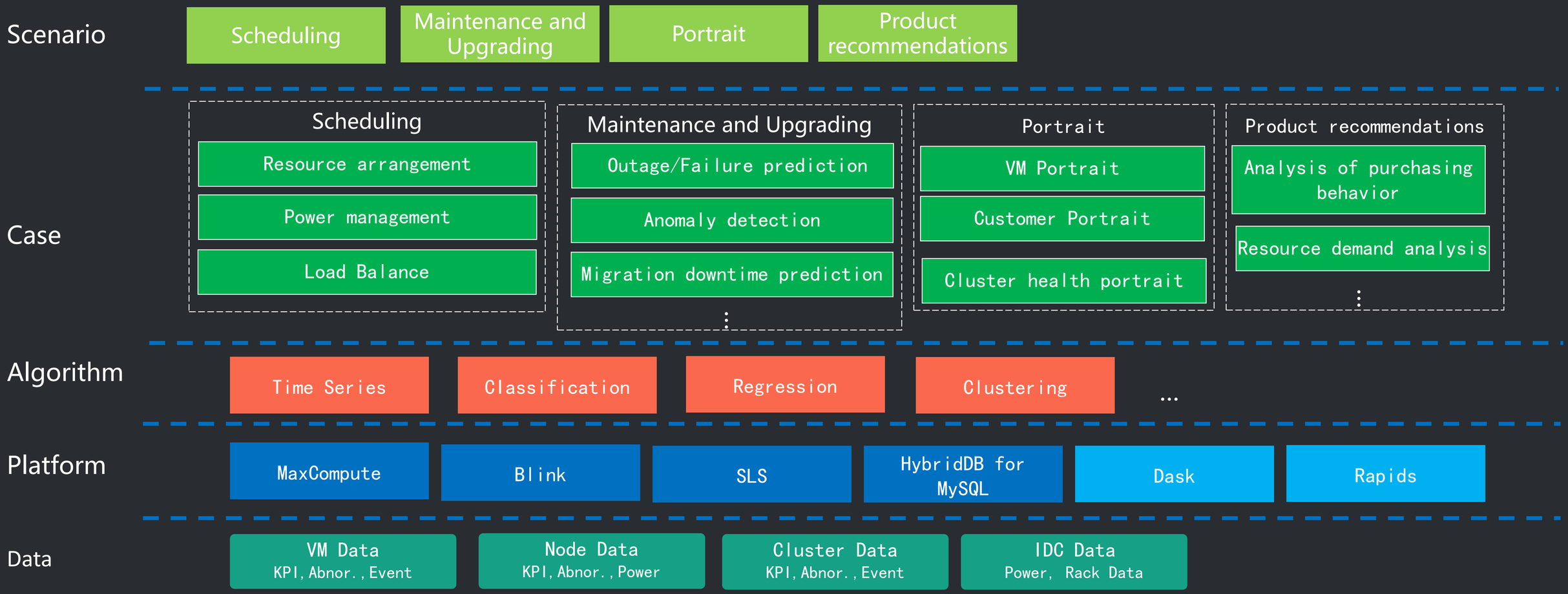
Alibaba Cloud
Official Cloud Services Partner

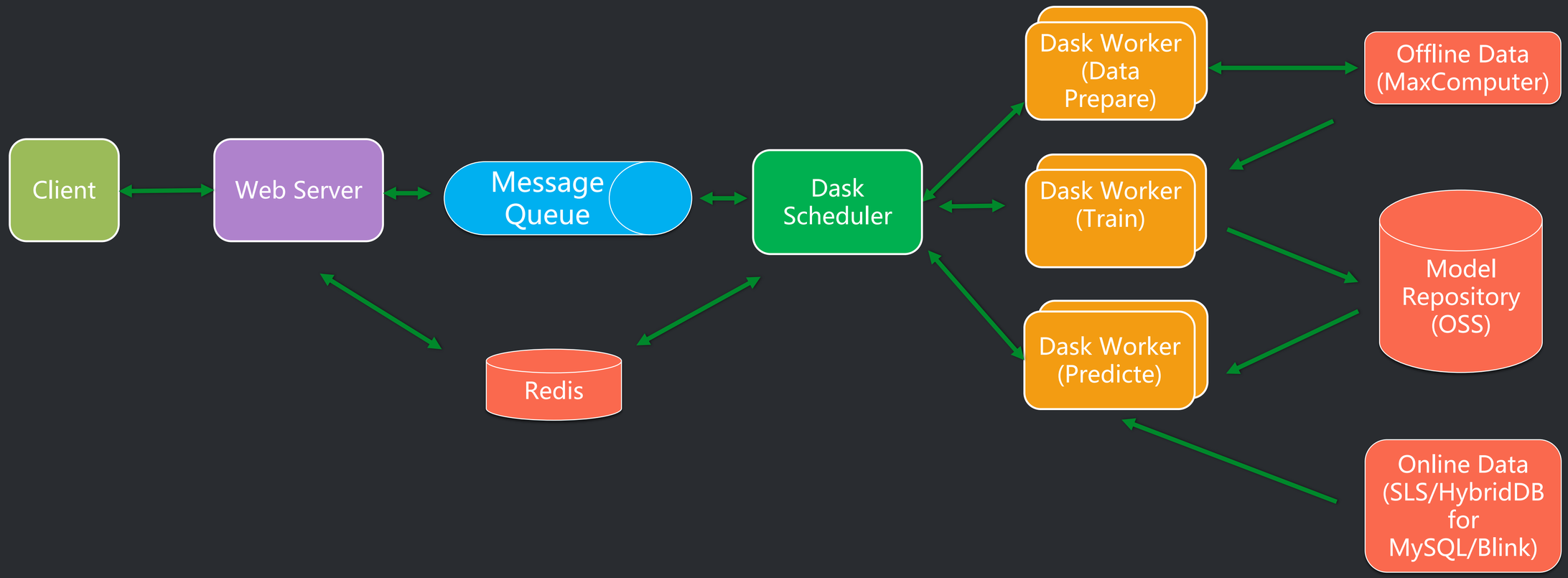
Intelligent Operation and Maintenance of Public Cloud Based on GPU-Accelerated Machine Learning

Feng Xie(stephen.xf@alibaba-inc.com)
Zhaowei Ouyang(zhaowei.oyzw@alibaba-inc.com)

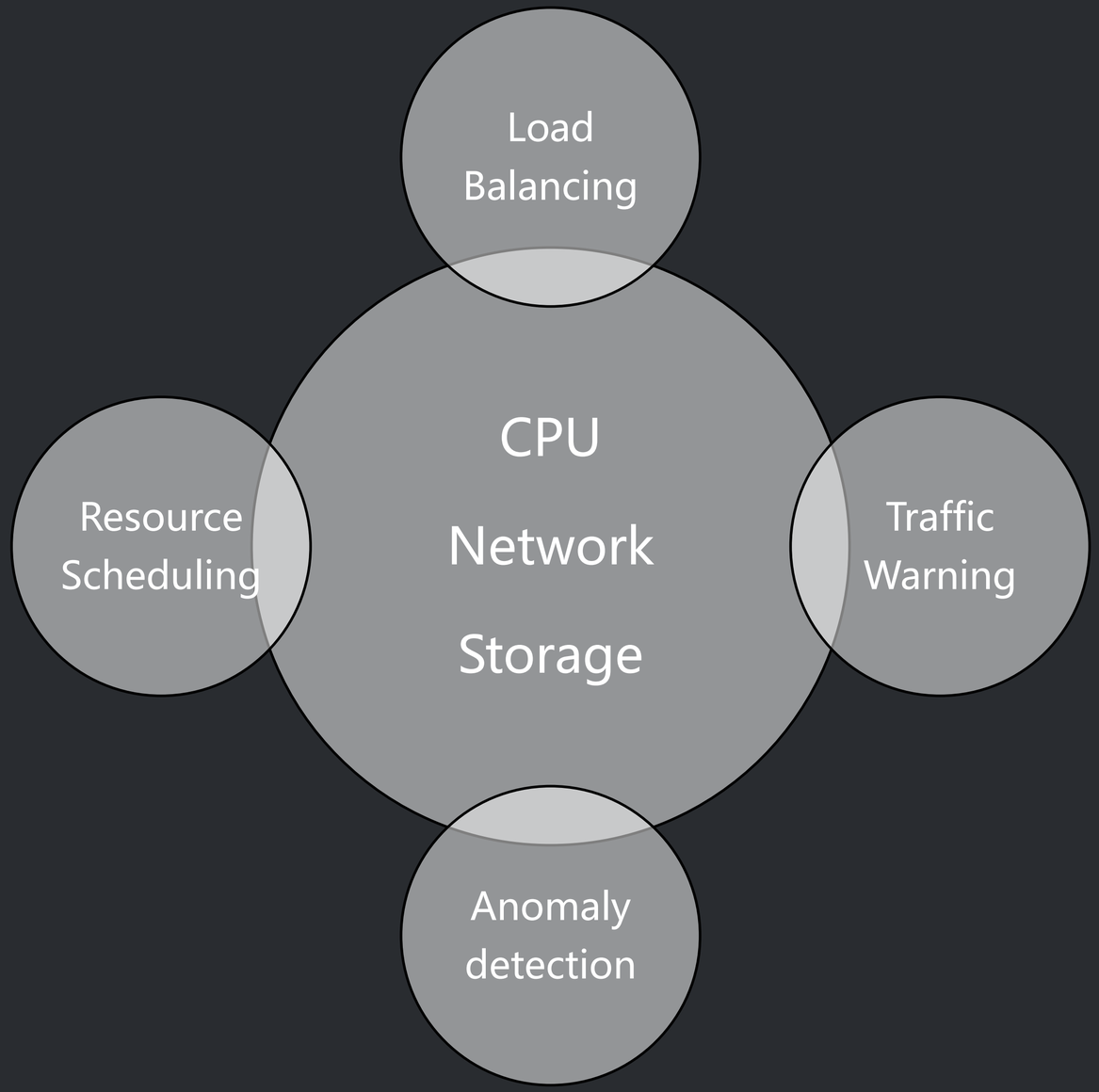
AIOps on Public Cloud



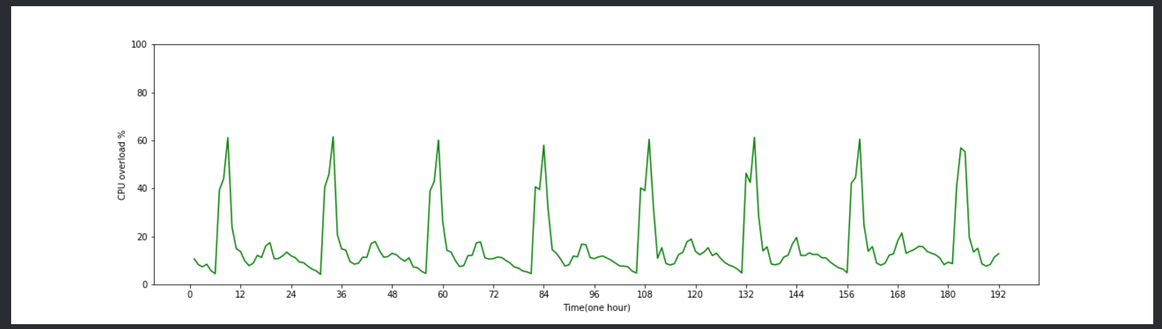
Machine Learning Platform Architecture

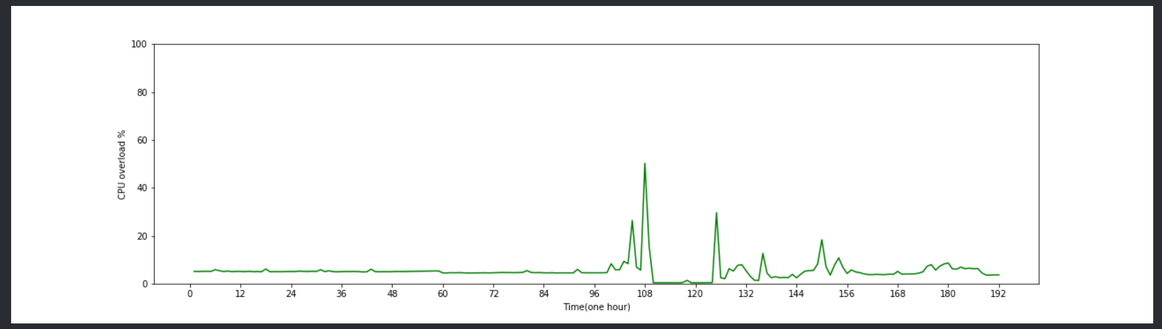


KPI Prediction

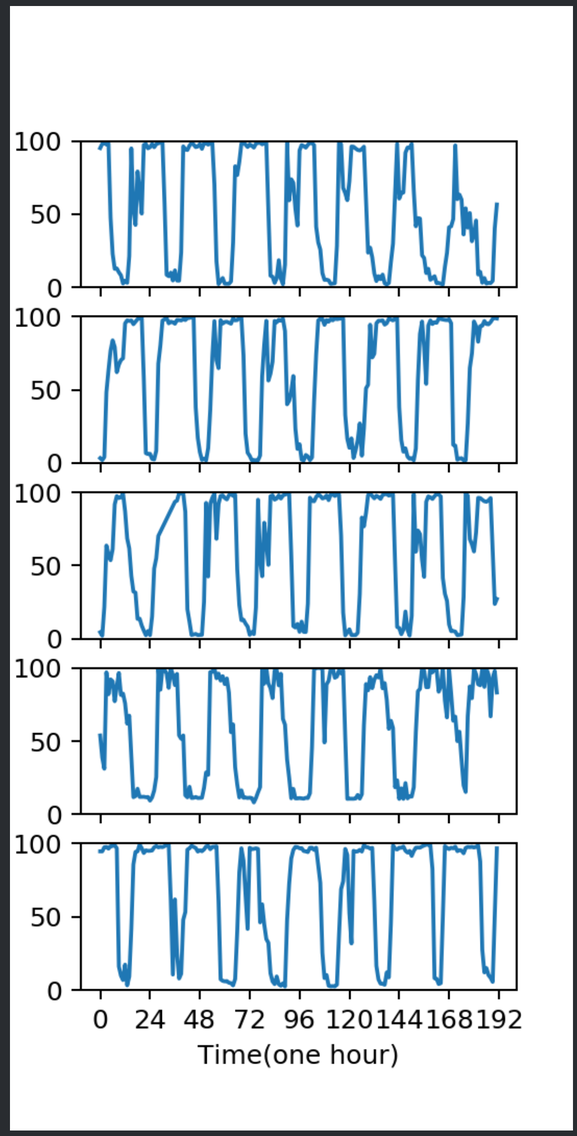


CPU Load Time Series



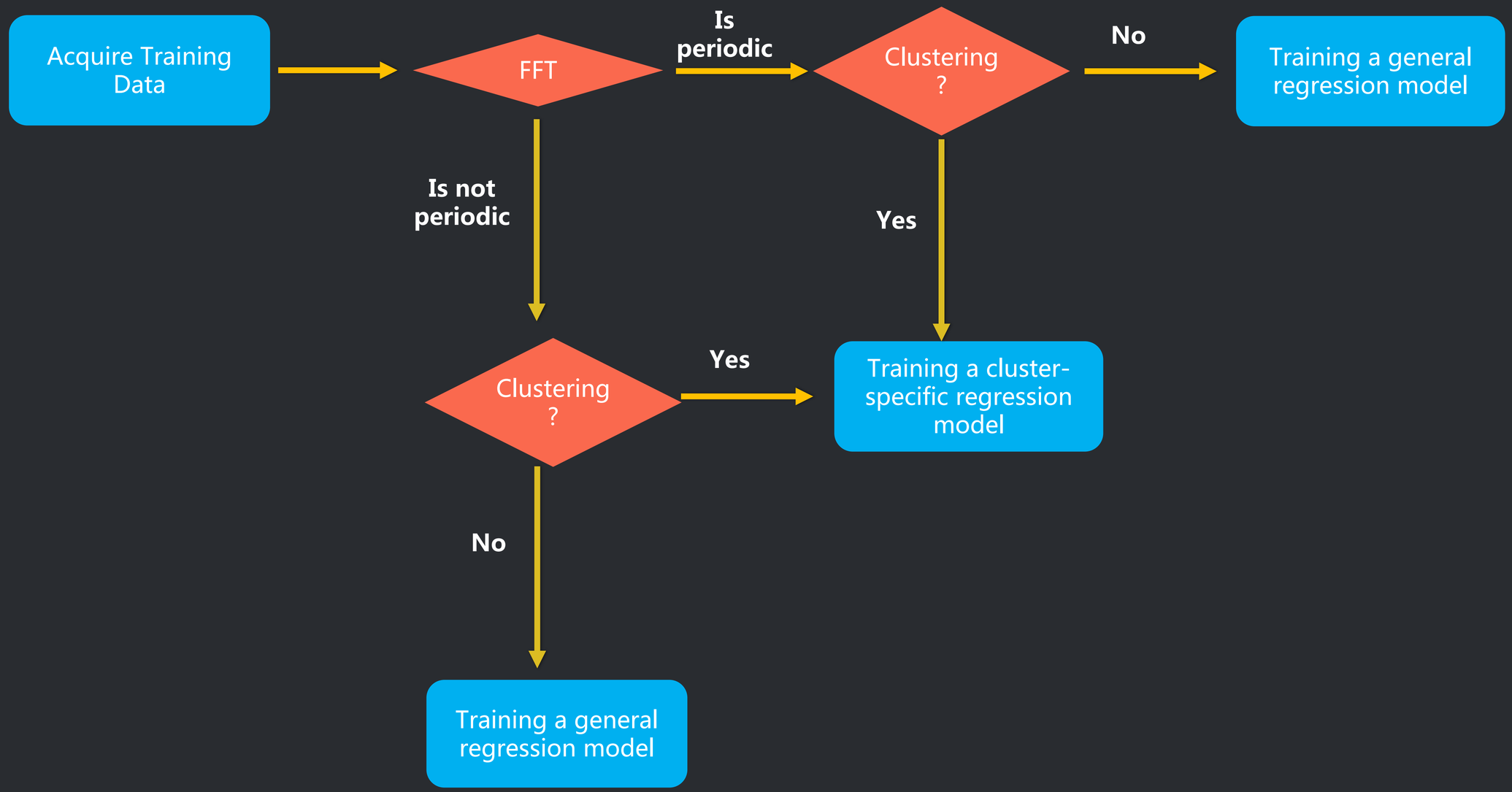


Periodicity

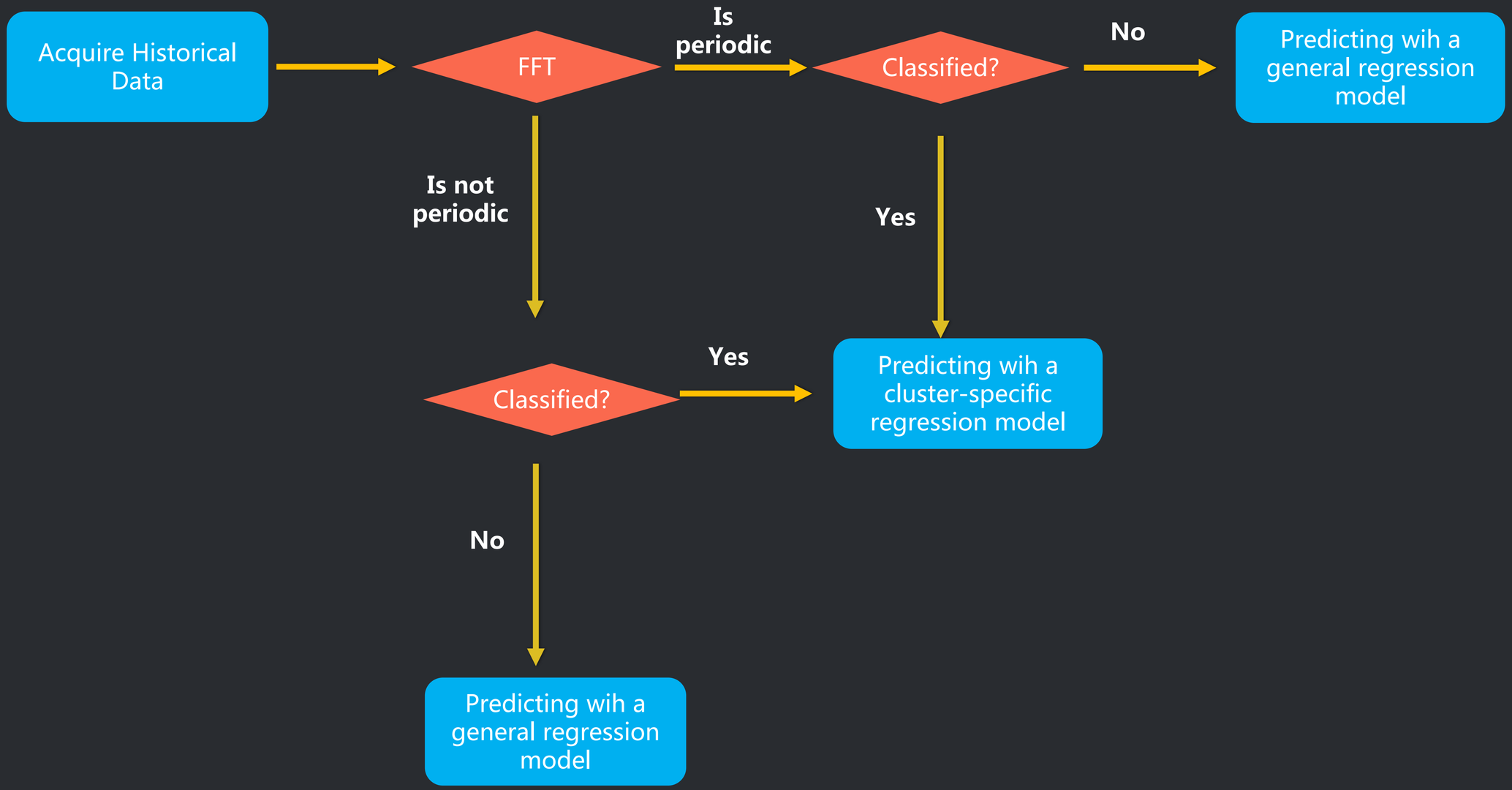


Similarity

Training Flow Chart



Predicting Flow Chart



Periodicity of Time Series

The Fast Fourier Transform (FFT) is used to transform the time series data from the time domain to the frequency domain. The frequency domain distribution is analyzed to determine whether it is periodic or not.

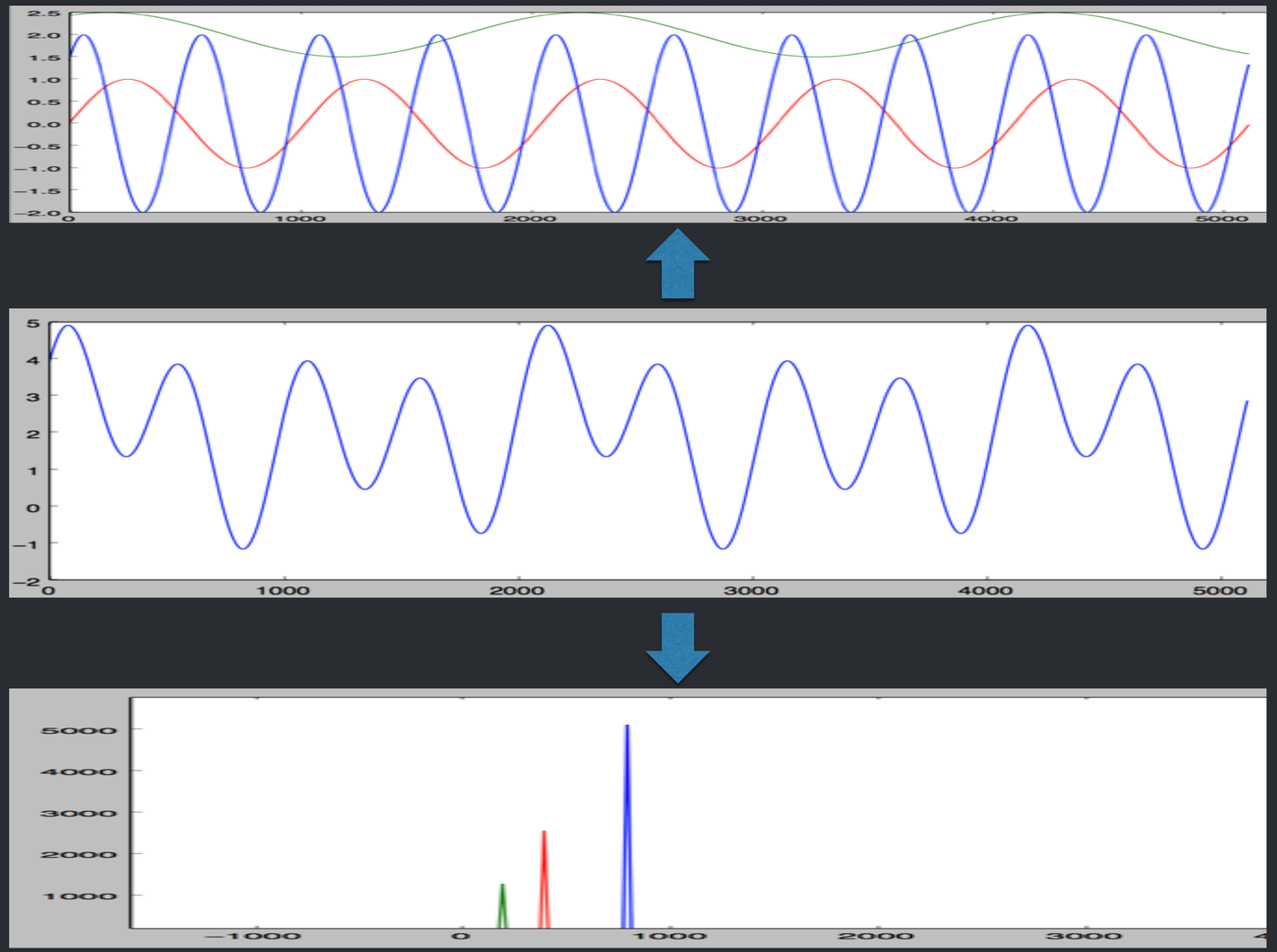
$$x_k = \sum_{n=0}^{N-1} x_n . e^{-i2\pi kn/N}$$

Basics of Signal Processing

Input time series

Original series

Frequency Domain



Similarity of Time Series

Use DTW distance as a measure of similarity between time series

$$\boldsymbol{M_{distance}} = \begin{pmatrix} 0 & d_{0,1} & \cdots & d_{0,n-1} & d_{0,n} \\ d_{1,0} & 0 & \cdots & d_{1,n-1} & d_{1,n} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ d_{n-1,0} & d_{n-1,1} & \cdots & 0 & d_{n-1,n} \\ d_{n,0} & d_{n,1} & \cdots & d_{n,n-1} & 0 \end{pmatrix}$$

$$\mathrm{K(i,j)} = e^{-\frac{d_{i,j}}{2\sigma^2}}$$

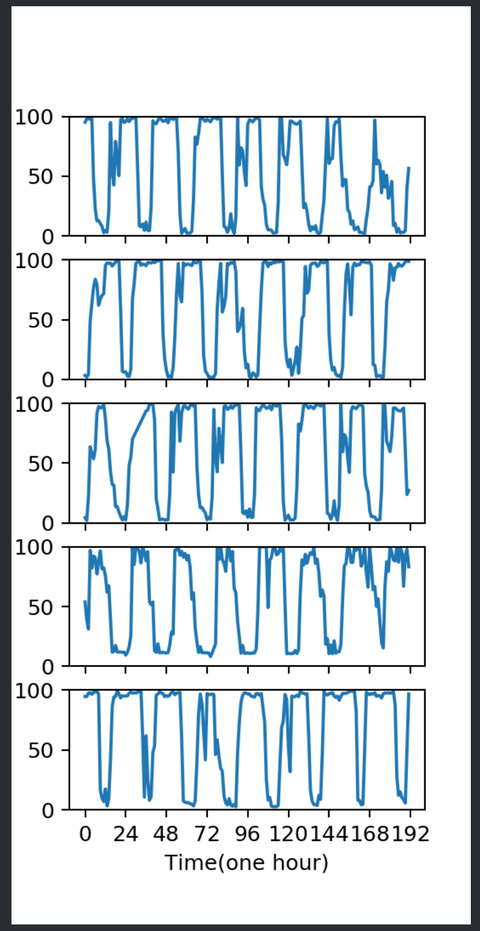
$$\boldsymbol{S_{similar}} = \begin{pmatrix} 1 & s_{0,1} & \cdots & s_{0,n-1} & s_{0,n} \\ s_{1,0} & 1 & \cdots & s_{1,n-1} & s_{1,n} \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ s_{n-1,0} & s_{n-1,1} & \cdots & 1 & s_{n-1,n} \\ s_{n,0} & s_{n,1} & \cdots & s_{n,n-1} & 1 \end{pmatrix}$$

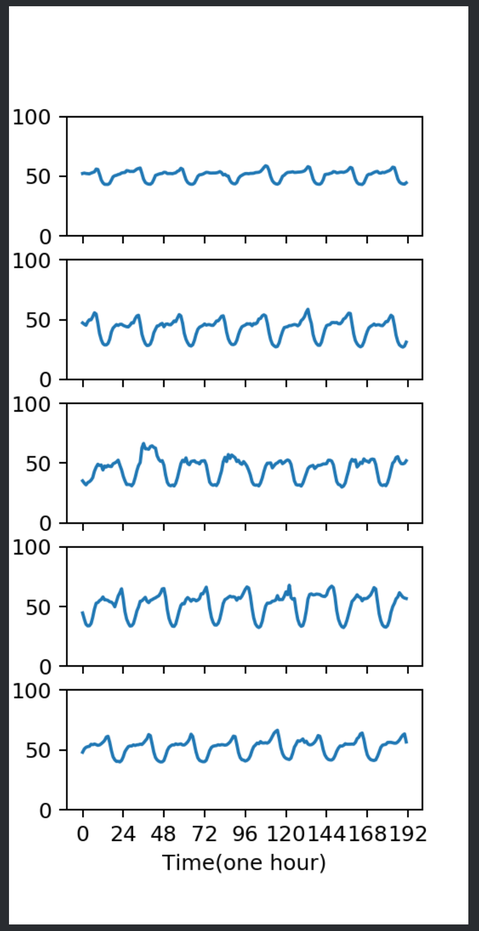
GPU-Accelearated FFT and DTW distance calculation

Use cuFFT to accelearate FFT calculation of massive time series data

The calculation of Dynamic Time Warping(DTW) distance
 is a task with high time and space complexity.
We can use the powerful parallel computing power of GPU to accelerate the calculation of DTW distance.

Clustering results



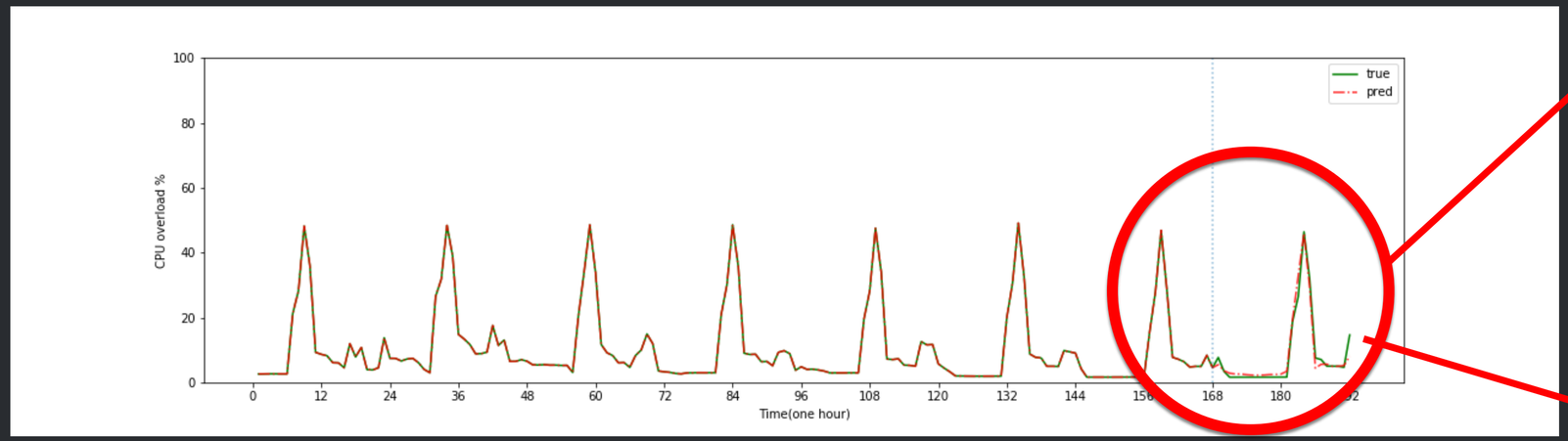


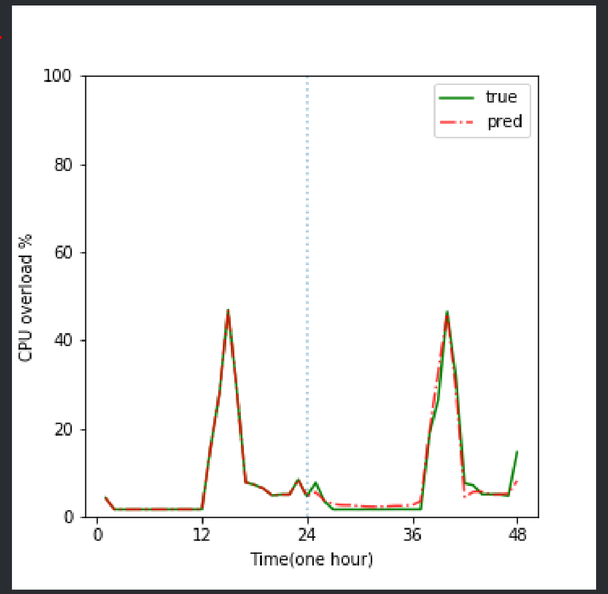
Time Series Regression Model

Model	Advantage	Disadvantage
ARIMA	Simple Hyperparameter Optimization	Low accuracy
LSTM	High accuracy	Complicated Hyperparameter Optimization Poor interpretability
XGBoost Regression Tree	High accuracy Good interpretability	Complicated Hyperparameter Optimization

Regression algorithm Result:XGBoost

Predict Next 24 Hours Result





Regression algorithm accuracy

Algorithm	RMSE	MAPE
ARIMA	70% <5	70% <0.5
XGB Regression tree	83% <5	83% <0.5

Migration downtime prediction

Use XGBoost Classification Tree to predict whether a VM is migration-sensitive

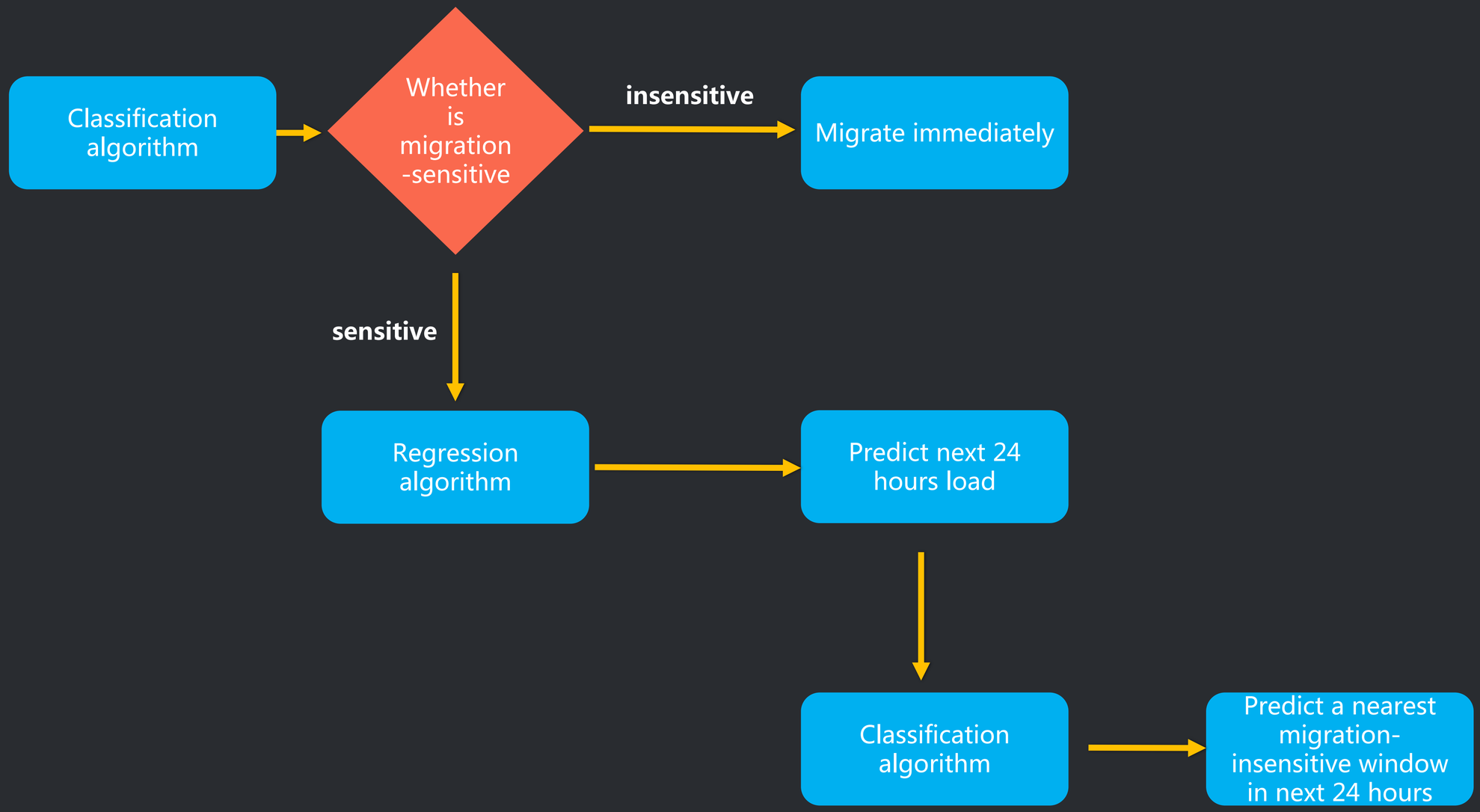
Feature

- Average vCPU utilization(1 hour before migration)
- Amplitude of fluctuation with vCPU utilization(one day before migration)
- VM Instance Type(How many vCPU/Memory?)
- ……

Result

- Migration-insensitive VM (downtime <= 100 ms)
- Migration-sensitive VM (downtime > 100 ms)

Migration Prediction Flow Chart

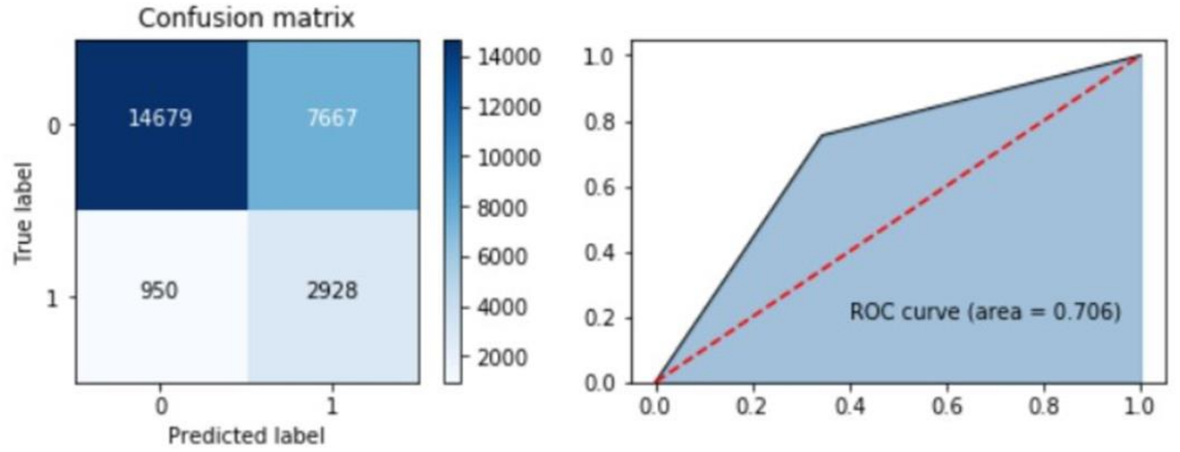


Classification Algorithm Accuracy:XGBoost

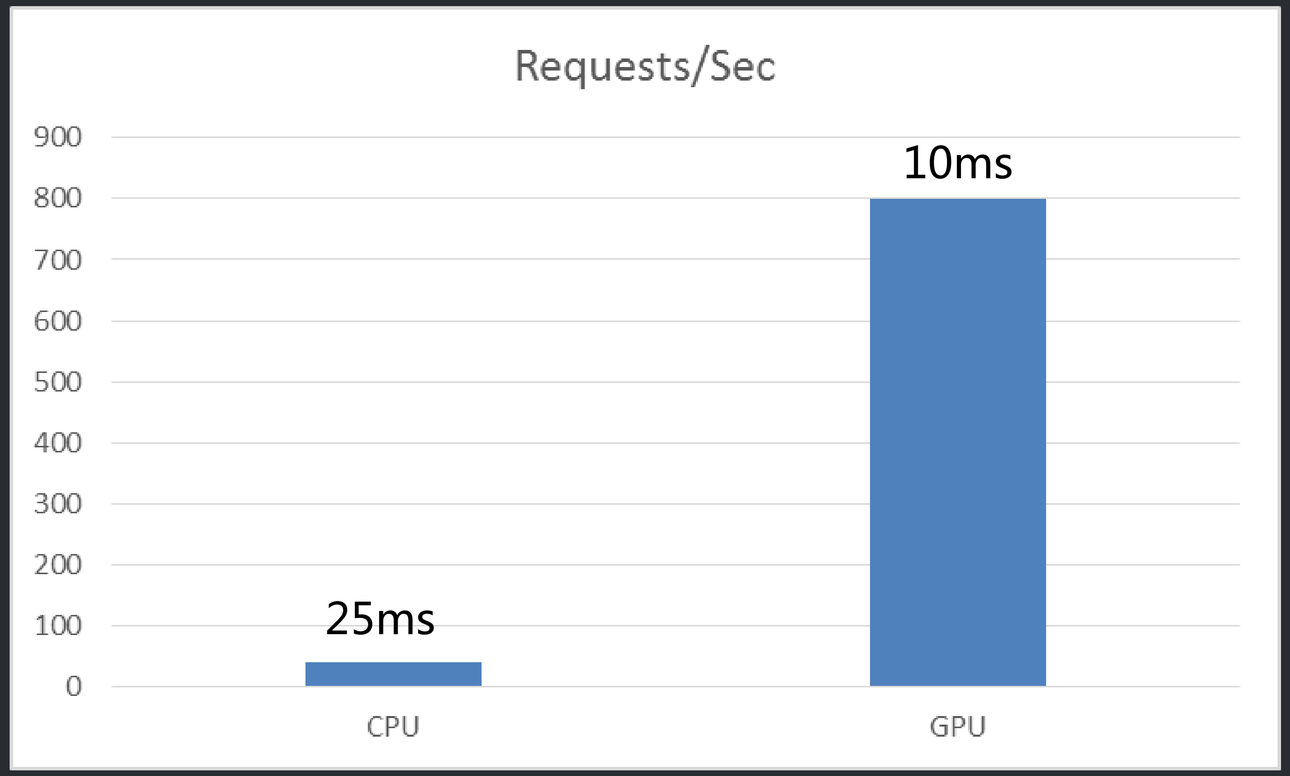
Accuracy ≈70%
Migration-sensitive Recal:76%

Classification_report:

	precision	recall	f1-score	support
0.0	0.94	0.66	0.77	22346
1.0	0.28	0.76	0.40	3878
micro avg	0.67	0.67	0.67	26224
macro avg	0.61	0.71	0.59	26224
weighted avg	0.84	0.67	0.72	26224



Classification Algorithm Performance:XGBoost



Latency:60% drop
Throughout:20x Speed-up

GPU:NVIDIA Tesla P100 * 8
CPU:2 Socket Intel Xeon E5-2682 v4 (Broadwell)

Questions?


Alibaba Cloud | MORE THAN JUST CLOUD