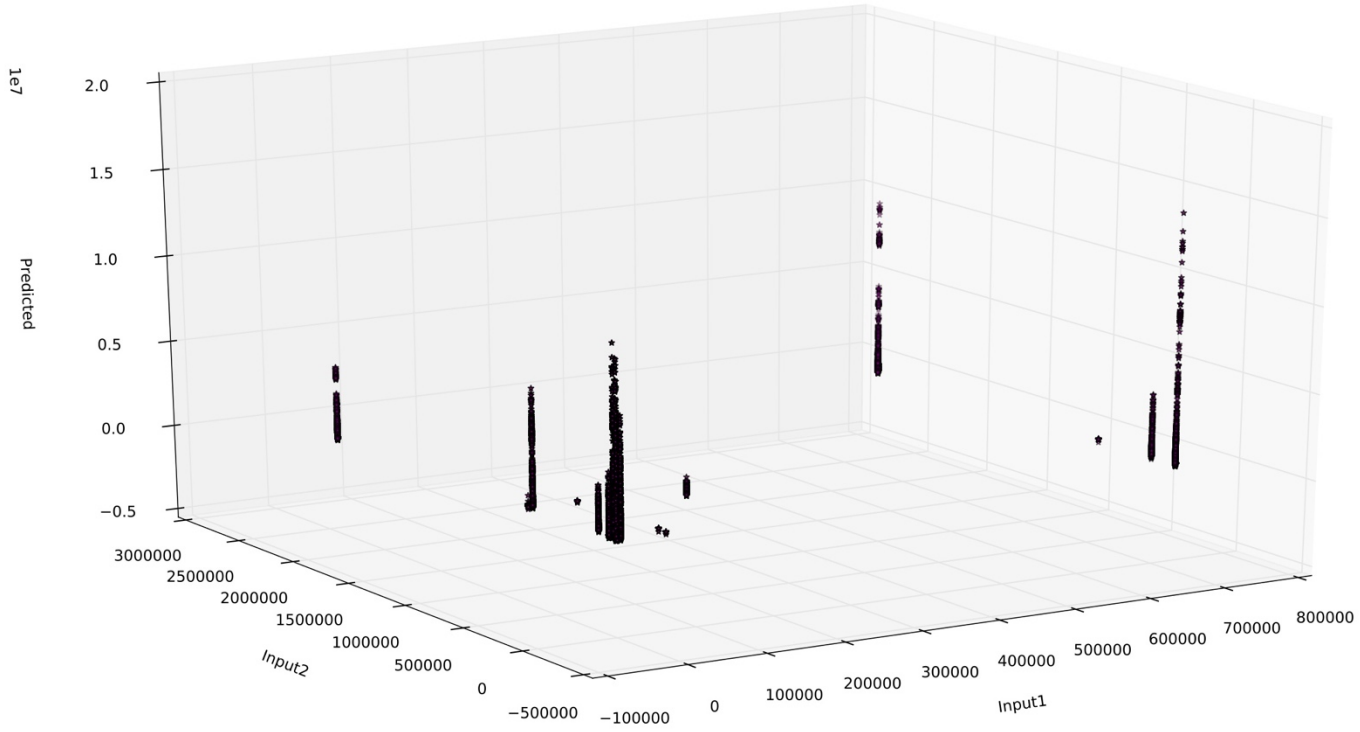
KernelRegression_WS96 on **SYNTH_D_CD_2000_1_50_1_13**

Targets vs Predictions by the 5 selected online learners



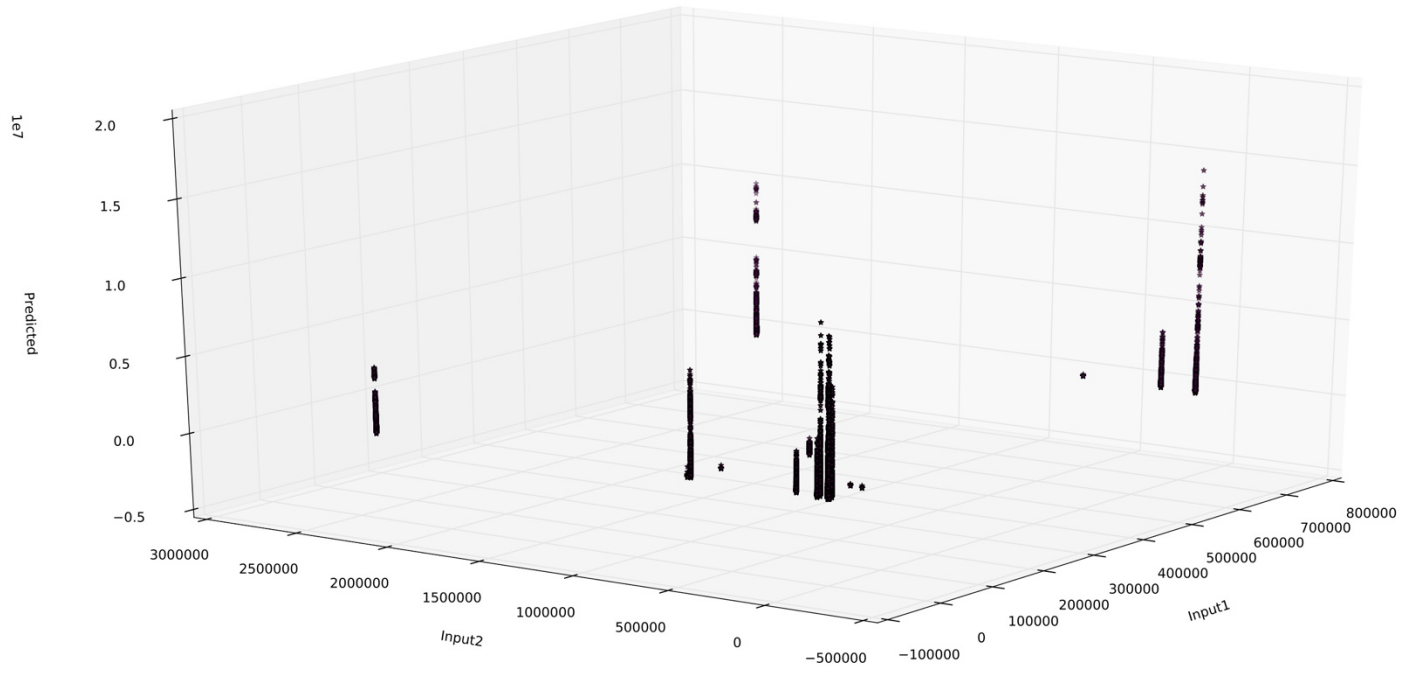
OCL_LEFTFETCHJOIN_GPU

Targets vs Predictions by the 5 selected online learners



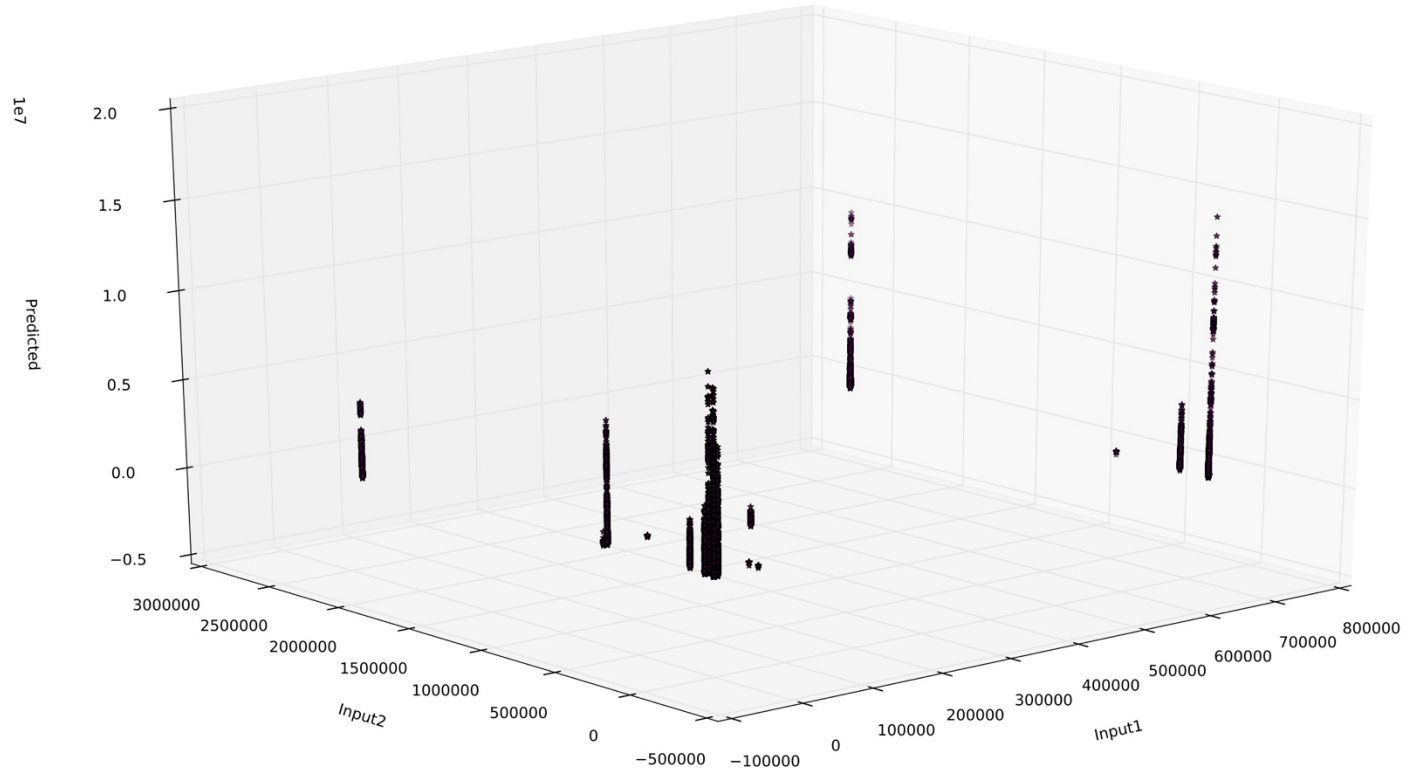
GPRegressionZeroMean_WS64 on **OCL_LEFTFETCHJOIN_GPU**

Targets vs Predictions by the 5 selected online learners



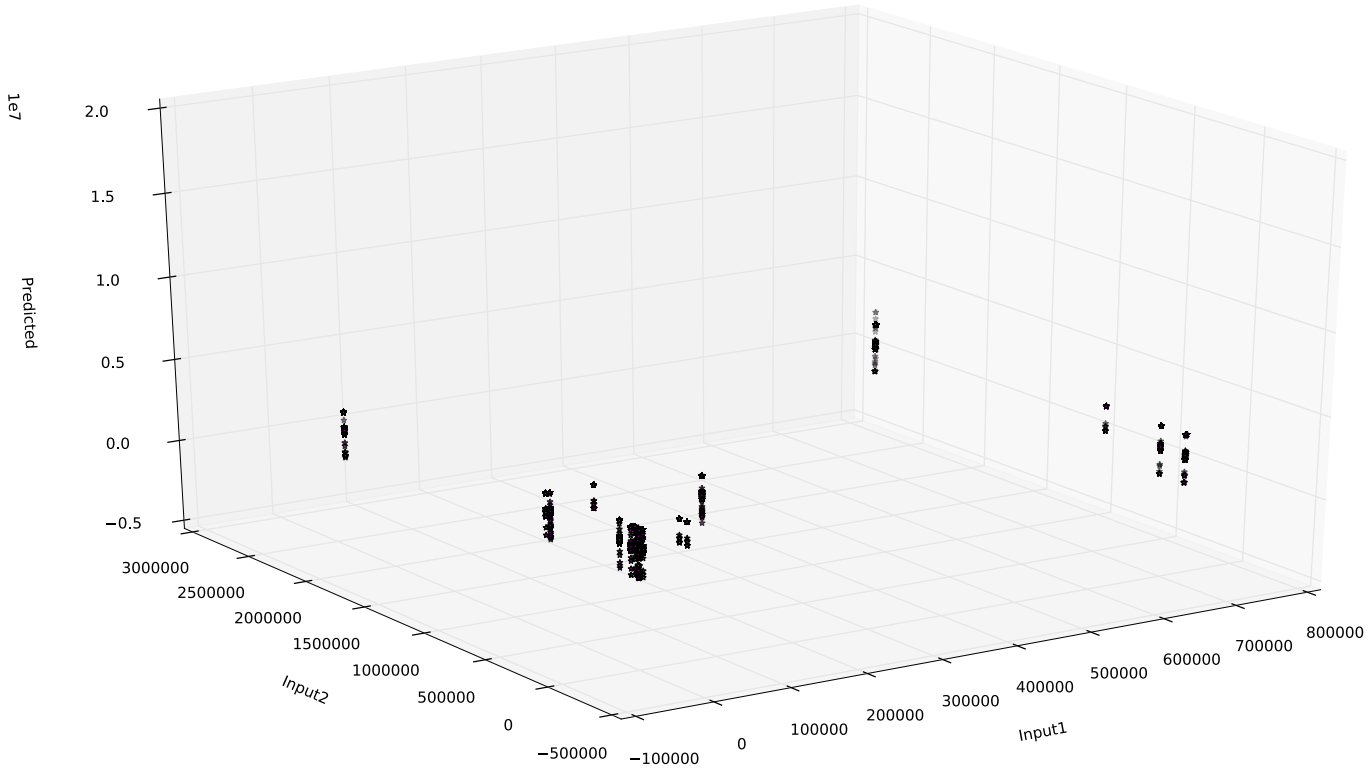
GPRegressionAvgMean_WS64 on **OCL_LEFTFETCHJOIN_GPU**

Targets vs Predictions by the 5 selected online learners



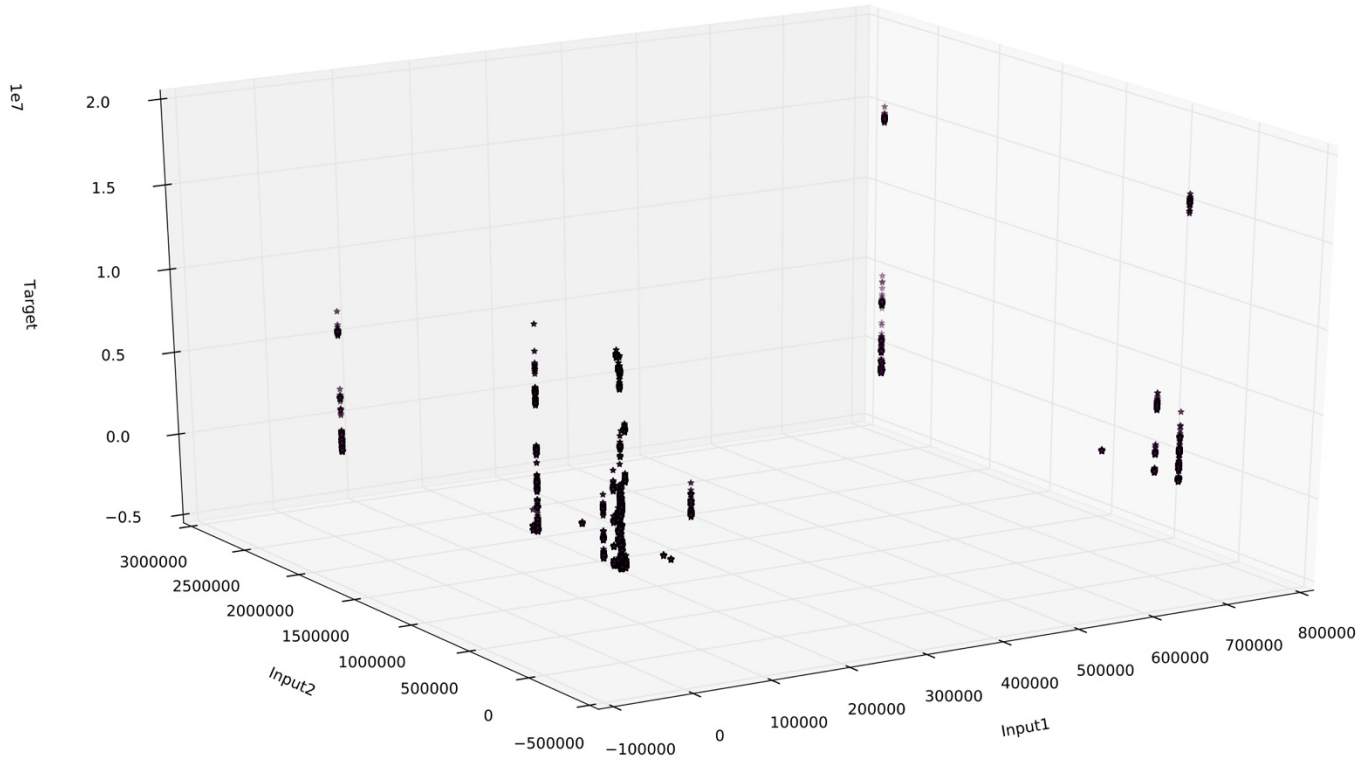
GPRegressionOLSMean_WS64 on **OCL_LEFTFETCHJOIN_GPU**

Targets vs Predictions by the 5 selected online learners



KernelRegression_WS64 on **OCL_LEFTFETCHJOIN_GPU**

Targets vs Predictions by the 5 selected online learners



KernelRegression_WS96 on **OCL_LEFTFETCHJOIN_GPU**

Conclusion

- Sliding-windowed non-parametric online learners are effective in learning from non-stationary data streams
- Sliding-window size has major implications on accuracy and time efficiency
- Forgetting-based parametric learners do not perform well on the drifting data
- Finding good hyperparameter configurations is a challenging task

Algorithm Comparison Table

	Accuracy	Prediction Bounds	Time Efficiency	Space Efficiency
MLE	+	+	+++	+++
MAP	+	+	+++	+++
GP Regression	+++	+++	+	++
Kernel Regression	+++	++	++	++

Future Work

- Online extension for Quantile Regression and SV-r
- Prediction Bounds Estimation mechanism for SV-r

References

Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, 2005. ISBN 0-262-18253-X.

Bernard W Silverman. Density estimation for statistics and data analysis, volume 26. CRC press, 1986.

Heimel, Max, et al. "Hardware-oblivious parallelism for in-memory column-stores." *Proceedings of the VLDB Endowment* 6.9 (2013): 709-720.

Joao Gama, Raquel Sebastiao, and Pedro Pereira Rodrigues. Issues in evaluation of stream learning algorithms. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '09, pages 329–337, 2009. doi: 10.1145/1557019.1557060.

Questions?

Thanks for listening! ☺